Big data for the social good

By Professor Dino Pedreschi, Fosca Giannotti, Valerio Grossi, and Roberto Trasarti – University of Pisa (Department of Computer Science), and Institute of Science and Technology of Informatics 'A. Faedo'

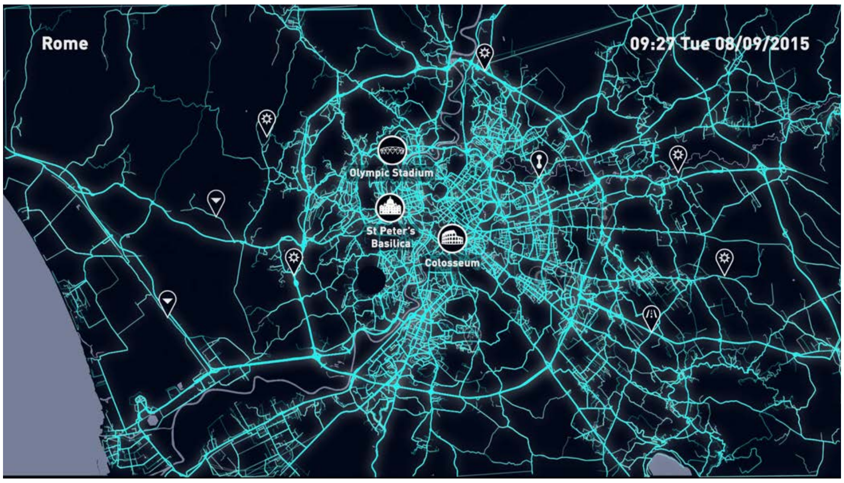
Source: Shutterstock / Cyber Kristiyan

Most people will agree that during the last two decades data have become an ever more present part of our lives, whether it is in our communication, our analysis or decision-making. Dino Pedreschi is Professor of Computer Science at the University of Pisa and a pioneering scientist in mobility data mining, social network mining and privacy-preserving data mining. Together with his colleagues Fosca Giannotti, Valerio Grossi and Roberto Trasarti – respectively director and researchers at CNR in Pisa – he gives an overview of some key issues relating to data. The authors consider the advent of big data is an opportunity for boosting social progress, and Artificial Intelligence tools are triggering new services with a clear impact on our daily life. They also touch upon how big data and AI help us to make more informed choices, underlining the need to achieve collective intelligence without compromising the rights of individuals.

From data to knowledge

In order to understand the role of data in our society, let's start by talking about the word *datum*. It comes from the Latin for 'something given,' and the Oxford English Dictionary defines it as a 'piece of information,' 'as a fact'. Our life is a generator of facts that can be stored and analysed. *Human interactions* leave traces of our phone calls or emails in our social networks. *Our lifestyle* is evidenced by records of purchases, while our movements are described in the records of our mobile phone and GPS tracks (see **Figure 1**).

Figure 1 – Mobility traces in Rome



Source: KDDLab & OctoTelematics

Artificial Intelligence (AI) is becoming an integral part of all areas of our daily lives, from smartphones and smart watches to personal digital assistants such as Amazon Echo and Google Home, to autonomous vehicles, smart cities, Industry 4.0, and beyond.

Transforming data into value; into knowledge it is not an easy task. On the contrary, it requires an interdisciplinary and pervasive paradigm where theories, models and artificial intelligence tools support each other. Experiments and analyses across massive datasets are essential, not only to the validation of existing theories and models, but also to the data-driven discovery of patterns emerging from data, which can help scientists design better theories and models, yielding a deeper understanding of the complexity of social, economic, biological, technological, cultural and natural phenomena.

In this context, a new revolution has happened due to three concurrent developments arising from the digital transformation of society:

- the advent of Big Data, which provide the critical mass of factual examples to learn from;
- the advances in data analysis and AI techniques that can produce predictive models and behavioral patterns from big data;
- the advances in scalable high-performance computing infrastructures that make it possible to ingest and manage big data and perform analytics.

The availability of data creates opportunities but also new risks. The use of data science techniques could expose sensitive traits of individual persons and invade their privacy. In particular, social mining approaches require access to digital records of personal activities that contain potentially sensitive information.

Depending on the course that this revolution takes, AI will either empower our ability to make more informed choices or reduce human autonomy. It will expand the human experience or replace it; create new forms of human activity or make existing jobs redundant; help distribute well-being for many or increase the concentration of power and wealth in the hands of a few; expand or endanger democracy in our societies. Europe bears the responsibility of shaping the AI revolution.

The choices we face today are related to fundamental ethical issues about the impact of AI on society; in particular, how it affects labour, social interaction, healthcare, privacy and fairness. The right direction cannot be found through a random selection of possible alternatives but only by a lucid public with renewed confidence in (scientific) research and technology.

Big data for science, industry and social good

The advent of big data is an actual opportunity to improve our society, to boost social progress, and the social good. It can support policy making, can offer novel ways to produce high-quality and high-precision statistical information, empower citizens with self-awareness tools, and it promotes ethical uses of big data. Several examples below show why big data are important for society.

Modern *cities* are perfect environments, densely traversed by large data flows. Using traffic monitoring systems, environmental sensors, GPS individual traces and social information, we can reorganise cities as a collective sharing of resources that need to be optimised, continuously monitored and promptly adjusted when needed. It is easy to understand the potentiality of big data exploitation by considering terms such as urban planning, public transportation, reduction of energy consumption, ecological sustainability, safety and management of mass events.

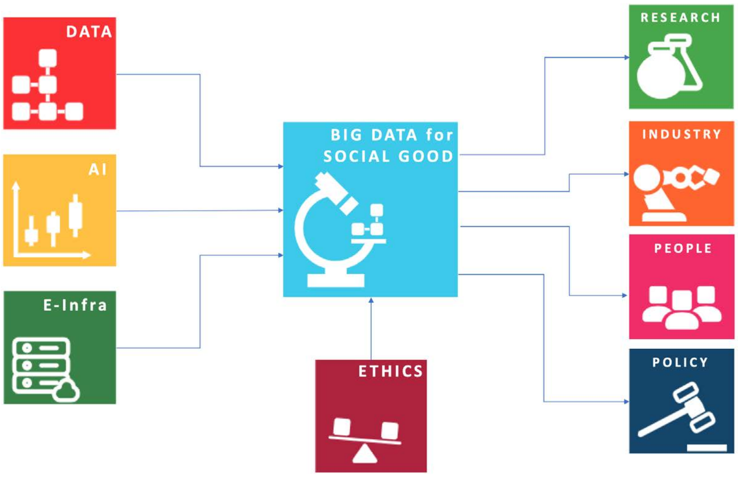
In the *biological sciences*, scientific data are stored in public repositories for use by other scientists. There is an entire discipline of bioinformatics that is devoted to the analysis of such data, e.g. network-based approaches to human disease can have multiple biological and clinical applications, especially in revealing the mechanisms behind complex diseases.

In *energy* and the *environment*, the digitisation of energy systems allows the acquisition of real-time, high-resolution data. Coupled with other data sources, such as weather data, usage patterns and market data, efficiency levels can be increased immensely.

In *manufacturing* and *production* with growing investments in Industry 4.0 and smart factories with sensor-equipped machinery that are both intelligent and networked (see internet of things, cyber-physical systems), in 2020 production sectors will be some of the major producers of data. The application of data science in this sector will bring efficiency gains and predictive maintenance.

These represent only the front line of topics that can benefit from an awareness of the gains that big data, in principle, might provide to stakeholders. As shown in **Figure 2**, scientific, technological and socio-economic factors play and interact, creating a complex ecosystem. The combination of data availability, sophisticated AI techniques, and scalable infrastructures is changing the way we do business, socialise, conduct research, and govern society.

Figure 2 – Impact of data, AI and E-infrastructures impacting key societal issues



Source: KDDLAb at University of Pisa & CNR

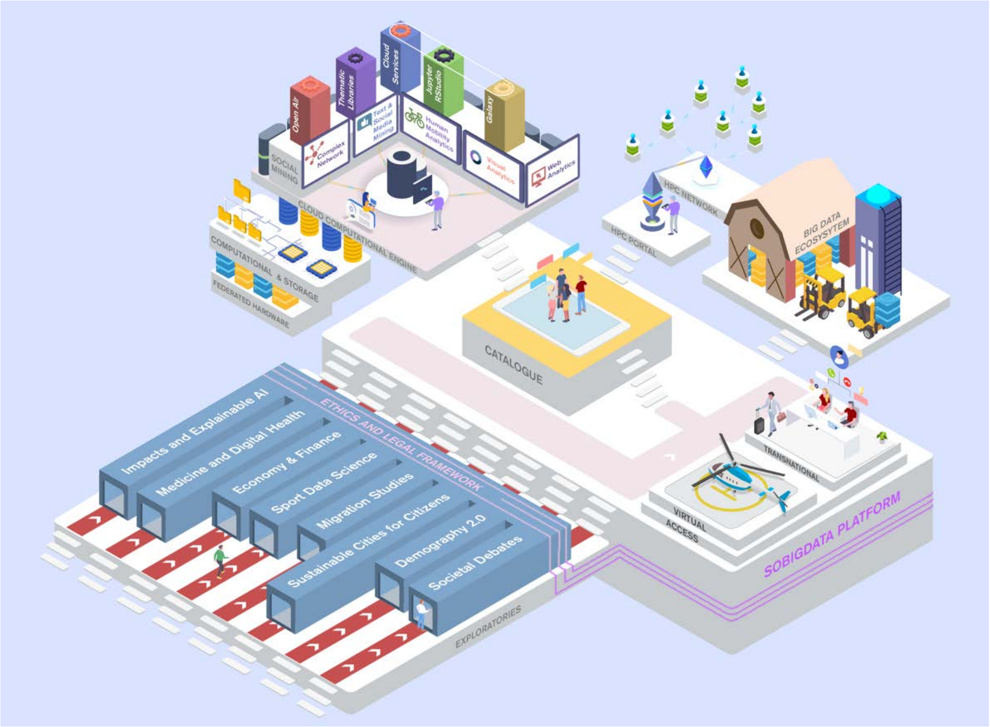
If we want to exploit data in order to face global challenges and make data a determinant factor in sustainable development and the social good, it is necessary to push towards an open global ecosystem for science and industrial and societal innovation, addressing multiple dimensions with interdisciplinary approaches. We need to build an ecosystem of socio-economic activities, where each new idea, product and service creates opportunities for further ideas, products and services. An open data strategy, a 'new deal on data,' open innovation, interoperability and suitable intellectual property rights can catalyse such an ecosystem and boost economic growth and sustainable development. This also requires 'networked thinking' and a participatory, inclusive approach.

Big Data ecosystem: the role of research infrastructures

Over the past decade Europe has developed world-leading expertise in building and operating e-infrastructures. They are large scale, federated and distributed via online research environments. They are meant to support unprecedented scales of international collaboration in science, both within and across disciplines, investing in economy-of-scale and common behaviour, policies, best practices, and standards. They shape a common environment where scientists can create and share their digital results, such as research data and research methods, by using a common 'digital laboratory' consisting of agreed-on services and tools.

Research infrastructures (RIs) play a key role in the advent and development of big data analytical tools for society. Resources - such as data and methods - help domain and data scientists in transforming a research or innovation question into a responsible data-driven analytical process. RI platforms offer easy-to-use means to define complex analytical processes and workflows, bridging the gap between domain experts and analytical technology. Well-defined thematic environments amplify new experiments' achievements towards the vertical scientific communities and potential stakeholders by activating dissemination channels to society.

Figure 3 – The SoBigData Research Infrastructure



Source: SoBigData Project

In this context, the 'SoBigData Research Infrastructure'[1] is an ecosystem of human and digital resources, comprising data scientists, analytics and processes. It is designed to enable multidisciplinary scientists and innovators to carry out experiments and to make them reusable by the community (see **Figure 3**).

The e-infrastructure of SoBigData provides researchers and practitioners with a uniform working environment where open science practices are transparently promoted; AI and data science practices can be implemented by minimising the technological integration cost. SoBigData provides users with domain-specific resources, and core services supporting data analysis and collaboration among users. It provides:

- a *shared workspace* to store and organise any version of a research artefact;
- a *social networking area* to have discussions on any topic and be informed on happenings;
- an *analytics platform* to execute processing tasks;
- a *catalogue-based publishing platform* to make the existence of a certain artefact public and disseminate this information.

SoBigData promotes and adopts ethically grounded collection, management and analysis of big data. Privacy enhancing tools ensuring that RI analyses are also fair and non-discriminatory. Scientists and other stakeholders, including citizens and policy makers, can access and use such facilities continuously and transparently, facilitating publishing of science according to Open Science principles of transparency and reproducibility.

[1] SoBigData is EU-Funded under the program Horizon 2020: Research and Innovation - Grant Agreement No. 654024 - http://www.sobigdata.eu/

Individual and collective intelligence

Social dilemmas occur when there is a conflict between individual and public interest. Such problems also appear in the ecosystem of distributed AI and humans, with additional difficulties due, on the one hand, to the relative rigidity of the trained AI system and the necessity of achieving social benefit, and, on the other, the necessity of keeping individuals interested. What are the principles and solutions for individual versus social optimisation using AI, and how can an optimum balance be achieved? The answer is still open, but these complex systems have to work on fulfilling collective goals, and requirements, with the challenge that requirements change over time, and change from one context to another.

Every AI system should operate within an ethical and social framework in understandable, verifiable and justifiable ways. Such systems must in any case operate within the bounds of the rule of law, incorporating protection of fundamental rights into the AI infrastructure. In other words, the challenge is to develop mechanisms that will result in the system converging to an equilibrium that complies with European values and social objectives (e.g. social inclusion) but without unnecessary losses in efficiency.

Interestingly, AI can play a vital role in enhancing desirable behaviours in the system, e.g., by supporting coordination and cooperation that is, more often than not, crucial to achieving any meaningful improvements. Our ultimate goal is to build the blueprint of a socio-technical system in which AI not only cooperates with humans but, if necessary, helps them to learn how to cooperate, as well as other desirable behaviours. In this context, it is also important to understand how to achieve robustness of the human and AI ecosystems in respect of various types of malicious behaviour, such as abuse of power and exploitation of AI technical weaknesses.

We conclude by paraphrasing Stephen Hawking in his *Brief Answers to the Big Questions*: the availability of data on its own will not take humanity to the future, but its intelligent and creative use will .