

# Big Data and Modernizing Federal Statistics: Update

Bill Bostic  
Associate Director  
Economic Programs Directorate

Ron Jarmin Ph.D.  
Assistant Director,  
Research and Methodology Directorate

August 2015

# Big Data Trends and Challenges

- Trends
  - Increasingly data-driven economy
  - Individuals are increasingly mobile
  - Technology changes data uses
  - Stakeholder expectations are changing
  - Agency budgets and staffing remain flat/reduced.
- The next generation of official statistics
  - Utilize broad sources of information
  - Increase granularity, detail, and timeliness
  - Reduce cost & burden
  - Maintain confidentiality and security
  - Measure impact of a global economy
- Multi-disciplinary challenges :
  - Computation, statistics, informatics, social science, policy

# MIT Workshop Series Objectives

- **Convene** experts –
  - in computer science, social science, statistics, informatics and business
- **Explore**
  - challenges to building the next generation of official statistics
- **Identify**
  - new opportunities for using big data to augment official statistics
  - core computational and methodological challenges
  - ongoing research that should inform the Big Data research program

# Workshop Organizers

## Series Conveners

- Census: Cavan Capps, Ron Prevost
- MIT: Micah Altman

## Workshop conveners

- First Workshop:  
Ron Duych (DOT), Joy Sharp (DOT)
- Second Workshop  
Amy O'Hara, Laura McKenna, Robin Bachman
- Third Workshop:  
Peter Miller, Benjamin Reist, Michael Thieme

## Workshop Sponsors

- William Bostic, Ron Jarmin

# Workshop Coverage

## Approach

- Examine a set of broad topical questions through a specific case
- Link broad issues to specific approach case
- Link specific challenge of case to broad challenge

## Topics

### Acquisition – Data Sources

*Using New forms of Information for Official Economic Statistics*

[August 3-4]

### Access -- Privacy Challenges

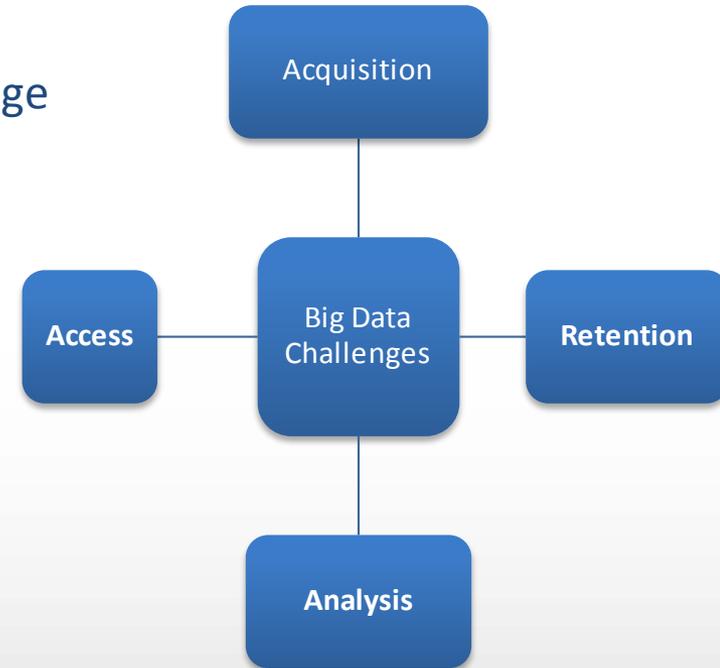
*Location Confidentiality and Official Surveys*

[October 5-6]

### Analysis – Inference Challenges

*Transparency and Inference*

[December 7-8]



# Preliminary Observations from First Workshop

- Topic:
  - Sources of Economic Big Data
- Use Case:
  - Commodity Flow Survey
- Observations:
- Different classes of decisions require different sources of data:
  - Designed survey data contributes baseline data for decisions about infrastructure and strategic planning
  - Transaction based big data could contribute frequency and granularity of estimates
- In big data, data sources are stakeholders
  - Businesses need to react quickly and predict the future – and need frequently updated detailed data
  - Critical to provide a value proposition to business
  - Critical to develop a trust relationship
- Some Potential sources
  - ERP and DRP operations data
  - EDI
  - Mobile Phone
  - Traffic Data

# Privacy and Big Data (Preview)

- Topic:  
Location Privacy
- Use Cases:
  - LEHD
  - Google Now
- Sample questions:
  - How can technical innovations such as differential privacy be adapted to protect public micro-data and statistical estimates?
  - How can we develop a modern privacy risk index for official data releases?

# Expected Outcomes

- Workshop Slides  
[August, October, December]
- Workshop Executive Summaries  
[September, October, January]
- Big Data White Paper  
[February]

[projects.informatics.mit.edu/bigdataworkshops](https://projects.informatics.mit.edu/bigdataworkshops)

# Retail Big Data Project Goal

- To explore the use of “Big Data” to supplement existing monthly/annual retail surveys to fill in data gaps and increase relevance. The primary focus is to try to produce subnational geographic area estimates more frequently than once every five years (from the Economic Census).

# Areas of Focus

- Point of Sale (POS) / Scanner Data (from companies like NPD, Nielsen, IRI)
- Credit Card Transaction Data (from credit card companies, 3<sup>rd</sup> party payment processors, banks, etc.)
- Direct Feeds from Companies

# Background on NPD Data

- 36 Months - January 2012 through December 2014
- Aggregated Sales by Geography, Product, Type of Merchant (Not Consistent)
  - Geography (varied by dataset)
    - Census regions
    - Metro areas
  - Product Level Aggregation by SKU
  - Type of merchant (jewelry and watches only)
- Names of Companies included in aggregated totals

# Evaluation Steps

- Analyzed NPD data to identify potential errors prior to use
  - Errors in geographic coding
  - Missing data
- Compared NPD data with Census Bureau estimates to obtain a rough assessment of coverage and to determine if the NPD data could serve as a potentially informative predictor
  - Aggregate levels (monthly or yearly totals)
  - Period-to-period changes (e.g., current-to-prior month, current-to-prior quarter)
- Census Bureau Data Used
  - Monthly Retail Trade Survey
  - Annual Retail Trade Survey
  - 2012 Economic Census
  - Business Register

# Major Findings

- 3 Datasets didn't help us achieve our goal
- Discrepancies Between NPD Data and Census Official Data
  - Geographic area differences
  - Less than full coverage of retail stores
  - Differences in industry/product classification
  - Under-representation of small firms
- Jewelry dataset highly correlated with national estimates from MRTS

# Next Steps with POS Data

- Explore the jewelry time series from NPD to better understand using big data in production models
- Continue discussions with scanner data companies on obtaining more useful data
- Evaluate the use of big data sources across all parts of the survey life cycle

# Credit Card Transaction Data

- Credit card transaction data (MasterCard, 3<sup>rd</sup> party processors, banks, etc.)
  - Joint effort with BEA
  - Test file obtained from MasterCard
  - Monthly Data from January 2009 through March 2015
  - Selected retail and service industry breakouts for 3 states
  - Research to be completed by January 2016
  - Discussions with other companies

# Data Feeds from Companies

- Data feeds directly from companies
  - Started meeting with companies
  - Received favorable feedback
  - Will proceed with working out details
  - Will test with a few companies in 2017 Economic Census

# Questions for CSAC Committee

- What is the reaction to name we have selected for new R&M Center?
- Is Census Bureau still headed in the right direction for its Big Data research?
- What could derail the Census Bureau's Big Data effort going forward?