Issue 3 / August 2019

Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Magazine

Editorial

The vision that led to the birth of SoBigData, in 2014, anticipated the rising demand for cross-disciplinary research and innovation on the multiple aspects of social complexity from combined data-driven and model-driven perspectives. SoBigData's vision in 2014 also predicted the rising importance of ethics and data scientists' responsibility as a pillar of trustworthy use of big data and analytical technology. SoBigData's initial vision has become today part of the mainstream discourse and may be summarised as follows.

• The necessary starting point to tackle the challenges is to observe how our society works, and the **big data originating from the digital breadcrumbs of human activities offer a huge opportunity to scrutinize the ground truth of individual and collective behaviour at an unprecedented detail and at a global scale.** This increasing wealth of data is a chance to understand social complexity, provided we can rely on **social mining, i.e., adequate means for accessing big social data together with models for extracting knowledge from them.**

[Continues on pag. 3]

Inside this issue

Content

Editorial

From SoBigData to Sobigdata-PlusPlus: toward an advanced community of data scientists

Fosca Giannotti, SoBigData Coordinator, ISTI-CNR, Italy

[continued]

• There is an urgency to thoroughly exploit this opportunity for scientific advancement and social good as currently the predominant exploitation of Big Data revolves around either commercial purposes (such as profiling and behavioural advertising) or – worse – social control and surveillance. **The main obstacle towards the exploitation of Big Data for scientific advancement and social good – besides the scarcity of data scientists – is the absence of a large-scale, open ecosystem where Big Data and social mining research can be carried out.**

• There is an urgency to develop strategies that allow the coexistence between the protection of personal information and fundamental human rights together with the safe usage of information for scientific purposes by different stakeholders with diverse levels of knowledge and needs. **There is a need to democratise the benefits of data science and Big Data within an ethical responsibility framework that harmonizes individual rights and collective interest.**

Thus, SoBigData was designed to promote large-scale, interdisciplinary social data mining which is both repeatable and open-science oriented, based on three pillars:

1. A continuously growing, distributed data ecosystem for procurement, access and curation of big social data within an ethic-sensitive context. This ecosystem is based on innovative strategies to acquire social big data for research purposes, using both opportunistic means provided by social sensing technologies and participatory means based on user involvement as prosumers of social data and knowledge.

2. A continuously growing, distributed platform of interoperable social data mining tools, methodologies and services for mining, analysing, and visualising massive datasets, together with associated data scientists' skills for the ethically safe deployment of big data analytics.

3. A 'social mining' community comprising scientific, industrial and third-party stakeholders, such as policymakers, supported by joint research, transnational and virtual access activities, as well as extensive networking and innovation actions (workshops, summer schools, datathons, training resources, knowledge transfer and industrial partnerships).

> After three and a half years since its inception, the approach developed by SoBigData has proven successful in creating a comprehensive **e-infrastructure**, which offers access to over **180 resources** designed to support data scientists in the execution of large scale experiments.
>
> *Fosca Giannotti*

SoBigData has integrated 12 key excellence centres at a European level (based in 7 different countries) in Big Data analytics and social mining during its starting phase, from September 2015 to August 2019. This integration was pursued to create a distributed, networked Research Infrastructure which leveraged each partners' scientific resources – such as Big Data, analytical tools and services, skills – into six thematic clusters: text and social media mining [TSMM]; social network analysis [SNA]; human mobility analytics [HMA]; web analytics [WA]; visual analytics [VA] and social data [SD]. This activity was a fundamental element in developing cutting-edge social mining experiments.

SoBigData has bootstrapped a unique and vibrant platform for open, ethically-minded social mining research and innovation that comprises a resource catalogue, storage and curation primitives for methods and datasets organised through the metaphor of Virtual Research Environments. Moreover, SoBigData has enforced Virtual Research Environments' privacy, confidentiality and security requirements.

SoBigData has made a relevant design choice for the creation of its e-infrastructure and community introducing an initial set of five ***exploratories***, i.e. vertical social mining research environments focused on broad societal challenges. The five exploratories, **City of Citizens, Societal Debates, Well-being and Economic Performance, Migration Studies, Sport Science,** were selected as showcase stories from a number of options at the beginning of SoBigData. The aim was to exemplify the cutting-edge, multidisciplinary research supported by the project and drive the integration into the e-infrastructure of the resources available within the national laboratories.

This organisational form has had a success beyond expectations, as exploratories have become environments where concrete, substantive multi-disciplinary social mining research has been carried out. Moreover, exploratories have served as drivers in

attracting many users, both via transnational access and virtual access, as well as students and innovators attending the project's training and innovation initiatives. Exploratories have been the vehicle for fostering cooperation and synergies across different lines of activity within the research infrastructure, promoting networking, access and joint research.

After three and a half years since its inception, the approach developed by SoBigData has proven successful in creating a comprehensive **e-infrastructure**, which offers access to over **180 resources** (social datasets, Big Data analysis algorithms and courseware) designed to support data scientists in the execution of large scale experiments. The SoBigData e-infrastructure is currently used by a vast and diverse pool of stakeholders, including **120 companies and over 2,500 registered users,** with daily peaks of accesses and executions in the millions. Insofar, thanks to the Transnational Access action, the project has supported **on-site visits by 35 scientists** outside of the consortium that were hosted to develop their projects. Other requests for experiments within the SoBigData platform were activated via dissemination and training actions connected to the exploratories. The activity stemming from the exploratories has produced **over 200 high-profile publications** stemming from Big Data experiments executed via resources provided by SoBigData. The project's training and innovation actions have produced courses for **over 700 students** and **120 pilot projects developed with companies** (53 Small and Medium Enterprises, 41 large companies and 26 public institutions).

SoBigData has also promoted the creation of **ethical guidelines and tools** (http://fair.sobigdata.eu/moodle/) regarding the protection of personal data and intellectual property while promoting the free flow of personal data as a common good. Extreme care has been taken in all parts of the SoBigData project and among all consortium partners not only to ensure compliance with relevant laws, regulations and codes of conduct, but also to prevent unethical outcomes of the envisaged cutting-edge research and applications. Our approach is based on a continuous analysis of moral challenges posed to existing laws, regulations and governance of information technology.

SoBigData has contributed to shape the debate on Big Data, influencing national research bodies and public authorities, becoming a reference initiative in this domain. Moreover, the project has originated two novel PhD programs, one focused on Data Science in Pisa and the other centered on Engineering Social Technologies for a Responsible Digital Future in Delft. Further information can be found at http://project.sobigdata.eu/material, which is the project's deliverable repository.

SoBigData has worked under **FAIR principles** (Findability, Accessibility, Interoperability and Reproducibility) starting with a focus on **Findability** and **Accessibility** by developing a comprehensive catalogue of resources, including data, methods and workflows, organised in a specifically-created ontology in order to facilitate retrieval and access. Regarding **Interoperability** and **Reproducibility, further work is needed beyond the initial stage, in order to render data science experiments more easily designed, enhancing adjustability and repeatability by domain experts that are not data scientists.** SoBigData has also worked along another set of principles, namely **FACT** (Fairness, Accuracy, Confidentiality and Transparency) making ethical, privacy-preserving and responsible data science research operational through the development of dedicated tools and training activities. However, **this process is yet to be completed as some tools need to be operationalized, such as methods for data anonymization, privacy risk assessment, audit, explanation and validation of machine learning models**, all of which are needed in order to deliver trustworthy solutions in social mining and associated decision making.

SoBigData has initiated an ambitious program of **curation and publication of unique Big Data assets**, fostering novel multi-disciplinary research performed by its user-scientists. **This activity must be consolidated in order to engage scientists to publish curated data and 'data papers' in order to link data and analytical workflows to scientists' publication output.** SoBigData has enabled novel, impactful social mining studies within its exploratories, addressing extremely relevant contemporary social issues, such as the assessment and possible solutions regarding misinformation within the media ecosystem or the analysis and nowcasting of migration flows.

We are now a mature community and we can move toward **SoBigData PlusPlus (SoBigData++),** a multi-disciplinary and cross-disciplinary European research community, aimed at using **social mining and big data** to understand the complexity of our contemporary, globally-interconnected society: e.g., disentangling urban sustainability and resilience, societal well-being and its multiple facets, the unequal distribution of resources and opportunities, the "ecological" problems of our information system, such as polarization and misinformation, the dynamics and economic drivers behind human migration, the disruptive impact of Big Data and Artificial Intelligence on the future of our society.

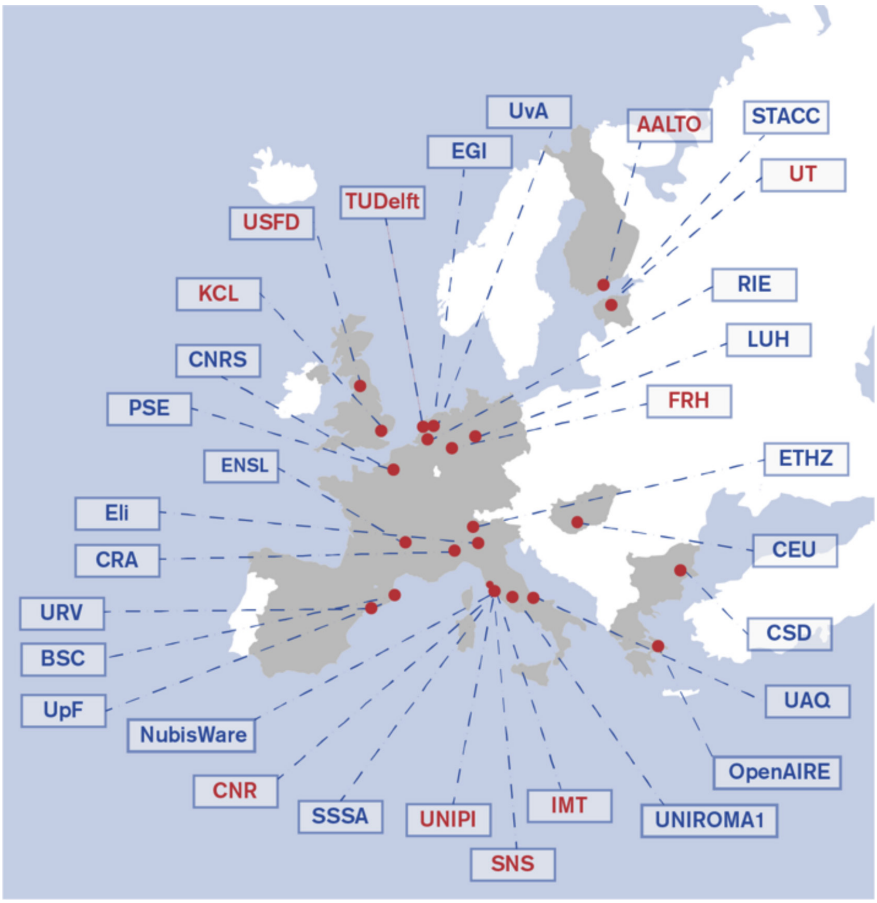
Becoming an advanced community, SoBigData++ will strengthen its tools

and services to help researchers and innovators to tackle these challenges, delivering **a platform for the design and execution of large-scale social mining experiments** accessible seamlessly on computational resources from the project cloud (align with European Open Science Cloud - EOSC) and on supercomputing facilities, open to users with diverse background.

SoBigData++ built on the solid foundations laid out by SoBigData in order to address future challenges and develop along the following general and intertwined objectives:

1. Moving forward from a repository of data and methods to a platform for the design and execution of complex social mining processes. This platform will be easily usable by interdisciplinary scientists and innovators and compliant with the European Open Science Cloud (EOSC).
2. Moving forward from a starting community of social mining pioneers to a wide and diverse movement capable of empowering the next generation of responsible social data scientists engaged into the challenging research questions of the exploratories and ambassadors of a critical data literacy aimed at facilitating data citizenship and data democracy.
3. Moving forward from the present awareness of ethical and legal challenges in social mining to the development of concrete tools that operationalise ethics with value-sensitive design, incorporating values and norms for privacy protection, fairness, transparency and pluralism.
4. Moving forward to an accelerator of data-driven innovation that facilitates the collaboration with industries and start-ups in order to develop pilot projects and proofs-of-concept.
5. Moving forward to an established pan-European Research Infrastructure to be recognised by ESFRI Roadmap and sustained by a SoBigData Association, which will design an appropriate form of legal entity.

SoBigData-PlusPlus (SoBigData++) has been approved for funding with grant n. 871042.



THE CONSORTIUM

A new enriched consortium will work on this ambitious project: **31 partners from 11 member countries of the European Union** (Bulgaria, Belgium, France, Greece, Spain, Hungary, Italy, UK, Germany, Estonia, Finland, The Netherlands) and Switzerland.

The core consortium has been designed to encompass the leading European research groups that contributed to shape the big social data analytics arena in the past decade. The core partners represent a mix of research centres that focus on big data from multiple perspectives: **E-infrastructures** and on-line services developers (CNR, EGI, OpenAIRE, BSC, Nubisware); **Big Data Analytics:** data mining and machine learning (CNR, UNIPI, AALTO, FRH, UT), text and social media mining (USFD, KCL, UNIPI, CNR, LUH), social network analysis (CNR, KCL, AALTO, UT, ETHZ, SNS, IMT), web intelligence (LUH, CNR, UPF), mobility analytics (CNR, FRH, ENSL, UNIPI, CRA), visual analytics (FRH, UvA, CNRS), complex network science (AALTO, SNS, IMT, ETHZ), computational sociology and social simulation (ETHZ); Complex Systems focussed on modelling social phenomena (ETHZ, ENSL, IMT, CEU, SNS); **Ethics and Legal aspects** of data protection (LUH, SSSA); **Privacy preserving techniques** (UNIPI, URV); **End Users** coming from several disciplines: Digital Humanities (KCL); Macroeconomics, Monetary Economics, Labour Economics and Inequalities (PSE, CEU, SSSA); Statistics, Demography and Quality of Life (UNIROMA1); Computational Social Science and Media Studies (ENSL, CNRS, UvA, UNIPI, KCL, CEU); Urban Planners and Environmental Sustainability (CRA, Eli, UAQ); **Think Tanks** aimed at observing the wealth of countries and providing new ideas for policy makers (RIE, CSD); Technology transfer company (STACC); and **Innovation managers** (SSSA).

Next SoBigData events

Data Science for Society

ACM Celebration of Women in Computing womENcourage 2019

"Diversity Drives Societal Change".

18th September 2019

MAXXI (National Museum of XXI Century Arts) - Rome

Nowadays, we socialize and search for information online, we sell and buy products and services thought web platforms, we carry on our mobile phone with us everywhere. Thanks to these digital breadcrumbs and the data science that allows us to make analysis, we are able to observe and understand complex social phenomena and take decisions that impact on our society. Data science is also changing the way scientific research is performed.

The purpose of this workshop is to encourage the research that will lead to the advancement of the social science. In particular, we will provide an overview of different methodologies and applications in several contexts: five successful female researchers will introduce their work and describe their experience in the respective fields.

Final SoBigData Conference

19th November 2019

European Parliament, Bruxelles

After three years since its start, the SoBigData approach revealed successful in delivering a comprehensive e-infrastructure offering access to 180+ social datasets, big data analysis algorithms and courseware, to support data scientists for executing large-scale experiments. The e-infrastructure is now used by a wide volume and diversity of stakeholders.
In the Final Conference we will present the scientific results of the project and we will provide an overview of the next steps: from SoBigData to SoBigData++ as advanced community.
Furthermore a special section will be dedicated to the researchers awarded by the ERC Advanced Grant.

More info soon published on www.sobigdata.eu. Stay tuned!

Transnational Access: an open call to visit the SoBigData R.I.

Interact with the local experts, discuss research questions, run experiments on non-public datasets and methods, present results at workshops/seminars.

The SoBigData project invites researchers and professionals to apply to participate in Short-Term Scientific Missions (STSMs) to carry forward their own big data projects. These opportunities are offered as part of SoBigData's Transnational Access (TNA) activities and calls for applications will be opened every six months.
We welcome applications from individuals with a scientific interest, professionals, startups and innovators that may benefit from training in data science and social media analytics. In order to apply you have to fill the Project Application Form.

Funding for a short-term scientific mission (2 weeks to 2 months) is available up to 4500 euros per participant (to cover the cost of daily subsistence, accommodation, and economy flights). STSM bursaries are awarded on a competitive basis, according to the procedure described in the application pack and eligibility criteria below, and based upon the quality of the applicant, the scientific merit of the proposed project, and their personal statement.

PRE-REQUISITES
Good understanding of social data and, ideally, track record of prior social data analysis projects
Experience with using at least one of machine learning, natural language processing, and/or complex networks algorithms

THE GOAL
The goal is to provide researchers and professionals with access to big data computing platforms, big social data resources, and cutting-edge computational methods.

Up to 4500 euros per participant, to cover all the cost incurred.

STSM visitors will be able to:
Interact with the local experts;
Discuss research questions;
Run experiments on non-public big social datasets and algorithms;
Present results at workshops/seminars.

The STSM visits will enable multi-disciplinary social mining experiments with the SoBigData Research Infrastructure assets: big data sets, analytical tools, services and skills.

RESEARCH CENTRES
Applications are invited for access at the following centres (infrastructures):

Gate *(Text and Social Media Mining), University of Sheffield;*

SoBigData.it *Pisa, Italy;*
Fraunhofer IGD, Darmstadt, Germany;

UT University of Tartu, Estonia;

L3S Research Center / Leibniz University Hannover;

Aalto University Finland;

ETHZ Zurich, Switzerland

The calls are published on the SoBigData website and the main social channels.

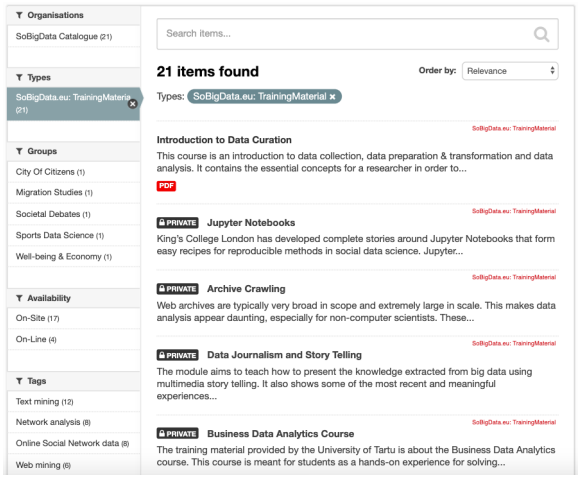
For more info, please visit the SoBigData website (*www.sobigdata.eu*)

The Fourth call on Transnational Access. Visit period: November 2018 / July 2019

The SoBigData e-learning area

Marco Braghieri, King's College London, UK | k1346642@kcl.ac.uk

Since the inception of the SoBigData project, training has been one of the main focus areas of both activities and the project's Research Infrastructure development. King's College London has led efforts in this area in



Work Package 4 and together with Work Package 7 and project partner CNR, it has created a dedicated area within the SoBigData Research Infrastructure, named e-Learning Area.

The e-Learning Area of the SoBigData infrastructure comprises all training materials that have been developed by the project's partners, including slides and lectures, interactive training materials, full training courses, and multimedia elements. In total, there are 22 training materials that have been integrated onto the Research Infrastructure by uploading them into the SoBigData Catalogue. This has allowed training materials to be housed within their own environment (the e-Learning Area) but also to be integrated within the six thematic areas which comprise the catalogue, which are text and social mining, web analytics, visual analytics, social network, human mobility and social data analysis. As per other resources, training materials are described by metadata and downloadable.

In detail there are 15 sets of lectures in slide format, 5 different interactive learning environments, 2 full courses and one video resource, produced by a variety of project partners. Project partner LUH has developed two interactive learning environments (Archive Crawling and Archive Spark); project partner TARTU has developed a full course in Business Data Analytics. Project partner ETHZ has developed a set of lectures on Data Mining and Machine Learning for Social Science and project partner FRH one on Visual Analytics for Data Scientists. Project partner KCL has developed two interactive learning environments based on the R programming language and Jupyter Notebooks, while project partner USFD has developed a full course on the usage of the GATE platform and a video on Online abuse of politicians. Moreover, project partners CNR, UNIPI have developed a wide range of lecture sets that comprise the SoBigData Program Post- Master in Big Data Analysis and Social Mining held at the University of Pisa, Italy.

Different stakeholders, once they have registered onto the SoBigData Research Infrastructure, can now browse, access and download training materials not only by title and description but also following some of the exploratories which have been developed within the project such as City of Citizens (smart cities, human mobility behavior analysis); Well-being & Economy (poverty indicators, spatial analysis); Societal Debates (text and social network analysis); Migration Studies (macroscopic human flows, social behavior analysis).

Despite being a recent addition to the SoBigData Research Infrastructure, although not in high numbers, registered users have increased over 5 times in less than a year, demonstrating the interest of the SoBigData community regarding training materials and its successful integration via the catalogue into the Research Infrastructure.

	Training Material	Developer
	Archive Crawling	LUH
	Archive Spark	LUH
	Business Data Analytics Course	TARTU
	Data Mining and Machine Learning for Social Science	ETHZ
	Visual Analytics for Data Scientists	FRH
	Jupyter Notebooks	KCL
	Interactive Learning Environments	KCL
	GATE Course	USFD
	SOS - Online Abuse of Politicians	USFD
	Text Analysis and Opinion Mining Module	CNR
	Social Network Analysis	CNR
	Social Network Analysis with Python	CNR
	Introduction to Data Science for Social Scientists	CNR
	High Performance and Scalable Analytics Module	CNR
	Data Management for Business Intelligence Module	UniPi
	Information Retrieval Module	UniPi
	Data Mining and Machine Learning Module	UniPi
	Data Visualisation and Data Analytics Module	UniPi
	Database Module	UniPi
	Efficiency - Effectiveness Trade-offs in Learning to Rank	CNR
	Data Journalism and Storytelling	CNR

ERC Advanced Grant: 4 SoBigData researchers awarded

ERC encourage the highest quality research in Europe through competitive funding and support investigator-driven frontier research across all fields, on the basis of scientific excellence.

European Research Council

Supporting top researchers from anywhere in the world

It is a pleasure to announce that 4 researchers of SoBigData, Marlon Dumas, Dirk Helbing, Aris Gionis and Fosca Giannotti have been awarded ERC Advanced Investigator Grant, an extremely competitive and particularly well funded Advanced Investigator Grants by the European Research Council.

Prof. Dirk Helbing, from ETH Zurich, has been granted for his work on how to use the wisdom of crowds to make cities smarter and, more specifically, how could networks of innovative cities contribute to the solution of humanity's existential problems.

"Given the on-going digital revolution and our present-day sustainability challenges, we have to reinvent the way cities are operated. We propose that the requirement of organizing societies in a more resilient way implies the need for more decentralized solutions, based on digitally assisted self-organization and that this concept is also compatible with sustainability requirements and stronger democratic participation." Says prof. Helbing

Prof. Marlon Dumas, at the University of Tartu, Estonia, intends to develop methods to discover opportunities for process improvement. Making business processes more efficient and agile is, in fact, of vital importance for modern organizations.

"My project", says prof. Dumas, *"will develop methods to analyze data extracted from enterprise systems in order to automatically discover opportunities for improving the quality and efficiency of business processes. The methods to be developed in this project will combine machine learning and optimization techniques to ensure that all possible improvement opportunities are considered and that the optimal combination of improvement opportunities is selected."*

Prof. Aristides Gionis, from the Computer Science Department of Aalto University, Finland, received the grant for studying how social media users shut themselves into closed off communities and explore the effectiveness of techniques designed to encourage users to be willing to engage with viewpoints which conflict with their own.
In fact, while online media is often promoted as a place which breaks information barriers and promotes diversity and democracy – in practice the opposite is often observed. Social media users tend to favour content

that agrees with their existing worldview and gets less exposure to conflicting viewpoints.
"Without any kind of intervention current social media platforms are gravitating towards a state in which net-citizens constantly reinforce their pre-existing opinions only." (A.Gionis)

Fosca Giannotti Research Director at National Research Council, Pisa, Italy and Project Coordinator of SoBigData, has been granted for her work on explaining and making human-understandable the decisions suggested by AI algorithms learned by data.

"To understand why it is so important to make human-understandable the decisions suggested by AI, let's imagine that situation: a friend of mine asks for a vacation credit card to his bank, to discover that the credit he is offered is surprisingly low. The bank teller cannot explain why. My stubborn friend continues his quest for explanation up to the bank executives, to discover that it was an algorithm that automatically lowered his credit score. Why? After a long, ad-hoc investigation, it turns out that this was due to ... bad credit by the former owner of the house where my friend lives", says F.Giannotti

In the project XAI, Science and Technology for the eXplanation of AI decision making, Prof. Giannotti focuses on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML systems, introducing a framework for theory and practice for black box explanation. Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic not only for lack of transparency but also for possible biases inherited by the algorithms from human prejudices and collection artefacts hidden in the training data, which may lead to unfair or wrong decisions.

The ERC's mission is to encourage the highest quality research in Europe through competitive funding and to support investigator-driven frontier research across all fields, on the basis of scientific excellence.

ERC grants are awarded through open competition to projects headed by starting and established researchers, irrespective of their origins, who are working or moving to work in Europe. The sole criterion for selection is scientific excellence. The aim is to recognise the best ideas and confer status and visibility on the best brains in Europe, while also attracting talent from abroad.

The ERC approach allows researchers to identify new opportunities and directions in any field of research, rather than being led by priorities set by politicians. This ensures that funds are channelled into new and promising areas of research with a greater degree of flexibility.

ERC Advanced Grant: Open Source Urbanism: Beyond Smart Cities

Open Source Urbanism can help mitigate the migration crisis and improve living conditions all over the world.

Dirk Helbing, ETH Zurich

The dream of building "good cities" is old[1]. Since the 20th century, there have been many attempts to create, develop or shape cities, sometimes even from scratch. Examples range from gigantic modernistic approaches known from Brasilia and Chandigarh, to more radical, but theoretical concepts aimed at changing society and engineering social order, such as Ecotopia or the Venus project. Recent developments are driven by the planetary trend towards urbanization, mass migration, and the need for sustainability. New visions of a global urban future were developed, such as "Sustainable", "Eco", or "Resilient" Cities, typically based on a top-down approach to the design of urban habitats.

Cities created from scratch heavily depend on massive private investments, for example, Songdo in South Korea or Lavasa in India. Despite ambitious goals and many technological innovations, their long-term success cannot be taken for granted, as they are often conceived by urban planners without the participation of people who later live in these cities. Such projects are typically implemented without much feedback from citizens. This makes it difficult to meet their needs. In fact, some of these cities have ended as "ghost cities".

In the wake of the digital revolution, data-driven approaches promised to overcome these problems. "Smart cities", "smart nations," and even a "smarter planet" were proposed. Various big IT companies decided to invest huge amounts of money into platforms designed to run the "cities of the future". Fuelled by the upcoming Internet of Things, cities would be covered with plenty of sensors to automate them and thereby turn them into a technology-driven "paradise." So far, however, these expectations have not been met.[2] Why?

> Given the digital revolution and the sustainability challenges, we now have to re-invent the way cities and human settlements are built and operated, and how cities can contribute to the solutions of humanity's present and future existential problems.
>
> *Dirk Helbing*

Geoffrey West points out that cities cannot be run like companies.[3] A company is oriented at maximizing profit, i.e. a single quantity, while a city must balance a lot of different goals and interests. This tends to make companies efficient, but vulnerable to mistakes. Cities are often less efficient, but more resilient. Driven by diverse interests, cities naturally do not put all eggs in one basket. This is why cities typically live longer than businesses, kingdoms, empires, and nation states.[4]

Importantly, cities are not just giant supply chains. They are also not huge entertainment parks, in which citizens consume premanufactured experiences. Instead, they are places of experimentation, learning, social interaction, creativity, innovation, and participation. Cities are places, in which diverse talents and perspectives come together, and collective intelligence emerges. Quality of life results, when many kinds of people can pursue their interests and unfold their talents while these activities inspire and catalyse each other. In other words, cities partly self-organize, based on a (co-)evolutionary dynamics.[5,6]

While rapid urbanization comes with many problems, such as the overuse of resources, climate change and inequality,[5] cities become ever more important, as they are motors of innovation.[3,5] Presently, more than half of humanity lives in cities, and the urban population is expected to increase to 68% by 2050. To meet the social, economic, and ecological challenges, innovation must be further accelerated, as the UN Agenda 2030 Sustainable Development Goals stress.

Given the digital revolution and the sustainability challenges, we now have to re-invent the way cities and human settlements are built and operated, and how cities can contribute to the solutions of humanity's present and future existential problems. In the past, we had primarily two ways of addressing such issues: (1) nation-states (and their organization in the United Nations) and (2) global corporations. Both have not managed to deliver the necessary solutions on time, e.g. to problems such as climate change and lack of sustainability. Therefore, we propose a third way of addressing global problems: through networks of cities.[7] So, how to unleash the urban innovation engine?

CITY CHALLENGES

"City Olympics" or "City Challenges" could boost innovation on a cross-city level involving all stakeholders. They would be national, international or even global competitions to find innovative solutions to important challenges. Competitive disciplines could, for example, be the reduction of climate change, the development of new, energy-efficient systems, sustainability, resilience, social integration, and peace. The solutions would be publicly funded and should be Open Source (for example, under a Creative Commons license) in order to be reused and developed further by a multitude of actors in all cities i.e. by corporations, SMEs and spin-offs, researchers, NGOs and civil society. In this way, the potential of trends such as Open Source Movement, Hackathons, Fablabs, MakerSpaces, Gov Labs and Citizen Science would be raised to an entirely new level, creating the potential for civil society solutions. The new success principles would be collaborative practices such as co-learning, co-creation, combinatorial innovation, co-ordination, co-operation, co-evolution, and collective intelligence.

Increasing the role of cities and regions as drivers of innovation would allow innovative solutions and initiatives to be launched in a bottom-up way. All interested circles could contribute to City Challenges. Scientists and engineers would come up with new solutions and citizens would be invited to participate as well, e.g. through Citizen Science. Media would continuously feature the efforts and progress made in the various projects. Companies could try to sell better products and services. Politicians would mobilize the society. Overall, this would create a positive, playful and forward-looking spirit, which could largely promote the transformation towards a digital and sustainable society. In the short time available (remember that the UN wants to accomplish the sustainability goals by around 2030), the ecological transformation of our society can only succeed if the majority of our society is taken on board, and if everyone can participate and profit.

OPENSOURCING URBAN INNOVATIONS

Cities are the places where the engagement of citizens can have the greatest impact. The most liveable cities manage to create opportunities to unfold the talents of many different people and cultures and to catalyse fruitful interactions among them. Opportunities for participation and co-creation are key for success.

> Increasing the role of cities and regions as drivers of innovation would allow innovative solutions and initiatives to be launched in a bottom-up way.
>
> *Dirk Helbing*

Alexandros Washburn[8] said about the design process of New York City that he could not control anything, but influence everything; successful urban design required the right combination of top-down and bottom-up involvement. It is therefore essential that urban development involves all stakeholders including citizens. Vauban, a quarter of the city of Freiburg, Germany, is a good example for this. The city council encouraged the citizens to actively participate in land-use planning and city budgeting. Sustainability and new energy-saving technologies were a primary focus of the planning strategy. In two new districts (Rieselfeld and Vauban), self-built and community architecture was created, which led to urban environments conceived and designed by future inhabitants according to their own vision. Now, Freiburg counts as benchmark city. Its concepts of sustainable urban planning and community participation are widely used by other cities all over the world.

So far, most urban planning professionals do not pay much attention to long-term involvement of citizens in urban development. With the ubiquity of information and communication technologies, our cities are getting smarter, but not automatically more inclusive, just, and democratic. The Citizen Score, a surveillance-based approach to control the behaviours of people, shows how easily technological progress may lead to technological totalitarianism. In the private sector as well, global corporations, geared towards profit, can turn into threats of democracy and human rights. When services are free, people are the product, data can fall into the wrong hands, and human dignity, autonomy, and freedom will be compromised. In data-rich societies, where people are measured and watched, profiled and targeted, this problem is quite significant. If cities of the future were run like businesses, based on surveillance, driven by data and controlled by algorithms, liberty, democracy, and human rights might quickly erode.

The application of open source principles to the co-creation of urban environments could overcome these problems by supporting active participation, technological pluralism and diversity. Thereby, it would also avoid technological lock-ins and dead-ends. The open source movement, which started with opening software (see the example of GitHub) now promotes the co-production of open content (Wikipedia, OpenStreetMap), open hardware (3D-printer RepRap), and even open architecture (WikiHouse). Open Source Urbanism would be the next logical step of this open source trend.

In 2011, Saskia Sassen wrote: "I see in Open Source a DNA that resonates strongly with how people make the city theirs or urbanize what might be an individual initiative. And yet, it stays so far away from the city. I think that it will require making. We need

to push this urbanizing of technologies to strengthen horizontal practices and initiatives."[4]

Yochai Benkler argues that open source projects indicate the beginning of a social, technological, organizational and economic transformation of the society towards a new mode of production.[9] This new mode, called commons-based peer production, is a collective activity of volunteers, usually coordinated via the Internet, producing free-to-use knowledge. Open Source Urbanism, as a new way of urban development, would therefore build on concepts such as Open Source Innovation and Commons-Based Peer Production.

In fact, citizens are keen to be not just consumers, but co-producers of their urban habitats. Some of them already experiment with open-sourcing urban design by collecting, improving, and sharing their Do-It-Yourself design blueprints and manuals on the Internet. The "Nation of Makers" initiative promotes community-driven design, prototyping, and fabrication as well in order to solve local and global challenges by improving lives in local communities around the planet.

Such examples are presently still rare and dispersed, and, therefore, not yet able to shift cities towards more inclusive urban development on a global scale. For this, one would need a socio-technical platform to consolidate and strengthen the nascent movement. Such a platform could promote the exchange of best practices and solutions to frequently occurring problems. The results would be digital commons designed to satisfy citizens' needs[10].

> "Glocalisation" thinking global but acting local is the new paradigm of globalisation
>
> *Dirk Helbing*

The proposed approach pushes for a new paradigm of globalisation, which one may call "glocalisation". It would be based on thinking global, but acting local (and diverse), on experimentation, learning from each other, and mutual support. The approach would be scalable. It would be more diverse and less vulnerable to disruptions. It would promote innovation and collective intelligence, while being compatible with privacy, freedom, participation, democracy, and a high quality of life. If cities would open up and engage in co-creation and sharing, they would quickly become more innovative and efficient. Open Source Urbanism could take our cities and societies to an entirely new level and also help to create better living conditions in developing countries and regions suffering from war more quickly.

References

(1) Sennett, R. Building and Dwelling: Ethics for the City. (Farrar, Straus and Giroux, 2018).
(2) Hugel, S. & Hoare, T. Disrupting cities through technology, Wilton Park. (2016).
(3) West, G. Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. (Penguin, 2017).
(4) Sassen, S. Open Source Urbanism. Domus (2011). Available at: http://www.domusweb.it/en/op-ed/2011/06/29/open-source-urbanism.html. (Accessed: 16th November 2016)
(5) Bettencourt, L. M. A. & West, G. A unified theory of urban living. Nature 467, 912–913 (2010).
(6) Batty, M. Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. (The MIT press, 2007).
(7) Barber, B. R. If mayors ruled the world: Dysfunctional nations, rising cities. (Yale University Press, 2013).
(8) Washburn, A. The nature of urban design: A New York perspective on resilience. (Island Press, 2013).
(9) Benkler, Y. Freedom in the Commons: Towards a Political Economy of Information. Duke Law J. 52, 1245–1276 (2003).
(10) Schrijver, L. in Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains (eds. van den Hoven, J., Vermaas, P. E. & van de Poel, I.) 589–611 (Springer Netherlands, 2015). doi:10.1007/978-94-007-6970-0_22

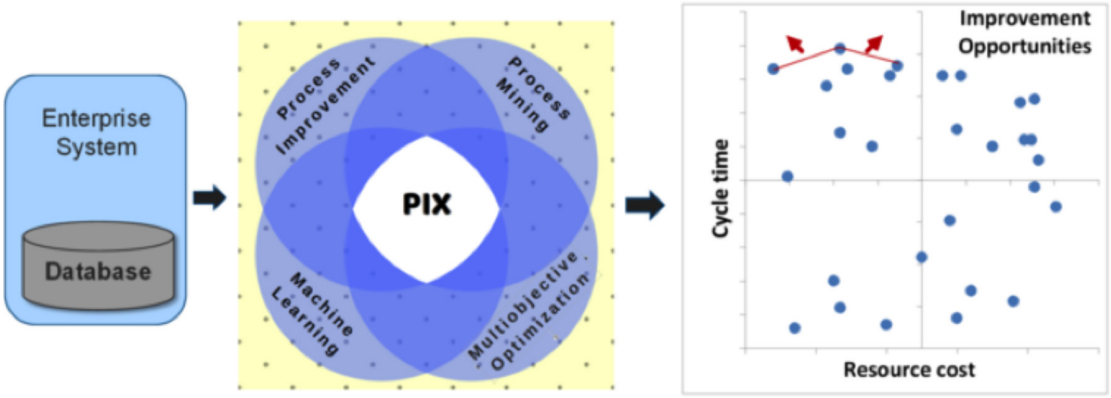
ERC Advanced Grant: The Process Improvement Explorer (PIX)

Marlon Dumas, University of Tartu

Business processes are the backbone of modern organizations. Their continuous improvement is key to the achievement of business objectives, be it with respect to efficiency, quality, compliance, or agility. Accordingly, a common task for analysts in companies is to discover and assess opportunities for business process improvement.

Current approaches to discover business process improvement opportunities are expert-driven. In these approaches, models and execution data are used to assess opportunities derived from experience and intuition rather than to discover them in the first place. Moreover, as the assessment of opportunities is manual, analysts can only explore a fraction thereof.

PIX is an ERC-funded project with the mission of building the foundations of a new generation of business process improvement methods that do not exclusively rely on guidelines and heuristics, but rather on a systematic exploration of a space of possible changes derived from process execution data. The PIX project will develop methods to analyze data extracted from enterprise systems in order to automatically discover opportunities for improving the quality and efficiency of business processes. These methods will combine machine learning and optimization techniques to ensure that all possible improvement opportunities are considered, and that the optimal combination of improvement opportunities is selected. The research will result in an open-source tool, namely Pix, which will automatically generate ideas for process improvement.



The Pix tool will extract data from enterprise systems and will apply advanced machine learning and optimization algorithms to search for optimal combinations of process improvement opportunities with respect to multiple performance measures.

Take for example an order-to-cash process. This is the process that starts when a company receives a purchase order from a customer and ends when the products ordered by the customer have been delivered and the customer has paid. The techniques we will develop in this project will be able to discover, for instance, that almost every time we send a purchase order to a supplier in Brazil without attaching a transport route, the delivery is delayed. On the other hand, when dealing with the Spanish supplier, the fastest solution is to let the local contact choose their own transport route. The Pix tool will recognise this difference by analysing the prior history of purchases and deliveries. It will then recommend to re-design the business process in such a way that a transport route will always be added to the order for the Brazilian partner but not for the Spanish one.

The main challenge here is that the number of ways in which a process can be changed or improved is very large. A typical order-to-cash process in a large company involves dozens of steps and each step involves several actions and decisions. We might deal with dozens if not hundreds of suppliers and customers, and each one has its own idiosyncrasies. There are millions of ways in which we could improve the process. To this end, Pix will incorporate advanced search algorithms to efficiently explore the space possible changes to a process, in order to identify combinations of changes that optimize one or more performance measures, such as cost, time, and defect rate.

Pix will be a major advance with respect to existing tools for data-driven process improvement, including process mining tools such as Apromore or Celonis. Unlike these tools, Pix will not just help analysts to identify bottlenecks and sources of defects, but it will also help them to come up with new improvement opportunities. To achieve this, the PIX project will borrow concepts and guidelines from established management theories, for example those found in Lean management, and it will "mechanize" these concepts and guidelines by mapping them to patterns that can be automatically found in sets of transactions recorded in enterprise systems.

ERC Advanced Grant: REBOUND: An algorithmic framework for reducing bias and polarization in online media

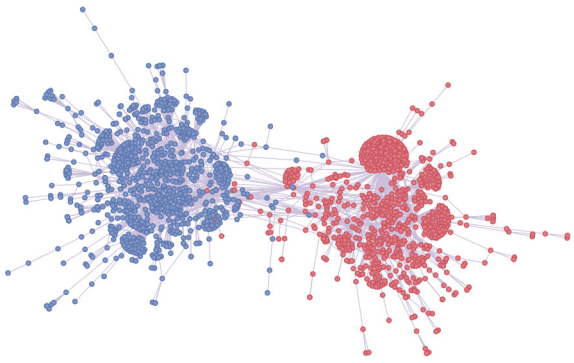
Aristides Gionis, Aalto University, Finland and KTH, Sweden

Social media play a critical role in today's information society, not only by connecting people with their friends, but also by providing a medium where information is disseminated and public opinion is shaped. Initially it seemed that giving ordinary citizens the means to create content of their own and share their opinion publicly can have only positive effects: increase the exposure to diverse ideas and improve the democratic process.

However, during the past few years we have witnessed that the rise of online media has led to a series of undesirable phenomena, such as creation of information silos and increased polarization. More and more digital citizens find themselves in online environments where they only see opinions that confirm their preexisting beliefs, and as a result, they become ideologically segregated.

These negative aspects of social media have drawn a lot of attention recently. There has been a considerable amount of criticism, skepticism, discussion in public forums, as well as calls to fix the problems. Prominent media outlets, researchers, and think tanks have contemplated whether *"social media is a threat to democracy."* Given all these concerns a natural question arises:
How can we address the deficiencies encountered in today's online platforms and how can we create environments in which online media enable exchange of alternative views and promote constructive deliberation?

This is a complex problem that requires parallel work and cooperation of multiple parties: owners of social-media platforms and traditional media outlets, journalists, policy makers, educators, as well as independent academics. It is also a cross-disciplinary research challenge that requires expertise in sociology, political science, economics, engineering, design, and computer science.



The "retweet" network of a polarized discussion that took place in twitter.

The REBOUND project aspires to contribute in addressing this complex problem by designing methods to perform large-scale data analysis so as to shed light to phenomena such as bias, polarization, conflict, and information silos in online media. It will also contribute with methods that give users easier access to opposing ideas and incentives to explore and understand alternative viewpoints.

Achieving the project goal is an extremely challenging task for a number of reasons. First, we will be dealing with noisy data, generated by a diverse set of actors who exhibit complex behavior and may have adversarial motives. Additionally, data are highly dynamic: new events are taking place and interactions between actors are constantly evolving over time. Second, we will deal with concepts that are not well-understood and we will need to develop novel abstractions. Third, we will be confronted with challenging computational tasks that lead to difficult optimization problems. Thus, it is desirable to develop methods that offer approximation guarantees or other theoretical properties.

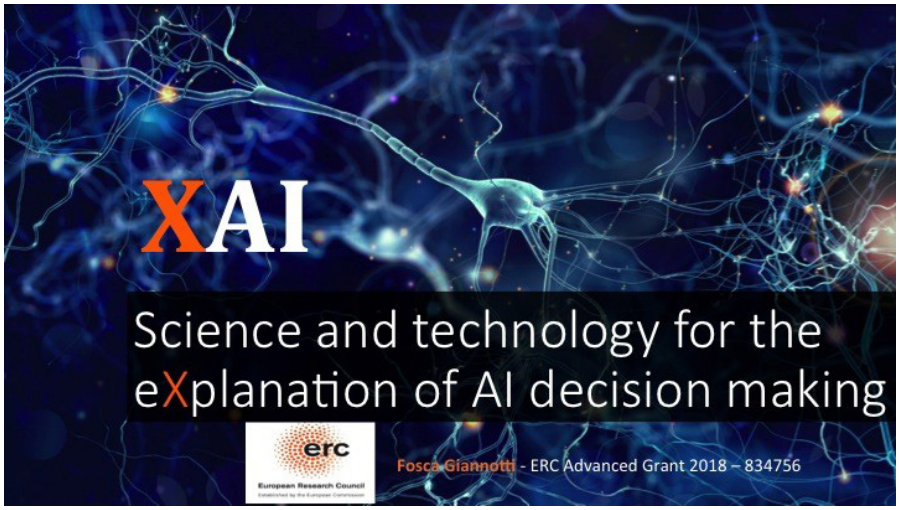
To achieve this ambitious objective we aim to consolidate existing approaches, including our recent work, and push significantly the state-of-the-art in terms of models, novel problem formulations, improved algorithmic techniques, and applications. In particular, we aim to develop theoretical foundations and a concrete set of algorithmic techniques to address deficiencies in today's online media.

We will develop methods to discover structure and patterns of segregation, conflict, and closeness in social-media systems. We will address the issues of reducing bias and polarization, breaking information silos, and creating awareness of users to explore alternative viewpoints.

ERC Advanced Grant: XAI: Science and technology for the eXplanation of AI decision making

Fosca Giannotti, SoBigData Coordinator, ISTI-CNR, Italy

A rather wealthy friend of mine asks for a vacation credit card to his bank, to discover that the credit he is offered is surprisingly low. The bank teller cannot explain why. My stubborn friend continues his quest for explanation up to the bank executives, to discover that it was an algorithm that automatically lowered his credit score. Why? After a long, ad-hoc investigation, it turns out that this was due to ... bad credit by the former owner of the house where my friend lives.



Black box AI systems for automated decision making, often based on machine learning over (big) data, map a user's features into a class or a score without exposing the reasons why. This is problematic not only for lack of transparency, but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions.

Many controversial cases have highlighted that delegating decision-making to black boxes is critical in many sensitive domains. To avoid such cases, the EU General Data Protection Regulation, introduces a right of explanation for individuals to obtain "meaningful information of the logic involved" when automated decision making takes place with "legal effects" on individuals. Without an enabling technology for explanation this right will either remain "dead letter", or will just outlaw opaque AI systems.

Therefore, it is dramatically urgent that Artificial Intelligence (AI) systems must be explainable and comprehensible in human terms; this is instrumental for validating quality and correctness of the resulting systems, and also for aligning the algorithms with human values and expectations, as well as preserving human autonomy and awareness in decision making. A missing step in the construction of a machine learning model generally used in AI system is precisely the explanation of its logic, expressed in a comprehensible, human-readable format, that highlights the biases learned by the model, validating their plausibility.

My project focuses on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML systems, introducing the local-to-global framework for black box explanation, articulated along three lines: (i) the language for expressing explanations in terms of expressive logic rules, with statistical and causal interpretation; (ii) the inference of local explanations for revealing the decision rationale for a specific case, by auditing the black box in the vicinity of the target instance; (iii), the bottom-up generalization of many local explanations into simple global ones, with algorithms that optimize for quality and comprehensibility. An intertwined line of research will investigate (i) causal explanations, i.e., models that capture the causal relationships among the (endogenous and exogenous) variables and the decision, and (ii) mechanistic/physical models that capture the detailed data generation behaviour behind specific deep learning models, by means of the tools of statistical physics of complex systems.

Alongside with the development of the models and algorithms for explanation, the XAI project will also develop: (i) an explanation infrastructure for the benchmarking of the methods developed within and outside this project, equipped with platforms for the users' assessment of the explanations and the crowdsourcing of observational decision data; (ii) an ethical-legal framework, both for compliance and impact of our developed methods on current legal standards; and (ii) a repertoire of case studies in explanation-by-design, with a priority in health and fraud detection applications.

Weak Signals in the Mobility Landscape: Car Sharing in Ten European Cities

Chiara Boldrini, Ubiquitous Internet Group, IIT-CNR, Italy | chiara.boldrini@iit.cnr.it

Raffaele Bruno, Ubiquitous Internet Group, IIT-CNR, Italy | raffaele.bruno@iit.cnr.it

Car sharing is one the pillars of a smart transportation infrastructure, as it is expected to reduce traffic congestion, parking demands and pollution in our cities. From the point of view of demand modelling, car sharing is a weak signal in the city landscape: only a small percentage of the population uses it, and thus it is difficult to study reliably with traditional techniques such as households travel diaries. In this work, we departed from the traditional approaches and we leveraged web-based, digital records about vehicle availability in 10 European cities for one of the major active car sharing operators.

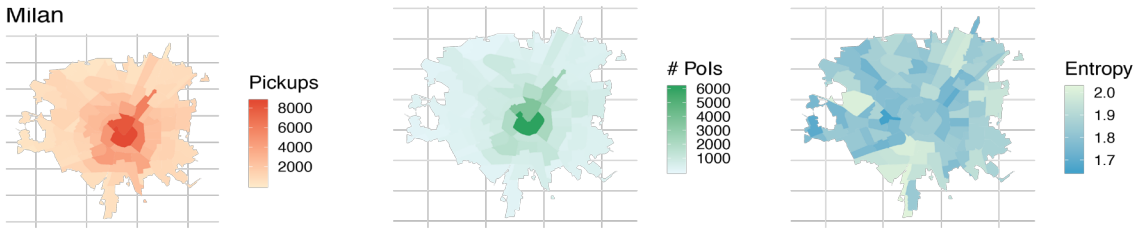


Figure 1: An example of the spatial distribution of the car sharing demand and urban fabric indicators in the city of Milan.

As a first step, we performed an explanatory analysis of the car sharing demand as a function of the socio-demographic and urban fabric (i.e., number, heterogeneity, and category of Foursquare Points of Interests - PoI) indicators associated with the cities of Milan, Rome, and Turin. While a single explanatory pattern does not emerge across the cities, the three cities share indeed several similarities. In fact, their car sharing demand is positively associated with high educational attainment (all Italian cities) and negatively correlated with commuting outside of the municipality area (Milan, Rome). These findings confirm the conclusion of the most recent sociodemographic surveys about car sharing services, but at a much finer spatial granularity and without relying on expensive and time-consuming interviews/questionnaires. With regards to the urban fabric indicators, the only PoI category that seems to have a statistically significant effect on car sharing demand is that of nightlife-related activities, suggesting that leisure is the most typical trip purpose.

Demand predictability is one of the crucial aspects for every transportation system. In car sharing, in particular, it is of utmost importance for vehicle redistribution, whose goal is in fact to proactively move vehicles in order to address the future demand. Thus, we have taken into consideration several approaches to demand forecasting, and we evaluated which are the best performing when it comes to car sharing pickups/drop-offs forecasting. Our results show that Random Forest yields consistently better results than simple average-based forecasting, time series forecasting, vanilla neural networks, and a popular custom approach proposed in the literature. However, prediction quality is in general quite good, even with the simplest solutions.

It is expected that different areas in a car sharing system are used differently by the users, but how many different usages can be identified? In order to answer this question, we have carried out a classification of districts based on their usage pattern. Four distinct car availability temporal patterns can be recognised in the cities considered in this study. We have labelled them day, night, neutral, and high-intensity behaviours, based on when they exhibit their peak availability and on the intensity of this peak. We also show that these patterns tend to be spatially autocorrelated, i.e., neighbouring districts are likely to feature the same behaviour.

A critical operational aspect for car sharing is how to perform cleaning and maintenance. When not done properly, it may even be a critical factor of the service shutdown, as in the case of Parisian car sharing Autolib. In order to perform cleaning and maintenance, the car sharing workforce is typically remotely dispatched to collect vehicles that are in need of either. However, moving workers around is expensive and inefficient. Better solutions can be found leveraging vehicle usage in the city. Specifically, we have shown how the airport zone can become a strategic asset for car sharing operators, due to the fact that the high volume of traffic generated by the area makes it extremely convenient to deploy cleaning and maintenance facilities there, thereby reducing the maintenance trips carried out by the car sharing workforce.

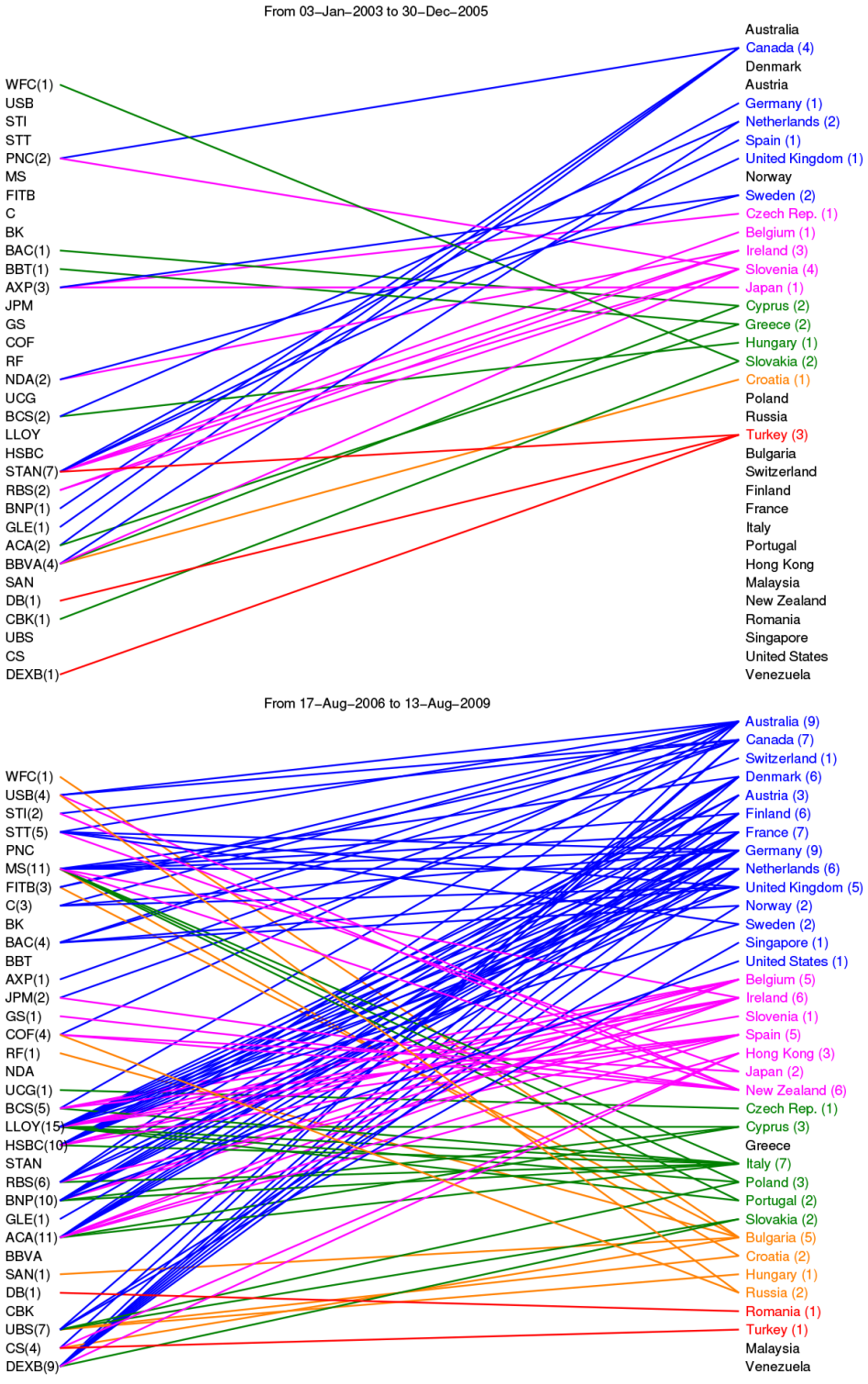
This time is different: monitoring the network of interconnections

Fabrizio Lillo, Università di Bologna, Italy | fabrizio.lillo@unibo.it

"This time is different" is the title of a famous book, written by Kenneth Rogoff and Carmen Reinhart in the aftermath of the 2008 financial crisis, claiming that the human history has been always plagued by financial crises. As each of us has directly experienced, these crises have a huge impact also on the "real" economy and on our everyday life. Understanding, monitoring, and possibly predicting these crises has become of paramount importance and subject of intense research by central banks, academicians, and financial institutions. The term systemic risk refers to the risk of collapse of a significant part of the financial system.

One of the key evidences of the studies of systemic risk is that interconnections between financial institutions, assets, companies, etc. are the main channel of propagation of risk, and can create a domino-like effect affecting large portions of the system. Monitoring interconnections is therefore very important and central banks, for example, started to collect large amount of data, after the crisis, exactly to this end. However understanding the role of interconnections and their capability to transmit risk is very complicated due to the complexity of the financial system.

**For this reason, in a recent paper(***) of the team of the Scuola Normale Superiore, funded in part by the European project SoBigData, we have proposed a new method to monitor the networks of interconnections by looking directly to the market data. In particular, we used a statistical technique, which allows deciding whether an extreme event (large or small price changes in the financial domain) in one time series helps predicting that there will be an extreme



Representation of the bipartite network of causality of risk between 33 global systemically important banks and 36 sovereign debt bonds. The above panel refers to a tranquil period, while the below one to the period of the Lehman crisis.

event of another time series in the future. This is called Granger causality in tail or in risk, since tail (extreme) events are clearly signaling very risky conditions. The existence of such causality indicates that risk in the first system spills over the second system. When many time series, each of them corresponding to a financial institution or investment, are considered, we obtained a Granger-tail network, which maps the propagation of risk in the system. By studying the topology of this network one can identify risk spreaders, risk absorbers, and in general identifying channels of risk propagation.

As an important case study, we considered 33 global systemically important banks and 36 sovereign debt bonds. Banks heavily invest in sovereign bonds, often of the country they belong to, and this creates a linkage between these two elements, since a large price swing of a bond affects the market value of the bank owning it. Moreover, as the sovereign debt crisis has shown, the deterioration of the quality of the bonds of some countries (for example the PIIGS countries) pushed banks disinvest in low quality bond and purchase high quality ones.

This massive and coordinated trading activity can lead to large price changes of bonds, prompting risk propagation. Combining measures of connectedness of Granger tail risk networks with the ratings of the sovereign bonds, the paper proposes a novel flight-to-quality indicator to identify periods of turbulence in the market. The measure clearly peaks at the onset of the European sovereign debt crisis, signaling the instability of the financial system.

Moreover the topology of the network dramatically changes, see the figure, displaying different, yet interpretable, patterns in the different phases of the crisis.

Finally, the connectedness measures of the networks can be used to forecast the quality of sovereign bonds. In conclusion, the method is able to identify channels of risk or extreme events propagation, not only in the financial domain, and can be fruitfully used for monitoring and prediction of systemic risk.

(*) Fulvio Corsi, Fabrizio Lillo, Davide Pirino, Luca Trapin, Measuring the propagation of financial distress with Granger-causality tail risk networks, Journal of Financial Stability 38, 18-36 (2018)

A perspective for electric mobility at regional level: a data-driven approach based on complex networks

Angelo Facchini, IMT Lucca, Italy

The expected electrification of mobility is one of the central strategies of the EU green energy transition, and according to the International Energy Agency about 19% of the GHG emission reduction should be attained through the mitigation of transport emissions.

In addition, as urbanization will increase in the next years, electric mobility especially in large urban areas like megacities will play a fundamental role. Indeed, in 2011 the world's gasoline share for mobility in 27 megacities was about 9%, and this share is expected to increase further. Consequently, an increase of GHG emissions and air pollution, as well as a growing dependency on fuel import is expected and should be mitigated, both for the environmental point of view and from the economic point of view.

Electric Vehicles (EV) fueled by renewable energy sources are considered a further winning strategy to reduce the GHG emissions from urban regions, and as the number of EV is expected to increase, given the potential high number of vehicles involved, a reliable planning and control of the charging infrastructure will be fundamental for its successful integration in the power system.

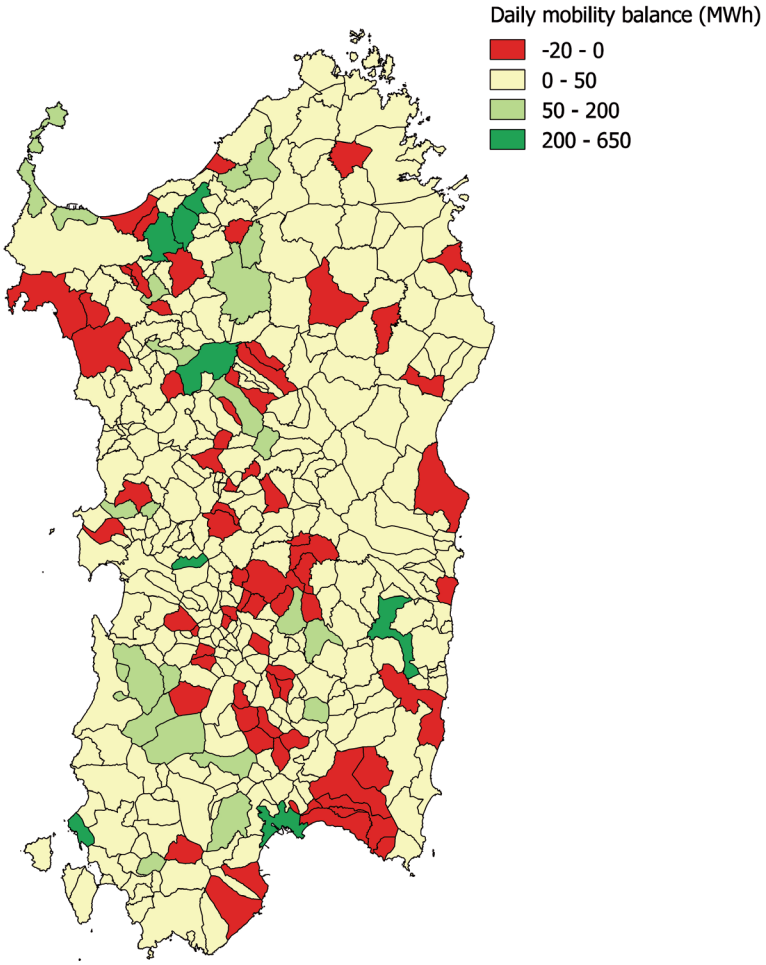


Figure 1: Energy balance map

In particular, a correct sizing of the needed resources, in term of electricity, charging stations, communication and management infrastructures must be based on the local mobility needs and on both the distribution infrastructures and electricity market constraints. In fact, electricity infrastructures structured on a hierarchical, fossil-based model of production and distribution need to be replaced out by a new, flexible model of generation and distribution whose solid grounds are based on renewable energy sources, integration, and decentralization.

In a recent paper (*) we performed a thought experiment: what might happen if over 700k cars present in the Sardinia region suddenly become electric? Would the regional infrastructures both for production and distribution be able to manage the increase demand for electricity? As a further challenge, we imposed that EV are only fueled by renewable generated power. To answer these questions we used an approach based on behavioural science, network science and data science, finding that electric mobility at regional level is possible but with some important caveats for policy makers and planners.

We study the co-existence of two different infrastructural network

layers, energy and mobility, to estimate the impact of electric cars, with the aim to quantify both short range and long range effects on the distribution network. Exploiting census data on commuting we estimate the origin-destination matrix for the trips between all the Sardinian municipalities, finding that commuting is mainly composed by short and medium range trips, which combines well with the limited ranges of EVs.

Furthermore, relying on maps of solar and wind generation plants, we identify those municipalities that are able to produce the low carbon electricity able to support the mobility of EV on regional scale. From these datasets, it is possible to estimate the energy need of charging services both on cities and on freeways, as well as the investigate the possible integration with the local grid and renewable generation plants. As a result we computed a set of maps reporting for each municipality the energy balance due to electric mobility. Figure 1 reports three cases:

1) Red, or negative balance: the energy needs are higher with respect to the renewable generation present in the municipality

2) Green, or positive balance: a surplus of energy is present

3) Light yellow: substantial energy balance

We notice that there are some regions in which both the surplus of production and consumption are relevant, especially (but not only) in the most urbanised areas of the region.

As a further step, we estimated the impact of the mobility infrastructure on the power grid by computing the daily energy flows associated to the energy balances using a set of algorithms that ensure local power balance in each node while minimizing the transmission costs over the power network. Figure 2 shows the energy flows associated to the unbalances: flows range from 0 to 7800 kWh and green municipalities indicate the ones contributing more to the energy production, while red municipalities are those showing unbalances up to 20MWh. We notice that the mobility needs of the two biggest cities of the territory can cause relevant energy flows in the regional grid, especially in the southern zone (where the capital city of Cagliari is located) we observe significant energy transactions extending (and challenging) transmission lines up to 100 km from the traffic attractor.

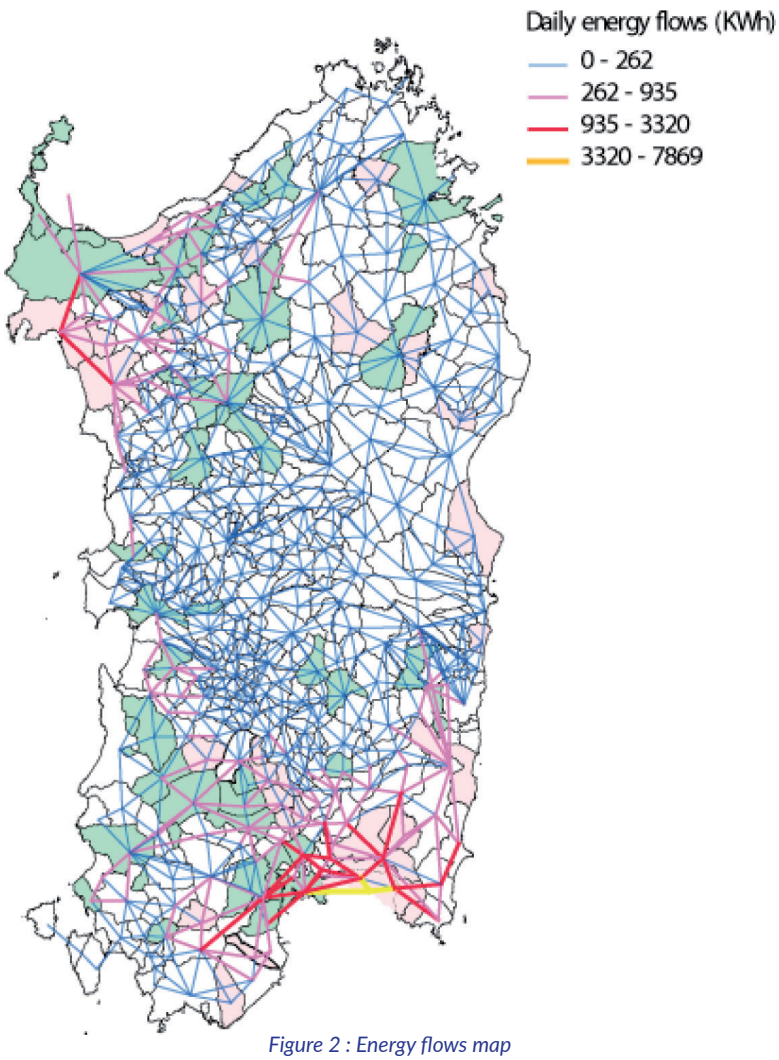


Figure 2 : Energy flows map

According to the general balance between distributed generation and electric mobility consumption, the actual production of the whole region is able to sustain the full deployment of EV for commuting, but attention must be devoted to the flows generated. As policy indication (and take-home message) we recommend that renewable generation plans should be located where they are actually needed, in order to reduce the stress on the power network, avoid transmission costs and infrastructural challenges.

(*) M. Mureddu, A. Facchini, A. Scala, G. Caldarelli, A. Damiano, A Complex Network Approach for the Estimation of the Energy Demand of Electric Mobility, Sci. Reports, 2018(8:268).

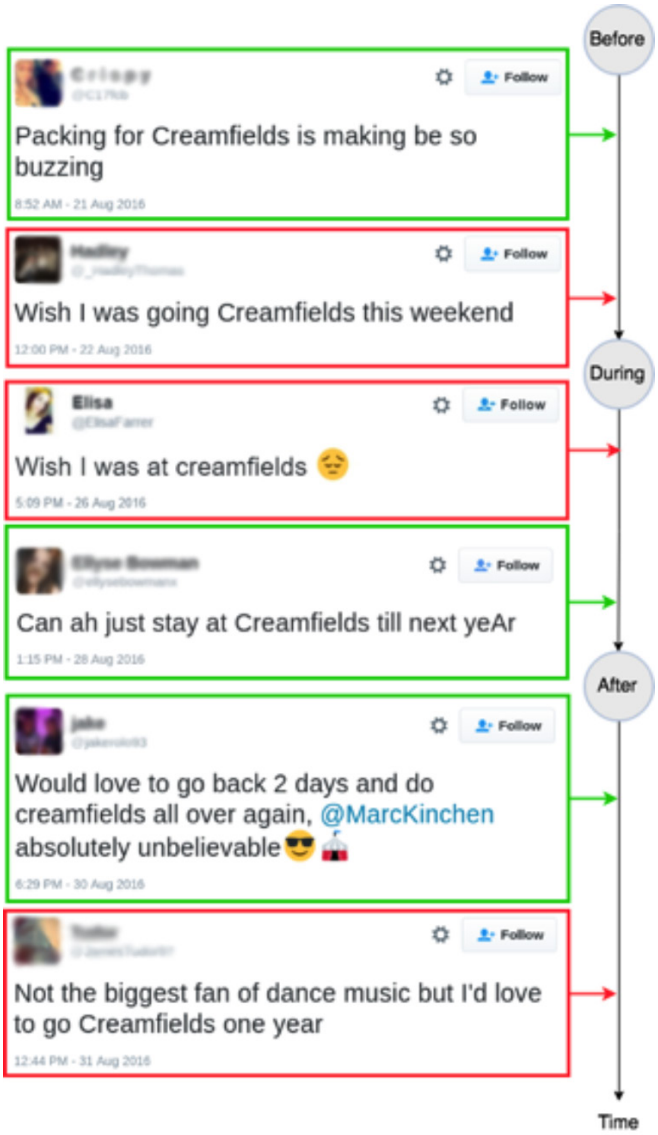
Event Attendance Prediction

Cristina Muntean, HPC Lab, ISTI-CNR, Italy | cristina.muntean@isti.cnr.it

Popular events are well reflected on social media, where people share their feelings and discuss their experiences. In our research we wanted to investigate the novel problem of exploiting the content of *non-geotagged posts* on social media in order to infer the users' attendance of large events. We detail the features used to train event attendance classifiers and report on experiments conducted on data from two large music festivals in the UK, namely the VFestival and Creamfields events. The work has been published in two important venues, the initial study at the ASONAM conference (1) and a more detailed study in the IPM journal (2).

An unprecedented amount of user-generated content about human activities has been created through the introduction of popular social media applications on smartphones (e.g. Facebook, Foursquare, Instagram, Twitter). The diversity and variety of content shared on these online social networks bears witness to the richness of these new forms of social interactions constituting nowadays an important source for an unprecedented outreach, speed and democratization of communication. Due to this vast applicability, social media analytics is a fast growing research area. Social media can be exploited to extract valuable information concerning human dynamics and behaviors. As a consequence, they can play a role in understanding modern life, including transportation and human mobility.

Music festivals, like many other popular events (e.g., important religious celebrations or sports matches) attract thousands of participants. Usually, they are well reflected in social media networks, allowing people to connect with "the event", expressing through posts their feelings, experiences or opinions well in advance of its planned date.



Given the attention to popular events reflected in social media, we address a novel challenging problem: *Is it possible to infer from Twitter posts the actual attendance of the user to the cited event?*. If we could classify user posts discussing an event on the basis of the actual attendance of the user to the event, we could enable or enhance several practical applications in the fields, for example, of targeted advertising and mobility management.

The simplest way of inferring the presence of users at events is to consider the geotag associated with their posts: the "check-in" or the user location in the event place at the time of the event can indeed be trivially associated with attendance. We observe however that this approach suffers from two drawbacks.

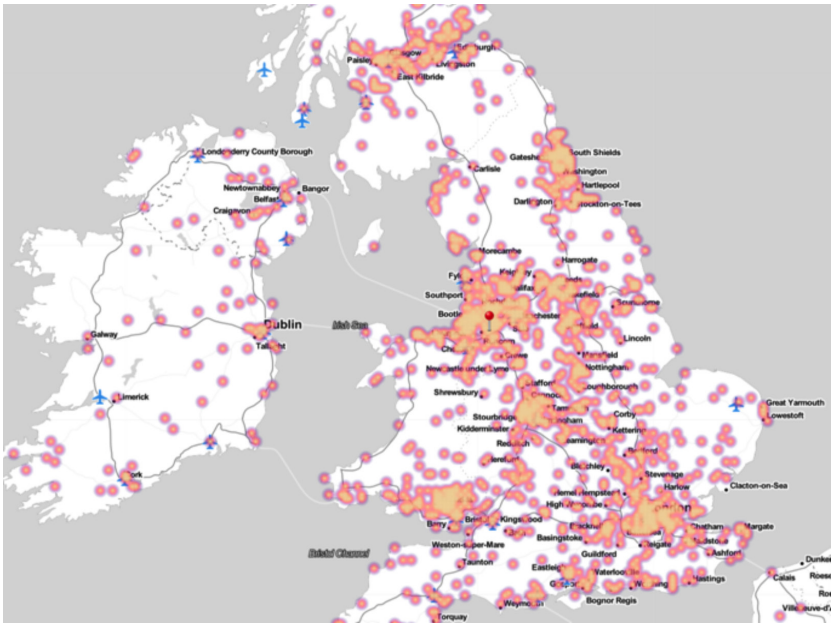
The first drawback is that very few social media users enable the geotagging of their posts (in Twitter the percentage of geotagged posts is about 2%). In fact, geolocation information is geographically accurate, but represents a very sparse data source. Learning attendance prediction classifiers based on sparse data would be extremely difficult and may lead to ineffective predictive models.

The second drawback of only using geolocated data is that they do not represent the intention of the user to participate in the event. Indeed, our aim is to infer the user's intention of participating to the event even before the event takes place.

To overcome these two aforementioned issues, we take a different direction by addressing the novel challenge of inferring the actual attendance of users to a mentioned event by only relying on the content of *non-geotagged posts*, without considering any spatial features. By not relying on {the} geotagged posts we can analyze a much larger number of posts to predict user attendance to a given event. The large {base} of users covered by our approach makes it a {good and realistic} candidate to enable innovative services and applications in the

field, for example, of transportation planning and crowd safety management.

Moreover, we perform our event attendance classification by distinguishing three temporal intervals identifying when the posts have been shared on social media: *before*, *during or after* the event. We propose three distinct classification tasks, one for each temporal interval. The analysis of posts shared before the event acts as a prediction for the users' actual attendance, the analysis of posts shared during the event reflects the actual participation of users at the event, while the analysis of posts shared after the event gives a view or a summary of past attendance. We exploit four different categories of features. Each category reflects a different dimension of social media, namely the: textual, temporal, social, and multimedia dimensions.



The preliminary results achieved highlighted that the features extracted from the textual content are those playing the most important role for attaining a good classification accuracy. This interesting insight encouraged us to explore in more in-depth these features and to study also how to enrich the textual features to help generalize across different events the learned classification models.

The "before" case is particularly interesting, since an early knowledge of the possible user attendance can be useful to enable innovative services and applications. For example, event organizers or third-party companies could precisely target their advertisement campaigns by offering specific or personalized services to the users predicted to participate in the event. Another example is in the field of transportation planning, where attendance prediction could allow the organizers or the local authorities to push potential attendees to use public transportation or can help bus and shuttle companies to plan and advertise collective transport services to the event.

During the event, people may express their feelings about the event, may report issues with the provided services or may also share photos and videos about the event. After the event, users may report feelings and comments on their experience at the event. Knowledge of the social media users who attended (or did not attend) the event can be very useful as well. Their posts can be used for example to understand attendees' profiles and provide insights, allowing to improve the future editions of the event. In addition, this knowledge can also be used to help in estimating crowd sizes and support transportation planning for the future version of the event.

Our classifiers attain very high accuracy with the highest result observed for the Creamfields festival. The results show that our approach provides a remarkably good performance, exhibiting ~91% accuracy at classifying users that have indicated their intention to attend the event. Our analysis showed that word embedding features contribute {saliently} to the performance. Additionally, we highlighted the most informative group of features and assessed the accuracy of our classifier even on an objective test set constituted by geo-tagged tweets. In Section \ref{sub:rq2}, we analyzed the generalization of the learned models across the datasets and {proposed} additional word embedding features to improve cross-dataset {performances}. {For example, when classifying the posts published after the event, by including both the embedding and {\em NFV} features, the GBDT has increased up to +7.8\% (from 73.3\% to 81.1\%) its generalization ability when trained on Creamfields dataset and tested on VFestival dataset}.

We study the most informative features for the tasks addressed and the generalization of the learned models across different events. We also discuss an illustrative application of the methodology in the field of transportation.

As future work, we aim to improve our results using information extracted from the visual content of the published photos or videos. The analysis of visual content is a growing trend in social media and could be {better} explored {in} our classification {process} through the use of {deep learning techniques}. Furthermore, we aim to further explore our methodology in the context of smart transportation applications.

(1) Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times: Exploring Social Media for Event Attendance. ASONAM 2017: 447-450

(2) Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times: Event attendance classification in social media. Inf. Process. Manage. 56(3): 687-703 (2019)

A glimpse into the intertwining of epistemic opacity and moral opacity in AI systems

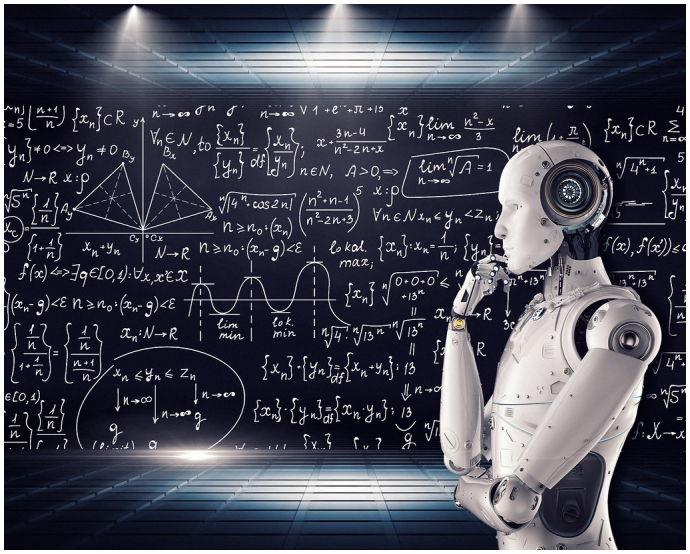
Juan Manuel Duran, TUDelft, Netherlands | J.M.Duran@tudelft.nl

An extended concern shared by researchers working with sophisticated algorithms and sheer amounts of data is that they have become "black boxes," a metaphor that attempts to capture the impossibility of a cognitive agent such as the researcher herself accessing the algorithm and its data. This is, indeed, a significant challenge for the reliability of algorithms, the ulterior trustworthiness of the results, and the moral justification for many of our actions.

On the face of it, there seem to be three possible alternatives to deal with epistemic opacity: first, we simply give up trying to open the black box. Algorithms are too complex for us, and we will never understand what is happening inside the machine. This option presents a dualistic scenario: either we blindly trust the results, or we don't. Unfortunately, this is virtually a self-defeating strategy. Imagine an AI system for medical care, where its complexity goes beyond any form of human comprehension. Now imagine the system recommending treatment for a given illness.

What should the clinician do? If she trusts the results and follows the recommended treatment without knowing why, or even being able to contend the results, she will be violating several basic medical principles, such as the autonomy of patients and the need for informed consent. On the other hand, if she does not follow the recommendation (remember, she cannot contend the result of an AI system), then why would we have such a system at all? It seems clear to me that this first option is a non-starter, for epistemic opacity straightforwardly leads to some form of moral opacity.

Alternatively, we decide to make an effort to survey the algorithm, similar to how a mathematician surveys a theorem and its proof. Attempts could include yet other pieces of algorithms (i.e., algorithm2) that will give some form of insight into how the original algorithm works (i.e., algorithm1). This is, by and large, the most popular approach among practitioners. Although a promising solution, this approach is not without problems.



The proposed strategy begs the question: how do we know that algorithm2 will provide the required insight about algorithm1? After all, epistemic opacity is not exclusive to algorithm1. The underlying idea is that algorithm2 is somehow simpler, and it only shows selected relations holding among stipulated variables in algorithm1.

Thus, if a clinician wants to know why the AI system recommends a given treatment, algorithm2 shows that certain variables and relations hold in algorithm1 which are relevant for the recommended treatment. Proponents of this solution argue that the clinician has now grounds for moral action based upon the knowledge offered by algorithm2. True, but not enough. Prima facie, the trust we have in algorithm1 is, in fact, trust in algorithm2, and that is all but epistemically and morally reassuring.

There is a third option, which consists of simply accepting epistemic opacity and embracing it, but also circumventing it to achieve epistemological and moral transparency. Taking this route means coming up with strategies that grant reliability to the AI system and trustworthiness to its results without surveying the algorithm. Today, the most promising framework to this end is called computational reliabilism. It consists of identifying epistemically self-grounded practices capable of granting reliability to the AI system. Examples are verification and validation methods and robustness analysis, just to mention two. Computational reliabilism, arguably, offers a better epistemic basis for morally transparent practices with AI systems.

Moral problems emerging in the context of new technologies such as AI system have an epistemological side, often hidden from sight. Black boxes are especially attractive because they bring these two traditionally separated domains to work in tandem for finding satisfactory solutions.

SoBigData Magazine is published under the project N° 654024 | Programme: H2020 | INFRAIA

Duration: 1/09/2015 - 30/08/2019

Editorial Secretariat

info@sobigdata.eu

Editorial Board

Fosca Giannotti
Beatrice Rapisarda
Roberto Trasarti
Valerio Grossi

Layout and Design

Beatrice Rapisarda

Copyright notice

Privacy statement

The personal data (names, email addresses...) and the other information entered in SoBigData Magazine will be treated according with the provision set out in Legislative Degree 196/2003 (known as Privacy Code) and subsequently integration and amendement.

Coordinator and Legal representative of the project: Fosca Giannotti | fosca.giannotti@isti.cnr.it

SOBIGDATA News is not for sale but is distributed for purposes of study and research and published online at http://www.sobigdata.eu/newsletter

To subscribe/unsubscribe, please visit http://www.sobigdata.eu/newsletter