Data-Intensive Research and Economics

Arnoldshain Seminar
Digitalization in Economics Roundtable

October, 2016

Ramiro H.Gálvez
Departamento de Computación, Universidad de Buenos Aires

The Rise of Data

Data analysis is on a hype right now.

The Rise of Data - Technology

On the one hand, this has been driven by a large increase in the capacity to store (+23% per year), communicate (+28% per year) and compute information (+58% per year).

The World's Technological Capacity to Store, Communicate, and Compute Information

Martin Hilbert[1]* and Priscila López[2]

We estimated the world's technological capacity to store, communicate, and compute information, tracking 60 analog and digital technologies during the period from 1986 to 2007. In 2007, humankind was able to store 2.9×10^{20} optimally compressed bytes, communicate almost 2×10^{21} bytes, and carry out 6.4×10^{18} instructions per second on general-purpose computers. General-purpose computing capacity grew at an annual rate of 58%. The world's capacity for bidirectional telecommunication grew at 28% per year, closely followed by the increase in globally stored information (23%). Humankind's capacity for unidirectional information diffusion through broadcasting channels has experienced comparatively modest annual growth (6%). Telecommunication has been dominated by digital technologies since 1990 (99.9% in digital format in 2007), and the majority of our technological memory has been in digital format since the early 2000s (94% digital in 2007).

Hilbert & López, *Science*, 2011

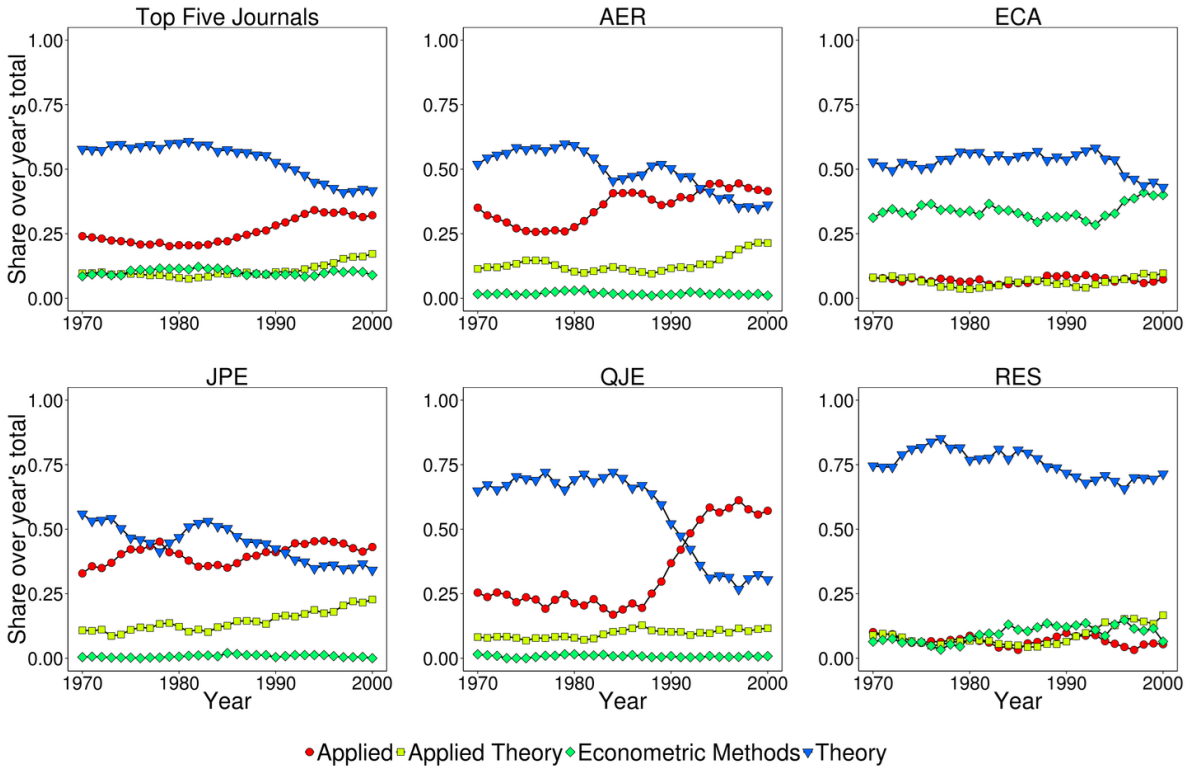
The Rise of Data - Applications in the Private Sector

On the other hand, this hype has been fueled by the private sector, which is giving a large importance (and assigning "big money") to the development of data analysis systems and techniques.

- Digital Advertisements
- Recommender Systems
- Speech Recognition
- Image Recognition
- Internet search
- Fraud and Risk Detection
- Churn analysis

Economists and Data Analysis

Data analysis is not new to economists.



Anauati, Galiani & Gálvez, *Economic Inquiry*, 2016

Big Data

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." (Gartner Inc., 2012)

Has this affected economic research?

Let see some examples

(Much of the rest of the presentation follows Einav & Levin, *Science*, 2014)

Real Time Analysis

First, as data is often available in real time, real-time analysis is possible.

Government surveys and statistics are released with a lag of months or years.

Private and administrative data that are continuously updated have great value for helping to guide economic policy and research.

Real Time Analysis

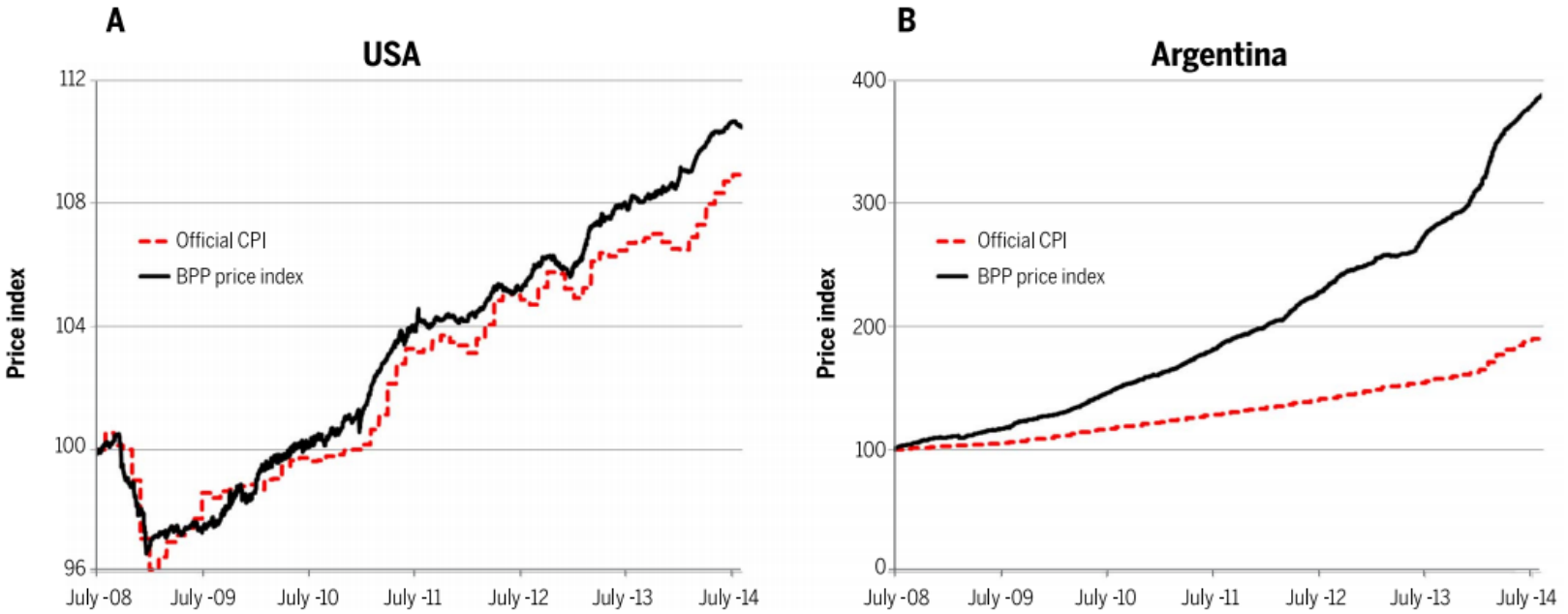
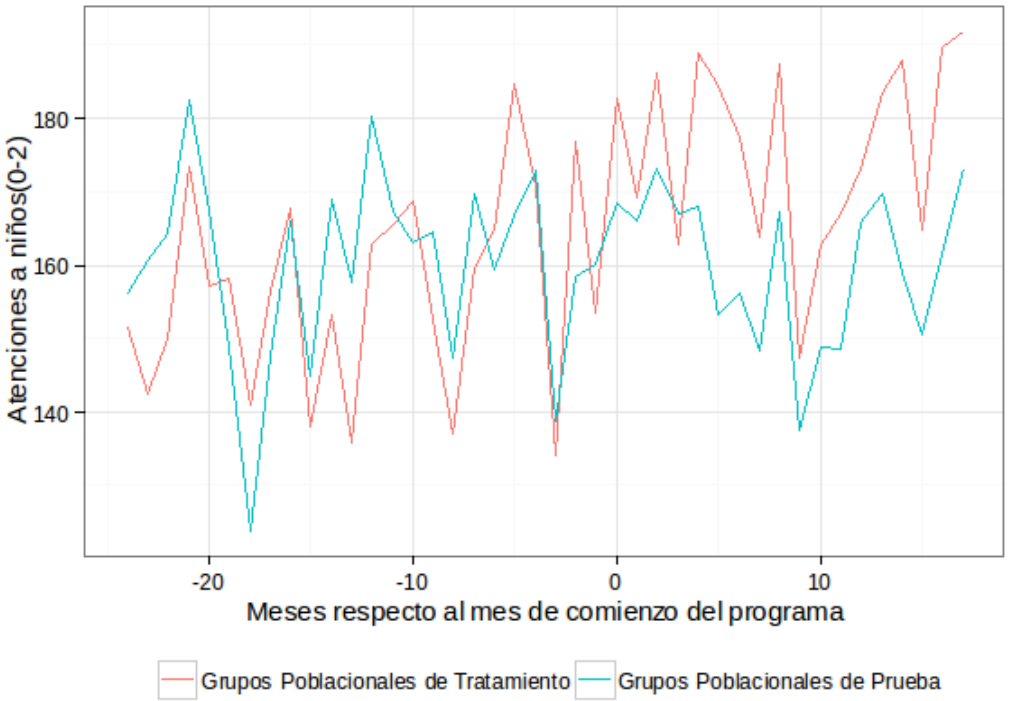


Fig. 2. BPP price index. Dashed red lines show the monthly series for the CPI in the United States (**A**) and Argentina (**B**), as published by the formal government statistics agencies. Solid black lines show the daily price index series, the "State Street's PriceStats Series" produced by the BPP, which uses scraped Internet data on thousands of retail items. All indices are normalized to 100 as of 1 July 2008. In the U.S. context, the two series track each other quite closely, although the BPP index is available in real time and at a more granular level (daily instead of monthly). In the plot for Argentina, the indices diverge considerably, with the BPP index growing at about twice the rate of the official CPI. [Updated version of figure 5 in (*18*), provided courtesy of Alberto Cavallo and Roberto Rigobon, principal investigators of the BPP]

Cavallo, *Journal of Monetary Economics*, 2013

Real Time Analysis



Galiani, Gálvez & Martínez, *IADB Technical Note (forthcoming)*, 2016

Data on Previously Unmeasured Activities

Second, data are available on previously unmeasured activities.

Much of the data now being recorded is on activities that were previously difficult to quantify, such as:

- Personal communications
- Social networks
- Search and information gathering
- Geolocation data, etc.

Data on Previously Unmeasured Activities

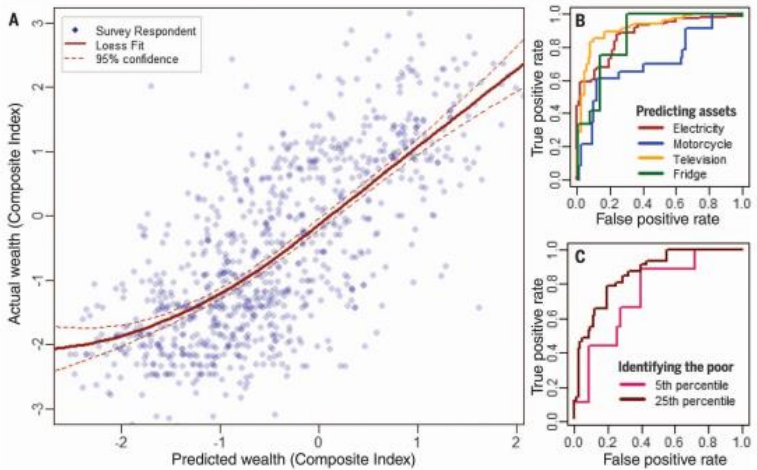


Fig. 1. Predicting survey responses with phone data. (**A**) Relation between actual wealth (as reported in a phone survey) and predicted wealth (as inferred from mobile phone data) for each of the 856 survey respondents. (**B**) Receiver operating characteristic (ROC) curve showing the model's ability to predict whether the respondent owns several different assets. AUC values for electricity, motorcycle, television, and fridge, respectively, are as follows: 0.85, 0.67, 0.84, and 0.88. (**C**) ROC curve illustrates the model's ability to correctly identify the poorest individuals. The poor are defined as those in the 5th percentile (AUC = 0.72) and the 25th percentile (AUC = 0.81) of the composite wealth index distribution.

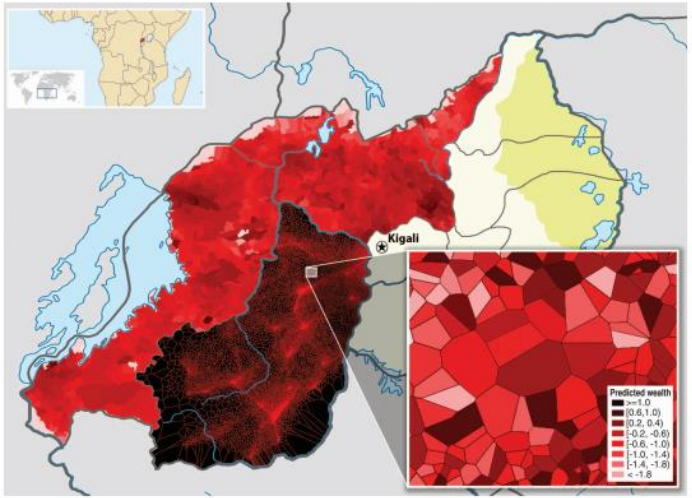
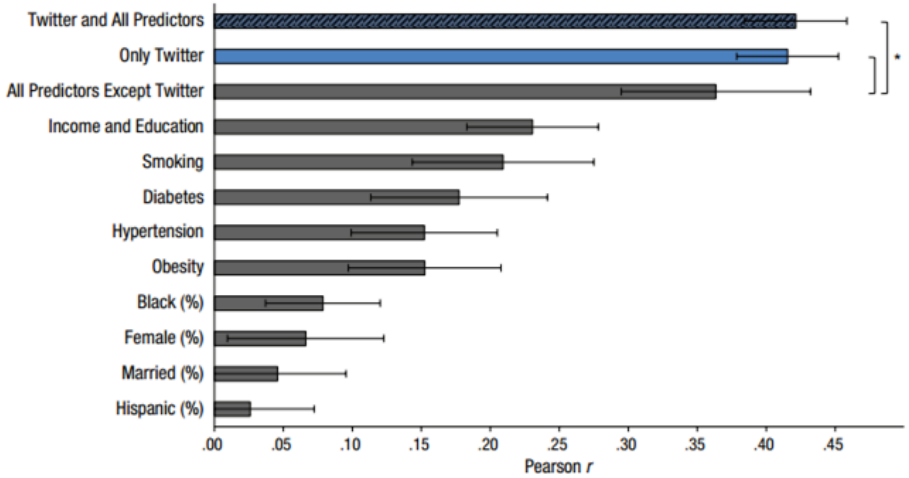
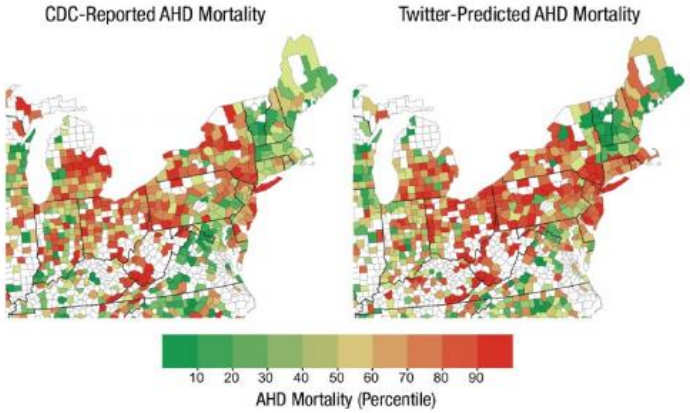


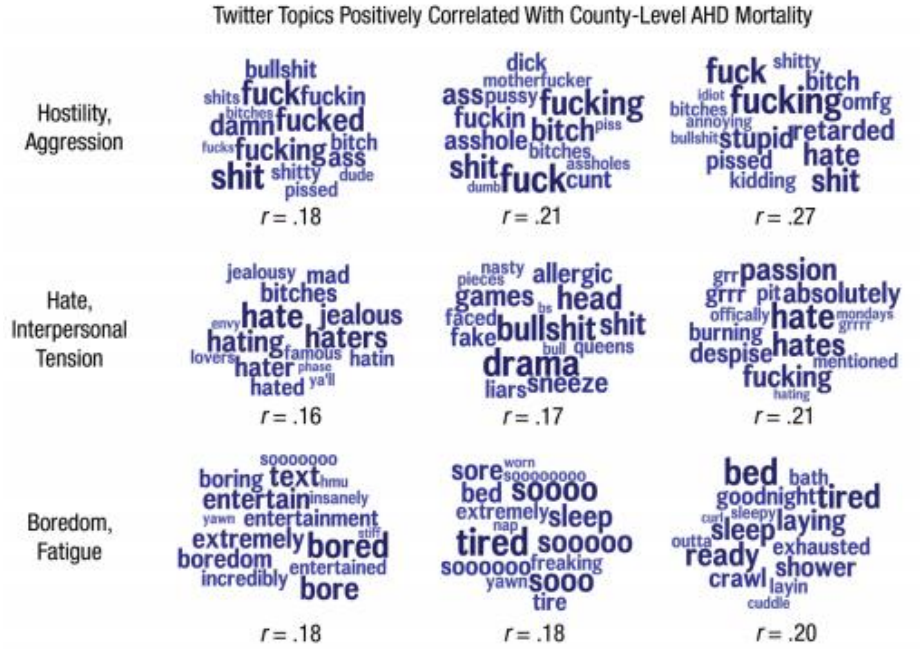
Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km^2 region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

Blumenstock, Cadamuro & On, *Science*, 2015

Data on Previously Unmeasured Activities







Eichstaedt et al., *Psychological Science*, 2015

Unstructured and High Dimensional Data

Data come with less structure.

Economists are used to working with "rectangular" data, with N observations and K << N variables per observation and a relatively simple dependence structure between the observations.

New data sets often have higher dimensionality and less-clear structure.

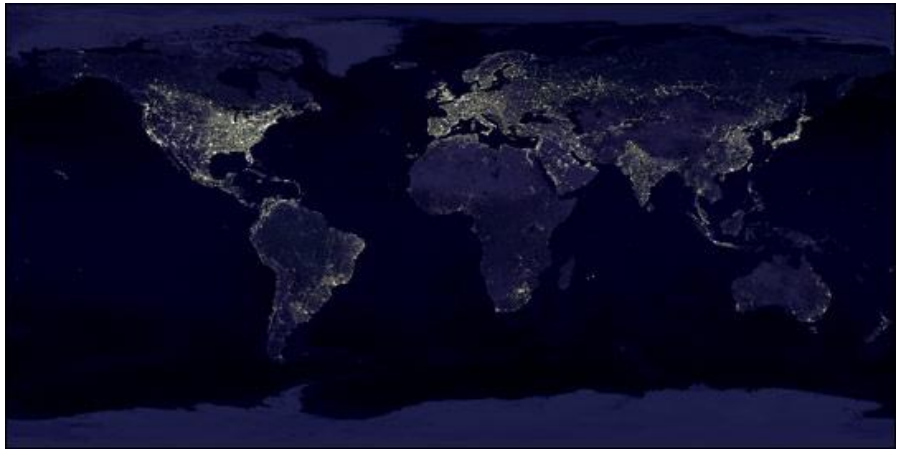
Unstructured and High Dimensional Data

Fig. 2. Example of a post.

Fig. 8. Influential tokens in the topics detected as predictive.

Gálvez & Gravano, *Unpublished Manuscript*, 2016

Unstructured and High Dimensional Data



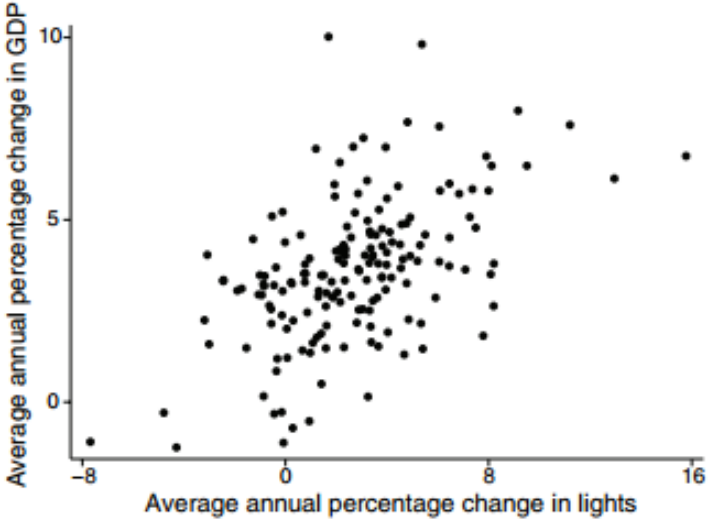


FIGURE 1. GDP VERSUS LIGHTS:
LONG DIFFERENCES 1992–1993 TO 2005–2006

Henderson, Storeygard & Weil, *The American Economic Review*, 2011

Unstructured and High Dimensional Data

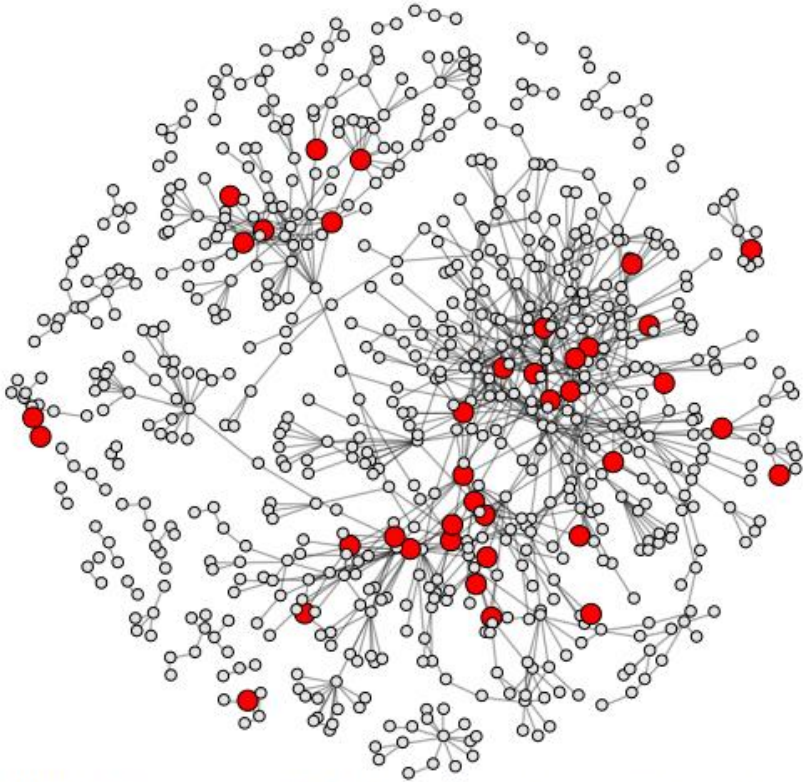
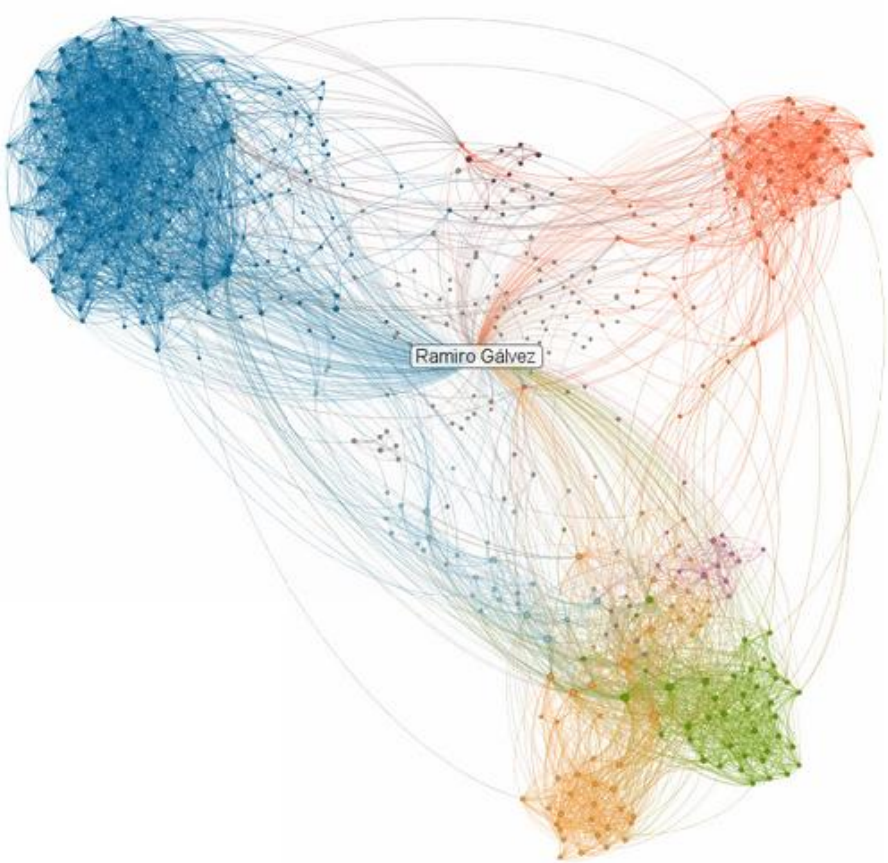


FIGURE 2. The social network of high-risk individuals in Cape Verdean community in Boston, 2008.

Papachristos, Braga & Hureau, *Journal of Urban Health*, 2012

Unstructured and High Dimensional Data



My Linkedin profile around 2011!

Prediction Policy Problems

Finally, everything points toward a greater importance of prediction in public policy problems.

Mullainathan et al. 2015 go as far as to name some policy problems as "Prediction Policy Problems".

TABLE 1—RISKIEST JOINT REPLACEMENTS

Predicted mortality percentile	Observed mortality rate	Futile procedures averted	Futile spending ($ mill.)
1	0.435 (0.028)	1,984	30
2	0.422 (0.028)	3,844	58
5	0.358 (0.027)	8,061	121
10	0.242 (0.024)	10,512	158
20	0.152 (0.020)	12,317	185
30	0.136 (0.019)	16,151	242

Mullainathan et al., *The American Economic Review*, 2015

Conclusions

It's a great moment for doing applied research!

There's plenty of data available.

Which skills are needed in order to deal with it?

- Proficiency in programming (Bye Stata!… R is good, but already kind of old… Python is probably the way to go).
- Feeling comfortable with statistics and data wrangling.
- Knowing how to manipulate large amounts of data.
- Knowing how to scrap online data is a must.
- Knowing machine learning is a plus.