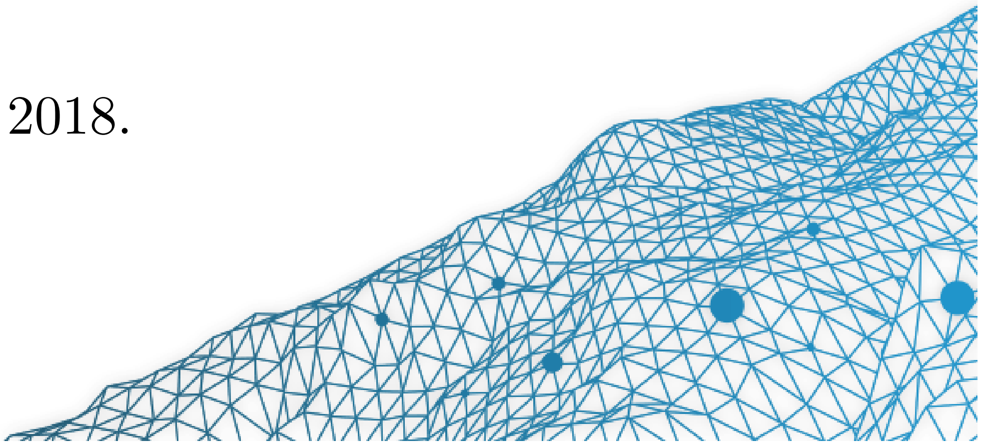
Big Data and Artificial Intelligence

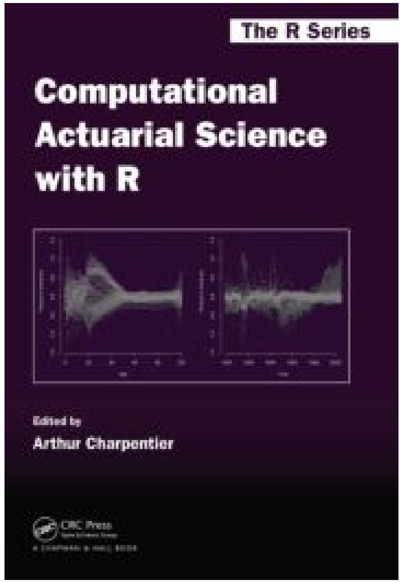
A. Charpentier (Université du Québec à Montréal)

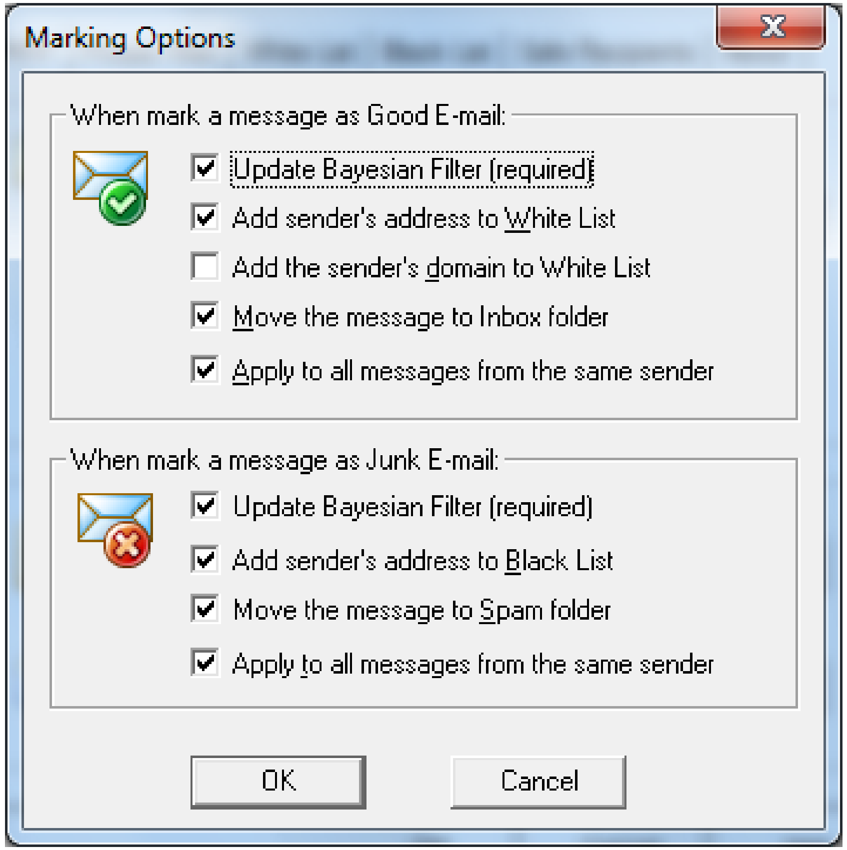
Ecole Supérieure de Commerce de Paris, 2018.



A. Charpentier (Université du Québec à Montréal)

Professor Mathematics Department, Université du Québec à Montréal
previously Econ. Dept, Université de Rennes & ENSAE Paristech
actuary in Hong Kong, IT & Stats FFA
director Data Science for Actuaries Program, Institute of Actuaries
PhD in Statistics (KU Leuven), Fellow of the Institute of Actuaries
MSc in Financial Mathematics (Paris Dauphine) & ENSAE
Research Chair :
ACTINFO (valorisation et nouveaux usages actuariels de l'information)
Editor of the freakonometrics.hypotheses.org's blog
Editor of Computational Actuarial Science, CRC
Author of Mathématiques de l'Assurance Non-Vie (2 vol.), Economica





Karmali (2017, Spam Classifier in Python from scratch)

Support The Guardian

Subscribe Search jobs Sign in Search

International edition

The Guardian

News Opinion Sport Culture Lifestyle More

World UK Science Cities Global development Football Tech Business Environment Obituaries

Spam filter
Google

Tired of texting? Google tests robot to chat with friends for you

With its new Reply system the firm is taking the art of conversation one step forwards - or should that be backwards?

HAL 90210
Wed 14 Feb 2018
10.38 GMT

2216

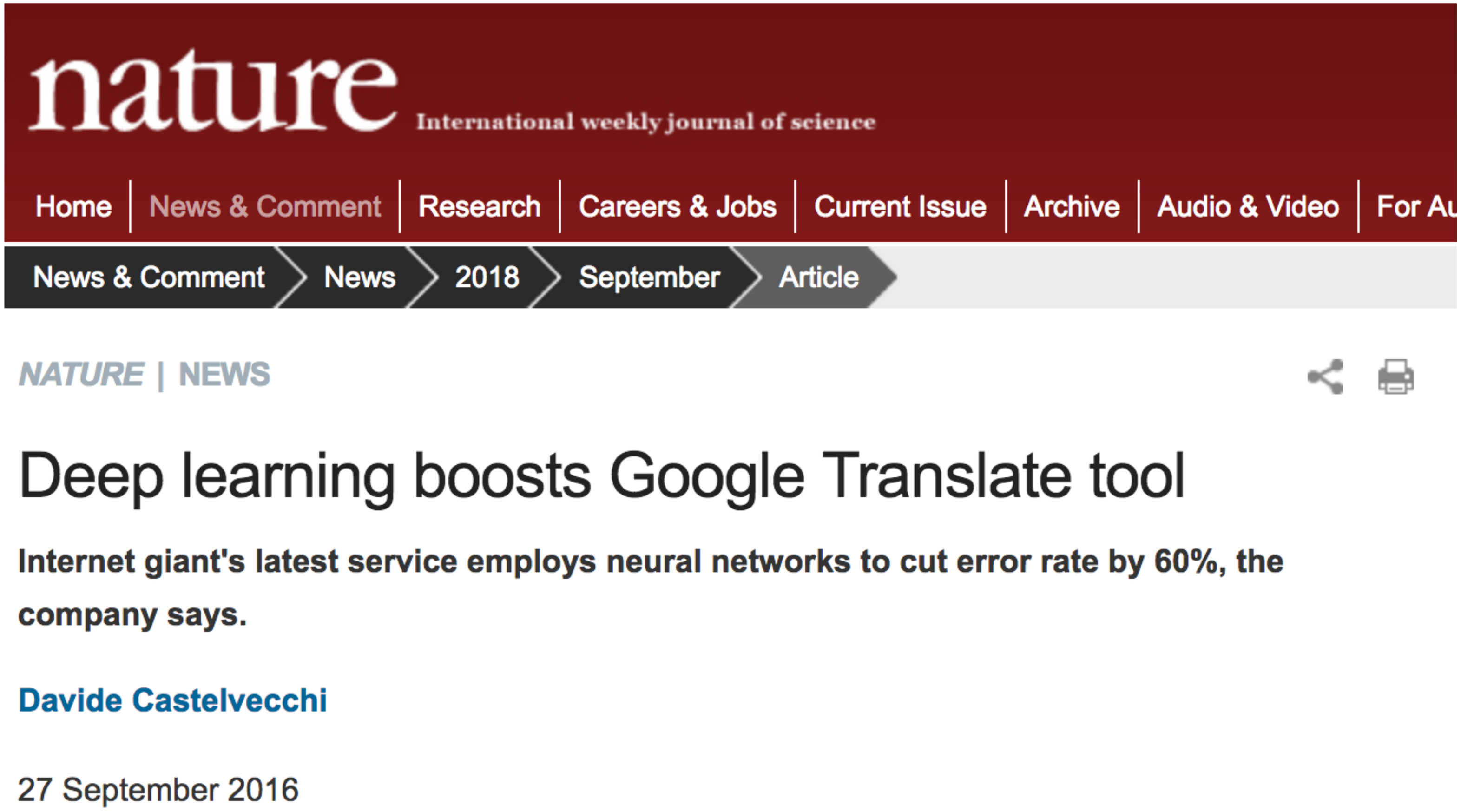
This article is over 8 months old

We aren't the robots ... but is one of these girls talking to one? Photograph: Richard Baker/Corbis via Getty Images

most viewed

- Trump and Obama trade blows as midterm elections loom
- Skin cancer deaths soar for men in developed nations
- Nurse who treated Pittsburgh shooter: 'I'm sure he had no idea I was Jewish'
- Xi Jinping promises to open Chinese markets in trade war joust with Trump
- Tower of London lights up in stunning Armistice memorial

Spam filter, The Guardian's tech diary column (2018, Tired of texting? Google tests robot to chat with friends for you)



Castelvecchi (2016, Deep learning boosts Google Translate tool) and Korbut (2017, Machine Learning Translation and the Google Translate Algorithm)

Support The Guardian | Subscribe | Search jobs | Sign in | Search | International edition

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

World UK Science Cities Global development Football **Tech** Business Environment Obituaries

Artificial intelligence (AI)

World's best Go player flummoxed by Google's 'godlike' AlphaGo AI

Ke Jie, who once boasted he would never be beaten by a computer at the ancient Chinese game, said he had 'horrible experience'

Agence France-Presse in Shanghai

Tue 23 May 2017 11.29 BST

964

This article is over **1 year old**

A screen at the Future of Go summit in Wuzhen, China, keeps score during the match between Chinese player Ke Jie and Google's AI AlphaGo. Photograph: Wu Hong/EPA

most viewed

Ivanka Trump shows yet again she's no friend to women *Arwa Mahdawi*

Storms raze forests, bringing northern Italy 'to its knees'

Trump emoluments case ruling opens way to financial records

Liverpool makes a stand and runs far-right marchers out of town

'Is our life just worth a photo?': the tragic death of a couple in Yosemite

Silver *et al.* (2016, Mastering the game of go without human knowledge)

The Guardian

News Opinion Sport Culture Lifestyle More

World UK Science Cities Global development Football Tech Business Environment Obituaries

Self-driving cars

Driverless taxi debuts in Tokyo in 'world first' trial ahead of Olympics

Minivan equipped with sensors makes four round-trips a day on a busy stretch of road, picking up paying passengers

Justin McCurry *in Tokyo*

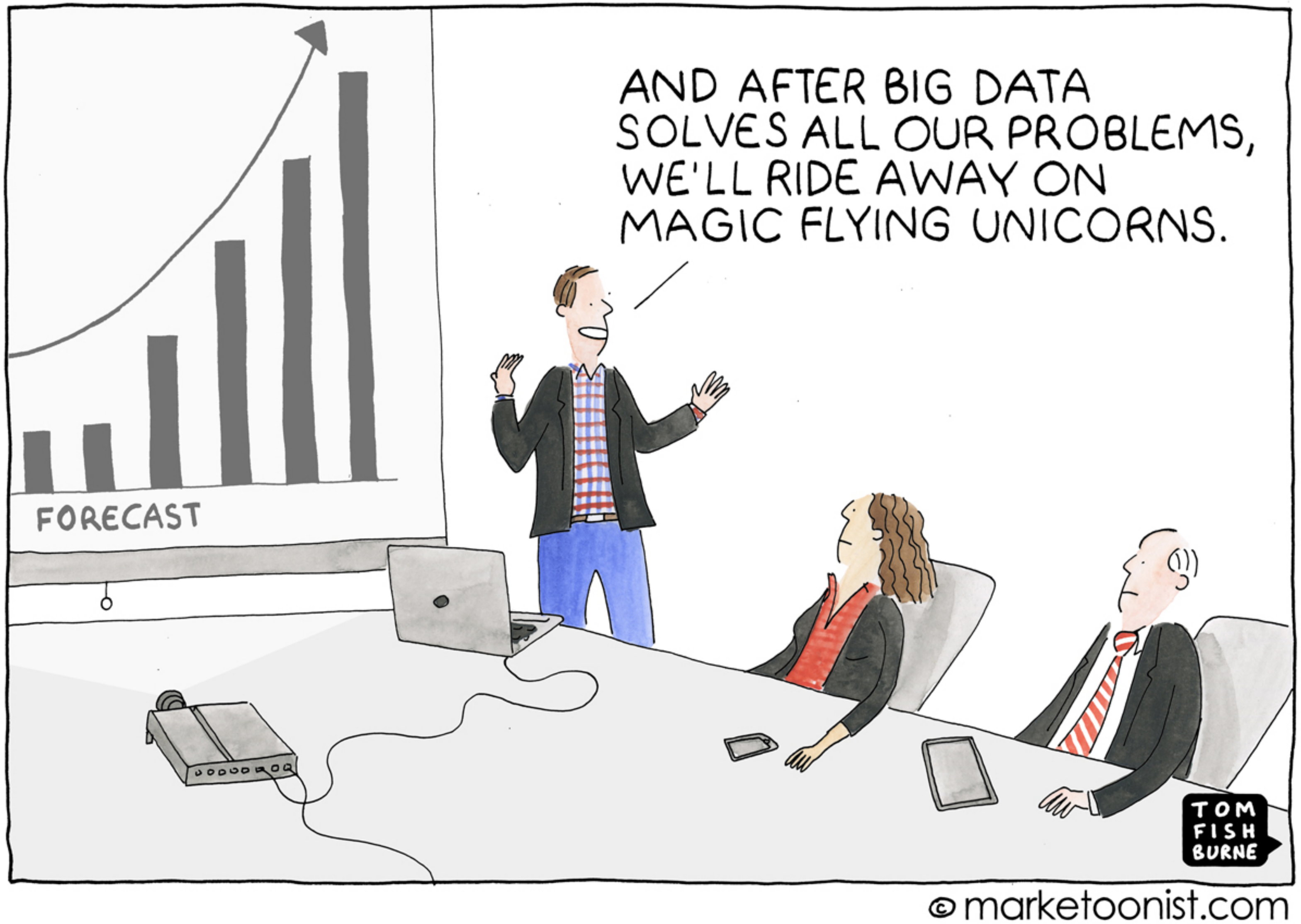
Tue 28 Aug 2018 06.23 BST

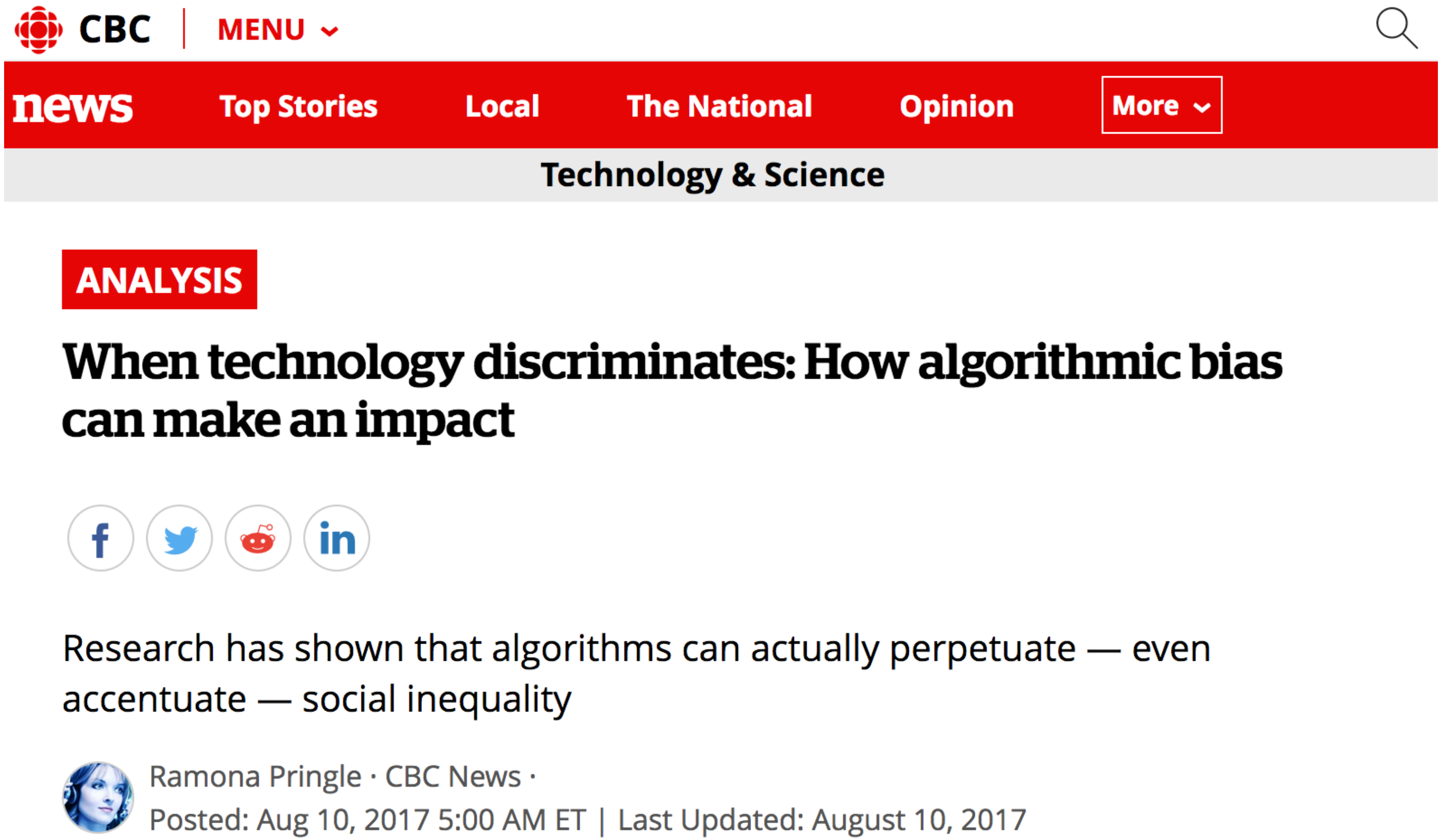
545

This article is over **2 months old**

The RoboCar MiniVan at the start of its test in Tokyo. Photograph: Toru Hanai/Reuters

Chen (2014, Deep Learning for Self -driving Car)


AND AFTER BIG DATA SOLVES ALL OUR PROBLEMS, WE'LL RIDE AWAY ON MAGIC FLYING UNICORNS.
FORECAST
TOM FISH BURNE
© marketoonist.com


CBC | MENU

news | Top Stories | Local | The National | Opinion | More

Technology & Science

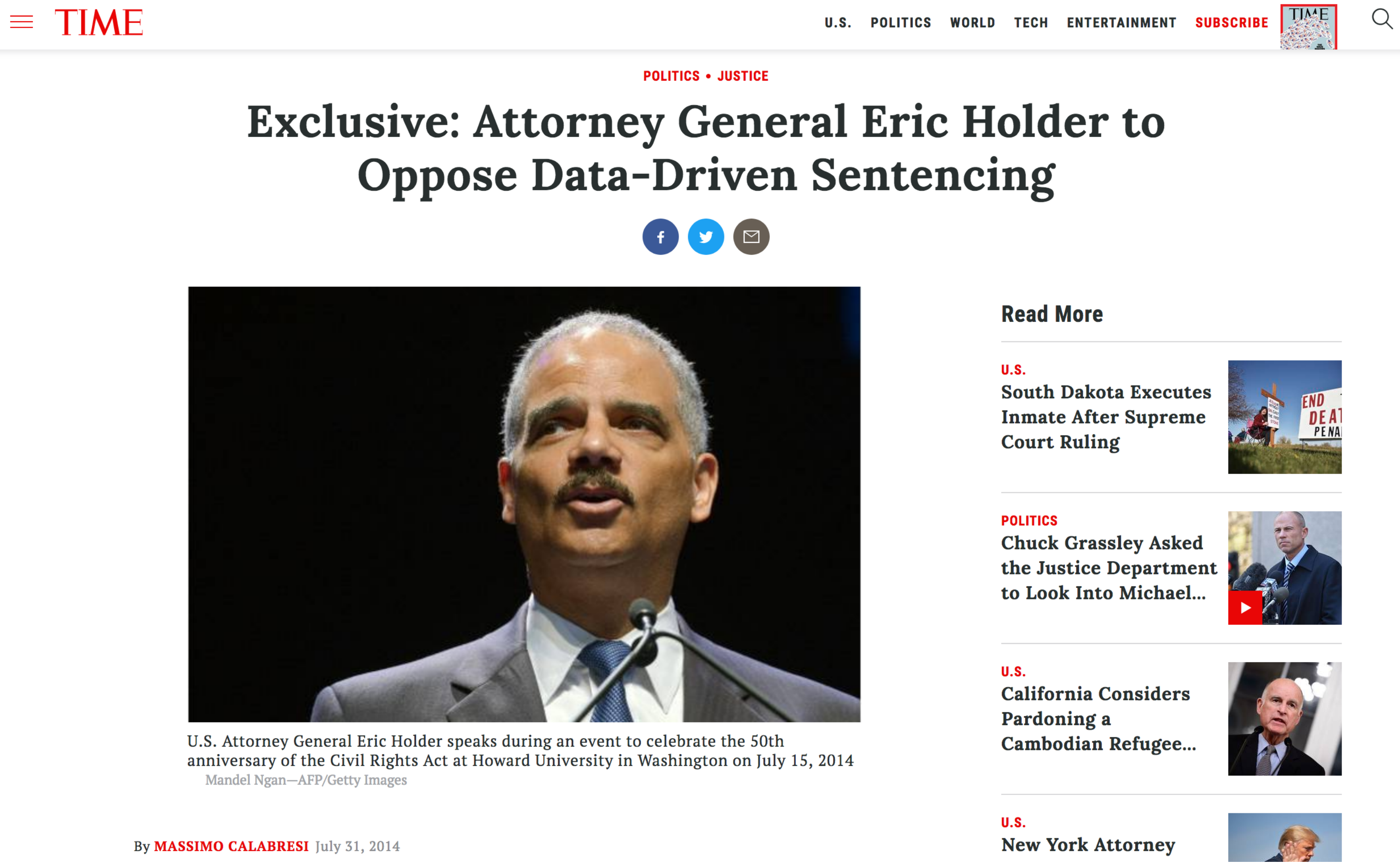
ANALYSIS

When technology discriminates: How algorithmic bias can make an impact

Research has shown that algorithms can actually perpetuate — even accentuate — social inequality

Ramona Pringle · CBC News ·
Posted: Aug 10, 2017 5:00 AM ET | Last Updated: August 10, 2017

O'Neil (2016, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy) and Eubanks (2018, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor)



TIME

U.S. POLITICS WORLD TECH ENTERTAINMENT SUBSCRIBE

POLITICS • JUSTICE

Exclusive: Attorney General Eric Holder to Oppose Data-Driven Sentencing

U.S. Attorney General Eric Holder speaks during an event to celebrate the 50th anniversary of the Civil Rights Act at Howard University in Washington on July 15, 2014
Mandel Ngan—AFP/Getty Images

By MASSIMO CALABRESI July 31, 2014

Read More

U.S.
South Dakota Executes Inmate After Supreme Court Ruling

POLITICS
Chuck Grassley Asked the Justice Department to Look Into Michael...

U.S.
California Considers Pardoning a Cambodian Refugee...

U.S.
New York Attorney

Starr (2018, Evidence-Based Sentencing and Scientific Rationalization of Discrimination)

BUSINESS INSIDER

TECH | FINANCE | POLITICS | STRATEGY | LIFE | ALL

PRIME | INTELLIGENCE

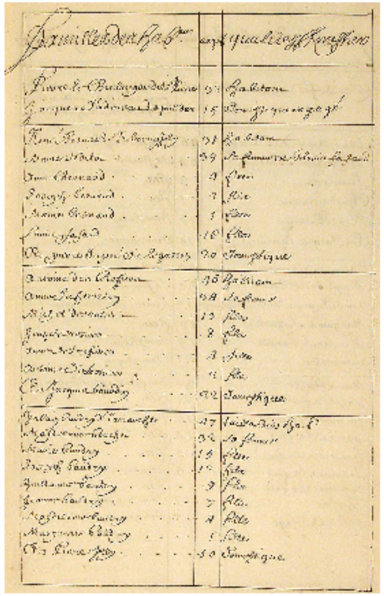
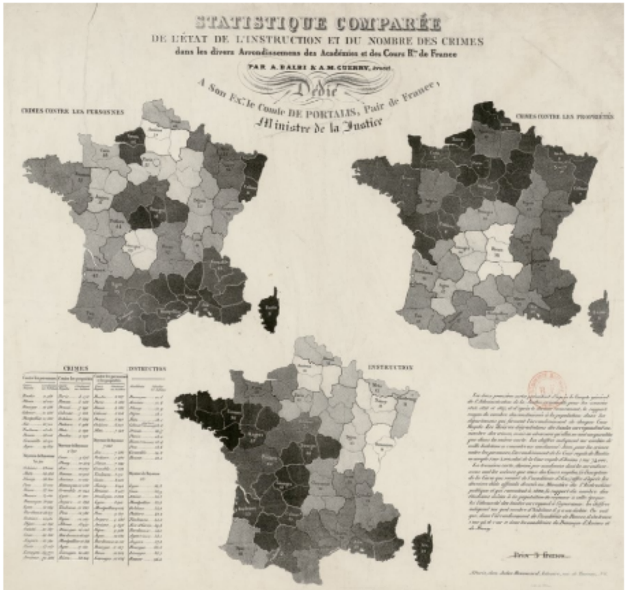
China has started ranking citizens with a creepy 'social credit' system — here's what you can do wrong, and the embarrassing, demeaning ways they can punish you

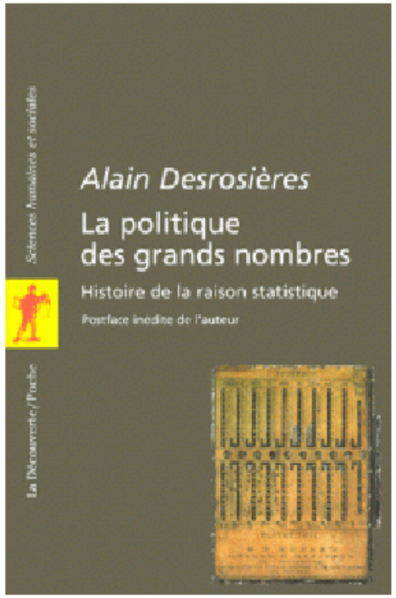
Alexandra Ma Oct. 29, 2018, 12:06 PM

An internet cafe in Wuhan, China. REUTERS/Stringer

Backer (2018, And an Algorithm to Bind Them All? Social Credit, Data Driven Governance, and the Emergence of Operating System for Global Normative Orders)

Data ? Exhaustive statistics ?

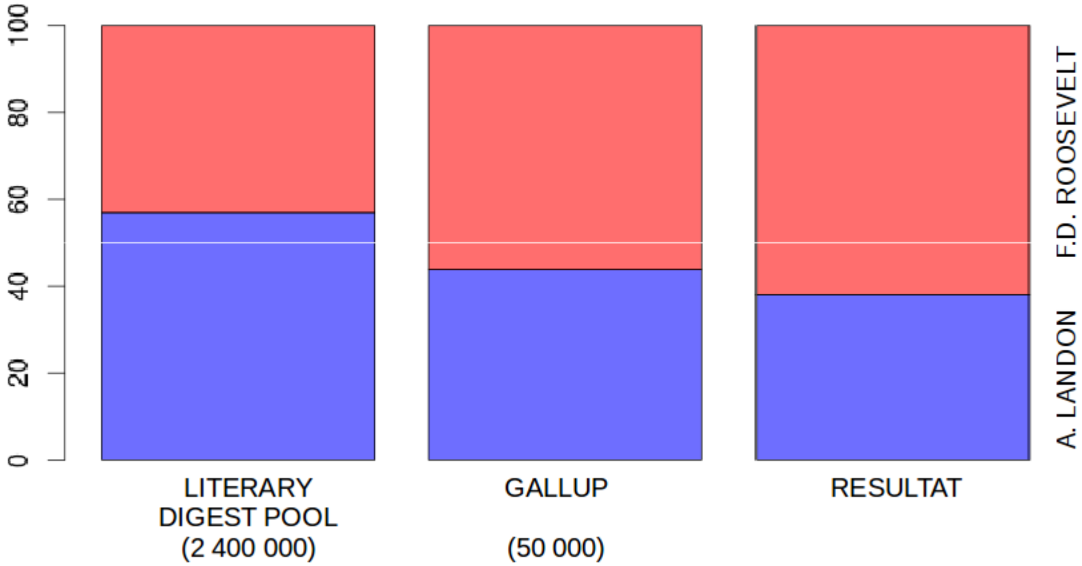



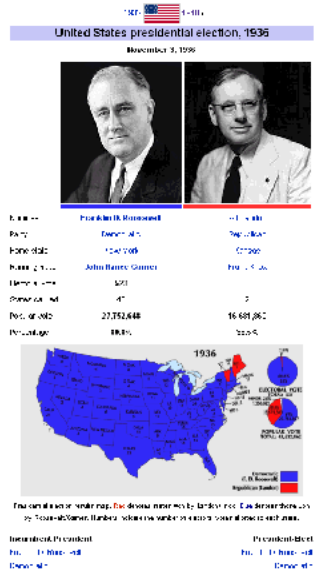


Historically, a political instrument, with the use of aggregates to describe the society, see Martin (2016, chiffrer pour évaluer).

Growing importance for public policy evaluation, and importance of censuses (exhaustivity)

Data ? Sampling Techniques



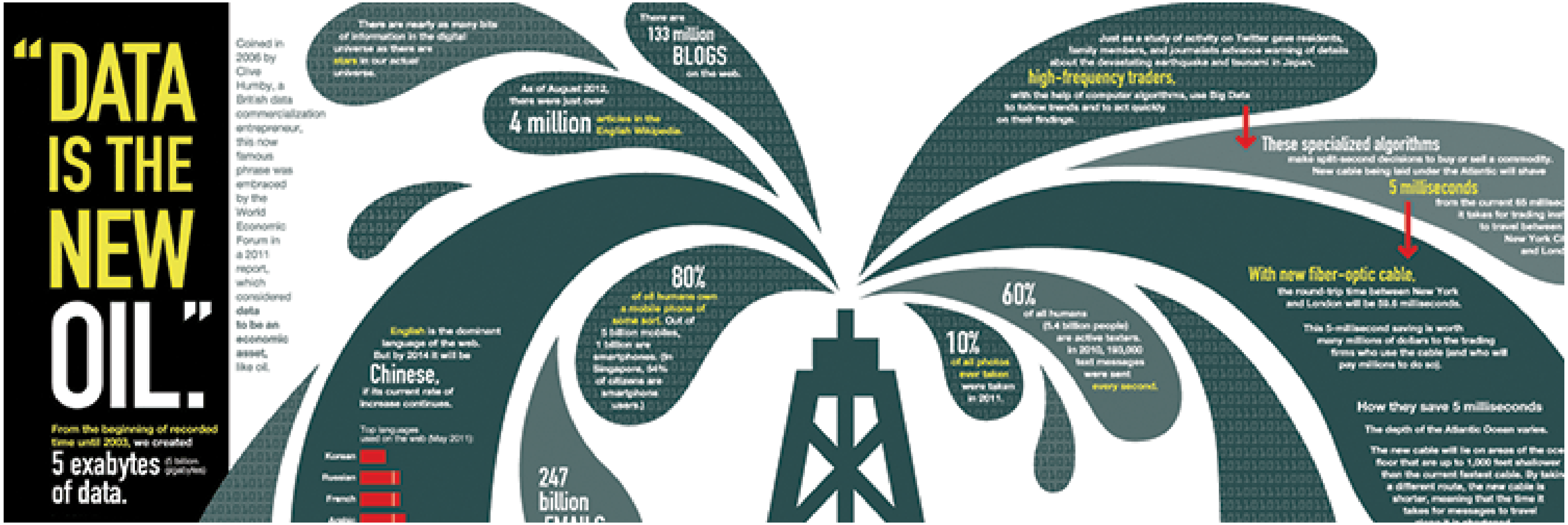


1936, US Presidential Election, "*the more, the better*" ?

Neyman (1934, On the two different aspects of the representative method) Cochran (1953, Sampling Techniques) or Deming (1966, Some Theory of Sampling).

Importance of mathematical statistics (and asmptotic properties)

Data ? The New Oil ?



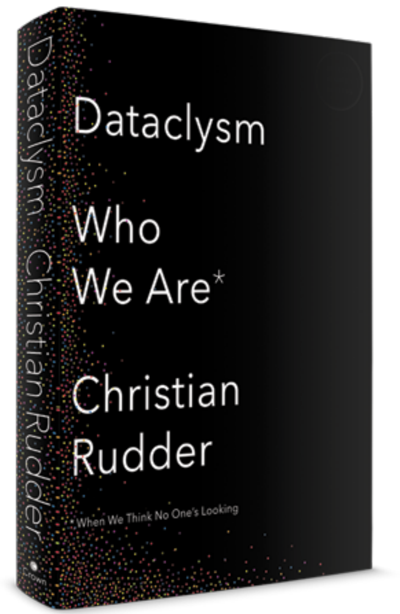
Rotella (2012, Data The New Oil?) or Toonders (2014, Data Is the New Oil of the Digital Economy)

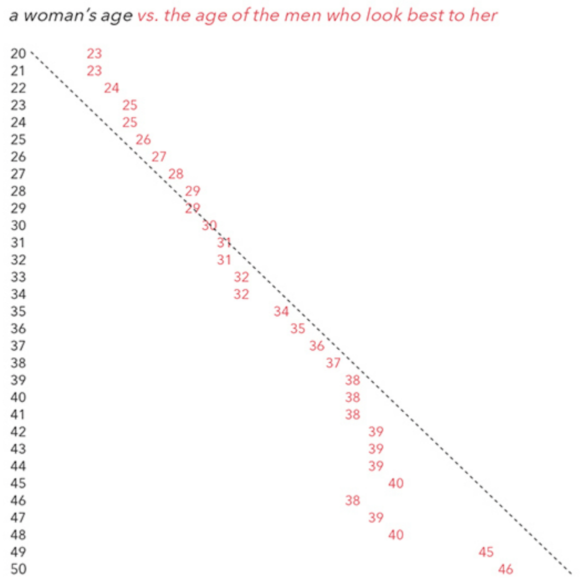
Not rare, infinite deposit, good not rival, no intrinsic value, difficult to protect, and easy to (re)produce, importance of flows (more than stocks)

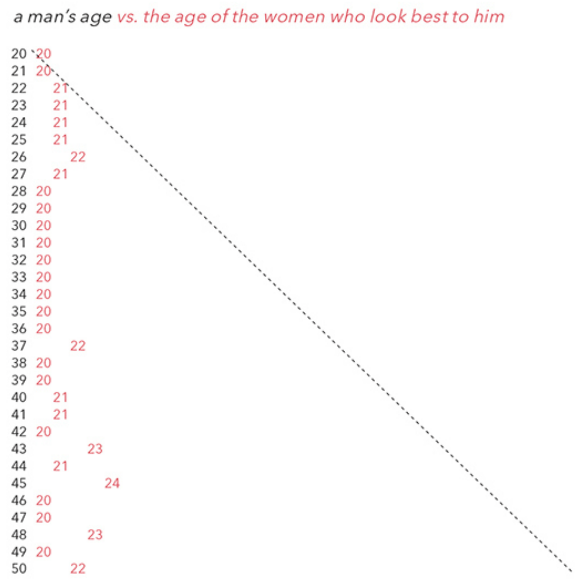
Data ? What For ?

"*People use statistics as the drunken man uses lamp posts - for support rather than illumination*" (Andrew Lang, or not). Notion of Data Driven

Big Data : What Data Say About Us (when we think no one's watching)



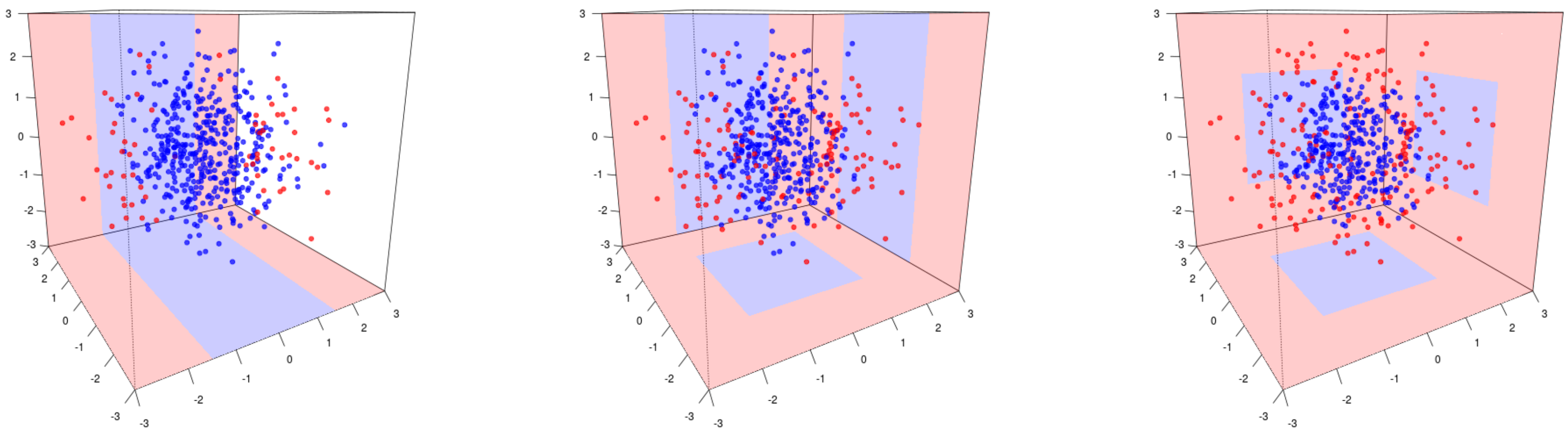




Donnelly (2014, Why OkCupid Users Don't Mind Being Lab Rats) about Rudder (2014, Dataclysm) and http://okcupid.com/

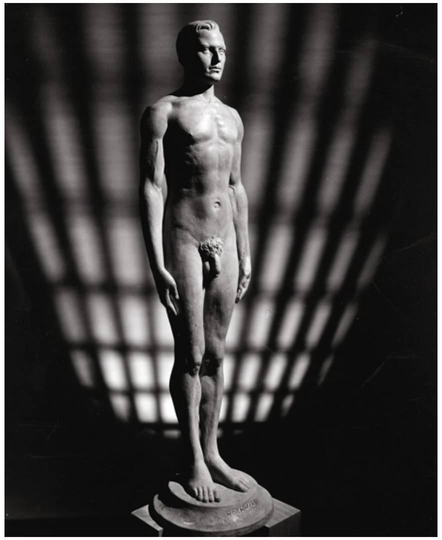
Big Data & Curse of Dimensionality

“*The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality*”



Rose (2016, The End of Average)

Big Data & Curse of Dimensionality : the average does not exist...



Norma & Normann, Cleveland (1943)
by artist Abram Belskie
and obstetrician Robert Dickinson

based on the measurements of 15,000 men and women between the ages of 21 and 25 compiled from a variety of sources, in white racial groups

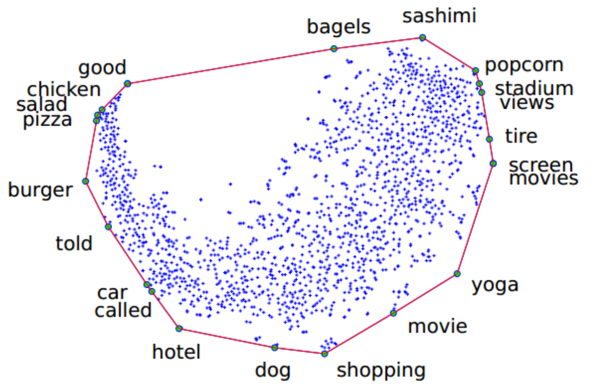
see Stephens (2004, The Object of Normality: The 'Search for Norma' Competition) or Cambers (2004, The Law of Averages : Norman and Norma)

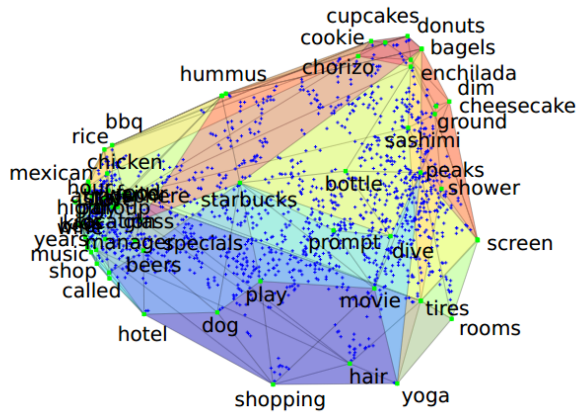
Big Data / New Data ? Text Based Data

- text analytics, web crawling and graph mining

e.g. yelp.com review corpus
(see Lee & Mimmo, 2014 Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference)

index i is a review
variable x_k indicates if review contains kth word
(e.g. yoga, dog or bbq)





Big Data / New Data ? Text Based Data

- co-clustering and text mining

Simultaneous clustering of rows and columns in a matrix

Seminar *noun* 'sem.I.na:ʳ

an occasion when a teacher or expert and a **group** of people meet to study and discuss something via dictionary.cambridge.org

a small **group** of students, as in a university, engaged in advanced study and original research under a member of the faculty and meeting regularly to exchange information and hold discussions via dictionary.reference.com

Big Data / New Data ? Text Based Data

- Recommendation system

What should you get when you search

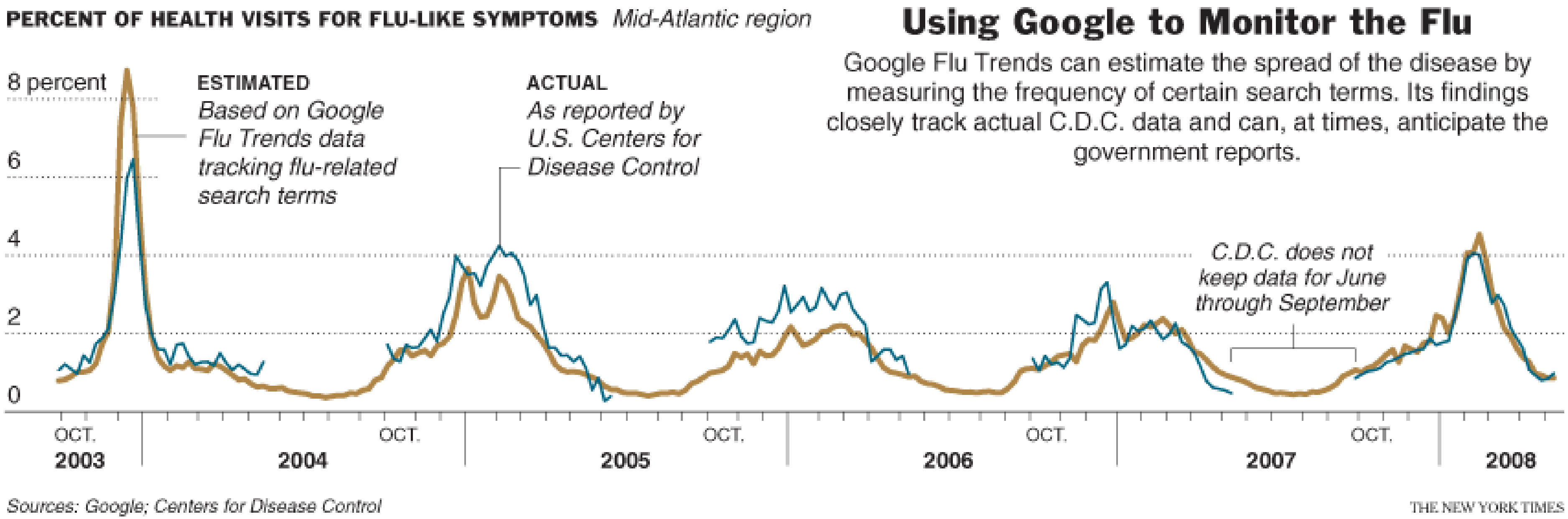
black iphone 5

and in which order should you sort items

see also Santini & Jain (2005, Similarity matching

Big Data / New Data ? Text Based Data

- Internet Browser Searches see Google Flu Project



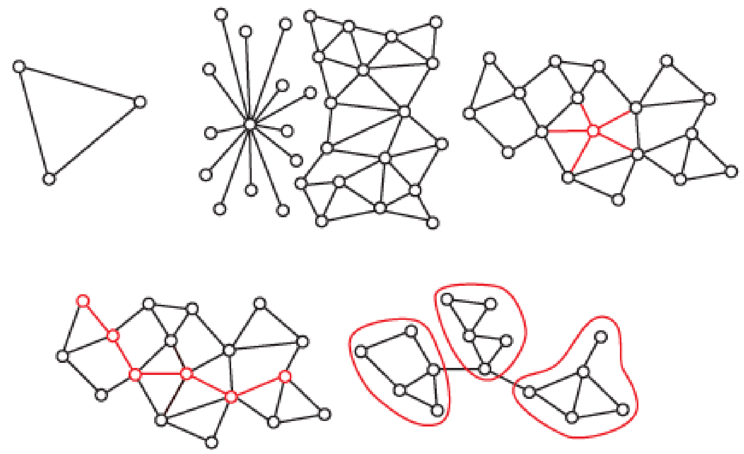
Ginsberg *et al.* (2009, Detecting influenza epidemics using search engine query data)

Butler (2013, When Google got flu wrong)

Big Data / New Data ? Network Data

Classical (individual based) data in econonmetrics $(y_i, \boldsymbol{x}_i)$, $(y_j, \boldsymbol{x}_j)$, etc supposed to be independent

Individuals are nodes of a network v_i, v_j, etc, that can be connected $e_{i,j} = 1$ or not $e_{i,j} = 0$.



See Easley, D. & Kleinberg, J. (2010) Networks, Crowds, and Markets Cambridge University Press, Jackson, M. (2008). Social and Economic Networks, Wasserman, S. & Faust, K. (1994) Social Network Analysis, Christakis & Fowler (2009, Connected : The Surprising Power of Our Social Networks and How They Shape Our Lives) or Can & Alatas (2017, Big Social Network Data and Sustainable Economic Development)

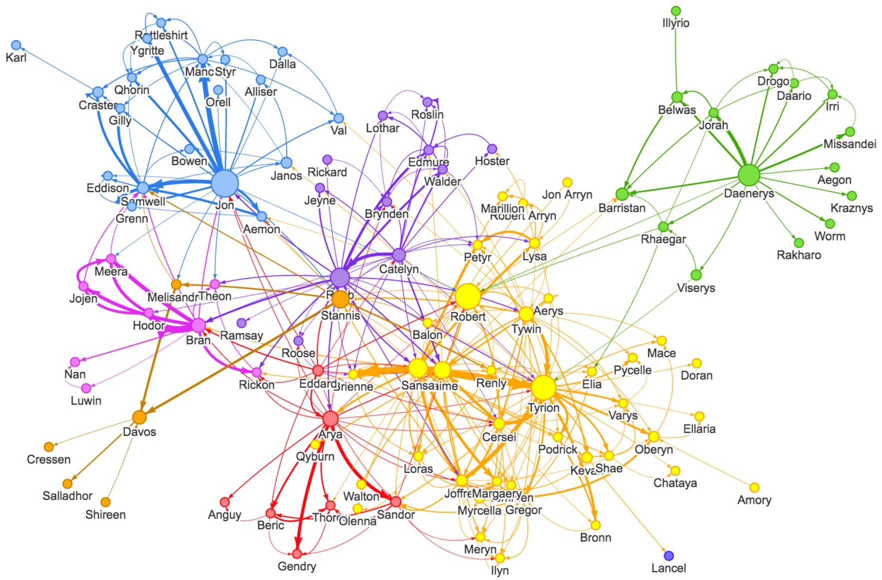
Working with networks can be complicated, and conter-intuitive

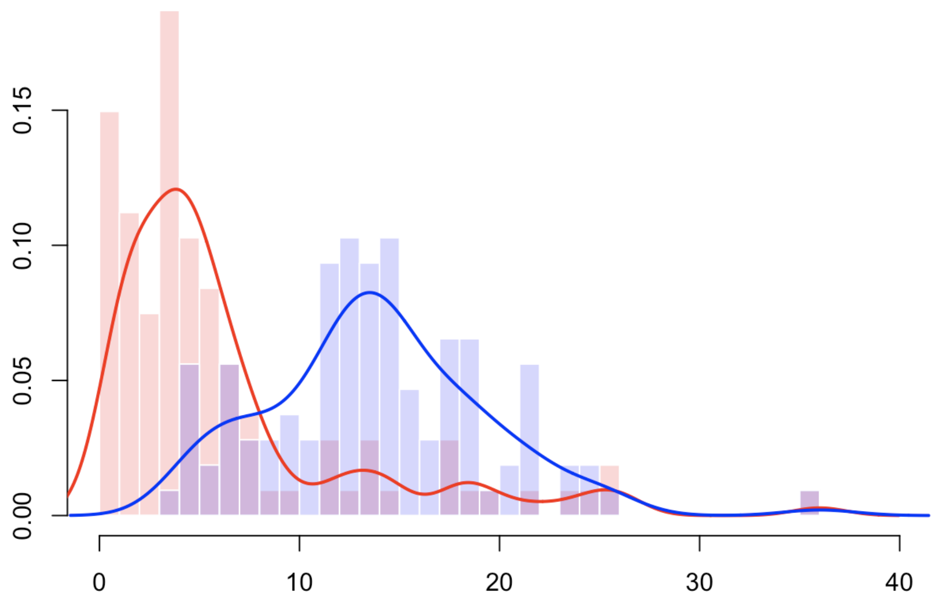
Big Data / New Data ? Network Data

- Friendship paradox

People on average have fewer friends than their friends (popular people are over-represented in the views of others)

See Game of Thrones network

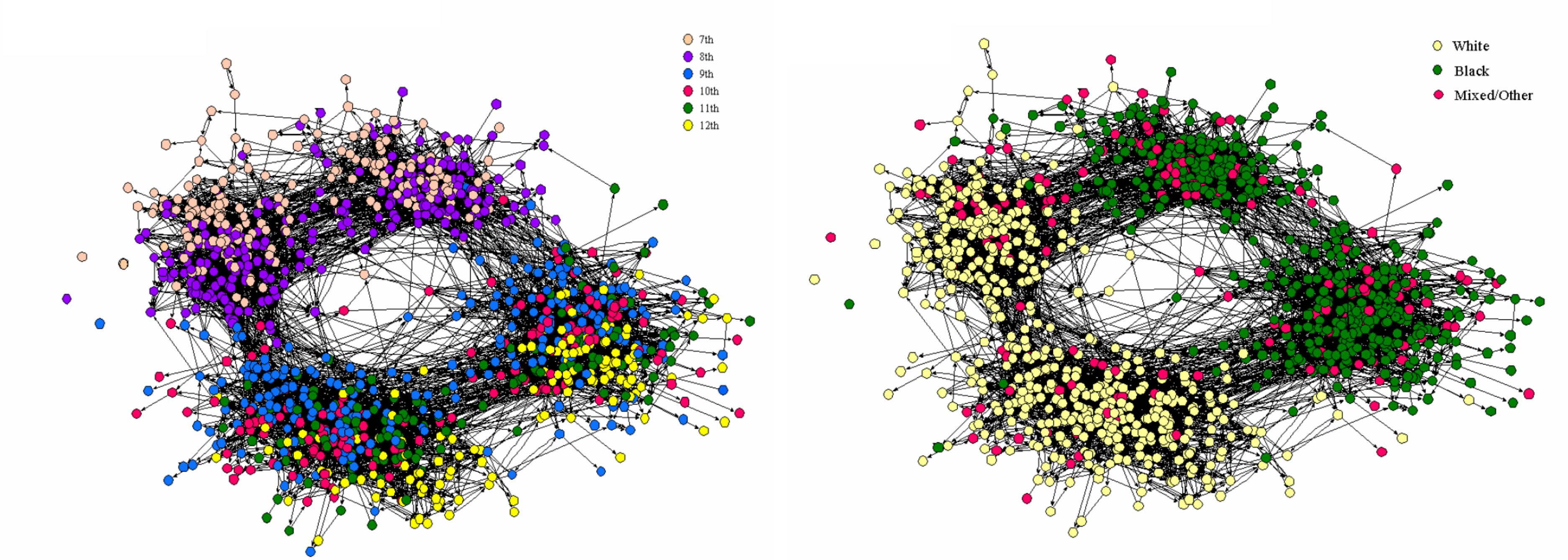




See Feld (1991) and Zuckerman & Jost (2001)

Big Data / New Data ? Network Data

- Homophily, “*birds of a feather flock together*”



from Moody (2001) Race, School Integration and Friendship Segregation in America

see also Being ‘wasted’ on Facebook may damage your credit score

Big Data / New Data ? Network Data

- Peer effect

see Angrist (2014, The perils of peer effects)

Table 1. Students' accuracy in perceiving the drinking norm at their school (comparing actual with perceived number of alcoholic drinks consumed the last time students "partied"/socialized)

Actual school norm (median drinks)	Accuracy of perceived drinking norm: Underestimate of ≥3 drinks (%)	Underestimate of 1-2 drinks (%)	Accurate estimate (%)	Overestimate of 1-2 drinks (%)	Overestimate of ≥3 drinks (%)	Total	*N* of respondents	*N* of schools
0	NA	NA	20.6	19.5	59.9	=100%	1,891	4
1	NA	10.5	3.8	28.5	57.2	=100%	2,526	6
2	NA	7.5	8.1	30.6	53.7	=100%	8,345	14
3	3.8	6.4	13.5	37.5	38.7	=100%	18,859	35
4	3.1	12.3	12.6	37.0	34.9	=100%	20,353	38
5	4.3	15.8	20.6	24.1	35.3	=100%	11,481	20
6	6.9	23.2	15.0	23.5	31.5	=100%	8,912	12
7	5.7	23.3	9.7	23.6	37.8	=100%	352	1
Total schools	3.4	11.8	13.8	31.9	39.1	=100%	72,719	130

Source : Perkins, Haines & Rice (2005, Misperceiving the college drinking norm and related problems)

Big Data / New Data ? Pictures Data

- image processing

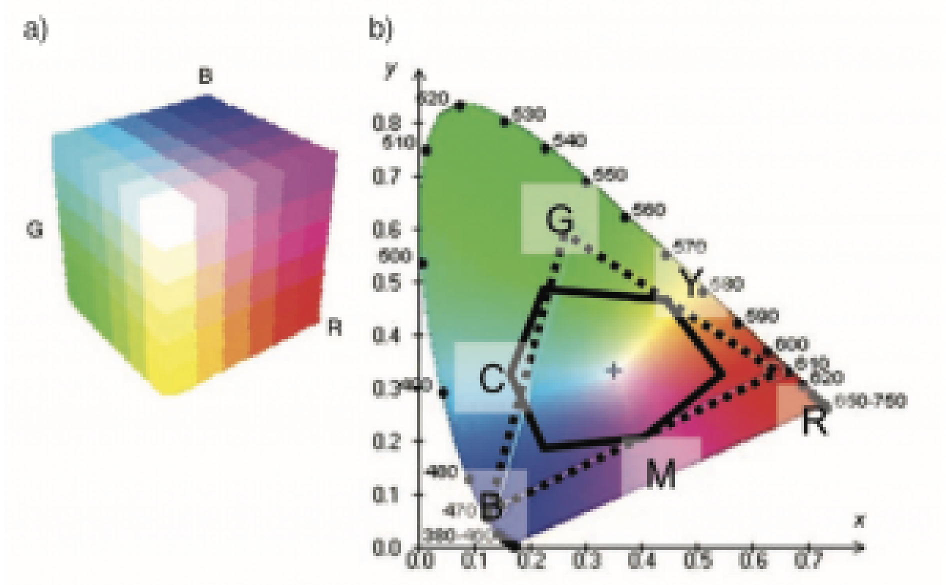
a picture is an array

```
        [,454]    [,455]    [,456]
[8,] 0.5137255 0.5176471 0.5411765
[9,] 0.5019608 0.5254902 0.5137255
        [,457]    [,458]    [,459]
[8,] 0.4666667 0.5921569 0.5529412
[9,] 0.4823529 0.5294118 0.6117647
        [,460]    [,461]    [,462]
[8,] 0.5960784 0.5764706 0.5529412
[9,] 0.5450980 0.6000000 0.5607843
```

Big Data / New Data ? Pictures Data

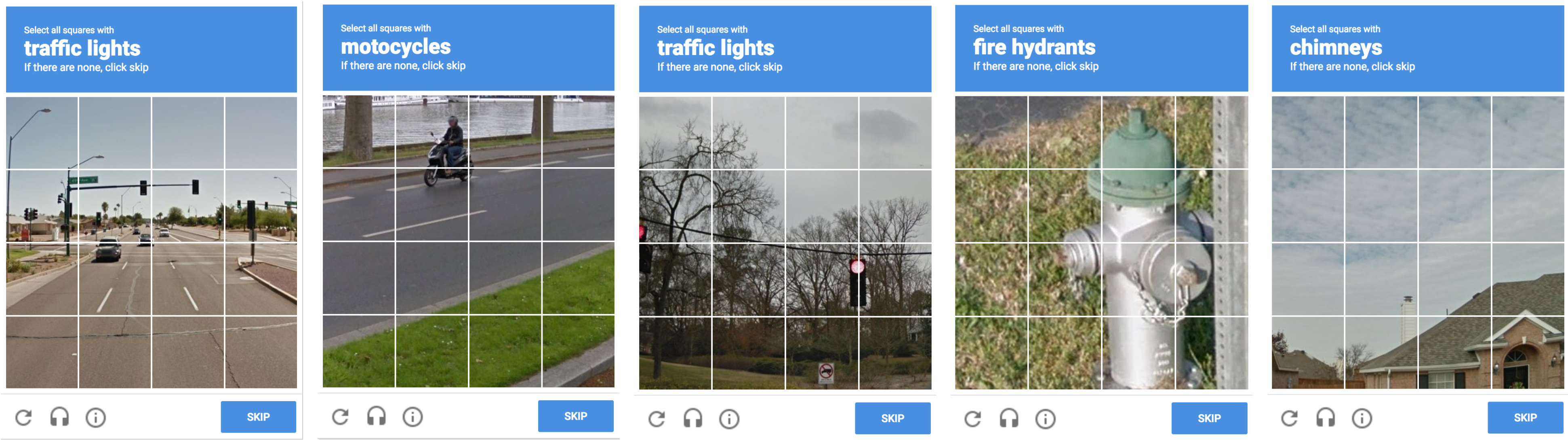
Lenna, Playboy Magazine, November 1972
with an RGB decomposition, wikipedia

can be used for feature detection (edges, SIFT - *scale-invariant feature transform*)



Big Data / New Data ? Pictures Data

Learning is based on tagged trained samples,



Goedegebuure (2016, You Are Helping Google AI Image Recognition)

O'Malley (2018, how you've been training AI for years without realising it

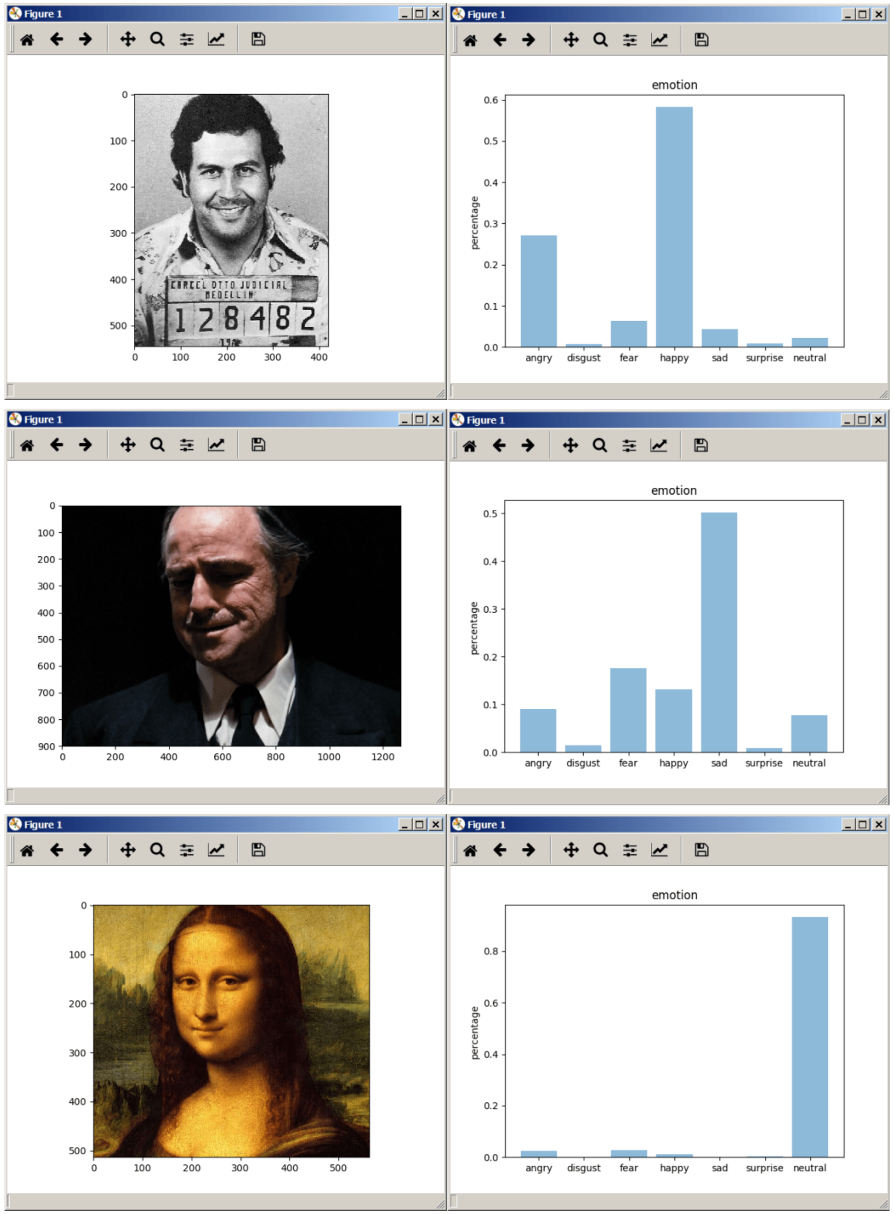
Big Data / New Data ? Pictures Data

Sometimes, things can be complicated...

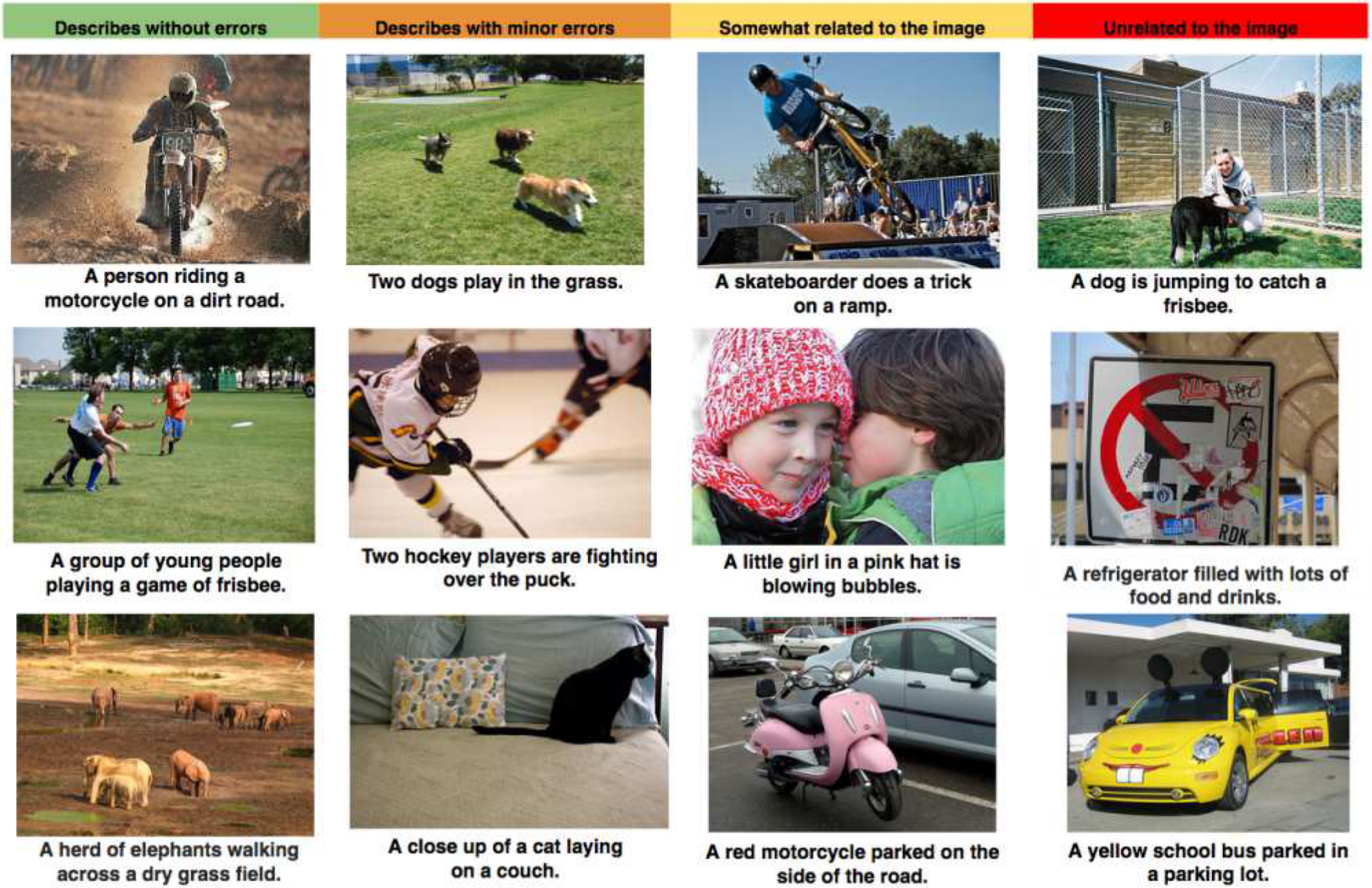
Labradoodle or fried chicken ? Sheepdog or mop ? Barn owl or apple ?

Big Data / New Data ? Pictures Data

- face or emotion recognition

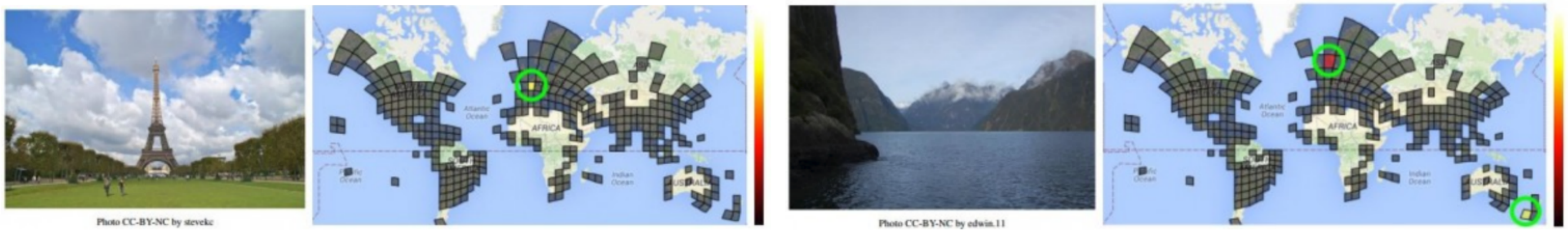


Big Data / New Data ? Pictures Data - Automatic Labeling



Big Data / New Data ? Pictures Data

can be used to locate a place from a picture



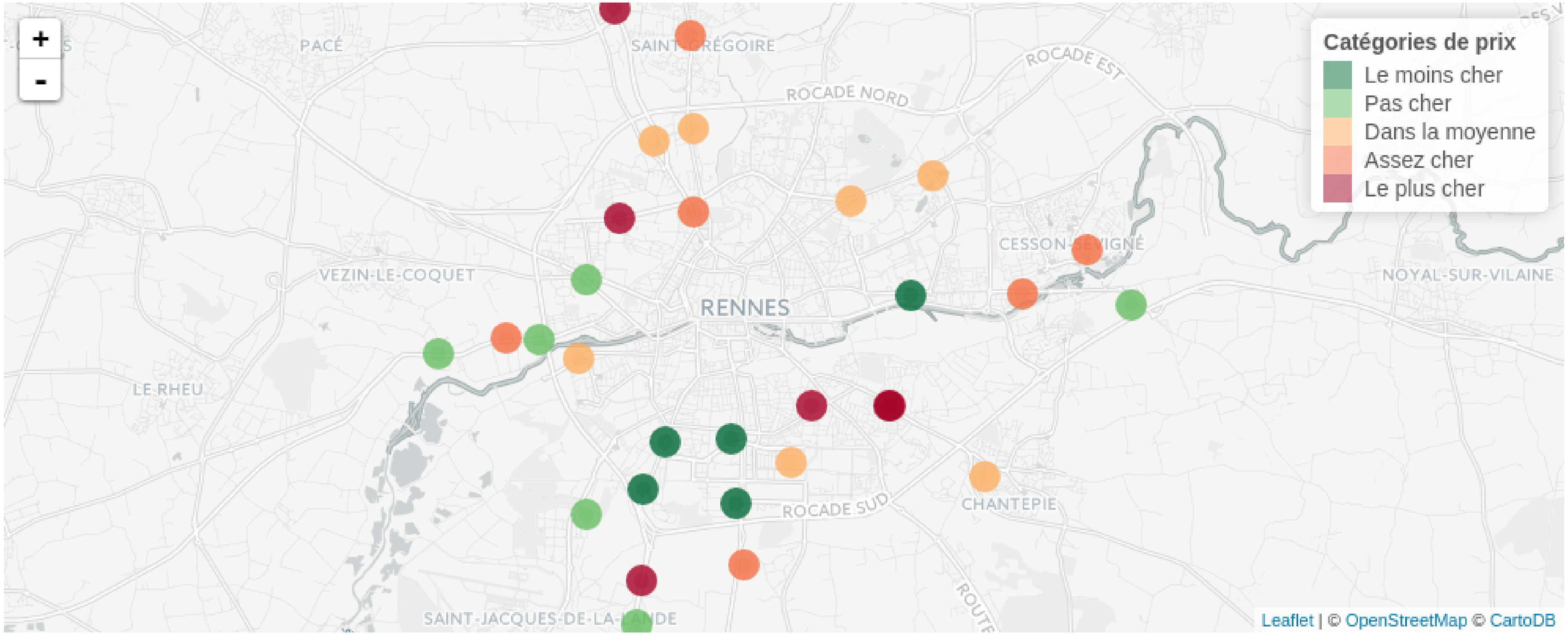
can be used to locate someone in the crowd ?

Big Data / Big Brother ?

Botsman (2017, Big data meets Big Brother as China moves to rate its citizens)

Big Data / Open Data ?

One can use open data to create an app, e.g. based on oil price, prixdescarburants.info and data.gouv.fr

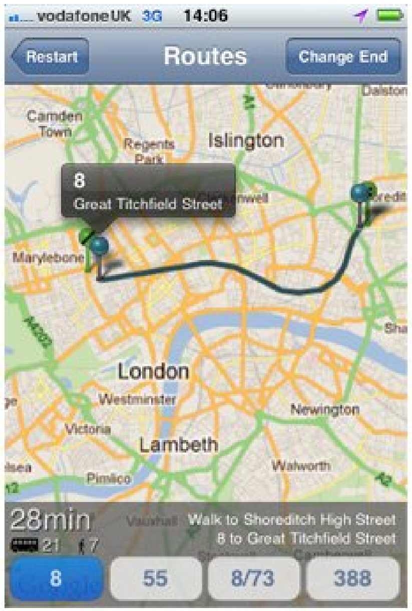


Big Data / Open Data ?

See citymapper, public transit app and mapping service

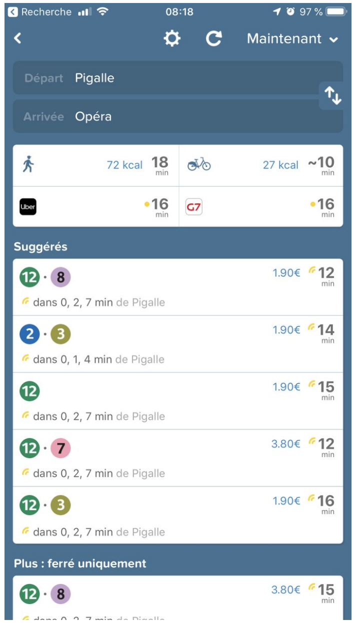
See Cohen (2018, The Guy Making Public Transit Smarter)
2011, Azmat Yusuf arrived in London to work for Google
Lost with London maps for public transportation,
creates "Busmapper" (that will become "Citymapper"

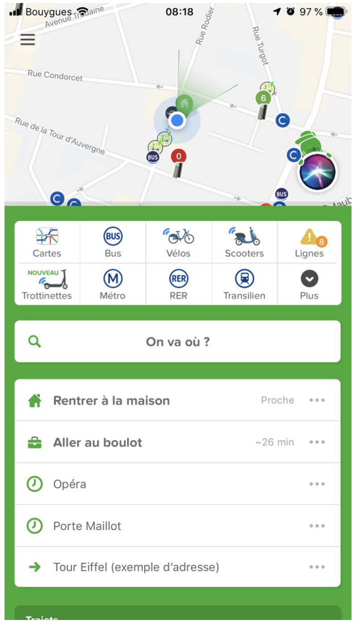
Citymapper

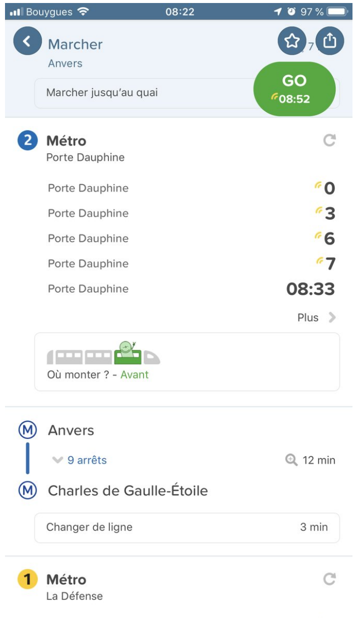


Big Data / Open Data ?

Based on open data, see the blog post Getting from A to (Series) B







2016: arrived in Paris, RATP did not want to open its data, bad buzz...

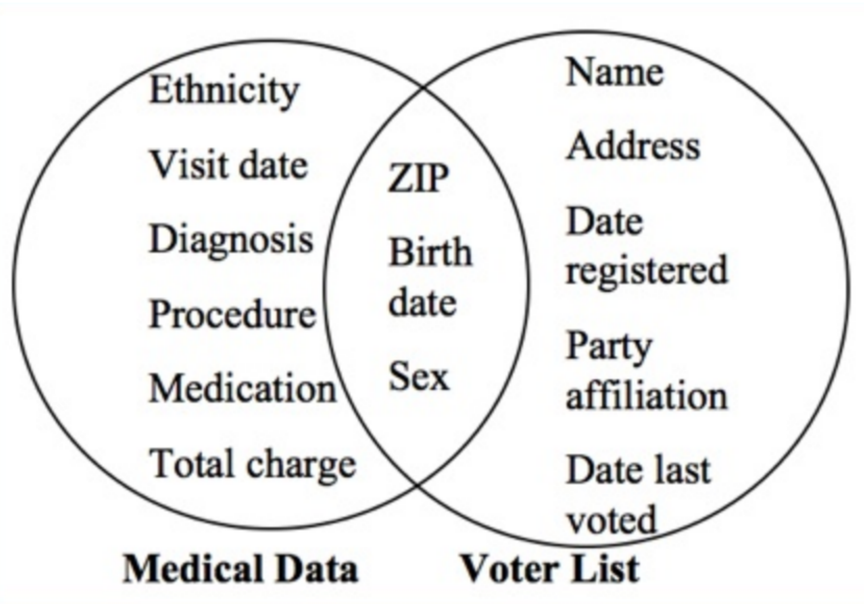
Collects billions of trajectories, used to compute real-time optimal journey

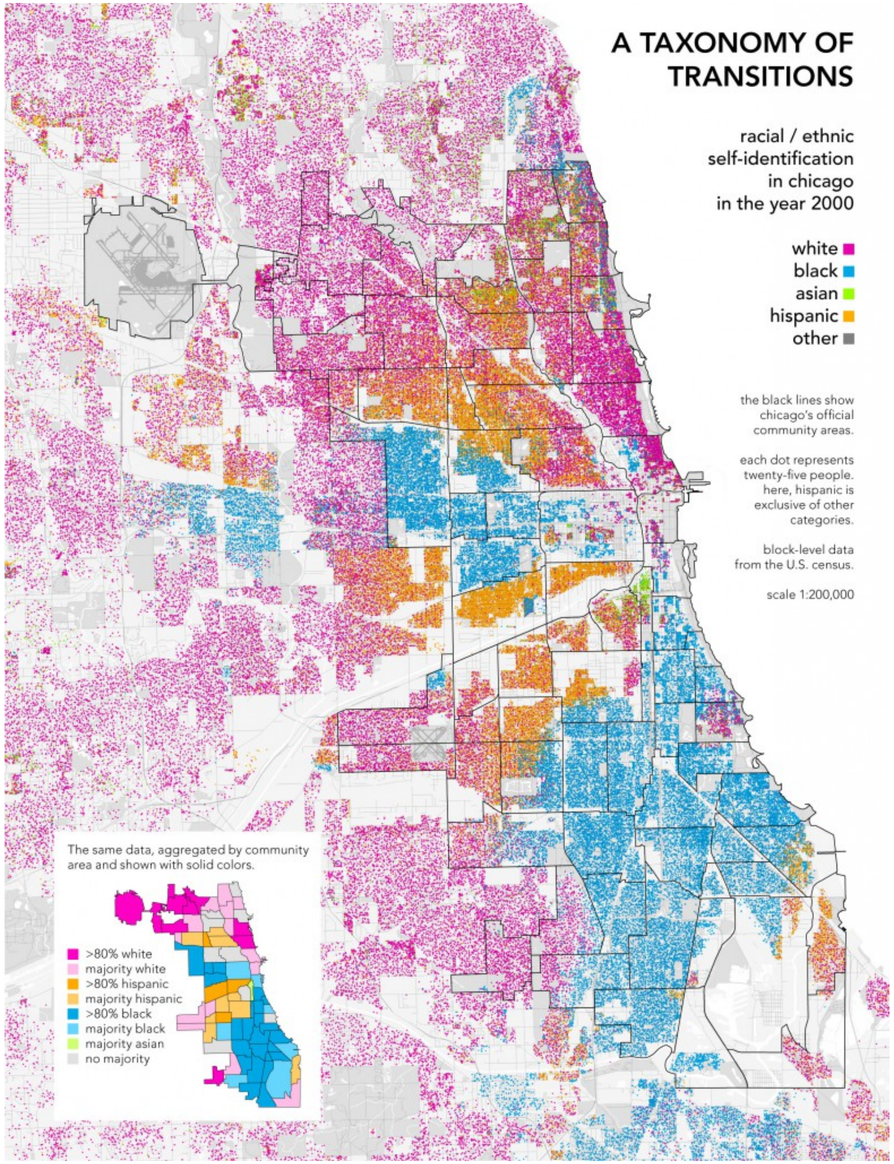
See also Building a city without open data: introducing Project Istanbul

Big Data / Open Data ?

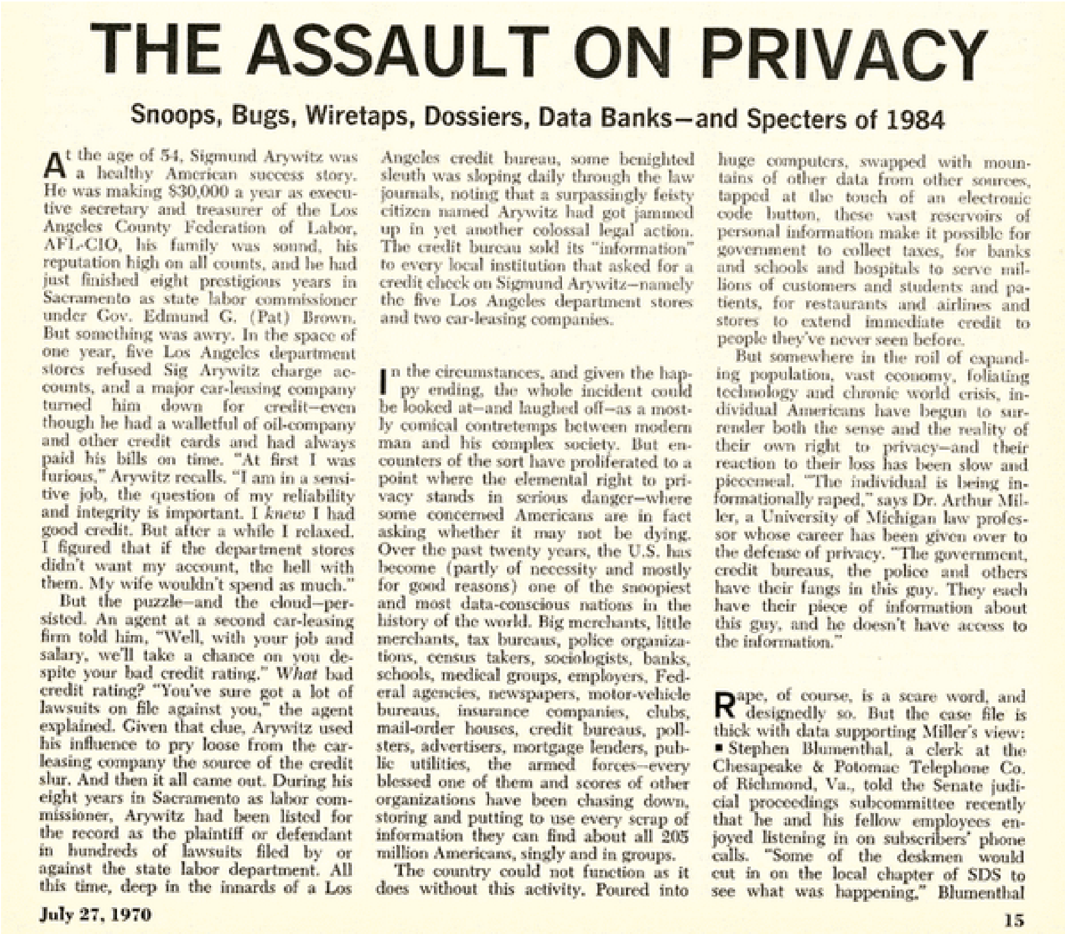
Open data with very low granularity
Rankin (2009) with census data.

See privacy breach,
Sweeney (2002, k-anonymity)





Privacy : A New Old Problem ?



THE ASSAULT ON PRIVACY

Snoops, Bugs, Wiretaps, Dossiers, Data Banks—and Specters of 1984

At the age of 54, Sigmund Arywitz was a healthy American success story. He was making $30,000 a year as executive secretary and treasurer of the Los Angeles County Federation of Labor, AFL-CIO, his family was sound, his reputation high on all counts, and he had just finished eight prestigious years in Sacramento as state labor commissioner under Gov. Edmund G. (Pat) Brown. But something was awry. In the space of one year, five Los Angeles department stores refused Sig Arywitz charge accounts, and a major car-leasing company turned him down for credit—even though he had a walletful of oil-company and other credit cards and had always paid his bills on time. "At first I was furious," Arywitz recalls. "I am in a sensitive job, the question of my reliability and integrity is important. I *knew* I had good credit. But after a while I relaxed. I figured that if the department stores didn't want my account, the hell with them. My wife wouldn't spend as much."

But the puzzle—and the cloud—persisted. An agent at a second car-leasing firm told him, "Well, with your job and salary, we'll take a chance on you despite your bad credit rating." *What* bad credit rating? "You've sure got a lot of lawsuits on file against you," the agent explained. Given that clue, Arywitz used his influence to pry loose from the car-leasing company the source of the credit slur. And then it all came out. During his eight years in Sacramento as labor commissioner, Arywitz had been listed for the record as the plaintiff or defendant in hundreds of lawsuits filed by or against the state labor department. All this time, deep in the innards of a Los Angeles credit bureau, some benighted sleuth was sloping daily through the law journals, noting that a surpassingly feisty citizen named Arywitz had got jammed up in yet another colossal legal action. The credit bureau sold its "information" to every local institution that asked for a credit check on Sigmund Arywitz—namely the five Los Angeles department stores and two car-leasing companies.

In the circumstances, and given the happy ending, the whole incident could be looked at—and laughed off—as a mostly comical contretemps between modern man and his complex society. But encounters of the sort have proliferated to a point where the elemental right to privacy stands in serious danger—where some concerned Americans are in fact asking whether it may not be dying. Over the past twenty years, the U.S. has become (partly of necessity and mostly for good reasons) one of the snoopiest and most data-conscious nations in the history of the world. Big merchants, little merchants, tax bureaus, police organizations, census takers, sociologists, banks, schools, medical groups, employers, Federal agencies, newspapers, motor-vehicle bureaus, insurance companies, clubs, mail-order houses, credit bureaus, pollsters, advertisers, mortgage lenders, public utilities, the armed forces—every blessed one of them and scores of other organizations have been chasing down, storing and putting to use every scrap of information they can find about all 205 million Americans, singly and in groups.

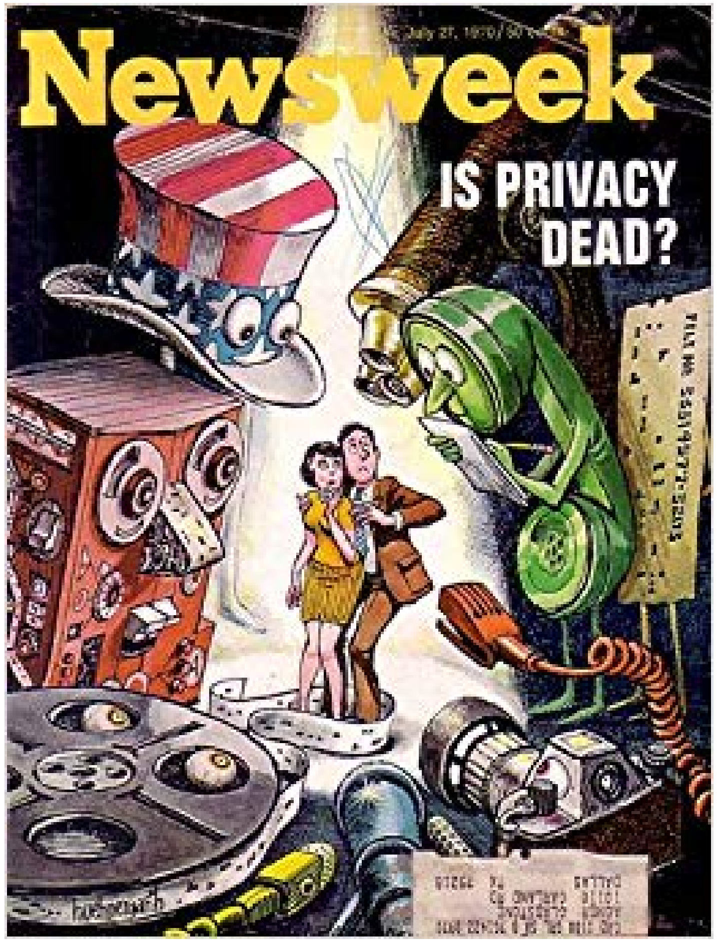
The country could not function as it does without this activity. Poured into huge computers, swapped with mountains of other data from other sources, tapped at the touch of an electronic code button, these vast reservoirs of personal information make it possible for government to collect taxes, for banks and schools and hospitals to serve millions of customers and students and patients, for restaurants and airlines and stores to extend immediate credit to people they've never seen before.

But somewhere in the roil of expanding population, vast economy, foliating technology and chronic world crisis, individual Americans have begun to surrender both the sense and the reality of their own right to privacy—and their reaction to their loss has been slow and piecemeal. "The individual is being informationally raped," says Dr. Arthur Miller, a University of Michigan law professor whose career has been given over to the defense of privacy. "The government, credit bureaus, the police and others have their fangs in this guy. They each have their piece of information about this guy, and he doesn't have access to the information."

Rape, of course, is a scare word, and designedly so. But the case file is thick with data supporting Miller's view:

- Stephen Blumenthal, a clerk at the Chesapeake & Potomac Telephone Co. of Richmond, Va., told the Senate judicial proceedings subcommittee recently that he and his fellow employees enjoyed listening in on subscribers' phone calls. "Some of the deskmen would cut in on the local chapter of SDS to see what was happening," Blumenthal

July 27, 1970 15

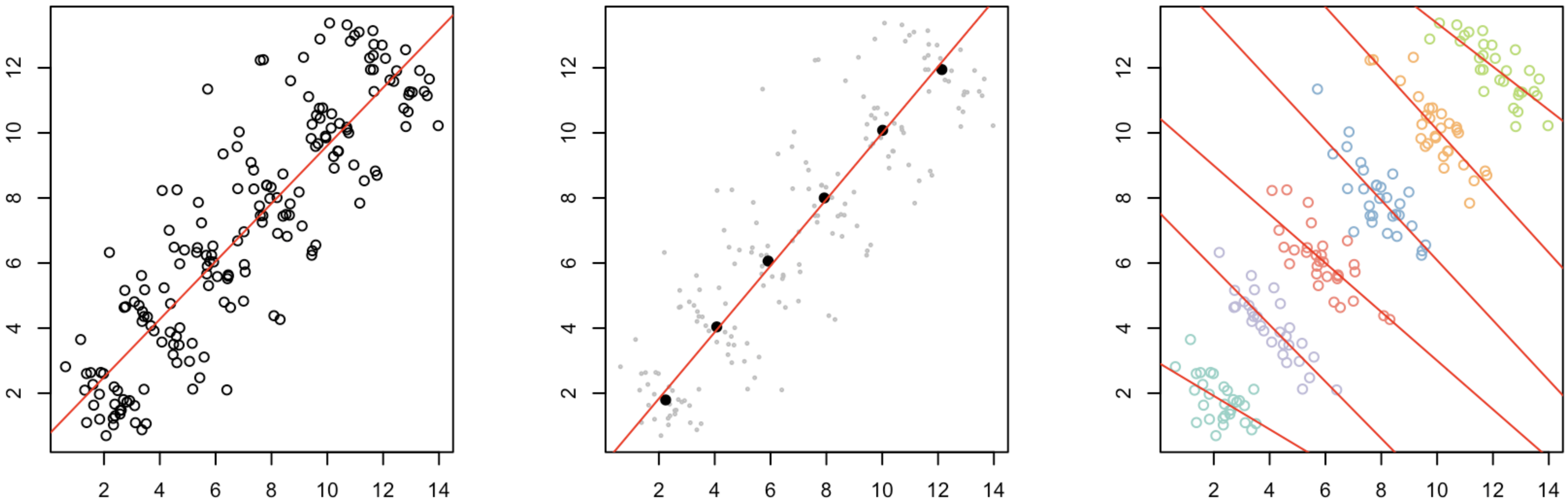


Boeth (1970, Is Privacy Dead?) - see also GDPR in Europe

Open - But Grouped - data ?

In order to avoid privacy issues, we do not acess individual data, $\boldsymbol{x}_i = (x_{1,i}, x_{2,i}, \cdots)$, but aggregated data, e.g. per spatial region $\overline{\boldsymbol{x}}_j = (\overline{x}_{1,j}, \overline{x}_{1,j}, \cdots)$ where $\overline{x}_{1,j}$ is the average of individuals i leaving in region j.

Problem, ecological fallacy - see wikipedia



see Gelman (2010, Red State, Blue State, Rich State, Poor State)

Open - But Grouped - data ?

See also Simpson's paradox - from Blyth (1972)

HOSPITAL	ADMISSION	SURVIVAL	DEATH	SURVIVAL RATE
A	1000	800	200	80%
B	1000	900	100	**90%**

For healthy people

HOSPITAL	ADMISSION	SURVIVAL	DEATH	SURVIVAL RATE
A	600	590	10	**98%**
B	900	870	30	97%

while for sick people

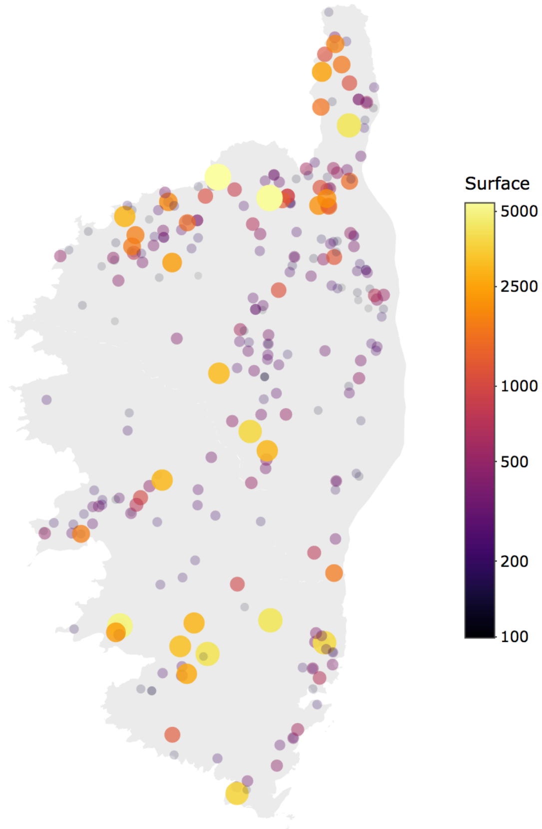
HOSPITAL	ADMISSION	SURVIVAL	DEATH	SURVIVAL RATE
A	400	210	190	**53%**
B	100	30	70	30%

Non-open data ?

Use scraping techniques,
e.g. Web Browser Automation with selenium

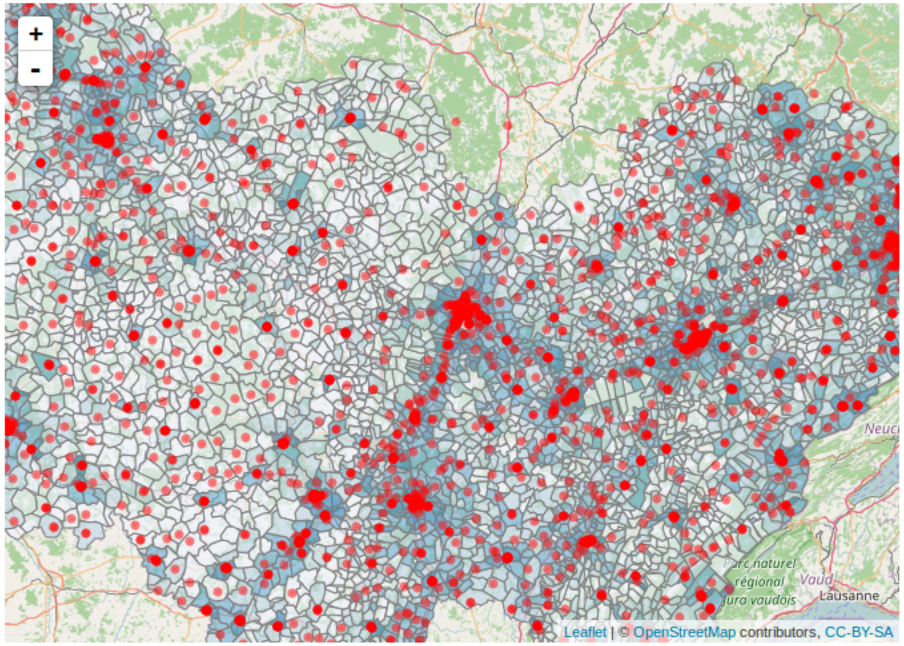
R code to scrap the Prométhée database
on fires, in France
see Scraper la base d'incendies de forêts

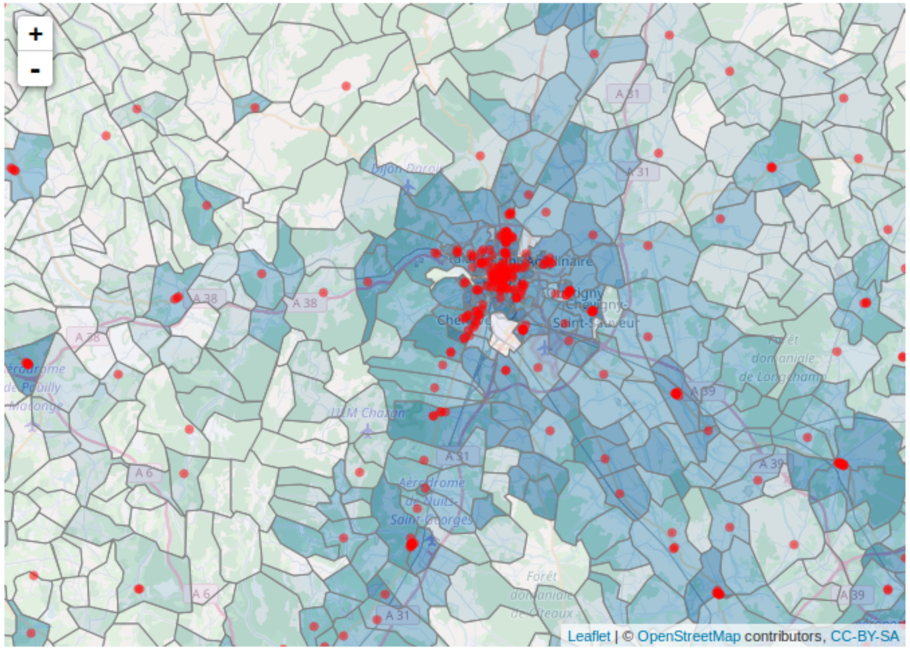
or Scraper pour avoir des infos sur les médecins sur Paris
on the ameli database, on doctors



Non-open data ?

see A quelle distance d'une banque habite-t-on ?



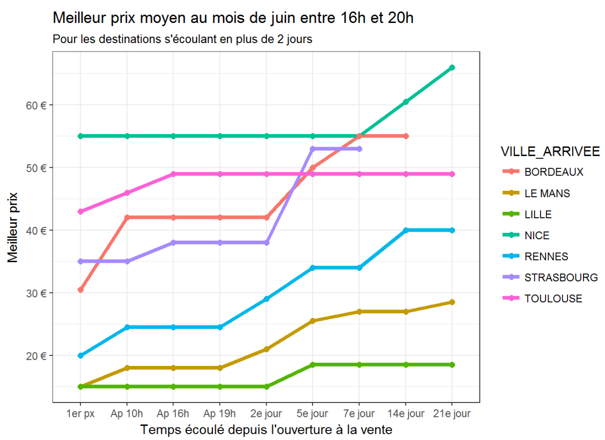


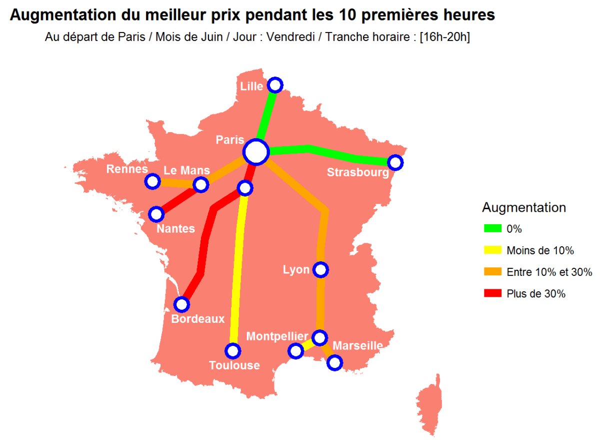
with a scrap of cbanque

Non-open data ?

see Acheter un billet de train (pas trop cher)

Ville de départ	Ville d'arrivée	1er prix	Après 10h	Après 16h	Après 19h	2ème jour	5ème jour	7ème jour	14ème jour	21ème jour
PARIS	BORDEAUX	30.5 €	42 €	42 €	42 €	42 €	50 €	55 €	55 €	—
PARIS	LE MANS	15 €	18 €	18 €	18 €	21 €	25.5 €	27 €	27 €	28.5 €
PARIS	LILLE	15 €	15 €	15 €	15 €	15 €	18.5 €	18.5 €	18.5 €	18.5 €
PARIS	LYON	38.5 €	47.5 €	50 €	50 €	50 €	—	—	—	—
PARIS	MARSEILLE	45.5 €	57 €	57 €	57 €	57 €	—	—	—	—
PARIS	MONTPELLIER	45.5 €	49 €	53 €	57 €	57 €	—	—	—	—
PARIS	NANTES	32 €	44 €	—	—	—	—	—	—	—
PARIS	NICE	55 €	55 €	55 €	55 €	55 €	55 €	55 €	60.5 €	66 €
PARIS	RENNES	20 €	24.5 €	24.5 €	24.5 €	29 €	34 €	34 €	40 €	40 €
PARIS	STRASBOURG	35 €	35 €	38 €	38 €	38 €	53 €	53 €	—	—
PARIS	TOULOUSE	43 €	46 €	49 €	49 €	49 €	49 €	49 €	49 €	49 €





based on casperjs, a browser emulator written in javascript.

Non-open data ?

Consider the case where datasets are located on various servers, and cannot be downloaded (e.g. hospitals), but one can run functions and obtain outputs.
see Wolfson *et. al* (2010, Data Shield)
or http://www.datashield.ac.uk/

Consider a regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Recall that $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ is the OLS estimator.

Is it possible to use parallel computations ? [spoiler: yes].

Big Data and Black Box Models

Black box models that solve a prediction problem:

- given an input $\boldsymbol{x}$
- predict an appropriate output y

E.g spam detection : $\boldsymbol{x}$ is an incoming email, $y \in \{\textsf{spam, not spam}\}$ (binary classification)

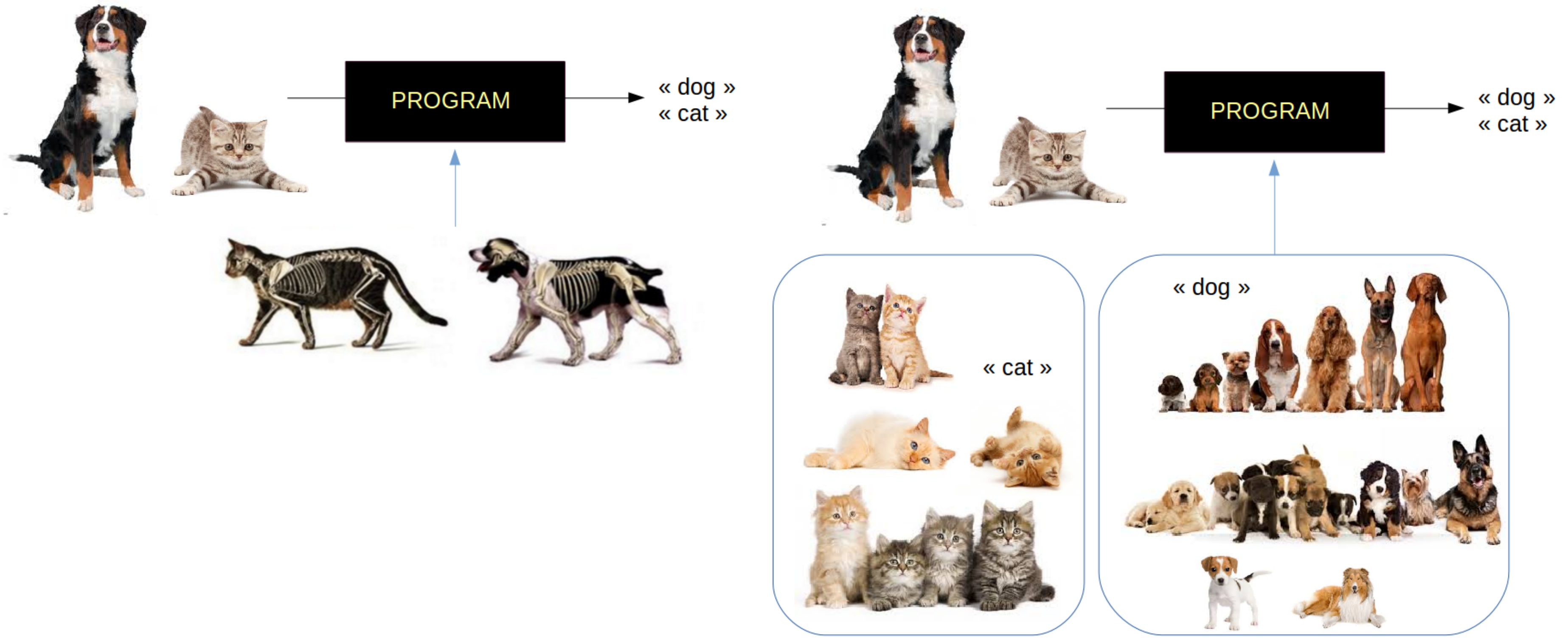
E.g medical diagnosis : $\boldsymbol{x}$ is the list of symptoms, y is the diagnosis (classification)

E.g finance : $\boldsymbol{x}$ is the history of stock's prices, y is the prediction of stock price for the next day/week/month (regression)

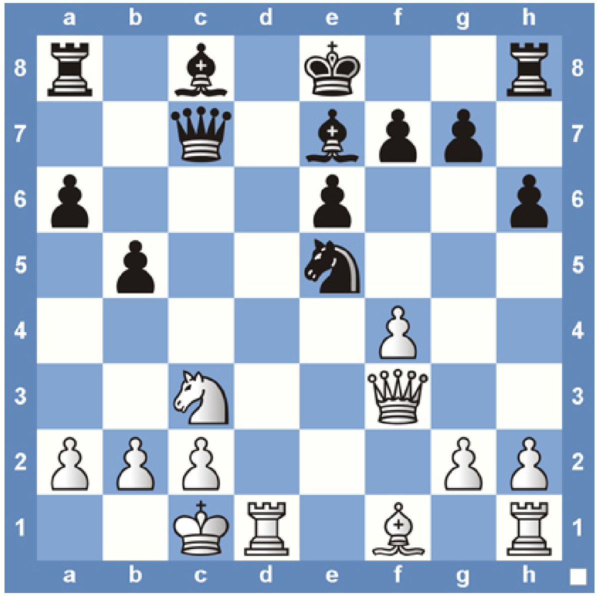
A prediction or a model is a function $m : \mathcal{X} \to \mathcal{Y}$.

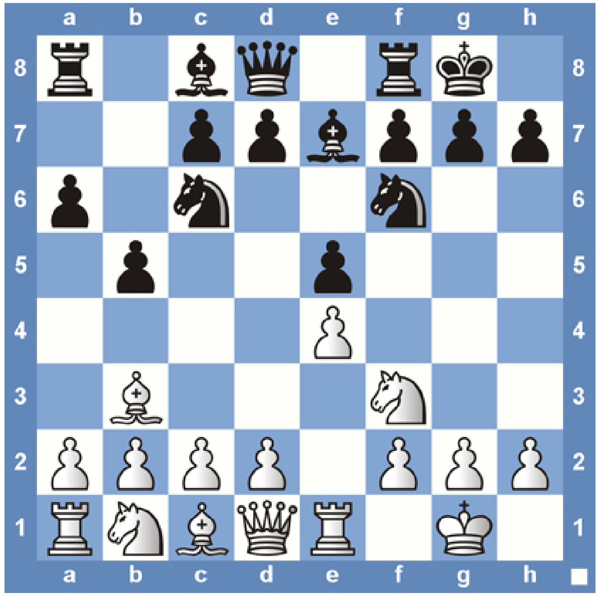
Big Data and Black Box Models

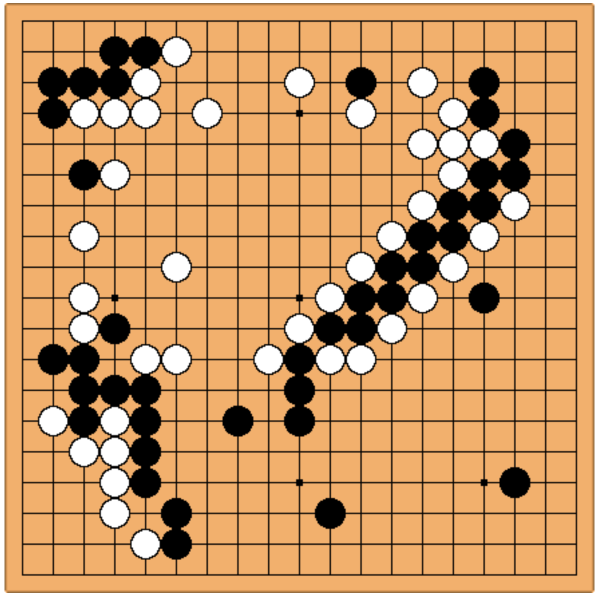
Historically, models were based on a rule-based approach ! (labor intensive to build, and hardly extendable to other situations)



Big Data and Black Box Models







Deep Blue, February 1996, 10^{47} board positions in chess, Campbell *et al.* (2002)

AlphaGo, March 2016, 10^{170} board positions in Go Silver *et al.* (2017) (19×19)

Artificial Intelligence : the End of Models ?

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse,

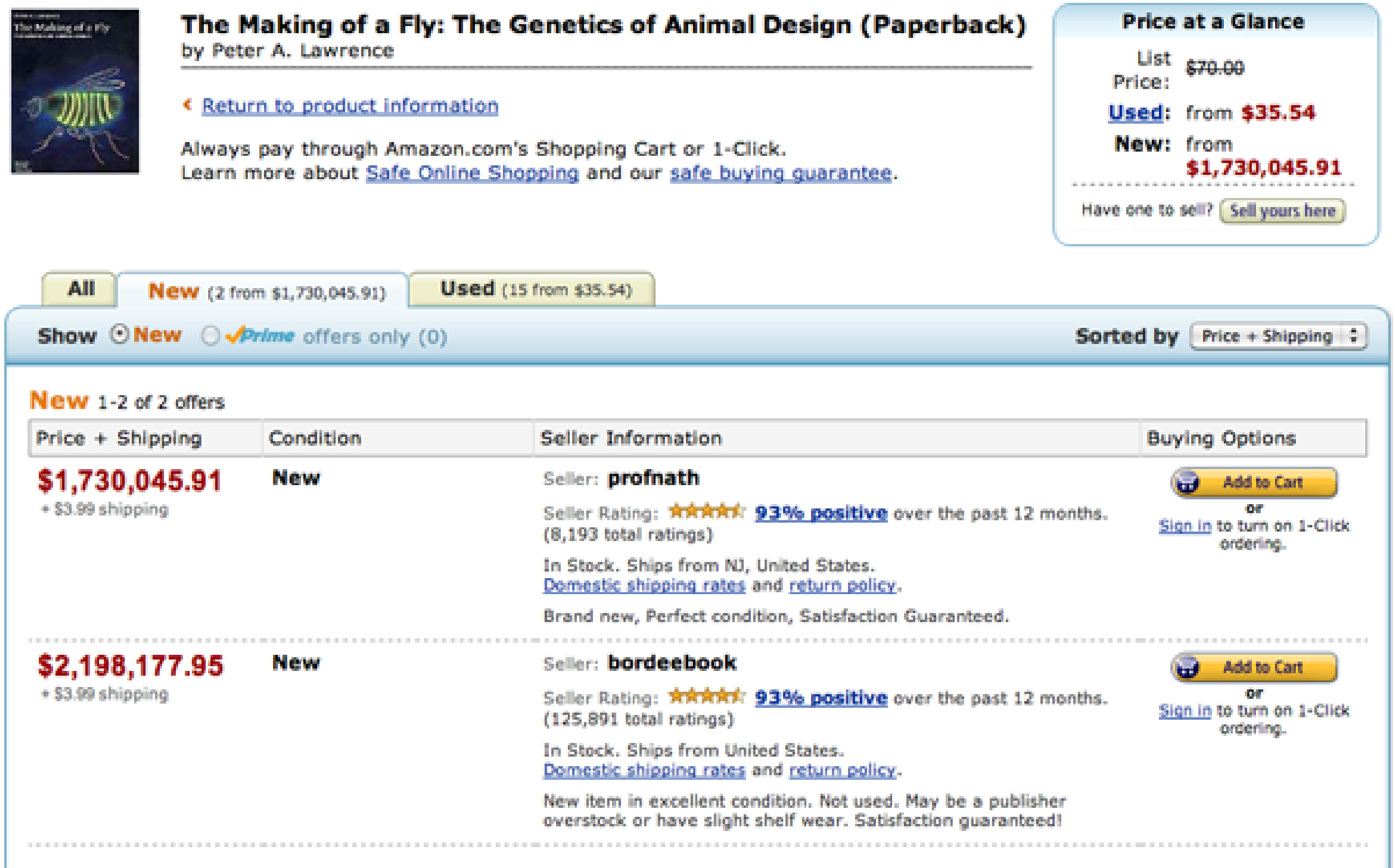
"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and

Chris Anderson (2008, The Data Deluge Makes the Scientific Method Obsolete)

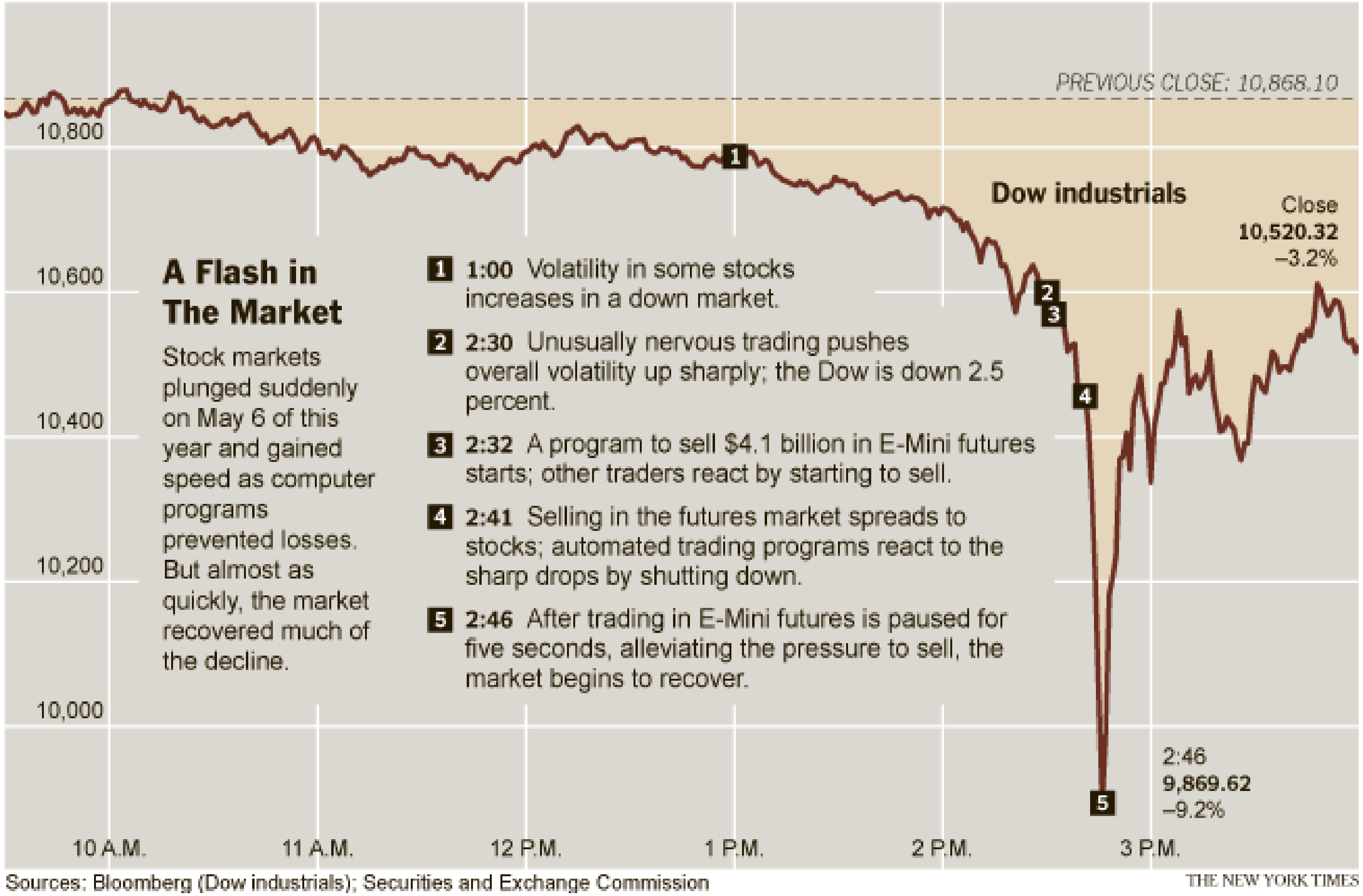
"Models" (or "Algorithms") can go wrong

See Amazon's prices, Eisen (2011, Amazon's $23,698,655.93 book about flies)



"Models" (or "Algorithms") can go wrong

See 2010 Flash Crash (see wikipedia)



(Mathematical) Statistics

Consider observations $\{y_1, \cdots, y_n\}$ from iid random variables $Y_i \sim F_{\boldsymbol{\theta}}$ (with "density" $f_{\boldsymbol{\theta}}$).
Likelihood is

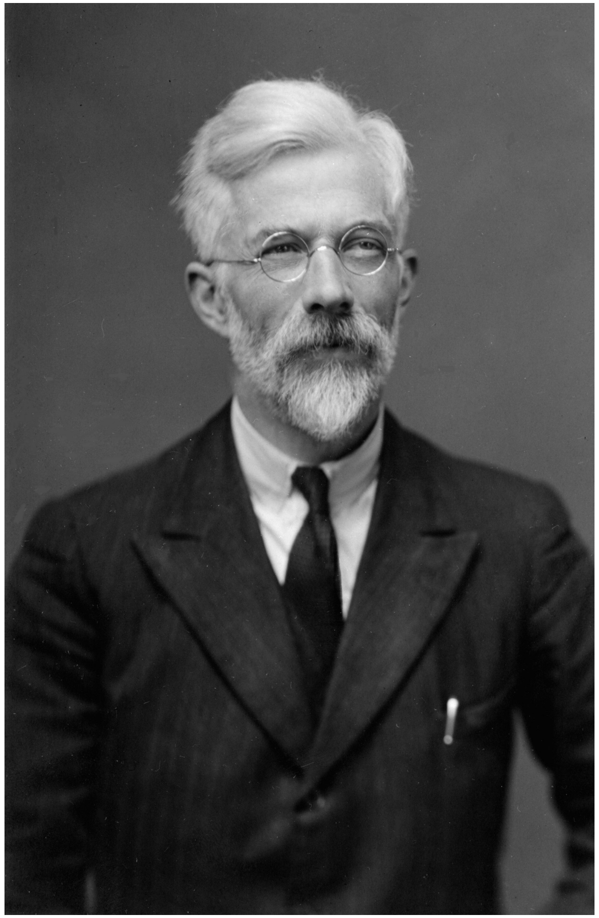
$$\mathcal{L}(\boldsymbol{\theta}\,; \boldsymbol{y}) \mapsto \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i)$$

Maximum likelihood estimate is

$$\widehat{\boldsymbol{\theta}}_n^{\text{mle}} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\{\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})\}$$

and, if $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^{\text{mle}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$$

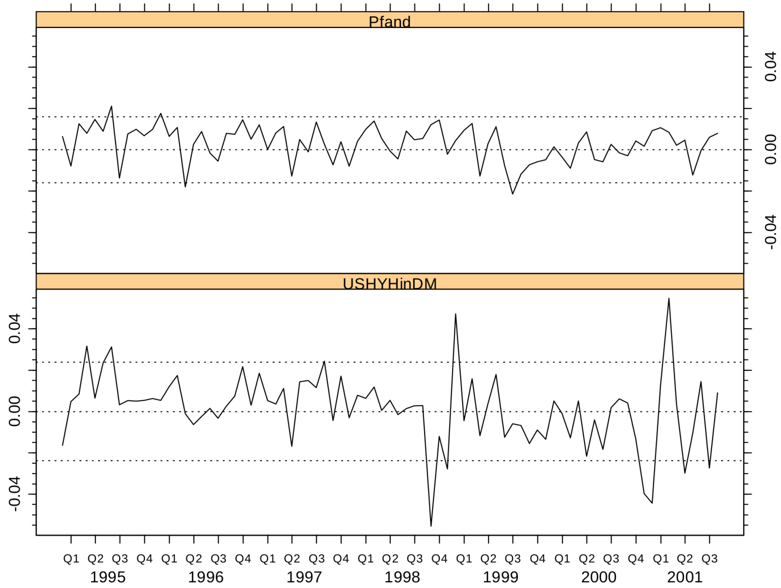


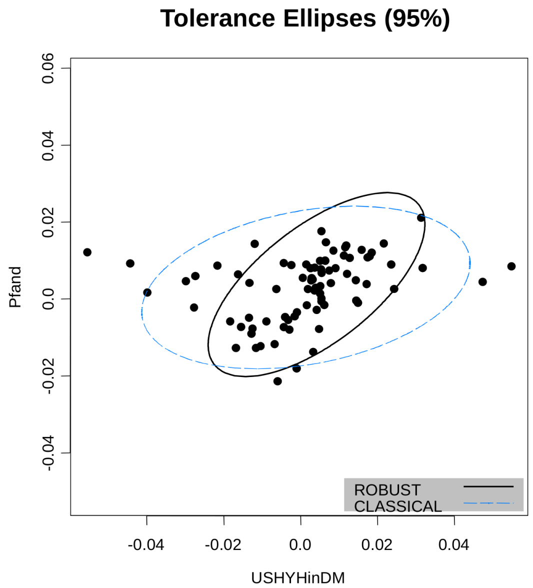
Fisher (1912, On an absolute criterion for fitting frequency curves).

Can we use statistical models in practice ?

Robust inference is important in real life applications (see Hubert, Rousseeuw & van Aelst (2004, Robustness))

How robust is that estimator ? See Martin (2014) on financial time series

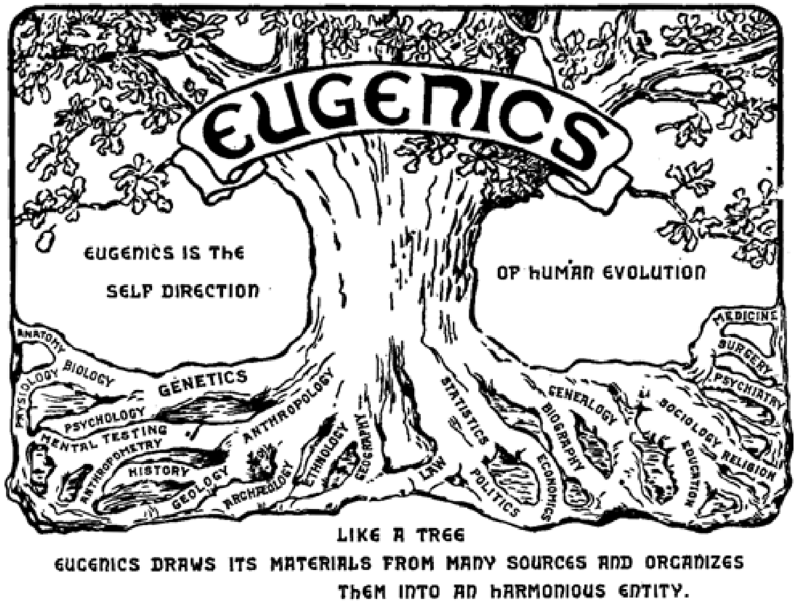


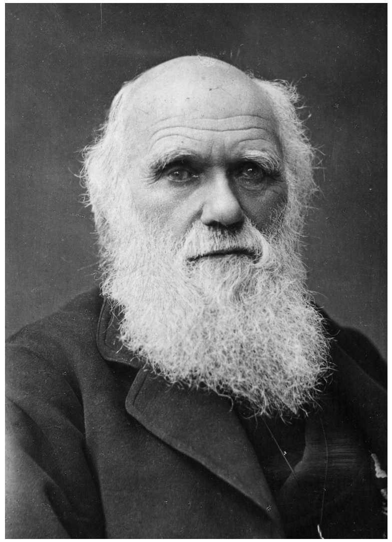
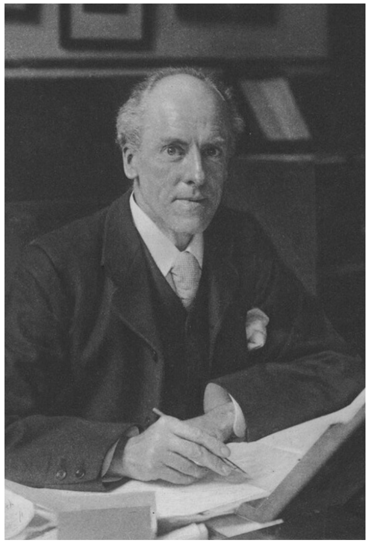
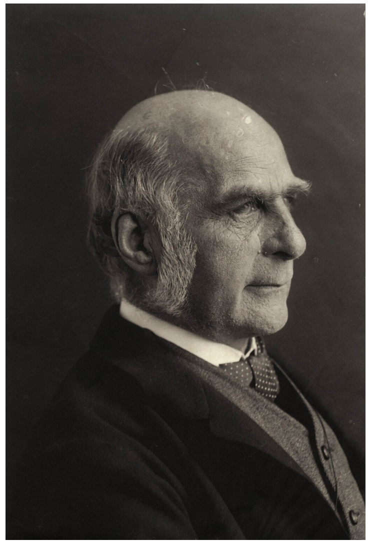


MLE (or classical) correlation estimator $\widehat{\theta} \sim 30\%$ while $\widehat{\theta}^{\text{mcd}} \sim 65\%$

(Linear) Regression ?

Adrien-Marie Legendre (1752-1833) least squares
Charles Darwin (1809-1882)
Francis Galton (1822-1922) regression
Karl Pearson (1857-1936) correlation



Predictive Models & Algorithms

Galton (1870, Heriditary Genius, 1886, Regression towards mediocrity in hereditary stature) and Pearson & Lee (1896, On Telegony in Man, 1903 On the Laws of Inheritance in Man) studied genetic transmission of characterisitcs, e.g. the heigth.

On average the child of tall parents is taller than other children, but less than his parents.

"*I have called this peculiarity by the name of regression*", Francis Galton, 1886.

Regression *towards* Mediocrity *in* Hereditary Stature.
By Francis Galton, F.R.S., &c.

Table 8.1. Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the mid-parent in inches	Height of the adult child														Total no. of adult children	Total no. of mid-parents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

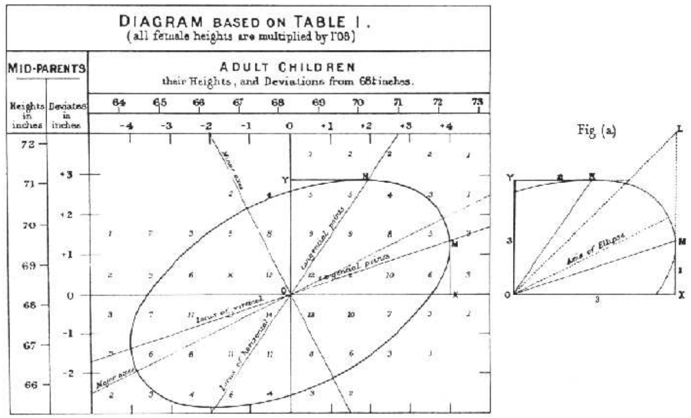


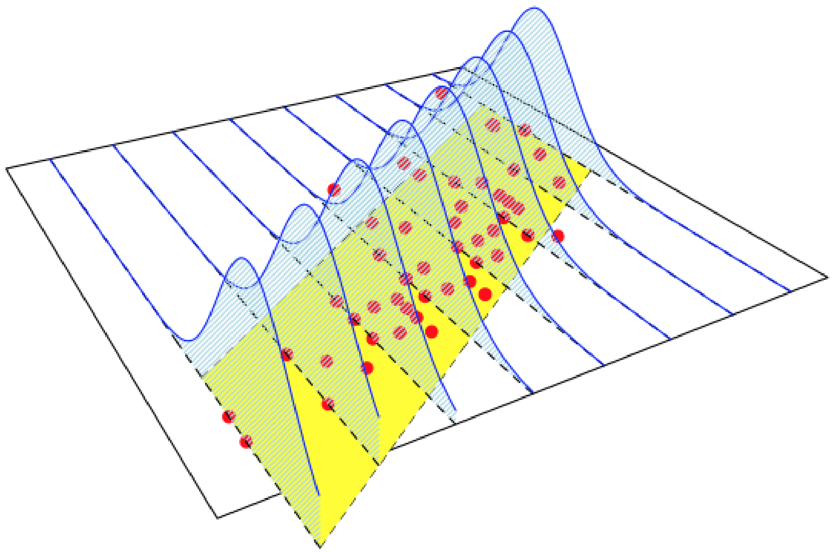
Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)

Predictive Models & Algorithms, $y \in \mathbb{R}$

$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
so that $\mathbb{E}[Y_i] = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

$$\widehat{\boldsymbol{\beta}}_n^{\text{ols}} = \text{argmin}\left\{\sum_{i=1}^n \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2\right\}$$

and prediction $\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_n^{\text{ols}}$.
Observe that $\widehat{\boldsymbol{\beta}}_n^{\text{mle}} = \widehat{\boldsymbol{\beta}}_n^{\text{ols}}$



Predictive Models & Algorithms, $y \in \{0,1\}$

Reed & Berkson (1929, The Application of the Logistic Function to Experimental Data)

Assume that $\mathbb{P}(Y_i = 1) = \pi_i$,

$$\text{logit}(\pi_i) = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta},\ \text{ where } \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right),$$

or

$$\pi_i = \text{logit}^{-1}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) = \frac{\exp[\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}]}{1+\exp[\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}]}.$$

see Classification from scratch, logistic regression)

Predictive Models & Algorithms, $y \in \{0,1\}$

Bliss (1934, The method of probits) suggested

$$\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) = H(\boldsymbol{x}^\top\boldsymbol{\beta}) \text{ where } H(\cdot) = \Phi(\cdot)$$

the c.d.f. of the $\mathcal{N}(0,1)$ distribution. This is the probit model.

This yields a latent model, $y_i = \mathbf{1}(y_i^\star > 0)$ where

$$y_i^\star = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \varepsilon_i \text{ is a non-observable score.}$$

In the logistic regression, we model the odds ratio,

$$\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y\neq 1|\boldsymbol{X}=\boldsymbol{x})} = \exp[\boldsymbol{x}^\top\boldsymbol{\beta}]$$

$$\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) = H(\boldsymbol{x}^\top\boldsymbol{\beta}) \text{ where } H(\cdot) = \frac{\exp[\cdot]}{1+\exp[\cdot]}$$

which is the c.d.f. of the logistic variable, see Verhulst (1845)

Table 3.2 Transformation of percentages to probits

%	0	1	2	3	4	5	6	7	8	9
0	—	2.67	2.95	3.12	3.25	3.36	3.45	3.52	3.59	3.66
10	3.72	3.77	3.82	3.87	3.92	3.96	4.01	4.05	4.08	4.12
20	4.16	4.19	4.23	4.26	4.29	4.33	4.36	4.39	4.42	4.45
30	4.48	4.50	4.53	4.56	4.59	4.61	4.64	4.67	4.69	4.72
40	4.75	4.77	4.80	4.82	4.85	4.87	4.90	4.92	4.95	4.97
50	5.00	5.03	5.05	5.08	5.10	5.13	5.15	5.18	5.20	5.23
60	5.25	5.28	5.31	5.33	5.36	5.39	5.41	5.44	5.47	5.50
70	5.52	5.55	5.58	5.61	5.64	5.67	5.71	5.74	5.77	5.81
80	5.84	5.88	5.92	5.95	5.99	6.04	6.08	6.13	6.18	6.23
90	6.28	6.34	6.41	6.48	6.55	6.64	6.75	6.88	7.05	7.33
—	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
99	7.33	7.37	7.41	7.46	7.51	7.58	7.65	7.75	7.88	8.09

Soit p la population : représentons par dp l'accroissement infiniment petit qu'elle reçoit pendant un temps infiniment court dt. Si la population croissait en progression géométrique, nous aurions l'équation $\frac{dp}{dt} = mp$. Mais comme la vitesse d'accroissement de la population est retardée par l'augmentation même du nombre des habitans, nous devrons retrancher de mp une fonction inconnue de p; de manière que la formule à intégrer deviendra

$$\frac{dp}{dt} = mp - \varphi(p).$$

L'hypothèse la plus simple que l'on puisse faire sur la forme de la fonction φ, est de supposer $\varphi(p) = np^2$. On trouve alors pour intégrale de l'équation ci-dessus

$$t = \frac{1}{m}\left[\log.\, p - \log.(m-np)\right] + \text{constante},$$

et il suffira de trois observations pour déterminer les deux coefficiens constans m et n et la constante arbitraire.

En résolvant la dernière équation par rapport à p, il vient

$$p = \frac{mp'\, e^{mt}}{np'\, e^{mt} + m - np'} \quad \ldots\ldots \quad (1)$$

en désignant par p' la population qui répond à $t = o$, et par e la base des logarithmes népériens. Si l'on fait $t = \infty$, on voit que la valeur de p correspondante est $P = \frac{m}{n}$. Telle est donc *la limite supérieure de la population.*

From Score Functions to 0/1 Classifier

$\boldsymbol{x} \mapsto \text{logit}^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta})$ and $\boldsymbol{x} \mapsto \Phi(\boldsymbol{x}^\top\boldsymbol{\beta})$ are called score functions.

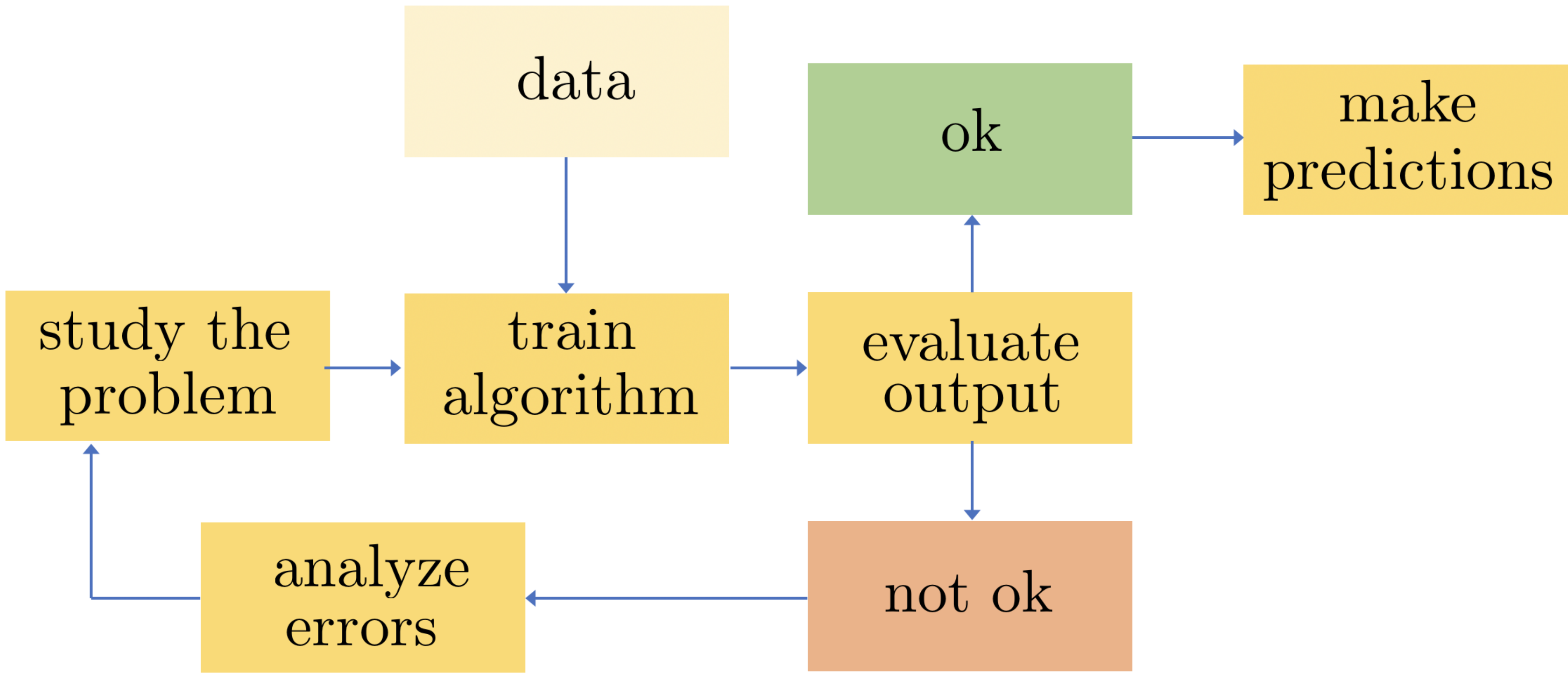
The score is interpreted as the probability that y takes value $+1$.

To go from a score to a class:

if $s(\boldsymbol{x}) > s$, then $\widehat{Y}(\boldsymbol{x}) = 1$ and $s(\boldsymbol{x}) \leq s$, then $\widehat{Y}(\boldsymbol{x}) = 0$

Plot $TP(s) = \mathbb{P}[\widehat{Y} = 1|Y = 1]$ against $FP(s) = \mathbb{P}[\widehat{Y} = 1|Y = 0]$

Predictive Models & Algorithms : The Process



Predictive Models & Algorithms : Machine Learning

Models : classification (binary or multi-class) and regression

Training data $(y_i, \boldsymbol{x}_i)$ that can be text documents, time series, image files, sound recordings, DNA sequences, etc. transformed into inputs in $\mathbb{R}^d$.

Feature extraction: (arbitrary) mapping from raw input to inputs in $\mathbb{R}^d$

E.g one-hot encoding (or dummy variable encoding)

To "estimate" and evaluate a prediction function, use a loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$,

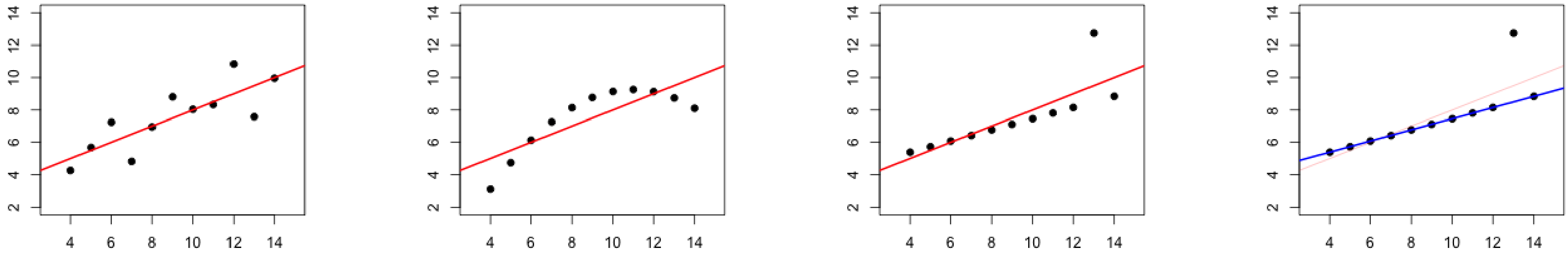
$$m^\star = \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(m(\boldsymbol{x}_i), y_i) \right\}$$

E.g. in classification, 0/1 loss (1 prediction is wrong, 0 prediction is correct)

E.g. in regression, square loss (or ℓ_2), (predicted – target)^2

Importance of Loss Function

What do you want to predict ? Least squares = “on average”



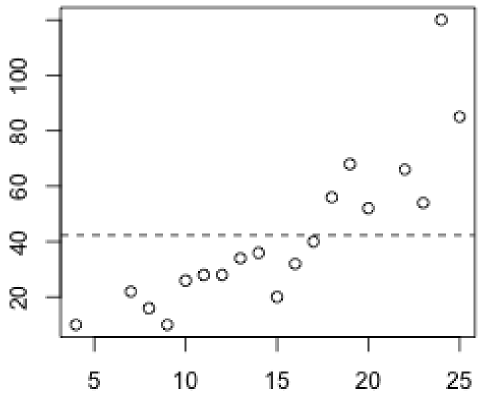
one can consider the least absolute deviation = “mediane”

Error is evaluated on a new dataset (the test or validation dataset).

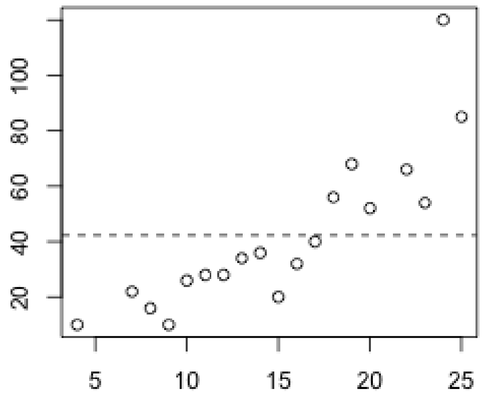
Split data into train and test.

Complexity and overfit

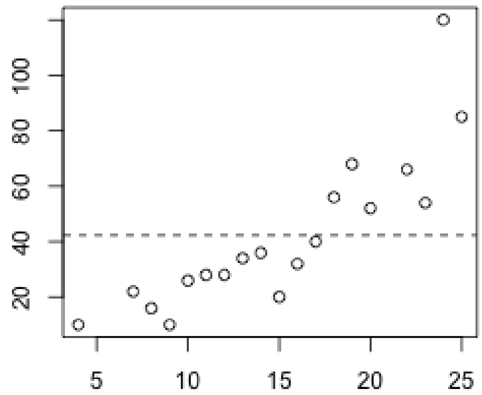
Without any explanatory variable, $\widehat{y} = \overline{y} = \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\sum_{i=1}^{n}(y_i - m)^2\right\}$



kernel-based regression



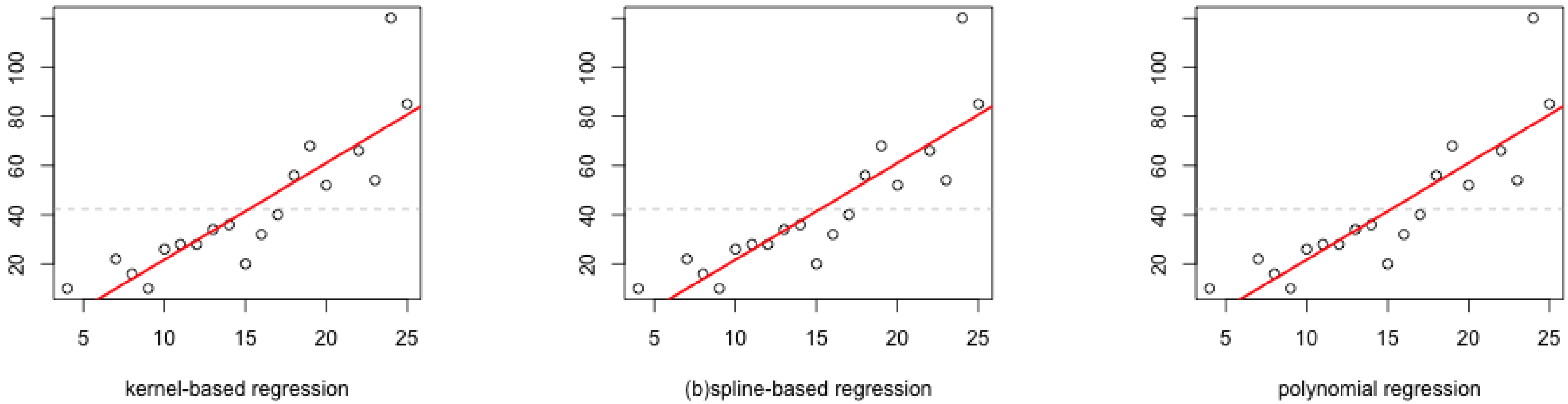
(b)spline-based regression



polynomial regression

Complexity and overfit

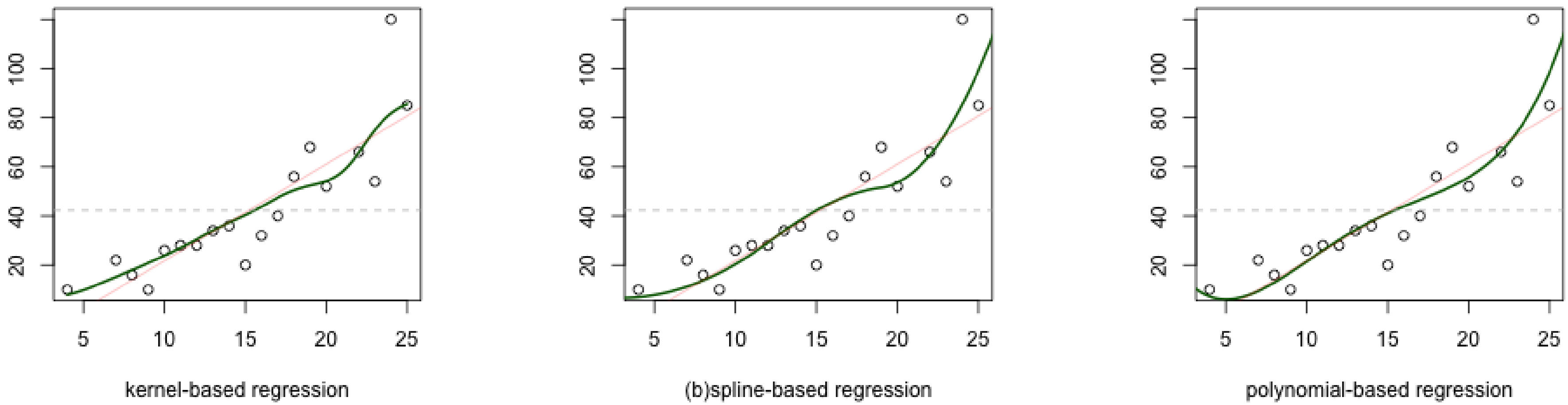
With one explanatory variable, $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$



the classical "linear regression" (from Sir Francis Galton in 1886)

Complexity and overfit

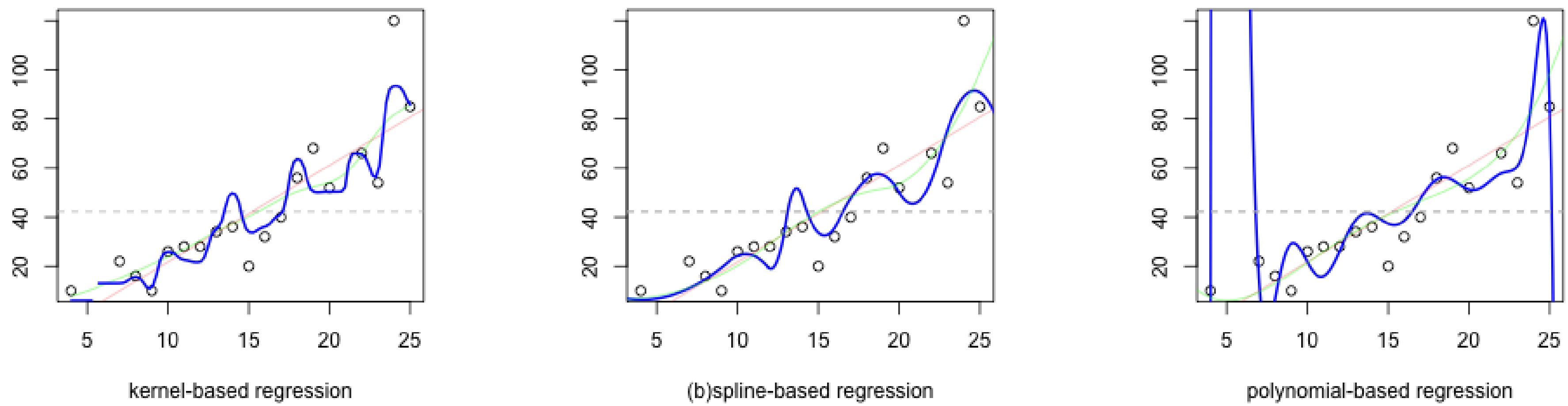
With one explanatory variable, $\widehat{y} = m(x)$



kernel based is a local regression (average on the neighborhood of x), while splines and polynomial are basis of functions (to approximate m)

Complexity and overfit

With one explanatory variable, $\widehat{y} = m(x)$, overfitting

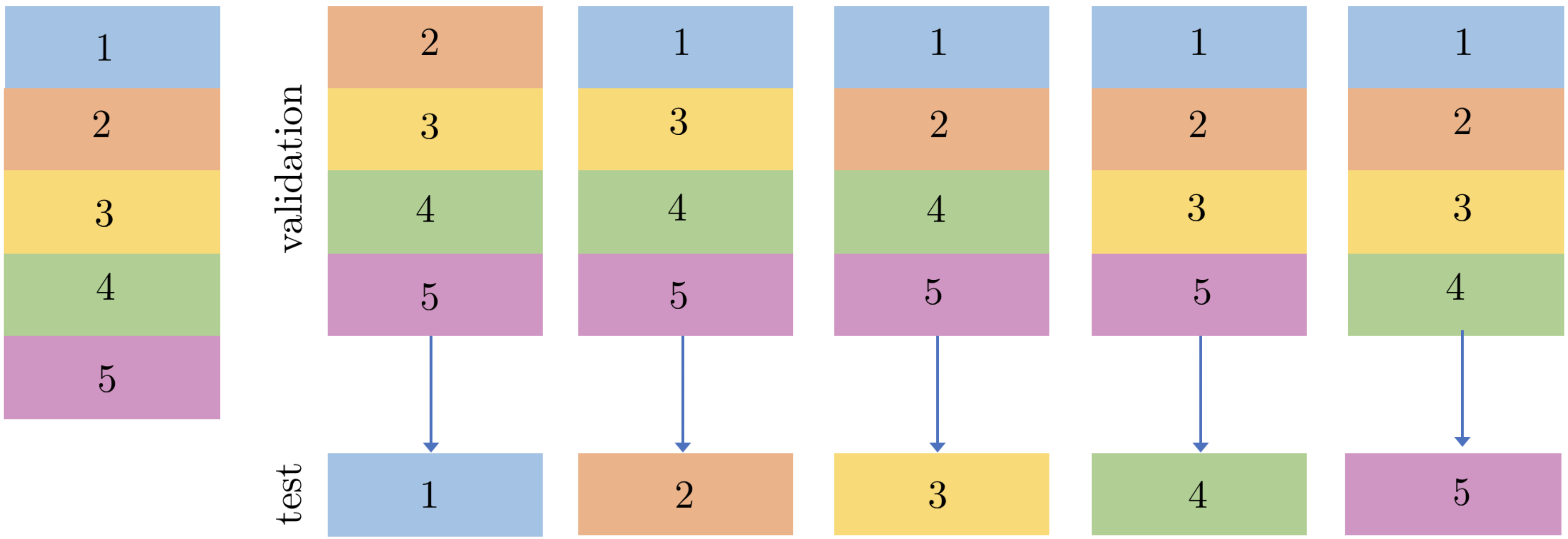


i.e. good (great) fit on the training dataset, but hard to imagine that it be generalized on new data...

Cross Validation

To ovoid oferfit, use leave-one-out or k-fold cross validation.

randomly partition data into k "folds" (of equal size), and train model



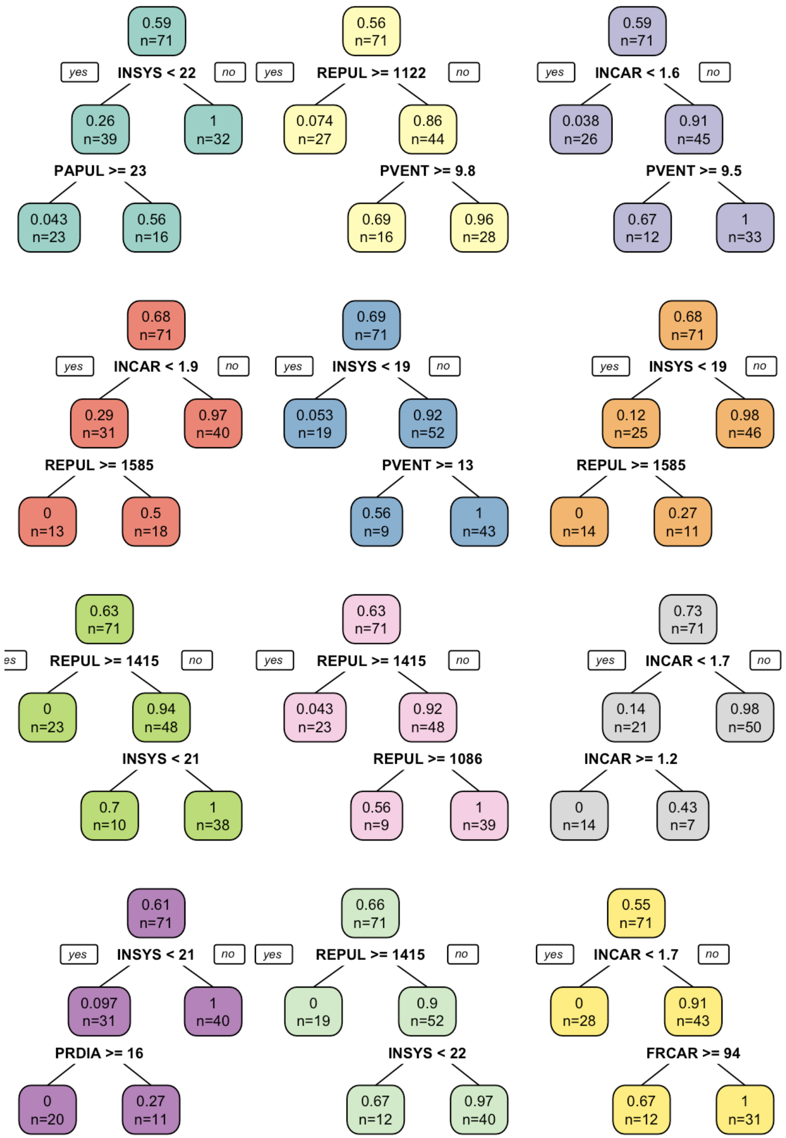
Ensemble Methods

"Ensemble learning" (see wikipedia)
"aggregation methods" or "stacking"
is combining predictive methods

$$m(\boldsymbol{x}) = \sum_{j=1}^{k} \widehat{m}_k(\boldsymbol{x})$$

See random forests, or bagging
(bootstrap + aggregation)

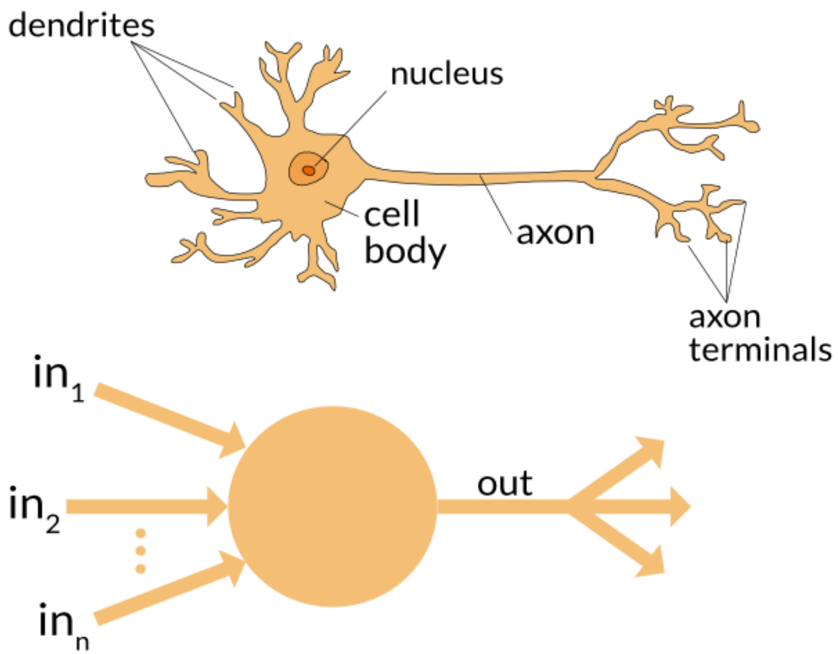
Classification from scratch, trees
and forests

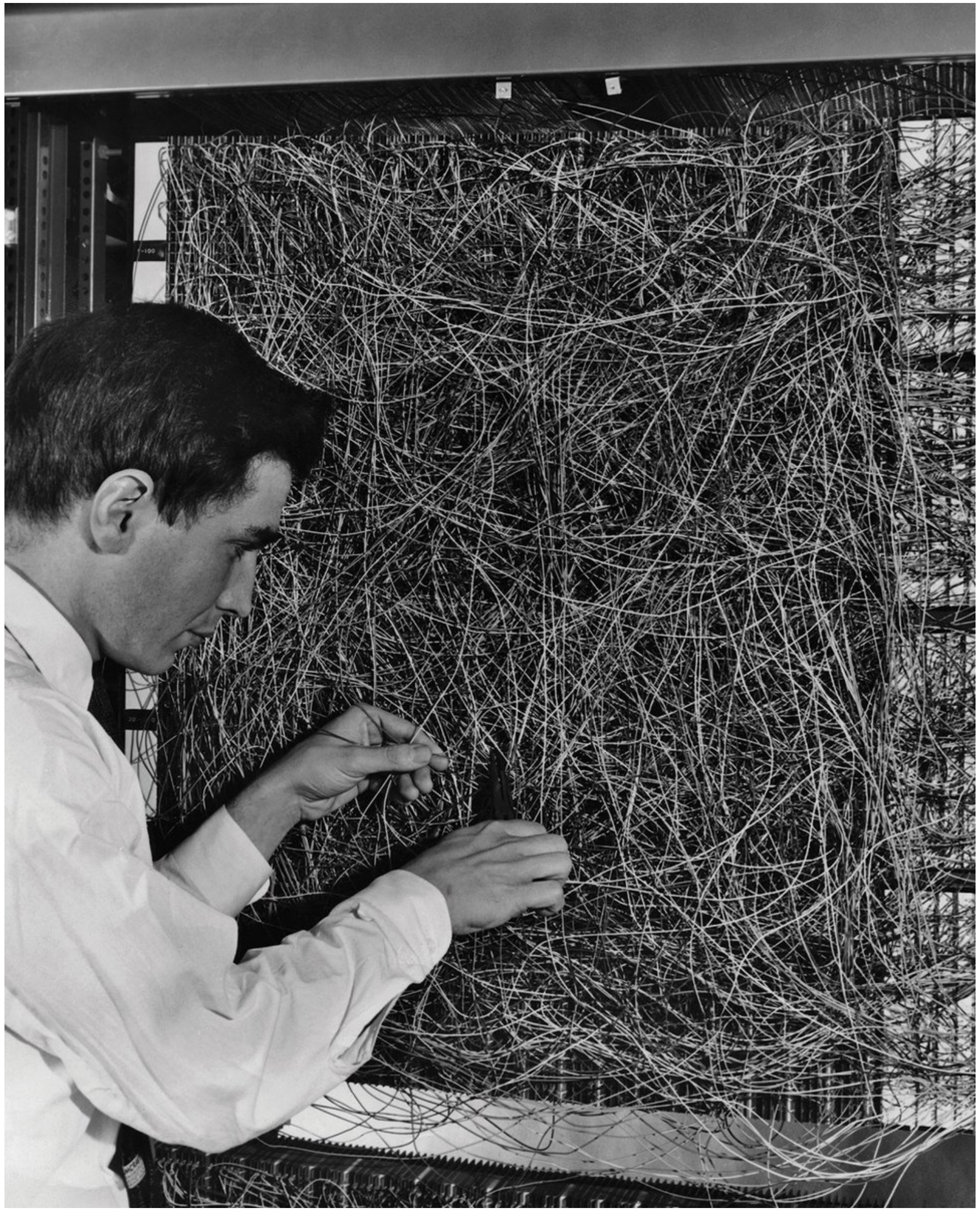


Neural Networks

Rosenblatt (1958, The Perceptron)

see Classification from scratch, neural nets





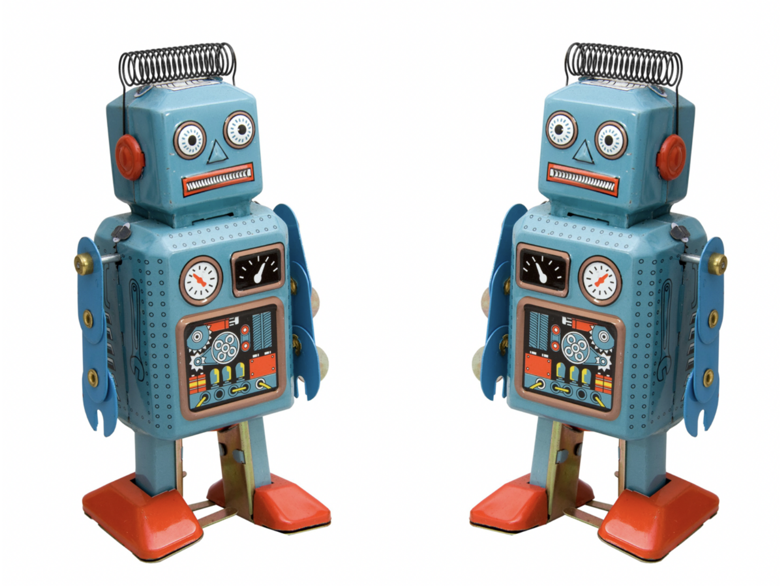
Incremental Algorithms and Reinforcement learning

Incremental algorithms (see wikipedia) allow a model to be updated using new observations as they arrive, without having to reprocess old ones.

We are talking about "data stream analysis", "on-line" learning (see wikipedia) as opposed to "batch" or "offline learning".

Useful for data streams and data too massive to be processed in their entirety

Reinforcement learning (see wikipedia)
Alphago: learned from a base of tens of thousands of games and 30 million moves, then by playing against himself (learning by reinforcement)



Data & Models : Selection Bias

We cannot differentiate data and model that easily...

"*After an operation, should I stay at hospital, or go back home ?*"

as in Angrist & Pischke (2008, Mostly Harmless Econometrics),

(health | hospital) − (health | stayed home) [observed]

should be written

(health | hospital) − (health | *had* stayed home) [treatment effect]

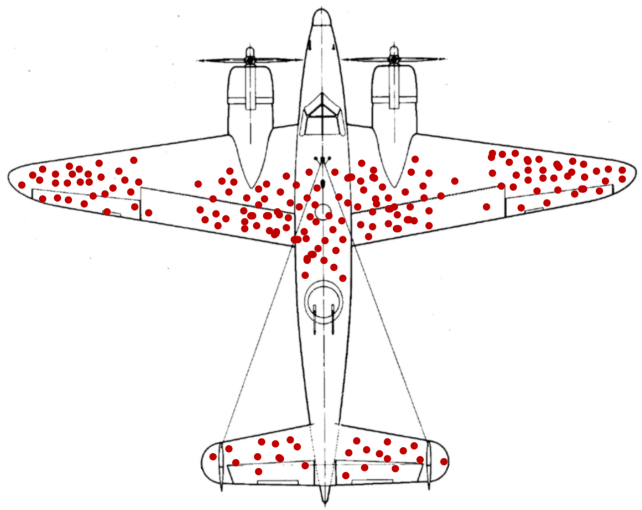
+ (health | *had* stayed home) − (health | stayed home) [selection bias]

Need randomization to solve selection bias.

see also Ioannidis (2005, Why Most Published Research Findings Are False).

Data & Models : Selection Bias

also called “survivor bias” (see wikipedia)
how to minimize bomber losses to enemy fire ?
study of the damage done to aircraft that had returned from missions



Which area should the Navy reinforce areas ?

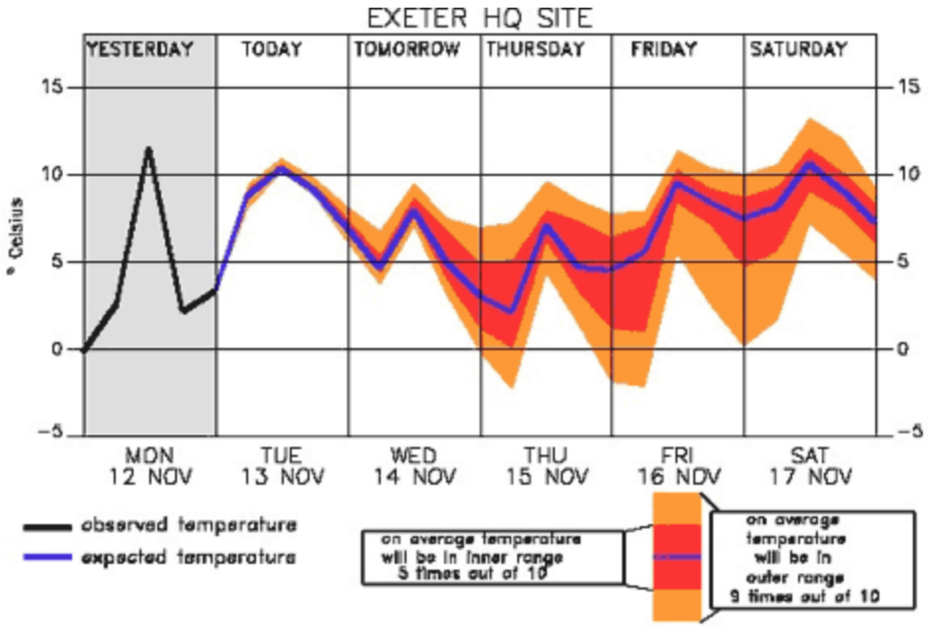
see Mangel & Samaniego (1984, Abraham Wald’s Work on Aircraft Survivability).

Probabilistic Forecasts

So far we did use ℓ on $\mathcal{Y}\times\mathcal{Y}$: $\ell(\widehat{y},y)$ which measure the difference between prediction $\widehat{y}$ and realization y

consider a propensity score (meteorology) $s(\widehat{F},y)$ where $\widehat{F}$ is the distribution of $\widehat{y}$

Used for time series, $s({}_{t-1}\widehat{F}_t,y_t)$ between realization y_t



and the forecast distribution, obtained at time $t-1$: ${}_{t-1}\widehat{F}_t=\mathbb{P}[Y_t|y_{t-1}]$

see Gigerenzer *et al.* (2005, "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts?)

Probabilistic Predictions with CitySense

Two use cases

- I'm new to the city : where does everybody hang out at night?
- I know the city : is there anything special going on tonight?

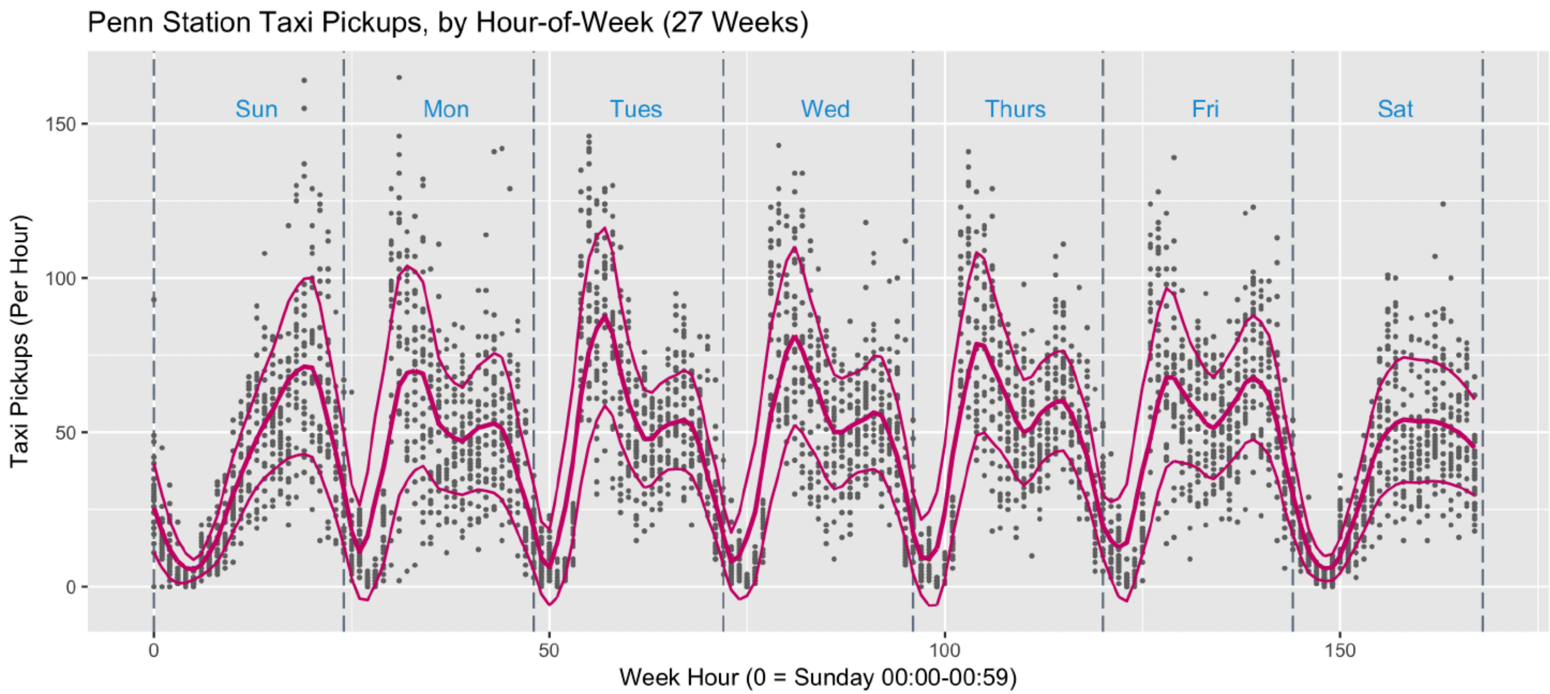
idea: use taxi GPS data in San Francisco & New York City (see Rosenberg (2017)

intuition: Taxi destinations are a proxy for where people are going

need to model "typical" behavior of each area of the city, and then detect the most unusual activities

Probabilistic Predictions with CitySense

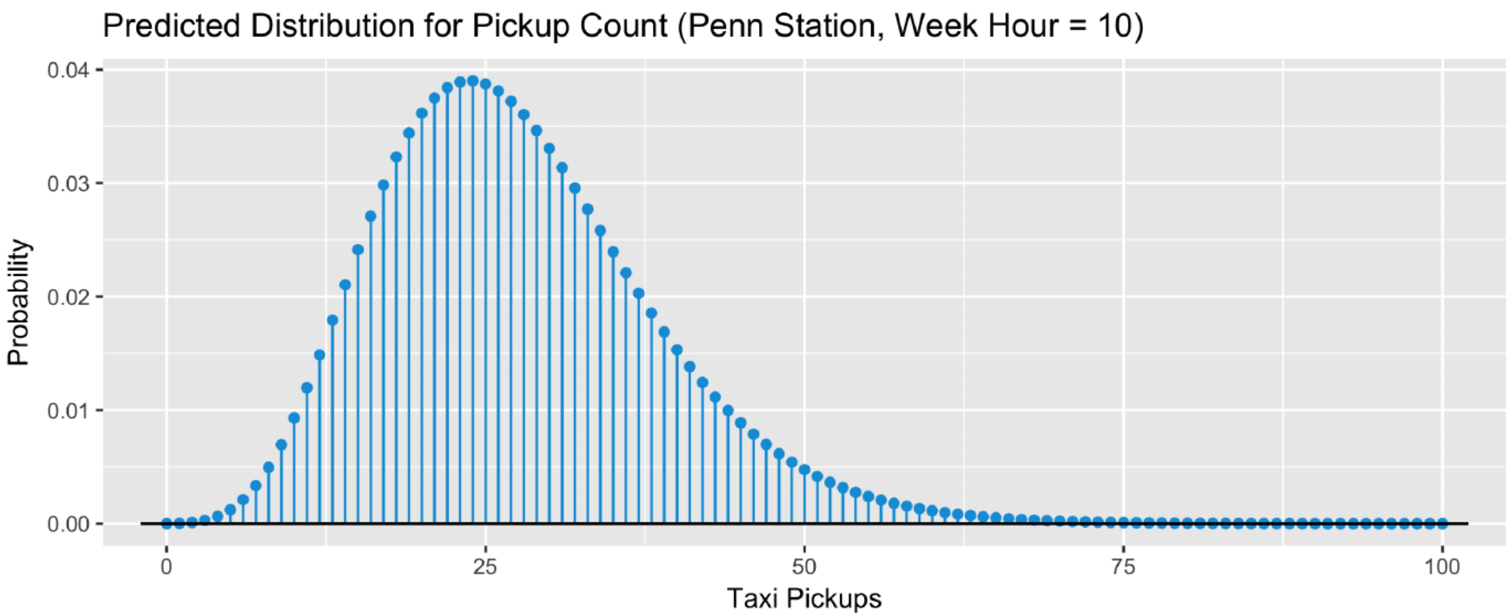
At some give location, given time, how many taxi pickup should you expect ?



y : taxi pickup (per hour), $\boldsymbol{x}$: time, location

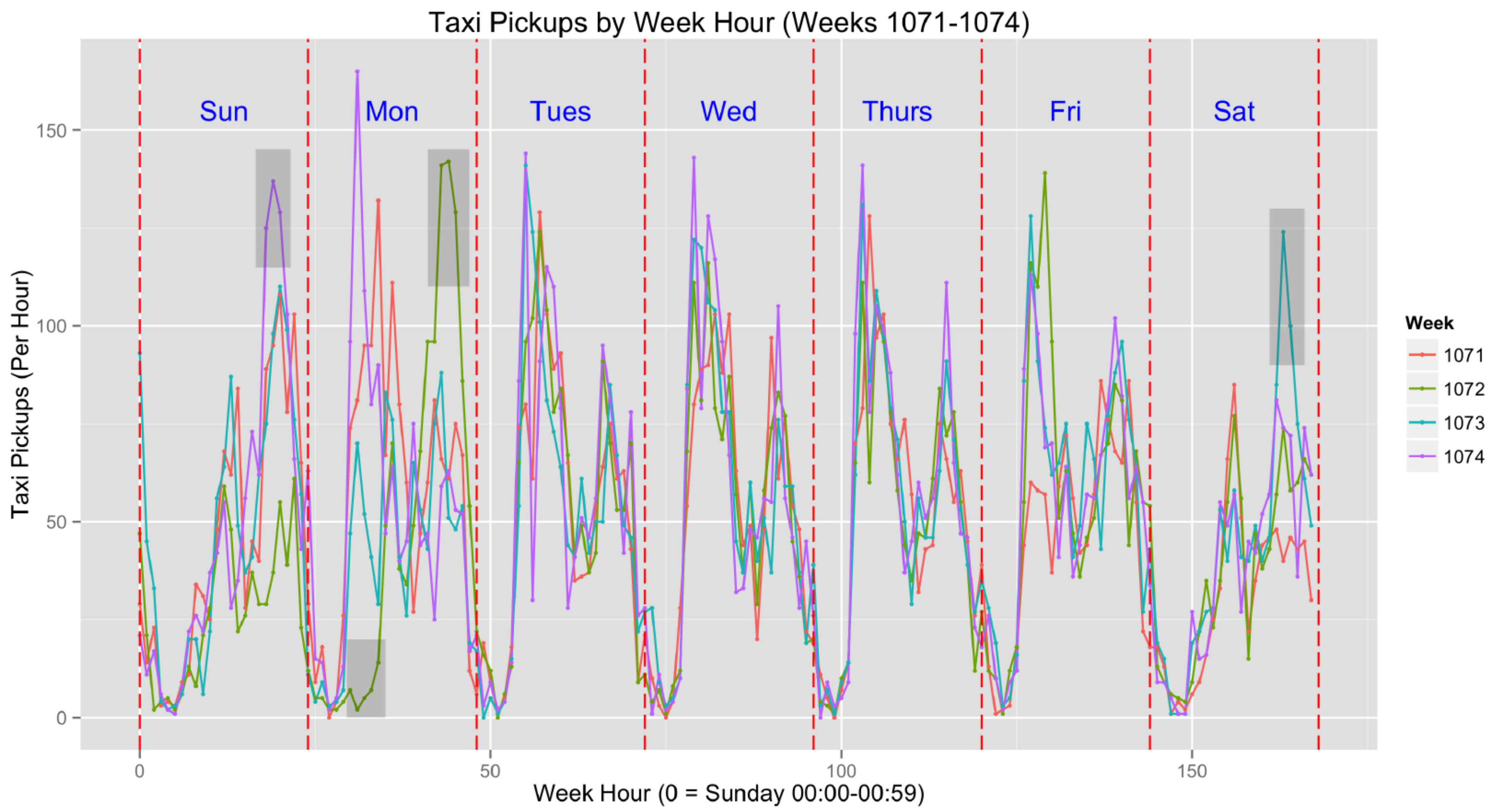
Probabilistic Predictions with CitySense

Consider $\boldsymbol{x}$ (time, location) such that $\mathbb{E}[Y|\boldsymbol{x}] = 30$, with 80% confidence interval $\mathbb{P}[Y \in [18, 42]] = 90\%$. Conditional distribution of $Y|\boldsymbol{x}$ is



what if we observe 90 pikups ? how unusual is it ?

Probabilistic Predictions with CitySense



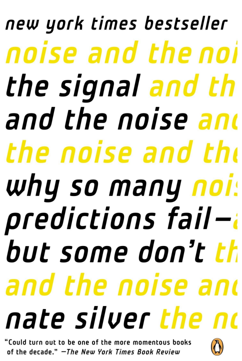
$\boldsymbol{x} = (x_1, x_2)$, x_1 is the time (in the week) and x_2 the grid location

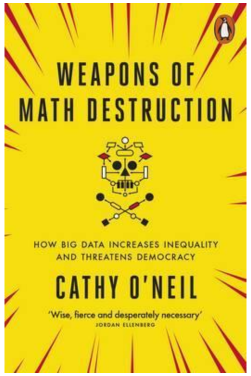
detection of outliers in a spatio-temporal problem

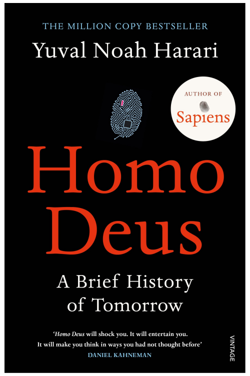
Conclusion

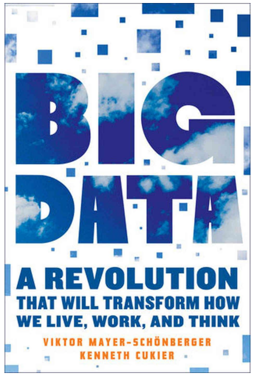
“*Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning*”
Winston Churchill

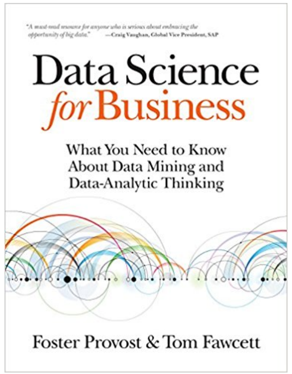
To go further,











References

- Shalev-Shwartz & Ben-David Understanding Machine Learning: From Theory to Algorithms (2014)
- Hastie, Tibshirani & Friedman The Elements of Statistical Learning (2017)
- L'intelligence artificielle dilue-t-elle la responsabilité (2018)
- Les modèles prédictifs peuvent-ils être loyaux et justes (2017)
- L'éthique de la modélisation dans un monde où la normalité n'existe plus (2017)
- Les dérives du principe de précaution (2016)
- Segmentation et mutualisation, les deux faces d'une même pièce (2015)
- La tarification par genre en assurance, corrélation ou causalité ? (2016)
- Big data : passer d'une analyse de corrélation à une interprétation causale (2015)

https://bloomberg.github.io/foml/lectures