BEA
BUREAU OF ECONOMIC ANALYSIS
U.S. DEPARTMENT OF COMMERCE

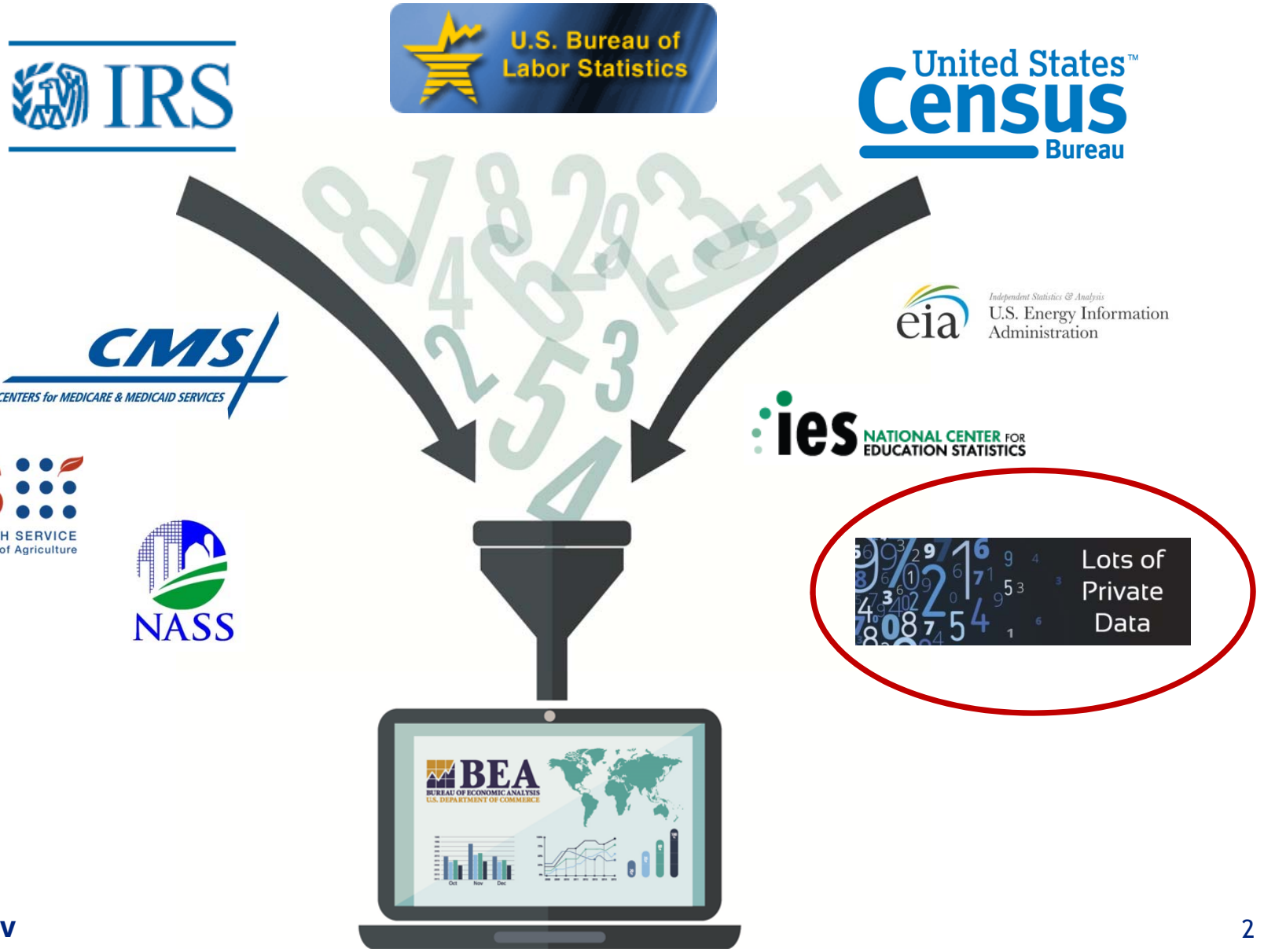
Commercial Big Data and Official Statistics

David Johnson

Federal Economic Statistics Advisory Committee

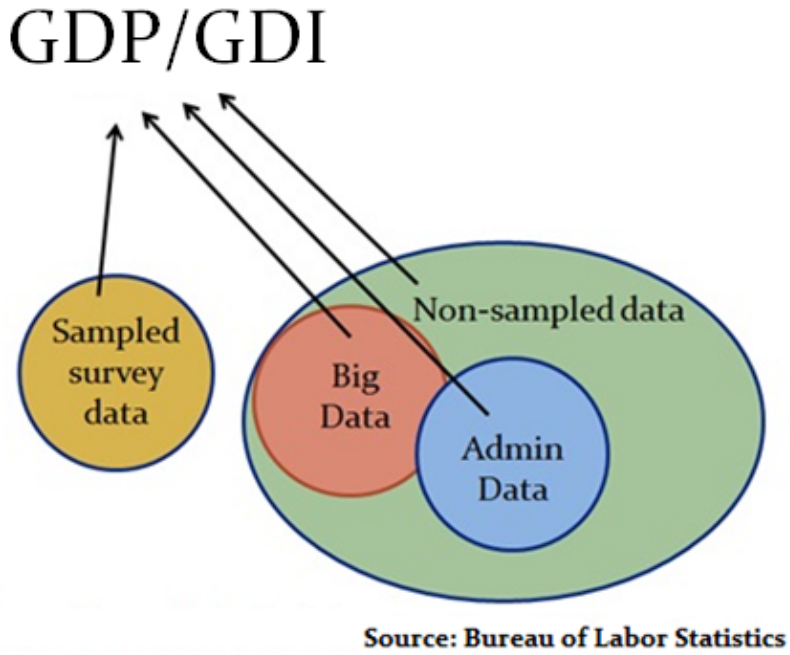
June 12, 2015

BEA Uses a Variety of Data



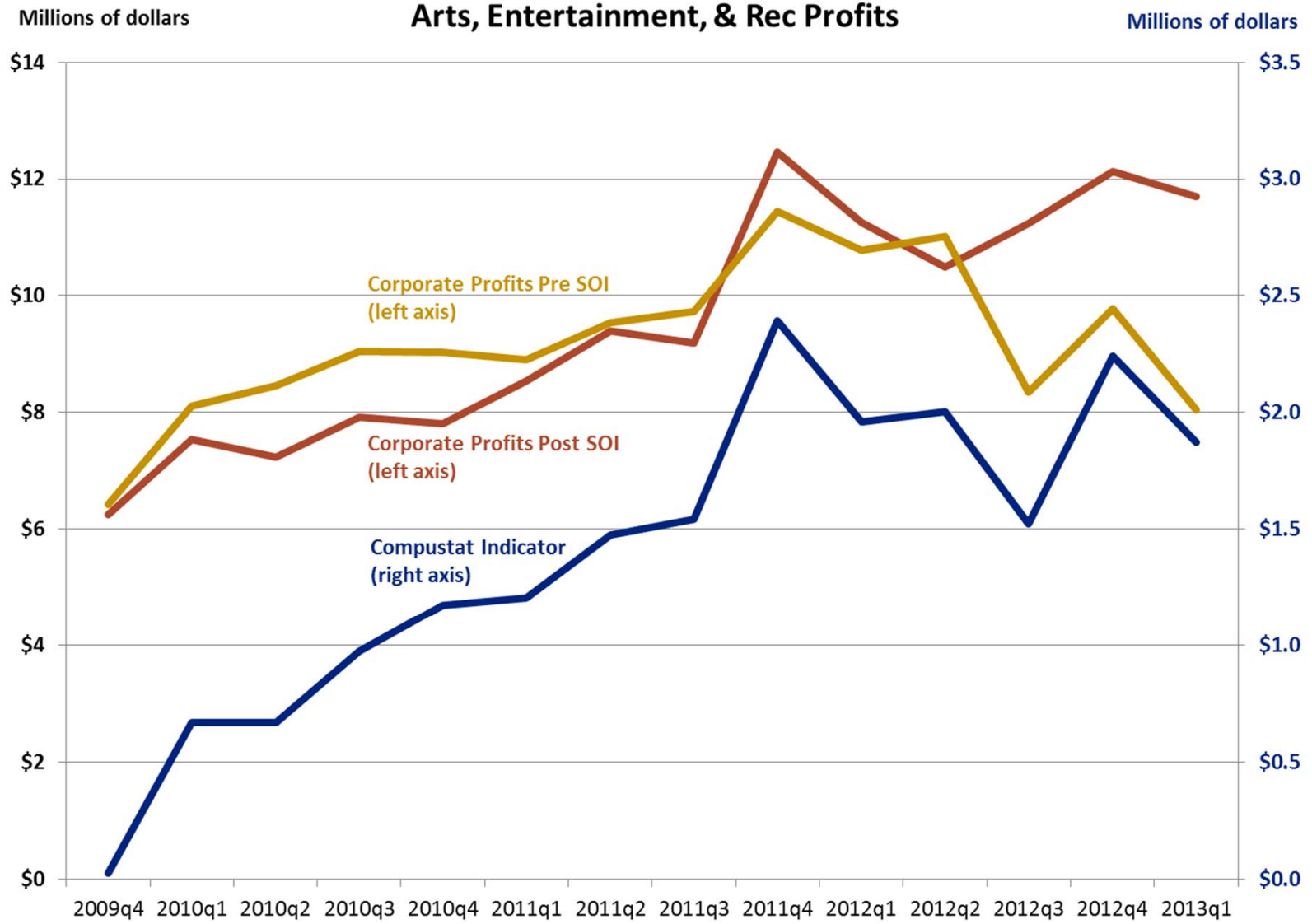
Current Uses of Commercial and Administrative Data

- BEA has a history of using nontraditional data for estimation (over 120 data sources)
- Current private source data include:
  - Ward's/JD Powers/Polk (auto sales/price/registrations)
  - Compustat (profits)
  - American Petroleum Institute (oil drilling)
- Current administrative source data include:
  - IRS, Statistics of Income
  - DOL, Unemployment Insurance data
  - FDIC, Commercial bank assets and liabilities data



Source: Bureau of Labor Statistics

Example: using Compustat for Corporate Profits



Challenges Using Commercial Data

- How representative are the data?
- Do the concepts match those needed for national accounts?
- Do the data provide consistent time series and classifications?
- Is it possible to fill gaps in coverage?
- How timely are the data?
- How cost effective?

BUT…

Opportunities for Commercial Data

- Provide indicators and extrapolators
- Fill data gaps
- Expand geographic detail
- Confirm trends
- Aid in future research efforts
  - E.g., distributional measures

Questions for Discussion

- How could the new data be used in official economic statistics?
  - Mint Bills
  - Paycycle and QuickBooks
  - CFPB Consumer Credit Panel
- Have the estimates from these new data been compared to official estimates?
- Are there suggestions for other possible data sources?
- What are the challenges in allowing agencies to access and use these data?

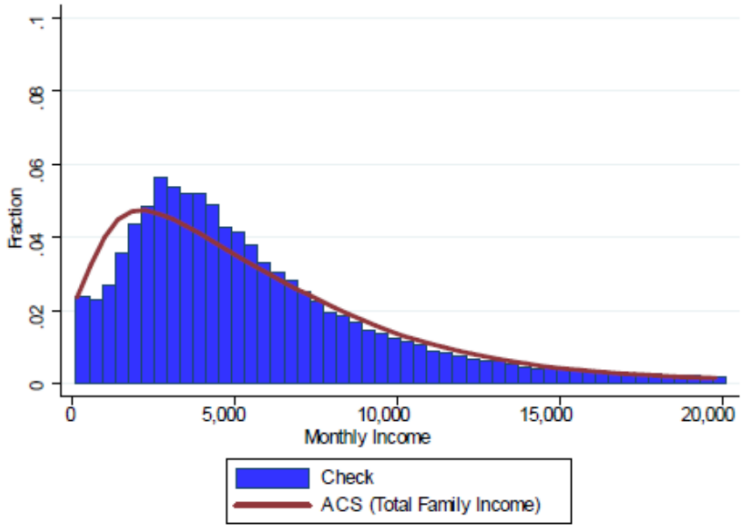
Mint Bills data are not representative of population – how does this affect measures

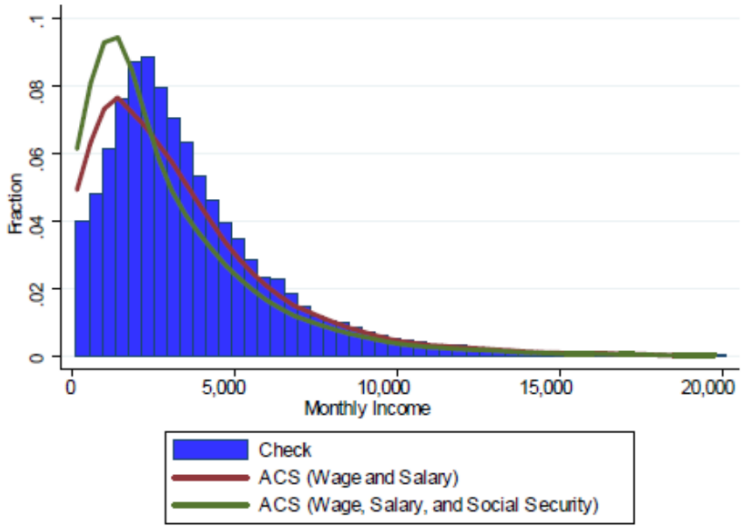
Table 1: Mint Bills vs. ACS Demographics

	Mint Bills	ACS
Female	40.07	51.41
Age		
18-20	0.59	5.72
21-24	5.26	7.36
25-34	37.85	17.48
35-44	30.06	17.03
45-54	15	18.39
55-64	7.76	16.06
65+	3.48	17.95
Highest degree		
Less than College	69.95	62.86
College	24.07	26.22
Graduate School	5.98	10.92
Census Bureau Region		
Northeast	20.61	17.77
Midwest	14.62	21.45
South	36.66	37.36
West	28.11	23.43

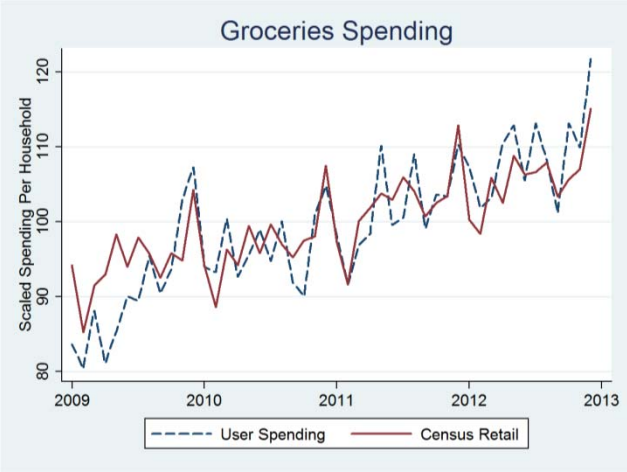
Note: The sample size for Mint Bills is 59,072, 35,417, 28,057, and 63,745 for gender, age, education and region, respectively. The sample size for ACS is 2,441,532 for gender, age, and region, and 2,158,014 for education

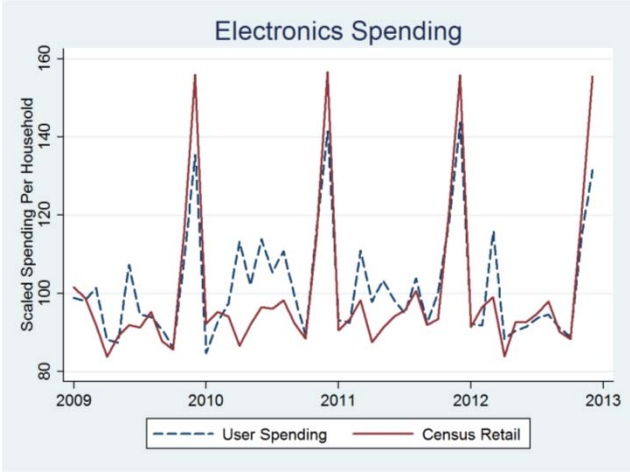
And income distribution is less skewed

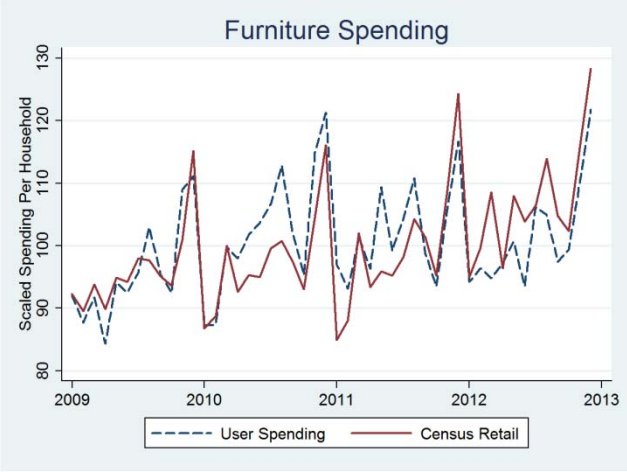




Need detailed spending to compare Mint Bills data to Census retail data, as in Baker (2014)

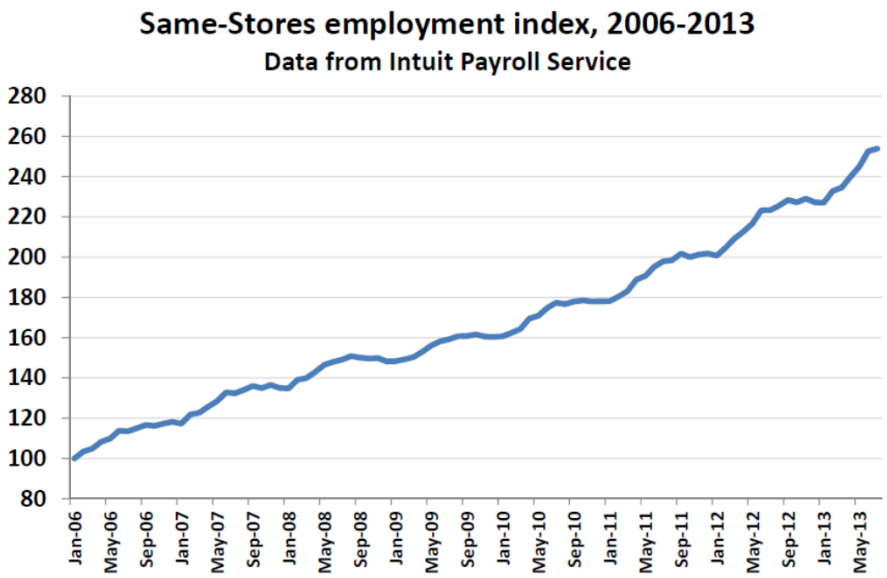


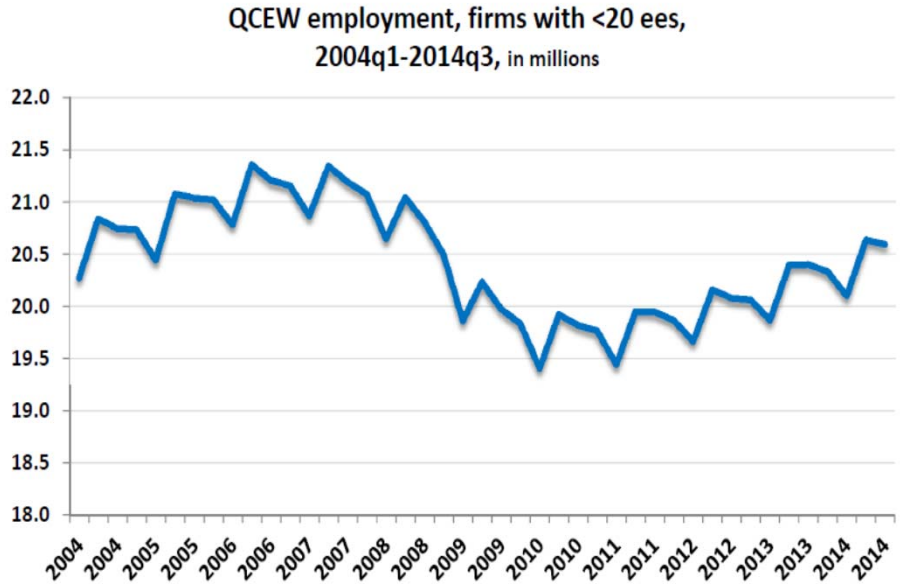




Source: “Debt and the Consumption Response to Household Income Shocks,” S. Baker, Kellogg School of Management, 2014

Why do the Paycycle small business growth rates differ from QCEW?





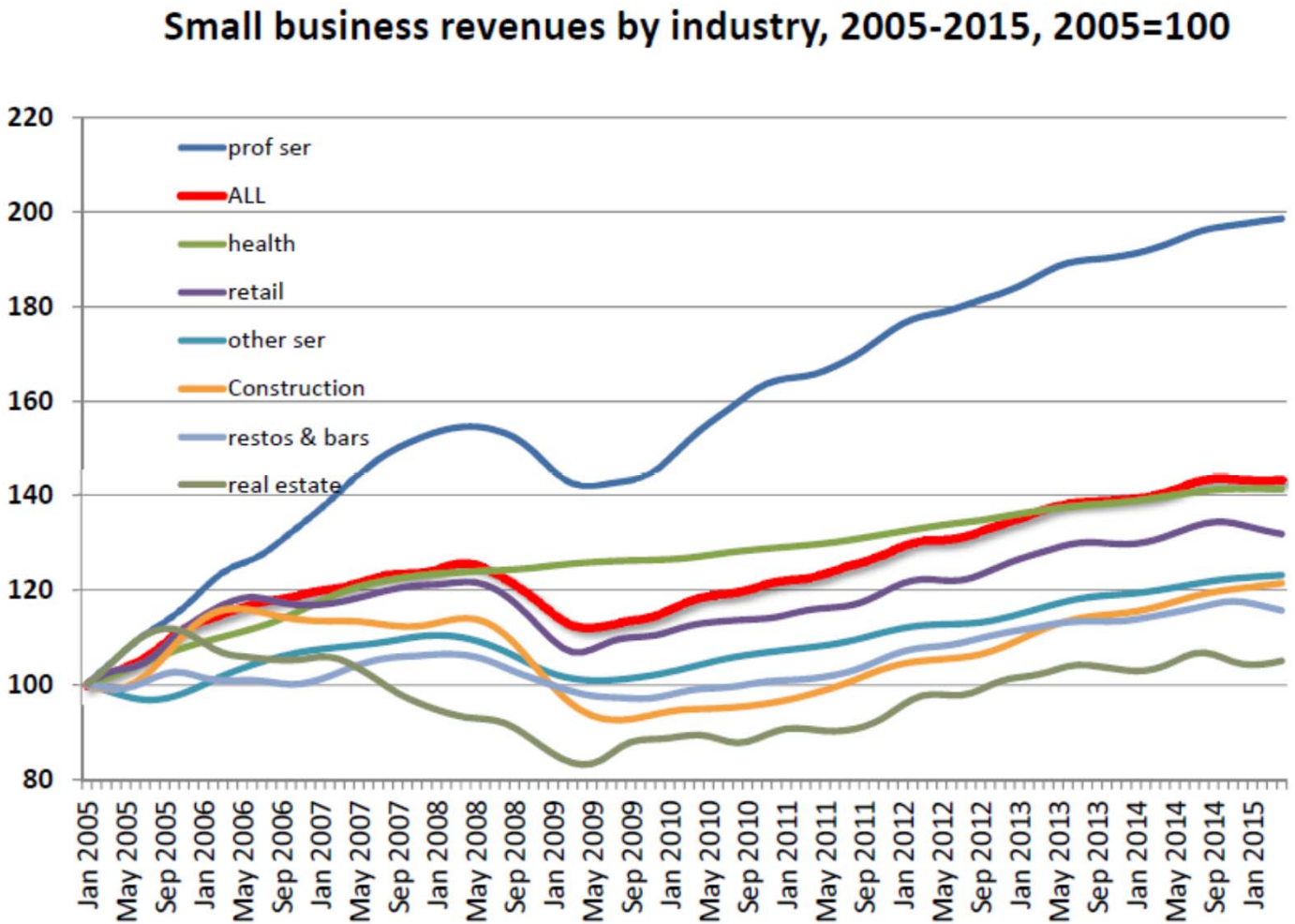
BEA initiative
Big Data for Small Business

Encourage small business growth

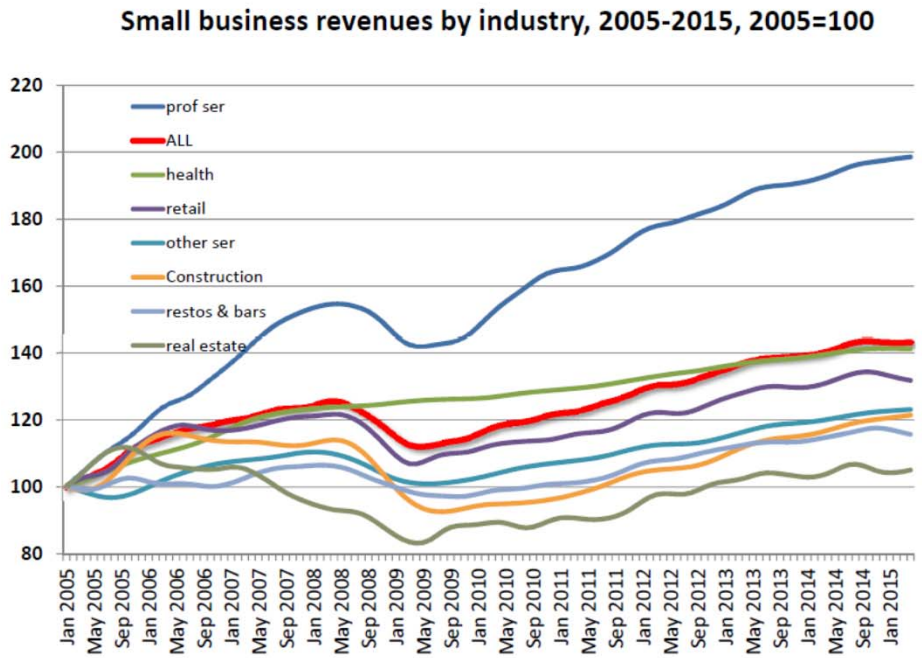
- Expanded information on small businesses would support the Department's and the Administration's goal to grow this important sector

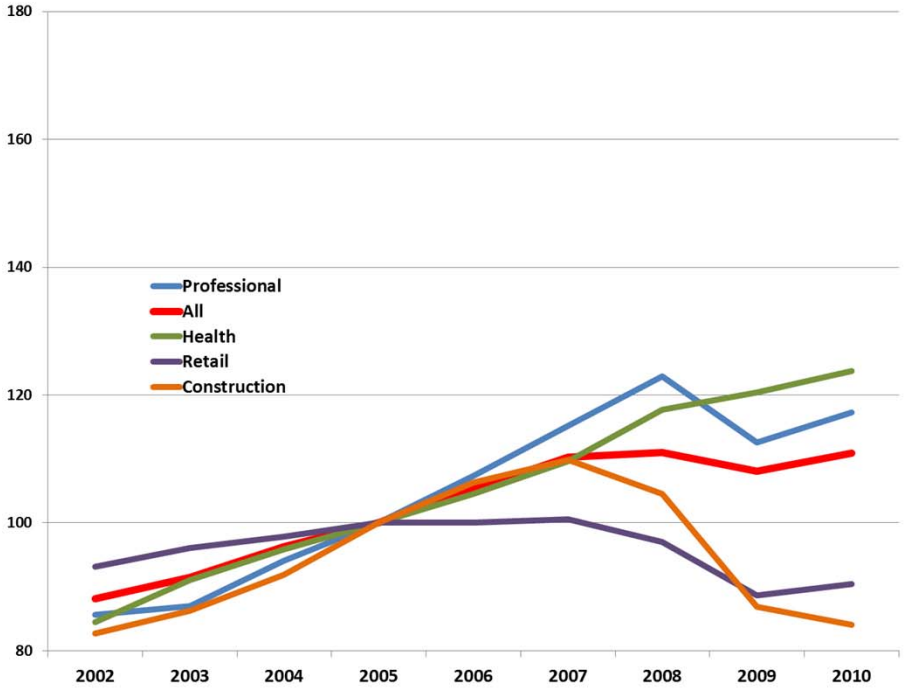
- Initiative will develop a new small business satellite account that would comprise:
  - Small business GDP
  - Small business GDP broken out by industry and by regions of the country
  - Distributional information of the employment and sales of small businesses
  - Information on the legal form, taxes, and net income of small businesses

Quickbooks data may provide insight into small business revenues



Need to Compare estimates to other sources, e.g. Small Business Administration





Source: "Small Business GDP: Update 2002-2010," Kathryn Kobe, Small Business Administration, 2012

How representative is the CFPB's Consumer Credit Panel (CCP)…

…a longitudinal sample of approximately 5 million de-identified credit records that is nationally representative of the credit records maintained by one of the nationwide credit reporting agencies (NCRA).

TABLE 1: EFFECTS OF SAMPLE EXCLUSIONS AND INCLUSIONS

	Scored Records	Credit Invisibles	Stale-Unscored	Insufficient-Unscored
Baseline Estimate	188.6	26	9.6	9.9
Exclusions:				
Missing in 2014 (Total)	-4.3	+11.7	-1.7	-5.7
Observed Merge	-3.0	+6.6	-1.2	-2.4
Disappeared	-1.2	+5.0	-0.5	-3.3
Missing Age	-6.5	+7.4	-0.4	-0.5
Inclusions:				
Debt Collection Only	+0.1	-2.1	+0.09	+1.9
Public Record Only	+0.01	-0.4	+0.01	+0.3

Source: "Data Point – Credit Invisibles," CFPB Office of Research, 2015

Need to compare the characteristics to another data source, as does the NY Fed CCP

Table 2. Comparison of 2008 Age Distributions Based on Extended Sample

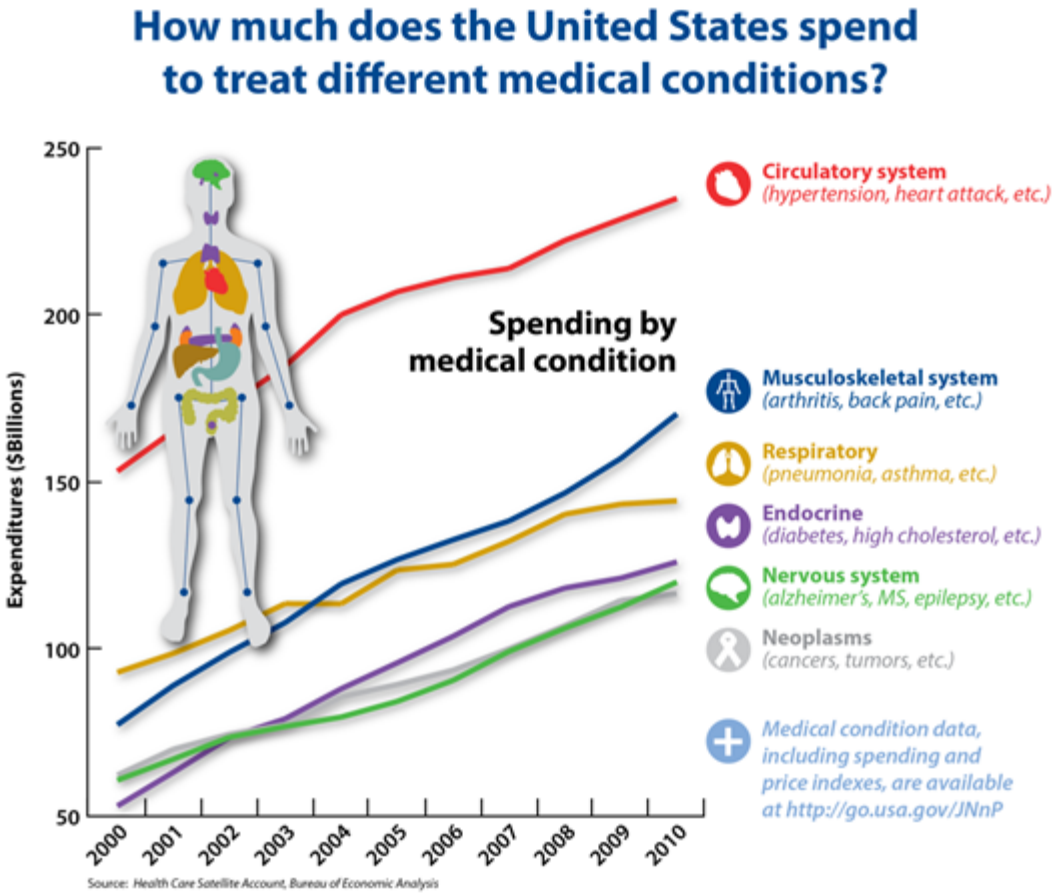
Age	US			NY State			NYC			Manhattan		
	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All	ACS Age≥18	ACS ≥20	FRBNY All
18-24	13.1	9.5	9.4	13.3	9.7	9.1	12.6	9.3	8.9	10.5	7.9	7.4
25-34	17.6	18.3	17.0	16.8	17.5	17.1	19.2	19.9	21.6	21.6	22.6	24.4
35-44	18.5	19.3	18.4	18.6	19.4	18.9	20.5	21.2	20.8	23.4	24.0	21.2
45-54	19.3	20.1	19.6	19.4	20.2	19.9	18.2	18.8	19.2	16.6	17.1	17.0
55-64	14.7	15.3	15.5	14.7	15.4	15.3	13.6	14.1	14.2	12.7	13.0	13.9
65-74	8.7	9.1	9.5	8.8	9.2	9.1	8.3	8.6	7.8	8.1	8.3	8.0
75-84	5.7	5.9	6.7	5.8	6.1	6.6	5.4	5.6	4.8	5.2	5.3	4.9
85+	2.4	2.5	4.0	2.6	2.7	4.0	2.3	2.4	2.8	2.3	2.4	3.2
Total (millions)	230.2	221.1	266.2	15.1	14.5	16.6	6.4	6.2	6.8	1.4	1.3	1.7

American Community Survey figures are 1-yr estimates for 2008 from tables B11016 Household Type by Household Size. The FRBNY figures are based on the Q4 2008 wave in the consumer credit panel. All counts are in millions.

Source: "An Introduction to the FRBNY Consumer Credit Panel," D. Lee and W. van der Klaauw, Federal Reserve Bank of NY Staff Report, 2010

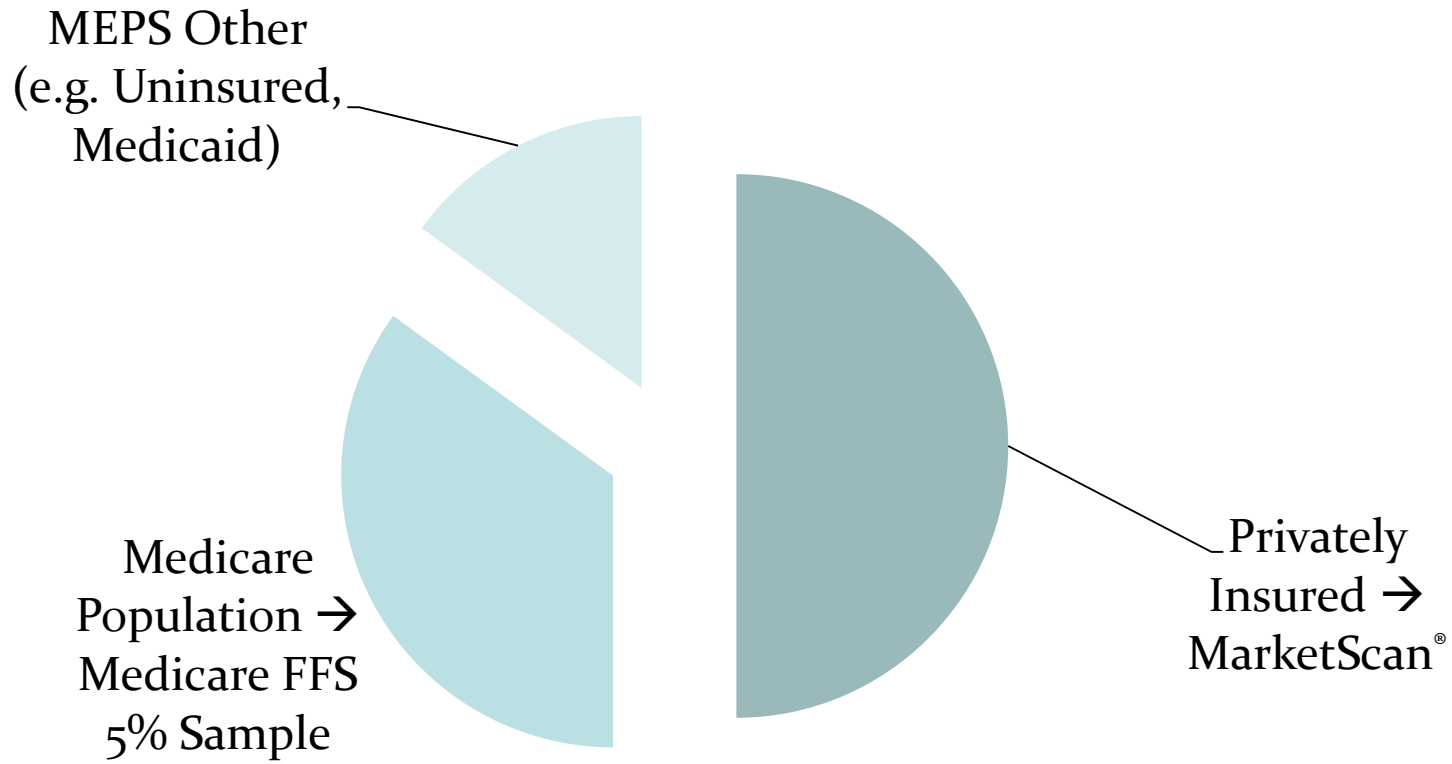
Example: Health Care Satellite Account

- Annual statistics for 2000-2010 that provide information on spending and price changes by disease category
- BEA combined billions of claims from both Medicare and private commercial insurance to determine the spending for over 250 diseases

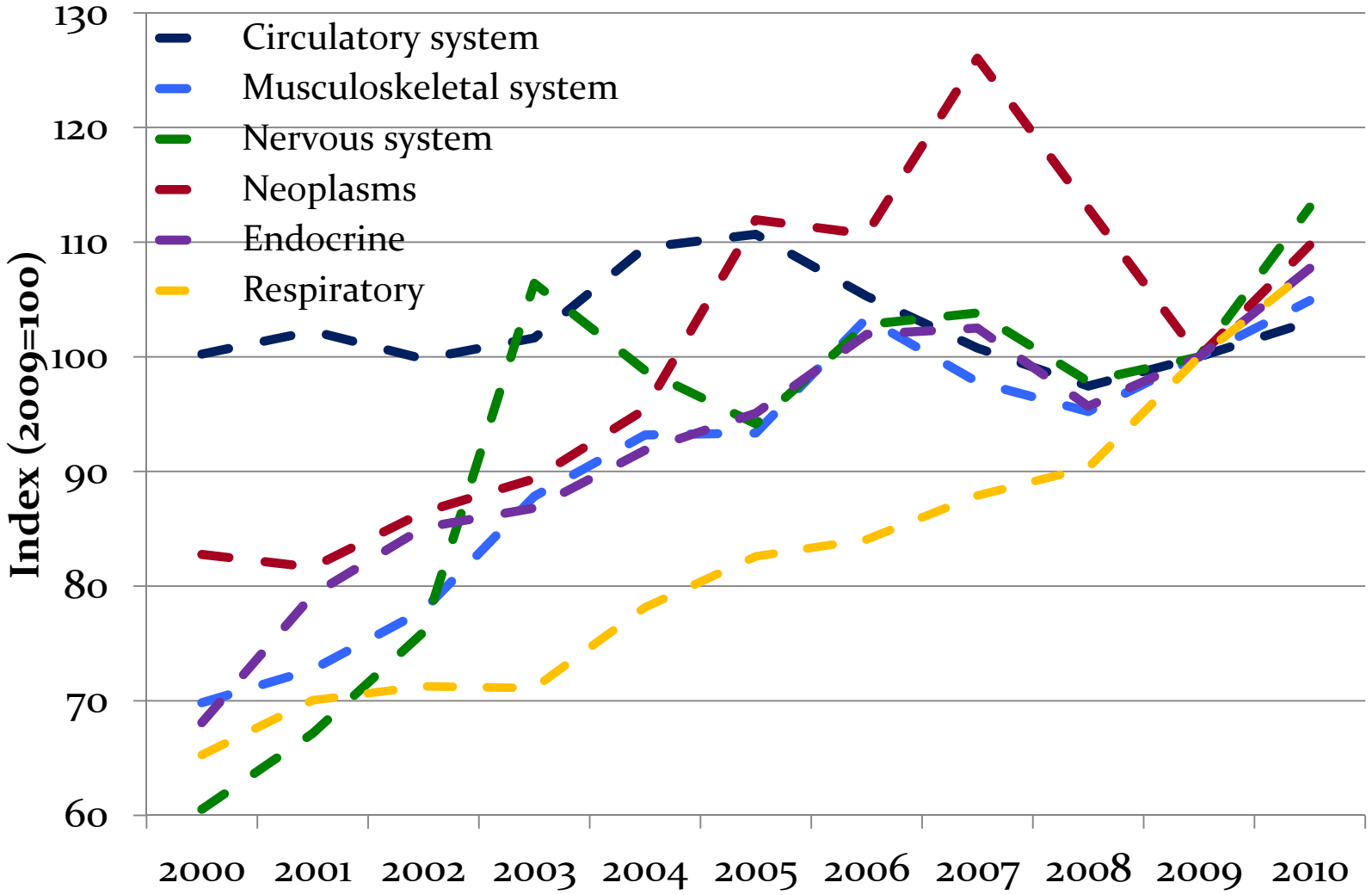


Construction of Blended Account

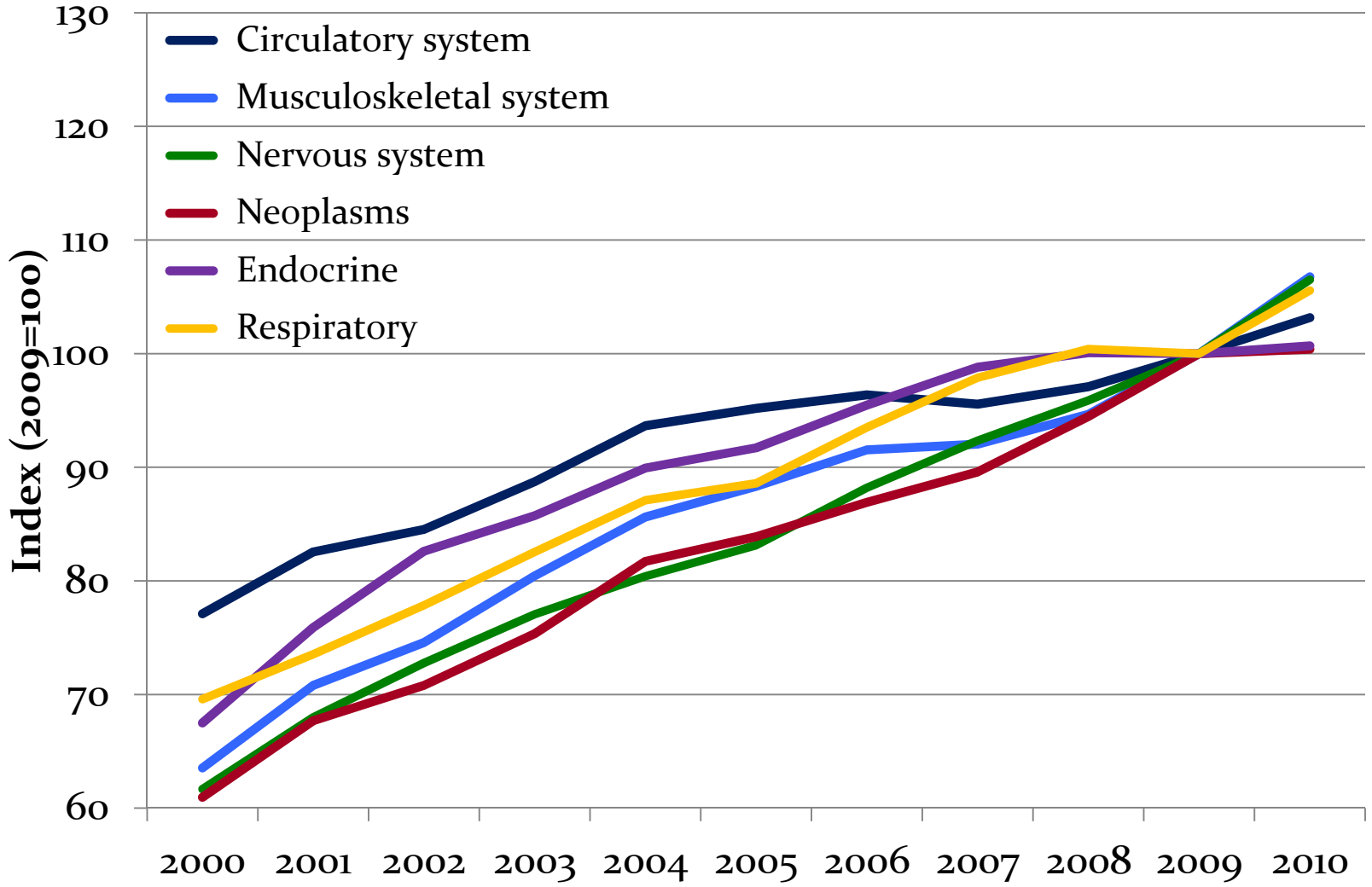
- Use survey population weights to fold in data from different sources



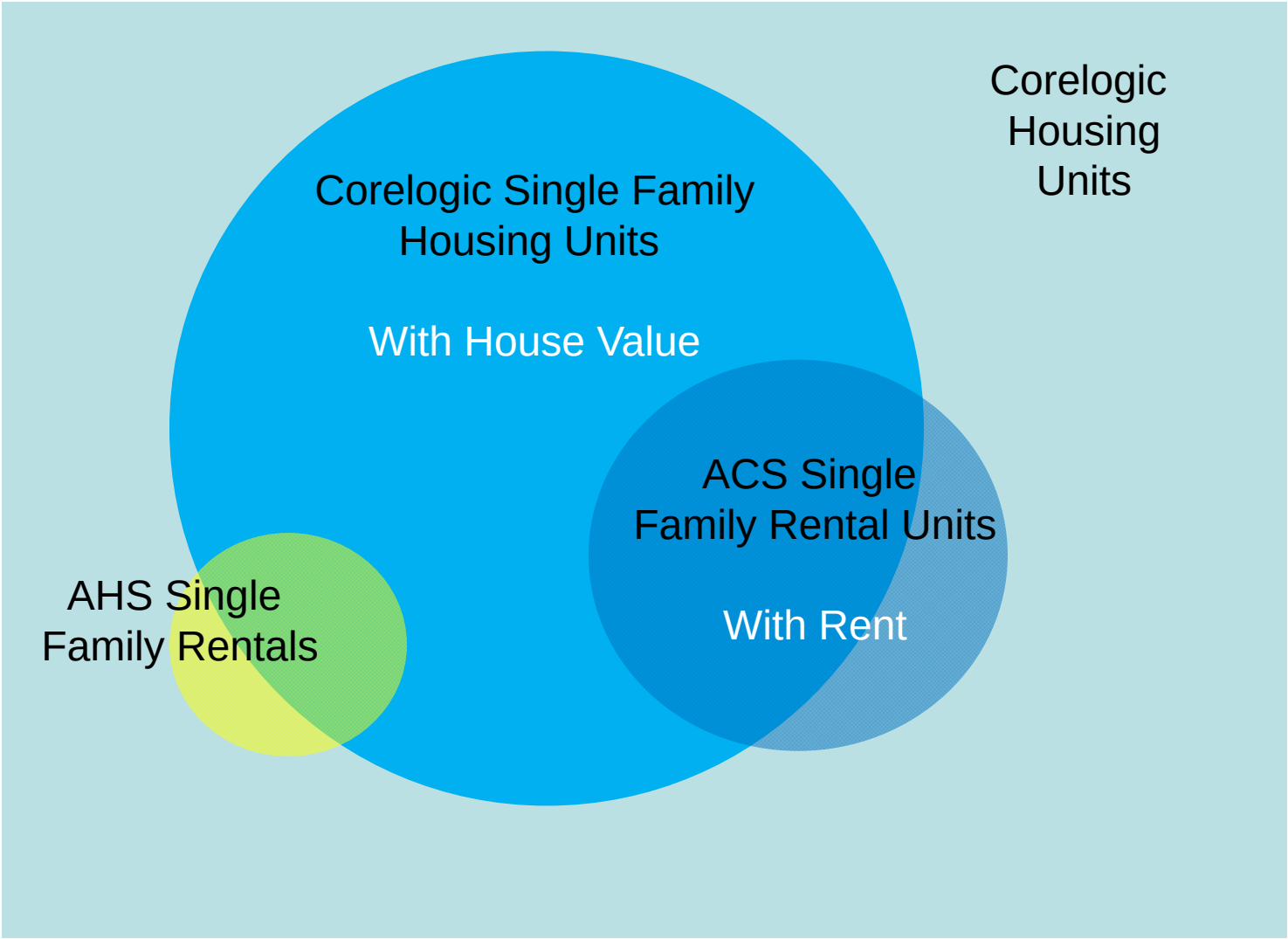
Health Care Satellite Account: Survey Data Only



Health Care Satellite Account: Survey + Big Data



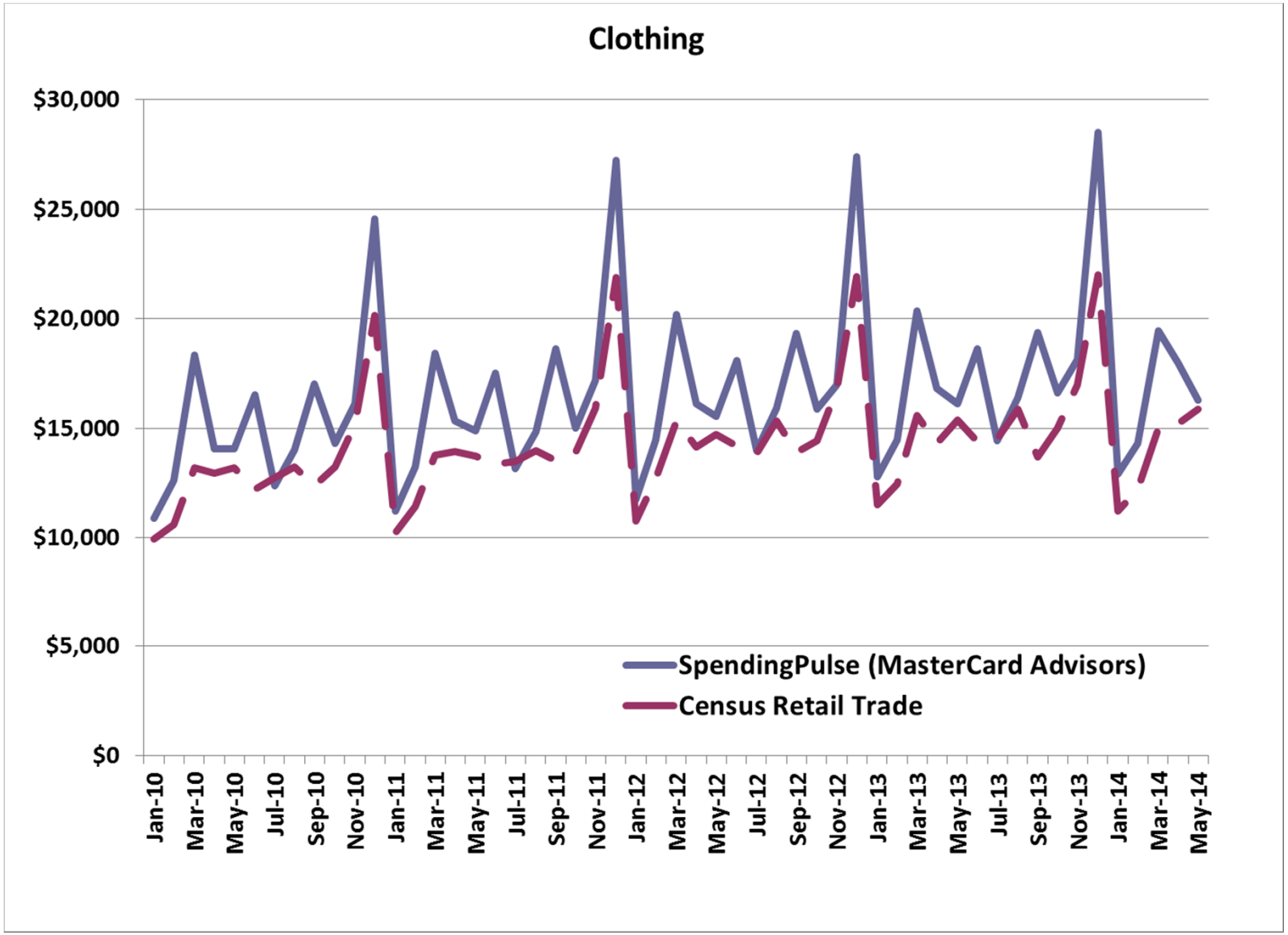
Example: Match Corelogic house prices to ACS rents to obtain rent to value ratio



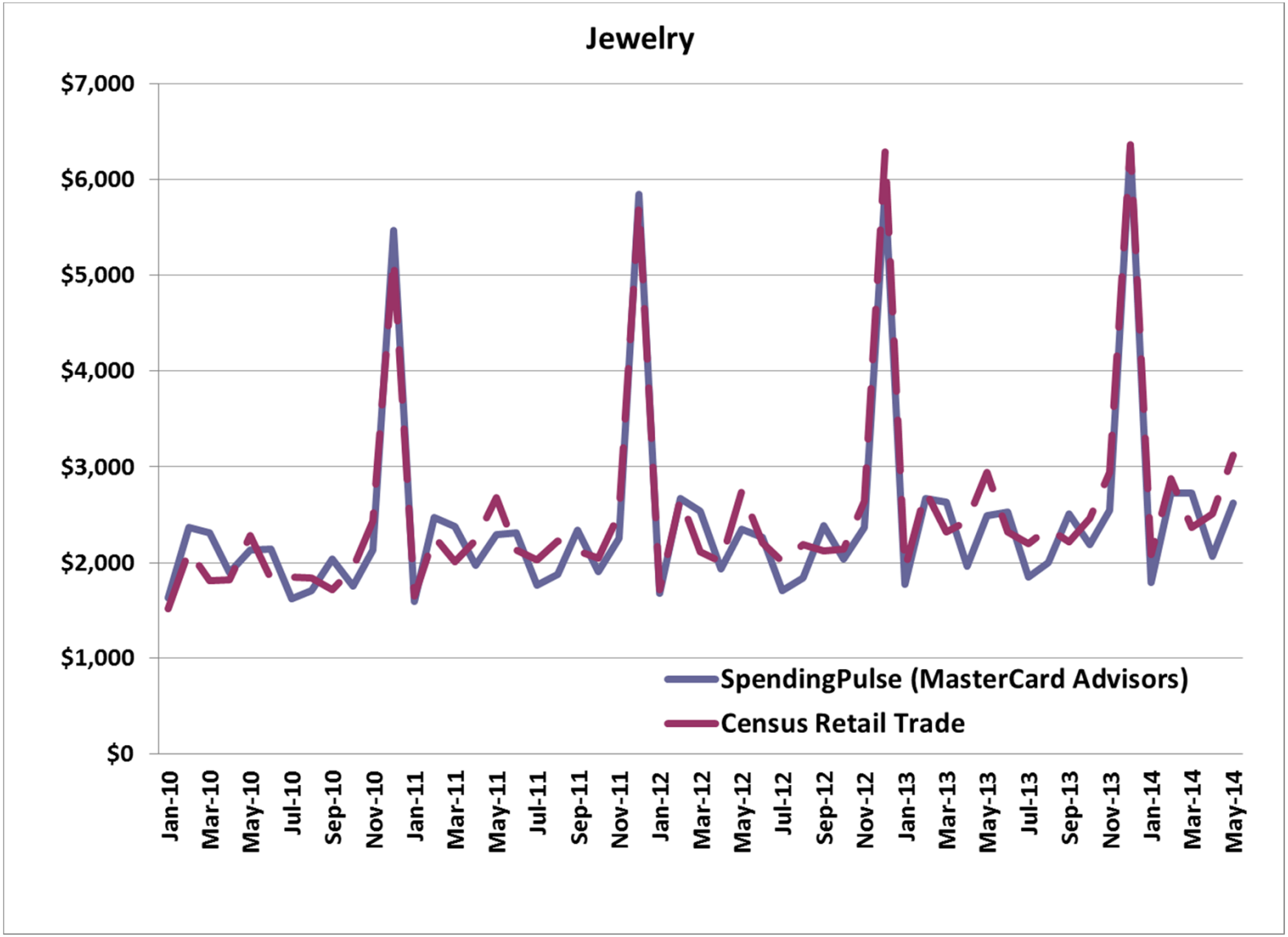
Example: Credit Card Data for Consumer Spending

- Using credit card data collected from the mandatory survey BE-150 to inform its estimates of international travel in the Balance of Payments Accounts
- Exploring use of credit card data to improve estimates of consumer spending, and to develop estimates at the metro area and county levels

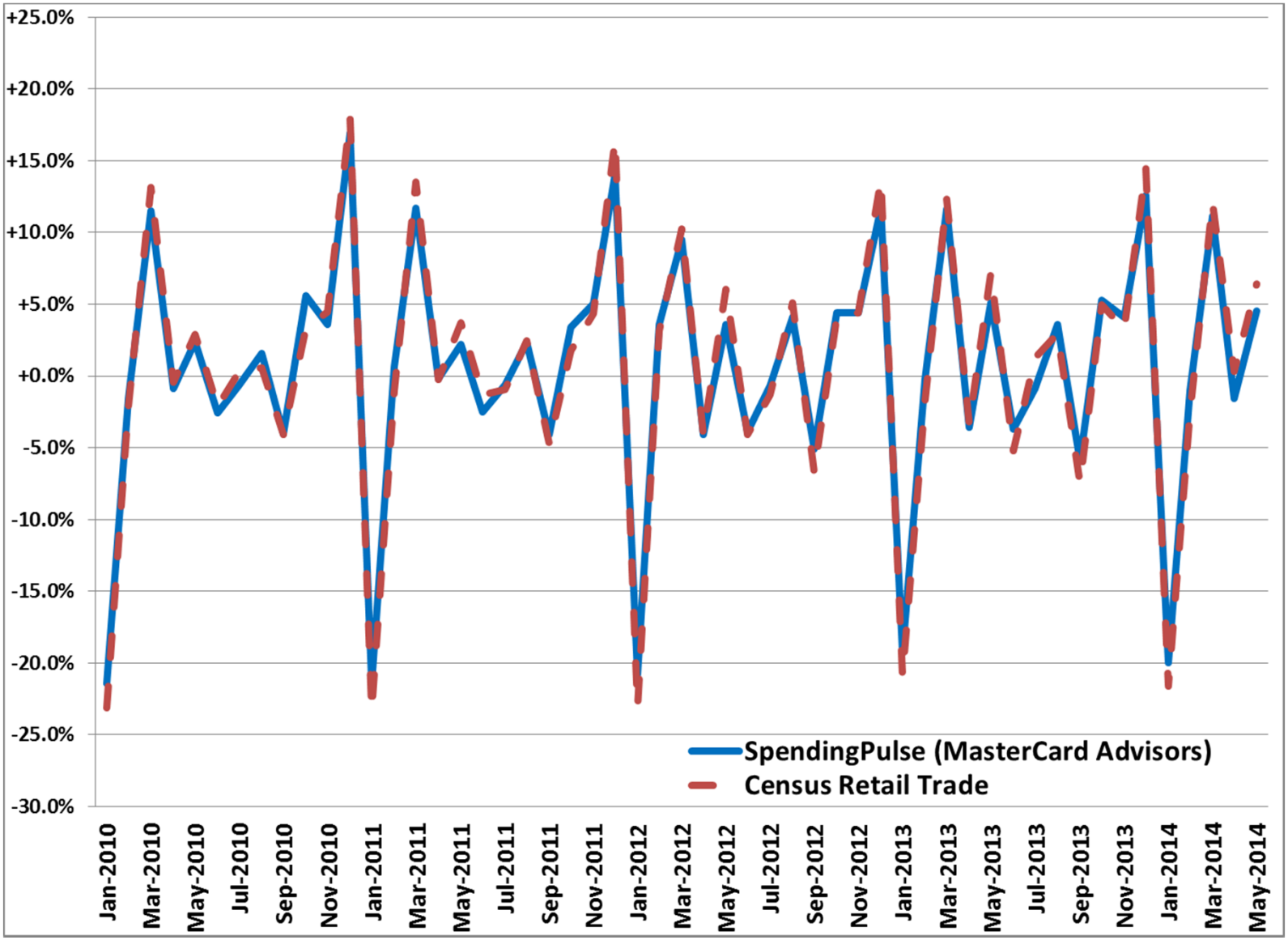
Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates



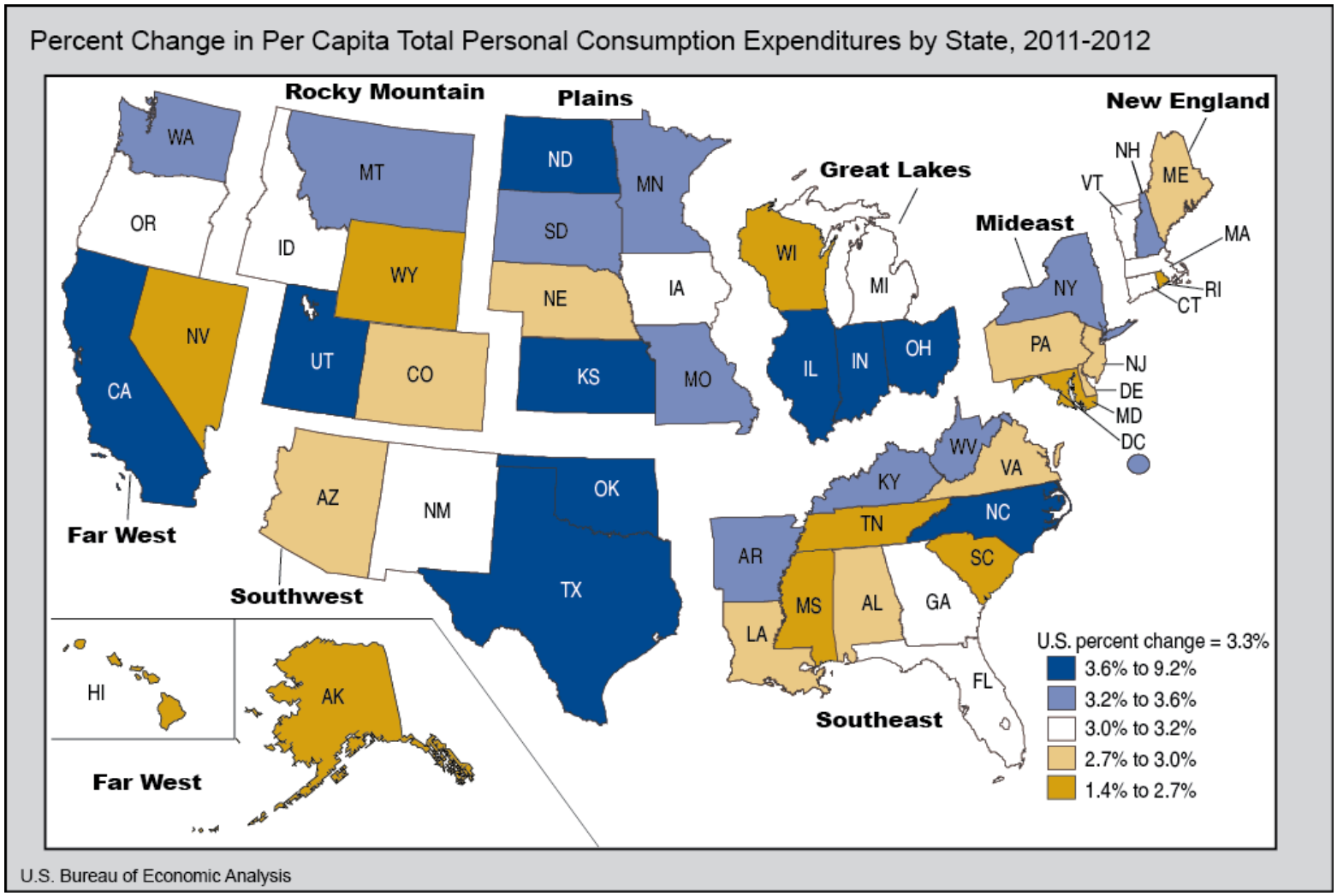
Estimates using Monthly Credit Card Data are similar to Retail Trade aggregates



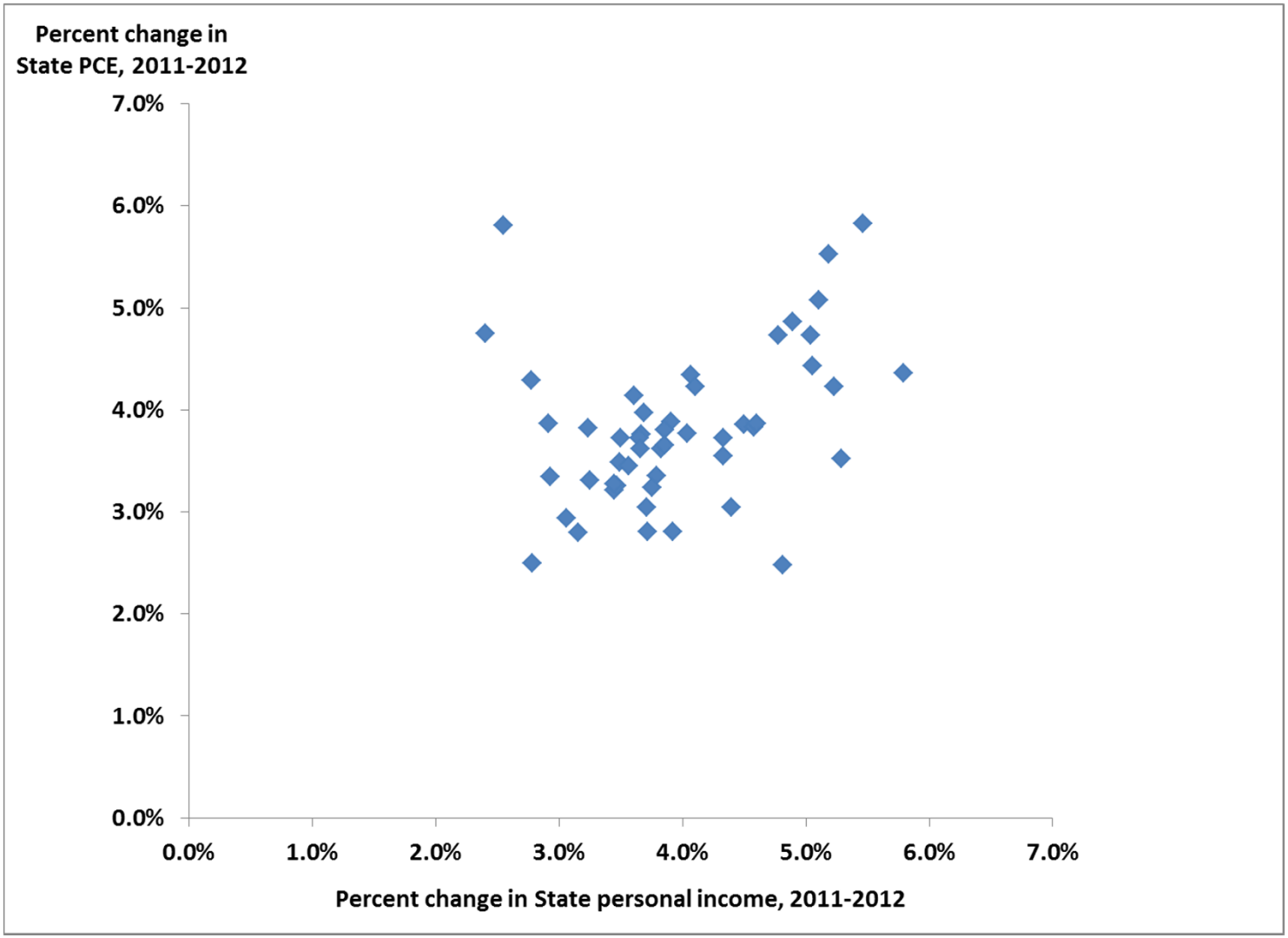
And the monthly changes in SpendingPulse total retail trade (less autos) are similar



Use Commercial data to improve State level estimates of PCE between Economic Censuses



Key is to link change in income to change in spending



Upcoming Meetings on Big Data

- CNSTAT Expert meeting on the use of commercial data in the national accounts, Oct/Nov 2015
  - Presentations by
    - Trivellore Raghunathan (University of Michigan)
    - Simon Wilkie (Microsoft)
    - Jonathan Parker (MIT)
    - Amir Sufi (University of Chicago)
- CNSTAT Panel "Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods"
  - Sponsor - The Laura and John Arnold Foundation
  - Chair – Robert Groves (Georgetown University)
  - Study Director – Brian Harris-Kojetin