BIG DATA

(a personal perspective)

Arthur Charpentier

charpentier.arthur@uqam.ca

http ://freakonometrics.hypotheses.org/

Institut des Actuaires, Paris, May 2014

"Big Data is like teenage sex :
everybody talks about it,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it..."

Dan Ariely, 2013 facebook.com/dan.ariely/

BIG DATA

(an actuarial & a statistical perspective)

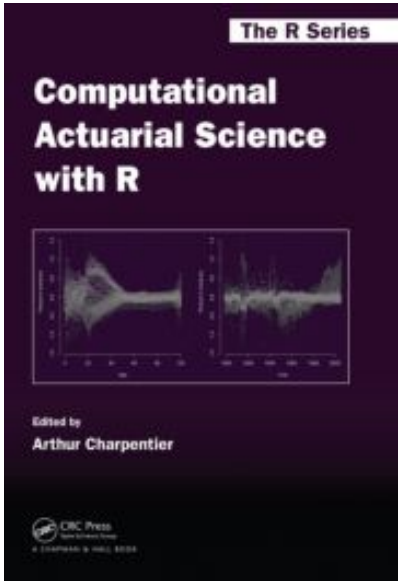
Arthur Charpentier

charpentier.arthur@uqam.ca

http ://freakonometrics.hypotheses.org/

Institut des Actuaires, Paris, May 2014

Professor of Actuarial Sciences, Mathematics Department, UQàM
(previously Economics Department, Univ. Rennes 1 & ENSAE Paristech
actuary in Hong Kong, IT & Stats FFSA)
PhD in Statistics (KU Leuven), Fellow Institute of Actuaries
MSc in Financial Mathematics (Paris Dauphine) & ENSAE
Editor of the freakonometrics.hypotheses.org's blog
Editor of (forthcoming) Computational Actuarial Science, CRC



Agenda

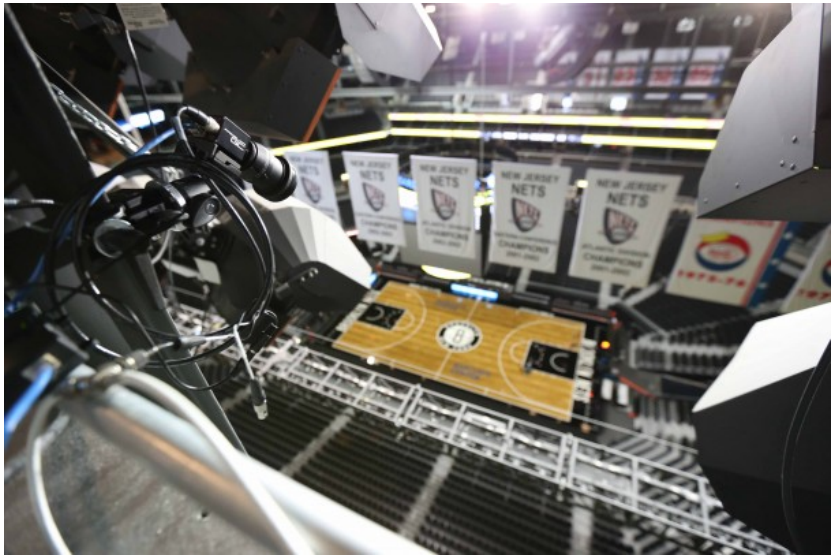
- Introduction
- **Examples of Big Data Issues**
- Basketball, Maps, Amazon, Netflix, Google, Wikipedia, and the Flu
- Some Thoughts about Big Data
- Defining Big Data
- **Volume**
- How Big is Big Data ?
- Correlation, Parametric and Non-Parametric Modeling
- **Velocity**
- High Frequency Trading
- **Variety**
- Textming, Graphs and Translation
- **Conclusion ?**

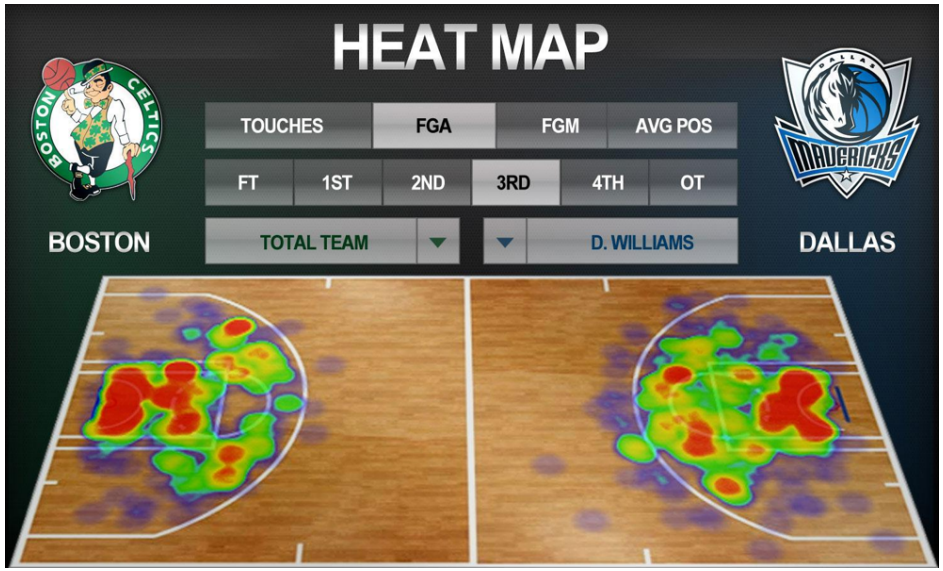
Basketball and Optical Tracking Data

Moneyball 2.0 : How Missile Tracking Cameras Are Remaking The NBA, 2012 fastcodesign.com

Predicting Points and Valuing Decisions in Real Time With NBA Optical Tracking Data. MIT Sloan, Sports Analytics Conference, Feb28-Mar1.

Basketball IRL and *Databall*, 2014 grantland.com





Basketball and Optical Tracking Data, the Matrix

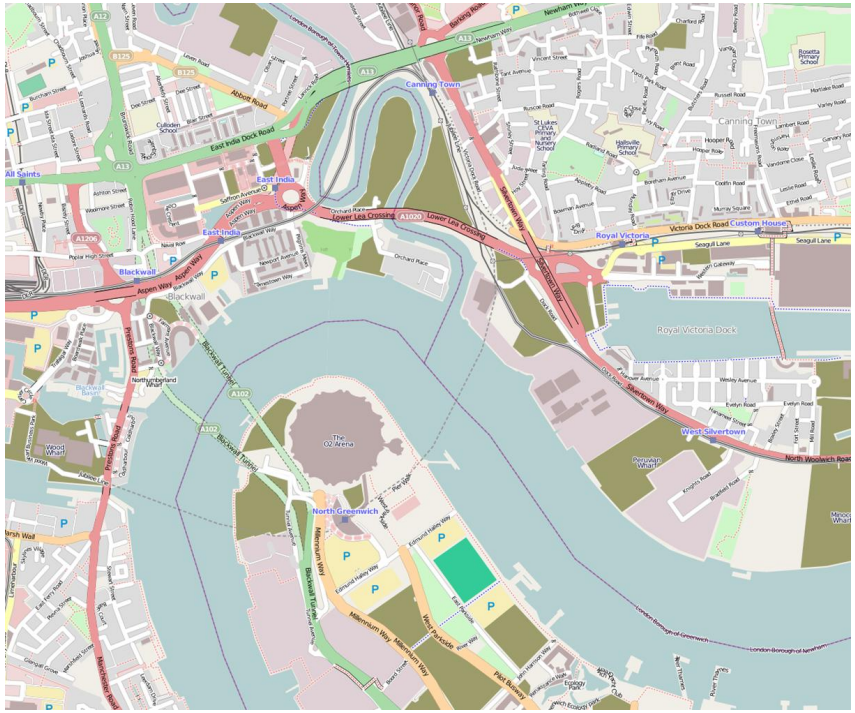
X	quarter	game_clock	ball_x	ball_y	ball_z	off1_ent	off2_ent	off3_ent	off4_ent	off5_ent	off1_x	off2_x	off3_x	off4_x	off5_x
1	1176094	4	9.24	30.46184	15.66440	5.73955	Parker	Neal	Leonard	Bonner	Duncan	31.20810	3.01034	5.71867	22.70688	22.66160
2	1176095	4	9.20	30.74898	16.56177	6.16081	Parker	Neal	Leonard	Bonner	Duncan	31.55756	2.90998	5.59025	22.60166	22.74272
3	1176096	4	9.16	31.14008	16.98810	6.08396	Parker	Neal	Leonard	Bonner	Duncan	31.89928	2.81066	5.47863	22.47178	22.89309
4	1176097	4	9.12	31.86791	17.09691	6.06588	Parker	Neal	Leonard	Bonner	Duncan	32.23151	2.71905	5.39309	22.31212	23.04843
5	1176098	4	9.08	31.99174	17.04352	5.80229	Parker	Neal	Leonard	Bonner	Duncan	32.54901	2.64163	5.33036	22.14823	23.20962
6	1176099	4	9.04	32.55396	16.93693	5.61020	Parker	Neal	Leonard	Bonner	Duncan	32.86782	2.56882	5.28070	21.97328	23.38824
7	1176100	4	9.00	33.01528	16.69451	5.24575	Parker	Neal	Leonard	Bonner	Duncan	33.17220	2.50098	5.24495	21.80377	23.58398
8	1176101	4	8.96	33.59858	16.52964	4.81964	Parker	Neal	Leonard	Bonner	Duncan	33.47203	2.44498	5.21824	21.61287	23.79921
9	1176102	4	8.92	33.75560	16.17879	4.53087	Parker	Neal	Leonard	Bonner	Duncan	33.76465	2.39688	5.19696	21.39753	24.03398
10	1176103	4	8.88	34.28415	15.67568	4.30612	Parker	Neal	Leonard	Bonner	Duncan	34.04971	2.34841	5.17501	21.17353	24.28544
11	1176104	4	8.84	34.26797	15.54423	3.69690	Parker	Neal	Leonard	Bonner	Duncan	34.32216	2.30455	5.15439	20.93744	24.55324
12	1176105	4	8.80	34.63560	15.57366	3.29938	Parker	Neal	Leonard	Bonner	Duncan	34.58345	2.26725	5.14478	20.68329	24.83165

	off1_y	off2_y	off3_y	off4_y	off5_y	def1_ent	def2_ent	def3_ent	def4_ent	def5_ent	def1_x	def2_x	def3_x	def4_x	def5_x	def1_y
1	18.74699	3.09636	47.53907	-1.30109	26.34201	Livingston	Ellington	Waiters	Gee	Zeller	28.39575	7.26153	6.61196	19.10779	18.49074	23.89819
2	18.44052	2.95225	47.67462	-1.23840	26.21422	Livingston	Ellington	Waiters	Gee	Zeller	28.75253	7.14682	6.52986	19.03464	18.65387	23.47161
3	18.16281	2.82953	47.79135	-1.19877	26.05057	Livingston	Ellington	Waiters	Gee	Zeller	29.10403	7.04102	6.45409	18.95038	18.82159	23.04556
4	17.92252	2.70648	47.88714	-1.16482	25.87835	Livingston	Ellington	Waiters	Gee	Zeller	29.45096	6.94894	6.38321	18.84988	18.99275	22.61536
5	17.70649	2.58385	47.96527	-1.12016	25.69984	Livingston	Ellington	Waiters	Gee	Zeller	29.77094	6.87188	6.31862	18.73036	19.16549	22.20077
6	17.52267	2.46749	48.03880	-1.06383	25.51339	Livingston	Ellington	Waiters	Gee	Zeller	30.08660	6.81373	6.26594	18.60247	19.33952	21.80239
7	17.35745	2.36363	48.10369	-1.00097	25.31308	Livingston	Ellington	Waiters	Gee	Zeller	30.39400	6.76269	6.21968	18.45502	19.52401	21.42181
8	17.20283	2.29945	48.15793	-0.94003	25.10529	Livingston	Ellington	Waiters	Gee	Zeller	30.68864	6.71973	6.18566	18.29982	19.71082	21.06233
9	17.07512	2.24979	48.20172	-0.87742	24.88480	Livingston	Ellington	Waiters	Gee	Zeller	31.00225	6.67899	6.16005	18.14136	19.90622	20.72579
10	16.97912	2.20291	48.23125	-0.80572	24.66389	Livingston	Ellington	Waiters	Gee	Zeller	31.31860	6.64573	6.15086	17.98013	20.11420	20.41285
11	16.90115	2.16288	48.25132	-0.72214	24.43810	Livingston	Ellington	Waiters	Gee	Zeller	31.63056	6.61147	6.14776	17.81079	20.32861	20.12016
12	16.84963	2.13311	48.26263	-0.63485	24.21707	Livingston	Ellington	Waiters	Gee	Zeller	31.93198	6.58326	6.15569	17.63143	20.55313	19.85651

	def2_y	def3_y	def4_y	def5_y	prob_pass1	prob_pass2	prob_pass3	prob_pass4	prob_pass5	prob_make	prob_miss	prob_TO	epv
1	7.87051	36.60250	2.23678	20.79599	NA	0.000939464	0.0005215383	0.0007800394	0.0011080469	2.061396e-05	3.530900e-05	4.582805e-05	0.9694958
2	7.78649	36.53292	2.26203	20.82908	NA	0.001121139	0.0006538326	0.0007140829	0.0009235663	1.484419e-05	2.593288e-05	4.617669e-05	0.9679251
3	7.72441	36.46256	2.29366	20.86528	NA	0.001297789	0.0007020684	0.0006851674	0.0007977597	1.005769e-05	1.793289e-05	4.619174e-05	0.9660304
4	7.67786	36.39285	2.33380	20.90569	NA	0.001430896	0.0007368078	0.0006935951	0.0007356085	6.828948e-06	1.242612e-05	4.780346e-05	0.9639335
5	7.64493	36.31915	2.38695	20.95015	NA	0.002003120	0.0009443742	0.0010669102	0.0007972303	9.493768e-06	1.762269e-05	4.879118e-05	0.9616730
6	7.61930	36.24331	2.44617	20.99211	NA	0.002028606	0.0009864967	0.0010422796	0.0007023052	6.369296e-06	1.206092e-05	5.125298e-05	0.9586997
7	7.60550	36.17378	2.51259	21.04188	NA	0.001883377	0.0010428576	0.0009959774	0.0005648587	4.288592e-06	8.279725e-06	5.458791e-05	0.9556066
8	7.60138	36.10760	2.59764	21.08655	NA	0.001768323	0.0011071285	0.0009526763	0.0005294989	2.934505e-06	5.770166e-06	5.815748e-05	0.9525035
9	7.60426	36.05122	2.66967	21.11940	NA	0.001577596	0.0011333265	0.0008995217	0.0005457132	2.002034e-06	4.016175e-06	6.117891e-05	0.9493675
10	7.61621	35.99961	2.74528	21.13542	NA	0.001415889	0.0011766095	0.0008518788	0.0005863730	1.391663e-06	2.850721e-06	6.527485e-05	0.9462897
11	7.63696	35.94937	2.85717	21.14073	NA	0.001238335	0.0011980263	0.0008016188	0.0006205814	9.694368e-07	2.028000e-06	6.893162e-05	0.9432457
12	7.66533	35.90460	2.97045	21.14716	NA	0.001071548	0.0012006243	0.0007474017	0.0006343324	6.795564e-07	1.451362e-06	7.255664e-05	0.9402737

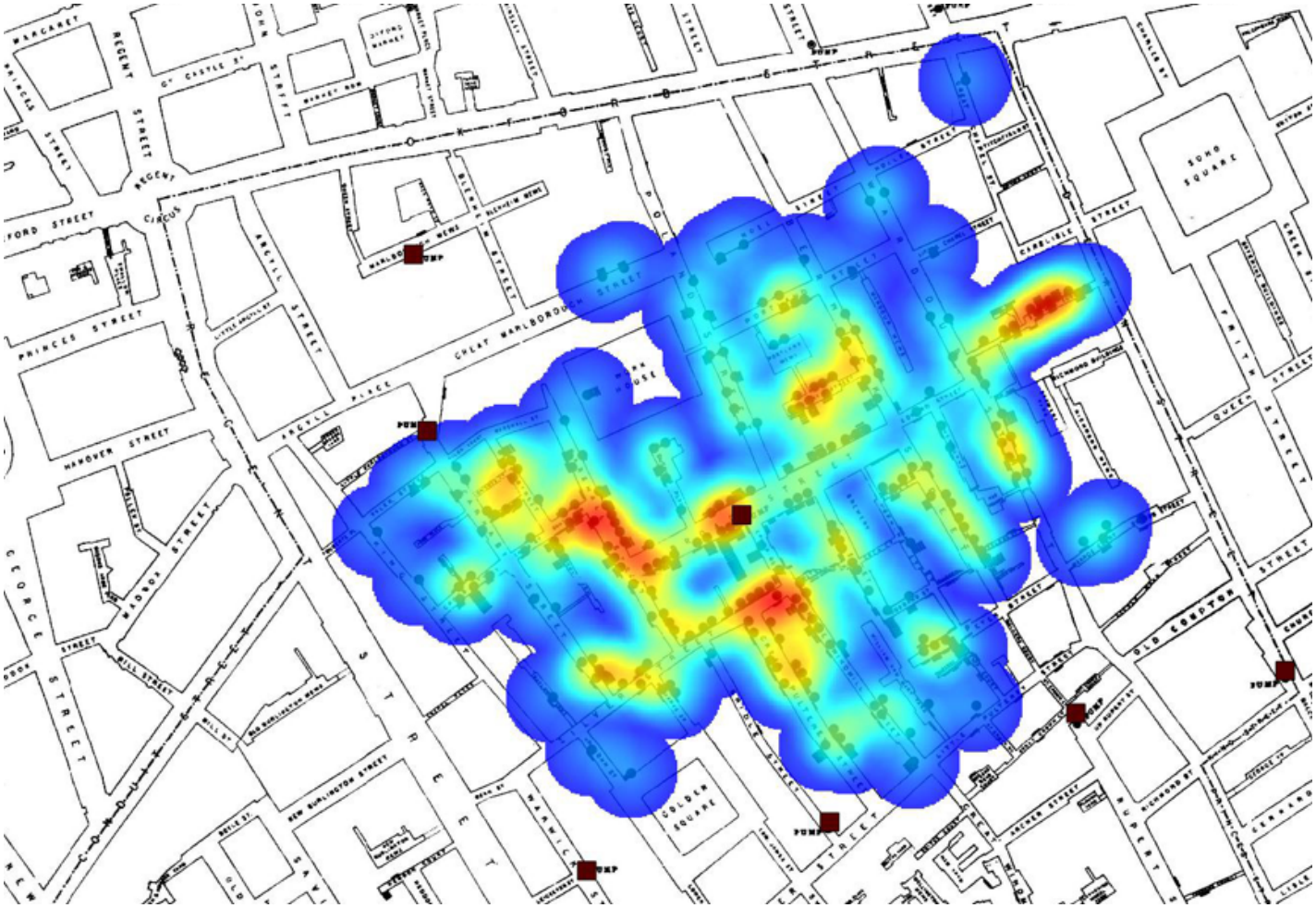
Collection Data for Maps, and Open Street Map

More than 1 million contributors on openstreetmap.org



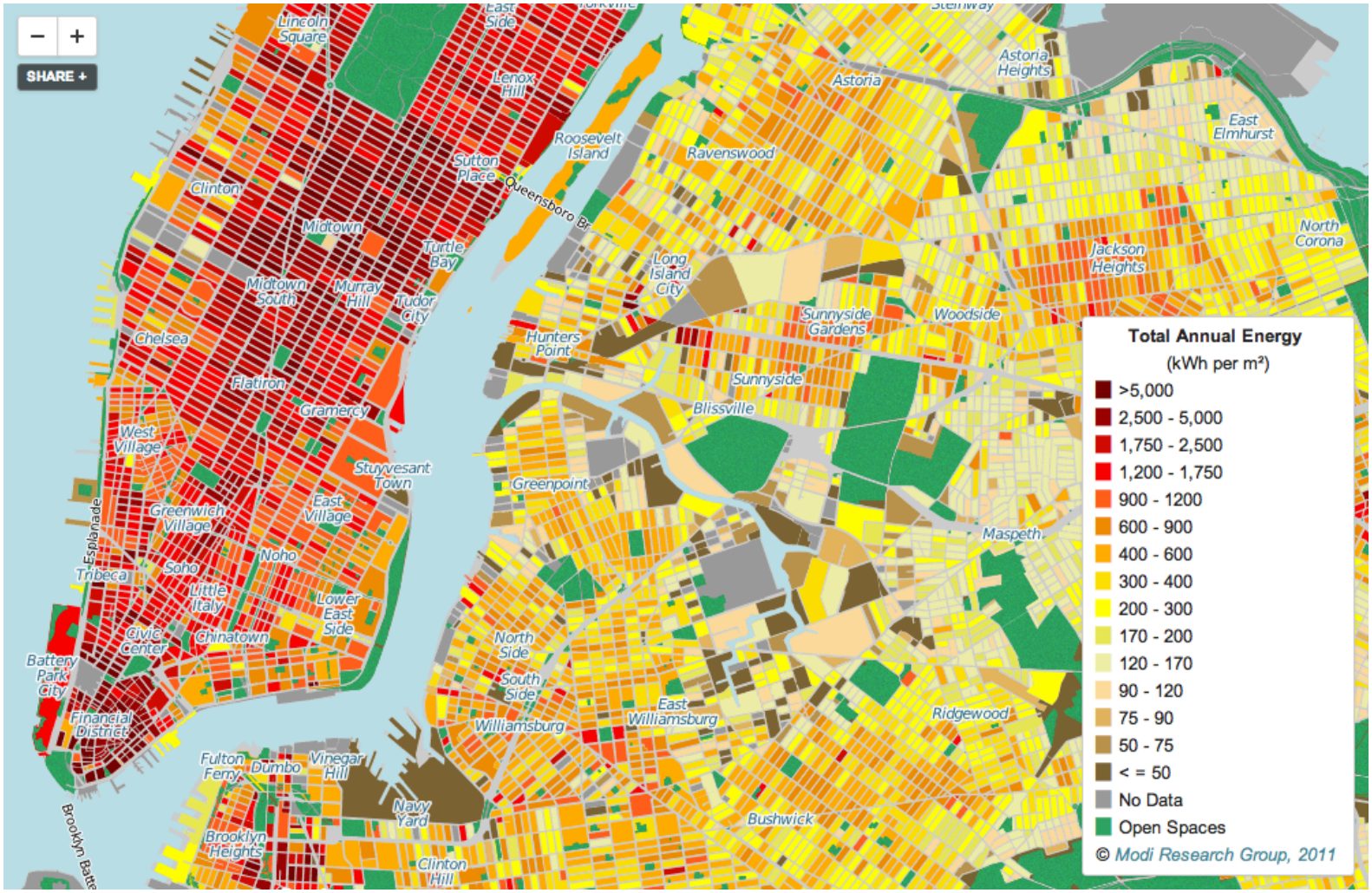
See *Over 100,000 buildings mapped in Guinea where Ebola broke out* mapbox.com and *OpenStreetMap mapping progress of Guéckédou to support doctors in Guinea after Ebola outbreak* datarep.tumblr.com.

Open Data, e.g. Open Street Map



Map by John Snow showing the clusters of cholera cases in the London epidemic of 1854, simonrogers.cartodb.com

Open Data, e.g. Open Street Map



Smart meters could bust marijuana nurseries, baltimoresun.com

Amazon, Netflix (and Sparse Matrices)

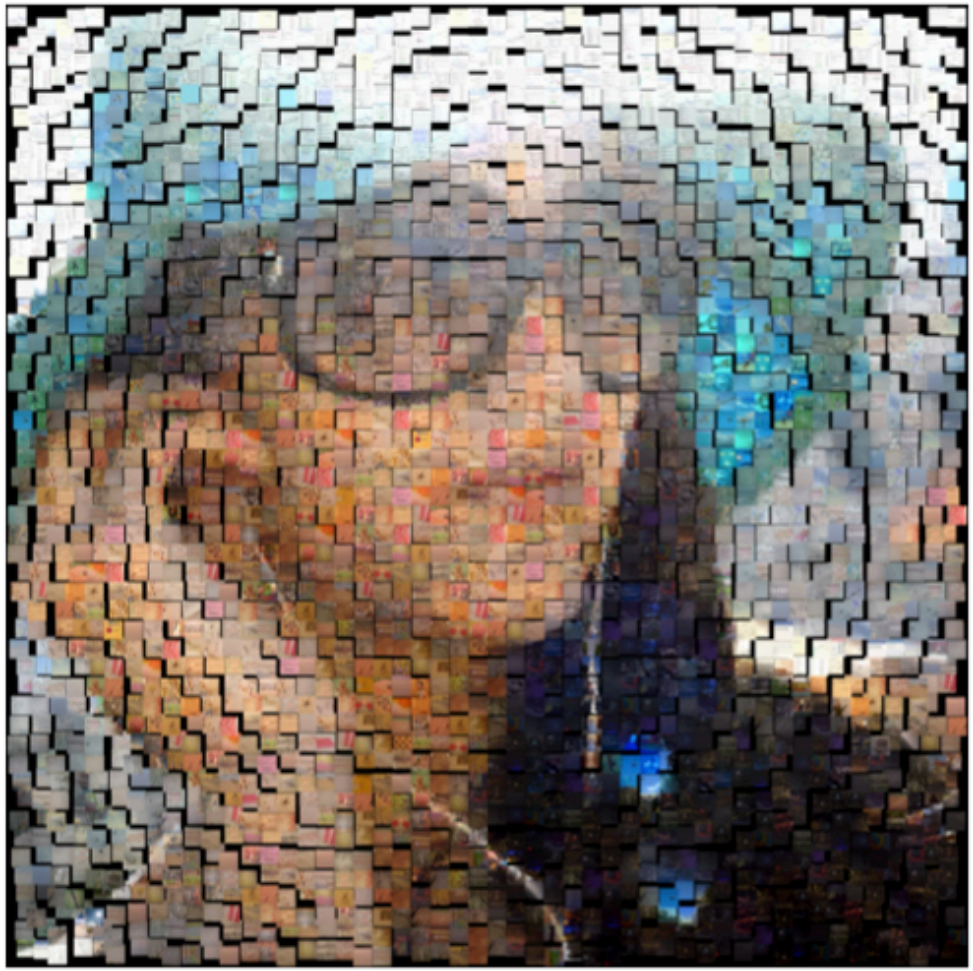
Hastie, 2009 use-R-2009

Example : Netflix problem.

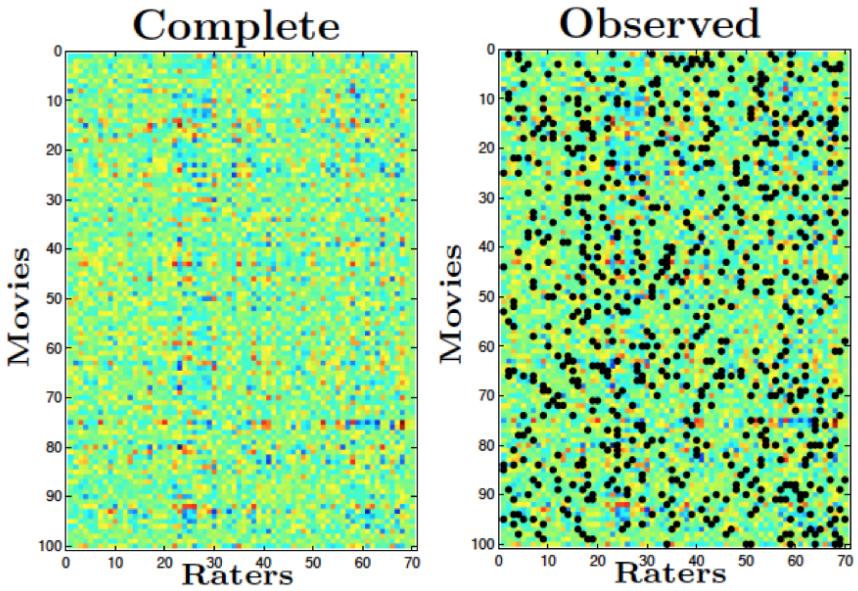
We partially observe a matrix of movie ratings (rows) by a number of raters (columns). The goal is to predict the future ratings of these same individuals for movies they have not yet rated (or seen)

Statistical significance of the Netflix challenge arxiv.org

Amazon, Netflix (and Sparse Matrices)



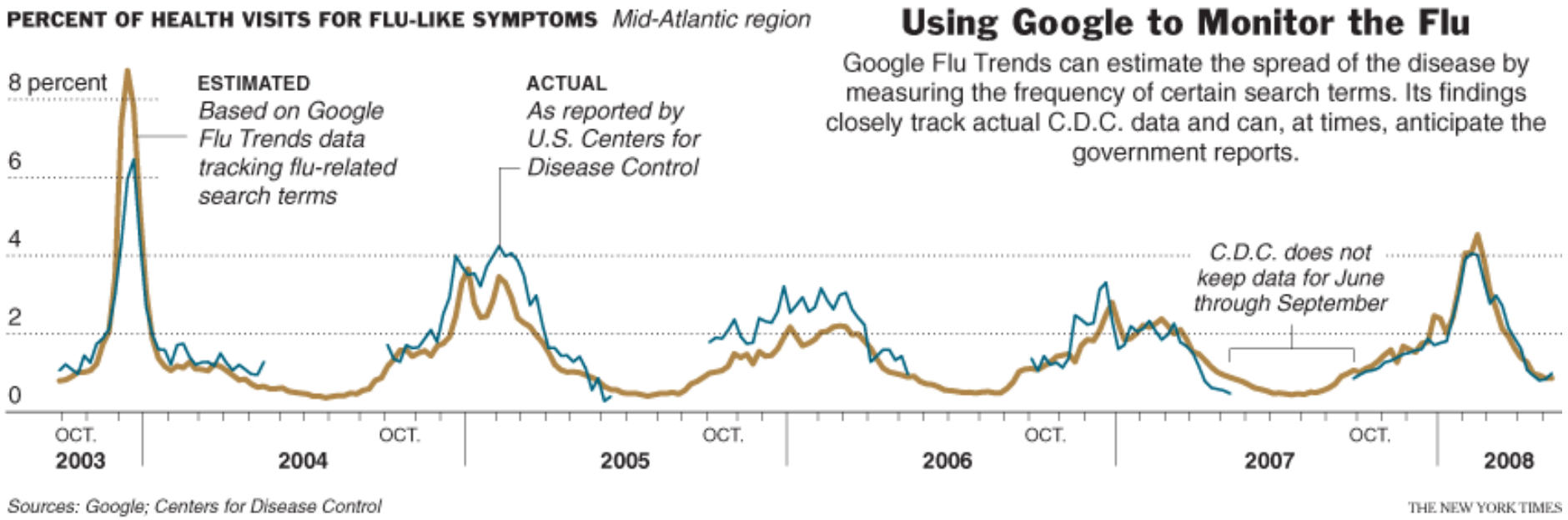
Hastie, 2009 use-R-2009



Google, Wikipedia and the Flu

2012, Google answers 100 billion search queries a month.

Detecting influenza epidemics using search engine query data nature.com

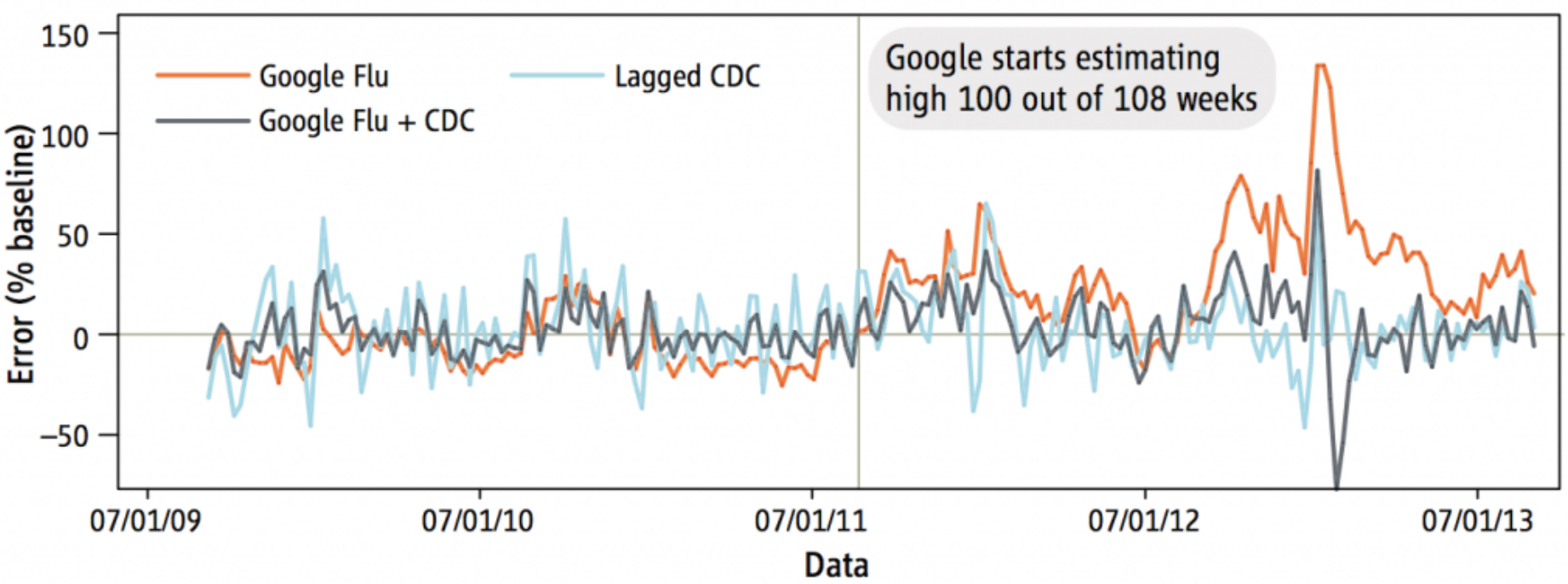


Predicting the present with Google Trends people.ischool.berkeley.edu

Nowcasting with Google Trends inan emerging market ideas.repec.org

Google, Wikipedia and the Flu

Google Flu Trends : The Limits of Big Data bits.blogs.nytimes.com



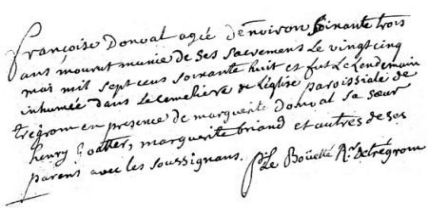
Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the U.S in Near Real-Time ploscompbiol.org

A Big Data Revolution ?

Mobile Provider & Home ISP, IT Office, GPS, Email, iTunes,

Twitter-Facebook, Cameras, Banks Loyalty Cards, etc.






A Big Data Revolution ?

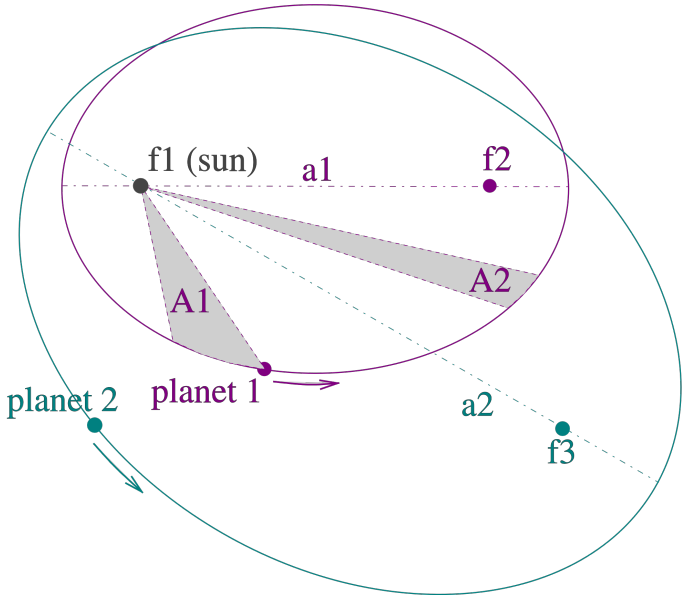
"*A data scientist is a statistician who lives in San Francisco*
Data Science is statistics on a Mac."

Jeremy Jarvis, 2014 twitter.com





Big Data in Sciences : Astronomy



Example : Kepler's laws (1609)

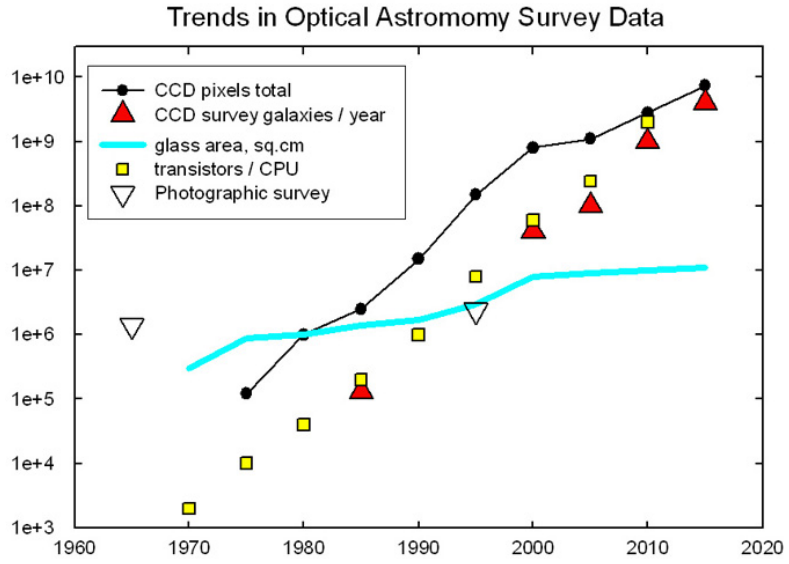
1. The orbit of a planet is an ellipse with the Sun at one of the two foci.

2. A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time (The two shaded sectors A1 and A2 have the same surface area and the time for planet 1 to cover segment A1 is equal to the time to cover segment A2.)

Data = Tycho Brahe's data (687 day intervals)

How Big Data Is Changing Astronomy (Again)
theatlantic.com

Big Data in Sciences : Astronomy

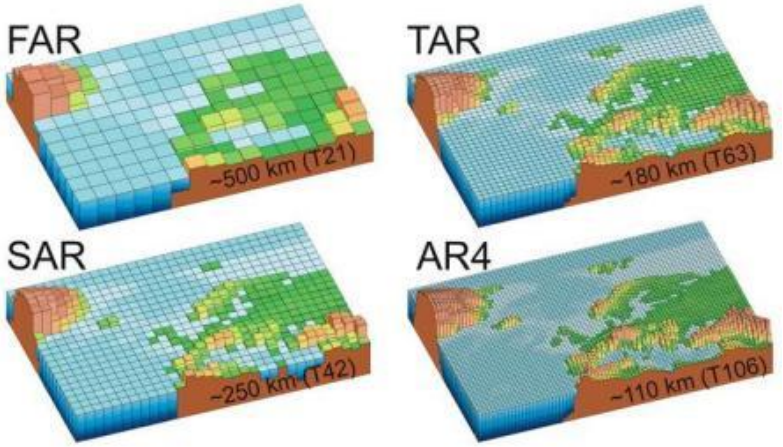


More Than Meets the Eye : How the CCD Transformed Science wired.com

"There are two reasons that astronomy is experiencing this accelerating explosion of data. First, we are getting very good at building telescopes that can image enormous portions of the sky. Second, the sensitivity of our detectors is subject to the exponential force of Moore's Law."

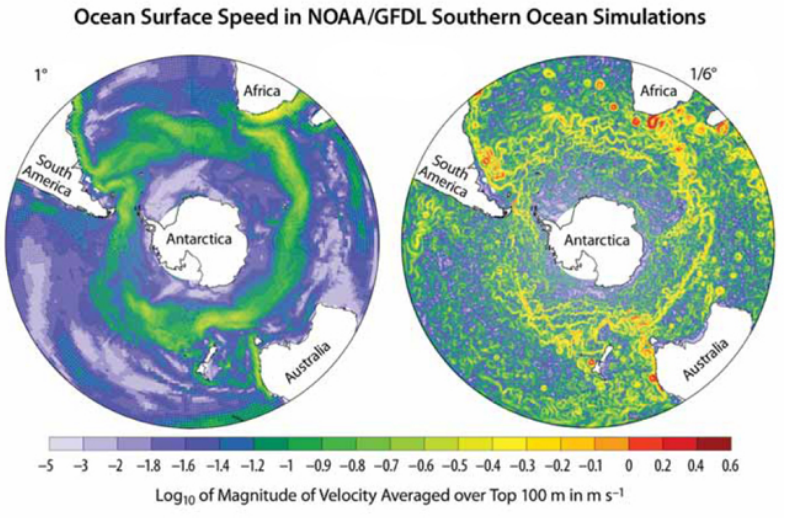
How Big Data Is Changing Astronomy (Again) theatlantic.com

Big Data in Sciences : Climate



→ Improvement of Climate Model resolution over the four IPCC reports clivar.org

Much more noisy data, need filtering techniques



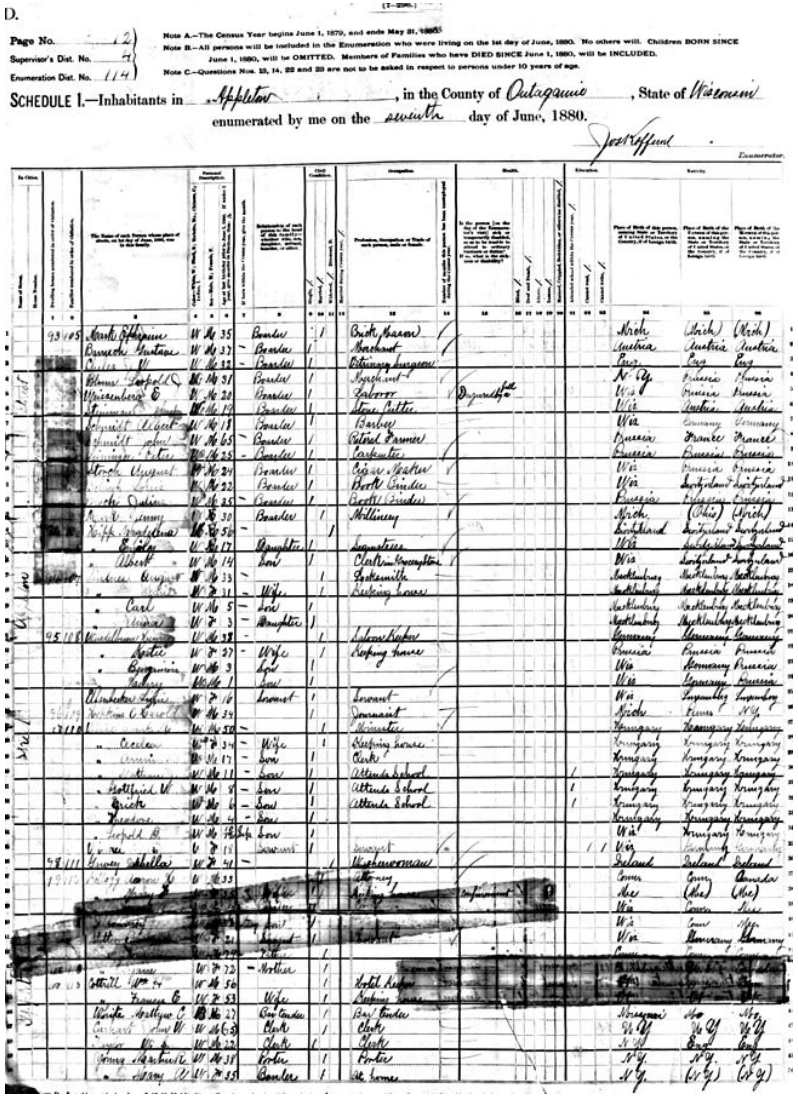
Big Data in Sciences : Genomic

Human Genomes and Big Data Challenges
osehra.org

"*Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze*"

The DNA Data Deluge spectrum.ieee.org

Big Data in Sciences : Demography


D.

Page No. 2
Supervisor's Dist. No. 7
Enumeration Dist. No. 114

Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.
Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880, will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be INCLUDED.
Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

SCHEDULE 1.—Inhabitants in Appleton, in the County of Outagamie, State of Wisconsin, enumerated by me on the seventh day of June, 1880.

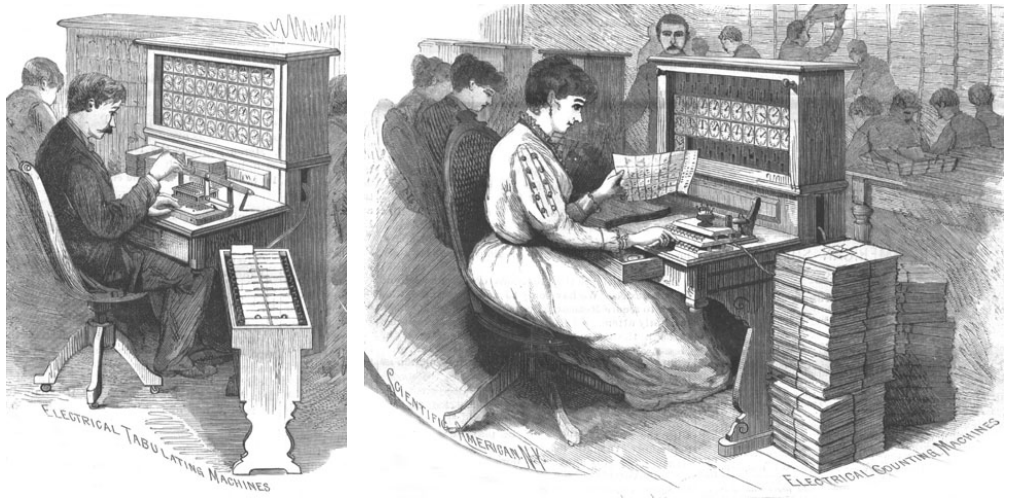
Enumerator.

US Census, 1880

50,189,209 persons censusrecords.com

8 years to tabulate

US Census, 1880



Hollerith Tabulating Machine (ex. IBM)

7 weeks to tabulate

Big Data History

More details on the history of Big Data :

A Very Short History Of Big Data forbes.com

A Very Short History Of Data Science forbes.com

Big Data and the History of Information Storage winshuttle.com

A brief history of big data, the Noam Chomsky way cnbc.com

Big Data and Actuarial Sciences

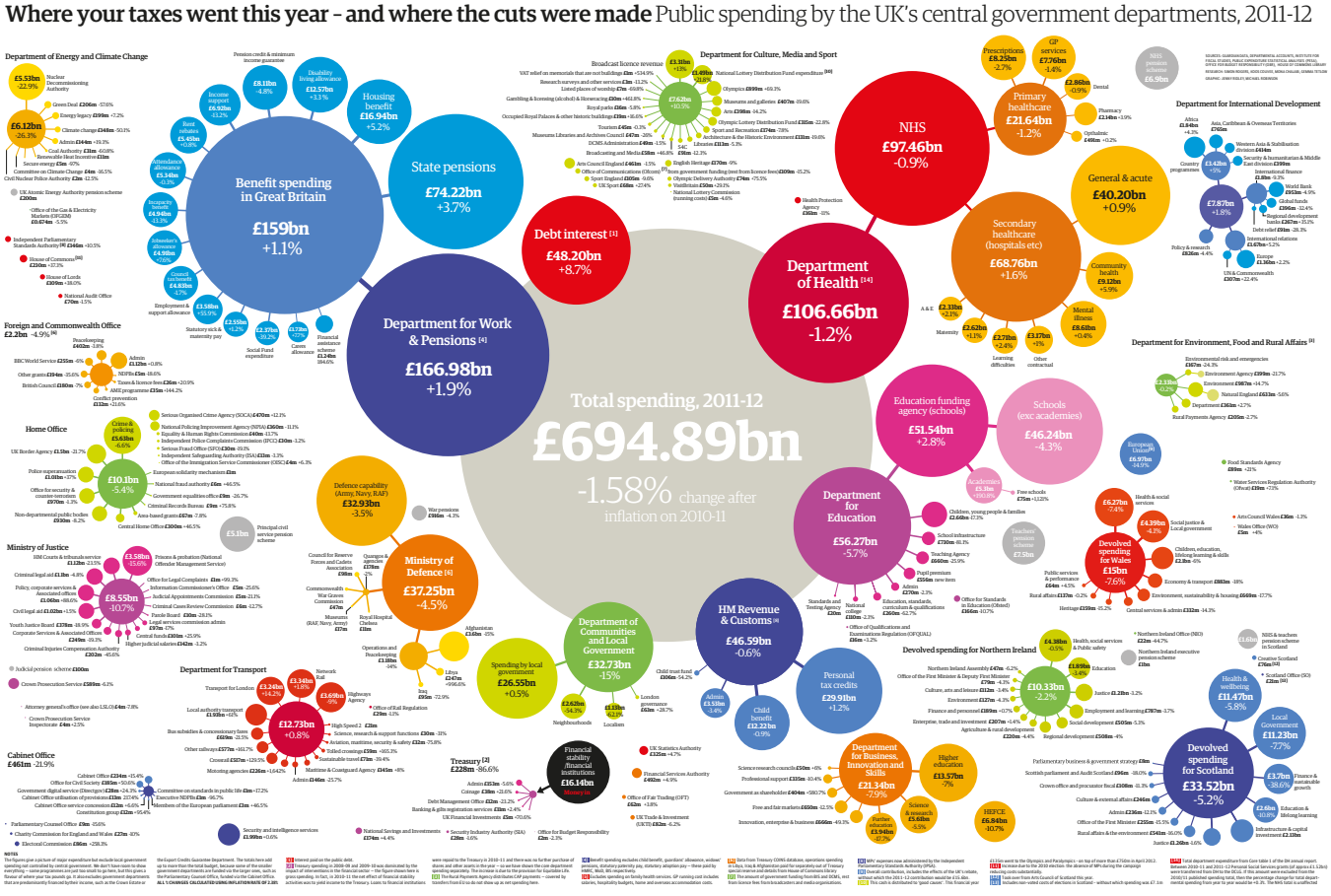
Pay-As-You-Drive, and Real-Time Pricing Models

Big Data Is My Copilot, business.time.com

Chaire *Stratégie Digitale et Big Data* hec.fr

Going Further

Many related topics, such as data viz'...



Government spending by department, 2011-12 theguardian.com/datablog/

Going Further

... web/data scraping, open data, etc.

Defining Big Data

"*The term itself is vague, but it is getting at something that is real*" Jon Kleinberg quoted in newyorker.com

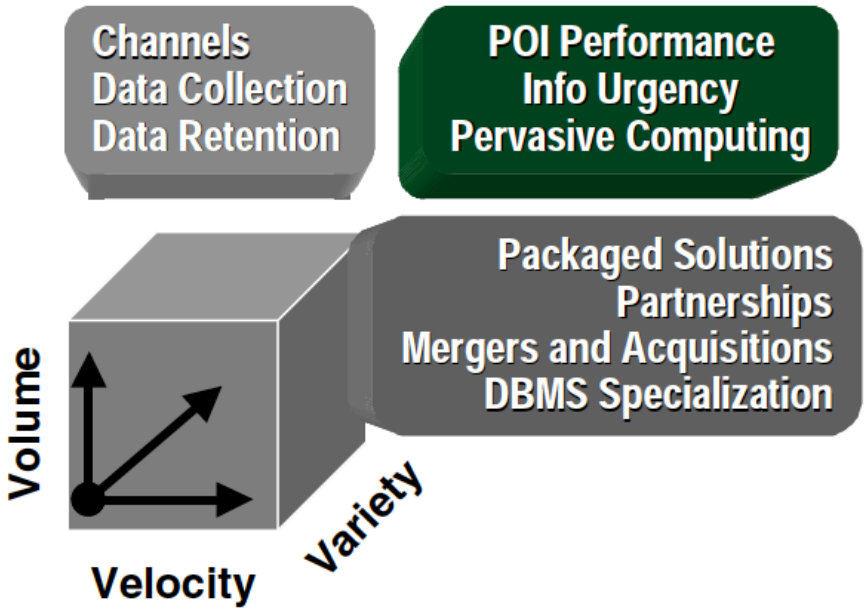
For Viktor Mayer-Schönberger, oii.ox.ac.uk,
"*Big Data is one where* $n = all$ "
(no sample, but the entire background population)
quoted by Tim Harford, ft.com

Defining Big Data

"*Large pools of data that can be brought together and analyzed to discern patterns and make better decisions.*" McKinsey Report, 2011, mckinsey.com

the 3Vs :

- increasing volume (amount of data),
- velocity (speed of data in and out),
- variety (range of data types and sources)



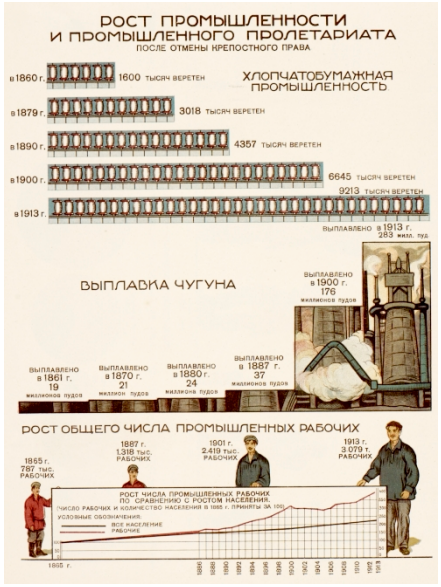
why not got up to 4 or 5Vs ?

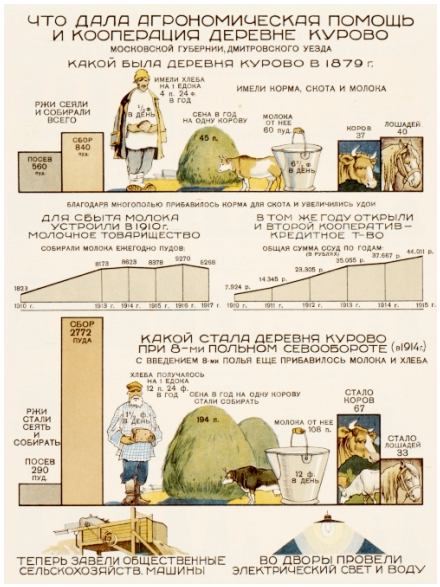
- veracity ? (uncertainty in the data)
- value ? (making big money weforum.org)

3D Data Management : Controlling Data Volume, Velocity and Variety
blogs.gartner.com/doug-laney/

Data (and Statistics)

Data emerged in 1646 as the plural of the Latin *Datum*



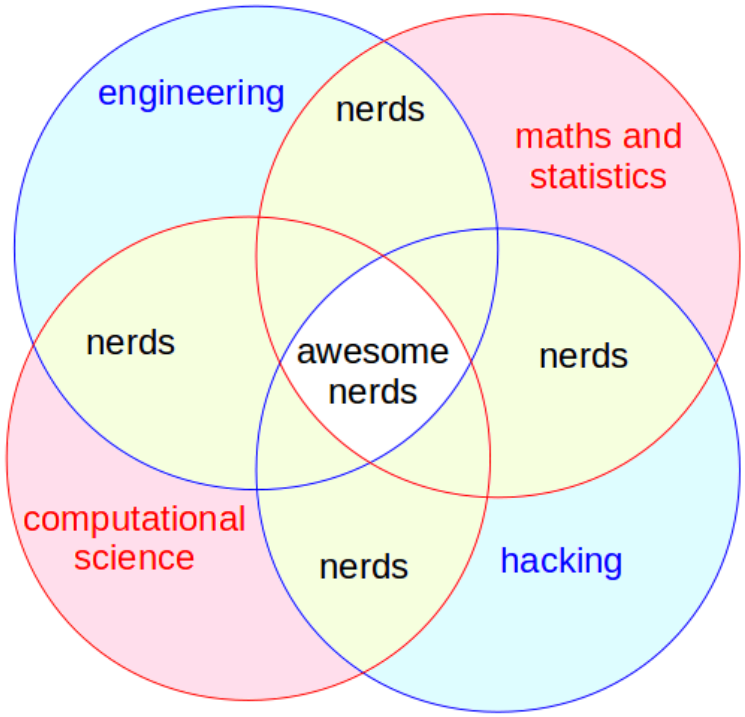


"*Big Data is misnamed in our (academic) world, because data sets have always been big. What is different is that we now have the technology to simply run every scenario.*" Chris Anderson, 2008 wired.com

Statistics, Correlations (and Econometrics)

"*statistics is the* grammar of data science.*It is crucial to making data speak coherently. But it takes statistics to know whether this difference is significant, or just a random fluctuation.* (...) *What differentiates data science from statistics is that data science is a holistic approach. We're increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.*" Mike Loukides, 2010

radar.oreilly.com



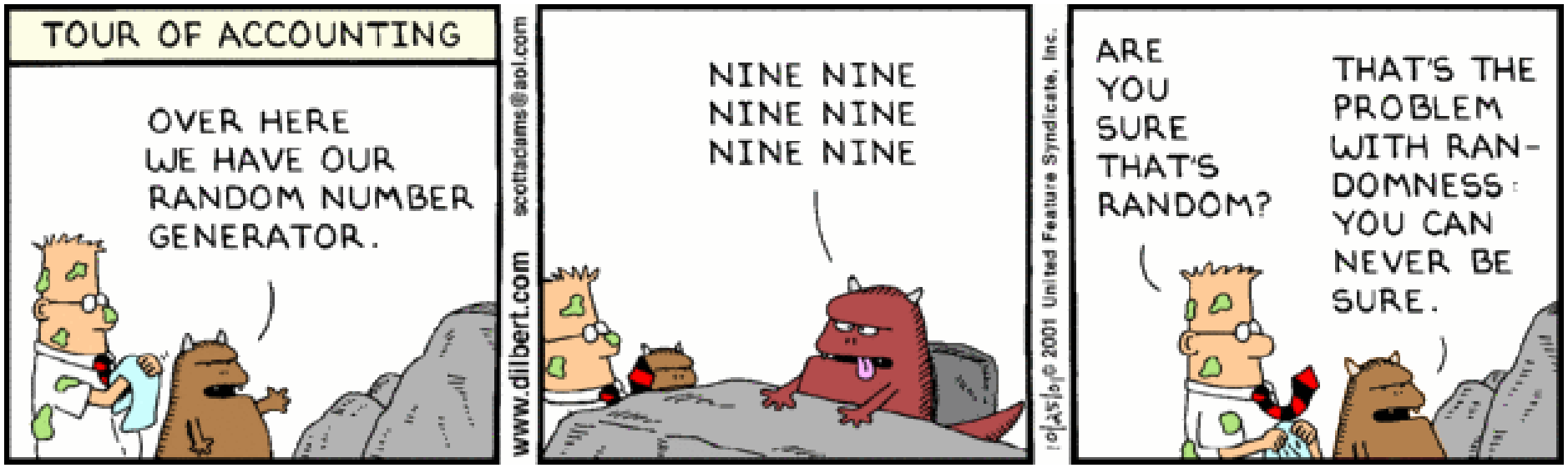
Statistics, Correlations (and Econometrics)

"*In short, the more we learn about biology, the further we find ourselves from a model that can explain it. There is now a better way. Petabytes allow us to say :* Correlation is enough. *We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.*" Narinder Singh, 2013 venturebeat.com

see also *I just ran two million regressions*, Xavier Sala-I-Marin, 1997, jstor.org
against *We Ran One Regression* David Hendry & Hans-Martin Krolzig, 2004, economics.ouls.ox.ac.uk

Statistics, Correlations (and Econometrics)

Epistemological problem : not more certainty, only high likelihood.



"*only an infinite sequence of events* (...) *could contradict a probability estimate*"
de Vries, *On Probable Grounds* filosofie.info

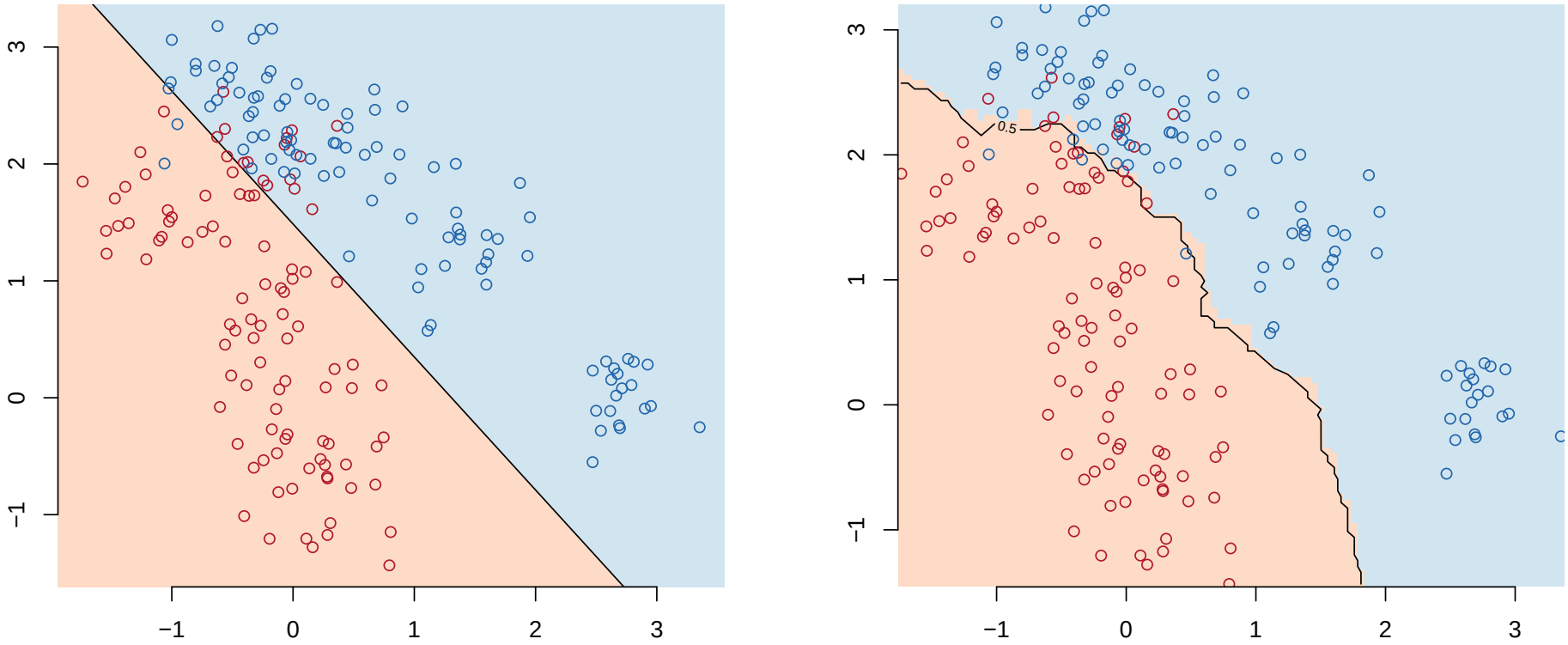
Statistics, Correlations (and Econometrics)

"*Data without a model is just noise. But faced with massive data, this approach to science - hypothesize, model, test - is becoming obsolete.*" Chris Anderson, 2008 wired.com

cf. structural models versus nonparametric statistics, *The Founding of the Econometric Society and Econometrica* jstor.org

$$\begin{cases} C_t = \gamma_{10} + \gamma_{11} P_t + \gamma_{12} P_{t-1} + \beta_{11} W_t + \varepsilon_{1t} \\ I_t = \gamma_{20} + \gamma_{21} P_t + \gamma_{22} P_{t-1} + \beta_{21} K_{t-1} + \varepsilon_{2t} \\ W_t = \gamma_{30} + \gamma_{31} A_t + \beta_{31} X_t + \beta_{32} X_{t-1} + \varepsilon_{3t} \\ X_t = C_t + I_t + G_t \\ P_t = X_t - T_t - W_t \\ K_t = K_{t-1} + I_t \end{cases}$$

Parametric and Non-Parametric Statistics



→ parametric model versus nonparametric model, and machine learning

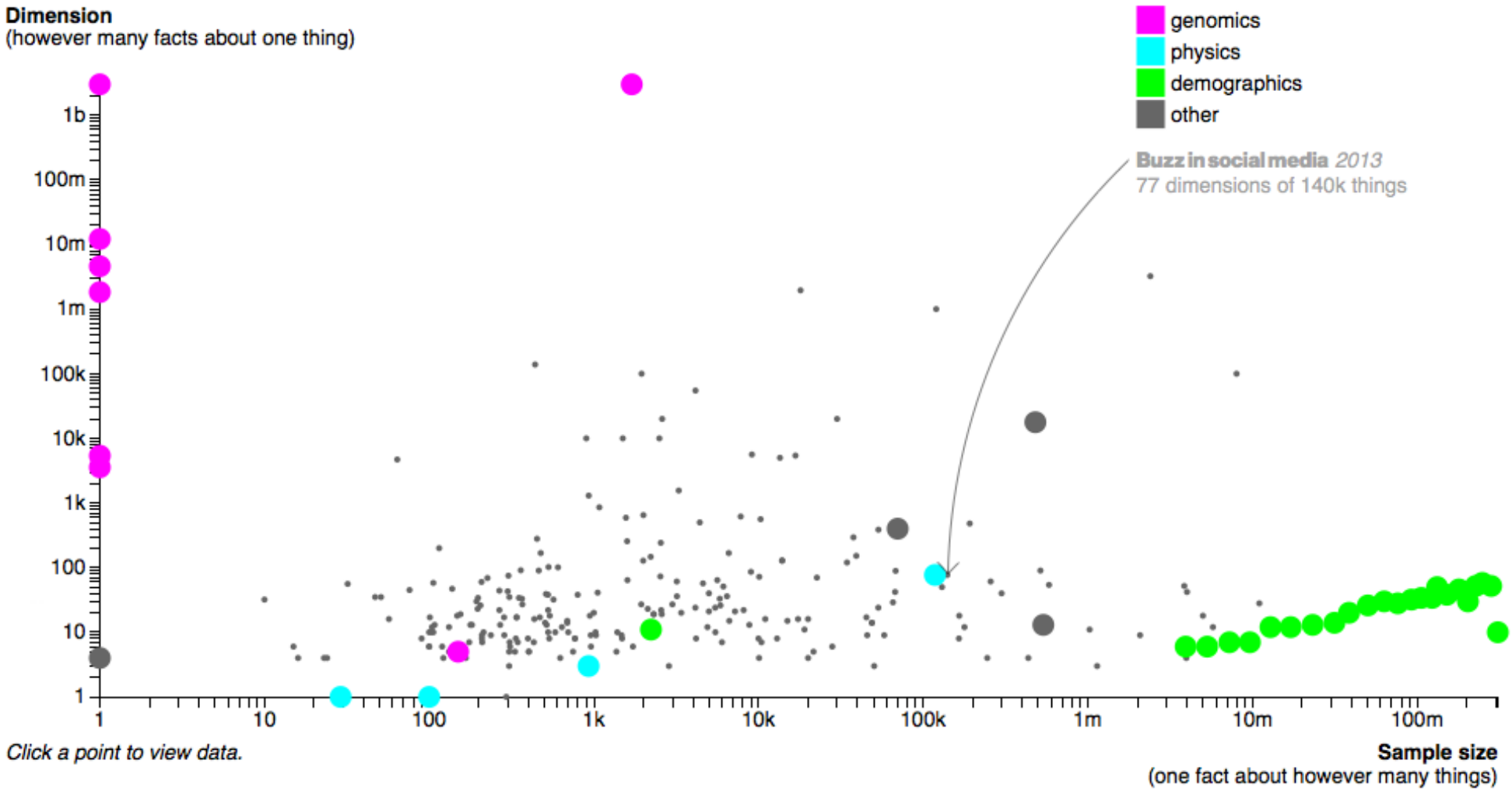
See Breiman's *Statistical Modeling : the Two Cultures* jstor.org

Statistics and Data Mining

"*Data mining, more stuffily* knowledge discovery in databases, *is the art of finding and extracting useful patterns in very large collections of data. It's not quite the same as machine learning, because, while it certainly uses ML techniques, the aim is to directly guide action (praxis !), rather than to develop a technology and theory of induction. In some ways, in fact, it's closer to what statistics calls* exploratory data analysis, *though with certain advantages and limitations that come from having really big data to explore.*"

Cosma Shalizi, 2013 vserver1.cscs.lsa.umich.edu/∼crshalizi/

Volume, How Big is Big Data ?



Interactive Graphic : How Big Is Big Data ?, 2014 businessweek.com

Statistics, Significance and p-values

"*A key issue with applying small-sample statistical inference to large samples is that even minuscule effects can become statistically significant. The increased power leads to a dangerous pitfall as well as to a huge opportunity. The issue is one that statisticians have long been aware of :* the p-value problem.*Chatfield (1995, p. 70) comments,* question is not whether differences are significant (they nearly always are in large samples), but whether they are interesting. Forget statistical significance, what is the practical significance of the results ?"
Mingfeng Lin, Henry Lucas, Jr. et Galit Shmueli , 2010 galitshmueli.com

"*Are there times, I ask, when you just have too much data ? When it gets in the way and confuses things ? He seems taken aback by this line of questioning.* More data is always better, *he says.*" Stephen Baker, *the Numerati.*

Volume, How Big could be Big Data ?

Consider some logistic model $Y_i \sim \mathcal{B}(p_i)$ with $\dfrac{p_i}{1-p_i} = \exp[\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}]$, with

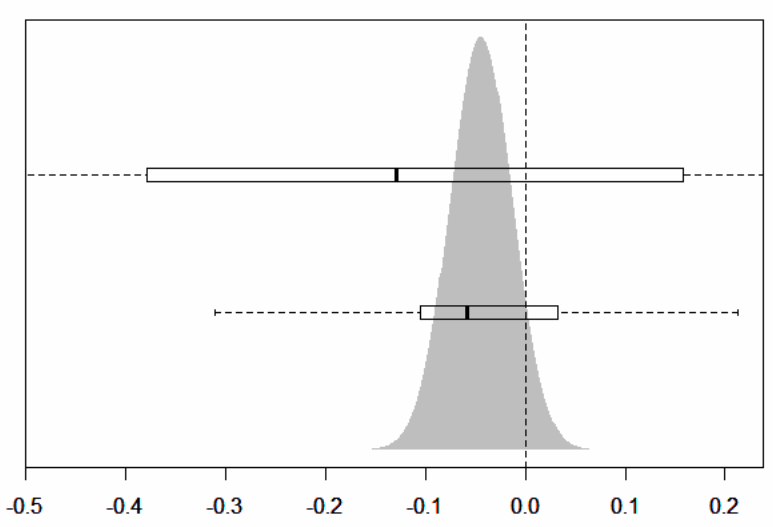
$$\underset{\uparrow\\ n\\ \downarrow}{} \begin{bmatrix} \vdots & \vdots & & \vdots \\ 1 & X_{1,i} & \cdots & X_{k,i} \\ \vdots & \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$
$$\leftarrow \quad k+1 \quad \rightarrow$$

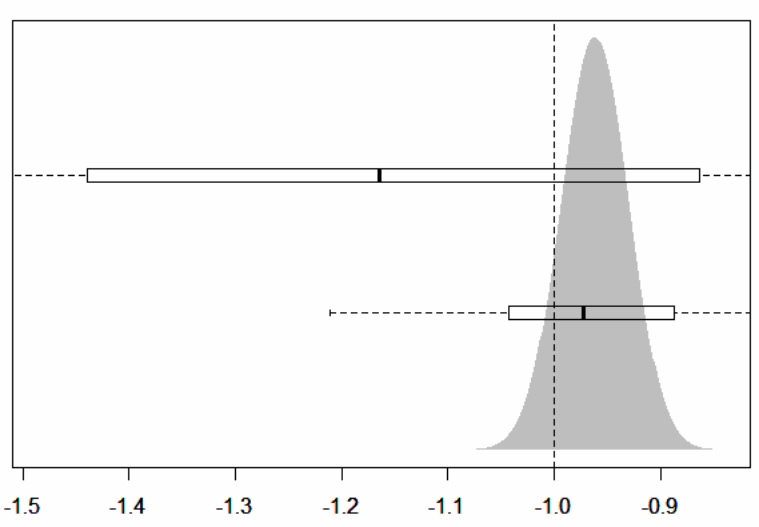
where $\boldsymbol{X}$ is some $n \times (k+1)$ matrix.

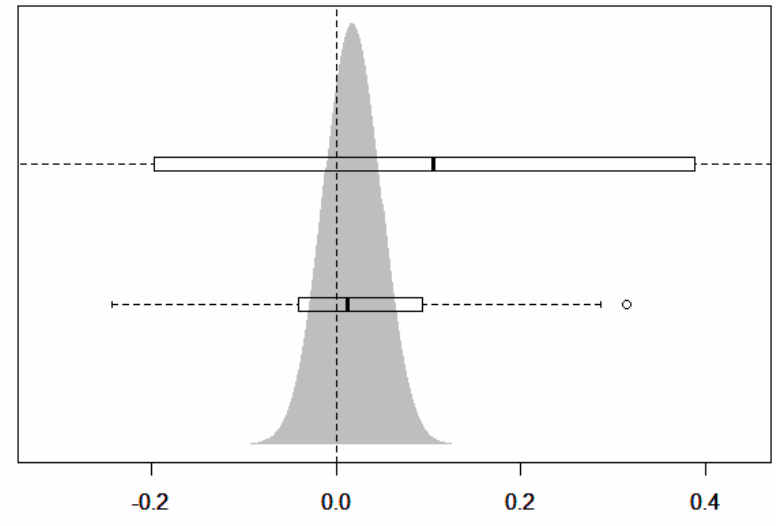
Monte Carlo simulation, with $n = 100,000$ and $k = 100$ (but only two β_j's are not null).

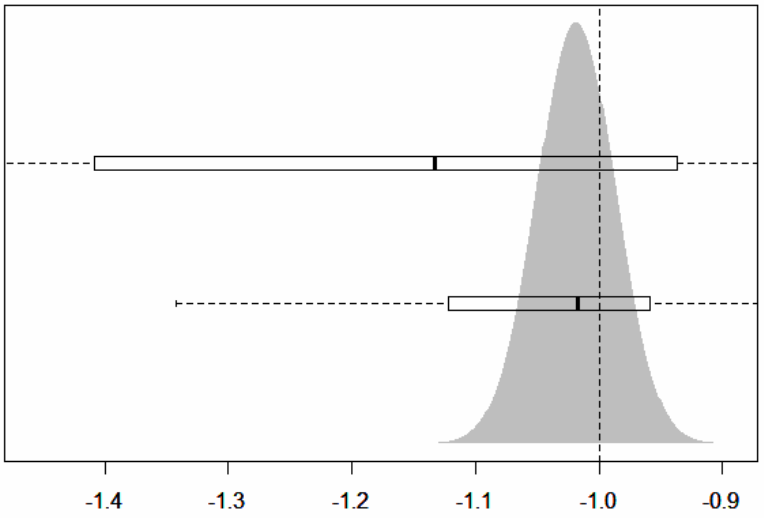
$\rightarrow$ use of subsampling techniques to estimate several models, $n/10$ and $n/100$.

Volume, How Big could be Big Data ?



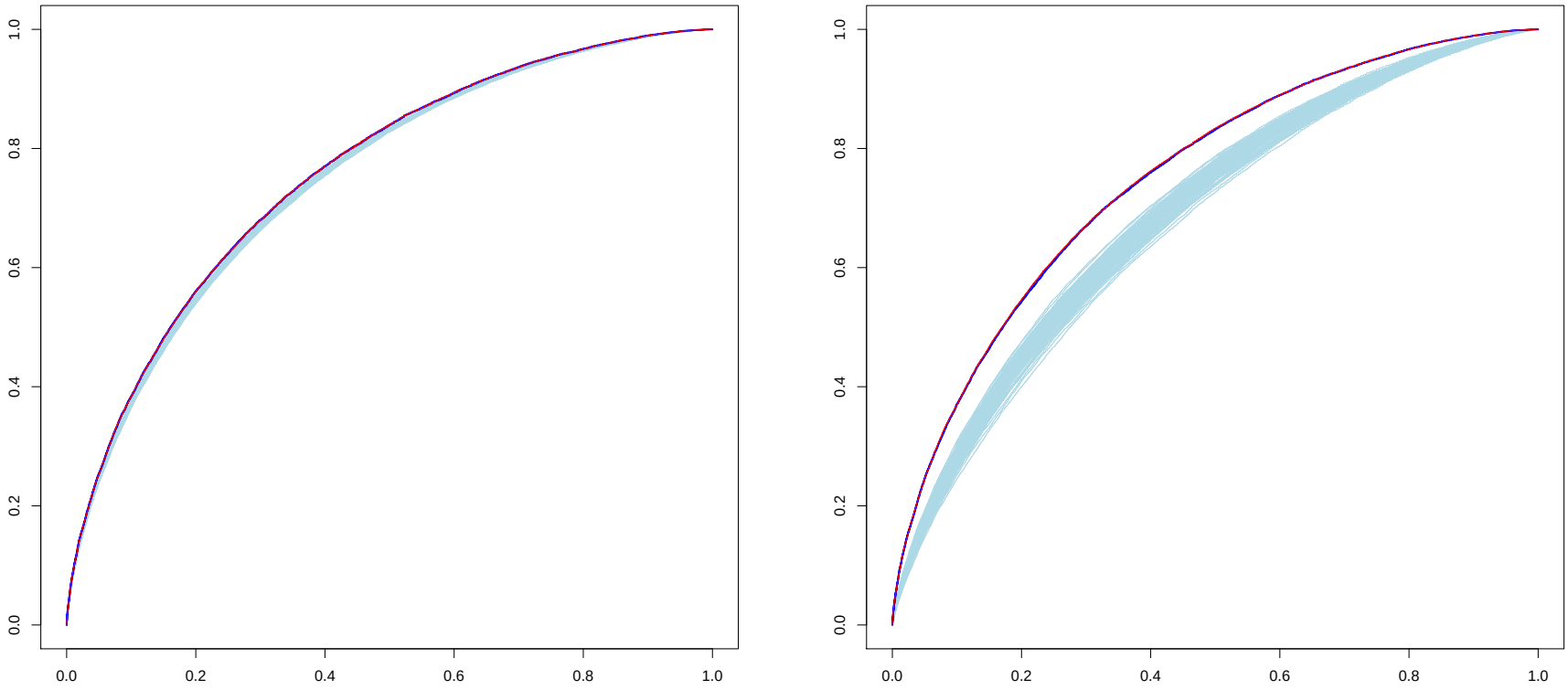






Volume, How Big could be Big Data ?

Look at ROC curves, instead of $\widehat{\beta}_j$'s



— average of the 100 regressions on datasets with $n/10$ and $n/100$ observations

Insurance : Personalization and Customization

Recall basic results on ratemaking and risk pooling.

No risk classification, identical premium

	Insured	Insurer
Loss	$\mathbb{E}[S]$	$S - \mathbb{E}[S]$
Average Loss	$\mathbb{E}[S]$	0
Variance	0	$\text{Var}[S]$

Insurance : Personalization and Customization

Perfect classification, (ultra) personalized premium

	Insured	Insurer
Loss	$\mathbb{E}[S\|\mathbf{\Omega}]$	$S-\mathbb{E}[S\|\mathbf{\Omega}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\mathbf{\Omega}]\Big]$	$\text{Var}\Big[S-\mathbb{E}[S\|\mathbf{\Omega}]\Big]$

$$\text{Var}[S]=\underbrace{\mathbb{E}\Big[\text{Var}[S|\mathbf{\Omega}]\Big]}_{\to\text{insurer}}+\underbrace{\text{Var}\Big[\mathbb{E}[S|\mathbf{\Omega}]\Big]}_{\to\text{insured}}.$$

Insurance : Personalization and Customization

Imperfect classification, personalized premium

	Insured	Insurer
Loss	$\mathbb{E}[S\|\boldsymbol{X}]$	$S-\mathbb{E}[S\|\boldsymbol{X}]$
Average Loss	$\mathbb{E}[S]$	0
Variance	$\text{Var}\Big[\mathbb{E}[S\|\boldsymbol{X}]\Big]$	$\mathbb{E}\Big[\text{Var}[S\|\boldsymbol{X}]\Big]$

$$\begin{aligned}\text{Var}[S] &= \mathbb{E}\Big[\text{Var}[S|\boldsymbol{X}]\Big]+\text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big]\\ &= \underbrace{\underbrace{\mathbb{E}\Big[\text{Var}[S|\boldsymbol{\Omega}]\Big]}_{\text{pooling}}+\underbrace{\mathbb{E}\Big[\text{Var}\Big[\mathbb{E}[S|\boldsymbol{\Omega}]\Big|\boldsymbol{X}\Big]\Big]}_{\text{solidarity}}}_{\rightarrow\text{insurer}}+\underbrace{\text{Var}\Big[\mathbb{E}[S|\boldsymbol{X}]\Big]}_{\rightarrow\text{insured}}.\end{aligned}$$

Velocity : High Frequency Trading

Continuous time model, $dS_t = \mu S_t\,dt + \sigma S_t\,dW_t$

$\rightarrow$ discretized time $dt = \Delta$, compound return, $r_t^{(\Delta)} = \log S_{t+\Delta} - \log S_t$

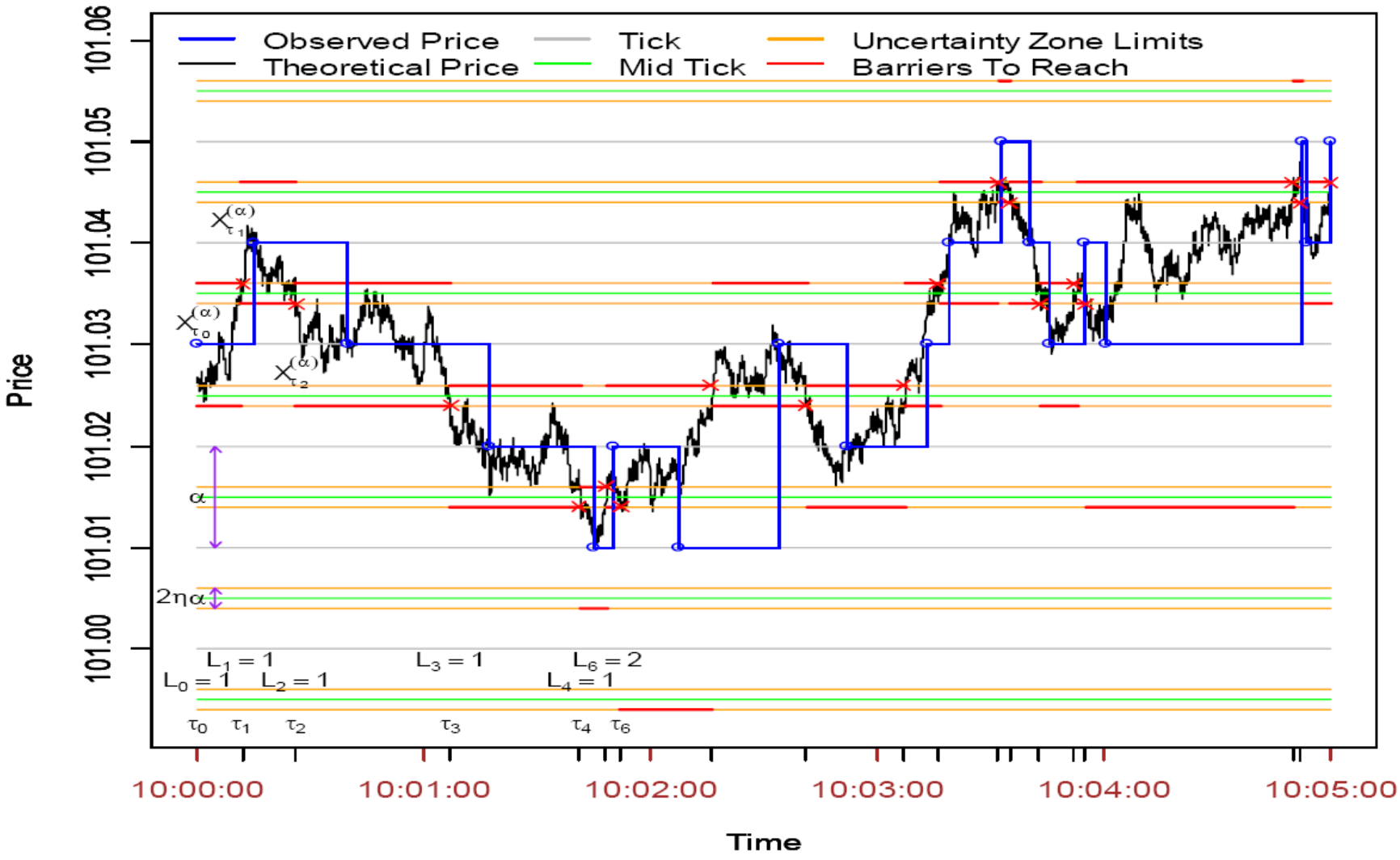
realized volatility over period $[T-h, T]$, $s_t^2 = \sum_{\tau=0}^{h/\Delta} [r_{t-\tau\Delta}^{(\Delta)}]^2$

The Wall Street Code : HFT Whisteblower Haim Bodek on Algorithmic Trading nakedcapitalism.com

High Frequency Trading : Threat or Menace ? blogs.hbr.org

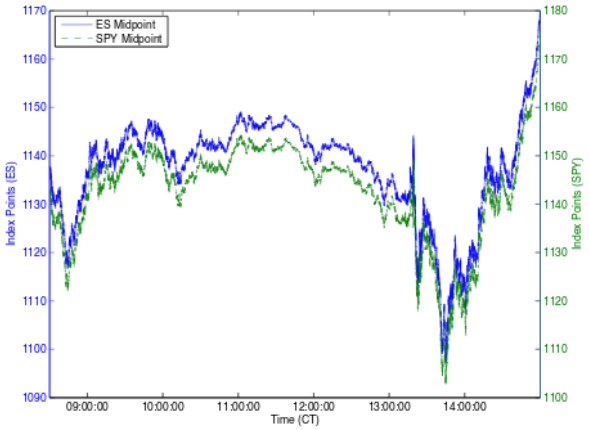
A healthy side effect of High Frequency Trading ? noahpinionblog.blogspot

The problem with high frequency trading blogs.reuters.com

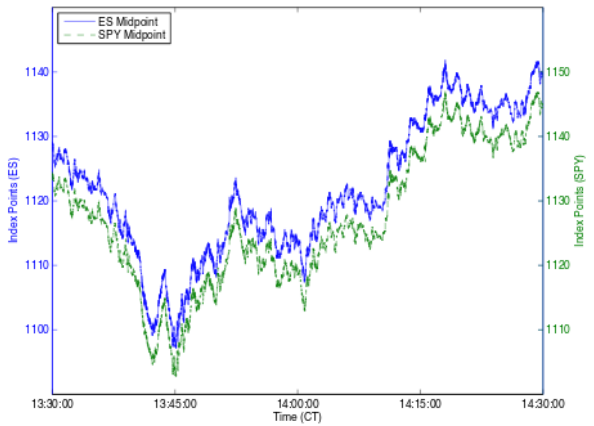


(via Mathieu Rosenbaum)

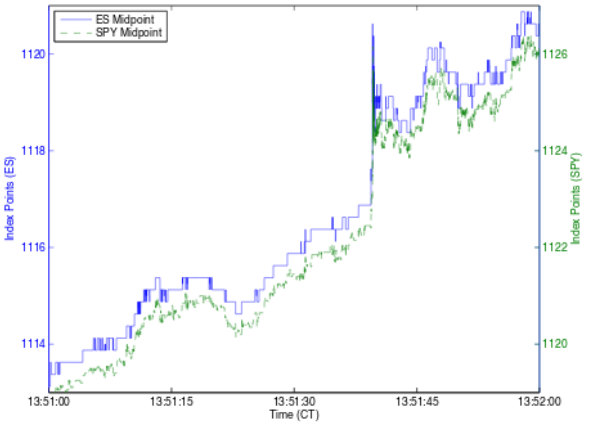
(a) Day



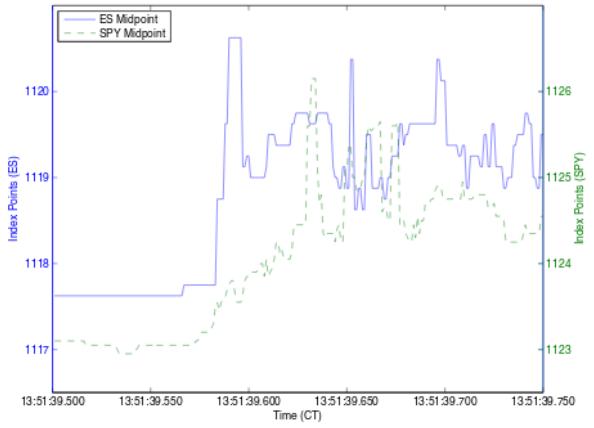
(b) Hour



(c) Minute



(d) 250 Milliseconds

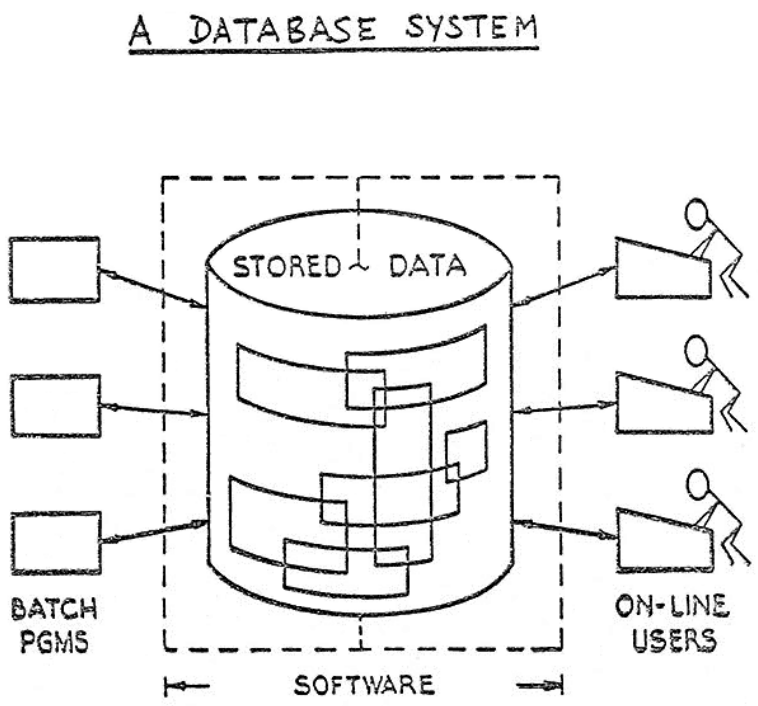


via *The High-Frequency Trading Arms Race : Frequent BatchAuctions as a Market Design Response* faculty.chicagobooth.edu

Variety ?


NEW CUYAMA
Population 562
Ft. above sea level 2150
Established 1951
TOTAL 4663

Variety : No More Datawarehouse ?



A Relational Model of Data for Large Shared Data Banks, by Edgar Codd, 1970, seas.upenn.edu

See NoSQL (e.g. nosql-database.org) that provides a mechanism for storage and retrieval of data, modeled in means other than the tabular relations used in relational databases.
Used for big data and real-time web applications.

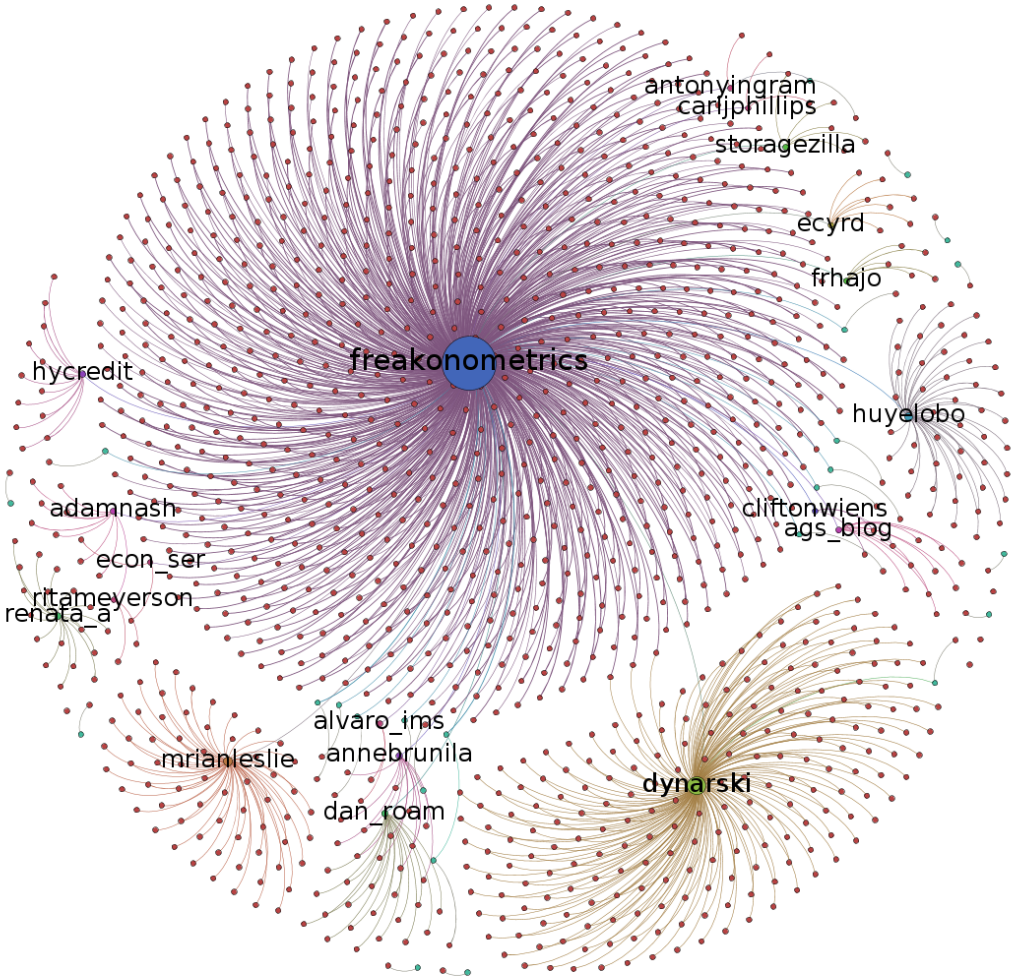
Remark NoSQL stands for "*Not only SQL*"

Variety : Graph Theory in Action

Classically, databases where structured (tables, relational databases)

Nowadays, most of the world's data is unstructured (text, image, video, voice)

RTs on Twitter



Variety : Translating Languages

"*Google can translate languages without actually* knowing *them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German)* Chris Anderson, 2008 wired.com

translate.google.com can translate back and forth between 71 languages

"*I have trouble learning languages, and that's precisely the beauty of machine translation : The most important thing is to be good at math and statistics, and to be able to program. (...) So what the system is basically doing (is) correlating existing translations and learning more or less on its own how to do that with billions and billions of words of text (...) In the end, we compute probabilities of translation*" Franz Och, quoted in spiegel.de

Variety : Text Mining and Sentiment Analysis

Automatically Extracting Dialog Models from Conversation Transcripts ieeexplore.ieee.org

Text Mining Medicine the-scientist.com

Text Mining Gun Deaths Data econometricsbysimulation.com from slate.com (crowdsourced) database

Twitter Mood Predicts the Stock Marker arxiv.org

Easy to use some simple Text Analytics to extract intent from Social Media,

> my car didn't start this morning ? should be late at work
>
> stuck in bed with back pain, again

$\longrightarrow$ possible to extract some personal information...

(need to detect jokes, sarcasm, amiguity, non-personal information, etc.)

A slide on Veracity

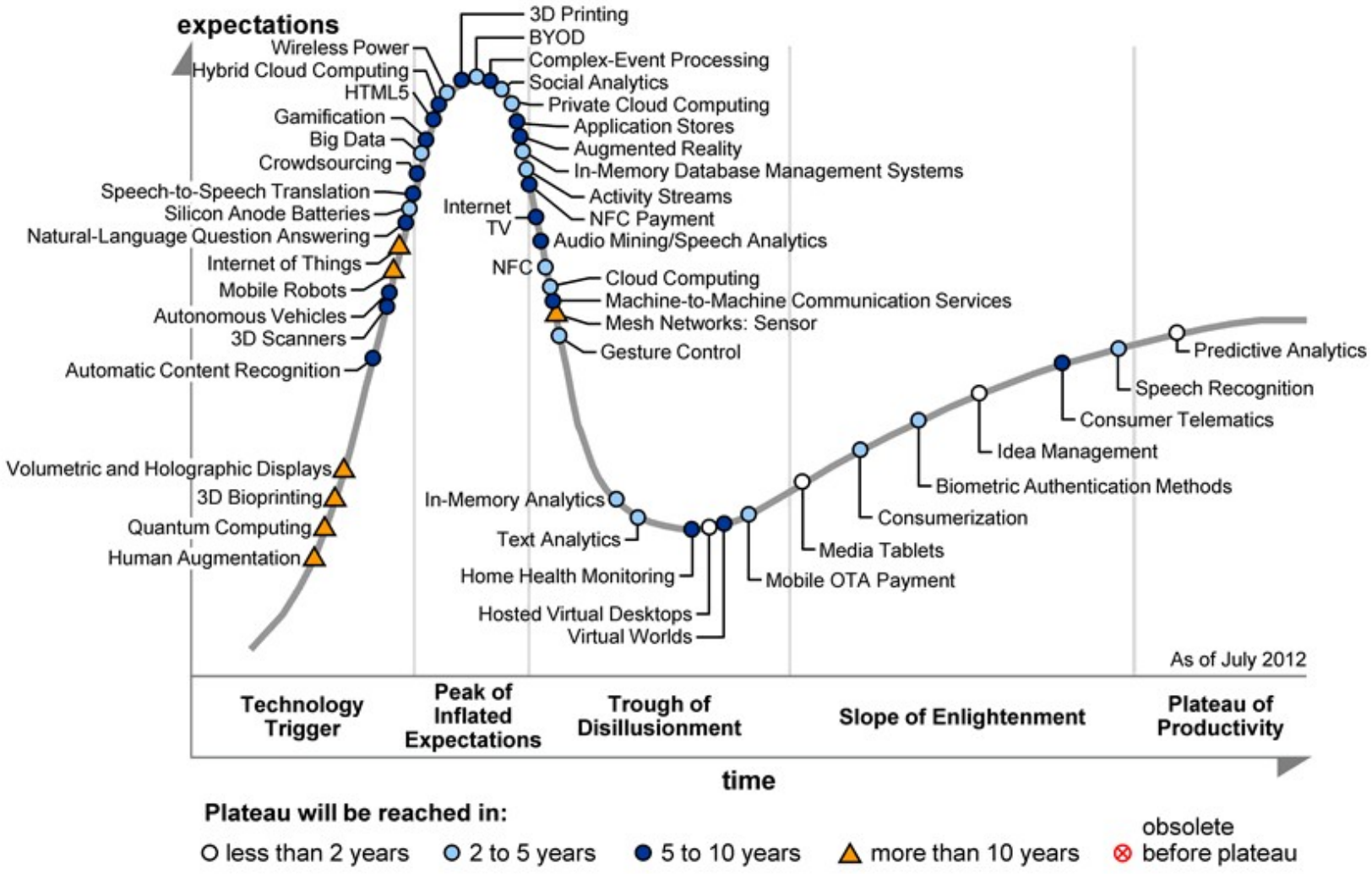
"*Not everything that counts can be counted, and not everything that can be counted counts.*"

"*Since much of the data deluge comes from anonymous and unverified sources, it is necessary to establish and flag the quality of the data before it is included in any ensemble.*" Arup Dasgupta, geospatialworld.net

How do you feel about online surveys ?

Conclusion, Big Data, the new Gold ?

Conclusion, Big Data, the new Gold ?



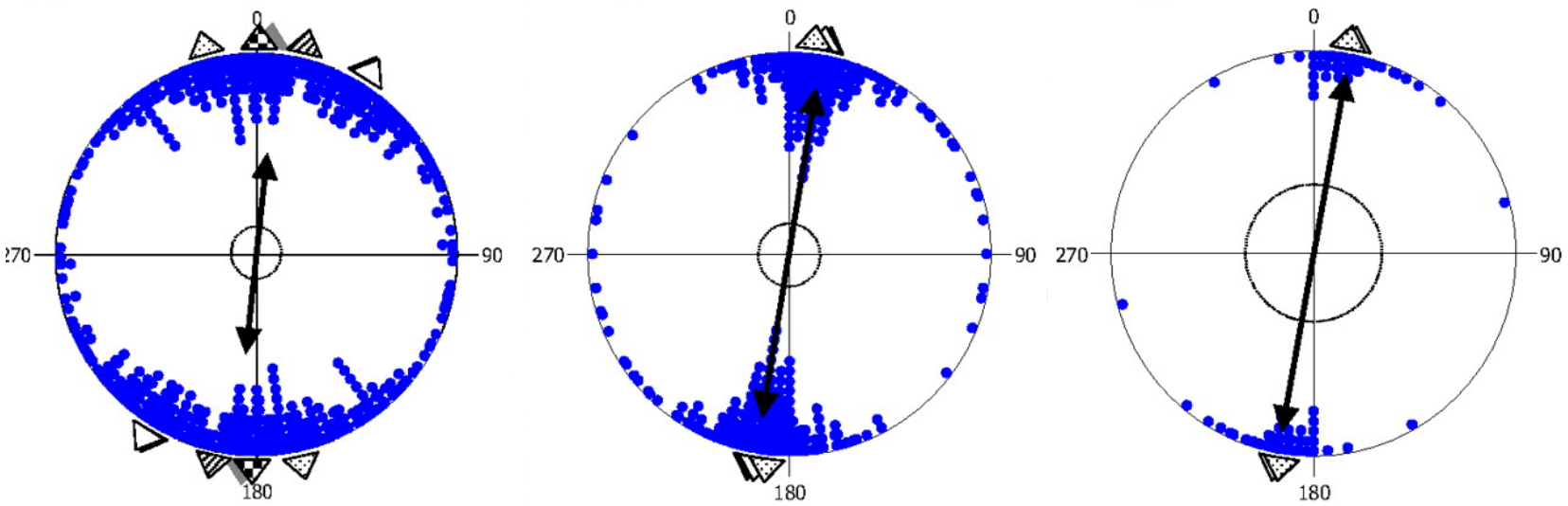
Hype Cycles, 2012 gartner.com

Conclusion, Big Data, the new Gold ?

"*Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media*" ft.com

huge data means increasing risk of *fake discoveries*, a theory-free analysis of mere correlation is inevitably fragile

"*many beats of straw look like needles*" Trevor Hastie, quoted in nytimes.com



see *Magnetic alignment in grazing and resting cattle and deer* pnas.org

Conclusion, Big Data, the new Gold ?

Big data : The next frontier for innovation, competition, and productivity, 2011 mckinsey.com/insights

Business Insider : Enterprises Aren't Spending Wildly on Big Data But Don't Know If It's Worth It Yet, 2012 businessinsider.com

Five myths about big data, 2013 washingtonpost.com

What Data Can't Do, 2013 nytimes.com

Big Data Boosts Customer Loyalty. No, Really, 2013 forbes.com

Conclusion, Big Data, the new Gold ?

... and some more recent posts and articles (March and April)

Future big data analysts will know everything you did today venturebeat.com

The backlash against big data economist.com

Big data and open data : what's what and why does it matter ? theguardian.com

Give us back our statistical datawashingtonpost.com

The rise of big data brings tremendous possibilities and frightening perils washingtonpost.com

Finding a great data scientist can feel like searching for Princess Peach. Shes always in another castle. hnews360.com

The Promise and Peril of Big Data aspeninstitute.org

Eight (No, Nine !) Problems With Big Data nytimes.com

What's Up With Big Data Ethics ? forbes.com

Conclusion, Big Data, the new Gold ?

Is data privacy an out of date concept ? nakedsecurity.sophos.com

Google Flu Trends : The Limits of Big Data bits.blogs.nytimes.com

Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the U.S in Near Real-Time ploscompbiol.org

Big data : are we making a big mistake ? ft.com

Data the buzzword vs. data the actual thing noahpinionblog.blogspot.com

Simplifying Data Analysis and Making Sense of Big Data scientificcomputing.com

Google may be a master at data wrangling, but one of its products has been making bogus data-driven predictions newscientist.com

Big Data Doesn't Have to Mean Big Brother recode.net

How Big Is Big Data ? businessweek.com

Conclusion, Big Data, the new Gold ?

We are moving from an era of private data and public analyses to one of public data and private analyses andrewgelman.com

Can Big Data Help Us Predict The Next Big (Snow) Storm ? business2community.com

You Know Who Else Collected Metadata ? The Stasi. propublica.org

Data is data, or are they ? stancarey.wordpress.com