Chapter 1
Big Data in Computational Social Sciences and Humanities: An Introduction

Shu-Heng Chen and Tina Yu

With the advance of information and digital technologies, specifically in relation to Web 2.0, ubiquitous computing, wearable devices, social media, and the Internet of things, a massive amount of information has been generated in the modern digital society. While the "big data gold rush" is under way in the business world (Peters 2012), government and social scientists are also interested in reaping the economic and scientific benefits from harnessing the power of big data. For example, the ESRC's Big Data Network in the UK and the US NSF's Big Data Research and Development Initiative are pouring fortunes into innovative projects to develop new methods and tools for capturing, managing, and exploiting enormous volumes of information. These major initiatives indicate that the governance of big data is essential for the advancement of human knowledge to accelerate economic growth and to provide a better quality of life for the people.

This edited book is about the potentials and challenges of big data for computational social sciences and the humanities.[1] It is prepared not only for computational

[1]Computational social sciences, as the title of this book series demonstrates, require little explanation. The term, computational humanities, however, is less popular. Gerhard Heyer distinguishes digital humanities from computational humanities as follows. The former is the creation, dissemination, and use of digital repositories, and the latter is the computer-based analysis of digital repositories using advanced computational and algorithmic methods (Biemann et al. 2014). Alternatively, "[c]omputational humanities is an emerging field that bridges the sciences and humanities with the goal of creating accurate computer simulations of historical, social, cultural, and religious events (Cruz-Neira 2003, p. 10)." See Gavin (2014) for a demonstration of the above two descriptions of computational humanities.

S.-H. Chen (✉) · T. Yu
AI-ECON Research Center, Department of Economics, National Chengchi University, Taipei, Taiwan

S.-H. Chen (ed.), *Big Data in Computational Social Science and Humanities*, Computational Social Sciences, https://doi.org/10.1007/978-3-319-95465-3_1

social scientists and humanists who are interested in using big data in their areas of research, but also for big data technologists who are interested in the governance of big data for social sciences- and humanities-related projects.

Big data means different things to different people. Regardless of the sources of the digital data, such as books, social media, databases, audio, and video, big data exhibit the characteristics of high-volume, high-velocity (speed of data in and out), and high variety (range of data types and sources). This new type of data enriches research prospects and has potential to advance research in the social sciences and humanities in the following ways:

- Advanced big data collection tools, such as web scraping, and innovative analytic techniques, such as machine learning, may help establish new research methodologies;
- New types of data may reveal new patterns and insights into human society, politics, and economics;
- New types of data may lead to new kinds of research questions that are beyond the perspectives of the established theories.

It is no wonder that many social scientists and humanists are turning to big data in their research.

With its great potential to revolutionize social science research, to what extent would big data inevitably replace more costly and time-consuming traditional methods of gathering information, e.g., surveys or in-depth interviews (Savage and Burrows 2007)? In addition, to what extent would big data fundamentally challenge traditional scientific practices? As a consequence, will the current standard research paradigms in the social sciences shift toward a new one? As one may learn from this volume, we consider that the change or the shift will gradually happen, but before that some words of caution have to be given. For example, Kleiner et al. (2015) believe that at this time big data should supplement, but not replace, traditional methods and data sources in the social sciences. "An important principle within the social sciences is that a study design should be optimal in producing valid and reliable data that fit and address the research questions of interest. Big data are usually not generated following a design intended to address specific research questions, and so generally do not easily lend themselves to use by social scientists" (Kleiner et al. 2015, p. 24). Hence, a related point is that the soundness of big data can never be established unless we can have a scientifically sound theory (model) for the generation process of big data (Chen and Venkatachalam 2017).

To promote the appropriate usage of big data in a scientific context, it would be beneficial to establish a research environment where the data utility, quality, and their accessibility are governed. The environment facilitates the sharing of big data best practices, such as suitable analytical strategies for a particular data type, and the early identification of error data sources. Furthermore, the ethical issues of using personal data can be addressed there.

In this volume, we are thrilled to have many distinguished scholars from diverse disciplines who have contributed to this book. The 18 chapters collected in this volume are organized into two parts, to be further elaborated in the first and second sections of this introductory chapter. In the first part of the book (Sect. 1.1), we present works that incorporate three major different kinds of big data, namely *geographic data*, *text corpus data*, and *social media data*, to conduct research on the social sciences in a wide range of fields, including anthropology, economics, finance, psychology, history, linguistics, political science, and mass communications. In the second part of the book (Sect. 1.2), we include two types of contribution: surveys that review published papers using big data to conduct research in various social science disciplines (Sect. 1.2.1), and articles that discuss challenges of using big data in social science research (Sect. 1.2.2). This is a book with many fascinating works. We hope that readers can enjoy it as much as we do.

1.1 Big Data for Computational Social Sciences and Humanities in Practice

There are 11 chapters that together constitute the first part of the book. These 10 chapters (Chaps. 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11) are further classified into three groups based on *the type of big data* employed. The three types of big data reviewed in this section are geographic data (Sect. 1.1.1), text corpus data (Sect. 1.1.2), and social media data (Sect. 1.1.3). While this is not an exhaustive list, these three give us the three most frequently seen types of big data used in the social sciences and humanities.

1.1.1 Geographic Data

Geographic data constitute information that has an implicit or explicit association with a location relative to the Earth. The data can be captured in many different ways, such as satellite remote sensing. Chapter 2, "Application of Citizen Science and Volunteered Geographic Information (VGI): Tourism Development for Rural Communities," authored by Jihn-Fa Jan, reports a successful application of using geographic data for tourism development in a rural community. In this case, the geo-referenced data are obtained through the global positioning system (GPS).

Chi-Shi is an agricultural-based community in the southern part of Taiwan, which is currently engaging in various activities to protect the natural environment and preserve the valuable cultural heritage. With additional interest in promoting

environmental education and cultural tourism, the community collaborated with the author to develop a web-based geographic information system (GIS) for community resources management and for tourism planning.[2] In particular, residents and visitors carried GPS loggers and digital cameras to record GPS coordinates and images while they traveled along the trails. These data were processed using various software tools and stored in a web-based GIS database. With the web interface, tour planners can dynamically query existing maps or create new richly annotated tour maps in real time using Google Maps applications. In addition to the geographic data, this chapter also provides a concrete illustration of how advances in information and communications technology (ICT), specifically, ubiquitous computing and wearable devices, have facilitated the development of citizen science, which has used crowdsourcing to gather information originally only sparsely disseminated, and then to share and process the pooled information so as to enhance decision-making and planning.[3]

Are cultural practices inventions or the propagation of existing ones? To answer that question, particularly the cultural practice of tombstone inscriptions in Taiwan, Oliver Streiter analyzed the tombstone inscriptions and their geographic information on more than 600 burial sites in Taiwan and in Penghu. Chapter 3, "Telling Stories through R: Geo-temporal Mappings of Epigraphic Practices on Penghu," presents this work. One particular place-name style on the tombstones, called *datanghao*, is most popular in recent Taiwan (after the Republic of China) and in Penghu (after the Japanese occupation). Given the localization of tombstones in time and space and their relatively large number, the *datanghao* style offers a unique opportunity to trace back the development of cultural practices under the influence of global political and economic changes.

The primary data collected are digital geo-referenced tombs photos, which contain in their meta-data the geolocation, the altitude, and the cardinal direction of the photo. Additionally, manual annotations and transcriptions values are assigned as attributes to the primary data. Using data collected in Penghu (3304 tombs), the Monte Carlo sampling method is applied to model the propagation of *datanghao* among the Penghu archipelago. In this case, a model is a directed graph where each node represents a burial site and the edges between the burial sites are directed from the early to the later to indicate the propagation direction. To make the models interpretable, in terms of their properties, some assumptions were made during the sampling. Among the sampled models, the one with the shortest average spatial

[2]For the related applications of GIS to the humanities, also see Chaps. 3, 4, and 14. In fact, these four chapters can together be read as part of the spatial humanities.

[3]For a general understanding of citizen science, also known as crowd science, and its recent development, the interested reader is referred to Cooper (2017) and Franzoni and Sauermann (2014).

distance calculated over the entire graph is selected as the final model. The model identifies Xiyu as the geographic origin of the *datanghao* in Penghu.

After interviewing tombstone carvers in Penghu and in Taiwan, Streiter hypothesized that the migration of carvers from Penghu to Taiwan might have caused the spread of *datanghao* to Taiwan. By expanding the Penghu model with data on 67,945 tombs collected in Taiwan, the new model supports the hypothesis. The research concludes that the *datanghao* place-name style on tombstones was invented most probably by a tombstone carver in Xiyu, Penghu to express loyalty toward the Qing dynasty without offending the Japanese during their occupation. Through the migration of carvers from Penghu to Taiwan, the *datanghao* spread and became an epigraphic practice in Taiwan.

1.1.2 Text Corpus Data

Text corpus data are digital data obtained from various sources (e.g., news, publications, and books), where the focus is on the text itself, and the texts are usually relatively long (big). Six chapters in this volume demonstrate the use of big data in this manner. The application domains covered here include history (Chap. 4), economics (Chap. 5), finance (Chap. 6), health (Chap. 7), literature (Chap. 8), and linguistics (Chap. 9). Not only do these six chapters show us how the use of the text corpus data can allow us to address some questions which are beyond the reachability of the conventional social sciences and humanities, but they also extend our generally neglected computational aspects of the social sciences and humanities. They, therefore, represent a new frontier in the social sciences and humanities.

As the title of this section suggests, corpus linguistics plays a threading role for these six chapters, but it serves only as the rudimental element (raw data). On top of it, different techniques, activities, and functions can be further added, such as map construction (Chap. 4), time series analysis (Chap. 5), sentiment analysis (Chap. 6), network modeling (Chap. 7), complexity analysis (Chap. 8), and event analysis (Chap. 9). In sum, these six chapters together show that the information hidden in the text can be very valuable for progress-making in the social sciences and humanities. In the following, we shall briefly give a sketch of each chapter.

1.1.2.1 History in Light of Dynamic Maps

In Chap. 4, "Expressing Dynamic Maps through 17th-Century Taiwan Dutch Manuscripts," Ann Heylen digitized a 17th-century Dutch handwritten manuscript (Church Minutes), which documented the presence of the Dutch community in Taiwan at that time. Based on the information, visualization GIS is developed to provide a better understanding of Taiwan's history in a global setting. The Dutch

United East India Company (VOC) established its presence in Taiwan in 1624 and continued until 1662. Since the VOC possessed quasi-governmental powers, tracing the mobility of their personnel, e.g., relocation and displacement, offers new insights into the social and economic changes taking place on the island during that period of time. Incidentally, the Church Minutes written by clergy contain such information, and hence these minutes are chosen for this research.

Initially, Heylen digitized the manuscripts (with Dutch and English translations). Next, names of persons and places in the text were extracted and standardized through the search/replace command using regular expressions. After that, place names were converted into geo-coding to produce the first blueprint map, where a "Personality Icon" was added to a place name if the person was mentioned at that location. To visualize the mobility of VOC personnel, users can select the person's "Personality Icon" at a location and a pop-up window will appear displaying the next location to which the individual moved. At the new location, one can select the person's "Personality Icon" there and the person's next location will appear. By continuing this process, one can obtain information on the spatial mobility of the individual throughout the time period.

1.1.2.2 Economics Paradigm Shift in Light of Corpus Linguistics

In Chap. 5, "Has Homo economicus Evolved into Homo sapiens from 1992 to 2014?: What Does Corpus Linguistics Say?" Yawen Zou and Shu-Heng Chen present an innovative big data application that studies the paradigm shift in economic research. In 2000, Richard Thaler predicted that economists would shift their research approach from the Homo economicus paradigm to the paradigm of *Homo sapiens*.[4] To check whether this prediction has been fulfilled or not, they adopted a corpus linguistic method to analyze their data.

Initially, they built a corpus using the abstracts of 51,285 economics research articles published from 1992 to 2014 in 42 mainstream economics journals. Next, they identified and selected two sets of keywords that are associated with each of the two paradigms. The keywords are in regular expression form, which supports the don't-care symbol *. For example, "cognit*" matches "cognitive," "cognition," "cognitively," "cognitivity," and so on. Since the two paradigms are opposite to each other, i.e., Homo economicus formulates the rationality of economic behavior in an ideal mathematical optimization framework while *Homo sapiens* emphasizes the consideration of the psychological, cultural, and social factors that constrain a human's rationality, the two keyword-sets are disjointed with little overlapping.

They then counted the frequencies of each keyword in the corpus and regressed these frequencies time series to see which keywords have an upper trend (increased usage) and which keywords have a downward trend (decreased usage) over time.

[4]Richard Thaler is the 2017 Nobel Laureate in Economics.

The results show that 81.4% of the *Homo sapiens* keywords have an upper trend and 65.2% of the Homo economicus keywords have a downward trend. From this observation, they concluded that Thaler's prediction is largely correct. Moreover, since the period studied is not long (22 years), they believe the paradigm shift is still happening and will continue for some time.

Using corpus linguistics to analyze economic texts has become a new research trend in economics. This trend can be considered to represent a small amount of progress in the less-explored interdisciplinary areas between economics and the humanities, the area that Deirdre McCloskey has long promoted, which is partially known as the rhetoric of economics (McCloskey 1983, 1998), and the Nobel Laureate in Economics, Robert Shiller, recently also advocated and coined the term *Narrative Economics* (Shiller 2017). How economists can learn from the humanities is also well illustrated in general by Morson and Schapiro (2017) and, specifically, in behavioral economics by Roy and Zeckhauser (2016). Probably the closest related study to this chapter is the recent application of computational linguistics to the study of the open market operations of Federal Reserve Banks (Hansen et al. 2018).

1.1.2.3 Financial Prediction in Light of Sentiment Analysis

One way to see the close connection between the humanities and the social sciences is through the sentiments or emotions derived from or extracted from the text or, as alternatively put, the psychology of reading the text. In the era of big data, humans will not read or handle the overwhelmingly generated texts themselves; instead, intelligent machines will take over the reading of the massive amount of texts. Machines and the embodied algorithms will simulate the emotions that humans may put into the text as an author or get from the text as a reader.[5] The replacement of humans with machines in reading or even in authoring is currently known as *sentiment analysis*.[6] What does this text mean? Does it imply something positive or negative? What are its implications for investors' emotions? This whole business overarching the humanities, psychology, computer science, and finance has defined a hot spot for the "big data gold rush" in the sense that if you know what the text means, you may know how to invest.

The financial industry has been greatly impacted by the big volume of data and by the technologies that generated these data. It is quite likely in the future that "quants" on Wall Street will no longer solely rely on numerical data to trace the ups and downs of prices; instead, they will increasingly rely on using data analytics, text

[5]There is a philosophical issue as to whether machines will evolve to have their own interpretations of the text and hence develop their own emotions which are different from those of general human beings under the governance of their own culture. More *positively*, would machines surpass humans by demonstrating the features of *positive psychology,* as advocated by Martin Seligman (Seligman 2004), more successfully than humans?

[6]There are already quite a few good references giving a panoramic guide to this fast growing field. The interested reader is referred to Liu (2015), Pozzi et al. (2016), and Cambria et al. (2017).

analytics, and the digital humanities to discover the underlying market sentiments as leading indicators for prices. In fact, a stream of the literature documenting these studies has already been piling up (see Chaps. 6 and 12).[7] In Chap. 6, "Big Data and FinTech," Jia-Lang Seng, Yao-Min Chiang, Pang-Ru Chang, Feng-Shang Wu, Yung-Shen Yen, and Tzu-Chieh Tsai first presented their works using online news to conduct sentiment analysis for the Taiwan stock market. They then discussed the strategies and the computer framework that they have developed to better serve financial institutes using real-time mobile/cloud computing.

The chapter presents two studies on the use of news sentiments to predict Taiwan stock market performance. The first study was based on an asset-pricing model, which incorporated news sentiments related to investment, macroeconomics, and politics. The authors reported that there was a positive correlation between the investment news sentiment and the Taiwan 50 (TW50) index returns. There was also a negative correlation between the political news sentiment and the returns of some stocks listed within the TW50 index. The second study on news sentiment analysis was performed using a software tool that graded the view of a news article as being positive or negative on a scale from $+5$ to -5. The kind of news articles that they were concerned with were also related to investment, macroeconomics, and politics. The authors reported that the news sentiments were correlated with the stock prices, but this result may have been affected by the subjective judgment of the software tool. In addition to sentiment analysis, the authors also discussed the strategies that financial business security and technology providers are using to better serve their customers. They proposed a cloud-based mobile computing framework that can handle complex data structures and large amounts of data within a short period of time.

1.1.2.4 Network Modeling of Health-Related Concepts and Institutions

As already seen in Chap. 6, from the perspective of the humanities, each text can be represented by its sentimental and emotional ingredients that present readers with a "mood" of the text. On the other hand, in a more concrete manner, a text can be represented by graphs (networks, word clouds, and maps), which present readers with *a geometry of the text*. Several chapters in this volume do work on the graphical representation of texts. Chapter 5 is an illustration of the *co-word network*, which is basically a single network, a network with one type of link. Chapter 7 extends this graphical representation of texts to *multiplex networks*, or networks with more than one type of link (relation), to which we now turn.

[7] While there are only two chapters collected in this volume, the interested reader may find more useful references in Peterson (2016) and the excellent collections edited by Mitra and Xiang (2016). However, sentiment analysis may go further, beyond what the current literature delineates, and can be further incorporated into agent-based computational finance and give new impetus to behavioral finance (Chen and Venkatachalam 2017).

The World Health Organization (WHO) and the Secretariat of the Convention on Biological Diversity recently decided to strengthen their collaboration, most notably in raising awareness of the complex linkages between biological diversity, ecosystems, and human health, acknowledging the strong connection between biodiversity and health. In a joint report, they highlight the fact that biodiversity loss constitutes a fundamental risk to the healthy and stable ecosystems that sustain all aspects of our societies (WHO-CBD 2015). Themes related to Health are increasingly cited by the COPs (Conferences of the Parties) of the CBD, the Convention on the Conservation of Migratory Species (CMS), and the Convention on International Trade in Endangered Species (CITES), encompassing dimensions of human health, animal health (domestic and wild fauna), and ecosystem health. Other ecological or environmental concepts, such as biodiversity, an ecosystemic approach, and risk assessment, favor the emergence of Health issues and their integration into the CBD.

In Chap. 7, "Health in Biodiversity-Related Conventions: Analysis of a Multiplex Terminological Network (1973–2016)," Claire Lajaunie, Pierre Mazzega, and Romain Boulet investigated which health themes have emerged from these three biodiversity-related conventions. They analyzed how concepts are used in a complete or partial form in each COP and how they are transmitted between COPs through a multiplex network, with each type of link in the network corresponding to a concept. They then identified the most central COPs and their gathering into communities in the process of emerging Health issues.

Their aim is to study the dynamic of health within the biodiversity-related conventions and to capture simultaneously the dynamic of the importance of COPs in the diffusion of health issues and the dynamic of health themes within their decisions and resolutions. The common dynamic shown, thanks to the use of multiplex analysis combined with text mining used in a big data perspective, facilitates in understanding how each concept contributes to the building of an integrative and multi-dimensional approach of Health issues in international environmental law.

In their approach, they first collected the textual corpus from the CBD, CMS, and CITES conventions (agreements) and from the decisions and resolutions published by their respective COPs from 1973 to 2016. Next, they extracted 213 terms that are related to health issues and divided them into 13 concept groups. After that, they counted the number of occurrences of the 13 concepts (occurrences of any of the constituent terms) in the textual corpus. They then analyzed these frequency data using network modeling. The network model consisted of two different kinds of nodes, namely 13 concepts and 27 COPs held from 1973 to 2016. A concept was linked to a COP if the concept was mentioned in the COP's publication. Moreover, two COPs were linked if the same concept was mentioned in both of their publications. By measuring the degree and betweenness centrality of the network, they identified the five most central COPs that played the most important role in disseminating health-related themes. Moreover, the temporal evolution shows that the three most frequently mentioned health-related concepts over time are risk and threat, health and security.

1.1.2.5 Linguistic Complexity of Shakespeare Plays

We have seen the "mood" and the geometry of a text. It is then natural to ask an even more fundamental question: What is the complexity of a text, considering that some texts are simple and some are not? One approach to this issue is to directly provide a complexity measure suitable for sentiment and another measure suitable for landscape, as one can easily imagine that the "mood" can be simple or complex, and the same for its shape. In fact, studies along these lines already exist.[8] However, there is a third approach that looks at this issue, which is presented in Chap. 8.

Although Shakespearean plays have been regarded as the finest works in the English language, they are not always easy to enjoy, because the language can be unfamiliar, and hence intimidating to the general public. Does the linguistic complexity of a Shakespearean drama play a role in its audience's acceptance and its commercial success? In Chap. 8, "How Does Linguistic Complexity in Shakespeare's Plays Relate to the Production History of a Commercial American Theatre?" Brian Kokensparger applied a computational method to the Shakespeare corpus to answer this question.

Kokensparger designed four measures to quantify the linguistic complexity of Shakespeare's 38 plays: average syllables per word, average words per sentence, percentage of complex words, and percentage of words not found in a standard dictionary. The plays were ranked from the lowest linguistic complexity score to the highest. Then these rankings were compared with the ranked production frequency of a commercial Shakespearean theater. The results indicated that the plays offering the highest frequency over the theater's history were also among the least complex of Shakespeare's plays. Therefore, there appears to be a relationship between linguistic complexity in the text of Shakespeare's plays and the commercial viability of offering those plays to a paying audience. As the linguistic complexity of a performed play affects the cognitive load on the audiences, it is reasonable to suggest that plays with the lowest linguistic complexity will be more frequently chosen for production than their counterparts with higher linguistic complexity for a theater that seeks to successfully entertain patrons and keep them coming back.

1.1.2.6 Evolution of "Language" in Light of Corpus Linguistics

Corpus linguistics is the study of languages as expressed in the corpora of real-world texts. In Chap. 9, "Language Communities, Corpora, and Cognition", Huei-Ling Lai, Kawai Chui, Wen-Hui Sah, Siaw-Fong Chung, and Chao-Lin Liu presented three case studies using corpus-based methods to investigate the linguistic patterns and the cognition of different community groups.

[8] For example, for the complexity measure for sentiments, see Joshi et al. (2014); for the complexity measure for networks, see Morzy et al. (2017).

The first study investigated how the lexicalized term <nganggiang stiff neck 硬頸>, a metonymy-based metaphor, is used in the news media. This body-oriented metaphor characterizes a person as being stubborn and tough by describing his/her body expression of making the neck stiff to show an unyielding attitude. To understand how this metaphorical expression has become entrenched and conventionalized to carry such a meaning, they collected online news data from four major newspapers in Taiwan, and counted the usage of the term in the corpus. Quite interestingly, they found that the frequency of the usage is correlated with the major election years in Taiwan. Moreover, while originally carrying a negative connotation, the term is now a positive phrase used to characterize Hakka-related matters. In addition, through the mechanisms of denotation extension, metonymy, and metaphor, its usage has increased over the years.

The second study analyzed the usage of gestures in Taiwan Mandarin conversations. In a collection of 15 recorded conversational excerpts, 2012 gestures were found across male-speaker conversations, female-speaker conversations, and mixed-gender conversations. They counted the frequency of five different kinds of gestures across the three types of interaction and found that each one differed from the others. This indicates that gender affects the usage of gestures.

The third study compared the oral narrative abilities of Mandarin-speaking children with and without the *autism spectrum disorder* (ASD). Among various indices of narrative abilities, referential choice is regarded as an important window to show a speaker's sensitivity to listeners' needs. The authors therefore used referential forms and pragmatic functions data to conduct their study. They found that both groups of children were comparable in using nominal forms such as introducing and reintroducing referents. However, null forms, rather than pronominal forms that are normally used to maintain reference as reported by other researchers, appeared to be the dominant device for both groups of children to maintain reference.

1.1.3 Social Media Data

While digital archives and the resultant text corpus data constitute the essential body of the big data, they are regarded as the "classical type" of big data in the sense that the original forms of texts are not digital and conversions are needed before placing them into the arrays of big data. Nowadays, thanks to the advances in ICT technology, the modern forms of texts are "born" digital; the online news data of Chap. 6 provide a case in point. Apart from that, social media networks built upon and further facilitated by the ICT technology have fundamentally revolutionized the way in which texts can be generated and also archived. Basically, social media and the Internet of everything can technically map and archive the entire world into its cyber counterpart, and thus new historiographies may be introduced. The conventional *discrete-in-time* historiographies will be challenged by the future *continuous-in-time* ones. With this irreversible trend, social media data will eventually monopolize the whole of big data.

Citizen science (crowd science or volunteer science), as we have already seen in Chap. 2, demonstrates how social media data can fundamentally change the way in which we do science by allowing people from all over the world to contribute to groundbreaking scientific discoveries. In addition to science, social media data can also exert great influences on the way in which the democratic system is operated; nevertheless, in addition to golden opportunities and promises presented to public administrators, challenges are also prevalent. Basically, it is still not clear whether we shall have more "wisdom of crowds" (information aggregation) or more "stupidities of herds" (noise amplification). The accumulated discussions are very long.[9] To some extent, all are concerned with the possibilities of building a good or better society using advanced digital technology, artificial intelligence, and big data. The following two chapters of this volume exhibit the typical flavor of this so-called *social media dilemma*; Chap. 10 is concerned with digital governance, whereas Chap. 11 is concerned with the grassroots politics.

1.1.3.1 Digital Governance Using Public Opinion Mining

Social media have become a popular channel for citizens to express their opinions and complaints regarding public policies. While the data volume is large, the analysis and interpretation of the data to produce meaningful insights is not an easy task. In Chap. 10, "From Naive Expectation to Realistic Progress – Government Applications of Big Data to Public Opinions Mining," Naiyi Hsiao, Zhoupeng Liao, and Don-Yun Chen presented their work on Internet public opinion analysis regarding the *Free Economic Pilot Zone* (FEPZ) policy in Taiwan.

In March 2014, the FEPZ bill was submitted to the Legislative Yuan for review and approval. The bill raised much controversy due to its *Free Trade Agreement* (FTA) with Mainland China, which has become a cause for concern due to fears of losing political independence under the increased economic dependency on Mainland China. To understand public opinion in relation to the bill, the Taiwan National Development Council commissioned the authors and a private technology company to collect and analyze unstructured public opinion data from various Internet media, including news websites, forums, blogs, social media (Facebook and Twitter), and PTT.[10] The analyzed results were then presented to the public officials for their feedback.

During the study conducted from May to November of 2014, many governmental officials participated in the project, for example, by providing keywords, key events, and the names of policy-relevant stakeholders for the team to query media data for

[9]Interested readers are referred to Bauerlein (2008), Sunstein (2008), Ceron et al. (2016), Thompson (2016), Helbing et al. (2017), O'Neil (2017), and Stephens-Davidowitz and Pabon (2017).

[10]The PTT Bulletin Board System is the largest terminal-based bulletin board system (BBS) based in Taiwan. For more information, see https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System.

various kinds of analysis: sentiments (positive/negative), volumes, popular media channels, and the public opinion leaders. However, feedback in regard to the analysis was mixed. Some argued that most netizens are not experts on economic policies; hence their opinions appear relatively useless. Others believed that the statistics and computer algorithms used to conduct the analysis were not adequate. Nevertheless, much has been learned from this study and it provides a baseline for future improvement.

1.1.3.2 Grassroots Politics

Social media have also become a critical platform for social activists to promote their movement. A case in point is the *Sunflower Student Movement* in Taiwan, where social media were the major vehicles used to disseminate information and to organize activities. Unlike the keyword mining approach used in the previous chapter, the study in Chap. 11 conducted by Hui-Wen Liu, I-Ying Lin, Ming-Te Chi, and Kuo-Wei Hsu analyzed the fan pages centered around the Sunflower Student Movement on Facebook, and is entitled *Understanding "the User-Generated": The Construction of the "ABC model" and the Imagination of "Digital Humanities."*

The Sunflower Student Movement evolved as a protest movement driven by a coalition of students and civic groups to protest the passing of the Cross-Strait Service Trade Agreement (CSSTA) by the then ruling party Kuomintang (KMT) at the legislature without a clause-by-clause review. The movement started on March 18, 2014 when a group of students and activists occupied the Legislative Yuan. Over the next 23 days, various Facebook fan pages were created, including that of the leading organization, *Black Island Youth Front* (BIYF), and two others that concentrated on this movement during the occupation (*New e-Forum* and *Anti-Media Monsters Youths*, AMMY). In addition, at least 13 other fan pages were created to support this movement.

The authors chose 16 fan pages for analysis. For each fan page, they first collected its activities (*fan*, *like*, *share*, and *comment*) during the movement from March 18 to April 11 in 2014. Next, they computed their proposed *ABC indices*. A represented "activity," summing up the activities of each user within a fan page. B represented "broadness," totaling the amounts of posts in which each user had participated. C referred to "continuity," calculating the duration time of the various users, from the first time they participated to the last time they left their digital footprints on one fan page. Using the ABC indices, they found diversified functions contributed by the fan pages during the Sunflower Student Movement.

All in all, through the social media data, this chapter is able to provide a lot of details regarding the university students' participation process in politics via social media networks. The findings have become valuable additions to the growing literature on this research stream, see, for example, Conover et al. (2013) and Loader et al. (2015).

1.2 Survey and Challenges

The first part of the volume focuses on the use of big data and the kinds and the flavor of the studies facilitated by the use of big data. Each of the ten chapters there works with one or more specific kinds of big data and addresses a set of intriguing questions, covering subjects belonging to anthropology, art and theater, citizen science, economics, finance, history, linguistics, literature, political science, public health, and sociology. Differing from the first part, the second part of the volume does not address a specific set of questions with a particular set of big data; instead, it provides a panoramic view of the development of big data in the computational social sciences and humanities (CSS&H), including its trendy directions and the evoked challenges. Hence, the two parts of the book are connected by employing the *specific-to-general* scheme. With this description, Part 2 is further divided into two sub-parts. We begin with a sketch of the general development of big data in CSS&H (Sect. 1.2.1). Four survey articles are written for this purpose, and they together cover some representative cases of the timely development of big data in business (Sect. 1.2.1.1), the social sciences (Sect. 1.2.1.2), the humanities (Sect. 1.2.1.3), and technology (Sect. 1.2.1.4). They are big data finance (Chap. 12), big data in psychology (Chap. 13), spatial humanities (Chap. 14), and cloud computing (Chap. 15). The second sub-part (Sect. 1.2.2) then presents an overview of some of the challenges associated with big data. Four challenges are presented in this book, namely the complexity of big data (Sect. 1.2.2.1) or the ontology and epistemology of big data, big data search (Sect. 1.2.2.2), big data simulation (Sect. 1.2.2.3), and big data risks (Sect. 1.2.2.4). While these challenges may have also been addressed in the literature, in this book these challenges are uniquely presented by social scientists from their work on mass communication (Chap. 16), anthropology (Chap. 17), history (Chap. 18), and sociology (Chap. 19); hence, their shared views are domain-specific and more focused.

1.2.1 Survey of Published Research

1.2.1.1 Big Data Finance

From the perspective of business, probably one the most astonishing examples of progress observed in the last decade is in the area known as *big data finance*. FinTech, as we have seen in Chap. 6, is a case in point, but, from a macrocosmic viewpoint, what can interest us is the following: with the avalanche of financial market data coming out every minute, in the form of Yahoo Finance, TRMI sentiment data, and other social media, have these financial big data changed financial market behavior? In Chap. 12, "Big Data Finance and Financial Markets," Dehua Shen and Shu-Heng Chen have surveyed those works which help define what we now know as big data finance.

There are two fundamental cornerstones in finance. The first one is the efficient markets hypothesis, basically, the unpredictability of financial returns. The second one is various characterizations of market activities, such as volatility, trading volumes, duration between trades, spreads, and depths; sometimes, their steady patterns are known as "stylized facts" (Chen 2008). With big data, Shen and Chen ask whether more light will be shed on studies of these issues. The first question has been addressed many times in similar forms in the history of economics, and no easy conclusion will be reached.[11] The second one is key because of the nature of big data.

Big data, by its very nature, is the archive of what people said, what they did, and what they thought (Chen and Venkatachalam 2017). It can inform us of lots of human intentions, interactions, emotions, attitudes, and hence decisions (see also Chap. 13 of this volume); to some extent, it functions as a mirror of the world, enabling us to see who we are and how we got here. In other words, we can now see the physical world around us through the lens of the cyber world, thanks to various measures derived from text mining, content analysis, sentiment analysis, and Google trends. Hence, one may reasonably expect to be exposed to an unprecedentedly high level of transparency, and not just be recipients of more information. The theory of general reflexivity (Soros 2013), level-k reasoning, or similar forms of reasoning (Chen 2013) may then suggest that this can further lead to some fundamental changes from microcosmic decision-making to macrocosmic market dynamics; for example, patterns of volatility can be distorted, as with many other familiar financial patterns, and therefore be destroyed. Hence, the survey provided by Shen and Chen is best viewed as the beginning of big data finance, and one still needs to figure out how big data finance can be fundamentally different from "small data" finance, by incorporating human cognition, emotions, social interactions and, above all, human psychology, a subject to which we now turn.

1.2.1.2 Big Data in Psychology

Psychology is the science of behavior and mind. It seeks to understand individuals and groups by establishing general principles and by studying specific cases. While many recognized research methods, such as controlled laboratory experiments, and electroencephalogram (EEG) and animal studies, have contributed to the advancement of the field, big data are providing new alternatives to study psychology. As already mentioned, the very nature of big data is the mapping of people's behavior in the physical world to its cyber counterpart. Therefore, it is anticipated that some human behavior which is not directly observable from the physical world may

[11] The conundrum has been well illustrated by the so-called *adaptive market hypothesis*, which endowed the efficient markets hypothesis with a dynamic and evolutionary interpretation (Lo 2004). In the vein of the agent-based fashion, the adaptive market hypothesis has been further studied in the form of the *market fraction hypothesis* (Chen et al. 2010).

be easier to observe in the cyber world. Hence, some forms of big data can be very useful for psychology, in particular, social media data, health tracker data, geolocation data, dynamic public records, travel route data, behavioral and genetic data, etc.

Currently, maybe the most ambitious project launched to explore this research opportunity is the Kavli HUMAN Project.[12] This project aims to gather a detailed array of measurements from 10,000 New York City residents over a 20-year span, allowing a team of scientists to monitor in intimate detail how these New Yorkers lead their lives over the course of 20 years, including where they go, what they eat, who they talk to, what they buy, and how their bodies grow, change, and deteriorate. Needless to say, this is a massive data-collection endeavor with the main pursuit being to learn how everything, from biology to behavior and the environment, affects the human condition; for example, how biological, medical, and social factors interact and impact the risks of cognitive decline from birth through to older age. We, as decision makers, make a large number of decisions (choices) in each typical day, some of which can substantially impact our well-being (Clark et al. 2018). Hence, if we can be better informed of how these choices are made, we may further improve our quality of life and increase our "happiness index."

In fact, much before the advent of the big data era, many man-made efforts to develop "big data" already existed. For example, it is known that the longest study of adult life that has ever been conducted is probably the Harvard Study of Adult Development (Vaillant 2008). In this unprecedented series of studies, Harvard Medical School has followed 824 subjects associated with different genders and different economic and social status, from their teens to old age. This "big data" study, containing subjects' individual histories, work, home lives, and health, which may be the most complete ever done anywhere in the world, has been used to illustrate the factors involved in reaching a happy, healthy old age. This pioneering study indicates that the demand for big data in psychology research has been much ahead of the times.

While psychologists have been aware of the indispensability of big data in their research, the responses made and the actions taken have by no means been sluggish. In Chap. 13, "Applications of Internet Methods in Psychology," Lee-Xieng Yang reviewed works that leverage big data to conduct research in modern psychology. In his survey, Yang grouped the existing works into three categories.[13]

First, crowdsourcing that has been used to conduct psychological surveys and experiments. Various works have reported that using Internet websites, such as

[12]This project is carried out within a collaboration between the Kavli Foundation, the Institute for the Interdisciplinary Study of Decision Making at New York University (NYU), and the NYU Center for Urban Science and Progress. For more details, the interested reader is referred to Azmak et al. (2015).

[13]The current use of big data in psychology is not just exhausted by the survey presented in this chapter. The journal *Psychological Methods* has published a special issue on this frontier (Harlow and Oswald 2016). For other developments, the interested reader is also referred to Cheung and Jak (2016) and Jones (2016).

Amazon's Mechanical Turk (MTurk), to recruit participants for psychological experiments has generated results that are consistent with those produced in a laboratory setting. Moreover, personality measures that survey data collected using Internet websites are more representative than traditional samples with respect to gender, socioeconomic status, geographic location, and age, and are about as representative as traditional samples with respect to race. However, there are also concerns about data quality due to duplication of participants and the cost caused by the large amount of post hoc data exclusion. Yang suggested that researchers consider their psychological study factors, such as the type of dependent variables and prior knowledge interference, to decide whether to conduct a survey or an experiment on Internet websites.

Second, Google news archives and the Wikipedia database that have been used to study psychology at the population level. Yang exemplified this direction of research with one work that used Google news archives to measure the amount of media attention a humanitarian crisis receives. They found that the more deaths that are involved in a crisis, the more media attention it receives. This result is consistent with previous studies reporting that humans generally have a diminishing sensitivity to the number of human fatalities. Another study used the entire English language Wikipedia corpus to train a computational model that represents human heuristics. The model was then used to answer a series of multiple choice trivia questions. One interesting observation is that the trained model mimics human behavior in making the probabilistic fallacies associated with the representativeness heuristic.[14]

Third, the influences of social networks on individuals. Yang referred to a study that has used Facebook data to study the emotional contagion of human beings. The results indicate that the emotion expressed in a user's Facebook friends' posts is a valid predictor of the emotional expression of the user's own posts. This, to some extent, is consistent with the so-called three degrees of influence in the literature on social networks (Pinheiro et al. 2014).

1.2.1.3 Spatial Humanities

Big data form the digital archive of what people did, what they said, and what they thought. What is implicit in this definition are the surroundings or the embeddedness, i.e., the places, the spaces within which these actions, narratives, perceptions, beliefs, and more extended social settings and social interactions are operated. The surroundings do not just refer to the physical settings, but, more importantly, to the information or the meanings associated with these settings, which in Chinese culture is broadly known as Feng Shui (Rossbach 1983; Webster 2012).

[14]The representativeness heuristic is one of the heuristics that has been carefully studied by psychologists and behavioral economists, regarding how human decisions or judgments are made under uncertainty (Kahneman and Tversky 1972).

Over the last few decades, the awareness of these surroundings, either known as the *spatial awareness* or the *network awareness* (the spatial thinking or the network thinking), has been widely received in parallel in many disciplines, covering both the humanities and social sciences. "Space," as it may literally suggest, becomes the engine to integrate subjects originally studied in isolation in different disciplines, which include people, time, events, history, beliefs, cultures, religion, politics, etc., as already exemplified in Chaps. 3 and 4. This integration forcefully shows the dynamic nature of space, motivates us to adopt a spatial approach to historiography, and promotes *spatio-temporal thinking* in the humanities and social sciences.

On the other hand, this "spatial turn" has also occurred within the discipline of geography itself. Beginning in the 1970s, many geographers were seeking alternative paradigms to rigorous geographical analysis that were not reducible to merely geometries. As Edward Soja (1940–2015) observed, "rather than being seen only as a physical backdrop, container, or stage to human life, space is more insightfully viewed as complex social formation, part of a dynamic process (Soja 2001)." This desire to make maps "deeper" with many time-layers of features associated with the same location further gained momentum with the technology innovation in geography around the same time, especially geographical information systems (GIS). These GIS refer to software that captures, stores, manages, displays, and analyzes information linked to a location on earth. GIS can relate different types of data—quantitative, textual, image, and audio—to each other based on their shared location. It also allows a visualization of these relationships on a map of the geographical space in which they all occur. The availability of GIS as well as other related technologies plus the spatial turn in geography have facilitated the spread of the idea of space, place, and place-making in the humanities, and have helped grow the interdisciplinary field, *spatial humanities*.

With the funding of the National Chengchi University President's Office and other sponsorships, the Asia-Pacific Spatio Temporal Institute (APSTI) was established in 2014. The Institute is a home for innovative GIS-based research on humanities-related subjects.[15] In Chap. 14, "Spatial Humanities: An Integrated Approach to Spatiotemporal Research," David Blundell, Ching-Chih Lin, and James Morris have summarized four thematic forms of research that are ongoing in APSTI and mark the nascent field of spatial humanities.[16]

First, there is the attempt to develop an interactive 3D visualization website to enhance the understanding of the diffusion of culture and oceanic navigation in Monsoon Asia. Second, there is the development of a GIS database that provides a heritage inventory and its management, with a specific application to the Nouli community in Taiwan. Third, there is the documenting and mapping of earth god shrines to establish the patterns of settlement, communal organization, and historical trade networks of communities in Taiwan, southern maritime China, Hong Kong,

[15]The interested reader is welcome to visit its home page: http://apsti.nccu.edu.tw/.

[16]For a general background of this fast-growing field, the interested reader is referred to Bodenhamer et al. (2010).

Macau, and outlying islands. Fourth, there is the development of methods and resources, such as GPS devices and visualization, to support the study of Chinese religion.

1.2.1.4 Cloud Computing

Cloud computing is an Internet-based form of computing that provides shared computer processing resources and data storage to other computers and devices on demand. Advocates claim that cloud computing allows organizations to avoid up-front infrastructure costs (e.g., the purchase of hardware servers). Moreover, individual researchers can enjoy the high-performance computer power from their own desktops and laptops. With the increased number of viable cloud computing providers on the market, adopting the cloud for big data research has become a very attractive option for computational social scientists.

In Chap. 15, "Cloud Computing in the Social Sciences and Humanities," Michael Gallagher has reviewed four major cloud-computing platforms, namely Amazon Web Services, Google Cloud Platform, Microsoft Azure, and Hewlett Packard Enterprises, from the perspective of a social science researcher. To address the deep learning curve and user-friendliness issues of using cloud computing, he has also provided step-by-step instructions to create an Amazon Web Services account and to set up the computing environment.

1.2.2 Challenges of Using Big Data for Research

In transitioning from the small-data research paradigm to the big-data research platform, social scientists and humanists face various challenges. Sect. 1.2.2.1 discusses about how to process/transform social media big data to be used to answer mass communication research questions. In Sect. 1.2.2.2, big database query/search issues, such as the ranking of relevancy to the query among a large number of possible matches, are addressed. Sect. 1.2.2.3 presents an approach to conduct big data simulation using agent-based models. Sect. 1.2.2.4 argues the risk of cyborgs, hybrids of humans and technologies, developed under the big data technologies era. In other words, our thinking will be a hybrid of biological and non-biological thinking. Would that be a curse or a blessing for the humanity in the long run? History will be the judge of that.

1.2.2.1 Big Data Complexity

Mass communications research is chiefly concerned with how the content of mass communications affects the attitudes, opinions, emotions, and ultimately behaviors of the people who receive the messages. Since social media have become important

vehicles to disseminate information, it is natural for mass communications scholars to become interested in analyzing their content. However, the data analytics developed so far is still inadequate to address the complexity of social media. Consequently, diving in the social media big data to search for the answers to their research questions is not an easy task for mass communications researchers. In Chap. 16, "Analysis of Social Media Data: An Introduction to the Characteristics and Chronological Process," Pai-Lin Chen, Yu-Chung Cheng, and Kung Chen discuss the challenges of social media data analytics.

Two challenges are addressed in this chapter. The primary challenge resides in the characteristics of social media data. First, social media data are enormous in scale and diverse in structure, which make the traditional research methods unworkable. Second, social media data are generated by humans, and hence are made complicated by the language usages and by the diverse human interaction patterns. Third, data integrity is not reliable because not all data sources are accessible. In addition, data transparency is insufficient, due to the black-box data algorithms. The authors suggest how these problems can be ameliorated or managed.

The second challenge concerns how to connect a research question to the data and to discover a problem-solving approach. To overcome this problem, the authors have proposed establishing a team of trans-disciplinary experts that consists of not only mass communications researchers but also data scientists who are familiar with the data processing tools. Together, the team can extract related data from the big data to address the research questions posted by researchers.

1.2.2.2 Big Data Search

There exist many social science and humanities databases that can support the sharing and reuse of information. One example is the HRAF, an abbreviation for Human Relations Area Files databases (eHRAF World Cultures and eHRAF Archaeology). Over the years, the databases have grown very rapidly, which has created a database search challenge: there are too many complex entries retrieved from each query. In Chap. 17, "Big Data and Research Opportunities Using HRAF Databases," Michael Fischer and Carol Ember discuss some potential and partial solutions to address this big data search problem. In addition, they have highlighted the challenges of integrating disparate ethnographic materials into HRAF databases to improve their reusability.

The HRAF databases contain texts that describe social and cultural life in past and present societies around the world. As of the spring of 2018, the two eHRAF databases contained almost three million "paragraph" units from over 8000 documents describing over 400 societies and archeological traditions. A typical keyword search normally returns around 20,000 paragraphs. Although there are simple strategies to reduce the number of returns, such as narrowing the search to a focal community and time period, or combining subject categories or adding keywords, these strategies are not effective due to the complexity of the

databases. They therefore proposed using big data technologies to develop a range of post-processing tools and methods to be applied to the returned text to help the users refine their search. An example of one method is search by example, where the user has selected entries from a search return that has been used as a basis for identifying similar entries from across the databases.

The HRAF has another challenge: how to achieve the interoperability across various kinds of ethnographic sources so that more documents can be shared. They address this problem at three levels: syntactic interoperability, semantic interoperability, and pragmatic interoperability. Currently, they are exploring research opportunities to resolve these issues.

1.2.2.3 Big Data Simulation

The third challenge for the use of big data is the frequent lack of a scientific or theoretical foundation for big data. As already seen in Chap. 16, big data are often naturally occurring (as archives of social processes); they are not obtained by being designed for the purpose of scientific inquiries. Hence, it will be desirable to place them in plausible contexts in which these big data can be generated. As Chen and Venkatachalam (2017) have argued, to the best of current knowledge, agent-based modeling is the only methodology that can help simulate big data, given that big data are often continuous in time and are individual-based (actor-based, agent-based). Despite this being so, it is still not often seen in the literature how an agent-based model is actually applied to simulate a real set of big data. In Chap. 18, "Computational History: From Big Data to Big Simulations," Andrea Nanetti and Siew Ann Cheong have echoed many of the big database challenges mentioned in the previous chapter. In addition, they have proposed new methodologies to take advantage of the big volume of historical data and have used agent-based model simulation to deepen our understanding of why our history has developed the way it has.

Their proposed method consists of two stages. It begins with restructuring the historical data, so that the narratives of events can be extracted and their relationship can be identified. In the second stage, the extracted narratives and identified relationship are employed to build an agent-based model and the agent-based simulation is performed to identify events as tipping points in a society's natural nonlinear life.

They used their research work on the Engineering Historical Memory (EHM) project to demonstrate their methodologies. The benefit of this method is that, based on the large number of simulated histories, we can identify the tipping point where the histories diverge. By comparing the key factors that led to the two different outcomes, and by understanding how they are different, we can have a better understanding of why our history developed the way it did.

1.2.2.4 Big Data Risks

As already mentioned, big data do not constitute a panacea, and their dark side should never be ignored. Not only shouldn't big data be regarded as a solution for everything, but they could further be a trigger for many problems, regardless of being new or existing data. First, we have not been assured that big data will enhance our decision-making capability and quality. Behavioral economists may tell us more on this. There is no clear picture as to how the stupidities of herds can be avoided. The floods of fake news may continue to characterize this year, next year, and the following years. So, despite the efforts being made by computer scientists (Conroy et al. 2015), we have not seen powerful remedies (algorithms) to take away these new forms of "virus." In fact, as we have learned from history, each technology-triggered problem can have human behavior as its catalyst; hence, psychology or behavioral economics may play an even more important role in the era of big data. The second one is more familiar. The very nature of big data implies little privacy and can threaten cybersecurity unless some cautions and actions are taken to protect against privacy rights (Lane et al. 2014).[17]

Nevertheless, the two fundamental challenges above do not exhaust the list of big data threats. Here comes another philosophical question: Who are we? Can we safeguard our identity? How do we know that we are loyal to ourselves? This series of "self" questions is not about cybersecurity, but about cyborgs.[18] In Chap. 19, "A Posthumanist Reflection on the Digital Humanities and Social Sciences," Chia-Rong Tsao has discussed a challenging reality: we are post-human cyborgs, hybrids of humans and technologies. Moreover, the symbiotic relationship between humans and technologies has transformed our cognition with the technologies that constitute the "extended cognitive system." In the big data era where technologies are playing an even more important role in the knowledge development process, Tsao has advocated that we should not ignore the dangers that the development of the digital humanities and computational social sciences may bring.

In conducting research on the digital humanities and social sciences, Tsao has argued that the relationship between researchers and digital tools can be examined from two perspectives. First, the researcher and the digital tool are regarded as collaborators within a hybrid network, which produces knowledge. Second, they not only collaborate with but also co-constitute each other; thus the digital tool changes the researcher when it magnifies and reduces the different dimensions of the world. In other words, the digital tool is not only a passive object manipulated by the researcher, but also a delegated actor that extends the agency of its user and

[17] This feature can be coined as the big data paradox, namely too big to be "small."

[18] In the development of the computational social sciences and humanities, the role of cyborgs is often ignored. For example, in social simulation or agent-based simulation, there is a clear distinction between human agents and software agents, but their possible hybridizations are left out. See Chen et al. (2018).

is capable of betraying users because of neglected rules or incorrect codes. To what extent would the danger be increased by such betraying behavior? Tsao seeks to investigate this question in future work.

1.3 Conclusion and Outlook

With massive amounts of digitized and born-digital data accessible to the public, many social scientists and humanists have explored such data to advance their research. In this book we present works that have incorporated geographic data (Chaps. 2 and 3), text corpus data (Chaps. 4, 5, 6, 7, 8, and 9), and social media data (Chaps. 10 and 11). Meanwhile, surveys of research leveraging big data in finance (Chap. 12), in psychology (Chap. 13), and in digital humanities (Chap. 14) are provided. Moreover, the transition of computational social sciences and digital humanities research from the small-data to the big-data scale presents many challenges Kleiner et al. (2015). We have also discussed some of them in this book (Chaps. 16, 17, 18, and 19).

Although big data research is not going to replace the traditional methods used in studies in the social sciences and humanities in the near future, its importance is undeniably growing. Currently, the number of social scientists embracing big data for new research opportunities is still relatively small. This might be due to the research questions that most social scientists are investigating being traditional ones, which can be solved using small data sets on their own laptops. To inspire more researchers to explore the advantages of big data, we need to provide more demonstrator projects, which can serve as examples to the rest of the community of what big data can provide for scientific development.

Another path to motivate more big data research involves providing assistance on big data tools and methods. For example, Chap. 15 provides a survey and instructions on using cloud computing. However, a more formal approach would be to establish an institution, which provides the research environment mentioned at the beginning of the chapter. In addition to data governance, the institution can offer assistance to social scientists and humanists who are interested in becoming part of the exciting big data research in their fields.

References

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using Big data to understand the human condition: The Kavli HUMAN project. *Big Data, 3*(3), 173–188.

Bauerlein, M. (2008). *The dumbest generation: How the digital age stupefies young Americans and jeopardizes our future (or, don't trust anyone under 30)*. London: Penguin.

Biemann, C., Crane, G. R., Fellbaum, C. D., & Mehler, A. (2014). Computational humanities-bridging the gap between computer science and digital humanities (Dagstuhl Seminar 14301). In *Dagstuhl reports* (Vol. 4, No. 7). Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington: Indiana University Press.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (Vol. 5). Heidelberg: Springer.

Ceron, A., Curini, L., & Iacus, S. M. (2016). *Politics and Big data: Nowcasting and forecasting elections with social media*. Didcot: Taylor & Francis.

Chen, S.-H. (2008). Financial applications: Stock markets. In B. Wang (Ed.), *Wiley encyclopedia of computer science and engineering* (pp. 1227–1244). Hoboken: Wiley.

Chen, S.-H. (2013). Reasoning-based artificial agents in agent-based computational economics. In K. Nakamatsu & L. Jain (Eds.), *The handbook on reasoning-based intelligent systems* (pp. 575–602). Singapore: World Scientific.

Chen, S.-H., & Venkatachalam, R. (2017). Agent-based modelling as a foundation for big data. *Journal of Economic Methodology, 24*(4), 362–383.

Chen, S. H., Kaboudan, M., & Du, Y. R. (2018). Computational economics in the era of natural computationalism. In S. H. Chen, M. Kaboudan, & Y. R. Du (Eds.), *The Oxford handbook of computational economics and finance*. New York: Oxford.

Chen, S.-H., Kampouridis, M., & Tsang, E. (2010). Microstructure dynamics and agent-based financial markets. In *International workshop on multi-agent systems and agent-based simulation* (pp. 121–135). Berlin: Springer.

Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 738 https://doi.org/10.3389/fpsyg.2016.00738.

Clark, A. E., Flèche, S., Layard, R., Powdthavee, N., & Ward, G. (2018). *The origins of happiness: The science of Well-being over the life course*. Princeton: Princeton University Press.

Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of occupy Wall Street. *PLoS One, 8*(5), e64679.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4.

Cooper, C. (2017). *Citizen science: How ordinary people are changing the face of discovery*. London: Gerald Duckworth & Co.

Cruz-Neira, C. (2003). Computational humanities: The new challenge for VR. *IEEE Computer Graphics and Applications, 23*(3), 10–13.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy, 43*(1), 1–20.

Gavin, M. (2014). Agent-based modeling and historical simulation. *DHQ: Digital Humanities Quarterly, 8*(4). Retrieved January 12, 2015, from http://www.digitalhumanities.org/dhq/vol/8/4/000195/000195.html

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics, 1*, 70. https://doi.org/10.1093/qje/qjx045.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447.

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., et al. (2017). Will democracy survive big data and artificial intelligence? *Scientific American, 25*. Retrieved February 27, 2017, from https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ (accessed 27 Feb, 2017)

Jones, M. N. (Ed.). (2016). *Big data in cognitive science*. Hove: Psychology Press.

Joshi, A., Mishra, A., Senthamilselvan, N., & Bhattacharyya, P. (2014). Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 2: Short papers)* (Vol. 2, pp. 36–41).

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kleiner, B., Stam, A., & Pekari, A. (2015). *Big data for the social sciences* (FORS Working Papers, 2015-2).

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge: Cambridge University Press.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Loader, B. D., Vromen, A., Xenos, M. A., Steel, H., & Burgum, S. (2015). Campus politics, student societies and social media. *The Sociological Review, 63*(4), 820–839.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, 30*, 15–29.

McCloskey, D. N. (1983). The rhetoric of economics. *Journal of Economic Literature, 21*(2), 481–517.

McCloskey, D. N. (1998). *The rhetoric of economics*. Madison: University of Wisconsin Press.

Mitra, G., & Xiang, Y. (2016). *Handbook of sentiment analysis in finance*. New York: Albury Books.

Morson, G. S., & Schapiro, M. (2017). *Cents and sensibility: What economics can learn from the humanities*. Princeton: Princeton University Press.

Morzy, M., Kajdanowicz, T., & Kazienko, P. (2017). On measuring the complexity of networks: Kolmogorov complexity versus entropy. *Complexity, 2017*, 3250301.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Broadway Books.

Peters, B. (2012). *The big data gold rush*. New York: Forbes Magazine.

Peterson, R. L. (2016). *Trading on sentiment: The power of minds over markets*. Hoboken: Wiley.

Pinheiro, F. L., Santos, M. D., Santos, F. C., & Pacheco, J. M. (2014). Origin of peer influence in social networks. *Physical Review Letters, 112*(9), 098702.

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). *Sentiment analysis in social networks*. Burlington: Morgan Kaufmann.

Rossbach, S. (1983). *Feng Shui, the Chinese art of placement*. New York: EP Dutton. Inc.

Roy, D., & Zeckhauser, R. (2016). Literary light on decision's dark corner. In R. Frantz, S. H. Chen, K. Dopfer, F. Heukelom, & S. Mousavi (Eds.), *Routledge handbook of behavioral economics* (pp. 230–249). Abingdon: Routledge.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*(5), 885–899.

Seligman, M. E. (2004). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York: Simon and Schuster.

Shiller, R. J. (2017). Narrative economics. *American Economic Review, 107*(4), 967–1004.

Soja, E. (2001). In different spaces: Interpreting the spatial organization of societies. In *Proceedings, 3rd international space syntax symposium* (p. 1-s1).

Soros, G. (2013). Fallibility, reflexivity, and the human uncertainty principle. *Journal of Economic Methodology, 20*(4), 309–329.

Stephens-Davidowitz, S., & Pabon, A. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. New York: HarperLuxe.

Sunstein, C. R. (2008). Neither Hayek nor Habermas. *Public Choice, 134*(1–2), 87–95.

Thompson, A. (2016). *Journalists and Trump voters live in separate online bubbles, MIT analysis shows*. New York: Vice News.

Vaillant, G. E. (2008). *Aging well: Surprising guideposts to a happier life from the landmark study of adult development*. Boston: Little, Brown.

Webster, R. (2012). *Feng Shui for beginners: Successful living by design*. Woodbury: Llewellyn Worldwide.

WHO-CBD. (2015). Connecting global priorities: biodiversity and human health: a state of knowledge review, p. 344.