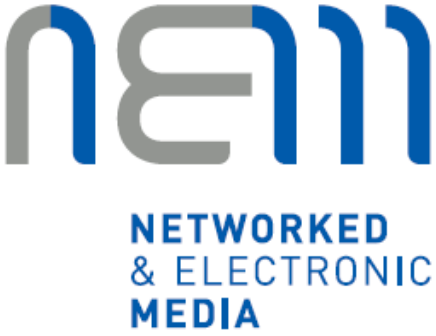


Big and Open data Position Paper

December 2013

This document has been drafted by a set of NEM members (see list of contributors below) under the leadership of Pierre-Yves DANET (Orange), and approved by the NEM Steering Board

Content

Executive Summary

From the NEM community perspective, Big and Open data is becoming a hot topic as far as most of the content will be stored in the future in data centers and there is a need to optimize the usage of such infrastructure. This huge amount of data provided by the content sector (business content but also user generated content) will have to be stored, manipulated and retrieved from any one in the easiest way and on any type of devices. In addition, content will become more and more heterogeneous due to the multiplication of formats used by the end user devices. This increasing complexity needs further research activities that are highlighted in this position paper.

This document addresses several aspects (but not all) of the Big and Open data domains and tries to provide a research agenda for the next years.

The NEM sector is vast and covers the entire data value chain, from creation, manipulation, distribution, search, and privacy; and these techniques are all very relevant for Big and Open Data. Big data technologies are nowadays mandatory to provide new forms of content in an ATAWAD (anytime, anyway, anydevices) seamless way, it enables providers to reach more and more people while optimizing the content storage in a sustainable and scalable way. Open data technologies are complementary as far as they allow organizations to offer to end users or to third parties the possibility to use and repurpose content for new usages and applications. It is very relevant for administrative content but also for any companies which could develop new business through data opening.

In addition to technical aspects, this paper also provides some useful input on societal impact mainly linked to privacy as far as content becomes more accessible to anyone when it is stored in the cloud. These data will be available to anyone unless we are able to provide mechanisms which offer to users the possibility to withdraw exhaustively their content and to provide secure data-tight clusters.

The last aspect addressed by that paper deals with business impact that is also a big deal for Big and Open Data. Many stakeholders are jumping on this new “gold rush” but as yet nobody knows exactly how to make successful business models in this new and evolving field. For Big Data, the objective is to optimize the data storage cost and exploitation while for Open data is to provide attractive information at a good price. Regulation has also a potentially huge and critical role in this domain and is still unstable, which brings additional difficulties and risks to users and content creators.

Finally, this paper proposes some research topics which should be addressed in the near future and for instance in the future Big Data Public Private Partnership that the European Commission is currently looking at the possiblilty of establishing (and which has not to be focus only on information technology data).

The NEM Initiative, (www.nem-initiative.org), is one of the recognized European Technology Platforms (ETP) of Horizon 2020. The NEM ETP aims at building sustainable European leadership in content, media, and the creative industries. With the launch of the Horizon 2020 programme, a renewed NEM platform will pursue its objective to promote an innovative European approach to convergent Media, Content and Creativity towards a Future Media Internet that will enhance the lives of European citizens through a richer media experience. The NEM constituency groups 900 members including all major European organisations working in the networked and electronic media area, comprising content providers, broadcasters, network equipment manufacturers, network operators and service providers, academia, standardisation bodies and government institutions. NEM delivers sustainable European leadership in the convergence of media, information and communication technologies, by leveraging the innovation chain to deliver rich user/citizen experiences and services using NEM technologies contributing to solve societal challenges.

Out of the several NEM strategic activities, Big Data and Open Data are one of the most relevant topics because it will be a key evolution in the NEM sector. That has been clearly identified through 2 dedicated workshops organized by NEM : at Future Internet Assembly (May 9 2013, Dublin) and at NEM summit 2013 (October 29 2013, Nantes) [see annex].

Contributors : Artur Krukowski (Intracom), Yiannis Kompatsiaris, Symeon Papadopoulos, Spiros Nikolopoulos (ITI), Francois Hanat, Nadia Echchihab (Cap Digital), Luis Rodero Merino (Gradiant), Ruben Casado (Treelogic), Pof. Dr. Katharina Morik (TU Dortmund University), Dimitrios Gunopoulos (University Athens), Carlos Bento, Rui Gomes (University Coimbra), Joachim Köhler (Fraunhofer/IAIS), Dev Audsin, Yves Raimond, Andy Bower (BBC), Petter Bae Brandtzæg (Sintef), Eric van Tol, Janienke Sturm (Fontys University), Hadmut Holken (Holken Consultants), Richard Jacobs (BT), Jean-Dominique Meunier (Technicolor), Jovanka Adzic (Telecom Italia).

I- Introduction

This paper provides a vision of Big and Open data domains from a “content and media” perspective. The idea is to draw a picture of the actual situation and to propose future evolutions from the NEM membership.

Big data is a collection of data sets so large and complex that it becomes difficult to process using hands-on database management tools or traditional data processing applications within a tolerable elapsed time; that is, when the size of the data becomes part of the problem itself The challenges include capture, curation, storage,[3] search, sharing, transfer, analysis,[4] and visualization. The trend to larger data sets is due to and should benefit from the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, enhance production and logistics, determine real-time roadway traffic conditions, and countless other applications.

Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "Open" movements such as open source, open hardware, open content, and open access. Open data should come from public administrations as well as industry.

The objective of this position paper is to define the view of the NEM community on these two complementary domains and to identify potential research topics in these fields.

This paper has the objective to establish the state of the art, to identify future evolution and to propose a set of future research activities that need to be studied from the NEM community perspective.

II- Challenges with BIG/OPEN DATA

Challenges with big-data:

As enormous volumes of data are being created every day, great interest has been placed on the theoretical and practical aspects of extracting knowledge from massive data sets (Bacardit and Llorμa, 2009), (Dean and Ghemawat 2008). For instance, big data streams coming from ubiquitous sensors are collected and aggregated in existing or newly emerging social networks creating a new generation of networked media. This information provides us with the opportunity to take decisions based on the data itself, rather than based on guesswork, or on artificially constructed models of reality. In other words Big Data carries the potential to become the main enabler of reality mining by driving nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. However, in achieving the ambitious objective of reality mining there are several challenges related to Big Data handling. Although the analysis intuitions behind big data are pretty much the same as in small data, having bigger data consequently requires new methods and tools for solving new problems, or solving the old problems in a much better way. Big data are characterized by their **variety** (i.e. multimodal nature of data ranging from very structured ones like ontologies to unstructured ones like sensor signals), their **velocity** (i.e. real-time and dynamic aspect of the data) and of course their **volume** (i.e. **we** used to speak in terms of megabytes and gigabytes of home storage – now we speak in terms of terabytes, while enterprises speak in terms of petabytes).

In using Big Data we may consider six different processing phases (Agrawal et al., 2012): a) *Data Acquisition and Recording***,** where we have to deal with the problem of filtering the redundant data without discarding useful information. b) *Information Extraction and Cleaning*, where we need to represent all data in a format suitable for analysis by our information extraction modules. c) *Data Integration, Aggregation, and Representation*, where differences in data structures and semantics need to be expressed in forms that are computer understandable. d) *Query Processing, Data Modeling, and Analysis***,** where declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments are applied on integrated, cleaned, trustworthy, and efficiently accessible data. e) *Interpretation*, where the analysis results are interpreted by decision-makers usually by examining all assumptions made and retracing the analysis. Based on the aforementioned phases we may identify the following challenges that underlie many, and sometimes all, of these phases (Agrawal et al., 2012):

Heterogeneity and Incompleteness that calls for novel data collection, fusion and aggregation methods able to combine vastly different modalities coming from different sources (Nikolopoulos et al. 2013; Lakka et al, 2011). Data analysis algorithms (e.g. machine learning) expect homogeneous and rather complete data. Indeed, structured data is typically required by many (traditional) data analysis algorithms since computers work more efficiently if they can store multiple items that are all identical in size and structure. However, since the less structured design is usually more effective for custom purposes, most real-world cases exhibit high levels of heterogeneity. As a consequence, data need to be carefully structured as a first step in (or prior to) data analysis. Similarly, the data stored in many real-world cases are incomplete or misleading, which may cause the data analysis algorithm to extract unreliable results. This incompleteness must be managed during data analysis and doing this correctly is particularly challenging especially when dealing with user-generated content (Chatzilari et al., 2012).

Scale that raises the need for scalable and efficient data stream mining, information extraction, indexing and retrieval approaches. Managing large and rapidly increasing volumes of data has always been amongst the most challenging goals of ICT systems. In the past, this challenge could be partly mitigated by throwing more hardware on the problem. Unfortunately, this is not the case today where data volume is scaling faster than computing resources, and processor

speeds are pretty much static. Nevertheless, there have been some developments lately that although promising, are not without their challenges.

Instead of having processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. Although the new technology is able to offer much more powerful systems, the challenge derives from the fact that the parallel data processing techniques that were applied in the past for processing data across nodes do not directly apply for intra-node parallelism, since the architecture is very different. Thus, we need to rethink how we design, build and operate data processing components.

At the same time, cloud computing is gradually becoming mainstream and is able to aggregate multiple disparate workloads with varying performance goals into very large clusters. However, this level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs. The challenge of system-driven holistic optimization is now becoming important.

Finally, the storage technology is also under a transformative change, moving away from the traditional I/O subsystems (i.e. hard disk drives (HDDs)) towards solid state drives (SSDs). SSDs have lower access time and less latency than HDDs but follow a different approach on how data is stored and retrieved. This poses the challenge of re-thinking how we design storage subsystems but also how we design various algorithms dealing, for instance, with query processing, query scheduling, database design, concurrency control methods and recovery methods.

Timeliness, that typically refers to the cases where the analysis outcome becomes useless unless delivered in a specific fragment of time. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. However, the full analysis of legacy data is rarely feasible in real-time and certain results will have to be pre-computed for meeting the real-time requirement. Thus, the challenge lies in designing methods where partial results can be computed in advance, so that only a small amount of incremental computation is needed when new data arrives. Indexing structures are typically the instrument employed to facilitate this goal when the searching criteria are specified. However, in the context of Big Data, new types of criteria may be specified or even change dynamically, calling for new index structures. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

Privacy, relating to the fact that given the nature of big data streams (i.e. usually generated by sensor-capable devices), the vast majority of potential services involve the storage and processing of users' private information. Moreover it is likely that, in many cases, the primary producers (i.e. the users of services and devices generating data) are unaware that they are doing so, and/or what it can be used for. For example, people routinely consent to the collection and use of web-generated data by simply ticking a box without fully realizing how their data might be used or misused. It is also unclear whether bloggers and Twitter users, for instance, actually consent to their data being analyzed. In addition, recent research showing that it has been possible to 'de-anonymise' previously anonymised datasets raises concerns. Finally, the privacy-related obstacles are magnified by the fragmentation of existing policies on handling private data and the absence of a clear Europe-wide directive on this issue. As a consequence users are reluctant to hand-over their personal information while companies and research organizations face many difficulties in collecting them at the appropriate scale.

Analytics-oriented challenges deriving from the need to work with new and high volume data sources. The question "what is the data really telling us?" is at the core of any social science

research and evidence-based policymaking, but there is a general perception that “new” digital data sources poses specific, more acute challenges. The challenges are intertwined and difficult to consider in isolation, but for the sake of clarity, they can be split into three distinct categories: (1) getting the picture right, i.e. summarizing the data (2) interpreting, or making sense of the data through inferences, and (3) defining and detecting anomalies (UN Global Pulse, 2012). The role of multimedia analysis is particularly critical in all aforementioned categories and initial efforts have already started to set the field and investigate potential solutions for certain problems (Nikolopoulos et al. 2013b). However, it may also be the case that the importance of analyzing big data has been underestimated by information technology experts. In order to meet these analytical challenges we need data scientists in addition to information scientists. Unfortunately this kind of scientist is difficult to find since there is no curriculum for this in educational institutions. In addition, large companies are often unwilling to open their data to the scientific community which makes the situation even worse. However, there is a reasonable expectation that since large industrials that own large datasets, will eventually need to hire employees that have experience with these data volumes, they have an interest to make available their datasets for research, and also put pressure on the educational institutions to include relevant courses in their curriculum. Similar attitude is also expected by the public bodies fostering research and innovation (i.e. European Commission) in offering the necessary resources for educating a new generation of scientists focused on data.

Agrawal, D., et al, White Paper: Challenges and Opportunities with Big Data, Feb 2012. (url: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf - last accessed Aug 2013)

Bacardit, J. and Llorμa, X. (2009) Large scale data mining using genetics-based machine learning. In GECCO '09: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference, pages 3381-3412, New York, NY, USA, 2009. ACM

Chatzilari, E., Nikolopoulos, S., Patras, I., and Kompatsiaris, I., (2012) Leveraging social media for scalable object detection, Pattern Recognition, Volume 45, Issue 8, August 2012, Pages 2962-2979,

Dean, J. and Ghemawa, S. (2008) Mapreduce: simplified data processing on large clusters. Comm ACM 51:1, pp 107-113, 2008

Lakka, C., Nikolopoulos, S., Varytimidis. C., and Kompatsiaris, I., (2011) A Bayesian network modeling approach for cross media analysis, Signal Processing: Image Communication 26 (2011), pages 175-193

Nikolopoulos, S, Zafeiriou, S., Patras, I., and Kompatsiaris, I. (2013a), High-order pLSA for indexing tagged images, Signal Processing Elsevier, Special Issue on Indexing of Large-Scale Multimedia Signals, Volume 93, Issue 8, August 2013, Pages 2212-2228

Nikolopoulos, S., Papadopoulos. S., and Kompatsiaris. Y., (2013b) Reality Mining in urban space, The Fourth International Conference on Information, Intelligence, Systems and Applications (IISA 2013), Workshop on Urban Computing & Modern Cities, Piraeus, Greece, July 10 – 12, 2013.

UN Global Pulse, White Paper: Big Data for Development – Challenges and Opportunities, May 2012, http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf - last accessed Aug 2013)

III- Technical approaches to Big Data

III-1. Introduction

Technology is ubiquitous and very much part of public and private organizations and individuals. People and things are becoming increasingly interconnected. Smartphones, buildings, cities, vehicles and other devices are filled with digital sensors, all of them creating evermore data. By 2013 numbers, everyday, we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone. The term *Big Data* applies to information that can't be processed using traditional processes or tools. Three characteristics define Big Data: *volume* (amount of data), *variety* (range of sources), and *velocity* (update time per new observation) - the "*3Vs*" model [1]. The term Big Data seems to indicate that the only challenge is the sheer size of information, but we must include the capture, storage, search, sharing, transfer, analysis and visualization of data.

Across industries and sectors (consumer goods, financial services, government, insurance, telecommunications, and more), companies are assessing how to manage their untapped information in an effort to find ways to make better decisions about their business. Big Data is expected to change the way things are done, how to gain insight, and how to make decisions. Science research, policy making and business opportunities are vast.

III-2. Technical elements

This section introduces the three main families of Big Data systems: batch oriented, real-time oriented, and hybrid.

Batch oriented systems

Currently batch big data analytics are applied to social networking applications, graph mining, scientific applications, etc. The main advance in the last decade is the development and widespread adoption of the MapReduce programming framework. It has several advantages: (i) allows a simple, unifying view of the data; (ii) it is inherently scalable; (iii) it effectively hides the complexity of programming distributed software, which is challenging due to potential hardware failures, fluctuations in network quality, device heterogeneity, etc. The MapReduce programming model supports the weak connectivity model of computations across open networks - such as mobile networks - which makes it very appropriate to use in a mobile setting. MapReduce also introduces some limitations or constraints in certain analysis settings that must be addressed in future work: (i) many analysis and mining tasks in real systems or applications have to run iteratively or on multiple rounds; this is difficult to do in MapReduce. Several recent implementations try to address this shortcoming; (ii) additional development for real-time and streaming computation, as well as optimisation for data access and indexing is required for efficient data analysis.

Real-time oriented systems

Several applications require real-time processing of data streams from heterogeneous sources, in contrast with the batch (unbounded in terms of time) approach of MapReduce. Some areas of importance where this is required are

- Smart cities, to organize transportation, energy supply, garbage collection, cultural tasks…

- Disaster management, especially through data gathered from citizens' usage of social networks. The EU INSIGHT[1] project combines twitter streams with other data for early warning of disasters
- Production and logistics, to use factories' sensors for quality control and resources saving; the application of real-time data allows for analytics that forecast the outcome of the production to correct it when needed
- Science, where it is applied in a broad set of research areas from biomedical research (e.g. genomics) to physics (being CERN a well-known example).
- Entertainment, where streaming data from music, TV and gaming platforms are to be analyzed for recommendations, analysis of users, and advertisement placement.

Hybrid systems (hybrid model)

Batch processing provides performance benefits since it can use more data and, for example, perform better training of predictive models. However it is not suitable for domains where a low response time is critical. Real time processing solves this issue, but the analyzed information is limited in order to achieve low latency. Many domains require the benefit of both batch and real time processing approaches so an hybrid model is needed. The architectural principles/requirements for such hybrid model (also known as *Lambda Architecture* [2]) are:

- Robustness the system has to be able to manage (hardware and human) errors
- Data immutability raw data is stored forever and it is never modified
- Recomputation results always can be obtained by (re)-computing the stored raw data

This requirements are implemented by a three-layer (four for some authors) architecture:

- Batch layer It contains the immutable, constantly growing master dataset stored on a distributed file system. With batch processing arbitrary *batch views* are computed from this raw dataset
- Serving layer It loads and exposes the *batch views* in a datastore so that they can be queried
- Speed layer It deals only with new data and compensates for the high latency updates of the serving layer. It computes the *real-time views*.

To obtain a complete result, the batch and real-time views must be queried and the results merged together. Synchronization, results composition and other non-trivial issues have to be addressed at this stage in which could be considered the **Combination layer**.

III-3. Available solutions

This section lists several Big Data solutions, grouped as in the previous section.

Batch oriented systems

Apache's Hadoop[2] framework is currently the industry-standard Big Data solution for batch processing. Hadoop is an open source Java-based system that supports data-intensive distributed applications. It has two main components: the Hadoop Distributed File System (HDFS) and the MapReduce Software Framework. Other MapReduce implementations include Disco [3] and Mars [4] which target more specialised environments. The recently developed Misco system [5] is a MapReduce framework tailored for cell phones and mobile devices. New research initiatives attempt to extend the MapReduce framework to target more complex computations which involve iterations. Most implementations of MapReduce, including Hadoop, have to create a number of successive MapReduce tasks to implement iterative computations. This forces applications to reload the associated data from disk, which often results in high

[1] http://www.insight-ict.eu/
[2] http://hadoop.apache.org/

execution times. New frameworks (e.g. Twister[3], Spark[4]) make a first attempt to address the problem of providing higher-level programming functionality in MapReduce by keeping data and programs in memory between iterations.

Real-time oriented systems

Several frameworks (Storm[5] and S4[6] being the most well-known) for the construction of systems for the real-time analysis of data are being actively developed. They are quite similar in the features they offer: easy construction of distributed data processing pipes, tolerance to failures, etc. On the other hand, present research is focusing on (i) moving code close to the data stream: feature extraction, summaries, sketches… from the data stream; (ii) combining different streams for analysis: distributed clustering, probabilistic models that cope with large volume and variety; (iii) methods that combine learned models and apply them to a stream [6].

Developing fast algorithms that directly learn from streaming data has been started by approaches like lossy counting [7] and clustering data streams [8]. Many applications require algorithms that exploit data as they arrive in real-time. Algorithms that process streaming data are on demand for all kinds of sensor measurements. The measurements from *mobile applications* (e.g., smart phones) characterize customer behavior, media use, or movement patterns. Similarly, measurements from *industrial production* are most often stored and analyzed in traditional computing environments but need to be applied in real-time [9]. *Smart cities and logistics* share the subject of resource-aware transportation: trucks, cars, or trains can be controlled to save energy using streaming data from in-vehicles sensors [10]. Finally, the search for *abnormalities*, *rare events*, and *early warning of disasters* combines diverse streams on the fly.

Hybrid systems (hybrid model)

The last generation of Big Data systems, where both batch and real-time approaches are combined, is just starting. This hybrid approach has practical problems to address:

- Two sets of aggregation logic have to be kept in sync in two different systems
- Keys and values must be serialized consistently between each system and the client
- The client is responsible for reading both datastores, aggregating data and serving the combined results

New software frameworks have to be developed to provide a general solution to these problems. Some promising technologies are *Lambdoop* [11] and *SummingBird* [12]. *Lambdoop* is a framework for easing developing Big Data applications by combining real time and batch processing approaches. It implements a Lambda based architecture that provides an abstraction layer (Java based API) to the developers. *SummingBird* defines a unique MapReduce where jobs can be run in batch, real-time or hybrid mode as the developer requires.

III-4. Future directions

Big Data systems have grown in importance, and all kinds of private and public organizations are increasingly aware of the potential benefits of Big Data as an enabler to exploit their (potentially vast) data. The IT industry has reacted by investing huge efforts in Big Data systems, however their limitations are becoming more and more evident. From a technical point of view, the future of Big Data will be shaped by the new solutions that deal with these limitations:

[3] http://www.iterativemapreduce.org/
[4] http://spark-project.org/
[5] http://storm-project.net/
[6] http://incubator.apache.org/s4/

- New systems that enable the analysis of both structured and unstructured data to be combined, i.e. able to combine multiple data sources (from social media to data warehouses) in a way that is manageable, not only for the professionals, but also more non-professional users and groups.
- New embedded analytics that exploits the streams of data in real time under strict resource restrictions of computing capacity, storage, energy and communication bandwidth.
- New paradigms that super-seed the 'pure batch' and 'pure real-time' approach of present Big Data tools.
- New application frameworks able to squeeze all distributed computing resources, allowing to run different types of tasks (batch, stream analysis, interactive) virtualizing all the underlying infrastructure and scheduling usage depending on the task requirements.
- New database systems able to handle huge datasets while keeping the transactional semantics of data operations available in traditional relational databases.
- New big data tools that are guiding and managing ethical, security and privacy issues in big data research.

---

[1] Gartner. *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. 2011; Available from: http://www.gartner.com/newsroom/id/1731916.
[2] N. Marz and J. Warren. *Big Data: Principles and best practices of scalable real-time data systems*. Manning Publication Co. 2013, ISBN: 9781617290343
[3] B. He, W. Fang, Q. Luo, N. K. Govindaraju, and T. Wang. Mars: a mapreduce framework on graphics processors. In PACT, ON, Canada, Oct 2008.
[4] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis. Evaluating mapreduce for multi-core and multiprocessor systems. HPCA, 2007.
[5] V. Tuulos. Disco. http://discoproject.org/
[6] Piatkowski, Nico and Lee, Sangkyun and Morik, Katharina (2013): Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation, Machine Learning Journal, Vol. 93, No. 1, p. 115 -- 139
[7] Manku, Gurmeet Singh and Motwani, Rajeev (2002): Approximate frequency counts over data streams, VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases
[8] Guha, Sudipto and Meyerson, Adam and Mishra, Nina and Motwani, Rajeev and O'Callaghan, Liadan (2003): Clustering Data Streams: Theory and Practice, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 3, p. 515 -- 528
[9] Uebbe, Norbert and Odenthal, Hans-Juergen and Schlueter, Jochen and Blom, Hendrik and Morik, Katharina (2013): A novel data-driven prediction model for BOF endpoint, The Iron and Steel Technology Conference and Exposition in Pittsburgh (AIST)
[10] Hillol Kargupta and Kakali Sarkar and Michael Gilligan (2010): MineFleet: an overview of a widely adopted distributed vehicle performance data mining system, Procs of the 16th ACM KDD
[11] R. Casado et al. *Lambdoop, a framework for easy development of Big Data applications*. NoSQL matters 2013 Conference, Barcelona, Spain. http://2013.nosql-matters.org/bcn/abstracts/
[12] S. Ritchie et al. *Streaming MapReduce with Summingbird*. 2013,

IV- Technical approaches to Open Data

IV-1. Introduction

Integrating and analyzing large amounts of data plays an increasingly important role in today's society. Often, however, new discoveries and insights can only be attained by integrating information from dispersed sources. Despite recent advances in structured data publishing on the Web (such as RDFa and the schema.org initiative) the question arises how larger Open Data sets can be published, described in order to make them easily discoverable and facilitate the integration as well as analysis.

One approach to address this problem are Open Data portals, which enable organizations to upload and describe datasets using comprehensive meta-data schemes. Similar to digital libraries, networks of such data catalogues can support the description, archiving and discovery of datasets on the Web. Recently, we have seen a rapid growth of data catalogues being made available on the Web. The data catalogue registry datacatalogs.org, for example, lists already 285 data catalogues worldwide. Examples, for the increasing popularity of data catalogues are Open Government Data portals, data portals of international organizations and NGOs as well as scientific data portals.

Also in the research domain, data portals can play an important role. Almost every researcher works with data. However, quite often only the results of analysing the data are published and archived. The original data, that is ground truth, is often not publicly available thus hindering repeatability, reuse as well as repurposing and consequently preventing science to be as efficient, transparent and effective as it could be. Some examples of popular scientific open data portals are the Global Biodiversity Information Facility Data Portal [1]. Also many international and non-governmental organizations operate data portals such as the World Bank Data portal [2] or the data portal of the World Health Organization[3].

Despite being a relatively new type of information system first commercial (e.g. Socrata[4]) and open-source data portal implementations (e.g. CKAN[5]) are already available.

---

[1] http://data.gbif.org

[2] http://data.worldbank.org

[3] http://www.who.int/research

[4] http://www.socrata.com/

[5] http://ckan.org

IV-2. Overview

Computers and the Internet are vital to modern day societies. The set of values fundamental to the development of the personal computer includes the idea that all information should be free or at least easily accessible and that computers can make the world a better place. Open data, in the same spirit as other "open" movements, like open software, can be freely used, reused and redistributed by anyone. Governments, public agencies, organizations and individuals can benefit on many levels from the availability of data. Unfortunately, there are still some issues preventing institutions from opening up their data: copyrights, fear of loss of control and lack of resources needed to open data, for instance [1]. Presently, technology is not seen as a major

challenge to open data. Political, organizational, financial and privacy issues are more of a concern [2].

Open data needs to be technically open as well as legally open. The key aspect is interoperability. This interoperability is key to dramatically enhanced ability to combine different data sets and thus to develop more and better products and services. But, open data and open knowledge are not only about availability. They're also about making data understandable and useful. The question arises as to how larger Open Data sets can be published and described in order to make them easily discoverable and facilitate the integration as well as analysis. A data portal is a Web based information system, that comprehensively collects, manages and preserves for the long depth of time machine-readable structured datasets as well as associated meta-data, and offers to its user communities specialized functionality on the datasets, of defined quality and according to comprehensive codified policies. We can roughly distinguish the following classes of data portals: (1) Data repository; (2) Data catalogue; (3) Comprehensive data portal.

Open Data from the public sector has been growing rapidly in many countries, with many governments actively opening their data sets [3]. The world is experiencing a new democracy-phenomenon, powered partly by the proliferation of social media, real-time, mobile and ubiquitous means of communication and the availability of uncensored information [1]. Open government data can help you to make better decisions in your own life, or enable you to be more active in society.

Open data in media has been supported by the EU for instance with the project Vista-TV http://vista-tv.eu/. It delivers real-time recommendation of media using the streams framework (see real-time processing above) and shows in real time the current numbers and some features of users watching a programme in internet television. An enrichment engine brings information from electronic programme guides and internet news together with real time image features of the shows. The enriched information is open linked data which is archived for further analysis.

In Europe, open data movement origins lie in the issuance of Directive 2003/98/EC, recently amended by Directive 2013/37/EU, of the European Parliament and the Council on access to and reuse of public sector information [4]. In 2004, the Science Ministers of all nations of the OECD signed a declaration, which, essentially, stated that all publicly funded archive data should be made publicly available. In 2009, the Obama administration gave also a big boost to the Open Data cause [5]. A few months after Data.gov from the US Government went live in May 2009, the UK followed by launching data.gov.uk in September 2009. The UK and the US Open Data Portal initiatives have been copied by many international, national and local public bodies, such as the World Bank [6]. A list of over 200 local, regional and national open data catalogues is available on the open source datacatalogs.org project, which aims to be a comprehensive list of data catalogues from around the world – prominent examples include the European Commission Data Portal (http://open-data.europa.eu/).

Along with the growth of the open data movement, there is a demand for smarter cities that are open for their smart citizens to contribute to their own living environment, in political decision making or developing new apps and digital public services, using not only static open data but also using dynamic data from a myriad of sensors spread over the cities, web cams and so on (see, for instance, [7] or [8]. The recent trend for apps contests in many cities around the world is a proof of this - two examples are the World Bank Apps contest [9] and NYC BigApps contest [10].

New mobility patterns are being induced by the increasing use and social integration of new technologies. The paradigm is shifting to more user-oriented transport services and open data offers an important source of information to promote sustainable mobility, in the sense that (a)

operators can better understand the mobility needs of different users and (2) users can make better/informed trip decisions based on up-to-date analysis of real-time data sets. This improves mobility and accessibility and, in the end, drives better and more sustainable transport systems.

Open data is also changing the way education can be seen. Open licensed educational resources, open courseware and online authoring environments enable new educational environments [11].

Economically, open data is of great importance as well. Several studies have estimated the economic value of open data at several tens of billions of Euros annually in the EU alone [12]. New products and companies are re-using open data in a wide spectrum of different areas, such as health, transportation, tourism, information services, research, marketing, and so on [1].

For instance, we are seeing an unprecedented increase in the time people spend interacting with machines and computational systems. These interactions take the form of time spent on social networking websites like Facebook, activities on search engines such as Bing or Google, or actions performed on mobile devices, and produce a lot of data about the user. This data can be leveraged through data analysis to not only understand the behaviour and preferences of users, but to also build intelligent interactive systems that are more effective and easy to use.

---

1. Pollock, R., et al., *The Open Book*, ed. J. Nissilä, K. Braybrooke, and T. Vuorikivi2013: The Finnish Institute in London.

2. Janssen, M., Y. Charalabidis, and A. Zuiderwijk, *Benefits, Adoption Barriers and Myths of Open Data and Open Government.* Information Systems Management, 2012. **29**(4): p. 258-268.

3. Goda, S., *Open data and open government.* INFORMACIOS TARSADALOM, 2011. **11**(1-4): p. 1.

4. Ferrer-Sapena, A., F. Peset, and Aleixandre-Benavent, *Access to and reuse of public data: open data and open government.* El Profesional de la Información, 2011. **20**(3): p. 260-269.

5. Peled, A., *When transparency and collaboration collide: The USA Open Data program.* Journal of the American Society for Information Science and Technology, 2011. **62**(11): p. 2085-2094.

6. de Vries, M., et al., *POPSIS - Pricing Of Public Sector Information Study - Open Data Portals* 2011, European Commission.

7. Desouza, K.C. and A. Bhagwatwar, *Citizen Apps to Solve Complex Urban Problems.* Journal of Urban Technology, 2012. **19**(3): p. 107-136.

8. Robinson, R., et al., *Street Computing: Towards an Integrated Open Data Application Programming Interface (API) for Cities.* Journal of Urban Technology, 2012. **19**(2): p. 1-23.

9. WorldBank. *World Bank Apps contest*. 2013 [cited 2013 01-09]; Available from: http://www.worldbank.org/appsfordevelopment.

10. NYC. *NYC BigApps*. 2013 [cited 2013 01-09]; Available from: http://www.worldbank.org/appsfordevelopment.

11. Tellez, A.G. *Authoring Multimedia Learning Material Using Open Standards and Free Software*. in *Multimedia Workshops, 2007. ISMW '07*. 2007.

12. OpenKnowledgeFoundation. *The Open Data Handbook*. [cited 2013 01-09]; Available from: http://opendatahandbook.org/en/index.html.

IV-3. Technical elements

The typical open data system contains the following data processing components:

Metadata. A core functionality of a data catalogue is to support the management of meta-data associated with the datasets being hosted or described in the catalogue. The metadata includes general dataset metadata (e.g. contact, creator, licensing, classification information), access metadata (e.g. dump download, webservice/SPARQL/API access to the datasets or URI/search lookup endpoints), structural metadata (e.g. vocabularies, classes, properties used, statistics and links to other datasets). The Data Catalogue Vocabulary (DCAT)[1] provides a high-level vocabulary for describing dataset catalogues, while the VoiD Vocabulary[2] can be used for describing individual datasets.

Archiving and Repository. Providing a repository for archiving and long term preservation is an important function of a data portal which, due to the resource requirements (in terms of storage space and bandwidth), is not always available.

Data integration. Many research questions can only be answered by integrating data from different sources. For example, in order to assess the impact of individual transport on air pollution, one might want to integrate a dataset published by a local authority on car traffic and congestion with a dataset published by an environmental agency with pollution measurements. To facilitate the integration, it is of paramount importance, that not only meta-data is attached to the dataset, but that also the semantic structure of a dataset is made explicit. Our experience with the datasets published at 30 European data portals, which are aggregated by publicdata.eu ist that the vast majority (>80%) of the datasets are tabular data (i.e. CSV files, Excel sheets) and only a small fraction (approx. 1%) are adhering to vocabularies and are represented in RDF.

Data Analytics. Open data is accumulating at such a rate that there are no longer enough qualified humans to analyse it. Data Mining is needed to make open data useful in essentially all sectors which are drowning in it. That is, we have to discover patterns and regularities in the big open data for instance by classification, association rule mining, subgroup discovery, graph mining, and clustering. However, the data is too big to cope with it without using new mining algorithms or technologies. A typical strategy for dealing with big data is sampling. Here, we assume that we can obtain an approximate solution using a subset of the examples only. Alternatively, we can use probabilistic techniques and model that quantify our uncertainty by estimating a distribution over the quantity of interest. Online approaches do not consider the data as a single a batch but cycle through it making sequential updates based on mini-batches of the data. Streaming learners, as already discussed above, incorporate examples as they arrive. And, one can use distributed approaches that rapidly process big data in parallel on large clusters of compute nodes.

Visualization. The potential of the vast amount of data published through Data portals is enormous but in most cases it is very difficult and cumbersome for users to visualize, explore and use this data, especially for lay-users without experience with Semantic Web technologies. Compared to prior information visualization strategies, there is a unique opportunity when datasets are semantically represented and described. The unified RDF data model enables to bind data to visualizations in an unforeseen and dynamic way. An information visualization technique requires certain data structures to be present. Data structures can be automatically derived and generated from reused vocabularies or semantic representations.

Quality. Data portals should play a key role in ensuring data quality. Datasets being available via data catalogues already cover a diverse set of domains. Specifically, biological and health care data is available as in a great variety covering areas such as drugs, clinical trials, proteins, and diseases. However, data on the Web also reveals a large variation in data quality. For example, data extracted from semi-structured or even unstructured sources, such as DBpedia, often contains inconsistencies as well as misrepresented and incomplete information.

Data quality is commonly conceived as fitness for use for a certain application or use case. However, even datasets with quality problems might be useful for certain applications, as long as the quality is in the required range. For example, in the case of DBpedia the data quality is perfectly sufficient for enriching Web search with facts or suggestions about common sense information, such as entertainment topics. In such a scenario, DBpedia can be used to show related movies and personal information, when a user searches for an actor. In this case it is rather neglectable, when in relatively few cases, a related movie or some personal fact is missing. For developing a medical application, on the other hand, the quality of DBpedia is probably completely insufficient. It should be noted that even the traditional, document-oriented Web has content of varying quality and is still perceived to be extremely useful by most people. Consequently, a key challenge is to determine the quality of datasets published on the Web and make this quality information explicit. Assuring data quality is particularly a challenge for federated data catalogues as it involves a set of autonomously evolving data sources. Other than on the document Web, where information quality can be only indirectly (e.g. via page rank) or vaguely defined, there are much more concrete and measurable data quality indicators available for structured information. Such data quality indicators include correctness of facts, adequacy of semantic representation or degree of coverage.

There are already many methodologies and frameworks available for assessing data quality, all addressing different aspects of this task by proposing appropriate methodologies, measures and tools. In particular, the database community has developed a number of approaches such as DQA, AIMQ, TDQM and CDQ. However, data quality on the Web of Data also includes a number of novel aspects, such as coherence via links to external datasets, data representation quality or consistency with regard to implicit information. Furthermore, inference mechanisms for knowledge representation formalisms on the web, such as OWL, usually follow an open world assumption, whereas databases usually adopt closed world semantics. Despite quality in data catalogs being an essential concept, few efforts are currently in place to standardize how quality tracking and assurance should be implemented. Moreover, there is no consensus on how the data quality dimensions and metrics should be defined.

---

[1] http://www.w3.org/TR/vocab-dcat/

[2] http://www.w3.org/TR/void/

[3] http://aksw.org/Projects/Sparqlify.html

IV-4. Available solutions

Governments and public administrations started to publish large amounts of structured data on the Web, mostly in the form of tabular data such as CSV files or Excel sheets. Examples are the US' data portal data.gov, the UK's data portal data.gov.uk, the data portal of Germany www.govdata.de, the European Commission's open-data.europa.eu portal as well as numerous other local, regional and national data portal initiatives. All of these data portals contain collections of data sets catalogued by specific topics (e.g. environmental data, transportation

data, etc.). Additionally these data hubs contain visualization applications to browse and explore the huge amount of data sets.

On the tool and application level …..

With CubeViz and SemMap two visualization widgets for different types of data are developed:

- *CubeViz* can visualize any statistical dataset, which adheres to the DataCube vocabulary. CubeViz allows users to select certain slices of the data to be visualized and generates various charts and diagrams.
- *SemMap* can be used for visualizing any kind of spatial data comprising longitude and latitude properties on a map. If the dataset additionally contains a taxonomy or class hierarchy for organizing entities, SemMap automatically adds faceted filtering options to the map.

In the area of media the approach of data journalism could be considered as open data application. The main idea of data journalism is to exploit open data for news generation and statistical based story telling. Some examples are the Guardian, …, …

---

[1] http://open-data.europa.eu

IV-5. Future directions:

We deem Open Data to be only the beginning of an era, where data portals are evolving into comprehensive distributed and federated knowledge hubs, which expose and connect the structured data fueling our information society. In the following, we describe our vision of how Open Data portals can facilitate three scenarios related to enterprises, science and society.

Data portals as enterprise knowledge hubs. While business-critical information is often already gathered in integrated information systems such as ERP, CRM and SCM systems, the integration of these systems itself as well as the integration with the abundance of other information sources is still a major challenge. Large companies often have hundreds or even thousands of different information systems and databases. After the arrival and proliferation of IT in large enterprises, there were various approaches, techniques and methods that tackled the data integration challenge. In the last decade, the prevalent data integration approaches were primarily based on XML, Web Services and Service Oriented Architectures (SOA). XML defines a standard syntax for data representation, Web Services provide data exchange protocols and SOA is a holistic approach for distributed systems architecture and communication. However, we become increasingly aware that these technologies are not sufficient to ultimately solve the data integration in large enterprises. In particular, the overhead associated with SOA is still too high for rapid and flexible data integration, which is a pre-requisite in the dynamic world of today's large enterprises. It can be argued that classic SOA architectures are well-suited for transaction processing, but more efficient technologies are available and can be deployed for integrating data. A promising approach is the use of semantics-based data portals for describing and integrating enterprise data. Similarly, as the data web emerged complementing the document web, data intranets can complement the intranets and SOA landscapes currently found in large enterprises. Ultimately, enterprises can establish reference data portals as hubs and crystallization points for the vast amounts of structured enterprise data and knowledge enabling to establish a Data Intranet extending existing Document-oriented Intranets.

Primary source providing ground truth. Data portals enable direct linking to the ground truth data for secondary (e.g. scientific publications) or tertiary (e.g. encyclopedias) sources. This enables

improved provenance tracking in those sources. It also allows automatic syndication of the data using SPARQL or simple REST queries, which enables a simpler verification of statements compared to the manual work, which would be necessary without Linked Data.

Mediatization of linked data: Currently most data hubs are pure text based or contain structured metadata. For the media world the availability of multimedia data hubs become more and more important. Hyper-Video-Linking and annotated and open image and multimedia libraries and repositories should be made available. This will enhance the functional spectrum of media applications in the area of media search, navigation and exploration (e.g. open media archives).

Declarative Mining and probabilistic programming: The quality of insights gained from open data is not only restricted by the size of the data but also by the complexity of the model we assume: a simple model can only encode simple patterns, more complex patterns require more complex models. And complex models are required since open data is actually structured: there are not only sets of data instances, but also observed relationships (e.g., hyperlinks) that naturally encode statistical dependencies among the data instances. Such open-link data abound from social networks, the World Wide Web, protein interaction networks, among others. And, even as mining technology is accelerating and can deal with big and structured data and patterns, every new application requires a Herculean effort. Teams of hard-to-find experts must build expensive, customized tools. Open data calls for a new mining methodology that supports the construction of integrated models across a wide variety of domains and tool types, reduces development time and cost to encourage experimentation even by non-experts, and in turn greatly increases the number of people who can successfully build mining applications respectively make data mining experts more effective. In other words, declarative data mining and probabilistic programming approaches have to be developed that separate the model and the solver.

V- Social impact

V-1. Current status

Social impact of big data

Social impact now:

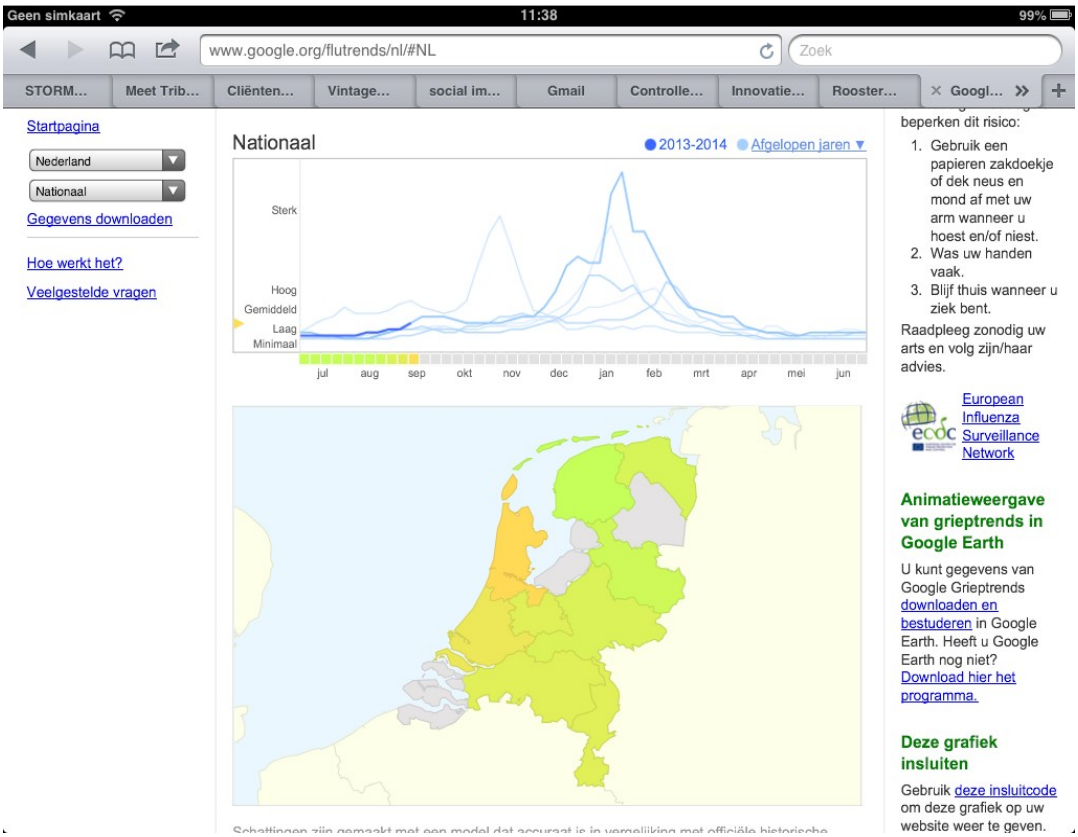
Big data and employment

Whether the rapid developments in big data will lead to more employment is topic of debate. Some companies forecast enormous growth in employment. Gartner for instance says that by 2015, 4.4 million IT jobs globally will be created to support big data, CEBR states that big data has the potential to add £216 billion ($327 billion) and 58,000 jobs to the UK economy by 2017 (Source: CEBR (Centre for Economics and Business Research) and supplier SAS), and McKinsey says that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired. Others state that increased productivity will not automatically lead to job growth, because digital technologies will take over human workers' jobs (Brynjolffson & McAfee). They show that while in the UK GDP has risen, median income has not. The counterargument, however, is that the money that a company saves through automation will eventually be fed back into the economy through lower prices, higher wages or more profit, which in turn will lead to other companies hiring more workers.(Robert D. Atkinson)

Big data and services

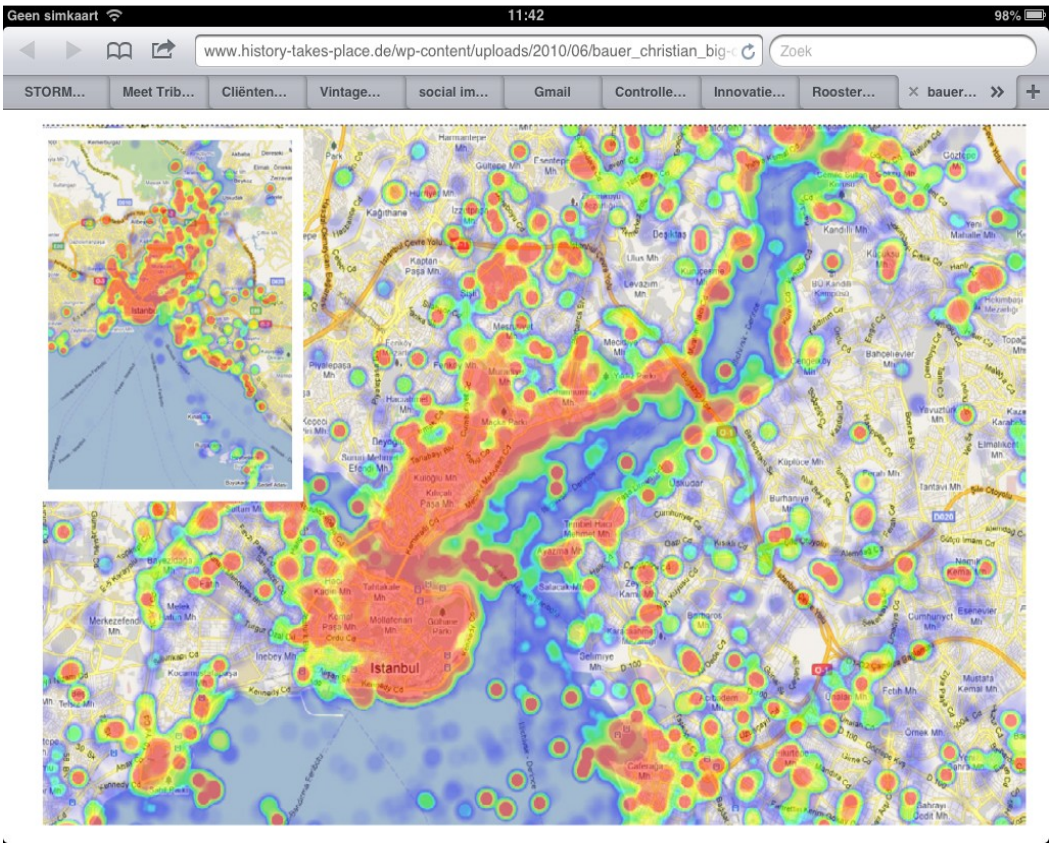
Large amounts of personal data, such as our browsing history, the hash tags we use in our tweets and the content of the emails we receive, contain a wealth of information for companies and marketeers to use in their sales and marketing strategies. Companies define consumer profiles based on this data, which are then used to determine which marketing strategy works best. For instance, people who conform to what other people do, may be persuaded to buy a certain book in an online bookstore because other people also bought the book. People who are more susceptible to the opinion of authorities may be persuaded by offering positive reviews by experts in the field (Kaptein etc.).

Big data is also applied to increase the quality of health care. Medical databases store a myriad of information about patients, ranging from information about the physical condition (heart scans, CT scans, X-rays) to information about their mental state (how happy is this patient) and data about their lifestyle (how many cigarettes does this person smoke a day). Using smart algorithms, data that is traditionally stored in different databases can now be collected, stored and combined in a secure way. Such clinical support systems allow practitioners and specialists to make more accurate diagnoses and informed decisions about possible treatments, for instance because they can look up how often a certain treatment has been effective with other patients.



Another use of big data is so called predictive analytics in healthcare. Some time ago researchers discovered that by analysing Google search terms they could predict a flu epidemic because they would see a large increase in the number of times search queries like “flu symptoms” and “flu treatment” would be used. By carefully analyzing search data in Google an epidemic could be detected much sooner than when using information provides by general practitioners and emergency rooms (http://www.google.org/flutrends). In another project, big data is used to predict suicide among veterans of the Vietnam War. Suicide rates are remarkably high among Vietnam veterans. In Durkheim research the Facebook and Twitter behaviour of Vietnam veterans is analysed by a machine-learning algorithm which has been fed with linguistic cues associated with suicide. Eventually the goal is to be able to prevent suicide and by the timely assistance of a psychologist (http://www.fastcolabs.com/3014191/this-may-be-the-most-vital-use-of-big-data-weve-ever-seen). Surveillance systems like these can help to detect early warning signs to put assistance programs into action timely and prevent damage.

Finally, big data can be used for good. For instance, to provide better help to people in crisis situations. An example is the language translation system researchers in Microsoft Research built for aid relief workers in Haiti after the 2010 earthquake. They built a statistical machine translation engine to translate Haitian Creole to English from scratch in under five days and delivered to aid workers in Haiti. Also, after the flood in Haiti, based on location data of SIM cards, the destination of 600.000 refugees was determined and communicated to government and humanitarian organisations so that they would know where to provide and coordinate help and relief. This shows how big data can support and enhance disaster and emergency response. (Big data, big impact whitepaper).



Big data is commonly used by companies to personalise their service to target customers. Information services that empower individuals to make informed choices are still scarce, are starting to appear. Toyota offer a real time traffic information service, which utilizes data such as vehicle locations and speeds, road conditions, and other parameters (http://www.greencarcongress.com/2013/05/tmc-201230529.html) to provide their customers with up-to-date traffic information. Doctors for Sale is an interactive map that shows all the Danish doctors who are sponsored by the medical industry. The information is designed to inform patients about which of their local doctors might be prescribing drugs based on their sponsorship rather than based on strict medical reasons or prescribing the cheapest drugs available. (http://googlemapsmania.blogspot.de/2012/04/danish-data-journalism-award.html. Other examples are maps showing average rental prices of almost every home in the USA (http://www.dataweek.co/2012-sf/index/reportdetail/report/96) and visualisations of how a city like Istanbul is viewed by tourists and by locals and heat maps of touristy spots in Istanbul (http://www.history-takes-place.de/wp-content/uploads/2010/06/bauer_christian_big-data-visualization-reader-mq.pdf).

Social impact: the future

The impact of Big Data is considered to be considerable, both for industry (IBM, 2012), innovation and research. Accoding to Kolb (2013) big data will be a key enabler of novel research insights, new theoretical approaches and research questions. Big data are also suggested to have a large impact on society in general, as the previous paragraphs show, both in terms of employment and in terms of services in various application domains. Society sees a transformational data deluge from which new scientific, economic and social value can be extracted (Mayer-Schonberger & Cukier, 2013). However the advancement and use of big data that affects societies needs to be carefully regulated. Hench, there is still a lot of debate on how big data should be collected, interpreted and managed to be used for social and scientific purposes, accurate quantitative research and representativeness (Kolb, 2013). Boyd and Crawford (2012) define six weaknesses in the current state-of-the-art of big data which should be guiding when referring to the social impact:

Definition of knowledge: Boyd and Crawford criticize the assumption that Big Data changes the definition of knowledge and argue for a continued need for conventional scientific methods. Hence, we will still need

to develop methods for big data that complement conventional social science methods, rather than serve as an complete alternative.

Objectivity and accuracy: The objectivity and accuracy of Big Data findings cannot be taken for granted. Bias may be introduced, for example, in pre-processing and interpretation; correlations may be spurious. A key concern is, therefore, rigorous method development related to big data.

Bigger data are not always better data: Big data, as "small data", may risk sampling bias. For example, Big Data sets of Twitter tweets only contain data for the tweeting public. Also, combining Big Data sets into still larger sets may threaten data quality due to a combining of errors from the individual data sets. Furthermore, current big data research tends to place a huge value on quantitative results, while devaluing the importance of qualitative data analysis on small data sets.

Context: Big data analysis depends on an understanding of the context where the data was gathered. This is for example seen in social network analysis. Historically, researchers tended to focus on an individual's personal network. Today, big data on individuals' networks includes personal contacts, acquaintances, co-workers, strangers and people who are brought together via communication channels, proximity analyses, and social media interactions. Guidelines for analysing big data with respect to context should therefore be established as part of future methods.

Limited access: Limited and difficult access to big data creates new digital divides and knowledge barriers for not only researchers, companies and people in general. Easy access to big data is in many cases restricted to commercial owners of online infrastructure, such as Google or Amazon (Anderson & Rainie, 2012). Furthermore use of big data may depend on storage and analysis capabilities. Big data-sets are increasingly available, but they may be difficult to access and make sense of in new domains, such as social science. As a response to this, future research should develop methods and tools to collect, analyse and visualize big data in the context of new and more democratized domains. In particular, the methods should target ordinary users but also the industry and academic research communities.

Ethical and privacy issues: Just because it is accessible does not make it ethical. Some big data sets, for example from social networks, may be difficult to make fully anonymous. Without explicit consent from individuals in the big data set, the ethics of big data analysis may be questionable. The future stakeholders of big data should investigate and establish guidelines for managing ethical and privacy issues in big data research.

Anderson, J. Q., & Rainie, H. (2012). *The Future of Big Data*. Pew Internet & American Life Project.

Kolb, J. (2013). *Secrets of the Big Data Revolution*. Applied Data Labs Inc. Boston.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society 15*(5), 662-679.

Mayer-Schonberger, V & Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think.* John Murray Publishers Ltd, New York.

The collection and analysis of personal data raises concerns in terms of ownership, privacy, and control. Privacy issues are involved when new technologies such as smart phones and Google glass are used to record video. All too often bystanders who appear in these clips are not aware that they are being filmed, let alone that they have given permission. In general, most people are not aware of how much personal data about them is collected and stored. All posts ever put on Facebook, all Twitter messages and all search queries entered are stored in big datasets. Also when you shop and pay through a debit card, or if you show your loyalty card, data are added to the big data collection. Most people do not realize that their personal digital footprint is gigantic and that it cannot be erased.

People have lost control over the data. The standard methods that Internet users deploy to ensure that their data are stored in an anonymous way are no longer sufficient. In the near future, smart algorithms will be able to identify users on not on the basis of identification data, but through their behaviour on the Internet or their interaction with technological artifacts (Eva Galperin). Also, companies like Google and Facebook influence the data that we get to see based on our (online) behavior. Two people searching the

web for the exact same information on Google may get different results, since Google looks at what you are doing, what computer you are using and what your browsing history is, to tailor the search results that you will be presented with. News sites also personalize their news pages. Increasingly, what people get to see is what companies think they want to see. According to Eli Pariser, this causes us to live in a filter bubble, with companies deciding what is going into the bubble and what is filtered out (Eli Pariser Ted Talk).

Blind belief in data-driven decision-making against uncontrollable complexity

Now that smart algorithms have been developed that can combine different data sets and find unpredicted correlations, we need to be aware of the danger of data-dogmatism and data dictatorship, which people are susceptible to. We tend to believe what the analysis of data shows us, even if we have reasonable grounds to suspect that something is not right. It is important to realize that the underlying data may be of poor quality and incomplete, and the analyses may be biased, misinterpreted or used to mislead people. (The dictatorship of data, Cukier)

Humans are not particularly good at unbiased interpretation of data. Kahneman has showed us how this bias works. Kahneman notes that many people do not solve problems with logic or statistical insight, but with rules of thumb / heuristics that allow a quick and easy way to find an answer.

We constantly ventilate our whereabouts, opinions and feelings on social media and even in office systems. If we can combine the data accompanying social data with the Big Data we want to interpret it could be possible to develop Big Data systems that suggest how to avoid bias?

In the urge to be the best, Big Data companies will fuel autonomous search for data with autonomous actions on this data. This rapid and autonomous growth of data can make Big Data systems uncontrollably complex, ending up with systems that we do not oversee or understand. Ciborra studied this phenomena and called it ‘drift’. The more sophisticated, integrated and standardised the technological platforms become, the more they tend to behave autonomously and drift. We can only fully figure out the meaning of new technology in business and institutions after the fact [drift]; and we plainly have to live with such impossibility and state of ignorance.The risk is not that we do not understand the Big Data systems that we run. The greatest risk is that we believe that we know how the Big Data systems work, but in fact we do not.

Ciborra’s answer was to embrace bricolage within an organization. Bricolage can be seen as the constant re-ordering of people and resources, the constant "trying out" and experimentation. Bricolage is not a random trying out: Ciborra emphasises that it is a trying out based on leveraging the world "as defined by the situation". If successful Big Data applications come from bricolage, tinkering, improvisation, hacking, and accidental side issues, then the question is how can we incorperate good bricolage within Big Data technologies and strategies?

In order to prevent being governed by the data, trained professionals are needed that know how to analyse data and interpret the correlations and patterns that are found. People who do not take the figures for granted, but who look behind the figures to discover the true story and are not misled by the data: “big data psychologists”.

The issues described in this chapter show that the advancements in the area of big data create a demand for new professions, new rules and new ways to deal with data. Ethics play an important role: we must determine how to treat the information that big data provides, and what we can and cannot use this knowledge for. Transparency, ethics and control mechanisms are key ingredients for a prosperous, safe and secure big data era.

V-2. Future directions

Big Data resulting from human behaviour on the internet in general, and in social media in particular, represents a potentially valuable data source for the social sciences in the future.

Examples of Big Data sets and their potential uses in social sciences are:

- Data from social media: Content and preferences from large amounts of users on blogs, Twitter and Facebook may all be used to explore important social science issues, such as bullying, hate messages, conflicts, democracy issues, migration and decline in civic engagement.
- Data from online newspapers: Newspaper-reading behaviour pooled on basis of large numbers of online newspaper readers may provide insight in interests and preferences, which may support new knowledge on societal behaviour.
- Data from internet-enabled mobile devices: Large sets of behaviour data, user generated content and preference data, including information on geographical location and social networks, may support descriptive analyses of human behaviour and the formation and development of social groups in local and global perspectives perspective.

Big Data may require rethinking of scientific method. While "small data" relies on statistical sampling, and emphasizes the reliability and accuracy of each measurement, Big Data typically samples the entire pool of activities within a certain area (Mayer-Schonberger & Cukier, 2013). Some even suggests that conventional statistical hypotheses testing gradually will become obsolete due to the increasing availability of Big Data (Anderson, 2008). Others claim a *hybrid approach* is needed, mixing conventional deductive, top-down, or theory-driven approaches, with inductive, bottom-up, or data-driven approaches (Settles & Dow, 2013).

Companies such as IBM, HP, Intel, Google, Netflix, Amazon and Facebook are harnessing online data (online searches, shopping behaviour, posts and messages and general usage patterns and user preferences) to increase their knowledge about their users and preferences (Lohr, 2012a; Kolb, 2013). Social science, however, lags behind commercial companies in utilizing Big Data to gain new insights into human behaviour and society (Brandtzæg, 2013; Eagel, 2010), though Big Data tools and technologies for the mapping of behavioural patterns and human preferences across the globe, present opportunities to address grand societal challenges (Lohr, 2012). For example, user created content on Facebook offers social science and media researchers the potential for new forms of analysis, using large amounts of data on demographic patterns, rather than sample-based surveys of what people think they did or might do (Hale, Margetts, & Yasseri, 2013).

In addition, future research should address how we should collecte, interprete and manage big data to be used for scientific purposes, accurate quantitative research and representativeness Future challenges should also be guided by the current weaknesses described by Boyd and Crawford (2012), which also include problems related to privacy issues.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine (16)7. Retrieved online: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Brandtzaeg, P.B.(2013). A big data approach to measuring civic engagement, gender differences in social media among young people. International Conference 'Communication and Digital Society, Madrid, Spain. 17.04.2013 [keynote].

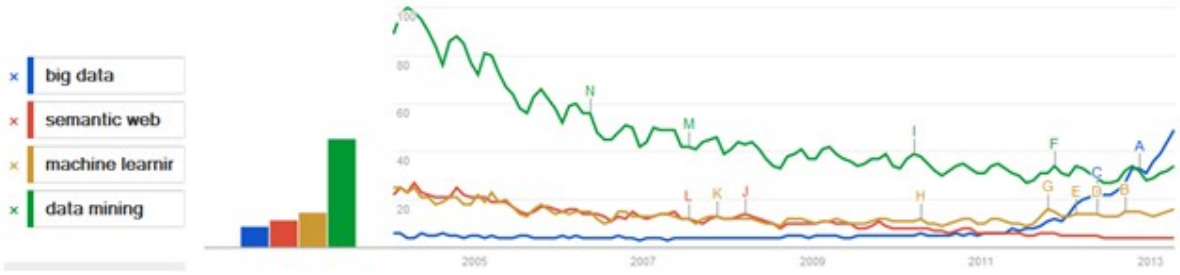
Mayer-Schonberger, V & Cukier, K. (2013) Big Data: A Revolution That Will Transform How We Live, Work and Think. John Murray Publishers Ltd, New York.

Settles, B. ,& Dow, S. (2013). Let's Get Together: The Formation and Success of Online Creative Collaborations. Proceedings of the Conference on Human Factors in Computing Systems (CHI). ACM, New York

VI- Business impact

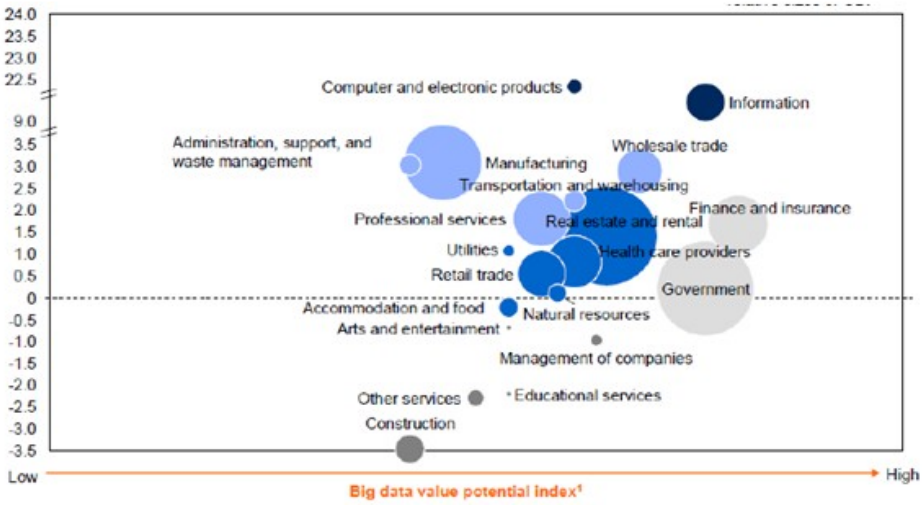
VI-1. Current status

There is general consensus on the fact that the amount of digital information created, consumed, and replicated will continue to explode. There are many factors that enable the big data torrent to grow in an enormous rate, ranging from the low cost of disk storage and the high penetration of mobile phones, all the way to social networks and the huge amount of shared information. As a consequence most public or private organizations recognize great value in these data and store them with the aim of benefiting from their intelligence. The growing popularity of big data on the web is another strong indicator of its impact, since the “big data” term has managed to overcome terms like the “semantic web”, “machine learning”, and only very recently “data mining”, as shown in the following diagram.



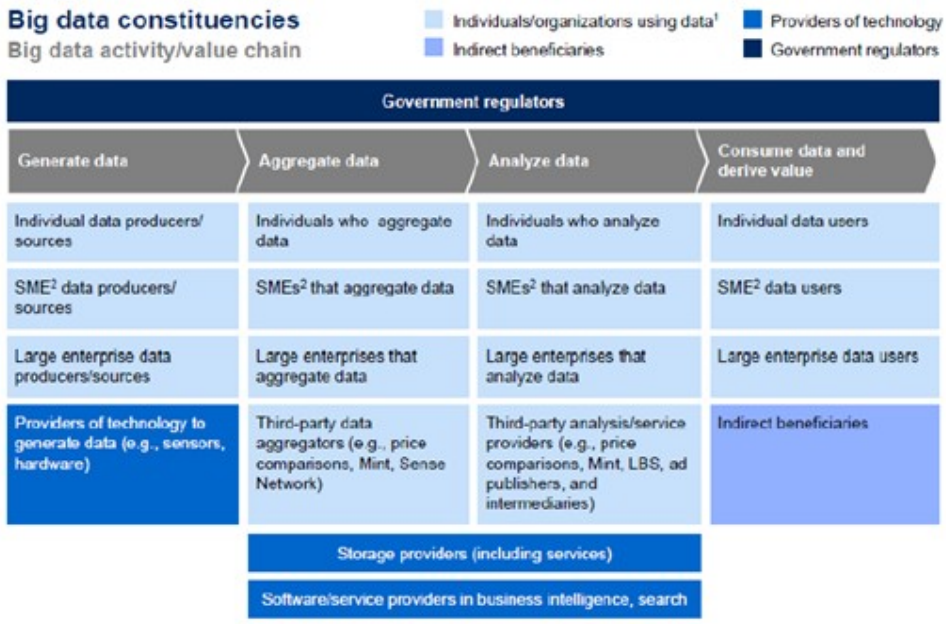
Source: Google Trends

Although the type of data generated and stored varies by sector, significant gains are expected from almost all sectors with some of them (e.g. Government, or Healthcare providers) having a biggest share than others.



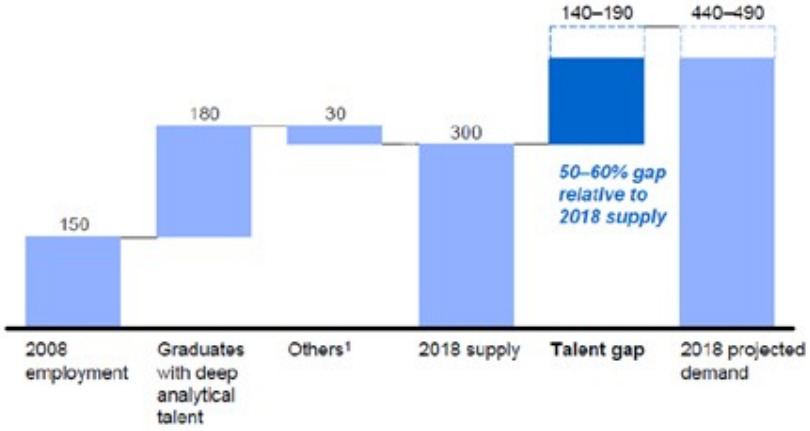
Source: US Bureau of Labor Statistics: McKinsey Global Institute analysis

In addition, the big data value chain involves a great number of different stakeholders that can range from: individuals/organizations using data, providers of technology, indirect beneficiaries, or governmental regulators that can be involved in any of the four phases (i.e. generate data, aggregate data, analyze data, consume data and derive value). Interesting is the fact that SMEs can take a prominent role in all different phases showing how the advancement of technologies for sense making out of big data can achieve great economic impact at all levels of industry.



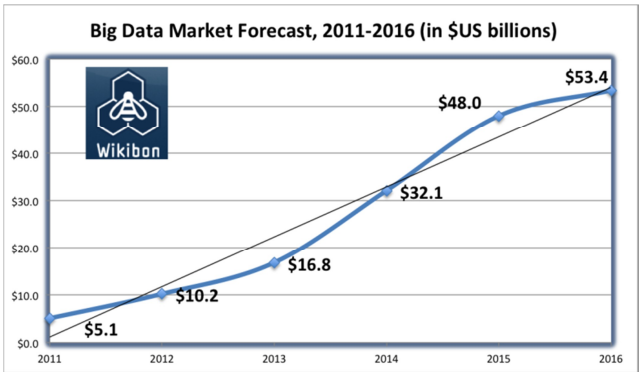
Source: McKinsey Global Institute Analysis

However, the volume and pervasiveness of information poses significant strains on our cognitive capacity, knowledge management systems and IT compliance infrastructures. It is evident that current businesses are faced with a multitude of challenges such as, how to find, understand, and synthesize the data needed to make better, faster decisions; or how to visualize data in richer and more meaningful ways. This growing need is very well shown in the following diagram that predicts the lack of talent for big-data related technologies. According to this graph the demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018.

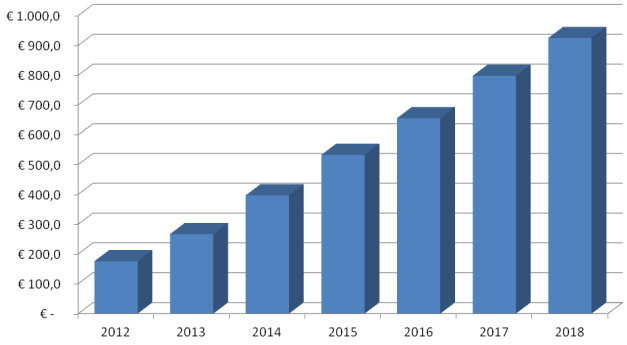


Source: US Bureau of Labor Statistics; US Census; Dun and Bradstreet; company interviews; McKinsey Global Institute

Big Data Market (BETTER AT ECONOMIC IMPACT)



Big Data Bestedingen Nederland, 2012-2018 (€M)

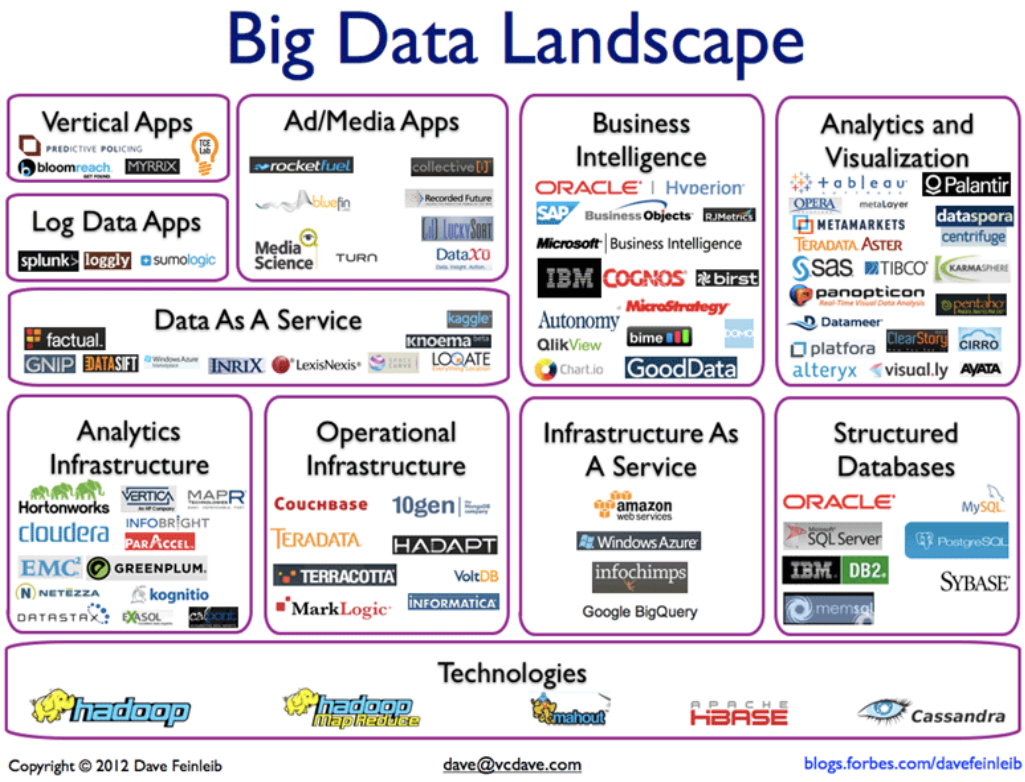


Bron: The METISfiles & Keala Consultancy, juli 2013

Most forecasts can differ dramatically depending on who does the analysis, but in general all predict high Big Data market growth.

The Wikibon report "Big Data Vendor Revenue and Market Forecast 2012-2017" lists over 60 big data vendors with total 2012 revenues of over $11 billion. Big data is a $5 billion dollar market today and is expected to top $50 billion by 2017, a 58% compound annual growth rate.

The Dutch Big Data market size was €176 million in 2012, an increase of 48% compared to 2011 and in 2013 the Dutch market is expected to grow 52% according to METISfiles and Keala.



European ICT suppliers do not play a significant role in Big Data. The only European player in the above figure is SAP and the European analytics tool RapidMiner, which is the world-wide favorite of data analysts (KDnuggets 2013) and offers a seamlessly integration to Hadoop/Hive (cf. Gartner report 2013).

The big and open data challenge

Companies certainly know better how to value “sleeping” big data rather than institutions and public bodies who are sitting on huge open data bases, try to find motivation, new markets and corresponding business models to open public data up to users they would like to get involved. But the users still have lots of difficulties in understanding what it is about, unless they are claiming the opening of public data and being (generally) volunteers in some kind of dedicated “open data” association. Associations are looking for *their* business models and also for that of smaller companies and developers they would like to get involved in the transformation process from available data to new services and business models. These associations positioned themselves as intermediaries between users, companies and public bodies. Awareness creation among all concerned parties remains one of the dominant issues in the open data fields.

The big and open data challenge is how to transform existing material and information into business impact. Before being able to create markets, the problem of large awareness creation needs to solved, and we propose in the following some considerations from the learning curve of associations and their municipality correspondents (municipalities are initiators of the open data movement in France, followed later by the government) who just glimpse solutions to come out of what they call the end of the open data blues.

Towards the end of the open data blues?

Idealized and expected for months and even some years (!), the birth of new services from open data provoking new models and markets was/is difficult. After the race of opening data, the euphoria of platform launches, the popularity of “hackathons”, and the frenzy of the applications’ competition... came the data baby blues, a symptom of post-data depression[1].

According to Libertic[2], the data blues matches to the fall of “interest rates” and initiatives around the global open data project. The interest of players of the value chain to get actively involved in the project, is estimated essential for the moral and the open data dynamic. However, this interest has been largely eroded by several findings[3]:

- Lack of data quality,
- No response to (potential) expectations of data re-users[4],
- Data structures changing without notice,
- Lack of accurate data (or budget lists or lists without identifier or lack of granularity),
- Lack of ambition in terms of not yet published data or transparency,
- Lack of interoperability of national data,
- More and more so-called open data licenses,
- Lack of sustainability for services created in competitions,
- Redundancy of services as and animations,
- Licenses versus free opening,
- Lack of data for democratic interest,
- Lack of diversified data,
- Lack of interaction with re-users,
- Limitations of short-term events,
- Lack of extension subject outside technical,
- Lack of re-appropriation, technical difficulties,
- Lack of open organizations, and others.

These difficulties to create quickly markets opened the field for the data blues for the short term. But in spite of these limitations, future open data markets will build on long-term dynamics.

The first phase of open data efforts highlighted the following:

- lack of data culture within organizations,
- poor quality of some data used as support for decision,
- partitioning of information,
- interest of external usages,
- possible gains through co-production efficiency.

After the blues, the next and second open data phase should take into account these “failures”, remove identified bottlenecks, counter obstacles, address expected benefits, sustain efforts and find the levers. This implies to develop new actions with new methodologies.

To do so, in the UK, the Open Data Institute offers comprehensive training courses, technical support through working groups and studies that go beyond open data and that consider the data world in general. Some months ago France started the Infolabs movement aiming at disseminating the culture of data through new mediation forms. And many other public processes are underway (revision of the governmental platform data.gouv.fr, organization of future thematic debates on health, creation of the Opendatafrance association for interoperability, etc.). Co-working is expected to include municipalities, associations, medias, companies, the learning environment, researchers and start-ups.

---

[1] The following part is taken and builds on the French association Libertic' observations and assessment of the French situation compared to some other European countries; 24/9/2013 in : http://libertic.wordpress.com/2013/09/24/vers-la-fin-du-baby-blues-de-lopen-data/

[2] http://libertic.wordpress.com/

[3] ibid

[4] People or companies being able to present data in a digest way and bringing it through new services to market usage

The Business Impact of Big Data

The current status in US and Europe

The Big Data beginnings in the US

History of Big Data

Initially Big Data Analytics was used by very large companies TELCO companies like Bell Atlantic Retail companies like WalMart,. It was a game with (expensive) specialists, such as SAS and Teradata, and very expensive proprietary software. Many new ideas and tools in the area of Big Data come from young internet companies, such as Google, Yahoo, Amazon and Facebook. These young companies have in turn produced various new start-ups. The traditional ICT suppliers are quickly adapting to the new situation: IBM with Watson, HP with the acquisition of Autonomy and Vertica, EMC Greenplum and SAP with development of in memory technology HANA. In the meantime Big Data analysis has also become possible for smaller companies. Cheap commodity hardware and open source software, together with Big Data cloud solutions, such as Big Query Google and Amazon Domino, ensure that Big Data is within the reach of SMEs.

Influence of the internet companies

In the late nineties a data explosion was caused by massive use of the World Wide Web. Indexing and querying extreme web content became necessary. The first serious 'Big Data' systems were built around 2000 by American Internet companies like Google, Facebook, Yahoo! and Amazon. These internet enterprises needed Big Data technologies that the traditional ICT suppliers (IBM, Oracle and Microsoft) could not deliver. Big Data Business in the USA is therefore paramount. The US Scientific community was relatively late to tap into this Big Data

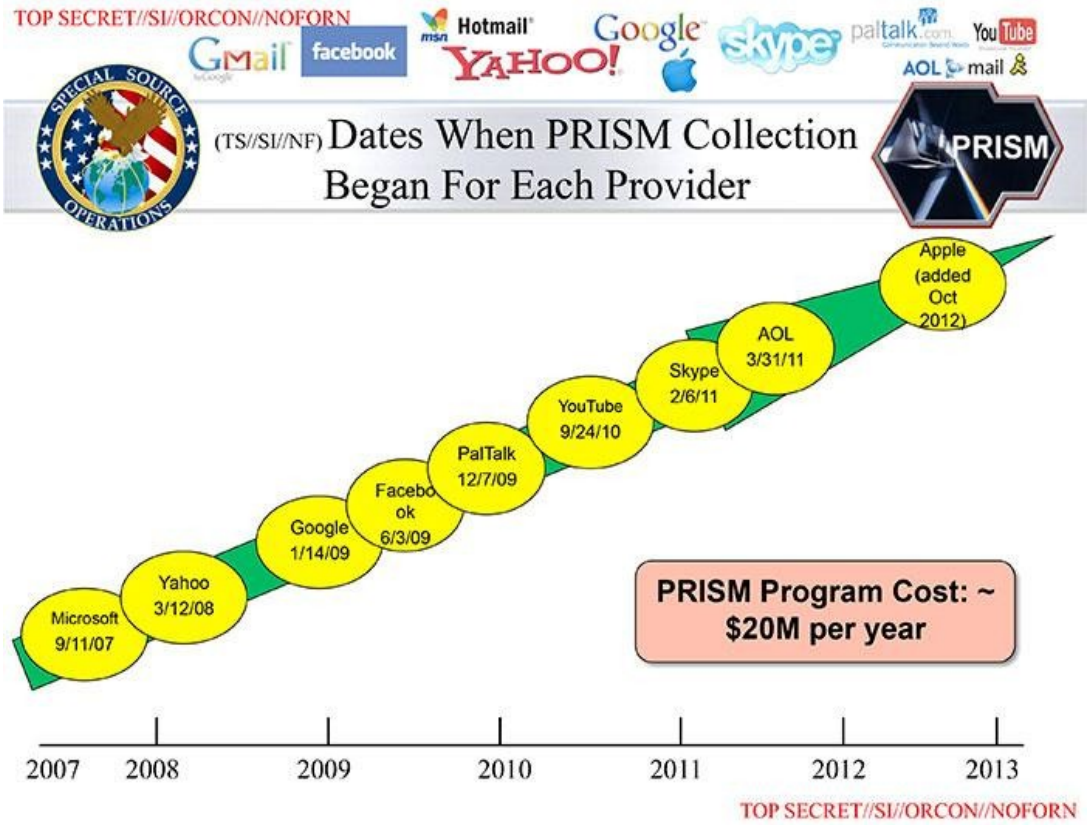
movement. Therefore Big Data innovations come mostly from commercial companies. Today a huge number of start ups appear as a spin-off thanks to the above mentioned. In the last two years, 119 database software vendors received $ 1.17 billion venture capital. Dow Jones Venture Source.

In short, Big Data is a USA dominated play-ground.

Because the traditional ICT suppliers could not deliver systems that would handle extremely large data sets, the internet enterprises propelled Open Source development on Big Data. As the owner of the largest database on earth, Google is at centre stage in Big Data. A lot of new open source developments start at Google and are sometimes published via Google papers. The other internet enterprises Facebook, Twitter and LinkedIn took these papers as a starting point for their own open source projects. As a consequence huge investments are made in Big Data open source software. Again, the traditional ICT companies adopt that open source technology, for instance integrating Hadoop in their current data management systems.

The role of the US government

The US Ministry of defence has direct influence on ICT innovation in the US and surely also on Big Data innovation.



The USA dominance is perfectly illustrated by the recent PRISM affair.

"Some of the world's largest internet brands are claimed to be part of the information-sharing program since its introduction in 2007. Microsoft was the first, with collection beginning in December 2007. It was followed by Yahoo in 2008; Google, Facebook and PalTalk in 2009; YouTube in 2010; Skype and AOL in 2011; and finally Apple, which joined the program in 2012. The program is continuing to expand, with other providers due to come online. Collectively, the companies cover the vast majority of online email, search, video and communications networks" *Date: 07-06-2013 Source: The Guardian*

Obama the first Big Data US President.

"Voter-registration files have been merged with vast quantities of bought consumer data, on top of which come bought or acquired e-mails, mobile and landline numbers, as well as data gathered through canvassing, phone banks and social-media pages. The campaigns are also making use of cookies, the crumbs of data people leave behind when they browse the net" Date: 01-11-2012 Source: The Economist Subject: Voters are being targeted in new and powerful ways

The US Federal Government announced in 2012 the National Big Data Research and Development Initiative, the Big Data Initiative featured more than $200 million in new commitments.

European adoption of Big Data

European ICT companies have a backlog (1 to 2 years) compared to the USA. European scientific institutions are also relatively late to become involved in Big Data research. There are only a few Big Data technology suppliers in Europe.

In Europe, most of the larger, ICT intensive companies look into Big Data. Some large companies already have limited experiments started. These pilots are mainly based on the Hadoop environment. Examples in the Netherlands can be found at Rabobank, BOL.com, ABN Amro and Marktplaats (eBay). But there are also examples of smaller Dutch companies that have recently focused on Big Data services such as SNMNT, Xebia, Synerscope, Crystaloids and Data Provider.

Interviews with about 90 decision makers (mostly in the Netherlands) show a lot of enthusiasm for the subject Big Data but not much action. Decision makers do not have to be convinced that Big Data holds possibilities, but most of them hesitate to start. There are too many unanswered questions. How to start? Which Big Data technology to choose? What is the Big Data business value? What is the Big Data business case? In short, Big data has the attention of most companies, but there aren't many real implementations. This is illustrated nicely with a study by Interxion and Gartner:

- *62 % of European organisations say that within 3 years Big Data has priority*
- *50% of European organisations say they have more pressing tasks*
- *7% (Benelux 13%) of European organisations give big data priority*
- *Only 25% of European organisations have a business case*

Source: Big Data - beyond the hype report from Vanson Bourne and INTERXION HOLDING NV (NYSE: INXN), a leading European provider of carrier-neutral colocation data centre services.

The survey in 2013 found that of the 64 percent of organizations investing or planning to invest in big data technology, 30 percent have already invested in big data technology, 19 percent plan to invest within the next year, and an additional 15 percent plan to invest within two years.

Source: September 2013 survey of 720 Gartner Research Circle members worldwide

Big Data technology and know how about how this technology can be deployed is still in its infancy. Asking the right questions, and finding the right accompanying data, is in full development. Experimentation is needed to let the proper Big Data tools and best practices crystallize.

Big Data requires a light and rapid development method. How do you realize an 'agile' approach? The consultancy firms such as IBM, KPMG, Nolan Norton and Sogeti have now their own "big data methods" pieced together. However, as yet, there are no solid public Big Data "best practices" known.

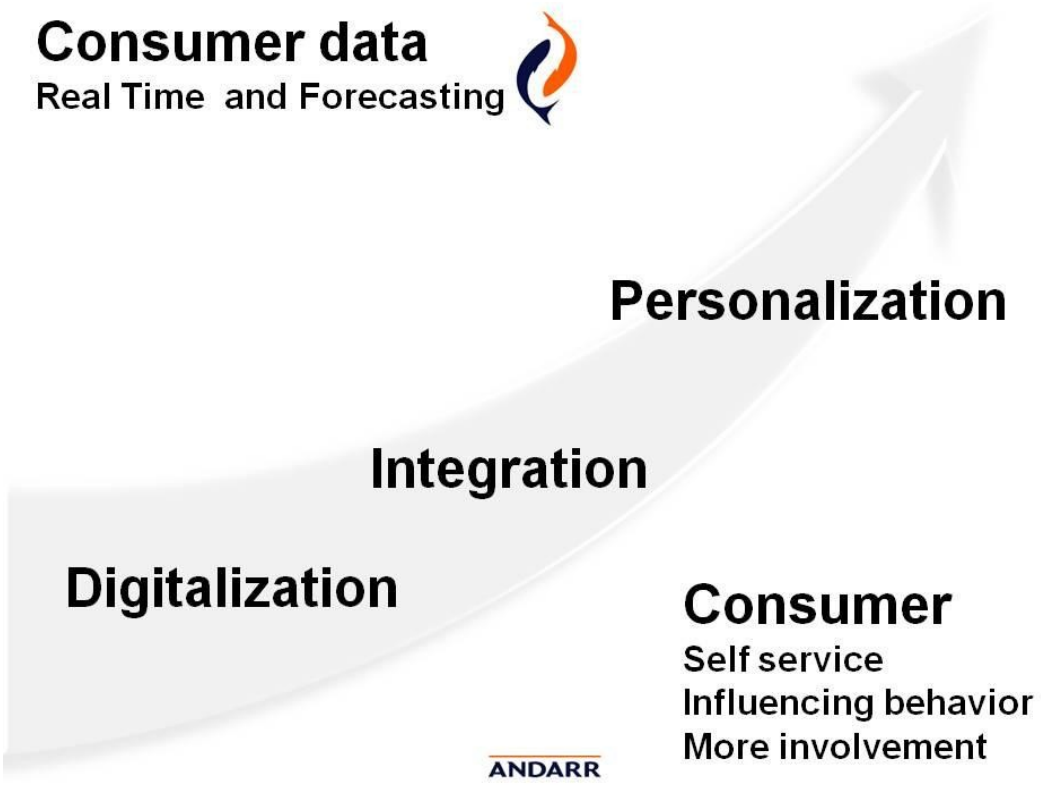
VI-2. Future directions

The future direction of Big Data: from digitalization, integration to personalisation

The "break point" of Big data is the sudden realization that we have collected 'critical mass' of valuable data we can make use of. The increasing flow of data brought the industry to a tipping point.

Big Data is a logical consequence of digitalization and the interconnecting of data sources (because data communication and computation is cheap). Most organizations are still integrating data sets (killing data silo's in enterprise) and some did not even digitize the necessary data (still scanning paper). Luckily, nearly all data is generated in digital format today and a lot of heterogenic data is collected. The early adopters already use Big Data to deliver personalized services. Examples of these early adopters are the traditional large retailers, financial institutions and telecom operators and now internet enterprises are the front runners. Organizations with a lot of consumers that are fast moving and information intensive will embrace Big Data first.

In the coming years nearly every industry will have useful Big Data around. All organizations will go through, or have already gone through, the steps from digitalization, integration to personalization. (see figure).



The ultimate business goal of big data is to deliver personalized services to the consumer. The consumer can be a student, a patient, a buyer, a citizen, etc. From the Big Data perspective organizations go roughly through this three-level evolutionary path, first digitized data has to be available, second the organizations have to be able to integrate different data sources (abandon silo's) and lastly organizations can personalize the data (do 1-1 marketing/service). To make personalized services work, Big Data analytics has to be more real time and more predictive.

The value for the consumer is more engagement/involvement or self-service, leading in the end to better services. The benefit for enterprises is to improve influence on consumer behaviour. Real time and predictive analytics solutions can sift through large amounts of data, understand, categorize and learn from this data, and then predict outcomes or recommend alternative product/services to consumers at the most useful moment. This drives highly targeted and localized consumer services and should improve ROI on Big Data investments.

"Learning to use a "computer" of this scale may be challenging. But the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There's no reason to cling to our old ways. It's time to ask: What can science learn from Google? "

Source: Chris Anderson (canderson@wired.com) *is the editor in chief of* Wired. "the data deluge makes scientific method obsolete" 2008

If a dataset is large enough, and up to date, the empirical approach often works better than a formula. Predictions based on correlation is the business success of Big Data. The more data you have, the more correlations you can find and the more accurate the pattern and the prediction become. In short, correlation is enough. We can stop making models in search of causation. You do not need a sample (n=150) when you can use all the data (n=all). Of course data cannot speak for itself, but thanks to the data explosion the emphasis lies on the data and algorithms rather than on the traditional models. Still, to be successful, domain knowledge is needed to judge the incoming data and the outcome of the analysis..

Some examples from the book "Big Data: A Revolution that Will Transform How We Live, Work and Think," of Google-like companies doing interesting things with big data:

- *Oren Etzioni, creator of one of the first search engines, MetaCrawler, scraped data from a travel website to build a tool called Farecast that predicted when an airfare price was likely to go down. Microsoft bought Farecast for $115 million in 2008. Today, it's part of Bing.*
- *Then Etzioni turned around and created Decide.com to help people predict when prices will drop on electronics.*
- *PriceStats tracks the prices of millions of products in over 70 countries to keep tabs on inflation and the economy.*
- *AirSage gathers 15 billion geolocation records daily from multiple cell phone carriers. It uses this for traffic analysis to help city planners.*
- *Similarly, Inrix collects traffic data from the sensors in the cars themselves via BMW, Ford, Toyota, and others and uses that to help city planners model traffic flow.*
- *Kaggle crowdsources the process of writing a big data algorithm by running contests.*
- *TheNumbers.com collects detailed records on Hollywood films, from box office sales to who worked together on which films. Producers use it to predict how much money a film is likely to make.*

Source: http://www.businessinsider.com/big-data-is-about-to-produce-a-whole-bunch-of-google-like-companies-heres-how-2013-8#ixzz2f4yBhieP

Source: "Big Data: A Revolution that Will Transform How We Live, Work and Think," written by Oxford professor Viktor Mayer-Schonberger and big data journalist for the Economist, Kenneth Cukier.

For the Dutch, the following topic is very important: Better water management

"IBM has just landed a €1 million ($1.3 million) 'big data' research project in the country to bring together disparate data sources related to water to help the authorities plan reactions to deluges, monitor water quality, improve internal navigation — in short pretty much anything to do with water in the country. The project, named Digital Delta, will investigate how to integrate and analyze water data from a wide range of existing data sources. These include precipitation measurements, water level and water quality monitors, levee sensors, radar data, model predictions as well current and historic maintenance data from sluices, pumping stations, locks and dams. "

Source: Wallstreet June 25, 2013, 11:28 AM The Netherlands Looks to Big Data to Tackle Flooding Source: Website video: http://www.digitaledelta.nl/

Another interesting topic is preventive and personalized healthcare by continuous monitoring and automatic diagnosing of the patient (example IBM Watson)

IBM, on the heels of its triumph last year with Watson, the Jeopardy-playing computer, is working on Watson for Healthcare. Physicians set aside five hours or less each month to read medical literature, while Watson can analyze the equivalent of thousands of textbooks every second. The program relies heavily on natural language processing. It can understand the nature of a question and review large amounts of information, such as a patient's electronic medical record, textbooks and journal articles, then offer a list of suggestions with a confidence level assigned to each.

Source: Date: 04-12-2012 The New York Times Subject: For Second Opinion, Consult a Computer?

Google as "Big Data business" acid test

Google runs everything according to data. That strategy has led to much of its success. Google has world data dominance. Google is the best Big Data company that knows how to use correlations. It keeps records of everything we do with its online services and finds new ways to use that data from Google translate to Google flu tracker. Google combines data location, calendar, mail, searches and video choices to offer individualized just-in-time services.

The acid test is: What if Google moves into ….. my industry?

Google growing 32 percent from 2011 to $50.2 2012 billion in annual revenue.

Google the largest database on earth.

Google makes up 25% of Internet traffic.

Google has world second largest network.

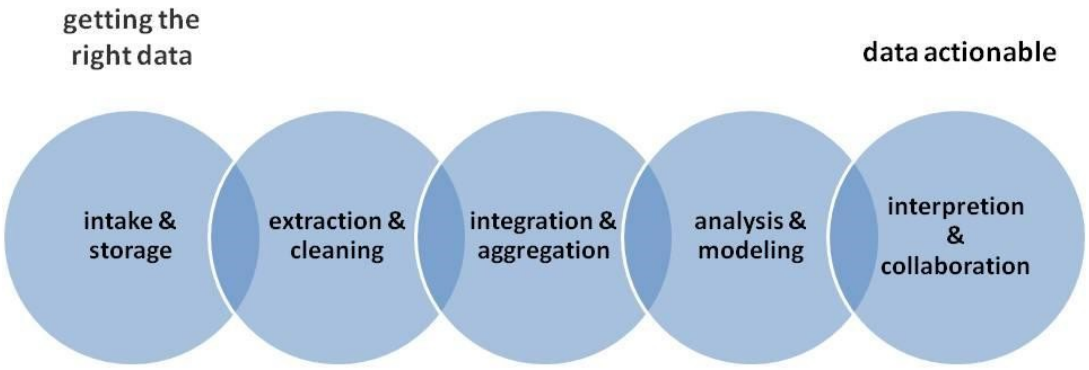
Google is the third server supplier for its own use.

Google spends $1 billion each quarter over the past few years on infrastructure investments.

The next frontier is to master personalization of services and products with Big Data

The company who gets the right data first and make it actionable the fastest has the business.

Organizations have to incorporate Big Data refining processes that take the right data and make it actionable within the organization. Big Data refining processes can be seen as a series of steps: intake and storage, extraction and cleaning, integration and aggregation, analysis and modeling, interpretation and collaboration. In business the first and last step are crucial. (see figure)



The decision maker has to select data and make others act on that data within the organization. The decision maker has the role of "Data Detective": the Big Data domain expert and analyst that translates a business question into a Big Data question. And is supported by the Hadoop Hacker: The DevOp (developer and engineer) with knowledge and experience of big data tools and methods. Hadoop Hacker has to oversee the whole refining process.

It is to be expected that Big Data tools are getting simpler to use and more people are gaining experience using Big Data. An ecosystem of data collectors, data analyzers, data providers, data advisors that support organizations with Big Data projects is emerging. These 'data suppliers' have a role in the ecosystem sourcing the whole or parts of the refining process.

Getting the right data for personalised services

The most business value comes from the combination of heterogenic data sets, especially the combination of internal and external data. A bank that checks your Facebook friends payment behaviour before you get a loan is an interesting example of combining external Facebook data with internal bank transaction data.

Typical internal data is : Website traffic data, financial transaction data, Customer Relationship data, Supply Chain Management data and Enterprise Resource Management data.

Typical external data is : mobile data, app data, search engine data, sensor data, geo data and social media data.

The issues in the hunt for datasets are:

- Overcome barrier of confidentiality and competitive sensitivity of the data.
- Social Media data will be increasingly difficult to get. No access at all or to high price to buy the data from social media parties like Twitter, Facebook or Google.
- Open Data movement getting momentum in Europe. Government will make more data available forced by EU regulations.
- Create your own datasets via engagement with consumers through strong social media presence (social media listening) or installation or use of sensor network (medical sensor, intelligent energy meters or Smartphone sensor apps).

Make personalised services work bases on data

Social analytics plays a major role in building strong personalized services. Gartner sees social analytics as the merging of context, sentiment and social network analysis.

- context (personal activity and location),
- sentiment (ratings, popularity, opinion and reputation monitoring),
- social network (social influence, value network, organization network analysis).

Getting correct data is vital. People lie about their health, 9% of Facebook profiles are fake, a lot tweets are generated by computers, sensors produce noise, etc. If the data is incorrect you have to be able to go back to the source, i.e. the person or machine/sensor that is the cause of the error.

Italian security researchers Andrea Stroppa and Carlo De Micheli say they found 20 million fake accounts for sale on Twitter this summer. That would amount to nearly 9% of Twitter's monthly active users. The Italian researchers also found software for sale that allows spammers to create unlimited fake accounts. The researchers decoded robot-programming software to reveal how easy it is for spammers to control the convincing fakes

Date:25-11-2013 Source:The New York Times Subject: Inside a Twitter Robot Factory

With suitable metadata generated at the source the data can be traced through the data refining process. In short, know the data provenance to handle error and uncertainty. Sometimes it is cheaper not to store the data, but ask the source again. (example when DNA sampling is cheaper than DNA data storage)

For sure there will be more data than most organizations can cope with. If the amount of data is too expensive to store, organizations have to reduce data volumes intelligently. How and which data to prune? Which data to retain? This is an art still to be developed.

Let the users interpret the data correct (deal with psychological bias) and making it possible to let the organization do the right action (actionable data). The best organizational practise will emerge in the coming years.

Issues that stretch the limits of personalised services with Big data:

- Proactively suggest patterns or correlations for consumer services.
- Automatic profiling of consumer.
- React real time on consumer behaviour. (google glass and steer interface with brain signals)
- Avoid crossing the creepy line from transparent influencing consumer behaviour to intrusive manipulation of consumer (or even perceived manipulation by the consumer).
- Social analytics put to work in the internal organization.
- The urge to be the first will fuel autonomous search for data and autonomous actions on the data

Considering the potential economic of big data, all stakeholders should start a broad discussion on effects impact societal values such as privacy and how to boost data-driven entrepreneurship.

VII- Conclusion and recommendation from NEM

Digitally generated and available information in public-private sector organizations as well as in the personal domain are becoming increasingly comprehensive and traceable. The volume and variety of data that is generated, collected and processed is growing rapidly, thanks to ubiquitous connectivity, increases in processing power and the growth of new digital data sources such as sensors and social media. The ability to collect, store, integrate, analyze this data, extracting the value out of it, is growing as big data technologies are refined.

The usage of digital data and digital identity are key components of business in data-intensive companies such as Facebook, Google, Amazon,...– By 2015, one quarter of the world's population will probably be a member of some social networks sharing personal information (demographical data, likes and preferences, images, video, location). It is already demonstrated and clear that, most end users are willing to share their personal data with public- and private-sector organizations for sufficient benefits (individual and social) and with proper privacy controls. Analysis of social network data provides key information on the users' social behaviors that can be exploited to design, optimize and dynamically operate social innovation services in a range of different domains.

As similar explosion is happening as a consequence of a variety of sensors, particularly ambient sensors (home, cars, open spaces, surveillance cameras,…) and of sensors embedded in objects (cell phones, appliances…). All of which will grow even more as the Internet of Things is established and matures.

Data are becoming, de facto, the new infrastructure superseding the physical communications infrastructure that could create value thanks to innovative effective tools able to extract relevant information out of huge amounts of heterogeneous data.

The usage of data is relevant to many sectors and to the whole economy. Retail, Finance and TELCO companies already use data extensively for internal enhancement and business process management (such as TELCO Fraud Detection, Churn Management, Client Profiling & CRM,…). Other private sectors and public services are at the initial steps of value generation from digitally available information, such as HealthCare providers migrating to digital medical records and setting up processes that leverage digital data along the whole value chain.

Leveraging internally generated data, organizations can reduce operational costs, create new revenue streams and expand old ones. Some of the major application domains for big data technologies are internal process automation and optimization together with creation of more personalized products and services.

While this trend of "internal optimization" will continue we expect the significant growth of "external" valorization and exploitation of data:

- new services for citizens based on analytics of heterogeneous information flows established "outside" of a single organization (including, among others, data coming from distributed sensors describing the "physical world" of the users, as well as data coming from social networks, describing the "virtual/cyber world" of the users);
- new ecosystem of data-related opportunities, such as sharing and selling information to third parties – Data Marketplace.

However, most of the potential value generation out of big data is at risk without an efficient and trusted flow of digital data.

The public sector and particularly services in Digital Cities / Territories and Medicine / HealthCare are expected to have significant benefits out of digital data flows leveraging big data technologies and related applications.

Beyond the below technical challenges and societal challenges which needs to be addressed at European level in a next future, NEM is recommending to be included in the discussion around the rational of establishing a Big Data Public Private Partnership (PPP) within the Media and data Directorate of DG Connect in order to have its community involved and to avoid to build an instrument to much focus on Information Technology data and not taking the huge amount of data generated by our new era of Digital Europe by our digital citizen.

VII-1. Technical challenges that have to be addressed now

Information flows, generated and collected often in real time, are the foundation for new services concerning individual and social-wellbeing, safety, environment, green and integrated transport, resource efficiency and social inclusion. Most of the potential value generation out of big data is at risk without an efficient and trusted flows of heterogeneous data related to particular territory (Data Governance). In order to enable information flows, the "data infrastructure" must be open in the input side (data sources shall be added in a plug-and-play way) as well as in the output side (data shall be accessible through common interfaces).

Some challenges @Data Management and Analytics Layer **- n**eed for flexible and reliable data management and analytics infrastructure - HW/SW data management and analytics architectures/solutions - presenting the following characteristics:

- scalable, cost efficient, environmental friendly and easy to manage multiple autonomous systems that support both structured and unstructured data;
- provide a mix of centralized in memory, distributed/parallel and device based analytics;
- include data bridges to construct common standards;
- data stream reasoning/management/storage;
- data cleaning, provenance and quality assessment and management;
- efficient search and retrieval;
- integrate real-time stream data mining and historical data mining;
- integrate analytics with data storage and data visualization.
- interactive and collaborative modeling and simulation of large and complex phenomena and systems;
- large scale social networks analysis;

Some challenges @Data Access Layer **-** The following are some of the main (technological and non) challenges here:

- Data ecosystem with Data Marketplace and new business models based on Open Data, Data vendors, Data customers;
- Data Visualization - new ways of presenting information - in order to make the knowledge derived from the data valuable and "actionable", new methods to explain and visualize the data, the new tools will be required, supporting interactivity, filtering, personalization, …
- User controlled Privacy based on awareness, participation and control and clear policies help to reduce the threshold for adapting new applications based on data sharing. Particular attention should be put on Accountability and Data Governance;

Analytics-oriented challenges need urgently to be tackled, since only they reward the efforts in collecting and storing the data. The known methods for data analysis require serious and demanding changes for their distributed, parallel and online processing of Big Data. Both batch (using primarily distributed and parallel data analysis) and real-time oriented systems (using primarily online processing on data streams) are needed.

Embedded analytics exploits the streams of data in real time under strict resource restrictions of computing capacity, storage, energy and communication bandwidth. This allows to harvest the value of data streams from services of mobile devices for logistics, traffic, medicine, production, and entertainment. The research to be done ranges from the infrastructure to analytics to visualization. Moreover, it should be extended to process control such that natural resources are saved and services are personalized and respect privacy.

Tailoring this broad subject into the application areas: logistics, traffic, medicine, production, and entertainment

Definition of the **European knowledge-based data architecture** between heterogeneous databases in a multiplayers environment.

- What place and role of the different actiors ?
- How to implement distributed calculations in an heterogeneous environment ?
- How to implement distributed calculations dealing with public and private databases ?
- How to insure security, SLA ?

Short/medium term :

- Raw data conversion into knowledge data
- Privacy insurance
- Open data APIs
- Distributed calculation in a shared infrastructure with third parties

Medium/long term :

- Semantic characterisation of data
- Qualification, traceability and certification of data
- Knowledge databases interconnection
- Ontology's alignment, semantic graphs

VII-2. Societal challenges that have to be addressed now

Companies such as IBM, HP, Intel, Google, Netflix, Amazon and Facebook are harnessing online data (online searches, shopping behaviour, posts and messages and general usage patterns and user preferences) to increase their knowledge about their users and preferences (Kolb, 2013). Social science, however, lags behind commercial companies in utilizing Big Data to gain new insights into human behaviour and society (Brandtzæg, 2013), though Big Data tools and technologies for the mapping of behavioural patterns and human preferences across the globe, present opportunities to address grand societal challenges (Lohr, 2012). For example, user created content on Facebook offers social science and media researchers the potential for new forms of analysis, using large amount of data on demographic patterns, rather than sample-based surveys of what people think they did or might do.

We recommend development of new and more accessible methods to analyze data as well as more open data that aims for high societal impact by providing industry and academics with improved capacities for applying Big Data in the social science domain, supporting scientific renewal;

Regarding privacy aspects, there is a huge need to develop tools suitable for any user (end users or organizations) to withdraw exhaustively any data stored in the cloud (data centers). We need to develop a tool able to analyse all content stored in any data center and to offer to the owner a "suppress" function.

Regarding Open data, several societal studies have to be set up in order to give a consistent environment. Open data should come from several sources such as public data, personal data and corporate data. These 3 sectors have to be studied in order to find commonalities and specificities from a social point of view. In addition these open data should be static data or real-time data, these 2 types of open data have specificities that have to be addressed. These actions should be :

- Intellectual property right
- Where is the value ?
- Big and open data as tools to preserve and improve democracy and security

- Raise the data value awareness through education and excellence (privacy charters, best practices, audits, certification)
- Demonstrable value of cross sector or vertical uses of data based systems (services, operations, decision making)

In addition, it appears that Europe has a lack of data scientists and there is an urgent need to set up a recovery plan in order to help establish master degrees in this domain.

Brandtzaeg, P.B.(2013). A big data approach to measuring civic engagement, gender differences and usage in social media among young people. *International Conference 'Communication and Digital Society,* Madrid, Spain. 17.04.2013 [keynote].

Kolb, J. (2013). *Secrets of the Big Data Revolution*. Applied Data Labs Inc. Boston.

.

Lohr. S. (2012). The age of big data. *New York Times.* Retrieved online: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html), Feb 2012.

[1] http://www.neo4j.org/
[2] http://thinkaurelius.github.io/titan/
[3] http://www.orientdb.org/
[4] http://www.tineye.com/
[5] http://www.moodstocks.com/
[6] http://www.kooaba.com/
[7] https://www.iqengines.com/
[8] http://www.ltutech.com/
[9] http://imagga.com/
[10] http://samasource.org/
[11] http://www.attrasoft.com/

Annexe

NEM workshop @ FIA 2013 :

NEM took the opportunity of the Future Internet Assembly (FIA) in Dublin May 8th -10th , 2013 to submit a workshop proposal dedicated to open data, which what accepted by the FIA organization. In fact, Future Internet offers the ground for accessing new content in a way that third parties exploitation and utilization can offer new innovative services. Open data belongs to this new content category and it is rather important to better know the borders, the access mode, the business models, the challenges in front of us.

This session organized May 10th 2013 had the objective to offer answers in questions arising for these Open Data tasks and at the same time to share a common vision on Open data. Questions were around:

- What is the definition of Open data ?
- Who are their providers ?
- Which are their access APIs ?
- Which are the key issues for their utilization ?
- Which are the most suitable business models ?

The audience (around 25 people) concerned:

- Researchers and developers from industry and academia engaged in Future Internet programs and projects, e.g. FI PPP program, FP7 Media projects, etc.
- Innovation leaders and entrepreneurs interested to know more about Open data and the usage
- Open data providers and stakeholders such as cities, ministers, local government and relevant authorities.

This workshop was the opportunity to set up the open data scene and get views from recognized experts, feedback from an experimentation conducted at Santanders (Spain), and different views trhough a round table and exchanges with the workshop attendees:

- Jean-Dominique Meunier, NEM Chairman, presented the NEM European Technology Platform dedicated to Media & Content. He outlined that Open Data subject is new to NEM community and most probably source of lot of opportunities. He was thanking the European Commission for allowing NEM to tackle this subject in a FIA workshop and to address several questions : is it linked to big data ? is it metadata ? How to have access to them /How to use them ? What is the EC position ? What' about standardization ? What are the experimental test feedback How Citizen are appropriating them ?
- Prof. Stefan Decker from Digital Entreprise Research Instituted, National University of Ireland Galway , pointed out that many open data initiatives are now existing in Europe : ie Apps4Fingal for creating apps which provide info particular at cities and at local level; also transport and travel info has exploited open data. There is open data movement in various fields : Archives / Europeanna/Libraries are now starting to publish their content in an open data format. Sciences also provide open access and open data (biology, climate change, open EI, etc). A network of knowledge is developed with RDF/Vocabularies - Examples given from Wikipedia/gov.ie to prove information aggregation, from Data catalog vocabulary emerging and SPARQL He concluded that a Network of Knowledge is more than just data, it It is people, communities…, And then it assists humans, organisations and systems with problem solving, enabling innovation and increased productivity
- James Clarke, from Waterford Institute of Technology – TSSG, Steering Board member of NEM made a presentation on open data challenges including those in related to trust and security. Privacy-by-design and privacy-by-default and security and trustworthiness issues were raised , where applicable in open data (secure protocols, cyber forensics, crypto, ..). The need to ensure that the Open Data Directive is in sync with the Data Protection Directive was emphasized.

Trustworthiness and a compilation of related open data projects and initiatives were highlighted, summarized and discussed.

- D. Iñigo de la Serna Hernáiz, Santander City Mayor from, Spain, presented SmartSantander project and its services along with its social innovation and new business models priorities. Smart Santander aims at providing a European experimental test facility for research and experimentation on architectures, key enabling technologies, services and applications (i.e., augmented reality, participatory sensing),for the Internet of Things (IoT) in the context of the smart city. 20,000 IoT devices were installed.
- The panel discussion allow the attendees to rise some specific issues such as how to control access to data (and recommend not to deliver raw material), how to encourage generic applications (and encourage going to standardization), how to balance close or open approach (and what are the corresponding added value and markets). Tools availability such as the one that FIRE projects can offer (in delivering open data) is key in the experimentation to conduct and as support for standardization.

The output of this workshop on Open data is clearly an opportunity for the NEM community to issue a position paper identifying the research trends and activities that would deserve attention in the future EU policy and research program(s). There is a need to ensure that Open Data Directive is in sync with Data Protection Directive. Key messages can be highlighted so far:

- Network of Knowledge is more than just data
- Stimulating Social Innovation through the Open Data Paradigm
- Privacy-by-design, privacy-by-default, security are key issues to tackle
- Open Data business model(s) remain to be invented
- Let's try to avoid over-regulation in the open data. Otherwise we would be killing the paradigm before making it a reality.

NEM workshop @ NEM summit 2013 :

The 2013 NEM Summit hosted the **Diverse Data Innovation Workshop** on October 29, 2013, chaired by Chris Thompson (Partnerships Director, The Connected Digital Economy Catapult (CDEC))

This workshop brought together some of the best European SME, public and corporate innovators currently working with diverse forms of big, open, diverse and integrated data, giving insight into the latest innovation across Europe in relation to data, spanning healthcare, media, creative, finance and cities sectors. The workshop also introduced new collaborations between SME's, entrepreneurs and large organisations to address data challenges and solutions and formed new cross-European partnerships to apply for Horizon 2020 data challenge related funding. The BBC and the Connected Digital Economy Catapult aim to keep the momentum going and will work together to develop new mechanisms to bring more data innovators and SME's spanning media, creative and technology sectors into future NEM and H2020 opportunities. The contributions of this workshop will help to consolidate a NEM 'position paper' being written on this subject.

Agenda and presentations

10:45	Registration
11:00	Welcome
11:10	*"The economic potential for data innovation"*, **Chris Thompson**, The Connected Digital Economy Catapult
11:25	*Keynote: "Open data challenges and opportunities across the public and research sectors"*, **Stuart Coleman / Gavin Starks**, ODI, UK – live from

	the ODI Summit in London
11:40	*Keynote: "Public data innovation challenges and opportunities"*, **Oliver Bartlett**, BBC, UK
11:55	*Keynote: "Corporate data innovation challenges and opportunities"*, **Patrick Launay**, Orange labs, France
12:10	*"SME case study: Innovating with open health data"*, **Neale Swinnerton**, Mastodon C, UK
12:25	*"SME case study: Enabling innovation with corporate data"*, **Adrian Hillary**, Sibdocity, UK
12:40	*"SME case study: Advances in data visualization"*, **Dr Rachel Jones**, CIKTN, UK
13:00	*Lunch*
15:00	*"SME case study: Innovating with city and citizens data"*, **Benoit Simard**, Libertic Nantes, France
15:15	*"SME case study: Innovating with real time public data"*, **François Bancilhon**, Data Publica, France
15:30	*"SME case study: Innovating with interactive, media and cities data"*, **Rob Aalders**, R.A.U.M, Holland
15:45	*"Learning, integration and confidentiality preservation for personal consumer data"*, **Marc Gelgon**, Nantes University, France
16:00	*Roundtable: "Policy, research, technology and regulation opportunities to drive data innovation and collaboration across Europe"*, **chaired by Chris Thompson,** The Connected Digital Economy Catapult
16:30	Closing remarks