# Big Data in the U.S. Consumer Price Index

## Experiences & Plans

**Crystal Konny, Brendan Williams, and David Friedman**
Bureau of Labor Statistics

Big Data for 21st Century Economic Statistics
March 15-16, 2019

# Potential Benefits

- Transaction prices
- Larger sample sizes
- Reduced collection costs
- Reduced or eliminated respondent burden
- Increased detail
- Real-time expenditures and weights

# Alternative Data

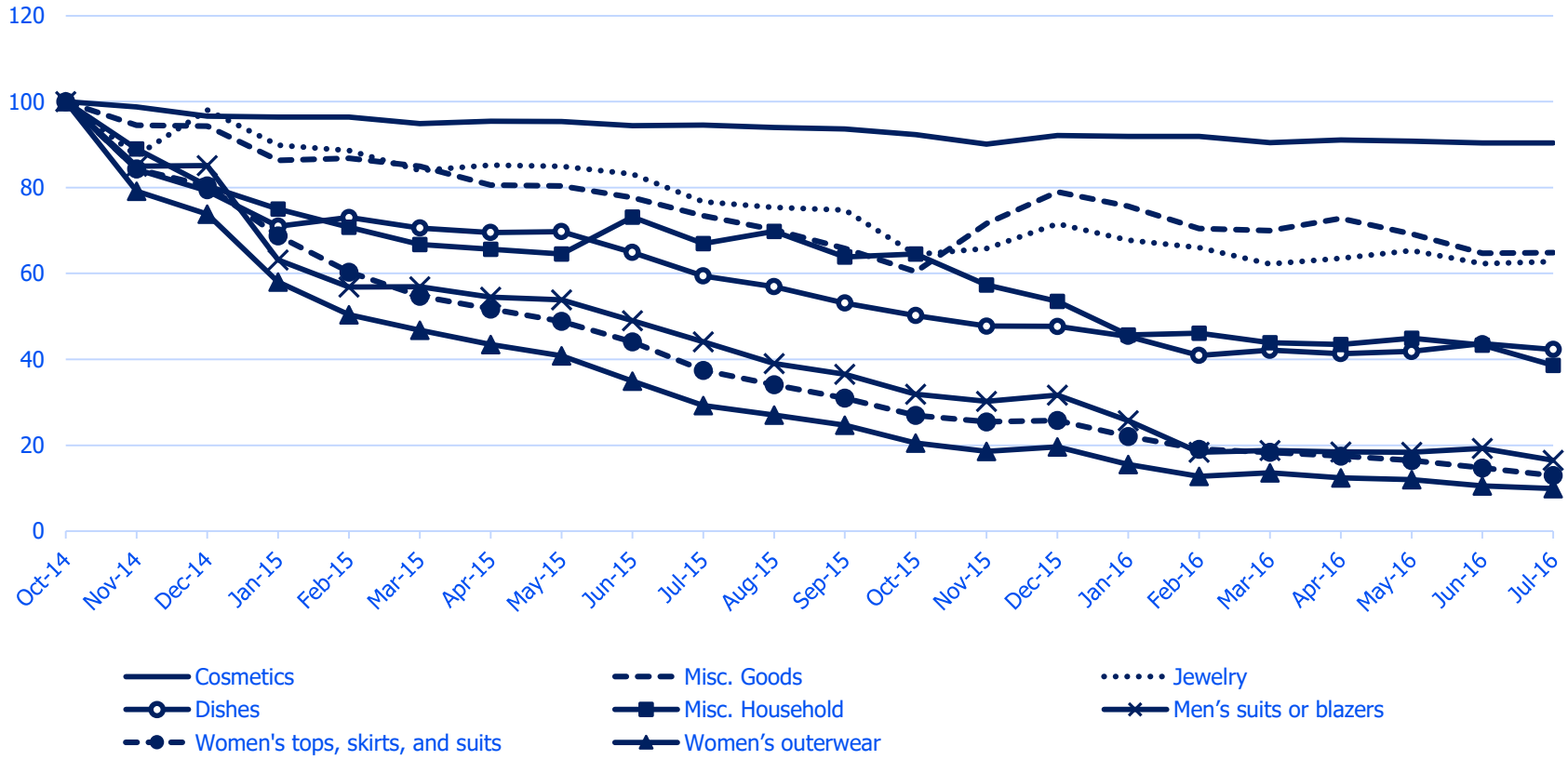
data not collected through traditional field collection procedures by BLS staff

1. Corporate supplied data
2. Secondary source data
3. Web scraped and API data

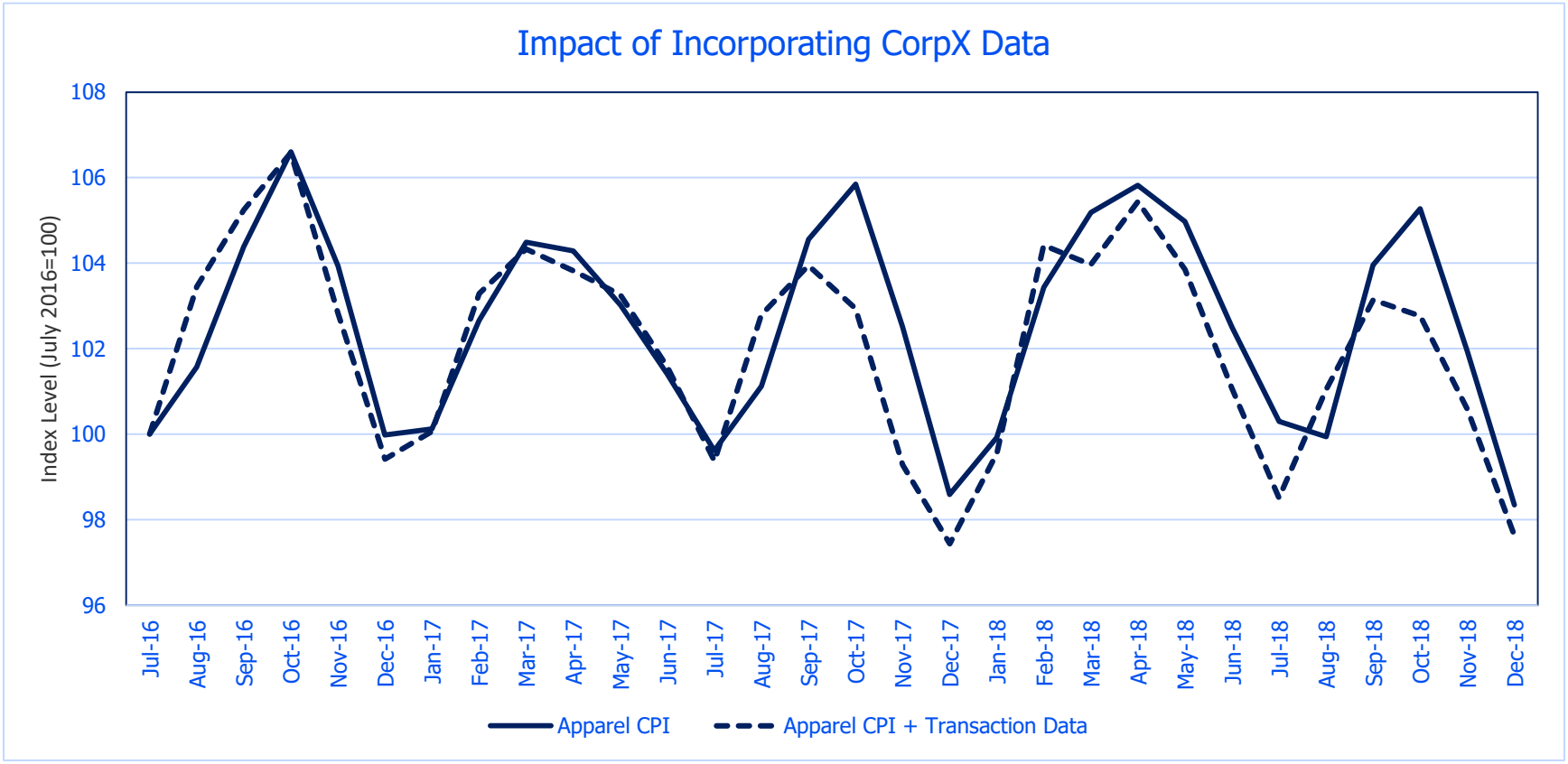
# Corporate

- CorpX
- CorpY

# CorpX


120
100
80
60
40
20
0
Oct-14
Nov-14
Dec-14
Jan-15
Feb-15
Mar-15
Apr-15
May-15
Jun-15
Jul-15
Aug-15
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Cosmetics
Misc. Goods
Jewelry
Dishes
Misc. Household
Men's suits or blazers
Women's tops, skirts, and suits
Women's outerwear

# CorpX


Impact of Incorporating CorpX Data
Index Level (July 2016=100)
108
106
104
102
100
98
96
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Jan-17
Feb-17
Mar-17
Apr-17
May-17
Jun-17
Jul-17
Aug-17
Sep-17
Oct-17
Nov-17
Dec-17
Jan-18
Feb-18
Mar-18
Apr-18
May-18
Jun-18
Jul-18
Aug-18
Sep-18
Oct-18
Nov-18
Dec-18
Apparel CPI
Apparel CPI + Transaction Data

# CorpY

- February 2012 refusal to initiate new prescription drug sample
- March 2015 agreement to supply data corporately
- May 2015 first use in index

# CorpY

| | CorpY | In-store |
|---|---|---|
| Item Selection | Probability Proportional to Size (PPS) over the past year nationally by sales excluding lowest 10% of transactions | PPS based on price of the last 20 prescriptions sold |
| Geography | National | Outlet Specific |
| Price | Average price of at least 100 transactions | Single price |
| | Insurance prices | Mostly cash prices |
| | National price | Outlet specific price |
| | Per pill price | Per prescription price |
| Patent Loss | Unit prices averaged across brand and generic | Based on analyst monitoring of patents for an NDC |
| Data Frequency | Bimonthly odd collection | Monthly and bimonthly odd/even collection |

# Secondary Source Data

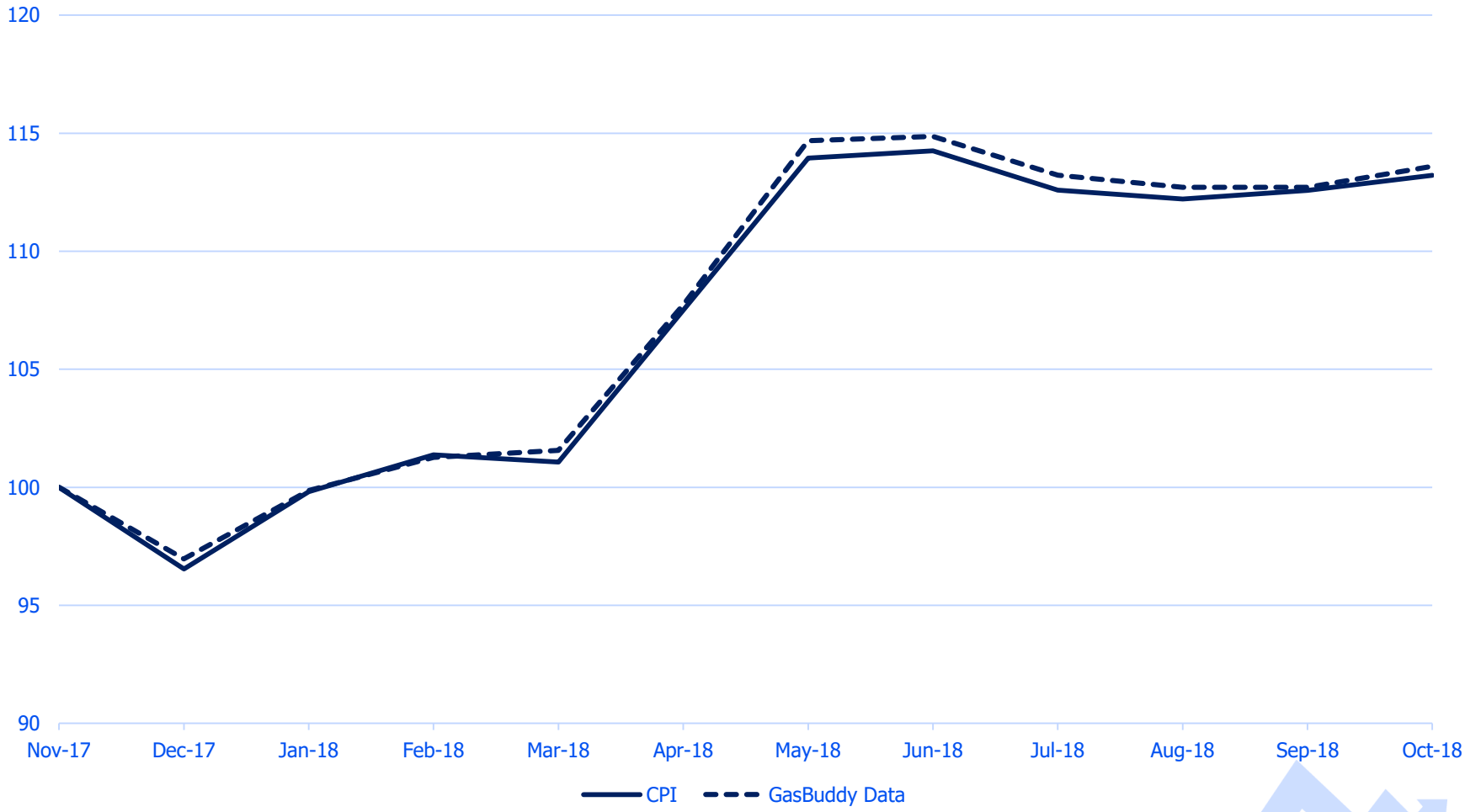
# Hospitals and Physicians' Services

- Relative Importance 4.04%
- Response Rate for Medical Care 48.1%
- 4,116 price quotes
- Cash price heavy
- High respondent burden
- High collection costs
- Difficult collection methodology

# Webscraped and API data

# Crowd Sourced Motor Fuels

## Regular Gasoline



# Establishing Priorities

- relative importance of the item
- number of quotes replaced
- cost of collection, cost of alternative data
- respondent relationship with BLS
- concentration of respondents in the sample
- ease of implementation
- Accuracy issues in the current index
  - response rates
  - collection of list prices rather than transaction prices
  - collecting prices on websites due to respondent request
  - restricted pricing at certain times of year
  - difficult collection methodology
  - degree of subjectivity in specification descriptions

# In the works

| Item | RI | # quotes | concentration | issues | priority | Source of data | % sample |
|---|---|---|---|---|---|---|---|
| Gasoline (all types) | 4.344 | 3,778 | M | L | H | scrape | 100 |
| Other motor fuels | 0.094 | 830 | M | L | H | scrape | 90 |
| New vehicles | 3.695 | 1,900 | L | H | H | sec | 100 |
| Physicians' services | 1.728 | 1,993 | L | H | H | sec | 75 |
| Hospital services | 2.312 | 2,123 | L | H | H | sec | 85 |
| Cable and satellite television service | 1.501 | 1,906 | H | H | H | sec | 95 |
| Wireless telephone services | 1.693 | 1,279 | H | H | H | sec | 98 |
| Land-line telephone services | 0.572 | 874 | H | H | H | sec | 95 |
| Internet services & electronic info providers | 0.780 | 773 | H | H | H | sec | 95 |

# In pursuit

| | RI | # quotes | concentration | issues | priority | Source of data | Experience | % sample |
|---|---|---|---|---|---|---|---|---|
| Prescription drugs | 1.316 | 4,641 | H | H | H | corp | some | |
| Limited service meals and snacks | 2.542 | 2,808 | M | L | M | corp | pursue | |
| Delivery services | 0.014 | 231 | H | L | | corp | pursue | |
| Airline fares | 0.683 | 1,745 | H | L | M | scrape, corp | research | |
| Used cars and trucks | 2.329 | 4,537 | H | H | H | sec | Prod, seek | 100 |
| Postage | 0.094 | 230 | H | L | | sec | prod | |
| Leased cars and trucks | 0.655 | 265 | L | H | M | sec | research | 100 |
| Electricity | 2.655 | 1,406 | M | M | H | | seek | |
| Utility (piped) gas service | 0.747 | 1,404 | M | M | H | | seek | |
| Rent and OER | 31.548 | | | | | | seek | |

# Conclusions

- Significant portion of the CPI based on alternative data within 5 years
- Substantial R&D on methodology needed
- Incremental improvements along with monthly publication

# Contact Information


## Crystal Konny

**Branch Chief**

**Branch of Consumer Prices**

**Konny.crystal@bls.gov**

## Brendan Williams

**Senior Economist**

**Branch of Consumer Prices**

**Williams.Brendan@bls.gov**

## David Friedman

**Associate Commissioner**

**Prices and Living Conditions**

**Friedman.david@bls.gov**