Data Science: Principles and Practice

Lecture 8: Advanced topics

Marek Rei

UNIVERSITY OF CAMBRIDGE

Data Science: Principles and Practice

01 Overview of Complementary ML Techniques

02 Ethics in Data Science

03 Replicability of Findings

04 Assignment

Overview of Complementary ML Techniques

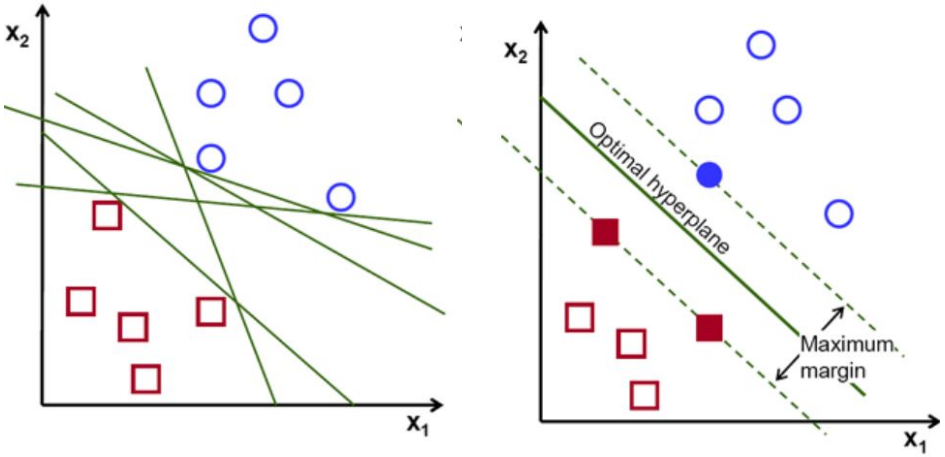
Support Vector Machines

Support Vector Machines (SVM) are a type of classification algorithm.

Logistic regression tries to maximize the probability of the correct class.

SVM tries to find a hyperplane that separates the closest points from both classes with the largest margin.

More details in *“Machine Learning and Bayesian Inference”* in the Easter term.



https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f

Decision Trees

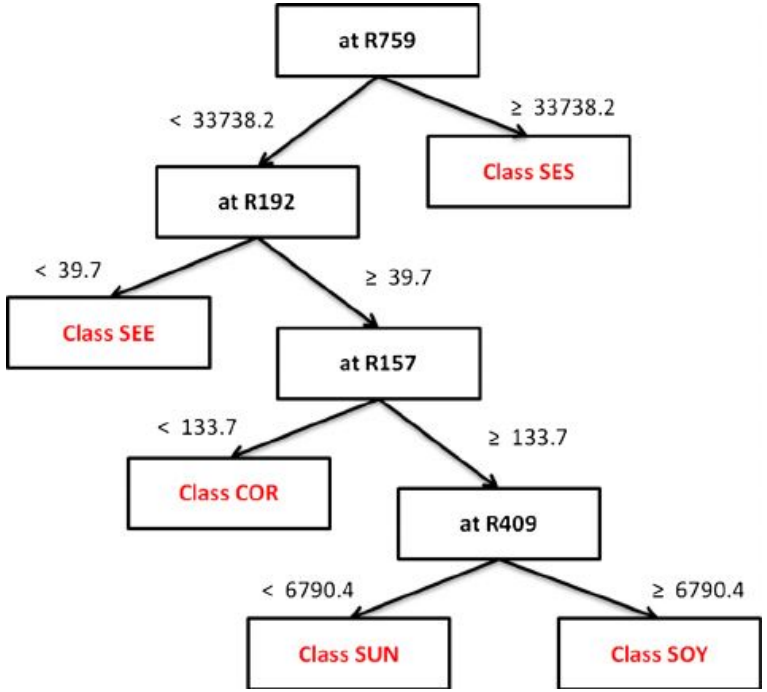
Recursively divide the data into smaller sections to perform classification.

Each node is a rule that splits the data.

Each leaf is a classification decision.

Provide an interpretable model (relatively).

Can easily overfit to the training data.



Ruiz-Samblás et al. (2014) Application of data mining methods for classification and prediction of olive oil blends with other vegetable oils.

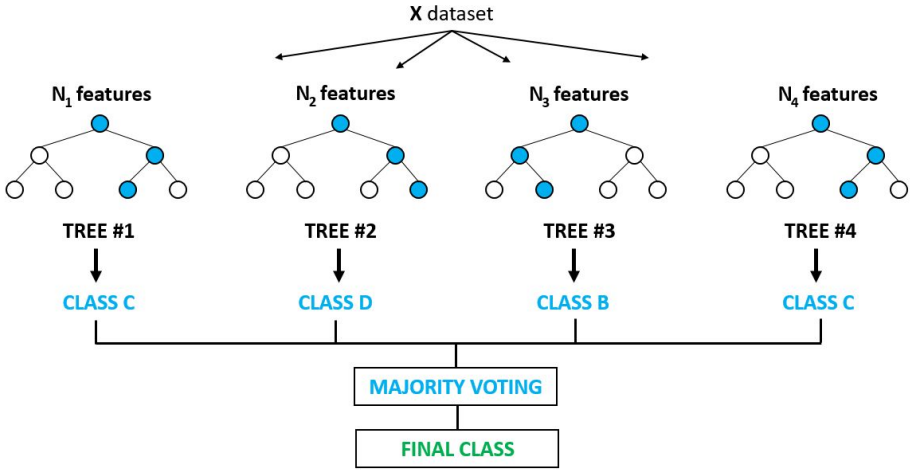
Random Forests

Combine many different decision trees together to make a single prediction.

Return either the most frequency predicted class or average the result.

Much more stable than a single decision tree - averages out the overfitting problem.

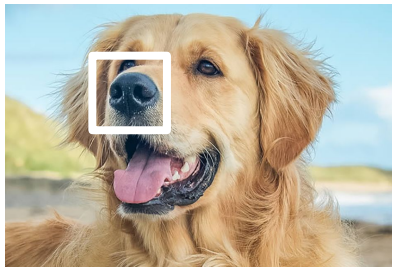
Works really well in practice!

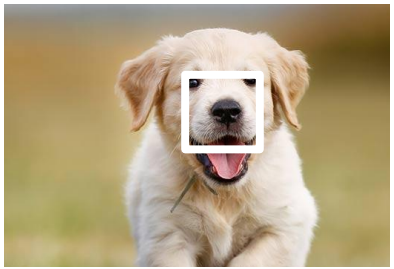


Convolutional Neural Networks

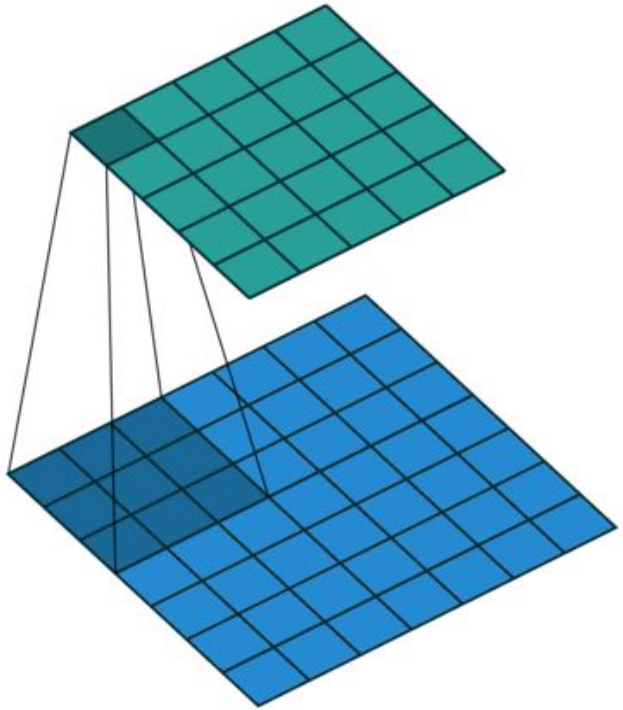
Neural modules operating repeatedly over different subsections of the input space.

Great when searching for feature patterns, without knowing where they might be located in the input.





The main driver in image recognition.
Can also be used for text.



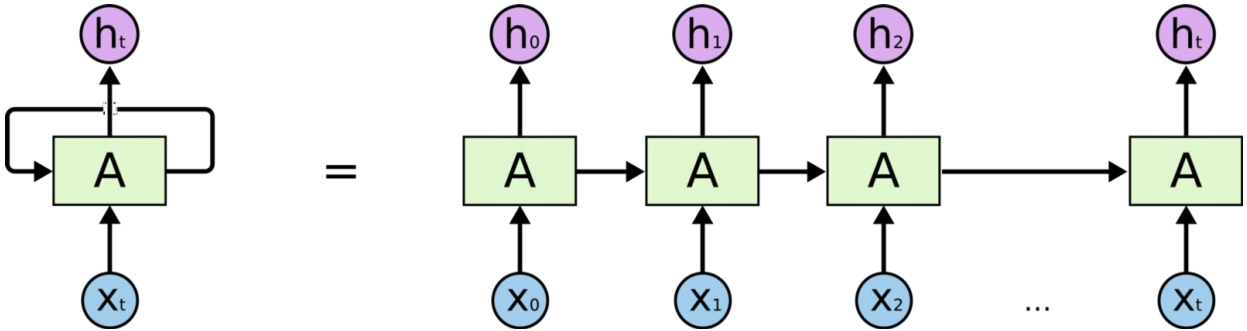
https://github.com/vdumoulin/conv_arithmetic

Recurrent Neural Networks

Designed to process input sequences of arbitrary length.

Each hidden state A is calculated based on the current input and the previous hidden state.

Main neural architecture for processing text, with each input being a word representation.



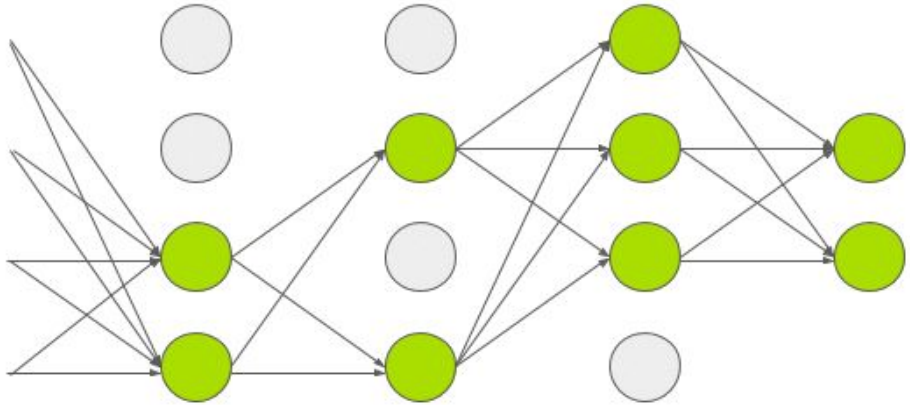
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Dropout

During training, randomly set some neural activations to zero.

Typically drop 50% of activations in a layer.

Form of regularization - prevents the network from relying on any one node.



https://www.learnopencv.com/understanding-alexnet/

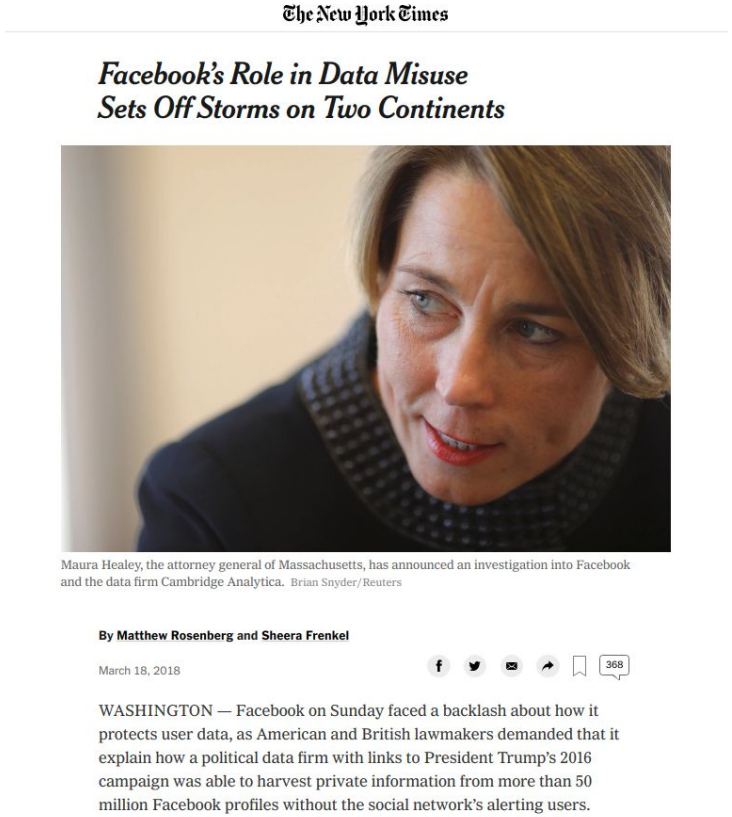
Ethics in Data Science

Privacy

1. Don’t collect or analyze personal data without consent!

2. Keep the data secure and f you don’t need the data, delete it!

3. If you release data or statistics, be careful - it may reveal more than you intend.



The New York Times

Facebook’s Role in Data Misuse Sets Off Storms on Two Continents

Maura Healey, the attorney general of Massachusetts, has announced an investigation into Facebook and the data firm Cambridge Analytica. Brian Snyder/Reuters

By Matthew Rosenberg and Sheera Frenkel

March 18, 2018

368

WASHINGTON — Facebook on Sunday faced a backlash about how it protects user data, as American and British lawmakers demanded that it explain how a political data firm with links to President Trump’s 2016 campaign was able to harvest private information from more than 50 million Facebook profiles without the social network’s alerting users.

https://www.nytimes.com/2018/03/18/us/cambridge-analytica-facebook-privacy-data.html

Privacy

Netflix released 100M anonymized movie ratings for their data science challenge.

In 16 days, researchers had identified specific users in the dataset.

1) Mapping movie scores to public accounts on IMDb.

2) Extracting the entire rental history based on a few rented movies.

Netflix tried to launch a sequel to the competition but were sued by a user.

movie	user	date	score
1	56	2004-02-14	5
1	25363	2004-03-01	3
2	855321	2004-07-29	3
2	44562	2004-07-30	4

Leaking Private Information

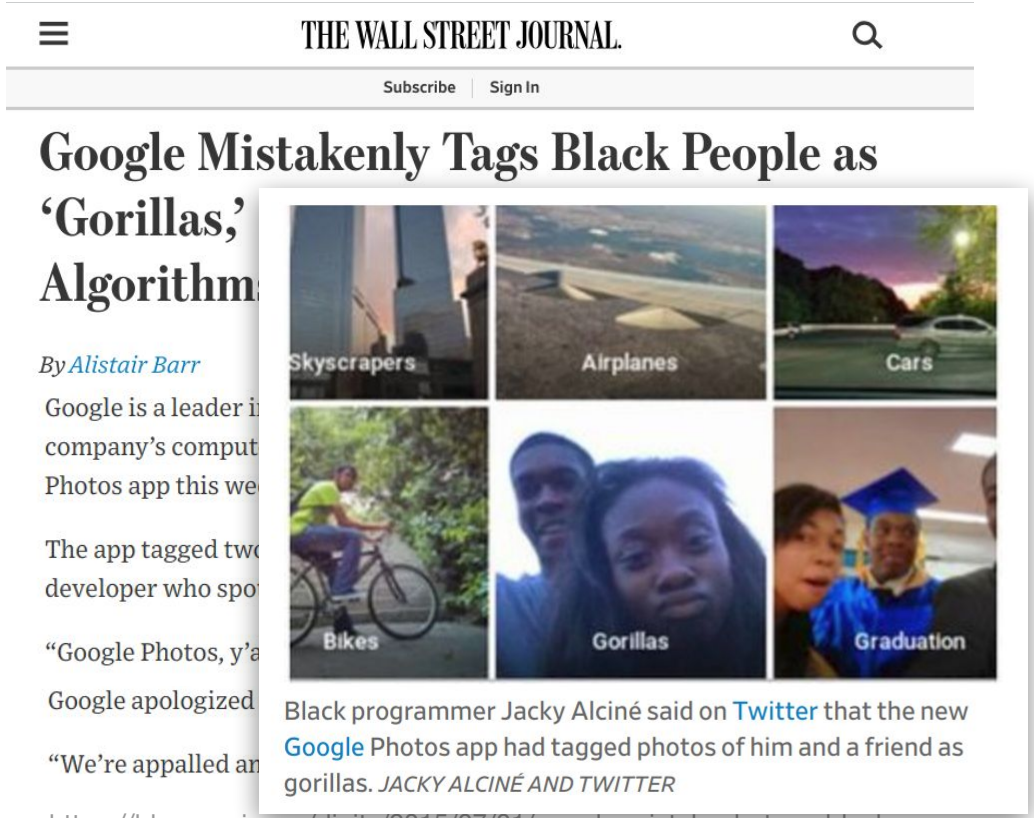
https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Bias in the Training Data

Machine learning models learn to do what they are trained to do.

The algorithms will pick up biases that are present in that dataset, whether good or bad.

Problem 1: The dataset is created with a bias and does not reflect the real task properly.



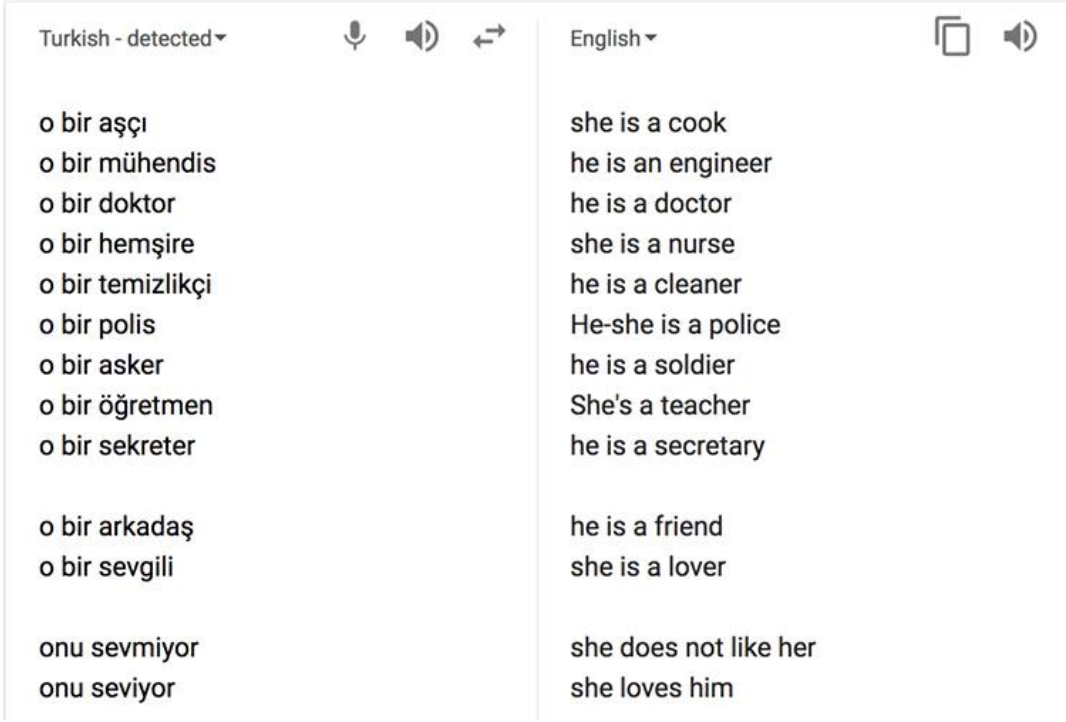
https://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

Bias in the Training Data

Problem 2: The data is representative but contains unwanted bias.

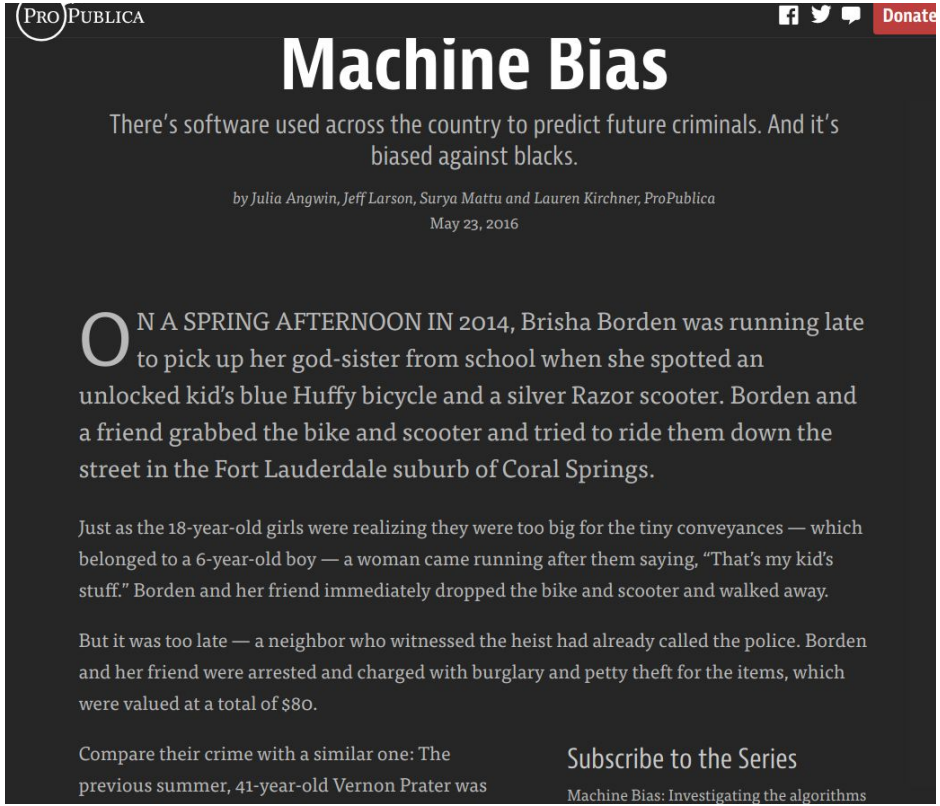
We don't want our models to be racist, sexist and discriminatory, even when the training data is.

Example: Turkish is a gender neutral language. Google Translate tries to infer a gender when translated into English.



https://twitter.com/seyyedreza/status/935291317252493312

Bias in the Training Data



PRO PUBLICA

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

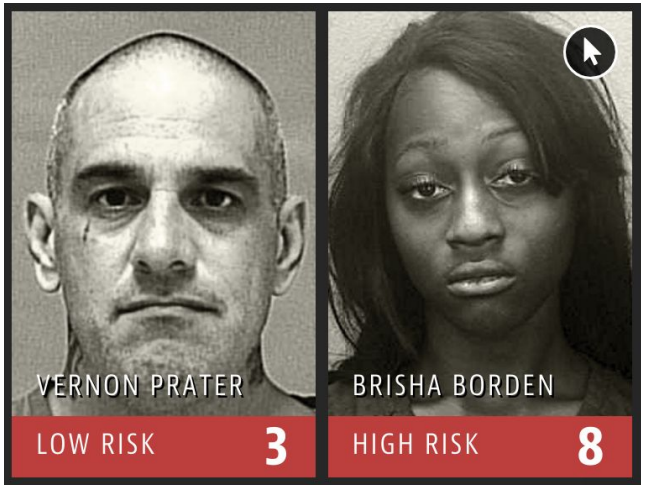
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was

Subscribe to the Series

Machine Bias: Investigating the algorithms



Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

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

Bias in the Training Data

Solution 1: just remove race as a feature.

Doesn't work!
Race is not used as a feature.

The problem: race is correlated with many other features that we may want to use in our machine learning system.

Solution 2: include race as a feature and explicitly correct for the bias.

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in 0, 1$$

Might need to accept lower accuracy for a more fair model.

Interpretability of our Models

For many applications we need to understand why the model produced a specific output.

EU law now requires that machine learning algorithms need to be able to explain their decisions.

Neural networks are notoriously unexplainable, black box models.

Bloomberg Opinion

Don't Grade Teachers With a Bad Algorithm

The Value-Added Model has done more to confuse and oppress than to motivate.

By Cathy O'Neil
15 May 2017, 12:00 BST *Corrected 16 May 2017, 15:01 BST*

Does not calculate. *Photographer: Paul J. Richards/AFP/Getty Images*

For more than a decade, a glitchy and unaccountable algorithm has been making life difficult for America's teachers. The good news is that its reign of terror might finally be drawing to a close.

LIVE ON BLOOMBERG
Watch Live TV >
Listen to Live Radio >
Bloomberg Television

Popular in Opinion

What If Democrats Have to Impeach the President?
by Francis Wilkinson
Pelosi would rather avoid the fight. Mueller may make that impossible.

Competition Is Dying, and Taking Capitalism With It
by Jonathan Tepper
We need a revolution to cast off monopolies and restore entrepreneurial freedom. First of two excerpts from "The Myth of Capitalism."

America Is Poorer Than It Thinks
by Noah Smith
Statistics don't quite capture the extent of U.S. poverty. A new measure could change that.

READ MORE FROM OPINION>

https://www.bloomberg.com/opinion/articles/2017-05-15/don-t-grade-teachers-with-a-bad-algorithm

Replicability of Findings

Replicability

We test a lot of hypotheses but report only the significant results.

This is fine - we can't publish a paper for every relation that doesn't hold.

But we need to be aware of this selection when analyzing the results.

Studies trying to replicate existing findings are rare and often fail.

Attempt to replicate major social scientific findings of past decade fails

Scientists and the design of experiments under scrutiny after a major project fails to reproduce results of high profile studies

▲ One finding which this study was unable to replicate was that people who viewed a picture of Rodin's sculpture The Thinker subsequently reported weaker religious beliefs. Photograph: Alamy

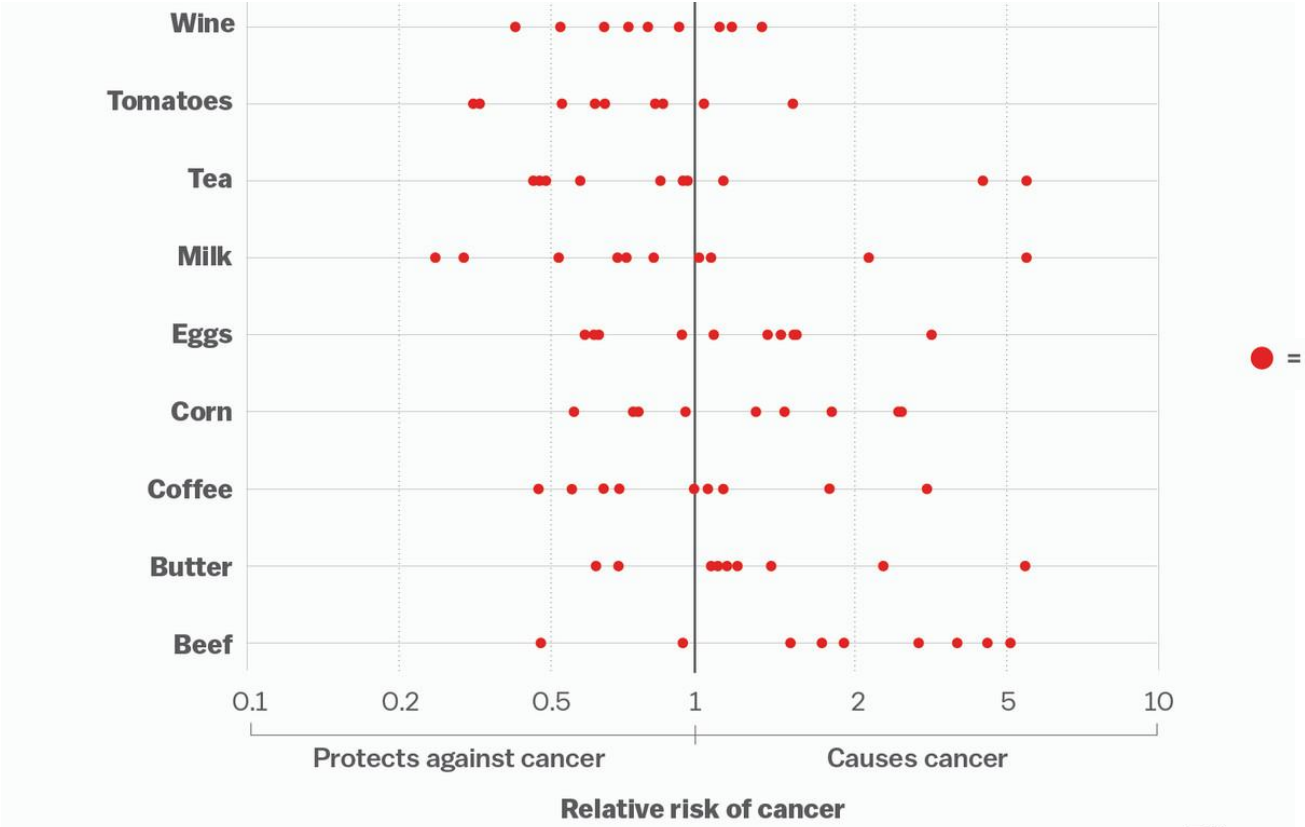
Some of the most high profile findings in social sciences of the past decade do not stand up to replication, a major investigation has found.

The project, which aimed to repeat 21 experiments that had been published in Science or Nature - science's two preeminent journals - found that only 13 of the original findings could be reproduced.

The research, which follows similar efforts in psychology and biomedical science, raises fresh concerns over the reliability of the scientific literature. However, the project's leaders say their results do not reflect a "crisis" in the social sciences.

https://www.theguardian.com/science/2018/aug/27/attempt-to-replicate-major-social-scientific-findings-of-past-decade-fails

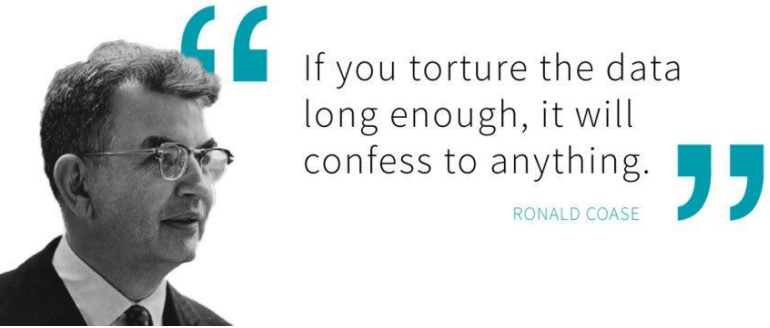
Contradicting Studies



https://www.vox.com/2015/3/23/8264355/research-study-hype

P-hacking

P-hacking is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no underlying effect.



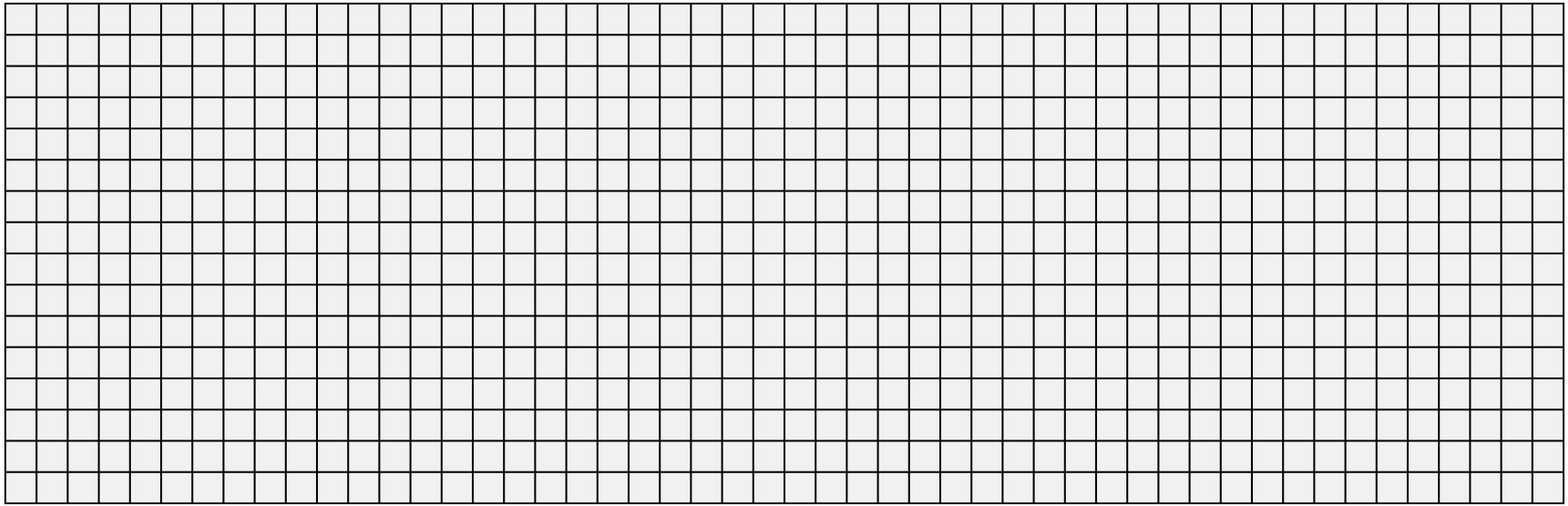
Done by running large numbers of experiments and only paying attention to the ones that come back with significant results.

Also known as '*data dredging*', '*data snooping*', '*data fishing*', etc.

Statistical significance is defined as being less than 5% likely that the result is due to randomness ($p < 0.05$).

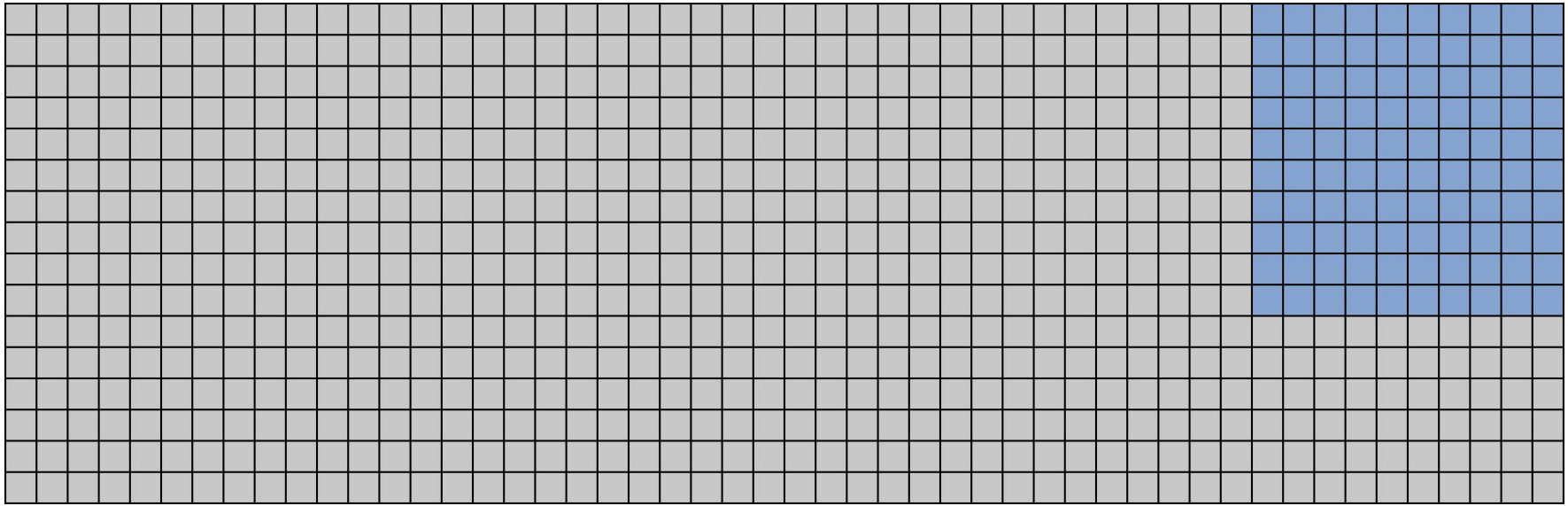
That means we accept that some "significant" results are going to be false positives!

P-hacking



Total 800 hypotheses to test

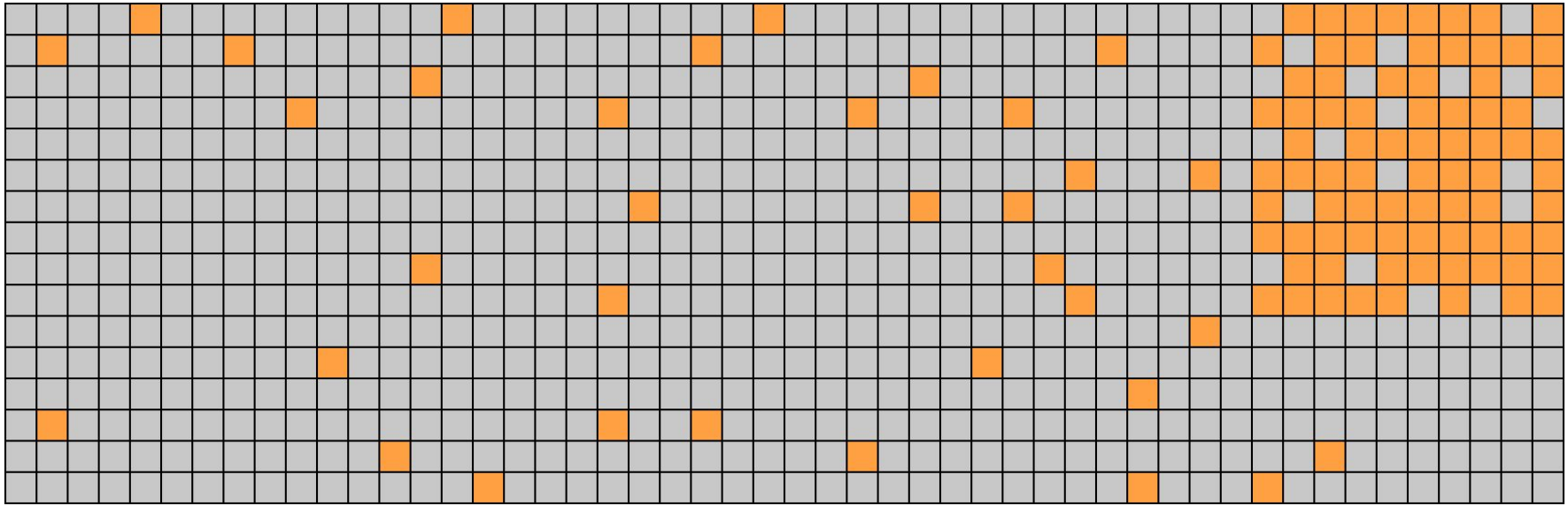
P-hacking



The true underlying distribution:

- Something going on in 100 configurations
- Nothing going on in the rest

P-hacking



For each hypothesis we test:

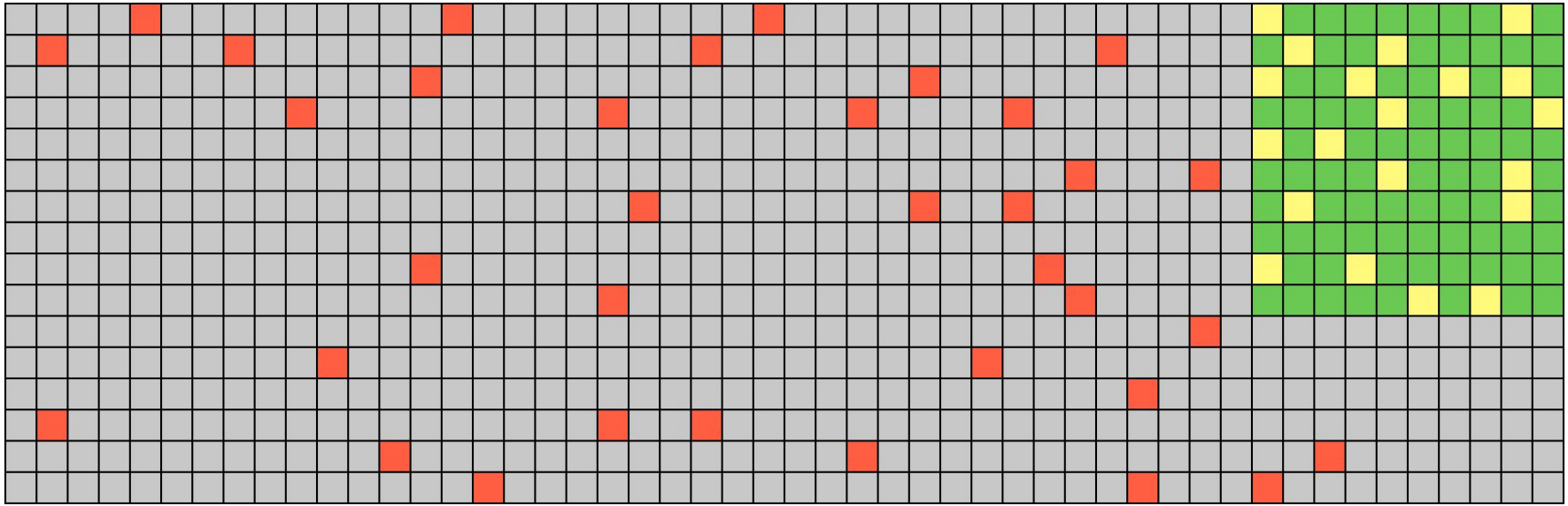
We discover something

We don't discover anything

P(false positive) = 0.05

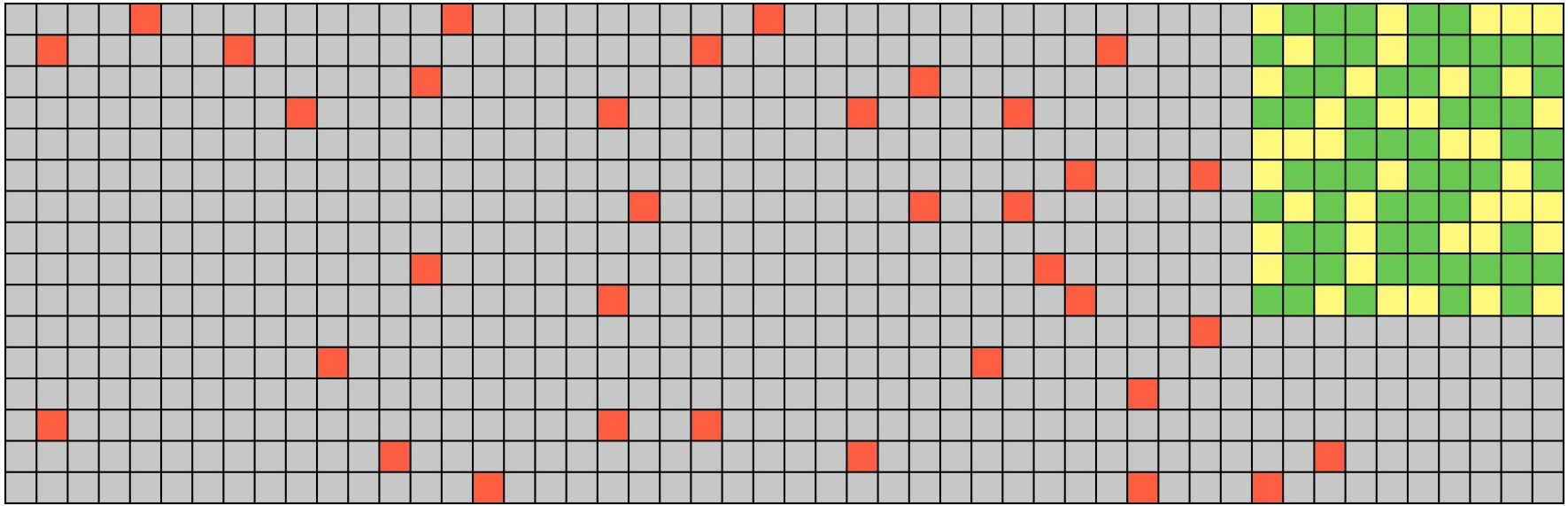
P(false negative) = 0.2

P-hacking



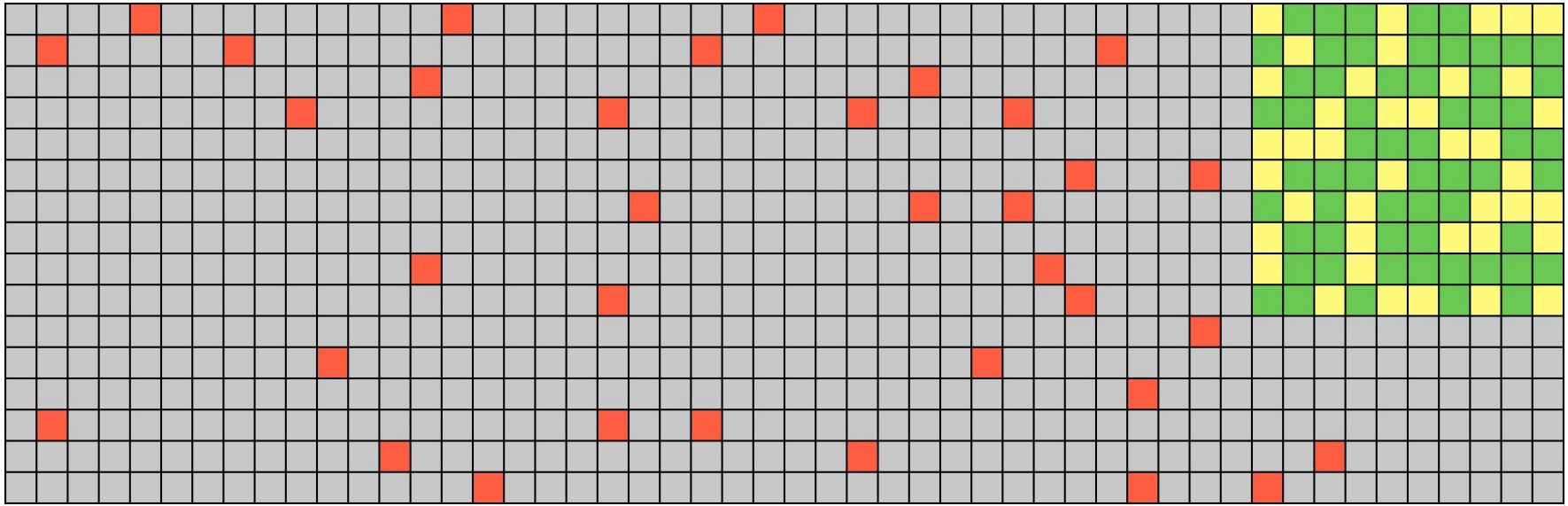
We made 80 true discoveries
We made 35 false discoveries
False Discovery Proportion = 35 / 115 = 0.3

P-hacking



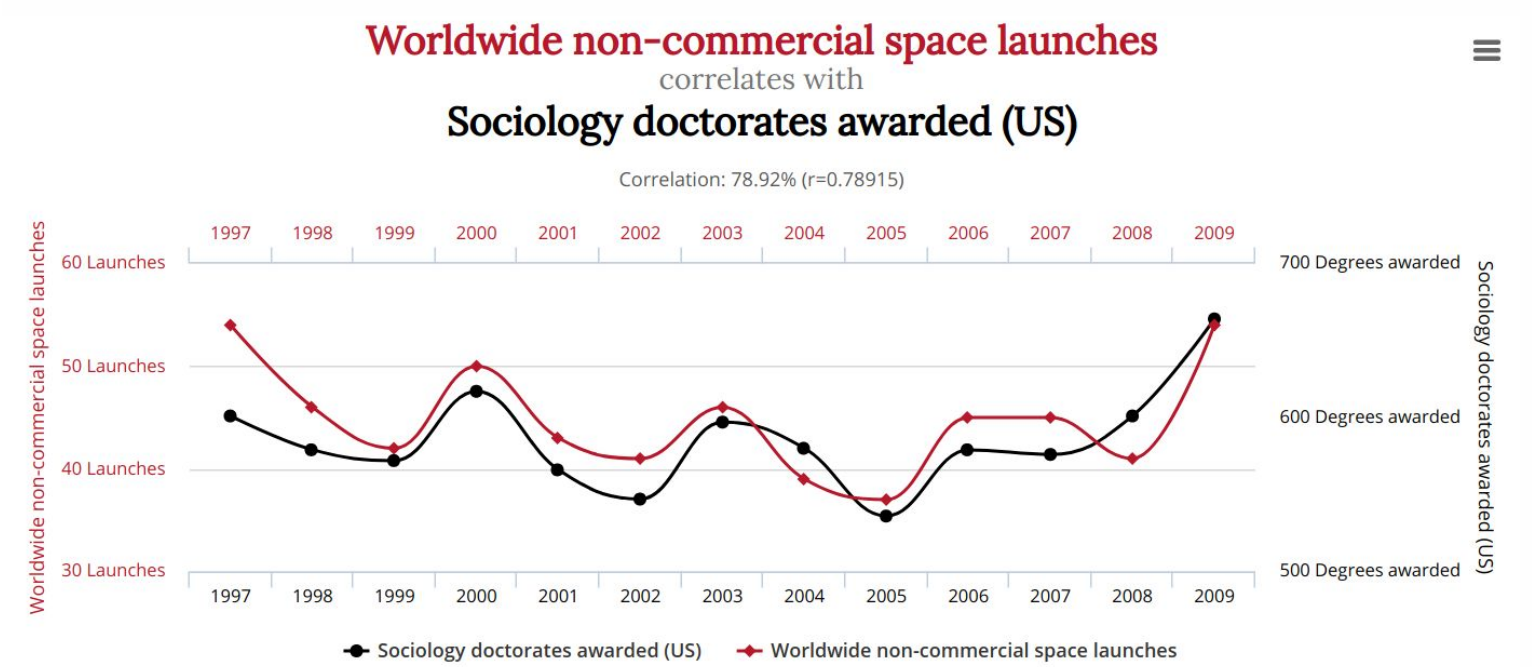
If P(false negative) = 0.4 and P(false positive) = 0.05
We made 60 true discoveries
We made 35 false discoveries
False Discovery Proportion = 35 / 95 = 0.37

P-hacking



If P(false negative) = 0.4 and P(false positive) = 0.05 over 1600 experiments
We made 60 true discoveries
We made 75 false discoveries
False Discovery Proportion = 75 / 135 = 0.56

Spurious Correlations



http://www.tylervigen.com/spurious-correlations

Spurious Correlations

A sample “study” with 54 people, searching over 27,716 possible relations.

Our shocking new study finds that ...

EATING OR DRINKING	IS LINKED TO	P-VALUE
Raw tomatoes	Judaism	<0.0001
Egg rolls	Dog ownership	<0.0001
Energy drinks	Smoking	<0.0001
Potato chips	Higher score on SAT math vs. verbal	0.0001
Soda	Weird rash in the past year	0.0002
Shellfish	Right-handedness	0.0002
Lemonade	Belief that “Crash” deserved to win best picture	0.0004
Fried/breaded fish	Democratic Party affiliation	0.0007
Beer	Frequent smoking	0.0013
Coffee	Cat ownership	0.0016
Table salt	Positive relationship with Internet service provider	0.0014

https://fivethirtyeight.com/features/you-cant-trust-what-you-read-about-nutrition/

Strategies Against P-hacking

Distinguish between verifying a hypothesis and exploring the data.

Benjamini & Hochberg (1995) offer an adaptive p-value:

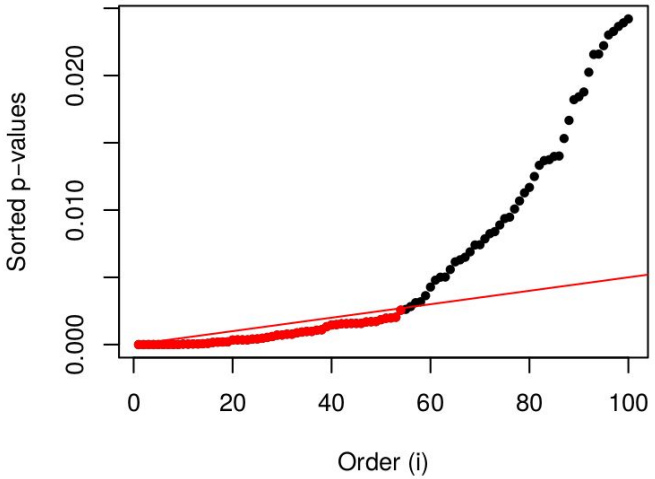
1. Rank p-values from M experiments.

$$p_1 \leq p_2 \leq p_3 \leq \ldots \leq p_M$$

2. Calculate the Benjamini-Hochberg critical value for each experiment.

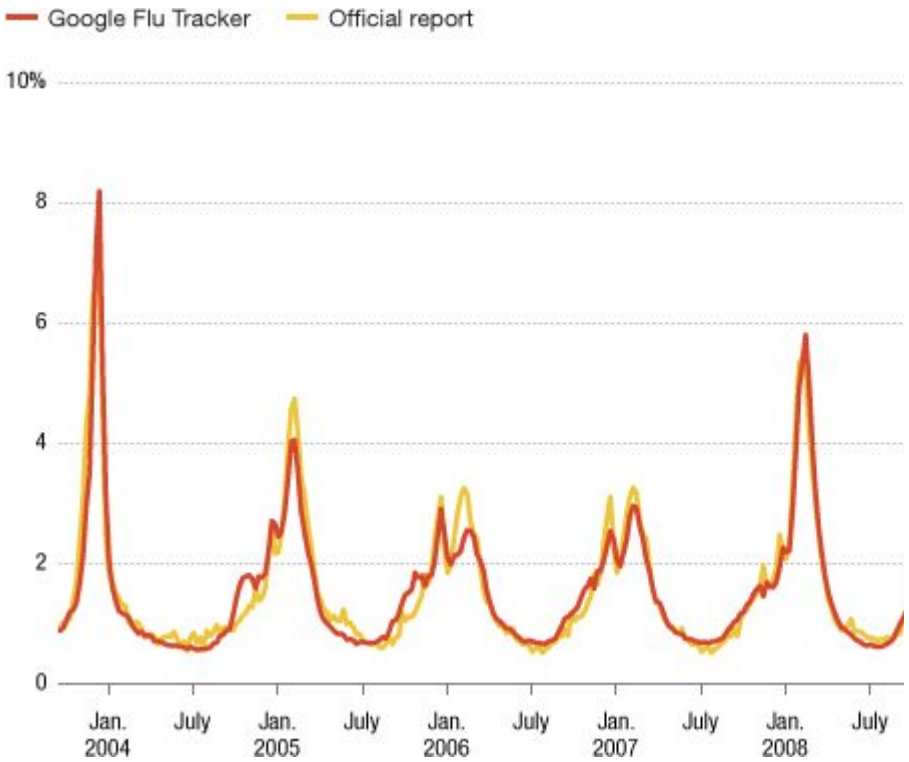
$$z_i = 0.05 \frac{i}{M}$$

3. Significant results are the ones where the p-value is smaller than the critical value.



https://web.stanford.edu/class/stats101

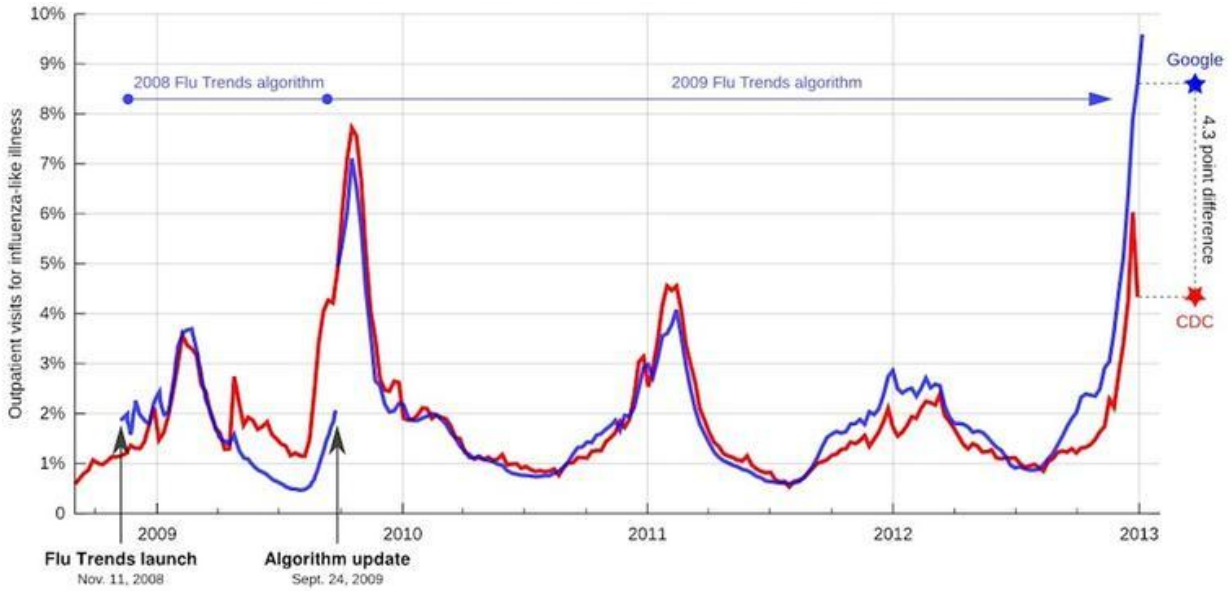
Google Flu Trends



Predicting flu epidemics based on online behaviour

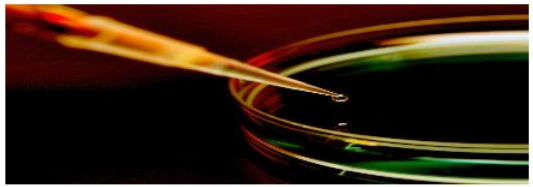
https://www.npr.org/sections/health-shots/2014/03/13/289802934/googles-flu-tracker-suffers-from-sniffles

Google Flu Trends



DAVID LAZER AND RYAN KENNEDY OPINION 10.01.15 07:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS



RAFE SWAN/GETTY IMAGES

EVERY DAY, MILLIONS of people use Google to dig up information that drives their daily lives, from how long their commute will be to how to treat their child's illness. This search data reveals a lot about the searchers: their wants, their needs, their concerns—extraordinarily valuable information. If these searches accurately reflect what is happening in people's lives, analysts could use this information to track diseases, predict sales of new products, or even anticipate the results of elections.

http://www.wbur.org/commonhealth/2013/01/13/google-flu-trends-cdc
https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/