Kazimieras Simonavičius
UNIVERSITY

Integrated Knowledge Management in the Field of Big Data Foresight and Associated Digital Platforms

WP2 Report
8 April 2019

Jari Kaivo-oja, Steffen Roth, Arūnas Augustinaitis,
Mikkel Stein Knudsen, Mathew Maavak, Levan Bzhalava, Darius Verbyla,
Austė Kiškienė and Theresa Lauraeus

www.ksu.lt

Report was developed under Project „Platforms of Big Data Foresight (PLATBIDAFO)“.

Project is implemented under funding of European Regional Development Fund.

Ataskaita parengta vykdant projektą ,,Didžiųjų duomenų platformos ateities įžvalgoms“, finansuojamas pagal priemonės Nr. 01.2.2-LMT-K-718 „Tiksliniai moksliniai tyrimai sumanios specializacijos srityje“ veiklą „Mokslininkų iš užsienio pritraukimas vykdyti mokslinius tyrimus“.
Nr. 01.2.2-LMT-K-718-02-0019

Projektas finansuojamas Europos Regioninės plėtros fondo lėšomis

Kuriame Lietuvos ateitį
2014–2020 metų Europos Sąjungos fondų investicijų veiksmų programa

The authors' language is uncorrected

Published by Kazimieras Simonavičius University
Dariaus ir Girėno st. LT-02189 Vilnius, Lithuania
E-mail: info@ksu.lt
www.ksu.lt

Free publication

ISBN 978-609-95634-9-7

Contents

1. Introduction 4

2. Big Data library development 7

3. Open innovation tools and crowdsourcing with Big Data 17

4. Knowledge management aspects of foresight analyses with Big Data 23

5. E-commerce tools and algorithms with Big Data and foresight processes 27

6. Platform economy and platform approach 30

7.Integration challenges in Big Data field: Ethical codes and other integration challenges 32

7.1. Ethical aspects 32

7.1.1. Issues of ownership and transparency 32

7.1.2. Data ecology issues: misinformation, overload, pollution 33

7.2. Further big data integration challenges 34

8. Methodological and theoretical challenges in the big data field 36

References 39

Annex 1: „Platforms for Big Data Foresight" Project Activities (Nov. 2018 – Mar. 2019) 46

Annex 2: „Platforms for Big Data Foresight" Project Publications (Nov. 2018-) 47

1. Introduction

Oxford Dictionaries define Big Data as: “Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions: much IT investment is going towards managing and maintaining Big Data” (Oxford Dictionaries, 2019).

This survey report focuses on knowledge management aspects of Big Data analytics and applications. Big data analytics is a special field that concerns how to (1) analyze, (2) systematically extract information from many data sources, or (3) otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

Big Data with many cases or data rows offer greater statistical power, while Big Data with higher complexity with more attributes or columns may lead to a higher false discovery rate because of problems of data management. Today typical Big Data key challenges include: (1) Capturing data, (2) creating data storage, (3) performing data analysis, (4) performing data search, (5) sharing data, (6) transferring data, (7) data visualization, (8) querying data, (9) updating data files, (10) data and information privacy and (11) data source(s).

Gartner defines big data as “high-volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making and process automation.” (LeHong and Laney, 2013). This needs to be augmented by aspects such as veracity (how much noise and uncertainty is in the big data) and value (value of big data) amongst others (Moorthy et al, 2015) as outlined in Fig. 1.

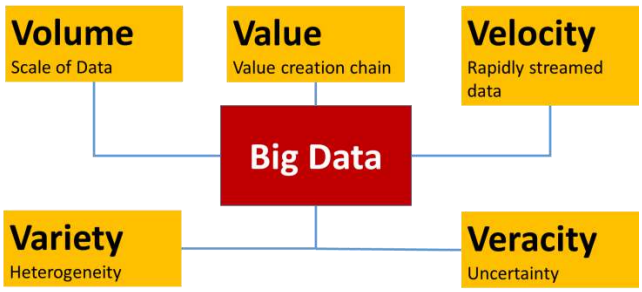

Figure 1. Big Data – conventional 5 Vs.

Before Big Data key concept of knowledge management was Business intelligence, which was dominating many discussions in the field of knowledge management. In Fig. 2 we can see how peoples´ interest in business intelligence has been developing in 2004-2019 (before 14.3.2019), when we analyze the big data base of Google Trends.

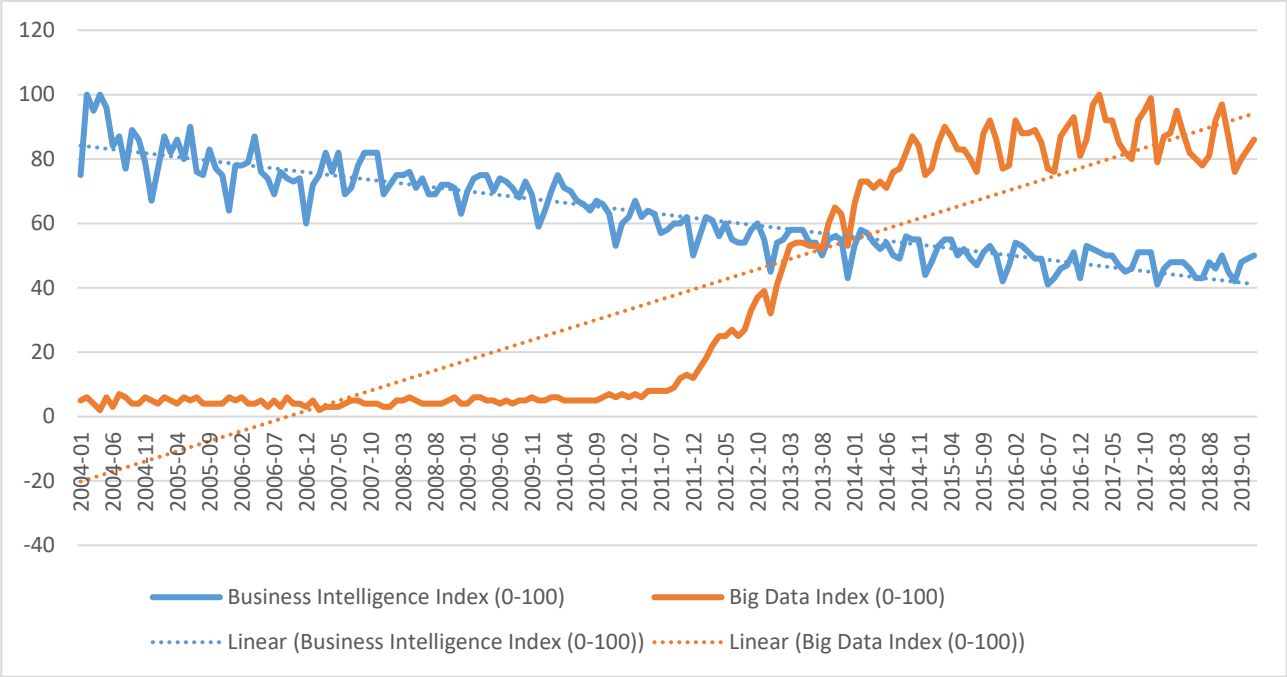

Figure 2. Business Intelligence Index and Big Data Index Trends in 2004-2019 with Linear Trend Lines (Index 0-100). Source: Google Trends (2019) Monthly Global Data from Google Trends 24.3.2019. https://trends.google.fi/trends/?geo=FI).

Liang and Liu (2018) have analyzed research landscape of Business Intelligence and Big Data analytics. They note that in the last five years, the trend of "Big Data" has emerged and become a core element of Business Intelligence research. Key high frequency words in this field are: "datamining", "social media, "information system". After 2016, new high frequency keywords have been "cloud computing", "data warehouse" and "knowledge management".

We can conclude that interest in Big Data analytics has development much especially after 2012, while interest in business intelligence has decreased in the long-run. Big Data analytics will have impacts on many fields of life. People are more and more interested about Big Data and opportunities which it will provide. Big Data is a revolutionary phenomenon, which is one of the most frequently, occurred topics in scientific and practical discussions nowadays. In Fig. 3 we have reported forecast about big data market worldwide in 2011-2027 in billion U.S. dollars.

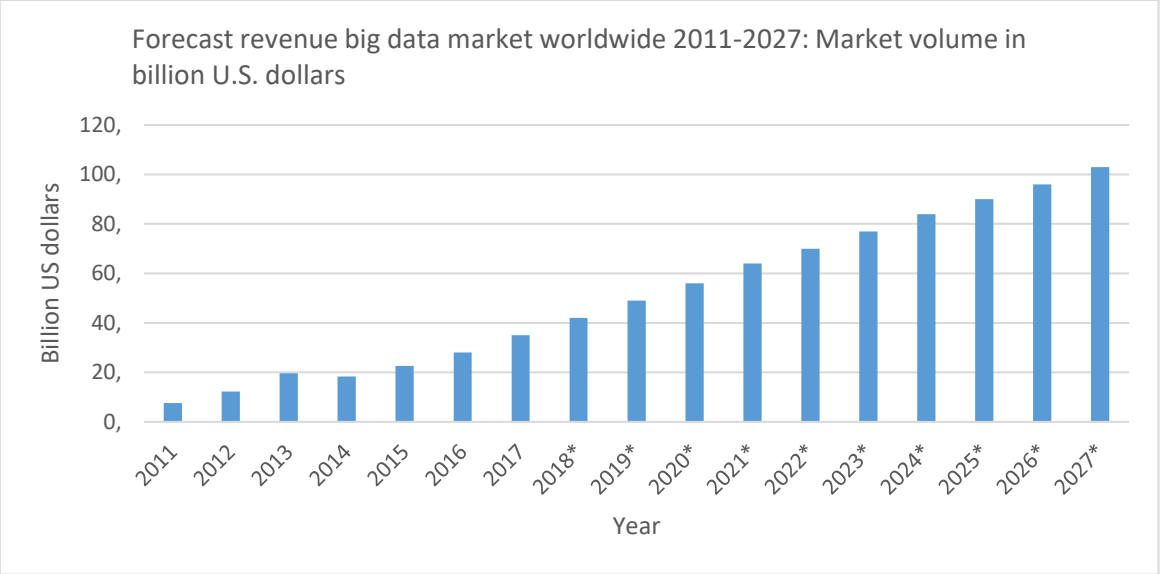

Figure 3. Big data market size revenue forecast worldwide from 2011 to 2027.: Market volume in billion U.S. dollars, Data source: Wikibon; Silicon ANGLE, Survey period March 2018.

We can expect that in the global economy and business the importance of knowhow in the field of Big Data analytics is growing. Many business organizations and corporations invest in Big Data competences and dynamic capabilities (see e.g. Eckroth, 2018). Key functions in organizations are (1) Big Data storage, (2) Big Data processing, (3) Big Data nalyzing and (4) Big Data transmissions. Current survey paper indicates that key issues in Big Data research are (1) descriptive analyses, (2) predictive analyses and (3) classification issues of Big Data (Shadroo and Rahmani, 2018, p. 27).

2. Big Data library development

Libraries have a long time been important social institutions that help people access various information resources. With the continuous development of information technologies (ITs), libraries have been evolving constantly, and this has greatly expanded library services and improved their efficiency and effectiveness.

The Online Computer literacy Centre (OCLC) was originally conceived in 1967 as the Ohio College Library Center to integrate varied and dispersed library resources into a single electronic database. Today WorldCat, formed by libraries around the world, has over 300 million records of physical and electronic books and journals, recordings, movies, maps and assorted data. However, these 300 million individual records still remain largely isolated in the Web, requiring integration for ease of access and analysis (Teets and Goldner, 2013).

The concept of a global digital library gathered further impetus after the introduction of the World Wide Web in the early 1990s.

After years of practical application and development, digital libraries have gradually become a very important aspect in the development of modern libraries. Digitalization of data and information systems is a trend in developed economies. Instead of analog storages, more and more digital storages can been developed in modern societies. We can note that mankind is on a quest to digitize the world. Especially, the concept of Digital Twin will an important part of digitalization of economic and human activity. Industry 4.0 approach is based on digitalization (see Kaivo-oja, Kuusi, Knudsen & Lauraeus, 2019 and Knudsen, Kaivo-oja & Lauraeus, 2019).

Nowadays data sets grow rapidly- in part because they are increasingly gathered by cheap and numerous information- sensing Internet of Things devices such as (1) mobile devices, (2) aerial (remote sensing), (3) software logs, (4) cameras, (5) microphones, (6) radio-frequency identification (RFID) readers and (7) wireless sensor networks (Hellerstein 2008, Segaran & Hammarbacher, 2009). Today these machines generate data a lot faster than people can do it, and their production rates will grow exponentially with Moore's Law. Storing this big data is cheap, and it can be mined for valuable information. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s (Hilbert & López 2011) as of 2012, every day 2.5 exabytes (2.5×1018) of data are generated. (IBM, 2013).

Based on an IDC report prediction, the global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020 (Hajirahimova & Aliyeva, 2017). By 2025, IDC predicts there will be 163 zettabytes of data (Reinsel, Gantz and Rydning, 2017).

Today demand for Big Data is growing in the digital library field (De Mauro, Greco, & Grimaldi, 2016). It is surprising that relatively few studies have investigated digital libraries relative to Big Data. One obvious reason for the lack of such studies is that many people still think that traditional database management systems can handle the daily data storage and business processing requirements of digital libraries (see Xu, Du, Wang, & Liu, 2017).

An “upper ontology” or knowledge graph approach has been advocated for the organization of data of such variety, volume and complexity. A knowledge graph is a model of relationships between entities or objects in a given space. As libraries are entrusted with organizing knowledge, the entities in a library knowledge graph must be quite broad and should include the following key entities (Teets & Goldner, 2013, pp. 432):

People: Traditionally involved people with formally published works. Increasingly, however, we must account for individuals involved in activities ranging from writing, reviewing, publishing to simply using library content.

Items: Physical items held within libraries (e.g. books and media) and electronic-only information such as e-books, journal articles and digital scans of real objects

Places: Geographic locations past and present must be maintained to understand context ofpublished works.

Events: Events can account for grand events such as a public performance by a rock music group, and also minor events such as a user viewing a book text;

Organizations: Organization entities encompass corporate names, publishers, political parties and bodies, and associations;

Concepts: Subject classification systems such as the Dewey Decimal Classification system and Library of Congress Class system are well-known concept organization schemes. More sophisticated data-linking systems are being developed.

In a nutshell, the reorganization of library databases and functions necessitate the adoption of Big Data Analytics (BDA) and associated restructuring tools. It has wider applications as well in the open source realm: Big data can be used to automatically discover relationships in innovation activities. An integrated system should be able to recognize and enable communications between geographically-dispersed researchers working on a similar subject. As Teets and Goldner (2013) observed: “Just making big data sets accessible is not a desired end point. It is about making the

data reusable in combination with other data sets across the Web.” For example, text mining of Big Data can be applied in many ways in innovation and foresight studies (see Bzhalava et al. 2019a, 2019b and 2019 c).

Some scholars like Chen et al. (2014) think that changes in five specific aspects are particularly obvious: (1) changes in the international data environment require digital libraries to manage Big Data; (2) changes to scientific research methods require digital libraries to support more data-driven research environments; (3) the transfer of the innovation model requires digital libraries to meet the needs of business development; (4) changes in user information literacy require digital libraries to meet the needs of knowledge search; and finally (5) digital libraries must adapt to development and changes in information technology to upgrade old service platforms to newer. In Fig. 4 we can present methods and trends of digital library transformation in the Big Data era.

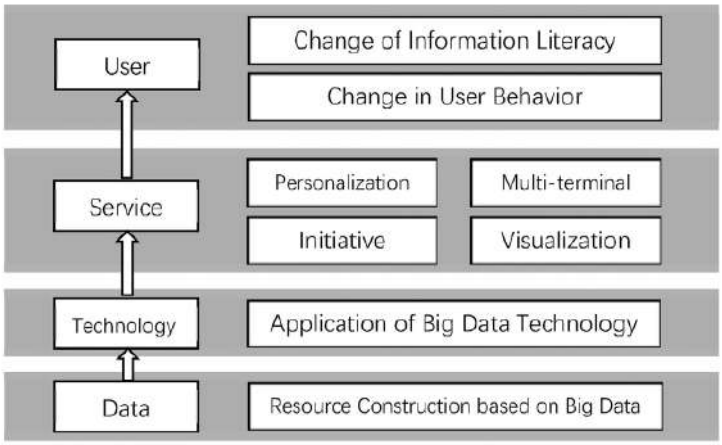

Figure 4. Methods and trends of digital library transformation in the Big Data era. (Li, Jiao, Zahng, & Xu, 2019, p. 28)

Big Data sources in the Internet of Things and trend development of Internet connected devices is visualized in Fig. 5. We can conclude that the business relevance of Big Data management is increasing in the future.

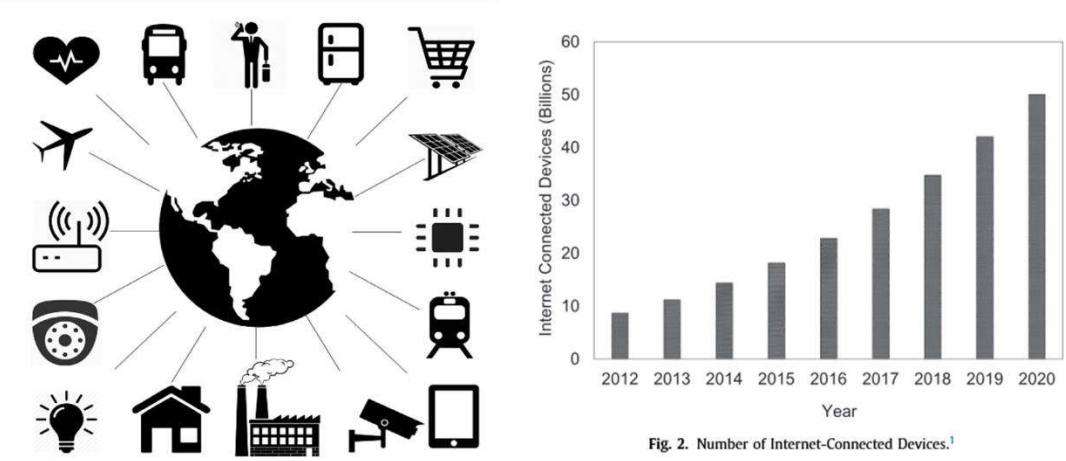

Figure 5. Big Data sources in the Internet of Things and number of Internet-connected devices (Ahmed et al., 2017).

In Fig. 6 we can visualize Big Data flow (1) from IoT Infrastructure to Big Data Platform and (2) finally to Big Data Analytics (Ahmed et al., 2017, Ge et al., 2018). Key processes from the IoT Infrastructure to the Big Data Platform are:

- Capture
- Integrate
- Store and
- Preprocess.

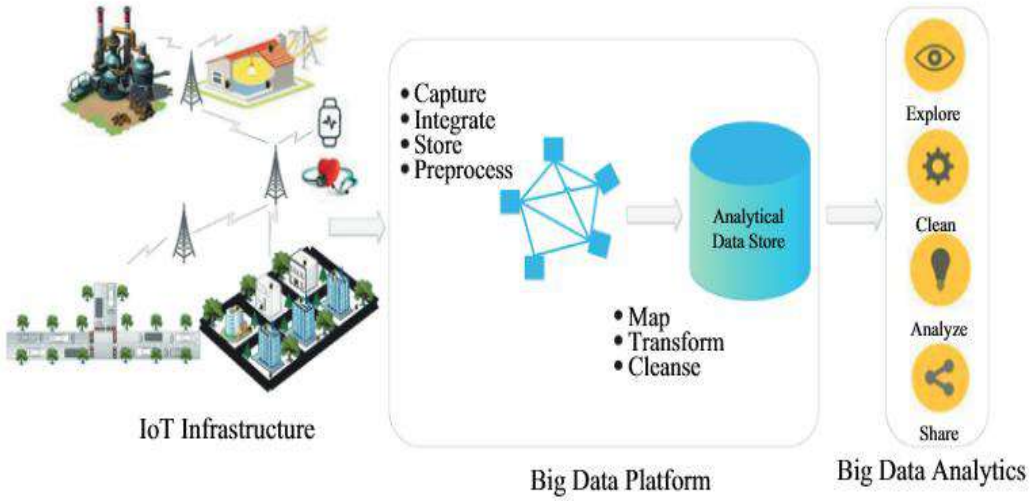

Figure 6. Big Data Flow (Ahmed et al., 2017)

Key processes from the Big Data Platform to the Big Data Analytics are:

- Mapping
- Transformation and
- Cleaning.

Key elements of Big Data Analytics are:

- Exploring
- Cleaning
- Analyzing and
- Sharing.

In Fig. 7 we have presented the key characters of Big Data with the fundamental idea that Big Data with various characters will leads to high quality foresight analytics.

In recent study about critical factors influencing effective use of big data, the following *thematic organizational issues* were identified to be important (Surbakti et al., 2019, p. 6):

1. Organizational cultural competence,
2. Talent management,
3. Change management program,
4. Strategic alignment,
5. Project management,
6. Performance management,
7. Organizational structure and size,
8. Interdepartmental collaboration,
9. Communication,
10. Top management support,
11. Environmental effect,
12. Clear goals, and
13. Focus on innovation.

In the study about critical factors influencing effective use of big data, the following *thematic systems, tools and technologies issues* were identified to be important (Surbakti et al., 2019, p. 7):

1. System quality,
2. IT Infrastructure, and
3. Vendor support.

In the study about critical factors influencing effective use of big data, the following *thematic people aspects* were identified to be important (Surbakti et al., 2019, p. 7):

1. People's knowledge and skills,
2. Trust,
3. Champions,
4. Employee engagement,
5. User participation, and
6. Individual characteristics.

In the planning and management processes of Big Data libraries and services these thematic issues are highly relevant. Naturally, factors of data and information quality are always important. Other relevant aspects are data completeness, data currency, data access, data relevance, data accuracy and data consistency (Surbakti et al., 2019, p. 9).

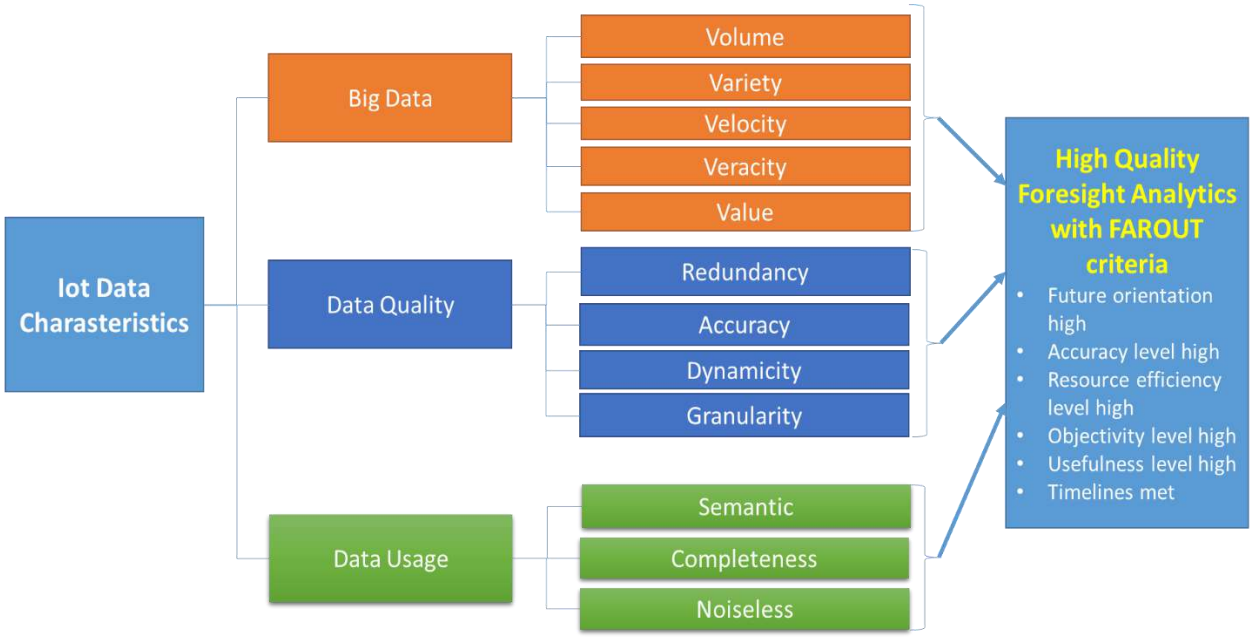

Figure 7. The Characters of Big Data (modification of Mahdavinejad et al., 2018).

In Fig. 7 the FAROUT refers to Future-orientation, Accuracy, Resource use, Objectivity, Usefulness and Timelines, 6 key quality criteria of futures and foresight studies (see Kaivo-oja and Roth, 2019).

Big Data Library research and development is a challenging issue. It is very clear that library contains big data which are valuable but wait for exploration. Big data research and development is different and relatively new field. In status and directions study Xu et al. (Xu et al., 2017) identified the following challenges:

1. Lack of data scientists,
2. Ability of adopting Big Data is low,
3. Budget issues,
4. Technical challenges,
5. Privacy challenges, and
6. Big Data is not for all organizations.

In order to solve all these challenges listed above, there is need to develop Big Data technology and educate new generation of Big Data scientists and specialists. There is often non-linear dynamics in Big Data R&D activities. Change of mind set is often needed, because old habits and models lead us conventional solutions, which do not help us to solve these challenges. However, a lot of libraries begin to embrace big data technology and many interesting pilot projects are going on. Such issues like (1) the needs of data-driven decision-making, (2) new data formats, (3) data standards and data modelling, (4) library data visualization, (5) user behavior studies and (6) big data technology possibilities are key issue in Big Data library projects (Xu et al., 2017, p. 84-85). Especially the perspective of personalized user services is critical issue and needs special attention (Li et al. 2019).

In Fig 8 the Big Data lifecycle is visualized. Big Data lifecycle starts from (1) Study and planning phase leading to (2) Data collection, (3) Data documentation and quality assurance, (4) Data integration, (5) Data preparation, (6) Data analysis, (7) Publishing and sharing and finally to (8) Data storage and maintenance and (9) Data reuse.

When a Big Data platform owner has implemented the Big Data lifecycle once, it is easier to make new data cycle rounds and new elements to Big Data analytics.

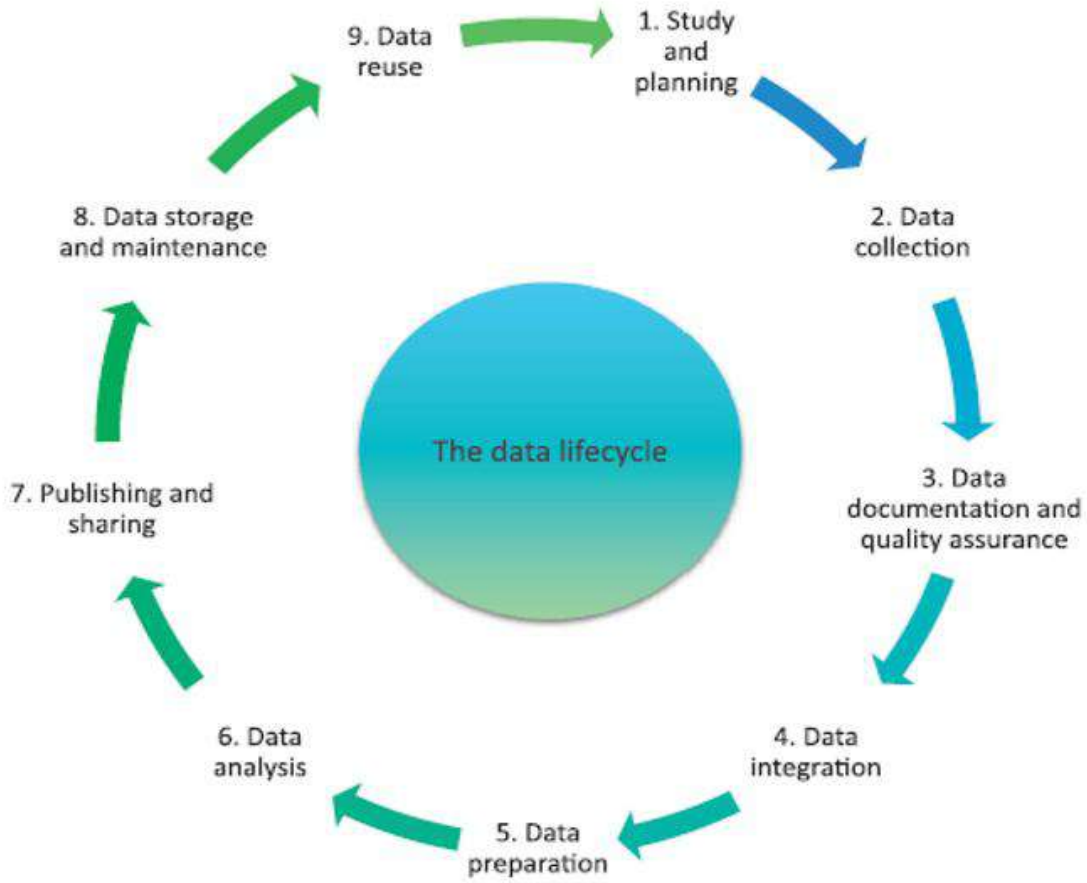

Figure 8. The Big Data Lifecycle (Blazquez and Domenech, 2018)

In Fig 9 we have presented a comprehensive classification of key sources of socio-economic Big Data.

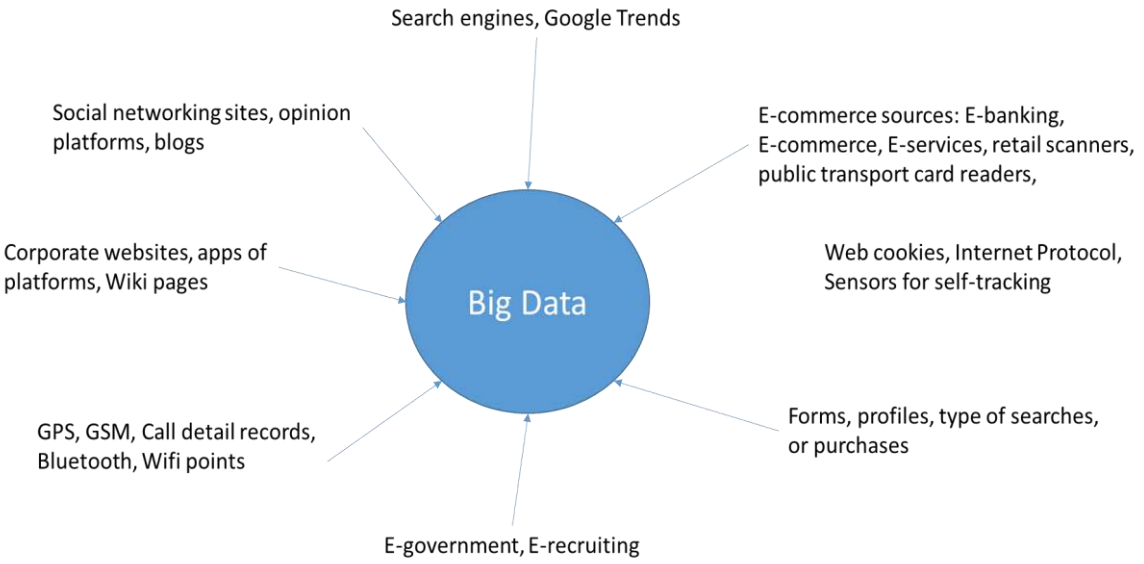

Figure 9. Classification of sources of socio-economic Big Data (modification of Blazquez and Domenech, 2018).

In Table 1 we present different data sources, short description of data sources and functions of users-

Table 1. Data sources, short descriptions of data sources and functions of users. (modification of Blazquez and Domenech, 2018).

Data sources	Short description	Functions of users
Search engines, Google Trends	The user aims to find information about a topic of his interest. Data is actively generated Search engines, Google Trends. The user interacts with an individual and/or machine to achieve an agreement in which the user demands and obtains a product or service in exchange for a financial or nonfinancial compensation. Data is actively generated.	Information search
E-commerce sources: E-banking, E-commerce, E-services, retail scanners,public transport card readers,	Event in which the user makes a payment to obtain a product or service	Financial transactions
Web cookies, Internet Protocol, Sensors for self-tracking	The simple fact of using any device generates data related to how, when and where an action has been done.	Usage
Forms, profiles, type of searches, or purchases	Personal data (age, sex, etc.) is generated consciously (e.g. filling a form to complete a purchase) or unconsciously (e.g. data about the type of information we look for is used to infer our incomes) as a consequence of using any device or tool to achieve a purpose.	Personal
E-government, E-recruiting	Event in which the user provides the counterpart with required information to obtain a product or service	Non-fiancial tranactions
GPS, GSM, Call detail records, Bluetooth, Wifi points	The use of mobile phones generates data particularly related to the position of the user.	Location
Corporate websites, apps of platforms, Wiki pages	The user aims to spread information or knowledge. This includes marketing purposes, in order to establish a public image of the user or the agent he represents. Data is actively generated.	Information diffusion
Social networking sites, opinion platforms, blogs	The user wants to share information, opinions and ideas with other users. Data isactively generated.	Social interaction

In Table 2 we present some examples of Big Data sources.

Table 2. Examples of Big Data sources

Data sources	Examples	Functions of users
Search engines, Google Trends	Google Trends: https://trends.google.ae/trends/?geo=AE Hostel Bookers: https://www.hostelbookers.com/blog/travel/fashion-trends/	Information search
E-commerce sources: E-banking, E-commerce, E-services, retail scanners,public transport card readers	Magento: https://demo.magento.com/	Financial transactions
Web cookies, Internet Protocol, Sensors for self-tracking	Insight article "Analyzing Internet Traffic Structure through Big Data Technology" by Keisuke Ishibashi, Shigeaki Harada, and Satoshi Kamei : https://www.ntt-review.jp/login/ntttechnical.php	Usage
Forms, profiles, type of searches, or purchases	Acxiom: https://www.acxiom.com/	Personal
E-government, E-recruiting	Big Data and e-Government: A review, https://ieeexplore.ieee.org/document/8080062 Adotion of E-governance applications towards Big Data approach, https://www.ripublication.com/ijaer17/ijaerv12n21_113.pdf	Non-financial transactions
GPS, GSM, Call detail records, Bluetooth, Wifi points	WiFi positioning and Big Data to monitor flows of people on a wide scale, the EU report, https://ec.europa.eu/jrc/en/publication/wifi-positioning-and-big-data-monitor-flows-people-wide-scale	Location
Corporate websites, apps of platforms, Wiki pages	Top 14 Big Data Companies: https://thinkmobiles.com/blog/best-big-data-companies/ Industrial Big Data, https://en.wikipedia.org/wiki/Industrial_big_data	Information diffusion
Social networking sites, opinion platforms, blogs	InsightsAtlas, https://insightsatlas.com/	Social interaction

In Fig. 10, genesis of Big Data applications is visualized. Genesis of Big Data applications is including the gradual development of the architecture of candidate applications from early desktop to recent versions (Abolfazli et al., 2014).

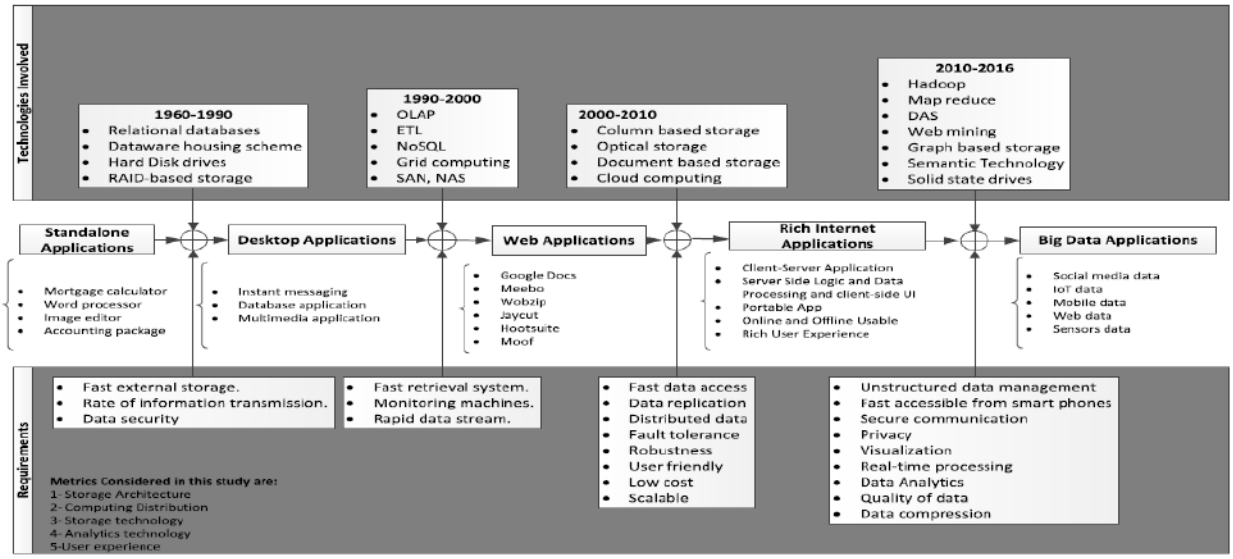

Figure 10. Genesis of big data applications, including the gradual development of the architecture of candidate applications from early desktop to recent versions (Abolfazli et al., 2014).

In Fig. 11, Big Data architecture for nowcasting and forecasting social and economic changes is visualized (see Blazquez & Domenech 2018).

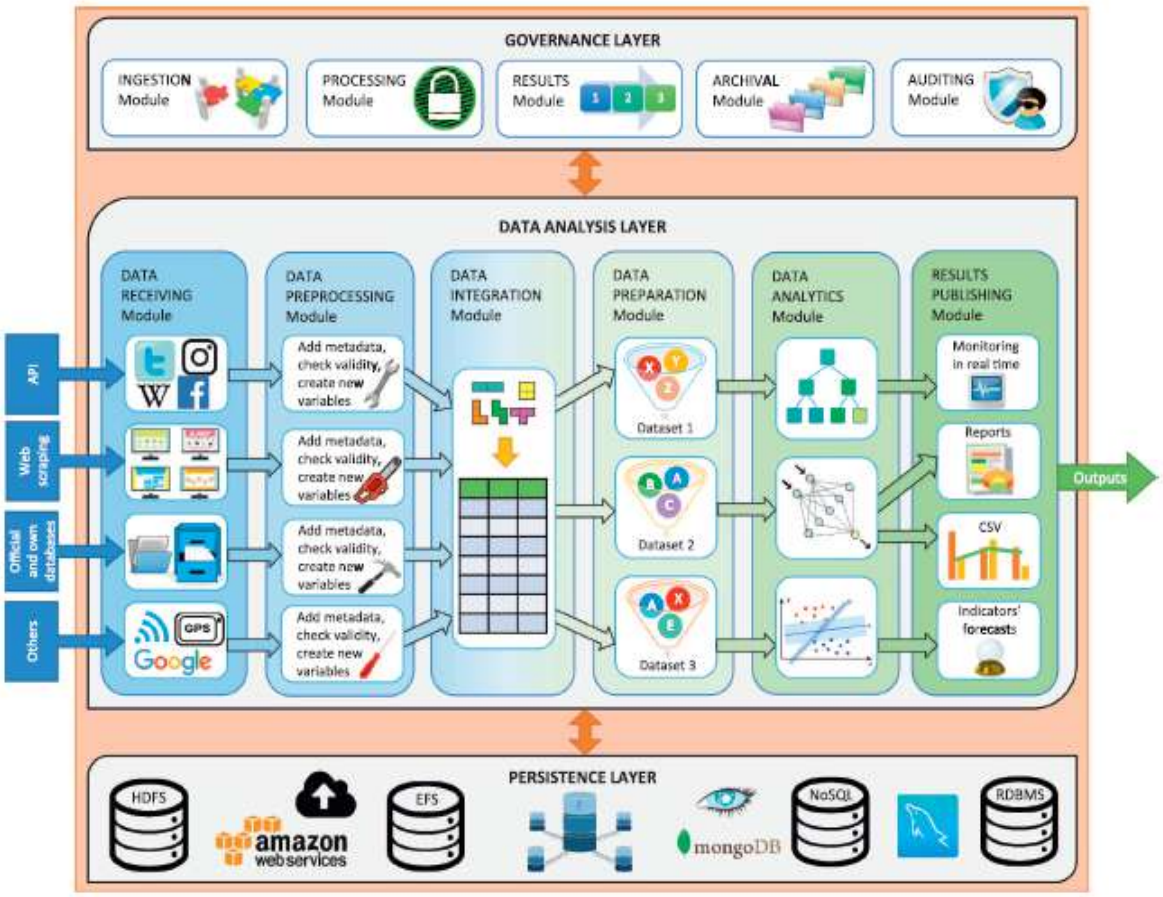

Fig. 4. Big Data architecture for nowcasting and forecasting social and economic changes.

Figure 11. Big Data Architecture for Nowcasting and Forecasting Social and Economic Changes (Blazquez & Domenech, 2018).

Fig. 11 informs us that in Big Data analysis, we must pay attention to (1) data receiving module, (2) data preprocessing module, (3) data integration module, (4) data preparation module, (5) data analytics module and finally to (6) results publishing module. The right combination of Big Data technologies must be selected in order to reach scalable services, better performance and accuracy. The question of horizontal and vertical scaling is a critical issue in the selection of Big Data analytics platforms. Each approach has advantages and disadvantages. Human-to-human, human-to-machine and machine-to-machine interactions need always special attention. Smart city applications must be tailored to local conditions (see e.g. Oussos et al., 2018, p. 431, Osman 2019, Kaivo-oja et al. 2019).

We can conclude that there are many activities in the field Big Data analytics. Next, we shall discuss more about innovation management and Big Data challenges.

3. Open innovation tools and crowdsourcing with Big Data

There are various Big Data Analysis techniques:

- Data mining
- Web mining
- Visualization methods
- Machine learning tools
- Optimization methods
- Social network analysis and
- Mixed Method Approach.

Fig. 12, we present basic dimensions of data analytics and quantitative and qualitative research. Our message is here that there is no reason to have “black-and-white” thinking concerning Big and Small Data or quantitative and qualitative data analyses. Both approaches can provide new insights to other “side of coin”. Methodological pluralism is needed to boost better foresight and innovation analyses.

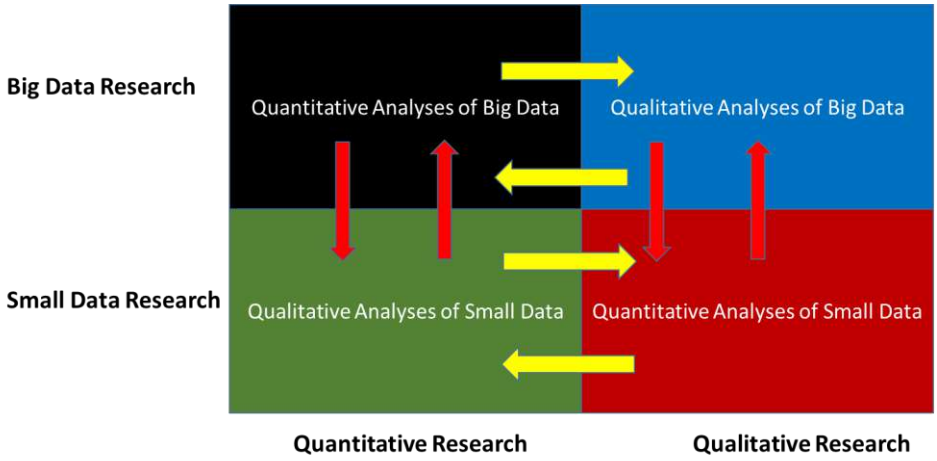

Figure 12. Big Data, Small Data, Qualitative Research and Qualitative Research.

As in modern days technological change is intensive and product life cycles are considerably shortened, companies engage in open innovation to cope with the increased risks and costs of R&D projects and to acquire the needed resources from external actors (Chesbrough, 2017, Bzhalava and Cantner, 2018) Moreover, firms combine foresight practice with innovation activities to scan external environment, detect technological opportunities and develop groundbreaking innovations (Kaivo-oja, 2012; Kaivo-oja et al., 2015) but many companies face problems to process massive amounts of external information, anticipating signals of impending technological changes and adjusting their innovation activities accordingly. As human decision makers have limited knowledge and attention span as well as show cognitive and motivational biases, they often make errors in

decision making (Goodwin and Wright, 2004, De Dreu, at al., 2008, Ocasio, 2011, van Knippenberg et al., 2015, Piezunka and Dahlander 2015). Moreover, many incumbent firms face problems to anticipate disruptive technology changes because they often originate outside companies' area of expertise (Day and Schoemaker, 2004; Heuschneider and Hersta, 2016). To deal with these issues, organizations increasingly seek to employ automated foresight and innovation management tools.

Previous studies develop text mining methods to analyze large volumes of literature and patent data as well as web news to automatically identify meaningful patterns and trends in technological development, to strengthen peripheral vision and, as a result, to advance strategic planning in organizations (Yoon, 2012; Thorleuchter and Van den Poel, 2013; Thorleuchter at al., 2014; Choi et al., 2012, Guo et al., 2016, Kim et al., 2016). These methods include text segmentation, term association, cluster generation, topic identification, and information mapping for deriving technology intelligence from scientific articles and patent data (Tseng et al., 2007). Moreover, to automate weak signal detection tools, the keyword-based text mining approach (e.g. keyword cluster map, keyword intensity map and keyword relationship map) is proposed (Yoon, 2012, Kim et al., 2016). The keyword cluster map approach identifies future-related homogeneous topics by classifying frequently co-occur keywords into clusters. Keyword intensity map assesses strength of the keywords for each topic by measuring keyword frequency (visibility) and the degree of diffusion based on document frequency (Yoon, 2012; Kim et al., 2016). Based on this, short-term, medium-term and long-term signals are classified, and keyword relationship map is built to categorize keywords into technology roadmaps. Furthermore, Thorleuchter and Van den Poel (2013) employ a latent semantic indexing text mining method, which considers the aspects of meaning and identifies similar textual patterns in different contexts, to recognize weak signals in the internet-based environment and to support strategic decision making.

As in modern days technological change is intensive and product life cycles are considerably shortened, firms have increased the degree of their openness in innovation to source knowledge from a wide set of external actors and, in this way, to develop innovation relatively quickly and inexpensively (Chesbrough, 2017). However, given that firms' absorptive capacity or knowledge base is limited, they face challenges to identify and utilize knowledge from new and unfamiliar technological areas (Laursen and Salter, 2006; Monteiro et al., 2017). This, in turn, negatively affects a firm's innovation performance. To cope with this issue, previous research proposes text mining tools to automatically select external partners for a given problem. In particular, Subject-Action-Object semantic analysis is developed to identify technological trajectory and opportunity as well as to search solutions in patent or internet-based environment by organizing text in a problem–solution format where the action–object (AO) states the problem and the subject (S) forms the solution (Choi et al., 2012, Guo et al., 2016).

Moreover, to improve market intelligence, the topic modeling method of text mining is used to calculate business proximity between firms for potential merging and acquisition. In particularly, Shi at al. (2016) analyze publicly available business descriptions of startups to identify potential successful matches between firms based on similarity of their business topic distribution. Text mining methods are also proposed for developing automated tools for smart specialization strategy management and analysis. In particularly, Bzhalava et al. (2018), use topic modeling method of text mining to identify key economic areas in which entrepreneurs from different EU regions see new business opportunities and in which areas they identify possibilities of smart specialization. The study shows that text mining can be used to identify dominant topics and trends in startup

entrepreneurial activities across regions and to develop automated management tools and planning approaches for Smart Specialization Strategy.

All these new analytical tools can be used in crowdsourcing process where a lot of crowdsourced data is available. Today crowdsourcing challenges are fast emerging as an effective tool for solving complex innovation problems. Howe (Howe, 2006, 2008, 2009) has classified the applications of crowdsourcing into the following four categories:

(1) *Collective intelligence* (or wisdom of the crowd). People (in a crowd) solving problems and providing new insights and ideas leading to product, process, or service innovations (e.g., see [18]).

(2) *Crowd creation* (or user-generated content). People creating various types of content and sharing it with others for free or for a small fee.

(3) *Crowd voting*. People giving their opinion and ratings on ideas, products, or services, as well as parsing, evaluating, and filtering information presented to them.

(4) *Crowdfunding*. This is a special model in which people can raise money for investment, donations, or for micro-lending of funds.

It is important to understand that there are many alternative crowdsourcing techniques. In Fig. 13 we have outline four alternatives: (1) Crowd-voting, (2) idea crowdsourcing, (3) micro-task crowdsourcing and (4) solution crowdsourcing.

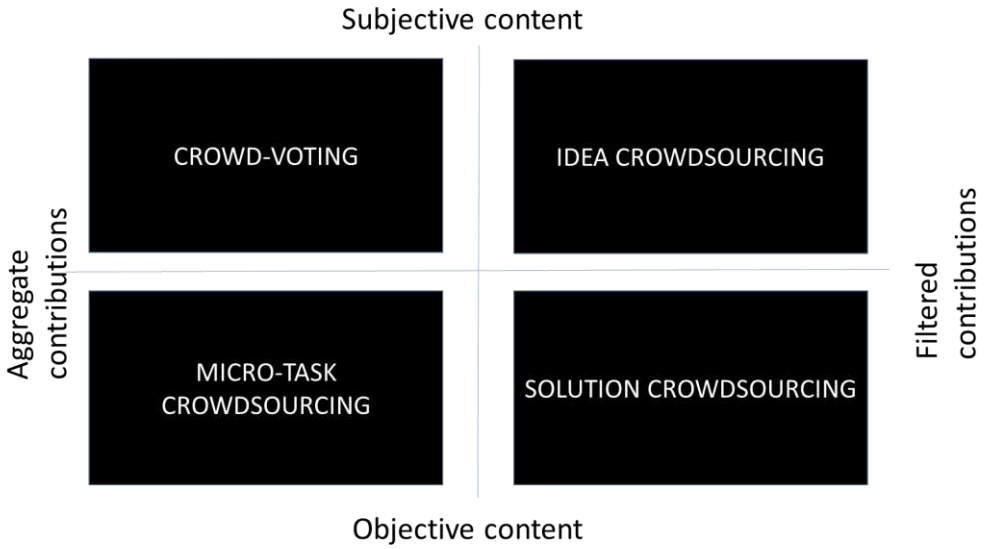

Figure 13. Crowdsourcing alternatives (Prpic´ et al., 2015, p. 79).

Crowd-voting means that an organization requests choices between alternatives and then aggregates the votes. *Idea crowdsourcing* means that an organization invites opinions for small or big questions and then evaluates the proposed ideas. *Micro-task crowdsourcing* means that an organization breaks problems into smaller jobs and then re-assembles the completed tasks. *Solution crowdsourcing* means that an organization invites and tests contributions for very specific problems and then adopts the best non-falsifiable solutions. (see Prpic´ et al., 2015).

When applying Crowdsourcing Delphi methodology (see Santonen and Kaivo-oja, 2019) we can make a choice concerning these four alternatives on the basis of (1) type of contribution (objective

or subjective) and (2) content (aggregate or filtered). Traditionally, the term 'crowd' has been used almost exclusively in the social context of people who are self-organized around a common purpose, idea, invention, innovation, emotion, or experience. Today firms and public agencies often refer to crowds in discussions of how collections of individuals can be engaged for organizational purposes.

Crowdsourcing means the use of information technologies to outsource business or other thinking responsibilities to crowds. Crowdsourcing can significantly influence an organization´s ability to leverage previously unattainable resources to build competitive advantage and novel products and services (Prpic´ et al., 2015). Furthermore Ross Dawson (2010) identified the following crowdsourcing platforms, which are in Table 3 aligned to Prpic´ et al. 2015 crowdsourcing alternatives a.k.a. types.

Table 3. Crowdsourcing platforms and types (Santonen and Kaivo-oja, 2019)

Platform type	Description	CS type
Distributed innovation platforms	Support innovation processes that cross organisational boundaries or take place entirely outside an organisation. Some notable resources include Innovation Exchange, which allows agencies and organisations to present innovation challenges to a community of innovators.	Solution CS
Idea platforms	Are used within a company context to be able to gather and filter and source ideas that are proposed. Often managers or workers submit ideas or proposals for cost savings, or new products, or new services, or process efficiencies, and then they collectively assess and rate and vote on and select and evolve and refine and build on those ideas to become the innovation that will drive that organisation forward..	Idea CS
Innovation prizes and Competition platform	Are challenges designed to catalyse new thinking and ingenuity. Typically, anybody anywhere can enter their own projects and ideas, and others can vote on them and build on them and use the wisdom of the crowd to make them more effective, and from all of those submissions somebody wins a big sum of money.	Idea CS Idea-voting
Content markets	Are platforms where people submit their content for people to purchase. A typical example is Threadless, which allows people to submit designs for t-shirts, with the community voting for the best ones to be made into actual t-shirts which can be ordered (Brabham, 2010).	Idea-voting
Prediction markets	Bring together many opinions to predict the future, often based on "stockmarket-type" mechanisms, which provide a value of a particular prediction that you can buy or sell to make points or potentially money as a result of it going the way you correctly predict.	Solution CS

Thus, we can note that distributed innovation platforms, idea platforms, innovation prizes and competition platforms, content markets platforms and prediction markets platforms can build on Big Data analytics. There is a logical link between these platforms.

The main strategic strength of the crowdsourcing model is that it can bring together large number of diverse people from all over the world to focus on solving a problem. Big Data can be used for collecting data from crowds. Big Data analytics can help to solve these challenges. Recent advances in digital technologies and services are fundamentally transforming the innovation practices of organizations. These technologies have blurred the boundaries of innovation processes and in turn have provided organizations with unprecedented opportunities in their search for ideas, inventions and innovations. For example, natural language processing and data matching have helped to structure data. New modelling tools like decision trees, artificial neural networks, support vector matrices, deep learning tools, ensemble algorithms, regularization methods and tools of Bayesian statistics have helped to model data in better ways. Also assessing methods have been improved by better robustness methods and performance methods (see Blazquez and Domenech, 2018).

Today in the field of Big Data analytics typical machine-learning algorithms and applications (see Mahdavinejad et al., 2018) are the following:

1. Classification methods, which lead to prediction and increase Data abbreviation.
2. Clustering methods, which lead to prediction and increase Data abbreviation.
3. Linear Regression methods, which lead to Real Time Prediction with reducing amount of data.
4. Support Vector Regression methods which are conventional forecasting methods.
5. Classification and Regression Trees analyses, which can be used in real time prediction and assessing consumption patterns.
6. K-Nearest Neighbors methodology, which helps us to analyze consumption patterns and improve efficiency of the learned metrics.
7. Naive Bayes methodology, which help us to analyse safety issue and improve Safety Metrics and understand consumption patterns and estimate the numbers of nodes of networks.
8. K-Means methodology, which is good tool for (1) outlier detection, (2) fraud detection, (3) analyze small data sets, (4) forecasting consumption, (5) identify consumption patterns, (6) stream data analyze and perform Weak Signal and Wild Card Analysis (WiWe methodology).
9. Density-Based Clustering, which helps in labeling data, fraud detection, and in the analyses of consumption patterns.
10. Feed Forward Neural Network analysis, which help in (1) reducing consumption, (2) in consumption pattern analyses, (3) in forecasting the states of key elements, (4) overcoming the redundant data and information.
11. Principal Component Analysis, which is tool for fault detection analyses.
12. Canonical Correlation Analysis => which is tool for fault detection analyses.
13. One-class Support Vector Machines methodology, which is tool for fraud detection and analysing emerging anomalies in the big data.

Big Data can be used in straightforward forecasting when we have a lot of time series data available. The source of the expression, ‘‘It’s difficult to make predictions, especially about the future,’’ is uncertain. This saying has been attributed variously to Niels Bohr, Samuel Goldwyn, Yogi Berra, and Mark Twain (Flostrand, 2017, p. 223).

On the basis of Big Data we can try to fight against this well-known expression. We can always say: “*Let Big Data speak to us*”. However, in many complex situations managers and leaders want to say and give their insights to other people and their organizations. Sometimes there are such messy problems that require expertise or diversity of opinion. In the field of foresight research much used methodology is the Delphi methodology. Delphi methodology works best when the foresight expert or Delphi manager has reasonable access to a community of experts who are willing participants. Delphi expert process requires a significant time investment. While the data collection phase itself is just a matter of mail exchanges and value consolidating so it passes quickly, the preliminary actions take much more time. It is quite easy to imagine a group of experts in a given special field, but Delphi methodology is only an available option when these experts individually have some interest in knowing the results of the research the foresight expert is proposing and have no aversions to sharing the knowledge they have. Delphi is also optimal when the researcher is seeking to identify and prioritize issues and when the foresight requires estimations of certainty. In many situations it is interesting to know how a median or average expert analyses the alternatives of decision-making.

Fig. 14 can help us to decide how to make a choice between Delphi methodology and crowdsourcing techniques.

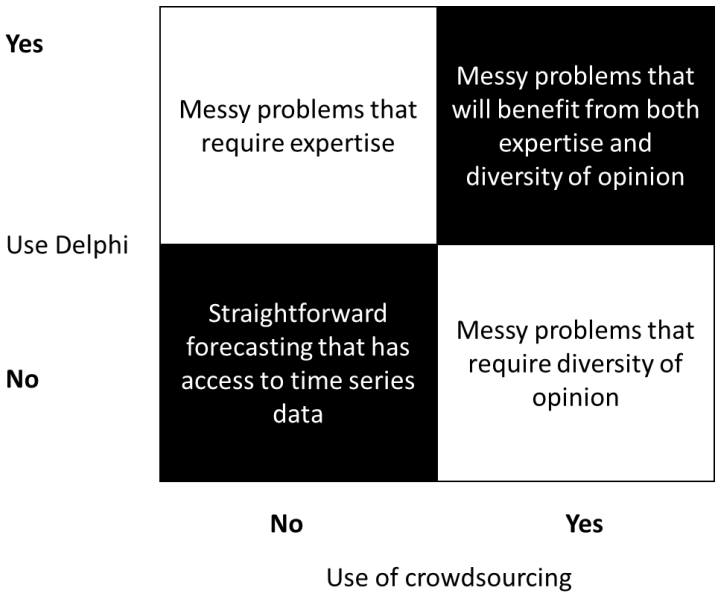

Figure 14. Delphi and crowdsourcing: Deciding between the two methods (Flostrand, 2017, p. 234).

4. Knowledge management aspects of foresight analyses with Big Data

In this section we summarize key aspects of knowledge management of foresight research with Big Data. Information and communication technology have quickly transformed the way entire sectors and industries access their data, documents, and information. Big Data challenge is one key element of ICT transformation and changes in knowledge management. In Fig. 15 a typical foresight process is outlined. We can see that the selection of foresight methods is a critical phase when we aim to perform Big Data analytics (see van der Duin, 2016, Kaivo-oja and Roth, 2019).

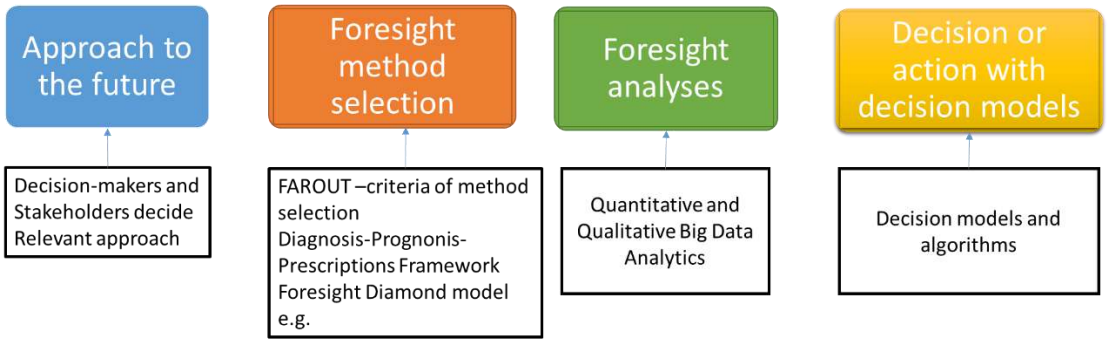

Figure 15. Typical foresight process with Big Data.

The most typical method in foresight analyses is scenario approach. In Fig. 16 we see foresight and scenario typology. We can present predictive scenarios, explorative scenarios and normative scenarios. Under different scenario types we can see specific scenarios. Big Data can be applied in forecasting, fore-sighting and in goal rational planning and management with normative scenario analyses.

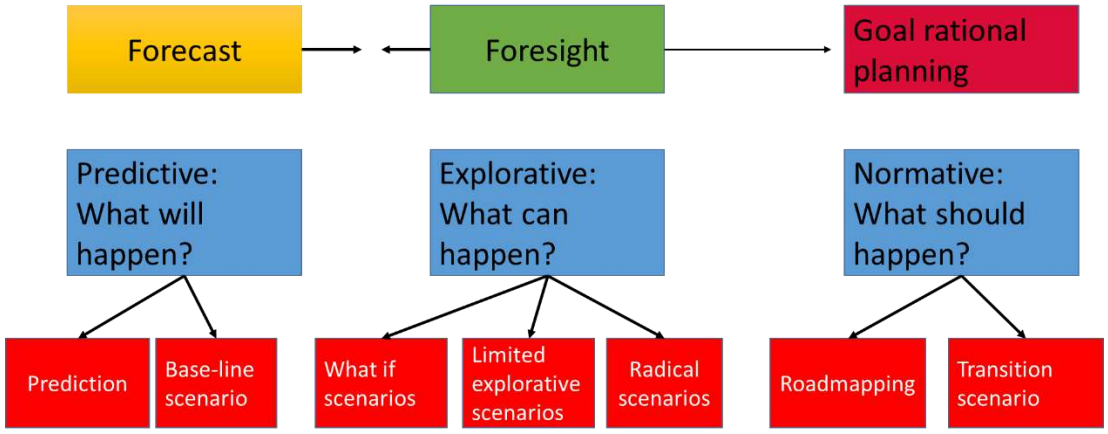

Figure 16. Foresight and scenarios typology (Nekkers 2014).

Big Data can be used in backcasting and forecasting scenario analyses. Forecasting scenarios can be build when we look at the past trend and present conditions then extrapolate the future values. In forecasting scenario process, we simply do the following steps:

- First, we collect past and present data or big data.
- Secondly, we take alternative future possibilities/ events in account for estimation, and
- Finally, we do extrapolation/ estimation for future with different statistical assumptions.

In Fig. 17 we present the basic logic of forecasting scenario analysis with Big Data analytics.

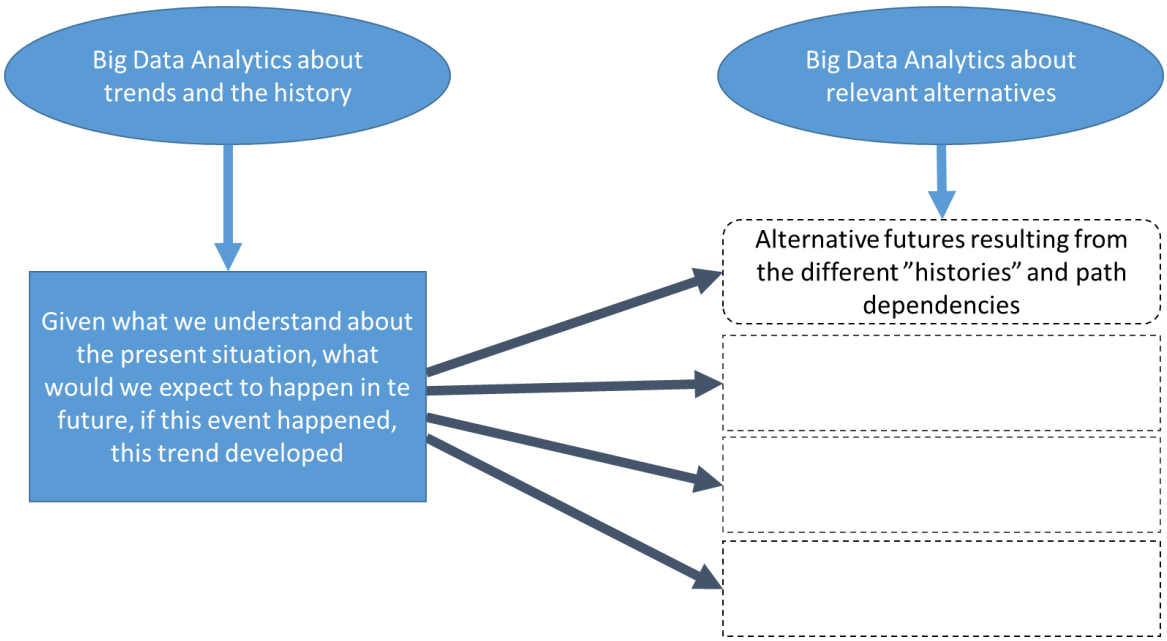

Figure 17. The basic logic of forecasting scenario analysis with Big Data analytics.

When we preform backcasting scenario analysis, we move almost opposite to forecasting. In this foresight technique we start with the point where we want to be in the future and then try to adjust everything in the present systems according to the goal or the target we want to achieve in the future. If we summarize the steps in which backcasting scenario analysis, it can be performed as follows:

- First, we define and establish the targets for future.
- Secondly, we collect data from past and present condition/ events.
- Thirdly, we analyze the collected data to finalize what changes and amendments are required to the present conditions so that ultimately it comes coordinated with the vision established.
- Thirdly, we establish a policy and strategy to implement the changes required

In Fig. 18 we present the basic logic of backcasting scenario analysis with Big Data analytics.

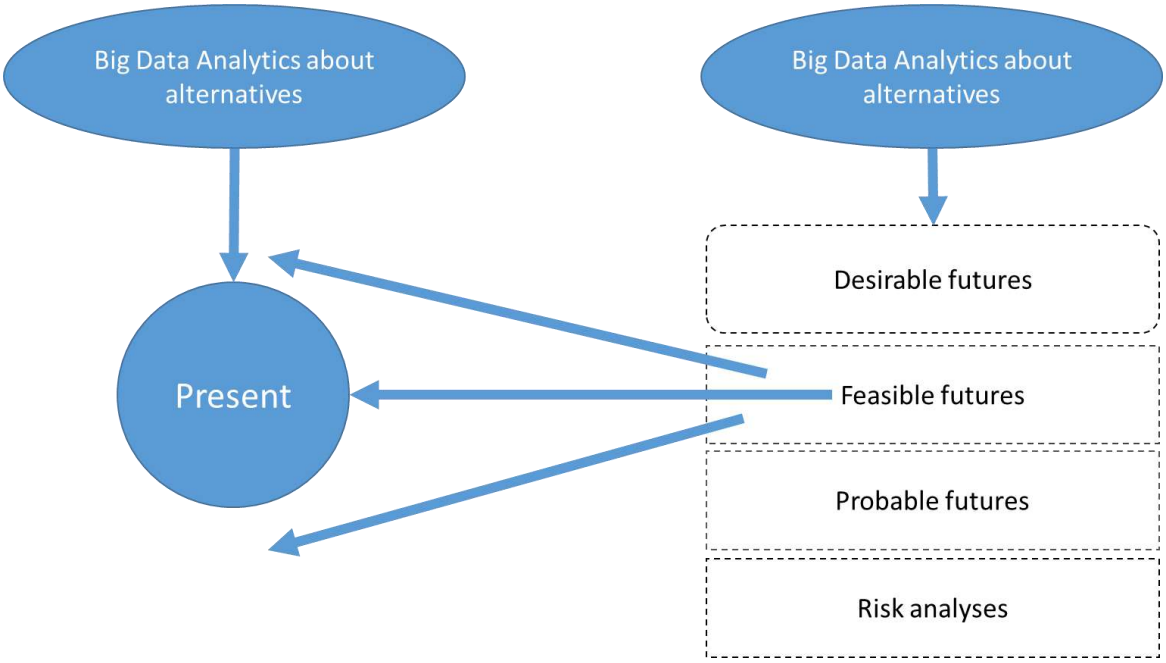

Figure 18. The basic logic of backcasting scenario analysis with Big Data analytics.

One way to summarize knowledge management challenges of Big Data is to link big data analytics to the Cynefin model of David Snowden, who has shown originality in his work on knowledge management. The Cynefin model includes four domains of systems: (1) Simple systems, (2) complicated systems, (3) complex systems and (4) chaotic systems. Big Data can provide value added in all these domains of systems. The degree of complexity varies, and foresight analyses needed in different decision-making situations depend on the degree of complexity of systems. For example, in chaotic systemic conditions, we do not need much foresight processes, but actions. Following the approach of Snowden (2002) and Snowden & Kurtz (2003) there are four decision models:

(1) Simple systems and Known systems where Sense - Categorise – Respond -model works;
(2) Complicated systems and for systems which are Knowable, where Sense – Analyse - Respond - model works;
(3) Complex systems and for systems which are Unknowable, complex, where Probe – Sense - Respond -model works; and
(4) Chaotic systems and Unknowable, chaotic: Act – Sense - Respond – model works.

These four decision and knowledge management models can be key elements of pragmatic governance and knowledge management of Big Data (Table 4).

Table 4. The Cynefin framework and the role of Big Data.

System	Knowledge management approach	The function of Big Data
Simple systems	Sense-Categorise-Respond	Create the Data base for benchmarking analyses of simple systems
Complicated system	Sense-Analyse-Respond	Create the Data base for statistical Big Data analyses (trend analysis, probability analysis, risk analysis and cause and response analyses)
Complex system	Probe-Sense-Respond	Create the Data base for complexity analyses like for morphology, scenario analyses and synergy analyses
Chaotic system	Act-Sense-Respond	Create the Data base for real time urgent crisis analyses and real-time monitoring analyses

There are various good books to master these key functions of Big Data analytics. For benchmarking studies good source is Tim Stapenhurst´s (2009) book "The Benchmarking Book". For more complicated data analyses a good basic source is John W. Freeman´s (2014) book "Data Smart", Joel Grus´s (2015) book "Data Science from Scratch" and Bart Baesens´s (2014) insightful book "Analytics in a Big Data World". In the field of complexity analysis good books are scenario analysis book "the Scenario Planning Handbook" by Bill Ralston and Ian Wilson (2006) and an updated article "Defining Scenario" by Spaniol and Rowland (2019). There are many other good sources of Big Data foresight research.

5. E-commerce tools and algorithms with Big Data and foresight processes

In the field of Big Data analytics, general model of foresight is the following, a process, where diagnosis phase, prognosis phase and prescription phase are implemented (see Fig. 19):

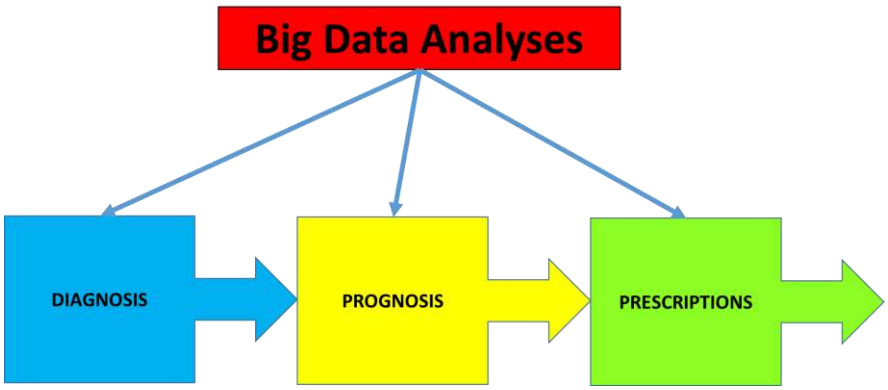

Figure 19. Big Data analyses and a typical foresight process.

Diagnosis phase is typically including descriptive statistical analyses with Big Data. Prognosis phase is typically including predictive and other future-oriented foresight analyses. Prescriptions phase is typically including decision model use and strategic and policy recommendations from diagnosis and prognosis phases. In all phases of foresight Big Data analytics can be useful. We can expect that Big Data analytics will be more extensively linked to many Artificial Intelligence applications. Already now we have various extensions in the fields of robotics, marketing, artificial life, health care etc. (Frankish & Ramsey 2014). Probably, so called *Digital Twin -approach*, which integrates physical systems to digital systems, will be more popular in the future (see Kaivo-oja et al., 2019).

Matching Tools with Big Data

The best way to match demand with supply in a marketplace naturally depends on the specifics of the marketplace in question. In different market places different matching algorithms are useful. Motivation to development matching algorithms is linked to strategic interests of stakeholders in the marketplace. Typical motivational factors are (1) increase bargaining power of an agent in the marketplace, (2) balance supply side factors, (3) balance demand side factors and (4) increase efficiency of market transactions and (5) save scarce resources in the marketplace transactions. The buyer can improve performance for example, by browsing the marketplace and selecting listings of interest, or by reviewing bids/proposals and selecting a preferred one(s). In the marketplace, offers are organized into some sort of catalogue structure with categories that the buyer can navigate. Browsing functionality allows buyers to identify specific offers they may be interested in (and matching themselves to sellers) and to signal this to the marketplace.

Demand side. Signaling buying interests to sellers is important for matching algorithms development. Sophisticated marketplaces may subsequently use this information for retargeting its offerings and supply side products. Within a given category, the marketplace may choose to order listings according to some relevance function.

Supply side. In marketplaces where products and services tend to need a high degree of customization, it's common for the buyer to provide details of what they're looking for (a request) and for potential sellers to provide proposals. Signaling the needs to provide proposals is important feature of supply side management and operations. The marketplace can then notify sellers about that request in a number of ways:

1. *Buyer invites* — the buyer invites specific sellers to respond (a matching action),
2. *Marketplace invites* — the marketplace invites specific sellers to respond (a matching action),
3. *Public listing* — the marketplace may allow sellers to browse or search for active requests and the sellers match themselves to the requests,
4. *Broad updates* — the marketplace sends an email to many or all sellers about the latest requests, often in daily/weekly batches (if done intelligently, the marketplace will be matching requests to relevant sellers)

The marketplace may then allow any seller to respond with a proposal or restrict it to those that have been specifically invited by some pre-selected criteria. In many marketplaces, sellers are selective about which requests they respond to. They are effectively matching themselves with suitable buyers. The seller can always improve performance by selecting categories for their listing(s) or by choosing which buyer requests to respond to. The marketplace can improve its performance by using algorithms/heuristics, whether automated or manual. Both manual and automated methods can work depending on skills of algorithm applications.

Matching algorithms are algorithms used to solve graph matching problems in graph theory. In theoretical level, a matching problem arises when a set of edges must be drawn that do not share any vertices. Graph matching problems are very common in daily activities. We all know that there are online matchmaking and dating sites, product and service matching, to medical residency placement programs, matching algorithms are used in pricing and purchasing, areas spanning scheduling, planning, pairing of vertices, and network flows. A few examples of algorithms at work are personalized services (like gym and health services), search help (like using Google services), matching functions (like in Uber taxi and hotel services), predictions (like stock performance predictions) and optimization tasks (like vehicle routing). In matching typically some filters and rankings are used. It is good to remember that all filtering algorithms operate on human input.

Two famous properties of matching algorithms are called (1) augmenting paths and (2) alternating paths, which are used to quickly determine whether a graph (1) contains a maximum, or contains a minimum, (3) matching, or (4) the matching can be further improved. Majority of realistic matching problems are much more complex than those presented above with mathematical simplicity. Added complexity often stems from graph labeling, where edges or vertices labeled with quantitative attributes, such as weights, costs, preferences or any other specifications, which adds constraints to potential matches. The more complex constraints are presented, more challenging is to identify and select good matches.

In matching process there can be both hard and soft filters, which constrain supply and demand in the marketplace.

Hard filters refer to criteria that help in defining a list of preferences that can be considered for a particular transaction. Some typical hard filters associated with many such platforms are: (1) Type of requested service, (2) location where the service must be performed, (3) time of service – i.e. scheduling preference by customer, (4) pricing principles, (5) specific preferences (like male/female therapist young or senior expert). (see e.g. Jungleworks 2019).

Soft filters will be applied after applying the hard filters. Matching logic is configured to apply certain soft filters to further narrow down or sort the list of matches in a particular business order. Typical soft filters are: (1) ETA (Tiers) and Request Dissemination, (2) Preferred Service Providers, (3) Curation/Ratings, (4) Freelance vs. Contracted, (5) Load Balancing and (6) Route Optimization. Both hard and soft filters are based on algorithms and available small or big data. (see e.g. Jungleworks 2019).

6. Platform economy and platform approach

During recent decade the rise of the platform economy has been phenomenal. Corporations like Amazon, Apple, Google, and Facebook together have an almost $3 trillion market cap and make up nearly 11% of the S&P 500 (see Fourkind 2019). We can point out that platform economy is relevant issue to discuss. We must remember that there are smaller and bigger platforms in the economy. Furthermore, there is still a strong tendency to consider platforms only from a technical viewpoint, consisting of APIs that are offered to the outside world, and paying less attention to the whole breadth of cultural, social, economic and governance aspects involved. The rise of platform economy is closely linked to Big Data developments.

There is an analogy between market square and the concept of platform. As we know, the market square enables food producers and consumers to interact without external intermediaries. For producers, it would be time- and resource-consuming to find all customers and present offerings for everybody separately. Also for consumers, it would be similarly inefficient to find various producers one by one. This situation is relevant for foresight and anticipation markets, where consumers and producers want to share knowledge intensive services and products. The market square as a platform hosts several market stalls that enable various producers to make their business offerings visible and attract consumers to see what is available and buy something and in final stage, in essence, interact with the platform. The market square is also synergic concept. The market square in itself accumulates growth. The more consumers there are present, the more producers are willing to participate, and vice versa; the more producers there are, the more consumers it attracts. In similar way, platform is a synergic concept. (Fourkind, 2019).

Inside platform we can also identify: (1) Platform architecture, (2) ecosystem architecture, and (3) apps architecture (Tiwana 2014).

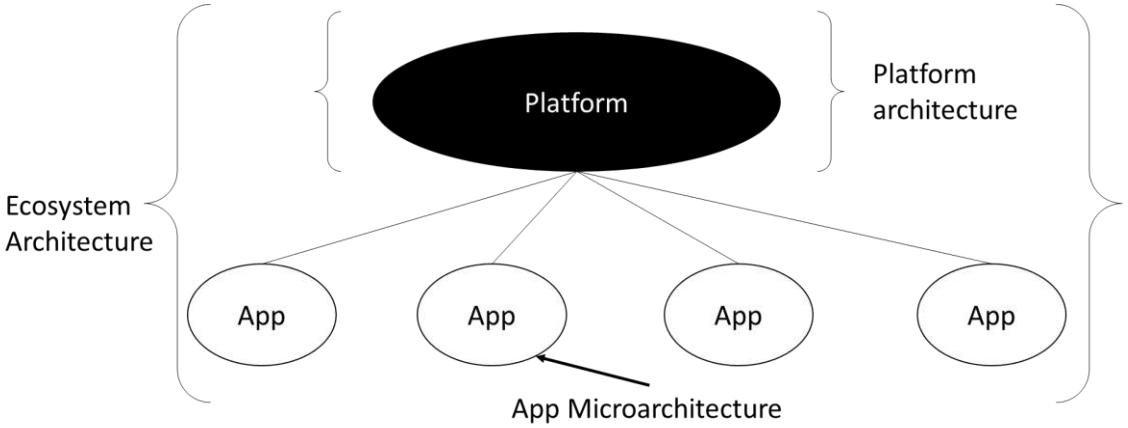

Figure 20. Ecosystem architecture as comprised of platform architecture and app microarchitecture (Tiwana 2014, p. 85).

Typical key elements of platform are: (1) Network, (2) community, (3) marketplace, (4) data and (5) infrastructure. For foresight and organizational purposes we can define these three architectures.

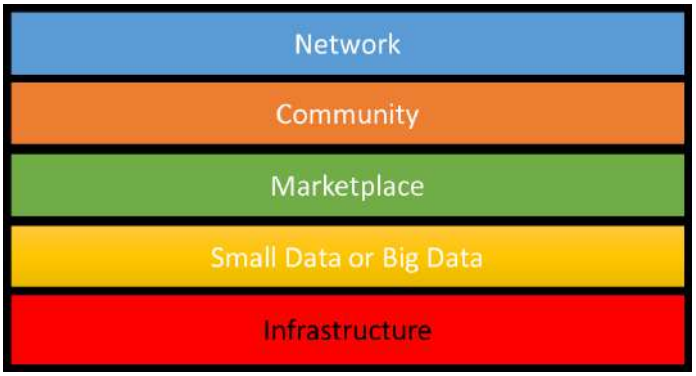

Figure 21. Key five elements of a platform (Choudary 2015).

If we plan platforms of Big Data foresight, we must define these five elements: What are critical networks we are serving? (2) What is our core community?, (3) what market places relevant for us? (4) what are data and data libraries need to have?; and (5) what kind of critical infrastructure we need?

7. Integration challenges in Big Data field: Ethical codes and other integration challenges

7.1. Ethical aspects

7.1.1. Issues of ownership and transparency

Nowadays, people are more and more aware of privacy and social media risks. Big Data-ethics will be critical topic of public and corporate ethics discussions. It has been argued that Big Data has the effect of shifting the focus of ethics away from the individual's ability to make moral judgments, and instead requires careful examination of those who have control over Big Data (Zwitter, 2014; Herschel & Miori, 2017). The following six principles are currently attributed to Big Data Ethics: (1) Ownership of small or big data - Individuals own their own data or sell their data with a contract. (2) Data Transaction Transparency - If individuals´ personal data is used, they should have transparent access to the algorithm design used to generate aggregate data sets, (3) Consent of data - If an individual or legal entity would like to use their personal data, one needs to be informed and explicitly expressed consent of what personal data moves to whom, when, how and for what purpose from the owner of the data, (4) Privacy of citizens - If data transactions occur all reasonable effort needs to be made to preserve privacy, (5) Currency – All individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions and (6) Openness - Aggregate data sets and data lakes should be freely available (see https://en.wikipedia.org/wiki/Big_data_ethics).

We can add other related 'Big Issues for Big Data' which needs to be added to the Big Data-ethics discussion (Raleigh, 2019): (7) Avoiding algorithm bias – Algorithm often unintendedly exacerbates underlying biases of real-world data and thereby harms specific populations, (8) Data longetivity – As data gains value through use, its reliability over the long-term becomes more important; this create emerging issues in cases of e.g. bankruptcy or decisions to discontinue management of data or data APIs.

On the other hand, data analytics can help us to manage some big risks like pandemic and climate change risks. Also, SMEs can have smarter business models and platforms with Big Data analytics. We are therefore also faced with dilemmas in which ethical boundaries might prevent us from achieving something we can inter-subjectively agree as valuable (Wiren, 2019). Value search, value configuration and value delivery can be improved by the five Vs of Big Data. Also, governments and academia and civil society organizations can improve their services and value delivery to citizens by the services and good based on Big Data analytics. McKinsey (2018) argues that '*Smart city applications can improve some key quality-of-life indicators by 10 to 30 percent'.* As an example of heeding this call, the Finnish city of Turku is developing big data and 'world class data science resources' as a local strategic flagship project (Piippo, 2019).

7.1.2. Data ecology issues: misinformation, overload, pollution

Another set of ethics-related challenges for big data research in general and big data foresight in particular consists in – often well-meaning – attempts to clean, curate, or preserve certain forms of big data for the sake of this or that higher good or stakeholder group (Shin and Choi, 2015).

Typical examples of this approach may be referred to under the umbrella terms information ecology or data ecology. These approaches aim at combining key concepts from ecological and information sciences so as to transfer and apply natural environment preservation techniques to the IT-related issues and challenges such as information overload (Floridi, 2012; de Mauro et al, 2015; Saxena and Lamest, 2018), information pollution (Baeth and Aktas, 2019, Helbing, 2019), and misinformation (Baeth and Aktas, 2019) including the notorious fake news (Vargo et al., 2018; Giglietto and Mesjasz, in press). The baseline idea behind these approaches boils down to the assumption that the information processing capacity of human brains and even larger human communities or organizations are necessarily limited and therefore require large-scale technological and socio-technological support with dealing with an ever-bigger data environment, sometimes referred to, in analogy to the ecosphere, as datasphere or infosphere (Kamppinen, 1998; Floridi, 2014, Beranger, 2016). Ambitions of protecting smaller or bigger data- or infosphere ecosystems are then inevitable.

The issues at stake, and thus the justifications for intervention, are manifold ranging from matters of transparency in health systems to problems of political propaganda. Next to problems with data leaks or information veracity, even less obvious cases of information pollution or overload such as hate speech filtering or hype-related media space saturation and the corresponding media space biases (Chun, 2016; Barnes, 2018) remain critical issues that require further attention and, according to information or data ecologists, also protectionist intervention. Needless to say, that recent big data research has also identified gender biases in the ways of how big data comes about in first place (Jia et al., 2016; Lansdall-Welfare et al., 2017).

Against this backdrop, data ecologists currently work towards strategies to mitigate the above and many further issues related to the pollution, and data ecology pioneers have already raised issues of limits to the growth of big dataspheres explicitly using the analogy to the classical natural ecology-relatedl limits to growth (Bergé and Grumbach, 2017).

Overall, quality decisions need quality data. One of the main problems following the process of big data system assimilation is the “sanity of data”. It is thus necessary to verify data quality before perform big data analysis (Bautista Villalpando et al., 2014). A data cleansing process of identifying and removing inaccuracies, incompleteness and inconsistencies of data (Kumar and Chadha, 2012) is therefore pivotal to BDA.

In this context, one key challenge of big data foresight is to raise awareness for the dark sides of the above well-meaning and probably also well-justified interventions in the data ecology in order to avoid that past and present forms of political correctness and other forms of social desirability corrupt the quality of big data views of the future. This issue boils down to the basic question of how much morality is there, and should there be, in big data?

7.2. Further big data integration challenges

We can summarize our discussion in this report noting that Big Data analytics is leading us to new era of knowledge discovery. Knowledge discovery processes have been based on Small Data, but Big Data analytics probably changes the processes of knowledge discovery. Our analysis indicates that Small Data platforms are still, at least, relevant for Big Data platforms, because classifications of data are not changing fast, but it may take time to create new and better classifications. We can partly build Big Data platform on the basis of existing platform, but not fully. Some Small Data platforms are not suitable for Big Data analytics. Novel and new innovative approaches are also needed. Methodologically, we shall need special methods and tools for quantitative data and qualitative data. We can expect that transform process from Business Intelligence to Big Data analytics is going to be challenging for many organizations and especially for organizational cultures. For organizations and their stakeholders, it is important to understand that cultural transformation function is a critical factor for the success of Big Data analytics and knowledge discovery process. There is need for national and international partners inclusion (e.g. Facebook is not sharing data nationally) and this international integration challenge is another aspect of integration. There can be also "Big Brother"-problems, threats and cybersecurity as the infamous case of Cambridge Analytica shows us.

There is always a problem in ICT-systems where systems are not technically integrated concepts and applications. This can be also a problem for platforms development in European Union. There is also possibility of the existence of worthless, highly integrated Big Data systems that are not creating any value to end-users and consumers.

It is also possible that quantitative and qualitative data are not matching with each other – and there is a question - how to integrate them? In field of foresight, Numbers and Narratives problem is still a problem: How to find an integrative bridge?

There can be also integration problems in data and information delivery. Big Data analyses are not always delivered and communicated to the government and other Quartet Helix partners with open innovation approach. This may create problems for competition policy in the European Union. Key question in this integration context is: Is National Open Innovation System (NOIS) getting Big Data results or not?

Countries succeeding in a global and competitive economy will be those whose innovation systems utilizes Big Data to generate continuous and updated information, and whose innovation ecosystems are able to adapt to these results dynamically.

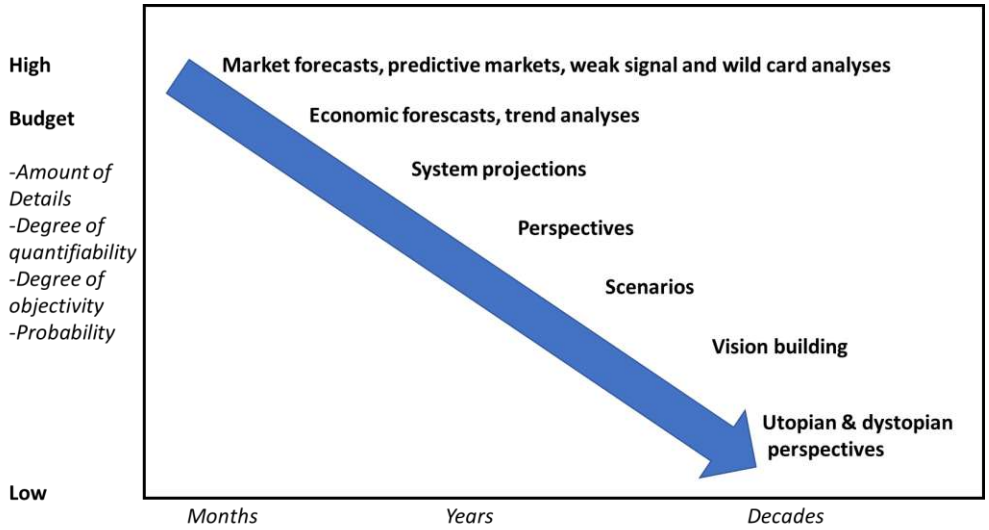

Figure 22. Time perspectives: months, years and decades

In Fig. 22 we can see different kind of foresight analyses. For decision-makers it is always important to understand the kind of foresight with which they are dealing. This forms the basis of common understanding in a joint decision-making situation. This is relevant aspect also in the contact of Big Data foresight. Big Data is not covering information for future months, years or decades, although it can create data and information base for futures-oriented analyses and foresight analysis. It is very important to create common language to understand the key results of foresight analyses.

8. Methodological and theoretical challenges in the big data field

When he was still Editor-in-Chief of Wired, C. Anderson (2008) heralded "The end of theory" as well as of research methodology as we know it in the age of big data. The basic idea behind his provocative thesis was that statistical models or speculative theories could become obsolete if complete datasets would provide researcher with complete information about the research issue at stake. Whereas previous models and theories have been, at best, incomplete (if not wrong), true to Anderson, big data seemed to allow for the drawing of both true and complete conclusions using sets of computational, algorithmic applications geared towards big data.

On the other hand, Anderson himself conceded that the mere ability to derive conclusions from the analysis of patterns in big data sets might also generate new research questions and thus be useful for the generation of new theories.

Just a few years after Anderson's seminal article, it has already been consensus that big data and the corresponding new forms of data analysis not only challenge established ways of doing research, but also hold the promise of the development of entirely new forms of theories, methodologies, and paradigms. A prominent topos in this context has been the observed shift from knowledge-driven to data-driven research in the digital humanities as much as in the digital social sciences (Kitchin, 2014).

As detailed in Hey et al. (2009), leading figures of the data revolution have early anticipated the emergence of a fourth research paradigm. According to this idea, the first paradigm of modern science has been descriptive experimental science; the second paradigm model-oriented theoretical science; the third paradigm that of computational simulation; and the forth paradigm that of data-intensive exploratory science.

The following Table 5 illustrates the paradigmatic evolution of scientific research

Table 5. The paradigmatic evolution of scientific research (Kitchin, 2014; Hey et al., 2009).

Paradigm	**Nature**	**Form**	**Period**
First	Experimental science	Empiricism; describing natural phenomena	Pre-Renaissance
Second	Theoretical science	Modelling and generalization	Pre-Computers
Third	Computational science	Simulation of complex phenomena	Pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Present

Data-driven exploratory science, however, obviously requires a huge amount of expertise especially if confronted with huge data. This is why the third and often under-estimated form of logical reasoning, abduction, is becoming increasingly prominent in the age of big data as this form of reasoning is particularly consistent with up-to-date standards in big data-driven research (Kelling et al., 2012) as well as the corresponding ambition of theory development and refinement (Wright, 2017). In management research, for example, the Academy of Management has recently launched an entire journal dedicated to exploratory, abductive reasoning, Academy of Management Discoveries, and the link between big data and the reignited interest in abduction has been explicitly and prominently mentioned as a justification of the need for this journal on several occasions, including editorials of the editors-in-chief such as in Bamberger (2018).

As is well known, Charles Sanders Peirce (1940, p. 151) proposed the following as the standard form of abductive reasoning: "The surprising fact, C, is observed; but if A were true, C would be a matter of course, [and hence], there is reason to suspect that A is true".

Also referred to as the art of educated or "intelligent guessing" (Popper, 1959), abduction is therefore the answer to the question of how researchers can react if confronted with phenomena and patterns as necessarily as hugely surprising as the ones that emerge from the observation of computer-/AI-generated huge data sets. In fact, abductive reasoning is necessarily also involved even before these surprising huge data-driven phenomena and patterns can at all be observed to emerge. This is true because all algorithmic or artificial form of intelligence that screens big datasets is based on an inevitably comparably smaller set of presumptions as to what phenomena or patterns, in short, what aspects of the huge dataset might be particularly informative or relevant to the research context.

Big data research is, therefore, yet another prime example of the observer effect, according to which any observation of a phenomenon necessarily shapes and changes this phenomenon. In the case of big data, this issue refers not only to the ultimately selective sampling of the raw data (as even the hugest dataset can only display those aspects of [social] life that can be datafied), but also to the design of the computer-/AI-based data analysis tools used for making sense out of what otherwise would remain a huge datafied white noise. In this context, it is critical to stress the difference between data and information, as data only turns into information if observed by an observer for whom the data constitutes "a difference that makes a difference" (Bateson, 1979). Any big data set in general and any large set of data lakes in particular is therefore as close as can be to an unmarked space (Spencer Brown, 1979), in the context of which it is the operations of the observer, i.e., the drawing of distinctions by that observer, that make the world in which they exist.

If really "complete" or at least considerably comprehensive, one and the same big data lake can be turned into a source of information for research interests as different as algorithmic finance and social credit systems or broadband scans of the past and data-driven multi-scenario visions of the future.

It is therefore, indeed, the choice of the distinctions that researchers draw to create the tools to make sense out of the unmarked space of big data white noise that makes the world in which we shall live in the imminent or even present big data age.

In this context, abduction actually emerges as the noblest form of inference insofar as it places high value on the role of the researchers' self-observation and thus on self-referential research methodologies that help to manage the third order risk (Godet, 1986) of giving right answers to wrong questions, a risk that is particularly prevalent and critical in foresight and futures studies where knowledge of future trends is regularly based on strong assumptions on the dependence on

and breaks in the paths the past has taken. In this sense, the marriage of big data research and abduction has the greatest potential whenever we understand that the primary purpose of big data-driven research is to inform rather than immediately confirm or challenge existing theories (see Luciano et al., 2018, p. 614).

The existence of big data research methods, therefore, facilitates and requires a complementary smartening of methodologies and theories, e.g., in the form of theories that act as a methodology, which may be created if we first follow for example Norbert Elias (1978) and challenge the categorical separation of theory and method, and then build theories that indicate how their observations come about and can be replicated. One example of (at least a prototype of) such as reflective, self-implicative theory clearly is social systems theory by Niklas Luhmann (1995; 2012; 2013).

Consequently, one major outcome of this initial literature review report is that the digital transformation of virtually everything along with the corresponding emergence of huge data sets or lakes calls for a profound digital transformation of the way of how present and future generations of researchers define and do theory. This digitalization or computerization of theory could then not stop by acts of mere copying and pasting of analog bodies of knowledge and research questions into digital contexts. Rather, theorists would have to scrutinize the ways of how analog contents and questions may actually be translated into digital ones. This exercise, for its part, would require the development of non-partisan theory machines that actually can perform these translations without dis-/favoring certain theory traditions or paradigms.

While BDA is poised to provide a welter of data-driven insights, it still needs to be contextualized with other elements in the subject or system under study i.e. existing theory, documents, small data studies, historical records, trends etc. to make sense of the patterns evident (Crampton et al., 2012).

To summarize our report we can present current understanding about challenges and opportunities of Internet of Things, Big Data and data mining.

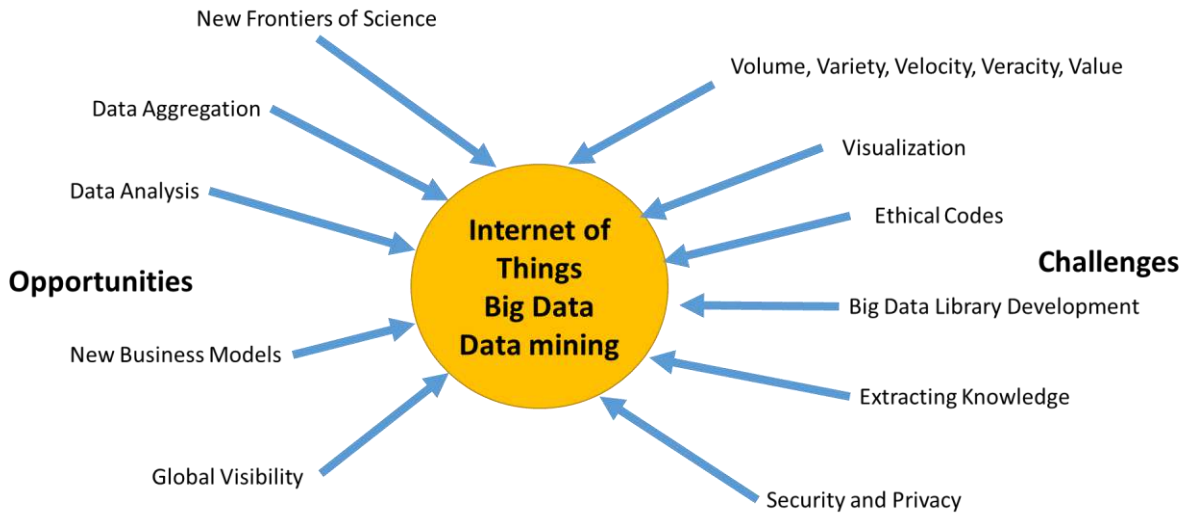

Figure 23. Challenges and opportunities (extended modification of Shadroo and Rahmani, 2018, p. 44).

Professional and specialists of Big Data are now facing both challenges and opportunities. The key task is to find balance between challenges and opportunities.

References

Abolfazli, S., Sanaei, Z., Gani, A., Xia, F., and Yang, L. T. (2014). Rich mobile appliciations: Genesis, taxonomy, and open issues. *Journal of Network and Computer Applications*. 40, 345-362.

Ahmed, E., Yaqoob I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., and Vasilakos, A. V. (2017). The role of Big Data Analytics in Internet of Things. *Computer Networks.* 129 (2), 459-471.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7), 16-17.

Baesens, B. (2014). *Analytcis in a Big Data World. The Essential Guide to Data Science and its Applications.* Hoboken, New Jerset: Wiley.

Baeth, M. J., & Aktas, M. S. (2019). Detecting misinformation in social networks using provenance data. *Concurrency and Computation: Practice and Experience*, *31*(3), e4793.

Bamberger, P. A. (2018). AMD—Clarifying what we are about and where we are going. *Academy of Management Discoveries, 4*(1).

Barnes, R. (2018). Weapons of mass distraction. *Northern Ireland Legal Quarterly*, *69*(4), 475-512.

Bateson, G. (1979). *Mind and Nature: A Necessary Unity*. New York: Dutton.

Bautista Villalpando, L.E., April, A., & Abran, A. (2014). DIPAR: A framework for implementing big data science in organizations. In: Mahmood,Z.(Ed.) *Continued Rise of the Cloud*. Springer, London, 177–194.

Béranger, J. (2016). *Big Data and Ethics: The Medical Datasphere*. London: Elsevier.

Bergé, J. S., & Grumbach, S. (2017). The Datasphere and the Law: New Space, New Territories. *Revista Brasileira de Politicas Pública, 7*(3), 1-19.

Brabham, D. C. (2010). Moving the Crowd at Threadless. *Information, Communication & Society*. 13 (8), 1122-1145.

Bzhalava, L., Hassan, S.S., Olsson, B.K. and Kaivo-oja, J. (2019). Text mining as a useful tool to detect trends in startup entrepreneurial activities. *Accepted paper* to the *39th Strategic Management Conference "Out of the Spotlight" Strategies*, in Minneapolis, US 19-22 Oct 2019.

Bzhalava, L. and Cantner, U. (2018). The journey towards open innovation: why do firms choose different routes? Eurasian Business Review, Vol. 8, Issue 3, 245–265. https://doi.org/10.1007/s40821-017-0101-9

Bzhalava, L., Kaivo-oja, J. and Hassan, S. (2018). Data-based startup profile analysis in the European smart specialization strategy: A text mining approach. *European Integration Studies*, 12, 118-128.

Bzhalava, L., Hassan, S.S., Olsson, B.K. & Kaivo-oja, J. (2019). Detecting trends in startup entrepreneurial activities using text mining. *Journal of Business Venturing*. Submitted. In review process.

Bzhalava, L., Hassan, S.S., Kaivo-oja, J., & Olsson, B.K. (2019). Mapping emerging industries by co-word and social network analysis. Full paper submitted to review process in March-April 2019.

Blazquez, D., & Domenech, J., (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, Vol. 130 (May 2018), 99-113.

Chen, C. F., Qian, O., & Dai, Y. Z. (2014). Study on the construction of digital library in the age of big data. *Library and Information Service*, 58(7), 40–45.

Chesbrough, H. (2017). The future of open innovation. *Research Technology Management*, 60(1), 35-38.

Choi, S., Park, H., Kang, D., et al. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications. An International Journal*, 39 (13), 11443–11445.

Choudary, P. (2015). *Platform Scale: How an Emerging Business Model Helps Startups Build Large Empires with Minimum Investment*. Platform Thinking Laps Pte.

Chun, W. H. K. (2016). *Updating to Remain the Same: Habitual New Media*. MIT Press. USA.

Crampton, J., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2012). Beyond the Geotag? Deconstructing ''Big Data'' and leveraging the Potential of the Geoweb. Available at: http://www.uky.edu/tmute2/geography_methods/readingPDFs/2012-Beyond-the-Geotag-2012.10.01.pdf

Day, G.S. and Schoemaker, P.J.H. (2004). Driving through the fog: Managing at the edge. *Long Range Planning,* 37(2), 127-142.

Dawson, R. (2010). Six tools to kickstart your crowdsourcing strategy. *MYCustomer* 1 July, 2010. Web: http://www.mycustomer.com/topic/customer-intelligence/ross-dawson-six-tools-start-your-crowdsourcing-strategy/109914

De Dreu, C. K. W., Nijstad, B. A. and van Knippenberg, D. (2008). Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, 12, 22–49.

De Mauro, A., Greco, M. and Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *AIP conference proceedings*, Vol. 1644, No. 1, 97-104.

De Mauro, A., Greco, M. and Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65(3), 122–135.

Eckroth, J. (2018). A course on big data analytics. *Journal of Parallel and Distributed Computing*, 118, 166-176.

Flostrand, A. (2017). Finding the future: Crowdsourcing versus the Delphi technique. *Business Horizons*, 60(2), 229-236.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology*, *25*(4), 435-437.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

Foreman, J.W. (2014). *Data Smart. Using Data Science to Transform Information to Insight*. Indianapolis: Wiley.

FOR LEARN (2019). Support to mutual learning between Foresight managers, practitioners, users and stakeholders of policy-making organisations in Europe. Institute for Prospective Technological Studies. Joint Research Centre. Web: http://forlearn.jrc.ec.europa.eu/index.htm

Fourkind (2019). Platform Economy Handbook. Web: https://fourkind.com/v2/wp-content/uploads/2018/08/Platform-Economy-Handbook.pdf

Frankish, K. and Ramsey, W.M. (2014). *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press.

Ge, M., Bangui, H. and Buhnova, B. (2018). Big Data for Internet of Things: A survey. *Future Generation Computer Systems*, 87, 601-614.

Giglietto, F., and Mesjasz, C. (2019). 'Fake news' is the invention of a liar: How false information circulates within the hybrid news system. *Current Sociology*, forthcoming.

Godet, M., (1986). Introduction to 'la prospective': Seven key ideas and one scenario method. *Futures*, 18, 134–157.

Google Trends (2019). Monthly Global Data from Google Trends. *Google Trends* 24.3.2019. Web: https://trends.google.fi/trends/?geo=FI

Goodwin, P. and Wright, G. (2004). *Decision Analysis for Management Judgement.* Third Edition. Chichester: John Wiley & Sons.

Graf, H.G. (2002). *Economic Forecasting for Management. Possibilities and Limitations*. Westport, Connecticut: Quorum Books.

Grus, J. (2015). *Data Science from Scratch. First Principles with Python*. USA: O´Reilly.

Guo, J., Wang, X., Li, Q. and Zhu, D. (2016). Subject–action–object-based morphology analysis for determining the direction of technological change. *Technological Forecasting and Social Change*, 105 (4), 27–40.

Hajirahimova, M. S. and Aliyeva, A. S. (2017). About Big Data measurement methodologies and indicators. *International Journal of Modern Education and Computer Science*, 9 (10), 1–9. Web: http://www.mecs-press.org/ijmecs/ijmecs-v9-n10/IJMECS-V9-N10-1.pdf

Heuschneider, S and Hersta, C. (2016). *External Search for Exploration of Future Discontinuities and Trends: Implications from the Literature Using Co-Citation and Content Analysis*. Hamburg University of Technology Working Paper No. 92. Hamburg, Germany.

Helbing, D. (2019). Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards Digital Enlightenment* , Springer, Cham, pp. 47-72.

Hellerstein, J. (2008). Parallel Programming in the Age of Big Data. Gigaom Blog. 9 November 2008). Web: https://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/

Herschel, R. & Miori, V. M. (2017). Ethics & Big Data. *Technology in Society*. 49: 31-36.

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Grey on eScience: A transformed scientific method. In: Id. (eds) The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft Research, pp. xvii–xxxi.

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 332 (6025), 60–65.

Howe, J. (2006). The rise of crowdsourcing, *Wired* 14 (6) (2006) 176–183.

Howe, J. (2008). *Crowdsourcing.* Crown Publishing Group, New York, 2008.

Howe, J. (2009). *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. Crown Business, New York.

IBM (2013). What is big data? – Bringing big data to the enterprise. *IBM*. Web: www.ibm.com. Retrieved 18 March 2019. IBM.

Jia, S., Lansdall-Welfare, T., and Cristianini, N. (2016). Gender classification by deep learning on millions of weakly labelled images. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 462-467.

Jungleworks (2019). How the Matching Algorithm works in the On-Demand Economy? Part Three of the *User Journey Series*. Web: https://jungleworks.com/matching-algorithm-works-demand-economy-part-three-user-journey-series/

Kaivo-oja, J., (2012). Weak signals analysis, knowledge management theory and systemic socio-cultural transitions. *Futures*, 44, 206–217.

Kaivo-oja, J. (2019). Introduction: The Challenges of Big Data Foresight. Lecture in Turku Science Park. *Dos and Don'ts of Big Data for Foresight*, Turku Science Park, Turku, Thursday 28.2.2019.

Kaivo-oja, J., Virtanen, P., Jalonen, H and Stenvall, J. (2015). The effects of the Internet of Things and Big Data to organizations and their knowledge management practices. In *Knowledge Management in Organizations, Lecture Notes in Business Information Processing*, Volume 224, Springer International Publishing, 495-513.

Kaivo-oja, J. and Stenvall, J. (2013). Foresight, governance and complexity of systems: on the way towards pragmatic governance paradigm. *European Integration Studies*, No 7, 28-34.

Kaivo-oja, J. and Roth, S. (2019). Strategic foresight for competitive advantage: A future-oriented business and competitive analysis techniques selection model. In review process after first round review reports. Pre-conditional acceptance. Second round corrections delivered in March 2019.

Kaivo-oja, J., Kuusi, O., Knudsen, M.S. and Lauraeus, T. (2019). Digital Twins approach and future knowledge management challenges: Where we shall need system integration, synergy analyses and synergy measurements? KMO2019 Conference. July 15-18, 2019 KMO 2019, *14th International Conference on Knowledge Management in Organisations. Theme: The synergistic role of knowledge management in organisations*, University of Salamanca, Zamora, Spain. *Accepted, forthcoming ©Springer Science*. Web: https://www.kmo2019.com/

Kamppinen, M. (1998). Evolution and culture: the Darwinian view on infosphere. *Futures*, *30*(5), 481-484.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies, *BioScience*, *59*(7), 613-620.

Kim, J., Park, Y., and Lee, Y. (2016). A visual scanning of potential disruptive signals for technology roadmapping: investigating keyword cluster, intensity, and relationship in futuristic data. *Technology Analysis and Strategic Management*, 28(10), 1–22.

Knudsen, M.S., Kaivo-oja, J. and Lauraeus, T. (2019). Enabling technologies of Industry 4.0 and their global forerunners: An empirical study of the Web of Science database. KMO2019 Conference. July 15-18, 2019 KMO 2019, *14th International Conference on Knowledge Management in Organisations. Theme: The synergistic role of knowledge management in organisations*, University of Salamanca, Zamora, Spain. *Accepted, forthcoming ©Springer Science*. Web: https://www.kmo2019.com/

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1-12.

Kumar, V. and Chadha, A. (2012). Mining association rules in student's assessment data. *International Journal of Computer Science Issues*, 9(5), 211–216.

Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., Team, F. N., & Cristianini, N. (2017). Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, *114*(4), E457-E465.

Laursen, K. and Salter, A. (2006). Open for innovation: the role of openness in explaining innovation performance among U.K. manufacturing firms. *Strategic Management Journal*, 27(2), 131–150.

LeHong, H. and Laney, D. (2013). *Toolkit: Board-Ready Slides on Big Data Trends and Opportunities*. Gartner, 1 March 2013, G00238695.

Li, S., Jiao, F., Zhang, Y., and Xu, X. (2019). Problems and Changes in Digital Libraries in the Age of Big Data from the Perspective of User Services. *The Journal of Academic Librarianship*. 45(1), 22-30.

Liang, T-P. and Liu, Y-H. (2018). Research landscape of business intelligence and Big Data analytics: A bibiometrics study. *Expert Systems with Applications*, 111, 2-10.

Luciano, M. M., Mathieu, J. E., Park, S. and Tannenbaum, S. I. (2018). A fitting approach to construct and measurement alignment: The role of big data in advancing dynamic theories. *Organizational Research Methods*, *21*(3), 592-632.

Luhmann, N. (1995). *Social Systems*. Palo Alto: Stanford University Press.

Luhmann, N. (2012). *Theory of Society, Vol. 1.* Palo Alto: Stanford University Press.

Luhmann, N. (2013). *Theory of Society, Vol. 2.* Palo Alto: Stanford University Press.

Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., and Sheth, A.P. (2018). Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, 4, 161–175.

McKinsey Global Institute (2018). *Smart Cities: Digital Solutions for a More Livable Future*. Executive Summary. McKinsey Global Institute.

Moe, S. and Kaivo-oja, J. (2018). Model theory and observing systems. Notes on the use of models in systems research, *Kybernetes,* 47(9), 1690-1703.

Monteiro, F., Mol, M. and Birkinshaw, J. (2017). Ready to be open? Explaining the firm level barriers to benefiting from openness to external knowledge. *Long Range Planning*, 50(2), 282-295.

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K. and Ghosh, P. (2015). *Big Data: Prospects and Challenges*. Vikalpa.

Ocasio, W. (2011). Attention to attention. *Organization Science*, 22, 1286–1296.

Osman, A.M.S. (2019). A novel big data analytics framework for smart cities. *Future Generation Computer Systems,* 91, 620-633.

Oussos, A., Benjelloun, F-Z., Laheen, A.A. and Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University – Computer and Information Services*. 30, 431-448.

Oxford Dictionaries (2019). *Big Data.* Web: https://en.oxforddictionaries.com/definition/big_data

Peirce, C. S. (1940). *Collected Papers Volume 1-6*. Cambridge, MA: Harvard University Press.

Piippo, T. (2019). Using world-class data science resources to create a smart and wise Turku. *Dos and Don'ts of Big Data for Foresight*, Turku Science Park, Turku, Thursday 28.2.2019.

Piezunka, H., and Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizaons' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, 58 (3), 856–880.

Popper, K. (1959). *The Logic of Scientific Discovery.* New York: Basic Books.

Prpić, J., Prashant P., Shukla, P.P., Kietzmann, J.H., & McCarthy, I.P. (2015). How to work a crowd: Developing crowd capital through crowdsourcing. *Business Horizons*, Vol. 58, Issue 1, January–February 2015, 77-85.

Raleigh, N. B. (2019). Current project insights: Potentials of big data for integrated territorial policy development in the European growth corridors. *Dos and Don'ts of Big Data for Foresight*, Turku Science Park, Turku, Thursday 28.2.2019.

Ralston, B. and Wilson, I. (2006). *The Scenario Planning Handbook*. Crawforsville, Indiana: Thomson.

Reinsel, D.; Gantz, J., & Rydning, J. (2017). Data Age 2025: The Evolution of Data to Life-Critical (PDF). Framingham, MA, US: International Data Corporation. Web: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Roth, S., Valentinov, V., Schwede, P., Pérez Valls, M., & Kaivo-oja, J. (2019). Harnessing big data for a multifunctional theory of the firm. Submitted. An article in review process.

Santonen, T. and Kaivo-oja, J. (2019). Crowdsourcing Delphi – Defining a novel methodological solution for foresighting. Submitted article to review process. An article in review process. 17 pages.

Saxena, D., and Lamest, M. (2018). Information overload and coping strategies in the big data context: Evidence from the hospitality sector. *Journal of Information Science*, *44*(3), 287-297.

Segaran, T. and Hammerbacher, J. (2009). *Beautiful Data: The Stories behind Elegant Data Solutions*. O'Reilly Media.

Shadroo, S. and Rahmani, A.M. (2018). Systematic survey of big data and data mining in internet of things. *Computer Networks*, 139, 19-47.

Shi, Z.M., Lee, G. and Whinston, A.B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, 40 (4), 1035-1056.

Shin, D. H., and Choi, M. J. (2015). Ecological views of big data: Perspectives and issues. *Telematics and Informatics*, *32*(2), 311-320.

Snowden, D.J. (2002). Complex acts of knowing: paradox and descriptive self-awareness. *Journal of Knowledge Management,* 6 (2), 100–111.

Snowden, D. J. and Kurtz, C. F. (2003). The new dynamics of strategy: Sense-making in a complex and complicated world. *IBM Systems Journal*, 42 (3), 462–483.

Spaniol, M.J. and Rowland, N.J. (2019). Defining Scenario. *Futures & Foresight Science*. 1(1), 1-13.

Stapenhurst, T. (2009). *The Benchmarking Book. A How-to-guide to Best Practice for Managers and Pactioners*. Oxford: Elsevier.

Surbakti, F.P.S., Wang, W., Indulska, M. and Sadig, S. (2019). Factors influencing effective use of big data: A research framework. *Information & Management*, forthcoming.

Tansley, S., and Tolle, K. (2009, Eds). *The Fourth Paradigm: Data-intensive Scientific Discovery*. Redmond: Microsoft Research, USA, pp. xvii–xxxi.

Teets, M., and Goldner, M. (2013). Libraries' role in curating and exposing Big Data. *Future Internet*. 5, 429-438.

Tiwana, A. (2014). *Platform Ecosystems. Aligning Architecture, Governance, and Strategy*. 1st Edition. Morgan Kaugman. Waltham, MA.

Thorleuchter, D. and Van den Poel, D. (2013). Weak signal identification with semantic web mining. *Expert Systems with Applications*, 40(12), 4978-4985.

Thorleuchter, D., Scheja, T. and Van den Poel, D. (2014). Semantic weak signal tracing. *Expert Systems with Applications*, 41(11), 5009-5016.

Tseng, Y., Lin, C. and Lin, Y. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, 43, 1216–1247.

van der Duin, P. (2016). *Foresight in Organizations. Methods and Tools*. New York: Routledge.

van Knippenberg, D. Dahlander, L., Haas, M. R. and George, G. (2015). Information, attention, and decision making. *Academy of Management Journal*, 58(3), 649–657.

Vargo, C. J., Guo, L. and Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028-2049.

Wikipedia (2019). *Big Data Ethics*. Located 31.3.2019 at: https://en.wikipedia.org/wiki/Big_data_ethics.

Wiren, M. (2019). Strategic Positioning in Big Data Utilization. *Dos and Don'ts of Big Data for Foresight*, Turku Science Park, Turku, Thursday 28.2.2019.

Wright, P. M. (2017). Making great theories. *Journal of Management Studies*, *54*(3), 384-390.

Xu, S., Du, W., Wang, C., and Liu, D. (2017). The library big data research: Status and directions. *The Journal of Academic Librarianship*, 5(3), 77–88.

Yoon, J. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. *Expert Systems with Applications*, 39(16), 12543-12550.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*. 1-6.

Annex 1: „Platforms for Big Data Foresight“ Project Activities (Nov. 2018 – Mar. 2019)

- *December 4th 2018*: Kickoff Meeting at Kazimieras Simonavicius University. More information at http://www.ksu.lt/en/news/kickoff-meeting-platforms-of-big-data-foresight/

- *February 28th 2019*: Open inception workshop “Dos and Don’ts of Big Data for Foresight”, Turku, Finland. More information at http://www.ksu.lt/en/news/open-inception-workshop-dos-and-donts-of-big-data-for-foresight/

- *March 2019*: Publication of WP2 Summary report: Integrated Knowledge Management in the Field of Big Data Foresight and Associated Digital Platforms

Annex 2: „Platforms for Big Data Foresight“ Project Publications (2018-)

Bzhalava, L., Text mining as a useful tool to detect trends in startup entrepreneurial activities. *Accepted paper* to the 39th Strategic Management Conference "Out of the Spotlight" Strategies, in Minneapolis, US 19-22 Oct 2019.

Bzhalava, L., Hassan, S.S., Olsson, B.K. & Kaivo-oja, J. (2019) Detecting trends in startup entrepreneurial activities using text mining. *Journal of Business Venturing*. Submitted. In review process.

Bzhalava, L., Hassan, S.S., Kaivo-oja, J., & Olsson, B.K. (2019) Mapping emerging industries by co-word and social network analysis. Full paper submitted to review process in March-April 2019.

Kaivo-oja, J. and Roth, S. (2019) Strategic foresight for competitive advantage: A future-oriented business and competitive analysis techniques selection model. *International Journal of Technology Management.* In review process after first round review reports. Pre-conditional acceptance. Second round corrections delivered in March 2019.

Knudsen, M.S., Kaivo-oja, J. & Lauraeus, T. (2019) Enabling technologies of Industry 4.0 and their global forerunners: An empirical study of the Web of Science database. KMO2019 Conference. July 15-18, 2019 KMO 2019, 14th International Conference on Knowledge Management in Organisations. Theme: The synergistic role of knowledge management in organisations, University of Salamanca, Zamora, Spain. *Accepted, forthcoming ©Springer Science*. Web: https://www.kmo2019.com/

Kaivo-oja, J., Kuusi, O., Knudsen, M.S. & Lauraeus, T. (2019) Digital Twins approach and future knowledge management challenges: Where we shall need system integration, synergy analyses and synergy measurements? KMO2019 Conference. July 15-18, 2019 KMO 2019, 14th International Conference on Knowledge Management in Organisations. Theme: The synergistic role of knowledge management in organisations, University of Salamanca, Zamora, Spain. *Accepted, forthcoming ©Springer Science*. Web: https://www.kmo2019.com/

Roth, S. (2019a) Digital transformation of social theory. A research update. *Technological Forecasting and Social Change*. Under review.

Roth S. (2019b) The open theory and its enemy. Implicit moralisation as epistemological obstacle for general systems theory, *Systems Research and Behavioral Science*, In press. DOI: 10.1002/sres.2590.

Roth S., Valentinov V., Kaivo-oja J., and Dana L.-P. (2018) Multifunctional organisation models. A systems-theoretical framework for new venture discovery and creation, *Journal of Organizational Change Management*, Vol. 31 No. 7, 1383-1400. (production started in during planning process of the project)

Roth, S., Leydesdorff, L., Kaivo-oja, J. and Sales, A. (2019) Open coopetition: When multiple players and rivals team up. *Journal of Business Strategy. Forthcoming*. ©Emerald. Web: https://www.emeraldinsight.com/journal/jbs

Santonen, T. and Kaivo-oja, J. (2019). Crowdsourcing Delphi – Defining a novel methodological solution for foresighting. Submitted article to review process. An article in review process. 17 pages.