REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
12-09-2017	Final Report	15-Sep-2016 - 14-Dec-2016

4. TITLE AND SUBTITLE
Final Report: MIT Institute for Data, Systems, and Society Launch Event

5a. CONTRACT NUMBER
W911NF-16-1-0578

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Massachusetts Institute of Technology (MIT
77 Massachusetts Avenue
NE18-901
Cambridge, MA 02139 -4307

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69869-NS-CF.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ali Jadbabaie-Moghadam
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-253-7339

RPPR Final Report
as of 18-Sep-2017

Agency Code:

Proposal Number: 69869NSCF **Agreement Number: W911NF-16-1-0578**
INVESTIGATOR(S):

Name: Ali Jadbabaie-Moghadam
Email: jadbabai@mit.edu
Phone Number: 6172537339
Principal: Y

Organization: **Massachusetts Institute of Technology (MIT)**
Address: 77 Massachusetts Avenue, Cambridge, MA 021394307
Country: USA
DUNS Number: 001425594 EIN: 042103594
Report Date: 14-Mar-2017 Date Received: 12-Sep-2017
Final Report for Period Beginning 15-Sep-2016 and Ending 14-Dec-2016
Title: MIT Institute for Data, Systems, and Society Launch Event
Begin Performance Period: 15-Sep-2016 **End Performance Period:** 14-Dec-2016
Report Term: 0-Other
Submitted By: Ali Jadbabaie-Moghadam Email: jadbabai@mit.edu
Phone: (617) 253-7339

STEM Degrees: **STEM Participants:**

Major Goals: Many of the most challenging problems that face the nation's military involve complex, interconnected systems and infrastructures formed by the interactions between social phenomena and large-scale engineering systems. Examples include autonomous systems, online social networks and their impact on resulting sociopolitical
phenomena, electronic voting systems, self driving cars and smart urban systems and services, as well as other complex social and engineered networks. Increasingly, one cannot understand the full picture of such complex, networked systems without understanding all of these components. What has made the above understanding possible is the availability of massive amounts of data on all relevant aspects. There
is an increasing push to to combine analytical tools that have been primarily, but not exclusively, in the purview of engineering with the study of human and social behavior. These aspects have been addressed traditionally in different disciplines.
As an example, consider the role of autonomy in urban systems. In addition to developing rigorous theory, modeling, and synthesis tools for smart buildings and autonomous, self-driving cars that are going to revolutionize urban systems and transportation, it is of paramount importance for researchers and policy
makers and practitioners to understand how the behavior of human users individually and in groups, will impact such a sociotechnical system. How should rules, and policies be designed to ensure efficient, robust, and safe operation of such systems?What are the social and economic implication of replacing humans with
autonomous systems? Even before answering such important questions, it is important to ask whether our current graduate educational systems designed within well-established departmental structures is adequate to train a new breed of scholars that can systematically address these questions.

To address these issues, The PI and colleagues have found the new MIT Institute for Data, Systems, and Society (IDSS), an interdisciplinary entity that is centered at the School of Engineering but spans all five Schools (Schools of Engineering, Sciences, Humanities and Social Sciences, Management, and Architecture).
The Institute, which operates like a graduate department with existing faculty and 16-16 new faculty lines, will address problems at the interface of complex engineering systems and social systems in a holistic manner, by integrating the engineering perspective, modern analytical tools (including advanced statistical methods and information and decision sciences), and the relevant social sciences, including economics,
political science, sociology, and anthropology. The scope of the Institute's research will be broad, encompassing diverse problems of significant societal importance such as systemic risk in financial and power networks, digitally driven socio-political change, social phenomena of cascades and contagion, and future social cyber-physical systems, just to name a few. The Institute will also have a new PhD program in Social and Engineering Systems, that will bridge the gap discussed above, produce innovative research of a different kind, train a next generation of

leaders, and provide a paradigm that will hopefully be emulated by other institutions.
To formally launch IDSS, we proposed a 2-day workshop scheduled for September 22 and 23, 2016 on MIT campus. An early draft of the workshop agenda is available at http://idss2016.mit.edu. The workshop consisted of luminaries from academia, industry and government, and will be open to the public with free registration.

The workshop was a 2-day event held at MIT Media Lab Conference Room on
September 22 and 23, 2016. The workshop was webcast live and videos of all sessions
are stored online and available for public viewing at https://idss2016.mit.edu/videos/.

Accomplishments: The workshop was organized along 9 scheduled over 2 days on September 22-23 on MIT campus at MIT Media Lab. Each session will have speakers and panelists with time for questions and discussion. Each
day was divided into morning and afternoon sessions with coffee and lunch breaks in between. Extreme care was been taken to create a balance between various societal domains and fundamental methodologies,as well as balance in terms of gender and diversity of speakers and panelists. All panelists and speakers were at the forefront of this emerging discipline and its related flagship domains. there will be a student session
showcasing the works of students. We recorded the workshop and had scribes taking notes
during the event in order to create a report which will be attached.
Schedule:

September 22 Morning Sessions:
1. Session 1: The future of voting
2. Session 2: Data-driven policy
3. Session 3: Risk in financial systems

September 22 Afternoon Sessions:
1. Session 4: Social phenomena and social networks
2. Session 5: The future electric grid
3. Session 6: Student session

September 23 Morning Sessions:
1. Session 7: Analyzing our health
2. Session 8: Driving smart cities forward

September 23 Afternoon Session
1. Session 9: From applications to theory

The workshop will open at 8:30 with registration and breakfast. First session will start at 9:00am with short remarks by MIT president and the director of the Institute. Each session will be roughly 60-90 minutes with a keynote speaker and then a discussion by panelists with audience engagement at the end.

Training Opportunities: The workshop was free for all participants. There was a special session for PhD students.

Results Dissemination: The workshop was broadcast live and all videos are available online at

https://idss2016.mit.edu/videos/

Honors and Awards: One of the keynote speakers won the Economics I prize a week after the workshop

Protocol Activity Status:

Technology Transfer: Workshop was attended by many people from government.

RPPR Final Report
as of 18-Sep-2017

IDSS Launch Event

September 22-23, 2016

Minutes of the Meeting

Workshop Location and Dates

The workshop was a 2-day event held at MIT Media Lab Conference Room on September 22 and 23, 2016. The workshop was webcast live and videos of all sessions are stored online and available for public viewing at https://idss2016.mit.edu/videos/.

Workshop Agenda

Each day was divided into morning and afternoon sessions.
Below is the full agenda of the 2-day event.

Thursday, September 22

Opening Remarks, 9:00-9:30am
MIT President L. Rafael Reif
Professor Munther Dahleh, Director of the Institute for Data, Systems, and Society

Morning Sessions:

Session 1: The Future of Voting, 9:30am-10:30am
The role of technology in voting has gained increasing prominence over the past decade, creating interdisciplinary collaborations between political, computer, and data scientists. Voting data contains an abundance of information that goes beyond the actual vote. This session will look at the complexity of voting, the usability of computing technologies (such as cryptography) in designing future voting systems, and how data is playing a role in understanding and predicting voting patterns and the outcome of elections.

- Moderator: Professor Charles Stewart, MIT
- Keynote: Mr. Nate Silver, fivethirtyeight.com
- Professor Michael Alvarez, Caltech
- Ms. Kassia DeVorsey, Chief Analytics Officer, Messina Group Analytics

Session 2: Data-Driven Policy, 10:30am-11:15am
While communities are collecting more data than ever before to measure effects of public policy, such data sets tend to be quite small. With the absence of a control group, the assessment of existing policies and the design of new ones utilizing such data bring new challenges to statistics and data science. This panel will explore such challenges and will highlight how data analysis has been quite effective in some applications.

- Moderator: Professor Alberto Abadie, MIT
- Keynote: Professor Enrico Giovannini, University of Rome Tor Vergata

Break 11:15am-11:45am

Session 3: Risk in Financial Systems, 11:45am-12:30pm
Recent research has been successful in deriving abstracted models of the interconnected financial systems that quantify systemic risk and address cascaded failures of such systems. However, combining such models with recorded data for the purpose of monitoring and mitigation continues to be a major research and practical challenge. This session will discuss such challenges, as well as the progress that has been made.

- Moderator: Professor Asu Ozdaglar, MIT
- Keynote: Professor Bengt Holmstrom, MIT

Lunch 12:30pm-1:30pm

Afternoon Sessions:

Remarks by Professor Ian A. Waitz, Dean of the School of Engineering, MIT, 1:30pm
Session 4: Social Networks, 1:45pm-2:50pm
Social networks through social media have brought to bear very large data representing people's preferences and opinions, and have highlighted effective incentive mechanisms. Such networks also impact and inform a variety of complex systems in our society. Such data has brought in new security and privacy challenges that have occupied much of the research in data science. This panel will look at new opportunities for understanding social networks and human behavior, as well as technological methods for ensuring security and privacy.

- Moderator: Professor Ali Jadbabaie, MIT
- Professor Jon Kleinberg, Cornell University
- Professor Matthew Jackson, Stanford University
- Dr. Jeannette M. Wing, Microsoft Research
- Dr. Cynthia Dwork, Microsoft Research

Session 5: Future Electric Grid, 3:00pm-4:00pm
The electric grid presents some of the most challenging engineering, social, and economic challenges of the future. With increased demands on electricity and increased penetration of renewable sources, the need for new innovations in dynamic demand response, spot markets, and distributed control is rapidly increasing. This session will discuss some of these challenges and current work.

- Moderator: Professor Bob Armstrong, MIT
- Professor William Hogan, Harvard University
- Professor Michael Greenstone, University of Chicago
- Professor Sally Benson, Stanford University
- Professor Steven Low, Caltech

Break 4:00pm-4:30pm
Session 6: Student Session, 4:30pm-5:15pm
Student Session Chair: Professor Sandy Pentland, MIT
Posters and Networking at the Media Lab
5:15pm-6:00pm
6:00pm-8:30pm Special Invite only Dinner
Introductory Remarks: Professor Martin Schmidt, MIT
Speaker: Professor Daron Acemoglu, MIT

Friday, September 23

Start Time: 9:00am
End Time: 2:45pm

Morning Sessions:

Remarks by Professor Melissa Nobles, Dean of the School of Humanities, Arts, and Social Sciences, MIT, 9:00am

Session 7: Analyzing our Health, 9:15am-10:30am

The collection, aggregation, and analysis of medical data presents possibilities for future healthcare developments, including opportunities for personalized medicine and patient care. The use of big data in medicine also raises serious questions about patient privacy. This session will discuss ways in which the practice of medicine is being transformed by data.

- Moderator: Professor Peter Szolovits, MIT
- Keynote: Dr. DJ Patil, U.S. Office of Science and Technology Policy
- Dr. John Halamka, MD,Chief Information Officer, Beth Israel Deaconess Medical Center
- Professor Deborah Estrin, Cornell Tech
- Dr. Elazer Edelman, MD, Brigham & Women's Hospital & Professor of Medicine at Harvard Medical School of Medicine (HMS) and MIT Health Sciences and Technology Program (HST).

Break 10:30am-11:00am

Session 8: Driving Smart Cities forward, 11:00am-12:25pm

Cities will become increasingly interconnected through an ever-expanding "internet of things," allowing governments, urban planners and engineers access to massive amounts of data about urban life. This data is being used to design, plan, and structure cities in the United States and around the world. This session seeks to explore the many facets of smart-cities research, design, planning, and transportation.

- Moderator: Professor Sarah Williams, MIT
- Keynote: Dr. Steven Koonin, NYU
- Professor Rob Kitchin, Maynooth University
- Professor Balaji Prabhakar, Stanford University
- Professor Susan Crawford, Harvard Law School
- Professor Alexandre Bayen, UC Berkeley

Lunch 12:30pm-1:30pm

Special Student Q&A Session for students and speakers.

Afternoon Sessions:

Session 9: From Applications To Theory, 1:30pm-2:30pm

While applications have their own nuances, there are overarching challenges that need to be identified and addressed. These include, among others, fundamental questions in prediction, robustness/risk, computation, system architecture, and privacy. This session will address some of the emerging challenges in these foundational fields in this new era of large data and complex systems.

- Moderator: Caroline Uhler, MIT
- Professor Allen Tannenbaum, Stony Brook University
- Professor Elchanan Mossel, MIT
- Professor David Tse, Stanford University
- Professor Vincent Blondel, Rector, Université catholique de Louvain

Closing Remarks

2:45pm End

Session 1: The Future of Voting:
Author: Allie Fero

The IDSS Launch Event kicked off with the session, "The Future of Voting," which discussed the intersection between data, technology, and societal interaction on issues related to voting—including polling, campaigning, and academic study. The use of large-scale data acquisition and analysis was a common thread amongst the three speakers. For example, every speaker mentioned an abundance of public data in the voting sector, especially because of the public "voter file," which profiles registered voters' demographics and past voting habits. On top of this public data, different interest groups can accumulate private, proprietary data for themselves. This abundance of information enables the analysis of voters and trends—underscoring the state of voting.

Professor Charles Stewart III (MIT) introduced the panel. He noted that the first panel's focus on voting and elections was an ideal starting point for an event launching the Institute for Data, Systems, and Society, since elections produce large quantities of data, which are consumed and processed through a variety of systems (both public and private). Data are not only about election returns, but also the log files of the machines we vote on, ballot image files, voter registration files, proprietary files kept by candidates, records of campaign contributions, individual response data to public opinion polls—and much more. All this data links together systems of candidates, activists, journalists, and pundits—each of whom has a different interest in the data. Stewart emphasized the importance of looking at who controls the data, and how they are used.

Keynote speaker Nate Silver (FiveThirtyEight) began the session with a discussion of polling and associated successes and challenges. The Clinton-Trump electoral race was used to highlight the nuances of polling. Silver pointed out that several competing statements could be made with accuracy: that the polls are close, that Clinton is up, that Trump can win, and that Clinton probably will win.

The data that FiveThirtyEight uses to produce its poll results is based on public information. The goal is to provide a macro level view from an outsider perspective, without relying on potentially biased campaign information.

However, polling is not done in a vacuum, and many factors besides direct voter responses affect the output. For example, economic conditions in general can be used as factors in the model based on historical trends of how voters respond to varied economic climates. Moreover, the existence of polls themselves affect poll results. There is a feedback loop amongst pollsters. When pollsters agree, there is greater potential for replicating each other's mistakes, and there is a fear of publishing results that are contradictory to the majority results. Finally, there is also a feedback loop between the voters and the pollsters, as various voters can change their attitudes towards a candidate based off of the polling results.

Professor Michael Alvarez discussed the role of data and modeling from the academic perspective, as well as the editor's role in data transparency. Advances in generation and computation of data have allowed for great strides in our ability to acquire and analyze data, but transparency is at issue. Professor Alvarez discussed his concern with the current lack of publically available data repositories that would allow for public consumption of data for both novel research and result replication.

In academia, broader access to data can be made possible by requiring that authors submit their data upon publication in academic journals. This promotes transparency and access to information. An example of this in action is the Dataverse project, an open source research data repository. However, not all data can be made public; thus requiring open access can limit the publication of good research. The tradeoffs between access, transparency, and privacy are fundamentally at issue when sharing results of data driven research.

Kassia DeVorsey addressed the movement towards analytical techniques for campaign strategies. Campaigns have three outcomes that they work towards: registering voters, persuading voters to their side, and reminding supporters to get to the polls. Campaigns strategize how to allocate resources amongst these three goals. Increasingly, data analytics are being used to determine how to allocate campaign resources. Using the voter file and other data sources, as well as data collected through directly contacting voters and identifying their characteristics, campaigns generate predictive models to drive resource allocation decisions. For example, on any given street, only a few houses will be targeted by door-to-door campaigners.

The use of data in campaigning highlights many of the key issues surrounding data, systems, and society. First, we have the issue of privacy. The existence of the voter file can be controversial, as some may argue that voter history and demographics should be considered personal information. Second is the issue of targeting. By profiling voters, or consumers more generally, campaigns and industries are able to specifically target individualized messages at different groups of people. Beyond privacy, this can be controversial, especially in campaigns, because it enables the campaign to present different messages to different parts of the population. Finally, by highlighting only a group of people that could affect elections, there are entire groups of the population that are left out of the dialogue entirely. In the case of campaigns, only "swing" voters get the majority of time and campaign resources.

In short, data is being used to assess and affect campaigns, elections, and voting. Modern abilities to handle and manage data are already being used to efficiently influence and promote policies. However, there are a series of concerns with privacy and feedback that remain unaddressed, as modernization in technology has outpaced the discussion. Campaigns are purposely manipulating the opinion of targeted populations, but even "neutral" uses such as polling can have feedback cycles that affect the outcome of elections. And additional pressures affect the pollsters, themselves who may be averse to publishing outlier results.

Session 2: Data-Driven Policy:
Author: Zach Needell

In his introduction to the IDSS Launch Event session titled, "Data Driven Policy," MIT professor Alberto Abadie described the Maeslant Barrier in the Netherlands. A Dutch infrastructure project built in the 1990s, the Maeslant Barrier is a massive, two-pronged storm surge barrier designed to protect the port city of Rotterdam from flooding during extreme storms. Shutting the two gates is a slow, laborious process that disrupts the normal commerce of one of the busiest ports in the world and therefore incurs significant costs. The flooding of the port of Rotterdam by an unusually powerful storm would, however, be many times more damaging. The decision of whether or not to deploy the barrier must weigh the guaranteed costs of temporarily shutting down the port against the speculative costs of an incoming storm surge flooding the city of Rotterdam.

This type of decision, politically charged and under great uncertainty, where costs and benefits must be weighed in an unbiased manner, is in theory the ideal case for an algorithmic solution. The logic behind the decision to open or close the Maeslant Barrier operates completely insulated from human intervention, and the parameters of the algorithm are worked out through the political system in advance. It is with this example that Abadie laid out the central question of the session: In a world where data and computational power are becoming increasingly available, what opportunities and risks are present as we incorporate data into political decision-making processes? Will further incorporation of data and measurement into the process of governance necessarily lead to better outcomes, or do the potential costs to political engagement and democracy outweigh the potential benefits?

In his opening remarks, Abadie categorized the increasing volume and diversity of data available to decision-makers. In the past, quantitative analysis of policy relied primarily on data collected for the expressed purpose of that analysis—surveys, censuses, and other direct measurements—for decisions of resource allocation and planning. This type of data has the benefit of being purposefully collected to answer a pre-defined set of questions, but it is expensive and slow to collect, limiting its use to a basic set of questions. As processing power and record keeping improved, researchers and policymakers gained additional access to administrative data—tax records, inspection results, and transit fare payment information, for example—produced in the everyday process of providing government services. While these data are typically not collected for the expressed purpose of policymaking, they have the benefit of consistent format and broad scope, allowing practitioners to better evaluate the impacts of policy proposals and quantitatively argue for their implementation or rejection. In recent years, however, access to "big data" offers the opportunity to incorporate by orders of magnitude more records of data into the decision-making process. This type of "big data," captured by cell phone sensors, public and private vehicles, and an increasing number of Internet-connected devices, allows, in theory, the increased automization and quantification of decision-making, a process of which the Maeslant Barrier is a simple example, and one that provides great opportunities and risks.

The keynote speaker, Professor Enrico Giovannini of the University of Rome Tor Vergata, echoed the importance of data and statistics in government. Indeed, Professor Giovannini suggests that data-driven policy decisions are crucial to building fair and just political processes, quoting from the UN data revolution report: "Data are the lifeblood of decision-making and the raw material for accountability." He argued that, in an abstract sense, the job of governments is to make decisions and set policies, under uncertainty, with the goal of increasing the general welfare. Increasing amounts of data available to these policymakers should, almost by definition, allow for greater effectiveness of these decisions. The increasing availability of data, at least in the abstract, presents a great opportunity for society–for people inside the government to better tailor their policies towards addressing the greater good, and for those outside the government to make sure that policymakers and agencies remain committed to that goal.

The picture, though, is never quite that simple. Professor Giovannini posed a related paradox—in a world where data is increasingly available and political decisions can be made with more and more empirical grounding, why does public trust in science and data seem to be decreasing? He quotes a *New York Times* op-ed by William Davies: "It is possible to live in a world of data, but no facts." Different political parties disagree not just on the interpretation of data, but on basic underlying truths. While we are becoming continuously better at predicting and measuring the state of the world and the impact of government decisions, public trust in the process has not kept up.

Professor Giovannini suggests that, in part, this growing disjunction between data and public discourse is because governments and researchers have not done a sufficiently good job of quantifying and arguing for the value provided by data and statistics. While, in theory, better information should lead to better decisions, the practical effects of these improvements remain abstract. Additionally, Giovannini argues that not enough thought has gone into the measurement of success. Far beyond gross national product, Giovannini suggests that greater availability of data should allow us to better quantify the objectives, such as health, happiness, and freedom, summarized by the UN's Sustainable Development Goals—allowing people and their values to be involved in the democratic process.

As such, Giovannini argues that policymaking, and civil society in general, is not just a measurement and optimization problem. Any successful integration of more data into government must recognize that government is a problem of society and culture, as well. While the ability of data to improve prediction and measurement is important, there is something fundamentally undemocratic about placing all political decision-making in the hands of a data-driven algorithm. The goal of researchers and public servants interested in making better policy should be not to remove humans from the decision loop entirely, but to better clarify the moral choices being made.

Session 3: Risk in Financial Systems:
Authors: Ali Makhdoumi and Denzican Vanli

The IDSS Launch Event session "Risk in Financial Systems" was moderated by Professor Asu Ozdaglar and featured keynote speaker Bengt Holstrom, Paul A. Samuelson Professor of Economics at MIT.

Ozdaglar offered some opening remarks on systemic risk and financial systems. Financial markets manage price and redistribute risks, but they also create other risks, in particular, risks associated with financial crisis as well as system-wide meltdowns in the process (such as the 2008 financial crisis). She presented three broad perspectives on systemic risk. The first (common in the media and in movies such as *The Big Short*), emphasizes the mistakes and shortcomings on the part of the participants in the financial crisis. In this view, the markets are inefficient and even dangerous. Ozdaglar said this doesn't provide for a satisfactory framework for understanding why several aspects of financial markets work well.

A second perspective focuses on how shocks to a few financial institutions can, under certain circumstances, create "domino effects" of systemic risk. The emphasis here is on interconnections. The creation of these large "macroshocks" from the "microshocks" actually leads to some interesting non-linearities and some surprising effects. When you have larger shocks, it's harder to contain single shocks, and all will spread. A complete network, with the most connected pathways, can lead to the most distress in the entire system. A weakly connected network is actually the most robust.

The third perspective, which Ozdaglar said would be addressed in detail by Bengt Holstrom, starts from emphasizing the complex problems markets are trying to solve, and also highlights how these efforts actually lead to some peculiar, distinct features that we can observe in financial markets. The huge volume of transactions and the speed of them, create a strong incentive for the participants to work within a certain "opacity"—which can then lead to series of problems.

Holstrom spoke on "The Purpose and Peril of Money Markets," and remarked that the talk could have the subtitle "When Ignorance Is Bliss." He explained the basics of what money markets are, and why they are used.

He stated that although there have been a lot of efforts made in understanding the 2008 financial crisis, there is still limited consensus on what caused the crisis. Many blame Wall Street greed and wrong incentives, the ratings agencies that appeared seriously off the mark, and the government for subsidizing subprime lending. It is particularly puzzling that Wall Street traded in securities that it knew little about. Many believe the purpose of opaque securities was to deceive investors—triggering a universal call for transparency. However, the characteristic of "no questions asked" is the hallmark of money-market liquidity. This is actually the way money markets are supposed to look when they are functioning well. Holstrom remarked that in this sense, "ignorance is *almost* bliss"—with the mention of the "almost," because, ultimately, "someone has to know something."

Money markets are "high-velocity" and trade in debt claims that are backed, explicitly or implicitly, by collateral. The purpose of money markets is to provide liquidity. People

often assume that liquidity requires transparency, but this is a misunderstanding. What is required for liquidity is symmetric information about the payoff of the security that is being traded, so that adverse selection does not impair the market. Without symmetric information, adverse selection may prevent trade from taking place or in other ways impair the market. Trading in debt that is sufficiently over-collateralized is a cheap way to avoid adverse selection. When both parties know that there is enough collateral, more precise private information about the collateral becomes irrelevant and will not impair liquidity. Unlike money markets, where the desire to circumvent price discovery is a natural consequence of lending, the importance of price discovery in stock markets goes hand-in-hand with the traders' incentives to acquire information about the value of a firm. Holstrom also presented the "dark side" of opacity, explaining that it can also hide systemic risk—as in the case of 2008 financial crisis.

Session 4: Social Networks:
Authors: Amin Rahimian and Samuel Chevalier

The session started by Professor Jadbabaie introducing the four speakers: John Kleinberg, Matthew Jackson, Jeannette M. Wing, and Cynthia Dwork. Following the short introductions, each speaker gave a ten-minute presentation of their major research findings in the past few years as well as their visions for the future.

Professor Kleinberg's talk emphasized the emergence of online social networks in the past decade, and how the prospects of big data computations on social data have revolutionized the perspective of social sciences. He highlighted the role of social networks as a transport mechanism for information, behavior, and opinions—and discussed how certain decisions, such as the probability of joining a group, can be predicted and analyzed as a function of the social connections and neighboring behaviors. He discussed the issue of weak and strong social ties: how they could be inferred from the social data and how they can be used to study the flow of information in an organization—for example in times of a shock, such as price changes in stocks. In particular, he pointed out empirical results that indicate that people in an organization turn back to their strong ties (as opposed to weak ones) in times of a major shock.

Kleinberg shared an example of how digital text can be used to determine how language and topic matter is a function of workplace stress. Email text correspondence between employees at a hedge fund company was analyzed against market price changes. These price changes correspond to periods of high, medium, and low stress. Potentially, this is an example where causal inference is very high since there are fewer variables at play: workplace focus and mood can be directly related to price changes.

Professor Jackson re-emphasized the idea that social data that has become available through emerging social media platforms plays a major role in many areas of social science, such as anthropology, sociology, developmental economics, etc. Data collection methods prior to the digital age were time-consuming and laborious—and rarely captured the dynamic of a large, complicated, and interconnected system. Today, we have the ability to visualize the dynamics of a full society. Even more valuable, we can access direct textual data as opposed to merely observed incidental data from poles or

questionnaires. Modern data collection is easy, we have better resolution, and we can trace how actions are correlated with a range of parameters.

Jackson then presented some of the empirical data from the fieldwork that his research group has done in villages in India, where they studied changes in the structure of a social network for kerosene and rice exchanges. They studied two villages: one with and one without a micro-finance scheme providing small loans to individuals. Accordingly, while the social structures in both villages showed a certain amount of social network erosion over time (possibly due to urbanization), the degree of social network erosion was far greater in the village where micro-loans were available. Notably, he observed that those individuals least likely to receive loans were more likely to lose access to the social structure than those who received loans. Jackson noted that those most likely to obtain loans were typically better educated, wealthier (relative to peers) and could have other advantages. Jackson remarked in general, that the way markets are evolving is having a significant impact on social structures, and that this is an area in need of further research.

Professor Jackson concluded his talk by pointing out some of the challenges and dangers of working in today's data-rich social platforms. In particular, the researchers should overcome the tendency to look for interesting patterns in data, rather than testing theories that explain them. Moreover, researchers may face challenges in accessing data from industry. Also, designers and policy makers should decide on what data are made accessible to researchers and the mechanism for data collection and publication.

Professor Wing opened her talk by portraying a path toward a theory of trust in networks of humans and computers. She emphasized the distinctive approaches of computer science and behavioral sciences, as well as economics, toward this issue—leading to notions of "computational trust" and "behavioral trust." While computational trust relies on techniques such as data encryption, proof correctness, redundancy, and fault resiliency, behavioral trust relies on concepts such as individual beliefs in trustworthiness, reputations, etc. She demonstrated some of these concepts in a game-theoretic framework and concluded her talk by highlighting how the neurologically implicated desire to punish cheaters may lead to betrayal aversion and how such ideas may find a way into computational trust and vice versa—leading to decreased risk aversion. Wing noted that if we train our machine-learning algorithms on biased data, the machines' predictions in future scenarios will be biased. When using network models to make policy decisions, this potential bias must be acknowledged. Since the predictions made by a model are only as good as the data it is trained on, we must be careful to present unbiased data so we can learn to trust our models. Without human trust, these models will only provide marginal benefits to the users.

Dr. Dwork began her talk by highlighting the need to protect the sensitive information that becomes available through social media platforms. Otherwise, sensitive information about individuals' private lives may be revealed and compromised. She explained why anonymization alone over network structure does not guarantee privacy. She presented an example of an anonymized network plot of nodes and edges, which represented the sexual partners of students from a particular high school. Names were removed, but

genders were indicated. She showed that several of the localized loops of the network could be used to identify particular students, despite the researchers' claims of anonymity. Even if an individual (a node) wishes to have no edges to another individual (another node)—such as a drug dealer who wishes to hide his distribution network—subgraphs can still be used in some cases to determine concealed connections.

Dwork proposed some mechanism to mitigate the issue of unique network neighborhoods that pose privacy risks with anonymized data. Next, she described her idea of differential privacy that data-processing algorithms should satisfy, in order to have good theoretical guarantees for privacy preservation over data sets. She concluded her talk by highlighting several theoretical and technical challenges in dealing with the differential privacy frameworks, and the opportunities for future research.

After presentations, speakers engaged in roundtable discussions. Professor Jadbabaie began by raising the issue of lack of a "physics" for social sciences and challenges of model building and prediction in social sciences. Several speakers contributed to the discussion by pointing out related aspects of the transition from data to models and decisions based on those models—and whether one can trust the models if they are based on biased data. For example, in the study of a recommendation system, one has to determine whether the research findings indicate aspects of human behavior or the properties of the recommendation system and the online media platform itself. Dr. Dwork raised the issue of fairness in online advertising and challenges of deciding what constitutes a fair performance, as ads are often targeted to a specific audience. Professor Jackson highlighted the need to understand the phenomenon through its underlying mechanisms, before deciding about norms and regulations. Finally, the speakers raised the issue of data sharing for research and public good. The speakers raised several challenges for future research such as network cascades, the need for sensitivity analysis of the models—as well as the legal, ethical and logistics issues that emerge when accessing social data.

Session 5: Future Electric Grid:

Authors: Devendra Shelar, Minghao Qiu, Magdalena Maria Klemun

The ongoing transformation of electric power systems is shaped by two competing forces: (1) rising population and affluence in emerging economies stimulate electricity demand growth and (2) decreasing emissions of greenhouse gases and other pollutants are required to mitigate climate change and health impacts of fossil fuel combustion. The convergence of end use sectors—transportation, manufacturing, buildings—through information and communications technologies and advanced controls makes electricity particularly important for decarbonizing the energy system as a whole. However, the required integration of time-variant, renewable energy sources transforms deterministic problems into more complex, stochastic modeling challenges. The "Future Electric Grid" session explored recent research findings across several of these dimensions, including: the use of health data to document relationships between air pollution and life expectancy in China; the selection of cost-optimal technology portfolios for grid-connected home

energy “hubs” with high renewables penetration in California and Germany; the development of computationally efficient methods to solve optimal power flow problems in large-scale, distributed electricity networks; and the computational and regulatory challenges associated with balancing supply and demand in day-ahead and real-time electricity markets with high penetration of renewable energy sources.

Michael Greenstone, Professor of Economics at the University of Chicago, highlighted the importance of natural experiments to understand causal links between air pollution and life expectancy changes in China. “No research ethics board is going to allow a randomized control trial where you can expose people to high levels of air pollution for a long time and a second group to no pollution,” said Greenstone. “As a consequence, many studies do not meet that kind of standard of evidence and have relied on short-term variation of pollution, or they have been carried out in countries with lower levels of air pollution than in China.”

In the study he presented, the randomized trial was replaced by a policy providing free, coal-based heating to people living North of the Huai River in China, but not to those in the South. Migration between the two areas was restricted. Using a time series of data on smoking and exercise habits, as well as mortality, Greenstone and collaborators found a 3.5-year loss of life expectancy of people in the North compared to the South, induced by particulate matter concentrations four times those in the U.S.

Sally Benson, Professor of Energy Engineering at Stanford University, presented a user-centric approach to the development of distributed energy infrastructure. “The home is currently a passive consumer of natural gas and electricity,” said Benson, “but in the future we can imagine a scenario where home energy management systems optimize the use of local resources, consider individual demand patterns and future weather, and provide electricity services back to the grid—there’s a big data and systems problem.”

Yet what are the economically optimal technology choices for electricity and heating, recognizing that elements of the portfolio—e.g. electric vehicles, solar panels and battery storage systems—are highly interconnected? To answer this question, Benson and colleagues used electricity demand and weather data from several communities in Germany and Northern California to simulate cost-optimal portfolios for homes during two sample years (2025 and 2035). “What you can see in all cases is that we are using a lot of solar in the cost-optimal solution, anywhere between 50-60% of total energy demand,” said Benson. Energy costs to the consumers were similar to the baseline case, despite large technology portfolio changes. In contrast to PV and battery storage, fuel cells performed less well, as system efficiency decreases from hydrogen production and storage needed to be offset by increased deployment of PV.

While Sally Benson’s work focused on individual homes, Stephen Low, Professor of Computer Science and Electrical Engineering at Caltech, highlighted large-scale architectural changes necessary to integrate distributed resources. “As vertically integrated, centrally controlled architectures are transformed into a layered architecture,

the power system will undergo a similar transformation as the telecommunication network and become the largest and most complex Internet of things," said Benson.

One key challenge, however, is the increasing number of "active endpoints"—non-dispatchable supply like wind turbines or intelligent loads like EVs—that make previous algorithms for economic dispatch inappropriate. Low explained, "But if we look at the network, it is solving these power flow equations for us, at scale, in real-time. We can exploit it as a power flow equation solver and develop algorithms that adapt to changing network conditions." One application of this idea, online optimization, is load-side frequency control incentivized by spot pricing, a combination that was first conceptualized in a paper by MIT's Fred Schweppe in 1980.

Bill Hogan, Raymond Plank Professor of Global Energy Policy at Harvard University, emphasized the societal challenges associated with overcoming the inertia in existing regulatory structures. One promising example can be found in the recent reforms to adopt a new business model for electric utilities in New York State, where distribution utilities are expected to act as Distributed System Platform providers to enable electricity market access for non-utility technology providers and end users. "It's easy to write the rule" said Hogan, "but the really hard part is going to be the sociological problem of persuading system operators that this will actually work. I'm quite confident this will not happen in my professional lifetime, so I'm not spending any time on this problem, but you can and MIT is right at the forefront."

Following the presentations, the discussion focused on challenges in developing countries. What can be learned from energy technology transitions in the developed world to meet growing electricity demand in emerging economies? What are the tradeoffs between less-costly, small-scale solutions like micro-grids, and large investments in high-voltage transmission lines and centralized power plants? Sally Benson said, "If developed countries are not provided the opportunity to develop a fossil fuel backbone, they are going to have suboptimal energy systems to support industrial development and provide well-being for their citizens." Is the value of centralized, dispatchable power, "underrated in the narrative about how the developing world can leapfrog the developed world," as Sally Benson puts it? Michael Greenstone responded, "The goal of energy systems is not only to reduce emissions—there are three goals: economic growth, don't kill yourself from air pollution, and don't burn up the planet. It's hard to have a system that does all three things at once, but what should not be lost is a central role of energy in advancing people's well-being. There are few substitutes for the power grid in terms of facilitating that."

[Session 6: Student Session]

Session 7: Analyzing our Health:
Author: Shreya Saxena

The advent of big data is a fairly recent phenomenon in the healthcare system. Many hospitals are still in the process of adopting Electronic Medical Records (EMRs), as well

as taking steps to digitize older paper records. These digital health records enable the accurate storage of patient data and simplify medical diagnoses based on patient history. Their secondary use, which is becoming more and more relevant for a wide variety of different applications, is population-based research studies.

With the introduction of EMRs as a common practice in most major hospitals, it is vital to be able to combine datasets in order to obtain patient-specific, personalized analytics on health data. For example, while the success rate of a certain therapy may be known across all people who try it, it is important to ask whether these statistics change when considering a patient's age and other characteristics. To have a big enough dataset to answer these questions, we need to go from multiple "small" datasets to one "big" dataset.

As in many other fields, the main issues when analyzing healthcare data are: (a) complexity and heterogeneity of data, (b) privacy, and (c) policy. While the need for technological growth is evident in addressing these questions, it can no longer be ignored that these are all societal problems as well.

The need for the combination of datasets implies a need for standardization across datasets collected in different hospitals and medical agencies. Standardization needs to be performed at a national and an international level, the latter being increasingly relevant in this age of global migration. This standardization requires an agreement in measurement procedures, data storage types, choice of software, etc. The wide range of complex data types and diseases makes this problem even more challenging.

The rewards may be great, however, since this standardization and combination of different datasets can turn data into wisdom if the right questions are asked. Personalized medicine, advancement of healthcare research, and advent of policy are just some of the possible advantages of having bigger datasets.

In order to empower the public, patients should be able to access their own health records that are available to doctors. This can lead to a "Live ICU," or better patient-facing apps that might lessen the burden on the medical practitioners. It would also be ideal to combine these personal medical records with other data readily available on most smartphones, like number of steps walked, etc. However, common frameworks would need to be established in order to turn these heterogeneous data streams into actionable insights.

While combining datasets has issues related to the complexity and heterogeneity of data, another important point arises when considering the privacy of the patients. Anonymization is a key aspect of alleviating trust issues while moving forward with combining different smaller datasets and sharing data across hospitals and medical agencies. Moreover, if the data is de-identified in a proper manner, it can be published in a common database for researchers from all fields to ask relevant questions. The potential paybacks from such a scheme are high, but the privacy of the patients is the main hurdle. Advantages to such a common database include the potential of researchers outside of the

medical field to apply sophisticated statistical models to answer relevant questions, as long as the information is available in a salient and anonymized manner. “Open” data can be very beneficial if care is taken to properly de-identify the patients, and a discussion of the interplay between technological and societal challenges is a key issue here.

From a governmental perspective, it is important to create policies that will responsibly unleash the power of data to all Americans, and healthcare is one of the most important fields for this. On the ground, the data that leads to policy decisions is often simplistic in form, with minimal data modeling implemented. For example, it was noticed in Cook County that there was a high degree of overlap between the people who require mental health services and those going in and out of prison. *(Note: View source here.)* This led to policy changes to provide further help to mental health patients to be rehabilitated, instead of spending more time in prison. In order to make these policy changes, we need on-the-ground solutions, for example, training the people in charge in crisis intervention.

In order to address healthcare issues, policy also needs to keep up with current problems. The governmental regulatory agencies are an integral part of bringing the technology and healthcare research into active use by the public. These agencies rely on researchers to carry out data collection and clinical trials, but they need a significant amount of data in order to test the technology. The collection of a large amount of data in a short amount of time can be aided by: (i) combining clinical trials by different bodies of research, (ii) embedding mechanistic studies within clinical trials, and (iii) the active collaboration between academia and industry. These steps will help the rapid, widespread adoption of technology in the healthcare sector.

More so with healthcare than most other fields, it is imperative to keep in mind that there are lives that hang in the balance when addressing the concerns raised in this session; if one misses some “edge cases,” these are human beings whose lives may be lost due to the policies or lack thereof. The issues discussed are crucial to consider while expanding the use of technology and data to advance research and policy in the healthcare field.

The Institute for Data, Systems, and Society is well-placed to research the use of healthcare data to answer questions at the intersection of highly technological systems and societal challenges.

Session 8: Driving Smart Cities Forward

Author: Ian Schneider

The IDSS Launch Event brought together a diverse group of experts to discuss the challenges cities face today and in the future. The speakers drew attention to the range of interesting and useful data that can be potentially extracted from cities, suggesting that effectively harnessing this data will be a key challenge as researchers and governments work to tackle urban problems in the future.

MIT Professor Sarah Williams opened the discussion by emphasizing that data will only

change the world if it used for social good. This point was underscored by each of the visiting speakers, who focused their talks on ways urban data can improve the public good while also focusing on the ethics of data to make sure it is not used for harm, nor excessively privatized at the expense of the very people by whom it is generated. Dr. Williams also emphasized the need for data literacy and for new visualizations and tools that improve the usefulness of data for public and government consumption.

Dr. Steven Koonin of NYU described the tremendous opportunity for data science to tackle problems in cities, and he explained how NYU's Center for Urban Science and Progress is working to harness that data while enhancing collaboration with government and other diverse stakeholders. Dr. Koonin argued that data in cities is already plentiful and diverse, driven by digitization, open data initiatives, and increasingly less expensive sensors. He explained that cities could be instrumented even further by collecting existing data flows or adding sensors or novel instrumentation to measure data related to infrastructure, environment, or people. Dr. Koonin also discussed some of the challenges of tackling data problems in the societal context. A defining problem is the difficulty of fusing the cultures of data scientists, social scientists, and civil servants. In order to make sure data can have a beneficial impact on policy and regulation, it is important to make sure that data experts and policymakers are increasingly comfortable working together.

While Dr. Koonin focused on harnessing city data in an efficient way, Dr. Rob Kitchin of Maynooth University and Dr. Susan Crawford of Harvard Law School emphasized important issues in safety, resiliency, and equity of data analysis and collection systems. The differences in these approaches underscored important concerns about ethics, risk, and equity in urban data systems as data collection increases and governments and researchers improve their ability to analyze the data. Dr. Kitchin introduced a range of ethics and security concerns related to data in smart urban centers, including surveillance and privacy erosion, control creep, and predictive profiling. He also expressed concern that online and data-driven management could lead to buggy, hackable urban systems.

Dr. Crawford emphasized two key questions related to big data: (1) "Is it good for democracy?" and (2) "Is it good for people?" She made the case for the need for responsive communities that emphasize the importance of digital justice and data stewardship, and for city and citizen control of their own data. Both Dr. Kitchin and Dr. Crawford highlighted concerns related to resiliency and ethics when the data citizens generate is owned and used by external parties without the consent of the very people that data measures. Dr. Crawford described how the Responsive Communities Initiative encourages students to work closely with mayors and cities. She focused on key issues in urban areas related to infrastructure, arguing that it is essential to address barriers to ubiquitous, affordable, high-speed Internet access as a matter of social justice, and advocated for a large expansion of open fiber Internet access. Dr. Crawford also focused on the need for leadership in the use of data, emphasizing the importance for engagement with policy makers. Other speakers at the event echoed this critical issue.

Dr. Balaji Prabhakar of Stanford University and Dr. Alexandre Bayen of U.C. Berkeley rounded out the panel by focusing on the specific application area of transportation as one

class of problems faced by urban centers. While the previous speakers discussed the collection of data, its use, and related concerns in a range of settings, these speakers focused especially on new work that has the potential to improve the efficiency of congested transportation systems in metropolitan areas. In this setting, the ability of data to drive changes in people's behavior is essential, so their work also touched on the challenges of influencing behavior through data-based and socially-oriented recommendations. Dr. Bayen explained how mobile measurements have revolutionized transportation routing, allowing software and users to gather and act on information that was previously unobservable. In particular, location data gathered from mobile phones allows for more fine-grained state estimation, enabling higher fidelity routing with decision points throughout the route. Dr. Bayen also discussed additional social implications of this sort of technology, for instance because citizens protest higher traffic density on small neighborhood streets. Both Dr. Bayen and Dr. Prabhakar discussed challenges related to changing human routing and driving behavior through incentives. Dr. Prabhakar described his work in Singapore to provide incentives for travelers to reduce peak time travel on Singapore's subway system, highlighting the ability of small rewards to have a major impact on peak demand.

The information and data density of cities means that there is tremendous opportunity for using data to help solve the problems of urban and metropolitan areas. The speakers on the panel highlighted problems related to governance, environmental health, and especially transportation as potential candidates for data solutions and increased collaboration between data scientists and regulators. While many panelists focused on the capacity of big data for furthering social good, some also discussed potential problems with increased data access with regards to privacy, ethics, and resilience. It will be exciting to see how researchers in programs like IDSS tackle these challenges while working to harness data to solve the problems of urban areas.

Session 9: From Applications to Theory

Author: Alfredo Guzman-Morales

This session, "From Applications to Theory," was moderated by Professor Caroline Uhler (MIT), and featured Professor Allen Tannenbaum (Stony Brook University), Professor Elchanan Mossel (MIT), Professor David Tse (Stanford University), and Professor Vincent Blondel (Université catholique de Louvain).

Big data has transformed the way we approach a variety of problems across multiple domains. In different fields, like finance or medicine, data stream in and enable the possibility of predictions and early time interventions. However, we still face multiple challenges when it comes to transforming data into action. One of the main challenges is our ability to create models to analyze the systems that underlie data. Fundamental sciences like mathematics and physics enable the transition from big data to models and theories. Models should be able to explain systems' behavior without making wrong

assumptions or compromising privacy, while being able to compute data efficiently and accurately. It is critical to develop methods to ensure these purposes.

We are usually interested in the systems that underlie data, rather than in data itself. The data represent nothing but a proxy to analyze the properties of systems. In order to make claims from data, we need models to structure our analysis. Models and abstractions provide a framework of understanding about systems beyond datasets, regardless of their size. Models create the opportunity for asking better questions and for designing tools to retrieve the right answers. In addition, good models and theories generally yield more robust and consistent results. Without such understanding, we can easily be fooled by the data, misleading the research and obtaining wrong outcomes. Numbers are meaningless if they lack the framework of understanding that models provide.

Historically, fundamental theories that have changed the course of science have been developed without any access to big data. For example, Riemann, father of differential geometry, outlined the way we understand the cosmos by looking at the curvature of objects. His studies served as basis for later scientific advances like the development of Einstein's theories or modern electronic and mobile communication. Another example is Boltzman, father of statistical mechanics and information theory. He understood that entropy has hidden information while describing the motion of millions of particles without ever looking at them. Finally, Shannon developed important theories about the interplay between information and computation. When Shannon's theories were developed, he looked at theorems of combination or computation in the abstract. These examples demonstrate that models and theories can explain systems without the need of detailed data.

However, models are not enough. Domain expertise is also needed in order to make the right assumptions and interpretations. Data-driven results might seem objective, but they hide crucial assumptions and decisions that are subjective. We should be able to balance the interplay between data and theory, as one may serve as feedback to revise the other. Good theories lead to good experiments and also to good data. Learning the nature of systems is crucial in order to develop tools for characterizing their behavior, and for making inferences, predictions, or strategies. Data science provides the toolbox for retrieving these patterns of information, but it does not explain their origins or implications.

Most of the real-world systems are complex. The behavior of their parts is not independent from one other and, rather, is connected in non-trivial ways—creating patterns of information at multiple scales of observation. Traditional statistics are not always valid in these kinds of systems. In multiple domains, problems arise from using traditional statistics and machine-learning techniques, given a misunderstanding of the underlying assumptions. For instance, in medicine, the number of samples is usually

greater than the number of features. Even if datasets are big, the data come from different laboratories, and are sporadic and not always consistent with each other. This creates problems when estimating probabilities, since we do not compare similar populations nor take into account the specific conditionals of each sample. Moreover, not all individuals respond equally to similar treatments, and extreme cases are not generally considered during experiments. Similar problems occur in finance. Misunderstanding fat-tailed distributions lead to the underestimation of risks and the probabilities of extreme events. As a result, fragile strategies are deployed that do not respond to the complex nature of the financial sector and often fail during volatile events which are not considered in models as part of the system. We should therefore understand the nature of systems in order to choose the adequate toolset for the analysis.

Privacy is another factor that must be taken into account when analyzing data. Tracking physical particles or cells is different than analyzing people's behavior. Communication, mobility, and medical records should be protected in order to ensure privacy. We need to develop methods to retrieve knowledge from data without compromising privacy. Aggregation can be a solution for such this. The large-scale properties of systems are often present at the collective behavior, and not in the individual details. The challenge is to determine the level of aggregation that ensures privacy without blurring out the important information.

In summary, models and theories structure the analysis of data and enable the design of good experiments. Domain experts are needed in order to evaluate assumptions, understand their implications, and create better interpretations. Data science provides the tools for retrieving patterns of information, but models and theories provide context and meaning. Therefore, data and models should complement each other in order to build a better understanding of the underlying systems.