# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

Can Azkan, Markus Spiekermann, Henry Goecke

"*Data is only as valuable as the decision it enables.*"
Ion Stoica
Computer scientist

Data-driven business models arise in different social and industrial sectors, while new sensors and devices are breaking down the barriers for disruptive ideas and digitally transforming established solutions. This paper aims at providing insights about emerging topics in the data economy that are related to companies' innovation potential. The paper uses text mining supported by systematic literature review to automatize the extraction and analysis of beneficial insights for both scientists and practitioners that would not be possible by a manual literature review. By doing so, we were able to analyze 860 scientific publications resulting in an overview of the research field of data economy and innovation. Nine clusters and their key topics are identified, analyzed as well as visualized, as we uncover research streams in the paper.

## Introduction

Due to rapid technological and organizational progress in digitalization, a diverse ecosystem of innovative technologies, platforms, and digital market players has emerged, leading to what we now call the "data economy". One characteristic of this data economy is the huge amount of available data, which is often referred to as the "big data" paradigm. There are many sources available in scientific and practitioner communities, which need to be analyzed in order to stay informed, and thus capable of acting. The volume of sources means that a complete manual analysis is time consuming. Information overload leads to challenges for scientists as well as practitioners to identify and track the main topics in which innovation might take place. The challenge, however, is a prerequisite for achieving sustainable competitive advantage, due to volatile market changes and disruptive innovation approaches.

This paper aims at facing this challenge and enables an automatized, repeatable way to identify topics of interest and track the fields of innovation as discussed in published research literature. By systematically reviewing scientific publications, major research streams and their (sub-)topics are revealed. This will help scientists and practitioners to identify potentials for innovation and give guidance regarding which topics could be of future interest for scientists on the one side as well as practitioners on the other side. Given the volume of publications, this paper uses a literature review and text mining approach to analyze keywords and abstracts of scientific publications in the context of the data economy in terms of their relevance, relation, and potential for automated innovation. In this paper, we provide the following results: on the one hand, we show what a text mining supported systematic literature review could look like. This approach can be easily adapted to analyze other research fields and topics. On the other hand, we provide content-related insights in the field of data economy and innovation.

## Background Information

*Data Economy*

Organizations invest a lot in digitalization programs and projects aiming to benefit from data economics. The discussion around digital business as "a business model whose underlying business logic deliberately acknowledges one or more characteristics of digitalization and aims to take advantage of them" (Otto et al., 2015), shows the growing importance of data within enterprises business (Moody & Walsh, 1999). Digitalization and advancing an organization's business model in this direction requires considering the opportunities and challenges that data and information

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

bring to value creation. Business models in the digital economy (Otto, 2015; Zimmerman, 2000) are characterized by developing products into hybrid or purely digital services. The close integration of digital and physical products in combination with a vast amount of internally and externally available data enables data-driven service offerings for traditional products, as well as innovations to add more value to tangible products (Yoo et al., 2010). However, what does the term “data economy” mean exactly? Despite, or perhaps even because of the high attention given to it, a common understanding of the term “data economy” is still missing.

Nevertheless, a number of definitions have formed in practice, which are presented as follows:
According to the German Association for the Digital Economy (BVDW), data economy deals with the monetization of information based on acquired data, which is transformed into valuable information using an algorithm, and then made accessible on the basis of business management functions. A data economy can be operated as its own business model or it can support, modify, or replace existing value creation models by increasing digitalization (German Association for the Digital Economy, 2018).

By the European Commission’s definition, data economy measures the overall impacts of the data market on the economy as a whole. It involves the generation, collection, storage, processing, distribution, analysis elaboration, delivery, and exploitation of data enabled by digital technologies (European Commission, 2019).

A study by Digital Reality, a worldwide leader in building data centers, defines data economy as the financial and economic value created by the storage, retrieval and analysis —using sophisticated software and other tools — of large volumes of business and organizational data at very high speeds (so-called ´big data´). This can involve, for example, realizing improved operational efficiency or implementing improved strategic decisions (Digital Reality, 2018).

This paper therefore defines “data economy” as an umbrella term, which includes digital business models independent of a particular industry, for example, data products and services, digital technologies, data value chains, and their technical implications for data creation, processing, provision, and use to gain benefits for an organization.

*Innovation*
Schumpeter's (1912) work on economic development theory, in which he describes an innovation as a "new combination" that asserts itself on the market and establishes a "creative redesign", is regarded as fundamental for introducing the concept of innovation. Numerous authors and scientists have taken up and interpreted innovation differently (Schumpeter, 1912). The following definitions reveal diverse understandings of the concept of innovation.

Barnett argues that innovation is a qualitative differentiation from existing ideas or objects. The distance or the extent of novelty is the decisive factor to distinguish between "non-innovation" and innovation (Barnett, 1953).

Many authors take up the characteristic of novelty in their definition of innovation, while nevertheless interpreting novelty in decisively different ways. Thus, Vedin sees innovation in the first application of the new idea, method, or use of a novel object (Vedin, 1980).

In his work on innovation diffusion theory, Rogers also takes up this approach, but adds a perspective that defines the concept more clearly. He thus interprets that something new only leads to an innovation if the adapting user perceives it in the same way (Rogers, 1983). This definition implies that (early) users adapt an innovation, which is to be understood as a first step in the later diffusion process. In addition to novelty, the concept of innovation is here linked with adaptation, that is, the application of a novel idea, method, or use of a new product by users. Following this definition, an innovation can be understood as a novel idea or invention that eventually finds commercial application. Zawislak et al. also define innovation as the application of knowledge to generate technical or organizational changes capable of offering advantages to the firm that accomplishes them (Zawislak et al., 2008).

Francis and Bessant view innovation from the perspective of the change that comes with innovation (Francis & Bessant, 2005). Regarding this view, Bessant and Tidd distinguish four categories of innovation. “Product innovation” refers to changes in the things (product/services) an organization offers. “Process innovation” implies changes in the way in which things are created and delivered. “Position innovation” refers to changes in the context in which things are introduced, while “paradigm innovation” describes changes in the underlying mental models that frame what the

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

organization does. These changes always lead to something new that creates some kind of value (Bessant & Tidd, 2007). Van der Kooij finds a generic definition for innovation and highlights the aspect of change as well. Here, an innovation is a change in the function of a system (product, process, organization, or society) that has a stepwise character. In short, it is the result of a process of human activity. The steps could be small, incremental, or large, and hence result in discontinuities (Van der Kooij, 2017).

An innovation is thus created by combining the two characteristics of novelty and use, as defined by Ahmed and Shepherd (2010). This paper follows the Ahmed and Shepherd's definition and consider innovation as a combination of something new that using or applying brings a change to the status quo.

**Research Design**

This paper adopts the methodology of systematic literature review (SLR) (Kitchenham, 2004, Figure 1). The SLR consists of three phases: planning, conducting, and reporting. Within the first phase, "planning the review", the goal is to create a basic framework and design the content arrangement. This involves identifying the need for a review, specifying the addressed research questions, and developing a review protocol for controlling the review. "Conducting the review" in the second phase means executing the review protocol designed in the planning phase, which includes the creation of a dataset. This begins with the selection of suitable publications as a first step, quality assessment and cleaning as a second step, and data extraction as a final step. The third and last phase, "Reporting the review", concludes with results that answer the predefined research questions (Kitchenham & Charters, 2007).

To obtain valid results it is important to follow a systematic search strategy while doing a literature review. This can be done by defining the objectives and formulating specific research questions to be answered by carrying out the review. The research questions addressed by this article are derived from the objectives mentioned in the introduction. Our research answers the following research question (RQ):

Which subject areas are relevant in the context of data economy innovation and what are the major research streams and (sub-) topics?

The first step to conduct a phase of the SLR is the study selection. As a first step, we focused on Elsevier's Scopus database as a source for exploring peer-reviewed publications. Scopus offers easy access for meta-data on publications and has one of the largest databases for scientific publications with over 70 million publications (https://www.elsevier.com/solutions/scopus). For the second step, we defined suitable keywords to meet the objectives of our review and to answer the above research questions. We used the keywords "digital economy", "data economy", "digital business model", "data driven business model", "digital business", "digital platforms", "data technologies", "digital disruption", and "digital transformation". These keywords, chained

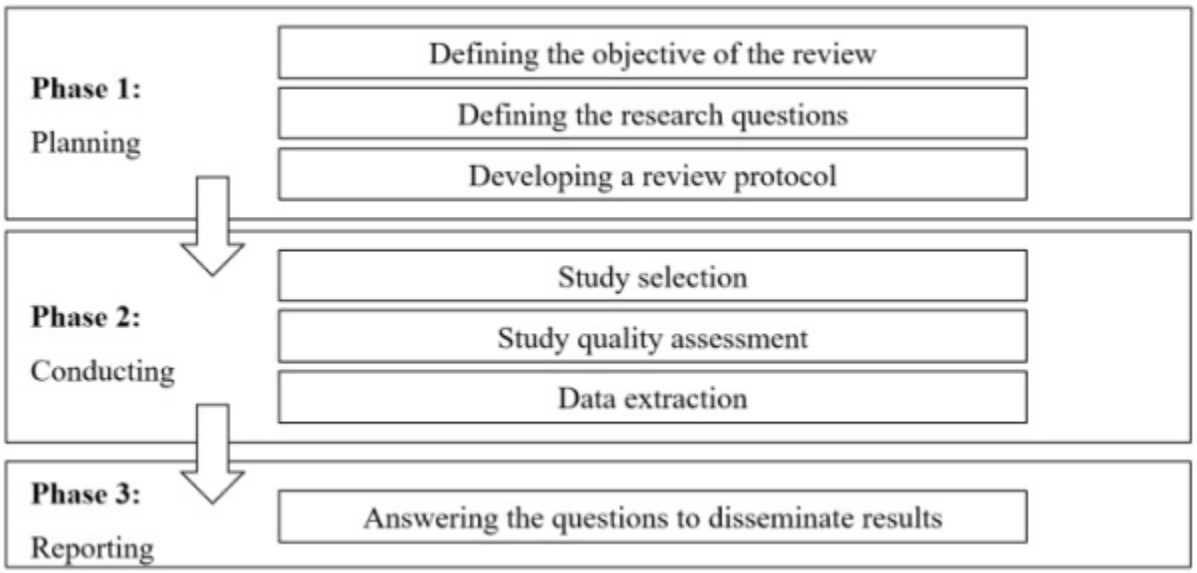


**Figure 1.** SLR process phases according to Kitchenham and Charters (2007)

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

```
TITLE-ABS-KEY ("digital economy" OR "data economy" OR "digital business model" OR
"data driven business model" OR "digital business" OR "digital platforms" OR "data
technologies" OR "digital disruption" OR "digital transformation") AND TITLE-ABS-KEY
("Innovation") AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar"))
AND (LIMIT-TO (LANGUAGE, "English" ))
```

**Figure 2.** Search Query from Scopus

as or-conditions, set the basis for retrieving appropriate publications for our review. Furthermore, we needed to ensure a connection to innovation. For this reason, we added the keyword "innovation" as a mandatory condition in the title, abstract, or keyword of the publication, and chained this as a prerequisite regardless of all the other keywords, that is, where that particular term has to be matched. With this combination of keywords, we ensured a focus on publications in the area of data economy and innovation. Following this approach, we were able to retrieve 1,163 publications as the foundational data set.

For the second step, we had to ensure the quality of our data set, and therefore combined the results with different filters and inclusion criteria, in order to gain a higher level of quality. In the second stage, 908 publications were returned, after limiting the result set to journal articles and conference papers. The third stage included only articles published in English, which returned 863 articles. Stage four excluded another three articles due to missing author names. We also consciously decided not to exclude subject areas in Scopus in order to cover a wide range of research. The final search string, as the result of combining our keyword search together with the limitation criteria, is depicted in Figure 2.

The final limitation on stage five was performed outside the Scopus search engine. We used Scopus' export functionality to export a BibTex formatted file of the search results, including the fields: Author (a), Document title (b), Year (c), Source title (d), volume, issue, pages (e), Citation count (f), Source & document type (g), DOI (h), Affiliations (i), Language (j), Abstract (k), Author keywords (l), and Index keywords (m). A python script was implemented to extract the information, which was exported in BibTex format within a relational database system in order to have a better structure for further analyses of relations. The script systematically loops through the BibTex file and stores article information, as listed above, for each entry.

In order to focus our analysis on innovation topics within the data economy, we limited our dataset by searching for specific words within the articles' abstracts, and excluded all articles that did not include these terms. For filtering, we choose the words (a) problem, (b) challenge, (c) demand, (d) requirement, (e) obstacle, (f) limit, (g) barrier, and (h) necessity. We argue that these terms, related to challenges and obstacles, within the abstracts enables through filtering the identification of novel approaches and applications to a specific problem. It was deliberately decided not to do a full-text analysis of the publications for two reasons:

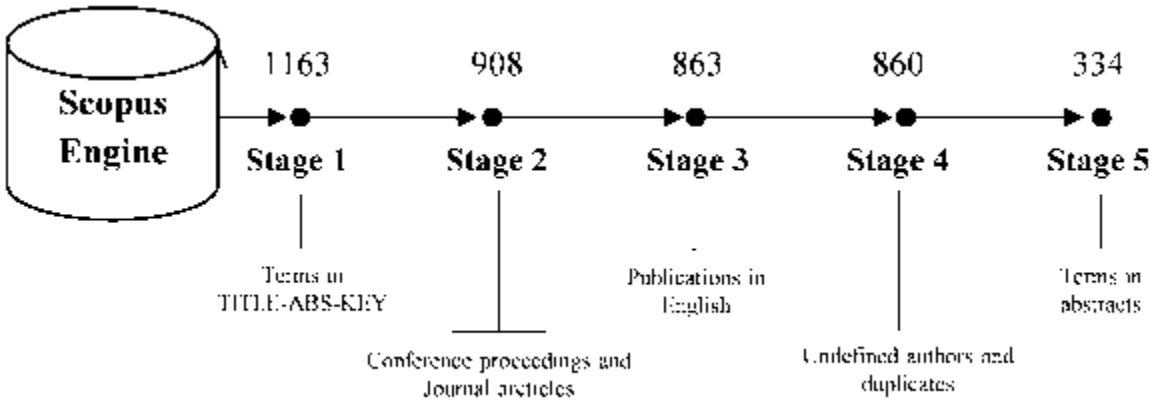


**Figure 3.** Study selection on the Scopus Engine

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

First, information density is highest in the abstract (Scheumie et al., 2004). Second, access to scientific full-text content for text mining is difficult due copyright and licensing reasons (O'Mara-Eves et al., 2015). Based on our approach, the final number of articles, as result of stage five, returned 334 documents. The full study selection and cleaning process is depicted in Figure 3.

In order to answer the specified research questions, we analyzed the abstracts and topic for data extraction. To get first insights into the data set, we used available general meta information. For example, we considered the year of publication in the area of data economy and innovation. In addition, in co-authored papers, author affiliations were analyzed, as well as country of work, in order to depict where research on the focus topic is done. From a content perspective, we took subject areas from different sources within our data set to get a distribution overview. For our main research, we focused on annotated keywords , since they can serve to help articulate a highly concise summary of a document (Siddiqi and Sharan, 2015). Within the available datasets, the keywords exported from Scopus database differed between indexed and author keywords. While author keywords are exempt from semantics rules and annotated directly by the authors, indexed keywords are assigned by Scopus using a taxonomy to form a semantic system and organize the platforms entries. Using this system enables a more consistent analysis through better comparability between different keywords. For further analyses, we used indexed keywords only.

To answer the specified research questions, we analyzed the relations between different keywords and formed clusters of different topics and sub-topics. The assigned relations between keywords were done by creating a correlation matrix. We looped through different keywords and selected all publications containing a specific keyword. After that we mapped all other keywords assigned to these publications and linked these relations within our database.

## Findings

*Summary and analysis of results*

It should be noted as a general result that the number of scientific publications in this field has increased more than 2,800% over the last 10 years (Figure 4). Although the publication date was not considered as a filter criterion in the search process, the following graph starts at 1998, because before 1998 only one article (in 1985) was published.

From a geographical point of view, most of the publications we studied where published within the United States of America, Germany, United Kingdom, China, and the Russian Federation, as seen in the following Figure 5.

Figure 6 shows the main subject areas: computer science (28%), business, management & accounting (16%), and engineering (14%). Surprisingly, the social science sector is also strongly represented with 11% of all scientific

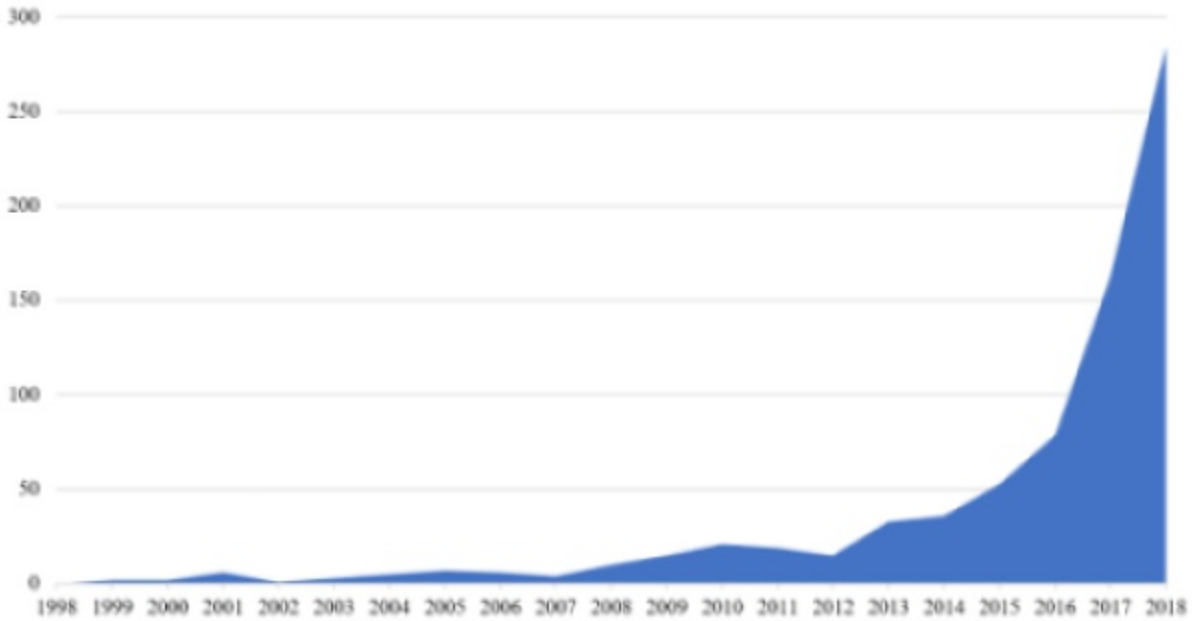


**Figure 4.** Number of scientific publications by year

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

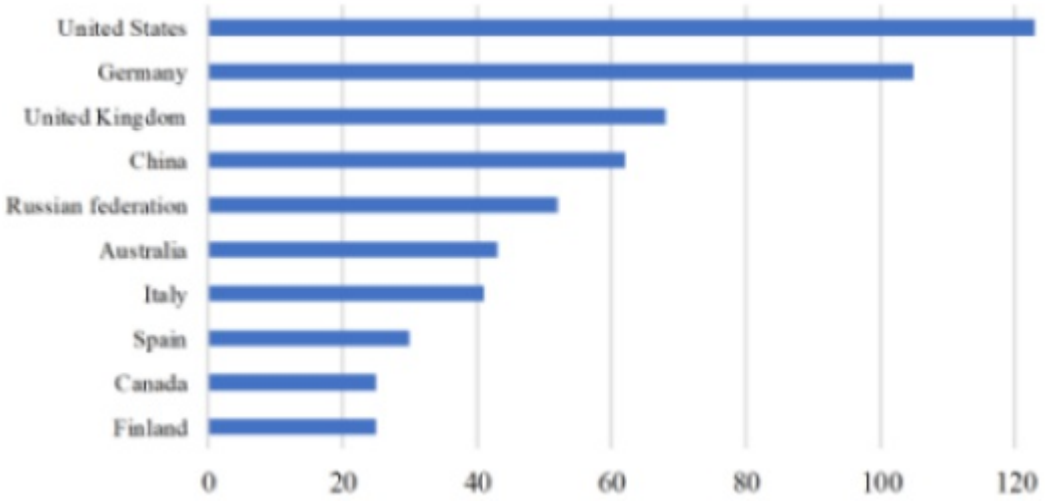


**Figure 5.** Number of scientific publications by country

publications on our topic. Figure 7 shows the leading research institutions in this field.

*Keyword relations in a network graph*

Gephi (http://gephi.org) software was used in order to identify and visualize subject areas and relations between keywords from the scientific literature. This software enabled the creation of a network graph, which illustrates the relations between keywords, as shown in Figure 8. In this graph, one can see so-called keyword nodes, as well as the edges that establish connections between nodes. The unfiltered graph includes 5,231 nodes and 114,414 edges. By using force-directed algorithms, where nodes repulse and edges attract each other, we identified nine relevant clusters.

By using filter techniques, such as a giant component (see Fulton et al., 2001) as used in the network theory, and a degree range setting of 65, only 658 nodes and 23,706 edges were left. In order to spatialize the network graph, the Forceatlas2 algorithm was used. Forceatlas2 is a force-directed layout where nodes repulse and edges attract (Jacomy et al., 2014). Furthermore, a modularity class filter was applied to examine the resulting

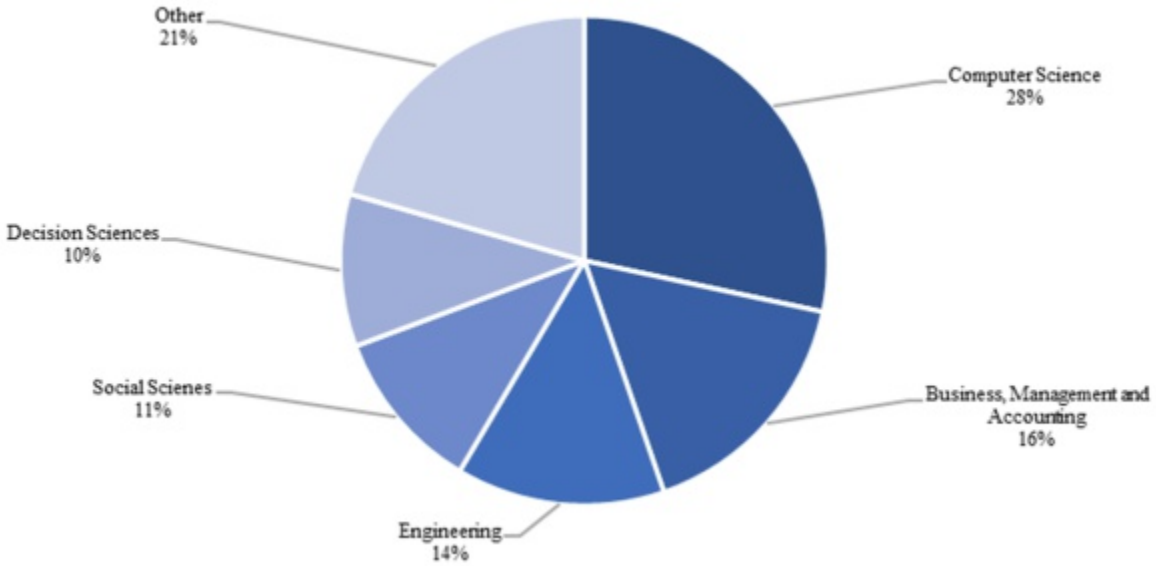


**Figure 6.** Portion of scientific publications by subject area

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

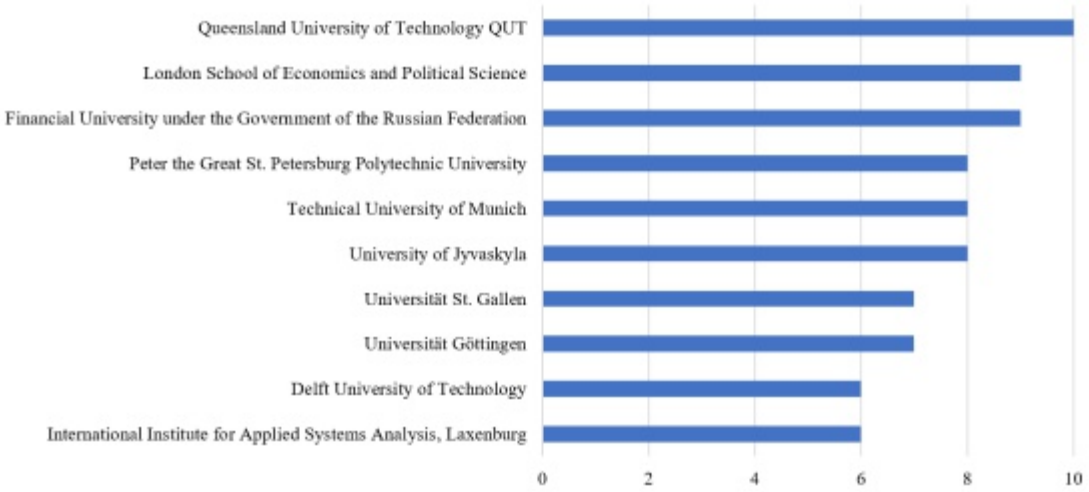


**Figure 7.** Number of scientific publications by affiliation

communities in the network (Blondel et al., 2014). As shown in Figure 8, the network graph has nine clusters. These were resized according to their degree of their interconnectedness to give a better presentation of the most relevant nodes.

Table 1 sums up the identified clusters, including the top keywords from each cluster. The name of the cluster is based on the node with the most incoming and outgoing edges. The number of all edges in total are given according to the keyword within the table. Based on the keywords related to the presented clusters, we derived a proposal for interpretation. This explanation was used to form a common understanding of the clusters in communicating preliminary results.

*Fields of innovation potential*

We argue that the nine clusters are to be regarded as categories for potential innovation within the overall data transformation towards a data economy. Organizations should pay attention to these topics, while transforming their business and developing digital services and business models. As well, they should track ongoing research to identify novel approaches and applications to different areas and topics.

In order to obtain more precise evaluation, we carried out a keyword comparison. For this purpose, we compared the number of keywords between the articles reviewed in stages 4 and 5. This was done to identify the ratio of keywords within all articles in order to discover possible articles for innovation topics.

Figure 9 shows a comparison between the top four article subject areas in data economy, as well as two selected subject areas (knowledge management and artificial intelligence), which are based on article abstracts. As can be clearly seen, many scientific publications address challenges within their abstracts, and therefore are rated as likely to provide insights about innovation potential. The topics of innovation management (76%), machine learning (71%), decision making (63%), and knowledge management (31%) are overrepresented in the publications dealing with challenges compared to all publications. However, contrary to expectations, we also found artificial intelligence having only a small increase (20%) in the number of mentions by comparison. While the ratio of artificial intelligence (AI) is relatively similar, the sub-class of machine learning reveals a considerable difference. This dominance of ML shows that authors prefer ML vs. AI terminology.

*Node analysis*

Analysis of the results shows that authors on this topic seem to assume that challenges are associated with data handling, innovation management, decision-making, and machine learning. We assume that more research in specific areas will lead also to higher innovation potential, especially in combination with other topics and technologies.

The following graph is presented in more detail for two concrete examples, first, for "Big Data" (Figure 10). The clearly visible connection to "Digital Transformation" and “Technological innovation” supports our argumentation about automatic identification of innovation potential.

Secondly, the node "machine learning" shows a very strong connection with the 5th cluster "Decision making" (Figure 11). Along with machine learning, one

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

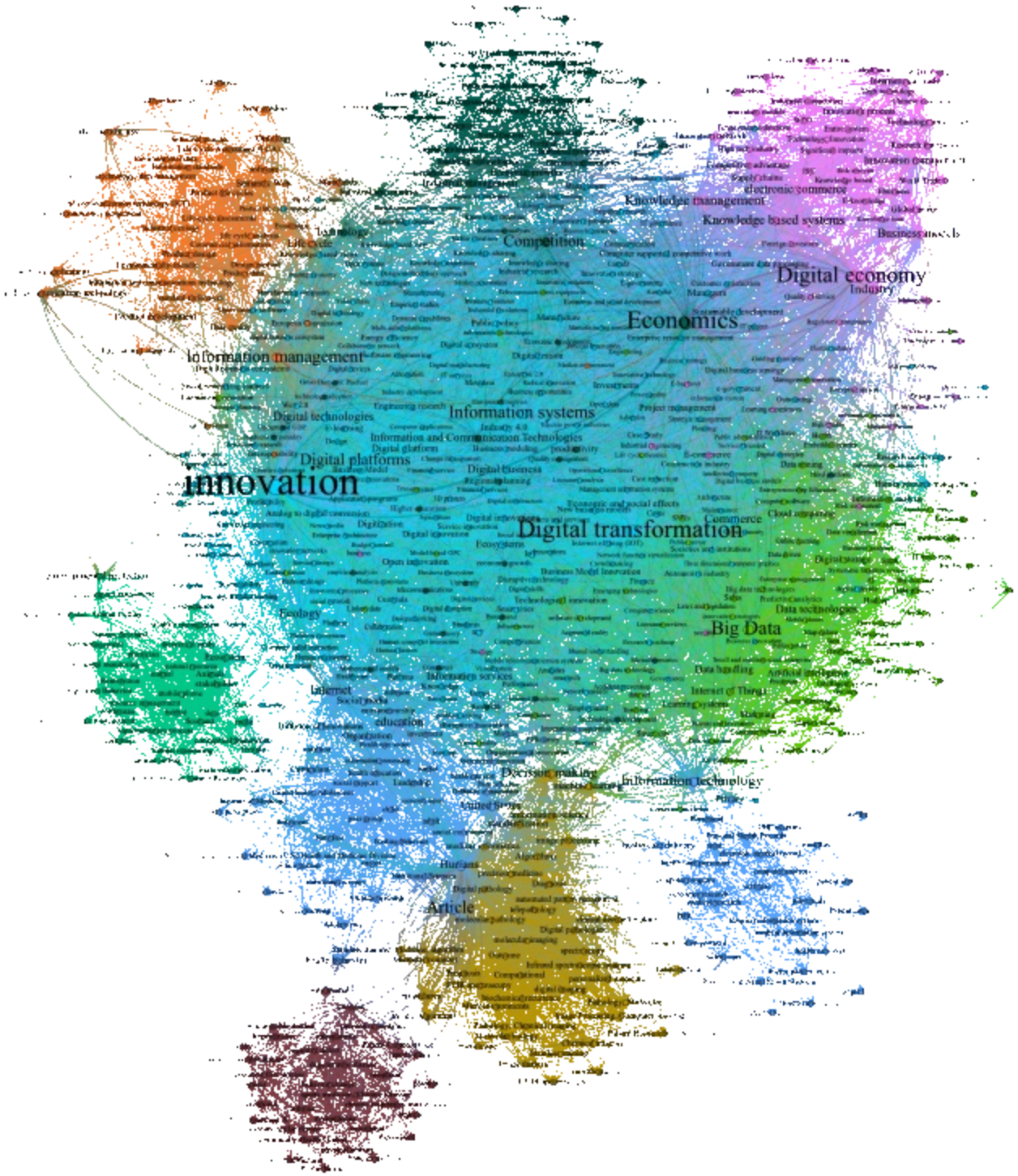


**Figure 8.** Network graph to visualize the relations between keywords

**Table 1.** Cluster results with related terms (below)

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

| Name | Terms | Interpretation |
|---|---|---|
| Digital transformation | Digital transformation (474) information systems (324), digital platforms (294), commerce (216), digital technologies (190), digital business (172), information services (148), ecosystems (136), open innovation (128), digital innovations (122), industry 4.0 (114) | The first cluster is the biggest cluster within the network graphic, and contains related terms with digital transformation. |
| Big data | Big data (362), data technologies (158), internet of things (142), data handling (142), learning systems (122), artificial intelligence (120), cloud computing (114), data mining (98), technological innovation (98), data privacy (88), Hadoop (68) | The cluster big data stands in close coloration with data ecosystems. This includes topics such as the handling of data, data processing, or big data technologies, such as Hadoop. |
| Digital Economy | Digital Economy (410), knowledge management (232), knowledge-based systems (230) business models (198), innovation management (148), international trade (124), competitive advantage (102), innovation network (102), High tech industry (78), Chinese companies (72) | Within the third cluster, the keyword digital economy has the most connections by far. This keyword is related, for example, to the development of new innovative business models in order to gain an advantage over competitors. |
| Economics | Economics (436), competition (246), information and communication technologies (160), productivity (102), regional planning (100), industrial management (92), economic growths (64), enabling technologies (52), industrial economics (42) | The topics of the Economics cluster are about general management topics, and processes as well as their adaption through digital components. This can be derived from strong relations to various keywords of topics within digital economy and innovation. |
| Decision making | Decision making (240), machine learning (136), information science (134), standardization (132), algorithms (110), image processing (110), image analysis (90), Prognosis (90) | The Gephi graphic shows that the use of decision-making algorithms, machine learning. Pattern recognition in the 5th cluster is very important for innovation and data. |
| Human | Human (536), information technology (210), internet (176), education (146), social media (86), united states (86), organization (70), efficiency (70), entrepreneurship (62), organizational innovation (60), international cooperation (46), leadership (44) | The cluster shows that roles and actors play an important role within data ecosystems. This is illustrated by keywords such as human, organization, and leadership. |
| Environmental Protection | Environmental protection (140), ecosystem (110), female (104), male (104), stakeholder (102), cooperation (98), conceptual framework (88), mass communication (88), natural resource (80), resource management (72) | This cluster shows close connections to environment & nature. Participants in natural ecosystems, as well as with resources seem to play a role here. |
| Information management | Information management (320), life cycle (162), information use (140), ontology (130), standards (128), semantic web (114), product design & development (112), software applications (94), design method (78), data quality (76), life cycle assessment (72) | The core of "information management" includes the handling and use of information. This is where the entire lifecycle of data plays a role: from data collection to the use of information for product development. |
| Healthcare | Exercise (78), preventive medicine (78), sports medicine (78), Patient-Centered Care (78), arthritis (72), behavior change (72), cardiovascular disease (72), cognitive defect (72), consensus (72), depression (72), diabetes mellitus (72), hypertension (72) | The ninth cluster is the only one with an industry focus. On the one hand, this cluster shows how important the use of data and information is in healthcare. On the other hand, it could be deduced that research and development in this sector is comparatively high. |

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

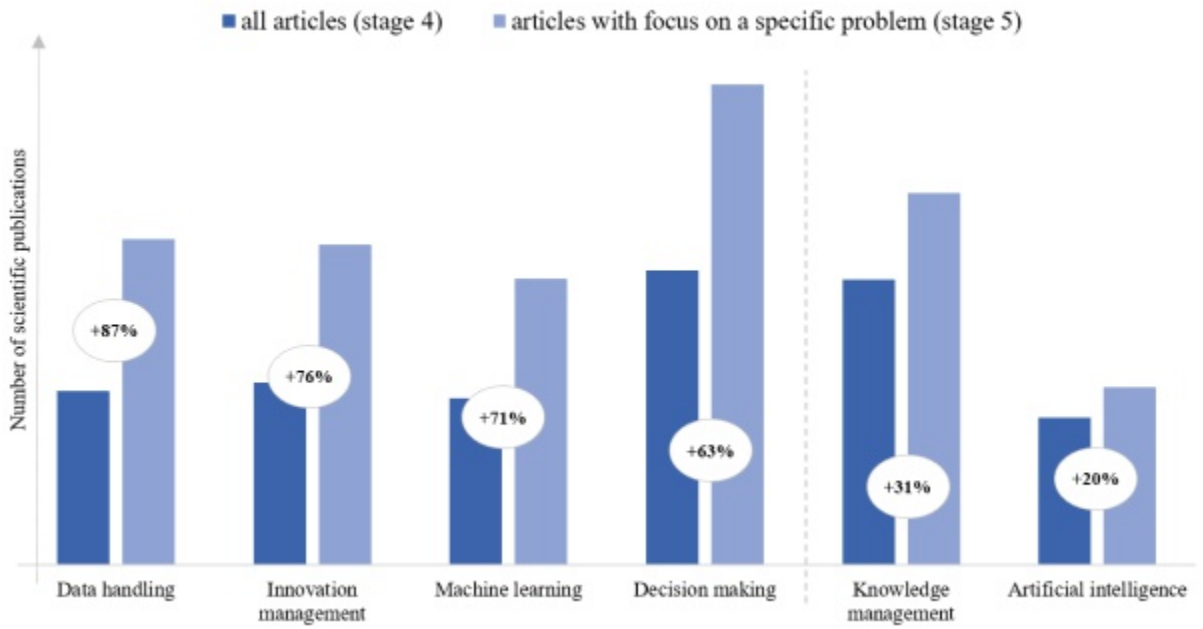


**Figure 9.** Keywords vs. keywords in context of challenges

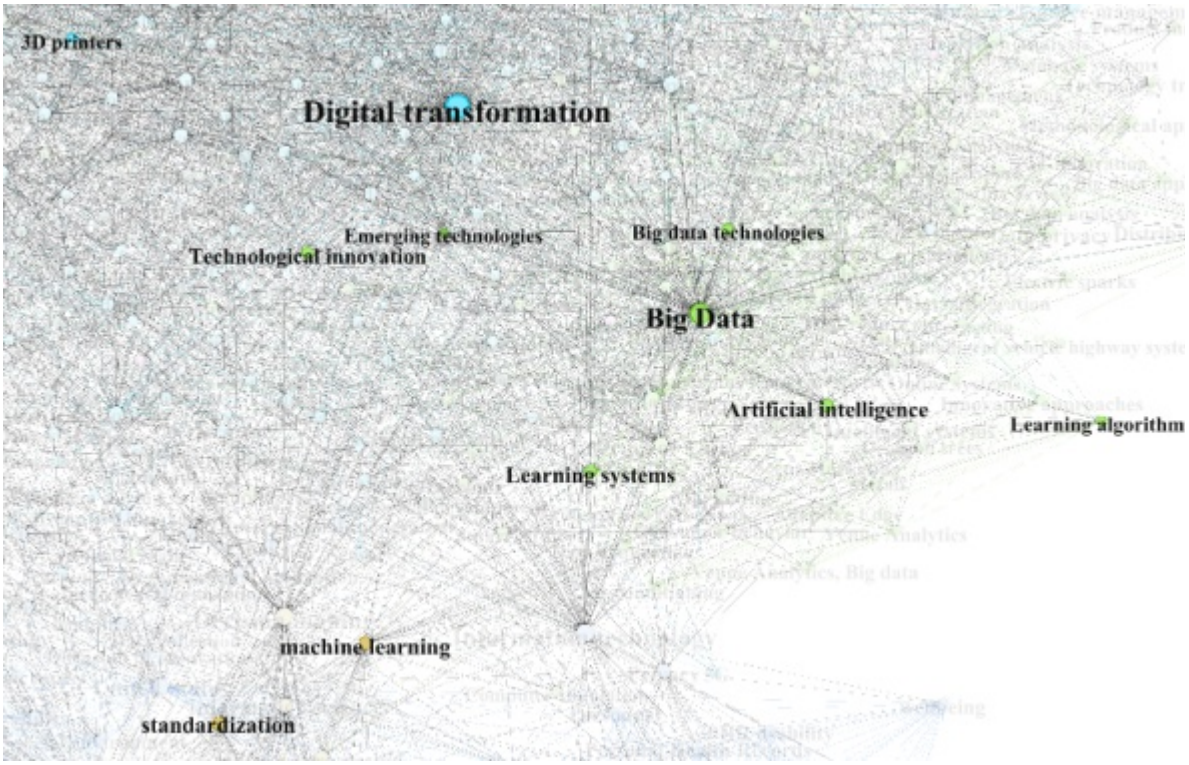


**Figure 10.** Gephi analysis of the node "Big Data"

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

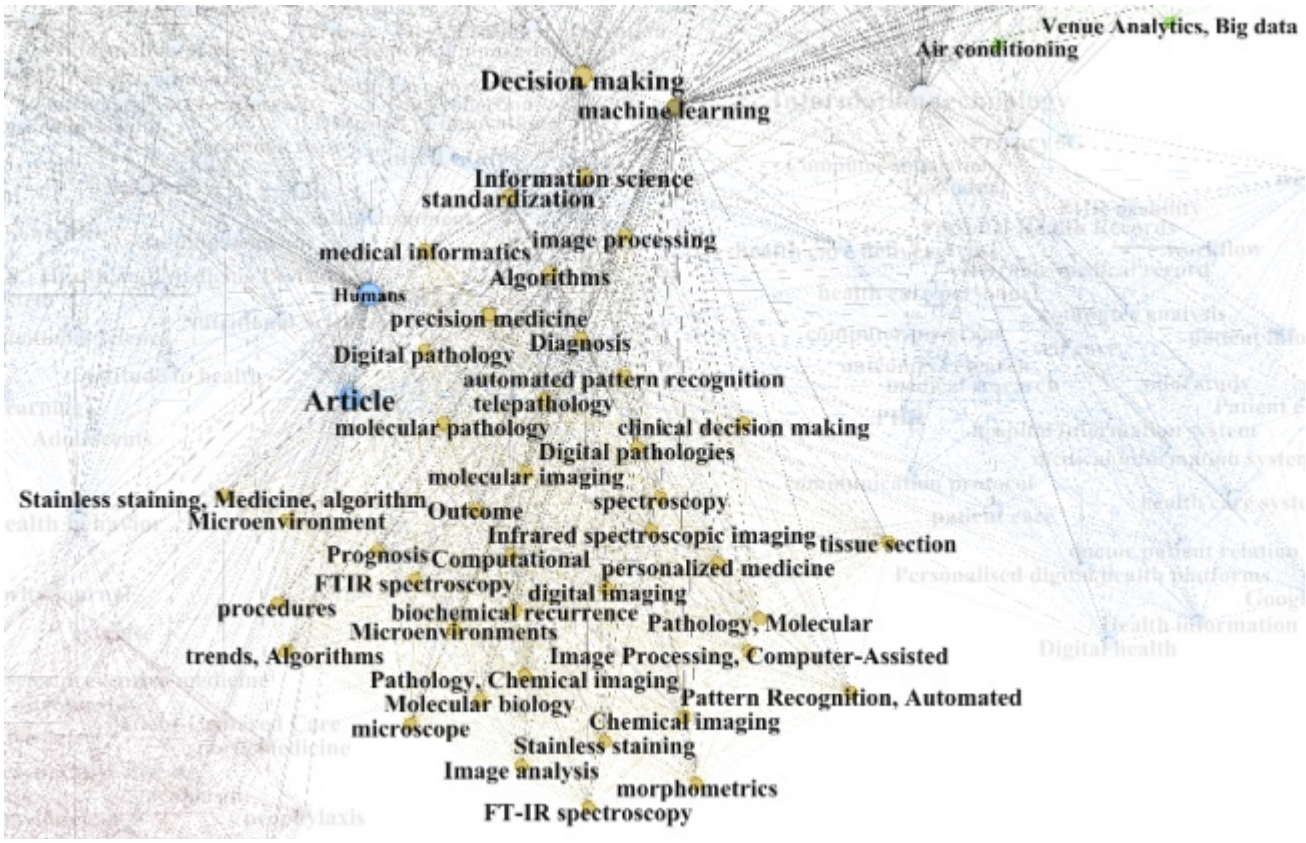


**Figure 11.** Gephi analysis of the node "machine learning"

can identify also strong connections to applications within the medical sector, shown by the nodes "medical informatics", "digital pathology", and "clinical decision making".

## Conclusion

This paper presents an automatized way to derive areas for innovation in the field of data economy. By conducting a systematic literature review in combination with basic text-mining methods, we identified 1,163 publications in the Scopus database. We analyzed them to identify a suitable dataset of publications containing terms related to challenges and requirements, as a way to answer our predefined research questions. We focused on these publications because abstracts dealing with challenges and related terms also refer to innovation topics. With pattern recognition based on text mining, we identified 334 articles based on abstracts that included specified terms for our analysis.

We then illustrated the development of topics and sub-topics related to data economy and innovation over the time, and depicted the main contributors in this area according to geography as well as affiliation. In addition, we identified major research streams by performing a network analysis and forming clusters based on the number of interconnections between different topics and their sub-topics. This provided an overview about relevant topics within the data economy that can help researchers derive topics where future research will probably emerge.

Researchers and practitioners are welcome to test the usefulness and applicability of our approach, especially evaluating our argumentation that derives innovation potential from challenges and requirement-related publications. Further research in the field of data economy may challenge our results with a more detailed view of specific clusters to gain even more insights.

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

***Reference***


Ahmed, Pervaiz K., Shepherd, Charlie. 2010. Innovation Management. *Context, strategies, systems and processes* (1st ed.). Harlow: Financial Times Prentice Hall.

Barnett, Homer G. 1953. *Innovation: The basis of cultural change*. New York: McGraw- Hill.

Blondel, Vincent D., Guillaume, Jean-Loup, Lambiotte, Renaud, and Lefebvre, Etienne. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*

Federal Association of the Digital Economy. 2018. Data Economy. *Datenwertschöpfung und Qualität von Daten*.

Coupland, Rob. 2018. Digital Realty. *Data Economy Report*.

Ding, Ying, Chowdhury, Gobinda, Foo, Schubert. 2001. Bibliometric cartography of information retrieval research by using co-words analysis. *Inf. Process.*

European Commission 2019. *Building a European data economy.* www.ec.europa.eu/digital-single-market/en/policies/building-european-data-economy.

Francis, Dave L. and Bessant, John. 2005. Targeting Innovation and Implications for Capability Development. *Technovation*, 25, (3): 171-183.

Bollobas, Bela. 2001. Random Graphs. *Cambridge Studies in Advanced Mathematics*, 73.

Jacomy, Mathieu, Venturini, Tommaso, Heymann, Sebastien, Bastian, Mathieu. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*. 9: 1–12.

Kitchenham, Barbara. 2004. Procedures for Undertaking Systematic Reviews. *Joint Technical Report*, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd.

Kitchenham, Barbara, and Charters, Stuart. 2007. Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3. *Engineering*, 45.

Moody, Daniel, and Walsh, Peter. 1999. Measuring the Value Of Information: An Asset Valuation Approach. *Seventh Eur. Conf. Inf. Syst.*: 1–17.

O'Mara-Eves, Alison, Thomas, James, McNaught, John, Miwa, Makoto, and Ananiadou,Sophia. 2015. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Syst. Rev.* 4: 1–22.

Otto, Boris. 2015. Quality and Value of the Data Resource in Large Enterprises. *Inf. Syst. Manag.* 32: 234–235.

Otto, Boris, Bärenfänger, Rieke, Steinbuß, Sebastian. 2015. Digital Business Engineering: Methodological Foundations and First Experiences from the Field. *Proc. 28th Bled eConference*: 58–76.

Rogers, Everett M. 1983. *Diffusion of Innovations* (3rd ed.). New York: Free Press.

Schuemie, Martijn J., Weeber, Marc, Schijvenaars, Bob JA, Mulligen, Rik M. Van, van der Eijk, C. Cristiaan, Jelier, C. Cristiaan, Mons, Barend, and Kors, Jan A. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20: 2597–2604.

Schumpeter, Joseph. 1912. *Theorie der wirtschaftlichen Entwicklung. Eine Untersuchung über Unternehmergewinn, Kapital, Kredit, Zins und Konjunkturzyklus* (9. Aufl.). Berlin: Duncker und Humbolt.

Siddiqi, Sifatullah, and Sharan, Aditi. 2015. Keyword and Keyphrase Extraction Techniques: A Literature Review. *Int. J. Comput. Appl.* 109: 18–23.

Tidd, Jie, and Bessant, John R. 2015. *Innovation and Entrepreneurship* (3rd ed.). Wiley.

van der Kooij, Bouke. 2017. Search for a Common Ground in the Fogs of Innovation Definitions. *SSRN Electronic Journal*.

Vedin, Bengt-Arne. 1980. *Management and organization for innovation: radical product renewal in large corporations*. Göteborg: Chalmers TH.

Yoo, Youngjin, Henfridsson, Ola, and Lyytinen, Ola. 2010. The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research*. Vol. 21, No. 4, December: 724–735.

Zimmerman, Hans-Dieter. 2000. Understanding the Digital Economy: Challenges for New Business Models. *AMCIS 2000 Proceedings*.

Zawislak, Paulo Antônio, Borges, Mauro, Wegner, Douglas, Santos, Andre and Cristina, Castro-Lucas. 2008. Towards the Innovation Function. *Journal of Technology Management & Innovation*.

# Uncovering Research Streams in the Data Economy Using Text Mining Algorithms

*Can Azkan, Markus Spiekermann, Henry Goecke*

### *About the Authors*

Can Azkan is a scientist and PhD candidate at the Fraunhofer Institute for Software and Systems Engineering ISST in Germany. He studied Mechanical Engineering at the Technical University of Dortmund and the San Diego State University, while he gained practical experience in the field of industrial engineering and digital business models in machine und plant engineering. His research at Fraunhofer ISST focuses on value co-creation in emerging data ecosystems and the management of data as a corporate asset.

Markus Spiekermann currently works as Head of Department "Data Business" at the Fraunhofer Institute for Software and Systems Engineering in Dortmund, Germany. He leads research projects and is active in several related advisory boards. His main research focuses on the topics of data engineering and data management, alongside on the valuation of data assets especially within data ecosystems. Before his time at Fraunhofer, he worked as IT-Professional and Software Engineer from 2008 to 2016. He obtained his Bachelor and Master of Science degree in the field of information systems with a focus on IT Management at the FOM University of Applied Sciences in Essen.

Since 2017 Dr. Henry Goecke has been head of the Research Group "Big Data Analytics" at the German Economic Institute. Previously he worked at the German Economic Institute as scientific assistant of the Director, at the IW Consult as Senior Economist, at the TU Dortmund University as research and teaching assistant as well as lecturer at the University of Cologne and the Hochschule Fresenius. He studied Economics at the TU Dortmund University, Strathclyde University of Glasgow, and the Leuphana University of Lüneburg. His research interests are on the impact of social media, artificial intelligence, big data, and data economy.

Citation: Azkan, C., Spiekermann, M., Goecke, H.. 2019. Uncovering Research Streams in the Data Economy Using Text Mining Algorithms. *Technology Innovation Management Review*, 9(10): 62-74.
http://doi.org/10.22215/timreview/1283



Keywords: Data Ecosystem, Data Economy, Digital Economy, Data Ecosystem, Digital Transformation, Data Market, Big Data, Literature Review, Network Graph, Text Mining.