BIG DATA DAY

May 11, 2018
Washington, DC

Sponsored by the Interagency Council on Statistical Policy

Hosted by CNSTAT

CONTENTS

EVENT SCHEDULE

TIME	EVENT	DETAILS
1:00 p.m. – 2:00 p.m.	**Exhibit Hall Opens** Light Refreshments See Floor Plan	**Posters and Demonstrations**
2:00 p.m.– 2:45 p.m.	**Welcome** Auditorium	**Brian Moyer**, Bureau of Economic Analysis, ICSP Employee Development Working Group Executive Sponsor
	CNSTAT Panel Auditorium	**Bob Groves**, Georgetown University, CNSTAT Chair **Sallie Keller**, Virginia Tech University **Jerry Reiter**, Duke University
2:45 p.m. – 4:00 p.m.	**Lightning Presentations** Auditorium	**Megan Sweitzer**, Economic Research Service Scanner Data in the Economic Research Service Consumer Food Data Program **Adley Kloth**, U.S. Census Bureau Census Data Lake **Anne Parker**, Internal Revenue Service Recommendation System Application for Anomaly Detection and Missing Value Imputation **Erik Friesenhahn**, Bureau of Labor Statistics Linking Inter-Agency Databases **Nanda Srinivasan**, U.S. Energy Information Administration Innovative Uses of Administrative Data (in collaboration with BLS)

EVENT SCHEDULE

—continued

TIME	EVENT	DETAILS
2:45 p.m. – 4:00 p.m.	**Lightning Presentations** —continued Auditorium	**Carol DeFrances**, National Center for Health Statistics "Big Data" for Health Care through the Collection of Electronic Health Record Data **Abe Dunn**, Bureau of Economic Analysis A Dive into U.S. Expenditures on Treatment by Disease, 2000–2014 **Carol Robbins**, National Center for Science Engineering Statistics New Opportunities to Observe and Measure Innovation, Modeling, Infrastructure, and Standards
4:00 p.m. – 5:00 p.m.	**Exhibit Hall** **Light Refreshments** See Floor Plan	**Posters and Demonstrations**

CNSTAT PANEL BIOGRAPHIES

Robert M. (Bob) Groves (NAS/NAM) is executive vice president and provost of Georgetown University, where he is also the Gerard J. Campbell professor in the Department of Mathematics and Statistics and a professor in the Department of Sociology. Prior to joining Georgetown as provost, he served as director of the U.S. Census Bureau from 2009 to 2012 after being appointed by President Barack Obama. Previously, he was director of the University of Michigan Survey Research Center and research professor at the Joint Program in Survey Methodology at the University of Maryland. He also served as associate director for research and methodology of the U.S. Census Bureau from 1990 to 1992. He is an elected member of the National Academy of Sciences (NAS) in the Social and Political Sciences Section and the National Academy of Medicine (NAM) in the Social Sciences, Humanities and Law Section. He served as a member of CNSTAT from 2000 to 2006, as a member of the Division of Behavioral and Social Sciences and Education (DBASSE) from 2014 to 2016, and has served on numerous Academies boards, panels, and committees, including chair of the Panel to Review Programs of the Bureau of Justice Statistics (BJS); member of the Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences: A Workshop; and the Workshop Steering Committee on Enhancing Research and Development for the Federal Statistical System, among others. He is an elected fellow of the American Statistical Association and an elected member of the International Statistical Institute and the American Academy of Arts and Sciences. He has an A.B. in sociology from Dartmouth College and an M.A. and Ph.D. in sociology from the University of Michigan.

Sallie Ann Keller is Director and Professor of Statistics for the Social and Decision Analytics Laboratory within the Virginia Bioinformatics Institute at Virginia Tech University. Her prior positions include Academic Vice-President and Provost at University of Waterloo, Director of the IDA Science and Technology Policy Institute, the William and Stephanie Sick Dean of Engineering at Rice University, Head of the Statistical Sciences Group at Los Alamos National Laboratory, Professor of Statistics at Kansas State University, and Statistics Program Director at the National Science Foundation.

Dr. Keller has served as a member of the National Academy of Sciences Board on Mathematical Sciences and Their Applications, the Committee on National Statistics, and has chaired the Committee on Applied and Theoretical Statistics. Dr. Keller's areas of expertise are social and decision informatics, statistical underpinnings of data science, uncertainty quantification, and data access and confidentiality. She is a leading voice in creating the science of all data and advancing this research across disciplines to benefit society. She is a fellow of the American Association for the Advancement of Science, elected member of the International Statistics Institute, fellow and past president of the American Statistical Association, and member of the JASON advisory group. She holds a Ph.D. in statistics from the Iowa State University.

Jerome P. (Jerry) Reiter is professor of statistical science at Duke University. Before joining Duke as an assistant professor, he was a lecturer of statistics at the University of California, Santa Barbara, and an assistant professor of statistics at Williams College. He participates in both applied and methodological research in statistics, and is most interested in applications involving social science and public policy. His methodological research focuses mainly on data confidentiality, missing data, and survey methodology. He previously served on the CNSTAT Panel on Addressing Priority Technical Issues for the Next Decade of the American Community Survey—First Phase; the TRB Committee on the Long-Term Stewardship of Safety Data from the Second Strategic Highway Research Program; and the CNSTAT-CPOP Panel on Collecting, Storing, Accessing, and Protecting Biological Specimens and Biodata in Social Surveys. He is a fellow of the American Statistical Association, an elected member of the International Statistical Institute, principal investigator of the Triangle Census Research Network (funded by the National Science Foundation to improve the practice of data dissemination among federal statistical agencies), and deputy director of the Information Initiative at Duke, an institute dedicated to research and applications in the analysis of large-scale (and not large-scale) data. He has a B.S. in mathematics from Duke University, and an A.M. and Ph.D. in statistics from Harvard University.

PRESENTATIONS

Megan Sweitzer
Economic Research Service
U.S. Department of Agriculture

Scanner Data in the Economic Research Service Consumer Food Data Program

The aim of this presentation is to describe (1) the requirements and challenges ERS faced using large commercial scanner datasets for research, (2) ERS's work to evaluate the statistical properties of the data, and (3) value-added projects to address data shortcomings by filling the gaps and linking with traditional data sources.

Adley Kloth, Nitin Natik, Kevin Reid
U.S. Census Bureau
U.S. Department of Commerce

Census Data Lake

The Census Data Lake (CDL) is a flexible Big Data management platform intended to provide the Census Bureau with a next-generation scaling capability to fulfill data management, storage, reporting, analytics, and security requirements while reducing costs associated with duplicative data silos. The 2020 instance of CDL plans to accomplish notable business goals.

Anne Parker
Internal Revenue Service
U.S. Department of Treasury

Recommendation System Application for Anomaly Detection and Missing Value Imputation

In this presentation, we describe the application of a collaborative filtering model to identify anomalous values among a population of millions of observations across multiple data fields. The model improves upon current methods of identifying anomalies within the IRS. An additional benefit of this approach is that it is both trained on, and applied to, the very same set of data.

Erik Friesenhahn
Bureau of Labor Statistics
U.S. Department of Labor

Linking Inter-Agency Databases: The Creation of Employment and Wage Estimates for Foreign-Owned Firms

A BLS project combines data from state unemployment insurance reports with surveys of foreign-owned business in the United States collected by BEA. As a result, the Quarterly Census of Employment and Wages (QCEW) and the Occupational Employment Statistics programs will be able to produce new estimates of employment and wages for foreign-owned businesses in the United States.

Nanda Srinivasan
U.S. Energy Information Administration
U.S. Department of Energy

Innovative Uses of Administrative Data: Leveraging the QCEW to Build a Frame of Petroleum Marketers

This presentation describes using interagency collaborations to leverage existing administrative data sources for a new purpose. We present the journey of developing a data-sharing agreement between the BLS and the EIA and detail how we achieved a "yes" from agency leadership. Technical issues and strategies to overcome them are also discussed.

Carol DeFrances
National Center for Health Statistics
Centers for Disease Control and Prevention
U.S. Department of Health

"Big Data" for Health Care through the Collection of Electronic Health Record Data

With the growing adoption and use of electronic health record (EHR) systems by health care providers, the NCHS is looking to leverage EHR data for its health care surveys, starting with the inpatient and ambulatory care settings. The move to EHR data offers tremendous analytical capabilities that raise the potential for new and exciting health services research.

Abe Dunn
Bureau of Economic Analysis
U.S. Department of Commerce

A Dive into U.S. Expenditures on Treatment by Disease, 2000–2014

We introduce detailed data on spending by condition to take an in-depth analysis of spending growth over the 2000 to 2014 period. The detailed spending data is built from a combination of survey data and large claims databases with millions of enrollees and billions of claims. The large sample size allows us to study spending by disease at more disaggregated levels than previously possible. Using these data, we report spending statistics for 261 conditions.

Carol Robbins
National Center for Science Engineering Statistics
National Science Foundation

New Opportunities to Observe and Measure Innovation, Modeling, Infrastructure, and Standards

Two case studies demonstrate the challenges and possibilities of developing measures of intangible assets using non-survey sources of data. The case studies focus on the value of organizational processes in a Fortune 500 company and on open-source software. Our goal is to assess the quality of the data to develop innovation measures that are scalable and repeatable.

DEMONSTRATIONS

Cavan Capps
U.S. Census Bureau
U.S. Department of Commerce

Secure Distributed Statistical Processing of Encrypted Streaming Data

This prototype employs machine learning (ML) to standardize product codes, parses encrypted standard business transaction files, tabulates data from multiple company donors, and aggregates and releases formally confidential results. Simple business intelligence queries can be done to inform business decisions in the shipping and logistics industry as well as inform national economic policy.

Jeff Chen
Bureau of Economic Analysis
U.S. Department of Commerce

Machine Learning Techniques for Economic Data

In this demonstration session, ML is introduced in the context of data from BEA among other sources. ML techniques are growing in influence in research and applied environments. Whereas techniques in the classic econometric toolkit are focused on causal inference, machine learning is a matter of honing predictive capabilities that are optimized for accuracy.

John Cuffe
U.S. Census Bureau
U.S. Department of Commerce

Developing Automated Industry Codes Using Publicly-Available Data

By combining information from IRS tax forms with publicly-available information on businesses, we are able to generate a new coding model for the North American Industry Classification System (NAICS), one that is more accurate and lower cost than previous methods. This research has significant cross-agency implications in our ability to improve NAICS and other classification models.

Ryan Farrell
Bureau of Labor Statistics
U.S. Department of Labor

Census of Fatal Occupational Injuries Public Data Management System

The Census of Fatal Occupational Injuries has created a new web-scraping application that identifies and stores potential source documents that utilize Google Alerts Service. This demonstration highlights the ability to search for relevant articles using keyword searches, state of incident, and date range as well as other features.

Ledia Guci
Bureau of Economic Analysis
U.S. Department of Commerce

Improving Regional Personal Consumption Expenditure Estimates Using Credit Card Transaction Data

Based on credit card data, we estimate consumption flow patterns for a variety of retail industries. We demonstrate that the consumption flows show predictable activity with consumers preferring nearby locations, larger economic markets, and popular vacation destinations. We also show how these estimates can be used to correct for border-crossing issues in the regional personal consumption expenditure state estimates.

Lisa Mavrogianis
U.S. Department of Veterans Affairs

Open Data Department of Veterans Affairs

The VA began the process of making its data available to the public in formats easily readable by computers, subject to restrictions allowable by law. This is in accordance with Executive Order 13642 released on May 9, 2013, which notes that "the default state of new and modernized Government information resources shall be open and machine readable." For further guidance, consult the Office of Management and Budget's Open Data Policy (M-13-13) and the Project Open Data website.

Joanne Pascale
U.S. Census Bureau
U.S. Department of Commerce

Classifying Health Insurance Type from Survey Responses Using Enrollment Data

The Census Bureau recently implemented a redesign of the health insurance module of the Current Population Survey Annual Social and Economic Supplement. Our research aims to inform development of an algorithm for combining answers to questions about health insurance from this module to maximize accurate categorization of coverage type.

Dipak Subedi
Economic Research Service
U.S. Department of Agriculture

The Challenges and Opportunities for a Federal Statistical Agency When Incorporating Code and Data Archival

In the last few years, economists at ERS implemented sweeping improvements to its farm income and wealth statistics data product, including version control and archiving of programming code used to create the data as well as archiving of all data for each release. This demonstration highlights the administrative and technical challenges faced along the way and the opportunities.

POSTERS

David Beede
Economics and Statistics Administration
U.S. Department of Commerce

Development of Winter Weather Forecast-Based Social Outcomes Risk Indices: A Machine Learning Approach

This poster demonstrates the potential for using machine learning techniques and environmental Big Data from federal government sources to develop risk indices of social outcomes. We combined data on weather, traffic accident fatalities, school closures, demographics, and economic conditions to produce prototype indices for the risk of winter-weather-related traffic fatalities and school closings.

John Bockrath
Bureau of Economic Analysis
U.S. Department of Commerce

Estimating Air Passenger Fares with Detailed Ticket Data

This study proposes enhancing the BEA's current methodology for estimating international trade in air passenger transport (APT) by using a dataset from the Airline Reporting Corporation. Average fares for foreign residents traveling to and from the United States are estimated at a highly disaggregated level to improve the accuracy of the APT estimates.

Claire Boryan
National Agricultural Statistics Service
U.S. Department of Agriculture

Operational Flood Monitoring of Agriculture During Hurricanes Harvey, Irma and Maria with Sentinel 1 Synthetic Aperture Radar

This study assesses the effectiveness of using freely available European Space Agency Sentinel-1 Synthetic Aperture Radar (SAR) data for operational agricultural flood monitoring in the United States. It concludes that the Sentinel-1 SAR is an effective, efficient, and affordable data source for operational disaster assessment.

Matthew Chambers
Bureau of Transportation Statistics
U.S. Department of Transportation

Bureau of Transportation Statistics Use of Automatic Identification System as a Big Data Source

BTS has partnered with the U.S. Army Engineer Research and Development Center to survey ferry service and create port performance metrics using Automatic Identification System (AIS) data in a variety of ways. AIS is a navigation tool that supplements voice radio communication or radar for collision avoidance.

Jennifer Davies
National Center for Education Statistics
U.S. Department of Education

EDFacts—Data Visualization and Data Dissemination

The EDFacts team uses data visualization to help bring large databases to life. The visualizations are helpful during data quality reviews to identify patterns at the state and national level and are also a valuable tool to help increase the accessibility of EDFacts data.

Jennifer Davies
National Center for Education Statistics
U.S. Department of Education

EDFacts—Data Quality and Evaluation

As the EDFacts system has matured over time, the level of scrutiny related to the quality of data submitted in terms of timeliness, completeness, and accuracy has increased as well. EDFacts business rules provide states with immediate feedback, and the EDFacts team conducts data quality reviews for program offices, providing a coordinated response for related data files.

Patrick Drake, Merianne Spencer
National Center for Health Statistics
Centers for Disease Control and Prevention
U.S. Department of Health

Analysis of Cause-of-Death Fields in Vital Records from the National Vital Statistics System (NVSS)

This poster describes efforts from NCHS to code and extract relevant cause-of-death information and specifically highlights two use-cases for analyzing literal text fields. Incorporating new methods to analyze and process literal text can allow for a better understanding of causes of death for improved data evaluation, timeliness, and quality of cause-of-death reporting.

Brian Dumbacher
U.S. Census Bureau
U.S. Department of Commerce

SABLE

Researchers at the Census Bureau are developing a collection of tools for web crawling, web scraping, and text classification called SABLE. The idea is to discover potential new data sources on the web, identify and scrape useful data, and then map the scraped data to Census Bureau terminology. Elements of SABLE involve machine learning and text analysis to perform text classification.

Irena Dushi
Office of Research, Evaluation and Statistics
Social Security Administration

Overview of SSA's Micro-Simulation Model "MINT"

Modeling Income in the Near Term (MINT) is a micro-simulation model developed by SSA to analyze the characteristics of future Social Security beneficiaries and simulate the distributional effects of proposed reforms to the Social Security program. MINT uses traditional SSA program data alongside non-SSA data as inputs to the model.

Rebecca Hutchinson
U.S. Census Bureau
U.S. Department of Commerce

Reducing Survey Burden Through Third-Party Data Sources

This poster details a successful pilot effort by the Census Bureau's Economic Directorate's Big Data Team along with the NPD Group, Inc. to compare NPD store- and national-level data feeds with data from the Monthly and Annual Retail Trade Surveys and with the 2012 Economic Census. Additionally, product data from the Economic Census was matched to the product-level data feeds provided by NPD.

Andrew Kerns
Economic Research Service
U.S. Department of Agriculture

Leveraging Big Data Tools and Cloud Storage for Research on Decades of Climate Data

This project used publicly available PRISM Climate Group data and Multivariate Adaptive Constructed Analogs data, with Amazon Web Services Cloud, to expand ERS's computing resources by allowing researchers to rapidly provision dedicated compute instances to process the data. Parallel processing enabled quicker production of derived products for economic analysis.

Tamara Lee, Jin Kim
National Center Veterans Analysis and Statistics
U.S. Department of Veterans Affairs

Veteran Population Projections 2017 to 2037

This poster presents (1) the annual change of the total Veteran population from 2017 to 2037, (2) the Veteran population by race and ethnicity over time, (3) the Veteran population by period of service over time, and (4) where Veterans live over time—the top 10 states and annual percent change by congressional district.

Jody McLean
National Center for Health Statistics
Centers for Disease Control and Prevention
U.S. Department of Health

More Biological Data, New Challenges

The NHANES Biospecimen Program allows researchers to use stored biospecimens for studies resulting in new microdata files. NHANES currently has Genome Wide Associations Studies data available (800,000+ variables), Oral Microbiome data forthcoming (110,000+ variables), and is anticipating data from a blood methylation pilot in the near future (480,000+ variables).

William McNary
U.S. Energy Information Administration
U.S. Department of Energy

Using Submetering Technologies to Measure Energy Usage in Homes

The EIA has begun to test the reliability and feasibility of submetering technologies on a small number of homes as an alternative strategy for estimating end uses. Rather than rely on respondents, energy suppliers' data and models, these sensing technologies may offer improved ways to objectively measure end uses.

Lisa Mirel
National Center for Health Statistics
Centers for Disease Control and Prevention
U.S. Department of Health

Linking Survey and Non-Survey Data for Health and Housing Research

A collaboration between the NCHS and U.S. Department of Housing and Urban Development (HUD) linked two surveys with HUD administrative data. The resulting linked data files enable researchers to examine relationships between the receipt of federal housing assistance and health.

Jennifer Ortman, David Raglin
U.S. Census Bureau
U.S. Department of Commerce

Making Administrative Records Key to Operational Agility for the American Community Survey

While there is great potential for the role of administrative records in the future of data collection and processing, there are also great challenges to using these data (e.g., issues with matching rates and geographic coverage). This poster details research on the use of administrative records to measure selected housing characteristics and income in place of survey response on the ACS.

Kevin Scott
Bureau of Justice Statistics
U.S. Department of Justice

Using Open-Source Tools to Improve Data Coverage—The Case of Arrest-Related Deaths

BJS has developed a hybrid, two-step process to improve coverage of law enforcement homicides. We present results on the utility of this hybrid process for both coverage and ability to reliably collect information about the attributes of identified arrest-related deaths.

Sean Simone
National Center for Education Statistics
U.S. Department of Education

Reducing Error and Burden in NCES Postsecondary Sample Surveys Using Administrative Data

This poster presents historic problems that the NCES has encountered with data provided from interview respondents (in terms of both quality and burden), how administrative data collections have addressed these problems, and the challenges that have emerged when using transactional data in Federal surveys.

Brian Smith
Bureau of Economic Analysis
U.S. Department of Commerce

Exploring Applications of Airbnb Listings in Modernizing Rent-To-Value Ratios

This interagency project, part of the ICSP Mentoring Program, finds that Airbnb listing data could be used to improve BEA estimates of rent-to-value ratios, which are used in estimating housing expenditures. However, to implement this process on an ongoing, publication-level scale, significant obstacles must be overcome.

Scott Wentland
Bureau of Economic Analysis
U.S. Department of Commerce

Monetary Policy and Home Prices: Big Data Applications for Macro Research

This presentation provides a brief overview of a new Big Data source at BEA, the Zillow "ZTRAX" data, and how daily-level data can be applied to key macroeconomic research questions. This data source allows researchers, who traditionally use aggregated monthly or quarterly data, to analyze macrophenomena (like a sector's response to monetary policy shocks) using daily-level microdata.

Avery Sandborn
National Agricultural Statistics Service
U.S. Department of Agriculture

CropScape and VegScape for Agricultural Geospatial Big Data Online Visualization and Dissemination

This presentation highlights two web-based GIS Big Data geospatial applications, CropScape and VegScape. Both showcase the technology capability for online visualization, analytics, dissemination, and web geoprocessing for geospatial Big Data.

Guangyu Zhang
National Center for Health Statistics
Centers for Disease Control and Prevention
U.S. Department of Health and Human Services

Augmenting the National Hospital Care Survey (NHCS) Sample with Non-Sampled Hospitals Registered for Meaningful Use (MU)—A Simulation Study

This study investigates whether augmenting NHCS data with non-sampled hospitals registered for the MU program can improve the NHCS national estimates. Total emergency department visits were derived and compared between a first-step sample and a combined sample.

Full abstracts are available at
http://sites.nationalacademies.org/dbasse/cnstat/dbasse_185840.

THANK YOU

Many people helped to make Big Data Day a reality, and we thank you. We also thank the following for their leadership and vision:

Brian Moyer, Director of the Bureau of Economic Analysis, executive sponsor of the Interagency Council on Statistical Policy's Employee Development Working Group

ICSP's Employee Development Working Group:

Laniera Jones, BEA, Big Data Day Chair
Maribel Aponte, NCVAS
Natalie Dupree, NCHS
Jennifer Edgar, BLS
Elimar Medina Figueroa, Census Bureau
Jael Jackson, SOI
Jaki McCarthy, NASS
John Popham, BJS
Susan Schneider, NCHS
Alia Shabazz, ERS
Marc Sinofsky, ORES
Nanda Srinivasan, EIA

*Special thanks to **Natalie Dupree**, the originator of the innovation showcase idea.*

The Committee on National Statistics:

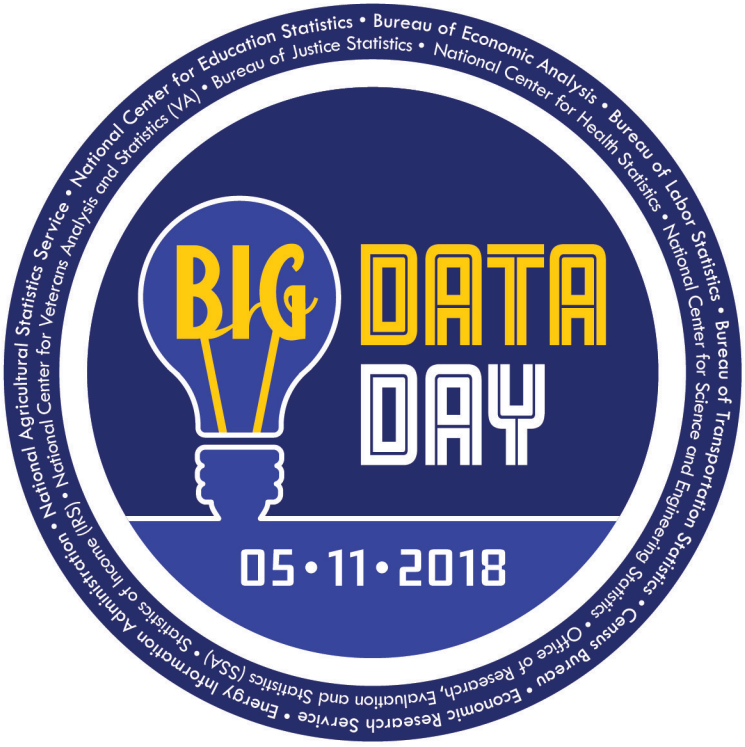
Brian Harris-Kojetin, Director
Eileen LeFurgy
Anthony Mann
George Schoeffel
Glenn White

BEA Big Data Day Working Group:

Jeannine Aversa
Ryan Byrnes
Lucas Hitt
Colby Johnson
James Kim
Gianna Marrone
Shaunda Villones

Thank you for participating! Please share your thoughts here:
https://www.surveymonkey.com/r/ICSPBDD

You may also scan this QR code to complete a survey.


National Agricultural Statistics Service • National Center for Education Statistics • Bureau of Economic Analysis • Bureau of Labor Statistics • Bureau of Transportation Statistics • Census Bureau • Economic Research Service • Energy Information Administration • National Center for Veterans Analysis and Statistics (VA) • Bureau of Justice Statistics • National Center for Health Statistics • National Center for Science and Engineering Statistics • Office of Research, Evaluation and Statistics (SSA) • Statistics of Income (IRS)
BIG DATA DAY
05•11•2018