

Federal Forecasters Conference

The role of 'big' data in forecasting

Michael Horrigan
Associate Commissioner
Office of Employment and
Unemployment Statistics

April 20, 2017

Outline

- What are 'big' data?
- A framework for thinking about 'big' data in forecasting
- How does BLS use 'big' data?
- Final thoughts on data gaps



What are 'big' data?

A view from outside
the statistical system



What are 'Big Data'?

A few examples

- Billion prices project
 - Daily CPIs in over 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

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

- $Y_t = f(\text{Search}_{t-1 \text{ to } t}, Y_{t-1, t-2, \dots})$

- Example: Initial claims

What are 'Big Data'?

A few examples

- Tweets University of Michigan Study database
 - Case study of job loss related tweets that examines the correlation with unemployment data to predict initial claims
- Intuit
 - Quicken payroll accounts - Time series of employment, compensation, hours worked, hourly rates of pay, % full time, new hire rate
 - Stratified by size, industries

What are 'Big Data'?

A few examples

- ADP Payroll

- Over the month change in payroll employment

- 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

What are 'Big Data'?

A few examples

● 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
- Power of 1% and the industrial internet
- 1% savings in fuel consumption in aviation would generate savings of \$30 billion
- 1% efficiency improvement in GE's global gas fire plant fleet would produce an estimated savings of \$66 billion in 15 years

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

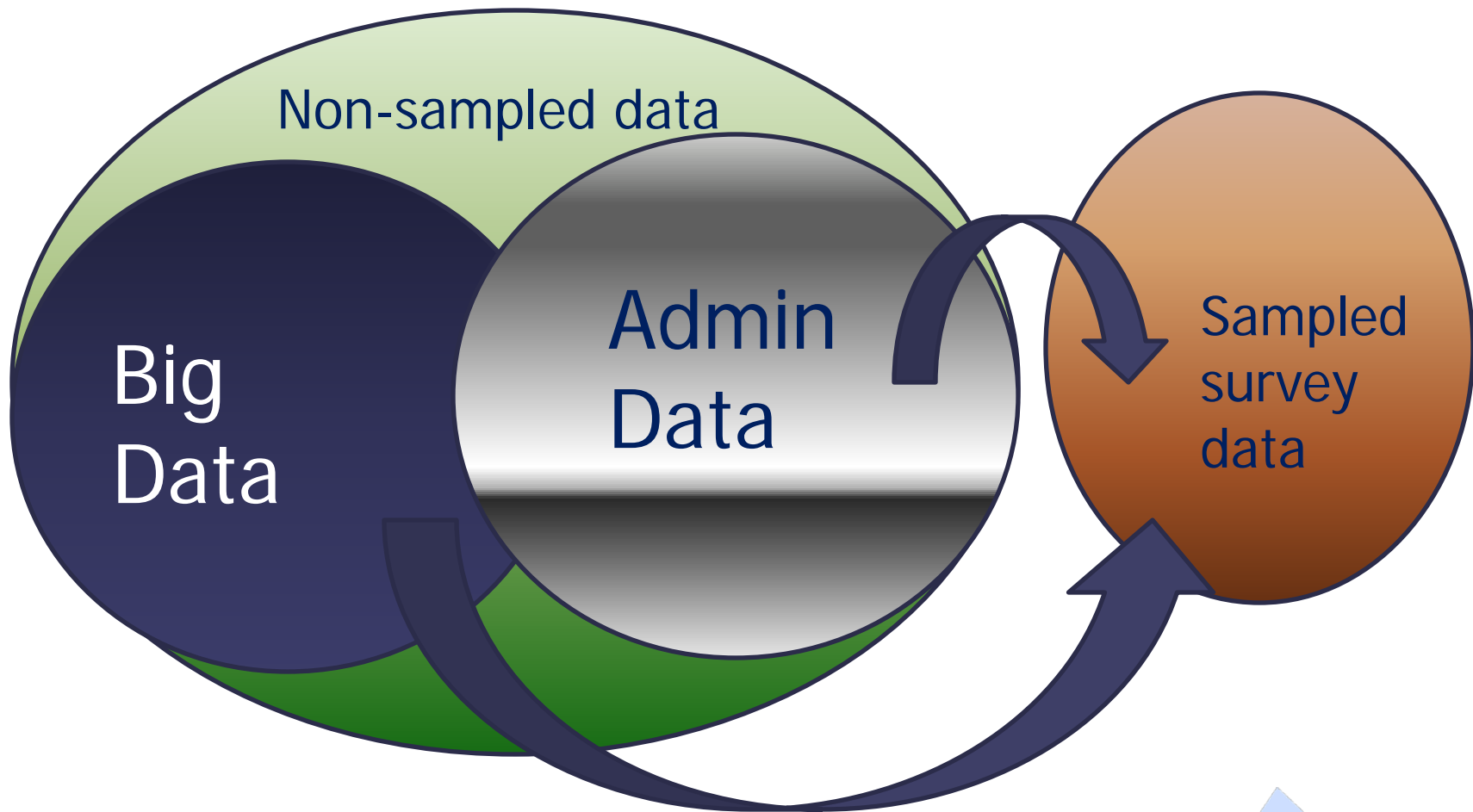
- Volume, Velocity, Variety
- SAS Institute – Variety not volume

- Transactional data

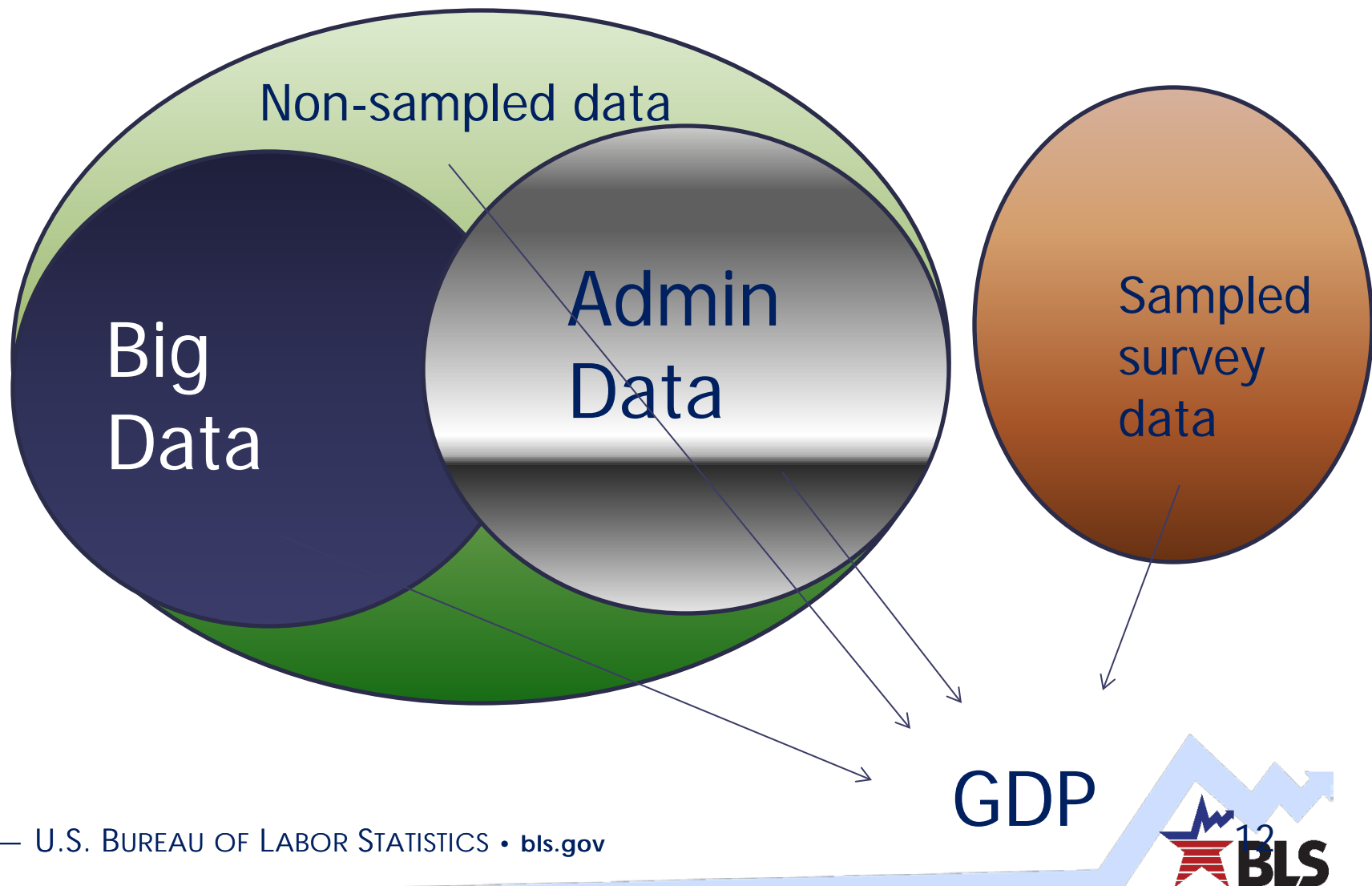
A framework for thinking about 'big' data in forecasting



Big Data and Official Statistics



Bureau of Economic Analysis



'Big' data and forecasting

● Forecasting

□ Time: out of sample forecasts

- Forecasts with error properties
- BLS projections example: natural rate of unemployment

□ Place

- Model subdomain detail
- Geography
- Strata subgroups

● Modelling

□ Causation versus correlation

□ Focus on inputs to improve forecasts



'Big' data and forecasting

- Quality measures

- Reduction of survey error
- Coherence checks
- Transparency of methods

- Big data methods

- Replace
- Blend
- Model / Use of 'big' data to ratio allocate to subdomains
 - Acceptable levels of Mean Squared Error
- Validate

Uses of alternative 'big' data at BLS



Types of alternative data and BLS uses

- Webscraped data
- Federal administrative data
 - Linking data sets
- Private Vendor data
- Modelling
- Corporate data
- Autocoding and text analysis

16

Webscraping

- Determine whether or not we need permission to scrape web sites.
- Examining the most promising areas for webscraping:
 - Food prices
 - Cable TV prices
 - Airline prices
 - Courier services

17



BLS uses of Federal Administrative Data

- Sampling frames
- Production of statistics.
 - Quarterly Data on employment and payrolls at nearly every establishment in the U.S.
 - Employment at startups, growing and contracting establishments

BLS uses of Federal Administrative Data

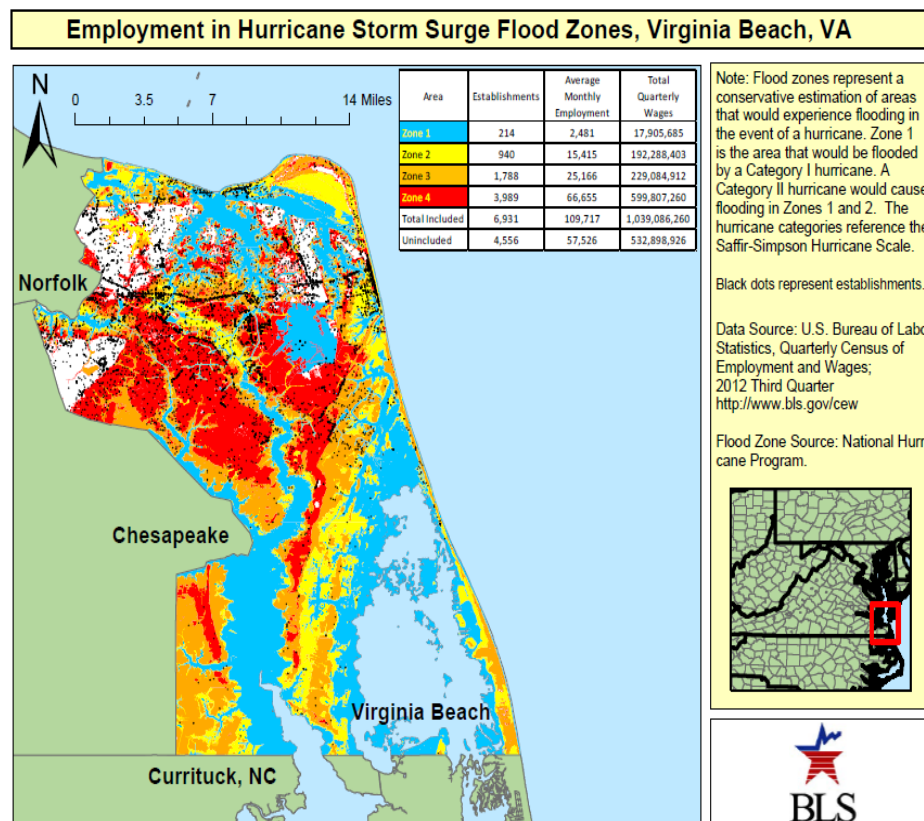
- Uses of Administrative data for direct estimation
 - Energy Information Agency --- Import Energy Price indexes
 - Department of Transportation baggages fees – Producer Price indexes
 - CMS – Medicare reimbursements for hospital and physician treatments – Producer Price Indexes

19



BLS uses of Federal Administrative Data

- Linking: Quarterly Census of Employment and Wages (QCEW) to:
 - Non profit data (IRS)
 - Hurricane maps (National Hurricane Program)
 - Foreign Direct Investment (BEA)
 - Occupational Employment Statistics Survey (OES)
 - Survey of Occupational Illnesses and Injuries (SOI) and OES



Private Vendor data: BLS uses

- Stock Exchange Security Trades – Producer Price Indexes
- JD Power – Consumer Price Indexes
- Scanner Data
 - Homescale, Nielson – Consumer Price Index research

21



Acquiring alternative data sets for use in estimation: future opportunities

- Supplement JOLTS data on vacancies with job openings data from private vendors (Snagajob, Burning Glass, Career Builder)
- Truven Health Analytics data for health care productivity measures
- Use of Compustat data to develop State level productivity estimates
- Use of credit card data collected by BEA to potentially use to create travel and tourism price indexes



Modelling

- Occupational Employment Statistics Program
 - Creation of time series data for the Occupational Employment Statistics program and imputing occupational staffing patterns to the universe of establishments
 - Develop estimates of annual changes in occupational employment and wages down to the state and possibly MSA level
 - Develop short-term forecasts / extrapolations

Modelling

- Job Opening and Labor Turnover (JOLTS)
 - Development of state based modelled estimates of vacancies, hires, quits, layoffs, and other separations
 - Potential future use of Unemployment Insurance wage records to model JOLTS data to finer levels of geography

Corporate data: BLS uses

- The Current Employment Statistics Survey (CES) collects data from 88 corporations at their Electronic Data Interchange facility in Chicago, IL
 - Accounts for nearly 10% of total weighted employment
 - Respondents submit electronic files in BLS formats
- The Occupational Employment Statistics (OES) Survey collects electronic data files from large firms that are also in the sample for the National Compensation Survey

25



Electronic data collection

- A large share of collected information in our establishment surveys comes from a small share of total establishments owing to the size concentration of economic activity
- In 2012, of the known value of U.S. exports that could be matched to specific companies:
 - the top 50 companies contributed nearly 31% of known value
 - the top 100 nearly 40%
 - the top 250 just over half
 - and the top 2000 nearly 78%

26



Electronic data collection: The future

- Allow firms to report using their formats and data bases
- Using autocoding learning models, or computational linguistics to convert firm based data and classifications to BLS concepts

27



Autocoding

- The Survey of Occupational Injuries and Illnesses is currently autocoding text fields on types of injuries into their classification system on injuries
- We are planning on introducing autocoding to classify job titles into the Standard Occupational Classification (SOC) system in the Occupational Employment Statistics (OES) Survey
- We are researching using text analysis in many other areas across BLS

28



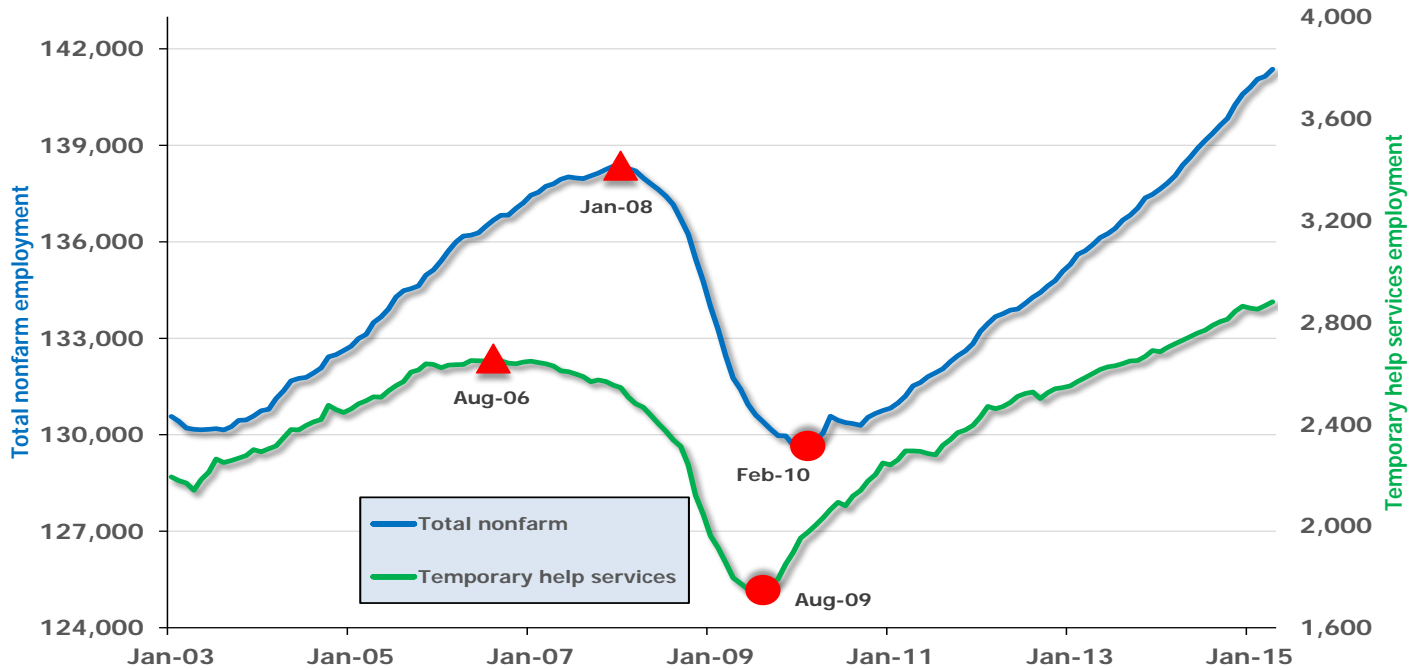
FINAL THOUGHTS ON DATA GAPS



Data Gaps

- What kinds of workers are used to produce goods or provide services?
 - The May 2017 Contingent Worker Survey (and the Katz/Kreuger study) are from the individual worker perspective
 - Need establishment data
- How much training is provided by firms?
- What other margins of adjustment are used such as domestic and foreign outsourcing?
- Industry of placement of temporary help agency workers
- Consistency of classification systems

Employment in temporary help services is considered a leading indicator for total nonfarm employment. However, one of the largest gaps in our data is not knowing the industry placement for workers in this industry.



Bureau of Labor Statistics, Current Employment Statistics survey, May 08, 2015.
 Shaded area represents recession as denoted by the National Bureau of Economic Research.
 Most recent 2 months of data are preliminary.

Contact Information

Michael W. Horrigan
Associate Commissioner
Office of Employment and
Unemployment Statistics

202-691-5735
horrigan.michael@bls.gov

