Informing Decisions while Protecting Privacy: The Future of the Federal Statistical System

Katharine G. Abraham
University of Maryland, NBER and IZA

Data Science for the Public Good Forum

September 17, 2019

It is the best of times, it is the worst of times…

- Hard to overstate importance of information produced by Federal statistical agencies for understanding our economy and society
- Observers often emphasize system's challenges … but also reason for optimism about new opportunities
- One notable event: Release of the final report of the Commission on Evidence-Based Policymaking two years ago this month
  - Commission grew out of bipartisan interest in better using data Federal government holds while respecting rights to privacy and confidentiality
  - Anniversary of Commission's report a good occasion to take stock of the statistical agencies and where they're headed

Challenges to “business as usual”

- Cost and quality of information based on surveys
- Privacy and confidentiality in a data-rich world

Federal statistics derive largely from surveys

- Much of the information produced by the federal statistical agencies comes from surveys of households and businesses. Some examples:
  - Poverty
  - Health insurance coverage
  - Crime victimization
  - Employment and unemployment
  - Wage rates and annual earnings
  - Retail sales

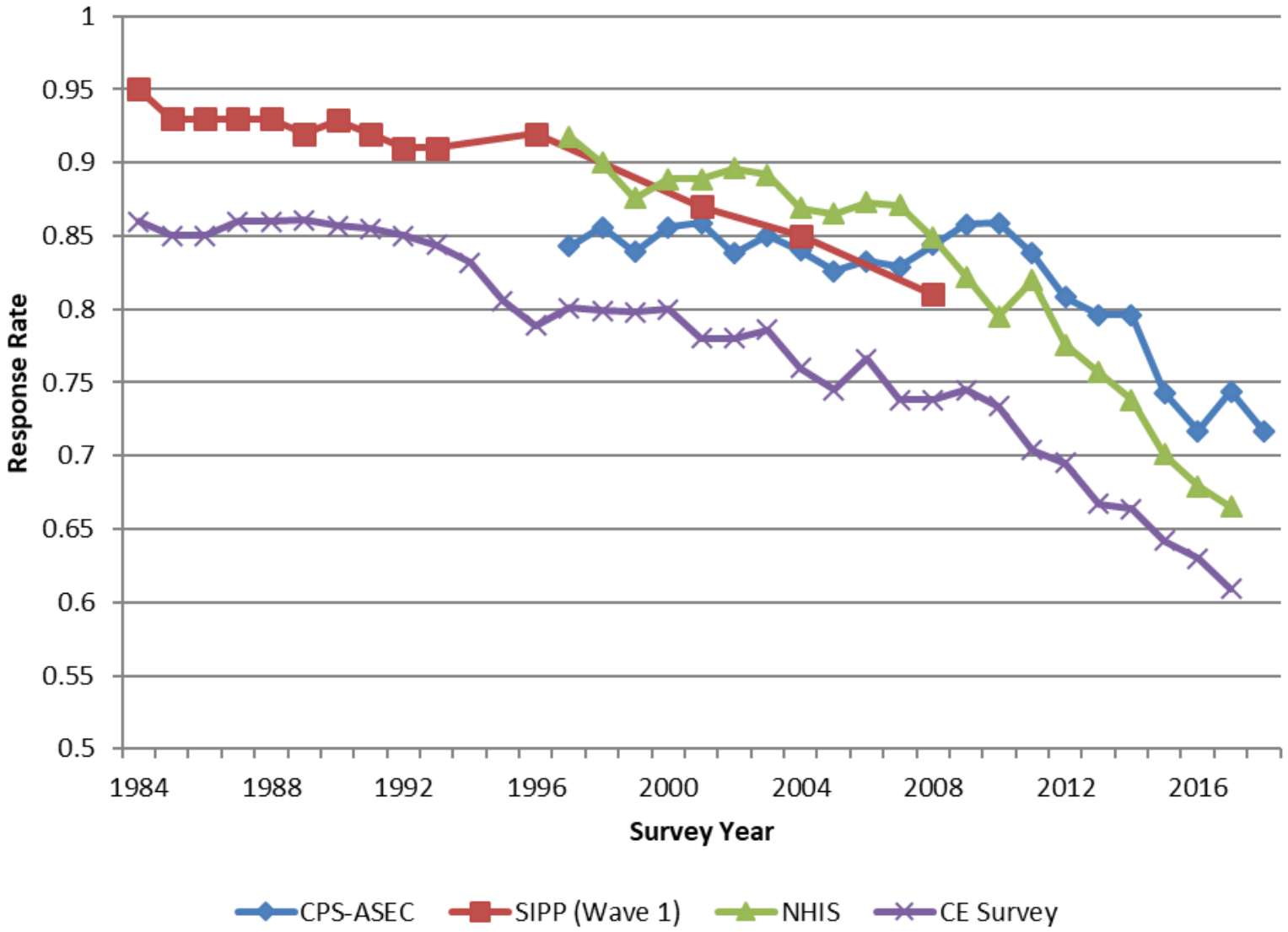
Survey model many strengths

- Methodology is transparent
- Results can be generalized
- Can ask exact questions needed to obtain desired information
  - Consistent questions should produce consistent estimates over time
- Rules for privacy and confidentiality are well developed
  - Respondents provide information under a pledge of confidentiality (though understanding of what it means to honor that pledge is evolving)

Pressures on survey model for data collection

- Increasingly difficult to obtain survey responses
- Growing concern about quality of information supplied by household survey respondents
  - Respondents less motivated?
- Increasing demand for more timely and more disaggregated data
  - Size of survey samples limits detail in published estimates
- Tightening agency budgets

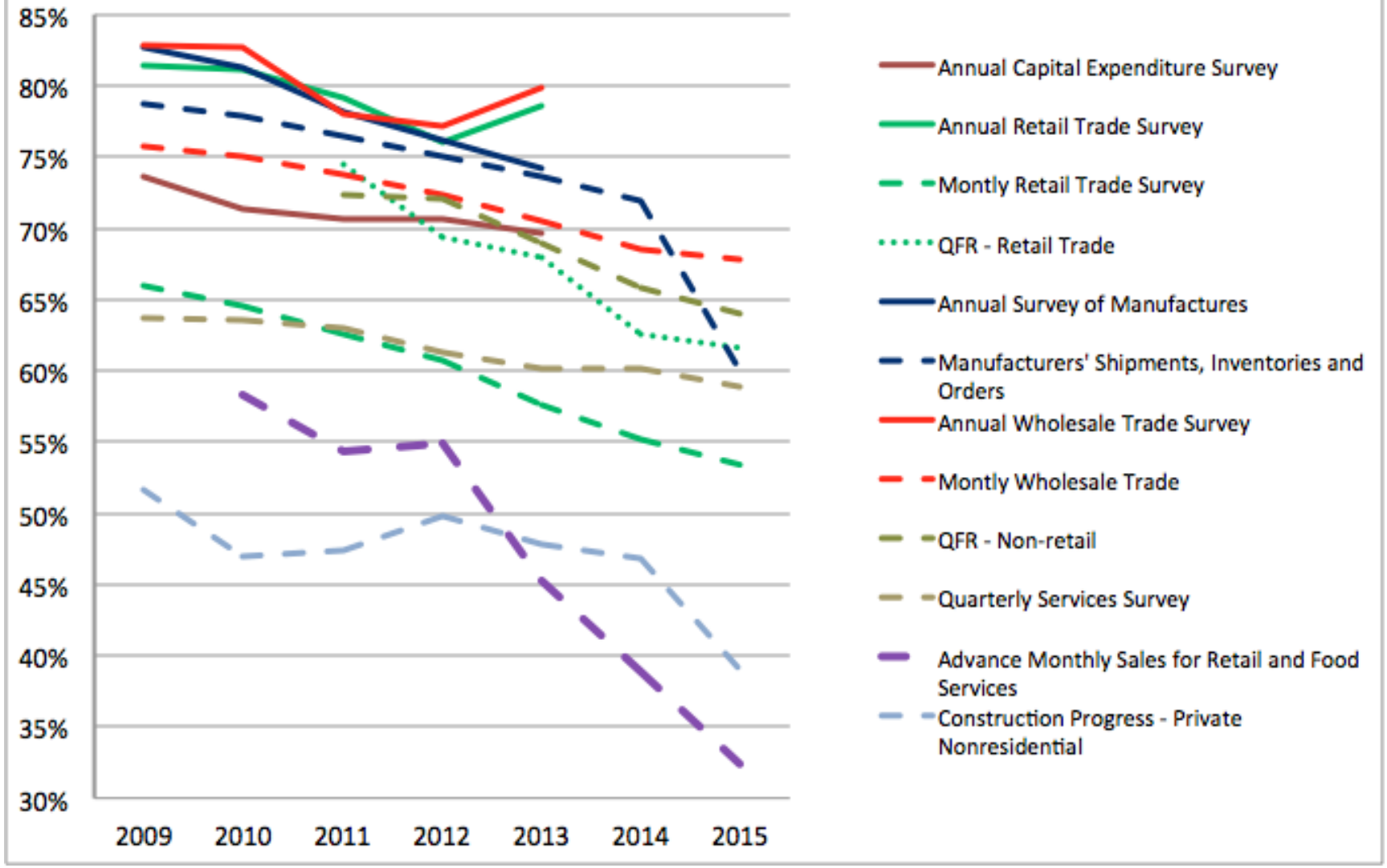
Unit response rates, selected household surveys



Source: Meyer, Mok and Sullivan (2015), adapted and updated

Economic Directorate Survey Unit Response Rates 2009-2015

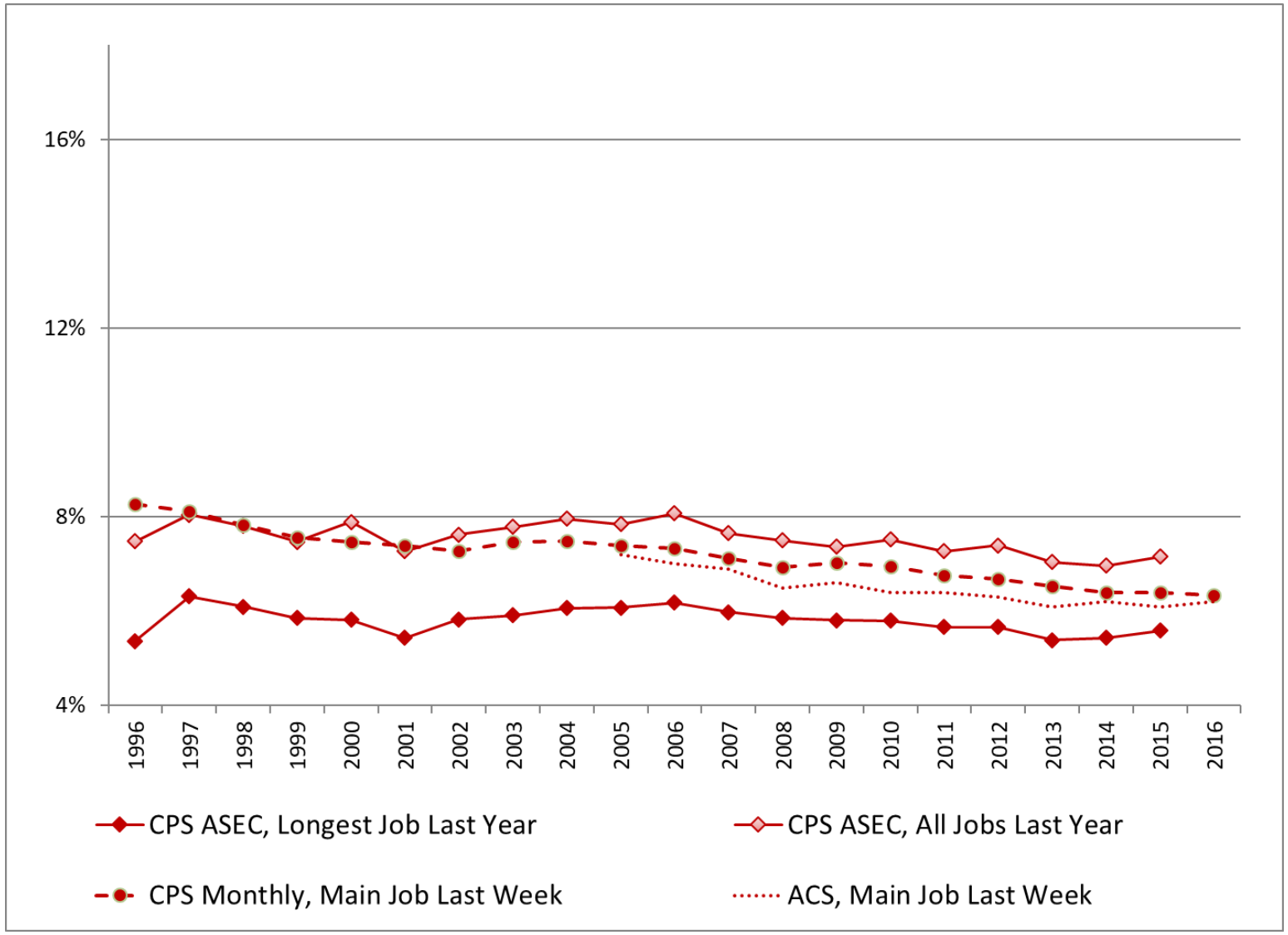
Annual(solid lines), Quarterly and Monthy (dashed lines). 2015 estimates based on data available so far this year.



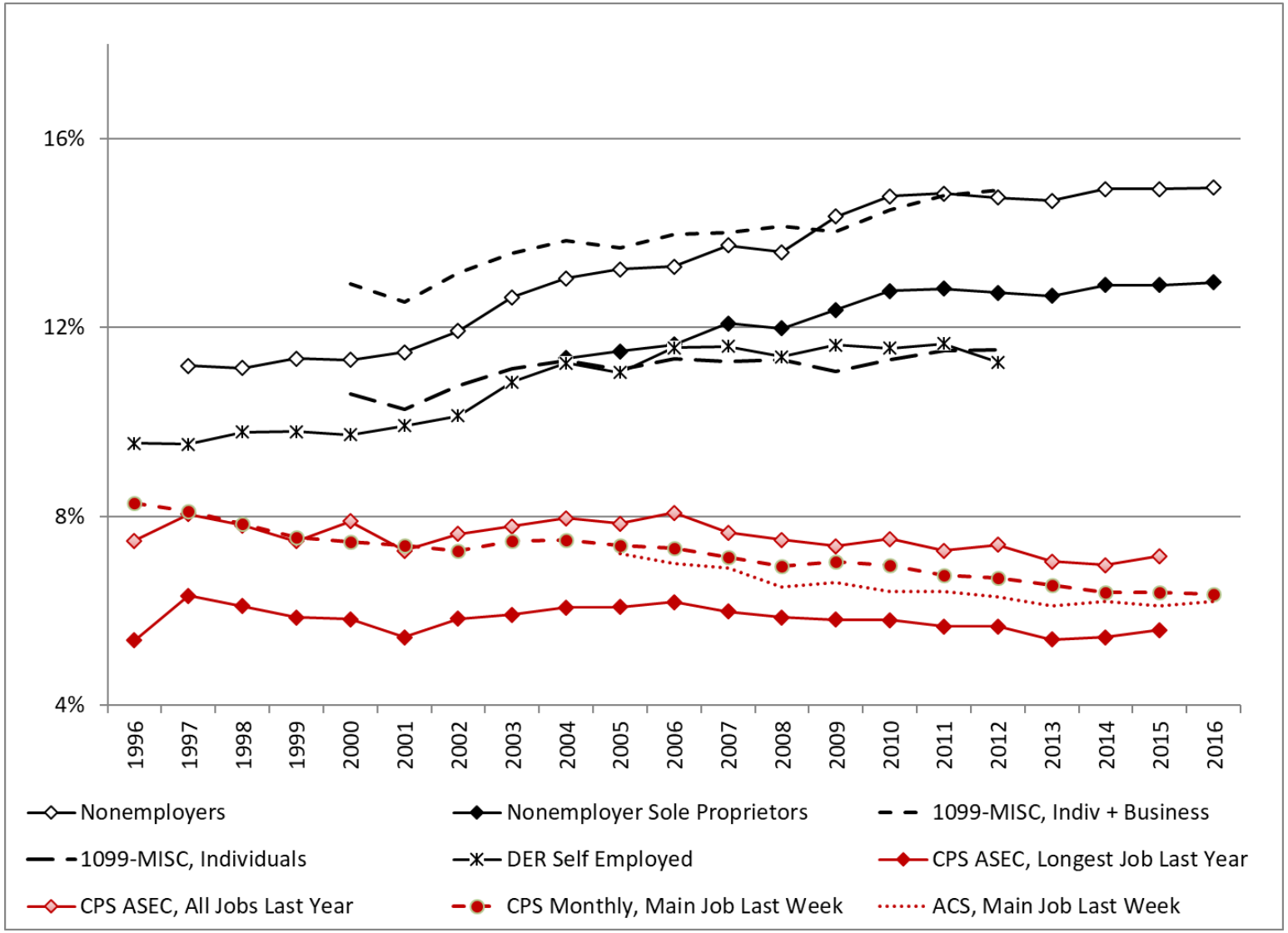
Pressures on survey model for data collection

- Increasingly difficult to obtain survey responses
- Growing concern about quality of information supplied by household survey respondents
  - Respondents less motivated?
- Increasing demand for more timely and more disaggregated data
  - Size of survey samples limits detail in published estimates
- Tightening agency budgets

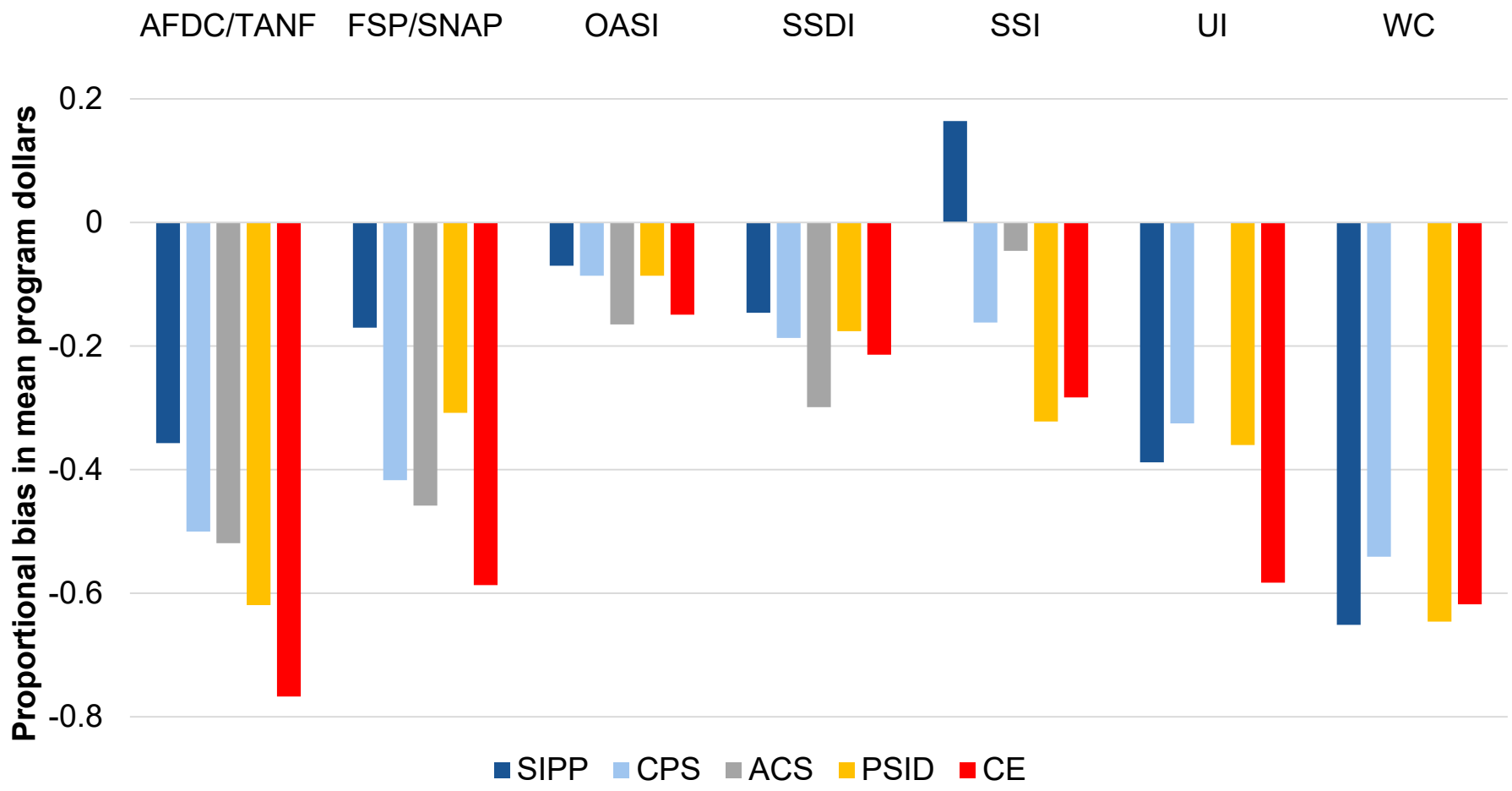
Surveys show flat or declining self-employment…



… but tax data show rising self-employment



Surveys understate income from government programs



Source: Meyer, Mok, and Sullivan (2015), by program and survey, 2000-2012

Pressures on survey model for data collection

- Increasingly difficult to obtain survey responses
- Growing concern about quality of information supplied by household survey respondents
  - Respondents less motivated?
- Increasing demand for more timely and more disaggregated data
  - Size of survey samples limits detail in published estimates
- Tightening agency budgets

Pressures on survey model for data collection

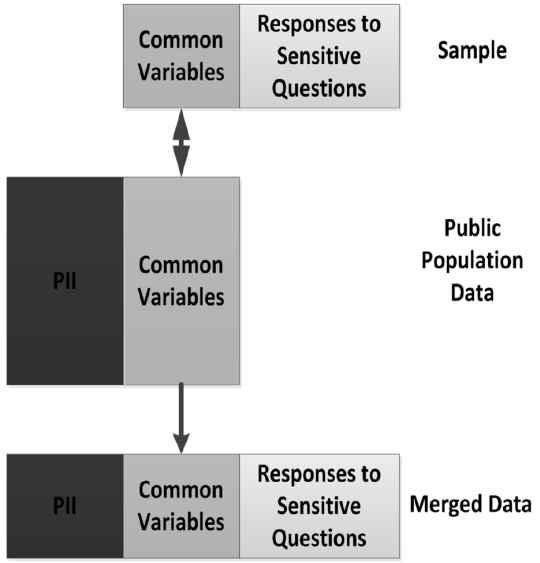
- Increasingly difficult to obtain survey responses
- Growing concern about quality of information supplied by household survey respondents
  - Respondents less motivated?
- Increasing demand for more timely and more disaggregated data
  - Size of survey samples limits detail in published estimates
- Tightening agency budgets

- Cost and quality of information based on surveys
- Privacy and confidentiality in a data-rich world

Statistical disclosure risks: Microdata

- Direct identifiers removed from statistical agency public use microdata
- Based on even small number of characteristics, many people unique in population
  - Example: In 1990, 87% of U.S. population had reported characteristics that likely made them unique based only on 5-digit ZIP2, gender, date of birth (Sweeney 2002)

Statistical disclosure risks: Microdata (continued)



- Disclosure may occur if variables on sample file can be matched to same variables in public records or other accessible information
  - Example: Identification in data released by Massachusetts Group Insurance Commission of hospital records for Governor William Weld, based on sex, date of birth and zip code linked to voter records
- Data breaches that increase amount of publicly available information increase risk of a disclosure

Source: Krenzke and Li (2019)

Statistical disclosure risks: Tabular data

- Allowing multiple queries against an underlying database may disclose individual information
  - Example: Query tool may preclude reporting for samples that are too small, but results that are individually acceptable may reveal information about smaller implicit samples
- Publishing multiple tables also may cause problems
  - More than 7.7 billion linearly independent statistics—or about 25 statistics per person—published from 2010 Census data
  - Can show possible to infer information about individuals through comparisons across tables (Garfinkel, Abowd and Martindale 2018)

Response to challenges to “business as usual”

- Increased use of administrative data
- Rethinking data release and publication

What are administrative data?

- Administrative records contain information collected for purpose of administering government programs.
- Some examples:
  - Income tax returns (household and business)
  - Unemployment insurance wage records
  - Social assistance program applications and benefit receipt histories (e.g., TANF, SNAP, housing assistance)
  - Social Security and Medicare records
  - School records
  - Customs declarations

Potential benefits to increased use of administrative data

- More accurate estimates
- More disaggregated estimates
- Lower respondent burden
- Lower cost (maybe)

Barriers to increased administrative data use

- Legal barriers
  - Census Bureau authorized to obtain administrative data
  - Other statistical agencies do not have same authority
- Lack of existing partnerships between statistical and administrative agencies
- Federal program data often collected and held by states

Is increased use of administrative data consistent with protecting privacy and confidentiality?

- Administrative data subjects have not given explicit permission to use their information for statistical purposes
- Ethical use of administrative data (Hart and Wallman 2018)
  - Transparency in use of data
  - Opportunity for public comment
  - Ensure that data releases do not reveal information about individuals

- Increased use of administrative data
- Rethinking data release and publication

Formal privacy protection methods

- Agencies take pledge to protect data subjects' confidentiality very seriously
  - For microdata: Coarsening categorical variables, top-coding continuous variables, noise infusion, data swapping
  - Tabular releases: Cell suppression (Swiss cheese tables), noise infusion and data swapping in underlying microdata, cell value rounding
- Existing methods neither guarantee protection of confidentiality nor optimize usefulness of information reported
- Differential privacy a formal method for quantifying risk of information disclosure associated with a data release
  - Measure pertains to most vulnerable case in data
  - Risk controlled by adding noise to output data

Drivers of change

Commission on Evidence-Based Policymaking

- Legislation to establish Commission jointly sponsored by House Speaker Paul Ryan (R-WI) and Senator Patty Murray (D-WA)
  - Signed into law March 30, 2016
- Key elements of Commission's charge:
  - Determine optimal arrangement under which administrative data, survey data, and related statistical data series may be integrated and made available for evidence building while protecting privacy and confidentiality.
  - Consider whether a clearinghouse for program and survey data should be established and how to create such a clearinghouse.
  - Make recommendations on how best to incorporate evidence building into program design.

Commission on Evidence-Based Policymaking (cont'd)

- Members appointed by the President, Speaker of the House, House Minority Leader, and the Senate Majority and Minority Leaders – 1/3 experts on privacy; 2/3 experts on program administration, data, or research
- Commission engaged in extensive fact-finding process, considered input received and distilled areas of agreement into 22 recommendations
  - Recommendations endorsed by all 15 Commissioners
- Report provided to President and the Congress on September 7, 2017

Major themes: Report of the Commission on Evidence-Based Policymaking

- Improved Access to Data – Improve access by administrators and researchers to data and facilitate linking of data sets
- Stronger Privacy Protections – protections today applied unevenly across government and not sufficiently dynamic in face of changing risks associated with use of data
- Greater Capacity – filling the existing capacity gaps across institutions and actors inside and outside government, including the establishment of a single entity to better support access and privacy

Foundations for Evidence-Based Policymaking Act

- In October 2017, proposed legislation based on Commission's recommendations co-filed in House by Speaker Ryan and in Senate by Senator Murray
  - Passed quickly through the House
  - Voted out of Senate on December 19, 2018
  - Law signed by President Trump on January 14, 2019
- Provisions address 11 of the Commission's 22 recommendations

Foundations for Evidence-Based Policymaking Act: Data access

- Directs agencies to develop inventories of data they hold
- Clarifies that, unless there is a legal prohibition, data assets can be made available to statistical agencies for use in building evidence (i.e., for statistical purposes)
- Establishes an Advisory Committee on Data for Evidence Building to make recommendations regarding the coordination and availability of data
- Directs OMB to establish a common application process for researchers, state and local governments and other entities to access data for evidence-building purposes

Foundations for Evidence-Based Policymaking Act: Privacy and confidentiality

- Designates a Chief Data Officer at each agency who will coordinate the management and governance of data at each agency in collaboration with the agency's statistical officials
- Reauthorizes the Confidential Information Protection and Statistical Efficiency Act of 2002
  - Protects information collected for statistical purposes
  - Violation a Class E felony (5 years in prison and/or $250,000 fine)
- Codifies OMB Statistical Policy Directive No. 1
  - Establishes responsibilities for statistical agencies
- Requires comprehensive risk assessments prior to data releases and analyses of data sensitivity

Foundations for Evidence-Based Policymaking Act: Evidence-building capacity

- Requires agencies to develop evidence-building plans (i.e., learning agendas)
- Requires agencies to designate a Chief Evaluation Officer

Federal Data Strategy aligns with Evidence Act

- Mission statement: “The mission of the Federal Data Strategy is to fully leverage the value of federal data for mission, service and the public good by guiding the Federal Government in practicing ethical governance, conscious design and learning culture.”
  - Ethical governance includes building in checks and balances; practicing effective data stewardship, protecting individual privacy, maintaining promised confidentiality, and ensuring appropriate access and use; and promoting transparency.
  - Conscious design includes protecting data quality and integrity; harnessing existing data; anticipating future uses when new data collections are designed; and demonstrating responsiveness.
  - Learning culture includes investing in data infrastructure and human resources; developing data leaders; and practicing accountability.
- Principles and practices issued June 4, 2019; action plan for first year to be issued Fall 2019

Looking to the future

Fuller implementation of Commission's recommendations

- Foundations for Evidence-Based Policymaking Act a *great* first step towards making better use of administrative data while protecting privacy
- Does not address all key recommendations of the Commission on Evidence-Based Policymaking. Still to be addressed
  - Legal barriers that preclude legitimate statistical uses of administrative data
  - Institutional capacity within the federal government to facilitate data linkages and drive implementation of new privacy protection methodologies
    - Expect Advisory Committee on Data for Evidence Building to make specific recommendations about how to do this

Private sector “big data”

- Next frontier with respect to alternative data sources: Private sector “big data”. Some examples:
  - Prices and product characteristics posted to the Web
  - Scanner data from retail outlets
  - Credit card transactions data (e.g., JP Morgan Chase data, Spending Pulse MasterCard data)
  - Medical records data
  - Sensor data (e.g., satellite imaging, traffic cameras)
  - GPS tracking data (e.g., tractors, trucks)

Private sector “big data” (continued)

- Many potential benefits to expanding uses
  - Fill in missing information (e.g., industry, franchise status)
  - Improve early estimates by providing timely information about recent trends
  - Inform modeled estimates for local geographies
  - More ambitiously, allow agencies to rethink how core estimates produced
    - Integration of data on prices and quantities

Private sector “big data” (continued)

- Some concerns for agencies
  - Cost of acquiring data
  - Suitability for use in producing official statistics
    - Concerns about availability and consistency over time
    - Need to document and archive non-designed data (transparency)
  - Privacy and confidentiality
- Will need private sector partnerships and a more coordinated or centralized agency approach
  - Inefficient and possibly counterproductive for agencies all to be developing separate relationships with private sector data providers

New models for data access and publication

- Release of many public use files unlikely to be sustainable
  - Expect new model to involve tiered access together with expanded capacity for external researchers to work behind(virtual) firewall
- Census Bureau plans to use differential privacy in publication of results from 2020 Census
  - Expect use of differential privacy to spread
- Need help from academia and private sector with developing methods
- Important *policy* questions related to appropriate tradeoff between privacy and information

Conclusion

- Strong imperative for statistical agencies to update methods used to produce the data their customers need
  - Better data
  - Stronger privacy and confidentiality protections
- Foundations for Evidence-Based Policymaking Act and Federal Data Strategy are exciting developments
  - Creating new opportunities
- Much more remains to be done!