Knowledge engineering

The 'Data Scientist'
A resource for the Intelligence

ANTONIO TETI

According to a survey made by the American EMC Corporations\(^1\) (EMC Data Scientist Study), in the years to come, the figure most sort by the companies and governmental organizations will be the Data Scientist - a person who is able to transform into "knowledge" the immense universe of information produced daily at world level. The value of this figure is already considered crucial for many businesses,….

The concept of knowledge engineering

The production and management of information is, perhaps, the subject which, more than any other, over the centuries, has fuelled discussion and conflict among human beings. Humanists, psychologists and scientists, such as Shannon, Weiner, von Bertalanffy and Turino have conducted, during these decades, important studies and research on the means of communication, which have generated significant discoveries in the sectors of mathematical logic, theory of information, cybernetics and the theory of the systems. With the advent and subsequent development of the electronic processors, information has assumed an ever increasing fundamental role in the life of man, thanks to the possibility of processing, filtering and crossing the unremitting quantity of data swallowed up by the media.

Therefore, the analysis and study of information, as a meaningful element on which to base previsions, analyses and decisions, becomes an activity of primary importance, irrespective of the sector or typology of organization in which the individual operates.

The breaking down of the information, the unceasing evolution of computer potentials and the development of the Internet network, have allowed us to reach the definition of a new scientific sector, in which we can study the integrations of knowledge, the use of the informatics systems for the processing and refining of the information to generate a system of advanced knowledge, usable for a multiplicity of purposes and objectives.

The advantage of the utilization of the informatics systems is fundamental to the possibility of making sophisticated elaborations of information which

\(^1\) www.emc.com/utilities/globalsiteselect.jhtml?checked=true
would, otherwise, require the use of extensive human resources and complex levels of specialization. The structured work of information engineering is incorporated within a well-defined scientific sector: knowledge engineering.

The term ‘knowledge engineering’, coined in 1977 by Edward Feigenbaum actually identifies that branch of artificial intelligence that concerns planning, realization and management of systems based on knowledge (KBS) or of informatics systems able to use the information contained in a Knowledge Base (KB) through automatic reasoning procedures. In the same scientific branch, it is possible to identify also the Expert Systems.

Usually, in the collective imagination, with the term “system based on knowledge” an informatics system is identified which is able to process and optimize the information contained in a knowledge database (information) utilizing specific computerized procedures of logical reasoning.

Nowadays, with the same term, we can identify those software applications realized not only to automate the analysis of the information, but also to assemble it through a logical type system based on semantics. In this way, without the constant work and intervention of a human operator, it is possible to construct and store the intelligence information containers, usable for multiple purposes.

In the book “Knowledge and the Flow of Information”, published in 1981, the philosopher, Frederick Irwin Dretske, known for his contributions to the study of the epistemology and philosophy of the mind, states that “Information is an asset able to produce knowledge: it is the vehicle of a signal that can trigger a learning process. Knowledge is belief produced from information”. From the statement of Dretske, it is possible to grasp two substantial aspects:

• information can produce knowledge;
• knowledge assumes the connotation of truth (belief).

In other words, through the analysis of information, it is possible to produce knowledge on a specific object/subject or event, which can assume, for the person who assimilates this information, a determining reference to establish actions, behaviour or methodologies to be adopted in order to reach a specific objective. The knowledge allows us to understand the reality that surrounds us, also to formulate predictions on the possible evolutions of the future. In function of this, the individual is enabled to operate in such a way as to

---

2 Edward Albert Feigenbaum, born in New Jersey in 1936, is a computer scientist, who has worked for many years in the artificial intelligence sector. Dubbed as the ‘father of the expert system’, created a PhD thesis at the Carnegie Mellon University on EPAM (Elementary Perceiver and Memorizer one of the first models of learning and memorization implemented as a computer programme.

3 Expert System. It is a branch of artificial intelligence and identifies a programme which attempts to reproduce the performances of one or more expert persons in a given sector.
make modifications on the reality so that effects are produced on future events. Therefore, the process of transforming information into knowledge represents a strategically very important step for anyone, both in establishing the correct perception of the surrounding reality and for the actions that can be exercised to modify advantageously the surrounding reality. Dretske himself identifies knowledge as information available for action. But only in function of the acquisition of ‘verified knowledge’ can the individual act in a rational manner and can, therefore, base his actions on real, tested and reasoned elements.

However, in view of the vast quantity of the available data, the individual can proceed to the intelligent analysis of the information only if he employs advanced informatics technologies and the Internet network.

It is also true that even though computers can handle enormous quantities of information, only a small part of it can be processed in an intelligent way, given the fact that the majority of the files contained in the memories of the processors is constituted by documents that use a natural language, which means that it is often impenetrable by the common software applications.

Therefore, it is apparent that computers although successful in solving many activities, in this case, they show their inability in the intelligent exploitation of the information in their possession. Therefore, to implement a system of information engineering, the human presence is again necessary, especially for those work flow processes, purposely studied for the identification of the useful phases usable for the realization of a system of cognitive awareness.

Types and sources of knowledge

Before proceeding to the analysis of the characteristics of a knowledge system, it is opportune to give some attention to the learning methodologies which contribute to the formation of the information container on which one is going to work.

A first distinction is made between ‘knowing what’ and ‘knowing how’, that is: between ‘what you know’ and ‘what you can do’. Actually, it was a British philosopher, Gilbert Ryle who coined these terms to distinguish what the individual knows, in terms of individual cultural and social learning, compared to their operative abilities acquired in different contexts and conditions.

It is possible to distinguish various modalities of acquisition of knowledge attributable to the cultural and social learning of the individual:

• **Knowledge based on concepts**: derives from the knowledge of possessed concepts which allow one to interpret reality. For example, if we think of an automobile, we imagine a vehicle that incorporates a motor or a propulsive mechanism capable of imparting a driving force;

• **Knowledge based on rules**: is based on the knowledge of regulations, more or less defined, which regulate the functioning of the entire planet. For example, if fuel runs out the vehicle ceases to function. Thus, as is obvious, the poisoning of a human being provokes death;

• **Knowledge based on the knowledge of facts**: is founded on the knowledge of information. For example, the individual knows the routes and times of the autobus, since he uses it to get to work every morning.

As far as **the sources of knowledge** is concerned, we know that these can come from **direct knowledge** (personal experience, information acquired directly from an individual); **from reasoning** (processing of much information which is processed and intersecting - elaboration of images, scenarios, films, discussions and so on); **from communication** (the use of languages, instruments and methods of communication to transfer information).

The sources serve, principally, to **gather** and **memorize** the information, but to be able to “work” it, it is indispensable to catalogue it according to schemes and criteria which must guarantee its maximum and immediate usage, to transform it into a **knowledge base**.

But the value of the knowledge is, above all, tied to the possibility of knowing how to interpret the reality in a correct way, to understand rapidly the situations and scenarios, to know how to interpret the facts of everyday life, also those which may seem irrelevant. This results in a capacity to predict the developments of the reality and the possible scenarios that may arise in contexts of major interest. The more reliable the predictions, the greater will be the ability to predict and improve the preventive actions to manage possible future events in an optimal fashion. Therefore, the preventive actions will translate into actions aimed at the modification of the reality, to produce advantageous effects.

**From the obtaining of the information to the Knowledge Engineering**

The term ‘**knowledge engineering**’ usually identifies that scientific sector dedicated to the integration of knowledge with the advanced informatics systems, to allow the solving of complex and sophisticated problems which require the support of personnel with particular levels of specialization. Nonetheless, within this branch of science, it is possible to identify a multiplicity of scientific veins, which go from the engineering of software to artificial intelli-
gence, from the intelligent use of the data bases to the utilization of the most advanced techniques of data mining, to then touch upon aspects of mathematical logic and disciplines that concern the cognitive sciences.

Since the beginning of the 70's, knowledge engineering has produced changes in many work places and, in particular, the industrial sector. Thanks to the informatics expansion, it has been possible to develop applications able to reproduce the work of specialized personnel in different sectors and activities, developing very high performance informatics systems, better known by the term 'expert system'. In the technological field, the term 'expert system' refers to a software application able to solve complex problems that can affect different sectors. For example, for the analyzing of enormous volumes of data; for the control or the diagnostics of gigantic industrial complexes or for plant design, it is possible to use these systems, entrusting the control and management of sizeable and complex work processes to a single hardware/software device. The substantial difference between the expert systems and the other applications used to facilitate the work of the human operator lies in the complexity of the technologies put into the field for these expert type systems, which are able to ensure all the logical steps (and automatic) which underlie the human decisions. In essence, they are able to manage, in absolute autonomy, the work of the human operator, using the accumulated historical information to improve the procedures utilized. It is precisely in this capacity that the elements of artificial intelligence can be recognized. Against this background, it is understandable how an expert system can also be adapted to the activities of Intelligence data.

It should be clarified that there are three elements in an expert system:

- **The knowledge base**: contains all the logical rules and procedures which serve the system for its functioning;

- **The inferential engine**: it is the core of the system and takes care of the management of the database and of the intelligent use of the data contained;

- **The user interface**: is the instrument that allows the user to interact with the system through a simplified software interface.

As already shown, in many cases the use of these systems has made it possible to facilitate and make simpler the automation processes of numerous complex activities. However, the applications of inference are not able to satisfy all the needs of the human operator, as in the case of the management of the information needs. This is due to the costs and complexity that surround the design and realization phase of an expert system, but also to the investment of human resource which can be particularly heavy in terms of cost.
Thanks to the advent of the Internet network, at the end of the 90’s, things became easier in terms of expert systems dedicated to the management of information. The possibility of creating enormous databases for the processing of data and information, to produce complete and structured information (knowledge) induced the international scientific community to concentrate studies and research on the development of information systems, based, above all, on the Web. At the same time, we see the development of a new scientific vein in which the research and implementation of knowledge engineering systems enables the centralization of data on the Network and the interoperability of the programmes and the Web technologies, leading to the development of an architectural model of intelligent search for information: thus, the semantic web is created.

Intelligence of information on the Network: the semantic web

The masses tend to consider the World Wide Web as a gigantic container of texts related, in some, to each other. The greatest characteristic of the Web is its universality, which makes it an extraordinary instrument of information, but at the same time, of extraordinary disinformation. Based on the functioning of hypertextual links, the Web permits access to a myriad of data and information, much of which can also be of little value or even useless for the navigators’ search of the Cyberspace. It is well to remember that the information entered on the Net is produced by worldwide users and, therefore, it is susceptible to personalization or conditioning of every kind. Therefore, a cybernaut cannot feel protected against false, altered or even misleading information.

For access to information in the cyberspace, we use the search engines; extraordinary instruments of access to knowledge which have revolutionized the whole concept of access to information. However, the irrepressible impulse of the individual to swallow up data and information can be driven by momentary needs and can also change over time, or it can be that the search is not precise and, consequently, the cognitive browsing may prove fallacious and completely useless.

The cybernauts seek information by probing the Web, basing their search on personal cognitive experience, which often leads to jumping from one portal to another, without well-defined search logic. But, above all, the ability of mnemonic recollection of words or key expressions is what considerably influences the searchers’ path.

For example, for a web navigator the ‘steered’ access to portals where he is sure to find the desired information, is not unusual, or the access to sites where the graphic aspects seem to offer guarantees in terms of availability of true and reliable information.
In any event, the basic instruments for the search of information of interest are always attributable to the same method: the use of specific words or key expressions. For these reasons, in 2001, Tim Berners-Lee hypothesized, through the use of the hypertextual pages HTML, the creation of applications able to interpret the content of the hypertextual documents for processing based on the semantic aspect. The idea was a winner and thus, the semantic web was created to search information on the Web, based not only on the possibility of intercepting particular key words included in the documents, but able to utilize specialized applications capable of constructing networks of relations and complex documental connections according to customizable logics.

Intelligent agents operate in the semantic web, applications able to understand the meaning of the texts present on the web sites, and to accompany the user, without useless deviations, towards the desired information. An intelligent agent must principally guarantee:

• the exact comprehension of the content of the documents present in the Network;

• the creation of guided search paths and the acquisition of information pertinent to the indications furnished by the user. The function of guide, which the agent must exercise, is essential to allow the cybernaut to reach only the information of interest;

• the logical navigation on the web portals to avoid accessing sites which do not contain the required information and to connect, in a structured manner, the data acquired during the navigation.

The semantic function, constructed on appropriate rules, must be based, essentially, on the meaning of the words in a particular context, of the sets of words related to a specific field of interest and of phrases and texts dealing with the same topics. Therefore, the semantic application must relate the linguistic expressions with what the content of such expressions ‘wants to mean’. For this reason, the most important process is represented by the correct interpretation of the meaning of the words within the documents.

It should be noted that what directs our mind in the search for the desired information, is the so-called domain of knowledge, i.e. the entirety of all the terms that our cerebral system connects to the desired information. The terms

---

6 HTML (Hyper Text Mark-up Language). It is the language used for the hypertextual documents available on the Web.
7 Semantics. It is that part of the language that studies the significance of the words, groups of words, phrases and documents. It closely relates to other scientific disciplines, such as semiology, logic, psychology, communications, the philosophy of language etc.
memorized in our mind are elaborated by cognitive processes which lead them to be linked to each other. For example, the terms, ‘car’, ‘automobile’, ‘vehicle’ or ‘van’ are united by the same association of meaning, and the instrument that allows us to make these cerebral connections is linguistic competence, on which the theory of communication by Noam Chomsky is based, in which he elaborated numerous and detailed studies.

For this reason, the search engines accessible on the Network are optimizing their functions to refine the users’ search (e.g. Google permits also visualization of the links that contain the synonyms of the word sought). More and more frequently, however, the ‘key words’ are not sufficient to guarantee an accurate and complete search.

For this reason, it is indispensable to focus one’s search on the ‘concepts’. The search model based on the concept is founded on the capacity of our brain to organize webs of word meanings, which contribute to the formation of geographical maps in which, for example, the cities are similar to the terms which we searching, and the streets correspond to connections with other terms related to the initial one.

These connections are explained by the concept of semantic proximity, for which a group of words and/or documents, in function of the use of special metrics, is similar in meaning or semantic content. These analogies between terms, which lead to knowledge, are known as semantic networks and permit simulation of the linguistic skills of the individual. A search system based on semantic networks is able to search and analyze all the documents present on the Web in which it is possible to identify the analogies with the real meaning of the terms sought. This method exploits the advantage of the adoption of a refined search system, based on concepts and domains of knowledge.

Nevertheless, even the semantic web is not completely devoid of ‘functional defects’ and, in some cases, may cause some problems in efficiency. Let us make an example: suppose we are conducting a search on the term ‘shot’ attributing to the word the concept of photograph (‘Did you see the shot of the dog in the swimming pool?’). The search engine makes a ‘rough’ analysis of the presence of documents that contain this word. The term also identifies a particular type of silk (shot silk), also the past participle of shoot (they were shot in the head), something that is ruined (the whole plan is shot), a brand name of jeans, the surname of a person etc. Therefore, at this point, we must add other terms that allow us to refine the search in progress. If we should then add other words, the list of the links proposed by the search engine would increase exponentially and we would be obliged to modify the semantic web application to further refine the search methods.

For this reason, the development of software applications that are capable of analyzing the Network is indispensable, thanks to the utilization of advan-

---

ced semantic networks that contain lexical databases that operate on a conceptual basis. These systems are able to correlate hundreds of thousands of terms, also of different languages, to understand unequivocally the meaning of the examined documents, acting, therefore, as reducers of ambiguities and conceptual anomalies of the examined texts.

These systems are based on the implementation of an ontological model, i.e. a complex conceptual scheme capable of understanding the criteria chosen by the individual for the relevancy of the documents sought by the system.

Obviously, a system based on an ontological model\(^9\), for its realization and subsequent use, necessitates a specific professionalism, with skills in multiple sectors and, preferably, possessing a substantial background experience. The human operator returns to take a decisive role in function of the high professionalism possessed.

The professional figure is that of the specialist of information, the only one able to understand the complex universe of the methods of the intelligent search of information, and the only one able to manage adequately the informatics applications capable of generating knowledge.

**The Data Scientist: an indispensable figure**

As we have understood, the valorisation of information, as a basis for the construction of knowledge, is a fundamental mission for any organization or company committed daily to compete in an ever-changing world.

Information is the cognitive fulcrum on which the decisional power of the individual is based, the correct interpretation of this information is essential. According to a study carried out in 2011, by IDC\(^{10}\) and commissioned by EMC\(^{11}\), in a matter of a few years, approximately 1.8 zettabytes of data will be created at planet level (a zettabyte is equivalent to 1000 billion gigabytes): a veritable “universe of information”.

This information overload will require, within 2020, all companies at world level, to expand their respective computer allocations for the memorization of data to an overall value of approximately ten times the present one.

The news has created considerable confusion among companies and orga-
nizations all over the world, which have admitted their total unpreparedness in facing this challenge.

The fears of Information Technology managers have been shown also by an investigation conducted by Gartner\textsuperscript{12} which underlined how the continual growth of information represents the most pressing problem for the IT infrastructure managers. Around 47% of those interviewed consider it among the top three most critical elements.

It should be emphasized that among the concerns most felt among the IT managers throughout the world is also the growing energy costs that must support the businesses to satisfy the energy ‘hunger’ needed by the modern Data Centers.

But what is more worrying for the Chief Information Officers is the unremitting data growth, which will make the ‘intelligent management’ of the information even more difficult. As a function of this, companies will be forced to enlist the aid of the data scientists, a strategic figure who will have the job of transforming this mare magnum of information into a knowledge provider. But what are the features and the responsibilities that must characterize a figure such as this? How can he/she best manage enormous and heterogeneous databases of information?

Among the various specializations, some are considered to be essential for this singular figure, in particular, the ability to:

- identify the best algorithms for the data mining operations;
- identify the most important criteria for analysis;
- develop new management and optimization of the data (data conditioning);
- manage, extrapolate, present and distribute the data and transform it into knowledge;
- identify new types of analytical databases in function of the type of data mining used;
- identify analysis instruments of the ‘high-end’ type, which are more predictive and usable by the organizations (e.g. for the prevention of fraud and to make prediction on the market trends and competition);
- identify the legal problems with regard to the treatment of confidential data or information protected by privacy;
- possess abilities in statistics, mathematics, calculation methods, calculation of probabilities and digital processing.

However, the professional abilities mentioned up to now must not be considered exhaustive. The data scientist, for example, must be able to carefully

\textsuperscript{12} Gartner Inc. is a multinational United States company specialized in analysis, research and events in the ITC sector. (www.gartner.com).

\textsuperscript{13} Information Technologies.
sift the information coming from different information sources, before deciding what data can be considered “useful” for his/her research. As well as to be able to preliminarily cross the data coming into his possession (originating from multiple sources), with particular attention to what comes from the Network, such as the social network, blogs, web servers or from the online registrations. He must also be able to manage more complex data (like geospatial data) and must be able to use more refined research algorithms capable of data mining immense databases containing terabytes of data, in a relatively short time; in addition he must be able to select the most suitable instrument of business intelligence, to perform the analyses required by his employees.

Even though it may seem ostensibly unusual, he must also possess a mentality which is oriented towards the arts and creativity, to enable him/her to elaborate visions on the methods of intelligent management of information and even on its possible use for purposes different from the original.

If we consider how the necessity for social media consultants has grown with the creation of the social network era, it will not be surprising if, in a few years, due to the imminent explosion of data, the data scientist will become the most sought after professional in the world. Thornton May, cultural and futurist anthropologist, actually describes this figure as ‘the hero of the future’.

On the surface, the multiple skills of the Data Scientist could seem “excessive” but, to the contrary, it is not so very difficult to identify those professionals who are in possession of these multiple skills. In fact, it should be noted that the data scientist must, above all, excel in certain characterial aspects, such as creativity, curiosity and determination in facing new and particularly complex situations. He should feel inspired by his organization and deployment of the information that is useful to the people for whom he works. His real mission is to transform data into value.

For a better understanding of the intersection level of expertise, we shall go to a classic diagram of Eulero-Venn\textsuperscript{14} (or simply, Venn) in which the different disciplines (or areas of competence) intersect each other, creating the ideal enclosure in which the data science is developed (Pic.1).

The different skills, when combined with others, can increase their value in terms of upgrading the information, but it is likewise true that they can also produce results which could be considered “dangerous”. In other words, the areas of competence must intersect in an intelligent manner, in such a way as to produce processes of information elaboration that is useful for the generation of knowledge.

The Danger Zone shown in Pic. 1. assumes a risk zone, which can derive from the commingling between personal experience and competence in the hacking sector. For example, it is not advisable to include an informatics ex-

---

\textsuperscript{14}The Eulero-Venn diagram is a diagram that graphically represents a set of elements enclosed within a closed line which does not intertwined with other. It may also be defined as a graphic used to represent algebra of sets.
pert who has a record of informatics crimes, in an analytical process of confidential information.

From the Data Scientist to the Intelligence Data Scientist

The ‘hunt’ for Data Scientists already began months ago. Multinational corporations, such as the American Express, since June 2012, have been concentrating their attention on the creation of a team of experts in the management of information. For this well-known U.S. company, which operates in the financial sector, the recruitment of these professionals seems even “crucial for the future” of the company. The top executives judge the work of these professionals as a central part of the “digital transformation” which will affect

the organization in the near future. The annual salary for this person, established at 160,000 U.S. dollars, is also indicative.

Also in Russia the attention given to the problem of handling “big data” has not been late in manifesting itself. In this case the announcement is of February 2012\(^\text{16}\) and concerns the opening of a new research center in which data analysis technologies will be developed with particular reference to certain sectors like bioinformatics and energy efficiency. The agreement was reached between EMC\(^\text{17}\) and Skolkovo Foundation\(^\text{18}\) and the center will be located in the Innovation Hub of the Muscovite Foundation. The center will develop methodologies of information analysis also for other sectors, such as molecular medicine, pharmaceutics, biomedicine, industrial biotechnologies, geopolitics etc.)

Also other Intelligence Agencies are aware of the necessity of introducing experts to the management of information.

‘Do you have a passion for creating data-driven solutions to the world’s most difficult problems? The CIA needs technically-savvy specialists to organize and interpret Big Data to inform US decision makers, drive successful operations, and shape CIA technology and resource investments’\(^\text{19}\).

This is the message that stands out on one of the web pages dedicated to the recruitment of personnel by the CIA portal (Central Intelligence Agency), in which all the different professionals sought by the Agency are listed. The message is clear: that also in the Intelligence sector, a Data Scientist is indispensible to interpret correctly the enormous databases which accumulate data on a daily basis.

Among the descriptions of the different characteristics required, the following passage is singular: ‘If you have experience in data analysis, informatics, mathematics, statistics, economics, operative research, calculation of the social sciences, quantitative finance, engineering or other fields of data analysis, consider a career as a data scientist with the CIA’.

Furthermore, it is possible to observe how the Agency attributes particular attention to the inherent and characterial abilities of this professional. ‘As a Data Scientist at the CIA, you will work with advanced software and hardware, with techniques to develop calculation algorithms and with statistical methods to find patterns and relationships for large volumes of data. In accordance with the global mission of the CIA, the Agency has access to particular sets of data which can be analyzed in a single computational environment. The chosen candidates will have sharp technical intuition, creativity, initiative and an inquiring mind. Data Scientists are expected to communicate their conclusions, in clear terms, to a heterogeneous public and to be-

\(^{16}\) www.01net.it/emc-dalla-russia-con-big-data/0,1254,1_ART_145976,00.html

\(^{17}\) www.emc.com/Utilities/globalsiteselect.jhtml?checked=true

\(^{18}\) www.sk.ru/en/Model/AboutFund.aspx

\(^{19}\) www.cia.gov/careers/opportunities/science-technology/data-scientist.html
come experts through continual training, participation in technical and academic conferences, and collaboration with the Intelligence Community’.

The Data Scientists are also expected to carry out a role of interfacing (for training-informative purposes), with technical and top executive structures of the Agency, without excluding the referents of other governmental structures. The work location is Washington DC, i.e. the power center of the Government. And this is certainly not a coincidence.

In 2011, the journalist, Kimberly Dozier of the Associated Press, during a visit conducted at the Open Source Center of the CIA20, has the opportunity to check some of the activities conducted by the first data scientists, recruited by the Agency. The first activity consists in the continual gathering of data and information produced in the Cyberspace in their mother language (i.e. in all the languages spoken all over the world), ranging from Arabic to Mandarin Chinese. Following this, one proceeds with the analysis of the gathered data: post suspects are investigated, angry users of tweet, e-mails containing strange phrases in clear documents, threats and insults in the social networks, anomalous or superficial blogs managed by improbably users. Then the information under attention is crossed with local newspapers, telephone conversations, photos, geostationary images, electronic post, sms, mms, etc., Subsequently, provision is made for the construction of ‘sought scenarios’, studies and useful reports for the highest levels of the White House, as in the case of the document in which was illustrated the state of mind of the people of the Pakistan region where the blitz occurred after the capture and killing of Osama Bin Laden.

In Dozier’s article, the statement of a CIA official, Doug Naquin is significant. He states how his team of analysts had predicted the revolt which occurred in Egypt last year, without, however, being able to predict the exact period in which it was to have developed. Nevertheless, the official assured that it was precisely in this direction that the data scientists are working, i.e. to be able to obtain predictive models able to establish even the moment in which a certain event will happen. In January 2012, also the FBI (Federal Bureau of Investigation) announces its search for data scientist21. The request of the Bureau for these professionals derives from the need to monitor the activity of the most popular social media used in the world (with particular attention to Facebook and Twitter) and to individuate software solutions able to improve the activities of analysis and intelligence. The objective is to realize a system capable of searching for information and useful data on the Net to inform the Authorities of possible future threats and new emerging risks which could influence internal scenarios of the Country. The system should allow a ‘stratification of data to correlate’, that it, the possibility of collecting together more local information (images from cameras arranged on the territory, maps of ro-

---

ads and motorways, sites of installations which are strategic and at risk, places where terrorist acts have been verified, location of people at risk etc.) as well as data of a predictive type, capable of elaborating forecasting models. These techniques of *data display* and *data integration* have already been used for some time by the Federal Authorities when they operate on particular difficult and complex cases. A typical example is constituted by the utilization of certain informatics applications which acquire and process information on fraud conducted in the USA, for the purpose of identifying the charactorial tendencies of the individuals who commit this type of crime.

Also the CHQ (Government Communications Headquarters), British governmental structure of intelligence, better known by the term British Intelligence Agency, is seeking these scientists\(^\text{22}\) who are, however, defined by a more glamorous name: Information Specialist. While retaining the specificities typical of a specialist of data, the Information Specialist, according to the British Agency, must guarantee knowledge in the following fields:

- Open Source Research
- Information Management
- Information Governance
- Information Risk Management
- Information Legislation
- Electronic Documentation and Records Management
- Intranet Management
- Technology Collaboration
- Knowledge Management

Identified as “*best practices*”, the possession of these skills, as well as for the Intelligence, is considered essential for the informatics infrastructure of any company.

If we move towards the East, following the route that conducts us to the discovery of those Countries that have understood the strategic importance of these new scientists, we are obliged to recognize that Nation which, perhaps, before any other, understood the importance of the contribution of the data scientist: China.

Above all, there are the companies, but also some governmental structures which are seeking *Chief Data Scientists and Senior Data Scientists*\(^\text{23}\). The fact is significant that they are seeking Chiefs of ‘groups’ of data scientists. An aspect which suggests that for some time they have had in operation ‘teams’ of professionals concentrated on the refining of information. And it is not surprising that since 2007, the Peoples’ Republic of China has had the *Center for Data* ...

\(^{22}\) [http://www.gchq-careers.co.uk/Jobs/Information-specialists.html](http://www.gchq-careers.co.uk/Jobs/Information-specialists.html)

\(^{23}\) [http://topic.csdn.net/u/20120830/17/99a40264-9541-4961-b970-c35e0a58c37c.html](http://topic.csdn.net/u/20120830/17/99a40264-9541-4961-b970-c35e0a58c37c.html)
Science and Dataology, a Center where research is done on theories, data mining, methods and technologies for the analysis of data in the Cyberspace. As can be understood from an interview granted by a professor and researcher of the structure, the Research Center focusses attention, above all, on the techniques of analysis of multi-sector data, such as finance, the economy, the insurances, bioinformatics and sociology. The Center is composed of seven units:

1. The Data Resource Service Office
2. Dataology and Data Science Research Lab
3. New Economy Development Strategy Research Lab
4. Bio-Medical Data Research Lab
5. Brain Informatics Research Group
6. Intelligent Transportation Data Research Group
7. Financial Data Research Group

All the personnel of the structure are rigidly selected in accordance with skills and the specifications of the job. From time to time, seminars and conferences are held (also at an international level) and the first monograph produced by its researchers, by the title ‘Dataology and Data Science’ goes back to 2009.

But the aspect which assumes the most relevance in the Shanghai structure is the variety and particularity of the fields of research. For example, studies are conducted on the techniques of data mining for sequencing of the gene; on the analysis of information; on the intelligence transport systems; on the reasons that lead to the creation of computer viruses and, even, on the psychological aspects that influence the stock market. In short, we are speaking about methods of analysis and intelligent processing of all the available information.

But the most important research project of the Center is the study of the founding theory of the science of data: the Theory of Data Science. According to Professor Yangyong Zhu, one of the heads of the Center, the term ‘data science’ identifies the science of the data in the Cyberspace and consists of two key dimensions: the first is to furnish a method of investigation that the researchers call Scientific Research Method with Data, usable for the natural and social sciences; the other is the research of phenomena and laws of nature data.

The latter term refers to the whole of the data available in the Cyberspace which reflects nature and human behaviour. In other words, it means identifying information judged to be tangible and credible from all of the other in-

---

24 http://datascience.fudan.edu.cn/s/98/t/316/main.htm
formation which does not have direct reference to real-life contexts (better identified as *rubbish data*). According to Professor Zhu, over the centuries two ‘data explosions’ have occurred. The first, the invention and fabrication of paper, the second, the invention of the computer and the Web.

Furthermore, he claims that the data mining and the analysis of data are only technologies relative to the treatment of data and cannot be considered a true “science”. In Data Science a strong emphasis is given, above all, to the management theories of data, rather than to the technologies utilized. Therefore, its mission is the study of the phenomena and the laws that govern the nature of the data. The Chinese research has led to the study of the ontology of the data, coining a new term: Dataology.

The basic areas of Dataology are:

- the founding theory of the science of data and Dataology;
- the methods of logical reasoning on the data and the experimentations;
- the theories and methods of operation in Dataology (ontology of the data applied to behaviour, to the human nervous system, to the motivations for the production of information etc.);
- the methods and technologies that use and exploit data as a resource.

The growing quantity of data can no longer be managed with the informatics systems and analysis methods which go back to a few decades ago. The storage of data in enormous databases, where applications structured for searches based on “key words” are working without interruption, no longer has any sense and very little real usefulness. New techniques are needed to handle the data in an intelligent manner. Certain structures which operate in the Intelligence are already developing new search theories and methods and intelligent treatment of information.

In a matter of a few years, we shall see the creation of research groups which will deal with the measuring of data, the algebra of data, the similarity of data, and algorithms will be developed which are capable of defining the accuracy of the information.

These innovations will produce considerable effects, especially in the Intelligence sector. At the present time there are already several software applications capable of using the semantic intelligence to exploit the advantages of the semantic search and of the natural language. Nevertheless, although very useful for the transformation of information into knowledge, it still does not fully satisfy the needs deriving from the intelligence activities, which are notoriously characterized by articulated and diverse information and which necessitate particularly complex intersecting and analysis.

For this reason, it is essential to proceed with the development of new algorithms able to expand the possible meanings and interpretations of the ac-
quired information, with the implementation of new rules and schemes of intersection of the data. In this respect, it is evident that the sector of the Intelligence that would be more concerned with the verticalized implementations on the search for information is the OSINT (Open Source Intelligence). As we have been able to see, the planning, the realization and the management of new systems of intelligent processing of information cannot prescind from the presence of persons who possess specific and complex skills. Therefore, the Intelligence Data Scientist assumes a fundamental role for any organization, but he becomes especially vital for the Intelligence Agencies, who must face the challenges of the near future in guaranteeing the internal and external security of their Countries. The forecast of a scenario projected towards the massive utilization of informatics technologies and the use of the Network, creates the preconditions for the configuration of a new scientific branch: Intelligence Data Science (Pic. 2).

As a consequence and in order to be prepared for the challenges of the near future, the Intelligence Agencies must, as soon as possible, activate proces-
ses of acquisition and training aimed at the inclusion of these new experts in the intelligent management of the data.

George Bernard Shaw said: “Science is always imperfect. Every time a problem is solved, at least ten new ones are created”. Perhaps it is not altogether false, but the important thing is to try not to meet them unprepared …

Bibliography


For further information, the author suggests…

What is Data Science
Author: Mike Loukides
Editor: O’Reilly Media, 2012

Chomsky’s Revolution in Linguistics
Author: John R. Searle

The entire or partial reproduction, reprint or quotation of the articles published hereby is not permitted without prior written consent of the editor