

de Prato, Giuditta; Simon, Jean Paul

Conference Paper

Is data really the new "oil" of the 21st century or just another snake oil? Looking at uses and users (private/public)

26th European Regional Conference of the International Telecommunications Society (ITS): "What Next for European Telecommunications?", Madrid, Spain, 24th-27th June, 2015

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: de Prato, Giuditta; Simon, Jean Paul (2015) : Is data really the new "oil" of the 21st century or just another snake oil? Looking at uses and users (private/public), 26th European Regional Conference of the International Telecommunications Society (ITS): "What Next for European Telecommunications?", Madrid, Spain, 24th-27th June, 2015, International Telecommunications Society (ITS), Calgary

This Version is available at:

<https://hdl.handle.net/10419/127134>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

What next for European telecommunications?
 26th European Regional Conference of the International Telecommunications Society,
 San Lorenzo de El Escorial, Spain, 24th – 27th June 2015

**Is data really the new
 “oil” of the 21st century¹ or just another snake oil?²
 Looking at uses and users (private/public).**

*Giuditta de Prato, EC JRC IPTS (Spain), Jean Paul Simon, JPS Public Policy Consulting,
 Spain*

*“Don’t ask for meaning; ask for the use”
 Usually attributed to Wittgenstein (*Philosophical Investigations*)*

Abstract:

The best “guesstimates” of the total amount of data in the world suggest that from 1987 to 2007 the total amount of analogue and digital data in the world grew from 3 billion gigabytes to 300 billion gigabytes. The so-called data explosion is driven by the combination of exponentially expanding amount of data available, the rapidly improving ability to process and analyse the data, and the deployment of the relevant infrastructures (broadband and ultra-high broadband networks, server farms, and devices such as smartphones, tablets and phablets).

This research paper aims at marshalling facts about a notion that have been spreading quickly without being, most of the times, properly defined thereby remaining vague and all-encompassing. This article tries first to put the phenomenon into perspective, it then takes a closer look at some leading users of big data and introduces some of its uses for evidence-based policy-making.

Key words: Analytics 3.0, Big Data, Fat Data, Digital Dragons, ICT Internationalisation Internet of Things, Machine-to-Machine communication, Microlevel Analysis, Policy 3.0, R&D.

The paper is built around to main parts. The first part which serves as an introduction, gives an overview of the phenomenon described as big data. It introduces some indication about its size.

However, neither the mere indication of the growth of the volume of data nor the more uncertain figures about the size of market(s) may be sufficient to understand the nature of the process. The question is what this phenomenon means, or as noted by the US Executive Office Report on big data: “*What really matters about big data is what it does*” (United States Executive Office, 2014a: 9). Therefore, the second part concentrates on users and uses.

The first section of the first part gauges the size of the data involved (market, volume), it offers glimpses of some of the assessment of the (potential) market linked to “Big Data”.

¹ The oil of the 21st century seems to proliferate: "Intellectual Property is the oil of the 21st century", Mark Getty, chairman of Getty Images, Water as the Oil of the 21st Century (Cleantech Alliance), and of course data.

² The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of their institutions.

The section offers a tentative definition of what the somewhat fuzzy notion covers while tracing back an early definition identifying a three dimensional data growth (3Vs): volume (amount of data: petabytes or above), velocity (speed of data in and out needed for real-time collection/analysis of data), and variety (range of data types, formats and services, collected from a variety of collection mechanisms). It introduces some of the basic components of “big data” as well as the value chain.

The second section questions the substance of the phenomenon trying to better underline its real present scope, to investigate beyond the present hype. It reveals that the benefits of big data are not always clear today, that the amount of valuable useful data is still low. Besides it highlights that despite industry hype, its real state of deployment is in average rather low, most organization have still to develop, implement or execute a big data strategy, as organizations continue to be wary of its impact. The third section reviews some of the major initiatives taken recently by governments in the EU (2014) and the US (2012).

The second part then aims at explaining what the phenomenon really means, fleshing out its real content through some field applications stemming from various players (users) as well as some research conducted by one of the author (uses). The first section of the second part starts with the IT players, so-called “digital dragons” (firms like Google, eBay, LinkedIn, and Facebook but also Alibaba or Tencent), firms that arguably were built around big data from the beginning. It concentrates on the case of Amazon. The paper considers other value added content providers with a focus on creative industries (Amazon, Klopotek, Netflix, Next Big Book, Pandora, Reed Elsevier, Zynga...).

The second section sums up an earlier attempt of using visualization techniques and data treatment to improve the production of scientific evidence to support policy making. It presents a research exercise by JRC IPTS IS Unit team for EC DG Connect dealing with the mapping of the European IT poles of excellence. The section closes with a quick overview of the IPTS GeoDIT project.

The paper concludes summing up the main output but delineating some limits of big data, identifies the issues the challenges ahead

The paper is based on desk research, a review of literature, review of the technical journals, analysis of annual reports, and meeting with experts and industry participants. It builds as well on the review of a "trial" project undertaken in 2014 in order to test tools and methods by applying them to real research issues so to test how a big data based approach could change the type and quality of the answers when techno-economic analysis is at stake. The paper is part of on-going set of research projects on R&D expenditures, patent analysis, internationalisation of ICT, and global networking.

I. Big data: an overview

I.1. Big amount (of data)/ Scarce deployment (so far)? The zettabytes³ waltz.

Big data can be traced back from a Meta Group (2001) report identifying a three dimensional data growth (3Vs) that are often quoted by other reports: volume (amount of data: petabytes or above), velocity (speed of data in and out needed for real-time collection/analysis of data), and variety (range of data types, formats and services, collected from a variety of collection mechanisms).

³ ExaByte, 10¹⁸ bytes, GigaByte, 10⁹ bytes, MegaByte, 10⁶ bytes, PetaByte: 10¹⁵ bytes, Zettabytes 10²¹bytes.

A decade after, in 2011, a widely-read McKinsey report on big data seems to have triggered part of what can be looked upon as some kind of hype. Other consultancies have been riding the same wave since, pumping up figures, urging the data-naïve to wise-up. However, the McKinsey report was cautious enough, providing a first tentative overview, dealing with selected case studies to illustrate the scope of activities and sectors implied.

To put it more bluntly like Harford (2014): “As with so many buzzwords, “big data” is a vague term, often thrown around by people with something to sell”. Or as Boyd and Crawford (2012: 3) put it, the notion blends technology (computation power and algorithmic accuracy), analysis (drawing on large data sets), and mythology: “the widespread belief that large data offer a high form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy”.

Indeed, there is no agreed definition of big data, no singular, internationally recognised definition of what constitutes ‘big data’, despite attempts and work being done at the ITU (ITU, 2013⁴). The terminology employed for the description is not an operational one and, as such, it makes it difficult to come up with clear views about what could constitute a distinct sector, occupation, process, etc. Hence the variations that are to be found among the reports. One way to flesh out the real meaning of big data is to document the way it is being dealt with through case studies as we will see in the next section.

Mummy where are big data coming from?

One can nevertheless give some indication about the amount of data it refers to as well as describing to of the basis component of the big data value chain. The best “guesstimates”⁵ of the total amount of data in the world suggest that from 1987 to 2007 the total amount of analogue and digital data in the world grew from 3 billion gigabytes to 300 billion gigabytes, an increase by a hundred by two decades (Haire, A., J., Mayer-Schönberger, V., 2014:5). The so-called data explosion is driven by the combination of exponentially expanding amount of data available (up to 7 “zettabytes” predicted for 2015), and the rapidly improving ability to process and analyse the data (Boston Consulting Group, 2012: 7).

It is enabled by the deployment of the relevant infrastructure (networks and devices). There are over 1.2 billion smart phones in the world, these devices are stuffed full of sensors and data collection features. One of the simplest indicator of the growth is the dramatic increase of the mobile traffic data that Cisco is monitoring, with its Cisco Visual Networking Index, signalling at the same time the dominance of video (nearly 79% of the total traffic predicted for 2018, Cisco VNI 2014: 14), and the leading role of consumers.

The data originate from various and heterogeneous different sources like people, machines or sensors. As emphasized by the UN Global Pulse, big data is both the information that is passively generated as by-products of people’s everyday use of technologies and the information people willingly communicate about themselves on the web⁶. “Data fusion,”

⁴ For further information on the work on big data carried out by the ITU Telecommunication Standardization Bureau (TSB), see <http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx>

⁵ One should be very careful about the data on “big data”, the figures collected and provided in this section are just meant to give some indications about trends. See section 2 for the lack of definition of the core terminology. As noted by the eSkill UK /SAS 2013 report: “reported adoption rates vary significantly, and in most cases observed are subject to significant caveats not always readily highlighted within the associated study documents” (e-skills UK/ SAS, 2013).

⁶ <http://www.unglobalpulse.org/about/faqs>

brings together disparate sources of data, “digital data” and “analogue data” (emanating from the physical world, but increasingly converted into digital format).

According to the EMC-IDC annual study (EMC Digital Universe study, 2014) the data created and copied annually will reach 44 zettabytes, or 44 trillion gigabytes. What they call the “digital universe” is growing 40% a year into the next decade, not only the increasing number of people and enterprises doing everything online, but also all the “things” (i.e. smart devices) are included.

Indeed, the combination of the devices and the relevant networks paves the way for a fastest growth of the “Internet of Things” (OIT⁷), heralded for some time, which is expected to generate huge amount of data (like with wearables⁸) through a networks of sensors and actuators. The pervasive integration of semiconductors, mobile communication, and “big data”/analytics is held as a way of propelling the Internet of Things (IoT) into the wider economy. McKinsey (2013: 4) predicts a potential economic impact by 2025 across all applications ranging between, US \$ 2.7 and 6.2 trillion. The growth of Machine-to-Machine (M2M)⁹ communication is contributing to this expansion of data produced. According to the Ericson Mobility Report (2014:28), M2M communication is taking off, driven by declining costs, improved coverage, more capable radio technologies, regulatory mandates and a growing range of successful applications and business models. The 2014 Ericson report estimates that, at the end of 2013, there were around 200 million cellular M2M devices in active use, a number is expected to grow 3–4 times by 2019. SAP (2014) predicts 2.1 billion connected devices worldwide by 2021 (SAP, 2014: 6).

The global market for “big data” was estimated at 6.3 billion US \$ in 2012 and is expected to reach US \$ 8.9 in 2014, but 48.3 billion by 2018 (a CAGR of 40.5% from 2012 on to 2018) (Transparent Market Research, 2013). Another specialised consultancy, IDC predicts that the market for big data will already reach \$16.1 billion in 2014¹⁰, growing 6 times faster than the overall IT market (quoted by Press, 2013). According to IDC, the “big data” market will grow to \$32.4 billion by 2020 (with an astonishing 212 billion devices connected up to the Internet of Things) (quoted by ??2014). IDC also predicted that by sub-segments 2020, the entire ICT industry will spend \$5 trillion, over \$1.3 trillion more than it does today, from which, “40% of the industry’s revenue, and 98% of its growth will be driven by 3rd Platform technologies¹¹ that today represent just 22% of ICT spending.” (IDC, 2014). Figure 1 gives the big data market forecast by segments over the period 2011-2017 (Wikibon, 2014) and clearly shows the strength of professional services.

Figure 1. Market by segments (2011-2017).

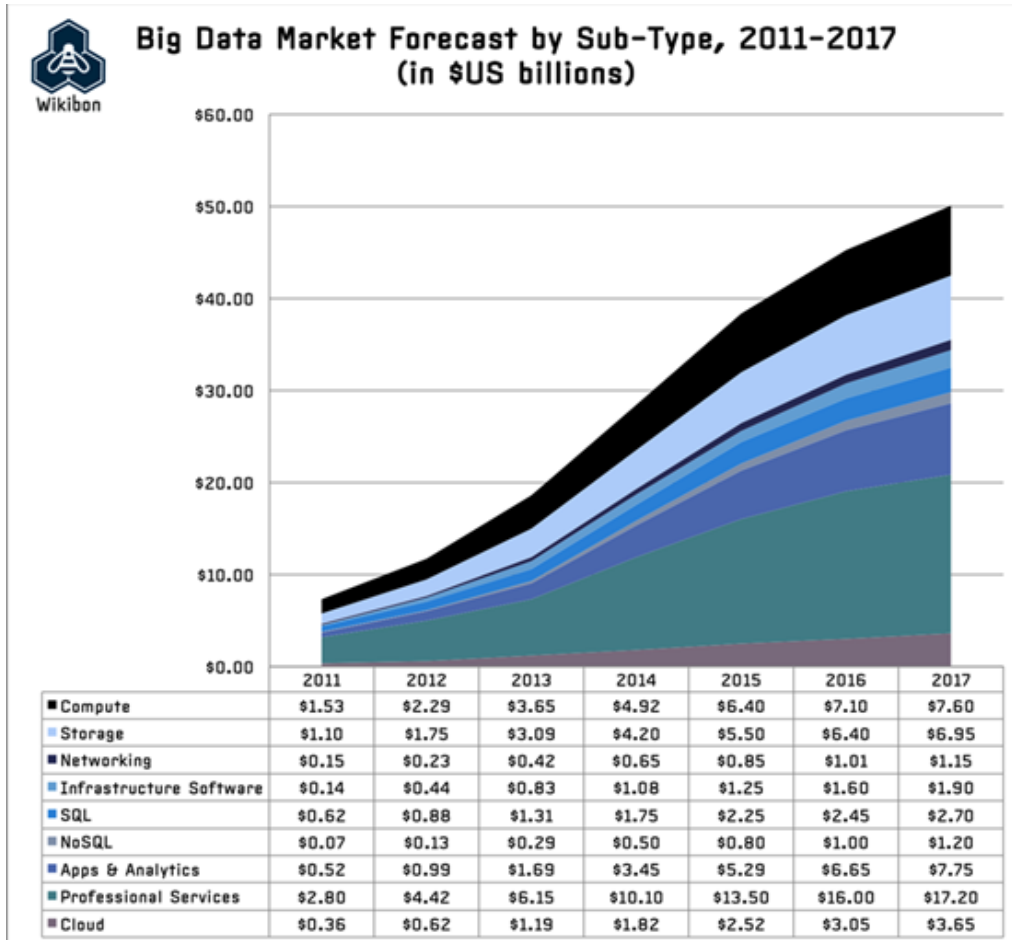
⁷ A blanket term referring to everyday objects like refrigerators, coffee machines, and TVs gaining network connections that enable them to talk to each other via standard protocols (i.e. the internet).

⁸ Apple released its watch in April 2015. Some Android Wear-rivalling smartwatches are or will be available. It remains to be seen what apps developers will be trying to connect with any smartwatch sensor.

⁹ Standards organisations formed OneM2M in 2012, with the aim of developing a common platform which could be used by service providers across a number of sectors, including smart grids, connected cars, eHealth, enterprise supply chain, home automation and energy management, and public safety. In 2015, South Korean operators SK Telecom (SKT) and LG U+ will make the first commercial Internet of Things deployments based on specifications from oneM2M.

¹⁰ IDC includes in this figure infrastructure (servers, storage, etc., the largest and fastest growing segment at 45% of the market), services (29%) and software (24%).

¹¹ They define the “3rd Platform” as built on the mobile network and apps market, cloud services, social media technologies, and big data analytics.



Source: Kelly (2014), Wikibon.

Bain (2013a) describes big data, as “*the mining and processing of petabytes worth of information to gain insights into customer behaviour, supply chain efficiency and many other aspects of business performance*” (Bain2014a: 1). Mc Kinsey (2011: 1) states that big data in many sectors will range from a few dozen terabytes to multiple petabytes (thousands of terabytes). The order of magnitude varies according to the source, letting “bytes” waltz. To the early 3Vs, other sources¹² are now adding veracity (the believability of the data itself) (Haire, A., J., Mayer-Schönberger, V., 2014: 6), visualization (Tascon, 2013), variability (temporal data peaks) and complexity (issues relating to linking/cleaning/editing data from different sources) (e-skills UK/ SAS, 2013:7).

Box 1. The big data value chain

The big data value chain can be broken down in the following fashion:

- Data Acquisition: Structured data, Unstructured data, Event processing, Sensor networks, Protocols, Real-time, Data streams, Multimodality.
- Data Analysis: Stream mining, Semantic analysis, Machine learning, Information extraction, Linked Data, Data discovery, whole world” semantics, Ecosystems, Community data analysis, Cross-sectorial data analysis

¹² Dataflog glossary (2015) goes up to seven Vs.

- Data Curation: Data Quality, Trust / Provenance, Annotation, Data validation, Human-Data Interaction, Top-down/Bottom-up, Community / Crowd, Human Computation, Curation at scale, Incentivisation, Automation, Interoperability.

- Data Storage: In-Memory DBs, NoSQL DBs, NewSQL DBs, Cloud storage, Query Interfaces, Scalability and Performance, Data Models, Consistency, Availability, Partition-tolerance, Security and Privacy, Standardization

- Data Usage: Decision support, Prediction, In-use analytics, Simulation, Exploration, Visualisation, Modelling, Control, Domain-specific usage.

Source: BIG, http://big-project.eu/sites/default/files/BIG_Introduction.pdf

The basic components of “big data” include software, hardware, and storage, with software and service being the larger share (IDC 2012: 3) (see figure 3, the Netflix architecture). New tools to deal with the data (extract, load, and transform) are emerging. Big data is grounded in technologies like Apache Hadoop¹³ (see box 2 for an illustration: the Hadoop platform, and again figure 3) and NoSql (Not Only SQL) (Akamai, 2014). The first one is an architecture platform, an open-source software framework that supports data-intensive distributed applications running on large clusters of commodity hardware. The second is a selective data based system for data to be retrieved easily, offering a quick and easy access. NoSQL databases are next generation databases, often “non-relational”, distributed and open-source as well as being horizontally scalable.

Box 2. The Hadoop Platform

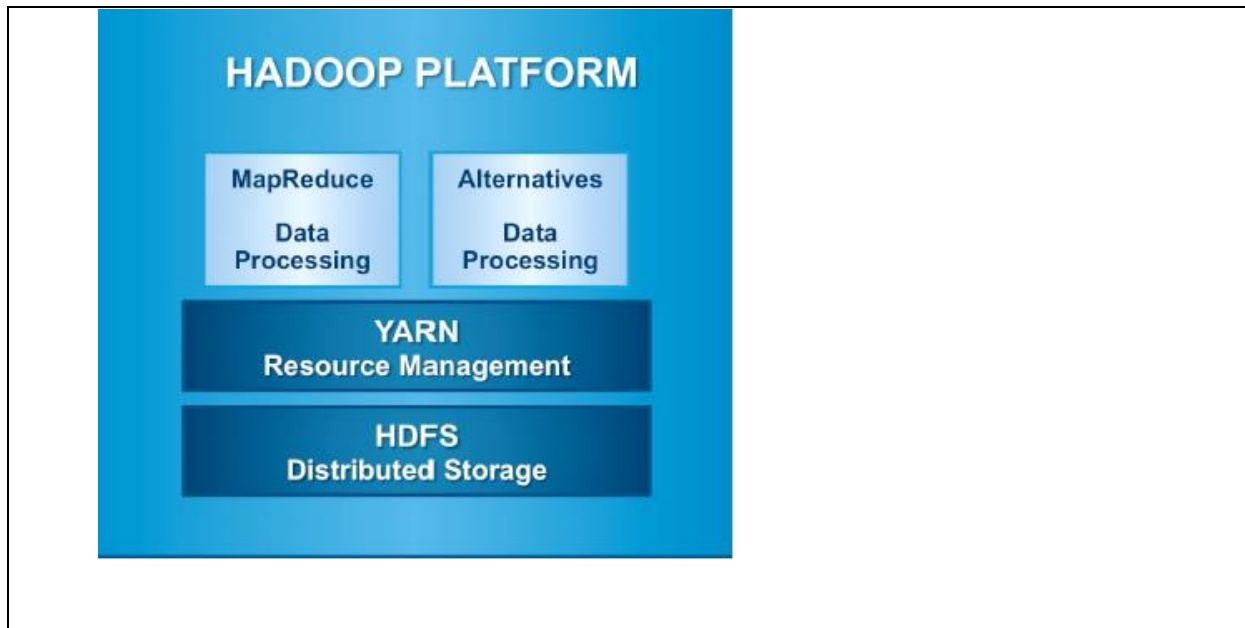
Hadoop is a Free Java software framework that supports distributed applications running on large clusters of commodity computers that process huge amounts of data. Hadoop consists of an open source implementation of Google's published computing infrastructure, specifically MapReduce and the Google File System (GFS).

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. The Hadoop framework is composed of the following modules:

- Hadoop Common contains libraries and utilities needed by other Hadoop modules,
- Hadoop Distributed File System (HDFS) is a distributed file system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster,
- Hadoop YARN is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users' applications,
- Hadoop MapReduce is a programming model for large scale data processing. It divides applications into smaller components and distributes them across numerous machines,
- Other open source components that augment HDFS, MapReduce and Yarn are Pig¹⁴, Hive, Hbase, etc..

¹³ See *Le Guide du Big Data. 2014-2015* for an historical account of the company: pp.24-28.

¹⁴ See Figure 3.



Source: SAS (2015: 1), <http://hadoop.apache.org/#sthash.fPho82IG.dpuf>

One of the issues, linked to this variety dimension, is that not only data collected are heterogeneous but they are structured (or not) differently: 51% of data is structured, 27% of data is unstructured, 21% of data is semi-structured, according to a report from the Tata Group (2013: 19, see box 3). Hence the role of companies such as Hadoop, offering a framework, i.e. a collection of software tools and applications, designed to allow organisations of any size to store and analyse huge amounts of information.

Box 3 .The dimension of data structure:

- Structured (retail, financial, geodata...): data that is identifiable as it is organized in structure like rows and columns. Data that resides in fixed fields (for example, data in relational databases or in spreadsheets)
- Unstructured (images, videos, sensor data, web pages): Data that does not reside in fixed fields (for example, free-form text from articles, email messages, untagged audio and video data, etc.). It does not have a formal structure like structured data. It does however have tags or other markers to enforce hierarchy of records
- _ Semi-structured (weblogs, e-mails, documents...): Data that does not reside in fixed fields but uses tags or other markers to capture elements of the data (for example, XML, HTML-tagged text). Unstructured data is regarded as data that is in general text heavy, but may also contain dates, numbers and facts

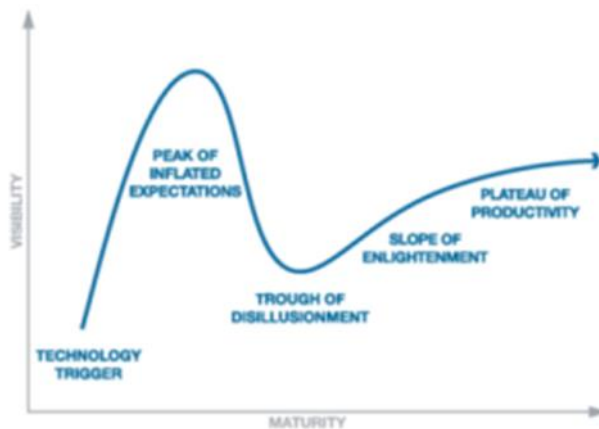
Source: compiled by authors from CSA (2014), Dataflop (2015), Tata (2013:18).

Uneven patterns of adoption and of use

IDC commented that the benefits of big data are not always clear today (Forbes, 2013). In sharp contrast with the impressive amount of data, the deployment of big data appear so far rather scarce and by and large uneven between sectors. In 2013, only 8 percent of organizations around the world have actually spent any money or time to build an actual big data application (such as Hadoop or a NoSQL cluster) in their shops, according to Gartner (2014) research findings. Gartner 2014 survey reveals this has increased to 13 percent in 2014. Woodie (2013) notes: “*the hype of big data has definitely exceeded the substance*”, but

adds following the Gartner's Hype Cycle (see figure 1) that we may be in what the consultancy calls a "*Trough of Disillusionment*" stage.

Figure 1. New technology adoption will often follow some variation of Gartner's Hype Cycle.

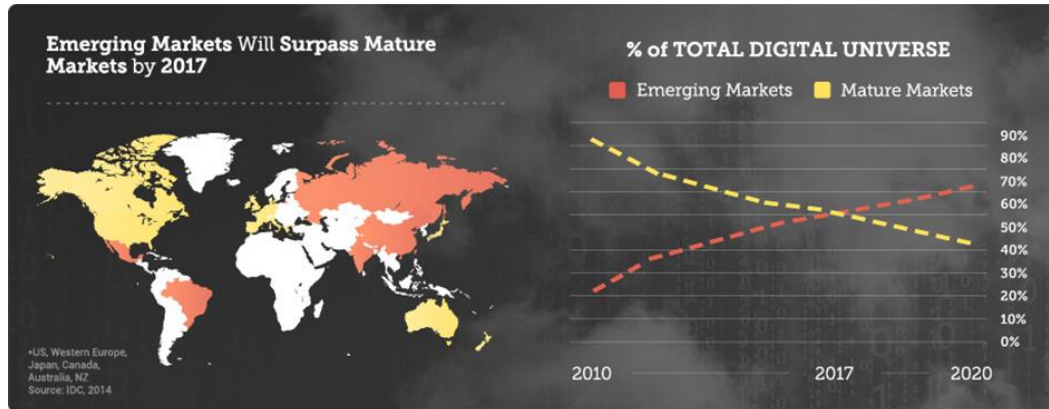


Source: Gartner quoted by Woodie (2013)

The McKinsey report already underlined that all sectors were far from being equal, that the propensity and likeliness of adoption vary. Focusing on the US, the report identifies (figure 2, Mc Kinsey 2011: 7) four clusters. The "usual suspects" (electronic products and information sectors) are sectors are set to gain substantially through access to huge pools of data (e.g., Internet companies collect vast amounts of online behaviour data). On the opposite, several sectors like construction, educational services, and arts and entertainment, "*have posted negative productivity growth*", the reports adds that it "*probably indicates that these sectors face strong systemic barriers to increasing productivity*". The Tata survey (2013:38) , two years later, find similar suspects: telecom, travel, high tech and banking firms spend the most on big data.

Not only one may notice inequalities between industrial sectors but inequalities between regions are also striking, as noted by Tata (2013: 7), big data was of particularly keen interest in certain countries: India, South Korea, the U.S., Australia, Canada, Western Europe, and Brazil. As noted by various sources, the US took an early lead (Tata 2013: 12, Hamel and Marguerit, 2013: 9). IDC reported to the EC that big data use was low in the EU, with only 6.9% usage amongst companies with more than 10 employees, and only 29% of European companies considering themselves ready for big data (IDC, 2013). Emerging markets are predicted to overpass (see Figure 2) mature markets in terms of % of the digital universe.

Figure 2. Evolution of the division of the digital universe between mature and emerging markets (2010-2020).



Source:IDC, 2014.

Besides, of the useful data, IDC estimates that in 2013 perhaps 5% was especially valuable, or “target rich.”(EMC IDC, 2014); adding in a more optimistic mode “*That percentage should more than double by 2020 as enterprises take advantage of new Big Data and analytics technologies and new data sources, and apply them to new parts of the organization*”.

By the same token, average M2M device penetration is around 2 percent of data subscriptions among measured networks, while it can reach 20 percent for those operators that focus on M2M. M2M communication represents a small share –around 0.1 percent – of total cellular traffic in terms of bytes. This traffic share will go up as LTE M2M devices and more powerful processors are included in high bandwidth and low latency-demanding applications such as consumer electronics, vehicles and billboards. However, the fast deployment of the new networks (4G, 5G¹⁵) will be key for the continuing growth of applications, to further unlock the development of all kind of connected devices, including connected cars. The deployment and its pace cannot be taken as granted though as there is a noted gap between the explosion of traffic and usages and limited streams of revenues especially in the EU. Therefore, it is likely to be unevenly spread globally with some region like Asia¹⁶ and the US¹⁷ taking the lead.

According to the SAS 2013 Big Data Survey Research Brief, despite industry hype, most organization have still to develop, implement or execute a big data strategy: only 12% of the surveyed organizations (SAS, 2013: 1). The same report concludes that while big data is a common theme is the market, organizations continue to be wary of its impact. A 2014 EY report noted as well that the big data revolution did not take place so far for most companies:

¹⁵ Still being standardized, the deployment is forecast for 2018-2020 introducing software defined 5G networks for “Anything as a Service”. Huawei predicts that 5G (10 Gbps) opens the “Internet of everything”, bridging human non-human, virtual and physical communities: by 2020, 10 billion mobile terminal, 100 billion global wireless connections, lower latency of 1 millisecond. *Source:* http://www.mobileworldlive.com/huawei-finalise-4-5g-branding-q1-commercialise-2016?utm_campaign=MWL_20141121&utm_medium=email&utm_source=Eloqua&elq=87ddd61125be487d834568408a84d783&elqCampaignId=2727

¹⁶ China Mobile for instance is leading with 4G with already over 500,000 4G base stations (China Telecom only). The increased scale has also driven down the cost of 4G smartphones, which are now below \$100 in China, and that has broadened the audience. *Source:* http://www.mobileworldlive.com/asias-scale-4g-rollouts-drives-innovation-cost-cuts?utm_campaign=MWL_AS_20141125&utm_medium=email&utm_source=Eloqua

¹⁷ In 2014, Google and Verizon announced that they are testing the capabilities for currently installed fiber networks to carry data even more efficiently—at 10 gigabits per second—to businesses that handle large amounts of Internet traffic.

63% of the French companies surveyed consider the notion was as too vague (EY, 2014: 27). As the Tata survey (2013:9) sums it up: “*a minority of companies spending massive amounts and a larger number spending very little*”.

Earlier, the McKinsey report (2011) was acknowledging the same issue about the scope of implementation when stressing that a company’s “data-driven mind-set” was to be a key indicator of big data’s value to companies. Obviously, one cannot expect all companies to be data-savvy and to display this specific mind-set, which means that this is likely to become a barrier to entry or even to an appropriate implementation. This will require new expertise¹⁸ and training. Besides, the new expertise is not exactly cheap with Silicon Valley siphoning most of it as noted for the case of the book publishing industry (McIlroy, 2014: 8) tilting the balance in favour of the larger organizations.

This is clearly why consultancies dealing with the issue are suggesting ways to initiate the transition. As Davenport and Tyche (2013:30) put it: “*organizations need to begin transitioning now to the new model*” (a model they call Analytics 3.0). They add that “*It means change in skills, leadership, organizational structures, technologies, and architectures*”, such a drastic change will require some time as the example of the “computerisation” of companies has illustrated¹⁹. As E.Brynjolfson rightly notes (quoted by McIlroy, 2014: 8) that in order to deal with big data to become competitive: technology is the prerequisite but not the harder, then comes the acquisition of the skills required which does not come easy, and the last and more difficult to achieve appears to be the cultural changes needed in organization.

I.2 Policy initiatives: the EC and the EU

Taking into account the expected potential benefits from big data, governments have started setting up policies. Part of the initiatives are designed to address the issue of the skills needed as reports and consultancies are pointing to the lack of expertise in the field. Prioritising data science in education and in training for early-career researchers is climbing up on the policy agenda as the lack of the requested expertise is regularly pointed at (e-skills UK, 2013). Other areas are being targeted like medicine with data-driven medicine²⁰.

In March 2012, the White House Office of Science and Technology Policy (OSTP) announced a Big Data Research and Development Initiative (also called the “Data to Knowledge to Action” initiative): six U.S. government agencies were allocated over \$200 million to help the government better organize and analyse large volumes of digital data. The initial goals were threefold:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

¹⁸ The 2011 Mc Kinsey report was stating that 1.5 million more data-savvy managers were needed to take full advantage of big data in the United States.

¹⁹ The literature on the topic is rather vast identifying the barriers, tensions; power struggles and the time-span involved.

²⁰ See the adoption in the US in 2009 of the HITECH Act: taxpayers and the medical industry have collectively invested more than \$100 billion in an information technology infrastructure.

Two years later the year, the US Executive Office released two reports²¹ (US Executive Office, 2014a, b): one dealing with the discussion on privacy, the other adding a Technology Perspective. A survey organised for these report revealed that respondents expressed a great deal of concern about big data practices, and communicated particularly strong feelings around ensuring that data practices have proper transparency and oversight

The report (US Executive Office, 2014b) concludes that “*technology alone cannot protect privacy*”, and that policies are needed to protect privacy. The research and development of privacy enhancing technologies has been indeed already (NITRD, 2014). a priority for the Obama Administration. Agencies across the Networking and Information Technology Research and Development (NITRD) program collectively spend over \$70 million each year on privacy research.

Some aspects of big data were already included in the 'Technologies for Information Management' portfolio that combined projects from the 7th Framework Programme (FP7) and the Competitive and Innovation programme (ICT-policy support programme)²². In July 2014, the European Commission outlined a new strategy on big data so as to support and accelerate the transition towards a data-driven economy in Europe. The coordinated action plan, between Member States and the EU, was focusing on economic aspects: accelerated innovation, productivity growth, and increased competitiveness. It was followed in October of that same year by the creation of a public-private partnership (PPP), a €2.5 billion partnership to master big data. The press release²³ stressed that: “*Mastering big data could mean: -up to 30% of the global data market for European suppliers; -100,000 new data-related jobs in Europe by 2020; -10% lower energy consumption, better health-care outcomes and more productive industrial machinery*”.

Besides, in Europe the “Big Data Public Private Forum (BIG: see Box 4) was set up to contribute to the definition and implementation of a clear strategy to define the efforts in terms of research and innovation. The final report notes that “*Europe in general has more appetite for personal data protection than the US, meaning that the needs of companies to innovate have to be balanced against the rights of citizens to have their data protected*” (Roadmap, 2014: 6). However, in that particular case, the strategy on big data does not appear to be accompanied by a report similar to the Podesta report.

Box 4. The “Big Data Public Private Forum” (BIG) (2012-2014).

The BIG mission can be summarized as follows:

- Build a self-sustainable Industrial community around Big Data in Europe:
 - Technical level establishing the proper channel to gather information,
 - Industrial-led initiative to influence adequately the decision takers.
- Promote adoption of earlier waves of big data technology,
- Tackle adequately existing barriers as policy and regulation issues.

Outcomes of this 3 years project (started in 2012-2014), funded by the European Commission (around 2.5 million euros), will be used as input for *Horizon 2020*. The group elaborated a sector roadmap available on the website. The roadmap provides key-findings, and covers the following sectors: healthcare, the public sector, finance and insurance, telecom, media and entertainment industries, energy, and transportation, manufacturing, and retail.

²¹ Podesta’s report. Prepared by the President’s Council of Advisors on Science and Technology (PCAST): www.whitehouse.gov/ostp/pcast

²² A portfolio of data projects aiming for more effective and efficient management of big data. See “Project Factsheets: data”: <http://ec.europa.eu/digital-agenda/en/node/72768>

²³ : http://europa.eu/rapid/press-release_IP-14-1129_en.htm

Source: <http://big-project.eu/>

The Executive Office of the United Nations Secretary-General, launched UNPulse Initiative to respond to the need for more timely information to track and monitor the impacts of global and local socio-economic crises, explore how new, digital data sources and real-time analytics technologies can help policymakers. Global Pulse is a flagship innovation initiative on big data. Global Pulse is working to promote awareness of the opportunities big data presents for relief and development, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs²⁴, and drive broad adoption of useful innovations across the UN System (UNPulse, 2014). Still at the UN level, the ITU has been doing work on big data carried out by the ITU Telecommunication Standardization Bureau (ITU, 2014a), and on regulation, holding meetings.

On both sides of the Atlantic, governments are dealing with the issues of data availability and ownership, of access to data generated through public funds. The US National Science Foundation in the US and the EU Horizon 2020 framework programme contain provisions for open data management in research projects.

On the scientific side, the International Council for Science (a non-governmental organisation that aims to strengthen cooperation, address major science issues), stressed the need to improve the understanding of big data sets through more close-knit global cooperation. It also calls for an improvement of policies and international guidelines surrounding big data (Codata, 2014).

II. Of users and uses.

II.1 Digital dragons and content providers

This section gives some examples of sectors and companies that are implementing a big data approach, but it concentrates in IT companies and content providers. These new players are betting on innovation in audience reach, looking for scale, and scale obviously means more data.

It opens up with the new players in this area, IT players, focusing on Amazon. The first organizations to embrace big data were online and start-up firms. Arguably, firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning. Indeed IT players, labelled “the digital dragons” by Atelier Parisbas (2013:78), are intrinsically well placed to benefit from this shift, to make the most of new analytics, of big data and the cloud (Amazon together with Google and Microsoft is a leading provider of third party cloud computing services with AWS). Furthermore, the Data Dragon have been leading in shaping the technologies, Cassandra was inspired by Facebook and Hadoop by Google.








For instance, Google now processes over 40,000 search queries every second on average, which translates to over 3.5 billion searches per day and 1.2 trillion searches per year worldwide. When Google was founded in September 1998, it was serving ten thousand search queries per day (by the end of 2006 that same amount would be served in a single second). In

²⁴ Pulse Labs bring together government experts, UN agencies, academia and the private sector to pioneer new methods and frameworks for using big data to support development goals.

September 1999, one year after being launched, Google was already answering 3.5 million search queries daily (Google Search Statistics, 2015).

Adopting targeted advertising as a business model required the ability to make the most out of the data collected through billions of searches that reveal the wants, needs, fears, and obsessions of users, what J. Battelle calls the “database of intentions” (2006: 1). This applies to the search aspect, but since then Google has introduced new products and services that are adding new set of recorded data (see Figure 3) about customers.

Figure 3. Recording data/ profiling customers.

services	recorded data	how it defines your profile
 search	your search patterns (volume, frequency), keywords and phrases	center of interests, passions, hobbies, likely occupation, level of education, location, intentions
 +  social mail	semantic and tone analysis of posts & mail, address book, friends	level of social activity, circles, general mood, state of mind, health condition, family situation, health condition, income indications, spending
 news	all the stuff you watch and click on	everything about your media diet, relation to news brands, favorite subject, media and authors your trust
 calendar	meetings and travel patterns	level and type of business activities, future business and personal trips
 +  location cell phone	gps data, commute length, traffic conditions, call patterns	social activities, travel data, much more personal stuff depending on application loaded

Source: F.Filloux, (2014).

We then follow developments in the content industries with examples from new disruptive players in the games industry and the distribution of video.²⁵ To accompany the changes first from a rental model to an ownership model, and then to a direct consumption model, these companies have shifted from a logic of supply, of prototypes deeply rooted in the culture of legacy players, toward a logic of demand. The role of big data in the shift has been pivotal as these new digital players also pioneer the use of data mining for compiling viewers’ recommendations (Amazon, Netflix, Pandora, Zynga...). The section closes with an overview of the book publishing industry, an industry characterised by sharp contrasts between sub-sectors and companies, which shows the role on new intermediaries, specialised technology companies (STC).

²⁵ Other examples can be found in De Prato and Simon (2015). For a comprehensive review see Mc Kinsey (2011), Tata (2013).

In China, Alibaba Pictures, the subsidiary of the Alibaba platform, expects to be able to create customized movies and TV programs while marketing and distributing them efficiently across Alibaba's platforms, by using big-data technology to analyze consumer shopping patterns and behavior on (e-tailers) Taobao and Tmall (Frater, 2015). By the same token the four Chinese Internet behemoths (Alibaba from e-commerce, Baidu from Web search, Tencent from social media, and Wanda from property development Alibaba) have reached a point where big data and network mass are being used to change how entertainment is conceived, sold and consumed.

Amazon: the strength of a sophisticated ecosystem

When Amazon.com opened its virtual doors on the World Wide Web in July 1995 J. Bezos (CEO and founder) made it clear: "We seek to be Earth's most customer-centric company". (Annual Report 2014: 3). Amazon, is one of the fastest growing business in the Internet's history (Simon, 2015, Distinguin, 2013, Blackman and Forge, 2012).

Being customer-centric meant data for J. Bezos, before Google, and long before Facebook, had realized that the greatest value of an online company lay in the consumer data it collected. Amazon intended to sell books as a way of gathering data on affluent, educated shoppers. The books would be priced close to cost, in order to increase sales volume. After collecting data on millions of customers, Amazon could figure out how to sell everything else at bargain prices on the Internet. Books were going to be the way to get the names and the data. The sales of books was the customer-acquisition strategy. Book sales in the U.S. now make up no more than seven per cent of the company's roughly seventy-five billion dollars in annual revenue, media accounts for 25% of the total sales (88, 98 US\$ billion) in 2014 (Annual Report 2014: 26).

The key building blocks of Amazon's success are their ability to use data and an eye for the right innovations and patents. When Amazon was primarily a book retailer, the company was the first to extensively use algorithms so that it could provide recommendations for customers: "*Customers who bought this item, also bought this one...*". The customer review section is one long-term feature, that has helped Amazon to understand its customers and so become the largest Internet retailer in the world. Using its data mining and profiling tools, the giant from Seattle tries to detect market trends early and then translates those trends and needs into new products and services. Management pays careful attention at the idea/creation stage to what customers are looking for in the products they choose. Forge & al (2013: 61) noted that "*Amazon is adept at reading the market, pursuing customers progressively into new areas, preferring to innovate incrementally while keeping a close eye on the innovations of competitors*".

On the distribution side, created new ways of buying and distributing books and went further with the introduction of its Kindle e-reader (De Prato and Simon, 2012: 13, Benghozi and Salvador, 2013: 22), a far cry from just shipping books from a warehouse in Seattle (Forge & al, 2013: 60). Hence the introduction of its own device, the Kindle reader, seemed logical to organize and structure its ecosystem. Amazon introduced his Kindle e-reader dealing directly with customers. Amazon is making the best out of the sophisticated ecosystem (recommendations, tools for self-publishing, on demand publishing, 14 imprints under the flagship of Amazon Publishing, a leading community of readers, Goodreads...) built around its Kindle since 2007.

The giant from Seattle is now trying to duplicate its successful experience around a new ecosystem for audio-visual contents based on its new device Fire, launched in 2014. Amazon

is now introducing its Android-based smartphone²⁶ Fire, with capabilities of sending TV shows and movies to Fire TV devices (the TV service of Amazon²⁷), giving mobile access to the company ecosystem with more than “33 million songs, apps, games, movies, TV shows and books” (Ferguson, 2014a).

As emphasized by Forge and Blackman (2012), Amazon, despite its apparently virtual presence in cyberspace, is very much a hardware company, needing heavy amounts of capital investment for the two key business processes:

- Its sales interfaces, which rely on some of the world’s largest data centres with intensive use of computer hardware and storage, running its order-taking, statistics, sales, cataloguing and customer profiling engines on a base of cloud –computing.
- Logistics chains of dispatch centres for warehousing and ground transport with a host of delivery services partners.

Amazon is moving into the production of contents (Simon, 2015) first with book production, and more recently with audio-visual but unsurprisingly using a data-based approach. Amazon Publishing is the full-service publishing arm of Amazon is now composed of 14 imprints²⁸. AmazonEncore imprint launched in May 2009 is interesting to look at as this is program whereby Amazon will use information such as customer reviews on Amazon.com to identify exceptional, overlooked books and authors with more potential than their sales may indicate. By producing its own original work, Amazon can sell more devices and sign up more Prime members, a major source of revenue.

Amazon Studios is Amazon.com's division, created in 2010, that develops comics, movies and television shows from online submissions and crowd-sourced feedback, is using its strength in data collection, an unusual way of producing television series that Netflix pioneered not so long ago. Amazon invited writers to submit scripts on its Web site, described as “*an open platform for content creators,*” Amazon put the pilots on its site, where customers could review them and answer a detailed questionnaire. (“*Please rate the following aspects of this show: The humor, the characters . . .*”) More than a million customers watched. Amazon Studios had received more than 10,000 feature screenplay submissions as of September 2012, and 2,700 television pilots as of March 2013.

Netflix: a data-driven ecosystem for Internet-connected screens.

²⁶ Fire runs a 2.2GHz quad-core Qualcomm processor, has a 4.7-inch screen, a 13 megapixel rear-facing camera and 2.1 megapixel front-facing sensor, and supports nine LTE bands. It is available with 32GB or 64GB of storage and uses Amazon’s latest forked version of Android, Fire OS 3.5.0. Fire, features a display which shows three-dimensional images that responds to movement, along with image, text and audio recognition. A dedicated processor and real-time computer vision algorithm then adjusts the image accordingly. The technology allows apps and games to be more immersive.

²⁷ The Amazon Fire TV box allows users to download and use app and games — as well as stream movies, television and music — to their HD TVs via their mobile device. *Source:* Ferguson, 2014b.

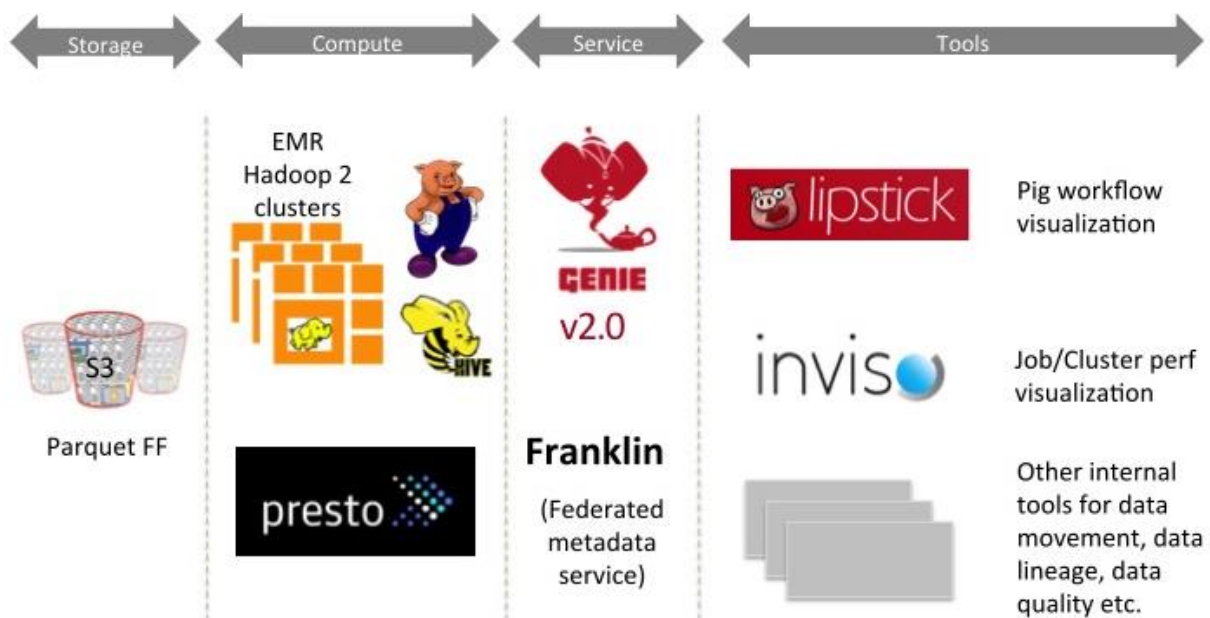
²⁸ AmazonEncore (Rediscovered Works), AmazonCrossing (Translated Works), Thomas & Mercer (Mystery, Thrillers, and Suspense), Montlake Romance (Romance), 47North (Science Fiction, Fantasy, and Horror), Little A (Literary Fiction), Skyscape (Teen and Young Adult), Two Lions (Children's Picture Books, Chapter Books, and Novels), Jet City Comics (Comics and Graphic Novels), Lake Union Publishing (Contemporary and Historical Fiction, Memoir and Popular Nonfiction), StoryFront (Short Fiction), Grand Harbor Press (Personal Growth and Self-Help), Waterfall Press (Christian Nonfiction and Fiction), Amazon Publishing (Nonfiction, Memoirs, and General Fiction).

Netflix provides an interesting case²⁹ of a niche provider, using a very antiquated distribution channel (the mail) to send VHS then DVD, later morphing into a global IT company. Netflix claimed to be (Annual Report, 2014: 1) “the world’s leading Internet television network with over 62 million members in over 50 countries enjoying more than 100 million hours of TV shows and movies per day³⁰”.

The company interacts now with various players within a complex network of commercial relationships, a network they described as “an ecosystem for Internet-connected devices” (Annual Report 2013: 1). Netflix designed its strategy under three main assumptions: Internet TV is replacing linear TV, Apps are replacing channels, and screens are proliferating. Accordingly, the company has built this ecosystem and sustained its growth around two linked technological developments: big data and the cloud. To offer instant streaming of content to various devices, the company uses services of third-party cloud computing providers -Amazon Web Services (AWS) in particular- and content delivery networks (Level 3 Communications) in order to stream this content to the consumer. Netflix has also built its own single-purpose content delivery network, Open Connect.

Over the past 7 years, Netflix streaming has expanded from thousands of members watching occasionally to millions of customers watching over two billion hours every month. Each time a customer starts to watch a movie or TV episode, a “view” is created in the data systems and a collection of events describing that view is gathered. Given that viewing is what customers spend most of their time doing on Netflix, having a robust and scalable architecture³¹ (see figure 3) to manage and process this data is critical. Their architecture enables Netflix to build a data warehouse of practically infinite scale in the cloud (both in terms of data and computational power).

Figure 3. Netflix big data platform architecture with Genie³² 2.0 at its core



Source: Netflix (2015), [http://techblog.netflix.com/search/label/big data](http://techblog.netflix.com/search/label/big%20data)

²⁹ See Annex B of the De Vinck, Lindmark cinema report (2012): 112-115. See as well for another case study, Grece (2014): 180-200.

³⁰ More than one billion hours of TV shows and movies per month (Annual Report 2013:1)

³¹ For a description of their current architecture see: <http://techblog.netflix.com/search/label/NoSQL>

³² Hadoop Platform as a Service.

Netflix has been transformed using big data analytics to ‘give people what they want’. Predicting what its customers will want to watch next is the primary goal of Netflix’s data strategy. The company is analyzing detailed viewing data from their 60 million subscribers to optimize their recommendations (Marr, 2015). To do this it has employed teams of movie buffs to scour years’ worth of film and TV and tag the common themes they contain. It also regularly runs crowd-sourced contests with million-dollar prizes for anyone who can come up with algorithms that more accurately predict audience viewing habits than its existing ones. Netflix uses a multitude of algorithms for doing personalization and recommendation including: how to predict the rating that a member will give a video, how to rank videos in each row, and how to create meaningful groupings of videos. Netflix started with rating prediction, evolved into personalized ranking of its catalog, and is now moving toward personalised page generation.

Netflix use personalization extensively and treat every situation as an opportunity to present the right content to each of its members. The main way a member interacts with the recommendations is via the homepage, the primary function being to help finding something to watch. The problem they face is that their catalog contains many more videos than can be displayed on a single page for each customer with his own unique set of interests. Thus, a general algorithmic challenge becomes how to best tailor each member's homepage to make it relevant, cover their interests and intents, and still allow for exploration of the catalog (<http://techblog.netflix.com/>)

Netflix is getting involved in content production, offering original programming such as Award winning “*House of cards*” (\$ 100 million cost of 13 episodes seasons), or original documentaries like “*Battered Bastards of Baseball*” and “*Mission Blue*”. Netflix produced such a hit show, “*House of Cards*”, after analysing the data from their consumer base (million “plays” per day, million searches, plus tags and other metadata) (Carr, 2013). In 2014 the company plans to substantially increase their investments in original content, to that end big data analytics will be used to find the next ‘*House of Cards*’ blockbuster.

The data driven road to 1 billion downloads: lessons from the games industry

The games industry has also been a pioneer in the field of big data, first with online games and since the last decade with mobile games especially social games for iOS and Android. Hit mobile games like *Candy Crush Saga*, from King Digital Entertainment, *FarmVille* from Zynga or *Angry Birds* from the Finnish company Rovio became success stories. The number of downloads reached by these games is often quoted as an indicator of such an achievement. “Angry Birds” franchise was launched in 2009, the Finnish start-up hit 1 billion downloads, with *Angry Birds*, in May 2012. However the link between these impressive figures and big data as a tool is not sufficiently highlighted.

This section concentrates on two examples, the well-known case of Zynga, described as “*A Big Data Company Masquerading as a Gaming Company*” (van Rijmenam, 2013), and the example of less well-known company that none the less hit the same 1 billion downloads mark, Storm8, selected as both companies explicitly stress the role of a data driven approach.

Zynga develops, markets and operates online social games as live services played over the Internet and on social networking sites and mobile platforms. Zynga games have been played by more than 1 billion people around the world and are available on a number of global platforms including Apple iOS, Google Android, Facebook and Zynga.com.

In its IPO prospectus, Zynga stated that “*Games should be data driven*” (IPO, 2011: 32). Indeed the company develops and operates the games as live services with daily, metrics-based player feedback. Zynga invested in Big Data and related technologies. In order to cope

with extreme high demands of data (on a regular day Zynga delivers one petabyte of content), Zynga has built a flexible cloud server centre that can easily add up to 1,000 servers in just 24 hours. Zynga's private and public cloud server park is known as one of the biggest hybrid clouds.

In other words Zynga is built around a metrics driven culture, as noted by van Rijmenam: "At Zynga everything revolves around metrics" (van Rijmenam, 2013), echoing the statement of the company: "Our culture combines the creative with the analytical" (IPO, 2011). Within, Zynga the designers are separated from those analysing the metrics. Analysts need to figure out what questions should be asked and the designers will develop/adjust the game around the answer. Their business model combines art (creating, developing and implementing an idea into a game) with science (listen to customers and find out whether the game is fun or not) to quote M. van Rijmenam stressing the example of the way *FarmVille* was modified to take into account customers' reactions: "In the original version of *Farmville*, animals were merely decoration. However, data showed that more and more people started interacting with the animals and even use real money to buy additional virtual animals. So, in *Farmville 2.0* animals were made much more central."

Besides, as stressed by Atelier Paribas (2013: 81), through the use of big data, Zynga intimately links game design and business models as the example of *FarmVille* reveals. As their business model is built on free to play games, the company generates revenue through the in-game sale of virtual goods and advertising (banner ads, video ads and product placement). Data analysis and mastering metrics become pivotal for the business models which was not the case for boxed games and even for on-line pay for games.

The business model of Storm8 is also the free-to-play model with virtual goods. Storm8 claims to be the leading developer of social games for iOS and Android, with more than 50 million monthly active users and more than 1 billion total downloads to date. The company from Redwood (Ca) also stresses the interaction with users: "keeping a laser focus on building and retaining our players who inspire us every day to develop the best games" (P.Tam, CEO, AppAnnie, 2015). The company opted for not relying on a hit game like Zynga or Rovio but to develop a portfolio strategy which meant that having mobile app market data became increasingly important in helping the company make intelligent roadmap choices, comparing offerings relative to the competition, understand overall market, game genres and geographic trends to better position new games and prioritize features. Developing games is based on handling data and constantly iterating on the feedback of customers.

Pioneers and laggards in the book publishing industry

These reviewed IT grown companies, digital natives, have resources to develop their own which is not the case of most content companies. If the cost of entry on the technological side does not appear to be too high, thanks to the scalability for instance of cloud solutions, it is only one aspect of the issue of shifting to the massive use of massive data. Technologies may be cheap enough but as noted expertise is not, which meant that for instance in the book publishing industry this shift will be triggered by the largest companies as stressed in a White paper from the book publishing industry (McIllroy/Digital Book World, 2014: 8). The challenge for the book publishing industry is to find appropriate ways to move from legacy printed content to digital content often more granular, frequently updated, on demand, searchable, hyperlinked, divisible and dynamic. Digital contents and more individualized geared to specific individuals needs rely on data.

Besides, within the three sectors of the book publishing industry, trade, educational and scientific, technical and medical (STM), some sectors are moving faster ahead like the STM which became digitised quicker and earlier including with e-books (De Prato and

Simon, 2012, De Prato, 2014). STM, taking into account its central role for research, may become the new publishing paradigm (see next section on evidence based policy making).

Reed Elsevier is a good example of such a fast-mover in the field. The company is a technology oriented scientific and technical book publisher, investing US \$ 500 million every year, claiming to be the fourth largest digital content provider in the world (Annual Report, 2013: 9) with global revenues of 7, 5 billion euros. Its High Performance Computing Cluster Systems (HPCC), HPCC Systems is one of the most advanced, fast-performing Big Data processing technologies available today according to the company. It was developed by LexisNexis Risk Solutions and currently powers core products from this division, which had 2013 revenues of £933 million. It is open source and used to solve large-scale, complex data and analytics challenges. HPCC Systems combines proven data processing methodologies with Reed Elsevier's proprietary linking algorithms

The latest version of SciVal, their scientific, technical & medical segment's tool for universities and other institutions to assess their relative performance, runs on HPCC Systems technology. SciVal provides analysis of over 30 million pieces of content and 350 million citations from 4,600 institutions in 220 countries. Lexis Advance provides lawyers with essential information and analytical tools covering all aspects of their daily work. The company claims it offers their customers ways to turn data into intelligence, better, faster and cheaper.

Smaller publishers do not have the same resources, especially the non-STM ones and have to rely on specialised technology companies (STC) such as Klopotek or Publishing Technology. Small publishers will usually rely on extensive manual intervention and so have difficulty streamlining their processes and guaranteeing consistent, quality metadata, meeting the technological prerequisites of STCs. Besides, as noted in the same White Paper, these technological entities tend to have privileged relationships with the leading companies within a highly concentrated industry with three/four companies leading in each sector.

For instance McMillan has been one of the first company to partner with Next Big Book, a recently founded (2014) subsidiary of data-centered analysis of music, Next Big Sound³³, to provide a dashboard for the publishing industry that draws sales, publicity, events, social media, web traffic and web trends data together on a daily basis in order to provide a holistic view of a book's trajectory in the marketplace and highlight which factors are most influential, from social signals to book tours.

Nevertheless, an array of companies have using STCs like Klopotek, a global software company that serves more than 350 publishers, with more than 14,000 users globally. Founded in 1992 in Berlin, Klopotek employs around 180 people in Europe and the USA with offices in Berlin, Munich, Amsterdam, London, Paris and New Jersey. The Klopotek software was used to process invoices or royalties totalling approximately 3.5 billion euros. This software supports a total of around 5 million subscriptions and 4 million products around the world. Klopotek's Product Planning and Management system (PPM) enables publishers and media companies to optimize the management of internal and external processes. Publishers of every size – from small publishers to global media corporations – plan, manage and market their complete print and digital product portfolios with PPM. With the support of modern workflows, PPM tools help to develop products and to turn content into business

³³A Manhattan based company founded in 2009, that analyzes all kinds of data for the music industry, based on over 4 years of public social data for hundreds of thousands of artists : YouTube and Spotify plays, social media stats, radio—and compares them against sales to determine which outlets have the most impact.

Specialised companies like App Annie, Distimo (bought by App Annie), and Flurry (bought by Yahoo) emerged over the last years and provide all kind of analysis and metrics to their customers to enable them dealing with big data. Sectoral STCs are emerging as well as seen in the case of the book publishing industry.

Online platform providers process and analyse big data to find meaningful correlations in order to specifically target products and services to individual consumers. Platforms are poised to become an increasingly important point of aggregation in the advertising market with unique reach, customer data and measurability. By combining their traditional expertise with new techniques including data analysis and viral marketing, publishers/ content aggregators can retain their valuable position as curators in the digital space and play a key role in ensuring quality content finds its audience. Big data will be used online and with mobile to power the content recommendation engine.

II.2 Big data for policy making: the case of European poles of excellence EIPE

Visualization techniques and data treatment used to improve the production of scientific evidence to support policy making such as in the case of the mapping of the European IT poles of excellence (formerly done in a previous research exercise by JRC IPTS IS Unit team for EC DG Connect) are summed up in this section.

The 2009 Commission Communication entitled "A Strategy for ICT R&D and Innovation in Europe: Raising the Game"³⁴ proposed reinforcing Europe's industrial and technology leadership in ICT. Building on Europe's assets, in particular its many ICT industrial clusters, the strategy seeks to step up the effort in ICT Research and Development and Innovation (R&D&I). In this context, in collaboration between DG CNECT and the JRC Institute for Prospective Technological Studies, the European ICT Poles of Excellence (EIPE) research project was set up to investigate the issues of growth, jobs and innovation. The EIPE study is developed around several tasks, whose results are presented in a series of IPTS reports. The set of studies attempts to identify ICT R&D&I-related agglomeration economies that would meet world-level excellence, and to identify weak signals that would indicate the dynamics of a changing ICT-related economic geography in Europe. Both of those identification processes are based on quantitative data, built on a set of relevant criteria leading to measurable indicators.

The series of reports offer a snapshot of the performance of regions that are identified as the main locations of ICT activity in Europe. It is meant to provide a comprehensive picture of how ICT activity is distributed across Europe and where its main locations are located (see figure 4). This information is expected to give a better overview of the European ICT landscape. In order to provide a dynamic access to the information gathered within the EIPE project, this report is accompanied by a freely accessible web visualisation tool showing the scores in all regions analysed³⁵.

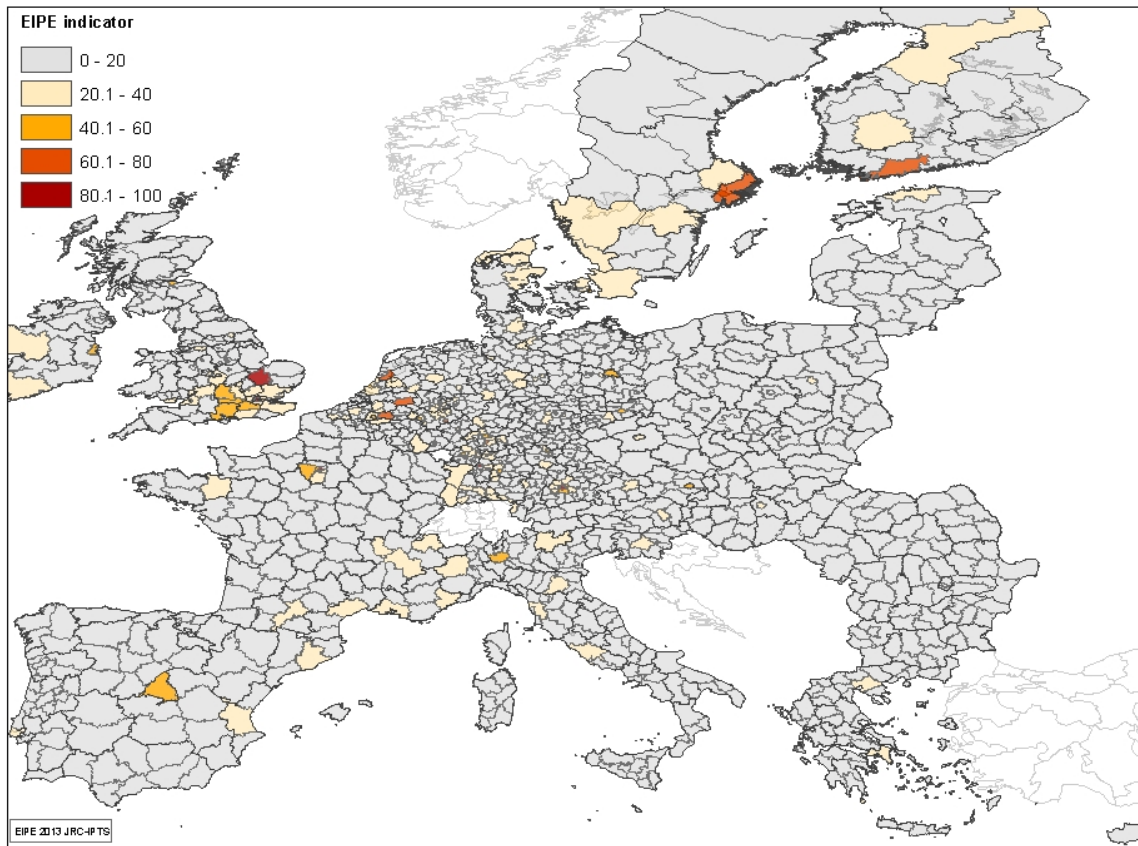
Carefully selecting on the basis of carefully elaborated framework of activities and their characteristics and the discussion on their empirical measurements, a list of indicators has been compiled for the EIPE project. This list includes 42 indicators that formed 3 sub-indicators (ICT R&D, Innovation and Business), which were then aggregated into one EIPE Composite Indicator. The EIPE Composite Indicator is fully based on a set of acknowledged,

³⁴ Available at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0116:FIN:EN:PDF>

³⁵ Available under: <http://is.jrc.ec.europa.eu/pages/ISG/eipe/atlas.html>

available and broad encompassing primary data sources, and a large set of documented state-of-the-art standard methods³⁶. The approach allowed generating the EIPE Identity Card of for each of the 1303 European NUTS 3 regions³⁷.

Figure 4: ICT activity in Europe according to the EIPE composite indicator.



Note: The map represents the geographical distribution of the EIPE Composite values across the 1303 European NUTS3 regions. Further methodological details can be found in De Prato and Nepelski (2013).

To sum up the main results (see figure 1), the mapping, consequent ranking of the EIPes and the in-depth analysis of selected EIPes show:

- A very strong concentration in a few places, and a few countries: in terms of location, actors, activities where money follows, among others, performance,
- Intense cross-border R&D and innovation, business and intensive internationalisation of all types of activities.
- A complex web of connections with different network structures emerging for activities, locations, displaying various roles, and positions.

From a methodological viewpoint, the research also indicated it was feasible to observe the ICT activity in Europe at a very fine-grain level with statistical data as initial input, despite the usual data limitation and constraints. In term of interpretation, the tool says what but no how and why, neither does it point at specific technologies.

³⁶ European coverage/ Global and broad scope: 27 Member States; World MNEs; Global networks.

³⁷ At NUTS 3 level. Nomenclature of Units for Territorial Statistics after the French “Nomenclature d’ Unités Territoriales Statistiques”. The current NUTS classification lists 97 regions at NUTS 1, 271 regions at NUTS 2, and 1303 regions at NUTS 3 level.

ICT Poles of Excellence emerge from the study as important, if not essential parts of ICT activity in Europe. Paradoxically, these world-class locations usually receive national and local acknowledgement and support, but not that much at the European level. The research therefore suggests a range of policies that could be tailored to the specific characteristics of each existing EIPE, while acknowledging and supporting a European ICT Poles of Excellence vision, mainly justified by the efficiency benefits expected from agglomeration and the role of global hubs. The EIPE research is just one example, it will be followed by the IPTS GeoDIT project (see box 5).

Box 5. The GeoDIT project (Geography of Digital Innovation and Technologies): an introduction.

This box is only providing a brief overview of the GeoDIT project building on accumulated data for evidence based support to policy making based on quantitative data exploration and analysis:

- Eurostat data mainly (National accounts, LFS, SES, SRD, SBS, CIS, Surveys),
- EC databases, FP7 etc,
- Other data sources (OECD, private providers, NSOs).

Up to now this initiative involved following public initiatives (EC: DIGIT, CNECT, JRC, ESTAT & ESS, NTTS, IPTS, and scientific networks i.e. Webdatanet COST Action), exploring market supply for tools and out-of-the-shelf solutions (IHS, Scival, Euromonitor, Pentaho, Hadoop..), mapping and testing state-of-the-art technologies and methodologies to improve data analysis capacity (indexing, pattern analysis, text data mining, sentiment extraction, semantics...), looking for potential new data sources (alternative data collection methods based on internet as a data source)

A feasibility checks and a pilot is being run to analyse ICT segments (ex.: IoT). In line with the EIPE research it investigates the dynamic networks of actors, technologies and locations. The project is meant to strengthen the areas identified were it works (or at least helps):

- Tracking emergence: emergent firms / emergent technologies,
- Startups and innovation spotting (with constraints!),
- Technology scouting,
- Value chain, ecosystem, platform description,
- Solving "institutional amnesia",
- Mapping and understanding,
- Increasing granularity,
- From descriptive to predictive.

Source: G. De Prato (2015).

Conclusion

The digital dragons are likely to continue acting as catalysts of the use of big data, although as stressed with some humour by C.Taylor (Wired, 2015) "*The irony of Big Data is that the places where success is quite clear are often the places where the term simply isn't used. In those places, the term may actually be disliked*". This of course stands in sharp contrast with the vagueness of what "*become an obsession with entrepreneurs, scientists, governments and the media*" (Harford, 2014). If the number we reviewed are impressive taking a look simply at the mere amount, the deployment less so. So far it has been characterised by an uneven deployment between companies, sectors and regions.

Nevertheless, the players and uses reviewed, the appearance of new intermediaries do document cases of innovation and growth areas. It should be stressed nevertheless that some of the most obvious examples are coming from improvement in marketing or from the e-

commerce sector. Indeed, from a marketing viewpoint, the information that companies get about the consumption of their products is growing exponentially thanks to technology, allowing a better and highly sophisticated understanding of customers. These examples are easy enough to understand as they represent a highly rationalized / updated demand analysis which clearly involved some innovation in the way demand is dealt with (through recommendations and other inputs from customers).

These achievements may not allow to jump to some more general cross-sectors conclusions. In the field of contents the idea that predictive analytics will yield the ability to produce future hits looks highly overoptimistic and probably a sheer phantasm grounded in a mechanical and technologically deterministic view, although not uncommon. For instance, in the case of Netflix it remains to be seen if the company will be able to duplicate the success story of *House of Cards*; and will not have to go back to the legacy business model of portfolio building in an economy of prototypes. *Marco Polo* does not appear as successful as *House of Cards*, (Miller, 2014). The same holds for the recent critical success of Amazon *Alpha's House*. There are clearly limits to predictive analytics. Besides in that particular case of contents determining customers' desires by analyzing surveys and viewing patterns may not offer the only path to artistic excellence. The effect of all this say corporatization, as with the replacement of independent booksellers by superstores, has been to privilege the blockbusters. Zynga however innovative has been accumulating losses since its IPO.

Some of the authors we quoted recognize the limitations of numbers, Cukier and Mayer-Schönberger for instance warn about falling prey to the “*dictatorship of data.*”, “*against overreliance on data.*” Boyd and Crawford (2012: 7) remind us that numbers don't speak for themselves, stressing that the claims to objectivity and accuracy are misleading (2012: 8), adding that bigger does not always mean better. The use of technology and data can both generate great value and create significant harm, sometimes simultaneously. The phenomenon may bring along new threats together with opportunities in several areas. Fully tapping the potential may hold much promise but is not likely to come as a straightforward route especially for policy-makers. Even if we tried to show (section two of part two) they will benefit also from the integration of new methodologies to have faster, better, smarter exploitation of heterogeneous data for better policy support. The GeoDIT project is an attempt to shift toward the integration of new methodologies to that end.

There as tensions that are inherent to the growth of this new industry and policy makers will be facing difficult choices as can be illustrated by two issues³⁸ (governance, security). For instance, in the case of Netflix, researchers for university of Texas at Austin were able to identify the publicly available and “anonymous” dataset of 500,000 Netflix subscribers by cross-referencing the data with the Internet Movie Database (IMDb) (Oracle White Paper, 2015)³⁹. Big data clearly raises concerns about the protection of privacy and other values as illustrated by the 2014 report for the second Obama administration (Podesta Privacy Report, 2014). A rethink of traditional approaches to data governance may be needed, particularly a shift from focusing away from trying to control the data itself to focusing on the uses of data. Prevalent data standard protection may have become higher as legal standard are now inadequate.

³⁸ Other issues are presented in De Prato and Simon (2015).

³⁹ This research was enabled by Netflix as on October 2, 2006, Netflix, announced the \$1-million Netflix Prize for improving their movie recommendation and to that publicly released a dataset containing 100, 480, 507 movie ratings, created by 480, 189 Netflix subscribers between December 1999 and December 2005 (Narayanan and Shmatikov, 2008:8).

Lastly, the rise of the “Data Barons” is triggering market concentration and data oligopolies issues what Haire and Mayer as the “*dark side of market concentration and data oligopolies*” (Haire and Mayer: 18).

References

Anderson, J., Rainie, L., (2012), *The Future of Big Data*, Pew Research, <http://www.pewinternet.org/2012/07/20/the-future-of-big-data/>

Atelier Paribas, (2013), *Big data, big culture? The Growing Power of the Data and its Outlook for the Economy of Culture*. Available at: http://www.forum-avignon.org/sites/default/files/editeur/EtudeATELIER_FA_2013.pdf

Bain, (2013), *Big Data: The Organization Challenge*. Available at: www.bain.com

Batelle, J. (2006), *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, Portfolio Trade, Penguin Group, New York.

Benghozi P, Salvador E, Simon J.P. (2015). *Models of ICT Innovation. A Focus on the Cinema Sector*. EUR 27234. Luxembourg (Luxembourg): Publications Office of the European Union; 2015. JRC95536.

Boston Consulting Group (2012), *The Value of Our Digital Identity*. Liberty Global Policy Series. Available at: https://www.bcgperspectives.com/content/articles/digital_economy_consumer_insight_value_of_our_digital_identity/

Boyd, D., Crawford, K. (2012). “Critical questions for big data: Provocation for a Cultural, Technological and Scholarly Phenomenon”. *Information, Communication & Society*, 15(5), 662–679. <http://www.google.es/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=7&ved=0CEEQFjAG&url=http%3A%2F%2Fwww.msr-waypoint.com%2Fpubs%2F228268%2FBigData-ICS-Draft.pdf&ei=IIVsVYnlH4ijU67XgbgC&usg=AFQjCNGS7fJJR-dpqVowAarGnBs6176bFA&bvm=bv.94455598,d.d24>

Brynjolfsson, E., Hitt, L.M., Kim, H.H., “Strength in numbers: How does data-driven decisionmaking affect firm performance?.” April 2011, available at SSRN (ssrn.com/abstract=1819486).

Cisco Visual Networking Index (2014), *Global Mobile Data Traffic Forecast Update, 2013–2018*. February 2014. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

Codata, International Council for Science (2014), “Big data needs global approach, says council”. <https://www.researchprofessional.com/0/rr/news/europe/regulation/2014/6/Big-data-needs-global-approach-says-council.html#sthash.qgSrShT8.dpuf>

CSA (Cloud Security Alliance) (2014), *Big Data Taxonomy*, Big Data Working Group. https://www.google.es/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=2&ved=0CCYQFjAB&url=https%3A%2F%2Fdownloads.cloudsecurityalliance.org%2Finitiatives%2Fbdwg%2FBig_Data_Taxonomy.pdf&ei=oPJqVY2lNsS5UYbCg7AE&usg=AFQjCNHMzvLjkecYR6cq7IVOGESpmqtcDg&bvm=bv.94455598,d.d24

Dataflog (2015), “An extensive glossary of big data terminology”. <https://dataflog.com/abc-big-data-glossary/>

Davenport, T.H., Dyché, J., (2013), *Big Data in Big Companies*. International Institute for Analytics (iianalytics.com). Available at: <http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf>

De Prato, G., Simon, J.P., (2015), “The next wave: “Big Data”?”, *Communications & Strategies*, no. 97, 1st Q. 2015, pp. 15-39.

De Prato, G., (2014) “The Book Publishing Industry” in De Prato, G., Sanz, E., and Simon, J.P. (ed), *Digital Media Worlds. The new media economy*, Palgrave Mc Milan, Houndmills Basingstock, pp.87-101.

De Prato, G., & Nepelski, D. (2014). *Mapping the European ICT Poles of Excellence. The Atlas of ICT Activity in Europe*. Seville: EC JRC-IPTS.

De Prato, G., & Nepelski, D. (2013). *Identifying European ICT Poles of Excellence. The Methodology*. Seville: JRC-IPTS.

De Prato, G., Simon, J.P., (2012), *The Book Publishing Industry* . <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=5702>

Distinguin, S., (2013), « [Amazon, l’empire caché](http://www.slideshare.net/faberNovel/amazoncom-the-hidden-empire), Fabernovel: <http://www.slideshare.net/faberNovel/amazoncom-the-hidden-empire>

European Commission ((EC) (2014a), *Towards a thriving data-driven economy*. <http://ec.europa.eu/digital-agenda/en/towards-thriving-data-driven-economy>

e-skills UK/ SAS, (2013).Big Data Analytics: An assessment of demand for labour and skills, 2012-2017.

EMC Digital Universe study, (2014), <http://www.emc.com/leadership/digital-universe/2014iview/index.htm>

Ericson Mobility Report (2014). *On the pulse of the networked society*. www.ericsson.com/ericsson-mobility-report

Ernst&Young (EY) (2014), *Etude Big Data 2014*. <http://www.ey.com/FR/fr/Services/Advisory/EY-etude-Big-data-2014-strategie-big-data>

Fabernovel, (2014), *GAFAnomics. New Economy, New Rules*, <http://fr.slideshare.net/faberNovel/gafanomics>

Filloux, F., (2014), “Google might not be a monopoly , after all”, http://www.mondaynote.com/2014/06/30/google-might-not-be-a-monopoly-after-all/?page_id=6396&print=pdf

(Frater, 2015),” China Rising: How Four Giants Are Revolutionizing the Film Industry”. *Variety*. <http://variety.com/2015/film/news/china-rising-quartet-of-middle-kingdom-conglomerates-revolutionizing-chinese-film-industry-1201421685/>

Gartner (2014), “Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years”.
<http://www.gartner.com/newsroom/id/2848718>

Haire, A., J., Mayer-Schönberger, V., (2014), *Big Data - Opportunity or Threat*, ITU GSR discussion paper, 2014.

Hamel, M.P, Marguerit, D. (2013), « Analyse des big data. Quels usages, quels défis », in Gilles, L., Marchandise, J.F. (ed) (2013), *La dynamique d’Internet. Prospective 2013*, Commissariat Général au Plan, Paris. Available at :
http://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/etude_internet_2030-web.pdf

IDC, (2013), *Business opportunities: Big Data*. https://ec.europa.eu/growth/tools-databases/dem/sites/default/files/page-files/big_data_v1.1.pdf

IDC, (2012), *Worldwide Big Data Technology and Services, 2012–2015 Forecast*. ITU, (2013).

International Telecommunication Union (ITU) (2014a). <http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx> (

[International Telecommunication Union \(ITU\), \(2014b\).](http://www.itu.int/en/ITU-T/Workshops-and-Seminars/bigdata/Pages/default.aspx)

<http://www.itu.int/en/ITU-T/Workshops-and-Seminars/bigdata/Pages/default.aspx>

Kelly, J. (2014), “Big Data Vendor Revenue and Market Forecast 2013-2017”,
http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Mayer-Schönberger, V., Cukier, K., (2013), *A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt

McIllroy, T. (2014) (2014), “11 Topmost Digital Book Publishing Trends & Opportunities”. A report for Digital Book World <http://www.digitalbookworld.com/2014/eleven-digital-publishing-trends-for-2015/>

McKinsey (2013), “A gallery of disruptive technologies”.
<http://www.mckinsey.com/tools/Wrappers/Wrapper.aspx?sid={21F95813-D665-4176-80BD-3823144E3FE2}&pid={A1D4B928-3A7B-4073-AFFA-6AD78525CDB1}>

McKinsey (2011), *Big Data: The next frontier for innovation, competition and productivity*.
http://www.mckinsey.com/~media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

Miller, S.L., (2014), “‘Marco Polo’ Disappoints History, ‘Game of Thrones’ Comparisons & Netflix’s Great Reputation”. <http://www.indiewire.com/article/review-marco-polo-disappoints-history-game-of-thrones-comparisons-netflixs-great-reputation-20141211>

Narayanan, A., Shmatikov, V., (2008), “Robust De-anonymization of Large Sparse Datasets”.
https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Nepelski, D., & De Prato, G. (2013). *Defining European ICT Poles of Excellence. A Literature Review*. Seville: JRC-IPTS.

Nepelski, D., & De Prato, G. (2014). *Analysing the European ICT Poles of Excellence. Case studies of Inner London East, Paris, Kreisfreie Stadt Darmstadt, Dublin and Byen Kobenhavn*. Seville: JRC-IPTS.

Networking and Information Technology Research and Development (NITRD) (2014), Report on Privacy Research within NITRD, April 2014, http://www.nitrd.gov/Pubs/Report_on_Privacy_Research_within_NITRD.pdf.

Oracle(2015), “Securing the Big Data Life Cycle”, *White Paper*. <http://files.technologyreview.com/whitepapers/Oracle-Securing-the-Big-Data-Life-Cycle.pdf>

SAP (2014), *Beyond Connectivity. A Guidebook for Monetizing M2M in a Changing World*, SAS, (2013), *2013 Big Data Survey Research Brief*. Available at: http://www.sas.com/resources/whitepaper/wp_58466.pdf

SAS (2015), “Bringing the power of SAS to Hadoop”, *White Paper*. http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/bringing-power-of-sas-to-hadoop-105776.pdf

Simon, J.P., (2015), “Le Big Data: un enjeu pour les industries créatives ». <http://www.inaglobal.fr/numerique/article/le-big-data-un-enjeu-pour-les-industries-creatives-8065>

Tata Consultancy Services Limited (2013), *The Emerging Big Returns on Big Data* .http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf

Taylor C.(Wired, 2015) reference?

United States Executive Office, (2014a), *Big Data: Seizing Opportunities, Preserving Values*. Big Data Privacy Report, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

United States Executive Office, (2014b), *Big Data and Privacy: A Technology Perspective*. http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

United Nations (2014), UNPulse. <http://www.unglobalpulse.org/about-new>

van Rijmenam, M., (2013), “Zynga: A Big Data Company Masquerading as a Gaming Company”. <http://smartdatacollective.com/bigdatastartups/116851/zynga-big-data-gaming-company>

Woodie, A. (2013), “The Big Data Market By the Numbers” . http://www.datanami.com/2013/10/03/the_big_data_market_by_the_numbers/

World Economic Forum (WEF) (2013); *Unlocking the Value of Personal Data: From Collection to Usage*; Prepared in collaboration with The Boston Consulting Group. http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf

Companies website:

Amazon, (2014, 2015) Annual Report 2013, http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-sec&control_selectgroup=Annual%20Filings

Google Search Statistics (2015). <http://www.internetlivestats.com/google-search-statistics/>

Hadoop: <http://hadoop.apache.org/#sthash.fPho82IG.dpuf>

Klopotek: <http://www.klopotek.com/en/homepage.html>

Next Big Book: <https://www.nextbigbook.com/>

Next Big Sound: <https://www.nextbigsound.com/about>

Netflix: <http://ir.netflix.com/long-term-view.cfm>

Reed Elsevier (2014), Annual Report 2013. <http://www.reedelsevier.com/mediacentre/pressreleases/2014/Pages/publication-of-annual-reports-and-financial-statements-2013.aspx>

Storm8: <http://www.storm8.com/about-us>

Publishing Technology: <http://www.publishingtechnology.com/about-us/>

Zynga (2011), IPO Prospectus. <http://investor.zynga.com/secfiling.cfm?filingID=1193125-11-341923&CIK=1439404>