FutureTDM

Explore . Analyse . Improve

REDUCING BARRIERS AND INCREASING UPTAKE OF TEXT AND DATA MINING FOR RESEARCH ENVIRONMENTS USING A COLLABORATIVE KNOWLEDGE AND OPEN INFORMATION APPROACH

Deliverable D3.1

Research Report on TDM Landscape in Europe

Project

Acronym: **FutureTDM**

Title: Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach

Coordinator: SYNYO GmbH

Reference: 665940

Type: Collaborative project

Programme: HORIZON 2020

Theme: GARRI-3-2014 - Scientific Information in the Digital Age: Text and Data Mining (TDM)

Start: 01 September, 2015

Duration: 24 months

Website: http://www.futuretdm.eu

E-Mail: office@futuretdm.eu

Consortium: **SYNYO GmbH**, Research & Development Department, Austria, (SYNYO)

Stichting LIBER, The Netherlands, (LIBER)

Open Knowledge, UK, (OK/CM)

Radboud University, Centre for Language Studies, The Netherlands, (RU)

The British Library Board, UK, (BL)

Universiteit van Amsterdam, Inst. for Information Law, The Netherlands, (UVA)

Athena Research and Innovation Centre in Information, Communication and Knowledge Technologies, Inst. for Language and Speech Processing**,** Greece**,** (ARC)

Ubiquity Press Limited, UK, (UP)

Fundacja Projekt: Polska, Poland, (FPP)

Deliverable

Number:	**D3.1**
Title:	**Research report on TDM Landscape in Europe**
Lead beneficiary:	**RU**
Work package:	WP3: ASSESS: Studies, Publications, Legal Regulations, Policies and Barriers
Dissemination level:	Public (PU)
Nature:	Report (RE)
Due date:	31.01.2016
Submission date:	29.04.2016
Authors:	**Maria Eskevich**, RU **Antal van den Bosch**, RU
Contributors:	**Marco Caspers,** UVA **Lucie Guibault,** UVA **Alessio Bertone**, SYNYO **Susan Reilly,** LIBER **Carmen Munteanu**, SYNYO **Peter Leitner**, SYNYO **Stelios Piperidis,** ARC
Review:	**Marco Caspers,** UVA **Lucie Guibault,** UVA **Alessio Bertone**, SYNYO

Acknowledgement: This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 665940.

Table of Contents

1 Introduction 6

1.1. What is text and data mining? 6

1.2. The growing importance of TDM 7

1.3. Goal and structure of the report 9

2 TDM framework 11

2.1. Data to be mined 11

2.2. Technology and R&D 16

3 TDM across economic sectors 21

3.1. Primary sector 22

3.2. Secondary sector 22

3.3. Tertiary sector 23

3.4. Quaternary sector 25

3.5. Quinary sector 26

4 Summary of the literature 27

5 Conclusion 40

References 41

LIST OF FIGURES

Figure 1. Number of different types of publications in the DBLP Computer Science bibliography. 8

Figure 2. Number of different types of publications on the European Life Sciences PMC website. 9

Figure 3. Macro view of TDM framework. 11

Figure 4. Micro view on TDM environment. 17

Figure 5. General economic structure and connection between TDM and all economic sectors. 22

Figure 6. Timeline of reports. 29

1. INTRODUCTION

Recent years have seen an exponential growth in the volume of digital data available to all levels of society. A further proliferation of digital or digitised content is expected in this, the era of Big Data. A number of interrelated factors have boosted this data *thrust* and the accompanying development of technologies to exploit this new type of information-rich ground material with text and data mining (TDM). First, the costs for data storage and cloud computing facilities continue to decrease, while computing power increases. Second, digitisation has moved from specific processes, such as archiving and optimizing industrial workflows, to nearly all daily aspects of societal activities. This process paves the way for *datafication* of this information, i.e. the creation of new information and knowledge, potentially with new value, on the basis of machine-readable content. TDM is the umbrella term for technologies that enable this goal.

The continuous growth of data[1] increasingly emphasises the question of how to best make use of it. Society expects, with good reason, that the implementation of intelligent methods to deal with BIg Data will lead to improved information access, faster knowledge discovery, higher productivity, and competitiveness. These expectations have inspired TDM researchers and companies across the world to explore novel algorithms for extracting information, discovering knowledge, and improving the ability to predict from data on a large scale. As data is a manifold concept, so is TDM. This report aims to chart and structure the TDM landscape.

1.1. What is Text and Data Mining?

Text and Data Mining was initially defined as "the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different (…) resources, to reveal otherwise hidden meanings" (Hearst, 1999), in other words, "an exploratory data analysis that leads to the discovery of heretofore unknown information, or to answers for questions for which the answer is not currently known" (Hearst, 1999). Nowadays, this umbrella term encompasses diverse techniques that allow interpretation of content of any type ranging from raw data, e.g. sensor data, text, images and multimedia, to processed content, e.g. diagrams, charts, tables, references, maps, formulas, chemical structures, and metadata from semi-structured sources, on a large scale through the identification of patterns.

TDM algorithms harbour vast potential for nearly all scientific fields and for a wide diversity of practical industrial and societal applications. However, this broad potential and variety of TDM applications complicates the landscape view, as even the technology itself is referred to using different terms within different fields. Within the business and managerial context, TDM can be referred to as *Business Intelligence Solutions* or *Qualitative Data Analysis*. To highlight the central

[1] It is expected that there will be 16 trillion gigabytes of data by 2020, which corresponds to an annual growth rate of 236 % in data generation (*Motion for a resolution further to Question for Oral Answer B8-0116/2016 pursuant to Rule 128(5) of the Rules of Procedure on "Towards a thriving data-driven economy" (2015/2612(RSP)*, 2016):
http://www.europarl.europa.eu/sides/getDoc.do?type=MOTION&reference=B8-2016-0308&language=EN

position of vast quantities of data, the term *Big Data (Processing)* is used. When the focus is on the discovery of hitherto undiscovered information, *Content Mining* appears to be a term of choice, while *Data Analytics* and *Text Analytics* emphasise the data-driven aspect of performing analyses and a focus on seeking analytical solutions to challenges. In the academic world, TDM is often considered to be composed of *Machine Learning* methods coupled with *Exploratory Data Analysis* methods such as visualisation. Machine Learning, in turn, arose from the post-war fields of *Artificial Intelligence*, *Information Theory*, and *Pattern Recognition*.

1.2. The Growing Importance of TDM

With the increase of computational power, the rise of Big Data, and the digitisation of vast amounts of data, the use of TDM promises to yield valuable societal and economic benefits in terms of new insights and cost savings that would otherwise not be possible: more scientific breakthroughs, a greater understanding of society, and countless business opportunities. The potential of TDM as a research technique and as an economic opportunity has been recognised at a political level in the EU (Hargreaves I. et al., 2014).[2] The benefits of TDM have also been noted by the research community and societal stakeholders e.g. via the Hague Declaration[3].

Originally rooted in academic research, TDM has left its academic childhood years and is now a global information technology (IT) industry. As an industry, TDM is fuelled by direct access to large amounts of data. Companies can either develop and run their own TDM software, or outsource this work to external experts who can customise TDM solutions. Industry is now rapidly rolling out the implementation of TDM for Big Data, as computing power grows and businesses find new ways to gather and access data. According to the International Data Corporation's Worldwide Big Data Technology and Services Forecast for 2013-2017, the growth rate of the Big Data market until 2017 will be six times faster than that of the overall information and communication (ICT) market. The Big Data Value Public-Private Partnership[2] predicts that TDM will reach a total of EUR 50 billion by 2017, and may result in 3.75 million new jobs. While many companies see the potential of TDM for their business development, only 1.7 % of companies are currently making full use of advanced digital technologies like TDM, despite the benefits that digital tools can bring.[2]

The application of TDM to science holds the promise of accelerating scientific discoveries. Mining the content present in research publication databases and the associated research data sets, as far as they are available for harvesting, will boost scientific research by increasing the productivity of researchers and leaving part of the scientific discovery process to computers, a trend called *Data Science*. The size of scientific publication databases varies across domains, but in each case, the number is beyond a value that would be feasible for a human to read and assess without digital support. Figures 1-2 give examples of the volume of content that has to be dealt with by the

[2] (*Motion for a resolution further to Question for Oral Answer B8-0116/2016 pursuant to Rule 128(5) of the Rules of Procedure on "Towards a thriving data-driven economy" (2015/2612(RSP)*, 2016): http://www.europarl.europa.eu/sides/getDoc.do?type=MOTION&reference=B8-2016-0308&language=EN; http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

[3] The Hague Declaration on Knowledge Discovery: http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/

academia and industry researchers in the Computer Science[4] and Life Sciences domains[5], respectively. TDM may help researchers in all scientific fields to regain control over the academic process, to aggregate information and to follow trends across fields beyond their own. Estimations of added economic value in this case, based on the time saving benefit of TDM alone, are an increase of 2 per cent, or EUR 5.3 billion, with long-term impact of overall gain in the range of at least EUR 32.5 billion (Filippov, 2014).

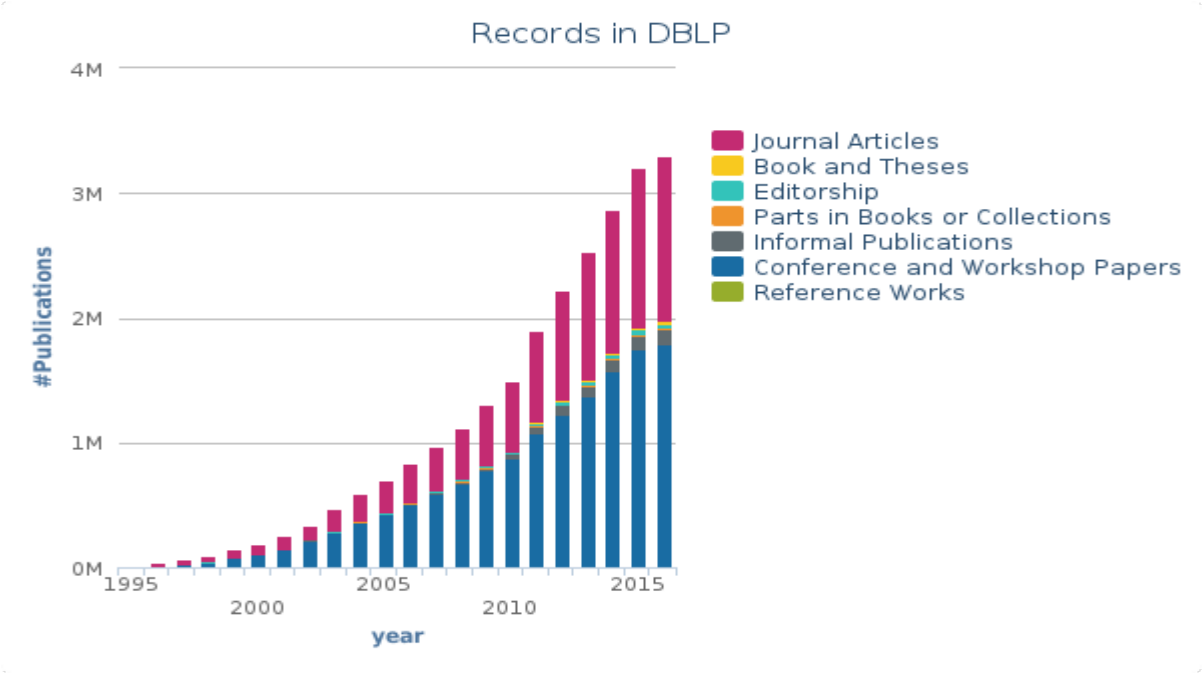

Figure 1. Number of different types of publications in the DBLP Computer Science bibliography.[6]

[4] http://dblp2.uni-trier.de
[5] https://europepmc.org
[6] http://dblp2.uni-trier.de/statistics/publicationsperyear.html [accessed on March 2016]

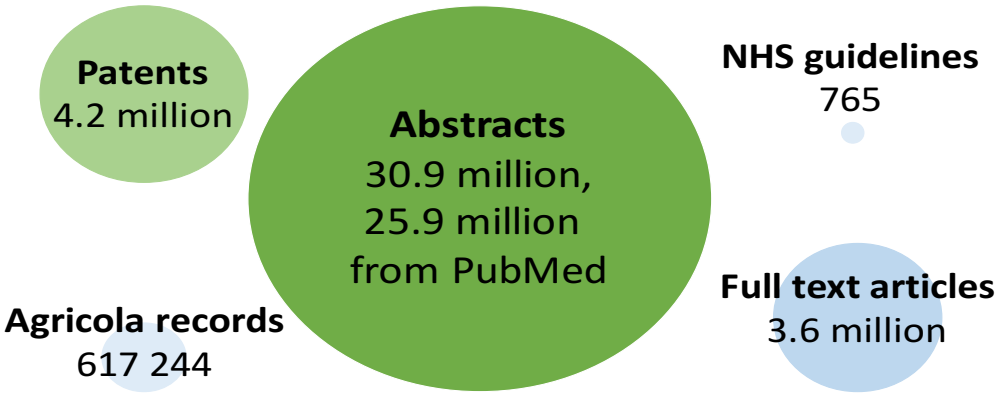

Figure 2. Number of different types of publications on the European Life Sciences PMC website.[7]

1.3. Goal and structure of the report

In this report, we outline the text and data mining landscape. We start by providing an overview of types of data and technologies that the TDM experts are working with. We look at the global economic structure of the knowledge-based society and highlight use cases for TDM uptake across all economic sectors. Following recent discussions on TDM perspectives within the scientific, political, and economic context, we highlight the importance of the technology and its potential.

TDM algorithms are developed globally. It is therefore our task to list and structure the important technology definitions, data types, underlying algorithms and leading research communities on a worldwide scale. However, the actual uptake of TDM and ease of practical implementation vary considerably depending on the country and the economic sector in question. In this report, we give a broad perspective on the general European Union (EU) landscape, and quote exceptional country-level directives only when these countries have legal regulations specifically addressing TDM, e.g. the case of statutory exception on copyright in the UK[8].

TDM is becoming increasingly ubiquitous on the global market, as each company that tracks its activities, products, and communications in some digital form has the potential to reuse this data for further analysis and improvement. For some sectors, countries and applications, it is already possible to estimate the potential profit or at least the size of the market. Other fields such as journalism,

[7] https://europepmc.org/About [accessed on March 2016]

[8] "Section 29A and Schedule 2(2)1D of the Copyright, Designs and Patents Act 1988": http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made

which is in the process of embracing the concept of *Data Journalism*, have to first agree on the conceptual changes required to their current work behaviour in order to merge the traditional work pattern with TDM (Lewis S.C. and Westlund O., 2014).

This report is structured as follows: in Section 2, we introduce the definitions of the overall TDM framework, classifications of data types, and the technologies that can be used in text and data mining process. In Section 3, we show how TDM is potentially related to all sectors of economics, and we illustrate this assumption with examples and potential use cases of TDM implementation and integration across sectors. In Section 4, we draw a timeline of scientific, industrial and political documents that emphasise the importance of and interest in TDM and the need for uptake across sectors, and we summarise the key messages. We conclude the report with Section 5, outlining the importance of TDM for the future of the economy and science.

2. TDM FRAMEWORK

The quantity of digital content currently produced by society and the potential ease of access to this data at any location via easy-to-use devices (with reliable Internet connection, and mobile devices) increase the need for content mining technologies that boost human capacities in terms of speed of data processing (Mayer-Schoenberger and Cukier, 2013). TDM technologies can support individuals as well as larger institutions and companies in decision-making processes and in the development of novel ideas and solutions. In this section we discuss diverse data classifications, technologies that underlie TDM implementations, and the research communities that bring together the researchers and developers from both academia and industry to improve and advance TDM algorithms.

2.1. Data to be mined

TDM is not defined through a single scientific domain or theory; it is a complex set of technologies with different underlying theories and assumptions, consisting of components that are developed within diverse disciplines. As Figure 3 shows, at a macro level, the TDM framework is centred around data. As we enter the Big Data era (Mayer-Schoenberger and Cukier, 2013), the efficient storage and management of content represent a significant challenge of their own, both in terms of the amount of information and in terms of the heterogeneous nature of the data, which can be numeric, textual, multimodal, etc. The more complex the data, the more creative and innovative the TDM components have to be. Data can be structured according to its size and legal accessibility, according to its sources, or by who produces it the content that constitutes the data collections.

In Section 2.1.1, we provide an overview of the types of sources that can generate and hold data, to illustrate the breadth of the sources on which TDM may operate. In Section 2.1.2, we summarise a newly proposed classification of datasets according to data size and types of legal access, in order to identify the types of datasets that could be more easily accessed and mined by for-profit and non-profit organisations, in the case that copyright restrictions are adjusted.

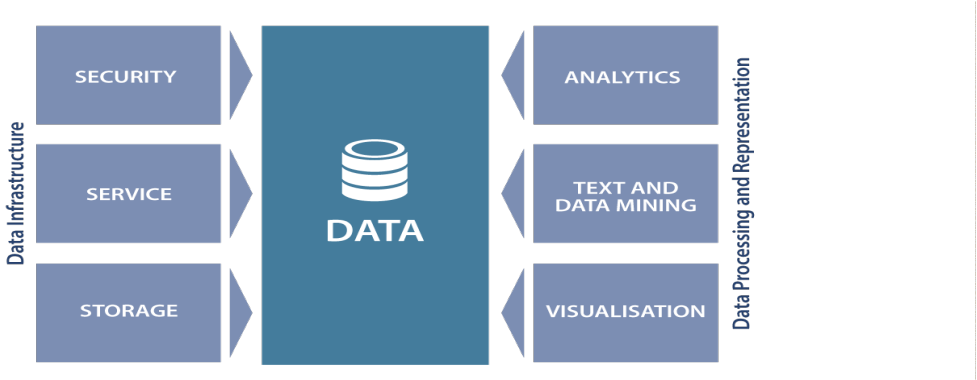

Figure 3. Macro view of TDM framework.[9]

[9] This Figure ('Macro view on TDM environment') by FutureTDM can be reused under the CC-BY 4.0 license.

2.1.1. Data sources and data holders

Data comes in countless forms. It is easy to think of data as intentionally collected text or scientific data, but data may be also produced by humans and businesses as they carry out datafied activities, leaving so-called "data exhaust" as a by-product of their actions and movements in the world, e.g. users' online interactions, corporate and personal websites, and released digital documents and archives (Mayer-Schoenberger and Cukier, 2013). Another type of data is represented by measurements of uncontrolled events happening in nature as recorded by weather sensors, telescopes, and other instruments across the globe.

Below we list different data sources, ranging from personal data, created or tracked from individuals, to professionally gathered and created content in forms of cultural knowledge data, scientific and business data, public sector information. All this data can be either shared publicly on the Internet, and stored offline. When shared publicly, the original content is enriched with a data description and further information that are available on a website.

Personal data

The traces that each Internet user leaves form a massive amount of information that has value when gathered and structured on a mass scale. Currently this information is mostly given by individuals either with or without their consent; in the first case often without them realising how profitable this information is, and in the second case, without them even realizing that their information is being collected in any way.

States also collect information about individuals in the country, and can often compel people to provide them with information, rather than having to persuade them to do so or offer something in return, as private companies would have to. This special status allows a state to gather large amounts of information that can be potentially mined by private or public institutions. This can then be used to optimise interactions between the state and individuals, to improve the services provided by the state, and to create business models around this data or to use this knowledge for the greater good of the society. At the same time, along with the access to personal data, states are responsible for protecting the privacy of the data producers when mining this data and releasing it to the general public.

Cultural heritage data

There has been enormous investment in the digitisation of heritage data, i.e. data that are considered worthwhile archiving, studying, or exhibiting. In 2010, it was estimated that it would take EUR 10 billion to digitise Europe's cultural heritage holdings.[3] By 2012, it was estimated that at least 20% of Europe's cultural heritage collections had been digitised[10]. At the same time libraries were investing in preserving digital collections and exploring ways to increase the accessibility of this content. Libraries use TDM to structure digital content, and to develop and explore further possibilities for knowledge discovery e.g. via topic modelling. Improvements in the accuracy of

[10] http://pro.europeana.eu/enumerate/statistics/results

optical character recognition (OCR), including for handwritten texts[11] will increase the impact and efficiency of applying TDM to digital content in order to enhance metadata or analyse collections. Google Books[12] and Google Art[13] operate on a worldwide scale, and digitise cultural (printed) heritage for their own technology development, as well as to make this data accessible to a wider audience. The HathiTrust Digital Library is a partnership model bringing together digitised collections from across US institutions and making them available to US researchers for analysis.[14] At a European level, Europeana, which collects metadata about heterogeneous data such as digitised artworks, artefacts, books, videos and sounds, provides access to this metadata and limited datasets via APIs[15]. The DARIAH[16] and CLARIN[17] infrastructures illustrate the value and potential of access to these extremely heterogeneous corpora for the research community. Both infrastructures support TDM across a wide-ranging corpora, e.g. CLARIN supports natural language processing and analysis of text (including newspapers, web fora, books), and speech (news broadcasts, sign language).

Scientific data

Scientific data consist of two main types: data that have been collected through experiments and studies (measurements of any kind, texts, videos, etc.), and the scientific publications that describe the work after analysis has been performed (containing data description, methods, experimental results, etc.).

Data used in experiments may be released to a general audience, along with the publication describing the study. This joint release of data, with corresponding publication and related metadata, is considered to have a beneficial side-effect: it promotes scientific research reproducibility and scientific integrity. An increasing number of researchers and international research organisations perceive this to be a positive step forwards (ICSU, ISSC, TWAS, IAP, 2015).[18]

Scientific data can be released purely as a data set collected for general purposes. In this way, researchers in business or academia can access the data and decide themselves on the type of experiments to run. An example of this approach is the open access release of the EU satellite data from Copernicus Earth observation[19]. The initiative to ask the questions and to answer them lies in the hands of the audience. Open data and data sharing in research is on the increase in Europe due to funder-mandates such as the H2020 Open Data Pilot, and the increasing availability of infrastructure for data sharing e.g. generic services such as EUDAT[20] and Zenodo[21]. On a worldwide

[11] http://transcriptorium.eu
[12] https://books.google.com
[13] https://www.google.com/culturalinstitute/project/art-project
[14] https://sharc.hathitrust.org
[15] http://labs.europeana.eu/api
[16] https://de.dariah.eu/tatom
[17] http://www.clarin.eu
[18] FORCE11 MANIFESTO: Improving Future Research Communication and e-Scholarship: https://www.force11.org/about/manifesto
[19] http://www.eea.europa.eu/highlights/eu-satellite-data-to-be
[20] http://www.eudat.eu

scale, an example is the Research Data Alliance (RDA)[22] , a forum where the infrastructures for data sharing across disciplines, institutions, and countries, are discussed, and guidelines are formulated. In addition, industry is also beginning to make their scientific datasets available, e.g. clinical trial data, often for ethical or transparency reasons.

Scientific publications can also be considered as a valuable dataset for TDM. These texts represent potential sources of new knowledge when the information they contain is summarised and analysed in comparison with other studies and lines of enquiry from other disciplines. Often containing figures and data visualisations, articles are of value not just because of the text they contain, but because they are important sources of raw data and context. TDM analysis of publications can help find connections between studies in different domains that are implicitly or indirectly connected (Swanson, 1986). In the field of bioinformatics, the literature has been mined to analyse the interactions of genes and proteins in various diseases. Often this type of analysis is performed on Medline[23] abstracts, as they are freely available but it has been noted that this is extremely limiting. Van Haagen et al. found that only 32% of known Protein Protein Interactions (PPIs) could be found in Medline abstracts, and therefore the rest could only be found by accessing the full text literature (van Haagen et al., 2009).

Scientific publications that are released via open access are, by definition, accessible for TDM and TDM-based applications. PubMed Central[24] has become a widely used source for TDM. The European open access publication infrastructure OpenAire[25] not only supports TDM by encouraging storing articles under open access licences and in interoperable machine readable formats, it also utilises TDM to infer links between data and publications, and to provide funders with data regarding publications arising from research funding and research trends[26]. Subscription content is less accessible as, with the exception of the UK, European researchers can only mine this content under conditions specified by licences. In contrast to this, in the US, researchers may mine content that they have legal access to, without the need for additional licence permissions.

Business data

Businesses are creating and collecting data from all possible sources, as mining data from combined sources improves insights into internal processes, strategies, and the market. These data can vary from energy consumption by business agents, such as shops and carriers, to personal information about customers and their digital and financial transactions. Most of this data constitutes value within the context of a knowledge-based economy, and is sensitive to data protection issues, thus, is used internally by companies and is not released publically.

[21] https://zenodo.org
[22] https://rd-alliance.org
[23] https://www.nlm.nih.gov/bsd/pmresources.html
[24] http://www.ncbi.nlm.nih.gov/pmc
[25] https://www.openaire.eu
[26] https://blogs.openaire.eu/?p=88

Public sector information (PSI)

PSI is data produced and made available by governmental institutions. Today, a growing number of governments realise that this data can bring more profit and benefit to society when released to a general audience. PSI data release initiatives allow commercial companies that are already experienced in data mining, to use this additional information source to build better products for their customers, consequently creating more jobs and boosting commercial success.

The European Union Open Data Portal[27] gives access to a growing range of PSI data from the institutions and other bodies of the European Union (EU); it is free for both commercial or non-commercial purposes. The EU Open Data Portal is managed by the Publications Office of the European Union. Implementation of the EU's open data policy is the responsibility of the Directorate-General for Communications Networks, Content and Technology of the European Commission.

Open data portals have also been set up by individual EU Member States.[28] The datasets of public documents that these portals share, are mostly under CC-0 license, i.e. the owners did not reserve any rights, and the data is free for sharing and use in any country, by any institutions. These datasets represent the documents in the native languages of each specific country.[29,30,31,32,33]

Internet data

Since the introduction of the first website in 1991, the World Wide Web had expanded to more than 45 billion pages with unique URLs in the late 2010s (Bosch et al., 2016). Many webpages contain valuable information that can be crawled for further mining. More generally, the Internet (the infrastructure on which the World Wide Web is based, but also services such as internet, email, and file transfer) allows access to large amounts of data of all types, that can be crawled with the proper credentials. Many data remains, of course, hidden behind passwords and firewalls. The so-called Deep Web, the part of the World Wide Web that is not directly accessible for indexing by search engines, has been estimated to hold at least 500 times as much data as the indexed ('surface') web.[34]

2.1.2. Data types

Any type of data considered to be properly available for TDM must first be available as, or converted to, a standardised and machine-readable form so that it can be easily processed. In their 2014 report to the European Commission entitled 'Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining', Hargreaves et al. (2014) identified five

[27] https://open-data.europa.eu/en/data
[28] http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=8340
[29] https://data.gov.uk
[30] https://data.overheid.nl
[31] https://www.data.gouv.fr
[32] https://www.govdata.de
[33] http://data.norge.no
[34] https://en.wikipedia.org/wiki/Deep_web

types of databases that lend themselves to TDM, according to their size and access policies, as well as their potential for revenue structure changes:[35]

1. **XXL**: all data behind firewalls, i.e. companies and organisations' internal databases, not accessible to the public.[36]

2. **XL**: all publicly accessible data not behind a firewall or paywall. e.g. freely accessible newspaper websites.

3. **L**: all publicly accessible data located behind a paywall. e.g. online newspaper pages that are accessible with a subscription fee.

4 **M**: all publicly accessible data behind a paywall whose clients are mainly researchers and companies in need of reports. e.g. Reuters, Bloomberg.

5. **S**: scientific publishers' data behind a paywall.

This classification covers data collections of different sizes that contain content of a diverse nature. We can make a distinction between high-level data that is likely to be subject to copyright protection, and low-level data that is highly unlikely to be protected under copyright law (although data protection law may apply in certain cases):

- Protected: text, image, sound, multimedia;
- Non-protected, unless the content becomes part of the database that the company or institution did put substantial investment into: e.g. clicks, likes, GPS coordinates, timestamps, and measurements of a different kind.

2.2. Technology and R&D

Text and data mining has gradually evolved into a large domain with complex approaches for a range of applications. Historically, the development of mining technologies has its roots in the mid-1700s, with the invention of mathematical models, regression methods, and statistical analysis. Due to the advent of computers and systems for storing larger quantities of data, the possibility to mine very large datasets and filter relevant information has greatly expanded what used to be laborious expert analysis work. Since the beginning of the 1990s, modern data mining tools have appeared on the market and in labs, allowing more advanced analysis, to discover hidden patterns and to extract implicit knowledge, as envisaged and tested by Swanson D.R. (Swanson, 1986). With the invention and adoption of the World Wide Web, researchers expanded their efforts to develop text and data mining methods to search and explore the web. This led to the creation of search engines, starting from simple interface-based browsers like Mosaic and Netscape, and developing into large-scale systems like Google and Bing. It represented the opening gambit towards more complex methods, not only based on keyword search, but also on their correlation, their similarity, their joint probabilities, and to the application of text mining to different fields (such as the biomedical domain,

[35] The original document uses the term 'database'; we replaced this by 'data'.
[36] These collections are outside the FutureTDM scope.

economics, social sciences, or humanities), and types of data and content (such as scientific publications and research data, social media, multimedia, or public sector information).

Figure 4 shows a zoom-in view of the TDM environment, outlining the main research fields that develop, test, and finally provide the algorithms for TDM implementations for each specific domain of use. The core scientific areas are ***machine learning***, which is machine learning of tasks and the discovery of patterns and structure in data; ***information retrieval and search***, which focuses on the optimisation of search methods that index vast amounts of content and allow the location and extraction of the most relevant facts for any question and user type; and ***natural language processing techniques***, which deal with human language in its textual and spoken form, covering humanity's past as well as reflecting current trends and topics. This scheme does not include an exhaustive list of techniques and research directions; its aim is to illustrate the scale and potential scientific interaction, and the broad knowledge required to create a successful, scientifically sound TDM application.

Below, we briefly introduce the main research directions that constitute the components of text and data mining applications.

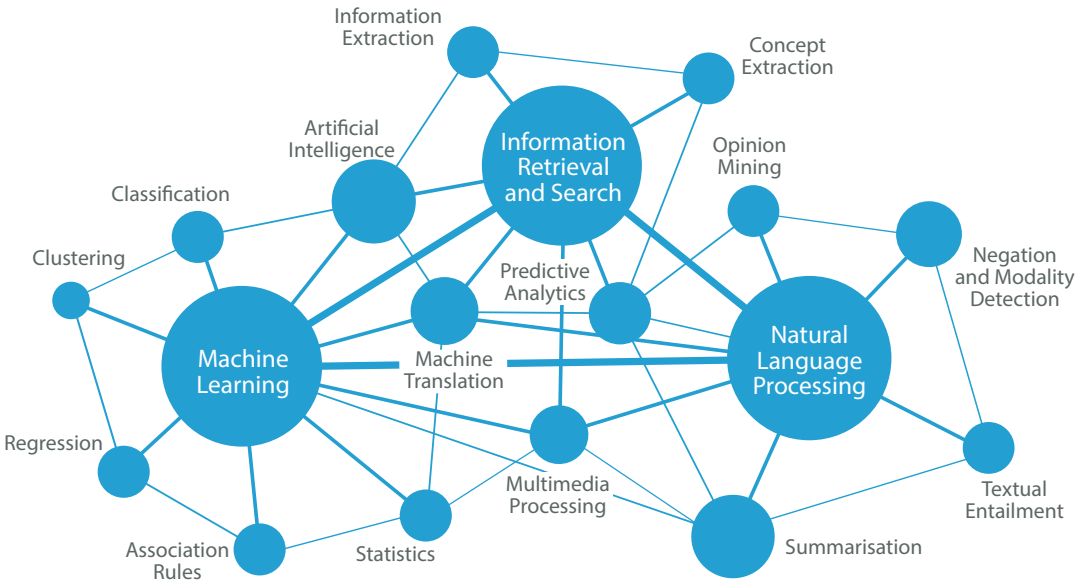

Figure 4. Micro view on TDM environment.[37]

Machine Learning (ML) studies the automated recognition of patterns in data, and develops algorithms able to learn tasks without explicit (human) instruction or programming. Learning is based

[37] This Figure ('Micro view on TDM environment') by FutureTDM can be reused under the CC-BY 4.0 license.

on data exploration and is aimed at generalizing knowledge and patterns beyond the examples that are fed to the algorithm at training stage.

Information retrieval (IR) can vary from simple retrieval of all the items (e.g. full documents) within a collection that contain user-requested words in metadata fields or structured databases, to more sophisticated search and ranking of the results based on the presence of combinations of important words and numbers in different sections of texts or data fields (Baeza-Yates and Ribeiro-Neto, 1999).

It is important to note that, although TDM uses information retrieval algorithms as one of the components within the TDM framework, it goes beyond the simple retrieval of whole documents or their parts. Within the IR framework, the task of knowledge processing and of new knowledge creation still resides with the user, while in the case of full TDM technology implementation, the extraction of novel, never-before encountered information is the main goal (Hearst, 1999).

Natural language processing (NLP) aims at developing computational models for the understanding and generation of natural language. NLP models capture the structure of language through patterns that map linguistic form to information and knowledge or vice versa; these patterns are either inspired by linguistic theory or automatically discovered from data. The NLP community is subdivided into groups that focus on spoken or written content, i.e. speech technologies (ST) and computational linguistics (CL). ST researchers work with audio signals, synthesising and recognising speech and paralinguistic phenomena. Once speech content has been transcribed in a textual representation, it can be further treated as text for TDM. CL researchers target syntactic and discourse structures; their algorithms help us understand the basic concepts, relations, facts, and events expressed in textual content.

The ***visualisation*** of the output of the ML, IR, and NLP algorithms is an intrinsic part of all the TDM constituent technologies. Visualisation is valuable for researchers in fields like bioinformatics where, for example, the mining of protein behaviour described in a set of publications is visualised in one scheme; or in text mining, when a word cloud, with words in different size and colours depending on their frequencies, gives a general impression of the most frequent terms used in a collection of documents.

Statistics, as a field of mathematics in general and machine learning in particular, provides its practitioners with a wide variety of tools for data collection, analysis, interpretation or explanation, and presentation.

Classification/Categorisation/Clustering of content within a collection in order to split (in a supervised or unsupervised manner) these into groups of certain types based on content, represent a TDM preprocessing step at the collection level, as these categories are used for further specific mining of the collection's content. Machine learning (ML) algorithms represent a dominant solution for this task, e.g. the Naïve Bayes approach (Lewis, 1998), decision rules (Cohen, 1995), and support vector machines (Joachims, 1998). While ***classification*** algorithms identify whether an object belongs to a certain group, ***regression*** algorithms estimate or predict a response as a value from a continuous set.

Association rule-learning algorithms allow for the discovery of interesting relations between variables within large databases (Agrawal et al., 1993). For example, when a large database of customer shopping behaviour is mined, association rules help to define what types of products or services are usually sold together. This information can help to adjust and tailor marketing campaigns, as well as make product arrangements in the stores more efficient.

Information extraction (IE) is a wide concept that covers the inference of information from textual data. Examples of types of extracted information are statistics on the usage of terms, the occurrence of named entities in text and their relations or associations according to certain classification schemes, and the presence of specific facts or logical inferences expressed in texts (e.g. interactions between genes, logical proofs, or the inference of logical consequences and entailment expressed in the text).

Term or Concept extraction is a first step for further IE, as it is necessary to define the terms of importance for the corpus to start its processing. Once these terms, being specific for the topic to conceptualise a given knowledge domain, are extracted, an ontology of the field can be created for further IE task implementations.

Negation and modality (or hedge) detection is an important aspect of TDM, as negation and hedging constructions in the text (such as "Our results cast doubt on finding X of Author Z" or "We found that finding Z does not apply to context Y") offer crucial evaluations of reported facts.

Sentiment analysis, or ***opinion mining***, aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, the affective state of the author, or the intended emotion the author intends to convey to the reader.

Summarisation of a given text or a set of texts is the task of creating a shorter text that captures the most crucial information from the original texts. First, the crucial content to be summarised is defined, and then, depending on the task, it is condensed into a new representation that may consist of sentences or parts from the original documents (*extractive* summarisation), or that may be a newly created text or multimedia document that rephrases the summarised message (*abstractive* summarisation).

Predictive analytics covers the methods that analyse previous usage or transactions statistics and, based on these results and trends, offers predictions on further systems development.

Machine translation targets human quality level of translation of written or spoken content from one language to another using computers. Although it originally started as rule-based systems taking linguistic knowledge into account, nowadays the pioneers in the field use Big Data size collections and deep learning algorithms to train their systems and to achieve performance improvements.

Multimedia processing allows further mining of all aspects of multimedia content, varying from the audio stream of a given video to visual concepts extraction and annotation, person and object discovery and localisation on the screen. This direction of research and applications has gained a

great deal of importance, due to the fact that contemporary Internet users share and create large quantities of multimedia content that, for example, reflect their opinions and spread information about events.

As depicted in Figure 4, the methods listed here are not insulated domains of research and application; they represent a complex interactive set of tools and methods that can be recombined in multiple ways, depending on the needs of the specific task and the availability of data and expertise.

3. TDM ACROSS ECONOMIC SECTORS

Since the beginning of humanity, the development of novel technologies has defined societal structure and its employment distribution. Agriculture and natural resource extraction form a primary sector, as this is an initial type of human joint-effort to produce food for survival. The invention of machines and elaborate tools led to the development of the secondary sector, which has evolved into complex machinery engineering, construction, and space exploration. The third sector comprises all the services that society members deliver to each other, varying from retail to transportation, and from medical healthcare support to entertainment. Until recently, this three component structure represented most of the societies in the world. However, the growth of data-driven business, services and research, has led to a discussion on the changes in the economic structure that should reflect a novel type of knowledge-based economy, where knowledge and data become a separate valuable commodity (OCDE, 1996). Thus recently, the research that supports knowledge sharing and growth, as well as education of the professionals that can carry out these activities, have been termed a quaternary economic sector. Moreover, the high level decision makers in governments, large industry companies, and education, who have the potential to shape the future of the entire sectors with their vision and following decisions, have been placed in a separate, quinary sector.

For the purpose of our report, we are interested in the current and potential presence of TDM in all economic sectors. We depict these relations in Figure 5. The primary, secondary, and tertiary sectors follow roughly the same type of interaction with TDM technologies. These sectors represent sources of all possible types of data that are available for mining, and at the same time, have tasks and challenges that require smart solutions. The quaternary sector is particularly central to TDM development and implementation, as it produces the TDM experts and advances the technology itself. Experts at the decision-maker level use TDM tools to test and prove their vision of economic development, as well as to hypothesise where a specific decision will lead to, and to measure and assess its progress.

In the sections below, we list examples per sector and subfield on how TDM is used, in order to illustrate the broadness of TDM uptake and its potential for development.

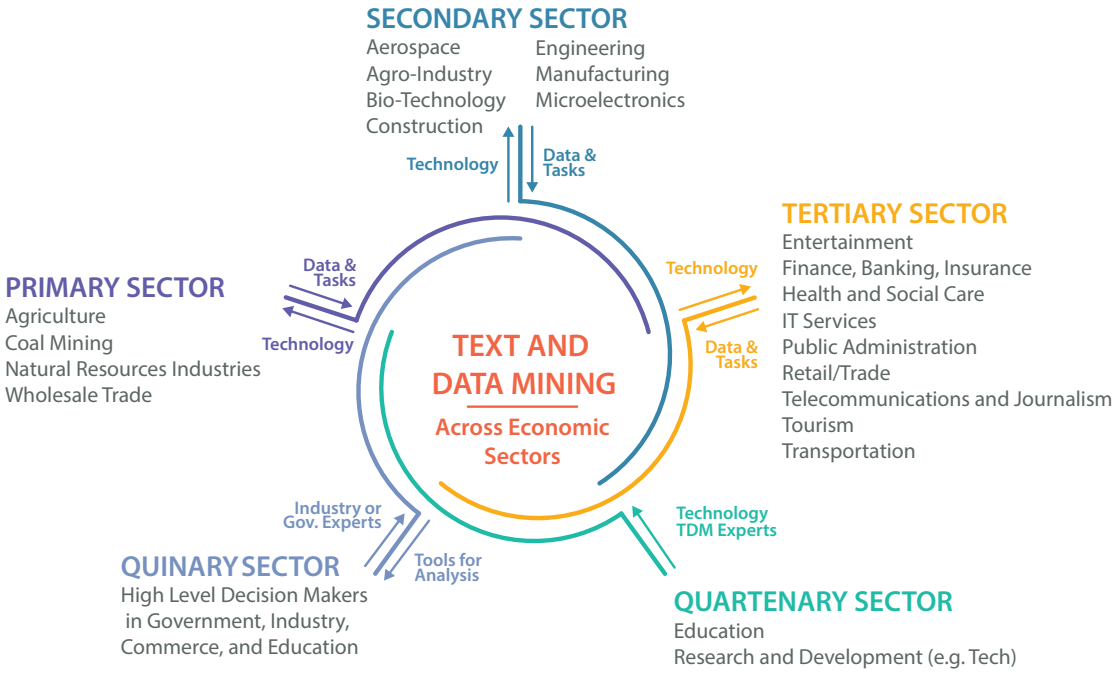

Figure 5. General economic structure and connection between TDM and all economic sectors.[38]

3.1. Primary sector

Making the best use of limited natural resources with a constantly growing population is one of humanity's ultimate challenges. Taking on the challenge requires the understanding of changing patterns in climate and environment; TDM has the potential to discover such patterns automatically from data. The primary data that can be used for TDM are collected in the form of satellite measurements, weather sensors on the ground and below, in the air, and in the water. Other data are derived from economic indicators from the primary sectors: production numbers, profits, stock market value patterns of mined goods such as oil, coal, and gold, and crops such as wheat[39].

3.2. Secondary sector

Any large-scale manufacturing plant or distributed network of pipelines is packed with various types of sensors that produce a continuous stream of measurements that are crucial for maintenance analysis, such as temperature, pressure levels, leak detections, etc. When analysed on a large scale, this information offers a basis for discovering and understanding strategies for cost savings and safety measures. For example, calculating optimal routes for product delivery can boost fuel efficiency, leading to a smaller ecological footprint as well as to costs savings; knowledge of electrical grid usage or gas pipelines integrity helps to schedule maintenance of the most sensitive parts of the

[38] This Figure ('General economic structure and connection between TDM and all economic sectors') by FutureTDM can be reused under the CC-BY 4.0 license.

[39] http://www.agroknow.com/agroknow

system (Mayer-Schoenberger and Cukier, 2013), (Auclair D., 2016); data-driven smart grid applications can significantly lower CO_2 emissions (ICSU, ISSC, TWAS, IAP, 2015).

3.3. Tertiary sector

The sector of provided services embraces TDM at a different pace depending on the TDM expertise and awareness in the field, BIg Data availability, and the sensibility of this data in terms of privacy, as much more personal data is mined within this sector.

Companies allocate great effort to interacting with their existing customers, as this allows them to monitor customer feedback on the products, and at the same time to profile customers for further targeted product promotion. On the other hand, information about users' Internet surfing behaviour is gathered in the form of click logs, website views and search queries. Once this data is categorised and profiled, the advertising network industry can match the right advertisement from the right company to the suitable client profile. Sentiment analysis of customer comments in global social networks, such as Twitter and internet forums, allows companies to track customer interest in their products, as well as to use these channels for targeted advertising of services and products.

3.3.1. Finance, Banking, Insurance

In the finance domain, predictive analytics play a crucial role. TDM based analyses help to manage risks and to make informed decisions. Banks and insurance companies use TDM to build consumer profiles and develop automatic classifiers to determine eligibility for financial products, and to set insurance premiums. Customer parameters can include diverse types of information gathered from different sources, ranging from consumers' zip codes, employment background and educational history, to their shopping history, social media usage and friendship network (Ramirez E. et al., 2016). Financial institutions typically purchase this information from consumer reporting agencies (CRA) that crawl data from the Internet and from special government and private databases.

3.3.2. Health and social care

Both health care providers and consumers benefit when treatment plans and medical advice are tailored to individual patients. TDM has gradually become part of the health care system: in routing and treatment tailoring, treatment monitoring, as well as in claims processing and insurance payments (Auclair D., 2016). In addition, Big Data is used by different medical institutions to predict the likelihood of hospital readmission for different categories of patients and types of disease combinations.

Medical treatments and predictions suggested by artificial intelligence and TDM are being integrated in decision support systems for doctors (e.g. Babylon[40] in the UK, IBM Watson[41] in the USA). Such systems, powered by AI, can help to lower the costs in regions with a shortage of specialty providers.

[40] http://www.wired.co.uk/news/archive/2014-04/28/babylon-ali-parsa

[41] http://www-03.ibm.com/press/us/en/presskit/27793.wss

Moreover, medical TDM systems can flag prescription anomalies based on patient records, thus helping to prevent medication errors (Auclair D., 2016).

3.3.3. Public administration

Public administration collect data about different aspects of the lives of the members of society, such as ownership of movable and immovable assets, use of the health care system, employment status, and family situation. This information is used to keep track of societal developments and changes, and to come up with adjustments to policies and reallocation of services. A substantial amount of this information stays offline and is not interconnected. If this content would be made open to the public through open government initiatives, many types of businesses and non-profit organisations would have a chance to put this data to use, e.g. by connecting the PSI data to their own data, or building new products and services on top of the released PSI (Mayer-Schoenberger and Cukier, 2013).

Like the finance industry, public administration can benefit from internal TDM implementations, as intelligent mining and automated auditing of the data can help to detect and to prevent fraud, currently leading to financial losses in tax collection; TDM may generally help to optimise the workflow.

3.3.4. Retail

In retail, tracking and mining of purchase transactions by customers allows companies to adjust the range of products on display according to the preferences of the customers in a particular area. In addition to better fitted product selection, food retailers can adjust their facilities arrangements, for example, refrigerator temperatures, based on large-scale statistics of all the shops in a chain, which in turn helps to save on the energy bills.

Some market places are set up as sales platforms for individuals and companies in completely virtual environments, like international giants eBay[42] and Amazon[43] or the French smaller size equivalent, Le Bon Coin[44]. In the context of online shopping, the mining of user behaviour (clicks, products selected or bought, time spent on the page, browsers used, time of the day when the website is accessed) is increasingly being used for product recommendations and reminders, which eventually increases sales.

3.3.5. Telecommunications and journalism

The telecommunications industry generates some of the largest multimedia datasets, and TDM helps to filter and to play already available content that is of interest to clients depending on their preferences. Mining customer preferences, in turn, influences decisions on allocating resources to the creation of new content that will potentially be successful.

[42] http://www.ebay.com
[43] https://www.amazon.com
[44] http://www.leboncoin.fr

Journalists always needed diverse and reliable information sources to report news events and to bring an analytical perspective to a general audience. Nowadays, journalists have to deal with an avalanche of data sources, thus mining is seen as a valuable tool for investigative journalism to find, confirm, and eventually to prove stories in a more transparent way (Diakopoulos N., 2014). Mining of blogs, websites and twitter feeds makes journalists aware of their audience's interests and level of knowledge, which helps them to shape their social interaction accordingly and to select relevant news pieces (Flaounas I. et al., 2013). At the same time, the journalists have to be careful when redefining their agenda using TDM. They need to cumulate the knowledge gained from TDM with the actual state of society, as not all society members are represented in the social media or web data available for crawling. In their work and use of data, journalists have to find a balance between data personalisation and the ecology of common knowledge (Lewis S.C. and Westlund O., 2014).

3.4. Quaternary sector

This information and knowledge-rich sector focuses on knowledge gathering, processing, and creation, via research practices and teaching society members. In this sector, TDM algorithms are both the subject of research and a set of tools that enable research.

3.4.1. Research

The speed of scientific progress depends to a large extent on the framework that supports the sharing of novel ideas and key findings within and across communities, the reproducibility and comparability of results, and the research impact measurements. TDM techniques have the potential to add value and increase efficiency for all these components.

As mentioned in Section 1 and illustrated with examples in Figures 1-2, researchers nowadays have to keep track of an enormous number of publications to retain an up-to-date understanding of the research in their field. Moreover, a growing amount of research happens at the intersection of several domains, or methods are being adopted across domains. Thus, the current volumes of cross-domain content render them infeasible to be read by humans. TDM can help researchers discover new connections and trends, allowing them to more optimally use their time and other resources.

The quality of scientific results shared within the scientific community relies and heavily depends on the systematic reviewing of submitted articles for publication. This rigorous process guarantees that the research outcome, when accepted, brings novel information to the field of knowledge, and the results have been achieved following the standard procedures that safeguard reproducibility and reliability. Ideally, high quality reviewing relies on full awareness of all the publications in the field. However, with the increasing number of published studies at a global level, it becomes unfeasible for any human researcher/reviewer to keep track of all studies that have been reported. Therefore, TDM can provide valuable support for reviewers, as it can reduce the workload and minimise any potential bias.

Furthermore, the quality of published research can be measured with using TDM. For example, smart mining of publication references that takes into account the discourse structure in case of each quotation, yields a more subtle and informative overview of the whole citations network.

Clinical studies require a meaningful sample size of patients with certain conditions in order to test the effectiveness of new drugs or medical procedure. Pharmaceutical companies that develop these drugs and treatments rely on intermediary companies that carry out the actual screening of medical records for suitable potential candidates. These intermediaries work on a large scale and rely on TDM technologies to establish patient-clinical matches, as well as to support post-trial pharmacovigilance through early detection of adverse drug events.

3.4.2. Education

Traditional educational institutions can use BIg Data techniques to identify students for advanced classes, while at the same time determine students who are at risk of dropping out or might be in need of early intervention strategies. The Gates Foundation has invested in a number of programmes to improve the quality and availability of educational data with the aim of developing policy and practice to improve the impact of education, and to boost student achievement[45].

Massive open online courses (MOOC), introduced in the late 2000s, are currently reshaping the whole concept of distance education, and have a potential to impact the whole educational system, as they make it possible to monitor and assess individual class performance based on large-scale mining of student profiles, attendance, task completion, and forum discussions.

3.5. Quinary sector

Funders and high-level decision makers rely on their general visions of technology and societal development paths, as well as the impact-metrics that assess potential economic, political, and societal value. These experts can use TDM to mine the current status of affairs, and to predict the outcomes of diverse types of interventions. Moreover, sentiment analysis of social networks activities can provide them with fast feedback from the population on actions taken.

[45] http://www.gatesfoundation.org/Media-Center/Press-Releases/2009/01/Foundation-Invests-in-Research-and-Data-Systems-to-Improve-Student-Achievement

4. SUMMARY OF THE LITERATURE

As text and data mining harbour a large potential across practically all sectors of the economy, their legal, economic, and political implications have been discussed in several expert reports and from the point of view of various stakeholders. This report is based on several of these reports; in this section we summarise the most prominent reports published in recent years for further reading and reference. We highlight the points which call for urgent legal and economic changes that would allow TDM to unleash its full potential. We also highlight cases where uptake success is mentioned, and we outline the direction in which discussions develop. Figure 6 depicts the general timeline of reports, their research questions and focus (Legal aspects, Economic factors, Examples, Technology). It is worth noting that reports created on behalf of governments and academia are more often released via open access, even when produced by private companies. Therefore, they are available for analysis, while private sector reports, usually created as a product for sale, stay behind a paywall, restricting our access to their content.

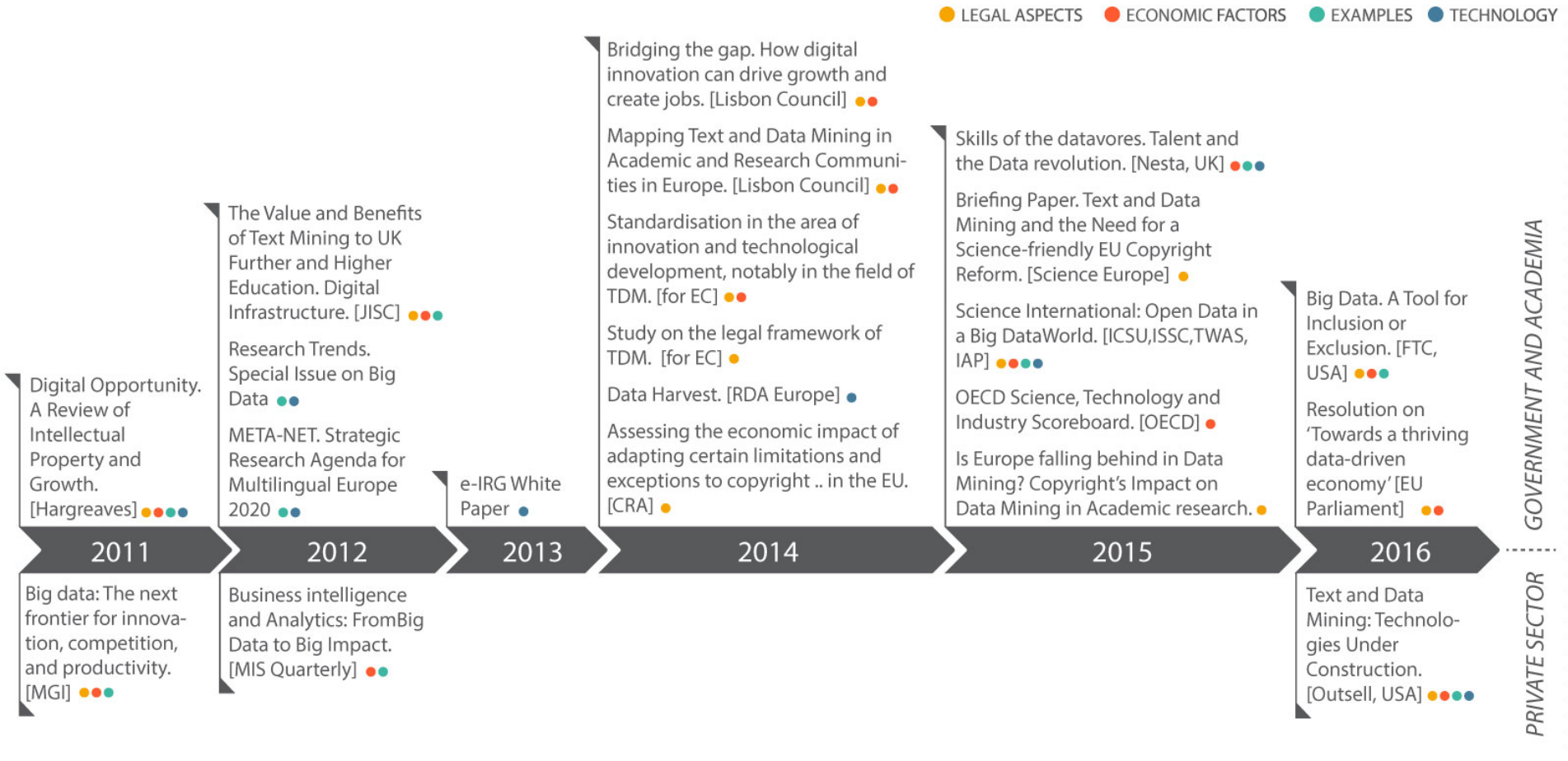

Figure 6. Timeline of reports.[46]

Year	*Title*	*Authors*	*Research questions/ Focus*	*Data and methods*	*Key findings*
2011	*Digital Opportunity. A Review of Intellectual Property and Growth.*[47]	I. Hargreaves	This report aims to find economic evidence to support changes in the Intellectual Property laws at the level of the UK, and further at a unified EU level.	Based on analysis of previous publications, reports, reported legal and economic practices.	Two important areas of utility of TDM: (1) cost savings, (2) wider economic impact generation
2011	*BIg Data: The next frontier for innovation, competition, and productivity.*[48]	*McKinsey Global Institute (MGI)* J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers	To explore Big Data potential when used in every economy sector, its innovation potential, and the added value.	Based on analysis of previous publications, reports, reported legal and economic practices, as well as internal MGI research.	This report gives a general perspective on the Big Data TDM being incorporated into all economic sectors, as datafication reaches all sectors of the global economy. The examples of potential profits, as well as raising issues, are listed for both EU and USA markets. - Issues: data policies in the context of Big Data, privacy protection; storage and access facilities and novel techniques that need to be implemented. - Big Data and its processing create additional value for businesses in a number of ways: improved and more user targeted service quality, overall transparency of performance, and the support of human decisions with reliable analytics.
2012	*The Value and Benefits*	D. McDonald, U.	To explore costs,	Consultations with stakeholders and	- The availability of material for text mining is limited. Even in

[46] This Figure ('Timeline of reports') by FutureTDM can be reused under the CC-BY 4.0 license.

[47] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32563/ipreview-finalreport.pdf

[48] http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

	of Text Mining to UK Further and Higher Education. Digital Infrastructure. JISC.[49]	Kelly	benefits, barriers and risks associated with text mining within UK further and higher education	case studies (literature review in systems biology, text mining to expedite research, increase of accessibility and relevance of scholarly content.	case of early adopters in the fields of biomedical chemistry research, the mining is limited to the Open Access documents. - Main benefits are increased researchers' efficiency, improved research process and quality. The report demonstrates how much money in terms of person time can be saved or used for proper research tasks when the scientific content is mined automatically, leaving the researchers with analysis challenges and pure research questions. - Barriers: legal uncertainty, inaccessible information.
2012	*Business intelligence and Analytics: From Big Data to Big Impact.*[50]	*MIS Quarterly*	The main focus of this revue is to describe and discuss Business intelligence and Analytics (BI&A) implemented for diverse market applications that use many data types (from user-generated content to Federal Reserve Wire Network information)	- Bibliometric study of critical Business intelligence and Analytics publications	- BI&A in its current state (2.0) processes web-based unstructured data, and the next step for the technology (3.0) will be to use mobile and sensor-based content.
2012	*Research Trends. Special Issue on Big Data.*[51]		*This special issue has a broad perspective varying from applications scenarios*	--	- Bibliometrics analysis confirms that 'Computer Science' and 'Engineering' are the leading topics in Big Data publications, and US leads the publications race in the field.

[49] https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining
[50] http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf
[51] http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf

			to infrastructure issues for TDM in Big Data context		
2012	META-NET. *Strategic Research Agenda for Multilingual Europe 2020.*[52]	*META Technology Council*	This report focuses on the current state of and the further potential for development of language technologies on a global scale, and more specifically, the European landscape.		- Multilingualism is a an important feature of Europe, as it means that there is need for the diversification and localisation of the tools that will be able to provide services at the same level for the whole variety of European languages and customers.
2013	*e-IRG White Paper.* [53]	*e-IRG*	This report focuses on the infrastructure initiatives that have already been taken and especially on those that are urgently needed in order to support TDM on a large scale across communities.		- Infrastructure needs support at national and EU level, and requires collaboration with business. - Open science should be supported through and via shared infrastructures.
2014	*Bridging the gap. How digital innovation can drive growth and create*	P. Hofheinz, M. Mandel	- This document outlines the growth of data at a global scale,	Based on analysis of previous publications, reports, reported legal and economic practices.	- Larger revolution in all economic sectors and aspects of societal life are predicted to be due to the Big Data era, but most importantly these will be shaped by data analytics for this

[52] http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf
[53] http://e-irg.eu/documents/10920/11274/annex_5.2_e-irg_white_paper_2013_-_final_version.pdf

	jobs. Lisbon Council Special Briefing.[54]		and aspects of data-driven economy like better conditions for creating an inclusive economy for all society members, better potential for cheaper individual education, and training for advanced manufacturing. - The authors focus on the 'data gap' between Europe and North America. - Strong focus is given to the discussion on policy regulations that need to be adjusted across the Atlantic, both in the EU and the USA.		data. - Europe is behind countries like South Korea, US, Canada, in terms of data use. - There are successful examples of open access policy introduction for public data. In Estonia, the state released publicly owned data including utilities and cell phone networks after having stripped private information and anonymization, which had positive impact on the development of data-driven businesses. - The authors argue that if the European data protection statutes are not lightened, the cost of data usage will only rise, which consequently will widen the gap between Europe and the USA, as well as the rest of the world.
2014	*Mapping Text and Data Mining in Academic and Research Communities in Europe. Lisbon Council Special Briefing.*[55]	S. Filippov	- This report continues the above-mentioned work on gathering information on the importance of TDM for the global economy	Based on analysis of previous publications, reports, reported legal and economic practices, as well as interviews, and bibliography and patent based analysis on ScienceDirect database.	- The difference in copyright access across the countries is reflected by the number of publications on text and data mining subjects. The US is leading the field, while the only European country at the top of the list is the UK, which has an exception for content mining for academic purposes. - The number of publications on TDM in English is several

[54] http://www.progressivepolicy.org/wp-content/uploads/2014/04/LISBON_COUNCIL_PPI_Bridging_the_Data_Gap.pdf

[55] http://www.lisboncouncil.net/publication/publication/109-mapping-text-and-data-mining-in-academic-and-research-communities-in-europe.html

			development, and focuses on the reasons why EU lags behind the US and some Asian countries in this domain, and hypothesises how this can change either positively or negatively, depending on the measures taken.		orders higher than in any other languages, covering 97% of the publications. - TDM is patented as algorithms in the US and other jurisdictions, while in Europe TDM is patented as being embedded in the software. China is rising strongly in this market in terms of patents.
2014	*Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. Report from the Expert Group, European Commission.*[56]	*Expert Group* I. Hargreaves, L. Guibault, C. Handke, P. Valcke, B. Martens, R. Lynch, S. Filippov			- This report lists the stakeholder types and the legal and economic issues that they encounter both in the EU and in the rest of the world. - The experts hypothesise on how the value chain might be adjusted when TDM is fully incorporated and legally allowed, arguing that it should have a more positive impact overall. The data is categorised according to both size and legal access pattern.
2014	*Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU. Analysis of specific policy options.*[57]	*Charles River Associates (CRA)* J. Boulanger, A. Carbonnel, R. De Coninck, G. Langus.	This report focuses on economic impacts of specific policy options in several topics of interest (digital preservation of and remote access to		-Introduction of IP exceptions for the case of TDM for science should not significantly adversely affect rightholders' incentives for content creation, content quality and TDM-specific investments.

[56] http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf
[57] http://ec.europa.eu/internal_market/copyright/docs/studies/140623-limitations-economic-impacts-study_en.pdf

			cultural heritage, e-lending for libraries, TDM for scientific purposes), in view of providing policy guidance on these topics.		
2014	*Study on the legal framework of text and data mining (TDM).*[58]	*De Wolf & Partners for the European Commission*	The focus is on the legal aspects of TDM.		Data analysis is the most encompassing term and is more preferable for use in a legal context. Classification of 4 different levels of access: "all to all" (web-date) / "many to many" (social networks) / "one to many" (contractual clauses) / "one to one" (confidential agreements)
2014	*The Data Harvest: How sharing research data can yield knowledge, jobs and growth.*[59]	*RDA Europe* F. Genova, H. Hanahoe, L. Laaskonen, C. Morais-Pires, P. Wittenburg, J. Wood	The report focuses on the measures that should be taken in order to support research using Big Data.		- Main incentives of the report include: -- data plan management; -- promotion of literacy in terms of all the aspects of Big Data use across society; -- develop tools and policies that support data sharing.
2015	*OECD Science, Technology and Industry Scoreboard 2015. INNOVATION FOR*	*Organisation for Economic Co-operation and Development,*	- This report aims to provide policy makers and analysts with the means to compare		- This is a broadly scoped report that has sections that discuss trends and features of knowledge economies. - It gives perspectives on the overall scientific excellence of EU countries compared to the rest of the world across all fields of

[58] http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf
[59] https://rd-alliance.org/sites/default/files/attachment/The Data Harvest Final.pdf

	GROWTH AND SOCIETY.[60]	*Europe*	economies with others of a similar size or with a similar structure, and to monitor progress towards desired national or supranational policy goals. - The main focus of this report is on the knowledge-based economies' development trends.		science, and the amount of investment in different domains, with special highlights on the new generation of ICT-related, Big Data based technologies and patents. - Across many metrics, European countries when combined are comparable to the results of USA, Japan, China, South Korea.
2015	*Skills of the datavores. Talent and the Data revolution.*[61]	*Nesta, UK* J. Mateos-Garcia, H. Bakhshi, G. Windsor	Introduces its own classification of the companies that work with Big Data, and then analyses the types of skills data scientists and engineers have to learn and master to boost the economy and scientific development. The analysis focuses on the UK market while the company	Internal NESTA analysis of the data analytics market and hiring strategies and challenges in the field that are based on more than 50 in-depth interviews (in the fields of manufacturing, retail, ICT, creative media, financial services, pharmaceuticals), collaboration with Creative Skillset.	- Classification of companies connected to data processing: -- 'Data-active' companies that are either data-driven (Datavores), work with large volumes of data (Data Builders), or combine data from different sources (Data Mixers), -- Dataphobes: other companies that restrict themselves to small size datasets. - These different types of companies have different behaviours while setting up their requirements, and then hiring the data experts. For example, datavores and data builders recruit most of their candidates from Computer Sciences, Engineering, and Mathematics, while Business disciplines are the source for management analytics positions.

[60] http://www.oecd.org/science/oecd-science-technology-and-industry-scoreboard-20725345.htm

[61] https://www.nesta.org.uk/sites/default/files/skills_of_the_datavores.pdf

			classification can be used globally.		
2015	*Briefing Paper. Text and Data Mining and the Need for a Science-friendly EU Copyright Reform.*[62]	*Science Europe*	Summarises the arguments in the debate for copyright law reform that should ensure that legally-accessed content could be freely mined without additional permission and cost within the EU.	Based on analysis of previous publications, reports, reported legal and economic practices.	- EU regulations for database protection (Directive 1996/9/EC) seriously impede researchers who aim to access the content to be mined via a database, as downloading a substantial part of the database content needs to be authorised by the database owner, even if none of the content is under copyright. - Scientific publishers start to introduce licences regulating the mining of subscribed content, which further limits the researchers in their immediate interaction with the content. - Exceptions in copyright law in countries like UK, Japan, USA, which allow their representatives to avoid difficulties with the legality of content mining, which puts EU researchers at a disadvantage, and complicates potential for international collaborations. - The authors suggest the following steps for research organisations: -- Avoid requesting that their researchers abstain from TDM; -- Refuse to sign TDM-licenses as part of subscription contracts, or negotiate more favourable condition; -- Empower their employees internally through training and TDM support, and externally by advocating copyright changes at national and European level.
2015	*Is Europe falling behind in Data Mining? Copyright's Impact on*	C. Handke, L. Guibault, J.-J. Vallbe	This paper focuses on the copyright arrangements	The authors measure research output in the number of publications in academic journal	Demonstrates that researchers from the EU/EEA countries have weaker performance in terms of publications in the growing field of data mining. This is partially due to the comparatively

[62] http://www.scienceeurope.org/uploads/PublicDocumentsAndSpeeches/WGs_docs/SE_Briefing_Paper_textand_Data_web.pdf

	Data Mining in Academic research.[63]		impacting data mining by academic researchers.	articles in Thomson Reuters's Web of Science (WoS) database.	strong copyright protection laws in these countries.
2015	*Science International: Open Data in a Big Data World.*[64]	*International Council for Science (ISCU), International Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP)*	-This report follows an accord of International Scientific organisations and focuses on the opportunities that the Big Data reality represents. - It discusses how text and data mining of this scientific content on a mass scale can help when the science addresses global challenges such as infectious disease, energy depletion, migration, sustainability and the operation of the global economy in general.	Based on analysis of previous publications, reports, reported legal and economic practices.	- Unprecedented opportunities for scientific development in current digital age rely on 2 pillars: Big Data (large volumes of complex and diverse data streaming in real time) and Linked Data (separate datasets logically related to a particular phenomenon allow computers to infer deeper relationships). - For the scientists, the open access to the scientific data and publications is a much-needed imperative. Therefore, authors insist on deregulating open data for scientific data usage. Data should be "intelligently open", i.e. discoverable on the Web, accessible and intelligible for further use, and assessable in terms of data producers' quality. This openness will also help replicability tests that test and scrutinise the reported results. -The authors note the ongoing discussion whether public data should be freely available to everyone, or just to the non-profit sector. They argue that it is primarily not appropriate or productive to introduce this discrimination in data access, and more importantly, robust evidence is accumulating to support the claim that this access to data for everyone brings diverse benefits, and broader economic and societal value. - The authors give examples of cross-country collaborations in open access initiatives in South America and Africa, while noting that sharing data allows more uniform development of technologies and their implementation on a global scale.

[63] http://elpub.architexturez.net/system/files/15_HANDKE_Elpub2015_Paper_23.pdf
[64] http://www.icsu.org/science-international/accord/open-data-in-a-big-data-world-short

2016	*Text and Data Mining: Technologies Under Construction.*[65]	*Outsell market performance report, USA* D. Auclair	- Focus of this report is on TDM aspects for scientific research, healthcare, and engineering.	The authors base the report on approximately 20 in-depth interviews with a range of stakeholders like content vendors and technology and tool providers, as well as on previously published information.	- The authors highlight the fact that data mining has been used longer for market research and commercial activities, as corresponding commercial websites contain large quantities of structured data, while academia and industry have to deal with high volume of unstructured texts and data, which made their progress in uptake slower. - The report provides a list of providers of TDM-related functions and examples of types of business that they are serving. - The authors list a number of actions for content providers and TDM tool vendors to follow, that could help them to benefit from TDM uptake in the near future: -- TDM solutions should be scalable -- Stored data should follow the standards at collection and storage stages. -- Privacy and security regulations are to be ensured.
2016	*Big Data. A Tool for Inclusion or Exclusion? Understanding the Issues.*[66]	*Federal Trade Commission (FTC) Report, USA* E. Ramirez, J. Brill, M.K. Ohlhaussen, T. McSweeny	- This report argues that it is not a question of *whether* companies should use Big Data analytics, but *how* to it that correctly, minimising legal and ethical risks. - Highlights need to identify and avoid potential biases and inaccuracies in TDM	FTC held a public one-day workshop 'Big Data: A Tool for Inclusion or Exclusion?' In order to get a feedback from involved stakeholders on the potential of Big Data in terms of created opportunities for consumers and policy issues, namely security breaches, and biases and inaccuracies in proper population/events representation in the data.	- The report includes a list of questions that are to be considered and investigated when companies are using Big Data analytics in their business models. This procedure could help them to avoid violating laws in terms of discrimination, and to prevent errors that might occur due to training data set inconsistencies that could lead to potential biases in offers and services. List of questions: - How representative is your data set? - Does your data model account for biases? - How accurate are your predictions based on Big Data? - Does your reliance on Big Data raise ethical or fairness

[65] http://www.copyright.com/wp-content/uploads/2016/01/Outsell-Market-Performance-22jan2016-Data-and-Text-Mining-Licensed.pdf

[66] https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

			applications, so that the Big Data is used to create novel opportunities for societal improvements, and especially to support low-income and underserved communities.		concerns?
2016	*European Parliament resolution on 'Towards a thriving data-driven economy'*.[67]	*EU Parliament* J. Buzek on behalf of the Committee on Industry, Research and Energy	This document aims to structure the knowledge-based economy landscape, with a focus on the EU, and to outline what are urgent measures need to be taken in order to be the leaders in the field at a global scale.	Based on analysis of previous publications, reports, and internally produced financial estimations of the market potential.	- This resolution outlines Big Data and data-driven economy potential till 2020, with a focus on the European issues and needs. - Announcement of the European 'Free Flow of Data' initiative - Stresses the need for a modern ICT infrastructure at a European scale.

[67] http://www.europarl.europa.eu/sides/getDoc.do?type=MOTION&reference=B8-2016-0308&language=EN

5. CONCLUSION

The uptake of text and data mining is growing across all sectors of knowledge-based economies, as the strengths of TDM and its economic advantages, such as time saving and scaling, are increasingly well recognised (ICSU, ISSC, TWAS, IAP, 2015), (Auclair D., 2016). While the development of TDM by the larger IT companies is directed at global markets, there is a long tail of TDM research and development in industry aimed at specific markets and niches. Most TDM industries makes use of proven TDM technologies, such as supervised machine-learning algorithms for classification or risks assessment (Ittoo et al., 2016), (Fleuren and Alkema, 2015), however there is a healthy relationship between industry and academic research on the development and uptake of new methods from fields like Artificial Intelligence (e.g. *deep learning* algorithms).

In general, industries have no trouble finding access to data, and manage to use TDM for economic profit. However, different types of barriers still occur. If barriers to the availability of full texts and datasets from academic domains and public sector information were lifted, the first applications that could be made possible would finally revive old visions from the second half of the 20th century on accomplishing automatic unconstrained scientific discovery from observational data (Simon et al., 1981), (Simon et al., 2007).

Large-scale global companies like Alphabet (Google), IBM, and Microsoft, currently invest heavily in TDM algorithms and infrastructure development, as much of the value they accrue nowadays is leveraged from the data provided to them by their customers, in combination with all the crawlable content on the web. For companies and research institutes based in the European Union, a competitive strategy necessarily involves a wider uptake of TDM, based on a better awareness of the potential of TDM across all economic sectors.

REFERENCES

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. ACM, Washington, D.C., USA, pp. 207–216.

Auclair D., 2016. Text and Data Mining: Technologies Under Construction. (Outsell market performance report).

Baeza-Yates, R.A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Cohen, W.W., 1995. Learning to classify English text with ILP methods. Advances in inductive logic programming 32, 124–143.

Diakopoulos N., 2014. Algorithmic accountability. Journalistic investigation of computatiional power structures. Digital Journalism 3, 398–415.

Filippov, S., 2014. Mapping Text and Data Mining in Academic and Research Communities in Europe.

Flaounas I., Ali O., Landsdall-Welfare T., De Bie T., Mosdell N., Lewis J., Cristianini N., 2013. Research methods in the age of digital journalism. Massive-scale automated analysis of news-content-topics, style, gender. Digital Journalism 1, 102–116.

Fleuren, W.W.M., Alkema, W., 2015. Application of text mining in the biomedical domain. Methods 74, 97–106. doi:10.1016/j.ymeth.2015.01.015

Hargreaves I., Guibault L., Handke C., Valcke P., Martens B., Lynch R., Filippov S., 2014. Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. Report from the Expert Group.

Hearst, M.A., 1999. Untangling Text Data Mining. College Park, Maryland, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99) 3–10.

ICSU, ISSC, TWAS, IAP, 2015. Science International: Open Data in a Big Data World. International Council for Science, International Science Council, The World Academy of Sciences, InterAcademy Partnership.

Ittoo, A., Nguyen, L.M., van den Bosch, A., 2016. Text analytics in industry: Challenges, desiderata and trends. Computers in Industry. doi:10.1016/j.compind.2015.12.001

Joachims, T., 1998. Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, in: Proceedings of the 10th European Conference on Machine Learning. Springer-Verlag, pp. 137–142.

Lewis, D.D., 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in: Proceedings of the 10th European Conference on Machine Learning. Springer-Verlag, pp. 4–15.

Lewis S.C., Westlund O., 2014. Big data and journalism: Epistemology, expertise, economics, and ethics. Digital Journalism 3, 447–466.

Mayer-Schoenberger, V., Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work and Think. John Murray Publishers.

Motion for a resolution further to Question for Oral Answer B8-0116/2016 pursuant to Rule 128(5) of the Rules of Procedure on "Towards a thriving data-driven economy" (2015/2612(RSP), 2016.

OCDE, 1996. The knowledge-based economy. Organisation for Economic Co-operation and Development, Paris.

Ramirez E., Brill J., Ohlhaussen M.K., McSweeny T., 2016. Big Data. A Tool for Inclusion or Exclusion? Understanding the Issues. (Federal Trade Commission (FTC) Report).

Simon, H.A., Langley, P.W., Bradshaw, G.L., 1981. Scientific discovery as problem solving. Synthese 47, 1–27. doi:10.1007/BF01064262

Simon, Uren, V., Li, G., Sereno, B., Mancini, C., 2007. Modeling naturalistic argumentation in research literatures: Representation and interaction design issues: Research Articles. Int. J. Intell. Syst. 22, 17–47. doi:10.1002/int.v22:1

Swanson, D.R., 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30.

van Haagen, H.H.H.B.M., 't Hoen, P., Bovo, A.B., de Morrée, A., van Mulligen, E., Chichester, C., Kors, J., den Dunnen, J., van Ommen, G.J.B., van der Maarel, S., Kern, V.M., Mons, B., Schuemie, M., 2009. Novel protein-protein interactions inferred from literature context. PLoS ONE 4. doi:10.1371/journal.pone.0007894