# Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy


Brian A. Harris-Kojetin, Deputy Director,
Committee on National Statistics
Washington, DC • June 5, 2017

# Acknowledgements

Funding for the panel was provided by

The Laura and John Arnold Foundation,

with additional support from the National Academy of Sciences Kellogg Fund.



Committee on National Statistics

The National Academies of SCIENCES · ENGINEERING · MEDICINE

# Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods

- **Robert M. Groves,** (Chair), Georgetown University
- **Michael E. Chernew,** Harvard University
- **Piet Daas,** Statistics Netherlands
- **Cynthia Dwork**, Harvard University
- **Ophir Frieder,** Georgetown University
- **Hosagrahar V. Jagadish**, University of Michigan
- **Frauke Kreuter**, University of Maryland
- **Sharon Lohr**, Westat, Inc.
- **James P. Lynch**, University of Maryland
- **Colm O'Muircheartaigh**, University of Chicago
- **Trivellore Raghunathan**, University of Michigan
- **Roberto Rigobon**, Massachusetts Institute of Technology
- **Marc Rotenberg,** Electronic Privacy Information Center



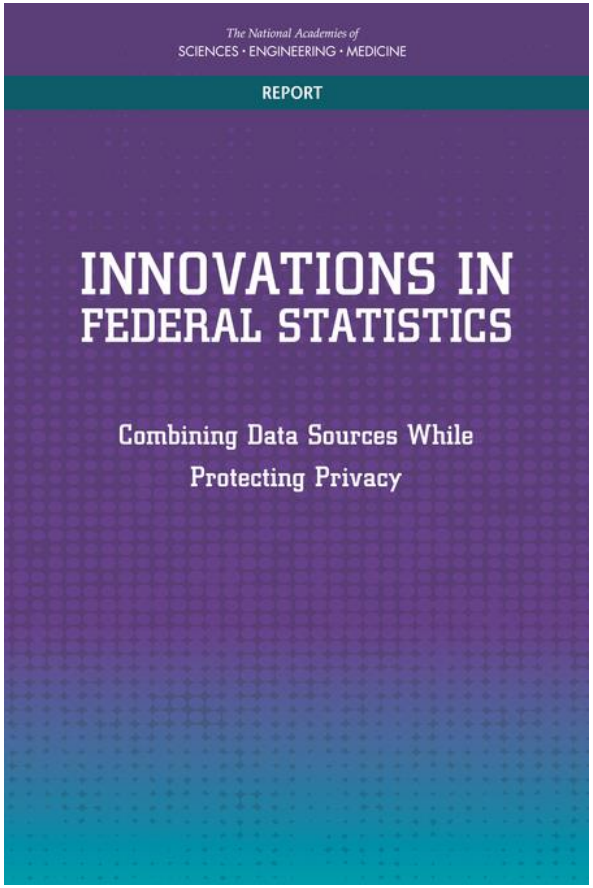
# Statement of Task

An ad hoc panel of nationally renowned experts in social science research, computing technology, statistical methods, privacy, and use of alternative data sources in the United States and abroad will conduct a study with the goal of fostering a paradigm shift in federal statistical programs. **In place of the current paradigm of providing users with the output from a single census, survey, or administrative records source, a new paradigm would use combinations of diverse data sources from government and private sector sources combined with state-of-the art methods to give users richer and more reliable statistics** leading to new insights about policy and socioeconomic behavior. The motivation for the study stems from the increasing challenges to the current paradigm, such as declining response rates and increasing cost and burden for surveys. **The panel will prepare two reports as part of this study.**



Committee on National Statistics
The National Academies of SCIENCES · ENGINEERING · MEDICINE

# First Report Released in January




**Available for free download at [www.nap.edu](http://www.nap.edu)**

**Second Report will be released in Summer 2017**

Committee on National Statistics
The National Academies of
SCIENCES · ENGINEERING · MEDICINE

# Contents

# Current Challenges and Opportunities in Federal Statistics

**Conclusion 2-3**: The way that statistics are currently produced by Federal statistical agencies faces threats from declining participation rates and increasing costs.

- Although generally higher than other surveys, federal statistical surveys face increasing nonresponse and increased costs of data collection to maintain response rates
- Agency budgets have decreased or remained flat
- Agencies face increasing demands for more timely and more geographically detailed information
- Increasingly alternative data sources are available that offer the potential of faster and more detailed information

# Using Government Administrative and Other Data for Federal Statistics

- **Conclusion 3-1** Administrative records have demonstrated potential to enhance the quality, scope, and cost efficiency of statistical products.
- **Conclusion 3-2** The use of administrative data can reduce the burden on survey respondents by supplementing or replacing survey items or entire surveys.

# Using Government Administrative and Other Data for Federal Statistics

- **Recommendation 3-1** Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits and risks of using administrative data. To this end, federal statistical agencies should create collaborative research programs to address the many challenges in using administrative data for federal statistics.

# Barriers to Use of Administrative Records

- **Conclusion 3-4** Legal and administrative barriers limit statistical use of administrative datasets by federal statistical agencies.
- **Conclusion 3-5** State and local governments may respond to incentives from the federal government to provide access to their administrative data by federal statistical agencies for statistical purposes.
- **Conclusion 3-7** Not enough is yet known about the fitness for use of administrative data in federal statistics. Coverage, missing information, lack of consistency, and continued availability present challenges with their use.

# Using Private Sector Data for Federal Statistics

- **Conclusion 4-1** Enormous amounts of private-sector data that are being generated every day have the potential to improve the timeliness and detail of national statistics.
- **Conclusion 4-2** The data from private-sector sources vary in their fitness for use in national statistics. Systematic research is necessary to evaluate the quality, stability, and reliability of data from each of these alternative data sources currently held by private entities for its intended use.

# Using Private Sector Data for Federal Statistics

- **Recommendation 4-1** Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits of using private-sector data sources.
- **Recommendation 4-2** The Federal Interagency Council on Statistical Policy should urge the study of private-sector data and evaluate both their potential to enhance the quality of statistical products and the risks of their use. Federal statistical agencies should provide annual public reports of these activities.

# Combining Multiple Data Sources

- **Conclusion 3-8** Combining multiple datasets allows for expansion of the number of attributes on units and thus can improve federal statistics, including the capacity to perform multivariate analysis for policy and evaluation studies.

# Combining Multiple Data Sources

- **Conclusion 3-9** There are statistical methods and models for combining information from multiple data sources using a variety of techniques.
- **Conclusion 3-10** Dealing with multiple data sources is more complex that dealing with a single dataset. A framework is needed to identify the error structure of each source and assess the utility of combining different data sources given their strengths and weaknesses.

# Current Context for Privacy

- **Conclusion 5-2** Combining multiple data sources increases risks to the public from data breaches and identity theft.
- **Conclusion 5-3** Privacy laws have established clear limitations on the collection and use of personally identifiable information for statistical purposes. There are also limits on the use of identifiers, such as Social Security numbers, that enable the linkage of distinct record systems. These laws reflect concerns about the use of personal data gathered by federal agencies.

# Challenges for Agencies

- **Conclusion 5-4** Federal Statistical agencies have a strong tradition of confidentiality and data stewardship. There are growing threats to data repositories and personal privacy that need to be addressed to support this tradition.
- **Conclusion 5-5** A continuing challenge for federal statistical agencies is to produce statistical products that safeguard privacy. This challenge is increased by the use of multiple data sources.

# Protecting Privacy and Confidentiality While Providing Access to Data

- **Recommendation 5-1** Statistical agencies should engage in collaborative research with academia and industry to continuously develop new techniques to address potential breaches of the confidentiality of their data.
- **Recommendation 5-2** Federal statistical agencies should adopt modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.

# Use of Computer Science and Cryptography to Protect Privacy

- Even if data breaches and security are solved, there still remain threats to privacy that come from the desired outputs of statistical data analysis
- There are fundamental mathematical limits on “how much” can be computed while maintaining any reasonable notion of privacy
- There are mathematical and algorithmic tools to formally quantify and control privacy loss

# Use of Computer Science and Cryptography to Protect Privacy

- **Conclusion 5-6** As federal statistical agencies move forward with linking multiple data sets, they must simultaneously address quantifying and controlling the risk of privacy loss.
- **Conclusion 5-7** Privacy-enhancing techniques and privacy-preserving statistical data analysis can potentially enable the use of private-sector data sources for federal statistics.

# Second Report

The panel will discuss these issues further in its second report, forthcoming Summer, 2017.

Additional topics to be discussed in Report 2:

- concepts, metrics, and methods for assessing the quality and utility of alternative data sources;
- statistical models for combining data from multiple sources and evaluate the quality of combined-information estimates;
- information technology infrastructure;
- additional privacy and confidentiality issues; and
- priorities for research needed for federal statistical agencies to advance a multiple-data-sources paradigm.

# THANK YOU!


For further information contact:
Brian Harris-Kojetin (bkojetin@nas.edu)