Embracing the future of the policy sciences: big data in pedagogy and practice

Goyal, N.; El-Taliawi, Ola; Howlett, Michael

DOI
10.4337/9781800376489.00009

Publication date
2021

Document Version
Final published version

Published in
The Future of the Policy Sciences

2. Embracing the future of the policy sciences: big data in pedagogy and practice

Nihit Goyal, Ola G. El-Taliawi and Michael Howlett

2.1 INTRODUCTION

The policy sciences emphasize the generation and mobilization of contextual, multi-method, and normative analyses in order to enable policymakers to bridge the knowledge–action gap and improve their problem-solving capacity and capability (Lasswell, 1970). Exactly how this is to be done, however, is controversial. In one assessment, for example, Peter deLeon (1994) attributed the mixed record of success of the field to its neglect of complexity and its overreliance on positivist, largely economic, methodologies and a simplistic technocratic perspective on government choice. A decade later, Pielke (2004) made much the same argument while, more recently, Cairney and Weible (2017) also have called for a "behavioral turn" in the policy sciences to better explain the effect of individual decision-making through multiple theoretical lenses and mixed methodologies.

Among the key dynamics currently unfolding in contemporary policy analysis and policy-making, however, are the development of new methods better able to deal with increasing complexity (Cairney, 2012). Generally lumped together under the label of "Big Data", these techniques show some promise in being able to better marshal very large amounts of information and conduct policy-relevant analyses that are better able to derive appropriate lessons and insights compared with more traditional techniques and methods.

"Big Data", in this sense, has been defined as a "cultural, technological, and scholarly phenomenon that rests on the interplay of: technology … analysis … and mythology" (Boyd & Crawford, 2012, p. 663). A belief in the ability of Big Data to offer a "higher form intelligence … and the aura of truth, objectivity, and accuracy" has no doubt contributed to its rise (Boyd & Crawford, 2012, p. 663). However, the phenomenon is driven by technological advancements that have permitted the collection of data from three types of sources, which hitherto eluded capture (Kitchin, 2013): (i) directed, such as digital surveillance; (ii) automated, such as traces from digital devices or transactions in a digital network; and, (iii) volunteered, such as crowdsourcing.

The data collected from such sources differs from traditional "small" data in several ways (Kitchin, 2013). This includes volume (size), velocity (speed), variety (structure), exhaustiveness (scope), resolution (granularity), relationality (ability to "merge" with other data), flexibility (ability to transform), and scalability (expansion in scale). In principle, these characteristics allow "users to do things at a large scale that cannot be done at a smaller one" (Bates et al., 2014; D'Orazio, 2017; Mayer-Schönberger & Cukier, 2013, p. 10).

The growth of Big Data, therefore, has implications for both public policy research and practice. Big data analytics can, for example, change or complement traditional approaches to examining political preferences (for example, through social media analytics), the assessment of policy alternatives (for example, beyond cost–benefit evaluation), the study of policy discourse (for example, through computational text analysis of policy documents and speeches), or the evaluation of policy outcomes (for example, using remote sensing). The utility of such approaches has recently been demonstrated, for example, by its use in policy analysis, policymaking, and policy implementation during the COVID-19 outbreak (C. J. Wang et al., 2020; Zou et al., 2020).

However, the prevalence of Big Data in public policy research and teaching remains limited at present (Goyal et al., 2020). This chapter is an evidence-informed appeal for the greater uptake of big data research and teaching in public policy settings. We start by briefly describing key machine learning techniques typically used in Big Data analysis (Section 2.2). Thereafter, we illustrate the applications of Big Data in policymaking and policy analysis using the case of COVID-19 (Section 2.3). In Section 2.4, we present key findings of a bibliometric review of the literature on Big Data in governance and public policy to identify three actions that scholars should take to further research. Section 2.5 then presents the findings from the assessment of policy curricula in programs around the world to argue for more engagement with Big Data analysis and techniques in public policy education. Finally, we conclude the chapter in Section 2.6.

2.2 BIG DATA AND MACHINE LEARNING

To process Big Data, significant analytical capabilities and computational power are necessary. Here, the growth of Big Data has been supported by the advent of machine learning. Machine learning has been implemented in numerous areas, including education, financial modeling, healthcare, manufacturing, marketing, and policing (Jordan & Mitchell, 2015).

In contrast to traditional statistical analysis, which assumes that a stochastic model underlies the data, machine learning uses algorithmic modeling to examine the data without assuming its underlying distribution as given (Breiman, 2001). The key types of machine learning techniques can be grouped as supervised learning, unsupervised learning, and reinforcement learning. While supervised learning includes classification and regression, unsupervised learning includes clustering and dimensionality reduction, among others. The popular modeling techniques for machine learning are gradient descent, artificial neural networks, support vector machines, decision trees, and random forests.

A key advantage of machine learning is that it can be used – with relative ease – to analyze unstructured data. This is evident, for example, in the case of text mining, or computational text analysis. Text mining can be defined as "a knowledge-intensive process in which a user interacts with a document collection … to extract useful information from data sources through the identification and exploration of interesting patterns … in the unstructured textual data in the documents in these collections" (Feldman & Sanger, 2006, p. 1). The key activities in text mining are preprocessing, information extraction and retrieval, classification, clustering, and visualization. Text mining has been used extensively in various disciplines, including medical research for relationship extraction and hypothesis generation (Cohen & Hersh, 2005; Spasic et al., 2005; Srinivasan, 2004), genetics for human phenome classification (van Driel

et al., 2006), information science for patent analysis (Tseng et al., 2007), and management to examine the relationship between customer experience and satisfaction (Xiang et al., 2015) and to investigate market structure based on online user-generated content rather than conventional data (Netzer et al., 2012).

The common techniques in text mining consider a word, or phrase, in the text as the basic unit of analysis and examine distributions, frequent sets, and associations in a document collection (Feldman & Sanger, 2006). As an example, frequency analysis can identify commonly occurring terms while co-occurrence analysis can detect the terms that co-occur frequently. Meanwhile, concordance analysis can shed light on the context in which a term is used, for instance, by identifying the terms that frequently precede or follow a certain term. Further, if the document contains time information, trend analysis can be useful for identifying the rise or fall in the prominence of key terms over time (Lent et al., 1997; Montes-y-Gómez et al., 2001). A term frequency–inverse document frequency (tf-idf) matrix – which represents the ratio of term frequency in a document to term frequency in the entire collection – can identify terms that are discriminative for documents in a collection (Salton & McGill, 1983). Latent semantic indexing has gained favor more recently due to its ability to account for synonymy and polysemy, i.e. multiple words having the same meaning and the same word having multiple meanings (Zhang et al., 2011).

Computational text analysis also relies on natural language processing to analyze text data. Natural language processing (NLP) combines artificial intelligence, computer science, and linguistics to use computational analysis to examine natural language based on knowledge of rules that structure linguistic expression, i.e. grammar (Allahyari et al., 2017). NLP comprises various techniques, such as parts of speech tagging, chunking, and named-entity recognition (Collobert et al., 2011). While parts of speech tagging labels the syntactic role of a term – for instance, adjective, adverb, noun, or verb – chunking identifies parts of a sentence, such as noun phrases or verb phrases, for further processing. Named-entity recognition, on the other hand, labels terms in a sentence based on their attributes, for example "actor", "action", "object", and "location". Thus, these techniques help decompose a text into its constituent elements, extract relevant elements, and identify the relationship amongst different elements. Amongst other areas, NLP has been used in biomedical research for data annotation, information retrieval, and decision support (Aronson, 2001; Kim et al., 2003; Noy et al., 2009).

Topic modeling is another text-mining technique for examining a large document collection. A topic model is based on the premise that a document in a document collection is composed of one or more *latent* topics, which are in turn composed of terms (or words). Topic modeling algorithms, then, use statistical modeling to "discover" topics in the document collection. Latent Dirichlet Allocation (LDA), developed by David M. Blei et al. (2003), was the first topic modeling algorithm and assumed that a document consisted of a "bag of words" (i.e., word sequence was not relevant) and a document collection consisted of a set of documents (i.e., document sequence was not relevant). Numerous topic modeling algorithms or techniques have been developed since then to account for correlation amongst documents by a common author (Rosen-Zvi et al., 2004), prior correlation amongst topics (for example, energy and environment) (Blei & Lafferty, 2007), word order and phrases in a document (X. Wang et al., 2007), temporal sequencing within the document collection (Blei, 2012), examining very short documents such as tweets, and associated structured data regarding documents within a collection (Roberts et al., 2014).

Another text mining technique that is gaining popularity is sentiment analysis, also known as opinion mining, which is "a computational study of opinions, sentiments, emotions, and attitude expressed in texts towards an entity" (Ravi & Ravi, 2015, p. 14). Sentiment analysis involves the detection, extraction, and classification of sentiments in text data (Cambria et al., 2013; Medhat et al., 2014), typically using a pre-defined lexicon (Taboada et al., 2011). Sentiment analysis has numerous applications, including for pricing product features (Archak et al., 2011) and understanding the political preferences of citizens.

2.3 APPLICATION TO PUBLIC POLICY: ILLUSTRATION USING THE CASE OF COVID-19

As one might expect, health policy has been a key area for application of machine learning in the case of COVID-19. Brinati et al. (2020), for example, argued that blood test analysis with classification techniques in machine learning is a feasible alternative to the reverse transcription polymerase chain reaction (rRT-PCR) test – the current gold standard in detection of the disease – especially for countries with limited health infrastructure. Focusing on patients diagnosed with COVID-19, F. Y. Cheng et al. (2020) developed a random forest model to predict – with about 75 percent accuracy – the likelihood of imminent transfer to intensive care. In a more macro-level application, Abdulmajeed et al. (2020) investigated the application of a multi-method "ensemble" approach, combining various machine learning techniques for forecasting the spread of COVID-19 in Nigeria.

Beyond this, machine learning has already been applied for detecting, tracing, and checking the spread of the outbreak in the community. Illustratively, Zou et al. (2020) found that the use of mobile technologies, big data, and artificial intelligence contributed to the success of the COVID-19 response in Shenzhen by increasing the accessibility of health services and curbing fake news. In addition, the Taiwanese government was able to respond rapidly to the crisis by integrating its national health insurance data with its immigration and customs data to create Big Data that enabled real-time alerting and case identification (C. J. Wang et al., 2020). This solution was layered with new technologies, including QR code scanning, to assess travelers' infectious risk at ports of entry.

Further, Big Data and machine learning have also been used to assess the fallout of the pandemic on other policy areas. Ou et al. (2020), for example, employed machine learning – specifically, the neural network technique – to examine the impact of COVID-19 on demand for gasoline under different scenarios by modeling variations in mobility based on demographics, pandemic spread, and policy response. To facilitate timely reactivation of tourism, Gallego and Font (2020) demonstrated the use of Big Data on air passenger searches from the travel website SkyScanner to assess willingness to travel in the short- and medium-term.

Moreover, the use of Big Data analytics is not limited to health policy analysis and has also informed its evaluation. Illustratively, Liu et al. (2020) used satellite data on night-time light to assess the impact of COVID-19 on social behavior in mainland China more generally and Wuhan more specifically. In another example, James et al. (2020) re-purposed existing smart city infrastructure – the Newcastle Urban Observatory – to create a dashboard for real-time policy evaluation using Big Data and machine learning. Meanwhile, Gupta et al. (2020) used Big Data from smart phone devices to examine the efficacy of social distancing measures in

the United States, and Vaid et al. (2020) used machine techniques to compare the effectiveness of policy rsponses to the spread of COVID-19 in Canada, Sweden, and the United States.

Several studies have advanced knowledge on the discursive aspect of public policy during the pandemic using machine learning. Lopez-Rico et al. (2020), for example, used affinity analysis, a data-mining technique, of online survey responses to examine polarization and trust among the Spanish citizenry. Sear et al. (2020) analyzed data on online communities on Facebook using topic modeling to investigate the 'infiltration' of online conversations on COVID-19 by those opposing vaccination. In another such application, Samuel et al. (2020) assessed the effect of the pandemic on the general public through text analysis of Twitter data using Naïve Bayes Method and logistic regression and then calculating fear sentiment using the processed text.

To facilitate comparative policy analysis and learning, C. Cheng et al. (2020) compiled a dataset of over 13,000 daily policy responses across more than 190 countries since 31 December 2019. In this effort, they used machine learning to automate data collection from news articles and natural language processing for identification and classification of policy announcements. Using this dataset with the OECD (2020) Country Policy Tracker, Capano et al. (2020) highlighted variation in the start, speed, and scope of the cross-national policy responses to COVID-19 through the application of natural language processing and topic modeling.

These numerous applications of Big Data analytics to public policy practice and research in a relatively short time in the case of COVID-19 highlight the potential of the field to contribute to the policy sciences through contextual, multi-method, and normative analysis, furthering knowledge of the policy process as well as knowledge in the policy process. In the next section, we highlight how the policy sciences can exploit this potential to move a step closer to Lasswell's dream.

2.4 ENGAGING WITH BIG DATA IN POLICY RESEARCH

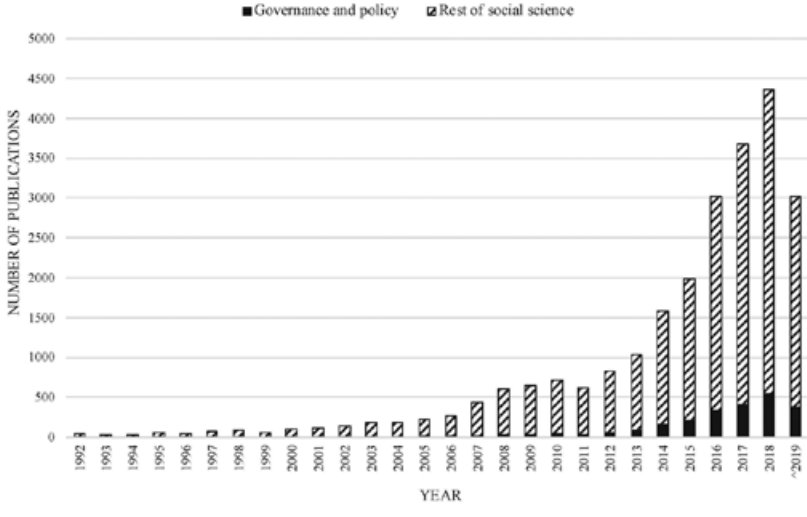
In this section, we identify opportunities for the policy sciences to engage with the Big Data phenomenon based on a bibliometric review of its current use in the field. The data for the bibliometric analysis were obtained using the following search query on the Web of Science database (social science research only): "automated text analysis" OR "big data" OR "computational text analysis" OR "data mining" OR "data science" OR "machine learning" OR "natural language processing" OR "opinion mining" OR "sentiment analysis" OR "text analytics" OR "text as data" OR "text mining" OR "topic model*." The search, conducted on August 9, 2019, returned 24,445 results. The data returned by the search were downloaded and processed using the Bibliometrix package in R (Aria & Cuccurullo, 2017). Publications potentially pertaining to governance or public policy were identified by searching for "govern*" OR "policy" OR "policies" in the title, abstract, and keyword list of publications in this dataset. This final analysis dataset included nearly 2,500 articles on the topic.

2.4.1 Upscale of policy relevant research on Big Data

The mention of Big Data in social sciences research more generally, and public policy research more specifically, has increased significantly over the years. In 1992, for example, only about

50 documents on Big Data were published in the Web of Science database. This increased to over 100 documents in the year 2000, over 500 documents in 2008, and over 1000 documents in 2013. Since 2016, over 3000 documents that mention Big Data as part of their topic have been published each year.

Within this body of research, the first article concerning governance or public policy was published in 1992. It is only after 2004 that 10 or more documents have been published each year on Big Data and governance or public policy. Since then, the volume of research has increased by over 30 percent per annum to reach annual production of over 150 documents in 2014 and over 500 documents in 2018. Thus, despite recent attention to Big Data, governance or public policy research comprises only approximately 10 percent of the annual social sciences production on the topic (Figure 2.1).



Source: Author's calculation.

Figure 2.1 Number of publications related to big data in the social sciences

This indicates that most of the research surrounding Big Data in the social sciences is not explicitly policy focused. This suggests that the use of Big Data in the policy sciences is still not mainstream. To further use it in the policy sciences, a concerted push at embracing "basic" science to tease out its policy relevance while also undertaking more applied research using Big Data and machine learning can, therefore, be fruitful.

2.4.2 Topics analyzed using Big Data in the policy sciences

Using structural topic modelling (Blei et al., 2003; Roberts et al., 2014), we clustered the existing research on Big Bata in governance or public policy into seven themes. Based on decreasing order of prevalence, these are: (i) Topic 7 on machine learning and predictive anal-

ysis; (ii) Topic 2 on big data for governance and public policy; (iii) Topic 3 on the governance of big data and its use in healthcare; (iv) Topic 4 on computational text analysis; (v) Topic 5 on the impact of big data on education and society; (vi) Topic 6 on analysis of high-frequency data; and, (vii) Topic 1 on the use of big data at the interface of private and public sectors. The prevalence of these themes along with frequently occurring terms that are exclusive to each theme are shown in Figure 2.2.

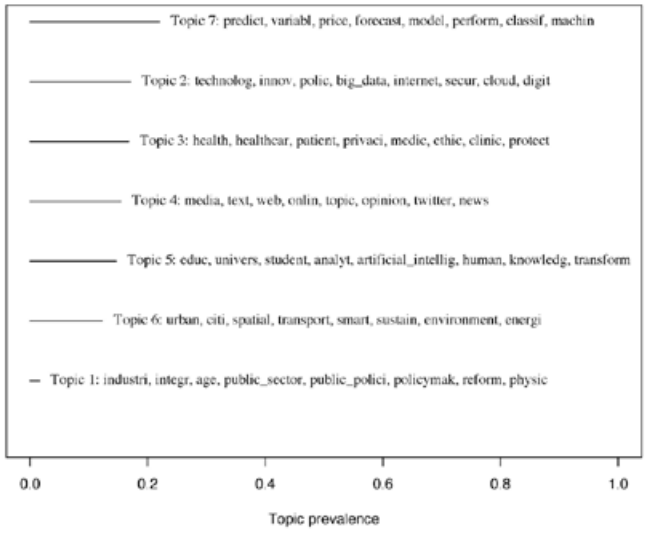


Figure 2.2 Topic prevalence and frequent but exclusive terms for each topic

A close examination of these themes reveals that they span policymaking, policy analysis, and policy studies. Topic 2, for example, emphasizes the potential of big data to contribute to governance and public policy. Within this broader theme, Zheng and Zhang (2016) explored the use of Big Data in government audits and shed light on the challenges and solutions therein. Meanwhile, X. D. Wang et al. (2017) highlighted the role of (commercial) big data in supporting economic governance and promoting growth and prosperity. With a focus on policing, Ning (2017) analyzed trends in cybercrime and provided recommendations for utilizing publicly available big data for ensuring cybersecurity. Other studies with high prevalence of this topic have highlighted the potential of Big Data for government and public management (Jin, 2015; Ping, 2018), urged the Chinese government to tap into the potential of Big Data (He et al., 2013), and examined its use in a local government in China (Guo et al., 2017).

Meanwhile, Topic 3 delves into the governance of big data and its use (especially) in healthcare. Canaway et al. (2019), for example, discussed issues around the collection and sharing of "big" health data and argued for a transparent approach towards risk analysis and communication. Relatedly, Dorey et al. (2018) examined the awareness of ethics associated with collection and processing of patient data among health staff in Switzerland, and Woolley et al. (2016) explored the ethical and social implications of citizen science in biomedicine. Outside the health sector, Gilmore (2016) analyzed the implications of the Big Data phenomenon on

the science of human development and argued for a holistic approach to its governance while Easton-Calabria and Allen (2015) highlighted the need for ethics governing the use of Big Data in civil society organizations.

Closely related to Topic 3, Topic 5 captures studies on the impact of Big Data on education and society more broadly. Gobert et al. (2015), for example, proposed the use of education data mining to measure inquiry skills among learners. In a more reflective essay, Gulson and Webb (2018) delved into the influence of computation on understanding life and analyzed its effect on education. Numerous studies in this cluster also focus on the relationship between artificial intelligence and law. Berman (2018), for example, examined whether reliance on artificial intelligence for policymaking and implementation is consistent with rule of law. Relatedly, Levendowski (2018) argued for a relook at copyright law to make otherwise excluded material available to computer algorithms and thereby reduce bias in artificial intelligence. Meanwhile, Pasquale (2019) argued that the flexibility and subtlety of legal language limits the scope for legal automation.

In contrast to the topics presented above, some other themes in the research concentrate predominantly on the use of Big Data for policy analysis. Topic 7, for example, pertains to the use of machine learning and predictive analysis for public policy. This is illustrated by Mitra and Chattopadhyay (2017), who used data-mining tools to predict the impact on monsoon rainfall on food inflation in India. In the area of finance, Emrouznejad and Anouze (2010) performed data envelopment analysis using classification and regression trees to analyze the productivity of banks and discuss implications for public policy. Meanwhile, Huber and Imhof (2019) combined machine learning with statistical screening to detect and predict anti-competitive bid-rigging in the Swiss construction sector. Other studies that exemplify research within this topic include estimation of energy demand (Sanchez-Oro et al., 2016), forecasting of housing prices (Plakandaras et al., 2015), temporal assessment of dengue incidence (Carvajal et al., 2018), and comparison of indicators for food security (Hossain et al., 2019) using machine learning.

While research within Topic 6 is also for policy analysis, it is concerned primarily with high-frequency spatial or temporal data (often in the urban context). Illustratively, Lim et al. (2019) combined spatial data on population distribution with national statistics to examine how patterns of urbanization influence greenhouse gas emissions. Similarly, Ahn and Sohn (2019) used information on energy consumption with geographic information system data in Seattle to investigate the effect of urban form at the neighborhood level on building energy consumption. Studies in this cluster have also used spatial and temporal data for transportation policy, for example, by analyzing the impact of investment on train system performance in Los Angeles (Giuliano et al., 2016), proposing sites for electric vehicle charging stations based on an assessment of data from taxis in Beijing (Cai et al., 2014), and examining the relationship between bus travel and cycling in the city of Tel Aviv (Levy et al., 2019).

Topic 1 is largely concerned with the use of Big Data in the private sector and its interface with the public sector, primarily in the case of industrial or social policy. Yang and Chen (2015), for example, analyzed the integration of informatization and industrialization in the big data era and delve into its implications for industrial policy. Smith et al. (2015) discussed the potential of commercial Big Data in informing social policy. Similarly, Roderick (2014) examined the use of big data in consumer finance and highlighted the need for more regulation on this front. In a different strand, Li (2017) used text mining to identify business opportunities

as well as areas for support in the Chinese old, aging industry. Meanwhile, Fan (2016) argued for greater use of information technology in education in the Big Data era to integrate research and teaching. In another application of text mining, Breit et al. (2017) studied the reporting of public sector information policy in newspapers in Queensland to examine the role of media in policy diffusion.

Different from the above – with an interest in the use of techniques such as opinion mining, sentiment analysis, and topic modeling – Topic 4 is closer to policy studies than policy analysis or policy practice. This is reflected by the most frequently occurring terms in this topic, such as news, online, opinion, social media, topic, and web. Prominent studies in this cluster focus on the application of computational text analysis to examine, for example, policy discourse, issue framing, public opinion, and research on innovation management (Alashri et al., 2016; Anas, 2018; Kawamura et al., 2019; Koltsova et al., 2016; Lashari & Wiil, 2016; Lee & Kang, 2018; Soroka et al., 2015).

The above analysis suggests that application of Big Data analytics for governance or public policy is even smaller than indicated by the size of this dataset. Only three of the six topics discovered here primarily demonstrate the use of such techniques for policy research: Topic 7, Topic 4, and Topic 6. On the other hand, Topic 2 and Topic 3 examine issues surrounding the governance of these technologies as well as their use in policymaking. Finally, Topic 5 and Topic 1 represent studies that delve into both policy analysis and governance in areas such as industrial and social policy. Thus, the topic modeling analysis reinforces the need for upscaling policy relevant research using Big Data.

More importantly, this analysis highlights the opportunity Big Data presents for bringing policy studies, policy analysis, and policymaking closer to one another. The seven themes discovered in this study provide evidence that – despite outstanding issues regarding its governance – Big Data can be applied for creating knowledge of the policy process as well as knowledge in the policy process. To take advantage of this opportunity, however, policy scientists should design and execute research that harnesses Big Data to cut across these themes and demonstrate the value added of contextual, interdisciplinary, and normative research in advancing scholarly knowledge and informing policymaking.

2.4.3 Diversity in institutional footprint and geographies of research

Not only is the general use of Big Data weak in the policy sciences but most of the scientific production using Big Data in a governance or policy perspective is limited to a few countries. Based on the institutional affiliation of the corresponding author, the countries with the most publications in the dataset are shown in Table 2.1. The United States, China, the United Kingdom, Australia, and South Korea constitute the top five of this list. Together, the top ten countries account for over 75 percent of the publications in this dataset. Outside the Organization for Economic Co-operation and Development (OECD), with the exception of China, the countries with the most publications in this dataset are: Brazil and Romania (25 publications each), Singapore (23 publications), Russia (22 publications), and India (21 publications). Meanwhile, South Africa (nine publications) is the only African country with more than five publications in the dataset.

An examination of the institutional footprint of research on Big Data in governance or public policy also highlights the highly uneven contribution of specific institutions to scientific

Table 2.1 Publications by country affiliation of the corresponding author

Country	Number of publications	Percent of total (%)
United States of America	652	27.3
China	464	19.4
United Kingdom	223	9.3
Australia	103	4.3
South Korea	97	4.1
Germany	65	2.7
Canada	64	2.7
Netherlands	64	2.7
Italy	60	2.5
Spain	58	2.4

Table 2.2 Publications by institutional affiliation of the authors

Author affiliations	Number of publications
Stanford University	57
Arizona State University	56
Tsinghua University	54
University of Oxford	45
Harvard University	44
University of Pennsylvania	43
University of Texas (Austin)	43
University of Hong Kong	41
University of Michigan	41
Delft University of Technology	39
Wuhan University	39
University of Tokyo	38
University of Queensland	37
University of Washington	36
University of California (Berkeley)	34
Yonsei University	32

knowledge in this field. While authors publishing on the topic represent over 2300 institutions worldwide, out of these approximately 900 institutions are represented just once in the dataset and only about 300 institutions have six or more publications in this dataset. The institutions with the most publications in the dataset are Stanford University, Arizona State University, Tsinghua University, University of Oxford, and Harvard University (Table 2.2). Outside North America and Europe, other countries with institutions that have published actively in this field include China, Japan, Australia, South Korea, and Singapore. Here too, we observe that, with the exception of China, no other low- or middle-income country is represented amongst the top 50 institutions in this dataset.

The limited geographical base of research in this field is possibly indicative of a broader trend in policy sciences. In a bibliometric review of publications in *Policy Sciences* – a premier journal to advance the policy sciences orientation – Goyal (2017) found that nearly 70 percent of the authors in the five-decade history of the journal were from institutions based in the US while hardly any were from institutions in Africa, South America, or the transition economies. This indicates a significant scope, as well as need, to diversify geographies and expand the institutional footprint of research in public policy generally which is also reflected in the patterns of use of Big Data in the context of governance or public policy. Policy scientists should explore new collaborations and partnerships involving scholars in the Global South to correct this anomaly.

2.5 ENGAGING WITH BIG DATA IN PUBLIC POLICY EDUCATION

Before such work can be undertaken, however, Big Data techniques and methods of analysis must be taught to aspiring analysts. In this section, we analyze Big Data course offerings of policy programs worldwide to determine the extent to which current pedagogy is giving sufficient attention to this topic. We find that while the use of Big Data as an analytical tool has begun to infiltrate many social science disciplines, including public administration, and is being increasingly used by government organizations, its use in policy analysis and policy education is trailing behind.

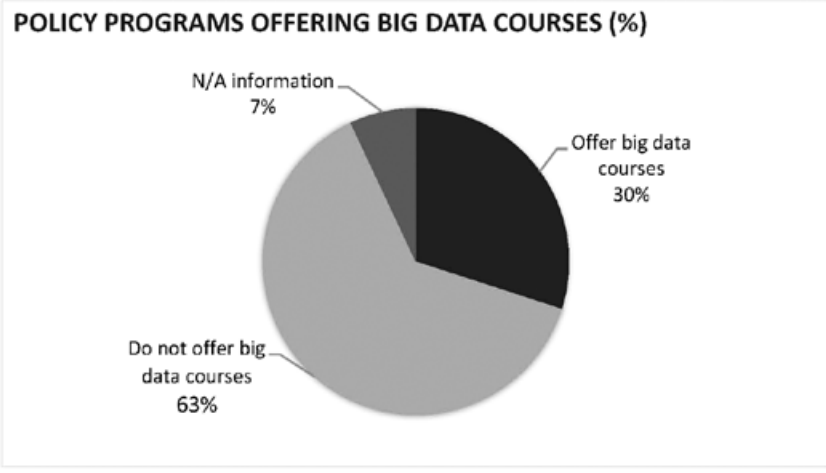
Recent attempts to embed public service teaching with Big Data techniques and methods include the 'Teaching Public Service in the Digital Age' project (https://www.teachingpublicservice.digital/). This project was recently launched by an international group of public policy and administration instructors. Their aim is to help equip future public servants with competences fit for the digital era by providing teaching materials for instructors, including an open-access syllabus. Other scholarly efforts to provide guidance include the article by Mergel (2016) in which she proposes a syllabus design for teaching Big Data in the public sector.

To explore the preponderance of policy programs offering Big Data analysis in their curriculum, we examined 122 programs, including the International Public Policy Association's (IPPA) institutional members. We excluded 33 results for not offering policy degrees. The remaining 88 programs were analysed as to their course offerings. We code a policy program offering Big Data pedagogy if it offers one or more, core or elective, modules that include the keywords "big data" or "informatics" or "analytics", either in the title or the course description. We used university websites as our data sources.

Among the results, we found that only 30 percent of programs included in our sample offer such courses, while 63 percent did not offer Big Data courses in their degree courses. The remaining 7 percent did not provide sufficient information on their websites to give a definitive conclusion (Figure 2.3).

When we analysed the geographic distribution of programs that did offer Big Data courses, we found predominance in the provision of such courses in North America (62 percent), followed by Europe (15 percent), with the remaining 23 percent was dispersed in universities in Asia, Africa, Australia and Latin America.

Further, three types of Big Data offerings were found in our sample: courses, degrees and concentrations, with the majority of programs offering courses only. Course titles include



Source: Authors' calculation.

Figure 2.3 Percentage of policy programs offering big data courses

variations, such as *Decision Support System and Data Analytics for Public Policy*; *Big Data and Government*; *Evidence and Analysis in Public Policy* with emphasis on Big Data and machine learning; and others. Less predominant in our sample are degrees and concentrations in policy analysis and Big Data (7 percent). Examples of such degrees offered include: *Master of Science in Policy Analytics*, while examples of concentrations include a *Concentration in Data Analytics and Program Evaluation*.

Since course syllabi are not accessible on university websites, we were unable to examine the pedagogical techniques used. Future content analysis of syllabi can be conducted to reveal the predominant techniques used by instructors, and the kind of pedagogical gaps that currently exist in the field. Qualitative and quantitative research methods modules can also be analyzed to determine whether big data techniques are embedded in their course learning objectives. Further, analyzing whether Big Data courses offered by policy programs are core or elective courses can reveal the extent to which such techniques are considered as essential components of the skills needed to be imparted on students. The rate of student uptake of these courses may also be indicative and can be assessed by means of institutional surveying.

With regards to our own findings, they point to a clear pedagogical gap in the offering of Big Data courses to future policy researchers and scholars. While further analysis may corroborate this, it again reflects room for advancement if the field were to catch up with this trend and make use of the potential that Big Data analytics offer.

2.6 CONCLUSION

While the emergence of Big Data has raised new concerns for governance and policymaking, it has also created opportunities for public policy research, teaching and practice from a policy

sciences perspective. Public policy research, whether on the policy process or policy analysis, can benefit from the application of Big Data analysis to complement traditional techniques such as polling, surveying, cost–benefit analysis, econometric evaluation, and content analysis. In this study, we provided a brief description of Big Data techniques and their applications in public policy, examined the use of these techniques in public policy research, and analyzed the extent to which such techniques are taught in policy programs around the world.

A survey of the literature on Big Data and a review of the key techniques and models in Big Data analytics as they have been applied in the COVID-19 pandemic indicate that Big Data has the potential to create a paradigm shift in public policy research and practice with the development of new epistemologies and new types of analyses.

While research on Big Data and governance or policy has been slow to take off, the number of journal articles and conference papers on the topic has increased significantly over the last decade. Although this is a promising trend, much research on the topic is still dominated by the USA, China, and the UK, and relatively elite institutions account for most scientific production on the topic. This research has witnessed limited transnational collaboration and the Global South, except for China, has largely been left out of this "revolution". Further, a closer examination of the sources in which the research has been published suggests that while governance and policy analysis have attracted some attention, policy studies have yet to make adequate use of this opportunity.

Much of this existing research has focused on issues surrounding the governance of Big Data analytics and techniques, such as ethics, privacy, and surveillance, rather than apply them to the better understanding and resolution of significant social problems.

In addition, our analysis of over 120 institutions indicated that only about 30 percent of these offer courses pertaining to Big Data analytics. And there is high geographic variation in the diffusion of Big Data analytics in public policy pedagogy as well – with over 60 percent of the programs in North America. The implication being that the potential of Big Data to fundamentally change public policy is unlikely to be realized if the status quo in policy-related pedagogy and practice continues.

Can this gap be addressed so that policy sciences may harness the potential of Big Data, and, if so, how?

Based on our analysis, several actions that policy scientists should take to address the situation can be discerned. First, scholars should focus on increasing the volume of policy relevant social science research using Big Data and machine learning. Second, policy scientists should aim to diversify the geographical and institutional footprint of such research. More multi-country and multi-institutional collaboration can potentially increase the diffusion of Big Data analytics in continental Europe and the Global South. In addition, multi-sectoral collaboration – especially involving fields that have been early adopters of Big Data analytics, such as communications, education, healthcare, management, and sustainability – can not only produce comparative research but also help identify potential applications in other areas. Third, a policy sciences perspective can encourage the production and dissemination of research using Big Data analytics in innovative ways that encompass policy studies and policy analysis. Existing research on Big Data has largely been published in – established or upcoming – sources that are not mainstream in policy sciences. A new journal focusing on Big Data analytics in public policy could help integrate knowledge of the policy process with knowledge in the policy process in the use of Big Data (analytics) in policy research. Fourth, we find

that course syllabi for public policy courses in general and Big Data courses in particular are not easily available online. Opening up the courseware can provide useful templates for institutions with lesser resources as well as scholars with expertise in traditional methods to engage with Big Data analytics and deploy it in research and teaching in new and innovative ways.

REFERENCES

Abdulmajeed, K., Adeleke, M., & Popoola, L. (2020). Online forecasting of Covid-19 cases in Nigeria using limited data. *Data in Brief*, 30, 8. doi:10.1016/j.dib.2020.105683

Ahn, Y., & Sohn, D. W. (2019). The effect of neighbourhood-level urban form on residential building energy use: A GIS-based model using building energy benchmarking data in Seattle. *Energy and Buildings*, 196, 124–133. doi:10.1016/j.enbuild.2019.05.018

Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., & Desouza, K. C. (2016). *An Analysis of Sentiments on Facebook during the 2016 US Presidential Election*. New York: IEEE.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint* arXiv:1707.02919.

Anas, A. (2018). An investigation of sentiment and themes from Twitter for Brexit-2016. In V. Cunnane & N. Corcoran (Eds), *Proceedings of the 5th European Conference on Social Media* (pp. 8–12). Nr Reading: Acad Conferences Ltd.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8), 1485–1509. doi:10.1287/mnsc.1110.1370

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap Program. *Journal of the American Medical Informatics Association*, 17–21. Retrieved from ://WOS:000172263400005

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123–1131.

Berman, E. (2018). A Government of Laws and Not of Machines. *Boston University Law Review*, 98(5), 1277–1355. Retrieved from ://WOS:000448668400003

Blei, D. M. (2012). Probabilistic topic models. *Communications of ACM*, 55(4), 77–84. doi:10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. *Annals of Applied Statistics*, 1(1), 17–35. doi:10.1214/07-aoas114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, 15(5), 662–679. doi:10.1080/1369118X.2012.678878

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. doi:10.1214/ss/1009213726

Breit, R., Fitzgerald, R., Liu, S., & Neal, R. (2017). How Queensland newspapers reported public sector information reform. *Media International Australia*, 162(1), 90–106. doi:10.1177/1329878X16680655

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, 44(8), 12. doi:10.1007/s10916-020-01597-4

Cai, H., Jia, X. P., Chiu, A. S. F., Hu, X. J., & Xu, M. (2014). Siting public electric vehicle charging stations in Beijing using big-data informed travel patterns of the taxi fleet. *Transportation Research Part D-Transport and Environment*, 33, 39–46. doi:10.1016/j.trd.2014.09.003

Cairney, P. (2012). Complexity theory in political science and public policy. *Political Studies Review*, 10(3), 346–358.

Cairney, P., & Weible, C. M. (2017). The new policy sciences: combining the cognitive science of choice, multiple theories of context, and basic and applied analysis. *Policy Sciences*, 50(4), 619–627. doi:10.1007/s11077-017-9304-2

Cambria, E., Schuller, B., Xia, Y. Q., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), 15–21. doi:10.1109/mis.2013.30

Canaway, R., Boyle, D. I. R., Manski-Nankervis, J. A. E., Bell, J., Hocking, J. S., Clarke, K., . . . Emery, J. D. (2019). Gathering data for decisions: Best practice use of primary care electronic records for research. *Medical Journal of Australia*, 210, S12–S16. doi:10.5694/mja2.50026

Capano, G., Howlett, M., Jarvis, D. S. L., Ramesh, M., & Goyal, N. (2020). Mobilizing policy (in) capacity to fight COVID-19: Understanding variations in state responses. *Policy and Society*, 39(3), 285–308. doi:10.1080/14494035.2020.1787628

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infectious Diseases*, 18, 15. doi:10.1186/s12879-018-3066-0

Cheng, C., Barceló, J., Hartnett, A. S., Kubinec, R., & Messerschmidt, L. (2020). COVID-19 government response event dataset (CoronaNet v.1.0). *Nature Human Behaviour*, 4(7), 756–768. doi:10.1038/s41562-020-0909-7

Cheng, F. Y., Joshi, H., Tandon, P., Freeman, R., Reich, D. L., Mazumdar, M., . . . Kia, A. (2020). Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *Journal of Clinical Medicine*, 9(6), 12. doi:10.3390/jcm9061668

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1), 57–71. doi:10.1093/bib/6.1.57

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug), 2493–2537.

D'Orazio, P. (2017). Big data and complexity: Is macroeconomics heading toward a new paradigm? *Journal of Economic Methodology*, 24(4), 410–429. doi:10.1080/1350178X.2017.1362151

deLeon, P. (1994). Reinventing the policy sciences: Three steps back to the future. *Policy Sciences*, 27(1), 77–95. Retrieved from www.jstor.org/stable/4532307

Dorey, C. M., Baumann, H., & Biller-Andorno, N. (2018). Patient data and patient rights: Swiss healthcare stakeholders' ethical awareness regarding large patient data sets – a qualitative study. *BMC Medical Ethics*, 19, 14. doi:10.1186/s12910-018-0261-x

Easton-Calabria, E., & Allen, W. L. (2015). Developing ethical approaches to data and civil society: From availability to accessibility. *Innovation-the European Journal of Social Science Research*, 28(1), 52–62. doi:10.1080/13511610.2014.985193

Emrouznejad, A., & Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree – a case of banking efficiency. *Expert Systems*, 27(4), 231–246. doi:10.1111/j.1468-0394.2010.00516.x

Fan, C. X. (2016). Study on the integration reform of teaching and research about the situation and policy course in Chinese colleges and universities. In X. Xiao, S. B. Tsai, & R. Feng (Eds), *Proceedings of the 2016 2nd International Conference on Social Science and Higher Education* (Vol. 53, pp. 132–134). Paris: Atlantis Press.

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Gallego, I., & Font, X. (2020). Changes in air passenger demand as a result of the COVID-19 crisis: Using Big Data to inform tourism policy. *Journal of Sustainable Tourism*, 20. doi:10.1080/09669582.2020.1773476

Gilmore, R. O. (2016). From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews – Cognitive Science*, 7(2), 112–126. doi:10.1002/wcs.1379

Giuliano, G., Chakrabarti, S., & Rhoads, M. (2016). Using regional archived multimodal transportation system data for policy analysis: A case study of the LA Metro Expo Line. *Journal of Planning Education and Research*, 36(2), 195–209. doi:10.1177/0739456x15604444

Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity*, 18, 81–90. doi:10.1016/j.tsc.2015.04.008

Goyal, N. (2017). A "review" of policy sciences: bibliometric analysis of authors, references, and topics during 1970–2017. *Policy Sciences*, 50(4), 527–537. doi:10.1007/s11077-017-9300-6

Goyal, N., El-Taliawi, O. G., & Howlett, M. (2020). The prevalence of big data analytics in public policy: Is there a research-pedagogy gap? In S. Nair & N. Varma (Eds), *Emerging Pedagogies for Public Policy Education in Asia*. London: Palgrave Macmillan.

Gulson, K. N., & Webb, P. T. (2018). 'Life' and education policy: Intervention, augmentation and computation. *Discourse-Studies in the Cultural Politics of Education*, 39(2), 276–291. doi:10.1080/01596306.2017.1396729

Guo, W. G., Xie, J. Q., Zhu, Z., Lin, Q. M., & Li, X. H. (2017). Analysis and research on the development and application of Big Data in Foshan. In N. Xin, K. ElHami, & Z. Kun (Eds), *Proceedings of the 2017 7th International Conference on Social Network, Communication and Education* (Vol. 82, pp. 527–531). Paris: Atlantis Press.

Gupta, S., Nguyen, T. D., Rojas, F. L., Raman, S., Lee, B., Bento, A., . . . Wing, C. (2020). Tracking public and private responses to the COVID-19 epidemic: Evidence from state and local government actions. *National Bureau of Economic Research Working Paper Series*, No. 27027. doi:10.3386/w27027

He, L., Zhou, Q. Q., & Hua, Q. S. (2013). Consideration on construction and development of national scientific and technological resources in the age of Big Data. In G. Lee (Ed.), *2013 International Conference on Management Innovation and Business Innovation* (Vol. 15, pp. 557–562). Singapore: Singapore Management & Sports Science Inst Pte Ltd.

Hossain, M., Mullally, C., & Asadullah, M. N. (2019). Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food Policy*, 84, 77–91. doi:10.1016/j.foodpol.2019.03.001

Huber, M., & Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65, 277–301. doi:10.1016/j.ijindorg.2019.04.002

James, P., Das, R., Jalosinska, A., & Smith, L. (2020). Smart cities and a data-driven response to COVID-19. *Dialogues in Human Geography*, 10(2), 255–259. doi:10.1177/2043820620934211

Jin, T. (2015). The innovation research of public management model based on Big Data. In Z. L. Yao & Y. Chen (Eds), *Proceedings of the 2015 International Conference on Economics, Social Science, Arts, Education and Management Engineering* (Vol. 38, pp. 200–203). Paris: Atlantis Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. doi:10.1126/science.aaa8415

Kawamura, K., Kobashi, Y., Shizume, M., & Ueda, K. (2019). Strategic central bank communication: Discourse analysis of the Bank of Japan's Monthly Report. *Journal of Economic Dynamics & Control*, 100, 230–250. doi:10.1016/j.jedc.2018.11.007

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19, i180–i182. doi:10.1093/bioinformatics/btg1023

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262–267. doi:10.1177/2043820613513388

Koltsova, O., Koltcov, S., & Nikolenko, S. (2016). Communities of co-commenting in the Russian LiveJournal and their topical coherence. *Internet Research*, 26(3), 710–732. doi:10.1108/IntR-03-2014-0079

Lashari, I. A., & Wiil, U. K. (2016). Monitoring public opinion by measuring the sentiment of retweets on Twitter. In C. Bernadas & D. Minchella (Eds), *Proceedings of the 3rd European Conference on Social Media* (pp. 153–161). Nr Reading: Acad Conferences Ltd.

Lasswell, H. D. (1970). The emerging conception of the policy sciences. *Policy Sciences*, 1(1), 3–14.

Lee, H., & Kang, P. (2018). Identifying core topics in technology and innovation management studies: A topic model approach. *Journal of Technology Transfer*, 43(5), 1291–1317. doi:10.1007/s10961-017-9561-4

Lent, B., Agrawal, R., & Srikant, R. (1997). *Discovering Trends in Text Databases.* Paper presented at the KDD.

Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. *Washington Law Review*, 93(2), 579–630. Retrieved from ://WOS:000445990600001

Levy, N., Golani, C., & Ben-Elia, E. (2019). An exploratory study of spatial patterns of cycling in Tel Aviv using passively generated bike-sharing data. *Journal of Transport Geography*, 76, 325-334. doi: 10.1016/j.jtrangeo.2017.10.005

Li, C. G. (2017). Market opportunity and policy support for Chinese old aging industry: An application of text mining. In J. Vachal, M. Vochozka, & J. Horak (Eds), *Innovative Economic Symposium 2017* (Vol. 39). Cedex A: E D P Sciences.

Lim, J., Kang, M., & Jung, C. (2019). Effect of national-level spatial distribution of cities on national transport CO_2 emissions. *Environmental Impact Assessment Review*, 77, 162–173. doi:10.1016/j.eiar .2019.04.006

Liu, Q., Sha, D. X., Liu, W., Houser, P., Zhang, L. Y., Hou, R. Z., . . . Yang, C. W. (2020). Spatiotemporal patterns of COVID-19 impact on human activities and environment in mainland China using night-time light and air quality data. *Remote Sensing*, 12(10), 14. doi:10.3390/rs12101576

Lopez-Rico, C. M., Gonzalez-Esteban, J. L., & Hernandez-Martinez, A. (2020). Polarization and trust in Spanish media during the COVID-19. Identification of audience profiles. *Revista Espanola De Comunicacion En Salud*, S77–S89. doi:10.20318/recs.2020.5439

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that will Transform how we Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. doi:10.1016/j.asej.2014.04.011

Mergel, I. (2016). Big Data in public affairs education. *Journal of Public Affairs Education*, 22(2), 231–248. Retrieved from www.jstor.org/stable/44114760

Mitra, S. K., & Chattopadhyay, M. (2017). The nexus between food price inflation and monsoon rain-fall in India: Exploring through comparative data mining models. *Climate and Development*, 9(7), 584–592. doi:10.1080/17565529.2016.1174662

Montes-y-Gómez, M., Gelbukh, A., & López-López, A. (2001). Discovering association rules in semi-structured data sets. Paper presented at the Proceedings of the Workshop on Knowledge Discovery from Distributed, Dynamic, Heterogeneous, Autonomous Data and Knowledge Source at the *17th International Joint Conference on Artificial Intelligence* (IJCAI'2001). Seattle, AAAI Press, Menlo Park, CA.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543. doi:10.1287/mksc.1120.0713

Ning, Y. (2017). *Big Data in Cybercrime Governance and the Legislation*. Marietta: American Scholars Press.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., . . . Musen, M. A. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37, W170-W173. doi:10.1093/nar/gkp440

OECD. (2020). Country policy tracker. Retrieved from https://www.oecd.org/coronavirus/country -policy-tracker/

Ou, S. Q., He, X., Ji, W. Q., Chen, W., Sui, L., Gan, Y., . . . Bouchard, J. (2020). Machine learning model to project the impact of COVID-19 on US motor gasoline demand. *Nature Energy*, 11. doi:10.1038/ s41560-020-0662-1

Pasquale, F. (2019). A rule of persons, not machines: The limits of legal automation. *George Washington Law Review*, 87(1), 1–55. Retrieved from ://WOS:000456618700001

Pielke, R. A. (2004). What future for the policy sciences? *Policy Sciences*, 37(3/4), 209–225. Retrieved from www.jstor.org/stable/4532628

Ping, J. (2018). Opportunities provided by Big Data technology for government management. In R. Green, I. Solovjeva, Y. Zhang, R. Hou, & E. McAnally (Eds), *Proceedings of the 3rd International Conference on Judicial, Administrative and Humanitarian Problems of State Structures and Economic Subjects* (Vol. 252, pp. 552–555). Paris: Atlantis Press.

Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the US real house price index. *Economic Modelling*, 45, 259–267. doi:10.1016/j.econmod.2014.10.050

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14–46. doi:https://doi.org/10.1016/j.knosys.2015.06.015

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. *Journal of Statistical Software*, 10(2), 1–40.

Roderick, L. (2014). Discipline and power in the digital age: The case of the US consumer data broker industry. *Critical Sociology*, 40(5), 729–746. doi:10.1177/0896920513501350

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. Paper presented at the *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Samuel, J., Ali, G., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information (Japan)*, 11(6), 22. doi:10.3390/info11060314

Sanchez-Oro, J., Duarte, A., & Salcedo-Sanz, S. (2016). Robust total energy demand estimation with a hybrid variable neighborhood search – extreme learning machine algorithm. *Energy Conversion and Management*, 123, 445–452. doi:10.1016/j.enconman.2016.06.050

Sear, R. F., Velasquez, N., Leahy, R., Restrepo, N. J., El Oud, S., Gabriel, N., . . . Johnson, N. F. (2020). Quantifying COVID-19 content in the online health opinion war using machine learning. *IEEE Access*, 8, 91886-91893. doi:10.1109/access.2020.2993967

Smith, A., Sparks, L., & Goulding, J. (2015). Using commercial Big Data to inform social policy: Possibilities, ethics, methods and obstacles. *Journal of Macromarketing*, 35(1), 125–150. Retrieved from ://WOS:000349623600080

Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *Annals of the American Academy of Political and Social Science*, 659(1), 108–121. doi:10.1177/0002716215569217

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3), 239–251. doi:10.1093/bib/6.3.239

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5), 396–413. doi:10.1002/asi.10389

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307. doi:10.1162/COLI_a_00049

Tseng, Y. H., Lin, C. J., & Lin, Y. I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5), 1216–1247. doi:10.1016/j.ipm.2006.11.011

Vaid, S., McAdie, A., Kremer, R., Khanduja, V., & Bhandari, M. (2020). Risk of a second wave of Covid-19 infections: Using artificial intelligence to investigate stringency of physical distancing policies in North America. *International Orthopaedics*, 44(8), 1581–1589. doi:10.1007/s00264-020-04653-3

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., & Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5), 535–542. doi:10.1038/sj.ejhg.5201585

Wang, C. J., Ng, C. Y., & Brook, R. H. (2020). Response to COVID-19 in Taiwan: Big Data analytics, new technology, and proactive testing. *JAMA*, 323(14), 1341–1342. doi:10.1001/jama.2020.3151

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. Paper presented at the *Seventh IEEE International Conference on Data Mining* (ICDM 2007).

Wang, X. D., Zhang, Y., Zhang, X. J., & Xue, L. (2017). Big Data application in commercial economy management. In X. Lin, B. Li, & J. Lamba (Eds), *Proceedings of the 2017 International Conference on Education Science and Economic Management* (Vol. 106, pp. 133–136). Paris: Atlantis Press.

Woolley, J. P., McGowan, M. L., Teare, H. J. A., Coathup, V., Fishman, J. R., Settersten, R. A., . . . Juengst, E. T. (2016). Citizen science or scientific citizenship? Disentangling the uses of public engagement rhetoric in national research initiatives. *BMC Medical Ethics*, 17, 17. doi:10.1186/s12910-016-0117-1

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 44, 120–130. doi:10.1016/j.ijhm.2014.10.013

Yang, Y. B., & Chen, M. (2015). *Analysis of Integration Quality of Informatization and Industrialization in Big Data Era*. Beijing: Science Press Beijing.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3), 2758–2765. doi:https://doi.org/10.1016/j.eswa.2010.08.066

Zheng, W., & Zhang, L. M. (2016). Research on the effects of government audit with Big Data – take China social security audit as example. In G. Lee (Ed.), *2016 3rd International Conference on Psychology, Management and Social Science* (Vol. 89, pp. 99–102). Newark: Information Engineering Research Inst, USA.

Zou, H. C., Shu, Y. L., & Feng, T. J. (2020). How Shenzhen, China avoided widespread community transmission: A potential model for successful prevention and control of COVID-19. *Infectious Diseases of Poverty*, 9(1), 4. doi:10.1186/s40249-020-00714-2