Parallel or Intersecting Lines? Intelligent Bibliometrics for Investigating the Involvement of Data Science in Policy Analysis

Yi Zhang,[1] Alan L. Porter,[2, 3] Scott Cunningham,[4] Denise Chiavetta,[3] Nils Newman[3]

[1]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia;
[2]Technology Policy and Assessment Centre, Georgia Institute of Technology, USA;
[3]Search Technology Inc., USA;
[4]School of Government and Public Policy, University of Strathclyde, UK;

Emails: yi.zhang@uts.edu.au; alan.porter@isye.gatech.edu; scott.cunningham@strath.ac.uk; dchiavetta@searchtech.com; newman@searchtech.com.

Abstract

Efforts to involve data science in policy analysis can be traced back decades but transforming analytic findings into decisions is still a far from straightforward task. Data-driven decision-making requires understanding approaches, practices, and research results from many disciplines, which makes it interesting to investigate whether data science and policy analysis are moving in parallel or whether their pathways have intersected. Our investigation, from a bibliometric perspective, is driven by a comprehensive set of research questions, and we have designed an intelligent bibliometric framework that includes a series of traditional bibliometric approaches and a novel method of charting the evolutionary pathways of scientific innovation, which is used to identify predecessor-descendant relationships in technological topics. Our investigation reveals that data science and policy analysis have intersecting lines, and it can foresee that a cross-disciplinary direction in which policy analysis interacting with data science has become an emergent area in both communities. However, equipped with advanced data analytic techniques, data scientists are moving faster and further than policy analysts. The empirical insights derived from our research should be beneficial to academic researchers and journal editors in related research communities, as well as policy-makers in research institutions and funding agencies.

Keywords: bibliometrics; science maps; policy analysis; data science.

Managerial Relevance Statement

This paper presents an intelligent bibliometric framework for empirically investigating the involvement of data science in policy analysis in five respects: 1) how much effort has been expended in the involvement, and how has that effort changed over time? 2) Who are the key players leading the involvement, and what are their geographical distributions and collaboration patterns? 3) Which academic journals publish this research, and what are their citation behaviors? 4) Which topics does the research involve and how do their efforts interact? 5) How does the involvement evolve over time, and which evolutionary pathways are apt to be cutting-edge in the near future? Our research shows that there are intersecting lines between data science and policy analysis. A cross-disciplinary direction interacting policy analysis with data science has become an emergent area in both communities. However, data scientists are equipping themselves with advanced data analytic techniques and are motived by a passion for developing new methods for examining real-world political issues, and so are moving further and faster than policy analysts.

Introduction

If one considers data science from a narrow view, i.e., information technology (IT)-based data analytics, one might observe that, historically, policy analysis and data science have operated in parallel lines. That is, traditional policy analysis is driven by actual political problems and emphasizes the use of inquiries and argumentations for proposing solutions [1], while innovation in information technology is usually motivated by an increase in data and its complexity, which means new ways to effectively and efficiently extract information and knowledge need to be developed [2]. Bi-directional efforts to foster interactions between data science and policy analysis can be traced back for decades. For example, information systems to support policy decision-making emerged in the early 1970s [3] and, in the 1990s, some policy analysts were pioneering bibliometrics as a way to profile science, technology, and innovation policy (STIP) [4, 5]. But it is clear that gaps still exist in transforming analytical results (e.g., predictions) into decisions, and that data-driven decision-making requires an understanding of the approaches and practices derived from decades of multidisciplinary research [6]. Given this, it is interesting to delve into those previous efforts to explore insights on how data science has been involved in policy analysis – what is the current status and what are the future research directions. Moreover, we argue that technology management is the private sector counterpart to policy analysis. Therefore, this research should be of interest to the technology management community as well.

Known as a toolkit with a series of statistical approaches for analyzing scientific documents (e.g., scientific articles, patents, and academic proposals) [7], bibliometrics has long been used as an analytic tool to either systematically review literature in a given technological area or scientific discipline [8], or handle specific issues within certain frameworks of technology management - e.g., technology roadmapping [9] and technology convergence [10, 11]. However, very few, if any, previous studies have constructed a framework of bibliometric analysis to delve into a case through answering questions such as "who", "where", "what", "why", and "how" within one system. Traditional bibliometrics profile key topics (i.e., "what") and players (i.e., "what" and "where") using citation/co-citation and co-word statistics, but fail to identify complicated relationships to explain "why" and "how". Novel bibliometric approaches, with the aid of advanced information technologies (e.g., machine learning and streaming data analytics), create new opportunities to uncover such relationships. Thus, applying these new techniques to discover the dynamics of such involvements over time would provide value-added information to help understand hidden mechanisms by identifying possible push and pull forces. Further, investigations of the involvement of data science in policy analysis is an emergent interest for the STIP community, but limited literature addresses this interface.

Aiming to construct a bibliometrics-based research framework for systematically exploring the activities that involve data science in supporting policy analysis, we specify the 'who, where, what, why & how' questions into five research questions:

1) WHAT: How much effort has been expended in the involvement, and how has that effort changed over time?
2) WHO & WHERE: Who are the key players leading the involvement, and what are their geographical distributions and collaboration patterns?
3) WHO: Which academic journals publish this research, and what are their citation behaviors?
4) WHAT: Which topics does the research involve and how do their efforts interact?
5) HOW & WHY: How does the involvement evolve over time, and which evolutionary pathways are apt to be cutting-edge in the near future?

To answer these questions, we developed a novel framework that draws together various techniques and strands of bibliometrics into a new sphere of research we term *intelligent bibliometrics*. Intelligent bibliometrics spans traditional bibliometric approaches, such as co-authorship, co-citation, and co-word analysis to profile trends in collaboration, coverage, and topics of research. It also integrates science maps for visually delivering analytic results and charts evolutionary pathways in topics and fields to identify the predecessor and descendant technologies [12].

Three datasets were constructed to explore these questions, each with a different purpose. The first comprises a set of articles published in *Nature* and *Science* as a reflection of novel attempts to involve data science in policy analysis. The remaining two datasets contain articles published in top-level journals in the two fields respectively, to capture the efforts contributed by the two research communities.

It is our intention to provide empirical insights on the involvement of data science in policy analysis to provide decision support for academic researchers and journal editors in these research communities. Policy-makers in research institutions and funding agencies should also significantly benefit from the findings of this study. Additionally, the designed framework, which integrates traditional and new bibliometric approaches, creates a reference for further bibliometric studies.

The rest of this paper is organized as follows: Related Works systematically reviews bibliometrics and its applications. The Data and Methodology section follows, which presents the details of the related datasets and methods. The Empirical Study reports the findings observed in this investigation, and the last section concludes the research and outlines its limitations and our future studies.

Related Works

Bibliometrics is known as an effective tool for exploiting statistical approaches for quantitatively analyzing scientific documents [7]. It has been widely used in STIP studies, e.g., profiling technological landscapes with multiple entities like individuals, institutions, and countries [13, 14]; identifying emerging topics in science and technology [15, 16]; and tracing the evolutionary pathways of research disciplines and emerging technologies [17, 18]. These bibliometric approaches mostly rely on bibliometric indicators such as citation/co-citation statistics, co-word statistics, co-authorships [19], coupled with science maps to visually illustrate the outcomes of the investigation [20, 21]. Compared to reviews based on expert knowledge, bibliometric case studies provide a solution to systematically review relatively large-scale documents and explore objective results while minimizing subjective bias [22].

As information technologies rapidly advance and artificial intelligence techniques, in particular, novel bibliometric approaches are emerging to handle challenging issues. For example, information visualization provides effective tools to vividly deliver and present bibliometric results [23, 24], and science maps further enhance the connections between bibliometric indicators and practical findings [19, 25]. Topic models and their benefits to bibliometric analysis have been well investigated since the early 2010s [26, 27], and network analytics has been integrated with science maps (especially co-authorship maps) to identify scientific activity (e.g., collaborative behaviors) [28, 29]. Machine learning techniques have significantly enhanced the efficacy of traditional bibliometric approaches in advanced data analytics [12, 30], and the rise of deep learning techniques has given rise to solutions for knowledge representation and deeper topic extraction [31].

That bibliometrics can be used as an important instrument to empirically investigate the development of science and technology has become common sense thinking. A wide range of sectors and scientific disciplines have seen its benefits [32], and these novel information technologies have only served to strengthen its advantages. From our literature review, we identified three key capabilities of bibliometrics for supporting decision making and policy analysis: 1) profiling key technological players, core technological components, and their relationships [8, 16, 33-35]; 2) tracing the evolutionary pathways of science and technology [36-39]; and 3) forecasting future technological trends with the aid of expert knowledge [9, 15, 40]. As highlighted in [41], bibliometrics is a way to detect what is emerging, to operationalize growth, to radicalize novelty, and to muster emerging technologies. Despite the fact that bibliometrics has been widely applied for empirically reviewing research disciplines and technological areas - e.g., bibliometrics-based trend analysis on knowledge management research [42] and investigations on China's nanoscience and nanotechnology [43], such studies usually concentrates on either an individual domain or certain highly related domains, rather than uncovering relationships among multiple domains. One representative bibliometric tool for exploring multidisciplinary interactions is a science overlay map [19], which illustrates such interactions at a macro level via overlays of nodes and edges on a format of science maps. Given the circumstance, it is feasible to foresee the potential of extending bibliometric studies from profiling individual disciplines to investigating the interactions between two disciplines and research domains, or even more, with the aid of advanced information technologies for identifying complicated relationships.

Data and Methodology

Our first step in investigating the involvement of data science in policy analysis was to specifically define the two fields. These were formulated from the literature review as follows:

- Based on the Lasswellian commitments [44] and a general definition given by William Dunn [1], we defined policy analysis as the use of "multiple methods of inquiry and argument to produce and transform policy-relevant information … to resolve policy problems".

- Data science, at a high level, is defined as "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data [45]". However, in this study, we focused on a relatively narrow concept of data science, which was data mining with an emphasis on the use of IT-based data analytics for information retrieval and knowledge discovery.

Given these definitions, we assumed there were two ways of involving data science in policy analysis: 1) A top-down approach, where cutting-edge policy analysis research that incorporates data science (or vice versa) initially appears in top-level journals, such as *Nature* and *Science*. More extensive follow-up studies then ensue. 2) A bottom-up approach where researchers from each of the two fields spontaneously come together to undertake a study. Hence, we constructed three datasets for our investigation:

- Dataset 1 – articles published in *Nature* and *Science*, representing a joint dataset with the two fields;
- Dataset 2 – articles published in top-level journals in the area of policy analysis;
- Dataset 3 – articles published in top-level journals in the area of data science.

Data

We collected the three datasets from the Web of Science (WoS)[1] on April 15, 2019, covering all articles published in the target journals before that date. The search strategy, including the target journals and the number of collected articles, appears in Table 1.

Table 1. Search Strategy for the Three Datasets

NO	***#A[1]***	***Search Strategy***
#1	82	TS[2]= (data AND policy) AND SO[3] = (*Nature* OR *Science*)
#2	1990	TS= (data SAME (big OR analy* OR science)) AND SO = (*American Journal of Political Science* OR *World Politics* OR *Journal of Public Administration Research and Theory* OR *Public Administration Review* OR *Review of International Political Economy* OR

[1] https://webofknowledge.com/

		Journal of Policy Analysis and Management OR *International Organization* OR *Political Analysis* OR *American Political Science Review* OR *British Journal of Political Science* OR *Research Policy*)
#3	597	TS= (policy SAME analy*) AND SO = (*MIS Quarterly* OR *Journal of Information Technology* OR *Information Sciences* OR *IEEE Systems Journal* OR *Journal of Strategic Information Systems* OR *Business & Information Systems Engineering* OR *Information & Management* OR *Decision Support Systems* OR *European Journal of Information Systems* OR *Information Systems* OR *Journal of Management Information Systems* OR *Journal of the Association for Information Science and Technology* OR *ACM Transactions on Information Systems* OR *Journal of the Association for Information Systems* OR *Knowledge-based Systems* OR *Expert Systems with Applications* OR *International Journal of Intelligent Systems* OR *Communications of The ACM*)
Timeline		All articles published before April 15, 2019

Note: 1) Number of articles; 2) Topic, including title, abstract, and keywords; and 3) Publication name.

Details of the three datasets follows.

- Dataset 1 should reflect cutting-edge knowledge to sharpen our research hypotheses on what data science offers to policy analysis. Thus, this dataset contained any article that addressed both data and policy issues in the world-leading journals *Nature* or *Science*. We manually browsed 82 articles meeting our criteria and observed that: *Nature* and *Science* provide an interesting selection of papers of potentially high-impact; and most of the articles concerned STIP studies, while some covered data analytics for STIP issues, e.g., [46, 47]. This is, admittedly, a small and specific sample but one that can offer exploratory work that could connect data science and policy analysis.
- Dataset 2 contained papers from the top 10 journals in 2016 in the WoS subject categories “political science” and “public administration”. Top ten means those with the highest impact factor. Additionally,

since the scope of the journal *Research Policy* is highly relevant to our study, we also included this title. It is clear that these are the leading journals in the area of policy analysis and, thus, we believe this dataset of 1990 articles could be the best source of information to reveal future research frontiers.

- Dataset 3 initially included journals aligning with the two WoS subject categories "computer science & information systems" and "computer science & artificial intelligence", but we are also fully aware that some journals within these categories are purely technical. Thus, with the help of several IT researchers from the Centre for Artificial Intelligence at the University of Technology Sydney, we selected 19 journals as representative of policy analysis-oriented data analytics. Each journal features the application of information systems and artificial intelligence techniques to decision/policy-making support.

Intelligent Bibliometrics

The research framework for investigating the involvement of data science in policy analysis is provided in Figure 1. The investigation is specified via five research questions for answering 'who, where, what, why & how' issues:

- Q1: How much effort has been expended in the involvement, and how has that effort changed over time? For statistically profiling what happened in this area;
- Q2: Who are the key players leading the involvement, and what are their geographical distributions and collaboration patterns? For identifying who and where they are;
- Q3: Which academic journals publish this research, and what are their citation behaviors? For identifying who are the involved research communities;
- Q4: Which topics does the research involve and how do their efforts interact? For discovering what happened in this area;
- Q5: How does the involvement evolve over time, and which evolutionary pathways are apt to be cutting-edge in the near future? For empirically investigating how topics evolved and why it happened.

We used descriptive analytics to answer Question 1 - specifically, the number of published articles in the three datasets per annum over the period. We drew on co-authorship analysis to explore Question 2 by generating maps of co-author affiliations to highlight collaborations among key players in these fields [48]. Question 3 was investigated through citation and co-citation maps at a journal level as a way to derive insights into which journals are highly influential [49]. We answered Questions 4 and 5 with term-based analysis and corresponding co-term maps [50] followed by tracing the evolutionary pathways in various topics [12]. The co-term map was used to identify key scientific topics and their semantic relationships, as well as to profile a topical landscape of the field. Scientific evolutionary pathways (SEP) were mapped with machine learning techniques and streaming data analytics.

Two pieces of software were used for data pre-processing and visualization. VantagePoint[2] was used for: 1) name disambiguation and consolidating sub-branches of institutions, e.g., the "Chinese Academy of Sciences" and the "Institute of Software, Chinese Academy of Sciences"; 2) stem-based fuzzy matching of terms and affiliation names; and 3) term clumping [51]. VoSviewer [23] was used to generate the science maps, i.e., the co-term, co-authorship, and co-citation maps.

Specifically, aiming to trace the evolution of scientific topics for Question 5, a modified version of the SEP approach [12] was proposed for identifying the predecessor-descendant relationships between scientific topics. The SEP approach follows the two definitions below:

Definition 1: a topic is a collection of articles and is geometrically represented as a circle, in which the centroid is identified as the article sharing the highest similarity with all other articles in the topic and the boundary is the largest Euclidean distance between the centroid and all articles.

Definition 2: if topic A appears after topic B in time and topic B is the one of all existing topics sharing the highest similarity with topic A, topics A and B are identified as descendant and predecessor in this relationship.

[2] VantagePoint is commercial software used in text mining and particularly in science, technology, and innovation text analysis. More details can be found on their website: https://www.thevantagepoint.com/

The stepwise algorithm of the modified SEP approach in this study is described as follows:

Step 1: Integrate the three datasets into one, and simulate the integrated dataset as a data stream that consists of 20 time slices based on the year of publication. Articles published in 2000 and before were set as the initial slice and based on our reading[3] and understanding on the three datasets, we manually set 'public policy' as the only initial topic.

Step 2: Process the data stream in an iterative flow - i.e., process one time slice in one iteration by analyzing their articles one by one.

Step 3: Measure the cosine similarity between a forthcoming article and the centroid of all existing topics and assign it to the most similar topic.

Step 4: Calculate the Euclidean distance between the article and the centroid of its assigned topic. If the distance is within a lower range (10% in this case) of the boundary, the article will be directly involved in this topic; if the distance is between the lower range and upper range (10%, as well) of the boundary, we set the article as "evolution." If it is larger than the upper range of the boundary, we consider it may be an irrelevant item and remove it from the topic.

Step 5: Group articles labeled with 'evolution' are assigned into certain sub-topics by using an unsupervised k-means approach [52], measuring the cosine similarity between these sub-topics and two sets of topics – i.e., their assigned topic and dead topics (motivated by the studies of 'sleeping beauty' identification [53]. We set a topic as 'dead' when it does not receive new assigned articles in two continuous time slices). In a common situation, a new sub-topic will share a higher similarity with its assigned topic, and that is the relationship between a descendant (i.e., the sub-topic) and its predecessor (i.e., the assigned topic). However, a 'dead' topic could resurge, if it shares a higher similarity with a new sub-topic, and thus, this 'dead' topic will become the predecessor of this new sub-topic.

[3] We screened the titles of around 200 articles and randomly read the abstracts of 20 articles from the integrated dataset, and noticed the majority of articles in the initial time slice relates to policy analysis.

Step 6: Update the centroid and boundary of all existing topics and return to Step 2 until the stream ends.

Compared to traditional bibliometrics, the use of streaming data analytics and machine learning techniques in the SEP approach benefits in identifying complicated relationships and tracing potential topic changes in a dynamic scenario. Here, in this research framework we highlight the development and application of intelligent models for recognizing patterns in bibliometrics and entitle this cross-disciplinary direction *Intelligent Bibliometrics.* Specifically, such intelligent models could be any computational models incorporating advanced data analytic approaches and/or artificial intelligence techniques, such as optimization, streaming data analytics, network analytics, fuzzy systems, and various machine learning techniques (e.g., neural networks). In fact, we have pursued studies along this direction for years - e.g., incorporating word embedding techniques for topic extraction [31], and network analytic models for scientific behavior recommendation and prediction [52, 54, 55]. This extends the disciplinary scope of bibliometrics from information and library science to broad business and computer science disciplines; it differs from traditional statistics-based bibliometrics (e.g., science maps) and qualitative approaches in technology management (e.g., technology roadmapping).

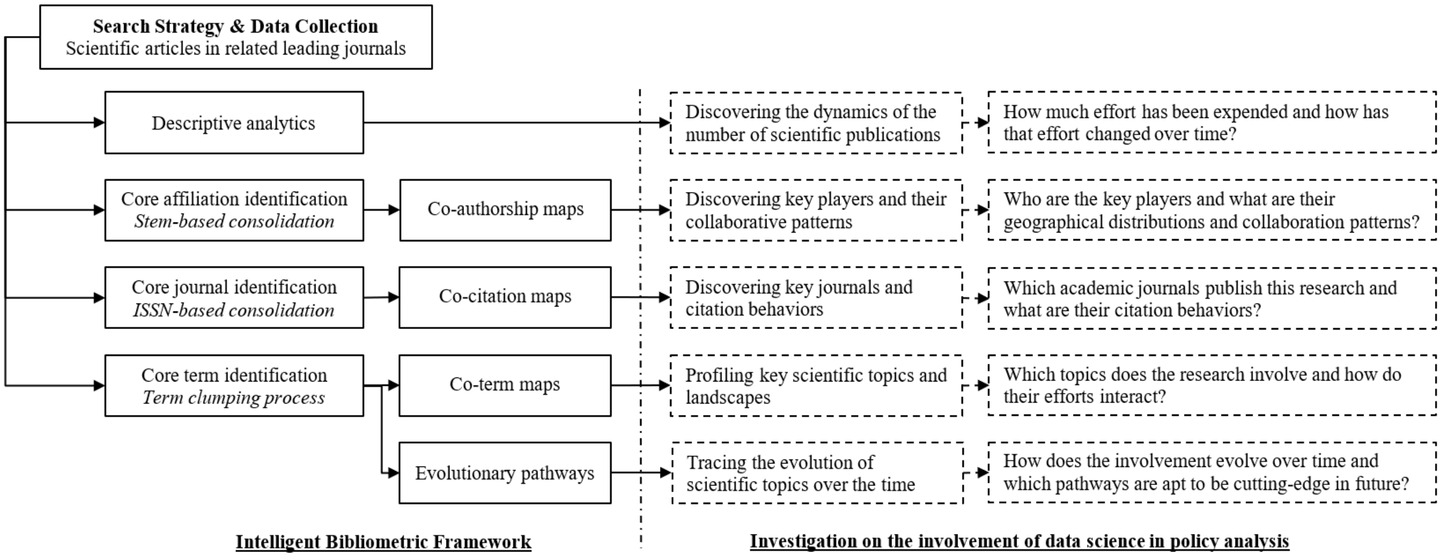

Search Strategy & Data Collection
Scientific articles in related leading journals
Descriptive analytics
Core affiliation identification
Stem-based consolidation
Co-authorship maps
Core journal identification
ISSN-based consolidation
Co-citation maps
Core term identification
Term clumping process
Co-term maps
Evolutionary pathways
Discovering the dynamics of the number of scientific publications
How much effort has been expended and how has that effort changed over time?
Discovering key players and their collaborative patterns
Who are the key players and what are their geographical distributions and collaboration patterns?
Discovering key journals and citation behaviors
Which academic journals publish this research and what are their citation behaviors?
Profiling key scientific topics and landscapes
Which topics does the research involve and how do their efforts interact?
Tracing the evolution of scientific topics over the time
How does the involvement evolve over time and which pathways are apt to be cutting-edge in future?
Intelligent Bibliometric Framework
Investigation on the involvement of data science in policy analysis

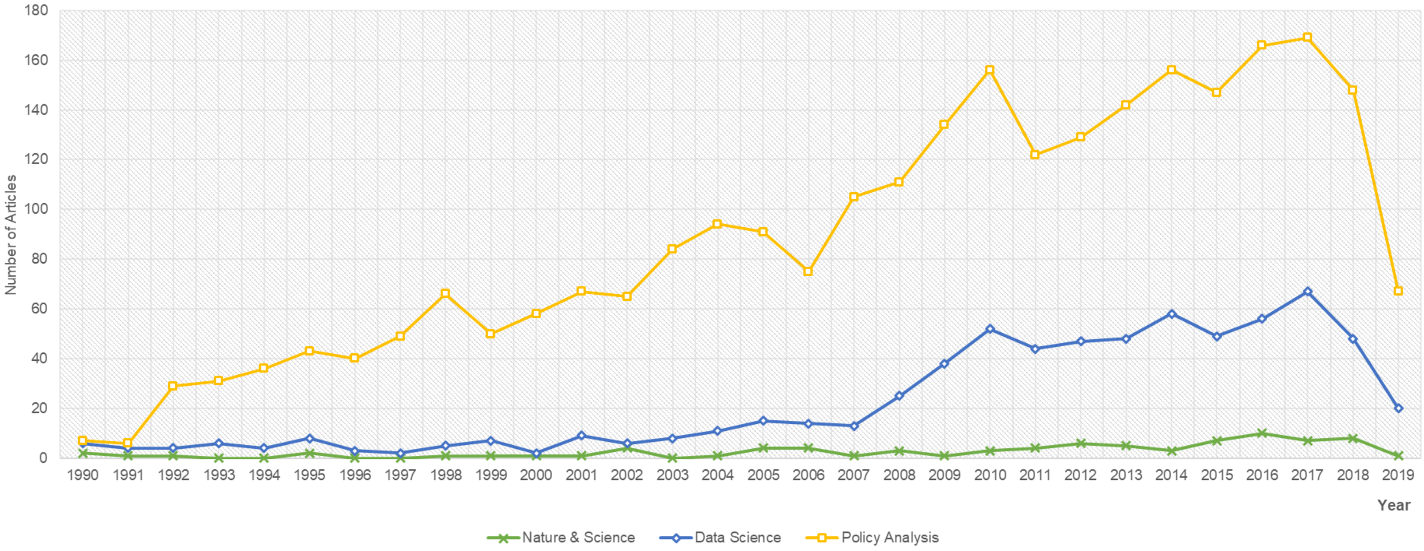
Figure 1. Intelligent Bibliometrics – a Research Framework

Empirical Study: Involvement of Data Science in Policy Analysis

Using the data and methods outlined in the previous section, this empirical study explores the five specified research questions on the involvement of data science in policy analysis.

Q1: How much effort has been expended in the involvement, and how has that effort changed over time?

We answered this question by tracing trends of the number of published articles in the three datasets. The results are shown in Figure 2. It is clear that the number of articles in Dataset 1 stayed relatively stable with a slight rise after 2009. Comparably, the number of articles in Datasets 2 and 3 steadily increased over the period, reaching a peak in 2017. Based on our observations, and the fluctuations at certain turning points in particular, we drew the following conclusions:



Note: 1) articles published prior to 1990 are combined with those published in 1990; and 2) publication lag is likely to account for the rapid decrease in 2019.

Figure 2. The Dynamics of the Number of Published Articles in the Three Datasets

- The interactions between data science and policy analysis are attracting increasing attention from both communities.
- A big data boom around 2008 accelerated these interactions, which resulted in the first peak in 2010.
- A natural correction occurred in 2011, followed by a second acceleration until 2014 likely due to rapid developments in artificial intelligence, and especially deep learning techniques.

- However, there is gloom ahead given the number of articles in 2018. The boom has clearly ended.

We should emphasize that our observations and understandings are only based on the number of published articles. While this is a fairly accurate reflection of the activities of the entire research community, it might not comprehensively indicate the depth of the interactions between the two research areas. Such concerns will be addressed through the evolutionary pathways examined in Question 5.

Q2: Who are the key players leading the involvement, and what are their geographical distributions and collaboration patterns?

251, 1055, and 702 affiliations were retrieved from the three datasets respectively, and a clean-up function in VantagePoint for light name disambiguation was applied for consolidating affiliations that align in the same organization but with different branch/department names. Thus, 249, 978, and 676 affiliations were identified respectively and the top 10 affiliations publishing the largest number of articles in the related datasets are listed in Table 2.

Table 2. Leading Affiliations in the Involvement of Data Science in Policy Analysis

No	*Nature & Science (D1)*	*Policy Analysis Journals (D2)*	*Data Science Journals (D3)*
1	Univ Calif Berkeley,	Harvard Univ (88)	City Univ Hong Kong (15)
2	Univ Edinburgh,	Texas A&M Univ (52)	Arizona State Univ,
3	Univ Washington (6)	Univ Michigan (50)	Hong Kong Polytech Univ (8)
4	Harvard Univ,	Univ N Carolina (49)	Natl Cheng Kung Univ,
5	MIT,	Ohio State Univ (46)	Natl Chiao Tung Univ,
6	Stanford Univ (5)	Princeton Univ,	Natl Univ Singapore,
7	Princeton Univ,	Stanford Univ (39)	Univ Florida,
8	Univ Calif San Diego (4)	Columbia Univ (38)	Univ Regina,
9	Columbia Univ,	Washington Univ (37)	Xi’dian Univ,

10	NIH, Univ Cambridge, Univ Exeter, Univ London Imperial, Univ Warwick, World Bank (3)	Yale Univ (36)	Yonsei Univ (7)

Note: 1) D1/2/3 = Dataset 1/2/3; and 2) the number in brackets represents the number of articles published by this affiliation in this dataset.

Certain interesting insights come from the geographical distribution of these leading affiliations: 1) Universities from the US and UK dominate the research published in *Nature* and *Science*. It is also surprising that all of the top 10 affiliations are located in the US. 2) The strength of Asian universities in computer and data science has expanded into policy analysis, as indicated by the 7 of 10 leading universities from these regions out of Dataset 3 – Hong Kong (China), Taiwan (China), Singapore, China, and South Korea. This phenomenon may be the result of a Western orientation toward particular mindsets as valuable – mindsets that are likely much more apparent in the social sciences, such as policy analysis, than in mathematics and computer sciences, such as data science. Such cultural gaps, coupled with language issues, may prevent Asian researchers from publishing articles in top-tier policy analysis journals. Conversely, publishing in high-quality data science journals may relatively easy.

For each dataset, VOSViewer identified a co-authorship network, as illustrated in Figure 3. Our observations and interpretations of the collaborative behaviors hidden in these maps follow.

Dataset 1 (*Nature* & *Science*)[4]:

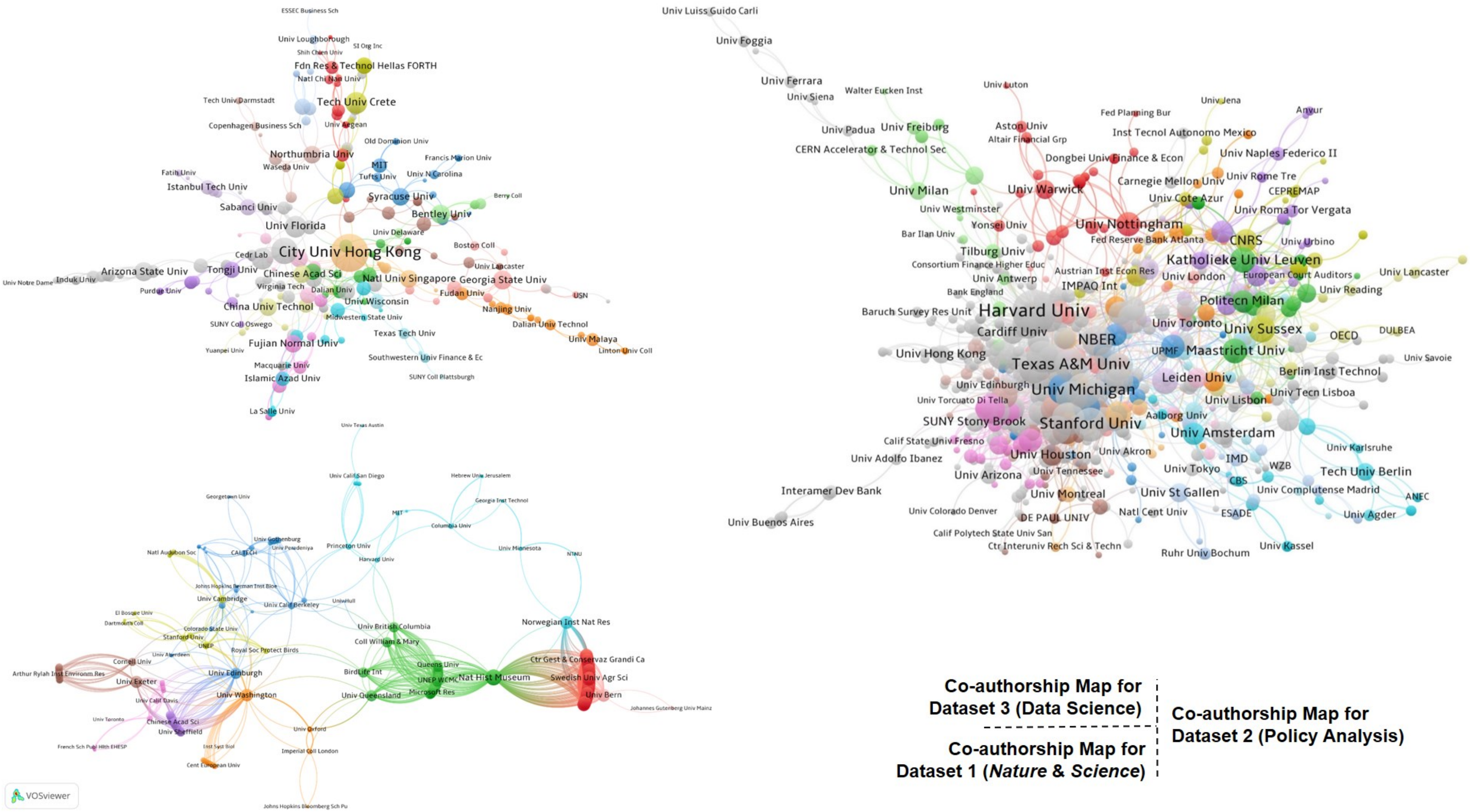
[4] It is essential to note that there were only 82 articles in this dataset, and our observations might be influenced by the small sample size.

- The thickest linkages indicate a collaborative project, with the participation of the Natural History Museum of Denmark and a number of European research institutions and universities;
- US universities (e.g., Princeton Univ, Harvard Univ, and Univ Calif Berkeley) appear to collaborate both domestically, forming their own in-country networks, and internationally – especially those on the west coast (e.g., Univ Calif Berkeley, Caltech, Stanford, and Univ Washington);
- Comparably, UK universities (e.g., Univ Edinburgh, Univ Cambridge, and Univ Sheffield) are more prone to international collaborations (e.g., Univ Toronto, Chinese Acad Sci, and Cent European Univ).

It is interesting to consider the co-authorship maps for Dataset 2 (Policy Analysis) and Dataset 3 (Data Science) together. On the surface, the patterns coincide with our observations from Table 2; that is, the diverse foci on research disciplines and topics between researchers from Western and Eastern countries.

- Research institutions (in particular universities) from the US and Europe show strength in the area of policy analysis with the involvement of data science, and their collaborations are active, interactive, and extensive.
- Comparably, Asian universities (especially those from China) dominate data analytics, underpinned by policy analysis.
- It is intriguing that the co-authorship map for Dataset 3 suggests researchers across the globe are collaborating in relatively isolated groups and with strong geographical preferences (i.e., Western vs. Asian universities). This is reflected as diverging chains/clusters in the map, e.g., the chain linking Fudan Univ, Nanjing Univ, Univ Malaya, etc. or the cluster of Syracuse Univ, MIT, Univ N Carolina, etc.

In summary, the leading universities from the US and UK are the key players in policy analysis. Their research is top-down from a policy analysis problem to a solution engaging data analytics. Cultural and language issues may explain why we do not see this route as suitable for researchers from non-English speaking countries, and especially Asian countries. Asian researchers also take a top-down approach but from data science to policy analysis as a way to evaluate and verify methods.



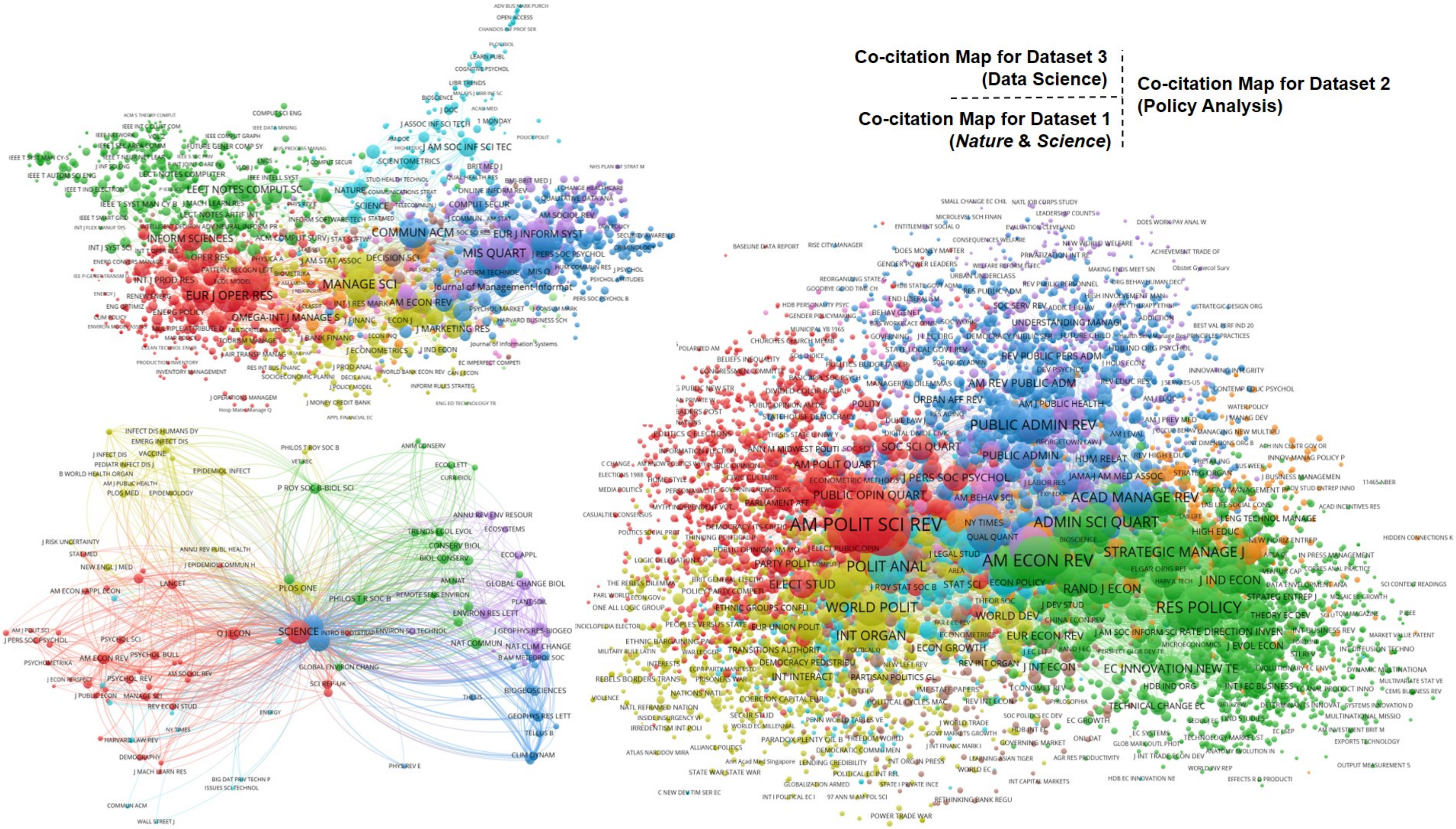
Note that the size of nodes indicates their collaborative strengths.

Figure 3. Co-authorship Maps for the Three Datasets

Q3: Which academic journals publish this research, and what are their citation behaviors?

Investigating journals that publish high-quality research articles on policy analysis through data science provides clues into potential knowledge flows among research disciplines and related communities who hold interests in this topic. Co-citation analysis for journals is based on the frequency with which two journals are cited together by other journals. Thus, we removed journals that were only cited once in each dataset since they could never be co-cited. 148, 6607, and 1954 co-cited journals from the three datasets were identified, and a series of co-citation maps were generated with VOSViewer, as shown in Figure 4. In general, the three datasets reveal extremely diverse citation behaviors.

Following the scope of *Nature* and *Science*, the co-citation map for Dataset 1 indicates significant multidisciplinary distributions. 1) In addition to *Nature* and *Science*, other multidisciplinary-oriented journals are located in the center of the map, e.g., *Proceedings of the National Academy of Sciences of the United States of America* (the label is hidden behind *Science* in Figure 4) and *PloS One*. 2) Economic journals account for a significant proportion of the cited journals, e.g., *Quarterly Journal of Economics*, *Management Science*, and *The American Economic Review*, in which econometrics serves as an indicator that data analytics is involved. 3) Journals in biology, medical science, and psychology are the other large groups and journals such as *Lancet*, *Proceedings of the Royal Society B*, and *Philosophical Transactions of the Royal Society B* are the bridges of these research communities. We assume that the form of data analytics used in these disciplines is bioinformatics and econometrics, but it is also reasonable to consider that expert knowledge-based recording, observations, and diagnosis of experimental data would be found. 4) It is not surprising that there was limited involvement of information technologies – the key to data science in our definition. Beyond several isolated but high-quality journals (e.g., *Communications of the ACM* and *Journal of Machine Learning Research*), the computer science community has not established a distinct or extensive research area that concentrates on policy analysis.



Note that the size of the nodes indicates their co-citation strengths.

Figure 4. Co-citation Maps for the Three Datasets

The co-citation map for Dataset 2 reveals relatively concentrated interests and research topics in these communities. A few groups actively interact, e.g., political science, public administration, and strategic management, but there are not clear terms to exactly indicate the use of information technologies in current policy analysis. In contrast, the co-citation map for Dataset 3 exactly profiles certain key sub-areas of data science, such as artificial intelligence and computer science (e.g., *Lecture Notes in Computer Science*), information systems (e.g., *MIS Quarterly*), operational research (e.g., *European Journal of Operational Research*), and information science (e.g., *Journal of the Association for Information Science and Technology*). It also links these data science-related techniques with a clear focus on actual issues (e.g., *Communications of the ACM* and *Management Science*). The interactions between methodological techniques and business and management issues found in Dataset 3 show much clearer correlations between data science and policy analysis.

In general, the willingness of data scientists to be involved in policy analysis is traceable. However, despite certain attempts for such involvement in *Nature* and *Science*, so far, the endeavors of policy analysts to exploit data analytics are relatively elusive, with their efforts being mostly dedicated to traditional approaches, such as econometrics.

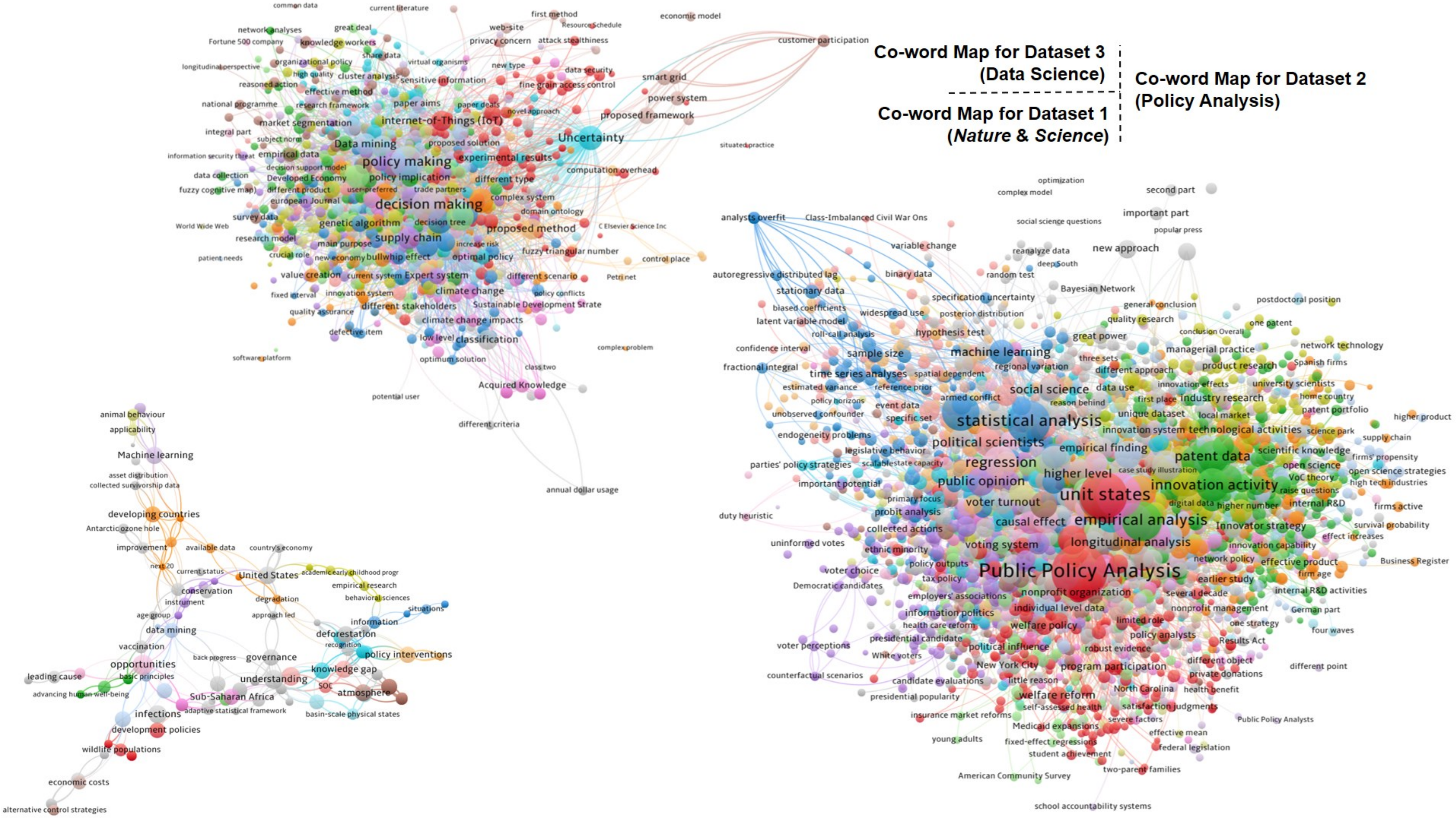
Q4: Which topics does the research involve and how do their efforts interact?

Topics are defined as a set of terms representing similar meanings, representing specific technological and research areas, and the semantic similarities between topics indicate their potential technological closeness [15]. The topics in this study consist of the terms derived from a term clumping process [51]. In brief, this process involved extracting raw terms from the data set using the natural language processing function integrated in VantagePoint. A set of thesauri and macros helped to remove meaningless or common terms (e.g., pronouns, conjunctions, and terms like “research framework”). The remaining terms were then consolidated based on stem (e.g., “co-word analysis” and “word co-occurrence analysis”). Lastly, any term that only appeared in a single article was removed. Using this method, 55, 3841, and 989 terms were identified as core terms from the three datasets, respectively. The co-word maps for the three datasets generated by VOSViewer are shown in Figure 5.

The key terms in Dataset 1 cross a broad range of disciplines as well as real-world issues: from governance to policy interventions; from individual countries (e.g., the United States) to specific regions (e.g., Sub-Saharan Africa); from populations to human well-being; and so on. Two data science-related terms appear frequently – machine learning and data mining – but beyond these sweeping concepts how data analytic techniques are used to support policy analysis in specific cases is still elusive.

The co-word map for Dataset 2 provides much more detail. Even though our definition of data science does not include econometrics, it is clear that econometrics has been used to study policy analysis through patent data and statistical analysis (such as regression) for decades. As another direction of statistical analysis, computer science-based data analytics, like machine learning, time series analyses, and Bayesian networks, are playing an increasingly active role in policy analysis. It is also surprising that compared to the limited number of data science-related journals shown in Figure 4, the interactive connections between machine learning and other terms are clearly observed in Figure 5.

Insights in the co-word map for Dataset 3 provide clues as to how data scientists have applied their data analytic approaches to support decision-making and policy-making. It might initially be appropriate for data scientists to use policy-making support rather than policy analysis to describe the applications of their methods, since, based on our definition, data scientists are highly unlikely to be able to resolve policy problems with the use of their methods. Yet their analytic results can provide supporting evidence for solutions to policy problems. What is most characteristic is that, time and again, data scientists highlight one issue hidden within policy-making data – uncertainty. They then proceed to address this data uncertainty with a series of computer science-related concepts, theories, and methods, such as expert systems, fuzzy logic, decision trees, and genetic algorithms. With these tools, data analytics perhaps offers a more feasible way to explore large-scale datasets than the common policy analysts' approach of applying econometrics to test a hypothesis.



Note that the size of nodes indicates their word co-occurrence strengths.

Figure 5. Co-word Maps for the Three Datasets

Overall, the co-word maps show that data science and policy analysis co-occur, but data scientists and policy analysts undertake their studies using their own methodological patterns and for ultimately different purposes.

Q5: How does the involvement evolve over time, and which evolutionary pathways are apt to be cutting-edge in the near future?

Aiming to track the involvement of data science in policy analysis over time, we used the scientific evolutionary pathways (SEP) method [12] to identify research topics and their predecessor-descendant relationships over time, and the 45 topics listed in Table 3 were generated. In the process of visualization, the 45 topics are 45 nodes, the weighted direct edge between two nodes indicates their predecessor-descendant relationship and their strength (i.e., cosine similarity). The nodes and edges were then visualized in Gephi [56] in Figure 6 - the color was set based on the modularity function integrated in Gephi and the layout was based on Gephi's 'ForceAtlas 2' style.

Two indicators were used to identify and evaluate core topics in a SEP: 1) term frequency-inverse document frequency (TFIDF), which highlights the distribution of key terms in involved documents. Note that old topics with a larger number of records and terms usually have a higher TFIDF score. 2) The resurgence of a "sleeping beauty" – a resurrected topic – could be due to the shortage of related theoretical and technical support studies on certain topics, and it thus could not be moved forward at that time. However, the invention of new concepts and technologies provides solutions for these unsolved issues, and those "dead" topics would then be enlightened [53]. The rise of deep learning around 2016 could be such an example. Given that, SEPs highlight two sets of topics: 1) core research areas, i.e., topics that are consistently studied without pause (and usually have high TFIDF scores); and 2) cutting-edge areas, i.e., either new topics at the end of a path or resurrected topics (both of which should have high TFIDF scores).

Table 3. Topics in the Scientific Evolutionary Pathways

ID	*Topic*	*Parent*	*TFIDF*	*Born*	*Death*	*Survival*	*Resurgence*
1	public policy	n/a	0.279	2000	n/a	20	-

2	decision-making	1	0.122	2001	2002	2	-
4	technology impacts	1	0.085	2001	2009	9	-
5	Bayesian analysis	2	**0.166**	2002	n/a	12	Yes
9	sustained development	1	0.065	2002	2006	5	-
11	measurements	5	0.089	2003	2006	3	Yes
14	social scientists	1	0.001	2003	2003	1	-
15	dependent variable	5	0.077	2003	2009	7	-
17	statistical analysis	11	**0.182**	2004	n/a	15	Yes
18	public administration	1	0.087	2004	2017	14	-
19	political economy	1	0.063	2004	2009	6	-
20	software engineering	1	0.014	2004	2008	5	-
22	technology system	1	**0.111**	2005	2013	5	Yes
24	macroeconomic performance	1	**0.171**	2005	n/a	12	Yes
26	privacy concern	1	0.030	2006	2010	5	-
27	software industry	9	0.058	2007	2008	2	-
29	political science	9	0.126	2007	n/a	13	-
32	Probability	11	0.130	2007	n/a	13	-
33	data mining	17	0.089	2008	2013	6	-
35	empiric evidence	24	**0.125**	2008	n/a	8	Yes
36	program implementation	1	0.101	2008	n/a	12	-
38	systematic approach	27	0.036	2009	2009	1	-
40	collaboration	29	0.087	2009	2016	3	Yes
44	empirical analysis	22	0.118	2009	n/a	11	-

45	authoritarian regime	18	0.012	2009	2010	2	-
46	observed data	38	0.044	2010	2012	3	-
48	prediction model	32	0.091	2010	n/a	10	-
50	Uncertainty	32	0.092	2010	n/a	10	-
51	developed country	19	0.114	2010	n/a	10	-
54	knowledge source	4	0.023	2010	2010	1	-
55	innovation performance	54	0.086	2011	2018	8	-
59	group interest	33	0.072	2011	n/a	9	-
62	efficiency	9	0.055	2011	n/a	9	-
63	policy-based systems	24	0.151	2011	n/a	9	-
64	scenario analysis	35	0.028	2012	2012	1	-
65	political science data	24	0.080	2012	n/a	8	-
70	financial decision-making	46	0.083	2013	n/a	7	-
72	labor market outcomes	1	0.039	2013	2018	6	-
73	knowledge management	59	0.046	2013	n/a	7	-
75	sensitive analysis	17	0.105	2015	n/a	5	-
78	big data	33	0.010	2016	2018	2	Yes
82	machine learning	48	0.044	2017	n/a	3	-
89	geospatial approach	44	0.006	2017	2018	2	-
90	social network	22	0.027	2017	n/a	3	-
93	performance evaluation	45	0.039	2018	n/a	2	-

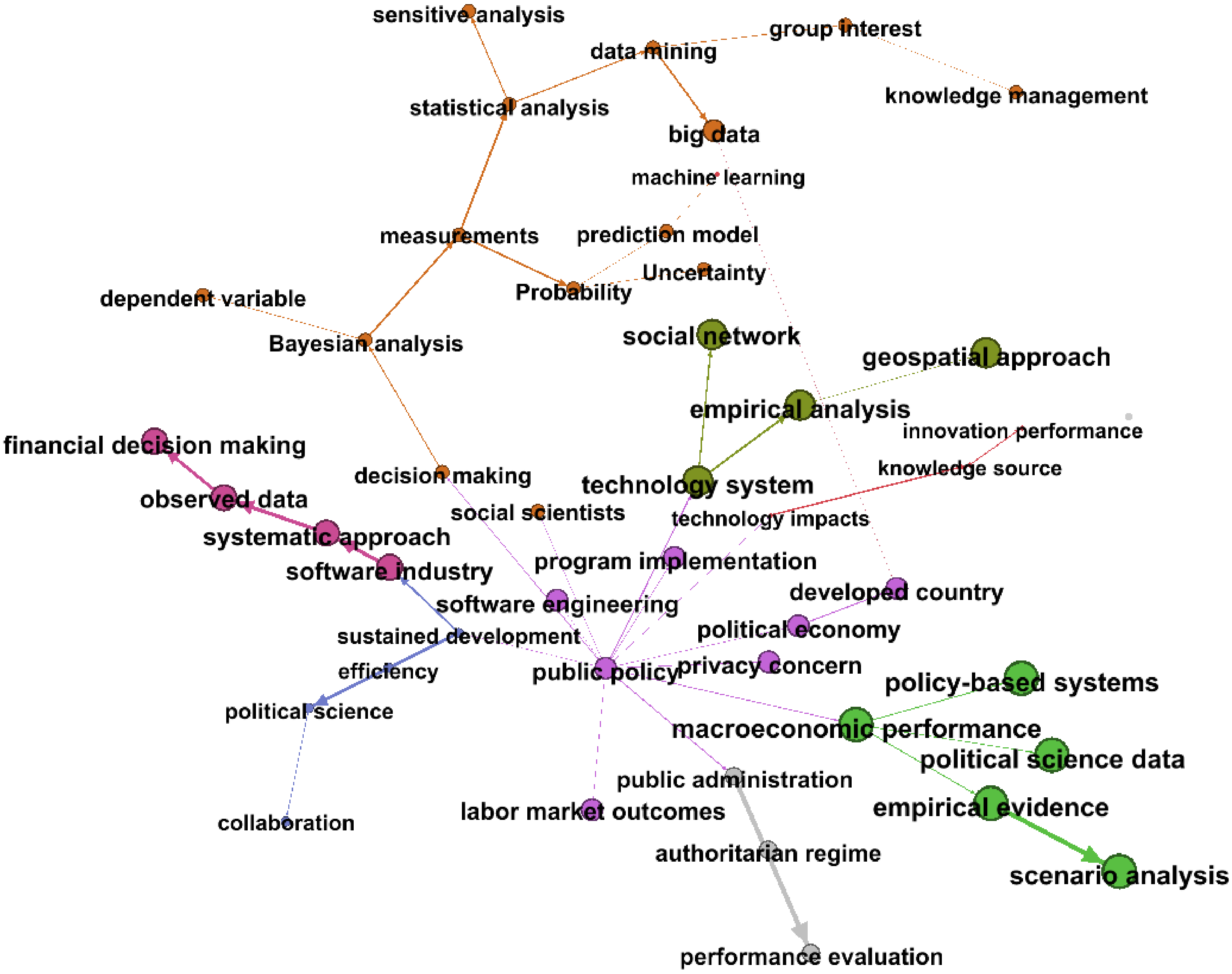

sensitive analysis
group interest
data mining
knowledge management
statistical analysis
big data
machine learning
measurements
prediction model
Uncertainty
Probability
dependent variable
Bayesian analysis
social network
geospatial approach
empirical analysis
innovation performance
financial decision making
knowledge source
decision making
technology system
observed data
social scientists
technology impacts
systematic approach
program implementation
software industry
developed country
software engineering
political economy
sustained development
privacy concern
efficiency
public policy
policy-based systems
political science
macroeconomic performance
political science data
public administration
empirical evidence
collaboration
labor market outcomes
authoritarian regime
scenario analysis
performance evaluation

Figure 6. Scientific Evolutionary Pathways of the Involvement of Data Science in Policy Analysis

From a bird's-eye view, Figure 6 shows two relatively isolated pathways – data science-based decision support (the orange chain and perhaps the pink chain) and policy analysis. That is to say, data science and policy analysis are relatively independent in their own pathways of development. These pathways are discussed below from the perspective of core and cutting-edge areas.

- Data science – decision support systems could be considered the backbone of this pathway, particularly their application to the financial sector. Information technologies, such as data mining and machine learning, have been used to enhance decision-making support, and related research frontiers include, not only big data, but also prediction models and uncertainty issues.
- Policy analysis – this pathway consists of a broad range of issues in political science and public administration. Two key areas are identified: 1) investigations into technology systems, impacts, and performance; and 2) political economy combined with multiple entities, such as organizations, countries, and regions. Considering topics at the end of related routes, social networks appear to be a new direction that involves data science and policy analysis, and policy-based systems might be another way to systematically integrate empirical scenarios and data analytic techniques in either econometrics or information technologies.

Discussion and Conclusions

In this paper, we presented an intelligent bibliometric framework for investigating the involvement of data science in policy analysis. We posed, explored, and answered five research questions: 1) How much effort has been expended in the involvement, and how has that effort changed over time? 2) Who are the key players leading the involvement, and what are their geographical distributions and collaboration patterns? 3) Which academic journals publish this research, and what are their citation behaviors? 4) Which topics does the research involve and how do their efforts interact? 5) How does the involvement evolve over time, and which evolutionary pathways are apt to be cutting-edge in the near future? We constructed and analyzed three datasets: a set of articles published in *Nature* and *Science* to represent the cutting-edge of research in data science and policy analysis; and two sets of articles published in top-level journals in data science and policy analysis, respectively.

Key findings: Intersecting lines between data science and policy analysis

From our investigation, we find intersecting lines between data science and policy analysis in terms of research content and geographical distribution.

(1) Interactions in Research Content

The articles published in *Nature* and *Science* suggest that the involvement of data analytics in policy analysis is becoming a cutting-edge direction in multiple disciplines, including computer science, information science, business, and management, etc.

Insights from Datasets 2 and 3 endorse claims that econometric approaches dominate current policy analysis, and information technologies are mostly used for supplementary support or explorative evidence. Comparably, data scientists hold strong interests in implementing novel information technologies to solve real-world issues, and support for policy-making is one such practice. Intriguingly, the co-citation maps at journal level consistently revealed the role of the bibliometric community in bridging data science with policy analysis, due to their advantages in both fields.

Two articles published in *Science* provide details on such interactions. One article [57] was published in 2015, in which machine learning techniques were developed to predict poverty and wealth from mobile phone metadata. The procedures were authored by a group of researchers with expertise in computer science. Another article [58], published in January 2018, showcases how machine learning and optimal matching techniques can be used to investigate refugee integration issues. Importantly, all the authors are policy analysts. These two articles offer support for our finding that data scientists are moving ahead further than policy analysts. They are equipping themselves with advanced data analytic techniques and, as such, seem to be finding it easier for realize applications for their expertise in policy analysis.

Based on the investigation, as well as the discussion given by Athey (2017) [6], one highlight for enhancing the involvement of data science in policy analysis moving forward is identified. That is the development of new information technologies (e.g., artificial intelligence techniques) is required to understand the approaches and practices from decades of multidisciplinary research, emphasizing empirical evidence for informing policy analysis.

(2) Geographical Distribution

We also found interesting insights from our geographical analysis that fall on two sides of the coin. Western countries (e.g., the US and European countries) are spearheading policy analysis through data science techniques. However, Anglo/English-speaking countries have the benefit of a similar and valued cultural mindset to the top journals in their field and distinct advantages in terms of language and communication requirements. For researchers in the West, an understanding of and their concerns for real-world political issues and policies are the main motivations for conducting research, and data science aids them in this cause. On the other side of the coin, Asian countries lack a comprehensive understanding of the political issues facing the West. They have cultural and language barriers to overcome, and a path of lesser resistance seems to be leveraging their strengths in computer science. Examining novel information technologies and computational models as ways to solve real-world issues, therefore, dominates published research. Decision support systems seem to have been a particularly effective way to achieve this goal.

Practical significance and possible applications

This study empirically investigated the interactions between data science and policy analysis in the past two decades. The overview, considering answers to Questions 1-4, systematically profiled the status of such interactions by discovering key players, research communities, technological landscapes, and their relationships, and the evolutionary pathways of research topics for Question 5 created exploratory thoughts on understanding why the interaction happened and how far it goes.

At a macro level, such insights provide a bird's-eye view for federal governments and research agencies to evaluate research disciplines (e.g., measuring research outcomes and locating the status of an entity's research strength in the globe), foresee research frontiers (e.g., tracking leading players' research interests and identifying emerging topics), and allocate research funding (e.g., emphasizing technical breakthroughs and strengthening research advantages).

At a micro level, this study could be a reference for both research communities to identify research interests (e.g., emerging research issues in policy analysis) and feasible solutions (e.g., effective data analytic approaches for policy analysis). In fact, this study noticed that despite a relatively active interaction between the two disciplines, such interaction still stays at an early stage - i.e., collaborations between policy analysts and data scientists are limited and the co-citations between journals in the two disciplines are rare. Thus, this investigation is expected to bridge the two communities and further enhance such interactions, deepening understanding of each other.

Limitations and future directions

The limitations and future directions of this study are discussed as follows. 1) Intelligent bibliometrics could and should be expanded, involving intelligent information technologies and bibliometrics from many diverse corners -- e.g., citation analysis combined with scientific evolutionary pathways to comprehensively trace technological change with citation linkages, topic analysis and topic models. 2) These findings should be considered as exploratory. Econometric models and further systematic examinations should be undertaken to support, refute, or contextualize our observations.

Acknowledgments

An early version of this paper has been presented at the 2018 Portland International Conference on Management of Engineering and Technology (PICMET). This work is partially supported by the Australian Research Council under Discovery Early Career Researcher Award DE190100994 and the United States National Science Foundation Award #1759960.

References

[1] W. N. Dunn, *Public policy analysis*. Routledge, 2015.

[2] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data: The management revolution.," *Harvard Business Review,* vol. 90, no. 10, pp. 61-67, 2012.

[3] G. A. Gorry and M. S. Scott Morton, "A framework for management information systems," 1971.

[4] A. L. Porter and M. J. Detampel, "Technology opportunities analysis," *Technological Forecasting and Social Change,* vol. 49, no. 3, pp. 237-255, 1995.

[5] M. Karki, "Patent citation analysis: A policy analysis tool," *World Patent Information,* vol. 19, no. 4, pp. 269-272, 1997.

[6] S. Athey, "Beyond prediction: Using big data for policy problems," *Science,* vol. 355, no. 6324, pp. 483-485, 2017.

[7] W. Hood and C. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics,* vol. 52, no. 2, pp. 291-314, 2001.

[8] Y. Guo, L. Huang, and A. L. Porter, "The research profiling method applied to nano ‐ enhanced, thin ‐ film solar cells," *R&d Management,* vol. 40, no. 2, pp. 195-208, 2010.

[9] D. K. Robinson, L. Huang, Y. Guo, and A. L. Porter, "Forecasting Innovation Pathways (FIP) for new and emerging science and technologies," *Technological Forecasting and Social Change,* vol. 80, no. 2, pp. 267-285, 2013.

[10] Y. Zhou, F. Dong, D. Kong, Y. J. T. F. Liu, and S. Change, "Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies," vol. 144, pp. 205-220, 2019.

[11] E. Kim, Y. Cho, and W. J. S. Kim, "Dynamic patterns of technological convergence in printed electronics technologies: patent citation network," vol. 98, no. 2, pp. 975-998, 2014.

[12] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Science evolutionary pathways: Identifying and visualizing relationships for scientific topics," *Journal of the Association for Information Science and Technology,* vol. 68, no. 8, pp. 1925-1939, 2017.

[13] Y. Guo, L. Huang, and A. L. Porter, "Profiling research patterns for a new and emerging science and technology: dye-sensitized solar cells," *2009 Atlanta Conference on Science and Innovation Policy,* pp. 1-7, 2009.

[14] L. Huang, Y. Zhang, Y. Guo, D. Zhu, and A. L. Porter, "Four dimensional science and technology planning: A new approach based on bibliometrics and technology roadmapping," *Technological Forecasting and Social Change,* vol. 81, pp. 39-48, 2014.

[15] Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, and J. Lu, "Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research," *Technological Forecasting and Social Change,* vol. 105, pp. 179-191, 2016.

[16] H. Small, K. W. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Research Policy,* vol. 43, no. 8, pp. 1450-1467, 2014.

[17] C. Chen, Z. Hu, S. Liu, and H. Tseng, "Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace," *Expert Opinion on Biological Therapy,* vol. 12, no. 5, pp. 593-608, 2012.

[18] Y. Huang *et al.*, "A hybrid method to trace technology evolution pathways: a case study of 3D printing," *Scientometrics,* vol. 111, no. 1, pp. 185-204, 2017.

[19] I. Rafols, A. L. Porter, and L. Leydesdorff, "Science overlay maps: A new tool for research policy and library management," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 9, pp. 1871-1887, 2010.

[20] R. Klavans and K. W. Boyack, "Toward a consensus map of science," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 3, pp. 455-476, 2009.

[21] A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human ‐ assigned subject classification," *Journal of the Association for Information Science and Technology,* vol. 67, no. 19, pp. 2464 ‐ 2476, 2016.

[22] Y. Zhang, Y. Guo, X. Wang, D. Zhu, and A. L. Porter, "A hybrid visualisation model for technology roadmapping: Bibliometrics, qualitative methodology and empirical study," *Technology Analysis & Strategic Management,* vol. 25, no. 6, pp. 707-724, 2013.

[23] L. Waltman, N. J. van Eck, and E. C. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *Journal of Informetrics,* vol. 4, no. 4, pp. 629-635, 2010.

[24] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology,* vol. 57, no. 3, pp. 359-377, 2006.

[25] L. Kay, N. Newman, J. Youtie, A. L. Porter, and I. Rafols, "Patent overlay mapping: Visualizing technological distance," *Journal of the Association for Information Science and Technology,* vol. 65, no. 12, pp. 2432-2443, 2014.

[26] W. Ding and C. Chen, "Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods," *Journal of the Association for Information Science and Technology,* vol. 65, no. 10, pp. 2084-2097, 2014.

[27] C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics,* vol. 100, no. 3, pp. 767-786, 2014.

[28] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *Journal of informetrics,* vol. 5, no. 1, pp. 187-203, 2011.

[29] E. Yan, Y. Ding, and Q. Zhu, "Mapping library and information science in China: A coauthorship network analysis," *Scientometrics,* vol. 83, no. 1, pp. 115-131, 2009.

[30] T. Tang and D. Popp, "The learning process and technological change in wind power: Evidence from China's CDM wind projects," *Journal of Policy Analysis and Management,* vol. 35, no. 1, pp. 195-222, 2016.

[31] Y. Zhang *et al.*, "Does deep learning help topic extraction? A kernel k-means clustering method with word embedding," *Journal of Informetrics,* vol. 12, no. 4, pp. 1099-1117, 2018.

[32] R. J. Funk and J. Owen-Smith, "A dynamic network measure of technological change," *Management Science,* vol. 63, no. 3, pp. 791-817, 2016.

[33] S. Lee, B. Yoon, and Y. Park, "An approach to discovering new technology opportunities: Keyword-based patent map approach," *Technovation,* vol. 29, no. 6, pp. 481-497, 2009.

[34] X. Li, Y. Zhou, L. Xue, and L. J. I. Huang, "Roadmapping for industrial emergence and innovation gaps to catch-up: a patent-based analysis of OLED industry in China," vol. 72, no. 1/2/3, pp. 105-143, 2016.

[35] Y. Zhou, X. Li, R. Lema, F. J. S. Urban, and P. Policy, "Comparing the knowledge bases of wind turbine firms in Asia and Europe: Patent trajectories, networks, and globalisation," vol. 43, no. 4, pp. 476-491, 2015.

[36] T. U. Daim, B.-S. Yoon, J. Lindenberg, R. Grizzi, J. Estep, and T. Oliver, "Strategic roadmapping of robotics technologies for the power industry: A multicriteria technology assessment," *Technological Forecasting and Social Change,* 2017.

[37] Y. Geum, H. Lee, Y. Lee, and Y. Park, "Development of data-driven technology roadmap considering dependency: An ARM-based technology roadmapping," *Technological Forecasting and Social Change,* vol. 91, pp. 264-279, 2015.

[38] Y. Zhang, X. Zhou, A. L. Porter, J. M. V. Gomila, and A. Yan, "Triple Helix innovation in China’s dye-sensitized solar cell industry: Hybrid methods with semantic TRIZ and technology roadmapping," *Scientometrics,* vol. 99, no. 1, pp. 55-75, 2014.

[39] X. Li, Y. Zhou, L. Xue, L. J. T. F. Huang, and S. Change, "Integrating bibliometrics and roadmapping methods: A case of dye-sensitized solar cell technology-based industry in China," vol. 97, pp. 205-222, 2015.

[40] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," *Technological Forecasting and Social Change,* vol. 73, no. 8, pp. 981-1012, 2006.

[41] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?," *Research Policy,* vol. 44, no. 10, pp. 1827-1843, 2015.

[42] P. Akhavan, N. A. Ebrahim, M. A. Fetrati, and A. J. S. Pezeshkan, "Major trends in knowledge management research: a bibliometric study," vol. 107, no. 3, pp. 1249-1264, 2016.

[43] J. Guan and N. J. R. P. Ma, "China's emerging presence in nanoscience and nanotechnology: A comparative bibliometric study of several nanoscience 'giants'," vol. 36, no. 6, pp. 880-886, 2007.

[44] H. D. Lasswell, *A pre-view of policy sciences*. Elsevier publishing company, 1971.

[45] F. Provost and T. J. B. d. Fawcett, "Data science and its relationship to big data and data-driven decision making," vol. 1, no. 1, pp. 51-59, 2013.

[46] T. Althoff, J. L. Hicks, A. C. King, S. L. Delp, and J. Leskovec, "Large-scale physical activity data reveal worldwide activity inequality," *Nature,* vol. 547, no. 7663, p. 336, 2017.

[47] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science,* vol. 353, no. 6301, pp. 790-794, 2016.

[48] M. M. Kessler, "Bibliographic coupling between scientific papers," *American Documentation,* vol. 14, no. 1, pp. 10-25, 1963.

[49] C. Calero-Medina and E. C. Noyons, "Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field," *Journal of Informetrics,* vol. 2, no. 4, pp. 272-279, 2008.

[50] M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," *Social Science Information,* vol. 2, no. 22, pp. 191-235, 1983.

[51] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, "“Term clumping” for technical intelligence: A case study on dye-sensitized solar cells," *Technological Forecasting and Social Change,* vol. 85, pp. 26-39, 2014.

[52] Y. Zhang, X. Wang, L. Huang, G. Zhang, and J. Lu, "Predicting the dynamics of scientific activities: A diffusion-based network analytic methodology," in *2018 Annual Meeting of the Association for Information Science and Technology*, Vancouver, Canada, 2018.

[53] A. F. van Raan, "Sleeping beauties in science," *Scientometrics,* vol. 59, no. 3, pp. 467-472, 2004.

[54] L. Huang, Y. Zhu, Y. Zhang, X. Zhou, and X. Jia, "A Link Prediction-Based Method for Identifying Potential Cooperation Partners: A Case Study on Four Journals of Informetrics," in *2018 Portland International Conference on Management of Engineering and Technology (PICMET)*, 2018, pp. 1-6: IEEE.

[55] X. Zhou, L. Huang, Y. Zhang, and M. Yu, "A hybrid approach to detecting technological recombination based on text mining and patent network analysis," *Scientometrics,* pp. 1-39.

[56] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," *Proceedings of International AAAI Conference on Web and Social Media,* vol. 8, pp. 361-362, 2009.

[57] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science,* vol. 350, no. 6264, pp. 1073-1076, 2015.

[58] K. Bansak *et al.*, "Improving refugee integration through data-driven algorithmic assignment," *Science,* vol. 359, no. 6373, pp. 325-329, 2018.

Bio

Dr **Yi Zhang** is a Lecturer (tenure) at the Centre for Artificial Intelligence, University of Technology Sydney (UTS), Australia, with dual PhD degrees (i.e., Management Science & Engineering and Software Engineering). His research interests align with bibliometrics, text analytics, and innovation & technology management, and has authored/co-authored more than 60 publications in related areas. He serves diverse roles (e.g., Associate Editor, Editorial Board Member, and Managing Guest Editor) in certain reputable international journals. He is a Member of the Advisory Board for ICSR, and Member of ASIS&T, ISSI, and IEEE. He received a 2019 Discover Early Career Researcher Award granted by the Australian Research Council.

Dr **Alan Porter** is Director of R&D for Search Technology, Inc., Norcross, GA (producers of VantagePoint and Derwent Data Analyzer software). He is also Professor Emeritus of Industrial & Systems Engineering, and of Public Policy, at Georgia Tech, where he is Co-director of the Program in Science, Technology & Innovation Policy (STIP). Dr. Porter is author or co-author of some 240 articles and books, including Tech Mining (Wiley, 2005) and Forecasting and Management of Technology (Wiley, 2011). He co-founded the International Association for Impact Assessment and later served as president. Research interests key on developing indicators of technological emergence (with current NSF support). Publications are available at: http://www.researchgate.net/profile/Alan_Porter4.

Professor **Scott Cunningham** worked at the Department of Multi-Actor Systems at the Delft University of Technology. There he engaged in a programme of research and teaching on the operation of socio-technical systems given the diverse needs, capabilities, and interests of system operators. The work encompassed diverse engineering systems including transport, logistics, telecommunications, and energy. The work embraced both human actors, with decision capability, as well as computational agents. The work was validated by research grants from the Dutch science foundation, the Next Generation Infrastructures Foundation, ProRail the national rail network provider, and the European FP7 and Horizon 2020 programmes.

Before joining TU Delft as a staff member, Scott worked as a software engineer for multiple Silicon Valley / Bay Area start-up companies. These companies used database and message bus technologies to aggregate consumer preferences to deliver automatic pricing capabilities, as well as channel, demand and relationship management services. Now at the University of Strathclyde, Scott holds a chair of Urban Technology Policy in the School of Government and Public Policy.

Denise Chiavetta is a Senior Consultant at Search Technology, Inc., Norcross, GA. With a BS in Electrical Engineering and an MS in Studies of the Future, she brings expertise in organizational applications of technology foresight developed over 20 years as a consultant as well as a professional inside Fortune 100 companies and government agencies.

Nils Newman is the President of Search Technology in Atlanta, Georgia, USA. For over twenty years, he has worked on the development of analytical tools to assist in the management of technology. His work focuses on the use of bibliographic and patent information in research evaluation, competitive intelligence, and strategic planning. Mr. Newman has a Bachelor of Mechanical Engineering and an MS in Technology and Science Policy from the Georgia Institute of Technology. In his spare time, he continues to pursue a PhD in Economics from UNU-MERIT at the University of Maastricht in the Netherlands studying the economics of technical change.