Project acronym:	BYTE
Project title:	Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities
Grant number:	285593
Programme:	Seventh Framework Programme for ICT
Objective:	ICT-2013.4.2 Scalable data analytics
Contract type:	Co-ordination and Support Action
Start date of project:	01 March 2014
Duration:	36 months
Website:	www.byte-project.eu

Deliverable D1.1:
Understanding and mapping big data

Author(s):	Rajendra Akerkar and Guillermo Vega-Gorgojo, *University of Oslo* Grunde Løvoll, *DNV GL AS* Stephane Grumbach and Aurelien Faravelon, *INRIA* Rachel Finn, Kush Wadhwa, and Anna Donovan, *Trilateral Research & Consulting* Lorenzo Bigagli, *National Research Council of Italy*

Dissemination level:	Public
Deliverable type:	FINAL
Version:	3.0
Submission date:	31 March 2015

Table of contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 ABSTRACT

Big data is an IT buzzword nowadays. Data is being collected at an extraordinary scale in very wide-ranging application areas. Big data analytics now steers almost every aspect of our society, including retail, manufacturing, life sciences, physical sciences, and financial services. One of the most common definitions of big data uses three terms starting with the letter V: volume, velocity and variety. Many leading big data companies seem to coalesce around this definition. But some are sceptical of that definition, too. This report examines and identifies several definitions of big data, including investigating how big data is understood in different disciplines, industries or contexts. Further, the report presents a working definition of big data.

One of the major challenges of big data is how to extract value from it. Creating value from big data is a multistep process. Most industry sectors know how to create it and store it, but they fall short when it comes to analysis and synthesis. Scientific research is also seeing dramatic impacts as data intensive fields such as bioinformatics, climate change, and physics more effectively use computation to interpret observations and datasets. Thus, this report presents a close-up view on big data, including big data applications, opportunities, and challenges in selected industry sectors.

Finally, we measure the data flows on the main intermediation platforms. These platforms are particularly relevant since they operate essentially in all countries in the world, and occupy in most countries the top position with a very large traffic. An analysis of the traffic on these platforms thus allows for a greater understanding of the global cross-country data flow.

1.2 PURPOSE AND SCOPE

The report summarises the activities performed in the context of Task 1.1 and Task 1.2 of the BYTE project. The aim of the report is to provide a useful reference for understanding big data, existing scientific and technological status and the immediate future activities, and to identify how BYTE can go beyond that point.

1.3 TARGET AUDIENCE

- Researchers in Data Science and related fields
- The deliverable is addressed to the European Commission for the purpose of mapping the current context in which big data is being utilised and the Project Consortium for the purpose of identifying the current status of big data landscape.

1.4 METHODOLOGY

This research attempts to clarify the concept of big data using an exploratory methodology and approach, given the novelty of the relatively new topic that is big data. The report is based upon a literature review of qualitative and quantitative data. The literature review prioritized the most relevant and up to date works because of the high velocity with which the field is evolving.

In addition to academic literature, the review also includes definitions and information from current industry leaders. Numerical data from secondary and tertiary sources are also used to further explore the challenges, opportunities and applications of big data. The criteria for choosing the literature and materials used were to only utilize the most relevant and up to date literature from reputable authors, industry players and publishers.

2 BIG DATA CONCEPT AND ITS ORIGIN

Data, or pieces of information, have been collected and used right through history. However, in contemporary world, advances in digital technology have considerably boosted our ability to collect, store, and analyse data. All of the data, however, are merely that—data—until they are analysed and used to inform decision-making.

The use of the term "big data" can be traced back to debates of managing large amount of datasets in both academia and industry during the 1980s. Big data arose due to the emergence of three major trends. First, it has become economical to generate a broad kind of data, due to inexpensive storage, sensors, smart devices, social software, multiplayer games, and the Internet of Things. Second, it has become inexpensive to process huge amounts of data, due to progresses in multicore CPUs, solid state storage, cloud computing, and open source software. Thirdly, not just database administrators and developers, but many more people (such as decision makers, domain scientists, application users, journalists, and ordinary consumers) have become involved in the process of generating, processing, and consuming data. This is known as a democratization of data.

As a result of these accelerating trends, there is now a widespread realization that an unprecedented volume of data can be captured, stored, and processed, and that the knowledge gleaned from such data can benefit everyone: businesses, governments, academic disciplines, engineering, communities, and individuals.

With varied data provisions, such as sensor networks, telescopes, scientific experiments, and high throughput instruments, the datasets increase at exponential rate (Szalay et al., 2006, Lynch 2008). The off-the-shelf techniques and technologies that are available to store and analyse data cannot work efficiently and adequately. The technical challenges arise from data acquisition and data curation to data analysis and data visualization.

Big data has changed the way that we adopt in doing businesses, managements and explorations. Data-intensive computing is coming into the world that aims to provide the tools that we need to handle the big data problems. Data-intensive science (Bell et al., 09) is emerging as the fourth scientific paradigm in terms of the previous three, namely empirical science, theoretical science and computational science. Long ago, researchers describing the natural phenomenon only based on human empirical evidences, so we call the science at that time as empirical science. It is also the beginning of science and classified as the first paradigm. Then, theoretical science emerged some hundreds years ago as the second paradigm. However, in terms of many complex phenomenon and problems, scientists have to turn to scientific simulations, since theoretical analysis is highly complicated and sometimes inaccessible and infeasible. Subsequently, the third science paradigm was the computational one. Simulations usually generate a large volume of data from the experimental science, at the same time; increasingly large data sets are created in various pipelines. There is no doubt that the world of science has changed just because of the increasing data-intensive applications.

In 2012, Gartner recorded the ‘‘Top 10 Strategic Technology Trends For 2013’’ (Savitz, 2012a) and ‘‘Top 10 Critical Tech Trends for the Next Five Years’’ (Savitz, 2012b), and big data is listed in the both places. Without doubt, in near future big data will transform many fields, including business, the scientific research, and public administration.

In order to discuss about various issues related to big data, it is necessary to understand different facets about big data. For the definition of the big data, there are various different explanations from 3Vs to 4Vs. Doug Laney used volume, velocity and variety, known as 3Vs (Laney, 2001), to present data related challenges. In literature, we come across definitions of big data using these 3 Vs:

- Volume – data sizes will range from terabytes to zettabytes.
- Variety – data comes in many formats from structured data, organized according to some structures like the data record, to unstructured data, like image, sounds, and videos which are much more difficult to search and analyse.
- Velocity – in several applications, like smart cities and smart planet, data continuously arrives at possible very high frequencies, resulting in continuous high-speed data streams. It is crucial that the time needed to act on such data be very small.

Occasionally, people use another V according to their special requirements. The fourth V can be value, variability, veracity, or virtual (Zikopoulos et al., 2011). In general, big data is a collection of massive data sets with an immense diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms.

In 2012, Gartner provided a more detailed definition (Laney, 2012) as:

> Big data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

Ultimately, a data set can be called big data if it is challenging to perform capture, curation, analysis and visualization on it at the existing technologies.

The above discussion underlines that volume alone is possibly the least difficult issue to address when dealing with big data. The genuine challenge arises when we have big volumes of unstructured and structured data uninterruptedly arriving from a large number of sources. Tackling this challenge require a new generation of technologies and architectures, designed to economically extract value from vary large volumes of a wide variety of data, by enabling high-velocity capture, discovery and analysis (O'Reilly, 2011).

Table 1 outlines the shifts required to move from traditional to the big data paradigm.

Table 1 Traditional paradigm to the big data paradigm

Traditional Paradigm	New Paradigm
Some of the data **E.g.** An online transaction records main data fields, a timestamp and IP address.	**All of the data** **E.g.** Clickstream and path analysis of web based traffic, all data fields, timestamps, IP address, geospatial location where appropriate, cross channel transaction monitoring from web.

Clean Data **E.g.** Data sets are typically relational, defined and delimited.	**Chaotic Data** **E.g.** Data sets are not always relational or structured.
Deterministic **E.g.** In relational databases, the data has association, correlation, and dependency following classic mathematical or statistical principles.	**Complex coupling** **E.g.** Data can be coupled, duplicative, overlapping, incomplete, have multiple meanings all of which cannot be handled by classical relational learning tools.
Examining of Data to Test Hypotheses **E.g.** Defined data structures induce the generation and testing of hypotheses against known data fields and relationships.	**Discovery of Insight** **E.g.** Undefined data structures induce exploration for the generation of insights and the discovery of relationships earlier unknown.
Lag-time Analysis of Data **E.g.** Data needs to be defined and structured prior to use, and then captured and collated. The period of extracting data will vary but often involves a delay.	**Real-time Analysis of Data** **E.g.** Data analysis takes place as the data is captured.

3 DEFINING BIG DATA

3.1 COMMON DEFINITIONS

Big data is still in its early stages, everyone is still trying to grasp its core nature and to define it scientifically and pragmatically. Nonetheless the precise meaning of the concept remains unclear and is often used synonymously with other related concepts such as Business Intelligence (BI) and data mining. Several stakeholders have created new definitions or revisions of existing definitions that best suit their interests. Nevertheless, to capture the core of big data, consistent themes can be found by examining various definitions provided by the industry gurus and related literature.

Among several definitions reported in the literature, the first formal, academic definition appears in a paper submitted in July 2000 by Francis Diebold in his work of econometrics and statistics (Diebold, 2000):

> "Big data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say, megabytes. Even data accruing at the rate of several gigabytes per day are not uncommon."

The most popular definition in recent years uses the "Three V's": volume (size of datasets and storage), velocity (speed of incoming data), and variety (data types). Reacting fast enough to deal with data velocity is a challenge for most organizations. As mentioned in Section 2, the concept was first raised by Doug Laney (2001) in his META Group research note that describes the characteristics of datasets that cannot be handled by traditional data management tools. With increasing interest and insight in the field, the "Three V's" have been expanded to "Five V's": volume, velocity, variety, veracity (integrity of data), value (usefulness of data) and complexity (degree of interconnection among data structures) (Armour, 2012). However, the soul of these V's remains within the extent of data characteristics per se.

Another noteworthy definition presented by Brian Hopkins and Boris Evelson. According to them, if we simply have high volume or velocity, then big data may not be appropriate. The definition described in "Expand your digital horizon with big data"[1]:"Big data: techniques and technologies that make handling data at extreme scale economical."

The two main characteristics are volume and velocity, while variety and variability shift the curve shown in the following Figure 1. In other words, extreme scale is more economical, and more economical means more people do it, leading to more solutions.

More comprehensive definitions and descriptions have also emerged. For example, in the report, "Demystifying big data", the big data commission at the TechAmerica Foundation offers the following definition:

[1] http://www.asterdata.com/newsletter-images/30-04-2012/resources/Forrester_Expand_Your_Digital_Horiz.pdf

> "Big data is a term that describes large volumes of high-velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" (TechAmerica, 2012).

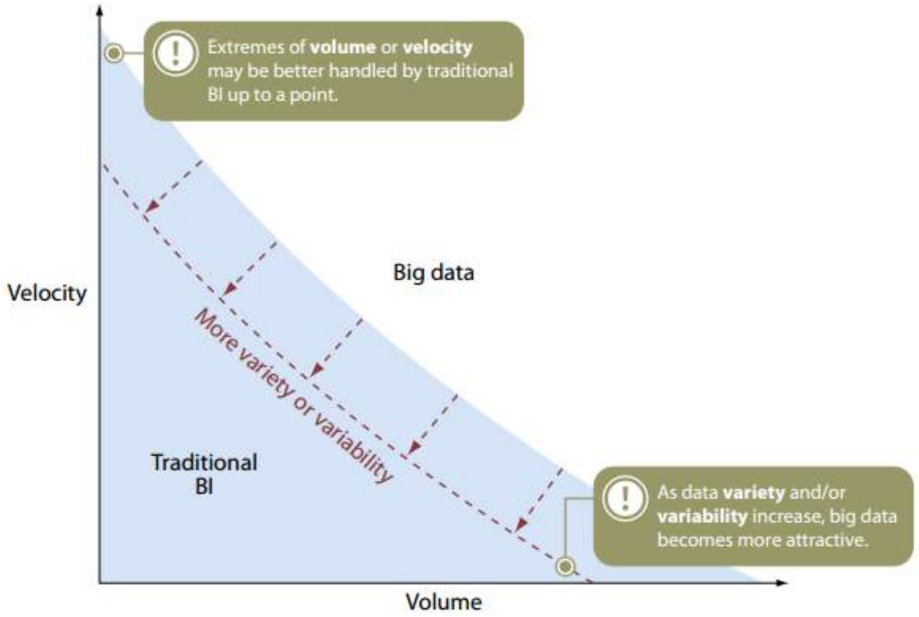

Extremes of **volume** or **velocity** may be better handled by traditional BI up to a point.
Velocity
Big data
More variety or variability
Traditional BI
As data **variety** and/or **variability** increase, big data becomes more attractive.
Volume

Figure 1 The graphics illustrating big data (Source: Expand Your Digital Horizon With Big data by Brian Hopkins and Boris Evelson[2])

Furthermore, researchers at McKinsey propose a *subjective* definition: "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse" (McKinsey, 2011).

The definition of big data can vary by sector, depending on what kinds of software tools are normally available and what sizes of datasets are common in a particular industry. Big data in several sectors at the moment range from a few dozen terabytes to multiple petabytes.

IBM[3] states that big data involves notable amounts of data that comes from a wide variety of sources. IBM highlights the increasing speed of data generation. "Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data."

Mike Gualtieri, Forrester Analyst, proposes a definition that attempts to be pragmatic and actionable for IT specialists; this definition doesn't depend on measurement of data characteristics:

> "Big data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers" (Gualtieri, 2012).

[2] www.asterdata.com/newsletter-images/30-04-2012/resources/Forrester_Expand_Your_Digital_Horiz.pdf

[3] www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Jerry Smith, Data Scientist Insights contributor, developed a mathematically sound definition of big data:

> "Big data represents the historical debri (observed data) resulting from the interaction of at between 70 and 77 independent variable/subjects, from which non-random samples of unknown populations, shifting in composition with a targeted time frame, can be taken" (Smith, 2012).

Some more definitions of big data presented in literature are as follows:
Association for Data-driven Marketing & Advertising (ADMA)[4], Sydney has provided global, commercial and marketing definitions of big data.

Global definition of big data
Big data is the collection of large volumes of varied information, used to extend our understanding of the environment, medicine, science, business and human experience.

Commercial definition of big data
Big data is the current term given to the wide use of data collected from digital, technological, and analogue sources. Big data is used to improve business understanding of markets, allowing improvements in customer experience and organisational performance.

Marketing definition of big data
Big data is the current term given to collecting, analysing and generating insights from a wide variety of customer, commercial and environmental information.

It is used to develop better understanding of customer preferences, habits and considerations in making transactions with different categories, brands and channels.

The successful use of data in marketing leads to improved customer experience, a better exchange of value between customers and organisations, and improved business performance.

SAP[5] offers a more promotion-oriented view on big data.

> "Big data is an opportunity to change how you work, play, and live by tracking new signals within digital noise. From major league sports and personalized shopping to carbon reduction and cancer treatment, organizations are using big data to re-imagine achieving what is possible."

Clearly, SAP is focusing on the benefits of big data rather than delivering a straightforward definition of the concept. Besides, SAP is underlining the fact that big data can provide value in a wide variety of fields.

SAS has added two additional dimensions to complement the original three Vs. These are variability and complexity (SAS, 2014). IBM, includes a fourth V; veracity. Furthermore, a fifth V, value is commonly associated with big data.

[4] http://www.adma.com.au/

[5] http://www.sap.com/bin/sapcom/en_us/downloadasset.2014-04-apr-24-19.sap-makes-big-data-real-real-time-real-results-pdf.bypassReg.html

Principal analyst for O'Reilly Radar, Edd Dumbill, has given alternative definition (Dumbill, 2012): “Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.”

In other words, there is no consensus on the exact characteristics of big data. Nevertheless there are multiple characteristics which most of the vendors in the field agree upon.

Despite the range and differences existing within each of the abovementioned definitions there are some points of similarity. Notably all definitions make at least one of the following assertions:

- Size: the volume of the datasets is a critical factor.
- Complexity: the structure, behaviour and permutations of the datasets are a critical factor.
- Technologies: the tools and techniques which are used to process a sizable or complex dataset is a critical factor.

Important point to be noted, while discussing of the concept of big data, is that the phrase can refer either to huge and/ distinct datasets, or to technologies processing such datasets. In literature, big data comes in two different types: static big data and real-time big data. Both types of datasets can be structured or unstructured.

3.2 Working big data definition for BYTE

From previous sub-section on big data definitions, big data can mean one of four things:

Big volumes of data and small analytics: Using traditional analytics (SQL, descriptive statistics, etc.) on datasets which are larger than possible with traditional tools.

Advanced analytics on big volumes of data: Using machine learning and other complex analytics on very large quantities of data.

Big (high) velocity: Refers to the increasing speed at which the data is created, and the increasing speed at which the data can be processed, stored and analysed. The possibilities of processing data in real-time is an area of particular interest, which allows businesses to do things like electronic trading, real-time ad placement on Web pages, real-time customer targeting, real time monitoring and optimization of assets, and mobile social networking.

Big variety: Many enterprises are faced with integrating a larger and larger number of data sources with diverse data. Most of data generated is unstructured; coming in all shapes and forms−from geo-spatial data, to tweets which can be analysed for content and sentiment, to multimedia data such as photos, audio and videos.

We can define big data as:

Big data is

using big volume, big velocity, big variety data asset to extract value (insight and knowledge),

and furthermore ensure veracity (quality and credibility) of the original data and the acquired information,

that demand cost-effective, novel forms of data and information processing for enhanced insight, decision making, and processes control.

Moreover, those demands are supported by new data models and new infrastructure services and tools which is able to procure and process data from a variety of sources and deliver data in a variety of forms to several data and information consumers and devices.

4 OPPORTUNITIES & CHALLENGES

While big data is transforming how research and business is conducted across the board. And in research it is leading to the emergence of a new paradigm of science based on data-intensive computing, in unison it poses a significant challenge for researchers and business. Big data technologies can derive value from large datasets in manner that were earlier impractical — truly, big data can generate insights which previously was not available and that researchers didn't even think to pursue[6]. Nonetheless the technical capabilities of big data have reached a level of complexity and pervasiveness that demands concern about how best to balance the opportunities offered by big data against the social and ethical questions these technologies raise.

4.1 OPPORTUNITIES

The opportunity that big data presents to all industry sectors is in the potential to unlock the value and insight contained in the data industries already hold via the transformation of information, facts, relationships and indicators. The value of big data for industries is limited by their ability to efficiently utilize big data and the ability to derive useful information from this data. With every opportunity there come barriers and sectors must overcome these to enable the benefits of big data to be realised.

Big data analysis may provide profound insights into a number of important areas including health care, medical and other sciences, transport and infrastructure, education, communication, meteorology and social sciences.

Important areas that big data may influence are described below:

Data management — there are potential savings in time and money if industries implemented smarter data management practices that were aware of the needs of big data analysis. Data sources from differing enterprises and operational areas would be of greater benefit to multiple industries and for multiple purposes if there were greater transparency. For example, through better business process management, redundant data collection processes can be reduced by reusing data collected from separate processes.

Personalisation of services — we have moved from an era of experiencing things at a macro level to experiencing things at a personal level. Big data analytics may produce value by revealing a clear picture of a customer. Big data is able to achieve this due to its characteristic granularity. This granularity may assist in unlocking the possibility of personalised services tailored to the individual and delivered by industry. The granularity in big data opens up new opportunities for personalising services. When a service provider knows something specific about a user then there is an opportunity to tailor the service offered accordingly. This will be most useful when the data in question relates to the user's needs, and when the personalisation is done in a manner that is prominent for the transaction being undertaken or service being used.

[6] http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Predictive analytics — the alliance of multiple datasets from disparate sources in combination with advanced analytics technologies will advance problem solving capabilities, and in turn will improve the ability of predictive analytics to reveal insights that can effectively support decision-making. In short, big data opens up the field of reliable predictive analytics. By assessing the relationships embedded in large datasets it is possible to construct a new generation of models explaining how things are likely to evolve in the future. This approach can be blended with scenario planning to develop a series of predictions for how a system will respond to distinct choices. The state of the art in predictive analytics can deliver forecasts for some domains with a very high degree of precision, offering an auditable, scientific basis for making decisions in complex systems.

Productivity and efficiency — the analysis of big data sources can be used to identify cost savings and opportunities to increase efficiency and reliability, which will directly contribute to an improvement in productivity. Where big data and analytics are used to detect cost savings and increase efficiency, they can contribute to a direct progress in productivity. This can in turn help to boost further innovation.

As per McKinsey (Manyika et al., 2012) report, the effectual use of big data has the primary benefits to transform economies, and delivering a new wave of productive growth. Drawing benefits of valuable knowledge beyond big data will become the key competition for enterprises and will create new rivals who are able to attract employees that have the critical skills on big data. Big data could produce $300 billion potential annual value to US health care, and €250 billion to European public administration (Manyika et al., 2012).

4.2 CHALLENGES

Big data brings many appealing opportunities and at the same time we are also facing a lot of challenges (Ahrens et al., 2011, Chen et al., 2014).

There are technical difficulties in data acquisition, storage, searching, sharing, analysis, and visualization. A key challenge exists in computer architecture for several decades, that is, CPU-heavy but I/O-poor (Hey et al., 2009). This system disproportion still curbs the progress of the discovery from big data. The CPU performance is doubling each 18 months following the Moore's Law, and the performance of disk drives is also doubling at the same rate. But, the disks' rotational speed has to some extent improved over the past decade. The consequence of this disproportion is that random I/O speeds have improved reasonably while sequential I/O speeds increase with density gradually. Information is growing at exponential rate at the same time, but the improvement of information processing methods is also rather slower. In a lot of important big data applications, the state-of-the-art techniques and technologies cannot preferably solve the real problems; this is mainly an issue for real-time analysis.

Typically, knowledge discovery life-cycle for big data is shown in Fig. 4.1, (Akerkar et al., 2007), which consists of the following phases:

1. Data Generation: Data may be generated by devices, experiments, sensors, or supercomputer simulations.
2. Data Processing and Organization: This phase involves sub-processes such as recording, cleaning, reduction, visualization, query analytics, and many other aspects

of data processing. This may also include combining data with external data or historical data, in principle creating a virtual data warehouse.

3. Data Analytics, Mining and Knowledge Discovery: Given the size and complexity of data and the need for both top-down and bottom-up discovery, scalable algorithms and software need to be utilized in this phase.
4. Actions, Feedback and Refinement: Insights and discoveries from preceding phases help close the loop by initiating new simulations, models, parameters, observations, thus, making the closed loop cycle for big data.

Challenges in big data analysis include data inconsistence and incompleteness, heterogeneity, scalability, timeliness and data security (Jagadish et al., 2014, Agrawal et al., 2011, Kouzes et al., 2009).

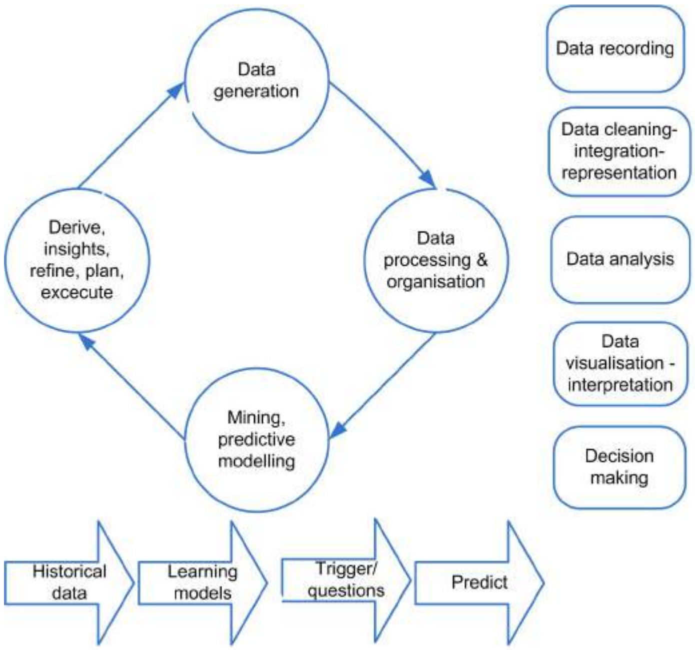

Data generation
Data processing & organisation
Mining, predictive modelling
Derive, insights, refine, plan, excecute
Data recording
Data cleaning-integration-representation
Data analysis
Data visualisation - interpretation
Decision making
Historical data
Learning models
Trigger/ questions
Predict

Figure 2 A knowledge-discovery lifecycle for big data

There are many technical challenges that must be addressed to realize the full potential of big data. (Jagadish et al., 2014) provide a comprehensive discussion of such challenges based on the notion of data analysis pipeline:

- Data Recording: it is critical to capture the context into which data has been created, to be able to filter out non relevant data and to compress data, to automatically generate metadata supporting precious data description and to track and record lineage.
- Information Extraction and Cleaning: data may have to be transformed in order to extract information from it and express this information in a form that is proper for analysis. Data may also be of poor quality and/or uncertain. Hence, data cleaning and data quality verification are critical.
- Data Integration, Aggregation and Representation: data can be very heterogeneous and may have different metadata. Data integration, even in more conventional cases, requires huge human efforts. Novel approaches that can improve the automation of data integration are critical as manual approaches will not scale to what is required for big data. Also different data aggregation and representation strategies may be needed for different data analysis tasks.

- Query Processing, and Analysis: methods appropriate for big data need to be able to deal with noisy, dynamic, heterogeneous, unreliable data and data characterized by complex relations. Though regardless of these difficulties, big data even if noisy and uncertain can be more precious for detecting more reliable hidden patterns and knowledge compared to tiny samples of good data. Also the relationships existing among data can represent an opportunity for cross-checking data and thus improve data trustworthiness. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures.
- Interpretation: analysis results extracted from big data needs to be interpreted by decision makers and this may require the users to be able to analyse the assumptions at each phase of data processing and perhaps re-tracing the analysis.

As the prior step to data analysis, data must be well-constructed. However, considering variety of data sets in big data problems, it is still a big challenge for us to purpose efficient representation, access, and analysis of unstructured or semi-structured data in the further researches. How can the data be pre-processed in order to improve the quality data and the analysis results before we begin data analysis? As the sizes of data set are often massive, occasionally gigabytes or more, and their origin from heterogeneous sources, current real-world databases are severely susceptible to inconsistent, incomplete, and noisy data. Hence, a number of data pre-processing methods, including data cleaning, data integration, data transformation and date reduction, can be applied to remove noise and rectify inconsistencies (Akerkar et al., 2007).

Different challenges arise in each sub-process when it comes to data-driven applications. In the following subsection, we give a brief overview of challenges facing for each sub-process.

4.2.1 Data acquisition and storage

Data sets grow in size because they are increasingly being collected by ubiquitous information-sensing mobile devices, sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on. There are 2:5 quintillion bytes of data created every day, and this number keeps rising exponentially (Hilbert et al., 2011). The global technological capacity to store information has somewhat doubled about every 3 years since the 1980s. In several disciplines, such as financial and medical, data is often deleted just because there is no enough space to store these data. These valuable data are created and captured at high cost, but ignored conclusively. The bulk storage requirements for experimental data bases, array storage for large-scale scientific computations, and large output files are reviewed in (Worlton, 2071).

Big data has changed the way we capture and store data (Oliveira et al., 2012), including data storage device, data storage architecture, data access mechanism. The accessibility of big data is on the top priority of the knowledge discovery process.

The existing storage technologies (Agrawal et al., 2011, Pirovano et al., 2003) cannot possess the same high performance for both the sequential and random I/O simultaneously, which requires us to rethink how to design storage subsystems for big data processing systems.

Direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN) are the enterprise storage architectures that were commonly used (Leong, 2009). However, all these existing storage architectures have serious shortcomings when it comes to large-scale distributed systems. Determined concurrency and per server throughput are the basic requirements for the applications on highly scalable computing clusters, and current storage systems lack both. Optimizing data access is a standard way to improve the performance of data-intensive computing (Ishii et al., 2009-11-12), these techniques include data replication, migration, distribution, and access parallelism. In (Bencivenni et al., 2008), the performance, reliability and scalability in data-access platforms were discussed. Data storage and search schemes also lead to high overhead and latency (Shen et al., 2011), distributed data-centric storage is a good approach in large-scale wireless sensor networks (WSNs). (Shen et al., 2011) proposed a distributed spatial–temporal similarity data storage scheme to offer efficient spatial–temporal and similarity data searching service in WSNs.

4.2.2 Data communication

Cloud computing has become mainstream. Enterprises have a multitude of cloud providers to choose from. Cloud computing has a wide variety of forms, including IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service). Moreover, the distinctions among IaaS, PaaS, and SaaS have started to blur. For example, IaaS providers nowadays provide manageability features that begin to resemble PaaS. From a data platform perspective, the ideal goal is to provide PaaS in its truest form. In a world with true PaaS for data, users would be able to upload data to the cloud, query it exactly as they do today over their SQL databases on the intranet, and selectively share the data and results easily, all without worrying about how many instances to rent, what operating system to run on, how to partition the databases across servers, or how to tune them. Although the emergence of services such as Database.com, Google Big Query, Amazon Redshift, and Microsoft Azure SQL Database, we are yet far away from that vision.

Here are some of the critical challenges in realizing the vision of Data Platform as a Service in the cloud. We know that the network bandwidth capacity is the bottleneck in cloud and distributed systems, especially when the volume of communication is large. On the other hand, cloud storage also leads to data security problems (Wang et al., 2011) as the requirements of data integrity checking. There are several schemes suggested under different systems and security models (Wang et al., 2009),(Oprea et al., 2005).

Data replication is another challenge. Although data replication has been studied extensively in the past, it is important to revisit it in the context of the cloud, keeping in mind the need for high availability, load balancing, and cost. Both elasticity and replication need to be considered not just within, but also across, geographically distributed data centres.

System administration and tuning is next challenge. A data platform in use as a cloud service will need extreme auto-tuning. In Data Platform as a Service, the traditional roles of database and system administrators do not exist. Therefore, all administrative tasks such as capacity planning, resource provisioning, physical data management, and admission control policy setting need to be automated while dealing with the variance that arises due to the elasticity of resources and their availability in the cloud setting.

Data sharing is another key challenge, as the cloud enables it at an unprecedented scale. The database community should seek to develop novel services that harness this potential. We have already seen services that enable collaborative productivity tools as well as the ability to share results of data analysis or visualization. There is an ample opportunity to explore deeper ideas in the context of data analytics.

Further, we must realise how we can support important services such as data curation and provenance when we want to perform such activities collaboratively in the cloud. Data sharing in the cloud will also raise new issues in leveraging data sets, such as how to find valuable public data, how to correlate your own data with public data to add context, how to find high-quality data in the cloud, and how to share data at fine-grained levels, as well as business issues, such as how to distribute costs when sharing computing and data and how to price data.

In the case of cyber-physical systems, as in the Internet of Things, where, e.g., cars will upload data into a cloud and obtain control information in return. Cyber-physical systems involve data streaming from multiple sensors and mobile devices, and must cope with intermittent connectivity and limited battery life, which pose difficult challenges for real-time and perhaps mission-critical data management in the cloud.

4.2.3 Data management and curation

In the perspective of data management, a number of aspects characterize big data, among them: the maximum size of the database, the data models, the capability of setting up distributed and clustered data management solutions, the sustainable rate for the data flow, the capability of partitioning the data storage to make it more robust and increase performance, the query model adopted, the structure of the database (relational, RDF (Resource Description Framework), reticular, etc.), etc. Considering data structures for big data there is a trend to find a solution using the so called NoSQL databases ("Not only SQL", Simple Query Language), even if there are good solutions that still use relational database (Dykstra, 2012). In the market and from open source solutions, there are several different types of NoSQL databases and rational reasons to use them in different situations, for different kinds of data. There are many methods and techniques for dealing with big data, and in order to be capable to identify the best choice in each case, a number of aspects have to be taken into account in terms of architecture and hardware solutions, because different choices can also greatly affect the performance of the overall system to be built. Related to the database performance and data size, there is the so called CAP Theorem that plays a relevant role (Brewer, 2001), (Brewer, 2012). The CAP theorem states that any distributed storage system for sharing data can provide only two of the three main features: *consistency*, *availability*, and *partition tolerance* (Fox and Brewer, 2099). Property of consistency states that a data model after an operation is still in a consistent state providing the same data to all its clients. The property of availability means that the solution is robust with respect to some internal failure, that is, the service is still available. Partition tolerance means that the system is going to continue to provide service even when it is divided in disconnected subsets, for example a part of the storage cannot be reached. To cope with CAP theorem, big data solutions try to find a trade-off between continuing to issue the service despite of problems of partitioning and at the same time attempting to reduce the inconsistencies, thus supporting the so called eventual consistency.

Furthermore in the context of relational database, the ACID (Atomicity, Consistency, Isolation and Durability) properties describe the reliability of database transactions. This paradigm does not apply to NoSQL database where, in contrast to ACID definition, the data state provides the so-called BASE property: Basic Available, Soft state and Eventual consistent. Therefore, it is typically hard to guaranteed an architecture for big data management in a fault-tolerant BASE way, since, as the Brewer's CAP theorem says, there is no other choice to take a compromise if you want to scale up.

4.2.4 Data analysis

Data analytics algorithms may range on data: ingestion, crawling, verification, validation, mining, processing, transcoding, rendering, distribution, compression, etc., and also for the estimation of relevant results such as the detection of unexpected correlations, detection of patterns and trends (for example of events), estimation of collective intelligence, estimation of the inception of new trends, prediction of new events and trends, analysis of the crowd sourcing data for sentiment/affective computing with respects to market products or personalities, identification of people and folk trajectories, estimation of similarities for producing suggestion and recommendations, etc. In the last few years, researchers made efforts to accelerate analysis algorithms to cope with increasing volumes of data and speed up processors following the Moore's Law. It is necessary to develop sampling, on-line, and multi-resolution analysis methods. Some researchers devote into this area (Tang et al., 2009, Hsiao et al., 2008, Hassan et al., 2087). As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift (Agrawal et al., 2011) in processor technology. Alternatively, processors are being embedded with increasing numbers of cores. This shift in processors leads to the development of parallel computing (Oehmen et al., 2006, Simeonidou et al., 2005, Garcia et al., 2011).

For those real-time big data applications, like navigation, social networks, finance, biomedicine, astronomy, intelligent transport systems, and Internet of thing, timeliness is at the top priority. It is still a big challenge for stream processing involved by big data.

It is right to say that big data not only have produced many challenge and changed the directions of the development of the hardware, but also in software architectures. One shift that is underway is the move towards cloud computing (Furht, 2011, Vouk, 2008, Adamov, 2012, Foster et al., 2008), which aggregates multiple disparate workloads with varying performance objectives (e.g. interactive services require that the data processing engine return back an answer within a fixed response time cap) into a large cluster of processors.

Data privacy and security emerges with great interests. Noteworthy security problems include data security protection, intellectual property protection, personal privacy protection, commercial secrets and financial information protection (Smith et al., 2012). Most developed and developing countries have already formulated related data protection laws to enhance the security.

4.2.5 Data visualization

One of the most valuable means through which to make sense of big data, and thus make it more approachable to most people is through data visualization. Data visualization is path, both literally, like the street signs that direct you to a road, and metaphorically, where colours, size, or position of abstract elements convey information. Data visualization can be categorized into two applications:

- *Exploration*: In the exploration phase, the data analyst will use several graphics that are mostly unsuitable for presentation purposes yet may reveal very interesting features. The amount of interaction needed during exploration is high. Plots must be created fast and modifications like sorting or rescaling should happen promptly so as not to interrupt the line of thought of the analyst.
- *Presentation*: Once the key findings in a data set have been explored, these findings must be presented to a broader audience interested in the data set. These graphics

often cannot be interactive but must be appropriate for printed reproduction. Besides, some of the graphics for high-dimensional data are all but trivial to read without prior training (say, in statistics), and thus probably not well suited for presentation purposes.

Obviously, the amount of interactivity used is the major dimension to discriminate between exploratory graphics and presentation graphics. The visual (graphics), when properly aligned, can offer a shorter path to help decision making and become a tool to convey information vital in all data analysis.

Having the ability to analyse big data is of restricted value if users cannot grasp the analysis[7]. Eventually, a decision-maker, provided with the result of analysis, has to interpret these results. It comprises assessing all the assumptions made and retracing the analysis. There are several probable sources of error: computer systems can have bugs, models mostly have assumptions, and results can be based on huge data. For all of these reasons, no responsible user will cede authority to the computer system. Rather he will try to understand, and verify, the results produced by the computer. The computer system must make it easy for him to do so. This is particularly a challenge with big data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in.

The purpose of data visualization (Simoff et al., 2008, Keim et al., 2004) is to represent knowledge more intuitively and effectively by using different graphics. To convey information easily by providing knowledge hidden in the complex and large-scale data sets, both visual form and functionality are necessary. Information that has been abstracted in some schematic forms, including attributes or variables for the units of information, is also valuable for data analysis.

Online marketplace eBay, have hundreds of million active users and billions of goods sold each month, and they generate a lot of data. To make all that data understandable, eBay turned to big data visualization tool: Tableau[8], which has capability to transform large, complex data sets into intuitive pictures. The results are also interactive.

Based on them, eBay employees can visualize search relevance and quality to monitor the latest customer feedback and conduct sentiment analysis.

For big data applications, it is difficult to conduct data visualization because of the large size and high dimension of big data. However, current big data visualization tools have poor performances in functionalities, scalability and response time. For instance, the history mechanisms for information visualization (Heer et al., 2008) are data-intensive and need more efficient tactics.

Uncertainty is a multi-faceted characterization about data, whether from measurements and observations of some phenomenon, and predictions made from them (Wu et al., 2012). It may include several concepts including error, accuracy, precision, validity, quality, variability, noise, completeness, confidence, and reliability. Uncertainty arises in any phases of a visual analytics process (Wu et al., 2012). Uncertainty is a key issue with geospatial data sets. The difficulty of visualizing data with uncertainty increases with the richness in which uncertainty

[7] http://wp.sigmod.org/?author=8

[8] http://www.tableausoftware.com/solutions/big-data-analysis

is represented (from scalars to distributions) and the dimensionality of the data. As more information needs to be displayed, it is natural to turn to techniques from multivariate and statistical visualization techniques such as Chernoff faces, scatter plots, star plots, box plots, etc. New framework for modelling uncertainty and characterizing the evolution of the uncertainty information are highly necessary through analytical processes.

Eventually, visualization and interactive exploration of large amounts of data has been challenging. In order to guarantee scalability, acceptable reaction times and support for multiple devices, developers generally want to think about efficient processing, rendering and brushing such data.

4.2.6 Human role in life cycle of big data

There has been a growing recognition of the increasing role of people in the data life cycle, of course, such as the work done in our community and elsewhere on crowdsourcing. However, the new need to “manage the people" is not just about crowdsourcing or micro-tasks. Today's setting requires the consideration of people (company culture and human factors) as they relate to: query understanding and refinement, identifying relevant and trustworthy information sources, defining and incrementally refining the data processing pipeline, and visualizing relevant patterns and obtaining query answers, all in addition to making the various micro-tasks doable by domain experts and end users. We can classify people's roles into four general categories: producers of data, curators of data, consumers of data, and community members.

Many people today are data producers, as virtually anyone can generate a flood of data: the use of mobile phones, social platforms and applications (e.g., Facebook, Twitter), and an increasing collection of wearable devices (e.g., Fitbit). One key challenge is to develop algorithms and incentives that guide people to produce and share the most useful data, while maintaining the desired level of data privacy.

Culture and organizations are also important factors. In order to succeed with big data there is a need for the right culture in addition to data, resources and competence. That means, cultural (as in company culture) aspects are just as crucial as competence and technical capabilities in big data success. The cultural aspects include: a culture which values facts and seeks facts from data, and a culture where sharing of data and idea is encourages (silo thinking is hindering getting value from data).

In today's data-driven world there is less central control over data. Data is no longer just in databases controlled by a DBA and curated by the IT department. Instead, as mentioned earlier, a wide variety of data is now being generated and a wide variety of people are now empowered to curate it. In particular, crowdsourcing has emerged as a promising curation solution[9,10]. Another key challenge, then, is to obtain high-quality data sets from a process based on often-imperfect human curators. Two related challenges are building platforms that allow people to curate data easily and extending relevant applications to incorporate such curation. For these people-centric challenges, data lineage and explanation will be crucial, as well considerations of privacy, security and legal aspects related to data ownership and

[9] https://cs.uwaterloo.ca/~ilyas/papers/StonebrakerCIDR2013.pdf

[10] http://ceur-ws.org/Vol-782/SimperlEtAl_COLD2011.pdf

ownership of insight derived from data. These aspects will be identified and examined in WP2 (Elements of societal impact).

People are data consumers as well. In the enterprise, data consumers have usually been people who know how to ask SQL queries, via a command-line interface or a graphical query tool, over a structured database. Today's data consumers may not know how to formulate a query at all. Here major challenge is to make it possible for such people to get their answers themselves, directly. This requires new query interfaces, e.g., interfaces based on multi-touch, natural language queries, and not just console-based SQL interfaces.

Numerous communities exist online, with more being created daily. Members of such communities often want to create, share, and manage data, and it is becoming increasingly easy for them to do so. In particular, members may want to collaboratively build community-specific knowledge bases, wikis, and tools to process data. For example, many researchers have created their own pages on Google Scholar, thereby contributing to this “community" knowledge base. Now challenge is to build tools to help communities produce usable data as well as to exploit, share, and mine it.

The big data researcher is a multi-skilled person that understands the domain of IT and business. Also, she has the right creativity to develop hard, technical solutions that indeed help a data-driven, information-centric organisation. Though, the fusion of talent at the crossroads of business, information technology, data sciences and operations can be challenging to identify and develop. For many enterprises it is a real challenge to find the right big data talent. The shortage of talent will be a significant constraint to capture values from big data (Manyika et al., 2012). This type of human resource is more difficult to educate. It usually takes many years to train big data analysts that must have inherent mathematical abilities and related professional (domain) knowledge. We believe that the same situation also happened in other nations, not matter developed or developing countries around the world. It is likely that there will be a hot competition for human resources in big data developments.

Finally, we provide some of the most common big data roles or job titles with the skills required.

- **Data scientists:** This title is similar to what a 2011 McKinsey report[11] calls "deep analytical talent". These people have backgrounds in mathematics or statistics. Some have experience or academic degrees in artificial intelligence, natural language processing or data management.
- **Data architects:** Programmers who are competent at working with complex data, disparate types of data, undefined data and lots of ambiguity. They may be people with usual programming or business intelligence backgrounds, with some background in statistics. They require the creativity and persistence to be able to harness data in new ways to create new insights.
- **Data visualizers:** Technologists who translate analytics into information a business can use. They harness the data and put it in context investigating what the data means and how it will impact the business.
- **Data change agents:** People who drive changes in internal operations and processes based on data analytics. They may have a Six Sigma background and proper communication skills to interpret jargon into terms others can understand.

[11] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

- **Data engineers/operators:** People who are the designers and managers of the big data infrastructure. They develop the architecture that helps analyse and process data in the way the business needs it.

To sum up this section, we present a table containing various challenges of big data in seven industry sectors. Exemplifying challenges among seven sectors, if we compare the historical productivity of sectors in Europe with the potential of these sectors to acquire value from big data, we see that patterns change from sectors to sectors. While these sectors will have to overcome challenges to capture value from the use of big data, challenges are structurally higher for some than for others. For instance, capturing value in healthcare faces challenges given the relatively low IT investment performed thus far. In the following table, we have used 'IT investment/tools' as the bandwidth.

	Volume of data	Velocity of data	Variety of data	Business models	Privacy/data protection	Lack of skills	IT investment/tools
Oil & gas	🟥	🟥	🟧	🟧	🟦	🟦	🟨
Healthcare	🟥	🟧	🟥	🟧	🟥	🟨	🟦
Environment	🟥	🟥	🟥	🟨	🟨	🟧	🟧
Crisis informatics	🟧	🟧	🟧	🟨	🟧	🟦	🟥
Smart cities	🟧	🟧	🟨	🟨	🟧	🟨	🟥
Shipping/Transport	🟦	🟦	🟨	🟨	🟨	🟧	🟥
Culture	🟨	🟨	🟨	🟥	🟦	🟨	🟨

Very high	🟥
High	🟧
Moderate	🟨
Less	🟦

5 BIG DATA APPLICATIONS

Nowadays several disciplines contain big data problems, varying from economy to administration and from scientific explorations to national security. In 2012, McKinsey institute report (Manyika et al., 2012) states transformative potentials of big data in five domains: health care of the United States, public sector administration of European Union, retail of the United States, global manufacturing and personal location data.

In the following subsections, we will briefly introduce some applications of and problems related to big data in business, science and public administration.

5.1 Big Data in business

Big data is transforming the business landscape, as companies tap into more and more broad varieties of structured and unstructured data with greater speed and complexity. Business has always wanted to derive insights from information in order to make better, smarter, real time, fact-based decisions: it is this demand for depth of knowledge that has stimulated the growth of big data tools and platforms.

The leading enterprises are now including big data from both within and outside the enterprise, incorporating structured and unstructured data, machine data, and online and mobile data to supplement their organizational data and offer the basis for historical and forward-looking views.

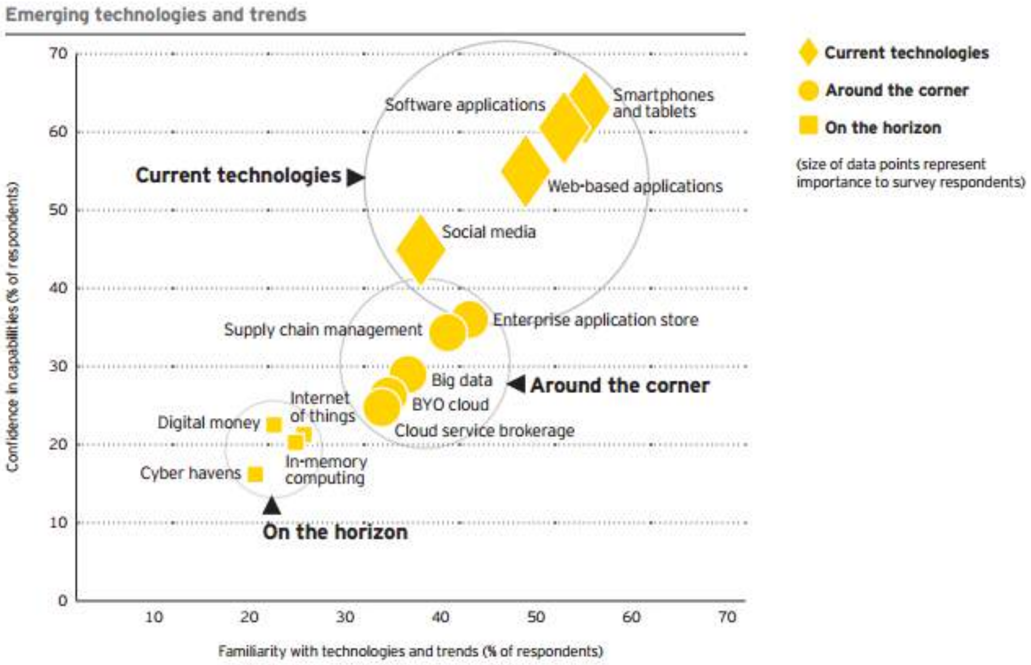


Figure 3 EY's 2013 Global Information Security Survey[12]

Companies that invest in and effectively derive value from their data will have a clear advantage over their competitors — a performance gap that will continue to grow as more relevant data is produced, emerging technologies and digital channels offer better acquisition

[12] http://www.ey.com/Publication/vwLUAssets/EY_-_2013_Global_Information_Security_Survey/$FILE/EY-GISS-Under-cyber-attack.pdf

and delivery mechanisms, and the technologies that enable faster, easier data analysis continue to develop.

Emergent technology has extended the potential of using data-driven results into every facet of an enterprise. But, though advances in software and hardware technology have empowered the big data era, enterprises need to take a complete view that recognizes that success is built upon the integration of people, process, technology and data.

EY's 2013 Global Information Security Survey results, shown in Figure 3, indicate that while adoption and use of big data is not yet widespread, there is growing confidence and familiarity with the technology.

Mostly business environment changes frequently and rapidly. Thus, in such environment, future prediction becomes more vital than the modest visualization of historical or contemporary perspectives. For effective future prediction, data analysis using machine learning and predictive modelling techniques may be applied to enhance and support the organization's business strategy. The gathering and aggregation of big data, and other information from outside the enterprise, enables the business to develop their own analytic capacity and capability, which for many years has only been available to a few larger organizations.

For the last few years we have seen that increasing data use are changing and transforming business all across the board. Businesses are increasingly adopting a data-driven attitude to answering business critical questions, increase efficiency and improve performance, conduct more targeted services, and reduce risks. Across the board programs to collect and store data from across the business and from sources outside the business are executed and analytics are done on the data to get new business insight and strengthen decision making. Some typical examples of big data use in traditional businesses are: fully automated credit scoring, churn avoidance in telecom, 360 degree customer view (all relevant customer data made available for support and sale), optimization of asset use, stock, and value chain, fraud detection in banking and credit sector, and utilization of sensor data for performance monitoring and condition based maintenance of equipment and systems.

It is not just traditional businesses which are harnessing the value of data. Over the last two decades we have seen the rise of a new industry, whose main asset is data. Armed with vast amounts of data and affordable capabilities for storage and processing, companies have been able to create new business and revenue streams from selling data or delivering services that are based on data analytics. The most prominent examples of new use of data occur in the new Internet-based industries. Here, companies exist that have created big business based entirely on access to data and their ability to use data. In addition to innovations in the business domain these companies have also moved "big data technology" forward by inventing and implemented the technology needed to solve the problems related to data handling and analyses in their own operations. Prime examples of such companies are Google, Facebook, Yahoo and Twitter.

As mentioned earlier companies in sectors such as telecom, retail, media, healthcare, insurance, finance and transport/logistics are being transformed by utilizing large data sets about their customers and internal work processes. Insights from these data streams are used for better assessment of utilization, to improve operational efficiency, and to increase sales and the efficacy of marketing.

For instance, available financial data sources include stock prices, currency and derivative trades, transaction records, high-frequency trades, unstructured news and texts, consumers' confidence and business sentiments hidden in social media and internet. Analysing such huge datasets helps measuring risks. It requires professionals who are familiar with sophisticated tools and techniques in portfolio management, securities regulation, proprietary trading, financial consulting, and risk management.

Big data capabilities enable companies to streamline their supply chain and improve overall efficiency and profitability. The optimization may include predictions using weather data, tracking data, traffic data, and other data that may influence the movement of cargo and parcels. An example of this is FedEx which is using data-based value chain optimizations for all its processes. At the core of the system lies the package tracking system that tracks and monitors every phase of the delivery cycle. Data from this system enables the company to maximize utilization of assets and personnel, while delivering on time. FedEx also use this to create value added services for customer, like online direct access to information about on-going deliveries. The FedEx system is integrated into a central system that collects and coordinates data from: airlines, connection hubs, positions of individual vehicles, weather forecasts, and real-time traffic information. This allows for real-time routing information to be pushed out to individual drivers and optimizations of pickup/delivery and asset utilization[13].

Study shows that the volume of business data worldwide, across all business sectors, doubles every 1.2 years (Manyika et al., 2012). As an example from the retail industry[14], there are around 267 million transactions per day in Wal-Mart's 6000 stores worldwide. For pursuing better competitiveness in retail, Wal-Mart teamed up with Hewlett Packard to establish a data warehouse which has a capability to store 4 petabytes of data, i.e., 4000 trillion bytes, tracing every purchase record from their point-of-sale terminals. By utilizing mature machine learning techniques to exploit the knowledge hidden in this huge volume of data, they effectively improve efficiency of their pricing strategies and advertising campaigns.

Banking and finance companies also have a lot of data available, in this industry taking the right decision fast enough is an additional challenge to the sheer volume of data. An example from this domain is FICO's falcon credit card fraud detection system[15] which monitor and manages transactions from over 2.1 billion valid accounts around the world.

By nature telecom generates and collect massive amount of behavioural data about their customers: who and how long people call, who send direct messages to whom, and where and when these activities take place. Traditionally these data have only been used for billing and more technical purposes like planning and debugging network infrastructure. In telecom loss of customers (often called churn) is also a big problem, and to counter this some players try to utilize and take targeted action based on their customer data in combination with external data sources (like social media data). By utilizing big data technology T-Mobile USA was able to utilize data from and about their 33 million customers to reduce churn rates by 50% in

[13] http://www.fedex.com/ma/about/overview/innovation.html
[14] http://istcolloq.gsfc.nasa.gov/spring2009/presentations/lopez.pdf
[15] http://www.fico.com/en/products/fico-falcon-fraud-manager/

just one quarter[16]. This was done by identifying key customers; customers with large social networks, and to target these individuals with tailor made offers to retain these influential high value customers.

When it comes to using and treating data as an asset and the foundation for business we see new, Internet based business innovation and paving the way. Google's main business model is based on its ability to use search phrases, and gathered personal profiles to connect advertisers effectively with potential customers. Twitter mines the content of Tweets to increase the value of advertising for their customers by enabling them to target relevant customer groups and measure the impact of their marketing campaigns. Twitter also sells raw data to insight aggregators and academia which is used to create new services and to research anything from epidemics to human behaviour[17].

Another interesting example is Cardlytics[18]. Cardlytics is an intermediary selling shopping habit and spending behaviour data from US banks and credit card providers to retailers. This data is then used to create targeted offerings to potential customers and the service is paid for by a commission for each sale.

As the examples above illustrates; while big data affect all industries[19], it does however have different kinds of impact and manifestations in different industries. And depending on the underlying business model[20] big data will have different impact on mode of operation and the opportunities and market positions one can take.

For traditional producers of goods the main use of data will be to facilitate improvement in the quality of the end product and to make the overall process more cost-effective. Typically this results in an increasingly automated production process. The benefits of process optimization has of course has been recognised for industrialized processes for a long time, but access to new technology, the availability of data have made it possible to take this to take this to new levels of sophistication.

In contrast, businesses in the knowledge industry (consulting would be a prime example of this) the main product is helping customers in solving their problems, and the key assets are the knowledge and capabilities of its workers. The main changes, challenges and opportunities posed by big data in this industry are related to: changes in the competitive landscape, data handling capabilities expected by the customers, and go to market strategies for data driven services. New technical capabilities have the potential to increase the efficiency of project delivery, to facilitate reuse of data, and enable scaling, through automation of the delivered services. At the same time; the possibility of encoding knowledge into models and use these models in combination with data can also enable competition from new players and thus completely alter the competitive landscape.

A third kind of businesses are network service providers, where a traditional example is telecom. These businesses provides a network where the customer can conduct their "own

[16] http://www.bigdata-startups.com/BigData-startup/t-mobile-usa-cuts-downs-churn-rate-with-big-data/

[17] http://dssg.io/2013/12/13/qcri-twitter-relief.html

[18] http://cardlytics.com/

[19] "Industry Analytic Services", Kurt Schlegel, Gartner research, July 2009, ID Number: G00167258

[20] "Configuring value for competitive advantage: On chains, shops, and networks",
Charles B. Stabell and Øystein D. Fjeldstad, Strategic Management Journal, Vol. 19, 413–437 (1998)
Casting off the chains", Øystein D. Fjeldstad and Espen Andersen, EBF issue 14, summer 2003

business". And the value of the network increases with size and the capabilities of the network. IT technology and the Internet have enabled the creation of many new such companies, and for some of them, like Facebook, Twitter, and Google, the product they sell is the data generated by their users and the insight (mainly used for targeted marketing) derived from these enormous amount of data.

When discussing big data impact on business it also make sense to talk about the "data driven" economy, and as indicated in the discussion above, there exist many different roles, market positions and business models within this economy.

5.1.1 Changing industries with big data

Recent big data evolution is transforming how industry sectors operate; we'll look closer at 3 examples: healthcare, finance, and life science.

Healthcare[21] industry presents immense opportunity for big data innovation. Physicians, patients, and other stakeholders have access to an astounding volume of rich clinical data that if combined and accurately mined, could help achieve healthcare's objectives such as providing better care experiences for patients; improving the health of the population served; and lowering per capita costs.

The healthcare industry must develop sophisticated strategies for managing emergent data sets. No longer is clinical data only accessed via electronic health records. Wearable medical devices like pacemakers or Fitbits are the fast growing sources of healthcare data. This information can be very valuable if gathered and harnessed effectively.

In financial industry, professionals had to spend excessive time and resources organizing disparate data into structured data for analysis. Analytics excellence is core to innovation across the financial industry. Business executives in the financial industry must view analytics and the ability to efficiently and effectively exploit big data, advanced modelling, and real-time decision making across channels and operations as a key capability that will eventually distinguish those that prosper in uncertain and uneven markets from those that mess up.

In comparison with other sectors, life sciences have been slower to adopt big data technology. Similar to healthcare, life sciences face exponentially growing sources of data. The industry's prime source of data is randomized controlled clinical trials, but electronic health records, electronic lab results, and even cell phones generate data leveraged by life sciences professionals. The diversity of disparate data sets makes life sciences data particularly chaotic. While other industries have been able to leverage existing big data tools and advance to the stage where predictive analytics delivers valuable results, life sciences lag behind.

5.1.2 Internet of things and big data

The advent of more use of sensor data and automated monitoring is the sign of the next wave of changes in business: "The Internet of things" (IoT). A future where an products and appliances not only are able to sense and act upon the environment but also exchange and act upon information and data shared by other devices (i.e. intelligent machine to machine

[21] http://c.ymcdn.com/sites/masstlc.site-ym.com/resource/resmgr/Files/Healthcare_MTLC-2014.pdf

communication). These technologies will certainly take automation to the next level and it will be crucial for future "smart grid solutions" in the power sector, and they will also generate large amounts of machine-to-machine interaction data which will be available across traditional barriers. The IoT trend will therefore open up new opportunities for advanced analytical offerings and services and this is the main rationale behind the large investments in IoT and analytics by companies like General Electric (GE) and Phillips. Another trend which plays into both commercial use of big data and IoT is the advent of a new generation of cognitive computing, like IBM's Watson technology, which is able to sift through, combine and infer knowledge form large collections of data sources[22]. (ref. to McKinsey report on disruptive technologies).

Internet of Things has replaced big data to be the most hyped technology as Gartner's Hype Cycle[23] for Emerging Technologies said. Things here can refer to uniquely identifiable embedded computing devices such as heart monitoring implants, smart thermostat systems, etc. Big data and Internet of Things together will help build valuable systems. The value of IoT will be in gaining timely, valuable insights from all of the data being generated by sensors, etc. The advent of the IoT does also drive development of big data technology, as an example of this ParStream's analytics database is a great fit for the speed and scale of IoT and is exactly the type of technology that can be the difference between big data and smart data driving business value.

Recently, ParStream[24] introduced the industry's first analytics platform designed to handle the massive volumes and high velocity of Internet of Things (IoT) data. The platform will help companies generate timely, actionable insights from IoT data by providing more innovative and efficient ways to analyse that data faster. Real time big data analytics is very crucial for certain fields, like stock data, sensors data, etc. Google also launched BigQuery for real-time big data this year.

5.1.3 The industrial Iinternet of things

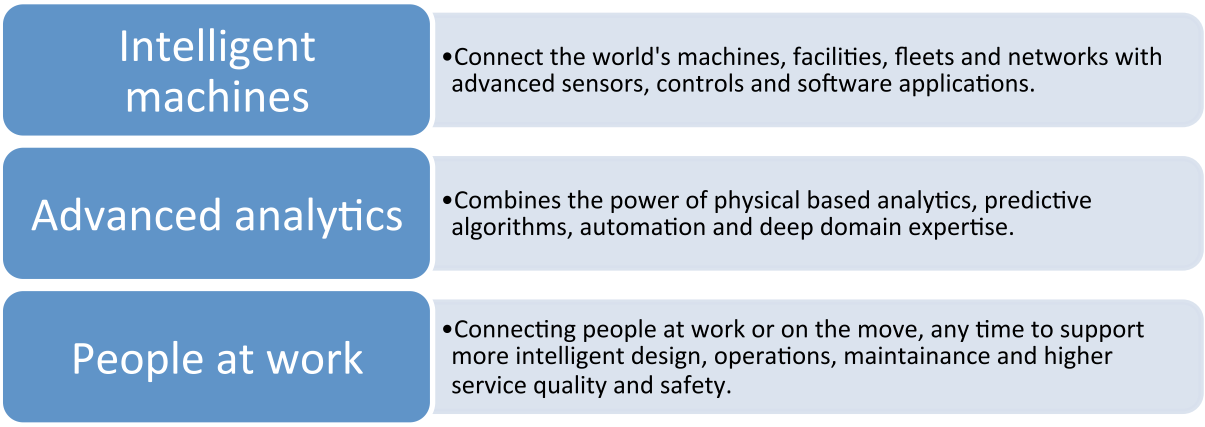
Mostly the discussion around big data has focused on clickstream data, sentiment analysis and consumer targeting. But behind the scenes, the capabilities enabled by machine-to-machine communication and advanced analytics stand poised to dramatically change the world around us. The Industrial Internet of Things (IIoT) is connecting the physical world of sensors, devices and machines with the Internet and, by applying deep analytics through software, is turning massive data into powerful new insight and intelligence. The Industrial Internet involves putting different kinds of sensors, sometimes by the thousands, in machines and the places they work, then remotely monitoring performance to maximize profitability. G.E.[25], one of the world's biggest makers of equipment for power generation, aviation, health care, and oil and gas extraction, has been one of its biggest promoters.

[22] "Disruptive technologies: Advances that will transform life, business, and the global economy", McKinsey white paper on disruptive technologies, http://www.mckinsey.com/insights/business_technology/disruptive_technologies

[23] http://www.gartner.com/newsroom/id/2819918

[24] https://www.parstream.com/

[25] http://www.ge-ip.com/industrial-big-data



Intelligent machines

- Connect the world's machines, facilities, fleets and networks with advanced sensors, controls and software applications.

Advanced analytics

- Combines the power of physical based analytics, predictive algorithms, automation and deep domain expertise.

People at work

- Connecting people at work or on the move, any time to support more intelligent design, operations, maintainance and higher service quality and safety.

Figure 4 General Electric's vision of industrial internet (source: http://www.ge.com/stories/industrial-internet)

The development in IIoT will allow us to connect intelligent machines, advanced analytics and clever people working in smarter ways to achieve extraordinary things. And the fuel driving this revolution will be big data – masses of information aggregated, analysed and put to work across sectors as diverse as energy, transport, aviation, healthcare, consumer goods, retail and even other professional services. The industrial internet promises great benefits. However it raises some sensitive legal and moral issues, which will become increasingly prominent as the use of data and analytics becomes more sophisticated.

5.1.4 Examples of typical big data use cases

Value chain optimization

Big data solutions help retailers optimize supply chains to reduce cost, improve service, and gain vital insight. Big data analytics are used to predict inventory positions in stores and distribution channels. This is achieved by utilizing demand plans and forecasts, sales history, external predictors of future performance such as category trends, weather patterns, local events and so on, allowing retailers to decrease both out-of-stocks and over-stocks. Big data analytics can also deliver full supply chain visibility by capturing real-time inventory positions across the enterprise and through the extended supply chain. Data to be leveraged include open purchase orders, in-transit inventory, or vendor and distributor inventory. Insight into channel behaviours are also used to identify renewal needs, by analysing customer sentiment, abandoned baskets, click through, time on page, etc.

Retail

For consumer marketers, the old adage still holds true: reach the right audience with the right message at the right time. Consider the Amazon's recommendation engine: it's an instant message to a singular person based on his or her actions, and the actions of similar people. This is a smart system, powered by big data.

The other end of the spectrum would be a prime time TV ad or a roadside billboard. They are mass marketing at a time when the person cannot make an immediate purchase. Email campaigns used to be like this. Nowadays, companies are harnessing campaign management tools to harness insights to increase the effectiveness of their outreach with personalization.

The next generation of consumer marketing offers multi-touch capabilities that seamlessly connect a personalized message across media, devices and locations to drive sales. Plus, the real-time data generated by sensors (which create an Internet of Things) means that consumer marketers will be able to take advantage of location-based data like Apple's iBeacon in stores as well external data like weather forecasts or social media sentiment that create a truly relevant and targeted buying environment.

Fraud detection

Fraud and financial crime is serious business, just ask the companies and individuals that have been victimized in the past year. With loads of transactions occurring per day, the financial data is most certainly big, making identifying a fraudulent purchase a challenge of pattern recognition. Patterns become stronger with more data points which give financial services companies, retails and other high-volume players the impetus to look at as much data as possible. However, when a bank has to process huge historical data along with real-time data nightly to detect fraud, it becomes nearly impossible to keep up with both the volume and velocity of data.

In addition, enormous sets of location data are often leveraged to increase the effectiveness of fraud detection. Past transactions indicate where and when a consumer usually shops. If a person purchased a plane ticket to London, it would make sense that transactions will start to appear in London during the trip.

While volume can be the key to recognizing fraud, velocity is the key to preventing fraud. New real-time big data platforms enable companies to process massive quantities of historical information and validate new transactions in real-time to spot patterns and halt a transaction before it occurs. By having real-time data at their fingertips, data scientists can also look at new information on the fly, evolving countermeasures just as criminals are adapting to security that's already in place. With every measure of fraud prevention, companies can lower their costs, protect their assets and customers.

Targeted marketing and churn avoidance

In marketing campaigns, segmenting customers is a part of typical objectives of increased satisfaction to prevent customers from churning. The key to a big data driven advanced analytics solution providing optimal churn prevention will be its ability to provide preventive churn actions in real time. Strategic use of big data and advanced analytics enables service providers to shift their business intelligence focus from looking back at old records to looking forward with current data in predictive and preventive fashion to determine things such as behaviour triggering churn events and steps to prevent a churn event.

Customer sentiment

Customer sentiment is being expressed for every enterprise, product and service in existence over numerous social channels at an increasing rate. Using social monitoring and text mining tools, there's a compelling opportunity to analyse what prospects and customers think about each of your products or services, as well as what they think about each of your competitors' products or services, and correlate this sentiment analysis to sales efforts, product mix, marketing spend, advertising expense, loyalty programs, market share, customer share, competitor programs and specific cost and profit measures. This type of correlation is powerful in manipulating company operating decisions to influence customer behaviours with predictive responses. There is also an opportunity to correlate customer sentiment analysis with broad economic factors, specific market indicators, competitor moves or other factors that may uncover patterns that permit companies to model changes for improved customer consumption and company performance.

Logistics

There have been many ways that enterprises have attempted to improve logistics efficiency through data, such as massive warehouse management system software and advanced planning systems that can predict the best quantity and location of inventory to avoid stock outs or dead inventory. The big data era builds on these concepts and takes advantage of advancements in data collection to make logistics faster, smarter and more efficient.

For instance, RFID tags and sensors create a passive network of communication, tracking each item and its location. This boosts productivity when goods are shipped, received, picked and packed because goods can move without the scanning of barcodes. This constant flow of real-time data is a goldmine of information that can be the catalyst for new insights and optimizations.

Big data and business models

In certain industry sectors, such as financial services, big data has urged completely new business models. For instance, algorithmic trading nowadays analyses huge amounts of market data on a minute-by-minute basis, detecting opportunities to capture value instantly. In the retail sector, big data expedites analysis of in-store purchasing behaviours in near real time. With such fast insight into demand shifts, stores can adjust supply, stock levels, and prices to maximize sales.

Big data can produce large data sets coupled with enormous processing capabilities to spur growth and reveal cost-reducing opportunities across industries. While every industry uses distinct approaches and emphases on different aspects from marketing to supply chain, almost all are engrossed in a transformation that leverages analytics and big data. As organizations evolve, so must their analytics capabilities, moving from basic to the more mature predictive analysis. Basic analytics provide a historic view of business performance: what happened, where it happened, how many times it happened. Anticipatory analytics identify unique drivers, root causes, and sensitivities. Predictive analytics perform business modelling and simulations and try to predict what will happen.

When tackling big data challenges organizations react and organize the response in different ways. We have identified the following three general patterns in how companies organize themselves to create value from big data. These patterns can be valuable tools for evolving from a data-and-information focus to a business insight-and-foresight focus. Each model has its pros and cons[26].

Decentralized services model: Every business has its own analytics group, which enables and encourages rapid decision making and execution. But normally there is no dedicated role for strategic planning or best-practice sharing, which can result in duplicate resources and infrastructure. This model increases focus, but the lack of an enterprise view can undermine opportunity.

Embedded shared-services model: It is a centralized model that spins under an existing business unit and serves the entire organization. It can speed execution and decision making, and its structure, support processes, and standards increase efficiency and IT proficiency.

[26] See "IT Innovation Spurs Renewed Growth" at www.atkearney.com

Standalone shared-services model: It is analogous to the embedded model but exists outside business entities or functions. It has direct executive-level reporting and elevates analytics to an imperative core competency rather than an enabling capability.

5.2 Big Data in science

There are several big data applications in scientific disciplines like astronomy, atmospheric science, medicine, genomics, geochemistry and other complex and interdisciplinary scientific studies. Web-based applications encounter big data frequently, such as social computing, Internet text and documents, Internet search indexing. Nowadays, there are numerous sensors around us, they produce seamless sensor data that need to be used, namely, intelligent transportation systems (ITS) (Zhang et al., 2011) utilize the analysis of large volumes of complex sensor data.

Several scientific disciplines are to a large extent data-driven (Szalay, 2011, Bryant, 2011) with the recent progress in computer sciences. Astronomy, social computing, meteorology (Wang et al., 2007), computational biology (McDermott et al., 2009, Akerkar, 2013) and bioinformatics (Lynch, 2008) are based on data-intensive discovery as huge amount of data with diverse types produced are utilized in these scientific domains (Fey et al., 2008).

The e-science evolving from big data is a transformed science, and its world is new to those of us trained in classical scientific paradigms. The e-Science represents to the large scale science that is progressively carried out through distributed world-wide collaborations enabled by the Internet. Main characteristics of such collaborations are that they will require access to very large data collections, very large scale computing resources and high performance visualisation. Big data has an intense liaison with e-Science (Hey et al., 2002) that is computationally rigorous science which is implemented in distributed computing systems. Several concerns on big data applications can be resolved by e-Science which require grid computing (Jakob et al., 2005). The e-Sciences include particle physics, bio-informatics, earth sciences and social simulations. It offers technologies that facilitate distributed collaboration, such as the Access Grid. Particle physics has a sophisticated e-Science infrastructure in particular because of its need for adequate computing facilities for the analysis of results and storage of data originating from the European Organization for Nuclear Research (CERN) Large Hadron Collider, which commenced acquiring data in 2009. e-Science is a broad notion with many sub-areas, such as e-Social Science which can be regarded as a prominent development in e-Science; and helps collecting, processing, and analysing the social and behavioural data.

The Large Hadron Collider (LHC) is a particle accelerator that can generate 60 terabytes of data per day (Brumfiel, 2011). The patterns in those data can give us a unique insight of the nature of the universe. The volume of human genome information is enormous, and it initially took a decade to process and decoding it. A lot of other e-Science projects (Hey et al., 2002) are planned or ongoing in wide-ranging research areas, such as environmental science, oceanography, geology and sociology. We can observe that enormous data sets generated in aforesaid fields greatly demand automated analysis. Furthermore, integrated repositories are essential as it is unrealistic to reproduce copies for remote research groups. Consequently, centralized storage and analysis approaches drive the entire system designs.

5.2.1 Examples of typical big data use cases

Bio-medical science

Mass spectrometers generate massive amount of complex proteomic data in a high-throughput manner, challenging traditional analytics systems and workflows. A moderate sized lab can easily generate several GB of raw data every day. Effective utilization of these data archives for deriving context and biological insights relies on the ability to efficiently store, query and analyse these data using computational and statistical methods.

Seismic exploration
By applying advanced analytics, such as pattern recognition, to a more comprehensive set of data collected during seismic acquisition, geologists may be able to identify potentially productive seismic trace signatures that have been overlooked in newly acquired or archived data Also using/combining with data from other disciplines could enhance exploration efforts. For instance, historical drilling and production data from a nearby well could help geologists and geophysicists verify their assumptions in their analysis of a field. This becomes particularly vital where environmental regulations restrict new surveys.

Climate studies
Climate change prediction research teams are regularly investigating natural and anthropogenic-induced patterns of climate variability and change by means of data analysis and simulations of the earth's climate system. With these model simulations, researchers are able to explore mechanisms of climate variability and change, as well as to detect and attribute past climate changes, and to project and predict future changes. The simulations are motivated by broad community interest and are widely used by the research communities.

Cosmology
One of the important challenges in cosmology is to provide a way to reduce photometric data in real-time for supernova discovery and to handle the large volume of observational data in conjunction with simulation data to reduce uncertainties in the measurement of the cosmological parameters via baryon acoustic oscillations, galaxy cluster counting and weak lensing measurements. Big data specific challenge is to perform analysis on both the simulation and observational data simultaneously.

Large scale geospatial analysis and visualization
As the number of geospatially aware sensors increase and the number of geospatially tagged data sources increases the volume geospatial data requiring complex analysis and visualization is growing exponentially. Traditional GIS systems are generally capable of analysing lots of objects. Today's intelligence systems often contain trillions of geospatial objects and need to be able to visualize and interact with millions of objects.

Big data and scientific research
The advent of enormous data sets and data-intensive science are profoundly changing the way researchers work in almost all scientific discipline. Physicists, Biologists, chemists, cosmologists, earth and social scientists are all gaining from access to the tools and technologies that will integrate the big data into standard scientific methods and processes. Researchers are gradually capable of collecting huge quantities of data through computer simulations, low-cost sensor networks and highly instrumented experiments, creating a *data deluge*.

The advantages of relating research with big data and the cloud have been most apparent in areas, for instance, genome studies. The considerable amounts of data are used to pinpoint latent links between a person's genome and traits such as the tendency to develop a disease or a specific response to a drug.

Environmental science is an excellent place for testing big-data initiatives. This research field has a broad variety and huge volumes of data, which need to be captured speedily. Nevertheless, the approaches and tools developed here can be applied more broadly in fields such as business, government and education. Climate scientists, for instance, may desire to use data to produce a model to make predictions – how climate change will modify ecosystems. Alternatively, they may want to use a model's predictions to create data, such as how changing ecosystems will influence further climate change.

Cosmologists have a lot of work ahead to uncover hidden secrets of our planet and solar system.

However, there are plenty of challenges to overcome first. Scientists are concerned that the data deluge will make it increasingly difficult to find data of relevance and to understand the context of shared data. The management of data presents ever more tough issues. The big question is: How do global, multidisciplinary and often competitive teams of researchers address challenges related to data management, the creation and use of metadata, ontologies and semantics, and still adapt to the principles of security, privacy and data integrity?

5.3 Big Data in public administration

Public administration too encompasses big data problems (Bryant, 2007). It is common that citizens in different age groups need different public services. For instance, kids and teenagers want education whereas the senior citizens require higher level of health care. Nevertheless, every individual in society generates a lot of data in every public service, thus the total number of data about public administration in one country is really massive.

Various Governments are dealing with adverse conditions to enhance their productivity. In such situations big data plays crucial role in public administration. As stated by McKinsey’s report (Manyika et al., 2012), big data functionalities, such as reserving informative patterns and knowledge, provide the public sector an opportunity to enhance productivity and efficiency while maintaining or increasing the quality and level of provided services. European’s public sector could potentially reduce expenditure of administrative activities by 15–20 percent, increasing 223 billion to 446 billion values, or even more.

Through its role in administering the tax system, social programs, and regulation, the federal government collects enormous amounts of granular administrative data. Examples include the abundant micro-level data sets maintained by the Social Security Administration, the Internal Revenue Service, and the Centres for Medicare and Medicaid. Although there is less uniformity, state and local governments similarly generate large amounts of administrative data, particularly in areas such as education, social insurance, and local government spending. Government administrative data are indeed underutilized, by government agencies and, because of limited and restricted access, by researchers and private data vendors who might use this data to uncover new facts. The major data sets also tend to be maintained separately, unlike in many European countries, which may have data sets that merge individual demographic, employment, and in some cases health data, for the entire population. Administrative data is a powerful resource. It typically covers individuals or entities over time, creating a panel structure, and data quality is high (Card et al., 2011).Besides , since the coverage is universal, administrative data sets can be linked to other, possibly more selective, data.

In modern business debates and decisions are usually informed by large amounts of data analytics, and in at least some companies, by extensive experimentation (Varian, 2010). Many government agencies are increasingly smart about using data analytics to improve their operations and services. However, most agencies almost surely lag behind the best private sector companies, and face challenges of both infrastructure and personnel demands. For example, a 2008 report by the JASON study group[27] expressed some of these challenges in the context of how the military must try to process and analyse the vast quantities of sensor data that have become available, such as from drone flights and communications monitoring.
The key challenge in consumer protection is to keep individuals from making decisions they will (predictably) come to regret without proscribing individual choice. Behavioural economics has emphasized that one way to strike this balance is through the framing of decisions (e.g., well-chosen defaults), and another way is through the careful presentation of information. For instance, people can end up making major financial decisions—buying a house, saving for retirement, planning health care spending—without good information about the financial consequences.

5.3.1 Examples of typical big data use cases

Cyber security
Government agencies face many challenges associated with protecting themselves against cyber-attacks, such as managing the exponential growth in network-produced data, database performance issues due to lack of ability to scale to capture this data, and the complexity in developing and applying analytics for fraud to cyber data. Agencies continue to look at delivering innovative cyber analytics and data intensive computing solutions. Government agencies are looking to incorporate multiple streams of data to benefit both human and automated analysis. Cyber data such as host, network, and information from the World Wide Web, are being combined with human oriented information such as psychosocial, political, and cultural information to form a more complete understanding of our adversaries, motives, and social networks.

Public transport
Through improved information and autonomous functions, big data has the potential to transform transportation in many ways. Distributed sensors on handheld devices, on vehicles, and on roads can provide real-time traffic information that is analysed and shared. This information, coupled with more autonomous functions in cars can let drivers to operate more safely and with less disruption to traffic flow.

Microdata services
Microdata services can arise from various government services by, for instance, charging a little fee for basic information queries, such as those from the land registry office, or providing access to decision models that help with applications for financial support, legal assistance, environmental assistance etc.

Tax

[27] http: // www.fas.org/irp/agency/dod/jason/data .pdf

Tax agencies need to minimize tax gaps and increase revenue collection by ensuring that all entities pay their required portion, refunds are issued only to those who legitimately qualify for them, and audits are performed on those most likely to be committing fraud, underreporting income or participating in other tax evasion schemes. Big data analytics can help tax agencies precisely determine who should be investigated for fraud or denied refunds by detecting new deception tactics, uncovering multiple identities and identifying suspicious behaviour.

Preventive policing

Preventive policing or algorithmic law enforcement is a recent field that predicts which areas are most vulnerable to crime, based on historical patterns, weather situations, the impact of certain events etc.

Threats and crimes

Two primary governmental functions are national security and public safety. Predicting potential threats and crimes before they happen and preventing them from occurring can significantly lower risks and improve public safety and national security. Big data analytics can help national security and law enforcement agencies improve intelligence by identifying threats and crimes before they happen, finding critical information faster, detecting associations between people and activities, improving the accuracy of threat and crime analysis, enabling information sharing and collaboration between investigative organizations, protecting sensitive facilities from attack and preventing emergent cyber-security risks.

Law enforcement efficiency

Law enforcement efficiency in relation to road traffic can be accomplished through analysis of traffic data obtained from smartphones and other location-aware devices, such as navigation units; it includes analysis of average speed.

Open data and public sector

The open data association and governments all over the world, including the EU, are devoted to make data publicly available and usable. The EU’s present review of the Public Sector Information Directive aims at unlocking the potential of big data held and accrued by government establishments both with regard to the public sector itself leveraging the potential and efficiencies that come along with a big data strategy, as well as to enable innovators and private enterprise to access big data held by public authorities.

Public entities harvest and hold huge amounts of data which is mostly sensitive or confidential in nature. Government and public institutions have an inherent interest in managing vigilantly this large amount of data, both to improve their performance and generate savings that allow for much sought-after spending cuts, but also to be able to provide open data to their citizens and business entities. At the same time, ensuring that private information is not disclosed.

Big data management is a key asset for the public sector to better conduct its public mandate as well as distribute knowledge and information to the public, empowering citizens and business with open data and information. Although big data is not restricted to data protection issues in many instances personal data plays no role at all, privacy concerns are however an important factor in any big data strategy. The areas of user sentiment and social data analysis, cross referencing and mixing of data acquired from varied sources trigger high demands for a safe and secure legal framework that can protect both data users and suppliers.

6 BIG DATA DEFINITIONS IN SELECTED SECTORS

6.1 Big Data Definition in Oil & Gas Sector

The energy industry is a very diverse sector in Europe and it basically covers:

- the petroleum industry, including oil companies, petroleum refiners, fuel transport and end-user sales at gas stations
- the gas industry, including natural gas extraction, distribution and sales
- the electrical power industry, including electricity generation, distribution and sales
- the coal industry
- the nuclear power industry, and
- the renewable energy industry, comprising alternative energy and sustainable energy companies, including those involved in hydroelectric power, wind power, and solar power generation, and the manufacture, distribution and sale of alternative fuels.

The sector is known for its rapid adoption and ability to adapt challenges of the digital age. The sector is responsible for the extraction of raw materials and using these materials to produce energy.

As assets' yields become harder to access and even harder to forecast, it is vital that the industry is collecting and maintaining its data effectively. Big data is relevant for the whole energy sector. However, since the industry sector is too vast, in this section, we will limit our discussion to Oil & Gas (O&G) industry.

The main objective of the O&G industry is to deliver sources of energy. Different segments of O&G industry are shown in the Figure 5.

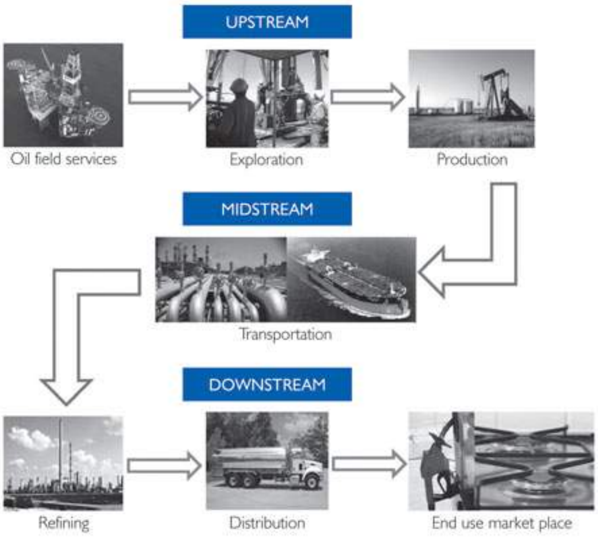


Figure 5 Oil & Gas supply chain (Source: S. Oladunjoye, R. Price, N. Rahman, and J. Wells, Transfer pricing Transfer pricing in the oil & gas sector- A primer, International Tax Review, July 2012)

Oil & Gas upstream industry is multifaceted, data-driven business with data volumes increasing exponentially (Feblowitz, 2012). The processes for Oil and Gas (O&G) exploration and production generate enormous amounts of data. The data volume and complexity grows day-to-day. With new data acquisition, processing and storage solutions, and the development of new devices to track a wider array of reservoir, wells, machinery and personnel performance, over-all data is growing fast. Upstream organizations work simultaneously with both structured and unstructured data. They must capture and manage more data than ever and are striving to store, analyse and get useful information from these huge volumes of data. Under these conditions, the traditional analysis tools would fail but with the appropriate infrastructure and tools, Oil & Gas companies can get quantifiable value from these data.

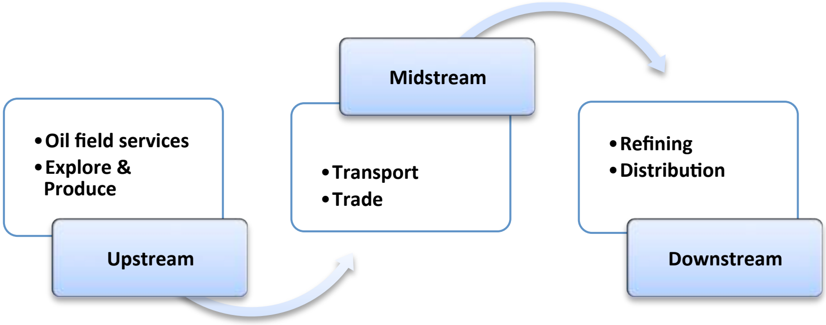


Figure 6 Type of trade and transaction flows in O&G industry

Oil & Gas companies use thousands of sensors installed in subsurface wells and surface facilities to provide continuous data collecting, real-time monitoring of assets and environmental conditions (Brulé, 2013). Nowadays, organizations are capturing a greater volume and variety of data, at a faster velocity, than ever before. Other than sensor data, big data includes large volumes of semi-structured and unstructured data— varying from high-frequency drilling and production measurements to daily, written operations logs—that rapidly produce terabytes of new data. It also comprises a huge collection of business data, such as internal financial results, and news on energy and petroleum competitors bidding on leases and making major capital investments. Therefore, right tools for data analysis should be used (Hems & al., 2013). To support the real-time decision-making, Oil & Gas industry need tools that integrate and synthesize diverse data sources into a unified whole. Being able to process big data makes it possible to derive insight from the relationships that will surface when all of these sources are processed as a whole. But to unlock this value, Oil & Gas companies need access to the appropriate technology, tools, and expertise.

6.1.1 Big Data Applications in Oil & Gas Sector

O&G industries are basically concerned with managing the massive complexity of Exploration and Production (E&P) data dominated by physical media and incompatible proprietary digital storage systems. Typically energy companies spend in E&P data management, handling streams of often incompatible data from different stages of a production operation lifecycle.

The industry is challenged by the time needed to process the data logs to their fullest extent as it requires enormous human intervention. It involves monotonous and time-consuming efforts and infrastructure cost to cleanse the data and derive meaningful context of the information embedded in these unstructured data in quick time.

For example, the big data solution can benefit in creating an Integrated Digital Oil Well which will provide a unified Well Life Cycle Management (WLM) methodology to assess the business process and recommend a Well Master Strategy for automating critical metrics and integrated workflows to apply Business Intelligence for optimizing business operations (Holdaway, 2009).

The following table provides some business benefits[28] which can be realized by using big data solutions in the O&G sector:

Table 2 Business benefits using big data solutions

Improved Operations	Drive combined insights from a single E&P data management platform for structured, unstructured and real-time data. Reduce the operational non-productive time and HSE (health, safety and environment), regulatory compliance cost due to "real-time risk management".
Unified Ontology for O&G sector	Provide personnel with access to searchable institutional knowledge that compensates for limited expert staffing and achieving accuracy and helping personnel find what they are looking for more quickly. The time saved in accessing and loading data is important given the shortage of experts.
Faster Production Rate	Accelerate time-to-production by minimizing data bottlenecks that reduce asset team productivity. Enable faster decision-making by Geologist & Geophysicists and operational teams as risk profiling and forecasting is performed.
Asset Development	Improve asset uptime and predict the need for asset related operational demands.
Enhanced safety and efficiency	Enhanced safety and efficiency in drilling operation by linking well and drilling data with physical models. This development is also paving the way for integrating these systems into the control system, which in turn can/will facilitate autonomous drilling.

By combination of big data and advanced analytics in Exploration and Production activities, experts can accomplish strategic and operational decision-making. The challenge in exploration is to provide quick, faultless, and automated access to structured and unstructured seismic data for geophysical interpretation. This linkage enables geotechnical professionals to understand the context in which seismic surveys were conducted, and it makes supplementary information available in real time to support the decision-making process. Additional benefits are gained when well master data is integrated with unstructured information. Correlating seismic and well production data is critical to enabling unified production and profitability analysis.

The areas where the big data analytics can benefit O&G exploration include:

- Historical drilling and production data help geologists and geophysicists verify their assumptions in their analysis of a field where environmental regulations restrict new

[28] http://www.igate.com/industries/energy

surveys (Feblowitz, 2012). Integrate enterprise data with real-time production data to deliver new insights to operating teams for enriching exploration efforts (Hems et al., 2013).

- Conceive competitive intelligence using Analytics applied to geospatial data, O&G reports and other syndicated feeds in order to bid for new prospects (Hems et al., 2013).
- By means of advanced analytics based on Hadoop and distributed Database for storage, quick visualization and comprehensive processing and imaging of seismic data to identify potentially productive seismic trace signatures previously (Hems et al., 2013).

Drilling and Completion

By using historical drilling data it is feasible to quantitatively identify best and worst practices that impact the target. The intent is that these insights will improve future drilling operations in unconventional plays and potentially in conventional fields.

Real-time information returned from supervisory control and data acquisition systems on well-heads can be used to grab prospects that boost asset performance and optimize production (Hems et al., 2013).

Associated fields where analytics can enhance geoscience include:

- Combine geologic measurement and scientific models into routine processes, such as shale development.
- Involve cutting-edge subsurface models and conduct detailed engineering studies on wells to identify commercial prospects earlier and with less risk.
- Utilise new models and simulators to leverage exploration activities to recognize the earth's subsurface better and to deliver more reasonable energy, safely and sustainably.

Production

Big data also plays an important role in production and operation work. Oil recovery rates can be improved by integrating and analysing seismic, drilling, and production data to provide self-service business intelligence to reservoir engineers.

- Big data analytics applied to seismic, drilling, and production data could help reservoir engineers map changes in the reservoir over time and offer decision support to production engineers for making changes in lifting methods. This approach could also be used to guide fracking in shale gas plays (Feblowitz, 2012).
- Identify how maintenance intervals are affected by variables such as pressure, temperature, volume, shock, and vibration to prevent failure and associated downtime.
- Forecasting production at thousands of wells. Aging wells where the forecast does not meet a predetermined production threshold are flagged for immediate remediation (Feblowitz, 2012).
- Real time analytical solutions provide the mission critical business needs like predicting the behaviour of a device under specific set of conditions and defining the appropriate action strategy. Real-time SCADA[29] and process control systems

[29] http://scada.com/

combined with analytics tools help O&G producer to optimize resource allocation and prices by using scalable computing technologies to determine optimum commodity pricing. They also, help to make more real time decisions with fewer engineers (Hollingsworth, 2013).
- By detecting well problems before they become critical.

Reservoir Engineering

O&G companies improve understanding of future strategy based on available oil for a better identification of reservoirs and reserves by integrating real-time data into the earth model both on the rig and in the office. They also predict the likelihoods of success of turning reservoir into a production well by:

- Engaging cutting-edge subsurface models and conduct comprehensive engineering studies on wells to identify commercial prospects earlier and with less risk (Feblowitz, 2012).
- Use big data tools to understand the earth's subsurface better and to deliver more affordable energy, safely and sustainably.

Equipment maintenance

In upstream, if pressure, volume, and temperature can be collected and analysed concurrently and compared with the past history of equipment failure, advanced analytics can be applied to predict potential failures. Several upstream operations are in remote locations or on ships, so being able to plan maintenance on critical assets is vital, especially if work requires purchase of specialized equipment (Feblowitz, 2012). Specialists often use data collected from pumps and wells to adjust repair schedules and prevent or anticipate failure. Better predictive maintenance becomes possible (Hems et al., 2013):

- Comprehend how maintenance intervals are affected by variables such as pressure, temperature, volume, shock, and vibration to prevent failure and associated downtime.
- Use this insight to predict equipment failures and enable teams to efficiently schedule equipment maintenance in the field.
- Integrate well and tool maintenance data with supply chain information to optimize scheduling of shop floor maintenance.

6.1.2 Big Data Challenges in O&G Sector

There are several challenges in the O&G sector. It is an imperative task to extract business-critical intelligence and insights from large volumes of data in a complex environment of legacy diverse systems and fragmented and decentralized solutions that are common in the O&G sector.

Like generic big data, the O&G Data is also characterized by the 5V (Baaziz, 13):

Table 3 5V in O&G sector

Volume	- Seismic data acquisition - Seismic processing
Variety	- Structured: standard and data models such Professional Petroleum Data Model[30] (PPDM) - Unstructured: images, log curves, well log, maps, audio, video, etc. - Semi-structured: processed data such as analysis, interpretations, daily drilling reports, etc.
Velocity	- Real-time streaming data from well-heads, drilling equipment, and sensors - Relevant data fragments needs to be automatically detected, assessed and acted upon.
Veracity	- Improve data quality - Run integrated asset models - Combination of seismic, drilling and production data - Drive innovation with unconventional resources (shale gas, tight oil) - Pre-processing to identify data anomalies
Value	- Increase speed to first oil - Enhancing production - Reduce costs, such as Non Productive Time (NPT) - Reduce risks, especially in the areas of Safety and Environment

Generally O&G companies are concerned with challenges associated in managing complexity of E&P data such as seismic, drilling, well and production. Furthermore upstream data is rising exponentially in the form of both structured as well as unstructured data. The common challenges and possible approaches to tackle these challenges are listed in the Table 4.

Table 4 Big Data Challenges in the O&G Industry

Challenges	Approach
Data from different sources (structured, unstructured & real-time)	Leverage the power of Hadoop , NoSQL databases for scalable information management systems in batch and near-real time streams to fulfil need for homogenous, integrated and perspective based information
Huge volume of domain specific information embedded in each data cluster	Agile big data techniques, distributed processing, data mining for Oil Well drill parameter configuration models
Use of different software products for data interpretation and decision	Agile big data techniques for consistent asset models, optimized OpEx and CapEx, effective

[30] https://www.ppdm.org/ppdm-standards

making	monitoring and integration between operation and business system
Difficulty in using data to quickly and efficiently respond to user needs	Analysing Oil Well productivity, planning, uncertainty in delivery of energy and managing storage
Huge expenses on E&P data management, handling streams of often incompatible data	Empower consumers with web, mobile enabled dashboards by easy slice and dice of data, planning innovative services and predictive risk modelling

6.2 Big Data Definition in Healthcare

The healthcare sector is facing tremendous challenges. On the one hand, there is a continuous demand to improve the provision of preventive, curative and rehabilitative medical services. On the other hand, healthcare is one of the main governments' expenditures, and there is ongoing pressure to control their growth – this is especially difficult in Europe with an aging population and the emergence of new and more expensive treatments.

Big data can give response to some of the previous challenges by exploiting the riches of medical datasets. There are four main sources of medical data that can be exploited[31]:

- Clinical data: electronic health records (EHRs) with patient diagnostics, laboratory values or medical images.
- Pharmaceutical R&D data: clinical trials, drug datasets, etc.
- Activity (claims) and cost data such as utilization of healthcare and cost estimates.
- Patient behaviour and sentiment data.

For all the promise of improving healthcare through data, the surprising reality is that big data is not utilised in healthcare settings[32]. In fact, data continues to be regarded by many as a "waste product" of the system: "In many industries, we collect a lot of data, and just haven't learned how to analyse and use it. In healthcare, arguably, we don't collect big data at all" – said Michael Chui of McKinsey & Co[2].

Nevertheless, some authors consider that the application of big data to healthcare is inevitable[33]. The digitization of medical records is a prerequisite, while the combination of medical data sources can significantly leverage the value of the healthcare data deluge. As a note of caution, the complexity of the medical domain is much higher than any other sector in which big data has consolidated, e.g. retailing.

In the remaining of this section we will succinctly describe the main applications of big data in healthcare, as well as the most relevant data challenges so far. We do so by conducting a literature review of some of the most relevant works that analyse big data in healthcare.

[31] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[32] Bollier, D. (rapporteur) (2010). The promise and peril of big data. Washington, DC, USA: Aspen Institute, Communications and Society Program.

[33] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. JAMA, 309(13), 1351-1352.

6.2.1 Big Data Applications in the Healthcare Sector

Table 5 presents the most promising big data applications in healthcare, along with the seminal works that identify them.

Table 5 – Main big data applications in healthcare

Generation of new knowledge	Bollier[34], Manyika et al.[35], Murdoch & Detsky[36]
Improved drug research and development	Bollier[4], Manyika et al.[5]
Information access and clinical decision systems	Løvoll & Kadal[37], Manyika et al.[5], Murdoch & Detsky[6]
Patient monitoring and management	Løvoll & Kadal[7], Manyika et al.[5]
Personalized medicine	Bollier[4], Manyika et al.[5], Murdoch & Detsky[6]
Epidemiology surveillance	Bollier[4], Løvoll & Kadal[7]
Transparency about medical data	Manyika et al.[5]
Participatory healthcare	Bollier[4], Manyika et al.[5], Murdoch & Detsky[6]

Generation of new knowledge

Medicine typically relies on experimental studies such as randomized trials for generating medical evidence. Big data offers the potential to derive further knowledge by analysing the data contained within EHRs – especially text-based annotations. This way, it is possible to obtain observational evidence for clinical questions that could not be possible otherwise.

Improved drug research and development

Clinical drug research often employs small sets of data, especially after drugs are introduced to the marketplace[4]. Access to larger datasets of patient populations can greatly improve drug surveillance. Other possible uses in this area include the predictive modelling for new drugs in order to detect the most promising allocation of resources, as well as analysing clinical trials and patient records to detect adverse effects[5].

Information access and clinical decision systems

[34] Bollier, D. (rapporteur) (2010). The promise and peril of big data. Washington, DC, USA: Aspen Institute, Communications and Society Program.

[35] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

[36] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. JAMA, 309(13), 1351-1352.

[37] Løvoll, G. & Kadal, J. G. (2014). Big data - the new data reality and industry impact. DNV GL Strategic Research & Innovation.

Since there are so many medical publications it is difficult for physicians to stay current. Medical guidelines and literature reviews are especially important for compiling clinical practice, although physicians still have problems for obtaining relevant information when dealing with complex or multiple illnesses[6]. Better query systems and visualization techniques can help to improve data access. In addition, clinical decision systems can check potential problems in a treatment by analysing EHRs and medical guidelines.

Patient monitoring and management

Chronically ill patients consume many hospital resources[7]. A remote monitoring system can be employed to treat such patients at their homes, since anomalies in their condition can be detected at an early stage.

Personalized medicine

Big data can help to translate personalized medicine practices into clinical practice. This is possible through the analysis of large datasets, e.g. genomics, with EHR data.

Epidemiological surveillance

One of the tasks of health authorities is to monitor the spread of certain diseases such as the flu. While this information is typically obtained from hospitals and clinics, it is possible to early detect the diffusion of a contagious disease by using social media data. Google Flu Trends[38] is an example of such a service that analyses live flu-related searches to early predict flu activity.

Transparency about medical data

Opening up medical data can help to identify performance opportunities for medical professionals, processes and institutions. In addition, patients can make more informed healthcare decisions, e.g. choosing a clinic. Moreover, data transparency can create a competition incentive to improve performance, even without a tangible reward.

Participatory healthcare

Patients can play a more active role in their healthcare by giving them direct access to healthcare data. They are increasingly using the Web to find information about their injuries and illnesses, while there are emerging social networks for exchanging healthcare information and providing support to each other, e.g. PatientsLikeMe[39].

6.2.2 Big Data Challenges in the Healthcare Sector

Despite the potential benefits of big data in healthcare, there are several challenges that should be addressed. The prominent ones from our literature review are shown in Table 6.

Table 6 Big data challenges in healthcare

Security and privacy rights protection	Bollier[40], Manyika et al.[41], Murdoch & Detsky[42]

[38] http://www.google.org/flutrends/

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

[40] Bollier, D. (rapporteur) (2010). The promise and peril of big data. Washington, DC, USA: Aspen Institute, Communications and Society Program.

[41] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Fragmentation and limited interoperability of EHR platforms	Bollier[10], Manyika et al.[11], Murdoch & Detsky[12]
Structural issues and resistance to change	Bollier[10], Manyika et al.[11], Murdoch & Detsky[12]

Security and privacy rights protection

Personal data in healthcare is especially sensitive, and there is potential for discrimination based on it. Therefore, security and privacy rights protection are critical[11]. Anonymation techniques should be taken and their efficacy tested before sharing personal healthcare data with another party[10].

Fragmentation and limited interoperability of EHR platforms

Murdoch & Detsky report that current EHR platforms are fragmented and have limited interoperability[12]. This is mainly due to the variety of medical data and the diversity of incompatible data formats. As a result, many datasets currently remain locked in silos that do not communicate.

Structural issues and resistance to change

While the previous challenge is related to technology, there are other structural issues that must be tackled for the adoption of big data in healthcare. For example, there are no strong incentives or champions for data use within hospitals or clinician groups[43]. Moreover, clinics, pharmaceutical companies, physicians and patients often believe that their interests will be harmed by the collection and use of data[44]:

- Patients worry that the disclosure of their health records could result in discriminatory treatment or have a negative effect in their insurance.
- Physicians are traditionally rewarded by frequent visits by patients; so preventive care and better health outcomes can negatively affect their income.
- Pharmaceutical companies are reluctant to carry out post-marketing drug surveillance in order to not reveal unnoticed adverse effects.
- Clinics could reduce their revenue for treatments found to be not effective, especially in comparison with other clinics.

We conclude this chapter with an assessment of the characteristics of medical data, according to the 5V model proposed in this deliverable. This assessment is depicted in Table 7.

Table 7 5V in Healthcare

Volume	- Massive datasets from electronic health records (EHRs) - Large drug datasets - Huge R&D datasets, e.g. genomics
Variety	- Quantitative data, e.g. laboratory values - Qualitative data, e.g. text documents, medical images, demographics

[42] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. JAMA, 309(13), 1351-1352.

[43] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. JAMA, 309(13), 1351-1352.

[44] Bollier, D. (rapporteur) (2010). The promise and peril of big data. Washington, DC, USA: Aspen Institute, Communications and Society Program.

	- Transactional data, e.g. records of medical delivery
Velocity	- Patient monitoring, especially in intensive healthcare
Veracity	- Improve data quality - Detection of inconsistencies and anomalies - Combination of medical data
Value	- Knowledge creation from the analysis of EHRs - Improved access to data, e.g. clinical guidelines - Improving efficiency and reducing operational costs

6.3 Big Data Definition in environment

Environment (and earth science) sector is rapidly entering the new paradigm of large scale, data-intensive analytics to understand our complex and ever changing planet. Environmental science is an excellent proving ground for big-data initiatives – it has a wide variety and large volumes of data, which need to be captured rapidly. Environmental data sets are growing in size, variety and complexity at an unprecedented rate, creating new challenges and opportunities for their access, manipulation and archiving. With the advent of modern, real-time analytics, massive environmental data sets are faster and cheaper to obtain than ever before, and the vast universe of big data applications is intersecting the realm of economic feasibility.

Big environmental data can be defined as massive or complex sets of structured data (e.g., databases of environmental data, such as chemistry measurements) or unstructured data (e.g., photos or historical records) that may be pertinent to an environmental issue or query at hand. These related data sets may not be readily available (i.e., in digital format) or apparently related at first glance, nor are they easily analysed by conventional channels. Examples include "long" data, such as many analytical results collected over decades from a lengthy remedial investigation, and "wide" data, such as water, sediment, soil, and tissue samples measured for chemistry, toxicity, and community structure, along with bathymetric records, historical site photos, and processing records.

6.3.1 Big data applications in environment sector

The remote sensing and Earth Observation communities are paying a lot of attention to big data since new sensors with high spatial, temporal and radiometric resolution are increasingly providing an ever growing data amount (ESA 2013). But it is expected that also in-situ observatories, including crowdsourcing-oriented platforms and mobile tools, providing a large amount of small heterogeneous datasets, will require big data tools in place. In urban settings, for example, big data strategies will surely require to enable access and analysis of the immense amount of social and environmental observation data stream that is collected from intelligent sensors and citizens (Provost & Fawcett, 2013).

Steed et al. (2013) describe a visual big data analytics system, called the Exploratory Data analysis ENvironment (EDEN), with specific application to the analysis of complex Earth system simulation data sets. Hampton et al. (2013) encourage ecologists to join large scientific communities and global initiatives to address scientific and societal problems by making their small data sets publish in big repositories to harness its collective power.

If ecologists' data, and in general scientists', have proved so valuable, surely equivalent benefits could be gained by sharing and integrating the data generated by people. The unprecedented combination of ubiquitous connectivity, today's mobile technology, and ubiquitous sensing allowed users to become citizen-scientist, i.e. to become actively involved in different stages of research projects (e.g. Citizen Science Alliance[45], Extreme Citizen Science[46]). From this standpoint, citizens at different level of technical expertise are empowered to collect, produce, and publish environmental observations in the interest of the research and society. In this context, participatory sensing and citizen science projects like eBird[47] demonstrate the value of sharing small, localized observations that, when aggregated in integrated in big data repositories, build a deeper and broader understanding of ecological phenomena.

Scholes et al. (2012) expose the need for combining different types of biodiversity data into an integrated, global system for biodiversity in the realm of the Group on Earth Observations Biodiversity network (GEO-BON). The authors stated that "a comprehensive and integrated observation system for biodiversity at several scales, from the subnational to the global for the purpose of protecting and improving biodiversity and human well-being. The system should help to compare the status of biodiversity at different places and track changes in biodiversity at a given place over time".

Along the same lines Havlik et al. (2011) observed the importance of user communities in generating valuable environmental observation data. They noted, however, that "these communities' environmental observations represent a wealth of information which is currently hardly used or used only in isolation and therefore in need of integration with other information sources, which will lead to a paradigm shift from a mere Sensor Web to an Observation Web" (Havlik et al., 2011; p. 3874). So, citizen science data (crowdsourcing, user-generated data) is essentially a form of data sharing, and the challenges of using citizen-science data reflect classic data-sharing and data integration challenges more generally present also in big data analytics, service-oriented architectures and in cloud computing.

In the big data and citizen science context, the ENVIROFI project[48] has paved a path towards the seamless integration of citizen-generated and research-generated environmental observation data. The set of ENVIROFI applications are clear examples of mobile cloud computing (Fernando et al., 2013) which take advantage of data context aspects (e.g. user's location, objects in the vicinity, etc.) and run mobile applications based on remote cloud-based services.

These aspects become even more challenging when multidisciplinary applications are concerned. Indeed the science of today is required to answer to complex and urgent questions involving several disciplinary domains. Combining weather and chemical models to estimate air quality and pollution and their impact on animal and human health, integrating climate change scenarios and ecosystems data to predict how biodiversity is affected (Nativi et al.,

[45] http://www.citizensciencealliance.org/
[46] http://www.ucl.ac.uk/excites
[47] http://www.birds.cornell.edu/citsci/
[48] http://www.envirofi.eu/

2009) or the possible invasion of alien species are just examples of answers that we currently need.

This imposes heavy requirements to the digital infrastructures supporting these scenarios since they need to grow becoming smart cyber-infrastructures capable to deal with big volumes of heterogeneous datasets, but also to combine Environmental Sciences models in workflows supporting complex scientific Business Processes (Nativi et al., 2013). ENVIROFI addressed these challenges in its pilots, exploring the use of FI-WARE[49] services for designing and running workflows, and in the design and development of specific mediation enablers (e.g. brokers) supporting different standard (e.g. from OGC or ISO) or community-specific service interfaces, metadata and data models.

6.3.2 Big data challenges in environment sector

According to many scientists and technologists big data would be able to support an entirely new approach to science based on data intensive scientific discovery, named the Fourth Paradigm (Hey et al., 2009). However, it requires innovative enabling technologies for data management, analytics, delivery, and so forth. Indeed many research efforts are now directed towards the development of new technologies or paradigms to support big data requirements.

While the solutions above address mainly the Volume and Velocity axes, other architectural and technological solutions are oriented to address the Variety due to heterogeneity of datasets. The brokered architectures, as an evolution of the Mediation-based approach (Bigagli, 2006), demonstrated a valuable solution for efficiently connecting existing infrastructures (Nativi et al., 2013) and providing heterogeneous resources to big System of Systems, such as the Global Earth Observation System of Systems (GEOSS).

In the scientific domain, several disciplinary areas are facing big data challenges as part of an innovative approach to science usually referred as e-Science. Environmental Sciences have been some of the disciplinary domains mostly pushing to, and potentially benefiting from, the e-Science approach, intended as “global collaboration in key areas of science, and the next generation of infrastructure that will enable it.” (Hey & Trefethen, 2002). They were in the forefront in many initiatives on distributed computing trying to realize the e-Science vision, including High Performance Computing, grid technologies (Petitdidier, Cossu, Mazzetti, Fox, Schwichtenberg, & Som de Cerff, 2009) and cloud services. The reason is that Environmental Sciences raise significant challenges in terms of storage and computing capabilities, as:

1. They encompass a wide range of applications: from disciplinary sciences (e.g. climate, Ocean, Geology) to the multidisciplinary study of the Earth as a System (the so-called Earth System Science). Therefore Environmental Sciences require:
    a. Covering of a diverse temporal range (such as for Climate and Geological studies);
    b. Supporting a wide spatial coverage (the whole Earth, for global studies, and beyond including planetary sciences);
    c. Modelling many different geospatial data types, including profiles, trajectories, regularly and irregularly gridded data, volumes, and so on;

[49] http://www.fi-ware.org/

2. They are based on Earth Observation, requiring handling observations and measurements coming from in-situ and remote-sensing data with ever-growing spatial, temporal, and radiometric resolution.
3. They make use of complex scientific modelling for deriving information from the large amount of observations and measurements.
4. They are bases on simulations to study complex scenarios (e.g. for Climate Change).

This points out that - referring to the big data V's - big Volume, big Variety, and high Velocity are typical issues of Environmental Sciences data systems.

6.4 Big Data Definition in Crisis Informatics

This sub-section examines the use of "big data" in crisis informatics. Crisis informatics is used as an umbrella term that "includes empirical study as well as socially and behaviourally conscious ICT development and deployment. Both research and development of ICT for crisis situations need to work from a united perspective of the information, disaster, and technical sciences".[50] Crisis informatics has links to a number of activity areas around crisis management. These include preparedness (training, baseline information gathering, simulations, and conflict prevention), response (coordination, information gathering, and provision of humanitarian relief or aid) and recovery (resource allocation, population monitoring, development). As such, the discussion in this piece is not limited to crisis management, but also includes literature and examples related to humanitarianism, emergency management, first response and socio-economic development. However, crisis informatics and its relationship with big data specifically, is a new and emerging area of research. The first use of crisis informatics is recorded in 2007[51], and its relationship with large-scale data has been the subject of serious investigation since 2011. Given the nascent nature of big data and crisis informatics, this examination relies heavily on grey literature, mass media and Internet resources, as these are the areas in which the most information on big data and crisis informatics can be located given the often protracted timeline associated with the academic publishing process. Nevertheless, this sub-section seeks to outline the relationship between big data and crisis informatics in order to begin to explore and delineate a definition of "big data" within crisis informatics, and to contribute to the overall effort to produce a BYTE definition of "big data".

6.4.1 Big data definitions and typologies

Reports from the UN, the International Federation of the Red Cross and other resources have offered a range of definitions and typologies of "big data" in areas related to crisis informatics. With respect to definitions, these have included attempts at specified definitions, as well as more general definitions via the identification of crisis points. In crisis-related documents, the UN Global Pulse initiative has made the most effort to define big data. First, the *Big Data for Development* report states that big data is "an umbrella term for the explosion in the quantity and diversity of high frequency digital data".[52] Later in the document, the authors expand upon this by stating that:

[50] Palen, L., S. Vieweg, J. Sutton, S.B. Liu & A. Hughes, "Crisis Informatics: Studying Crisis in a Networked World", *Third International Conference on e-Social Science*, Ann Arbor, Michigan, October 7-9, 2007.

[51] Ibid.

[52] UN Global Pulse, *Big Data for Development: Challenges & Opportunities*, United Nations, New York, May 2012, p. 4. http://unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

> "Big Data" is a popular phrase used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques. The characteristics which broadly distinguish big data are sometimes called the "3 V's": more volume, more variety and higher rates of velocity. This data is known as "big data" because, as the term suggests, it is huge in both scope and power.[53]

Here, UN Global Pulse specifically makes reference to the original, Gartner definition of big data, by invoking the three Vs. Similarly, although Patrick Meier, a key figure in digital humanitarianism, does not explicitly seek to define big data, he also invokes the Vs in his discussion about a crisis point his team experienced in relation to their capacity to analyse data. Describing the Ushahidi platform used to map text messages and social media posts after the Haitian earthquake, he notes, "We quickly realized that our platform was not equipped to handle this high volume and velocity of urgent information".[54] Thus, for Meier and Ushahidi, it appears that "big data" emerged at the point of crisis in established computing infrastructures. Others have declined to specifically attempt to define big data and have instead opted to describe its effects. This may include a focus on the ability of high volumes of complex data to assist in "decision-making processes"[55], to a focus on leveraging "complex real-time" data "for the common good"[56].

These different definition streams largely conform to attempts to define "big data" outside of the specific domain of crisis management. As indicated in Section 3, many definitions rely upon the Gartner definition of the three Vs, sometimes even expanding this to include additional Vs. Thus, the Gartner definition remains foundational. Additionally, there is also a sub-set of crisis informatics stakeholders that view "big data" as a continuation and augmentation of existing processes that is not worthy of the "hype" with which they are associated.[57] However, what is not present in these definitions, or texts associated with definition, is a perspective common to many sociology, legal and privacy experts that foregrounds potential human rights infringements and unanticipated consequences of collecting, linking and mining large data sets; and the specific intention to use this information to segregate and profile individuals.[58] Yet, the focus of crisis informatics and related disciplines is often the potentials and possibilities of big data to alleviate human suffering and assist in protecting human rights. Furthermore, these issues are recognized within the literature surrounding big data in crisis informatics more broadly, despite the fact that they are not specifically considered in relation to definitions.

[53] Ibid., p. 13.

[54] Meier, Patrick, "Harnessing the Power of Big Data to Deliver Humanitarian Response", *Forbes Magazine,* 2 May 2013. http://www.forbes.com/sites/skollworldforum/2013/05/02/crisis-maps-harnessing-the-power-of-big-data-to-deliver-humanitarian-assistance/

[55] Heaton, Brian, "Harnessing big data: Emergency managers can benefit from big data during the early stages of a disaster", *Emergency Management*, July/August 2013, p. 44.

[56] Letouzé, Emmanuel Patrick Meier and Patrick Vinck, "Big Data for Conflict Prevention: New Oil and Old Fires", in Franceso Mancini (ed.), *New Technology and the Prevention of Violence and Conflict*, International Peace Institute, New York, April 2013, p. 7

[57] Heaton, op. cit., 2013, p. 44.

[58] See for example, Lyon, David, "Surveillance, Snowden, and Big Data: Capacities, consequences, critique", *Big Data & Society*, Vol. 1, July–December 2014, pp. 1–13, Boyd, Danah and Kate Crawford, "Critical questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon", *Information, Communication and Society*, Vol. 15, No. 5, 2012, pp. 662–679, and Crawford, Kate, "Think Again: Big Data", *Foreign Policy*, 9 May 2013. www.foreignpolicy.com/articles/2013/05/09/think_again_big_data

Typologies

In addition to these definitions, specific types of data are particularly associated with big data and crisis. These include:

Table 8 Types of data associated with big data and crisis

• Social media data (text, visual, moving image, audio)	• Mass media data (text, visual, moving image, audio)
• Geographical Information System data (satellite and/or drone)	• Official publications (text)
• Global positioning data (principally associated with mobile phones)	• Climate information
• Transaction data from smart phones, cash programmes or other on-line transactions	• Digital health records
• Humanitarian or other organisations' databases	• Mapping information

The *Big Data for Development* report, specifically, has also developed typologies to assist in circumscribing "big data" as opposed to other types of data. According to the report, big data often include the following features:

- **Digitally generated** – i.e. the data are created digitally (as opposed to being digitised manually), and can be stored using a series of ones and zeros, and thus can be manipulated by computers.
- **Passively produced** – a byproduct of our daily lives or interaction with digital services
- **Automatically collected** – i.e. there is a system in place that extracts and stores the relevant data as it is generated
- **Geographically or temporally trackable** – e.g. mobile phone location data or call duration time.
- **Continuously analysed** – i.e. information is relevant to human well-being and development and can be analysed in real-time[59]

In addition, the report constructs a taxonomy of relevant digital sources, again with specific relation to big data for development, including:

- **Data Exhaust**—passively collected transactional data from people's use of digital services like mobile phones, purchases, web searches, etc., and/or operational metrics and other real-time data collected by UN agencies, NGOs and other aid organizations to monitor their projects and programs (e.g.,, stock levels, school attendance); these digital services create networked sensors of human behavior;
- **Online Information** – web content such as news media and social media interactions (e.g., blogs, Twitter), news articles obituaries, e-commerce, job postings; this approach considers web usage and content as a sensor of human intent, sentiments, perceptions, and want;

[59] UN Global Pulse, op cit., 2012, p. 15. The authors of the Big Data for Development report note that real-time in this context should be understood as a relatively short and relevant time period.

- **Physical Sensors** – satellite or infrared imagery of changing landscapes, traffic patterns, light emissions, urban development and topographic changes, etc; this approach focuses on remote sensing of changes in human activity;
- **Citizen Reporting or Crowd-sourced Data** – Information actively produced or submitted by citizens through mobile phone-based surveys, hotlines, user generated maps, etc; While not passively produced, this is a key information source for verification and feedback.[60]

The UN Office for the Coordination of Humanitarian Affairs (OCHA) has stated that big data sets often originate from three different sources, individuals, governments and the private sector, and they construct the following matrix to associate data sources with particular types of data:

Table 9: OCHA matrix of data sources and types[61]

Source	Data type
Individuals	Data "exhausts" from devices Social media SMS
Governments	Census and geo-data Tax information Public indicators (e.g. health)
Private sector	Transaction data Spending information GSM aggregate data

This matrix is particularly useful in demonstrating that big data and crisis informatics requires different types of stakeholders to work together to make big data available to crisis managers, humanitarians, first responders and other actors. It also demonstrates the different varieties of data that crisis informatics professionals are working with in this sphere. As the discussion of applications below demonstrates, GIS data and social media data, particularly for mapping purposes are particularly visible in this area. According to Heaton, this is because such information is easiest to come by[62]; it is often open source data and thus accessible immediately and without restriction to authorities and humanitarian organizations who often do not have budget or time to negotiate with private companies about proprietary data. In terms of whether this data is considered "big", research by the Woodrow Wilson Center has indicated "Hurricane Sandy in 2012 generated more than 20 million tweets, several terrabytes of satellite and aircraft imagery, and an incalculable number of emails, SMS/test messages, and documents."[63] While this volume may not be considered "big data" in some contexts, governments, humanitarian organizations, emergency managers, local

[60] Ibid., p. 16.

[61] UN Office for the Coordination of Humanitarian Affairs (OCHA), *Humanitarianism in the network age: Including world humanitarian data and trends 2012*, United Nations, 2013, p. 26.

[62] Heaton, op cit., 2013.

[63] Crowley, John, *Connecting Grassroots and Government for Disaster Response*, Commons Lab, Woodrow Wilson International Center for Scholars, Washington DC, 2013, p. 22.

authorities and other organisations are not themselves data processing experts, nor do they have the resources to hire such experts for occasional occurrences. Given this lack of expertise, attempting to identify useful information from this relatively high volume of different types of data, being generated over a relatively short period of time and under conditions of enormous pressure certainly represents a crisis point in data processing that signals a need for new systems, processes, architectures and organisations to meet this challenge. As such, although it is a new field, crisis informatics certainly represents as a specific case of "big data" practice.

6.4.2 Big data applications in crisis informatics

There are a number of big data analytics in the crisis informatics sector when the data sets above are combined and mined to create new insights. Many of these applications are in a very early stage of development given the relative immaturity of the big data in this area. However, information gathered from a number of recent reports, media articles and web resources demonstrates that big data for crisis informatics has become a diverse and dynamic area. Furthermore, these applications are being used both to bring information in to professional stakeholders and to disseminate and exchange information with members of the public.

The following is not intended to be an exhaustive taxonomy of data applications in crisis management and response. Specifically, some practices – e.g., the use of drones to aid search and rescue, mapping, etc. – have generated significant media attention, but they are not big data applications, as such. They may support, or be fed into, big data processing, but they are not "big data" in and of themselves as the relative immaturity of the drone sector means that they are often restricted to collecting one or two items of data.[64] As such, the following categories focus on the combination of different data resources in order to produce new information that would not have been available by focusing on data sets in isolation.

The International Federation of the Red Cross and Red Crescent Societies (ICRC) *World Disasters Report* offers an initial categorization of the use of big data in crisis informatics. This includes the following categories:

- Situational analysis
- Needs analysis
- Coordination and resource allocation
- Awareness raising
- Community-driven response[65]

This 2013 categorization demonstrates that the ICRC recognizes that big data can be used to bring data in to professionals to assist them in decision making in terms of understanding the situation, identifying needs and coordinating personnel and other resources in order to respond effectively. It also recognizes that members of the community are key stakeholders in the crisis response domain, and that other applications include the use of big data to raise awareness about risks, needs and to encourage members of the public to get involved. However, given the context in which it is produced, this categorization understandably

[64] This is set to change as drone technology develops and matures, and in the near future drones may emerge as big data platforms in their own right.

[65] International Federation of Red Cross and Red Crescent Societies, *World Disasters Report: Focus on technology and the future of humanitarian action*, Geneva, 2013.

focuses on response activities and the immediate and mid-term aftermath of the event. The categorization offered by this report is based on multiple resources and recognizes that big data can assist before, during and after a crisis, often via a cyclical process, as indicated in Figure 7 below:

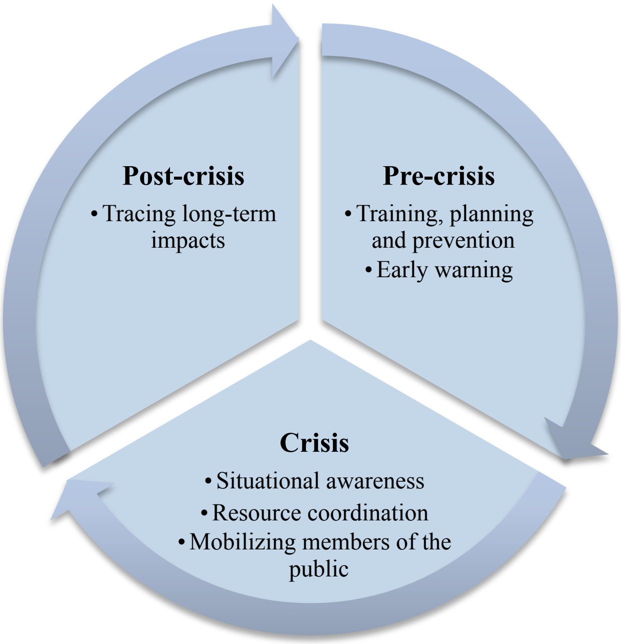

Post-crisis
• Tracing long-term impacts
Pre-crisis
• Training, planning and prevention
• Early warning
Crisis
• Situational awareness
• Resource coordination
• Mobilizing members of the public

Figure 7: Big data application areas in crisis informatics

Each of these application areas demonstrates two key issues. First, as noted above, ordinary citizens are key stakeholders in this process. They act as "seeded" volunteers who are organized before crisis events or "crowd-sourced" volunteers spontaneously emerging during events. Second, that realizing the potential applications of big data in crisis management requires participation from crisis informatics stakeholders (first responders, emergency managers, authorities, and humanitarian organizations), members of the public (as already noted) and industry.

Training, planning and prevention

In the pre-crisis phase, big data can be used to assist in training, planning and prevention activities. Specifically, Heaton argues that a key area of application of big data analytics is in training exercises that prepare responders and other authorities for crisis situations. He notes that big data can build better simulation platforms as these platforms "can at times suffer from a lack of statistical information to fuel predictive models. So where earthquakes, hurricanes or even shoreline erosion events are being trained for, large-volume data sets could help increase the accuracy and reliability of those models."[66] Data from the 2013 Japan earthquake is also being stored and analyse to see if the trends and details can help to develop better tools for future crises. The data being collected, stored and integrated includes the experiences of those who were impacted by the crisis as well as data from Google, Twitter Japan, NHK (mass media), Asahi Shimbun (mass media), Honda and a mapping company called Zenrin[67]. Finally, in the US data from volunteers, geographical surveys, weather services and previous crises is used by the US Army Corps of Engineers to inform the future

[66] Heaton, op. cit., 2013, p. 44.

[67] Appleby, Lois, *Connecting the last mile: The role of communications in the great East Japan earthquake*, Internews, London, 2013.

engineering of bridges, spillways and dams.[68] Thus, the collection of data in previous crises works in a cyclical fashion to inform preparation for future crises.

Early warning

Also in the pre-crisis phase, big data can be harnessed to provide early warning to authorities, crisis stakeholders and ordinary citizens. This early warning may be medium-term immediate (e.g., areas under threat of flooding) or more immediate (e.g., minutes before an earthquake). These early detections may relate to early warnings about physical, topographical changes as well as social or health changes related to members of the public. The ICRC has noted a number of potential early warning services based on big data processing:

> Advances in high-performance computing and the availability of a large number of computers in the cloud (a network of remote servers) have made it possible to compute more complex models for hydrological and seismological risks. This allows decision-makers to make better-informed decisions sooner about which areas to evacuate. Emergency managers have used tools that take advantage of computing technology, for example, the Global Disaster Alert and Coordination System (GDACS), (and) the Humanitarian Early Warning Service (HEWS).[69]

Data generated by citizens, coming in at fast rates and in high volumes can be used to identify the location and intensity of earthquakes. Information from social media "is converted to a real-time map hosted on the USGS (United States Geological Service) website, allowing for the public and responders to witness the distribution of shaking from an earthquake."[70] This can assist scientists and crisis managers to understand the potential impacts of the earthquake, the regions that might be most affected and the type of relief that might be necessary. It can also assist in notifying members of the public. Furthermore it saves time by enabling experts to process this information in less than one minute rather than up to 20 minutes. Finally, the USGS uses this "user-generated content from Twitter to produce a key (public) service, and pushes information back to the public through the same medium."[71] Crowd-sourced information from social media may also provide early warning about natural disasters. In the northeast of the US, individuals monitoring their twitter feed may have received warning of an impending earthquake minutes before they actually felt it.[72] Combining machine processing of data from sensors, Global Positioning and seismic data can also result in the detection of "surface changes caused by natural disasters".[73]

With respect to social and public health issues, social media, sometimes combined with other data, can also be used to provide early warnings. In political crises, big data, such as that coming from social media feeds can also assist in providing early warning about social unrest. Letouzé et al. report for the International Peace Institute reports that social media feed from Iran's post-election crisis can be used to "detect web-based usage of terms that reflect a general shift from awareness/advocacy toward organization/mobilization, and eventually action/reaction within the population (and, thus) model and predict social upheavals and

[68] Bowser, Anne, and Lea Shanley "Community Collaborative Rain, Hail, and Snow Network", *New visions in citizen science*, Woodrow Wilson International Center for Scholars, Washington DC, November 2013, p. 14.

[69] IFRC, op. cit., 2013, p. 104.

[70] Bowser, Anne, and Lea Shanley, "Did You Feel It? and Twitter Earthquake Detection", *New visions in citizen science*, op. cit., 2013, p. 32.

[71] Ibid.

[72] OCHA, op. cit., 2013.

[73] Bowser, Anne, and Lea Shanley, "The Advanced Rapid Imaging and Analysis Project: Validating maps for disaster response", op. cit., 2013, p. 34.

revolutions".[74] OCHA has also indicated that real-time monitoring of Twitter messages combined with proprietary GSM information from mobile phones indicating population movement in Haiti "could have predicted the October/November 2010 cholera outbreaks two weeks earlier than they were detected" by authorities.[75]

Situational awareness

In crisis situations, big data can be used to provide situational awareness both to authorities and members of the public using information coming from scientists, private companies and members of the public. This further demonstrates that members of the public are key stakeholders in the big data and crisis informatics ecosystem.

With respect to providing situational awareness to authorities, big data coming from multiple sources is a key innovation. Principle among these innovations is the combination of GIS and social media data to produce maps of crisis-affected areas. This practice began in Kenya with the open-source Ushahidi platform, but really only began to integrate "big data" during the 2010 earthquake in Haiti. The Ushahidi platform used text-based information provided by volunteers and members of the public to map human rights abuses in Kenya in 2007. In Haiti, the system was expanded to also include SMS messages and social media messages being generated by the hundreds of thousands by members of the public that were automatically mapped onto satellite imagery by the Humanitarian OpenStreetMap community.[76] The map pinpointed the locations, on a street-by-street level and in almost real-time, of damage, missing persons, search and rescue needs, displaced persons, humanitarian infrastructure and refugee camps.[77] The maps were so effective in assisting authorities in understanding the scale of the crisis and the needs of members of the public, which "the administrator of the US Marine Corps even claimed that the live crisis map of Haiti helped them save hundreds of lives".[78] More recently, this automated processing has been augmented with human intelligence and processing to provide more detailed information and reduce false positives. The Artificial Intelligence for Disaster Response (AIDR) system developed by the Qatar Computing Research Institute (QCRI) combines machine learning with human processing by a corps of volunteers to classify information coming in from social media (damage assessment, shelter needs, search and rescue, etc.) and to score the information by relevance.[79]

In addition to mapping based primarily on social media, other resources are also being used to map crisis events and provide situational awareness. GSM data from mobile phones can be used to map people's movements and identify areas of congestion or over-crowding during crisis that can impact the resources that are needed.[80] SARWeather provides on-demand, high-definition weather forecasts for crisis areas, which allow "emergency managers to make operational decisions based on weather conditions".[81] In a final example, scientists at Georgia Tech have developed a landslide detection system called LITMUS that combines data from physical sensors from the "USGS seismic network, NASA TRMM Rainfall network, Twitter,

[74] Letouzé, et al., op cit., 2013, p. 13.
[75] OCHA, op. cit., 2013, p. 27.
[76] IFRC, op. cit., 2013.
[77] Ibid.
[78] Ibid., p. 75.
[79] Qatar Computing Research Institute, "What is AIDR?", no date. http://aidr.qcri.org/intro/
[80] IFRC, op. cit., 2013.
[81] Ibid., p. 104.

YouTube and Instagram".[82] In 2013, the system resulted in "detection of all 11 landslides reported by USGS and 31 more landslides unreported by USGS."[83]

While all of the above systems provide useful information to authorities, they may also provide useful information to members of the public. The mapping activities described above also provide information on the location of shelters, humanitarian headquarters and outposts and information about developing situations to members of the public. In addition to these, big data analytics can be used to bridge the gap between international humanitarian organisations and local people via the provision of information in local languages. In one example, an organisation called Translators without Borders was asked to make information about the crisis in Syria available to Arabic speaking audiences. This information was relevant both for those affected by the crisis directly as well as media outlets in the Arabic-speaking world, and enabled people to make an informed decision using information in their own language.[84]

Coordination and resource allocation

In addition to providing situational awareness for authorities, big data can also assist in coordinating human and other resources during a crisis to meet immediate needs. This information can be used to "accurately assess damage to people, property and the environment"[85] and to "inform the design and targeting of programmes and policies"[86]. The crisis mapping exercises described above have clear impacts for resource allocation and decision-making, as they can enable authorities to target specific areas for search and rescue work and identify the best places to set up camps for displaced people or stations for aid provision[87]. In addition, this information can enable authorities to assess damage before they are able to access specific locations. This means that they can have aid and other provisions ready once regions are accessible, rather than having to wait for access to determine levels of need. UN Global Pulse also notes that big data can provide near-to real-time feedback, which enables authorities to monitor the population and identify policies and programmes that are succeeding, failing or having undesired impacts and enable them to make adjustments accordingly.[88]

Mobilising members of the public

According to the ICRC, "disaster-affected communities today are increasingly likely to be 'digital communities' as well – that is, both generators and consumers of digital information".[89] This is obvious from the massive amounts of data being generated by members of the public in disaster situations. However, as well as gathering information from members of the public via social media, this tool can also be used to push information and share information with members of the public to aid in response during a crisis. This is

[82] Musaev, Aibek, De Wang and Calton Pu, "LITMUS: Landslide Detection by Integrating Multiple Sources", in S.R. Hiltz, M.S. Pfaff, L. Plotnick, and P.C. Shih (eds.), *Proceedings of the 11th International ISCRAM Conference*, University Park, Pennsylvania, USA, May 2014, pp. 677-86 [p. 677].
[83] Ibid.
[84] IFRC, op. cit., 2013, p. 75.
[85] Oxendine, Christopher E., Emily Schnebele, Guido Cervone, Nigel Waters, "Fusing Non-Authoritative Data to Improve Situational Awareness in Emergencies", in S.R. Hiltz, M.S. Pfaff, L. Plotnick, and P.C. Shih (eds.), op. cit., 2014, pp. 762-6 [p. 762].
[86] UN Global Pulse, op. cit., 2013, p. 39
[87] Meier, Patrick, "How UAVs Are Making a Difference in Disaster Response", *iRevolution*, 5 December 2013. http://irevolution.net/2013/12/05/uavs-in-disaster-response/
[88] UN Global Pulse, op. cit., 2013.
[89] IFRC, op. cit., 2013, p. 74.

particularly important, as the ICRC also recognizes that members of the public often act as first responders in crisis situations, well before aid or assistance is available.[90] Many of the systems described above, including early warning systems, situational awareness systems and training systems also include systems for disseminating information to members of the public, or are particularly intended as collaborative information sharing platforms. Taking crisis mapping as a specific example, the system can be set up in a matter of hours, long before humanitarian or other organizations can arrive. As such, the information can be used to enable members of the public to meet one another's needs in the gap between the incident and the official response.

Tracing long-term impacts

Finally, circling back to preparedness and training activities, big data can be used after a crisis to prepare for the next one. The section above has already demonstrated how data from previous crises is being used to analyse the incident and extract useful information for future events. However, these activities tend to focus on the next acute crisis. Data collection and processing using data from the East Japan earthquake, including displaced persons registers, documentation of experiences, health records, information on the movement of people, radiation logs and readings as well as other data is being used to inform future healthcare planning given the large population of people who may have been exposed to radiation.[91] These are often referred to as "cascading effects" of a crisis, and can persist for many years or even decades after crisis events. In addition to health, these may also implicate sectors such as taxation, public works, education, infrastructure and other areas.

These different application areas for big data in crisis informatics demonstrate two key findings. First, crisis management is cyclical and many of the data collection and processing activities that happen during one phase of crisis management feed directly into other phases. For example, information collected from those affected during the crisis is needed for post-disaster planning, information collected after disasters is useful in training for and preventing future crises and information collected outside of crises are essential for management of personnel and resources during an event. Second, there are many stakeholders that are involved in generating, collecting and processing big data during crises. These stakeholders include authorities, government agencies, humanitarian organisations, first responders, crisis managers, the media, members of the public and many others. Effective applications in big data and crisis management require the cooperation of all of these different actors. This complex ecosystem drives many of the challenges associated with big data and crisis informatics, and these are examined in the next section.

6.4.3 Big data challenges in the crisis informatics sector

The use of big data for crisis informatics results in a number of serious challenges, primarily because crises are complex, unexpected emergency situations and because much of the data generated, collected and processed comes from members of the public, often via social media. Many reports on big data in crisis informatics outline the challenges that are involved in collecting and processing such large volumes of information coming from a variety of sources and sometimes at a challenging pace. The ICRC, UN Global Pulse and OCHA have all identified the following challenges:

- Privacy, data protection, ethics and data security

[90] Ibid.

[91] This also has implications for the health of those who might be impacted by radiation. Appleby, op. cit., 2013.

- Validity / accuracy of the data
- Issues related to interpretation, bias and unequal power relations

BYTE deliverable 3.2 examining big data in crisis informatics will examine each of these challenges in more detail and in relation to a specific instance of big data practice. However, the information here provides preliminary information about these challenges taking a broad perspective.

Privacy, data protection, ethics and security

Because people caught up in crisis situations are particularly vulnerable and because standard procedures may need to be adapted during crises, privacy, data protection, ethics and security emerge as important challenges both to enable collecting accurate and relevant data, and to protect those who have been impacted by the crisis. The first issue emerge around the fact that while there are many protection measures in place to deal with traditional data sets, combining these data sets can often present additional challenges that are already well-known to big data practitioners in general. For example, many big data and crisis informatics practitioners use Twitter as their key source of social media data because the company's privacy settings make it clear that the data produced is open and can be accessible by anyone. However, research has consistently demonstrated that most users of social media applications do not read the terms of service and do not understand how their data can be accessed and used.[92] Some organisations have responded to this by arguing that the data they produce is aggregated and thus de-linked from personal information. In addition, other organisations use specific processes to de-link personal information on social media from the information generated in respect of crisis informatics. The following process by the US Geological Service, described by the Woodrow Wilson International Center for Scholars, provides an example:

> The [US] Privacy Act of 1974 establishes policies and procedures pertaining to the collection, protection, maintenance, utilization, and dissemination or federal records containing personally identifiable information (PII). On Twitter, all Tweets are linked to a username, or the unique identifier of an account holder; in some cases, this username may contain PII such as the full name of the person controlling a Twitter account. When collecting tweets for TED, USGS uses a one-way encryption technique to replace usernames with a different identifier that effectively anonymizes the sender of the Tweet. This technical solution is sufficient to comply with The Privacy Act of 1974.[93]

Nevertheless, re-identification can often be achieved using only four data points from a particular individual[94], and these solutions, while legally compliant, have raised concerns about their efficacy, particularly as big data analytics and data linking practices develop. In addition, the risks around re-identification may have specific consequences in crisis situations. Thus, Letouzé et al., note that:

> the practice of big data remains in its infancy regarding standards and guidelines. In this respect, it is clear, for example, that the "general" challenge of big data privacy

[92] Andrejevic, Mark, "Exploitation in the data mine", in Fuchs, C., Boersma, K., Anders, A. and Sandoval, M. (Eds.), *Internet and Surveillance: The Challenges of Web 2.0 and Social Media*, Routledge, London, 2012, pp. 71-88.

[93] Bowser, Anne, and Lea Shanley, "Did You Feel It?", op. cit., 2013, p. 32.

[94] de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen and Vincent D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility, *Nature Scientific Reports*, Vol. 3, No. 1376, 25 March 2013.

> can soon turn into a security risk in conflict contexts, which poses the larger question of production, dissemination, analysis, use and archival within conflict zones.[95]

In addition, many of the new technologies currently being deployed for crisis informatics have specific and serious privacy and data protection risks. For example, scholars, privacy activists and regulators have heavily critiqued the use of drones or other remotely piloted vehicles in general.[96] Using these technologies in crisis or disaster situations can compound their negative impacts, particularly because they are being deployed in situations where the population is extremely vulnerable.

Data validity and accuracy

When collecting and processing large amounts of data, the accuracy and validity of that data is of utmost importance to ensure that the decisions being based on that data will adequately meet the needs of people and organizations caught up in a crisis. While data quality is a recognised challenge for big data in general, in crisis informatics data validity and accuracy is entwined with the human-centric aspect of the sector. Specifically, collecting data from humans raises specific quality and accuracy issues, while at the same time, human computing plays an essential role in validating data collected and processed automatically by mechanical sensors.

With respect to the use of data originating from humans, often via social media, data can be incomplete, inaccurate or be unrelated to the issue in question. Specifically, relying upon crowd sourced data by ad hoc volunteers or incidental data (e.g., social media information not directly intended for a crisis management audience) can result in low-quality and incomplete data.[97] For example, people may be mistaken about locations, they may mis-read a situation or the data they produce might be only tangentially related to the issue being examined (e.g., "My bedroom reminds me of the destruction in XX"). Outside of social media, those without professional training, e.g., citizen scientists, can use measurement tools incorrectly or misunderstand the readings generated.[98] In order to mitigate such issues, some organisations or networks "seed" volunteers in specific contexts who can be mobilised in the event of an incident. However, in these scenarios, volunteer participation requires or should be optimised thorough training and organisation. This has the benefit of ensuring the right population of volunteers are involved, that they collect high-quality data in a rigorous manner and that they understand their tasks and responsibilities during a crisis situation. However, this also requires funding for training, organisation and infrastructure creation.[99]

On the other side, volunteers can be essential for interpreting data and assisting in machine learning with respect to automated processing of data. The many mapping platforms use volunteers to "score" information coming from social media in terms of accuracy and relevance. This helps automated processing algorithms to distinguish between false positive matches and true matches. In the USGS project, machine readings and automated processing resulted in a high rate of false positives that significantly undermined the accuracy of the

[95] Letouzé, et al., op cit., 2013, p. 22.

[96] Finn, Rachel and David Wright, "Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications", *Computer Law & Security Review*, Vol. 28, No. 2, 2012, pp. 184-194.

[97] Bowser, Anne, and Lea Shanley, *New visions in citizen science*, Woodrow Wilson International Center for Scholars, Washington DC, November 2013.

[98] Bowser, Anne, and Lea Shanley, "Community Collaborative Rain, Hail, and Snow Network", *New visions in citizen science*, op. cit., 2013.

[99] Ibid.

system.[100] In one example, the absence of cars parked in a large parking lot was incorrectly tagged by the system as infrastructure damage. In response, volunteers were used to score the data and reduce the overall rate of false positive alerts. Thus, while volunteers can be a source of inaccuracy, they can also be essential elements of a system that reduces inaccuracies.

Interpretation, bias and unequal power relations

Finally, relying upon technological solutions to assist in solving problems with a significant social element almost always introduces social issues around discrimination and inequality. Access to technology and the ability to use it invoke inequalities related to power and resources that may exist within societies or between one society and another. Many applications of big data in crisis informatics rely upon social media, and use of such technology raise significant issues around digital inequality. Specifically, data from social media may be incomplete as it necessarily requires Internet connectivity, digital skills and the resources to acquire hardware. On a global scale, this is disproportionately skewed against individuals who are rural, poor, elderly, female and educated to a lower level.[101] The ICRC notes that while Internet connectivity and smart phone penetration is certainly increasing in the developing world, "in many countries, men may be more likely to own the only mobile phone in the family".[102] In addition, relying on one social media source alone can also result in a sample that is biased towards particular groups of people. As Crawford notes, findings from the PEW Research Centre indicate that "only 16 percent of online adults in the United States use Twitter, and […] they skew younger and more urban than the general population."[103] Finally, linking this data with other resources may result a compounding of under-representation of particular groups as:

> Citizens, equipped with mobile digital devices and smart cards, and registered in – to name but a few systems –electoral rolls, electronic health systems (EHR), and social media networks naturally generate information that can be used for a range of different purposes, from e-government, to corporate services, to crisis management.[104]

Those who do not appear in these additional systems may be further excluded in crisis situations.

In addition to digital inequality, other biases and unequal distributions can impact crisis management. For example, Letouzé et al. note that in some contexts deliberate and targeted attempts at skewing the data can combine with digital inequality to direct resources to particular groups or interests.[105] With respect to private versus public organizations, crisis informatics stakeholders need to rely upon data being held by private organizations, especially telecom companies.[106] These organizations may be unwilling to share their data, or they may only release sub-sets of the data that can lead to particular biases or inaccurate

[100] Bowser, Anne, and Lea Shanley, "The Advanced Rapid Imaging and Analysis Project: Validating maps for disaster response", *New visions in citizen science*, op. cit., 2013.

[101] McKinsey and Company, *Offline and falling behind: Barriers to Internet adoption*, Aug 2014.

[102] IFRC, op. cit., 2013, p. 130.

[103] Crawford, Kate, op. cit., 2013.

[104] Buscher, Monika, Markus Bylund, Pedro Sanches, Leonardo Ramirez and Lisa Wood, "A New Manhattan Project? Interoperability and Ethics in Emergency Response Systems of Systems", in T. Comes, F. Fiedrich, S. Fortier, J. Geldermann and T. Müller, (Eds.), *Proceedings of the 10th International ISCRAM Conference*, Baden-Baden, Germany, May 2013 2013, 426-31 [p. 426]. See also Easton, Catherine, "The digital divide, inclusion and access for disabled people in IT Supported Emergency Response Systems: A UK and EU-based analysis", in S.R. Hiltz, M.S. Pfaff, L. Plotnick, and P.C. Shih, eds., *Proceedings of the 11th International ISCRAM Conference*, University Park, Pennsylvania, USA, May 2014, pp. 280-83.

[105] Letouzé, et al., op. cit., 2013.

[106] Ibid.

representation.[107] Finally, the use of digital technologies is also dependent on the resources available to different humanitarian and response organizations. Thus, Letouzé et al. argue, large, multi-national and national relief organizations, rather than local organizations are more likely to be able to afford big data tools; furthermore, this has a knock-on effect on development of local capacity in crisis situations, which can reinforce inequalities on a global and local scale.[108]

6.5 Big Data Definition in Smart Cities

Smart cities are complex systems of resource infrastructures such as energy, transport, and information. The many stakeholders of a smart city ecosystem, from infrastructure to service providers to end users, require a common understanding of these complexities and the potential synergies. The concept of smart cities has been developing since the 90s with every technology push to enhance efficiency and welfare, to optimize costs and resource efficiency, and to engage more effectively and actively with its smarter citizens. The last waves since 2004: Web2.0 ("the participatory Web"), Cloud Computing, Big Data Analytics, and Internet of Things can be subsumed under the greater umbrella of digitalization. Only if synergies in resource usage across all interdependent infrastructure of a city are leveraged with the aid of digitalization, can the complexities associated with big data in smart cities be addressed, i.e. potentially massive amounts of data coming from intelligent infrastructures and especially always connected end users giving way to unnecessary data storage and potential profiling.

The number of smart cities worldwide will quadruple between 2013 and 2025 according to a report from IHS Technology. In this report smart cities are described as the integration of information, communications and technology (ICT) solutions across three or more different functional areas of a city mobile and transport, energy and sustainability, physical infrastructure, governance, and safety and security[109]. ICT solutions can be narrowly defined as data, communication, management and analytics algorithms efficiently enabled through platforms. Analytics, i.e. the discovery, conveying, and usage of meaningful patterns (see Figure 8), are playing a bigger role in the recent discussions as it becomes evident, that analysed insights from big data and open data in the cities are defining the *smartness*, of decisions made by connected humans and machines:

> "Two of the biggest technological trends of the last five years -- analytics, including big-data analytics, and the Internet of Things, represented by sensors, smart meters, and even our smartphones -- are converging to reshape our urban environment drastically"[110]. Open data and sensor data, from digitalized processes as well as crowd sourced, augmented with social media data, provide the foundation for making cities smarter by enabling new services and acting as a feedback loop for improving existing services[111]. Linked data could address some of the issues of co-relating social data and open data. This is also happening at a grassroots level i.e. by empowering citizens and hackers, so-called civic hackers, to create apps and new services to solve

[107] boyd and Crawford, op. cit., 2012.

[108] Letouzé, et al., op. cit., 2013.

[109] http://press.ihs.com/press-release/design-supply-chain-media/smart-cities-rise-fourfold-number-2013-2025

[110] http://www.ubmfuturecities.com/author.asp?section_id=459&doc_id=526799

[111] http://www.opengardensblog.futuretext.com/wp-content/uploads/2012/08/Big-Data-for-Smart-cities-How-do-we-go-from-Open-Data-to-Big-Data-for-Smart-cities.pdf

a specific problem[112]. Many cities are currently co-relating Open data and Smart cities for example through hackathons[113] and for public institutions like the NHS[114], or Berlin[115] .

The city data is manifold, multidimensional: data streams in real-time with time synchronization and geo-positioning expose flows of electricity, gas, water, heating, flows of people and goods along digitized infrastructure. The dataspace will extend to spatial financial market data: housing markets and point of sales data pertaining to other kinds of consumption. Analysing and interpreting this data will require new theories about the short-term behaviour of people at a very fine spatial scale[116]. Building a digital copy of a city's infrastructures and movements therein – this will have profound consequences. If the issue is only addressed technically, as storage and computing scalability issue, or only economically, as how to create value from data, then we are missing the point that the system that can be predicted and optimized includes us, the users, the citizens.

In the following sub-section, we will list applications of big data in the smart city context, with a special focus on big data applications that do have potential for cross-optimization across different functional areas of a city. These cross-domains not only expose the immense efficiency potentials both in big data as well as smart city, but also crystallize the challenges of big data in the city that need to be carefully considered in constructing sustainable solutions for liveable cities.

[112] Anthony M. Townsend, "Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia," November 2013.

[113] http://www.sfgate.com/bayarea/article/Hackathon-aims-to-make-Oakland-more-open-3725229.php

[114] http://www.guardian.co.uk/public-leaders-network/blog/2012/jul/20/nhs-open-data-challenge

[115] http://opendataberlin.files.wordpress.com/2010/10/auswertung-online-voting-zu-open-data.pdf

[116] Reades J (2013) Big data's little secrets: part 1. Placetique: People, Data, Place, http://www.reades.com/2013/05/ 31/big-data-little-secret

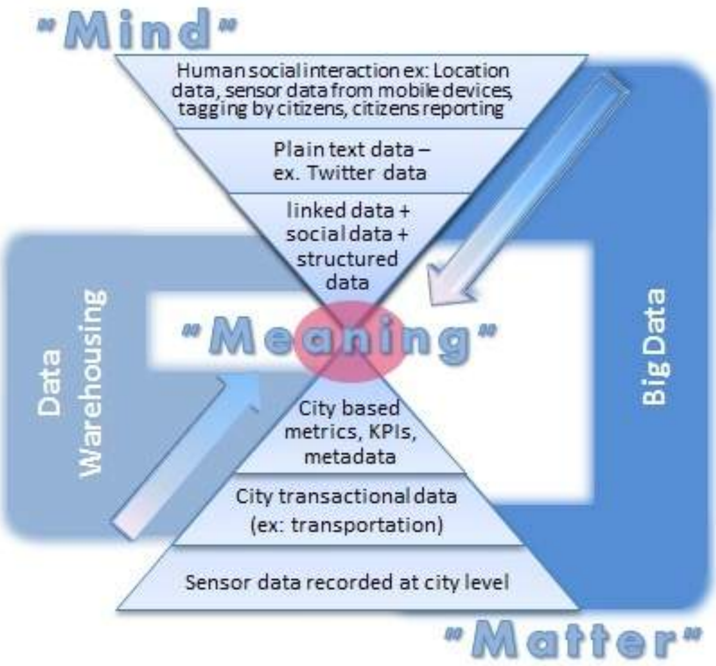

"Mind"
Human social interaction ex: Location data, sensor data from mobile devices, tagging by citizens, citizens reporting
Plain text data – ex. Twitter data
linked data + social data + structured data
"Meaning"
City based metrics, KPIs, metadata
City transactional data (ex: transportation)
Sensor data recorded at city level
"Matter"
Data Warehousing
Big Data

Figure 8 Overview of generating meaning[117] via analytics (Data Warehousing vs. Big Data) in the context of a smart city.

6.5.1 Big Data Applications in the Smart Cities Sector

The problems to be solved will be complex and not known in advance: this delivers the underlying motivation to think in terms of big data.

City planning will move from longer term strategic planning to short-term operational optimization. For example, transport or utility planning has always been concerned with peak daily flows, but these are assumed to pertain to a much longer period. The theory and planning has been focussed on what happens to cities over planning horizons that relate to years the short term being 5 years and the long term 20 or 50 years. This is significantly changing through big data technology which enables real-time insights and decisions or actions. Big data applications in smarter cities can be characterized as a new top-down discipline that is more 'open' to bottom-up responses from a multiple range of actors. It also looks to limit choice but still allow infinite possibilities. It is therefore, by its nature, freedom within constraints.[118]

Digitalized "Physical" Infrastructure

Energy, Mobility, and Information networks make up the digitalizing physical infrastructure of smarter cities. Situational awareness on the multimodality and cross-optimization or -utilization is a secondary trend, which is allowed by the convergence of technological advancements and platforms. Each entity or mode in these infrastructures, can be seen as a big data application in itself: a solution that makes use of big data and adds to the body of data.

[117] http://www.opengardensblog.futuretext.com/archives/2012/08/big-data-for-smart-cities-%E2%80%93-for-hackers-data-scientists-and-citizens.html

[118] http://engagingcities.com/post/5012064472/massive-small-the-operating-system-for-smart-urbanism

Smarter mobility: Traffic can be reduced with dynamic road pricing or smart parking: systems that detect where the nearest available parking slot is[119], i.e., providing timely information to locate parking slot quickly in order to save time and fuel. Available traffic resources can be utilized more efficiently and/or more personalized by utilizing all modes available in a customizable, real-time adaptive manner: Origin and destination information is transformed into a real-time multi-layered map of the different means of transport sorted by the individuals real-time preferences across multiple dimensions such as environmental friendliness, time, or mood (matching the weather or the music the person is currently listening to via spotify[120]).

Smarter energy: Current technology allows generation and consumption of energy to be optimized not only on the level of buildings (i.e. prosumers, e.g. a household capable of generating power via distributed energy resources, covering its own consumption need at times or exceeding it and hence becoming a producer) but also across the different modes energy is being utilizes, such as gas, water, heating, and electricity – both locally and system-wide. Acquiring, time-synchronizing this information via smart metering of energy, making it available via platforms, and utilizing it in new energy-related services is not necessarily confined to the borders of a city – however, again, the cross-optimization potentials across the different modes can be today best taken advantage of by local municipal utilities who still do own all or many of the energy modes. Of course through platforms and ICT the cross-optimization should be feasible across competing organizations as well. But this will still take time.

In a smart city, another cross-optimization potential through big data applications becomes feasible: the cross-optimization of energy and mobility, e.g. in scenarios like energy-efficient city logistics: In energy *efficient city logistics* the flow of goods, vehicles, and electricity is forecast, optimized, monitored, and controlled both long-term and in real-time. City hubs and logistics consolidation nodes play an important role for coordination, as they are where goods are stored and sorted for the following optimized in-city distribution. These logistics consolidation hubs offer options for energy efficiency, through cooling and heating, water management, as well as electricity optimization. There are already energy demand response service providers targeting these logistics centres to become part of so-called “energy saving fleet,” which then offer flexibility (i.e. saved energy when power supply is low) to local utilities. Together with fleets of smart electric trucks the margin of efficiency increase is of course much higher per “user.” An integrated positioning and on-trip vehicle re-routing based on recent traffic, order, and weather data leads to improved transparency in fleet management. The utilization of electric vehicles, energy management system, dynamic vehicle routing, order management, and electric mobile city hubs can reduce the environmental impact of commercial traffic in urban areas. Integration of an order management, for logistic processes, hubs and service models enables cities, logistics service providers, and logistics hubs as well as major internet retailers to save resources and gain efficiency, enable better quality of life in cities.

Participatory Sensing

Smart Santander project[121], a public-private partnership, is placing sensors around various European cities in order to gather data, as well as take advantage of what citizens are willing

[119] http://www.scielo.org.mx/pdf/jart/v11n5/v11n5a11.pdf

[120] http://apps.moodagent.com/spotify

[121] http://www.smartsantander.eu/

to contribute through their smartphones and specialized software[122]. The Smart Santander project includes 'participatory sensing.' The participatory sensing service aims at exploiting the use of citizens' smartphones to make people to become systematic observers and contributors of data. It takes advantage of the ability of these devices to be connected to people as well as to the core network. Data analytics, IoT, and some aspects of social media is blended so that problems are found in real time and conveyed back those who can fix them. For example a malfunctioning streetlamp is reported to all the users that have previously subscribed to this type of event via their mobile phones. One of these users could be a municipality technician that after receiving the notification can log in a repair job. The result would be a well-functioning city made smarter by the people living in it, sharing their data.

Linked City

Dublinked[123] is an innovative new data-sharing network. The network is seeking to link data, sectors, skills and people to generate new commercial opportunities for the Dublin Region. Dublinked will also provide the Dublin Region's first Open Data Platform which makes public data available for research and reuse. The city is explicitly utilizing data as a resource to invite new data-driven economy actors: "The most valuable data is often valuable because someone invested a lot of money in collecting it. These forms of data are often the most valuable for the creation of new services, and Dublinked is focussed on making this data as accessible to as many people and companies as possible for the purposes of research and development. This creation of a pool of high-value data for research purposes, in combination with all the normal "open" data, is unique in the world and will give Dublin-based companies a significant advantage."[124]

Public Safety

Open data and cloud-based big data analytics can be used to improve the efficiency of police and fire services by capturing and correlating the analysis of all data coming from various systems installed in the city, including surveillance cameras, emergency vehicle GPS tracking, and fire and smoke sensors[125]. Predictive policing uses historical crime data to automatically discover trends and patterns in the data. Such patterns help in gaining insights into crime related problems a city is facing and allow a more effective and efficient deployment of mobile forces[126] and significant decrease in crime.

Big Data Challenges in the Smart Cities Sector

The predictive policing is a good example for showing both sides of big data: the promise and the peril, as depicted popularly in the movie Minority Report. The good and the bad is easily conveyed to the majority of citizens, without the need for technology foresight: it is great to decrease number of crimes by predicting, but it does diminish the primitive that a person is innocent unless proven otherwise.

Additionally, in conclusion, one can say, that the main challenges are not related to big data but to trust and new rules and regulations for data access, usage, and sharing, as well as the mind-set that big data businesses are being carried out in ecosystems not by individual companies.

[122] http://www.ubmfuturecities.com/author.asp?section_id=459&doc_id=526800

[123] http://www.dublinked.ie/

[124] http://www.dublinked.com/?q=aboutus

[125] http://www.accenture.com/us-en/blogs/technology-blog/archive/2015/02/11/iiot-and-big-data-analytics-help-smart-city-development.aspx

[126] http://www.predpol.com/

User Acceptance, Privacy & Confidentiality Concerns

No infrastructure – whether it is a road, a building, a broadband network or an intelligent energy grid – will have a transformative effect on a city unless it engages with individuals in a way that results in a change of behaviour. One strain of big data analytics application is to implement dynamic pricing (road, water, electricity) based upon real-time demand data. In many cases this meets the resistance from citizens, and needs a very thorough communication strategy and involvement of users.

Maintaining the privacy and security of the data being collected is also a very important challenge. Different stakeholders need to be allowed access to different portions of the data being stored and collected and this security must be maintained at all levels of the network. The data must also be anonymized sufficiently so that the customers cannot be individually identified even after the data analysis. This is very difficult as the following example shows[127]: "Only four spatio-temporal points, approximate places and times, are enough to uniquely identify 95% of 1.5M people in a mobility database. The study further states that these constraints hold even when the resolution of the dataset is low, mobility datasets and metadata circumvent anonymity."

Cost-effectiveness Better Models vs. More Data

All one needs to look for in big data, so the argument goes, are more and more correlations vs. what we need to look for in big data can only ever be discovered through the lens of theory[128]. However, if the analytics should also give insights about what actions need to be undertaken (i.e. prescriptive analytics) then theory of system behaviour is absolutely necessary. At the same time the system may be changing in different ways than initial models can predict, hence a real-time monitoring of both data and model is necessary to capture so-called concept drift.

Model- and data-driven analytics is at the core of the required smart data infrastructure as opposed to the purely data-driven approach of big data. Data-driven analytics, e.g. data mining and machine learning, can be is used to reveal characteristics of the systems not known before or to learn solely based on the data, when the programming of rule-based algorithms are infeasible. Data-driven analytics is required when dealing with data-rich but theory-poor domains such as online communities and neuroscience.

The city, however, is a planned, constructed, and engineered system, consisting of increasingly digitized physical infrastructure. Models based on physical laws such as the flow network models use known external knowledge of the physical processes. At the same time, in today's complex systems and increasing dynamics through liberalized economic transactions, end user participation with their shared resources – e.g. cars to provide transportation, or PV installation to provide energy – numerical analysis to solve these models becomes very hard.

The digitization also extracts digital copies of domain know-how entered by domain or planning experts using software tools, e.g. how a distribution network for electricity is setup. Finally, the digitization of infrastructure not only enables combining domain models, e.g. the concrete topology implementation, with physical models, multimodal flow networks to

[127] de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel (2013, March 25): "Unique in the Crowd: The privacy bounds of human mobility". Nature srep. doi:10.1038/srep01376

[128] http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

explain the system with static but hard facts, but also to analyse real-time data coming from that infrastructure to discover unknown facts caused by stochastic and behavioural processes such as end user participation.

Finally, the power lies in semantically capturing the existing knowledge as well as the knowledge discovered from model- and data-driven analytics. This continuous semantic knowledge modelling allows continuous model improvement through real-time and historical data. Real-time data, thus, is not only used for determining when to take corrective actions according to the prescriptive analytics but also to improve models and the precision of the prescribed actions. As such with model- and data-driven analytics, more data leads to better models, and better models lead to smarter data – enabling actionable knowledge without invading privacy or compromising confidentiality.

These are all new frontiers, which will require years of research before producing feasible answers.

City-wide Information Infrastructure & Investments

FI-WARE "future internet platform" project, addressed this topic and identified the specific challenges that local innovators need help to overcome, and that could be provided by city information infrastructures. The challenges included: real-time access to information from physical city infrastructures; tools for analysing "big data"; and access to technologies to ensure privacy and trust[129].

Investment into ICT infrastructure is still the biggest question: Whether the above applications are profitable is an open question. The stakeholders are still assessing and investigating pilots instead of investing. For example, while there are a growing numbers of mobile ticketing success stories for public transportation, retrofitting stations with the ability to accept payments from smartphones is a significant investment for transit agencies[130].

Although there are many benefits to energy efficient electric multi-modal city logistics, the cases are supported by external funding. But if freight companies, store owners and municipality join forces, there should be no reason for why this successful and popular solution could support its own.

The addition of sensors and data collection and transmission infrastructure adds some additional maintenance requirements to the system. These costs, if not offset by the benefit provided by the advanced diagnostics, could make the project impractical to implement.

Variety of Data & Data Sharing

Transport for London (TfL) has very detailed data on buses and trains that give us precise geo-positioning, times and delays with respect to timetables. In principle, these can be matched against the travellers using the Oyster card. These demand and supply data sets, however, are entirely incompatible[131]. This vision of big data enabled urban mobility hinges upon transportation providers — both public and private — sharing data, collaborating, and supporting innovation. In the existing cases challenges in data sharing results in inability to plan a single trip that uses multiple modes, which is important both in personal mobility as well as city logistics.

[129] http://theurbantechnologist.com/2012/08/13/the-amazing-heart-of-a-smarter-city-the-innovation-boundary/
[130] http://www.citylab.com/commute/2014/04/true-future-transportation-has-two-big-barriers-entry/8933/
[131] http://www.complexcity.info/files/2013/12/BATTY-DHG-2013.pdf

Hence, central to the development of an intelligent energy efficient transportation system in cities for both people and goods, is open data, that is compatible yet feasible to manage in a distributed loosely coupled way. To understand and monitor the complete network of available transportation modes – and energy modes –, each modal provider must be willing to make their data available to an aggregator or to develop open apis for open (linked) data. The closer that data is to real time the better it is for the system operations.

Many private-sector service providers see a competitive advantage in keeping data proprietary[132]. This may be the case in certain situations, but there are many benefits associated with a private company opening its data: these include increasing system efficiency, expanding the market of users, fostering innovation ancillary to the service, and other benefits associated with transparency (e.g., emission and fuel-use reductions).

Skills & Socio-demographic Silos

We need a different skill set than that of a data scientist for cities[133] i.e. someone who understands city services, data analysis, co-relation of data etc. This is similar to how digitalization of healthcare and the need for specialized information science resulted in bioinformatics. The analytics provided through big data allows for more informed system planning, however, without employees capable of using the new information properly, the benefits of using big data can be lost.

The "skills shift" is a phenomenon also encountered in the smart city scenarios: The skilled workers are currently retiring in waves. This enforces the trend of digitization and automation to model expert systems and model knowledge into the infrastructures. Hence, a lot of the new jobs that are being created along this trend are in the knowledge and information intensive segments. However, there is also a "skills mismatch" that the EU struggles to fill these ICT jobs.

The dependability of many of the big data applications on smart phones highlights the establishing of socio-demographic cohorts, the digital and socio-demographic divide in the world[134]: e.g. Of U.S. residents who are making less than $30,000 per year, less than half own a smartphone. Conversely, over 78 percent of people earning $75,000 or more own one. Educational attainment and age show a similar correlation, with older and less educated people less likely to own a smartphone than younger, more educated individuals. In Europe the numbers must be similar.

6.6 Big Data Definition in Shipping

The shipping industry is a very diverse sector and it often includes actors from different corners of the world. There is no clear cut boundary/definition of shipping industry. At the centre are, of course, sectors such as ship owner, ship operator, ship yards, naval authorities (national & international), class societies, port authorities, naval academies, etc.; which exclusively serve shipping whereas other actors may also serve other industries such as machinery providers, equipment producers, travel agencies, etc. Therefore, there will be no

[132] http://mappable.info/one-week-of-carsharing/ was forced to take data from car2go offline.

[133] http://www.opengardensblog.futuretext.com/archives/2012/08/big-data-for-smart-cities-%E2%80%93-for-hackers-data-scientists-and-citizens.html

[134] http://www.citylab.com/commute/2014/04/true-future-transportation-has-two-big-barriers-entry/8933/

crisp common understanding what is meant by big data in the shipping sector but at most a rather fuzzy comprehension of this term.

The sector at large is known as a very slow adopter of new technological solutions. The main reason lies in the simple fact that the primary focus of their business model is on earning money through profitable transport deals and less on cutting costs. Introduction of new technologies is to a large extent about doing things faster, better, cheaper, which is mainly about cutting cost and will therefore get secondary priority. In other words, there is reluctance in adopting new technology unless they have to. In shipping it has become very clear that authorities have a central and importing role in the introduction of new technologies by forcing ship owner through legislation. It must be pointed out that the authorities' sole focus is on technologies increasing safety and environmental protection.

Some of the current changes in the maritime industries seem to, by the new data reality and the increased connectedness of ships and players in the maritime industry (more satellites and reduced prices enable 24/7 ship-to-shore connection).

To our understanding a good definition of big data in Shipping can be formulated as following: data with *high-volume, -variety, -velocity, -veracity and -value information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.*

Big data often means collections of data sources that produce data in such a speed and volume that the capturing, the storage and the analysis exceed the limitations of traditional solutions, such as standard database management systems. If novel forms of data management are successfully applied, new information, facts, relationships, indicators and trends can be extracted from the vast amount of data entries. The ability of effectively manage information and derive knowledge of it is now seen as a key competitive advantage.

A classic definition of big data identifies three to five key characteristics (the ***5Vs***):

- *Volume*: the amount of data that needs to be ingested and analysed. Typical data set sizes today vary between the scale of a few terabytes to hundreds of petabytes or even exabytes. However, this is dependent on industry and reflects its maturity.
- *Variety*: the different types of data to be combined or integrated, for deriving information. New, earlier hidden, unexplored, or undiscoverable information from big data are often obtained today by effectively combining different data types.
- *Velocity*: how fast data is being produced, changed and the speed with which data needs to be transferred, recorded, and processed. Typical high speed figures here are today around gigabits per second and teraflops concerning transmission and processing speeds, respectively.
- *Veracity:* It refers to the quality, provenance and trust of the data. For deriving knowledge out of volumes of data, the accuracy of the data sources (or even of the data entries) needs to be evaluated as well.
- *Value:* potential gain for an organisation when exploiting the data

6.6.1 Big Data Applications in the Shipping Sector

There have been a number of applications where data driven services have implemented within the following main areas of the shipping industry;

- Technical operation and maintenance
- Energy efficiency (cost and environment)
- Safety performance
- Management and monitoring of accident and environmental risks from shipping traffic
- Commercial operation (as part of a logistics chain)
- Automation of ship operation (long-term development)

Technical operation and maintenance
Remote operation and maintenance has been enabled by increased use of sensors in components and systems on-board ships. It has allowed implementing improved monitoring component; this allowed vendors implement advanced analytics for conditions of systems and advise ship management on system operation and predictive maintenance.

This may eventually lead to new business models where one can lease equipment and systems or subscribe to specific functions performed by the system and components as is occurring in aviation. Some examples are as follows:

- Electronic Power Design Inc., which provides diesel electric propulsion for some of the most sophisticated DP3 vessels in the world;
- Rolls Royce Hemos System, a system that draws on their system in aviation, for transmitting sensor data from various components to land-based service centres where system specialists create health asset reports for the customers;
- Wärtsilä's Propulsion Condition Monitoring Service, which enables detection of maintenance requirements some 2–6 months in advance.
- Engineering Software Reliability Group (ESRG) that has built a commercial cross-silo analytic platform for ships, the OstiaEdge® Monitoring Suite for real-time analytics. This is derived on past engagements with the US Navy, capturing and analysing more than 5000 data points for more than 120 USNavy ships.

Energy efficiency and environmental performance
Improving fuel efficiency of ships is one of the easiest ways of reducing exploitation costs for ships; consequently, many ship owners have try using big data solutions and advanced analytics. For example, Maersk has created a proprietary solution for collecting, analysing, and presenting data from all the vessels in their fleet to drive continuous energy efficiency improvement.

Another driver for adopting technologies specific for big data are new environmental requirements for emission control by 2018. Currently, EU is establishing guidelines for MRV reporting and has initiated projects to develop systems to manage such reporting. These initiatives are supported by national associations, such as the Norwegian Ship-owners' Association, are encouraging and enabling their members to document emissions and prepare for the new legislation. Concomitantly, retailers, such as IKEA, require that shipping companies document their emissions for their product footprint documentation. It is likely to be a boost in availability of data and adoption of big data solutions by all actors in the shipping industry.

The market is composed from:

- Niche vendors like Marorca, SRG group (offering energy efficiency optimisation strategies) and Maersk (offering a version of their own internal system in the market).
- Class societies like ABS and DNV GL have been well established in this field for a long time with traditional advisory services founded in their in depth knowledge and competence and class customer base. The class societies are now actively seeking to move this to a more analytically based platform. Through the acquisition of a PMS vendor ABS has attained a position on board the vessels for data capture and analysis of performance (ABS Nautical Systems). Societies like DNV GL and NK have developed on-board reporting solutions from scratch or together with partners like NAPA.
- Component-manufacturers/system integrators for example Rolls Royce, Wärtsilä and ESRG, provide offerings to providing energy efficiency advice and intelligent operations.

Safety performance

Class societies have as main activity domain the safety, performance, verification and assessment of ships; through their role as standard setters, delegated body and advisory offerings.

These offerings are transitioning from a traditional approach based on empirical and analytical rules and *ad hoc* processing of disparate datasets towards a more data-supported offering for assessing and monitoring ship safety. Availability of data from a multitude of sources on-board ships may become available for safety assessment activities on a continuous basis.

Regulatory bodies, for example EU pushes for increased transparency of safety of shipping, by requiring disclosure of more data and information directly from the ships. There has been established an initiative for developing methodologies for improving existing risk management procedures and processes for inspections, incident detection and recording, compliance monitoring, contingency plans and emergency responses. This requires dynamic collection, processing and use of real-time information from ships. System vendors of on-board systems for operational support push services such as weather routing, or hull monitoring.

Some actors in the market are:

- RightShip has offered safety-rating services that are based on available data outside shipowners’ organizations (vetting, port state control, class, ship registration data, etc.).
- Lloyd’s List Intelligence, offer advisory services founded in similar datasets.
- Ocean Intelligence, are more concerned with financial risk aspects, use similar datasets as input to, for example, debt, and credit analyses.
- Shipping KPI is another major initiative that provides independent safety assessments and performance management schemes for the shipping industry based on data reported from the ships.

Management and monitoring of accidents and environmental risks from shipping traffic

Ship tracking data emerges as a platform for many new services. Examples of such applications could include continuous monitoring of emissions from shipping, and real-time monitoring of accident and emergency response risks (oil recovery, tug preparedness, pilot schemes etc.). Operational and navigational risk monitoring is one of the problems addressed in the European Maritime Safety Agency (EMSA) EU initiative.

Busy ports and shipping lanes have used ship-tracking information to implement real-time traffic control centres to avoid collisions and accidents; we expect that similar approaches will be adopted other places also. These solutions will expand to include more complex risk models, enriched with data from multiple sources, including data from the ships themselves. Highly important will be to combine real-time information about course, speed etc., with facts concerning the safety condition and the weather, which will allow identifying those areas most exposed to risk accident.

In the next years, the global satellite network exactEarth will cover all the oceans and the poles. This will improve the reliability and the availability of ship tracking data, which will change the way services are delivered, including monitoring the safety of ships based partly on knowledge of weather conditions to which the ship has been exposed at any time.

Commercial operation (as part of a logistics chain) of ships and fleets

Shipping companies have begun using analytics as part of the means to optimise their place in the value chains, and to optimise their own operations. Combining and analysing data about the availability of cargoes, space for cargoes, port slots, weather, ship performance, fuel prices, etc. could result in enormous business cases.

Automation of ship operations

Ships will become more automated, and, in the long-term, autonomous. Transmission of increasing amounts of sensor data to shore will require onshore operations centres. This trend will challenge regulatory aspects, as well as safety performance assessment and verification processes and roles.

Big Data Challenges in the Shipping Sector

There are several challenges in the shipping sector. Some of the most important are extraction of business-critical intelligence and insights from diverse data sources with different availability, in a complex environment of legacy diverse systems and fragmented and decentralized solutions that are common in the Shipping sector.

Like generic big data, the Shipping Data is also characterized by the 5V:

Table 10 5V in Shipping

Volume	- High resolution AIS - Engine and hull vibration - Note: most data sources are considered small volumes
Variety	- Structured: sub-systems measurements - Unstructured: video, audio - Semi-structured: voyage optimization

Velocity	- Real-time AIS
Veracity	- Unreliable data sources - Unverifiable data from measurements due to suppliers competition
Value	- Reduce costs of exploitation, fuel costs in general

Generally, Shipping companies are concerned with challenges associated with reducing exploitation cost of vessels and fleets. The industrial landscape is very wide and fragmented, often involved actors have conflicting interests.

The common challenges and possible approach to tackle these challenges are listed in the Table 11.

Table 11 Big Data Challenges in the Shipping Industry

Challenges	Approach
Data from different sources (structured, unstructured & real-time)	Need for low cost and low maintenance solutions such as RDBS, Hadoop, NoSQL databases for scalable information management systems in batch. Need for opportunistic and delay tolerant streaming to fulfill need for heterogeneous infrastructures.
Conflicting interests	Improved international legislation and regulatory support, with public funds support for adoption of new technologies for all actors involved with a high focus on the end user.
Advanced decision support to enable use of data to quickly and efficiently respond to current situation	Enlarge education focus during exploitation of vessels to include advanced monitoring systems.
Remote operations and maintenance	Implement autonomous safety systems that can provide adaptive safe states for the systems. Changes in the legislation to create a viable support for implementing the necessary mechanisms. Including safety related requirements.

6.7 BIG DATA DEFINITION IN CULTURE

This section examines the meaning of big 'cultural' data, both in terms of reference to the BYTE project case study on big cultural data, and with a broader view that encompasses big data applications in the cultural sector. Understanding big cultural data will also assist in developing a meaningful and cross-sector definition of big data as one of the outputs of the

BYTE project. A definition of big cultural data that encompasses both the digitisation of works and their metadata, and the data that is generated by applying big data applications to the cultural sector to generate data for commercial use is pertinent as big cultural data matures.

The cultural sector is facing new ways of disseminating cultural works, including literature, manuscripts, sound recordings and a variety of images. Big cultural data can thus refer to the digitisation of private and public collections of such works and their associated metadata. This interpretation bears much relevance to the BYTE project case study for big cultural data, which is focussed on the Europeana.[135] For example, Europeana deals largely with open linked metadata for digitised copies of works, such as text, images, audio, manuscripts etc., held by cultural heritage organisations, including national libraries museums and archives. Beyond this understanding of big cultural data, it can also extend to encompass big data applications in the cultural sector. These include, but are not limited to, social media data relating to culture and cultural events, as well as cultural user behaviour and sentiment data. However, there does not appear to be a universally accepted definition of big cultural data, and the application of big data practices within the cultural sector is very much in its infancy. It follows that big cultural data as a means of creating value in a traditional economic sense has, to some extent, been overlooked, and there is limited literature about big cultural data per se and/ or big data applications in the cultural sector that could perhaps generate value in the commercial sense. Lilley observes,

> The current approach to the use of data in the cultural sector is out-of-date and inadequate. The sector as a whole and the policy and regulatory bodies that oversee it are already failing to make the most of the considerable financial and operational benefits which could arise from better use of data

Instead, the value of big cultural data lies in its cultural and social contributions to society. This approach also reiterates the difficulty in understanding big cultural data as well as providing insight into prospective approaches to applying big data practices in the cultural sector in a way that could see commercial value derived from these data more in line with approaches in other industries and sectors.

6.7.1 Defining big cultural data

In light of the underdeveloped nature of big cultural data generally, there is license to view big cultural data in a narrow sense, namely the digitisation of cultural works and their associated metadata, or more broadly, by making reference to big data applications within the cultural sector that source social media traces and other data relating to culture and cultural events in societies. As there remains demand by educational and cultural institutions, as well as members of the public, for access to these digitised works, big cultural data will inevitably mature. The European Commission, in particular, is committed to supporting the growth and popularity of European culture by supporting the digitisation of cultural records.[136] In that context, cultural data has been referred to as including: “statistical and economical cultural

[135] Europeana, “About”, no date. http://www.europeana.eu/portal/aboutus.html

[136] See for example, European Commission, Cultural Heritage Digitisation, Online Accessibility and Digital preservation, Report on the Implementation of Commission Recommendation 2011/711/EU, 2011-2013.

data, metadata, visual files from public domain works, etc.".[137] In addition, Manovich recognises the importance of understanding big cultural data as it relates to digital humanities to involve, "Analysing massive amounts of cultural content and peoples' conversations, opinions, and cultural activities online – personal and professional websites, general and specialized social media networks and sites."[138] This broader view of what cultural data entails is more appropriate than the narrow view, which refers to the digitisation of works and their metadata, as the field of big cultural data expands.

When understanding what is meant by big cultural data, we can also consider the extent to which big data in the cultural sector contends with the 5Vs definition adopted by this Deliverable, which is an extension of Gartner's foundational meaning of big data. The 5Vs include: *Volume; Variety; Velocity; Veracity;* and *Value*. These Vs can be met by either stand-alone collections of cultural data held by cultural heritage institutions and organisations (or sizeable private collections if and where in existence) or the linking and aggregation of these data to form larger datasets. The 5Vs as they contend with cultural data are as follows:

- *Volume* can be indicated by: massive datasets from aggregating cultural metadata; or large datasets of metadata of cultural items available at cultural heritage institutions (museums, libraries, galleries) and organisations.
- *Variety* can be indicated by: quantitative data, e.g. cataloguing of metadata and indexed cultural datasets; qualitative data, e.g. text documents, sound recordings, manuscripts, images across a number of European and international cultures and societies; and transactional data, e.g. records of use and access of cultural data items.
- *Velocity* can be indicated by: monitoring user behavioural and sentiment data, social media traces, and access rates of cultural data etc.
- *Veracity* can be indicated by: improved data quality; and a combination of cultural data or user data from within the cultural sector at large.
- *Value* can be indicated by: knowledge creation from the access and potential re-use of digitised cultural items; improved access to metadata and data, e.g. historical texts; and improving efficiency for students, researchers and citizens wishing to access the data and reducing overall operational of cultural institutions and organisations.

The above indicators of the 5Vs are not an exhaustive list and as big cultural data measures, so too will the applications of big data to the cultural sector as prospective determinants of the value of big cultural data. Furthermore, whether we view big cultural data in the narrow sense or interpret it as all-encompassing of culturally-related data from within the sector, the extent to which is it is 'big' is variable. It is also relevant that that the immaturity of the big cultural data sector is indicated by the existence of few genuinely large data sets.[139]

Having made reference the immaturity of big cultural data, it is important to consider why this is so and indeed the prospects of the sector in order to fully appreciate how and why big cultural data has not been universally defined or developed. First, the cultural sector inhabits

[137] This reference to 'cultural data' was made with reference to a partnership between the ministry of Communication and the Open Knowledge Foundation when referring to the development of a public domain calculator, which was developed with the assistance of the French National Library: European Commission, Cultural Heritage Digitisation, Online Accessibility and Digital preservation, Report on the Implementation of Commission Recommendation 2011/711/EU, 2011-2013, p. 27.

[138] Manovich, Lev, "How and Why Study Big Cultural Data", *Slidesharenet*, 9 March 2012. http://www.slideshare.net/formalist/how-and-why-study-big-cultural-data

[139] Lilley, Anthony, "What Can Big Data Do for the Cultural Sector? An article Exploring the characteristics and Potential of Big Data for the Cultural Sector", *Audience Finder,* no date.

the public sector, and the process of digitising works is carried out largely by public sector institutions and organisations. This means that these processes are subject to policy and funding restrictions. Second and again related to the public positioning of the cultural sector, there is a strong focus on deriving cultural and social value from the cultural data rather than monetising these data or applying big data applications to generate profit in a commercial sense. Thus, big cultural data is currently understood as a publicly funded investment in culture creation and preservation. As such, the value in big cultural data is difficult to assess in an economical sense, the method which is most commonly used to indicate the value of data in other sectors.[140] Thus, as is the case with cultural data (big or other), "there exists a lag in its understanding due to an apparent conflict between the value of big cultural data being tangible only to the extent that it perpetuates and encourages culture, rather than in commercial, monetary terms."[141] What this means is that the social value of big cultural data is still viewed as the appropriate measurement tool. Lilley affirms that big cultural data "can contribute social capital and cultural value creation to digital societies." Social capital was defined by the originator of the term, Hanifan, in 1916 when he remarked, "I do not refer to real estate, or to personal property or to cold cash, but rather to that in life which tends to make these tangible substances count for most in the daily lives of people, namely, goodwill, fellowship, mutual sympathy and social intercourse among a group of individuals and families who make up a social unit…"[142] However, Lilley observes that despite these longstanding notions of cultural value and social capital, cultural data can also offer, "an opportunity to begin to measure both the economic benefits and those such as the formation and reinforcement of social and cultural capital which also arise from cultural activity."[143]

Importantly, what this means is that big data, and the possibilities it presents, has implications for culture and the arts. Manovich suggests that big cultural data, "offers unprecedented opportunities to understand cultural processes and their dynamics and develop new concepts and models which can be also used to better understand the past."[144] Lilley also opines, "Assuming that the foundation of the raw materials is strong enough, the analytics robust and the people in the room willing to listen; data-driven decision-making could and perhaps should be a key element of increasing artistic impact and commercial resilience both for individual organisations and for the sector as a whole."[145] In that same paper, it is suggested that sentiment or semantic analyses to measure aspects of artistic impact could also be an important new tool for the cultural sector.[146] Thus, ultimately, studying big cultural data can assist by moving from (incomplete) knowledge to actual cultural data.[147] This recognised potential can also involve collaborative prospects between the public and private sectors to achieve the wide application of big data practices in the cultural sector. Whilst limited in number, current and prospective applications of big data in the cultural sector can positively impact upon society as addressed below.

[140] Ordinarily, value is measured through ways in which data can add by value, such as the five ways identified by McKinsey: segmenting audiences to customise activity; creating transparency; supporting/ replacing human decisions/ enabling experimentation; and innovating new business models and services: Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers, "Big Data: The Next Frontier for Innovation Competition, and Productivity", The McKinsey Global Institute, 2011, p.5.

[141] Lilley and Moore, op. cit., 2013, p.23.

[142] Ibid., p.36.

[143] Ibid.

[144] Manovich, op. cit., 9 March 2012.

[145] Lilley, op. cit., no date.

[146] Ibid.

[147] Manovich, op. cit., 2012.

6.7.2 *Big data applications in the cultural sector*

The following are some of the current and potential applications of big date in the cultural sector:

- Broader dissemination of knowledge and culture
- Access to greater cultural resources
- Improved research and development for students, academics, practitioners and citizens
- User behaviour and sentiment monitoring
- Transparency of public institutions

Broader dissemination of knowledge and culture

Understanding cultures and history depends on the ability to access information about them. Big cultural data in the sense that cultural works are digitised and thus accessible in that manner can enable efficient and widespread access to cultural data, as well as providing information about where to locate and access additional resources or provide access via linked data and metadata. Thus, the combination of metadata and linked metadata can increase the value of cultural data simply by making it accessible. Further, the aggregation of all existing metadata enables access to a more complete picture of the subject matter. By facilitating access to big cultural data in a way that supports re-use and sharing of works and their dissemination, new discoveries, views and interpretations become possible.

Access to greater resources

Virtual access to big cultural data enables greater access to resources. In the public sector, initiatives such as Europeana are making cultural data open and accessible to all internet users and in turn, adding cultural and social value to the digital economy through virtual access to millions of items from a range of Europe's leading galleries, libraries, archives and museums. Another good example of re-use of cultural data is the CHContext widget developed by PSNC and released as open source.[148] This JavaScript-based widget is,

> able to provide links to cultural heritage materials (from Europeana or Digital Public Library of America or Polish Digital Library Federation) based on predefined item of a website on which it is embedded (via given JQuery HTML selector). The widget can be used by anyone who has a website, but it may be especially valuable for cultural heritage institutions which would like to enrich their online catalogues or websites with links to Europeana. It may also be useful for bloggers who are writing about culture and related topics.[149]

Outside of Europe, a number of players in digital humanities are working with digitised historical cultural archives, which were created by libraries and universities with funding from NEH and other institutions.[150] Big data practices are enabling efficient access to resources in these areas by supporting a variety of access options, including linked and open access. A McKinsey Global report provides, "Simply making big data more easily accessible to relevant stakeholders in a timely manner can create tremendous value."[151]

[148] "Git Hub", no date. https://github.com/psnc-dl/chcontext

[149] "CH Context Widget", *Europeana Labs*, no date. http://labs.europeana.eu/apps/ch-context-widget/

[150] Manovich, op. cit., 2012.

[151] Manyika, James, et al., op.cit., 2011, p.5.

Improved research and knowledge for students, academics and citizens

Access to larger datasets of cultural data (including metadata), especially by way of linked data, can greatly improve research activities. This in turn fosters the development of new perspectives and contributions to discourse not previously made. This can also result in new insights into cultural data. Studying big cultural data in the broader sense that also includes data about societies through the social media traces using social computing and by studying society itself, produces a more inclusive understanding of history and the present.[152] This is especially so when using much larger samples that also enable stakeholders to map cultural variability and diversity. Large-scale cultural patterns can also be detected, as well it being the "best way" to follow global professionally produced digital culture and understand new developed cultural fields.[153]

User behaviour and sentiment monitoring and

User behaviour and sentiment data encourage thoughts about user participation in the arts and culture as well as being able to indicate usage levels and behaviours. Such data includes search queries for cultural objects, data that exists about cultural operations and events, e. tickets sales or access to cultural items through linked metadata, and social media traces of cultural activity. This is a result of the increasingly-sophisticated approaches to the measurement of such data that make it increasingly possible to track, measure and influence the spread of ideas and the coming together of groups of people and associated changes in their behaviour both on- and off-line.[154] The emerging trend of computer scientists working in the area of computational humanities that analyse social media signals can also be applied to the cultural sector. Further, Transactional data can be of assistance because,

> If more transactional data were gathered and analysed by cultural organisations, it would give them the potential to run genuine experiments to discover the efficacy of, for instance, sales and marketing techniques and approaches. Knowing more about your audience allows you not only to segment more accurately but also, as a consequence, affords the ability to determine two similar groups and use one as a control, in the scientific sense, against which to test anything from a new slogan to the effectiveness of a marketing channel, even a casting decision.[155]

Big data applications in the cultural sector provide a number of positive outcomes. These practices also expand the meaning of cultural data in a way that promotes the generation of value, in the traditional economic sense, for stakeholders.

Transparency of public institutions

Analysing big data from the cultural sector can assist cultural institutions and organisations in identifying performance opportunities for employees and institutions, as well as ensuring that these institutions meet public needs. This means that funding bodies and governments can also make more informed decisions about the allocation of funds in the cultural sector. Thus, data analytics for this purpose can be built into organisational plans in the pursuit of data driven decision-making.[156] Further, big cultural data can assist the accountability of public funds in the arts and culture, although Lilley observes that currently, "a significant

[152] Manovich, op. cit., 2012.
[153] Ibid.
[154] Donovan, Anna, Rachel Finn, Kush Wadhwa, Lorenzo Bigagli, Guillermo Vega Gorgojo and Martin Georg Skjæveland, *Open Access to Data*, BYTE Project D2.3, 30 September 2014, p62.
[155] Lilley and Moore, op. cit., 2013, p.18.
[156] Lilley, op. cit., 2012.

opportunity to better understand and possibly increase the cultural and social impact of public expenditure is going begging."[157] Nevertheless, there is great potential in aggregating data about the behaviour of cultural consumers could provide powerful new arguments both for the provision and allocation of public funding and for the measurement of its impact. [158]
However, realising the potential of big data applications in the cultural sector and the facilitation of use and re-use of big cultural data (in the context of it referring to digitised works) to positively impact society, by for example the broader dissemination of information, raises challenges that are addressed below.

Big 'cultural' data challenges

Despite the potential benefits of access to big cultural data and the application of big data practices to the cultural sector, challenges have been identified that include:

- Restrictions associated with the funding environment of the cultural sector;
- A limited understanding of or interest in the use of data at senior levels in the cultural sector;
- Inherent threats to intellectual property rights; and
- In the case of user data, the risks to personal data and information privacy.

This is not an exhaustive list of challenges and the extent of their impact have not yet been fully realised given the immaturity of the field of big cultural data.

The funding environment
The cultural sector inhabits the public sector making big data applications subject to funding constraints. The funding environment has limited the extent to which big data has been mobilised or considered in the cultural sector. In that regard, Lilley observes: "Too often, the gathering and reporting of data is seen as a burden and a requirement of funding or governance rather than as an asset to be used to the benefit of the artistic or cultural institution and its work".[159] Funding restrictions can also translate into a lag of technological tools and resources required to keep with the trends in analytics in other sectors. It also limits the budgets spent on salaries and re-training current employees, which can lead to a fragmentation of systems. However, there is some evidence that despite the current approach to supporting the extraction of value from big cultural data, the analysis of big data in other sectors is starting to uncover the possibility of new ways of measuring the impact of arts and cultural investment on our wider society in terms of social capital and cultural value creation.[160]

Limited understanding of or interest in the use of big data in the cultural sector
Big data in the cultural sector can enable access to cultural and historical records maintaining their relevance to contemporary society and also facilitating new discoveries by way of reuse. Aside from the recognised benefits of big cultural data, there is limited interest beyond this: "For many, the potential of data in the cultural sector is at best a 'known-unknown' or worse goes entirely unappreciated."[161] This may also be the reason that,

> in the cultural sector, there are few genuinely massive data sets currently available. Audience Finder is an example, as is Channel Four's 4oD, video-on-demand service, given the broadcaster's public service status. The interaction

[157] Lilley, and Moore, op. cit., February 2013, p.3.
[158] Lilley, op. cit., no date.
[159] Lilley and Moore, op. cit., 2013, p.3.
[160] Donovan, et al., op. cit., 2014, p62.
[161] [161] Lilley and Moore, op. cit., 2013, p.4.

> between these data sets and others and their continued expansion and refinement are key planks of a truly big data approach to cultural policy and decision-making in the future.[162]

Related to little interest is the little perceived value in, and understanding of, the metadata that represents the main data asset held by cultural institutions and organisations. Lilley observes,

> Put simply, almost all cultural organisations already live in a world of greater volume and variety of data – even if they don't yet harness it. Many are exploring the opportunities and challenges of velocity (for instance through the liveness of Twitter). But very few have an integrated strategic approach and the skills and tools to make the most (or even much sense) of the potential that they are faced with.[163]

Furthermore, deriving potential from big cultural data in general has also largely been overlooked and is "currently underused".[164]

However, limited interest and understanding may also be borne out of the contention that surrounds the idea of creating value from cultural data, when it is accepted as adding value by creating cultural value and social capital as discussed in the introduction of this report. Nevertheless, a solution to increasing interest and understanding of big data applications in the cultural sector could be attempted through funding R&D activity to help arts and cultural organisations understand their data 'assets' and systems and look at the relationship between cultural value/ social capital formation.[165] This could also involve bringing data scientists from other, non-cultural fields into the sector as an important way to explore the needs of cultural organisations and to build capacity.[166]

Licensing issues

For cultural data to be lawfully re-used it needs to be done so in accordance with the relevant intellectual property legal framework. Arranging the necessary licensing agreements to enable re-use of cultural data can be a barrier to capturing the full value of the data in terms of it leading to new discoveries and innovations. This not only includes the technological challenge of making the data truly open and accessible, but also necessitates an attitudinal shift amongst traditional rights holders, as well as cultural heritage organisations that hold cultural data. Licensing arrangements in the sector are commonly tackled through applying a Creative Commons licensing regime. Europeana Creative provides a good example of how transparent licensing arrangements can support open cultural data, which enables re-use and the benefits that flow that re-use.

Privacy and data protection issues

In so far as the meaning of big cultural data extends to include big data applications in the cultural sector, such as collecting and analysing user behaviour and sentiment data, personal data protection issues will inevitably arise. For example, analysis of user information, including identifiable information can attract data protection rules safeguarding personal data. However, at this stage, the underuse of big data applications in the cultural sector makes this an issue to be aware of as a potential challenge for stakeholders.

[162] Lilley, "op. cit., no date.

[163] Ibid.

[164] Ibid.

[165] Lilley and Moore, op. cit., 2013, p.8.

[166] Lilley, op. cit., no date.

7 CARTOGRAPHY OF DATA FLOWS

In this section we have mapped data as a resource on an international scale, examining where data originates, where it flows and where it is being processed. It examines which countries, industries, actors and companies are deriving economic and other benefits from big data internationally, as well as which regions, companies and actors are losing out. Here we have examined scientific, security, commercial as well as other types of data flows, and determined whether big data is being better utilised as a resource in some disciplinary contexts than in others.

We measure the data flows on the main intermediation platforms. These platforms are particularly relevant since they operate essentially in all countries in the world, and occupy in most countries the top position with a very large traffic. An analysis of the traffic on these platforms thus allows us to gain a better idea of the global cross-country data flow.

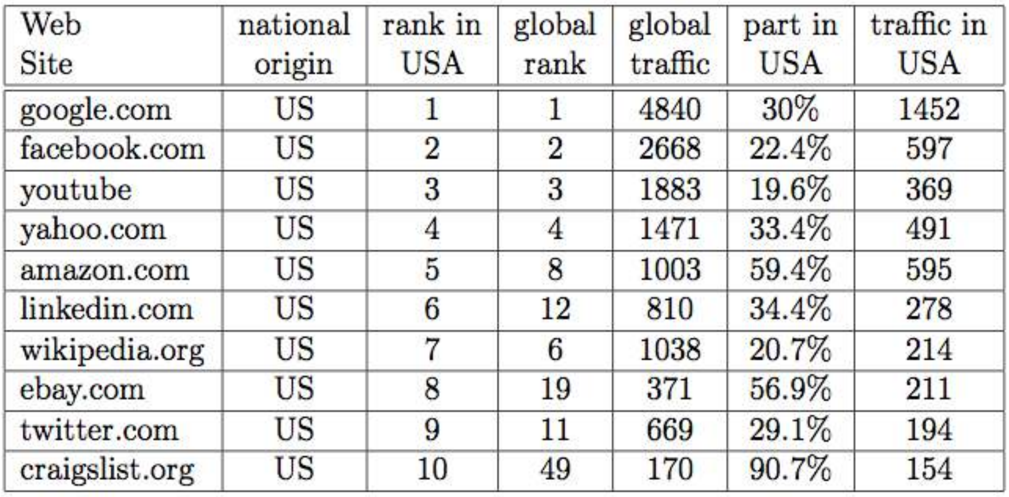
We have considered in addition to the USA and China, countries representative of the different regions in the world, Egypt, Brazil, France and Korea. For each country, we consider the first ten sites locally. It might seem too restrictive a view, but we believe that it is meaningful. Indeed, the traffic on sites decreases very fast. In France for instance, the Top 10 sites represent a third of the traffic of the Top 500 sites. It is thus a good approximation of global activity. Second, the quality of the data harvested on top sites is most often higher and diversified than on platforms of lesser importance which are often more specialised.

7.1 COUNTRY STUDY

For each country, we consider the Top 10 sites, and present some analysis in a table. For each site, we recall its national origin, that is the location of it's headquarter. Its *global rank* is extracted from Alexa, which produces a monthly rank calculated using a combination of average daily visitors and page views over the past month. The *national rank* is deduced from Alexa by taking the percentage in the given country. The *global traffic* is obtained from Traffic estimate, which is based on the number of monthly visits, and not the number of visitors, which generates subtle variations, but doesn't change the big picture. Traffic is measured in millions of visited[167]. For sites which have both a .com and a .xx, where xx in a country code, we indicate the ranks of both. As an example, Google.fr occupies global rank 29th, while Google.com is number 1.

[167] The statistics of visits have been obtained the week of November, 4, 2013 from www.alexa.com and www.trafficestimate.com.

Table 12 Top 10 sites in the US



Web Site	national origin	rank in USA	global rank	global traffic	part in USA	traffic in USA
google.com	US	1	1	4840	30%	1452
facebook.com	US	2	2	2668	22.4%	597
youtube	US	3	3	1883	19.6%	369
yahoo.com	US	4	4	1471	33.4%	491
amazon.com	US	5	8	1003	59.4%	595
linkedin.com	US	6	12	810	34.4%	278
wikipedia.org	US	7	6	1038	20.7%	214
ebay.com	US	8	19	371	56.9%	211
twitter.com	US	9	11	669	29.1%	194
craigslist.org	US	10	49	170	90.7%	154

7.1.1 USA

Let's consider the Top 10 sites in the USA. The situation is quite regular, all of them are American. Moreover, as exhibited in Table 12, all of them are global systems, with a global cover. In average, national access represents only 31% of all accesses for the Top 7 sites.

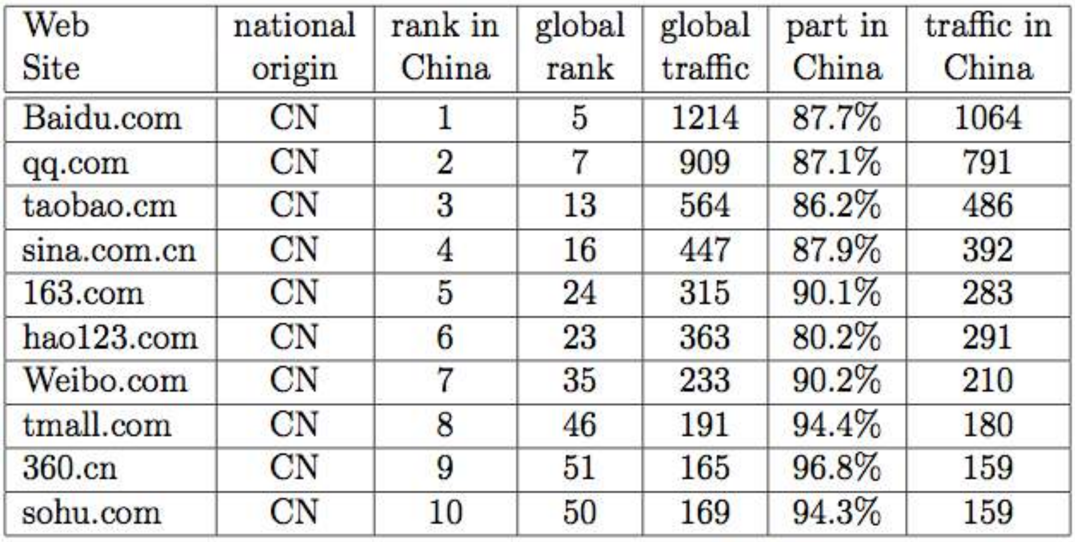
7.1.2 China

For the second power online, China, a similar picture holds with only national sites among the Top 10 as shown in Table 13. Some of these sites are also global leaders, such as Baidu, second global search engine, with 18% global marketshare[168], after Google which has 65%, but far ahead of other competitors. Google is on the other hand the only foreign site in the Top 20 in China. What drastically differs from the US at this stage is the global impact, although undoubtedly, China has strong international ambition, cf. Alibaba.

If 87% of the activity of the Top 7 sites is national, a ratio close to countries like France, given the size of the country, the activity abroad is important. Baidu for instance, with 87.7% of activity on the national stage, and traffic of 1214 million monthly visits, enjoys 150 million monthly visits abroad. The e-business platform Alibaba, 63rd global rank, with only 43% of its activity locally, is ahead of the national industry for the international coverage and ambition, with a volume of good exchanged whose values is higher that eBay and Amazon combined.

[168] http://www.netmarketshare.com

Table 13 Top 10 sites in China



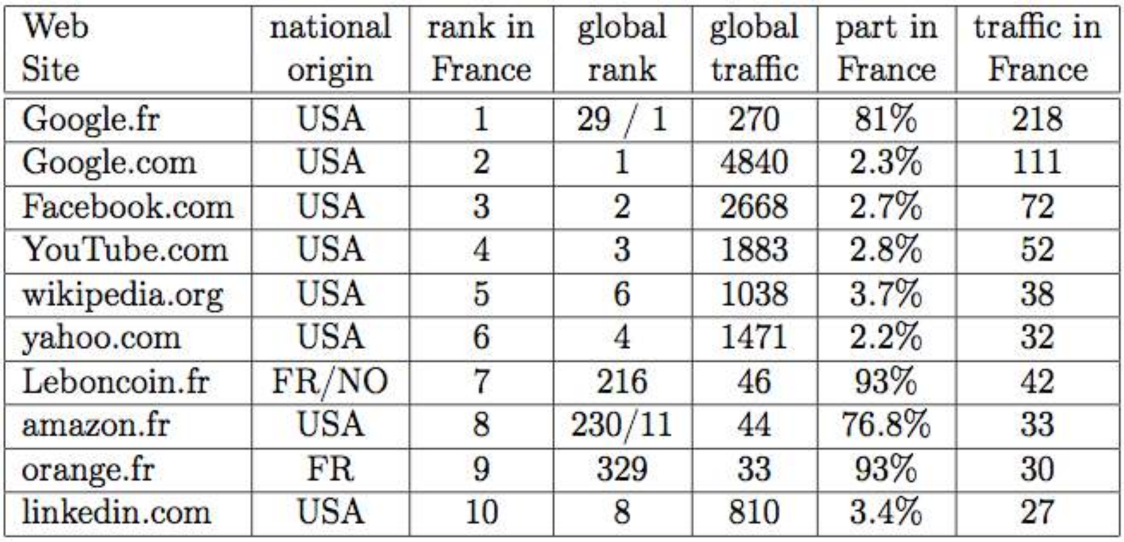
Web Site	national origin	rank in China	global rank	global traffic	part in China	traffic in China
Baidu.com	CN	1	5	1214	87.7%	1064
qq.com	CN	2	7	909	87.1%	791
taobao.cm	CN	3	13	564	86.2%	486
sina.com.cn	CN	4	16	447	87.9%	392
163.com	CN	5	24	315	90.1%	283
hao123.com	CN	6	23	363	80.2%	291
Weibo.com	CN	7	35	233	90.2%	210
tmall.com	CN	8	46	191	94.4%	180
360.cn	CN	9	51	165	96.8%	159
sohu.com	CN	10	50	169	94.3%	159

7.1.3 France

Let's consider now a country representative of Europe, namely France. As shown in Table 14, most Top sites used in France are American, including all of the Top 6. The Top 10 sites have traffic of 655 million monthly visits, 583 million on the Top American sites, and only 72 million on the 2 European sites. Thus 89% of the visits are made on US sites. The ratio of French sites increases slightly beyond the Top 10.

When considering the activity of the Top 7 French sites, 78% of their activity is national. Leboncoin, top French site is not even in the Top 500 in the USA. Dailymotion.com, which is really an exception in the French landscape, belongs to the Top 500 in the USA (180th), enjoys a large territory of activity, including countries such as beyond the USA and France, Japan, India, Pakistan, etc. It makes only 11% of its activity at the national level, 86 million monthly visits abroad far ahead of Leboncoin with million monthly visits.

Table 14 Top 10 sites in France



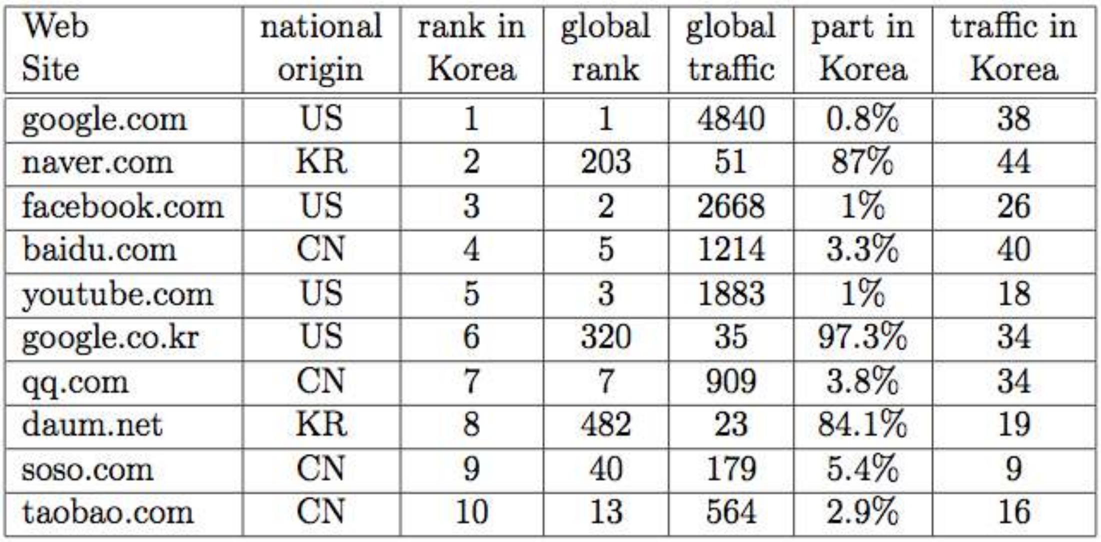
Web Site	national origin	rank in France	global rank	global traffic	part in France	traffic in France
Google.fr	USA	1	29 / 1	270	81%	218
Google.com	USA	2	1	4840	2.3%	111
Facebook.com	USA	3	2	2668	2.7%	72
YouTube.com	USA	4	3	1883	2.8%	52
wikipedia.org	USA	5	6	1038	3.7%	38
yahoo.com	USA	6	4	1471	2.2%	32
Leboncoin.fr	FR/NO	7	216	46	93%	42
amazon.fr	USA	8	230/11	44	76.8%	33
orange.fr	FR	9	329	33	93%	30
linkedin.com	USA	10	8	810	3.4%	27

7.1.4 Korea

Korea occupies a remarkable position, with powerful national sites, such as the Naver portal, which offers the first Korean search engine, or the communication portal daum.net. At the international level Korea enjoys more diversity, relying on sites from both the USA and China as shown in Table 15. Like for most countries, Korean sites have little visibility abroad, and the linguistic area is restricted. Naver though is among the Top 20 sites in Japan together with its competitor Baidu.

Even in the influence of China is strong in Asia; its impact goes beyond the regional borders as shown by the following examples of Egypt and Brazil for instance. The Chinese portal hao123.com occupies in these two countries the respectively the 8th and the 20th positions for instance.

Table 15 Top 10 sites in Korea

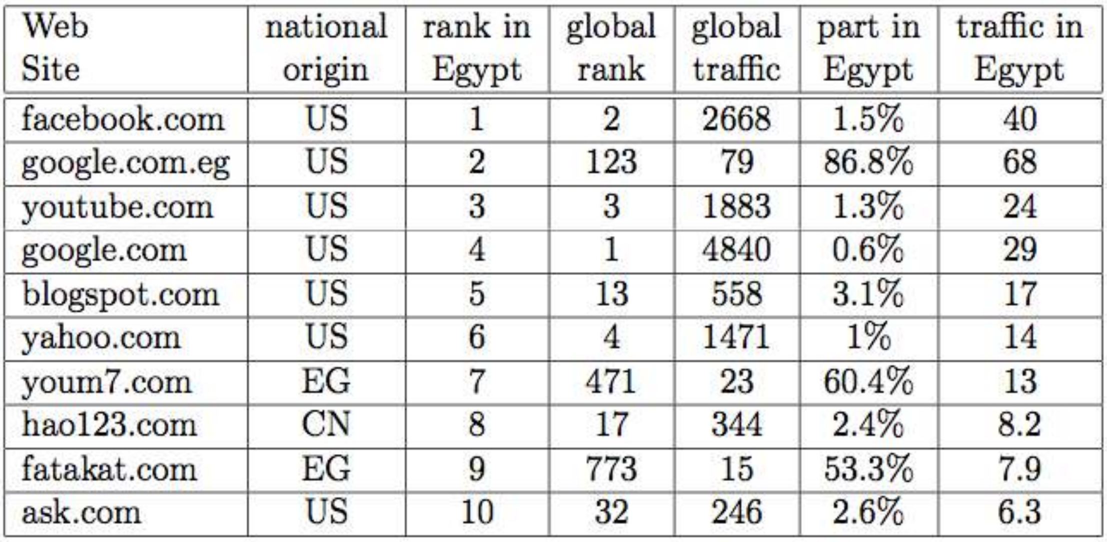


Web Site	national origin	rank in Korea	global rank	global traffic	part in Korea	traffic in Korea
google.com	US	1	1	4840	0.8%	38
naver.com	KR	2	203	51	87%	44
facebook.com	US	3	2	2668	1%	26
baidu.com	CN	4	5	1214	3.3%	40
youtube.com	US	5	3	1883	1%	18
google.co.kr	US	6	320	35	97.3%	34
qq.com	CN	7	7	909	3.8%	34
daum.net	KR	8	482	23	84.1%	19
soso.com	CN	9	40	179	5.4%	9
taobao.com	CN	10	13	564	2.9%	16

7.1.5 Egypt

Egypt is massively dependent on American sites, which are dominant in the region, but Egypt has also sites which have a national as well as regional influence in its linguistic sphere, among which the news site youm7.com, and the women portal - fatakat.com, which has a strong visibility in the Arabic world.

Table 16 Top 10 sites in Egypt



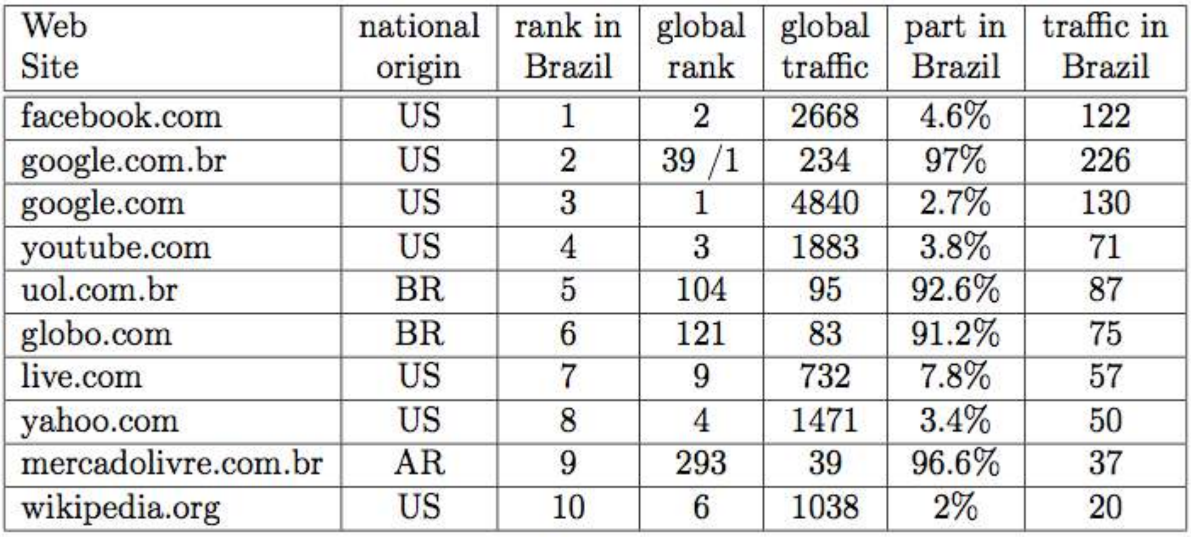
Web Site	national origin	rank in Egypt	global rank	global traffic	part in Egypt	traffic in Egypt
facebook.com	US	1	2	2668	1.5%	40
google.com.eg	US	2	123	79	86.8%	68
youtube.com	US	3	3	1883	1.3%	24
google.com	US	4	1	4840	0.6%	29
blogspot.com	US	5	13	558	3.1%	17
yahoo.com	US	6	4	1471	1%	14
youm7.com	EG	7	471	23	60.4%	13
hao123.com	CN	8	17	344	2.4%	8.2
fatakat.com	EG	9	773	15	53.3%	7.9
ask.com	US	10	32	246	2.6%	6.3

7.1.6 Brazil

The Web in South-America resembles its counterpart in other regions, with a strong American domination, and some local sites with limited regional influence as illustrated by the example of Brazil.

Among the Top 10 sites in Brazil[169], there are 7 Americans, and 3 regional from Brazil or Argentina.

Table 17 Top 10 sites in Brazil



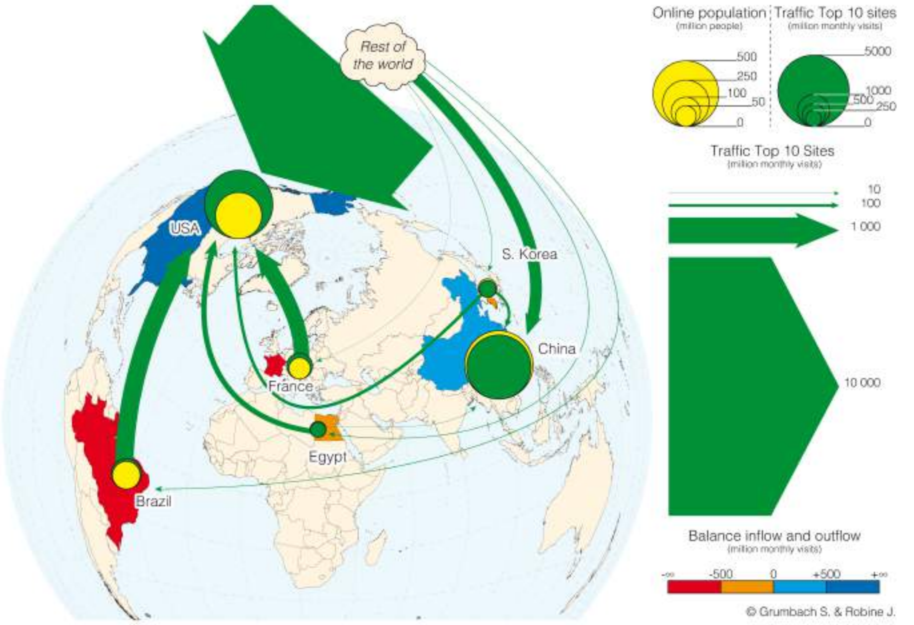
Web Site	national origin	rank in Brazil	global rank	global traffic	part in Brazil	traffic in Brazil
facebook.com	US	1	2	2668	4.6%	122
google.com.br	US	2	39 /1	234	97%	226
google.com	US	3	1	4840	2.7%	130
youtube.com	US	4	3	1883	3.8%	71
uol.com.br	BR	5	104	95	92.6%	87
globo.com	BR	6	121	83	91.2%	75
live.com	US	7	9	732	7.8%	57
yahoo.com	US	8	4	1471	3.4%	50
mercadolivre.com.br	AR	9	293	39	96.6%	37
wikipedia.org	US	10	6	1038	2%	20

7.1.7 Global perspective

The Figure 9 shows the flows of data between the above-mentioned countries. It relies on the figures presented in Table 18, which exhibits a representative sample of the flows between

[169] Alexa, 12/12/13

these six countries. For each country, we indicate the size of the online population, the traffic of the Top 10 sites in that country, the outgoing traffic on foreign sites as well as the incoming flow.

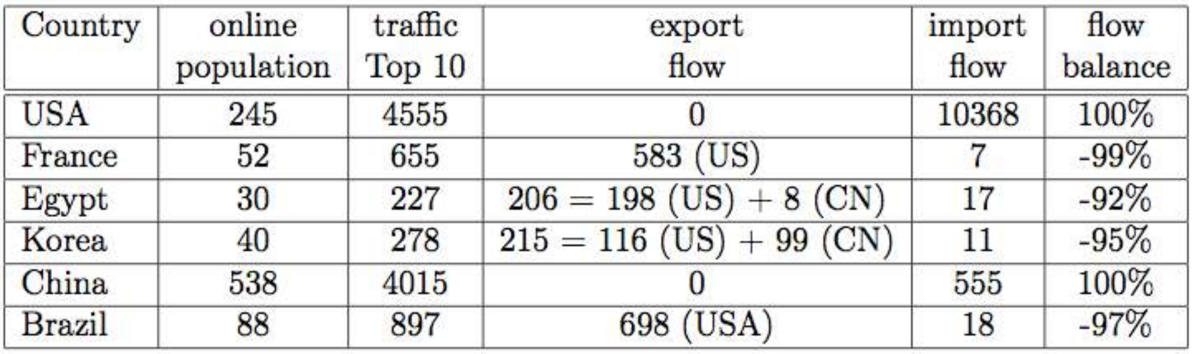


© Grumbach S. & Robine J.

Figure 9 Map of World Data Flow

These measures do not take into account sites that would not be in the Top local sites, but have important activity abroad, such as DailyMotion for France.

Table 18 Online population and flow balance



Country	online population	traffic Top 10	export flow	import flow	flow balance
USA	245	4555	0	10368	100%
France	52	655	583 (US)	7	-99%
Egypt	30	227	206 = 198 (US) + 8 (CN)	17	-92%
Korea	40	278	215 = 116 (US) + 99 (CN)	11	-95%
China	538	4015	0	555	100%
Brazil	88	897	698 (USA)	18	-97%

8 SUMMARY

Big data has become the so called *mother of invention*, forcing different industries to take a fresh look at their data and ask themselves whether they are using it strategically. In order to uphold competitive advantage, industries must focus on a well-defined business goal, and persistently assess the business case for intensifying their analytics activities to encompass big data.

A data driven world has the potential to improve the efficiencies of industry sectors and improve the quality of human life. No doubt big data brings new opportunities to modern society and, at the same time, challenges to data community. For instance, on one hand, big data hold great promises for discovering subtle patterns and heterogeneities that are not possible with small-scale data. On the other hand, the massive sample size and high dimensionality of big data introduce unique challenges, including scalability and storage bottleneck, and measurement errors. In this report an overview on the salient features of big data as well as a brief overview on big data problems, including opportunities and challenges in selected industry sectors, is presented.

The Oil and Gas sector, known for its ability to adapt to challenges of the digital age, is entering a new generation of data driven transformation. The oil and gas industry is greatly dependent on data to make critical strategic decisions – yet they do not fully realise the value of this data. As industry captures petabytes of data daily, it is the ability to understand analytics trends, correctly interpret all geological, engineering, production and equipment data performance data swiftly and efficiently that warrants success. The ability to access and draw actionable insights from data sets is at the heart of profitability in this industry and in an industry where success relies on how rapidly one can predict potential and keeping costs low to realise that success. The oil and gas industry has undergone significant growth in unconventional resources markets. This increase in focus has not only brought on greater competition for assets, but a smaller margin for error. With projects demanding more expensive drilling and production technology and profound changes in government regulations and commodities, companies need to work-out operational wisdom and strategic foresight to ensure success.

Data is taking over in an eloquent manner, and is transforming the healthcare industry. There is more data available than ever before, and applying accurate analytics can spur growth. Benefits extend to patients, providers, and physicians, and the technology can make integrated patient management a reality. Healthcare data definitely meets the definition of big data. The challenges enveloping the complete aggregation and use of healthcare data are not insuperable. Sustaining those challenges will require a culture shift in healthcare both internal to providers and between providers and other sections of the industry. The major challenge is determining the proper balance between protecting the patient's information and maintaining the integrity and usability of the data. Robust information and data governance programs will address a number of these challenges. The sharing of data between organizations must be addressed before the full potential of big data in health care may be unlocked. Recall that Gartner defines big data as "high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Laney, 2012). Healthcare data meets the "3Vs" of the big data definition. By recognising the second part of the Gartner definition of big data, namely "innovative forms of information processing for enhanced insight and decision making," (Laney, 2012) will make a considerable impact on not only the healthcare delivery system but the Europe as a whole.

Big Data is also expected to help the world solve some of its most intractable environmental problems. While solutions addressed in the Section 6.3 primarily the volume and velocity axes, other architectural and technological solutions are oriented to address the variety due to heterogeneity of datasets. The key utilization of big data is its ability to help assess environmental risks, both in real time and in the future. In order to tackle environmental issues, very different types of models need to be combined.

The deployment of big data tools in crisis informatics can have significant, positive impacts on the life chances of individuals caught up in sudden, complex and difficult circumstances. Big data can assist across the crisis informatics lifecycle to help actors prepare for disasters, respond during disasters and evaluate the situation in the medium to long term. A review of these applications has indicated that big data is cyclical, and can have self-perpetuating impacts across the whole lifecycle of a crisis event. However, experts in the field note that the evidence of potential positive impacts of big data analytics in this area are mainly drawn from anecdotal evidence and there has not yet been any empirical evaluation, baseline assessment or systematic learning about these interventions.[170] This is essential in order to realistically evaluate their impacts. Furthermore, Letouzé et al., point out that while big data offers descriptive information can help in understanding what is happening, it does not yet offer much information about why it is happening. Thus in crisis informatics contexts, significant further research needs to be done to understand the areas in which big data can offer analytic insight as well as descriptive insight.

Moreover, the close integration of ordinary people in crisis management activities, and the proliferation of Internet-based communication in general, has been heralded as more participatory and democratic.[171] This analysis has demonstrated that there is an ecosystem of stakeholders involved in big data and crisis informatics, which include authorities, humanitarian organisations, first responders, crisis managers and other professionals as well as members of the public. Yet, while ordinary people are getting more involved in information gathering and response activities, this does not necessarily mean that they are trained appropriately, that they are accessing people in most need of help, nor that they are acting responsibly and with due attention to best practice (including privacy and security). This is particularly significant as new information and communication technologies often reflect and reinforce existing power relations between authorities and citizens, and between different groups of people.[172] As such bringing any new technology into use has to be sensitive to these potentials, and directly and proactively address them. In the area of crisis management, development and associated areas, the use of big data can have significant impacts on life chances, and these should be distributed as equally and equitably as possible.

Smart cities and big data are buzz themes in recent times, but the implications of how the city is being wired, how it is generating new data, how this data might force new theories and models relevant to our understanding, how to use intelligence to plan the city, building on this new understanding−these are all key questions to be explored. The shift from cities to Smart Cities depends on the efficiency with which information is shared among citizens and private and public companies. This information brings challenges, and, following the big data revolution, novel processing schemes must be adopted to enable the possibilities that exist of

[170] OCHA, op. cit., 2013, p. 7.

[171] Ibid.

[172] McCahill, Michael and Rachel L. Finn, *Surveillance, Capital and Resistance: Theorizing the surveillance subject*, Routledge, London, 2014.

this domain. All the possibilities enabled by smart cities, like improved quality of life or energy efficiency, shall build on top of efficient data processing and users' privacy protection schemes.

To enable better decision-making, efficiency, and cost savings for its customers, shipping industry analyses live sea traffic conditions and metrics from multiple carriers. These data sets are massive, and the speed at which they must be analysed is primarily real time. The data is also varied, coming from wired and wireless devices. The big shipping data sets significant demand on data storage, event processing, and analytics. The shipping industry is beginning to embrace the opportunities provided by technical, engineering data that allows condition based maintenance, efficiency and engine health monitoring, but the data prospect is far wider. The smart collection and analysis of massive commercial datasets are driving transparency across the industry.

Big cultural data is largely undefined, and research suggests that the potential of big cultural data is yet realised. However, in the context of the BYTE project case study on big cultural data, it refers to public or private collections of digitised cultural works, including their associated metadata, as held by cultural organisations and institutions. However, as the field matures, a pertinent definition of big cultural data is one that also includes big data applications in the cultural sector and the related data that is captured and used by these applications. Whilst the numbers of current applications appear minimal, and are perhaps restricted by the longstanding sentiment that cultural data is a cultural investment rather an investment that can be quantified in economic terms, the potential for big cultural data is far reaching.

The broader definition adopted by this report reflects the reach of big data. As this field of big data matures, so too will the understanding and adoption of the number of potential uses of big data in various sectors, as well as applications of big data practices in the various sectors. This will produce positive societal impacts, both in terms of adding social capital and value, as well as commercial oriented outcomes. It follows that challenges will be better identified and overcome.

Finally, the data flows on the main intermediation platforms are described. We have considered different countries representing different regions in the world. An analysis of the traffic has provided a better understanding of the global cross-country data flow.

REFERENCES

Adamov, A. Distributed file system as a basis of data-intensive computing, in: 6th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–3 (October 2012).

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Michael Franklin, Johannes Gehrke, Laura Haas, Jiawei Han Alon Halevy, H.V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakon stantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom, Challenges and Opportunities with big data, CYBER CENTER TECHNICAL REPORTS, Purdue University, 2011.

Ahrens, J.P., Hendrickson, B., Long, G., Miller, S., Ross, R., Williams, D. Data-intensive science in the us doe: case studies and future challenges, Comput. Sci. Eng. 13 (6) (2011) 14–24.

Akerkar, R. Improving Data Quality on Big and High-Dimensional Data. Journal of Bioinformatics and Intelligent Control, Vol. 2, No. 1, pp 155-162 (8), 2013.

Akerkar, R., Lingras, P. Building an Intelligent Web: Theory & Practice, Sudbury: Jones & Bartlett Publishers, 2007.

Anderson, C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008. <http://www.wired.com/science/discoveries/magazine/16-07/pb-theory>.

Armour, F. Introduction to big data, presentation at the symposium Big Data and Business Analytics: Defining a Framework, Center for IT and Global Economy, Kogod School of Business, American University, Washington, DC. (September 21, 2012).

Baaziz A., Quoniam L., 2013. The information for the operational risk management in uncertain environments: Case of Early Kick Detection while drilling of the oil or gas wells, International Journal of Innovation and Applied Studies (IJIAS), Vol. 4 No. 1, Sep. 2013.

Baumann, P., Holsten, S., 2012. A Comparative Analysis of Array Models for Databases. International Journal of Database Theory and Application 5(1), 89-120.

Bell, G., Hey, T., Szalay, A. Beyond the data deluge, Science 323 (5919) (2009) 1297–1298.

Bencivenni, M. , Bonifazi, F., A. Carbone, A. Chierici, A. D’Apice, D. De Girolamo, L. dell’Agnello, M. Donatelli, G. Donvito, A. Fella, F. Furano, D. Galli, A. Ghiselli, A. Italiano, G. Lo Re, U. Marconi, B. Martelli, M. Mazzucato, M. Onofri, P.P. Ricci, F. Rosso, D. Salomoni, V. Sapunenko, V. Vagnoni, R. Veraldi, M.C. Vistoli, D. Vitlacil, S. Zani, A comparison of data-access platforms for the computing of large hadron collider experiments, IEEE Trans. Nucl. Sci. 55(3) (2008) 1621–1630.

Brewer E., *“CAP Twelve Years later: How the Rules Have Changed”*, IEEE Computer, Pages 23-29, February 2012.

Brewer E., *“Lesson from Giant-Scale Services”*, IEEE Internet Computing, Pages 46-55, July/Aug 2001.

Brulé M., Tapping the power of big data for the oil and gas industry, IBM Software White Paper for Petroleum Industry, May 2013.

Brumfiel, G. High-energy physics: down the petabyte highway, Nature (469) (2011) 282–283.

Bryant, R. E. Data Intensive supercomputing: The Case for Disc. Technical Report CMU-CS-07-128, 2007.

Bryant, R. E. Data-intensive scalable computing for scientific applications, Comput. Sci. Eng. 13 (6) (2011) 25–33.

Card, D, Chetty, R., Feldstein, M. and Saez, E. 2010. “Expanding Access to Administrative Data for Research in the United States.” NSF SBE 2020 White Paper, National Science Foundation Directorate of Social, Behavioral, and Economic Sciences, Arlington, VA.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., Gruber, R. E., 2006. Bigtable: A Distributed Storage System for Structured Data. Google Research Publications. Available at http://research.google.com/archive/bigtable-osdi06.pdf

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G.M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., Schweitzer, R., in press. The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2013.07.002.

Diebold, F.X. (2000). Big data dynamic factor models for macroeconomic measurement and forecasting. Discussion read to the 8th World Congress of the Econometric Society, Seattle, August. http://www.upenn.edu/~fdiebold/papers107/ABCD_HOED.pdf.

Dumbill, E., 2013. Making sense of Big Data, Big Data 1(1), 1-2, DOI: 10.1089/big.2012.1503.

Dykstra D., *"Comparison of the Frontier Distributed Database Caching System to NoSQL Databases"*, Computing in High Energy and Nuclear Physics (CHEP) Conference, May 2012.

ESA, 2013. Big Data from Space – Event Report. Available at http://www.congrexprojects.com/docs/default-source/13c10_docs/13c10_event_report.pdf?sfvrsn=2.

Feblowitz J., 2012, The Big Deal About Big Data in Upstream Oil and Gas, Paper & presentation, IDC Energy Insights, October 2012.

Fernando, N., Loke, S.W., Rahayu, W., 2013. Mobile cloud computing: A survey. Future Generation Computer Systems 29(1), 84–106.

Fey, P., Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, Seung Yon Y. Rhee, Doug Howe, Maria Costanzo, Big Data: the future of biocuration, Nature 455 (7209) (2008) 47–50.

Foster, I., Zhao, Y., Raicu, I., Lu, S. Cloud computing and grid computing 360-degree compared, in: Grid Computing Environments Workshop, 2008, GCE'08, 2008, pp. 1–10.

Fox A.; Brewer E.A., *"Harvest, Yield, and Scalable Tolerant Systems"*, Proceedings of the Seventh Workshop on Hot Topics in Operating Systems, Pages 174-178, March 1999.

Furht, B. Armando Escalante, Handbook of Cloud Computing, Springer, 2011.

Garcia, A.O. , Bourov, S., Hammad, A., Hartmann, V., Jejkal, T., Otte, J.C. , Pfeiffer, S., Schenker, T. , Schmidt, C. , Neuberger, P., Stotzka, R. , van Wezel, J., Neumair, B. , Streit, A. Data-intensive analysis for scientific experiments at the large scale data facility, in: 2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV), 2011, pp. 125–126.

Gualtieri, M., The pragmatic definition of big data. Forrester Research Blog. http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data.

Hampton, S.H., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11, 156–162

Hassan, K., Mahmoud, F. An incremental approach for the solution of quadratic problems, Math. Modell. 8 (1987) 34–36.

Havlik, D., Schade, S., Sabeur, Z.A., Mazzetti, P., Watson, K., Berre, A.J., Mon, J.L., 2011. From Sensor to Observation Web with Environmental Enablers in the Future Internet. Sensors 11(4), 3874-3907.

Heer, J., Mackinlay, J.D., Stolte, C., Agrawala, M. Graphical histories for visualization: supporting analysis, communication, and evaluation, IEEE Trans. Visual. Comput. Graph. 14 (6) (2008) 1189–1196.

Hems A., Soofi A., Perez E., Drilling for New Business Value: How innovative oil and gas companies are using big data to outmanoeuvre the competition, A Microsoft White Paper, May 2013

Hey, T., & Trefethen, A. E. (2002). The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*, 1017–1031.

Hey, T., Tansey, S., Tolle, K. (edited by), 2009. The Fourth Paradigm - Data-Intensive Scientific Discovery. Microsoft Research, 2009. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf.

Hey, T., Trefethen, A. E., The UK e-science core programme and the grid, Future Gener. Comput. Syst. 18 (8) (2002) 1017–1031.

Hilbert, M., Lopez, P. The world's technological capacity to store, communicate, and compute information, Science 332 (6025) (2011) 60–65.

Holdaway, K. R. "Exploratory Data Analysis in Reservoir Characterization Projects," SPE/EAGE Reservoir Characterization and Simulation Conference, 19–21 October 2009, Abu Dhabi, UAE.

Hollingsworth J., 2013, Big Data for Oil & Gas, Oracle Oil & Gas Industry Business Unit, March 2013.

Horne, A., Shah, S., Capella, J. Good Data won't Guarantee Good Decisions, 2012. <http://hbr.org/2012/04/good-data-wont-guaranteegood-decisions>.

Hsiao, W-F., Chang, T.M. An incremental cluster-based approach to spam filtering, Expert Syst. Appl. 34 (3) (2008) 1599–1608.

Ishii, R.P., de Mello, R.F. A history-based heuristic to optimize data access in distributed environments, in: Proc. 21st IASTED International Conf. Parallel and Distributed Computing and Systems, 2009.

Jacob, B., Brown, M., Fukui, K., Trivedi, N. Introduction to Grid Computing, IBM Redbooks Publication, 2005.

Jacobs, A. The pathologies of big data, Commun. ACM 52 (8) (2009) 36–44.

Jagadish, H. V. , Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R. and Shahabi, C. 2014. Big data and its technical challenges. *Commun. ACM* 57, 7 (July 2014), 86-94.

Keim, D. A. Christian Panse, Mike Sips, Visual data mining in large geospatial point sets, IEEE Comput. Graph. Appl. 24 (5) (2004) 36–44.

Kouzes, R.T., Anderson, G.A., Elbert, S.T., Ian Gorton, Gracio, D.K. The changing paradigm of data-intensive computing, Computer 42 (1) (2009) 26–34.

L. Bigagli, S. Nativi, P. Mazzetti. *Mediation to deal with information heterogeneity*. European Geosciences Union, Advances in Geosciences (ADGEO), vol. 8, Earth System Science Data access, distribution and use for education and research, pp. 3-9. SRef-ID: 1680-7359/adgeo/2006-8-3, 2006.

Laney, D. 3d Data management: controlling data volume, velocity and variety, Appl. Delivery Strategies Meta Group (949) (2001). http://refhub.elsevier.com/S0020-0255(14)00034-6/h0650.

Laney, D. "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012.

Laney, D., 2001. 3D Data Management. Controlling Data Volume, Velocity, and Variety in Application Delivery Strategy, META Group, February 2001.

Leong, D. A new revolution in enterprise storage architecture, IEEE Potentials 28 (6) (2009) 32–33.

Lynch, C. Big data: how do your data grow?, Nature 455 (7209) (2008) 28–29.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2012.

McDermott, J., Samudrala, R., Bumgarner, R., Montgomery, K. Computational Systems Biology, Humana Press, 2009.

McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. (May 2011).

Muhleisen, H., Dentler, K. Large-scale storage and reasoning for semantic data using swarms, IEEE Comput. Intell. Mag. 7 (2) (2012) 32–44.

Nativi, S., Mazzetti, P., Geller, G. N., 2013. Environmental model access and interoperability: The GEO Model Web initiative. Environmental Modelling and Software 39, 214-228.

Nativi, S., Mazzetti, P., Saarenmaa, H., Kerr J., Tuama E. Ó., 2009. Biodiversity and climate change use scenarios framework for the GEOSS interoperability pilot process. Ecological Informatics 4(1), 23-33.

Nicholson R., 2012, Big Data in the Oil & Gas Industry, IDC Energy Insights, September 2012.

O'Reilly Radar Team, Big Data Now: Current Perspeectives from O'Reilly Radar. O'Reilly, 2011.

Oehmen, C., Nieplocha, J., Scalablast: a scalable implementation of blast for high-performance data-intensive bioinformatics analysis, IEEE Trans. Parallel Distrib. Syst. 17 (8) (2006) 740–749.

Oprea, A., Reiter, M. K., Yang, K. Space efficient block storage integrity, in: Proc. 12th Ann. Network and Distributed System Security Symp. (NDSS 05), 2005.

Petitdidier, M., Cossu, R., Mazzetti, P., Fox, P., Schwichtenberg, H., & Som de Cerff, W. (2009). Grid in Earth Sciences. *Earth Science Informatics, 2*, 1-3.

Philip Chen, C. L., Zhang, C-Y. *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data*, Information Sciences, Volume 275, 10 August 2014, Pages 314-34.

Pirovano, A. , Lacaita, A.L. , Benvenuti, A. , Pellizzer, F. , Hudgens, S. , Bez, R. Scaling analysis of phase-change memory technology, IEEE Int. Electron Dev. Meeting (2003) 29.6.1–29.6.4.

Provost, F., Fawcett, T. *Data Science and its Relationship to Big Data and Data-Driven Decision Making*, Big Data. March 2013, 1(1): 51-59.

Qian Wang, Cong Wang, Kui Ren, Wenjing Lou, Jin Li, Enabling public auditability and data dynamics for storage security in cloud computing, IEEE Trans. Parallel Distrib. Syst. 22 (5) (2011) 847–859.

Qian Wang, Kui Ren, Wenjing Lou, Yanchao Zhang, Dependable and secure sensor data storage with dynamic integrity assurance, in: Proc. IEEE INFOCOM, 2009, pp. 954–962.

Renato Porfirio Ishii, Rodrigo Fernandes de Mello, An adaptive and historical approach to optimize data access in grid computing environments, INFOCOMP J. Comput. Sci. 10 (2) (2011) 26–43.

Renato Porfirio Ishii, Rodrigo Fernandes de Mello, An online data access prediction and optimization approach for distributed systems, IEEE Trans. Parallel Distrib. Syst. 23 (6) (2012) 1017–1029.

Savitz, E., Gartner: 10 Critical Tech Trends for the Next Five Years, October 2012. <http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/>.

Savitz, E., Gartner: Top 10 Strategic Technology Trends for 2013, October 2012. <http://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top-10-strategic-technology-trends-for-2013/>.

Scholes, R.J., Walters, M., Turak, E., Saarenmaa, H., Heip, C.H.R., Tuama, É.Ó., Faith, D.P., Mooney, H.A., Ferrier, S., Jongman, R.H.G., Harrison, I.J., Yahara, T., Pereira, H.M., Larigauderie, A., Geller. G., 2012. Building a global observing system for biodiversity. Current Opinion in Environmental Sustainability 4(1), 139–146.

Seshadri M., 2013, Big Data Science Challenging The Oil Industry, CTO Global Services, EMC Corporation, March 2013.

Shah N.H. and Tenenbaum J.D.: *J Am Med Inform Assoc.* 19(e1): e2-e4 (2012).

Shen, H., Zhao, L., Li, Z. A distributed spatial-temporal similarity data storage scheme in wireless sensor networks, IEEE Trans. Mobile Comput. 10 (7) (2011) 982–996.

Simeonidou, D., Nejabati, R., Zervas, G., Klonidis, D., Tzanakaki, A., Mahony, MJO. Dynamic optical-network architectures and technologies for existing and emerging grid services, J. Lightwave Technol. 23 (5) (2005) 3347–3357.

Simoff, S. Bohlen, M.H., Mazeika, M. Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, Springer, 2008.

Simone Ferlin Oliveira, Karl Furlinger, Dieter Kranzlmuller, Trends in computation, communication and storage and the consequences for dataintensive science, in: IEEE 14th International Conference on High Performance Computing and Communications, 2012.

Smith, J.A. FIELD NOTE: What Makes Big Data Big – Some Mathematics Behind Its Quantification, Data Scientist Insights. http://datascientistinsights.com/2012/12/19/field-note-what-makes-big-data-big-some-mathematics-behind-its-quantification/.

Smith, M., Szongott, C., Henne, B., von Voigt, G., Big data privacy issues in public social media, in: 2012 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), 2012, pp. 1–6.

Steed, C.A., Ricciuto, D.M., Shipman, G., Smith, B., Thornton, P.E., Wang, D., Shi, X., Williams, D.N., 2013. Big data visual analytics for exploratory earth system simulation analysis. Computers & Geosciences 61, 71–82.

Szalay, A. S. Extreme data-intensive scientific computing, Comput. Sci. Eng. 13 (6) (2011) 34–41.

Szalay, A., Gray, J. Science in an exponential world, Nature 440 (2006) 23–24.

Tambe P. *Big data know-how and business value. Working paper,* NYU Stern School of Business, NY, New York, 2012.

Tang, K., Lin, M., Minku, F.L., Yao, X. Selective negative correlation learning approach to incremental learning, Neurocomputing 72 (13-15) (2009) 2796–2805.

TechAmerica Foundation. Demystifying Big Data. Washington, DC. (2012).

Tkacz, E., Kapczyn´ski, A., Internet: Technical Development and Applications, Springer, 2009. http://refhub.elsevier.com/S0020-0255(14)00034-6/h0950

Varian, Hal. 2010. "Computer- Mediated Transactions." *American Economic Review Papers and Proceedings* 100 (2): 1–10.

Vettiger, P., Cross, G., Despont, M., Drechsler, U., Durig, U., Gotsmann, B., Haberle, W., Lantz, M.A., Rothuizen, H.E., Stutz, R., Binnig, G.K. The millipede –nanotechnology entering data storage, IEEE Trans. Nanotechnol. 1 (1) (2002) 39–55.

Vouk, M. A. Cloud computing – issues, research and implementations, in: 30th International Conference on Information Technology Interfaces, 2008, ITI 2008, 2008, pp. 31–40.

Wang, Fei-Yue, Zeng, D., Carley, K. M., Mao, W. Social computing: from social informatics to social intelligence, IEEE Intell. Syst. 22 (2) (2007) 79–83.

Worlton, W. Bulk storage requirements in large-scale scientific calculations, IEEE Trans. Magn. 7 (4) (1971) 830–833.

Wu, Y., Yuan, G., Ma, K-L., Visualizing flow of uncertainty through analytical processes, IEEE Trans. Visual. Comput. Graph. 18 (12) (2012) 2526–2535.

Zhang, J., Wang, F-Y., Wang, K., Lin, Y-H., Xu, X., Chen, C. Data-driven intelligent transportation systems: a survey, IEEE Trans. Intell. Trans. Syst. 12 (4) (2011) 1624–1639.

Zikopoulos, P., Eaton, C. Understanding Big Data*:* Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill Professional, 2011.