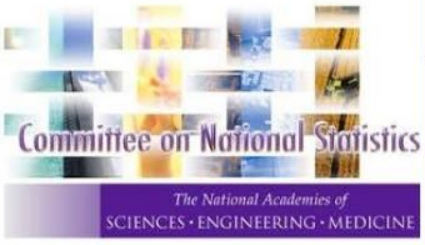
National Statistical Programs: Directions and Challenges

Brian Harris-Kojetin, Ph.D.
Director, CNSTAT
Washington, DC • August 9, 2018

The Committee on National Statistics

- Established in 1972 as a standing unit of the National Academies on the recommendation of the President's Commission on Federal Statistics to provide an independent, objective resource for evaluation and improvement of federal statistical methods and operations.
- There are about 50 standing units like CNSTAT in the Academies.



What is the Committee on National Statistics?

- CNSTAT's mission is to improve the statistical methods and information on which public policy decisions are based. It also serves as a coordinating force in the highly decentralized U.S. federal statistical system.
- Over its 46-year history, CNSTAT has produced over 270 consensus, interim, letter, and workshop reports.
- Every October and May, CNSTAT holds a public seminar on a topic of broad interest to the federal statistical and research communities and has a substantive luncheon with heads of major statistical agencies.

Who Serves on CNSTAT?

Robert Groves (*chair*), Math/Statistics

Mary Ellen Bock, Math/Statistics

Anne C. Case, Health Economics

Michael Chernew, Health Economics

Janet Currie, Welfare Economics

Don Dillman, Survey Research

Tom Mesenbourg, Sr. Federal Statistics Management

Sarah Nusser, Survey Research

Colm O'Muircheartaigh, Survey Research

Jerome Reiter, Applied Statistics

Roberto Rigobon, Economic Statistics, Big Data

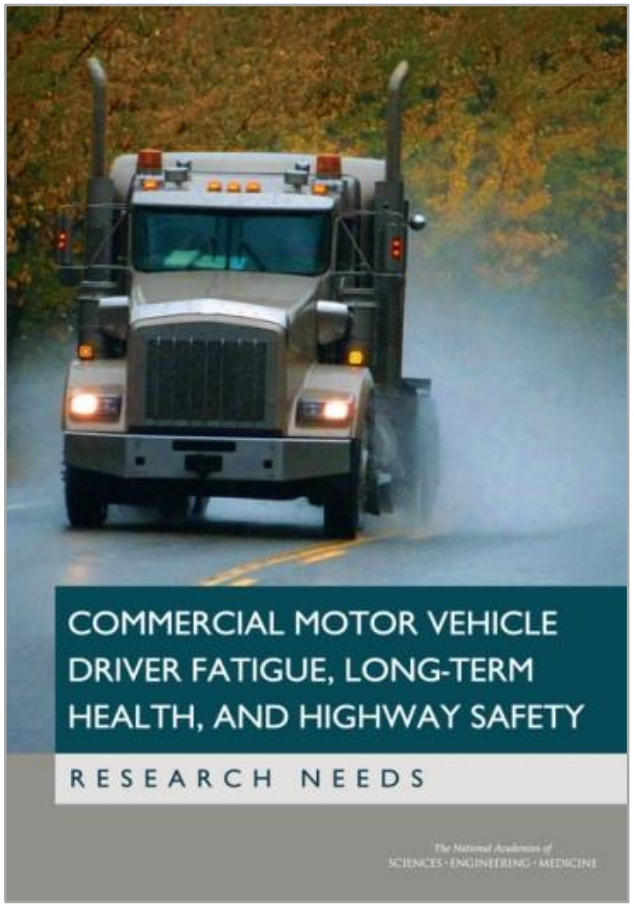
Judith Seltzer, Sociology

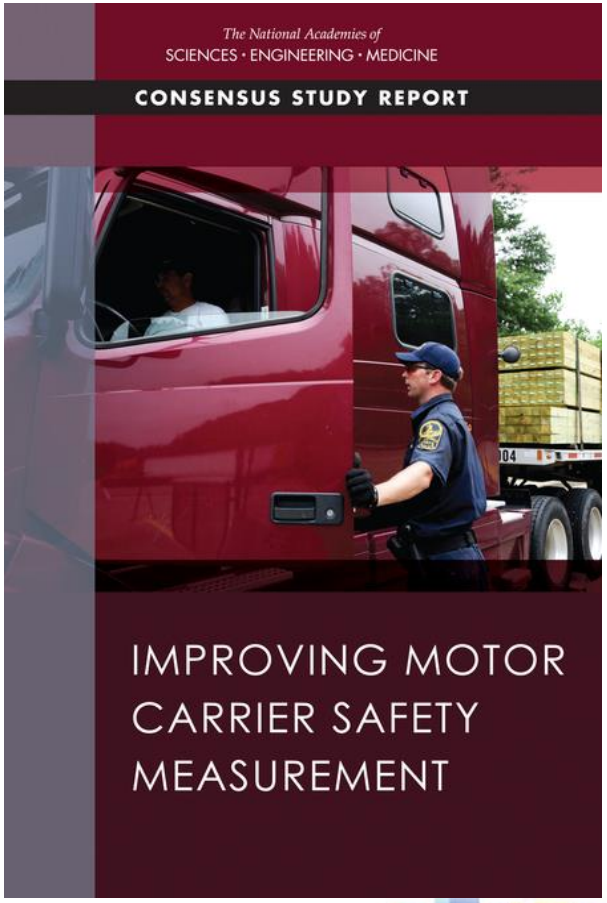
Note: all members serve *pro bono*

The CNSTAT Portfolio

- Coordinating and Sustaining Federal Statistics
- Decennial Census Coverage and Quality and the American Community Survey
- Economic Measurement
- Federal Household and Business Surveys
- Health and Social Welfare
- Science, Technology, and Innovation Indicators
- Statistical Methods and Estimates for Policy Use

Recent Projects with TRB





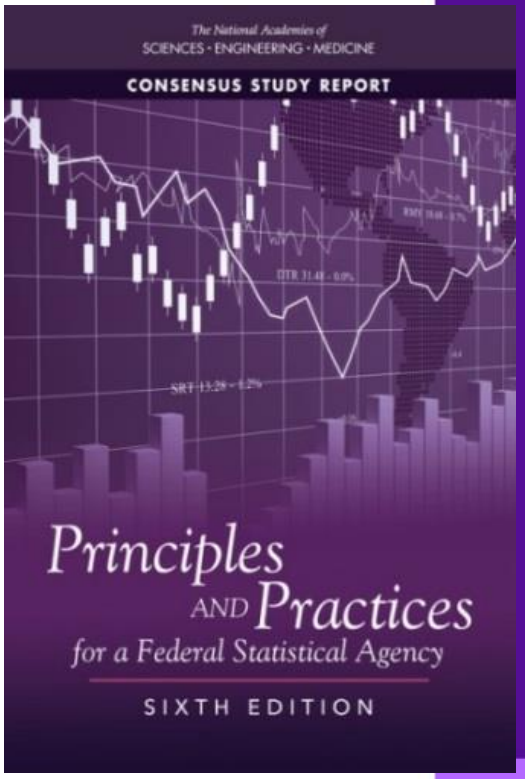
Committee on National Statistics

The National Academies of SCIENCES · ENGINEERING · MEDICINE

The CNSTAT Portfolio

Principles and Practices for a Federal Statistical Agency –

Now in its 6th edition (released July 2017); emphasizes statistical agency independence to make possible the production of relevant, high-quality, timely, accessible, and nonpartisan statistics.



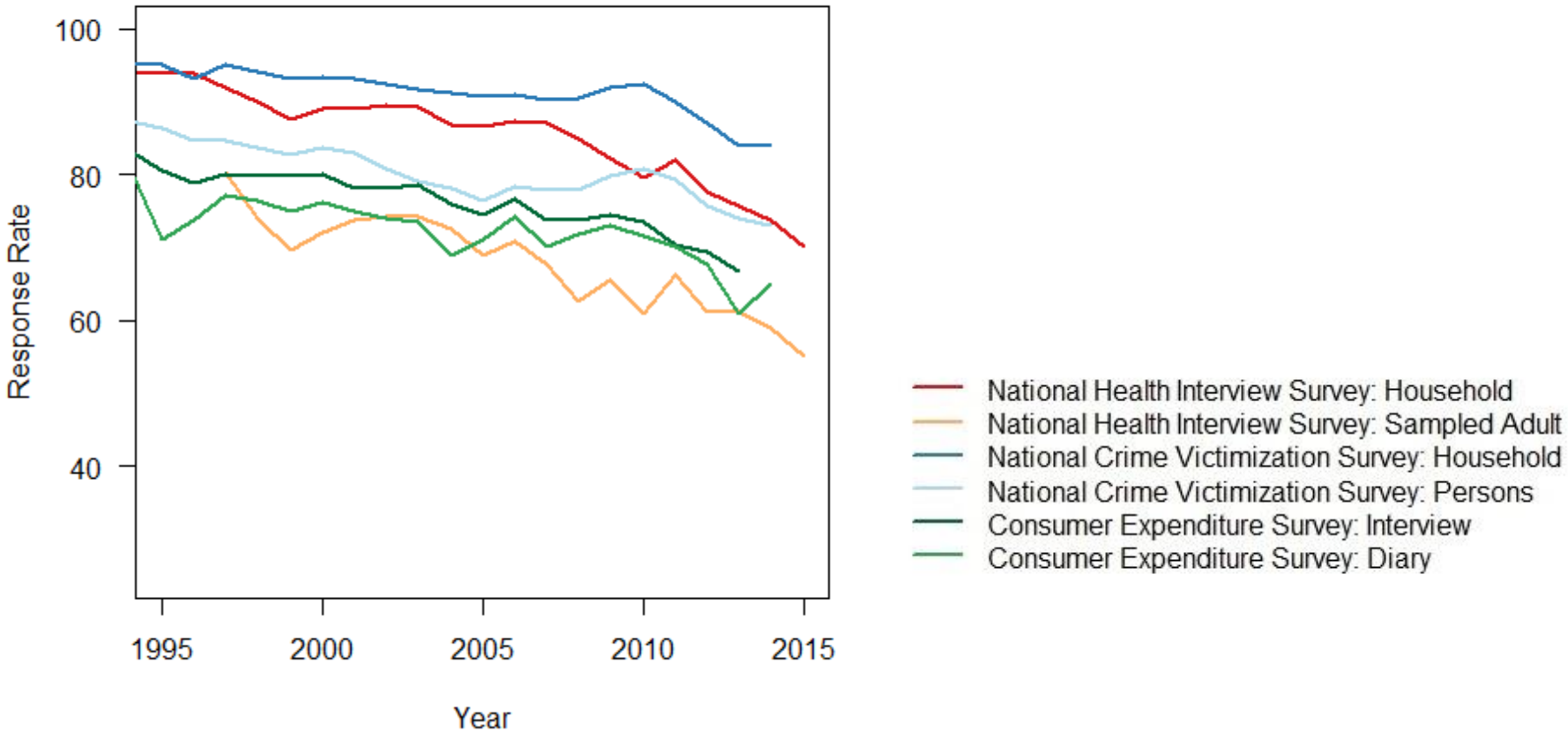
Committee on National Statistics
The National Academies of
SCIENCES · ENGINEERING · MEDICINE

Current Challenges and Opportunities in Federal Statistics

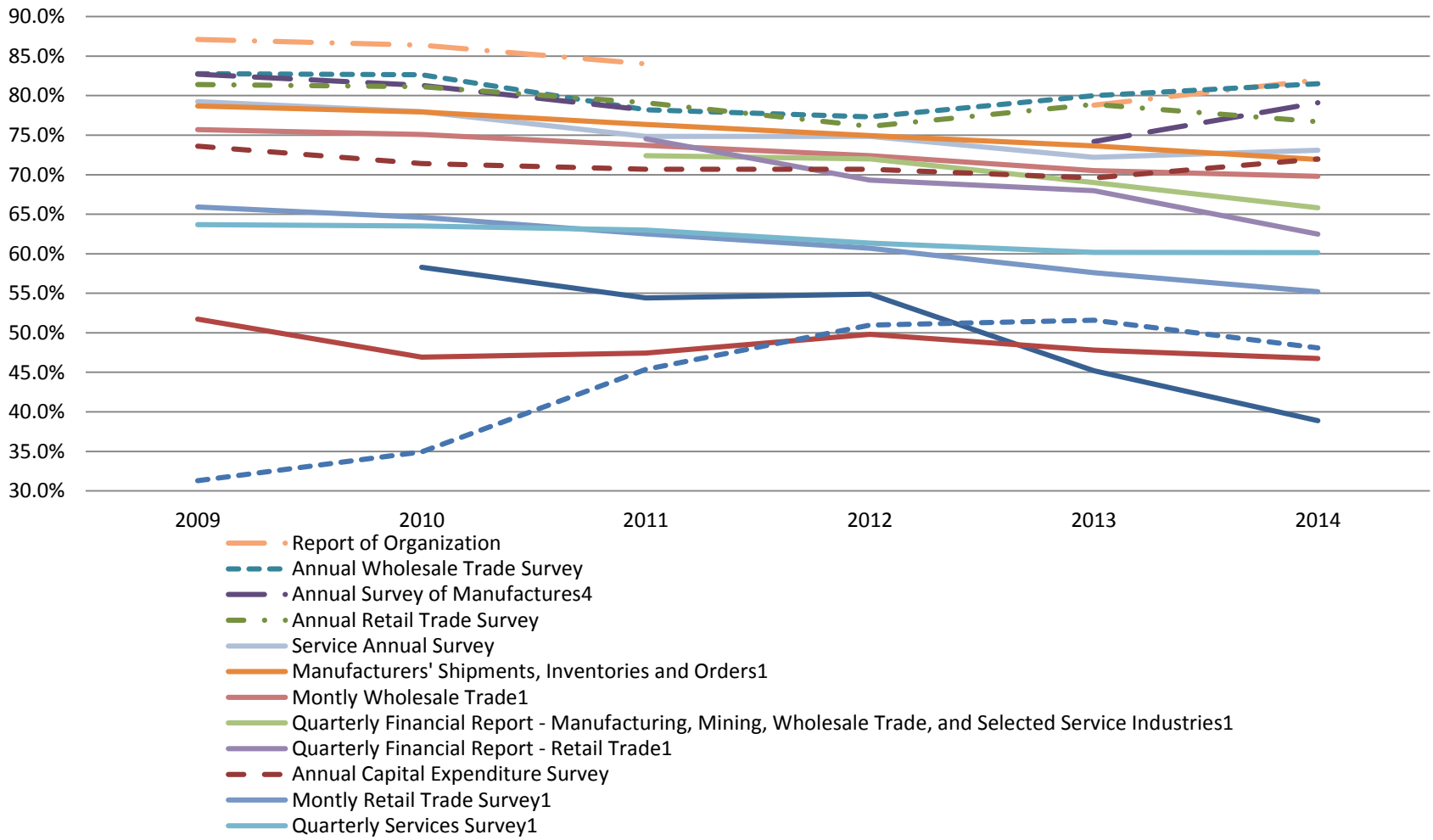
The way that statistics are currently produced by Federal statistical agencies faces threats from declining participation rates and increasing costs.

- 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

Response Rates for Three Surveys in which at least One Interview is Done in Person



Census Bureau Economic Directorate Survey Unit Response Rates 2009-2014



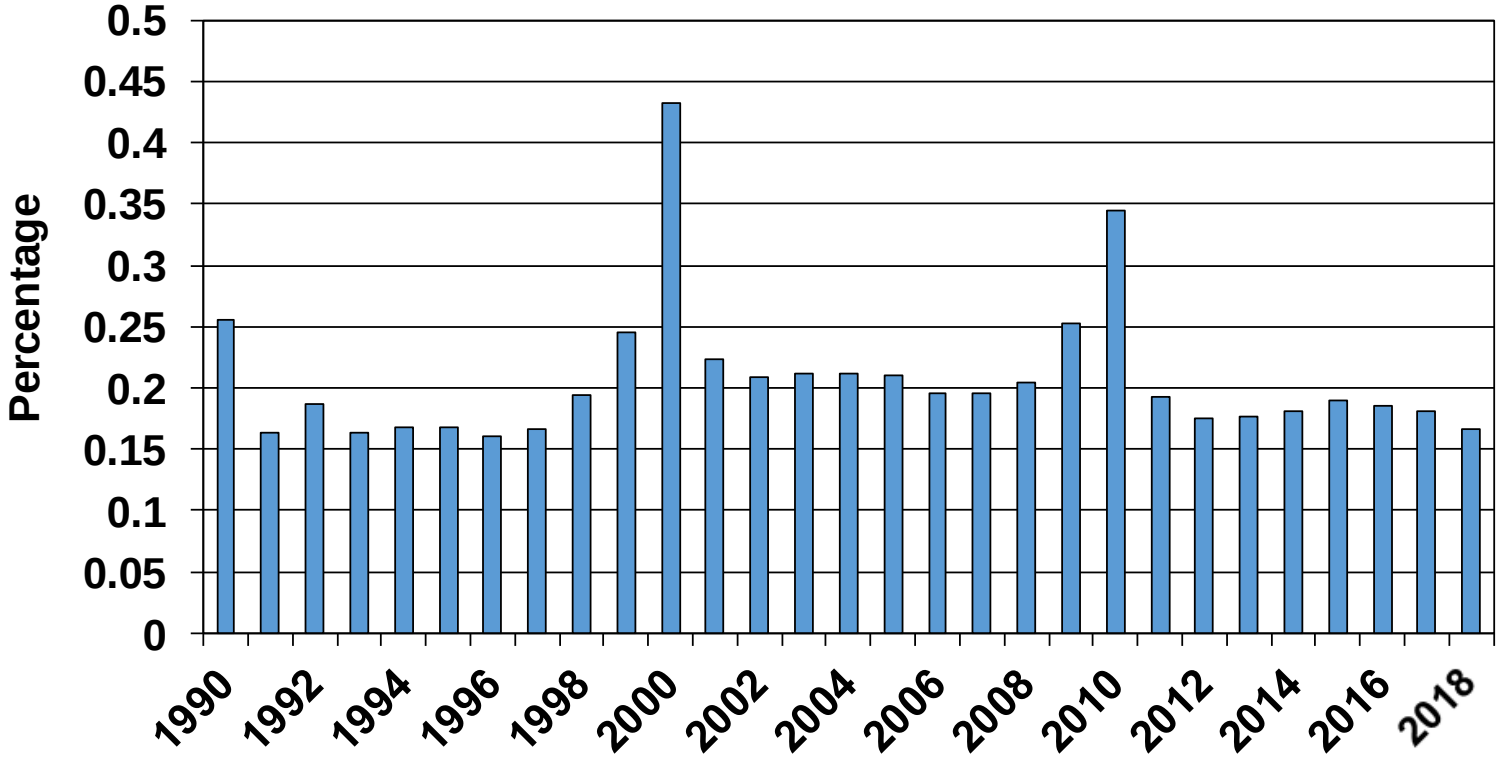
Overview of U.S. Federal Statistical System

- The U.S. decentralized system includes about 100 agencies with an FY 2018 budget of about $6.5 billion (excluding decennial census)
  - About $2 billion contracted to private sector or states
- A substantial portion of official statistics are produced by 13 principal statistical agencies.
  - Account for approximately 40 percent of resources dedicated to federal statistics

Federal Statistical System Decentralized

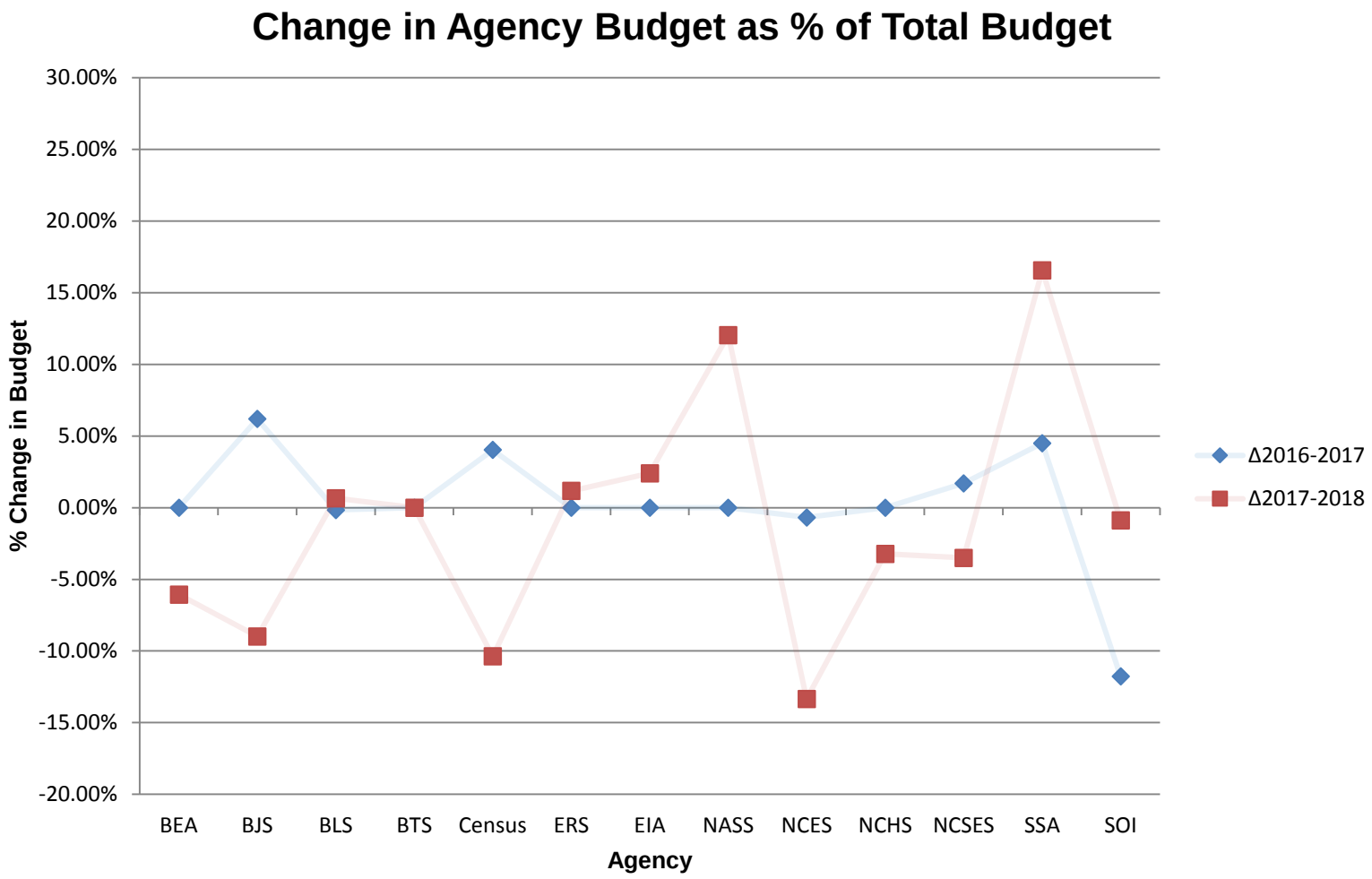
Statistics Account for 0.15% to 0.45% of the Federal Budget

(U.S. Government Expenditures on Statistical Programs, 1990-2017, including Decennial Census, as Percent of Total Federal Outlays)



Nominal dollars

Decreasing and Flatlining Budgets

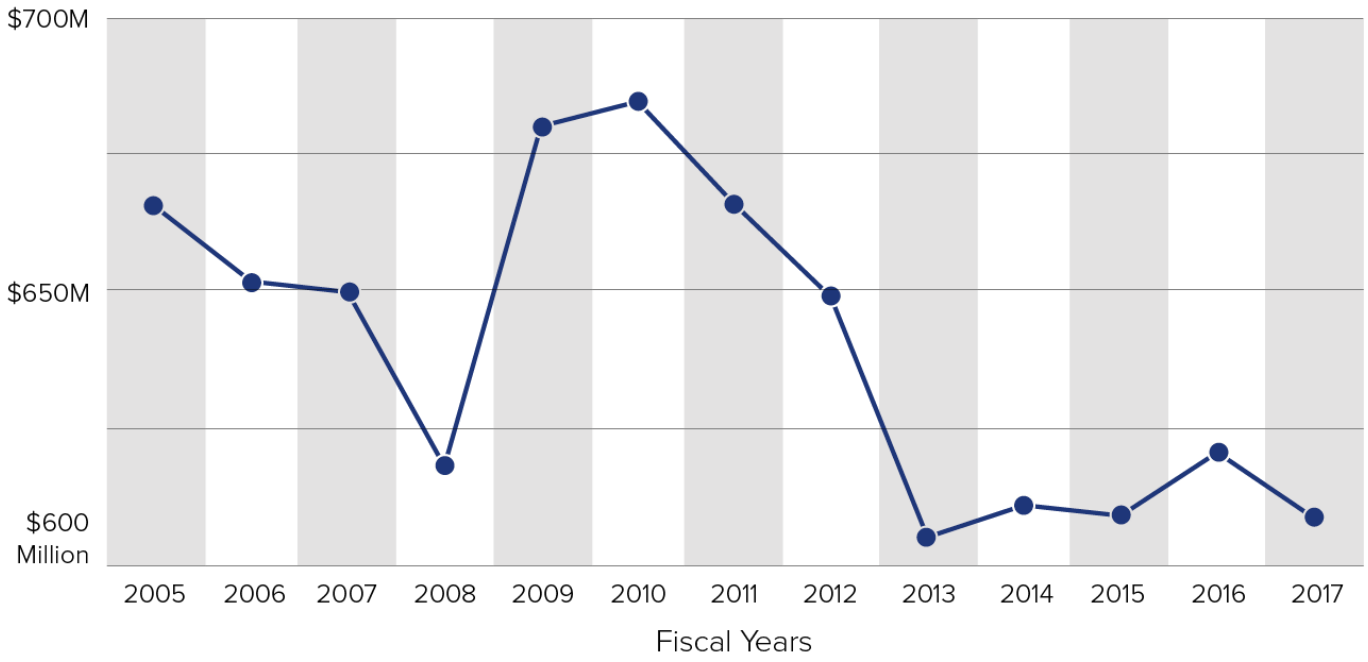


Note: Census Bureau figures reflect ongoing programs, not periodic programs.

Decrease in Labor Statistics Budget

Labor statistics on a diet

Congress has cut the inflation-adjusted budget for the Bureau of Labor Statistics by almost 10 percent since 2010, forcing the agency to pull back on ideas for new, updated measures of the economy—and even cut some surveys altogether.



Source: Statistical Programs of the United States, FY2005-2017 | **Graphic by Christina Animashaun**

POLITICO

Priority Issues for the Future of U.S. Federal Statistics

- **New data sources—**
  - How can agencies use administrative records and private-sector data sources to enhance federal statistics?
  - What privacy issues need to be addressed when combining data sources?
  - What quality frameworks and metrics are appropriate for these new sources and for blended estimates?
  - How can agencies collaborate to access, evaluate, and use these new data sources?

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.**

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

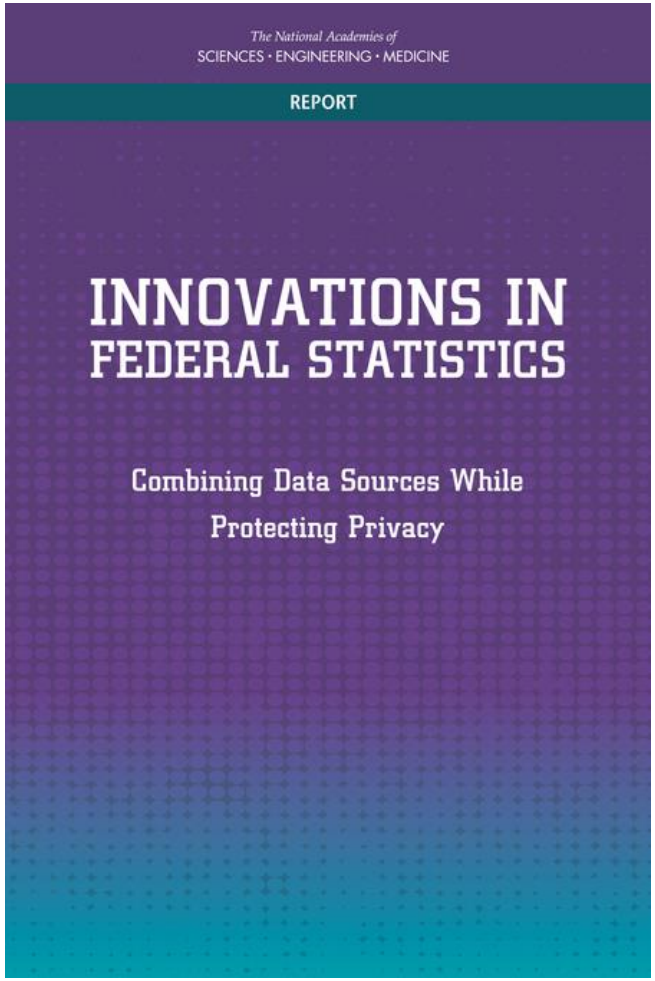
Acknowledgements

Funding for the panel was provided by

The Laura and John Arnold Foundation,

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

The Panel's First Report:



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

Contents

- Chapter 1: Introduction
- Chapter 2: Current Challenges and Opportunities in Federal Statistics
- Chapter 3: Using Government Administrative and Other Data for Federal Statistics
- Chapter 4: Using Private-Sector Data For Federal Statistics
- Chapter 5: Protecting Privacy and Confidentiality While Providing Access to Data for Research Use
- Chapter 6: Advancing the Paradigm of Combining Data Sources

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.

Structured Data: Censuses and Probability Surveys	Structured Data: Administrative Records	Other Structured Data	Semistructured Data	Unstructured Data
Collecting data from the universe or a sample of that population and estimating their characteristics through the systematic use of statistical methodology	Data collected by government entities for program administration, regulatory, or law enforcement purpose	Data that are highly organized and can easily be placed in a database or spreadsheet. They may still require substantial scrubbing and transformation for modeling and analysis	Data that have structure, but also permit flexibility in structure so that they cannot be placed in a relational database or spreadsheet. Transformation into structured form requires decisions with regard to the way in which to standardize the observed variety in structure. The scrubbing and transformation for modeling and analysis is usually more difficult than for structured data	Data, such as in text, images, and videos, that do not have any pre-defined structure. Information of value must first be extracted from such data, after which the extracted information can be placed in a structured table for further processing and analysis

• **Decennial Census of Population and Housing**
• **Economic Census**
• **Agriculture Census**
• **Federal statistical surveys**
  ○ **ACS**
  ○ **CPS**
  ○ **NHIS**
  ○ **AHS**
  ○ **NCVS**
	• Federal Records
	  ○ Income Tax
	  ○ Social Security
	  ○ Unemployment
	  ○ Medical Records
	• State Records
	  ○ SNAP data
	• Police accident reports
	• County Records
	• Other jurisdictions' data
		• Weather sensors
		• Traffic sensors
		• Water quality sensors
			Web-scraped quantitative data
			• Web logs
				• Traffic videos
				• Satellite images
				• Blogs and comments
				• Input in free text fields

Current Barriers to Use of Alternative Data Sources

- Conclusion 3-4: **Legal and administrative barriers** limit statistical use of administrative datasets by federal statistical agencies.

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.

Definition and Examples	Structured Data from Censuses and Probability Surveys	Structured Data from Administrative Records	Other Structured Data	Semistructured Data	Unstructured Data
Definition	Data from a population or a sample of that population used to estimate the population's characteristics through the systematic use of statistical methodology	Data collected by private companies from transactions, process control, or financial human resource records	Data that are highly organized and can easily be placed in a database or spreadsheet, though they may still require substantial scrubbing and transformation for modeling and analysis	Data that have structure, but also permit flexibility in structure so that they cannot be placed in a relational database or spreadsheet; the scrubbing and transformation for modeling and analysis is usually more difficult than for structured data	Data, such as in text, images, and videos, that do not have any structure So that information of value must first be extracted and then placed in a structured table for further processing and analysis

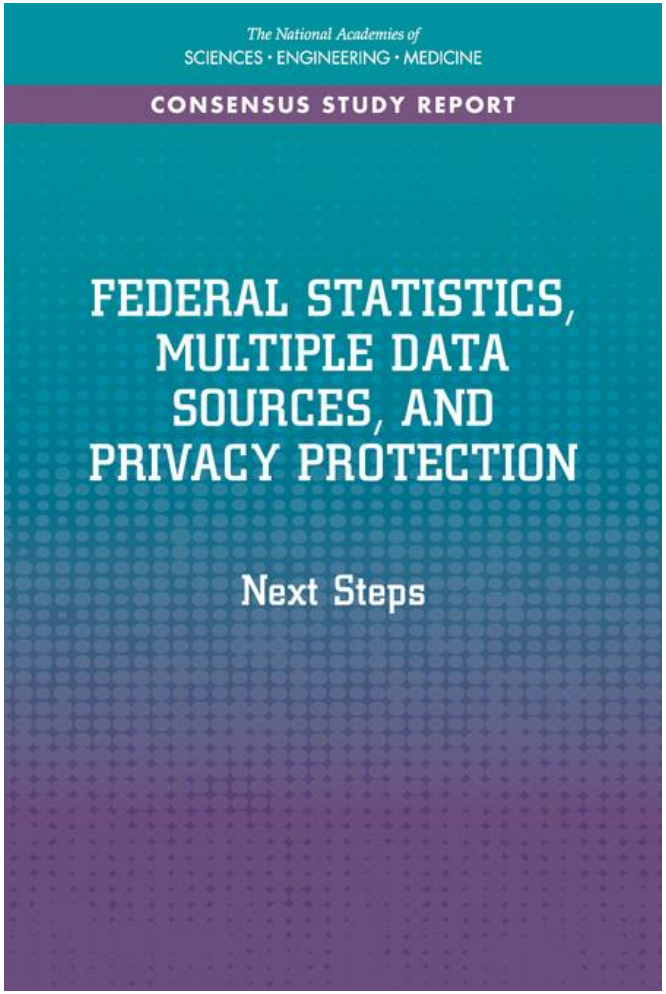
Private-Sector Examples	• Customer satisfaction surveys
	• Marketing research surveys
	• Media use surveys
	• Academic surveys
		• Data Produced by businesses
		  ○ Commercial transactions
		  ○ Banking and stock records
		  ○ Credit cards records
		  ○ Medical records
		• University and other nonprofits grant transactions
			• E-commerce transactions
			• Mobile phone location sensors
			• Global Positioning System sensors
			• Utility company sensors
			• Weather, pollution sensors
				• Extensible Markup Language (XML) files
				• Data from computer systems
				  ○ Logs
				  ○ Web logs
				• Mobile phone content: text messages
				• E-mail
				• Internet of things*
				• Sport activity sensors (from watches, etc.)
					• Social network data (Facebook, Twitter, Tumblr, etc.)
					• Internet blogs and comments
					• Documents
					• Pictures (Instagram, Flickr, Picasa, etc.)
					• Internet searches
					• Traffic webcams
					• Security/surveillance videos/images
					• Satellite images
					• Drones
					• Radar images

Advancing the Paradigm of Combining Data Sources

RECOMMENDATION 6-1: A new entity or an existing entity should be designed to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics

RECOMMENDATION 6-2: The proposed new entity should maximize the utility of the data for which it is responsible while protecting privacy by using modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.

The Panel's Second Report:



Committee on National Statistics

The National Academies of SCIENCES • ENGINEERING • MEDICINE

Table of Contents

Next Steps for Combining Data Sources

RECOMMENDATION 2-3 Current statistical methods should be adapted to the extent possible and new methods should be developed to harness the statistical information from multiple data sources for analysis.

RECOMMENDATION 2-4 Federal statistical agencies should ensure their statistical staff receive training for the new skills needed for combining data from different sources.

RECOMMENDATION 2-5 Federal statistical agencies should develop partnerships with academia and external research organizations to develop methods needed for design and analysis using multiple data sources.

Data Processing Issues

CONCLUSION 3-3 Creating statistics using multiple data sources often requires complex methodology to generate even relatively simple statistics. With the advent of new and different sources and innovations in statistical products, federal statistical agencies need to figure out ways to provide transparency of their methods and to clearly communicate these methods to users.

Personnel Staffing and Skills

RECOMMENDATION 3-1 Because technology changes continuously and understanding those changes is critical for the statistical agencies' products, federal statistical agencies should ensure that their information technology staff receive continuous training to keep pace with these changes. Training programs should be set up to meet the current and expected future training needs for technology, and recruitment plans should account for future technology demands.

Privacy Implications for Federal Statistical Agencies

RECOMMENDATION 4-1 Because linked datasets offer greater privacy threats than single datasets, federal statistical agencies should develop and implement strategies to safeguard privacy while increasing accessibility to linked datasets for statistical purposes.

Privacy Implications for Federal Statistical Agencies

RECOMMENDATION 5-1 Federal statistical agencies should ensure their technical staff receive appropriate training in modern computer science technology including but not limited to database, cryptography, privacy-preserving, and privacy-enhancing technologies.

Broader Frameworks for Assessing Quality

CONCLUSION 6-3 Timeliness and other dimensions of granularity have often been undervalued as indicators of quality; they are increasingly more relevant with statistics based on multiple data sources.

RECOMMENDATION 6-1 Federal statistical agencies should adopt a broader framework for statistical information than total survey error to include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability.

Assessing the Quality of Administrative and Private Sector Data

RECOMMENDATION 6-2 Federal statistical agencies should outline and evaluate the strengths and weaknesses of alternative data sources on the basis of a comprehensive quality framework, and, if possible, quantify the quality attributes and make them transparent to users. Agencies should focus more attention on the tradeoffs between different quality aspects, such as, trading precision for timeliness and granularity, rather than focusing primarily on accuracy.

RECOMMENDATION 6-3 Federal statistical agencies should ensure their statistical and methodological staff receive appropriate training in various aspects of quality and the appropriate metrics and methods for examining the quality of data from different sources.

A New Entity to Provide Vital Information through Enhanced Federal Statistics

- Attributes of the New Entity
  - Organizational Location
  - Functions
  - Technological Environment for Data Access
  - Access by Outside Researchers
  - Privacy
  - Transparency
  - Financing
  - Governance
- Implementation

Core Requirements for New Entity

- It has to have **legal authority to access data** that can be useful for statistical purposes. The legal authority needs to span cabinet-level departments and independent agencies.
- It has to have **strong authority to protect the privacy of data** that are accessed and prevent misuse. At minimum, that authority needs to be commensurate with existing laws (CIPSEA, the Privacy Act), but it may also require new legislation.
- It has to have **authority to permit appropriate uses** for the extraction of statistical information from the multiple datasets relevant to program evaluation and the monitoring of policy-relevant social and economic phenomenon. The authority needs to delimit what uses are forbidden as well as what uses are encouraged.
- It needs to be **staffed with personnel whose skills fit the needs of the recommended entity**, including advance IT architectures, data transmission, record linkage, statistical computing, cryptography, data curation, cybersecurity, and privacy regulations.

Implementation

- A strategic plan will be needed for expanding the data sources accessible through the entity.
- This plan will need to be carefully structured in phases, detailing outcomes for each phase and decision points.
- The first phase might cover 5 years, at which time it would be useful to have a comprehensive review.
- The first phase needs to include expanded access to federal administrative and operational data that could be useful for federal statistics.
- Private-sector data could be included as part of the new entity in a later phase.
- How this entity is created and how it functions will determine its ability to be an effective resource of and for the federal statistical system.

Thank You!

Contact Information

- **Brian Harris-Kojetin, CNSTAT Director**
  bkojetin@nas.edu

- **CNSTAT Website**
  http://www.nationalacademies.org/cnstat

CNSTAT reports are available at the National Academies Press:
http://www.nap.edu

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