DE GRUYTER
OPEN

REM - Research on Education and Media
Vol. 8, N. 2, Year 2016
ISSN: 2037-0830 – DOI: 10.1515/rem-2016-0020

Think big.

Learning contexts, algorithms and data science

Michele Baldassarre

University of Bari, Italy, michele.baldassarre@uniba.it, ORCID 0000-0002-5209-0613

Abstract

Due to the increasing data growth available in recent years, all areas of research and the management of institutions and organizations, and specifically in schools and universities feel the need to give meaning to this availability of data. This article after a brief reference to the definition of big data, intends to focus attention and reflection on their type in order to proceed to an extension of their characterization. One of the hubs to make feasible the use of Big Data in operational contexts is to give a theoretical basis to which to refer. The DIKW model (Data, Information, Knowledge, Wisdom) correlates these four aspects concluding in Data Science which in many ways could revolutionize the established pattern of scientific investigation The Learning Analytics applications on online learning platforms can be tools for evaluating the quality of teaching. And that's where some problems arise. It becomes necessary to handle with care the available data. Finally, a criterion for deciding whether it makes sense to think of an analysis based on Big Data can be to think about the interpretability and relevance in relation to both institutional and personal processes.

Keywords: *learning context; education; data science.*

Not everything that counts can be counted, and not everything that can be counted counts
Cameron, W. B. (1963).

I put out my hand and what do I feel? I know there's a thousand things you analyze every ten minutes. Patterns, ratios, indexes, whole maps of information. I love information. This is our sweetness and light. It's a fuckall wonder. And we have meaning in the world.
DeLillo, D. (2003)

Introduction

In many sectors, including education, the amount of data has greatly increased during the last 15 years. It is necessary to make sense of this data to improve decision-making abilities of school and university.

"We have entered an era of Big Data. Many sectors of our economy are now moving to a data-driven decision making model where the core business relies on analysis of large and diverse volumes of data that are continually being produced. This data-driven world has the potential to improve the efficiencies of enterprises and improve the quality of our lives. However, there are a number of challenges that must be addressed to allow us to exploit the full potential of Big Data" (Jagadish et al., 2014, p. 94)

This paper is a contribution to the reflection about the potential and the identification of the most relevant and interesting elements of big data for learning contexts.

1. Big data: a definition

The development and diffusion of digital technologies and exponential growth in tablets, smartphones, computers and the Internet of Things generate huge quantities of data. This data constitutes a challenge for the future, and a source of value for citizens and as-yet under-exploited potential growth for companies. (The Industry of the Future 2015 p. 41) A challenge is: how can we create added value from the unprecedented volumes of data now available?

"High volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." (Gartner, 2012). This is one of the most diffused definition of big data. But *Big data* is a trending buzzword in an emerging research area where common terminology is still evolving. The term is used to describe a wide range of concepts: from the technological ability to store, aggregate, and process data, to the cultural shift that is invading business and society, both drowning in information overload. The figure 1 shows a static tag cloud visualization of the most popular words (obtained by counting the appearance frequency of words included in the abstracts) in 1,437 abstracts of conference papers and articles, which contained the full term "Big Data" in either the title or within the author-provided keywords. (De Mauro et al., 2015, p. 98)

Fig. 1. Static tag cloud visualization of Big Data-related papers (source: De Mauro et al., 2015, p.98)

If we analyze the evolution of the definition of the term, we can observe that the primary idea temporally associated with *big data* is that the volume of information had grown so large that the quantity being examined no longer fit into the memory that computers use for processing, so engineers needed new tools to analyzing it. This is the origin of new processing technologies like Google's MapReduce and Yahoo's Hadoop. These let one manage far larger quantities of data than before, and the data -this is remarkable- need not be placed in tidy rows or classic database tables. One way to think today about *big data* is that the words refers to things one can do at a large scale and that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments. (Mayer-Schönberger & Cukier, 2013, pp. 6-7)

2. Types of big data

In data science and big data we can use many different types of data, and for each of them are often required different tools and techniques. The main categories of data are these:

2.1. Structured (Data in rigid formats).

The term structured data in the big data environment usually refers as the data has arranged in a definite form and resides in a fixed field within a record. It's often easy to store structured data in tables (rows and columns) within databases or Excel files. The important characteristics of this data are all the entities of the same group have same

attributes, format, length and follow the same order; everything is labelled and easy to access, and can be easily organized and processed by data mining tools. In the literature many of the authors expressed that generally this kind of data accounts about 20 percent out of whole data in all key domains. (Rao, et al., 2015, p. 25).
Example: Databases, Excel sheets.

2.2. Unstructured (No particular pattern/format)

The term *unstructured data* refers to information that doesn't fit in a traditional row-column database the data. It's the exact opposite of *structured data*. Data that is stored in a format that is easy readable for human but is very difficult or almost impossible for a computer to understand because of irregularities and ambiguities that make it difficult; Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. Unstructured data can be also, images, videos, audios.
One typical example of unstructured data is a regular email. Although email contains structured elements such as the sender, title, and body text, it's difficult for computers to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example (Cielen et al., 2016., p. 5).
Examples: Word documents, email messages, audio-video files.

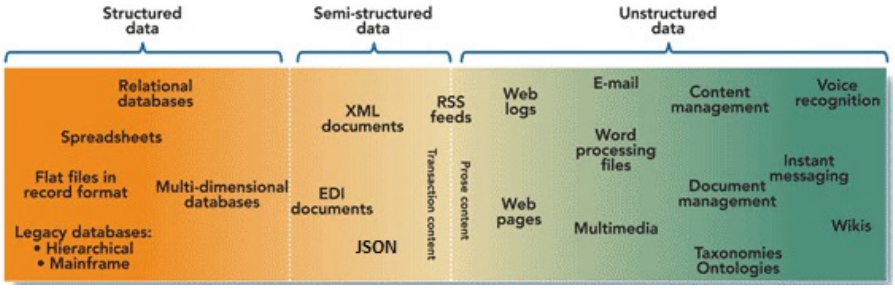

Fig. 2. Structured, semi structured and unstructured data (adapted from Klein J. 2014)

2.3 Semi-structured (Unstructured data with a format).

Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables. In some applications, data is collected in an ad-hoc manner before it is known how it will be stored and managed. This data may have a certain structure, but not all the information collected will have identical structure. Semi-structured data is a cross between unstructured and structured. This is data that may have some structure that can be used for analysis but lacks the strict data model structure. In semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, a Facebook post can be categorized by author, data, length and even sentiment but the content is generally unstructured. Another example is word processing software that includes metadata detailing the author's name, when it was created and amended but the content of the document is still unstructured (Marr, 2015, pp. 61-62).
Examples: Twitter- feeds, tags in videos, Web Pages, XML.

2.4 The UNECE taxonomy

The United Nations Economic Commission for Europe (UNECE) in 2013 proposed a taxonomy to classify Big Data based on the source of data; it is shown analytically in figure 3 and can be summarized in three points (UNECE, 2013):

- *Social Networks* (human-sourced information): this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video.
- *Traditional Business systems* (process-mediated data): these process record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc.

- *Internet of Things* (machine-generated data): derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world

1. Social Networks (human-sourced information)
Social Networks
Blogs and comments
Personal documents
Pictures: Instagram, Flickr, Picasa
Videos: Youtube etc.
Internet searches
Mobile data content: text messages
User-generated maps
E-Mail
2. Traditional Business systems (process-mediated data)
Data produced by Public Agencies
Medical records
Data produced by businesses
Commercial transactions
Banking/stock records
E-commerce
Credit cards

3. Internet of Things (machine-generated data)
Data from sensors
Fixed sensors
Home automation
Weather/pollution sensors
Traffic sensors/webcam
Scientific sensors
Security videos/images
Mobile sensors (tracking)
Mobile phone location
Cars
Satellite images
Data from computer systems
Logs
Web logs

Fig. 3. Classification of Types of Big Data developed by UNECE (source: De Francisci, 2015, p. 16)

3. From Four V's to seven V's

Big Data are often explained and characterized using the "Four V's": Volume, Velocity, Variety and Veracity. (v. fig. 4).

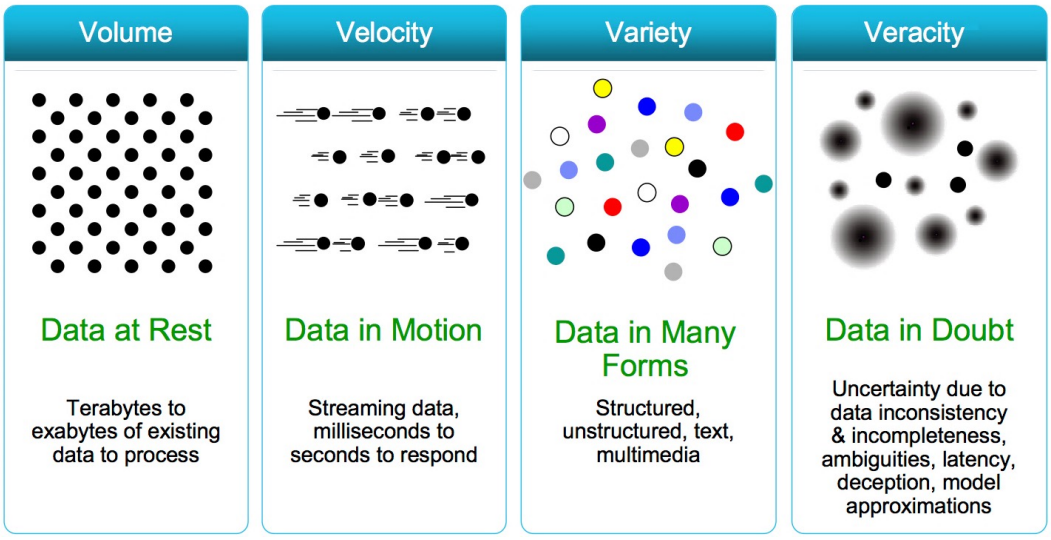

Fig. 4. The Four V's of Big Data (source: Minor, 2013)

- **Volume**: *How many data?* The size is a particularly important feature for the data generated by automated systems; for example the sensors installed on a single engine of an airliner generate about 20 terabytes of data each hour, which multiplied by the number of in flight airliners every day (about 25000) leads to about 500 petabytes of data every day (MIUR, 2016, p.16).
In 2011 we created 1.8 Zettabyte data. By 2020, the world will generate 50 times that amount of data, according to a study done in 2011 by IDC (Gantz & Reinsel, 2011). One of the most remarkable and significant contributions in terms of quantity will come from IoT (Internet of Things), with data created every second by the sensors of all devices connected worldwide.

- **Velocity:** How fast new data are generated? In processes where time is an important factor, such as in identifying online fraud, or in rescue operations in situations emergency, it is necessary to analyze big data flowing almost in real time, to maximize its value. The volume of data from social networks is reduced compared to the machine-generated data, however, the great speed with which they can be produced in any case generates large volumes of data (it is estimated that in 2011 Twitter generated about 8 terabytes of data every day; this value in 2015 was around 12 terabytes).

- **Variety**: *How many different types of data*? In traditional dataset information it is usually generated in the same way by a small number of sources and according following an established pattern (Lemberger, et al, 2015, p.16). In the past, all the data were structured data that fit perfectly in rows and columns, but today is no longer so. With big data, the variety of sources are heterogeneous and evolving and we find any kind of data - structured and unstructured data such as, for example, text data, sensor data, audio data, video data, click streams, log files, and more. Each type of data requires a different type of analysis and different tools to interpret (Van Rijmenam, 2014, p.6).

- **Veracity:** *How the data is accurate?* The generation and collection of a large amount of high-speed data are useless if the data is invalid or biased. Incorrect or biased data cause problems for organizations. This is even more true in automated decision-making processes, in which no human being is involved, and the algorithm is king. We should be sure that the data are accurate, by implementing procedures to prevent invalid data accumulate in the collection systems. This challenge is sometimes called "*garbage in, garbage out*[1]".

During last years 4 V's have been updated with Variability, Visualization, and Value. In total count we have 7 V's to characterize *big data*:

- Volume,
- Velocity,
- Variety,
- Veracity,
- Variability,
- Visualization,
- Value.

Let's explore the last three:

- **Variability:** *How many meanings can data have?* The Big Data are extremely variable. Words have different meanings depending on the context. Variability should not be confused with the variety. If a bakery sells ten different types of bread, we talk about variety. If, instead, the same type of bread has a taste and a smell different every day, that is variability.
Variability is a very important aspect in performing sentiment analysis. In similar tweet, a word can have a completely different meaning. To perform a correct sentiment analysis, algorithms need to decipher the exact meaning of a word in its context.

- **Visualization:** *How to represent the big date*? Visualization is essential, but perhaps the most difficult thing to achieve for big data. Using charts and graphs you can display large amounts of complex data, in a much more effective way if compared to spreadsheets full of numbers and formulas. The large amount of data must be made understandable and easy to read and interpret.

- **Value:** *How is the value of big data?* Of course, the data itself is not at all precious. The value is in the analyses performed on these data and how data is transformed into information and then into knowledge and wisdom (see § 4). The value is in the way organizations use data to make decisions on the basis of the information derived from their analysis. McKinsey says that the potential annual value of big data, only for the US health care industry, is 300 billion dollars (Manyika et al. 2011).In the same report it states that the big data have a potential annual value of 250 billion euro for the European public administration. The challenge is therefore to identify and extract what they can create value.

4. From data to wisdom

An attempt to give a theoretical basis to the analysis of big data is the DIKW model (*Data, Information, Knowledge, Wisdom*), introduced by Ackoff. This model brings together data, information, knowledge and, in a more

[1] *Garbage in, garbage out*: the terms refer to the fact that computers, since they operate by logical processes, will unquestioningly process unintended, even nonsensical, input data ("garbage in") and produce undesired, often nonsensical, output ("garbage out")

nuanced way, wisdom. In everyday language, usually data and information are synonymous, while according to the DIKW model (Ackoff 1989):

- **Data** are the result of a more or less accurate observation, and may or may not be inspired by a problem to be solved. The data are objective facts, signs, numbers, and they do not need relationships with other elements to exist, but if you take them individually they don't communicate anything and do not contain any meaning.

Data is something perceived by the senses (or sensors) but that has no intrinsic value until you put in a context. Data becomes information only when they are placed in context, through: contextualization (in fact), categorization, processing, correction, synthesis.

- **Information**, deduced from the data, includes them, giving them a meaning and gaining added value compared to the data. Information is the choice of an individual to put some data in a context, fixing some as premises, and making a series of inferences, drawing conclusions. These conclusions are called *information* but don't become *knowledge* if they are not related to the knowledge and experience of a specific person.

- **Knowledge** is the combination of data and information, to which is added the opinion of espert persons, competence and experience, to build a valuable asset that can be used to aid decision-making. The knowledge can't be lost in the same way in which one can lose data and information. We are in the domain of competence, and the more you move from data to knowledge, the greater is the dependence on the context (Blair, 2002). Davenport & Prusak offer their definition: "Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norm. (Davenport & Prusak,1998, p.5).

Knowledge is always individual and cannot be transmitted because it is generated from the individual's previous experience and knowledge; what we can transmit is only the narration of the experience.

-**Wisdom** is immaterial, intangible. Wisdom is the judgment, the ability to add value and is unique and personal. Wisdom is something that goes beyond the concepts of information and knowledge and embraces both assimilating and transforming into individual experience. Wisdom accompanies the knowledge and allows us to make the best choices.

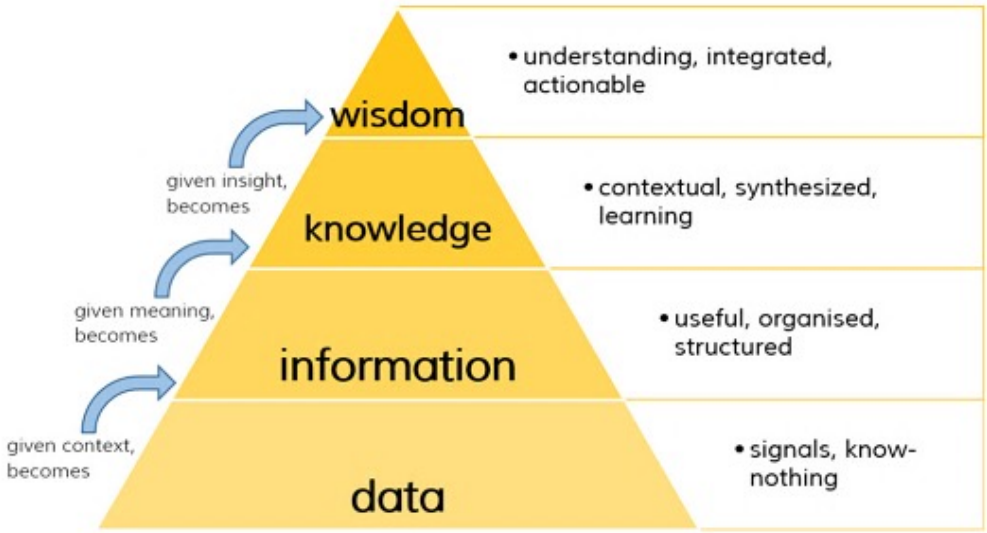

Fig.5. The DIKW Pyramid. source: Soloviev, K. (2016).

The DIKW model is an attempt to categorize and simplify the key concepts involved in cognitive processes, especially when we have to manage large amounts of data. This theoretical model provides a hierarchy, called DIKW (Data-Information-Knowledge-Wisdom), consisting of a very large base of raw data which, going towards the top of the pyramid, are subject to an aggregation-contextualization process (information) and application-testing (knowledge). On top of the pyramid, as showed in figure 5, is confined wisdom that assumes a level of knowledge that is beyond the scope of a specific application. These cognitive states were then connected in a hierarchical way assuming that between them there can be a smooth transition from the bottom to the top.

Besides that as pyramid, there is an effective representation of the DIKW model on a Cartesian plane (figure 6); Ackoff had originally indicated only one axis, the understanding one, but several authors (Rowley, 2007, Fricke 2009) later showed that it is also important to the size of the *context* or connection. The figure 6 highlights the rising value, from data to wisdom.

In figure 6 you can see that the first three categories refer to the past; they deal with what *has been* or what *was* known. Only the fourth category, *wisdom*, deals with the future because it incorporates vision and design. With wisdom, we can imagine the future and not just grasp the present and the past. But achieving wisdom is not easy; onet must go through all the other categories.

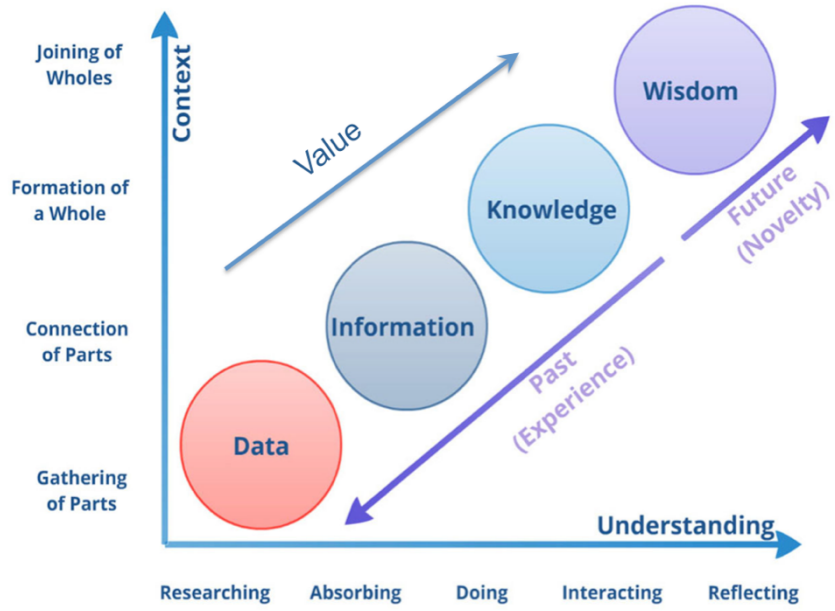

Fig. 6 DIKW model on a cartesian plane. Source: adapted from Omid, 2014

5. Data science

Big data is a blanket term for any collection of data sets so large or complex that is difficult to analyse using traditional data management techniques such as, for example, the RDBMS (relational database management systems). The widely adopted RDBMS has long been regarded as a one-size-fits-all solution, but the demands of handling big data have shown otherwise. *Data science* involves using methods to analyse massive amounts of data and extract the knowledge it contains. You can think of the relationship between big data and data science as being like the relationship between crude oil and an oil refinery. Data science and big data evolved from statistics and traditional data management but are now considered to be distinct disciplines (Cielen et al., 2016, p.1).
Every day we create several brontobyte[2] of data, so that 90% of the data in the world were created in the last two years. We have to imagine that in just a minute, 150 million emails, 347.222 tweets, 527.760 Snapchat images, 21 million Whatsapp messages are sent according to the consulting and technology society Excelacom (Leboeuf K. (2016).
This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data (Zikopoulos et al 2013).

How can we observe in real time the spread of an epidemic? How can we prevent crime and improve the security of the city? Can we learn about the emotions and moods of an entire nation? May our passions dangerously threaten our privacy? Big data, and moreover data science are the answer to all these questions: offering the possibility to act on all the information and not just on statistical samples, we can develop faster, economic and extraordinarily more precise responses about the world around us.

"What makes our era different is that many of the inherent limitations on the collection of data no longer exist. Technology has reached a point where vast amounts of information often can be captured and recorded cheaply. Data can frequently be collected passively, without much effort or even awareness on the part of those being recorded. And because the cost of storage has fallen so much, it is easier to justify keeping data than discarding it. All this makes much more data available at lower cost than ever before. Over the past half-century, the cost of digital storage has been roughly cut in half every two years, while storage density has increate 50 million-fold" (Mayer-Schönberger &. Cukier, 2013, pp. 100-101)

[2] 1 brontobyte is 1 000 000 000 000 000 000 000 000 000 bytes, that is 10^{27} byte

The widespread availability of digital technologies and the Internet is producing a real paradigm shift in the statistics, in public communication, in the functioning of society, and in social research (Ayres, 2008). In no previous age it has never been so easy and so cheap to produce quantitative information, opinion surveys and statistical data; we are seeing the development of big data, i.e. data generated from transaction systems, interaction, monitoring and localization.

The availability of powerful computer and algorithms capable of analysing this huge amount of data - the order is of Exabyte (one billion billion bytes) is going to produce changes in scientific paradigms.

In 1976 the famous British statistical George Box wrote:

«*All models are wrong, but some are useful*» (Box, 1976).

During the O'Reilly Emerging Technology Conference of 2011, Peter Norvig, Google's research director, gave an update to George Box's quote:

«*All models are wrong, and increasingly you can succeed without them*».

The scientific method is built around testable hypotheses. These models, generally, are systems visualized in the minds of scientists and researchers. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years. Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

Chris Anderson, editor-in-chief of Wired magazine) wrote an article titled *"The end of theory. Will the Data Deluge Makes the Scientific Method Obsolete?*" (Anderson, 2008) where we can read:

«This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves».

In the article Anderson says that recent and important progress, not only in the strictly scientific field, has been made completely ignoring the meaning and structure of what was being investigated; in his opinion, the development of knowledge is going to orientate in this direction, obtaining the general theories of the phenomena from the analysis of huge amounts of data available electronically, that computers are able to analyse at speeds impossible for humans.

Mayer-Schönberger & Cukier take up the idea:

«Not only is the world awash with more information than ever before, but that information is growing faster (…) The era of big data challenges the way we live and interact with the world. Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing *why* but only *what*. This overturns centuries of established practices and challenges our most basic understanding of how-to make decisions and comprehend reality» (Mayer-Schönberger & Cukier, 2013, pp. 6-7).

We are moving towards a paradigm shift that involves all subject: a *data driven* scientific approach; there is new synthesis between the top-down modelling of phenomena and the discovery of bottom-up knowledge, emerging from large masses of data available; experiments with data are data not validation of theories and models, but they are considered the discovery of emerging pattern that suggest scientists new theories and new models able to explain deeply the complexity of social, biological, technological, cultural economic phenomena.

It seems no longer necessary, according to the traditional model of scientific inquiry, draw up ex ante assumptions about the functioning of a certain phenomenon, and then make the necessary checks in order to ensure their reliability. In fact it is possible to process enormous amounts of data looking for correlations independently from knowledge of their content, and bring out meaningful relationships by extremely complex systems without the result is influenced by the type of research question or even in absence of any research question (figure 7).

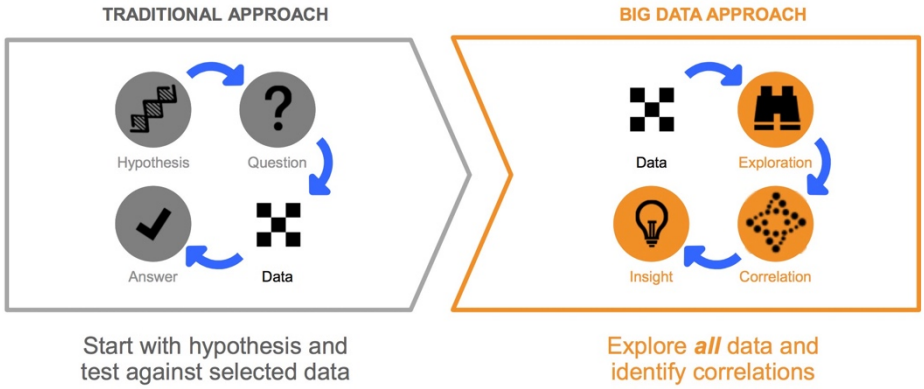

Fig. 7 Traditional and big data approach in research. Source: Kristensen, 2014

The Data Science Association defined in October 2013 the terms 'data science' and 'data scientists' as follows (Data Science Association, 2013):

- Data science is the scientific study of the creation, validation and transformation of data to create meaning.
- A data scientist is a professional who uses scientific methods to liberate and create meaning from raw data.

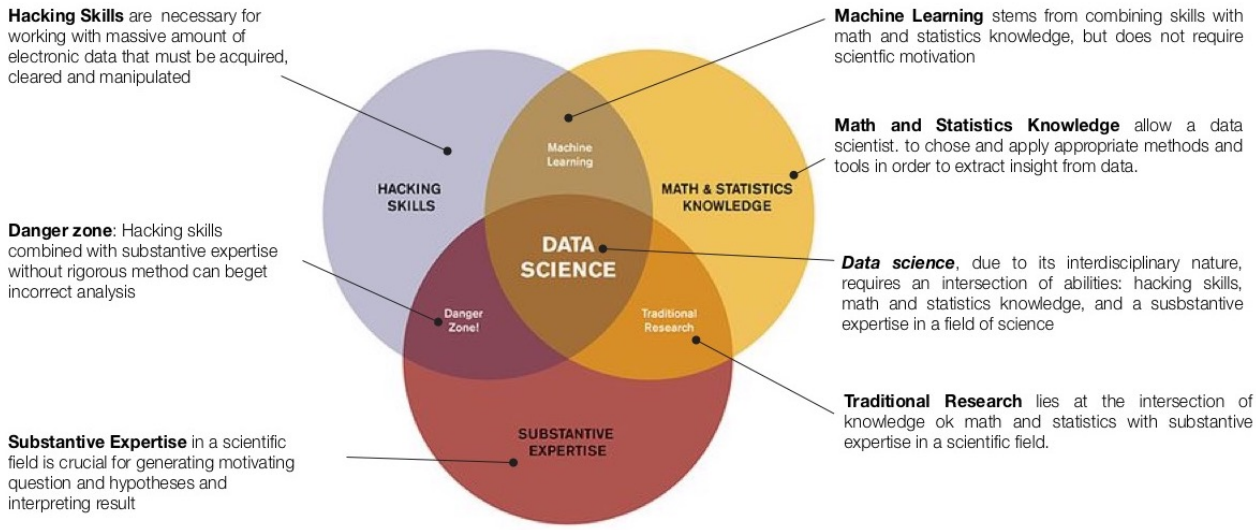

Fig. 8. The data science venn diagram (graphic by author, based on Conway, 2010).

What distinguishes data science from statistics? Statistics traditionally is concerned with analysing primary or experimental data that have been collected to check specific hypotheses. We use statistics for primary data analysis or top-down (confirmatory) analysis, *hypothesis evaluation* or *testing*. On the other hand Data science is typically concerned with analysing secondary or observational data that have been collected for other reasons. We use data mining and data science for secondary data analysis or bottom-up (exploratory) analysis, and for *hypothesis generation* and *knowledge discovery*.

As you can see in figure 9, a *data analyst* focuses on the interpretation of data, typically with a focus on the past and present, while a *data scientist* focuses on summarizing data to provide forecasting based on the patterns identified from past and current data.

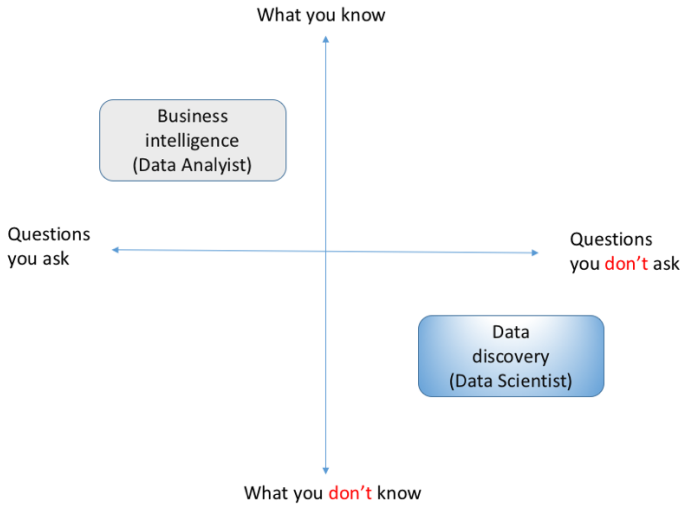

Fig. 9 Data Scientist vs Data Analyst

Table 1 summarizes the main phases of the data science process; it includes the actions to take during a project. Often there is not a linear way from first to seventh phase, but one has to regress and iterate between the different steps.

Table 1. Main phases of the data science process.

#	Phase
1	Research goal
2	Data retrieval
3	Data preparation
4	Data exploration
5	Data modeling
6	Presenting results
7	Automating the analysis

- The first phase is setting a research goal. We have to understand the what, how, and why of the research project, and we have to have clear vision (and write down) the project mission and context, the resources we expect to use, the way we are going to perform the analysis, the list of deliverables a timeline.
- The second step of the process is data retrieval. We need data available for analysis, so we have to find suitable data and getting access to the data from the data owner. We'll found data in many forms, from tables in a database to text files. The result of this step is data in raw form. It's probable we need to polish and transform before use.
- Once we have the raw data, we have to prepare it, that is transforming the into a form into directly usable in our models. To compete this step, we must find and eventual errors that are the data, merge data that comes from different sources, and transform it.
- The fourth phase is data exploration and visualization, to reach understanding of the data. We have to use visual and descriptive techniques to look for correlations, patterns, to gain the insights.
- The fifth step is data modelling; we try to build a model. We have to verify if the insights gained can make the predictions written in the first step.
- The sixth phase of the data science process is presenting the results
- The last step, (to act if needed, to save time when the process has to be replicated over and over again), is the automation of the analysis.

The figure 9 shows the steps of the data science process; it includes the actions to take during a project (Cielen et al., 2016., p. 23). In the figure the steps we indicated as 6th and 7th in table 1 have been associated in a single phase.

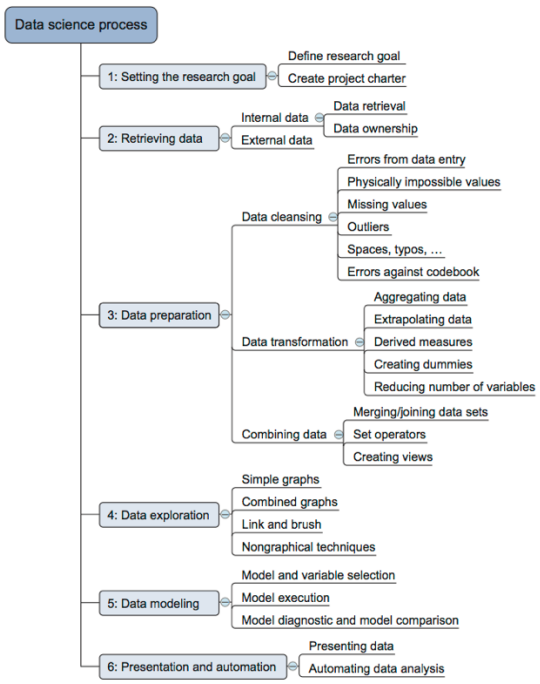

Fig. 9 The six steps of the data science process. Source: Cielen et al.,2016., p. 23

A new interdisciplinary field of inquiry is emerging during the last years: the *Educational data science* (EDS)[3]. It brings together computer science, education, statistics and other social sciences to examine and understand social and technical phenomena. EDS researchers and practitioners utilize various sets of procedures and techniques to gather, organize, manipulate and interpret rich educational data sources. EDS also presents techniques for merging voluminous and diverse data sources together. ensuring consistency of these data sets. and creating unified visualizations to aid in understanding of complex data. Further, in this field, educational data scientists build mathematical models and use them to communicate insights/findings to other educational specialists and scientists in their team and if required to nonexpert stakeholders (Cordoba, 2016, p. vii)

6. Learning and evaluating with big data

Big data gives us unprecedented insight into what works and what doesn't. It is a way to improve student performance by showing aspects of learning that were previously impossible to observe. Lessons can be personally tailored to students' needs, boosting their comprehension and grades. It helps teachers identify what is most effective: it doesn't take away their jobs but makes their work more productive, and probably more fun too. It helps school administrators and policymakers provide more educational opportunities at lower cost, important factors for reducing income gaps and social disparities in society. For the first time, we have a robust empirical tool with which to understand both how to teach, and how to learn. (Mayer-Schönberger & Cukier 2014, p.4).
With big data we can develop models that, according to Daniel can be *descriptive, predictive and prescriptive* (figure 10).
Descriptive models are grounded in the analysis of transactional and interactional data about teaching or learning. They can he used to identify trends such as student enrolment, graduator rates and patterns likely to trigger important dialogue on improving student learning. The presentation of descriptive models alone is inadequate. Institutions need to be able to examine their present performances and be able to predict future outcomes.
Predictive models provide institutions with the ability to uncover hidden relationships in data and predict future outcomes with a certain degree of accuracy. For instance, they enable institutions to identify students who are exhibiting risky behaviours during their academic programme.
Prescriptive models are actionable tools built based on insights gained from both descriptive and predictive models. They are intended to help institutions to accurately assess their current situation and make informed choices on alternative course of events based on valid and consistent predictions (Daniel 2016, p.3).

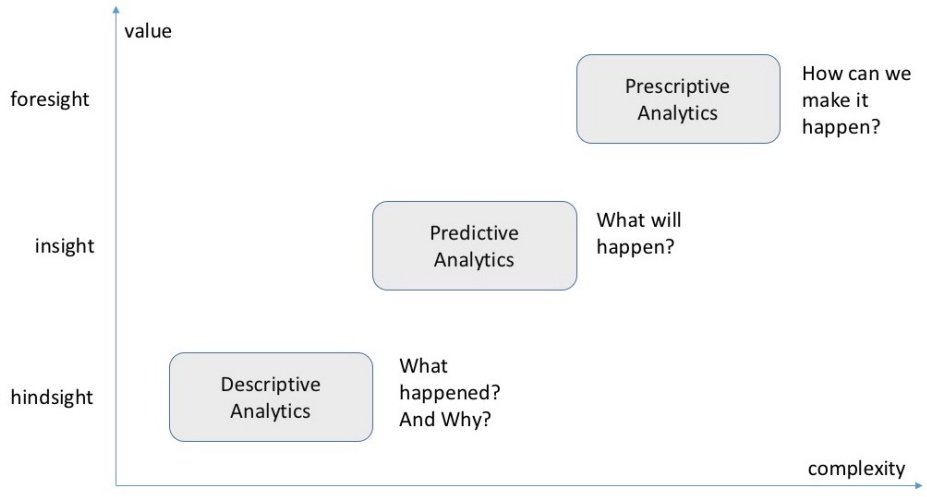

Fig. 10. Descriptive, Predictive, Prescriptive Analytics (source: modified from Alahuhta 2014, p.7, and based on Daniel 2016)

[3] As a subdiscipline of data science, EDS originated from discussions held during several workshops between years 2000 and 2007, mainly from the Educational Data Mining (EDM) Conference in 2008. EDM itself as a field of research is concerned with developing methods for exploring increasingly large-scale educational data, to better understand students and the settings in which they learn. In the last years other two international conferences were held. focusing on EDS themes. Learning and Knowledge Analytics (LAK2011) was the first conference, followed by Learning at Scale (L@S) in 2014. Lately, conferences in this area focus discourse on exploring the impact of big data and learning analytics in fostering learning and teaching and in engaging the growing community of researchers, practitioners. and learners within higher education to build tools, procedures and techniques to explore and solve complex learning problems. (Cordoba, 2016)

Big data bring to learning contexts three main changes (cf. Mayer-Schönberger & Cukier 2014, p.4).

- You can collect feedback data that was impractical or impossible to amass before.
- You can individualize learning, tailoring it not to a cohort of similar students, but to the individual student's needs.
- You can use probabilistic predictions to optimize what they learn, when they learn, and how they learn.

The real novelty offered by big data in education is concerned with the analysis of data relating to interactions, to the path of interfacing of learners with online platforms (Learning Management Systems, or social learning platforms).

The data are immediately available to analysis to provide an overall picture of the behaviour of the class and optimize the learning strategy in almost real time. The *Learning Analytics* applications on online learning platforms can be tools for evaluating the quality of teaching on the basis of the cognitive response of the students, and even to develop customized courses based on data obtained from the actions carried out by the individual learner.

Big Data can be used for the whole sector improvements - from a single school, to governmental directions. With distance learning and online courses, all that data acquires a completely new meaning. Big Data could support changes in education that would revolutionize all educational aspects - the way students learn, teachers teach, administrations manage, employers select and develop staff. (Kabakchieva & Stefanova, 2015, p. 54).

We could synthesize the 5 main goals of implementing Big Data within the educational system (Kabakchieva & Stefanova, 2015, pp. 53-55):

- *Improve student results;*
- *Creating mass student-centric programs;*
- *Improving the learning experience;*
- *Increasing results and reducing drop-outs;*
- *Improving education guidance through longitudinal study during university and after student graduation.*

6.1. Improve student results

Student performance is today only measured by the answers to assignments and exams. We don't know what is hidden behind the formal results. Gathering data from different sources, for different students, for different processes and for all participants, is now possible to monitor all student actions – to analyse how long they need to answer a question, what sources they prefer to use, which questions they skip, which teaching resources work best for which student, what is the relation between teaching materials and questions, etc. (Kabakchieva & Stefanova, 2015, p. 54).

6.2. Creating mass student-centric programs

With many students enrolled, universities create big amounts of data that are often not used in the best way. The collected data related to the different educational process aspects could be a basis for creating customized programs for each student. We could give students opportunities to design their own personalized program, following classes from their interest, while having the possibility for directed communication with professors, administrators and colleagues.
If we think to MOOCs, the Massive Open Online Courses, which have surely democratized access to education, we must remark that there is one aspect of MOOCs that is new and powerful: the data they generate. The data we can collect teach us what is most effective; they can tell us things we couldn't know before, since there was no way to unlock its secrets. (Mayer-Schönberger & Cukier 2014, p.4).
At the moment, most of the MOOC courses are still mass produced, but in the future they could be mass customized.

6.3. Improving the learning experience in real time

Big Data can give insights on how each student learns, and it could be very important for improving the learning effectiveness and results. Some students are very concentrated and learn very efficiently, while others may be slow and extremely inefficient. Delivering course materials online allows the whole learning process to be monitored and analysed. "Using innovative Big Data analytical tools, the professor could monitor students in real time and direct their interest in deeper topics of choice. This approach would give students an opportunity to gain a better understanding of the subjects and be confident of constructing individual approach. When students are monitored in real time, digital textbooks and course outlines can be improved. Special algorithms could monitor how the students read the texts,

including identification of which parts are difficult, which are easy, and which are unclear. Changes could be based on analysis of how often a text is read, how long it takes a text to be read, how many questions are asked about a specific topic, how many links are clicked when looking for more information, and how many and which sentences and paragraphs are underlined. If this information is provided in real time, professors could adapt their textbooks to meet the needs of students, in this way improving the overall education process results" (Kabakchieva & Stefanova, 2015, p. 55).

6.4. Increasing results and reducing drop-outs

All analyses on educational big data wants to improve student results and reduce drop-out rates. Closely monitoring students according to the outlined key performance indicators and receiving instant feedback about their personal needs, and delivering guidelines accordingly, could help to reduce the number of drop-outs.

6.5 Improving education guidance through longitudinal study during university and after student graduation.

It's time to think about a longitudinal study during and after university; big data educational analysis should continue after student graduation and monitoring could be used to analyse how students perform in their jobs. This information would improve education and career guidance, giving future students better insights in order to choose the appropriate university.

7. Big Data: handle with care

The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Data-driven predictions can succeed — and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves. (Silver, 2012)

We live in the era of big data. It may be spontaneous to collect more and more data. But the data must be of a certain type to really add value to an organization. Big data does not necessarily mean more information: the belief, often widespread, that *more data = more information* does not always correspond to the true (cf. Interaction Design Foundation, 2016)

Among big data there are obviously interpretable data and data that we cannot be interpreted (sometimes because they lack the meta-data or place/time references). Among the interpretable data we have relevant data (the "signal") and irrelevant ("noise") for our aims. Relevance is a characteristic of data, not only subjective (what is for me "signal" to another could be "noise"), but also contextual (what may be relevant to me depends on the context that I am analysing). This concept is represented in figure 11.

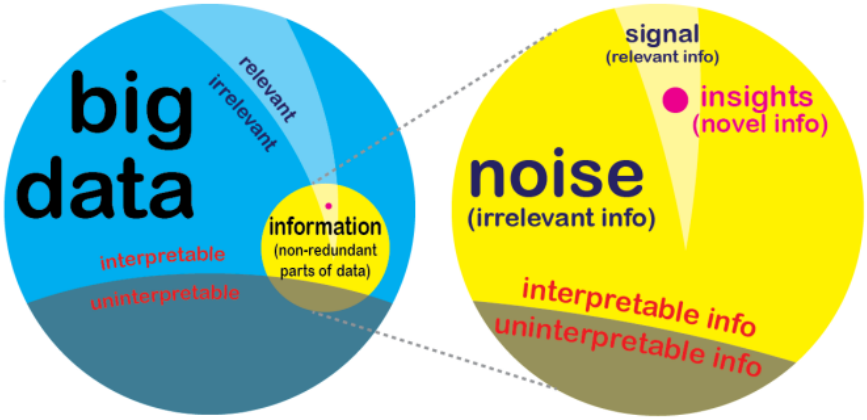

fig. 11. (source: Wu, 2012)

So a criterion to decide if it makes sense to think of an analysis based on big data would be to think about the interpretability, relevance, and if the process could extract from the mass of data really new information.
I would close this paper with a quote that can give us at the same time an alert and a perspective:
"In the twentieth century, education was the great equalizer. Now, with big data, there is a risk that our predictions of potential outcomes, probabilistic outcomes, may make education the setting that widens inequalities." (Mayer-Schönberger & Cukier, 2014).

References

Ackoff, R. L. (1989). From Data to Wisdom, *Journal of Applies Systems Analysis*, Vol. 16, 3-9.

Alahuhta P. (2014), Big Data Analytics –Business Opportunities and Challenges. Digitalization-Key to Growth-Seminar in Espoo, Finland 24.9.2014, Retrieved from http://www.slideshare.net/petterialahuhta/alahuhta-big-dataandanalytics24sep2014

Anderson, C., (2008). The end of theory. Will the Data Deluge Makes the Scientific Method Obsolete?, *Wired* Magazine, 16.07, Retrieved from https://www.wired.com/2008/06/pb-theory/
Box, G. E. P. (1976), Science and Statistics, *Journal of the American Statistical Association*, Vol.71, pp. 791-799

Ayres I. (2008), Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart, New York: Random House Publishing Group.

Blair, D. C. (2002). Knowledge management: hype, hope, or help?. *Journal of the American Society for Information Science and Technology*, *53*(12), 1019-1028

Cameron, W. B. (1963). Informal sociology: A casual introduction to sociological thinking. New York: Random House.
D. Cielen, D., Meysman, A. D. B. ,Ali, M. (2016). Introducing Data Science-Big data, machine learning, and more, using Python tools, New York: Manning, Shelter Island

Conway, D. (2010). The data science venn diagram. Dataists Retrieved, from http://www.dataists.com/2010/09/the-data-science-venn- diagram/.

Cordoba, R (2016). Foreword. In Daniel, Big data and learning analytics in higher education: Current theory and practice.(pp. vii-viii). Switzerland: Springer

Daniel, B. K. (Ed.) (2016). Big data and learning analytics in higher education: Current theory and practice. Switzerland: Springer

Data Science Association (2013). Terminology. Retrieved from http://www.datascienceassn.org/code-of-conduct.html
Silver, N. (2012). The Signal and The Noise: Why Most Predictions Fail but Some Don't. New York, NY: The Penguin Press

Davenport, T. H., & Prusak, L. (1998). Working knowledge: How organizations manage what they know. Boston: Harvard Business Press.

De Francisci S. (2015). La visualizzazione dei Big Data. *Documenti ISTAT*. Retrieved from http://www.istat.it/it/files/2015/05/Big-Data-Visualization-ForumPA2015-finale1.pdf

De Mauro, A. & Greco, M. & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics, *AIP Conference Proceedings*, 1644, 97-104. http://dx.doi.org/10.1063/1.4907823

DeLillo, D. (2003). Cosmopolis: A novel. New York: Scribner.

Elliott M. (2013). Big learning data. Alexandria, VA: ASTD Press.

Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. *Journal of information science*, *35*(2), pp. 131-142.

Gantz J. & Reinsel D. (2011). Extracting Value from Chaos. Retrieved from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Gartner. (2012). Big Data. Retrieved from http://www.gartner.com/it-glossary/big-data/

The Industry of the Future (2015). Ministère de l'Économie et des Finances Français. Retrieved from http://www.economie.gouv.fr/files/files/PDF/pk_industry-of-future.pdf

Information Resources Management Association. (2016). Big data: Concepts, methodologies, tools, and applications. Hershey, PA: Information Science Reference.

Interaction Design Foundation (2016). *Three Common Problems in Enterprise System User Experience*, Retrieved from https://www.interaction-design.org/literature/article/three-common-problems-in-enterprise-system-user-experience

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

Jordan M. (2015). *Modelos DIKW conceptuales valiosos*, Retrieved from http://informationxdummies.blogspot.it/2015/05/modelos-dikw-conceptuales-valiosos.html

Kabakchieva, D., & Stefanova, K. (2015). Big Data Approach and Dimensions for Educational Industry. *Economic Alternatives*, (4), pp. 47-59.

Klein J. (2014). Relational Data Lake, SQLBlog, Retrieved from http://sqlblog.com/blogs/jorg_klein/archive/2014/12/18/relational-data-lake.aspx

Kristensen A. (2014). Big Data Platform. Retrieved from http://www.slideshare.net/ibmsverige/ibm-big-dataplatform
Leboeuf K. (2016). What happens in one internet minute?. Excelacom. Retrieved from http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute

Lemberger, P., Batty, M., Morel, M., Raffaëlli J. (2015), Big Data et machine learning: Manuel du data scientist, Paris: Dunod

Marr, B. (2015). Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance. Chichester (UK):John Wiley & Sons.

Manyika J. et al. (2011). Big data: The next frontier for innovation, competition, and productivity. Mckinsey Digital. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, V. & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt.

Mayer-Schönberger, V. & Cukier, K. (2014). Learning with big data: The future of education. Boston: Houghton Mifflin Harcourt.

Minor, K. (2013). How Big Data and Cognitive Computing are Transforming Insurance. Retrieved from http://www.ibmbigdatahub.com/blog/how-big-data-and-cognitive-computing-are-transforming-insurance-part-2

MIUR (2016). Rapporto del gruppo di lavoro Miur sui big data del 28.7.2016. Retrieved from http://www.istruzione.it/allegati/2016/bigdata.pdf.

Omid, M. (2014) How to characterize DIKW (Data, Information, Knowledge, Wisdom) hierarchy?. Retrieved from http://www.researchgate.net/post/How_to_characterize_DIKW_Data_Information_Knowledge_Wisdom_hierarchy

Petro B. (2011) Welcome to the Zettabyte Era, Info Exponential. Retrieved from http://infox.billpetro.com/2011/06/05/welcome-to-the-zettabyte-era/

Rao, V. M., Kumari, V. V., & Silpa, N. (2015). An extensive study on leading research paths on big data techniques & technologies. *Technology*, *6*(12), 20-34.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), pp. 163-180.

Silver, N. (2012). The signal and the noise: Why so many predictions fail-but some don't. New York: Penguin Press.

Soloviev, K. (2016). 3 Steps to a Data-Driven Content Quality Approach. *Contentquo*. Retrieved from http://www.contentquo.com/blog/3-steps-to-data-driven-quality-approach/

UNECE - United Nations Economic Commission for Europe (2013), Classification of Types of Big Data. Retrieved from http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

UNECE -United Nations Economic Commission for Europe (2014), How big is Big Data? Exploring the role of Big Data in Official Statistics. Retrieved from http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307

Van Rijmenam, M. (2014) Think Bigger: Developing a Successful Big Data Strategy for Your Business, New York: AMACOM Div American Mgmt Assn.

Ward, J.S., Barker, A., (2013). Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821. Retrieved from https://arxiv.org/abs/1309.5821v1

Wu, M (2012), *The Big Data Fallacy And Why We Need To Collect Even Bigger Data*, Techrunch, Retrieved from https://techcrunch.com/2012/11/25/the-big-data-fallacy-data-≠-information-≠-insights/

Zikopoulos P.C. et al (2013) Harness the Power of Big Data. The IBM Big Data Platform. New York: Mc Graw Hill