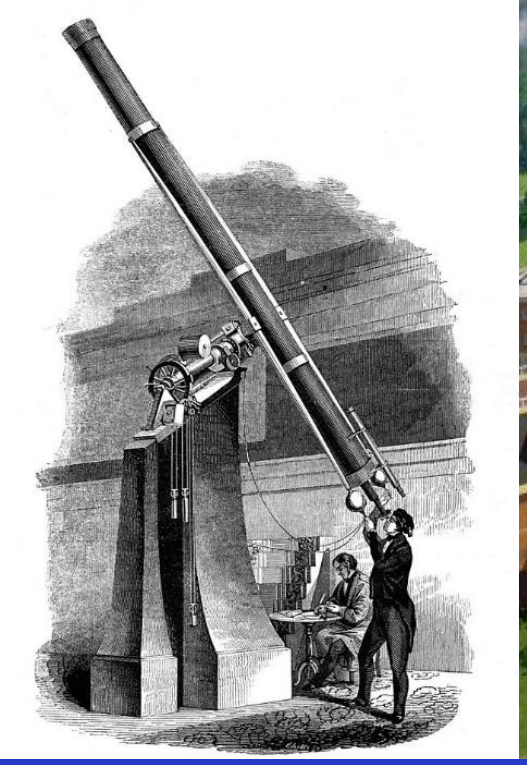


Curses and Blessings of an (almost) Data-Complete Science: Big Data and the Social Sciences

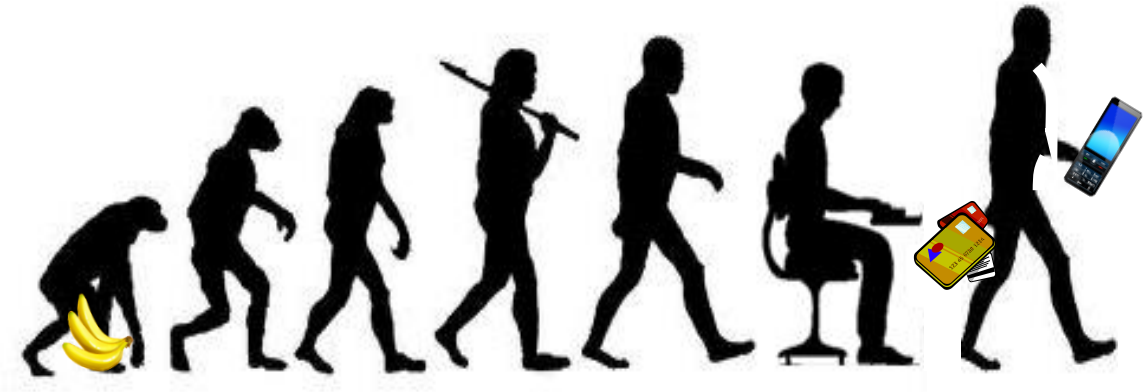
UCDAVIS
UNIVERSITY OF CALIFORNIA

Martin Hilbert
Department of Communication
hilbert@ucdavis.edu

Abstract:

Big Data has turned the social sciences from a traditionally data-poor science into arguably the most data-complete science to date – and this basically “overnight”. With over 99% of all of human kinds’ technologically mediated information in digital format, and a mobile penetration of 98% worldwide, the digitalization of human interaction produces an impressively detailed digital footprint of everything that’s relevant for the social sciences. Each and every digital communication inevitably leaves a trace that can be analyzed to better understand and influence social conduct. This renders many traditional survey and data collection and production processes obsolete. While creating unprecedented opportunities for private actors and lots of low hanging fruits for academic research, it also creates challenges that call for a profound paradigm shift in our relation to data.

Computational Social Science



Volume 493 Number 7432 pp271-446 17 January 2013

"The **biggest stumbling block**… **is obtaining the data** to parameterize and validate…
…using automated cameras and image recognition…motion-activated cameras… continuous plankton recorders towed beneath ships…"

A hyena surveys a flock of flamingos in South Africa.

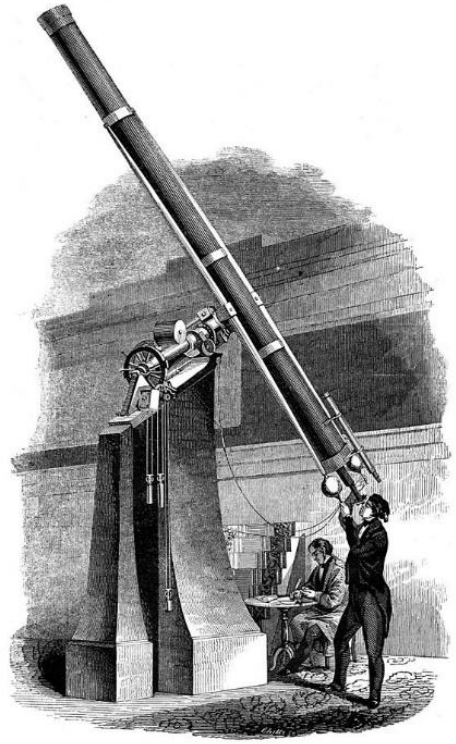
FAQ's

What, in a nutshell, is 'The Madingley Model'?

A huge computer simulation of all life on Earth.

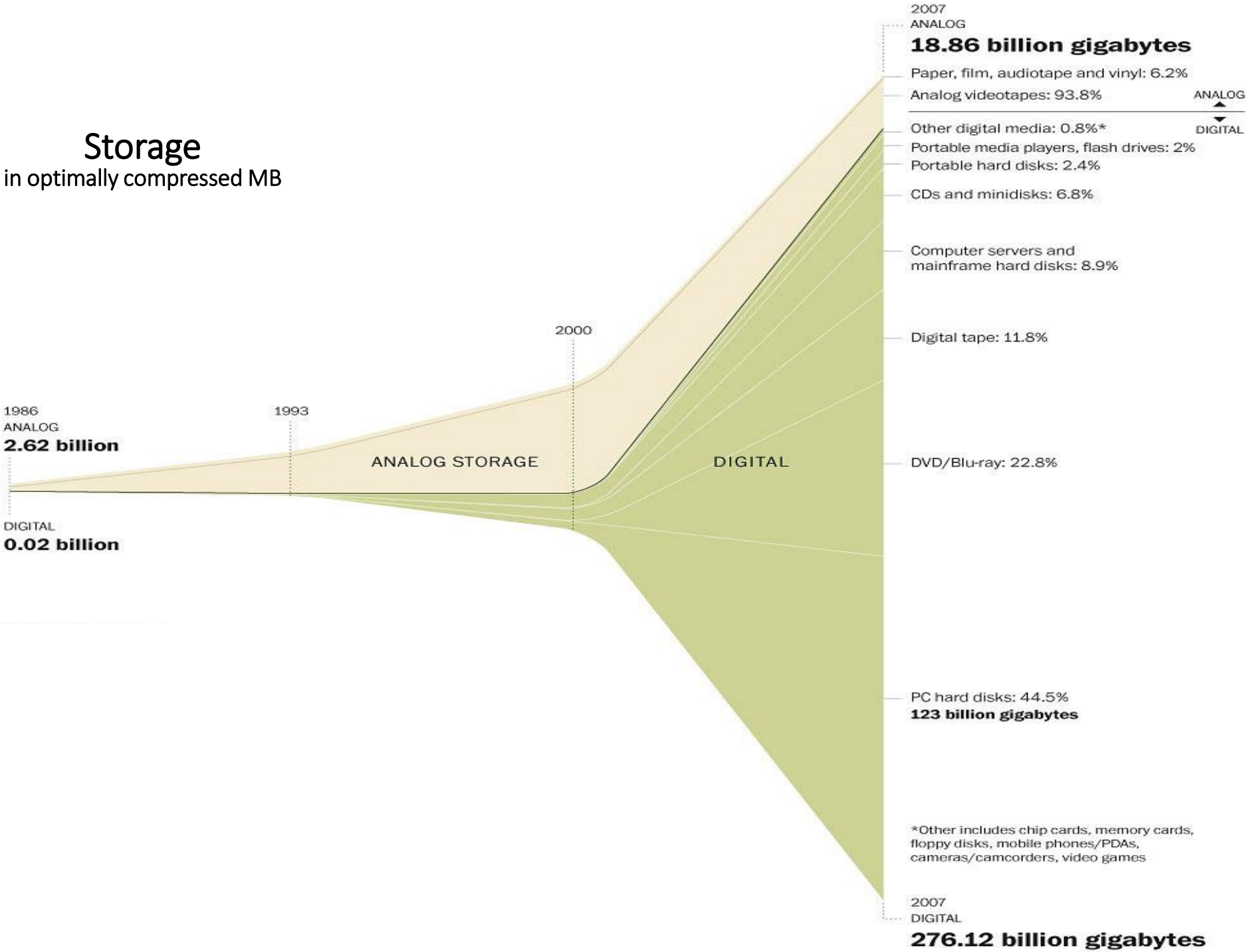
Time to model all life on Earth

To help transform our understanding of the biosphere, ecologists — like climate scientists — should simulate whole ecosystems, argue Drew Purves and colleagues.



Storage

in optimally compressed MB



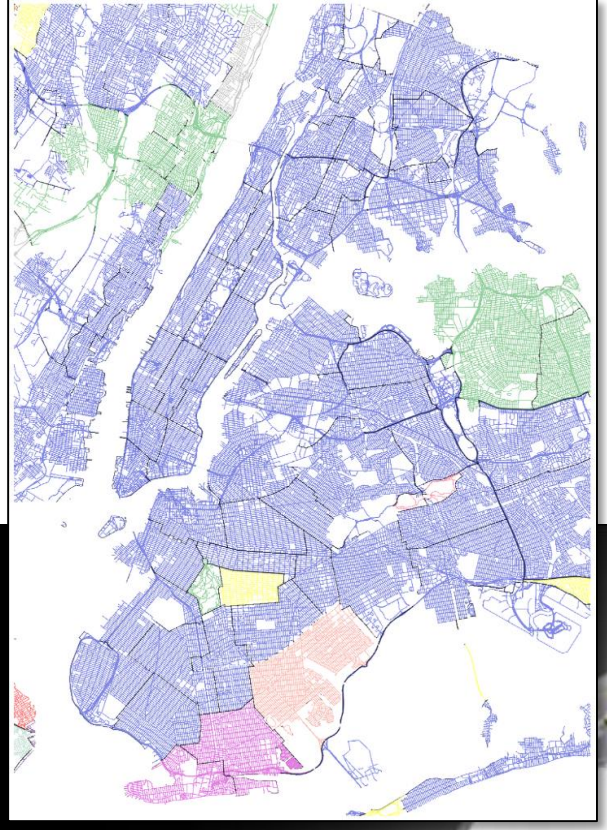
Source: animation by Washington Post, based on
Hilbert, M., P. López (04/2011). The world's technological capacity to store, communicate and compute information. Science, 332, 6025, 60-65
www.martinhilbert.net/WorldInfoCapacity.html

- **Digital Footprint** (real interactions recorded anyways)
- **n = N** (no sampling, but potential bias)
- **Data-fusion** (unstructured and incomplete)
- **Real-time** (dynamic)
- **Spooky accuracy through ML** (no need for theory)

Source: animation by Washington Post, based on
Hilbert, M., P. López (04/2011). The world's technological capacity to store, communicate and compute information. Science, 332, 6025, 60-65
www.martinhilbert.net/WorldInfoCapacity.html

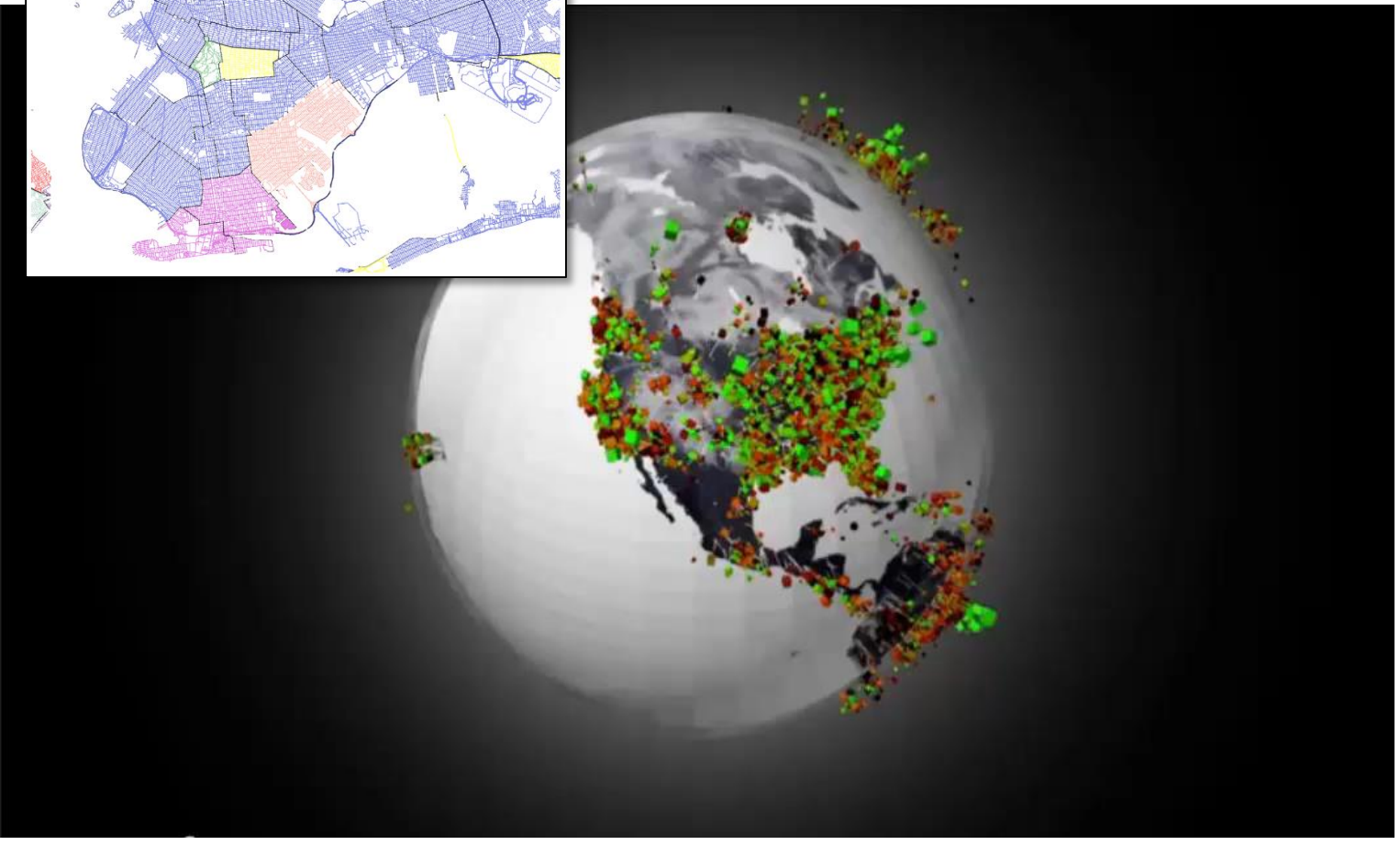
Source: Hilbert, M. (2015). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*.

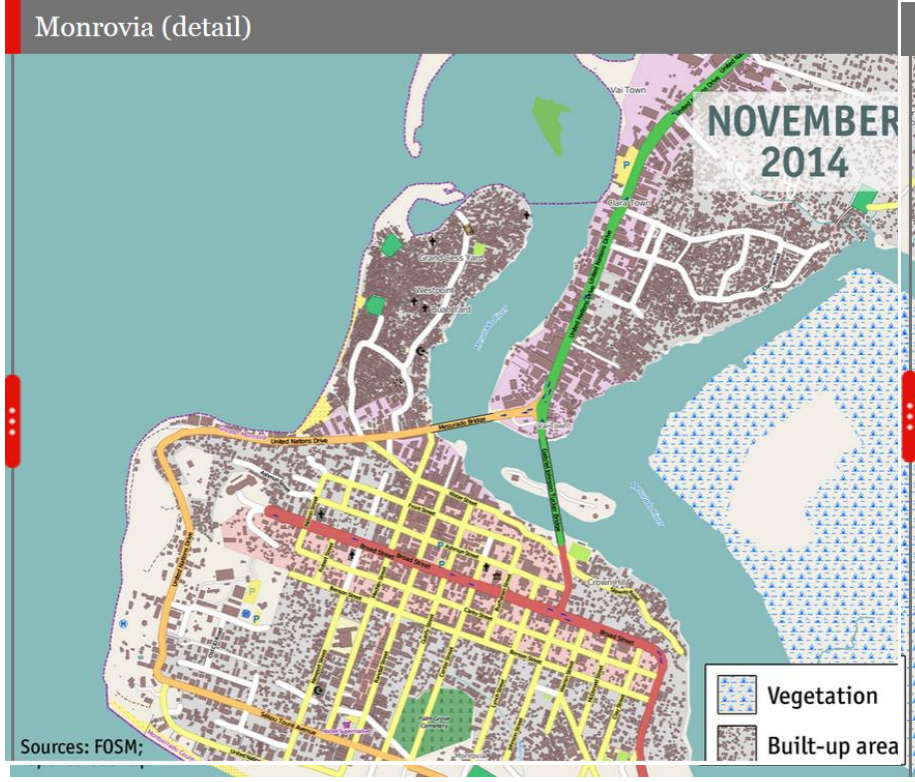
Digital Footprint



TWITTER: 2nd language

Blue - Spanish,
Light Green - Korean,
Fuchsia - Russian,
Red - Portuguese,
Yellow - Japanese,
Pink - Dutch,
Grey - Danish,
Coral - Indonesian.





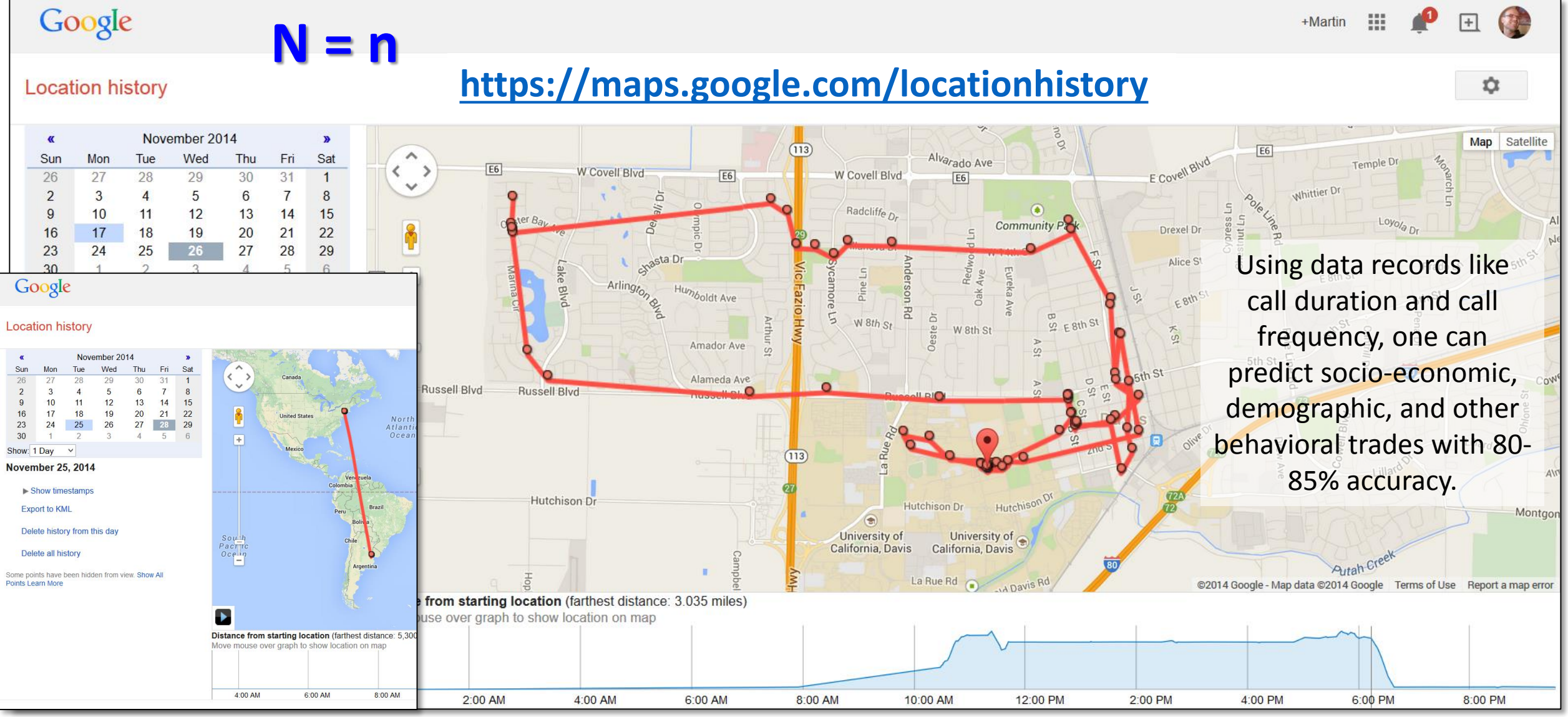
TED-Ed. (2013). Visualizing the world's Twitter data - Jer Thorp. http://www.youtube.com

The Economist. (2014, November 15). Off the map. *The Economist*. http://www.economist.com

Mocanu, et al.(2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, *8*(4), e61981.

N = n

https://maps.google.com/locationhistory



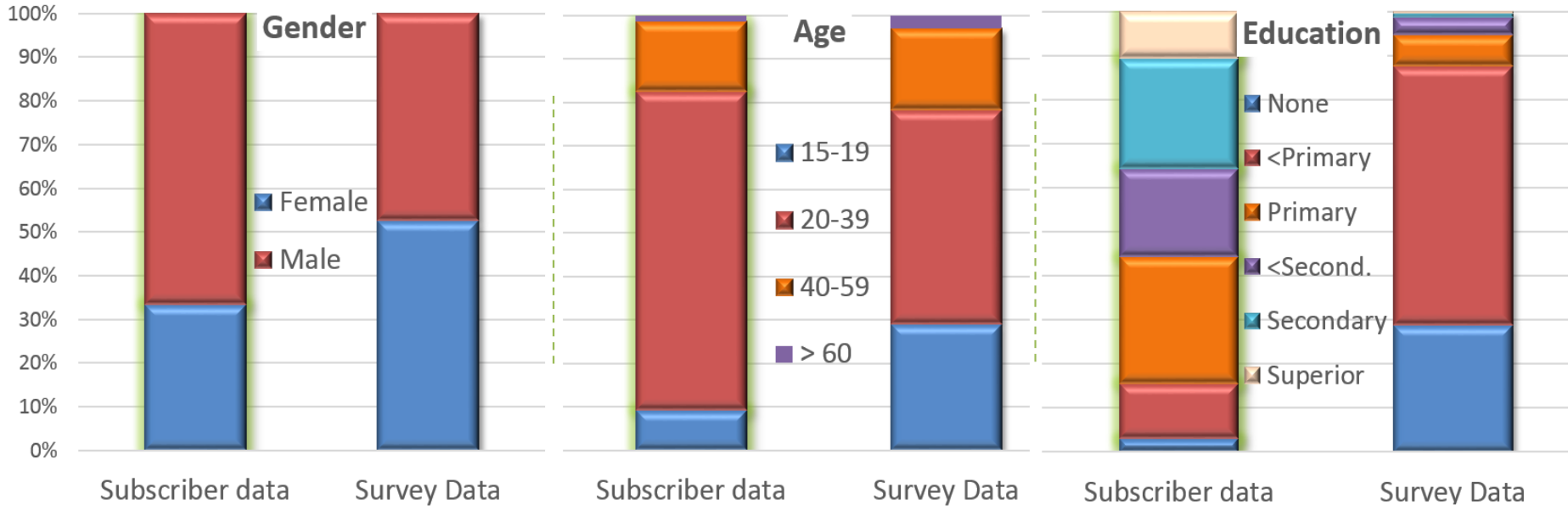
"…human mobility traces are highly unique. …four spatio-temporal points are enough to uniquely identify 95% of the individuals. "

Sources: Raento, et al. (2009). Smartphones: *Sociol. Methods & Research*, *37*(3), 426–454. Frias-Martinez, et al. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engin. Appl. of Artificial Intell.*, 35, 237–245. Frias-Martinez, et al. (2013). Cell Phone Analytics: *ITID*, 9(2), pp. 35–50. Frias-Martinez, et al. (2010). A Gender-centric Analysis of Calling Behavior…. AAAI 201 *Artificial Intelligence for Development*. Blumenstocket al. (2010). Who's Calling? AAAI 201 *Artificial Intelligence for Development.* De Montjoye, et al. (2013). Unique in the Crowd: Scientific Reports, 3

N = n ?

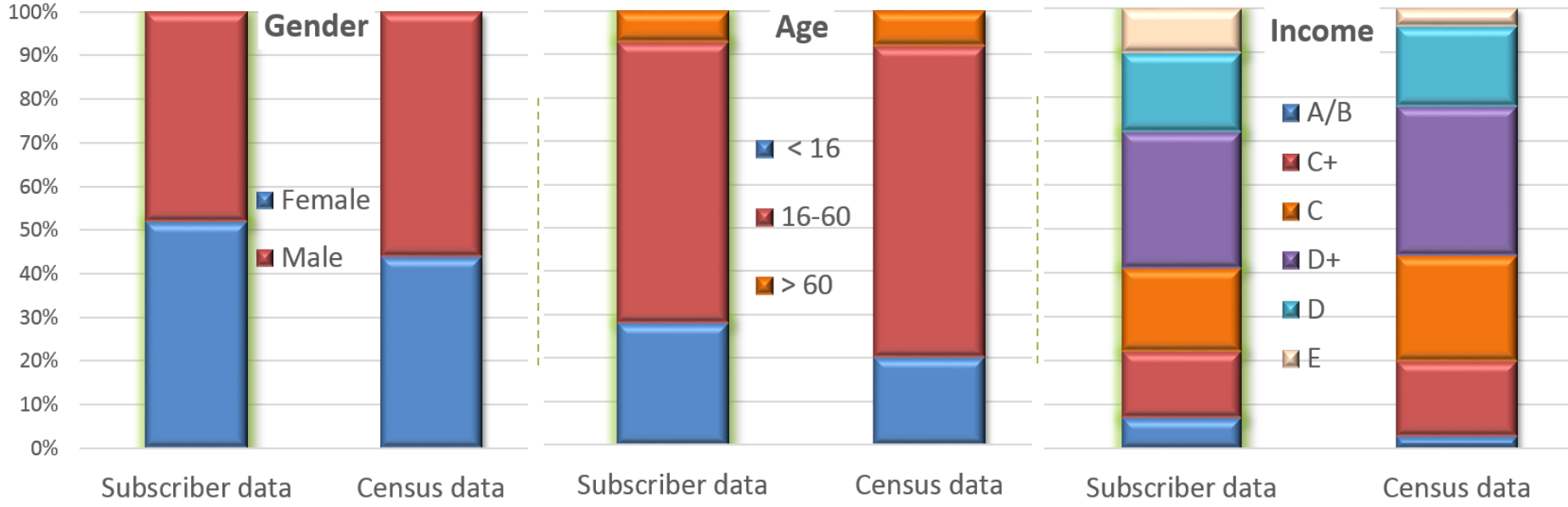
(a) Rwanda 2005/09:

mobile phone penetration of 2-20%



(b) LatAm economy 2009/10:

mobile phone penetration of 60-80%



Source: (a) Blumenstock and Eagle (2012); (b) Frias-Martinez and Virseda (2013).

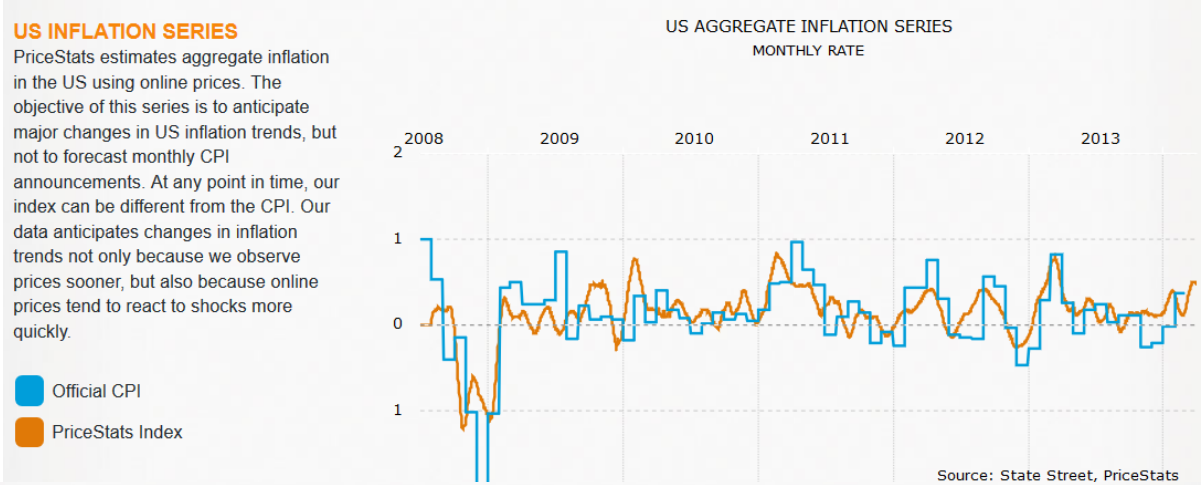
Real-time

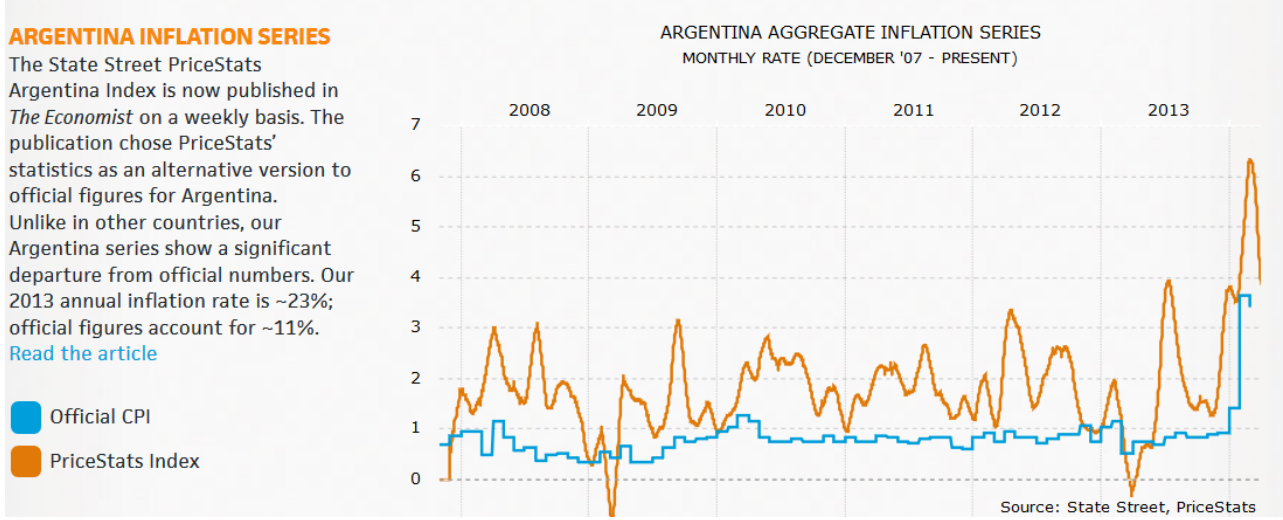
U.S. Bureau of Labor Statistics

- > 100s staff visiting > 90 cities
- 80,000 prices
- Cost: US$ 250 million/year

PriceStats

- 17 staff
- 300 online retailers; > 70 countries
- daily 5,000,000 prices

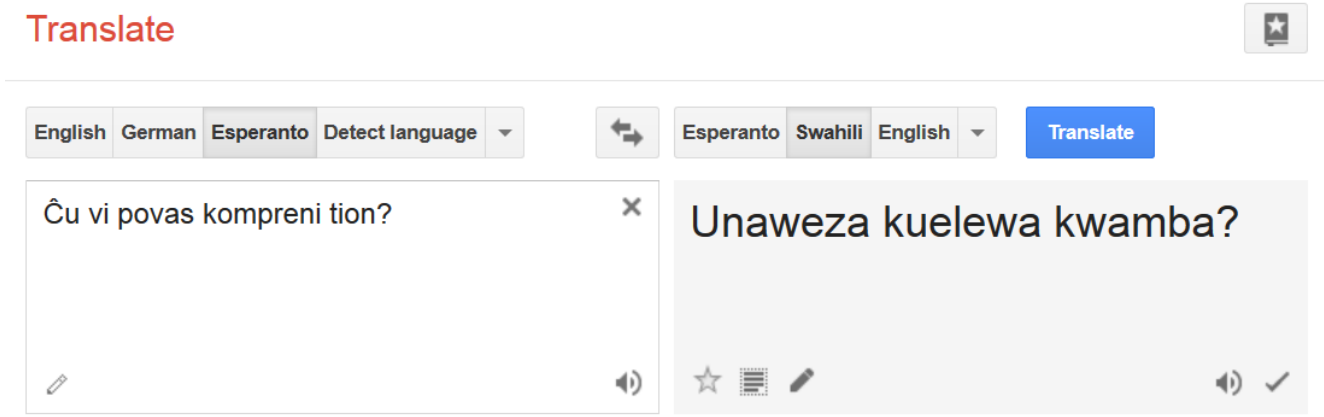




Sources:
http://www.pricestats.com/about-us/meet-the-team ;
www.economist.com/node/21548242 ;
http://www.inflacionverdadera.com

“Big” doesn’t need to know Why

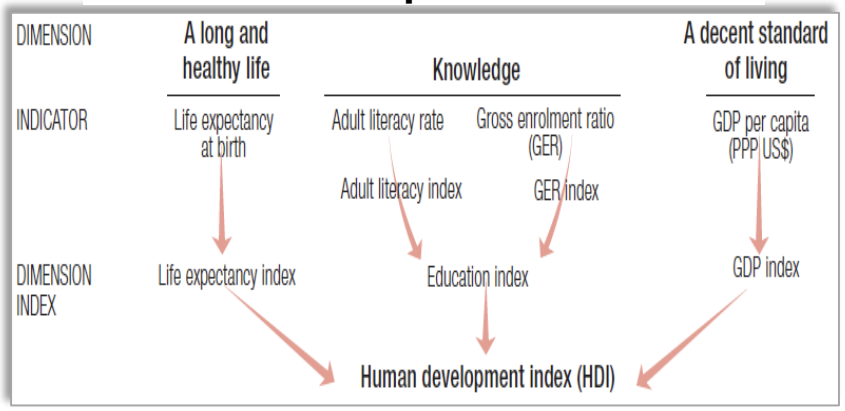
Choose a representation that can use unsupervised learning on unlabeled data, which is so much more plentiful than labeled data.



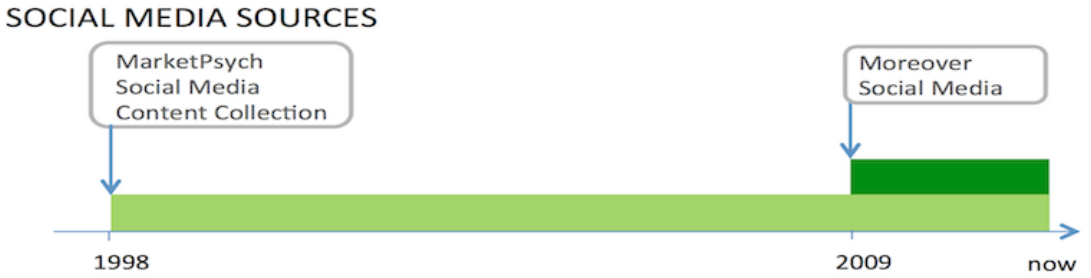
Afrikaans	Cebuano	Finnish	Hmong	Korean	Nepali	Somali	Welsh
Albanian	Chinese (Simplified)	French	Hungarian	Lao	Norwegian	Spanish	Yiddish
Arabic	Chinese (Traditional)	Galician	Icelandic	Latin	Persian	Swahili	Yoruba
Armenian	Croatian	Georgian	Igbo	Latvian	Polish	Swedish	Zulu
Azerbaijani	Czech	German	Indonesian	Lithuanian	Portuguese	Tamil	
Basque	Danish	Greek	Irish	Macedonian	Punjabi	Telugu	
Belarusian	Dutch	Gujarati	Italian	Malay	Romanian	Thai	
Bengali	English	Haitian Creole	Japanese	Maltese	Russian	Turkish	
Bosnian	**Esperanto**	Hausa	Javanese	Maori	Serbian	Ukrainian	
Bulgarian	Estonian	Hebrew	Kannada	Marathi	Slovak	Urdu	
Catalan	Filipino	Hindi	Khmer	Mongolian	Slovenian	Vietnamese	

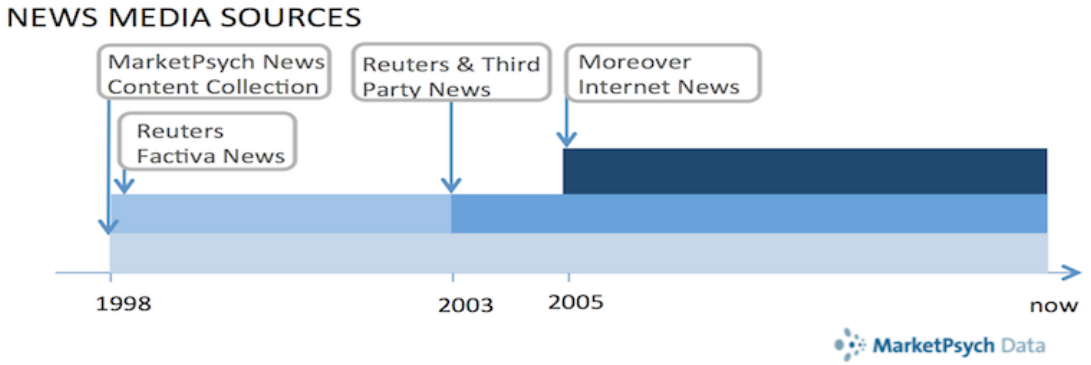
Sources: Banko, M., & Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 26–33). Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12.

Human Development Index HDI



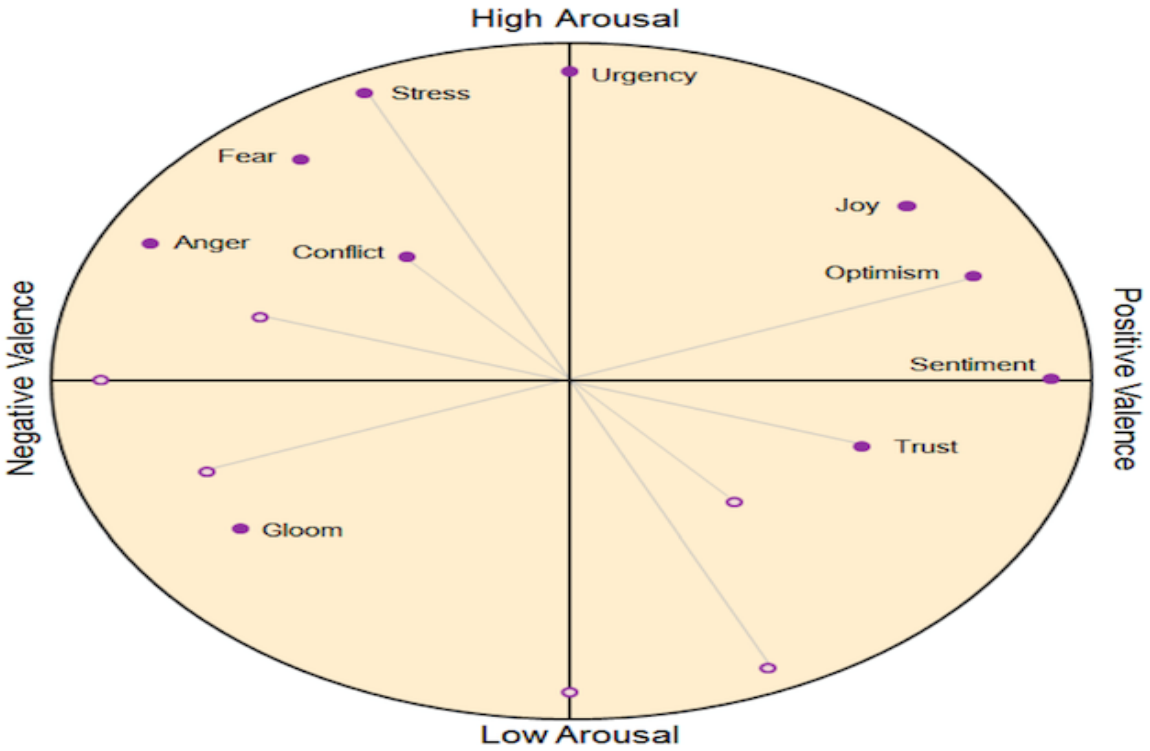
Data-fusion





Thomson Reuters MarketPsych Indices (TRMI)

18,864 separate indices, 119 countries, updated each minute (!)



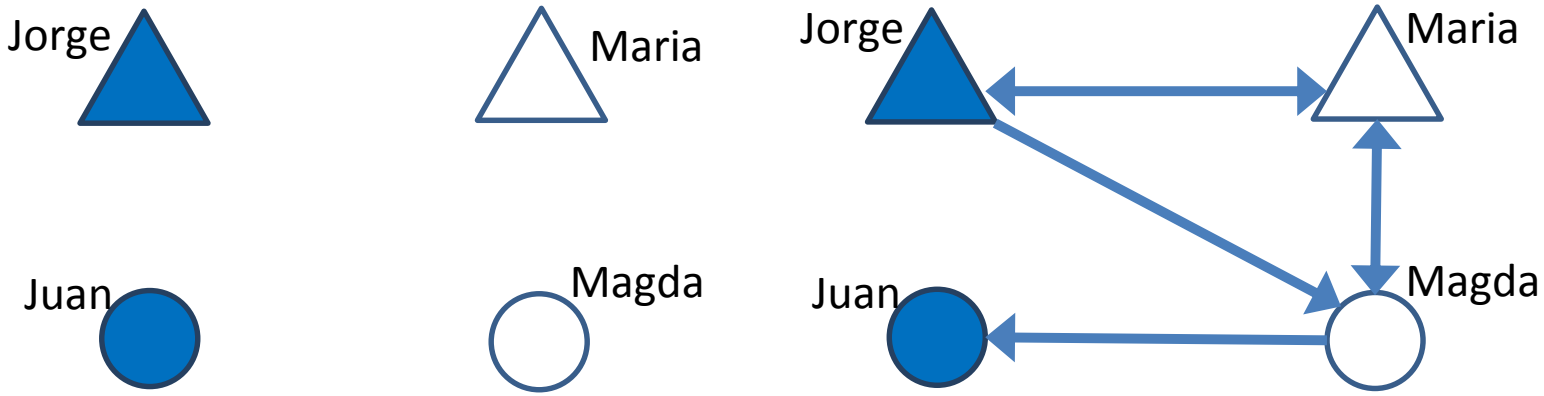
Network Data fusion

Traditional database of attributes

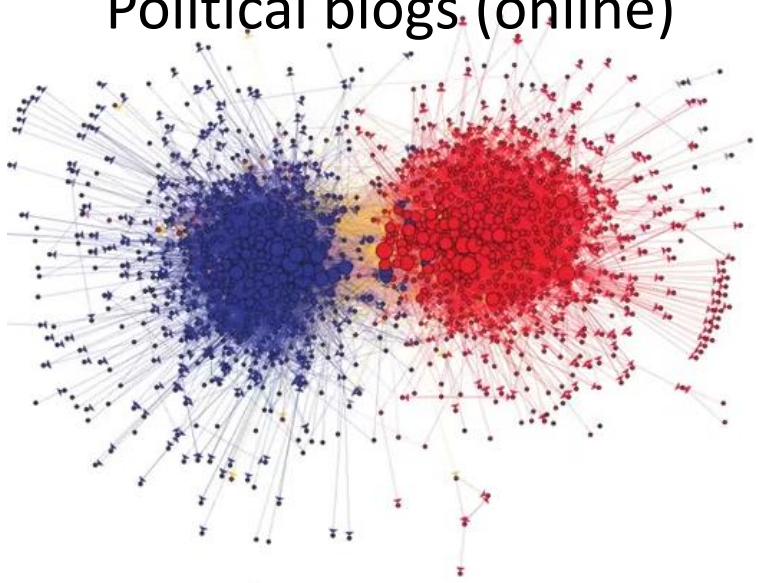
	Gender	Location	Income	Educat.
Jorge	M	Urban	700	Tertiary
Maria	F	Urban	500	Second.
Juan	M	Rural	300	Primary
Magda	F	Rural	200	---

Network database of links

	Jorge	Maria	Juan	Magda
Jorge	Self	↑	---	↑
Maria	↑	Self	---	↑
Juan	---	---	Self	---
Magda	---	↑	↑	Self



Political blogs (online)



Without any information about a Facebook user beyond a list of his friends, one can accurately predict his sexual orientation.

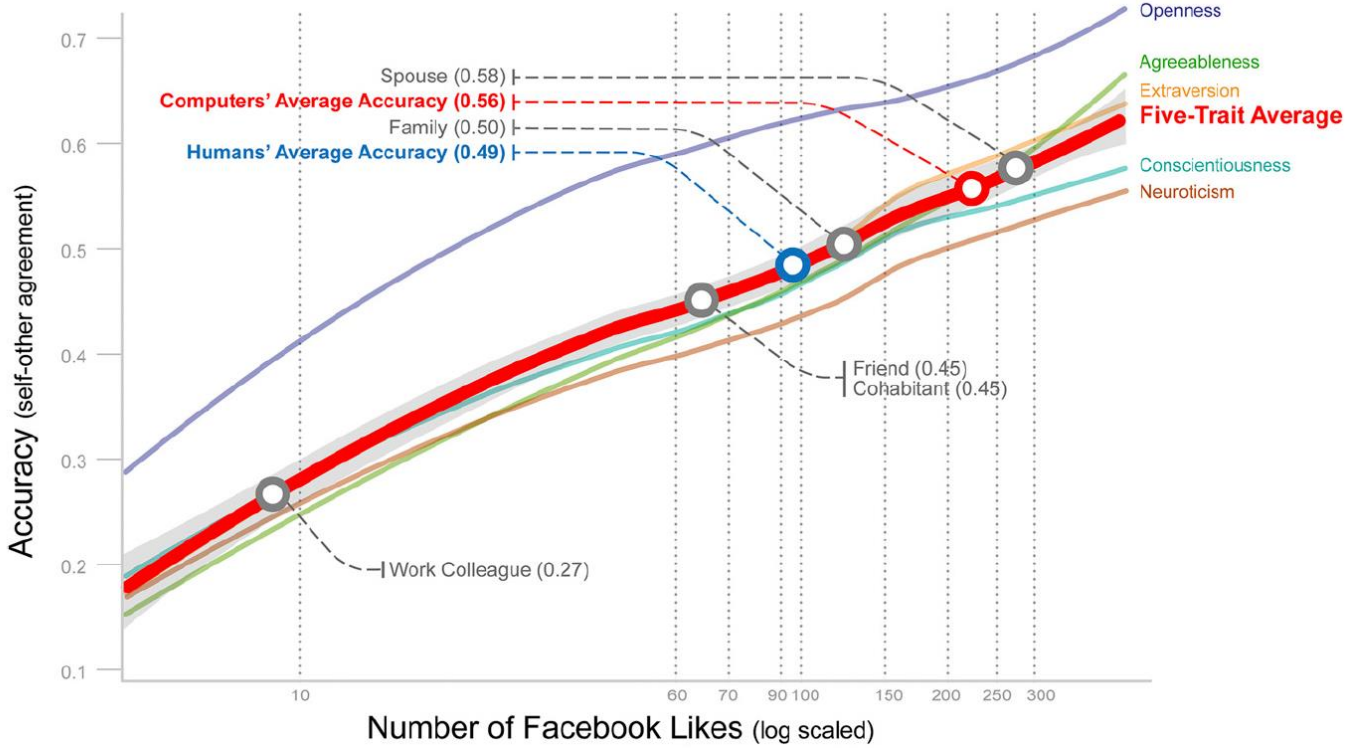
Table 6: Subjects presented in Table 5 with private profiles accurately classified as gay male.

Name	Profile privacy setting	Reported sex	Reported orientation	Percentage gay friends	Classified as gay
A	Private	Unknown	Unknown	13.21%	True
H	Private	Unknown	Unknown	4.56%	True
I	Private	Unknown	Unknown	4.19%	True
K	Private	Unknown	Unknown	3.80%	True
P	Private	Unknown	Unknown	2.86%	True
R	Private	Unknown	Unknown	2.65%	True

Source: Adamic L. (2012), Social Network Analysis, Coursera; Jernigan & Mistree (2009). Gaydar. *First Monday*, *14*(10).

Big Network Data:

Digital Footprint + N=n + Data-fusion + Real-time + ML



"Facebook Likes, can be used to automatically and accurately predict…: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender…"

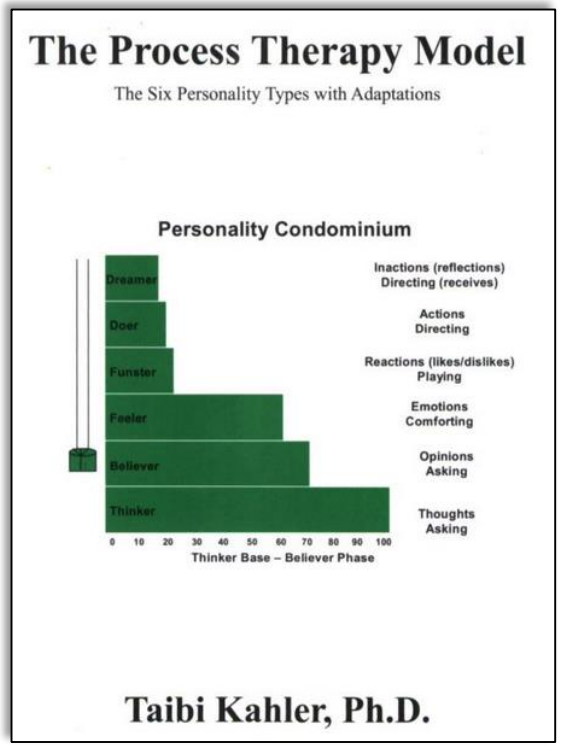
"(i) computer predictions based on … Facebook Likes are more accurate (r = 0.56) than those made by the participants' Facebook friends (r = 0.49);

(iii) computer personality judgments have higher external validity when predicting life outcomes such as substance use, political attitudes, and physical health; for some outcomes, they even outperform the self-rated personality scores…"

Source: Stephens-Davidowitz, S. (2015). Searching for Sex. *The New York Times*. 2015, January 24. Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, 201418680. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, *110*(15), 5802–5805.

"This call might be recorded for quality and training purposes."

Social transparency, *homophily* & polarization



Meaningful business impact

Mattersight, a leader in enterprise analytics focused on customer-employee interactions, offers analysis showing that a behavioral *mismatch* between a customer and employee, as opposed to a favorable behavioral *connection*, has a significant impact on business outcomes across industries, even in highly specialized contact center functions:

Business Outcome	Impact of Behavioral Connection
Sales Conversion Rate	85% to 230%
Customer Retention/Attrition Rate	25% to 50%
Customer Service Cost/ Efficiency	35% to 45%
Student Enrollment Rate (Private Sector Education)	200% to 700%
Debt Cure Rate (Collections Organization/Function)	Over 1400 Basis Points

Matching Personality Types:

- ✓ Call average from 10 min to 5 min
- ✓ Customer Satisfaction from 47 % to 92%

EMOTIONS-DRIVEN (30% of the population)
THOUGHTS-DRIVEN (25%)
REACTIONS-DRIVEN (20%)
OPINIONS-DRIVEN (10%)
REFLECTIONS-DRIVEN (10%)
ACTIONS-DRIVEN (5%)

http://www.eloyalty.com ; http://www.mattersight.com/ ; http://www.fastcompany.com/1706766/how-personality-test-designed-pick-astronauts-taking-pain-out-customer-support ; http://www.ssca.com/resources/articles/104-the-history-of-the-process-communication-model-in-astronaut-selection ; http://www.forbes.com/forbes/2011/0214/entrepreneurs-kelly-conway-software-eloyalty-your-pain.html
Cook, Scott (October 2013). "Personality Matters: Behavioral analytics is now a reality in contact centres". Direct Marketing Magazine 26 (3): 5.

Obama 2012 campaign

YOLO: MEET THE OBAMA CAMPAIGN'S CHIEF TECHNOLOGY OFFICER

The President hugging Harper Reed as shown on his Instagram feed.

➢ Data

- **US$1 billion investment**; core group of **40 engineers**
  (from Twitter, Google, Facebook, Craigslist, stem cell, professional poker players…)
- Project Narwhal: **16 million unique voter profiles**:
  email sign-ups, zip codes, profession, voter registrations, volunteering & donation record, Tweets, Facebook postings and network ties, TV Watching behavior through 20 million set-top boxes, etc.
- **62,000 computer simulations** of likely voter behavior

➢ Outcome

- Identified the 20% of Obama's 2008 vote that shifted into the undecided column, ranking them on a 0-10 persuasion score
- Obama paid 35% less per broadcast commercial than Romney
  (40,000 more spots on the air, spending $90 million less!)
- Present tailor made campaign promises (agreeable adds; etc)
- Guide volunteers in phone and door-to-door campaigns
- Email donation requests, raising $181 $^{\text{million}}/_{\text{month}}$
- Predict States voting outcome at an accuracy of 0.5 percent
- **Change voting behavior of 78 % of targeted undecided voters through Facebook**

Sources: Woodie, A. (2013, June 7). Big Data Analytics Give Electoral Edge. Datanami. Kolb, J., & Kolb, J. (2013). The Big Data Revolution. CreateSpace Independent Publishing Platform. Madrigal, A. C. (2012, November 16). When the Nerds Go Marching In. The Atlantic. Rutenberg (2013), Data You Can Believe In The Obama Campaign's Digital Masterminds Cash In; NYT.

Big Data as commodity

GfK Telefonica Dynamic Insights Smart Steps

Source: http://www.youtube.com/watch?v=wqjKTW3wJZ8

acxiom AOS Audience Operating System

Experian A world of insight

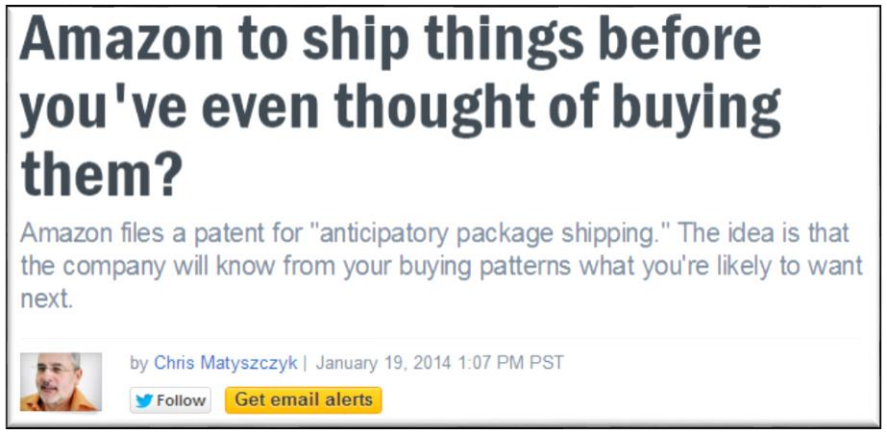
epsilon

Consumers' financial vulnerability:

- *"Social Influencer"*
- *"Rural and Barely Making It"*
- *"Ethnic Second-City Strugglers"*
- *"Retiring on Empty: Singles"*
- *"Tough Start: Young Single Parents"*
- *"Credit Crunched: City Families"*
- *"Transitory lifestyles: military personnel"*
- *"Elderly Opportunity Seekers: elderly looking for ways to make money"*
- *"Oldies but Goodies: gullible, want to believe their luck can change"*

Source: US Senate. *A Review of the Data Broker Industry: Collect, Use, and Sale of Consumer Data for Marketing Purposes*, 2013)

The long-begun end of “free will”?



Predictive Policing LADP & SantaCruz

- Data on crimes, weather, buses, parks…
- Models from Earthquake aftershock
- Predictions to 500^2 feet / 50^2 m
  - ✓ Crimes down 13 %; burglaries 11 %; car theft 8 % (while other districts went up during same period)

Homicide Parole candidates

- dataset > 60,000 crimes
- with some 300 predictors (nature of crime, *age*, repetition)
  - ✓ 60 – 70 % correct who commit homicide

Pre-punishment vs. Free Will?

- already insurance premiums per age or gender (*punishment*)

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: a high stakes application of st learning. *Journal of the Royal Stat.Soc.: Series A*, *172*(1), 191–211. http://spectrum.ieee.org/podcast/at-work/innovation/can-software-predict-repeat-offenders ;http://www.spiegel.de/netzwelt/web/in-santa-cruz-sagen-computer-verbrechen-voraus-a-899422.html ;http://www.sfgate.com/default/article/Sci-fi-policing-prec crime-before-it-occurs-3725708.php ; Wikipedia Commons

Proxies are just proxies

Fruit prices to detect violence in Jalalabad Afghanistan

…a “*big breakthrough*” (DARPA’s Director Regina Dugan) that impressed a group of four-star generals...

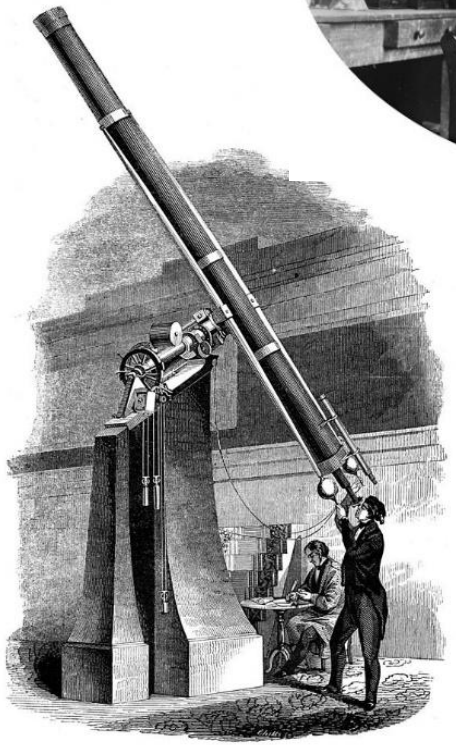
JSOC drone operator: “*It’s of course assumed that the phone belongs to a human being who is nefarious and considered an ‘unlawful enemy combatant.’ This is where it gets very shady…*”

Sources: Shachtman, N. Exclusive: Inside Darpa’s Secret Afghan Spy Machine | Danger Room. Wired (2011). http://www.wired.com/2011/07/darpas-secret-spy-machine/all/
Swisher, K. DARPA’s Regina Dugan Takes It to Mach 20: The Full D9 Interview (Video). AllThingsD, (2011). http://allthingsd.com/20110627/darpas-regina-dugan-takes-it-to-mach-20-the-full-d9-interview-video/
Scahill, J., & Greenwald, G. (2014). The NSA’s Secret Role in the U.S. Assassination Program. *The Intercept*. https://firstlook.org/theintercept/article/2014/02/10/the-nsas-secret-role/

Computational Social Science

I. Big Data Blessings

- Produced anyways
- n = N (Volume)
- Data-fusion (Variety)
- In real-time (Velocity)
- Spooky accuracy through ML



II. Big Data Curses:

- Social Transparency & Polarization
- Data as commodity
- Confusing statistical variables (proxies) with reality