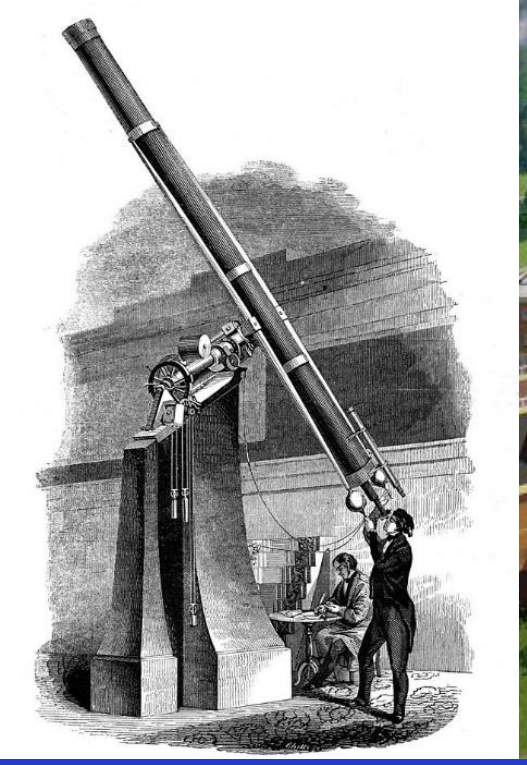


The Theory, Practice and Limits of Big Data for the Social Sciences

UCDAVIS
UNIVERSITY OF CALIFORNIA

Martin Hilbert
Department of Communication
hilbert@UCDavis.edu

Storage

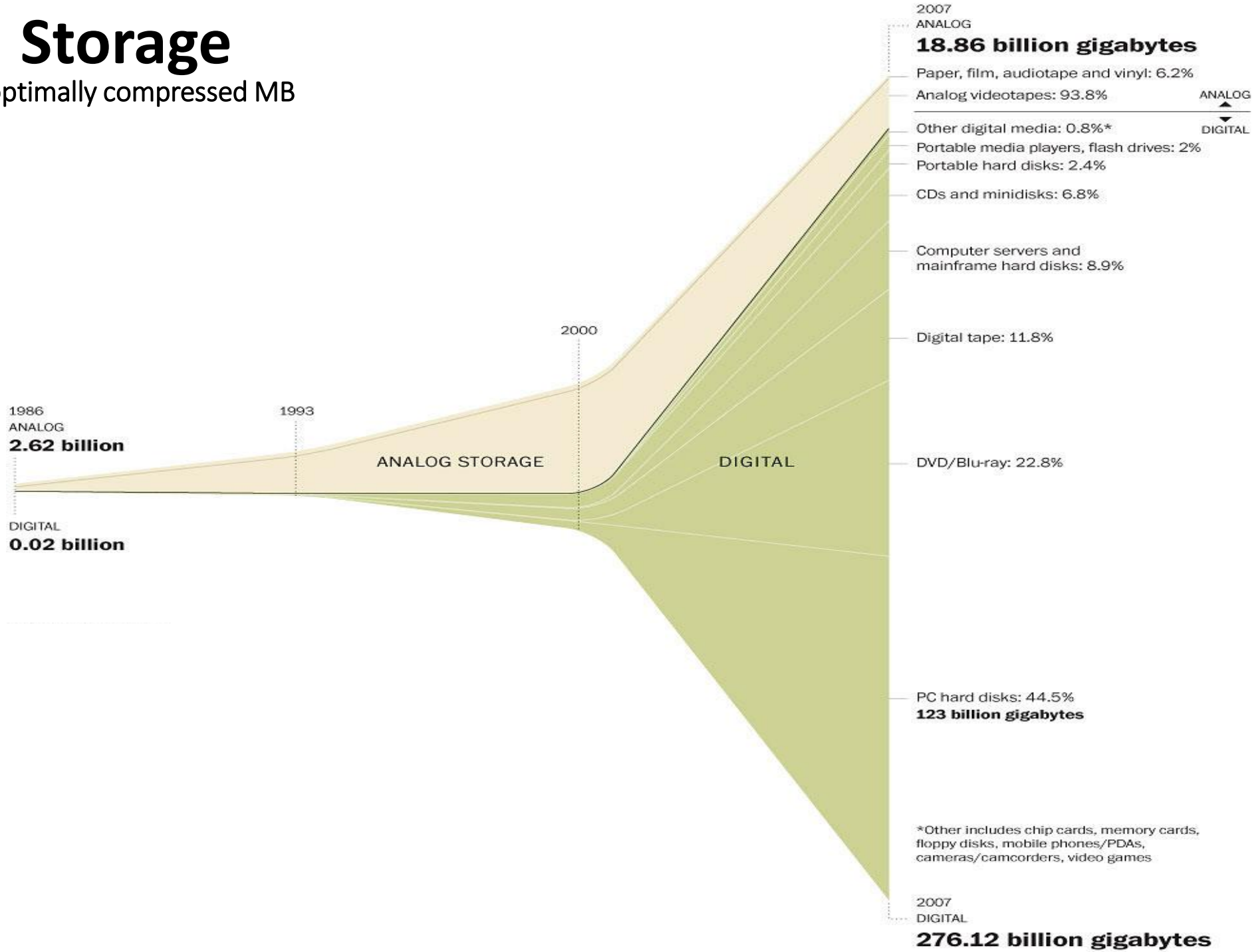
in optimally compressed MB

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

annual growth rate
1986-2007
25 %

Storage

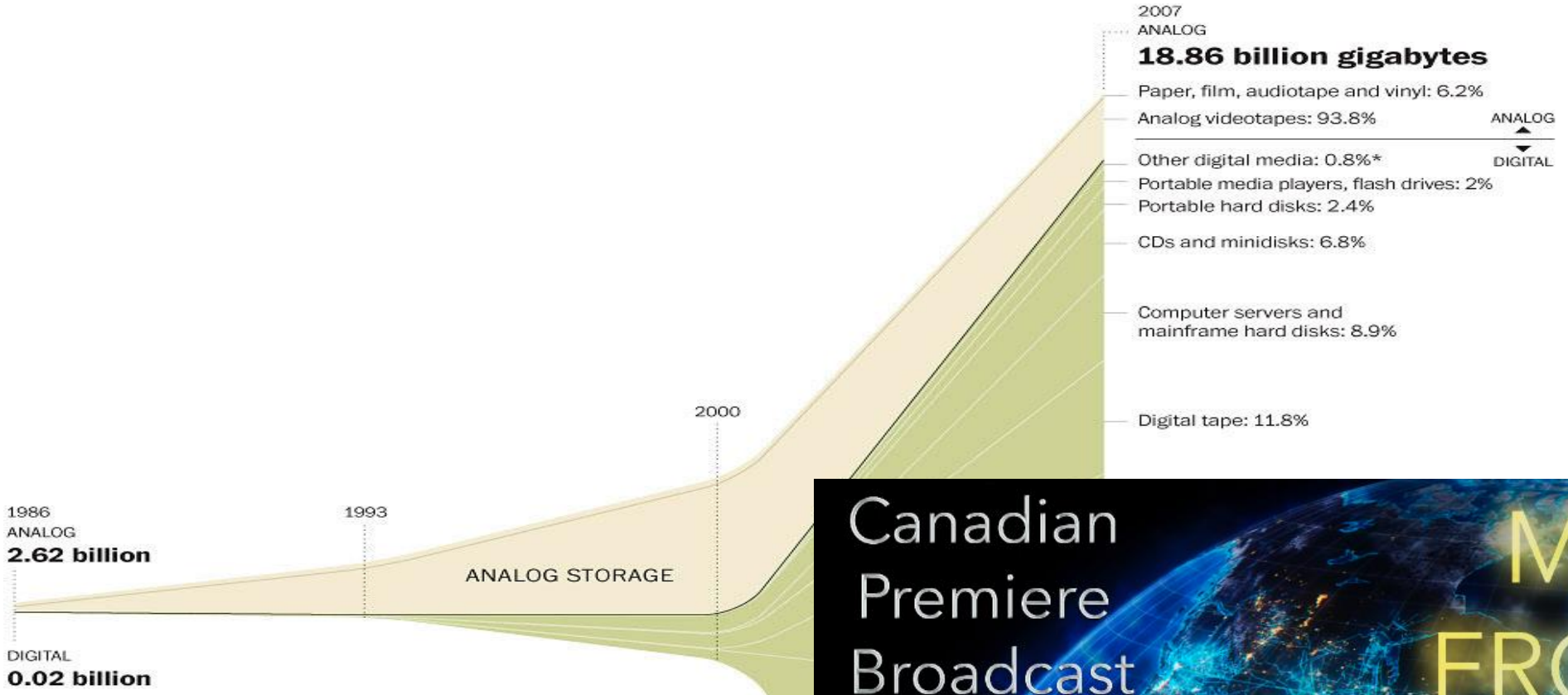
in optimally compressed MB



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

www.martinhilbert.net/WorldInfoCapacity.html

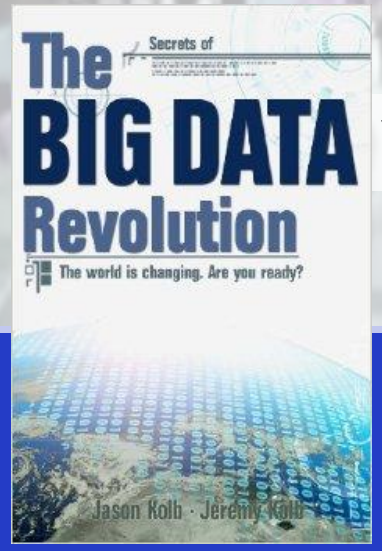
OECD

BETTER POLICIES FOR BETTER LIVES

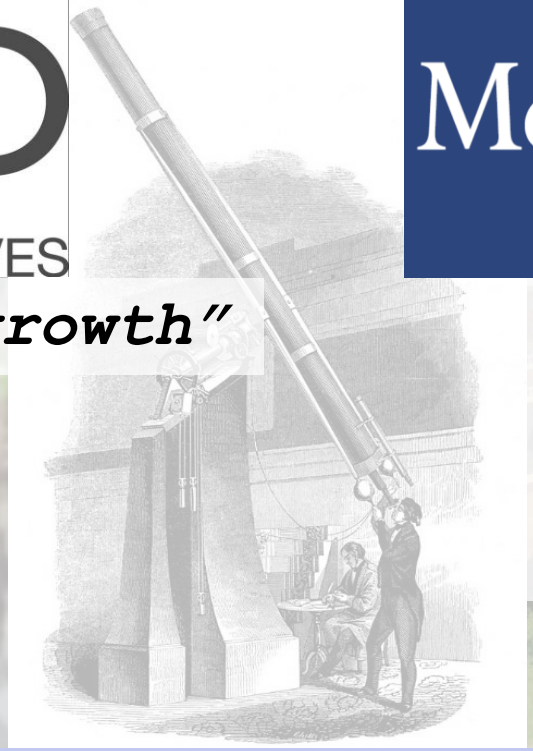
McKinsey&Company

"data as a new source of growth"

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

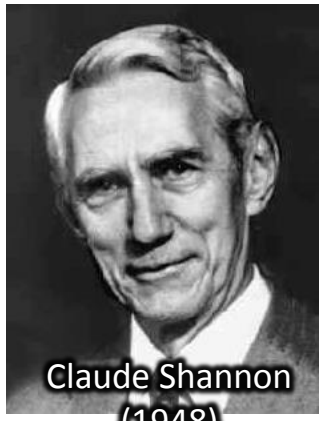


"the new oil"



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

Information & Growth



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

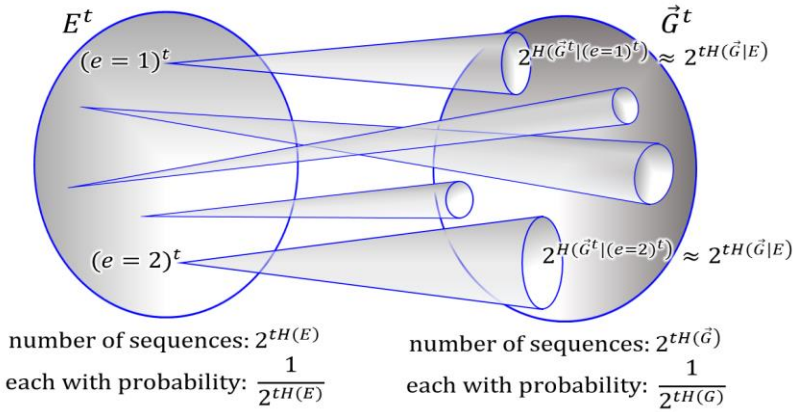
1 bit of **information** = reduction of **uncertainty** by half

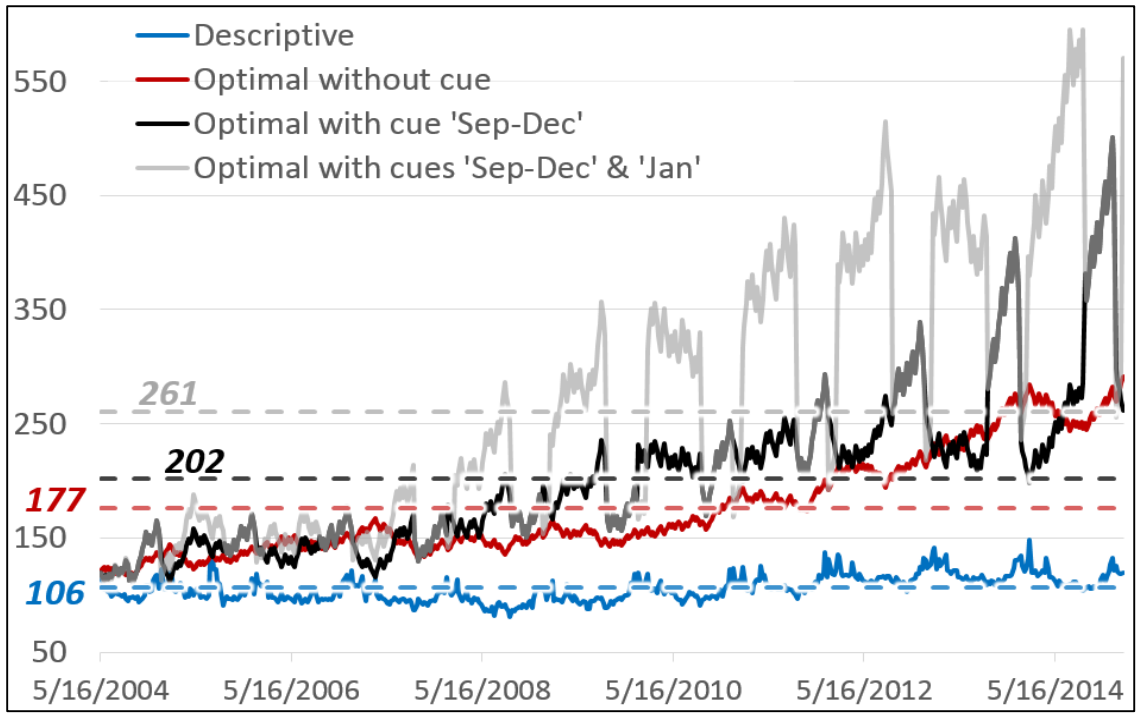
½ * uncertainty = 1 bit of information = **2** * Growth

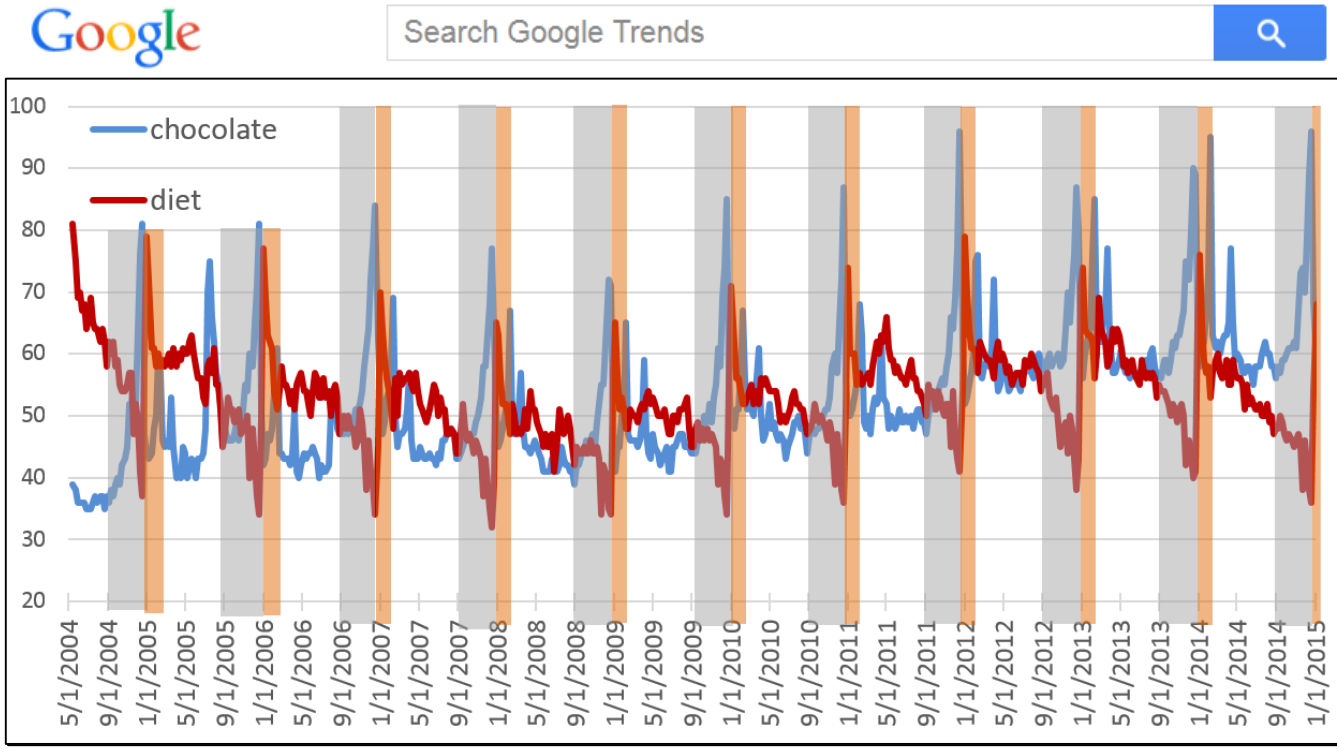
$$Growth = E_e[\log {}^d W] - H(E|\vec{G}) - D_{KL}(\vec{P}(e|g)\|P(e|m)) - I(E;\vec{G})$$

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

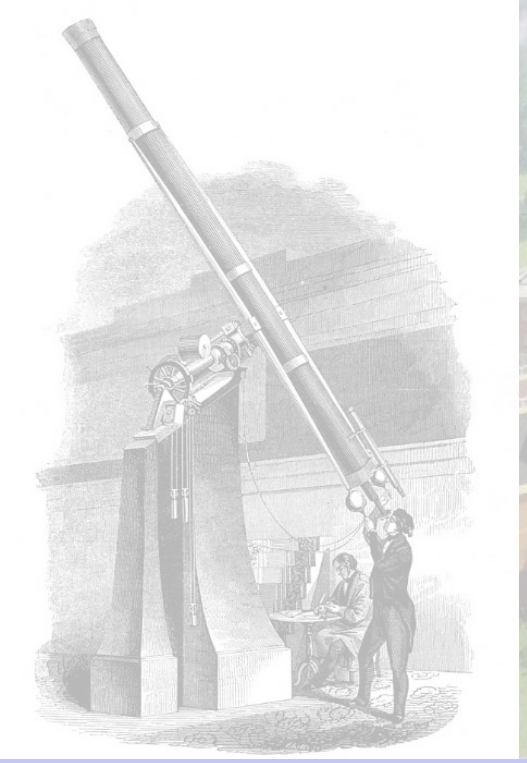






$$Growth = E_e[\log {}^d W] - H(E|\vec{G}) - D_{KL}(\vec{P}(e|g)\|P(e|m)) - I(E;\vec{G})$$

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

Digital Footprint

https://maps.google.com/locationhistory

Google

Location history

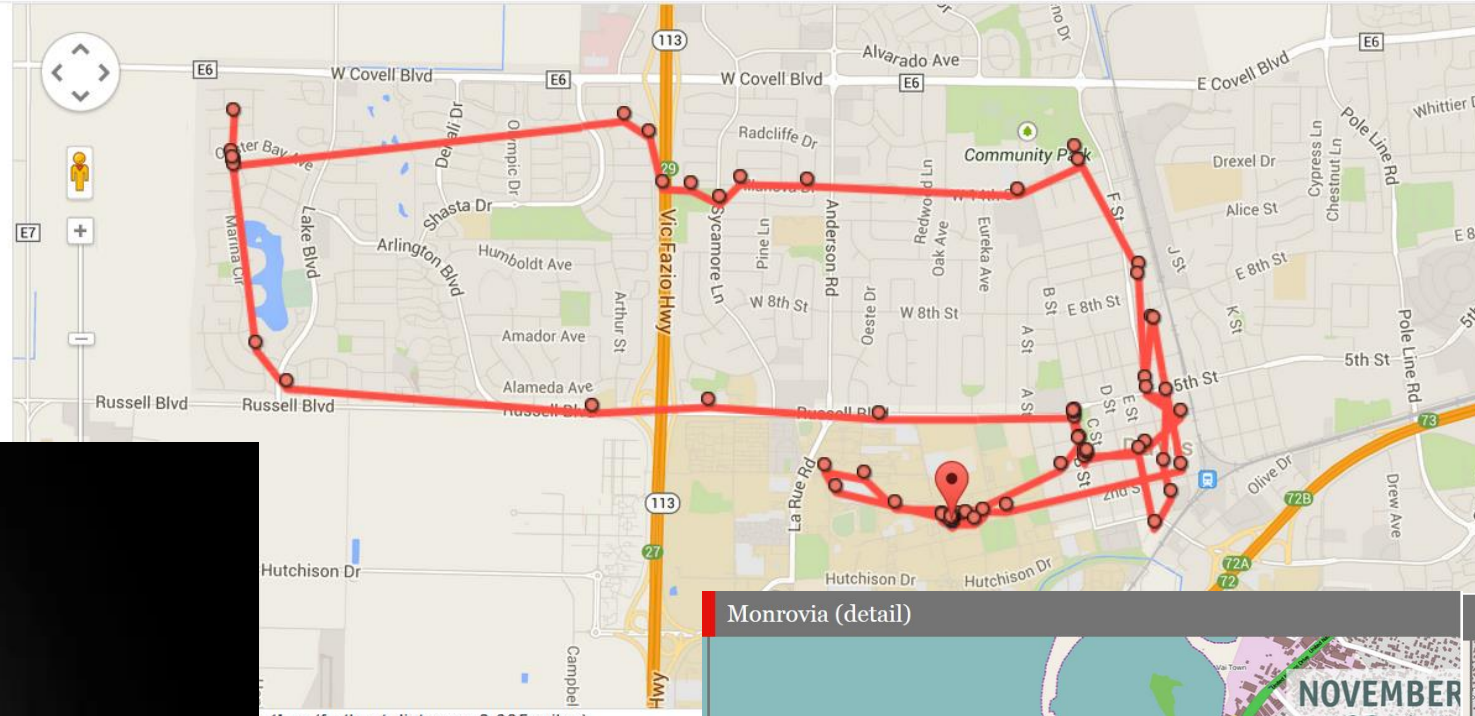
«	November 2014					»
Sun	Mon	Tue	Wed	Thu	Fri	Sat
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6

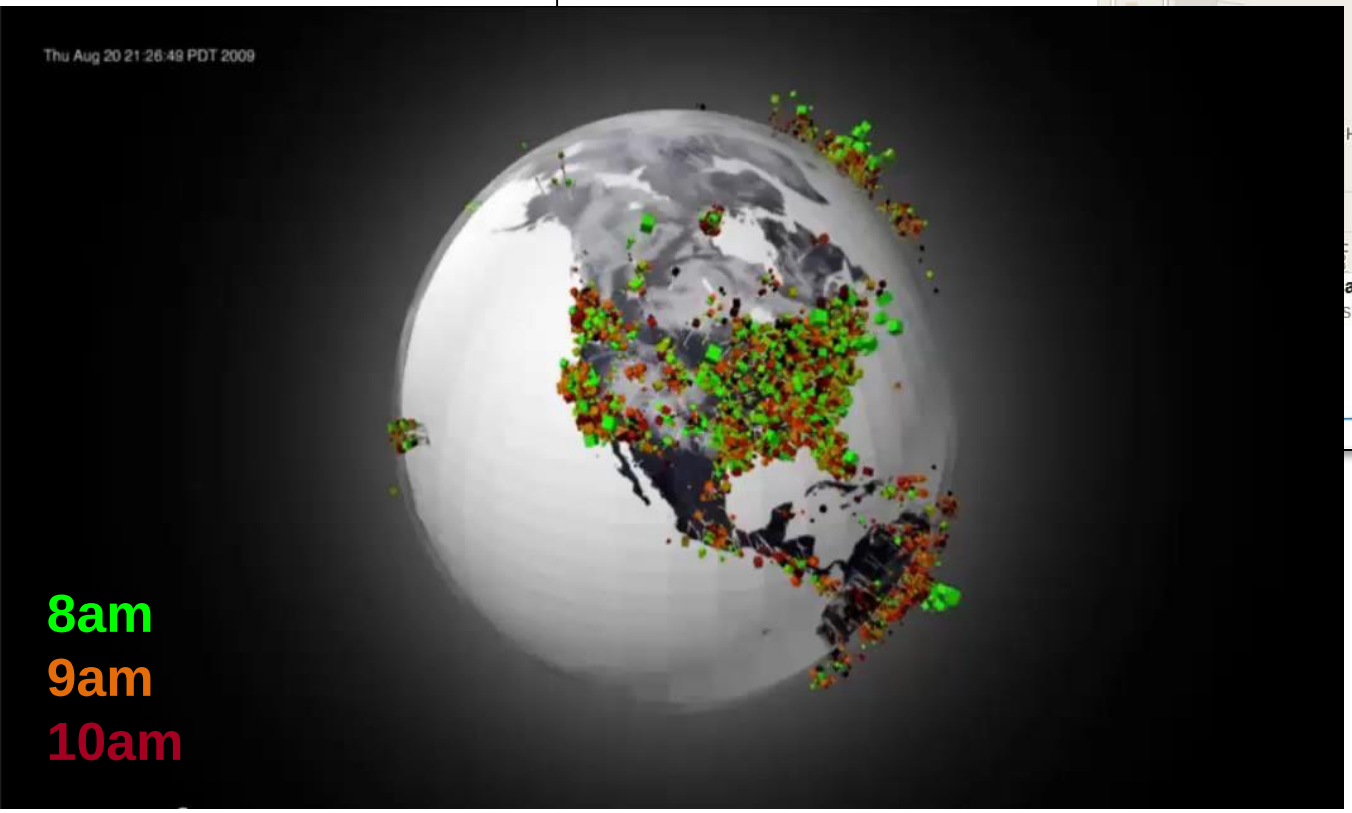
Show: 1 Day

November 17, 2014

▶ Show timestamps

Export to KML



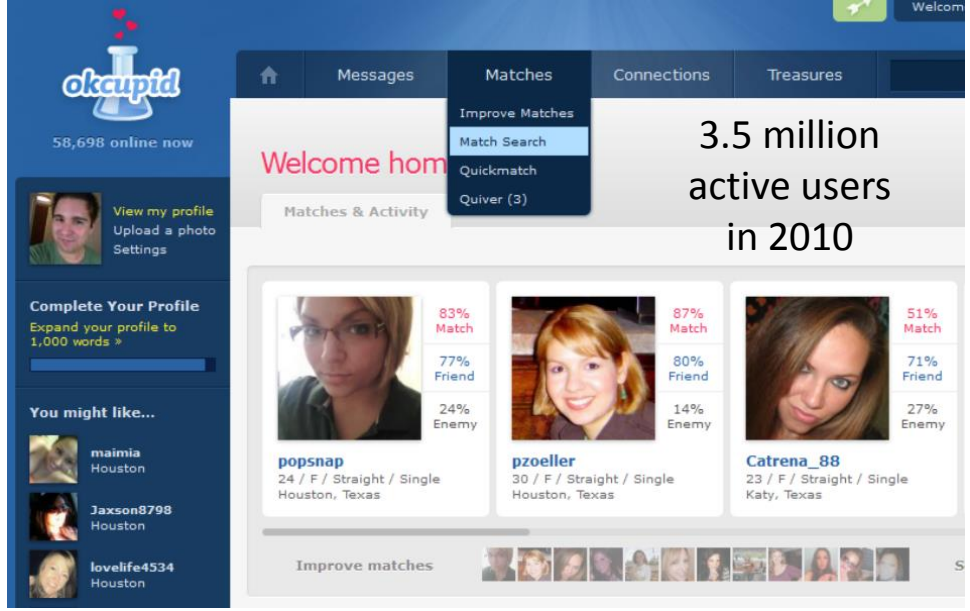


ation (farthest distance: 3.035 miles)

show location on map

4:00 AM 6:00 AM 8:00

TED-Ed. Jer Thorp(2013). Visualizing the world's Twitter data;
The Economist. (2014,). Off the map.



3.5 million active users in 2010

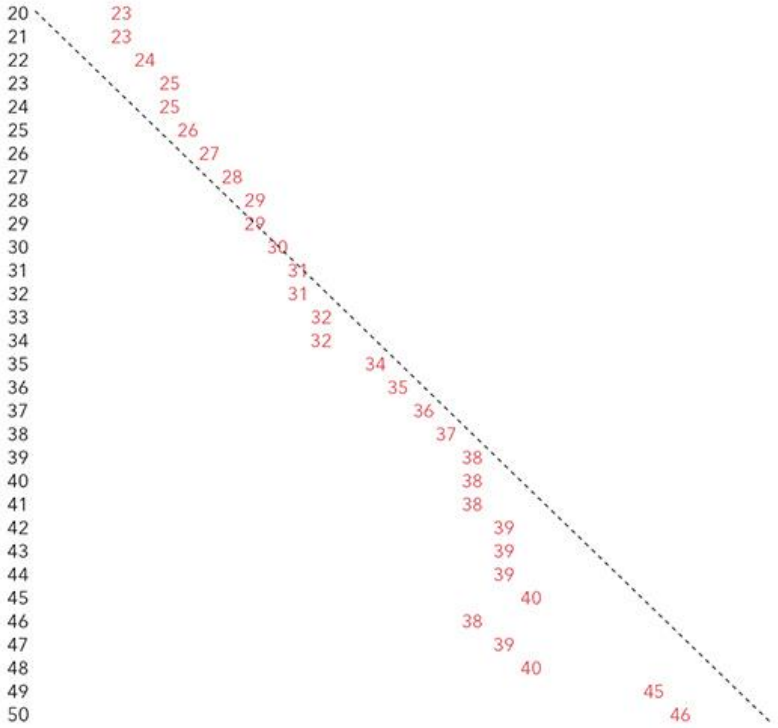
"_____ marriage"

		AVERAGE MONTHLY SEARCHES
sexless	**21,090**	
unhappy	6,029	
loveless	2,650	
sex starved	**1,658**	
no sex	**1,300**	

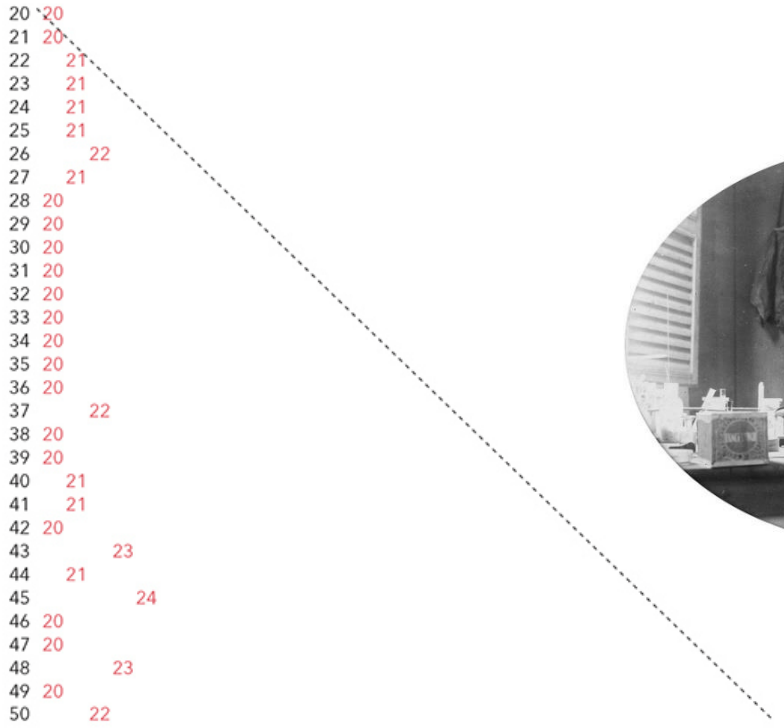
	"my husband . . .	"my wife. . .
. . . won't have sex with me"	**972**	**1,048**
. . . won't talk to me"	49	78
. . . won't touch me"	45	50

	"my boyfriend . . .	"my girlfriend . . .
. . . won't have sex with me"	**805**	**413**
. . . won't talk to me"	218	209
. . . won't text me back"	137	85

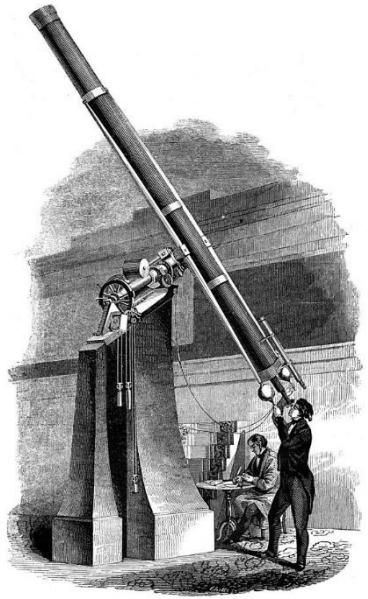
a woman's age vs. the age of the men who look best to her



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



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



arXiv.org > physics > arXiv:1312.7128

Physics > Popular Physics

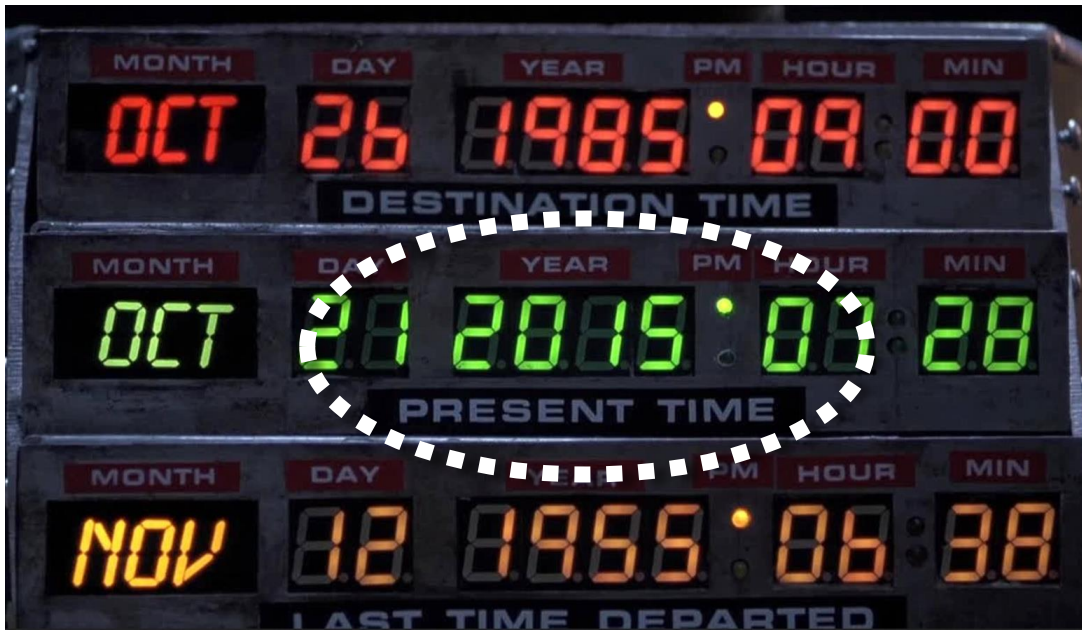
Searching the Internet for evidence of time travelers

Robert J. Nemiroff, Teresa Wilson

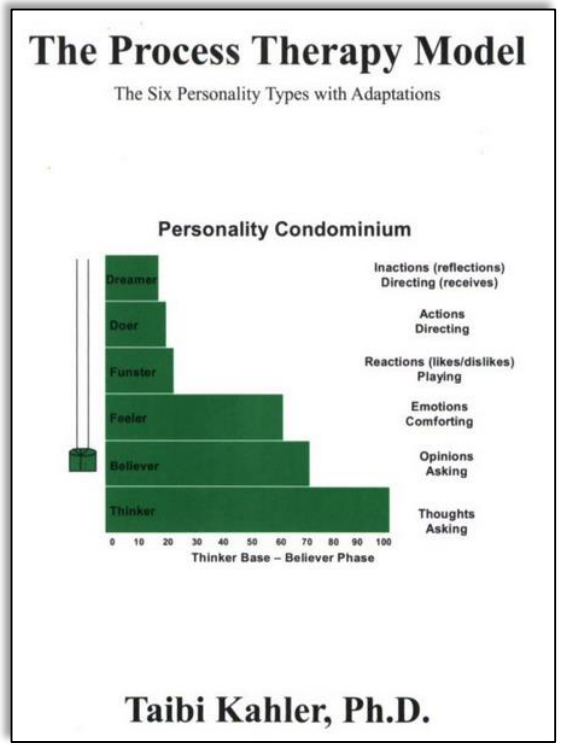
(Submitted on 26 Dec 2013)

"…prescient content placed on the Internet… prescient inquiries submitted to a search engine… direct Internet communication…

No time travelers were discovered…"



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



Matching Personality Types:

✓ Call average from 10 min to 5 min

✓ Customer Satisfaction from 47 % to 92%

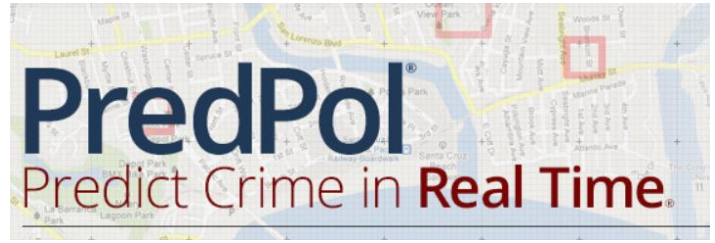
EMOTIONS-DRIVEN (30% of the population)
THOUGHTS-DRIVEN (25%)
REACTIONS-DRIVEN (20%)
OPINIONS-DRIVEN (10%)
REFLECTIONS-DRIVEN (10%)
ACTIONS-DRIVEN (5%)

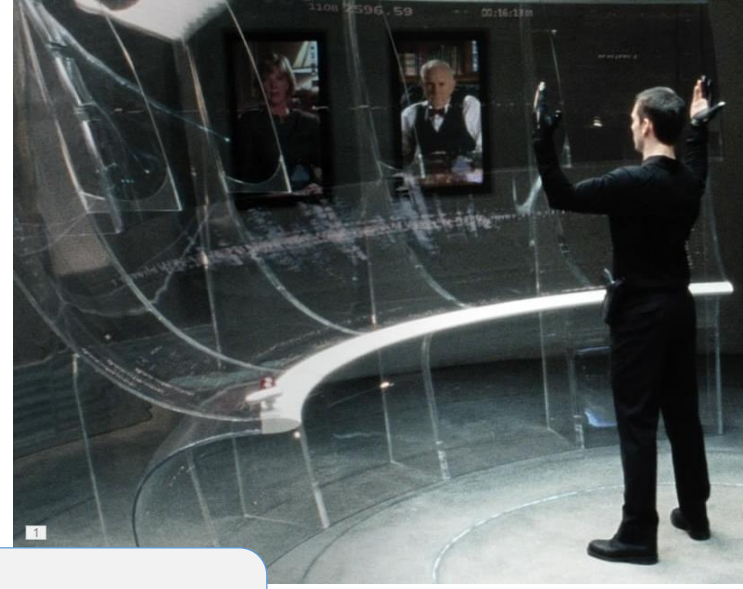
http://www.eloyalty.com ; http://www.mattersight.com/ ; http://www.fastcompany.com/1706766/how-personality-test-designed-pick-astronauts-taking-pain-out-customer-support ; http://www.ssca.com/resources/articles/104-the-history-of-the-process-communication-model-in-astronaut-selection ; http://www.forbes.com/forbes/2011/0214/entrepreneurs-kelly-conway-software-el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

Proxies vs. Reality

Predictive Policing LADP & SantaCruz

- Predictions to 500^2 feet
  - ✓ Crimes down 13 %; burglaries 11 %; car theft 8 %
    (while other districts went up during same period)

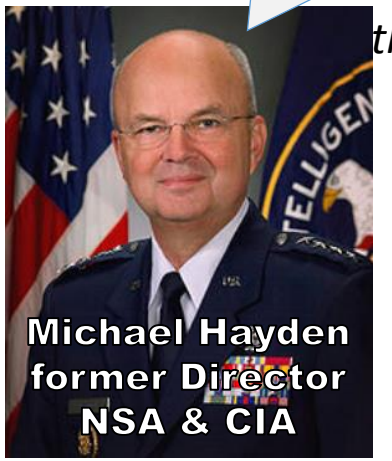




Homicide Parole candidates

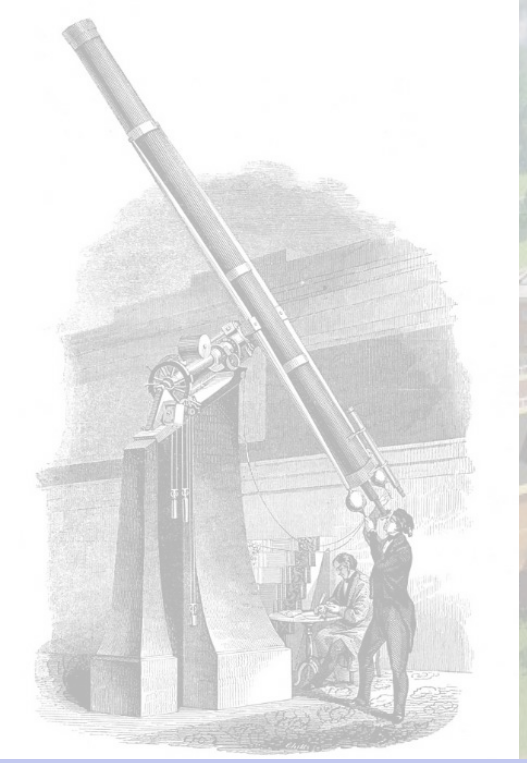
- 60 – 70 % correct who commits homicide

"We kill people based on metadata"



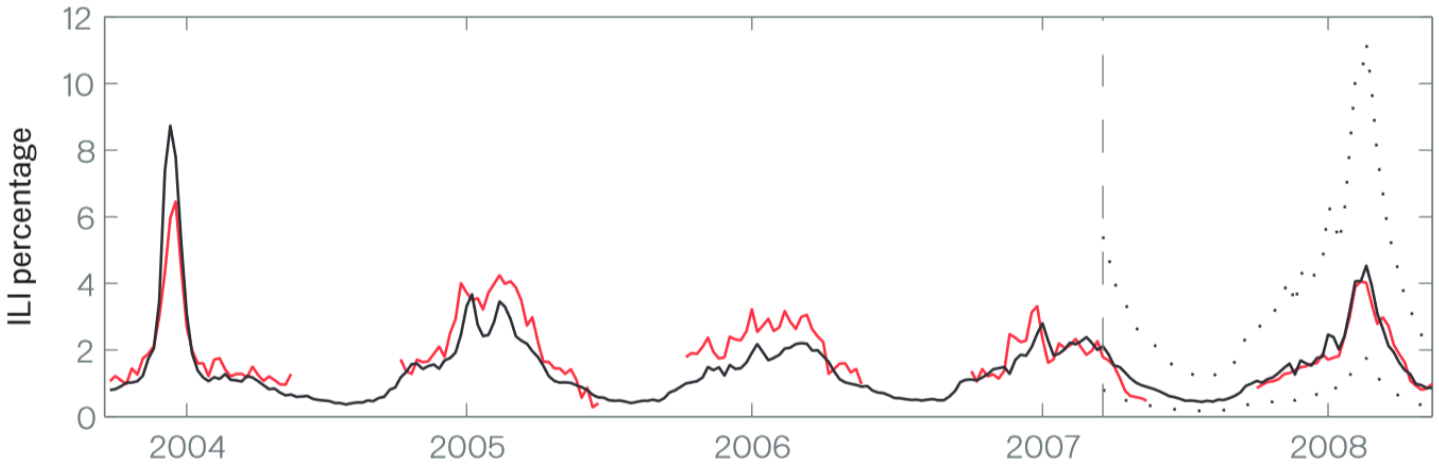
JSOC drone operator: "*It's of course assumed that the phone belongs to a human being who is nefarious and considered an 'unlawful enemy combatant.' This is where it gets very shady…*"

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*.



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

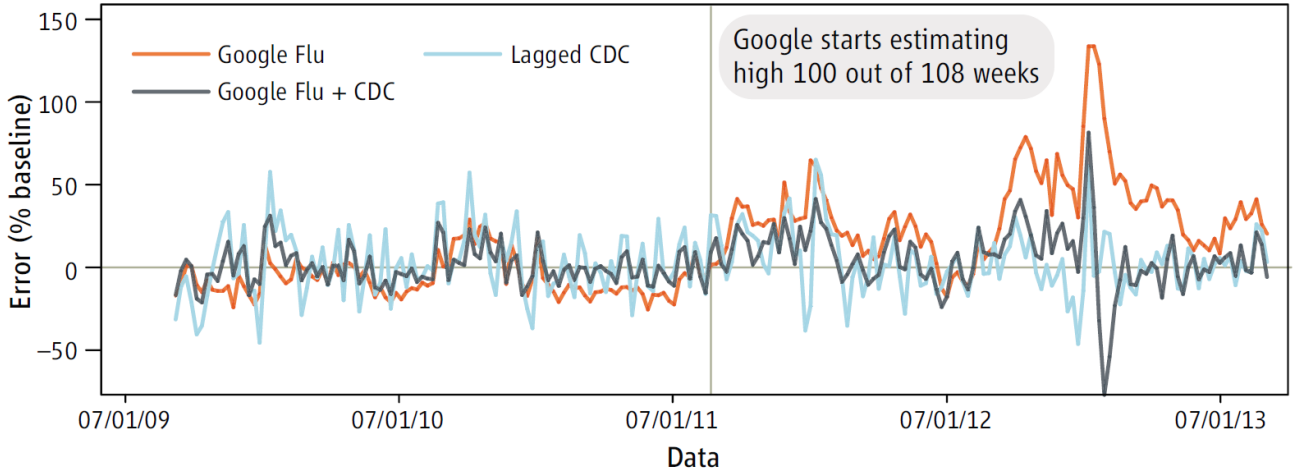
Data (from the past) has problems with changing futures



ECONOMETRIC POLICY EVALUATION: A CRITIQUE

Robert E. Lucas, Jr.

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



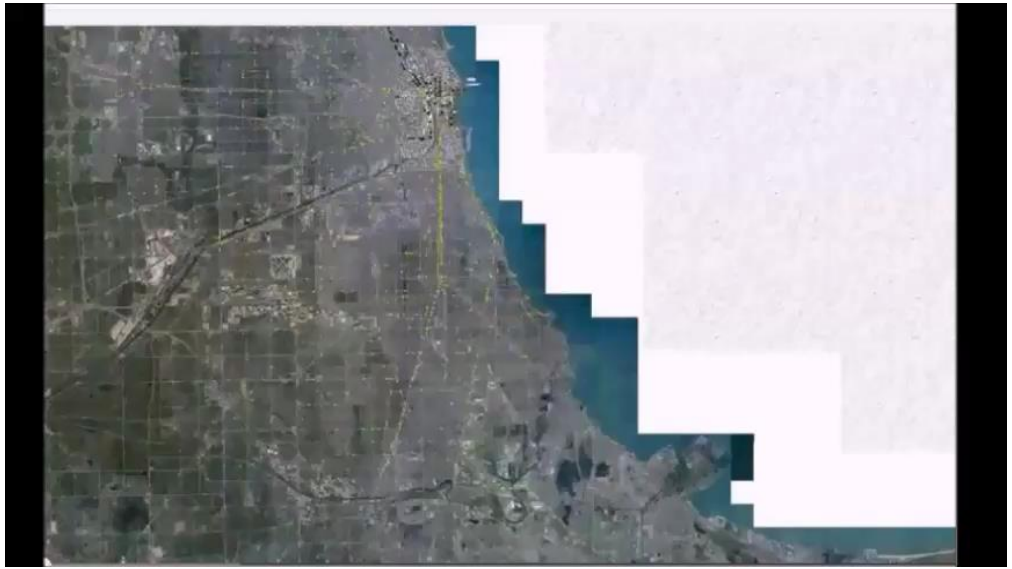
Sources: Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. ***Nature*** 457, 1012–1014 (***2009***).
Lazer et al. The Parable of Google Flu: Traps in Big Data Analysis. ***Science*** 343, 1203–1205 (***2014***).

Theoretical Models can deal with with changing futures!

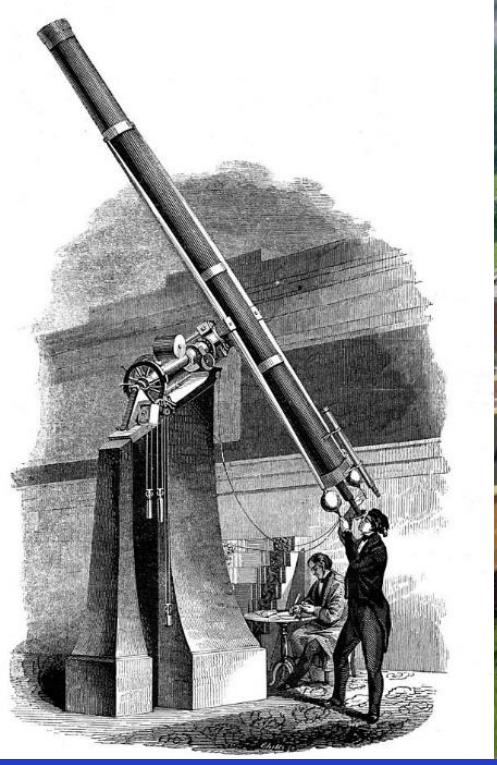
ECONOMETRIC POLICY EVALUATION: A CRITIQUE

Robert E. Lucas, Jr.

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



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

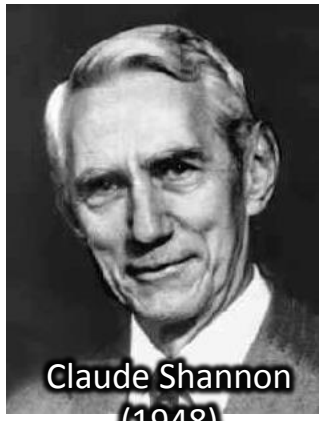


The Theory, Practice and Limits of Big Data for the Social Sciences

UC DAVIS
UNIVERSITY OF CALIFORNIA

Martin Hilbert
Department of Communication
hilbert@UCDavis.edu

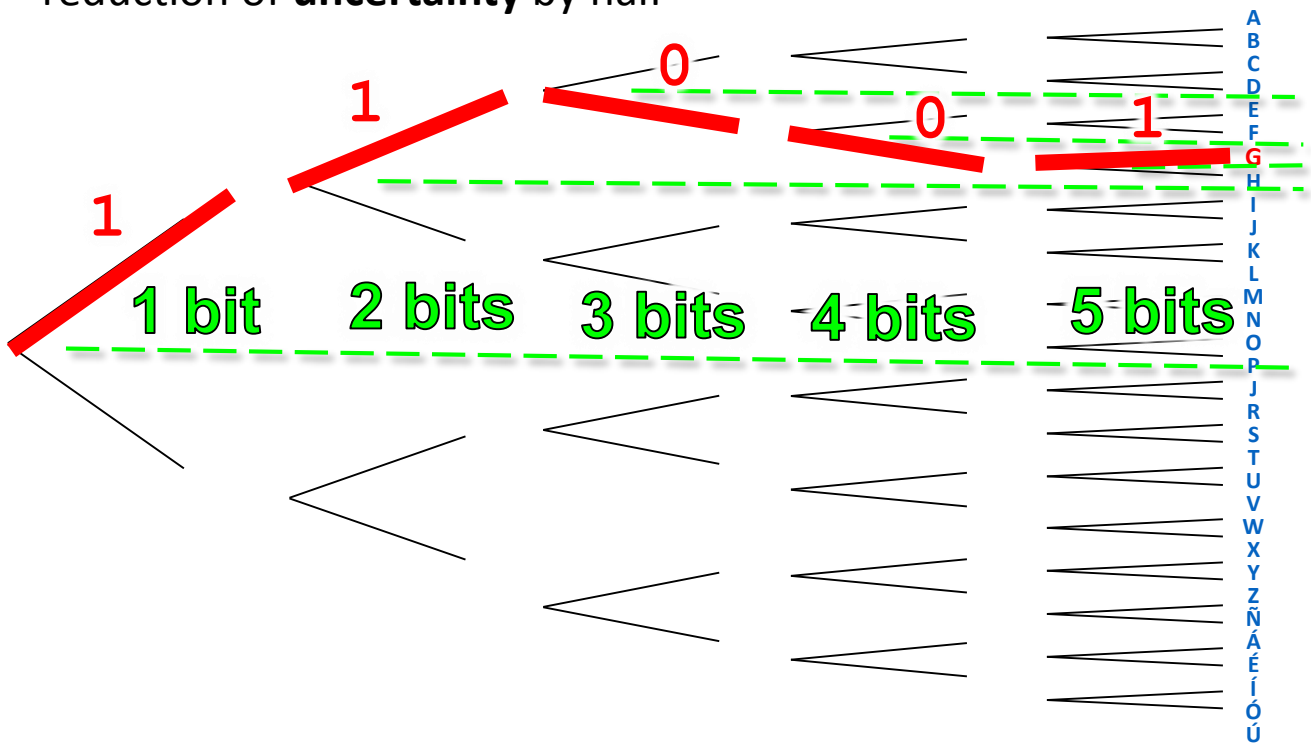
Information & Growth



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

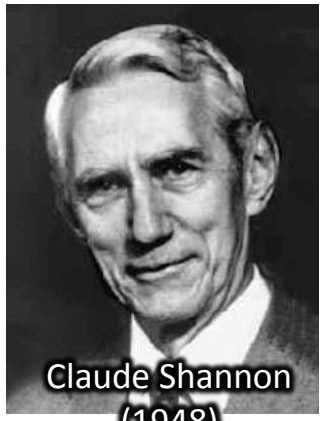
1 bit of **information** = reduction of **uncertainty** by half

transmit genius: "g"



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



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

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

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