Data&Society
RESEARCH INSTITUTE

NEW YORK UNIVERSITY
Information
LAW
Institute

THE SOCIAL, CULTURAL & ETHICAL DIMENSONS OF "BIG DATA"

March 17, 2004 · NYU School of Law
Full Day Schedule

The Data & Society Research Institute, the White House Office of Science and Technology Policy, and New York University's Information Law Institute are pleased to welcome you to the "The Social, Cultural, & Ethical Dimensions of 'Big Data' ". The purpose of this event is to convene key stakeholders and thought leaders from across academia, government, industry, and civil society to examine the social, cultural, and ethical implications of "big data," with an eye to both the challenges and opportunities presented by the phenomenon.

This event is one of three conferences that OSTP is co-hosting with academic institutions across the country that will examine key questions on the use of "big data" and the future of privacy. Other events include a conference organized by the Massachusetts Institute of Technology (MIT) Big Data Initiative at CSAIL, and the MIT Information Policy Project that focused on the technical aspects underpinning privacy, and an event organized by the School of Information with the Berkeley Center for Law and Technology at University of California-Berkeley which will explore the legal and policy issues raised by big data. These are all part of ongoing efforts by the Obama Administration to review the implications of collecting, analyzing and using massive or complex data sets for privacy, the economy, and public policy.

SPONSORS

The Data & Society Research Institute acknowledges the generous gifts and institutional support of the following sponsors for making this event possible: Alfred P. Sloan Foundation, Ford Foundation, John D. and Catherine T. MacArthur Foundation, the John S. and James L. Knight Foundation, Microsoft Research, and Robert Wood Johnson Foundation. Outcomes from this event will help inform the anticipated National Science Foundation-supported Council on Social, Legal, and Ethical aspects of Big Data. These funds were not solicited or collected on behalf of the Office of Science & Technology Policy (OSTP), or the White House. Acknowledgment of a contributor by the Data & Society Research Institute does not constitute an endorsement by OSTP or the White House.

SCHEDULE

Daytime Event (invited guests only)

- **10:00-10:30**: Conference Check-in
- **10:30-10:45**: Introduction from danah boyd, Data & Society
- **10:45-11:15**: Fireside Chat with John Podesta, White House
- **11:15-12:00**: Firestarters
    - **Tim Hwang:** On Cognitive Security
    - **Nick Grossman:** Regulation 2.0
    - **Nuala O'Connor:** The Digital Self & Technology in Daily Life
    - **Alex Howard:** Data Journalism in the Second Machine Age
    - **Mark Latonero:** Big Data and Human Trafficking
    - **Corrine Yu:** Civil Rights Principles for the Era of Big Data
    - **Natasha Schüll:** Tracking for Profit; Tracking for Protection
- **12:00-1:00**: Lunch
- **1:00-1:30**: Firestarters
    - **Kevin Bankston:** The Biggest Data of All
    - **Alessandro Acquisti**: The Economics of Privacy (and Big Data)
    - **Latanya Sweeney:** Transparency Builds Trust
    - **Deborah Estrin:** You + Your Data
    - **Clay Shirky:** Analog Thumbs on Digital Scales
- **1:30-2:30**: Discussion, moderated by danah boyd & Nicole Wong, White House
- **2:30-3:00**: Break & Transition

Workshops (invited guests only)

- **3:00-5:00**: Workshops
    - Data Supply Chains
    - Inferences and Connections
    - Predicting Human Behavior
    - Algorithmic Accountability
    - Interpretation Gone Wrong
    - Inequalities and Asymmetries

Public Plenary

- **5:00-5:30**: Check-in for Public Plenary
- **5:30-5:38**: Welcome from danah boyd, Data & Society
- **5:38-5:40**: Video from John Podesta, White House
- **5:40-6:00**: Keynote by Nicole Wong, White House
- **6:00-6:30**: Statements by Plenary Panel
    - **Kate Crawford,** Microsoft Research and MIT
    - **Anil Dash,** Think Up and Activate (*moderator*)
    - **Steven Hodas,** NYC Department of Education
    - **Alondra Nelson,** Columbia University
    - **Shamina Singh,** MasterCard Center for Inclusive Growth
- **6:30-7:00**: Moderated Plenary Panel
- **7:00-7:30**: Public Comments & Q&A

Evening Reception

- **7:30-8:30**: Reception

FIRESTARTER TALK ABSTRACTS

Alessandro Acquisti: The Economics of Privacy (and Big Data): Facts, Myths, and Unknowns

A substantial amount of as of yet untested assumptions pervades the debate over the economic benefits of big data and its potential privacy trade-offs. I will use past and current findings from the economics literature to highlight a) which facts concerning the economics of privacy and personal data we have a decent grasp of; b) which myths we should try to dispel; and c) which crucial questions have not yet been solved, or even addressed, but should be.

Alex Howard: Data Journalism in the Second Machine Age: on Accountability, Algorithms and Transparency

Journalists have been adapting and adopting technology to gather information, report news and hold the powerful to account for centuries. In the 20th century, computer-assisted reporting provided investigative journalists with new ways to apply an empirical lens to creating acts of journalism. In the 21st century, data journalists are using and creating powerful new tools and platforms to understand the world, tell stories and inform the public.

Clay Shirky: Analog thumbs on digital scales: Hard-to-detect self-dealing by data-centric organizations

There are some risks of self-dealing by data-driven organizations that can't be solved through transparency. When algorithmic selection relies on random or frequently re-weighted inputs, an organization can run millions or billions of trial runs, then select the one it prefers, allowing them to optimize for privately held preferences. Even an auditor with complete access to both the data and the algorithm couldn't detect this form of self-dealing. This points more generally to the need to consider governance alongside transparency as a check on abuse.

Corrine Yu: Civil Rights Principles for the Era of Big Data

On February 27, a broad coalition of leading civil rights and media policy organizations came together to endorse Civil Rights Principles for the Era of Big Data. For the coalition as a whole, and for many of the signatories, this is a historic first step into the national conversation that's starting now around big data. Through these principles, the signatories

highlight the growing need to protect and strengthen key civil rights protections in the face of technological change. They call for an end to high-tech profiling; urge greater scrutiny of the computerized decision-making that shapes opportunities for employment, health, education, and credit; underline the continued importance of constitutional principles of privacy and free association, especially for communities of color; call for greater individual control over personal information; and emphasize the need to protect people, especially disadvantaged groups, from the documented real-world harms that follow from inaccurate data.

Deborah Estrin: You and Your Small Data

Each time you use their smartphone, social media, search engine, mobile game, or loyalty card, they implicitly generate a trail of digital breadcrumbs that together form a digital trace of your activities. We call these data that are particular to an individual, their small data. These data are used in aggregate by digital services to improve system performance, tailor service offerings, conduct research, and target ads. But these highly personalized, data can also be analyzed to draw powerful inferences about your health and everyday behaviors. In the future we are building towards at Cornell Tech and Open mHealth, and more broadly in the community, you would be able to opt in to access to your digital traces through Personal Data APIs, and to choose apps that privately process, fuse, and filter your Small Data for you.

Kevin Bankston: The Biggest Data of All

When talking about big data, it's worth stepping back to talk about the biggest data set of all, available only to ISPs and governments--the Internet and voice traffic that transits the Internet backbone. What are the current rules and practices when it comes to accessing and using that data? What types of data are we talking about, and what kinds of distinctions between types of data make sense? For example, does the longtime distinction between communications "content" and "metadata", with the latter receiving weaker protection, make sense in an age of bulk data analysis where metadata can often be just as revealing, if not more revealing, than the actual content of your communications? And what responsibilities do ISPs and governments have when it comes to handling this vast stream of information that includes the mass of our personal and private communications and data?

Latanya Sweeney: Transparency Builds Trust

Our national security and largest hi-tech companies have found value in personal data sharing. They trust the quality of the personal data they receive to make revenue and

national security decisions. The opposite is true too. Lack of trust breeds distrust. Not knowing what others do with the personal data they receive leaves the public distrustful and vulnerable. When data sharing is transparent (e.g., thedatamap.org), society can understand risks and harms and pose effective remedies and technical solutions. Approaches might be reporting arrangements recursively along data sharing chains or allowing individuals to audit their own data sharing chains (e.g., HIPAA).

Mark Latonero: Big data and human trafficking

This presentation will discuss the socio-technical complexities that arise as big data and analytics are applied to human rights interventions, using human trafficking as a specific example. The promises of big data approaches for human trafficking interventions raise important questions about data privacy, monitoring, and interpretation. What are the responsibilities of data and technology companies or computer and social science researchers (monitoring and analyzing a number of human sensor networks) when they are able to detect evidence of the most pressing human rights violations? Can policies be developed that strike the right balance between privacy, data ownership, and human rights benefits?

Natasha Schüll: Tracking for profit; tracking for protection

Casinos' data-intensive player-tracking systems, developed to gather behavioral intelligence, run predictive analytics, and more profitably market to gamblers (by mail or in the moment), are in some cases also equipped with algorithms that can detect patrons,Äô problematic gambling behavior and prompt real-time intervention. In some jurisdictions, player-tracking systems are set up to perform both roles simultaneously, despite their apparent conflict. This case, in which a single data-monitoring platform serves at once to persuade and to protect consumers, brings to the fore the challenging tensions at the heart of "big data" as it plays out in our everyday life transactions.

Nick Grossman: Regulation 2.0: Bringing an Internet Approach to Real-World Regulation

The last 15 years on the internet has been an experiment in generating trust among strangers. Along the way, Internet-based networks, from eBay to Airbnb, have invented increasingly sophisticated trust and safety mechanisms that in many ways mirror the intentions of government regulations, but achieve their goals through intensive use of peer review and data, rather than up-front granting of permission. This approach to regulation is in itself a major innovation. So, as more and more internet-based businesses bump into real-world regulatory regimes, the question is: should we regulate them the old way or the Internet way?

Nuala O'Connor: The Digital Self & Technology in Daily Life

The increasing ubiquity of technology in daily life leads us to take for granted the sometimes opaque decision making inherent in seemingly personalized algorithms. O'Connor poses the questions "do we know what motivates the machine" and what is the impact on individuality, on learning and distribution of knowledge, on our digital selves.

Tim Hwang: On Cognitive Security

Data is the raw material of prediction. To that end, the increasing availability of social data opens the possibility of predicting and influencing the behavior of large groups. This data is -- and increasingly will be -- used by actors to shape (and reshape) social systems at scale. This talk will briefly explore what would appear to be the logical next step: what occurs when quantitative methods of influence and the parties that use them come into competition with one another? It will discuss the concept of "cognitive security," and how big data is poised to inform the future landscape of persuasion.

WORKSHOP DESCRIPTIONS

Algorithmic Accountability

Facilitators: Nicholas Diakopoulos and Gideon Lichfield
Rapporteur: Malte Ziewitz

Accountability is fundamentally about checks and balances to power. In theory, both government and corporations are kept accountable through social, economic, and political mechanisms. Journalism and public advocates serve as an additional tool to hold powerful institutions and individuals accountable. But in a world of data and algorithms, accountability is often murky. Beyond questions about whether the market is sufficient or governmental regulation is necessary, how should algorithms be held accountable? For example, what is the role of the fourth estate in holding data-oriented practices accountable?

Data Supply Chains

Facilitators: Anne Washington & Tim Hwang
Rapporteur: Seda Gurses

As data moves between actors and organizations, what emerges is a data supply chain. Unlike manufacturing supply chains, transferred data is often duplicated in the process, challenging the essence of ownership. What does ethical data labor look like? How are the

various stakeholders held accountable for being good data guardians? What does clean data transfer look like? What kinds of best practices can business and government put into place? What upstream rights to data providers have over downstream commercialization of their data?

Inequalities and Asymmetries

Facilitators: Neil Richards and Nichole Pinkard
Rapporteur: Elana Zeide

The availability of data is not evenly distributed. Some organizations, agencies, and sectors are better equipped to gather, use, and analyze data than others. If data is transformative, what are the consequences of defense and security agencies having greater capacity to leverage data than, say, education or social services? Financial wherewithal, technical capacity, and political determinants all affect where data is employed. As data and analytics emerge, who benefits and who doesn't, both at the individual level and the institutional level? What about the asymmetries between those who provide the data and those who collect it? How does uneven data access affect broader issues of inequality? In what ways does data magnify or combat asymmetries in power?

Inferences and Connections

Facilitators: Janet Vertesi and Tim Sparapani
Rapporteur: Karen Levy

Data-oriented systems are inferring relationships between people based on genetic material, behavioral patterns (e.g., shared geography imputed by phone carriers), and performed associations (e.g., "friends" online or shared photographs). What responsibilities do entities who collect data that imputes connections have to those who are implicated by association? For example, as DNA and other biological materials are collected outside of medicine (e.g., at point of arrest, by informatics services like 23andme, for scientific inquiry), what rights do relatives (living, dead, and not-yet-born) have? In what contexts is it acceptable to act based on inferred associations and in which contexts is it not?

Interpretation Gone Wrong

Facilitators: Ritu Agarwal and Gilad Lotan
Rapporteur: Heather Patterson

Just because data can be made more accessible to broader audiences does not mean that those people are equipped to interpret what they see. Limited topical knowledge, statistical skills, and contextual awareness can prompt people to read inferences into, be afraid of, and

otherwise misinterpret the data they are given. As more data is made more available, what other structures and procedures need to be in place to help people interpret what's available?

Predicting Human Behavior

Facilitators: Seeta Peña Gangadharan and Jason Schultz
Rapporteur: Solon Barocas

Countless highly accurate predictions can be made from trace data, with varying degrees of personal or societal consequence (e.g., search engines predict hospital admission, gaming companies can predict compulsive gambling problems, government agencies predict criminal activity). Predicting human behavior can be both hugely beneficial and deeply problematic depending on the context. What kinds of predictive privacy harms are emerging? And what are the implications for systems of oversight and due process protections? For example, what are the implications for employment, health care and policing when predictive models are involved? How should varied organizations address what they can predict?