Holding out the promise of Lasswell's dream

Big data analytics in public policy research and teaching

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

DOI
10.1111/ropr.12448

Publication date
2021

Document Version
Final published version

Published in
Review of Policy Research

Citation (APA)
El-Taliawi, O. G., Goyal, N., & Howlett, M. (2021). Holding out the promise of Lasswell's dream: Big data analytics in public policy research and teaching. *Review of Policy Research*, *38*(6), 640-660. https://doi.org/10.1111/ropr.12448

Received: 8 March 2020 | Revised: 14 June 2021 | Accepted: 22 July 2021

DOI: 10.1111/ropr.12448

ORIGINAL ARTICLE

RPR REVIEW OF POLICY RESEARCH

Holding out the promise of Lasswell's dream: Big data analytics in public policy research and teaching

Ola G. El-Taliawi[2] | **Nihit Goyal**[1] | **Michael Howlett**[3]

[1]Faculty of Technology, Policy and Management, Delft University of Technology, Delft, Netherlands

[2]Department of Political Science, Faculty of Public Affairs, Carleton University, Ottawa, ON, Canada

[3]Department of Political Science, Simon Fraser University, Burnaby, BC, Canada

Correspondence
Nihit Goyal, Faculty of Technology, Policy and Management Delft University of Technology (TU Delft), Jaffalaan 5, 2628 BX Delft, Netherlands.
Email: nihit.goyal@tudelft.nl

Abstract

While the emergence of big data raises concerns regarding governance and public policy, it also creates opportunities for diversifying the toolkit for analysis for the policy sciences as a whole, i.e., research concerning policy analysis as well as policy studies. Further, it opens avenues for practice, which together with research requires adaptation in teaching curricula if policy education were to remain relevant. However, it is not clear to what extent this opportunity is being realized in public policy research and teaching. In this study, we examine the prevalence of big data analytics in public policy research and pedagogy using bibliometric analysis and topic modeling for the former, and content analysis of course titles and descriptions for the latter. We find that despite significant scope for application of various big data techniques, the use of these analytic techniques in public policy has been largely limited to select institutions in a few countries. Further, data science has received limited attention in policy pedagogy, once again with significant geographic variation in its prevalence. We conclude that, to stay relevant, the policy sciences need to pay more attention to the integration of big data techniques in policy research, pedagogy, and thereby practice.

KEYWORDS

bibliometric review, big data analytics, machine learning, pedagogy, policy sciences, public policy, topic modeling

INTRODUCTION

The research on public policy straddles several disciplines and numerous policy areas. Arguably, from the policy sciences perspective, emphasizing a focus on knowledge of policy as well as knowledge in policy encompasses the research in public policy (Lasswell, 1970, 1971), regardless of its disciplinary orientation or issue focus. The policy science perspective, however, is at a key stage in its evolution and this is an opportune time to take stock of its past and (re-)imagine its future.

The future of policy sciences is intertwined with several other dynamics that are unfolding. These include the increasing complexity of policy making (Cairney, 2012; El-Taliawi & Hartley, 2020), evolution of the state-society relationship and the shift from government to governance (Hysing, 2009), and a slowly but steadily increasing interest in public policy research and teaching in the Global South (El-Taliawi et al., 2021). One such change in the environment that will influence the future of the policy sciences is the rise of big data.

Big data 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), indicating that it is not only about size and computation power but also the belief in its ability to offer a "higher form intelligence... and the aura of truth, objectivity, and accuracy" (Boyd & Crawford, 2012, p. 663). The key features of big data are volume, velocity, variety, exhaustiveness, high granularity, relationality, i.e., its ability to be combined with other big or small data, flexibility, and scalability (Kitchin, 2013). The sources of such big data can be: (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.

While big data raises several concerns surrounding control, ethics, power, and privacy (Acquisti et al., 2015), it also creates possibilities that otherwise would not exist for both research and practice (Bates et al., 2014; D'Orazio, 2017). As Mayer-Schönberger and Cukier (2013, p. 10) note: "big data allows users to do things at a large scale that cannot be done at a smaller one... by changing the amount we can change the essence." The authors highlight three shifts associated with big data analytics: (i) the possibility of analyzing the entire population (rather than a sample); (ii) the ability to accept more measurement error due to reduction in sampling error; (iii) the change in focus from causality to prediction.

These shifts have important implications for data analysis in public policy research. One such implication is the ability to combine existing administrative data with more granular or dynamic data for decision-making. Another implication is the use of computational analysis for unconventional, unstructured data. Illustratively, text mining (Feldman & Sanger, 2006) and natural language processing (Allahyari et al., 2017), which extract useful information from textual data, have been employed in various disciplines, such as business management (Netzer et al., 2012; Xiang et al., 2015) and medicine (Aronson, 2001; Cohen & Hersh, 2005; Kim et al., 2003; Noy et al., 2009; Spasic et al., 2005; Srinivasan, 2004). A third implication is the growth of machine learning for data analysis. In contrast to traditional statistical analysis, which assumes that a stochastic model underlies the data, machine learning techniques use algorithmic modeling to

examine the data without assuming its underlying distribution as given (Breiman, 2001). These techniques have been implemented in numerous areas, including education, health care, marketing, and policing (Jordan & Mitchell, 2015).

However, the extent to which such new data sources or analytical techniques have been adopted in public policy research is not clear. Is big data analytics being used extensively in public policy research? If so, where and by whom? Further, is it facilitating new types of analyses or research that combine knowledge *of* policy with knowledge *in* policy to uncover hidden relationships that other techniques have been less apt to detect? The first objective of this study is to examine the prevalence of big data analytics in public policy research and shed light on its content. To undertake this, we conducted a bibliometric analysis and a topic modeling analysis of research at the intersection of big data and public policy.

Moreover, as big data have significant implications for public policy research, they are increasingly being used by governments in harnessing the power of information to improve policies. At present big data analytics is used by several U.S. federal agencies including the Social Security Administration, the Food and Drug Administration, and others (IBM Center for Business of Government, 2015). Therefore, curriculum aimed at providing future policy analysts and researchers with a relevant toolkit that holds out the promise of Laswell's dream requires a reflection of this uptake in the use of big data analysis in teaching. Big data are thus an opportunity and a threat for policy education, where failure to up-skill future graduates would lead to divergence in teaching on the one hand, and practice and scientific research on the other (Donovan, 2008).

However, despite the relevance of teaching to research and practice and the implications of equipping policy scholars with the necessary tools to harness the potential of big data, it is yet unclear the extent to which policy education has incorporated big data into its curricula. Therefore, the second objective of this study is to examine the prevalence of big data analytics in policy pedagogy around the world. To undertake this, we analyzed big data course offerings by policy programs worldwide.

The remainder of this article is structured as follows. In Section "Research Methods", we describe the methods for this study. Section "An Overview of Research on Big Data Analytics in Public Policy" provides an overview of the bibliometric analysis. Subsequently, in Section "Themes in Big Data Research", we examine the key themes in research on big data analytics in public policy using topic modelling and differentiate the discussion on big data from their use specifically for public policy research. Section "Pedagogical Gaps" presents the findings from the assessment of policy curricula in programs around the world. Finally, we conclude the article in Section "Conclusion".

RESEARCH METHODS

We combined bibliometric analysis, topic modeling, and content analysis for this study. To gauge the prevalence of publications mentioning terms related to big data analytics in their title, abstract, or keywords in the literature on public policy, we first estimated the approximate volume of the literature on governance or public policy. For this, we conducted a search for (*governance* OR *policy*) in the title, abstract, or keywords of publications in the Social Sciences Citation Index (SSCI), the Conference Proceedings Citation Index–Social Sciences and Humanities (CPCI-SSH), and the Book Citation Index–Social Sciences and Humanities (BKCI-SSH) of the Web of Science database. We also excluded items published in the year 2021 from this search for a more accurate picture of the annual trend. The search was conducted on April 20, 2021, and returned 748,279

results. The annual trend in the number of publications was documented based yearly breakdown provided by the Web of Science.

To obtain publications pertaining to big data analytics (in public policy research), we used keywords spanning the concept itself ("big data"), epistemology ("data science"), analytic approach(es) (e.g., "machine learning" or "natural language processing"), new data sources (e.g., "text as data"), and modelling techniques (e.g., "artificial neural networks" or "random forests"). The following search string was developed iteratively based on discussion among the authors, the results of the search, and feedback from the reviewers: ("artificial neural network*" OR "automated content analysis" OR "automatic content analysis" OR "automated text analysis" OR "automatic text analysis" OR "big data" OR "computational text analysis" OR "data science" OR "deep learning" OR "machine learning" OR "natural language processing" OR "opinion mining" OR "random forest*" OR "sentiment analysis" OR "supervised learning" OR "support vector machine*" OR "text analytics" OR "text as data" OR "text mining" OR "topic model*" OR "unsupervised learning") AND (*governance* OR *policy*).

This search query was conducted on the title, abstract, or keywords of publications in the SSCI, the CPCI-SSH, and the BKCI-SSH of the Web of Science database. Items published in the year 2021 were excluded from the search for a more accurate picture of the annual trend. The search was conducted on April 20, 2021 and returned 3680 results. This dataset of 3680 publications was downloaded in 'bibtex' and 'plain text' formats for further processing. A limitation of our approach, however, is that publications in our sample might merely mention the search terms without necessarily focusing on the intersection of big data analytics and public policy.

The bibliometric analysis was conducted using the *bibliometrix* package in the R programming language (Aria & Cuccurullo, 2017). Bibliometrix is an open-source library that provides various functions for importing bibliometric data, mapping scientific activity, and examining the relationships among different publications. For this study, we focused specifically on four characteristics of research pertaining to big data and public policy: (i) country-wise scientific production and inter-country collaboration; (ii) universities involved and inter-university collaboration; (iii) types of sources featuring the publications; and (iv) keywords used by authors to describe the study. As countries or universities involved in a study are not captured separately in a bibliometric dataset, the *bibliometrix* package infers these based on authors' affiliations. Therefore, the results of the analysis are likely to be indicative rather than exact.

We used topic modeling to explore the main themes in the dataset. Topic modeling is a text mining technique for examining large document corpora. It is based on the premise that each document in a document corpus comprises one or more *latent* topics, which are in turn composed of terms, i.e., words or phrases (Blei, 2012). Topic modeling algorithms use statistical modeling to 'discover' these hidden topics in the document corpus. While numerous topic modeling algorithms have been developed to account for different characteristics of the document corpus (Blei & Lafferty, 2007; Rosen-Zvi et al., 2004; Wang et al., 2007), we use structural topic modeling as it can use document metadata for discovering topics, allow for correlation among topics, and be easily implemented in the R programming language using the *stm* package (Roberts et al., 2014).

The dataset was preprocessed through several steps before executing the topic model. First, we corrected typographical errors and harmonized spellings using the *hunspell* package (Ooms, 2018). Second, we annotated the text in the publication titles and abstracts using the *udpipe* package (Wijffels, 2020). This included parts of speech tagging and lemmatization of the input. Third, we identified 'n-grams' in the text through a combination of rapid automatic keyword extraction (Wijffels, 2020), noun phrase detection, and manual parsing; where applicable, we replaced sequences of words in the input with (lemmatized) phrases. Fourth, we identified and removed

'stop words' (that add little insight for our analysis) in the input based on parts of speech, the *stopwords* package (Benoit et al., 2020), and a manual scan of terms that occurred more than 10 times in the dataset. Fifth, iteratively with step four, we created a dictionary to harmonize terms in the input (for example, abbreviations and their expanded form or plural and singular form). Finally, we stemmed terms to reduce them to their root form, using the *SnowballC* package (Bouchet-Valat, 2020). Subsequently, we selected the number of topics—an input to the topic modeling algorithm—as 8, based on the log likelihood and semantic coherence of models with 5 to 50 topics.

To explore the preponderance of public policy programs offering big data analysis in their curriculum, we examined 75 programs to explore the trend of big data courses offered in policy programs worldwide. This sample was selected based on the results of the bibliometric analysis. We included in our sample the top 75 institutions with the highest publishing record on the topic. We further excluded 16 results for not offering public policy degrees, specializations, or concentrations. Websites of the remaining 59 programs were examined to gauge their big data course offerings.

We coded a policy program offering big data pedagogy if it offers one or more, core or elective, modules that include the following keywords, either in the title or the course description: "analytics", "artificial neural network," "automated content analysis," "automatic content analysis," "automated text analysis," "automatic text analysis," "big data," "computational text analysis," "data science," "deep learning," "informatics," "machine learning," "natural language processing," "opinion mining," "random forest," "sentiment analysis," "supervised learning," "support vector machine," "text analytics," "text as data," "text mining," "topic model," "unsupervised learning."

The limitation of this method lies in its non-generalizability to the universe of policy programs worldwide, since our sample was not drawn from a comprehensive sampling frame of policy programs. Therefore, our findings are limited to the programs included in our sample. Further testing may be done to test whether our findings would hold in a larger random sample. Additionally, while some of the big data analysis skills may be taught in computer science classes, the focus of this study is on analyzing the content of policy programs. Hence, we did not include in our analysis big data courses that may be made available to public policy students through computer science classes, which may be the subject of further investigation.

AN OVERVIEW OF RESEARCH ON BIG DATA ANALYTICS IN PUBLIC POLICY

The tradition of public policy research in the social sciences goes back to over a century and has steadily gained momentum during this period. This is indicated by the increase in the number of publications in the Web of Science mentioning the term *governance* or *policy* in their titles, abstracts, or keywords during this time (Figure 1). The number of publications per year on the topic increased from approximately 100 in the 1930s to about 1000 in the 1960s, crossing 10,000 in the 2000s, and exceeding 50,000 since 2018. Within these, the number of publications mentioning terms relevant to big data analytics has grown exponentially as well. The first such publication appeared in this dataset in 1992. However, the subsequent years witnessed relatively little activity in this area and it was only in 2007 that the number of publications in the year went beyond 10. Since then, the field has developed rapidly and the number of publications per year was over 100 by 2014 and nearly 1000 by 2020. Yet, publications pertaining to big data analytics constituted less than 2% of the total volume of research on public policy in 2020.

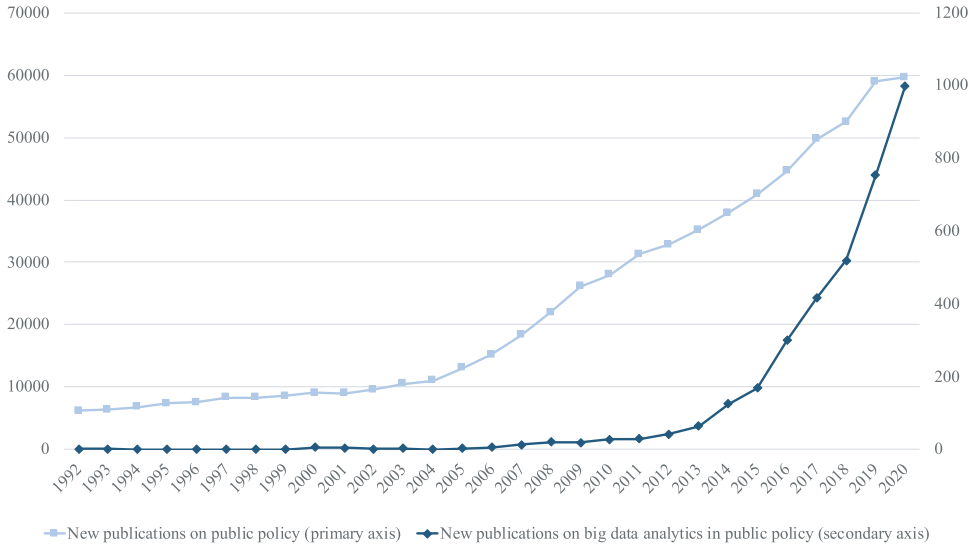

FIGURE 1 The volume of scientific production on big data analytics in public policy research

Further, most of the scientific production in this field is limited to a few countries (Figure 2). Based on the affiliation of the corresponding author, the United States has been involved in over 850 publications, China and United Kingdom have been involved in over 650 and approximately 360 publications, respectively. No other country has been involved in more than 150 publications on the topic. Among the rest, Australia, Republic of Korea, Germany, Italy, the Netherlands, Canada, and Spain account for the most scientific production in this field. With the exception of China, OECD countries dominate research on the topic, and India and Brazil are the only other non-OECD country among the top 15. Other non-OECD countries with notable scientific production are Iran, Russia, Singapore, South Africa, and Malaysia.

In addition, most research on the topic is conducted within a single country. While about 80% (approximately 3000) publications in this dataset are co-authored, only about 30% involve multi-country partnership, possibly due to the ethics and legality of sharing big data across geographic boundaries. Among countries with the highest scientific production in this field, Switzerland, France, Germany, Italy, and the United Kingdom have the highest share of publications with multi-country partnerships (Table 1). Although the share of publications with multi-country partnerships is relatively low for the United States and China, they still have the highest number of publications with multi-country partnership in this dataset. In terms of absolute volume, most multi-country partnerships are observed between China and the United States (128 publications), the United Kingdom and the United States (101 publications), Canada and the United States (55 publications), Australia and China (42 publications), and Germany and the United Kingdom (41 publications).

Although authors who have published on big data analytics in public policy represent over 3500 institutions worldwide, out of these approximately 1500 institutions occur just once in the dataset and only about 300 institutions occur 10 or more times in this dataset (multiple occurrences by an author and occurrences by multiple authors from an institution in the same publication are counted as distinct). As one might expect, institutions in the United States, China, and United Kingdom feature prominently in this dataset. In fact, the institutions with the most occurrences in the dataset—Stanford University, the University of Oxford, the University of Pennsylvania,

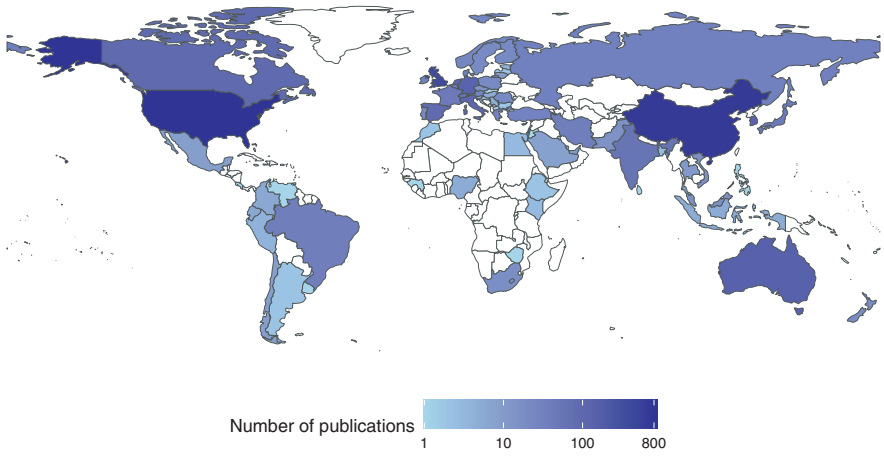

FIGURE 2 Country-wise scientific production based on institutional affiliation of the corresponding author

TABLE 1 Country-wise volume of scientific production based on single country partnership (SCP) and multi-country partnership (MCP) for countries with the highest scientific production on big data analytics and public policy

Country	SCP	MCP	Share of MCP
United States of America	702	168	19%
China	490	170	26%
United Kingdom	223	141	39%
Australia	95	54	36%
Republic of Korea	111	25	18%
Germany	76	57	43%
Italy	66	47	42%
Netherlands	73	35	32%
Canada	65	29	31%
Spain	62	27	30%
India	47	11	19%
Switzerland	22	36	62%
France	18	29	62%
Japan	33	11	25%
Brazil	32	7	18%

the University of Michigan, Arizona State University, and the University of Cambridge—are all located in one of these countries. Institutions from other countries with significant research on this topic include the Delft University of Technology, the University of Queensland, the National University of Singapore, the University of Toronto, and the University of Melbourne. Outside of China, hardly any institutions in non-OECD countries are very active in publishing on big data analytics in public policy.

A collaboration network among the top 50 institutions is shown in Figure 3. The institutions are grouped into four clusters based on the collaborations (links) between the institutions (nodes), using the Louvain method. The largest cluster—in terms of the number of institutions—comprises universities from various countries, including Australia (e.g., the University of Sydney), Canada (e.g., University of Toronto), China (e.g., Southeast University), Japan (the University of Tokyo), the Netherlands (e.g., Leiden University), and the United States (e.g., Harvard University). Another large cluster also consists of universities in China (e.g., Tsinghua University), the Netherlands (e.g., the Delft University of Technology), Singapore (the National University of Singapore), the United Kingdom (e.g., the University of Cambridge), and the United States (e.g., Arizona State University). The remaining two clusters indicate more regional collaboration, with one centered largely around universities in China and the other around universities in the United States.

Although research on big data analytics in public policy has been presented at or published in over 1500 sources, nearly 1000 sources have only a single publication on the topic and only about 140 sources have five or more publications in this dataset. In fact, the top 15 sources account for over 15% (640 publications) of the existing research (Figure 4). As this figure shows, the journal *Sustainability* has published, by far, the most articles on the topic. Other entries in the top 15 also indicate that topics related to climate change and energy have received much attention in the current research on big data analytics, with the presence of sources such as the *Journal of Cleaner Production*, *Energies*, and *Energy Policy*. Further, the prominence of sources such as *Land Use Policy*, *Remote Sensing*, and *Computers*, *Environment*, *and Urban Systems* is suggestive

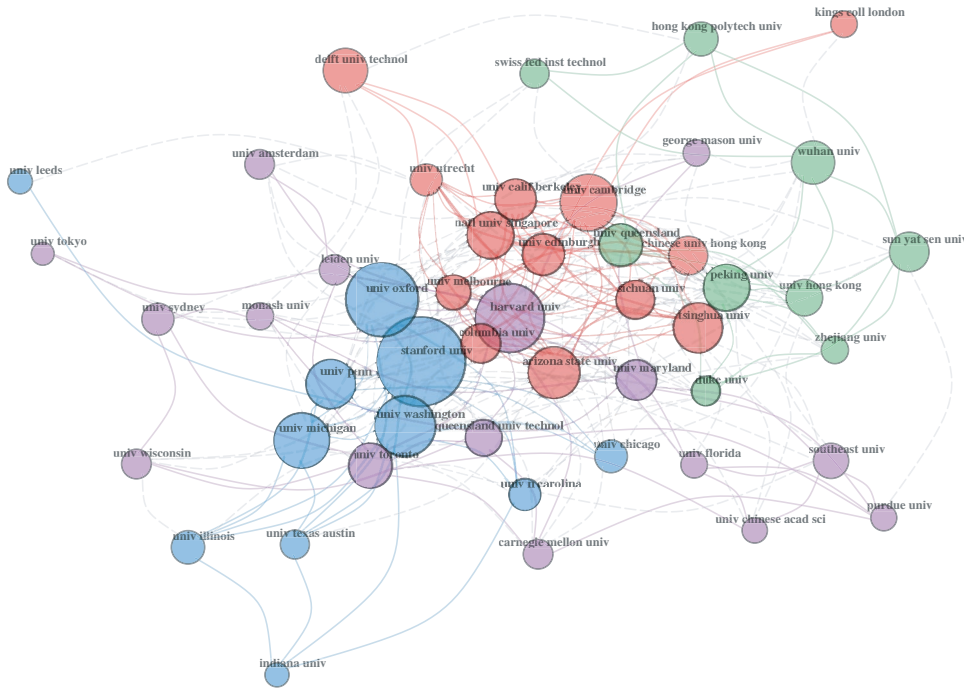

FIGURE 3 Institutional network of scientific research on big data analytics in public policy. A node depicts an institution. A link connecting two nodes indicates a collaboration between authors from the two institutions. The size of the node indicates the degree (number of collaborations) of the node. The color of the node indicates its cluster. The nodes as clustered using a clustering algorithm on the collaboration network

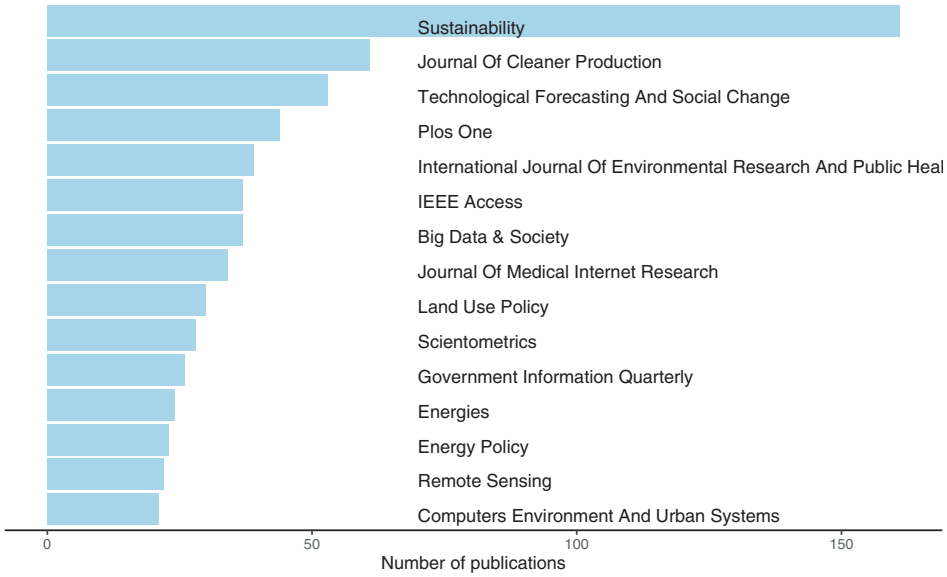

FIGURE 4 Cumulative scientific production on big data analytics in public policy by source

of the use of big data analytics for spatial analysis. The influence of big data on society (e.g., *Big Data & Society*), the impact of information and communication technologies on governments (e.g., *Government Information Quarterly*), the relationship between technological development and social change (e.g., *Technological Forecasting and Social Change*), and public health (e.g., the *International Journal of Environmental Research and Public Health*) appear to be the focus of scientific activity on big data analytics in public policy.

THEMES IN BIG DATA RESEARCH

For a more granular analysis of the key themes in research on big data analytics in public policy and to understand the extent of applied research within this dataset, we clustered publications into eight themes. These themes are depicted—in decreasing order of prevalence—along with the most exclusive and most frequently occurring terms in each theme in Figure 5.

The most prevalent theme in this dataset is Theme 1 on Digital technology. The commonly occurring terms in this theme include 'challenge', 'practice', 'digital', 'social', 'knowledge', 'smart city', 'analytics', 'concept', and 'sustainability' while the terms most exclusive to this theme are 'dashboard', 'lab', 'knowledge management', 'smart urban', 'digital age', and 'IoT'. These indicate that publications in this theme focus primarily on new digital technologies, which generate big data, and their effects on societies. Constantinides et al. (2018), for example, review the literature on digital infrastructure to identify emerging themes, while Nieminen (2016) delves into the social implications of the growth of big data and Clarke and Craft (2017) highlight the possibility of marginalization of citizens in digital era policy design due to the layering of instruments such as big data, crowdsourcing, open data, and robotics on existing policy tools. Further, several scholars study the big data phenomenon specifically in the urban context, for instance, by examining the impact of data-driven urbanism on sustainability (Bibri, 2019), applying a systems thinking

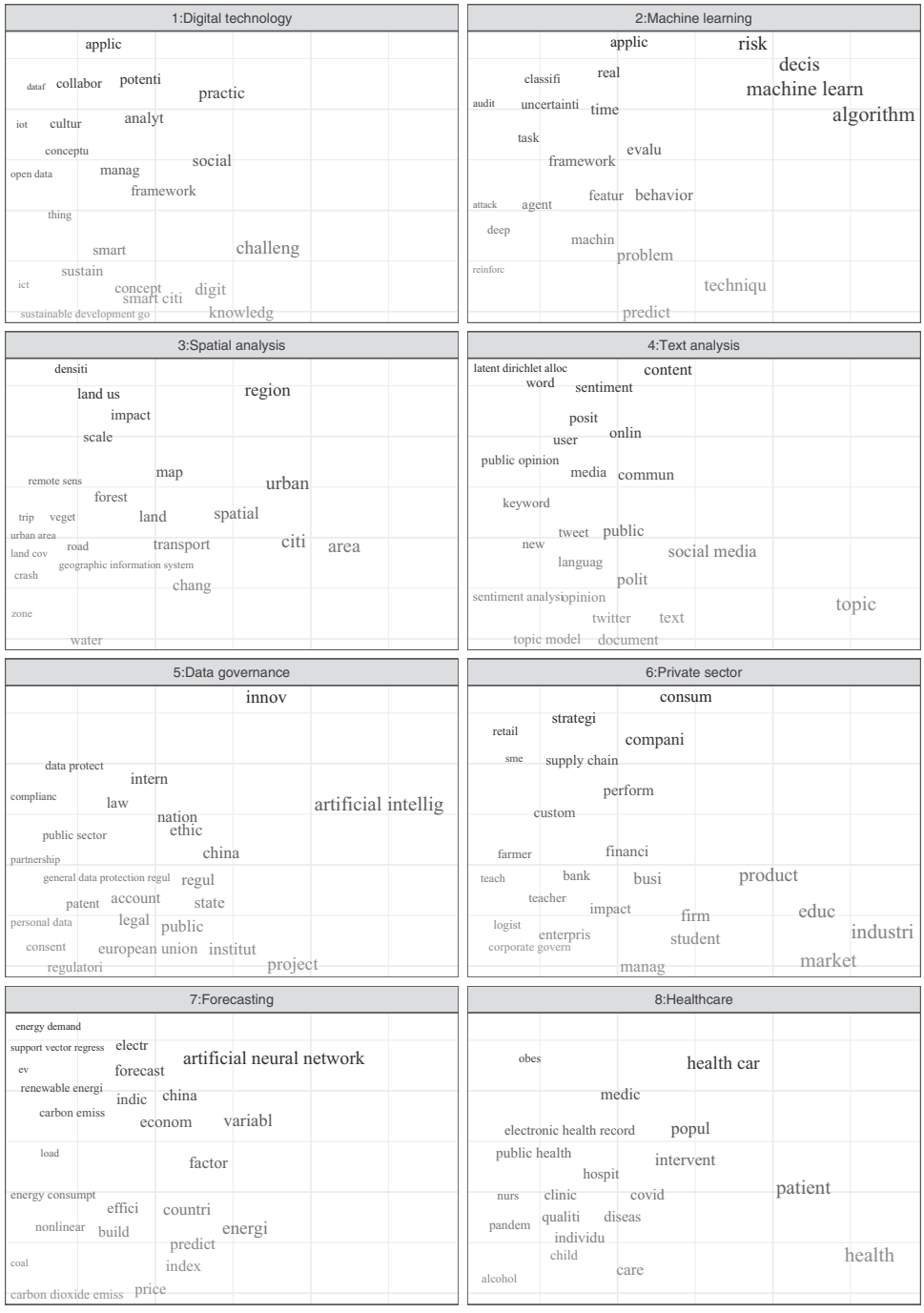

FIGURE 5 Themes in research on big data analytics in public policy. Each sub-graph shows terms associated with a theme. The position of a term on the *x*-axis (as well as its size) is indicative of the probability of occurrence of the term within that theme while its position on the *y*-axis (as well as its color intensity) is indicative of the exclusivity of occurrence of the term to that theme. The themes are numbered in descending order of prevalence in the dataset

perspective on the effect of smart technologies on urban governance (Caputo et al., 2019), developing a framework for smart city design by incorporating the influence of digital technologies (Hämäläinen, 2020), analyzing the instrumental use of the concept of smart city by actors (Desdemoustier et al., 2019), and unpacking the paradoxes in the concept of smart urbanism through a study of the smart nation initiative in Singapore (Kong & Woods, 2018).

Closely related to this theme, Theme 5 on Data governance explores issues surrounding the governance of big data and the regulation of artificial intelligence. This can be observed in the most commonly terms—'artificial intelligence', 'innovation', 'institution', 'China', 'public', 'ethics', 'state', and 'nation'—and the most exclusive terms—'data protection', 'law', 'data subject', 'intellectual property', 'Canadian', 'patent', and 'European Commission'—in this theme. Illustratively, Zharova and Elin (2017) discuss big data in the context of personal data privacy and Arkhipov and Naumov (2016) examine the legal definition of personal data in the big data era, both in the case of Russia. Numerous publications focus on the effect of data protection legislation, for instance, on open data policy in the case of mapping data (van Loenen et al., 2016), governance of digital health (Marelli et al., 2020), medical data processing (Mostert et al., 2016), and biobanking (Townsend & Thaldar, 2019). Meanwhile, Marley (2019) analyzes the issue of data sovereignty in policies of the Institutional Review Board with the growth of big data. In a different vein, Sanga (2019) examines over a million filings with the Securities and Exchange Commission to create a database of 800,000 contracts using machine learning.

Theme 2 on Machine learning is the second most prevalent theme in this dataset. Among the most commonly occurring terms in this theme are 'algorithm', 'machine learning', 'decision', 'risk', 'technique', 'prediction', 'behavior', and 'evaluation'. Also, the most exclusive terms in this theme include 'library', 'deep reinforcement (learning)', 'markov decision process', 'reinforcement (learning)', 'optimal policy', 'deep neural network', and 'agent'. Publications in this theme typically focus on the use of machine learning techniques for data analysis. This is depicted, for instance, in the case of transportation by Gerum et al. (2019) in their predictive analysis of defects in railways using a discounted Markov decision process, Ying et al. (2020) in their use of deep reinforcement learning for metro train scheduling, and Rasouli and Timmermans (2014) in their prediction of modal choice in the Netherlands using a model ensemble with decision trees. Illustrative applications in other policy areas include the use of reinforcement learning for evaluation of dialogue strategies (Rieser & Lemon, 2011), implementation of A-Deep Q-Learning clustering algorithm for transmission line tower fault prediction (Jung & Huh, 2019), and integration of machine learning with quasi-random assignment to analyze its potential to enhance judicial decision-making (Kleinberg et al., 2018). Some publications in this theme also go beyond the direct application of machine learning. Illustratively, Kwakkel and Cunningham (2016) draw inspiration from the random forest modeling to enhance scenario discovery for decision-making under uncertainty, while Berk and Bleich (2013) examine the effect of decision boundaries on forecasting accuracy of machine learning techniques in the case of criminal justice, and Janssen et al. (2020) advocate caution against the indiscriminate use of machine learning and propose the adoption of 'explainable' artificial intelligence and 'hybrid intelligence' to decrease the chance of error.

While publications in the theme on machine learning emphasize prediction, publications in Theme 7 focus more on forecasting, especially in the context of energy and environment. This can be observed in the most exclusive terms—'artificial neural network', 'carbon dioxide emission', 'energy use', 'fossil fuel', 'ARIMA' (Autoregressive Integrated Moving Average), 'autoregression', 'forecast', 'energy sector', and 'sustainable energy'—as well as other commonly occurring terms in this theme, such as 'energy', 'variable', 'country', 'factor', 'economy', 'prediction', 'China',

and 'index'. Illustratively, Liu and Wu (2017) study the heterogeneous effect of emissions tax on carbon dioxide, methane, and nitrous oxide in the case of China using big data from a national census. Numerous scholars use machine learning models such as artificial neural networks or support vector machines, for example, to analyze the factors that influence carbon dioxide emissions (Li et al., 2017), forecast wind generation and facilitate renewable energy integration (Nazir et al., 2020), predict policy effect on primary energy production and consumption in the future (Sözen & Arcaklioğlu, 2011), and forecast energy-related carbon dioxide emissions (Wen & Cao, 2020; Zhao et al., 2017). Other scholars have also compared the performance of artificial neural networks with more 'traditional' techniques such as ARIMA and linear regression (Adeyinka & Muhajarine, 2020; Bilgili et al., 2012).

Theme 3 focuses predominantly on spatial analysis. Consequently, the most commonly occurring terms in this theme are 'area', 'city', 'urban', 'region', 'spatial', 'change', 'transport', 'land', and 'map'. Meanwhile, the most exclusive terms in this theme are: 'land', 'land cover change', 'remote sensing', 'urban expansion', 'vegetation', 'zone', 'coastal', 'convolutional neural network', and 'density'. In comparison with the themes above, the emphasis here is less on the techniques of analysis and more on the type of data used, such as satellite data, radar information, and LiDAR data. Conrad et al. (2017), for example, use satellite data to spatially map crop diversity in Uzbekistan and Akodéwou et al. (2020) analyze the intensity and trajectory of historical land use and land cover change in Togo. Some scholars also use spatial data to pursue more explanatory analysis, such as examining the effect of land use and urbanization on land surface temperature (Athukorala & Murayama, 2020; Ranagalage et al., 2019), estimating the impact of residential property (re-)development on canopy cover change in New Zealand (Guo et al., 2019), investigating the relationship between climatic variables and landscape characteristics on deforestation in Peru (Bax & Francesconi, 2018). In a different vein, De Alban use a random forest classifier to combine radar and satellite data to enhance land use classification and evaluate land cover change in Myanmar (De Alban et al., 2018). A few studies have also used unconventional data, for instance geolocation from social media, to perform spatially explicit analysis (Chun et al., 2020; Moyano et al., 2018).

Theme 4 on Text analysis differs from the above themes in its use of unstructured text as key data for research. This can be observed in the most commonly occurring terms in this theme: 'topic', 'social media', 'text', 'politics', 'public', 'content', 'communication', 'online', and 'document'. The terms most exclusive to this theme—'lexicon', 'dictionary', 'hashtag', 'Latent Dirichlet Allocation', 'sentiment', 'topic model', 'Facebook', and 'keyword'—also demonstrate this focus. Numerous publications apply these techniques to examine the politics of policymaking. Greene and Cross (2017), for example, trace the evolution in the agenda of the European Parliament by developing a dynamic topic model of speeches from Members of European Parliament during 1999–2014. In another application, Rabini et al. (2020) develop a coding scheme for leadership trait analysis in the German language to examine the influence of leadership on German foreign policy using computational text analysis. Text analysis has also been used to study issue framing—such as in the case of biofuel policy in Brazil (Talamini & Dewes, 2012)—as well as its influence on public attitude, for example, in the case of austere fiscal policy in Britain (Barnes & Hicks, 2018). In addition, scholars have employed text analysis techniques to investigate the interplay between the public and the media, illustratively, in response to mass shootings (Croitoru et al., 2020) or the presidential election (Alashri et al., 2016) in the United States. Bibliometrics and scientometrics represent other active domains for the application of computational text analysis (Lyu & Costas, 2020).

The remaining two themes, Theme 6 on Private Sector and Theme 8 on Healthcare, predominantly feature publications that use big data analytics for policy-relevant—but not necessarily

public policy—research. The most exclusive terms in the theme on Private sector include 'e-commerce', 'retail', 'firm', 'corporate governance', 'supply chain', and logistics. Other commonly occurring terms in this theme are 'industry', 'market', 'education', 'product', 'student', and 'business'. Publications within this theme show an interest in big data from diverse perspectives within this context, such as strategy (Huo & Hung, 2015), financing (Zhang & Destech Publicat, 2017), marketing (Liu & Yi, 2017), supply chain management (Gawankar et al., 2020; Lai et al., 2018), and sales (Du et al., 2018). Meanwhile, the most exclusive terms in the theme on Healthcare include 'health care', 'clinic', 'comorbidity', 'diagnosis', 'disease', 'child', and 'COVID'. Other commonly occurring terms in this theme are: 'health', 'patient', 'population', and 'intervention'. Scholars in this area show an interest in big data analytics broadly to study its strengths and weaknesses (Howie et al., 2014), to examine a large volume of information (Aznar-Lou et al., 2018; Zenk et al., 2018), to analyze unstructured data such as text or video (Bayen et al., 2017; Hernandez-Boussard et al., 2020), to combine multiple types of data, such as geographic information and medical records (Pandey et al., 2020), and to apply machine learning for impact assessment (Rikin et al., 2015).

The above analysis suggests that application of big data analytics in public policy research is even smaller than indicated by the size of this dataset. Only four of the eight themes discovered here unequivocally demonstrate the use of such techniques for public policy research: theme 2 applies machine learning to shed light on policy problems and solutions, theme 3 uses various methods to analyze (big) spatial data, theme 4 employs a range of text mining techniques using unconventional data, and theme 7 concerns the use of machine learning models for forecasting and time-series analysis. While themes 6 and 8 also involve the application of big data analytics—in the private sector and in healthcare, respectively—their focus is not on public policy per se. Meanwhile, theme 1 examines the effect of digital technologies that generate big data on societies and theme 5 focuses on issues surrounding big data governance—such as ethics and privacy—and their implications for public policy.

PEDAGOGICAL GAPS

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, the use of big data in policy analysis and policy education is trailing behind. We surveyed 59 university websites, identifying public policy and public affairs programs, courses, and specializations in which big data analytics was a focus. Among the results, we found that only 30% of programs included in our sample offer such courses, while 69% did not offer big data courses in their degree courses. The remaining 1% did not provide sufficient information on their websites to give a definitive conclusion (Figure 6).

When we analyzed the geographic distribution of programs that did offer big data courses, we found predominance in the provision of such courses in North America (50%), followed by Europe (18%) and Australia (18%), with the remaining 14% in Asia (Figure 7). It would be worth exploring systematically what factors lie behind such variation.

Further, four types of big data offerings were found in our sample: courses, degrees, certificates, and concentrations, with the majority of programs offering courses only. Course titles include 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, certificates and concentrations in policy analysis and big data (12%). Examples of degrees offered include *Master of Science*

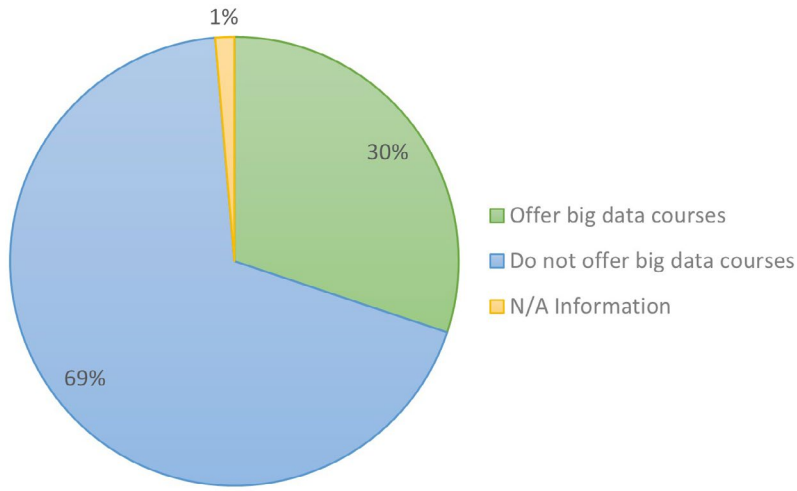

FIGURE 6 Percentage of policy programs offering big data courses (%)

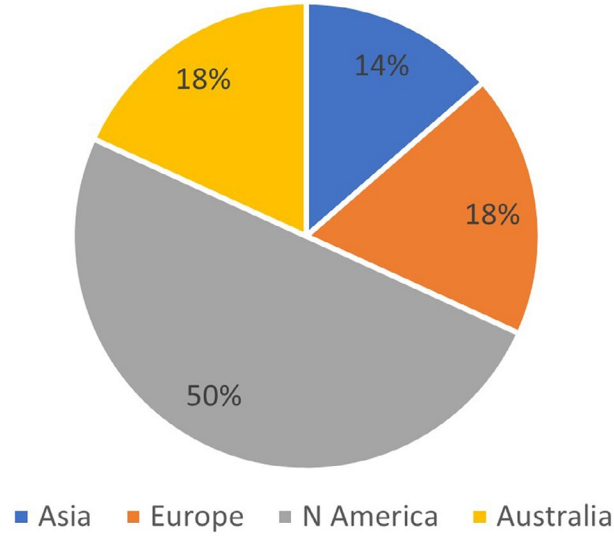

FIGURE 7 Public policy programs offering big data courses by continent (%)

in Policy Analytics and *Big Data for Public and Global Governance*. Additionally, examples of concentrations and certificates include *Concentration in Data Analytics and Program Evaluation* and a *Certificate in Data Science and Policy Analysis*.

In terms of course content, we were unable to examine the pedagogical techniques used, since course syllabi are not accessible on university websites. A content analysis of syllabi would reveal the predominant techniques used by instructors, and the kind of pedagogical gaps that exist. The content of qualitative and quantitative research methods modules should also be analyzed in order to determine whether they offer big data techniques embedded in their course learning objectives. Further, offering core modules, as opposed to elective courses, in the use of big data analytics in policy analysis would reveal that such techniques are increasingly being considered

by policy programs as essential components of the skills needed to be imparted to students. The rate of student uptake of these courses may also be assessed by means of surveying.

CONCLUSION

While the emergence of big data has raised new concerns for governance and policymaking, it has also created opportunities for public policy research 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. However, the extent to which such techniques have been utilized in public policy research or taught in public policy programs around the world is unclear. 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 indicate that big data have the potential to create a paradigm shift public policy research and practice with the development of new epistemologies and new types of analyses that fulfil Laswell's dream of not only creating new knowledge but also mobilizing it to inform policymaking. While research on big data and governance or policy was 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, it is not clear whether it is adequate for disrupting the field or just reinforcing status quo. For instance, 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, with the exception of 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 still attracted some attention, policy studies are yet to make use of this opportunity.

The dominance of governance and policy analysis is corroborated by a discovery of the key topics in the dataset based on a topic modeling analysis of publication titles, abstracts, and keywords. Much of the existing research has focused on issues surrounding the governance of big data analytics and techniques, such as ethics, privacy, and surveillance. While this is definitely required and not a problem in itself, it also indicates that the volume of research *applying* these techniques for public policy is even lesser than the size of this dataset might suggest. Further, this preliminary analysis suggests that the use of big data analytics has typically been in policy analysis rather than policy studies.

Further, our analysis of 75 institutions active in public policy research indicated that only about 59 of them offer a public policy program or degree, and only about 30% of these offer courses pertaining to big data analytics. There is high geographic variation in the diffusion of big data analytics in public policy pedagogy as well—while 50% of the programs in North America have courses related to big data, this number is about 18% or less in Europe and the rest of the world. This is a worrying trend that suggests that policy research and practice in the future are also likely to miss out on the promise of big data. Thus, the potential of big data in fundamentally changing public policy not only remains unrealized but also is unlikely to be realized if status quo continues.

Can this gap be addressed and, if so, how? First, more multi-country and multi-institutional collaboration can potentially increase the diffusion of big data analytics in continental Europe and the Global South. Key issues surrounding sharing of data—such as ethics, legality, and privacy—however, will need to be addressed for this to happen. Second, 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, 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 from 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. 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.

CONFLICT OF INTEREST

The authors declare none.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Web of Science Database. Restrictions apply to the availability of these data, which were used under license for this study.

ORCID

Ola G. El-Taliawi https://orcid.org/0000-0002-1615-6021
Michael Howlett https://orcid.org/0000-0003-4689-740X

REFERENCES

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514. https://doi.org/10.1126/science.aaa1465

Adeyinka, D. A., & Muhajarine, N. (2020). Time series prediction of under-five mortality rates for Nigeria: Comparative analysis of artificial neural networks, Holt-Winters exponential smoothing and autoregressive integrated moving average models. *BMC Medical Research Methodology*, *20*(1), 1–11. https://doi.org/10.1186/s12874-020-01159-9

Akodéwou, A., Oszwald, J., Saïdi, S., Gazull, L., Akpavi, S., Akpagana, K., & Gond, V. (2020). Land use and land cover dynamics analysis of the togodo protected area and its surroundings in Southeastern Togo, West Africa. *Sustainability (Switzerland)*, *12*(13), 5439. https://doi.org/10.3390/su12135439

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. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 795–802).

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*.

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Arkhipov, V., & Naumov, V. (2016, December). The legal definition of personal data in the regulatory environment of the Russian Federation: Between formal certainty and technological development. *Computer Law & Security Review*, *32*(6), 868–887. https://doi.org/10.1016/j.clsr.2016.07.009

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap Program. Proceedings. AMIA Symposium, 17–21.

Athukorala, D., & Murayama, Y. (2020). Spatial variation of land use/cover composition and impact on surface urban heat island in a tropical Sub-Saharan city of Accra, Ghana. *Sustainability (Switzerland)*, *12*(19), 7953. https://doi.org/10.3390/su12197953

Aznar-Lou, I., Pottegard, A., Fernandez, A., Penarrubia-Maria, M. T., Serrano-Blanco, A., Sabes-Figuera, R., Gil-Girbau, M., Fajo-Pascual, M., Moreno-Peral, P., & Rubio-Valera, M. (2018, November). Effect of copayment policies on initial medication non-adherence according to income: A population-based study. *BMJ Quality & Safety*, *27*(11), 878–891. https://doi.org/10.1136/bmjqs-2017-007416

Barnes, L., & Hicks, T. (2018, April). Making austerity popular: The media and mass attitudes toward fiscal policy. *American Journal of Political Science*, *62*(2), 340–354. https://doi.org/10.1111/ajps.12346

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. https://doi.org/10.1377/hlthaff.2014.0041

Bax, V., & Francesconi, W. (2018). Environmental predictors of forest change: An analysis of natural predisposition to deforestation in the tropical Andes region, Peru. *Applied Geography*, *91*, 99–110. https://doi.org/10.1016/j.apgeog.2018.01.002

Bayen, E., Jacquemot, J., Netscher, G., Agrawal, P., Noyce, L. T., & Bayen, A. (2017). Reduction in fall rate in dementia managed care through video incident review: Pilot study. *Journal of Medical Internet Research*, *19*(10), e339. https://doi.org/10.2196/jmir.8095

Benoit, K., Muhr, D., & Watanabe, K. (2020). *stopwords: Multilingual Stopword Lists*. In (Version R package version 2.0). https://CRAN.R-project.org/package=stopwords

Berk, R. A., & Bleich, J. (2013, August). Statistical procedures for forecasting criminal behaviour: A comparative assessment. *Criminology & Public Policy*, *12*(3), 513–544. https://doi.org/10.1111/1745-9133.12047

Bibri, S. E. (Ed.). (2019). The underlying technological, scientific, and structural dimensions of data-driven smart sustainable cities and their socio-political shaping factors and issues. In *Big data science and analytics for smart sustainable urbanism*, Advances in Science, Technology & Innovation (IEREK Interdisciplinary Series for Sustainable Development). (pp. 95–129), Cham, Springer. https://doi.org/10.1007/978-3-030-17312-8_5

Bilgili, M., Sahin, B., Yasar, A., & Simsek, E. (2012). Electric energy demands of Turkey in residential and industrial sectors. *Renewable and Sustainable Energy Reviews*, *16*(1), 404–414. https://doi.org/10.1016/j.rser.2011.08.005

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

Blei, D. M., & Lafferty, J. D. (2007, June). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Bouchet-Valat, M. (2020). *SnowballC: Snowball stemmers based on the C 'libstemmer' UTF-8 library*. In (Version R package version 0.7.0). https://CRAN.R-project.org/package=SnowballC

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/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. https://doi.org/10.1214/ss/1009213726

Cairney, P. (2012). Complexity theory in political science and public policy. *Political Studies Review*, *10*(3), 346–358. https://doi.org/10.1111/j.1478-9302.2012.00270.x

Caputo, F., Walletzky, L., & Štepánek, P. (2019). Towards a systems thinking based view for the governance of a smart city's ecosystem. *Kybernetes*, *48*(1), 108–123. https://doi.org/10.1108/K-07-2017-0274

Chun, J., Kim, C.-K., Kim, G. S., Jeong, J., & Lee, W.-K. (2020). Social big data informs spatially explicit management options for national parks with high tourism pressures. *Tourism Management*, *81*, 104136. https://doi.org/10.1016/j.tourman.2020.104136

Clarke, A., & Craft, J. (2017, December). The vestiges and vanguards of policy design in a digital context. *Canadian Public Administration-Administration Publique Du Canada*, *60*(4), 476–497. https://doi.org/10.1111/capa.12228

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

Conrad, C., Löw, F., & Lamers, J. P. A. (2017). Mapping and assessing crop diversity in the irrigated Fergana Valley, Uzbekistan. *Applied Geography*, *86*, 102–117. https://doi.org/10.1016/j.apgeog.2017.06.016

Constantinides, P., Henfridsson, O., & Parker, G. G. (2018, June). Platforms and infrastructures in the digital age. *Information Systems Research*, *29*(2), 381–400. https://doi.org/10.1287/isre.2018.0794

Croitoru, A., Kien, S., Mahabir, R., Radzikowski, J., Crooks, A., Schuchard, R., Begay, T., Lee, A., Bettios, A., & Stefanidis, A. (2020). Responses to mass shooting events: The interplay between the media and the public. *Criminology & Public Policy*, *19*(1), 335–360. https://doi.org/10.1111/1745-9133.12486

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

De Alban, J. D. T., Connette, G. M., Oswald, P., & Webb, E. L. (2018). Combined Landsat and L-band SAR data improves land cover classification and change detection in dynamic tropical landscapes. *Remote Sensing*, *10*(2), 306. https://doi.org/10.3390/rs10020306

Desdemoustier, J., Crutzen, N., Cools, M., & Teller, J. (2019, September). Smart City appropriation by local actors: An instrument in the making. *Cities*, *92*, 175–186. https://doi.org/10.1016/j.cities.2019.03.021

Donovan, S. (2008). Big data: Teaching must evolve to keep up with advances. *Nature*, *455*, 461. https://doi.org/10.1038/455461d

Du, S. F., Tang, W. Z., Zhao, J. J., & Nie, T. F. (2018, November). Sell to whom? Firm's green production in competition facing market segmentation. *Annals of Operations Research*, *270*(1–2), 125–154. https://doi.org/10.1007/s10479-016-2291-4

El-Taliawi, O. G., & Hartley, K. (2020, October). The COVID-19 crisis and complexity: A soft systems approach. *Journal of Contingencies and Crisis Management*, *29*, 104–107. https://doi.org/10.1111/1468-5973.12337

El-Taliawi, O. G., Nair, S., & Van der Wal, Z. (2021). Public policy schools in the global south: A mapping and analysis of the emerging landscape. *Policy Sciences*, *54*, 371–395. https://doi.org/10.1007/s11077-020-09413-z

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Gawankar, S. A., Gunasekaran, A., & Kamble, S. (2020). A study on investments in the big data-driven supply chain, performance measures and organisational performance in Indian retail 4.0 context. *International Journal of Production Research*, *58*(5), 1574–1593. https://doi.org/10.1080/00207543.2019.1668070

Gerum, P. C. L., Altay, A., & Baykal-Gürsoy, M. (2019). Data-driven predictive maintenance scheduling policies for railways. *Transportation Research Part C: Emerging Technologies*, *107*, 137–154. https://doi.org/10.1016/j.trc.2019.07.020

Greene, D., & Cross, J. P. (2017, January). Exploring the political agenda of the European parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77–94. https://doi.org/10.1017/pan.2016.7

Guo, T., Morgenroth, J., Conway, T., & Xu, C. (2019). City-wide canopy cover decline due to residential property redevelopment in Christchurch, New Zealand. *Science of the Total Environment*, *681*, 202–210. https://doi.org/10.1016/j.scitotenv.2019.05.122

Hämäläinen, M. (2020). A framework for a smart city design: Digital transformation in the Helsinki smart city. In V. Ratten (Ed.), *Entrepreneurship and the community* (pp. 63–86), Contributions to Management Science. Cham, Springer. https://doi.org/10.1007/978-3-030-23604-5_5

Howie, L., Hirsch, B., Locklear, T., & Abernethy, A. P. (2014, July). Assessing the value of patient-generated data to comparative effectiveness research. *Health Affairs*, *33*(7), 1220–1228. https://doi.org/10.1377/hlthaff.2014.0225

Huo, D., & Hung, K. (2015). Internationalization strategy and firm performance: Estimation of corporate strategy effect based on big data of Chinese it companies in a complex network. *Romanian Journal of Economic Forecasting*, *18*(2), 148–163. https://ipe.ro/rjef/rjef2_15/rjef2_2015p148-163.pdf

Hysing, E. (2009). From government to governance? A comparison of environmental governing in Swedish forestry and transport. *Governance*, *22*(4), 647–672. https://doi.org/10.1111/j.1468-0491.2009.01457.x

IBM Center for Business of Government. (2015). Five examples of how federal agencies use big data. Extracted on June 14, 2021 from https://www.businessofgovernment.org/blog/five-examples-how-federal-agencies-use-big-data

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, 0894439320980118. https://doi.org/10.1177/0894439320980118

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

Jung, S. H., & Huh, J. H. (2019, July). A novel on transmission line tower big data analysis model using altered K-means and ADQL. *Sustainability (Switzerland)*, *11*(13), 25, Article 3499. https://doi.org/10.3390/su11133499

Hernandez-Boussard, T., Blayney, D. W., & Brooks, J. D. (2020). Leveraging digital data to inform and improve quality cancer care. *Cancer Epidemiology, Biomarkers & Prevention*, *29*(4), 816–822. https://doi.org/10.1158/1055-9965.EPI-19-0873

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

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

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018, February). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kong, L., & Woods, O. (2018, March). The ideological alignment of smart urbanism in Singapore: Critical reflections on a political paradox. *Urban Studies*, *55*(4), 679–701. https://doi.org/10.1177/0042098017746528

Kwakkel, J. H., & Cunningham, S. C. (2016, October). Improving scenario discovery by bagging random boxes. *Technological Forecasting and Social Change*, *111*, 124–134. https://doi.org/10.1016/j.techfore.2016.06.014

Lai, Y. Y., Sun, H. F., & Ren, J. F. (2018). Understanding the determinants of big data analytics (BDA) adoption in logistics and supply chain management: An empirical investigation. *International Journal of Logistics Management*, *29*(2), 676–703. https://doi.org/10.1108/ijlm-06-2017-0153

Lasswell, H. D. (1970). The emerging conception of the policy sciences. *Policy Sciences*, *1*(1), 3–14. https://doi.org/10.1007/BF00145189

Lasswell, H. D. (1971). *A pre-view of policy sciences*. American Elsevier Publishing Company New York.

Li, J., Zhang, B., & Shi, J. (2017). Combining a genetic algorithm and support vector machine to study the factors influencing CO_2 emissions in Beijing with scenario analysis. *Energies*, *10*(10), 1520. https://doi.org/10.3390/en10101520

Liu, L. C., & Wu, G. (2017, January). The effects of carbon dioxide, methane and nitrous oxide emission taxes: An empirical study in China. *Journal of Cleaner Production*, *142*, 1044–1054. https://doi.org/10.1016/j.jclepro.2016.08.011

Liu, P., & Yi, S. P. (2017, October). Pricing policies of green supply chain considering targeted advertising and product green degree in the Big Data environment. *Journal of Cleaner Production*, *164*, 1614–1622. https://doi.org/10.1016/j.jclepro.2017.07.049

Lyu, X., & Costas, R. (2020). How do academic topics shift across altmetric sources? A case study of the research area of Big Data. *Scientometrics*, *123*(2), 909–943. https://doi.org/10.1007/s11192-020-03415-7

Marelli, L., Lievevrouw, E., & Van Hoyweghen, I. (2020). Fit for purpose? The GDPR and the governance of European digital health. *Policy Studies*, *41*(5), 447–467. https://doi.org/10.1080/01442872.2020.1724929

Marley, T. L. (2019, May). Indigenous data sovereignty: University institutional review board policies and guidelines and research with American Indian and Alaska native communities. *American Behavioral Scientist*, *63*(6), 722–742. https://doi.org/10.1177/0002764218799130

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. UK, John Murray.

Mostert, M., Bredenoord, A. L., Biesaart, M., & van Delden, J. J. M. (2016, July). Big Data in medical research and EU data protection law: Challenges to the consent or anonymise approach. *European Journal of Human Genetics*, *24*(7), 956–960. https://doi.org/10.1038/ejhg.2015.239

Moyano, A., Moya-Gomez, B., & Gutierrez, J. (2018, December). Access and egress times to high-speed rail stations: A spatiotemporal accessibility analysis. *Journal of Transport Geography*, *73*, 84–93. https://doi.org/10.1016/j.jtrangeo.2018.10.010

Nazir, M. S., Alturise, F., Alshmrany, S., Nazir, H., Bilal, M., Abdalla, A. N., Sanjeevikumar, P., & M. Ali, Z. (2020). Wind generation forecasting methods and proliferation of artificial neural network: A review of five years research trend. *Sustainability (Switzerland)*, *12*(9), 3778. https://doi.org/10.3390/su12093778

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

Nieminen, H. (2016, February). Digital divide and beyond: What do we know of Information and Communications Technology's long-term social effects? Some uncomfortable questions. *European Journal of Communication*, *31*(1), 19–32. https://doi.org/10.1177/0267323115614198

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M. A., Chute, C. G., & Musen, M. A. (2009, July). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, *37*, W170–W173. https://doi.org/10.1093/nar/gkp440

Ooms, J. (2018). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. In (Version R package version 3.0.). https://CRAN.R-project.org/package=hunspell

Pandey, A., Mereddy, S., Combs, D., Shetty, S., Patel, S. I., Mashaq, S., Seixas, A., Littlewood, K., Jean-Luis, G., & Parthasarathy, S. (2020). Socioeconomic inequities in adherence to positive airway pressure therapy in population-level analysis. *Journal of Clinical Medicine*, *9*(2), 442. https://doi.org/10.3390/jcm9020442

Rabini, C., Brummer, K., Dimmroth, K., & Hansel, M. (2020). Profiling foreign policy leaders in their own language: New insights into the stability and formation of leadership traits. *The British Journal of Politics and International Relations*, *22*(2), 256–273. https://doi.org/10.1177/1369148120910984

Ranagalage, M., Murayama, Y., Dissanayake, D., & Simwanda, M. (2019). The impacts of landscape changes on annual mean land surface temperature in the tropical mountain city of Sri Lanka: A case study of Nuwara Eliya (1996–2017). *Sustainability (Switzerland)*, *11*(19), 5517. https://doi.org/10.3390/su11195517

Rasouli, S., & Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research*, *14*(4), 412–424.

Rieser, V., & Lemon, O. (2011). Learning and evaluation of dialogue strategies for new applications: Empirical methods for optimization from small data sets. *Computational Linguistics*, *37*(1), 153–196. https://doi.org/10.1162/coli_a_00038

Rikin, S., Glatt, K., Simpson, P., Cao, Y., Anene-Maidoh, O., & Willis, E. (2015, November–December). Factors associated with increased reading frequency in children exposed to reach out and read. *Academic Pediatrics*, *15*(6), 651–657. https://doi.org/10.1016/j.acap.2015.08.008

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

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (487–494), Banff, Canada

Sanga, S. (2019). A new strategy for regulating arbitration. *Northwestern University Law Review*, *113*(5), 1121–1161. https://scholarlycommons.law.northwestern.edu/nulr/vol113/iss5/5

Sözen, A., & Arcaklioğlu, E. (2011). Empirical analysis and future projection of the ratio of primary energy production to consumption (RPC) in Turkey and upgrading policies. *Energy Sources, Part B: Economics, Planning, and Policy*, *6*(3), 263–279. https://doi.org/10.1080/15567240802534268

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

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

Talamini, E., & Dewes, H. (2012, February). The macro-environment for liquid Biofuels in Brazilian science and public policies. *Science and Public Policy*, *39*(1), 13–29. https://doi.org/10.3152/030234212x13214603531923

Townsend, B. A., & Thaldar, D. W. (2019). Navigating uncharted waters: Biobanks and informational privacy in South Africa. *South African Journal on Human Rights*, *35*(4), 329–350. https://doi.org/10.1080/02587203.2020.1717366

van Loenen, B., Kulk, S., & Ploeger, H. (2016, April). Data protection legislation: A very hungry caterpillar: The case of mapping data in the European Union. *Government Information Quarterly*, *33*(2), 338–345. https://doi.org/10.1016/j.giq.2016.04.002

Wang, X., McCallum, A. & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA.

Wen, L., & Cao, Y. (2020). Influencing factors analysis and forecasting of residential energy-related CO_2 emissions utilizing optimized support vector machine. *Journal of Cleaner Production*, *250*, 119492. https://doi.org/10.1016/j.jclepro.2019.119492

Wijffels, J. (2020). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. In (Version R package version 0.8.4-1). https://CRAN.R-project.org/package=udpipe

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

Ying, C.-S., Chow, A. H., & Chin, K.-S. (2020). An actor-critic deep reinforcement learning approach for metro train scheduling with rolling stock circulation under stochastic demand. *Transportation Research Part B: Methodological*, *140*, 210–235. https://doi.org/10.1016/j.trb.2020.08.005

Zenk, S. N., Tarlov, E., Powell, L. M., Wing, C., Matthews, S. A., Slater, S., Gordon, H. S., Berbaum, M., & Fitzgibbon, M. L. (2018, March). Weight and Veterans' Environments Study (WAVES) I and II: Rationale, methods, and cohort characteristics. *American Journal of Health Promotion*, *32*(3), 779–794. https://doi.org/10.1177/0890117117694448

Zhang, X. Y., & Destech Publicat, I. (2017). Selection of Small & Micro-enterprise Financing Methods Under the Background of Internet Finance. In *3rd International Conference on Social Science and Development* (pp. 123–127). Destech Publications, Inc. https://doi.org/10.12783/dtssehs/icssd2017/19191

Zhao, H., Guo, S., & Zhao, H. (2017). Energy-related CO_2 emissions forecasting using an improved LSSVM model optimized by whale optimization algorithm. *Energies*, *10*(7), 874. https://doi.org/10.3390/en10070874

Zharova, A. K., & Elin, V. M. (2017, August). The use of Big Data: A Russian perspective of personal data security. *Computer Law & Security Review*, *33*(4), 482–501. https://doi.org/10.1016/j.clsr.2017.03.025

AUTHOR BIOGRAPHIES

Ola G. El-Taliawi is a Postdoctoral Fellow at the Department of Political Science, Faculty of Public Affairs, Carleton University.

Nihit Goyal is an Assistant Professor at the Faculty of Technology, Policy and Management, Delft University of Technology (TU Delft). His research interests lie in comparative public policy, computational social science, and the sustainable energy transition.

Michael Howlett is the Burnaby Mountain Professor and Canada Research Chair at the Department of Political Science, Simon Fraser University.

How to cite this article: El-Taliawi, O. G., Goyal, N., & Howlett, M. (2021). Holding out the promise of Lasswell's dream: Big data analytics in public policy research and teaching. *Review of Policy Research*, 00, 1–21. https://doi.org/10.1111/ropr.12448