Issue 6 / June 2021

Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Magazine

A roadmap to the future

ESFRI, the European Strategy Forum on Research Infrastructures, has selected SoBigData to be part of the Roadmap 2021. This ambitious recognition given to the most important AI and big data research infrastructure in Europe lays the groundwork to receive the largest funding for a research consortium dedicated to improving artificial intelligence for social good. ESFRI ensures long-term sustainability to SoBigData RI and opens new opportunities to grow and establish itself more and more as a reference point for research in AI and big data, not only in Europe but worldwide. Indeed, SoBigData's goal is to become a world-leading institution in its field.

SoBigData has gained political support from Italy, Estonia, Switzerland, and Bulgaria. The Italian Ministry of Research and the Italian National Research Council (CNR), and ten other European entities have been granted financial support, 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,

[Continues on pag. 3]

Inside this issue

Editorial

News

Events Highlights

Research Highlights

TransNational Access

Exploratories Highlights

SoBigData RI joins ESFRI

The Editorial Board

Austria, Germany, France, Spain, and the United Kingdom. The overall cost of the ESFRI SoBigData RI is estimated at more than 150 million €, which includes both the build-up and operational phase. The preparation phase started in 2020, and the RI will be operative until 2050.

ESFRI has chosen AI as the next landmark technology to be developed in Europe. Big data analytics and AI are fundamental tools for sustainable socio-economic development, and all European countries are involved in the transformation that AI implies. In this context, for the SoBigData Research Infrastructure (RI) is a very successful result to become part of the ESFRI RoadMap 2021, considering that SoBigData is the only European RI that binds AI and Society.

The uniqueness of SoBigData is represented by its ability to connect heterogeneous scientific communities, such as data science and artificial intelligence. The ESFRI SoBigData RI, with its network of prestigious data science nodes, has the ambition and chance to become a strategic European resource worldwide in data-set, experiment and research skill, and computational resource sharing.

"Our mission is to create a research infrastructure to support the analysis of Big Data and the development of artificial intelligence for the future of Europe, and at the same time guarantee the protection of the individual's privacy and promote ethical principles such as transparency and repeatability at every level of analysis. The entry into the ESFRI roadmap makes this goal more concrete and closer".

The coordinator of SoBigData RI from the National Research Council (CNR) of Italy

A challenge for the exploitation of Big Data potential

Challenge Us is a brilliant opportunity for companies interested in exploring the potential of big data in their business. This initiative's aim is to help small and medium-sized companies to manage their (big) data problems and provides companies with opportunities to further benefit from the data they bring for analysis. The program – which will run for the next three years – is organized by the SoBigData++ consortium and will help companies by providing free of charge services in accomplishing their proposals' proof of concept.

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

Rajesh Sharma, University of Tartu | rajesh.sharma@ut.ee

In an increasingly digitalized and interconnected world, the amount of data produced every day has reached levels unimaginable until a few years ago. Companies produce a lot of data [R1] on a daily basis through various devices connected to the network such as smartphones, industrial technological machinery, sensor networks, smart cars, web search engines, web searches and social networks [R2]. Every time a machine is in operation, its internal computer records its consumption, wear, productivity and interactions with other machines. An analysis of the data can help to improve the efficiency of the machine, predict failures and reduce the number of defective parts brought to market. For example, many companies underestimate the fact that their presence on the Internet - through a company website and social network accounts - produces a whole series of data which, if properly collected and analyzed, can be useful in predicting new markets and increasing competitiveness. Exploitation of existing data may increase firms' innovation and production capabilities.

However, being able to exploit this data correctly is not an easy challenge for companies, especially those that are not "digital natives". It is necessary to develop skills and acquire mastery of cutting-edge technologies, integrating them later in the organizational and production processes. For this reason, companies increasingly tend to rely on experts who can guide them in this transformation process.

In order to exploit the potential of data, however, firms need to rethink their business models and develop competences and skills which helps them in successfully lead this transformation. To facilitate this transition, the SoBigData++ consortium has proposed the "Challenge Us Program", a free consultancy service provided by various European research institutions and universities to companies wishing to explore the potential of data they generate during the production, marketing, interactions with customers and the supply of their products.

As highlighted by researchers involved in the "Challenge Us Program" [http://www.sobigdata.eu/challenge-us-2021], companies must

be able to exploit this data to continue to remain globally competitive through the offer of new products / services and to make their processes more efficient. The use and analysis of this large amount of data can in fact be crucial to understand and predict customer behaviour, predict when a machine will need maintenance, or add new services to improve its offer.

The proposal selection connected with a problem or a potential idea of data exploitation generated by a business has been developed by a group of experts belonging to the SoBigData++ consortium. The selected companies will then be assisted in the resolution of the "challenge" by expert researchers in the field of big data belonging to the SoBigData ++ consortium.

The first challenge us program has been organized in Spring 2021 (which is the project's second year) and this year is organized by Scuola Superiore Sant'Anna in Pisa, Italy. All the information about the program and how to participate can be found at the following link: http://www.sobigdata.eu/challenge-us-2021

Links:
L1: http://www.sobigdata.eu/challenge-us-2021

References:
[1] 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.
[2] Urbinati, A., Bogers, M., Chiesa, V., & Frattini, F. (2019). Creating and capturing value from Big Data: A multiple-case study analysis of provider companies. Technovation, 84, 21-36.

Photo_credit_Franki_Chamaki_Unsplash

Introducing the novel SoBigData Literacy database

Paving the way for consolidating a shared knowledge base for the SoBigData community as a whole.

Giorgia Pozzi, TuDelft |G.Pozzi@tudelft.nl

The working environment "SoBigData literacy" has been created in the previous months as a part of the SoBigData catalogue. It is conceived to support the collaborative development of a curated collection of literature of interest for the SoBigData community as a whole.
The collection consists of a catalogue service enacting authorized members to publish literature of interest and organize the selected contents to facilitate discovery and access, promoting the exchange of relevant information among the members of the project.

The SoBigData Literacy has been divided into different categories according to different types of publishing. Therefore, to allow the users of the catalogue to browse through it in an easy and agile way, the different items are grouped in journal articles, book chapters, conference papers, and research articles.

The topics of the literature gathered in the SoBigData Literacy are particularly broad and comprehensive since the catalogue aims at mirroring the thematic and conceptual variety present within the project, which touches upon different aspects of higly relevant issues related to, among others, big data and artificial intelligence-based developments. In order to cope with the task of representing the thematic diversity of the project, the literature repository includes works on, broadly, five different categories: Privacy, Fairness and Justice, Legal and Ethical concerns, Transparency and Explainability, and Accountability and Responsibility. Of course, also more general topics, that exceed these categories and that are deemed to be relevant for the whole SoBigData community, are not missing in the catalogue.

Under the umbrella category "Ethics and legality", the users of the SoBigData Literacy are able to find a broad range of cutting-edge publications covering topics related to, among others, issues concerning machine-learning fairness and legal challenges related to data protection and anti-discrimination laws. Also, questions concerning algorithmic fairness and the trust we are justified in attributing to automated systems have been integrated into the catalogue. Moreover, publications treating problems related to possible data biases, discrimination, and the exploration of ethical values, that should be considered central in assessing the impact of artificial intelligence and new technologies, are surely a central part of this thematic group of the catalogue.

Of course, also research articles regarding how to achieve explainability and transparency in machine-learning systems are present in the catalogue, addressing the crucial question of how to make interpretability conceptually coherent and technically feasible. Strongly connected to these issues are also questions regarding responsibility attribution and accountability related to the use of automated systems. A thematic cluster in the catalogue has also been dedicated to the latter thematic category. Of course, also issues related to information technology and the challenges represented by the protection of personal data are represented in the thematic category dedicated to privacy. Additionally, publications related to justice and fairness build a category on their own, in which literature treating, among others, the measurement

of algorithmic fairness and addressing ways of how to preserve democracy in a digitalized era plays a central role. Above and beyond the general organization of the literature in the thematic groups previously described, tags are assigned to each item so that also a more fine-grained thematic classification is possible, which renders the search for contents of interest particularly straightforward.

For every item in the catalogue, an abstract and all relevant metadata (author(s), DOI, publisher, journal/conference, source etc.) are available at a glance. This provides the users with a useful overview of the piece of research they could be interested in reading in its entirety.

Moreover, the catalogue can also be very easily used for citation purposes since every item is endowed with a BibTeX citation string that can be easily exported and integrated in one's own private literature management system. Besides, a direct link to the publication is present so that the physical paper/research article can be retrieved immediately.

Furthermore, a social networking area enables users of the environment to collaboratively contribute to the development of the selected literature by commenting on published contents as well as suggesting new contents of interest.

At the moment, 120 items have been integrated in the SoBigData Literacy, but the goal is to enrich the literature repository with new literature and updated research as an ongoing process, which should guarantee organizational accuracy and broad thematic representation. To achieve this aim, we would like to strongly encourage the active participation to the development and expansion of the SoBigData Literacy so that its relevance can be exploited at the most from the whole SoBigData community as a highly useful tool contributing to create a common knowledge base within the project.

Photo_credit_Chuttersnap_Unsplash

The SoBigData JupyterHub service

The SoBigData JupyterHub service has reached the production stage and it's now accessible from the SoBigData portal. The article describes the service's design and architecture, delivered in the context of the SoBigData++ project by the EGI Foundation.

Enol Fernandez, EGI Foundation | enol.fernandez@egi.eu

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

Notebooks are becoming the de-facto standards in order to provide an easy online coding system combined with interactive computing. In the context of the SoBigData++ project, the deployment and operation of a Notebook service integrated with the SoBigData e-infrastructure has been designed to satisfy the following requirements:

- The service should be integrated and accessible via a number of SoBigData++ Virtual Research Environments (VREs);
- The service should be integrated with the SoBigData Authentication/Authorization system;
- The access to the SoBigData Workspace should be implemented, in order to provide access to VRE Workspaces users files within the notebooks' environment;
- The users should be able to access different flavours of notebooks both in terms of software pre-installed and resource needed;
- In addition to the SoBigData Workspace, users need persistent storage to save a working copy of their notebooks across sessions;
- The users should be able to publish the notebook to the SoBigData catalogue.

Given the requirements, the EGI Notebooks service [1] has been selected in order to offer a notebook solution with seamless access from the SoBigData++ VREs graphical environment.

Figure 1 shows the reference architecture of the JupyterHub solution made available on the SoBigData e-infrastructure. The EGI Notebooks service is offered via a Kubernetes deployment and it has been selected to provide a way to automatically provision notebooks servers as containers, with the ability to select the image flavours to run and the resource limits (in terms of CPU and RAM). Kubernetes offers an easy way to scale the deployment, by adding new workers to the existing installation, thus allowing extending the capacity of the service if needed.

The EGI Notebooks service relies on the project Zero to JupyterHub[2] for the deployment of a JupyterHub cluster, and it offers many customizable hooks that are needed for the integration into the SoBigData environment. First of all, the possibility to extend the authentication system for the JupyterHub in order to use the SoBigData Authn/Authz framework. A new JupyterHub Authenticator class has been implemented for the purpose of taking a user personal token from a VRE and using it to get user account details stored in JupyterHub DB. JupyterHub extensions specific for SoBigData are available [2], together with other customizations developed for EGI needs. Further integration activities were required to expose the JupyterHub GUI within the VRE portal, this was needed in order to include the GUI as an iframe.

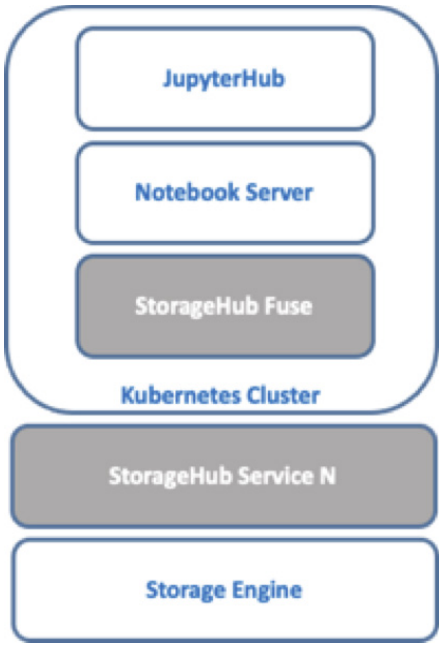


figure 1- The JupyterHub solution reference architecture

JupyterHub is also providing customizable hooks to implement a selection of Notebooks images to run and related flavours. A default image has been configured initially derived from the DataScience notebooks distribution, with further libraries installed by default. The image is automatically built and published to DockerHub [3] so it can be easily made available to the Kubernetes cluster. A default profiles list has also been included, so users can select

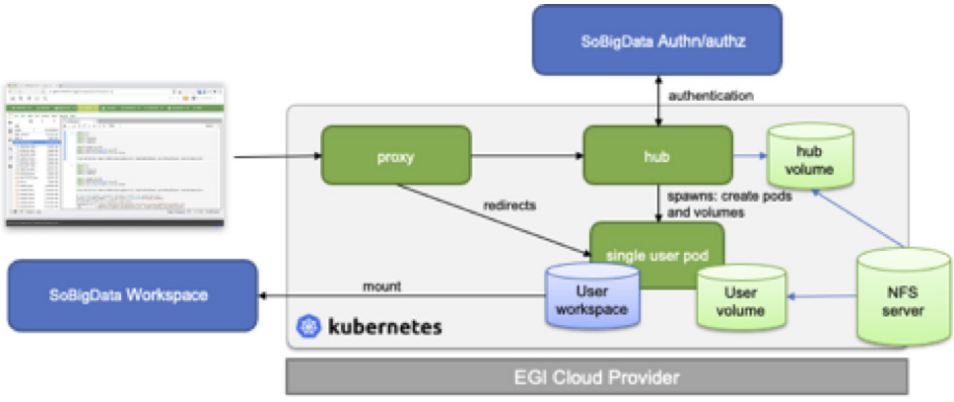
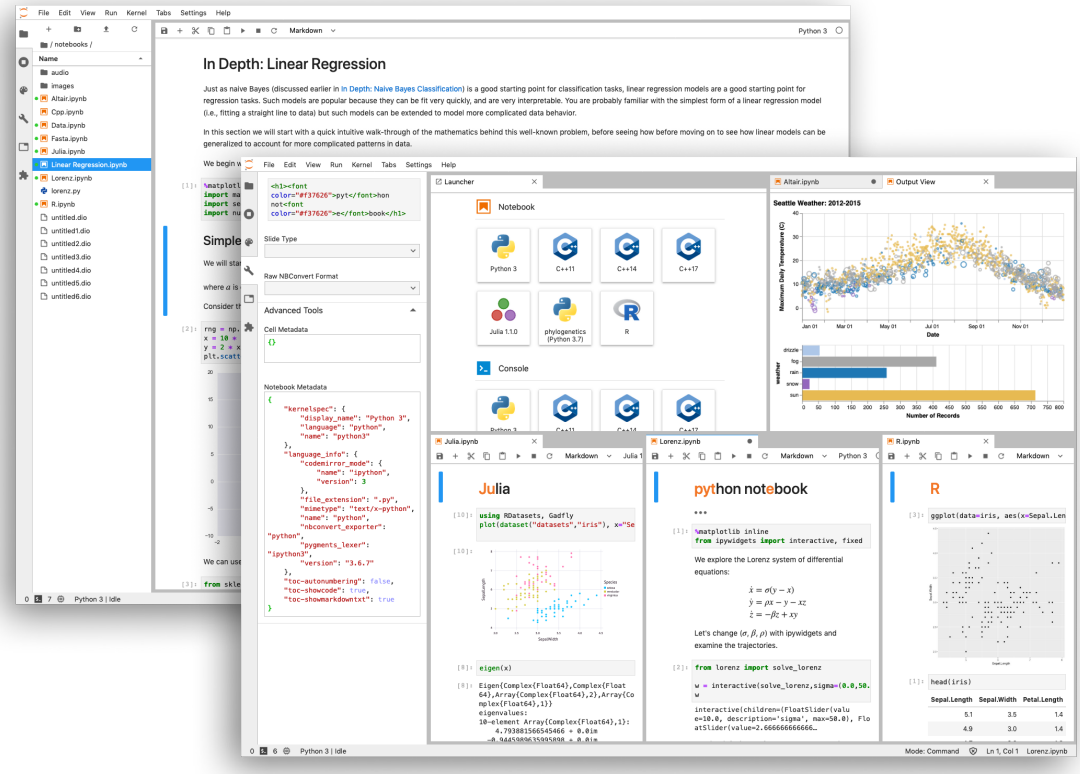


Figure 2 The JupyterHub deployment for the SoBigData e-infrastructure

the flavours of the notebooks to run.
Further integration steps have been performed in order to implement access to the D4Science workspace. The access is implemented via the StorageHub Fuse library, which allows mounting the user workspace folders in the notebook environment. The library integration allows the workspace to be mounted in a sidecar container of the user's notebook pod and made available under the user's home.
Finally, the Kubernetes cluster has been configured with an NFS server which is needed to host the users' volumes in order to offer persistent working folders. (More details of the integration activities performed are shown in Figure 2).
The first implementation of the service has been made available as a pilot to users of the SoBigDataLab VRE in July 2020 and selected members of the project have been able to do initial tests. In parallel, members of the SoBigData++ WP8 have also developed and requested the integration of new libraries as part of the notebooks default image and finally the production availability has been announced in March 2021. Since then, service has been actively used by the students of the Master in Big Data Analytics & Social Mining from the University of Pisa.
The currently deployed cluster has enough capacity to offer concurrent access to 80 users each with notebooks of 8 GB RAM (the RAM available on the cluster is the limiting factor). The initial uptake of the service will be monitored and the resources available on the cluster extended accordingly.
From the functional point of view, one of the requirements of the project is to let the user publish the notebooks to the SoBigData catalogue. Sharing of notebooks is already available through the Workspace and also publishing into the catalogue can be performed via the Workspace. However, a future extension to publish notebooks directly from the Jupyter environment will be analyzed and possibly implemented to further simplify and promote the publication of notebooks.
The current deployment of the service does not support multi-site clusters; hence the Kubernetes resources are provided by one D4science.org site (GARR Napoli). Investigations will be performed in order to support the deployment over multiple sites, which will also improve the availability of the service in case of upgrades and site incidents.



JupyterHub_photo_credit_jupiter.org

[Links]
[L1] https://www.egi.eu/services/notebooks/
[L2] https://github.com/EGI-Federation/egi-notebooks-hub
[L3] https://hub.docker.com/r/eginotebooks/single-user-sobigdata

SoBigData Events: virtually there!

From large conferences to smaller webinars, multiple events have continued to take place virtually this year. Here are some of the highlights demonstrating the project's ability to be flexible and accommodating whilst continuing to disseminate knowledge and cutting-edge research.

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

With the international travel situation continuing to be highly unpredictable and many countries still experiencing restrictions in some form or another, the SoBigData++ project has continued to plan and host all events in a virtual format. There has been a variety of events in 2021 – some of the highlights are detailed below.

DATA IN SOCCER: AN ATHLETIC TRAINER'S POINT OF VIEW

This Webinar took place on 12 February 2021 and introduced the concept of using data to tailor training programmes for athletes and maximise the effectiveness of the training schedule. Cristoforo Filetti (athletic trainer from Paris Saint-Germain) spoke for 30 minutes about how he uses data during workdays in order to schedule a training program. After the presentation, there was a 15 minute Q&A session where the moderators of the event asked questions about his work.

The recording can be seen via the SoBigData++ YouTube channel posted on both the SoBigData++ website and Twitter account. The webinar itself had 40 participants and 88 people have watched the recording on the SoBigData++ YouTube channel [L1].

3RD SOBIGDATA++ AWARENESS PANEL:

Medical Device Regulation and Digital Health: Problems and Perspectives

The 3rd Awareness panel [L2] organised by SSSA was held on 15 February 2021 and concerned the issues and considerations that arise regarding the use of digital devices for health purposes. Digital Humanities is an emerging field that brings together social scientists and computer scientists to encourage collaboration and allow research on a far larger scale than ever before. However, alongside advancements in technology there needs to be a parallel advancement in the regulations, ethical decisions and a robust legal framework to protect patients' privacy. This panel raised these problems and perspectives and opened up a discussion that proposed some interesting questions.

There were approximately 15 participants, mostly from a teaching or academic background including PhD Researchers, early career researchers and academics. The event also attracted individuals working in industry and policy makers.

ECIR 2021 (EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL)

This five day event [L3] took place virtually between 28 March – 1 April 2021. It was organised by CNR and was aimed at teaching and academic institutions, researchers and industry. Although the conference had to convert to a virtual format it maintained the same structure of previous ECIR conferences. The first day was devoted to tutorials and the Doctorial Consortium. The next 3 days were occupied by the main conference and the last day consisted of satellite workshops and the industry day.

On the first day, two full-day tutorials and six half-day tutorials were offered. The three days of the main conference featured 3 keynote talks, 50 full paper presentations (out of 211 submissions, for a 23.7% Acceptance Rate(AR)), 11 presentations from the Reproducibility Track (with a 47.8% AR), 7 presentations of papers recently published on the Information Retrieval Journal, 12 short-paper presentations of the CLEF 2021 Labs, 39 poster presentations of short papers (with a 28.5% AR), 15 demos (with a 48.4% AR), and a panel on the theme of "OpenAccess and IR Literature". Finally, the last day of the conference hosted the Industry Day and five satellite workshops.

This conference was aimed at researchers from academia and industry and the anticipated number of participants was approximately 200. Converting the conference to a virtual event meant the fee structure needed to be reconsidered. It was decided to charge a flat nominal fee of €150 to authors only and to grant a free registration to all other attendees. It was hoped this decision would maximise participation – especially from developing countries. This strategy proved effective with over 1,100 registrations from 63 countries. Of this number, 1,028 participants logged into the conference at least once. This jump in numbers represents a five-fold increase on expected numbers meaning the conference reached a far larger number of individuals than anticipated.

ROMCIR 2021 - REDUCING ONLINE MISINFORMATION THROUGH CREDIBLE INFORMATION RETRIEVAL

This workshop [L4] was organised by IMT as part of ECIR 2021 and took place on 1 April 2021. The central topic of ROMCIR concerns providing access to users to credible and/or verified information, to mitigate the information disorder phenomenon. In this context, "information disorder" means all forms of communication pollution, from misinformation due to ignorance, to the intentional sharing of false content. This topic is so broad as it concerns different contents (e.g., web pages, news, reviews, medical information, online accounts, etc.), different web and social media plat-

forms (e.g., microblogging platforms, social networking services, social question-answering systems, etc.), and different purposes (e.g., identifying false information, accessing information based on its credibility, retrieving credible information, etc.).
The event consisted of two keynotes speakers and six users presenting their contributions. The contributions were published on CEUR [L5].The event was directed at teaching and academic institutions, researchers and data analysts. It is estimated that there were approximately 80 participants.

THEMATIC WORKSHOP ON MIGRATION AND BIG DATA

This event was organised by Institut Convergences Migrations Paris [L6] and although was not a SoBigData++ event, it involved researchers who delivered talks which acknowledged SoBigData++.
It was the inaugural event of a series of 3 workshops and a training course dedicated to the subject of migration. The objective was to introduce the audience to the general topic and present state of the art research using big data for the study of migration and to offer a critical overview of the potential, as well as the challenges, posed by this type of data. The series of workshops will also identify the training needs in the methods used in this type of research for both early career researchers and experienced researchers.
This first workshop was a series of four talks, Alina Sîrbu's, 'Human Migration: the big data perspective' and Simone Bertoli's "Tell me what you eat, I will tell you who you are" both acknowledge the SoBigData++ project. The audience was interdisciplinary and included both students and researchers. There were 55 participants of which 30 were female and 25 were male.

AISC (AGGREGATE INTELLECT SCIENCE) - BENCHMARKING AND SURVEY OF EXPLANATION METHODS FOR BLACK BOX MODELS

Organised by SNS, this webinar took place online of 28 April 2021 [L7]. It was a 45 minute event with a 30 minute presentation of recent work on artificial intelligence followed by a Q&A session of approximately 15 minutes.
The widespread adoption of black-box models in Artificial Intelligence has enhanced the need for explanation methods to reveal how these obscure models reach specific decisions. Retrieving explanations is fundamental to unveil possible biases and to resolve practical or ethical issues. Nowadays, the literature is full of methods with different explanations. This study shows a visual comparison among explanations and a quantitative benchmarking of various explainers.
Participants were very interested in the topic asking multiple questions both during the Q&A session and after the event by email. Requests for follow ups and details about future works were also received. There were approximately 100 participants, of which 80 were male and 20 were female. The event was aimed at teaching and academic institutions, researchers and industry.

MOVING ON

There are many more events currently being planned and prepared for the coming months. It is certain that as soon as travel restrictions are lifted, many previous and new SoBigData++ partners and participants will be looking forward to enjoying 'in person' events and returning to the old normal. However, the lessons learned and the benefits that have been revealed during this unprecedented time will not be forgotten and many aspects of the virtual domain are sure to be incorporated into future events.

Screenshot of the OpenAccess and IR Literature panel.

Links
[L1] https://youtu.be/40PAAPum9V4
[L2] https://www.lider-lab.it/2021/02/12/third-sobigdata-awareness-panel-medical-device-regulation-and-digital-health-problems-and-perspectives-3/
[L3] https://www.ecir2021.eu/
[L4] https://romcir2021.disco.unimib.it/
[L5] http://ceur-ws.org/Vol-2838/
[L6] https://www.icmigrations.cnrs.fr/2021/03/10/conf-thematic-workshop-big-data-and-migration-wednesday-7-april-2021-9h30-12h30-online/
[L7] https://aisc.ai.science/

Special issue on Social Mining and Big Data Ecosystem for Open, Responsible Data Science

A collection of papers in the special issue entitled "Social Mining and Big Data Ecosystem for Open, Responsible Data Science", just published in the International Journal of Data Science and Analytics (JDSA). The special issue provides an account of a global scientific trend that is profoundly transforming science, providing it better means to foster social good.

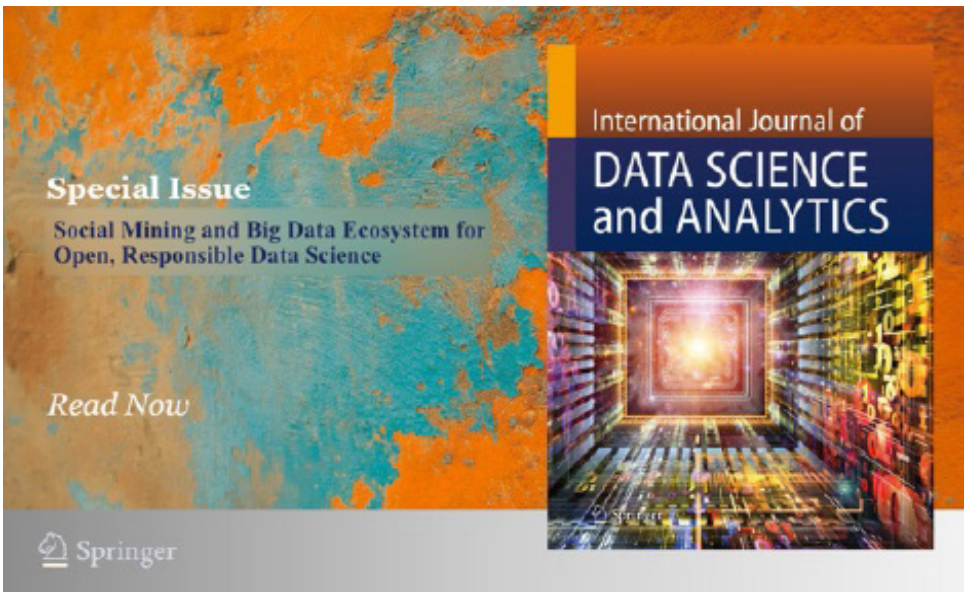
Valerio Grossi, ISTI - CNR Pisa | valerio.grossi@isti.cnr.it

Luca Pappalarto, ISTI - CNR Pisa | luca.pappalardo@isti.cnr.it

Dino Pedreschi, University of Pisa | dino.pedreschi@unipi.it.

Data became a part of our lives during the last two decades. They are generated as a by-product of our communication, analysis, or decision-making. In this perspective, Data Science is rapidly changing the way we do business, socialize, and govern society, and the way we make scientific research. A new paradigm is emerging, where theories and models and the bottom-up discovery of knowledge from data mutually support each other. If properly oriented at social good and human values, Data Science and Artificial Intelligence (AI) can help tackle the global challenges facing humanity, well represented in the Sustainable Development Goals (SGDs). The SGDs were set forth by the United Nations, and dramatically highlighted by the pandemic situation. In SoBigData RI, we also touch upon how big data and AI help us make informed choices, underlining the need to achieve collective intelligence without compromising the rights of individuals.

In this context, it is a pleasure to introduce the resulting collection of papers in the special issue entitled "Social Mining and Big Data Ecosystem for Open, Responsible Data Science" [L1], just published in the International Journal of Data Science and Analytics (JDSA). The special issue provides an account of a global scientific trend that is profoundly transforming science, providing it better means to foster social good.



Special Issue: A Social Mining and Big Data Ecosystem for Open, Responsible Data Science, Volume 11, issue 4, May 2021

Social Mining and Big Data Ecosystems for Open, Responsible Data Science are necessary to create global, inter-disciplinary communities of social data scientists fostered by extensive training, networking, and innovation activities. Such ecosystems are based on three pillars: infrastructures granting access to large datasets, analytical and AI tools, and data-driven experiments; communities of data scientists and AI experts; broad communities of users and stakeholders. The special issue is designed to target contributions that tackle socially relevant challenges in original and ethical ways, from data collection to model exploitment. Another objective of the special issue is to understand whether and how the emerging research infrastructures and the associated data science research ecosystems are enhancing the capacities of social data-driven research.

We are thankful to the SoBigData++ community for providing many excellent concrete examples of socially relevant and impactful data-driven research, also showing how the idea of a Social Mining and Big Data Ecosystems for Open, Re-

sponsible Data Science can boost a transformative effect in basically all scientific disciplines. Several of these examples are described in this special issue, which was launched with the idea of soliciting contributions from researchers and practitioners in data mining and other disciplines to share their research in big data analytics and data science applications.

We thank the Editor-in-Chief of the International Journal of Data Science and Analytics (JDSA), Professor Longbing Cao, for the opportunity to guest-edit this collection. We are also very thankful to the contributing authors and the reviewers who carefully examined the papers. We ended accepting six papers, covering a broad spectrum of challenging issues:

DATA SCIENCE: A GAME-CHANGER FOR SCIENCE AND INNOVATION.
This paper introduces how data science impacts science and society, including ethical and governance issues connected with managing data that touch upon aspects of human behavior. https://link.springer.com/article/10.1007/s41060-020-00240-2

MEASURING OBJECTIVE AND SUBJECTIVE WELL-BEING: DIMENSIONS AND DATA SOURCES.
This work illustrates the approaches for measuring well-being. The authors distinguish between objective and subjective well-being and surveys the theoretical background, the relevant dimensions of well-being, the new data sources for measurement, and relevant recent studies. https://link.springer.com/article/10.1007/s41060-020-00224-2

(SO) BIG DATA AND THE TRANSFORMATION OF THE CITY.
This paper discusses the main issues of urban data analytics, focusing on privacy issues, algorithms, applications, and georeferenced data from social media. As concrete case studies of urban data science tools, the authors leverage the results obtained in the "City of Citizens" thematic area of the SoBigData initiative, which includes a virtual research environment with mobility datasets and urban analytics methods by several institutions around Europe. https://link.springer.com/article/10.1007/s41060-020-00207-3

HUMAN MIGRATION: THE BIG DATA PERSPECTIVE.
In this paper, the authors answer the question "How can big data help to understand the migration phenomenon?" through an analysis of various phases of migration, comparing traditional and novel data sources and models at each phase. They focus on three phases of migration - the journey, the stay, and the return - at each phase describing state of the art and recent developments and ideas. https://link.springer.com/article/10.1007/s41060-020-00213-5.

A WORKFLOW LANGUAGE FOR RESEARCH E-INFRASTRUCTURES.
This work outlines the HyWare language and platform. The language is an extension of the traditional workflow languages enabling the definition of workflows, including automatic and manual analytical steps to replicate and build large-scale data-driven experiments. https://link.springer.com/article/10.1007/s41060-020-00237-x

AN ETHICAL AND LEGAL FRAMEWORK FOR SOCIAL DATA SCIENCE.
This paper provides a framework for research infrastructures that enable ethically sensitive and legally compliant data science, helping data scientists frame the appropriate self-assessment questions to ensure an ethical, responsible design, implementation, and deployment of data science projects. https://link.springer.com/article/10.1007/s41060-020-00211-7

SoBigData RI is proud to offer this collection to the attention and scrutiny of the scientific community, and we recall that all the papers are published under Open Access rules.

Links:
[L1]: https://link.springer.com/journal/41060/volumes-and-issues/11-4
The other links are near to the article, but if you prefer for formatting reasons you can move altogether here.

All the ties that bind: a socio-semantic analysis of the Italian Twittersphere

At the intersection between social sciences and statistical physics, our new methodological approach paves new ways for defining online collective identities in a fully data-driven fashion.

Tommaso Radicioni, Scuola Normale Superiore | tommaso.radicioni@sns.it

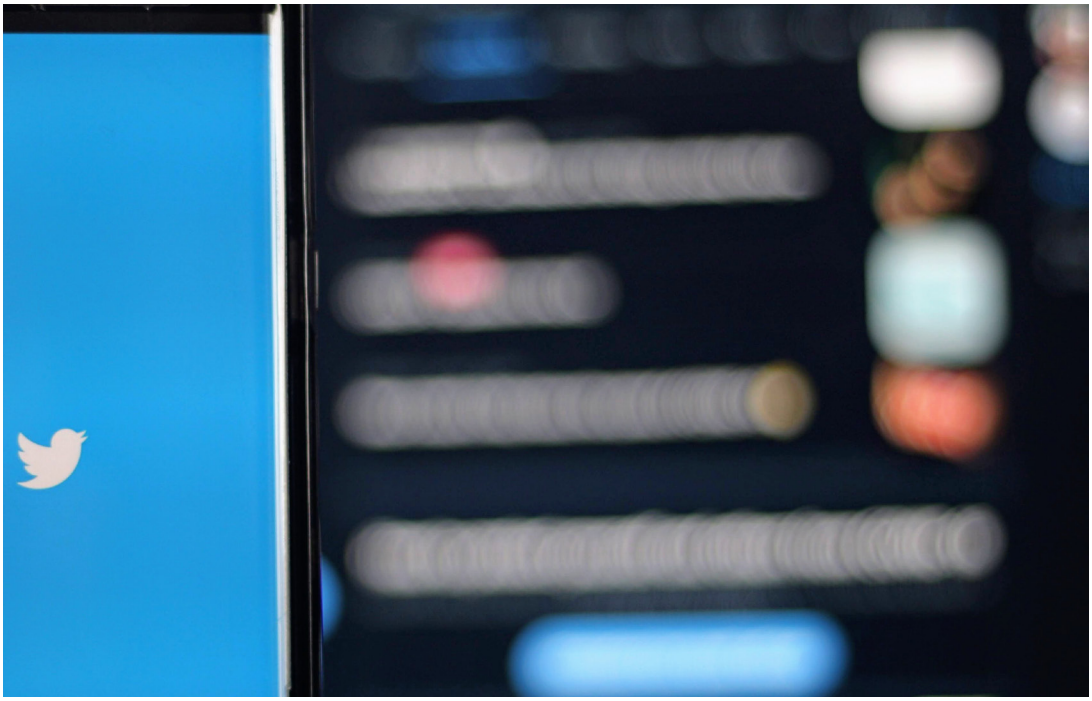
Tiziano Squartini, IMT School for Advanced Studies | tiziano.squartini@imtlucca.it

Fabio Saracco, IMT School for Advanced Studies | fabio.saracco@imtlucca.it

The huge amount of data made available by the massive usage of social media has opened up an unprecedented possibility: studying the online human behaviour in a fully data-driven fashion. Our research, supported by SoBigData++ and lying at the intersection between social sciences and statistical physics, paves new ways for defining online collective identities. According to the Euro-barometer [L1], the percentage of EU citizens who employ social media on a daily basis has increased from 18% in 2010 to 48% in 2019. Although the growth of social media usage has dramatically increased the availability of data, it has also required the definition of novel modelling frameworks capable of extracting useful information out of the aforementioned data deluge.

Amongst the stylized facts that have most attracted the attention of researchers, a special mention is devoted to the phenomenon known as echo-chambers formation. The latter, defined as homophilic groups of social media users, are a consequence of the combined effect of several socio-technological mechanisms, such as the online political fragmentation, the tendency of individuals to select media contents adhering to their opinions and the algorithmic biases induced by social media recommendation engines. All these mechanisms reinforce the identity of those communities of users who share a common narrative, in turn increasing their group polarization, i.e. the tendency of moving towards more extreme opinions or contents while rejecting political talk with opposing viewpoints [R1]. We are interested in understanding the effects that the formation of these polarized communities have on both the social and the semantic dynamics within a social media platform such as Twitter.

More in detail, our work expands the current analyses of online conversational dynamics by adopting a filtering procedure, rooted into statistical physics, that allows us to project the initial bipartite networks on the users and the hashtags layers, thus obtaining two monopartite networks. On the one hand, our technique identifies discursive communities as groups of users with a similar (re)tweeting behaviour; on the other, it allows us to identify the most prominent contents circulating within each of them by extracting the corresponding se-



Photo_Credit_joshua-hoehne_Unsplash

mantic network - induced by its own users via hashtagging practices.

Our methodological approach, based on the constrained optimization of Shannon entropy [R2], is innovative in a three-fold way: first, it guarantees that the similarity between any two users, as well as that between any two hashtags, is inferred in a statistically unbiased way; second, it does not require any manual intervention such as that of labelling the users features or the sharing patterns; third, it is completely data-driven, as it takes as input data the users' (re)tweeting behavior - hence, implementing a bottom-up instead of a top-down approach which typically characterizes previous researches on communtiies of social media users [R3]. The data we have employed for our analyses concern the online conversations that unfolded on Twitter (i) in the run up of the 2018 Italian Elections and during the two weeks after the Election day [R4] and (ii) about immigration policies across the period May-November 2019 [R5].

On the social side, the main result of our work concerns the identification of discursive communities that not necessarily match with the ones expected solely on the basis of the political leaning of their users. While this is indeed the picture emerging from the data concerning the 2018 electoral campaign - during which our discursive communities overlap with the political coalitions running for the 2018 Italian general elections (see the left panel of Figure 1) - this is no longer true for the 2019 discussion on migration policies: as clearly shown in the right panel of Figure 1, the online behavior characterizing the users interacting with the Five Stars Movement and the League verified users induces two distinct communities also during the period in which these two parties were formally allied (and jointly shared seats within the first Conte government).

On the semantic side, our analysis sheds light on the cognitive dimension of partisan dynamics [R6]. Different discursive communities are, in

Image by Free-Photos from Pixabay

fact, characterized by different conversational dynamics at both the daily and the monthly time-scale. One of the main findings of the analysis of the daily discursive dynamics, during the 2018 electoral campaign, concerns the way the topological structure of semantic networks "reacts" to the so-called mediated events (e.g. TV debates, the media coverage of offline events, etc.), revealing the diverse behaviour of right-wing and left-wing users towards these events. Particularly insightful is the analysis of our semantic networks at the mesoscale: what emerges is the presence of a densely connected bulk, surrounded by a periphery of loosely inter-connected hashtags, indicating that semantic networks are typically characterized by few relevant hashtags to which far less relevant ones attach.

Albeit non representative of the entire Italian population, our analysis of Twitter data concerning two distinct online political conversations (about the 2018 electoral campaign and the 2019 discussion about migration policies) provides interesting results about the construction of online collective identities - that can be also thought as a starting point to develop more fine-grained analyses of the voters' political opinion and behaviour. Besides, the generality of our approach makes it suitable to study large-scale Twitter conversations in all domains and regardless of the language in which they take place.

Links:
[L1]: https://op.europa.eu/en/publication-detail/-/publication/c2fb9fad-db78-11ea-adf7-01aa75ed71a1/

References:
[R1]: C. R. Sunstein. The law of group polarization. Journal of Political Philosophy, 10(2):175–195, 2002.
[R2]: F. Saracco, M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, and T. Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. New Journal of Physics, 19(5):053022, May 2017.
[R3]: M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pages 192–199, Oct 2011.
[R4]: T. Radicioni, E. Pavan, T. Squartini, and F. Saracco. Analysing Twitter Semantic Networks: the case of 2018 Italian Elections. arXiv e-prints, page arXiv:2009.02960, Sep 2020.
[R5]: T. Radicioni, T. Squartini, E. Pavan, and F. Saracco. Networked partisanship and framing: a socio-semantic network analysis of the Italian debate on migration. ArXiv e-prints, page arXiv:2103.04653, Mar 2021.
[R6]: M. Stella. Cognitive network science for understanding online social cognitions: A brief review. ArXiv e-prints, page arXiv:2102.12799, Feb 2021.

Data Inquiries: an Innovative Model for Dathatons

In this short communication we present a new initiative recently launched by the Center for Internet and Society of the CNRS in collaboration with the Public Data Lab to facilitate the encounter between data experts and civil society actors around concrete projects of data interventions.

Tommaso Venturini, CIS-CNRS | tommaso.venturini@cnrs.fr

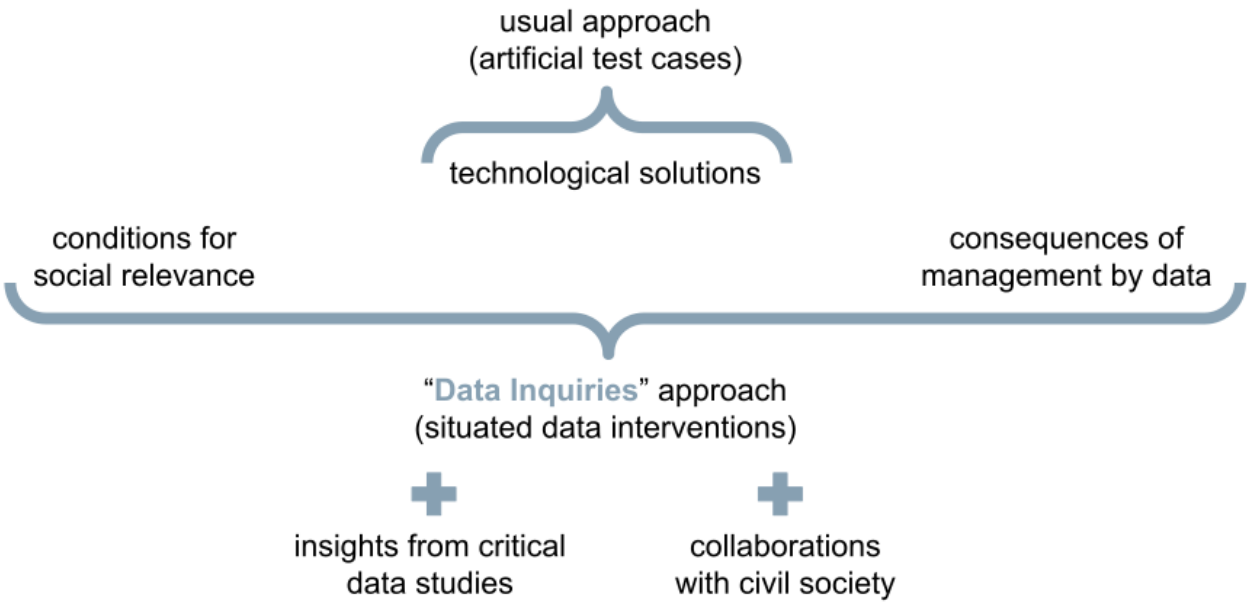
Axel Meunier, CIS-CNRS | tommaso.venturini@cnrs.fr

The role of the Center for Internet and Society of the CNRS (CIS-CNRS) in the SoBigData++ project is to offer advice and support to other Consortium members wishing to organise datathons and other workshopping activities as a way to put social big data at the service of societal needs. Over the years, we have acquired a tried and tested expertise in this kind of interventions, and we are among the funders of the Public Data Lab, an interdisciplinary and international network exploring what difference the digital makes in attending to public problems [L1]. The object of the PDL is precisely to develop materials and formats for collective inquiry with and about digital data, digital methods and digital infrastructures.
Recently, the Public Data Lab and the CIS-CNRS launched a new initiative called "Data Inquiries" which intends to bring together these materials and formats and to make them available to teachers and practitioners of data science [L2].

The Data Inquiries website is meant to grow into a resource center addressed to different actors interested to social big data interventions:

- To teachers and students, it offers a combined training in data literacy and critical data studies, as well as occasions to engage in real-life data projects.
- To data scientists, it offers occasions to put their skills and ingenuity at the service of societal causes, not in abstract but in the context of actual social interventions.
- To civil society groups, it offers a platform to promote their initiatives and open them to external contributions.

The website already offers a detailed description of a workshopping format



Data Inquiries Chart

Photo_Credit_nubelson-fenandes-unsplash

we develop to organize data intervention (e.g. the Data Sprints, [L3]) and an example of a syllabus to turn this approach into a teaching module [L4].

But Data Inquiries is also more than a resource center and introduces a very specific approach to social data science that we wish to propose to the members of the SBD++ consortium. Data Inquiries suggests, in particular, to shift the attention away from data and technologies and to the social and political implications of data research. Drawing on test cases, standard datasets and pre-defined challenges, many datathons end up hiding the complexity that lies upstream and downstream most data projects:

• Upstream, in the task of matching the data available (or obtainable) to research questions and social interventions that are actually useful for society.

• Downstream, in the consideration of the political consequences of managing social phenomena through their computational representation.

Bracketing out of such complications facilitates the development of technical solutions and the learning of technical skills, yet it also disconnects data sciences from its social conditions and consequences. Against this risk Data Inquiries draw attention to the social life of data (see fig.1) both conceptually and practically:

• Conceptually, it proposes to connect the practice and teaching of data research to the transdisciplinary literature on critical data studies and its description of the implicit assumptions and side effects of data infrastructures.

• Practically, it suggests experimenting and teaching data research not in abstract challenges, but in actual social situations, that is in collaboration with civil society groups using data in their projects.

Finally, Data Inquiries is what we will make of it together. If you have an idea about how to intervene in a specific social situation through data collection, analysis or visualization, please get in touch – we will be happy to provide advice on how to turn this idea into a workshop.

Links:
[L1]: https://publicdatalab.org
[L2]: http://datainquiries.publicdatalab.org
[L3]: http://datainquiries.publicdatalab.org/WorkshopExample.html
[L4]: http://datainquiries.publicdatalab.org/SyllabusExample.html

Investigating the effect of Academic Networks on Academic Migration using evidence from Scholarly Big Data and the Iron Curtain

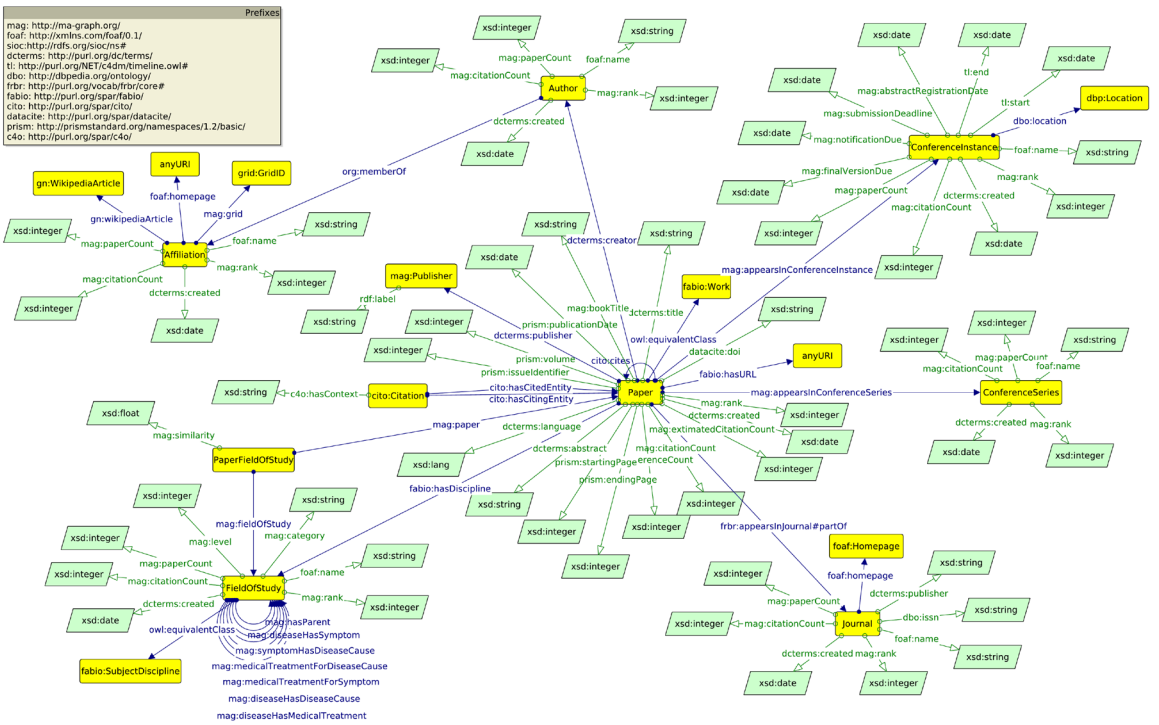
Iron Curtain and Big Data are two words usually used to denote completely different eras. Yet, the context the former offers and the rich data source the latter offers, complement each other in a perfect manner enabling us to causally identify the effect of networks on migration.

Hillel Rapoport, Paris School of Economics | hillel.rapoport@psemail.eu

Donia Kamel, Paris School of Economics | kamel.donia@psestudent.eu

The relationship between migration and networks is an inherently hard task, empirically. The reason behind this difficulty is two-fold. One lies in the data and how traditional data sources for each are flawed. Traditional migration data sources provide a stock measure for each year and thus limits the ability to anticipate immigrants' movement, which, as you expected from the title, could be done by closely examining the individual's social network. Additionally, data on social networks are very hard to collect due to expansiveness and lengthiness of the process. The second difficulty lies in the endogenous nature of networks that are usually utilised to facilitate migration by individuals, thus, to identify the sole effect of networks on migration, ruling out reverse causality (where migration affect network) and the scope for other variables affecting network and hence migration, is a difficult task.

This research abstracts from these issues by using Microsoft Academic Knowledge Graph that provides detailed information on academics from 1960-2020, and the big data aspect comes from having millions of observations. Such information is presented graphically in figure 1. Figure 1 helps in understanding the data extraction process: for each paper published, an ID is provided, the authors, of this paper, are stated, with their respective IDs, year, field, and affiliation, citation count, paper count, rank and more. As such, networks based on several characteristics can be mapped. By focusing on academics with affiliations



This is the schema that shows the nature of the data and how information is extracted. Based on this diagram, identification of authors, their network, affiliations and consequently locations of each, is achieved. Based on locations and quality proxies, the network size and quality dimensions are built.

Photo_Credit_serhat-beyazkaya-unsplash

of countries behind the Iron Curtain and those who are observed in the data from 1980-1988, and their networks, we are able to achieve identification, i.e., the causal effect of network on migration.

This is due to the fact that the Iron Curtain served a barrier halting migration between the East and the West. Thus, its existence inhibited the anticipation of migration and the consequent manipulation of networks at destination (defined as countries migrants migrated to between the period of 1989 (marking the fall of Berlin Wall) to 2003 (marking the year many Eastern European countries signed treaties and held referendums to join the EU and thus another migration wave). In other words, intuitively and also empirically, networks at home and at destination were not strategically altered by academics to facilitate migration, thus providing evidence in support of the identification strategy. Additionally, the Iron Curtain ensures that the affiliation of academics, hence their location and nationality pre the fall of the Berlin Wall/Iron Curtain, coincides with their actual nationality, simply because they could not migrate to other countries.

The second concern when it comes to causal identification is ruling out that other variables that shape networks affect also migration decision. To mitigate selection into migration, age (proxied), home-destination pair, quality, prior migration to EE countries that later became part of the EU, field, language are controlled for (De la Croix et al., 2019; Gould and Moav, 2014) [R1, R2]. Robustness tests are done to ensure the robustness and validity of the results. Looking at summary and descriptive statistics we find that the majority of migrants go to the US, that migrants are younger and that they have larger networks prior the fall of the Berlin Wall. By investigating the effect of the size of the network at home and at destination we see that as the size of the network at home increases, the probability to migrate post 1988 falls, and as the size of the network at destination increase, the probability to migrate post 1988 falls. The sign and statistical significance of both effects remain the same throughout various specification and econometric models of estimation, giving support for the results. The economic significance, which looks at the magnitude of coefficients, is also important to investigate. An increase in the size of the home network only decreases the probability to migrate by 0.03% whilst an increase in the size of destination network increases the probability to migrate by 1.5%. Looking at the quality of the network at home and destination is also important and is an advantage of this data as we have information on citations, paper count and rank. It is important to note that by focusing on the size and quality of the networks prior the fall of the Berlin Wall/Iron Curtain, we abstract from the implicit assumption that is extensively made in migration literature that assumes that migrants benefit equally from networks abroad. An important exception to this literature is Blumenstock (2019) [R3] which also uses big data in the form of phone call records in Rwanda and investigates how the size of the network and other characteristics of it affect migration decisions of migrants. The data for this quality-network-analysis is still underway, so stay tuned for more insightful evidence on the effect of networks on migration.

This research is part of a microproject initiative in WP10 and is the current micro-project in the Migration Exploratory, under the leadership of Professor Hillel Rapoport. This is a collaboration between Paris School of Economics and University of Pisa (especially Dr. Laura Pollacci, which provided tremendous support in terms of data extraction and cleaning, and her personal work will culminate into another dataset and a data paper). The research was accepted in January 2021 and started in February 2021 with an expected end date of July 2021.

Links:
[L1]: https://makg.org/
References:
[R1]: Gould, E.D. and Moav, O., 2016. Does high inequality attract high skilled immigrants? The Economic Journal, 126(593), pp.1055-1091.
[R2] : De la Croix, D., Docquier, F., Fabre, A. and Stelter, R., 2020. The Academic Market and the Rise of Universities in Medieval and Early Modern Europe (1000-1800).
[R3]: Blumenstock, J.E., Chi, G. and Tan, X., 2019. Migration and the value of social networks.

TransNational Access is now open! Submit your application!

We welcome applications from individuals with a scientific interest, professionals, startups and innovators that may benefit from training in data science and social media analytics.

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.

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.

> 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!

Researchers, professionals, start-ups and innovators are encouraged to apply. You will benefit from data science and ocial 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.

City indicators for Mobility Data Mining

How can we define the characterization of a geographical area? In this article, the authors describe how this can be achieved through a range of quantitative measures that provide a multilayer description of urban regions and are a means for displaying differences between cities, municipalities, or other geographical units.

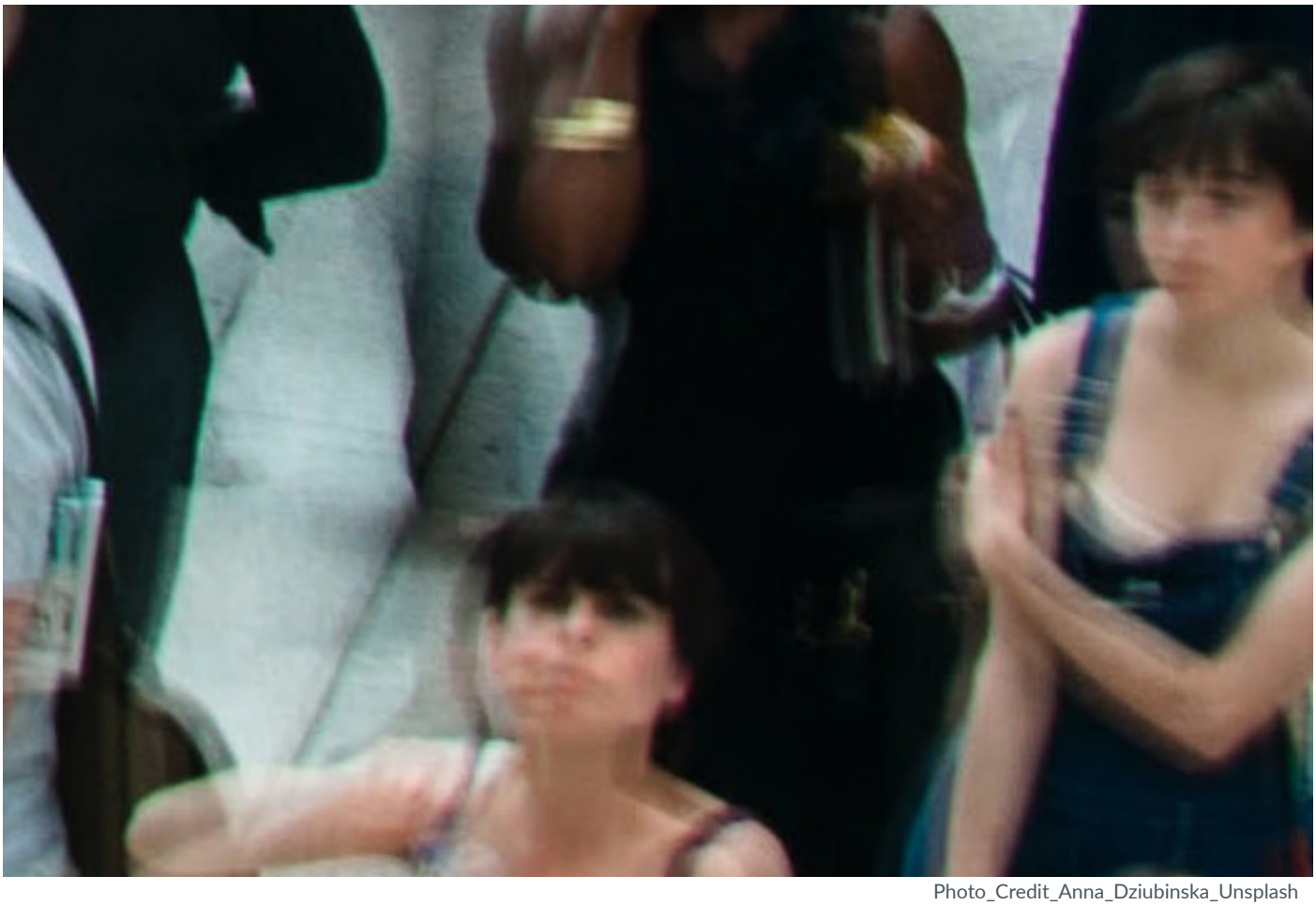
Mirco Nanni, CNR | mirco.nanni@isti.cnr.it

Agnese Bonavita, Scuola Normale Superiore | agnese.bonavita@sns.it

Riccardo Guidotti, Università di Pisa | riccardo.guidotti@unipi.it

Classifying a geographical territory into semantic categories is one of the most common tasks in research areas such as urban geography, urban planning and mobility data analytics [R1]. Characterizing human mobility is a key component of this process, and it is well known that mobility often does not work the same way across different regions. A movement pattern in a mountainous countryside may have other implications than the same pattern has in the suburbs of a large town. The movement trajectories in a planned city with rectangular streets and strict zoning laws might be completely different than the ones in a town that has grown organically without any clear structure. Therefore, any kind of property that was learned in a particular area, in general cannot simply be assumed to hold in another one.

The paper presented in this article aims at making a first step towards the characterization of a geographical area. That is achieved through a range of quantitative measures that provide a multilayer description of



Photo_Credit_Anna_Dziubinska_Unsplash

urban regions and are a means for displaying differences between cities, municipalities, or other geographical units. Such a numerical description of urban areas can have a wide spectrum of applications. Among them, the measures presented in this work can be used as an input for geographical transfer learning, that is the transformation of knowledge gained in one geographical region in order to apply it to another region. This problem will be considered as a case study for the extracted indicators.

Authors consider two main approaches: (i) computing features that describe each area isolated from the others, that we call local city indicators; and (ii) computing features that describe its relation with the others, named global city indicators.

The first group covers four different families of measures: spatial concentration indexes of human activities; network features of intra-city traffic flows; mobility characteristics of the individual mobility, obtained from networks that represent the places and movement of single users; last, characteristics of road networks and how traffic is distributed in them. The group of global city indicators, instead, looks at the mobility between cities as a graph, where each city is represented by a node, and extracts network features for each node. Both the complete network and the ego-network for each city are considered.

After describing all the city indicators, a mobility prediction problem is introduced, and authors use it to test how many predictive models are transferable across different regions. In particular, we study the relationship between transferability between two areas, i.e. the performances of a model built on one area and used to make predictions on the other one, and their similarity in terms of city indicators. The results confirm our hypothesis that cities with similar indicators are more likely to be transfer-compliant, this providing a first guide to understand which predictive models can be reused in other areas.

Finally, a key feature of this work is that all methods are implemented in a way that makes it possible to automatically calculate all characteristics for hundreds of different cities and entire regions. The resulting software (a Python library) enables the user to process an unlimited amount of data simply by passing a database with trajectories and a list containing the positions of the geographical areas of interest as an input.

The paper describing this research has been accepted to the workshop "BMDA 2021: Fourth International Workshop in Big Mobility Data Analytics". This work is partially supported by the European Community H2020 programme under the funding scheme Track &Know (Big Data for Mobility Tracking Knowledge Extraction in Urban Areas), G.A. 780754, [L1] and SoBig-Data++, G.A. 871042, [L2].

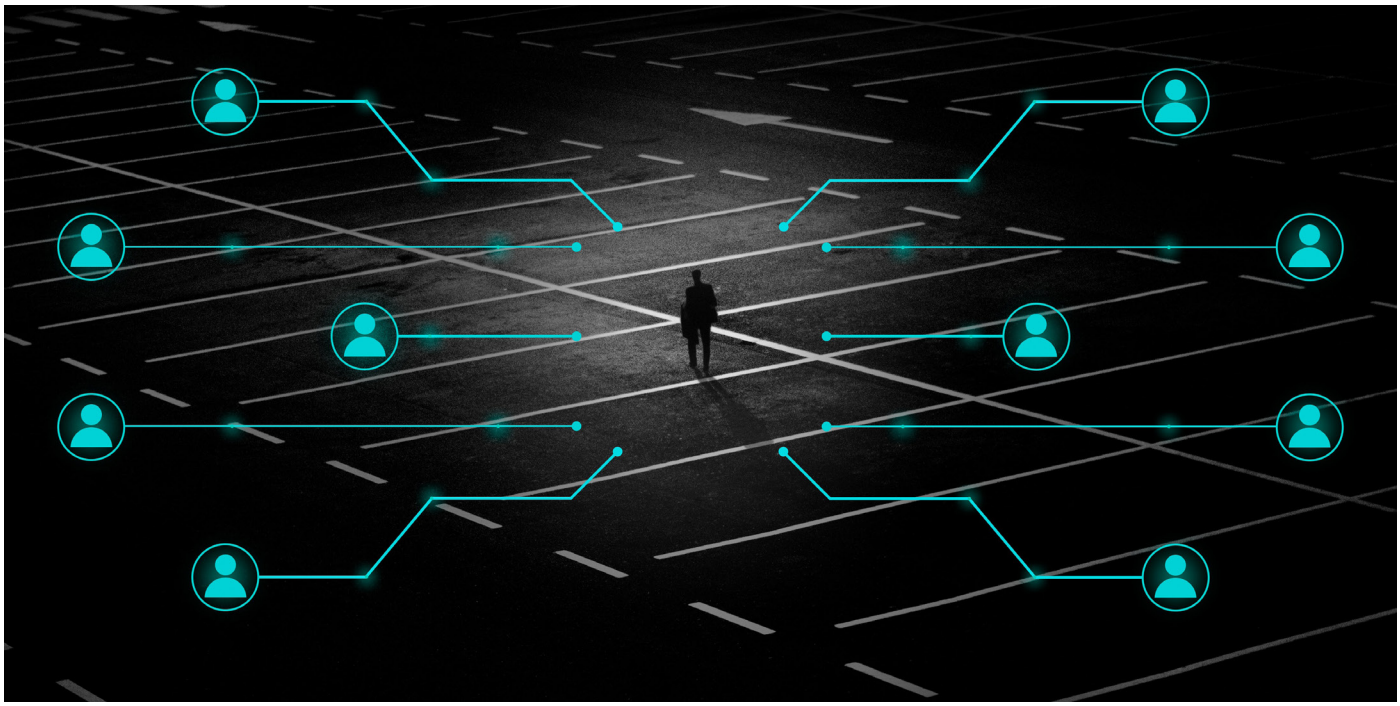


Image by Buffik from Pixabay

Links:
[L1] https://trackandknowproject.eu/
[L2] http://www.sobigdata.eu

References:
[R1] Harold Carter. 1995. The Study of Urban Geography. E. Arnold publications.
[R2] Gennady Andrienko et al. 2020. (So) Big Data and the transformation of the city. International Journal of Data Science and Analytics (2020).
[R3] European Commission and European Investment Bank, 2016. "Smart Cities & Sustainable Development" Program in Europe.

Predicting seasonal influenza using supermarket retail records

Generating real-time epidemic forecasts through novel digital data streams and machine learning approaches has become increasingly relevant. In this article (here proposed as an excerpt of the full text), the authors propose a novel, high quality data source, particularly retail market data, as a proxy for seasonal influenza nowcasts and forecasts.

Ioanna Miliou, Unipi, CNR

Dino Pedreschi, Unipi

Recent years have seen a growing interest in generating real-time epidemic forecasts through novel digital data streams and machine learning approaches. Seasonal influenza forecasting approaches are leading the way in this rapidly advancing research landscape. Seasonal influenza is still a major burden to the health care systems of countries with 3 to 5 million infected, and 290,000 - 650,000 deaths caused by influenza worldwide every year [L1]. For this reason, the US Centers for Disease Control and Prevention (CDC) formally pioneered infectious disease forecasting by starting the Flusight consortium focused on prediction of seasonal flu incidence. The CDC seasonal influenza challenge has been remarkably successful in maintaining momentum for a coordinated focus on the operational implementation of disease forecasting.

Simultaneously, it fuels the research on developing forecasting models based both on traditional surveillance systems such as influenza-like illness (ILI) incidence captured by the network of outpatient clinics, and novel digital data streams such as search engine queries and social media [R1–

Photo_Credit_Josh_Cameron_Unsplash

2]. In this context the use of machine learning techniques has received considerable attention [R3], and although the use of novel digital data streams as proxy data for disease forecasting did show evident limitations in early approaches, the use of multiple data sources and ensemble of models is now defining the second generation of forecasting tools defining the state of the art in the field.

[...]

Here we propose a novel, high quality data source, particularly retail market data, as a proxy for seasonal influenza nowcasts and forecasts. The assumption behind the use of this dataset is that items purchased in a shopping cart are a good proxy of consumers' behavioral changes, thus allowing to capture the spread of seasonal flu reflected in a specific set of supermarket purchases. More specifically, we first identify a set of sentinel products whose volume of purchase is historically correlated with the previous flu season. In order to avoid the use of spurious correlations and seasonal predictors (items generally available during the flu season but not related to flu), we consider the whole purchase history of customers buying sentinel products. This allows the identification - with an Apriori algorithm - of sentinel baskets, i.e., products bought together that we can use as a proxy for the actual seasonal flu.

By using sentinel baskets purchases, we develop a nowcasting and forecasting algorithm that provides seasonal flu incidence in Italy estimates up to 4 weeks ahead of the regular surveillance system. We make use of the Support Vector Regression (SVR) model to produce our predictions. We need to emphasize that the most important component in our framework is the data proxy - sentinel baskets - and that any other forecasting method can be applied in this framework.

Our results show that exploiting the information hidden in the retail market data can contribute to predicting the future incidence of influenza. Our findings indicate that the seasonal influenza forecast accuracy improves with the use of retail records and our predictive framework outperforms the baseline autoregressive model with historical ILI reports. More specifically, with two-week and three-week forecasts ahead, forecast performance indicators improve consistently with error estimates decreasing of about 50%. In order to support the rationale behind our choice of sentinel baskets as a proxy for predicting seasonal influenza, we introduce a second baseline using single products' time series of retail market data. Forecasts obtained by using sentinel baskets are significantly more accurate than those obtained using single products' time series. It's not the predictive power of our framework that is important, but rather the increase of the predictive power when we add the sentinel baskets that capture hidden human behaviors adapted to ongoing influenza epidemics.

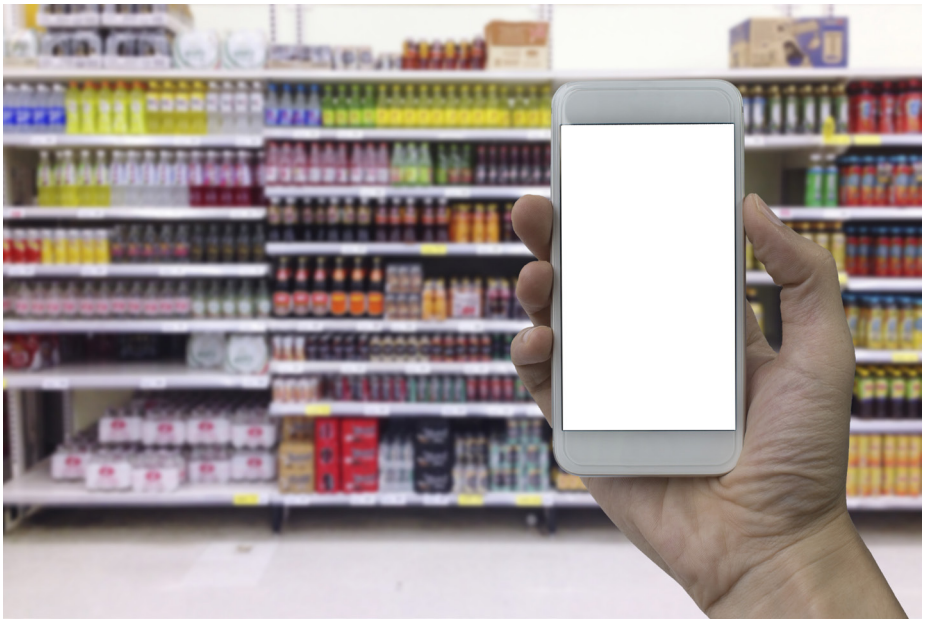


Image by achirathep from Pixabay

The presented work shows quantitatively the value of incorporating retail market data in forecasting approaches, adding one more dataset to the armory of proxy signals that can be used for the real-time analysis of epidemics. The framework developed in this paper has shed lights on the great potential of combining other predictive approaches (e.g., mechanistic models and/or deep learning models) and assimilating algorithms based on different proxy data [R4], thus defining ensemble forecasting methodologies that have proven to achieve the reliability required in the policy-making process.

Full text available as pre-print on arXiv [L2]

[Links]

[L1] https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)
[L2] https://arxiv.org/abs/2012.04651

[References]

[R1-2] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences. 2012;109(50):20425–20430 and Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PloS one. 2010;5(3):e9450
[R3] Adhikari B, Xu X, Ramakrishnan N, Prakash BA. Epideep: Exploiting embeddings for epidemic forecasting. In: KDD 2019; 2019. p. 577–586.
[R4] Perrotta D, Tizzoni M, Paolotti D. Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. In: Proceedings of the 26th International Conference on World Wide Web - WWW '17. ACM Press; 2017. p. 303–310.

Data on migration and integration policies and trends in Europe

Researchers have carried out systematic comparative analyses of migration and integration trends as well as analysed policies on migration and integration which are key as they not only influence migrants' ability to enter a country, but also the possibility to remain in the country, and their quality of life there.

by Giacomo Solano, Migration Policy Group – MCG | gsolano@migpolgroup.com

Photo_Credit_Christian_Lue_Unsplash

Migration and integration issues are at the center of the political and public debate in Europe. The scale of international migration has increased further in recent years. According to the International Organization for Migration (IOM) [L1], the number of international migrants is estimated to be almost 272 million globally (3.5% of the world population). For this number of migrants, migration and integration policies are key as they not only influence their ability to enter a country, but also the possibility to remain in the country, and their quality of life there.

Researchers have carried out systematic comparative analyses of migration and integration trends (e.g. migration inflows and integration outcomes of migrants) as well as analysed policies on migration and integration. Several data sources are available to this end.

Regarding migration and integration policy, over the last twenty years, researchers [L2] have undertaken systematic comparisons of migration policies by creating sets of indicators at the national level, and then aggregating them into an index. Indicators and indexes have the purpose of (i) understanding the nature of migration policy, (ii) allowing for cross-country comparison over time, and (iii) monitoring the evolution of policy frameworks. Researchers have developed these indicators on a wide range of areas, including admission policies [e.g., IMPIC (L3)], citizenship acquisition [e.g., GLOBALCIT (L4)], and integration policies [e.g., MIPEX (L5)].

Eurostat [L6], which is the statistical office of the European Union, provides a wide range of statistics on migration trends and integration outcomes. Data is gathered through National statistical offices and EU-surveys. Figures are generally available on both (a) people that were born in a country different from the country where they reside, and (b) people that have the citizenship of a country other than the one where they reside. Data are available on three main areas:

- Demography and migration: which includes statistics on foreign-born and/or foreign-national population, acquisition of citizenship, and marriages.
- Asylum and managed migration: including statistics on asylum applicants, residence permits, enforcement of immigration legislation, and children in migration.
- Migrant integration: which includes information on different domains of integration of migrants in their country of destination (e.g. education, employment, active citizenship, social inclusion, housing, poverty, and health).

Furthermore, several European-wide social surveys, either commissioned by EU institutions or coordinated at European level, contribute to a better understanding of integration-related issues. They include general population surveys such as the EU Labour Force Survey (EU-LFS) [L7], the EU Statistics on Income and Living Conditions Survey (EU-SILC) [L8], the European Social Survey (ESS) [L9], Eurobarometer [L10], and the European Values Study (EVS) [L11].

In conclusion, further research employing data from these sources would help in understanding the role that migration and integration policies play in influencing migration trends and migrants' integration outcomes.

Links:
[L1] https://publications.iom.int/system/files/pdf/wmr_2020.pdf
[L2] https://migrationresearch.com/migration-policy-indicators
[L3] http://www.impic-project.eu/data/
[L4] https://globalcit.eu/citizenship-law-indicators/
[L5] https://www.mipex.eu/
[L6] https://ec.europa.eu/eurostat/data/database
[L7] https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey
[L8] https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology
[L9] https://www.europeansocialsurvey.org/
[L10] https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm
[L11] https://europeanvaluesstudy.eu/

Explaining Any Time Series Classifier

While there is a growing interest, in defining eXplainable AI (XAI) methods to describe the behavior of black-box models used by AI decision systems, there is a surprising lack of research on black-box models for tasks regarding sequential data. Here, a method is assessed in order to provide an explanation that is easily understandable from a human standpoint.

Francesco Spinnato, Scuola Normale Superiore | francesco.spinnato@sns.it

Artificial Intelligence (AI) systems are increasingly outclassing human performances in many different fields and applications. Unfortunately, due to the usage of opaque Machine Learning (ML) algorithms, the logic behind their predictions often is difficult, if not impossible, to understand from a human standpoint and, for this reason, they are often called "black-box" models. As a consequence, in recent years, there has been an ever-growing interest, from a technical, social, and legal point of view, in defining eXplainable AI (XAI) methods to describe the behavior of black-box models used by AI decision systems.

Photo_Credit_Kyle_Sung_Unsplash

There are different approaches for explaining ML models depending on the input data, the black-box and the kind of explanation required. However, while interpretability is widely studied for tabular data [1], images [2] and only partially for texts [3], there is a surprising lack of research on black-box models for tasks regarding sequential data. Furthermore, the increasing availability of data stored in the form of time series such as electrocardiogram records, motion sensors data, climate measurements, stock indices, and so on, contributed to the diffusion of a wide range of time series classifiers employed in high-stakes decision making, where the explanation aspect becomes the crucial building brick for a trustworthy interaction between the human expert and the AI system.

For these reasons we tackled the problem of interpretability for opaque time series classifiers, i.e. black-box models that take as input a time series and predict its label, by proposing a Local Agnostic Subsequence-based Time Series explainer (LASTS)[7], whose objective is to return an explanation that is easily understandable from a human standpoint. The human mind usually tends to reason

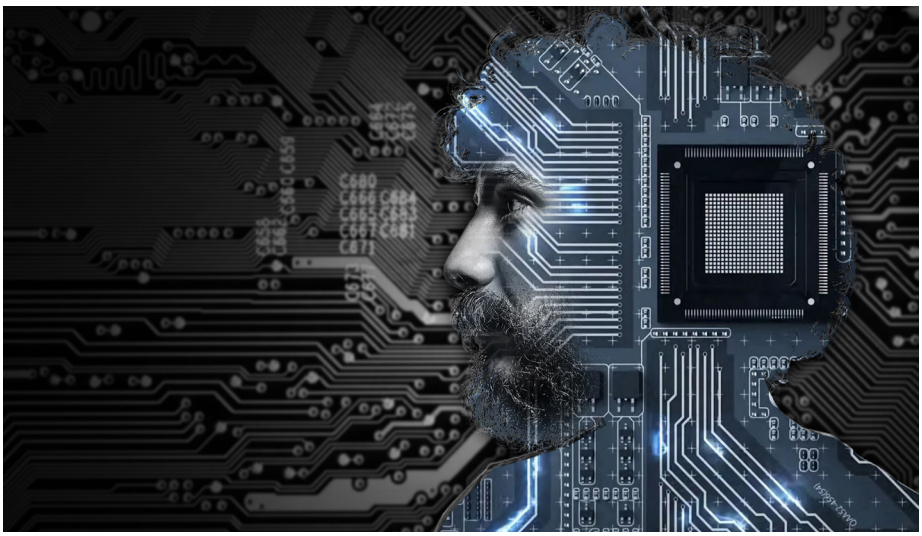


Image by Rox_buwa from Pixabay

in terms of counterfactuals, i.e. modifications in the input that change the prediction outcome [4]. In fact, while "direct" explanations such as decision rules are crucial for understanding the reasons for a certain outcome, a counterfactual reveals what should be different in the classified instance in order to have a different black-box outcome.

Our goal is to provide an explanation that offers the most possible complete way to understand the decision of a time series black-box classifier. Therefore, the output explanation of LASTS is composed both by a factual and a counterfactual subsequence-based rule that show the reasons for the classification in terms of logic conditions, indicating the subsequences of the time series that must, and must not, be contained in order to have a specific label returned by the black-box. In addition, the explanation contains also a set of exemplars and counterexemplars time series. Exemplars are instances classified with the same label as the time series being explained, i.e. prototypes highlighting similarities and common parts responsible for the classification. On the other hand, counterexemplars are instances similar to the one being explained but with a different label, and provide evidence of how the time series could be "morphed" to be classified with a different label, giving an intuitive idea of the decision boundary of the black-box, as pictured in [L1] for the dataset Cylinder-Bell-Funnel. [A morphing matrix for the dataset Cylinder-Bell-Funnel generated by modifying horizontally and vertically the two latent features obtained with the encoder compression. In green are depicted instances of the class bell, slowly morphing into instances having a different label, depicted in red].

LASTS uses an autoencoder, i.e. a neural network trained in order to compress and reconstruct instances as well as possible, to first encode a time series into a latent, simpler, representation. Then, in this latent space, through a genetic algorithm [5], it generates a neighborhood of new instances similar to the input instance and learns a local decision tree classifier to find factual and counterfactual rules. Latent instances respecting these rules are then decoded into exemplar and counterexemplar time series. These extracted time series are used to learn an interpretable shapelet-based decision tree that imitates the black-box, explaining its prediction in terms of subsequences that must, and must not, be contained in the time series. In [L2] we show an example of the explanation for an instance of the dataset Cylinder-Bell-Funnel. [An example of the explanation for an instance of the dataset Cylinder-Bell-Funnel. The instance x is correctly classified by the black-box as belonging to the class bell. In the second and third rows we can see, respectively, the exemplars (green) and counterexemplars (red) and the shapelet factual and counterfactual rules. The rules indicate that the black-box is "looking" at the slope of the central part of the time series].

We tested LASTS on four datasets and three black-boxes and showed that it effectively challenges existing state of the art explainers, like SHAP [6], in terms of fidelity and stability, providing also easily interpretable explanations.

Links:

[L1] https://lh6.googleusercontent.com/5bebUvTJSyPhQd1lXmjZjEtbPT1IUKwr-B2EAFVg6ULcQeyGLsMP7UGBwe_ArYFpFx-DeJRchR_cVQq3qs10wZZi8uAiruXHMiA1as-u4XuE0xSvPgIvoLm5FZGSBQpr7109Y7jhKfj

[L2] https://lh6.googleusercontent.com/Z8oKUcSUGb2HoIsp-9LztB5GiN-ayAmo-yNGIEJU90wq17zufOiC_b5fYyG1t4HA1U-fI5CIeT_tj3hGr-Xb2OerArSx-_L6fS-c1laF3ZMd--Os4syJ5x0IS-ix4IjLII0hT45EkgY

References:

[R1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models, 2018.

[R2] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 2019.

[R3] H. Liu, Q. Yin, and W. Y. Wang. Towards explainable NLP: A generative explanation framework for text classification. InProceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5570–5581, Florence, Italy, July 2019. Association for Computational Linguistics.

[R4] R. M. Byrne. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 6276–6282, 2019.

[R5] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems.CoRR,abs/1805.10820, 2018.

[R6] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[R7] R. Guidotti, Anna Monreale, F. Spinnato, D. Pedreschi and F. Giannotti, 2020. Explaining Any Time Series Classifier. IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), 2020.

About the role of social media platforms in the modern society

Two days after the riots at the US Capitol, on 8 January Twitter banned former US President Donald Trump from its platform with other online social networks soon following. This re-ignited the not-novel discussion about control and "super-power" of modern social media platforms, highlighting the tension between accountability and the freedom of expression

Francesca Pratesi, CNR | francesca.pratesi@isti.cnr.it

The first days of 2021 were characterized by very serious episodes that "struck the latest and perhaps most savage blow to America's reputation as a paragon of liberal democracy" according to David Smith on The Guardian [L1]. The riot at the United States Capitol [L2,3,4] has driven the public debate about justice on Online Social Network (OSN), after President Trump was subjected to ban by the major OSNs [L5] (and, even indirectly, by other social media apps being taken offline, such as Parler [L6]).

Social media Directives [L7,8] explained and justified their decision with the danger of a potential escalation of violence, since Trump's communications seemed to "contribute rather than diminish the risk of ongoing violence" [L9]. Many journalists and experts [L10,11,12] agreed with this resolution, maintaining that labelling posts and tweets is simply an inadequate solution to fight misinformation about the US election [L13], especially considering the huge number of insults that former US President Trump posted on social media [L14], and they argued that the terms of use of social media platforms are clear enough about violent speech [L15]. Some journalists pointed out that rules normally applied to common people were being enforced for one of the most powerful individuals in the world [L16,17]. Moreover, the same journalists highlighted the hypocrisy of some of their colleagues who are usually silent when real censoring acts are inflicted on users, for example to Middle Eastern bloggers or activists [L18].

However, the de-activation of Donald Trump's OSN accounts reignited the not-novel discussion about control and "super-power" of modern social media platforms. Indeed, fake news and disinformation are issues that have been amplified by the democratization of information, by the massive use of OSNs, by the information overload, and by the presence of bots and echo chambers [L19]. One of the key reasons that these issues are so difficult to untangle is that "social media is fundamentally different from traditional media" (in terms of bandwidth, oversight, and availability) and "so traditional approaches to regulation have largely fallen short" [L20].

These issues are mainly due to the possibility to remain anonymous on

Photo_Credit_Marco_Verch_Flickr_Creative_Commons_Attribution 2.0 Generic (CC BY 2.0)

Image by Thomas Ulrich from Pixabay

OSNs. In the "real" world, actions have different consequences with respect to digital life, and real life often has some form of self-regulatory system (a person who often lies, would soon be seen as unreliable). Thus, some experts in the ethics field [L21] advocated to lever individual responsibility for each action and opinion. This would be possible with the help of online social networks themselves, who are not necessarily acting as a "bad guy" who tries to break the rules. Some argue that it also seems "not completely fair that OSNs should be the only entities in charge of supervising users' content, and the only entities who have the responsibility to establish whether certain content is illegal". An independent authority could help, taking part in these disputes; thus, "outsourcing of legal decisions might be a solution" [L21]. Nevertheless, OSNs are owned by private companies and this allows them to autonomously decide whether to publish a content or not. Yet, it does not remove the responsibility or the role of the company itself in society. Recently, some steps have been made by private companies to enforce control and integrity of published content [R1].

Susan Ness, an American attorney and former commissioner of the Federal Communications Commission, states that "Platform regulation should focus on transparency, not content" [L22]. And, then, she continues arguing that "it is possible to tackle hate speech and disinformation without trampling on free expression, if the U.S. and Europe work together: by mandating transparency — with accountability — instead of regulating content. Require social media companies to provide greater transparency of their content moderation rules and procedures, including how their algorithms influence what users see, and enforce these disclosures through robust oversight. [L...] Internet platforms are a black box. Mandating transparency would increase public pressure on platforms to improve users' online experience. Researchers and regulators would gain access to essential information, resulting in better rules and oversight based on evidence, not assumptions. And online companies would be prodded to examine problems like algorithmic bias that they might prefer to ignore".

She is, of course, not the only one to advocate for "robust reforms in a slate of technology regulation areas including privacy, market competition, and algorithmic transparency" [L20], and even involved companies have started a process in this direction, acknowledging the "need for more perspective and accountability" [L24].

The tension between accountability and the freedom of expression and information clearly represents a dilemma. However, it is relevant to underline that law must usually balance between opposing rights. It appears needed for governments to face this topic, and including public opinion in the debate can be a path to follow, starting from here.

[Links]

[L1] https://www.theguardian.com/us-news/2021/jan/06/trump-mob-capitol-clash-police-washington

[L2] https://www.washingtonpost.com/technology/2021/01/09/trump-twitter-banned-apps/

[L3] https://eu.usatoday.com/story/tech/2021/01/08/twitter-permanently-bans-president-trump/6603578002/

[L4] https://www.ctpost.com/news/slideshow/Q-A-How-can-Twitter-ban-Trump-215406.php

[L5] https://www.independent.co.uk/life-style/gadgets-and-tech/trump-twitter-ban-facebook-b1785378.html

[L6] https://www.bbc.com/news/technology-55624630

[L7] https://www.facebook.com/zuck/posts/10112681480907401

[L8] https://blog.twitter.com/en_us/topics/company/2020/suspension.html

[L9] https://twitter.com/guyro/status/1346950532372393985

[L10] https://www.linkiesta.it/2021/01/trump-twitter-facebook-censura/

[L11] https://twitter.com/alexstamos/status/1346932573235077121

[L12] https://www.theverge.com/2021/1/6/22217894/deplatform-trump-twitter-ban-facebook-youtube-congress-capitol-riots and https://www.platformer.news/p/its-time-to-deplatform-trump

[L13] https://edition.cnn.com/2020/12/08/tech/facebook-twitter-election-labels-trump/index.html

[L14] https://www.nytimes.com/interactive/2021/01/19/upshot/trump-complete-insult-list.html?smid=tw-nytimes&smtyp=cur#

[L15] https://www.ilpost.it/carloblengino/2021/01/18/il-paradosso-dei-social-parte-i/

[L16] https://www.valigiablu.it/trump-social-media-regole-ban/

[L17] https://www.valigiablu.it/deplatforming-trump-facebook-twitter/

[L18] https://anchor.fm/valigiablu/episodes/Trump--i-social-media-e-lipocrisia-di-giornalisti-e-politici--Lintervista-di-Selvaggia-Lucarelli-su-Radio-Capital-ad-Arianna-Ciccone-eou3k4

[L19] https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/

[L20] https://hbr.org/2021/01/are-we-entering-a-new-era-of-social-media-regulation

[L21] https://kdd.isti.cnr.it/esme2019/

[L22] https://slate.com/technology/2020/12/platform-regulation-european-commission-transparency.amp

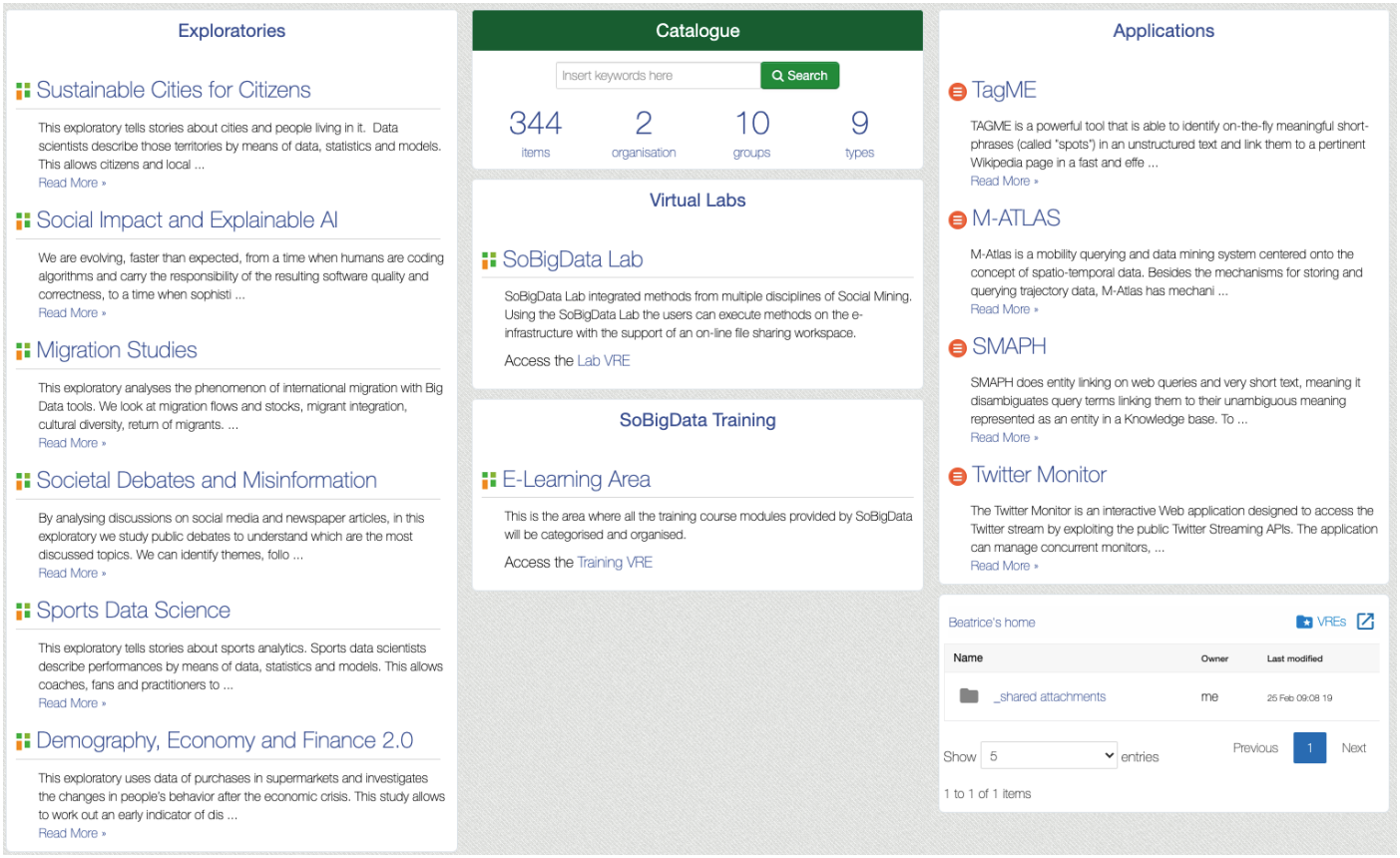
[L23] https://medium.com/whither-news/the-case-of-trump-v-facebook-1d82cc7dc193

[References]

[R1] Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, Ves Stoyanov. Preserving Integrity in Online Social Networks, arXiv:2009.10311v3 [cs.SI] (September 25, 2020)

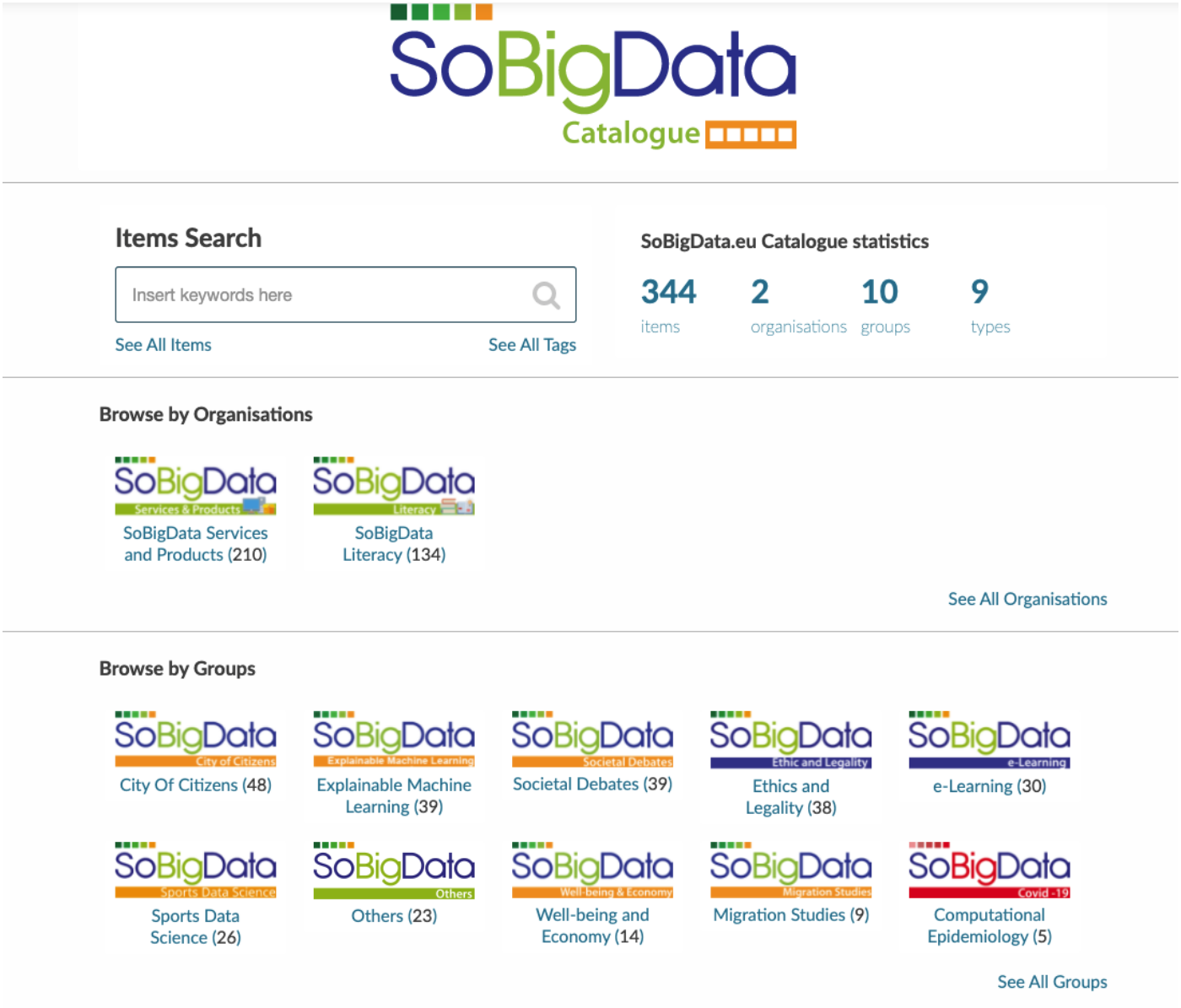
Join the SoBigData Community!

http://www.sobigdata.eu



Join the SoBigData Community!

http://www.sobigdata.eu



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