# Location Confidentiality and Official Surveys -- Executive Summary from the 2nd Workshop in the Census-MIT Big Data Workshop Series on Using New forms of Information for Official Statistics

***Mar 31, 2016***

*Micah Altman, MIT; Cavan Capps & Ronald Prevost, U.S. Census Bureau*[1]

## 1. Workshop Series Overview

**Trends and opportunities**. In today's increasingly data-driven economy, technology changes data uses, and stakeholder expectations increase, while statistical agency budgets and staffing remain flat. Big Data offers both tremendous promise and new challenges for official statistics. New forms and scales of data may offer opportunities to enhance and strengthen official statistics by improving estimates; lowering costs and helping agencies improve the frequency and timeliness of data releases. Achieving this promise requires innovative integration of methods from many disciplines and the expertise of many sectors.

**Workshop approach.** The workshops series brings together select groups of experts in universities, industry, and the U.S. government. Each of these three workshops focus on a different set of issues relates to big data -- potential sources; data privacy and security; and barriers to statistical inference. During the workshop, the experts are guided to *explore* the challenges involved in building the next generation of official statistics; *identify* new opportunities to use big data in Statistical Organizations, and synergistic work in the discipline; and to *examine* broader questions through their application to an exemplar use case.

## 2. Workshop Overview: Location Confidentiality and Official Surveys

**Workshop motivation.** Based on mobile devices alone, commercial entities have the potential to collect extensive, fine grained, continuous, and identifiable records of a person's location and movement history, accompanied with a partial record of other mobile devices (potentially linked to people) encountered over that history. This information is increasingly used for commercial purposes, such as targeted advertising, and for scientific research. In contrast, large-scale

[1] Authors are listed in alphabetical order. We describe contributions to the paper using a standard taxonomy. (Allen, Liz, et al. "Credit where credit is due." *Nature* 508.7496 (2014): 312-313.) All authors take equal responsibility for the article in its current form. MA and CC authored early versions of the manuscript; all authors contributed to review and revision; and in the conception of the article (including core ideas, analytical framework, and statement of research questions). All authors contributed to the project administration and to the writing process through direct writing, critical review, and commentary.

surveys, such as those conducted by federal Statistical Organizations, have yet to incorporate fine-grained locational data into data collection, or to augment address-based sampling with to person-based sampling methods. At the same time, protecting individual privacy is a central value of the US Statistical Organizations, and protecting fine-grained locational information is particularly difficult both because human mobility patterns are highly predictable and that these patterns have unique signatures, making them highly linkable even in the absence of associated identifiers.[2]

In general, the growth of big data sources have changed the threat landscape of privacy and statistics in at least three major ways. First, when surveys were initially founded as the principal source of statistical information, whether one participated in a survey was largely unknown. Now, as government record systems and corporate big data sources are increasingly used that include all or a large portion of a given universe, that privacy protection is eroded. Second, in the past, little outside information was generally available to match with published summaries. Now the ubiquity of auxiliary information enables many more inferences from summary data. Third, in the past, typical privacy attacks relied on linking outside data through well-known public characteristics -- PII or BII. Now, datasets can be linked through behavioral fingerprints.

**Workshop use case**. This workshop, the second in the series, asked the question of how fine-grained data collected from personal mobile devices might be used to augment the census surveys, and what new approaches to privacy protection might be needed. The workshop engaged these issues through an examination of a hypothetical use case – how might census products be augmented with information collected through the *Google Now* service.

“Google Now” is an innovative product developed by Google to function as an intelligent personal assistant – and was named innovation of the year by Popular Science in 2012. Google Now’s primary innovation is that it pro-actively makes recommendations to individuals based on their location history, web search history, email activities, travelling patterns, and other individual behavior.[3] It is both an example of how much information third-parties can effectively gather about individuals, and how useful that data can be in predicting (and providing assistance with) individual behavior.

**Workshop participants.** Twenty experts participated in the workshop. These experts were drawn from senior leadership in the privacy field, industry associations, federal government, and national experts in academia.

Discussion was conducted under Chatham-house rules, which restricts attribution of individuals, and individual statement without prior explicit approval. A number of participants agreed to identify themselves for the purposes of this public summary:

- John Abowd; Edmund Ezra Day Professor of Economics; Cornell University

[2] See Gonzalez, Marta C., Cesar A. Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns." Nature 453.7196 (2008): 779-782.
[3] CITE

- Micah Altman; Head/Scientist, Program on Information Science; Massachusetts Institute of Technology (Facilitator)
- Robin Bachman, Chief, Policy Coordination Office at U.S. Census Bureau
- Elizabeth Bruce; Executive Director, Institute for Data, Systems and Society; Massachusetts Institute of Technology
- Cavan Capps, Big Data Lead, U.S. Census Bureau (Workshop Organizer)
- Jennifer Childs; Research Psychologist; Center for Survey Measurement, Research and Methodology Directorate; U.S. Census Bureau
- Aref Dajani; Mathematical Statistician; U.S. Census Bureau
- Yves-Alexandre de Montjoye; Postdoctoral Fellow; Harvard University; School of Engineering and Applied Sciences
- Benjamin Fung; Associate Professor of Information Studies; McGill University
- Ronald Jarmin; Deputy Associate Dir. for Research & Methodology; U.S. Census Bureau
- Christa D. Jones; Deputy Chief; Congressional and Intergovernmental Affairs; US Census Bureau
- Kobbi Nissim; Professor of Computer Science; Ben-Gurion University and Senior Research Fellow; Harvard University; Center for Research on Computation & Society
- Amy O'Hara; Chief, Center for Administrative Records Research & Applications
- Alex "Sandy" Pentland; Toshiba Professor of Media, Arts, and Sciences
- Massachusetts Institute of Technology
- Satyam Priyardarshy; Chief Data Scientist; Haliburton
- Ronald Prevost; Senior Statistician; Research & Methodology; U.S. Census Bureau (Workshop Organizer)
- Stephanie Shipp; Deputy Director and Research Professor; Social and Decision Analytics Laboratory
- Vitaly Shmatikov; Professor of Computer Science; Cornell Tech.
- Ryan T Wright; Associate Professor Operations & Information Management; Isenberg School of Business; University of Massachusetts, Amherst
- Alexandra Wood; Fellow; Berkman Center for Internet and Society

**Workshop focus questions**. Participants in the workshop were asked to focus on a number of specific questions. These included:

- How do different methods of collecting location information affect individual privacy (control over means and manner of disclosure)?
- What are the confidentiality risks to individuals associated with location information? What are common attitudes towards sharing location information? What approaches are being used to make data collection and use transparent?
- What are key measures of data utility and what are the key tradeoffs between confidentiality and utility?
- How can the risks of reidentification from geographic information be characterized? What is the state of the art and practice for limiting disclosure when disseminating data products that are based on location information?

- In cases where open access to collected geo-location information creates substantial privacy risks, what other methods of access are feasible? What might be done to facilitate wider analysis of protected geo-location information? What are the limitations on data utility created by restrictions on access?
- What laws, regulations, policies, and principles are most relevant to geo-location data? How do organizations make the decisions on what geo-location data to collect and use?

## 3. **Workshop Discussion**

Discussion among workshop participants highlighted both the opportunities for using big data on persons and business and the challenges such uses pose. Throughout the discussion a number of recurring themes were emphasized.

### ***Mission of the Census is to create data for decision making***

Participants noted that "The Census Bureau's *mission* is to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly."

Public data release is critical to the Census Bureau mission, and releases must preserve privacy and confidentiality. Big data threats to privacy must not be allowed to threaten the release of quality data for decision support, and continuing support for research is needed.

### ***Opportunities for Incorporating Mobile Data***

Participants first drew attention to "low hanging fruit" – opportunities to realize immediate and substantial benefit from integrating locational information. In general, participants suggested that using existing data to reducing survey administrative costs was likely to be tractable, and yield significant cost savings.

Participants noted three potential applications in which survey costs are high, and external data is available, if privacy concerns can be addressed:

- Location data from mobile phones could be used to adapt survey implementation, and to target survey contacts. For example, location data could be used to assist in the identification of occupied houses, and to determine when occupants are most likely to be available for contact. In large data collections efforts, such as the decennial Census and the American Community Survey reducing the number of contacts required would have a large financial impact.[4]

[4] S Konicki, T. Adams. 2015. Adaptive Design Research for the 2020 Census
, Working Paper. Presented at Join Statistical Meeting 2015. https://www.census.gov/content/dam/Census/library/working-papers/2015/dec/DSSD-WP2015-02.pdf Schouten, Barry, Melania Calinescu, and Annemieke Luiten. "Optimizing quality of response through adaptive survey designs." *Survey Methodology*39.1 (2013): 29-58.

- External data could be used to replace some measures that respondents consider most intrusive. For example, respondents to the American Community Survey have reacted most strongly to questions on plumbing, commuting, income, and disability.[5] Of those, measures of commuting can be constructed through mobile phone location data; plumbing measures are available through external data sources aggregated by third parties, such as Zillow; and income measures are available through IRS data. Use of these sources could be used to replace or pre-fill these measures, reducing respondent burden and intrusion.
- External data could also be used to reduce respondent burden, while increase the granularity and timeliness of existing surveys. For example, the monthly survey of retail trade typically has a low response rate[6] – however sales for many of the respondents could be estimated through credit-card and debit-card transactions.

However, participants also noted that this low hanging fruit taps only a small part of the potential use of this data. Current computational social science research is using data in new and exciting ways, developing inferential possibilities that will exemplify new perspectives and possibilities in social science research and policy oriented decision making.

And that the exciting work exemplified by computational social science is using data in new and of inferential possibilities that computational social science research exemplifies.

## 4. Challenges for Privacy and Confidentiality

Historically, national statistical organizations have managed learning ("identification") risks by limited and/or masking data releases such that the likelihood of a certain identification based on that release is either zero or minimal. (In more formal terms, these organizations reduced the probability of deterministic record-linkage to the population, based on specified quasi-identifiers.) This *threshold-based approach* is reflected both in organizational practice; and in the accompanying legal frameworks' focus on preventing identification, and the sharp distinction it makes between personally identifiable information and other (e.g., "anonymized") information.

Participants noted that academic research in cryptography theory, statistics, and information science has demonstrated a number of deep challenges related to this approach, and to protecting privacy in the modern world.

The first challenge, as demonstrated by research led by Pentland and de Montjoye, is that many human behaviors leave behind distinct behavioral fingerprints in the data -- even in the complete absence of traditional identifiers (or quasi-identifiers). For example, de Montjoye and Pentland demonstrated that most individuals can be reidentified from datasets containing location history if an attacker can obtain just a small number of external observations (auxiliary information) of

[5] D. Cohn, "Census may change some questions after pushback from public**",** Factank. April 22, 2014 <http://www.pewresearch.org/fact-tank/2014/04/22/census-may-change-some-questions-after-pushback-from-public/>

[6] See "Monthly Retail Trade Survey Methodology" <https://www.census.gov/retail/mrts/how_surveys_are_collected.html#nonsampling_error>

location information for that person.[7] Further, using aggregation to prevent such reidentification is ineffective -- because the aggregation needed (at the observation level) scales geometrically with the number of external points available to the attacker. Nor is behavioral fingerprinting unique to location information -- for example, people have been successfully reidentified using characteristics of their writing style, typical credit card purchases, and movie rankings.[8] This creates a problem for many traditional statistical disclosure limitation methods, which are based on an analysis of designated characteristics of a subject (quasi-identifiers) and do not consider the potential for reidentification arising from independent observation of the behaviors that the survey is designed to measure.

A second challenge is that when data is released that is protected by traditional statistical disclosure control methods privacy risks to individuals from that data release continue to grow in the future as new external data is released. This is because traditional methods whatever modification they make to the data (e.g. swapping, aggregation, suppression, or generalization, topcoding) are designed to remove specific identifiers, prevent deterministic record linkage, or reduce the probability of a complete reidentification.[9] While these methods may be sufficient for controlling what can be learned about an individual from a specific data set – modern privacy research shows that such approaches can not provide any strict bounds on the amount that can be learned from composing with independent auxiliary information.[10] Furthermore, many traditional methods, such as local suppression, and topcoding, cause subsequent estimates of population parameters and model estimates to be systematically biased.[11]

A third challenge revealed by modern privacy research is that every release of data, if it has any utility, no matter how it is protected, inevitably leaks some private information, and this leakage

[7] de Montjoye, Yves-Alexandre, et al. "Unique in the Crowd: The privacy bounds of human mobility." *Scientific reports* 3 (2013).

[8] Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA law review. 2010 Aug 13;57:1701.; Narayanan A, Shmatikov V. Myths and fallacies of personally identifiable information. Communications of the ACM. 2010 Jun 1;53(6):24-6.; de Montjoye, Yves-Alexandre, Laura Radaelli, and Vivek Kumar Singh. "Unique in the shopping mall: On the reidentifiability of credit card metadata."*Science* 347.6221 (2015): 536-539.

[9] Willenborg L, De Waal T. Elements of statistical disclosure control. Springer Science & Business Media; 2001.

[10] Dwork, C., 2006. Differential privacy. In *Automata, languages and programming* (pp. 1-12). Springer Berlin Heidelberg.

[11] See. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E.S., Seri, G. and Wolf, P., 2010. Handbook on statistical disclosure control. *ESSnet.* for comments on the estimation properties of many traditional techniques. On top-coding bias in particular see also Rigobon, R. and Stoker, T.M., 2009. Bias from censored regressors. *Journal of Business & Economic Statistics*, *27*(3), pp.340-353; and Crimi, N. and Eddy, W., 2014. Top-Coding and Public Use Microdata Samples from the US Census Bureau. *Journal of Privacy and Confidentiality*,*6*(2), p.2.

increases with each release.[12] In other words – there is no free lunch with respect to information privacy, you always have to buy it with utility.

Thus a modern approach to information privacy incorporates an explicit analysis of privacy loss versus information utility across all releases. In contrast, traditional statistical disclosure control, and traditional legal approaches to privacy protection, both assume that privacy risk can be effectively eliminated (e.g. by preventing direct identifications) and analyze each data release in isolation.

Fourth, while approaches to privacy provide a formal measure of learning risk, the potential harm to participants that may result from others learning their private information is highly dependent on the types of information collected, and the potential contexts in which it might be used. In order to design data releases that provide a good tradeoff between social data benefits and the potential individual costs requires a systematic evaluation of the informational harms that could occur from unauthorized use.

These challenges have a number of implications within the context of data collection and dissemination by national Statistical Organizations. In the next section, we discuss the implied concerns and promising approaches to mitigate them.

## 5. Promising Approaches

The threshold-based approach to managing learning risks is no longer reliable – because of advances in statistical and computational methods, combined with increased availability of data (as described above). In other words, incremental learning risks can no longer be ignored -- they must be managed.

A modern, scientific approach to managing the privacy risks that flow data releases requires a systematic evaluation of the potential harms and benefits flowing from each release. In particular organizations should evaluate information leakage -- what third parties can learn about individuals from the release because of individuals' participation in the data collection; the potential harms that could arise to the individuals from such learning events; and the utility of the information made available for its intended audiences.[13]

*With respect to managing inferential leakage, or "learning risks" -- participants recommended exploring "differential privacy" as an approach to measuring and controlling learning risks.* Differential privacy is a strong, quantitative notion of learning risk that is provably resilient to a

[12] Dwork, C., 2006. Differential privacy. In *Automata, languages and programming* (pp. 1-12). Springer Berlin Heidelberg.

[13] Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law. Forthcoming. < http://informatics.mit.edu/publications/towards-modern-approach-privacy-aware-government-data-releases>

very large class of potential misuses.[14] As a robust privacy framework that addresses both known and unforeseeable attacks, differential privacy represents a solution that moves beyond the penetrate-and-patch approach that is characteristic of traditional de-identification approaches. Furthermore, the differential privacy approach is provably robust to all auxiliary information – and thus protects against inferential disclosures even in the presence of a "mosaic" of additional information. Moreover, at present, it is the only framework in practical use that provides general formal bounds on inferential disclosure in the presence of mosaic effects (i.e. all other frameworks in current practice make strong assumptions on the limits of third-party auxiliary knowledge).

In contrast to evaluation of learning risks, no systematic framework currently exists for National Statistical Organizations that wish to systematically evaluate the informational harm that potentially arise from private information leakage.  Although some populations are deemed "sensitive" within the census organization, and subject to heightened scrutiny – both the criteria for constituting a sensitive population, and the analysis of harms to those populations are ad-hoc. Further, although most harms stem from the release of new information, NSO's routinely allocate effort to protect information about individuals that is already widely publicly available (such as most people's residential addresses). A more systematic analysis of harm from information leakage is necessary whether the leakage was the result of an unanticipated security breach or the inevitable small but incremental leakage of personal information that results from publishing aggregate results

Further, the discussion highlighted that much of the anticipated potential arising from information leakage was harm to the reputation of the census as an institution. This harm is particularly important because it affects individual's willingness to participate in data collection and the honesty of their responses; and this has a direct effect on the reliability of information products; and on the cost and ultimate utility of those estimates. However, information leakage is not a reliable proxy for this type of harm, because the effect that such leakage has on institutional reputation depends on individual's perception and awareness of the information leakage; their attitudes and values; and on the larger context of the information leakage (e.g. whether it is necessary to promote a recognized value); and on the framing of the interactions between participants and agencies.[15] *Participants recommended that the census engage with social scientists to conduct applied research into how individuals recognize, and are affected by privacy harm.*

[14] Dwork C. 2011, "A firm foundation for private data analysis**"**, *Communications of the ACM* v54

[15] For one evaluation of reputation effects see: Acquisti, A., Friedman, A. and Telang, R., 2006. Is there a cost to privacy breaches? An event study. *ICIS 2006 Proceedings*, p.94.
. For a more general review of many of these effects see Acquisti, A., 2012. Nudging privacy: The behavioral economics of personal information. *Digital Enlightenment Yearbook 2012*, pp.193-197.

When statistical agencies collect information about organizations rather than individuals, the types of risks and harms differ. First, when data releases or summaries leak information about an identified commercial firm, the primary harm is more likely to be loss of some competitive advantage. Second, where some industries are dominated by a single or small number of firms, it is impossible to provide useful aggregate statistics unless one leaks more than minimal amounts of information about those dominant firms: And this utility-privacy tradeoff may be justified because of the importance of the aggregate measurements for policy.

*Participants recommended that NSO's explore new methods of collecting and disseminating data be explored, such as secure multiparty computation*. The current modal practice of statistical organizations is to gather all relevant data centrally to link and analyze it, and then to release it in the form of summary statistics and, less frequently, in redacted and aggregate "public use" form. This practice is increasing inconsistent with the distributed and dynamic nature of data sources, and with the diversity of stakeholder needs for access.

Many potential sources of external data are commercial. Many of these sources are not able to simply export all relevant data to an NSO – either because of legal restrictions or technical limitations. Many other commercial sources are unwilling to provide this type of comprehensive direct access. In contrast, companies are increasingly using distributed, API-based methods to compute over their own and their partners' data.  even their own internal data, and using API's

Private API's enable a data provider to limit access to portions of data, to meter access, and to monitor. Secure multiparty computation goes beyond a private API to control and audit exactly what can be computed across a distributed set of data sources. Secure multi-party computations thus enables data to be shared for particular calculations and analysis, without the need to trust others with your data – only the final results of the agreed-upon computations are shared in unencrypted form.[16] Developing the capability to use secure multiparty computation to generate the analysis needed for NSO's would substantially reduce the "trust barriers" to incorporating commercial data.

The federal open data policy officially promotes making data available to the broadest public for decision making.[17]  Historically, public use micro-data files and aggregated data have been trusted in legal cases, in some cases by low-income non-profits. Future public use data files may rely upon using differential privacy techniques to create synthetic data sets. In order to insure that public synthetic data releases support the needed legal trust, research into the development of systematic quality checking and perhaps quality metrics for synthetic datasets need to be developed and formalized.

Finally, participants observed that no single technology or approach will cover all uses of data. Data from NSO's is used by a range of stakeholders whose needs range from quick lookups, to the estimation of highly complex statistical models applied to data linked from many sources.

[16] See for a review of SMC for data mining, Lindell, Yehuda, and Benny Pinkas. "Secure multiparty computation for privacy-preserving data mining." Journal of Privacy and Confidentiality 1.1 (2009): 5.

[17] See "Open Data Policy-Managing Information as an Asset", 2013, OMB < https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf >

Further, not only do stakeholders require the results of the calculations – they may also require assurance, or evidence of authenticity, transparency, provenance, and reproducibility. Although methods such as differential privacy and secure multiparty computation can address a range of these requirements – but analyses using the cutting edge statistical methods, integrating novel data source, or demonstrating the highest levels of transparency will require direct access to the collected data. Thus NSO's will always need to make available to some stakeholders secure research enclaves, and other tiers of access to data.

## 6. Recommendations

### *Next Steps*

The workshop participants reached a consensus on three next steps towards incorporating a modern approach to data privacy.

First, participants recommended developing a pilot project to actively collaborate with external researchers in applying new privacy research to census data publications. These collaborations are vital to translate state of the art research into robust methods; adapt these methods so that they can be applied to census surveys; and to promote technology transfer and build capacity within the Census organization.

Second, the bureau should engage external expertise by publicizing both the key capabilities that the organization possesses. Because of its resources, reputation, and enabling law, the Census is uniquely positioned to function as a hub/integrator for statistical analysis across the Federal government; to integrate computations over commercial data; and to develop a network of expert

Third, the bureau should publicize and support research and pilot projects in several key areas:

- *Formal approaches to privacy and secure computation.* Methods such as secure multiparty computation and differential privacy are able to provide strict guarantees on (respectively) the types of analysis that can be performed on data, and on the privacy of the individual confidentiality leaked from any inferences made on that analysis. These methods are being applied now – but in a limited way, using largely bespoke systems, to provide protected access to limited query results.[18] In contrast, the Bureau now supports a richer array of accessing the information it collects, including: as summary statistics; generalized contingency tables; simplified public use files; and, for vetted researchers, fully integrated databases that are open to analysis with state of the art computational methods. Research and development is needed to adapt formal methods to support these multiple modes of access, and incorporating them into robust off-the-shelf software.

[18] For a survey of static data publishing approaches, including some differentially privacy approaches see: Fung, et. al. 2010, *Introduction to Privacy-Preserving Data Publishing*, CRC Press. For a recent application to big data see: Chen, R., Fung, B.C., Philip, S.Y. and Desai, B.C., 2014. Correlated network data publication via differential privacy. The VLDB Journal, 23(4), pp.653-676.

- *Measuring informational harms*. Modern approaches to privacy provide a formal measure of learning risk, but not of the potential harm to participants – which is highly dependent on context. The harms that were identified as most relevant to the Bureau's analysis of data releases are organizational harms to the bureau's reputation (which in turn affects both the cost and value of the data product), and information harms to vulnerable groups. Neither category of harm is well characterized, either within the bureau, or in external research[19]: Empirical work in social science, particularly sociology, economics, and psychology would enable the bureau to move from ad-hoc restrictions on data release toward a systematic risk-benefit balance.
- *Public education and understanding.* There is an increasing recognition that in the age of big data, individuals lack meaningful understanding of how their data is used, and what risks this generates; and at the same time, that these the complexity of data risks makes standard approaches to notice and consent unworkably burdensome.[20] Research in education, psychology, and behavioral economics is needed to support meaningful notice and real understanding. A potential benefit of this is that the reputational effects of data releases will be better calibrated to the actual impact on individuals, rather than its perceived impact.

### *Strategies for National Statistical Organizations*

The first report in this series[21] observed that broad new sources of information have the potential to bring increased granularity, detail, and timeliness to the next generation of official statistics -- while reducing survey burden. The next generation of official statistics will utilize broad sources of information, potentially linked together, to provide increasing granularity, detail, and timeliness, while reducing cost and burden. The report observed also that incorporating big data into Statistical Organizations will require agencies to broaden the current narrow focus on data collection to include a more general focus on information provisioning; and that utilizing big data requires creating new relationships with data providers. Many of the primary sources of big data are businesses that use data intensively to guide decisions: Big data sources are also critical

[19] See e.g Altman M, Wood A, O'Brien D, Vadhan S, Gasser U. Towards a Modern Approach to Privacy-Aware Government Data Releases. Berkeley Journal of Technology Law. Forthcoming.
< http://informatics.mit.edu/publications/towards-modern-approach-privacy-aware-government-data-releases>
; Vayena E, Gasser U, Wood A, O'Brien D, Altman M. Towards a New Ethical and Regulatory Framework for Big Data Research, in Beyond IRBs: Ethical Review Processes for Big Data Research. Washington & Lee Law Review; Forthcoming. <
http://informatics.mit.edu/publications/towards-new-ethical-and-regulatory-framework-big-data-research>
Export

[20] See, e.g. McDonald, A.M. and Cranor, L.F., 2008. Cost of reading privacy policies, the. ISJLP, 4, p.543.

[21] Altman, Micah and Capps, Cavan and Prevost, Ronald C, Using New Forms of Information for Official Economic Statistics -- Examining the Commodity Flow Survey: Executive Summary from the 1rst Workshop in the Census-MIT Big Data Workshop Series (October 1, 2015). Available at SSRN: http://ssrn.com/abstract=2670209 or http://dx.doi.org/10.2139/ssrn.2670209

stakeholders. In order to obtain access to data from businesses that create and use it, it is critical to both provide a value proposition to the business and to develop a trust relationship.

Implementing modern methods for privacy evaluation and protection could help to assure potential data sources that the information they provide will be used only as intended. Further, much of this data will need to remain under the control of the provider. So increasingly, agencies will need to develop the capacity to compute over distributed data – rather than acquiring and linking all data in-house.

This second workshop highlights the need for national Statistical Organizations to understand and strengthen trust relations with the *participants* in data collection. Trust is critical because it affects participation in data collection, which in turns drives costs, affects sample representativeness, and contributes to the reliability of the statistical products produced – and this reliability is the hallmark of NSO value.

Whether or not NSOs use big data directly, the increasing availability of big data from many sources creates challenges to participant trust. As data about individuals accumulates, and becomes available as "auxiliary information" to third parties, these parties are increasingly able to make additional inferences about individuals from "small" data releases – even if these releases are deidentified using traditional methods. This is because traditional methods of deidentification (and other traditional statistical disclosure limitation methods), while they protect against what can be learned from a specific release, do not systematically address the combination (or "composition") of releases. Thus, as more data is available in the world, individuals' ability to trust in the privacy of any incremental data release decreases – unless that release is protected using formal methods.

Modern research in privacy demonstrates that all useful releases of data involve a measurable loss of privacy – which must be balanced against the utility of the information learned. This implies that the core issue of how NSOs will address privacy does not depend primarily on how NSOs choose to use big data. Instead, the widespread existence of big data highlights the need for NSOs to modernize the approach they take in both assessing privacy and utility across their entire range of data releases.

The current state of the practice in privacy lags well behind the state of the art in this area. Most commercial organizations, and most NSOs in other countries continue to rely (at most) on traditional aggregation and suppression methods to protect privacy – with no formal analysis of privacy loss or of the utility of the information gathered.  The U.S. Census Bureau, because of its size, institutional capacity, and strong reputation for privacy protection could establish leadership in modernizing privacy practices.

What is required to accomplish this is to invest in the development of new expertise in areas such as differential privacy and secure multiparty computation. Much of this expertise cannot simply be purchased or hired, but must be acquired through collaboration -- since much is highly specialized and exists primarily in research universities.

In sum, the increasing accumulation of data about people outside Statistical Organizations highlights the need to modernize agency practice to systematically address both the privacy and utility of data. To do this will require moving from a model where all data is centrally owned and all expertise is located within the agencies to a distributed model of expertise and information.