# Big Data for 21st Century Economic Statistics

Katharine G. Abraham, University of Maryland

Ruggles Lecture

August 23, 2021

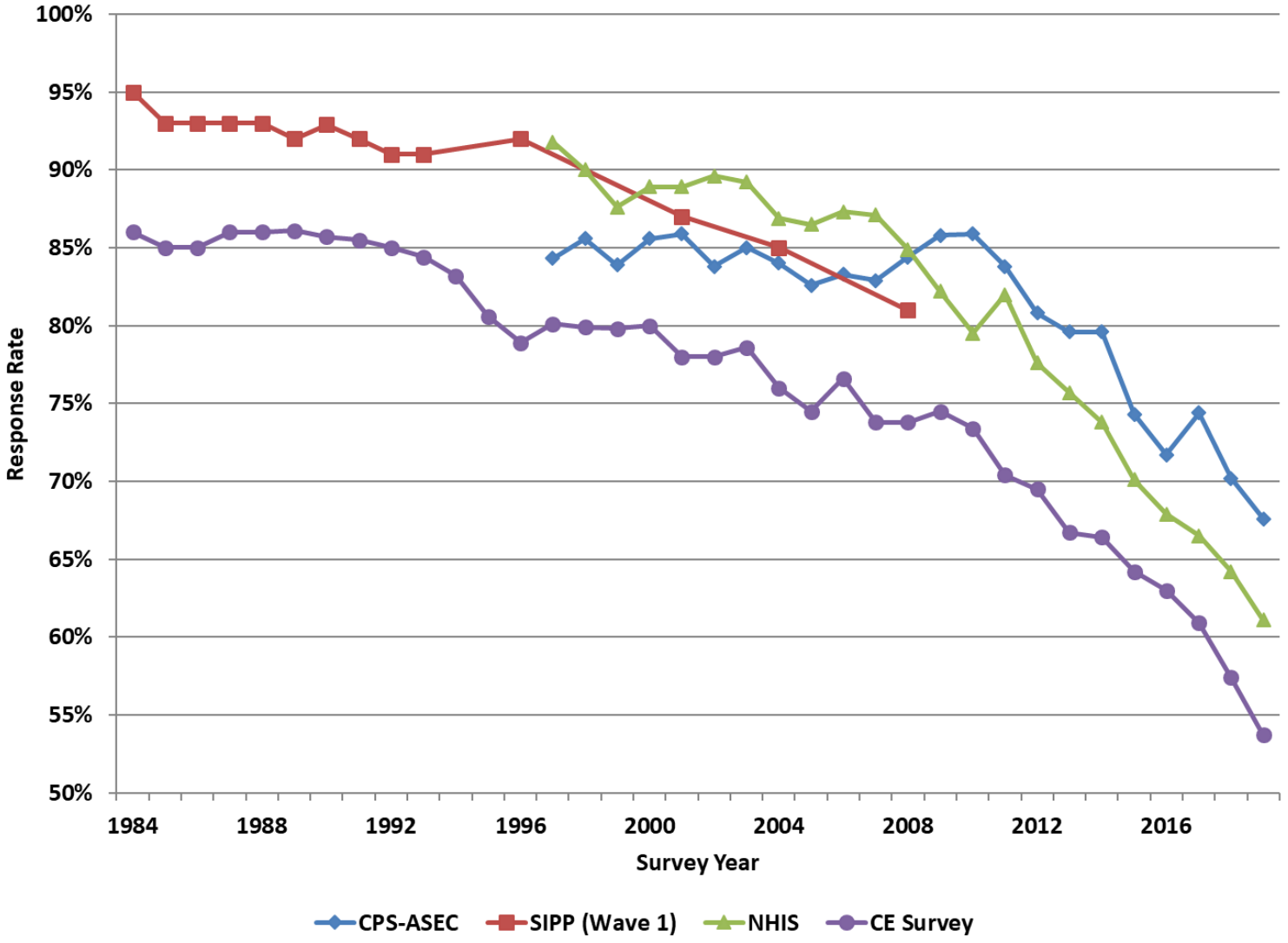
# Infrastructure and methods for U.S. economic statistics developed in years after WWII

- For decades, surveys based on probability samples have provided reliable estimates at lower cost than complete enumerations
  - Surveys underlie estimates of employment, unemployment, earnings, labor turnover, job openings, production, sales, prices, …
  - Samples designed to represent the population of interest
  - Questionnaires designed to collect desired information
- Periodic censuses and administrative data provide benchmarks
- Tasks allocated across several statistical agencies
  - Bureau of Labor Statistics (BLS), U.S. Census Bureau and Bureau of Economic Analysis (BEA) primary responsibility for economic data

# Model has served nation well, but subject to growing pressures

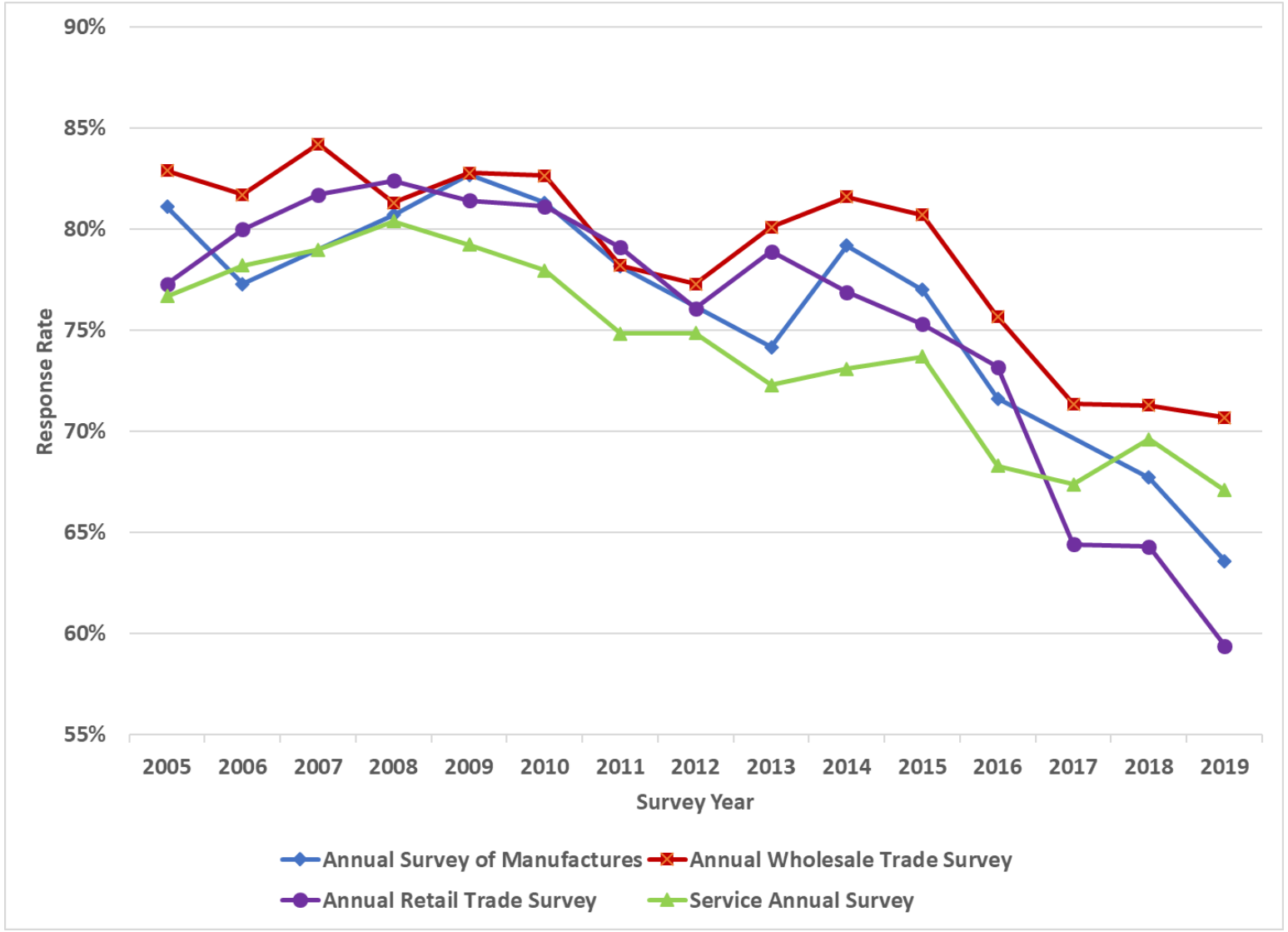
- Increasing difficulty of obtaining survey responses

# Unit response rates, selected household surveys




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

# Unit response rates, selected annual business surveys




Source: U.S. Census Bureau

# Model has served nation well, but subject to growing pressures

- Increasing difficulty of obtaining survey responses
- Increasing demand for more timely data

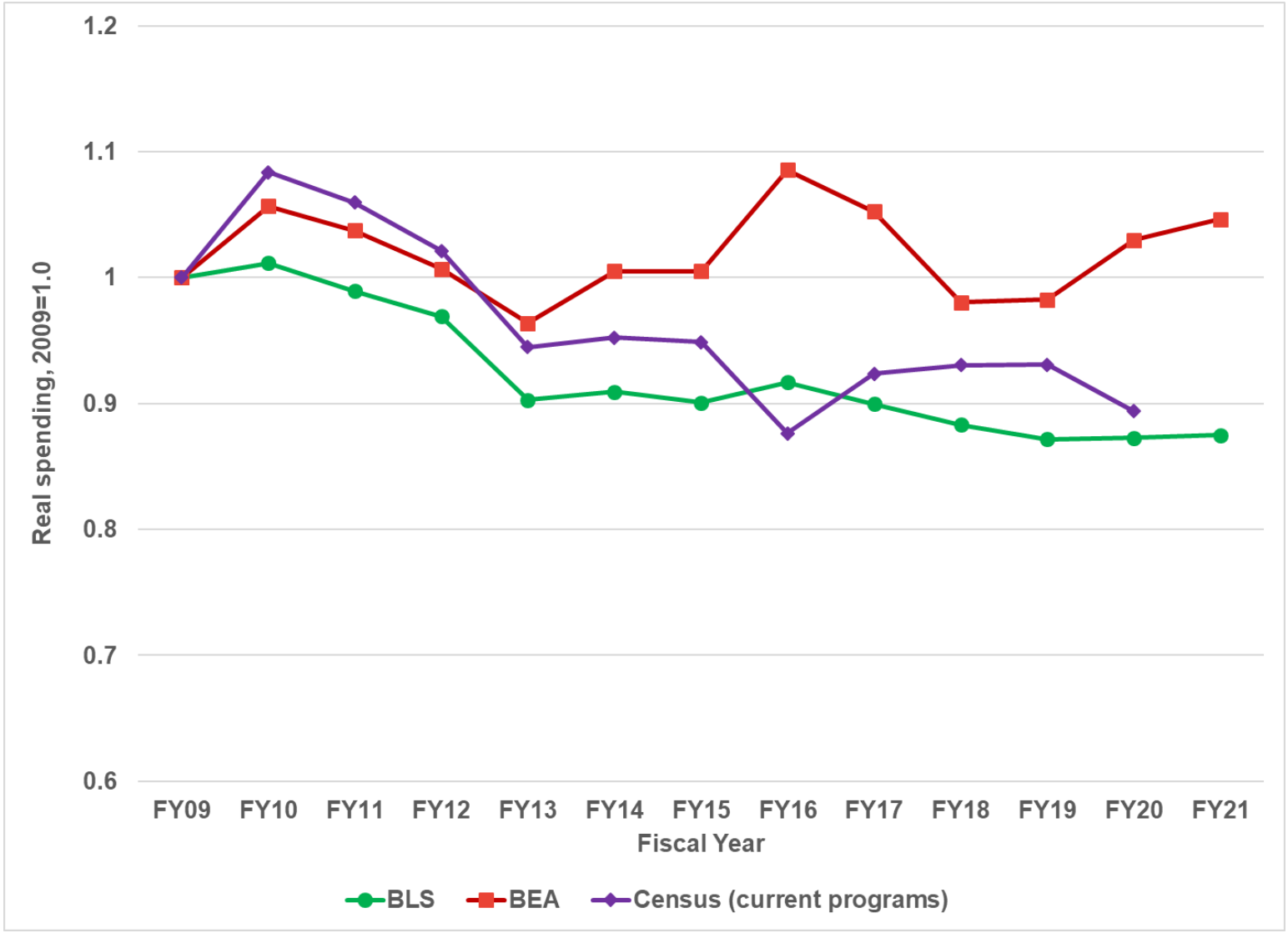
# Model has served nation well, but subject to growing pressures

- Increasing difficulty of obtaining survey responses
- Increasing demand for more timely data
- Increasing demand for more disaggregated data

# Model has served nation well, but subject to growing pressures

- Increasing difficulty of obtaining survey responses
- Increasing demand for more timely data
- Increasing demand for more disaggregated data
- Stagnant or declining agency budgets

# Economic statistics agency real funding trends (2009=1)




Source: American Statistical Association; Statistical Programs of the U.S. Government, various years

# Big data to the rescue?

- Natively digital data have proliferated in recent years
- Economic statistics agencies considerable experience with *administrative* big data
- New frontier: Use of naturally occurring data from *private sources* in the production of economic statistics to:
  - Reduce respondent burden
  - Increase timeliness and/or reduce revisions of published data
  - Increase the granularity of published data
  - Lower statistical agency costs?

# Wealth of naturally occurring private data

- Scanner data from retail outlets
- Prices, product characteristics and other information on the Web
- Credit card transactions data (e.g., JP Morgan Chase data, Spending Pulse MasterCard data)
- Payroll processing and scheduling data
- Sensor data (e.g., satellite imaging, traffic cameras)
- GPS tracking data (e.g., tractors, trucks)

# Considerations in incorporating naturally occurring data into official statistics

## Survey data

- Small but representative share of target population observed directly
- Data elements selected to meet statistical needs
- Quality control central to survey process, though errors in measurement may arise

## Naturally occurring data

- Large but not necessarily representative convenience samples
- Data elements reflect needs and constraints of business processes
- Data elements relevant to business processes most likely to be accurate

# Considerations in incorporating naturally occurring data into official statistics (continued)

## Survey data

- Comparability of data over time controlled by survey statistician
- Data records designed for statistical analysis; typically well documented
- Data "owned" by statistical agency, typically collected from respondents under a pledge of confidentiality

## Naturally occurring data

- Comparability of data over time may be disrupted by changes in business requirements
- Data records reflect business purposes; may or may not be well documented
- Data "owned" by business where it was generated; obtaining data may be expensive or raise legal, business or other concerns (including concerns about relying on a monopoly provider)

# Considerations in incorporating naturally occurring data into official statistics (continued)

## Survey data

- Agencies' physical and human infrastructure developed for collection and processing of survey data

## Naturally occurring data

- Naturally occurring data sets require enhancements to computing capacity and additional staff skills

# Potential criteria for deciding when to adopt big data for official statistics

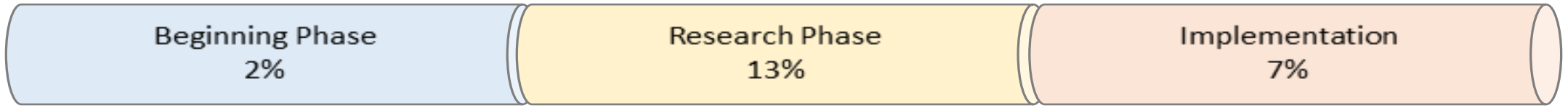
- Collecting data using current methods has become difficult or is proving inadequate to meet users' demands
- Alternative data a good fit for the intended purpose
- Quality of estimates of similar or better quality
- Costs are lower or added cost can be justified based on improvements to estimates
- Risk of relying on 3rd-party data suppliers can be mitigated

# BLS: Using big data to improve the Consumer Price Index (CPI)

- CPI price data collected by surveying businesses and rental units
  - Commodities and Services Survey: ~94,000 prices per month
  - Housing Survey: ~8,000 rental housing unit quotes per month
  - Majority of data collected by personal visit
- Program underway to substitute data from alternative sources where feasible and cost-effective
  - Similar programs underway at ONS and Statistics Canada

# CPI Alternative Data Pipeline

April 2021




**= 22%**

### 1. Identify new sources

### 2. Collecting data 2%

**Apparel**
Web-scraping, one footwear retailer

**General merchandise**
Web-scraping, many item categories

**Food away from home**
Corporate data, one fast food company

### 3. Developing methodology 13%

**Medical services**
Purchased data, insurance payments to physician's services, hospital services

**Wireless phone service**
Purchased web-scraped data, offer prices for new plans

**Residential telecommunication service**
Purchased web-scraped data, offer prices for new plans

**Airfare**
Web-scrape aggregator site, near full item coverage

**Vehicle leasing**
Purchased data (JD Power), near full item coverage

**Hotels**
Web-scrape aggregator site, near full item coverage

**Housing**
HUD administrative data, government subsidized rental properties

### 4. Seeking approval 0%

### 5. Approved for implementation 4%

**New vehicles**
Purchased data (JD Power), full item coverage, targeted deployment 2022

**Airline**
Corporate data, one airline

### 6. In development 3%

**Motor fuel**
Corporate data, full item coverage, targeted deployment June 2021

### 7. Parallel testing 0%

### 8. In production 3%

**Used cars**
Purchased data, longtime source

**Postage**
Publicly available data, longtime source

**CorpY**
Corporate data, March 2018

**CorpX**
Corporate data, March 2019

Source: Paplomatas (2021)

# BEA: Using big data to improve early GDP estimates

- GDP "advance" estimates released one month after end of quarter
  - Data from Census Quarterly Services Survey (QSS) not yet available
  - Estimated services spending for that release extrapolated from past data
- BEA researchers explored methods for "nowcasting" QSS estimates
  - Tested models based on different algorithms and different predictor variables

$$y_{it} = f_m[g_k(X_t, Y_{i,t-p})]$$

  - $y_{it}$ quarterly growth in industry *i*, $f_m$ algorithm, and $g_k$ variable selection operator
  - Candidate X's include traditional data (employment, prices) and nontraditional data (credit card transactions, Google search queries)

# BEA: Using big data to improve early GDP estimates (continued)

- Goal: Reduce revisions between advance and 3rd estimates
  - Greatest reductions: Ensemble models using employment, credit card data
  - Significant improvements in predictions for a number of sectors
- Nowcasting has been incorporated into GDP production process
  - Most often used for health care services and software investment

# Census Bureau: Using big data to produce state-level retail sales estimates

- Monthly retail sales data collected from a survey sample of ~13,000 retail and food services businesses
  - Data collected at company level; no geographic component to design
- Census has explored use of point-of-sale data from 3rd party vendor NPD to reduce respondent burden and improve national estimates
- NPD data key input to new experimental monthly state-level estimates
  - Estimates for total retail excluding non-store retailers, 11 specific sectors
  - Top-down estimates: National sales allocated based on share of industry's annual payroll in each state
  - Bottom-up estimates: Sum of sales for pre-selected multi-unit businesses from NPD, survey reporters operating in a single state, and imputed values for other retailers
  - Composite estimates: Weighted sum of two estimates; weights based on relative variances

# Census Bureau: Using big data to produce data on residential construction

- Building Permits Survey (BPS) tracks residential building permits
  - Monthly data for states and metropolitan areas
- Will be replaced with 3rd party data (~70% of single family units) and online permit information
  - Beginning January 2022, monthly census of all jurisdictions
  - Studying possibility of weekly estimates
  - Studying possibility of estimates for smaller geographic areas (e.g., zip codes)

# Census Bureau: Using big data to produce data on residential construction (continued)

- Survey of Construction tracks residential building starts, completions, sales and unit characteristics
  - Monthly data on housing starts, completions and sales for Census regions
  - Annual data on housing characteristics for Census regions
- Work in progress on using satellite images to measure building starts, completions, and selected unit characteristics

# Image Categorization

**Pre-constructions**
*90 days before permit authorization date or earlier*



Ground untouched and no major delimitations or excavations.

**Construction Starts**
*Between 30 days after permit authorization date and 125 days after it.*



Visible excavation or foundation

**Construction Completions**
*270 days after permit authorization date.*



Completed roof covering the area where there was a foundation or excavation previously.

Shape your future

United States Census

Source: Studds and Abriatis (2021)

# Census Bureau: Using big data to produce data on residential construction (continued)

- Survey of Construction tracks residential building starts, completions, sales and unit characteristics
  - Monthly data on starts, completions and sales, annual data on unit characteristics
  - Published for Census regions
- Work in progress on using satellite images to identify building starts completions, and selected unit characteristics
  - Field collection of supplemental information not obtainable via satellite will continue
  - Reduced collection costs will allow for larger sample, additional geographic and type of construction detail

# Examples cover a spectrum of uses for big data in production of official statistics

- Substituting for selected survey observations
- Improving early estimates
- Producing more disaggregated estimates
- Replacing survey data entirely!

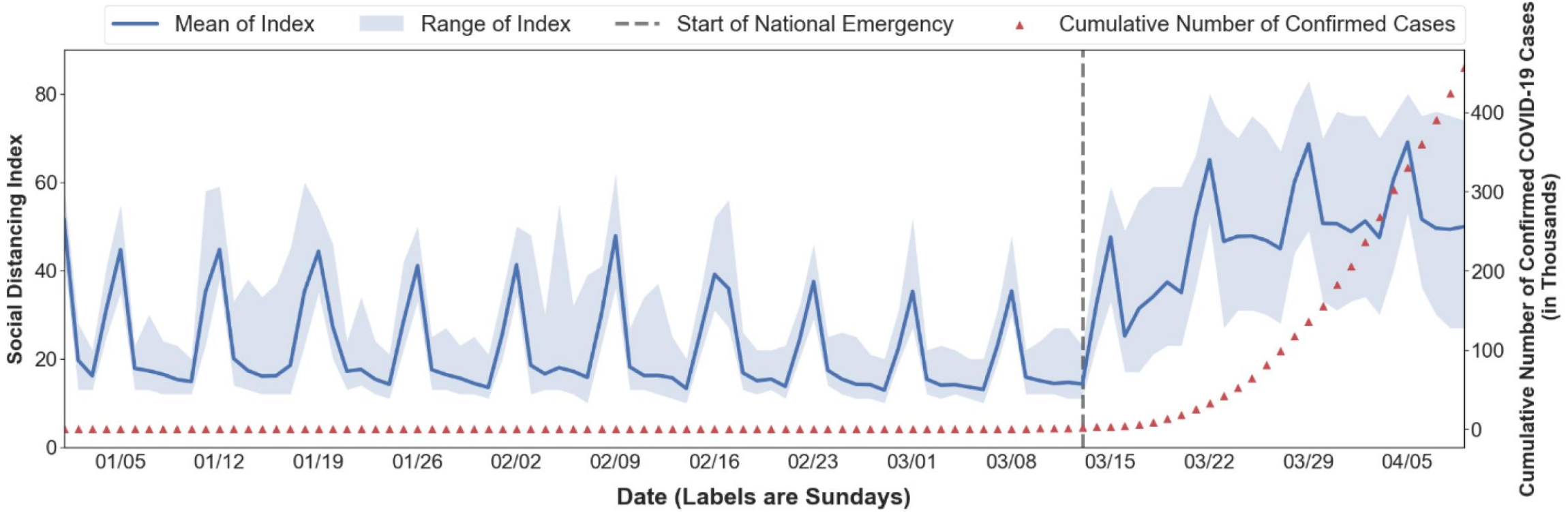
# Pandemic accelerated growth in interest in nontraditional data

- Intense demand for real-time data
- Intense demand for local area data

# Tsunami of U.S. academic research on the impact of the crisis!

- To consider a few examples:
  - Mobile phone data showed early increase in social distancing, but it was far from uniform (Pan et al. 2020)

# Evolution of social distancing index over time



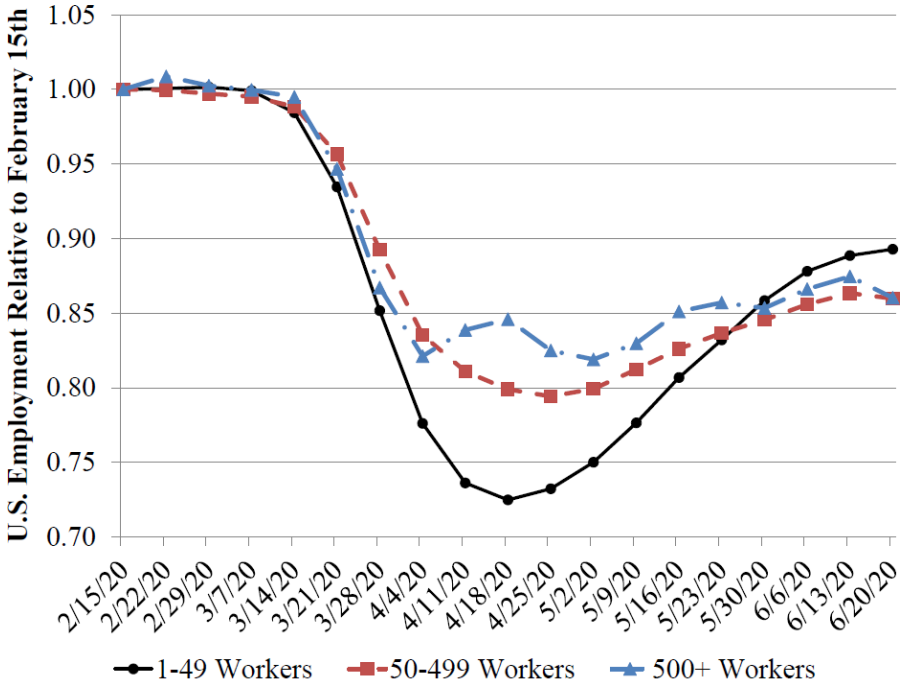

Source: Pan et al. (2020)

# Tsunami of U.S. academic research on the impact of the crisis!

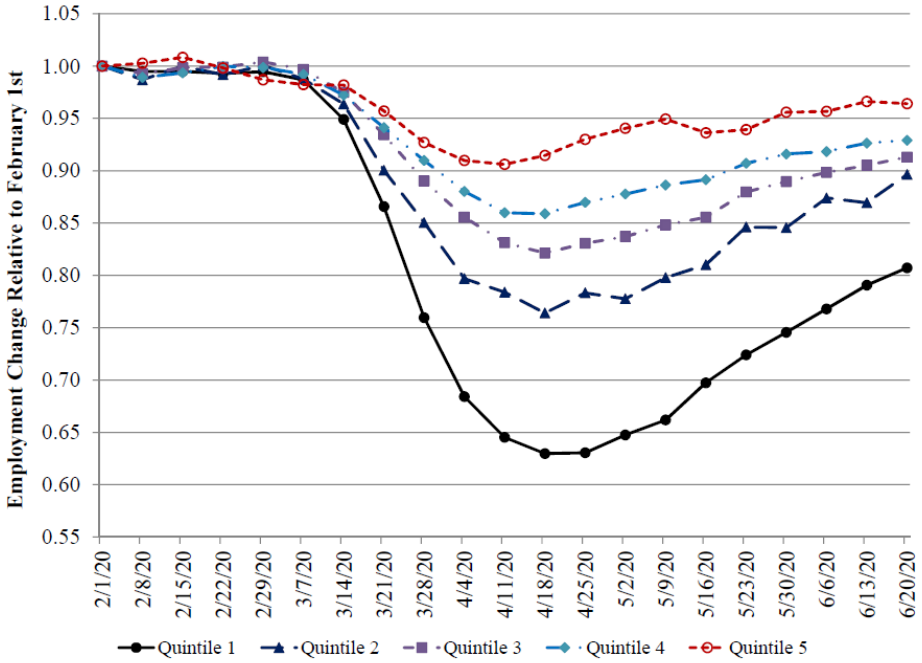
- To consider a few examples:
  - Mobile phone data showed early increase in social distancing, but it was far from uniform (Pan et al. 2020)
  - Data from ADP, a large payroll processing firm, showed larger initial impacts on employment for small firms and lower paid workers (Cajner et al. 2020)

# Pandemic impacts on employment

## By firm size




## By wage quintile

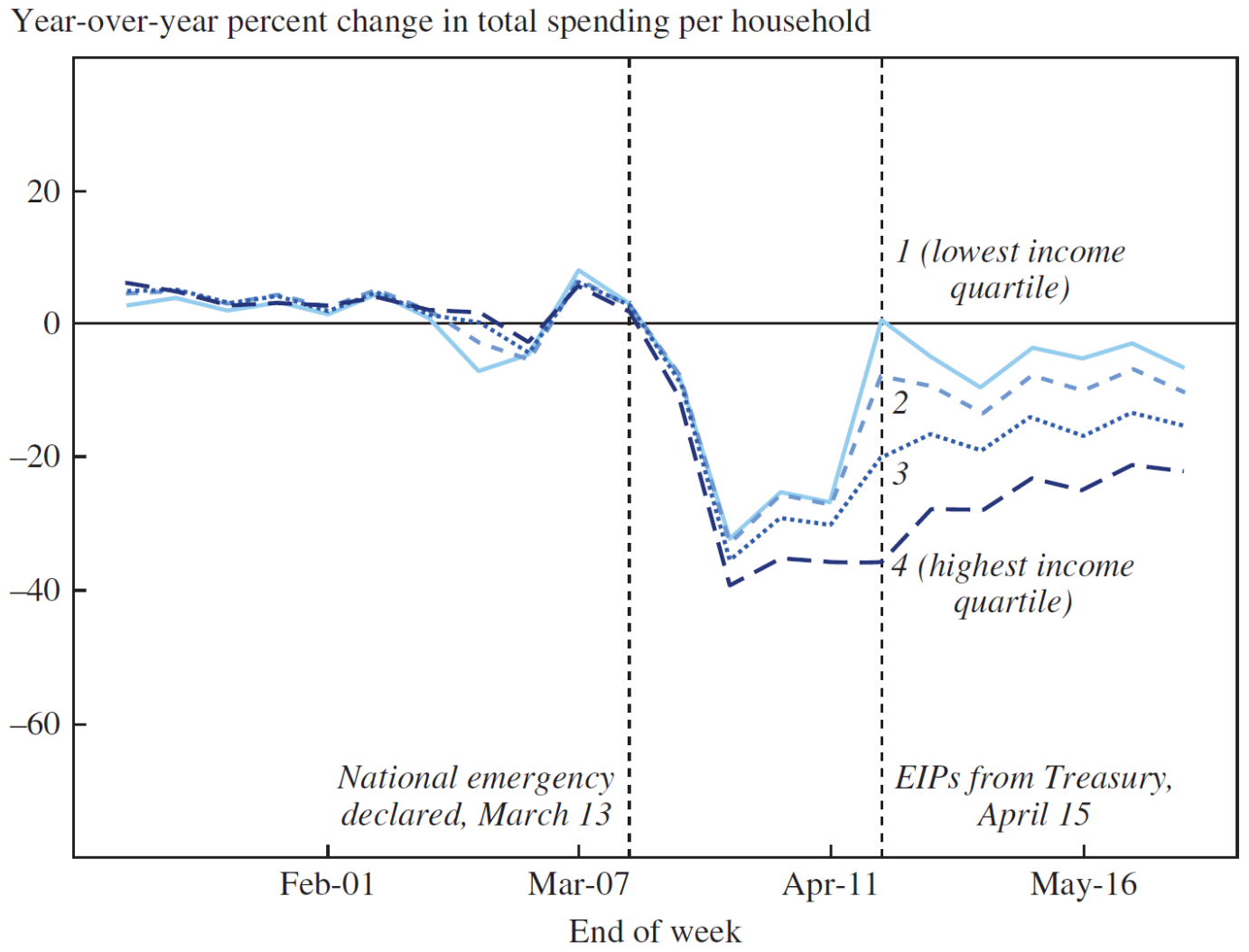



Source: Cajner et al. (2020)

# Tsunami of U.S. academic research on the impact of the crisis!

- To consider a few examples:
  - Mobile phone data showed early increase in social distancing, but it was far from uniform (Pan et al. 2020)
  - Data from ADP, a large payroll processing firm, showed larger initial impacts on employment for small firms and lower paid workers (Cajner et al. 2020)
  - Credit card, debit card and checking account data for customers of JPMorgan Chase showed stimulus payments limited financial impact of job loss on lower-income households (Cox et al. 2020)

# Pandemic impacts on household spending



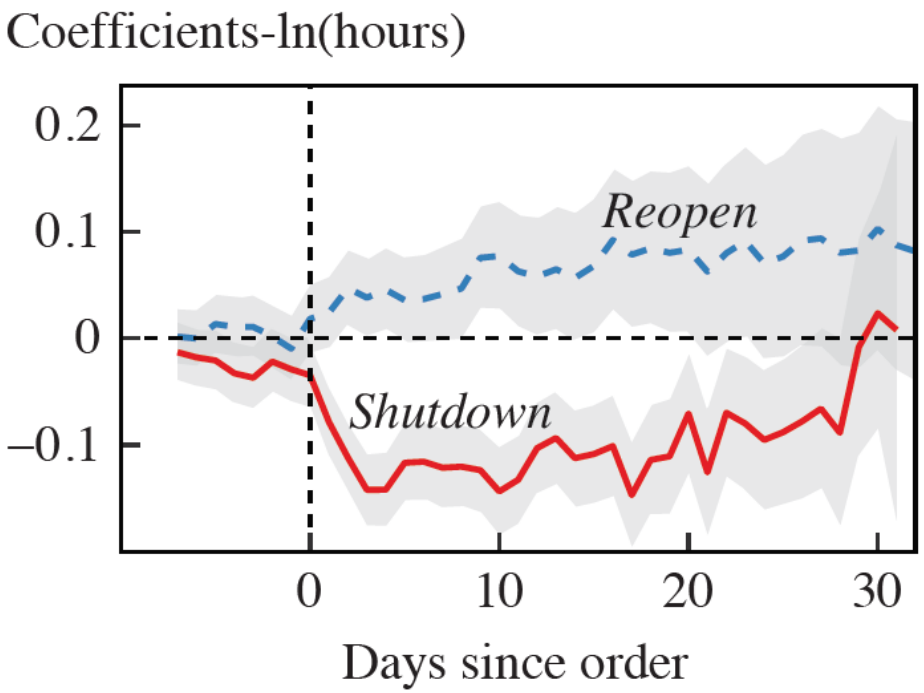

Source: Cox et al. (2020)

# Tsunami of U.S. academic research on the impact of the crisis!

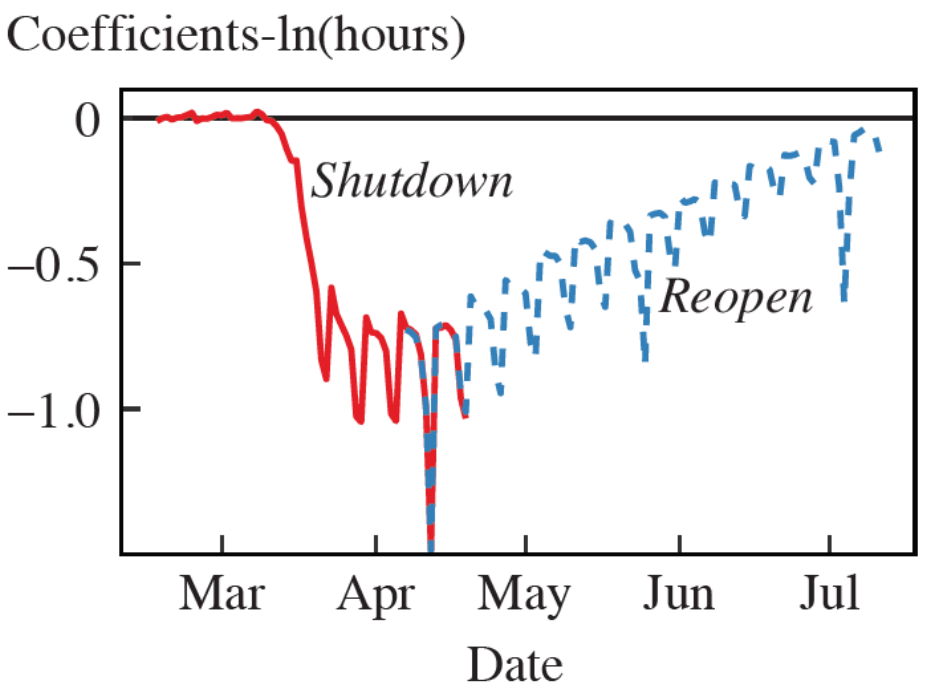
- To consider a few examples:
  - Mobile phone data showed early increase in social distancing, but it was far from uniform (Pan et al. 2020)
  - Data from ADP, a large payroll processing firm, showed larger initial impacts on employment for small firms and lower paid workers (Cajner et al. 2020)
  - Credit card, debit card and checking account data for customers of JPMorgan Chase showed stimulus payments limited financial impact of job loss on lower-income households (Cox et al. 2020)
  - Data from Homebase, a small business scheduling software company, showed modest shutdown order effects on hours (Bartik et al. 2020)

# Shutdown order and common time effects on hours early in pandemic

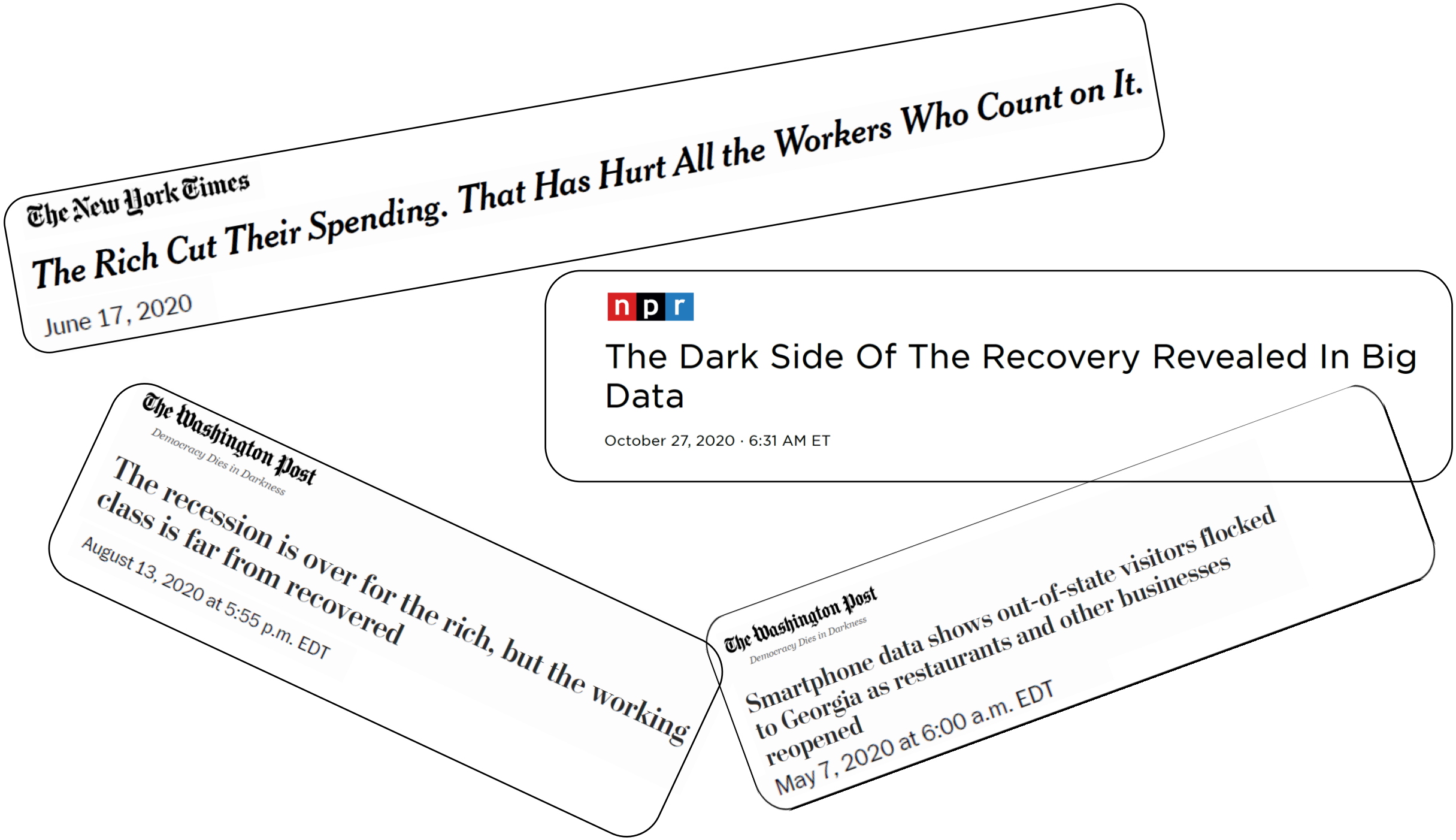
**Event study estimates of shutdown order effects on hours**




**Calendar time effects on hours**




Source: Bartik et al. (2020)


The New York Times
The Rich Cut Their Spending. That Has Hurt All the Workers Who Count on It.
June 17, 2020
npr
The Dark Side Of The Recovery Revealed In Big Data
October 27, 2020 · 6:31 AM ET
The Washington Post
Democracy Dies in Darkness
The recession is over for the rich, but the working class is far from recovered
August 13, 2020 at 5:55 p.m. EDT
The Washington Post
Democracy Dies in Darkness
Smartphone data shows out-of-state visitors flocked to Georgia as restaurants and other businesses reopened
May 7, 2020 at 6:00 a.m. EDT

# U.S. statistical agencies responded too… though in large part by expanding traditional data collection

- U.S. Census Bureau launched new Household Pulse and Small Business Pulse surveys in April 2020
- BLS added questions to Current Population Survey in May 2020 and fielded new Business Response Survey in July-September 2020
- Since June 2020, Bureau of Economic Analysis has used credit card transaction data to produce weekly estimates of retail spending (exclusive of non-store retailers)

# A number of national statistical offices are pursuing expanded big data agendas

- Statistics Netherlands created Center for big data Statistics in 2016
  - Focus on satellite data, social media data and sensor data
- ONS Data Science Campus established in 2017
  - Faster Indicators program seeks to use real-time big data to provide more timely and more granular economic insights
  - Projects undertaken during pandemic have included
    - Using Barclaycard data to produce near-real-time information on consumer spending
    - Using data from Google Community Mobility reports to produce usable information on mobility patterns (e.g., travel to work, travel to retail establishments)
    - Using text extracted from business websites to learn about how they are responding to the pandemic

# What is the role of a national statistics office?

- Traditional vision: Produce portfolio of high-quality official statistics with well-documented properties published on a regular schedule
  - Private big data incorporated into existing structure in cases where it can be shown to be preferable on grounds of respondent burden, quality or cost
- Expanded vision: Serve as a source of information that can best inform important current national, state and local policy decisions
  - Larger role for national statistics office in using private naturally occurring data to shed light on questions official statistics cannot answer

- My view: Agencies that do not adapt to demand for more timely and more granular information risk being perceived as less relevant
  - Statistical offices no longer have a monopoly on data provision

# Looking to the future

- Good reasons to rethink both the production of official statistics and the range of information national statistical offices produce
- Do not mean to understate the challenges to doing this!
  - Technical difficulties
  - Appropriate resources
  - Consistent leadership and buy-in
  - In the United States, decentralized agency structure may be a speed bump
- Viewed in a positive light, it's an exciting time to be an economic statistician!

Katharine G. Abraham
University of Maryland
kabraham@umd.edu