The Responsibility Challenge for Data

H. V. Jagadish
University of Michigan, USA
jag@umich.edu

Francesco Bonchi
ISI Foundation, Italy
francesco.bonchi@isi.it

Tina Eliassi-Rad
Northeastern University, USA
tina@eliassi.org

Lise Getoor
UC Santa Cruz, USA
getoor@soe.ucsc.edu

Krishna Gummadi
Max Planck Institute for Software Systems, Germany
gummadi@mpi-sws.org

Julia Stoyanovich
New York University, USA
stoyanovich@nyu.edu

ACM Reference Format:
H. V. Jagadish, Francesco Bonchi, Tina Eliassi-Rad, Lise Getoor, Krishna Gummadi, and Julia Stoyanovich. 2019. The Responsibility Challenge for Data. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3299869.3300079

As data science and artificial intelligence become ubiquitous, they have an increasing impact on society. While many of these impacts are beneficial, others may not be. So understanding and managing these impacts is required of every responsible data scientist. Nevertheless, most human decision-makers use algorithms for efficiency purposes and not to make a better (i.e., fairer) decisions. Even the task of risk assessment in the criminal justice system enables efficiency instead of (and often at the expense of) fairness. So we need to frame the problem with fairness, and other societal impacts, as primary objectives.

In this context, most attention has been paid to the machine learning of a model for a task, such as recognition, prediction, or classification. However, issues arise in all parts of the data eco-system, from data acquisition to data presentation. For example, the majority of the population is not white and male, yet this demographic is over-represented in the training data. It is challenging for a data scientist to satisfactorily discharge this broad responsibility.

One good way to think about these problems is to consider them in the context of societal-scale human-computer systems that facilitate interactions and knowledge exchange between individuals, organizations, and governments in our society. These systems involve networking of humans and computers, where human societal behaviors affect computations and vice-versa. We can conduct research to understand, predict, and control the behaviors of their constituent human users and computer systems.

How can machine learning and AI systems reason effectively about complex dependencies and uncertainty? Furthermore, how do we understand the ethical and societal issues involved in data-driven decision-making? There is a pressing need to integrate algorithmic and statistical principles, social science theories, and basic humanist concepts so that we can think critically and constructively about the socio-technical systems we are building. For example, what is a good null model for fairness? The current definitions (group vs. individual fairness) have not been informed by the vast literature on fairness from humanities and social sciences.

Furthermore, if we really want to have impact on society, we should use computer science to expose bad policy. For example, can we accurately reconstruct the intent of a policy such as the Stop-and-Frisk Program in NYC by analyzing its data? If so, then we should show the difference between the intent of the policy (as written in executive orders) and its implemented intent; and with public support, cause change.

Finally, we need to consider regulatory frameworks and their impact on the data science systems of interest to us. Governments are starting to recognize the need to regulate data-driven algorithmic technology. Three prominent examples are the European Union's General Data Protection Regulation (GDPR), the New York City Automated Decisions Systems (ADS) Law, and the Net Neutrality principle, that aim to protect the rights of individuals who are impacted by data collection and analysis.

In this panel, we will explore the responsibility challenge from multiple perspectives and identify research areas that sorely need the attention of the database research community.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3300079

The panel itself comprises five experts with diverse expertise, in addition to a moderator, as described below.

Francesco Bonchi is Deputy Director at the ISI Foundation, Turin, Italy, with responsibility over the Industrial Research area. At ISI Foundation, he is also Research Leader for the "Algorithmic Data Analytics" group. He is also (part-time) Research Director for Big Data Data Science at Eurecat (Technological Center of Catalonia), Barcelona. Before he was Director of Research at Yahoo Labs in Barcelona, Spain, where he was leading the Web Mining Research group.

His recent research interests include mining query-logs, social networks, and social media, as well as the privacy and fairness issues related to mining these kinds of sensible data. He has more than 200 publications in these areas. He also filed 15 US patents, and got granted 8 US patents.

Francesco is member of the Steering Committee of ECML PKDD and IEEE DSAA. He has served as program co-chair of the first and second ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD 2007 and 2008), the 1st IEEE International Workshop on Privacy Aspects of Data Mining (PADM 2006), and the 4th International Workshop on Knowledge Discovery in Inductive Databases (KDID 2005). He is co-editor of the book "Privacy-Aware Knowledge Discovery: Novel Applications and New Techniques". For more information visit http://www.francescobonchi.com/

Tina Eliassi-Rad is an Associate Professor of Computer Science at Northeastern University in Boston, MA. She is also a core faculty member at Northeastern University's Network Science Institute. Prior to joining Northeastern, Tina was an Associate Professor of Computer Science at Rutgers University; and before that she was a Member of Technical Staff and Principal Investigator at Lawrence Livermore National Laboratory. Tina earned her Ph.D. in Computer Sciences (with a minor in Mathematical Statistics) at the University of Wisconsin-Madison. Her research is rooted in data mining and machine learning; and spans theory, algorithms, and applications of big data from networked representations of physical and social phenomena. She has over 70 peer-reviewed publications (including a few best paper and best paper runner-up awardees); and has given over 180 invited talks and 13 tutorials. Tina's work has been applied to personalized search on the World-Wide Web, statistical indices of large-scale scientific simulation data, fraud detection, mobile ad targeting, cyber situational awareness, and ethics of machine learning. Her algorithms have been incorporated into systems used by the government and industry (e.g., IBM System G Graph Analytics) as well as open-source software (e.g., Stanford Network Analysis Project). In 2017, she served as the program co-chair for the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (a.k.a. KDD, which is the premier conference on data mining) and as the program co-chair for the International Conference on Network Science (a.k.a. NetSci, which is the premier conference on network science). In 2010, she received an Outstanding Mentor Award from the Office of Science at the US Department of Energy. For more details, visit http://eliassi.org.

Lise Getoor is a professor in the Computer Science Department at UC Santa Cruz and the director of the UC Santa Cruz D3 Data Science Center. Her research areas include machine learning and reasoning under uncertainty; in addition she works in data management, visual analytics and social network analysis. She has over 200 publications and extensive experience with machine learning and probabilistic modeling methods for graph and network data. She is a Fellow of the Association for Artificial Intelligence, an elected board member of the International Machine Learning Society, serves on the board of the Computing Research Association (CRA), has served as Machine Learning Journal Action Editor, Associate Editor for the ACM Transactions of Knowledge Discovery from Data, JAIR Associate Editor, and on the AAAI Council. She was co-chair for ICML 2011, and has served on the PC of many conferences including the senior PC of AAAI, ICML, KDD, UAI, WSDM and the PC of SIGMOD, VLDB, and WWW. She is a recipient of an NSF Career Award and eleven best paper and best student paper awards. In 2014, she was recognized as one of the top ten emerging researchers leaders in data mining and data science based on citation and impact according to KDD Nuggets. She is on the external advisory board the San Diego Super Computer Center, and the scientific advisory board for the Max Planck Institute for Software Systems, and has served on the advisory board for companies including Sentient Technologies. She received her PhD from Stanford University in 2001, her MS from UC Berkeley, and her BS from UC Santa Barbara, and was a professor at the University of Maryland, College Park from 2001-2013

Krishna Gummadi is a tenured faculty member and head of the Networked Systems research group at the Max Planck Institute for Software Systems (MPI-SWS) in Germany. He also holds an honorary professorship at the University of Saarland. He received his Ph.D. (2005) and B.Tech. (2000) degrees in Computer Science and Engineering from the University of Washington and the Indian Institute of Technology, Madras, respectively.

Krishna's research interests are in the measurement, analysis, design, and evaluation of complex Internet-scale systems. His current projects focus on understanding and building social computing systems. Specifically, they tackle the challenges associated with (i) assessing the credibility of information shared by anonymous online crowds, (ii) understanding

and controlling privacy risks for users sharing data on online forums, (iii) understanding, predicting and influencing human behaviors on social media sites (e.g., viral information diffusion), and (iv) enhancing fairness and transparency of machine (data-driven) decision making in social computing systems.

Krishna's work on online social networks, Internet access networks, and peer-to-peer systems has been widely cited and his papers have received numerous awards, including SIGCOMM Test of Time, IW3C2 WWW Best Paper Honorable Mention, and Best Papers at NIPS ML Law Symposium, ACM COSN, ACM/Usenix SOUPS, AAAI ICWSM, Usenix OSDI, ACM SIGCOMM IMC, ACM SIGCOMM CCR, and SPIE MMCN. He has also co-chaired AAAI's ICWSM 2016, IW3C2 WWW 2015, ACM COSN 2014, and ACM IMC 2013 conferences. He received an ERC Advanced Grant in 2017 to investigate "Foundations for Fair Social Computing".

Julia Stoyanovich is an Assistant Professor in the Department of Computer Science and Engineering at the Tandon School of Engineering, and the Center for Data Science. Julia's research focuses on responsible data management and analysis practices: on operationalizing fairness, diversity, transparency, and data protection in all stages of the data acquisition and processing lifecycle. She established the Data, Responsibly consortium (dataresponsibly.github.io), and serves on the New York City Automated Decision Systems Task Force (by appointment by Mayor de Blasio). Julia developed a technical course on Responsible Data Science, with all materials publicly available online (dataresponsibly.github.io/courses/spring19/).

In addition to data ethics, Julia works on management and analysis of preference data, and on querying large evolving graphs. She holds M.S. and Ph.D. degrees in Computer Science from Columbia University, and a B.S. in Computer Science and in Mathematics and Statistics from the University of Massachusetts at Amherst. Julia is a recipient of an NSF CAREER award.

H. V. Jagadish (Panel Moderator) is Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science at the University of Michigan in Ann Arbor, and Director of the Michigan Institute for Data Science. Prior to 1999, he was Head of the Database Research Department at ATT Labs, Florham Park, NJ.

Jagadish is well known for his broad-ranging research on information management, and has approximately 200 major papers and 37 patents. He is a fellow of the ACM, "The First Society in Computing," (since 2003) and of AAAS (since 2018). He served on the board of the Computing Research Association (2009-2018). He has been an Associate Editor for the ACM Transactions on Database Systems (1992-1995), Program Chair of the ACM SIGMOD annual conference (1996), Program Chair of the ISMB conference (2005), a trustee of the VLDB (Very Large DataBase) foundation (2004-2009), Founding Editor-in-Chief of the Proceedings of the VLDB Endowment (2008-2014), and Program Chair of the VLDB Conference (2014). Since 2016, he is Editor of the Morgan Claypool "Synthesis" Lecture Series on Data Management. Among his many awards, he won the ACM SIGMOD Contributions Award in 2013 and the David E Liddle Research Excellence Award (at the University of Michigan) in 2008. His popular MOOC on Data Science Ethics is available on both EdX (https://www.edx.org/course/data-science-ethics-0) and Coursera (https://www.coursera.org/learn/data-science-ethics).