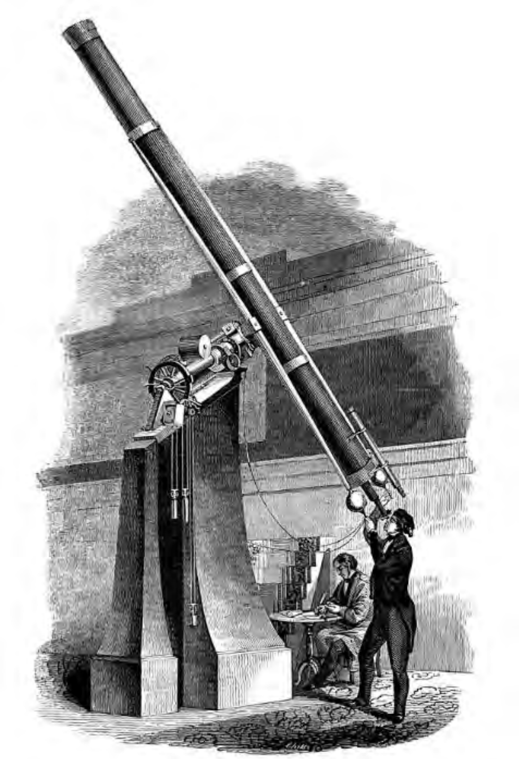


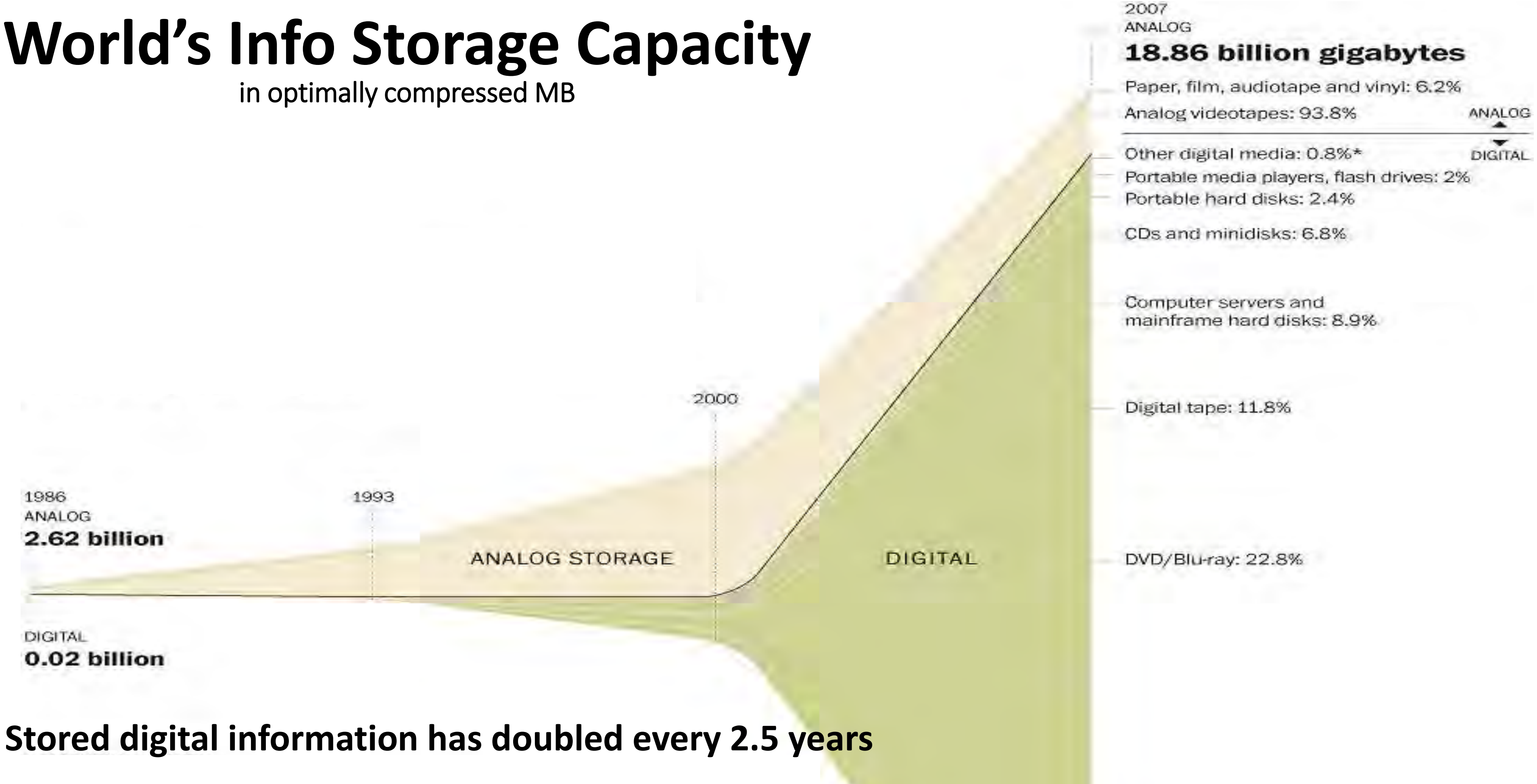
The Theory, Practice and Limits of Big Data for the Social Sciences

UCDAVIS
UNIVERSITY OF CALIFORNIA

Martin Hilbert
Department of Communication
hilbert@UCDavis.edu

World's Info Storage Capacity

in optimally compressed MB



*Other includes chip cards, memory cards, floppy disks, mobile phones/PDAs, cameras/camcorders, video games

Hilbert & López (2011). The world's technological capacity to store, communicate and compute information. *Science*, 332, 6025, 60-65

www.martinhilbert.net/WorldInfoCapacity.html

Stored digital information has doubled every 2.5 years

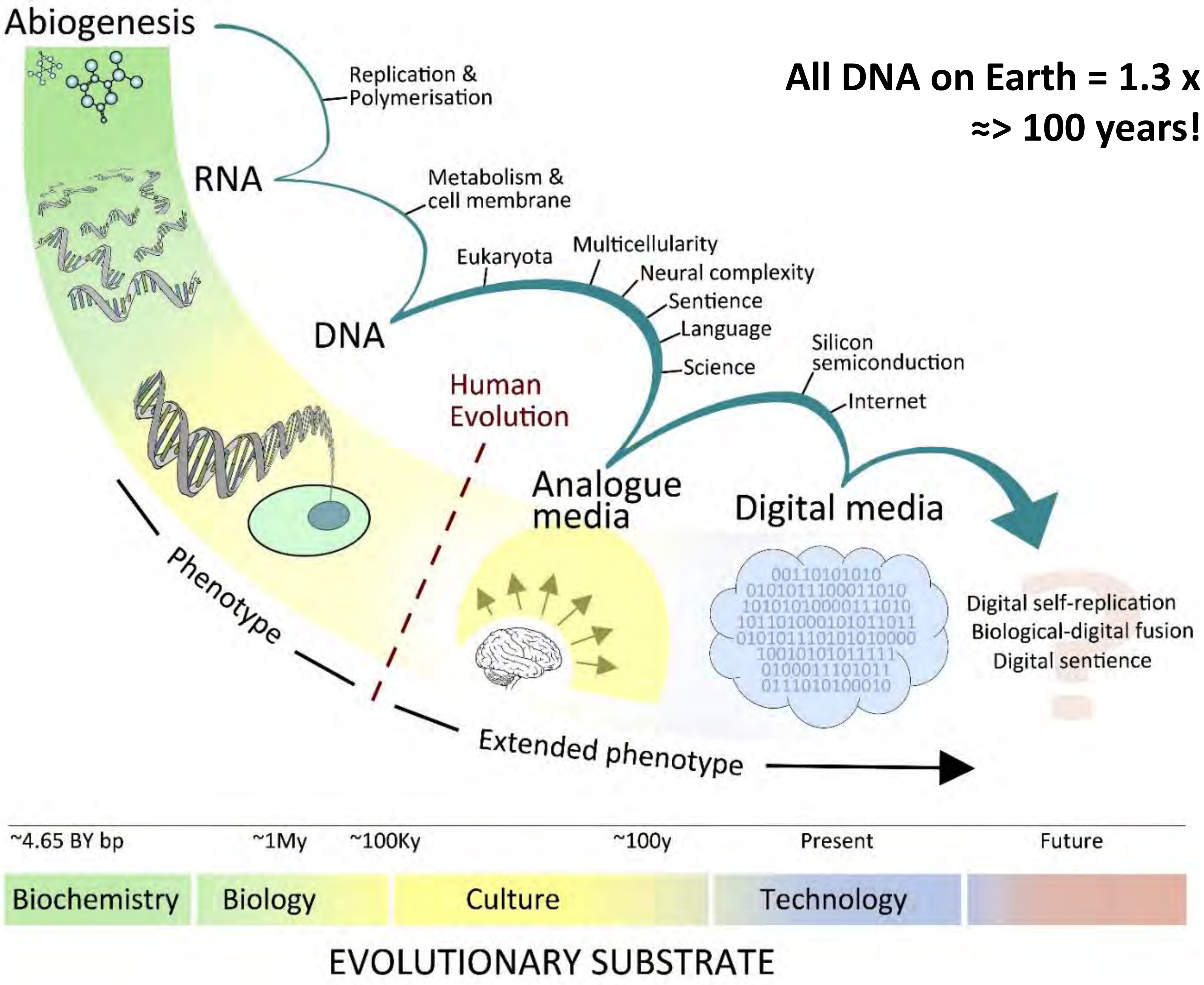
≈ 5 ZB in 2014 (5 x 10^{21} Bytes)

4,500 piles of printed books

Ken Costello

All DNA on Earth = 1.3 x 10^{37} Bytes
≈> 100 years!

Gillings, Hilbert & Kemp (2016)
Information in the Biosphere:
Biological and Digital Worlds.
Trends in Ecology & Evolution,
31(3), 180–189
www.martinhilbert.net/information-in-the-biosphere/



7.2 bn humans * 6.2 bn nucleotides = 1 x 10^{19} Bytes vs. 5 x 10^{21} Bytes

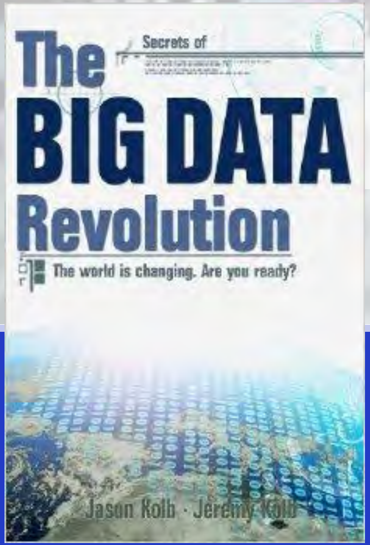
OECD

BETTER POLICIES FOR BETTER LIVES

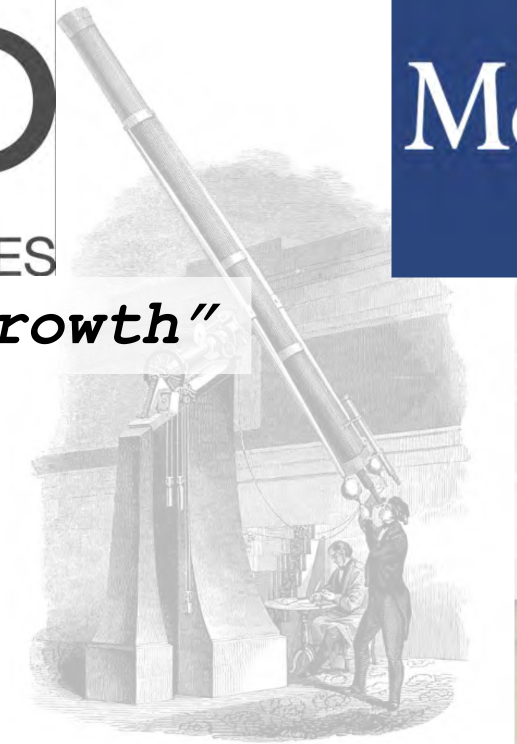
McKinsey&Company

“data as a new source of growth”

“need to recognize the potential of harnessing big data to unleash the next wave of growth”



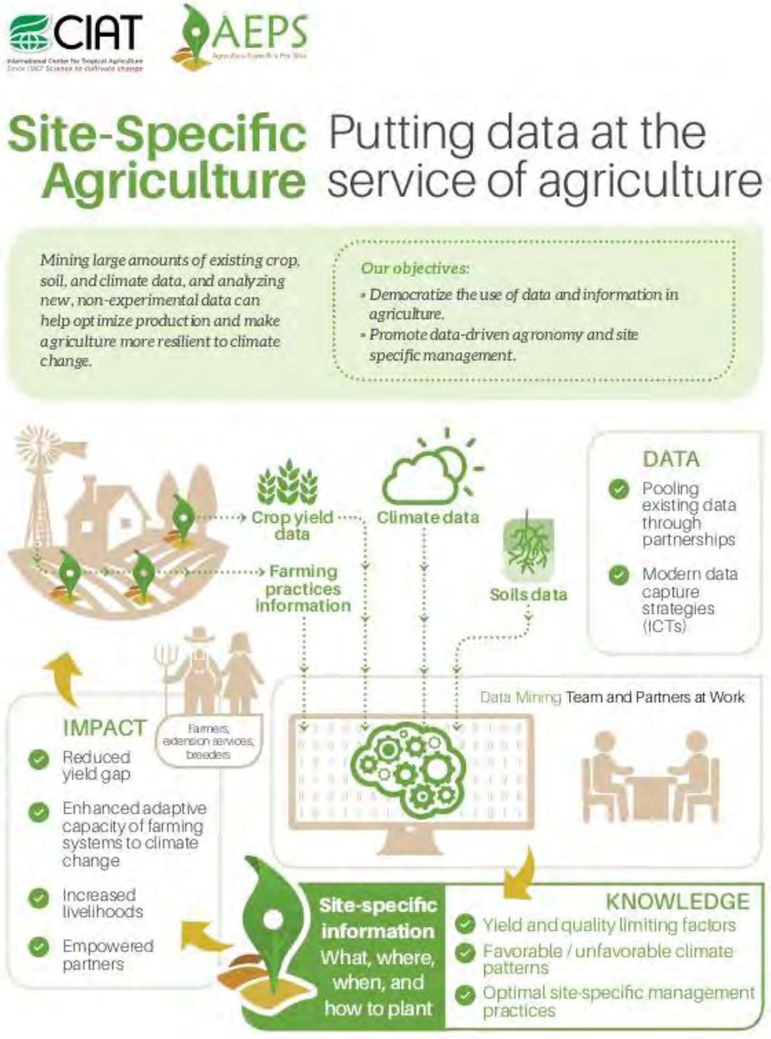
“the new oil”



The Theory, Practice and Limits of Big Data for the Social Sciences

SDG Goal 2.4: …ensure sustainable food production systems and implement resilient agricultural practices that increase productivity

International Center for Tropical Agriculture; Colombian Government Agriculture and Food Security; Colombia's National Federation of Rice Growers



CIAT International Center for Tropical Agriculture

AEPS

Site-Specific Agriculture Putting data at the service of agriculture

Mining large amounts of existing crop, soil, and climate data, and analyzing new, non-experimental data can help optimize production and make agriculture more resilient to climate change.

Our objectives:

- Democratize the use of data and information in agriculture.
- Promote data-driven agronomy and site specific management.

DATA

- Pooling existing data through partnerships
- Modern data capture strategies (ICTs)

Data Mining Team and Partners at Work

IMPACT

- Reduced yield gap
- Enhanced adaptive capacity of farming systems to climate change
- Increased livelihoods
- Empowered partners

Site-specific information What, where, when, and how to plant

KNOWLEDGE

- Yield and quality limiting factors
- Favorable / unfavorable climate patterns
- Optimal site-specific management practices

[weather data] + [RICE crops data] +

+ algorithms from neuroscience / biology =>

=> ***climate change***

Results **localized** for towns:

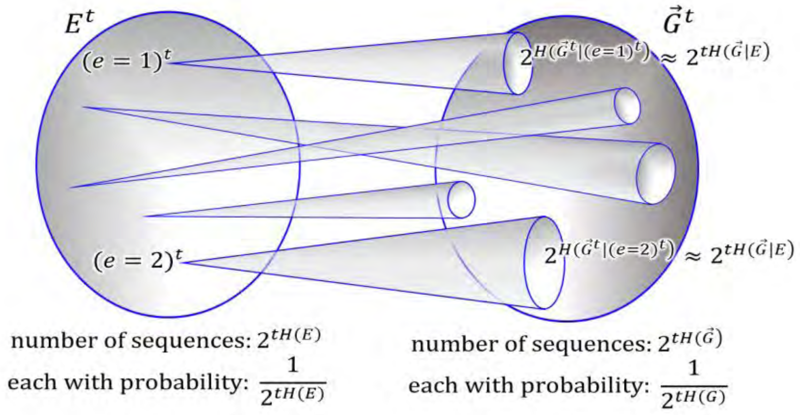
- Saldaña: solar radiation during the grain-ripening stage
- Espinal: sensitivity to warm nights

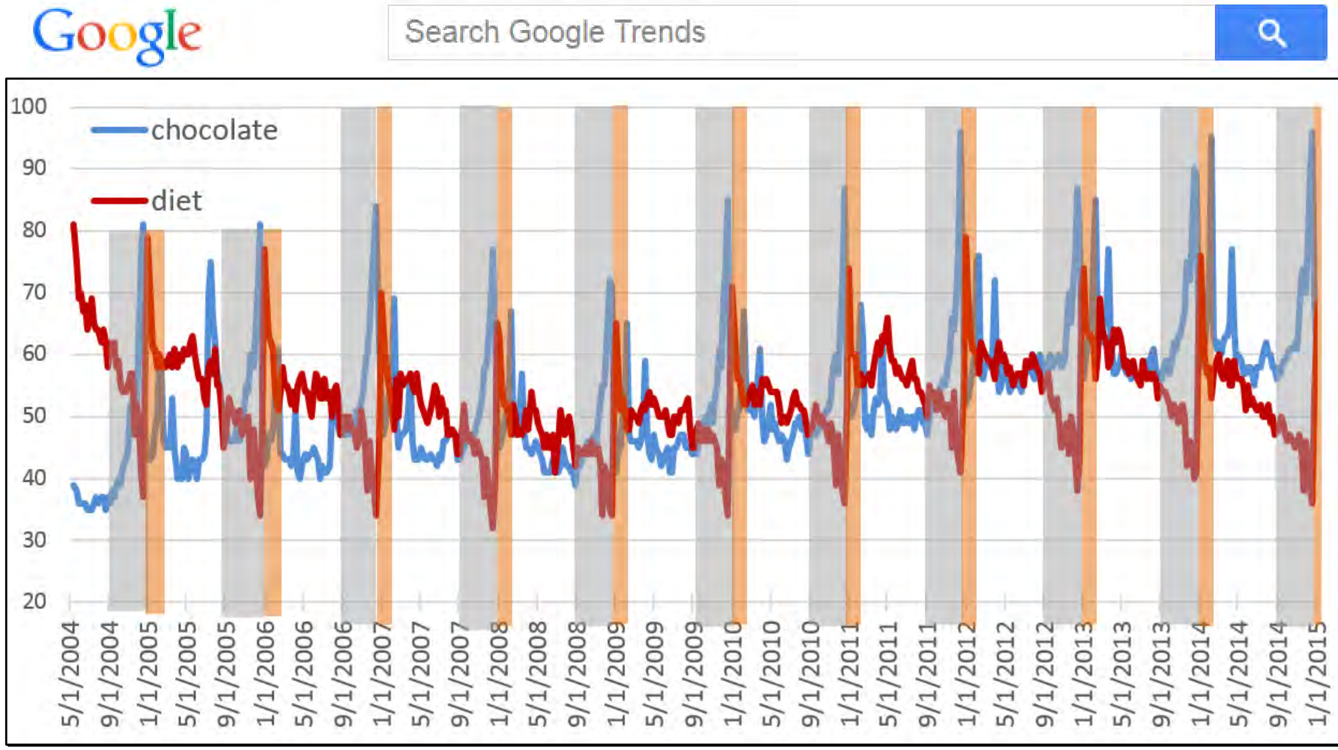
⇒**Low cost** solutions: sowing crops in right period of time

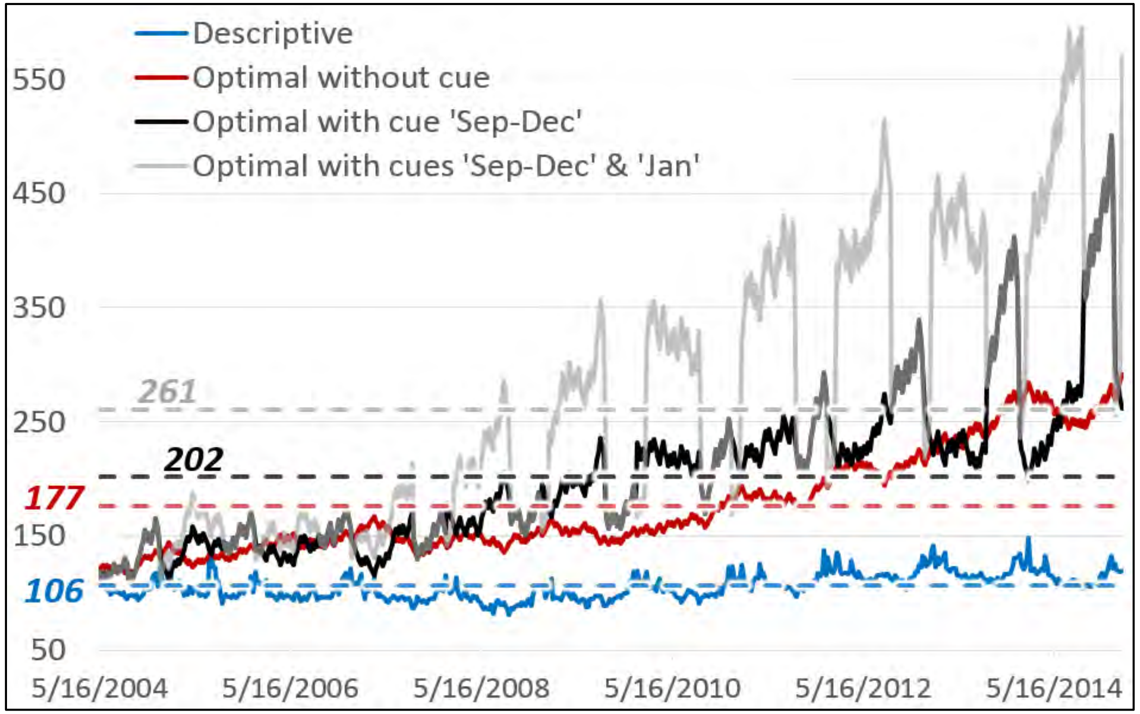
⇒**Impact**: 170 farmers avoided direct losses of $ 3.6 million + productivity from 1 to 3 tons per hectare.

….now being scaled out through Colombia, Argentina, Nicaragua, Peru and Uruguay.

Information & Growth


E^t
$\vec{G}^t$
$(e=1)^t$
$2^{H(\vec{G}^t|(e=1)^t)} \approx 2^{tH(\vec{G}|E)}$
$(e=2)^t$
$2^{H(\vec{G}^t|(e=2)^t)} \approx 2^{tH(\vec{G}|E)}$
number of sequences: $2^{tH(E)}$
each with probability: $\frac{1}{2^{tH(E)}}$
number of sequences: $2^{tH(\vec{G})}$
each with probability: $\frac{1}{2^{tH(G)}}$

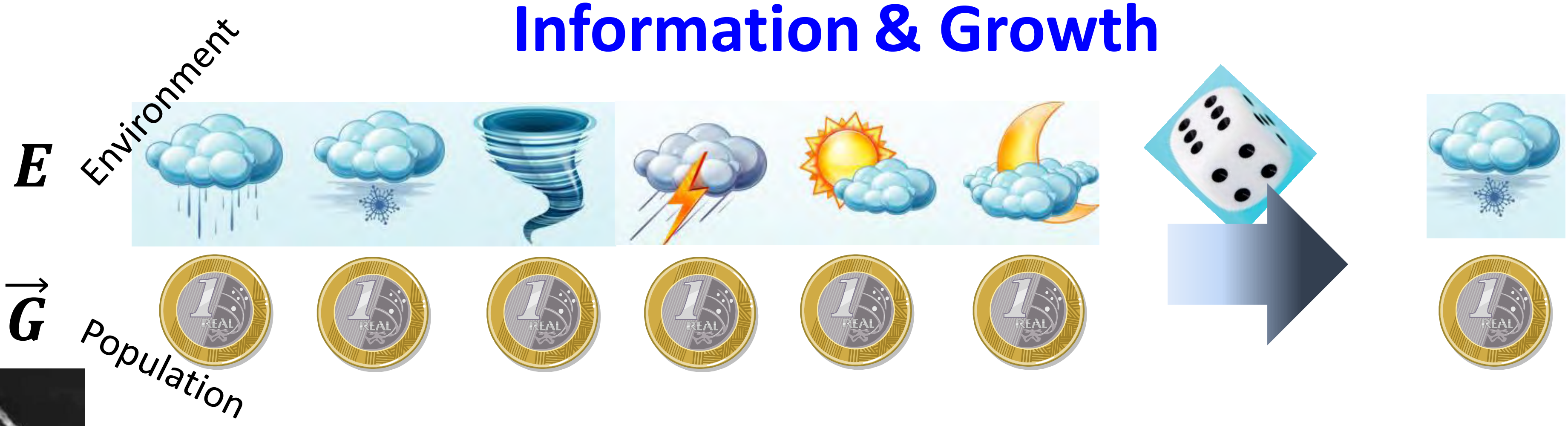


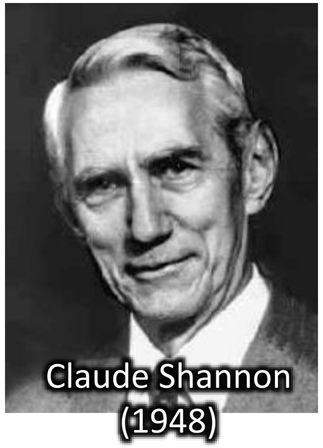


$$Growth = E_e\left[\log {}^{d}W\right] - H\left(E\middle|\vec{G}\right) - D_{KL}\left(\vec{P}(e|g)\middle\|P(e|m)\right) - I\left(E\,;\,\vec{G}\right)$$

Hilbert, M. (2015). An Information Theoretic Decomposition of Fitness: Engineering the Communication Channels of Nature and Society (SSRN Scholarly Paper No. ID 2588146). Social Science Research Network. http://papers.ssrn.com/abstract=2588146

Information & Growth





Claude Shannon (1948)

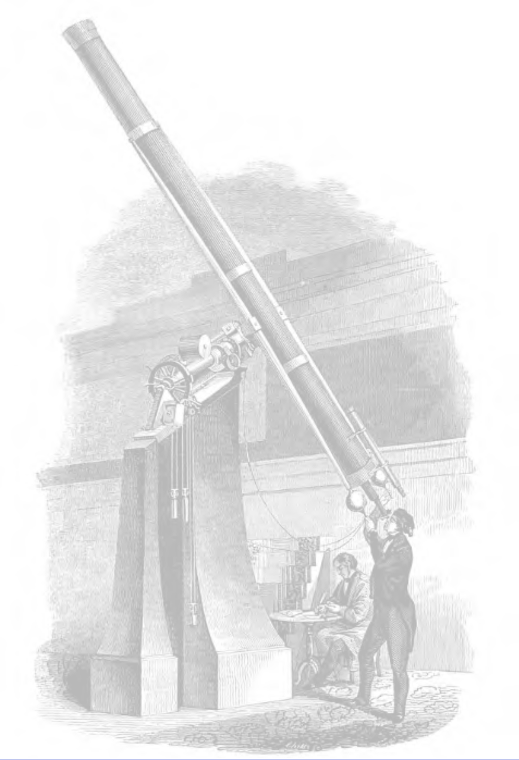
A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656.

1 bit of **information** = reduction of **uncertainty** by half

½ * uncertainty = 1 bit of information = **2** * Growth

$$Growth = E_e\left[\log {}^d W\right] - H\left(E|\vec{G}\right) - D_{KL}\left(\vec{P}(e|g) \| P(e|m)\right) - I\left(E ; \vec{G}\right)$$

Hilbert, M. (2015). An Information Theoretic Decomposition of Fitness: Engineering the Communication Channels of Nature and Society (SSRN Scholarly Paper No. ID 2588146). Social Science Research Network. http://papers.ssrn.com/abstract=2588146



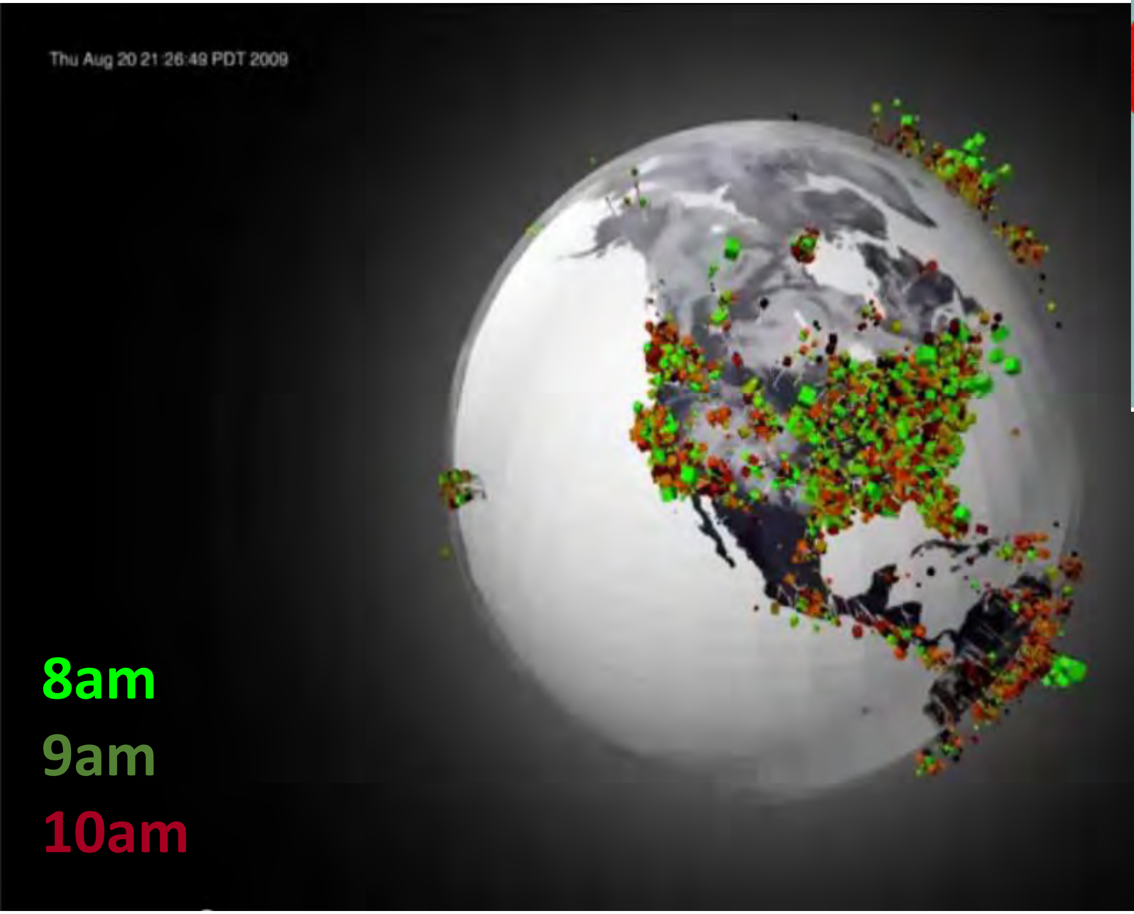
The Theory, Practice and Limits of Big Data for the Social Sciences

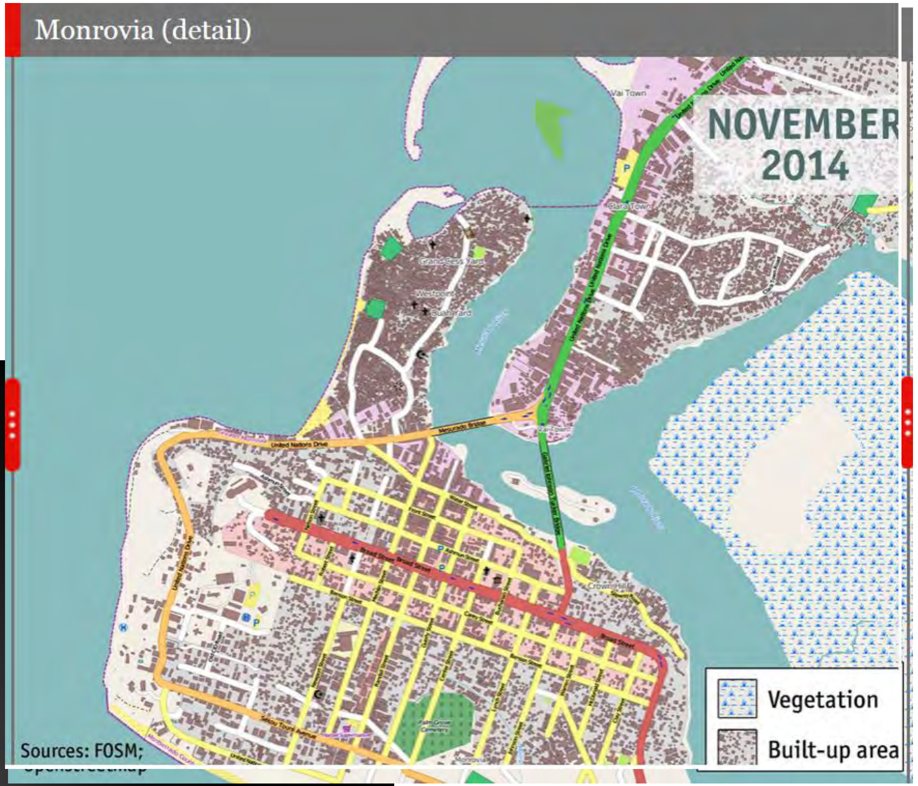
Characteristics of Big Data

- **Digital footprint** (produced anyways for free)
- **n = N** (no sampling, but potential bias)
- **Data-fusion** (unstructured and incomplete)
- **In real-time** (dynamic)
- **Machine Learning** (no need for theory)

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

Digital footprint



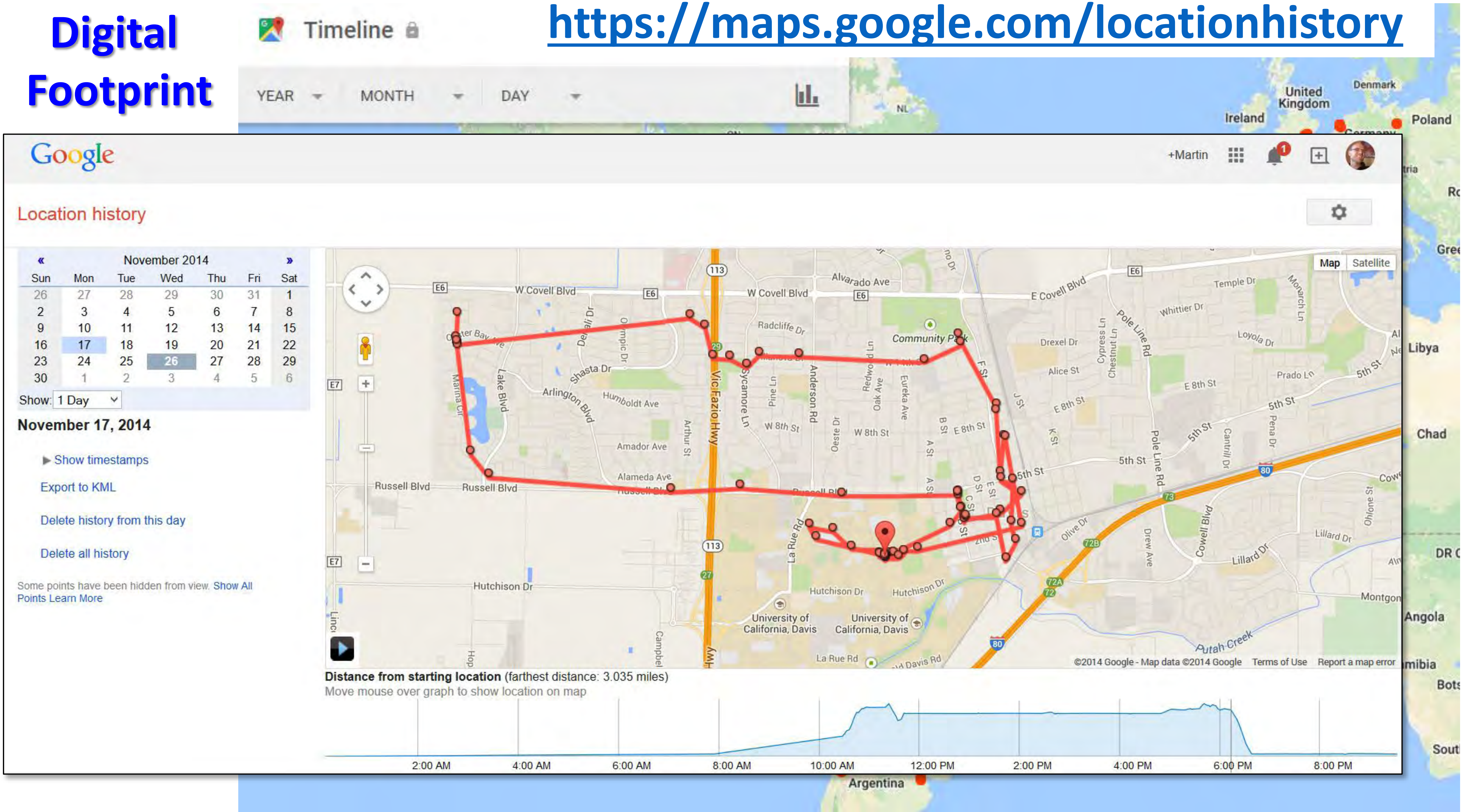


Source:
TED-Ed. (2013). Visualizing the world’s Twitter data - Jer Thorp. http://www.youtube.com

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

Digital Footprint

https://maps.google.com/locationhistory



Timeline

YEAR ▾ MONTH ▾ DAY ▾

Google +Martin

Location history

«			November 2014			»
Sun	Mon	Tue	Wed	Thu	Fri	Sat
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6

Show: 1 Day

November 17, 2014

▶ Show timestamps

Export to KML

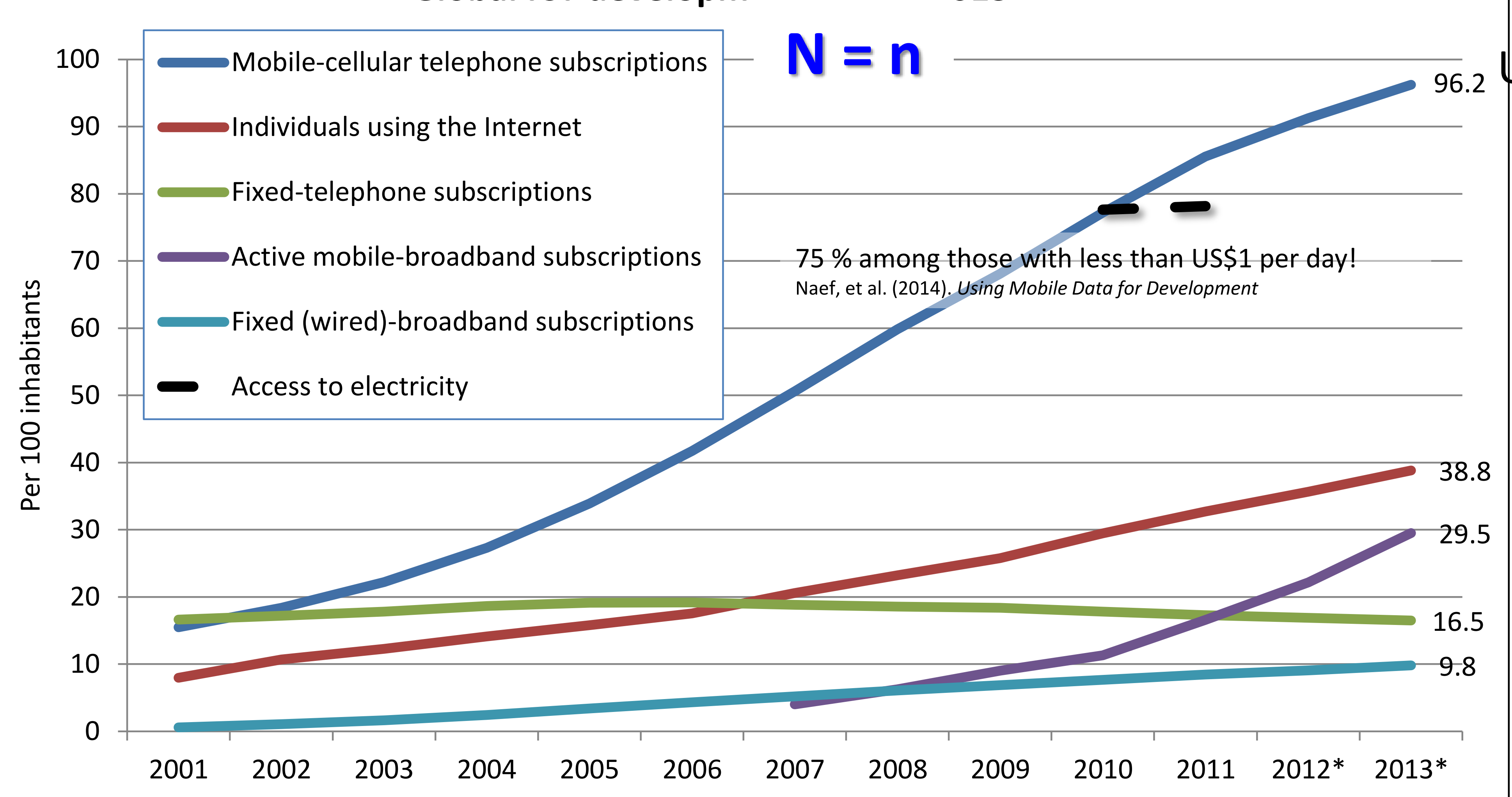
Delete history from this day

Delete all history

Some points have been hidden from view. Show All Points Learn More

Distance from starting location (farthest distance: 3.035 miles)

Move mouse over graph to show location on map



N = n

75 % among those with less than US$1 per day!

Naef, et al. (2014). *Using Mobile Data for Development*

Note: * Estimate

Source: ITU World Telecommunication /ICT Indicators database

Using data records like call duration and call frequency, one can predict socio-economic, demographic, and other behavioral trades with 80-85% accuracy.

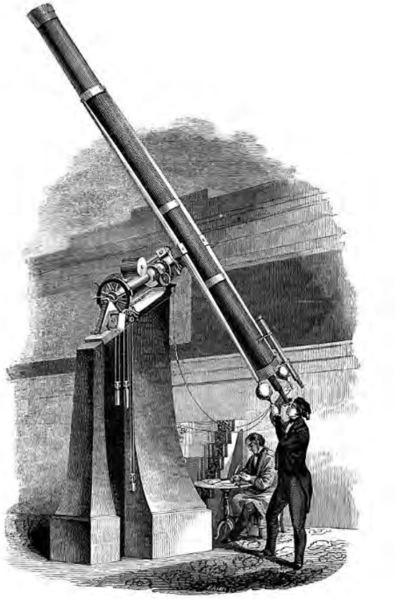
Sources: Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An Emerging Tool for Social Scientists. *Sociol. Methods & Research*, *37*(3), 426–454.
Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engin. Appl. of Artificial Intell.*, 35, 237–245.
Frias-Martinez, V., & Virseda, J. (2013). Cell Phone Analytics: Scaling Human Behavior Studies into the Millions. *ITID*, 9(2), pp. 35–50.
Frias-Martinez, V., Frias-Martinez, E., & Oliver, N. (2010). A Gender-centric Analysis of Calling Behavior…. AAAI 201 *Artificial Intelligence for Development*.
Blumenstock, J. E., Gillick, D., & Eagle, N. (2010). Who's Calling? Demographics of Mobile Phone Use in Rwanda. AAAI 201 *Artificial Intelligence for Development.*

Source: Stephens-Davidowitz, S. (2015). Searching for Sex. *The New York Times*. 2015, January 24. Rudder, C. Dataclysm: Who We Are. (Crown, 2014).

social *science*

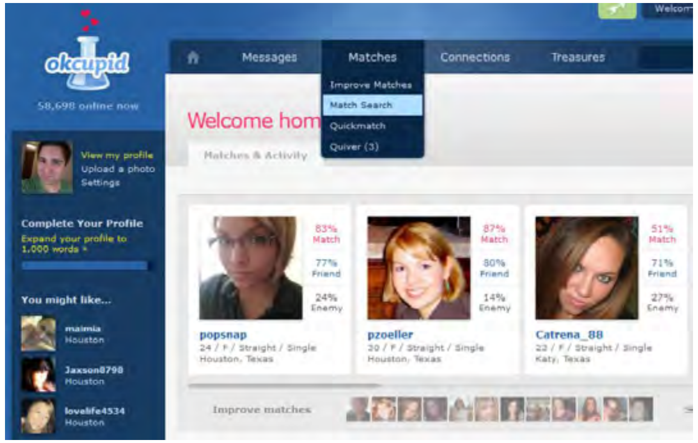
a woman's age vs. the age of the men who look best to her

a woman's age	the age of the men who look best to her
20	23
21	23
22	24
23	25
24	25
25	26
26	27
27	28
28	29
29	29
30	30
31	31
32	31
33	32
34	32
35	34
36	35
37	36
38	37
39	38
40	38
41	38
42	39
43	39
44	39
45	40
46	38
47	39
48	40
49	45
50	46



a man's age vs. the age of the women who look best to him

a man's age	the age of the women who look best to him
20	20
21	20
22	21
23	21
24	21
25	21
26	22
27	21
28	20
29	20
30	20
31	20
32	20
33	20
34	20
35	20
36	20
37	22
38	20
39	20
40	21
41	21
42	20
43	
44	21
45	
46	20
47	20
48	
49	20
50	22



Data Fusion

GfK Telefónica

Dynamic Insights

Smart Steps

acxiom AOS Audience Operating System

Experian A world of insight

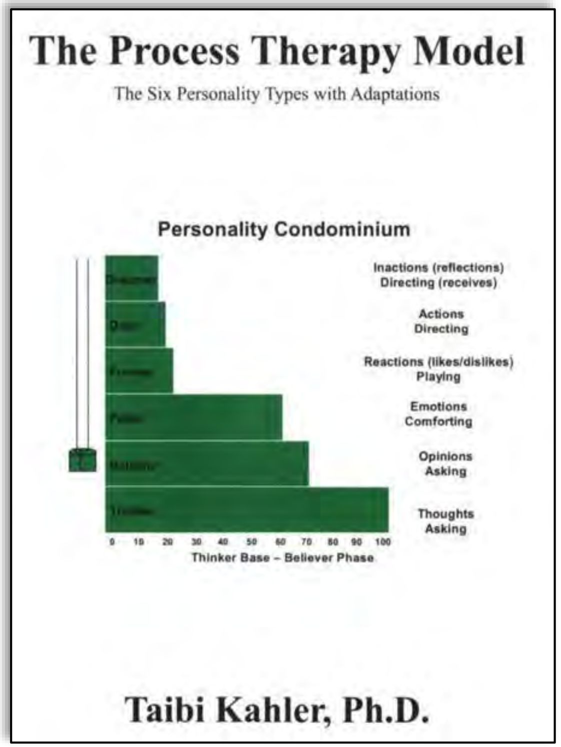
epsilon

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: http://www.youtube.com/watch?v=wqjKTW3wJZ8 US Senate. *A Review of the Data Broker Industry: Collect, Use, and Sale of Consumer Data for Marketing Purposes*, 2013)

Real time



The Process Therapy Model

The Six Personality Types with Adaptations

Taibi Kahler, Ph.D.

Matching Personality Types:

✓ Call average from 10 min to 5 min

✓ Customer Satisfaction from 47 % to 92%

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

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

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.
- Ranking the 20% of Obama’s 2008 vote that shifted to undecided on a 0-10 persuasion score
- **62,000 computer simulations** of likely voter behavior

YOLO: MEET THE OBAMA CAMPAIGN'S CHIEF TECHNOLOGY OFFICER

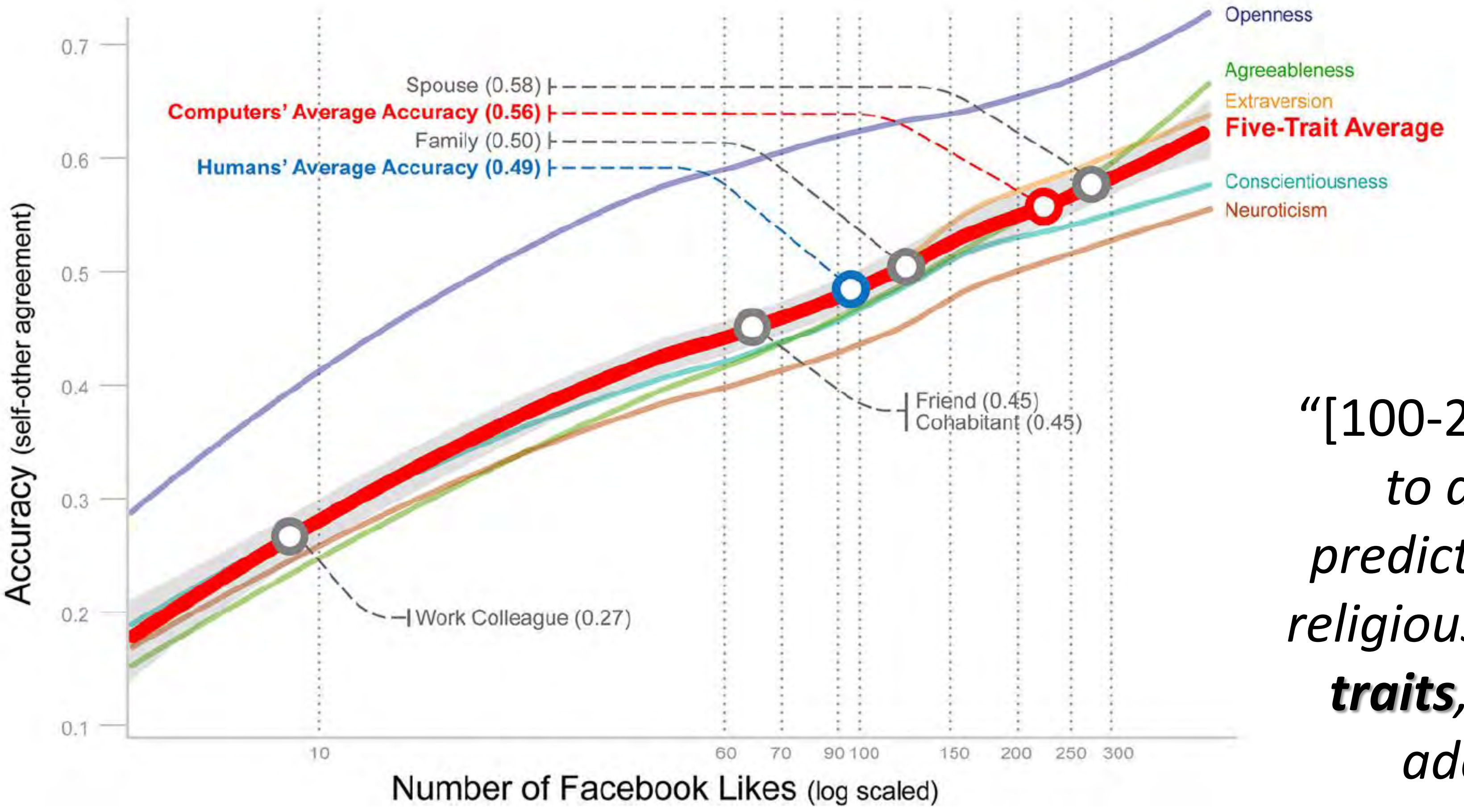
The President hugging Harper Reed as shown on his Instagram feed.

Outcome

- Obama paid 35% less per broadcast commercial than opponent 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.

Machine learning knows us better than we ourselves



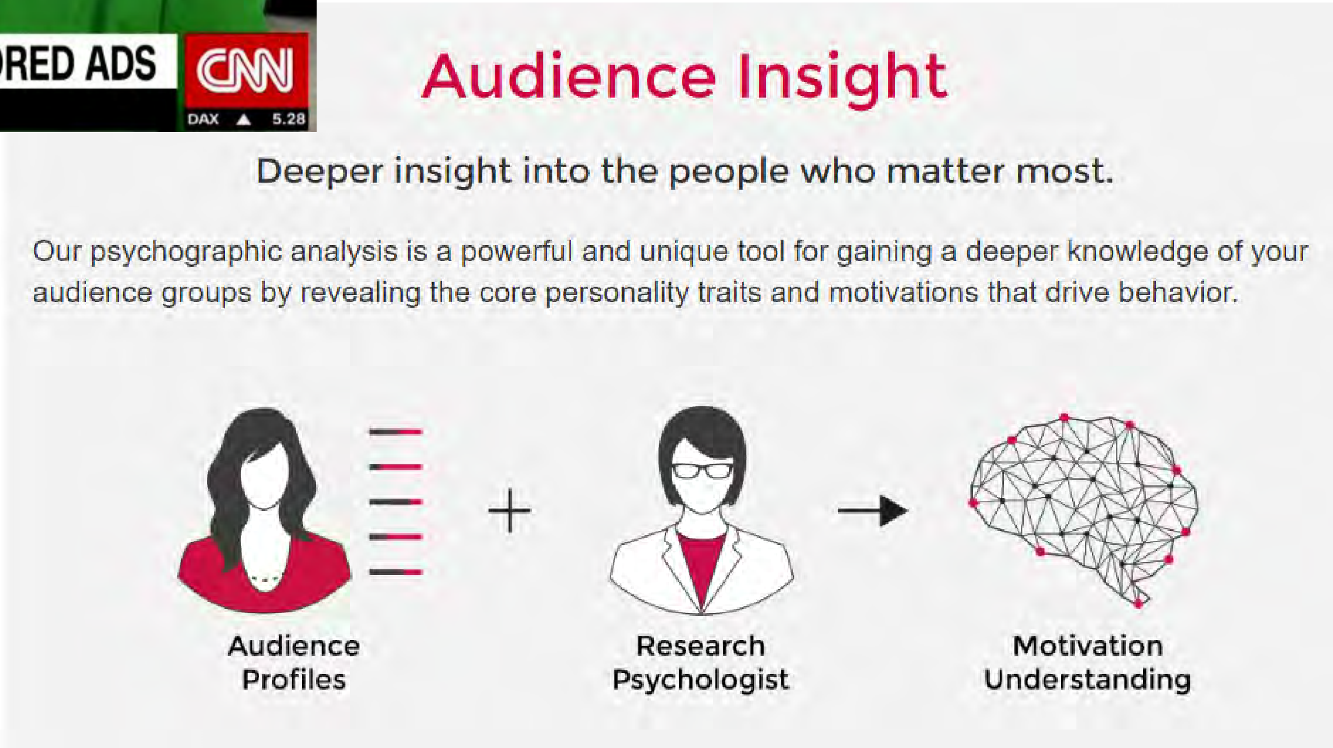
"[100-250] *Facebook Likes, can be used to automatically and accurately predict…: sexual orientation, ethnicity, religious and political views,* ***personality traits****, intelligence, happiness, use of addictive substances, parental separation, age, and gender…"*

Source: 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.

July 2016:	$100,000
August:	$250,000
September:	$ 5 million

Cambridge Analytica

32 personality types in 17 states



Audience Insight

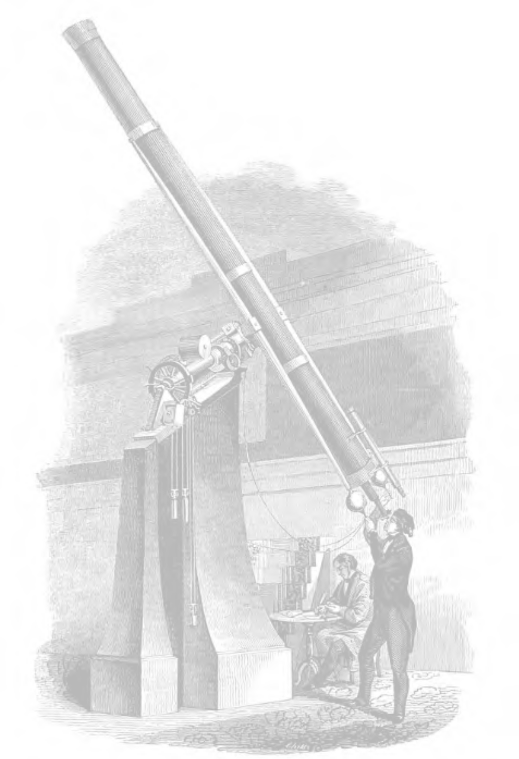
Deeper insight into the people who matter most.

Our psychographic analysis is a powerful and unique tool for gaining a deeper knowledge of your audience groups by revealing the core personality traits and motivations that drive behavior.

Ownership of American-built car (phone app)
Haitians: Clinton Foundation Haiti Earthquake
Afro-Americans: Clinton’s superpredators soundbite

Psych: 2nd amendment: fear or tradition?

3rd Debate: 175,000 variations of Trump’s arguments
Differences in title, subtitle, color, picture, video, etc



The Theory, Practice and Limits of

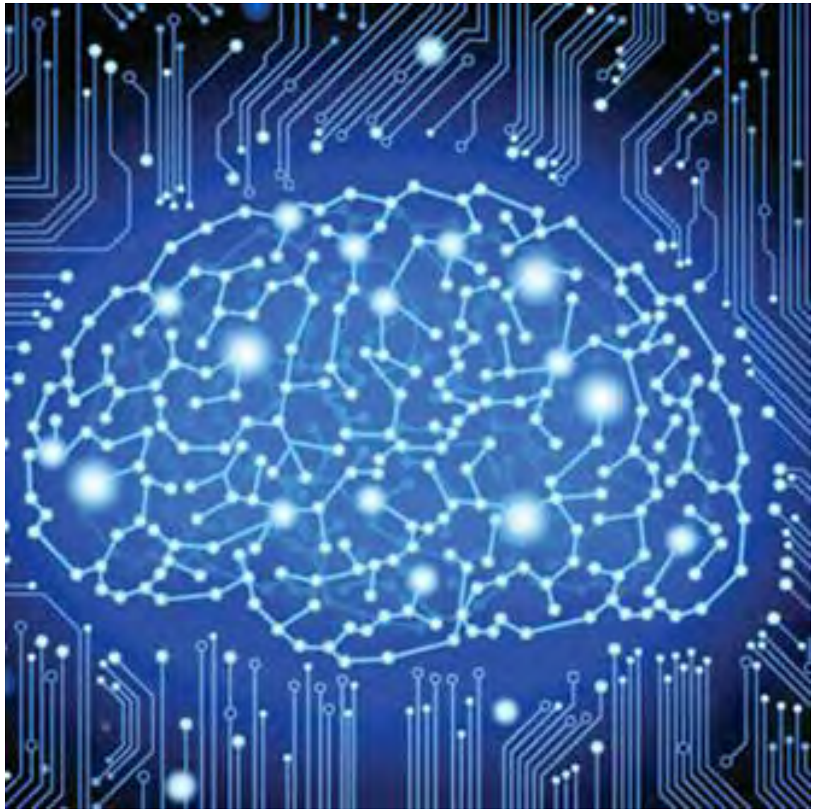
Big Data for the Social Sciences

⚠ Data ≠ Reality

⚠ Meaning ≠ Meaningful

⚠ Correlation ≠ Causation

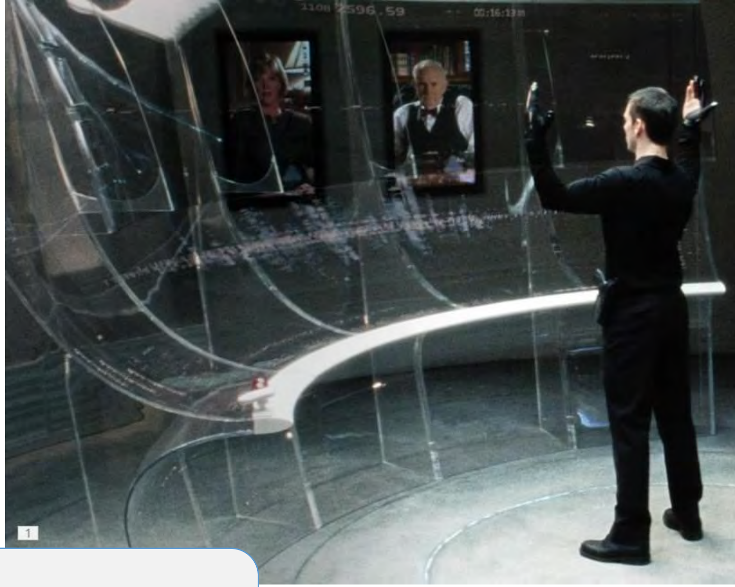
⚠ Past ≠ Future



Data ≠ Reality

Homicide Parole candidates

- 60 – 70 % correct who commits homicide



"We kill people based on metadata"

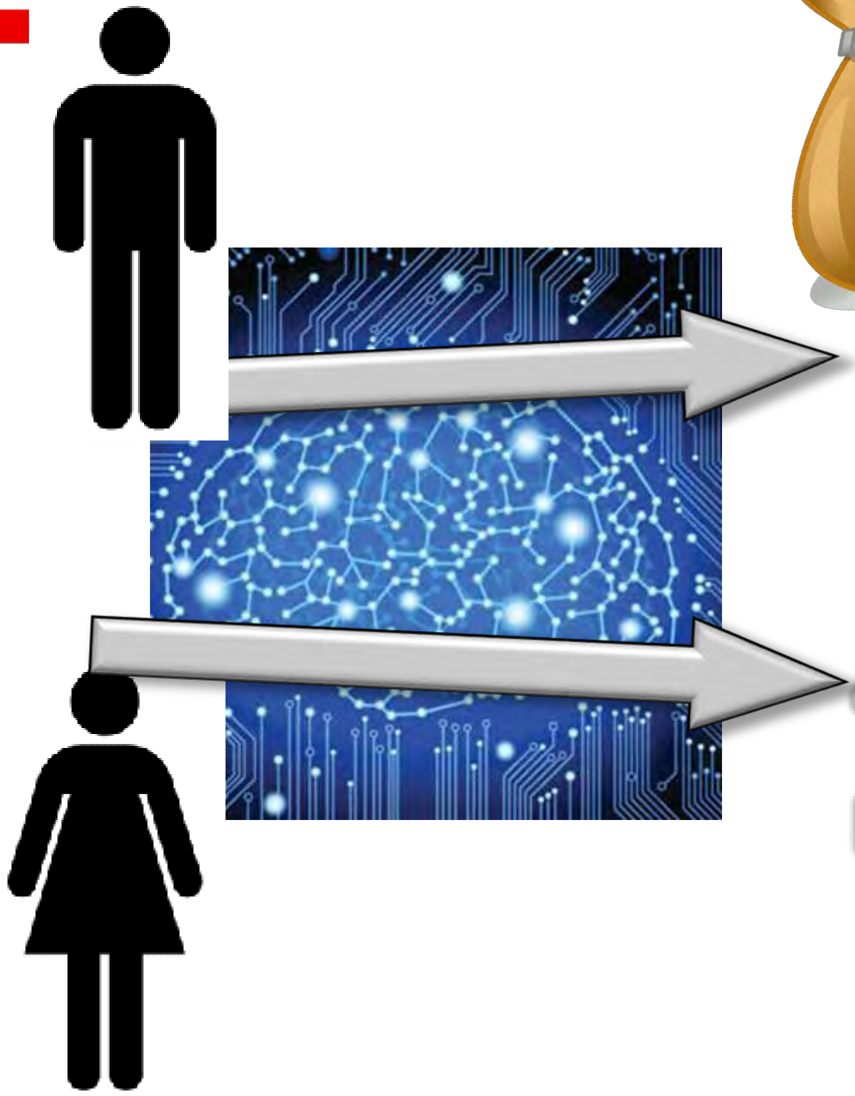
Michael Hayden
former Director
NSA & CIA

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…”*

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 statistical 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-predicting-crime-before-it-occurs-3725708.php ; Wikipedia Commons; Scahill, J., & Greenwald, G. (2014). The NSA’s Secret Role in the U.S. Assassination Program. *The Intercept*.

Meaning ≠ Meaningful

“John,
Paul,
Mike,
Kevin,
Bill”



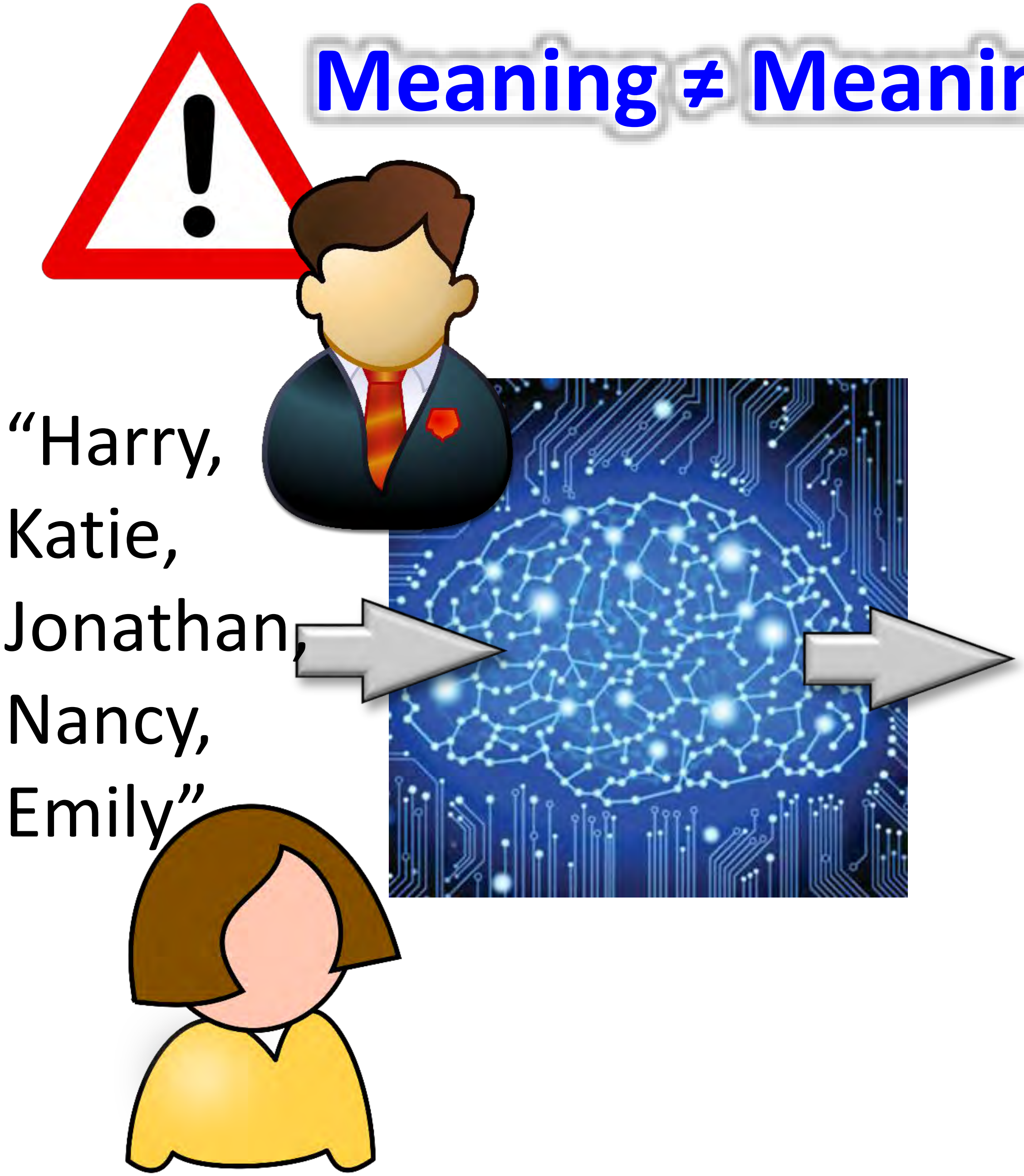
“executive, management, professional, corporation, salary, office, business, career.”

“Amy,
Lisa,
Sarah,
Diana,
Ann”

“home, parents, children, family, marriage, wedding, relatives”

Caliskan-Islam, Bryson, Narayanan (2016). Semantics derived automatically from language corpora necessarily contain human biases. *arXiv:1608.07187 [Cs]*

Meaning ≠ Meaningful

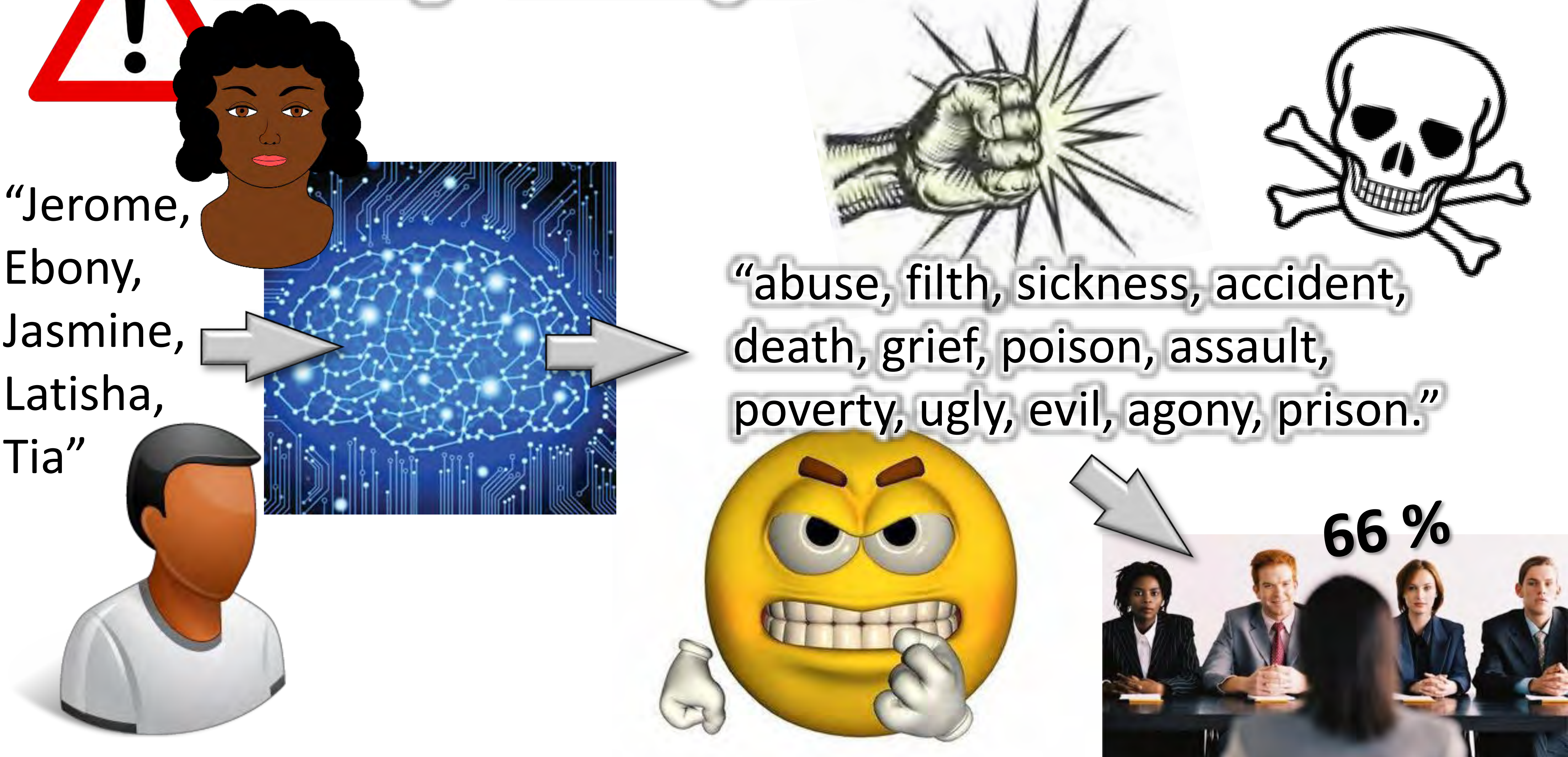


"Harry,
Katie,
Jonathan,
Nancy,
Emily"

"freedom, health, love, peace, friend, heaven, gentle, loyal, lucky, diploma, happy, laughter, vacation"

Caliskan-Islam, Bryson, Narayanan (2016). Semantics derived automatically from language corpora necessarily contain human biases. *arXiv:1608.07187 [Cs]*

Meaning ≠ Meaningful



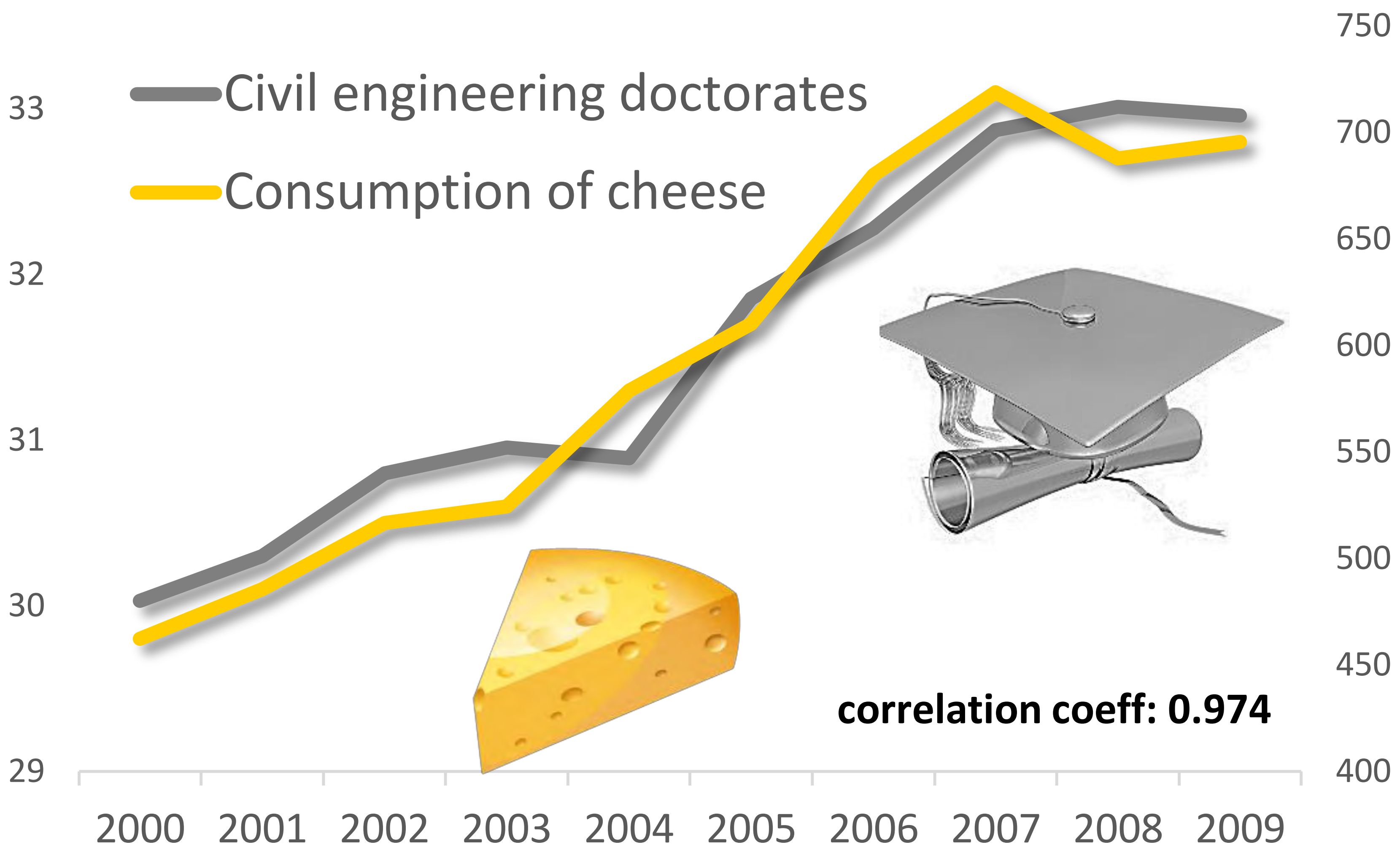
“Jerome, Ebony, Jasmine, Latisha, Tia”

“abuse, filth, sickness, accident, death, grief, poison, assault, poverty, ugly, evil, agony, prison.”

66 %

Caliskan-Islam, Bryson, Narayanan (2016). Semantics derived automatically from language corpora necessarily contain human biases. *arXiv:1608.07187 [Cs]*

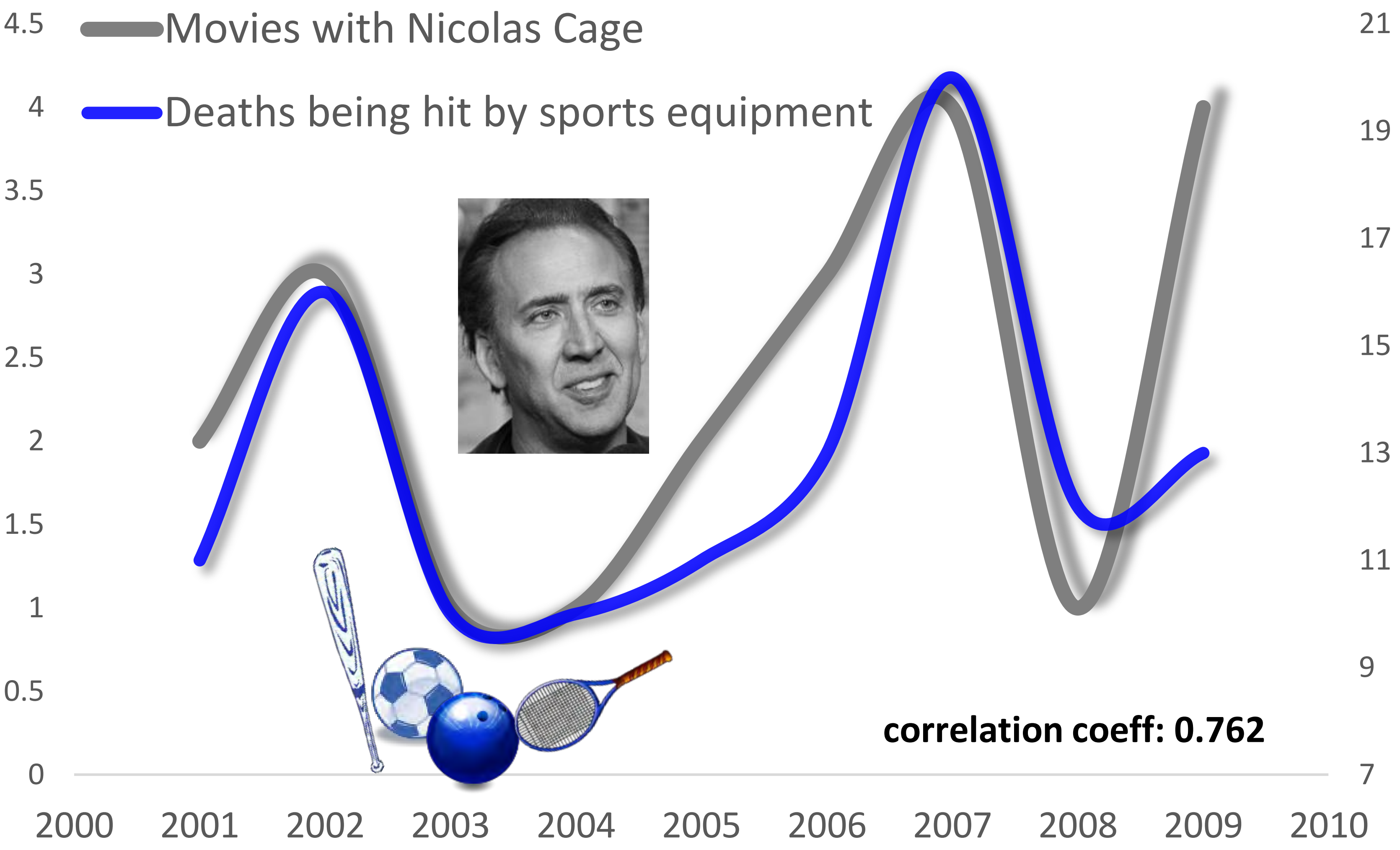
Correlation ≠ Causation



correlation coeff: 0.974

http://tylervigen.com/discover

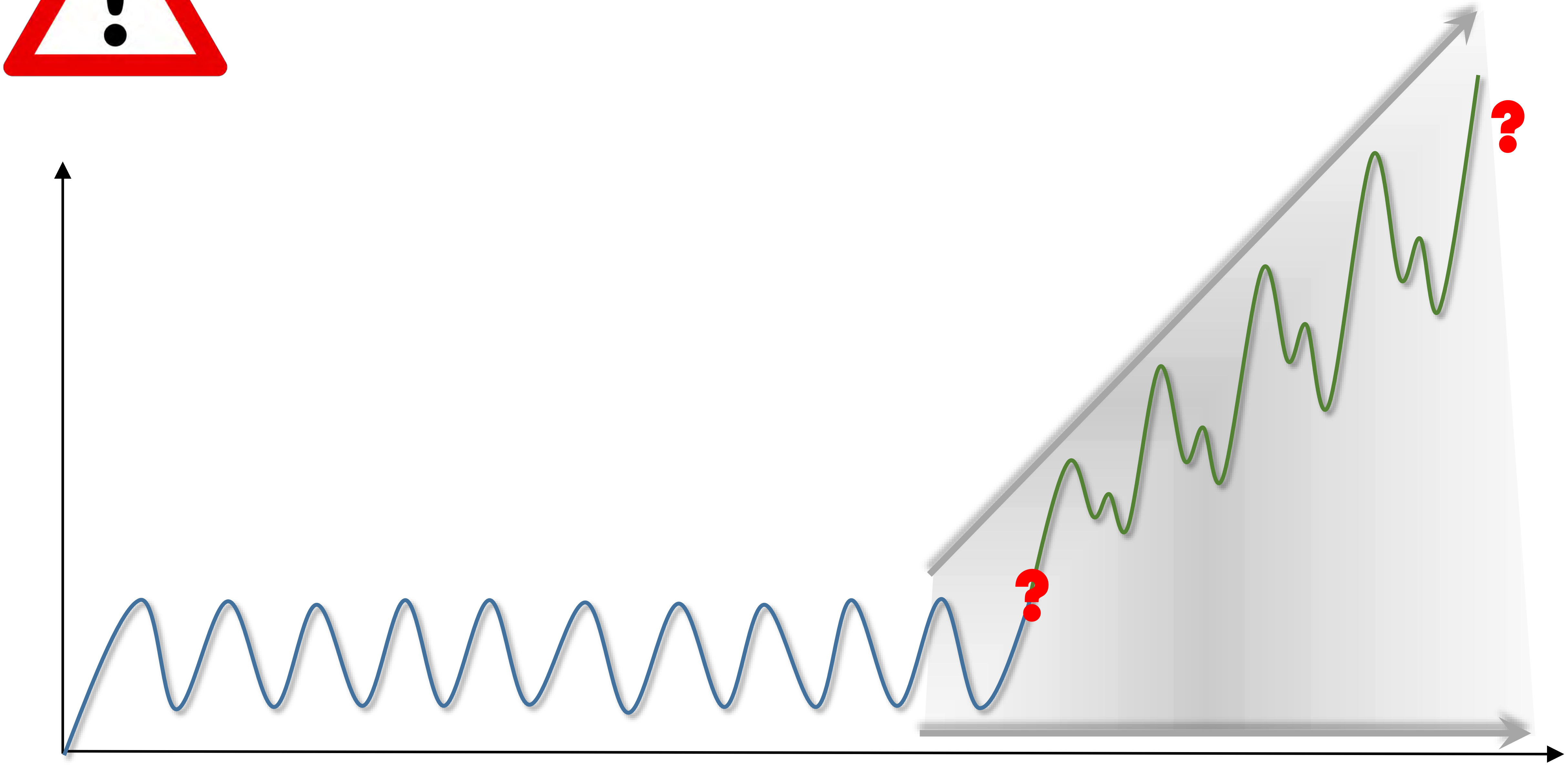
Correlation ≠ Causation



correlation coeff: 0.762

http://tylervigen.com/discover , Wikipedia Commons

Past ≠ Future



Past ≠ Future

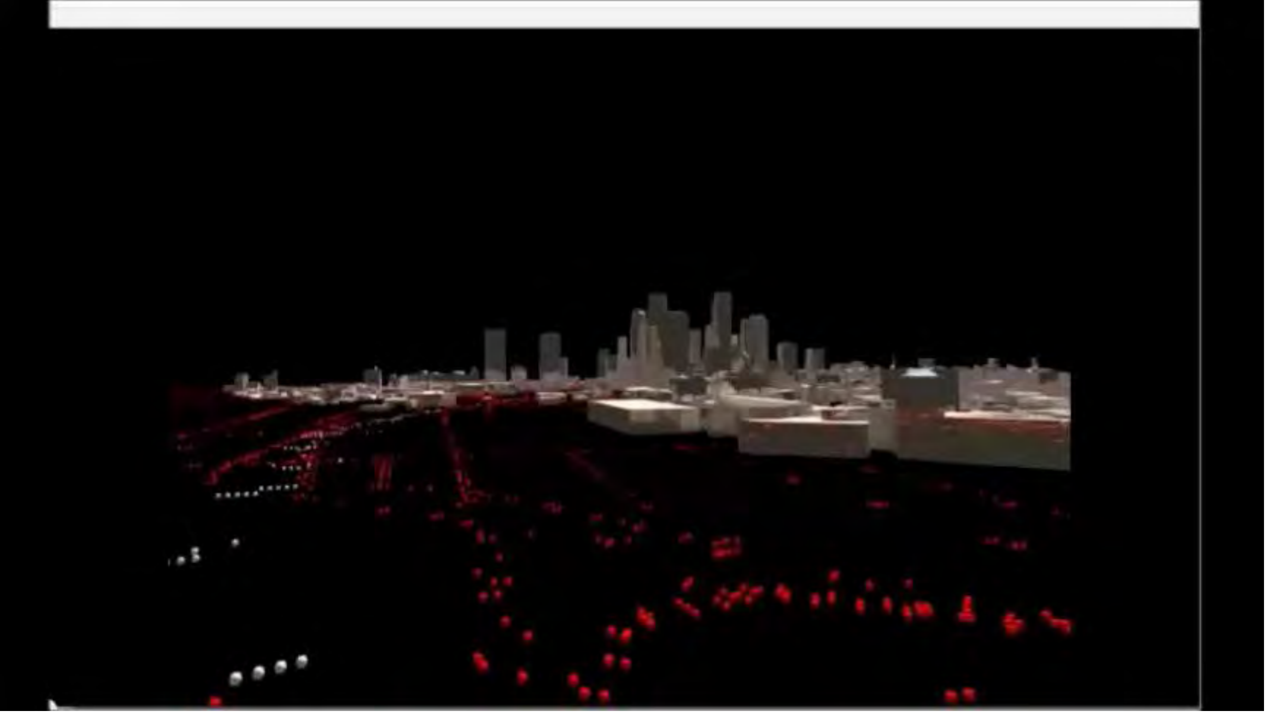
1989

Pollution very hi
Heavy Traffic re ed.
Crime very high.
Jan 2016: Score 460, capital population 55420.
Inadequate rail system.

Wikipedia Commons

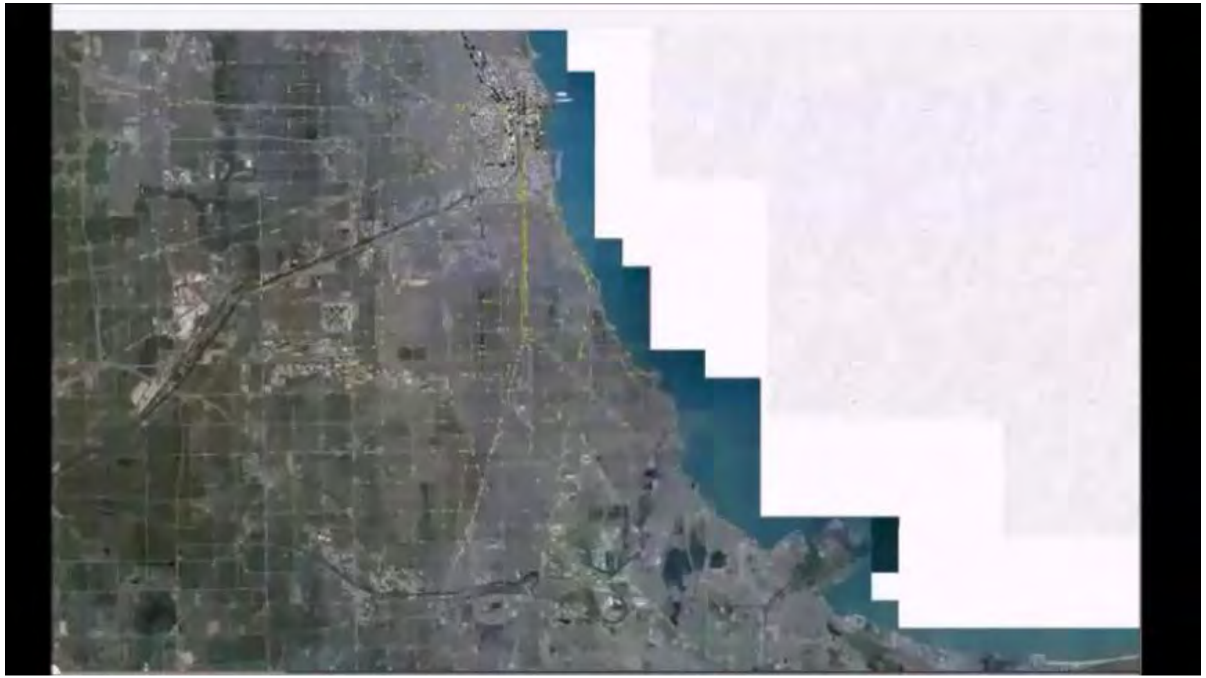
2013

SIMCITY edu



SimCentric Crowd Ambience
CROWD AMBIENCE
Press Esc to exit full screen mode.
SimCentric Technologies

Sources: Bohemia Interactive Simulations, http://youtu.be/G9P9bUTCdpA ; TRANSIMS: http://www.youtube.com/watch?v=mN7kq0ITAys ; Epstein, http://www.youtube.com/watch?v=wZZJCIGtVkw



Computational Social Science

ECONOMETRIC POLICY EVALUATION: A CRITIQUE

Robert E. Lucas, Jr.

"…any change in policy will systematically alter the structure of econometric models"
(1976)

Main References:

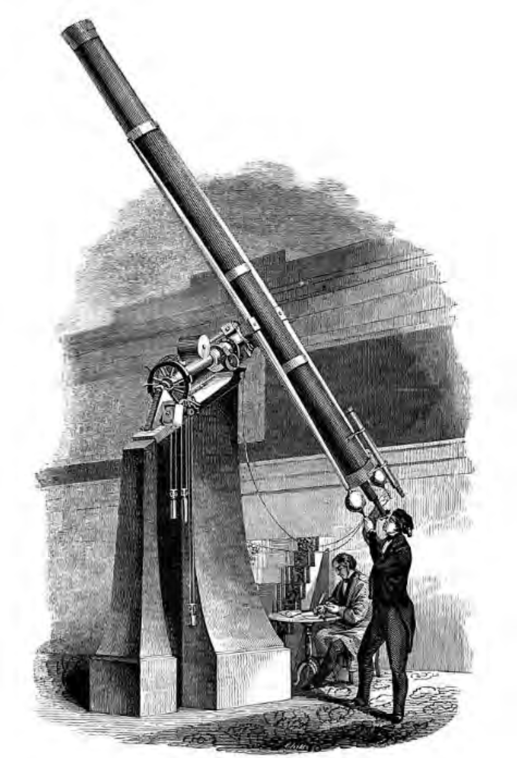
> Hilbert (2016). Big Data for Development: A Review of Promises and Challenges. ***Development Policy Review***, 34(1), 135–174. https://doi.org/10.1111/dpr.12142

> Hilbert, M. (2015). ICT4ICTD: Computational Social Science for Digital Development. 48th (HICSS) (pp. 2145–2157). ***IEEE Computer Society***. https://doi.org/10.1109/HICSS.2015.258

> Gillings, Hilbert & Kemp (2016). Information in the Biosphere: Biological and Digital Worlds. ***Trends in Ecology & Evolution****, 31(3), 180–189* www.martinhilbert.net/information-in-the-biosphere/

> Hilbert & López (2011). The world's technological capacity to store, communicate and compute information. ***Science***, 332, 6025, 60-65 www.martinhilbert.net/WorldInfoCapacity.html

> Hilbert (2016). The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. ***Telecommunications Policy***. www.martinhilbert.net/the-bad-news-is-that-the-digital-access-divide-is-here-to-stay



The Theory, Practice and Limits of Big Data for the Social Sciences

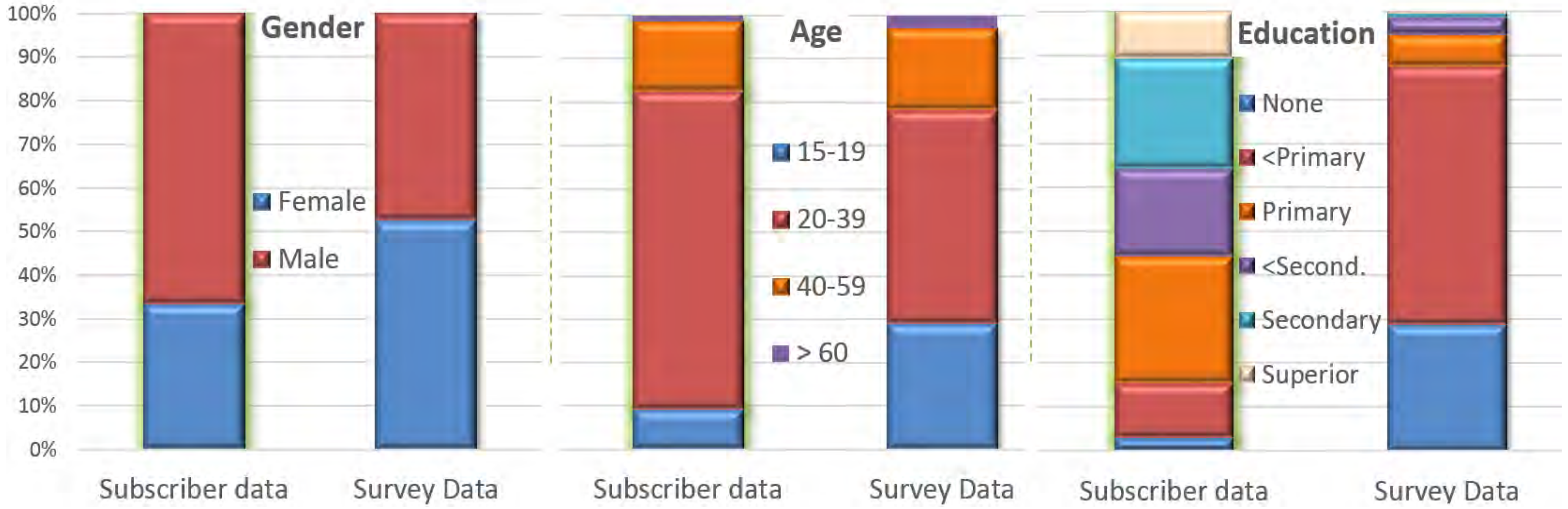
UCDAVIS
UNIVERSITY OF CALIFORNIA

Martin Hilbert
Department of Communication
hilbert@UCDavis.edu

N = n ?

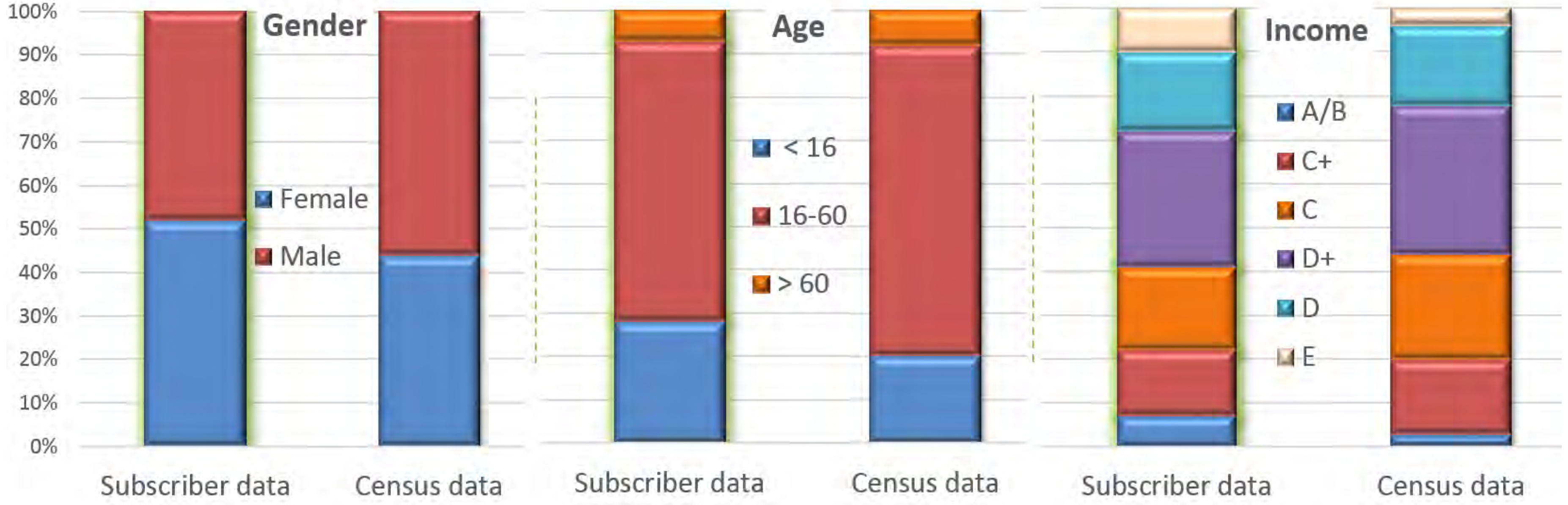
(a) Rwanda 2005/09:

mobile phone penetration of 2-20%

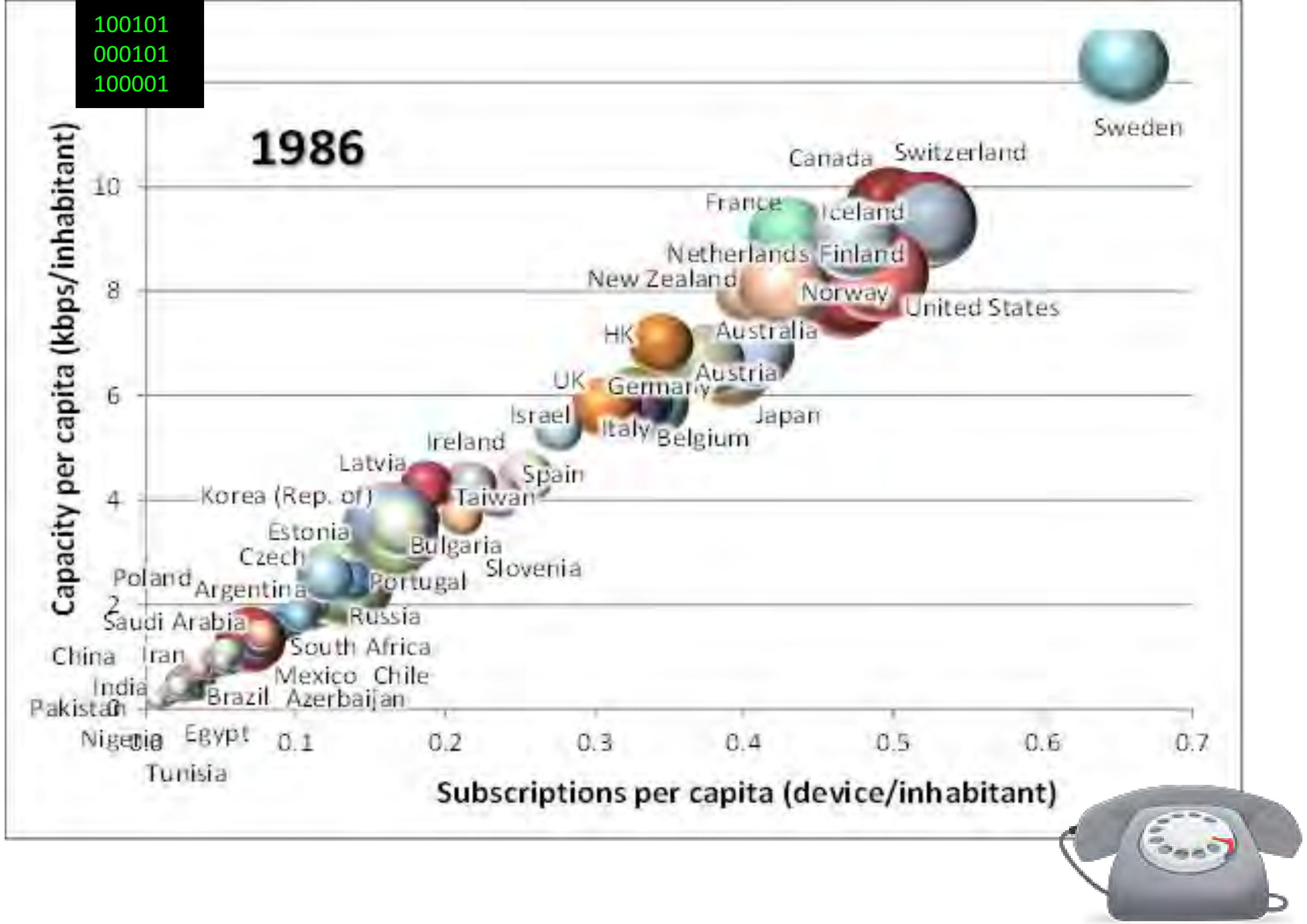


(b) LatAm economy 2009/10:

mobile phone penetration of 60-80%



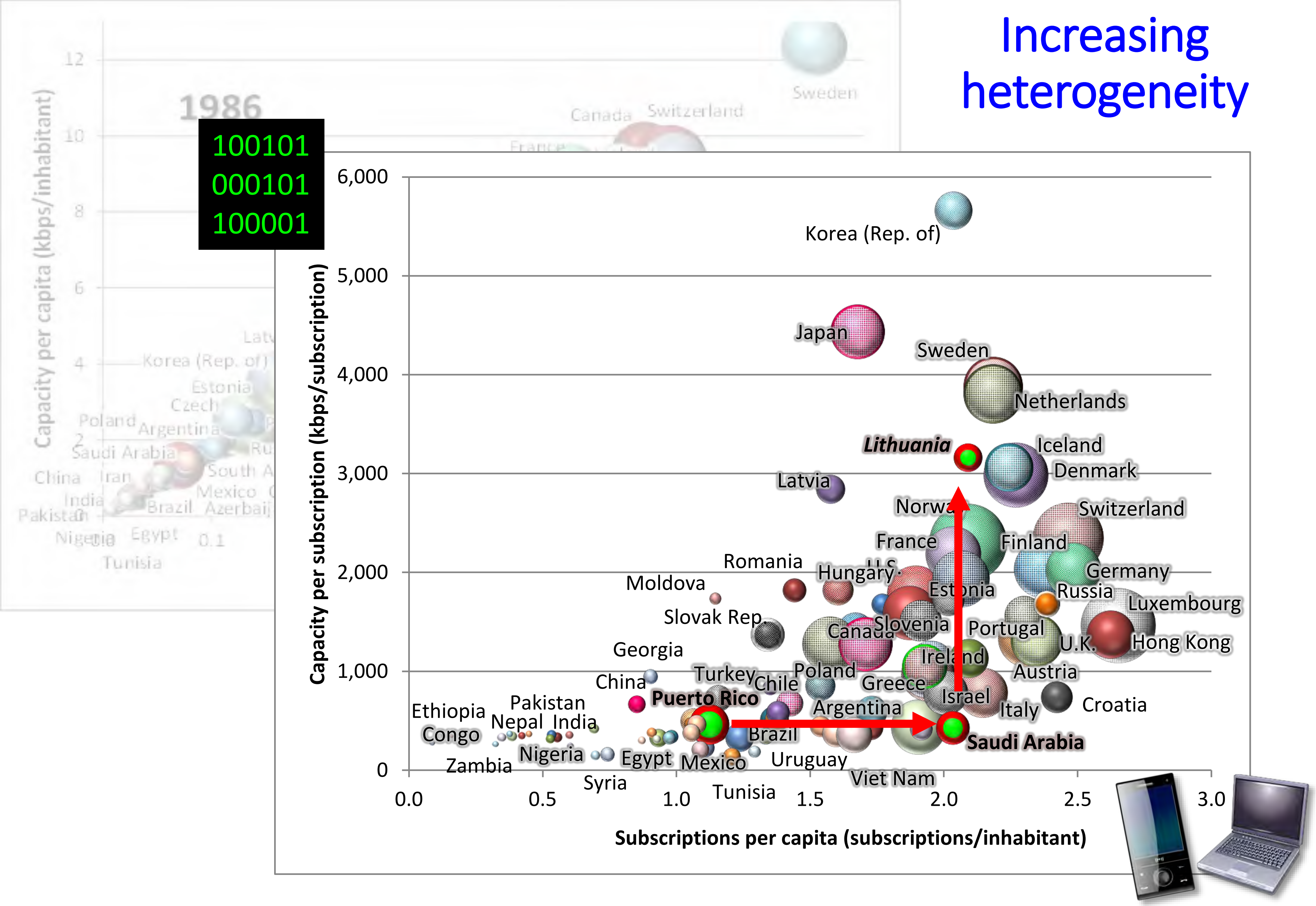
Source: (a) Blumenstock and Eagle (2012); (b) Frias-Martinez and Virseda (2013).



So we're all finally good now! …(?)

Source: Hilbert, M. (2013), Technological information inequality as an incessantly moving target: The redistribution of info. and communication capacities between 1986 and 2010. *Journal of the Assoc. for Info. Science and Technology*. http://www.martinhilbert.net/TechInfoInequality.pdf

Increasing heterogeneity

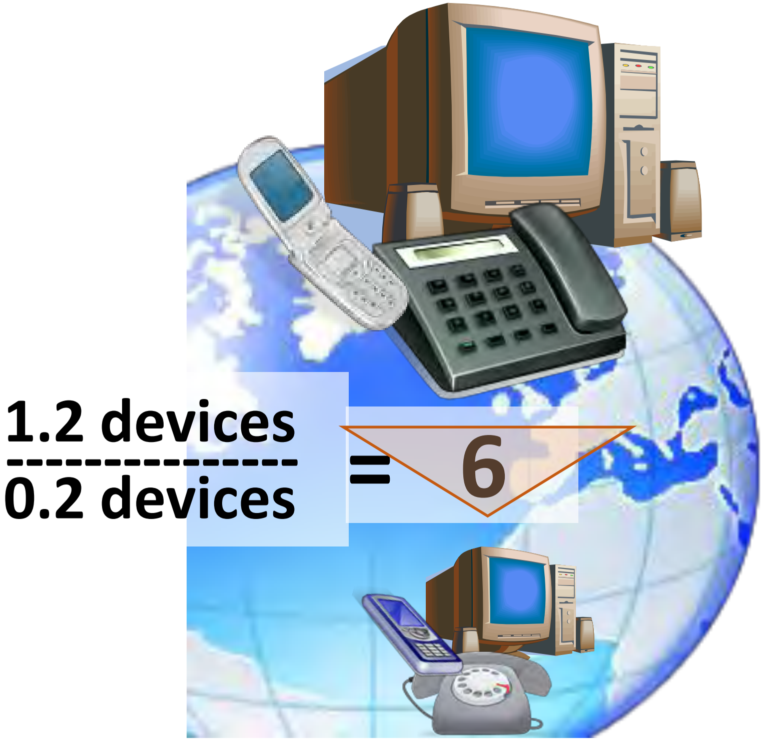


Source: Hilbert, M. (2013), Technological information inequality as an incessantly moving target: The redistribution of info. and communication capacities between 1986 and 2010. *Journal of the Assoc. for Info. Science and Technology*. http://www.martinhilbert.net/TechInfoInequality.pdf

Number of subscriptions of countries

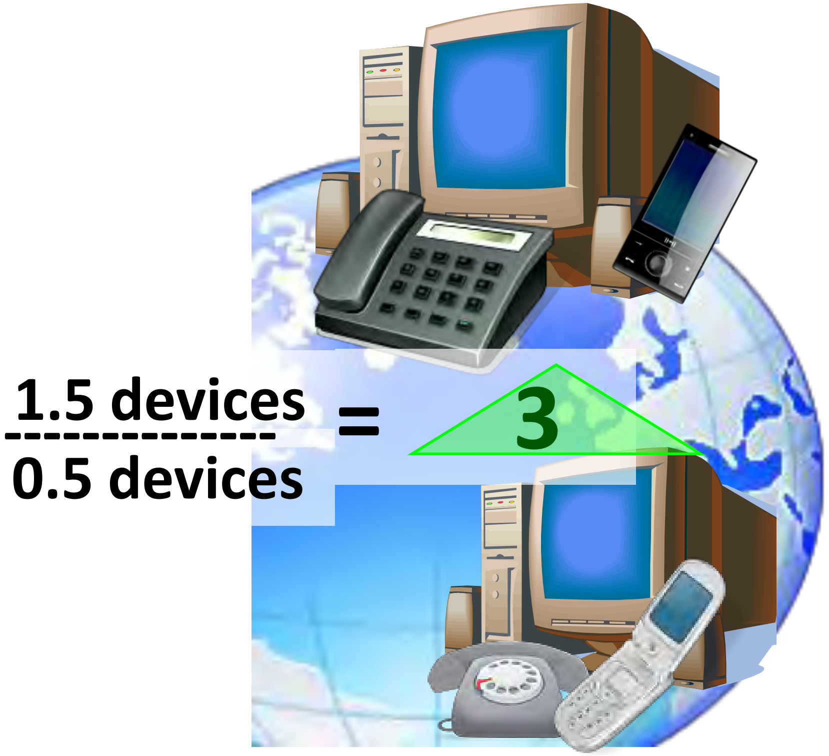
Telecom: OECD vrs. the rest of world

(fixed and mobile Internet and telephony SUBSCRIPTIONS per capita)



$\frac{1.2 \text{ devices}}{0.2 \text{ devices}} = 6$

2001



$\frac{1.5 \text{ devices}}{0.5 \text{ devices}} = 3$

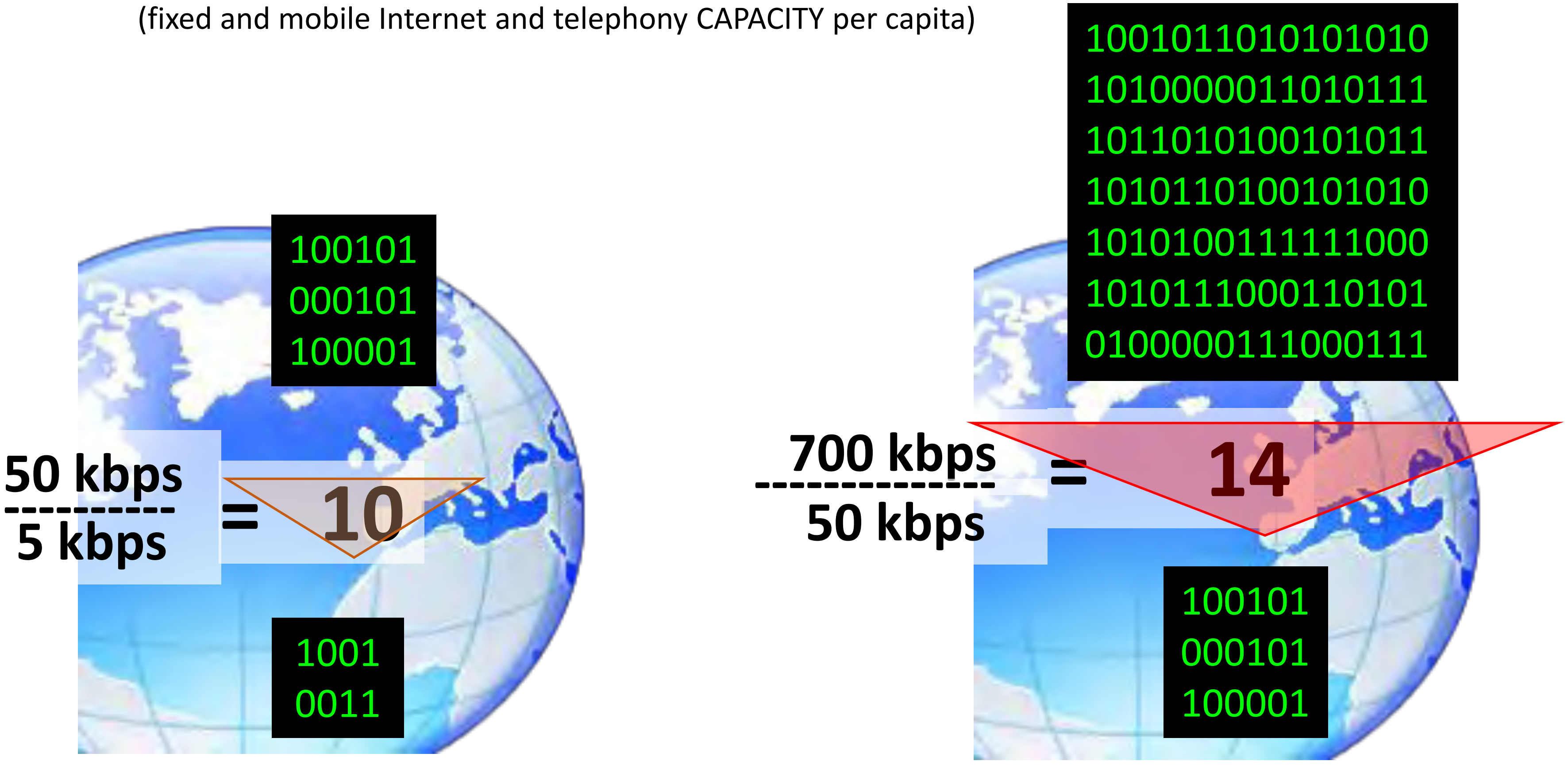
2006

Source: Hilbert, M. (2013), Technological information inequality as an incessantly moving target: The redistribution of info. and communication capacities between 1986 and 2010. *Journal of the Assoc. for Info. Science and Technology*. http://www.martinhilbert.net/TechInfoInequality.pdf

Telecommunication capacity of countries

Telecom: OECD vrs. the rest of world

(fixed and mobile Internet and telephony CAPACITY per capita)



100101
000101
100001

50 kbps
----
5 kbps
= 10

1001
0011

2001

1001011010101010
1010000011010111
1011010100101011
1010110100101010
1010100111111000
1010111000110101
0100000111000111

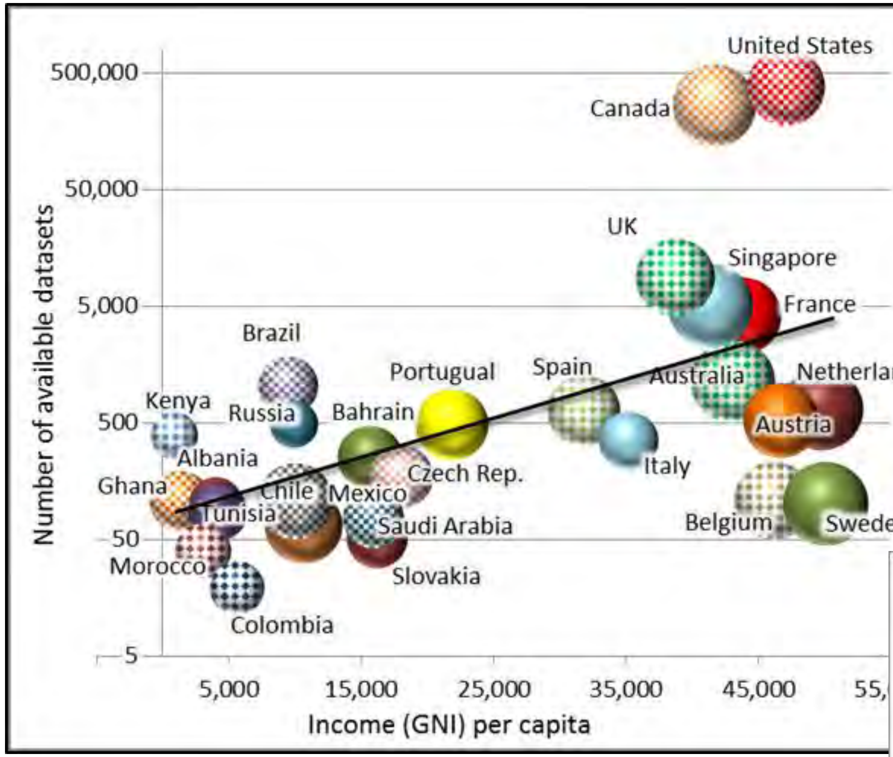
700 kbps
----
50 kbps
= 14

100101
000101
100001

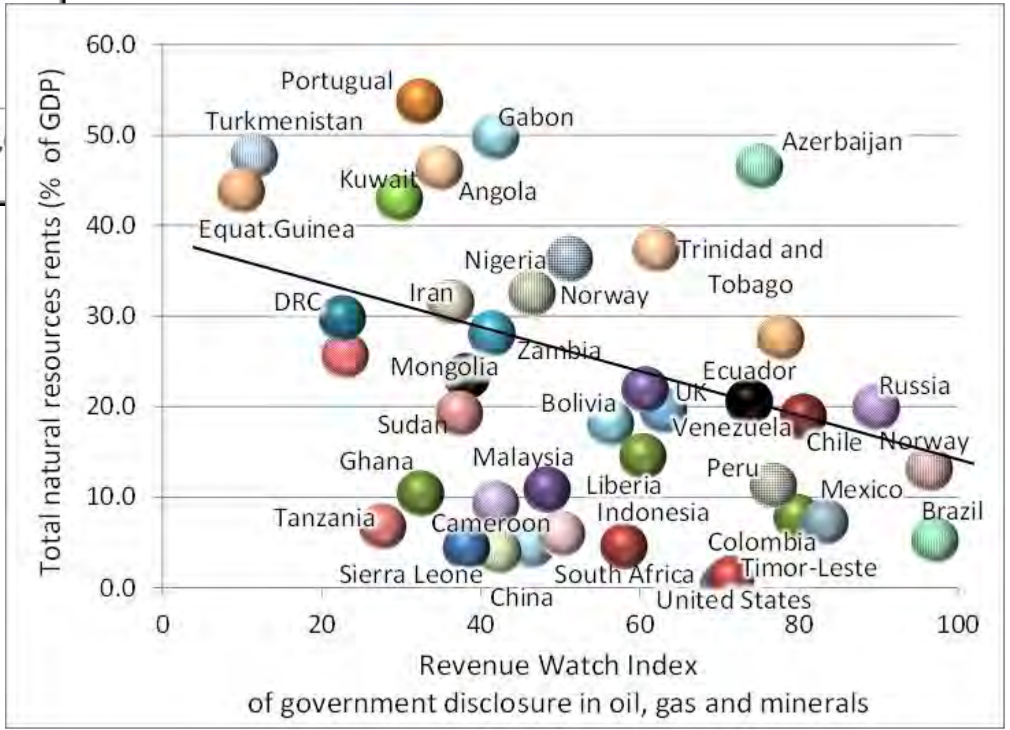
2006

Source: Hilbert, M. (2013), Technological information inequality as an incessantly moving target: The redistribution of info. and communication capacities between 1986 and 2010. *Journal of the Assoc. for Info. Science and Technology*. http://www.martinhilbert.net/TechInfoInequality.pdf

Content Divide



Open Government data: *Number of datasets provided on central government portal (vertical y-axis, logarithmic scale), Gross National Income per capita (horizontal x-axis), Corruption Perception Index (size of bubbles: larger bubbles, more transparent) (year=2011; n=27).*



Public data on natural resource extraction: *Natural resource rent vs. government data disclosure (year=2010; n=40).*