Issue 5 / December 2020

Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Magazine

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance along such ambitious lines thanks to SoBigData, the predecessor project that started this construction in 2015.

Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments open to users with diverse background, accessible from the European Open Science Cloud and on supercomputing facilities. Pushing the FAIR principles further, SoBigData++ will render social mining experiments more easily designed, adjusted and repeated by domain experts that are not data scientists. SoBigData++ will move forward from a starting community of pioneers to a wide and diverse scientific movement, capable of empowering the next generation of responsible social data scientists, engaged into the grand challenges of the exploratories:

[Continues on pag. 3]

Inside this issue

03 **EDITORIAL**
Roberto Trasarti, CNR, Italy

07 **ESFRI ROADMAP 2021**
Valerio Grossi, CNR, Italy

13 **EVENTS HIGHLIGHTS**
Multiple authors

18 **RESEARCH HIGHLIGHTS**
Multiple authors

23 **TRANSNATIONAL ACCESS**
Editorial Board

24 **EXPLORATORIES HIGHLIGHTS**
Multiple authors

Content

Editorial

SoBigData++ : an advanced community of big data researchers and innovators. 3

News

ESFRI Roadmap 2021. 7

Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. 8

SoBigData goes virtual!. 9

Events Highlights

Real Time Epidemic Datathon. 13

EGI Conference 2020: SoBigData Research infrastructure. 15

SocInfo 2020: a virtual edition of the Social Informatics conference 17

Research Highlights

Proposal for a regulation on AI: our feedback..18

Visual Analytics for Data Scientists. 19

Knowledge Sharing And Network Dynamics In A European Research Project Setting. The Case Of Sobigdata++ 20

TransNational Access

TNA visits: Ready when the world is open. 23

Exploratories Highlights

Transparency Issues in Tracing COVID-19. . . . 24

Private Sources of Mobility Data Under COVID-19. 26

How digital data is changing how we measure well-being and happiness. 28

Human migration: the Big Data perspective. . 30

Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. 32

Intensity vs Accuracy: Technical-tactical differences between male and female football teams. 34

SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics

Roberto Trasarti, Porject Manager of SoBogData++ ISTI-CNR, Nationale Research Council of Italy | roberto.trasarti@isti.cnr.it

[Continued]

Societal Debates and Online Misinformation, Sustainable Cities for Citizens, Demography, Economics & Finance 2.0, Migration Studies, Sport Data Science, Social Impact of Artificial Intelligence and Explainable Machine Learning. SoBigData++ will advance from the awareness of ethical and legal challenges in social mining to concrete tools that operationalise ethics with value-sensitive design, incorporating values and norms for privacy protection, fairness, transparency and pluralism.

> 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.

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.

- **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 policy makers, 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).

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. 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. Nowadays with SoBigData++, as an advanced community built on the solid predecessor project we are addressing the future challenges and goals:

> The project is now entering a new phase with the deployment of new libraries in the e-infrastructure usable through Jupyter notebooks and the possibility of executing them on our SoBigData++ computational network.

- **Wider, simplified,** and more efficient access to the best research infrastructures: SoBigData++ will firstly transform the nascent SoBigData research infrastructure, to create the most comprehensive research infrastructure in large-scale social data mining. Efficient access will be provided through new cloud- and HPC-based components and services, with sustainability and uptake enhanced further through alignment with EOSC. Access will be simplified through the enhanced provision of SoBigData software-as-as-a-service, made further easier to compose through a new workflow language and editor. Access will also be widened very significantly beyond computer science, through incorporation and engagement of new research communities from multiple complementary fields, e.g., social and political science, digital humanities, economics, journalism and media. The core SoBigData computer science partners will help the new partners to include their activities and services in the RI, harmonizing them with the existing platform as well as extending

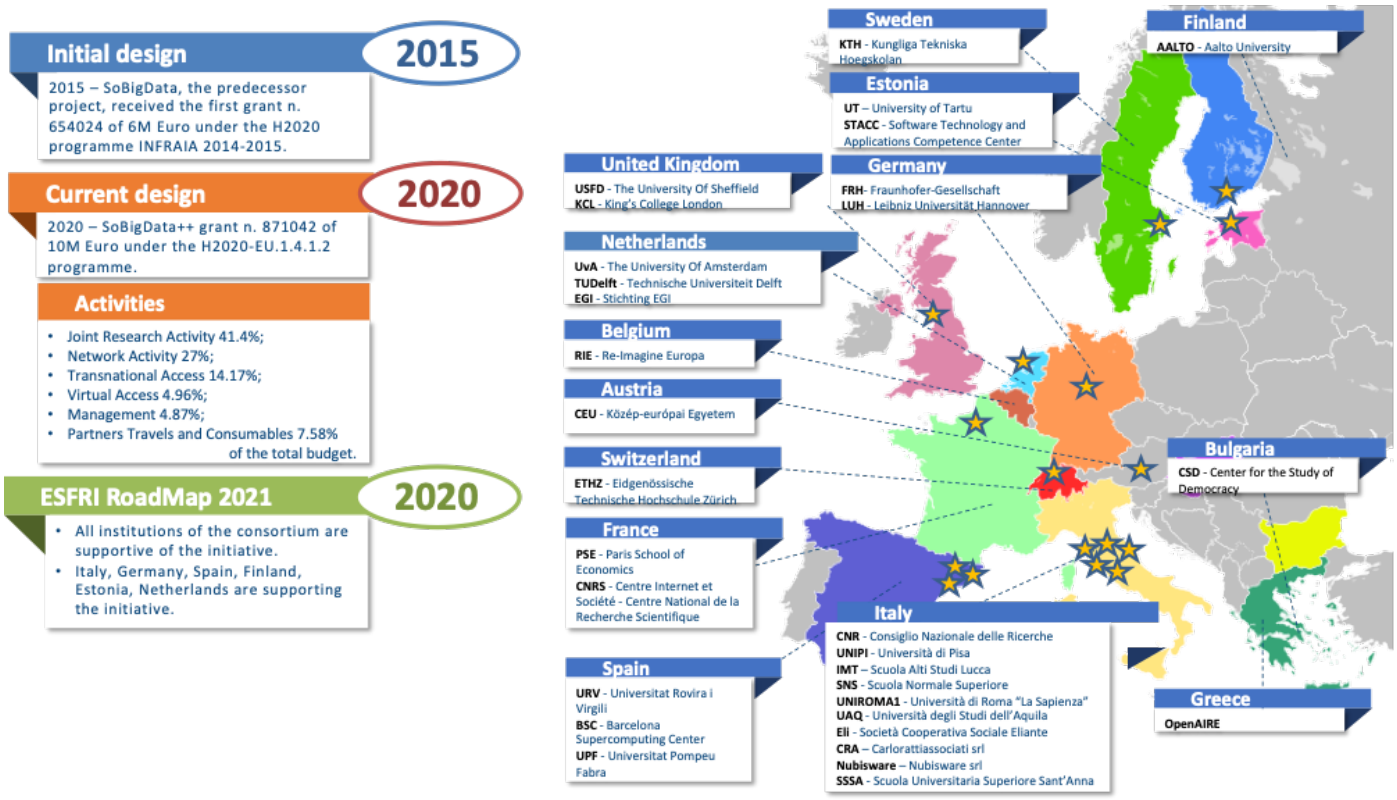


Figure 1. Timeline of the Projects and evolutions

Figure 2. A glimpse of SoBigData++ project architecture

its capabilities. In this way also, other currently disjoint hubs of knowledge will be made accessible to researchers irrespective of their nationality, country of work, and facilities at their organisation. This will benefit in particular researchers from smaller research institutions, newly associated states, and early career researchers, all of whom are currently facing difficulties in competing against researchers working in big European centres and industrial research labs.

• **More advanced research** infrastructure services, enabling leading-edge or multidisciplinary research, used a wider user community: SoBigData++ will advance the existing SoBigData infrastructure from an online repository of data and methods to a platform centred on the design and execution of complex social mining processes. We will provide new, advanced services, based on novel methods arising from cutting-edge research in Big Data analytics and Artificial Intelligence that project partners are pursuing in other projects (i.e. AI4EU, HumaneAI, WeVerify), such as methods for human-comprehensible explanation of complex machine models and methods for detection and analysis of online misinformation. SoBigData++ will widen significantly the existing user community of computer scientists, towards new users from the fields of digital humanities, social and political sciences, digital journalism and media, economics, and beyond. In particular, a new generation of multi-disciplinary data scientists will be trained to exploit in the best way the new research infrastructure, with its wealth of datasets, social data analytics services, and visual research environments.

• **Develop synergies** and integrate complementary capabilities with related infrastructures: The impact of SoBigData++ will be multiplied even further through establishing collaborations and pursuing integration activities with relevant research infrastructures. In particular, OpenAIRE is already a partner in the enlarged consortium, USFD are ensuring alignment and tool interoperability with CLARIN and the European Language Grid (ELG), while DARIAH, EHRI, and RISIS have declared their intention to make use of the enhanced SoBigData++ RI as part of its extension towards their respective communities of digital humanities, social scientists, political scientists, and information scientists.

• **Foster innovation** through partnership with industry: The SoBigData starting grant already established

a strong track record in promoting innovation and knowledge transfer through open science and partnerships with industry. Successful previous and ongoing collaborations include, inter alia, Sky, the BBC, the Press Association, Thompson Reuters, the British Library, the UK National Archives, Amazon, Microsoft, Google, and Skype. These will be strengthened and widened very significantly in SoBigData++ as detailed in WP5 Accelerating Innovation.

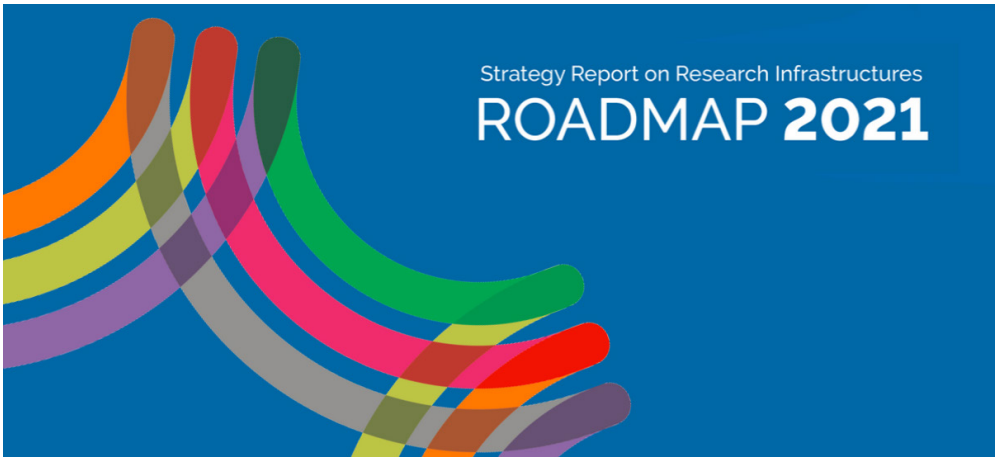
- **Educate a new generation** of researchers to optimally exploit all the essential tools for their research: The transnational and virtual access activities in SoBigData++ will continue to offer new research and pan-European collaboration opportunities, to help develop a new generation of data scientists. Through exposing them to mobility opportunities and transnational collaborations, coupled with making available a rich set of training materials and events, the project will ultimately aim to improve the employment opportunities of European data scientists.

- **Facilitate closer interactions** between larger number of researchers to facilitate cross-disciplinary fertilisations and a wider sharing of information, knowledge and technologies across fields and between academia and industry: Thanks to the transnational access, joint research, training, and innovation activities, truly multi-disciplinary, pan-European cooperation will arise, focused on tackling the hot research challenges in big social data mining. By widening the consortium towards new scientific fields, research communities, and companies in diverse vertical domains, SoBigData++ will stimulate novel multidisciplinary research and cooperation, including sharing of best practices in research infrastructure deployment, data standards and management of large-scale data, provision of high-quality services, and novel services and research experiments. Going beyond data scientists, SoBigData++ with an enlarged consortium (w.r.t. SoBigData), will impact on other research fields, including but not limited to web science, digital humanities, social informatics, and economics. They will benefit from the substantial quantity of data and analytical services integrated within SoBigData++.

- **Integrate knowledge-based** resources (collections, archives, structured scientific information, data infrastructures, etc.) to ensure better management of the continuous flow of data: SoBigData++ will integrate a wealth of new resources, over and above the already provided over 80 social data collections. The project will create not only some new social data collections, but also integrate existing relevant archives and structured scientific information from the OpenAIRE, RISIS, CLARIN and ELG infrastructures/online platforms.

- **Extend and Improve** the Exploratories. From the existing ones, SoBigData++ introduced new vertical environments to include the hot topics in research: Impacts and Explainable AI, Medicine and Digital Health, Economy and Finance, Sport Data Science, Migration Studies, Sustainable Cities for Citizens, Demography 2.0 and Societal Debates. Those Exploratories are the tools for the creation of active communities with the collaboration of top-level researchers in Europe and are able to produce advancement in the fields.

The SoBigData++ project started in January 2020 and went through the Covid-19 pandemic situation in its early stages, the impact on the project is important due the collaborative and community building nature of the project which are more difficult without travels and the possibility of having face-to-face meeting between the partners (especially the new ones) and the large community of users. Anyway, new ways of collaborative working were activated and the usage of the e-infrastructure to exchange ideas gave the possibility of producing results and to move forward with the planed activities. The project is now entering a new phase with the deployment of new libraries in the e-infrastructure usable through Jupyter notebooks and the possibility of executing them on our SoBigData++ computational network. Moreover, the Project submitted the request to become part of the ESFRI and was considered eligible, therefore and will face the last evaluation step in March 2021.



ESFRI Roadmap 2021

An opportunity to guarantee long-term sustainability to our SoBigData Research Infrastructure.

Valerio Grossi, ISTI-CNR National Research Council of Italy | valerio.grossi@isti.cnr.it

Being part of the ESFRI Roadmap 2021 is a fundamental element to guarantee long-term sustainability to the SoBigData Research Infrastructure. Our aim in taking part in this process is simple: we would like the SoBigData Research Infrastructure to become a CERN-level institution, becoming competitive worldwide.

ESFRI, which is the acronym of 'European Strategy Forum on Research Infrastructures', is 'a strategic body established in 2002 by the Council of the European Union to support a coherent and strategy-led approach to policy-making on Research Infrastructures in Europe' as is described on its website. ESFRI Roadmaps are designed to support the best European science facilities. Its process is based on a thorough evaluation and selection procedure. All ESFRI Roadmap projects have proven to be very influential and provided a truly strategic guidance for investments from both Member States and Associated Countries, going well beyond the Research Infrastructure domain [1].

The relationship between SoBigData and ESFRI originates in 2015, at the time of the first SoBigData Horizon2020 project. This relationship has evolved and consolidated with SoBigData++ and, on 9 September 2020, we submitted our proposal to create a new Research Infrastructure within ESFRI. The ESFRI SoBigData Research Infrastructure proposal is the most ambitious project that has ever involved our consortium. The overall cost of the ESFRI SoBigData Research infrastructure is estimated in an excess of 150 million €, which includes both the build-up and operational phase. The preparation phase has started in 2020 and the Research Infrastructure will be operative until 2050, for more than 20 fully operational years. In this new adventure for SoBigData, we have gained political support from three states, Italy, Switzerland and Estonia. Financial support has been granted by the Italian Ministry of Research and the Italian National Research Council (CNR) and ten other European entities, while the consortium comprises 27 partners. The proposal focuses on creating a Central Hub in Italy with ten nodes in the following countries: Netherlands, Estonia, Switzerland, Finland, Sweden, Austria, Germany, France, Spain, and the United Kingdom.

Each ESFRI phase is proving to be an interesting experience. The design phase, which will end in 2020, the first year of the current Horizon2020 SoBigData++ Project. The preparation phase is devoted to the preparation of the financial and legal aspects (for both the central hub and the national nodes). Moreover, this phase comprises designing, experimenting and testing the technical aspects of the permanent research infrastructure. Likewise, it includes the cost of becoming a legal entity (ERIC – European Research Infrastructure Consortium), which will have a registered office in Italy. This phase, which has started in 2020, will terminate at the end of 2024. Next comes the implementation phase, which aims to render operational all the management and legal structures involved with the construction of the Research Infrastructure. Moreover, in this phase, a cost book is to be implemented as is service access, which will be defined and modelled in regard to the SoBigData Research Infrastructure. This phase will start in 2025 and terminate in 2029. The implementation phase comprises the establishment of the Research Infrastructure as an ERIC, software development, the set-up of physical spaces and training related to the implementation of each national node and the central hub. The SoBigData Research Infrastructure mission is expected to retain relevance for at least two decades.

The ESFRI SoBigData project represents an opportunity for creating new working positions, allowing recruitment of high-level staff. Most key staff at the national nodes are already in place within the Research Infrastructure consortium institutions. Furthermore, the central organisation will require an office for administrative, legal, financial, communication and project management support. Scientific leadership will also be key for the services that will represent the core of the SoBigData Research Infrastructure activities.

The uniqueness of SoBigData is represented by its ability to connect in synergy the effort of heterogeneous scientific communities, such as data science and artificial intelligence. The ESFRI SoBigData Research Infrastructure, with its network of prestigious data science nodes, has the ambition and chance to become a strategic resource at a European level in dataset, experiment and research skill and computational resource sharing. Hence it will enhance the comprehension of current societal transformations, including aspects and practices revolving around ethics and privacy issues that these transformations imply. Paraphrasing Steven Hawking's Brief Answers to the Big Questions [2]: the availability of data on its own will not take humanity to the future, but its intelligent and creative use will. The best is yet to come!

[1] https://www.esfri.eu
[2] Hawking, S. Brief Answers to the Big Questions (2018) London, Hodder & Stoughton.

Give more data, awareness and control to individual citizens, and they will help COVID-19 containment

Mirco Nanni, Gennady Andrienko, Albert-László Barabási, Chiara Boldrini, Francesco Bonchi, Ciro Cattuto, Francesca Chiaromonte, Giovanni Comandé, Marco Conti, Mark Coté, Frank Dignum, Virginia Dignum, Josep Domingo-Ferrer, Paolo Ferragina, Fosca Giannotti, Riccardo Guidotti, Dirk Helbing, Kimmo Kaski, Janos Kertesz, Sune Lehmann, Bruno Lepri, Paul Lukowicz, Stan Matwin, David Megías Jiménez, Anna Monreale, Katharina Morik, Nuria Oliver, Andrea Passarella, Andrea Passerini, Dino Pedreschi, Alex Pentland, Fabio Pianesi, Francesca Pratesi, Salvatore Rinzivillo, Salvatore Ruggieri, Arno Siebes, Roberto Trasarti, Jeroen van den Hoven, Alessandro Vespignani

The rapid dynamics of COVID-19 calls for quick and effective tracking of virus transmission chains and early detection of outbreaks, especially in the phase 2 of the pandemic, when lockdown and other restriction measures are progressively withdrawn, in order to avoid or minimize contagion resurgence. For this purpose, contact-tracing apps are being proposed for large scale adoption by many countries. A centralized approach, where data sensed by the app are all sent to a nation-wide server, raises concerns about citizens' privacy and needlessly strong digital surveillance, thus alerting us to the need to minimize personal data collection and avoiding location tracking. We advocate the conceptual advantage of a decentralized approach, where both contact and location data are collected exclusively in individual citizens' "personal data stores", to be shared separately and selectively, voluntarily, only when the citizen has tested positive for COVID-19, and with a privacy preserving level of granularity. This approach better protects the personal sphere of citizens and affords multiple benefits: it allows for detailed information gathering for infected people in a privacy-preserving fashion; and, in turn this enables both contact tracing, and, the early detection of outbreak hotspots on more finely-granulated geographic scale. Our recommendation is two-fold. First to extend existing decentralized architectures with a light touch, in order to manage the collection of location data locally on the device, and allow the user to share spatio-temporal aggregates - if and when they want, for specific aims - with health authorities, for instance. Second, we favour a longer-term pursuit of realizing a Personal Data Store vision, giving users the opportunity to contribute to collective good in the measure they want, enhancing self-awareness, and cultivating collective efforts for rebuilding society.

https://arxiv.org/abs/2004.05222

SoBigData

Covid -19

SoBigData goes virtual!

An overview on the events that took place during the Covid-19 pandemic period. Despite the situation, the SoBigData consortium have found new ways to connect and share teaching, learning and supporting in a variety of events.

Joanna Wright, The University of Sheffield | Joanna.wright@sheffield.ac.uk

SoBigData++ Events have forged ahead despite the Covid-19 pandemic and have found new ways to connect with participants to share teaching, learning and supporting in a variety of events all which help to promote the main project and disseminate its message whilst taking 'social distancing' to a whole new level.

Covid-19 has shaped our existence for many months and changed the way we see and understand the world. Sharing and collaborating in a virtual format has long been entertained by academics and is often the preferred and only method of interacting when working with researchers or experts in another city or even country. However, researchers, scientists and experts have always been able to participate in face to face events such as conferences, workshops, lectures and colloquiums enabling them to network, share findings, make connections and benefit from creative and productive meetings and engage with cutting edge research.

SoBigData++ had many events planned for this year – most of which have had to be reimagined in a virtual format. The various organisers from multiple institutions have had to totally reconfigure their plans continually to keep up to date with the ever-changing international travel restrictions. SoBigData++ would like to thank everyone involved in organising and promoting these events and congratulate them in successfully delivering all the planned events in a fully virtual environment.

Here are just a selection of events that have taken place.

VI SCUOLA NAZIONALE DI CHIMICA DELL'AMBIENTE E DEI BENI CULTURALI

http://www.scuolacabc.it/

This event was one of the few this year that was a face-to-face event as it occurred in February 2020 before the international lockdowns. It was organised by IMT and was held at the University of Siena, Italy.

This school was dedicated to the analysis of general knowledge and, in particular, chemical-environmental knowledge relating to the study of the climate system, the role of anthropic contribution to greenhouse effect and its consequences on the distribution of chemicals, ecosystems and the technosphere.

This event was aimed at Masters and PhD students as well as Post Doc researchers. It attracted approximately 120 participants from various EU Universities. It was a five-day programme encompassing around 20 hours of lectures from experts in the fields of environmental chemistry, ecological economics and sustainability studies. The event was highly successful.

Our thanks and appreciation go to Angelo Facchini (IMT) and other organisers.

REAL-TIME EPIDEMIC DATATHON

https://www.epidemicdatathon.com

ETH Zurich seized the opportunity to work with the real time data of the unfolding Covid-19 pandemic and provide an opportunity for data science researchers to work on a collective open source real-time forecasting challenge.

The Datathon was open to everyone (individuals or teams) and ran from April 2020 to July 2020. The Datathon used publicly available data and encouraged data scientists to contribute to the global open-source scene by releasing real-time epidemic forecasting models.

The event attracted 37 individuals and in line with the project's aim to encourage more female participation in data science, approximately one third of participants were female.

Our thanks and appreciation go to Nino Antulov (ETHZ) and other organisers.

DATA SCIENCE IN TECHNO-SOCIO-ECONOMIC SYSTEMS ONLINE WORKSHOP 2020

https://www.eth-courant-workshop.com

ETH Zurich also ran a Workshop on 10-11 June 2020, which was open to students, academics and practitioners in the field of data science for techno-socio-economic systems and quantitative finance.

The aim was to attract approximately 300 participants. The Workshop actually had 348 registered participants and the event was well received. The organisers adjusted the schedule for the online format from full days to half days to ensure the participants had a more positive experience.

Although the vast majority of the participants were from Europe, the event reached far further with individuals from America, South America, Asia and the Middle East taking part. The majority of the participants were from academia; however, there were also participants from Industry including several CEOs and Company Directors. This demonstrates how the project is reaching out further than academia to individuals who have influence in the financial sector.

Our thanks and appreciation go to Nino Antulov (ETHZ) and other organisers.

EPIDEMICS AND THE CITY: HOW HUMAN MOBILITY AND WELL-BEING CHANGED DURING THE COVID-19 ERA

http://sobigdata.eu/events/epidemics-and-city-how-human-mobility-and-well-being-changed-during-covid-19-era

This was a dissemination webinar that was quickly organised to capture the

threads of the topical Covid-19 crisis and the impact it has had on society from the perspective of Data Science and Environmental Epidemiology. It took place on 3 July 2020.
The webinar was aimed at experts, stakeholders of the SoBigData++ project and it was also open to the general public. Approximately 70 people were involved and the event was well received.
Experts discussed the effect of Covid-19 on mobility, the impact on people's well-being and on virus transmissibility as well as the quality of life of citizens. It also looked at co-benefits in relation to lockdown.
Our thanks and appreciation go to Angelo Facchini (IMT), Luca Pappalardo (CNR) and other organisers.

DATA PROTECTION FOR RESEARCH AND STATISTICAL PURPOSES: TOWARDS LEGALLY ATTENTIVE DATATHONS

This Webinar took place on 22 July 2020 and was organised by Scuola Superiore Sant'Anna in Pisa, Italy as an open event. The program included 3 short talks from experts in the field followed by a Q&A session. The topics covered were 'Data Processing for Scientific Research and Statistics and the SoBigData++ Framework', 'Datathons: Risks & Opportunities' and 'Legally Attentive Datathons: Ready for the Check list'. Two of the three speakers were female following one of the main aims of the SoBigData++ project to showcase females in Data Science to inspire and encourage more females into this research sector. The event attracted 34 participants.
Our thanks and appreciation go to Prof.Giovanni Comandé, Dr. Giulia Schneider and Dr.Denise Amram, all of Scuola Superiore Sant'Anna.

XKDD WORKSHOP (PART OF ECML PKDD) - EXPLAINABLE AI

XKDD Workshop (part of ECML PKDD) - Explainable AI

https://kdd.isti.cnr.it/xkdd2020/
This workshop was the last event of a series of initiatives over a 5 year period organised by UNIPI, CNR and the University of Warsaw. The purpose of XKDD, eXplaining Knowledge Discovery in Data Mining, is to encourage principled research that will lead to the advancement of explainable, transparent, ethical and fair data mining and machine learning.
This event was open to the research community on Machine Learning and Data Mining and attracted 213 participants – far more than had been originally planned for.
There were 2 invited speakers, 7 accepted papers involving numerous authors and the event encompassed a four-hour program. The event went smoothly and generated useful questions and proposals for future research. As a direct result of the workshop there was a commitment made to set up a communication group to work with the research issues which were raised and community building actions that will be implemented.
Our thanks and appreciation go to Salvo Rinzivillo (CNR) and other organisers.

CMF 2020 - COMPLEXITY MEETS FINANCE: DATA, METHODS AND POLICY IMPLICATIONS (Part of NetSci 2020 17-25 Sept 2020)

https://sites.google.com/view/cmf20/home
This satellite aimed to bridge the gap between the fields of complex networks theory and finance by bringing together experienced researchers and young scholars interested in interdisciplinary research to discuss state-of-the-art work, share knowledge and create opportunities for novel and fruitful collaborations.
The event also intended to bring cutting-edge academic research in contact with industry experience and impact. Therefore CMF 2020 was open to researchers, scholars, industry stakeholders and also policy makers demonstrating SoBigData++'s aim to reach out and mesh together the academic realm with policy makers and the economic world.
The event attracted just under 60 participants.
Our thanks and appreciation go to Fabio Saracco (IMT) and other organisers.

ROME II - REDUCING ONLINE MISINFORMATION EXPOSURE (Part of NetSci 2020 17-25 Sept 2020)

https://sites.google.com/imtlucca.it/rome-ii/home
The aim of the ROME II satellite is to convey state of the art research on

the analysis and comprehension of the information system of OSNs (Online Social Networks).
OSN's role in shaping the political debate is crucial: misinformation has been demonstrated to distort and divert public discourse. The spread of hoaxes, propaganda, and rumours has an impact on different areas of social interest, such as political elections, public health and social tensions. It is therefore necessary to develop the proper tools in order to detect the various facets with which the false information spreads on the web and to measure its effect and pervasiveness.
This Satellite was aimed primarily at academia – in particular - PhD Students and researchers. It attracted approximately 30 participants. As with many other events, speakers were asked to pre-record their presentations and be online during the satellite to answer questions and enter into discussions.
Our thanks and appreciation go to Fabio Saracco (IMT) and other organisers.

THE 7TH SATELLITE ON QUANTIFYING SUCCESS (Part of NetSci 2020 17-25 Sept 2020)
http://www.onurvarol.com/netsci20-qs7/
This satellite took place on 17 September 2020.
The event was aimed at computational social scientists interested in understanding the relationships between performance and success in several contexts, from scientific publication to sports, cinema and writing.

The aim was to bring together scientists and researchers from different disciplines and give them the opportunity to present and discuss their recent results, identify open questions and new challenges, and develop common languages for solving problems in the emerging, fascinating field of the "science of success". It was an opportunity to discuss different, multi- and inter-disciplinary approaches to quantify success across a variety of scientific domains, with a main focus on data-driven methods. The event attracted approximately 30 participants.
Our thanks and appreciation go to Luca Pappalardo (CNR) and other organisers.

"More than 70 attendees of 17 countries, representing national statistical offices, universities, research centers and companies, shared their last contributions on privacy models, microdata protection, protection of statistical tables, protection of interactive and mobility databases, record linkage and alternative methods, synthetic data, data quality, and case studies. The conference has been carried out without incidences and, given the circumstances, we are very glad of the high level of achieved interaction between the attendees, that is one of the main goals of organizing a conference. Despite the organization success, we hope the next edition PSD2022 will be hold face-to-face." *Jesús Manjón |URV | Organizer PSD 2020*

PSD 2020 - PRIVACY IN STATISTICAL DATABASES
https://unescoprivacychair.urv.cat/psd2020/
The SoBigData++ sponsored event PSD 2020 went ahead on 23-25 September. This event takes place every 2 years and this year was converted to an online conference.
The event is aimed at researchers and statistical agencies and although the conference originated in Europe, it is open to all and contributions and attendees from across the globe are welcomed. This year's event was planned for approximately 40-50 attendees and contrary to expectations, the event attracted 72 participants.
The organisers were able to maintain the structure of the event, albeit losing the social aspects, and every speaker had the opportunity to pres-

ent their paper and answer questions from the audience. The organisers were also able to utilise a feature of Zoom that enabled them to set up parallel rooms for some attendees so that they could engage in more detailed discussions about a specific presentation. This facility turned out to be an advantageous benefit of holding the conference online.
Our thanks and appreciation go to Jesús Manjón (URV) and other organisers.

CAN BIG DATA BRIDGE GAPS IN MIGRATION STATISTICS?

This webinar was organised by UNIPI and CNR and was born from the HumMingBird project which looks into human migration. It took part on 29 September 2020.
Traditional statistical data on international migration suffers from the problems (gaps) of inconsistency in definitions, differences in geographical coverages, absence of reasons for migration, timeliness and limitations in demographic characteristics.
Although there is a novel list of potential data sources that could provide valuable, real-time insights, these remain largely untapped for the time being. There are sensitivity obstacles, legal issues, availability, accessibility, purpose of the data, etc. However, improving migration data is a crucial step to improving migration governance since better data is needed in order to bring about sustainable social and economic development and national migrant data strategies are needed to inform good policies. This talk discussed the existing gaps and shortcomings of the migration statistics and the potential utilisation of Big Data analytics for bridging these gaps.
It was expected that this event would attract participation of around 30 people; however, 45 individuals accessed the webinar – an increase of 50% on plan, demonstrating the success of the project's outreach. Furthermore, over a third of the participants were female demonstrating an improving ratio between the genders. The event was aimed at experienced and non-expert researchers and academics interested in studying human migration, but coming from different research fields, e.g., sociology, computer science, and demography. The organisers were very satisfied with the level of participation and there were various interesting questions and discussions that arose at the end of the presentation.
Our thanks and appreciation go to Laura Pollacci (CNR & UNIPI) and other organisers.

The SoBigData Project++ also offers support in alternative ways to students and the online element of these means they can be attended by individuals who are not geographically near and can include experts from anywhere in the world. An example of this is detailed below:

DATA SCIENCE COLLOQUIUM

https://datasciencephd.eu/events/data-science-colloquium-2020
20 May – 8 June
Organised by SNS, CNR, IMT & UNIPI. The Colloquium consisted of 2 hour sessions, 3 times a week from 20 May to 8 June. It included seminars held by professors and 3rd year PhD students to support and guide 1st year students through their research projects. The sessions were also open to any interested party.
Our thanks and appreciation go to Dino Pedreschi (UNIPI), Ioanna Miliou (UNIPI) and Luca Pappalardo (CNR) and other organisers.

It is perhaps not a surprise that the modern world can find ways around face to face meetings and technology is such that geography doesn't dictate involvement. However it is still impressive that a whole range of events have been successfully reformatted to work around the pandemic and, rather than hosting reduced numbers, has often meant more attendees have had the opportunity to become involved. This shows extensive and impressive organising and also demonstrates the resilience of the students, researchers and experts to find a way to 'meet', discuss, share, question and provide feedback to each other's work and research. The whole SoBigData++ community has come together from all over the world and can rightly feel proud of their achievements during this unprecedented time.
Moving forwards we will certainly take what we have learned from this time to make sure future face to face events also contain a strong element of virtual participation for those that require it. For most people, there is no substitute for being present at conferences, and making social connections is arguably just as important for collaboration. However, the virtual aspect adds another dimension that can be exploited for its benefits (e.g. removing geographical or financial barriers, being ideal for individuals who find it hard to travel or mix with others, or individuals with time/family pressures that mean conferences/workshops are usually out of the question). The pre-recorded talks will prove extremely useful for reassessment, a second listening to deepen understanding and can also be shared with other groups at other times as appropriate. This can only help spread the message further and wider and encourage more discussion and collaboration within the community.
We are confident that, no matter how long the pandemic affects our lives and continues to inhibit travel, the SoBigData++ community will only become stronger for its resilience and determination to forge ahead and push through the challenges that the Covid-19 pandemic has presented.

Real Time Epidemic Datathon

Epidemics + Hackaton + Data as open-source real time forecasting challenge that aims at joining forces to develop real-time and large-scale epidemic forecasting models.

Nino Antulonov, ETHZ | nino.antulov@gess.ethz.ch

After the initial spread of SARS-CoV-2 in Hubei (China), the World Health Organization declared the coronavirus disease COVID-19 a pandemic on 11 March 2020. The goal of the Epidemic Datathon which started in March 2020 and ended in July 2020 was to use publicly available data to accelerate scientific innovation in modeling and forecasting the evolution of COVID-19 cases in different countries as well as to evaluate possible response measures. Quantifying the forecast accuracy and uncertainty in real time is used as a benchmark for epidemic forecasting models and hopefully can provide additional insights into the COVID-19 outbreak dynamics. Finally, in the planning phase of our Epidemic Datathon we aimed to contribute to the global open-source scene by releasing real-time epidemic forecasting models.

SCIENTIFIC MOTIVATION

Are mechanistic epidemic models able to make good predictions or do purely data-driven approaches outperform standard epidemiological frameworks? Better performances of data-driven models that incorporate various datasets may help to determine missing features in standard epidemic models. Large deviations in the predictive accuracy of standard epidemic models can indicate wrongly estimated disease parameters and containment strategies.

We encouraged participants to get inspired by the state-of-the-art research in epidemiology, network science, statistics, data science, and other fields and make scientific contributions. We also constructed a disclaimer and ethics section to integrate information provided to the participants.

EPIDEMICDATATHON OUTCOMES

The event, in its live phase, attracted 40 individuals divided in 8 teams – and approximately one third of participants were female. Our starting point was the observation that simple data-driven methods can outperform mechanistic epidemic model predictions, highlighting a need for policymakers, domain experts, and researchers to be aware of possible limitations of common epidemic forecasting frameworks.

Live dashboard -- https://submit.epidemicdatathon.com/#/dashboard
We deployed baseline epidemic forecasting models for providing real-time predictions. By selecting a specific country, one could observe historical epidemic forecasts along with real-time updates.

During the period from March – June 2020, 8 teams with 40 persons were continuously working on epidemic forecasting tasks as part of Epidemic Datathon. They have used different classes of models: (i) Time-series: Logistic Growth model, Linear Regression models, Parametric Curve Fitting, AR, ARIMA, Exponential Smoothing; (ii) Epidemic models: SIR, SEIR, SIRD; (iii) Machine Learning: Logistic Regression, Neural Networks, LSTM, Random Forest and others.(fig.1).

STEFANELLI'S TEAM (INDIVIDUAL PARTICIPANT) ABOUT EPIDEMIC DATATHON

Stefanelli's team used an ARIMA-based time-series model to predict the evolution of COVID-19 case numbers.

Marcello Stefanelli (33, individual participant) obtained a degree in Economics and Statistics at Università degli Studi di Pavia. In the following paragraphs, Marcello describes his motivation to participate in Epidemic Datathon and his forecasting approach in more detail.

" I'm a Business Strategy Manager in a global consulting firm. I'm passionate about public and private health care systems management and of course data science.
The approach I applied topredict daily time series data is based on the Autoregressive Integrated Moving Average (ARIMA) model, a classical statistical model for analyzing and forecasting time-series data. As highlighted by Columbia University Mailman School of Public Health, the ARIMA model finds various applicationsin epidemiology, ranging from patient zero detection to the evaluation of population-level health interventions. .
However, there are some constraints that we have to take in account: (i) ARIMA models do not predict rare "black swan" events. (ii) Data samples must be consistent; the performance

Solutions from different teams -- https://github.com/ninoaf/baseline_epi_predict.git

Team name	Link to team solution
CFRSS	https://bit.ly/2K9kYwZ
GNTM	https://bit.ly/2LfGgcV
MAE	https://bit.ly/33WgnVY
PandeML	https://bit.ly/3qJQQJM
Quaranteam	https://bit.ly/39UhCJe
QuaranteedSucess	https://bit.ly/3ozwnVY
Stayhome	https://bit.ly/3guoYEL
ToBeAnMa	https://bit.ly/39PC65Y

Fig.1 - Stefanelli's team (individual participant) about epidemicdatathon

of the model could be biased with a reduced number of observations, which may happen in the early stages of an epidemic.

I started adopting the ARIMA model with the aim of assessing the existence of some degree of correlation between COVID-19 case numbers and daily economic variables' volatility (e.g., stock market volatility transmission modeled with GARCH time series analysis)."

For more details, check the full blog post on *https://www.epidemicdatathon.com/post/bottom-up-approach-in-forecasting-covid-19-outbreak-adopting-arima-time-series-model*

Selected scientific publications related to Epidemic Datathon

Böttcher, Lucas, and Nino Antulov-Fantulin. "Unifying continuous, discrete, and hybrid susceptible-infected-recovered processes on networks." Physical Review Research 2.3 (2020): 033121.

Böttcher, Lucas, Mingtao Xia, and Tom Chou. "Why case fatality ratios can be misleading: individual-and population-based mortality estimates and factors influencing them." Physical Biology 17.6 (2020): 065003.

Böttcher, Lucas, D'Orsogna, Maria R., and Tom Chou, Using excess deaths and testing statistics to improve estimates of COVID-19 mortalities, preprint

Organisation
Official datathon for the ETH Zurich class: 851-0585-38L Data Science in Techno-Socio-Economic Systems & EU SoBigData++ datathon.

Organising committee:
Nino Antulov-Fantulin (Computational Social Science, ETH)
Dirk Helbing (Computational Social Science, ETH)
Lucas Böttcher (Computational Medicine, UCLA & Institute for Theoretical Physics, ETH)
Zhang Ce (DS3-LAB, Computer Science, ETH)
David Dao (DS3-LAB, Computer Science, ETH)

Advisory board:
Hans Gersbach (Macroeconomics: Innovation and Policy, ETH)
Christopher Dye (Oxford Martin School, University of Oxford)
Dino Pedreschi (University of Pisa, Italy)
Fabrizio Lillo (Department of Mathematics, University of Bologna, Italy)
Mile Sikic (A*STAR Genome Institute of Singapore, FER University of Zagreb)
Petter N. Kolm (NYU Courant, USA)

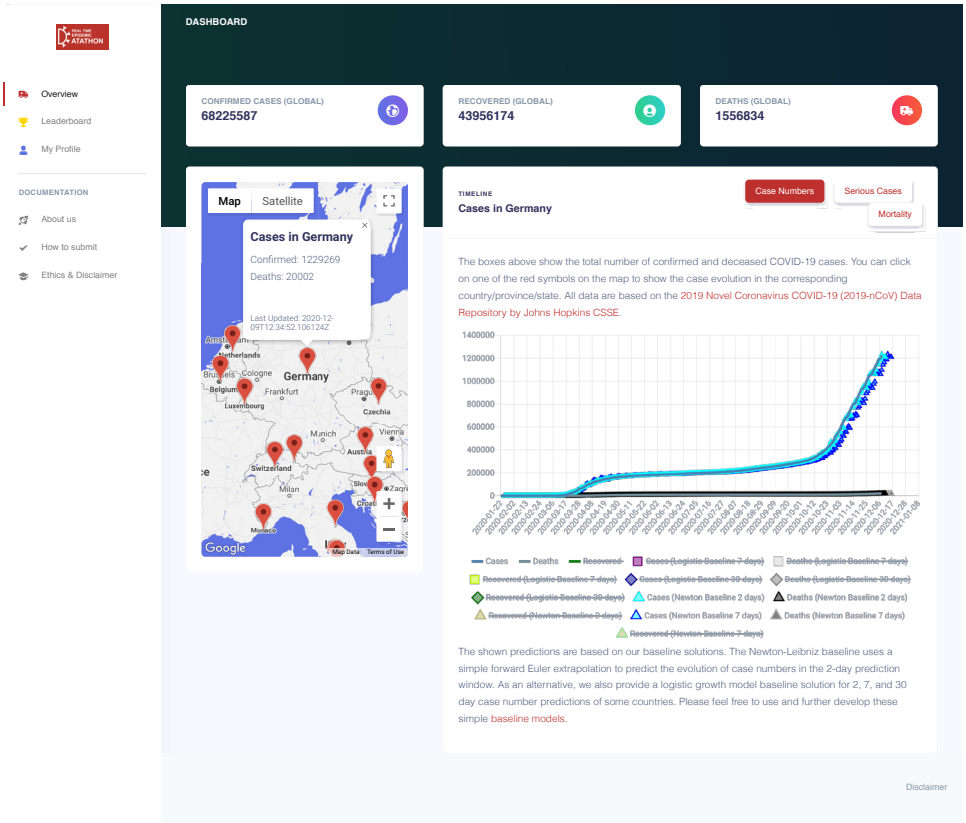


Fig. 2 - Live dashboard -- https://submit.epidemicdatathon.com/#/dashboard

EGI Conference 2020: SoBigData Research Infrastructure

The workshop, organized at the EGI Conference, gave an overview of the important aspects needed in the design and implementation of a distributed research infrastructure for big social data analytics such as SoBigData RI.

Andrea Manzi, EGI Foundation | andrea.manzi@egi.eu

Iulia Popescu, EGI Foundation | iulia.popescu@egi.eu

The EGI Conference is an annual meeting organised by the EGI Foundation (L1) to gather scientists, representatives of scientific communities, service providers, data providers, scientific software/platform developers orbiting around the EGI e-infrastructure.

The EGI Federation is an international e-infrastructure set up to provide advanced computing and data analytics services for research and innovation.

The EGI Conference [L2] took place virtually from 2 to 4 November and attracted more than 750 participants from 52 countries (see fig 2). The SoBigData workshop has been chaired by *Beatrice Rapisarda (CNR)* and has started with an introduction of the Sobigdata++ project by *Roberto Trasarti (CNR)*. The new project has as main objective the consolidation and enrichment of the research infrastructure, delivered in the context of SoBigData project, for the design and execution of large-scale social mining experiments accessible seamlessly on computational resources from the European Open Science Cloud (EOSC) and on supercomputing facilities. SoBigData++ integrates a community of 31 key excellence centres at a European level in Big Data analytics and social mining. Roberto presented the plan for SoBigData to become a research infrastructure recognized by ESFRI RoadMap 2021 and sustained by a ERIC legal entity.

The workshop continued with the talk from *Luca Pappalardo (CNR)*, who discussed how big data originating from the human digital activities is offering an opportunity to scrutinize the ground truth of individual and collective behaviour at an unprecedented detail and at a global scale. In particular, he highlighted the activities of the explanatories within the SobigData++ project as the means to create and integrate new services within the infrastructure.

The EGI Foundation is a member of the SoBigData++ consortium, dealing with coding and workflow systems integration in the infrastructure, and has invited members of the SoBigData community to organize a workshop to describe the main highlights of the

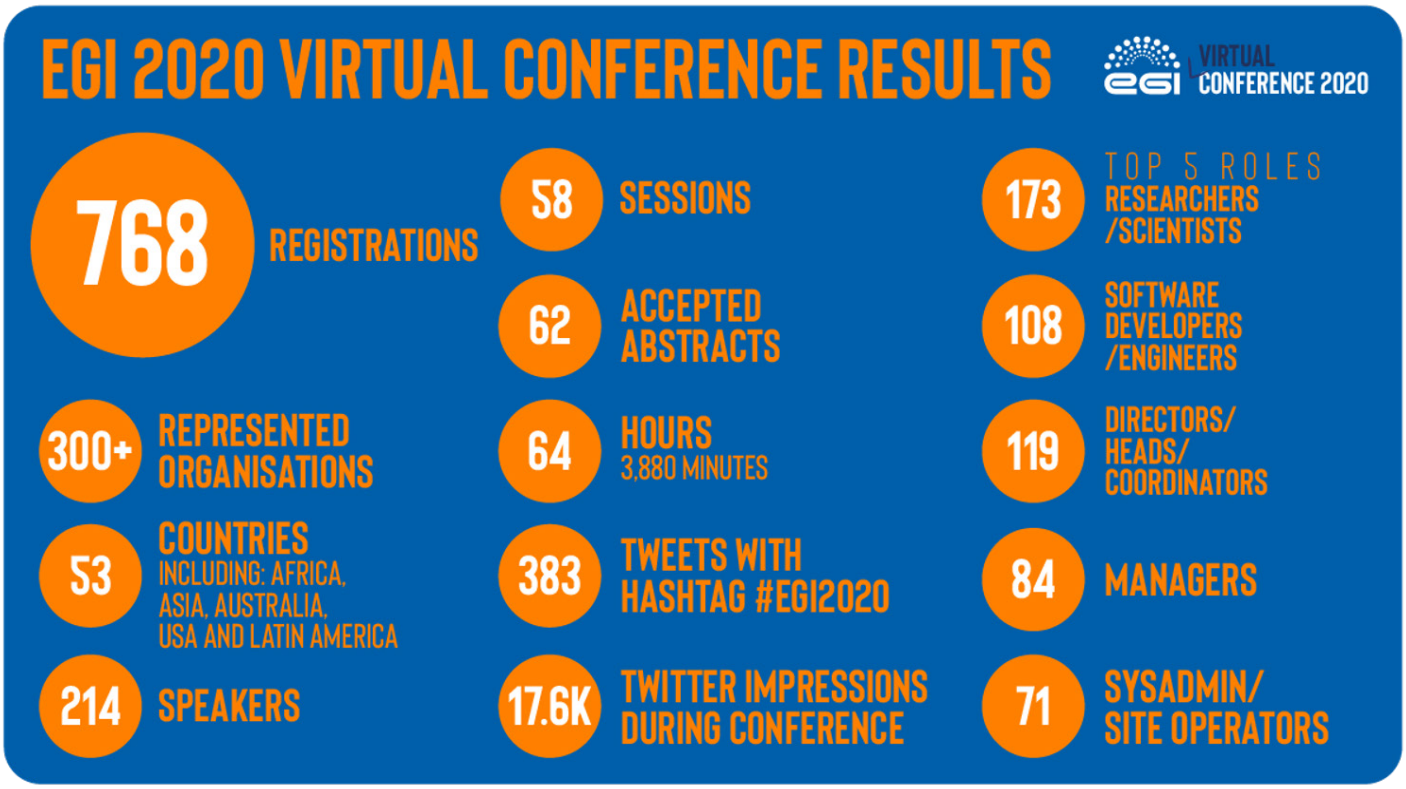
The following talk by *Francesca Pratesi (CNR)*, was focused on the issue of privacy and ethics when dealing with social data. Francesca highlighted the need for strategies that allow the co-existence between the protection of personal information and fundamental human rights together with the safe usage of information for scientific purposes by different stakeholders with diverse levels of knowledge and needs.

The workshop continued with a talk from *Dr. Mark Coté (King's College London)*. According to Coté, 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. Therefore, one of the main aspects of the SoBigData infrastructure resides on the training area for data scientists which is integrated in the catalogue of the infrastructure. The SoBigData++ project will further promote training materials and has already organized and will organize online events and datathons and will perform improvements in the e-Learning area.

Finally, *Paolo Ferragina (University of Pisa)* presented TagMe, one of the success stories of the SoBigData infrastructure, which boosted its popularity and usage. The system is a powerful tool that is able to identify on-the-fly meaningful short-phrases in an unstructured text and link them to a pertinent Wikipedia page in a fast and effective way. The tool once integrated in the SoBigData infrastructure has seen highly relevant peaks of daily accesses (1 billion of accesses since 2016). The future plan is to develop more tools in SoBigData++ that are similarly simple and effective to use.

The workshop originally included a round table with participants, in order to discuss in detail the abovementioned projects, but unfortunately due to time constraints it was not possible to host it. The EGI Foundation would like to thank all the speakers and the chair for their availability and for the interesting talks and it's looking forward to collaborating for the success of SoBigData++.

For more info see:
https://www.egi.eu
https://indico.egi.eu/event/5000/overview



SocInfo 2020: a virtual edition of the Social Informatics conference

The 2020 edition of the Social Informatics conference, was held virtually between 6 and 9 October 2020. The conference proceedings were published by Springer.

Marco Braghieri, King's College London, marco.braghieri@kcl.ac.uk

The Social Informatics conference held its 12th edition virtually, between 6 and 9 October 2020. With over 120 registered participants, the event was in general well-received, and keynote speakers attracted over a hundred virtual attendees as their speeches were part of the 9th edition of the Internet Festival which was held between 8 and 11 October 2020.

SocInfo 2020 hosted four great keynote speeches, from Dr. Leticia Bode, Provost Distinguished Associate Professor at Georgetown University, Washington DC, USA; Professor Virginia Dignum from the Department of Computing Science at Umeå University, Sweden; Professor Alessandro Vespignani, Director of Network Science Institute, Northeastern University, Boston, USA and Dr. Bruno Lepri, head of the Mobile and Social Computing Lab at Bruno Kessler Foundation, Italy. The conference program included a number of workshops and tutorials were held with an average of 15 to 30 individuals, whereas regular sessions were attended by between 45 and 60 participants.

Initially designed as an in-person conference, SocInfo2020 was then transformed into a virtual event due to the impact of the COVID-19 disease, which was declared a pandemic by the World Health Organization on 11 March 2020. Despite the difficulties brought by the ongoing pandemic, SocInfo2020 managed to adapt its organisation and scientific production, centered on complex social systems using computational methods, or explore socio-technical systems using social sciences methods.

Before the conference, a call for papers was issued in order to produce a volume of conference proceedings [R1], which was published by Springer [L2] during the conference. SocInfo2020 received 99 papers 294 authors from 37 different countries. A key element of pre-conference work was carried out by committee members: 27 senior program committee members along with 152 program committee members and the steering commitee, representing a diverse and broad interdisciplinary background. At the end of the double-blind peer review process, 30 full and 3 short papers were selected to be part of the proceedings, along with 14 submissions that were featured as posters during the conference.

During the conference, a number of awards were given to the best paper, best runner-up paper, best poster and best reviewer. The best paper was 'Social Capital as Engagement and Belief Revision' by Gaurav Koley, Jayati Deshmukh and Srinath Srinivasa. The best runner-up paper was 'Facebook Ads: Politics of Migration in Italy' by Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, Andre Panisson and Daniela Paolotti. The best paper was 'Who Won it Online? A comparative study of 2019 Indian General Elections on Twitter' by Kanay Gupta, Avinash Tulasi, Omkar Gurjar, SathvikSanjeev Buggana, Paras Mehan, Arun Balaji Buduru, Ponnurangam Kumaraguru, whereas the best reviewer award was given to Ryota Kobayashi.

A number of individuals collaborated as chairs in order to grant that SocInfo2020 and are all listed here [L1]. Moreover, the conference was held thanks to the support of many sponsors, such as SoBigData++, XAI, AI4EU, HumMingBird, WeVerify, NoBias, HumanAI, the Internet Festival, and Springer for providing generous support. A special thanks to Asti for their help with logistics and planning.

The 33 papers featured in the proceedings contribute to the scientific contribution developed by SocInfo2020, which despite all difficulties related to the ongoing pandemic, managed to be successful in engaging a significant number of participants and feature insightful works on complex social systems using computational methods, or explore socio-technical systems using social sciences methods.

Links:

[L1] https://kdd.isti.cnr.it/socinfo2020/index.html
[L2] https://www.springer.com/gp/book/9783030609740

References:

[R1]: Social Informatics, 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings, Editors: Aref, S., Bontcheva, K., Braghieri, M., Dignum, F., Giannotti, F., Grisolia, F., Pedreschi, D. (Eds.)

Proposal for a regulation on AI: our feedback

LIDER-Lab Scuola Superiore Sant'Anna provided a feedback to the EU Commission on the Proposal for a Regulation of the European Parliament and the Council laying down requirements for Artificial Intelligence.

Denise Amram, LIDER-Lab, Scuola Superiore Sant'Anna (Pisa-Italy), denise.amram@santannapisa.it

Giovanni Comandé, LIDER-Lab, Scuola Superiore Sant'Anna (Pisa-Italy), giovanni.comande@santannapisa.it

Authors describe the legal feedback provided to the EU Commission on the "Proposal for a Regulation of the European Parliament and the Council laying down requirements for Artificial Intelligence".

The EU Commission launched an inception impact assessment aimed at assessing multiple sets of legislative tools as possible options in order to address the risks linked to the development and use of certain AI applications [L1]. These sets of options ranged from an exclusively "soft law" approach (i.e. not binding rules) to comprehensive EU-level legislation initiatives. The main purpose is to shape the future digital age for Europe in order to make it a global leader within a sustainable technological innovation process.

Under the WP2 activities of SoBigData++ project, we provided a feedback to the EU Commission on such a "Proposal for a Regulation of the European Parliament and the Council laying down requirements for Artificial Intelligence" [R1].

Our remarks focused on two main issues. Firstly, we provided operational tools to link the ethics and the legal dimension of a Trustworthy AI avoiding risks of the so called "ethics washing" phenomenon [R2]. Secondly, we highlighted the role that the EU Regulation 2016/679 (General Data Protection Regulation, hereinafter "GDPR") may play to achieve the purposes of the EU Strategy on Artificial Intelligence[R3]. These two grounds of analysis brought us to suggest a multilevel legal framework solution that may include a General Regulation on Artificial Intelligence and specific safeguards both in terms of national and domain legislation, as well as in terms of soft law instruments.

In our feedback, we argued in particular in favor of Option 4 (i.e. to develop any legislative options taking into account the different levels of risk that could be generated by a particular AI application), using Option 3.c (i.e. to adopt EU legislative instrument establishing mandatory requirements, covering all AI applications) of the EU Commission Proposal [R1] as a basis for a General AI Regulation able to bridge lawfulness, robustness and ethics concerns in AI while remaining flexible enough to avoid roadblocks to innovation and uptake for AI solutions. In addition, we briefly proposed some contents for a General AI Regulation, according to a risk-based oriented system of check and balance aimed at ensuring the development of any AI-solutions in light of fundamental rights protection.

From this perspective, a new Regulation on AI shall consider the peculiarities emerging within the different domains and therefore provide the opportunity for each "AI controller" [R2] to allocate tasks, roles, and responsibilities. Under this system, independent authorities shall provide assistance and promote awareness and trustworthiness among data subjects/end-users/stakeholders.

From this feedback, we confirmed the impactful role played by SoBigData++ activities to boost the cultural and inclusive challenge that the AI is driving within the current society.

Links:
[L1]: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements
[L2]: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F551050
[L3]: https://www.lider-lab.sssup.it/lider/notizia/proposal-for-a-regulation-of-the-european-parliam

References:
[R1] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements.
[R2]: Giovanni Comandé. Unfolding the legal component of trustworthy AI: a must to avoid ethics washing. In Annuario di diritto comparato, ESI, 2020, forthcoming (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690633)
[R3]: Denise Amram. The Role of the GDPR in Designing the European Strategy on Artificial Intelligence: Law-Making Potentialities of a Recurrent Synecdoche. Opinio Juris in Comparatione, [S.l.], jul. 2020. ISSN 2281-5147. Available at: http://www.opiniojurisincomparatione.org/opinio/article/view/145/153

Visual Analytics for Data Scientists

Presents the main principles, techniques and approaches of visual analytics in a practice-oriented way

Natalia and Gennady Andrienko, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Germany

Georg Fuchs, Big Data Analytics and Intelligence division, Fraunhofer IAIS, Germany

Aidan Slingsby, University of London, UK

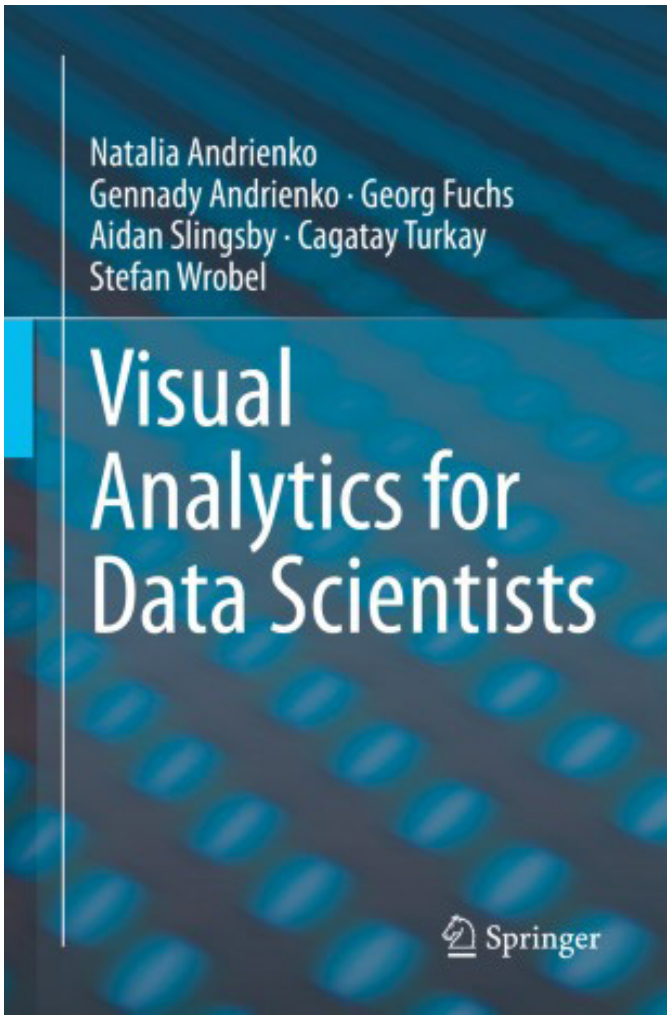
Cagatay Turkay, Centre for Interdisciplinary Methodologies, University of Warwick, UK

Stefan Wrobel, University of Bonn

This textbook presents the main principles of visual analytics and describes techniques and approaches that have proven their utility and can be readily reproduced. Special emphasis is placed on various instructive examples of analyses, in which the need for and the use of visualisations are explained in detail.

The book begins by introducing the main ideas and concepts of visual analytics and explaining why it should be considered an essential part of data science methodology and practices. It then describes the general principles underlying the visual analytics approaches, including those on appropriate visual representation, the use of interactive techniques, and classes of computational methods. It continues with discussing how to use visualisations for getting aware of data properties that need to be taken into account and for detecting possible data quality issues that may impair the analysis. The second part of the book describes visual analytics methods and workflows, organised by various data types including multidimensional data, data with spatial and temporal components, data describing binary relationships, texts, images and video. For each data type, the specific properties and issues are explained, the relevant analysis tasks are discussed, and appropriate methods and procedures are introduced. The focus here is not on the micro-level details of how the methods work, but on how the methods can be used and how they can be applied to data. The limitations of the methods are also discussed and possible pitfalls are identified.

The textbook is intended for students in data science and, more generally, anyone doing or planning to do practical data analysis. It includes numerous examples demonstrating how visual analytics techniques are used and how they can help analysts to understand the properties of data, gain insights into the subject reflected in the data, and build good models that can be trusted. Based on several years of teaching related courses at the City, University of London, the University of Bonn and TU Munich, as well as industry training at the Fraunhofer Institute IAIS and numerous summer schools, the main content is complemented by sample datasets and detailed, illustrated descriptions of exercises to practice applying visual analytics methods and workflows.



https://www.springer.com/gp/book/9783030561451

Knowledge Sharing and Network Dynamics in a European Research Project setting: the Case Of Sobigdata++

The goal of the contribution is to expand the current knowledge about two topics that in the last decades have been growing in importance: knowledge sharing and network theory. Organisational studies are exploring the power of connecting different minds in a shared environment and let them work on intangible assets to create value for customers and the research community. In order to achieve that, SoBigData++ consortium has been used as a case study.

Nicola Del Sarto, SSSA Scuola Superiore Sant'Anna | nicola.delsarto@santannapisa.it

Alessandro Marchesin, SSSA Scuola Superiore Sant'Anna | alessandromarchesin@gmail.com

The literature on this topic is split into three main pillars: big data, network theory and knowledge management.

First, big data is shaping almost every human life: volume, variety and value of data are increasing exponentially, letting businesses take critical decision on a more substantial basis. Five concepts are essential to unlock data value: leadership, human resources, technologies, information systems and culture.

Network theory is an essential key to understand the inner working of modern economies: it can describe several everyday dynamics as the interaction of a network's nodes. Network theory was born in 1736, with the 'Kalingrad Bridge' problem, and it has been developed significantly in the last century with the introduction of the concept of non-randomness, centrality, scale-free, flow direction, and so on.

Lastly, knowledge management is a topic that has been on the raise in the last decades, due to the growth of knowledge workers and the value of intellectual-capital-driven firms. According to knowledge management, there are two kinds of knowledge: explicit and tacit one. The first can be expressed in written forms while tacit knowledge is often called sticky: it is highly personal, it is impossible to write or teach in its entirety. Knowledge creation is a process, in which there is a continuous interaction between explicit and tacit knowledge that leads to a virtuous cycle of wisdom generation.

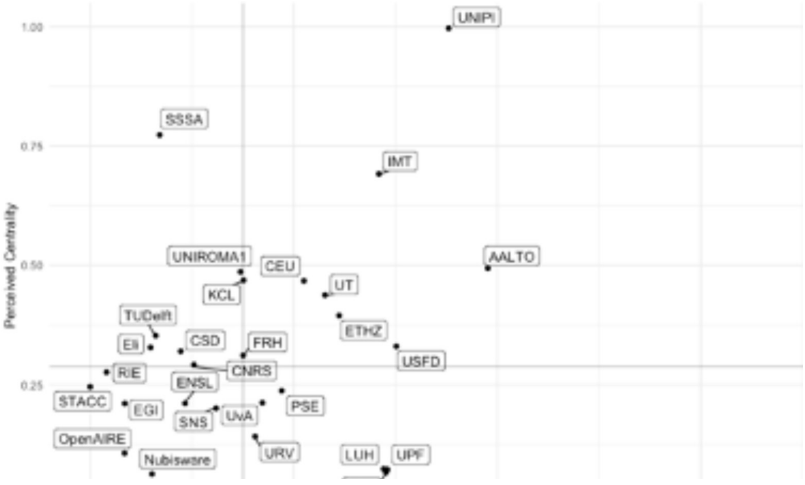


Fig. 1 - The participants plotted on a cartesian plain, with the two measure of centrality on the axes

METHODOLOGY

Data has been collected through a survey regarding consortium knowledge management practices and network's linkage formation and then elaborated.

A network analysis was performed, considering the centrality of the nodes. Two measures of centrality were created: the one coming from the role planned from the project and the one created from the impressions of the other participants. Crossing these two measures, the participants have been put into four centrality categories. In the follow image it is possible to see the participants plotted on a cartesian plain, with the two measure of centrality on the axes (fig.1).

The result is that organisations can fall in four different categories of centrality:

1. The ones formally and subjectively central

2. The ones formally central but subjectively secondary
3. The ones formally secondary but subjectively central
4. The ones formally and subjectively secondary.

The second part of the analysis regarded knowledge management processes: using Nonaka's theory of knowledge creation, it was possible to create a clear picture of how participation in the project influences internal knowledge management processes.
Thirdly, the competences brought from everyone has been measured in terms of uniqueness, by analysing how frequently they appear, and then put together at partner level. At partner level, the procedure has been to select the maximum and minimum score of competence held by each one.
Then each partners' uniqueness score has been plotted against their centrality score and their knowledge management characteristics (fig.2).

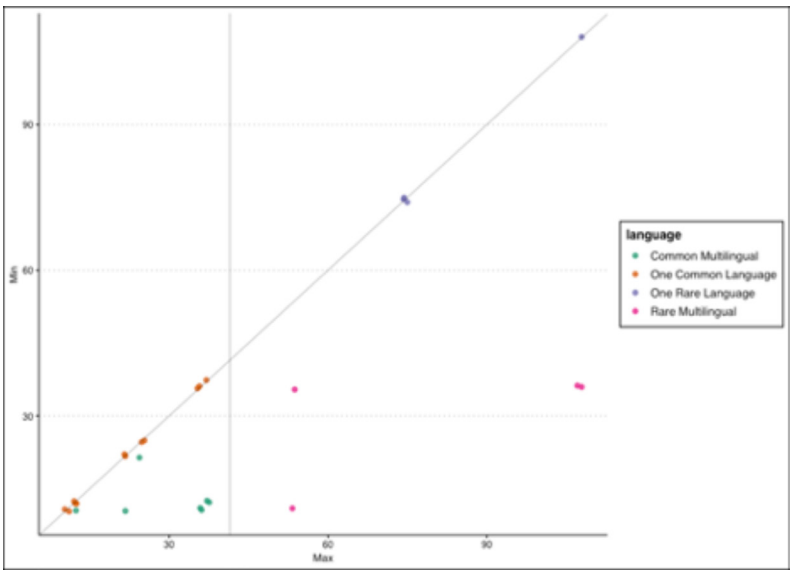


Fig. 2 - Partners' uniqueness score

resources and physical resources, while it is not one hundred per cent clear whether inner culture is strong enough to ensure correct behaviour among the members of the organisation.

In the second place, following Barabasi suggestions, the consortium seems to have some characteristics of a scale-free network: when generating a histogram of both centralities, it is possible to recognise a Poisson distribution shape of the curve. One of the follow-ups of this research would be to demonstrate whether it is a real Poisson distribution mathematically (fig.3).
Another point to regards Powell and Grodal's Proposal, where medium levels of knowledge codification enhance the probability of knowledge sharing, while when knowledge is too codified or too implicit, the incentives to share are low. According to the results it may be inferred that highly central entities are highly connected organisation because of their above-the-average level of codification, while the other participants have below-the-average levels of codification.
Moreover, it was interesting to see how organisations seem to maintain their own attitude towards knowledge creation in a consortium project. Universities tend to focus on conversation and codification as they do every day while production organisations tend to exploit their know-how through routine and solution creation. Lastly, consulting appears to focus mainly on solution creation as they would do for a client (fig.4).

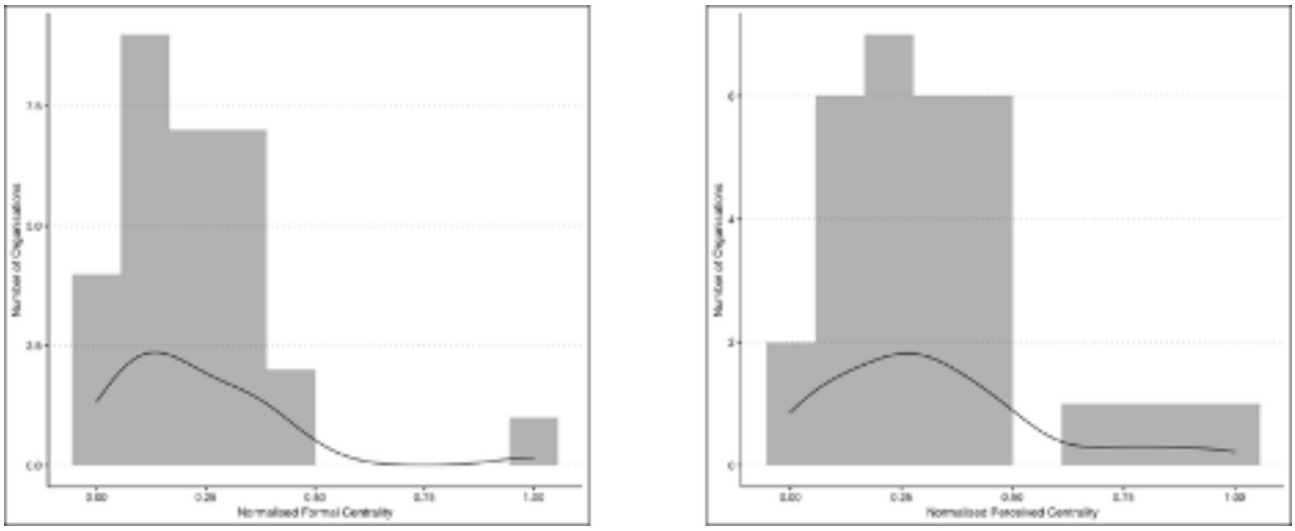


Figure 3 - demonstration of the Poisson distribution

RESULTS

The consortium seems to be ready to face the challenges that big data present: a leadership in charge responsible for innovating the standard internal procedure coherently, human

last point, it is relevant to underline how utilising common languages helps organizations gain a central role in the network, while it seems that participants that occupy a less central role are there because of their lower

A s a

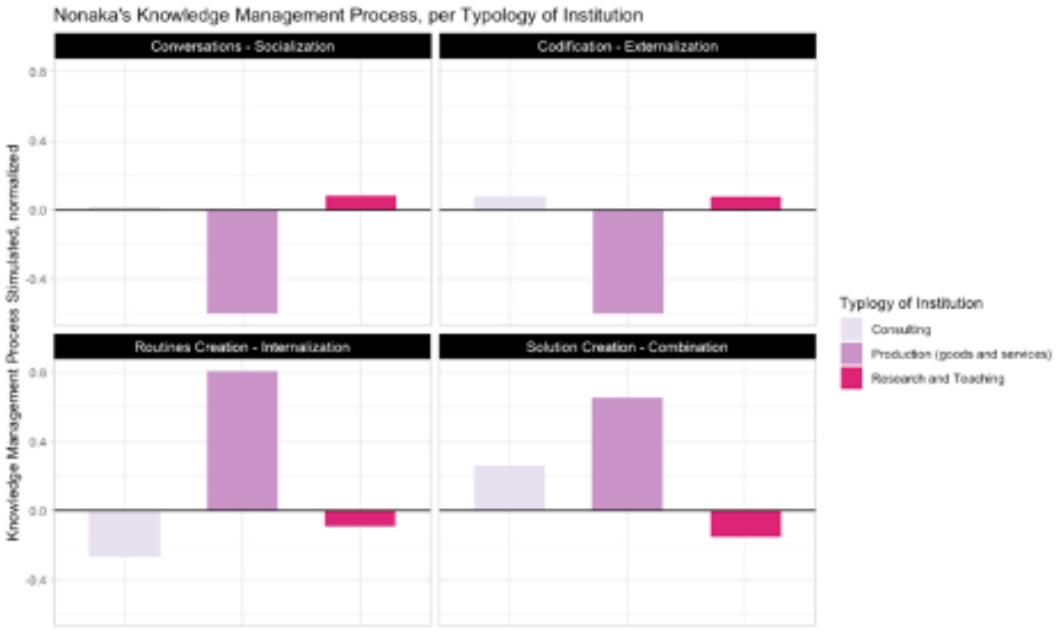


Figure 4 - knowledge creation in a consortium project

ability of speaking shared languages. This intuition, of course, will be expanded once other language dimensions will be added (fig.5).

CONTRIBUTIONS

This document proposes interesting contributions by exploring an unusual context, which is the one of European projects.

As first, it was possible to notice that also research public-funded projects present the same challenges of big data: according to Andrew McAfee and Erik Brynjolfsson (2012) highlighted leadership, data science, system integration and culture are driving factors for the execution of a process, and also in SoBigData+ + context they received a huge amount of focus.

Secondly, as network dynamics seem to apply also in the research context: some characteristics of scale-free networks highlighted by Barabasi (2009) and other contributors such as Poisson distribution, are also present in SoBigData++ Consortium.

Thirdly, this study highlights that network centrality may have a subjective dimension. In the studies of Derrible (2012), centrality was measured from a single point of view. In this research centrality has been split into perceived and formal one, and it has been clear that perceived and formal measure can assume very distant values. This dissonance has led to very different behaviours, i.e., organisations with the same formal centrality but different perceived centrality had very different outcome in Nonaka's knowledge management processes.

References:

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. science, 325(5939), 412-413.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Derrible, S. (2012). Network centrality of metro systems. PloS one, 7(7), e40575.

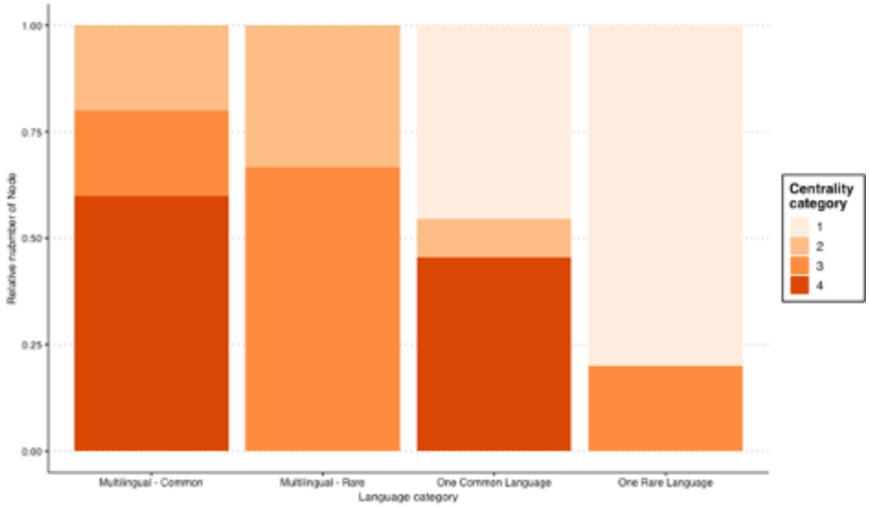


Figure 5 - language dimension

TNA visits: ready when the world is open

Due to the pandemic, very few TNA visits have been able to go ahead in 2020. However, dependent on worldwide travel restrictions, we are hopeful that these will hopefully recommence sometime in 2021.

Joanna Wright, The University of Sheffield | Joanna.wright@sheffield.ac.uk

Trans National Access (TNA) visits offer an opportunity to spend some time in one of numerous institutions throughout Europe conducting a Short-Term Scientific Mission. The host will provide big data computing platforms, big social data resources, and cutting-edge computational methods where you can run experiments on non-public datasets and algorithms. You will also have access to local experts and will be able to discuss your research questions.
Due to the pandemic, very few TNA visits have been able to go ahead in 2020. However, dependent on worldwide travel restrictions, we are hopeful that these will recommence in early 2021. If you would like to get involved then please have a look at the link above to check out past Calls. The next Call will be out as soon as we are able to release it (when travel restrictions are lifted).

Your visit can be as short as a week, or up to 8 weeks and will allow you access not only to datasets and facilities that would ordinarily be unavailable to you, but also to experts who can guide and support you in your research. Researchers, professionals, start-ups and innovators are encouraged to apply. You will benefit from data science and social mining training and experiment ideas and will have access to non-public data sets. We are especially interested in receiving applications from female researchers as SoBigData++ has a commitment to increase the number of females becoming involved in Data Science.

Funding is provided to cover your stay at a host site and is awarded once your application has been approved by the host and an external reviewer. Your application will be assessed on various grounds – such as the scientific merit and originality of the proposed project and your personal statement. The procedure will also include an Ethical review as SoBigData++ is determined to promote responsible and ethical data mining.

We are expecting a high level of applications, as this year we have many more host institutions offering to share their facilities and expertise with visitors. The research project you choose could be linked to the multidisciplinary themes specified in the Calls; it could address resources offered by specific institutions, or you could be planning a blue-sky experiment and wish to explore the infrastructure for your own research project.

As part of your commitment to the project, you will be expected to report on your research, provide feedback on your visit and produce a blog which may be included in one of the SoBigData++ newsletters. Check out previous copies of our newsletter here to see previous blogs: http://www.sobigdata.eu/newsletter.

These visits are a great opportunity to learn something new, meet experts in your field, make connections with similar minded researchers and try out experiments to test your theories as well as the added bonus of visiting a new city. Why not take the opportunity and see where it takes you!

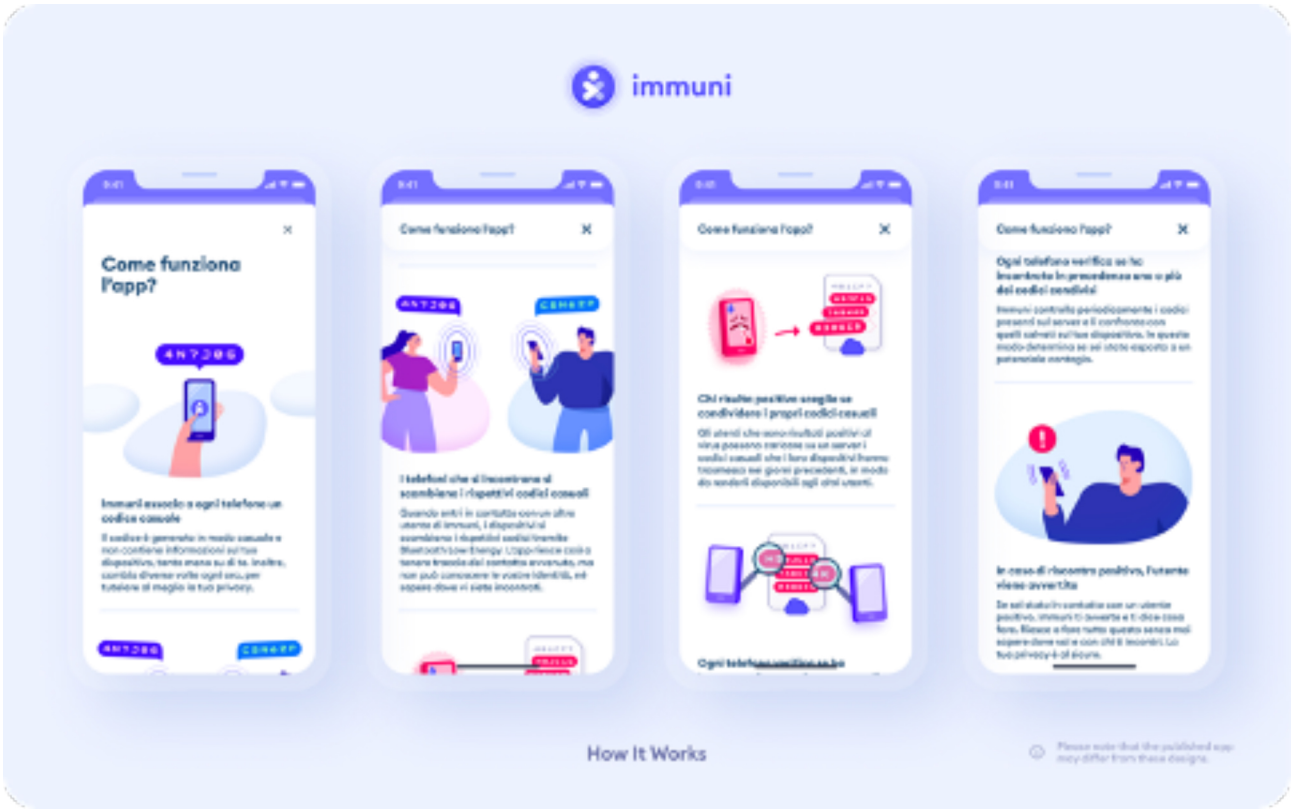
Transparency Issues in Tracing COVID-19

A need for new technological tools has emerged during the current Sars-Cov-2 pandemic. In particular, several mobile applications based on digital tracking and contact tracing have been developed, with ethical implications that have been addressed differently by a number of countries.

Marco Braghieri, Kings College of London, UK | marco.1.braghieri@kcl.ac.uk

Francesca Pratesi, ISTI-CNR National Research Council of Italy | francesca.pratesi@isti.cnr.it

The SoBigData community published a white paper, entitled "Give more data, awareness and control to individual citizens, and they will help COVID-19 containment" (https://bit.ly/whitepaper_covid_sobigdata). The white paper states:
"contact-tracing apps are being proposed for large scale adoption by many countries. A centralized approach, where data sensed by the app are all sent to a nation-wide server, raises concerns about citizens' privacy and needlessly strong digital surveillance, thus alerting us to the need to minimize personal data collection and avoiding location tracking. We advocate the conceptual advantage of a decentralized approach, where both contact and location data are collected exclusively in individual citizens' "personal data stores", to be shared separately and selectively, voluntarily, only when the citizen has tested positive for COVID-19, and with a privacy preserving level of granularity."



A screenshot of how the Italian Government Immuni App works

Aside from the numerous critical evaluations of this technology, a number of countries have published the source code for their official tracking applications, allowing a further degree of transparency and ethical evaluation. What follows is a list of tracking applications and their source code locations, which allow any interested party to evaluate these applications from an ethical point of view.

Australia – CovidSafe (https://www.health.gov.au/resources/apps-and-

tools/covidsafe-app)
Source code (https://github.com/AU-COVIDSafe/mobile-android).

Austria – Stopp Corona (https://www.roteskreuz.at/site/meet-the-stopp-corona-app/)
Source code (https://github.com/austrianredcross/stopp-corona-android

Czekia – eRouška (https://erouska.cz/)
Source code (https://github.com/covid19cz/erouska-android)

France – StopCovid (https://www.economie.gouv.fr/stopcovid)
Source code (https://gitlab.inria.fr/stopcovid19)

Germany – Corona Warn App (https://www.coronawarn.app/en/)
Source code (https://github.com/corona-warn-app)

Iceland – Rakning C-19 (https://www.covid.is/app/en)
Source code (https://github.com/aranja/rakning-c19-app)

India – Aarogya Setu Mobile App (https://www.mygov.in/aarogya-setu-app/)
Source code (https://github.com/nic-delhi/AarogyaSetu_Android)

Israel – Hamagen (https://govextra.gov.il/ministry-of-health/hamagen-app/download-en/)
Source code (https://github.com/MohGovIL/hamagen-react-native)

Italy – Immuni (https://www.immuni.italia.it/)
Source code (https://github.com/immuni-app)

Norway – Smittestop (https://helsenorge.no/coronavirus/smittestop-p?redirect=false)
Reverse engineered source code (https://github.com/djkaty/no.simula.smittestopp/)

Poland – Protego (https://safesafe.app/)
Source code (https://github.com/ProteGO-Safe/web)

Singapore – TraceTogether (https://www.tracetogether.gov.sg/)
Source code (https://github.com/opentrace-community)

Spain – OpenCoronavirus (https://opencoronavirus.app/)
Source code (https://github.com/open-coronavirus/open-coronavirus)

Switzerland – SwissCovid (https://ethz.ch/services/en/news-and-events/solidarity/pilot-swiss-covid-app.html)
Source code (https://github.com/DP-3T/dp3t-app-android-ch)

United Kingdom – NHS Covid19 (https://covid19.nhs.uk/)
Source code – (https://github.com/nhsx/COVID-19-app-Documentation-BETA)

Alongside national applications, Apple and Google developed an Exposure Notification system, providing sample code and information at https://www.apple.com/covid19/contacttracing. Moreover, an updated list of apps and source codes is currently maintained at http://open-source-covid-19.weileizeng.com/.

Besides individual citizens, experts in the field are studying and evaluating these apps. One of the most distinguished institution is the Massachusetts Institute of Technology (MIT), which published an article summarizing the various apps on the basis of five major questions:

Is it voluntary? In some cases, apps are opt-in, but in other places many or all citizens are compelled to download and use them.
Are there limitations on how the data gets used? Data may sometimes be used for purposes other than public health, such as law enforcement — and that may last longer than COVID-19.
Will data be destroyed after a period of time? The data the apps collect should not last forever. If it is automatically deleted in a reasonable amount of time (usually a maximum of around 30 days) or the app allows users to manually delete their own data, we award a star.
Is data collection minimized? Does the app collect only the information it needs to do what it says?
Is the effort transparent? Transparency can take the form of clear, publicly available policies and design, an open-source code base, or all of these.
All the aforementioned apps can be downloaded and used on a voluntary basis (like the large majority of the ones considered in the list of the MIT article, where the only exceptions were Bahrain, China, India, Qatar, and Turkey). Actually, also Australia, France, Norway, Spain and Switzerland (not listed as transparent in the MIT summary) released their source code, guaranteeing a better transparency. From the MIT report, 17 apps rely on Bluetooth technologies, 7 will collect locations, while 5 of them will use an hybrid approach.

For more information:
https://www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/

Private Sources of Mobility Data under COVID-19

The COVID-19 pandemic is changing the world in unprecedented and unpredictable ways. Human mobility is at the epicenter of that change, as the greatest facilitator for the spread of the virus.

Dario Garcia Gasulla, HPAI group, Computer Science Department, Barcelona Supercomputing Center (BSC) | dario.garcia@bsc.es

To study the change in mobility, to evaluate the efficiency of mobility restriction policies, and to facilitate a better response to possible future crisis, we need to properly understand all mobility data sources at our disposal. This post regards a work dedicated to the study of private mobility sources, gathered and released by large technological companies. This data is of special interest because, unlike most public sources, it is focused on people, not transportation means. i.e., its unit of measurement is the closest thing to a person in a western society: a mobile phone. Furthermore, the sample of society they cover is large and representative. On the other hand, this sort of data is not directly accessible for anonymity reasons.

We consider the use of private data sources (Google and Facebook) for assessing the levels of mobility in Spain. By doing so, we draw conclusions on two fronts. First, on the behavior and particularities of private data sources. And second, on how mobility changed during the COVID-19 pandemic in Spain.
Regarding private data sources, we have shown the differences between using an absolute measure (like Facebook) and a relative measure (like Google). Both of them have limitations when used in isolation. The former lacks a contextualization of its values, while the latter depends entirely on the baseline used. When used together, they provide a visualizing of mobility where consistent patterns can be easily identified (as presented later in this section). For specific purposes, using a single data source may suffice, as long as it fits the goal:
An absolute measure like Facebook's can be very useful for epidemiologic purposes, as it provides a pure measurement of mobility. That includes estimating number of contacts in a society, modeling the spread of the virus, and measuring the impact of policies on absolute mobility;

A relative measure like Google's can

City vector created by GarryKillian

be very useful for socio-economic purposes, as it provides a contextualized measurement of mobility. That includes understanding the change caused by the new normality, and the economic impact of mobility restriction policies.

Regarding the analysis of Spanish mobility during the COVID-19 pandemic, we extract several conclusions. On one hand, data shows a huge mobility containment, sustained for a month and a half (March 15 to May 1st, approximately), very close to its theoretical limit (as represented by mobility during the hard-lockdown). This duration was sufficient to contain the spread of the virus and bring infection numbers down to traceable scale. In hindsight, the policies implemented in Spain seem appropriate and proportional to the severity of the situation. That being said, the role, timing and convenience of the hard-lockdown remains to be further discussed. We show a relatively modest contribution of this policy to mobility reduction. On the other hand, the hard-lockdown may have had an effect on prolonging adherence.
We identify mild differences between regions during the three months of restricted movement. Certain regions had a stronger adherence to confinement than others, mostly in relative terms. This may be caused by regional differences in pre-pandemic mobility, which is used as baseline for the relative measurement. A similar artifact are the inverted peaks of weekends, where a relative measure spikes down and an absolute measure spikes up. As demonstrated, this is the result of combining a measure relative to the weekday with an absolute measure.

We also saw significant differences among days. Weekends exhibit the highest volume of mobility reduction in absolute terms, even during the hard-lockdown, when work-related trips were forbidden for all except essential services. At the same time, weekends have the smallest mobility reduction in relative terms, indicating that the effort society had to make in this regard with respect to its previous patterns was smaller. Fridays and Sundays are particularly relevant days, the first because it represents the biggest change from normal behavior, the second because it represents the biggest absolute decrease in mobility. These particularities could be exploited for the general good.

Finally, we analyzed the new normality by looking at the weeks of de-confinement, up until June 27, a week after the state of alarm was lifted on the whole of Spain. In this period, we found Saturdays and Sundays to be already at pre-pandemic levels of mobility. In contrast, working days (Monday to Friday) still show significant differences. The new normality also shows differences between regions, particularly for working days. Regions with large metropolitan areas exhibit a reduction in mobility between 4% and 14% after restrictions were lifted. Indeed, the new normality is most new on urban working days.

The paper [1] describing this research has been recently published and it is available here: http://arxiv.org/abs/2007.07095

[1] Raquel Pérez Arnal, David Conesa, Sergio Alvarez-Napagao, Toyotaro Suzumura, Martí Català, Enric Alvarez, Dario Garcia-Gasulla, “Private Sources of Mobility Data Under COVID-19”, 7 2020.

How Digital Data is changing how we measure Well-Being and Happiness

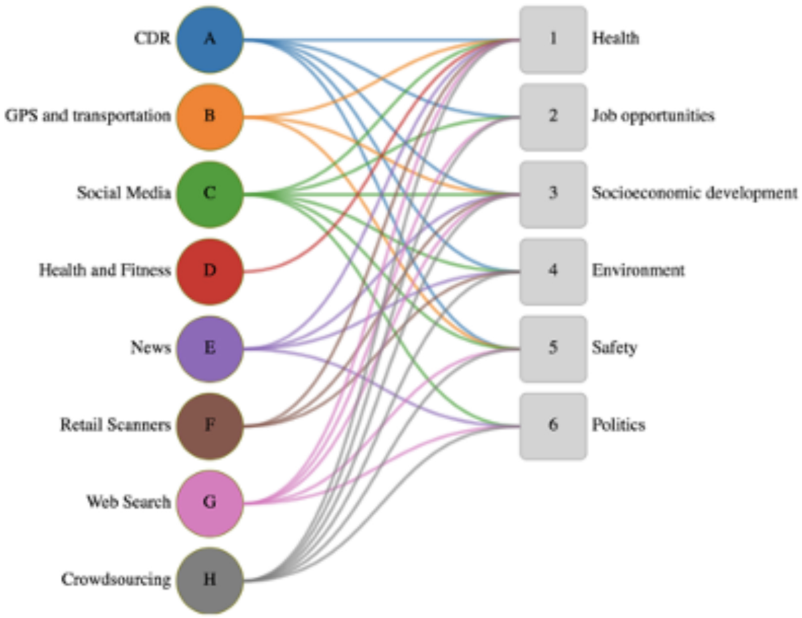
What is well-being, and how can we measure it? This complex question has fascinated philosophers and thinkers since ancient times. Now we can measure well-being and happiness analyzing digital data.

Vasiliki Voukelatou, SNS Scuola Normale Superiore, Pisa, Itlay | vasiliki.voukelatou@sns.it

What is well-being, and how can we measure it? This complex question has fascinated philosophers and thinkers since ancient times. For example, Aristotle has expressed his interest on the topic claiming that human well-being, labeled as eudaimonia (greek: ευδαιμονία: Eu=Good, Daimon=spirit), is an activity of the soul expressing complete virtue [11].

In modern times, economists and policy-makers have traditionally considered Gross Domestic Product (GDP) as a good indicator of well-being in society. Unfortunately, GDP cannot measure many aspects of what makes people's life worth living, and lately researchers of various backgrounds have started instead measuring well-being considering it as an index of societal progress and an effective indicator for public policing [1].

However, talking about well-being generally can be misleading, given the complexity that this concept conveys. For this reason, researchers generally distinguish between **objective well-being** and **subjective well-being** [2, 3, 4, 5]. Both definitions, and their relevant dimensions, have been traditionally captured with self-report surveys [6]. Although traditional data have been considered accurate and valid, they bring some considerable disadvantages, such as time limitations and high costs. Therefore, taking advantage of the big data revolution, researchers and nonprofit organisations have started to use many novel data sources to eliminate the limitations brought from traditional data and to contribute to the exploration of well-being and its relevant dimensions.(https://www.nature.com/articles/d41586-020-01747-1?proof=true19)



Sources of data (left) and dimensions (right) of objective well-being.

MEASURING OBJECTIVE WELL-BEING

Objective well-being investigates the objective dimensions of a good life, and therefore researchers have focused on determining and exploring its dimensions. In fact, the Organisation for Economic Co-operation and Development (OECD) [2], the United Nations Development Programme (UNDP) [3], and the Italian Statistics Bureau (ISTAT) [4] work in this direction. Thus, taking into consideration their documentation, we can identify six major objective and observable dimensions for its measurement: health, job opportunities, socioeconomic development, environment, safety, and politics. All these dimensions together represent the objective well-being, which is assessed through the extent to which these "needs" are satisfied.

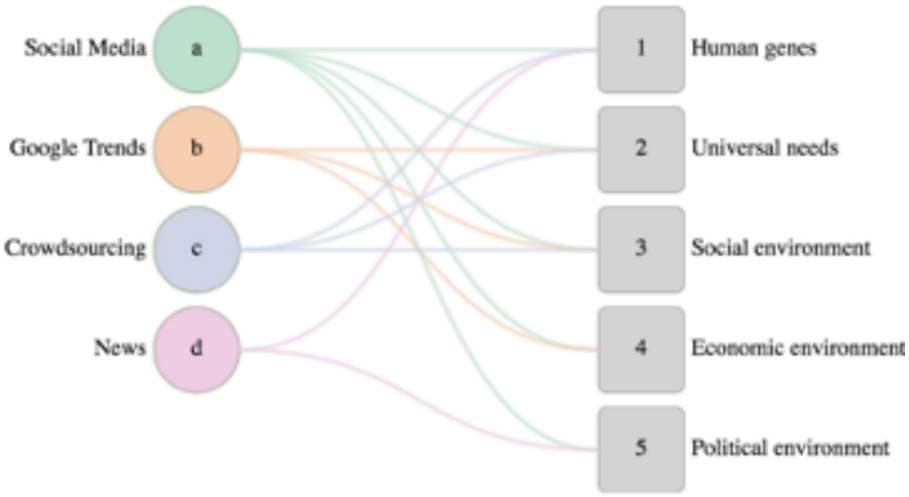
The last few years have witnessed a change in the way of measuring objective well-being. We identify eight important novel data sources that are lately used for the exploration of objective well-being and its relevant dimensions: CDRs, GPS and transportation, Social Media, Health and Fitness, News, Retail Scanners, Web Search, and Crowdsourcing. The figure below describes the new data sources (left) that have been used to estimate one or more dimensions of objective well-being (right). The presence of a link in the figure between a data source and a dimension indicates that there are papers in the literature on monitoring that dimension with that data source. For example, Pappalardo et al. [7] use CDRs to capture the employment rate of French cities (A2).

MEASURING SUBJECTIVE WELL-BEING

Subjective well-being examines people's subjective evaluations of their own lives. In 2013 the OECD [8] recognized the importance of taking into consideration people's perceived well-being, labeled as subjective

well-being or happiness, when investigating overall well-being. Studies using traditional data to measure happiness have identified the main determinants of well-being (see e.g. [9]) that we divide in five main dimensions: the role of human genes, which seem to be fairly heritable, universal needs, meaning basic and psychological needs, social environment, such as education and health, economic environment, including a lot of research on income, and political environment, such as democracy and political freedom.

Similarly to objective well-being, researchers use novel data sources to explore subjective well-being. In particular, Social Media, Google Trends, Crowdsourcing, and News are used from researchers for the exploration of subjective well-being and its relevant dimensions. The figure below describes the new data sources (left) that have been used to estimate one or more dimensions of subjective well-being (right). The presence of a link in the figure between a data source and a dimension indicates that there are papers in the literature on monitoring that dimension with that data source. For example, Dodds et al. [10] construct the Hedonometer to measure temporal patterns of societal happiness, as influenced by basic needs (a2), social (a3), economic (a4) and political (a5) determinants.



Sources of data (left) and dimensions (right) of the subjective well-being.

WHY FREQUENT MEASUREMENT OF WELL-BEING IS IMPORTANT

At this critical moment that the global society is under socioeconomic and political crisis and instability, policy-makers and social good organisations need frequent updates of well-being. This is the reason they are attracted by the intellectual opportunities that novel data sources offer to explore well-being. In particular, novel data sources supplement the traditional data by making the estimation of well-being cost-efficient and almost real-time. Only by having frequent well-being estimations, policy-makers can timely react on applying the right policies to prevent detrimental societal effects and contribute to societal progress.

This post is based on a paper[A1], supported by SoBigData, which provides the theoretical background on objective and subjective well-being, including their relevant dimensions. Additionally, it presents to researchers the new data sources used for capturing well-being, discusses indicative existing studies, and sheds light on still barely unexplored dimensions and data sources that constitute opportunities for future research on well-being.

[A1]
V. Voukelatou, L. Gabrielli, I. Miliou, S. Cresci, R. Sharma, M. Tesconi, and L. Pappalardo, "Measuring objective and subjective well-being:dimensions and data sources,"International Journal of Data Science and Analytics (JDSA), 2020, https://doi.org/10.1007/s41060-020-00224-2

REFERENCES

[1] Marc Fleurbaey. Beyond gdp: The quest for a measure of social welfare. Journal of Economic literature, 47(4):1029–75, 2009.
[2] Organisation for Economic Co-operation and Development.How's life?: measuring well-being. OECD Paris,2011.
[3] UNDP. Sustainable Development Goals. https://sustainabledevelopment.un.org/sdgs, 2015. (Online;accessed October 2019).
[4] Rapporto, BES. Il benessere equo e sostenibile in Italia, 2015. ISTAT.[5] R Veenhoven. Conditions of happiness, Reidel (now Springer), Dordrecht, The Netherlands, 1984.
[5] R Veenhoven. Conditions of happiness, Reidel (now Springer), Dordrecht, The Netherlands, 1984.
[6] Angus Deaton.The analysis of household surveys: a microeconometric approach to development policy. TheWorld Bank, 1997
[7] L. Pappalardo, D. Pedreschi, Z. Smoreda, and F. Giannotti. Using big data to study the link between humanmobility and socio-economic development. In2015 IEEE International Conference on Big Data (Big Data),pages 871–878, Oct 2015.
[8] Organisation for Economic Co-operation and Development (OECD).OECD guidelines on measuring subjectivewell-being. OECD Publishing, 2013.
[9] Paul Dolan, Tessa Peasgood, and Mathew White. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being.Journal of economic psychology,29(1):94–122, 2008.
[10] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher MDanforth. Temporal patterns of happiness and information in a global social network: Hedonometrics andTwitter. PloS one, 6(12):e26752, 2011.
[11] Burger R (2009) Aristotle's Dialogue with Socrates:On the" Nicomachean Ethics". University ofChicago Press

Human Migration: the Big Data perspective

Human migration is a constant phenomenon in human history, and its study involves numerous research fields.To date, both traditional and novel models and data are employed to understand the mechanisms of the different stages of migration (the journey, the stay, and the return).

Laura Pollacci, ISTI-CNR National Research Council of Italy | laura.pollacci@isti.cnr.it

Matteo Bohm, ISTI-CNR National Research Council of Italy | matteo.bohm@gmail.com

Human migration is a constant phenomenon in human history, and its study involves numerous research fields. To date, data not typically used for studying migration are increasingly available. These include the so-called social Big Data: digital traces left by humans through cell phones, online social networks, and online services. More and more technologies can be employed to extract information from these large datasets. However, how can Big Data help to understand the migration phenomenon?

To date, both traditional and novel models and data are employed to understand the mechanisms of the different stages of migration (the journey, the stay, and the return).

THE JOURNEY: MIGRATION FLOWS AND STOCKS

Tracking international migrants' flows and stocks is a task as important as it is challenging. Researchers and policymakers relying on traditional data sources such as official statistics or administrative data often meet various limitations. These limitations are typically due to the involvement of various nations in the migration process; i.e., data may be inconsistent across different countries' databases. While traditional data are useful to study the journey of migrants, social Big Data may help researchers to overcome the limitations of traditional data and may allow in real-time analyses (see for instance, [1,2,3]).

The use of social Big Data to study the immigrants' journey is increasing. Various data types fall under this category; between these, Twitter data,

Skype Ego networks, Google Trend Index (GTI) [4], LinkedIn data [5], publications in academic journals [6], ORCID data (A recent line of work in the SoBigData project is to understand, by using ORCID data, what was the effect of the Brexit referendum on scientific migration), and long-term origin-destination data. For instance, Twitter data can be used to quantify diversity in communities [7] and estimate user nationality; Skype Ego networks data can be used to explain international migration patterns [8].

As well as traditional data, unconventional Big Data has its limitations, including bias and privacy issues. Thus, new methods are developing to address issues and take advantage of the almost worldwide data coverage. The hope is that merging knowledge from both traditional and novel datasets may lead to overcoming issues and building more and more accurate models to nowcast immigrants' journeys and immigration rates.

THE STAY: EFFECTS ON COMMUNITIES, IMMIGRANT INTEGRATION

The study of immigrants' integration and the effect of migration on the communities is complex and challenging. Integration and cultural changes have been traditionally analyzed using census data, administrative registries, and surveys.

Integration has been analyzed from multiple viewpoints, including marriage relationships [9,10], social relations [11], labor market [12], and language adoption [13]. On the other side, educational expectations [14,15], economic prosperity [16], cultural distance with the origin country, school class composition [17], and ethnic attitudes are used to study the effects on the local population due to integration.

As for the journey, Big Data can help to analyze the stay producing real-time results. Several works have been done using Call Detail Records (CDRs) in understanding individual [18] and group mobility [19], even during environmental disasters [20]. These data can be used to describe social interaction, mobility, and segregation. However, CDR may lead to coverage issues when analyzing international migration flows.

Retail data, such as those from a supermarket chain, may help understand how immigrants adopt habits and whether they are converging to or diverging from the norms of the destination country [21].

Also, Online Social Networks (OSNs) data, for instance, can help study social integration looking at the opinions of the locals related to migration topics. The language used on OSNs can be used to depict the worldwide linguistic geography [22], detect linguistic variation [23], identify patterns in language usage, analyze the language diversity [7], changes in the local language, and sentiment towards immigrants [24, 25].

THE RETURN: MIGRANTS RETURNING TO THE COUNTRY OF ORIGIN

Migration can also be considered as a temporary phenomenon. Since return migration is increasing in several countries, it has been extensively investigated wrt different aspects, such as decreasing violence [29].

Together with factors involved in the decision of return, scholars also investigated the benefits that return migration brings to the countries of origin. Advantages fall in various fields. Economically, new skills learned abroad may help returned migrants to start their new one business in the origin country; and, the money sent from migrants to their families is a valuable incoming [30,31]. Benefits also affect educational attainment and health conditions. For instance, regarding education, return migrants can be associated with increases and improvement of educational attainment [34], and social practices introduced by return migrants positively affect healthcare [35].

Other studies [28, 36, 37] focuses on electoral participation. These works suggest that local policies are typically positively affected by returning migrants since they contribute to increase political participation and enhance political accountability.

Especially in recent times, much of the research has focused on the relationship between return migration and personal skills. In particular, researchers investigated the "brain gain'" provided by the return of high-skilled individuals, such as scientists returning in the origin country [32, 33].

DISCUSSION

Human Migration can be studied following three lines of research. Today social Big Data can complement existing approaches, but these models still need to be validated and refined. The issue is the lack of gold standards as exact current immigration rates with which to validate nowcasting models. The hope is that better relations between policies and immigration could be a breakthrough in solving this problem.

On the other hand, research needs to consider issues with the data that is being used, be it traditional or unconventional. An additional issue relies on the ethical dimension of collecting and processing personal data, including sensitive personal data, describing human individuals and activities.

Now more than ever, collecting, preprocessing, and analyzing data need to be managed with ethical and legal values such as privacy and data protection. The context of migration is sensitive to the ethics dimension since individuals described in the data may be particularly vulnerable.

Main Reference:

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., ... & Pappalardo, L. (2020). Human migration: the big data perspective. International Journal of Data Science and Analytics, 1-20. https://link.springer.com/article/10.1007/s41060-020-00213-5

Black Box Explanation by Learning Image Exemplars in the Latent Feature Space

Nowadays, Artificial Intelligence (AI) systems for image classification are generally based on effective machine learning methods such as Deep Neural Networks (DNNs). These models are recognized to be "black boxes" because of their opaque, hidden internal structure, which is not human-understandable.

Riccardo Guidotti, ISTI-CNR National Research Council of Italy | riccardo.guidotti@isti.cnr.it

Nowadays, Artificial Intelligence (AI) systems for image classification are generally based on effective machine learning methods such as Deep Neural Networks (DNNs). These models are recognized to be "black boxes" because of their opaque, hidden internal structure, which is not human-understandable [1]. As a consequence, there is an increasing interest in the scientific community in deriving eXplainable AI (XAI) methods able to unveil the decision process of AI based on black box models [2]. Explaining the reasons for a certain decision can be very important. For example, when dealing with medical images for diagnosing, how can we validate that an accurate image classifier built to recognize cancer actually focuses on the malign areas and not on the background for taking the decisions?

In [3], we have investigated the problem of black box explanation for image classification. In the literature, such a problem is addressed by producing explanations in forms of saliency maps through different approaches. A saliency map is an image where the color of each pixel represents a value modeling the importance of that pixel for the prediction, i.e., they show the positive (or negative) contribution of each pixel to the black box outcome. Gradient-based attribution methods

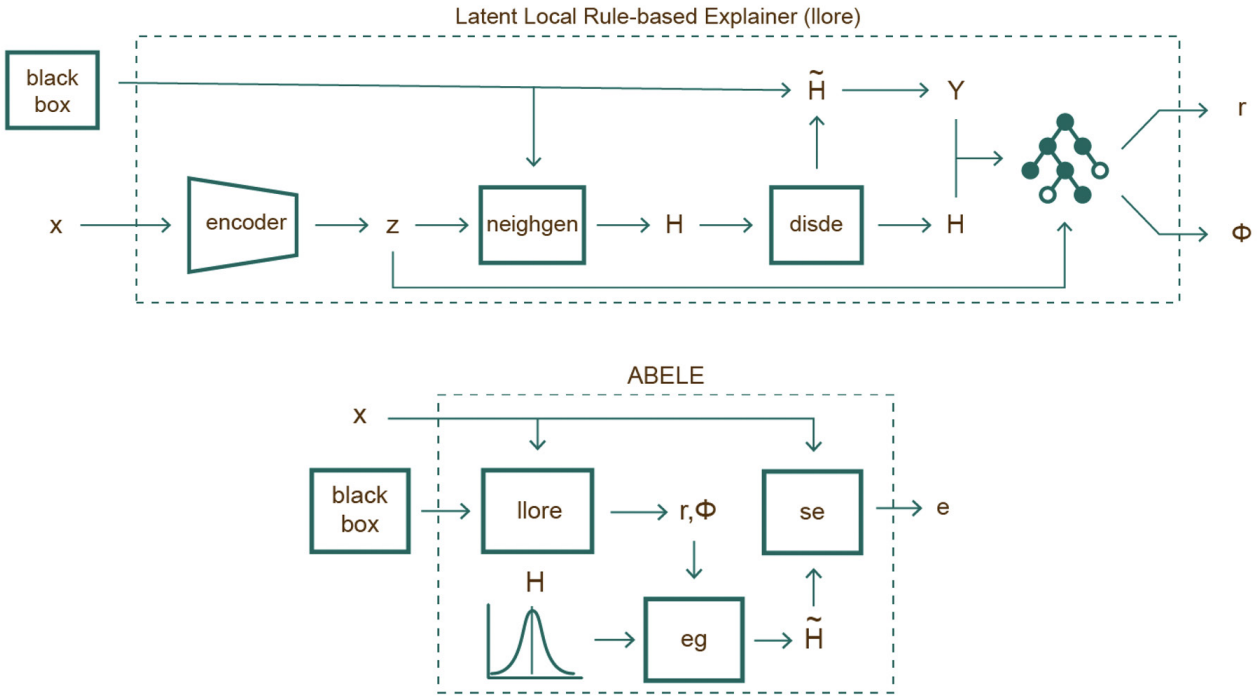


Figure 1 - Saliency map

[4,5] reveal saliency maps highlighting the parts of the image that most contribute to its classification. These methods are model-specific and can be employed only to explain specific DNNs. On the other hand, model-agnostic approaches can explain through a saliency map the outcome of any black box [6,7]. However, these methods exhibit drawbacks that may negatively impact the reliability of the explanations: (i) they do not take into account existing relationships between pixels during the neighborhood generation, (ii) the neighborhood generation may not produce "meaningful" images.

ABELE, Adversarial Black box Explainer generating Latent Exemplars [3], is a local, model-agnostic explanation method able to overcome the existing limitations by exploiting a latent feature space for the neighborhood generation process. Given an image classified by a black box model, ABELE provides an explanation for the classification that consists of two parts: (i) a saliency map highlighting the areas of the image that contribute to its classification, and the areas of that push it towards another outcome, (ii) a set of exemplars and counter-exemplars illustrating, respectively, instances classified with the same label and with a different label than the instance to explain.

The explanation process of ABELE involves the following steps. First, ABELE generates a neighborhood in the latent feature space exploiting the encoder of the Adversarial Autoencoder (AAE) [8]. Then, it learns a decision tree on the latent neighborhood providing local factual and counter-factual rules r and [9]. Figure 1-left shows this workflow. After that, ABELE generates exemplars and counter-exemplars respecting r and by exploiting the decoder and discriminator of the AAE. Finally, (Figure 1) the saliency map is obtained as the median value of the pixel-to-pixel-difference between the image analyzed and the exemplars.

The experiments illustrated in [3] show that ABELE overtakes state of the art methods by providing relevant, coherent, stable and faithful explanations. In Figure 2 are reported the explanations of a DNNs for the mnist dataset classified as "9" and "0", respectively. The first column contains the saliency map provided by ABELE: the yellow areas must remain unchanged to obtain the same label, while the red and blue ones can change without impacting the black box decision. With this type of saliency map we can understand that a "9" may have a more compact circle and that a "0" may be more inclined. The rest of the columns contain the explanations of other explainers: red areas contribute positively, blue areas contribute negatively. For LIME, nearly all the content of the image is part of the saliency map, and the areas have either completely positive or completely negative contributions. The other gradient-based explanation methods [4,5] return scattered red and blue points not clustered into areas. It is not clear how a user could understand the decision process with these other explanations. In Figure 2-right are shown two exemplars and two counter-exemplars. We notice how the label "9" is assigned to images very close to a "4" but until the upper part of the circle remains connected, it is still classified as a "9". On the other hand, looking at counter-exemplars, if the upper part of the circle has a hole or the lower part is not thick enough, then the black box labels them as a "4" and a "7", respectively.

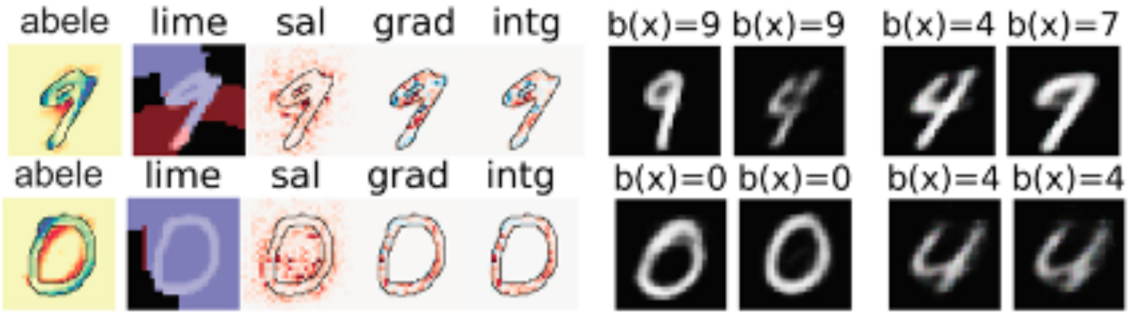


Figure 2 - explanations of a DNNs for the mnist dataset classified as "9" and "0"

References

[1] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

[2] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys.

[3] Guidotti, R., Monreale, A., Matwin, S., & Pedreschi, D. (2019). Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In ECML-PKDD.

[4] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

[5] Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. arXiv:1605.01713.

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. In SIGKDD.

[7] Lundberg, M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In NIPS.

[8] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv:1511.05644.

[9] Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems.

Intensity vs Accuracy: Technical-tactical differences between male and female football teams

Women's football resurfaced in the 1960s in the Nordic countries of Europe, and it is now spreading all over the world. From 2012 the number of women academies has doubled, with around 40 million girls and women playing football worldwide nowadays.

Luca Pappalardo, ISTI-CNR National Research Council of Italy | luca.pappalardo@isti.cnr.it

Alessio Rossi, University of Pisa, Italy | alessio.rossi@di.unipi.it

Paolo Cintia, University of Pisa, Italy | paolo.cintia@di.unipi.it

Women's football took its first steps since the early twentieth-century. Unfortunately, the ostracism from the English Football Association drastically slowed down its development, which experienced a long period of stagnation. Women's football resurfaced in the 1960s in the Nordic countries of Europe, and it is now spreading all over the world. From 2012 the number of women academies has doubled, with around 40 million girls and women playing football worldwide nowadays.

The gain of the popularity of women's football has stimulating exciting question: What are the differences between women's and men's football? In principle, the rules and requirements of the two games are the same. In practice, as in other sports, there are natural differences between men and women in terms of physical skills, while technical-tactical differences between male and female players' are not deeply investigate yet.

To this aim, we analysed an extensive data set of soccer-logs describing all the spatio-temporal events that occur during the last men's and women's World Cups: 64 and 44 matches, respectively, and 32 men's and 24 women's teams with 736 male players and 546 female players. We quantified the performance of players and teams in several ways, from the number of events generated during a match to the proportion of accurate passes, the velocity and fluidity of the game, the quality of individual performance, and the collective behaviour of teams.

Men's matches have, on average, more events than women's ones. Specifically, women's matches have, on average, more free kicks, duels, accelerations, clearances, and ball touches but fewer passes and fouls than men's matches. Furthermore, men's passes are on average more accurate than women's ones. More-

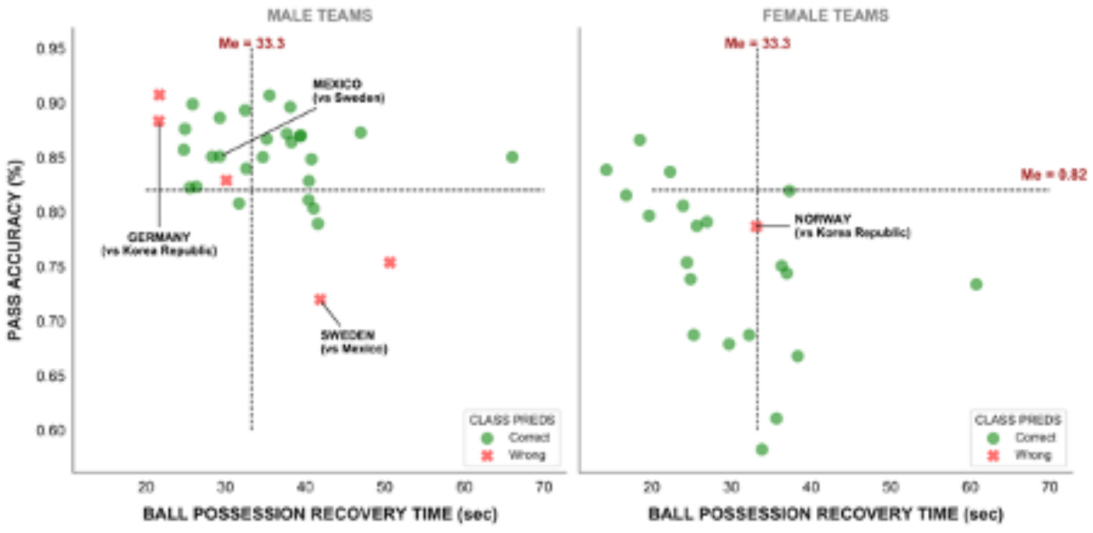


Figure. Scatter plots displaying pass accuracy as a function of recovery time, among male national teams (left) and female national teams (right). The green circle indicates a team correctly classified, in a game; the red cross indicates a mistake. The dashed lines are at the median values for the two variables over the entire data set.

over, on average, men kick the ball from a greater distance than women, where we measure shooting distance from the shot's origin to the opponent's goal center. Furthermore, pass velocity is lower for a female team than a male one, while woman regain the ball possession faster than men. In contrast, there is no difference in the time elapsed between two shots, between male and female teams, and men's passes are on average longer than women's ones.

We then ask the following question: Can a machine distinguish a male team from a female one, based on their technical performance? Our answer, based on the use of a supervised classifier, reveals that men's and women's football do have apparent differences in terms of technical features, which we investigate through the inspection of the classifier. In particular, the most important features that permits to discriminate between male and females football teams are:
the percentage of accurate passes in the match (AccP);
the average ball possession recovery time;
how long a team waits after a game stop before restarting the game with a free-kick, a corner kick or a throw-in;
the average time elapsed between two passes.

Interestingly, we find that the number of passes and shots, generally recognized as important metrics for a team's performance, are considered less important in discriminating between a male and a female team.

This inspection allows us to highlight the characteristics of the female teams that are misclassified as male ones and vice versa, revealing interesting information about the national teams' playing styles. For example, the figure below shows the predictions of the Decision Tree where in one case out of 21, the model misclassified a female team as a male one, while on five cases out of 31, a male team is misclassified as a female one. For example, in match Sweden vs. Mexico of the men's World Cup, Mexico is correctly classified as a male team (pass accuracy = 85%, recovery time = 30 s), while Sweden is misclassified as a female team (pass accuracy = 72%, and recovery time = 42 s). Note that the accuracy of passes of Sweden in that game is lower than the average pass accuracy of male teams (84%), making Sweden more similar to a female team than to a male one (average pass accuracy for females is 75%).

In conclusion, the way female teams play is more intense but less accurate and more fragmented: in a women's football match, the time elapsed between two consecutive passes is lower, and so is pass accuracy, and female teams tend to regain ball possession faster than male ones.

The paper describing this research has been just submitted to the workshop "MLSA 2020: Machine Learning and Data Mining for Sports Analytics".

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

SoBigData Magazine is published under the

project N° 871042 | Programme: H2020 - INFRAIA

Duration: 01/01/2020 - 31/12/2023