Risks of Data Science (What could go wrong?)

Joana Gonçalves de Sá

June 30th 2022

ROADMAP

1. (Some) problems with methods

2. (Some) problems with data

3. Examples of applications gone wrong

4. (Some) possible solutions / approaches

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events

Breast cancer prevalence is quite low, with only 1.4% of women having it

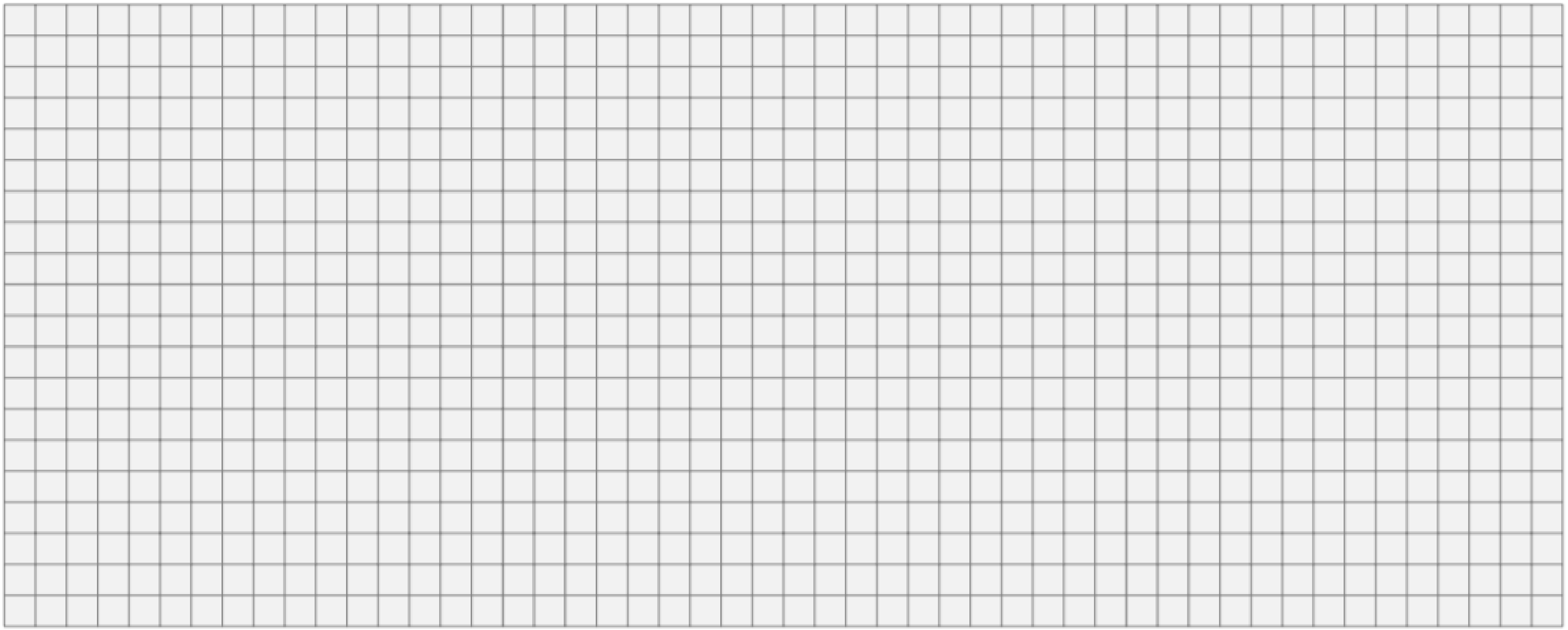
If a woman does not have cancer (NC) the probability of having a positive (+) mammogram is 10%

If a woman has cancer (BC), it will be detected by the mammogram 75% of the time

What is the probability of having cancer, given that the mammogram was positive?

Adapted from The Signal and the Noise, Nate Silver, Penguin Books

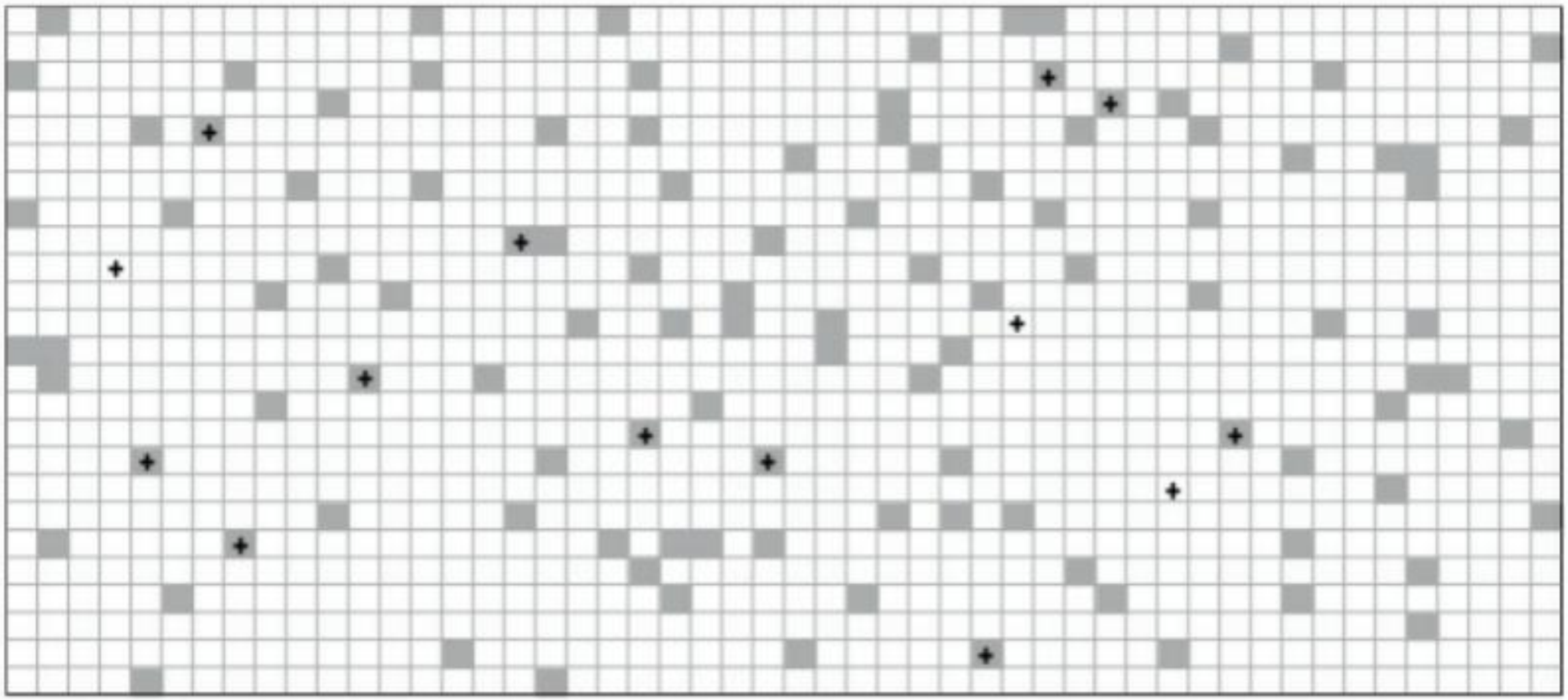
FISHER vs BAYES



1000 hypotheses to test

Statistics for the Big Data Era , Emmanuel Candès, https://www.ljll.math.upmc.fr/IMG/pdf/ljll170314candes.e-mc1-5.5mo.pdf

FISHER vs BAYES



Women with Breast Cancer (14 of 1000)

- Positive mammogram (true positive) (11 of 14)
- Negative mammogram (false negative) (3 of 14)

Women Without Breast Cancer (986 of 1000)

- Positive mammogram (false positive) (99 of 986)
- Negative mammogram (true negative) (887 of 986)

True Positives ~ 79%

Adapted from The Signal and the Noise, Nate Silver, Penguin Books

$$P(A|B)=P(B|A)P(A)/P(B)$$

Breast cancer prevalence is quite low, with only 1.4% of women having it

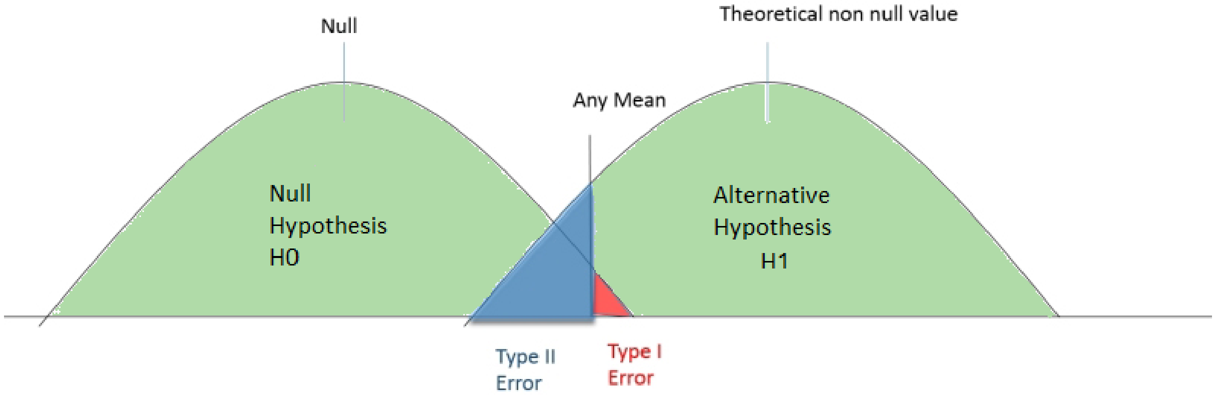
If a woman does not have cancer (NC) the probability of having a positive (+) mammogram is 10%

If a woman has cancer (BC), it will be detected by the mammogram 75% of the time

What is the probability of having cancer, given that the mammogram was positive?

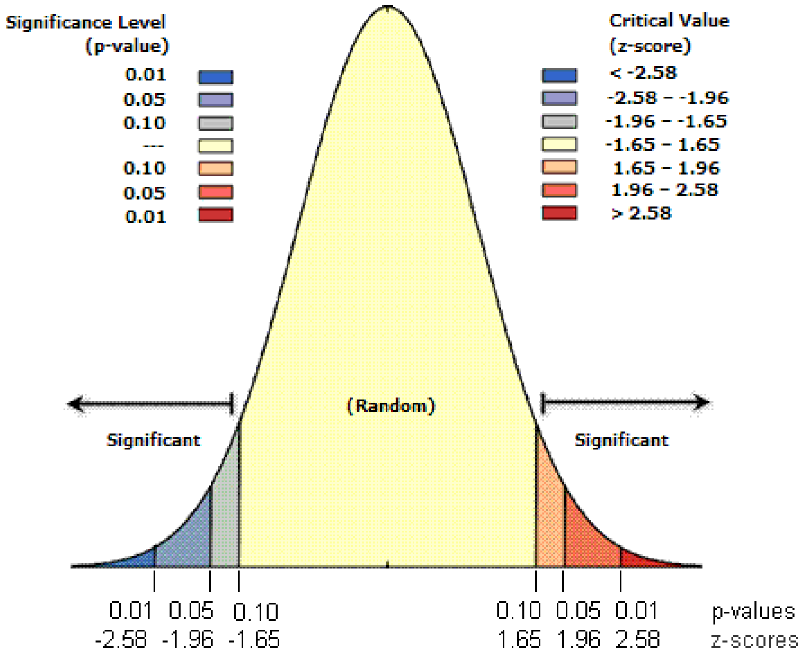
$$P(BC|+)=P(+|BC)P(BC)/P(+) \qquad \sim 0.1$$

FISHER vs BAYES



	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

FISHER vs BAYES



$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

population mean: μ
hypothesized mean: μ_0
population standard deviation: σ
sample mean: $\bar{x}$
sample size: n

You have a randomly selected sample.
The sample is significantly smaller that the population.
The variable in question has a Normal distribution.
We “know” the population standard deviation.

http://blog.minitab.com/blog/statistics-and-quality-data-analysis/large-samples-too-much-of-a-good-thing

P(detecting an effect when there is none) = α

P(detecting an effect when it exists) = $1 - \alpha$

P(detecting an effect when it exists on every experiment k) = $(1 - \alpha)^k$

(k=50)

P(detecting an effect when there is none on at least one experiment) = $1 - (1 - \alpha)^k$

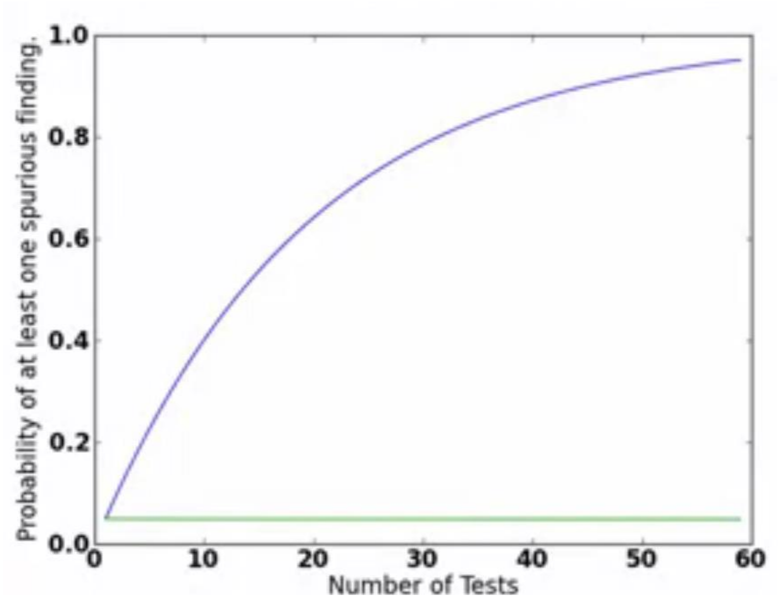
P(detecting an effect when there is none) = α **0.05**

P(detecting an effect when it exists) = $1 - \alpha$ **0.95**

P(detecting an effect when it exists on every experiment k) = $(1 - \alpha)^k$

(k=50) **0.077**

P(detecting an effect when there is none on at least one experiment) = $1 - (1 - \alpha)^k$ **0.932**



Example 1: Sample size = 10

```
Two-sample T Results

       N    Mean  StDev  SE Mean
C4    10   5.011  0.748     0.24
C5    10   5.020  0.803     0.25

Difference = mu (C4) - mu (C5)
Estimate for difference:   -0.009

95% CI for difference:  (-0.741, 0.723)
T-Test of difference = 0 (vs not =): T-Value = -0.03  P-Value = 0.979  DF = 17
```

With 10 observations, the difference (-0.009) is not statistically significant

Example 2: Sample size = 1,000,000

```
Two-sample T Results

           N   Mean  StDev  SE Mean
C1   1000000   5.01   1.00   0.0010
C2   1000000   5.02   1.00   0.0010

Difference = mu (C1) - mu (C2)
Estimate for difference:   -0.00912

95% CI for difference:  (-0.01189, -0.00635)

T-Test of difference = 0 (vs not =): T-Value = -6.45  P-Value = 0.000  DF =1999994
```

With a million observations, the same difference (-0.009) is statistically significant!

"So why did Fisher dismiss the theory? One reason may have been that he was a paid consultant of the tobacco companies. Another may have been that he was a lifelong smoker himself. And Fisher liked to be contrarian and controversial, and disliked anything that smacked of puritanism. In short, he was biased, in a variety of ways."




Should we have hypothesis?

"All who drink of this remedy recover in a short time except those whom it does not help, who all die"

"It is obvious, therefore, that it fails only in incurable cases."

Galen

Randomized control trials
Training and testing
Validation

associations: a study of astrological signs and health

Peter C. Austin, Muhammad M. Mamdani, David N. Juurlink, Janet E. Hux

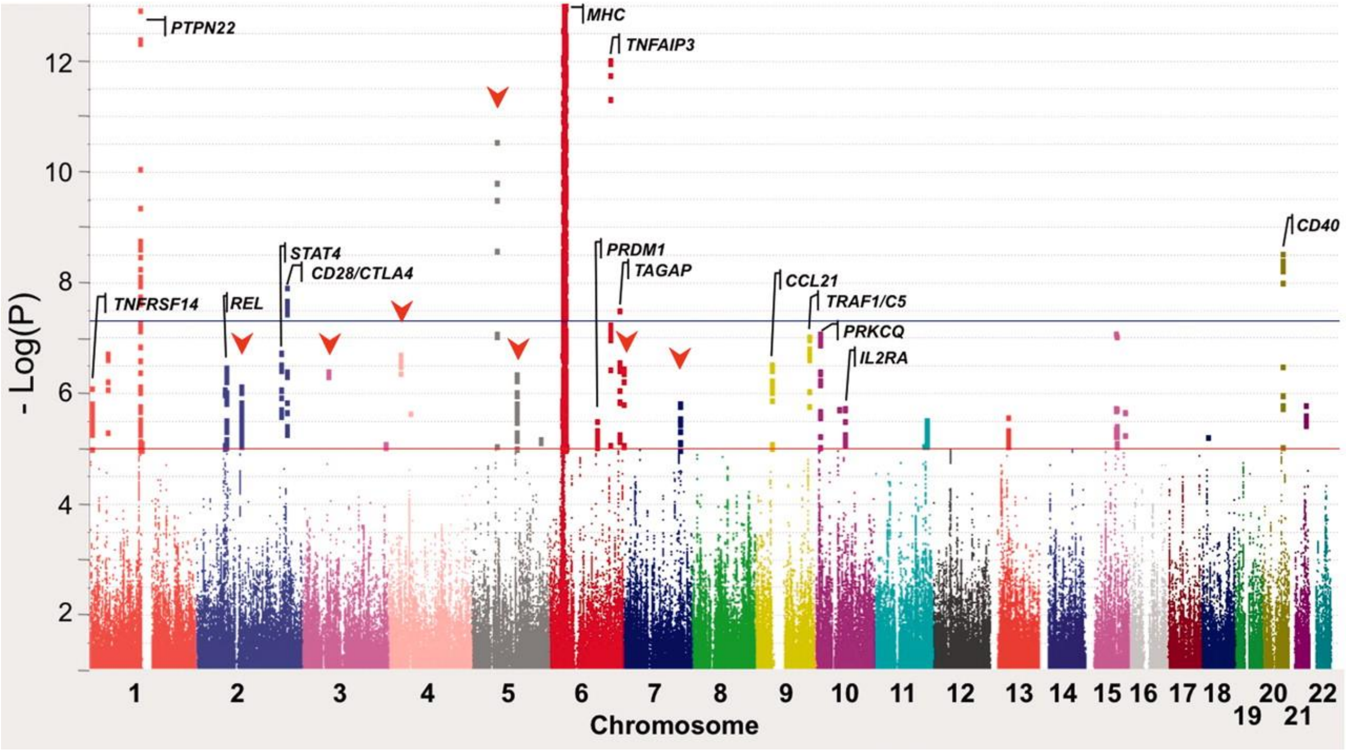
Study Design and Setting

We conducted a study of all 10,674,945 residents of Ontario aged between 18 and 100 years in 2000. Residents were randomly assigned to equally sized derivation and validation cohorts and classified according to their astrological sign. Using the derivation cohort, we searched through 223 of the most common diagnoses for hospitalization until we identified two for which subjects born under one astrological sign had a significantly higher probability of hospitalization compared to subjects born under the remaining signs combined ($P < 0.05$).

Results

We tested these 24 associations in the independent validation cohort. Residents born under Leo had a higher probability of gastrointestinal hemorrhage ($P = 0.0447$), while Sagittarians had a higher probability of humerus fracture ($P = 0.0123$) compared to all other signs combined.

NEW TOOLS, OLD STATS II



Stranger BE, et al., Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics, Genetics, 2011

Manhattan plot for RA GWAS meta-analysis. Statistical strength of association (-Log10P) is plotted against genomic position with the 22 autosomal chromosomes in different colors. The blue horizontal line indicates the genome-wide significance threshold of P = 5 × 10−8; the red line is a threshold for "suggestive" association (P = 10−5). SNPs at 5 of 29 loci known from previous studies (gene symbols shown), and one of the 10 new loci identified in this study (marked by red triangles), achieved genome-wide significance in this meta-analysis (prior to the replication phase of the study). Over 200 SNPs representing 35 loci achieved P <10−5, versus roughly 10 expected by chance.

NEW TOOLS, OLD STATS III

1. We don’t need hypothesis
2. We often don’t have samples
3. Old statistics, new methods

Example 1: Sample size = 10

```
Two-sample T Results

      N    Mean   StDev   SE Mean
C4   10   5.011   0.748      0.24
C5   10   5.020   0.803      0.25

Difference = mu (C4) - mu (C5)
Estimate for difference:   -0.009

95% CI for difference:   (-0.741, 0.723)
T-Test of difference = 0 (vs not =): T-Value = -0.03  P-Value = 0.979  DF = 17
```

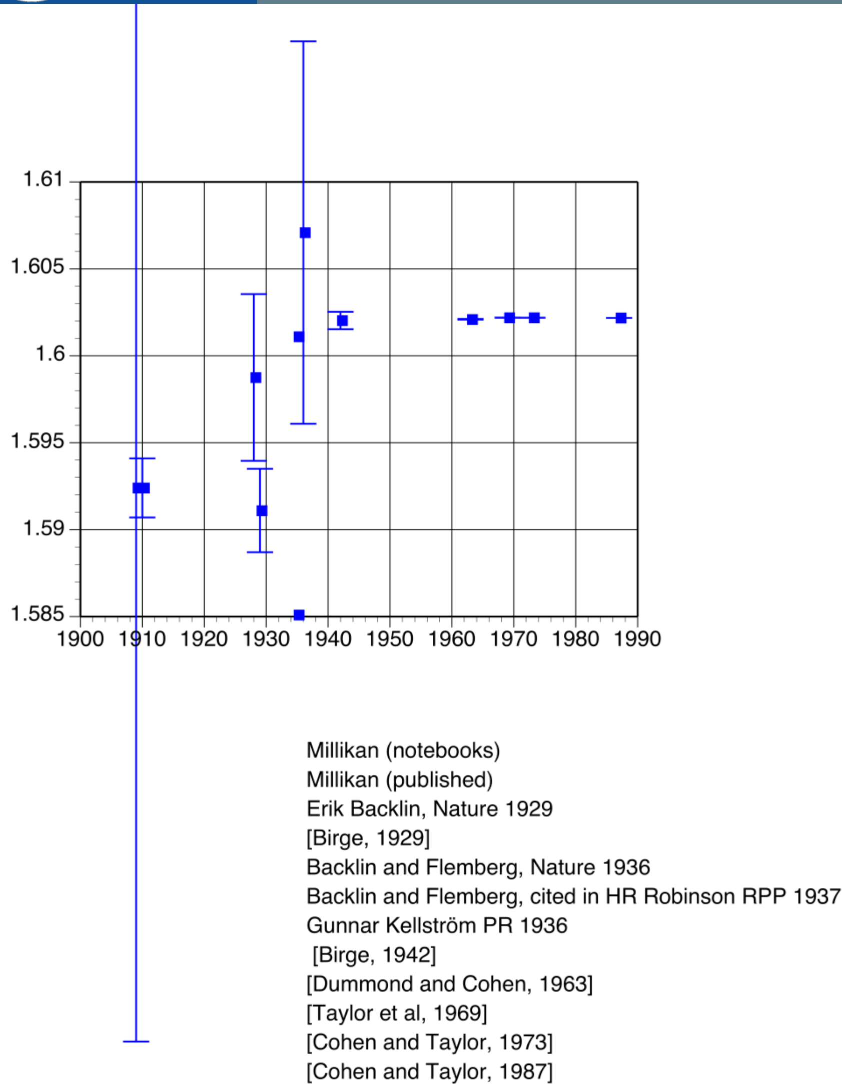
With 10 observations, the difference (-0.009) is not statistically significant

Example 2: Sample size = 1,000,000

```
Two-sample T Results

           N   Mean   StDev   SE Mean
C1   1000000   5.01    1.00    0.0010
C2   1000000   5.02    1.00    0.0010

Difference = mu (C1) - mu (C2)
Estimate for difference:   -0.00912

95% CI for difference:   (-0.01189, -0.00635)

T-Test of difference = 0 (vs not =): T-Value = -6.45  P-Value = 0.000  DF =1999994
```

With a million observations, the same difference (-0.009) is statistically significant!

NEW TOOLS, OLD BIAS III



https://hsm.stackexchange.com/questions/264/timeline-of-measurements-of-the-electrons-charge

We have learned a lot from experience about **how to handle some of the ways we fool ourselves**.

(…) Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It's a little bit off because he had the incorrect value for the viscosity of air. It's interesting to look at the history of measurements of the charge of an electron, after Millikan. If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.

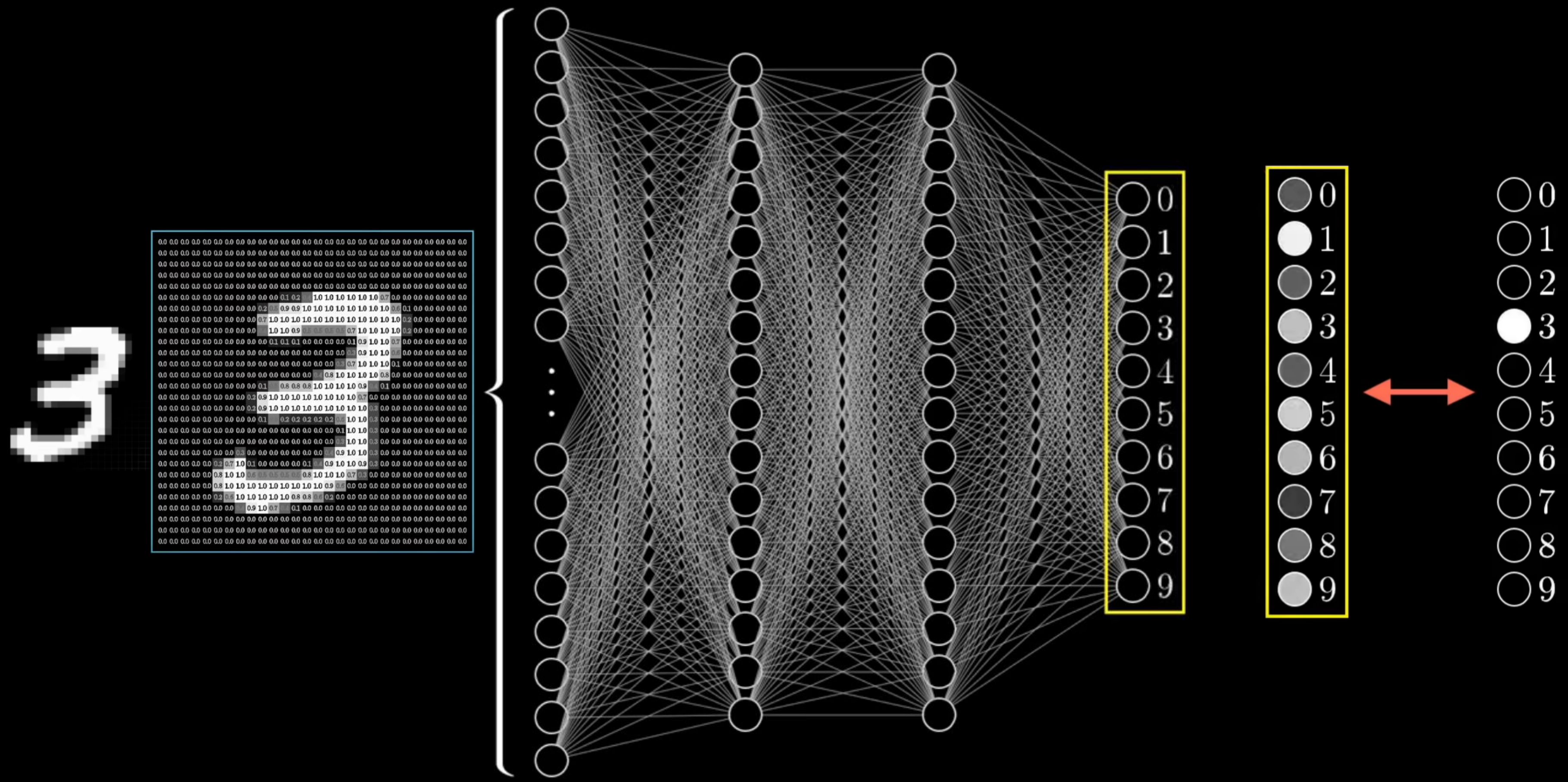
Why didn't they discover the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that"

Richard Feynman "Surely you're joking Mr. Feynman!" 1997

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events

CAN WE VALIDATE THEM?



CAN WE VALIDATE THEM?

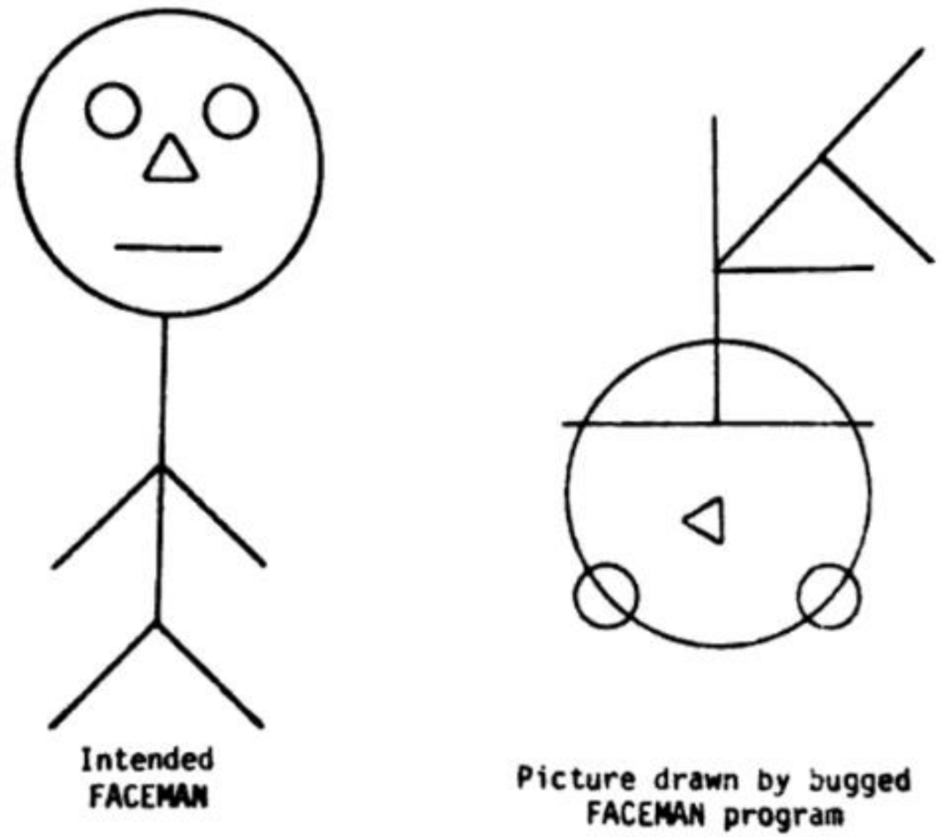
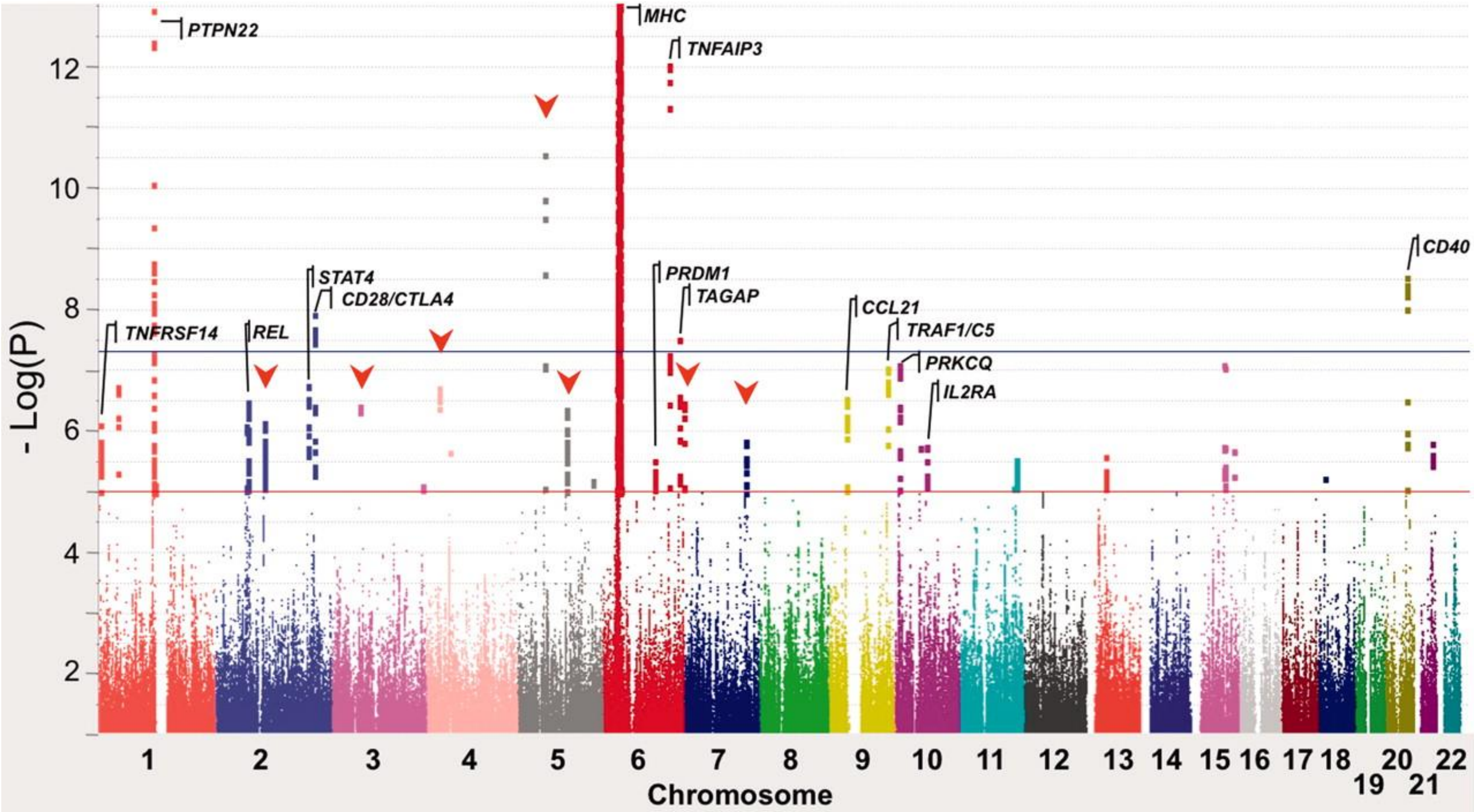


Figure 9. Stick men drawn by LOGO programs (from Sussman, 1973)

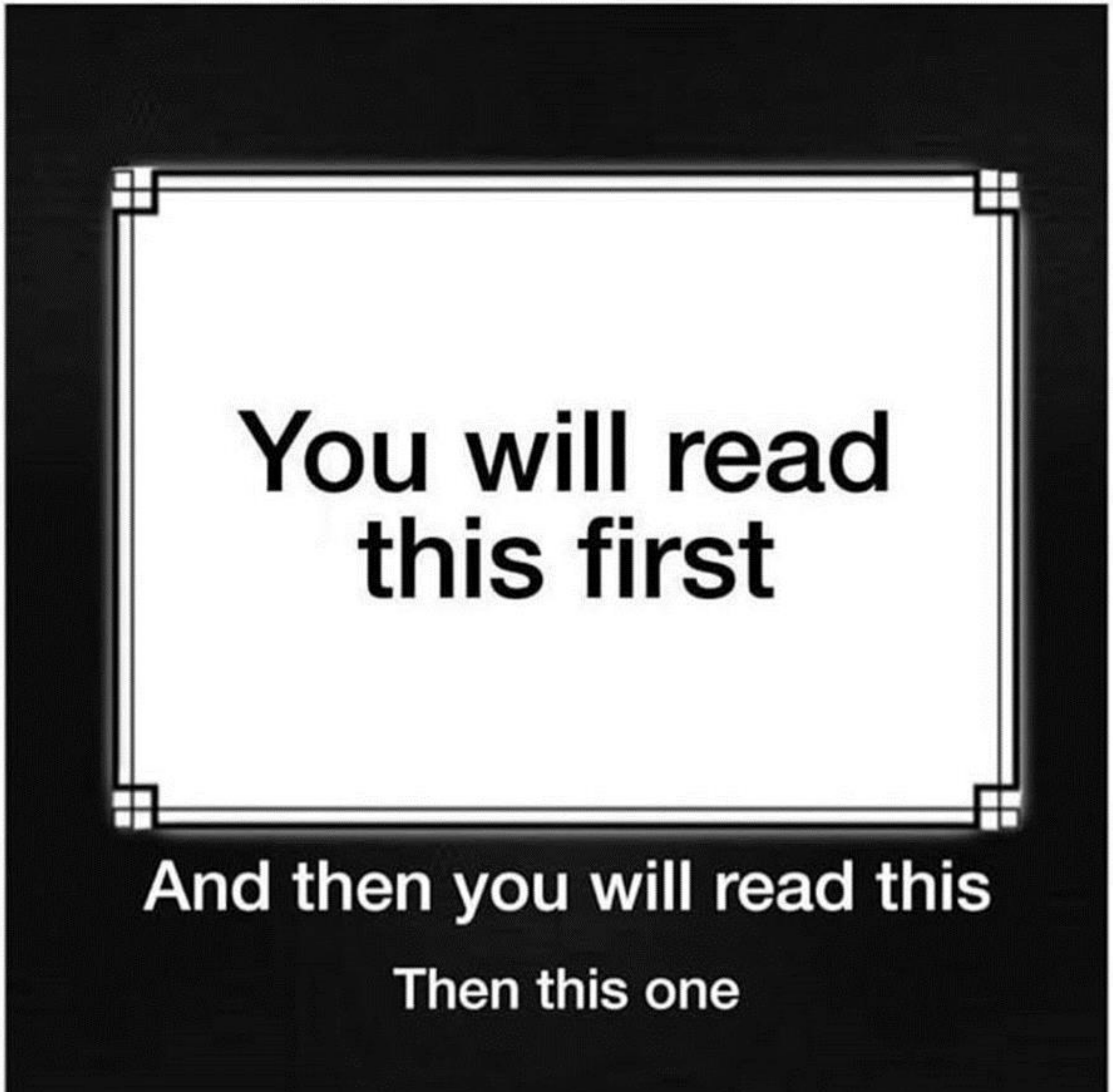
Thinking: Readings in Cognitive Science, (1978) edited by P. N. Johnson-Laird, P. C. Wason, page 26.

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events

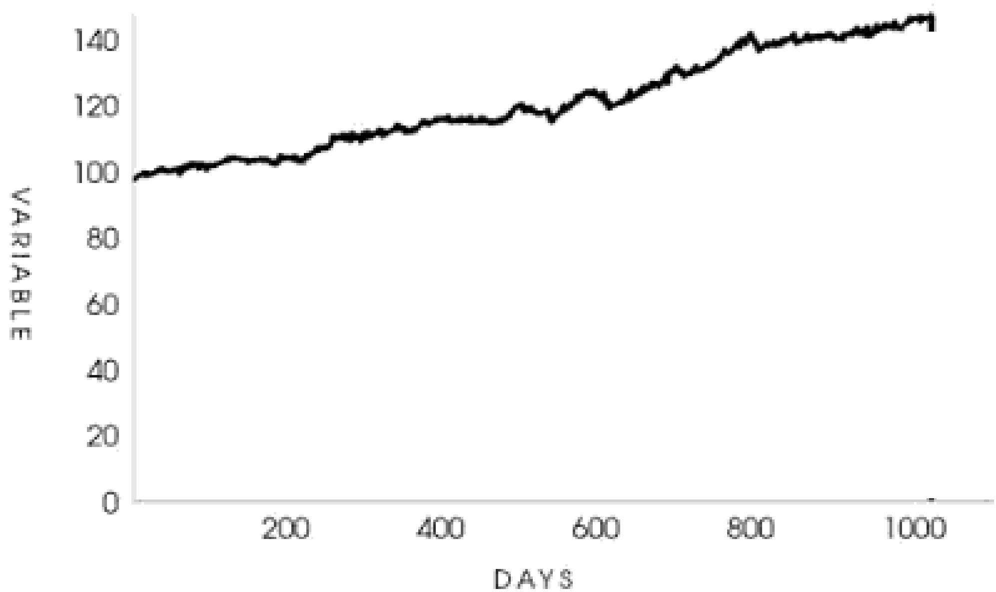

PTPN22
MHC
TNFAIP3
12
10
CD40
STAT4
PRDM1
CD28/CTLA4
TAGAP
CCL21
8
TNFRSF14
REL
TRAF1/C5
PRKCQ
IL2RA
6
4
2
- Log(P)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

And you will read this at the end



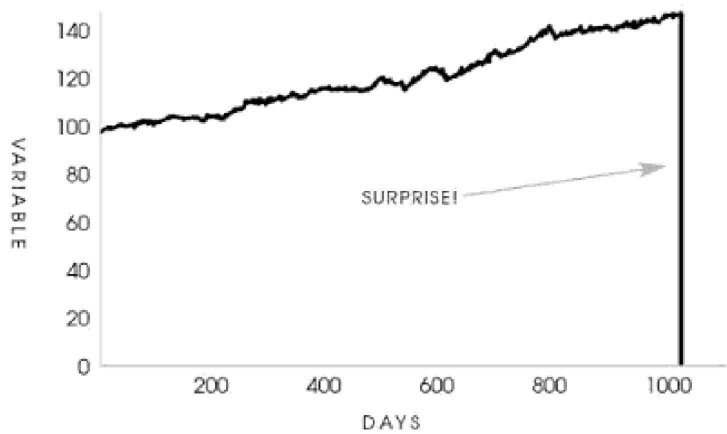
SAME AS THE PAST

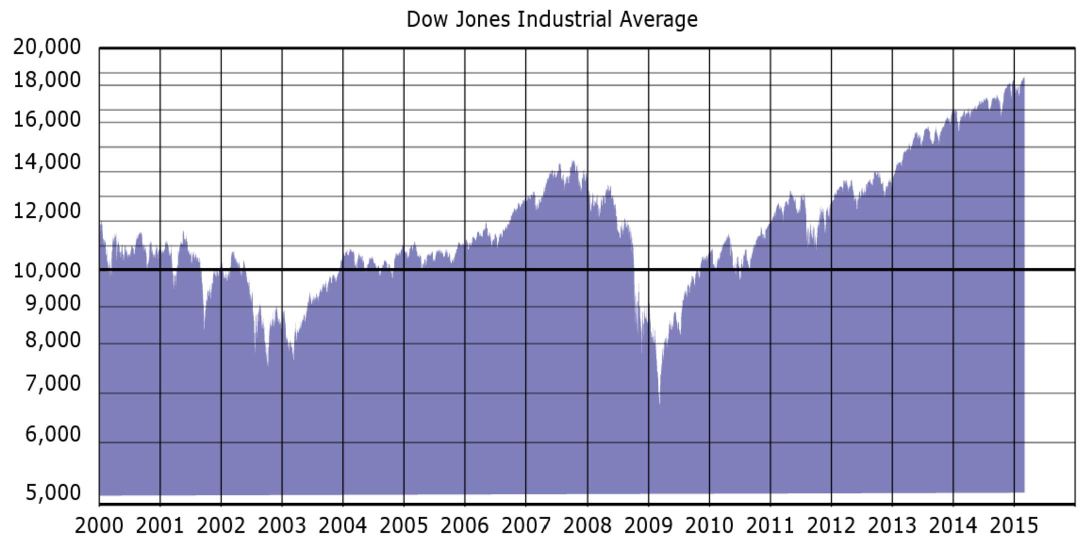
FIGURE 1: ONE THOUSAND AND ONE DAYS OF HISTORY

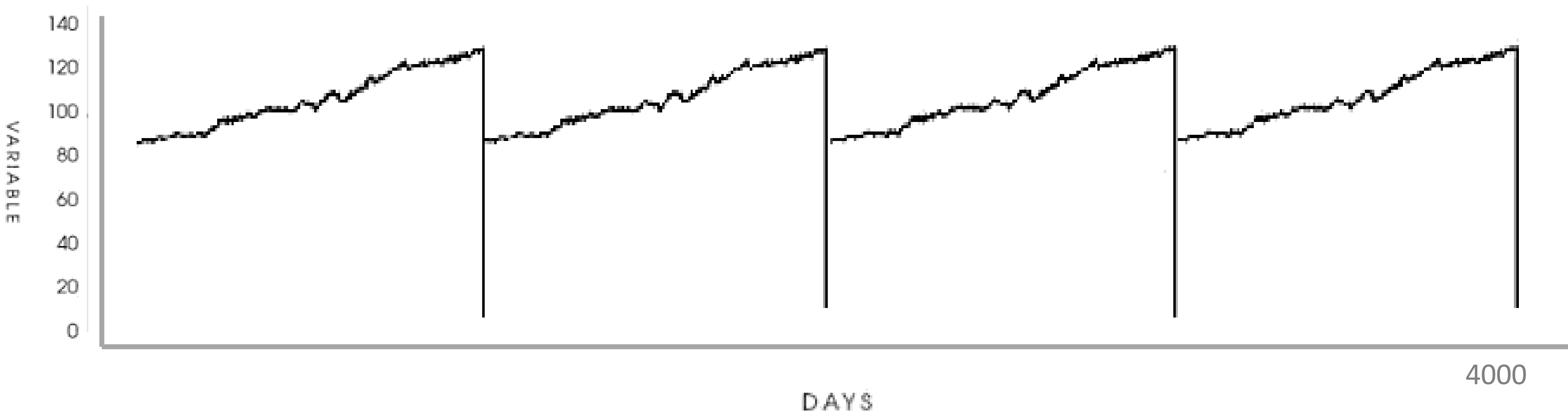


Nassim Taleb

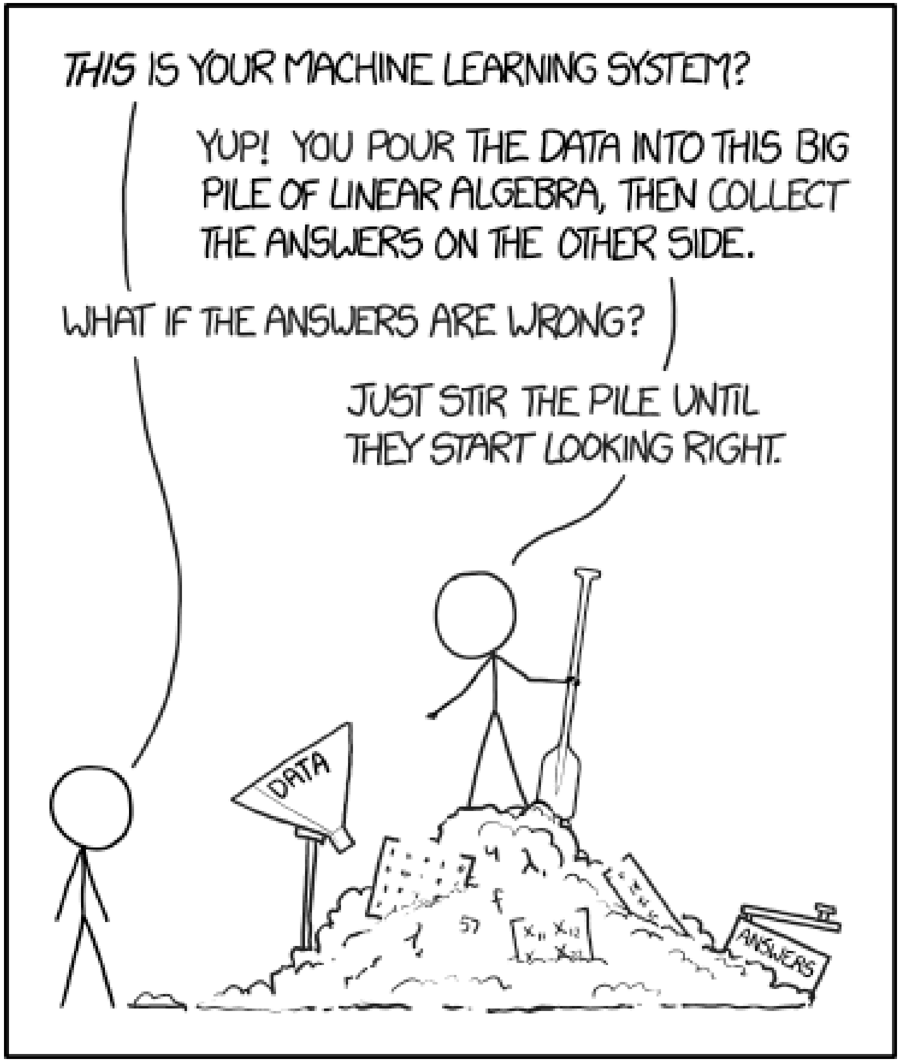
SAME AS THE PAST







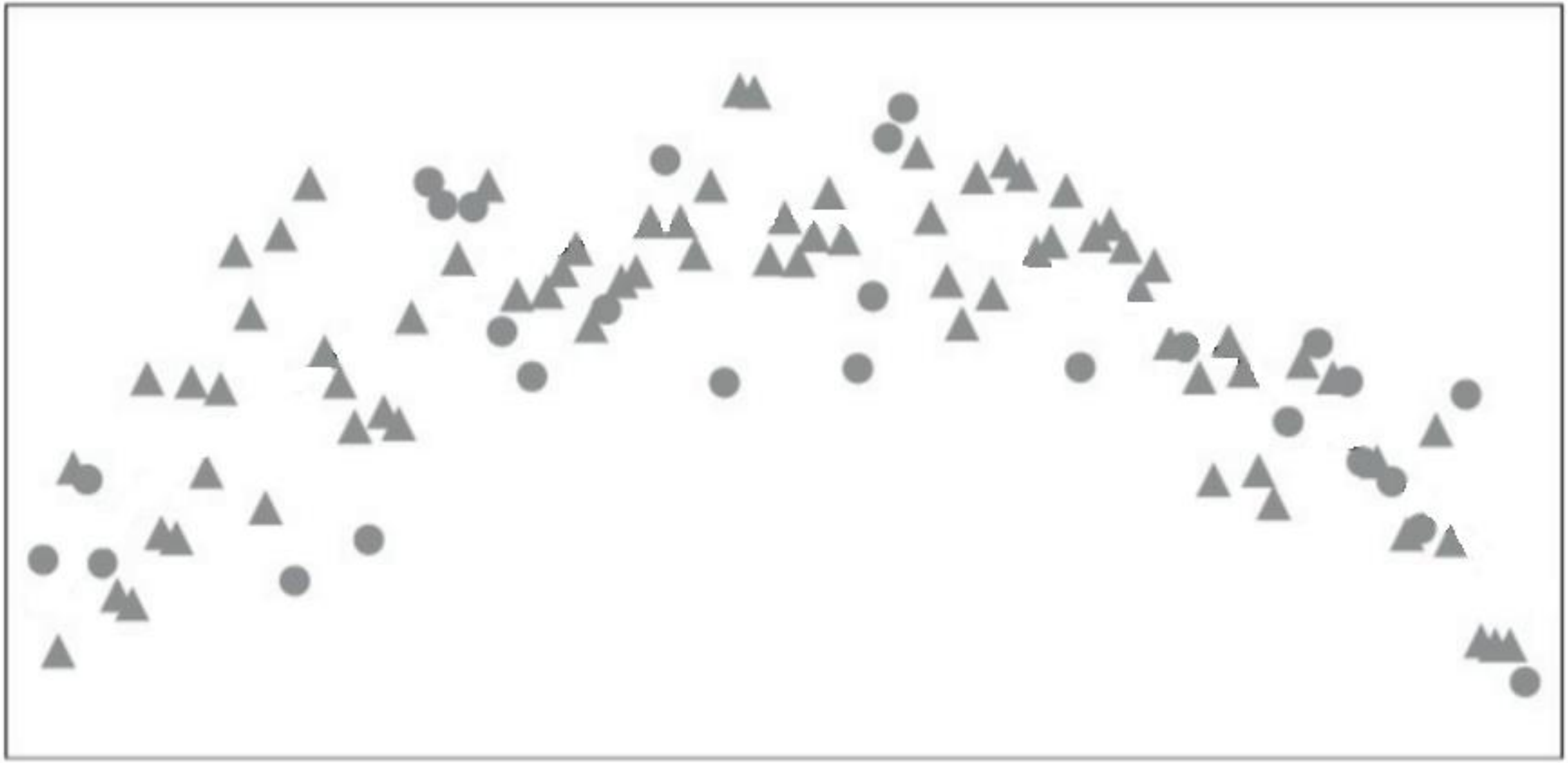
IT ALWAYS GIVES A RESULT



ROADMAP

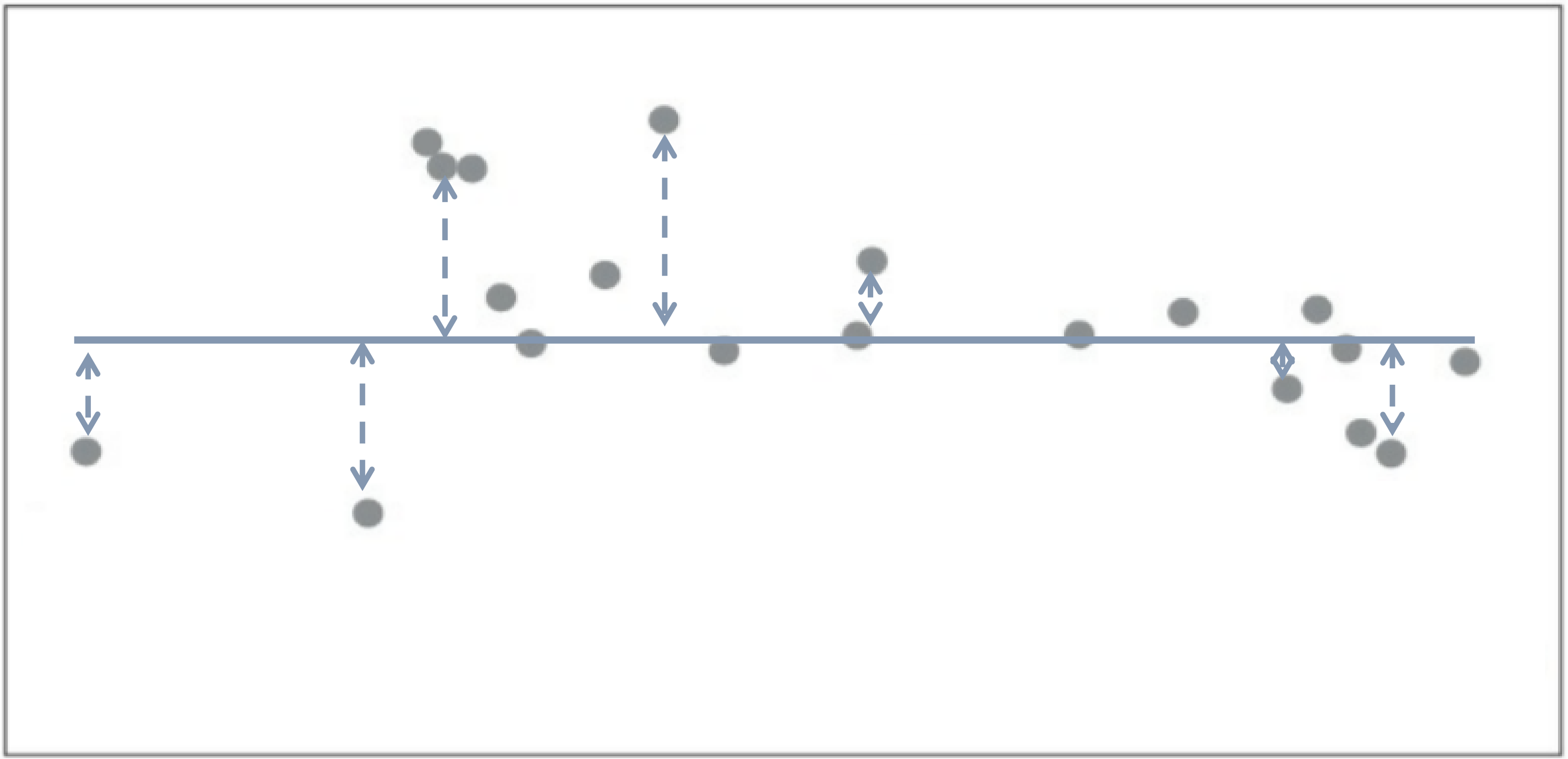
1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting

FIGURE 5-5: TRUE DISTRIBUTION OF DATA



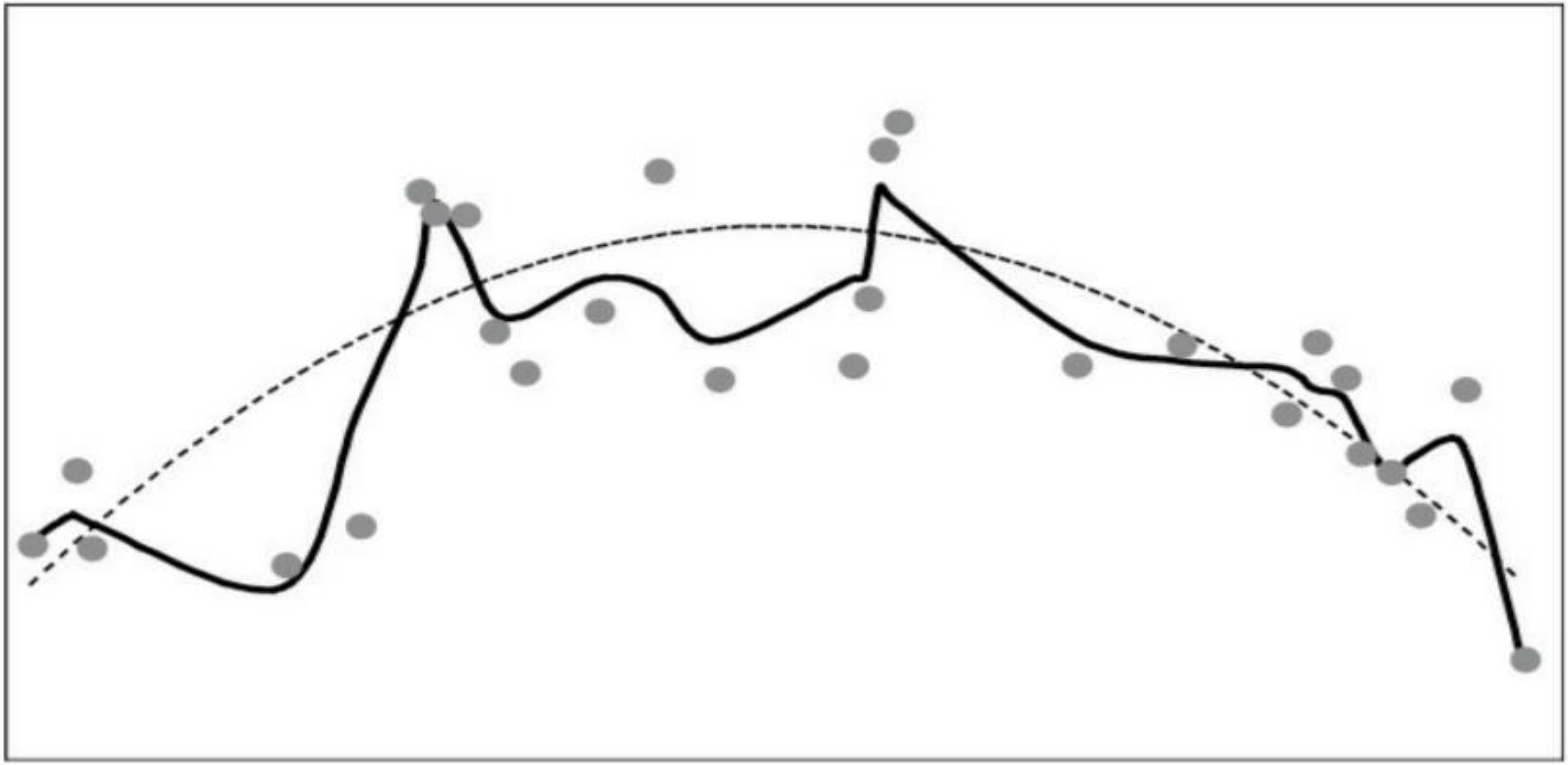
Adapted from The Signal and the Noise, Nate Silver, Penguin Books

BIASED DATASETS



Adapted from The Signal and the Noise, Nate Silver, Penguin Books

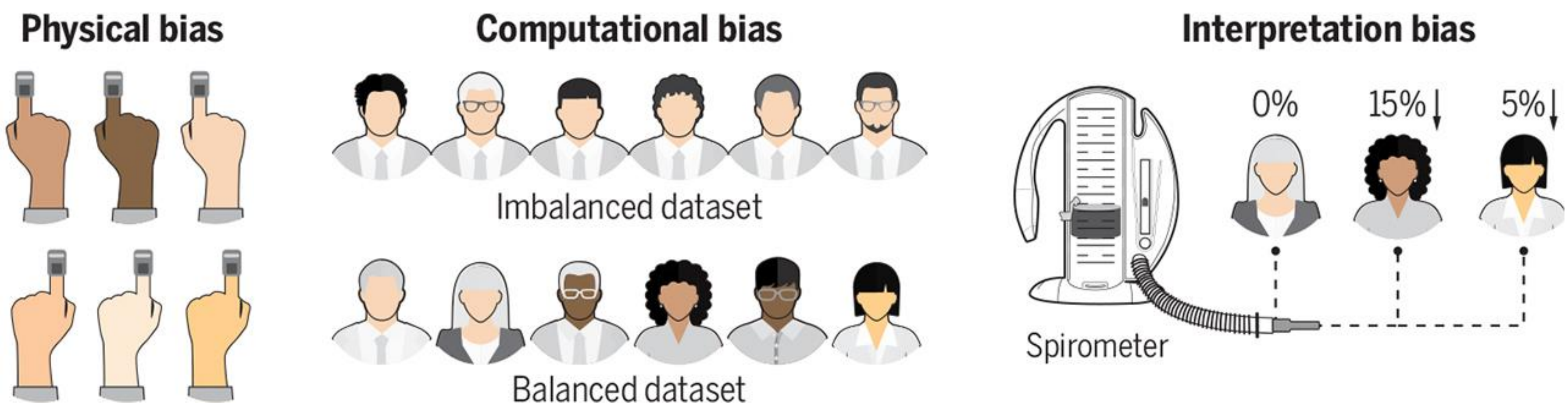
BIASED DATASETS



Adapted from The Signal and the Noise, Nate Silver, Penguin Books

Bias in medical devices

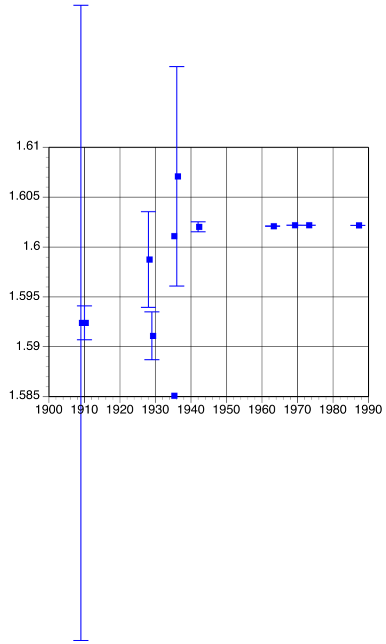
A device can be biased if its design disadvantages certain groups on the basis of their physical attributes, such as skin color. For example, pulse oximeters (see the photo) detect changes in light passed through skin and are less effective in people with dark skin. Computational techniques are biased if training datasets are not representative of the population. Interpretation of results may be biased according to demographic groups, for example, with the use of "correction factors."

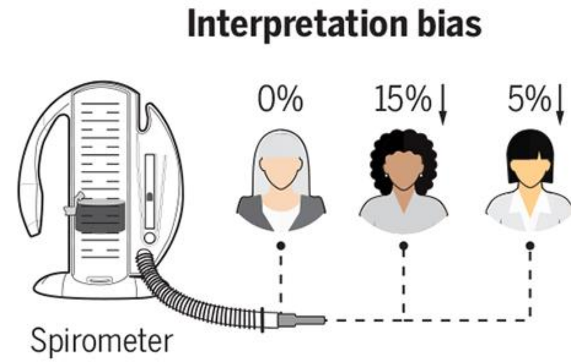


Kadambi, Achieving fairness in medical devices, *Science* Apr 2021

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting





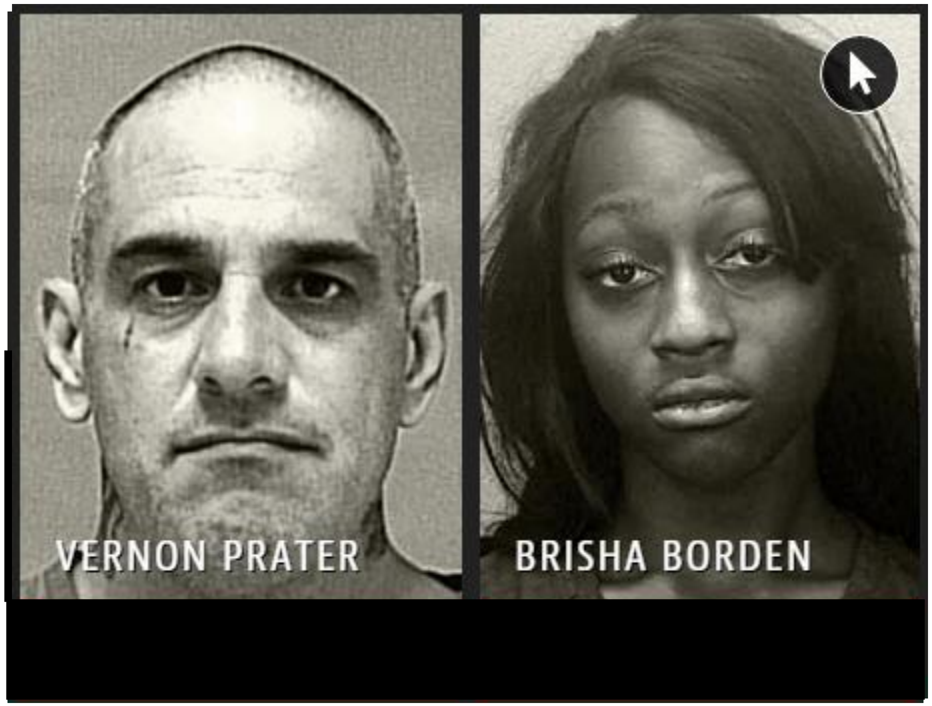
Anchoring Effect

How happy are you with your life?

How many dates did you have last month?

Priming and communication: Social determinants of information use in judgments of life satisfaction

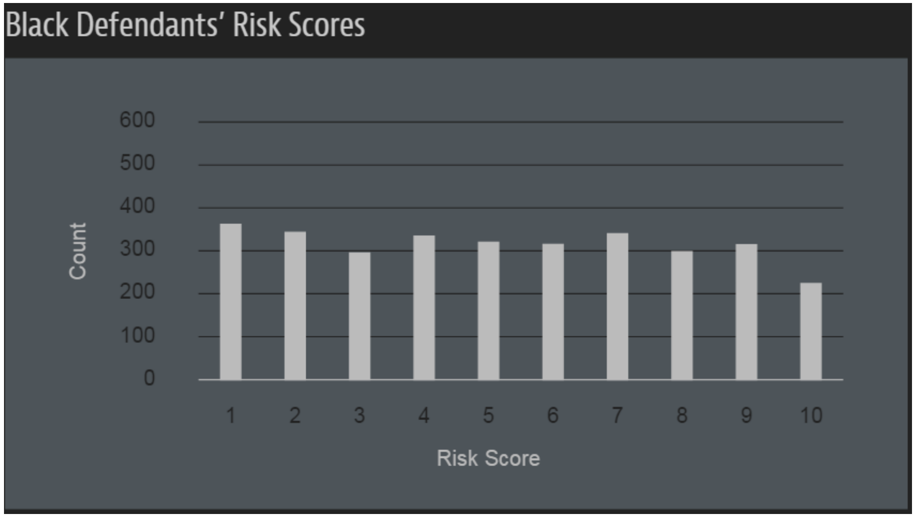
BIASED DATASETS

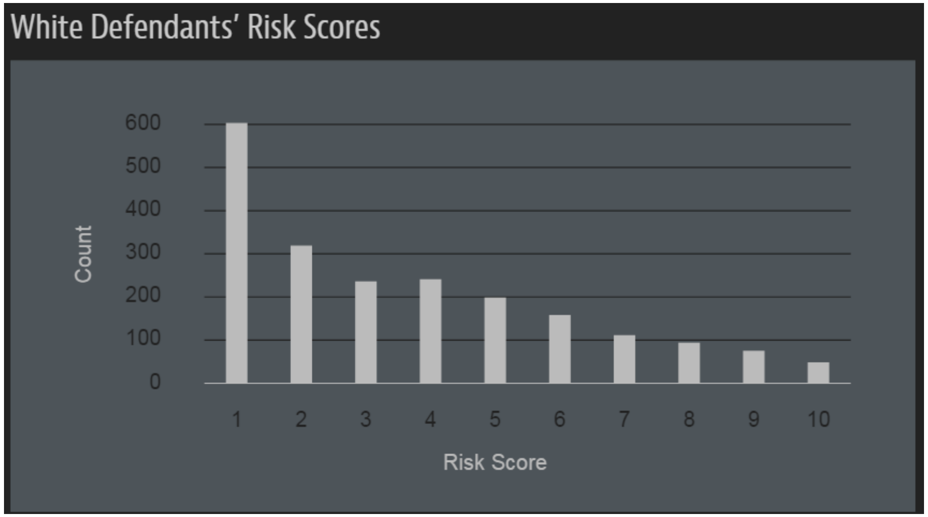


https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- Particularly likely to falsely flag black defendants as future criminals, wrongly labelling them this way at almost twice the rate as white defendants.

- White defendants were mislabelled as low risk more often than black defendants.





Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

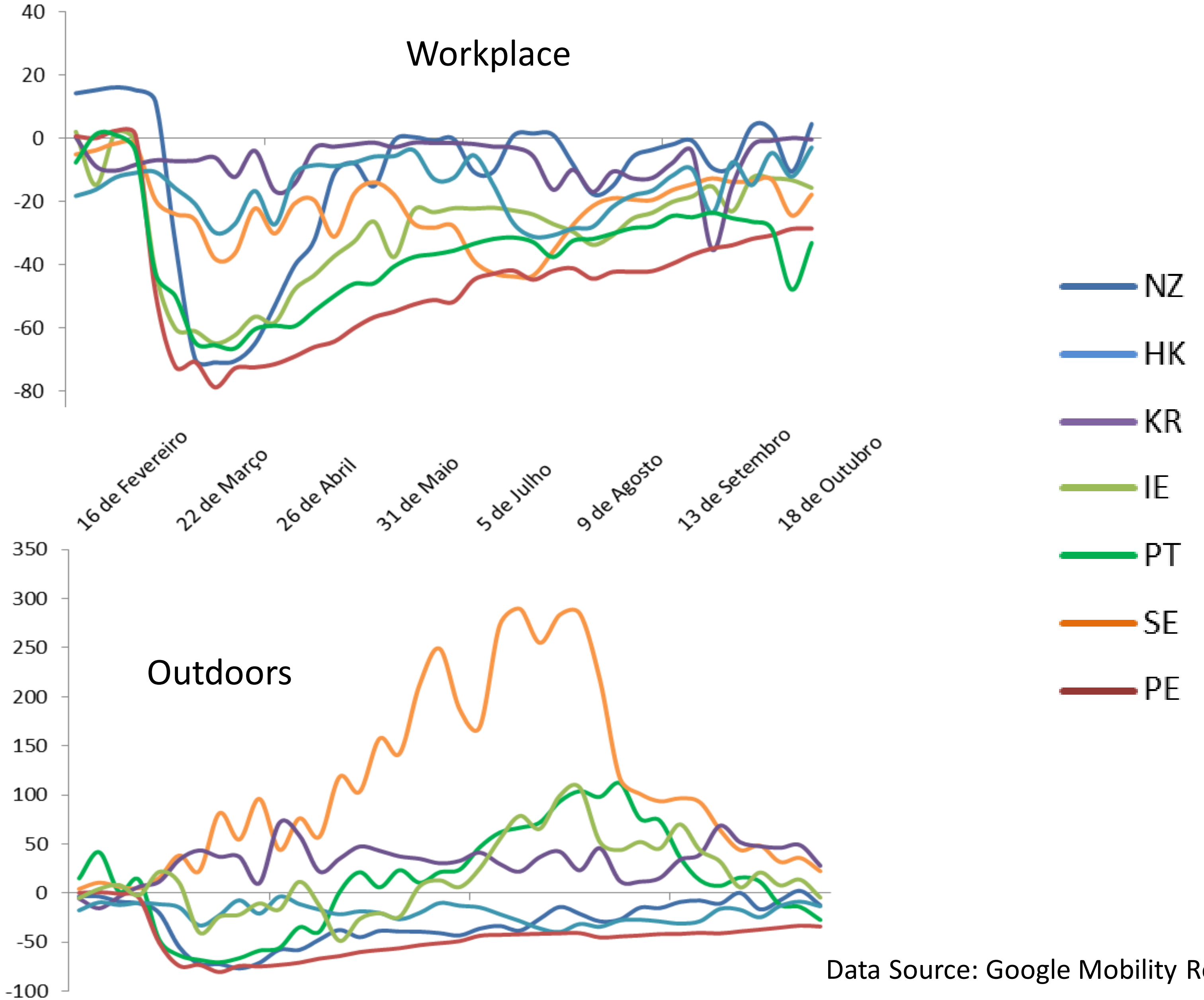
CNN World Africa Americas Asia Australia More

Dutch government resigns over child welfare fraud scandal

By Rosanne Roobeek, James Frater and Niamh Kennedy, CNN

Updated 1622 GMT (0022 HKT) January 15, 2021

BIASED DATASETS

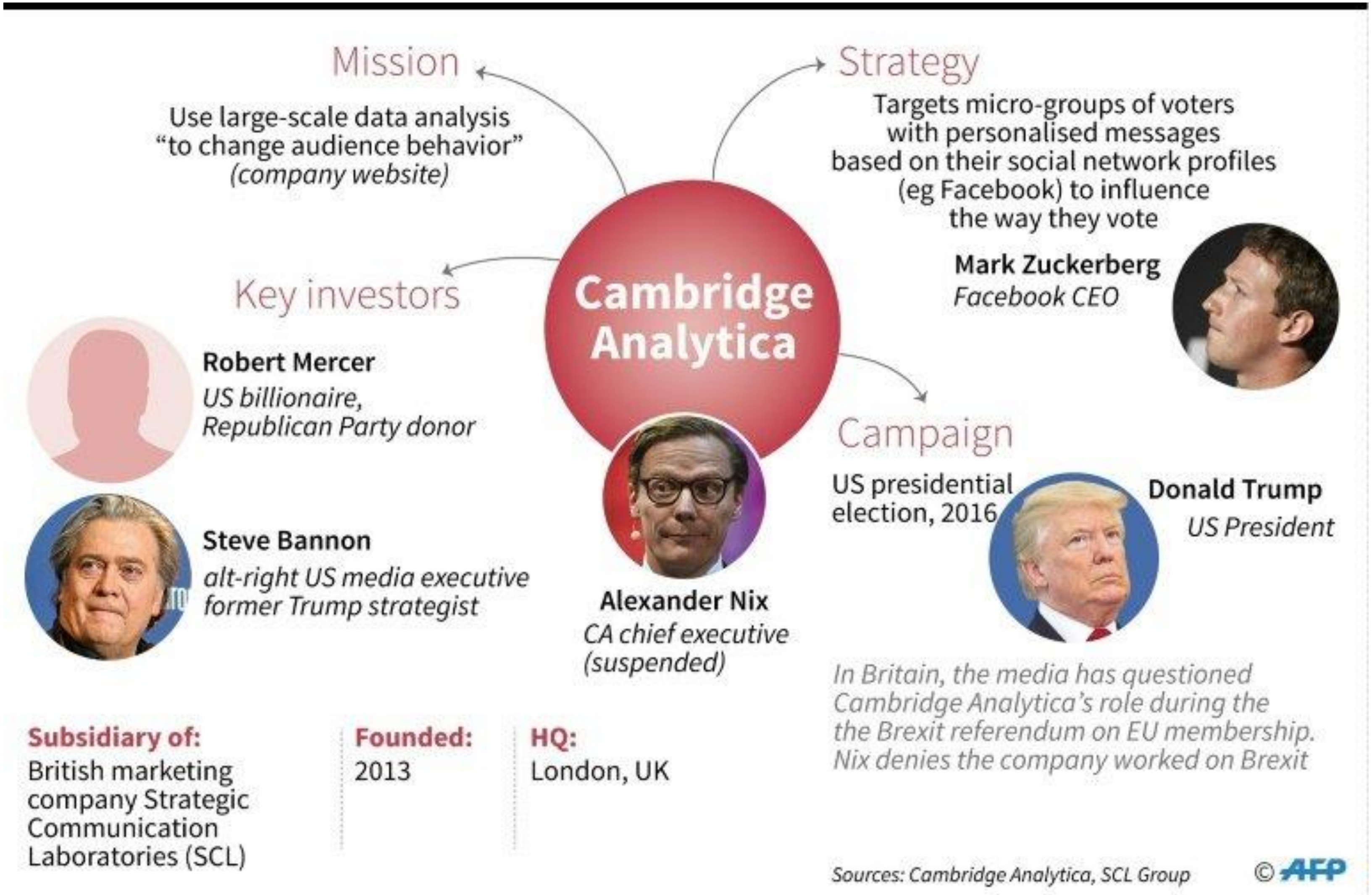


Data Source: Google Mobility Reports

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting

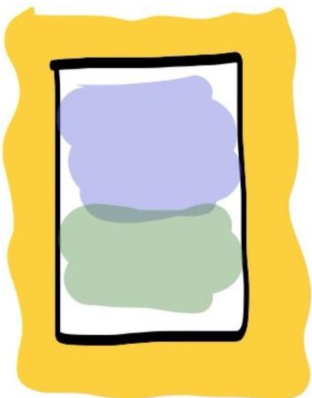
TARGETTING

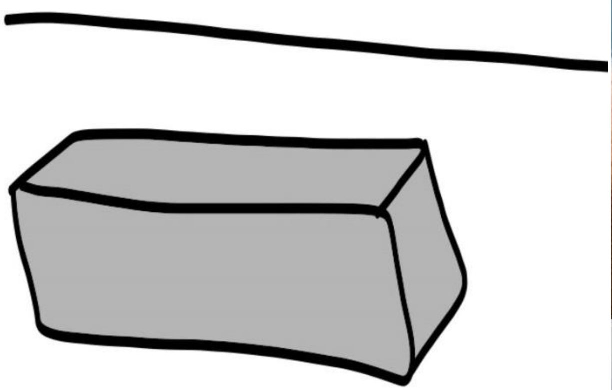

Mission
Use large-scale data analysis
"to change audience behavior"
(company website)
Strategy
Targets micro-groups of voters
with personalised messages
based on their social network profiles
(eg Facebook) to influence
the way they vote
Mark Zuckerberg
Facebook CEO
Key investors
Cambridge Analytica
Robert Mercer
US billionaire,
Republican Party donor
Steve Bannon
alt-right US media executive
former Trump strategist
Alexander Nix
CA chief executive
(suspended)
Campaign
US presidential
election, 2016
Donald Trump
US President
In Britain, the media has questioned
Cambridge Analytica's role during the
the Brexit referendum on EU membership.
Nix denies the company worked on Brexit
Subsidiary of:
British marketing
company Strategic
Communication
Laboratories (SCL)
Founded:
2013
HQ:
London, UK
Sources: Cambridge Analytica, SCL Group
© AFP

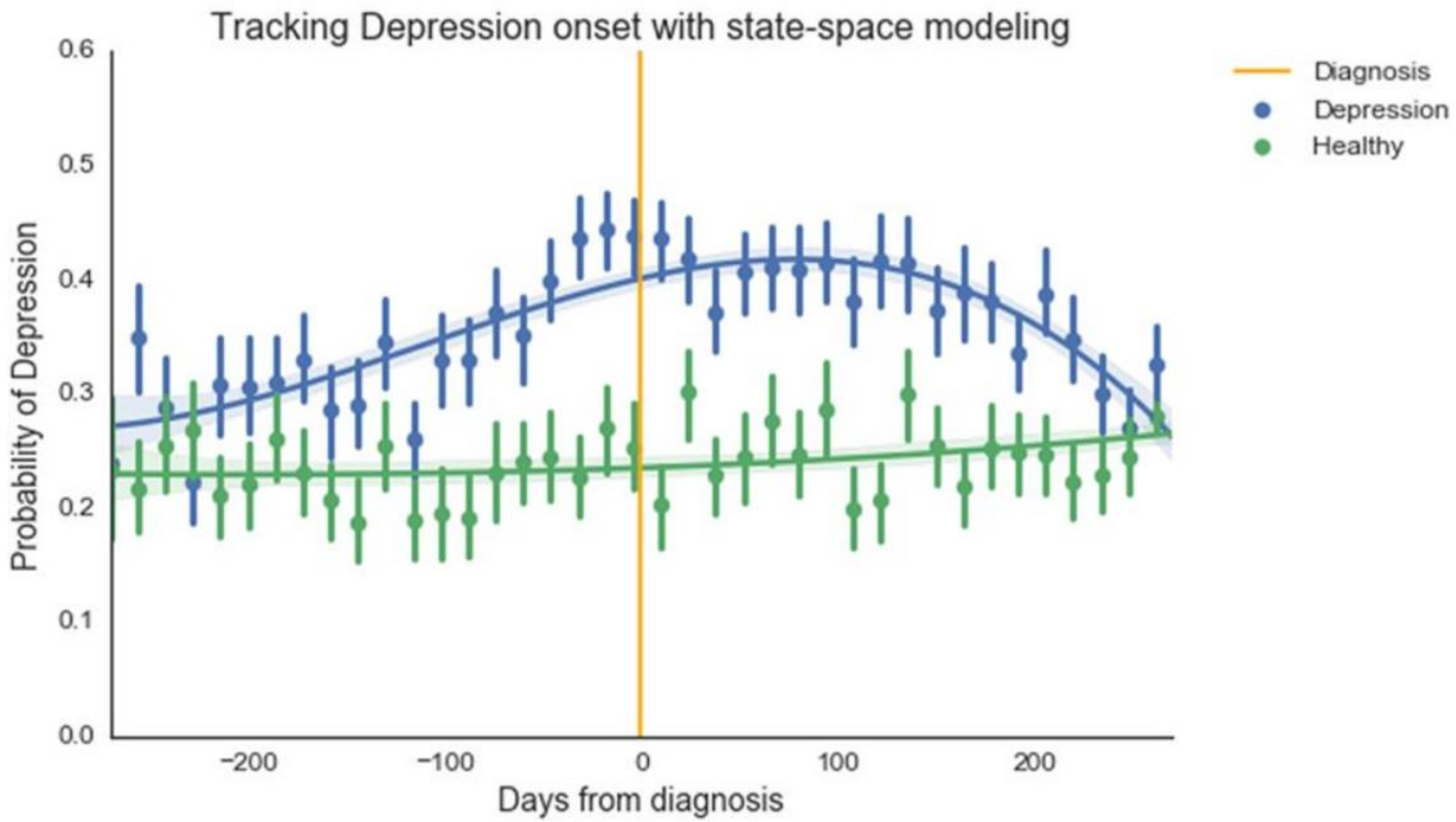
WILL IoT HELP US BECOME HAPPIER?

What if a picture knew it was making you feel calmer, more mindful...just happier?

Using sensors, apps & museums to enable wellbeing







https://www.nature.com/articles/s41598-017-12961-9

@mjoanasa
@DPolicyLab

why am i so

why am i so **tired**
why am i so **ugly**
why am i so **gassy**
why am i so **thirsty**
why am i so **angry**
why am i so **itchy**
why am i so **sad**
why am i so **hungry**
why am i so **emotional**
why am i so **bloated**

como posso ser

como posso ser **amigo de alguem**
como posso ser **feliz**
como posso ser **inteligente**
como posso ser **uma pessoa melhor**
como posso ser **salvo**
como posso ser **rico**
como posso ser **feliz sozinho**
como posso ser **um hacker**
como posso ser **popular no facebook**
como posso ser **cantora**

pourquoi je suis

pourquoi je suis **moche**
pourquoi je suis **triste**
pourquoi je suis **toujours fatigué**
pourquoi je suis **célibataire**
pourquoi je suis **toujours célibataire**
pourquoi je suis **devenu rebelle pdf**
pourquoi je suis **seule**
pourquoi je suis **toujours fatiguée**
pourquoi je suis **jalouse**
pourquoi je suis **triste sans raison**

how to

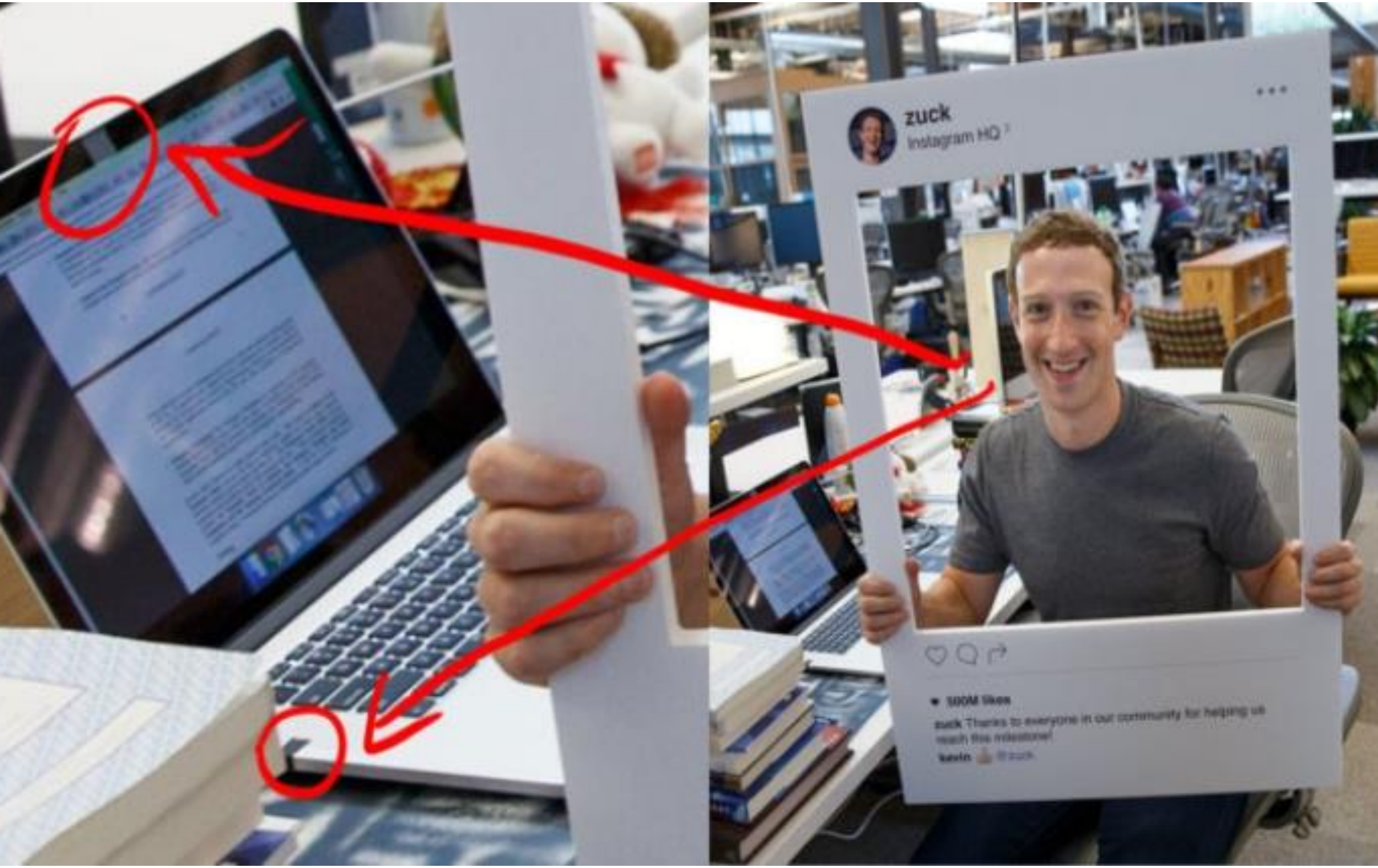
how to **make slime**
how to **tie a tie**
how to **buy bitcoin**
how to **lose weight**
how to **draw**
how to **buy ripple**
how to **kiss**
how to **make pancakes**
how to **mine bitcoin**
how to **train your dragon**

como é que se

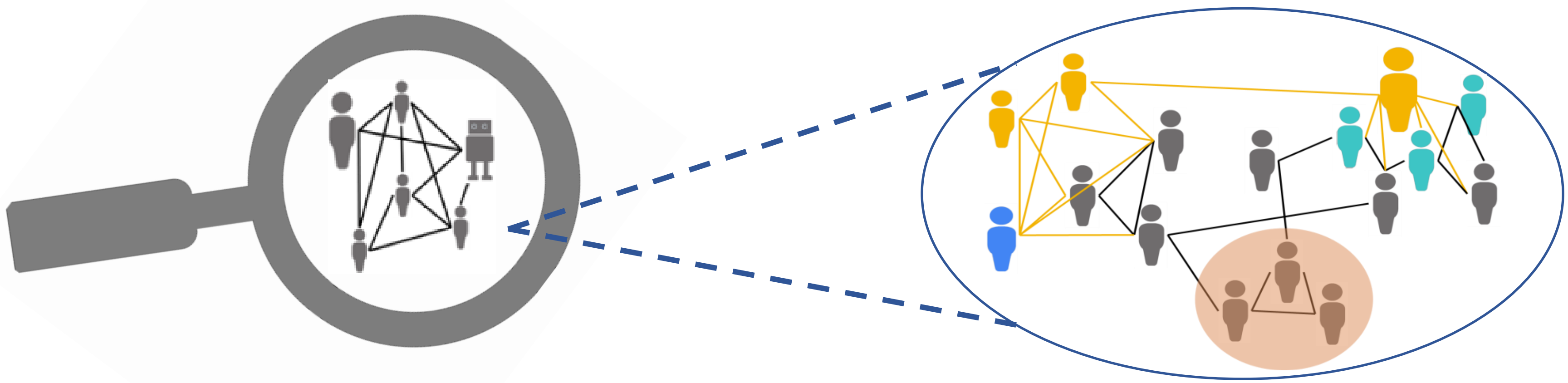
como é que se **beija**
como é que se **diz eu te amo**
como é que se **beija de lingua**
como é que se **engravida**
como é que se **beija na boca**
como é que se **escreve**
como é que se **beija pela primeira vez**
como é que se **faz um facebook**
como é que se **faz um relatório**
como é que se **faz panquecas**

comment faire

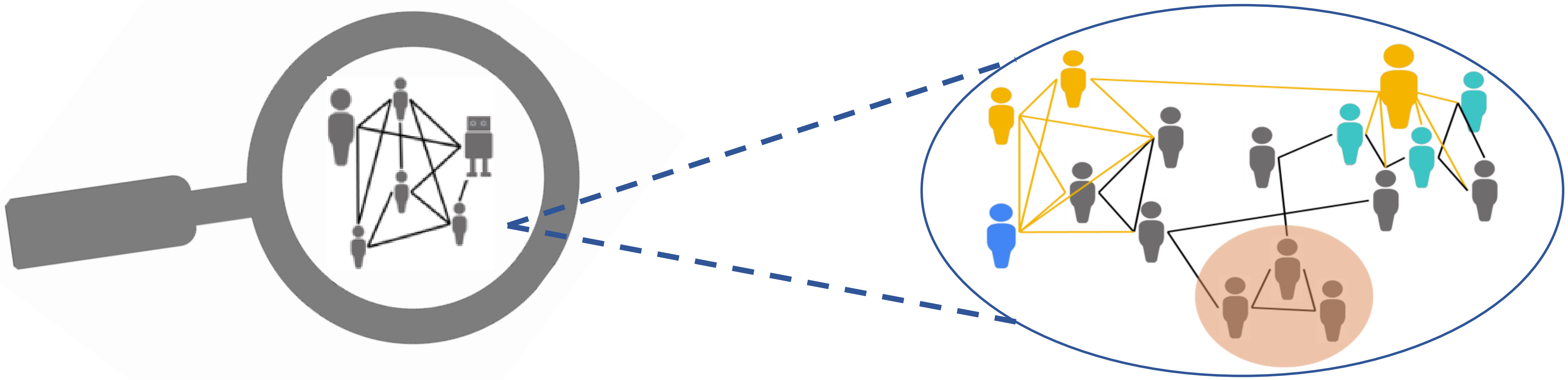
comment faire **du slime**
comment faire **un cv**
comment faire **des crepes**
comment faire **une dissertation**
comment faire **une capture d'écran**
comment faire **une bibliographie**
comment faire **un gateau**
comment faire **du caramel**
comment faire **de la glue**
comment faire **du pain**


zuck
Instagram HQ

WE CREATED A MACROSCOPE



WE CREATED A MACROSCOPE



- SECURITY BREACHES
- PRIVACY CONCERNS
- ETHICAL CONCERNS
- DATASET BIAS
- ALGORITHMIC BIAS
- INSTRUMENTATION BIAS

ROADMAP

1. (Some) problems with methods
    1. Fisher or Bayes (p-values, priors)?
    2. Should we have hypothesis?
    3. How can we deal with randomness?
    4. Can we really validate results that are non-interpretable?
    5. Signal to noise identification in rare events
2. (Some) problems with data
    1. Incomplete
    2. Non-random biases (sampling, prejudice, systematic value distortion)
    3. Private/personal
3. Examples of applications gone wrong
    1. Human Bias / interpretation
    2. When recommendation systems fail
    3. When recommendation systems work
4. (Some) possible solutions / approaches
    1. Auditing
    2. Myth busting

WHAT CAN BE DONE?

1. ACCEPTANCE – We have a problem
2. TRAINING – Like today
3. INCLUSION AND DIVERSITY – In teams, projects, etc
4. AUDITING – Know your data
5. DE-BIASING – When possible
6. TRANSPARENCY – If you can’t be right, be honest
7. INTERACTION – Teach, engage, change the ones, legislate

Bias and Fairness Audit Toolkit

The Bias Report is powered by Aequitas, an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

Upload Data → Select Protected Groups → Select Fairness Metrics → The Bias Report

False Positive Rate Parity: Failed

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have False Positive parity if every group has the same False Positive Error Rate. For example, if race has false positive parity, it implies that all three races have the same False Positive Error Rate.	If your desired outcome is to make false positive errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is punitive and has a risk of adverse outcomes for individuals. Using this criteria allows you to make sure that you are not making false positive mistakes about any single group disproportionately.	**For race** (with reference group as **Caucasian**) Asian with 0.37X Disparity African-American with 1.91X Disparity Native American with 1.60X Disparity Other with 0.63X Disparity

http://aequitas.dssg.io/

From the Industrial Revolution to the Digital Revolution

First

Water and steam power is used to create mechanical production facilities.

Second

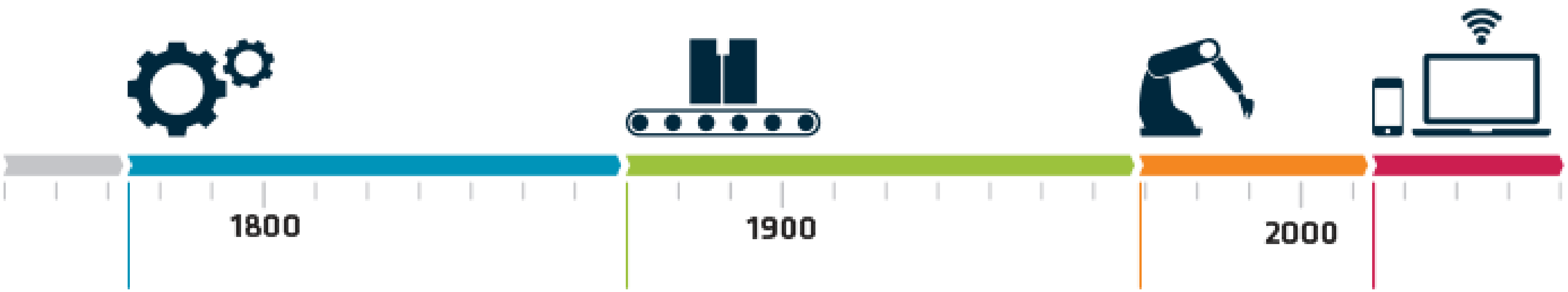
Electricity lets us create a division of labor and mass production.

Third

IT systems automate production lines further.

Fourth

IoT and cloud technology automate complex tasks.



1784: First mechanical loom

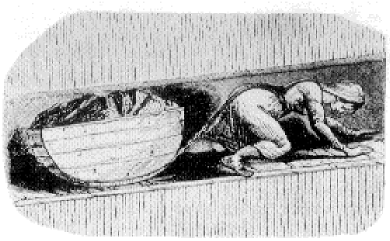
1870: First assembly line

1969: First programmable logic controller

Today

Source: https://mjolner.dk/2015/01/14/realizing-the-fourth-industrial-revolution/

From the Industrial Revolution to the Digital Revolution



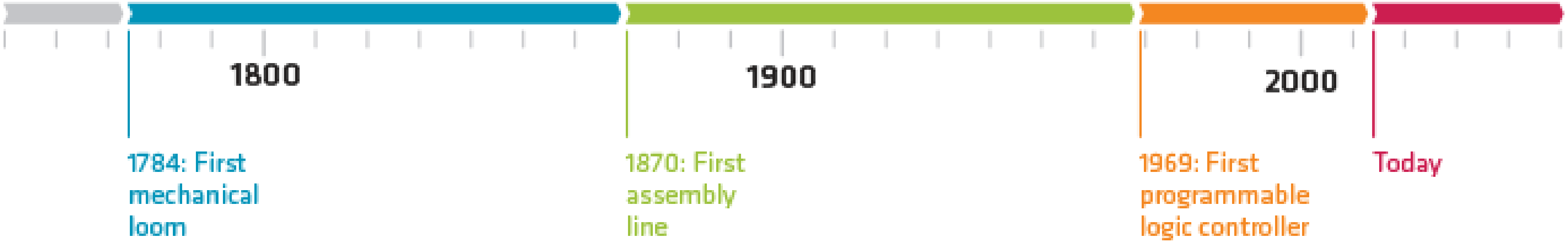
Child Labor in the Industrial Revolution

- 9 year old limit
- 9-13 yo should not work > 9h a day
- 13-18yo should not work > 12h a day
- Four inspectors

1833 UK Factory Act

1938 US Federal Fair Labor Standards Act

1973: ILO Conference



Source: https://mjolner.dk/2015/01/14/realizing-the-fourth-industrial-revolution/

Some myths

1. Models are neutral
2. More (data) is always better
3. If the model returns a highly likely results it must be true
4. ML models facilitate extrapolation
5. There is nothing we can do about privacy

Ask ourselves:

1. Am I using proxies and are they fare?
2. What happens if I get it wrong? What is the worst thing that can happen? – Punitive models
3. Can I update my model continuously?
4. Does the algorithm itself impact the results?

FURTHER READING

Popular reading on Data Science/Statistics/Social Physics:

- Taleb, Nassim Nicholas. *The black swan: The impact of the highly improbable*. Vol. 2. Random house, 2007.
- Silver, Nate. *The signal and the noise: the art and science of prediction*. Penguin UK, 2012
- Harford, Tim. "Big data: A big mistake?." *Significance* 11.5 (2014): 14-19.
- Pentland, Alex. *Social physics: How good ideas spread-the lessons from a new science*. Penguin, 2014.
- Lazer, David, et al. "Social science. Computational social science." Science (New York, NY) 323.5915 (2009): 721-723

Popular reading on AI risks

- Zuboff S (2019) The age of surveillance capitalism: The fight for a human future at the new frontier of power. Public Affairs, New York
- O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Publishers, New York

Some specific examples:

- Amnesty International (2021) Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal.
- Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (20118) Aequitas: A bias and fairness audit toolkit
- Kadambi, Achieving fairness in medical devices, *Science* Apr 2021

THANK YOU

socialcomplexity.eu
joanagsa@lip.pt
@mjoanasa
@spac_lab