BIG DATA MINING AND COMPLEXITY

BRIAN C. CASTELLANI
RAJEEV RAJARAM

THE SAGE QUANTITATIVE RESEARCH KIT

Contents

1

Introduction

1.1 The Joys of Travel

For those who know us, it goes without saying that one of the things we (and our families) enjoy most in life is travel. And the more often we get to do so, the better. There is little in life like traveling somewhere to catch a glimpse, albeit briefly, of how other people live. As Rick Steves, the travel guru, says, 'Travel is freedom... one of the last great sources of legal adventure' [1]

Traveling, however, is not the same as taking a vacation. For Rick Steves – and for many of the travel bloggers we follow [2] – while vacations are great, traveling is different. As a photojournalist recently put it, if your trip photos are mostly selfies, you took a vacation; if your photos are of the places you visited, then you traveled. Vacations are about relaxing, which we all regularly need. But travel is about taking an adventure, which we also need. And we need travel because it pushes us to see the world and our lives in new and different ways. As Rick Steves says, 'Travel destroys ethnocentricity. It helps you understand and appreciate different cultures. Travel changes people.'[3]

1.2 Data Mining and Big Data Travel

Rick Steve's philosophy of travel is also true of the current book, insomuch as it is a travel-log of our adventures into the fields of data mining and big data, which we seek to share with readers; all in an effort to see if we, together, can weave a new way of understanding the planet and ourselves *vis-a-vis* the global complexities of the data saturated world(s) in which we live.

The challenge, however, is getting to our destination. A few years back, after some eighteen hours of travel from the States – including two plane delays, a several-hour layover in New York City, a cramped international flight, several tube connections, two UK trains, and a taxi ride – we arrived exhausted in northern England for a research seminar. Fortunate for us, we had several of our British friends awaiting us with dinner and drinks. At which point, one of them asked, 'Why on earth do you do it?' After a bit of a pause, one of us (Rajeev) replied, 'We like traveling, we just don't like getting there.'

We think it fair to say that, while people enjoy traveling to new places – be it to learn new methods, new ideas, or experience new frontiers in research – the biggest hurdle is getting there. Hence our book's more technical purpose: we seek to make the journey into the new world of big data and data mining as painless as possible, knowing that the journey, while worth it, presents a series of challenges. We have organised these challenges into two major journeys – that is, parts 1 and 2 of the current book.

1.3 Part 1: Thinking Complex and Critically

The first theme concerns thinking about big data and data mining from a complex and critical perspective. The worlds of data and method have changed, expanding far beyond the confines of conventional research (Burrows and Savage, 2014, Veltri, 2017). But, because of this expansion, data and method have also broadened in their usage, becoming part and parcel of the daily life of most companies and public-sector organisations (Raghavan, 2014). In other words, data and method are no longer under the strict purview of academia. In fact, one is more likely to read about the cutting-edge of big data in Wired than most journals in the social sciences (Cukier and Mayer-Schoenberger, 2013). As a result, knowledge of big data and data mining within academia varies considerably (Castellani, 2014). For example, while the social sciences continue to study relatively static datasets using conventional linear statistics, other fields such as physics regularly study highly dynamic temporal/spatial datasets using the latest advances in data mining (Castellani et al., 2016).

And, it is here, with this imbalance in awareness, particularly amongst the social sciences, that we arrive at our first major challenge. In fact, some go so far as to call it a crisis (Savage and Burrows, 2007, Burrows and Savage, 2014), which they articulate as follows: the significant variance in knowledge of the tools of data mining and computational modelling and big data leaves many within the social sciences disadvantaged and discredited when it comes to the complex and critical discussions surrounding our currently data-saturated globalised world(s) – which is a problem for all involved, as social science is critical to such discussions (Burrows and Savage, 2014).

For example, many experts in data mining and big data see their respective fields as more than a simple advance on method, treating data science instead as an epistemological transformation, constituting an entirely new approach for data acquisition, management, modelling, analysis, output, results and decision-making. A paradigm shift, if you will, in social scientific thinking (Kitchin, 2014). Some critics, however, push back, arguing that data mining and big data are empty buzzwords for little more than the latest trend in data management or methodological technique – for a review of these critiques, see (Kitchin, 2014). Other critics, however, take this 'paradigm shift' claim seriously, arguing that while data mining and big data are touted as intellectual revolutions, they are at best only useful (albeit limited) attempts to deal more effectively with the data saturated global world in which we now live. And, while some of these new attempts are innovative and thoughtful, others are not; which is key to why social scientists need to be involved in such discussions (Byrne and Callaghan, 2013). In fact, some of the latest advances in data mining and big data are seen as downright dangerous and foolhardy, potentially leading to terrible outcomes and decision-making (Mahnke and Uprichard, 2014). However, to be fair, many of the current conventions for data acquisition, management, analysis and modelling – in particular statistics – are often equally foolhardy and dangerous; hence part of the reason data mining and big data emerged in the first place. Put simply, on both sides of the debate we need better scientific tools (Castellani, 2014).

1.3.1 Organisation of Part I

So, we are left with a challenge, which Part I of our book was written to address. For this section, our goal is to determine critically what it is about data mining and big data that is useful, and to what extent, and in what ways, and in what contexts? Part I is therefore organised as follows:

- Chapter 2 begins with a critique of conventional method in the social sciences, specifically statistics. The goal is to explain how and why data mining and big data are presently overwhelming the social sciences; as well as exploring the consequences of the social sciences not overcoming this problem – from policy and polity to economy and scientific practice.
- Chapter 3 provides a quick but similarly critical overview of the field of big data and the conflicted arguments surrounding its development.
- Chapter 4 does the same with data mining.
- Chapter 5, finally, provides a fast survey of the complexity sciences, mainly to demonstrate (albeit, again, critically) the utility of this field for providing – at least to us – the best methodological framework for advancing the data mining of big data.

The purpose of Chapter 5 takes us to our next point. For us, while the term big data has proven, on some levels, to be useful, in the end it is too simplistic, as it suggests that the only real difference in the world of data, circa 2020, is more of it. But, as Uprichard points out (2013), the data saturated world(s) in which we all presently live are not just big; they are complex. In other words, big data today are comprised of a multitude of different factors (e.g., ecological, geographical, social, economic, political, psychological, medical, etc), which are distributed, interdependent, multi-level (macroscopic to microscopic), dynamic and evolving (often in real-time), and spatial, self-organising, emergent, nonlinear, and network-like in their organisation. In short, data are complex.

Equally important, the globalised world(s) these data 'represent' are likewise complex (Capra and Luisi, 2014). Case in point: there are few social science topics that do not sit at the intersection of multiple data sets, governmental agencies or areas of concern. For example, food safety in a metropolitan area links to poverty, which connects with the region's ecology and infrastructure, which connects to its economy and political stability; which, in turn, links to other such issues as inequality and racism and access to education and women's rights and so forth.

In terms of understanding such complex issues what we need, then, is an epistemological and ontological shift in thinking. As Stephen Hawking famously quipped, science in the 21st century is all about complexity. More specifically, science needs to embrace a complex systems view of the world (Byrne and Callaghan, 2013, Capra, 1996, Mitchell, 2009). As Capra and Luisi state, 'As the twenty-first century unfolds, it is becoming more and more evident that the major problems of our time – energy, the environment, climate change, food security, financial security – cannot be understood in isolation. They are systemic problems, which means that they are all interconnected and interdependent (2014, xi).' Hence the need for the complexity sciences, which are fast becoming the guiding framework and critical carrier for the journey of many scholars into the worlds of data mining and big data. All of which takes us to the second section of our book.

1.4 Part 2: Learning New Tools and Techniques

In terms of the challenges and hurdles associated with our journey, the second theme revolves around learning new tools and techniques. No matter one's background, including mathematics, learning new methods is always hard going; particularly when it comes to data mining and big data – which have, over the last few decades, amassed into a rather significant number of new approaches. As shown in our book's index, this list ranges from machine intelligence and textual analysis to geospatial modelling and network analysis. Still, even with such a long list, learning about these methods need not be any more difficult than necessary.

Hence our rationale for Part 2 of our book, which seeks to provide readers two things. First, we seek to provide an ontological view of complexity, as seen through the lens of data mining and big data. And, second, we seek to provide a set of user-friendly mathematical formalisms that, once reviewed, should prove helpful in making sense of the wider fields of data mining and computational modelling. To accomplish these two tasks, we will ground the whole of our review within the framework of *case-based complexity* and its methodological ontology (Byrne and (eds.), 2013, Castellani et al., 2016).

By way of a brief introduction, case-based complexity represents one of the main avenues of research – as well as one of the most developed views on social complexity – within the complexity sciences. As Byrne and Callaghan note (2013), case-based complexity is based on the key insight that, as pertains to social life, the cases we study are best viewed in complex systems terms; and, in turn, the complex social topics (a.k.a social systems) we study are best modelled as cases or sets of cases. Let us explain.

Regardless of the technique used, data mining and big data analytics are ultimately about modelling, exploring, clustering or cataloguing cases, based on key characteristics or etiological differences. For example, smart machines can be used to identify tumor or disease types; predictive analytics can explore public policies and their multiple outcomes; artificial intelligence can identify reliable investment opportunities; genetic algorithms can detect subtle changes in weather or traffic patterns; and network analyses can find the fastest route from a search question to its answer. And all of them (albeit to varying degrees) can be counted as an improvement on conventional statistics, mainly because they avoid aggregate-based one-size-fits-all solutions; focusing, instead, on identifying multiple case-based trends; which, in turn, they catalogue and examine based on differences in their respective profile of key factors. In short, all of these techniques treat the topics they study as evolving sets of complex cases.

The problem, however – which brings us to the heart of the challenge in Part 2 – is that few of the authors of these tools and techniques identify their work as such. Nor do they note their technique's similarity with other data mining and computational modelling techniques. For example, if you go to the SAS website [4] – a leading software package for statistical and computational analysis – it defines predictive analytics as 'the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data.' In turn, it defines data mining as 'the process of finding anomalies, patterns and correlations within large data sets to predict outcomes' – again, based on a broad range of techniques. As the reader can see, these two definitions, minus a few words of emphasis, are otherwise identical, as the goal in both instances is to group cases, based on profile differences, for the purposes of prediction. More important, they basically use almost the same set of techniques: machine intelligence, artificial neural nets, etc.

A similar argument can be made of machine learning and artificial intelligence. For example, according to Wikipedia, 'Artificial intelligence (AI, also machine intelligence, MI) is intelligence exhibited by machines, rather than humans or other animals.' In other words, machine learning is really artificial intelligence, which is really part of predictive analytics, which is really part of data mining.

In short, and as these examples hopefully illustrate, while the techniques and tools of data mining and computational modelling are numerous, they are also highly similar in focus and design. And that is a good thing, as readers will see, as it allows us to create, based on the framework of case-based complexity, a mathematical and methodological shortcut. Hence, again, the purpose of Part 2, which is organised as follows:

- Chapter 6 provides readers with an ontology and set of mathematical formalisms for thinking about data mining and big data, based on the theoretical framework of case-based complexity.

- Chapter 7 overviews the techniques of classification, including cluster analysis.

- Chapter 8 reviews machine intelligence and machine learning, with specific emphasis on neural nets, including the famous Kohonen topographical self-organising map.

- Chapter 9 examines the tools of predictive analytics, including Bayesian statistics, decision trees, and regression.

- Chapter 10 deals with longitudinal and temporal data analysis, with specific emphasis on differential equations, dynamical systems theory and growth mixture modelling.

- Chapter 11, in turn, deals with geospatial data, exploring the techniques used to collect such data and, in turn, analyse them.

- Chapter 12 deals with text and video mining, including such techniques as sentiment analysis, issue mapping and fitness landscapes.

- Chapter 13 addresses the topic of complex networks.

Before proceeding, however, two caveats are necessary.

1.4.1 Sage Quantitative Methods Kit

While the number of techniques reviewed in Part 2 is somewhat exhaustive, our summary of these methods is by no mean in-depth. In other words, our primary goal in this book is to use our mathematical/methodological shortcut to demonstrate the continuity and inter-linkage of these data mining techniques *vis-a-vis* the challenges of modelling and studying complex datasets. Also, given the breadth of our review involved in Part 2 of our book, we will not have time to delve into such important big data details as how to best build a big data database or run a specific technique. Such concerns (along with a more in-depth analysis of specific techniques) is, however, a major goal of the *SAGE Quantitative Research Kit*. As such, we recommend readers explore these books as well. All of which takes us to our next caveat.

1.4.2 COMPLEX-IT and the SACS Toookit

In addition to reading the other books in the *Sage Quantitative Methods Kit*, for those interested in taking the next step toward actually data mine their own big data, in Chapter 6 we provide a brief introduction to the SACS Toolkit and COMPLEX-IT. The utility of this combined methodological platform is that, while the SACS Toolkit provides the methodological framework for employing most of the tools and techniques we review in Part 2, COMPLEX-IT provides a free R-studio software package for running and integrating several of them. As such, the SACS Toolkit and COMPLEX-IT function as a methodological/software companion to the current book. (For more on COMPLEX-IT, see `https://www.art-sciencefactory.com/complexit.html`; and for more on the SACS Toolkit, see `https://www.art-sciencefactory.com/cases.html`.)

1.5 The Airline Industry: A Case Study

In addition to the mathematical formalisms of case-based complexity – and their corresponding software – we thought another way to tie the book together is to pick a case study that would have wide appeal. And, given our focus on traveling, what better example than one with which most are familiar – the airline industry! In terms of a complex system of study that globally exemplifies the challenges of data mining and big data, one could hardly pick a better topic than the airline industry. Or, at least, that is the conclusion that we (Brian and Rajeev) reached one day while sitting in our favorite teashop in Oxford, UK. The name of the place is *Cafe Loco*, themed on the famous Mad Hatter's tea party. Turns out that Brian's brother, Warren, was with us that day, as he was part of our long-delayed trip to Northern England, which we mentioned above, and was equally perplexed by the complexities of international travel. Which got us thinking: how does such a complex global system work, and with all that complex big data?

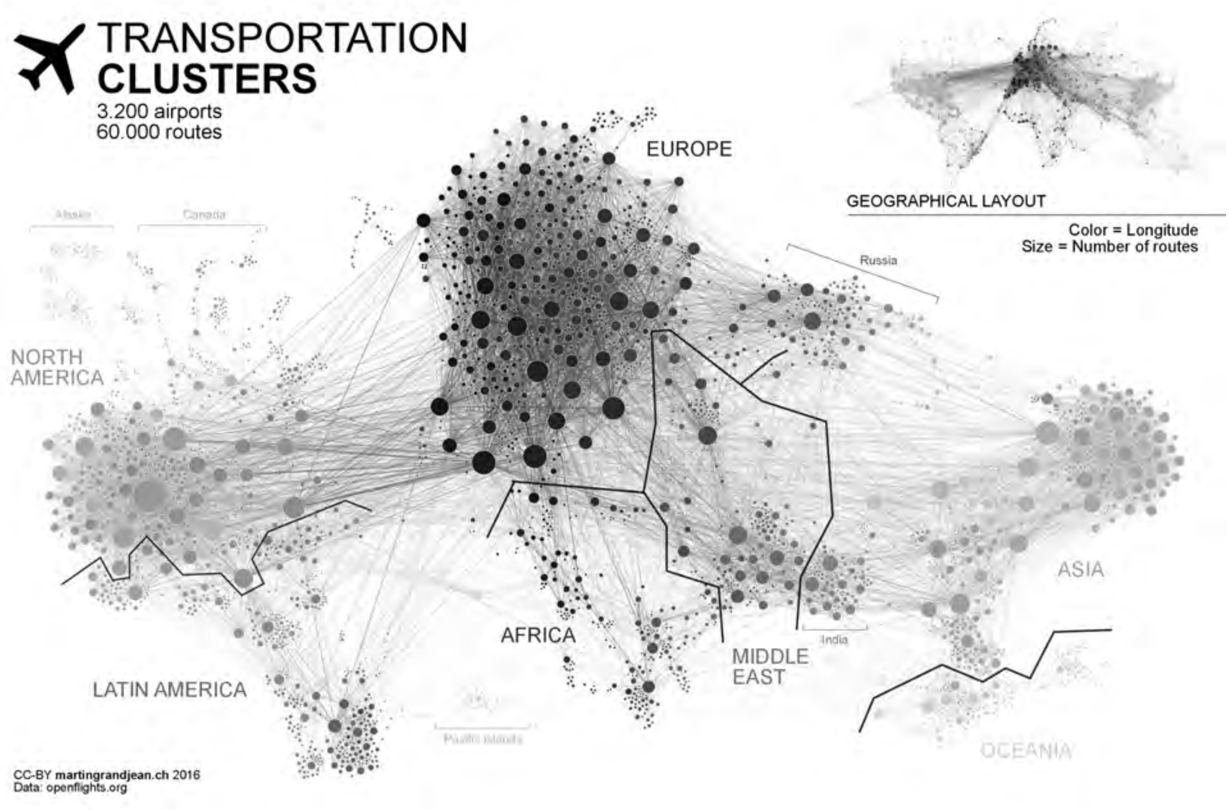

Figure 1.1: Global Air-Traffic Network

Consider, for example, Figure 1, which was created by Martin Grandjean[5]. It is a visualization of the world's air-traffic network, based on airports located worldwide. How is such a complex network of data and activity managed, we wondered, as there is obviously no one command center or central database?

And, that is just one aspect of the airline industry's complexity, *vis-a-vis* the issues of data acquisition, management, modelling, analysis, output, results and decision-making. Think also about the number of planes in the air at any given moment and their complex traffic patterns and schedules. Or how about the massive number of people traveling worldwide; let alone all of their baggage. How does it all reach its destination? And how about all of the discount travel websites and blogs and Facebook pages, and travel agents and related industries, such as tourism and hotels and restaurants? Or how about the credit card world and their point systems and travel discounts and all those databases that need to link-up to one another? And then there is security and policing and safety data, as well as airport surveillance cameras and so forth; not to mention all of the data for each plane that is being compiled, managed, analysed and used to make sure it arrives safely at each destination? Or how about the other fifty things we forgot to mention?

Sipping his tea, Warren looked at us both and said, 'So, along with various other examples, why not make the airline industry the case study for your book?' Great idea we thought. Sometimes the challenges of life just seem to work out. Or, should we say, travel is worth it. Or that is the hope, at least, for our journey through the current book.