Data Science: Principles and Practice

Lecture 10: Challenges in Data Science

Ekaterina Kochmar[1]

[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

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

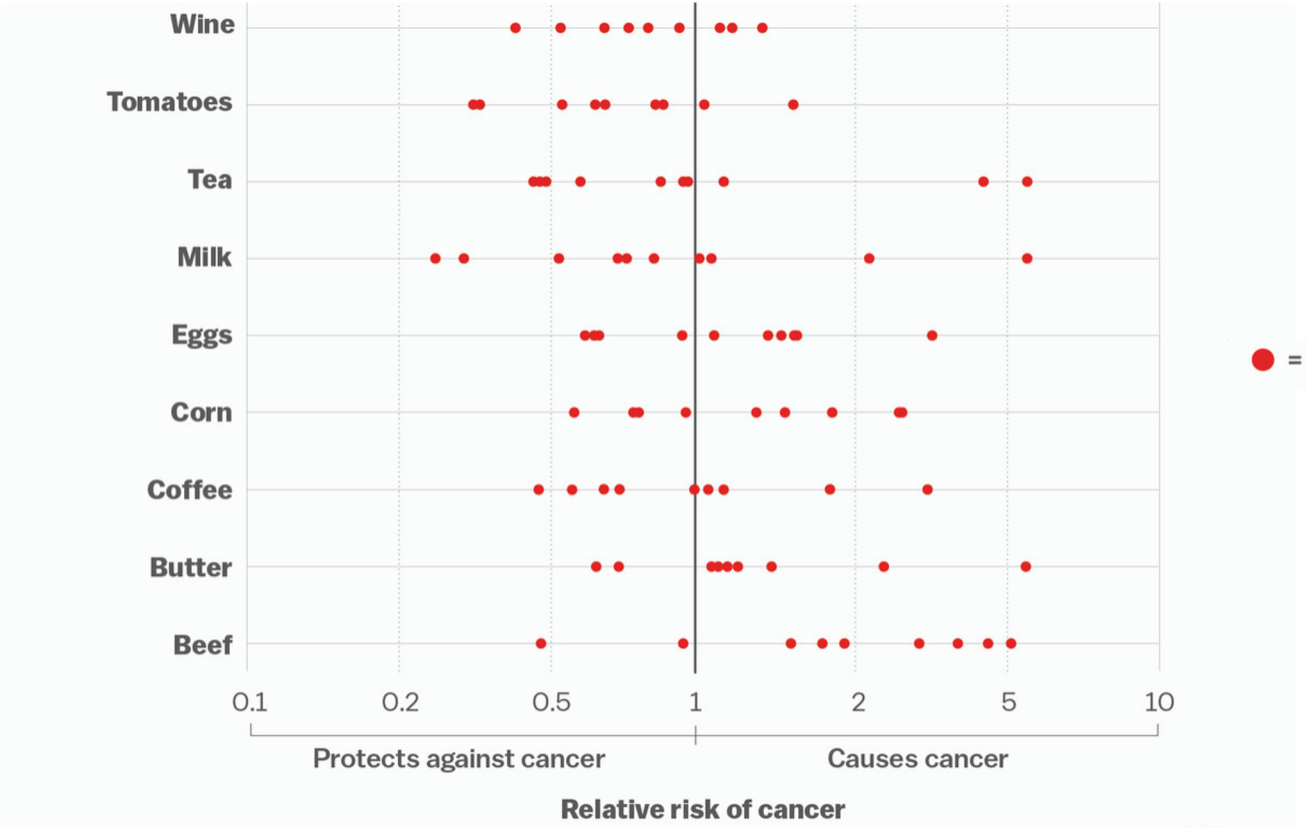
Some of the most high profile findings in social sciences of the past decade do not stand up to replication, a major investigation has found.

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.

> If you torture the data long enough, it will confess to anything.
>
> RONALD COASE

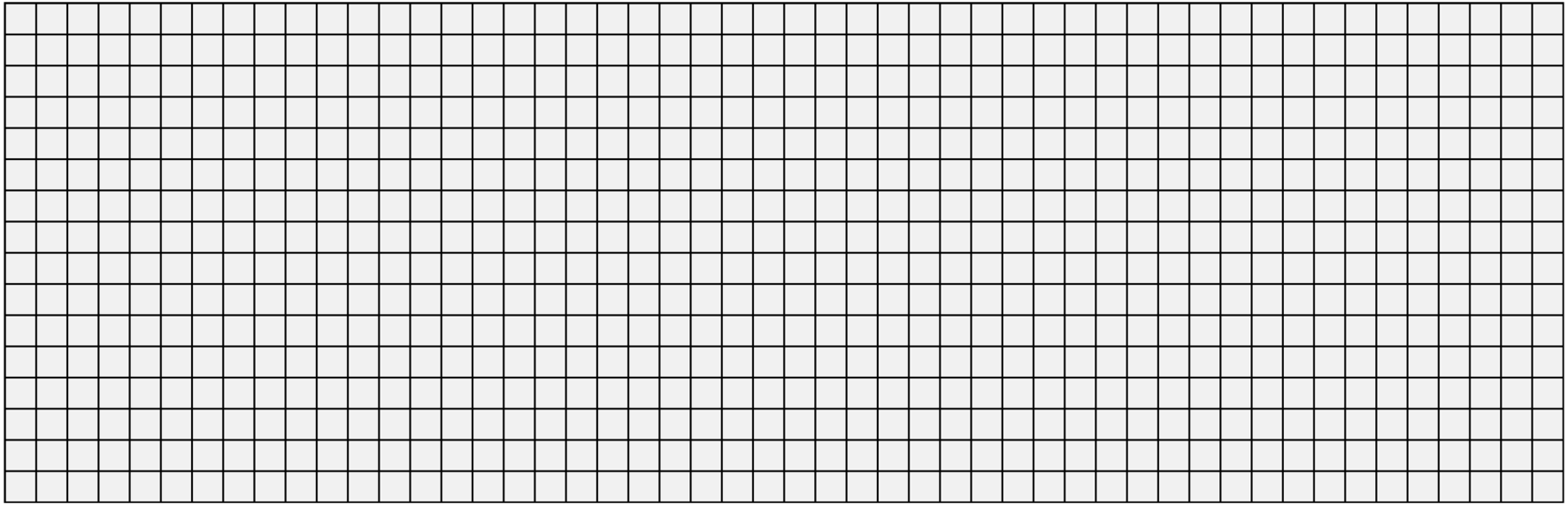
Done by running large numbers of experiments and only paying attention to the ones that come back with significant results.

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 ($p < 0.05$).

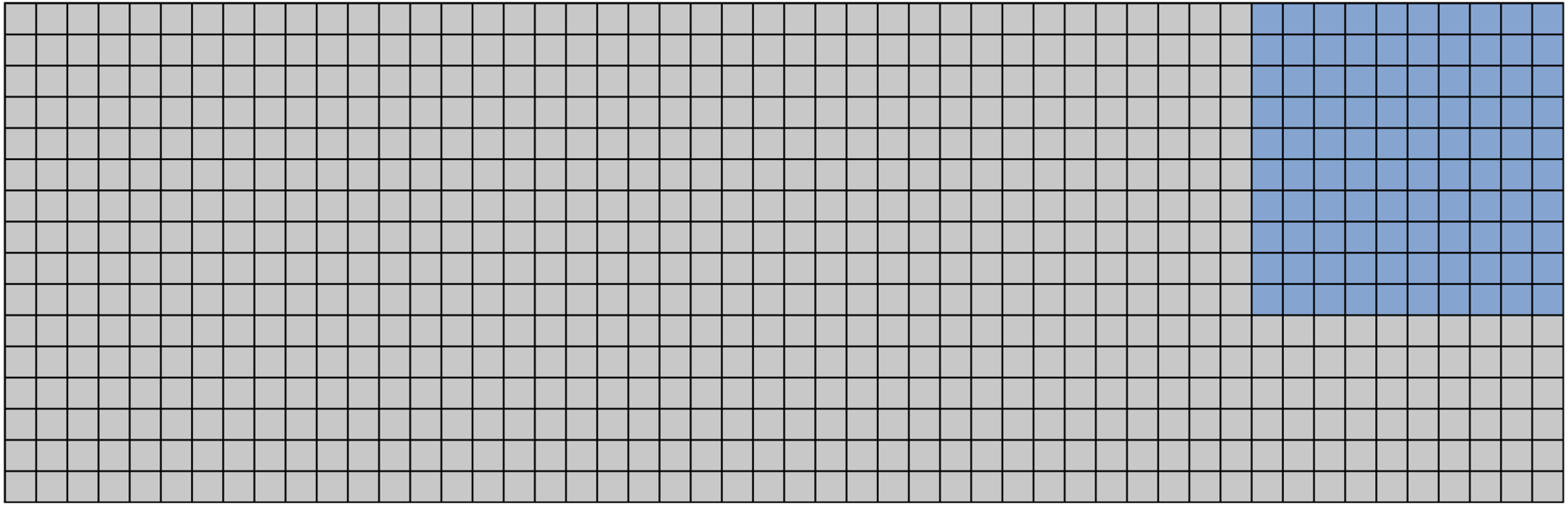
That means we accept that some "significant" results are going to be false positives!

P-hacking



Total 800 hypotheses to test

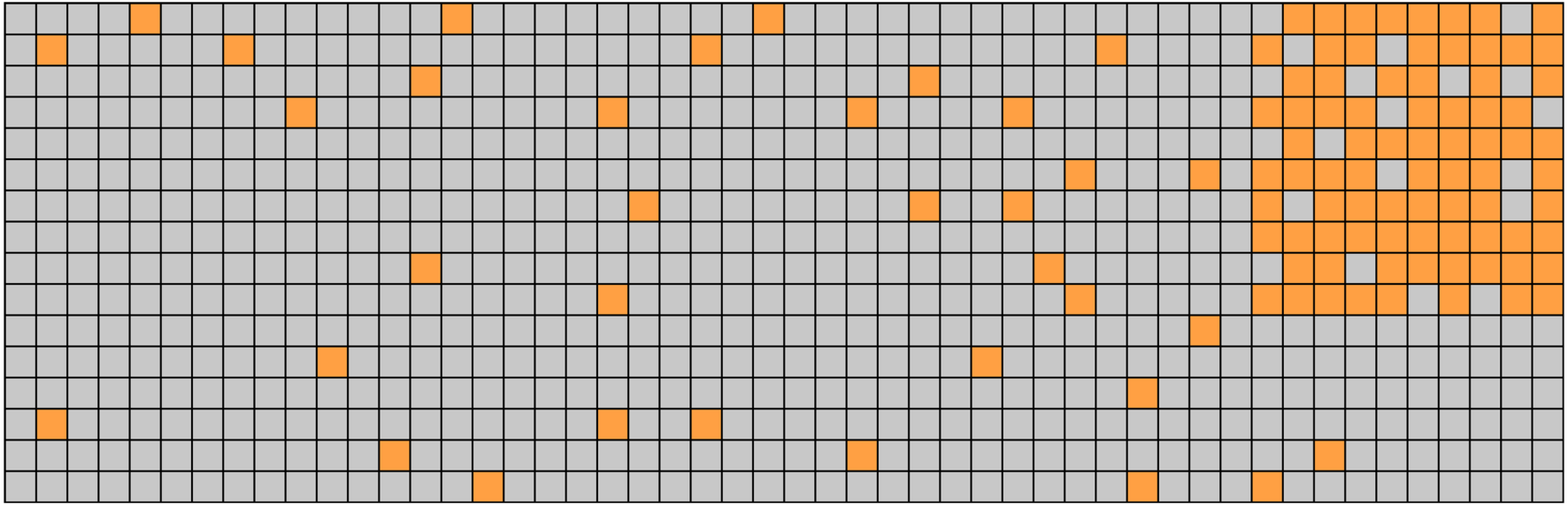
P-hacking



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:

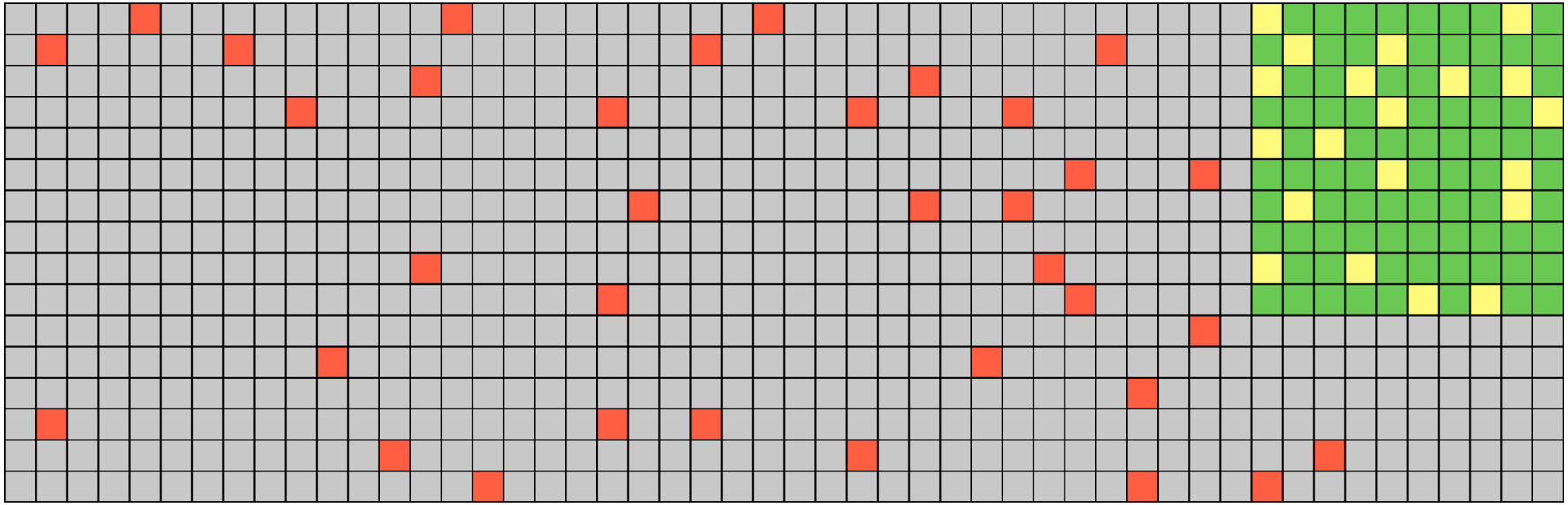
We discover something

We don't discover anything

P(false positive) = 0.05

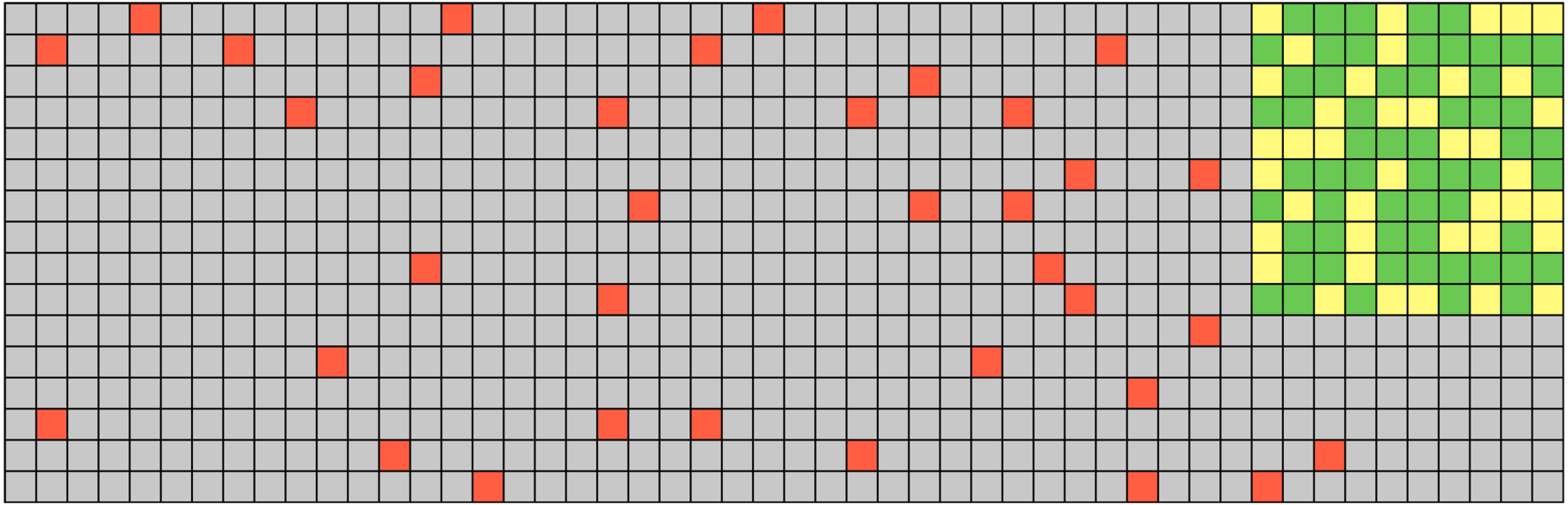
P(false negative) = 0.2

P-hacking



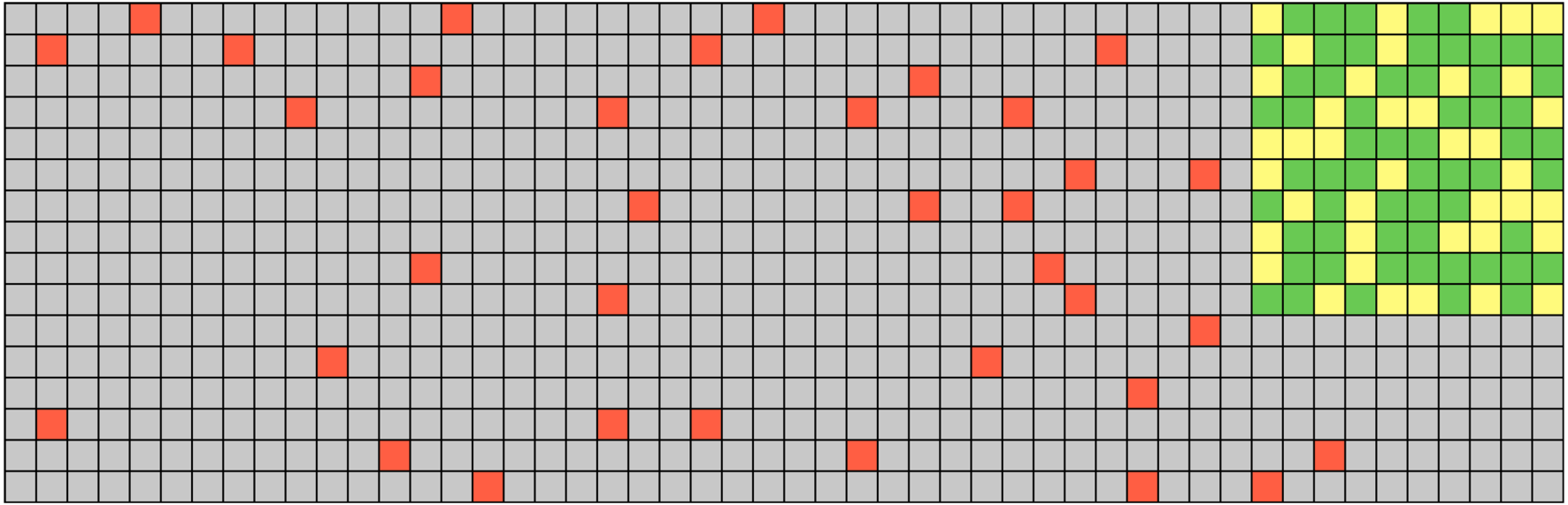
We made 80 true discoveries
We made 35 false discoveries
False Discovery Proportion = 35 / 115 = 0.3

P-hacking



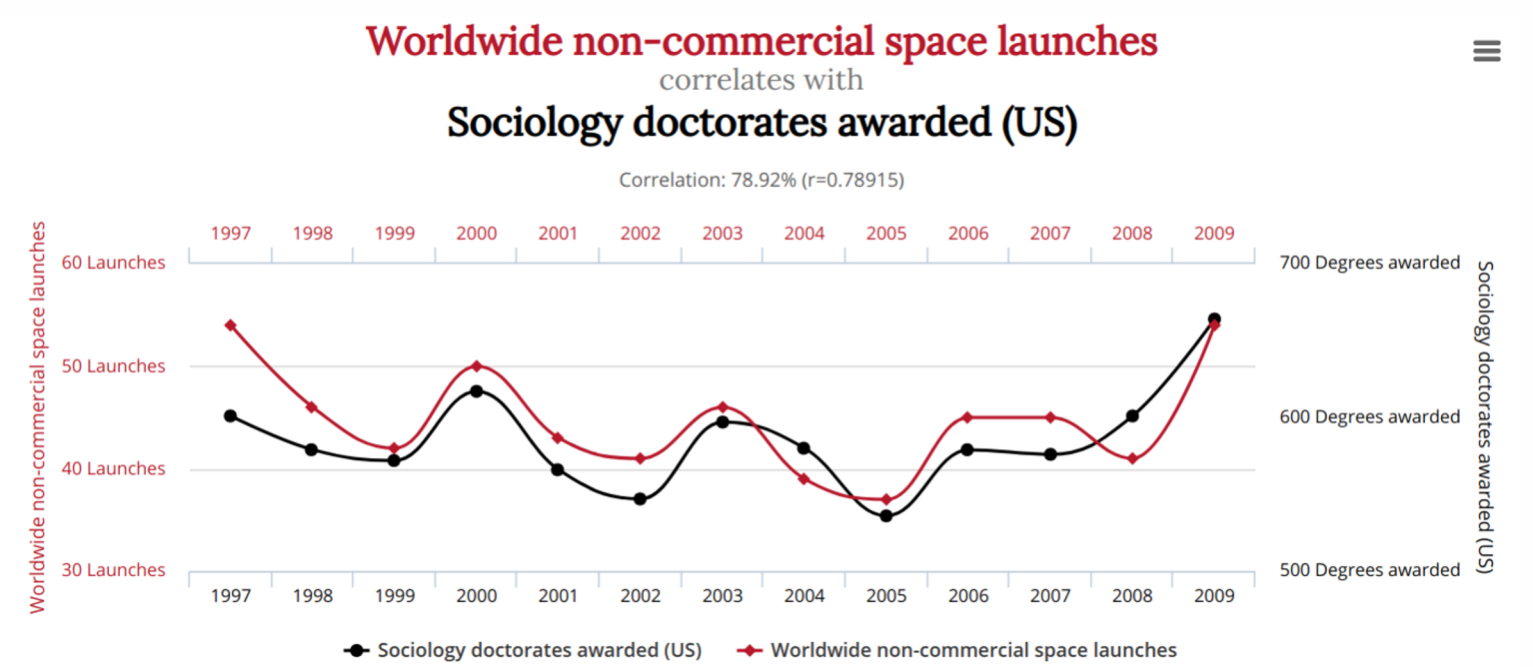
If P(false negative) = 0.4 and P(false positive) = 0.05
We made 60 true discoveries
We made 35 false discoveries
False Discovery Proportion = 35 / 95 = 0.37

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:

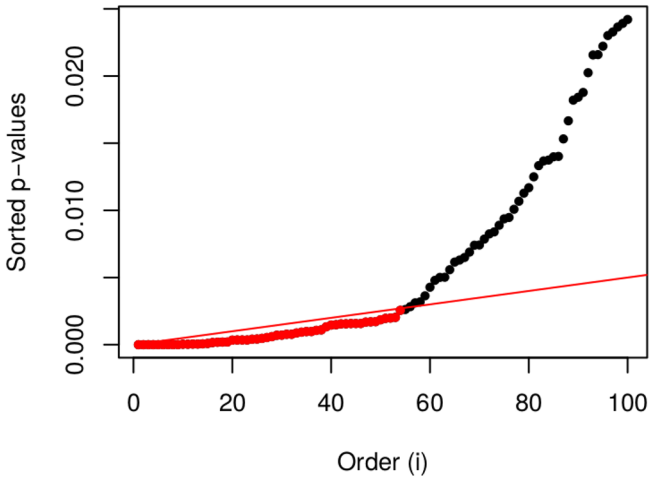
1. Rank *p*-values from M experiments.

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

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

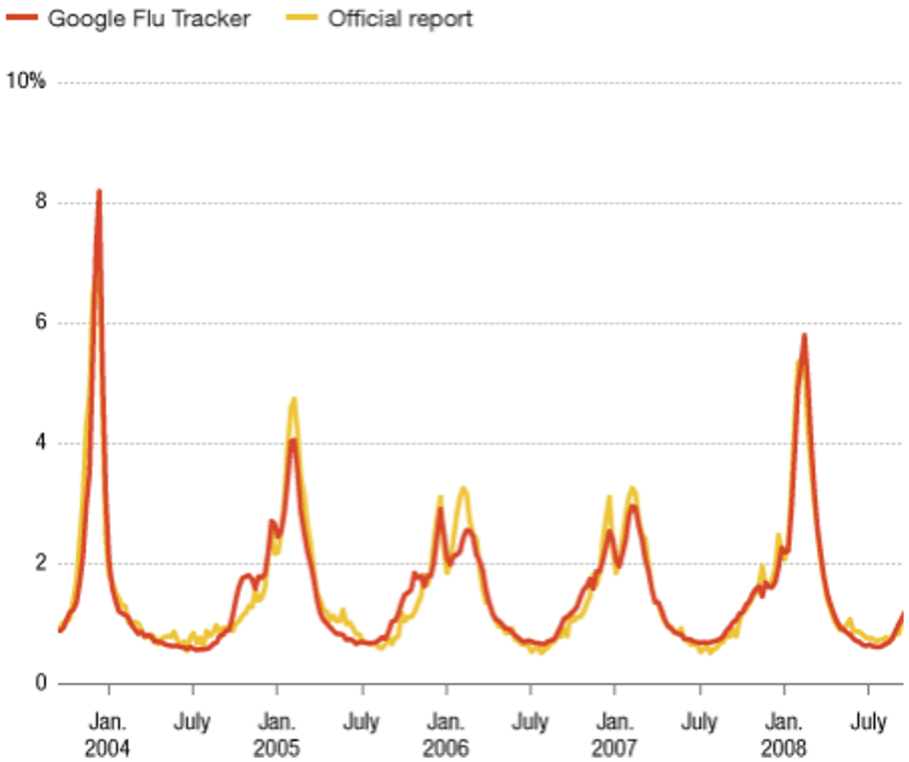
$$z_i = 0.05\frac{i}{M}$$

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



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

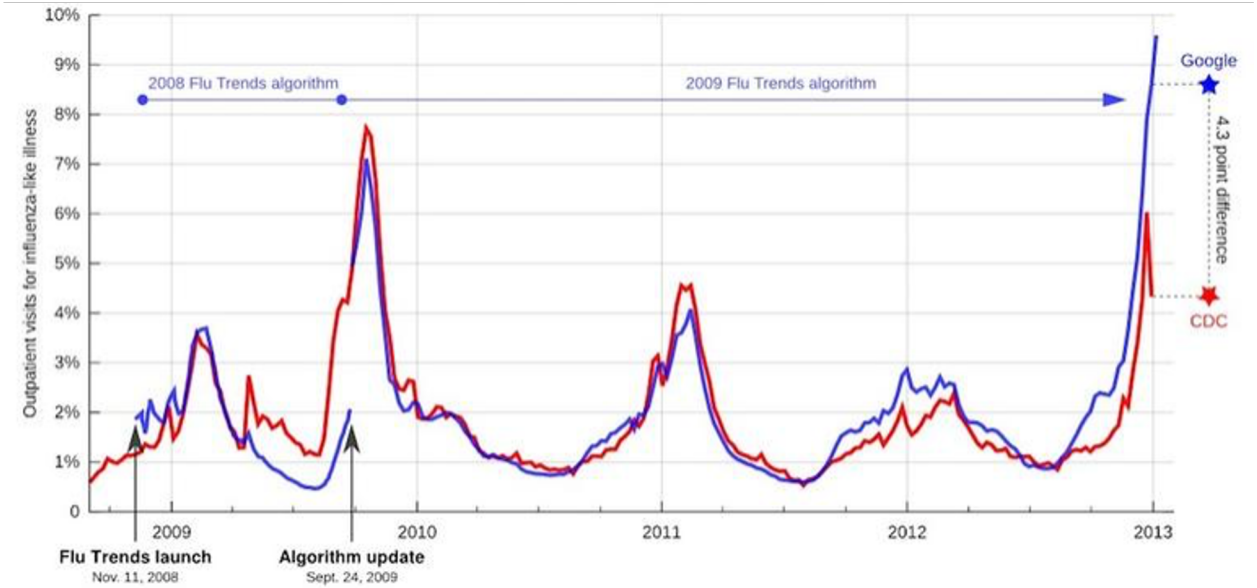
Google Flu Trends



Predicting flu epidemics based on online behaviour

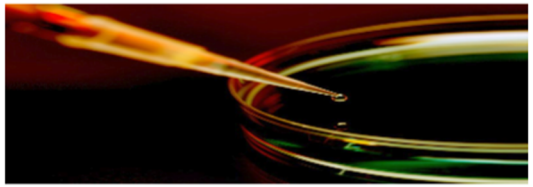
https://www.npr.org/sections/health-shots/2014/03/13/289802934/googles-flu-tracker-suffers-from-sniffles

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.

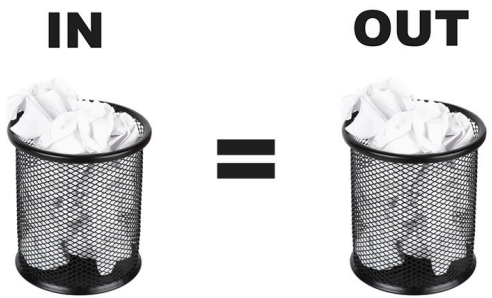
http://www.wbur.org/commonhealth/2013/01/13/google-flu-trends-cdc
https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Summary of Challenges in Data Science

Crucial Components

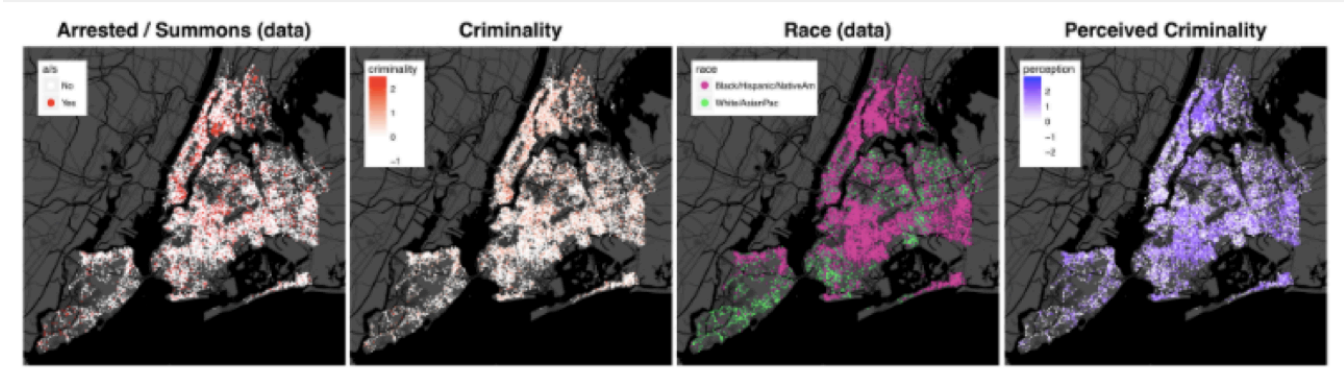
Data:

- 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*

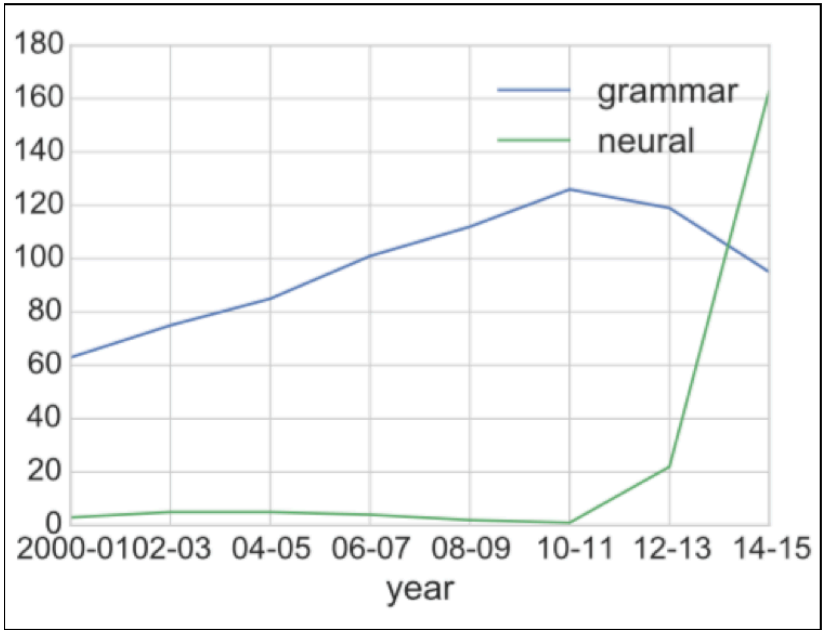
- Problem widely known in machine learning: false diagnosis, false fraud detection, …

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

(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!)



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

(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

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

(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

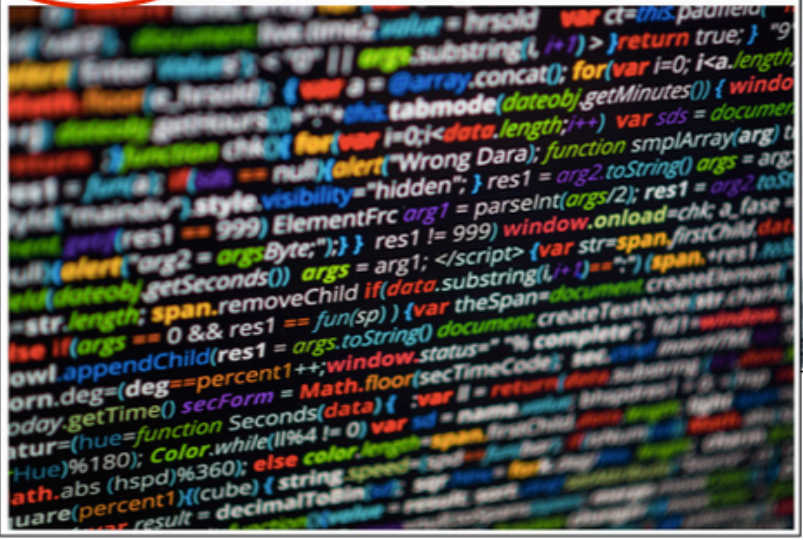
TECH

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

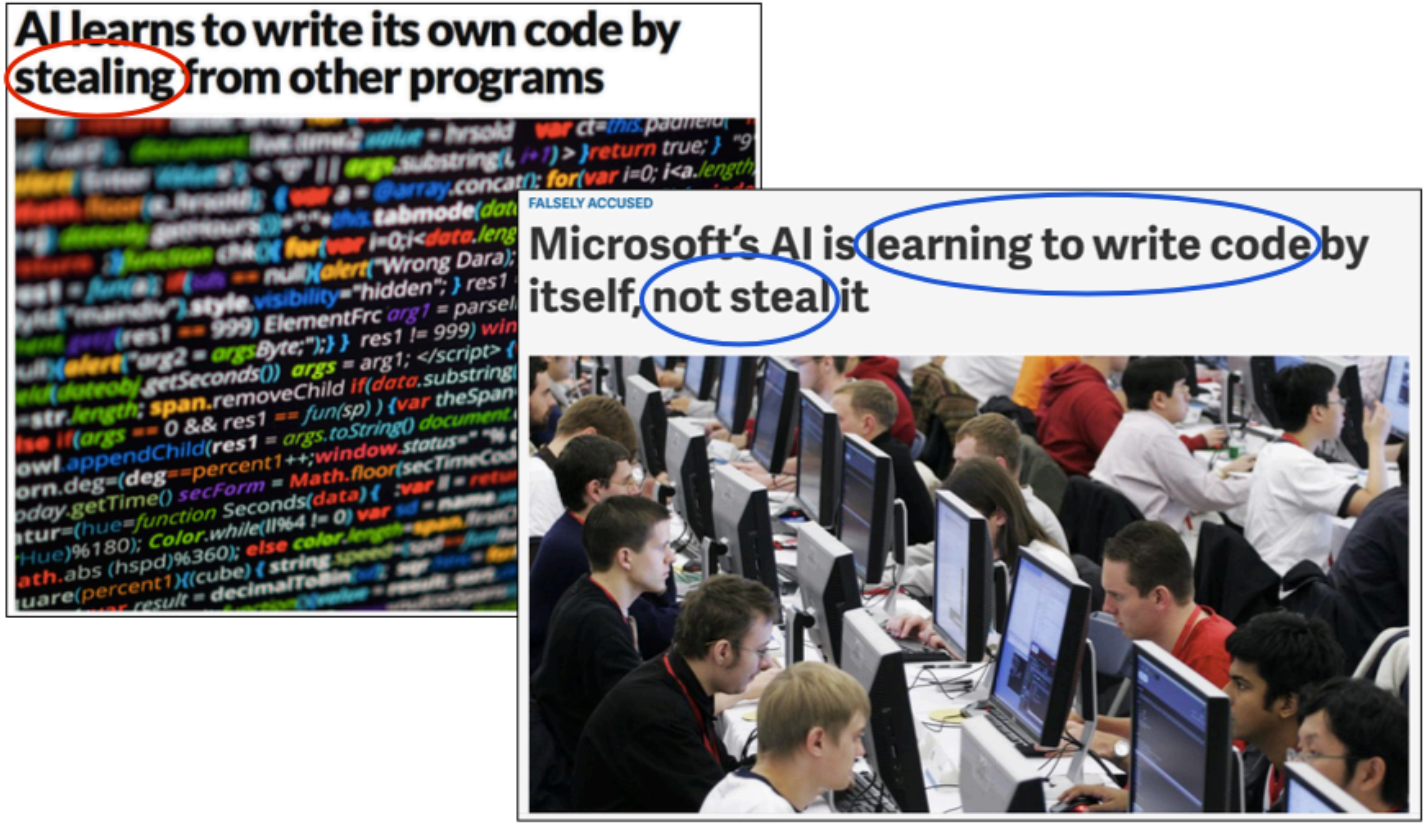
BY DANIEL STARKEY 02.16.2017 :: 3:45PM EDT @DCSTARKEY

3.4K SHARES

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



A Good Example of a Negative Topic Bias

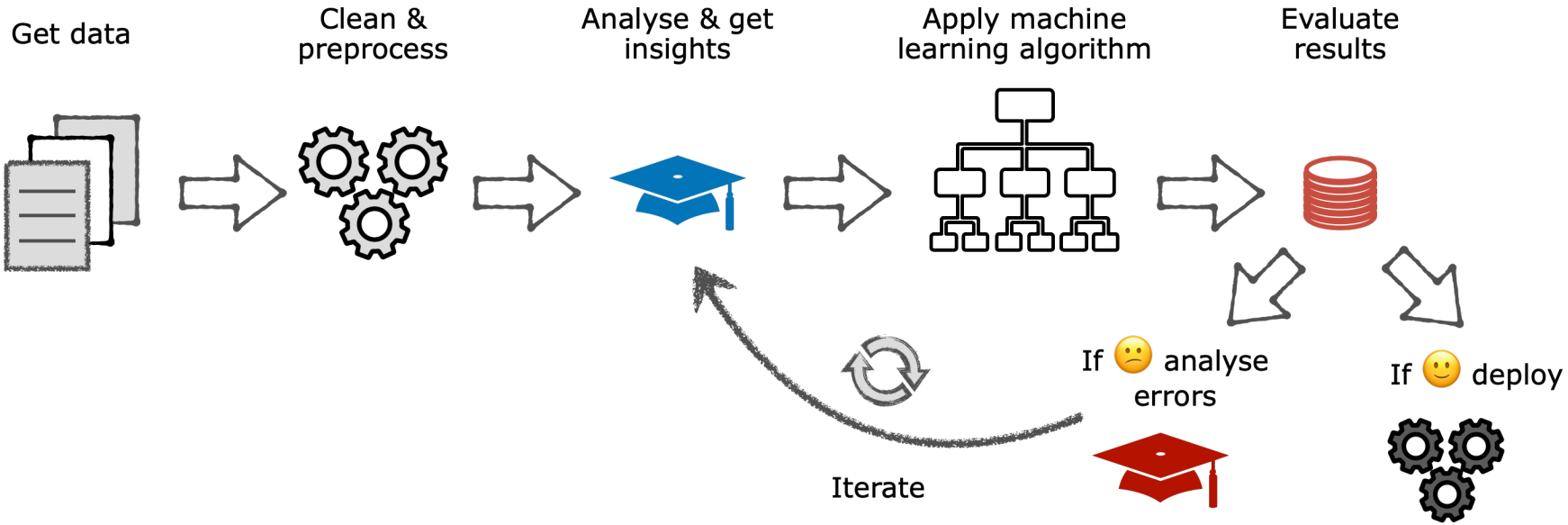


Instead of the Conclusion

- 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

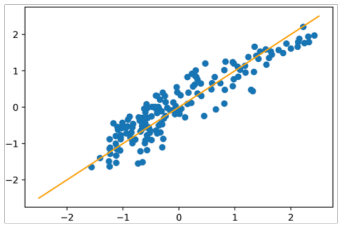
Structuring your DS Project



Machine Learning Overview

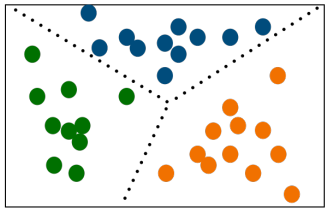
Supervised Learning

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



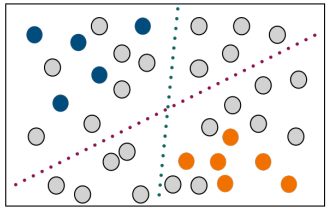
Unsupervised Learning

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



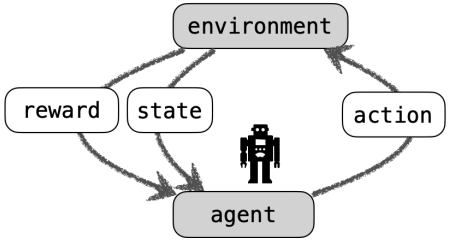
Semi-supervised Learning

- 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

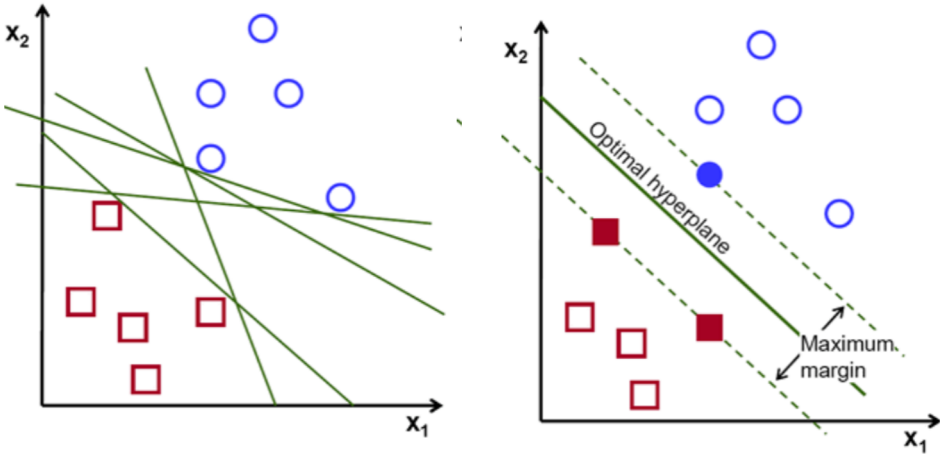
- You've learned how to combine multiple algorithms into ensembles (**Lecture 4**, **Practical 3**)
- We've talked about the advances in the field brought about by Deep Learning (**Lecture 5**)
- We've looked into a number of Deep Learning algorithms (**Lectures 6 & 7**) and you've implemented them in 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*