BIG DATA: ISSUES AND AN OVERVIEW IN SOME STRATEGIC SECTORS

Massimiliano GIACALONE[1]
PhD, Researcher, Department of Economics and Statistics,
University of Naples Federico II, Italy

E-mail: massimiliano.giacalone@unina.it

Sergio SCIPPACERCOLA[2]
Associate Professor, Department of Economics, Management, Institutions,
University of Naples Federico II, Italy

E-mail: sergio.scippacercola@unina.it

Abstract

Big Data is a new technology with a model that works with a large amount of various type data (structured, semi-structured and unstructured) differently from static data being stored in warehouse. The data are generated from a variety of instruments, sensors and mainly by computer transactions. They are constantly updated with a high frequency and become more and more accurate and precise with the passage of time. Main purpose of this paper is to bring into light the new technologies, process and statistical analysis to extract values and results from Big Data. This work, in the first part, introduces the main characteristics of Big Data and its basic management. Important suggestions are developed for a quality control before to extract significant samples for subsequent analysis. Follows, in the second part, the comparison with other traditional techniques. In the last part, the paper highlights the growing role of Big Data and the key benefits in some strategic sectors (Education, Health Care and Banking Industry). Common to our interest fields, the principles of ethics and privacy, to be observed, are also mentioned.

Key words: Big Data, Data Quality, Data Mining, E-learning, Learning Analytics, Health Care, Banking Industry, Ethics, Privacy

Introduction

Big Data were born because of the massive proliferation of elementary data from multiple sources. Sets of images, e-mail, GPS data, or information obtained from web sites (such as access, permanence, etc.) can be defined Big Data (Snijders, et. al., 2012). One of the fundamental characteristics of Big Data is the heterogeneity of data sources: these are data sets, or frequently of dynamic flows of 'metadata', from heterogeneous databases (Rezzani, 2013). Table 1 shows, in a non-exhaustive list, the main sources from which they are taken Big Data. Almost all data is generated typically at sub-daily basis: hours / minutes / seconds / milliseconds. For example, not only censuses, surveys, interviews, or questionnaires; but also information collected from the Internet, by telephone networks, by satellite, or for transport, may be part of the same set of data.

Big Data is often confused with simple digital traces of human activities mediated by information and communication technology, as are the recordings of access to services which are called the log of service (phone calls, messages exchanged with identification of the applicant, short texts, and geolocation). Even the logs can be considered data of a Big Data system.

Table 1. Some Data Source of Big Data and type of data

Data source	Type of data generated
E-mail, SMS, instant message, YouTube, WhatsApp, Web	Textual, graphical, and audio video
Electronic medical instruments, scientific experimental and observational data	Numerical (i.e. temperature, pressure, etc.), and diagnostic images (i.e. computerized tomography, ECG, etc.)
Environmental sensors	Numerical, textual, graphical, audio-video
Financial transactions	Textual and Numerical
Traditional Database and Datawarehouse	Numerical and textual
Satellite	Numerical and graphical

A common definition of Big Data is that offered by Doug Laney (Laney, 2001), which is based on the paradigm of the three V (Volume, Velocity, Variety) (Table 2):

- **Volume**: it is estimated that by 2020 a measure of 35 thousand billion gigabytes of data will been generated. With regards the enumerations of the bases of Big Data, it the measure would have gradually proceeded to extend the measuring units that gradually proceeded in extension of the average-sized volumes in place, arriving today with volume orders of magnitude expressed in 'Zettabyte', equal to one billion terabytes and Yottabyte equal to one trillion of bytes.
- **Velocity:** Once extracted, the data must be analyzed promptly, not to become obsolete, and therefore unnecessary to make a "decision". The fast acquisition and access to required data is therefore essential. Just think that it is not uncommon the need to acquire 'live data' (for example, access to sites, search engines in the Internet, or share data in television), to process on a daily basis, and, mainly at sub-daily basis.
- **Variety:** Data have highly heterogeneous nature (eg., texts, images, videos, web searches, financial transactions, email, post on blogs and social networks, etc.), and each size requires a dedicated treatment. This characteristic of Big Data may require scaling operations or conventional classifications (for example catalog of images for the chronological date, or for chromatic scale or according to another ordinative scale) (Manyika et al., 2011).

Some scholars suggest adding to the definition of Big given two more V:

- **Variability**: the data must be contextualized, as their meaning can vary depending on the context.
- **Virality:** the growth of Big Data is exponential, like wildfire.

These peculiar features and specifications require that, with respect to storage, the constituents Big Data, are both structured to unstructured, and are expressed on different measurement scales, or are also qualitative (Table 2).

Therefore, Big Data is not just a lot of data, but it is a **System to handle a large amounts of data of any type.**

Table 2. Features of Volume, Velocity and Variety of Big Data

Feature	Size, Time and Type of data
VOLUME	Size: TB (terabyte = 1024^4 byte) - PB (petabyte =1024^5 byte) - EB (exabyte = 1024^6 byte) - ZB (zettabyte =1024^7 byte) -YB (yottabyte = 1024^8 byte)
VELOCITY	Time: Results in real time: fast acquisition and access to data is essential, especially for 'live data' that must be processed on daily or sub-daily basis.
VARIETY	Type of data: Structured – Semi-structured – Unstructured (qualitative)

In this complex panorama, the aim of this work is to bring into light the new technologies, process and statistical analysis to extract values and results from Big Data.

The paper is structured into three parts: introduction to Big Data, comparison with other traditional techniques, main applications to some strategic sectors.

In the first part, after the present introduction, the difficulties inside the Big Data management are presented. Due to the large amount of data produced in continuity and to the need to work on samples drawn from the population it is essential to carry out a preliminary Data Quality Statistical Control (Section 2).

In the second part, we compare the Data Mining methods already used for some time, with the new frontiers opened by Big Data (Section 3). This part is devoted to the applications of Big Data in strategic sectors as E-learning, Learning analytics, Healt Care, Banking Industry (Sections from 4 to 7).

In the last part, the principles of ethics and privacy in the era of Big Data are discussed (Section 8) and the main benefits of using Big Data in the analysed sectors are underlined (Section 9).

Finally, in the conclusions, the perspectives that today offer the Data Science including Big Data are emphasized.

2. Quality Assessment Process for Big Data

The management of Big Data is very complex because many are the data stored and sometimes the Big Data are erroneously also referred to as large the data set. We must filter from this very lot of data selecting only those that meet the quality control requirements. The filtered data become statistical sample to which it is possible to apply inference or traditional analysis methods of Data Mining. Otherwise, for work directly with Big Data we need only apply special parallel algorithms. In this context it is useful, also, the adoption of 'genetic algorithms' capable of operating a meeting of non-metric also data from dynamic sources coagents in the process of dataset formation. These algorithms can be used for categorical or probabilistic selection methods (selection of 'Boltzmann') (Koza, 1992; Wright, 1991).

The progress made in the meantime by the scientific and technological research in hardware and software area have ensured satisfactory performance in terms of efficiency, access to Big Data and power and effective processing speed.

Big Data collection requires to acquire and analyze data from several sources and with various researchers. For this reason decision-makers have gradually realized that this massive amount of information has benefits for understanding customer needs, improving service quality, and predicting and preventing risks. It is also logical that use and analysis of accurate high-quality data is a necessary condition for generating value from Big Data. Therefore, we analyzed the challenges faced by Big Data and a quality assessment framework and assessment process for it.

In the last years, Xi'an Jiaotong University set up a research group of information quality that analyzed the challenges and importance of assuring the quality of Big Data and response measures in the aspects of process, technology, and management (Zong & Wu, 2013).

Big Data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it.

An appropriate quality assessment method for Big Data is necessary to draw valid conclusions. In this paragraph, we propose an effective data quality assessment process with a dynamic feedback mechanism based on Big Data's own characteristics, shown in Figure 1.

Different tasks like filtering, cleaning, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.

After the quality assessment preparation is completed, the process enters the data acquisition phase. If the analysis results meet the goal, then the results are outputted and fed back to the quality assessment system so as to provide better support for the next round of assessment. If results do not reach the goal, the data quality assessment baseline may not be reasonable, and we need to adjust it in a timely fashion in order to obtain results in line with our goals. Poor Big Data quality will lead to low data utilization efficiency and even bring serious decision-making mistakes.

The application of SPC methods to Big Data is similar in many ways to the application of SPC methods to regular data. However, many of the challenges inherent to properly studying and framing a problem can be more difficult in the presence of massive amounts of data.

There exist several frameworks for solving problems in the Total Quality Management (TQM), Statistical Process Control (SPC), or Six Sigma area (Montgomery, 2013).

The classical tools are the Plan Do Check Act (PDCA) Deming cycle, or the Define, Measure, Analyze, Improve, Control (DMAIC) cycle can be applied to Big Data (Qiu, 2014). For example, the Cross Industry Standard Process for Data Mining (CRISP-DM) and knowledge discovery in data mining (KDD) were recently introduced.

It is important for researchers in statistical surveillance to consider processing speed when developing and refining methodologies. Another challenge in monitoring high dimensional data sets is the fact that not all of the monitored variables are likely to shift at the same time; thus, some method is necessary to identify the process variables that have changed (Megahed & Jones-Farmer, 2015).

Another important challenge when using SPC methods with big data applications is that, traditionally, SPC methods were developed for numeric data. While there are some attributes control charts, these tend to be a distant choice to using methods designed for quantitative variables.

However, one of the great challenges of big data is the ability to process and analyze unstructured data. Most of big data applications are concerned with non-numeric data obtained from several databases.

In the next future, more complex hierarchical structure of a data quality system could be analyzed and proposed to evaluate the Big Data quality framework.

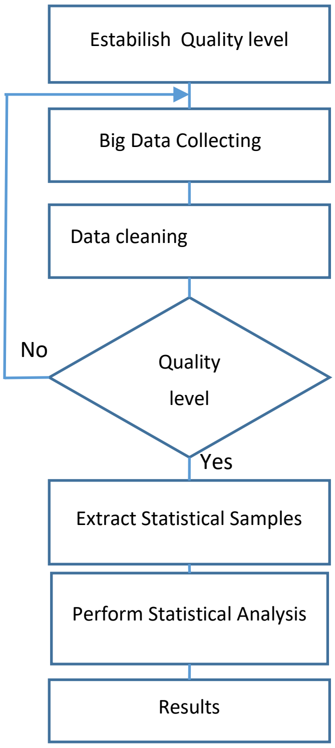

Figure 1. Big Data Statistical Quality Control

3. Data Mining and Big Data

Data Mining is part of Business Intelligence, and indicates the process of exploration and analysis of a set of data to identify any regularity, extracting new knowledge and meaningful applicantion rules.

The main objective of the "Data Mining" is to "extract information" useful from a database and turn them into a data structure (pattern) for further use survey. Among the main applications of Data Mining we can highlight the summary description of the data, the associations and correlations, classifications, and evolutionary analysis (regularity of data that changes over time). The techniques of data mining are adopted in various fields as Statistics, Sciences of Education, Economics, Medicine, etc.

There are clear similarities found among the "Big Data" and "Data Mining". The latter could be considered the **old Big Data** because it responds at least in part to two of the characteristics of Big Data that are the size and velocity, but lacks the third V (variety) as the Data Mining is often extracted knowledge only by means of the Database or Data Warehouse and Data Mart that are retrospective static type unlike Big Data constantly updated with a high frequency and become more and more accurate and precise with the passage of time. Therefore, Data Mining could be considered the **old Big Data** and Big Data could be considered the **new Data Mining.**

Another interesting aspect of Big Data that differentiates it from the Data Mining is the **structural diversity** (Fig. 2). Some data have a well-defined format, in the classic way of files / records / fields, such as, for example, in the transactions recorded in a database other data may be of very different type (i.e. Municipal data, Driving Licence Data, etc.); such as a collection of texts on a blog, or tables, or images, or audio recording, or video.

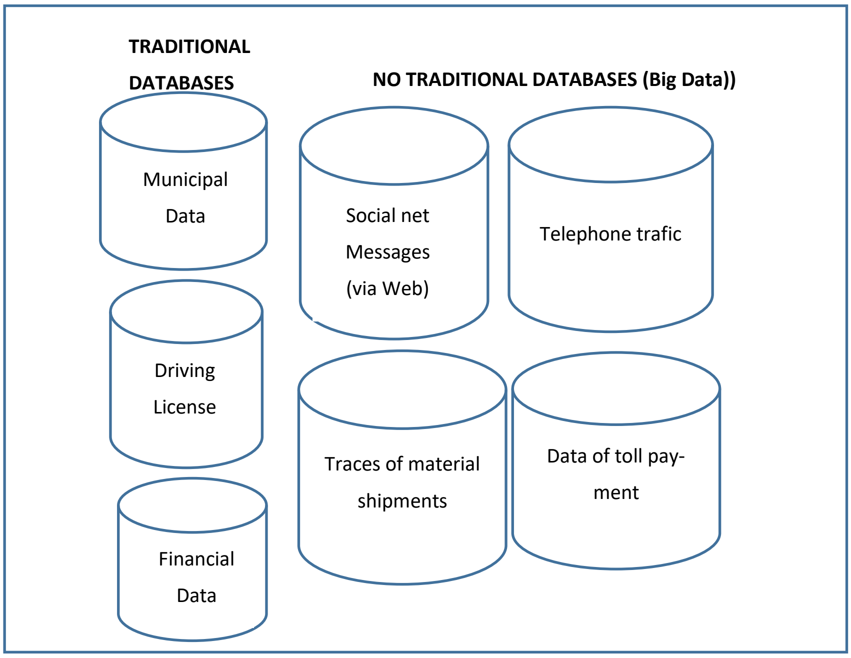

Figure 2. Examples of structural diversity between Data Base and Big Data

From the point of view of the architecture and engineering of the dataset and data structures, the latest models of Big Data are based on highly scalable methodologies, and type of **No Structured Query Language** (***NoSQL***) solutions (Vaish, 2013). It is intended for *No Structured Query Language* a set of technologies forming a different new data management system from the traditional *Relational Data Base Management System*, because the relational model is not used, it does not have an explicit scheme and the system is designed to work quickly and well in the cluster.

The Big Data have redundant informations (**redundancy conditions**) and it is preferable to work with samples. A preliminary inferential approach to aggregation comes beore the actual creation of databases and datasets useful for the statistical analysis (Manoochehri, 2013).

Briefly, the aggregation of mixed numerical sources is addressed by operating on data streams in parallel (*approach map*) then subjected to reduction treatment, filtering and 'clean' data eliminating those untrue or unnecessary (*Data garbage*) before to operate combinations and reorganizations in the final dataset (Reiss et al., 2012).

4. Big Data for e-Learning

The impact that Big Data in education - both with reference to teaching, which learning - is relevant, not only in the design of the modules, but also in terms of refinement of learning objectives already predefined (Gutierrez-Santoz et al., 2012). The Big Data can be used in multiple sectors and e-learning is one of them.

The traditional or *e-learning training* can be evaluated at four progressive levels (Kirkpatrick, 1979) (see Table 3, with our adaptation to domain of the education). Before the delivery of the training, we should have identified our strategy, completed an assessment and built a plan. Then, during the delivery of the education solution, we need to manage a number of factors to ensure success. After the delivery, we have to evaluate the success of the implementation in terms of the originating need and strategy (Giacalone, Scippacercola, 2016)

The E-learning teaching materials should be built ad hoc to ensure the four main characteristics of online education:

- Modularity: course material should consist of "learning modules", also called *Learning Objects;*
- Interactivity: the student must interact with the system by providing his answers that are properly recorded;
- Exhaustiveness: each module should contain a complete topic;
- Interoperability: instructional on any platform and technology to ensure traceability of the training.

Currently the most common standard is the Shareable Content Object Reference Model (SCORM) (Bohl et. al, 2002). Technological progress has led to the creation of the Learning Content Management Systems (LCMS) that deal with the content management both in the process of creation and during the delivery: they can be considered a complete platform for e-learning. Today we are able to track and collect this data also through social networks and any other media.

Table 3. Evaluation of e-learning training

Action	Evaluation and measurement of
Reaction	personal reaction to the training
Learning	the increase in knowledge
Behavior	changes in on-the-classroom behavior
Results	obtained vs desidered results

Each time that the learners (students) interact with the content of a course, in fact, they produce data.

Beside the usual 'assessment of end-over', by means of the satisfaction questionnaires proposed to learners, it grows and becomes relevant the need to acquire real-time information always more detailed and organized on the various areas of teaching evaluation. For example, accesses ('visits') to Web pages are data that can be purchased on-line with other data, to compose patterns useful for teaching evaluation.

By Big Data, the e-Learning responsible teachers can receive information to make teaching more effective, or to correct any defects. For example, access to websites, the data collected from social networks, the content of the web searches, and online learning modules, Big Data can be useful to assess the information use by learners and their behaviors in the learning phase.

An interesting prerogative is given by the possibility, using special software programs or power tool immediately discard the data not useful from the information point of view. The use of mathematical models and statistical methods on data of e-Learning, once organized the same in databases or 'metadata', allows us to produce models of understanding or even useful prediction refinement or simple evaluation of teaching methods (Chatti et al. 2012).

Another approach to the use of Big Data, is to evaluate different parameters of prefixing didactic training for each variable appropriate 'threshold values' or 'levels-target' to achieve the educational objectives. (Siemens et al. 2011).

Christopher Pappas (Pappas, 2014) listed in this regard five benefits that can be drawn from the analysis of data related to the use of a course and e-learning:

1. The data analysis allows you to identify which type of teaching is more effective in achieving the objectives of the course.
2. It becomes possible to identify improvements of the educational path. For example, if a large number of learners taking too long to complete a certain module, means that the form must be made more slender and usable
3. And it is possible to monitor what are the forms displayed the most shared links with other learners.
4. The data resulting from the traces of the learner are immediately available and there is no need to wait for the final evaluation of the test results to know the situation. In this way, teachers can get an overall picture of learners' behavior and can optimize the learning strategy in near real time.
5. Based on the data it is possible to make predictions about the successes and failures of learners and develop in a way that courses that students have always the possibility of obtaining the best possible result (Pappas, 2014).

In summary, the main advantage of collecting and analyzing Big Data in e-learning, is mainly the possibility of obtaining useful information to customize the learning experience based on the needs and learning styles of learners (Giacalone, Scippacercola, 2016).

5. Big Data for learning analytics

The term *learning analytics* identifies an important sector within the Technology-Enhanced Learning emerged in recent years and is closely related to several disciplines such as Business Intelligence, Web analytics and Educational Data Mining (EDM). The term learning analytics refers to the measurement, collection, analysis and presentation of data on students and their contexts for understanding and optimization of learning and the environments in which it takes place (Baker et al., 2014) (Ferguson, 2012, 2014).

The transfer of the knowledge through *learning objects* in the various environments is missing of reference standards for the assessment. There are various products (like Google Analytics, Omniture SiteCatalyst, WebTrends, Coremetrics, etc.) that allow the retrieval of information transmitted over the web.

The evaluation of the dissemination of knowledge via web can be done by traffic parameters that can be listed as inferred from such a traffic controller that can detect what in slang is called the *Visitors Overview*. This traffic overview allows you to view in detail the aspects of quality (ie average pageviews, time spent on site, bounce rate) and characteristics (for example, first time visitors, return visits) of visits. The traffic indicators can be classified in two types (Scippacercola, 2012):

– indirect (the number of accesses to the module, the usage time of a session, the mode of use, flow of the navigation in the website, etc.)

– direct (the average response time to questions, the number of attempts before you answer correctly, etc.) The direct indicators are derived often by user surveys.

Exist metrics that allow evaluating, ex-post, the personal reactions to the training and permit to evaluate the validity of the a web page or for directing the eventual reengineering. Assume that a web site (with four pages *P_i*) (i = 1, 2, 3, 4), illustrated schematically in Fig. 3 in the box inside, is visited by three (A, B, C) hypothetical students that enter the website, and consider, for example, the following actions:

- Student A sees P_1, P_2, P_1 and then exits from the website;
- Student B sees P_4, P_2 and then exits;
- Student C sees P_3 and immediately exits.

Using the above indicators it is possible evaluate the traffic of students and the analytic behavior on the same network.

Referring mainly to the most widely used Web analytics (Google) (Clifton, 2010; Vasta, 2009) we list the main indicators that it gives us (Fig. 3) (Scippacercola, 2012):

- *Entrance*: the *number of inputs* to the page P_i**;**
- *Pageviews*: is the *total number of requests* for loading a P_i of the website;
- *Unique Pageviews*: is the *number of sessions* in which a page was viewed more than once;
- *Average Time on Page*: is one way of measuring visit quality. A high Average Pageviews number suggests that visitors interact extensively with the web site;
- *Bounce Rate:* is the *percentage* of single-page visits (i.e. visits in which the person left your site from the entrance page). The percentage of visits where the visitor enters and exits at the same page without visiting any other pages on the site in between. The Bounce rate is one way of measuring visit quality. A high bounce rate generally indicates that the entry pages (landing) is not relevant to your visitors.
- *Exit*: is the *percentage of users* who exit from a page.

The approach here considered can be classified as a theoretical approach to **ex-post non-interactive**. Conversely other approaches tend to interact during the learning phase (**interactive approach**). In Ferguson is reported, for example, the Signals Project (Ferguson, 2014), developed by the Purdue University explores large datasets and apply statistical tests to predict, during the courses, students who risk being left behind.

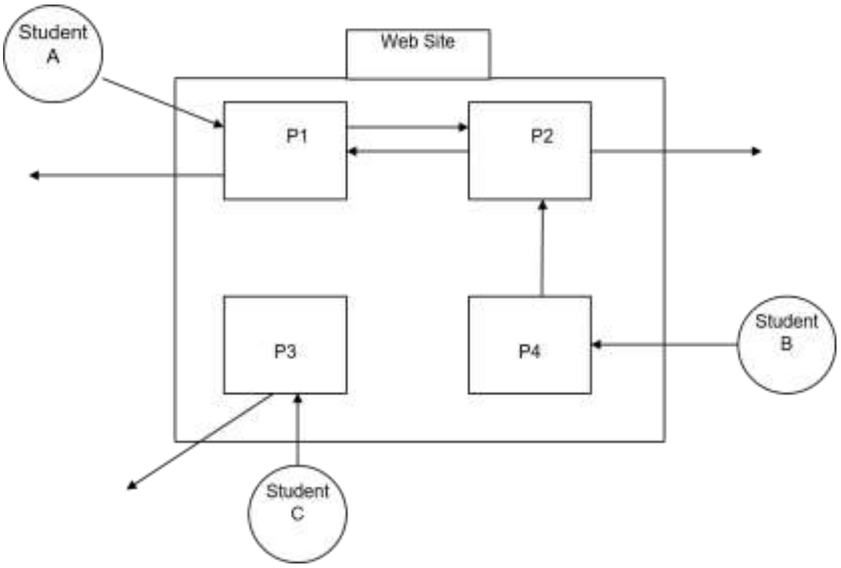

Figure 3. Students which link to a website to make navigation in four pages (P_1, P_2, P_3 and P_4)

The goal is to produce actionable intelligence, guiding students to appropriate resources and explaining to them how to use them. A traffic light shows students if things are going well (green), or if they were classified as high risk (red) or moderate risk (yellow) (Alan et al., 2010). The reported results are promising although the system as data and software may not be entirely comparable with a Big Data system.

From a technological point of view learning analytics is an emerging discipline and its connections with Big Data, despite some significant proposals in American College (Picciano, 2012), it remains to be developed.

6. Big Data in Health Care System

In the Health Care sector the information is the most important aspect, and the human body, in particular, is the major source of production of data. Consequently, the new challenge for health care world is, knowing how to take advantage of these huge amounts of unstructured data between them. Electronic medical records include within them data with each other very heterogeneous in terms of size: audio recordings, magnetic resonance imaging, computerized tomography and other diagnostic images, electrocardiograms, and the list goes on indefinitely. Nevertheless, electronic medical records have to be designed to process and manage data characterized by high volumes, generating speed and wide variety of sources (Sanchez et al., 2014) (Murdoch et al, 2013).

Organize Big Data health means being able to sort the huge amount of information about the medical history of each patient. A concrete example is the electronic medical file that will soon replace the medical records. A single support will allow the patient to store in one device prescriptions, drugs, diagnostic tests, laboratory analysis findings, emergency department, hospital, and the doctor to rebuild quickly and accurately the state of overall health and especially the patient, in addition to being able to share information with other doctors in the case of diseases that require more expertise.

The *Big Data analytics*, cloud computing, social networking and the emergence of micro-sensors are the main technologies improving predictive analysis in the medical field and the patient's quality of care. The traditional Data Warehouse strategies are not easily and quickly scalable and they provide a retrospective view and not in real time or predictive. Through data analytics we can classify data, make predictions, and greatly increase the understanding of patient's clinical data (Raghupathi, 2014).

The **Clinical Intelligence** (Fig. 4) (Groves et al., 2013) consist of all the analytical methods, made possible through the use of computer tools in the set of processes and disciplines of the mining and processing of raw clinical data into meaningful insights, new discoveries and knowledge that ensure greater efficiency clinical and better health-related decisions (Harrington, 2011). Clinical intelligence is the set of electronic methods, processes and disciplines extraction and transformation of raw data into meaningful clinical insights, new discoveries and knowledge that affect the clinical decision-making and the decisions in the health sector.

The clinical intelligence differs from *business intelligence* for the following considerations. The business intelligence deals with raw economic data, often structured, and provide insights and information on the decision-making process in the economic field. In contrast the clinical intelligence deals with clinical data and requires statistical methods and analysis much more sophisticated than that used by business intelligence.

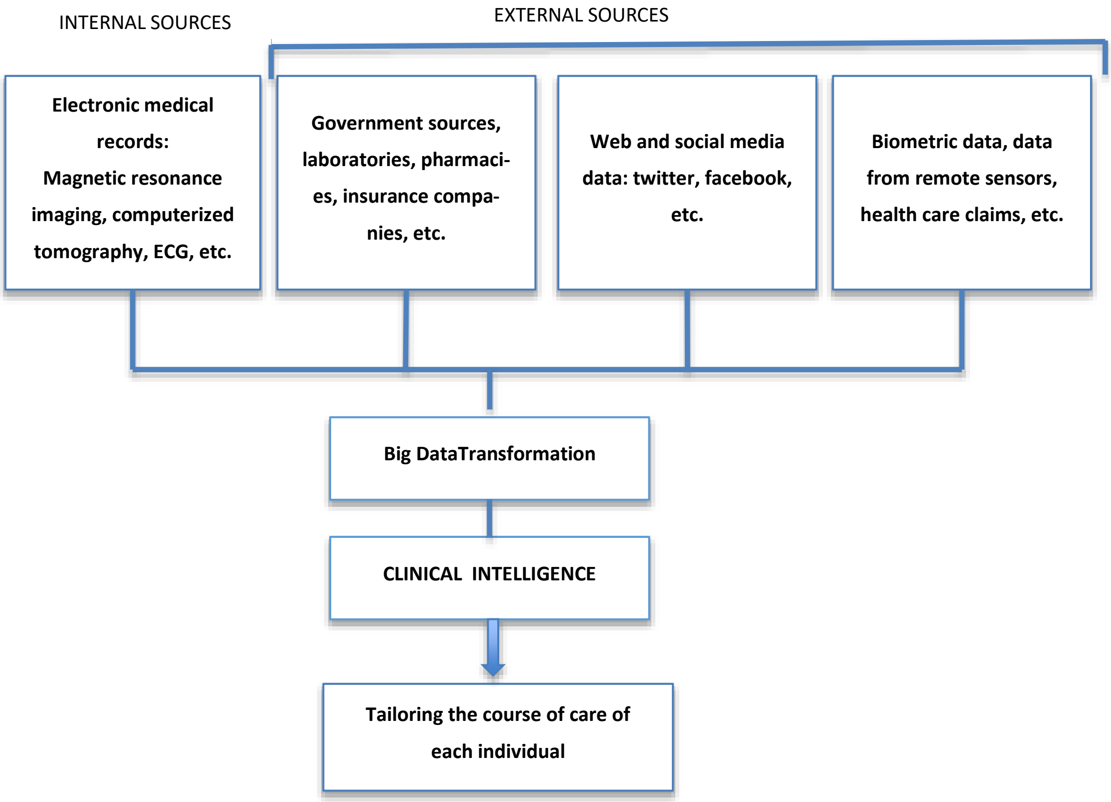
INTERNAL SOURCES
EXTERNAL SOURCES
Electronic medical records: Magnetic resonance imaging, computerized tomography, ECG, etc.
Government sources, laboratories, pharmacies, insurance companies, etc.
Web and social media data: twitter, facebook, etc.
Biometric data, data from remote sensors, health care claims, etc.
Big DataTransformation
CLINICAL INTELLIGENCE
Tailoring the course of care of each individual

Figure 4. Some main sources for Clinical Intelligence

The clinical intelligence often has to do with types of data unstructured and much more complex as data that often arise ambiguous, incomplete, conditional and inconclusive. The clinical intelligence uses sophisticated methods to study the data and interpret the results as machine learning techniques, non-linear and multi-algorithm approaches. The clinical intelligence allows a more sophisticated classification of patients' based solely on demographic variables such as age, sex, lifestyle but also on relevant medical and clinical features related to certain diseases, medical conditions, genetic predispositions and the likelihood of therapeutic response (Chawla et al., 2013).

The clinical intelligence makes it possible to optimize and custom *tailor the course of care of each individual* patient by basing it on a multitude of factors that define the medical protocol of care: previous medical history, known allergies, personal risk factors, genetic traits, lifestyle and business, management of personal safety. The clinical intelligence allows implemention of multi factorial analysis to determine the effective utility associated to different treatment courses (Kayyali et al, 2013). These analysis allow doctors to identify the most appropriate treatment for a particular patient as well as specific indicators for measuring outcomes. Analytical tools, when applied to medicine, are able to suggest medical care plans and clinical pathways providing a prediction of the results corresponding to them. These tools allow you to compare treatment options which are complementary, manage the risks associated with each treatment plan and select the most appropriate course of care (Murdock et al., 2013).

Therefore, Big Data are radically changing the world of health and medicine (Raghupathi et al., 2014; Growes et al., 2013), allowing for more personalized care pathways (patient experience), effective and less subject to clinical risk through new mechanisms for control and governance of processes. In the opposite, while the structured data provide the "what" of a disease or medical treatment, rarely they offer a "why" behind the decisions taken. In many cases, the collection of unstructured data remains the best option to capture the details in depth, for example, a medical record as they contain valuable information on the health of the patient.

7. Big Data for banking

Banks generate a large amount of data: paper documents with signatures, checking accounts, mobile banking, credit and debit cards, loans, etc. Today many data comes from the Bank's contacts with customers in *online mode*. Such data can be classified into structured data like e-mails, chat logs, feeds, posts, web logs and semi-structured data such as customer reviews. Among others the bank's objectives, we present, in particular, the **fraud detection** (Russom, 2011; Hipgrave, 2013), the **money laundering** and the **risk analysis** (Boinepelli, 2015).

Some of **frauds in the banking** are in the "online banking (credit card, internet or mobile transactions)" where the fraudster performs the transactions with the same code or sign of customer. The bank, in order to prevent the success of such fraudulent transactions while running you must provide a customer profile based on the story of its financial transactions. Through the analysis of all the transactions made by customers you can create a "**customer profile**" and its relations with other correspondents and payment methods used. Each new transaction is compared to the profile to identify if an operation does fit into the usual ones and if not, to report suspicion of fraud. To create the profiles are indispensable and techniques ranging from simple statistics (mean, standard deviation, minimum and maximum values of the transactions, dates, etc.) to advanced (inference) will be required.

Money laundering continues to be a local, regional and global concern. To combat money laundering is essential to have databases containing multidimensional data on financial transactions and the database of the police. Clustering, classification, outlier identification, data visualisation tools are used to detect behavior fraudster in transactions with large amounts of money between accounts. Both of these databases are the platform to reveal associations and patterns of activities that help identify the fraudster suspicious.

The fraud detection is in real time by means of suitable software that should be used for example to prevent unlawful claims before they are completed, and analyze business data where fraud has been committed in the past.

The third application of Big Data in the banking is the **risk analysis**. The predicton of default on loan and credit card accounts is important for banking. The model that allows you to make predictions includes a data warehouse that contains historical data of customers, the realized contracts and the observed results. The model provides for the application of traditional Data Mining techniques so as to achieve a financial risk analytics framework. In addition to the Data Warehouse in the model are considered all contacts between the customer and the bank in the period between the loans granted and the end result. Using the results obtained from the aforementioned model, the bank can identify and classify more safely customers with a risk of insolvency.

8. Ethics and privacy

The Big Data collection activities have different effects in the sensitive area of the issues pertaining to the processing of personal data, called 'sensitive' under the legislation and the overall system of constraints and responsibilities, governed by privacy laws.

Relevant are the ethical issues arising from the management of personal data of users, with reference also to the possibilities of diffusion, sharing and use of information 'sensitive' about learners themselves.

There are critical issues that have already led professional organizations to a serious reflection on the contents and on the delimitation of the operational boundaries of e-learning activities in order to reach the fulfillment of plans the collection and interventions 'legal' ie capable of qualifying the various aspects of the activities in place (the collection, conservation, management, use and publication of personal data) in observance of 'best practices' default (Slade et al., 2013).

One of the most important issues is the communication of the intent and purpose of the collection of data for the sectors analysed, in order to achieve preliminary authorization and legitimacy in accordance with local legislation, making clear communication to users interested in each of the which require explicit consent to the processing of their personal data.

If there are 'stakeholder' or external customers of data collection, in the same way the organizations and professionals that carry out studies or data analyzes in the specific sector should carry out all acts of communication and producing relative contracts in accordance with applicable regulations in order the processing of personal data, both with regard to any estate regarding the dissemination of personal data obtained or evicted from the activities in question.

Another aspect is not secondary to the safe preservation and accessibility of personal data in a server equipped with procedures, protocols, and active and passive safety standards, as required by the regulations in force for years, and international safety standards (ISO, EN) (Corposanto et al., 2014).

With regard to the preservation and accessibility of data, technologies in support of Big Data are highly reliable, low-cost and scalable. For example, Hadoop (Hadoop, 2014) is an adequate system to Big Data, because it allows you to store huge volumes of data and then process them when more is appropriate. Hadoop also allows the distribution of data on multiple nodes, reducing the computational and storage costs for storage and analysis of Big Data, and masking hardware failures. It has been estimated by Zedlewski (Zedlewski et al., 2003) that the cost of a data management system based on Hadoop, considering the cost of hardware, software, and other expenses, amounts to about $ 1,000 per terabyte, or by one-fifth to a twentieth of the cost of other data management technologies (Giacalone, Scippacercola, 2016).

9. Benefits of Big Data in the analyzed fields

Benefits for e-Learning, and Learning Analytics. The Big Data are currently used by various companies for training and also in the university: with the help of Big Data, we can watch the learners and examine the traces of their individual paths. For example, we can identify the web pages on which learners will entertain more or which are more learning difficulties, those that often revisit, and determine the days and times they work out more, etc. Therefore, Big Data help us to understand the true role models for learners, much more than now occurs through the traditional education. These models lead to interesting information about what and how they learn. Thus, helping to make informed decisions about learning programs and to identify courses with design flaws. However, the real power of Big Data lies in their power to help predict or forecast scenarios to take preventive measures. For example, with the help of Big Data, it is also possible to make predictions as to what are the concepts that are difficult parts to students, the topics that generate confusion and difficulty in learning. Big Data today is presented as an effective platform that revolutionizes the traditional way in which e-learning was born. By Big Data it is possible to design more personalized learning plans and suitable for students. Monitoring becomes the main element used by educators. They are used the same results achieved by students to improve their training.

Benefits for Health Care. The Health Big Data are conceived as a real digital collection of all that the patient had, assumed or required in the medical field. The challenge, which is also the main difficulty of Big Data, is that the same should also merge any news on the patient's health that it communicates via social networks, like Facebook or Twitter, to his friends or acquaintances. In our life of patients each of us generates Big Data every time it receives a prescription, buy a drug, requires a health service, access to the ER, undergoing a diagnostic examination or laboratory uses Facebook, Twitter and other social network to communicate to friends their health. If it was a constant cross-analysis of all this information for each client: the doctors would have an overall picture of the health of the person, both in general and for a given period; policy makers, hospitals and clinics could include medical bills, preventing the most common diseases, and select healthcare services according to the real needs of the population in a given territory.For example, just an exchange of Tweet among people in constant contact to allow the data scientists to outline the possibility of infection and spread of a disease and then define the most appropriate prevention measures or to better manage health care demand. If the opportunities of Big Data will result in profiling of diagnostics and personalized therapies it will contribute to making the "health service" more effective and more sustainable also in financial terms (Roski et al., 2014).

Benefits for Banking Industry. Most of Big Data are useful for the customer management. The Bank's objective is to use this data to identify the customer profile more precisely (micro-segmentation) than with other methods. New services as tailored accounts to offer to customers are designed. The dialogue with consumers (sentiment analytics, multi-channel customer sentiment) is useful to identify the products that you could develop. The Bank can detect when a customer is about to leave the bank and it is possible to perform risk analysis better fraud detection more precise than before the advent of Big Data.

10. Conclusions

Nowadays we see the confluence of Big Data, Data Mining, Statistics, Mathematics, Computer Sciences, the Data Warehouse, the Artificial Intelligence, and neural networks in a new paradigm, which takes its name Data Science, and that promises to revolutionize the world, affecting all sectors, from health care, up to the academic world. In this perspective, *Data Science* will also modify the way of analyzing data.

The Data Science paradigm consists of extracting data of each type existing "in the world", applying appropriate formats, obtaining descriptive analysis of the phenomena, re-entering the results in the world circuit and so on, always perfecting more knowledge useful to the domain of the applications. From Data Science, it comes a new profession, the Data Scientist, who has the task of analyzing the data to provide useful information to make decisions to the customer.

The Data Scientist is the common professional figure for all analyzed sectors, drawing from the analysis of Big Data new and more effective strategies; he will have to learn to process information and be the responsible for statistical analysis, to determine what changes to make and to suggest choices to improve the processes.

Therefore, the role of Big Data is not only to be able to quickly handle large volumes of different data of various types, but is also given by the opportunity that these technologies offer us for new and remarkable applications, even in education, health care and banking.

With today's tools (mobile, tablets, smart phones, cloud-based technologies, etc.) the infrastructure is well-established. The Data Analytics allow us to get a much better picture of tracking than in the past with conventional methods used so far.

References

1. Alan, F. K., Sanil, A. P., and Sacks Arnold, K. E. **Signals: Applying Academic Analytics.** Educause Quarterly, Vol. 33, No. 1, 2010, pp. 1-10
2. Baker, R. S. and Inventado, P. S. **Educational data mining and learning analytics.** Learning Analytics, Springer, New York, 2014, pp. 61-75
3. Bohl, O., Scheuhase, J., Sengler, R. and Winand, U. **The sharable content object reference model (SCORM) a critical review.** Computers in education. Proceedings. International conference on IEE, 2002, pp. 950-995
4. Boinepelli, H. **Application of Big Data.** in Mohanty, H., Bhuyan, P. and Chenthati, D. (eds.) "Big Data: A Primer". Vol. 11. Springer, 2015
5. Chatti, M. A., Dyckhoff, A.L., Schroeder, U. and Thüs, H. **A reference model for learning ana-**

lytics. International Journal of Technology Enhanced Learning (IJTEL), Vol. 4, No. 5-6, 2012, pp. 318-331

6. Chawla, N. V., and Davis, D. A. **Bringing Big Data to personalized healthcare: a patient-centered framework.** Journal of general internal medicine, Vol. 28, No. 3, 2013, pp. 660-665
7. Clifton, B. **Advanced Web Metrics with Google Analytics,** 2nd ed., Wiley Publishing, Inc., Indianapolis, Indiana, 2010
8. Corposanto, C. and Lombi, L. **E-Methods and web society,** Università Cattolica del Sacro Cuore, Milano, 2014
9. Ferguson, R. **Learning Analytics: fattori trainanti, sviluppi e sfide.** TD tecnologie didattiche, Vol. 22, No. 3, 2014, pp. 138-147
10. Ferguson, R. **Learning Analytics: drivers, developments and challenges.** International Journal of Technology Enhanced Learning, Vol. 4, No. 5/6, 2012, pp. 304-317
11. Giacalone, M., and Scippacercola, S. **Il ruolo dei Big Data nelle strategie di apprendimento,** Atti Conferenza Didamatica, AICA ed., 2016, pp. 1-10
12. Groves, P., Kayyali, B., Knott, D., and Van Kuiken, S. **The 'Big Data'revolution in healthcare.** McKinsey Quarterly, Vol. 2, 2013
13. Gutierrez-Santos, S., Geraniou, S., Pearce-Lazard, S. D., and Poulovassilis, A. **Architectural Design of Teacher Assistance Tools in an Exploratory Learning Environment for Algebraic Generalisation.** IEEE Transactions of Learning Technologies, Vol. 5, No. 4, 2012, pp. 366-376
14. Hadoop, **http://Hadoop.apache.org/2014**.
15. Harrington, L. **Clinical intelligence.** Journal of Nursing Administration. Vol. 41, No. 12, 2011, pp. 507-509
16. Hipgrave, S. **Smarter fraud investigations with Big Data analytics,** Network Security, Vol. 12, 2013, pp.7-9
17. Kayyali, B., Knott, D. and Van Kuiken, S. **The big-data revolution in US health care: Accelerating value and innovation.** Mc Kinsey & Company, 2013, pp. 1-13
18. Kirkpatrick, D., L. **Techniques for evaluating training.** Training & Development Journal, Vol. 33, No. 6, 1979, pp. 78-92
19. Koza, J. R. **Genetic programming: on the programming of computers by means of natural selection,** vol 1. MIT Press: Cambridge, MA, 1992
20. Laney, D. **3D data management: Controlling data volume, velocity and variety.** Vol. 2, META Group Research Note, Vol. 6, 2001, p. 70
21. Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R. C., Roxburgh, C., and Byers, A. H. **Big Data: The next frontier for innovation,competition, and productivity.** McKinsey Global Institute, 2011
22. Manoochehri, M. **Data Just Right: Introduction to Large-Scale Data & Analytics.** Addison-Wesley Professional, 2013
23. Megahed, F. M., and Jones-Farmer, L. A. **Statistical Perspectives on "Big Data".** Frontiers in Statistical Quality Control, 11, 2015, pp. 29-47
24. Montgomery, D.C. **Introduction to Statistical quality control,** 7th edn.,Wiley, Hoboken, N.J., 2013
25. Murdoch, T. B., and Detsky, A. S. **The inevitable application of Big Data to health care.** Jama, Vol. 309, No. 13, 2013, pp. 1351-1352
26. Pappas, C. http://elearningindustry.com/Big-data-in-elearning-future-of-elearning-industry, 2014
27. Picciano, A. G. **The Evolution of Big Data and Learning Analytics in American Higher Education.** Journal of Asynchronous Learning Networks, Vol. 16, No. 3, 2012, pp. 9-20
28. Qiu, P. **Introduction to Statistical Process Control,** Boca Raton, FL: Chapman Hall/CRC, 2014
29. Raghupathi, W., and Raghupathi, V. **Big Data analytics in healthcare: promise and potential.** Health Information Science and Systems, Vol. 2, No. 1, 2014, p. 1

30. Reiss, C., Tumanov, A., Ganger, G. R., Katz, R. H., and Kozuch, M. A. **Heterogeneity and dynamicity of clouds at scale: Google trace analysis.** Proceedings of the Third ACM Symposium on Cloud Computing, ACM, 2012, p. 7

31. Rezzani, A. **Big Data: Architettura, tecnologie e metodo per l'utilizzo di grandi basi di dati.** Maggioli editore, 2013

32. Roski, J., Bo-Linn, G. W. and Andrews, T. A. **Creating value in health care through Big Data: opportunities and policy implications.** Health Affairs, Vol. 33, No. 7, 2014, pp. 1115-1122

33. Russom, P. **Big Data analytics.** TDWI Best Practices Report, Fourth Quarter, 2011, pp. 1-35

34. Sanchez, F. M., Gray, K., Bellazzi, R., and Lopez-Campos, G. **Exposome informatics: considerations for the design of future biomedical research information systems.** Journal of the American Medical Informatics Association, Vol. 21, No. 3, 2014, pp. 386-390

35. Scippacercola, S. **Metrics-based markov chains for web analytics.** Statistica & Applicazioni, Vol. 1, No. 12, 2012, pp. 55-66

36. Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S.B., Ferguson, R., Duval, E., Verbert, K. and Baker, R. S. J. D. **Open Learning Analytics: an integrated & modularized platform.** Doctoral dissertation, Open University Press, 2011

37. Slade, S., and Prinsloo, P. **Learning Analytics Ethical Issues and Dilemmas,** American Behavioral Scientist, Vol. 57, No. 10, 2013, pp. 1510-1529

38. Snijders, C., Matzat, U. and Reips, U. D. **Big Data: Big gaps of knowledge in the field of Internet.** International Journal of Internet Science, Vol. 7, 2012, pp. 1-5

39. Vaish G. **Getting Start with NoSQL,** Packt Publishing, 2013

40. Vasta, D. **Web Analytics,** Apogeo, Milano, 2009

41. Wright, A. H. **Genetic algorithms for real parameter optimization.** Foundations of genetic algorithms, Vol. 1, 1991, pp. 205-218

42. Zedlewski, J., Sobti, S., Garg, N., Zheng, F., Krishnamurthy, A. and Wang, R. Y. **Modeling Hard-Disk Power Consumption.** FAST, Vol. 3, 2003, pp. 217-230

43. Zong, W., and Wu, F. **The Challenge of Data Quality in the Big Data Age.** Journal of Xi'an Jiaotong University (Social Sciences), Vol. 33, No. 5, 2013, pp 38–43

[1] He is currently **Researcher in Statistics** and teaching staff member of the *Department of Economics and Statistics, University of Naples* "Federico II". Graduate in "Statistics and Economics Sciences", magna cum laude (Faculty of Economics – University of Palermo) he received his *PhD in Computational Statistics and Applications* from University of Naples "Federico II", Department of Mathematics and Statistics. During the PhD, he was *visiting scholar* of Prof. A.H. Money (Henley Management College). His research area encompasses the following topics: *Norm-p nonlinear regression - Multidimensional Data Analysis - Data Quality Control - Applications of Statistics in Medicine and in Justice*. Local component of the research groups funded by the University of Naples "Federico II" and co-financed by the relevant Ministry, he attended many Statistical Conferences organized by various national and international institutions, presenting numerous communications and papers.
Membership of "Società Italiana di Statistica", "International Association for Statistical Computing", "International Biometric Society" and "American Statistical Society", he gained over the years a considerable teaching experience as **Adjunct Professor** of "Statistics", "Probability", "Statistical Inference", "Medical Statistics", at various Italian Universities (Bologna, Naples "Federico II", Palermo, Catanzaro "Magna Graecia", Cosenza "University of Calabria", Messina, Catania). He is author of about eighty published works in Methodological and Applied Statistics.

[2] Associate **Professor of Statistics**, formerly **Professor of Information Processing Systems**, and he is a member of the teaching staff of the *Department of Economics, Management, and Institutions of the University of Naples* "Federico II". Graduate in Physics, he received the *Post-graduate in Theories and Techniques for the Use of Computers* from the Faculty of Engineering, University of Naples "Federico II". His research fields include *Multivariate Methods, Cluster Analysis, Business Intelligence and Decision Support Systems.*
He enjoys membership of the *American Statistical Society, Società Italiana di Statistica, CIRDIS, AICA* and of the *Working Group of the Italian Society of Statistic for Customer satisfaction and evaluation of services*.
He is Author of numerous publications in the Multivariate Statistical Analysis. He designed and developed models and algorithms for Decision Support Systems and for Ultra-Metrics.