BIG Data and OFFICIAL Statistics

International Year of Statistics
November 13, 2013

Michael W. Horrigan

Associate Commissioner

Office of Prices and Living Conditions

BLS

Big Data and Official Statistics

- What are big data?
  - A few examples
  - Definition / scope
- How is BLS using big data?
- Examining big data through the lens of data quality frameworks
- What is the future of using big data by statistical agencies?

BLS

What are 'Big Data'? A few examples

- Billion prices project
  - Daily CPIs in 20 countries
  - Webscraping technology
- Google
  - Tools to create large data files that combine publicly available data on social and economic activity stratified by geography, and social-demographic characteristics

BLS

What are 'Big Data'? A few examples

- Google
  - Modeling form combines Google search index data in the current period with past values of an economic measure from the statistical system to predict a future value of the same concept.
  - $Y_t = f(\text{Search}_{t-1 \text{ to } t}, Y_{t-1, t-2, \ldots})$
  - Example: Initial claims

BLS

What are 'Big Data'? A few examples

- UPS
  - Using telematic sensors in over 46,000 vehicles, big data on route selection, speed, and direction
  - Estimated savings of 8.4 million gallons of fuel by cutting off 85 million miles of route driven in 2011.
- GE
  - Use of real time monitoring of machines with big data analytic techniques to improve productivity of electricity generating machines, aviation, rail transportation, and health care.

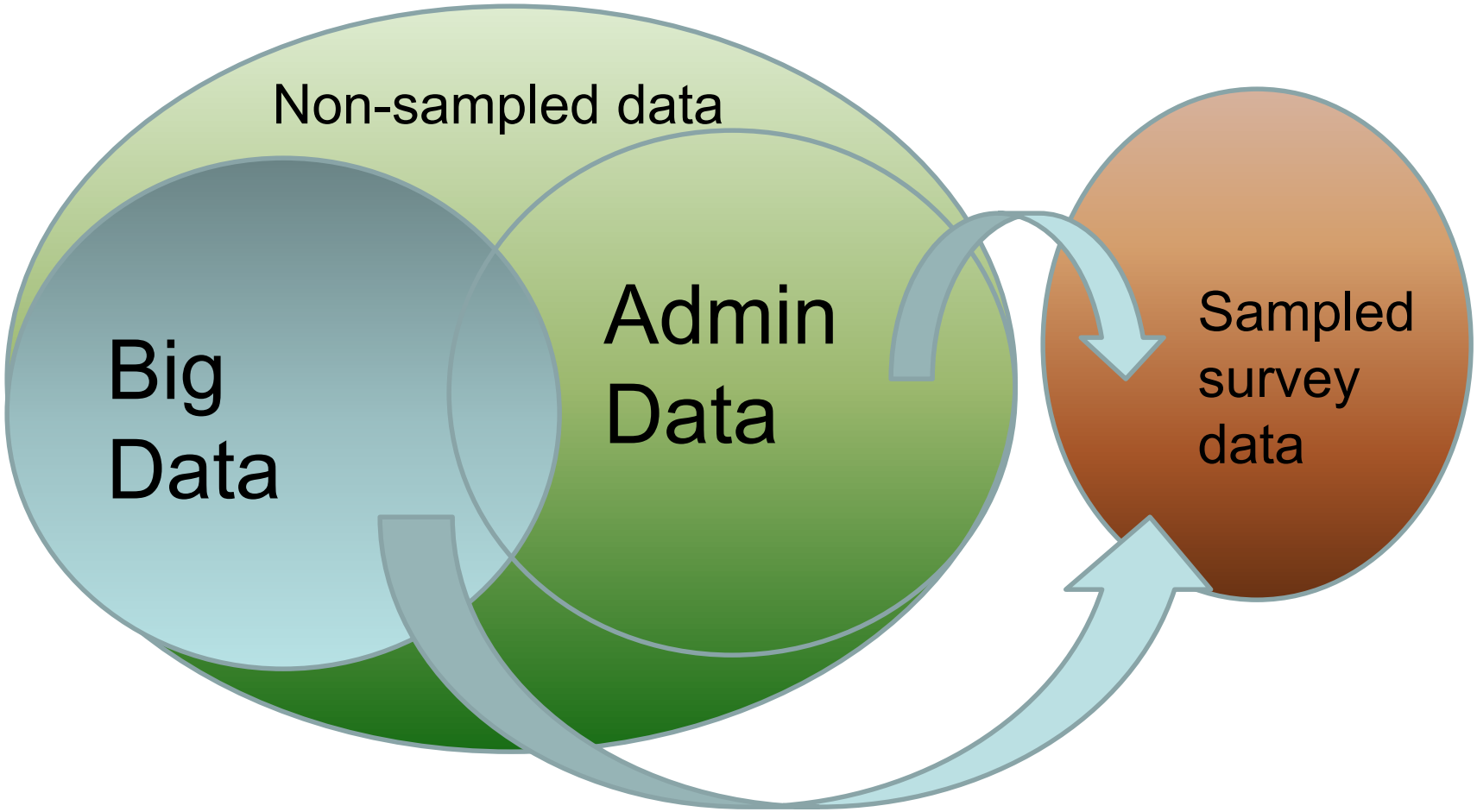
BLS

Big Data – Definitions/Scope

- Wikipedia
  - Big data is term for the collection of data sets so large and complex that it becomes difficult to process using hands-on data base management tools or traditional data base processing applications.
- 3V definition
  - Volume, Velocity, Variety

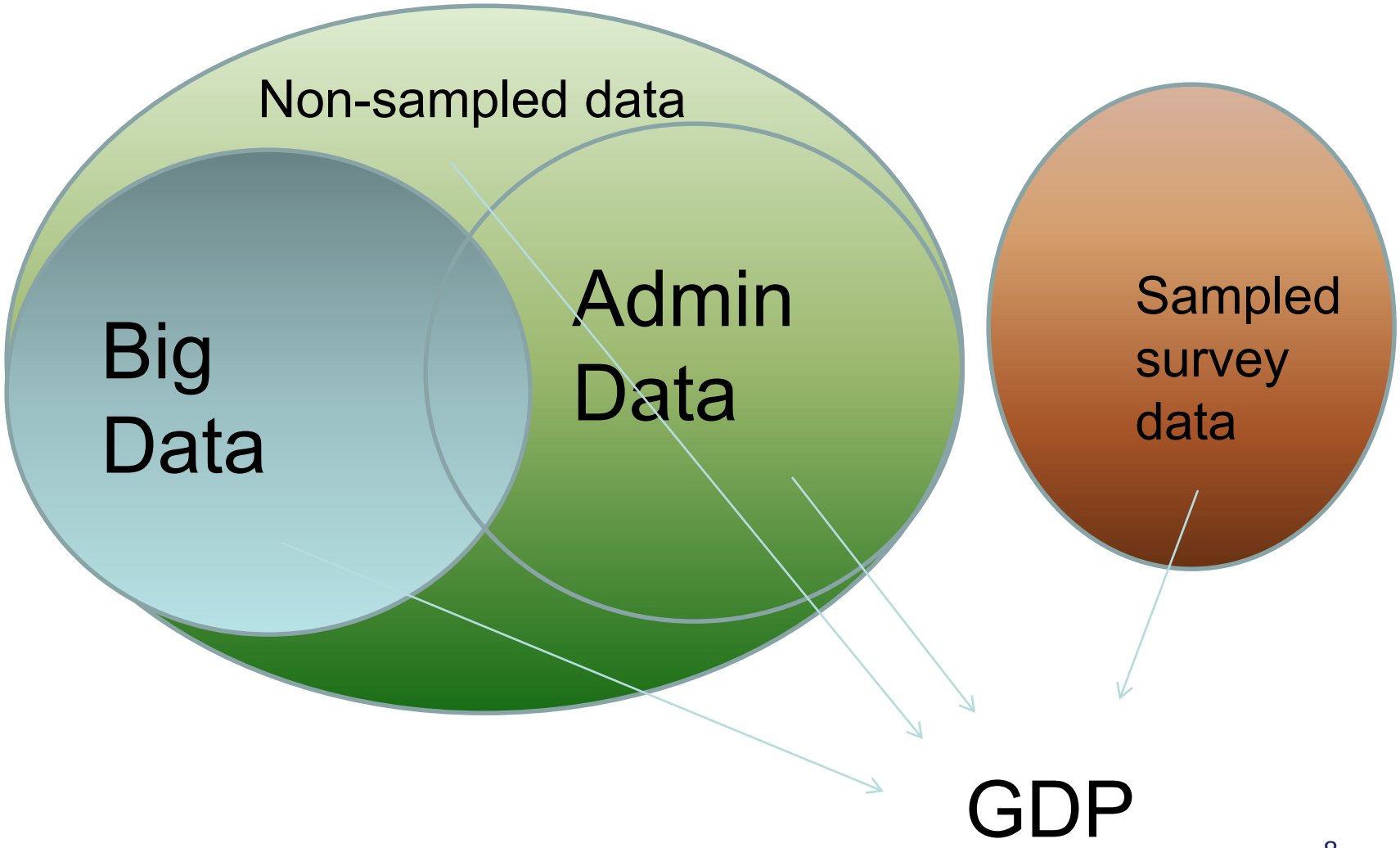
BLS

Big Data and Official Statistics



BLS

Bureau of Economic Analysis



Big Data and Official Statistics

- What are big data?
  - A few examples
  - Definition / scope
- How is BLS using big data?
- Examining big data through the lens of data quality frameworks
- What is the future of using big data by statistical agencies?

BLS

How are Big Data being used?

- Webscraping – BLS CPI
  - Create data base of product characteristics for use in quality adjustment hedonic models
    - Televisions
    - Camcorders
    - Camera
    - Washing Machines
  - Research to expand use to collect prices for cable TV plans and airline prices

BLS

How are Big Data being used?

- Scanner data: Homescan, Nielson
  - Actual sales transactions
  - Comparison of national distribution of selected products with results from CPI disaggregation process
- JD Power
  - Used car frame for CPI
  - Researching use for CPI production of new car price indexes

BLS

How are Big Data being used?

- Medicare part B
  - PPI and CPI use reimbursements to doctors by procedure code in indexes
- Claims data
  - Validation of MEPS and CPI inflation rates
  - Note: CPI constructs experimental disease based price indexes using annual weights from the MEPS household survey data

BLS

How are Big Data being used?

- Stock Exchange Security Trades
  - PPI receives a monthly census of all bid and ask prices and trading volume for all traded securities as of market close for 3 selected days of the month.
  - These data are used for index estimation
- Ratio allocation
  - CPI cost allocation weights and gasoline data from EIA

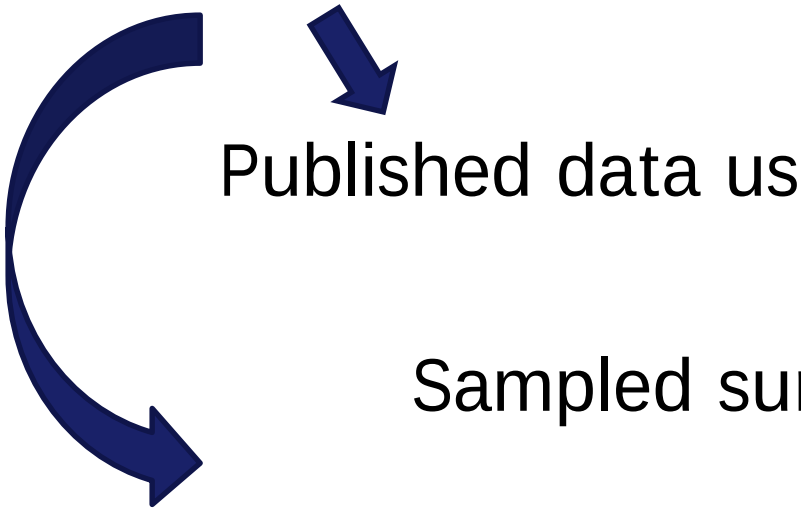
BLS

How are Big Data being used?

- Company provided data – Corp X
  - Research by CPI to use company provided data on all register transactions for sampled outlets
  - Challenges:
    - Can the matched model requirement be satisfied
    - Accounting for substitutes
    - IT production requirements
    - Risk of losing access
  - Opportunity:
    - Use of corporate data in direct estimation

BLS

How are Big Data being used?



How are Big Data being used?

- BLS Quarterly Census of Employment and Wages: Some examples of uses:
  - BLS sampling: PPI, NCS, CES, OES, OSH, JOLTS, Green Jobs
  - Imputation: State based estimates use QCEW data to impute for key non-respondents
  - Use of QCEW data to develop forecasts that are used in the CES birth death model

How are Big Data being used?

Administrative data

Used directly in estimation

- IPP uses EIA data on crude petroleum for their import indexes
- PPI uses Department of Transportation data on baggage fees
- CPI uses SABRE data for airline prices

BLS

How are Big Data being used?

Administrative data

Linking

- Census Bureau's Longitudinal Establishment….
- BLS Business Employment Dynamics
- Linking within agencies
- Sharing across agencies: CIPSEA

Big Data and Official Statistics

- What are big data?
  - A few examples
  - Definition / scope
- How is BLS using big data?
- Examining big data through the lens of data quality frameworks
- What is the future of using big data by statistical agencies?

BLS

Assessing Big Data through the lens of Quality frameworks

- Statistical agencies use a variety of quality dimensions to judge the efficacy of their direct data collection programs.
- It is reasonable to ask how the use of Big Data by Billion Prices, Google, Intuit and others fare along the same dimensions
- The use of external data sets (Big, Administrative, Other surveys) by statistical agencies to produce 'blended' estimates should come under the same scrutiny

BLS

Quality as a three-level concept

Product Quality

- Timeliness
- Relevance
- Objectivity
  - Clear, unbiased
- Accuracy – sampling errors
  - Calculated, published, used in analysis
- Accuracy – non sampling errors
  - Coverage
    - Primary challenge to statistical systems
    - Often an advantage of Big Data

BLS

Product Quality

- Accuracy – non sampling errors
  - Non response bias
    - Significant concern of statistical systems about their own data and for Big Data
  - Classification/specification
    - Lack of cross walks across different classification systems across statistical systems, administrative data, firm data, big data
- Metadata/transparency/interpretability
- Coherence / comparability

Big Data and Official Statistics

- What are big data?
  - A few examples
  - Definition / scope
- How is BLS using big data?
- Examining big data through the lens of data quality frameworks
- What is the future of using big data by statistical agencies?

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Budgets for the U.S. statistical system are flat or declining in real terms

- Faced with budget cuts, such as through the sequester, most agencies cut programs
  - For example, in FY 2013, BLS eliminated the International Comparisons Program, the Mass Layoff Survey Program, and the Green Jobs Employment Program

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Big Data
  - Potential for large volumes of data that, except for up front ($) investment in infrastructure and skill enhancement needed for handling big data, potentially relatively less expensive than direct data collection.
  - Replace, enhance, expand or validate

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Replace
  - Replace directly collected data with big data estimates
    - Administrative data less subject to selection bias
      - Use of DOT baggage fees, ask/bid prices on securities
    - Self selection bias from social media sources very problematic for top side estimates
      - Use reliable survey methods for control totals and ratio allocate biased 'proportions' to sub categories
      - Replace the goal of unbiased estimation for detailed estimates that minimize Mean Square Error
    - Bias is not always an issue – QA model example

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Enhance
  - Use direct survey (or high quality blended) estimates for periodic repeated cross-sectional or time series estimates (monthly, quarterly, semi-annual, annual) and redefine the relevance and mission of the statistical system by publishing (noisier and blended) data on a more frequent basis (weekly, daily)
    - EIA weekly gasoline prices
    - CPI's three pricing periods

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Expand
  - Use big data, especially administrative data and corporate data, to expand coverage of the economy.
  - As of March 2013, 4.3% percent of businesses were of size 50 or more, accounting for just under 60% of employment.
  - Could IRS administrative records for the remaining 95.7% be used as a substitute for expensive direct data collection from small firms and greatly increase their coverage?

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Expand
  - In the first quarter of 2013, there were just under 16,000 firms with 250 or more employees, accounting for 17% of total employment in the US
  - Explicit strategies to gain access to corporate records from the 16,000 firms could again reduce the direct data collection burden from such large firms.
  - Corporate records have the potential of greatly increasing the quantity, quality and timeliness of data we collect.

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Expand
  - Two largest gaps in our coverage of the economy is in the service providing sector and in global production processes.
  - Could corporate data record keeping systems, especially firms with global production processes, be leveraged to fill in these gaps?

BLS

Leveraging Big Data and the Future of the U.S. statistical system

- Validate
  - Use of big data as a check on findings from direct survey collection such as validating the CPI market basket for select items against scanner data
  - Development of control/treatment groups
    - Machine learning by Google
    - Experimental design for social interventions

BLS

Contact Information

Michael Horrigan
Associate Commissioner
Office of Prices and Living Conditions
www.bls.gov
202-691-6960
horrigan.michael@bls.gov

What are “Big Data”?