

BIG Data

Ron S. Jarmin
Assistant Director for Research and Methodology
U.S. Census Bureau

Talk less about “Big Data” and more about using technology to modernize how Census collects, processes and disseminates data to improve economic and social measurement.

Need to modernize

- o Costs
 - § Both to reduce costs, but also to
 - § Free up resources to expand and improve the data we provide to our users
- o Declining response rates
- o Competitive challenge from alternative sources of information
- o Possibly co-equal opportunities
- o Our users demand new data products and that data be more timely, available for smaller domains, and incorporate new or substantially modernized items.

How to modernize?

- o Technology offers many opportunities – sometimes referred to as Big Data
- o Many new sources and means of capturing raw source data with which to construct economic and social statistics:
 - § New sources of administrative data (e.g., real estate records)
 - § Social media
 - § Sensor data (e.g., traffic patterns, commodity flows)
 - § Passive collections (e.g., APIs from large firms and organizations, QuickBooks from smaller ones)
- o These newer sources free us from the constraints of paper survey forms
 - § But unstructured (or at least less structured) data pose challenges, including the challenge of developing at least an approximate sampling frame that will support weighting adjustments or melding with traditional survey data.
- o Move from admin data that supports surveys to surveys that support admin/non-survey data

To successfully employ these new opportunities to modernize, we need to make progress on the following distinct but inter-related components:

- o Methodological - how to produce scientifically valid estimates and uncertainty measures of economic and social statistics from data collected from a wide variety of sources, most of which were not designed to produce inputs to the production of official statistics. Empirical and methodological research on bias, variance, and total survey error, will be necessary to support this requirement.
- o Computational - how to develop the hardware and software infrastructure to compute and disseminate statistics constructed from a variety of sources including surveys, administrative sources, transaction data, social media, sensors, and so on.
- o Policy - how to secure legal permissions and stakeholder buy-in to utilize non-traditional sources of data for the production of official statistics. Requirements include legal

agreements with data providers, and engaging the complete set of stakeholders in the legal and privacy space in a transparent way to ensure all understand the cost, benefits and risks of expanding the capabilities of this next-generation federal statistical system.

- o Outreach and marketing - the data products produced employing new data sources and techniques will differ from traditional survey-based statistics. Users will need to be satisfied that these new statistics actually accurately measure the phenomena we intend them to and may in some cases users may need to be educated in how to properly draw inferences from estimates constructed in novel ways.

What are we doing to make progress on these issues and modernize ?

- o External collaborations are key
 - § NCRN
 - § RDCs
 - § Other academic collaborations – e.g., MIT, Georgetown, VT, Stanford, Chicago, AIR
 - § Private sector – e.g., Google, ESRI, UPS, First Data
 - § Other agencies – FRB, NIST, OSTP, USPTO
- o Training existing staff
 - § Census / U of Chicago Big Data Class
- o Recruiting new staff with the right skill sets
- o CEDCap and CEDSi
- o Projects –
 - § 2020
 - Admin records
 - Field reengineering
 - § BDS/Patent data linkages
 - § New pilot project with AIR and Big 10 universities to measure downstream impact of federally funded research with “fat pipe” of information on grants.

Expected Outcomes

- o Enhanced understanding of the trade-offs in using all data
 - o Identification of important enhancements to data that will improve quality, for example enhanced meta and para data for non-traditional sources
 - o Changed business processes and staff development
 - o Modernization of hardware, software and personware
 - o Maintenance or enhancement of the role of federal statistical agencies and their data products
-