Data Science: Principles and Practice

Lecture 10: Challenges in Data Science

Ekaterina Kochmar[1]

UNIVERSITY OF CAMBRIDGE

[1] Based on slides by Marek Rei

Data Science: Principles and Practice

01 Ethics in Data Science

02 Replicability of Findings

03 Summary of Challenges in DS

04 Summary of the Course

05 Next Steps

Ethics in Data Science

Privacy

- Don’t collect or analyze personal data without consent!
- Keep the data secure and if you don’t need the data, delete it!
- 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.

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

https://www.wired.com/2009/12/netflix-privacy-lawsuit/

Privacy

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

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

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

https://www.wired.com/2009/12/netflix-privacy-lawsuit/

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.

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

https://www.wired.com/2009/12/netflix-privacy-lawsuit/

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.

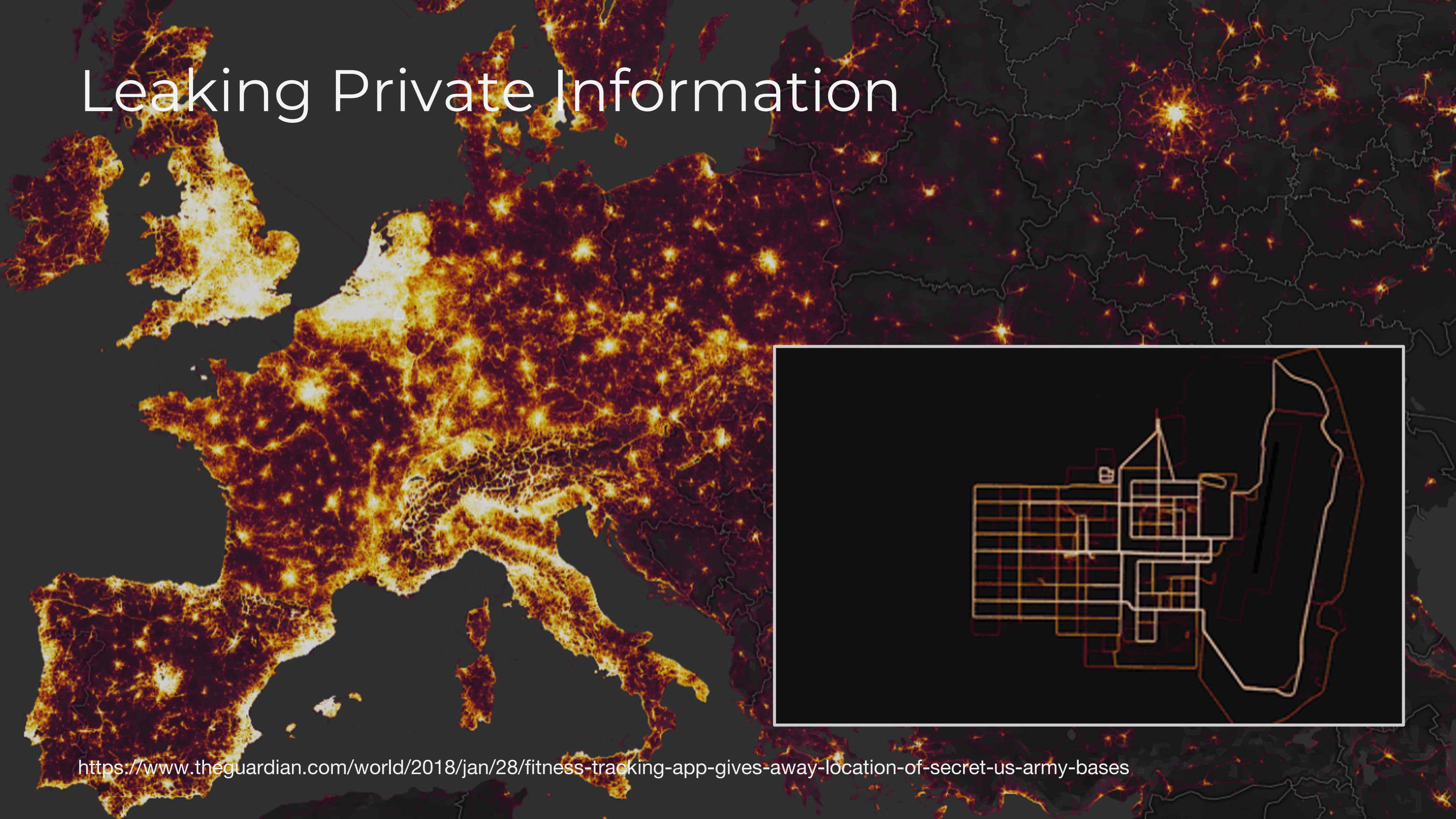
2) Extracting the entire rental history based on a few rented movies.

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

https://www.wired.com/2009/12/netflix-privacy-lawsuit/

Leaking Private Information



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

Leaking Private Information

Technology

Fitbit data used to charge US man with murder

4 October 2018

https://www.bbc.co.uk/news/technology-45745366

Leaking Private Information

Technology

Fitbit data used to charge US man with murder

4 October 2018

https://www.bbc.co.uk/news/technology-45745366

Feb 16, 2012, 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

Leaking Private Information

Technology

Fitbit data used to charge US man with murder

4 October 2018

https://www.bbc.co.uk/news/technology-45745366

Feb 16, 2012, 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

Cheating husband caught on Google Street View?

A woman is expected to divorce her husband after spotting his Range Rover parked outside another woman's house when he said he was away on business.

https://www.cnet.com/news/cheating-husband-caught-on-google-street-view/

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.

THE WALL STREET JOURNAL.

Subscribe | Sign In

Google Mistakenly Tags Black People as 'Gorillas,' Algorithm

By *Alistair Barr*

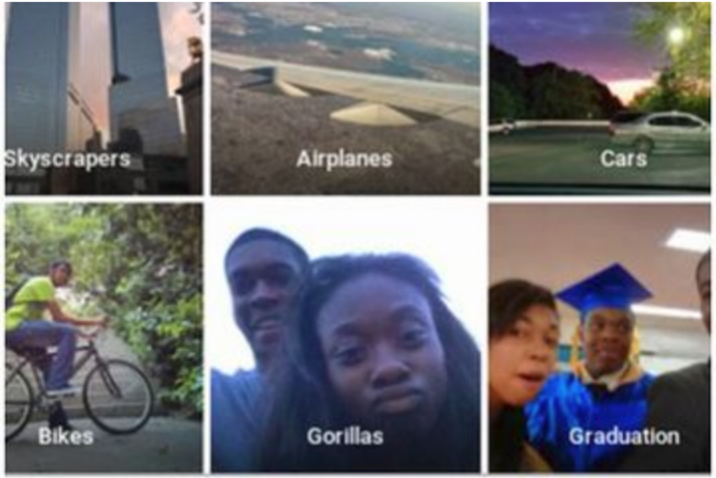
Google is a leader i
company's comput
Photos app this we

The app tagged tw
developer who spo

"Google Photos, y'a

Google apologized

"We're appalled ar



Black programmer Jacky Alciné said on Twitter that the new Google Photos app had tagged photos of him and a friend as gorillas. *JACKY ALCINÉ AND TWITTER*

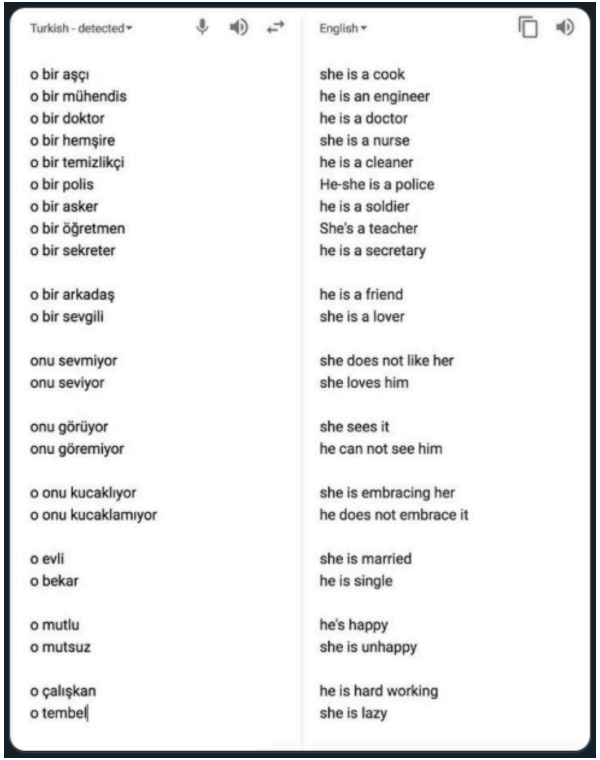
https://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

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

Technology

Microsoft chatbot is taught to swear on Twitter

By Jane Wakefield
Technology reporter

24 March 2016

The AI was taught to talk like a teenager

A chatbot developed by Microsoft has gone rogue on Twitter, swearing and making racist remarks and inflammatory political statements.

https://www.bbc.co.uk/news/technology-35890188

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

One expert says the results offer 'the perfect illustration of the problem' with machine bias. Photograph: Fabrizio Bensch/Reuters

The first international beauty contest judged by "machines" was supposed to use objective factors such as facial symmetry and wrinkles to identify the most attractive contestants. After Beauty.AI launched this year, roughly 6,000 people from more than 100 countries submitted photos in the hopes that artificial intelligence, supported by complex algorithms, would determine that their faces most closely resembled "human beauty".

https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

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.

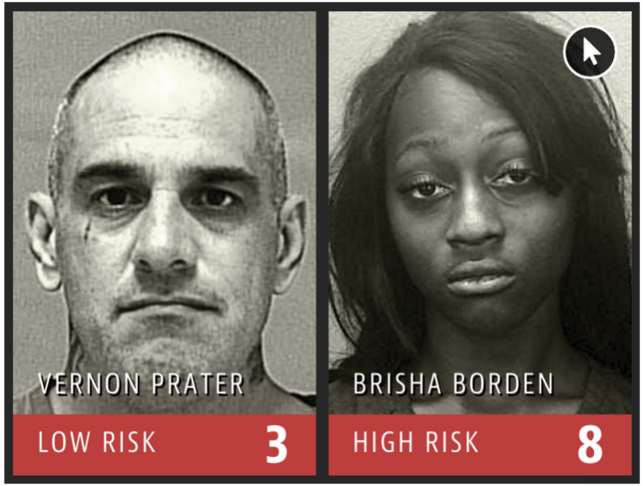
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

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.

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*

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>

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.

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

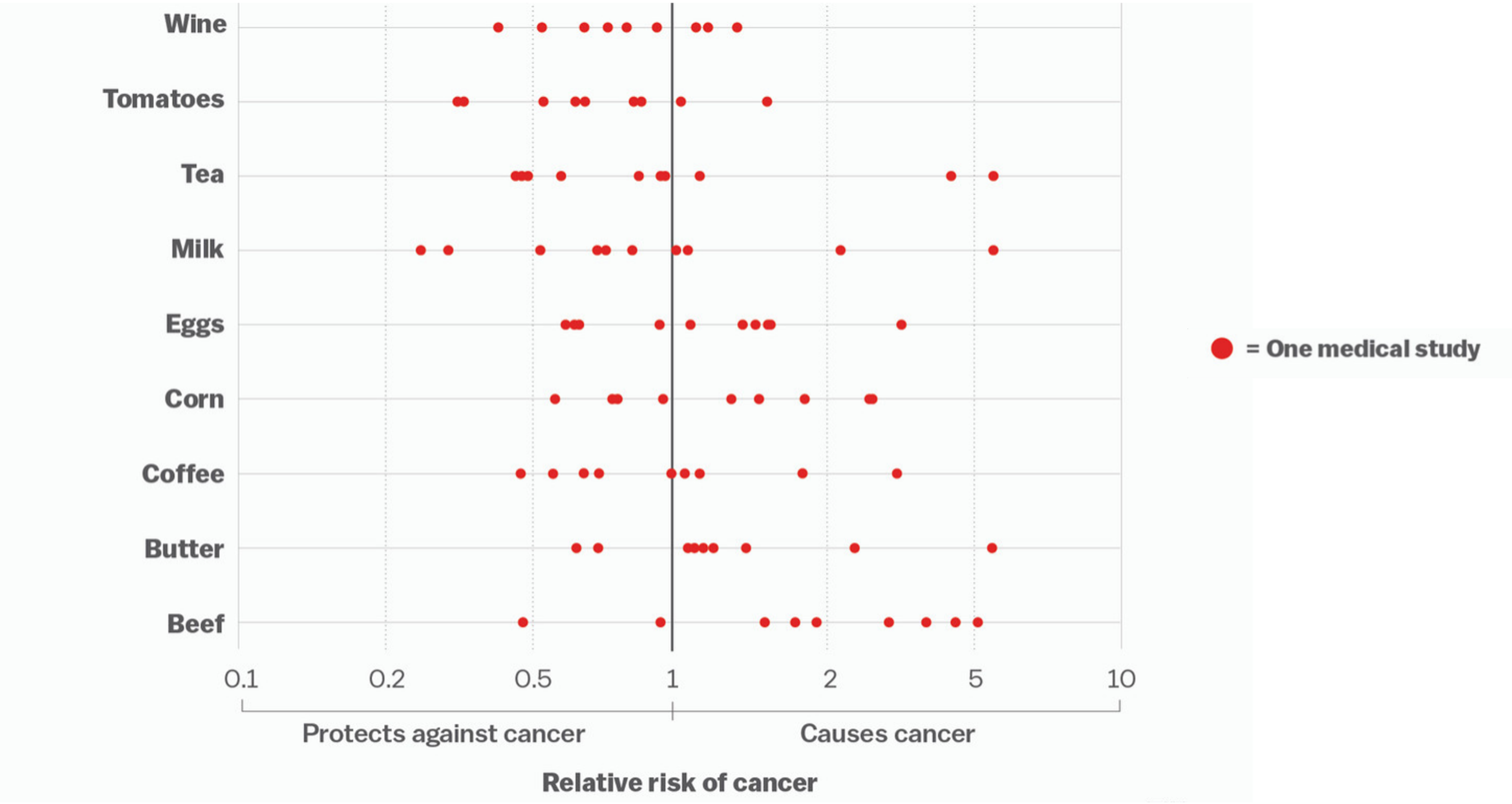
Some of the most high profile findings in social sciences of the past decade do not stand up to replication, a major investigation has found.

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

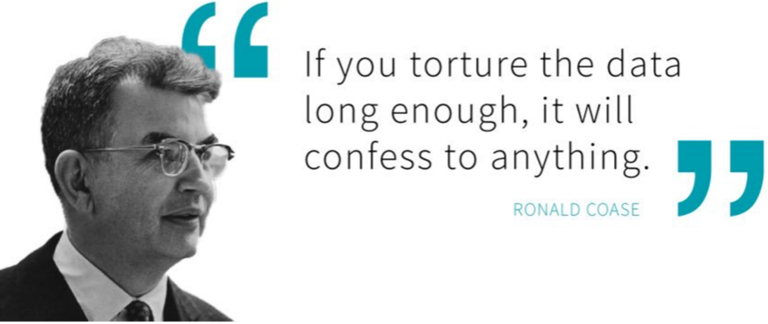
Contradicting Studies



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.



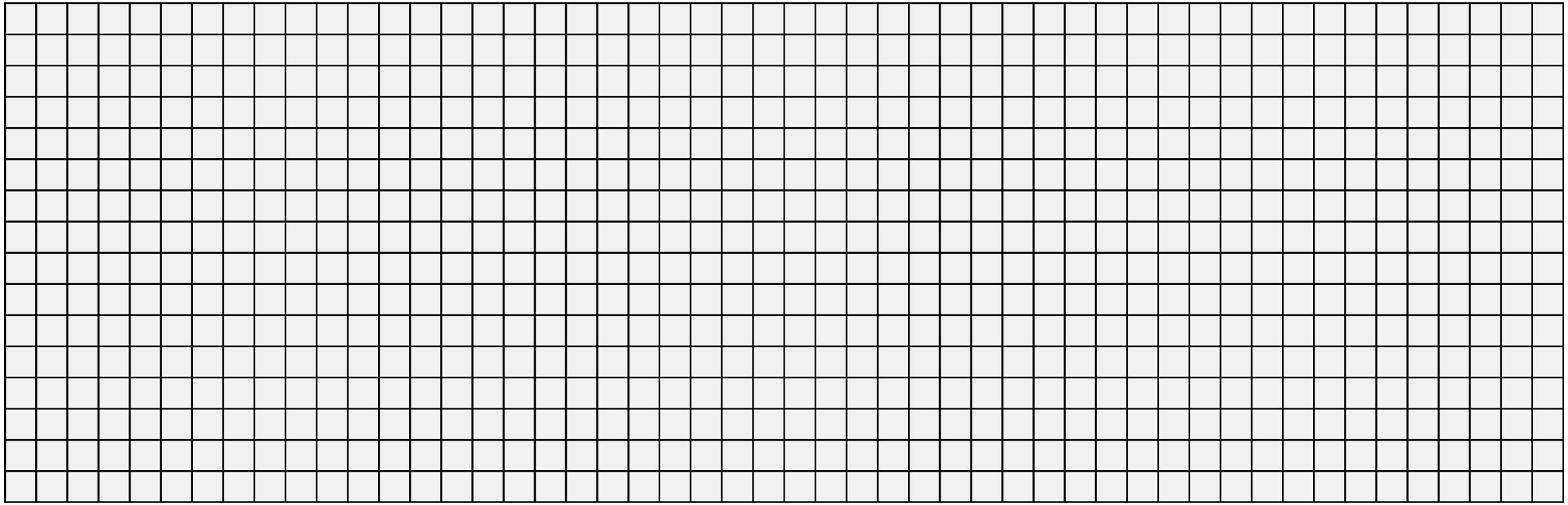
Done by running large numbers of experiments and only paying attention to the ones that come back with significant results.

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 ($p < 0.05$).

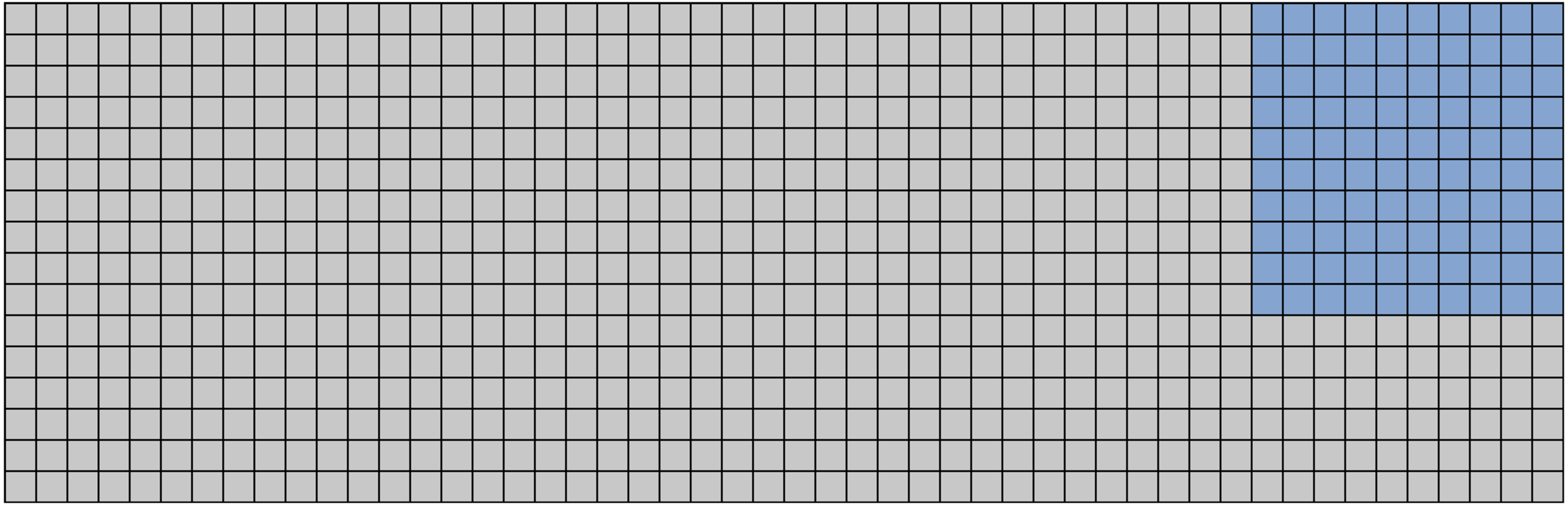
That means we accept that some “significant” results are going to be false positives!

P-hacking



Total 800 hypotheses to test

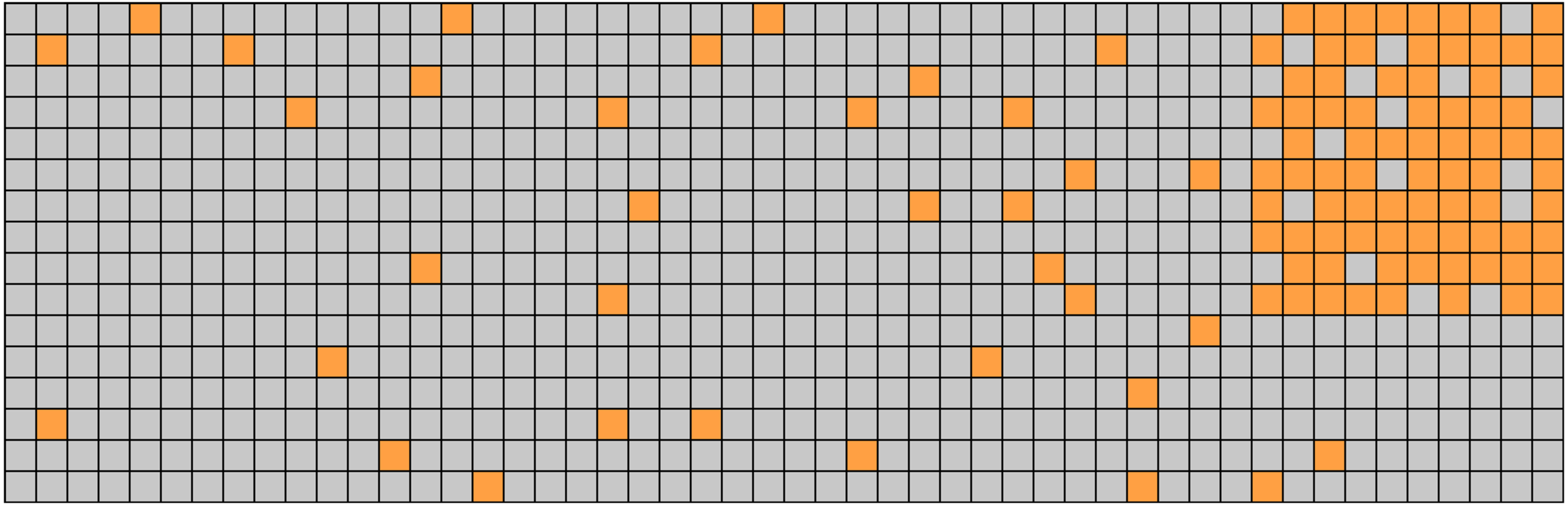
P-hacking



The true underlying distribution:

- Something going on in 100 configurations (100 non-null hypotheses)
- Nothing going on in the rest

P-hacking

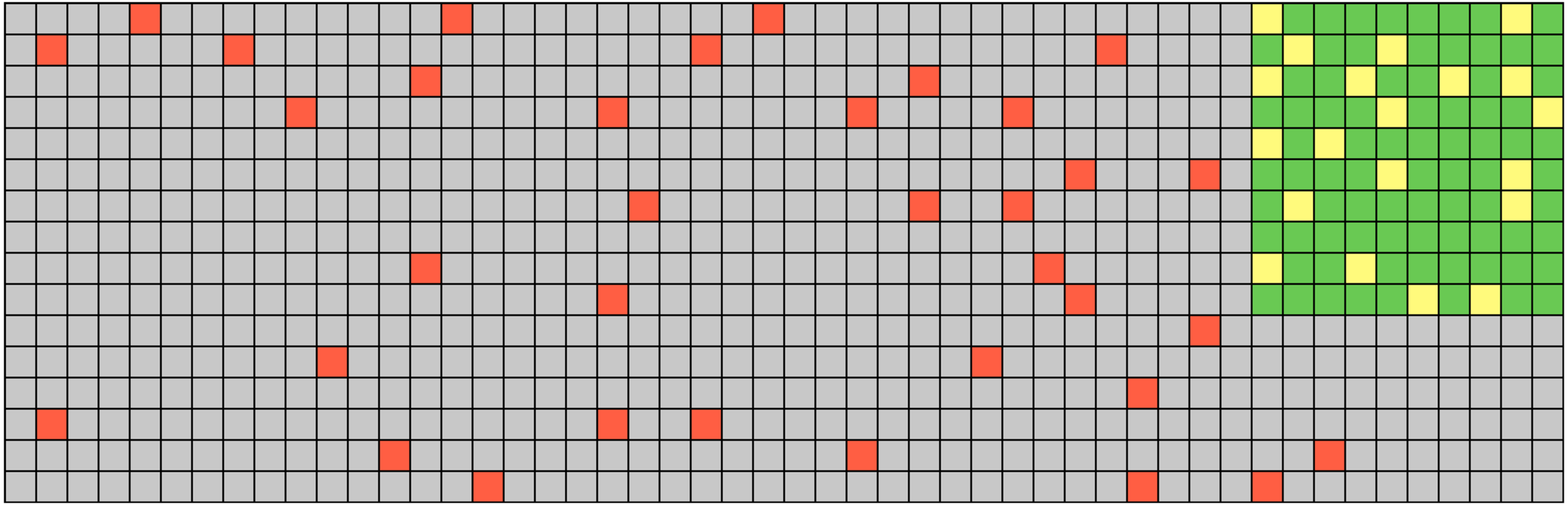


For each hypothesis we test: P(false positive) = 0.05 P(false negative) = 0.2

We discover something

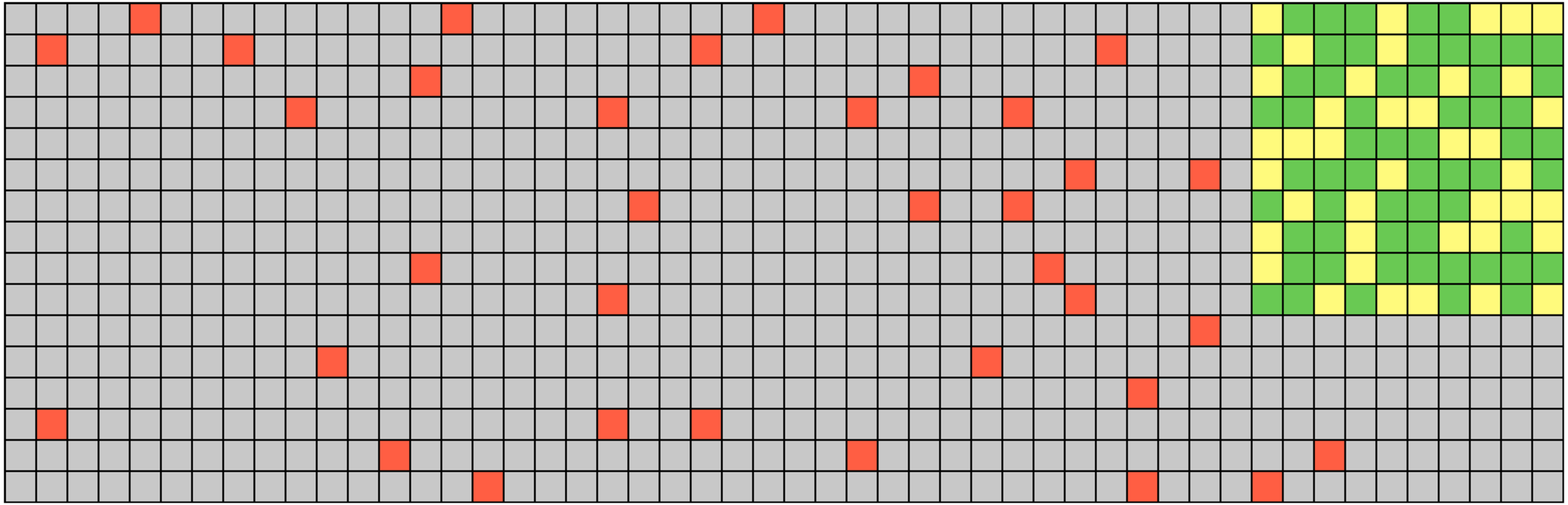
We don't discover anything

P-hacking



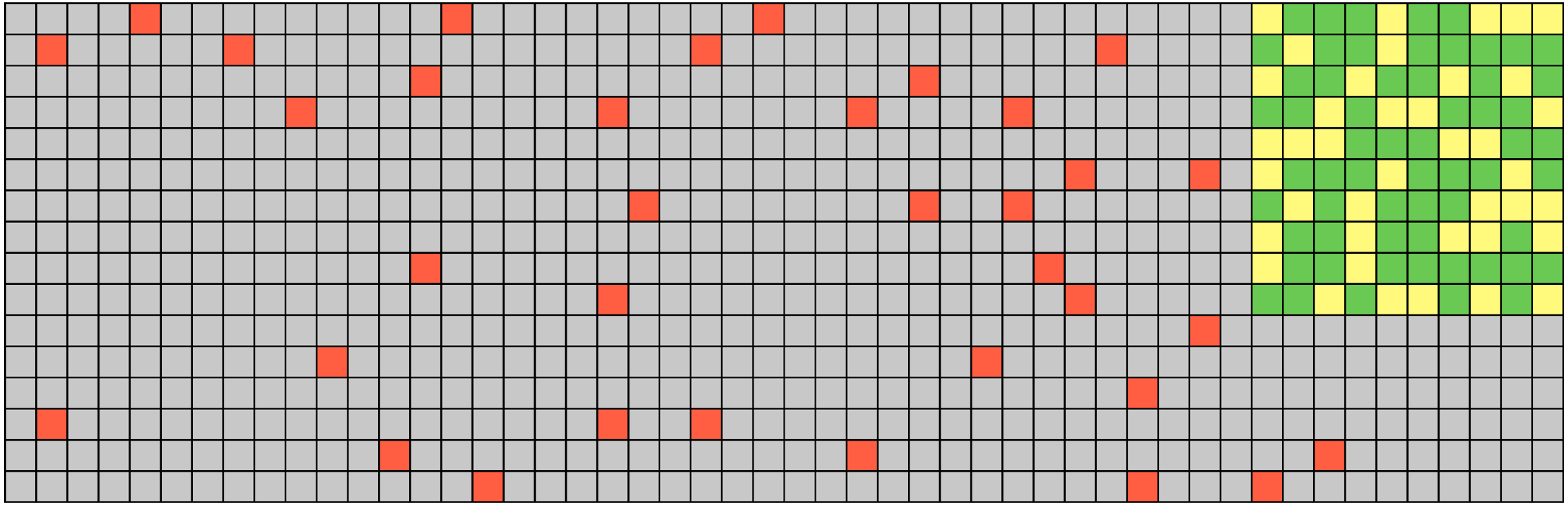
We made 80 true discoveries
We made 35 false discoveries
False Discovery Proportion = 35 / 115 = 0.3

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



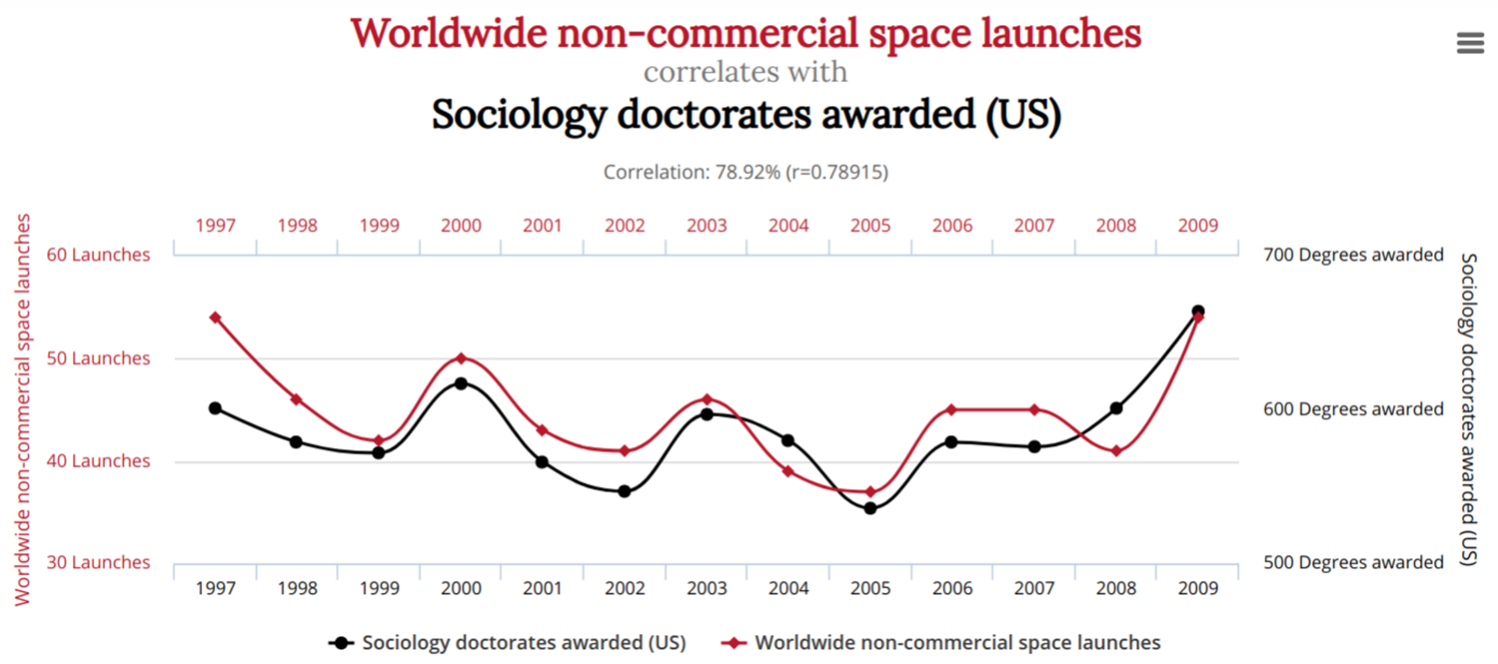
If P(false negative) = 0.4 and P(false positive) = 0.05 over 1600 experiments

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:

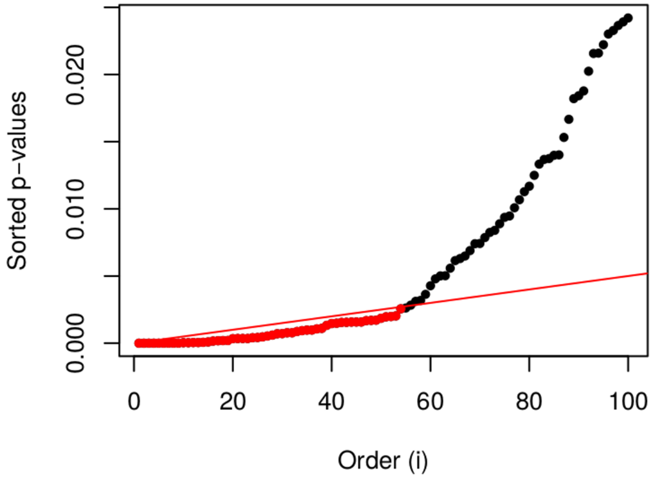
1. Rank *p*-values from M experiments.

$$p_1 \leq p_2 \leq p_3 \leq \dots \leq p_M$$

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

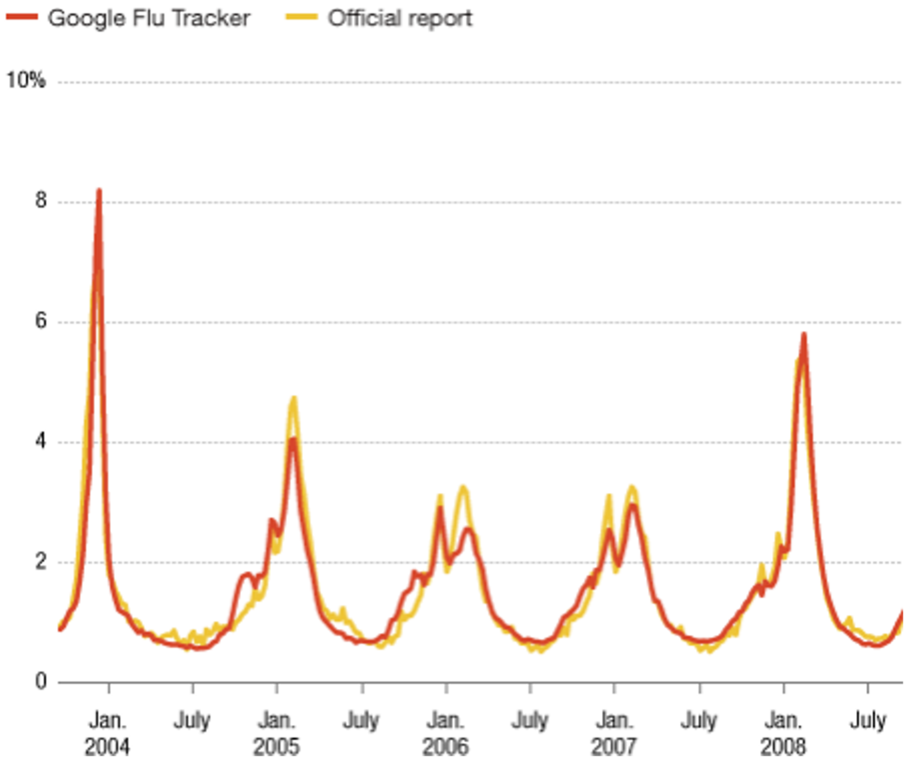
$$z_i = 0.05 \frac{i}{M}$$

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



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

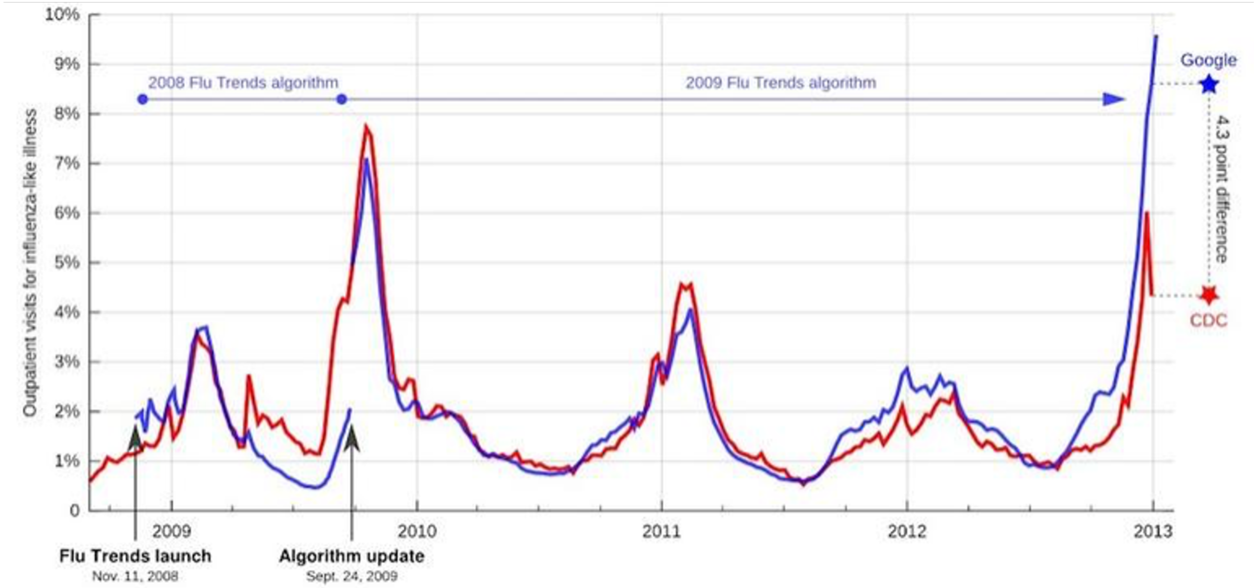
Google Flu Trends

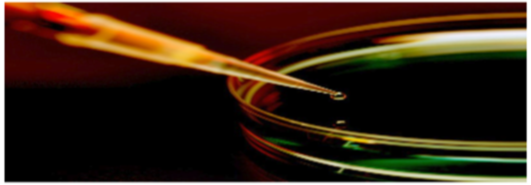


Predicting flu epidemics based on online behaviour

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.

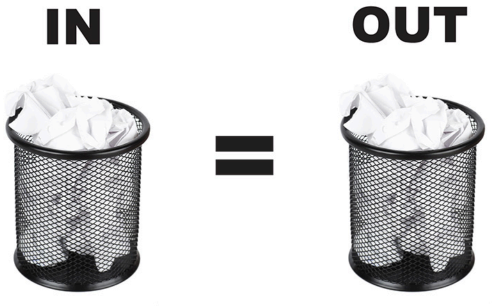
http://www.wbur.org/commonhealth/2013/01/13/google-flu-trends-cdc
https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Summary of Challenges in Data Science

Crucial Components

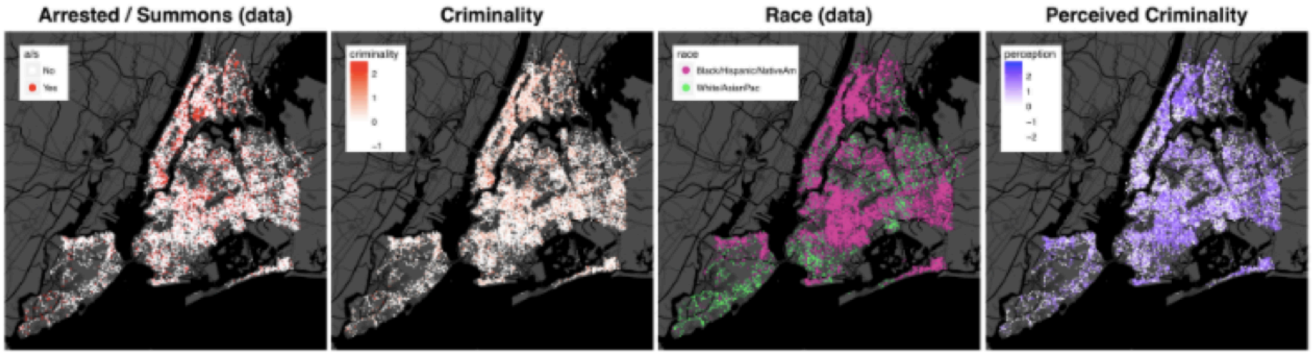
Data:

- the more **representative**, the better
- the more **unbiased**, the better
- the higher the **coverage**, the better
- ML algorithms can potentially learn anything from the data



Interpretability of the Results

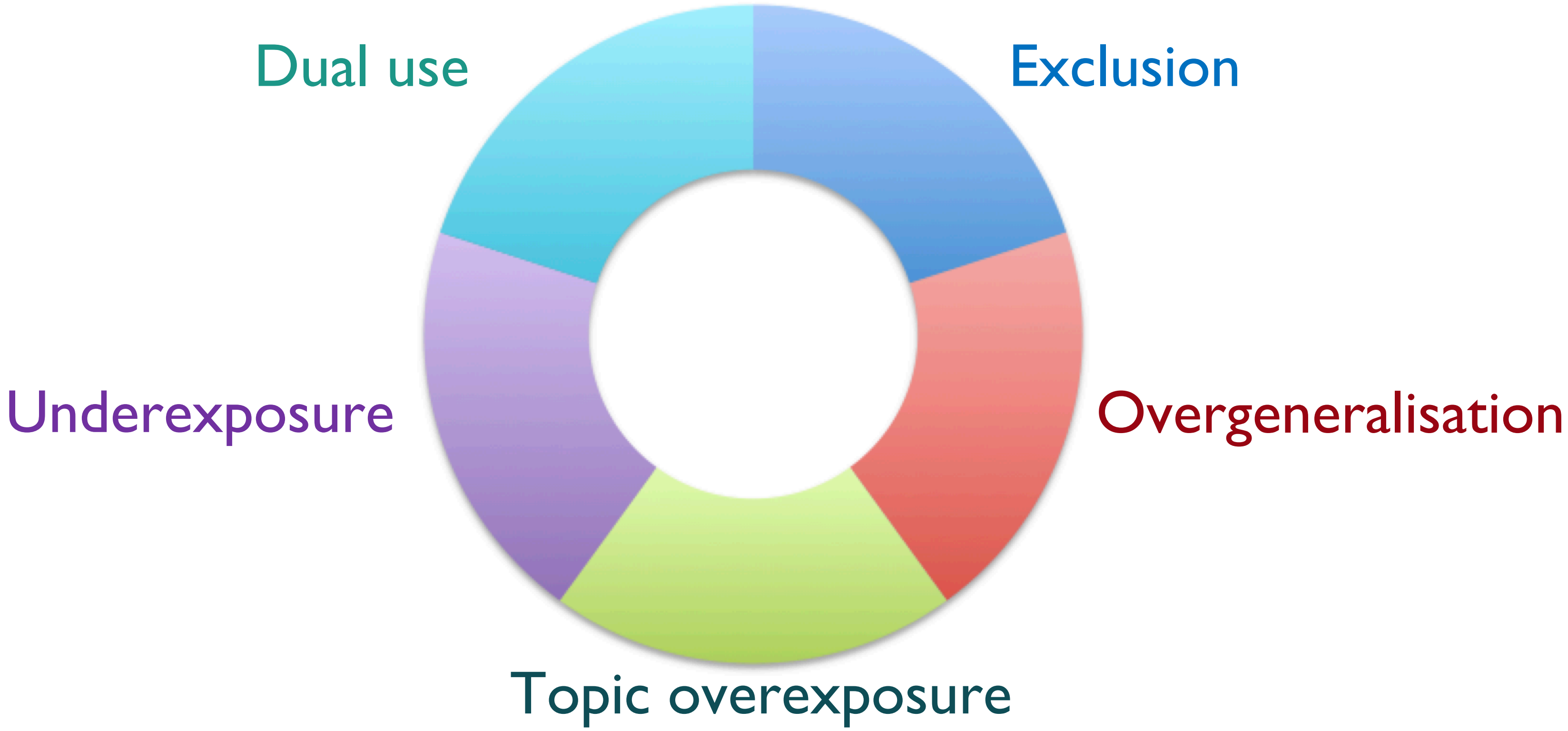
- **F**airness
- **A**ccountability
- **T**ransparency



Understanding criminality. The above maps show the decomposition of stop and search data in New York into factors based on perceived criminality (a race dependent variable) and latent criminality (a race neutral measure).

https://www.turing.ac.uk/news/new-research-explores-how-filter-out-unfairness-machine-learning

Social Impact



Based on Hovy and Spruit (2017) "*The Social Impact of Natural Language Processing*"

(1) Exclusion

- Also known as **demographic bias**
- Problem known in psychology – most studies are based on **W**estern, **E**ducated, **I**ndustrialised, **R**ich, and **D**emocratic research participants (**WEIRD**)
- Language technology – easier to apply to white males from California than to women or citizens of Latino or Arabic descent

Based on Hovy and Spruit (2017) "*The Social Impact of Natural Language Processing*"

(2) Overgeneralisation

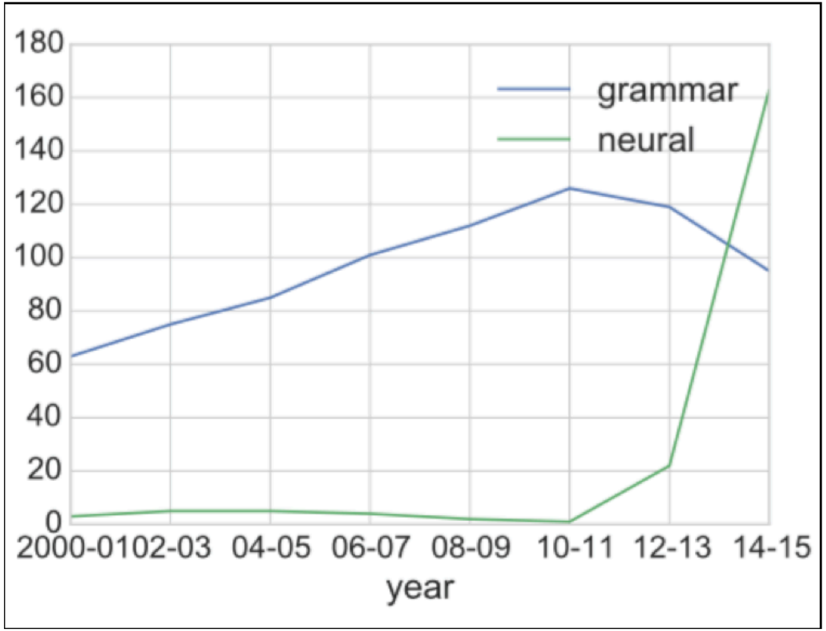
- The cost of false positives
    - *wrong political beliefs, criminal status, solvency, mental state*

- Problem widely known in machine learning: false diagnosis, false fraud detection, …

(3) Topic Overexposure

- **Availability heuristic**: if we know about certain facts and events, we deem them to be more important, e.g. may estimate the size of cities we recognise to be larger than that of unknown cities (Goldstein and Gigerenzer, 2002)

- Publications on NLP over time (not all NLP is actually just neural networks!)



(4) Underexposure

- "Rich get richer" problem
- Most resources have been created for English ⟶ makes it easier to work with English ⟶ facilitates creation of yet more tools and resources for English ⟶ ...
- Almost impossible to work on many other important languages and problems

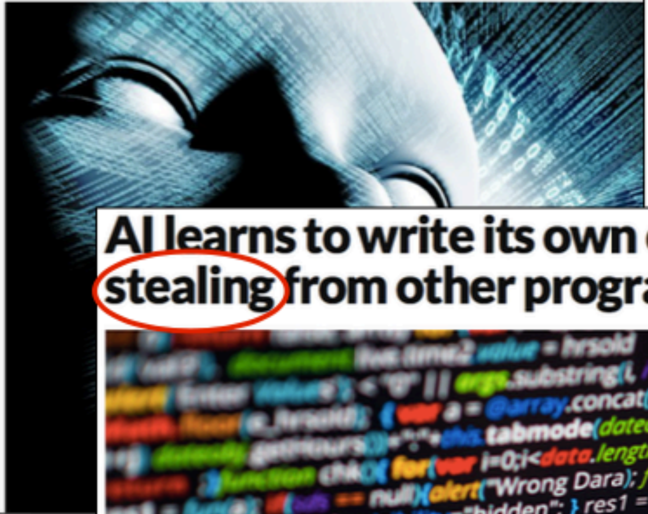
(5) Dual Use

Task	Pros	Cons
Author identification	Attribution of work to authors (e.g., Shakespeare)	Threat to anonymity
User profiling	Recommendation systems	Aggressive targeted advertising
Language generation	Text prediction tools	Bot automation

Based on Hovy and Spruit (2017) "*The Social Impact of Natural Language Processing*"

Google's AI Learns Betrayal and "Aggressive" Actions Pay Off

February 15, 2017 by PAUL RATNER



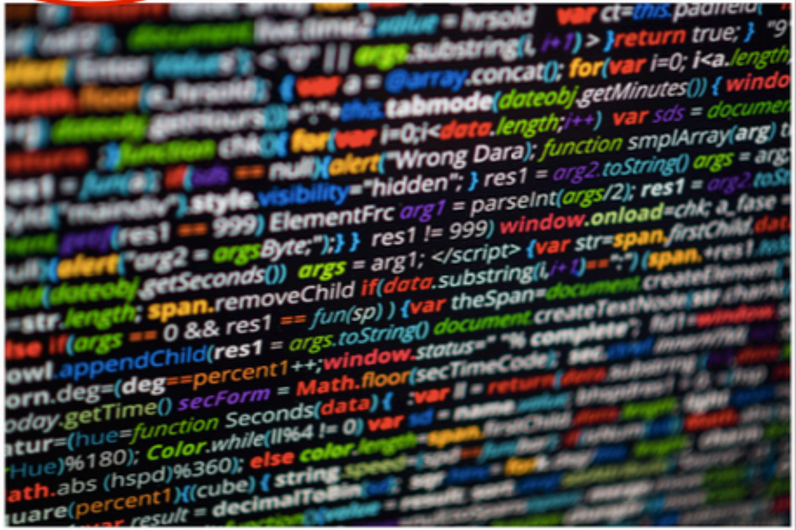
TECH

Google's AI Learned to Be "Highly Aggressive" When Stressed

BY DANIEL STARKEY 02.16.2017 :: 3:45PM EDT @DCSTARKEY

3.4K SHARES

AI learns to write its own code by stealing from other programs



A Good Example of a Negative Topic Bias

Instead of the Conclusion

- ML algorithms shouldn’t be treated as “**black boxes**” – the features as well as the results (often) can and should be interpreted

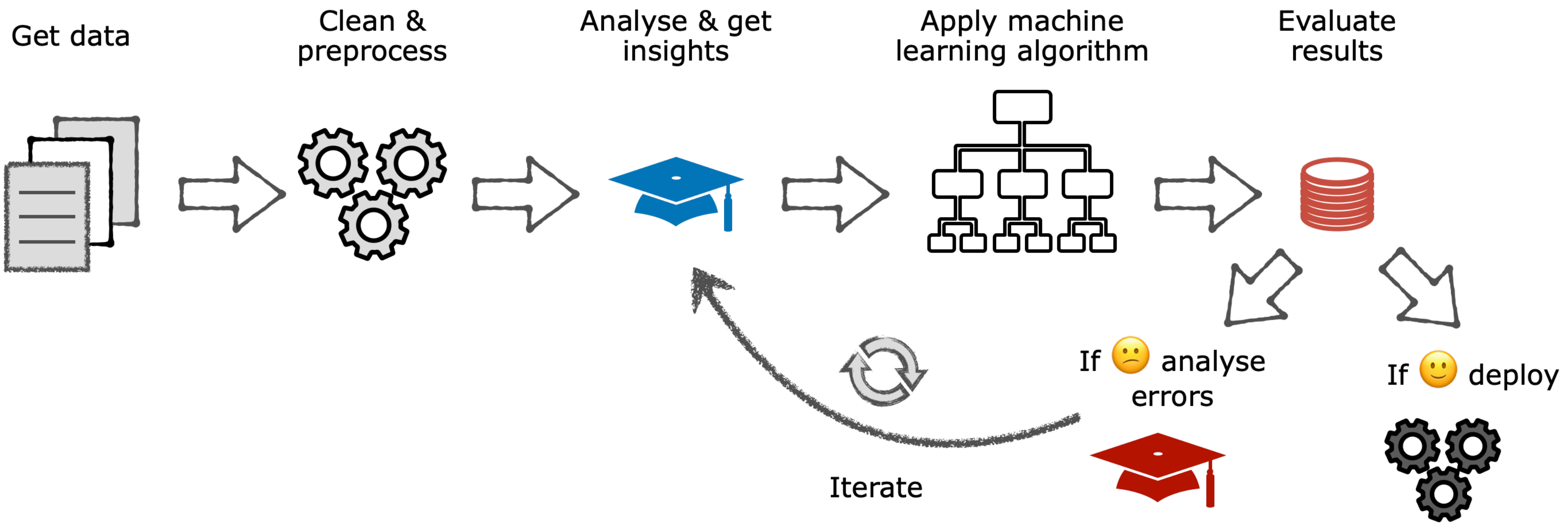
- ML algorithms do not substitute humans but supplement them (“**human-in-the-loop**”)

- ML algorithms can and will learn successfully from the data but the **data should be of an appropriate quality** (representative, unbiased, etc.)

- **No free lunch theorem**: no algorithm outperforms any other algorithm on an infinite number of problems

Summary of the Course

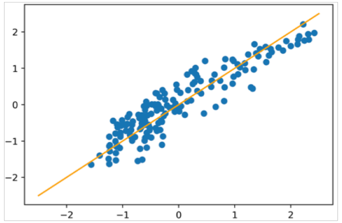
Structuring your DS Project



Machine Learning Overview

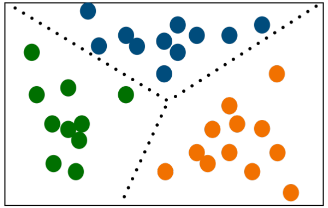
Supervised Learning

- You have access to training data with desired labels
- Learn a function to map observations to labels



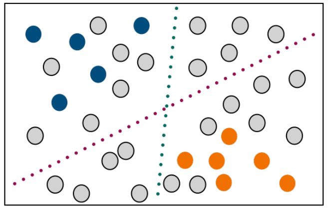
Unsupervised Learning

- Your training data is unlabelled
- Discover structure in data, (ir)regularities, groups of similar instances, etc.



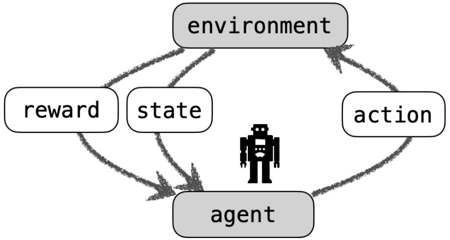
Semi-supervised Learning

- You have a small amount of labelled data and a lot of unlabelled data
- Combine the strengths of both supervised and unsupervised approaches



Reinforcement Learning

- A learning system (*agent*) observes the environment, selects and performs actions, and gets *rewards / penalties* in return
- Learns the best strategy (*policy*) by itself



This course will focus on supervised and unsupervised techniques

What we’ve covered

- We’ve talked about real-life applications of Data Science (**Lecture 1**)
- We’ve discussed and seen in practice how to set up a data science project
- You’ve learned how to pre-process and get insights from data
- You’ve learned about a range of machine learning algorithms
- We’ve looked into regression tasks (**Lecture 2**, **Practical 1**) and classification tasks (**Lecture 3**, **Practical 2**)

What we’ve covered

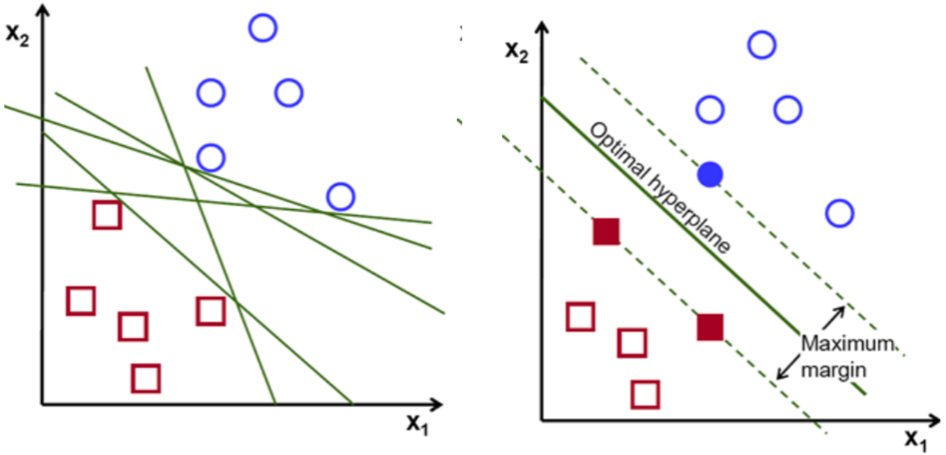
- You’ve learned how to combine multiple algorithms into ensembles (**Lecture 4**, **Practical 3**)
- We’ve talked about the advances in the field brought about by Deep Learning (**Lecture 5**)
- We’ve looked into a number of Deep Learning algorithms (**Lectures 6 & 7**) and you’ve implemented them in practice (**Practicals 4 & 5**)
- We’ve discussed the importance of good visualisation practices and talked about the best strategies when visualising different data scales (**Lecture 8**)

What we’ve covered

- We’ve looked into dimensionality reduction techniques and why they are important (**Lecture 9**)
- You’ve implemented some dimensionality reduction techniques in practice (**Practical 6**)
- We’ve talked about unsupervised and semi-supervised learning, and discussed embeddings
- Finally, we’ve talked about the challenges in Data Science (**Lecture 10**)

What we haven’t covered

- Other ”traditional” ML algorithms – e.g., Support Vector Machines (*“Machine Learning and Bayesian Inference”* course), Gaussian Processes (*“Probabilistic Machine Learning”* course)
- Other DL architectures and techniques
- More in-depth unsupervised learning techniques, semi-supervised learning, transfer learning
- Reinforcement learning
- …



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

Next Steps

Practical Data Science

- Kaggle datasets (https://www.kaggle.com/datasets)
- Data Science competitions (https://www.drivendata.org)
- UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/)
- Registry of Open Data on AWS (https://registry.opendata.aws)
- A Comprehensive List of Open Data Portals from Around the World (http://dataportals.org)
- Financial and economic datasets (https://www.quandl.com)
- Wikipedia's list of Machine Learning datasets (https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)
- Datasets subreddit (https://www.reddit.com/r/datasets/)

Finally, your own data and projects

References

- For **practical skills**:
  - Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow,* and *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*
  - Chollet, F. (2017). *Deep Learning with Python*

References

- **Theoretical Background**:

  - Bishop, C.M. (2008). *Pattern Recognition and Machine Learning*
  - MacKay, D.J. (2003). *Information Theory, Inference and Learning Algorithms*
  - Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*
  - Norvig, P. and Russell, S. J. (2020). *Artificial Intelligence: A Modern Approach*
  - Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*