Breakthroughs on Cross-Cutting Data Management, Data Analytics, and Applied Data Science

Silvia Chiusano¹ · Tania Cerquitelli² · Robert Wrembel³ · Daniele Quercia^{4,5}

Accepted: 16 November 2020 / Published online: 28 November 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

1 Introduction - Emerging Data-Centered Ecosystems

In recent years, topics like (advanced) *data analytics* and *data science* have been constantly gaining popularity in research and business domains (Cerquitelli et al. 2020; Romero et al. 2020). *Data analytics* refers to the process of analyzing data by means of On-Line Analytical Processing techniques and Machine Learning (ML) algorithms, which are implemented in multiple procedural and declarative languages. *Data science* refers to techniques applied in the whole workflow (a.k.a. data processing pipeline) of data preparation for analysis and the analysis itself. The workflow typically includes the following tasks: data acquisition,

transformation, integration, cleaning, pre-processing, labeling (for ML), data analysis, and sophisticated visualizations, as reported in the recent Gartner report (Gartner 2020).

Presently, a huge and heterogeneous amount of data are continuously generated by humans and machines. These data are commonly referred to as big data (Ceravolo et al. 2018; Mauro et al. 2015). They are characterized mainly by: (1) the heterogeneity of data models and structures - from fully structured to unstructured, (2) a very high speed of creation, and (3) exceptionally large volumes.

Using modern data integration and storage architectures (including polystores and data lakes), based on Map-Reduce as a processing model, distributed file systems as storage, in-memory data processing engines, and NoSQL data model, big data can be efficiently collected, integrated, stored, managed, and analyzed for novel and more interesting data-driven applications.

The growing relevance of non-traditional domains such as bioinformatics (Liu et al. 2009), social networking (Zadeh et al. 2019), mobile computing (Deng et al. 2020), sensor applications (Strous and Cerf 2019), smart cities (Kar et al. 2019), and gaming (Clua et al. 2018) are generating increasing quantities of data that are complex in contents, heterogeneous in formats, and often order of terabytes in amount. These novel domains result in new ecosystems that include business areas such as: human resources, business processes, processes of data and information, IoT, mobile equipment, which definitely impact societies (Gupta et al. 2018).

Such abundance of data and technologies enables researchers to conduct significant research on data management and analytics to ultimately add artificial intelligence in real settings, i.e., *applied data science*. It also raises many new challenges for data management, storage, processing, and mining. Heterogeneous data collected from different sources should be adequately combined, integrated,

✉ Silvia Chiusano
silvia.chiusano@polito.it

Tania Cerquitelli
tania.cerquitelli@polito.it

Robert Wrembel
robert.wrembel@cs.put.poznan.pl

Daniele Quercia
daniele.quercia@kcl.ac.uk

¹ Interuniversity Department of Regional and Urban Studies and Planning, Politecnico di Torino, Turin, Italy

² Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy

³ Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan, Poland

⁴ King's College, London, UK

⁵ Nokia Bell Labs, London, UK

Particular attention should also be devoted to the involvement of the end-user in the entire process to enhance her/his exploitation and understanding of data and knowledge as well as to make the user a valuable data provider. Data management and analytics solutions will perform better when users can interact with systems and the systems take account of the cognitive and physiological characteristics of the people involved to provide personalized data processing and analytics. Examples of approaches tailored to the awareness of the people who form part of the processes are: interactive query refinement, data visualization, human-assisted data processing, and crowd-powered data infrastructure.

A wide range of modern scientific and more invasive applications could therefore benefit from those advances, covering all facets of the knowledge discovering process. Furthermore, the exploitation of these advanced technical solutions in real-life application domains could highlight new research issues to be addressed, thus opening interesting and novel research questions which need to be addressed. A variety of modern real-life settings along with academic settings could benefit from the dissemination of those advances and novel paradigms, covering all facets of the data management and analytics process. Industries and modern applications could explore the possibility of using these new research findings in real life settings, thus keeping pace with the latest technologies. Academics could identify open research issues coming from the industrial and real-life domains to continuously support the innovation process with the required methodological and technological solutions. The characteristics of big data and new needs to get insights from the data, induce new research and technological challenges.

This special issue includes 9 papers selected after a thorough reviewing process from 46 submitted to the call. The special issue received significant attention from both

The characteristics of big data and new needs from the business to get insights from the data, induce new research and technological challenges. The papers presented in this special issue cover a wide range of hot topics related to:

- The wordcloud in Fig. 1, generated through a traditional text-mining pipeline by analyzing titles, abstracts, and keywords of all accepted papers, provides an overview of the special issue content. By showing the top-60 frequent words characterizing the accepted papers, Fig. 1 highlights the macro-topic addressed (e.g., data, processing, model, learning, system) as well as it offers a detailed characteristic of all data management and data analytics

techniques (e.g., natural language, query, view management, machine learning) and the scenarios where they have been exploited (e.g., social network, text).

The content of the special issue is briefly introduced in the following subsections.

2.1 Efficient Data Models, Data Processing Pipelines and Architectures

One of the main challenges in managing big data is to build architectures for integrating big data sources as well as ingesting and transform big data into formats suitable for OLAP processing and machine learning algorithms. It is now widely agreed that the currently available architectures that allow to integrate big data, are mainly the data lake (Ceravolo et al. 2018; Nargesian et al. 2019) and polystore architectures (Alotaibi et al. 2020; Gadepally et al. 2016; Tan et al. 2017). Both of them apply physical and virtual integration. In both the architectures, metadata are crucial to build an interoperable data integration layer. Executing queries in data integration architectures is challenging for the following reasons. First, the information is constructed by crossing data from numerous sources, schemas, and formats. Second, a query in a global schema needs to be decomposed into queries on particular data sources. Third, such a query needs to be efficiently executed in a distributed and heterogeneous environments, whereby multiple users submit queries concurrently.

In this special issue, paper (Jovanovic et al. 2020) contributes an integration platform (called Quarry) for standard and big data sources. The platform supports a semi-automated construction of and end-to-end data processing pipeline. The platform adapts the concept of the mediated architecture where a global schema is mapped to local schemas by means of the Local-as-View (LAV) approach. Quarry heavily relies on metadata, which are organized as a hypergraph to facilitate searching and the integration of new metadata. Quarry has been evaluated in an international project for the World Health Organization.

To gain a more efficient resource utilization and better aggregate performance in shared environments, where queries are concurrently submitted by multiple users, Multi-Query Optimization (MQO) techniques are adopted in paper (Michiardi et al. 2020). The proposed system extends the SparkSQL Catalyst optimizer to provide a general approach to MQO for distributed computing frameworks that support a relational API. The system demonstrated significant improvements in terms of aggregate query execution times, while fulfilling the memory budget given to the MQO problem.

2.2 Predictive Analytics

Predictive analytics refers to a variety of statistical and analytical techniques used to develop models that predict future events or behaviors (Tsai et al. 2015). Classification models are common predictive models that aim at categorizing information based on historical data. Instead, regression analysis is used for forecasting, time series modelling and finding the causal effect relationship between a dependent (target) and independent variable (s) (predictor). Predictive analytics finds applications in a wider range of different domains. Although a variety of predictive algorithms and models are available in the literature, the application of these solutions in specific real domains always raises some challenging questions to be addressed related to different aspects, such as data distribution, data dimensionality, or availability of a sufficient amount of labeled data for training a model.

In this context, paper (Gardino et al. 2020) proposes a method for predicting product demand in the fashion industry. The proposed prediction method, called multi-VIEW Bridge Estimation (VIBE), takes advantage of the existence of multiple views on items, i.e., sets of homogeneous features. Such views compensate the problem of missing data by means of learning the interactions (or dependencies) between common latent features. The authors show how their method is able to reconstruct the missing views for products whose sales are not known, with the goal to forecast the future sales based on the reconstructed views.

Classification models are also used for the realization of sensor-based human activity recognition (HAR) to make a prediction of well-defined human activities out of the sensor data. However, for HAR a sufficiently large amount of annotated data is needed to realize an accurate classification model. Since the performance of these models degrades when they face new samples coming from unseen distributions, several domain adaptation methods have been proposed to minimize the need for labeled data. To address this issue, paper (Prabono et al. 2020) proposes a novel two-phase autoencoder-based approach on domain adaptation for sensor-based HAR. The proposed approach learns a latent representation, which minimizes the discrepancy between domains by reducing statistical distance. The effectiveness of the proposed approach has been tested in cross-domain sensor-based HAR.

2.3 Text Mining

As reported in Grohe (2019) and Marr (2019), a volume of unstructured data increases much faster than the volume of structured data. Among them, text data constitute

a substantial counterpart. For text analytics and Natural Language Processing (NLP), text data are processed by complex pipelines, which typically include noise removal, homogenizing lower-upper cases, resolving abbreviations, acronyms, and synonyms, stemming, lemmatization, stop-word removal, and (optionally) tagging. However, several challenging research issues are still open in text mining. Although the adaptation of NLP technologies brings attractive benefits to processing and understanding documents on a large-scale, the specific setting of each domain often prevents the adaptation of general-purpose NLP solutions or the adaptation of domain-specific NLP pipelines. Moreover, to test the quality of texts produced by such pipelines, one needs a ground truth data, i.e., benchmarks. Another issue to be solved is how to efficiently store and access text data.

Paper (Bordino et al. 2020) focuses on designing a novel NLP-based solution for garnishment - a specific use case in the legal domain so far overlooked in the NLP literature. The proposed GarNLP framework can automatically analyze and annotate various kinds of documents exchanged in a garnishment process, to classify documents into a predefined taxonomy of categories and automatically extract from the text, information on the garnishment procedure. This work was motivated by a request of one of a big pan-European commercial bank.

Paper (Truică et al. 2020) proposes a generic benchmark, called TextBenDS, for assessing the performance of text storage systems. At the conceptual level, the benchmark models text data (documents) as a cube. This model is then transformed by the benchmark into three different logical models to be implemented in distributed Hive, Spark, and MongoDB. TextBenDS includes a number of aggregation queries of different complexities, selectivities, and selection patterns to compute term weighting schemes applied to extracting top-k keywords and documents. TextBenDS has been applied in practice to evaluate the performance of Hive, Spark, and MongoDB on a text corpus including 2500000 tweets.

2.4 Social Media Mining

Social media includes a set of tools, such as blogs, social networking sites, and forums, which enable communication and cooperation. They facilitate relationship forming between users of diverse backgrounds, resulting in a rich social structure and a valuable source of information to capture user perception and to support decision making processes in a variety of scenarios (Kapoor et al. 2018). However, social media and their data represent complex ecosystems characterized by challenging issues from the data science perspective. Information is exchanged using heterogeneous data types (i.e., pictures, videos, text), thus requiring proper data processing techniques to turn out the

social data content into useful insights. Another remarkable aspect is the the risks associated with the use of social media, due to malicious behavior such as bots, sock puppets, creation and dissemination of fake news, Sybil attacks, and actors hiding behind multiple identities.

Social networks are analyzed as a valuable source of information to support business and management. One of the application is an intelligent transportation systems in large cities in paper (Vallejos et al. 2020). ML and NLP techniques are combined in a novel approach, named Manwë, for detecting, interpreting, geo-locating and disseminating traffic incidents written in natural language reported in social networks. Manwë is a complementary approach to other traffic-specific social networks such as Waze, acquired by Google in 2013.

Social media are exploited to dig deep the Internet of Things (IoT) scenario in paper (Ustek-Spilda et al. 2020). Nowadays sensing technologies allow collecting and transmitting data for algorithmic processing and ML in various domains. However, the potential consequences of pervasive connectivity along with concerns about security, privacy, and trust are abundant and are still widely debated. The paper proposes a method developed within the VIRT-EU project that integrates social media data and data from qualitative research and network analysis. The authors aim at: (1) analyzing the main actors in IoT-related discussions in Europe, and (2) identifying geographical clusters and specific topics that emerge within IoT, and the matters of concern discussed by the technology developers, designers and entrepreneurs of IoT on Twitter.

2.5 Privacy Issues in Distance Learning Solution

Distance learning solutions are becoming increasingly popular as they allow enriching traditional teaching and reaching students spread over vast geographical areas, but they also represent a valuable alternative when face-to-face teaching is not feasible (as for example due to critical events such as Codiv-19 pandemic). In such a systems, the adequate monitoring of student engagement requires a considerable amount of computational resources and network capacity to process audiovisual, interaction and physiological data of the audience in near-real time. In this context, critical privacy concerns arise since the continuous tracking and centralized analysis of sensitive personal data may invade the privacy of remote students.

To address the aforementioned concerns, paper (Preuveneers et al. 2020) proposes a solution based on well-known and proven enabling technologies for decentralized data analytics as well as best practices for preserving privacy. A multi-modal engagement model is proposed that runs in the cloud and at the edge to easily scale with a growing number of participants. Federated behavior data processing and the

application of secure multi-party computation are exploited to allow privacy enhanced analysis of student engagement.

3 Conclusions and open issues

Open science Data-driven methods require the data before making any conclusions. Unfortunately, only a few real, large enough, and fully reliable datasets are barely available. To significantly enhance and promote open science a privacy-preserving subset of data used in the research activities and the developed code must be released as open resources to allow researchers and practitioners to advance the research outcomes.

Enriching the Knowledge Discovery in Databases (KDD) Pipeline with Self-learning Capability Tailoring the KDD pipeline to a specific application setting always requires a lot of data science expertise and intense experimental activity. Cutting-edge techniques that can support the analyst in the automatic configuration of the KDD pipeline for the use case under analysis could significantly increase the spread of the data science in a wide range of real-life applications.

Privacy-Preserving Data Analytics Privacy issues have always been critical aspects to be considered in any ICT solutions. These aspects are of great importance in data-driven solutions, where therefore ad-hoc strategies have to been integrated to guarantee the right level of data protection. Preliminary solutions on federated ML solutions seem promising by offering privacy-preserving data strategies, besides having accurate performance, but many research issues are still open.

Personalized Data Visualization With the final aim to add intelligence to business applications, the value extracted from data must be provided to subjects with different levels of expertise through more adaptive and personalized strategies to support the decision-making process effectively. An exciting research challenge is to provide data analytics tools with personalized visualization strategies to deliver data and knowledge stories in an automated way based on user expertise.

Efficient Integration of Big Data A big data integration architecture has to handle new challenges as compared to a standard architecture. First, it processes much more complex data (e.g., graphs, natural language, non-tabular data in general). Second, data volumes are much larger. Third, big data change faster and arrive faster to the integration system (e.g., in the form of data streams). For these reasons, the development of a big data integration

architecture calls for resolving multiple research and technological issues. In our opinion, the most crucial ones include: (1) automatic data source discovery to be included in the integration architecture, (2) metadata management and standardization, (3) automatic integration of new data sources, (4) handling the evolution of data sources at the integration layer, (5) optimizing performance of data processing pipelines.

We hope that the readers will find the topics of the selected papers cutting-edge and exciting, and that they could lead to the development of a creative and technological-advanced range of future research activities to address some of the highlighted open issues.

Acknowledgements The Guest Editors thank all friends and colleagues who contributed to the success of the special issue. First of all, we appreciate the effort of all the authors who were attracted by the topics of the special issue and submitted a scientific contribution.

Special thanks go to the Editor-In-Chiefs Prof. Ram Ramesh and Prof. H. Raghav Rao, for their professional and precious support and to the Springer staff, in particular to Kristine Kay Canaleja and Aila O. Asejo-Nuique for their timely work.

We appreciated the excellent and timely work done by the reviewers who offered their expertise in assessing the quality of the submitted papers by providing constructive comments to the authors. The names of the reviewers for this special issue are available at <https://www.cs.put.poznan.pl/events/ISF2020-spec-issue.html>.

References

- Alotaibi, R., Cautis, B., Deutsch, A., Latrache, M., Manolescu, I., Yang, Y. (2020). ESTOCADA: towards scalable polystore systems. *Proceedings of VLDB Endowment*, 13(12), 2949–2952.
- Bordino, I., Ferretti, A., Gullo, F., Pascolutti, S. (2020). Garnlp: a natural language processing pipeline for garnishment documents. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09997-0>.
- Ceravolo, P., Azzini, A., Angelini, M., Catarci, T., Cudré-Mauroux, P., Damiani, E., Mazak, A., van Keulen, M., Jarrar, M., Santucci, G., Sattler, K., Scannapieco, M., Wimmer, M., Wrembel, R., Zaraket, F.A. (2018). Big data semantics. *Journal on Data Semantics*, 7(2), 65–85.
- Cerquitelli, T., Chiusano, S., Vargas-Solar, G. (2020). Editorial special issue on data analytics for engineering, science and society. *Computing*, 102(5), 1093–1096. <https://doi.org/10.1007/s00607-020-00810-z>.
- Clua, E., Roque, L., Lugmayr, A., Tuomi, P. (Eds.) (2018). *Int. Conf. on Entertainment Computing at the 24th IFIP World Computer Congress (WCC 2018)*, Vol. 11112. Springer Science: LNCS.
- Deng, S., Wu, H., Yin, J. (2020). *Mobile service computing* Vol. 58. Berlin: Springer.
- Gadepally, V., Chen, P., Duggan, J., Elmore, A.J., Haynes, B., Kepner, J., Madden, S., Mattson, T., Stonebraker, M. (2016). The bigdawg polystore system and architecture. In *IEEE high performance extreme computing conference*, pp. 1–6.
- Gardino, G.B., Meo, R., Craparotta, G. (2020). Multi-view latent learning applied to fashion industry. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-10005-8>.
- Gartner (2020). Magic quadrant for data science and machine learning platforms. Report.

- Grohe, K. (2019). Unstructured data growth in the enterprise. <https://www.stellus.com/blog/unstructured-data-growth-in-the-enterprise/>. Accessed Jul, 2020.
- Gupta, A., Deokar, A.V., Iyer, L.S., Sharda, R., Schrader, D. (2018). Big data & analytics for societal impact: recent research and trends. *Information Systems Frontiers*, 20(2), 185–194.
- Jovanovic, P., Nadal, S., Romero, O., Abelló, A. (2020). Quarry: a user-centered big data integration platform. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-10001-y>.
- Kapoor, K.K., Tamilmani, K., Rana, N.P., Patil, P., Dwivedi, Y.K., Nerur, S. (2018). Advances in social media research: past, present and future. *Information Systems Frontiers*, 20(0), 531–558.
- Kar, A.K., Ilavarasan, P.V., Gupta, M.P., Janssen, M., Kothari, R. (2019). Moving beyond smart cities: digital nations for social innovation & sustainability. *Information Systems Frontiers*, 21(3), 495–501.
- Liu, Y., Chan, L.W., Shyu, C., Liu, Y. (2009). Editorial for the special issue of knowledge discovery and management in biomedical information systems. *Information Systems Frontiers*, 11(4), 345–347.
- Marr, B. (2019). What is unstructured data and why is it so important to businesses? <https://www.forbes.com>. Accessed Jul, 2020.
- Mauro, A.D., Greco, M., Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *Proc. of int. Conf. on integrated information*, pp. 97–104.
- Michiardi, P., Carra, D., Migliorini, S. (2020). Cache-based multi-query optimization for data-intensive scalable computing frameworks. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09995-2>.
- Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q., Arocena, P.C. (2019). Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment*, 12(12), 1986–1989.
- Prabono, A.G., Yahya, B.N., Lee, S.L. (2020). Atypical sample regularizer autoencoder for cross-domain human activity recognition. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09992-5>.
- Preuveneers, D., Garofalo, G., Joosen, W. (2020). Cloud and edge based data analytics for privacy-preserving multi-modal engagement monitoring in the classroom. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09993-4>.
- Romero, O., Wrembel, R., Song, I. (2020). An alternative view on data processing pipelines from the DOLAP 2019 perspective. *Information Systems*, 92.
- Strous, L., & Cerf, V.G. (Eds.) (2019). *Internet of Things. Information Processing in an Increasingly Connected World - First IFIP International Cross-Domain Conference at the 24th IFIP World Computer Congress (WCC 2018), IFIP Advances in Information and Communication Technology*, Vol. 548. Berlin: Springer.
- Tan, R., Chirkova, R., Gadepally, V., Mattson, T.G. (2017). Enabling query processing across heterogeneous data models: a survey. In *Bigdata* (pp. 3211–3220).
- Truică, C.O., Apostol, E., Darmont, J., Assent, I. (2020). Textbends: a generic textual data benchmark for distributed systems. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09999-y>.
- Tsai, C., Lai, C., Chao, H., Vasilakos, A.V. (2015). Big data analytics: a survey. *Journal of Big Data*, 2(21). <https://doi.org/10.1186/s40537-015-0030-3>.
- Ustek-Spilda, F., Vega, D., Magnani, M., Rossi, L., Shklovski, I., Lehuède, S., Powell, A. (2020). A twitter-based study of the european internet of things. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-10008-5>.
- Vallejos, S., Alonso, D.G., Caimmi, B., Berdun, L., Armentano, M.G., Soria, Á. (2020). Mining social networks to detect traffic incidents. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-09994-3>.
- Zadeh, A.H., Zolbanin, H.M., Sharda, R., Delen, D. (2019). Social media for nowcasting flu activity: spatio-temporal big data analysis. *Information Systems Frontiers*, 21(4), 743–760.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Silvia Chiusano (PhD) is associate professor in computer engineering at the Politecnico di Torino, Italy. Her research interests focus on the design of innovative solutions for large-scale data management and mining with interest on health care, smart cities and healthy cities application domains. She served on the program committee of several international conferences and workshops on data mining research area (e.g., KDD, ICDM, ECML-PKDD, PAKDD). She is associate editor of “Computing” Journal (Springer) and member of the editorial board of “Internet of Things – Engineering Cyber Physical Human Systems” Journal (Elsevier). She was co-chair of the workshop track for the international conference ADBIS 2018. She organized, jointly with Tania Cerquitelli (Politecnico di Torino, Italy) and Genoveva Vargas Solar (CNRS, France), four editions of the DARLI-AP international workshop on data analytics for real-life applications co-located with the international conferences EDBT/ICDT (2020, 2019, 2018) and IEEE Smart data (2017). Silvia has been involved in many National and European research projects addressing research issues in machine learning and data analytics in different application domains such as citizen security and urban mobility.

Tania Cerquitelli has been an associate professor at the Department of Control and Computer Engineering of the Politecnico di Torino, Italy, since March 2018. Her research activities have been mainly devoted to fostering and sharing research and innovation on automated data science and machine learning in different real-life settings, explainable artificial intelligence solutions, and interpretable predictive and descriptive modeling techniques. Tania has been involved in many European and Italian research projects addressing different research issues related to machine learning and data analytics for energy-related data and Industry 4.0. Tania has been a member of the editorial board of the Future Generation Computer Systems journal (Elsevier) and the Computer Networks journal (Elsevier) since March 2020. She has been serving on the program committee of some major international conferences (e.g., KDD since 2018, ECML-PKDD since 2017, ICDM since 2020) on data mining and machine learning research area. She got the master degree with honours in Computer Engineering (in 2003) and the PhD degree (in 2007) from the Politecnico di Torino, Italy, and the master degree with honours in Computer Science (in 2003) from the Universidad De Las Américas Puebla, Mexico.

Robert Wrembel (PhD, Dr. Habil.) is an associate professor in the Faculty of Computing and Telecommunications, at Poznan University of Technology (PUT), in Poland. He specializes in databases and data warehouses; involved in 8 research and 8 industrial projects in the area of data processing technologies; visiting scholar in BI company Targit (USA) (2013); consultant at software house Rodan Systems (2002-2003); lecturer at Oracle Poland (1998-2005); PUT coordinator of Erasmus Mundus Joint Doctorate Information Technologies for Business Intelligence - Doctoral College (2013-2020); a regular editorial board member of Data & Knowledge Engineering. Has visited numerous research and education centers, including: Universitat Politècnica de Catalunya - BarcelonaTech (Catalunya), Université Lyon 2 (France), Université de Tours (France), Universidad de Costa Rica (Costa Rica), Klagenfurt University (Austria), University of Maribor (Slovenia), Loyola University (USA), INRIA Paris-Rocquencourt (France), and Université Paris Dauphine (France). In 2010 he received IBM Faculty Award for highly competitive research, in 2019 - IBM Shared University Research Award, and in 2019 - International Federation for Information Processing (IFIP) Service Award. He represents Poland in IFIP Technical Committee TC 2 - Software: Theory and Practice; he is a chair of IFIP Working Group 2.6 (Database).

Daniele Quercia (<https://www.researchswinger.org/>) is Professor of Urban Informatics at the Center of Urban Science and Progress (CUSP), King's College London. He is also Department Head at Nokia Bell Labs in Cambridge (UK). He is interested in the relationship between online and offline worlds and his work has been focusing in the areas of urban informatics. He was General Chair for AAAI ICWSM and Track Chair for ACM WWW. He has been co-editor of Computer Communications Journal Special Issue on Online Social Networks (Elsevier) 2014 and for the Special issue on Personality in Personalized systems (UMUAI) 2014. His research has been published in leading venues including ICSE, Ubicomp, ICDM, CSCW, RecSys, WSDM, and WWW, received a best paper award from ACM Ubicomp and from AAAI ICWSM, and an honorable mention from AAAI ICWSM, and has been featured on more than 90 international news outlets. He spoke at TED, wrote for BBC, and has been named one of Fortune magazine's 2014 Data All-Stars. He was a Postdoctoral Associate at the Massachusetts Institute of Technology where he worked on social networks in a city context, then a Horizon senior researcher at The Computer Laboratory of the University of Cambridge and, after that, a Research Scientist at Yahoo Labs in Barcelona. He received his PhD from UC London.