October 17, 2014

Suzanne H. Plimpton, Reports Clearance Officer
National Science Foundation
NCO, Suite II-405
4201 Wilson Blvd
Arlington, VA 22230

RE: Request for Information (RFI)-National Privacy Research Strategy

Dear Ms. Plimpton,

On behalf of the Information Technology and Innovation Foundation (ITIF), we are pleased to submit these comments in response to the National Science Foundation's (NSF) request for information concerning a national privacy research strategy for the Networking and Information Technology Research and Development (NITRD) Program.[1]

ITIF is a nonprofit, non-partisan public policy think tank committed to articulating and advancing a pro-productivity, pro-innovation, and pro-technology public policy agenda internationally, in Washington, and in the states. Through its research, policy proposals, and commentary, ITIF is working to advance and support public policies that boost innovation, e-transformation, and productivity.

The ability to share data is a key component of data-driven innovation.[2] Not only does widespread data sharing offer enormous economic benefits—McKinsey Global Institute estimates open data has the potential to unlock $3 trillion annually—it also offers widespread societal benefits, including

[1] "Request for Information (RFI)-National Privacy Research Strategy," *Federal Register*, September 18, 2014, https://www.federalregister.gov/articles/2014/09/18/2014-22239/request-for-information-rfi-national-privacy-research-strategy.
[2] Daniel Castro and Travis Korte, "Data Innovation 101," *Center for Data Innovation*, November 3, 2013, http://www.datainnovation.org/2013/11/data-innovation-101/.

saving lives.[3] For example, Google CEO Larry Page has estimated that better data sharing in health care could save an additional 100,000 lives per year.[4] However, privacy and security challenges threaten to stall the deployment of "big data" initiatives that offer many potential economic and quality-of-life benefits to consumers.[5]

To ensure the potential benefits of data-driven innovation are attained, the U.S. federal government should support research efforts to address the most pressing privacy and security research questions faced by industry and government.[6] The U.S. government funds millions of dollars in research and development for computer science, including efforts to improve privacy enhancing technologies, but these efforts have not been aligned with the strategic needs of industry and government. To maximize the benefits of federally-funded privacy research, a clear set of research goals and objectives is needed.

The following filing will describe five areas—health care, transportation, criminal justice, education, and social networks—where additional research is needed on how to share data while best preserving privacy.

Health Care

Data has the potential to vastly improve health care services in the United States. Increased sharing and use of health care data will produce a variety of benefits for this sector, including more personalized and coordinated care, faster treatment development, increased efficacy of treatments,

[3] Michael Chui, Diana Farrell, and Kate Jackson, "Open data: Unlocking innovation and performance with liquid information," *McKinsey Global Institute*, October 2013, http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.
[4] Alex Hern, "Google: 100,000 Lives a Year Lost Through Fear of Data-Mining," *The Guardian*, June 26, 2014, http://www.theguardian.com/technology/2014/jun/26/google-healthcare-data-mining-larry-page.
[5] Daniel Castro, "The Need for an R&D Roadmap for Privacy," *The Information Technology and Innovation Foundation*, August 2012, 1, http://www2.itif.org/2012-privacy-roadmap.pdf.
[6] Ibid.

and lower costs.[7] However, to achieve these goals, it will be important to find more ways to improve the privacy and security of health data when it is shared between different stakeholders, such as clinicians, insurers, and researchers. In fact, a 2014 survey found that, with appropriate anonymity, 94 percent of social media users in the United States with a medical condition would be willing to share their health data to help doctors improve care.[8]

There is still a long way to go to provide patients assurance that the health data they share is protected. The U.S. Department of Health and Human Services reports that since 2009 there have been almost 1,000 data breaches affecting the health care records of more than 30 million total individuals.[9] In addition, consumers need assurance that the data collected by popular mobile health and fitness applications have been properly secured and de-identified.[10]To address the challenges associated with sharing health care data, additional research is needed for effective ways to use a mix of new technologies to ensure that data is properly safeguarded and consumers are protected. For example, health care providers could limit the way that the data they share can be used or researchers may want to build databases of aggregated patient information that provide certain privacy guarantees while also facilitating new discoveries.

Transportation

The prevalence of in-car and mobile navigation systems, as well as roadway sensors, traffic cameras, and other sensors integrated into intelligent transportation systems, allows for increased collection of mobility data, i.e. data about individual movements. In addition, in the near future, there will be a

[7] Daniel Castro, "The Rise of Data Poverty in America," *Center for Data Innovation,* September 10, 2014, http://www2.datainnovation.org/2014-data-poverty.pdf.
[8] Francisco Grajales et al., "Social Networking Sites and the Continuously Learning Health System: A Survey," *Institute of Medicine, National Academy of Sciences*, February 4, 2014, 17, http://www.iom.edu/~/media/Files/Perspectives-Files/2014/Discussion-Papers/VSRT-PatientDataSharing.pdf.
[9] Jason Millman, "Health care data breaches have hit 30M patients and counting," *The Washington Post*, August 19, 2014, http://www.washingtonpost.com/blogs/wonkblog/wp/2014/08/19/health-care-data-breaches-have-hit-30m-patients-and-counting/.
[10] Craig Michael Lie Njie, "Technical Analysis of the Data Practices and Privacy Risks of 43 Popular Mobile Health and Fitness Applications," *Privacy Rights Clearinghouse*, August 12, 2013, http://www.privacyrights.org/mobile-medical-apps-privacy-technologist-research-report.pdf.

substantial amount of data generated from connected vehicles—cars that are able to wirelessly link together and warn each other of traffic, accidents, and many other dangers.[11] Collecting and sharing transportation data has many potential benefits, including reducing motor vehicle deaths and injuries, improving traffic flows, and increasing fuel efficiency.[12] This data also allows city planners to better provide municipal services, including transit options, to citizens.

In December 2013, the Government Accountability Office (GAO) did a study analyzing the extent of data collected by auto manufacturers, portable navigation device companies, and map and navigation application developers.[13] GAO found that while all of these companies collected location data, only a few shared this data, doing so only when personally identifiable information has been removed.[14] However, since transportation data includes geo-location information, it is difficult to strip the data of personally identifiable information while still preserving the data's utility.[15] For example, mobility data will be identifiable for some people since it may contain information about where individuals work and live, as well as sensitive information about the places they visit. Better tools to de-identify high-dimensional data, such as mobility data, would make it easier for organizations and researchers to share these types of data sets privately and securely.

Criminal Justice

The U.S. criminal justice system allows law enforcement agencies, including police departments, courts, and prisons, to collect and analyze many different types of data to create a "virtual picture" of

[11] Andreas Mai and Dirk Schlesinger, "A Business Case for Connected Cars," *Cisco*, April 2011, http://www.cisco.com/web/about/ac79/docs/mfg/Connected-Vehicles_Exec_Summary.pdf.
[12] "The Importance of Sharing Data," *National Highway Traffic Safety Administration*, March 2007, http://www-nrd.nhtsa.dot.gov/Pubs/810687.pdf.
[13] "Companies Are Taking Steps to Protect Privacy, but Some Risks May Not Be Clear to Consumers," *U.S. Government Accountability Office*, December 2013, http://www.gao.gov/assets/660/659509.pdf.
[14] GAO reported that none of these companies sold personally identifiable location data to, or shared this information with marketing companies or data brokers.
[15] Ann Cavoukian and Daniel Castro, "Big Data and Innovation, Setting the Record Straight: De-identification Does Work," *The Information Technology and Innovation Foundation*, June 16, 2014, http://www2.itif.org/2014-big-data-deidentification.pdf.

individuals, which can then be shared by the various agencies within the system.[16] This data includes, but is not limited to, criminal history information, criminal intelligence information, juvenile justice information, and supplemental information. Supplemental information consists of non-criminal information including tax records, credit reports, organization affiliations, and various other types of data.[17] Some individuals have expressed concern about the availability of this data, especially since it may be reproduced or used beyond its original purposes even after it is withdrawn from official government databases. For example, a judge may order that court records be expunged, but references to the court records outside of the government database may still exist. Law enforcement agencies need the ability to efficiently and effectively share data, doing so in ways that allow them to protect against misuse. Therefore, processes need to be developed to ensure that law enforcement officials can share data sets electronically in ways that are irrevocably linked to self-enforcing data-handling policies, similar to the way digital rights management (DRM) technology enforces certain data-handling restrictions on multimedia content.

Education

Data gathered by schools will not only help government leaders create more effective and efficient education policy, but it will also allow families to find the best school and teachers to create personalized lesson plans. There are many ways using data can improve K-12 education, including adaptive learning software to personalize concepts to individual students and learning styles, performance-based measures to reward the best teachers, and predictive analytics to assist with teacher hiring or selecting colleges.[18] The foremost beneficiary of all of these improvements are the students.

[16] Paul Kendall, Neal Swartz, and Anne Gardner, "Gathering, Analysis, and Sharing of Criminal Justice Information by Justice Agencies: The Need For Principles of Responsible Use," *U.S. Department of Justice*, 1994, http://www.justiceprivacy.com/pdf/Kendall_Principles_responsibleUse.pdf.

[17] Ibid.

[18] Daniel Castro, "The Rise of Data Poverty in America," *Center for Data Innovation,* September 10, 2014, 4, http://www2.datainnovation.org/2014-data-poverty.pdf.

As schools begin collecting, sharing, and using more data, they need to be able to develop secure systems to protect the privacy of students while enabling access to a broad set of stakeholders who can use this data for beneficial purposes. One important component will be ensuring that schools have access to identity management tools to enable secure, multi-party access to sensitive information and granular access controls for different data sets.

Social Networks

Social networks allow users to share content and other information with each other. These social networks offer a rich array of data that can be used for useful purposes, including advanced personalization of content, connecting like-minded people, and connecting people to businesses and government. But there are also many emerging opportunities to use social network data for research and other socially beneficial purposes. One study shows how real-time social networks like Twitter can be used to track HIV incidence and drug-related behaviors with the intention of detecting and preventing outbreaks.[19] An example of this in practice is HealthMap, a collaborative epidemiological mapping effort that incorporates social media data and news reports to track diseases.[20] HealthMap was able to detect the ongoing West African Ebola outbreak nine days before World Health Organization authorities became aware of it.[21] However, the operators of social networks need better tools to enable much of this beneficial research while respecting user privacy. Doing so requires the ability to use techniques, such as privacy-preserving data mining, where more technical research is needed.[22]

[19] Mark Stoove and Alisa Pedrana, "Making the most of a brave new world: Opportunities and considerations for using Twitter as a public health monitoring tool," *Preventative Medicine*, Vol. 63, June 2014, 109-111, http://www.sciencedirect.com/science/article/pii/S0091743514001029.
[20] "About," *HealthMap*, accessed October 15, 2014, http://healthmap.org/site/about.
[21] Zoë Schlanger, "An Algorithm Spotted the Ebola Outbreak Nine Days Before WHO Announced It," *Newsweek*, August 11, 2014, http://www.newsweek.com/algorithm-spotted-ebola-outbreak-9-days-who-announced-it-263875.
[22] Yehuda Lindell and Benny Pinkas, "Securing Multiparty Computation for Privacy-Preserving Data Mining," *Journal of Privacy and Confidentiality*, 2009, 59-98, http://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=jpc.

Assessing a Research Framework for NITRD

ITIF applauds the NSF's approach to seeking comment on increasing investments for research and development in privacy-enhancing technologies and encouraging multi-disciplinary research. Through this research, the NSF can help unlock additional data-driven innovation by giving both the public and private sectors the tools they need to share data privately and securely.[23] By developing a roadmap for its research goals, NSF will be able to maximize the social and economic benefits of federal funds and coordinate research efforts across multiple federal agencies.

In addition, ITIF recommends the NSF begin to document progress based on this roadmap. To enable this information to be better aggregated, all federal funding agencies should start identifying the research they fund on privacy-related subjects and how it relates to this roadmap. This information could be used to gain a baseline understanding of what privacy-related research is already being funded, and in the future, it could be used to identify progress toward specific privacy-research objectives. Cross-agency monitoring would cut down on duplicative studies and allow agencies to identify the latest research on any particular subject. This would also allow researchers in the field to easily find funding opportunities, and developers to easily learn about new research to integrate into their products and services.

There are many areas where technology is rapidly changing that are ripe for additional privacy research. In particular, we recommend NITRD consider funding research in the following areas:

1. De-identification: This area of research explores techniques to remove personally identifiable information from data sets.[24]
2. Differential privacy: This area of research explores algorithms that can produce statistical information about a data set without compromising the privacy of the individuals represented in that data.[25]

[23] Daniel Castro, "The Need for an R&D Roadmap for Privacy," *The Information Technology and Innovation Foundation*, August 2012, 1, http://www2.itif.org/2012-privacy-roadmap.pdf.
[24] Ann Cavoukian and Daniel Castro, "Setting the Record Straight: De-identification Does Work," *The Information Technology and Innovation Foundation*, June 2014, http://www2.itif.org/2014-big-data-deidentification.pdf.

3. Self-enforcing data policies: This research looks at techniques that allow data sets to be released only under certain conditions, such as automatically expiring after a certain amount of time.[26]
4. Privacy-preserving data mining: This research explores data mining algorithms that can analyze large databases while not revealing private information.[27]
5. Usability and accessibility of privacy-enhancing technologies: This research analyzes how human factors affect the successful application of privacy-enhancing technologies.[28]
6. Interoperable digital credentials: This research examines how to make systems and organizations work together by establishing trust through digital credentials.[29]
7. Privacy metrics: This research explores standards of measuring and comparing the privacy of a specific activity or process, which may be useful evidence of compliance for legislative or regulatory requirements, as well as organizations' internal privacy policies.[30]

While the primary focus of NITRD should be on developing technical capabilities that are currently lacking, finding comprehensive solutions for these problems will require bringing together

---

[25] Cynthia Dwork, "Differential Privacy: A Survey of Results," *Microsoft Research*, 2008, http://research.microsoft.com/pubs/74339/dwork_tamc.pdf.
[26] Philippe Golle, Frank McSherry, and Ilya Mironov, "Data Collection With Self-Enforcing Privacy," *Microsoft Research*, 2006, https://crypto.stanford.edu/~pgolle/papers/selfprivacy.pdf.
[27] Charu C. Aggarwal and Philip S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," *Advances in Database Systems*, Vol. 34, 2008, 11-52, http://www.polyteknisk.dk/related_materials/9780387709918_chapter_1.pdf, and Yehuda Lindell and Benny Pinkas, "Securing Multiparty Computation for Privacy-Preserving Data Mining," *Journal of Privacy and Confidentiality*, 2009, 59-98, http://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=jpc..
[28] Alma Whitten and J.D. Tygar "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0," *In Proceedings of the 9th USENIX Security Symposium,* August 1999, http://www.cs.berkeley.edu/~tygar/papers/Why_Johnny_Cant_Encrypt/OReilly.pdf.
[29] Ting Yu, Marianne Winslett, and Kent Seamons, "Supporting Structured Credentials and Sensitive Policies through Interoperable Strategies for Automated Trust Negotiation," *Journal of ACM Transactions on Information and System Security*, Vol. 6, Issue 1, February 2003, 1-42, http://www4.ncsu.edu/~tyu/pubs/tissec03.pdf, and "Interoperable Identity Credentials for the Air Transport Industry," *Smart Card Alliance*, October 2008, http://www.smartcardalliance.org/resources/lib/Air_Transport_ID.pdf.
[30] Rasika Dayarathna, " Taxonomy for Information Privacy Metrics," *Journal of International Commercial Law and Technology*, Vol. 6, No. 4, 2011, http://www.jiclt.com/index.php/jiclt/article/view/139.

researchers from different disciplines outside of computer science, including industrial design, economics, behavioral sciences, public policy, and law.

Conclusion

Organizations, both public and private, will continue to collect massive amounts of data on individuals, and technology will continue to improve, opening many more opportunities for privacy violations. NSF should lead the effort to ensure our privacy protections keep up with this inevitable wave of innovation and not in spite of it. A roadmap for privacy research goals and objectives would help ensure that more federal research dollars are used effectively for our most pressing privacy challenges and offer government another tool to help protect consumer privacy.

Sincerely,

Robert D. Atkinson
President and Founder

Daniel Castro
Senior Policy Analyst

Alan McQuinn
Research Assistant

The Information Technology and Innovation Foundation