NATIONAL ACADEMIES *Sciences Engineering Medicine*

Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good – Private Sector Data: An Essential Component

FCSM Session J-4: The Importance of Private Sector Data to Federal Statistics

Thomas L .Mesenbourg Jr , CNSTAT Study Director

CNSTAT

Background

NSF grant to produce 3 complementary reports to help develop a vision for a new national data infrastructure for federal statistics and social and economic research.

- Report 1: The components and key characteristics of a 21st century data infrastructure. (Final report released March 3, 2023)
- Report 2: The implications of using multiple data sources for major statistical programs. (Prepub released March 2023)
- Report 3: Approaches to Data Governance and Protecting Privacy. (Workshops in May 22, 23, & 25, 2023, report later this year)

** Vision is a work in progress, challenges abound, years of work **

Panel on the Scope, Components, and Key Characteristics of a 21st-Century Data Infrastructure

Robert M. Groves (Chair)
Georgetown University

danah boyd
Microsoft Research, Data & Society

Anne C. Case
Princeton University

Janet M. Currie
Princeton University

Erica L. Groshen
Cornell University

Margaret C. Levenstein
University of Michigan

Ted McCann
American Idea Foundation

C. Matthew Snipp
Stanford University

Patricia Solís
Arizona State University

Interpretation of the Charge

- Improve national statistics, social and economic research, and evidence building for the common good.
- Use evidence from previous expert groups, the CEP Report, 2018 Evidence Act, the work of the ACDEB, and international data initiatives.
- December 2021 workshop focused on private sector data use.
- Synthesize evidence into consensus conclusions informing a vision for a new national data infrastructure

Vision

- Stat agencies & approved users:
- Use blended, relevant cross-sector data assets, for soley statistical purposes, to:
  - -Improve the timeliness, granularity, & usefulness of national statistics
  - - Facilitate more rigorous resesearch
  - - Support evidence-based policymaking
- Access & statistical uses comply with existing laws & regs and is governed by established rules & policies
- Operations & decisions guided by explicit values and key attributes
- High trust, transparent environment balances expanded use of multiple data sources with strengthened privacy & confidentiality protection & responsible, ethical data use

Data Infrastructure Components

1. Data assets
2. Technologies
3. Skilled staff & expertise
4. Standards, policies, rules for data access, use, & protection
5. Organizations
6. Communities of interest, data holders & subjects whose data are shared & impacted by uses

PILOT OF INDVSTRY • CONQVEROR OF DISEASE • MVLTIPLIER OF THE H

Attributes

1. Safeguards and advanced privacy-enhancing practices to minimize individual harm
2. Statistical uses only, for common good information, freely shared with all
3. Mobilization of relevant digital data assets, blended in aggregates, providing benefits to data holders, with societal benefits proportionate to risks & costs
4. Reformed legal authorities protect all parties interests
5. Governance framework and standards effectively support operations
6. Transparency to the public about operations
7. State-of-the-art practices continuously improved

Catalysts for Including Private Sector Data

- Evidence Act’s focus on federal data assets is too narrow
- Private sector data are a huge but largely untapped data asset
- Unlike federal & state admin data, statistical uses not limited by statute
- Federal statistical agencies recognize importance of private sector data
- Recent reports, foundations, and professional organizations support using blended private sector data for official statistics
- Other countries recognize the promise private sector data

ENCE PILOT OF INDVSTRY · CONQVEROR OF DISEASE · MVLTIPLIER OF THE H

Challenges and Impediments

Private Sector Data

- Collected for non-statistical purposes, incomplete coverage & documentation
- Nonrepresentative, bespoke, costly, no inherent sustainability, little incentive to share
- Privacy-protecting behaviors variable & unregulated
- FSS private data acquisition, access, and use is siloed, inefficient, & largely uncoordinated
- No cohesive plan for making blended private sector data a national data infrastructure reality

Key Takeaways Related to Private Sector Data

Takeaway

1

U.S. Needs a New National Data Infrastructure

- Credible statistical info key to democracy, just as important as physical infrastructure
- Survey-centric paradigm no longer sustainable (declining response, increasing costs)
- Explosion of alternative data sources including private sector data provides an opportunity to mitigate these threats
- Blending multiple sources can overcome limitations of
- a single source
- Survey data remain an important source, but blending can improve existing products & create new ones

Takeaway

2

Mobilizing the Nation’s Relevant Data Assets

- Need to expand the Evidence Act scope beyond federal; add relevant state, local, tribal govts; private sector enterprises; nonprofits & academic institutions; & others
- OMB evaluating ACDEB recommendations related to federal & state data assets
- Private sector data offers huge opportunities but
- New mechanisms & relationships needed to assess and acquire relevant private sector data assets
- Data acquisition should be limited to what is needed to satisfy pre-specified statistical uses
- Not building a data warehouse; a shared service permits authorized users to conduct temporary linkages for exclusively statistical purposes

Takeaway

3

Incentivizing Data Holders

- Benefits must go beyond improved statistics to reciprocal sharing
- Data sharing provides data holders with direct, tangible benefits informing their operations and activities
- Data holders must understand how they and society benefit from expanded data sharing
- Societal benefits must justify the costs and risks of expanded data sharing
- Benefits, costs, and risks must be quantified

Takeaway

4

Respecting Data Subjects’ Rights and Interests

- Attention to how data use affects data subjects’ lives; minimize harm
- Consider issues of data subjects’ autonomy re how data are used
- Data subjects understand how they benefit from expanded data sharing
- Data infrastructure activities respect data subjects and engages them in matters that impact them
- Data are safeguarded and secure, while privacy is preserved

Takeaway

5

Legal & Regulatory Change Needed

- Current legal framework limits what data can be shared, with whom, and for what purposes
- Implement Evidence Act regs & guidance
- Identify legislation/regulatory priorities regarding CEP state-related recommendations
- Develop legislative strategy for sharing business data among Census, BEA, and BLS
- Identify legal options that would incentivize data holders to share data for statistical purposes.
- Provide uniform privacy preservation, protect confidentiality, ensure autonomy, and prevent data sharing abuses.

Takeaway

6

Transparency & Accountability

- The public, data holders, and data subjects should be able to understand:
  - - how their data are used
  - - by whom
  - - for what purposes, and
  - - to what societal benefits
- Public can express concerns and seek redress
- Ongoing engagement with stakeholders including data holders and data subjects
- Data infrastructure operators are held accountable

Takeaway

7

Multiple organizational structures can support a new data infrastructure, but including private sector data has implication for organizational **options** and **warrant further study and dialogue with the private sector**.

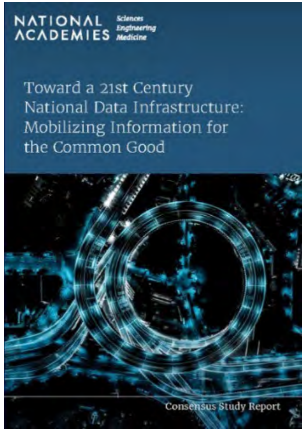
Suggested Action Steps Re Private sector Data

1. Identify private sector data holders’ incentives & disincentives to sharing for statistical purposes
2. Identify legal options that may incentivize data sharing
3. Determine criteria, characteristics, & guidance for deciding what data assets to include in a national data infrastructure
4. Monitor & learn from ongoing stat agency/private sector projects
5. Document costs, risks, and benefits of blending private sector data
6. Assess organizational implications of including private sector data
7. Initiate bi-partisan, multi-sector dialogue identifying options for governing private sector data use for official statististics

Report Resources

Publication and resources

- Report
- Report highlights
- Policy brief
- Private sector data - Issue brief



NATIONAL ACADEMIES Sciences Engineering Medicine

Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good

Consensus Study Report

Interactive site

- Accessible, digestible format
- FAQ section

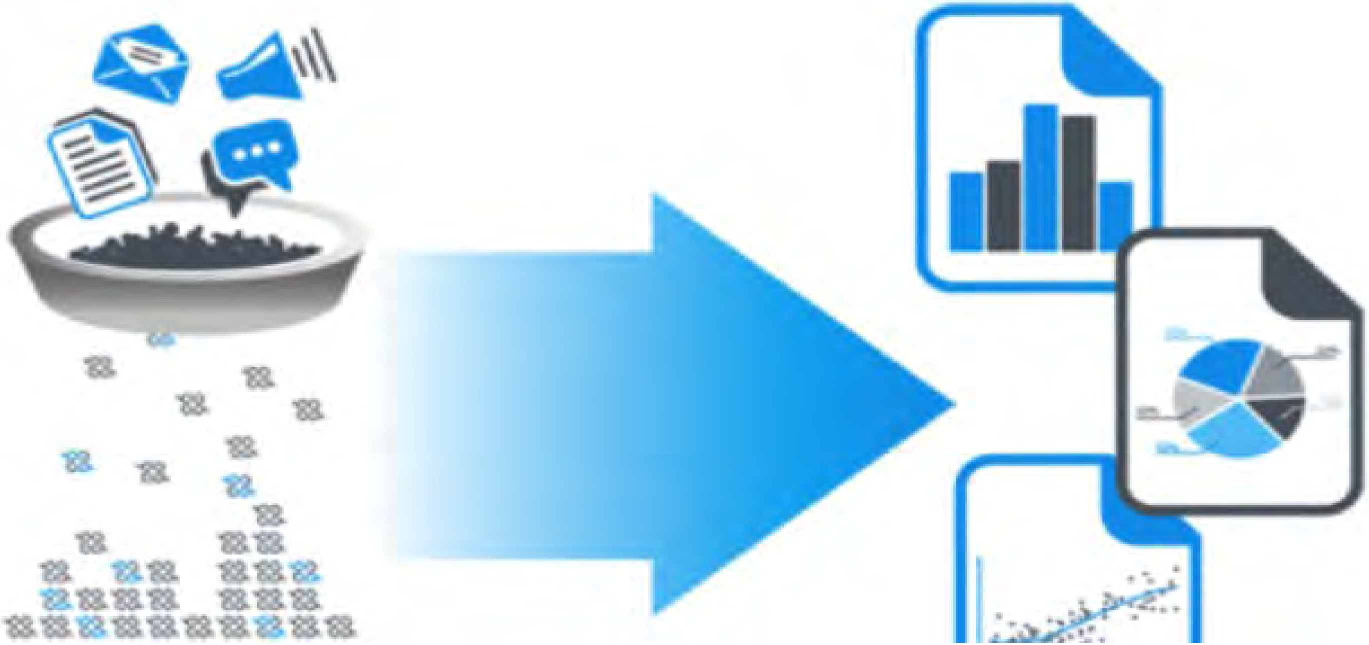
Toward a 21st Century

BLS Use of Private Data for Airline Fares

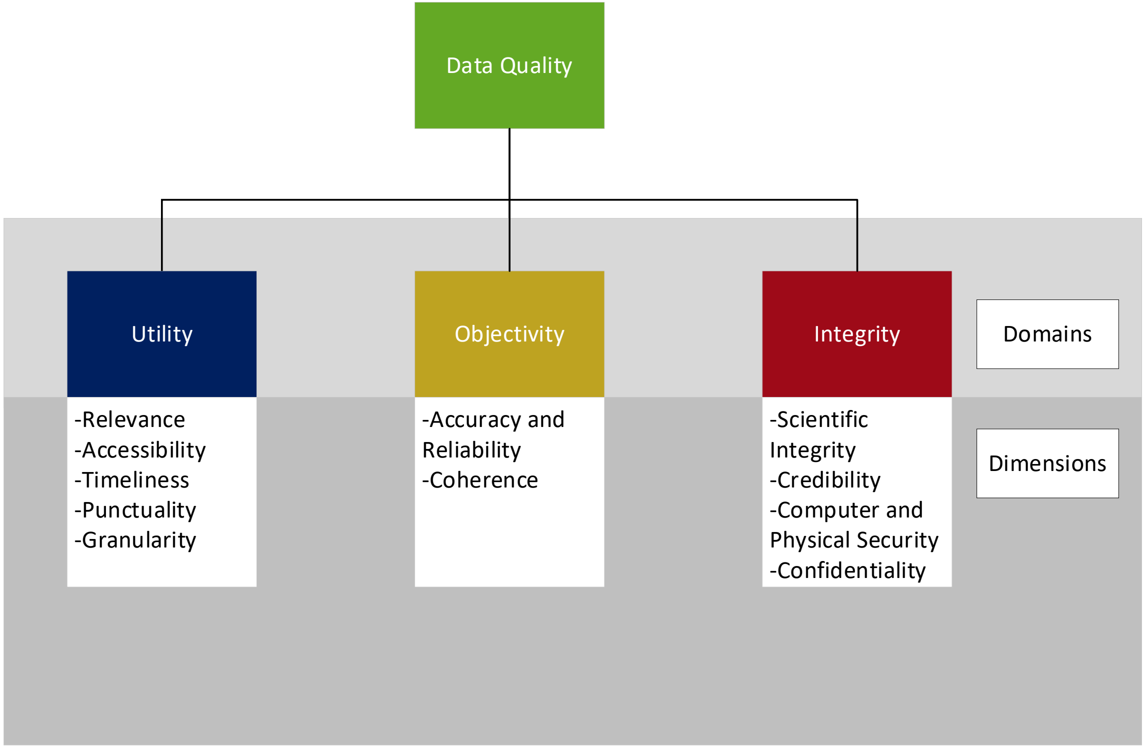
Bill Wiatrowski

Acting Commissioner

Bureau of Labor Statistics



Assessing Quality using FCSM Framework



BLS and Private Data for Airline Fares

Airline Fares in BLS Price Statistics

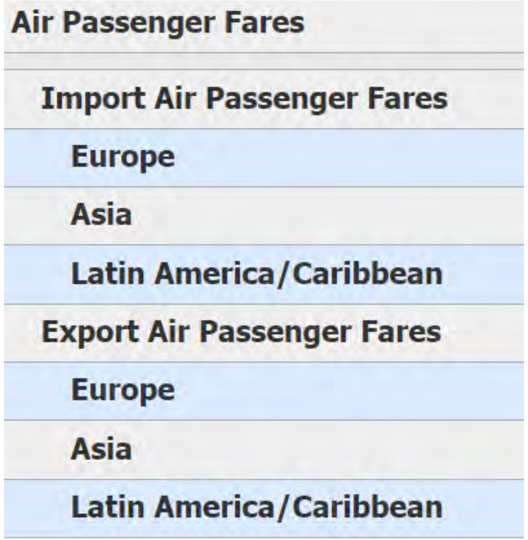
Survey characteristic	Consumer Price Index (CPI)	Producer Price Index (PPI)	Import and Export Price Indexes (MXPI)
Survey definition	Average change in prices paid by consumers for market basket of goods and services.	Average change in selling prices received by domestic producers for their output.	Average change in prices of nonmilitary goods and services traded between the U.S. and the rest of the world.
What is being measured?	Change in the **price that consumers paid** for future travel.	Change in **price that domestic producers receive** for tickets with scheduled departure dates in the reference month.	Change in **price that producers receive for air passenger travel** to and from the U.S. for tickets with scheduled departure dates during the reference month. Prices paid by U.S. residents on foreign carriers to foreign destinations are imports. Prices paid by non-U.S. residents on U.S. carriers between foreign destinations or to the U.S. are exports.
What is included in the measure?	Regularly scheduled domestic and international commercial airline trips.	Regularly scheduled domestic and international commercial airline trips on certified US domestic carriers for travel to and from the U.S.	Fares to and from the United States and U.S. territories and to and from approximately 230 foreign country markets flown on approximately 200 U.S. and foreign carriers.
How is a price defined?	**Base price** of the flight inclusive of taxes, government fees, and one checked bag.	**Average revenue** per passenger including fees (baggage, etc.) and excluding taxes and government fees.	**Average revenue** per passenger excluding fees and including taxes.

Background on Private Data

- Third-party company with a comprehensive database of airline transaction data – actual sales
- Includes ticketing data from U.S. travel agencies and market data from airlines
- Covers approximately over 70% of global ticket sales and over 60% of U.S. ticket sales
- Uses a masking technique to prevent knowing competitor’s prices

MXPI Use of the Private Data

The airline data are used in production and are published at these aggregated levels in the BLS Import/Export Price Indexes (MXPI):



Air Passenger Fares
Import Air Passenger Fares
Europe
Asia
Latin America/Caribbean
Export Air Passenger Fares
Europe
Asia
Latin America/Caribbean

Why Alternative Data for MXPI Airline Fares?

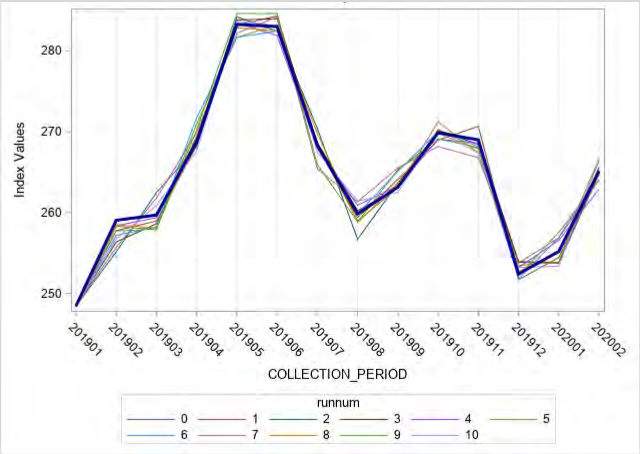
- MXPI previous data source was secondary data
- Price Concept and Time Coverage Improvement
- Index Quality and Coverage Improvement
- Details about airline data in published MXPIs are available here:
  - https://www.bls.gov/mxp/publications/additional-publications/air-passenger-facts.htm

CPI Evaluation

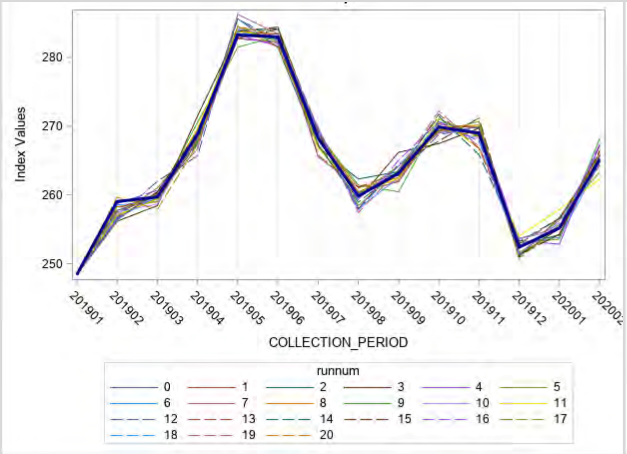
- Research phase
- Investigated the effect of masking prices
  - Designed a Monte Carlo simulation
  - Similar results to the CPI production index values
  - Concluded that the masking process will not be a major deterrent
- Calculating research indexes; looking at changes in methodology

CPI Evaluation: Production & Experimental Index Values

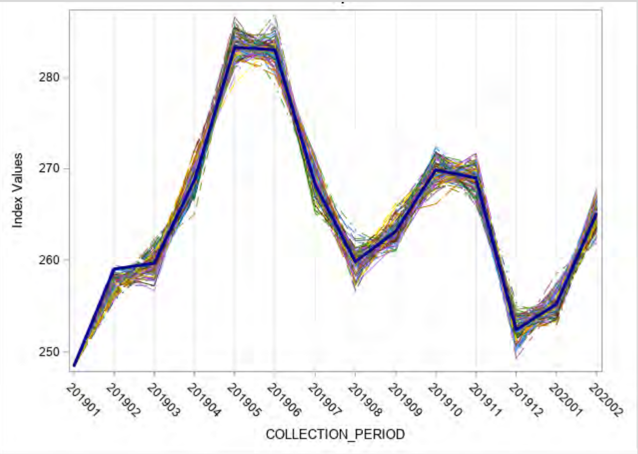
10 Simulation Runs



20 Simulation Runs



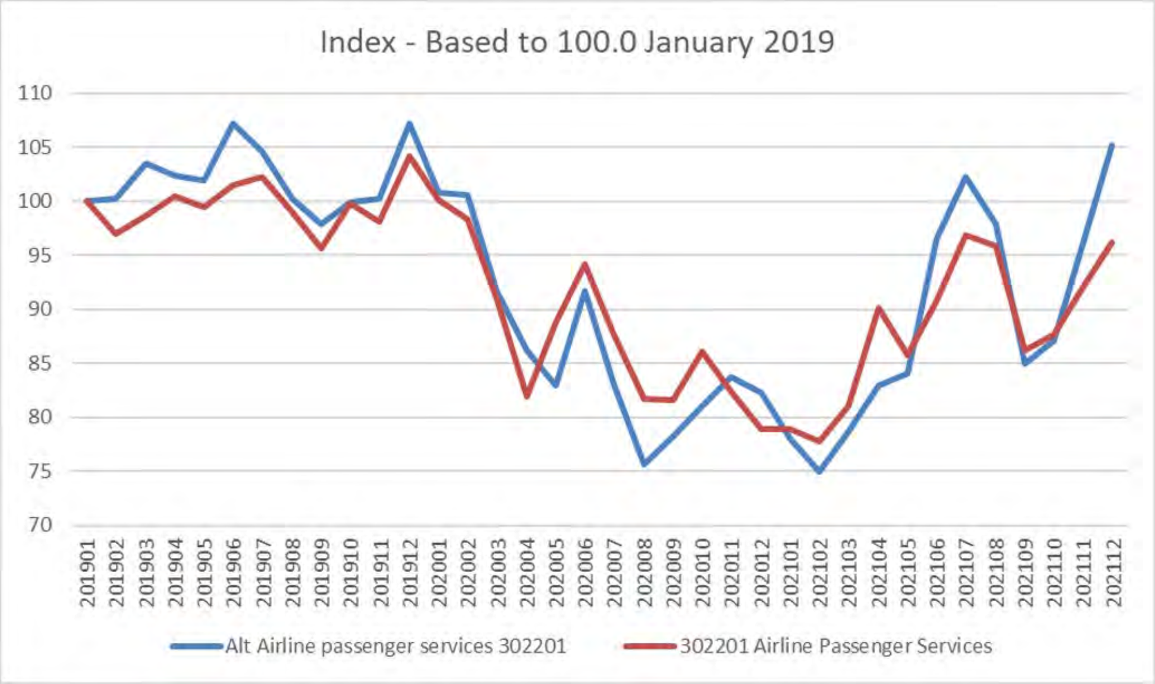
100 Simulation Runs



PPI Evaluation

- Research phase
- Private data provided needs to be in-scope for PPI:
  - US Airline exclusively
  - Based on month and year of departure date
- PPI Counterfactual index calculations replicated Average Revenue per Passenger (ARPP) pricing methodology

PPI Airline Passenger Services: Published PPI compared to Alternative Data



Contact Information

www.bls.gov/cpi

www.bls.gov/ppi

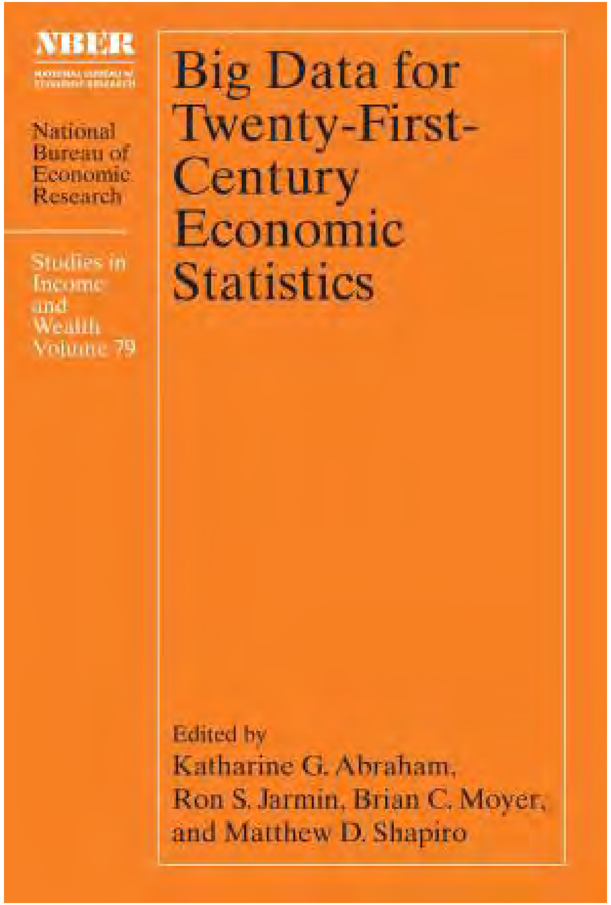
www.bls.gov/mxp

Wiatrowski.William@bls.gov

Modernizing Official Statistics: the Role of Private Sector Data

Ron Jarmin, Deputy Director
U.S. Census Bureau

FCSM
October 26, 2023



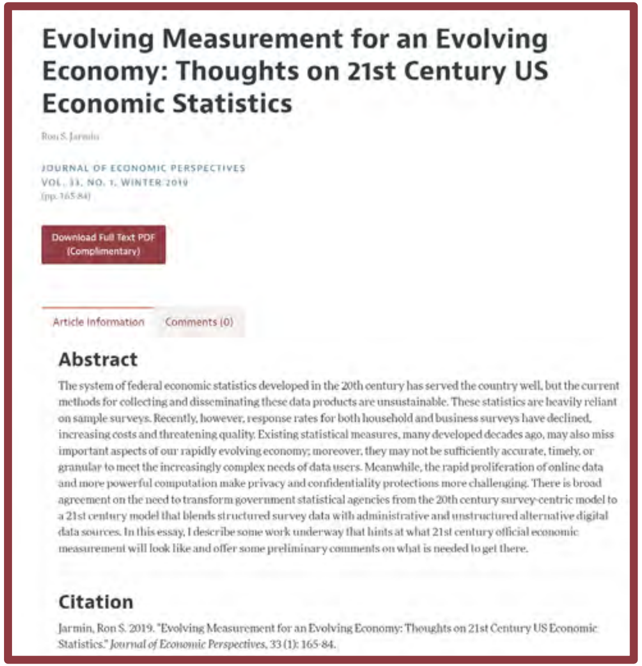
NBER
NATIONAL BUREAU OF ECONOMIC RESEARCH

National Bureau of Economic Research

Studies in Income and Wealth
Volume 79

Big Data for Twenty-First-Century Economic Statistics

Edited by
Katharine G. Abraham,
Ron S. Jarmin, Brian C. Moyer,
and Matthew D. Shapiro



Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics

Ron S. Jarmin

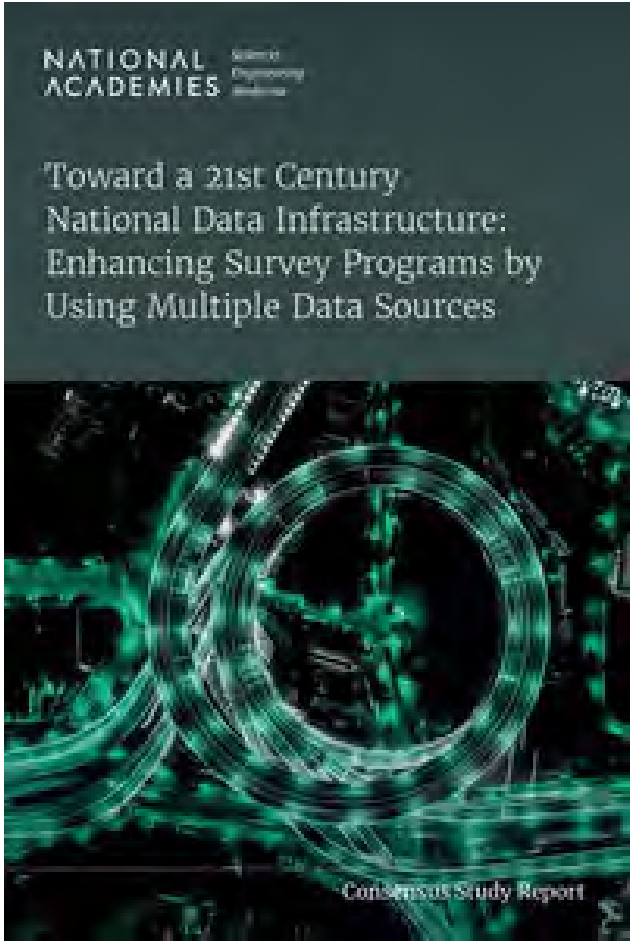
JOURNAL OF ECONOMIC PERSPECTIVES
VOL. 33, NO. 1, WINTER 2019
(pp. 165-84)

Abstract

The system of federal economic statistics developed in the 20th century has served the country well, but the current methods for collecting and disseminating these data products are unsustainable. These statistics are heavily reliant on sample surveys. Recently, however, response rates for both household and business surveys have declined, increasing costs and threatening quality. Existing statistical measures, many developed decades ago, may also miss important aspects of our rapidly evolving economy; moreover, they may not be sufficiently accurate, timely, or granular to meet the increasingly complex needs of data users. Meanwhile, the rapid proliferation of online data and more powerful computation make privacy and confidentiality protections more challenging. There is broad agreement on the need to transform government statistical agencies from the 20th century survey-centric model to a 21st century model that blends structured survey data with administrative and unstructured alternative digital data sources. In this essay, I describe some work underway that hints at what 21st century official economic measurement will look like and offer some preliminary comments on what is needed to get there.

Citation

Jarmin, Ron S. 2019. "Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics." *Journal of Economic Perspectives*, 33 (1): 165-84.



NATIONAL ACADEMIES Sciences Engineering Medicine

Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources

Consensus Study Report

United States Census Bureau

Why Modernize?

We are faced with continuing challenges and new opportunities:

- Declining survey/census response rates
- Increasing survey/census costs
- Increased demand for statistical information in both volume and complexity
- Improved computation
- The proliferation of alternative source data for producing statistics.

United States® Census Bureau

Why Modernize?

We are faced with continuing challenges and new opportunities:

- **Declining survey/census response rates**
- **Increasing survey/census costs**
- Increased demand for statistical information in both volume and complexity
- Improved computation
- The proliferation of alternative source data for producing statistics.

Threats to existing statistical paradigm.

United States® Census Bureau

Why Modernize?

We are faced with continuing challenges and new opportunities:

- Declining survey/census response rates
- Increasing survey/census costs
- **Increased demand for statistical information in both volume and complexity**
- **Improved computation**
- **The proliferation of alternative source data for producing statistics.**

Opportunities to modernize the statistical paradigm.

United States® Census Bureau

Why Modernize?

We are faced with continuing challenges and new opportunities:

- Declining survey/census response rates
- Increasing survey/census costs
- **Increased demand for statistical information in both volume and complexity**
- **Improved computation**
- **The proliferation of alternative source data for producing statistics.**

Opportunities to modernize the statistical paradigm.

Statistical Agencies Around the World Have Taken On This Challenge. But…

- Progress is varied based on several factors.
  - Varied institutional arrangements across countries.
    - Combined or federated statistical units.
    - Access to administrative data.
  - Different relationships with the private sector.
  - Different relationships with academia.
  - Different relationships with the public.
  - Different capacities for change – aka "fixing the plane while its in the air."

United States® Census Bureau

Statistical Agencies Around the World Have Taken On This Challenge. But…

- Progress is varied based on several factors.
    - Varied institutional arrangements across countries.
        - Combined or federated statistical units.
        - Access to administrative data.
    - **Different relationships with the private sector.**
    - Different relationships with academia.
    - Different relationships with the public.
    - Different capacities for change – aka "fixing the plane while its in the air."

Relationships with data providers are key, and difficult.

- We should expect that investments in relationships with private sector providers may need to be as large as those stat agencies have made with administrative data providers.
  - In many cases, these investments can be substantial and require ongoing maintenance. Census-IRS relationship is a great example.
- Examples of the costs and activities associated with these arrangements include:
  - Identifying firms and organizations with data useful for statistical purposes.
  - Negotiating access agreements.
  - Engineering costs to curate non-design data and facilitate statustica. computation.
  - Possibly paying for access to monetized data assets.

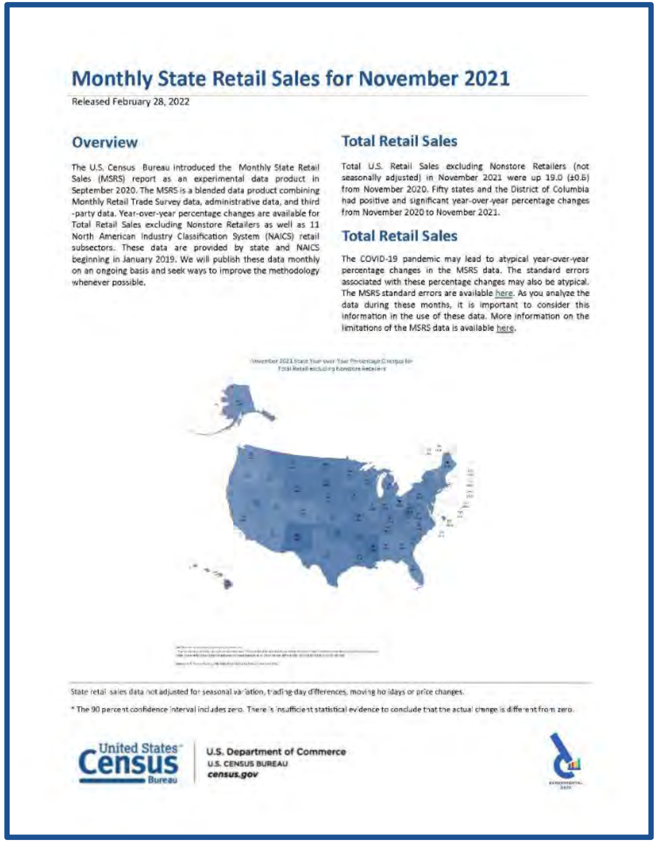
Current Census Bureau Uses of (non-survey) Private Sector Data

- NPD/Circana
- Nielsen
- SAP
- Dodge Data & Analytics
- Imagery (Maxar, Google Earth Engine, Regrid, Lightbox)
- Compustat

United States®
Census
Bureau

Example with NPD/Circana and Nielson: Blended Monthly State Retail Sales

- More timely state-level retail sales are among the most requested data by our data users.
- In September 2020, the Census Bureau released the new blended Monthly State Retail Sales (MSRS) data product.
- First version of these experimental data.
- MSRS was created using existing survey data, administrative data, and third-party/alternative data sources as its inputs. No new data were collected.



Monthly State Retail Sales for November 2021

Released February 28, 2022

Overview

The U.S. Census Bureau introduced the Monthly State Retail Sales (MSRS) report as an experimental data product in September 2020. The MSRS is a blended data product combining Monthly Retail Trade Survey data, administrative data, and third-party data. Year-over-year percentage changes are available for Total Retail Sales excluding Nonstore Retailers as well as 11 North American Industry Classification System (NAICS) retail subsectors. These data are provided by state and NAICS beginning in January 2019. We will publish these data monthly on an ongoing basis and seek ways to improve the methodology whenever possible.

Total Retail Sales

Total U.S. Retail Sales excluding Nonstore Retailers (not seasonally adjusted) in November 2021 were up 19.0 (±0.6) from November 2020. Fifty states and the District of Columbia had positive and significant year-over-year percentage changes from November 2020 to November 2021.

Total Retail Sales

The COVID-19 pandemic may lead to atypical year-over-year percentage changes in the MSRS data. The standard errors associated with these percentage changes may also be atypical. The MSRS standard errors are available here. As you analyze the data during these months, it is important to consider this information in the use of these data. More information on the limitations of the MSRS data is available here.

November 2021 State Year-over-Year Percentage Changes for Total Retail excluding Nonstore Retailers

State retail sales data not adjusted for seasonal variation, trading-day differences, moving holidays or price changes.

* The 90 percent confidence interval includes zero. There is insufficient statistical evidence to conclude that the actual change is different from zero.

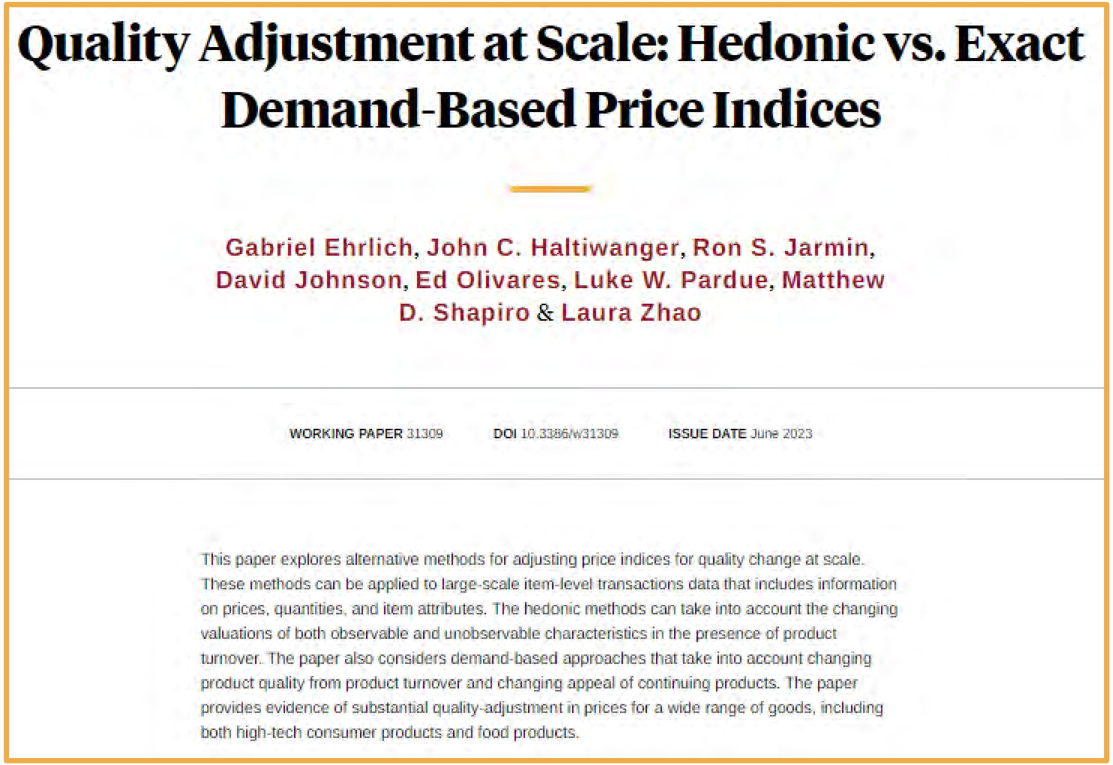
U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Example with NPD/Circana: RESET Project

UNIVERSITY OF MICHIGAN

UNIVERSITY OF MARYLAND 1856

United States® Census Bureau



Quality Adjustment at Scale: Hedonic vs. Exact Demand-Based Price Indices

Gabriel Ehrlich, John C. Haltiwanger, Ron S. Jarmin, David Johnson, Ed Olivares, Luke W. Pardue, Matthew D. Shapiro & Laura Zhao

WORKING PAPER 31309 DOI 10.3386/w31309 ISSUE DATE June 2023

This paper explores alternative methods for adjusting price indices for quality change at scale. These methods can be applied to large-scale item-level transactions data that includes information on prices, quantities, and item attributes. The hedonic methods can take into account the changing valuations of both observable and unobservable characteristics in the presence of product turnover. The paper also considers demand-based approaches that take into account changing product quality from product turnover and changing appeal of continuing products. The paper provides evidence of substantial quality-adjustment in prices for a wide range of goods, including both high-tech consumer products and food products.

RESET Project: Lessons Learned & Next Steps

- Using item-level P and Q transactions data with attributes can be used to produce
  - Internally consistent nominal sales
  - Price deflators that adjust for quality
  - Quality adjustment at scale using machine learning
- Next Steps
  - Expand methods to create new indicators at
  - scale on timely basis.
    - Demonstrate to statistical agencies this is feasible and yields improvements.
    - Deliver RESET estimates for entire retail goods sector
  - Robustly and efficiently scale to new partners
    - Information aggregators such as Nielsen and NPD/Circana + Private Firms (aiming for largest firms + sample of smaller)
- https://ebp-projects.isr.umich.edu/RESET/

Use ALL Data Assets

Going Beyond the Survey Data We Collect

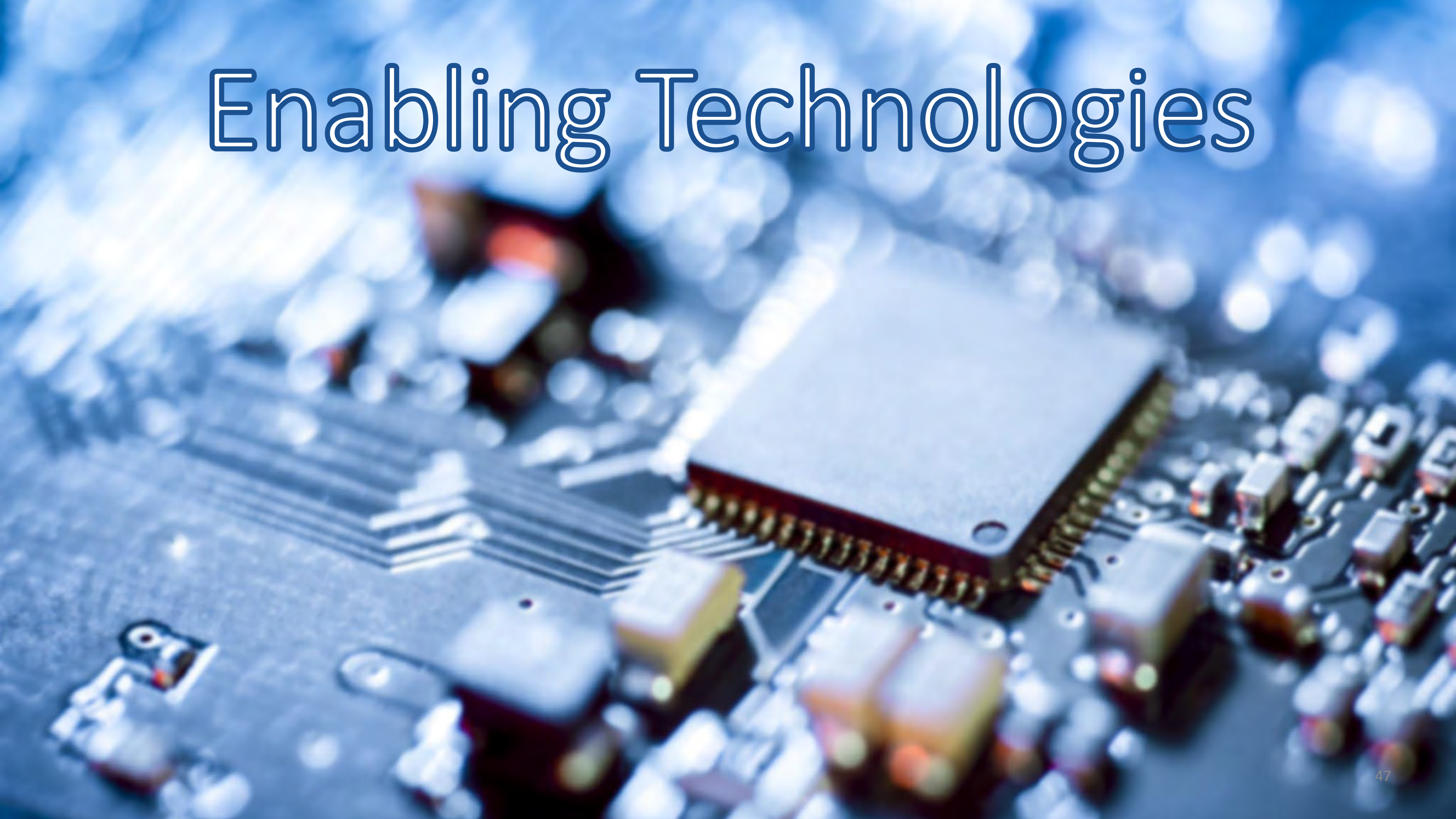
Designed Data

Administrative Data

Opportunity Data

Procedural Data

Enabling Technologies



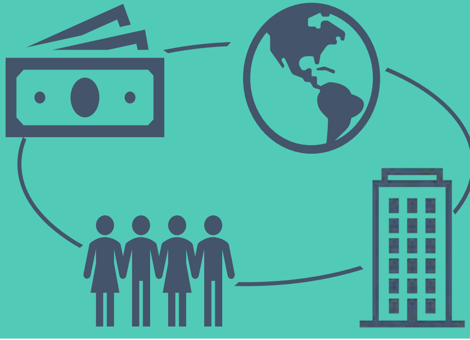
Data Ingest and Collection for the Enterprise (*all* data assets)

Enterprise Data Lake

Data processing, computing, and management

Enterprise Linked Frames

4 seamless and linkable frames



Enterprise Dissemination Services

We’re Using These Tools NOW!

ANNUAL INTEGRATED ECONOMIC SURVEY (AIES)

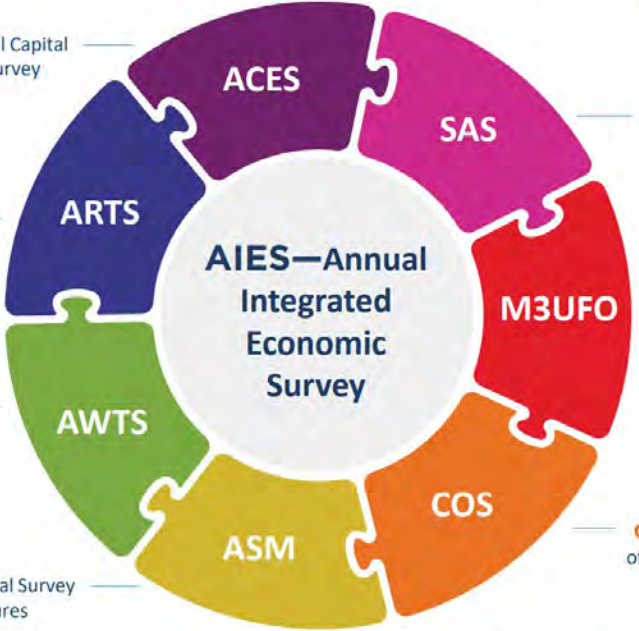
Current State

- Industry focused.
- Overlapping content.
- National estimates.
- Independent samples.
- Different reporting units by program.
- Inefficient processes and operations.
- Varying classification systems.

Future State

- Alignment to enterprise programs.
- Consolidated processing system.
- Integrated frame and sample.
- National and state geographic estimates.
- Standardized and rotating content.
- Respondent centric.
- Coordinated collection and instruments.
- Harmonized reporting units.
- Leveraging of alternative data.
- Holistic company analysis.
- Economy-wide data products.

Integration of Seven Annual Economic Surveys



ACES—Annual Capital Expenditures Survey

SAS—Service Annual Survey

ARTS—Annual Retail Trade Survey

M3UFO—Manufactures' Unfilled Orders Survey

AWTS—Annual Wholesale Trade Survey

COS—Report of Organization

ASM—Annual Survey of Manufactures

✓ 2020	✓ 2021	✓ 2022	2023	2024	2025
Test coordinated collection, evaluate existing content.	Harmonize and test content, create frame prototype.	Test sample selection, finalize content, and conduct pilot.	Develop instrument and systems, finalize frame and sample.	Launch the AIES, collect data.	Analyze and release the AIES data.

What Does This Give Us?

- The tools to manage the “estimate” (and the dataset) rather than the collection platform.
- Coverage and quality assessment for third-party data.
- Weight adjustments.
- Targeted sampling (e.g., households likely to have bar/restaurant workers in the Household Pulse Survey).
- Better resilience of survey estimates to pandemics and other shocks.
- Solving longstanding issues like undercount of groups in the census.
- New data products.
- Improved research infrastructure.

Challenges

- Access to Private Sector Data
- Resourcing innovation
- Reluctance to Change (internally and externally)
  - Introducing New Products much easier than changing existing ones.
  - Need to support other data programs
- Institutional Challenges
  - Internal organization and silos
  - Staff morale and energy
  - Executive Branch bureaucracy

United States® Census Bureau

Thank You