INFO 7470/ECON 7400/ILRLE 7400 Alternate Data Sources of the 21st Century

John M. Abowd and Lars Vilhuber

March 4, 2013 and April 4, 2016

Rising Cost of Decennial Census

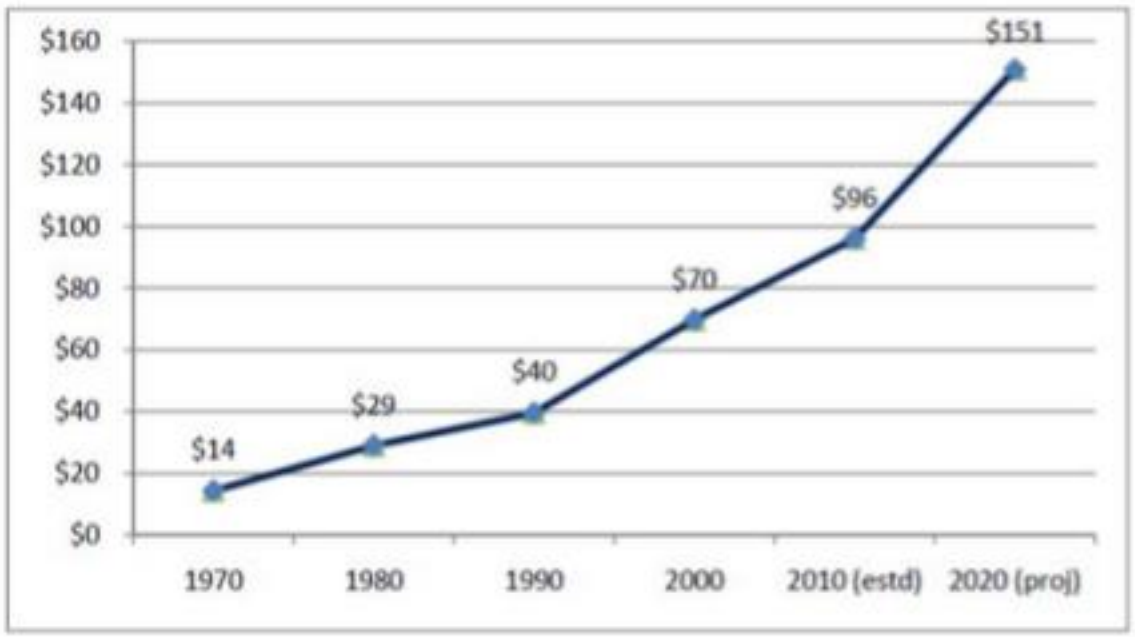


Figure 1: Cost per Housing Unit by Census Year, 1970 - 2020 (2010 dollars)

Vitrano and Chapin, 2012

Where Are Official Statistics Heading?

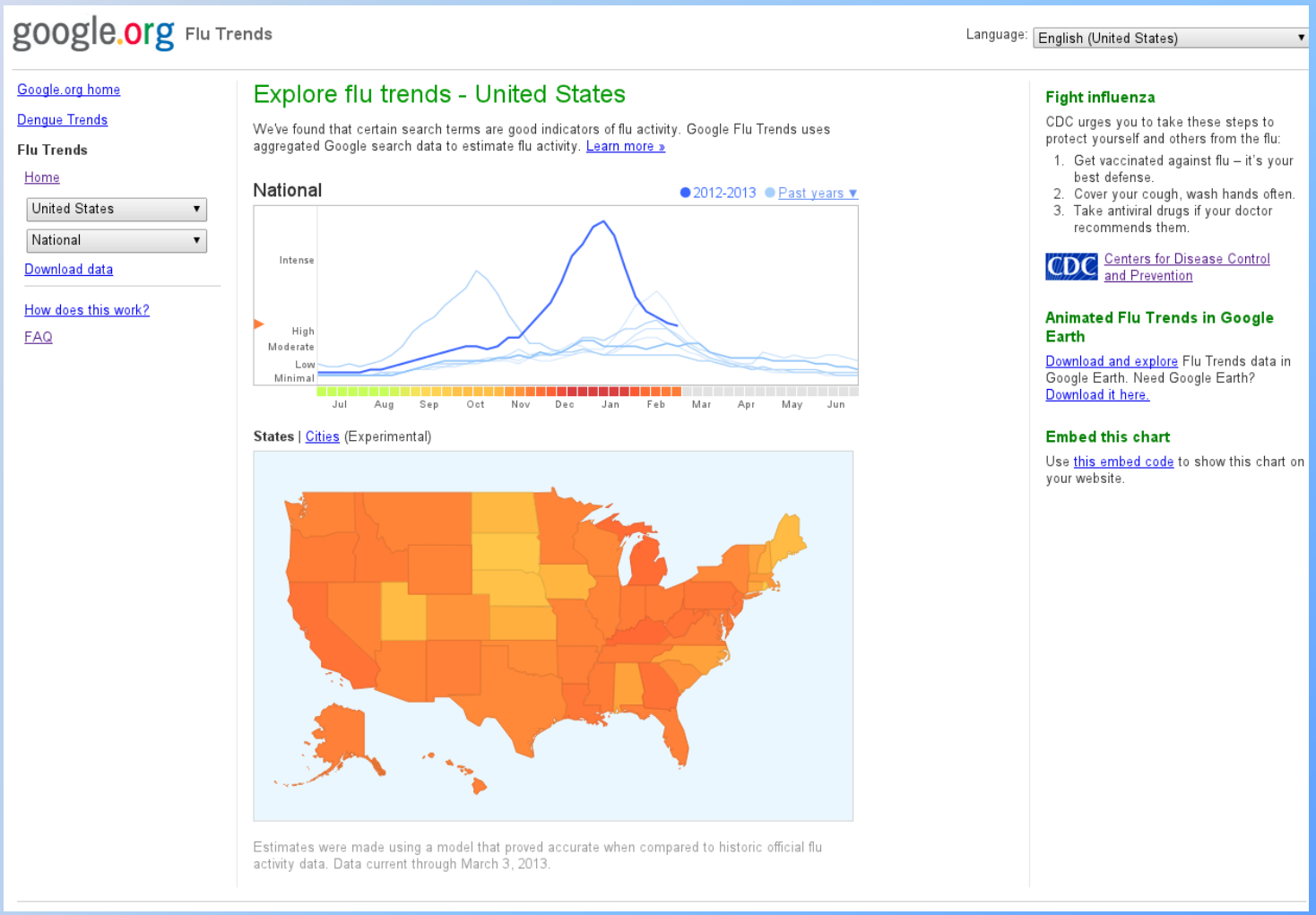
- On the negative side:
  - Increasing (?) privacy and confidentiality concerns
  - Decreasing response rates
    - Canada's Census long form is no longer compulsory
  - Increasing costs
- On the positive side:
  - Far more data available in general – petabytes per minute
  - "Big Data" is of commercial interest – ability to buy-in data

Consider Mapping

- In the 1980s
  - Census was the key provider of detailed demographic maps (TIGER)
- In 2013
  - Navteq, Google, Bing, *etc*., etc. provide highly detailed maps
  - Private provision of satellite imagery (SPOT, 1986)
    - GeoEye-1 provides 0.50m resolution for commercial customers

Google Flu Index

- http://www.google.org/flutrends/



Google Flu Index

- Compares to CDC's flu tracking http://www.cdc.gov/flu/weekly/overview.htm

Seasonal Influenza (Flu)

Flu Activity & Surveillance

Reports & Surveillance Methods in the United States

[Situation Update: Summary of Weekly FluView]

[Full FluView Report]

[FluView Interactive]

[Overview of Influenza Surveillance in the United States]

[Past Weekly Surveillance Reports and Historical Data]

Email page link
Print page
CDC on Facebook
CDC Flu on Twitter
Get email updates
Subscribe to RSS
Listen to audio/Podcast

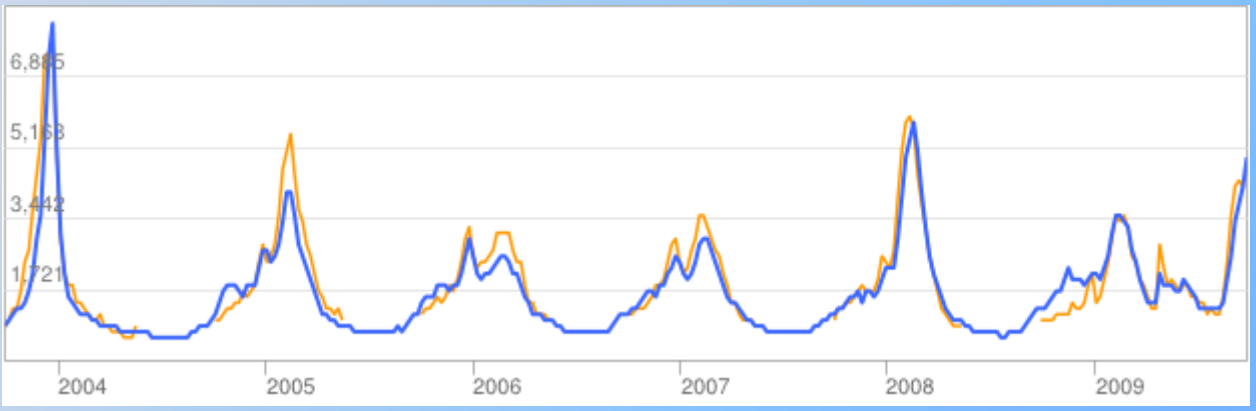
Contact Us:

Centers for Disease Control and Prevention
1600 Clifton Rd
Atlanta, GA 30333
800-CDC-INFO
(800-232-4636)
TTY: (888) 232-6348
Contact CDC-INFO

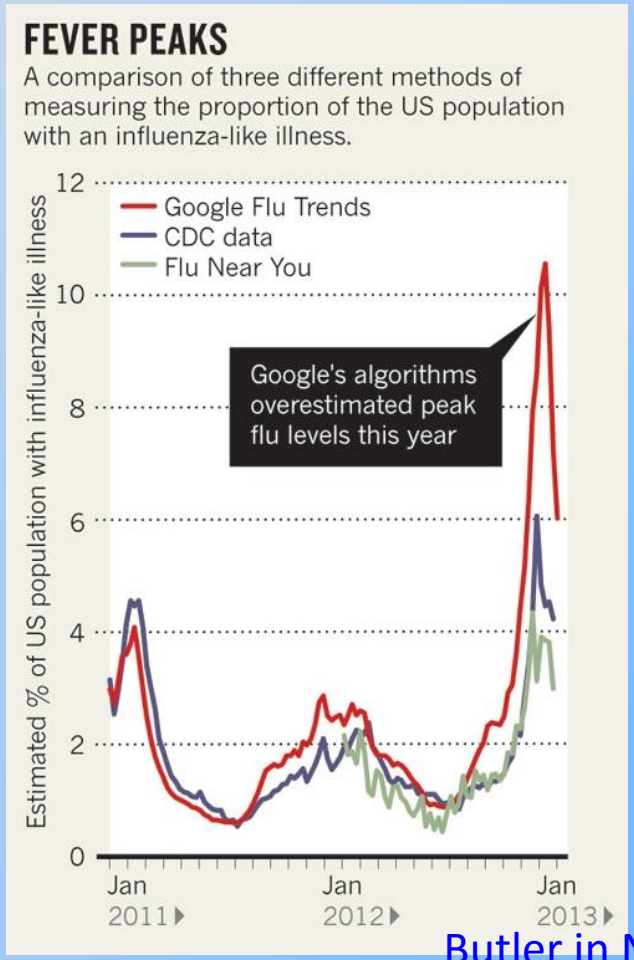
Google Flu Index

- How it works:
  - Uses search terms for flu-related topics
  - Validated against historical “high-quality” data (in the US: CDC, see doi:10.1038/nature07634)
- Advantages:
  - Very current (vs. some lag at CDC)
- Disadvantage:
  - May get side-tracked by changes in behavior/indicators *

Google Trends by Google



Latest Google Trends



Butler in Nature, Vol 494, Issue 7436

The Use of Non-traditional Data in Statistics

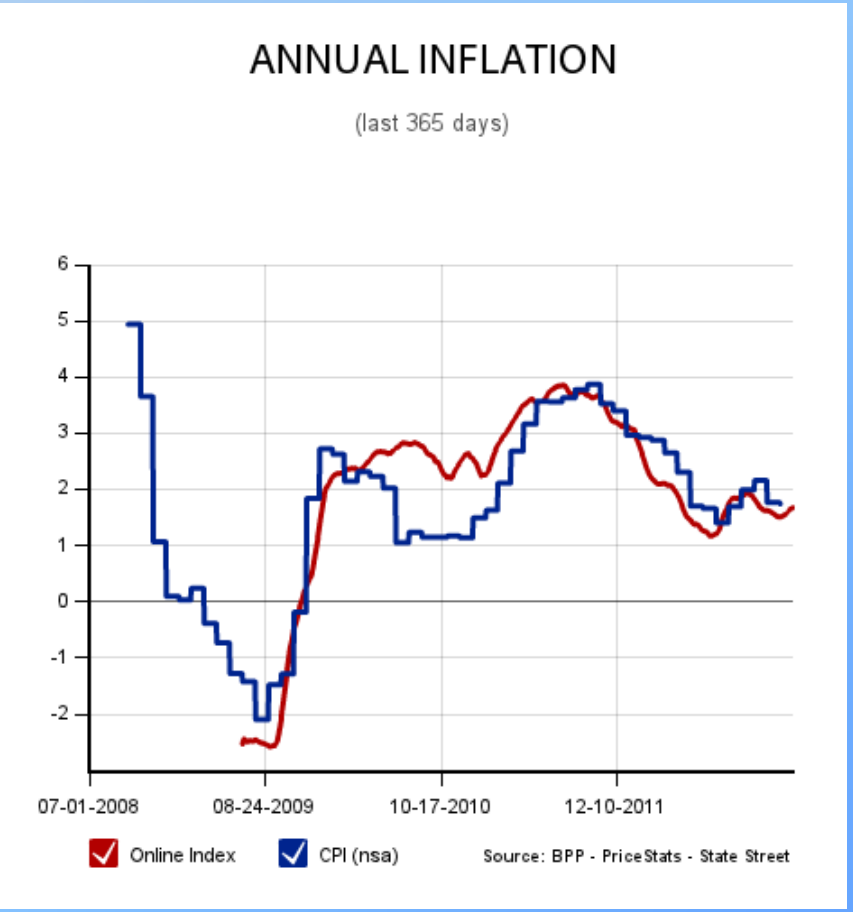
- Powerful, timely
- Power obtained in part through validation against "old-style" statistics
- Difference between
  - "raw data" (Google search terms)
  - First-order analytics (Google Flu trends)
  - Representativeness of the data – deep analysis and interpretation

Google Price Index

- Tracks database of web shopping data
- *The GPI shows a "pretty good correlation" with the CPI for goods such as cameras and watches that are often sold on the web, but less so for others, such as car parts, that are infrequently traded online.* (FT 2010)
- General approach: Choi and Varian (2011)

Billion Prices Project

- http://bpp.mit.edu
  - Alberto Cavallo and Roberto Rigodon (MIT)
- Methodology
  - Collected every day from online retailers
  - DB contains prices on the full array of products sold by these retailers, product descriptions, package sizes, brands, special characteristics (*e.g.*, "organic"), and whether the item is on sale or price control. *



Billion prices project

- Availability
  - PriceStats commercially sells data through State Street (for instance, used in the *Economist* for Argentina)
  - Research data available with a lag
  - Do not claim to cover 100% of CPI of goods and services (depend on online availability)
  - "Standard NSF funding would cover the costs for 1 day of web scraping"

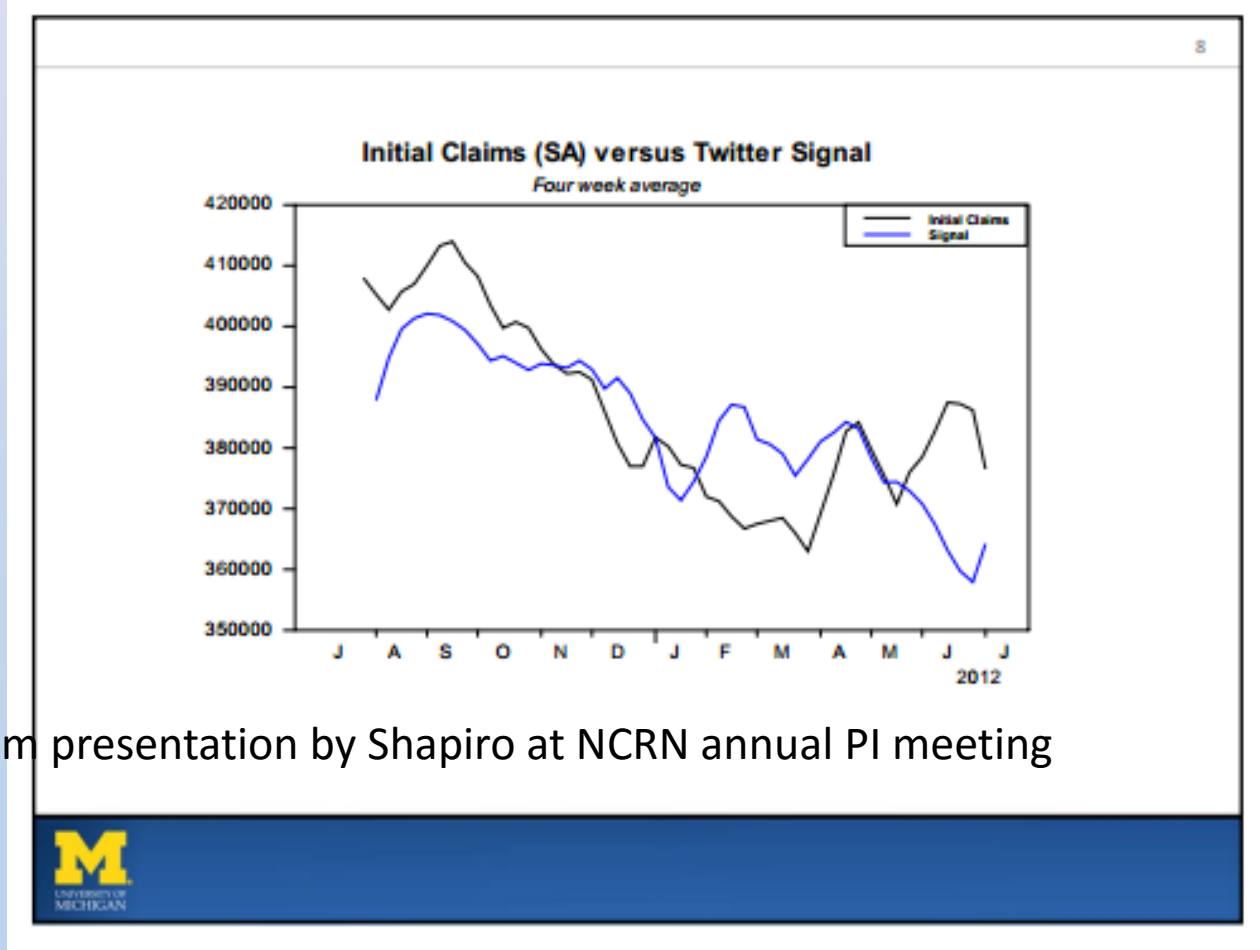
Billion prices project

- Use of similar methods by BLS to supplement CPI (*Amstat* article)
  - Quality adjustment models for televisions, camcorders, cameras, and washing machines.
  - Investigating retail scanner data for use in CPI

Twitter as a Statistical Data Source

- Work at the Michigan NCRN node by Matthew D. Shapiro and co-authors
- Use Twitter activity to create indicators of new job loss
- Benchmarked to weekly UI claims
- Critical: classification, search phrases

Initial UI Claims vs. Twitter Signal



Excerpt from presentation by Shapiro at NCRN annual PI meeting

New Sources, New Methods, …

- New unstructured data
- Convergence of qualitative and quantitative sciences
- Needs traditional statistics (registers!), but role may increase over time
  - All of the uses here are validating against official statistics where available
- Can supplement or replace official statistics where the latter are of lower quality (BPP for Argentina!)

… New Skill Sets

- Social scientists need to learn new tools
- INFO7470 students 2013
- Who knows how to leverage a 15,000 node Hadoop cluster?

	None	Advanced
SAS	31%	11%
Stata	25%	16%
R	56%	4%
Python	87%	1%
C, Fortran	67%	2%