

REINVENTING THE
SOCIAL SCIENTIST AND HUMANIST IN
THE ERA OF BIG DATA

A PERSPECTIVE FROM SOUTH AFRICAN SCHOLARS

Susan **BROKENSHA**
Eduan **KOTZÉ**
Burgert A **SENEKAL**

REINVENTING THE
SOCIAL SCIENTIST AND HUMANIST IN
THE ERA OF BIG DATA

A PERSPECTIVE FROM SOUTH AFRICAN SCHOLARS

Susan **BROKENSHA**

Eduan **KOTZÉ**

Burgert A **SENEKAL**

sb **SUNBONANI
SCHOLAR**

Reinventing the Social Scientist and Humanist in the Era of Big Data:

A Perspective from South African Scholars

Published by Sun Media Bloemfontein (Pty) Ltd.

Imprint: SunBonani Scholar

All rights reserved

Copyright © 2019 Sun Media Bloemfontein and the authors

This publication was subjected to an independent double-blind peer evaluation by the publisher.

The author and the publisher have made every effort to obtain permission for and acknowledge the use of copyrighted material. Refer all inquiries to the publisher.

No part of this book may be reproduced or transmitted in any form or by any electronic, photographic or mechanical means, including photocopying and recording on record, tape or laser disk, on microfilm, via the Internet, by e-mail, or by any other information storage and retrieval system, without prior written permission by the publisher.

Views reflected in this publication are not necessarily those of the publisher.

First edition 2019

ISBN: 978-1-928424-36-9 (Print)

ISBN: 978-1-928424-37-6 (e-book)

DOI: <https://doi.org/10.18820/9781928424376>

Set in Garamond Pro 10.5

Cover design, typesetting and production by Sun Media Bloemfontein

Research, academic and reference works are published under this imprint in print and electronic format.

This printed copy can be ordered directly from: media@sunbonani.co.za

The e-book is available at the following link: <https://doi.org/10.18820/9781928424376>

Contents

| | |
|----------------------------------------------------------------------------------------------------------------|------------|
| Acknowledgements | i |
| Foreword | iii |
| Introduction | 1 |
| Chapter 1 The (fuzzy) origins of big data and the dangers of ignoring history | 6 |
| 1.1 A messy affair | 6 |
| 1.2 The history of (big) data storage | 7 |
| 1.3 The emergence of statistical analysis | 10 |
| 1.3.1 <i>Big business, big data</i> | 12 |
| 1.4 The digital revolution and events surrounding big data | 13 |
| 1.5 Big data's history lessons | 16 |
| 1.5.1 <i>Revolution versus evolution</i> | 16 |
| 1.5.2 <i>The past in big data</i> | 16 |
| 1.5.3 <i>Two foundational narratives</i> | 18 |
| Chapter 2 Locating big data in the (digital) humanities and (computational) social sciences | 19 |
| 2.1 How big data is framed | 20 |
| 2.1.1 <i>Two dominant metaphors</i> | 21 |
| 2.1.2 <i>A collaborative effort</i> | 23 |
| 2.2 Digital humanists and computational social scientists | 24 |
| Chapter 3 Big Data, big despair: Myths debunked and lessons learned | 29 |
| 3.1 Epic fails | 29 |
| 3.2 Big data lessons | 30 |
| 3.2.1 <i>Lesson 1</i> | 30 |
| 3.2.2 <i>Lesson 2</i> | 34 |
| 3.2.3 <i>Lesson 3</i> | 36 |
| 3.2.4 <i>Lesson 4</i> | 37 |
| 3.2.5 <i>Lesson 5</i> | 40 |
| 3.2.6 <i>Lesson 6</i> | 42 |
| 3.2.7 <i>Lesson 7</i> | 43 |

| | | |
|------------------|------------------------------------------------------------------------------------------------|-----------|
| Chapter 4 | Big Data needs big ethics | 45 |
| 4.1 | Big data <i>faux pas</i> of the millennium | 45 |
| 4.2 | A review of the literature | 47 |
| | 4.2.1 <i>Current challenges</i> | 47 |
| | 4.2.2 <i>The controversy surrounding human subjects</i> | 49 |
| | 4.2.3 <i>The public-private space conundrum</i> | 50 |
| | 4.2.4 <i>The culture of informed consent on the Internet and anonymisation</i> | 52 |
| | 4.2.5 <i>Ethics and the problem of representativeness</i> | 53 |
| | 4.2.6 <i>A new digital ecosystem</i> | 54 |
| 4.3 | Data justice | 55 |
| Chapter 5 | Does big data visualisation make our endeavours less humanistic? | 57 |
| 5.1 | Data visualisation, human cognition, and human perception | 58 |
| 5.2 | Technology and the humanities | 64 |
| 5.3 | Data as <i>capta</i> | 66 |
| 5.4 | The emotional and social pitfalls of visualisation | 69 |
| | Visualisation and the problem of data power | 71 |
| Chapter 6 | Data power in the era of big data: Friend or foe? | 72 |
| 6.1 | Big data's shadow side | 72 |
| 6.2 | Engaged humanists and social scientists | 75 |
| 6.3 | Big data as an obstacle/bridge to humanitarian projects | 83 |
| 6.4 | Size revisited | 84 |
| Chapter 7 | The place of qualitative data analysis software (QDAS) programmes in a big data world | 87 |
| 7.1 | Software programmes and the qualitative researcher | 87 |
| 7.2 | Two ways of thinking about QDAS | 88 |
| 7.3 | QDAS and blended reading | 92 |
| 7.4 | QDAS, qualitative content analysis, and big data | 95 |
| 7.5 | Beyond traditional databases | 96 |

| | | |
|----------------------|---------------------------------------------------------------------------------------------------|------------|
| Chapter 8 | The nitty-gritty: Big data infrastructure | 98 |
| 8.1 | Managing the unmanageable | 98 |
| 8.2 | Big data systems | 100 |
| 8.3 | Data generation | 101 |
| 8.4 | Data acquisition | 103 |
| 8.5 | Data storage | 104 |
| | 8.5.1 <i>Distributed file systems</i> | 104 |
| | 8.5.2 <i>NoSQL databases</i> | 105 |
| | 8.5.3 <i>Programming models</i> | 107 |
| 8.6 | Data analysis | 107 |
| 8.7 | The Hadoop ecosystem | 110 |
| | 8.7.1 <i>Hadoop's core components</i> | 110 |
| 8.8 | The Hadoop software stack | 112 |
| 8.9 | An example of a Hadoop big data system | 115 |
| 8.10 | Commercial big data systems and cloud big data | 116 |
| 8.11 | A reminder | 117 |
| Chapter 9 | Leveraging social scientific and humanistic expertise in the world of (big) data science | 118 |
| 9.1 | What is data science? | 119 |
| 9.2 | Marrying (big) data science and the humanities and social sciences | 124 |
| Chapter 10 | An example: Big data analysis in the humanities in South Africa | 126 |
| 10.1 | Introduction | 126 |
| 10.2 | Identifying the problem | 126 |
| 10.3 | Data gathering | 127 |
| 10.4 | Text pre-processing | 131 |
| 10.5 | Performing analytics on the data | 132 |
| | 10.5.1 <i>Introduction to sentiment analysis</i> | 132 |
| | 10.5.2 <i>Applying sentiment analysis</i> | 133 |
| | 10.5.3 <i>Sentiment analysis results</i> | 134 |
| 10.6 | Visualising the results | 135 |
| 10.7 | Conclusion | 140 |
| The last word | | 141 |
| References | | 142 |
| Index | | 187 |

List of figures

| | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| Figure 1.1: Milestones in the era of big data since the introduction of the first commercial microprocessor | 15 |
| Figure 3.1: Cai and Zhu's (2015:7) quality assessment process | 33 |
| Figure 5.1: Cape Town's average maximum temperatures in degrees Celsius | 60 |
| Figure 5.2: Semantically incongruent colour choices to represent temperatures | 60 |
| Figure 5.3: The Stroop effect | 60 |
| Figure 5.4: Gestalt principle of similarity | 62 |
| Figure 5.5: Gestalt principle of proximity | 62 |
| Figure 5.6: Gestalt principle of enclosure | 62 |
| Figure 5.7: Example of analogical reasoning (adapted from Daugherty & Mentzer 2008:10) | 64 |
| Figure 5.8: John Snow's map of cholera outbreaks in London (drawn by Snow circa 1854, and taken from Stamp's (1964) <i>The geography of life and death</i>) | 68 |
| Figure 5.9: John Snow's map reinvisaged (Drucker 2011:19, with credit to Xárene Eskandar for the graphic) | 68 |
| Figure 7.1: An example of a word cloud from South Africa's Life Esidimeni Arbitration Hearings, 24 and 25 January 2018 | 90 |
| Figure 8.1: The Hadoop ecosystem (adapted from Hu, Wen, Chau & Li 2014:678) | 112 |
| Figure 8.2: A big data system (adapted from Ibarra 2012:3) | 115 |
| Figure 9.1: The skills and knowledge of a data scientist | 122 |
| Figure 9.2: The various tasks performed by a data scientist | 122 |
| Figure 10.1: A comparison of sentiment classifiers | 135 |
| Figure 10.2: Tweets about the Afrikaner by country | 136 |
| Figure 10.3: Examples of negative tweets | 137 |
| Figure 10.4: Google Image search of 'South African squatter camps' | 138 |
| Figure 10.5: Negative tweets from within South Africa | 139 |

List of tables

| | |
|------------------------------------------------------------------------------------------------|-----|
| Table 2.1: Two approaches to big data analysis (O’Sullivan 2017:15) | 28 |
| Table 3.1: Cai and Zhu’s (2015:5) big data quality framework | 32 |
| Table 4.1: Measuring ethics variation in digital research (Jang & Callingham 2012:75) | 51 |
| Table 10.1: The use of the word ‘Afrikaner’ in some sample languages..... | 129 |
| Table 10.2: Language distribution of tweets | 130 |
| Table 10.3: Results of the sentiment analysis (without re-tweets, n = 4505) | 134 |

Acknowledgements

We would like to express a special thanks of gratitude to the University of the Free State for an Interdisciplinary Research Grant that enabled us to undertake the big data project. We are also indebted to Theo du Plessis and to Lani de Lange for their unwavering support over the last few years. Thank you too to our family members and friends for their steadfast encouragement.

Foreword

The mantra of “Big Data” – the two capital letters in this expression epitomise the massive nature of the data involved – has increasingly gained traction in recent years. When this mantra first appeared, it had the aura of academic disciplines, and almost every sphere of business began dipping into the sea of big data. During that trail-blazing period, most scholars in the social and human sciences, owing largely to their academic training, felt immobilised and technologically hapless at the prospect of big data. This point is poignantly contextualised in the current book at the beginning of Chapter 2: “The very notion of big data creeping into their research spaces casts an intimidating shadow over traditional humanists and social scientists, who may fear human behaviour being reduced to mere mathematical models” (p. 19). Much information about the latter would emerge from the Internet and from big commercial publishing houses as is still the case even now. Again, during that period, there was not the slightest chance that any scholar from these two cognate sciences, least of all scholars from the South African context, would ever dabble in the nascent field of big data. But not anymore. This is the background against which the current book, aptly titled, “Reinventing the Social Scientist and Humanist in the Era of Big Data: A Perspective from South African Scholars”, should be viewed. Not only is this book’s title apposite, but the putting together of the book itself is a fitting and welcome scholarly event in the South African higher education ecosystem, and more so given the related mantra of the Fourth Industrial Revolution (4IR) – of which big data is an integral part (see Chaka 2019; Chaka forthcoming) – which is gaining currency in the South African higher education system.

Given the points highlighted above, a pertinent question to pose is: what role does big data have to play in the social and human sciences? Before delving into and providing a bespoke answer that the book attempts to offer, I venture to say that big data has a role to play in every aspect of the social and human sciences. In the main, the book is “about big data aimed specifically at academics in the humanities and social sciences” (p. 1) and is written by three scholars from South Africa, and dare I add, by three scholars from the global South. Overall, the book boasts ten chapters whose lowdown has been eloquently captured in the Introduction. One of the telling points of the book is its brutal honesty about the fact that despite concerted efforts by South African scholars to create a nexus between big data and the digital humanities (DH), the applications of big data in the “traditional humanities and social sciences” (p. 3) are, at best, few and far between, and at worst, “an alien phenomenon at local universities” (p. 3). In fact, argues the book, at these universities, big data analytics is almost the exclusive preserve of computer science disciplines.

Launching its first chapter by delineating the fuzziness that characterises the genealogy of big data and by mapping a historical trajectory of “(big) data storage” (p. 7), the book meticulously provides its well-articulated successive themes and sub-themes in an integrated whole. A few examples are “Two dominant metaphors” (p. 21); “Digital humanists and computational social scientists” (p. 24); “Big Data, big despair: Myths debunked and lessons learned” (p. 29); “The nitty-gritty: Big data infrastructure” (p. 98); “The Hadoop ecosystem” (p. 110); and “Marrying (big) data science and the humanities and social sciences” (p. 124). It then rounds off its discussion by presenting a real-world big data-based chapter to demonstrate how some of the aspects of big data may be applied to tweets (about the word *Afrikaner*) harvested from Twitter as one example of a big data generating platform. The book could not have come at a better moment for advancing big data scholarship for human and social sciences and for experimenting with *datafication* (Chaka 2019) as it pertains to these two cognate sciences.

Chaka Chaka
December 2019

Introduction

Will large quantities of data transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means?

– Danah boyd and Kate Crawford (2012:663)

One of the authors of this book encountered the concept of a ‘data scientist unicorn’ for the first time in a position article on (big) data analytics in 2015.¹ She was intrigued by the romantic notion that a data scientist was expected to be a *wunderkind* in the sense of being exceptionally well versed in all the myriad areas that encompass data science. Her fascination with the idea of a data scientist as being some kind of mythical creature stemmed partly from the fact that she is a linguist housed in an English department in a humanities faculty; as a linguist she is interested, amongst other things, in the (covert) effects of metaphorical framing of ideas, practices, and disciplines that are regarded as emergent. Since data scientists are increasingly working in the arena of big data, she also began to explore how big data itself is conceptualised by the mass media and academics, and what she found was that this phenomenon is variously described as “portentous” (Lupton 2015:2) and “perverse”, (Lupton 2015:2) or as “gold” (van Dijck 2014:199) and “oil” (Dean 2014:9). As insightful as these descriptions were, they did not tell her anything about the ontological attributes of big data, and so she turned to two colleagues employed at the same institution for answers. One of these colleagues is also a humanist, but with a great deal of experience in using big data tools to conduct research in the fields of complex networks and information technology; the other is a data scientist working towards developing technologies for extraction and analysis of sentiment from microblogging sites by employing a combination of natural language processing, text mining, and machine learning techniques. As these three academics began to discover some of the advantages of big data research, they also began to wonder if and how big data could be harnessed by traditional humanists and social scientists in South Africa, particularly in light of the fact that many of these scholars remain sceptical about its ultimate import. It is through countless debates and emails that they decided to write a book about big data aimed specifically at academics in the humanities and social sciences.

Big data has changed virtually every aspect of society over the last two decades. In business, big data has facilitated better marketing campaigns and monitoring of sales and manufacturing processes (Davenport 2014). Most large international corporations such as

1 ‘Chasing the data science unicorn’ by David Stodder (<https://tdwi.org/articles/2015/01/06/chasing-the-data-science-unicorn.aspx>).

Amazon, Google, Facebook, Coca-Cola and Walmart use big data in some way or another, as do major banks and financial institutions.

Governments have also seized opportunities created by big data. In March 2012, the United States (US) government launched their Big Data Research and Development Initiative, involving various agencies in infrastructure development to store, manage, and analyse large-scale data (Lazar 2012:47; Chen, Mao & Liu 2014:175). Big data also played an important role in Barack Obama's election campaign to rally individual voters (Jin, Wah, Cheng & Wang 2015: 61). Jin *et al.* (2015:60) anticipate "that future economic and political competitions among countries will be based on exploiting the potential of big data, among other traditional aspects. In short, the research and applications of big data are of strategic importance and significance for improving the competitiveness of any country". The Defense Advanced Research Projects Agency (DARPA) – responsible for the creation of the Internet – is one of the organisations once again involved in this technological development through one of their programmes called ADAMS (Anomaly detection at multiple scales). The US Department of Defense has a programme called MAPD (Mathematics for the analysis of petascale data) (Lazar 2012:48), which has been developed to extract insights from huge scientific datasets. The National Security Agency (NSA) in the US exploits big data through its Planning Tool for Resource Integration, Synchronization and Management (PRISM) programme, while the United Kingdom (UK) does the same through Tempora (Lyon 2014:2). Even the United Nations – which Davenport (2014:17) notes is not known for technological innovation – has a big data programme called HunchWorks.

In science, big data has resulted in what some call the fourth paradigm² (Park & Leydesdorff 2013:757; Abreu & Acker 2013:549; Hitzler & Janowicz 2013:233). Big data has become so important that journals have emerged over the past two decades that focus specifically on this field. These include *Annals of Data Science*, *International Journal of Data Science and Analytics*, *EPJ Data Science*, *GigaScience*, and *Big Data & Society*.

From the outset, it should be noted that this book neither hails big data as the game changer of the twenty-first century nor dismisses the phenomenon outright. Instead, it interrogates the multiple facets of and (controversial) issues surrounding big data with specific reference to the humanities and social sciences. Amongst other things, the book challenges the notion that big data is a monolith; asks critical questions about its epistemology; and delves into important aspects of the phenomenon related to its so-called objectivity and accuracy, emerging ethical concerns surrounding its design and implementation, and very real anxieties about the new digital divide it is creating. In a sense the book also seeks to allay the fears that traditional humanists and social scientists might have about big data signalling

2 The first paradigm was empirical science, the second theoretical science, and the third computer-driven science (Chen & Zhang 2014:315).

the end of theory and resulting in a data-driven world in which individuals simply keep generating new data before developing new algorithms to process it.

We would like to clarify why the book's sub-title includes the words 'a perspective from South African scholars'. Although we refer to the situation in South Africa as it pertains to the advent of the digital age and to links between big data and the humanities/social sciences, the book does not reflect an explicit South African contextualisation across *all* chapters. Selecting a sub-title such as 'a South African perspective' would have been misleading on our part. The wording in the sub-title reflects the fact that we are South African scholars writing about big data. In addition, and where possible, we have discussed technology and big data as they relate to the South African context, and the final chapter showcases a big data project that focuses specifically on a South African phenomenon. The main reason why each chapter does not constitute a more deliberate South African contextualisation is that, aside from efforts by South African scholars to link big data and the digital humanities³, the application of big data in the traditional humanities and social sciences is, for the most part, an alien phenomenon at local universities. Indeed, at these universities, big data analytics is typically taught in computer science/information systems departments or faculties, for example.⁴

Chapter 1 provides a brief history of big data by examining the foundational narrative behind it. This chapter puts paid to the notion that big data does not have a lengthy past and offers the caveat that all scholars – no matter what their discipline – should not ignore big data's history because "[even] in a petabyte world, history matters" (Barnes 2013:298).

In Chapter 2, the focus shifts to positioning big data in the (digital) humanities and (computational) social sciences. It considers how, if at all, big data has affected the epistemologies of these disciplines. This chapter also evaluates the now widely adopted practice of collaboration between big data scientists and humanists/social scientists, since each type of researcher possesses knowledge and skills that the other does not have.

Chapter 3 contemplates what some scholars have referred to as big data hubris. To this end, it begins with a description of one of the most significant big data 'fails' of the millennium before attempting to identify and resolve some of the numerous myths that surround big data, myths that have, at least to some degree, reduced the appetite that scholars in the humanities and social sciences may have for pursuing big data studies in

3 *Voices from the South* (2018), edited by Amanda du Preez and published by AOSIS, is an excellent volume that illustrates how humanities and arts scholars in South Africa are utilising data-driven research.

4 Examples are the University of Pretoria, the University of the Western Cape, and the University of the Witwatersrand. Interestingly, the latter university does offer an MA degree through its School of Social Sciences in the Humanities Faculty which focuses on data-driven methods in the humanities and social sciences.

their fields. Each common misconception is linked to a particular lesson that humanists and social scientists may harness to avoid some of the errors that have been committed by big data scholars in the past.

Ethical considerations constitute the focus of Chapter 4, and are reviewed against the background of some disturbing big data fiascos that only serve to signal the ongoing tendency within the big data ecosystem to neglect or entirely ignore ethical research.

Big data visualisation decisions have to be made in terms of ethical guidelines, and a detailed discussion of visualisation is contained in Chapter 5. A chapter has been devoted to this element given the complex cognitive, social, and emotional risks associated with graphical representations of data. The chapter has been written mainly with humanists in mind, since some scholars are of the view that data visualisation and the humanities cannot be reconciled. Social scientists and scholars in other disciplines should nevertheless find the information contained in the chapter quite useful.

In Chapter 6, the notion of big data power is probed against the background of studies conducted by humanists and social scientists. Although the chapter touches on big data's dark side, it also reflects on the positive contributions made to big data research by scholars in fields such as psychology, literature, history, political science, and the spatial humanities. While the chapter takes into account research that revolves around large volumes of data, it also challenges what big data scientists regard as 'big' since data size is a relative concept in the humanities and social sciences.

Since traditional humanists and social scientists tend to be qualitative researchers, Chapter 7 critically appraises the various qualitative data analysis software (QDAS) tools they may utilise to conduct big data research. The chapter dispels a frequently held myth that the use of such tools eliminates the need for human interpretation of data.

Chapter 8 focuses on the various components that make up what big data scientists refer to as big data architecture. It is necessary to take a closer look at the more technical aspects of the big data ecosystem since QDAS, while able to manage fairly large datasets, cannot adequately accommodate big data as defined by big data scientists.

In the world of big data, data science has become an important buzzword, but needs to be interrogated in light of the fact that it is a nascent rather than established discipline that is defined in ways that tend to perplex academics from different fields. Chapter 9 is thus devoted to deconstructing what data science and data scientists embody. The chapter also pays some attention to best practices for academic data science with a view to establishing a socially informed, "human-centred" (Neff, Tanweer, Fiore-Gartland & Osburn 2017:95) data science that also involves ethical thinking.

Chapter 10 focuses on a specific big data study in the humanities in order to illustrate how big data may be employed to answer specific research questions. The study

revolves around considering how tweets – both in and outside South Africa – have depicted Afrikaners in light of recent social and political events such as racist incidents and the government’s decision to expropriate land without compensation.

Currently, big data is a fraught phenomenon that either excites academics or generates feelings of anxiety and confusion. It is entirely up to scholars whether or not they would like to carry out big data research, and the various issues in this book may help them critically interrogate the phenomenon and understand its complex and evolving nuances.

The (fuzzy) origins of big data and the dangers of ignoring history

... the past remains potent for big data and ... proponents ignore it at their peril.

– Trevor Barnes (2013:297)

Does big data have a past or is it a phenomenon that simply exploded onto the global electronic scene with the advent of the microprocessor, the Internet, and the World Wide Web? If big data does indeed have a foundational narrative, what can we glean from this narrative, if anything? Does big data's emergence constitute a revolution or an evolution? If it is “a disruptive force” (Bollier 2010:28) to be reckoned with, then how should traditional humanists and social scientists view its place in their research environments? We attempt to answer these questions in this first chapter (and of course in subsequent ones) by navigating through the somewhat cloudy waters of big data's past.

1.1 A messy affair

“Big data” is a relative term depending on who is discussing it.

– Keith Foote (2017:1)

Any researcher who attempts to trace the history of big data will find this exercise both challenging and frustrating, as it is a phenomenon that simply does not fit into neat boxes. In the absence of academic papers that attempt to accurately chart the origins of big data, a survey of (visual) timelines of this phenomenon on the Internet has signalled that individuals in different fields hold opposing views on the brainchild behind it. Some recognise Roger Mougla (director of market research at O'Reilly Media) as having coined the term big data in 2005, while others argue that credit belongs to John Mashey (1998), who was chief computer scientist at Silicon Graphics, Incorporated (SGI) in the 1990s. Still others have pointed out that academic references to the term appeared for the first time in studies by Cox and Ellsworth (1997), Weiss and Indurkha (1998), and Diebold (2003). To muddy the waters even further, the origins of big data have variously been described as “intriguing and a bit murky” (Diebold 2012:2), “brief” (Marr 2015a:1), “uncertain” (Gandomi & Haider 2015: 138), and even “disputed” (Kaplan 2015:2). Since the history of big data is

somewhat messy, with researchers and technology journalists vehemently contesting the role specific individuals have played in its evolution, it is perhaps easier to think about it in terms of (big) data storage before and after the advent of the digital age; the emergence of statistical analysis and business intelligence (BI); the digital revolution; as well as significant events surrounding big data, with some acknowledgment going to Gil Press (2013) and Bernard Marr (2015a) who have identified key milestones in these areas.

1.2 The history of (big) data storage

Although it might be easy to forget, our increasing ability to store and analyze information has been a gradual evolution. – Bernard Marr (2015a:1)

Attempting to uncover the origins of big data would not be complete without first considering how people through the ages have stored information they consider to be valuable. What we know is that the history of data storage goes back to approximately 18000 BCE when Palaeolithic tribespeople in East Africa made notches on sticks or bones – presumably to facilitate counting when it came to their food supplies and trading activities (Igarashi, Altman, Funada & Kamiyama 2014:3). A fascinating example of an ancient storage device is the Caral quipu, a system of knotted strings that archaeologists believe may have been used 5000 years ago by the Incas to store massive amounts of data about their culture and civilisation (Chrisomalis 2009:67). Between 2500 and 2400 BCE, dwellers in the ancient city of Ebla (an early kingdom in Syria) used argyle tablets to store huge volumes of information pertaining to their economic, diplomatic, and commercial activities (Kaplan & di Lenardo 2017:4). Other ancient storage tools include coins (thought to have been used for the first time in the sixth or fifth century BCE) to store economic data and the Antikythera mechanism, which scholars speculate might be an analogue astronomical computer the Greeks used in circa 82 BCE to track planetary movements. Big data storage was first attempted by the Babylonians (in circa 2400 BCE) and the Alexandrians (between 200 BCE-48 CE) when they constructed libraries to store large collections of tablets and scrolls, respectively. From 130-115 BCE, Rome’s *pontifex maximus*, Publius Mucius Scaevola, had his archives stored in the *Annales maximi* in order to record details about events such as battles and famines (Forsythe 2012). Unlike argyle tablets which could be erased with ease, the *Annales* were more durable in terms of wear and tear, showing that from a technological perspective, storage devices improved over time (Kaplan & di Lenardo 2017:4).

What all these ancient devices have in common “is their capacity to deal with an open-ended stream of information and reorganize it to fit a given information paradigm” (Kaplan & di Lenardo 2017: 3) – they are “regulated representations ... governed by a set of production and usage rules” (Kaplan & di Lenardo 2017:3-4). A coin, for example, is

a regulated representation reflecting information about a particular civilisation's economy, trade, and social organisation, amongst other things.

When it comes to the Renaissance period in Europe, what Kaplan and di Lenardo (2017:3) call *data acceleration* became pronounced: “[not] only were new editions of ancient texts starting to be printed and circulated but also a deluge of ‘how-to books’ explaining previously secret arts and methods ... This sudden increase in knowledge and exposure to new practices created a well-documented feeling of information overload” (Kaplan & di Lenardo 2017:3). In terms of technology, the Renaissance period is noted for improvements in search and retrieval; in this respect, scholars point to the invention of indexes, chronologies, and accounting tables (Kaplan & di Lenardo 2017:3). A major step in the development of data storage was the evolution of the printing industry which contributed significantly to open-ended streams of data which then became part of the data deluge (Kaplan & di Lenardo 2017:2). The Gutenberg Press invented around 1440 by Johannes Gutenberg of Germany is often hailed as the first movable type printing press of its kind. However, Chinese artisan Bi Sheng is credited with having invented a printing press out of Chinese porcelain between 1041 and 1048 (Gunaratne 2001:467).

The eighteenth and nineteenth centuries are marked by “shared knowledge systems” as well as by “the rise of standardization [which] reinforced the idea that the fate of every dataset is to become, sooner or later, a shared resource by which new predictions and patterns can be established” (Kaplan & di Lenardo 2017:11). In the eighteenth century, rapid developments occurred in the field of lexicography which saw the publication of dictionaries such as *Encyclopédie* edited by Denis Diderot (1751) and Samuel Johnson's *A Dictionary of the English Language* published on 4 April 1755. These dictionaries “[illustrated] a growing need to not just collect but classify, categorise and order information to make it both meaningful and useful” (Robertson & Travaglia 2015:2), although it must be added that eighteenth-century dictionaries were not particularly “enlightened” (McIntosh 1998:3), since they were “more likely to represent the subjective impressions and prejudices of the editor or his sources than the objective documentation of language which became a feature of work from the mid- to late nineteenth century” (Simpson 1989:181-182). The first Significant industrial application of data storage in the eighteenth century may be attributed to Basile Bouchon, a textile worker in Lyon, who invented a perforated paper tape mechanism for storing patterns to be used on cloth in 1725. By the early part of the nineteenth century, which is also referred to as the pre-digital era (Robertson & Travaglia 2015:2), Joseph Marie Jacquard had refined Bouchon's device with a view to further simplifying the process of manufacturing textiles (1804). In terms of the automation of computers, historians contend that huge strides were made in 1822 when Charles Babbage proposed the Difference Engine to perform differential equations and again in 1837 when Babbage described a mechanical

computer called the Analytical Engine. What makes this computer particularly significant is that its co-designer was Augusta Ada Byron King, Countess of Lovelace (1815-1852), who is regarded as the first computer programmer (Stanley 2016).

One of the most important contributions made to the evolution of digital computers and artificial intelligence was that of George Boole, who was an English philosopher, mathematician, and logician. In *Laws of thought* (1854), Boole began developing an algebra for logical inference, and today we still employ Boolean algebra which employs binary numbers (that is, 0 and 1) to operate computers.

When it comes to the twentieth century, methods for storing large amounts of data included Fritz Pfleumer's magnetic tape invented in 1928 (Levaux 2017), the Selectron tube (1946) which could hold between 32 and 512 bytes of data, the hard disk drive developed by IBM in 1956, the cassette tape produced by the Phillips Company in 1962, and the floppy disk drive conceived in 1967 by Alan Shugart who worked for IBM. By the 1990s, floppy disks could store approximately 250 megabytes of data. In the 1980s and 1990s, compact disks (CDs) and digital versatile discs (DVDs) dominated the global market until universal serial bus (USB) flash drives and secure digital (SD) cards made their way onto the scene in 2000.

Currently we are reaping the benefits of network-based or cloud computing for processing, storing, and distributing big data. The origins of this technology can be traced back to the 1960s when Joseph Carl Robnett Licklider introduced the concept of an "intergalactic computer network" (Kaufman 2009:61) at the Advanced Research Project Agency (ARPA).⁵ However, others dispute this, arguing that cloud computing made its debut in 2006 when former Google CEO Eric Schmidt referred to it at a search engine conference (Regalado 2011:4). The current prediction is that the shape of cloud computing will change radically in the near future with some "[envisioning] a hybrid model where cloud users begin deploying small-scale data [centres] in strategic geographic locations that move data processing capabilities closer to the user – but are still centrally managed" (Froehlich 2017:1).

5 In an office memorandum sent to his colleagues in 1963, Licklider referred to this experimental computer network as an open networking system that would be "the main and essential medium of informational interaction for governments, institutions, corporations and individuals" (cf. Garreau 2006:22). Needless to say, this version of Licklider's vision called ARPANET was the forerunner of the Internet.

1.3 The emergence of statistical analysis

Errors using inadequate data are much less than those using no data at all.

– Charles Babbage (circa 1850)

In his essay on the history of big data, sociologist David Beer (2016a:1) is of the view that “we already have a history of big data that can be found in accounts of the use of statistics to know and govern populations”, and it is for this reason that we conduct a brief survey of the emergence of statistics through the centuries. Standard statistics was not a reality in prehistoric times, but scholars contend that its beginnings can nevertheless be found in ancient Sumar, Crete, China, and Egypt, for example, in the form of censuses related to livestock, capitation tax, the construction of buildings, and the like (cf. Kitchin & Lauriault 2014:2). Essentially, statistics was restricted to collation of the population or to the recording of trade activities in most early empires. It appears that the earliest reference to statistics appeared in *Manuscript on deciphering cryptographic messages* written by Sheikh Al-Kindi in the ninth century (Singh 2000).

Before 1600, the origins of statistics are not entirely clear (Creighton 2012), but *Liber de ludo aleae* (*Book on games of chance*) written by Gerolamo Cardano (1501-1576) and published in the sixteenth century was hailed as a sophisticated approach to probability calculus (Bellhouse 2005:180). Then, in 1577, a Dominican theologian, Bartolomé de Medina (1527-1581), defended the doctrine of probabilism which states that if certainty about a particular position or issue is not possible, then the best criterion to follow is probability (cf. Decock 2013:75).

In the seventeenth century, John Graunt (1620-1674) made scientific inferences based on the London bills of mortality⁶ (cf. Kotz 2005:140) and is particularly famous for providing statistical details about outbreaks of bubonic plague in London (Sartorius, Jacobsen, Törner & Giesecke 2006:181) in his book entitled *Natural and political observations, mentioned in a following index, and made upon the bills of mortality*.⁷ French mathematicians Blaise Pascal and Pierre de Fermat proposed probability theory, which originated out of

6 In 1603, James I instructed the Company of Parish Clerks to publish weekly accounts of London's births and deaths and this culminated in the bills of mortality (Morabia 2013:1). Interestingly, during the London plague of 1665, the Company of Parish Clerks employed mainly elderly women called 'searchers' to count the number of people who died and to determine their causes of death (Slauter 2011:9). At this time, opponents of the bills of mortality heavily criticised searchers for their lack of medical training, maintaining that some of them could not be trusted to keep accurate records because they often misreported deaths (Slauter 2011:9).

7 John Graunt's full text was edited by Walter Francis Willcox in 1939 and published by Johns Hopkins Press. The text is regarded as constituting pioneering work on the science of demography (Brimblecombe 2017:18).

a dispute over a popular game of dice in 1654. This mathematical theory is extensively employed for big data analysis today (Singpurwalla & Landon 2014:19). Statistical analysis advanced further thanks to Sir Henry Furnese who employed business intelligence (or BI) to collect and analyse data about the markets in order to make a profit before his business rivals could do the same. His data analysis is described in Richard Millar Devens' *Cyclopaedia of commercial and business anecdotes* (1865). In an interesting article on big data in historical perspective, and basing her thoughts on Hacking's (1991) essay on the history of statistics, Meg Ambrose (2015:203) points to "the avalanche of printed numbers that flooded Europe between 1820 and 1840", a period known as the probabilistic revolution in terms of statistical analysis. Ambrose (2015:203) observes important similarities between this period and what is occurring today:

Between 1820 and 1840, a flood of data from across society became available, aggregated, analyzed, and acted upon. From this period, a series of similarities to big data can be extracted: datafication issues, big data lures, and structural changes. A number of social issues surfaced in the 1800s that have resurfaced today: governability, classification effects, and data-based knowledge. Enthusiasm for big data during both periods was driven by particular lures: standardized sharing, objectivity, control through feedback, enumeration, and the discovery versus production of knowledge. Both periods also experience(d) structural changes: division of data labor, methodological changes, and a displacement of theory.

Data visualisation techniques also began to emerge such as those in the form of John Snow's maps employed to track outbreaks of cholera in London in 1854 and Florence Nightingale's (1858) sunburst graph designed to record, amongst other things, the mortality rates of British soldiers fighting on the Crimean Peninsula between 1854 and 1855. In 1880, statistical computation took a giant leap forward when data processing pioneer Herman Hollerith, an engineer employed by the US Census Bureau, responded to data acceleration in the form of "an initial *societal stress*" (Kaplan & di Lenardo 2017:3) which, in the case of North America, involved a growing population that had to be counted. Conducting a census of the population in 1880 took a total of eight years to complete, but in the meantime, Hollerith invented a punch card that helped to significantly reduce tabulation time for the 1890 US census to two years (Austrian 1982). It is for this reason that some scholars call Hollerith the father of our modern automatic computation,⁸ although other scholars argue that this epithet should be attributed to either Charles Babbage (Cooper 2004:4)

8 The company that Hollorith founded later became known as IBM (International Business Machines Corporation).

or to Alan Turing (Daylight 2015:205), a pioneer of theoretical computer science and artificial intelligence.

Twentieth century statistics is equated with Karl Pearson who famously presented the notion of the Chi square distribution which constituted “a qualitative leap in applying powerful new mathematics (matrix theory) to statistical reasoning” (Efron 2003:32). Other influential scholars include chemist and statistician William Gosset, who used his work on the t-distribution (1908) to employ methods of sampling in Guinness’ Brewery in Ireland; Ronald Fisher, who introduced fundamental concepts such as sufficiency, efficiency, and optimality in *Statistical methods for research workers* (1925); Neyman and Pearson, who published a seminal paper on optimality theory for testing problems in 1932; and Gene Glass, who coined the term ‘meta-analysis’ in 1976.

1.3.1 *Big business, big data*

In order to understand the role of big data in big companies, it’s important to understand the historical context for analytics and the brief history of big data.

– Thomas Davenport (2014:194).

We noted in the section preceding this one that the term ‘business intelligence’ (BI) appeared in Devens’ (1865) *Cyclopaedia of commercial and business anecdotes* which among other things described how Henry Furnese utilised BI to outperform his competitors when playing the stock markets. What is significant about the study carried out by Furnese is that it is regarded by some to constitute the first attempt to exploit data about business for commercial purposes (cf. Marr 2015a). In 1958, Hans Peter Luhn developed the notion of BI even further when he wrote a paper in *IBM Journal of Research and Development* outlining the development of an intelligence system that would “utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the ‘action points’ in an organization” (Luhn 1958:314). It was only in the 1990s, however, that BI was widely exploited in the context of business (cf. Wang 2016:673) and today it is being reshaped by big data (cf. Fan, Lau & Zhao 2015:28).

As a term, ‘big data’ appears to be fairly new in the world of business (Hashem, Yaqoob, Mokhtar, Gani & Khan 2015:100), but from the 1960s onwards, businesses across the globe began to design centralised computing systems to cope with the data deluge and to carry out analytics, which may at a basic level be defined as “the systematic computational analysis of data or statistics” (*Oxford English Dictionary* 2019). Thomas Davenport and Jill Dyché (2013:26) point out that it is useful to think about big business and big data in terms of what they refer to as Analytics 1.0, 2.0, and 3.0. Between 1954 and 2009, business organisations employed traditional Analytics 1.0, which Davenport and Dyché (2013)

observe was characterised by internally sourced data that was small and structured and that was analysed mainly through descriptive analytics (reporting). The picture changed quite dramatically between 2005 and 2012 when huge Internet-based companies such as Google and eBay began to exploit the digitisation of huge amounts of data (cf. Stubbs 2010:3). Davenport and Dyché (2013) refer to this period as the advent of Analytics 2.0, which was characterised by a rather narrow analytical focus on data that allowed companies to carry out customer reviews and analyse their product data in real time. What made the Analytics 2.0 era distinct from Analytics 1.0 was that data was both internally and externally sourced, while data was very large and/or unstructured. In addition, companies began to use big data processing frameworks such as Hadoop, Spark, and Storm (see Chapter 8) to manage and analyse enormous amounts of data with a view to achieving a competitive advantage over their rivals. We now appear to be entering Analytics 3.0, which Davenport and Dyché (2013) claim combines big data and traditional analytics, but this time with a greater emphasis on predictive or prescriptive analytics. Predictive analytics “is about forecasting and providing an estimation for the probability of a future result, defining opportunities or risks in the future” (Vassakis, Petrakis & Kopanakis 2018:10). Thus, for example, a company may choose to apply this kind of analytics in order to predict future trends and patterns as they relate to customer behaviour. Prescriptive analytics, by contrast, is aimed at forecasting the impact of future actions in order to improve strategic decision-making and generate solutions for problems related to costs or the development of new products, for example (Vassakis *et al.* 2018:10).

1.4 The digital revolution and events surrounding big data

From the dawn of civilization to 2003, five exabytes of data were created. The same amount was created in the last two days. – Google CEO Eric Schmidt (2010:1)

Whether we are exploring the history of data storage or unearthing the origins and development of statistical analysis, what is clear is that it has always been necessary to find effective and efficient ways of collecting, processing, and managing data: “[appeals] to the ‘data deluge’ and ‘information overload’ are not as new as we might be led to believe” (Levin 2018:668), given that “every age was an age of information, each in its own way” (Darnton 2000:1). As early as 1941, the huge volumes of data to be managed were described in terms of an “information explosion” in the *Lawton Constitution*, and this explosion has only been exacerbated by the digital revolution which could be described in terms of three acts, namely, “by the microprocessor and the power to compute, ... by the network and the power to connect, [and by] the third [which] will be defined by data and the power to predict” (Richards & King 2014:397). To cope with the tsunami of data, there are just over half a

million data centres worldwide.⁹ Currently, there are 54 data centres in Africa, and 21 of them are located in South Africa.¹⁰

South Africa's own digital evolution can be traced back to 1988 when Francois Jacot-Guillarmod, Dave Wilson, and Mike Lawrie set up an email link to the Internet at Rhodes University to improve access to the university and to academia.¹¹ These pioneers used a dial-up system which was then replaced with Internet access in 1991. Thereafter, commercial Internet Service Providers (ISPs) arrived in quick succession, which "propelled South Africa to one of the 20 most Internet-connected countries in the world by the mid-1990s" (Horowitz & Currie 2007:451). With the arrival of the Internet, a wave of developments took place, including the formation of the country's first commercial Internetworking Company of Southern Africa, Tisca (1993), which expanded Internet access even further. By 1997, dial-up connections of 56 kilobytes per second (kbps)¹² had gained traction, followed shortly afterwards by Telkom's 64 kbps Internet service (Friedenthal 2015:2). Wireless broadband arrived in 2004 and currently, South Africans have access to broadband speeds of over 100 megabits¹³ (Friedenthal 2015:2), while universities enjoy access through the cyber infrastructure for big data provided by the South African National Research Network, (Friedenthal 2015:2).

Collating information from Press (2013) and Marr (2015a), the timeline in Figure 1.1 below visually represents some of the most significant events in the evolution of big data identified by Richards and King (2014:397) since the emergence of microprocessors, although these events certainly intersect, given that inventions and advances in computing overlap. We have included quotations from big data proponents to provide examples of how big data has been framed in the mass media, since framing may assist us in understanding the values and assumptions encoded in big data (cf. Portmess & Tower 2015:3) which we interrogate further in Chapter 2.

9 Emerson Network Power.

10 <http://www.datacentermap.com/africa/>.

11 "The history of Internet access in South Africa" (<https://mybroadband.co.za/news/internet/114645-the-history-of-internet-access-in-south-africa.html>). We relied heavily on this particular site to glean information about the history of the data revolution in South Africa, which is sketchy to say the least. The site draws on a report entitled 'Internet access in South Africa 2010', written by Arthur Goldstuck, managing director of World Wide Wox.

12 Kilobits per second is a measure of bandwidth (data transfer speed); one kbps is equivalent to 1000 bits of bandwidth per second.

13 One megabit equals one million bits.

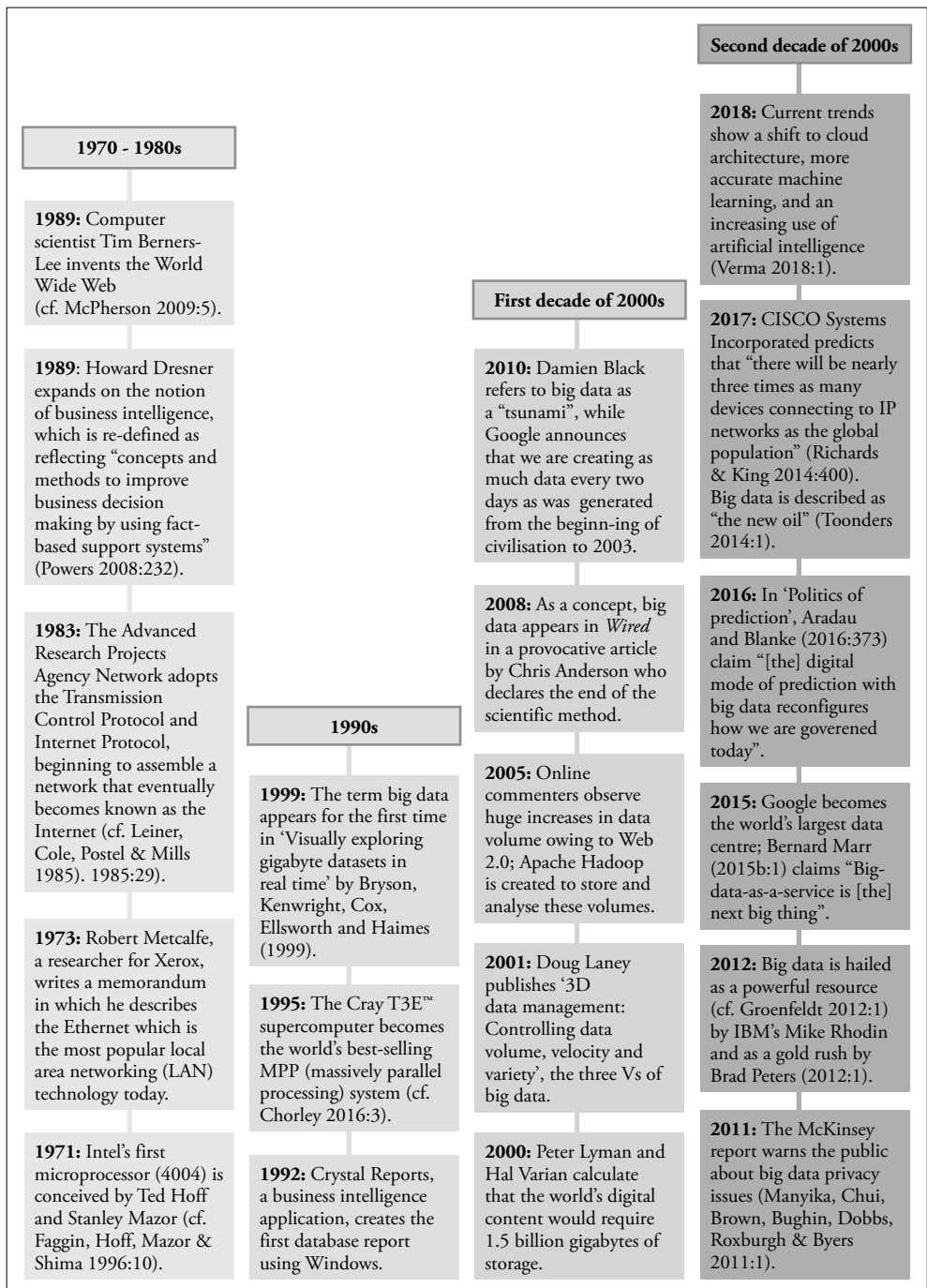


Figure 1.1: Milestones in the era of big data since the introduction of the first commercial microprocessor

1.5 Big data's history lessons

1.5.1 *Revolution versus evolution*

Digital transformation: human evolution, not technological revolution.

– Richard Mullins (2017:1)

Big? Smart? Clean? Messy? – Christof Schöch (2013:2)

Given these events or milestones, whether “[knowingly] or unknowingly, with every Google search, every Facebook post, and ... every time we simply turn on our smartphones ... we produce metadata” (Richards & King 2014:402), or data about that data (Schöch 2013:3), which in turn has resulted in the creation of “a kind of big metadata computer” (Richards & King 2014:403) as it were. However, we need to pause momentarily and ask ourselves, *Are we experiencing a big data revolution?* In her essay on the lessons we can glean from the era of big data, Ambrose (2015) casts doubt on big data constituting a revolution. Instead, she argues that what we are currently seeing is a revolution as it pertains to the *sheer volumes of data* we are currently faced with: “big data is not the revolution – big data is the avalanche of numbers which was intimately intertwined with the probabilistic revolution” (Ambrose 2015:204) referred to earlier on. That the big data transition is emergent is echoed not only by Steven Finlay (2014:17), who contends that “[big data is] a process of evolution not revolution”, but also by Lazar, Kennedy, King and Vespignani (2014:1203), who suggest that we should focus not on a big data revolution, but “on an ‘all data’ revolution”, where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of the world”. This latter statement will no doubt constitute good news to scholars in the humanities and social sciences who are accustomed to working with traditional small data which “often offer information that is not contained (or containable) in big data” (Lazar *et al.* 2014:1205).

1.5.2 *The past in big data*

It is by acknowledging the long history of the accumulation of data about individuals and populations that we can begin to make a departure into seeing the different ways that data are presented in conceptual terms. – David Beer (2016a:4)

What the brief history in this chapter also tells us is that the notion that big data has emerged out of a vacuum is a common misconception (O’Sullivan 2017:4): “the sense that we are being faced with a deluge of data about people is not something that is entirely new” (Beer 2016a:2), given “the great explosion of numbers that ... occurred during the 1820s and 1830s”

(Porter 1986:11), for example. As Kaplan and di Lenardo (2017:1) so succinctly put it, “history is punctuated by several *Big Data moments* which are characterized by a widespread, shared sense of information overload alongside rapid societal acceleration accompanied by the invention of new intellectual technologies”. Yet for some big data scholars, the history behind big data is unfortunately not particularly important. In a cautionary essay entitled ‘Big data, little history’, Trevor Barnes (2013) passionately criticises individuals who believe that history is disconnected from a petabyte world. Barnes (2013:298) specifically targets individuals such as Chris Anderson when he argues that:

... [In so far] as I can contribute to the discussion about big data, it is to caution that things are not as different as they might seem. It is not quite the brand new day that Chris Anderson supposes. In Chris Anderson’s account, history drops out. It might be big data, but it is little history. This neglect of history is a typical modernist move. The past is ignored because nothing must constrain or limit what is to come. Only the bright new future matters. ‘History is more or less bunk’ as the iconic modernist Henry Ford once famously put it.

What lies at the crux of Barnes’ (2013) message is that the problems and challenges that accompanied the deluge of numbers in the past have not disappeared – that any scholar who chooses to explore big data needs to be mindful not only of the history of a particular phenomenon, but also of the methodological pitfalls experienced by past researchers. Kitchin (2014a:136) presents an example from the field of social physics to illustrate what can occur if history is ignored. Some scholars in this field have exploited big data analysis to draw conclusions about social and spacial processes as they occur in cities (Bettencourt, Lobo, Helbing, Kühnert & West 2007). Unfortunately, these scholars have “often wilfully [ignored] a couple of centuries of social science scholarship ... the result [being] an analysis ... that is largely reductionist, functionalist and ignores the effects of culture, politics, policy, governance and capital, and a rich tradition of work that has sought to understand how cities operate socially, culturally, politically, and economically” (Kitchin 2014a:136). These scholars have therefore replicated the limitations inherent in the studies of social scientists of the mid-twentieth century (Kitchin 2014a:136).

1.5.3 *Two foundational narratives*

[Two] narratives contribute to structuring a multifaceted definition of the bigness of Big Data. – Frédéric Kaplan and Isabella di Lenardo (2017:1)

A useful point of departure might be to consider that the history of big data points to two significant foundational narratives, namely, the so-called *data deluge narrative* and the *big science narrative* (Kaplan & di Lenardo 2017:1). The narrative behind the big data deluge is, as we have seen, that big data has originated out of the limitless possibilities offered up by the Internet and online communication networks (Kaplan & di Lenardo 2017:1), and that we are now attempting what Mayer-Schönberger and Cukier (2013:73) have coined *datafication* – the process of transforming vast quantities of real-time digitised data into “structured knowledge systems that document the lives of people, companies and institutions, aggregating information about places, topics, or events” (Kaplan & di Lenardo 2017:2). The big science narrative is an older narrative that tells the story of how researchers in diverse fields have been compelled to devise new methodologies to manage huge scientific archives known collectively as *big science* (Kaplan & di Lenardo 2017:2).

What makes these narratives insightful is that they point to what Kaplan and di Lenardo (2017:1) refer to as “an epochal paradigm shift”. Basing his thinking on that of Kuhn (1996), Kitchin (2014b:3) observes that “[periodically] ... a new way of thinking emerges that challenges accepted theories and approaches”: as a “disruptive innovation” (Kitchin 2014b:10), big data is certainly compelling scholars in the (digital) humanities and (computational) social sciences to think about using alternative epistemologies in their research. However, the various shapes that these epistemologies will take in these disciplines are still disputed by scholars (Kitchin 2014b:3), and it is this contestation that is interrogated throughout this book. For now, it suffices to say that it is unlikely that the approaches behind big data – empiricism and data-driven science – will become substitutes for approaches employed in the traditional social sciences and the humanities, given their philosophical underpinnings (cf. Friese 2016:35). Kitchin (2014b:3) predicts that what will happen instead is that while “[big data] will enhance the suite of data available for analysis and enable new approaches and techniques ... [it] will not fully replace traditional small data studies”. This sentiment is shared by Danah boyd and Kate Crawford (2012:670), who contend that “it is increasingly important to recognize the value of ‘small data’. Research insights can be found at any level, including at very modest scales”.

We turn now to a more detailed discussion of the role of big data in the (digital) humanities and (computational) social sciences before attempting to dispel some of the myths that scholars in the humanities and social sciences might harbour about employing big data in their fields of study.

Locating big data in the (digital) humanities and (computational) social sciences

In [the] humanities there [has] always been big data. – Amalia Levi (2013:33)

The rise of big data ... represents a watershed moment for the social sciences.

– Daniel McFarland, Kevin Lewis and Amir Goldberg (2016:12)

The very notion of big data creeping into their research spaces casts an intimidating shadow over traditional humanists and social scientists, who may fear human behaviour being reduced to mere mathematical models (Schirrmacher 2015). Some humanists, for example, are of the view that the collection of vast amounts of quantitative data by digital humanists is akin to the loss of qualitative meaning (cf. Heuser & Le-Khac 2011:84), while social scientists argue that because big data approaches are essentially descriptive they will not be able to answer the *why* and *how* questions that pertain to the data they are interested in (Serfass, Nowak & Sherman 2017:341).

Notwithstanding the fact that the humanities and social sciences have enjoyed a long association, they fundamentally differ in terms of their areas of focus and philosophical underpinnings (Kitchin 2014b:7). While scholars in the humanities use critical and analytical approaches to study the human condition or experience, those in the social sciences generally employ empirical methods to explore human behaviour, although other types of research such as theoretical, historical, analytical, and conceptual-philosophical research are also carried out (Punch 2013:3). The emergence of big data is now radically transforming the landscapes of both disciplines, but the “trajectory” (Kitchin 2014b:7) of big data epistemology in the two disciplines remains a topic of heated debate among scholars. In this chapter, we consider the literature and what it tells us about the nature and impact of big data on the epistemologies of the humanities and social sciences. We also look at the new and contested big data connections being forged in the digital humanities and computational social sciences.

2.1 How big data is framed

What the hell is big data anyway? – FabCom (2013:1)¹⁴

Torture the data, and it will confess to anything. – Ronald Coase¹⁵

Dominique Boullier (2016:3) correctly observes that when Mike Savage and Roger Burrows published ‘The coming crisis of empirical sociology’ in 2007, little did they know how prophetic their predictions would be. Amongst other things, they wrote, “sociologists have not adequately thought about the challenges posed to their expertise by the proliferation of ‘social’ transactional data which are now routinely collected, processed and analysed by a wide variety of private and public institutions” (Savage & Burrows 2007:885). The same could of course be said of scholars working in the humanities (Schöch 2013:2). Reflecting on their claims in a 2014 paper entitled ‘After the crisis?’, Savage and Burrows point out that “[what] must have read as new, innovative and important in 2007 ... now reads to us as a pretty mainstream position, not just in sociology but also across the cognate social sciences more generally” (Savage & Burrows 2014:1). However, we question how mainstream many humanists and social scientists find the big data deluge to be. In contrast to reactions from those in the digital humanities¹⁶ and computational social sciences, “the response seems mixed in the more traditional branches of the social sciences and humanities” (Puschmann & Burgess 2013:1691). Much of the big data critique centres around the absence of theoretical grounding, the decontextualisation of social phenomena (Puschmann & Burgess 2013:1691), and the difficulties inherent in having to grapple with data analytics, concerns which are constantly acknowledged and addressed throughout this book. A discussion of these concerns makes more sense if we first consider how big data is framed in the mass media through specific (and seemingly harmless) metaphors because “when we approach any novel phenomenon, we begin to understand it first by metaphor. We start to make sense of the phenomenon by treating it as if it were like some other phenomenon with which we are more familiar” (Adamson & Bakeman 1982:224, cf. Penzold & Fischer 2017:2). Big data is framed in contradictory ways, reflecting both fear of it and excitement about the research possibilities it opens up to scholars (cf. Lupton 2015:2).

14 <https://www.fabcomlive.com/strategic-marketing-agency/wp-content/uploads/What-The-Hell-Big-Data-White-Paper.pdf>.

15 Nobel Prize winning economist. The date of this quotation is unknown, although Coase himself claims he said this in the 1960s.

16 We would like to acknowledge the work being done by The South African Centre for Digital Language Resources (SADiLaR). On this platform, which is aimed at managing digital resources around language-related studies, the digital humanities programme makes use of state-of-the-art data-driven methods to carry out research in the humanities and social sciences (See <https://www.thesouthafrican.com/lifestyle/government-launches-south-african-centre-for-digital-language-resources/>).

2.1.1 *Two dominant metaphors*

...the people who get to impose their metaphors on the culture get to define what we consider to be true – George Lakoff and Mark Johnson (1980:160)

In the previous chapter, we traced significant big data moments, particularly after the advent of the Internet, and noted that the media made use of statements pertaining to the “controlling [of] data volume” (Laney 2001:70), “the end of theory” (Anderson 2008:1), and to data as a “tsunami” (Black 2010:1), a “powerful natural resource” (Mike Rhodin in Groenfeldt 2012:1), a “gold rush” (Peters 2012:1), “the new oil” (Toonders 2014:1), and “big data-as-service” (Marr 2015b:1). Such pronouncements are significant as they highlight two metaphors Puschmann and Burgess (2014:1691) argue have become dominant, namely, that “big data is a force of nature to be controlled” and that “big data is nourishment/fuel to be consumed” (cf. Penzold & Fischer 2017:2). The message behind the first conceptual metaphor is that although big data is overwhelming, it can nevertheless be converted into a resource if effectively controlled. However,

[data] is not a natural resource that replenishes itself, but ... is created by users with intentions entirely unrelated to its use as a valued commodity. It is created by humans and recorded by machines rather than being discovered and claimed by platform providers or third parties. At the same time, it is generally not used for the purpose for which it was collected. Its mass makes it easy to deliberately ignore individual items in favor of aggregate properties (Puschmann & Burgess 2014:1691).

In addition, Puschmann and Burgess (2014:1699) point out that big data cannot be perceived as a value-neutral resource because in the first place, its value differs depending on whether we are referring to its ‘owners’,¹⁷ generators or collectors, and in the second, its value is borne out of analysis: it is not “inherent in some sort of natural form of consumption” (Puschmann & Burgess 2014:1699). The gold rush metaphor is a particularly odd one, since “[suggesting] that the intrinsic meaning of data is, like nuggets of gold, already there, just waiting to be uncovered, means distancing the interpretation from the interpreter and her subjectivity” (Puschmann & Burgess 2014:1699). This is problematic since scholars in the humanities and social sciences who employ qualitative methodology rely on subjective processes to understand social phenomena (Ratner 2002).

The message behind the second metaphor is essentially that big data is a resource that needs to be consumed to secure survival and that it is, in a sense, a fuel that drives

17 In Chapter 4, we will discuss the notion that information is owned by information empires such as Twitter, Apple, and Facebook because they trade in information generated by their users.

businesses and institutions (Puschmann & Burgess 2014:1700). Again, it is a metaphor that is far from innocuous, since it conveys the notion that “the consumption of data strengthens the company or institution while requiring no or very little conscious interpretation or reflection” (Puschmann & Burgess 2014:1700; cf. Crawford, Miltner & Gray 2014:1669). Dawn Holmes (2017:20) adds that the framing of big data as oil is essentially a marketing tool exploited by data analytics to sell their products and services. She adds that “the metaphor only holds so far. Once you strike oil you have a marketable commodity. Not so with big data; unless you have the right data you can produce nothing of value” (Holmes 2017:20).

These two metaphors reflect a conflict between two paradigms – the big data paradigm in which data is seen as neutral, existing independently of context (Kitchin 2014a:145) and “an older [paradigm]” (Puschmann & Burgess 2014: 1702) in which data is perceived to be socially constructed.

In marked contrast to the second paradigm, the first suggests that meaning surpasses context, “that anyone with a reasonable understanding of statistics should be able to interpret [data] without context or domain-specific knowledge” (Kitchin 2014b:2). Kitchin (2014b:5) calls this “a conceit” on the part of some data scientists who are now undertaking social science and humanities research, sometimes without regard to the subject matter expertise required in these fields. This lament has also been voiced by data scientist Jake Porway (2013:2), who remarks that “[as] data scientists, we are well equipped to explain the ‘what’ of data, but rarely should we touch the question of ‘why’ on matters we are not experts in”.

On the subject of conceit, qualitative researcher Annette Markham (2013:1) notes that she was at one stage asked the following questions by a big data analyst:

How can we make qualitative research more important in the arena of big data? If big data is the purview of quantitative and computational analysts and qualitative researchers do not want to be left behind, how can they better inform big data research and researchers?

Markham (2013) understandably goes on to state that these questions troubled her since they are based on two faulty premises. With regard to the first mistaken assumption – that it is only computational analysts and quantitative researchers who use big data – data in the humanities and the social sciences can also be big (cf. Levi 2013:33). However, as we will see in Chapter 3, (1) “[it is] not just about [the] size of the data” (Lazar *et al.* 2014:1204) and (2) size is in any case only one of several defining dimensions of big data (Gandomi & Haider 2015:143). The second erroneous supposition reflected in the set of questions is that qualitative research has no place in big data, but a number of studies put paid to this notion. Kathy Mills (2017:15) observes that “[qualitative] researchers are well-positioned to generate research questions, and to select, curate, interpret and theorize big data away

from reductionist claims”, and provides many examples of instances in which qualitative research has been used to bolster big data studies. In the field of communication research, for instance, Christine Lohmeier (2014:75) argues that combining qualitative research methods with big data analytics yields fruitful results for illustrating the dynamics of large online communities. In the area of the sociology of technology and innovation, Snijders, Matzat & Reips (2012:2) are of the view that in order to understand the micro-processes that occur in social media networks, employing mathematical models is not sufficient and that analysts should also use social science theories to guide their research. Parker, Saundage and Lee (2011:6) recommend that social informaticians combine data analytics and qualitative content analysis to deepen their insights into how people appropriate social media discourse: “social informaticians require a qualitative research method which gives them the flexibility to allow their research questions and units of analyse to emerge (or change) inductively throughout the process”.¹⁸

2.1.2 *A collaborative effort*

...analysis of big data is an interdisciplinary research, which requires experts in different fields to harvest the potential of big data.

– Min Chen, Shiwen Mao and Yunhao Liu (2014:176)

The above studies have been carried out by researchers who have a background in data analytics, which is certainly not the case for many humanists and social scientists, who, like Kevin Lewis (2015:3), might describe themselves as “statistically challenged, Python-ignorant digitalphobe[s]”. It is also not helpful when messages with a slightly threatening undertone are bandied about such as Lev Manovich’s (2012:472) “[you had] better have knowledge of [big data analytics]” if you want to understand big social datasets”. What is often overlooked is that many data analysts do not have any expertise in the humanities and social sciences, and “[without] subject matter experts available to articulate problems in advance, you get [poor] results ... Subject matter experts are doubly needed to assess the results of [a study], especially when you’re dealing with sensitive data about human behavior” (Porway 2013:2). This conundrum is also touched on by Salah, Manovich, Salah and Chow (2013:411) in the context of analyses of social network sites or SNSs: “[social network sites] are studied so far either by social scientists, who lacked the necessary tools and expertise to conduct research on large-scale datasets, or by physicists who lacked the research goals of social scientists in exploring the SNSs for inquiries about social phenomena”.

One solution to this dilemma has been offered by scholars such as Stockmann (2016:23), Mills (2017:9), and Ford (2014:1), who recommend that big data analysts come

18 See Chapter 7 for a brief discussion of content analysis and big data.

together with social scientists and humanists in order to overcome gaps in their knowledge and expertise. Ford's (2014) contribution reflects an interesting narrative about how, as an ethnographer working on projects related to Wikipedia sources, she collaborated with data scientists Dave Musicant and Shilad Sen:

... Dave and Shilad had the necessary skills and resources to extract over 67 million source postings from about 3.5 million Wikipedia articles ... [and] I was able to contribute ideas about different ways of slicing the data in order to gain new insights. Dave and Shilad had access to sophisticated software and data processing tools for managing such a high volume of data, and I had the knowledge about Wikipedia practice that would inform some of the analyses that we chose to do on this data (Ford 2014:3).

In doing this kind of interdisciplinary research, Ford (2014:2) discovered that her skills and those of the two data scientists complemented each other, while making collective discoveries about the data yielded far more fruitful results than attempting to work in isolation. More importantly, she notes that contrary to her pre-conceived notions as to how big data scientists go about conducting research, her colleagues actually preferred to follow an inductive approach to analysing the Wikipedia dataset, and were also systematic in challenging any initial assumptions they had made about the data (Ford 2014:2).

2.2 Digital humanists and computational social scientists

What is humanistic about visualisation? – Elyse Graham (2017: 449)

... [Mounting] evidence suggests that many of the forecasts and analyses being produced [by powerful computational resources] misrepresent the real world.

– Derek Ruths and Jürgen Pfeffer (2014:1063)

Encouraging humanists/social scientists to collaborate with digital humanists/computational social scientists, on the other hand, might prove to be more difficult. If we look for a moment only at the picture in the digital humanities, it appears that traditional humanists feel obligated to keep justifying their research activities, criticising digital humanists for bowing to the pressure to prove that their research has utilitarian value, given that many institutions of higher learning have become corporatised. In this regard, a number of South African universities too appear to have surrendered to corporatisation (Clare & Sivil 2014:60). In 'The dark side of the humanities', Richard Grusin (2014:87) goes so far as to contend "that it is no coincidence that the digital humanities has emerged as 'the next big thing' at the same moment that the neoliberalisation and corporatisation of higher education has intensified

in the first decades of the 21st century”. Grusin (2014:87) also identifies another serious tension that exists owing to the notion that some humanists hold that digital humanists tend “to ‘make things’ rather than ... critically comment on issues”. In this respect, one of the more contentious issues for scholars in the literary field is that digital humanists use statistical processing and data aggregation techniques to analyse massive amounts of data to ‘read’ a text, which is a process referred to as distant reading (Moretti 2005), as opposed to hermeneutic close reading. Advocates of the former kind of reading assert that machine reading helps uncover aspects of a text (related to vocabulary, genre, and themes, for example) on a scale that is impossible for human beings to achieve. They also contend that visualisation techniques, which entail the graphical display of aspects of a text in the shape of maps, tag clouds or graphs to name a few, assist them in making sense of that text (Jockers 2013; Jänicke, Franzini, Cheema & Scheuermann 2015) and that big data can significantly improve on these visualisation techniques. Visualisation techniques are certainly not new as noted in Chapter 1 when we referred to John Snow’s and Florence Nightingale’s maps and graphs, respectively. In defense of (big) data visualisation in the digital humanities, Graham (2017:450) argues that traditional humanists mistakenly assume that data visualisation is merely a “discovery tool” when it actually “serves ... to refine arguments already made or illustrate conclusions already drawn”. Similar to the recommendation made by Ford (2014), Stockmann (2016), and Mills (2017) for team-work between traditional humanists and data scientists, Graham (2017:455) calls for closer collaboration between literary scholars and digital humanists, although, interestingly, this kind of suggestion is frowned on by some scholars. Urszula Pawlicka (2017:456), for instance, argues that collaboration “is less a development arising from new technologies than a response from within humanities departments to ‘neo-economic’ forces that have driven humanities departments to a point of crisis”. Nevertheless, a review of the literature shows fruitful collaboration between literary scholars and digital humanists in a few instances, particularly when it comes to using visualisation to conduct both close *and* distant readings of texts (cf. Wang Boldonado, Woodruff & Kuchinsky 2000) (see Chapter 7).

Karin van Es and Mirko Schäfer (2017) suggest two possible solutions to the concerns traditional humanists have when it comes to the digital humanities. First, they assert that these scholars should not lose sight of the fact that the digital humanities is not a new field – that the term “is merely the *nom de guerre* of the computational turn in the humanities”, which is another reminder not to be swept up in the hype of the big data evolution (van Es & Schäfer 2017:15). Indeed, these scholars anticipate that rather than replacing existing methodologies in the traditional humanities, the digital humanities will simply expand traditional humanists’ current methods, which is a very similar prediction to the one made by Kitchin (2014b:10) that big data methods will not become a substitute

for the methodologies employed in the humanities and social sciences. Second, traditional humanists should embrace the opportunities the digital humanities has to offer with a view to studying society based on large quantities of data:

Because datafication is taking place at the core of our culture and social organization, it is crucial that humanities scholars tackle questions about how this process affects our understanding and documentation of history, forms of social interaction and organization, political developments, and our understanding of democracy (van Es & Schäfer 2017:14).

Proponents of digital humanities are divided into those who still follow the first wave of digital humanities (cf. Schnapp, Presner & Lunenfeld 2009), which (from the late 1990s to the early 2000s) was characterised by quantitative research, the belief being that the use of techniques such as textual mining, graphing, and mapping result in “methodological rigour and objectivity” (Kitchin 2014b:7-8), and those in the second wave, who assert that these kinds of techniques should only supplement traditional humanities methods (Kitchin 2014b:8). Individuals in the first camp are criticised for being mechanistic, encouraging superficial analysis “rather than deep, penetrating insight” (Kitchin 2014b:8) and unfortunately, as Kitchin (2014b:8) points out, computational social scientists are also sometimes guilty of reductionist techniques. By way of illustration, he describes a study in which a group of researchers mapped millions of tweets over a three-year period to determine which geographical areas in a city were more multilingual than others (Rogers 2013). The map became “the end-point” (Kitchin 2014b:6), since the researchers did not go beyond this preliminary stage to utilise social theory and consider context with a view to answering pertinent questions related to people’s social and economic backgrounds, for example.

However, the fact that big data is being utilised by social scientists cannot be ignored. Youtie, Porter and Huang (2017:65) correctly point out that little research has been carried out to determine the role that social scientists (and humanists) are beginning to play in informing the emergence of big data technologies. Drawing on 488 research papers, these scholars conducted a comprehensive review of early social science and humanities research in the area of big data. They concluded that the majority of the papers drew on specific knowledge sources when researching big data, namely, the Internet and society, big data and medicine, law and privacy, and business impacts studies (Youtie *et al.* 2017:67).

Political scientists, communication scholars, sociologists, and psychologists, to name a few, are beginning to consider how computational tools may best be exploited to help them gain deeper insights into people and their interactions with one another (Shah, Cappella & Neuman 2015:6). When we look at the connection between big data and computational

social science and at the challenges this entails, it might be a good idea to first define the latter in relation to the former. Shah *et al.* (2015:7) define the term as follows:

It is an approach to social inquiry defined by (1) the use of large, complex datasets, often – though not always – measured in terabytes or petabytes; (2) the frequent involvement of “naturally occurring” social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability to social theory in a variety of domains from the study of mass opinion to public health, from examinations of political events to social movements.

As is the case for traditional social scientists, some computational social scientists remain sceptical about big data’s usefulness, while others have adopted an extreme view, arguing that the methodology behind big data is superior to that employed by computational social scientists. Proponents of such a view include Matthew Hindman (2015:48), who provocatively states that “[analytic] techniques developed for big data have much broader applications in the social sciences, outperforming standard regression models even” and Jimmy Lin (2015:33), who is of the view that “[if] the end goal of big data use is to engineer computational artifacts that are more effective according to well-defined metrics, then whatever improves those metrics should be exploited without prejudice”. Some scholars take a more moderate stance, such as Shah *et al.* (2015:9), who cast doubt on the notion that big data will entirely replace surveys, clinical trials, and analysis.

Mirroring the trepidations of humanists and social science scholars unfamiliar with big data, computational social scientists’ concerns revolve around ethics, the subordination of theory to data, and the very real danger of data validity being compromised, given the oftentimes questionable representativeness of social data, concerns we interrogate more closely in Chapters 3 and 4. For geography scholar David O’Sullivan (2017:9), a greater problem – and one that Kitchin (2015a, 2015b) has brought up more than once – lies in how big data represents *processes* in computational social science. What he finds problematic is that huge datasets only create the impression of capturing dynamism, and that many big data scientists never actually explain the processes behind the changes that must surely be taking place in the data: “[process] and change are ... rendered as ‘one damn thing after another’ with no notion of process or mechanism in the data themselves” (O’Sullivan 2017:9). In his view, a bottom-up, emergentist approach (which is commonly associated with complexity science) rather than a top-down aggregate one is better suited to process and explanation.¹⁹ Table 2.1 illustrates the main differences between complexity science and big data:

19 Digital humanities is associated with the former approach. (See Burdick, Drucker, Lunenfeld, Presner & Schnapp’s (2012) *Digital_humanities* in this regard.)

Table 2.1: Two approaches to big data analysis (O’Sullivan 2017:15)

| Complexity science | Big data |
|------------------------------------------------|-----------------------------------------|
| Embeds theory in models | Correlation and classification |
| Represents process | Temporal snapshots |
| Open-ended exploration of process implications | Exploration of already-collected data |
| Bottom up | Top down |
| Multiple levels and scales | Two levels: aggregate and individual |
| Many alternative histories (or futures) | ‘Just the facts’ (or optimal solutions) |

As the table illustrates, complexity science and big data appear to reflect divergent methodological perspectives and intellectual styles in the context of the digital humanities and computational social sciences. Yet it would be a mistake to conclude that the two traditions are worlds apart. Indeed, “[both] are about fitting simple models to observation” (O’Sullivan 2017:17), while complexity science may complement big data. Of course this does not mean that complexity science is *the* answer to the various questions posed by researchers: “[just] as it is foolish to believe that data-mining Big Data can provide answers to every social science [and humanist] question, it would be foolish to argue that simple complexity science models can answer every question” (O’Sullivan 2017:18).

We hold the view that complexity theory is of value to both (digital) humanists and (computational) social scientists, given that it helps foster a deeper understanding of human and social phenomena. In this respect, Youngman and Hadzikadic’s (2014) scholarly publication *Complexity and the human experience* discourses at length on quantitative analyses currently being undertaken in the humanities and social sciences. Their work may be regarded as ground-breaking since it applies the principles of complexity theory to research fields we would not generally associate with it. These fields include literature, anthropology, political science, and sociology. However, we also acknowledge the value of a qualitative approach to the analysis of big data as several chapters in this book reveal. Before doing so, we first turn to uncovering some of the common misperceptions about big data analysis that persist in the humanities and social sciences.

Big Data, big despair: Myths debunked and lessons learned

The temptation to form premature theories upon insufficient data is the bane of our profession. – Sherlock Holmes²⁰

The misconceptions that surround big data are numerous, which is perhaps not surprising given the fact that scholars unfamiliar with big data are regularly confronted by messages such as *Volume is all that truly counts* (cf. Jagadish 2015), *Big data does not need theory* (cf. Bowker 2014), and *Big data is not generally associated with qualitative research methods* (cf. Chen & Zhou 2017). On the other hand, scholars are also bombarded by messages that contradict these admonitions. It is somewhat confusing for researchers to also encounter claims such as *Small data sources can become big data sources* (cf. Ferguson, Nielson, Cragin, Bandrowski & Martone 2014) and *Qualitative research methods may be employed to analyse large datasets* (cf. Housley, Dicks, Henwood & Smith 2017; Mills 2017).

3.1 Epic fails

Data, in the wrong hands, whether malicious, manipulative or naïve can be downright dangerous. – Donald Clark (2013:1)

Globally, the term ‘big data’ has a tendency to engender fear, mistrust and – perhaps most worrisome of all – inertia among many humanists and social scientists, and the situation appears to be no different in South Africa. At the time this chapter was drafted, an Internet search of what South African universities are doing to prepare humanists and social scientists for the era of big data yielded very little information.²¹ What appears to partially drive negative attitudes towards big data has to do with a number of myths surrounding this meme, myths which we aim to explore and debunk in and across several chapters in this book. At the risk of sounding contradictory, we would like to invite humanists and social scientists to be sceptical about big data – at least for now. Scepticism is not an unhealthy inclination

20 Arthur Conan Doyle (1915), *The valley of fear* (George H. Doran Company).

21 In South Africa at least, big data analysts appear to be mainly big data scientists who have expertise in new statistical and programming skills. Lack of computational skills may be one reason why social scientists and humanists have been slow to tap into the benefits of big data theory.

in the face of the big data deluge (Davenport 2014:2). Indeed, “big-data related efforts have had as many failures as they have [had] successes” (Schilling & Bozic 2014: 3272). To date, Google Flu Trends or GFT is arguably the most well-known big-data foible of the millennium (Lazar, Kennedy, King & Vespignani 2014:1203). Attempting to accurately forecast the outbreak of influenza in some 25-odd countries from 2008 onwards, this web service was 140 percent off in its predictions in 2013 due to flawed algorithmic dynamics.²² In addition to human error and misinterpretation of data, GFT did not take into account that online users searching for ‘cough’, ‘headache’ or ‘fever’, for instance, might have been looking for information on topics not related to influenza-related illnesses. In this chapter, we interrogate what can be mined (pun intended) from the GFT failure and others as well as from the misconceptions surrounding big data, centering our discussion on how various lessons may be exploited by social scientists and humanists carrying out big data research.

3.2 Big data lessons

3.2.1 Lesson 1

With big data comes big noise. – Beth Mole (2015:1)

Above all else, assess data quality.

In ‘The parable of Google Flu’, Lazar *et al.* (2014) caution against what they refer to as big data hubris, a term they employ to describe the penchant among some analysts to entirely replace traditional data collection and data analysis methods with big data. Relying exclusively on massive quantities of data generated by Internet services, what developers of the first version of GFT failed to do was to employ instruments aimed at establishing valid and reliable data:

The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC [Centre for Disease Control] data, such as those regarding high school basketball ... (Lazar *et al.* 2014:1203).

22 In fact, “the algorithm has been criticised for overfitting a small number of cases and masking a simple question, namely, does it predict flu or is it merely reflecting the incidence of winter...” (Agarwal & Dhar 2014:446).

They add that “[t]his should have been a warning that the big data were overfitting the small number of cases, a standard concern in data analysis” (Lazar *et al.* 2014:1203). The first lesson for social scientists and humanists? To be of value to scholarship, any big data collected and analysed should be accurate and of high quality. Is this easier said than done? To answer this question, we reviewed a study by Cai and Zhu (2015) who have proposed what we perceive to be a useful data quality framework and accompanying dynamic assessment process to assess the quality of data. Any scholar new to big data needs to first be aware that big data cannot be simplistically defined as data that is too large to fit into an Excel spreadsheet, for example (Kitchin 2014b:1). Rather, big datasets reflect distinct features or vectors commonly referred to as the five Vs,²³ namely, volume, velocity, variety, veracity, and value (Katal, Wazid & Goudar 2013). Volume is self-explanatory, whether it is measured in the number of records, the amount of storage space it requires or the completeness of the dataset. Velocity has to do with the need to analyse big datasets in a timely fashion, since data is produced at breakneck speed and changes just as quickly. A defining and third feature of big data has to do with the variety of sources and content available to researchers, and the data are in turn divided into structured, semi-structured, and unstructured data demanding sophisticated data processing capabilities (see Chapter 8). Veracity refers to the trustworthiness of the data collected and analysed, and finally, value refers to the worth of the data, since huge amounts of data have no value if worthwhile information and knowledge cannot be extracted from them. Being aware of these features is crucial if data quality is to be accurately assessed.²⁴ With these vectors in mind, Cai and Zhu (2015)²⁵ suggest that analysts use their hierarchical data quality framework illustrated in Table 3.1

23 We only touch on big data’s four Vs here; Chapter 8 provides a more detailed, technical account of the technologies required to manage huge volumes of data.

24 Some users of big data have added additional Vs depending on their research objectives, and these are visualisation, variability, and value (see Chen & Zhang (2014) for a definition and explanation of each vector). Some scholars have also added exhaustivity which demands that the data collected represent the entire population under investigation (Kitchin & Lauriault 2015:464).

25 Of interest is the observation by some big data analysts that researchers favour specific dimensions over others, depending on their disciplines. Researchers who focus on the Internet, for example, emphasise velocity, while humanists and social scientists highlight value and veracity. In this respect, see Hitzler & Janowicz (2013).

Table 3.1: Cai and Zhu's (2015:5) big data quality framework

| Dimensions | Elements | Indicators |
|----------------------|---------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Availability | <p><i>Accessibility:</i></p> <p><i>Timeliness:</i></p> | <p>Whether a data access interface is provided</p> <p>Data can be easily made public or easy to purchase</p> <p>Within a give time, whether the data arrive on time</p> <p>Whether data are regularly updated</p> <p>Whether the time interval from data collection and processing to release meets requirements</p> |
| Usability | <p><i>Credibility:</i></p> | <p>Data come from specialized organisations of a country, field, or industry</p> <p>Experts or specialists regularly audit and check the correctness of the data content</p> <p>Data exist in the range of known or acceptable values</p> |
| Reliability | <p><i>Accuracy:</i></p> <p><i>Consistency:</i></p> <p><i>Integrity:</i></p> <p><i>Completeness:</i></p> | <p>Data provided are accurate</p> <p>Data representation (or value) well reflects the true state of the source information</p> <p>Information (data) representation will not cause ambiguity</p> <p>After data have been processed, their concepts, value domains, and formats still match as before processing</p> <p>During a certain time, data remain consistent and verifiable</p> <p>Data and the data from other data sources are consistent or verifiable</p> <p>Data format is clear and meets the criteria</p> <p>Data are consistent with structural integrity</p> <p>Data are consistent with content integrity</p> <p>Whether the deficiency of a component will impact use of the data for data with multi-components</p> <p>Whether the deficiency of a component will impact data accuracy and integrity</p> |
| Relevance | <p><i>Fitness:</i></p> | <p>The data collected do not completely match the theme, but they expound one aspect</p> <p>Most datasets retrieved are within the retrieval theme users need</p> <p>Information theme provides matches with users' retrieval theme</p> |
| Presentation quality | <p><i>Readability:</i></p> | <p>Data (content, format, etc.) are clear and understandable</p> <p>It is easy to judge that the data provided meet needs</p> <p>Data description, classification, and coding content satisfy specification and are easy to understand</p> |

The data quality assessment process itself involves following a number of specific steps outlined in Figure 3.1.

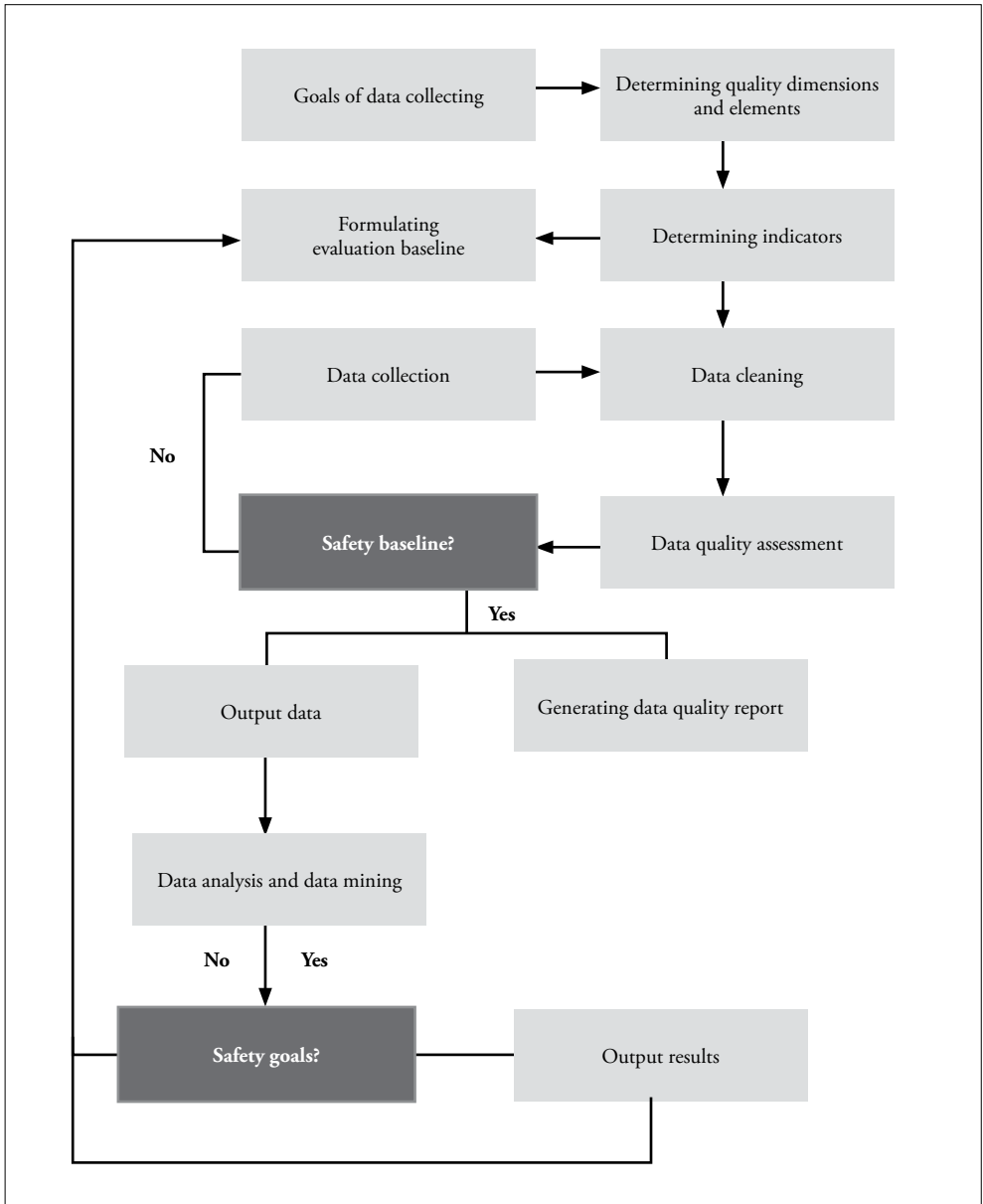


Figure 3.1: Cai and Zhu's (2015:7) quality assessment process

Although both the framework and process were specifically designed for commercial purposes, we argue that they may be adapted to suit different research environments in the sense that academics may simply select data quality dimensions that will help them achieve their research objectives. Let us consider a scenario in which social scientists would like to explore the attributes of civil and uncivil discourse online. To achieve this goal, they decide to collect posts generated by commenters on online news sites. When it comes to determining quality dimensions, the need for them to take timeliness, accuracy, and completeness into account will obviously be prioritised (Cai & Zhu 2015:7). Since social media data is raw data, it will also be necessary to assess its credibility (Cai & Zhu 2015:7). Dimensions such as consistency and integrity will not be particularly useful to them, since social media data is generally unstructured (Cai & Zhu 2015:7). The next step will be to select indicators for every dimension chosen. Thus, for example, in terms of timeliness and completeness, the researchers will have to be aware that commenters' posts on online news sites are regularly updated, and that they will therefore have to make sure that they access and download the latest posts so that their dataset is complete. They will also have to ensure that they do not miss the window of opportunity for downloading data, since online news outlets may close down comments sections at specific times and without prior warning. Assessing the quality of each dimension will allow the researchers to determine if they are satisfied with the baseline standard. If the baseline standard is satisfactory, they will draft a quality report and then enter the data acquisition phase during which the data is scrubbed or cleaned with the aim of “[detecting] and [removing] errors and inconsistencies from data in order to improve their quality” (Cai & Zhu 2015:7) (see Chapter 7 as well). A little later on in this chapter, we more closely consider how the use of big data complicates the research process when it comes to assessing data quality, but for now we assume that the researcher is ready to analyse his data, which brings us to the next lesson.

3.2.2 Lesson 2

[In the age of big data] [q]ualitative research can illuminate questions that need asking and real problems that need solving. – Elizabeth Kaufer (2016:1)

The growth of quantitative technique may be counterbalanced by qualitative understanding. – Joshua Fairfield and Hannah Shtein (2014:49)

Think qualitatively (too).

Probably one of the most important lessons for social scientists and humanists is that the birth of big data is not, in fact, the end of the scientific method, a provocative pronouncement made by Chris Anderson in 2008 when he was the editor-in-chief of *Wired*. Anderson (2008:5)

claimed, amongst other things, that “[p]etabytes allow us to say: ‘Correlation is enough.’ We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot”. In a tongue-in-cheek response to the end of theory declaration, Geoffrey Bowker (2014:1797-1798) puts it aptly when he observes that “this is a massive reduction of what it means to ‘know’” and that “any ‘thing’ we create (object, way of looking at the world) embodies theory and data”. Interestingly, the notion that big data is the re-birth of empiricism is one that anti-big-data researchers cannot seem to shake, yet the paradox lies therein that “it is not just empirical-quantitative” (Cope & Kalantzis 2015:226) in nature. In fact, scholars have echoed Bowker’s (2014) sentiments, reiterating that “[big data] demands more conceptual, theoretical, interpretative, hermeneutical – indeed qualitative – intellectual work than ever” (Cope & Kalantzis 2015:227).

The importance of qualitative thinking is echoed by a number of other researchers who are interested in exploring big data. Two of these are boyd and Crawford (2012:670), who argue that “there remains a mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts”. They offer the caveat that such narrow thinking means that “[big data] risks reinscribing established divisions in the long running debates about scientific method and the legitimacy of social sciences and humanistic inquiry” (boyd & Crawford 2012:67). Mazzochhi (2015:1253) also expresses it well when he maintains that

[scientific] research does not take place in a purely theoretical and rational environment of facts, experiments and numbers. It is carried out by human beings whose cognitive stance has been formed by many years of incorporating and developing cultural, social, rational, disciplinary ideas, preconceptions and values, together with practical knowledge.

Where does the notion come from that big data makes theoretical assumptions and hypotheses redundant? We know that this is certainly not a new idea, and that it harks back to Francis Bacon’s (1620) *Novum organum* in which he subordinated the testing of theory to observation and analysis. Isaac Newton (1713) too was unimpressed by theory, famously remarking “hypotheses non fingo” (“I frame no hypothesis”) in the second edition of the *Principia*. In the last few years, some advocates of big data have argued quite convincingly in favour of correlation being superior to causation, notably Mayer-Schönberger and Cukier (2013) in their book entitled *Big Data: A revolution that will transform how we live, work, and think* (Houghton Mifflin Harcourt). Mayer-Schönberger and Cukier (2013:14) maintain that “correlations may not tell us precisely why something is happening, but they alert us

that it is happening. And in many situations this is good enough”. We support Mazzocchi’s (2015:1252) view that the “no theory” thesis is not good enough for academics – that “understanding the why is crucial for reaching a level of knowledge that can be used with confidence for practical applications and for making predictions”. As social scientists or humanists, we thus need to analyse big data within specific theoretical and methodological limitations in order “to assign them a meaning and to distinguish between meaningful and spurious correlations” (Mazzocchi 2015:1252).

We would not want to create the impression that the majority of big data advocates are anti-theory, however. A number of researchers have called for big data projects to be driven or enhanced by theory (Frické 2013; Coveney, Dougherty & Highfield 2016; Sparks, Ickowicz & Lenz 2016; Wu, Buyya & Ramamohanarao 2016; Olshannikova, Olsson, Huhtamäki & Kärkkäinen 2017). Some researchers in the social sciences, notably, Monroe, Pan, Roberts, Sen and Sinclair (2015:71), maintain that big data and social science research methods are not entirely unsuited, and that big data may enhance these methods and “[enable] us to answer new questions”. In fact, they go on to argue that “it is the responsibility of social scientists to assume their central place in the world of big data” (Monroe *et al.* 2015:71) because a great deal of big data reflects social data. How much data is sufficient in the world of social scientists though?

3.2.3 Lesson 3

More data do not necessarily generate more knowledge.

– Fulvio Mazzocchi (2015:1253)

Be mindful that size is relative.

From a computational point of view, big data is information that is so large it can only be retrieved and organised with the assistance of special tools discussed in Chapter 8. However, and here we debunk yet another myth, “from a social sciences perspective big data is not always that big” (Beneito-Montagut 2017:915). What is important is not size, but rather accessing “more data, quicker and richer than before” (Beneito-Montagut 2017:915). What scholars should bear in mind is that the size or volume of data differs from industry to industry, and is therefore not the defining characteristic of big data (Gandomi & Haider 2015:143). Indeed, in the context of the humanities or social sciences, “[big data] can safely be reduced to medium-size data and still yield valid and reliable results” (Mahrt & Scharkow 2013:28), provided that validity is achieved when it comes to the sampling process. Of course, the immediate question then becomes, *But what about the generalisability of one’s findings?* It may be surprising to academics working in the social sciences and in the humanities to learn that more data from a particular source is not needed to achieve generalisability. What *is* required

is “horizontal expansion” (Mahrt & Scharnow 2013:26) or more data from multiple data sources (Mahrt & Scharnow 2013:26). To understand what this means in practical terms, we briefly describe a research study by Gray, Jennings, Farrall and Hay (2015) significantly entitled “Big small data”.

Exploring social and economic change at a national level in Britain, Gray and her team (either criminologists or political scientists) were mindful of the need to make use of large datasets while staying true to “the essential social and cultural heredity that is intrinsic to the human sciences” (Gray *et al.* 2015:1). One of their preliminary findings in the context of economy and crime rates was that property crime spiked when unemployment levels rose, which in turn increased citizens’ fear of crime as well as the government’s attention to that crime. What the researchers needed in order to examine attitudes towards other kinds of crime was an integrated model of analysis that would allow them to construct a series of connected, multi-layered datasets that would help them not only to observe attitudinal shifts in relation to specific crimes, but also to examine changes at aggregate levels over a 30-year period. Ultimately, the dataset comprised of individual-level data such as victimisation and social attitudes and aggregate-level data such as socio-economic indicators, official statistics on crime, public opinion data, and policy documents. Although not hailing big data as the magic bullet, Gray and her colleagues were able to harness the power of *computational technology* and *statistical techniques* to exploit the richness of high-volume datasets.

Given the need for computational skills and in light of the previous chapter’s focus on, amongst other things, the divide between social and data scientists, readers may at this stage be asking whether they too will be able to develop the necessary skills-set to make use of big data. This takes us to the next lesson which has to do with a radical shift in the way social scientists and humanists think about research.

3.2.4 Lesson 4

For the analysis of big data to truly yield answers to society’s biggest problems, we must recognise that it is as much about social science as it is about computer science.
– Justin Grimmer (2015:80)

Develop specific skills to manage and analyse big data, but remember that these skills are tied to a number of caveats.

One of the major implications of the big data evolution is that scholars working in the humanities and social sciences will require a fairly unique combination of skills that draw on the social sciences, computer science, and statistics (Miller 2011:1815). These skills are explored when we critically appraise big data software tools (in Chapters 7 and 8), but suffice it to say at this early stage, analysing big data involves thorough data profiling, “thoughtful

measurement ... [as well as] careful research design, and the creative deployment of statistical techniques” (Grimmer 2015:80). Not surprisingly, each of these phases comes with its own myriad set of pitfalls discussed below.

When it comes to data profiling and measurement, political theorists John Patty and Elizabeth Penn (2015:100) remind scholars that “‘data’ is nothing until we use it to measure something, and ‘measurement’ is accordingly *inherently* theoretical”. Expressed a little differently, the decision as to what to describe and what to omit makes theory unavoidable (cf. Lemire & Petersson 2017). In their study of how novice data users made sense of large-scale datasets, Faniel, Kriesberg and Yakel (2012) found that one of the major concerns of social scientists measuring big data had to do with first putting their data in context because without it, data has no meaning (cf. Keim, Qu & Ma 2013:21). Putting data into context has to do, amongst other things, with profiling, which encompasses evaluating the content and quality of datasets. We touched upon assessing quality a little earlier on in this chapter, but in the age of big data, it is not simply a case of transferring data into a statistical package, describing the basic features of the data, and then summarising those features. More often than not, social scientists who attempt to obtain (or ‘scrape’) big datasets are confronted by reams of raw or unstructured data. More than a decade ago, Geoffrey C. Bowker (2005:183-184), an expert in informatics at the University of California, wrote that “[raw] data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care”. Cooking the data is not without its challenges since the process is inescapably subjective in nature. Almost a decade ago, Bollier (2010:13) raised this concern when he asked “[c]an the data represent an ‘objective truth’ or is any interpretation necessarily biased by some subjective filter or the way that data is ‘cleaned’?”. Coupled to this subjectivity is the problem of what boyd and Crawford (2012:668) refer to as the tendency to practise apophenia – “seeing patterns where none actually exist”. They refer to physicist, mathematician, and computational scientist Leinweber (2007:1), who describes how, together with his colleagues, he was able to show that data mining techniques can easily be manipulated to produce ridiculous correlations in the financial world:

The example in this paper is intended as a blatant example of [a] totally bogus application of data mining in finance. We first did this several years ago to make the point about the need to be aware of the risks of data mining in quantitative investing. In total disregard of common sense, we showed the strong statistical association between the annual changes in the [Standard & Poor’s] 500 stock index and butter production in Bangladesh, and other farm products. Reporters picked up on it, and it has found its way into the curriculum at Stanford Business School and elsewhere. We never published it since it was supposed to be a joke.

An important lesson behind this hoax is that social scientists and humanities scholars need to be explicit about their methodological processes, elucidating how the interpretation of their data is clouded by their biases. Boyd and Crawford (2012:668) suggest that one way to be accountable for bias is to “[recognize] that one’s identity and perspective [inform] one’s analysis” every step of the way, which is not an alien notion for anyone carrying out qualitative research.

An additional challenge when it comes to data profiling pertains to the huge amounts of unreliable datasets available in cyberspace. In this regard, Desouza and Smith (2014:42) point to an ominous practice by the American Petroleum Institute in 2011 which involved manipulating sentiment via Twitter to such a degree that they were able to create the impression that vast numbers of farmers, environmentalists, and landowners supported a pipeline project between Alberta in Canada and Texas in the USA. The Rainforest Action Network (RAN) became suspicious when they discovered an unusual spike in pro-pipeline tweets. They were ultimately able to prove that many of these messages were generated via automated twitter bots.²⁶

On the topic of Twitter, scholars need to be cautioned about using social media platforms in general to collect and analyse data because these platforms entail a series of troubling methodological concerns. Below are a number of caveats issued by Boyd and Crawford (2012:669) when it comes to mining data from Twitter. Their warnings could just as easily apply to other social media platforms.

1. Twitter does not represent the global population.
2. Some Twitter accounts are generated by social bots that people may be convinced are authentic internet personas.
3. Twitter accounts may reflect either active users or passive participants – individuals who actively post tweets versus those who are simply “listeners” (Crawford 2009:525).
4. An individual who tweets may hold multiple Twitter accounts, while the reverse may also be true: an account could be used by multiple individuals.
5. Data from Twitter may be skewed in the sense that its gatekeepers are able to filter out posts deemed to be uncivil or racist, for example.
6. Even more interestingly, people have ‘fire-hose’, ‘garden-hose’, or ‘spritzer’ access to Twitter. Twitter Incorporated’s so-called fire-hose supposedly contains all public tweets that have been posted to date, while garden-hose access reflects approximately ten percent of all public tweets. A spritzer contains a mere one percent of these tweets. It may surprise (qualitative) scholars unfamiliar with the nature of big data to know that very few researchers actually have access to the fire-hose.

26 These bots autonomously perform tweeting and re-tweeting; they follow tweeters and control their accounts via the Twitter API. API is an abbreviation for application programming interface; it is a software intermediary enabling two applicants to communicate with each other.

At the heart of the matter, then, is that determining data context does not simply hinge on exploiting what technology has to offer in terms of web scraping, text mining, data integration, pattern recognition, and the like. Instead, big data analysis demands special domain knowledge and expertise, not to mention the proper skills to analyse the given dataset accurately (boyd & Crawford 2012).

When it comes to research design and data analysis, both digital and social media researchers have not been swift to take advantage of what some scholars refer to as the computational turn or “data gold rush” (Kennedy, Moss, Birchall & Moshonas 2015:172). In fact, after conducting a meta-analysis of big data social media research by communication studies scholars, Mylynn Felt (2016) discovered that very few of them opted to exploit big data analytics. Felt (2016:13) speculates that low usage of data analytics may be due to the methods being so different from traditional social science research methods. Yet data analytics can certainly complement more traditional methods. Bogdan Batrinca and Philip Treleven (2015) provide, amongst other things, a useful taxonomy of social media analytics tools for analysis of large datasets. This taxonomy includes scientific programming tools, business toolkits, social media monitoring tools, text analysis tools, and data visualisation tools. In subsequent chapters, we review these tools (as well as the relevant techniques and platforms associated with them), placing emphasis not only on the methodologies that underlie them, but also on *critiquing* these tools, the majority of which are unfortunately “commercial, expensive and difficult for academics to obtain full access [to]” (Batrinca & Treleven 2015:92).

Now which data analytics tool to select depends largely on the epistemological and theoretical underpinnings of any given scholar’s research (Felt 2016), and in this regard we turn to Rob Kitchin (2014b) for several lessons regarding emerging epistemological positions in the context of the social sciences and humanities.

3.2.5 Lesson 5

Big Data and new data analytics are disruptive innovations which are reconfiguring in many instances how research is conducted. – Rob Kitchin (2014b:1)

Consider developing a reflexive epistemology in the context of big data projects.

Earlier in this chapter, we cautioned scholars to also think qualitatively about big data. Thinking qualitatively involves, amongst other things, a careful re-consideration of – and critical reflection on – “the epistemological implications of the unfolding data revolution” (Kitchin 2014b:1). Kitchin (2014b) observes that there are currently two paths in the natural, life, engineering, and physical sciences, paths that reflect entirely dissimilar epistemologies, namely, empiricism and data-driven science, which we touched upon a little earlier on. While empiricism involves collecting data and then letting that data speak for itself (*sans* theory),

data-driven science adheres to the principles of the scientific method. Presently, many scholars advocate the need for data-driven science, predicting that it “will become the new paradigm of scientific method in an age of big data because the epistemology favoured is suited to extracting additional, valuable insights that traditional, ‘knowledge-driven science’ would fail to generate” (Kitchin 2017:32).

Significantly, and as we noted in the previous chapter, big data and new analytics will probably not result in entirely new paradigms for researchers in either (computational) social science or the (digital) humanities, “given the diversity of their philosophical underpinnings” (Kitchin 2014b:12). Kitchin (2014b:9) makes the following observations about computational social science and the digital humanities in this regard:

Whereas most digital humanists recognize the value of close readings, and stress how distant readings complement them by providing depth and contextualization, positivistic forms of social science are oppositional to post-positivist approaches. The difference between the humanities and social sciences in this respect is because the statistics used in the digital humanities are largely descriptive – identifying and plotting patterns. In contrast, the computational social sciences employ the scientific method, complementing descriptive statistics with inferential statistics that seek to identify associations and causality. In other words, they are underpinned by an epistemology wherein the aim is to produce sophisticated statistical models that explain, simulate and predict human life.

Although neither big data empiricism nor data-driven-science appears to be making significant inroads when it comes to the epistemologies of these two disciplines (Kitchin 2014b:7), big data nevertheless “presents a number of opportunities for social scientists and humanities scholars” (Kitchin 2014b:10) who need to view big data far more critically than commercial analysts do (Mahrt & Scharkow 2013:25). When critically reflecting on the kinds of epistemologies that could be combined with big data, we take a leaf out of the book of Mahrt and Scharkow (2013:30), who suggest that scholars should focus on combining data-driven and theory-driven operationalisation strategies, but with a greater emphasis on the latter. Coupled to this recommendation is a proposal related to research design as well as to the methodological training that social scientists and humanists should receive:

In general, we should resist the temptation to let the opportunities and constraints of an application or platform determine the research question; the latter should be based on relevant and interesting issues – regardless of

whether something is available through an API platform or seems easily manageable with a given analytical tool. Methodological training ... should not only focus on computational issues and data management, but also continue to stress the importance of methodological rigor and careful research design. This includes a strong need for theoretical reflection, in clear contrast to the alleged ‘end of theory ...’ (Mahrt & Scharkow 2013:30).

Both digital humanities and computational social science have their detractors. The former is frequently criticised for its “weak, surface analysis ... [and its] overly reductionist and crude ... techniques” (Kitchin 2014b:8). The latter is seen “as being mechanistic, atomizing, and parochial, reducing diverse individuals and complex, multidimensional social structures to mere data points” (Kitchin 2014b:8).

3.2.6 Lesson 6

Thick Data can rescue Big Data from the context-loss that comes with the processes of making it usable. – Tricia Wang (2013:4)

... there has been a lack of attention paid to the ethical obligation of transparent and complete reporting of studies using large-scale ... datasets.

– Stuart Nicholls, Sinéad Langan and Eric Benchimol (2016:339)

Big (social) data needs thick description and transparent reporting.

Social scientists and humanists usually think about making small datasets “thick” (Latzko-Toth, Bonneau & Milette 2017:199), but according to Tricia Wang (2013:9), who describes herself as a global tech ethnographer, using thick description “is a great opportunity for qualitative researchers to position [their] work in the context of [big data’s] quantitative results” as well. This view is echoed by a number of social scientists such as Felt (2016:14), who argues that computational analysis (combined with qualitative methods) can benefit from thick description because it enables scholars to see “both the big picture and the close, critical view”. This in turn enhances transparency, which “mandates that researchers have an ethical obligation to facilitate the evaluation of their evidence-based knowledge claims” (Moravcsik 2014:665). Graham (2017:449) points out that “while the scientist’s methods can be paraphrased without any loss, in the humanities, the description itself is understood to be part of the method”. In this respect, thick description is crucial; it forms part of the researcher’s thinking and sense-making process.

Tellingly, a review of the literature over the last decade signals a paucity of research on the connection between thick description and big data analysis. Most studies of this research concept continue to be tied to qualitative work and to small datasets. Even more significant is the fact that very few big data studies even mention transparency obliquely, although to their credit, Lazar *et al.* (2014:1205) explicitly refer to lack of transparency as one of the most critical mistakes made by the Google Flu Trends team referred to a little earlier. Both transparency and thick description are critical if we want to achieve ethical research and this brings us to the next important big data lesson.

3.2.7 Lesson 7

But it's already public, right? – Emily Wolfinger (2016:1)

Accessible online data does not always translate into ethical practice.

Perhaps a rather unpalatable truth to have to digest is that although vast quantities of data are now available online, sound ethical practice cannot be disregarded by anyone carrying out big data research. The assumption persists that online data is public data and that ethics is therefore not an issue. Yet we cannot ignore what constitutes a 'public' or a 'private' space in Internet-based research (IBR). A literature review of IBR ethics supports the necessity of obtaining informed consent in private domains and waiving it in public domains (Convery & Cox 2012:51), but the private-public space distinction is a contentious one as borne out by questions such as the following: If an online participant regards his/her posting as private, but communicates in a public space such as Twitter or Instagram, is the space deemed to be private or public? Can a domain be semi-private or semi-public? The literature provides us with some partial – and in some cases, ambiguous – answers which are related to the issue of informed consent in IBR:

1. “Public space is defined as the space that applies no restriction to interaction and communication, whereas isolated space (private space) is one that completely constrains communication (Georgiou, 2006)” (Jang & Callingham 2012:76).
2. “A study with archival data and pre-existing resources publicly available might waive obtaining informed consent if the study does not include individual information and sensitive topics” (Jang & Callingham 2012:76).
3. “One position is that data posted in open spaces without password or membership restrictions would usually be considered in the public domain and so available for research use without the need for informed consent from individual contributors ...” (Beninger, Fry, Jago, Lepps, Nass & Silvester 2014:6).

4. Currently, there is no consensus as to whether or not scholars should obtain informed consent if they wish to conduct Internet-based research (Gao & Tao 2016:185).

To complicate matters, some researchers working in the field of IBR contend that we cannot ignore what an online participant perceives to be a private or public domain. A study that focused on chat-room participants, for example, found that these participants perceived the chat-room space to be private (Hudson & Bruckman 2004). In a similar study, researchers discovered that bloggers may regard their blogs as constituting private diaries, and may therefore not be in favour of public scrutiny of their thoughts (Gao & Tao 2016). We feel so strongly about the importance of abiding by ethical guidelines in big data research that ethics is the focus of the next chapter.

Big Data needs big ethics

[Big data technology] is inherently ethics-agnostic. – Mandy Chessell (2014:1)

The above statement appears to be provocative, but Chessell (2014:1) goes on to qualify it by arguing that researchers themselves need to be more thoughtful when it comes to *how* they use the technologies at their disposal. In this chapter, we interrogate the ethical issues behind big data research, beginning with big data blunders that underscore the urgent need for a sound code of ethical practices. Thereafter, we consider some of the major challenges of big data ethics that pertain to the use of human subjects, the dilemma inherent in the public-private space debate, the need for informed consent and anonymisation of data, the problem of representativeness,²⁷ and the ethical drawbacks to the new digital ecosystem.

4.1 Big data *faux pas* of the millennium

Outrage has a way of shaping ethical boundaries. – Sarah Zhang (2016:1)

Big data is not simply about major epistemological adaptations, but also about significant changes at the level of ethics (Ambrose 2015:214). Yet, the world has seen quite a number of embarrassing and serious gaffes in terms of big data ethics in the last decade, one of the biggest being the 2016 release by Danish researchers of 70000 OKCupid profiles, complete with information relating to users' genders, personality traits, and sexual preferences (Kirkegaard & Bjerrekaer 2016). What should concern big data scholars is that in an attempt to preempt any criticism, the researchers inserted a disclaimer to the effect that “[s]ome may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form” (Kirkegaard & Bjerrekaer 2016:2). What is more, a quick scrutiny of the tweets that followed the publication of the study shows that the researchers did not immediately respond to users' concerns about their lack of ethical consideration for OKCupid users.²⁸ The only response that was forthcoming came from the lead researcher, Emil Kirkegaard, who tweeted, “No. Data is already public”.²⁹ Following public outrage, OKCupid filed a

27 See Chapter 3 as well.

28 <https://twitter.com/KirkegaardEmil/status/730449904909324289>.

29 Ibid.

Digital Millennium Copyright Act (DMCA) and the OKCupid dataset was subsequently removed by the Open Science Framework (Resnick 2016).

The “it’s-in-the-public-domain” argument is not new according to privacy and Internet ethics scholar, Michael Zimmer (2010). In fact, Zimmer (2010) exposed an ethics scandal in 2008 when Harvard sociologists Lewis, Kaufman, Gonzalez, Wimmer and Christakis (2008), in an attempt to track the evolution of friendships over time, published “Tastes, ties and time” which comprised an ostensibly anonymous dataset of 1700 college students. Unfortunately, the database could without too much effort be traced back to Harvard’s college class of 2009. The “already public” disclaimer was also employed by a former Apple engineer, Pete Warden, who, in attempting to create his own search engine, exploited a flaw in Facebook’s architecture to collect 215 million Facebook users’ accounts in 2010. Shortly after this mining exercise, Warden announced that he would make this database public for purposes of academic research (cf. Zimmer 2016:2). He deleted the entire database following Facebook’s decision to take legal action against him (cf. Zimmer 2016:4).

Clearly, Zimmer’s (2010) earlier warnings about ethics have not been heeded by all; as early as 2010, and in the context of the Harvard Facebook scandal, Zimmer (2010:323-324) wrote:

The ... research project might very well be ushering in “a new way of doing social science,” but it is our responsibility as scholars to ensure our research methods and processes remain rooted in long-standing ethical practices. Concerns over consent, privacy and anonymity do not disappear simply because subjects participate in online social networks; rather, they become even more important.

Zimmer (2010:323-324) is quick to point out that the aim of his study was not to condemn or disparage the Harvard analysts, but rather to present their research as a case study of the ethical challenges inherent in big data research.

Perhaps the worst ethics scandal to rain on big data research to date is reflected in a 2012 study published in *Proceedings of the National Academy of Sciences* by Kramer, Guillory and Hancock (2014) in which 700000 Facebook users’ news feeds were subtly manipulated to create either negative or positive messages. Kramer *et al.* (2014:8788) concluded that “[e]motional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness”. What is highly problematic is that Facebook users were not informed that their news feed algorithms would be tweaked, resulting in major outcries from these users after the study was made public. The most that

Adam Kramer would say on Facebook was that “[in] hindsight, the research benefits of the paper may not have justified all of this anxiety”.³⁰

One of the oddest big data studies to showcase ethical dilemmas is by Michelle Hauge and her colleagues published in the *Journal of Spatial Science* in 2016. Using a statistical inference technique known as geographical profiling, which is essentially employed to find individuals suspected of serious crimes such as rape, murder, and terrorism, Hauge, Stevenson, Rossmo and Le Comber (2016) attempted to track down a well-known pseudonymous graffiti artist based in the United Kingdom, Banksy, who prefers to stay entirely out of the public eye. The researchers even mined electoral rolls and other public websites to trace not only the artist’s former residential addresses, but also those of his wife. As Metcalf and Crawford point (2016:2) out,

[t]here are many questions that could be asked of this study, not least about the correlation between graffiti and terrorism. But for our purposes, we will only focus on the ‘ethical note’ that appeared at the end of the article: “the authors are aware of, and respectful of, the privacy of [subject name removed] and his relatives and have thus only used data in the public domain” (Hauge *et al.* 2016: 5). This claim is particularly striking, as it is difficult to see how tracking a specific individual (and their family) to such an invasive degree could be considered respectful of their privacy.

Metcalf and Crawford (2016:2) go on to argue that this research offers a case study for the invasiveness of public data and for the mistaken belief among big data analysts that data’s existence in the public domain means that ethical clearance is never necessary.

4.2 A review of the literature

4.2.1 Current challenges

When applying ethical inquiry to any new method or practice, it can easily shift into ethical relativism, where no fixed principles universally apply to any situation that may arise. – James Willis, John Campbell and Matthew Pistilli (2013:7)

Eventually, ethicists will have to continue to discuss ... how we can prevent the abuse of Big Data as a new found source of information and power.

– Andrej Zwitter (2014:5)

30 https://web.facebook.com/akramer?_rdc=1&_rdr.

A plethora of studies have been published over the last few years that detail researchers' concerns about some big data analysts' disregard for ethical considerations.³¹ Legal scholars Richards and King (2014:395), for example, argue that

[w]e are building a new digital society, and the values we build or fail to build into our new digital structures will define us. Critically, if we fail to balance the human values that we care about, like privacy, confidentiality, transparency, identity, and free choice, with the compelling uses of big data, our big data society risks abandoning these values for the sake of innovation and expediency.

Other researchers who echo these sentiments include Davis (2012), boyd and Crawford (2012), Lyon (2014), and Metcalf and Crawford (2016). The latter two scholars identify emerging problems that are currently related to ethics and big data research. The first has to do with the ever-increasing divide between research ethics in traditional disciplines and the methodologies underlying big data: “[big] data research methods exacerbate a long-standing tension between the social sciences and research regulations that are geared to the methods and harms of biomedical research” (Metcalf & Crawford 2016:1). In traditional disciplines, we have become used to exercising moral agency – protecting human subjects so as not to cause unjustified harm (Zwitter 2014:2). The second problem, which we have already touched upon, is one that Metcalf and Crawford (2016) paint as a North American dilemma – that US research policy regards studies that exploit digital data to pose minimal risks to human subjects because the data is in the public domain. In South Africa, the scenario is not quite the same, but it is nevertheless worrying that most websites dedicated to big data ethics focus exclusively on research for commercial purposes or government projects, while a search of South African universities' stance(s) on big data ethics in the humanities/social sciences has signalled that ethics review boards have no formal guidelines on conducting big data research: “there are very few examples of how institutions respond to the ethical challenges and issues [in big data research]” (Prinsloo & Rowe 2015:61). South Africa's (PoPI) Protection of Personal Information Act (Act No. 4 of 2013) is not at all useful, making no mention of big data whatsoever.³² Indeed, South African attorney Jared Nickig (2017:1) argues that “PoPI ... is somewhat lacking in the nuance and sophistication needed to tackle the type of issues that might arise in the digital world”. On a positive note, South African scholars have begun to address ethical concerns about the use of student data in the digital age. Prinsloo and Rowe (2015), for example, have called on scholars to ensure that

31 A Google Scholar search of “big data ethics” from 2010 to 2018 yielded only 438000 results.

32 https://www.saica.co.za/Portals/0/Technical/LegalAndGovernance/37544_pro25.pdf.

universities' use of student data should at all times be legal, ethical and fair (Prinsloo & Rowe 2015:59). Drawing on international trends in ethical concerns and taking the South African higher learning context into account, these scholars consider a number of best practices that researchers should adhere to in the field of learning analytics. These best practices include taking into account the benefits and unintended consequences of using big data, accepting that collecting, analysing and using student data is "a moral practice and duty" (Prinsloo & Rowe 2015:60), appreciating that "[learning] analytics should be student-centric" (Prinsloo & Rowe 2015:60), and adhering to transparent collection, analysis and utilisation of student data. Several other papers have explored the ethics of learning analytics in South Africa including those by Jordaan and Van Der Merwe (2015), Lemmens and Henn (2016), and Prinsloo and Slade (2017).³³

4.2.2 *The controversy surrounding human subjects*

Where are human subjects in big data research?

– Jacob Metcalf and Kate Crawford (2016:1)

In an attempt to try to understand why we find it difficult to align big data with ethical research, it may be helpful to consider that we are in all likelihood accustomed to thinking about ethics in terms of protecting individuals from physical harm or data discrimination, for example, and that big data tends to divorce us from these concerns: "[it creates] an abstract relationship between researchers and subjects, where work is being done at a distant remove from the communities most concerned, and where consent often amounts to an unread terms of service or a vague policy" (Metcalf & Crawford 2016:2). A major dilemma in this regard is that the disciplines that have preceded data science, such as computer science, statistics, and mathematics, have not regarded themselves as having studied human subjects (Metcalf & Crawford 2016:2).

Social scientists and humanists who make use of big data should be mindful that human subjects generate the data they are interested in (Rosenberg 2010:28), and therein lies one of the intricacies reflected in the ethics of digital research: scholars have to decide if what they are researching is person-based or text-based (McKee & Porter 2009:5), and if the former is at play, how human subjects' rights will be protected. In addition, researchers need to determine whether informed consent should be sought or whether it can be waived. In the social sciences and humanities, what constitutes a human subject is debatable and it seems that the concept of 'human subject' is not a particularly useful one, given the diverse and oftentimes confusing definition of this concept in digital research (cf. Markham &

33 Prinsloo and Slade's (2016) paper considers both South African and British case studies of learning analytics.

Buchanan 2012:6). According to the *Stanford Encyclopedia of Philosophy*, a human subject may be defined as “a living individual about whom an investigator ... conducting research obtains 1. data through intervention or interaction with the individual, or 2. identifiable private information”.³⁴ Some ethics scholars such as Lipinski (2009:58) argue that if there is no direct interference or interaction with the human subjects under investigation, then the subjects are not regarded as human and “informed consent ... is not relevant”. In terms of Lipinski’s (2009) framework, if a researcher collects data in the form of users’ comments posted below a YouTube clip in order to analyse the linguistic features of the comments, and he/she does not interact with those users, then they are not human subjects. As Rourke, Anderson, Garrison and Archer (2001:13) put it, “a researcher analyzing transcripts of a [computer] conference, without participating in the conference, has not intervened in the process and thus has not placed them in the position of research participants”. This stance is supported by other researchers who have examined ethical issues in the context of web research (e.g., Madge 2007; Rosenberg 2010; Whiteman 2010; Thelwall 2010).

4.2.3 *The public-private space conundrum*


... the online status of public and private is ambiguous and contested.

– David Berry (2004:326)

At this point, we would like to state that this does not mean that informed consent is never required when digital research is carried out. Here, it is necessary to take into account what constitutes a public or a private space, and this too is often a debatable issue as we have already observed. Since it is not possible to view the private-public space in terms of a straightforward dichotomy, a good practice is to perceive it along a continuum in order to assess any risks to the human subjects under investigation. We believe that the following scale (Table 4.1) developed by Jang and Callingham (2012) is a useful one for researchers to employ when it comes to assessing ethics variation and risk:

34 <http://plato.stanford.edu/entries/ethics-internet-research/#HumSubRes>.

Table 4.1: Measuring ethics variation in digital research (Jang & Callingham 2012:75)

| Risk level | Researcher's acts | Research field and sub-contexts | Participants' acts | Data content |
|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------|----------------------------------------------------------------------------------------------------|---------------------------------------------------|--------------------------------------------|
| LOW  HIGH | Auto-participant (being a participant) | Public (open to anyone, no registration required to access information) | Overt without identifiable information (public) | Published without sensitive topic |
| | Participant observer (being one of the participants) | Semi-public (open to most people, but registration is required for participation) | Overt with identifiable information (public) | Published with sensitive topic |
| | Open observer (participants know they are being observed) | Semi-private (registration is always required to access information, but people can sign in or up) | Covert without identifiable information (private) | Unpublished/hidden without sensitive topic |
| | Passive observer (lurking) | Private (registration is always required and only people with permission can get access) | Covert with identifiable information (private) | Unpublished/hidden with sensitive topic |

If a researcher opts to collect commenters' postings to *Mail & Guardian Online*, an online news site, then in terms of the scale in Table 4.1, the site constitutes a semi-public space for a number of reasons. First, *Mail & Guardian Online* is open to anyone, but readers who wish to discuss an article are required to first sign in with Facebook, Twitter, Google, Disqus or Thought Leader. Second, their comments can be read by anyone who visits the news site.³⁵ Clearly then, commenters' acts are overt with identifiable information. Third, comments made may be sensitive in nature if they are made in relation to a contentious topic discussed in a given online article. According to some scholars such as Convery and Cox (2012:54) and Warrell and Jacobson (2014:28), it is only research into private and semi-private domains (such as closed chat rooms or email correspondence) that requires informed consent from the participants. Discussing research by Sveningsson Elm (2009), Rosenberg (2010:33) argues that either public or semi-public spaces can be studied without obtaining informed consent.

³⁵ Another space open to readers is eNCA's online comments section.

4.2.4 *The culture of informed consent on the Internet and anonymisation*

... it is increasingly argued that consent distorts results, that consent is prohibitively burdensome for many studies; and that perhaps, in the era of “big data”, consent can be removed to facilitate major discoveries ...” – John Ioannidis (2013:40)

In addition to the challenge of trying to define private versus public spaces, big data researchers who explore aspects of social media also encounter potential problems when it comes to the culture of informed consent on social media platforms. In her study of Facebook and consent, Anja Bechmann (2014:22) questions the legality of the agreement between this site and its users, expressing doubt that it constitutes informed consent – that users actually read and understand its privacy policies, and that they are therefore aware that what they post online could be employed by third-party stakeholders. Her view is certainly borne out by a number of researchers, notably Acquisti and Gross (2006) and Govani and Pashley (2005), who have found, amongst other things, that the majority of users have never read Facebook’s privacy policies or terms of services and that they have a limited understanding of the information contained in these documents. Bechmann (2014:33) states that this behaviour points to a non-informed consent culture because “[in] a legal sense, one could argue that ‘informed consent’ does not take place”.

A user’s acceptance of Facebook’s privacy policies does not necessarily mean that most researchers are exempt from asking that user for his/her informed consent to participate in a particular study. However, this is easier said than done even in a small-scale study. This is because it is not simply a matter of asking individual users for their consent: “many secondary persons (e.g., friends of the Facebook participants, and conversation partners of Twitter profile holders) are involved” (Lomborg & Bechmann 2014:21) and so metadata is generated. It would not be practical to ask *all* users to provide their informed consent. Lomborg and Bechmann (2014:21) advise researchers to obtain consent from primary users and “[to] assess possible privacy problems on behalf of [secondary] users and only use the data as documentation to the extent that other measures of privacy protection, such as anonymization, are put to use”. When it comes to big data analysis of social media sites, it is not feasible to seek informed consent; instead, “the legal and ethical challenges ... revolve around how data is anonymized both to the researcher and when presenting results” (Lomborg & Bechmann 2014:21). Strategies to protect users include scrubbing the data by removing identifying names and protecting the archives collected and stored by means of encryption. Storage safeguards and de-identification of individuals are strategies that are in line with South Africa’s Protection of Personal Information Act 4 of 2013.

Yet, even anonymised data can be de-anonymised which we saw in the case where Michael Zimmer was able to trace Harvard sociologists’ dataset to a specific group of college

students. We question if it is at all practical to try to anonymise data in the era of big data when it is so easy to trace identities via the semantic web (cf. Dawson 2014) even when a user's direct quotes have not been published. Dawson (2014) recommends that in light of this reality, it may be best to explicitly advise research participants of the risks involved. He adds that institutional review boards generally acknowledge that it is not entirely possible to preserve the anonymity of users in online spaces.

4.2.5 *Ethics and the problem of representativeness*

Big data continues to present blind spots and problems of representativeness, precisely because it cannot account for those who participate in the social world in ways that do not register as digital signals. – Kate Crawford, Kate Miltner and Mary Gray (2014:1667)

We briefly touched on another ethical issue in the previous chapter when we discussed the fact that data gleaned from social media platforms such as Twitter may be unreliable when it comes to representativeness. Some of the concerns we raised are echoed by Kari Steen-Johansen and Bernard Enjolras (2015:127):

Twitter studies have become very popular internationally, especially due to the availability of data. However, questions may be raised as to what analyses of Twitter posts represent. An obvious challenge is that Twitter users only constitute a certain selection of the population. Other issues are linked to the fact that there is no one-to-one-relationship between user accounts and actual people. One person can have several accounts, several people can use the same account, and accounts can also be automated – so-called 'bots'.

Coupled to these problems is the phenomenon of lurkers on Twitter – users who read other people's tweets without posting any messages, a phenomenon which has also been observed on Facebook (Brandtzæg 2012). A possible solution to this complication "might be to argue that the use of [big data] makes it possible to analyze social media sites like Twitter or Facebook on an aggregate rather than an individual level, and in this way paint a picture of these social media as public spheres based on whatever topics are being discussed and distributed" (Steen-Johansen & Enjolras 2015:128).

4.2.6 *A new digital ecosystem*

The current ecosystem around Big Data creates a new kind of digital divide: the Big Data rich and the Big Data poor. – Danah boyd and Kate Crawford (2012:674)

Another critical ethical issue pertains to the fact that the computational turn has led to the creation of what boyd and Crawford (2012:674) call a big data ecosystem in which we can now distinguish between the big data haves and have-nots. In this respect, it is possible to identify two important elements that have contributed to this divide. The first relates to the reality that data has been privatised; it is ‘owned’ by companies such as Facebook, Google, and Twitter, which makes data access for academic researchers quite difficult (Steen-Johansen & Enjolras 2015:131). What is more, researchers who work for a particular company enjoy full access to all available data and are not obligated to obtain informed consent from users who would have had to accept specific terms and conditions in order to use the company’s services in the first place (Steen-Johansen & Enjolras 2015:131). The second element is an interesting one that has to do with the fact that everyone can now conduct research, given the growing disconnect between research activity and traditional research institutions (Steen-Johansen & Enjolras 2015:131). An accompanying and worrying trend is that the kind of big data research being carried out is not necessarily based on appropriate theoretical foundations, while data quality and representativeness are being thrown out the window (Steen-Johansen & Enjolras 2015:131).

What “[big data] collectors, [big data] utilizers, and [big data] generators” (Zwitter 2014:3) are currently grappling with in the big data ecosystem is datafication (van Dijck 2014). In the first chapter, we introduced readers to the notion of datafication, the process whereby social action is transformed into huge amounts of computerised data to be studied (Mayer-Schönberger & Cukier 2013). Interestingly, van Dijck observes that in the evolution of big data, datafication has become normalised: “[datafication] as a legitimate means to *access, understand* and *monitor* people’s behavior is becoming a leading principle, not just amongst the technoadepts, but also amongst scholars who see datafication as a revolutionary research opportunity to investigate human conduct” (van Dijck 2014: 198). In terms of ontology and epistemology, big data proponents claim that datafication is a neutral and thus innocuous paradigm for understanding social behaviour (van Dijck 2014:198) particularly in the context of predictive analytics, which uses techniques such as data mining, artificial intelligence, and statistics to make predictions about the future or about unknown events, thereby also generating new personal information and knowledge about individuals. This picture is insidious for a number of reasons, not least of which is the fact that this information has become “a commodity that is sold and traded among information empires and data brokers” (Mai 2016:192); big data companies have control over who may and may not access the information they generate (cf. boyd & Crawford 2012). Further, the

ordinary citizen “does not have the power ... to control the flow of information into and among information empires and data brokers” (Mai 2016:197). An additional problem may be couched in the form of a question: *Can we trust predictive analytics?* On one level, flawed methodologies in big data analysis may be harmful if they result in inaccurate predictions. On another level, what Kate Crawford and Jason Schultz (2014:94) refer to as “predictive privacy harms” may also occur:

... [In] 2012, a well-publicized *New York Times* article revealed that the retail chain Target had used data mining techniques to predict which female customers were pregnant, even if they had not yet announced it publicly. This activity resulted in the unauthorized disclosure of personal information to marketers. In essence, Target’s predictive analytics “guessed” that a customer was pregnant and disclosed her name to their marketing department, manufacturing [personally identifiable information] about her instead of collecting it directly. Although the customers likely knew that Target collected data on their individual purchases, it is doubtful that many considered the risk that Target would use data analytics to create such personal customer models to send advertising material to homes (Crawford & Schultz 2014:94).

Finally, in addition to privacy concerns, predictive analytics could lead to probability harms. Dennis Hirsch (2014:351) points out that according to Mayer-Schönberger and Cukier (2013), algorithms could forecast the likelihood of an individual suffering a heart attack, renegeing on a home loan or committing a crime, for example. In this way, the individual’s chances of obtaining an insurance policy, buying a home or securing a job could be compromised.³⁶

4.3 Data justice

...an idea of data justice – fairness in the way people are made visible, represented and treated as a result of their production of digital data – is necessary to determine ethical paths through a datafying world. – Linnet Taylor (2017:1)

Professor of data ethics Linnet Taylor (2017:1) urges scholars to support the notion of data justice—treating the participants under investigation fairly and transparently—by constructing three pillars she refers to as *visibility, engagement with technology, and non-discrimination*.

³⁶ Having said this, even in data mining the target variable is based on probability, and the data miner needs huge amounts of high-quality data to perform accurate data mining/prediction. This means that it is not that easy for an individual to be compromised.

The first pillar relates to privacy and representation, which Taylor (2017:9) argues should be accompanied by visibility and respect for informational privacy. Furthermore, this pillar calls for awareness of the risks that might occur if, through, collective profiling, group privacy is not protected (cf. Floridi 2016; Raymond 2016). The second pillar has to do with rejecting the notion of being treated as subalterns – individuals should enjoy the freedom to choose *not* to use specific technologies and to determine to what degree they would like to be visible to data markets. The final pillar which is non-discrimination is made up of “the power to identify and challenge bias in data use, and the freedom not to be discriminated against” (Taylor 2017:9).

The ethical considerations discussed here merely scrape the surface, and we will return to these considerations as well as to the idea of data justice in Chapter 6 when we assess how “data power” (Kennedy & Hill 2017a:769) affects not only the ever-increasing divide between scholars who collect data and individuals targeted by data collection, but also data visualisation, which is the focus of the next chapter. An entire chapter is devoted to data visualisations because they “may ... incorporate conscious or unconscious bias in those who have prepared them” (Fuller 2017:93).

Does big data visualisation make our endeavours less humanistic?

...visuals allow for simple and powerful communication of data, while also serving as a tool for research development. – Malu Gatto (2015: 9)

It will come as no surprise to practitioners of data science that for many humanists, visualisation in the era of big data remains a foreign element. Although visualisation is a useful tool for illustrating aggregate data in a comprehensible way,³⁷ questions that scholars in the fields of literary studies and English studies, for example, are currently asking, point to a level of mistrust of data visualisation techniques. These questions include the following:

How do literary keywords such as style, influence and genre demand rethinking in the context of quantitative analysis? Do we necessarily abandon a theoretical commitment to foregrounding the material basis of texts when we use digital methods of “distant reading” to understand textual transmission? ... how can visual tools and practices rework, rather than refigure, verbal information? (Graham 2017:449).

What may at first appear to be discouraging is that data visualisation and the humanities *are* indeed on the face of it irreconcilable for the simple reason that scholars in the humanities disciplines generally subordinate objective measurements to interpretation (Bradley, Mehta, Hancock & Collins 2016:1). Yet, many researchers in the humanities have in recent years taken up the challenge of marrying the humanities and visualisation in ways that are both innovative and productive. Before we consider their suggestions, we need to take into account the fact that if a particular visualisation is to be effective, in the sense of enabling individuals to not only comprehend it, but also interact with it, then their designers need to consider both the cognitive and perceptual processes and limitations of the human mind (Olshannikova, Ometov, Koucheryavy & Olsson 2015:9). On a cognitive level, poorly designed visual displays may lead, amongst other things, to ambiguous interpretation of data (Burkhard & Eppler 2005), cryptic encoding (Tufte 1986), the obscuring of important insights (Few 2006; Kosslyn 2006), over-complication in how information is represented

37 It should be noted that data visualisations may be visual and (occasionally) auditory.

(Few 2006), and the absence of adherence to Gestalt principles (Tufte 1986). These and many other consequences of poor design are discussed in the section below to highlight what Kennedy and Hill (2017a:769) have referred to as “the pleasure and pain” of visualisation.

5.1 Data visualisation, human cognition, and human perception

Visualisation is not unique to the computer science domain.

– Maureen Stone (2009:44)

With data ever-increasing in quantity and becoming integrated into our daily lives, having effective visualizations is necessary. – Michelle Borkin (2014:iii)

There is no question that visual representations of highly intricate information and knowledge demand higher-level cognitive functions which include, but are not limited to, recall, reasoning, understanding, and insight (Patterson, Blaha, Grinstein, Liggett, Kaveney, Sheldon, Havig & Moore 2014:42). Patterson *et al.* (2014:44) theorise that these functions are enabled through top-down cognitive processes, a theoretical perspective which challenges the traditional notion that the connection between perception and cognition is relatively simple, involving “bottom-up engagement of low-level, feature-detection processes that sequentially feed into higher-order cognition” (Patterson *et al.* 2014:44). While these researchers foreground top-down processing in their human cognition framework, they nevertheless argue that bottom-up and top-down processes interact in a dynamic way: the former type of processing essentially guides the way in which the latter type is processed, resulting in the activation of so-called “organized knowledge structures in long-term memory” (Patterson *et al.* 2014:44). This interplay has been observed by other researchers, notably, Grill-Spector and Kanwisher (2005), Oliva and Torralba (2006), and Biederman (1981), who have carried out tests to determine how fast and/or accurately individuals are able to detect meaning from visual stimulation.

Building on dual-process theory, Patterson *et al.* (2014:44) go on to propose that human cognition involves interaction between dual systems – *an analytical or reasoning system* and *an autonomous or intuitive system*. The analytical system allows for pattern recognition, analogical reasoning as well as deductive reasoning, while the intuitive system by contrast is responsible only for pattern recognition. Both systems rely on top-down and bottom-down processing as well as *encoding* during which a new visual image is converted into a neural representation in a person’s short-term memory (Patterson *et al.* 2014:45). What is important for scholars designing big data visualisations to consider is the fact that it is only information extracted during encoding that can be utilised for any processes that follow. *Attention*, a process during which the brain filters and selects specific information,

and working *memory*, the ability to retain that information, interact during encoding. For visualisations to be effective, they need to be designed in such a way that there are no attention distractions present, particularly when the individual viewing the visualisation is performing another task at the same time (Patterson *et al.* 2014:45). One of the consequences of attention interference is that memory retrieval may be impeded (Dudukovic, Dubrow & Wagner 2009:953), which is problematic given that long-term memory, which interacts with working memory and pattern recognition, is ultimately required to store neural representations of information (Patterson *et al.* 2014:46). Attention interference may also impact negatively on the decisions an individual makes about a data visualisation (cf. Craik 2014:4).

Given these complex cognitive processes, Patterson *et al.* (2014:47) recommend that data visualisations be designed with specific leverage points or strategies in mind, the first of which is intended to achieve bottom-up, stimulus-driven attention: a visualisation should reflect “salient cues to drive exogenous attention, alerting users to changes in or important attributes of [the] visualisation”. What this essentially means is that the visualisation should be designed so that it draws an individual’s attention to a stimulus; if this does not happen then the result is “inattentional blindness” (Simons 2000:147) – the individual’s failure to consciously register specific attributes of the visualisation because his/her attention has been directed elsewhere. To overcome inattentional blindness, Patterson *et al.* (2014:48) suggest that the researcher incorporate striking design elements or cues into a visualisation that the individual cannot easily ignore. One such cue is colour, and a review of the literature provides several interesting findings about why and how specific colours should be incorporated into data visualisations (cf. Lin, Fortuno, Kulkarni, Stone & Heer, 2013). Silva, Santos and Madeira (2011:326) argue that colour carries specific cultural connotations and that being mindful of this fact helps reduce a viewer’s cognitive load. For instance (and stating the rather obvious), when it comes to the visualisation of average maximum temperatures in Cape Town over six months, viewers will be drawn to Figure 5.1 below because South Africans associate shades of red and orange with high temperatures and blue shades with low temperatures. The choice of these colours helps focus the viewer’s attention (cf. Stone 2009:48). Figure 5.2 illustrates the same data, but this time, the colours are reflected in the adjectives ‘hot’, ‘warm’, ‘mild’, ‘cool’, and ‘cold’. However, these colour choices are semantically incongruent when they should be semantically resonant; different shades of blue, for example, would not be evocative of hot or warm temperatures for most people and would therefore cause strong attention distraction (Lin *et al.* 2013:2). In cognitive science, contrasting colour cues result in interference, and this is referred to as the Stroop effect (see Figure 5.3), named after J. Ridley Stroop (1935) who discovered that processing one stimulus feature will hamper the simultaneous processing of another stimulus feature (cf. MacLeod 1991:163).

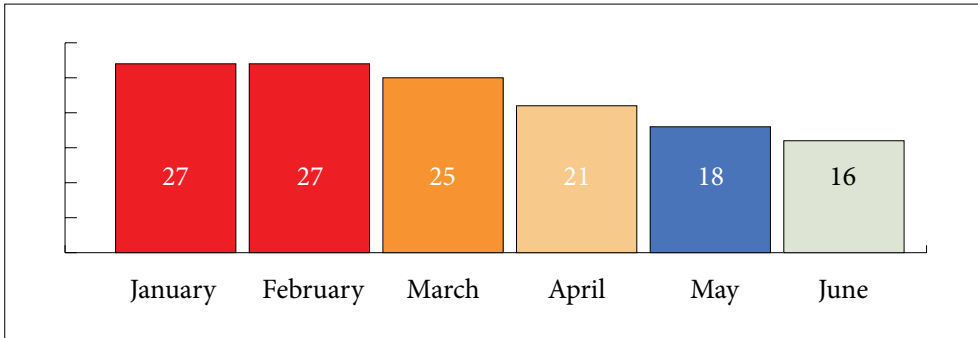


Figure 5.1: Cape Town’s average maximum temperatures in degrees Celsius

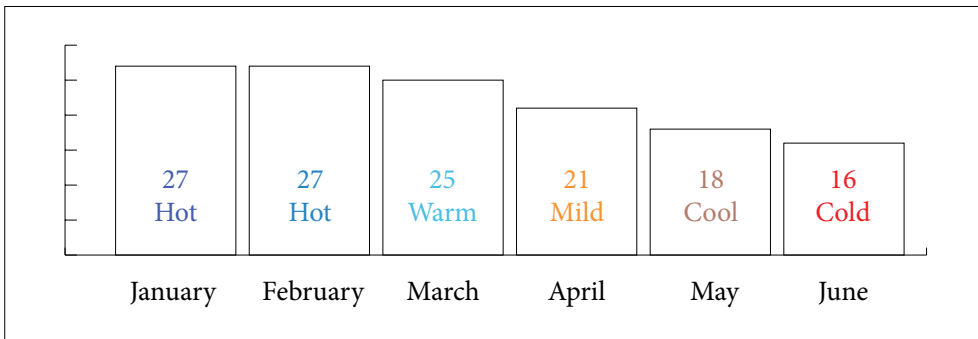


Figure 5.2: Semantically incongruent colour choices to represent temperatures



Figure 5.3: The Stroop effect

The second leverage point identified by Patterson *et al.* (2014:47) is geared towards achieving top-down, voluntary attention, and so data visualisation needs to reflect “appropriate organization of material or interaction options to assist endogenous attention and minimize distracting information” (Patterson *et al.* 2014:48). To focus attention as well as reduce any distractions so that information can be encoded and then stored in working memory, Patterson *et al.* (2014:48) recommend the use of what they call “endogenous attentional resources” such as clear labels or arrows that guide the viewer towards the

relevant information as well as the elimination of extraneous information. As far as the latter is concerned, a number of studies have explored how visual clutter impacts a viewer's cognitive processing (Barvo & Farid 2004; Doolittle, McNeill, Terry & Sheer 2005; Ellis & Dix 2007; Rosenholtz, Li & Nakano 2007): "if too much visual information is presented ... then the visual channel's capacity will be exceeded, leading to insufficient processing of that visual information" (Doolittle *et al.* 2005: 198). Techniques for reducing clutter include limiting data (Sula 2012), removing uninformative or redundant detail, and experimenting with different layouts which also affect cognitive processing (Hornof 2004).

The third strategy to bear in mind to facilitate cognitive processing of a visualisation has to do with chunking, which entails "[choosing] visualisation parameters that provide strong grouping cues ... which will minimize the effects of working-memory capacity limitations" (Patterson *et al.* 2014:48). Well-known chunking techniques are the use of common image parameters in the form of specific colours or shapes, for example (Patterson *et al.* 2014:48), as well as adherence to the principles of Gestalt theory (Wertheimer 1938), which proposes that as human beings, we tend to organise visual information into groups or patterns in an attempt to comprehend the picture as a whole (Quinn & Bhatt 2015:691). More digestible visual representations that follow Gestalt principles are those that exploit similarity, proximity, and enclosure.³⁸ Similarity is a principle according to which viewers of a visualisation will group visual attributes that they perceive as possessing similar characteristics related to colour, texture, and size, to name a few (cf. Kobourov, Mchedlidze & Vonessen 2015). Thus, Figure 5.4 will automatically be perceived as having two elements – green and blue rectangles. The Gestalt law of proximity states that objects will be grouped together if they have been arranged close to one another (Figure 5.5), and this applies even if the objects reflect diverse characteristics (although Guberman (2015:28) points out that the objects still need to be similar in some sense). Enclosure/ closure states that we are predisposed "to close up objects that are not complete" (Olshannikova *et al.* 2015:17). In this respect, the World Wide Fund for Nature's iconic panda symbol is a good example. Data visualisations may also utilise the principle of closure by surrounding specific groups with visual elements. In Figure 5.6, viewers can easily identify two distinct groups based on the fact that the designer has enclosed related elements in visually dissimilar boxes.

38 Other Gestalt principles include continuity/continuation (where the visualisation enables the eye to be drawn from one object to another) and connectedness (where there are clear connections between various objects as typically seen in genealogical charts).

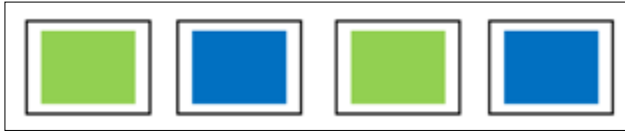


Figure 5.4: Gestalt principle of similarity



Figure 5.5: Gestalt principle of proximity

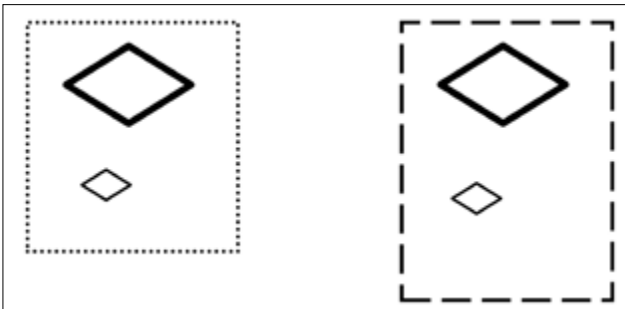


Figure 5.6: Gestalt principle of enclosure

Organising information based on mental models (structural analogies of the world) is another tactic that may be useful to consider in visualisations because this kind of information supposedly activates “strong retrieval cues for knowledge structures in long-term memory to aid reasoning” (Patterson *et al.* 2014:49). In the literature, theories that attempt to explain how human beings reason abound, and include the theory of the meaning of conditionals (Johnson-Laird & Byrne 2002), or what Patterson *et al.* (2014:49) refer to as a mental model of “imagined possibilities” according to which inferences are generated; the probabilistic approach to human reasoning (Oaksford, Chater & Larkin 2000); and mental logic theory (O’Brien 2009). Although these theories are controversial,³⁹ many scholars concede that mental models play a significant role in reasoning (Johnson-Laird & Byrne 2002; Johnson-Laird 2010; Johnson-Laird & Khemlani 2015):

39 For a critique of the mental model theory proposed by Johnson-Laird and Byrne (2002), see Evans, Over and Handley (2005), and for a critique of Oaksford, Chater and Larkin’s (2000) conditional probability model, refer to Schroyens and Schaeken (2003). López-Astorga (2016) provides a sound critical appraisal of the mental logic theory.

Human reasoning is not simple, neat, and impeccable. It is not akin to a proof in logic. Instead, it draws no clear distinction between deduction, induction, and abduction, because it tends to exploit what we know. Reasoning is more a simulation of the world fleshed out with all our relevant knowledge than a formal manipulation of the logical skeletons of sentences. We build mental models, which represent distinct possibilities, or that unfold in time in a kinematic sequence, and we base our conclusions on them (Johnson-Laird 2010:18249).

Several researchers have exploited what we know about mental models and reasoning to design more comprehensible visualisations. We have already touched on Lin *et al.* (2013), who have experimented with semantically resonant colours they believe will evoke specific associations in the minds of individuals because they are grounded in well-known linguistic and cultural conventions. In an interesting study, another team of researchers used photographs, clipart pictures, and icons as visual embellishments, and found that they facilitated both memorisation and concept comprehension (Borgo, Abdul-Rahman, Mohamed, Grant, Reppa, Floridi & Chen 2012).

The fifth leverage point recommended by Patterson *et al.* (2014:51) has to do with creating visual displays that reflect analogous patterns related to viewers' mental models with a view to facilitating analogical reasoning. Analogical reasoning involves using the attributes of a source domain that is well understood to facilitate the understanding of a target domain that is not so well understood (Figure 5.7). Not to be confused with metaphorical graphics or visualisations,⁴⁰ Risch (2008:4) argues that analogical visualisations reflect specific characteristics, the most important being “[to] express systematic relations among the elements of a target domain in terms of those of some source domain”. In addition, the elements of the source and target domains must be aligned (Risch 2008:4). Thus, for example, geographic maps and pictures are graphical analogues: the internal structure of these visualisations very closely resembles that of the phenomena they represent, and therefore allows viewers to bridge the gap between the known and the unknown.

40 In metaphorical graphics, abstract concepts are expressed in such a way that the target domain is semantically distant from the source domain (Risch 2008:4).

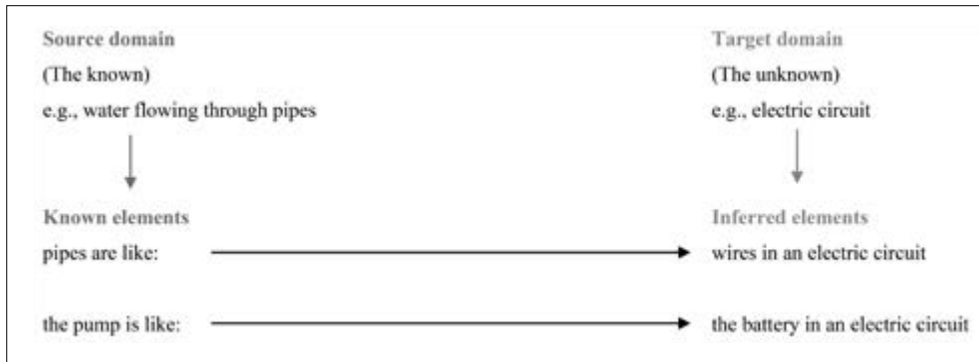


Figure 5.7: Example of analogical reasoning (adapted from Daugherty & Mentzer 2008:10)

The five recommendations just outlined are useful for scholars whether their research is informed by quantitative or qualitative methods, but humanists require even clearer direction since, as noted at the beginning of this chapter, the humanities and data visualisation appear to be ill-matched. Let us return now to what the literature tells us about the ways in which humanists may exploit data visualisations to provide insights into their qualitative findings.

5.2 Technology and the humanities

If we believe that technology determines our choices and that we are therefore not responsible but act only in accordance with a technological imperative, we let technology decide for us in the sense that we renounce responsibility for our actions.

– Bjørn Hofmann (2006:2)

What adds to the divide between humanists and big data scholars is not only the mistaken assumption that data visualisation is the sole domain of quantitative research, but also the fact that the visual display of qualitative data does not have a particularly long tradition. In the last decade, researchers who have explored how to visually express qualitative information include Slone, (2009), Verdinelli and Scagnoli (2013), Henderson and Siegal (2013), and Chandler, Anstey and Ross (2015). What humanists may find frustrating about these studies is that although they provide practical advice on displaying qualitative data through flow charts (Draucker & Martsof 2008), ladders (Eriksson, Starrin & Janson 2008), matrices (LeGreco & Tracy 2009), networks (Cheyney 2009), and the like, some researchers have begun challenging the emphasis placed on how data should be displayed when “it must be understood that data visualization for the humanities needs to be built more for experience than demonstration of fact” (Bradley *et al.* 2016:3). In Drucker’s (2014:125) view, graphical

tools introduced to the humanities and then embraced by scholars “are a kind of intellectual Trojan horse, a vehicle through which assumptions about what constitutes information swarm with potent force”.

For Bradley *et al.* (2016:2), a most useful starting point for humanists is to consider Martin Heidegger’s (1966, 1977) conception of technology since it provides an analogy for understanding the interaction between texts and technology.⁴¹ Looking through a phenomenological lens, Heidegger argues that technology should not be reduced to its instrumental value – that its value depends very much on its contexts of use⁴² (cf. Ihde 2010:152). In the humanities, computing’s instrumental purpose lies in its ability to process data as quickly as possible, *but* this does not constitute interpretation: “what is necessary is an understanding that the expressed purpose of technology itself (its instrumental existence) and our engagement with it (its anthropological potential) are two separate parts of a technological whole” (Bradley *et al.* 2016:2).

In terms of this understanding, Fish (2012:1) contends that one of the flaws inherent in many digital humanities projects is the tendency to “first run the numbers, and then ... see if they prompt an interpretative hypothesis. The method, if it can be called that, is dictated by the capability of the tool” (cf. Bradley *et al.* 2016:2). To overcome this deficiency, Bradley *et al.* (2016:3) recommend an approach according to which humanists adopt what they call “slow analytics”, which takes into account both the time and space required to cognitively process information.

To test this approach, these researchers recruited PhD students and academics who teach and/or publish on literary criticism or poetics. These participants were asked to analyse two poems using Livescribe Anoto (digital) pens⁴³ and paper so that the researchers could determine the processes involved in each participant’s analysis. Annotation of the poems (which were selected on the basis of being unknown to each participant to avoid expertise bias) was voluntary, and if a participant opted to annotate a poem, they were also asked to explain how their annotations functioned during a self-reflective discussion with the

41 We acknowledge that Heidegger’s views on technology are not without their idiosyncrasies, and that they are sometimes obscure to say the least. In this respect, Godzinski (2005) provides an insightful overview of Heidegger’s controversial philosophy of technology.

42 Of course, we concede that this argument should be moderated since “[t]echnologies reveal different things for different people” (Loukissas 2012:22). In addition, Heidegger himself did not view technology as something to be avoided, a view exemplified in his ‘Memorial address’: “For all of us, the arrangements, devices, and machinery of technology are to a greater or lesser extent indispensable. It would be foolish to attack technology blindly” (Heidegger 1966:53).

43 Digital pens allow the user to record what they write, and all data generated can be transmitted to a digital device through wireless technology.

researchers.⁴⁴ Not unexpectedly, what the researchers found was that literary analysis is a painstakingly slow, methodical, and iterative process, in which reflection is essential, helping generate meaning-making and insights about the texts under investigation (cf. Srivastava & Hopwood 2009:76). In a subsequent phase, the researchers designed what they call a “metatation” interface or system, generating visualisations of participants’ analyses only *after* their sense-making phase. Thus, for instance, if one participant underlined specific words in a poem to make sense of synonyms and antonyms, the metatation system augmented the analysis by adding additional synonyms and antonyms, in this way drawing the participant’s attention to words he/she had missed. This prototype interface allows scholars to interact with texts in their own time and space, and because it is introduced only after the analyses have been carried out, it does not impede the thinking and sense-making process in any way, thus respecting the performative nature of interpretation in the humanities.

5.3 Data as capta

Capta is not data as we typically understand data. Capta represents what is seen, thought and felt. – Bryan Beverly (2017:2)

In order to reconcile the humanities and visualisation, the latter needs to be re-conceptualised to preserve the integrity of humanistic data, and here, Drucker’s (2011:2) “polemic”, as she refers to it, involves strongly urging humanists to re-conceive of data as capta. This call has been made by many other scholars over the past few years (Clement 2012; Kitchin 2014; Maier & Deluliis 2015; Enslin 2016; Furner 2016), one of the reasons being that the etymology of the term data in the context of the humanities is fairly problematic. ‘Data’, originally derived from the Latin word *datum*, literally means ‘something given’, which erroneously creates the impression that data is always a given (Owens 2011). In an earlier chapter, we mentioned the fact that the notion of (big) datasets as constituting raw information is rather troublesome, given that data does not just come into existence: capturing data and then interpreting it is based on specific choices on the part of the researcher that encompass judgement and discernment, amongst other things. Yet, in an increasingly datafied society, the realist view of knowledge continues to persist, presenting various phenomena as existing outside of the observer when, as correctly pointed out by Drucker (2011:1), “[r]endering *observation* (the act of creating a statistical, empirical, or subjective account or image) as if it were *the same as the phenomena observed* collapses the critical distance between the phenomenal world and its interpretation, undoing the basis of interpretation on which humanistic knowledge production is based” (cf. Ambrosio 2015:137; Kennedy, Hill, Aiello & Allen 2016:719).

44 In the humanities, an annotation constitutes information employed to classify, code or comment on the data sources collected (Evers 2018:64).

To generate a more nuanced understanding of data visualisation in the humanities, *capta* (literally meaning ‘taken’ in Latin) appears to be a far better term, since it captures the constructivist notion that knowledge is “taken” in the sense that it is partial, situated, and constitutive (Drucker 2011:2).⁴⁵ Levi (2013:34) expresses it succinctly when she says that “[h]umanistic data are as un-data as they can get” ... [because] [h]umanities data are not generated by instruments, but by people in the process of going about their everyday life”. Drucker (2011:2) proposes a number of fundamental principles to be followed when *capta* is constituted and displayed, going so far as to argue that if these principles are ignored, what will ultimately be compromised is the very authority of the knowledge generated by humanists: “[t]he digital humanities can no longer afford to take its tools and methods from disciplines whose fundamental assumptions are at odds with humanistic method”.⁴⁶

Drucker (2011:13) refers to John Snow’s well-known map of cholera outbreaks (discussed in Chapter 1) as an excellent example of how visuals may be (re-)designed to reflect co-dependent constructivism rather than observer-independent realism. Snow’s map as depicted in Figure 5.8 quite effectively illustrates the role of the pump in the number of people who died of cholera in London. In Drucker’s (2011:14) view, therefore, the map served its purpose. However, she goes on to wonder *who* the dots are, arguing that the addition of demographic features such as an individual’s age, health, and family role could have provided a more nuanced, complex statistical view of the epidemic. Drucker (2011:14) does not stop there, suggesting that the rate of deaths and their frequency could be integrated into the map on a temporal axis to reflect “increasing panic”. This is a fairly novel notion, namely, that “[t]he display of information about... affective experience can easily use standard metrics” (Drucker 2011:7). The terrain could also be re-drawn from the perspective of, for instance, an individual who has lost a loved one, not only to illustrate the urban streetscape of nineteenth century London, but also to highlight what Drucker (2011:14) describes as “[features] of the graphical representation of humanistic interpretation”.⁴⁷ The reinvisaged graphic is illustrated in Figure 5.9.

45 This does not mean that humanists and scientists are engaged in some kind of intellectual battle. Drucker (2011:2) concedes that her data-*capta* distinction “is not a covert suggestion that [...] only the humanists have the insight that intellectual disciplines create the objects of their inquiry. Any self-conscious historian of science or clinical researcher in the natural or social sciences insists the same is true for their work”.

46 We contend that the humanities and not only the digital humanities falls under this admonition.

47 Drucker (2011:7) recognises that what she is suggesting are “subjective methods”, and risks the observation that “[r]ecognizing that such subjective methods are anathema to the empirically minded makes me even more convinced that they are essential for the generation of graphical displays of interpretative and interpreted information”.

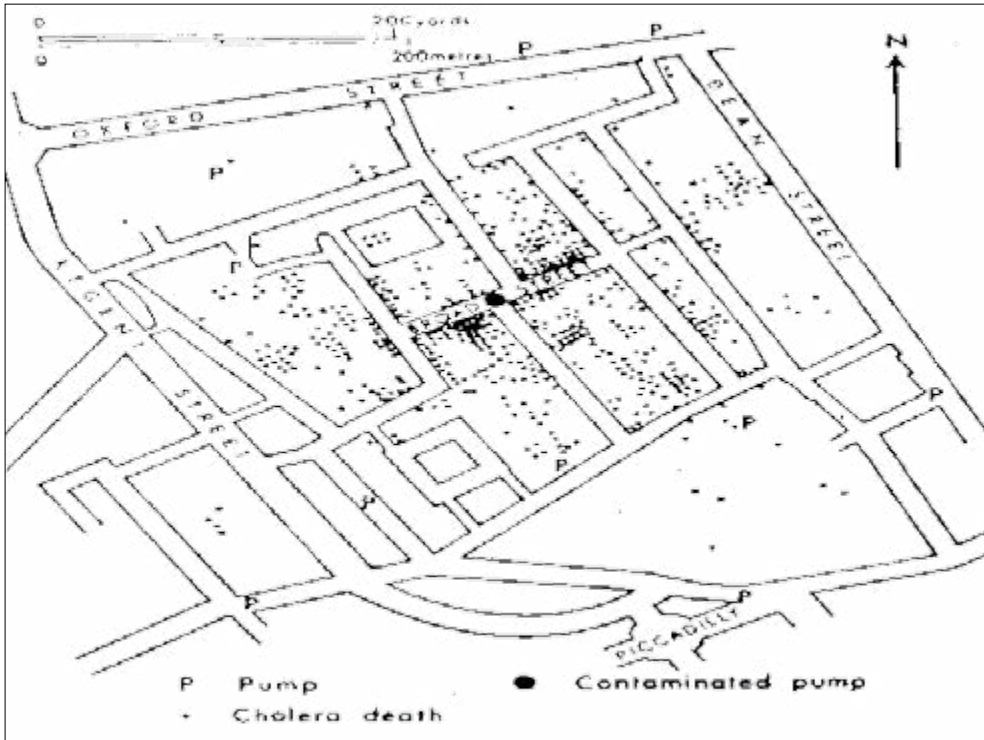


Figure 5.8: John Snow's map of cholera outbreaks in London (drawn by Snow circa 1854, and taken from Stamp's (1964) *The geography of life and death*)

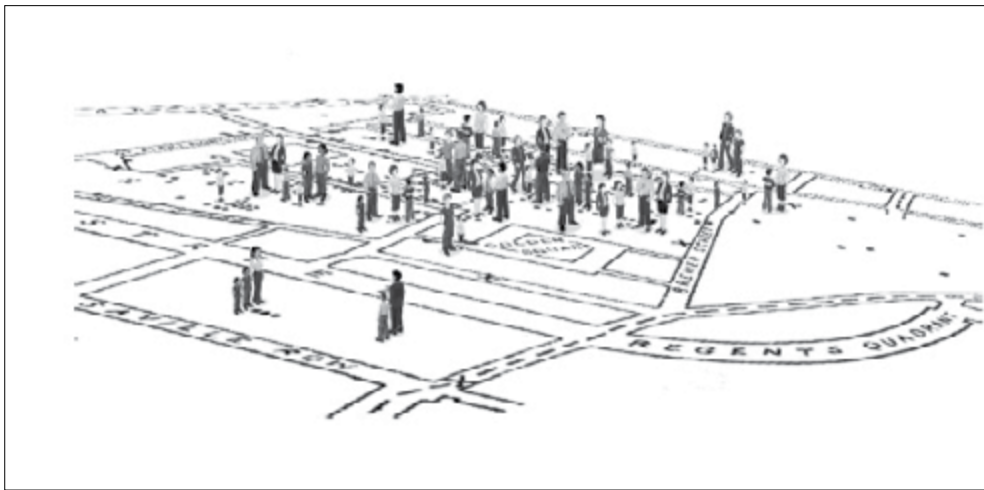


Figure 5.9: John Snow's map reinvisaged (Drucker 2011:19, with credit to Xárene Eskandar for the graphic)

An important tenet implicit in Drucker's (2011) description, and one that highlights the distinction between data and *capta*, is that graphical displays in the humanities should always follow *a humanistic approach* which entails infusing the displays with affect through expressive graphics and metrics. This tenet is exemplified in the suggestions that John Snow's map would be even more striking if it (1) reflected rates of death to show "increasing panic" (Drucker 2011:14) and (2) included features that reminded the viewer of what the streets of London looked like in the 1850s. What runs like a golden thread through Drucker's (2011) suggestions is respect for the interpretative nature of knowledge represented in *qualitative displays*, which is another tenet she upholds.

Although they are still in the minority, some scholars working with large volumes of data have also called for more attention to be paid to the emotional dimensions of graphical representations. Kennedy and Hill (2017b) recently conducted an empirical study with a view to illustrating how individuals engage on an emotional level with data and their visualisation. The conclusion reached was that visualisations should inform individuals' minds and hearts – that "it is not only numbers but also *the feeling of numbers* that is important" (Kennedy & Hill 2017b:1). This argument is advanced by other researchers such as Norman (2004), Grosser (2014), and Kirk (2016), since emotions not only play a key role in our social and cultural experiences (cf. Kennedy & Hill 2017b), but also support reason and rational thinking (cf. Stratton 2012). Kennedy and Hill (2017b:3) speculate that the neglect of emotional dimensions could, at least in part, be due to the history of visualisation which may be traced back to the Age of Enlightenment which was preoccupied with reason rather than subjectivity, tradition or passion.⁴⁸

5.4 The emotional and social pitfalls of visualisation

Data are not just numeric – they are both statistical and visual. In part because of this entanglement, data stir up emotions. – Helen Kennedy and Rosemary Hill (2017b:2)

Up to this point, we have discussed the fact that data visualisations may have cognitive drawbacks if they are not designed with a human cognition framework in mind. Additional caveats to bear in mind, particularly in light of Drucker's (2011) call for visual displays imbued with affect, pertain to what researchers may not be aware of, namely, the emotional and social risks of visualisation, risks which are often overlooked or dismissed in favour of the cognitive effects of visualisation (cf. Roos, Bart & Statler 2004:551). As far as the emotional dimension is concerned, and based on a multidisciplinary literature review of data visualisations, Bresciani and Eppler (2015:4) point out that visualisations may unwittingly

48 This is certainly an over-simplification; van Holthoorn (2017:8) quotes Hume, who claimed in *A treatise of human nature* (1739) that "[r]eason is and ought only to be the slave of passions".

cause viewers to experience negative or inappropriate feelings. In terms of emotion and encoding, their study⁴⁹ has revealed that viewers may find visualisations disturbing (Tufte 1990; Cawthon & Moere 2007) uninteresting (Cawthon & Moere 2007) or unattractive (Cawthon & Moere 2007; cf. Krum 2013). As far as decoding is concerned, some visualisations (such as those that flicker or are striped) have been found to cause visual stress to the point of causing viewers to feel ill (Ware 2007). Viewers may also favour some visualisations rather than others based on their personal likes and dislikes of certain visual elements (Tversky 2005). Finally, viewers may experience negative emotions when decoding an image if they have encountered this image before and experienced it in a negative way (Chen 2005; Avgerinou & Pettersson 2011).

When it comes to the social effects of encoding, the role of hierarchy and the exercise of power involved in the process of designing a visualisation cannot be dismissed. In this regard, visual representations may be manipulated in such a way that viewers have access to some information, but are barred from possessing knowledge about certain aspects of that information based on the designer's choice to (un)intentionally include or exclude information and knowledge, a risk outlined in a study by Ewenstein and Whyte (2007). What we should never forget is that "data visualisations are not neutral windows onto data: they privilege certain viewpoints, perpetuate existing power relations and create new ones and, as such, they do ideological work" (Kennedy & Hill 2017a:773). Another area of concern pertains to data visualisations that are designed in such a way that a specific point of view is emphasised to such a degree that there is no room for individuals to generate alternative views or invent other options (Whyte, Ewenstein, Hales & Tidd 2007). Bresciani and Eppler (2015) have also reviewed the literature to identify problems in the area of decoding. In the context of the use of visualisations in group interaction, some researchers have recorded altered behaviour on the part of members of the group. For instance, Eppler and Platts (2009) have observed that in cases where a graphical representation is generated in a group, the opinions of the members of that group may be suppressed by an individual who is regarded as dominant. Another (perhaps unintended) consequence of badly designed visuals is that they may be misinterpreted in different cultural contexts, given that symbols and colours are not universal in their meanings, a consequence that is well documented in the literature (Nisbett 2003; Ewenstein & Whyte 2007; Avgerinou & Pettersson 2011; Bresciani 2014; Forsythe 2014; Jahns 2014). The recency effect is an additional drawback of information visualisation observed by Tufte (1986, 2006) and Nisbett (2003), amongst others. In terms of the recency effect, a viewer's interpretation of a graphical representation may be coloured by a recent experience or event.

49 We have added additional studies to Bresciani and Eppler's (2015) list.

5.5 Visualisation and the problem of data power

Is big data analytics good or evil? – Bill Franks (2015:1)

If humanists thought that designing data visualisations was fraught with myriad challenges such as those just highlighted, then they have not yet considered the problems inherent in data power which can be exploited for good or ill: data displays present arguments and explanations put forward as objective facts, thus shaping the ways in which we perceive the world we live in (Boehnert 2016:1; cf. Kitchin 2014; Williamson 2017). Furthermore, and advancing critical approaches to visualisation, Boehnert (2016:18) observes that scholars are increasingly compelled to work within what she calls “the ideological scaffolding of the neoliberal political project”, which is becoming all too familiar to South African humanists and social scientists (Clare & Sivil 2014; Le Grange 2016). One of the negative consequences of this project is that it subjects human endeavours to “market principles” (Brown 2011:118), forcing academics to generate reams of research for the sake of the knowledge economy.⁵⁰ Under pressure to publish as much and as frequently as possible in a datafied world, academics are bowing to “the hegemony of metric power” (Feldman & Sandoval 2018:219),⁵¹ sacrificing the interpretative element of research typically undertaken by humanists and social scientists to data visualisation: “the desire for data visualisations can be understood as motivated by the need to operate within a market, as visualisations are seen as a means to ‘sell’ the research capabilities of university departments, as they market themselves to external organisations” (Kennedy & Hill 2017a:776). We re-iterate here that while data visualisations are an important element of the research process, they must not become mere “graphical primitives” (Manovich 2011:47) of the artefacts under investigation.

In the next chapter, we take a closer look at the notion of data power in an era of big data, but this time in the context of studies carried out in the humanities and social sciences. Both disciplines face threats, “not least because ‘ethnography’ is often presented as the Other to big data” (Boellstorff 2013:2). Yet there are also many opportunities for humanists and social science scholars to showcase the kinds of contributions their disciplines can make to the big data phenomenon.

50 Diane Powers Dirette (2016:2) expresses dismay when she observes that “[i]f statistical significance is found [in an article], the research article, regardless of the size of the data, is more likely to be published”.

51 Kennedy and Hill (2017b:6) define metric power as “the growing prevalence of numbers, data and measurement in contemporary forms of government and control” (cf. Beer 2016b).

Data power in the era of big data: friend or foe?

Datafication ... harbors both threats and opportunities for civic engagement.

– Gutiérrez and Milan (2017:95)

In her insightful book entitled *Weapons of math destruction*, mathematician and data scientist Cathy O’Neil (2016) offers a number of sobering thoughts on the dangers of big data power in the absence of social justice. Two choice comments she makes are that “Big Data has plenty of evangelists, but I’m not one of them” and “[while] Big Data, when managed wisely, can provide important insights, many of them will be disruptive”. With respect to the latter comment, O’Neil (2016) laments both the fact that data scientists tend to present the problems reflected in ecosystems without also providing solutions to them, while many big data researchers employ algorithms they do not share with the public; “we see only the results of the experiments researchers choose to publish” (O’Neil 2016:148). We cannot ignore the fact that big data has a dark side too.

6.1 Big data’s shadow side

We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead. – Cathy O’Neil (2017:256)

In a paper published more than a decade ago, Ivor Baatjies (2005:30) painted a somewhat stark picture of scholarship undertaken at institutions of higher learning in South Africa when he wrote that “[c]orporate mentality and ... neoliberal fatalism have no regard for any form of research in favour of social justice, oppression and exploitation”. It appears that the emergence of big data may only have made the situation worse. Scholars refer to the many instances of big data being used as a tool to threaten democracy and increase inequality (O’Neil 2017). Gangadharan (2012), for instance, warns how large-scale commercial data profiling in terms of race and ethnicity has excluded vulnerable individuals from receiving economic, social or political benefits in North America. Indeed, Mayer-Schönberger and Cukier (2013) argue that we may be heading for big-data authoritarianism in which Big Brother will create even greater asymmetry of power between the privileged and the

oppressed.⁵² The idea that Big Brother is watching our movements and tracking our personal data has also been raised by the South African media. In 2017, investigative journalist Heidi Swart⁵³ wrote the following in *Daily Maverick*:

South Africa's intelligence community outsources the analysis of social media platforms to the private security sector. The social media posts that people choose to make public ... can be used by intelligence services to accurately assess the sentiments, thoughts, movements and plans of people, groups, or institutions. Such analysis, called data mining, produces SOCMINT – social media intelligence.

In addition, the South African government currently uses geofencing technology⁵⁴ utilised by a software programme referred to as Media Sonar to monitor citizens' social media activity, the aim being to identify potential threats to the country's security (Swart 2018).

South Africa experienced its biggest data leak in 2017 when an Australian information security expert, Troy Hunt, exposed the fact that 60 million South Africans' personal data had been leaked. The data breach appears to have emanated from Jigsaw Holdings, a holding company for real estate franchises. According to Skolmen and Gerber (2015:4), South African organisations still have little understanding of the basic conditions of protecting personally identifiable information in terms of the country's Protection of Private Information (POPI) Act signed into law in 2013 (Burke & van Heerden 2017:85).

Other examples of the abuse of big data power include the now infamous Snowden disclosures about North America's National Security Agency and its surveillance practices (van Dijck 2014), the establishment of "spurious correlations"⁵⁵ (Calude & Longo 2017) in large datasets, the manipulation of consumers' buying choices through robotic "nudging"⁵⁶

52 This is of course not a new notion: as Gangadharan (2012:1) aptly puts it, "old forms of prejudice and injustice can be grafted onto these new tools".

53 It is worth noting that Heidi Swart's article was commissioned by the University of South Africa's Department of Communication and the University of Johannesburg's Department of Journalism (through its Media Policy and Democracy Project).

54 Geofencing allows users to cordon off specific geographical locations using GPS technology (cf. Luxhoj 2016).

55 Calude and Longo (2015:13) define a correlation as spurious "if it appears in a randomly generated database".

56 In *Nudge: Improving decisions about health, wealth, and happiness*, Thaler and Sunstein (2008) assert that people can be "nudged" into make better choices about various aspects of their lives. While some scholars have lauded nudging theory (Cohen 2013; Saghai 2013), others maintain that it is not as innocuous as it seems and that it is riddled with ethical dilemmas (Selinger & Whyte 2011; Leggett 2014; Borenstein & Arkin 2016).

(Helbing 2015), and the engineering of public opinion during election campaigns (Tufekci 2014; Bessi & Ferrara 2016).

In all probability the most ominous big data project we have recently encountered is China's so-called "social credit" programme, which the Communist Party claims will be fully operational by 2020 to monitor its 1.4 billion citizens. Referring to the creation of a digital dictatorship, ABC journalist Matthew Carney (2018) reports that China has already launched a pilot programme in which each citizen has been assigned a default social credit score of 800 points which then changes depending on his/her behaviour which is tracked via the country's 200 million state-of-the-art CCTV cameras designed with facial recognition, body scanning, and geo-tracking capabilities.⁵⁷ In addition, the Chinese government intends collecting data about its citizens' behaviour from their smartphones as well as from their medical, financial, and purchasing records. In trial areas, this Orwellian project has already resulted in the punishment of approximately 10 million people who have lost points owing to "anti-government" behaviour such as bad driving, smoking in non-smoking areas, and defamation of the ruling party. Punishment is meted out in the form of fines and through lack of access to job promotions, travel visas, and the like. Those with high scores, on the other hand, are rewarded by way of incentives which include faster and easier access to housing, jobs or job promotions, and good schools. In an insightful article entitled 'Engineering the public', Zeynep Tufekci (2014:12) observes that "[s]tarting an empirically informed, critical discussion of data politics now may be the first important step in asserting our agency with respect to big data that is generated *by* us and *about* us, but is increasingly being used *at* us". (We consider Tufekci's (2014) call in Chapter 9 when we discuss the need for scholars to approach big data science through a critical data studies lens that challenges stealthy practices such as dataveillance and big data privacy breaches by interrogating the social, political, ethical and economic implications of big data projects.)

Academics too are beginning to harness the power of big data in ways that do not necessarily put the well-being of individuals first (Lewis *et al.* 2008; Kramer *et al.* 2014; Kirkegaard & Bjerrekær 2016), an insidious practice we noted in Chapter 4. What appears to be exacerbating this trend is the pressure that some journals are placing on qualitative researchers to limit their discussion of empirical data and abbreviate any explanation of their data collection methods owing to these journals' space constraints (cf. Chandler *et al.* 2015:1; Messner, Moll & Strömsten 2017:440).⁵⁸

57 <https://www.abc.net.au/news/2018-09-18/china-social-credit-a-model-citizen-in-a-digital-dictatorship/10200278>

58 These space limitations may also inadvertently encourage the creation of graphical representations that are condensed to such a degree that they become over-simplifications of complex information.

Although this situation sounds anything but encouraging, big data nevertheless provides humanists and social scientists (who already have a long tradition of using their scholarship to study cultural, political, and social issues)⁵⁹ with new opportunities to foster humanistic and social scientific inquiry.⁶⁰ Some scholars may ask, *Are humanists and social scientists who work with big data indeed creating awareness in the areas of humanity and society and/or advocating change in these areas?* Below, we discuss just a few studies in the humanities and social sciences to give an indication of some of the fields that are beginning to make a contribution in the context of the big data movement. We focus predominantly on those studies in which the researchers have harnessed big data techniques, but without sacrificing traditional, qualitative methods. We are mindful that humanists and social scientists use diverse methodological approaches, and that our discussion does not do justice to covering the full range of what each discipline entails. Our purpose here is simply to give readers a taste of what humanists and social scientists who exploit big data are currently doing.

6.2 Engaged humanists and social scientists

Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide.

– Cathy O’Neil (2017:256)

In the era of big data, humanities scholars and social scientists may be struggling between what they perceive to be only two choices: to continue analysing texts in a slow, methodological way but without analysing the vast amounts of information at their disposal, or to exploit big data methods that sacrifice interpretation to metrics (Gregory, Cooper, Hardie & Rayson 2015:151). However, it is not as simple as choosing one option over the

59 These kinds of studies are diverse, ranging from the exploration of gender patterns in reading literacy (Zuze & Reddy 2014) to the examination of petrocultures (Szeman 2017) and food security (Pradhan & Rao 2018).

60 According to Milan and Gutiérrez (2015:121), the era of big data has fuelled a type of activism referred to as data activism among citizens and ordinary people as opposed to only hackers and open-source activists. They distinguish between pro-active data activism and re-active data activism, although both conceive of information as “a constitutive force in society able to shape social reality” (Milan 2018:155). Re-active data activists resist digital censorship, control, and surveillance through technical practices such as encrypting personal communications, activating anonymous browsing, and blocking advertisements on websites (Milan & van der Velden 2016:57). Pro-active data activism is a little more difficult to conceptualise as it is a relatively new empirical phenomenon, but activists in this domain typically utilise tactics that “range from technology development projects and platforms for the manipulation of data and the visualization of data patterns for campaigning and advocacy” (Milan & Gutiérrez 2015:129). Strictly speaking, humanists and social scientists do not fit into either of these categories, but we contend that many regard themselves as activists if their work is geared towards social change (see Section 6.3 in this respect).

other since scholars need to strike a fine balance between close reading, which cannot be divorced from qualitative approaches to research, and big data methodologies (Gregory *et al.* 2015:151). This recommendation is echoed by DeLyser and Sui (2012:3) who call on researchers to be informed by scholarship in the traditional humanities and social sciences to ensure that they not only resist “superficial number crunching” of data, but that their findings are also contextualised. This advice has been taken to heart by a number of scholars with fruitful results.

In the spatial humanities, for instance, Porter, Atkinson and Gregory (2015) have successfully combined two disparate methodologies to analyse and map nineteenth- and early twentieth-century disease mortality patterns in infants recorded in official population reports for Britain and Ireland from 1850 to 1911. These paired methodologies are geographic information systems or GIS (used to capture, store, manage, analyse, and display spatial or geographic patterns) and corpus linguistics, traditionally employed by linguists to analyse vast volumes of digital texts. Porter *et al.* (2015:27) chose to supplement GIS with corpus linguistics because the latter enabled them to both quantitatively and qualitatively analyse a digital text containing two and a quarter million words. The innovative merging of the methodologies, which the researchers call Geographical Text Analysis (GTA), uncovered new insights into how the population reports were linked to changing mortality patterns in infants over a period of time. The researchers maintain that GTA could be used to make comparisons between the geographies mentioned in the texts and spatial distributions of specific events (Porter *et al.* 2015:33). GTA techniques could also be exploited to help researchers understand other historical and contemporary documents such as library archives, newspapers, diaries, photographs, and literary narratives (Porter *et al.* 2015:34). The kind of qualitative GIS employed by Porter *et al.* (2015) is growing rapidly and contributing to the big data movement in many fields such as sociology, history, and the digital humanities in light of the realisation that big qualitative data such as images and texts also need to be coded for computation (Pavlovskaya 2016:1).

Big data has also partnered quite productively with the field of history and indeed, big history courses have become popular at universities in Australia, the Netherlands, and the United States (Spier 2014:171) because it “offers a fundamentally new understanding of the human past, which allows us to orient ourselves in time and space in a way no other form of history can match” (Spier 2014:172).⁶¹ In this regard, big history enables scholars

61 An extensive search of the web indicates that South African historians have not yet embraced big history. We did find reference to big history on a website hosted by PASCAP, an after-school care organisation that aims to bring big history to children in the Western Cape. On the site, Chris Wheeler makes the following comment with specific reference to an integrated history of the cosmos, earth and humanity: “I can’t help but be inspired by the values and the spirit of wonder Big History embodies, as something that could help empower and galvanise a new scientifically literate generation to add their voices, histories, insights and knowledge to the greatest story ever told”.

to see big patterns they would not ordinarily have seen had they chosen to explore only small procedures (Spier 2014:172). Spier (2014:179) speculates that it is this ability to see new patterns that has encouraged many big historians across the globe to collapse several approaches into a single approach, typically combining scientific views and historical accounts into a single large narrative accompanied by rigorous theory. Walter Alvarez's (2016) book, *A most improbable journey: A big history of our planet and ourselves*, is a reminder that big data should never be staid or one-dimensional in nature and that they may generate enthusiasm for history among academics and students alike. Alvarez (2016) provides an ambitious big history of the universe, foregrounding geological history, but also incorporating cosmic history, an account of earth's major geographical features, and even a discussion of our human bodies to generate a personal narrative about the planet's history.⁶² Like other big data scholars, big historians also struggle to access large datasets from commercial providers when they "[find] themselves locked out of the digital archive by paywalls and search structures which are unhelpful for the kind of analysis that they wish to conduct" (Maxwell-Stewart 2016:360-361). Fortunately, Maxwell-Stewart (2016:362) signals that there are ways around this problem, and that the goal is not necessarily to attempt to capture vast amounts of information from a single source. What many big data scholars are doing instead is to extract information from multiple sources online, thus collecting large quantities of metadata on a particular topic (cf. Chapters 1 and 4 in this book). Historian Catherine Hall (2016), for instance, trawled through vast amounts of online data to find information on slave compensation pay-outs, settler colonialism in Australia, and the migration of imperial families to achieve a deeper understanding about the movement of human and financial assets into Australia between the 1830s and 1840s. Collecting data from multiple digital sources is one way in which scholars may overcome the lack of access to big data discussed in previous chapters.

Although some political scientists view big data with a degree of scepticism, arguing that big data itself cannot help them achieve valid causal inferences and that theory building is a key aspect of their research (Grimmer 2015:81), others have produced both insightful and useful projects when they have combined big data with descriptive inferences, which in turn have helped them create new theories (Grimmer 2015:80). One such project is the VoteView project (Pool & Rosenthal 1997; McCarthy, Pool & Rosenthal 2006) in the United States that, among other things, makes use of NOMINATE scores to map US House and Senate representatives' ideological positions. According to Grimmer (2015:81), an interesting finding to come out of NOMINATE is that ideological polarisation between Republicans and Democrats has increased significantly over the last four decades. Other political scientists have harnessed the power of big data to conduct US presidential

62 See J. Daniel May's (2017) review of the book in *Journal of Big History*.

election forecasting (Linzer 2013) and conflict forecasting (Brandt, Freeman & Schrodt 2011). Social and political scientists Colleoni, Rozza and Arvidsson (2014) have combined machine learning and social network analysis to predict whether Twitter users are Democrats or Republicans. One of their main findings is that Democrats who are not inclined to track Twitter's official accounts tend to demonstrate higher levels of homophily – the tendency to connect with similar individuals – while activists who do follow official accounts have lower levels of homophily. The researchers suggest that in order to better understand political homophily, Twitter users' political and cultural practices need to be considered as well. The lesson here is that big data scholars doing social media research should not focus solely on the social media platform under investigation: “[such] a turn away from ... treating the Internet or SNS as a separate reality and towards a focus on the Internet as one among many aspects of social reality in general ... might open up interesting and fruitful avenues for big data analysis” (Colleoni *et al.* 2014:329).

Digital data in education is another area in which big data has made significant inroads, particularly given the shift in recent years from traditional classrooms to blended and online learning which employs learning management systems such as Moodle, Blackboard, LearnUpon, and Fuse Universal (cf. Reyes 2015:75).⁶³ As is the case in areas of big data science and business intelligence, education has also undergone datafication.⁶⁴ As Selwyn (2015:66) puts it, “schools, colleges, universities and other educational contexts now function increasingly along ‘data driven’ lines”. In fact, digital data work in education has become normative and this is particularly evident in how learning analytics (which measures, collects, and analyses student data with a view to improving teaching and learning) has been embraced by educators (Greller & Drachsler 2012; Siemens 2012). We include a discussion of digital data in education here because sociologists have become increasingly concerned about some of the ethical facets of big digital data in education. One of these facets pertains to data inequality because control, social power, and inequality may be reinforced through processes driven by data (Selwyn 2015:71). Another cause for concern is related to the fact that digital data tends to reinforce what Selwyn (2015:71) refers to as an “increase in managerialism within education”. One bleak study has highlighted how school authorities, after being informed by data generated at a historically African-American school, proposed to close it down, ignoring the fact that the sensibilities of the surrounding community should also be taken into account in the decision-making process (Khalifa, Jennings, Briscoe, Oleszweski & Abdi 2014:148). Of deep concern is that digital data in

63 We refer to educational studies here, since scholars in the humanities may regard this field as a humanistic endeavour.

64 Data could include, among other things, naturally occurring data generated via a learning management system, data based on library use, assessment grades, and the like.

education could be exploited to conduct “dataveillance” (Selwyn 2015:73). Unfortunately, dataveillance practices have been observed by a number of scholars including Land and Bayne (2005), Knox (2010), Rosenzweig (2012), and Taylor (2013). Sociologists remain perturbed that school authorities may be more interested in the operationality that digital data demonstrates than in the social meanings it generates. Selwyn (2015:79) calls on all interested stakeholders – academics, educators, parents, and pupils – to take a stand against “the ‘politics of [big] data’ in education, and not to take them at “face value”.

Social scientists studying different aspects of social media have also begun exploring the benefits of complementing their traditional research methods with those from big data. An enlightening study in this respect is one by Williams and Burlap (2016) who combined their criminology and computer science skills to examine cyber hate on Twitter following the murder of Lee Rigby by Islamist extremists in 2013. What makes this study innovative is not only the fact that it reflects a new interdisciplinary methodology the researchers call “computational criminology” (Williams & Burlap 2016:217), but also that it is one of the first to exploit sophisticated big data techniques and analytics to analyse contemporary crimes such as cyber hate. Another informative study using big data analytics is one conducted by Innes, Roberts, Preece and Rogers (2016) aimed at analysing social media reactions to the murder of Lee Rigby. In contrast to Williams and Burlap (2016) who paid attention to quantified measures of cyber hate speech following the terrorist attack, Innes *et al.* (2016) focused on understanding the content of social media communications through detailed qualitative coding of users’ tweets. Reviewing Felt’s (2016) article on the intersection between social science and big data analytics in the context of social media research, it is clear she would in all likelihood favour the study by Innes and his colleagues rather than that by Williams and Burlap (2016) as she contends that only traditional, qualitative methods would “enable both the big picture and the close, critical view” of the data collected.

Psychology is yet another discipline that could make a meaningful contribution to big data research, and like researchers in other disciplines, those conducting psychological research are of the view that psychologists should harness their own competencies rather than try to develop new computational skills. Two psychologists who support this position are Cheung and Jal (2016:4) who point out that most psychologists are sufficiently competent to make use of data analytics to analyse large datasets because they are trained in psychological theories, statistics, and psychometrics. These researchers propose that scholars make use of what they call the “SAM” or “split/analyse/meta-analyze approach” to test their theories on big datasets that are based on human behaviour. Since standard computers cannot accommodate huge amounts of data, the first step entails splitting the data into many datasets. In the analysing step, common statistical analyses are employed such as regression analysis and multilevel analysis. The third step entails carrying out a meta-

analysis so that statistical inferences can be made (Cheung & Jal 2016:4). At the same time, Cheung and Jal (2016:10) argue that psychologists should “lower the threshold for engaging in big data research” because data analytics is complex. They also suggest that psychologists collaborate with researchers who have sophisticated computational skills, a suggestion that, as we have seen, has been made by many scholars since the advent of the big data phenomenon (Chapter 2).

A recent study to emerge out of the digital humanities is one by Brown, Mendenhall, Black, van Moer, Lourentza, Flynn, McKee and Zevai (2016)⁶⁵ who employed black feminist theory to make sense of large datasets. What makes this data study quite fascinating is that it simultaneously interrogates the biases in computational analysis and the digital humanities while exploiting both to recover and preserve black women’s narratives. Challenging “the embedded whiteness and maleness of computational analysis” and critiquing the digital humanities for failing to examine issues of identity and power, Brown *et al.* (2016) searched through approximately 800 000 documents – articles, books, and newspapers – archived in two digital libraries, JSTOR and the HathiTrust, in an attempt to identify black women’s perceptions about and lived experiences in the United States. In order to make sense of the large datasets, Brown *et al.* (2016) made use of MALLEET for topic modelling, comparative text mining, and data visualisation. Topic modelling allowed the researchers to explore and interrogate different genres of text such as sociology or poetry in the large datasets, while comparative text mining helped them identify latent and common themes across all data collected. Isolating a sub-set of all the data collected, Brown *et al.* (2016) discovered that many texts by or about black women do not have metadata tags,⁶⁶ which means that scholars navigating the Internet will not know of their existence. The research team’s ultimate goal is to identify all untagged volumes and then make the entire corpus available online.

In the field of linguistics, a number of big data studies have begun focusing on artificial intelligence fields (such as machine learning and natural language processing) to detect and/or prevent cyberbullying and suicide. The former, defined as “wilful and repeated harm inflicted through the medium of electronic text” (Patchin & Hinduja 2006:152), has reached perturbing levels owing to the rapid growth of social networking globally, and for this reason several researchers have turned to big data analytics and combined it with computational linguistics with a view to creating language models to automatically detect cyberbullying content. For instance, in a recent study by van Hee, Jacobs, Emmerly, Desmet, Lefever *et al.* (2018), trained linguistics employed a fine-grained annotation scheme to

65 This research team was made up of humanities scholars, social scientists, and data researchers.

66 Tags are metadata (keywords/terms/snippets of text) about a particular resource. For example, ‘South African novelist’ and ‘playwright’ would be tags for Zakes Mda; they describe the author, thus providing the browser with useful information. Many websites provide user-friendly tutorials on how to create metadata tags.

analyse 192 085 social media posts from a site called ASKfm. The main aim of the research team was to model bullying attacks and reactions from both victims and bystanders so that any signals of potential cyberbullying events could be quickly investigated and prevented by human moderators.⁶⁷ The researchers' annotation scheme took into account that existing parental control tools, while useful in detecting keywords that reflect profanity or insulting words, are unable to perceive covert forms of cyberbullying, particularly when these forms do not contain explicit vocabulary. The annotation categories thus encompassed expressions related, amongst other things, to threats/blackmail, insults, curses, defamation, sexual talk, bystander/victim defense, and encouragement in support of the harasser (Hee *et al.* 2018:8-9). The development of fine-grained annotation schemes is crucial, given that some forms of cyberbullying such as defamation may be fairly difficult to recognise.⁶⁸ Other researchers who have combined big data analytics and computational linguistics to detect cyberbullying include Dinekar, Reichart and Lieberman (2011), Spitzberg and Gawron (2016), Power, Keane, Nolan and O'Neill (2017), and Lee, Lee, Park and Han (2018). These researchers and many others would agree that what makes the automatic detection of cyberbullying difficult is the deliberate obfuscation of abusive language, hence the need for more sophisticated detection systems. What scholars have found is that abusers use various strategies to avoid exposure by replacing a single character in an offensive word (cf. Pitsilis, Ramampiaro & Langseth 2018) or by making use of newly invented words (Lee *et al.* 2018:23).

As far as assessing risk of suicide is concerned, the linguistic analysis of suicide notes is not new and can be traced back to *Clues to suicide* (1957) edited by Edwin Shneidman and Norman Farberow (cf. Desment & Hoste 2013:6352). Employing discourse analysis for the most part, contributors to this book analysed 66 authentic or fabricated suicide notes in an attempt to distinguish between real and false notes.⁶⁹ Essentially early work on suicide risk assessment was based on analyses of surface-level elements such as the individual's choice of verbs, adverbs, modals, and auxiliaries (Osgood & Walker 1959; Gleser, Gottschalk & Springer 1961; Desment & Hoste 2013:6352). A number of researchers have made use of computational methods to classify suicide notes as genuine or elicited (inauthentic). Jones

67 Research by van Royen, Poels, Daelemans and Vandebosch (2014) suggests that most online users are not averse to automatic monitoring of cyberbullying on condition that their privacy and autonomy are guaranteed.

68 For example, Hee *et al.* (2018:10) point out that while an encouragement in support of the harasser such as "I agree we should send her hate" is explicit and thus easily recognisable, one such as "hahaha" or "LOL" is not.

69 For example, research shows that in contrast to simulated notes, real suicide notes tend to be longer. They also contain more pronouns as well as more references to people and social phenomena (Fernández-Cabana, Ceballos-Espinoza, Mateos, Aeresá Alves-Pérez, Gómez-Reino Rodríguez & García-Caballero 2015:147).

and Bennell (2007), for example, classified suicide notes in terms of structure variables such as sentence length and parts of speech as well as in terms of content features such as the individual's instructions and explanation for his/her intention, while Handelman and Lester (2007) analysed the semantic content of words in suicide notes using a text analysis programme developed by Pennebaker, Fancis and Booth (2001).⁷⁰ In the last decade, researchers have begun experimenting with machine learning techniques in order to classify suicide notes (Pestian, Nasrallah, Matykiewicz, Bennett & Leenaars 2010; Yang, Willis, De Roeck & Nuseibeh 2012; Cheng, Li, Kwok, Zhu & Yip 2017; Just, Pan, Cherkassky, McMakin, Cha, Nock & Brent 2017; Walsh, Ribeiro & Franklin 2017). It appears that traditional statistical approaches to predicting suicide attempts in clinical psychology are not as accurate as machine learning techniques since they generally employ logistic regression; commenting on work carried out by researchers such as Colin, Walsh, Ribeiro and Franklin (2017) and Just *et al.* (2017), O'Connor and Kirtley (2018:7) argue that “[r]ecent advances in machine learning techniques allow the computation of optimized risk algorithms, from hundreds of different individual variable pathways, to suicidal thoughts and behaviour”.

In South Africa, scholars in the humanities and in the social sciences have embarked on exciting research projects through technological and data-driven research. Karli Brittz (2018), for example, who works in the Department of Visual Arts at the University of Pretoria, explores data artists' works which reflect the use of big data analytics as well as visualisation techniques with a view to making big data more accessible and useful to society. What makes Brittz' (2018) study particularly interesting is that she attempts to reconcile art and dataism.

A review of the scholarly literature reveals that it is in the field of learning analytics that South African scholars have begun harnessing the power of big data. Matsebula and Makandla (2019), for example, have considered how big data architecture in the context of higher education needs to be conceptualised and tailor made for specific institutions so that analysts are able to extract meaningful insights from educational data. In their study of how learning analytics is practised in South Africa, Lemmens and Henn (2016: 250) offer the caveat that this dimension of institutional research should not result in a “data silo”, which describes “a situation where fragmented data is collected, analysed and stored on personal or distributed systems”. Other scholars who have explored learning analytics in South African higher education settings are Jordaan and Van der Merwe (2015), who have reviewed best practices for implementing a learning analytics strategy in institutions of higher learning and Prinsloo and Rowe (2015), who call for analysts to be mindful of the ethical issues surrounding the use of student data. We do not offer a detailed opinion on learning analytics

70 Handelman and Lester (2007) classified words in terms of the use of future tense verbs, metaphysical references, social references, and negative or positive emotions, to name a few variables.

from an African perspective “considering that the African continent comprises 54 sovereign states, each with its unique regulatory framework, development agenda, information and communications technology (ICT) infrastructure, and state of adoption of online learning” (Prinsloo 2018:25).

6.3 Big data as an obstacle/bridge to humanitarian projects

Big data and increased connectivity allow humanitarian organizations to better understand where to target humanitarian assistance. – Helena Kamper

An increasingly significant online trend is the use of big data in big digital humanitarianism, which may be defined as “the *enacting* of social and institutional networks, technologies, and practices that enable large, unrestricted numbers of remote and on-the-ground individuals to collaborate on humanitarian management through digital technologies” (Burns 2014:52). The purposes of digital humanitarianism are myriad, ranging from managing social/natural disasters and uncovering truths or untruths to understanding why disasters occur in the first place (Barnett (2008:259) – all of which are not too far removed from what many social scientists (and some humanists)⁷¹ do (cf. Barnett 2008:259). A simple Google search using the keywords *digital humanitarianism/humanitarians* and *social sciences/social scientists* yields an interesting result, namely, that digital humanitarians are calling on social and data scientists to assist them in dealing with what Patrick Meier (2015:99) describes as “the overflow of information generated during a disaster [and which] can be as paralyzing to humanitarian response as the absence of information”. In this regard, a major crisis that the big data deluge has created pertains to the flood of misinformation or disinformation generated on social media platforms during disasters. Bruce Lindsay (2011:7), for example, points out that false, inaccurate or malicious social media posts hinder or delay humanitarian response efforts and in some cases create an unsafe environment for both the community and first responders. To pre-empt or at least reduce these kinds of consequences, we argue for social science scholars to collaborate with data scientists and to employ crowdsourcing with a view to verifying or invalidating huge amounts of information generated via social media platforms during times of disaster. In the context of disaster management, crowdsourcing “is the volunteer-generated, decentralized contribution of [crisis] information online” (Harrison & Johnson 2016:17), and several researchers have begun tapping into this kind of information to glean valuable and accurate information about disasters (cf. Barton

71 Linguistic analyses of huge volumes of online posts are widely employed to extract useful information during disasters. Sarah Vieweg (2012), for example, carried out a linguistic analysis of verbs in tweets with a view to showing how such an analysis may augment situational awareness of mass emergency situations. Cresci, Tesconi, Cimion and Dell’Orletta (2015) also used linguistic analysis to detect social media messages vital for accurately assessing damage during natural disasters.

2018). In one interesting study, a team of researchers made use of automatic methods to extract useful information from text-based microblogging messages generated when the so-called Joplin 2011 tornado struck Joplin, Missouri in the United States with catastrophic results (Imran, Elbassuoni, Castillo, Diaz & Meier 2013). What the research has yielded is an automatic system that filters out information irrelevant to a given disaster while detecting informative messages that will facilitate what digital humanitarians refer to as situational awareness of the disaster. Since this 2013 study, many others have appeared, focusing on, amongst other things, ebola outbreaks (Odlum & Yoon 2015; Kim, Jeong, Kim, Kang & Song 2016), rockslides (Dammeier, Moore, Hammer, Haslinger & Loew 2016), and floods (Lo, Wu, Lin, Hsu 2015).

6.4 Size revisited

Big data: Does size matter? – Timandra Harkness (2016:1)

To return to data size, the questions uppermost in readers minds might be these: is it necessary for qualitative researchers to collect enormous amounts of data to achieve their research objectives? In other words, is the analysis of large datasets theoretically justified? Will the demand for big data projects result in the demise of small data studies? Mahrt and Scharrow (2013:23) partially answer both questions when they contend that researchers need to determine if coding a huge data set has any inherent value since it may have limitations in terms of validity and scope (Mahrt & Scharrow 2013:27). We noted in Chapters 3 and 4 that a given sample, no matter how large, may not be representative of a specific population, making generalisability of results difficult if not impossible. In the field of social media research, for example, Manovich (2012:465) laments that “[p]eople’s posts, tweets, uploaded photographs, comments, and other types of online participation are not transparent windows into their selves; instead, they are often carefully curated and systematically managed”. Big data experts advise researchers to first generate a small data sample to glean preliminary insights before collecting large amounts of data (Rojas, Kery, Rosenthal & Dey 2017:26). When determining just how much data to collect, researchers should always take the (social) context of data into account since the data cannot speak for itself (cf. Frické 2015). Throughout this book, we have seen numerous examples of studies that rely on the context of the data collected to make sense of the content in that data. For media and communications scholar Shani Orgad (2009:34), a fundamental question to be asked when doing a qualitative analysis of data taken from the Internet is the following: “what does ‘the Internet’ stand for in a particular context, for particular agents”? She rightly observes that cyberspace is not a monolithic space; it is “a collection of locations much like the real world” (Reips, Buchanan, Krantz & McGraw 2015:141). Thus, for example,

when Orgad (2005) explored how breast cancer patients communicate in online spaces, she initially mapped what she calls “the landscape of breast cancer patients’ communication” (Orgad 2009:34) by identifying the spaces or arenas in which these participants engage which may be (online) message boards and (offline) personal diaries.

To come back to the question of data size, and in the context of social media research, Mahrt and Scharkow (2013:20) have noted that smaller-scale analyses of online users’ messages and/or behaviour may yield meaningful insights, provided that researchers employ sampling, measurement, and analytical procedures that are sound. Kaplan, Chambers and Glasgow (2014:342) offer the caveat that we should not be seduced by the notion that huge datasets, as opposed to small ones, will yield more reliable and meaningful findings. Kaplan *et al.* (2014) provide several examples from epidemiology, clinical trials, and health service research to illustrate that very large datasets could result in sampling biases⁷² and significant inferential errors.⁷³

Scholars have argued that since size is relative in the world of big data and that it is not linked solely to volume, small data can in fact be quite large in size. We have noted, for instance, that in their study of black women’s narratives, Brown *et al.* (2016) collected 800 000 periodicals. This is a large dataset, although some big data analysts would argue that while it reflects variety in the sense that it is made up of books, articles, and newspapers, it lacks the other fundamental ontological attributes that big data should have, namely, velocity and volume. The solution to this problem is to scale small data into data infrastructures – that is, “pool ... and link small data in order to create larger datasets” (Kitchin & Lauriault 2015:463):⁷⁴

Whilst the scaling of small data into data infrastructures does not create big data, in the sense that the data still lack velocity and exhaustivity, it does make them more big data-like by making them more extensive, relational and interconnected, varied, and flexible. This enables two effects to occur. First, it opens scaled small data to new epistemologies and, in particular, to new forms of big data analytics ... [and second], it facilitates small data being conjoined with big data to produce more complex, inter-related and wide-ranging data infrastructures (Kitchin & Lauriault 2015:470).

72 For example, Kaplan *et al.* (2014:343) describe a nurses health study of just over 48000 postmenopausal women that did not take into account the atypical nature of the sample under investigation. Researchers in the study erroneously concluded that hormone replacement therapy significantly reduces coronary heart disease.

73 See Chapter 3 in which we referred to Leinweber’s (20007) warning that large datasets can be manipulated to yield questionable correlations.

74 Bollier (2010:12) points out that “[it] is generally safer to use larger [datasets] from multiple sources” to avoid the risk of drawing the wrong conclusions.

In their discussion of the value that small data still holds, boyd and Crawford (2012:670) describe the work of Veinot (2007), who studied only one individual, a blue-collar worker at a hydroelectric power plant, with a view to studying workplace information practices. They conclude that small data may be superior to big data in some instances: “[Veinot’s] work tells a story that could not be discovered by farming millions of Facebook or Twitter accounts”. These sentiments are echoed by Bollier (2007:14) in *The promise and peril of big data* when he argues that more is not necessarily better; he quotes Stefaan Verhulst (currently co-founder and chief of research at the Governance Laboratory at New York University), who observes that “[people] quite often fail to understand the data points that they actually need, and so they just collect everything or just embrace Big Data. In many cases, less is actually more.”

In further defense of the use of smaller data inputs, it is worthwhile noting that very big datasets may generate “dirty” or “biased” data (Kitchin & Lauriault 2015:466) which in turn will impact negatively on validity. Although big data proponents may be critical of small data for failing to reflect volume or velocity, “[s]mall data studies ... seek to mine gold from working a narrow seam, whereas big data studies seek to extract nuggets through open-pit mining, scooping up and sieving huge tracts of land” (Kitchin & Lauriault 2015:466).

The place of qualitative data analysis software (QDAS) programmes in a big data world

Comprehension of [QDAS] as a facilitator and data management system, and not an alternative for data immersion and analysis, will serve the qualitative researcher well.

– Diane Cope (2014:323)

Despite being fairly well established, little is known about how qualitative researchers employ qualitative data analysis software (QDAS) programmes to analyse their datasets (cf. Woods, Paulus, Atkins & Macklin (2016:597), while even less information is available when it comes to the use of these programmes in the arena of big data. We thus consider the place of QDAS tools in the era of big data in this chapter, not only discussing recent developments in their design, but also critically appraising their usefulness to ‘big’ qualitative researchers by reviewing their advantages as well as their pitfalls.

7.1 Software programmes and the qualitative researcher

... [w]hile qualitative data analysis software ... will not do the analysis for the researcher, it can make the analytical process more flexible, transparent and ultimately more trustworthy. – Florian Kaefer, Juliet Roper and Paresha Sinha (2015:1)

According to political scientist and software inventor Stuart Shulman (2014), scholars consider themselves to be either purists, who prefer to focus exclusively on rich and in-depth interpretation of data, pluralists, who favour experimental, mixed methods, or positivists, who rely on scientific quantitative methods to achieve validity, reliability, and objectivity. A large number of qualitative researchers position themselves in between purists and pluralists (Evers 2018:66), and are seeking to employ qualitative data analysis software (QDAS) tools that support the analytic process that takes place in the mind (Evers 2018:65). Selecting an appropriate software tool is fairly challenging, given that many software tools such as Tableau, Statistical package for the social sciences (SPSS), and the free, open-source software referred to as R, are designed to support quantitative rather than qualitative research. Furthermore, the myth persists that different QDAS programmes reflect different methodological stances: ATLAS.ti, for example, appears to promote hermeneutics and grounded theory (Friese 2014), while MAXQDA is regarded as supporting mixed methods research (Guetterman, Creswell

& Kuckartz 2015). The assumption that QDAS tools tend to champion grounded theory in particular is challenged in the literature: “the perception of a grounded theory bias within [QDAS] is countered not only by a reminder that the functions offered by [QDAS] are employed within other methodological frameworks ... but also by a reminder that grounded theory is an ambiguous methodology” (Tummons 2014:5). Tummons (2014:5) notes that “theoretical vagueness” in qualitative research has helped fuel this erroneous assumption. He offers the caveat that researchers should not regard software programmes as driving the research process: “[QDAS] can provide the tools, but it cannot do the analysis” (Tummons 2014:5). Indeed, it is no coincidence that Tummons (2014) qualifies the term QDAS with the adjective *computer-assisted* (CAQDAS) to emphasise the fact that software tools can only *aid* the researcher to carry out a variety of tasks such as data collection, storage, coding, and data visualisation (cf. Friese 2016:2).

7.2 Two ways of thinking about QDAS

... we too often rely on our intuition and routine thinking for big decisions when we should actually slow down and become more analytical. – John Reeves (2014:2)

By now it should be clear that we are not advocating for big data analysis to replace a close, interpretative analysis of information. Instead, we are proposing that scholars merge computational and interpretative instruments as several scholars have done (Drucker 2011; Friese 2016; van Dijck 2016; Taylor, Gregory & Donaldson 2017). Van Dijck (2016:13) contends that such a merging is not new, and makes use of two analogies to prove his point: the magnetic resonance imaging (MRI) scanner has not entirely replaced X-rays, computed tomography (CT) scanners, and ultrasound, since all these instruments complement as well as overcome one another’s limitations by offering different and unique diagnostics. In addition, interpretation of each of these device’s images does not occur automatically, but is the result of many years’ commitment to interpretation and fine-tuning of their features (van Dijck 2016:13). Similarly, the microscope will not supplant the telescope as they are not competing instruments (van Dijck 2016:13)⁷⁵ and both require human interpretation of what they reveal to the naked eye. In the same way, combining interpretative and digital methods “does not mean that ... [we] ‘surrender’ to a new methodological paradigm” (van Dijck 2016:17) in either the humanities or social sciences.

Friese (2016:35) offers researchers working in a qualitative paradigm useful advice when she argues that they should consider thinking about the development of QDAS tools

75 Indeed, Charles Perrault (1693) observed that “with the help of the telescope and microscope it was possible to discover the immeasurable space in the largest and the smallest bodies, which gives an almost infinite extent to science, which engages with them”.

in terms of Nobel Prize winner Daniel Kahneman's (2011) distinction between fast and slow thinking. In terms of Kahneman's (2011:23) model drawn from behavioural psychology, "System 1" describes the brain's ability to make rapid, intuitive, and automatic decisions, while "System 2" or slow thinking encompasses making choices and decisions at a more leisurely pace. Friese (2016:36) links these two systems to innovations in QDAS, asserting that in the past, software tools tended to accommodate only System 2, since they demanded a considerable amount of time and energy on the part of the researcher to read through data and then to manually interpret that data. She goes on to argue that in an era of big data, what is needed are tools that support System 1 as well, and indeed, many are now designed to do so, reflecting features that allow for rapid, automatic coding⁷⁶ and for the creation of word clouds (see Figure 7.1), word trees, tables, and a variety of graphical displays. ATLAS.ti, NVivo, MAXQDA, and QDA Miner are all examples of software tools that can store fairly large (but not enormous) datasets⁷⁷ whether text-, audio- or video-based, facilitate coding of data, generate word clouds and word trees, for instance, and create visualisations of data.

76 When it comes to automatic coding, software tools have been designed in such a way that they allow for the coding of strings of words and for the identification of themes in the data collected.

77 As noted in the previous chapter, qualitative researchers may work with very large datasets that big data analysts do not regard as reflecting all the Vs, namely, volume, velocity, variety, and veracity. Nevertheless, qualitative scholars tend to create large datasets by collecting and then linking small datasets (Kitchin & Lauriault 2015:463) that several QDAS tools are able to accommodate.



Figure 7.1: An example of a word cloud from South Africa's Life Esidimeni Arbitration Hearings, 24 and 25 January 2018⁷⁸

We are not suggesting that System 2 should be subordinated to System 1, however. Friese (2016) describes her own challenges when doing a research study which clearly illustrates that both fast and slow thinking are essential to the qualitative researcher, since an analysis based on the former type of thinking cannot be guaranteed to be accurate.

78 Source: <https://www.youtube.com/watch?v=bsoO8pkkt6o>. A scandal and national tragedy hit South Africa on 1 February, 2017 when Health Ombudsman Prof. Malegapuru Makgoba released a report detailing the deaths of mentally ill patients at psychiatric facilities located in Gauteng. A total of 144 individuals died, and neglect and starvation were listed as some of the causes of death. It is possible that more people died, but the exact number is not yet known. The hearings took place between September 2017 and March 2018, and were presided over by retired Deputy Chief Justice, Dikgang Moseneke.

Frieser (2016:37-38) employed System 1 tools to create a word cloud from a survey she conducted based on open-ended questions, with a view to determining which words occurred with specific themes. Although she made use of auto-coding, during which the software automatically sorted the words initially identified by theme, she realised that she nevertheless could not entirely trust the coding as she herself had to employ System 2 thinking to manually check the auto-coded segments for accuracy. The situation becomes more complicated when one takes into account that QDAS tools (such as NVivo) have a tendency to either take a very long time to code data or to crash under huge amounts of data. Best practices to resolve or pre-empt these problems include storing data outside the given QDAS programme and cleaning a dirty database. With respect to data storage, Kaefer *et al.* (2015:7) suggest that frequent back-ups be made using Dropbox, Microsoft's OneDrive or Google Cloud Platform, to name a few, although Li, Gai, Qiu, Qiu and Zhao (2016:103) warn that serious security issues still surround these cloud-based storage spaces in the sense that cloud service operators may access sensitive data.⁷⁹ When it comes to data cleaning, the literature at this stage remains heavily skewed in favour of quantitative data cleaning (Rousseuw & Leroy 2005; Hellerstein 2008; Aggarwal 2003), but Chu and Ilyas (2016) provide taxonomies of qualitative error detection and data repair techniques for scholars working with big data. The first taxonomy pertains to detecting so-called surface anomalies by determining what, how, and where these anomalies occur. In this respect and in a detailed article, Chu and Ilyas (2016) provide scholars who have minimal training in software use with a user-friendly tutorial on how to answer the three questions posed. In terms of the second taxonomy, these researchers also provide a detailed explanation of what, how, and where to repair an erroneous database. Both error detection and error repair techniques may be either automatic or guided by humans.⁸⁰ Using these cleaning techniques is essential to enhance the quality of data as discussed in Chapter 3 (cf. Cai & Zhu 2015; Fan, Xiao & Yan 2015).

Another design challenge for QDAS developers which qualitative researchers need to be aware of lies in the field of sentiment analysis which involves coding opinions in a piece of text as either positive, negative or neutral. Friese (2016:38-39) remarks that what is highly problematic is that QDAS may sometimes incorrectly classify opinions. A major hurdle in this regard involves sarcasm and irony in social media, two phenomena which sentiment coding cannot easily detect (Farias 2017). Thus, for example, a detection engine may not classify *Of course I'd love to spend the last Rand I have on an expensive TV* as being sarcastic.

79 These researchers have proposed a unique cryptography method that effectively blocks cloud operators from accessing users' data on cloud servers. Essentially, the method entails dividing sensitive data into two encrypted components and then storing them on different cloud servers.

80 If data has been duplicated by mistake, for example, a machine will not be able to detect this as accurately as a human being can (Chen, Zobel, Zhang & Verspoor 2016).

Indeed, “[sarcasm] ... is usually ignored in social media analysis, because it is considered too tricky to handle” (Maynard & Greenwood 2014:4238).⁸¹

7.3 QDAS and blended reading

Both close and distant readings reveal not just what is in a data set, but how that data might be enacted. – Yanni Loukissas (2016:4)

Besides the above challenges, the voluminous amounts of data to be managed remain problematic and no software programme exists that, at the click of a mouse, has the ability to churn out accurate analyses of the information contained in the data. Initially conceived by Lemke (2014), a number of researchers have suggested that qualitative researchers employ what they call blended reading to analyse large data sets (Lemke, Niekler, Schaal & Wiedemann 2015; Hammond, Brooke & Hirst 2016; Stulpe & Lemke 2016; Loukissas 2016). Blending reading combines distant and close reading, thus “integrating quantitative and qualitative analyses of complex data progressively” (Lemke *et al.* 2015:7) to provide a multifaceted view of data. A number of scholars have found a combination of close and distant reading of texts to be fruitful in film history (Hoyt 2014), digital research (Wills 2016), and literary-historical scholarship (Taylor *et al.* 2017), to name a few areas. Here, we briefly describe studies by Rauscher (2014) and Reiberg (2016) to give some indication of how distant and close reading may be combined.

In his exploration of literary representations of cities in crime novels, Rauscher (2014) combined a qualitative close reading of fairly large quantities of literary texts with computer-assisted QDAS analysis, corpus linguistics, and distant reading in an iterative research process, moving constantly between close and distant reading. Rauscher’s (2016) study is a useful one in the sense that it shows how crime novels have the potential to serve as a rich “data basis for urban sociology and interdisciplinary research questions about the distinctiveness of cities” (Rauscher 2014:68). Significantly, Rauscher (2014) exploited thick description to analyse patterns of discourse in and across 240 novels, a method that is becoming increasingly important to enrich (big) data analytics (Felt 2016) and which we discussed in Chapter 3. Using the term thick analysis, but with acknowledgement to Clifford Geertz (1973) for having coined the term thick description, Evers (2015:1) contends that although time consuming, thick analysis is useful for “[enhancing] the depth and breadth of data analysis by creatively combining several analysis methods, allowing for a more comprehensive analysis” of data. Evers (2015:7) also asserts that QDAS programmes

81 Having said this, researchers who adopt a computational linguistic approach to sentiment analysis have actively worked on identifying irony and sarcasm (Reyes, Rosso & Veale 2013; Sulis, Farias, Rosso, Patti & Ruffo 2016; van Hee, Lefever & Hoste 2018).

that offer hyperlinking, visualisation, and annotation tools, for example, will help augment thick description. A number of qualitative software programmes allow for thick description including NVivo (cf. Mwangi 2017), MAXQDA (cf. Goldenberg, Darbes & Stephenson 2017), and ATLAS.ti (Ang, Embi & Yunus 2016). What is particularly advantageous about these programmes is that if a team of researchers is working on a particular study, each researcher has the opportunity to code the data independently and to explain his/her reasoning via thick description of the data via a given programme's memo function.⁸² As Brokensha and Conradie (2016:9) observe, the memo function “not only [allows] researchers to immerse themselves in the data, thus avoiding the so-called ‘tactile-digital divide’ (Gilbert 2002, 216), but also affords greater transparency because researchers can share insights, challenges, and doubts (Tracy 2010, 842)” they may harbour about their individual interpretations (cf. Tummons 2014:8).

Reiberg (2016) explored the first five years of Germany's Internet policy with a view to gaining insights into how this policy was socially constructed by political actors in debates reflected in two large corpora, namely, parliamentary minutes (from 1996 to 1998) and approximately 37 000 articles published in five major newspapers between 1994 and 1998. Since the corpora were so large, making it impossible to conduct a close reading of the content in them, Reiberg (2016) initially carried out distant reading of 129 articles from one newspaper to determine the frequency of terms related to the word ‘Internet’ and its synonyms/synonymous terms. Next, he selected content that pertained strictly to Internet policy, further reducing the corpus through a list of keywords that were synonyms of Internet policies or reflected the names of organisations that focused on Internet policy. This list was then employed to both select and process the remaining articles, and the corpus was analysed in terms of the names of Internet policies. Identifying only those Internet policies that were enacted between 1996 and 1998, Reiberg (2016) searched for terms that were synonymous with these policies before moving on to the second phase of his research. In this phase, Reiberg (2016) carried out both distant and close reading in the sense that he employed a distanced look to identify broad themes in the corpora before using qualitative content analysis to conduct a close reading of the reasoning used by political actors to generate two types of statements – either societal problem statements or demands for state intervention in the context of the domain of Internet policy. This close reading allowed Reiberg (2016) to generate a codebook to describe the features of each type of statement. In

82 Memos, along with graphical representations and search results, for example, are referred to as secondary documents in the literature on qualitative data analysis software. The practice of generating both primary and secondary documents is referred to as system closure. Managing both kinds of documentation in a given software programme “makes it a simple task for the researcher to search and then code his or her ongoing analytical or explanatory material using the same coding structure as has been used for the primary data” (Tummons 2014:7).

a third and final phase, each type of statement was analysed to identify their similarities and determine political actors' shared understandings of the new Internet policy domain. What Reiberg (2016:10) ultimately concluded was that the term "information society"⁸³ was one frequently used by political actors to construct a shared understanding of Internet policy. It is notable that Reiberg made use of MAXQDA to carry out his study, since it is a software programme that allows the researcher to carry out System 1 and System 2 analyses of coded data. In addition, and as Friese (2016:40) puts it, such a software programme "[supports] qualitative data analysis rather than offering to [analyse] the data".

Studies by Rauscher (2014), Reiburg (2016), and others offer a number of important lessons, one being that a close reading of large datasets is not an impossible task when combined with computational analysis (cf. Hammond *et al.* 2016). However, a more important lesson is that close reading is by no means obsolete in a big data world, contrary to what Jockers (2013:7) maintains in the context of literary history when he writes that "[close] reading is not only impractical as a means of evidence gathering ... but big data render it totally inappropriate as a method".⁸⁴ Hammond *et al.* (2016:74) are of the view that rather than perceiving close and distant reading to be parallel structures for interpretation as Jockers (2013) does, scholars should "use computational analysis to test, probe and enliven human close readings". That is, they should adopt a hybrid approach which reflects a feedback loop in which distant and close readings are in a continual and reciprocal dialogue. This is the approach that Rauscher (2014) adopted when he studied how cities are depicted in crime novels: he refers, for example, to employing a feedback loop in which interpretations based on close and distant readings constantly challenge one another. Rauscher (2014:96) observes that for in-depth interpretation of data, "distant reading and visualization alone are not sufficient", while a "particular and limited close reading can (and should) be enhanced and enlarged to a wider context" through distant reading.

Ultimately, qualitative researchers need to decide for themselves whether or not they would like to work with much larger datasets. Friese (2016:44) calls on scholars to at least consider using CAQDAS when analysing their data, paraphrasing Hitchcock (2014) when she states that "CAQDAS and 'Big Data' absolutely need to have a conversation. The subject of that conversation should be on how (or whether) to integrate close reading and small data, and distant reading and large data". Qualitative software tools are in a sense at an embryonic stage, and much work still needs to be done to refine them so that they are

83 This terms "refers to an inevitable, positive change societies in general are undergoing at different speeds" (Reiberg 2016:10).

84 Interestingly, and in the next sentence, Jockers (2013:7) confusingly remarks that "[this] is not to imply that scholars have been wholly unsuccessful in employing close reading to the study of literary history". He cites the works of Ian Watt and Erich Auerbach as excellent examples of close readings of texts.

able to accommodate increasingly large amounts of information. Nevertheless, these tools may open doors scholars never imagined existed. Van Dijck (2016:17) puts it eloquently when he observes that “[as] academic guardians of the arts, culture, language, heritage, and the traditions of humanities thinking, we will have to engage in multifarious ways with the interrelatedness of digital technology in all kinds of cultural practices”. This statement applies equally well to scholars working in the social sciences (cf. Smith 2017).

7.4 QDAS, qualitative content analysis, and big data

In the era of “big data”, the methodological technique of content analysis can be the most powerful tool in the researcher’s kit. – Steven Stemler (2015:1)

Although content analysis as a method reflects a number of limitations related, amongst other things, to issues of validity and intercoder reliability⁸⁵ as well as to the problem of being time-consuming, several scholars recommend that QDAS tools be combined with qualitative content analysis since this method allows scholars “to systematically transform a large amount of text into a highly organised and concise summary of key results” (Erlingsson & Brysiewicz 2017:94; cf. Elo & Kyngäs 2008:113; Lewis, Zamith & Hermida 2013:34; Renz, Carrington & Badger 2018:824). Indeed, Kaefer, Roper and Sinha (2015:1) argue that when combined with QDAS tools, this method improves both the transparency and trustworthiness of the qualitative research process in the context of big data. However, there are a number of guidelines that researchers should follow when combining QDAS and qualitative content analysis (QCA) which we briefly address here.

Unlike *quantitative* content analysis, which emphasises objective, quantitative, and systematic descriptions of messages (Berelson 1952:18), QCA focuses instead on context, aiming for “the subjective interpretation of the content of text data” (Hsieh & Shannon 2005:1278) which avoids “rash quantification” (Mayring 2000:1). Kaefer *et al.* (2015:1-2) suggest that scholars make a distinction between qualitative data analysis (QDA) and QCA because the former generates interpretations based on an entire body of texts, while the latter reflects a (usually smaller) quantitative component. Two major criticisms levelled at QCA in the context of big data research are – paradoxically – that the use of software has a distancing effect in the sense that it removes the researcher from the data, while being too deeply immersed in the data prevents the researcher from seeing the bigger picture, as it were (cf. Bazeley 2007; Ryan 2009). We noted in the previous section that a number of

85 Validity may be called into question if the coding process is neither consistent nor coherent (cf. Renz *et al.* 2018:825), while establishing reliability becomes a problem the moment coders working independently of each other are unable to re-code the data and classify categories membership in the same way (cf. Bolognesi, Pilgram & van den Heerik 2017:1988).

scholars have resolved these problems by making use of blended reading (e.g., Lemke *et al.*, 2015; Hammond *et al.* 2016; Loukissas 2016). Kaefer *et al.* (2015:1) suggest that a similar approach be used when they refer to “[a] multi-level approach to QDAS-assisted analysis” because it enables scholars to achieve both “closeness for familiarity ... [and] distance for abstraction and synthesis” (Bazeley 2007:8):

[U]sually the researcher closest to the data brings in the awareness of the complexities in the data, while other researchers bring in the abstraction necessarily for synthesis at the final stage of the analysis. However, they do this without being familiar with the coding process and the data. Thus abstraction is imposed on to the data. In this case as the coding and analysis processes are transparent, the focus on familiarity can be switched to abstraction, followed by synthesis as some members will still remain distant from the data, while being fully aware of how the data was coded (Kaefer *et al.* 2015:17).

A significant pitfall of QCA pertains to the tendency that scholars relatively unfamiliar with content analysis might have to over-code the data they have collected, particularly because software makes it so easy to code that data. Marshall (2002:61) summarises this dilemma succinctly when she asks “If ‘there is always something to be found’, then even if one has reached ‘theoretical saturation’ where no new themes emerge, has one finished coding?”. Solutions to over-coding at the analysis stage are to have both “well-defined questions and a clear, step-by-step procedure” in place before the analysis commences (Kaefer *et al.* 2015:17).

7.5 Beyond traditional databases

...the sheer volume and variety of [primary research material] make it difficult to access through the traditional approaches. – Dmytro Karamshuk, Frances Shaw, Julie Brown and Nishanth Sastry (2017:33)

In conclusion, and with respect to data management and storage, software tools such as ATLAS.ti, NVivo, MAXQDA, and QDA Miner are all examples of software tools that can manage and store relatively large amounts of data in, for example, text-, audio-, and video-based formats. As far as analysis is concerned, “while qualitative data analysis software ... will not do the analysis for the researcher, it can make the analytical process more flexible, transparent and ultimately more trustworthy.” (Kaefer *et al.* 2015:1).

In this chapter as well as in the previous one, we have argued that appropriate data size in a big data world depends very much on (a) whether researchers regard data as big only if it conforms to all four Vs or on (b) whether, as is the case for traditional humanists and

social scientists, the argument is that volume is not the only attribute that makes big data “big”. Nevertheless, we cannot deny that QDAS is unable to handle and store big data as defined by researchers in category (a), since it “pushes at the limits of traditional databases as tables of rows and columns, and requires new ways of querying and leveraging data for analysis” (Amoore & Piotukh 2015:4). It is for this reason that we take a closer and far more technical look at how the big data ecosystem is geared towards managing huge datasets in the next chapter.

The nitty-gritty: Big data infrastructure

Infrastructure is the cornerstone of Big Data architecture. – Eileen McNulty (2014:1)

Thus far, we have interrogated big data from a number of perspectives, but without exploring the nitty-gritty of big data – the nuts-and-bolts of this phenomenon, as it were. Its emergence reflects rapid advancements in the development of new technologies often coupled with “overlapping technology waves” (Demchenko, Grosso, Laat & Membery 2013:1), and in this environment, it is vital that humanists and social scientists interested in examining huge amounts of data familiarise themselves with the various big data architectural components which make up the big data ecosystem.

8.1 Managing the unmanageable

Computer systems have become a vital part of the modern research environment, supporting all aspects of the research lifecycle. – Savas Parastatidis (2009:165)

As discussed in previous chapters, big data refers to a collection of data that is growing at such a rapid rate that traditional database technologies cannot accommodate the collection, hence the need for newer types of technology (Provost & Fawcett 2013; Vaisman & Zimányi 2014). Big data therefore differs from small data and traditional analytics. As we have already noted, big data is generally associated with three features or Vs, namely, volume, variety, and velocity (Storey & Song 2017). These three features are used in the literature to highlight the fact that big data involves more than just massive amounts of data that are generated, stored, and analysed. Big data also takes on many different disparate and incompatible data formats, which include structured, semi-structured or unstructured data sources and which are created in (near) real-time (Davenport, Barth & Bean 2012; Jagadish, Gehrke, Labrinidid, Papakonstantinou, Patel, Ramakrishnan & Shahabi 2014; Wamba, Akter, Edwards, Chopin & Gnanzou 2015).⁸⁶ A major challenge inherent in big data research is that datasets need to be cleaned and processed before they can be integrated into a big data

⁸⁶ Near real-time (or NRT as it is also referred to) describes data integration that takes place on a regular basis. Thus, data, may be collected on a daily basis or every few minutes or hours: “[t]he time taken between when data arrives and is processed is very small, close to real time” (Chandio, Tziritas & Xu 2015:9). Real-time data, on the other hand, refers to data that “arrives and is processed in a continuous manner, which enables real-time analysis” (Chandio *et al.* 2015:9).

system. In response to this challenge, IBM coined the fourth factor called *veracity*, to address the element of uncertainty that arises when it comes to data quality. Touched on in Chapter 3, veracity entails managing data quality in an adequate fashion (cf. Buhl, Röglinger, Moser & Heidemann 2013:67), and simply refers to “the truthfulness of the data” (Powers Drette 2016:2).

To use big data as an information asset, innovative data processing technologies are required to generate insights and facilitate decision-making. Scholars unfamiliar with big data may assume that technologies have existed since the dawn of the computer age to store and process massive datasets. What makes big data unique, however, is the nature of the overwhelming data flows which in turn drive the need for fundamental changes to be made to computing architectures and data processing mechanisms. Jim Gray, a data software pioneer, called this driving force *the fourth paradigm* as far back as 2007, and pointed out that this new paradigm, or “data-intensive science” (Bell, Hay & Szalay 2009:1298) constitutes the only way to cope with the management and visualisation of huge datasets.⁸⁷ These datasets also demand the repetitive activity of storing data over a long period of time. Examples include a logger writing millions of visits to a webpage into a weblog, or a cellphone database storing the details of each call from all handsets every 15 seconds (Jacobs 2009).

Not only must a big data system handle large amounts of continuously generated data, but it must also provide a stable and scalable environment for storing, analysing, and mining the given datasets (Hu, Wen, Chua & Li 2014).⁸⁸ This provides an interesting challenge to researchers who rely on traditional data storage systems, since these systems are used to store structured data and are repeatedly queried by a relational database management system (RDBMS).⁸⁹ With the arrival of semi-structured and unstructured data, these systems are stretched far beyond their original system design, since they are now required to handle ad hoc queries, as well as a single batch query which could take several hours to complete. It is widely accepted that traditional RDBMSes and structured query language

87 Jim Gray’s (2007) presentation can be found at http://microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt. In a presentation made to the Computer Science and Telecommunications Board in California on 11 January 2007, Gray remarked that “[t]he techniques and technologies for ... data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration” (<http://itre.cis.upenn.edu/myl/JimGrayOnE-Science.pdf>). Sadly, it is the very last presentation posted by Gray as he was lost at sea on 28 January 2007.

88 Scalability is the ability of a network, software or process to manage huge amounts of data.

89 A relational database management system (RDBMS) is a tool data analysts utilise to create, update, and manage the structured data they collect which is then stored in tables, very much like Excel spreadsheets.

(SQL)⁹⁰ cannot be used in this way owing to their relational architecture and adherence to the ACID⁹¹ (atomicity, consistency, isolation and durability) properties of a RDBMS (Krishnan 2013; Hu *et al.* 2014). Furthermore, traditional RDBMSes are incapable of managing the volume and velocity of the new types of data generated by sensor networks, machine loggers, clickstream analysers and e-commerce platforms, since relational data is stored in pre-defined schemas (Jukić, Sharma, Nestorov & Jukić 2015).

As a result nearly all major Internet companies such as Oracle, IBM, Microsoft, Google, Amazon, Facebook, and Yahoo, to name a few, have been compelled to initiate big data projects. These projects have yielded ground-breaking and emerging fourth paradigm technologies to manage as well as analyse and visualise big datasets. Some of these technologies include, but are not limited to, Apache Hadoop, MapReduce, Hive, and Spark. In subsequent sections, we shed some light on how the fourth paradigm may play a role in the humanities and social sciences by providing an overview of innovative and state-of-the-art big data technologies and their associated applications.

8.2 Big data systems

...big data presents unique systems engineering and architectural challenges.

– Edmon Begoli and James Horey (2012:215).

The architectural layout of big data systems is often complex and consists of several layers which include applications, data tools, information and communications technology (ICT) infrastructure, and service level agreements (SLA). This complexity can be bewildering to researchers who do not have a computer science background. In order to gain a better understanding of these layers, it is sometimes easier to conceptually view big data systems as a culmination of technology, people, data, and processes instead of a complex layered system often depicted in the literature (Kim, Jeong & Kim 2014). In this way, a more traditional value-chain systems-engineering approach can be followed which divides a big data system into several consecutive phases, which include data generation, acquisition, storage, and processing (Hu *et al.* 2014). However to appreciate these phases, the applications, data tools, and ICT infrastructure often associated with big data systems need to be introduced first.

The *application layer*, the first entry point into a big data system, typically employs a programming model to implement various data analyses functions, which include querying,

90 SQL is a language that enables data analysts to interact with relational databases. What this means in practical terms is that a data analyst is able to access, insert, update or delete data to and from a relational database.

91 “... ACID ... is a set of properties that guarantee[s] that transactions are processed reliably” (Miller 2013:145).

statistical analysis or recommendation engines (Zhang, Yang, Chen & Li 2018:72). Potential applications will then tap into the layer to gain insight into and value from the given big datasets. Domains that have invested heavily in big data systems include healthcare (Wang & Hajli 2017; Wang, Kung & Byrd 2018), the public sector (Desousa & Jacob 2017; Maciejewski 2017), the retail sector (Li & Wang 2017), and the financial industry (Seddon & Currie 2018).

In the social sciences, big data systems are currently being utilised in disciplines such as geography (Gao, Li, Li, Janowicz & Zhang 2017), cultural sociology (Bail 2014), political science (Colleoni *et al.* 2014), and history (Graham, Milligan & Weingart 2016). In the context of humanities scholarship, the use of big data systems is becoming popular in education/learning analytics (Kellen, Recktenwald & Burr 2013), musicology (Pugin 2015) and in the digital archiving of literary texts or art (Matusiak, Meng, Barczyk & Shih 2015). (See Chapter 6 in which the applications of big data to social science and humanities scholarship have been discussed in greater detail.)

The next layer is the *computing layer*, which consists of data tools (Hu *et al.* 2014). The toolset includes tools to integrate, manage, and make data accessible for the programming model or application layer. These tools include, among others, state-of-the-art and innovative distributed file systems (cf. Howard *et al.* 1988; Ghemawat, Gobiuff & Leung 2003), NoSQL ('not only SQL') databases (cf. Cattell 2011), MapReduce (cf. Dean & Ghemawat 2008), YARN (cf. Kulkarni & Khandewal 2014) Hadoop (cf. White 2015), and Spark (cf. Reyes-Ortiz, Oneto & Anguita 2015).

With the application and computing layer conceptually defined, an *infrastructure layer* is required with consists of several ICT resources. This layer may be physical infrastructure on-site, or it may use cloud computing that enables virtualisation. What is important to note is that this layer is responsible for data storage, networking, and computation functions. Direct attached storage (DAS), Network attached storage (NAS) and Storage area network (SAN) devices are organised into a network architecture of storage systems. The storage systems can either be disk oriented (DAS), file oriented (NAS) or block oriented (SAN) (Hu *et al.* 2014). With the three layers conceptually defined, a more in-depth discussion is required about the phases in big data systems which include data generation, acquisition, storage, and processing.

8.3 Data generation

You may ... think of [velocity] as the frequency of data generation or the frequency of data delivery. – Philip Russom (2011:7)

Data generation is a highly diverse phase often producing complex datasets generated by distributed data sources. These data sources include databanks, webpages, social media, sensors, and mobile data. Essentially, data is distributed across a number of data sources and

currently, big data sources are businesses or enterprises, computer networks that include the Internet and Internet of Things (IoT)⁹² as well as scientific applications.

The internal data of business or enterprises mainly consists of operational data sources such as human resource data, production data, inventory data, sales data, and so forth. Most of these data sources are structured and historical in nature and focus on capturing day-to-day activities which are managed by RDBMSes (Ponniiah 2010). In order to turn this data into strategic information, IT departments have worked hand-in-hand with the business sector over the past few decades with a view to improving profitability through better decision-making. However, to continuously increase the value of strategic information, real-time analysis is required. For example, the retail corporation Walmart both processes and stores approximately 2.5 petabytes of customer data per hour, which is the equivalent of more than one million transactions every hour (Kitchin 2014a:71, cf. Marks 2016:194). Similarly, Facebook loads more than 60 terabytes of new data and stores more than 15 petabytes of information on a daily basis (Khan, Naqvi, Alam & Rizvi 2015:298). According to the US-based IT research and advisory company Gartner, approximately 8.4 billion devices were connected to the Internet of Things in 2017; it is predicted that this number will increase to 20.4 billion within the next two years⁹³ (cf. Li, Da Xu & Zhao 2018). In terms of scientific data, applications are generating increasingly large datasets that rely on big data analytics for insight creation. For example, the Centre for European Nuclear Research (CERN) recorded storing more than 200 petabytes of data in its tape libraries on 29 June 2017⁹⁴ (Gaillard & Pandolfi 2017:1).

However, to be viewed as a big data generation source, large volumes of data should be generated at a very high velocity. Furthermore, the data formats should include semi-structured or unstructured formats instead of the traditional structured format. Structured data refers to data entities organised and stored in a structured manner such

92 The term *Internet of Things* is not easy to define (Wortmann & Flüchter 2015:221), since it is an evolving paradigm. One definition is that it “is used as an umbrella keyword for covering various aspects related to the extension of the Internet and the Web into the physical realm, by means of the widespread deployment of spatially distributed devices” (Miorandi, Sicari, DePellegrini & Chlamtax 2012:1497). It also been defined as “a paradigm where everyday objects can be equipped with identifying, sensing, networking and processing capabilities that will allow them to communicate with one another and with other devices and services over the Internet to accomplish some objective” (Whitmore, Agarwal & Da Xu 2015:261).

93 <https://www.gartner.com/newsroom/id/3598917>.

94 This data emanated from CERN’s Large Hadron Collider (LHC). The LHC’s computers store approximately 15 petabytes of data annually which, according to Harford (2014:14), translates into “15,000 years’ worth of [someone’s] favourite music”.

as XML documents⁹⁵, or database tables in a relational database system (White 2015). Like structured data, semi-structured data has a schema or structure, but is less organised or structured (Katal, Wazid & Goudar 2013; Kim, Trimi & Chung 2014). A spreadsheet from Microsoft Office or Google Sheets is a good example of a semi-structured source. Examples of unstructured data include social media messages, webpages, photographs, and audio files (Minelli, Chambers & Dhiraj 2013; White 2015).

8.4 Data acquisition

Some data sources ... can produce staggering amounts of raw data. Much of this data is of no interest ... – Alexandros Labrinidis and Hosagrahar Jagadish (2012:2032)

Currently a number of data sources are capable of producing overwhelming amounts of raw data, but the data acquired may be useless for a number of reasons. For example, if a researcher's aim is to continually analyse real-time data, it may be unusable if it is not captured on time. In this regard, real-time processing software remains in its infancy (Cai & Zhu 2015). Other challenges which we have noted in this book are those related to improper data representation (Chen, Mao & Liu 2014:175) and “dirty” data, which is duplicate data, data that contains errors or data that is incomplete/outdated (Liu, Wang, Li & Gao 2017:644). The acquisition of big datasets reflects three phases which are data collection, data transmission or transportation, and data pre-processing (Chen, Mao, Zhang & Leung 2014). Three of the most popular approaches for acquiring big data include log files, sensors, and web crawlers (Hu *et al.* 2014). Log files are one of the most widely adopted approaches for collecting big data since they are generated by source systems to record activities taking place on database systems and web servers. Sensor data, on the other hand, is the output of digital devices after detecting an input in the physical environment. These sensors include accelerometers, photo sensors or smart-grid sensors. Sensors are an integral component of the Internet of Things environment (Chen *et al.* 2014). Network data, which includes web pages, is acquired through a web crawler which is often applied in search engines or web caching.

When data is acquired, the raw data is transferred to a data storage facility for processing and analysis. These data storage facilities, commonly referred to as data centres, consist of physical media such as fibre optic cables and an interconnected network that offers high throughput and low latency. Owing to the diverse variety of data sources which

95 Similar to HTML, in that it also contains markup symbols that are employed to describe the content of a file (or page), XML (Extensible Markup Language) reflects rules that enable documents to be encoded in human- and machine-readable formats. (The markup symbols in XML are unlimited, which is not the case when it comes to HTML.)

often contain noise (meaningless information), redundancies, and anomalies, data should be pre-processed to avoid transferring and storing data that cannot be used. Pre-processing techniques of big datasets include data integration, data cleansing, and redundancy elimination (Hu *et al.* 2014). These techniques are well established in the field of data warehousing, and even though they are referred to variously in the literature as extraction, transformation, and loading (ETL) (Santoso 2017:95), all ultimately deal with improving data quality (Kimball & Caserta 2004), which we have already addressed in this book.

8.5 Data storage

Big data storage requirements are complex and thus needs a holistic approach to mitigate its challenges. – Rajeev Agrawal and Christopher Nyamful (2016:1)

Once data is acquired and pre-processed, it must be prepared for storage, analysis, and value extraction by a big data platform. We have already noted that this poses a significant dilemma for any traditional data storage system, since traditional RDBMSes are not capable of handling very large amounts of data (Patel 2017:125). Coupled to this is the challenge inherent in most datasets being either semi-structured or unstructured in nature. For these reasons, a data storage infrastructure should be flexible and reliable, providing a scalable access interface for data processing and queuing (Hu *et al.* 2014).

Storage technologies, such as direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN), are responsible for storing the data collected (Hu *et al.* 2014). However, an additional challenge comes into play since the data stored should also be organised in an efficient way so that it can be processed effectively. Big data management frameworks are known for their ability to facilitate this process and are comprised of three layers, which are distributed file systems, NoSQL databases, and programming models.

8.5.1 Distributed file systems

File systems are the foundation of any computing system and are employed to control how data is stored and retrieved. Considerable research, mainly driven by large Internet companies, has been devoted to improving these systems for the era of big data. Both the Google File System (GFS) and Facebook’s Haystack are well-known examples of distributed file systems developed in an attempt to meet data processing needs. The GFS was the result of the so-called “Big Files” effort by Google co-founders Larry Page and Sergey Bin (Lydia & Swarup 2015:391), and was designed for system-to-system interaction and not for user-to-system interaction (Gemayel 2016:67). The GFS utilises inexpensive community servers, commonly known as computer clusters, to provide a scalable distributed file system for large distributed data-intensive applications (Ghemawat *et al.* 2003). The GFS is thus regarded

as a popular distributed file system (Lee, Lee, Choi, Chung & Moon 2011), with many GFS clusters currently deployed world-wide. Released for the first time in 2010, a new version of the Google File System (the GFS2) is code named Colossus, and promises to provide a considerable performance advantage over the first version. Facebook, by contrast, designed Haystack to store and process the massive amounts of photographs in their big data application (Beaver, Kumar, Li, Sobel & Vajgel 2010). The number of photographs uploaded to Facebook is staggering; at present about 14.58 million images are uploaded each hour, which translates into 350 million every day (Aslam 2018:3).

8.5.2 NoSQL databases

NoSQL (or ‘no SQL’) databases together with distributed file systems have become the de facto standard to store and manage big datasets (Cattell 2011). NoSQL databases can be roughly grouped into four types, namely, key-value stores, column-oriented stores, document databases, and graph databases. Each of these types of NoSQL databases organise data in a different data model.

A key-value NoSQL database, for instance, has a simple data model where data is stored as a key-value pair and each key is unique (Hu *et al.* 2014). Well-known examples of this type of database include Voldemort (developed by LinkedIn.com) and Dynamo, used by Amazon’s e-business platform to obtain data-driven recommendations from major data (Provost & Fawcett 2013; Chen *et al.* 2014).

Column-oriented databases store and process data by columns rather than by rows as is the case in relational databases and was inspired by Google’s Big Table (Hu *et al.* 2014). Popular examples of these kinds of databases include Cassandra (Laksham & Malik 2010; Haseeb & Puttun 2017) and HBase (Perkins, Redmond & Wilson 2018). Cassandra was developed by Anish Lakshman (one of the authors of Amazon’s Dynamo) and Prashant Malik at Facebook to power the Facebook inbox search feature and was open-sourced in 2008. Cassandra⁹⁶ is a fault tolerant, decentralised database, meaning that it has no single point of failure, and excels at real-time transactions and data analytics. HBase is an open-source clone of Google’s Big Table and leverages the distributed data storage capabilities provided on top of Hadoop and the Hadoop Distributed Files Systems. The capabilities of HBase are elaborated upon when we introduce Hadoop in Section 8.7.

Unlike key-value stores, document databases are able to support more complex data structures because the data is stored as documents and represented in JSON format (Krishnan 2013:86). JSON or JavaScript Object Notation is a recognized data-interchange format and easy for researchers to read and write (Izquierdo & Cabot 2016:52). What makes the JSON format quite appealing for researchers is that it is computer language independent

96 <http://cassandra.apache.org/>.

and thus employed among a wide range of computer programming languages. Standard examples of document stores include MongoDB, Couchbase, CouchDB, as well as Riak (Krishnan 2013:86). According to DB-Engines,⁹⁷ which ranks database management systems, MongoDB is currently the most popular NoSQL document database among users (cf. Columbo & Ferrari 2015:146), the main reason being that the database does not require a predefined schema, allowing for flexible storage strategies when changing and updating documents.

The final type of NoSQL database is called a graph database. In graph theory, a graph is a mathematical structure that consists of nodes (vertices) connected by edges (arcs), while a node is usually represented as a dot and edges by a link in diagrams (Celko 2014:208). Graph databases are known for their ability to represent links and relationships between relevant data nodes using graph theory. However, graph databases are not suitable for computations and aggregations, and are therefore predominately used to store and represent data as a graph (Celko 2014:208). It is also viewed as one of the most complex NoSQL database types and has developed as a consequence of the rapid growth in data from social media data (Krishnan 2013:97). Well-known examples of this type of database include Neo4j (Jordan 2014:11) and OrientDB (Küçükkeçeci & Yazici 2018:35). Neo4j⁹⁸ is a widely-used graph database according to DB-Engines;⁹⁹ it is implemented in Java and uses Cypher Query Language to query the data.¹⁰⁰ All objects in Neo4j are stored as either an edge, node or an attribute.¹⁰¹ Unlike Neo4j, OrientDB¹⁰² is an open-source NoSQL multi-nodal database that combines graph theory with key-value as well as document- and object-oriented models into a single database. In other words, OrientDB is a graph database where every edge and node is a document. This allows for increased functionality and flexibility, making OrientDB a second generation NoSQL database. It is envisaged that multi-nodal databases will soon replace traditional RDBMes as the preferred big data store, due to their ability to accommodate multi-format datasets of very high volume (Assay 2015).

97 <https://db-engines.com/en/ranking>.

98 <http://neo4j.com/>.

99 Ibid.

100 Cypher Query Language is a declarative, SQL-inspired language employed by researchers to visually describe patterns in data using circles for nodes and lines for relationships.

101 Nodes represent entities such as people, accounts, or businesses, and are very much related to a record in a relational database. Edges are the lines that connect the nodes and represent relationships, while attributes constitute the information about the nodes. In 'John likes Baroque', for example, 'John' and 'Baroque' are the nodes, 'like' is the relationship between the two nodes, and 'music' is the attribute or property.

102 <http://orientdb.com/>.

8.5.3 Programming models

As noted earlier, the application layer make uses of the computing layer as the bridge between the application and infrastructure layer. Programming models are known for their ability to link the underlying hardware (infrastructure layer) with the software as well as for their tools designed not only to integrate and manage data, but also to make it accessible (application layer). In other words, a programming model offers NoSQL databases and distributed file systems the functionality for querying and analysing big data sets. The programming model is therefore a critical component in any big data architecture. Prior to the big data era, traditional programming models such as OpenMP (Dagum & Menon 1998) and MPI¹⁰³ (Walker & Dongarra 1996) were used as parallel models. However, these models differ from the model required to implement parallel computations on GFS, for example. What is required is a generic process model to process and store big data sets. Some of the most important programming models include MapReduce, Dryad, Pregel, GraphLab, S4, and Storm (Hu *et al.* 2014). It should be pointed out that a programming model is not a programming language and is designed to be used by programmers, rather than business users. In addition, a programming model exists independently of a programming language.

8.6 Data analysis

The massive amounts of high-dimensional data bring both opportunities and new challenges to data analysis. – Jianqing Fan, Fang Han and Han Liu (2014:293)

Data analysis is arguably the most important stage of any big data value system since its goal is to extract value from the big data sets. This can then be used to improve decision-making in an organisation or to improve organisation inefficiencies. In the context of an academic environment, big data analysis may be employed to predict, for instance, future performance and examine patterns of student performance over time, since large quantities of longitudinal student data can be stored (Daniel 2015). Data visualisation (see Chapter 3), statistical analysis, and data mining are currently successfully employed in several big data applications for these purposes (Hu *et al.* 2014). Some of these applications include text mining, web mining, multimedia analytics, and structured data analytics.

Text mining, which is also referred to as text analytics, describes the techniques exploited to extract meaningful information from unstructured textual data such as emails, blogs, social media content, webpages, online forms, documents or call centre logs (He, Zha & Li 2013). The techniques include computational linguistics, statistical analysis as well as machine learning, and the goal is to identify models, trends, and patterns from textual data that may be useful to the researcher (Sabherwal & Becerra-Fernandez 2011:89-90).

103 Message passing interface.

Some of the major applications of text mining include *information extraction*, *text summarisation*, *question answering (QA)*, and *sentiment analysis* or *opinion mining* (Gandomi & Haider 2015). Information extraction employs information retrieval (IR) systems to extract relevant facts from documents. In other words, it extracts structured data from unstructured text. Text summarisation uses algorithms to produce tags for a document by analysing words, sentences, and phrases in a document. These tags are then employed to product a summary of a document which could be an email, blog, or news article, for example (Kim *et al.* 2014). Question and Answering systems make use of NLP techniques¹⁰⁴ to provide answers to questions that human pose. Examples of these type of systems include Apple's Siri and IBM's Watson. Sentiment analysis is used to determine whether a text, or part of it, is subjective or not and, if subjective, whether it expresses a positive, negative or neutral view (Liu 2015). Sentiment analysis may be performed at the document level (Zhang, Zeng, Li, Wang & Zuo 2009), sentence level (Riloff & Wiebe 2003; Appel, Chiclana, Carter & Fujita 2016) or even at aspect or entity level (Popescu & Etzioni 2007).

Whereas text mining focuses on the process of extracting information from unstructured data, web mining aims to retrieve and extract information from online web data, which could include text, HTML (Hypertext Markup Language), hyperlinks or even multimedia data (Kim *et al.* 2014). Web mining makes use of similar techniques used in text mining such as IR and natural language processing (NLP). Web mining can be categorised into three areas of interest: web content mining, web structure mining, and web usage mining (Kosala & Blockeel 2000). Web content mining is utilised to analyse the content of web pages to discover relationships among documents or to the analyse the text content itself. Web structure mining, by contrast, examines how web documents are structured and determines the hierarchy of the underlying hyperlinks. This is particularly useful in uncovering relationships between a web site and similar web sites. Web usage mining, also known as clickstream analysis, examines web server logs to expose surfers' behaviour and patterns, the ultimate goal being to uncover paths customers follow through a company's website. By analysing these paths, companies are able to identify web pages that are infrequently visited, or reveal broken hyperlinks on web pages (Sabherwal & Becerra-Fernandez 2011:92). Clickstream analyses are widely employed by academic researchers too. In one study, for example, a team of researchers used clickstream analysis to glean insights into which attributes of online social networks tend to attract and retain their participants (Scheider, Feldmann, Krishnamurthy & Willinger 2009). In another useful study, scholars analysed, among other things, the clickstream patterns of tertiary students

104 Natural language processing (NLP) entails giving computers the ability to process human language. NLP techniques allow computers to analyse and understand text, and include lexical acquisition, word sense, disambiguation, part-of-speech (POS) tagging, probabilistic context free grammars, and probabilistic parsing (Manning & Schütze 1999).

majoring in science with a view to developing better video-assisted learning practices for teachers (Giannakos, Chorianopoulos & Chrisochoides 2015).

Multimedia analytics examines multimedia data such as images/photos, videos, and audio. Mining multimedia data is generally seen as highly complex, and requires more computational power than mining numeric or textual data (Kim *et al.* 2014). The main aim of multimedia analytics is to extract interesting knowledge from multimedia and to gain insights into the semantics as captured in the data (Hu *et al.* 2014). Significant multimedia analytics research focuses on multimedia summaries and multimedia annotation, to name just two areas of interest. When it comes to the former, for example, Bian, Yang, Zhang and Chua (2015) have generated multimedia summaries of social events in a given microblog stream. This kind of summarisation is useful for a number of reasons. First, it affords researchers the opportunity to obtain a preliminary overview of the data at their disposal before an exhaustive analysis is made. Second, it “allow[s] for ... targeted access to data, and ... [builds] the basis for visualization techniques” (Blank, Henrich & Kufer 2016:67). With respect to annotation, a team of researchers has shown how user-generated comments (UGCs) on You Tube may be used for what they call “multimedia annotation verification” of videos (Bajaj, Kavidayal, Srivastava, Akhtar & Kamaraguru 2016:53). They have discovered that without accurate verification, You Tube users are unable to successfully retrieve specific videos they may be searching for. What makes this particular study useful is that while researchers have focused on verifying the credibility of textual content generated by online users (Eidenbenz 2012; Nguyen, Yan & Thai 2013; Hocevar, Flanagan & Metzger 2014), few have focused on detecting misleading metadata as it relates to videos (cf. Bajaj *et al.* 2016).

Structured data analytics entails the analysis of large quantities of structured data gathered by business or scientific applications. As mentioned earlier, structured data sources are managed by RDBMSes and data warehouses are often used to integrate all the data sources (Inmon 2005). In terms of taxonomy, data analytics can be achieved according to (1) descriptive, (2) predictive, and (3) prescriptive analytics. Examples of these type of analysis include using regression to find a trend in historical data, or predicting future probabilities and trends, or just address decision-making and improve efficacy. Technologies used to perform data analytics include online analytical transaction processing (OLAP), data visualisation, data mining, and statistical analysis (Ponniiah 2010).

A number of scholars working in the field of education/learning analytics have applied or combined the three different kinds of analytics to study their unique contexts (cf. Daniel 2014). Some scholars have used descriptive analytics to predict and improve students’ success (Dietz-Uhler & Hurn 2013), for instance, while others have employed predictive analytics to track at-risk students in an attempt to pre-empt course failure (Milne, Jeffrey, Suddaby & Higgins 2012; Zacharis 2015). Prescriptive analytics utilises both

descriptive and predictive analytics to “[help] institutions of higher education assess their current situation and make informed choices on alternative ... [courses] based on valid and consistent predictions” (Daniel 2015:915).

8.7 The Hadoop ecosystem

[Hadoop] is a flagship technology which became the center of gravity for an entire ecosystem. – Fullestop¹⁰⁵

During the last few years, the computing infrastructure used to process and store big data has changed significantly. This includes NoSQL databases (Cattell 2011), Hadoop-related databases (White 2015), Spark data processing engines (Zaharia *et al.* 2010), computer clusters, in-memory databases, and massively parallel-processing (MPP) databases,¹⁰⁶ to name a few. Regarded as one of the most widely adopted frameworks in the world and the cornerstone of modern big data systems (Chen *et al.* 2014; Mavridis & Karatza 2017), Hadoop was introduced in 2007 as an open-source implementation of Google’s MapReduce programming model (White 2015), but has grown into a web of projects consisting of several open-source software components.¹⁰⁷ Currently, Hadoop provides scalable, distributed, and parallelised computing on clusters of inexpensive servers for data collection, storage as well as processing (Eckerson 2011; Rahman & Iverson 2015). This complex system which comprises several related projects, tools, and technologies, is often called the *Hadoop ecosystem*, with some tools making the writing task easier or orchestrating more complex tasks.

8.7.1 Hadoop’s core components

According to Hadoop’s website,¹⁰⁸ the “Hadoop project” consists of three main components, which are the Hadoop Distributed File System (HDFS), MapReduce, and YARN (or “Yet Another Resource Negotiator”). The three base components form the basis of a Hadoop ecosystem; however, the ecosystem may consist of many other technologies as well, such as an NoSQL database (HBase), a data warehousing system (Hive), a platform to manipulate the data (Pig), a tool to efficiently transfer bulk data (Sqoop), or a tool for machine learning (Mahout). This set of technologies complements one another and should not be viewed as separate components. Each of the three main components is explored below and this is followed by a discussion of what constitutes a Hadoop software stack.

105 <https://www.fullestop.com/blog/a-peek-into-the-hadoops-ecosystem/>.

106 MPP is a parallel processing hardware option where multiple processes work on different parts of a programme.

107 See section 8.5.3 for an explanation of programming models.

108 <http://hadoop.apache.org/>.

The HDFS

The Hadoop Distributed File System or HDFS was designed to store large amounts of data across multiple nodes of commodity hardware in an HDFS cluster (White 2015). The HDFS cluster usually consists of (1) a single NameNode that manages the file system's metadata, and (2) a collection of DataNodes that store the actual data (Hu *et al.* 2014). Since the HDFS is file-based, it does not require a data model (as in the case of an RDBMS) to save or process the data and can store files in any format. Once a file is uploaded onto the HDFS, the file is divided into blocks. The blocks are then distributed between computers within the HDFS cluster and are duplicated to store multiple copies of each block on multiple computers within the HDFS cluster (White 2015). With this ability, HDFS provides the perfect environment for storing structured, semi-structured and unstructured data, while simultaneously enabling parallel processing via MapReduce applications (Watson 2014).

MapReduce

MapReduce is considered to be a popular data processing engine since it is easy to use, is powerful, and enables the automatic parallelisation of computations on large clusters of commodity servers (White 2015). MapReduce was originally developed by Google for the generation of data for its production web services, for sorting and machine learning, and to scale over large clusters of machines (Dean & Ghemawat 2008). At present, however, MapReduce is employed as a programming model that enables the implementation of applications associated with processing and generating large datasets. This is mainly possible because the programming model uses the distributed file system to store and process the data and consists of machine code for processing and generating large datasets (Sharda, Delen & Turban 2014:587). This allows programmers with limited experience in parallel and distributed systems to develop applications that are automatically parallel and distributed in nature. These applications may be developed using either C, Java, Ruby or Python¹⁰⁹ (Watson 2014; White 2015). A typical MapReduce application consists of two parts, a *Map* phase and a *Reduce* phase. The map phase converts raw data into value-key pairs, while the reduce phases processes the data in parallel using a cluster (Landset, Khoshgoftaar, Richter & Hasanin 2015). The end product of these phases is a file that may either be loaded into a data warehouse or analysed through the use of big data analytic tools such as Tableau¹¹⁰ or Gephi.¹¹¹ Tableau is a commercial data visualisation tool often used by business intelligence

109 C, Java, Ruby, and Python are programming languages which allow researchers to use a set of instructions to produce various kinds of outputs. These instructions will generally receive input and implement a specific algorithm.

110 <https://www.tableau.com/>.

111 <https://gephi.org/>.

professionals to build interactive and shareable dashboards, scoreboards, and reports. Gephi, on the other hand, is an open-source and free data visualisation tool suited for research such as social network analysis.

YARN

YARN, which is Hadoop's cluster resource management system, has been used since Hadoop 2.0 to improve MapReduce implementation performance (White 2015). Prior to the introduction of YARN, Hadoop and MapReduce were tightly coupled and MapReduce was responsible for both cluster resource management as well as data processing. In YARN (regarded as MapReduce 2), MapReduce handles only data processing, while YARN is now responsible for cluster resource management. This division of tasks means that the new ecosystem does not only scale better, but can also accommodate more nodes, thus improving on the original Hadoop 1.0 ecosystem.

8.8 The Hadoop software stack

Choosing the right technologies and tools, such as the right solution stack, is an important part of the architectural challenges we try to solve. – Tom Smith (2016:1)

As mentioned earlier, the Apache Hadoop software library consists of several projects, tools, and technologies often collectively referred to as the *Hadoop ecosystem* illustrated in Figure 8.1 below. To simplify this vast web of projects, tools, and technologies, the structure of an ecosystem is described in this section in terms of its storage, processing, and management layers.

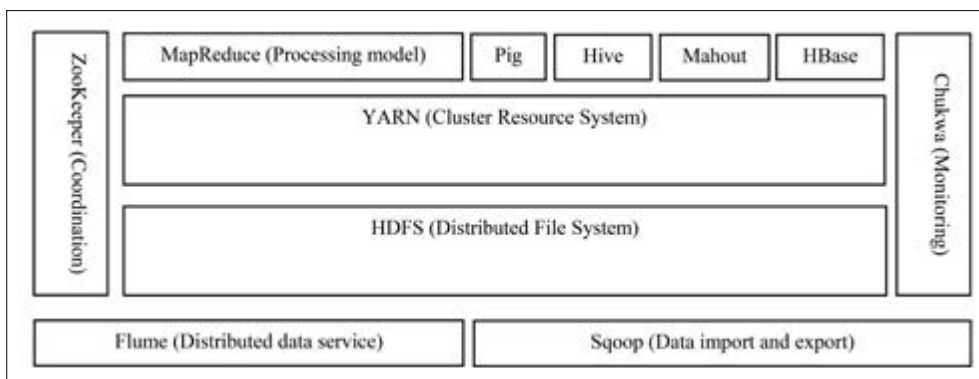


Figure 8.1: The Hadoop ecosystem (adapted from Hu, Wen, Chau & Li 2014:678)

The storage layer is the lowest level of the software stack and includes the HDFS, (described previously) and HBase, a NoSQL database. Both the HDFS and HBase are responsible for data storage. It is important to re-iterate that HDFS is not a database, but a file storage system designed for scalability and fault tolerance. HBase, a column oriented non-relational distributed database runs on top of the HDFS and uses its scalability to provide a distributed storage engine. Records are stored in tables and columns and like a RDBMS, the tables must have a primary key (PK) to retrieve records when queried. The difference is that many attributes are stored as column families (Landset *et al.* 2015). Since HBase is a NoSQL database, it does not support SQL and when any SQL-like command is executed, the command must be translated into a Java-like equivalent (Rahman & Iverson 2015).

The processing layer is where the actual processing and analysis take place, and its foundation is MapReduce and YARN. In addition to these processing frameworks, this layer includes tools for data acquisition such as Flume¹¹² and Sqoop¹¹³ (Basha, Kumar & Babu 2016:126), which were developed to assist with data movement and integration. Flume is a distributed service that collects and aggregates large amounts of data from multiple sources and then loads it to a centralised data store or HDFS. Sqoop, on the other hand, handles the import and export of data between relational databases and Hadoop. For example, Sqoop can send data from a MySQL or Oracle database¹¹⁴ to HDFS, perform a MapReduce task, and send the HDFS MapReduce results back as an import to a relational database (Intel 2013; Celko 2014). Sqoop therefore plays an important role in importing data from a relational database to Hadoop and vice versa. It also enables researchers to consolidate structured data (from different sources) with unstructured data (from NoSQL sources) on a single Hadoop storage system. Other tools include a query engine such as Hive (Story & Song 2017) and a scripting language such as Pig¹¹⁵ (Gates & Dai 2016). Hive was developed by Facebook to bring the concepts of tables, columns and SQL (from the relational database world) to the Hadoop ecosystem. Hive also allows users to organise and partition big data sets into tables and provides HiveQL, an extension of ANSI SQL to write queries. However, one of the drawbacks of MapReduce is that algorithms and code developed in Java, Python, or C, for example, can become very complex, particularly for users not familiar with MapReduce programming. Pig Latin was developed by Yahoo! as a high-level declarative language to offer abstractions and hide the complexities associated with programming MapReduce jobs. Pig supports user-defined functions written in Python, Java, and JavaScript, and translates MapReduce jobs internally to MapReduce tasks, without the programmer having to manage

112 <http://flume.apache.org/>.

113 <http://sqoop.apache.org/>.

114 MySQL and Oracle are examples of relational database management systems (RDBMS).

115 <http://pig.apache.org/>.

the conversation. Pig is therefore ideal for programmers who do not enjoy programming in a higher-level language and are accustomed to developing scripts. In a sense, both Hive and Pig are “SQL-like” languages providing data warehousing capabilities to the Hadoop ecosystem.

The management layer includes tools for high-level organisation such as scheduling, monitoring, and co-ordination. Example of these tools include ZooKeeper and Chukwa which are utilised to monitor and manage distributed applications performed on Hadoop (Hu *et al.* 2014). ZooKeeper was originally developed by Yahoo! to make it easier for applications to access configuration information, but has grown to such a level that it can now co-ordinate and synchronise applications across distributed computer clusters (Warden 2011:10). Chukwa is a Hadoop sub-project that serves as a data collection system to monitor and manage large-scale systems (Krishnan 2013). Chukwa is built on HDFS and the MapReduce programming model, and has flexible and powerful tools for displaying, monitoring and analysing collected data (Chen *et al.* 2014).

Even though most big data systems rely on Hadoop, there are scenarios that require real-time data streaming and processing instead of batch processing. Two stream processing systems include Spark¹¹⁶ and Storm¹¹⁷ (Shoro & Soomro 2015), which make it possible to run real-time, distributed computations on streams of data store and emit the results to Hadoop (White 2015). Spark was originally developed by the University of California, Berkeley on the MapReduce framework (Zaharia *et al.* 2010), but is now an Apache project. Spark is seen as a new generation distributed processing engine and is faster and more flexible than MapReduce (Landset *et al.* 2015). In conjunction with its batch processing option, Spark also offers micro-batching using Spark Streaming. This approach partitions an incoming stream into chunks of data, which can then at a later stage be batch processed (Shahrivari & Jalili 2014). This is, however, not true real-time streaming, but does allow load balancing and also offers integration of stream and batch processing for an online application such as clickstream analysis (Zaharia *et al.* 2010). A typical Apache Spark ecosystem will consist of Spark SQL for structured data, GraphX for graph processing, MLlib for machine learning and Spark Streaming for micro-batching incoming data (Databricks 2016).¹¹⁸ Whereas Spark does not entail true real-time streaming, Storm was developed as an open-source distributed stream processing engine with the purpose of processing data in real-time. This option allows researchers to eliminate sources of latency, as well as deploy real-time analytics and online machine learning. In words, transactions can be processed in seconds and the data can be transferred to a big data system from where real-time predictions can be made using machine learning. Storm is already successfully used by companies such as Groupon, Alibaba, and Twitter Analytics.

116 <https://spark.apache.org/>.

117 <https://storm.apache.org/>.

118 <https://databricks.com/spark/about>.

8.9 An example of a Hadoop big data system

... our traditional systems are not capable enough [to perform] the analytics on ... data which is constantly in motion. – Avita Katal, Mohammad Wazid and Rayan Goudar (2013:404)

Figure 8.2 below provides a good example of what a Hadoop big data system looks like. Specifically, it illustrates a myriad of both structured and unstructured big data sources processed by ETL tools and then stored in HDFS. On the right-hand side of the illustration, batch processing via Hadoop and/or Hive is illustrated. The output is then used for analytics, business intelligence, and data visualisation using tools such as Excel or Tableau. The left-hand side of the visual representation illustrates how streaming data is processed in real-time by Storm and stored in real-time structured databases such as HBase and Cassandra. This processed data can be transferred to Hadoop and/or Hive using Sqoop and used by analytics tools such as Business Objects and Microstrategy.

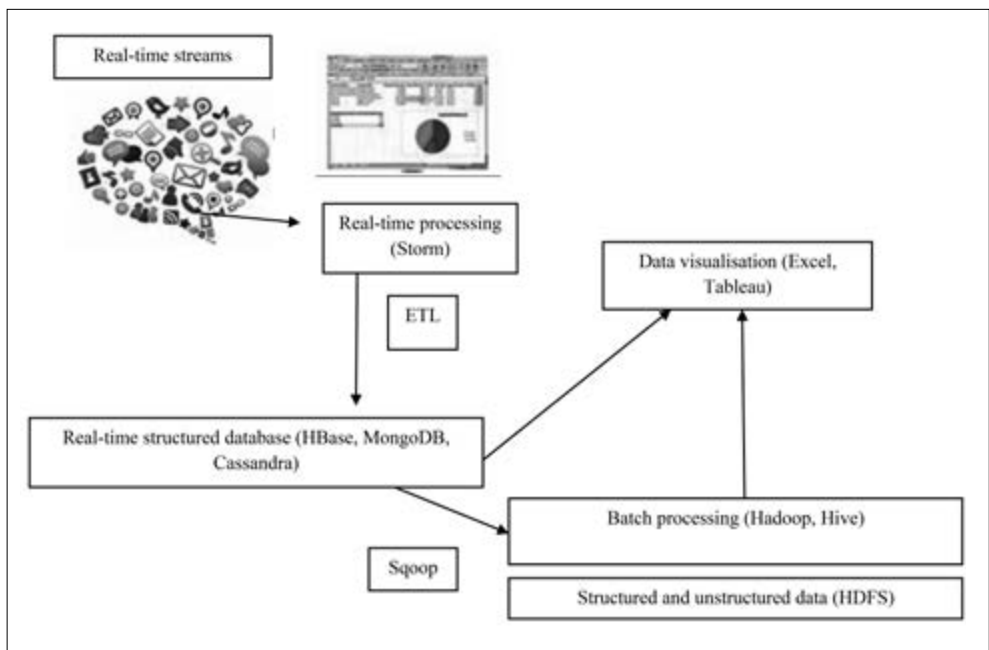


Figure 8.2: A big data system (adapted from Ibarra 2012:3)

8.10 Commercial big data systems and cloud big data

Building scalable, assured big data systems is expensive. – Ian Gorton (2014:6)

Not all big data systems are available as open source software platforms such as Apache Hadoop or Apache Spark. Several of the large traditional database vendors also offer their own big data systems, usually packaged with their proprietary RDBMS software, such as Oracle Big Data¹¹⁹ and IBM DB2 for Big Data.¹²⁰ Another commercial data platform is SAP HANA,¹²¹ an in-memory, column-oriented RDBMS for real-time analytics. However, commercial big data systems are not limited to the database or enterprise resource planning vendors. Quite a number of companies have created commercial big data systems that employ their own versions of Hadoop, recognising that one of the burning issues with regard to an open source platform is that despite being free for anyone to use or modify, it still requires specialist knowledge to set it up. Exploiting this need, several commercial versions have appeared on the marketplace with vendors creating their own versions which are easier to use. Some of the commercial versions include Cloudera Hadoop, Hortonworks Hadoop, EMC Hadoop, Microsoft Hadoop, Intel Hadoop, and MapR (Davenport 2014). It is interesting to note that some of the most important applications that make use of commercial big data systems include targeted marketing, social media analytics, and website recommendation engines (Loshin 2013).

Purchasing and setting up big data infrastructure can be very expensive, but fortunately, alternative solutions are available and such infrastructure can be set up using cloud computing services. Cloud computing services are essentially a new style of delivering applications, data, and resources over the Internet because they allow researchers to rely on a third party (cloud provider) who uses a number of interconnected computers to provide a data service. These data services are currently offered by companies such as Amazon EC2¹²² and Microsoft Azure,¹²³ who charge a user a fee based on storage space and processing time (Gunarathne, Wu, Qiu & Fox 2010). Alternatively, cloud computing infrastructure could be used for the infrastructure layer (instead of a pool of ICT resources) and enabled by virtualisation technology (Hu *et al.* 2014), which refers to the ability to create a virtual rather than actual version of technology, such as virtual computer hardware platforms.¹²⁴

119 <http://www.oracle.com/>.

120 <http://www.ibm.com/>.

121 <https://www.sap.com/products/hana.html>.

122 <http://aws.amazon.com/ec2/>.

123 <http://azure.microsoft.com/>.

124 Virtual machines (VM) are a well-known example of hardware virtualisation.

8.11 A reminder

The mythic power of big data is part of what unifies it as a concept and informs its legibility as a set of tools. – Kate Crawford, Kate Miltner and Mary Gray (2014:1664)

We would not want to create the impression that the above quotation constitutes a rejection of the big data ecosystem. Throughout this book, our message has been clear; it is a call to consider using big data in the social sciences and humanities, given the huge amounts of data available, while at the same time interrogating big data's methods and assumptions, with a view to improving the ways in which research is conducted (cf. Crawford et al. 2014:1665). In Chapters 2 and 5, we considered how big data is framed and noted that the era of big data has “[not] been precipitated by technology alone” (Crawford et al. 2014:1664), although some might argue that the advent of Apache Hadoop “[reinforces] the idea of big data's newness” (Crawford et al. 2014:1663).

What we cannot refute are that current trends in technology such as wearable computers, the Internet of Things, and mobile sensors are all contributing to mountains of “big” heterogeneous data. Turning this data into useful information is vital in our quest to obtain actionable knowledge that can be used to improve the world around us – that is, knowledge that will not only aid us in decision making, but also help us analyse complex problems (cf. Cao 2015:288). Throughout history, we have continued to overcome many hurdles in fulfilling this quest. Now in the year 2019, we are facing a new obstacle, which is essentially too much data coupled with too little wisdom. Big data technologies offer us the ability to improve our collective learning, and provide an architecture that facilitates data collection, storage, processing, and analysis of massive amounts of information. We are also of the opinion that the emergence of cloud computing will become an increasingly important information technology platform (as an alternative to expensive computer clusters). With the promise of faster networks, as can be seen in the design of 5G wireless networks, for example, there is a real possibility of big data becoming mobile. New business models such as Big Data as a Service (BDaaS), or Big Data Analysis as a Service (BDaaS) are exciting possibilities for the social scientist and humanist who would like to embrace big data and technology to promote scientific progress. Some of these possibilities pertain to exploring crowdsourcing (O’Leary 2014; Mulder, Ferguson, Groenewegen, Boersma & Wolbers 2016), big data social networks (Scott 2017; De Nooy, Mrvar & Batagelj 2018), and “smart-cities” (Li, Cao & Yao 2015; Giest 2017), all of which provide useful insights into human activities that could potentially foster more “self-aware” societies.

Leveraging social scientific and humanistic expertise in the world of (big) data science

Humanists have distinct abilities to examine data from multiple angles while interpreting trends and outliers. – James Shulman (2017:1)

Integration of social science into research is crucial. – Ana Viseu (2015:292)

With the vast amounts of data available these days, companies and academic institutions alike are striving to find new ways to exploit it to gain a competitive advantage over their rivals, but this requires deep insights which cannot be generated by machines, but only by humans with rich analytical skills. Previously, institutions would make use of statisticians or modellers to analyse and explore their datasets and this was often done employing manual methods. As computers have become increasingly powerful, more data has been produced and at the same time, more powerful algorithms¹²⁵ have been developed to connect new datasets and enable deeper exploration. This has given rise to a fairly new practice called “data science”,¹²⁶ which involves the extraction of information and insights from complex data in order to detect meaningful patterns (Wang *et al.* 2018:8). Of course, this is a rather simplistic definition which leaves humanists, social scientists, and even data scientists a little confounded.

Before we take a look at the confusion surrounding the term data science, we would like to point out that we have opted to use it as an umbrella term for data analytics, machine learning, and data mining, to name a few. We argue that traditional humanists and social scientists who choose to work with big data may see themselves as conducting data science and/or data analytics. Those who regard themselves as carrying out data science research will be interested in *all* aspects of data science, from data cleansing and preparation to data analysis, employing statistics, programming, mathematics, and the like (cf. Bařkarada & Koronios 2017; Kelleher & Tierney 2018). Researchers who are interested in data analytics, by contrast, will be concerned with generating meaningful insights from their datasets (cf. Concessao 2017); the majority of humanists and social scientists tend to fall into the latter category.

125 Algorithms, which may be defined as sets of instructions for solving tasks, have been in existence for almost 4000 years. The first known algorithm was created around 2000 BCE in Mesopotamia. Classical algorithms include Sumerian-Babylonian root extraction (circa 1700 BCE), the Euclidian algorithm (fourth century BCE), and the approximation of the circle number π (third century BCE) (Brudener 2018).

126 As a scientific term, “data science” came into being approximately 17 years ago (Cao 2016:1).

9.1 What is data science?

What's in a name? – Peter Lake and Robert Drake (2014:104)

In order to understand what *data science* entails, it is useful to deconstruct the term as it is open to different interpretations (cf. Lake & Drake; 2014; Zhu & Xiong 2015) owing mainly to the fact that it is still regarded as an emerging discipline (Rose 2016:3; Donoho 2017:745). According to the *Oxford English Dictionary* (2019), “data is facts and statistics collected together for reference or analysis”, and its origin can be traced back to the mid-seventeenth century when *data* was used as the plural of *datum*, which literally refers to “a piece of information”. *Science* on the other hand is defined as “the intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment” (*Oxford English Dictionary* 2019). We could thus argue that data science refers, at least to some extent, to *the ability to analyse and interpret distinct pieces of information (i.e., data) about the natural and physical world employing a systematic approach that uses both observation and experimentation*. However, this definition does not do justice to the diverse viewpoints adopted by scholars who variously argue (1) that data science is the study of data produced and employed in scientific studies (Dhar 2013), (2) that it involves the study of business data (Provost & Fawcett 2013:56) (3) often with a view to solving scientific and business problems (Svolba 2017), and (4) that it reflects the integration of computing technology, statistics, and artificial intelligence (Dhar 2013) (Zhu & Xiong 2015:2-3).

The term *data scientist* is equally difficult to delineate¹²⁷ because academic research about data scientists and what they do is virtually non-existent (Choi & Tausczik 2017:2). In an earlier chapter, we found it insightful to explore the various metaphors used in the media and in academic environments to define big data with a view to understanding how this phenomenon is framed, and we repeat this exercise with the term data scientist. Usefully, research professor in open and distance learning at the University of South Africa, Paul Prinsloo (2016:348), has summarised a few of the fascinating metaphors used to describe data scientists. In popular media, data scientists are variously labelled as “gods” (Bloor 2012), “high priests” (Dwoskin 2014), “game changers” (Chatfield, Shlemoon, Redublado & Rahman 2014), and even “rock stars” (Sadkowsky 2014). In the current literature on data science and big data, academic scholars are also increasingly referring to data scientists as “unicorns” (Anderson 2015; Harris & Eitel-Porter 2015; Baškarada & Koronis 2017) – individuals who are steeped in the whole range of skills required to

127 While the *Oxford English Dictionary* makes no reference to data science, it does define a data scientist as “a person employed to analyse and interpret complex digital data ...”.

carry out (big) data science research.¹²⁸ Significantly, these scholars are not calling for data scientists to become unicorns, however. Instead, the metaphor is being used to show how such scientists tend to be romanticised in the literature: “[a] perfect data scientist is often described as a ‘unicorn’ because it is impossible for an individual to have all the skills needed. Renowned data scientists have urged their field to make use of more teams because it is so difficult for any individual to gain a complete skillset” (Choi & Tausczik 2017:2). The myth of the data scientist as unicorn is unfortunately driven by industry and position articles in popular media. Articles with titles such as ‘How to become a unicorn data scientist’,¹²⁹ ‘The hunt for unicorn data scientists lifts salaries’ (Press 2015:1), and ‘What’s the secret source to transforming into a unicorn in data science?’ (Kesari 2018:1) are becoming quite prevalent on the Internet. To be fair, there are just as many articles online that are challenging the existence of such a mythical creature, and in the academic environment, researchers are also questioning the feasibility of being able to train data scientists who are au fait with all areas of data science (van der Aalst 2016; Grant 2017; Ohri 2017; Stadelmann, Stockinger, Bürki & Braschler 2018).

While it is not problematic to define what specific scientists such as political, social or climate scientists do given that they are the products of easily identifiable disciplines, this is not the case when it comes to data scientists, since they come from a wide variety of different fields such as mathematics, statistics, data analysis, and information engineering (Rose 2016:3). Doug Rose (2016:4) contends that individuals from these fields might appear to be “a better fit for the title ‘data scientist’ than others”, yet “[i]f [researchers have] worked with numbers and [know] a little about data, [they] could call themselves data scientist[s]”. Irrespective of where they come from, what all data scientists have in common is “[their] focus on the science and not the data” (Rose 2016:4). Interestingly, a common misconception persists that data scientists work only with large datasets and that the terms data science and big data are thus interchangeable (Jagadish 2015:51). Believing that data science equates to big data is a common error in light of the fact that the two are often discussed in the same breath (cf. Rose 2016:6).¹³⁰

Given that there are numerous views when it comes to what data scientists are, not all scholars are in agreement as to what their skills and responsibilities should encompass. In 2014, Vincent Granville distinguished between the “horizontal data scientist” and the “vertical data scientist” in his book *Developing analytic talent*. The former type of data scientist

128 The term “unicorn” appears to have been coined in 2013 by Aileen Lee, founder of the venture capital company *Cowboy Ventures* (Ohri 2017:3).

129 <https://whatsthebigdata.com/2015/10/17/how-to-become-a-unicorn-data-scientist-and-make-more-than-240000/>.

130 As Jagadish (2015:51) observes, using the two terms as though they were synonymous “is not completely inappropriate: [however] the primary difference between the two terms is their perspective: ‘Big Data’ begins with the data characteristics ... whereas ‘Data Science’ begins with data use ...”.

is someone who possesses deep technical knowledge in a narrow field. Thus, for example, computer scientists may be regarded as horizontal data scientists because they are familiar with programming languages, algorithms, data structures, software development methodologies, and the like. In addition, these scientists have cross-disciplinary knowledge (Granville 2014:76) because they are able to blend a number of fields such as computer science, statistics, and machine learning in addition to possessing domain expertise. By contrast, vertical data scientists are individuals with a narrow set of technical skills coupled with very little domain expertise. Such individuals are, according to Granville (2014:78), “fake data scientist[s]”. What is rather interesting is that in Granville’s (2014:5) view, a researcher analysing 10 000 rows of data (as opposed to 10 million, for example) is conducting “fake data science” because the amount of data does not conform to all the Vs identified in the big data arena. However, as we have noted numerous times, big data is relative; as Jagadish (2015:50) so aptly puts it, size is “a moving target” because what counts as big is forever shifting: “What we consider big now is not the same as what we considered big five years ago or what we will consider big five years from now” (Marsden, Shirai & Wilkinson 2018:33). What makes datasets regarded as small by some big data scientists “big” in the humanities and social sciences is that scholars in these disciplines are discovering that they can no longer use traditional methods to analyse them. More importantly, *complex data* rather than big data is the focus for these scholars, since they are often compelled to work with data from multiple sources (cf. Marsden *et al.* 2018:35).

According to well-known computer and data scientist Wil van der Aalst (2014), the main rationale behind data science is to provide answers to a number of generic questions such as “What happened?”, “Why did it happen?”, and “What will happen?” using big data as the new oil. Based on the literature, the essential skills a data scientist should acquire to answer these questions are those in machine learning, data visualisation, advanced data management, (data) storytelling,¹³¹ and problem-formulation/problem-solving skills, while a background in mathematics, statistics and/or computer science accompanied by domain expertise are also important (Patil 2011; Davenport & Patil 2012; Dhar 2013; Harris, Shetterley, Alter & Schnell 2013; Anderson, Bowring, McCauley, Pothering & Starr 2014; Holtz 2014; Mills, Chudoba & Olsen 2016). The typical skillset and areas of knowledge of a data scientist are depicted in the form of a mind map in Figure 9.1, while the mind map in Figure 9.2 highlights many of the tasks required of a data scientist. It should be noted that our focus is on the skills and knowledge of (big) data scientists, rather than on those of big data developers, engineers or business analysts (De Mauro, Greco, Grimaldi & Nobili 2016).

131 In the era of big data, “data explorers have become the narrators of the stories that data is trying to tell. The power of narrative in scientific data lies behind transforming information into knowledge that provides better understanding of complex matters. This enables scientific data explorers to achieve the goal of creating engagement and raising awareness of the message that is being communicated” (Arboleda & Dewan 2017:38; cf. Yoder-Wise & Kowalski 2003:37). Data visualisation is one component of storytelling.

Chapter 9

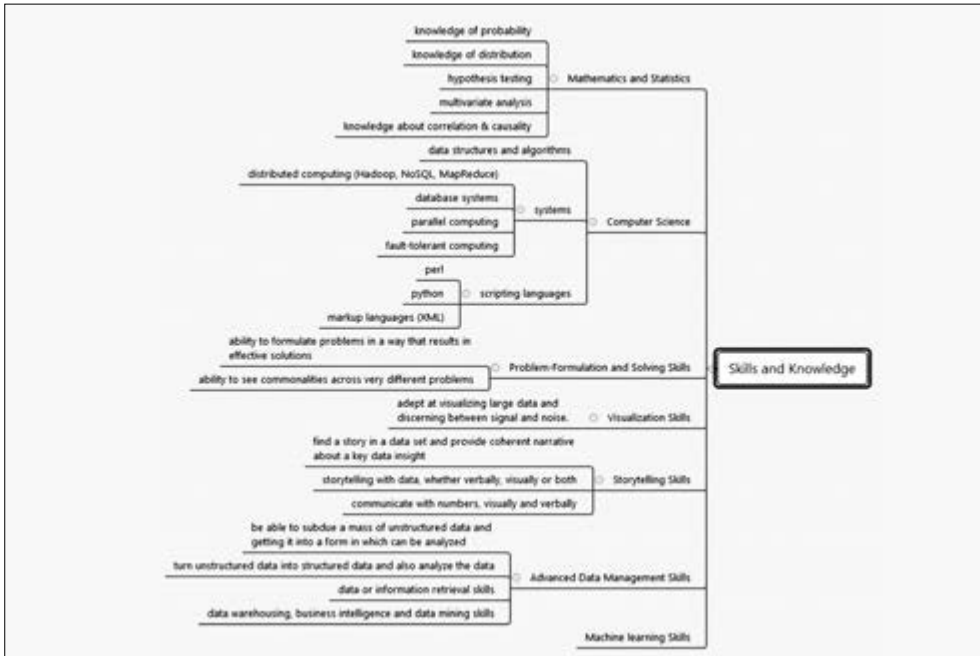


Figure 9.1: The skills and knowledge of a data scientist

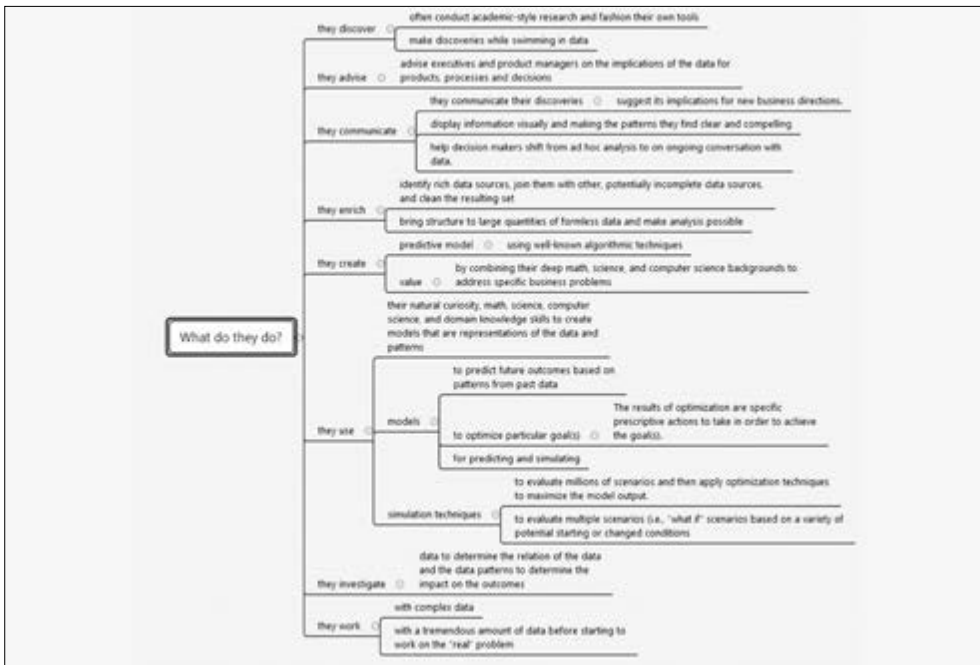


Figure 9.2: The various tasks performed by a data scientist

Ben Stenhaus, a “data science for social good” fellow at Stanford University, is mindful that it is challenging for anyone who has no background in (big) data science to learn about the skills and tasks summarised in Figures 9.1 and 9.2 in any formal way. Indeed, Stenhaus (2017:3) points out that many graduate students at Stanford “[are] struggling to cobble together their own data science education”. The situation is not very different in South Africa since data science is still a practice rather than an established discipline (cf. Nongxa 2017:1). Both in South Africa and abroad, data science degrees are not geared towards helping non-data scientists easily transition into (big) data science, while many resources around data science are simply “too deep [and] too fast” (Stenhaus 2017:3) for the novice data scientist. This does not, however, mean the end of the road for anyone interested in carrying out data science research. Statistician and data miner Meta Brown (2017:2-3) argues that based on her own teaching experiences, “bare minimum requirements” are needed to *begin* conducting (big) data science research, and these requirements include a knowledge of general mathematics (linear algebra), probability and statistics, familiarity with at least one computer language (such as Python, R or SQL), competence in applications such as spreadsheets and word processing, and good communication skills. Brown (2017:3) goes so far as to add that when hotly challenged about this advice by an individual who contended that a Master’s degree in data science is essential, she simply replied, “Swear all you like, brother, I know what I’m talking about”. Academics at UC Berkeley’s Division of Data Sciences would certainly agree with Brown: they offer a course in analysing cultural data through a combination of humanities and machine learning approaches, and students do not need a background in data science or digital humanities to complete the course.¹³² The same division also offers a number of other courses such as ‘Foundations of Data Science’ which reflects no data science, computer science or statistics prerequisites.¹³³ Stanford University offers a ‘Data Science Minor for the Humanities and Statistics’ course that does not require any programming or statistical background and that is aimed at helping students develop data-analytic methods that are directly related to their fields of interest.¹³⁴

A number of scholars suggest that humanities and social science graduates consider a career as a data scientist (cf. Antonijević 2015; Salmon 2017). A good example of an individual who has combined a humanities degree with data science is a computational linguist who makes use of both linguistics and computer science (Jurafsky & Martin 2009) as we noted in Chapter 6. In linguistics, a scholar’s focus may fall on sociolinguistics, dialectology or corpus linguistics, to name a few branches of this discipline. In recent times, these three areas have begun harnessing advanced quantifiable methods to analyse big

132 <http://data.berkeley.edu/making-sense-cultural-data>.

133 <https://data.berkeley.edu/education/courses/data-8>.

134 <https://mcs.stanford.edu/news/new-data-science-minor-humanities-statistics-course>.

datasets, methods which are often associated with natural language processing (NLP) which is a sub-field of computer science. It is well known that text and web mining of social media data are popular data science approaches, often leading to new insights about language use and variation (Liu, 2011, 2012; Liu & Zhang, 2012), for example.

9.2 Marrying (big) data science and the humanities and social sciences

...creating opportunities for bringing social scientific and humanistic expertise into data science practice simultaneously will advance both data science and critical data studies. – Gina Neff, Anissa Tanweer, Brittany Fiore-Gartland & Laura Osburn (2017:85)

A number of researchers have pointed out that it is not sufficient for data scientists to have skills and knowledge in the areas just summarised, but that a more holistic view of data science is called for which acknowledges the human perspective as well. In this respect, Blei and Smyth (2017: 8690) argue that since data science cannot be fully automated, “significant human judgment and deep disciplinary knowledge” are needed when scientific questions are posed. This is echoed by data scientist and machine learning researcher Matthew Mayo (2018:2):

Human involvement, for the foreseeable future, is paramount, not only for overseeing and correcting [the] course for any level of automation, but also to kick off searches for insight. We may be able to automate exploratory investigations of what questions we should be looking to potentially apply the data science process to in the hopes of answering, and even have this phase augmented by facts and figures, but the human element will need to make nuanced decisions on which courses of action are worthy of pursuit.

The so-called human angle cannot be ignored as this is exactly what is required to “understand the context of data, appreciate the responsibilities involved in using private and public data, and clearly communicate what a dataset can and cannot tell us about the world” (Blei & Smyth 2017:8691). Humanists and social scientists are uniquely positioned to meet these requirements and to resist the notion that (big) data is a monolith (cf. Bailey 2016: 169).

In this respect and in the context of so-called (big) “data for good projects” (Neff, Tanweer, Fiore-Gartland & Osburn 2017:85), there is an urgent call to improve research practices in both data science and critical data studies (CDS).¹³⁵ The aim of scholars such

135 Critical data studies (CDS) are studies that interrogate the challenges of big data that pertain to culture, politics, and ethics (Dalton, Taylor & Thatcher 2016). These studies “[question] the many

as boyd and Crawford (2012), Provost and Fawcett (2013), Iliadis and Russo (2016), and Dalton *et al.* (2016), to name a few, is not to label data scientists as being either reflexive or unreflexive about their research, but to identify best practices that would serve these scientists as well as CDS researchers. In their ethnographic research on the culture and practices of data scientists, communication scholars Neff *et al.* (2017:85-86) have observed that many data scientists acknowledge the political, cultural, and ethical challenges inherent in producing and analysing data, although they tend not to be as explicit about these challenges as humanists and social scientists are. They therefore suggest that humanities and social science scholars be positioned in research teams so that “[they] can help data scientists understand the layers of social, organizational, political, ethical, and emotional complexities embedded in their work” (Neff *et al.* 2017:95).¹³⁶ They describe a series of conversations they initiated to bring together social scientists, librarians, science and technology study scholars, and data scientists to talk about aspects of their research such as privacy, transparency, and the democratisation of data science. What they discovered was that the data scientists appreciated the initiative not only because it fostered opportunities for collaborative sense making, but also because it improved their own critiques of data science. With respect to these critiques and in Neff *et al.*'s (2017:85) view, (big) data science studies benefit from the acknowledgement that (1) communication should not be marginalised at any stage in the data science process; (2) sense making is a collaborative effort; (3) data is a starting point and not an end in itself; and (4) data originates from and is exchanged via sets of stories.¹³⁷

In the world of big data, there is also an urgent call for data scientists to adhere to a number of fundamental principles that allow them to extract information and knowledge from complex datasets in principled ways (Provost & Fawcett 2013:56). These principles include but are not limited to (i) extracting useful information from datasets with a view to solving complex problems, (ii) being sensitive to the contexts in which their data-science results will be employed, (iii) using information technology to extract informative data items from a much larger body of data, (iv) paying careful attention to so-called “confounding” or even unseen factors before drawing causal conclusions, and (v) detecting and thus avoiding “overfitting” a dataset – that is, finding effects in a specific sample that cannot be generalised to a larger population (Provost & Fawcett 2013:560-57).

assumptions about Big Data that permeate contemporary literature on information and society by locating instances where Big Data may be naively taken to denote objective and transparent informational entities” (Iliadis & Russo 2016:1).

136 Kitchin & Lauriault (2014:1) refer to these crucial aspects of data science as the sociotechnical “data assemblages” of big data.

137 Neff *et al.* (2017:94) refer to the field of engineering to illustrate what they mean by sets of stories: “the stories that data can tell begin with the stories that shape the production of data or the stories that help make sense of the potential desired outcome and need for data”.

An example: Big data analysis in the humanities in South Africa

... [Wh]at questions are we asking of our big data sets, and what data are we using? The answers are important, and point to the need for a humanist's touch in big data projects. – Alex Woodie (2015:1)

10.1 Introduction

This chapter uses big data methods discussed throughout this book to illustrate the value of using big data to answer questions posed in the humanities. We follow the data analytics project life cycle, which reflects the following stages:

1. Identifying the problem
2. Gathering the data
3. Pre-processing the data
4. Performing analytics on the data
5. Visualising the results

The following sections provide an illustrative example of a typical data analytics project life cycle using sentiment analysis.

10.2 Identifying the problem

The Afrikaner was placed in the spotlight in January 2018 when Hoërskool Overvaal in Gauteng was accused of racism after refusing to accept pupils who were not Afrikaans speaking (Mitchley 2018). Large-scale protests by the Economic Freedom Fighters (EFF) and the African National Congress (ANC) followed, but eventually the North Gauteng High Court turned down the appeal of Gauteng Education MEC¹³⁸, Panyaza Lesufi, to have 55 English-speaking pupils admitted to the school (Masinga 2018).

Another major event took place on 28 February of that year when Parliament undertook to re-consider Article 25 of the South African Constitution to allow for the expropriation of land without compensation. The discourse centred on white/black identities, the 'haves' versus the 'have-nots', the privileged as opposed to the underprivileged, the land

138 Member of the executive council.

thieves against the dispossessed, and the legacy of apartheid – the Afrikaner’s dominance of South African politics from 1948 to 1994 (Eloff 2017; Chung 2018; Mkokeli 2018; Osborne 2018; Roelf, 2018). Julius Malema for instance remarked: “We must ensure that we restore the dignity of our people without compensating the criminals who stole our land” (Chung 2018). On 4 December 2018, Parliament voted in favour of the decision to consider the possibility of amending the constitution. Eventually, ‘expropriation without compensation’ became South Africa’s catchphrase of the year (Grobler 2018; Sekhotho 2018).

Coupled with the land issue, Afrikaners also made headlines through numerous incidents of racism and alleged racism. One of the episodes was the sentencing of Vicky Momberg for *crimen injuria* on 28 March 2018 following her racist tirade against black police officers (Fihlani 2018; Pijoo 2018; Ritchie 2018). After several other incidents, the year ended with a race row at Clifton Fourth Beach in Cape Town (Nombembe 2018; Chambers 2019).

A further issue that put the Afrikaner in the spotlight was farm attacks. In May 2018, the Afrikaner civil rights group, AfriForum, visited the USA to obtain support for what it referred to as “white genocide” and a “fight against land expropriation without compensation” (Thamm 2018:1; cf. Du Toit 2018; Kriel 2018). Journalists such as Lauren Southern and Katie Hopkins made documentary films about farm attacks and the Afrikaner’s position in South Africa. Eventually, this media exposure led to a diplomatic incident between the ANC and Australia after Peter Dutton, Australia’s Home Affairs Minister, declared that white farmers should be welcomed in Australia (Gous 2018; Killalea, 2018), and then between the ANC and the US (Steinhauser 2018) after US President Donald Trump tweeted: “I have asked Secretary of State @SecPompeo to closely study the South Africa land and farm seizures and expropriations and the large scale killing of farmers”.

But how is the Afrikaner generally regarded in post-apartheid South Africa? This is an important question given that the term ‘Afrikaner’ is not a homogenous one (Visagie 2018:5). We argue that social media analytics, and more specifically sentiment analysis, may offer some insights into public sentiment and opinion surrounding the Afrikaner.

10.3 Data gathering

In the literature on Afrikaner identity, the word ‘Afrikaner’ is highly contested. According to Verwey and Quayle (2012:555), “[many] prominent (and often dissident) Afrikaner writers have engaged with the question of whether Afrikaners are African”, for example. In this regard, they point to Breyten Breytenbach, who observed in 1983 that “he both belong[ed]

and [did] not belong to Africa”¹³⁹ (Verwey & Quayle 2015:555) and to Max du Preez, who wrote in 2003 that he was both an African and an Afrikaner.¹⁴⁰

Bearing in mind that the word ‘Afrikaner’ is fraught with contradictions, we collected tweets from Twitter that mentioned the word ‘Afrikaner’ from 6 November 2018 to 6 February 2019. Out of 21379 users, a total of 56565 tweets were amassed, including 23270 unique tweets. These tweets were generated at an average of 734.6 tweets per day by an average of 277.9 users per day.

The first challenge was to filter out irrelevant tweets. What is problematic is that the word ‘Afrikaner’ is used to denote different groups in different languages. In Afrikaans, English and the African languages, the term refers to “a South African person whose family was originally Dutch and whose first language is Afrikaans” (the *Cambridge Dictionary* 2019). The South African trade union Solidarity (2018:5) defines the Afrikaner in clearer terms as “mense wat hulle ras as wit en hulle taal as Afrikaans aangee” [“people who define themselves as white and have Afrikaans as a first language”].¹⁴¹ Although not stated explicitly in racial terms, the *Cambridge Dictionary’s* reference to “a South African person whose family was originally Dutch” implies that this definition is similar to that used by Solidarity. We concede that it is difficult to find just one definition of the term. Steyn (2016:2) puts it eloquently when he says that it is a “slippery” act to use the term ‘Afrikaner’ to refer to all white people whose mother tongue is Afrikaans. Steyn (2016: 2) asks, “What, for example, should the test be to determine whiteness? [...] some ‘Afrikaners’” choose to distance themselves from any identification with ‘other’ Afrikaners or with some form of ‘Afrikaner identity’, while other ‘Afrikaners’, in their everyday lives, speak little or no Afrikaans”.

In German, Danish, Swedish, Norwegian, and Polish, on the other hand, the word ‘Afrikaner’ is used to denote “someone of African descent”. To test dictionary definitions against colloquial use, the language of tweets was identified using Google Translation API, and then translated into English using the same API. A total of 77 languages were identified in this manner. Table 10.1 provides some sample tweets with an automatic English translation reflected in the right-hand column. What is also illustrated is that the dictionary definitions remain valid for German, Swedish, Norwegian, Danish, Polish, and Dutch, but that the word Afrikaner carries the same meaning in French as it does in English and Afrikaans.

139 *The true confessions of an Albino terrorist* (Mariner Books).

140 *Pale native: Memories of a renegade reporter* (Zebra Press).

141 <https://solidariteit.co.za/en/who-are-we/>.

Table 10.1: The use of the word ‘Afrikaner’ in some sample languages

| Language | Tweet | English translation |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Swedish | Tre afrikaner häktade för gruppvåldtäkt på Södermalm - Nyheter Idag | Three Africans arrested for gang rape in Södermalm - News Today |
| | En vacker och intelligent vit 22-åring tjej med A i alla ämnen blir gruppvåldtagen och mördad av ett gäng svarta afrikaner 😞😞😞😞😞 Är det detta ni vill? Är det värt det med massinvandring? Dela om du vill se en förändring! 😊 #svpol #migpol #sd2019 #AfS2019 | A beautiful and intelligent white 22-year-old girl with A in all subjects are gang-raped and murdered by a gang of black afrikaner 😞😞😞😞😞 Is that what you want? Is it worth it to mass immigration? Share if you want to see a change! 😊 #svpol #migpol # sd2019 # AfS2019 |
| Norwegian | Hvis du aldri har blitt forbanna på den innvandringspolitikken som føres i landet så har du heller aldri lest en personlig fortelling fra ei 15 år gammel jente om hvordan det føles å voldtas med kniv mot strupen av en afrikaner. | If you've never been pissed off at the immigration policies pursued in the country you have never read a personal story of a 15 years old girl about how it feels to be raped with a knife at the throat of an African. |
| German | Freiburg Hauptschule ... ein Afrikaner schlägt einem Deutschen Schüler das Pausenbrot aus der Hand .. und schlägt Ihm in das Gesicht der Deutsche Schüler sagt Zitat : scheiss Neger... und schlägt zurück ... Ergebnis ein Schulverweis ... er muss such eine neue Schule suchen | Freiburg main school ... an African beats a German student's lunchbox from the hand .. and hits him in the face the German student says Quote: fucking nigger ... and fighting back ... Result an expulsion ... he must search to find a new school |
| | Wenn Gott gewollt hätte,dass Europa ein Ort für Afrikaner sei,hätte er sie weiss gemacht.!!! | If God had willed that Europe is a place for Africans, he would have made them white.!!! |
| Dutch | Congolese(54) meegelift/carrière gemaakt bij linkse partij in Italië, stapt uit, richt “Afrikaanse partij op alleen voor Afrikanen” met 100%racistisch programma, links juicht | Congolese (54) hitched / career made by left-wing party in Italy, get off, dir “African Party for Africans only” 100% racist program, left cheers |
| | Zo begint de #terreur - en met officiele #toestemming! De kleinste Nederlandse kinderen gaan uitgescholden & belaagd worden op sinterklaasfeestjes. Hoe verschilt dat van belaagde #Afrikaner kleintjes op lagerescholen in #ZuidAfrika? - | Thus begins the #terreur - and with official #toestemming! The smallest Dutch children are abused and attacked to Saint Nicholas parties. How is that different from endangered #Afrikaner children in lower schools #ZuidAfrika? - |

| | | |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Polish | Erytrejczyk i dwaj Somalijczycy aresztowani za gwałt zbiorowy w centrum Sztokholmu | Eritrean and two Somalis arrested for gang rape in central Stockholm |
| | Niemcy. Migrant z Afryki subsaharyjskiej przeciął twarz młodej kobiety nożem tylko dlatego, że nie miała papierosa! Kirchheim Teck 19-letnia kobieta w sobotni wieczór o godzinie 23:15 została poważnie oszpecona nożem. | Germany. Migrant from sub-Saharan Africa cut the face of a young woman with a knife because she did not have a cigarette! Kirchheim Teck 19-year-old woman on Saturday night at 23:15 was severely disfigured knife. |
| French | 1/7 des afrikaners vie dans des bidonvilles (500 000) La politique actuelle est anti blanche. Effectivement il y a un certain pouvoir blanc mais cela ne doit pas de résoudre comme cela . C'est pour ça que je prenne une zone de l'Afrique du sur pour les afrikaner et les coloured | 1/7 Afrikaners living in slums (500,000) Current policy is anti white. Sure there are some white power but this does not solve like that. That's why I take an African area on for Afrikaner and colored |
| | ce matin: Deuxième temps ce mardi de notre série sur l'Afrique du Sud. Aujourd'hui Victor Macé de Lépinay nous guide, en compagnie de ses deux invitées, dans le Voortrekker Monument de Pretoria, lieu symbolique de l'histoire afrikaner. | this morning Second time on Tuesday in our series on South Africa. Today Victor Mace Lépinay to guide us with his two guests in the Voortrekker Monument in Pretoria, symbolic place of Afrikaner history. |

Table 10.2 shows the number of tweets collected by language.

Table 10.2: Language distribution of tweets

| Language | Number of tweets | Percentage of tweets | Number of users | Percentage of users |
|-----------|------------------|----------------------|-----------------|---------------------|
| German | 27917 | 49.22% | 9192 | 42.96% |
| English | 16684 | 29.42% | 9106 | 42.56% |
| Spanish | 3622 | 6.39% | 390 | 1.82% |
| Afrikaans | 3017 | 5.32% | 1083 | 5.06% |
| Swedish | 2729 | 4.81% | 1272 | 5.95% |
| Dutch | 705 | 1.24% | 501 | 2.34% |
| Italian | 187 | 0.33% | 150 | 0.70% |
| Frensch | 168 | 0.30% | 81 | 0.38% |
| Danish | 163 | 0.29% | 140 | 0.65% |
| Polish | 119 | 0.21% | 96 | 0.45% |

Given that tweets in English constitute the largest relevant category, and since German tweets refer to a different understanding of ‘Afrikaners’, we focused our analysis on English tweets.

10.4 Text pre-processing

With the introduction of user-generated content on Internet platforms, one of the main bottlenecks in any text analytics system is the handling of text data that is often noisy. Noisy text typically contains spelling errors, *ad hoc* abbreviations, contractions, improper casing, and incorrect punctuation (Dey & Haque 2009). The text should therefore be “cleansed” or pre-processed before analytical operations can be performed on the data. In the context of natural language processing, text pre-processing is an essential step in any text analytics project, since the words and characters identified during pre-processing will be passed on to subsequent processing stages. According to Palmer (2010:9), text [pre-processing] “is the task of converting raw text, which can be described as a sequence of digital bits, into well-defined sequences of linguistically meaningful units”. Twitter data is known to be particularly noisy and contains additional noise such as incomplete or poorly structured sentences, irregular expressions, ill-formed words and terms that do not appear in a dictionary (Jianqiang & Xiaolin 2017). Before the text can be used for analysis, a series of pre-processing steps must be completed to reduce the amount of noise. These pre-processing series of steps include data cleansing, tokenisation, and syntactic parsing (Dey & Haque 2009). During our data cleansing phase, URLs were replaced with a tag `||HTTP_URL||` and targets (e.g. @John) with tag `||AT_USER||`. Usernames (identified by @ sign) and any external links were removed. Elongated words, which are words in which one character is repeated multiple times (e.g., aaaaangry) were shortened. All punctuation marks (such as full stops, commas, question marks, exclamation marks or emoticons) as well as special characters (\$, %, and, #, etc.) were removed. Re-tweets (i.e., tweets that are re-distributed and begin with ‘RT’) were also removed from the corpus. For tokenisation, we split text on white spaces¹⁴², since all emoticons, HTML tags, URLs, re-tweets, and user mentions were removed from the text. Finally, all stop words (i.e., the most common words in a language such as ‘is’ and ‘the’) were removed from the corpus, and the remaining tokens were converted to lowercase. This included the words ‘Afrikaner’ and ‘Afrikaners’ since either of these words could be present in a tweet.

142 White spaces are defined as “the markers that [separate] each word” (<http://javadevnotes.com/java-string-split-space-or-whitespace-examples>). In other words, they are the unused spaces between, for example, paragraphs or graphics. Tokenisation is the “process of breaking a sentence by the white spaces or any kind of special symbols” (Karpurapu & Jololian 2017:210).

10.5 Performing analytics on the data

After the text was pre-processed and available in the required format, data analytics operations were performed. Typically, these data analytics operations are used to discover meaningful nuggets of knowledge from data or information. For the purpose of this illustrative example we make use of sentiment analysis.

10.5.1 Introduction to sentiment analysis

Sentiment analysis is a growing subfield of natural language processing (NLP), and is often combined with text analysis and computational linguistics to identify, extract, and study any subjective data source (Pang & Lee 2004). In the context of big data, these data sources, which are generally referred to as user-generated content (UGC), include blogs, tweets, and web reviews. In order to extract information from UGC about, for example, people's opinions and attitudes about topics or products they purchase, sentiment analysis is regularly employed (Serrano-Guerrero, Olivas, Romero & Herrera-Viedma 2015:18), and entails classifying these opinions/attitudes as positive, negative or neutral. Sentiment analysis can also be used to monitor how global news events affect public opinion. For example, a real-time sentiment analysis model was used in a study by Wang, Can, Kazemzadeh, Bar and Narayanan (2012) to explore public opinion regarding the 2012 US presidential election. In a more recent study, the entire life-cycle of a large hydro project was assessed through examination of public opinion and critique (Jiang, Lin & Qiang 2015). Other real-world applications include We Feel (Milne, Paris, Christensen, Batterham & O'Dea 2015), which continually monitors Twitter for emotional content, as well as StockTwits and The Stock Sonar (Feldman, 2013), which provide traders, investors, and entrepreneurs with stock market advice.

Based on a review of the literature, two main approaches to carrying out a sentiment analysis are lexicon-based and machine learning approaches. The lexicon-based approach is a rule-based approach that relies on an existing sentiment dictionary or lexicon. Specifically, an existing lexicon is employed which should contain the given word and its polarity (e.g., 'awesome' is positive and 'horrible' is negative). When a new sentence is classified, words in the sentence are matched to words in the lexicon, and using pre-defined rules, the values are aggregated into a sentiment score for the sentence. The aggregation of positive or negative values produces the semantic orientation of the sentence (Taboada, 2016). Such an approach has been successfully used to analyse conventional texts such as blogs, forums, and product reviews (Turney 2002; Kim & Hovy 2004). Machine learning, on the other hand, analyses and interprets patterns or structures in data: it "allows the user to feed a computer algorithm an immense amount of data and have the computer analyse and make data-driven

recommendations and decisions based on only the input data”.¹⁴³ Machine learning approaches make use of supervised or unsupervised methods to determine the polarity of texts (in the form of a document, sentence, or phrase). Supervised learning models require that the algorithm learn on a labelled dataset. These labels can be assigned manually by human evaluation or resources that explicitly define ratings (Giachanou & Crestani 2016). Examples of machine learning algorithms that classify texts as positive, negative or neutral include Naive Bayes (NB), Support Vector Machines (SVM), Maximum Entropy (MaxEnt), Random Forests, and Logistic Regression. By contrast, unsupervised learning learns from training data that is not labelled or classified; classification is based on fixed syntactic patterns (Turney 2002) or on a sentiment lexicon (Taboada, Brooke, Tofloski, Voll & Stede 2011).

10.5.2 *Applying sentiment analysis*

The sentiment analysis employed in the current study followed the lexicon-based approach owing to lack of sufficient training data. Additionally, the fact that the lexicon-based approach can be employed across different domains without changing the dictionaries makes it an attractive approach for Twitter sentiment analysis (Taboada *et al.* 2011). Furthermore, the lexicon-based approach has been shown to be particularly useful for analysing conventional texts such as blogs and tweets (Thelwall, Buckley & Paltoglou 2012; Mohammad, Kiritchenko & Zhu 2013).

The sentiment analyser used two existing lexicon-based dictionaries, namely, Bing Liu’s Opinion Lexicon (Hu & Liu 2004), and the National Research Council (NRC)-Canada Hashtag Sentiment Lexicon (Mohammad *et al.* 2013). The Opinion Lexicon consists of 6789 words and is divided into two lists of words; the one list contains positive ($n = 2006$) words and the other negative words ($n = 4783$). The NRC-Canada Hashtag Sentiment Lexicon consists of 54129 unigram words associated with positive and negative sentiment and was generated automatically from tweets with sentiment-word hashtags such as #amazing and #terrible.

A Python 2.7 application was developed to handle the pre-processing and sentiment analysis of each tweet message collected. The sentiment score of each tweet was derived using the polarity scores of each word found in the lexicon dictionaries as mentioned earlier. We first computed a sentiment score by identifying the sentiment words in each of the sentiment lexicons. A semantic orientation score of +1 is assigned to a positive word and a semantic orientation score of -1 is assigned to a negative word. The sentiment score is then calculated as the sum of scores of its sentiment words divided by the number of scored words to produce an average score. The average score was then used in a classification method to classify the

143 <https://www.netapp.com/us/info/what-is-machine-learning-ml.aspx>

polarity of the tweets into either positive, negative or neutral categories. The sentiment function would finally return a polarity score (positive/negative) between -1.0 and $+1.0$.

Special care was also taken with negation and modifiers. Negation refers to the task of converting a positive to a negative (or the reverse) through special words such as ‘never’, ‘no’, and ‘not’. For example, if ‘not good’ was detected in the dataset, the polarity score of the following word (i.e., ‘good’) was reversed. A similar approach was followed with modifiers such as ‘much’, ‘very’, and ‘really’ (as in ‘very happy’, for example). The polarity of the following word (i.e., ‘happy’) was adjusted by a factor of 1.3. This would either increase or decrease the polarity score of the word.

10.5.3 Sentiment analysis results

Three sentiment classifiers were used and compared with one another. The classifiers included our own polarity classifier, AFINN¹⁴⁴ (Nielsen 2011), and Pattern for Python (De Smedt & Daelemans 2012). A threshold of zero was used to classify the tweets into positive, negative or neutral groupings. In other words, if the score was 0, which indicates no sentiment value, the tweet was classified as neutral. A score of $+0.1$ was considered positive, and a score of -0.1 was considered negative. The results of the sentiment analyser in terms of polarities are shown in Table 10.3.

Table 10.3: Results of the sentiment analysis (without re-tweets, n = 4505)

| Method | Lexicon | Negative | Neutral | Positive | Negative proportion (%) |
|----------------|-------------------------------|----------|---------|----------|-------------------------|
| Own classifier | Opinion Lexicon | 1508 | 1796 | 1201 | 33.5% |
| Own classifier | NRC Hashtag Sentiment Lexicon | 2595 | 1130 | 780 | 57.6% |
| AFINN | - | 1705 | 1334 | 1466 | 37.8% |
| Pattern | - | 758 | 2488 | 1259 | 16.8% |

Based on Table 10.3, it is evident that the results vary, but this is not unexpected since a lexicon is dependent on a particular domain. The Opinion Lexicon was extracted from customer reviews (that were not limited to 140 characters), and the Hashtag Sentiment Lexicon was generated from sentiment-word hashtags contained in a tweet. AFINN contains only 2477 words, while the Hashtag Sentiment Lexicon contains 54129 words, which means that it has fewer words with polarity scores. Figure 10.1 provides a breakdown of results obtained using these four classifiers.

¹⁴⁴ AFINN is an affective lexicon developed by Finn Årup Nielsen, who is a senior researcher at the Technical University of Denmark.



Figure 10.1: A comparison of sentiment classifiers

For the purpose of our study, the Hashtag Sentiment Lexicon was used: it contains the largest corpus (32048 positive and 22081 negative words) and was generated automatically from Twitter data which contained sentiment-related hashtags such as #terrible and #amazing.

10.6 Visualising the results

Using the NRC Hashtag Sentiment Lexicon sentiment classifier, we could investigate where users were registered and whether they tweeted about the Afrikaner in a positive or negative light. Figure 10.2 below illustrates the countries where users were located according to sentiment. We show only those countries with more than ten mentions of the word 'Afrikaner'.

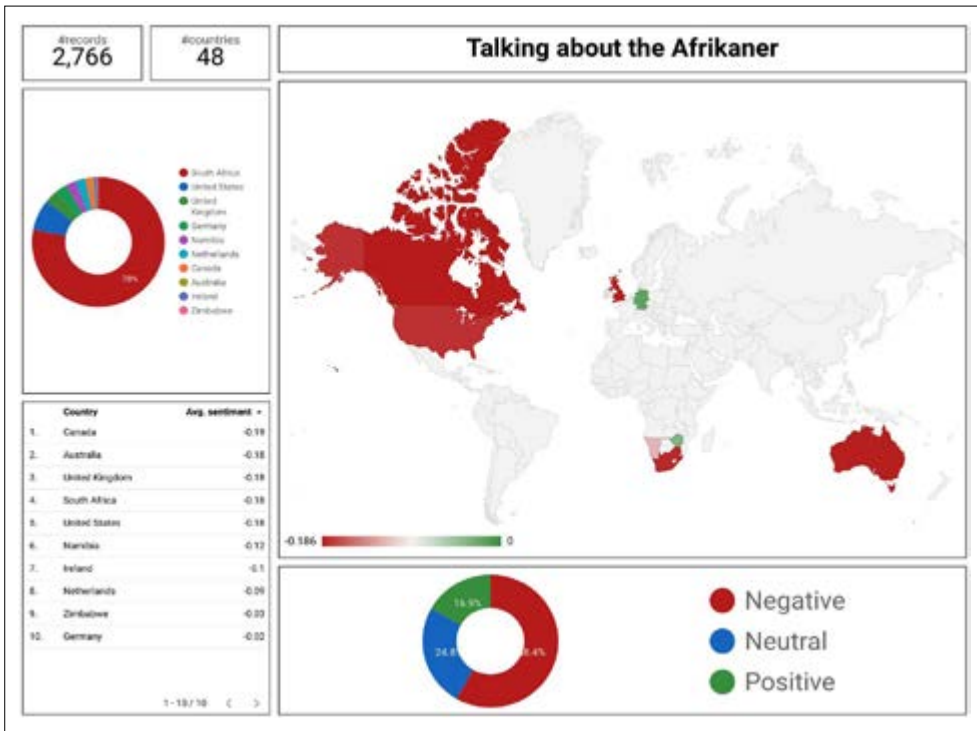


Figure 10.2: Tweets about the Afrikaner by country

Figure 10.2 indicates that out of the 48 countries that referred to the Afrikaner, only ten mentioned the Afrikaner more than ten times. The donut chart in the top left-hand corner shows that 78% of users were based in South Africa, while the country with the second-most users who mentioned the Afrikaner was the United States with a contribution of 7.7% of all tweets collected. While the figure clearly illustrates that discourse about the Afrikaner was concentrated in South Africa, it also shows that users in countries to which a large number of South Africans have emigrated such as the United States, Australia, the United Kingdom, Canada and Ireland also contributed to the discourse. Keep in mind the 2018 incidents relating to the political concerns that the US and Australia had with South Africa over Afrikaner issues mentioned in section 10.2. Countries with strong historical ties to the Afrikaner such as the Netherlands and Germany also contributed significantly to the discourse on the Afrikaner. Lastly, Namibia and Zimbabwe contributed to the discourse on the Afrikaner, which is understandable given these countries' proximity to South Africa. Notably, the rest of Africa, South America, the Middle East, the Far East and Russia did not contribute to this discourse in a significant way. Overall, the figure shows that for Twitter users in neighbouring countries, countries with historical ties to South Africa, and

countries where a large number of Afrikaners have settled since 1994, the Afrikaner was relevant. Nevertheless, almost 80% of the discourse was confined to South Africa, which also illustrates a limited global relevance.

Users in all these countries generally tweeted about the Afrikaner in a negative way: on average and as the donut chart at the bottom shows, approximately 58% of tweets about the Afrikaner were negative, while only 16.9% were positive. We checked the distribution of positive, negative, and neutral tweets per country, and all approximated this distribution. The most negative users were located in Canada, followed by users in Australia and the United Kingdom. Users in Germany and Zimbabwe were the least negative as indicated in green on the world map. The *context* in which these tweets were generated however is significant: we investigated what was said in a given context, and in most cases, the negative tweets referred to *a negative context*. In other words, negative tweets referred to the Afrikaner in a negative context (and not in a negative light) such as in the context of farm attacks, land expropriation or white poverty. In many cases, users were sympathetic to the Afrikaner, but it is the context that made the tweets negative. Figure 10.3 shows a few examples from outside South Africa.

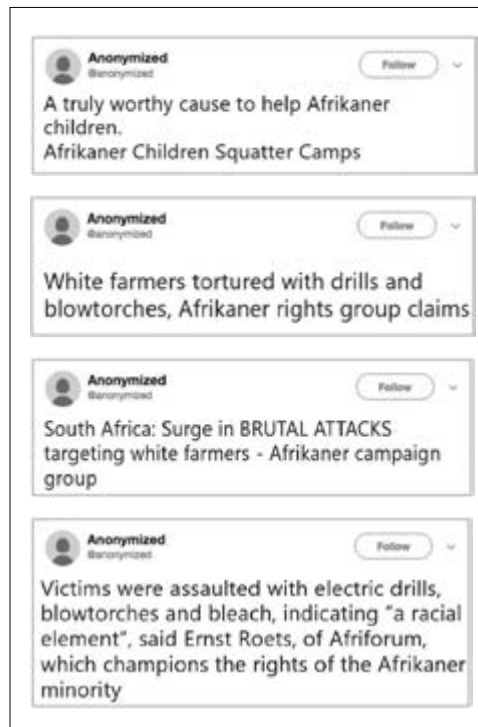


Figure 10.3: Examples of negative tweets

What we can see from the examples in Figure 10.3 is that the issues that brought the Afrikaner to international attention in 2018 contributed to the context in which the Afrikaner was regarded that year. The picture painted of Afrikaners on Twitter from outside South Africa is one in which the Afrikaner is seen as being at risk owing to poverty, economic exclusion, and farm attacks. As an aside, the issue of white poverty is so widely reported on outside South Africa that a Google Image search of the term ‘South African squatter camps’ delivers results such as the one in Figure 10.4. (The screenshot was taken on 20 February 2019 and any facial features have been blurred).

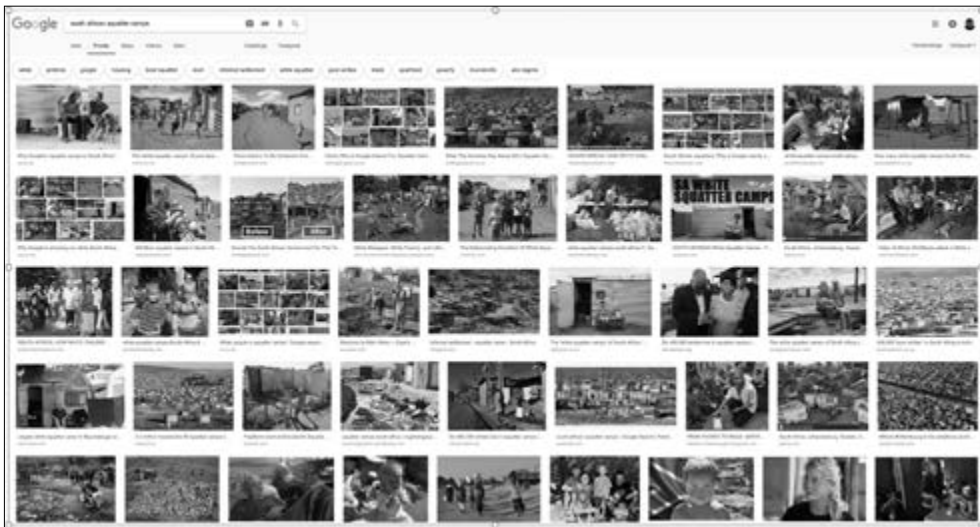


Figure 10.4: Google Image search of ‘South African squatter camps’

What is fascinating – and disturbing – about this Google Image search is that when a Google spokesperson was asked why a search of ‘squatter camps in South Africa’ overwhelmingly yields the kinds of images seen in Figure 10.4, he replied, “Because our systems are surfacing and organising information and content from the web, [a] search can mirror biases or stereotypes that exist on the web and in the real world. [...] We [...] will continue to work, to improve image results for all of our users” (Tembo 2018:1).¹⁴⁵ According to cyber security expert Catalin Cimpanu (2019:1), Google has structured its URL parameters in such a way that “allows threat actors¹⁴⁶ a way to essentially edit search

145 We do not dismiss the existence of white informal dwellings which are well documented in the literature (Sibanda 2012; Kruger 2016). According to the South African Human Rights Commission’s *Equality Report 2017/2018*, 1% of whites in South Africa live in poverty. Approximately 64% of blacks, 41% of coloureds, and 6% of Indians are poverty-stricken. (See https://www.sahrc.org.za/home/21/files/SAHRC%20Equality%20Report%202017_18.pdf).

146 In cyber security, a threat actor is an individual who perpetrates malicious online acts.

results, which is a dangerous issue”. Journalist Lynsey Chutel (2018:2) points out that many South Africans were irate when they encountered the search results for ‘squatter camps in South Africa’, and took to social media to voice their opinions. Significantly, she observes that people’s anger at Google’s search results is misdirected, “since algorithms learn what humans teach them through their [behaviour]”. She adds that “[the] search results are a reflection on a broader conversation on race and poverty” (Chutel 2018:2).

Figure 10.5 below shows examples of negative tweets from *within* South Africa. We found that in South Africa, Afrikaners were depicted as oppressors and as racist and corrupt, while the outside view was one of Afrikaners as victims of negative circumstances (hence the negativity score of tweets).

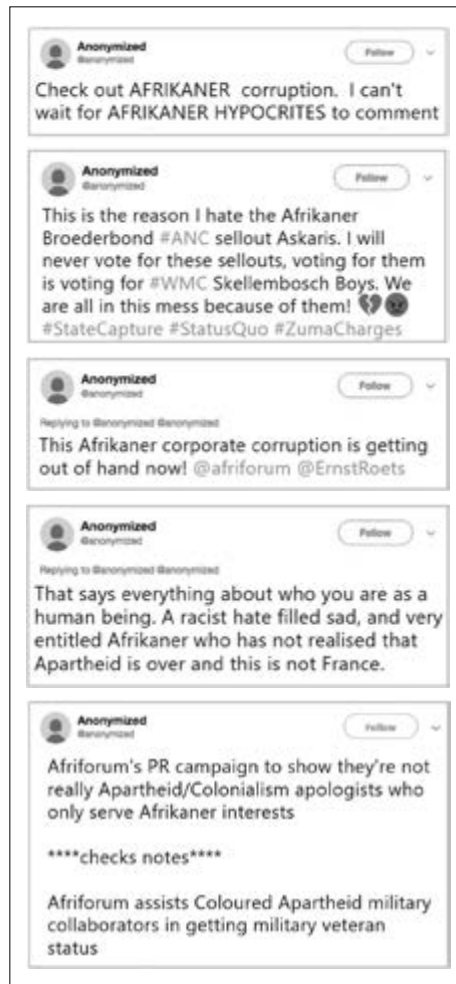


Figure 10.5: Negative tweets from within South Africa

Expressed a little differently, while 58,4% of tweets from across the globe were negative when they referred to the Afrikaner, a significant nuance was detected: tweets from within South Africa portrayed the Afrikaner in a *negative light*, while tweets from outside South Africa depicted the Afrikaner as inhabiting a *negative space*. These findings are noteworthy as they point to what Theunissen (2015:2) refers to as “incongruities and ambiguities” surrounding Afrikaner identity, which is not surprising since this concept is not stable or homogenous (Theunissen 2015:3; cf. Blaser 2012:11). We would like to emphasise that our study merely illustrates how big data may be employed to show how Afrikaners are depicted on Twitter in South Africa and abroad. Quoting a study by Pretorius (2014:21), we recognise that Twitter users both in and outside South Africa may frame or justify the experiences of Afrikaners such as those related to farm attacks as “a Boer genocide” or as “a form of colonial struggle/restitution” that “remains rooted in totalising Afrikaner and black nationalisms respectively”. Referring to a cult of white victimhood, social media expert and journalist Hannelie Booysens (2018:68) argues that social media platforms constitute “the perfect vehicle” for instilling fears about issues such as farm attacks. Booysens (2018:68) observes that in terms of statistics, “farmers are no more likely to be victims of violent crime than any other demographic group” in South Africa.¹⁴⁷ A big data study of *how* Twitter users refer to Afrikaners also needs to include reasons *why* Afrikaners are framed as they are. We argue that humanists who choose to examine how Afrikaners are portrayed on social media platforms need to take into account that “complex cognitive processes and intergroup behaviour [are] at play” (Theunissen 2015:3) in Afrikaner identity. This brings us back to the important message we underscored in Chapter 9 – that we need to understand the social, political, and economic dimensions of any dataset under investigation because the dataset itself cannot help us make inferences about the world (Blei & Smyth 2017:8690).

10.7 Conclusion

While the study is merely an illustration and the dataset too small to generate generalisable conclusions, it nevertheless indicates that big data allows researchers in the humanities to ask new questions, and in novel ways. It shows how a humanities question such as *How is the Afrikaner viewed on Twitter?*, can be answered using big data. The study also signals that researchers need to be cautious about the authenticity of online data, since individuals and organisations may deliberately disseminate misinformation and disinformation, where the former refers to false information designed to deceive online viewers and the latter to false information that is generally geared towards propaganda that promotes a specific agenda or viewpoint (Kumar & Geethakumari 2014:3).

147 This does not mean that we do not acknowledge that farm attacks occur. Indeed, according to the South African Police Service’s crime statistics, the most dangerous province for farm murders is Gauteng (Head 2018). According to Kate Wilkinson (2019:1) of *Africa Check*, two major problems that made it difficult to accurately calculate the farm murder rate in South Africa in 2015/2016 were that (1) the South African Police Service’s definition of what a farm murder entails was vague and (2) a breakdown of the status of victims of farm attacks was not analysed by the South African Police Service.

The last word

While there is no need to educate social scientists and humanists ... in the technical detail and inner workings of ... [big] data tools, it is absolutely critical to introduce to them its capabilities and ability to trigger their imagination, assisting them in asking the “right questions.” – Dominic Lam (2014:4)

It would be easy to discourse at length about the flaws reflected in a field that is not yet fully understood because it finds itself at the early stages of practice and experimentation. Yet, the many successful collaborations between humanists/social scientists and (big) data scientists discussed throughout this book point to the importance of not dismissing big data science unconditionally. Kaplan (2015:1) correctly observes that “[most] of the methods needed to study ... [large] datasets still need to be invented, as they are currently not mastered [either] by humanists or computer scientists”. We could add that social scientists too are feeling their way through big data. We closely examined many of the ontological and epistemological challenges of big data which van Dijck (2014:206) urges humanists and social scientists to address, particularly “[if] predictive analytics and real-time data analytics become the preferred modes of scientific analysis of human behavior”. Big data needs social scientists and humanists. ‘Reinventing the social scientist and humanist in the era of big data’ is not about these scholars being compelled to radically alter what they already do. Given that both the humanities and the social sciences are currently being hybridised with technological and data-driven research, reinvention is more about humanists and social scientists retooling themselves and working in collaboration with data scientists in order to respond to the era of big data. We call on these scholars not only to carry out “big data science for good” projects, but also to critically assess the societal, ethical, and political impacts of big data. We call on them to work closely with data scientists in order to advance a big data environment in which there is greater sensitivity to these impacts. Finally, and without wishing to sound contradictory, we urge humanists and social scientists to consider pursuing a small data research agenda within a big data world: “small data studies will continue to flourish because they have a proven track record of answering specific questions” (Kitchin & Lauriault 2015:464).¹⁴⁸ What is changing and thus opening up small data to big data analytics is the pooling and linking of this data into data infrastructures in order to make data not only accessible and stimulating, but also transparent.

148 In fact, Kitchin and Lauriault (2015:473) state that “[the] pressure to harmonize, share and reuse small data will continue to grow as research funders seek to gain the maximum return on their investment through new knowledge and innovations”.

References

- Abreu A & Acker A. 2013. Context and collection: A research agenda for small data. *iConference 2013 Proceedings*:549-554. <https://dx.doi.org/10.9776/13275>
- Acquisti A & Gross R. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: G Danezis & P Golle (eds). *Privacy enhancing technologies*. Berlin, Heidelberg: Springer. 36-58. <https://doi.org/10.1007/11957454>
- Adamson L & Bakeman R. 1982. Affectivity and reference: Concepts, methods, and techniques in the study of communication development of six- to 18-month-old infants. In: T Field & A Fogel (eds). *Emotion and early interaction*. Hillsdale, New Jersey: Lawrence Erlbaum. 213-236.
- Admin. 2013. What the hell is big data anyway?. *FabCom*, 20 November. Available: <https://www.fabcomlive.com/strategic-marketing-agency/wp-content/uploads/What-The-Hell-Big-Data-White-Paper.pdf> (Accessed 8 May 2018).
- Aggarwal CC. (ed). 2013. *Managing and mining sensor data*. Berlin, Germany: Springer. <https://doi.org/10.1007/978-1-4614-6309-22>
- Agrawal R & Nyamful C. 2016. Challenges of big data storage and management. *Global Journal of Information Technology*, 6(1):1-10. <https://doi.org/10.18844/gjit.v6i1>
- Alvarez W. 2016. *A most improbable journey: A big history of our planet and ourselves*. New York, NY: W.W. Norton & Company. <https://doi.org/10.1016/j.pgeola.2016.10.0088>
- Ambrose ML. 2015. Lessons from the avalanche of numbers: Big data in historical perspective. *ISJLP*, 11(2):201-277.
- Ambrosio C. 2015. Objectivity and representative practices across artistic and scientific visualization. In: A Carusi, AS Hoel, T Webmoor & S Woolgar (eds). *Visualization in the age of computerization*. London, UK: Routledge. 118-144. <https://doi.org/10.4324/97802030669733>
- Amoore L & Piotukh V. 2015. Life beyond big data: Governing with little analytics. *Economy and Society*, 44(3):341-366. <https://doi.org/10.1080/03085147.2015.1043793>
- Anderson C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June. Available: <https://www.wired.com/2008/06/pb-theory/> (Accessed 21 May 2017).
- Anderson C. 2015. *Creating a data-driven organization: Practical advice from the trenches*. Sebastopol, CA: O'Reilly Media, Inc.
- Anderson P, Bowring J, McCauley R, Pothering R & Starr C. 2014. An undergraduate degree in data science: Curriculum and a decade of implementation experience. In: J Dougherty, K Nagel, A Decker, K Eiselt (eds). *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. New York, NY: ACM:145-150. <https://doi.org/10.1145/2538862.25389366>
- Ang CK, Embi MA & Yunus MM. 2016. Enhancing the quality of the findings of a longitudinal case study: Reviewing trustworthiness via ATLAS.ti. *The Qualitative Report*, 21(10):1855-1867.
- Antonijević S. 2016. *Amongst digital humanists: An ethnographic study of digital knowledge production*. New York, NY: Palgrave Macmillan. <https://doi.org/10.1057/9781137484185>

- Appel O, Chiclana F, Carter J & Fujita H. 2016. A hybrid approach to the sentiment analysis problem at sentence level. *Knowledge-Based Systems*, 108:110-124. <https://doi.org/10.1016/j.knosys.2016.05.040>
- Aradau C & Blanke T. 2017. Politics of prediction: Security and the time/space of governmentality in the age of big data. *European Journal of Social Theory*, 20(3):373-391. <https://doi.org/10.1177/1368431016667623>
- Arboleda SA & Dewan A. 2017. Unveiling storytelling and visualization of data. In: R Smedinga, R Biehl & F Kramer (eds). *Proceedings of the 14th SC@RUG 2016-2017*. Groningen, Netherlands: Rijksuniversiteit:38-42.
- Armbrust M, Das T, Davidson A, Ghodsi A, Or A, Rosen J, Stoica I, Wendell P, Xin R & Zaharia M. 2015. Scaling spark in the real world: Performance and usability. In: C Li & V Markl (eds). *Proceedings of the VLDB Endowment*, 8(12):1840-1843. <https://doi.org/10.14778/2824032.2824080>
- Aslam S. 2018. Facebook by the numbers: Stats, demographics & fun facts. *Omnicores*, 1 January. Available: <https://www.omnicoreagency.com/facebook-statistics/> (Accessed 18 May 2018).
- Assay M. 2015. A new breed of database hopes to blend the best of NoSQL and RDBMS. Tech Republic, 21 September. Available: <https://www.techrepublic.com/article/a-new-breed-of-database-hopes-to-blend-the-best-of-nosql-and-rdbms/> (Accessed 6 July 2018).
- Auerbach E. 1953. *Mimesis*. Translated by WR Trask. Princeton: Princeton University Press.
- Austrian GD. 1982. *Herman Hollerith: Forgotten giant of information processing*. USA: Columbia University Press.
- Avgerinou MD & Petterson R. 2011. Toward a cohesive theory of visual literacy. *Journal of Visual Literacy*, 30(2):1-19. <https://doi.org/10.1080/23796529.2011.116746877>
- Baatjes IG. 2005. Neoliberal fatalism and the corporatisation of higher education in South Africa. *Quarterly Review of Education & Training in South Africa*, 12(1):25-33.
- Bacon F. 1853. *Novum organum in The physical and metaphysical works of Lord Bacon, Book I*. London, UK: H.G. Bohn.
- Bail CA. 2014. The cultural environment: Measuring culture with big data. *Theory and Society*, 43(3-4):465-482. <https://doi.org/10.1007/s11186-014-9216-5>
- Bailey M. 2016. Will big data diminish the role of the human in decision making? In: CR Sugimoto, HR Ekbia & M Mattioli (eds). *Big data is not a monolith*. Cambridge, MA: The MIT Press. 164-180.
- Bajaj P, Kavidayal M, Srivastava P, Akhtar MN & Kumaraguru P. 2016. Disinformation in multimedia annotation: Misleading metadata detection on You Tube. In: MF Moens, K Pastra, K Saenko & T Tuytelaars (eds). *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*. New York, NY: ACM:53-61. <https://doi.org/10.1145/2983563.2983569>
- Barnes TJ. 2013. Big data, little history. *Dialogues in Human Geography*, 3(3):297-302. <https://doi.org/10.1177/2043820613514323>
- Barnett M. 2008. Humanitarianism as a scholarly vocation. In: M Barnett & TG Weiss (eds). *Humanitarianism in question: Politics, power, ethics*. USA: Cornell University Press. 235-263. <https://doi.org/10.7591/9780801461538-012>
- Barton S. 2018. Big data and humanitarianism. *Innovation Enterprise*, 26 May. Available: <https://channels.theinnovationenterprise.com/articles/127-big-data-and-humanitarianism> (Accessed 29 May 2018).
- Basha SJ, Kumar PA & Babu SG. 2016. Storage and processing speed for knowledge from enhanced cloud computing with Hadoop frame work: A survey. *IJSRSET*, 2(2):126-132.

References

- Baškarada S. & Koronios A. 2017. Unicorn data scientist: The rarest of breeds. *Program*, 51(1):65-74. <https://doi.org/10.1108/PROG-07-2016-0053>
- Batrinca B & Treleaven PC. 2015. Social media analytics: A survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89-116. <https://doi.org/10.1007/s00146-014-0549-4>
- Beaver D, Kumar S, Li HC, Sobel J & Vajgel P. 2010. Finding a needle in Haystack: Facebook's photo storage. *OSDI*, 10(2010):1-8.
- Bechmann A. 2014. Non-informed consent cultures: Privacy policies and app contracts on Facebook. *Journal of Media Business Studies* 11(1): 21-38. <https://doi.org/10.1080/16522354.2014.11073574>
- Beer D. 2016a. How should we do the history of big data? *Big Data & Society*: 1-10. Available: <http://journals.sagepub.com/doi/pdf/10.1177/2053951716646135> (Accessed 17 April 2018). <https://doi.org/10.1177/2053951716646135>
- Beer D. 2016b. *Metric power*. London, UK: Palgrave Macmillan. <https://doi.org/10.1057/978-1-137-55649-3>
- Begoli E. & J. Horey. 2012. Design principles for effective knowledge discovery from big data. In *Software architecture (WICSA) and European conference on software architecture (ECSA), 2012 Joint working IEEE/IFIP conference*. Helsinki, Finland: IEEE. 215-218. <https://doi.org/10.1109/WICSA-ECSA.212.32>
- Bell G, Hay T & Szalay A. 2009. Beyond the data deluge. *Science*, 323(5919):1297-1298. <https://doi.org/10.1126/science.1170411>
- Bettencourt LMA, Lobo J, Helbing D, Kühnert C & West GB. 2007. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301-7306.
- Beverly B. 2018. Capta: The data of conscious experience. *InformationWeek*, 8 August. Available: <https://www.informationweek.com/big-data/big-data-analytics/capta-the-data-of-conscious-experience/a/d-id/282625?> (Accessed 8 May 2018).
- Bian J, Yang H, Zhang H & Chua TS. 2015. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216-228. <https://doi.org/10.1109/TMM.2014.2384912>
- Biederman I. 1981. On the semantics of a glance at a scene. In: M Kubovy and J Pomerantz (eds). *Perceptual organisation*. Hillsdale: Lawrence Erlbaum Associates. 213-253. <https://doi.org/10.4324/9781315512372-8>
- Black D. 2010. Big data: Dealing with the data tsunami. *SQLstream*, June. Available: <http://sqlstream.com/2010/06/big-data-dealing-with-the-data-tsunami/> (Accessed 7 December 2017).
- Blank D, Henrich A & Kufer S. 2016. Using summaries to search and visualize distributed resources addressing spatial and multimedia Features. *Datenbank-Spektrum*, 16(1):67-76. <https://doi.org/10.1007/s13222-015-0210-5>
- Blaser T. 2012. 'I don't know what I am': The end of Afrikaner nationalism in post-apartheid South Africa. *Transformation: Critical Perspectives on Southern Africa*, 80(1):1-21. <https://doi.org/10.1353/trn.2012.0048>
- Blei DM & Smyth P. 2017. Science and data science. *PNAS*, 114(33):8689-8692. <https://doi.org/10.1073/pnas.1702076114>
- Bloor R. 2012. Are the data scientists future CEOs?. *Inside Analysis*, 12 December. Available: <https://insideanalysis.com/2012/12/are-the-data-scientists-future-ceos/> (Accessed 9 September 2018).
- Boehnert J. 2016. Data visualisation does political things. *DRS2016: Design + research + society: Future-focused thinking*. Brighton, UK: Design Research Society. <https://doi.org/10.21606/drs.2016.387>

- Boellstorff T. 2013. Making big data, in theory. *First Monday*, 18(10):1-17. Available: <http://ojphi.org/ojs/index.php/fm/article/view/4869/3750> (Accessed 20 March 2018). <https://doi.org/10.5210/fm.v18i10.4869>
- Bohn RE, Short JE & Baru C. 2011. How much information? 2010 report on enterprise server information. *Global Industry Center*. University of California, San Diego. Available: <http://clds.sdsc.edu/sites/clds.sdsc.edu/files/pubs/ESI-Report-Jan2011.pdf> (Accessed 8 December 2017).
- Bollier D. 2010. *The promise and peril of big data*. Washington, DC: The Aspen Institute.
- Borenstein J & Arkin R. 2016. Robotic nudges: The ethics of engineering a more socially just human being. *Science and Engineering Ethics*, 22(1):31-46. <https://doi.org/10.1007/s11948-015-9636-2>
- Borgo R, Abdul-Rahman A, Mohamed F, Grant PW, Reppa I, Floridi L & Chen M. 2012. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759-2768. <https://doi.org/10.1109/TVCG.2012.197>
- Borkin MA. 2014. Perception, cognition, and effectiveness of visualizations with applications in science and engineering. Unpublished Doctoral thesis. USA: Harvard University.
- Boullier D. 2016. Big data challenges for the social sciences: From society and opinion to replications. arXiv preprint arXiv:1607.05034. Available: <https://arxiv.org/abs/1607.05034> (Accessed 13 December 2017).
- Bowker GC. 2005. *Memory practices in the sciences*. Cambridge, MA: MIT Press.
- Bowker GC. 2014. The theory/data thing. *International Journal of Communication*, 8(2043):1795-1799.
- Boyd D & Crawford K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662-679. <https://doi.org/10.1080/1369118X.2012.678878>
- Bradley AJ, Mehta H & Collins C. 2016. Visualization, digital humanities, and the problem of instrumentalism. *Workshop on visualization for the digital humanities, IEEE VIS*. 24 October 2016. Baltimore, Maryland, USA: IEEE. 1-4.
- Brandt PT, Freeman JR & Schrodt PA. 2014. Evaluating forecasts of political conflict dynamics. *International Journal of Forecasting*, 30(4):944-962. <https://doi.org/10.1016/j.ijforecast.2014.03.014>
- Brandtzæg P. 2012. Social networking sites: Their users and social implications – a longitudinal study. *Journal of Computer-Mediated Communication*, 17(4): 467-488. <https://doi.org/10.1111/j.1083-6101.2012.01580.x>
- Bravo MJ & Farid H. 2004. Search for a category target in clutter. *Perception*, 33(6):643-652. <https://doi.org/10.1068/p5244>
- Bresciani S & Eppler MJ. 2015. The pitfalls of visual representations: A review and classification of common errors made while designing and interpreting visualizations. *Sage Publications Open*, 5(4):1-14. <https://doi.org/10.1177/2158244015611451>
- Breytenbach B. 1984. *The true confessions of an albino terrorist*. Cape Town, South Africa: Taurus.
- Brimblecombe P. 2017. Early episodes. In: P Brimblecombe (ed). *Air pollution episodes*. London, UK: World Scientific Publishing Europe Ltd. 11-26. <https://doi.org/10.1142/q0098>
- Brittz K. 2018. The bigger picture: What digital humanities can learn from data art. In: A. Du Preez (ed). *Voices from the South: Digital arts and humanities*. Cape Town, South Africa: AOSIS. 177-205. <https://doi.org/10.4102/aosis.2018.BK79.07>

References

- Brokensha SI & Conradie T. 2016. Facilitating critical enquiry about race and racism in a digital environment: Design considerations. *South African Journal of Higher Education*, 30(1):17-41. <https://doi.org/10.20853/30-1-550>
- Brown MS. 2017. You don't need a fancy education to start a data analytics career. *Forbes*, 29 June 2017. Available: <https://www.forbes.com/sites/metabrown/2017/06/29/you-dont-need-a-fancy-education-to-start-a-data-analytics-career/#a06e2e930181> (Accessed 10 September 2018).
- Brown, NM, Mendenhall R, Black ML, van Moer MV, Zerai A & Flynn K. 2016. Mechanized margin to digitized center: Black feminism's contributions to combatting erasure within the digital humanities. *International Journal of Humanities and Arts Computing*, 10(1):110-125. <https://doi.org/10.3366/ijhac.2016.0163>
- Brown W. 2011. Neoliberalized knowledge. *History of the Present*, 1(1):113-129. <https://doi.org/10.5406/historypresent.1.1.0113>
- Brudener H. 2018. Algorithms have been around for 4,000 years. Communications of the ACM, 13 July. Available: <https://cacm.acm.org/blogs/blog-cacm/229543-algorithms-have-been-around-for-4000-years/fulltext> (Accessed 25 July 2018).
- Bryson S, Kenwright D, Cox M, Ellsworth D & Haimes R. 1999. Visually exploring gigabyte data sets in real time. *Communications of the ACM*, 42(8):82-90. <https://doi.org/10.1145/310930.310977>
- Buhl HU, Röglinger M, Moser F & Heidemann J. 2013. Big data: A fashionable topic with(out) sustainable relevance for research and practice? *Business & Information Systems Engineering*, 5(2):65-69. <https://doi.org/10.1007/s12599-013-0249-5>
- Burdick A, Drucker J, Lunenfeld P, Presner T & Schnapp J. 2012. *Digital Humanities*. Cambridge, MA: MIT Press.
- Burkhard R & Eppler M. 2005. Knowledge visualization. In: DG Schwartz (ed). *Encyclopedia of knowledge management*. Hershey, PA: IGI. 551-560. <https://doi.org/10.4018/978-1-59140-573-3.ch072>
- Burke I & van Heerden RP. 2017. Treating personal data like digital pollution. In: M Scanlon & LK Neihn-An (eds). *ECCWS 2017 16th European conference on cyber warfare and security*. Reading, UK: Academic Conferences and Publishing Limited. 82-91.
- Burns R. 2014. Moments of closure in the knowledge politics of digital humanitarianism. *Geoforum*, 53:51-62. <https://doi.org/10.1016/j.geoforum.2014.02.002>
- Cai L & Zhu Y. 2015. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14(2):1-10. <https://doi.org/10.5334/dsj-2015-002>
- Calude CS & Longo G. 2017. The deluge of spurious correlations in big data. *Foundations of Science*, 22(3):595-612. <https://doi.org/10.1007/s10699-016-9489-4>
- Cao L. 2015. *Metasynthetic computing and engineering of complex systems*. London, UK: Springer. <https://doi.org/10.1007/978-1-4471-6551-4>
- Cao L. 2016. Data science and analytics: A new era. *International Journal of Data Science and Analytics*, 1(1):1-2. <https://doi.org/10.1007/s41060-016-0006-1>
- Cardano G. 1953. *The Book on games of chance*. Translated by SH Gould. New York, NY: Princeton University Press. <https://doi.org/10.1017/S0950563600000993>
- Carney M. 2018. Leave no dark corner. ABC. Available: <https://www.abc.net.au/news/2018-09-18/china-social-credit-a-model-citizen-in-a-digital-dictatorship/10200278> (Accessed 31 October 2018).
- Cattell R. 2011. Scalable SQL and NoSQL data stores. *ACM SIGMOD Record*, 39(4):12-27. <https://doi.org/10.1145/1978915.1978919>

- Cawthon N & Moere AV. 2007. The effect of aesthetic on the usability of data visualization. In: E Banissi, RA Burkhard, G Grinstein, U Cvek, M Trutschl, LStuart, TG Wyeld, G Andrienko, J Dykes, M Jern, D Groth & A Ursyn (eds). *Information visualization, 2007. IV'07. 11th international conference*. Los Alamitos, CA: IEEE. 637-648. <https://doi.org/10.1109/IV.2007.147>
- Celko J. 2014. *Complete guide to NoSQL*. Waltham, MA: Morgan Kaufmann. <https://doi.org/10.1016/C2012-0-03536-3>
- Chaka C. 2019. Re-imagining literacies and literacies pedagogy in the context of semio-technologies. *Nordic Journal of Digital Literacy*, 14(1-2):54-69. <https://doi.org/10.18261/issn.1891-943x-2019-01-02-05>
- Chaka C. Forthcoming. Skills, competencies and literacies attributed to 4IR/Industry 4.0: Scoping review. *International Federation of Library Associations and Institutions Journal*.
- Chambers D. 2019. This is what really happened on Clifton Fourth Beach: Security firm boss. *Timeslive*, 5 January. Available: <https://www.timeslive.co.za/news/south-africa/2019-01-05-this-is-what-really-happened-on-clifton-fourth-beach-security-firm-boss/> [Accessed 8 January 2019].
- Chandio AA, Tziritas N & Xu CZ 2015. Big-data processing techniques and their challenges in transport domain. *ZTE Communications*, 1(010):1-21. <https://dx.doi.org/10.3969/j.issn.1673-5188.2015.01.007>
- Chandler R, Anstey E & Ross H. 2015. Listening to voices and visualizing data in qualitative research: Hypermodal dissemination possibilities. *Sage Publications Open*, 5(2):1-8. <https://doi.org/10.1177/2158244015592166>
- Chatfield AT, Shlemoon, VN, Redublado W & Rahman F. 2014. Data scientists as game changers in big data environments. In: W Wang & D. Pauleen (eds). *Proceedings of the 25th Australasian Conference on Information Systems*. Auckland, New Zealand: Australasian Conference on Information Systems:1-11.
- Chen C. 2005. Top 10 unresolved information visualization problems. *IEEE Computer Graphics and Applications*, 25(4):12-16. <https://doi.org/10.1109/MCG.2005.91>
- Chen CP & Zhang Y. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314-347. <https://doi.org/10.1016/j.ins.2014.01.015>
- Chen H & Zhou L. 2017. The myth of big data: Chinese advertising practitioners' perspective. *International Journal of Advertising*: 1-17. <https://doi.org/10.1080/02650487.2017.1340865>
- Chen M, Mao S & Liu Y. 2014. Big data: A survey. *Mobile Networks and Applications*, 19(2):171-209. <https://doi.org/10.1007/s11036-013-0489-0>
- Chen M, Mao S, Zhang Y & Leung VCM. 2014. Related technologies. In: M Chen, S Mao, VCM Leung & Y Zhang (eds). *Big data: Related technologies, challenges and future prospects*. Heidelberg: Springer. 11-18. <https://doi.org/10.1007/978-3-319-06245-7>
- Cheng Q, Li TMH, Kwok CL, Zhu T & Yip PSF. 2017. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. *Journal of Medical Internet Research*, 19(7):e243. <https://doi.org/10.2196/jmir.7276>
- Chen Q, Zobel J, Zhang X & Verspoor K. 2016. Supervised learning for detection of duplicates in genomic sequence databases. *PLoS One*, 11(8):e0159644. Available: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159644> (Accessed 25 April 2018). <https://doi.org/10.1371/journal.pone.0159644>

References

- Chessell M. 2014. Ethics for big data and analytics. Somers: IBM Corporation. Available: http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD&A.pdf (Accessed 11 December 2017).
- Cheyney MJ. 2008. Homebirth as systems-challenging praxis: Knowledge, power, and intimacy in the birthplace. *Qualitative Health Research*, 18(2):254–267. <https://doi.org/10.1177/1049732307312393>
- Choi J & Tausczik Y. 2017. Characteristics of collaboration in the emerging practice of open data analysis. In: CP Lee, S Poltrock, L Barkhuus, M Borges & W Kellog (eds). *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. New York, NY: ACM:835–846. <https://doi.org/10.1145/2998181.2998265>
- Chorley, JC. 2016. Plasma physics computations on emerging hardware architectures. Unpublished Doctoral thesis. USA: Harvard University. UK: Durham University. Available: <http://etheses.dur.ac.uk/11912/> (Accessed: 14 November 2018).
- Chrisomalis S. 2009. The origins and coevolution of literacy and numeracy. In: DR Olson & N Torrance (eds). *The Cambridge handbook of literacy*. New York, NY: Cambridge University Press. 59–74. <https://doi.org/10.1017/CBO9780511609664.005>
- Chu X & Ilyas IF. 2016. Qualitative data cleaning. In: S Chaudhuri & J Haritsa (eds). *Proceedings of the VLDB Endowment*, 9(13):1605–1608. <https://doi.org/10.14778/3007263.3007320>
- Chung F. 2018. ‘The time for reconciliation is over’: South Africa votes to confiscate white-owned land without compensation. *News.com.au*, 28 February. Available: <https://www.news.com.au/finance/economy/world-economy/the-time-for-reconciliation-is-over-south-africa-votes-to-confiscate-whiteowned-without-compensation/news-story/a8a81155995b1adc1c399d3576c4c0bc> (Accessed 8 January 2019).
- Chutel L. 2018. Why googling squatter camps in South Africa returns pictures of white people. *Quartz Africa*, 15 June. Available: <https://qz.com/africa/1306782/why-googling-squatter-camps-in-south-africa-returns-pictures-of-white-people/> (Accessed: 7 February 2019).
- Cimpanu C. 2019. Google search results listings can be manipulated for propaganda. *ZDNet*, 9 January. Available: <https://www.zdnet.com/meet-the-team/us/catalin.cimpanu/> (Accessed 6 March 2019).
- Clare J & Sivil R. 2014. Autonomy lost: The bureaucratisation of South African HE. *South African Journal of Higher Education*, 28(1):60–71. <https://doi.org/10.5840/ijap2014121735>
- Clark D. 2013. When big data goes bad: 6 epic fails. *Blogspot*, 7 November. Available: <http://donaldclarkplanb.blogspot.co.za/2013/11/when-big-data-goes-bad-6-epic-fails.html> (Accessed 8 May 2018).
- Clement T. 2012. Methodologies in the digital humanities for analyzing aural patterns in texts. In: JE Mai, J Furner & P Marty (eds). *Proceedings of the 2012 iConference*. New York, NY: ACM:287–293. <https://doi.org/10.1145/2132176.2132213>
- Cohen S. 2013. Nudging and informed consent. *The American Journal of Bioethics*, 3(6):3–11. <https://doi.org/10.1080/15265161.2013.781704>

- Colleoni E, Rozza A & Arvidsson A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2):317-332. <https://doi.org/10.1111/jcom.12084>
- Colombo P & Ferrari E. 2015. Privacy aware access control for big data: A research roadmap. *Big Data Research*, 2(4):45-154. <https://doi.org/10.1016/j.bdr.2015.08.001>
- Concessao R. 2017. *Big data analytics: Derive insights*. USA: CreateSpace Independent Publishing Platform.
- Convery I & Cox D. 2012. A review of research ethics in Internet-based research. *Practitioner Research in Higher Education*, 6(1):50-57.
- Cooper SB. 2004. *Computability theory*. New York, NY: Chapman and Hall/CRC Press.
- Cope B & Kalantzis M. 2015. Interpreting evidence-of-learning: Educational research in the era of big data. *Open Review of Educational Research*, 2(1):218-239. <https://doi.org/10.1080/23265507.2015.1074870>
- Cope DG. 2014. Computer-assisted qualitative data analysis software. *Oncology Nursing Forum*, 41(3):322-323. <https://doi.org/10.1188/14.ONE322-323>
- Coveney V, Dougherty ER & Highfield RR. Big data need big theory too. *Philosophical Transactions*, 374(2080):1-11. <https://doi.org/10.1098/rsta.2016.0153>
- Cox M & Ellsworth D. 1997. Application-controlled demand paging for out-of-core visualization. In: R Yagel & H Hagen (eds). *Proceedings of the 8th Conference on Visualization '97*. Los Alamitos, CA: IEEE:235-244. <https://doi.org/10.1109/VISUAL.1997.6638888>
- Craik FI. 2014. Effects of distraction on memory and cognition: A commentary. *Frontiers in Psychology*, 5:1-4. <https://doi.org/10.3389/fpsyg.2014.00841>
- Crawford K. 2009. Following you: Disciplines of listening in social media. *Continuum*, 23(4): 525-535. <https://doi.org/10.1080/103043109033003270>
- Crawford K, Miltner K & Gray M. 2014. Critiquing big data: Politics, ethics, epistemology. *International Journal of Communication*, 8:1663-1672.
- Crawford K & Schultz J. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*, 55(1): 93-128. <https://dx.doi.org/sp.2007.54.1.23>
- Creighton JH. 2012. *A first course in probability models and statistical inference*. USA: Springer Science & Business Media. <https://doi.org/10.1007/978-1-4419-8540-8>
- Cresci S, Tesconi M, Cimino A & Dell'Orletta F. 2015. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In: A Gangemi, S Leonardi & A Panconesi (eds). *Proceedings of the 24th International Conference on World Wide Web*. New York, NY: ACM:1195-1200. <https://doi.org/10.1145/2740908.2741722>
- Cukier K & Mayer-Schönberger V. 2013. The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, 92:28-40. <https://doi.org/10.2469/dig.v43.n4.65>
- Dagum L & Menon R. 1998. OpenMP: An industry standard API for shared-memory programming. *IEEE Computational Science & Engineering*, 5(1):46-55. <https://doi.org/10.1109/99.660313>
- Dalton CM, Taylor L & Thatcher J. 2016. Critical data studies: A dialog on data and space. *Big Data & Society*, 3(1):1-9. <https://doi.org/10.1177/2053951716648346>

References

- Dammeier F, Moore JR, Hammer C, Haslinger F & Loew S. 2016. Automatic detection of alpine rockslides in continuous seismic data using hidden Markov models. *Journal of Geophysical Research: Earth Surface*, 121(2):351-371. <https://doi.org/10.1002/2015JF003647>
- Daniel B. 2015. Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, 46(5):904-920. <https://doi.org/10.1111/bjet.12230>
- Darnton R. 2000. An early information society: News and the media in eighteenth-century Paris. *American Historical Association*, 5 January. Available: <https://www.history.org/about-aha-and-membership/aha-history-and-archives/presidential-addresses/robert-darnton> (Accessed 17 April 2018). <https://doi.org/10.2307/26524333>
- Databricks. 2016. Apache Spark ecosystem. Available: <https://databricks.com/spark/about> (Accessed 21 June 2016).
- Daugherty J. & Mentzer N. 2008. Analogical reasoning in the engineering design process and technology education applications. *Journal of Technology Education*, 19(2):7-21.
- Davenport T. 2014. *Big data at work: Dispelling the myths, uncovering the opportunities*. USA: Harvard Business Review Press. <https://doi.org/10.15358/9783800648153>
- Davenport T, Barth H & Bean R. 2012. How 'big data' is different. *MIT Sloan Management Review*, 54:22-24.
- Davenport T & Dyché J. 2013. Big data in big companies. *International Institute for Analytics*, 3:1-31.
- Davenport T & Patil DJ. 2012. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, October. Available: <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century> (Accessed 9 September 2018).
- Davis K. 2012. *Ethics of big data: Balancing risk and innovation*. USA: O'Reilly Media, Inc. <https://doi.org/10.1109/CBI.2015.27>
- Dawson P. 2014. Our anonymous online research participants are not always anonymous: Is this a problem?. *British Journal of Educational Technology*, 45(3): 428-437. <https://doi.org/10.1111/bjet.12144>
- Daylight EG. 2015. Towards a historical notion of 'Turing – The father of computer science'. *History and Philosophy of Logic*, 36(3):205-228. <https://doi.org/10.1080/01445340.2015.1082050>
- Dean, J. 2014. Big data: Accumulation and enclosure. *Theory & Event*, 19(3):1-22.
- Dean J & Ghemawat S. 2008. MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107-113. <https://doi.org/10.1145/1327452.1327492>
- Decock W. 2013. *Theologians and contract law: The moral transformation of the Ius Commune (ca. 1500-1650)*. Leiden: Brill. <https://doi.org/10.1163/15718190-08134P20>
- DeLyser D & Sui D. 2013. Crossing the qualitative-quantitative divide II: Inventive approaches to big data, mobile methods, and rhythm analysis. *Progress in Human Geography*, 37(2):293-305. <https://doi.org/10.1177/0309132512444063>
- De Mauro A, Greco M, Grimaldi M & Nobili G. 2016. Beyond data scientists: A review of big data skills and job families. In: JC Spender, G Schiuma & JR Noennig (eds). *Proceedings of IFKAD 2016 Towards a New Architecture of Knowledge: Big data, Culture and Creativity*. Dresden, Germany: IFKAD:1844-1857.

- Demchenko Y, Grosso P, De Laat C & Membrey P. 2013. Addressing big data issues in scientific data infrastructure. Paper presented at the *2013 International Conference on Collaboration Technologies and Systems (CTS)*. 20-24 May 2013. San Diego, USA: CTS. 48-55. <https://doi.org/10.1109/CTS.2013.6567203>
- De Nooy W, Mrvar A & Batagelj V. 2018. *Exploratory social network analysis with Pajek*. USA: Cambridge University Press. <https://doi.org/10.1017/9781108565691>
- De Smedt TD & Daelemans W. 2012. Pattern for Python. *Journal of Machine Learning Research*, 13:2063-2067.
- Desmet B & Hoste V. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351-6358. <https://doi.org/10.1016/j.eswa.2013.05.050>
- Desouza KC & Jacob B. 2017. Big data in the public sector: Lessons for practitioners and scholars. *Administration & Society*, 49(7):1043-1064. <https://doi.org/10.1177/0095399714555751>
- Desouza KC & Smith KL. 2014. Big data for social innovation. *Stanford Social Innovation Review*: 39-43. Available: <https://communityengagement.uncg.edu/wp-content/uploads/2014/08/Big-Data-for-Social-Innovation.pdf> (Accessed 25 April 2018).
- Devens RM. 1865. *Cyclopaedia of commercial and business anecdotes*. New York, NY: D. Appleton and Company.
- Dey L & Haque S. 2009. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJ DAR)*, 12(3):205-226. <https://doi.org/10.1007/s10032-009-0090-z>
- Dhar V. 2013. Data science and prediction. *Communications of the ACM*, 56(12):64-73. <https://doi.org/10.1145/2500499>
- Diebold FX. 2003. 'Big data' dynamic factor models for macroeconomic measurement and forecasting. In: M Dewatripont, LP Hansen & S Turnovsky (eds). *Advances in economics and econometrics. Eighth world congress of the Econometric Society*. Cambridge: Cambridge University Press. 115-122. <https://doi.org/10.1145/2500499>
- Diebold FX. 2012. On the origin(s) and development of the term "big data". *PIER Working Paper*, 12(037). Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152421 (Accessed 4 December 2017). <https://doi.org/10.2139/ssrn.2152421>
- Dietz-Uhler B & Hurn JE. 2013. Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of Interactive Online Learning*, 12(1):17-26.
- Dinakar K., Reichart R & Lieberman H. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web*, 11(2):11-17.
- Donoho D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745-766. <https://doi.org/10.1080/10618600.2017.1384734>
- Doolittle PE, McNeill AL, Terry KP & Scheer SB. 2005. Multimedia, cognitive load and pedagogy. In: S Mishra & RC Sharma (eds). *Interactive multimedia in education and training*. London, UK: Idea Group, Inc. 184-212. <https://doi.org/10.4018/978-1-59140-393-7.ch010>
- Doorn P. 2014. Big data in the humanities and social sciences. *Science Node*, 5 February. Available: <https://sciencenode.org/feature/big-data-humanities-and-social-sciences.php> (Accessed 8 May 2018).
- Doyle AC. 1915. *The valley of fear*. New York, NY: George H. Doran Company.
- Draucker CB & DS Martsolf. 2008. Storying childhood sexual abuse. *Qualitative Health Research*, 18(8):1034-1048. <https://doi.org/10.1177/1049732308319925>

References

- Dudukovic NM, DuBrow S & Wagner AD. 2009. Attention during memory retrieval enhances future remembering. *Memory & Cognition*, 37(7):953-961. <https://doi.org/10.3758/MC.37.7.953>
- Drucker J. 2011. Humanities approaches to graphical display. *Digital Humanities Quarterly*, 5(1):1-21.
- Drucker J. 2014. *Graphesis: Visual forms of knowledge production*. Cambridge, MA: Harvard University Press.
- Du Preez, A. (ed). 2018. *Voices from the South: Digital arts and humanities*. Cape Town: AOSIS. <https://doi.org/10.4102/aosis.2018.BK79>
- Du Preez M. 2003. *Pale native: Memories of a renegade reporter*. Cape Town, South Africa: Zebra Press.
- Du Toit P. 2018. Men on a mission: AfriForum's Kriel And Roets en route to US to talk about land, crime. *Huffpost*, 2 May. Available: https://www.huffingtonpost.co.za/2018/05/02/men-on-a-mission-afriforums-kriel-and-roets-en-route-to-us-to-talk-about-land-crime_a_23425166/ (Accessed 8 January 2019).
- Dwoskin E. 2014. Big data's high priests of algorithms. *The Wall Street Journal*, 11 August. Available: <https://datascienceguru1.wordpress.com/2014/08/11/big-datas-high-priests-of-algorithms-wall-street-journal/> (Accessed 9 September 2018).
- Eckerson W. 2012. Big data analytics: Profiling the use of analytical platforms in user organisations. *BeyeNetwork*. Available: http://docs.media.bitpipe.com/io_10x/io_103043/item_486870/Big%20Data%20AnalyticsMarkLogic.pdf (Accessed 11 November 2015).
- Efron B. 2001. The statistical century. In: J Panaretos (ed). *Stochastic musings: Perspectives from the pioneers of the late 20th century*. New York, NY: Psychology Press. 29-44. <https://doi.org/10.4324/9781410609120>
- Ellis G & Dix A. 2007. A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1216-1223. <https://doi.org/10.1109/TVCG.2007.70535>
- Eloff T. 2017. Who owns the land?. *Politicsweb*, 2 May Available: <http://www.politicsweb.co.za/opinion/who-owns-the-land> (Accessed 10 April 2018).
- Enslin JA. 2016. Visual interpretations and humanistic interfaces. *MATLIT: Materialidades da Literatura*, 4(2): 279-282. https://doi.org/10.14195/2182-8830_4-2_14
- Eppler MJ & Platts KW. 2009. Visual strategizing: The systematic use of visualization in the strategic-planning process. *Long Range Planning*, 42(1):42-74. <https://doi.org/10.1016/j.lrp.2008.11.005>
- Equality Report 2017/2018. South African Human Rights Commission. Available: https://www.sahrc.org.za/home/21/files/SAHRC%20Equality%20Report%202017_18.pdf (Accessed 5 February 2019).
- Eriksson U, Starrin B & Janson S. 2008. Long-term sickness absence due to burnout: Absentees' experiences. *Qualitative Health Research*, 18(5):620-632. <https://doi.org/10.1177/1049732308316024>
- Evans JSB, Over BT, David E & Handley SJ. 2005. Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review*, 112(4):1040-1052. <https://doi.org/10.1037/0033-295X.112.4.1040>
- Evers JC. 2015. Elaborating on thick analysis: About thoroughness and creativity in qualitative analysis. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 17(1):1-28. Available: <http://www.qualitative-research.net/index.php/fqs/article/view/2369/3924> (Accessed 10 April 2018). <http://dx.doi.org/10.17169/fqs-17.1.2369>

- Evers JC. 2018. Current issues in qualitative data analysis software (QDAS): A user and developer perspective. *The Qualitative Report*, 23(13):61-73.
- Ewenstein B & Whyte JK. 2007. Visual representations as 'artefacts of knowing'. *Building Research & Information*, 35(1):81-89. <https://doi.org/10.1080/09613210600950377>
- Faggin F, Hoff ME, Mazor S & Shima, M. 1996. The history of the 4004. *IEEE Micro*, 16(6):10-20. <https://doi.org/10.1109/40.546561>
- Fairfield J. & Shtein H. 2014. Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics*, 29(1):38-51. <https://doi.org/10.1080/08900523.2014.863126>
- Fan C, Xiao F & Yan C. 2015. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction*, 50:81-90. <https://doi.org/10.1016/j.autcon.2014.12.006>
- Fan J, Han F & Liu H. 2014. Challenges of big data analysis. *National Science Review*, 1(2):293-314. <https://doi.org/10.1093/nsr/nwt032>
- Faniel IM, Kriesberg A & Yakel E. 2012. Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology*, 49(1):1-10. <https://doi.org/10.1002/meet.14504901068>
- Feldman R. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82-89. <https://doi.org/10.1145/2436256.2436274>
- Feldman Z & Sandoval M. 2018. Metric power and the academic self: Neoliberalism, knowledge and resistance in the British university. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, 16(1):214-233. <https://doi.org/10.31269/triplec.v16i1.899>
- Felt M. 2016. Social media and the social sciences: How researchers employ big data analytics. *Big Data & Society*, 3(1):1-15. <https://doi.org/10.1177/2053951716645828>
- Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE & Martone ME. 2014. Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, 17(11):1442-1448. <https://doi.org/10.1038/nn.3838>
- Fernández-Cabana M, Jiménez-Féiz J, Alves-Pérez MT, Mateos R, Gómez-Reino Rodríguez I & García-Caballero A. 2015. Linguistic analysis of suicide notes in Spain. *The European Journal of Psychiatry*, 29(2):145-155. <https://doi.org/10.4321/S0213-61632015000200006>
- Few S. 2006. *Information dashboard design*. Cambridge, MA: O'Reilly Media, Inc.
- Fihlani P. 2018. Vicky Momberg: South African estate agent jailed for racist abuse. *BBC News*, 28 March. Available: <https://www.bbc.com/news/world-africa-43567468> (Accessed 8 January 2019).
- Fineberg D. 2016. Extract, transform, and load big data with Apache Hadoop. *Intel*, 19 February. Available: <https://software.intel.com/en-us/articles/extract-transform-and-load-big-data-with-apache-hadoop> (Accessed 28 May 2018).
- Finlay S. 2014. *Predictive analytics, data mining and big data: Myths, misconceptions and methods*. New York, NY: Palgrave Macmillan. <https://doi.org/10.1057/9781137379283>

References

- Fish S. 2012. Mind your ps and bs: The digital humanities and interpretation. *nytimes.com*, 23 January. Available: <https://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/> (Accessed 4 March 2018).
- Floridi L. 2016. On human dignity as a foundation for the right to privacy. *Philosophy and Technology*, 29(46):307-312. <https://doi.org/10.1007/s13347-016-0220-8>
- Foote K. 2017. A brief history of big data. *Dataversity*, 13 December. Available: <http://www.dataversity.net/brief-history-big-data/> (Accessed 8 May 2018).
- Ford H. 2014. Big data and small: Collaborations between ethnographers and data scientists. *Big Data & Society*: 1-3. Available: <http://journals.sagepub.com/doi/pdf/10.1177/2053951714544337> (Accessed 13 December 2017). <https://doi.org/10.1177/2053951714544337>
- Forsythe A. 2011. The human factors of the conspicuous Babel fish; dyadic referencing through icons. *Journal of Visual Literacy*, 30(3):91-115. <https://dx.doi.org/10.1080/23796529.2011.11674691>
- Forsythe G. 2012. *Annales maximi. The encyclopedia of ancient history*. USA: Blackwell Publishing Ltd. <https://dx.doi.org/10.1002/9781444338386.wbeah08010>
- Forster C. 2017. Humanities graduates should consider data science. *Towards Data Science*, 31 August. Available: <https://towardsdatascience.com/humanities-graduates-should-consider-data-science-d9fc78735b0c> (Accessed 25 July 2018).
- Franks B. 2015. Is big data analytics good or evil?. *Venturabeat*, 15 June. Available: <https://venturebeat.com/2015/06/15/is-big-data-analytics-good-or-evil/> (Accessed 8 May 2018).
- Frické M. 2015. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651-661. <https://doi.org/10.1002/asi.23212>
- Friendenthal J. 2015. Human capital development for big data in South Africa. Available: <https://globalstatement2015.wordpress.com/2015/09/23/big-data-in-south-africa/> (Accessed 12 December 2017).
- Friese S. 2014. *Qualitative data analysis with ATLAS.ti*. Thousand Oaks, CA: Sage Publications. <https://doi.org/10.17583/qre.2016.2120>
- Friese S. 2016. Qualitative data analysis software: The state of the art. *KWALON*, 21(1): 34-45. Available: https://www.tijdschriftkwalon.nl/scripts/shared/artikel_pdf.php?id=KW-21-1-5 (Accessed 5 April 2018).
- Froehlich A. 2017. Will edge computing replace the cloud?. 23 May 2017. *InformationWeek*, 23 May. Available: <https://www.informationweek.com/cloud/will-edge-computing-replace-the-cloud/a/d-id/1328929> (Accessed 6 December 2017).
- Fuller M. 2017. Big data, ethics and religion: New questions from a new science. *Religions*, 8(5):88-97. <https://doi.org/10.3390/rel8050088>
- Fullestop. 2016. A peek into the Hadoop ecosystem. *Fullestop*, 5 April. Available: <https://www.fullestop.com/blog/a-peek-into-the-hadoops-ecosystem/> (Accessed 21 May 2018).
- Furner J. 2016. "Data": The data. In: M Kelly & J Bielby (eds). *Information cultures in the digital age*. Wiesbaden: Springer VS. 287-306. https://doi.org/10.1007/978-3-658-14681-8_17
- Gaillard M & Pandolfi S. 2017. CERN Data Centre passes the 200-petabyte milestone. CERN Document Server, 7 June. Available: <http://cds.cern.ch/record/2276551> (Accessed 15 November 2018).
- Gandomi A & Haider M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137-144. <https://doi.org/10.1016/j.ijinfomgt.2014.10.007>

- Gangadharan S. 2012. Digital inclusion and data profiling. *First Monday*, 17(5):1-11. Available: <http://journals.uic.edu/ojs/index.php/fm/article/view/3821/3199> (Accessed 17 March 2018). <https://doi.org/10.5210/fm.v17i5.3821>
- Gao S, Li L, Li W, Janowicz K & Zhang Y. 2017. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*, 61:172-186. <https://doi.org/10.1016/j.compenvurbsys.2014.02.004>
- Gao X & Tao G. 2016. Ethical challenges in conducting text-based online applied linguistics research. In: P DeCosta (ed). *Ethics in applied linguistic research: Language researcher narratives*. New York, NY: Routledge. 181-194. <https://doi.org/10.4324/9781315816937-11>
- Garreau J. 2006. *Radical evolution: The promise and peril of enhancing our minds, our bodies – and what it means to be human*. USA: Random House Digital, Inc.
- Gates A & Dai D. 2016. *Programming Pig: Dataflow scripting with Hadoop*. USA: O'Reilly Media, Inc.
- Gatto M. 2015. Making research useful: Current challenges and good practices in data visualisation. Reuters Institute for the Study of Journalism (with the support of the University of Oxford's ESRC Impact Acceleration Account in partnership with Nesta and the Alliance for Useful Evidence). Available: <https://www.alliance4usefulevidence.org/assets/Making-Research-Useful-Current-Challenges-and-Good-Practices-in-Data-Visualisation.pdf> (Accessed 1 October 2018).
- Geertz C. 1973. *The interpretation of cultures; selected essays*. New York, NY: Basic Books.
- Gemayel N. 2016. Analyzing Google file system and Hadoop distributed file system. *Research Journal of Information Technology*, 8(3):66-74. <https://doi.org/10.3923/rjit.2016.66.74>
- Ghemawat S, Gobioff H & Leung S. 2003. The Google File System. *ACM Sigops Operating Systems Review*, 37(5):29-43. <https://doi.org/10.1145/1165389.945450>
- Giannakos MN, Chorianopoulos K & Chrisochoides N. 2015. Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *The International Review of Research in Open and Distributed Learning*, 16(1). Available: <http://www.irrodl.org/index.php/irrodl/article/view/1976/3198> (Accessed 18 May 2018). <https://doi.org/10.1109/FIE.2014.7044485>
- Giest S. 2017. Big data analytics for mitigating carbon emissions in smart cities: Opportunities and challenges. *European Planning Studies*, 25(6):941-57. <https://doi.org/10.1080/09654313.2017.1294149>
- Gilbert LS. 2002. Going the distance: 'Closeness' in qualitative data analysis software. *International Journal of Social Research Methodology*, 5(3):215-228. <https://doi.org/10.1080/13645570210146276>
- Glass GV. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):3-8. <https://doi.org/10.2307/1174772>
- Gleser GC, Gottschalk LA & Springer KJ. 1961. An anxiety scale applicable to verbal samples. *Archives of General Psychiatry*, 5(6):593-605. <https://doi.org/10.1001/archpsyc.1961.01710180077009>
- Georgiou M. 2006. Architectural privacy: A topological approach to relational design problems. Unpublished Master's dissertation. London, UK: University College London.
- Godzinski Jr, R. 2005. (En)Framing Heidegger's philosophy of technology. *Essays in Philosophy*, 6(1):1-9.

References

- Goldenberg T, Darbes LA & Stephenson R. 2017. Inter-partner and temporal variations in the perception of sexual risk for HIV. *AIDS and Behavior*: 1-15. Available: <https://link.springer.com/content/pdf/10.1007%2Fs10461-017-1876-5.pdf> (Accessed 4 April 2018). <https://doi.org/10.1007/s10461-017-1876-5>
- Goldstuck A. 2010. Internet access in South Africa 2010. *World Wide Worx*, July 2010. Available: <http://www.worldwideworx.com/wp-content/uploads/2010/07/Exec-Summary-Internet-Access-in-SA-2010.doc> (Accessed 13 December 2017).
- González-Bailón S. 2013. Social science in the era of big data. *Policy & Internet*, 5(2):147-160. <https://doi.org/10.1002/1944-2866.POI328>
- Gorton I. 2014. Software architecture: Trends and new directions. *Software Engineering Institute/Carnegie Mellon University*. Available: <https://pdfs.semanticscholar.org/presentation/b02b/06bb5da8b89dd365a1e0174c03f07135f664.pdf> (Accessed 21 May 2018).
- Gous N. 2018. AfriForum in Australia to talk about farm attacks and murders in SA. *Timeslive*, 14 October. Available: <https://www.timeslive.co.za/news/south-africa/2018-10-14-afriforum-in-australia-to-talk-about-farm-attacks-and-murders-in-sa/> (Accessed 8 January 2019).
- Govani T & Pashley H. 2005. Student awareness of the privacy implications when using Facebook. Unpublished paper presented at the "Privacy Poster Fair" at the *Carnegie Mellon University School of Library and Information Science*, 9:1-17. Available: <http://lorrie.cranor.org/courses/fa05/tubzhlp.pdf> (Accessed 1 October 2018).
- Graham E. 2017. Introduction: Data visualisation and the humanities. *English Studies*, 98(5):488-458. <https://doi.org/10.1080/0013838X.2017.1332021>
- Graham S, Milligan I & Weingart S. 2015. *Exploring big historical data: The historian's microscope*. London, UK: Imperial College Press. <https://doi.org/10.1142/p981>
- Grant R. 2017. Statistical literacy in the data science workplace. *Statistics Education Research Journal*, 17(1):17-21.
- Granville V. 2014. *Developing analytic talent: Becoming a data scientist*. Indianapolis, IN: John Wiley & Sons.
- Graunt J. 1662. *Natural and political observations, mentioned in a following index, and made upon the bills of mortality*. Edited with an introduction by WF Willcox. Baltimore: Johns Hopkins Press, 1939.
- Gray E, Jennings W, Farrall S & Hay C. 2015. Small big data: Using multiple datasets to explore unfolding social and economic change. *Big Data & Society*, 2(1):1-6. <https://doi.org/10.1177/2053951715589418>
- Gray J. 2007. Jim Gray on eScience: A transformed scientific method. Available: http://microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt (Accessed 15 May 2018).
- Gregory I, Cooper D, Hardie A & Rayson P. 2015. Spatializing and analysing digital texts: Corpora, GIS and places. In: D Bodenhamer, J Corrigan & T Harris (eds). *Deep maps and spatial narratives*. Bloomington, IN: Indiana University Press. 150-178. <https://doi.org/10.2307/j.ctt1zxxzr2.11>
- Greller W & Drachler H. 2012. Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, 15(3):42-57.
- Grill-Spector K & Kanwisher N. 2005. Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, 16(2):152-160. <https://doi.org/10.1111/j.0956-7976.2005.00796.x>

- Grimmer J. 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, 48(1):80-83. <https://doi.org/10.1017/S1049096514001784>
- Grobler R. 2018. No compensation for guessing what SA's 'word' of the year is. *News24*, 16 October. Available: <https://www.news24.com/SouthAfrica/News/no-compensation-for-guessing-what-sas-word-of-the-year-is-20181016> (Accessed 8 January 2019).
- Groenfeldt T. 2012. IBM and Ohio State University get analytical. *Forbes*, 29 November. Available: <https://www.forbes.com/sites/tomgroenfeldt/2012/11/29/ibm-and-ohio-state-university-get-analytical/#7c1d815f7bb> (Accessed 13 December 2017).
- Grosser B. 2014. What do metrics want? How quantification prescribes social interaction on Facebook. *Computational Culture*, 4. Available: <http://computationalculture.net/article/what-do-metrics-want> (Accessed 15 March, 2018). <https://dx.doi.org/10.1007/s13398-014-0173-7.2>
- Grusin R. 2014. The dark side of digital humanities: Dispatches from two recent MLA conventions. *differences*, 25(1):79-92. <https://doi.org/10.1215/10407391-2420009>
- Guberman S. 2015. On Gestalt theory principles. *Gestalt Theory*, 37(1):25-44.
- Guetterman T, Creswell JW & Kuckartz U. 2015. Using joint displays and MAXQDA software to represent the results of mixed methods research. In: MT McCrudden, G Schraw & C Buckendahl (eds). *Use of visual displays in research and testing: Coding, interpreting, and reporting data*. Charlotte: Information Age Publishing. 145-176.
- Gunarathne, T, Wu TL, Qiu J & Fox G. 2010. Cloud computing paradigms for pleasingly parallel biomedical applications. In: P Dinda (ed). *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. New York, NY: ACM:460-469. <https://doi.org/10.1145/1851476.1851544>
- Gunaratne SA. 2001. Paper, printing and the printing press: A horizontally integrative macrohistory analysis. *Gazette (Leiden, Netherlands)*, 63(6):459-479. <https://doi.org/10.1177/0016549201063006001>
- Gutierrez M & Milan S. 2017. Technopolitics in the age of big data. In: FS Caballero & T Gravante (eds). *Networks, movements & technopolitics in Latin America: Critical analysis and current challenges*. USA: Palgrave Macmillan. 95-109. https://doi.org/10.1007/978-3-319-65560-4_5
- Hacking I. 1991. How should we do the history of statistics?. In: G Burchill, C Gordon & P Miller (eds). *The Foucault effect*. Chicago: The Chicago University Press. 181-195.
- Halevi G & Moed H. 2012. The evolution of big data as a research and scientific topic: overview of the literature. *Research Trends*, 30(1): 3-6.
- Hall C. 2016. Writing history, making 'race': Slave-owners and their stories. *Australian Historical Studies*, 47(3):365-380. <https://doi.org/10.1080/1031461X.2016.1202291>
- Hammond A, Brooke J & Hirst G. 2016. Modeling modernist dialogism: Close reading with big data. In: S Ross & J O'Sullivan (eds). *Reading modernism with machines*. London, UK: Palgrave Macmillan. 49-77. https://doi.org/10.1057/978-1-137-59569-0_3
- Handelman LD & Lester D. 2007. The content of suicide notes from attempters and completers. *Crisis*, 28(2):102-104. <https://doi.org/10.1027/0227-5910.28.2.102>

References

- Harford T. 2014. Big data: Are we making a big mistake?. *FT Magazine*, 28 March. Available: <https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2014.00778.x> (Accessed 25 May 2018).
- Harkness T. 2016. *Big data: Does size matter?*. USA: Bloomsbury Publishing.
- Harris J & Eitel-Porter R. 2012. Data scientists: As rare as unicorns. *The Guardian*, 12 February. Available: <http://www.theguardian.com/media-network/2015/feb/12/data-scientists-as-rare-as-unicorns> (Accessed 9 September 2018).
- Harris J, Shetterley N, Alter AE & Schnell K. 2013. The team solution to the data scientist shortage. *Accenture*. Available: https://www.accenture.com/r20150923T082247Z__w__ie-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Indurities_17/Accenture-Team-Solution-Data-Scientist-Shortage.pdf (Accessed 9 September 2018).
- Harrison SE & Johnson PA. 2016. Crowdsourcing the disaster management cycle. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 8(4):17-40. <https://doi.org/10.4018/IJISCRAM.2016100102>
- Haseeb A & Pattun G. 2017. A review on NoSQL: Applications and challenges. *International Journal of Advanced Research in Computer Science*, 8(1):203-207. <https://dx.doi.org/10.26483/ijarcs.v8i1.2885>
- Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A & Khan SU. 2015. The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, 47:98-115. <https://doi.org/10.1016/j.is.2014.07.006>
- Hauge MV, Stevenson MD, Rossmo DK & Le Comber SS. 2016. Tagging Banksy: Using geographic profiling to investigate a modern art mystery. *Journal of Spatial Science*, 61(6):185-190. <https://doi.org/10.1080/14498596.2016.1138246>
- Head, T. 2018. Farm murders: Six surprising facts we learned from crime stats 2017/2018. *The South African*, 11 October. Available: <https://www.thesouthafrican.com/farm-murders-crime-stats-2017-2018/> (Accessed: 5 February 2019).
- He W, Zha S & Li L. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464-472. <https://doi.org/10.1016/j.ijinfomgt.2013.01.001>
- He X, Liu P, Zhang W & He K. 2016. Study on the mobile cloud framework for sociology: An empirical implementation. Paper presented at *Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on Cloud Computing and Big Data Analysis*. 5-7 July 2016. Chengdu, China: IEEE. 9-14. <https://doi.org/10.1109/ICCCBDA.2016.7529526>
- Heidegger M. 1966. Memorial address. Translated by JM Anderson & EH Freund. *Discourse on thinking*. New York, NY: Harper and Row. 43-57.
- Heidegger M. 1977. Science and reflection. Translated by W Lovitt. *The question concerning technology, and other essays*. New York, NY: Harper Collins. 155-182.
- Helbing D. 2015. Societal, economic, ethical and legal challenges of the digital revolution: From big data to deep learning, artificial intelligence, and manipulative technologies. Available: <https://arxiv.org/ftp/arxiv/papers/1504/1504.03751.pdf> (Accessed 17 March 2018). <https://doi.org/10.2139/ssrn.2594352>

- Hellerstein JM. 2008. Quantitative data cleaning for large databases. *White Paper, United Nations Economic Commission for Europe (UNECE)*. Available: <http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf> (Accessed 1 October 2018).
- Henderson S & Segal EH. 2013. Visualizing qualitative data in evaluation research. *New Directions for Evaluation*, 2013(139):53-71. <https://doi.org/10.1002/ev.20067>
- Heuser R & Le-Khac L. 2011. Learning to read data: Bringing out the humanistic in the digital humanities. *Victorian Studies*, 54(1):79-86. <https://doi.org/10.2979/victorianstudies.54.1.79>
- Hindman M. 2015. Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1): 48-62. <https://doi.org/10.1177/0002716215570279>
- Hirsch DD. 2014. That's unfair! Or is it? Big data, discrimination and the FTC's unfairness authority. *Kentucky Law Journal*, 103: 345-361.
- Hitchcock T. 2014. Big data, small data and meaning. *Historyonics*, 9 November. Available: http://historyonics.blogspot.sg/2014/11/big-data-small-data-and-meaning_9.html (Accessed 13 April 2018).
- Hitzler P & Janowicz K. 2013. Linked data, big data, and the 4th paradigm. *Semantic Web*, 4(3):233-235. <https://dx.doi.org/10.3233/SW-130117>
- Hocevar KP, Flanagan AJ & Metzger MJ. 2014. Social media self-efficacy and information evaluation online. *Computers in Human Behavior*, 39:254-262. <https://doi.org/10.1016/j.chb.2014.07.020>
- Hofmann B. 2006. When means become ends: Technology producing values. *Seminar. Net*, 2(2):1-12.
- Holmes DE. 2017. *Big data: A very short introduction*. New York, NY: Oxford University Press. <https://doi.org/10.1093/actrade/9780198779575.001.0001>
- Holtz D. 2014. 8 skills you need to be a data scientist. *Udacity*, 7 November. Available: <https://blog.udacity.com/2014/11/data-science-job-skills.html> (Accessed 7 October 2015).
- Hornof AJ. 2004. Cognitive strategies for the visual search of hierarchical computer displays. *Human-Computer Interaction*, 19(3):183-223. https://doi.org/10.1207/s15327051hci1903_1
- Horwitz RB & Currie W. 2007. Another instance where privatization trumped liberalization: The politics of telecommunications reform in South Africa – A ten-year retrospective. *Telecommunications Policy*, 31(8):445-462. <https://doi.org/10.1016/j.telpol.2007.05.008>
- Housley W, Dicks B, Henwood K & Smith R. Qualitative methods and data in digital societies. *Qualitative Research*, 17(6):607-609. <https://doi.org/10.1177/1468794117730936>
- Howard JH, Kazar ML, Menees SG, Nichols DA, Satyanarayanan M, Sidebotham RN & West MJ. 1988. Scale and performance in a distributed file system. *ACM Transactions on Computer Systems*, 6(1):51-81. <https://doi.org/10.1145/37499.37500>
- Hoyt E. 2014. Lenses for Lantern: Data mining, visualization, and excavating film history's neglected sources. *Film History: An International Journal*, 26(2):146-168. <https://doi.org/10.2979/filmhistory.26.2.146>
- Hu H, Wen Y, Chua TS & Li X. 2014. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2:652-687. <https://doi.org/10.1109/ACCESS.2014.2332453>

References

- Hudson, JM & Bruckman A. 2004. 'Go away': Participant objections to being studied and the ethics of chat room research. *The Information Society*, 20:127-139. <https://doi.org/10.1080/01972240490423030>
- Ibarra F. 2012. 4 architecture considerations for big data analytics. *VFabric*, 28 August. Available: <https://blogs.vmware.com/vfabric/2012/08/4-key-architecture-considerations-for-big-data-analytics.html> (Accessed 12 May 2018).
- Idhe D. 2010. A phenomenology of technics. In: C Hanks (ed). *Technology and values: Essential readings*. Singapore: John Wiley & Sons. 134-156.
- Igarashi Y, Altman T, Funada M & Kamiyama B. 2014. *Computing: A historical and technical perspective*. USA: CRC Press. <https://doi.org/10.1201/b17011>
- Iliadis A & Russo F. 2016. Critical data studies: An introduction. *Big Data & Society*, 3(2):1-7. <https://doi.org/10.1177/2053951716674238>
- Imran M, Elbassuoni S, Castillo C, Diaz F & Meier P. 2013. Extracting information nuggets from disaster-related messages in social media. In: T Comes, F Fiedrich, S. Fortier, J Geldermann & Y Yang (eds). *Proceedings of the 10th International ISCRAM Conference*. Baden-Baden, Germany: ISCRAM Association:1-10.
- Inmon WH. 2005. *Building the data warehouse: Getting started*. Indianapolis, IN: John Wiley & Sons.
- Innes M, Roberts C, Preece A & Rogers D. 2016. Ten "Rs" of social reaction: Using social media to analyse the "post-event" impacts of the murder of Lee Rigby. *Terrorism and Political Violence*: 1-21. Available: <https://www.tandfonline.com/doi/full/10.1080/09546553.2016.1180289> (Accessed 25 April 2018). <https://doi.org/10.1080/09546553.2016.1180289>
- Ioannidis JP. 2013. Informed consent, big data, and the oxymoron of research that is not research. *The American Journal of Bioethics*, 13(4):40-42. <https://doi.org/10.1080/15265161.2013.768864>
- Izquierdo JLC & Cabot J. 2016. JSONDiscoverer: Visualizing the schema lurking behind JSON documents. *Knowledge-Based Systems*, 103:52-55. <https://doi.org/10.1016/j.knsys.2016.03.020>
- Jacobs A. 2009. The pathologies of big data. *Communications of the ACM*, 52(8):36-44. <https://doi.org/10.1145/1536616.1536632>
- Jagadish HV. 2015. Big data and science: Myths and reality. *Big Data Research*, 2(2):49-52. <https://doi.org/10.1016/j.bdr.2015.01.005>
- Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R & Shahabi C. 2014. Big data and its technical challenges. *Communications of the ACM*, 57(7):86-94. <https://doi.org/10.1145/2611567>
- Jahns V. 2014. *Information visualization: perception for design by Colin Ware*. ACM SIGSOFT Software Engineering Notes, 39(2):43-44. <https://doi.org/10.1145/2579281.2579288>
- Jang SH & Callingham R. 2012. Conducting research in social media research: Ethical challenges. In: SI Fan, T Lê, Q Lê & Y Yue (eds). *Conference Proceedings: Innovative Research in a Changing and Challenging World*. Launceston: Australian Multicultural Interaction Institute:70-80.
- Jänicke S, Franzini G, Cheema MF & Scheuermann G. 2015. On close and distant reading in digital humanities: A survey and future challenges. Paper presented at the *Eurographics Conference on Visualization (EuroVis)-STARs*. 25-29 May 2015. Cagliari, Italy: The Eurographics Association. <https://dx.doi.org/10.2312/eurovisstar.20151113>

- Jiang H, Lin P & Qiang M. 2015. Public-opinion sentiment analysis for large hydro projects. *Journal of Construction Engineering and Management*, 142(2):1–12. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001039](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001039)
- Jianqiang Z & Xiaolin G. 2017. Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, 5:2870–2879. <https://doi.org/10.1109/ACCESS.2017.2672677>
- Jin X, Wah BW, Cheng X & Wang Y. 2015. Significance and challenges of big data research. *Big Data Research*, 2: 59–64. <https://doi.org/10.1016/j.bdr.2015.01.006>
- Jockers ML. 2013. *Macroanalysis: Digital methods and literary history*. USA: University of Illinois Press. <https://doi.org/10.16995/dscn.62>
- Johnson-Laird PN. 2010. Mental models and human reasoning. *PNAS*, 107(43):18243–18250. <https://doi.org/10.1073/pnas.1012933107>
- Johnson-Laird PN & Byrne RM. 2002. Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, 109(4):646–678. <https://doi.org/10.1037//0033-295X.109.4.646>
- Johnson-Laird PN, Khemlani SS & Goodwin GP. 2015. Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19(4):201–214. <https://doi.org/10.1016/j.tics.2015.02.006>
- Jones NJ & Bennell C. 2007. The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research*, 11(2):219–233. <https://doi.org/10.1080/13811110701250176>
- Jordaan AJJ & Van der Merwe A. 2015. Best practices for learning analytics initiatives in higher education. Universities South Africa. In: WR Kilfoil (ed). *Moving beyond the hype: A contextualised view of learning with technology in higher education*. Pretoria, South Africa: Universities South Africa. 53–58.
- Jordan G. 2014. *Practical Neo4j*. Berkeley, CA: Apress. <https://doi.org/10.1007/978-1-4842-0022-3>
- Jukić N, Sharma A, Nestorov S & Jukić B. 2015. Augmenting data warehouses with big data. *Information Systems Management*, 32(3):200–209. <https://doi.org/10.1080/10580530.2015.1044338>
- Jurafsky D & Martin JH. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall. <https://doi.org/10.1162/089120100750105975>
- Just MA, Pan L, Cherkassky VL, McMakin DL, Cha C, Nock MK & Brent D. 2017. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, 1(12):911–919. <https://doi.org/10.1038/s41562-017-0234-y>
- Kaefer F, Roper J & Sinha P. 2015. A software-assisted qualitative content analysis of news articles: Example and reflections. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 16(2):1–20. Available: <http://www.qualitative-research.net/index.php/fqs/article/view/2123/3815> (Accessed 4 April 2018). <http://dx.doi.org/10.17169/fqs-16.2.2123>
- Kahneman D. 2011. *Thinking, fast and slow*. New York, NY: Farrar, Straus & Giroux. <https://doi.org/10.1086/674372>
- Kamper H. 2017. Digital humanitarianism: Using big data. *The Borgen Project*, 24 February Available: <https://borgenproject.org/digital-humanitarianism/> (Accessed 29 May 2018).
- Kaplan F. 2012. How books will become machines. In: CM Jérôme, V François & V Joseph (eds). *Lire demain. Des manuscrits antiques à l'ère digitale*. Lausanne, Switzerland: PPUR. 25–41. <https://doi.org/10.3389/fdigh.2015.00001>

References

- Kaplan, F. 2015. A map for big data research in digital humanities. *Frontiers in Digital Humanities*, 2. Available: <https://www.frontiersin.org/articles/10.3389/fdigh.2015.00001/full> (Accessed 12 December 2017). <https://doi.org/10.3389/fdigh.2017.00012>
- Kaplan F & di Lenardo I. 2017. Big data of the past. *Frontiers in Digital Humanities*, 4(12):1-12. <https://doi.org/10.1111/cts.12178>
- Kaplan RM, Chambers DA & Glasgow RE. 2014. Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science*, 7(4):342-346. <https://dx.doi.org/10.1111/cts.12178>
- Karamshuk D, Shaw F, Brownlie J & Sastry N. 2017. Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, 1:33-43. <https://doi.org/10.1016/j.osnem.2017.01.002>
- Karpurapu BSH & Jololian L. 2017. A framework for social network sentiment analysis using big data analytics. In: SC Suh & T Anthony (eds). *Big data and visual analytics*. Cham, Switzerland: Springer. 203-218. https://doi.org/10.1007/978-3-319-63917-8_12
- Katal A, Wazid M & Goudar RH. 2013. Big data: Issues, challenges, tools and good practices. Paper presented at the *2013 Sixth international conference on contemporary computing*. 8-10 August 2013. Noida, India: IEEE. 404-409. <https://doi.org/10.1109/IC3.2013.6612229>
- Kaufner E. 2016. Qualitative research in the age of big data. *Electronic Ink*, 22 June. Available: <http://electronicink.com/qualitative-research-in-the-age-of-big-data/> (Accessed 5 October 2017).
- Kaufman LM. 2009. Data security in the world of cloud computing. *IEEE Security & Privacy*, 7(4):61-64. <https://doi.org/10.1109/MSP.2009.87>
- Keim D, Qu H, & Ma KL. 2013. Big-data visualization. *IEEE Computer Graphics and Applications*, 33(4):20-21. <https://doi.org/10.1109/MCG.2013.54>
- Kelleher JD & Tierney B. 2018. *Data science*. Cambridge, MA: The MIT Press. <https://doi.org/10.7551/mitpress/11140.001.0001>
- Kellen V, Recktenwald A & Burr S. 2013. Applying Big Data in higher education: A case study. Arlington, MA: *Cutter Consortium*, 13(8):1-39.
- Kennedy H & Hill RL. 2017a. The pleasure and pain of visualizing data in times of data power. *Television & New Media*, 18(8):769-782. <https://doi.org/10.1177/1527476416667823>
- Kennedy H & Hill RL. 2017b. The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology*. Available: <http://eprints.whiterose.ac.uk/105061/11/Kennedy%20-%20The%20Feeling%20of%20Numbers%20-%20AFC%202016-09-14.pdf> (Accessed 15 March 2018). <https://doi.org/10.1177/0038038516674675>
- Kennedy H, Hill RL, Aiello G & Allen W. 2016. The work that visualisation conventions do. *Information, Communication & Society*, 19(6):715-735. <https://doi.org/10.1080/1369118X.2016.1153126>
- Kennedy H, Moss G, Birchall C & Moshonas S. 2015. Balancing the potential and problems of digital methods through action research: Methodological reflections. *Information, Communication & Society*, 18(2):172-186. <https://doi.org/10.1080/1369118X.2014.946434>

- Kesari G. 2018. What's the secret source to transforming into a unicorn in data science?. *Towards Data Science*, 7 June. Available: <https://towardsdatascience.com/whats-the-secret-sauce-to-transforming-into-a-unicorn-in-data-science-94082b01c39d> (Accessed 9 September 2018).
- Khalifa M, Jennings M, Briscoe F, Oleszweski A & Abdi N. 2014. Racism? Administrative and community perspectives in data-driven decision making. *Urban Education*, 49(2):147-181. <https://doi.org/10.1177/0042085913475635>
- Killalea D. 2018. South Africa: Peter Dutton's 'white farmer' comments anger Pretoria. *News.com.au*, 16 March. Available: <https://www.news.com.au/finance/economy/world-economy/south-africa-peter-duttons-white-farmer-comments-anger-pretoria/news-story/a6a48505f72dabf517e961efa58242be> (Accessed 8 January 2019).
- Kim B, Trimi S & Chung J. 2014. Big-data applications in the government sector. *Communications of the ACM*, 57(3):78-85. <https://doi.org/10.1145/2500873>
- Kim EHJ, Jeong YK, Kim Y, Kang KY & Song M. 2016. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science*, 42(6):763-781. <https://doi.org/10.1177/0165551515608733>
- Kim SM & Hovy E. 2004. Determining the sentiment of opinions. In: E Yuste, SJ Jekat, AK Pantli & G Massey (eds). *In Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics:1367. <https://doi.org/10.3115/1220355.1220555>
- Kim W, Jeong OR & Kim C. 2014. A holistic view of big data. *International Journal of Data Warehousing and Mining*, 10(3):59-69. <https://doi.org/10.4018/ijdw.2014070104>
- Kimball R & Caserta J. 2004. *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data*. New York, NY: John Wiley & Sons.
- Kirk A. 2016. *Data visualisation: A handbook for data driven design*. London, UK: Sage Publications. <https://doi.org/10.1177/2399808317715320>
- Kirkegaard EOW & Bjerrekær JD. 2016. The OKCupid dataset: A very large public dataset of dating site users. *Open Differential Psychology*, 46:1-10. <https://doi.org/10.26775/ODP.2016.11.03>
- Kitchin R. 2014a. *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand Oaks, California: Sage Publications. <https://doi.org/10.1111/jors.12293>
- Kitchin R. 2014b. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1-12. <https://doi.org/10.1177/2053951714528481>
- Kitchin R & Lauriault TP. 2014. Towards critical data studies: Charting and unpacking data assemblages and their work. *The Programmable City Working Paper 2*. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2474112 (Accessed 13 September 2018).
- Kitchin R & Lauriault TP. 2015. Small data in the era of big data. *GeoJournal*, 80(4):463-475. <https://doi.org/10.1007/s10708-014-9601-7>
- Knox D. 2010. Spies in the house of learning: A typology of surveillance in online learning environments. Paper presented at the *EDGE 2010 – e-Learning: the horizon and beyond conference*. 12-15 October 2010. St. John's, Newfoundland and Labrador, Canada.

References

- Kobourov SG, Mchedlidze T & Vonessen L. 2015. Gestalt principles in graph drawing. *International Symposium on Graph Drawing and Network Visualization*. 24-26 September 2015. Los Angeles, CA: Springer. 558-560. https://doi.org/10.1007/978-3-319-27261-0_50
- Kosala R & Blockeel H. 2000. Web mining research. *ACM SIGKDD Explorations Newsletter*, 2(1):1-15. <https://doi.org/10.1145/360402.360406>
- Kosslyn SM. 2006. *Graphic design for the eye and the mind*. New York, NY: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311846.001.0001>
- Kotz S. 2005. Reflections on early history of official statistics and a modest proposal for global coordination. *Journal of Official Statistics*, 21(2):139-144.
- Kramer AD, Guillory JE & Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, 111(24):8788-8790. <https://doi.org/10.1073/pnas.1320040111>
- Kriel K. 2018. No Mr Ramaphosa, we're in the US fighting for SA – AfriForum. *Politicsweb*, 8 May. Available: <https://www.politicsweb.co.za/news-and-analysis/no-mr-ramaphosa-were-in-the-us-fighting-for-sa--af> (Accessed 8 Januarie 2019).
- Krishnan K. 2013. *Data warehousing in the age of big data*. Amsterdam: Elsevier, Morgan Kaufmann. <https://doi.org/10.1016/C2012-0-02737-8>
- Kruger C. 2016. (Dis)empowered whiteness: Un-whitely spaces and the production of the good white home. *Anthropology Southern Africa*, 39(1):46-57. <https://doi.org/10.1080/123323256.2016.1157026>
- Krum R. 2013. *Cool infographics: Effective communication with data visualization and design*. Indianapolis, IN: John Wiley & Sons.
- Küçükkeçeci C & Yazici A. 2018. Big data model simulation on a graph database for surveillance in wireless multimedia sensor networks. *Big Data Research*, 11:33-43. <https://doi.org/10.1016/j.bdr.2017.09.003>
- Kuhn T. 1996. *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226458106.001.0001>
- Kumar KK & Geethakumari G. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1):1-22. <https://doi.org/10.1186/s13673-014-0014-x>
- Labrinidis A & Jagadish HV. 2012. Challenges and opportunities with big data. In: A Saçan & N Tatbul (ed). *Proceedings of the VLDB Endowment*, 5(12): 2032-2033. <https://doi.org/10.14778/2367502.2367572>
- Lake P & Drake R. *Information systems management in the big data era*. USA: Springer. <https://doi.org/10.1007/978-3-319-13503-8>
- Lakoff G & Johnson M. 1980. *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Lakshman A & Malik P. 2010. Cassandra: A decentralized structure storage system. *ACM SIGOPS Operating Systems Review*: 1-6. Available: http://www.cl.cam.ac.uk/~ey204/teaching/ACS/R212_2014_2015/papers/lakshman_ladis_2009.pdf (Accessed 28 May 2018). <http://dx.doi.org/10.1145/1773912.1773922> <https://doi.org/10.1145/1773912.1773922>
- Lam D. 2018. Big data challenges in social sciences & humanities research. *Datanami*, 8 September. Available: <https://www.datanami.com/2014/09/08/big-data-challenges-social-sciences-humanities-research/> (Accessed 10 September 2018).
- Land R & Bayne SS. 2005. Screen or monitor: Issues of surveillance and disciplinary power in online learning environments. In: R Land & S Bayne (eds). *Education in cyberspace*. London, UK: Routledge. 165-178.

- Laney D. 2001. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6. Available: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (Accessed 8 December 2017).
- Landsat S, Khoshgoftaar TM, Richter AN & Hasanin T. 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1):2-36. <https://doi.org/10.1186/s40537-015-0032-1>
- Latzko-Toth G, Bonneau C & Millette M. 2017. Small data, thick data: Thickening strategies for trace-based social media research. In: L Sloan & A Quan-Haase (eds). *The SAGE PUBLICATIONS handbook of social media research methods*. New York NY: Sage Publications. 199-214. <https://doi.org/10.4135/9781473983847.n13>
- Lazar D, Kennedy R, King G & Vespignani A. 2014. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203-1205. <https://doi.org/10.1126/science.1248506>
- Lee HS, Lee HR, Park JU & Han YS. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22-31. <https://doi.org/10.1016/j.dss.2018.06.009>
- Lee KH, Lee YJ, Choi H, Chung YD & Moon B. 2011. Parallel data processing with MapReduce: A survey. *ACM SIGMOD Record*, 40(4):11-20. <https://doi.org/10.1145/2094114.2094118>
- Leggett W. 2014. The politics of behaviour change: Nudge, neoliberalism and the state. *Policy & Politics*, 42(1):3-19. <https://doi.org/10.1332/030557312X655576>
- Le Grange L. 2016. Decolonising the university curriculum: Leading article. *South African Journal of Higher Education*, 30(2):1-12. <https://doi.org/10.20853/30-2-709>
- LeGreco M & Tracy SJ. 2009. Discourse tracing as qualitative practice. *Qualitative Inquiry*, 15(9):1516-1543. <https://doi.org/10.1177/1077800409343064>
- Lehrer J. 2010. A physicist solves the city. *New York Times*, 17 December. Available: http://www.nytimes.com/2010/12/19/magazine/19Urban_West-t.html (Accessed 9 December 2017).
- Leiner B, Cole R, Postel J & Mills D. 1985. The DARPA internet protocol suite. *IEEE Communications Magazine*, 23(3):29-34. <https://doi.org/10.1109/MCOM.1985.1092530>
- Leinweber D. 2007. Stupid data miner tricks: Overfitting the S&P 500. *The Journal of Investing*, 16(1):15-22. <https://doi.org/10.3905/joi.2007.681820>
- Lemire S & Petersson GJ. (In press). Big bang or big bust? The role and implications of big data in evaluation. In: GJ Petersson & JD Breul (eds). *Cyber society, big data, and evaluation: Comparative policy evaluation*. New Jersey, New Brunswick: Transaction Publishers.
- Lemke M. 2014. Frequenzanalyse und Diktionsansatz. *eTMV*, 1(5). Available: http://www.epol-projekt.de/wp-content/uploads/2014/10/eTMV_1.pdf (Accessed 10 April 2018). <https://doi.org/10.1007/s13222-014-0174-x>
- Lemke M, Niekler A, Schaal GS & Wiedemann G. 2015. Content analysis between quality and quantity. *Datenbank-Spektrum*, 15(1):7-14. <https://dx.doi.org/10.1007/s13222-014-0174-x>
- Lemmens JC & Henn M. 2016. Learning analytics: A South African higher education perspective. In: J Botha & NJ Muller (eds). *Institutional Research in South African higher education. Intersecting contexts and practices*. Stellenbosch, South Africa: AFRICAN SUN MeDIA. 231-253. <https://doi.org/10.18820/9781928357186/12>

References

- Levaux C. 2017. The forgotten history of repetitive audio technologies. *Organised Sound*, 22(2):187-194. <https://doi.org/10.1017/S1355771817000097>
- Levi AS. 2013. Humanities 'big data': Myths, challenges, and lessons. In: X Hu, TY Lin, V Raghaven, B Wah, R Baeza-Yates, G Fox, C Shahabi, M Smith, Q Yang, R Lempel & R Nambiar (eds). *Big Data, 2013 IEEE International Conference Proceedings*. Sana Clara, CA: IEEE:33-36. <https://doi.org/10.1109/BigData.2013.6691667>
- Levin N. 2018. Big Data and biomedicine. In: M Meloni, J Cromby, D Fitzgerald & S Lloyd (eds). *The Palgrave handbook of biology and society*. London, UK: Palgrave Macmillan. 663-681. https://doi.org/10.1057/978-1-137-52879-7_28
- Lewis K. 2015. Three fallacies of digital footprints. *Big Data & Society*, 2(2): 1-4. <https://doi.org/10.1177/2053951715602496>
- Lewis K, Kaufman J, Gonzalez M, Wimmer A & Christakis N. 2008. Tastes, ties, and time: A new social network dataset using Facebook. *Social networks*, 30(4):330-342. <https://doi.org/10.1016/j.socnet.2008.07.002>
- Li D, Cao J & Yao Y. 2015. Big data in smart cities. *Science China Information Sciences*, 58(10):1-12. <https://doi.org/10.1007/s11432-015-5396-5>
- Li D & Wang X. 2017. Dynamic supply chain decisions based on networked sensor data: An application in the chilled food retail chain. *International Journal of Production Research*, 55(17):5127-5141. <https://doi.org/10.1080/00207543.2015.1047976>
- Li S, Da D Xu & Zhao S. (In press). 5G internet of things: A survey. *Journal of Industrial Information Integration*.
- Li Y, Gai K, Qiu L, Qiu M & Zhao H. 2017. Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Information Sciences*, 387:103-115. <https://doi.org/10.1016/j.ins.2016.09.005>
- Lidlicker JCR. 1963. Memorandum for: Members and affiliates of the intergalactic computer network. *Kurzweil Network*. Available: <http://www.kurzweilai.net/memorandum-for-members-and-affiliates-of-the-intergalactic-computer-network> (Accessed 11 October 2018).
- Life Esidimeni arbitration hearings: Qedani Mahlangu I. 2018. *SABC News*, 24 January. Available: <https://www.youtube.com/watch?v=bsO8pkkt6o> (Accessed 2 February 2018).
- Life Esidimeni arbitration hearings: Qedani Mahlangu II. 2018. *SABC News*, 25 January. Available: <https://www.youtube.com/watch?v=InQtYsktdfg> (Accessed 2 February 2018).
- Lin J. 2015. On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):33-47. <https://doi.org/10.1177/0002716215569174>
- Lin S, Fortuna J, Kulkarni C, Stone M & Heer J. 2013, June. Selecting semantically resonant colors for data visualization. *Computer Graphics Forum*, 32(3):401-410. <https://doi.org/10.1111/cgf.12127>
- Lindsay BR. 2011. Social media and disasters: Current uses, future options, and policy considerations. *CRS Report for Congress*, 6 September. Available: https://ofti.org/wp-content/uploads/2012/07/42245_gri-04-11-2011.pdf (Accessed 29 May 2018).

- Linzer DA. 2013. Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501):124-134. <https://doi.org/10.1080/01621459.2012.737735>
- Lipinski JD. 2009. Emerging legal issues in the collection and dissemination of Internet-based research data: Part II, Tort law issues involving defamation. *International Journal of Internet Research Ethics*, 2(1):58-72.
- Liu B. 2011. Social network analysis. In: B Liu (ed). *Web data mining*. Berlin, Heidelberg: Springer. 269-309. https://doi.org/10.1007/978-3-642-19460-3_7
- Liu B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1-167. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139084789>
- Liu B & Zhang L. 2012. A survey of opinion mining and sentiment analysis. In: C Aggarwal & C Zhai (eds). *Mining text data*. Boston, MA: Springer. 415-463. https://doi.org/10.1007/978-1-4614-3223-4_13
- Liu XL, Wang, HZ, Li, JZ & Gao H. 2017. Entity Manager: Managing dirty data based on entity resolution. *Journal of Computer Science and Technology*, 32(3):644-662. <https://doi.org/10.1007/s11390-017-1731-1>
- Lo SW, Wu JH, Lin FP & Hsu C.H. 2015. Cyber surveillance for flood disasters. *Sensors*, 15(2):2369-2387. <https://doi.org/10.3390/s150202369>
- Lohmeier C. 2014. The researcher and the never-ending field: Reconsidering big data and digital ethnography. In: M Hand & S Hillyard (eds). *Big data? Qualitative approaches to digital research: Studies in qualitative methodology*. Bingley, UK: Emerald Group Publishing Limited. 75-89. <https://doi.org/10.1108/S1042-319220140000013005>
- Lomborg S & Bechmann A. 2014. Using APIs for data collection on social media. *The Information Society*, 30(4):256-265. <https://doi.org/10.1080/01972243.2014.915276>
- López-Astorga M. 2016. Some arguments that the mental logic theory needs to clarify to continue being an alternative to the mental models theory. *Civilizar Ciencias Sociales y Humanas*, 16(31):235-248. <https://doi.org/10.22518/16578953.652>
- Loshin D. 2013. *Big data analytics: From strategic planning to enterprise integration with tools, techniques, NoSQL, and Graph*. Waltham: Elsevier, Morgan Kaufmann.
- Loukissas YA. 2012. *Co-designers: cultures of computer simulation in architecture*. London and New York, NY: Routledge. <https://doi.org/10.4324/9780203123065>
- Loukissas YA. 2016. A place for big data: Close and distant readings of accessions data from the Arnold Arboretum. *Big Data & Society*, 3(2):1-20. <https://doi.org/10.1177/2053951716661365>
- Luhn HP. 1958. A business intelligence system. *IBM Journal of Research and Development*, 2(4):314-319. <https://doi.org/10.1147/rd.24.0314>
- Lupton D. 2015. The thirteen Ps of big data. *This Sociological Life*. Available: https://www.researchgate.net/profile/Deborah_Lupton/publication/276207564_The_Thirteen_Ps_of_Big_Data/links/5552c2d808ae6fd2d81d5f20.pdf (Accessed 10 May 2018). <https://dx.doi.org/10.13140/RG.2.1.2900.8800>

References

- Luxhoj JT. 2016. System safety modeling of alternative geofencing configurations for small use. *International Journal of Aviation, Aeronautics, and Aerospace*, 3(1):1-27. <https://doi.org/10.15394/ijaaa.2016.1105>
- Lydia EL & Swarup MB. 2015. Big data analysis using Hadoop components like flume, mapreduce, pig and hive. *International Journal of Science, Engineering and Computer Technology*, 5(11):390-394.
- Lyon D. 2014. Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data & Society*, 1(2):1-13. <https://doi.org/10.1177/2053951714541861>
- Maciejewski M. 2017. To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, 83(1_suppl):120-135. <https://doi.org/10.1177/0020852316640058>
- MacLeod CM. 1991. Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2):163-203. <https://doi.org/10.1037//0033-2909.109.2.163>
- Madge C. 2007. Developing a geographer's agenda for online research ethics. *Progress in Human Geography*, 31:654-674. <https://doi.org/10.1177/0309132507081496>
- Mahrt M & Scharkow M. 2013. The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, 57(1):20-33. <https://doi.org/10.1080/08838151.2012.761700>
- Mai JE. 2016. Big data privacy: The datafication of personal information. *The Information Society*, 32(3):192-199. <https://doi.org/10.1080/01972243.2016.1153010>
- Maier CT & Deluiliis D. 2015. Recovering the human in the network: Exploring communicology as a research methodology in digital business discourse. In: E Darics (ed). *Digital business discourse*. London, UK: Palgrave Macmillan. 208-225. <https://doi.org/10.1080/01972243.2016.1153010>
- Makgoba MW. 2017. The Life Esidimeni disaster: The Makgoba report. Available: <https://www.sahrc.org.za/home/21/files/Esidimeni%20full%20report.pdf> (Accessed 1 December 2017).
- Manning CD & Schütze H. 1999. *Foundations of statistical natural language processing*. USA: MIT Press. <https://doi.org/10.1017/S1351324902212851>
- Manovich L. 2011. What is visualisation?. *Visual Studies*, 26(1):36-49. <https://doi.org/10.1080/01472586X.2011.548488>
- Manovich L. 2012. Trending: The promises and the challenges of big social data. In: MK Gold (ed). *Debates in the digital humanities*. Minneapolis, MN: University of Minnesota Press. 460-475. <https://doi.org/10.1080/01973762.2013.761126>
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C & Byers AH. 2011. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*. Available: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation (Accessed 8 December 2017).
- Markham A. 2013. Undermining 'data': A critical examination of a core term in scientific inquiry. *First Monday*, 18(10):1-14. Available: <http://uncommonculture.org/ojs/index.php/fm/article/view/4868/3749> (Accessed 6 December 2017). <https://doi.org/10.5210/fm.v18i10.4868>

- Markham A & Buchanan E. 2012. Ethical decision-making and internet research: Recommendations from the aoir ethics working committee (version 2.0). Available: <https://pure.au.dk/ws/files/55543125/aoirethics2.pdf> (Accessed 11 December 2017). <https://doi.org/10.5210/fm.v18i10.4868>
- Marks S. 2016. *The information nexus: Global capitalism from the Renaissance to the present*. London, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9781316258170>
- Marr B. 2015a. A brief history of big data everyone should read. *World Economic Forum*, 25 February. Available: <https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/> (Accessed 5 December 2017).
- Marr B. 2015b. Big data-as-service is next big thing. *Forbes*, 24 April. Available: <https://www.forbes.com/sites/bernardmarr/2015/04/27/big-data-as-a-service-is-next-big-thing/#77965ab633d5> (Accessed 13 December 2017).
- Marsden JH, Shirai Y & Wilkinson VA. 2018. Big data analytics and corporate social responsibility (CSR): Adding quantifiable and qualifiable sustainability science to the three P's. *Studies in Informatics, Shizuoka University*, 23:29-43. <https://doi.org/10.1109/ProComm.2018.00019>
- Mashey JR. 1998. Big data and the next wave of infraS-tress. *Computer Science Division Seminar*. University of California, Berkeley.
- Masinga L. 2018. Concourt ruling on Hoërskool Overvaal a blow to non-racialism – Lesufi. *IOL*, 27 July. Available: <https://www.iol.co.za/news/politics/concourt-ruling-on-hoerskool-overvaal-a-blow-to-non-racialism-lesufi-16281981> (Accessed 7 March 2019).
- Matsebula F & Makandla E. 2017. A big data architecture for learning analytics in higher education. In: DR Cornish (ed). *IEEE africon: Science, Technology and Innovation for Africa*. Cape Town, South Africa: IEEE AFRICON: 951-956. <https://doi.org/10.1109/AFRCON.2017.8095610>
- Matusiak KK, Meng L, Barczyk E & Shih CJ. 2015. Multilingual metadata for cultural heritage materials: The case of the Tse-Tsung Chow collection of Chinese scrolls and fan paintings. *The Electronic Library*, 33(1):136-151. <https://doi.org/10.1108/EL-08-2013-0141>
- Mavridis I & Karatza H. 2017. Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark. *Journal of Systems and Software*, 125:133-151. <https://doi.org/10.1016/j.jss.2016.11.037>
- Maxwell-Stewart H. 2016. Big data and Australian history. *Australian Historical Studies*, 47(3):359-364. <https://doi.org/10.1080/1031461X.2016.1208728>
- Mayer-Schönberger V & Cukier k. 2013. *Big data: An evolution that will transform how we live, work, and think*. Boston and New York, NY: Houghton Mifflin Harcourt.
- Maynard, D. & M.A. Greenwood. 2014. Who cares about sarcastic tweets? Investigating the Impact of sarcasm on sentiment analysis. In: N Calzolari, K Choukri, T Declerck, H Loftsson, B Maegaard, J Mariani, A Moreno, J Odijk & S Piperidis (eds). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: ELRA:4238-4243.
- Mayo M. 2018. Automated machine learning vs automated data science. *KDdnugets*, July. Available: <https://www.kdnuggets.com/2018/07/automated-machine-learning-vs-automated-data-science.html> (Accessed 9 September 2018).

References

- Mayring P. 2000. Qualitative content analysis. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 1(2):1-7. Available: <https://utsc.utoronto.ca/~kmacd/IDSC10/Readings/text%20analysis/CA.pdf> (Accessed 4 December 2017).
- Mazzocchi F. 2015. Could big data be the end of theory in science?. *EMBO Reports*, 16(10):1250-1255. <https://doi.org/10.15252/embr.201541001>
- McCarty N, Poole K & Rosenthal H. 2006. *Polarized America: The dance of inequality and unequal riches*. Cambridge, MA: MIT Press. <https://doi.org/10.1007/s00712-007-0295-x>
- McFarland DA, Lewis K & Goldberg A. 2016. Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist*, 47(1): 12-35. <https://doi.org/10.1007/s12108-015-9291-8>
- McIntosh C. 1998. Eighteenth-century English dictionaries and the Enlightenment. *The Yearbook of English Studies*, 28:3-18. <https://doi.org/10.2307/3508753>
- McKee HA & Porter JE. 2009. *The ethics of internet research: A rhetorical, case-based process*. New York, NY: Peter Lang. <https://doi.org/10.1016/j.compcom.2010.03.003>
- McPherson SS. 2009. *Tim Berners-Lee: Inventor of the World Wide Web*. Twenty-First Minneapolis, MN: Century Books.
- Meier P. 2015. *Digital humanitarians: How big data is changing the face of humanitarian response*. Boca Raton, FL: CRC Press. <https://doi.org/10.1007/s11673-017-9807-8>
- Mendenhall R, Brown N, Black ML, Van Moer M, Lourentzou I, Flynn K, Mckee M & Zerai A. 2016. Rescuing lost history: Using big data to recover black women's lived experiences. In: P Navrátil, M Dahan, D Hart, A Romanella & N Sukhija (eds). *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*. New York, NY: Article 56. <https://doi.org/10.1145/2949550.2949642>
- Messner M, Moll J & Strömsten T. 2017. Credibility and authenticity in qualitative accounting research. In: Z Hoque, LD Parker, MA Covaleski & K Haynes (eds). *The Routledge companion to qualitative accounting research methods*. New York, NY: Routledge. 432-444.
- Metcalf J & Crawford K. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, 3(1):1-14. <https://doi.org/10.1177/2053951716650211>
- Milan S. 2018. Data activism as the new frontier of media activism. In: V Pickard & G Yang (eds). *Media activism in the digital age*. New York, NY: Routledge. 151-163. <https://doi.org/10.4324/9781315393940-13>
- Milan S & Gutiérrez M. 2015. Citizens' media meets big data: The emergence of data activism. *Mediaciones*, (14):120-133. <https://doi.org/10.26620/uniminuto.mediaciones.11.14.2015.120-133>
- Milan S & van der Velden L. 2016. The alternative epistemologies of data activism. *Digital Culture & Society*, 2(2):57-74. <https://doi.org/10.14361/dcs-2016-0205>
- Miller G. 2011. Social scientists wade into the tweet stream. *Science*, 333(6051):1814-1815. <https://doi.org/10.1126/science.333.6051.1814>

- Mills CA. 2017. What are the threats and potentials of big data for qualitative research? *Qualitative Research*: 1-27. Available: <http://journals.sagepub.com/doi/pdf/10.1177/1468794117743465> (Accessed 17 April 2018). <https://doi.org/10.1177/1468794117743465>
- Mills RJ, Chudoba KM & Olsen DH. 2016. IS programs responding to industry demands for data scientists: A comparison between 2011-2016. *Journal of Information Systems Education*, 27(2): 131-141.
- Milnea D, Paris C, Christensen H, Batterham P & O'Deac B. 2015. We Feel: Taking the emotional pulse of the world. In G Lindgaard & D Moore (eds). *Proceedings 19th Triennial Congress of the IEA*. Melbourne, Australia: IEA:9-15.
- Milne JD, Jeffrey LM, Suddaby G & Higgins A. 2012. Early identification of students at risk of failing. In: M Brown, M Hartnett & T Stewart (eds). *Future challenges, sustainable futures*. Wellington, New Zealand: ASCILITE: 657-661.
- Minelli M, Chambers M & Dhiraj A. 2013. *Big data, big analytics: Emerging business intelligence and analytic trends for today's business*. Hoboken, NJ: John Wiley & Sons. <https://doi.org/10.1002/9781118562260>
- Miorandi D, Sicari S, De Pellegrini F & Chlamtac I. Internet of things: Vision, applications and research challenges. *Ad Hoc Networks*, 10(7):1497-1516. <https://doi.org/10.1016/j.adhoc.2012.02.016>
- Mitchley A. 2018. Hoërskool Overvaal denies allegations of racial segregation. *News24*, 9 January. Available: <https://www.news24.com/SouthAfrica/News/hoerskool-overvaal-denies-allegations-of-racial-segregation-20180109> (Accessed 8 January 2019).
- Mkokeli S. 2018. South Africa's path to land reform is riddled with pitfalls. *Bloomberg Businessweek*, 23 November. Available: <https://www.bloomberg.com/news/articles/2018-11-23/south-africa-s-path-to-land-reform-is-riddled-with-pitfalls> (Accessed 8 January 2019).
- Mohammad S, Kiritchenko S & Zhu X. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of Tweets. *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013)*:321-327.
- Mole B. 2015. New flu tracker uses Google search data better than Google. *Ars Technica*. Available <https://arstechnica.com/science/2015/11/new-flu-tracker-uses-google-search-data-better-than-google/> (Accessed 5 October 2017).
- Monroe BL, Pan J, Roberts ME, Sen M & Sinclair B. 2015. No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, 48(1):71-74. <https://doi.org/10.1017/S1049096514001760>
- Morabia A. 2013. Observations made upon the bills of mortality. *BMJ*:346"e8640. <https://doi.org/10.1136/bmj.e8640>
- Moravcsik A. 2014. Trust, but verify: The transparency revolution and qualitative international relations. *Security Studies*, 23(4):663-688. <https://doi.org/10.1080/09636412.2014.970846>
- Moretti F. 2005. *Graphs, maps, trees: Abstract models for a literary history*. London, UK: Verso.
- Mulder F, Ferguson J, Groenewegen P, Boersma K & Wolbers J. 2016. Questioning big data: Crowdsourcing crisis data towards an inclusive humanitarian response. *Big Data & Society*, 3(2):1-13. <https://doi.org/10.1177/2053951716662054>

References

- Mullins R. 2017. Digital transformation: Human evolution, not technological revolution. *BusinessLIVE*, 8 December. Available: <https://www.businesslive.co.za/redzone/news-insights/2017-12-08-digital-transformation-human-evolution-not-technological-revolution/> (Accessed 8 May 2018).
- Mwangi CAG. 2017. Partner positioning: Examining international higher education partnerships through a mutuality lens. *The Review of Higher Education*, 41(1):33-60. <https://doi.org/10.1353/rhe.2017.0032>
- Neff G, Tanweer A, Fiore-Gartland B & Osburn L. 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data*, 5(2):85-97. <https://doi.org/10.1089/big.2016.0050>
- Nelson B. 2014. The data on diversity. *Communications of the ACM*, 57(11):86-95. <https://doi.org/10.1145/2597886>
- Newton I. 1999. *The Principia: The mathematical principles of natural philosophy*, Book II. Berkeley, CA: University of California Press.
- Neyman J & Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(1933):289-337. <https://doi.org/10.1098/rsta.1933.0009>
- Nguyen NP, Yan G & Thai MT. 2013. Analysis of misinformation containment in online social networks. *Computer Networks*, 57(10):2133-2146. <https://doi.org/10.1016/j.comnet.2013.04.002>
- Nicholls SG, Langan SM & Benchimol EI. 2016. Reporting and transparency in big data: The nexus of ethics and methodology. In: B Mittelstadt & L Floridi (eds). *The ethics of biomedical big data. Law, governance and technology series*. Switzerland: Springer. 339-366. https://doi.org/10.1007/978-3-319-33525-4_15
- Nickig J. 2017. Does PoPI offer adequate legislation in the digital age? *Bizcommunity*, 22 May. Available: <https://www.bizcommunity.com/Article/196/717/162190.html> (Accessed 18 November 2019).
- Nightingale F. 1858. *Notes on matters affecting the health, efficiency, and hospital administration of the British army, founded chiefly on the experience of the late war*. London, UK: Harrison.
- Nisbett RE. 2003. *The geography of thought*. New York, NY: The Free Press.
- Nombembe P. 2018. Sheep slaughtered on Clifton beach as animal rights activists protest. *Timeslive*, 28 December. Available: <https://www.timeslive.co.za/news/south-africa/2018-12-28-sheep-slaughtered-on-clifton-beach-as-animal-rights-activists-protest/> (Accessed 8 January 2019).
- Nongxa LG. 2017. Mathematical and statistical foundations and challenges of (big) data sciences. *South African Journal of Science*, 113(3-4):1-4. <https://doi.org/10.17159/sajs.2017/a0200>
- Norman DA. 2004. *Emotional design: Why we love (or hate) everyday things*. New York, NY: Basic Books.
- Northrop D. 2014. Other globes. In: D Northrop (ed). *A companion to world history*. USA: Wiley-Blackwell. 497-526. <https://doi.org/10.1002/9781118305492.ch33>
- Oaksford M, Chater N & Larkin J. 2000. Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:883-889. <https://doi.org/10.1037/0278-7393.26.4.883>
- O'Brien DP. 2009. Human reasoning includes a mental logic. *Behavioral and Brain Sciences*, 32(1):96-97. <https://doi.org/10.1017/S0140525X09000429>

- O'Connor RC & Kirtley OJ. 2018. The integrated motivational-volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B*, 375(1754):1-10. <https://doi.org/10.1098/rstb.2017.0268>
- Odlum M & Yoon S. 2015. What can we learn about the Ebola outbreak from tweets?. *American Journal of Infection Control*, 43(6):563-571. <https://doi.org/10.1016/j.ajic.2015.02.023>
- Ohri A. 2017. *Python for R users: A data science approach*. Hoboken, NJ: John Wiley & Sons. <https://doi.org/10.1002/9781119126805>
- O'Leary DE. 2014. Embedding AI and crowdsourcing in the big data lake. *IEEE Intelligent Systems*, 29(5): 70-73. <https://doi.org/10.1109/MIS.2014.82>
- Oliva A & Torralba A. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23-36. [https://doi.org/10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2)
- Olshannikova E, Olsson T, Huhtamäki J & Kärkkäinen H. 2017. Conceptualizing big social data. *Journal of Big Data*, 4(1):1-19. <https://doi.org/10.1186/s40537-017-0063-x>
- Olshannikova E, Ometov A, Koucheryavy Y & Olsson T. 2015. Visualizing big data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, 2(1):1-27. <https://doi.org/10.1186/s40537-015-0031-2>
- O'Neil C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown.
- Orgad S. 2005. The transformative potential of online communication: The case of breast cancer patients' Internet spaces. *Feminist Media Studies*, 5(2):141-161. <https://doi.org/10.1080/14680770500111980>
- Orgad S. 2009. How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In: AN Markham & NK Baym (eds). *Internet inquiry: Conversations about method*. Los Angeles, CA: Sage Publications. 33-53. <https://doi.org/10.4135/9781483329086.n4>
- Osborne S. 2018. South Africa votes through motion that could lead to seizure of land from white farmers without compensation. *Independent*, 1 March. Available: <https://www.independent.co.uk/news/world/africa/south-africa-white-farms-land-seizure-anc-race-relations-a8234461.html> (Accessed 8 January 2019).
- Osgood CE & Walker EG. 1959. Motivation and language behavior: A content analysis of suicide notes. *Journal of Abnormal and Social Psychology*, 59:5-67. <https://doi.org/10.1037/h0047078>
- O'Sullivan D. 2017. Big data: Why (oh why) this computational science? In: J Thatcher, A Shears & J Eckert (eds). *Geography and the geoweb: Rethinking research in the advent of big data*. 1-27. UC Berkeley. Available: <https://escholarship.org/uc/item/0rn5n832> (Accessed 29 December 2017).
- Owens T. 2011. Defining data for humanists: Text, artifact, information or evidence?. *Journal of Digital Humanities*, 1(1):1-4. Available: <http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/> (Accessed 25 April 2018).
- Palmer D. 2012. Text Preprocessing. In: N Damerau & FJ Indurkha (eds). *Handbook of natural language processing*. New York, NY: CRC Press. 9-30.
- Pang B & Lee L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: D Scott, W Daelemans & MA Walker (eds). *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona: Association for Computational Linguistics:1-8. <https://doi.org/10.3115/1218955.1218990>

References

- Parastadidis S. 2009. A platform for all that we know: Creating a knowledge-driven research infrastructure. In: T Hey, S Tansley & K Tolle (eds). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research. 165-172. <https://doi.org/10.1108/13673270210440839>
- Park HW & Leydesdorff L. 2013. Decomposing social and semantic networks in emerging “big data” research. *Journal of Informetrics*, 7(3):756-765. <https://doi.org/10.1016/j.joi.2013.05.004>
- Parker C, Saundage D & Lee CY. 2011. Can qualitative content analysis be adapted for use by social informaticians to study social media discourse? A position paper. In: P Seltisikas, D Bunker, L Dawson & M Iddulka (eds). *Proceedings of the 22nd Australasian Conference on Information Systems: Identifying the Information Systems Discipline*. Sydney, Australia: Association of Information Systems:1-7.
- Patchin JW & Hinduja S. 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2):148-169. <https://doi.org/10.1177/1541204006286288>
- Patel K. 2017. Big data, its issues and challenges. *IJEDR*, 5(3): 124-126.
- Patil DJ. 2011. *Building data science teams*. Sebastopol, CA: O'Reilly Media, Inc.
- Patterson RE, Blaha LM, Grinstein GG, Liggert KK, Kaveney DE, Sheldon KC, Havig PR & Moore JA. 2014. A human cognition framework for information visualization. *Computers & Graphics*, 42:42-48. <https://doi.org/10.1016/j.cag.2014.03.002>
- Patty JW & Penn EM. 2015. Analyzing big data: Social choice and measurement. *PS: Political Science & Politics*, 48(1):95-101. <https://doi.org/10.1017/S1049096514001814>
- Pavlovskaya M. 2017. Qualitative GIS. In: D Richardson, N Castree, M Goodchild, A Kobayashi, W Liu & RA Marston (eds). *The international encyclopedia of geography*. USA: John Wiley and Sons, Ltd. 1-11. <https://doi.org/10.1002/9781118786352.wbieg1156>
- Pawlicka U. 2017. Data, collaboration, laboratory: Bringing concepts from science into humanities practice. *English Studies*, 98(5):526-541. <https://doi.org/10.1080/0013838X.2017.1332022>
- Pedone R, Hummel JE & Holyoak KJ. 2001. The use of diagrams in analogical problem solving. *Memory & Cognition*, 29(2):214-221. <https://doi.org/10.3758/BF03194915>
- Pennebaker JW, Francis ME & Booth RJ. 2001. *Linguistic inquiry and word count (LIWC 2001)*. Mahwah, NJ: Erlbaum.
- Pentzold C. & Fischer C. 2017. Framing big data: The discursive construction of a radio cell query in Germany. *Big Data & Society*, 4(2):1-11. <https://doi.org/10.1177/2053951717745897>
- Perkins L, Redmond E & Wilson J. *Seven databases in seven weeks: A guide to modern databases and the NoSQL movement*. USA: Pragmatic Bookshelf.
- Perrault C. 1964. *Parallèle des anciens et des modernes en ce qui regarde les arts et les sciences*. Paris: Jean Baptiste Coignard.
- Pestian J, Nasrallah H, Matykiewicz P, Bennett A & Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*, 3:19-28. <https://doi.org/10.4137/BII.S4706>
- Peters B. 2012. The big data gold rush. *Forbes*, 21 June. Available: <https://www.forbes.com/sites/bradpeters/2012/06/21/the-big-data-gold-rush/#59809a0eb247> (Accessed 4 December 2017).

- Pijoo I. 2018. Vicki Momberg sentenced to an effective 2 years in prison for racist rant. *News24*, 28 March. Available: <https://www.news24.com/SouthAfrica/News/vicki-momberg-sentenced-to-an-effective-2-years-in-prison-for-racist-rant-20180328> (Accessed 8 January 2019).
- Pitsilis GK, Ramampiaro H & Langseth H. 2018. Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433. Available: <https://arxiv.org/abs/1801.04433> (Accessed 18 October 2018).
- Ponniiah P. 2010. *Data warehousing fundamentals for IT professionals*. Hoboken, NJ: John Wiley & Sons. <https://doi.org/10.1002/9780470604137>
- Pool K & Rosenthal H. 1997. *Congress: A political-economic history of roll call voting*. Oxford: Oxford University Press.
- Popescu AM & Etzioni O. 2007. Extracting product features and opinions from reviews. In: A Kao & SR Poteet (eds). *Natural language processing and text mining*. London, UK: Springer. 9-28. <https://doi.org/10.3115/1220575.1220618>
- Porter C, Atkinson P & I. Gregory. 2015. Geographical Text Analysis: A new approach to understanding nineteenth-century mortality. *Health & Place*, 36:25-34. <https://doi.org/10.1016/j.healthplace.2015.08.010>
- Porter TM. 1986. *The rise of statistical thinking 1820–1900*. Princeton, NJ: Princeton University Press.
- Portmess L & Tower S. 2015. Data barns, ambient intelligence and cloud computing: The tacit epistemology and linguistic representation of Big Data. *Ethics and Information Technology*, 17:1–9. <https://doi.org/10.1007/s10676-014-9357-2>
- Porway J. 2013. You can't just hack your way to social change. *Harvard Business Review*, 7 March. Available: <https://hbr.org/2013/03/you-cant-just-hack-your-way-to> (Accessed 6 November 2017).
- Power A, Keane A, Nolan B & O'Neill B. 2017. A lexical database for public textual cyberbullying detection. *Revista de Linguas para Fines Especificos*, 23(2):157-186. <http://dx.doi.org/10.20420/rlfe.2017.177>.
- Power, DJ. 2008. Decision support systems concept. In: F Adam & P Humphreys (eds). *Encyclopedia of decision making and decision support*. Hershey, PA: Information Science Reference. 232-232. <https://doi.org/10.4018/978-1-59904-843-7.ch027>
- Powers Dirette D. 2016. Why the veracity of data matters in health care research. *The Open Journal of Occupational Therapy*, 4(4):1-4. <https://doi.org/10.15453/2168-6408.1324>
- Pradhan M & Rao N. (In press). Gender justice and food security: The case of public distribution system in India. *Progress in Development Studies*.
- Press G. 2013. A very short history of big data. *Forbes*, 9 May. Available: <https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#5ffc7ccb65a1> (Accessed 5 November 2017).
- Press G. 2015. The hunt for unicorn data scientists lifts salaries for all data analytics professionals. *Forbes*, 9 October. Available: <https://www.forbes.com/sites/gilpress/2015/10/09/the-hunt-for-unicorn-data-scientists-lifts-salaries-for-all-data-analytics-professionals/#2cd9c1fd5258> (Accessed 9 September 2018).
- Pretorius J. 2014. "Dubula ibhunu" (shoot the boer): A psycho-political analysis of farm attacks in South Africa. *Psychology in Society*, (47):21-40.
- Prinsloo P. 2016. Some provocations for a future of institutional research: Evidence-based decision-making and séance. In: J Botha & NJ Muller (eds). *Institutional research in South African higher education: Intersecting contexts and practices*. Stellenbosch: SUN MeDIA. 337-360. <https://doi.org/10.18820/9781928357186>

References

- Prinsloo P & Rowe M. 2015. Ethical considerations in using student data in an era of 'big data'. In: W Kilfoil (ed). *Moving beyond the hype: A contextualised view of learning with technology in higher education*. Pretoria, South Africa: Universities South Africa. 59-64.
- Prinsloo P & Slade S. 2017. An elephant in the learning analytics room: The obligation to act. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. Vancouver, British Columbia, Canada: ACM 46-55. <https://doi.org/10.1145/3027385.3027406>
- Prinsloo P. 2018. Context matters: An African perspective on institutionalizing learning analytics. In: CP Lim & VL Tinio (eds.). *Learning analytics for the global South*. Quezon City, Philippines: Foundation for Information Technology Education and Development.
- Protection of Personal Information Act 2013 (Act No. 4 of 2013). *SAICA*. Available: https://www.saica.co.za/Portals/0/Technical/LegalAndGovernance/37544_pro25.pdf (Accessed 11 December 2017).
- Provost F & Fawcett T. 2013. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51-59. <https://doi.org/10.1089/big.2013.1508>
- Pugin L. 2015. The challenge of data in digital musicology. *Frontiers in Digital Humanities*, 2:1-3. <https://doi.org/10.3389/fdigh.2015.00004>
- Punch KF. 2013. *Introduction to social research: Quantitative and qualitative approaches*. Thousand Oaks, CA: Sage Publications.
- Puschmann C & Burgess J. 2014. Big data, big questions: Metaphors of big data. *International Journal of Communication*, 8:1690-1709.
- Quinn PC & Bhatt RS. 2015. Development of perceptual organization in infancy. In: J Wagemans (eds). *The Oxford handbook of perceptual organization*. New York, NY: Oxford University Press. 691-792. <https://doi.org/10.1093/oxfordhb/9780199686858.013.016>
- Rahman N & Iverson S. 2015. Big data business intelligence in bank risk analysis. *International Journal of Business Intelligence Research*, 6(2):55-77. <https://doi.org/10.4018/IJBIR.2015070104>
- Ratner C. 2002. Subjectivity and objectivity in qualitative methodology. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 3(3):1-8. Available: <http://www.qualitative-research.net/index.php/fqs/article/view/829> (Accessed 4 December 2017). <http://dx.doi.org/10.17169/fqs-3.3.829>
- Rauscher J. 2014. Grasping cities through literary representations. A mix of qualitative and quantitative approaches to analyze crime novels. *Historical Social Research/Historische Sozialforschung*, 39(2):68-102. <https://dx.doi.org/10.12759/hsr.39.2014.2.68-102>
- Raymond NA. 2016. Beyond "do no harm" and individual consent: Reckoning with the emerging ethical challenges of civil society's use of data. In: L Taylor, L Floridi & B van der Sloot (eds). *Group privacy: New challenges of data technologies*. Cham, Switzerland: Springer. 62-82. https://doi.org/10.1007/978-3-319-46608-8_4
- Reeves J. 2014. 9 critical investing lessons from a Nobel prize winner. *The Motley Fool*, 24 November. Available: https://www.fool.com/slideshow/these-15-states-produce-94-us-natural-gas/?fs_test=False (Accessed 8 May 2018).

- Regalado A. 2011. Who coined “cloud computing”? *MIT Technology Review*, 31 October. Available: <https://www.technologyreview.com/s/425970/who-coined-cloud-computing/> (Accessed 5 December 2017).
- Reiberg A. 2016. The construction of an internet policy domain in German parliamentary debates and newspaper articles. Paper presented at the *24th IPSA World Conference on Political Science*. 23-28 July 2016. Poznan, Poland: 1-17. <https://dx.doi.org/10.1002/epa2.1001>
- Reips UD, Buchanan T, Krantz JH & McGrawK. 2015. Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica*, 15(2):139-148. <https://doi.org/10.21697/sp.2015.14.2.09>
- Resnick B. 2016. Researchers just released profile data on 70,000 OkCupid user without permission. *Vox*, 12 May. Available: <https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release> (Accessed 11 December 2017).
- Reyes A, Rosso P & Veale T. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239-268. <https://doi.org/10.1007/s10579-012-9196-x>
- Reyes JA. 2015. The skinny on big data in education: Learning analytics simplified. *Tech Trends*, 59(2):75-80. <https://doi.org/10.1007/s11528-015-0842-1>
- Reyes-Ortiz JL, Oneto L & Anguita D. 2015. Big data analytics in the cloud: Spark on hadoop vs mpi/openmp on beowulf. *Procedia Computer Science*, 53:121-130. <https://doi.org/10.1016/j.procs.2015.07.286>
- Richards NM & King JH. 2014. Big data ethics. *Wake Forest Law Review*, 49:393-432.
- Riloff E & Wiebe J. 2003. Learning extraction patterns for subjective expressions. In: D Yarowsky, T Baldwin, A Korhonen, K Livescu & S Bethard (eds). *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*. Seattle, Washington: Association for Computational Linguistics:105-112. <https://doi.org/10.3115/1119355.1119369>
- Risch JS. 2008. On the role of metaphor in information visualization. *The Computing Research Repository.arXiv preprint arXiv:0809.0884*: 1-20.
- Ritchie K. 2018. Vicki Momberg shouldn't be alone for long. *IOL*, 1 April. Available: <https://www.iol.co.za/sundayindependent/dispatch/vicki-momberg-shouldnt-be-alone-for-long-14183453> (Accessed 8 January 2019).
- Rob P, Coronel C, Crockett K & Morris S. 2013. *Database principles: Fundamentals of design, implementation and management*. London, UK: Cengage Learning EMEA.
- Robertson H & Travaglia J. 2015. Big data problems we face today can be traced to the social ordering practices of the 19th century. *Impact of Social Sciences Blog*:1-6. Available: <http://blogs.lse.ac.uk/impactofsocialsciences/2015/10/13/ideological-inheritances-in-the-data-revolution/> (Accessed 21 November 2018).
- Roelf W. 2018. South African parliament endorses report on disputed land reform. *Reuters*, 4 December. Available: <https://www.reuters.com/article/us-safrica-land/south-african-parliament-endorses-report-on-disputed-land-reform-idUSKBN1O31WL> (Accessed 8 January 2019).
- Rogers S. 2013. Twitter's languages of New York mapped. *The Guardian*, 21 February. Available: <https://www.theguardian.com/news/datablog/interactive/2013/feb/21/twitter-languages-new-york-mapped> (Accessed 4 December 2017).

References

- Rojas JAR, Kery MB, Rosenthal S & Dey A. 2017. Sampling techniques to improve big data exploration. Paper presented at the *2017 IEEE 7th symposium on large data analysis and visualization (LDAV)*, 2 October 2017. Phoenix, Arizona: IEEE:26-35. <https://doi.org/10.1109/LDAV.2017.8231848>
- Romero C & Ventura S. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12-27. <https://doi.org/10.1002/widm.1075>
- Roos J, Bart V & Statler M. 2004. Playing seriously with strategy. *Long Range Planning*, 37:549-568. <https://doi.org/10.1016/j.lrp.2004.09.005>
- Rose D. 2016. *Data science: Create Teams That Ask the Right Questions and Deliver Real value*. Berkeley, CA: Apress.
- Rosenberg A. 2010. Virtual world research ethics and the private/public distinction. *International Journal of Internet Research Ethics*, 3(12):23-27.
- Rosenholtz R, Li Y & Nakano L. 2007. Measuring visual clutter. *Journal of Vision*, 7(2): 17-17. <https://doi.org/10.1167/7.2.17>
- Rosenzweig P. 2012. Whither privacy?. *Surveillance & Society*, 10(3-4):344-347. <https://doi.org/10.24908/ss.v10i3/4.4528>
- Rourke L, Anderson T, Garrison DR & Archer W. 2001. Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education*, 12:8-22.
- Rousseeuw PJ & Leroy AM. 2005. *Robust regression and outlier detection*. Canada: John Wiley & Sons.
- Russom P. 2011. Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, 19(4):1-34.
- Ruths D & Pfeffer J. 2014. Social media for large studies of behavior. *Science*, 346(6213):1063-1064. <https://doi.org/10.1126/science.346.6213.1063>
- Sabherwal R & Becerra-Fernandez I. 2011. *Business intelligence: Practices, technologies and management*. USA: John Wiley & Sons.
- Sadkowsky T. 2014. Data scientists: The new rock stars of the tech world. *Techopedia*, 2 July. Available: <https://www.techopedia.com/2/28526/it-business/it-careers/data-scientists-the-new-rock-stars-of-the-tech-world> (Accessed 9 September 2018).
- Saghai Y. 2013. Salvaging the concept of nudge. *Journal of Medical Ethics*, 39(8):487-493. <https://doi.org/10.1136/medethics-2012-100727>
- Salah AA, Manovich L, Salah AA & Chow J. 2013. Combining cultural analytics and networks analysis: Studying a social network site with user-generated content. *Journal of Broadcasting & Electronic Media*, 57(3):409-426. <https://doi.org/10.1080/08838151.2013.816710>
- Salmon M. 2017. Emily Robinson, from social scientists to data scientist. *Forwards*, 2 February. Available: <https://forwards.github.io/blog/2017/02/07/emily-robinson-from-social-scientist-to-data-scientist/> (Accessed 11 October 2018).
- Santoso LW. 2017. Data warehouse with big data technology for higher education. *Procedia Computer Science*, 124:93-99. <https://doi.org/10.1016/j.procs.2017.12.134>
- Sartorius B, Jacobsen H, Törner A & Giesecke J. 2006. Description of a new all cause mortality surveillance system in Sweden as a warning system using threshold detection algorithms. *European Journal of Epidemiology*, 21(3):181-189. <https://doi.org/10.1007/s10654-005-5923-6>
- Savage M & Burrows R. 2007. The coming crisis of empirical sociology. *Sociology*, 41(5):885-899. <https://doi.org/10.1177/0038038507080443>

- Schilling PL & Bozic KJ. 2014. The big to do about “big data”. *Clinical Orthopaedics and Related Research*, 472(11):3270-3272. <https://doi.org/10.1007/s11999-014-3887-0>
- Schirrmacher F. 2015. *Ego: The game of life*. Translated by N Somers. Cambridge, UK: Polity Press.
- Schmidt E. 2010. Every 2 days we create as much information as we did up to 2003. *TechCrunch*, 4 August. Available: <https://techcrunch.com/2010/08/04/schmidt-data/> (Accessed 7 December 2017).
- Schnapp J, Presner T & Lunenfeld P. 2009. The digital humanities manifesto 2.0. Available: http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (Accessed 7 December 2017).
- Schöch C. 2013. Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, 2(3):2-13.
- Schroyens W & Schaeken W. 2003. A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(1): 140-149. <https://doi.org/10.1037/0278-7393.29.1.140>
- Schüssler R. 2005. On the anatomy of probabilism. In: JKraye & R Saarinen (eds). *Moral philosophy on the threshold of modernity* (Vol. 57). Dordrecht: Springer. 91-113. https://doi.org/10.1007/1-4020-3001-0_5
- Scott J. 2017. *Social network analysis*. Los Angeles, CA: Sage.
- Seddon JJ & Currie WL. 2017. A model for unpacking big data analytics in high-frequency trading. *Journal of Business Research*, 70:300-307. <https://doi.org/10.1016/j.jbusres.2016.08.003>
- Sekhotho K. 2018. SA's word of the year 2018 is.... *EWN*, 16 October. Available: <https://ewn.co.za/2018/10/16/sa-s-word-of-the-year-2018-is> (Accessed 8 January 2019).
- Selinger E & Whyte K. 2011. Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass*, 5(10):923-935. <https://doi.org/10.1111/j.1751-9020.2011.00413.x>
- Selwyn N. 2015. Data entry: towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1):64-82. <https://doi.org/10.1080/17439884.2014.921628>
- Serfass D, Nowak A & Sherman R. 2017. Big data in psychological research. In: RR Vllacher, SJ Read & A (eds). *Computational Social Psychology*. New York, NY and London, UK: Routledge. 332-348. <https://doi.org/10.4324/9781315173726-15>
- Serrano-Guerrero J, Olivás JA, Romero FP & Herrera-Viedma E. 2015. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18-38. <https://doi.org/10.1016/j.ins.2015.03.040>
- Shah DV, Cappella JN & Neuman WR. 2015. Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, 659(1):6-13. <https://doi.org/10.1177/0002716215572084>
- Shahrivari, S. & S. Jalili. 2014. Beyond batch processing: Towards real-time and streaming big data. *Computers*, 3(4): 117-119. <https://doi.org/10.3390/computers3040117>
- Sharda R, Delen D & Turban E. 2014. *Business intelligence and analytics: Systems for decision support*. Harlow, UK: Pearson Education.
- Shneidman ES & Farberow NL. (eds). 1957. *Clues to suicide* (Volume 56981). USA: McGraw-Hill Companies.
- Shoro AG & Soomro TR. 2015. Big data analysis: Apache spark perspective. *Global Journal of Computer Science and Technology*, 15(1):1-9. <https://doi.org/10.14445/22312803/IJCTT-V19P103>

References

- Shulman J. 2017. A good job for humanists. *The Andrew W. Mellon Foundation*, 6 February. Available: <https://mellon.org/resources/shared-experiences-blog/good-job-humanists/> (Accessed 4 September 2018).
- Shulman SW. 2014. Measuring reliability and validity in human coding and machine classification. Paper presented at the *CAQDAS Conference: Past, Present and Future – 25 Years of CAQDAS*. 1-3 May, 2014. East Horsley, Surrey, United Kingdom.
- Sibanda, O. 2012 'Social pain and social death': Poor white stigma in post-apartheid South Africa, a case of West Bank in East London. *Anthropology Southern Africa*, 35(3-4): 81-90. <https://doi.org/10.1080/23323256.2012.11500027>
- Silva S, Santos BS & Madeira J. 2011. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320-333. <https://doi.org/10.1016/j.cag.2010.11.015>
- Simons DJ. 2000. Attentional capture and inattention blindness. *Trends in Cognitive Sciences*, 4(4):147-155. [https://doi.org/10.1016/S1364-6613\(00\)01455-8](https://doi.org/10.1016/S1364-6613(00)01455-8)
- Simpson JA. 1989. Nathaniel Bailey and the search for a lexicographical style. In: G James (ed). *Lexicographers and their works*. Exeter: University of Exeter Press. 181-182.
- Singh S. 2000. *The code book: The secret history of codes and code-breaking*. London, UK: Fourth Estate.
- Singpurwalla ND & Landon J. 2014. Solving a system of high-dimensional equations by MCMC. In: SE Ahmed (eds). *Perspectives on big data analysis: Methodologies and applications*. USA: American Mathematical Society. 11-20. <https://doi.org/10.1090/conm/622/12437>
- Skolmen DE & Gerber M. 2015. Protection of personal information in the South African Cloud Computing environment: A framework for Cloud Computing adoption. Paper presented at *Information security for South Africa (ISSA) 2015*. 12-13 August, 2015. Johannesburg, South Africa: IEEE:1-10. <https://doi.org/10.1109/ISSA.2015.7335049>
- Slauter W. 2011. Write up your dead: The bills of mortality and the London plague of 1665. *Media History*, 17(1):1-15. <https://doi.org/10.1080/13688804.2011.532371>
- Slone DJ. 2009. Visualizing qualitative information. *The Qualitative Report*, 14:488-497.
- Smith N. 2017. Cultivating the technological imagination. *Cultural Studies*, 31(5): 712-714. <https://doi.org/10.1080/09502386.2015.1057857>
- Smith T. 2016. The software stack explained. *DZone/ToT Zone*, 19 May. Available: <https://dzone.com/articles/using-vr-to-test-urban-designs> (Accessed 21 May 2018).
- Snijders C, Matzat U & Reips UD. 2012. Big data: Big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, 7(1):1-5.
- Sparks R, Ickowicz A & Lenz HJ. 2016. An insight on big data analytics. In: N Japkowicz & J Stefanowski (eds). *Big data analysis: New algorithms for a new society*. Switzerland: Springer. 33-48. https://doi.org/10.1007/978-3-319-26989-4_2
- Spier F. 2014. Big history. In: D Northrop (eds). *A companion to world history*. USA: Wiley-Blackwell. 171-184. <https://doi.org/10.1002/9781118305492.ch11>
- Spitzberg BH & Gawron JM. 2016. Toward online linguistic surveillance of threatening messages. *Journal of Digital Forensics, Security and Law*, 11(3):43-78. <https://doi.org/10.15394/jdfsl.2016.1418>

- Srivastava P & Hopwood N. 2009. A practical iterative framework for qualitative data analysis. *International Journal of Qualitative Methods*, 8(1):76-84. <https://doi.org/10.1177/160940690900800107>
- Stadelmann, T., K. Stockinger, G.H. Bürki & M. Braschler. 2018. Data scientists. (In press). In: M. Braschler, TK Stadelmann & K Stockinger (eds). *Applied data science: Lessons learned from the data-driven business*. Springer. https://doi.org/10.1007/978-3-030-11821-1_3
- Staff reporter. 2014. The history of Internet access in South Africa. *MyBroadband*, 30 November. Available: <https://mybroadband.co.za/news/internet/114645-the-history-of-internet-access-in-south-africa.html> (accessed 10 December 2017).
- Stamp LD. 1965. *The geography of life and death*. 5th Edition. Ithaca, New York, NY: Cornell University Press.
- Stanley D. 2016. *Ada Lovelace, poet of science: The first computer programmer*. USA: Simon and Schuster.
- Steinhaus G. 2018. Trump Tweet on South African land overhaul draws government's ire. *The Wall Street Journal*, 23 August. Available: <https://www.wsj.com/articles/trump-tweet-on-south-african-land-reform-draws-governments-ire-1535017460> (Accessed 8 January 2019).
- Stenhaus B. 2017. Teaching data science is broken. *Towards Data Science*, 22 August. Available: <https://towardsdatascience.com/teaching-data-science-is-broken-4d551440df59> (Accessed 10 September 2018).
- Steyn AS. 2016. A new laager for a "new" South Africa: Afrikaans film and the imagined boundaries of Afrikanerdom. Unpublished Doctoral dissertation. Stellenbosch: University of Stellenbosch.
- Stockmann D. 2016. Towards area-smart data science: Critical questions for working with big data from China. *Policy & Internet*: 1-32. <https://doi.org/10.2139/ssrn.2718120>
- Stone M. 2009. Information visualisation: Challenge for the humanities. *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*: 43-56. Washington, DC: Council on Library and Information Resources.
- Stroop JR. 1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643-661.
- Storey VC & Song IY. 2017. Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering*, 108:50-67. <https://doi.org/10.1016/j.datak.2017.01.001>
- Stratton AK. 2012. The role of emotion in rational decision-making. Unpublished Doctoral dissertation. Australia: University of Adelaide.
- Stubbs, E. 2014. The value of business analytics. In: J Liebowitz (ed). *Business analytics: An introduction*. New York, NY: CRC Press. 1-28. <https://doi.org/10.1002/9781118983881.ch1>
- Stulpe A & Lemke M. 2016. Blended reading. In: M Lemke & G Wiedemann (eds). *Text Mining in den Sozialwissenschaften*. Wiesbaden: Springer. 17-61. https://doi.org/10.1007/978-3-658-07224-7_2
- Sula CA. 2012. Visualizing social connections in the humanities: Beyond bibliometrics. *Bulletin of the Association for Information Science and Technology*, 38(4):31-35. <https://doi.org/10.1002/bult.2012.1720380409>
- Sulis E., Fariás DIH, Rosso P, Patti V & Ruffo G. 2016. Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, 108:132-143. <https://doi.org/10.1016/j.knsys.2016.05.035>

References

- Sveningsson Elm M. 2009. How do various notions of privacy influence decisions in qualitative Internet research? In: AN Markham & NK Baym (eds). *Internet inquiry: Conversations about method*. Los Angeles, CA: Sage Publications. 69-87. <https://doi.org/10.4135/9781483329086.n7>
- Svolba G. 2017. *Applying data science: Business case studies using SAS*. Cary, NC: SAS Institute.
- Swart H. 2017. Social media: Big Brother, and his brother, are watching you via data mining. *Daily Maverick*, 11 October. Available: <https://www.dailymaverick.co.za/article/2017-10-11-social-media-big-brother-and-his-brother-is-watching-you-via-data-mining/#.Wqzxqk0h05t> (Accessed 17 March 2018).
- Swart H. 2018. Government surveillance of social media is rife. Guess who's selling your data? *Daily Maverick*, 25 April. Available: <https://www.dailymaverick.co.za/article/2018-04-25-government-surveillance-of-social-media-is-rife-guess-whos-selling-your-data/#.WuBJc4h97cs> (Accessed 25 April 2018).
- Szeman I. 2017. On the politics of extraction. *Cultural Studies*, 31(2-3):440-447. <https://doi.org/10.1080/09502386.2017.1303436>
- Taboada M. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325-347. <https://doi.org/10.1146/annurev-linguistics-011415-040518>
- Taboada M, Brooke J, Tofloski M, Voll K & Stede M. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267-307. https://doi.org/10.1162/COLI_a_00049
- Taylor E. 2013. *Surveillance schools*. Basingstoke: Palgrave Macmillan. <https://doi.org/10.1057/9781137308863>
- Taylor JE, Gregory IN & Donaldson CE. 2017. Combining close and distant reading: A multiscalar analysis of the English Lake District's historical soundscape. *International Journal of Humanities and Arts Computing*. Available: http://eprints.lancs.ac.uk/89167/1/IJHAC_REVISEDTaylorGregoryDonaldson_Sound.pdf (Accessed 5 April 2018).
- Taylor L. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2):1-14. <https://doi.org/10.2139/ssrn.2918779>
- Tembo T. 2018. 'White people in squatter camps': Google responds. *IOL*, 20 June. Available: <https://www.iol.co.za/capeargus/news/white-people-in-squatter-camps-google-responds-15468240> (Accessed: 21 June 2018).
- Thaler RH & Sunstein CR. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. USA: Yale University Press.
- Thamm M. 2018. The imperative of challenging the 'white genocide' and land expropriation narrative abroad. *The Daily Maverick*, 18 May. Available: <https://www.dailymaverick.co.za/opinionista/2018-05-18-the-imperative-of-challenging-the-white-genocide-and-land-expropriation-narrative-abroad/> (Accessed 9 March 2019).
- Thelwall M. 2010. Researching the public web. *eResearch Ethics*, 12 July. Available: <https://www.ehumanities.nl/researching-the-public-web/> (Accessed 25 April 2018).
- Thelwall M, Buckley K & Paltoglou G. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163-173. <https://doi.org/10.1002/asi.21662>
- Theunissen PS. 2015. Being 'Afrikaans': A contested identity. Paper presented at the *International Communication Association Annual Conference*, San Juan, Puerto Rico. 21-25 May 2015:1-7.

- Toonders J. 2014. Data is the new oil of the digital economy. *Wired*, July. Available: <https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/> (Accessed 2 December 2017).
- Tracy SJ. 2010. Qualitative quality: Eight 'big-tent' criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10):837-851. <https://doi.org/10.1177/1077800410383121>
- Tufekci Z. 2014. Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19(7):1-16. <http://www.firstmonday.dk/ojs/index.php/fm/article/view/4901/4097> (Accessed 17 March 2018). <https://doi.org/10.5210/fm.v19i7.4901>
- Tufte ER. 1986. *The visual display of quantitative information*. Cheshire, CT: Graphic Press.
- Tufte ER. 1990. *Envisioning information*. Cheshire, CT: Graphic Press.
- Tufte ER. 2006. *Beautiful evidence*. Cheshire, CT: Graphic Press.
- Tummons J. 2014. Using software for qualitative data analysis: Research outside paradigmatic boundaries. In: M Hand & S Hillyard (eds). *Big data? Qualitative approaches to digital research*. UK: Emerald Group Publishing Limited. 155-177. <https://doi.org/10.1108/S1042-319220140000013010>
- Turney PD. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics:417-424. <https://doi.org/10.3115/1073083.1073153>
- Tversky B. 2005. Visuospatial reasoning. In: K Holyoak & R Morrison (eds). *Handbook of reasoning*. Cambridge, UK: Cambridge University Press. 209-249.
- Vaisman A. & Zimányi E. 2014. *Data warehouse systems design and implementation*. Heidelberg: Springer-Verlag. <https://doi.org/10.1007/978-3-642-54655-6>
- van der Aalst, WM. 2014. Data scientist: The engineer of the future. In: K Mertins, F Bénaben, R Poler & JP Bourrières (eds). *Enterprise interoperability VI: Interoperability for agility, resilience and plasticity of collaborations* (Vol 7). Switzerland: Springer Science & Business Media. 13-26. https://doi.org/10.1007/978-3-319-04948-9_2
- van der Aalst WM. 2016. *Process mining: data science in action*. Heidelberg: Springer-Verlag. <https://doi.org/10.1007/978-3-662-49851-4>
- Van der Westhuizen, C. 2016. Afrikaners in post-apartheid South Africa: Inward migration and enclave nationalism. *HTS Theological Studies*, 72(4):1-9. <https://doi.org/10.4102/hts.v72i4.3351>
- van Dijk J. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2):197-208. <https://doi.org/10.24908/ss.v12i2.4776>
- van Dijk J. 2016. Big data, grand challenges: On digitization and humanities research. *KWALON*, 21(1):8-18. Available: <https://dspace.library.uu.nl/handle/1874/350785> (Accessed 5 April 2018).
- van Es K & Schäfer MT. 2017. New brave world. In: MT Schäfer & K van Es (eds). *The datafied society: Studying culture through data*. Amsterdam: Amsterdam University Press. 13-22. <https://doi.org/10.1515/9789048531011-003>
- van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, De Pauw G, Daelemans W & Hoste V. 2018. Automatic detection of cyberbullying in social media text. arXiv preprint arXiv:1801.05617. Available: <https://arxiv.org/pdf/1801.05617.pdf> (Accessed 18 October 2018).

References

- Van Hee C, Lefever E & Hoste V. 2018. We usually don't like going to the dentist. Using common sense to detect irony on Twitter. *Computational Linguistics*, pp.1-63. https://doi.org/10.1162/coli_a_00337
- van Holthoorn F. 2017. *A case for the Enlightenment, ten essays*. Berlin, Germany: Logos Verlag. on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, 32(1):89-97. <https://doi.org/10.1016/j.tele.2014.04.002>
- Vassakis K, Petrakis E & Kopanakis I. 2018. Big data analytics: Applications, prospects and challenges. In G Skourletopoulos, G Mastorakis, CX Mavromoustakis, C Dobre & E Pallis (eds). *Mobile big data: A roadmap from models to technologies*. Cham, Switzerland: Springer. 3-20. <https://doi.org/10.1016/j.tele.2014.04.002>
- Veinot TC. 2007. 'The eyes of the power company': Workplace information practices of a vault inspector. *The Library Quarterly*, 77(2):157-179. <https://doi.org/10.1086/517842>
- Verdinelli S & Scagnoli NI. 2013. Data display in qualitative research. *International Journal of Qualitative Methods*, 12(1):359-381. <https://doi.org/10.1177/160940691301200117>
- Verma A. 2018. Big data trends in 2018. *Whizlabs*, 22 January. Available: <https://www.whizlabs.com/blog/big-data-trends-in-2018/> (Accessed: 15 November 2018).
- Verwey C & Quayle M. 2012. Whiteness, racism, and Afrikaner identity in post-apartheid South Africa. *African Affairs*, 111(445):551-575. <https://doi.org/10.1093/afraf/ads056>
- Vieweg SE. 2012. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications. Unpublished Doctoral dissertation. Boulder, Colorado: University of Colorado.
- Visagie, R. 2018. Struggle(s) for self-determination: Afrikaner aspirations in the twenty-first century. Unpublished Master's dissertation. Stellenbosch: University of Stellenbosch
- Viseu A. 2015. Integration of social science into research is crucial. *Nature*, 525(7569):291. <https://doi.org/10.1038/525291a>
- Walker D & Dongarra J. 1996. MPI: A standard message passing interface. *Supercomputer*, 12:56-68.
- Walsh CG, Ribeiro JD & Franklin JC. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3):457-469. <https://doi.org/10.1177/2167702617691560>
- Wamba SF, Akter S, Edwards A, Chopin G & Gnanzou D. 2015. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234-246. <https://doi.org/10.1016/j.ijpe.2014.12.031>
- Wang Baldonado MQ, Woodruff A & Kuchinsky A. 2000. Guidelines for using multiple views in information visualization. In: M De Marsico, S Levialdi & E Panizzi (eds). *Proceedings of the Working Conference on Advanced Visual Interfaces*. 23-26 May 2000. New York, NY: ACM:110-119.
- Wang CH. 2016. A novel approach to conduct the importance-satisfaction analysis for acquiring typical user groups in business-intelligence systems. *Computers in Human Behavior*, 54:673-681. <https://doi.org/10.1016/j.chb.2015.08.014>
- Wang H, Can D, Kazemzadeh A, Bar F & Narayanan S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: M Zhang (ed). *Proceedings of the ACL 2012 System Demonstrations*. 8-14 July 2012. Stroudsburg, PA: Association for Computational Linguistics:115-120.

- Wang T. 2013. Big data needs thick data. *Ethnography Matters*, 13 May. Available: <http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/> (Accessed 24 April 2018).
- Wang Y & Hajli N. 2017. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, 70:287-299. <https://doi.org/10.1016/j.jbusres.2016.08.002>
- Wang Y, Kung L & Byrd TA. 2018. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3-13. <https://doi.org/10.1016/j.techfore.2015.12.019>
- Warden P. 2011. *Big data glossary*. USA: O'Reilly Media, Inc.
- Warrell JG & Jacobsen M. 2014. Internet research ethics and the policy gap for ethical practice in online research settings. *The Canadian Journal of Higher Education*, 44(1):22-37.
- Watson HJ. 2014. Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34(1):1247-1268. <https://doi.org/10.17705/1CAIS.03465>
- Watt IP. 1957. *The rise of the novel: Studies in Defoe, Richardson and Fielding*. Berkeley, CA: University of California Press.
- Weiss SM & Indurkha N. 1998. *Predictive data mining: A practical guide*. USA: Morgan Kaufmann Publishers, Inc.
- Wertheimer M. 1938. Gestalt theory. In: WD Ellis (ed.). *A source book of Gestalt psychology*. London, UK: Routledge and Kegan Paul. 71-88.
- Wheeler C. 2016. Big history in South Africa? *Pascaptrust*, 23 May. Available: <http://pascap.org.za/big-history-in-south-africa/> (Accessed 20 March 2018).
- White T. 2015. *Hadoop: The definitive guide*. Sebastopol, CA: O'Reilly Media, Inc.
- Whiteman N. 2010. Control and contingency: Maintaining ethical stances in research. *International Journal of Internet Research Ethics*, 3(12):6-22.
- Whitmore A, Agarwal A & Da Xu L. 2015. The Internet of Things – A survey of topics and trends. *Information Systems Frontiers*, 17(2):261-274. <https://doi.org/10.1007/s10796-014-9489-2>
- Whyte JK, Ewenstein B, Hales M & Tidd D. 2007. Visual practices and the objects used in their design. *Building Research & Information*, 35(1):18-27. <https://doi.org/10.1080/09613210601036697>
- Wilkinson K. 2017. Comment: Does AfroForum care about getting farm murder statistics right? *Africa Check*, 24 November. Available: <https://africacheck.org/2017/11/24/comment-does-afriforum-care-about-getting-farm-murder-statistics-right/> (Accessed 7 March 2019).
- Williams ML & Burnap P. 2016. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211-238. <https://doi.org/10.1093/bjcr/azv059>
- Williamson B. 2017. *Big data in education: The digital future of learning, policy and practice*. London, UK: Sage Publications.
- Willis III JE, Campbell J & Pistilli M. 2013. Ethics, big data, and analytics: A model for application. *EDUCAUSE Review*, 6 May. Available: <https://er.educause.edu/articles/2013/5/ethics-big-data-and-analytics-a-model-for-application> (Accessed 8 May 2018).
- Wills T. 2016. Social media as a research method. *Communication Research and Practice*, 2(1):7-19. <https://doi.org/10.1080/22041451.2016.1155312>
- Woodie A. 2015. The humanist's emerging role in big data. *Datanami*, 21 January. Available: <https://www.datanami.com/2015/01/21/humanists-emerging-role-big-data/> (Accessed 7 March 2019).

References

- Woods M, Paulus T, Atkins DP & Macklin R. 2016. Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. *Social Science Computer Review*, 34(5):597-617. <https://doi.org/10.1177/0894439315596311>
- Wolfinger E. 2016. “But it’s already public, right?": The ethics of using online data. *Data Driven Journalism*, 25 November. Available http://datadrivenjournalism.net/news_and_analysis/but_its_already_public_right_the_ethics_of_using_online_data (Accessed 5 October 2017).
- Wortmann F & Flüchter K. 2015. Internet of things. *Business & Information Systems Engineering*, 57(3):221-224. <https://doi.org/10.1007/s12599-015-0383-3>
- Wu C, Buyya R & Ramamohanarao K. 2016. Big data analytics= Machine learning+ Cloud computing. *arXiv preprint arXiv:1601.03115*:1-27. Available <https://arxiv.org/ftp/arxiv/papers/1601/1601.03115.pdf> (Accessed 4 November 2017).
- Yang H, Willis A, De Roeck A & Nuseibeh B. 2012. A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights*, 5:17-30. <https://doi.org/10.4137/BII.S8948>
- Yoder-Wise PS & Kowalski K. 2003. The power of storytelling. *Nursing Outlook*, 51(1): 37-42. <https://doi.org/10.1067/mno.2003.2>
- Youngman PA & Hadzikadic M (eds). 2014. *Complexity and the human experience: Modeling complexity in the humanities and social sciences*. Singapore: Pan Stanford Publishing. <https://doi.org/10.1201/b16877>
- Youtie J, Porter AL & Huang Y. 2017. Early social science research about big data. *Science and Public Policy*, 44(1):65-74. <https://doi.org/10.1093/scipol/scw021>
- Zacharis NZ. 2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27:44-53. <https://doi.org/10.1016/j.iheduc.2015.05.002>
- Zhang C, Zeng D, Li J, Wang FY & Zuo W. 2009. Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474-2487. <https://doi.org/10.1002/asi.21206>
- Zhang S. 2016. Scientists are just as confused about the ethics of big-data research as you. *Wired*, 20 May. Available: <https://www.wired.com/2016/05/scientists-just-confused-ethics-big-data-research/> (Accessed 11 December 2017).
- Zhang Q, Yang LT, Chen Z & Li P. 2018. High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. *Information Fusion*, 39:72-80. <https://doi.org/10.1016/j.inffus.2017.04.002>
- Zhu Y. & Xiong Y. 2015. Towards data science. *Data Science Journal*, 14(8):1-7. <https://doi.org/10.5334/dsj-2015-008>
- Zimmer M. 2010. “But the data is already public”: On the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313-325. <https://doi.org/10.1007/s10676-010-9227-5>
- Zimmer M. 2016. OkCupid study reveals the perils of big-data science. *Wired*, 14 May. Available: <https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/> (Accessed 11 December 2017).
- Zuze TL & Reddy V. 2014. School resources and the gender reading literacy gap in South African schools. *International Journal of Educational Development*, 36:100-107. <https://doi.org/10.1016/j.ijedudev.2013.10.002>
- Zwitter A. 2014. Big data ethics. *Big Data & Society*, 1(2):1-6. <https://doi.org/10.1177/2053951714559253>

Index

- ACID 100
- ADAMS (Anomaly detection at multiple scales) 2
- AFINN 134
- algorithmic solution(s) 27
- Alibaba 114
- Amazon 2, 100, 116
- analogical reasoning 58, 63-64
- Apache Hadoop 100, 112, 116-117
- API (Application programming interface) 39, 42, 128
- apophenia 38
- Apple 21, 46
- application layer 100-101, 107
- ATLAS.ti 87, 89, 93, 96
- BDaaS (Big data as a service) 117
- BI (Business intelligence) 7-8, 11-12, 78, 111, 115
- big data**
 - big data 1-7, 9-32, 34-49, 52-55, 57-58, 64, 71-89, 91, 94-107, 110-111, 113-121, 124-126, 132, 140-141
 - big data analytics 3, 23, 40, 71, 79-82, 85, 102, 141
 - big data architecture 4, 82, 98, 107
 - big data failures/faux pas 45
 - big data generation 102
 - big data storage 7, 104
 - big data systems 100-101, 110, 114, 116
- Big Table 105
- blended reading 92, 96
- BO (Business Objects) 115
- Boolean algebra 9
- capta 66-67, 69
- Cassandra 105, 115
- causal inferences 77
- causality 41
- CERN (European Organisation for Nuclear Research) 102
- Chukwa 114
- clickstream analysis 108, 114
- close reading 25, 76, 92-94
- Coca-Cola 2

- Colossus 105
- column-oriented stores 105
- complexity science 27-28
- computational social science**
 - computational social science 26-27, 41-42
 - computational social scientist(s) 24, 26-27
- computer science**
 - computer science 3, 12, 37, 49, 58, 79, 99-100, 121, 123-124
 - computer scientist(s) 6, 121, 141
- computing layer 101, 107
- content analysis 23, 93, 95-96
- correlation(s) 35-36, 38, 47, 73, 85
- Couchbase 106
- Cypher Query Language 106
- DARPA (Defense Advanced Research Projects Agency) 2
- data**
 - data acquisition 34, 103, 113
 - data brokers 54-55
 - data cleaning 91
 - data cleansing 104, 118, 131
 - data deluge 8, 12, 18, 20, 30, 83
 - data-driven 3, 18, 20, 40-41, 82, 105, 132, 141
 - datafication 11, 18, 26, 54, 72, 78
 - datafied 66, 71
 - data-intensive science 99
 - data justice 55-56
 - data mining 38, 54-55, 73, 107, 109, 118
 - data power 4, 56, 71-73
 - democratisation of data science 125
 - raw data 34, 103, 111
 - real-time data 98, 103, 114, 141
 - unstructured data 31, 38, 98-99, 103, 108, 111, 113
- data scientist(s)**
 - big data scientist(s) 3-4, 24, 27, 29, 121
 - fake data scientist(s) 121
 - horizontal data scientist(s) 120-121
 - unicorn data scientist(s) 120
 - vertical data scientist(s) 120-121
- descriptive analytics 13, 109

digital

- digital ecosystem 45, 54
- digital humanitarianism 83
- digital humanities 3, 19-20, 24-28, 41-42, 65, 67, 76, 80, 123
- disinformation 83, 140
- distant reading 25, 57, 92-94
- distributed file systems 101, 104-105, 107
- document databases 105

domain

- domain 40, 43-44, 47-48, 58, 63-64, 75, 93-94, 121, 134
- domain expertise 121
- domain knowledge 3, 8, 11, 18, 22-24, 26, 31, 35-36, 40, 42, 54, 58, 62-63, 66-67, 69-71, 76, 109, 116-117, 121-125, 132, 141
- Dropbox 91
- Dynamo 105
- Elongated words 131

empiricism

- empirical methods 19
- empiricism 18, 35, 40-41

ethics

- ethical considerations 4, 48, 56
- ethics 27, 43-46, 48-51, 53, 55, 124
- exhaustivity 31, 85
- Facebook 2, 16, 21, 46-47, 51-54, 86, 100, 102, 105, 113
- fire-hose 39
- Flume 113
- fourth paradigm 2, 99-100
- garden-hose access 39
- generalisability 36, 84
- Gephi 111-112

Gestalt theory

- Gestalt principles 58, 61
- Gestalt theory 61
- gold rush metaphor 21

Google

- GFS (Google File System) 104-105, 107
- GFT (Google Flu Trends) 30, 43
- Google 2, 9, 13, 16, 30, 43, 48, 51, 54, 83, 91, 100, 103-105, 111, 128, 138
- Google Cloud Platform 91
- Google Image search 138
- Google Sheets 103

Google Translation 128
graph databases 105-106
GraphLab 107
GraphX 114
Groupon 114
Hadoop 13, 100-101, 105, 110-117
Hashtag Sentiment Lexicon 133-135
Haystack 104-105
HBase 105, 110, 113, 115
Hive 100, 110, 113-115
humanist(s) 1-4, 6, 19-20, 23-31, 34, 36-37, 41-42, 49, 57, 64-67, 71, 75, 83, 96, 98, 117-118, 124-125, 140-141
HunchWorks 2
IBM (International Business Machines Corporation) 9, 11-12, 99-100, 116
ICT (Information and communications technology) 83, 100-101, 116
infrastructure layer 101, 107, 116
Intel 113, 116
IoT (Internet of Things) 102-103, 117
IR (Information retrieval) 108
Java 106, 111, 113
JSON (JavaScript Object Notation) 105
kbps (Kilobytes per second) 14
key-value stores 105
lexicon-based approach (to sentiment analysis) 132-133
linear algebra 123
linguistics 76, 80-81, 92, 107, 123, 132
Logistic Regression 82, 133
machine learning 1, 78, 80, 82, 107, 110-111, 114, 118, 121, 123-124, 132-133
Mahout 110
MAPD (Mathematics for the analysis of petascale data) 2
MapReduce 100-101, 107, 110-114
markup language
 HML (Hypertext markup language) 108
 XML (Extensible markup language) 103
MaxENT (Maximum Entropy) 133
MAXQDA 87, 89, 93-94, 96
megabits 14
metadata 16, 52, 77, 80, 109, 111
Microsoft Azure 116
Microstrategy 115
misinformation 83, 140

- MMP (Massively parallel-processing) 110
- MongoDB 106
- MPI (Message passing interface) 107
- multimedia analytics 107, 109
- Naive Bayes 133
- Neo4j 106
- Neoliberal**
 - Neoliberal 71-72
 - Neoliberal fatalism 72
 - Neoliberalisation 24
 - Neoliberal political project 71
- NLP (Natural language processing) 1, 80, 108, 124, 131-132
- Noisy text 131
- NoSQL 101, 104-107, 110, 113
- NRT (Near real-time) 98
- NSA (National Security Agency) 2, 73
- nudging 73
- NVivo 89, 91, 93, 96
- OLAP (Online analytical transaction processing) 109
- OneDrive 91
- OpenMP 107
- Opinion Lexicon 133-134
- Oracle 100, 113, 116
- OrientDB 106
- over-coding 96
- Pattern for Python 134
- Pig 110, 113-114
- PoPI (Protection of personal information) 48, 52, 73
- predictive analytics 13, 54-55, 109-110, 141
- Pregel 107
- pre-processing 103-104, 126, 131, 133
- Prescriptive analytics 13, 109
- PRISM (Planning tool for resource integration, synchronization and management) 2
- privacy 26, 46-48, 52, 55-56, 74, 81, 125
- probability**
 - probabilism 10
 - probabilistic 11, 16, 62, 108
 - probability 10, 13, 55, 62, 74, 123
 - probability calculus 10
- programming models 104, 107, 110
- public-private space 45, 50

- Python 111, 113, 123, 133-134
- QDA Miner 89, 96
- QDAS (Qualitative data analysis software) 4, 87-89, 91-93, 95-97
- Random Forests 133
- RDMS (Relational database management system) 99
- real-time**
 - real-time data 98, 103, 114, 141
 - real-time streaming 114
- reductionist 17, 23, 26, 42
- reliability 87, 95
- representativeness 27, 45, 53-54
- research**
 - qualitative research 22-23, 29, 39, 87-88, 95
 - quantitative research 26, 64
 - research 1-6, 9, 12, 14, 18-20, 22-24, 26, 28-31, 34-51, 53-54, 57, 64, 71-72, 76-82, 84-88, 90, 92-93, 95-96, 98, 102, 104, 109, 112, 117-120, 123-125, 133, 141
- Ruby 111
- S4 107
- SAP HANA 116
- scalability 99, 113
- semi-structured data 103
- sentiment analysis 91-92, 108, 126-127, 132-134
- SLA (Service level agreements) 100
- small data**
 - small data 16, 18, 29, 37, 84-86, 94, 98, 141
 - small datasets 42-43, 89
- social network**
 - social network analysis 78, 112
 - social network(ing) sites 23
- South African National Research Network 14
- Spark 13, 100-101, 110, 114, 116
- spreadsheets 99, 123
- SPSS (Statistical package for the social sciences) 87
- SQL (Structured query language) 99-100, 113-114, 123
- Sqoop 110, 113, 115
- statistical analysis 7, 10-11, 13, 101, 107, 109
- Stock Sonar 132
- StockTwits 132
- stop words 131

storage

- DAS (Direct attached storage) 101, 104
- NAS (Network attached storage) 101, 104
- SAN (Storage area network) 101, 104
- Storm 13, 107, 114-115
- storytelling 121
- Stroop effect 59-60
- structured data 99, 102-103, 107-109, 113-114
- suicide 80-82
- SVM (Support Vector Machine) 133
- Tableau 87, 111, 115
- Tempora 2
- text mining 1, 40, 80, 107-108
- threat actor(s) 138
- tokenisation 131
- transparency 42-43, 48, 93, 95, 125
- trustworthiness 31, 95
- Twitter 21, 39, 43, 45, 51-54, 78-79, 86, 114, 128, 131-133, 135-136, 138, 140
- UGC (User-generated content) 131-132
- validity 27, 36, 84, 86-87, 95
- variety 20, 27, 31, 85, 88-89, 96, 98, 103, 120
- velocity 31, 85-86, 89, 98, 100-102
- veracity 31, 89, 99
- visualisation 4, 11, 24-25, 31, 40, 56-61, 64, 66-67, 69-71, 80, 82, 88, 93, 99, 107, 109, 111-112, 115, 121
- Voldemort 105
- volume 3, 21, 24, 29, 31, 36, 85-86, 89, 96-98, 100, 106
- Walmart 2, 102
- web mining 107-108, 124
- We Feel 44, 132
- white space(s) 131
- word cloud 90-91
- word processing 123
- Yahoo 100, 113-114
- YARN 101, 110, 112-113
- ZooKeeper 114

This book explores the big data evolution by interrogating the notion that big data is a disruptive innovation that appears to be challenging existing epistemologies in the humanities and social sciences. Exploring various (controversial) facets of big data such as ethics, data power, and data justice, the book attempts to clarify the trajectory of the epistemology of (big) data-driven science in the humanities and social sciences.

Susan Brokensha is a Senior Lecturer in the Department of English at the University of the Free State and has a PhD in Applied Language Studies. As a teacher, she is passionate about how the pedagogical value of learning management systems may be exploited to enhance teaching and learning. As a researcher, she is particularly interested in both linguistic and non-linguistic aspects of behaviour and communication in offline and online contexts. Against the backdrop of the fourth industrial revolution, she has recently begun interrogating how big data tools and technologies reflect perils and possibilities for scholars carrying out qualitative research in the humanities and social sciences.

Eduan Kotzé holds a PhD in Computer Information Systems and is a Senior Lecturer and the Academic Departmental Head of the Department of Computer Science and Informatics at the University of the Free State. Dr Kotzé has over 15 years technical and managerial experience in Information Technology, specialising in database management systems, data warehousing, business intelligence, and text mining. His research focuses mainly on algorithms and methods to process big datasets while employing a data science approach. These datasets are predominately in unstructured natural language text formats.

Burgert Senekal has been associated with the University of the Free State since 2008. After completing his Master's degree in Afrikaans, Dr Senekal completed a Master's degree in English on contemporary British fiction. In 2013, he obtained his PhD on counterinsurgency literature at the University of the Free State. He has published more than 50 peer-reviewed articles and is an NRF-rated researcher. His research interests include alienation, information technology, big data, data science, and complex networks.