Big Data: at the Crossroad of Technology, Business and Regulation

Roberto Saracco, *Senior Member, IEEE*

Abstract – **The technological evolution in the last 20 years has steadily increased the amount of data in digital form, sensed from the environment, produced by private companies or collected from individuals' activities. Communications becoming more and more pervasive make these data available to any application, as if they were in a single database. The real value comes from information and services derived from data correlation, tailored to specific interest. We claim that the availability of a Living Open Data Framework enables the exploitation of data through services, created by many parties, in several contexts. However, data correlation is fraught with perils, from the obvious of privacy breaching to more subtle ownership and value protection. These issues are of fundamental importance and have to be addressed to create a viable business based on data and require advances in technology and in the regulatory framework. This is what is being pursued by the Italian EIT ICT Labs Trento Node through the cooperation of several parties, including the Autonomous Province of Trento, Telecom Italia, and several universities and SMEs. The paper aims at reporting the directions and the results so far obtained.**

Index Terms — **Data Analysis, Data Correlation, Open Data**

I. INTRODUCTION

Last year researchers at Stanford published a study on the analyses of millions of medical records of USA patients indicating that the use of two drugs, one for lowering cholesterol level and the other fighting depression, led to higher blood glucose levels [1]. Spotting this was quite easy, while determining this effect through clinical studies would have been almost impossible.

In fact, how could a pharmaceutical company clinically test all possible outcomes from all possible use of drugs? There are thousands of them and the possible permutations of concurrent use go up in the millions.

In a nutshell, this is the power of Big Data: their correlation and analyses generate valuable information.

If the potential is evident, the way to exploit it is not straightforward: on the one hand we have more and more data being produced but not all of them are readily available nor through secure channels. This is mostly due to a cultural resistance of data producers to share their data because so far data ownership has been considered a source of control: it is true for Telco companies as well as for single individuals. On the other hand the power provided by correlation can be misused, can affect privacy of individuals, the value created is often unbalanced among the data chain stakeholders. The latters have recently raised new concerns about data ownership and value redistribution.

Manuscript received 13 March, revised 24 March, accepted for publication 31 March.

R. Saracco is with EIT ICTLabs Trento Node, 38123 Trento, Italy.
E-mail: roberto.saracco@ictlabs.eu

Technology is progressing rapidly, both in the creation and in the correlation of data, with distributed computing enabling the processing of huge amount of data, unthankable just five years ago. Riding many positive forecasts, business is endeavoring to find ways to exploit the potential offered but regulation is lagging behind.

In the following we provide a roadmap to address Big Data, describing the context and the activities carried out in the Trentino area within the EIT ICT Labs framework.

II. IT'S BIG, AND IT'S GETTING BIGGER

The amount of data being created every day is staggering. Only in Italy we have exceeded 3.5 billion data records that can be harvested every day, including call data records, digital power meters, mailing and banking transactions, health care prescriptions, traffic monitoring, security camera feeds, ambient sensors, etc.

This amount is growing every day, as more and more sensors are deployed, individuals track their actions in a digital form, the fabric of commerce turns to digital transactions (mobile payment substituting cash), health care becomes heavily personalized and requires continuous monitoring, safety and energy concerns move the focus from vehicles to infrastructure and so on.

Our expectation is that by the end of this decade Italy alone will be producing 10 billion "accessible" transactions a day.

This figure may not still considered Big Data, if one compares it to the TB generated by the LHC for a single experiment. But looking at all the correlations generated over those multi-source data, the amount of data can be compared in computational complexity with the ones of the LHC or the KAT-7 Radio Telescope.

The following are some application scenarios we envision, where big data and correlation of multi-data-sources go hand in hand.

A. Our life in bits

My Life in Bits was a research program initiated by Microsoft a few years ago. It assumed that our individual life could be recorded in digital form. At that time the focus was on the creation of devices that would capture the various moments of our life and on the storing and retrieval of those moments.

Today, capturing in real time what we see, what we hear, what we do is no longer a challenge. Products like Pivot Head let us wear a pair of glasses that embed a camcorder, so tiny that you won't even notice it, at a cost that compares to the one of some cool sun-glasses. Data storage is no longer a challenge either.

Sorting out and extracting knowledge from the collected data: this is now the challenge that research is taking up.

Recording our life in bits create a sort of black box that can be used for a variety of purposes, from helping to remember, to increase the effectiveness of education and training, to increase safety.

But the records of our life can also be used as the starting point for correlating other data and creating information. You've got a running nose. An analyses of the data recorded 2 days ago, correlated with data from the place you have been (temperature, other people health status, ambient particulate,...) may pinpoint the probable cause and single out a normal (annoying) cold from a potential serious disease at a stage when prompt actions reduce risks (and cost).

B. The rise of digital transport infrastructures

Cars are becoming a bunch of computers on wheels: entertainment of passengers, fuel consumption optimization software, safety control systems, etc., making use of tens of sensors and computers that continuously monitor the vehicle and sometimes share the data with the outside world.

Actually, much more in terms of safety and energy consumption can be done if data are shared and correlated to create a digital transport infrastructure. The infrastructure can talk to the car navigator to get information on the car destination (eventually it may become compulsory to use the navigator and input a "drive plan" as pilot are requested to input a "flight plan") and based on information received from all cars on the road and road conditions (man at work, accidents, weather hazard...) can instruct the engine to drive the car at a certain speed. No need to hurry to find yourself in a queue sooner!

In the future cars (people) may pay a variable road toll (like some "pay as you go" profile that insurance broker already offer), with higher rates charged for specific itineraries and offer of cheaper solution on the fly. They may also opt for even higher traffic fees if they want to get priority routing....

By the way, being aware of how much a trip would cost you at that particular time of the day (taking into account driving time, fuel consumption and taxes) may change our driving behavior.

C. Understanding a living environment

Many of today's products are born with several sensors embedded. Be it a toaster (heath sensor), a television (light sensor) a lamp (proximity sensor), an Mp3 player (accelerator sensor), a digital camera (light, accelerometer sensors) ... and more are coming. These sensors are used for "local" action (like detection of a smiling face for a camera or the dusk for a street lamp) but they may, in principle, provide data to the world, via Internet. Data are coming from sensors in orchards to warn on the presence of harmful bugs, they come from tunnels if temperature of a passing car is too high, from switches if it gets too cold and ice may become a danger, from water reservoirs, rivers and lake warning for pollutants...

Through these data and through their correlation we are getting a better understanding of the environment, and we can start to take action sooner, decreasing overall cost, increasing safety and environment protection.

D. Just in time production

The digitalization of the supply and delivery chains, along with the flexibility offered by robotized production is changing the way we think about factories. The growing percentage of software in products further contributes to a shift in the production paradigm.

Monitoring sales leads to change the production in real time mostly for local productions and for some kind of goods. If something sells better in blue, then that is the color for the next batch. These kinds of information are usually inferred by a posteriori analysis of reports and surveys that companies produce starting from their data. "Just in time" production requires the synchronization of many processes, and in turns the correlation of many data. Just few years ago these data where the result of guessing, now they are hard data reflecting what is going on.

A more complex and more efficient analysis that correlated multiple data source such as the mobile payment, aired ads, blogs and tweets, mood derived from security cameras, etc. let companies have a deeper understand of customers' needs and preferences in real time. This provides plenty of "hints" to design marketing strategies in real-time.

E. Society Well being

Defining and understanding the well-being degree in a society is not easy because it is often the result of multiple factors. Although poverty is not a basic recipe for happiness, wealth is not a guarantee at all: European northern countries are considered leaders of good services and social support, although often they are plagued with a high suicide rate, an indicator of a general discomfort.

The availability of multi-source Big Data enables the studies of the factors and the correlation that affect the well-being.

The supporting actions are focused on people with specific disabilities or diseases e.g. Autisms (or more generally ASD). This is essential but for every person with ASD you should take into account two parents, an educator (or more), colleagues and so on. It is this fabric, or social network, that actually has to be considered and there is a need to discover the relevant social network and work on improving its well-being.

Discovering social networks and sensing the mood of a social network is now within the realm of possibilities and much information can be derived from data analyses.

III. TECHNOLOGY EVOLUTION

Technology and Big Data interplay in many different ways.

Collection. Sensors of many kind transform changes in the environment into digital data, that can be "stamped" with origin (localization) time, identity; actions, like paying with your phone, writing a prescription, generate data; photographing, filming generate data. All of this is getting easier and cheaper thanks to technology evolution.

Transport & Storage. Data have to be transported from the point of generation to the point of storage, and pervasive communications is making this ever more effective. Storage can, to a certain extent, be considered an integral part of the communications technology, particularly after the rising of cloud technologies.

Mining. Stored data need be retrieved and processed. Here, Data Mining and modeling technologies play the lion's share. Data mining had its first presence in the spot light in the nineties, in conjunction with artificial intelligence studies. Today, after some time in the twilight Data Mining again is taking the forefront with a radically different approach. It is no longer based on Artificial Intelligence but on statistical analyses. Clearly this is made possible by Big Data, and, in turns, stimulate the attention on Big Data.

Visualization. The data visualization plays a crucial role in our understanding the hidden semantics of Big Data. Availability of graphic displays, and in perspective of haptic interfaces, let us immerse in data, feel them and physically interact with them. A good part of these progresses is enabled by increased processing capacity and lower cost.

Control. Data, and the derived information, are fraught with potential mismanagement. Privacy is clearly at the forefront of concerns every time we deal with personal data. Correlation makes these concerns even bigger by circumventing normal safeguards, like anonymity. Ownership comes second. If data can generate value shouldn't their owner be entitled to some sort of compensation? Digital rights management didn't work very well for music, it is even more difficult in case of data management and correlation. What if the sources of data are public (freely available) and their correlation is able to generate significant value. Is this value to be reaped by the one who made the correlation or should it extend to the ones whose data made correlation possible?

Technology evolution helps addressing these issues, although it often creates new issues, since more sophistication in correlation worsen many of the above listed issues.

IV. TRENTINO OPEN LIVING DATA (TOLD) FRAMEWORK

As it emerges from the above statements, exploiting the potential of Big Data raises a lot of issues, that technology can help to fix, but alone it is not enough. What we need is a "framework" that, relaying on technology and innovative solutions for big data management, takes into account the related process innovation opportunities, as well as the stakeholders' role and constraints, the business potential and so on.

We talk about *Living* and *Open* data: Living data because they are real-time, geo-localized and Open data because free to use, reusable and distributable. For sure, technology evolves faster than the regulatory framework, and the differences among different frameworks are exploited by players in this field, particularly those that are subject to lower constraints.

The institutions of the Trentino Region have taken the decision of managing the territorial data in an open fashion and have created an Open Living Data Framework, not unique but surely one of the first to approach this issue and develop a "policy" on open data.

This goes along with the statement given on December 12th, 2011, by the European Vice President Neelie Kroes "The best way to get value from data is to give them away". Indeed, Open Public Data are expected to boost EU's economy by 40 billion euros each year and the EU is investing 100 million euros in the 2011-2013 period to fund research on data handling technologies. And these estimated are referring just to data owned by public administrations. Just imagine the boost to the economy if also private companies would make their data available.

This is actually the aim of the Open Living Data Framework set up by the Trento Region: take the lead to steer private companies to jump on the bandwagon of Open Data by releasing their own data to get back value from information and services deriving it from multi-source data correlation. This should overcome the general difficulty of other Open Data initiatives to get momentum due to a scarce involvement of private subjects, to the lack of a defined semantics of the available data and the need of common uniformed format of data exchange. Since Trento satisfies the requisites for this framework, the idea is to use this (relatively small) pilot area to apply and verify the approach and then extend it at Country level.

The regulation provides the legal framework onto which data can be opened, and in parallel funding is given to research technologies to solve issues related to the opening of data. The Telecom Italia research lab, SKIL, is part of this global approach.

At the time of writing, April 2012, a first core of Companies have joined this initiative making available (part of) their data. Among them Telecom Italia, providing billions of anonymized call data records every month, the local highway management company, providing in and out of vehicles through its highway system, the local public transportation authority, providing hundreds of thousands of trip data per month, the Autonomous Province Public Authority, providing ambient sensors data, environmental geo data and safety related statistical data, the Area Electric Power Company, providing individual smart metering data every day, the National Postal Service, providing money transaction and goods delivery data.. The first result has been achieved: having the commitment of private entities to share their data within this initiative means having broken the first cultural reluctance that made most of similar

Data are going to be collected through several channels, each provided in its own format and subject to the constraints decided by the owner. That makes them difficult to use and requires the system to bring them in a common shared format.

In the Open Living Data framework SKIL [2], the aforementioned Telecom Italia Research Lab focused on semantic annotation and knowledge extraction, will be delegated to convert the raw data stream into a semantically annotated data stream and to correlate the data streams. The generated correlations are made accessible by the open API, thus decoupling the original raw data from the actual use.
This is one of the tricks being used to neutralize data, not just at the point of origin (this is done by the data provider) but also at the point of usage. This neutralization is a fundamental part of the Open Living Data Framework.

Correlation is also based on other nation-wide data provided by ISTAT, the Italian Institute of Statistics, a government organization.

The Trentino Open Living Data Project is moving the first steps towards the direction of exploring how to get value from multi-source Big Data and how this can be leveraged by the different stakeholders. Such kind of initiatives need to be "socialized" with the Territory, in order to let the citizens provide feedbacks and be familiar with the importance of open data. We planned to achieve this by visualizing the results of a number of correlations on big screens in places where people are likely to roam.

Example visualizations are thematic maps, such as those on car accidents, where people can easily detect the most dangerous areas, understand the underlying factors and eventually see the actions taken by the public administration to decrease the risk.

Other maps could show cultivated areas, their yield and the increase in productivity made possible by understanding negative factors impacting the products; others showing the spikes of some diseases and correlation points out probable causes.

V. Creating Business Opportunities

The framework provided by Trentino Open Living Data will be proposed as the starting point to generate new business through SMEs (Small Medium Enterprises). This is where the EIT ICT Labs [3], of which Trento is one of the six nodes.

The ICT Labs is a European Initiatives aiming at stimulating European competitiveness by fostering innovation and helping SMEs to put successful products into market. This is done using in an auto-sustaining way the tools of education, research and business innovation.

More specifically, in the context of Big Data, there is a need for research to create the required tools for data neutralization, security, privacy, authentication, accountability and value tracking. There is a need to support SMEs to manipulate those date and create services that are conforming to European and local regulation, to submit patent and to commercialize the product. This latter is usually a component in a complex network and through its wide reach ICT Labs can be a winning partner for these enterprises.

Finally, there is a need for high level education in this area, engineers need to understand what wealth can be leveraged out of Big Data and what are the issues related to that.

It is often voiced the concern of an imbalanced situation once comparing the strict rules usually imposed on European Companies when dealing with data, compared with the much looser ones imposed in other world areas. This generates, it is claimed, unfair competition. However, the more stringent rules applied in Europe can be turned to market competitive advantage, since they provide a context for better security and trust and these have commercial value.

SMEs can work at different level in the Open Living Data Value Chain; actually they usually operate at the Open Data Ecosystem level, by acting independent of one another and all together increasing the overall value and biz opportunities.

Data are pervasive and support many application areas, from Digital Cities to Smart Spaces, from Health Care and Well Being to Intelligent Transportation Systems. These are all areas addressed in the EIT ICT Labs. Furthermore, the area Security, Privacy and Trust is right on the spot to help tackling and leveraging Big Data, acting as a catalyst to speed up innovation in the afore mentioned areas.

Innovation is expected both in correlation (finding meaning in the data) and in the presentation of correlation result (visualization on a variety of media, big screens, cell phones, television, kiosks...) and each innovation is basically a service that can generate an attractive biz for an SME.

VI. Conclusions

We have presented a case for working on Big Data, by saying that technology evolution has already provided us with huge amount of data and this amount will increase in the future. Technology is also providing us means to deal with correlation, and this is the place where value is generated, and to deal with issues connected to data, like privacy, ownership, accountability.

We have also indicated that a strong steering from an institutional and political entity is required and this is what is happening in the Trentino Region.

Furthermore, the Big Data is a platform to sustain a new economy where the wealth is mined from data analyses and usage, a wealth that requires both a "data infrastructure" usually provided by big companies, and many SMEs creating and delivering services.

What we did not say, and we are now doing, is that Big Data leverage results in better life for each of us, for the citizen of the Information Society.

Correlation makes us aware of phenomena that are normally hidden, explains behavior in terms of causes and let us take corrective actions when needed.

The Information Society delivers at a much lower cost than the Industrial Society (that was in turn delivering at a lower cost than the agricultural and artisan society). This lower cost means lower barriers to enter business, hence more people entering the business as entrepreneurs.

We need to respond to these opportunities by preparing our students to become entrepreneurs, education is a fundamental stepping stone. Research is another fundamental building block, particularly research that can act as catalyst for the development of business.

EIT ICT Labs ambition is to be a main player in this context. It is in its infancy, since it has started in 2011 (the Trento node started on January 1st, 2012) so the jury is still out. For sure the commitment and enthusiasm is very high.

Acknowledgement

The contribution of ideas and text revision by Fabrizio Antonelli, Director of the SKIL Lab is gratefully acknowledged.

References

[1] http://www.technologyreview.com/business/38484/?nlid=nldly&nld=2011-09-19
[2] http://skil.telecomitalia.com/
[3] http://eitictlabs.eu

Open Data:
at the Crossroad of Technology, Business and Regulation

Roberto Saracco is the President of EIT Italy (European Institute for Innovation and Technology) and Node Director of EIT ICTLabs Italy, based in Trento, one of the 6 European Labs set up by the European Commission to transform research results in the Information and Telecommunication areas into market innovation. Up to December 2011 he was the Director of the Future Centre in Venice, responsible for innovative telecommunications architectures and scientific communications reporting directly to the Strategy Officer of Telecom Italia. Roberto chaired the Visionary Group (1996-1997) on Super Intelligent Networks to steer the co-operative research at the European Union (EU) level beyond the year 2000. He has recently served as member of the Internet 2020 Strategy Group and European Research Network (GEANT) expert group. In the eighties, Roberto led research in Telecommunications Management in CSELT, and actively participated in standardization activities at CCITT, and in a number of international standardization organizations including OSI, ETSI and T1M1. His leadership includes chairing an EU-level group for planning, leading European research activities in the area of software technologies, and the EURESCOM group in designing the framework for European co-operation on TMN. He has published over 100 papers in journals and magazines, six books – including "The Disappearance of Telecommunications," which was published in the USA by IEEE press, He has also delivered speeches and keynotes at many international conferences. He currently lectures at the Turin Polytechnic on the aspects of multimedia and telecommunications.

He is a senior member of IEEE that he joined over 20 years ago. In the last 15 years he has held several leading roles and conducted a number of DLTs and DSPs. Currently he is the Director of the Sister and Related Societies of COMSOC. Previously he has served as VP of Member Relations, Director of Marketing, Chair of the CNOM and Enterprise Management TCs.

Call for Papers – Special Issue on Quantum Communications

Although quantum computers are going to be the applications of the far future, there are already a few algorithms to solve problems which are very difficult to handle with traditional computers. Quantum computing is based on various quantum effects in physics and offers revolutionary solutions for different problems e.g., prime factorization, searching in unsorted database, key distribution and information coding. The power of quantum parallelism allows us to solve classically complex problems, and the quantum entanglement leads to quantum communication algorithms like teleportation and super dense coding. The quantum cryptography provides new ways to transmit information with unconditional security by using different quantum key distribution protocols. The special issue of Infocommunications Journal will focus on quantum communications with the following scope:

- Quantum Communications
- Quantum Interferometry and Quantum Sensors
- Entanglement as a Resource of Quantum Technology
- Quantum Cryptography
- Quantum Processors and Computers Design
- Quantum Programming Languages and Semantics

Authors are requested to send their manuscripts via electronic mail to László Bacsárdi at *bacsardi@hit.bme.hu* until 30 September, 2012. The manuscripts should follow the IEEE format with maximum length of 8 journal pages.

Guest Editors:

OZGUR B. AKAN (M'00–SM'07) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University and Middle East Technical University, Ankara, Turkey, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, in 2004.

He is currently a Professor with the Department of Electrical and Electronics Engineering, and the Director of Next-generation and Wireless Communications Laboratory (NWCL), Koc University, Istanbul, Turkey. His current research interests include wireless communications, acoustic communications, nano communications, quantum communications and information theory. Dr. Akan is an Associate Editor for the IEEE Transactions on Vehicular Technology, the International Journal of Communication Systems (Wiley), the European Transactions on Telecommunications, and the Nano Communication Networks Journal (Elsevier). He served as an Editor for ACM/Springer Wireless Networks (WINET) Journal from 2004 to 2010, as an Area Editor for AD HOC Networks Journal (Elsevier) from 2004 to 2008, as a Guest Editor for several special issues. He currently serves as the General Co-Chair for ACM MobiCom 2012, General Co-Chair for IEEE MoNaCom 2012, and TPC Co-Chair for IEEE ISCC 2012. He is the Vice President of the IEEE Communications Society – Turkey Section. He is a Senior Member of the IEEE Communications Society (COMSOC), and a member of ACM. He is a COMSOC Distinguished Lecturer (2011–2012).

He received the IEEE COMSOC Outstanding Young Researcher Award for EMEA Region 2010 (as runner-up), the IBM Faculty Award twice in 2010 and 2008, and the Turkish Academy of Sciences Distinguished Young Scientist Award 2008 (TUBA-GEBIP).

LÁSZLÓ BACSÁRDI obtained his MSc degree in computer engineering from Budapest University of Technology and Economics (BME) in 2006. He holds an assistant professor position at the University of West Hungary. He wrote his PhD Thesis on the possible connection between space communications and quantum communications at the BME Department of Telecommunications. His current research interests are in mobile ad hoc communication, quantum computing and quantum communications. He is Secretary General of the Hungarian Astronautical Society (MANT), which is the oldest Hungarian non-profit space association founded in 1956. He is member of the board of a Hungarian scientific newspaper ('World of Nature') and he is the publisher of a non-profit Hungarian space news portal ('Space World'). Furthermore he is member of IEEE and HTE. He has joined the Space Generation Advisory Council (SGAC) as well, and is currently active as the Hungarian National Point of Contact.

SÁNDOR IMRE was born in Budapest in 1969. He received the M.Sc. degree in Electrical Engineering from the Budapest University of Technology (BUTE) in 1993. Next he started his Ph.D. studies at BUTE and obtained Dr.Univ. degree in 1996, Ph.D. degree in 1999 and DSc degree in 2007. Currently he is carrying his activities as a Professor and a Head of Department of Telecommunications at BUTE. He is a member of Telecommunication Systems Committee of the Hungarian Academy of Sciences. He participates in the Editorial Board of two journals: Infocommunications Journal and Hungarian Telecommunications. He was invited to join the Mobile Innovation Centre as R&D director in 2005. His research interests includes mobile and wireless systems. His main research interests and contributions are in the areas of various wireless access technologies, mobility protocols and reconfigurable systems.