The Role of Federal Statistical Agencies in 2020

- The role of alternative data
  - Types of alternative data and BLS uses
  - Cautions on the use of alternative data sets
  - A ‘draft’ vision for the use of alternative data sets at BLS

- Other vision elements
The role of alternative data

- What should be the BLS’s highest priorities in investing our scarce and declining real resources in terms of the uses of alternative data and techniques?

- For each instance in which we use alternative data in the production of our economic statistics, what are the tradeoffs in terms of data quality and transparency, and are those tradeoffs worth making the investment?
Types of alternative data and BLS uses

- Webscraped data
- Internet search data
- Social network data
- Federal administrative data
- Private vendor data
- Corporate data
- Private sector process data
Webscraped data

- Billion Prices project
  - My initial interest in big data
  - Daily CPIs in 22 countries

- Some BLS uses
  - 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 used and new books.
Internet Search Data

- Google
  - Tools to create large data files that combine publicly available data on social and economic activity stratified by geography, and social-demographic characteristics
  - 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.
  - **No active BLS use of this alternative data source**
Social network data

- Tweets – Matthew Shapiro et al., University of Michigan Study
  - Case study of job loss related tweets that examines the correlation with unemployment data to predict initial claims
  - No active BLS use of this alternative data source
Federal Administrative Data

- Sampling frames used by statistical agencies for drawing stratified probability samples and in the construction of estimation weights
  - Cross agency use of sampling frames for drawing samples
- Use of administrative data for imputation and benchmark revisions
- Use of administrative data for estimation replacing direct data collection
- Linking administrative data to other administrative data and surveys
QCEW Hurricane maps

- Combines detailed QCEW on total employment, total wages, and the count of establishments with flood zones (geographical areas) that have been created by the U.S. Corp of Engineers and State emergency management authorities. These maps are now on the BLS public web site. 
  http://www.bls.gov/cew/hurricane_zones/home.htm
Private Vendor data

- Stock Exchange Security Trades - PPI
- JD Power - CPI
- Scanner data: Homescan, Nielson - CPI
- Health Claims data – PPI, CPI
- Credit card data - BEA
Corporate data: BLS uses

- 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

- More generally, corporate data may take the form of data extracts from company data systems that are not translated into BLS formats.
  - Example: OES collection
Corporate data: BLS uses

- CPI is also examining the potential of using corporate data records.
  - Matched model requirements or some version of unit value pricing
  - Difficulty in capturing quality change
  - Actual recorded transactions, including all coupons and discounts
  - Processing challenges associated with large volumes of data
  - Potential for larger samples than from original sampling draw
Private sector process data

- **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.
Private sector process data

- GE
  - 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.

No active BLS use of these alternative data
The Role of Federal Statistical Agencies in 2020

- The role of alternative data
  - Types of alternative data and BLS uses
  - **Cautions on the use of alternative data sets**
    - A ‘draft’ vision for the use of alternative data sets at BLS

- Other vision elements
Some Cautions

- A natural question that arises in considering the use of alternative data sets is to ask, to what extent does the use of alternative data bring us into conflict with these goals?
- The previous section, however, shows that we have already made the choice of using blended data.
  - We must produce and maintain transparent methodological documentation in our use of blended data sources.
Some Cautions

- One of the biggest challenges in using alternative data is in knowing (or not knowing) the relationship between the scope of alternative data and how it relates to the target population under study.

- In the cases where the alternative data does not represent a census or universe of units or transactions, do we have sufficient information to determine their weights or relative importances in the construction of estimates? Under what circumstances do we decide to use or not use such data?
Some Cautions

- At what level of aggregation do we use alternative data?
  - Surveys for top side
  - Ratio allocation
  - MSE criterion

- Finally, under what conditions is it not appropriate legally or by statistical principle to use alternative data sets?
  - In the special case of webscraping, does BLS need to seek permission from the web sites we scrape for the purposes of collecting data?
The Role of Federal Statistical Agencies in 2020

- The role of alternative data
  - Types of alternative data and BLS uses
  - Cautions on the use of alternative data sets
  - A ‘draft’ vision for the use of alternative data sets at BLS

- Other vision elements
A draft ‘vision’ for the use of alternative data at BLS

- Linking
- Electronic data collection
- Acquiring alternative data to replace direct data collection
- Webcrawling
Linking

- Linking across BLS establishment data sets to the QCEW or other Federal administrative data bases has been underutilized
  - QCEW (9 million) and OES (1.2 million over 3 years)
    - Example: OES as a times series
    - Examination of occupations with rising wages and employment by industry employment growth and further stratification down to the MSA level (or lower using modelling)
  - Similar linkages of QCEW to other BLS establishment data bases
Linking

- Linkages of QCEW to other statistical agency’s establishment data bases
  - Custom Bureau sampling frame for exports matched to the QCEW
  - Currently IPP gets export trade volumes from the Custom Bureau for sampled units – extend to all units?

- PPI use of Census establishment frames to draw samples based on product revenue
  - Current research using multi-establishments
  - Extension to small firms with CIPSEA amendments to allow access to IRS data
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%.
Electronic data collection

- Move beyond our current approach to collecting electronic records from firms using our survey forms through the EDI center or the BLS Internet Data Collection Facility
- Allow firms to report using their formats and data bases
- Using autocoding learning models and computational linguistics to convert firm based data and classifications to BLS concepts
Acquiring alternative data sets for use in estimation

- Acknowledging the need to develop statistical approaches to blending data, there are a lot of opportunities for acquiring alternative data sets that remain.

- Employment
  - Supplement JOLTS data on vacancies with job openings data from private vendors (Snagajob, Burning Glass, Career Builder)
Acquiring alternative data sets for use in estimation

- **Productivity**
  - Truven Health Analytics data for health care productivity measures
  - American Short Line and Regional Railroad Association data for the potential development of productivity measures for Short Line railroads (and complete coverage for Rail Transportation);
  - Data from Compustat to potentially produce State level productivity estimates;
Acquiring alternative data sets for use in estimation

- Safety and Health
  - Data from National Institute for Occupational Safety and Health that combines fatal highway accident data with the Census of Fatal Occupational Injuries to obtain greater detail about these fatalities;
  - OSHA is proposing a regulation to require web-based reporting of injuries and illnesses, which would create a new web based source of administrative data on these concepts.
Acquiring alternative data sets for use in estimation

- Prices
  - Use of credit card data collected by BEA to potentially use to create travel and tourism price indexes.
  - Use of secondary source data on education to develop import and export education price indexes.
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
Other elements of a vision

- Quick response surveys
- Classification systems
- Global supply chains and corporate data
- Geography
Contact Information

Michael Horrigan
Associate Commissioner
Office of Employment and Unemployment Statistics

www.bls.gov
202-691-6400
horrigan.michael@bls.gov
What are “Big Data”? 

- Consultants say three quintillion bytes of data are created every day.
- It comes from everywhere. It knows all.
- According to the book of Wikipedia, its name is "Big Data."
- Big data lives in the cloud. It knows what we do.
- In the past, our company did many evil things.
- But if we accept big data in our servers, we will be saved from bankruptcy.
- Let us pay.
- Is it too late to side with evil? Shhh! It hears you.