Applying Machine Learning and AI Techniques to Data

Dr David Tarrant
@davetaz
theODI.org

Our founders

Dr Jeni Tennison
CEO

Sir Nigel Shadbolt
Chairman

Sir Tim Berners-Lee
President

Founded in 2012, the Open Data Institute (ODI) is an international, independent and not-for-profit organisation based in London, UK.

Me

Dr David Tarrant
Learning & Technology

12+ years experience in Open Data

Established first degree level module in Open Data at the University of Southampton

Part of the team that established the ODI

Have since helped transform governments and unlock over $15m for startups

Our mission

We work with companies and governments to build an open, trustworthy data ecosystem.

Our vision

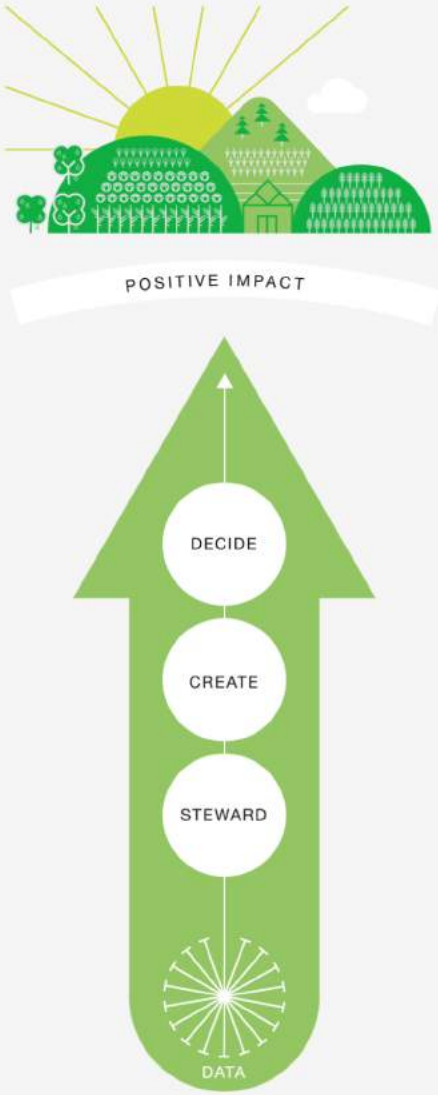
We want people, organisations and communities to use data to make better decisions and be protected from any harmful impacts.

Our theory of change

How value is created from data

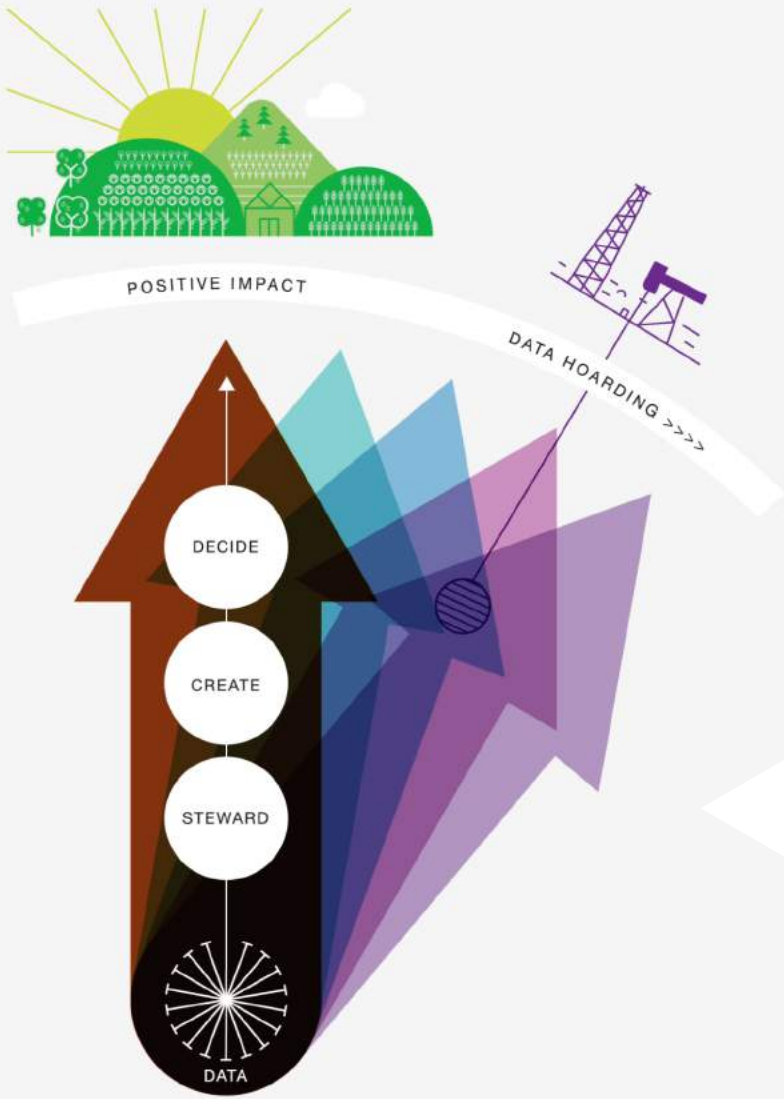
Scenario:

What happens when we hoard data – the oil field



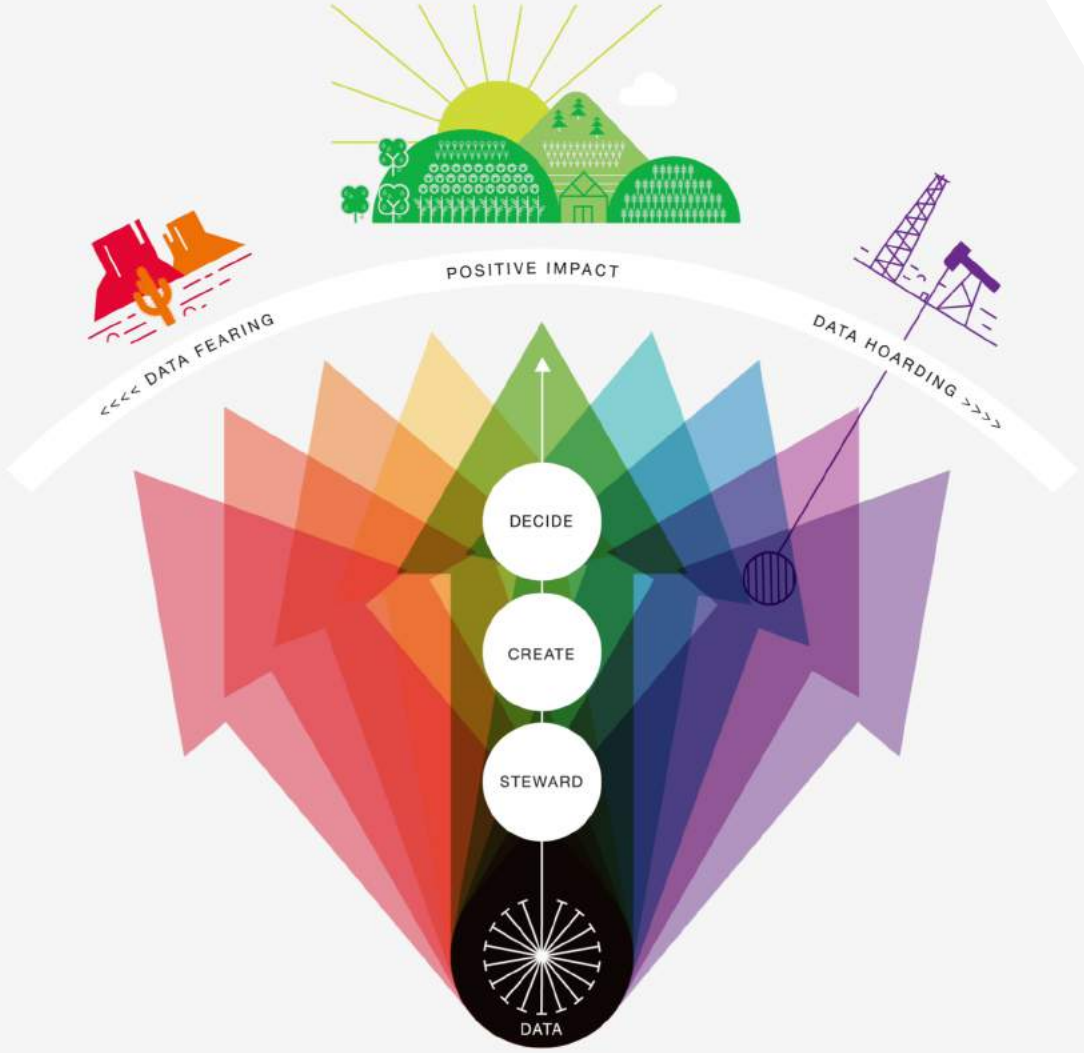
Scenario:

What happens when we fear data – the wasteland



Our theory of change

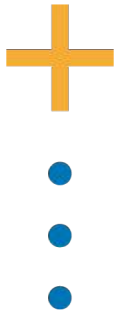
We are one of many organisations working towards a good balance between encouraging and restricting how data is collected and used.

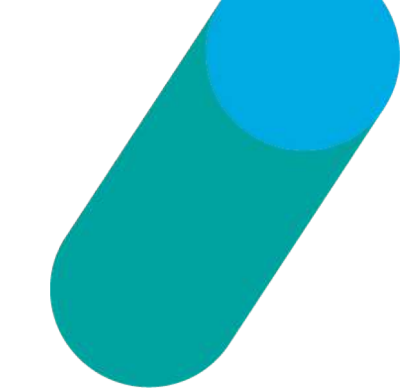


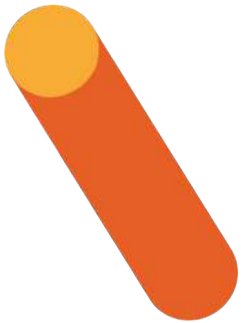
Aim

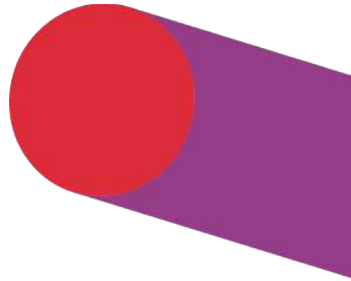
Enable you to apply machine learning and AI techniques to data and discover how ethical frameworks can help you avoid teaching your machines bad habits.

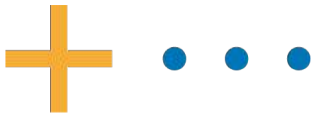
What skills/knowledge do you need to engage effectively with machine learning?

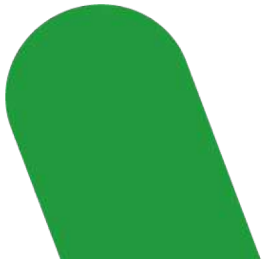












Machine learning is a subset of **artificial intelligence** in the field of **computer science** that often **uses statistical techniques** to give computers the ability to "learn" with data, without being explicitly programmed.

Wikipedia

Machine learning *allows computers to* "learn" with data, without being explicitly *told the answers*.

Dr David Tarrant

Machine learning is a subset of **artificial intelligence** in the field of **computer science** that often **uses statistical techniques** to give computers the ability to "learn" with data, without being explicitly programmed.

Wikipedia

Morning!

What can you tell me about the dataset on your desk?

Work in pairs

open
data
institute

Data, it turns out, has shape.
That shape has meaning.

The shape of data tells you everything you need to know about your data from its obvious features to its deepest secrets.

Averages

You each have a set of cards containing salaries for the players in the red bulls soccer team.

How would you establish the shape of this data?

33,750	33,750	33,750	33,750	44,000
44,000	44,000	45,566	65,000	95,000
103,500	112,495	138,188	141,666	181,500
185,000	190,000	194,375	195,000	205,000
292,500	301,999	4,600,000	5,600,000	

Averages

33,750	33,750	33,750	33,750	44,000
44,000	44,000	45,566	65,000	95,000
103,500	112,495	138,188	141,666	181,500
185,000	190,000	194,375	195,000	205,000
292,500	301,999	4,600,000	5,600,000	

What problems does this dataset have when working out the average?

How might we solve these problems?

Is the middle value (with data points ordered) better than the average?

Outliers

$ 33,750.00	$ 33,750.00
$ 44,000.00	$ 33,750.00
$ 138,188.00	$ 33,750.00
$ 45,566.67	$ 33,750.00
$ 44,000.00	$ 44,000.00
$ 141,666.67	$ 44,000.00
$ 292,500.00	$ 44,000.00
$ 5,600,000.00	$ 44,000.04
$ 103,500.00	$ 45,566.67
$ 190,000.00	$ 65,000.00
$ 65,000.00	$ 95,000.00
$ 33,750.00	$ 103,500.00
$ 195,000.00	$ 112,495.50
$ 44,000.04	$ 138,188.00
$ 4,600,000.00	$ 141,666.67
$ 194,375.00	$ 181,500.00
$ 33,750.00	$ 185,000.00
$ 112,495.50	$ 190,000.00
$ 95,000.00	$ 194,375.00
$ 301,999.00	$ 195,000.00
$ 181,500.00	$ 205,000.00
$ 33,750.00	$ 292,500.00
$ 185,000.00	$ 301,999.00
$ 205,000.00	$ 4,600,000.00
$ 44,000.00	$ 5,600,000.00

Average (mean): $518,311.64

Median: ~$125,000

16% of 25 = 4

16% trimmed mean: $128,109.09

=trimmean(RANGE,0.16)

Averages

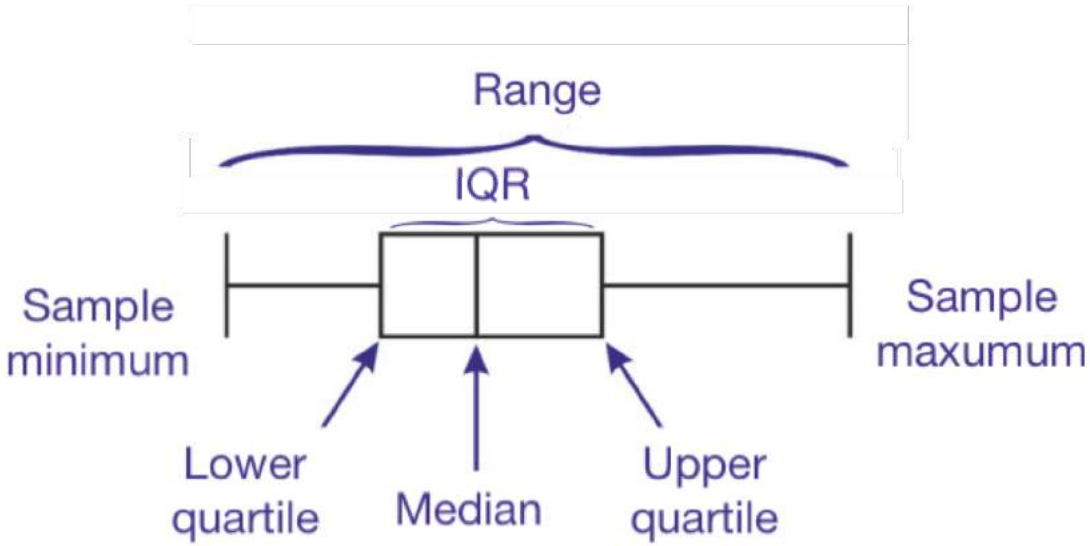
33,750	33,750	33,750	33,750	44,000
44,000	44,000	45,566	65,000	95,000
103,500	112,495	138,188	141,666	181,500
185,000	190,000	194,375	195,000	205,000
292,500	301,999	4,600,000	5,600,000	

Ensure your cards are in order.

- Find the value ¼ of the way through
- Find the value ¾ of the way through

The five-number summary

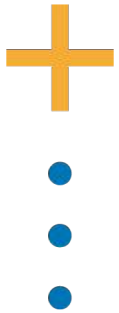
1. the sample minimum (smallest value)
2. the lower quartile (value ¼ through the list)
3. the median (middle value in the list)
4. the upper quartile (value ¾ through the list)
5. the sample maximum (largest value)

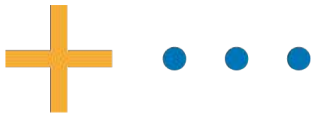


Exercise

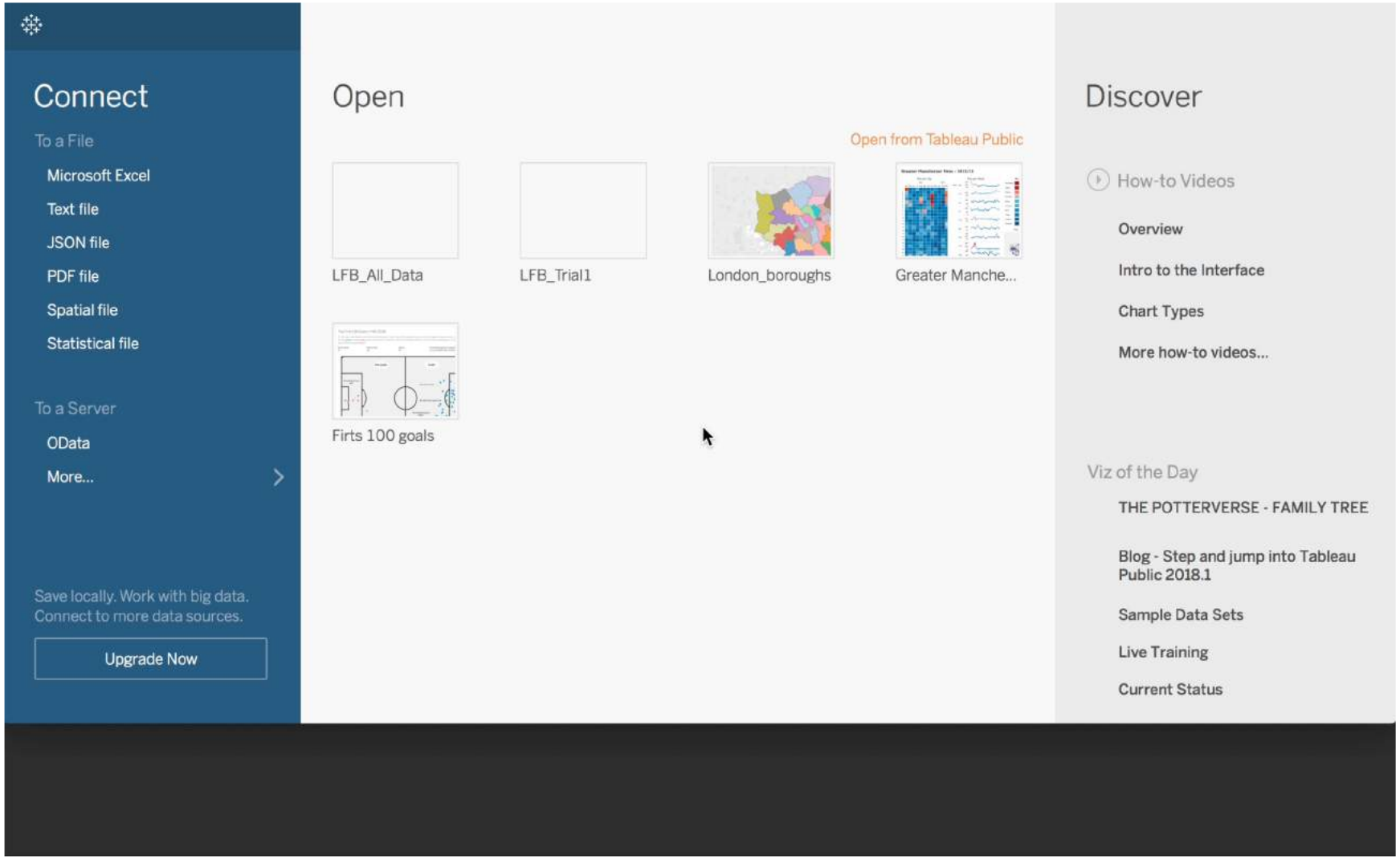
The five-number summary in Tableau

http://bit.ly/datashape





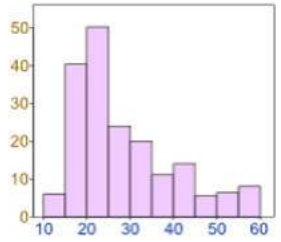
Five-number summary in Tableau



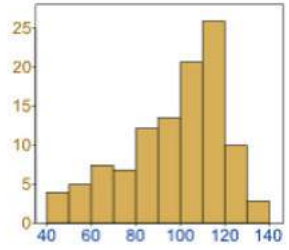
The distribution of data

Another important aspect to consider is the distribution of data. It is always a good practice to know the distribution of your data before analysing it further. Certain analyses require certain distributions.

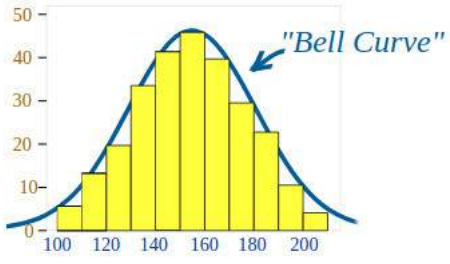
The examples here show different types of distributions of data.



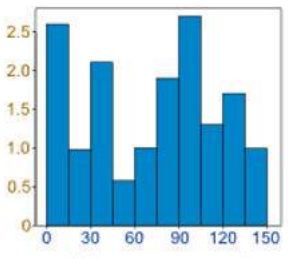
Positively skewed distribution



Negatively skewed distribution



Normal distribution

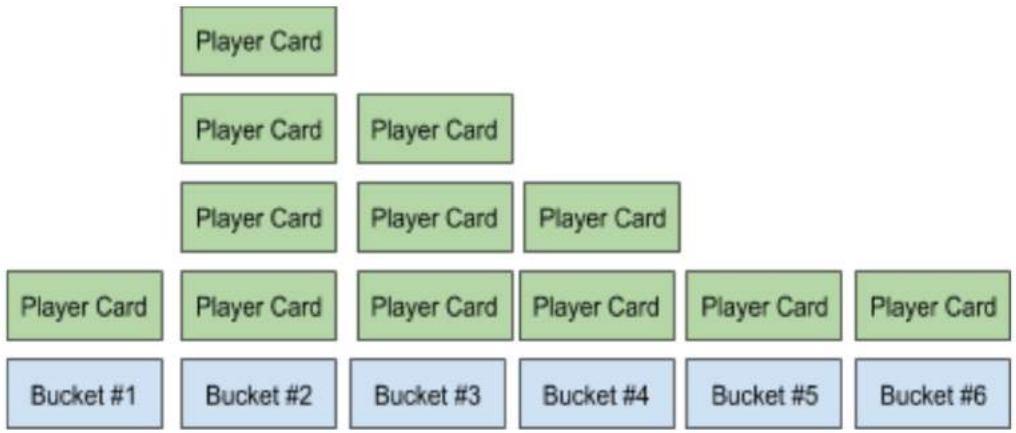


Non-normal distribution

The distribution of data

Take a look at the salaries of the New York Red Bulls.

Divide the data into 6 evenly distributed buckets of salaries and stack the cards as per the diagram on the right.



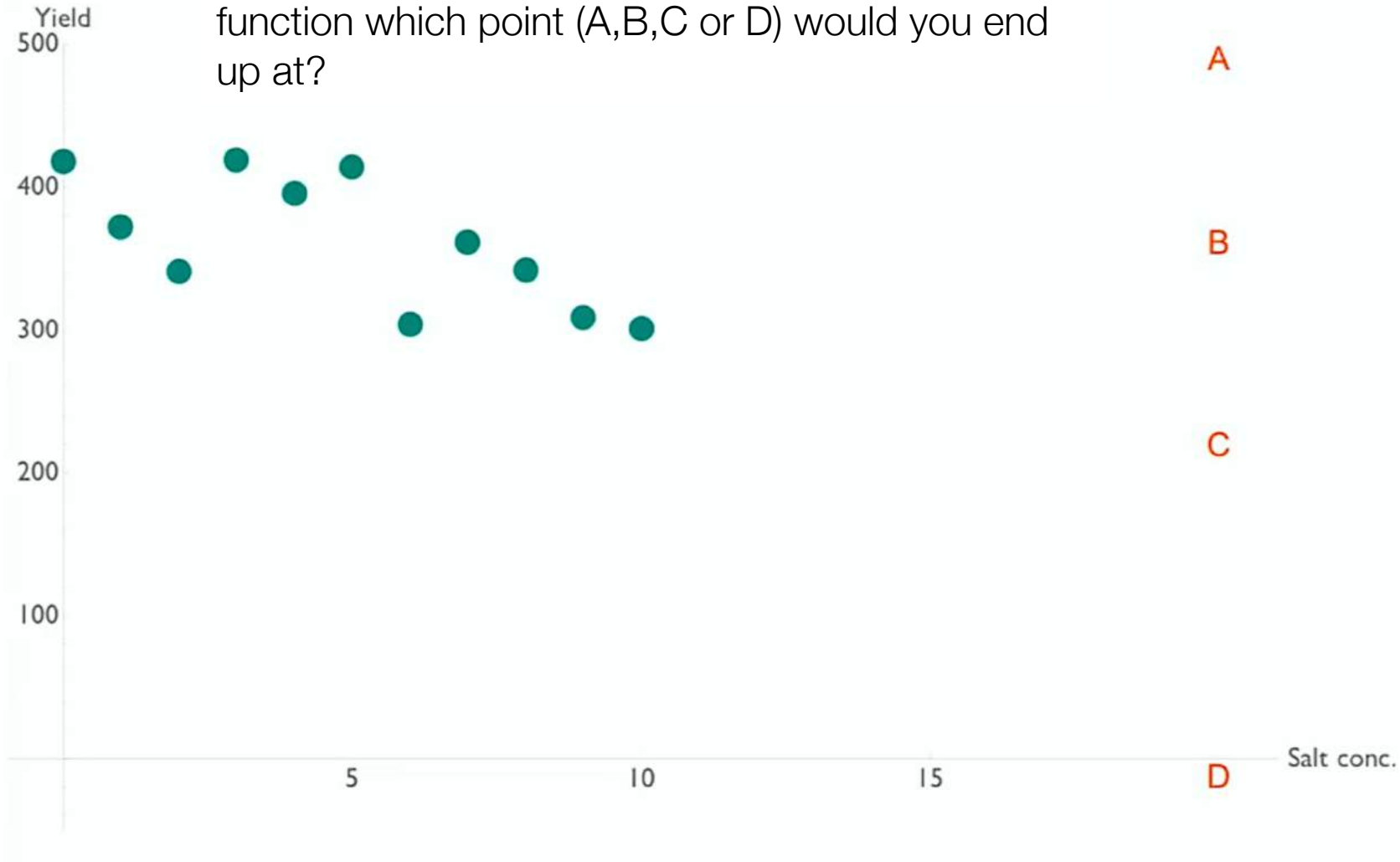
Trends in data

One of the key aspects of statistical analysis is that of spotting trends. Trends give an indication of the direction in which something is changing and can also help us predict what is going to happen next. However when trying to predict or classify we need to be confident our prediction is correct. This section looks at some of techniques that can be used to increase you confidence in making predictions and classifications from data.

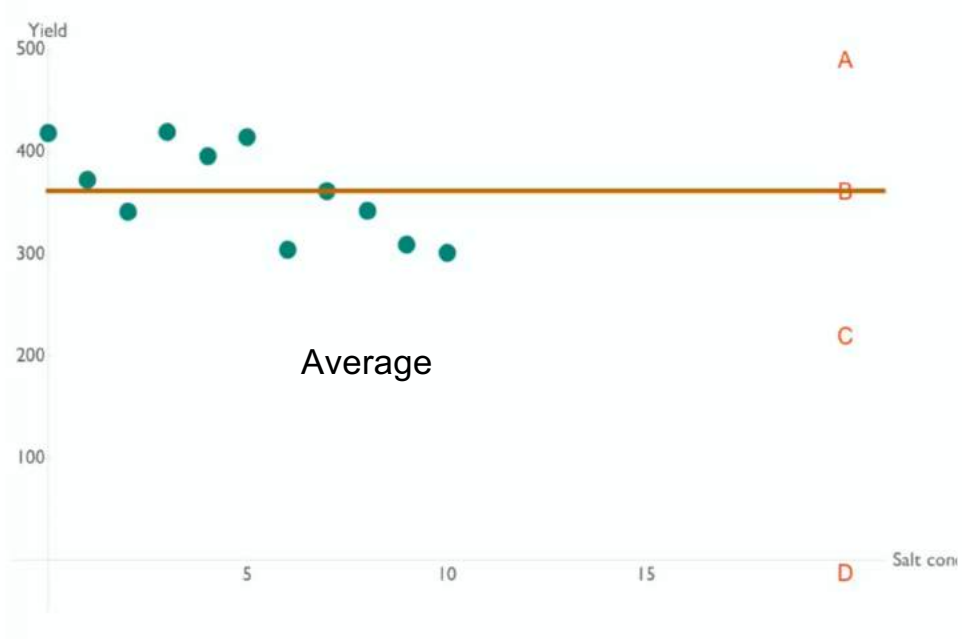
Overfitting

Plotted here are 10 points in a series that have been generated using a function.

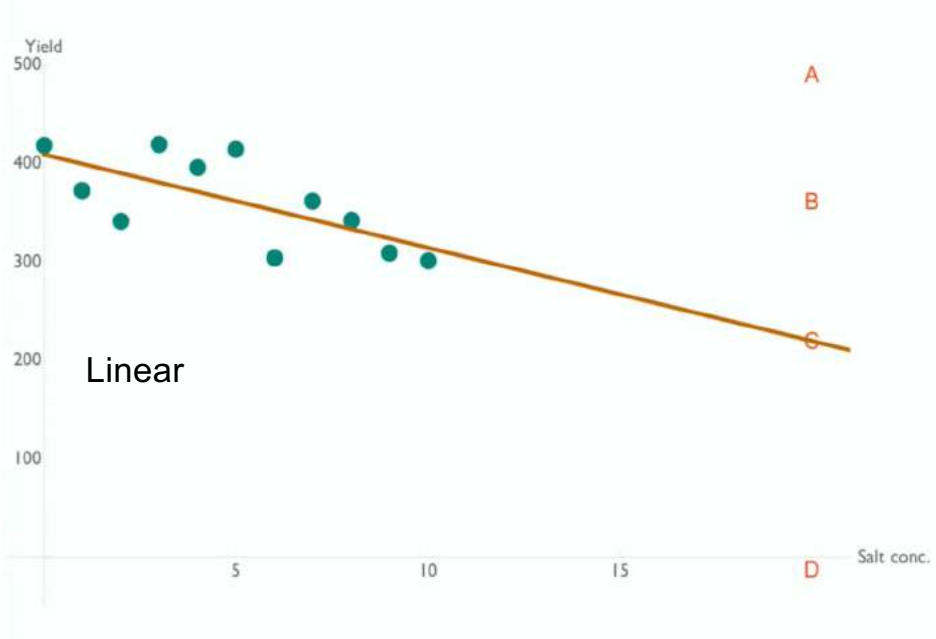
If another 10 points were plotted of the same function which point (A,B,C or D) would you end up at?



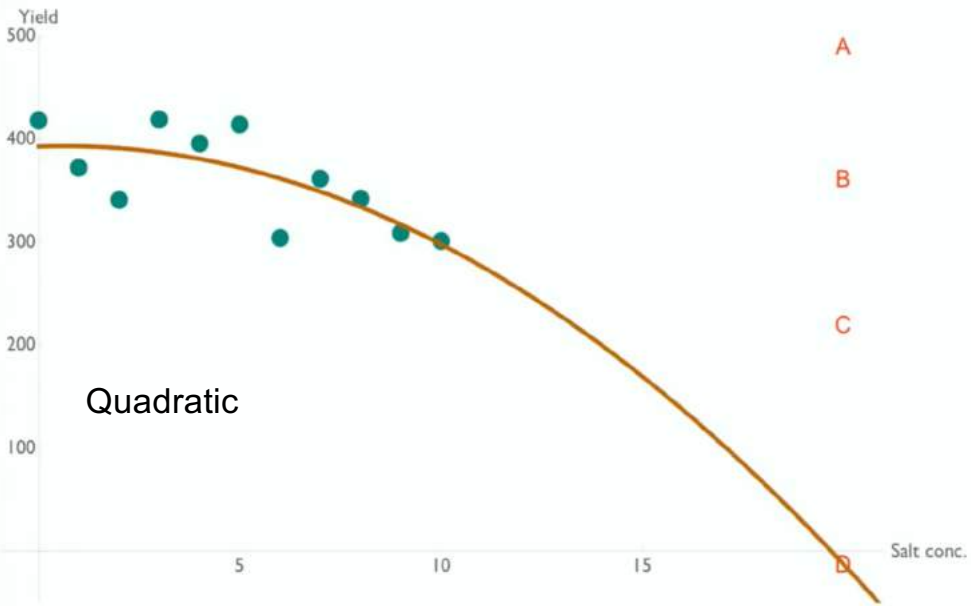
Overfitting



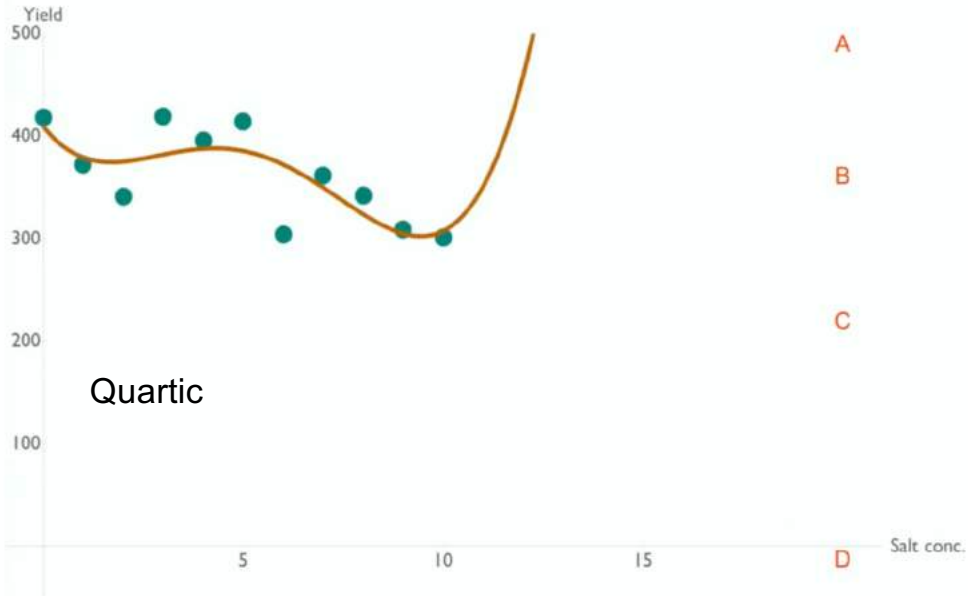
Average



Linear

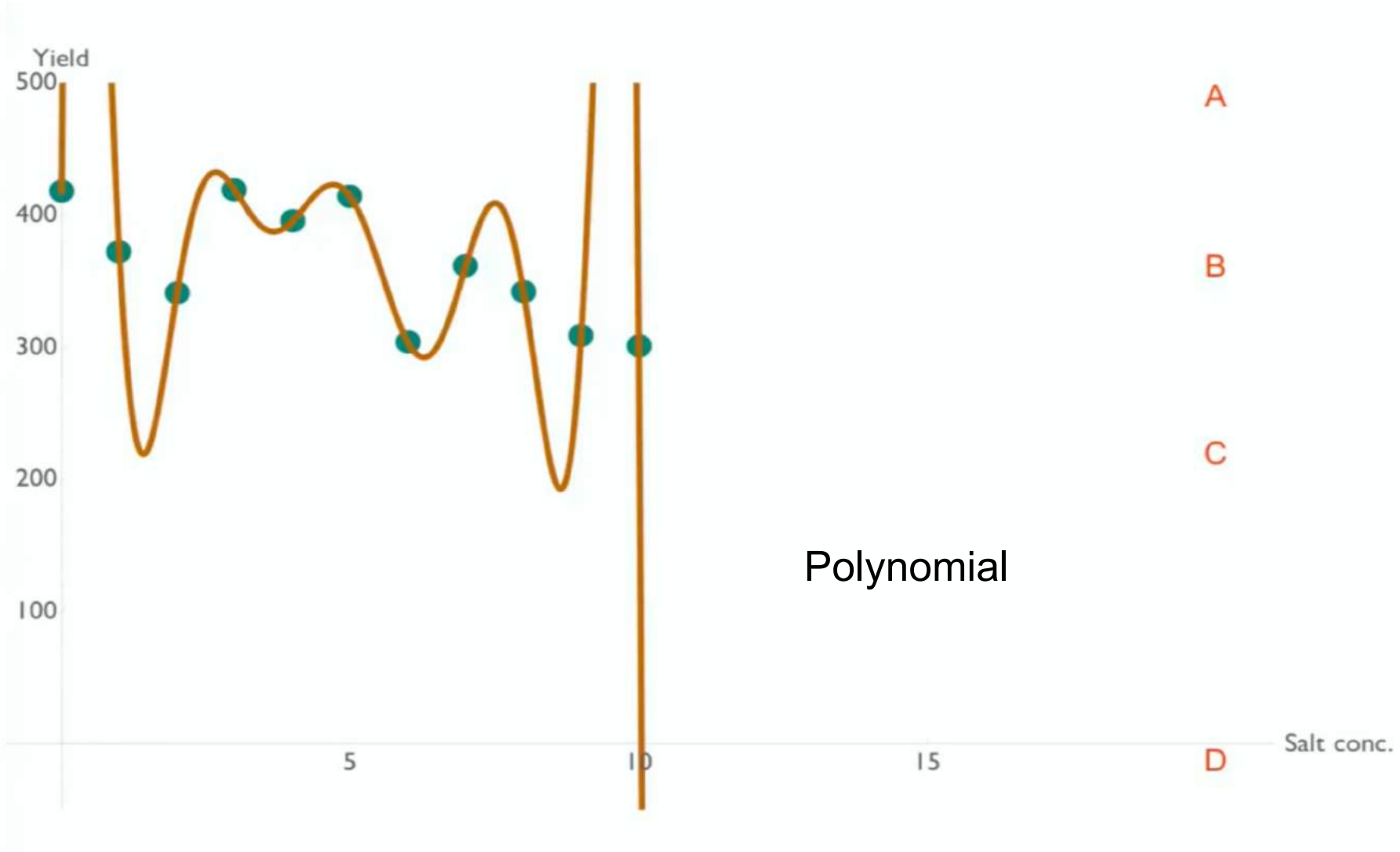


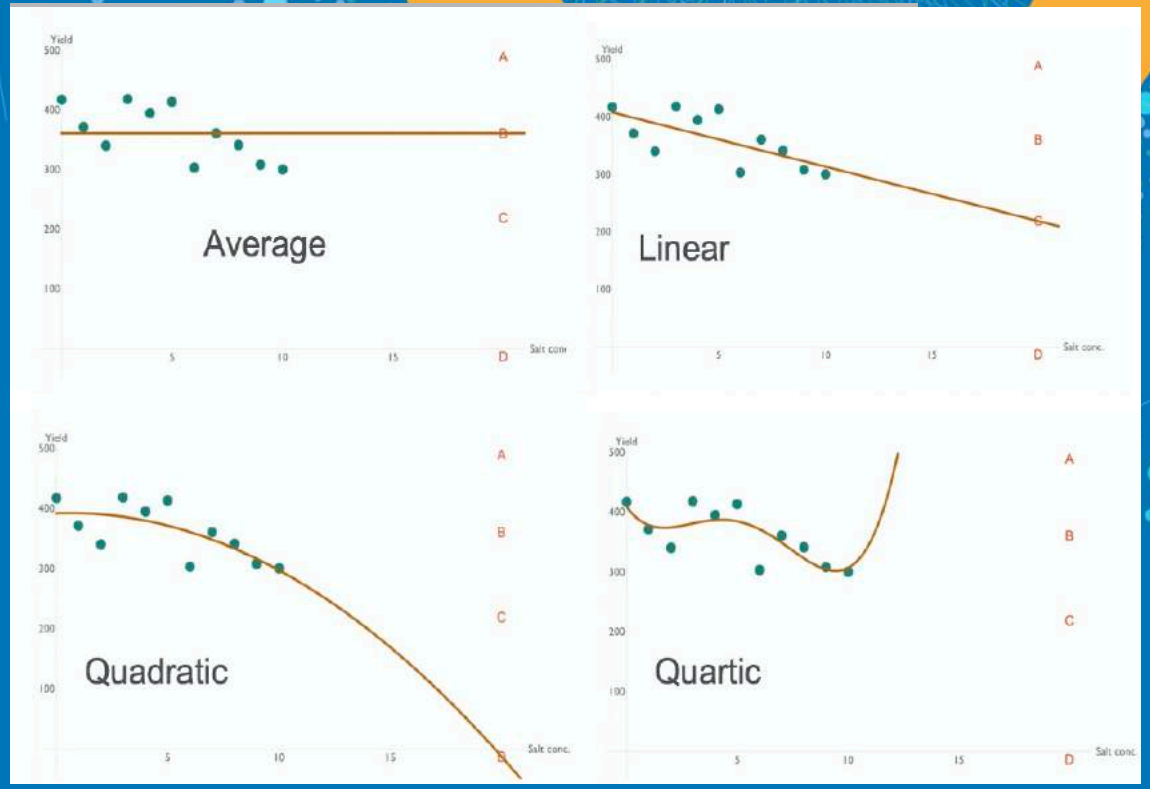
Quadratic



Quartic

Overfitting

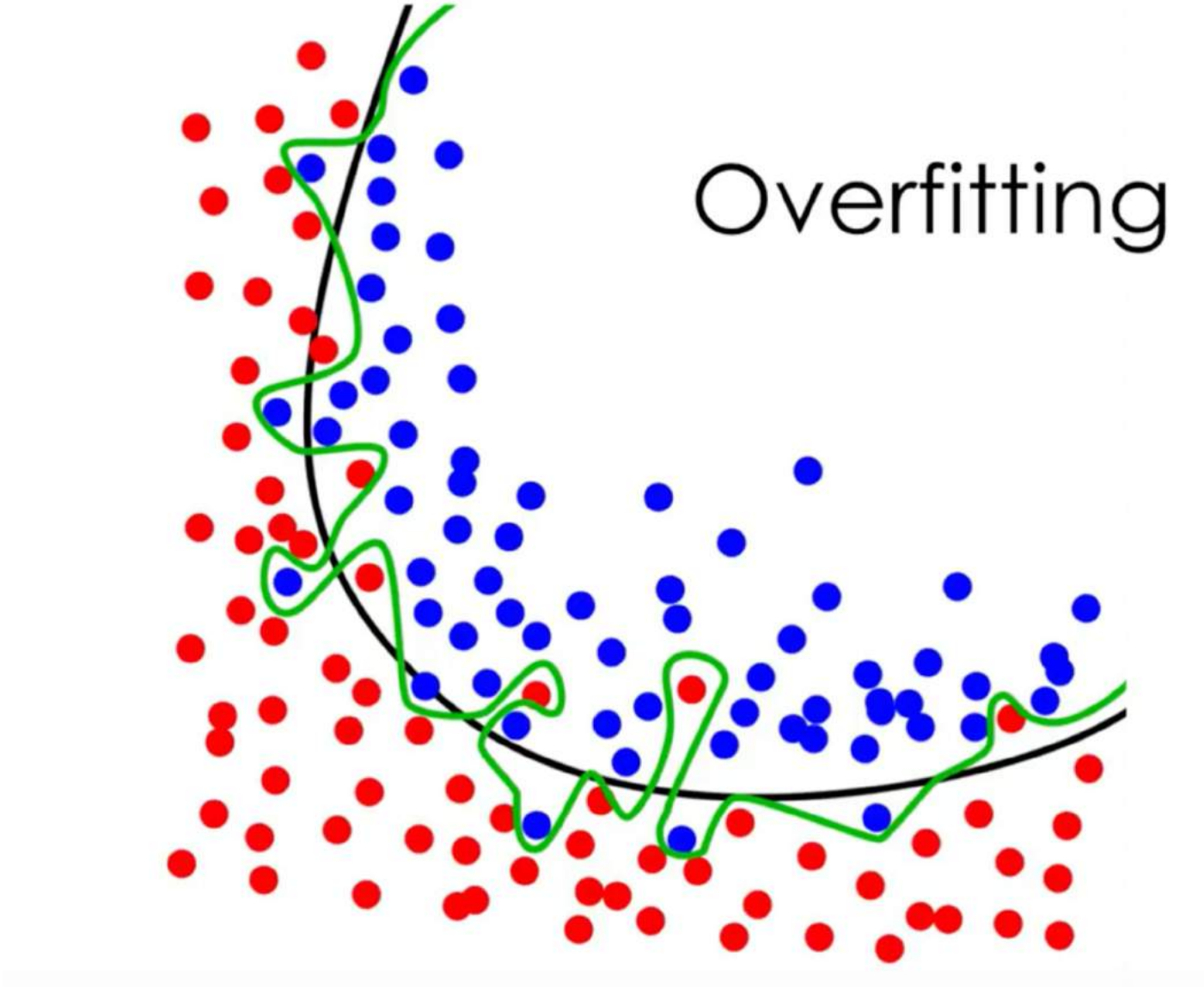




Regression

Used to establish the relationship between two independent variables by fitting a line of best fit.

Overfitting



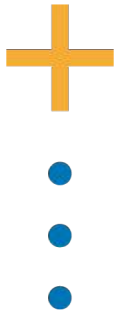

New Glarus
Brewing Co.

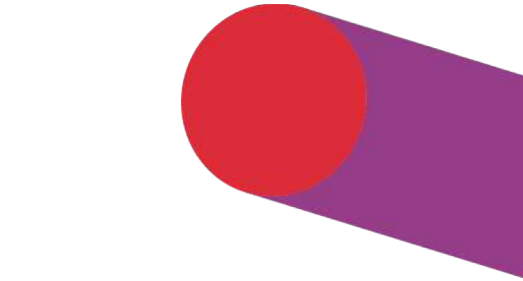
Contents:

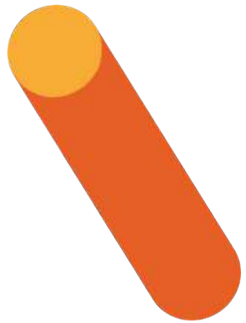
Exercise

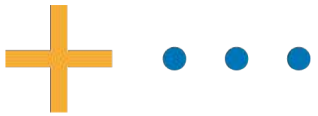
What is big data to you?

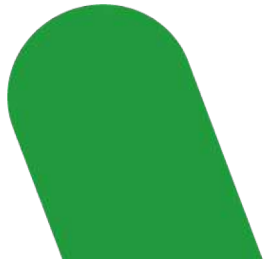
What are the key characteristics that make something big data?











Big data is changing the world

Whenever you work with big data you must eliminate the effect of any of these aspects on your result.

They are like outliers.

Big data is changing the world

nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

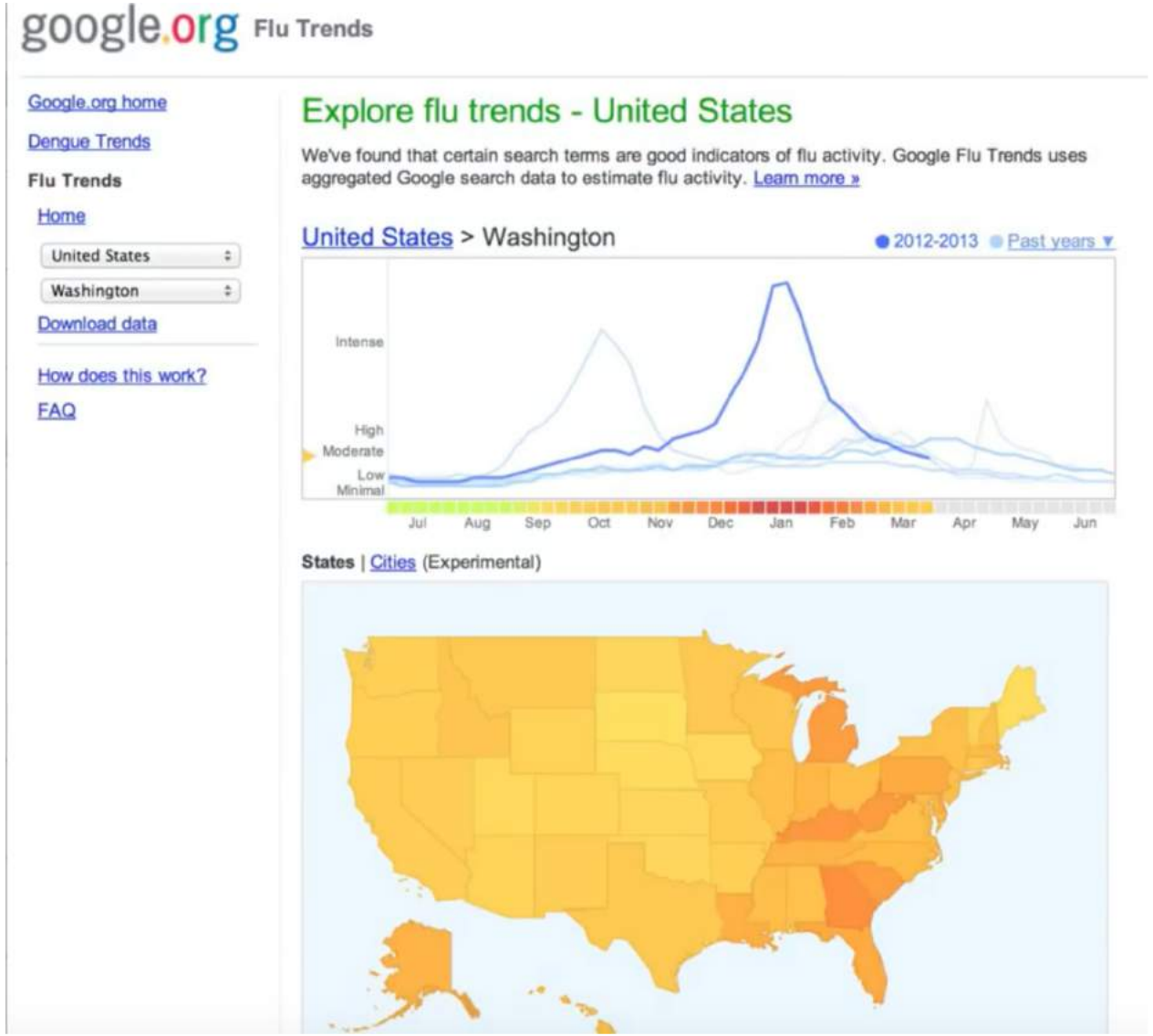
Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human trans-

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each

Confidence

97%

Google's claimed accuracy when compared to Centers for Disease Control data.

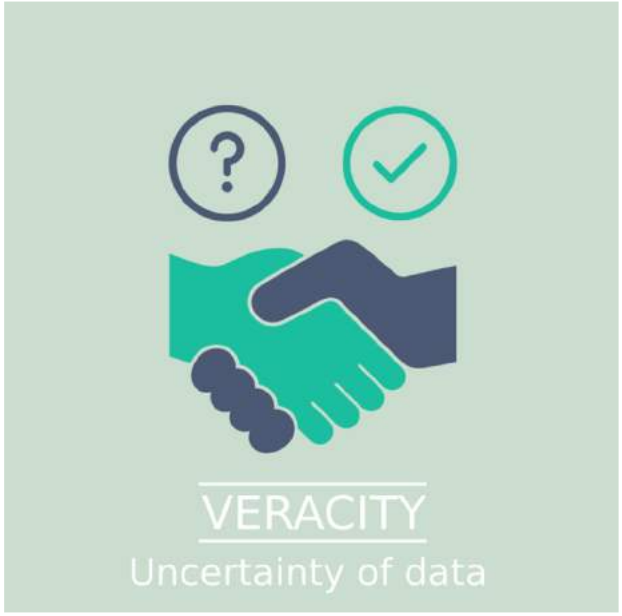


Problems?


VARIETY
More data sources
VOLUME
More data collected
BIG DATA
VERACITY
Uncertainty of data
VELOCITY
Faster than ever


Google
fever aching joints headache
All
Shopping
Images
News
Videos
More
Settings
Tools
About 369,000 results (0.52 seconds)
WebMD Symptom Checker helps you find the most common medical conditions indicated by the symptoms chills, fever, headache and joint aches including Lyme disease, Acute sinusitis, and Aseptic meningitis (adult). ... Osteoarthritis happens when the cartilage in your joints breaks down causing pain, stiffness, and swelling.
Chills, Fever, Headache and Joint aches: Common Related Medical ...
symptomchecker.webmd.com/multiple-symptoms?...chills%7Cfever%7Cheadache%7Cjo...
About this result • Feedback

Problems?



People making flu-related Google searches may know very little about how to diagnose flu?

Does a search for flu mean they have flu or just an interest in it?

Why did google flu trends fail?

- The data was not reliable (counting problem)
- Used nth degree polynomial (overfitting)
- Solely relied on this data over scientific method

For more on big data search YouTube for “calling bullshit on big data”

Excellent lecture series from the University of Washington

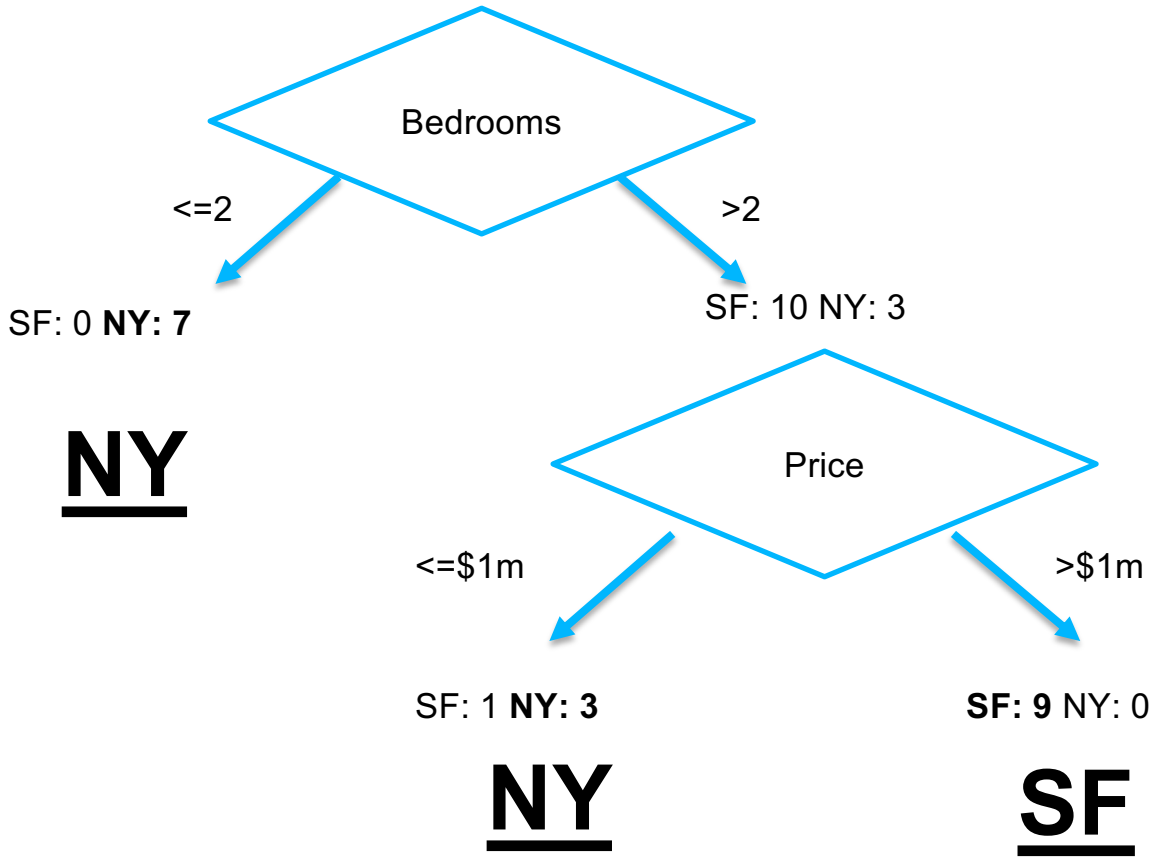
Each table has a set of “Top Trump” cards relating to properties in two cities.

Build a decision tree to sort them into “New York” and “San Francisco”.

You cannot use the name of the city to sort them.

Example decision tree

95%
Confidence



Correct: 19
Wrong: 1

Checking your result

Use your decision tree to classify the red cards into two piles (NY and SF)

Also record your prediction for each red card in a table →

Card No.	Prediction
#23	SF
#59	NY
#233	SF

Decision trees

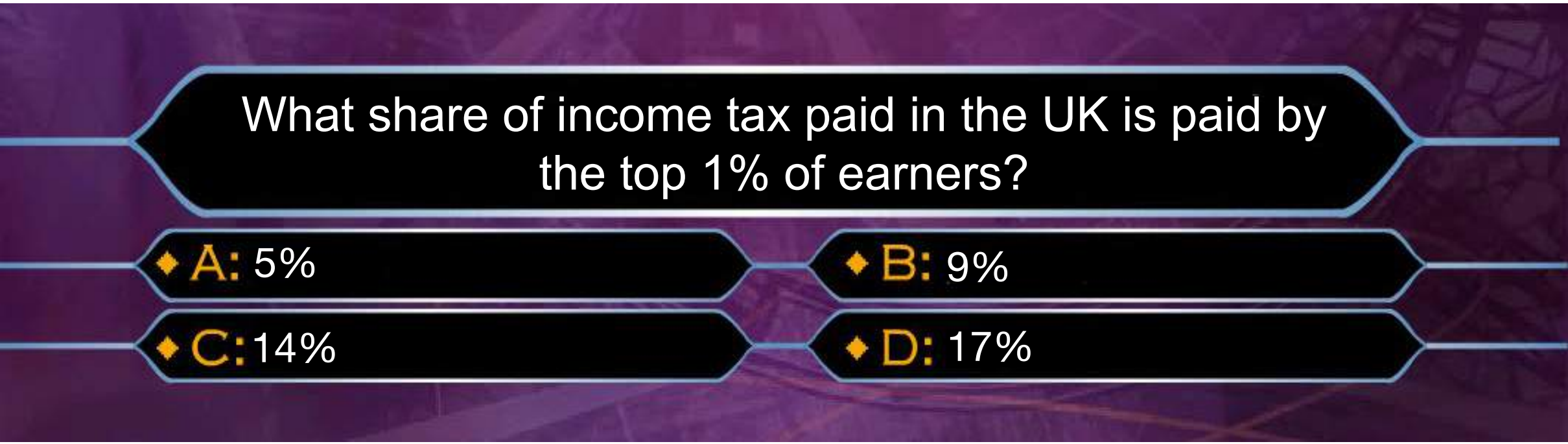
A decision tree is a uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Approaches

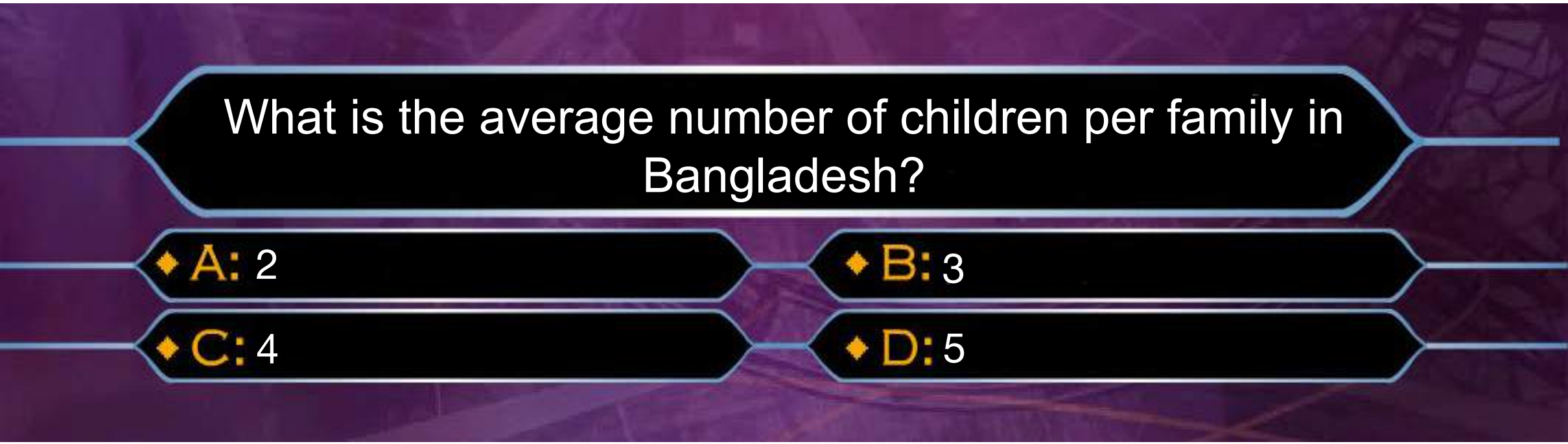
Data first

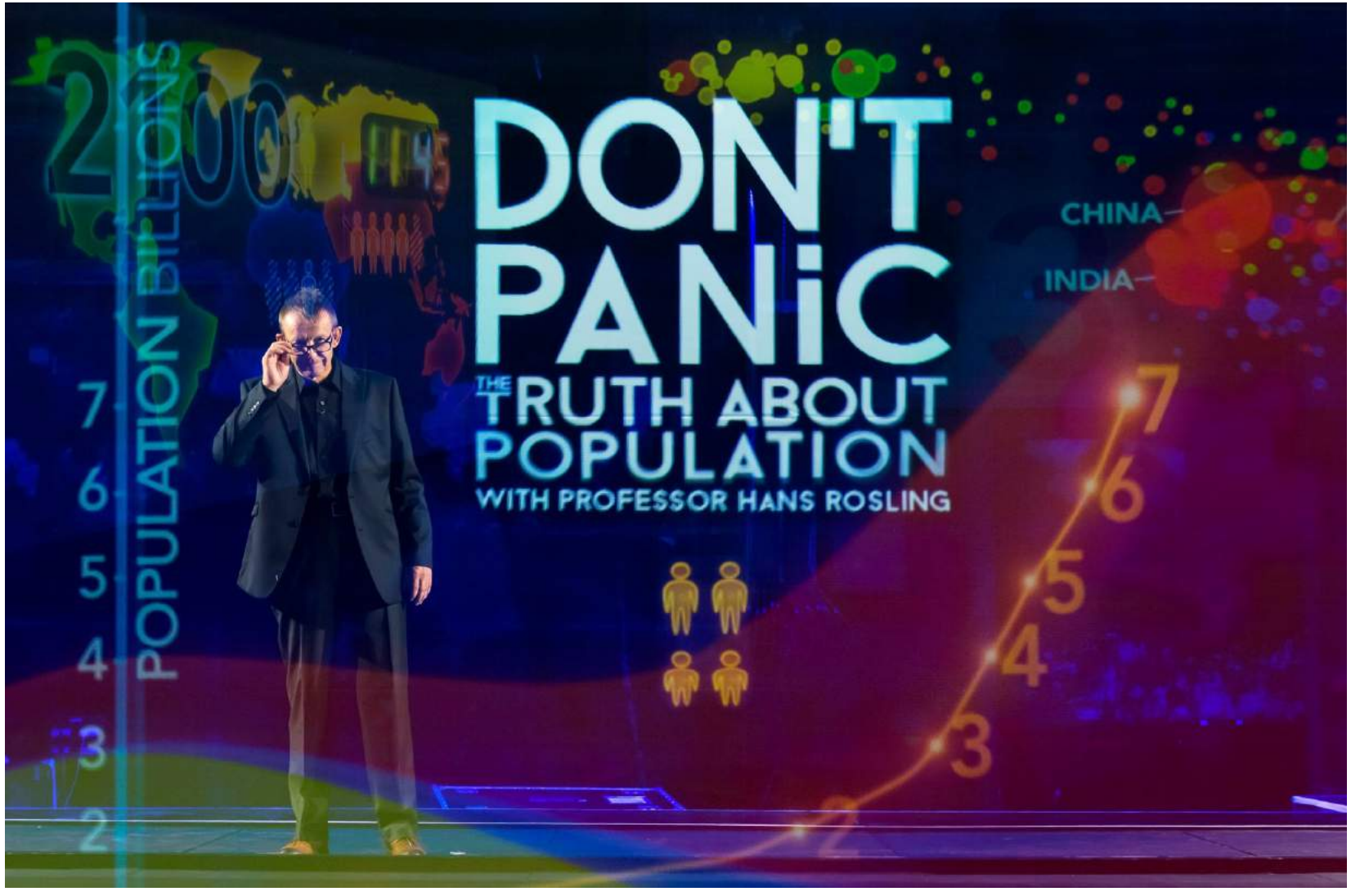
Knowledge first

Assumption


What share of income tax paid in the UK is paid by the top 1% of earners?
A: 5%
B: 9%
C: 14%
D: 17%

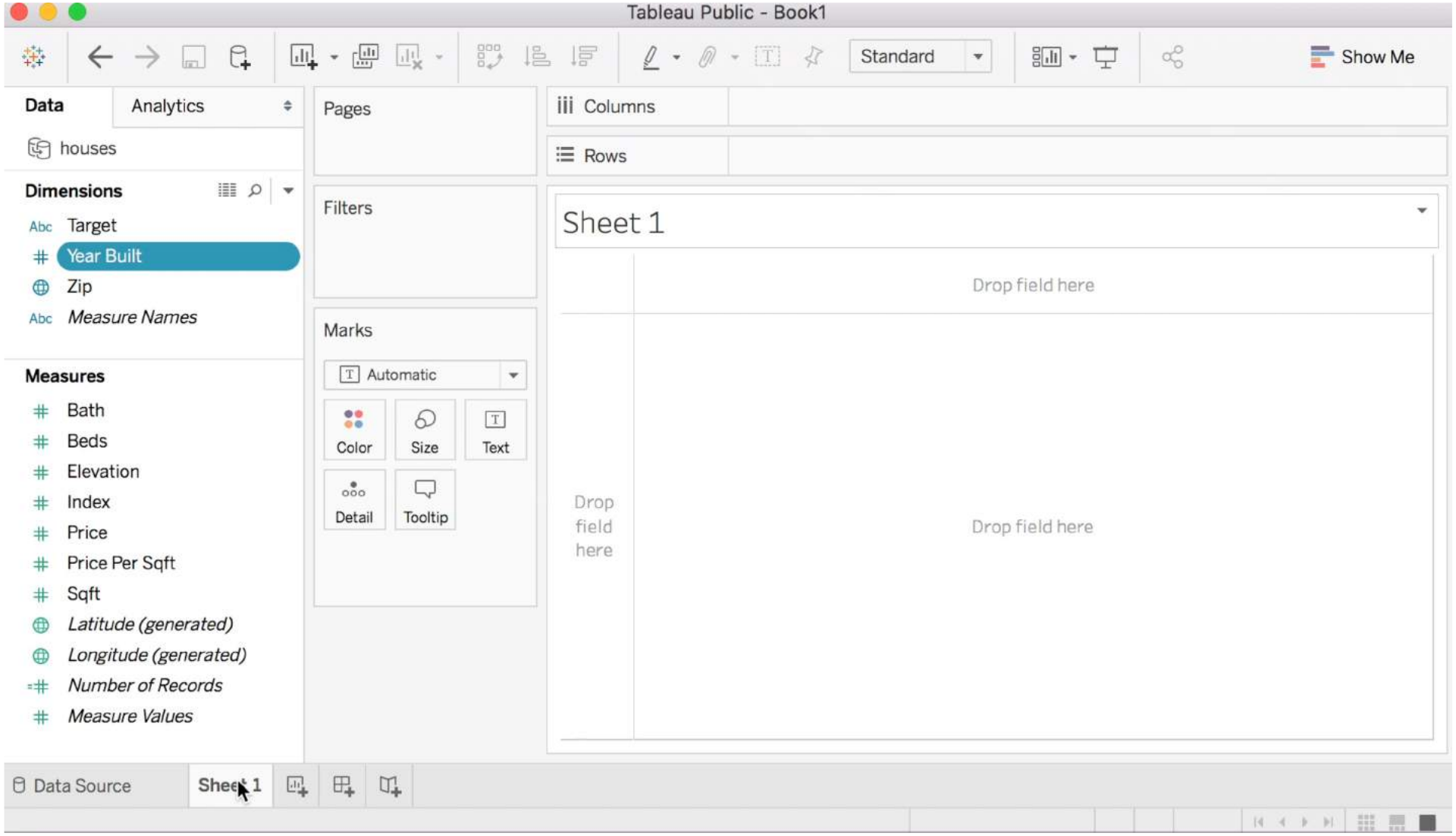
Assumption


What is the average number of children per family in Bangladesh?
A: 2
B: 3
C: 4
D: 5



gapminder.org

Data analysis for the houses dataset in Tableau



Options?

Grow the decision tree until…

1. every classification is perfect?
2. confidence level is above 80%?
3. it is too big to process the data in reasonable time?
4. until the evaluation and training set have the same confidence?
5. you have used all principal components?

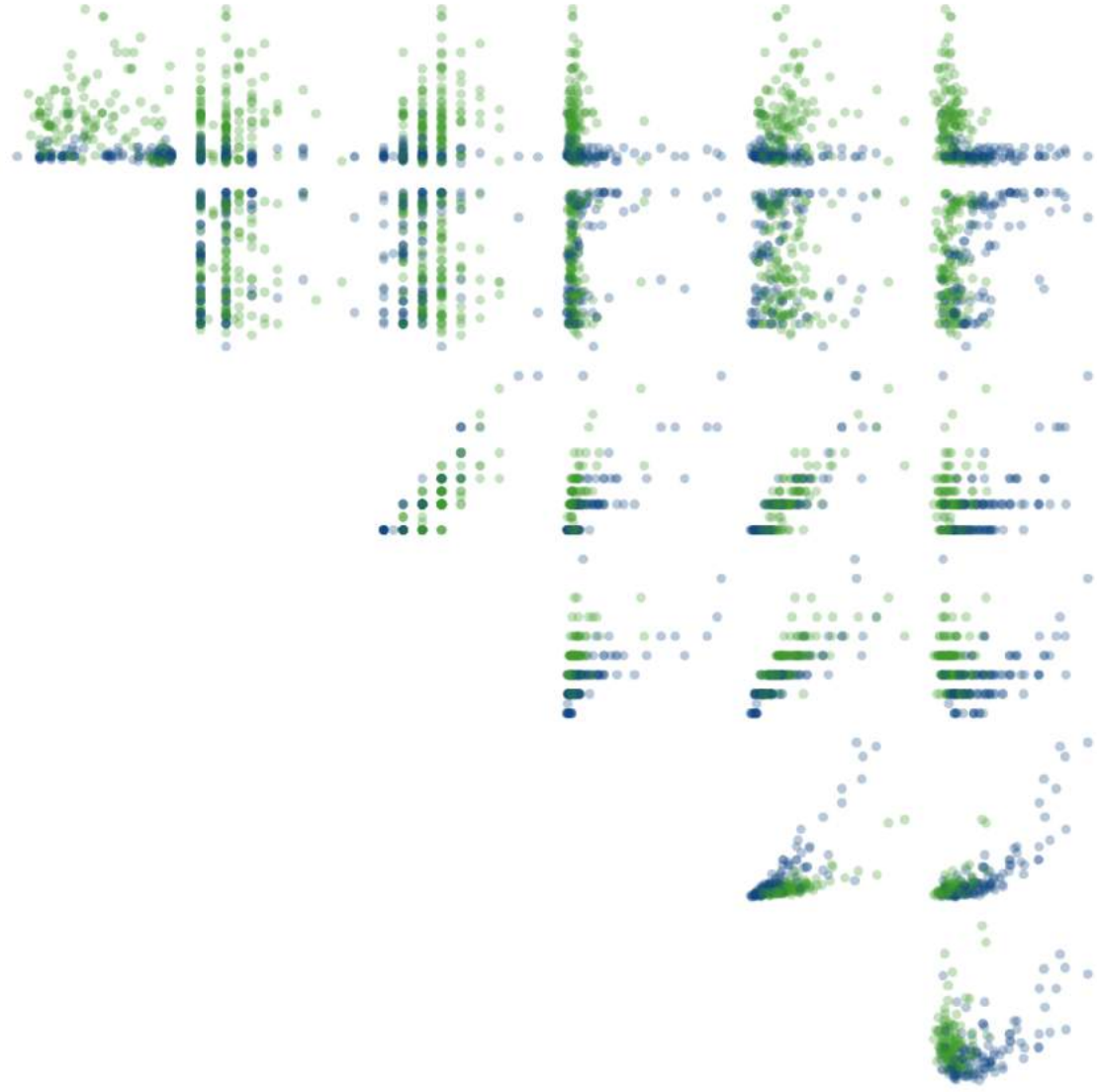
Solutions

Combine the data with existing knowledge. But make sure the existing knowledge it correct!

Grow the tree around the principal components (using PCA). Stop when only minor improvements are achieved on the training set. Don’t overfit!

A visual introduction to machine learning

r2d3.us

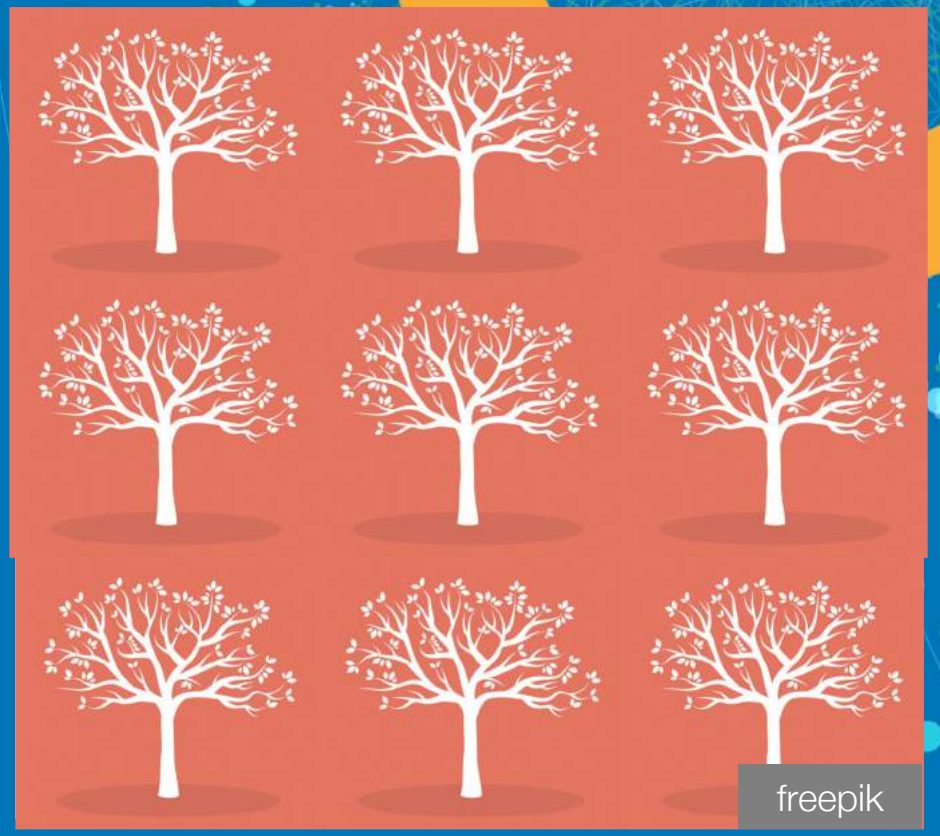


Combined result

Now give your red cards and your results table (right) to another group to see what result their tree gives for each card.

(Repeat this to get up to 3 results.)

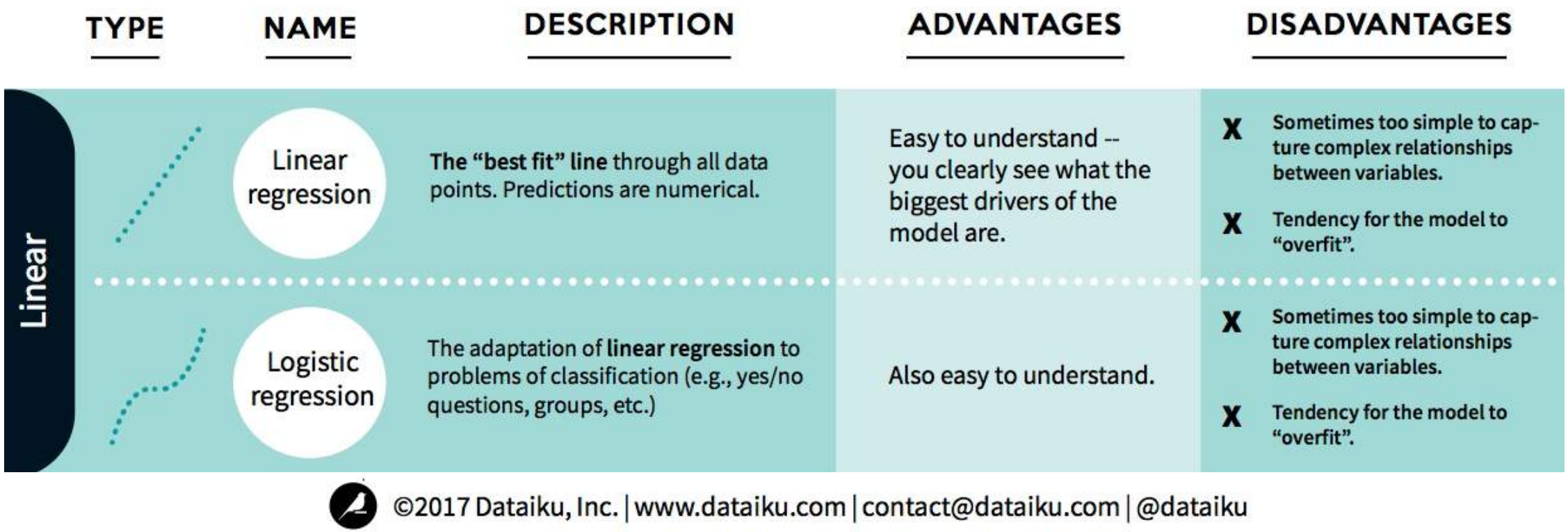
Card No.	Group #1	Group #2	Group #3
#23	SF	…	…
#59	NY		
#233	SF		



Random Forests

To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

Algorithm summaries



Algorithm summaries

Tree-based

Decision tree

A graph that uses a **branching method** to match all possible outcomes of a decision.

Easy to understand and implement.

- X **Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.**

Random Forest

Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**

A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.

- X **Can be slow to output predictions relative to other algorithms.**
- X **Not easy to understand predictions.**

Gradient Boosting

Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**

High-performing.

- X **A small change in the feature set or training set can create radical changes in the model.**
- X **Not easy to understand predictions.**

Algorithm summaries

Neural networks

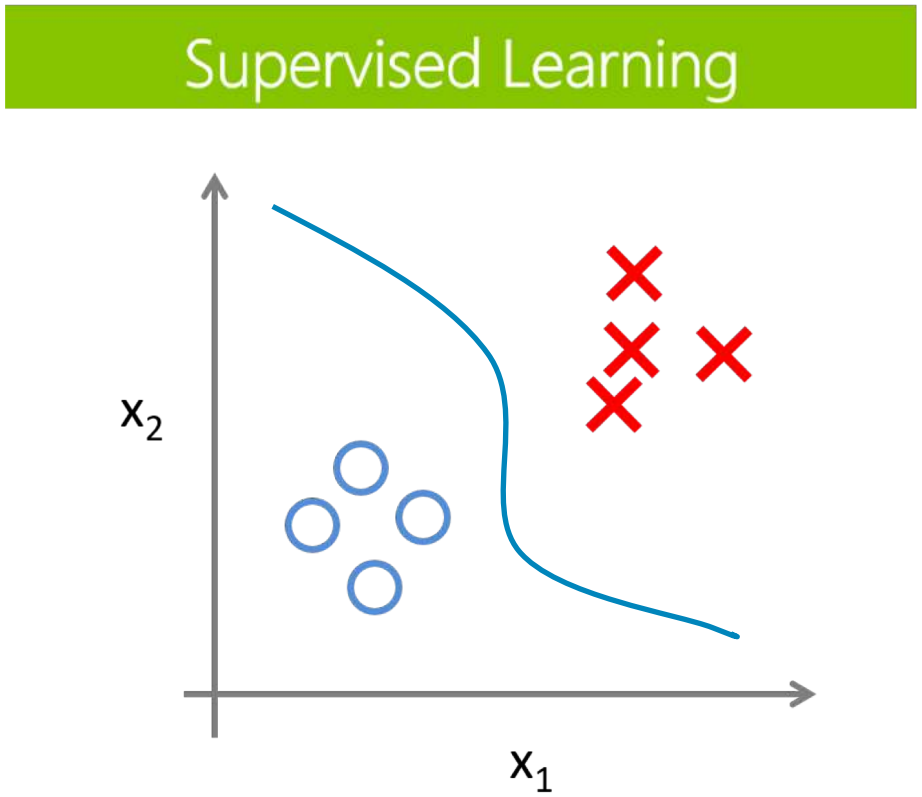
Neural networks

Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several **layers of neural networks put one after the other.**

Can handle extremely complex tasks - no other algorithm comes close in image recognition.

- X **Very, very slow to train, because they have so many layers. Require a lot of power.**
- X **Almost impossible to understand predictions.**

Supervised Learning



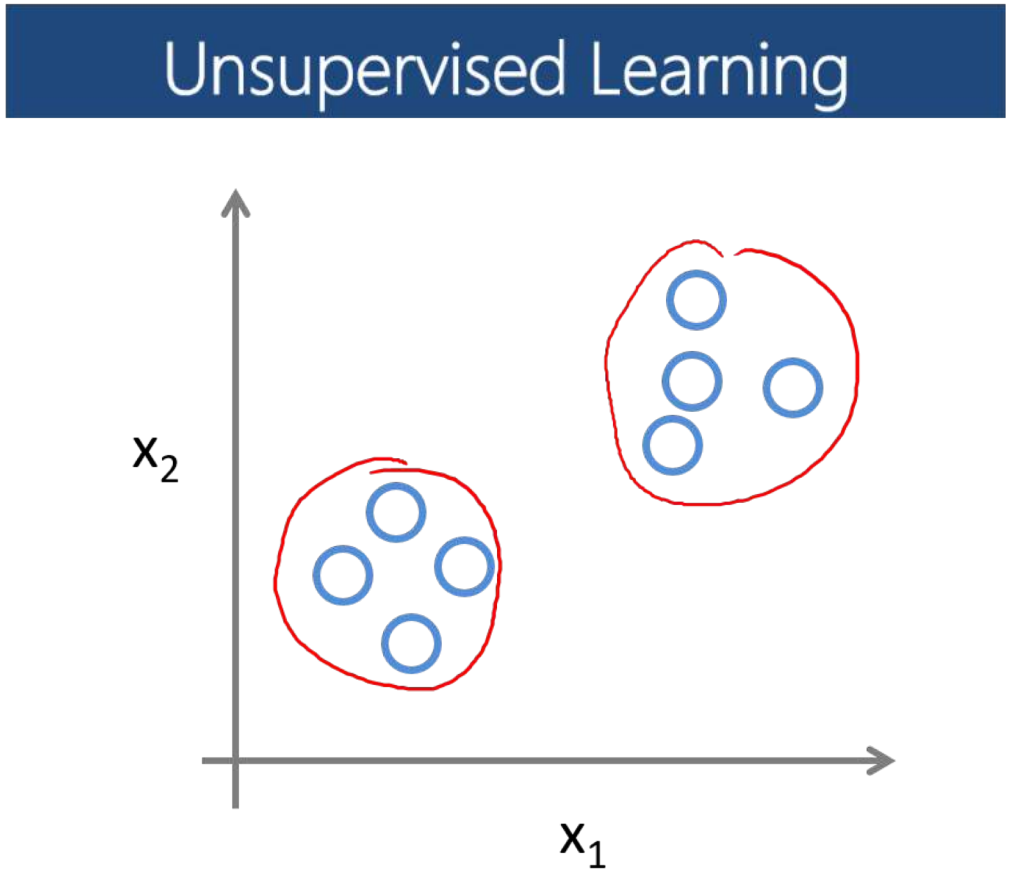
Supervised Learning

- Known target

Reinforcement Learning

- Learns from experience
- e.g. 80% of the that tactic failed (e.g. Markov decision process).

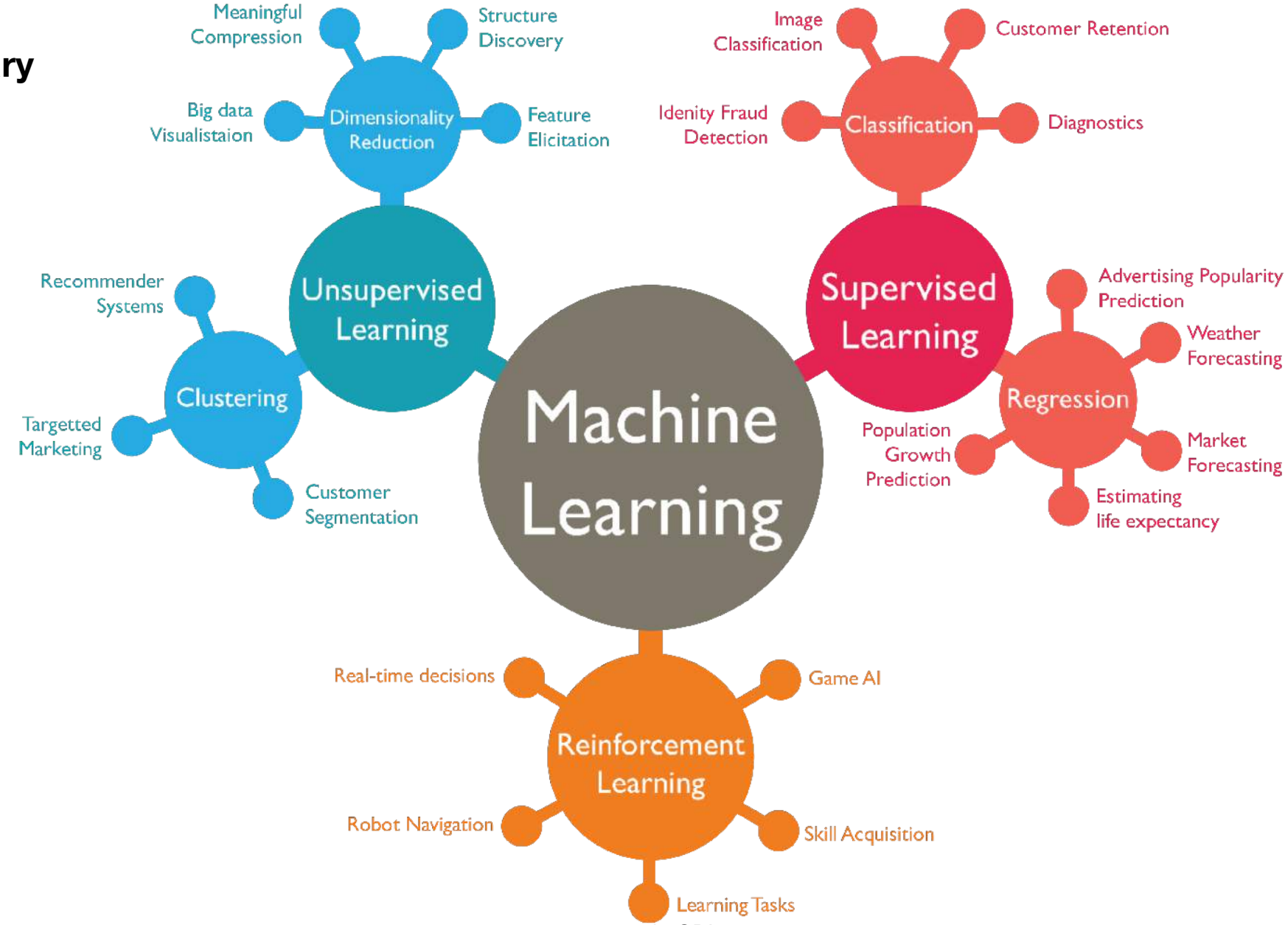
Unsupervised learning



It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.

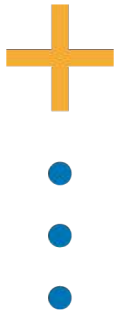
Examples of Unsupervised Learning: Apriori algorithm, K-means.

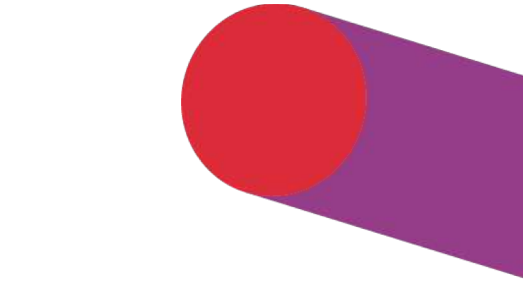
Summary

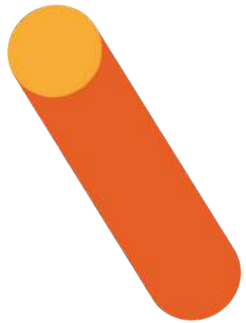

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

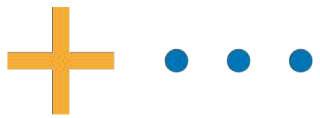
Contents:

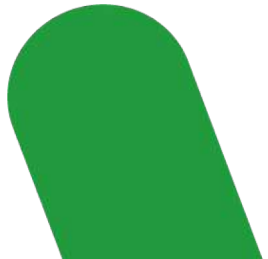
What happens when computers learn bad habits?











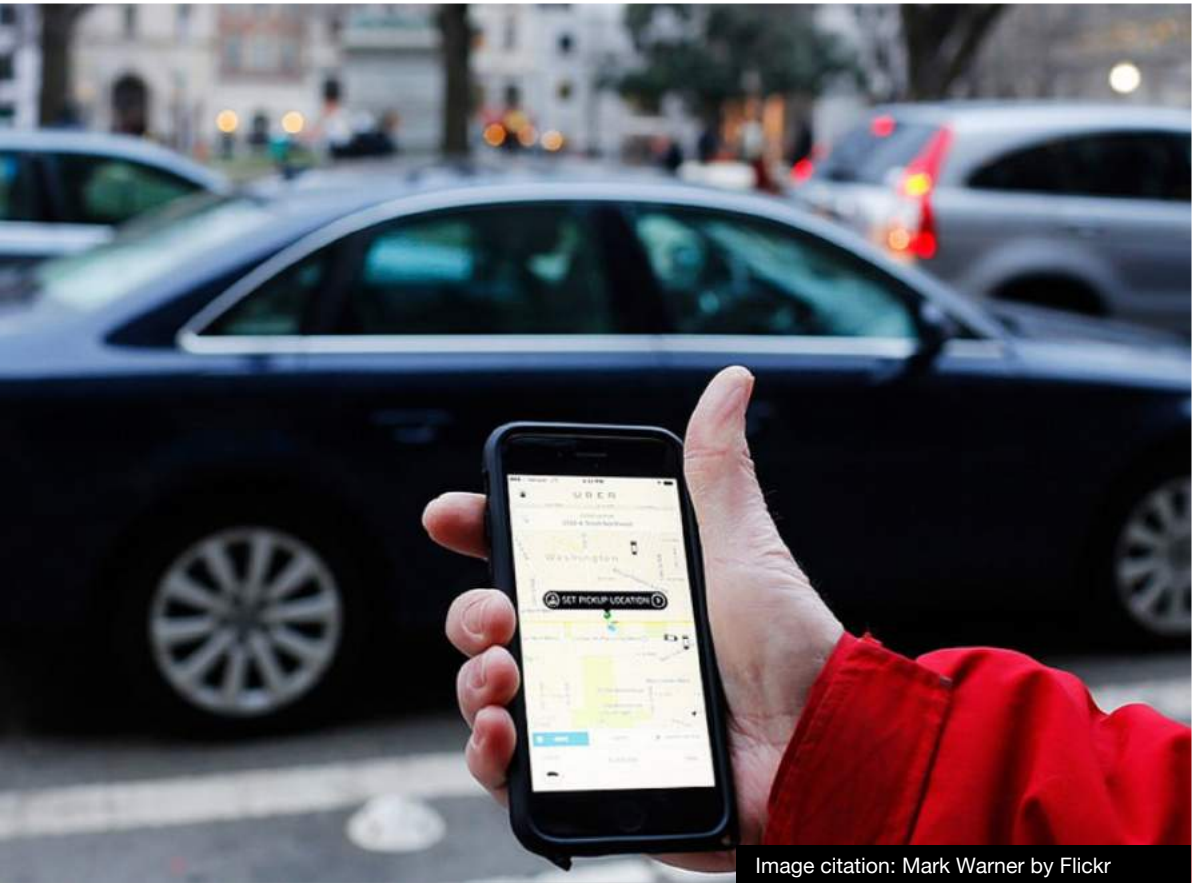


Image citation: Mark Warner by Flickr

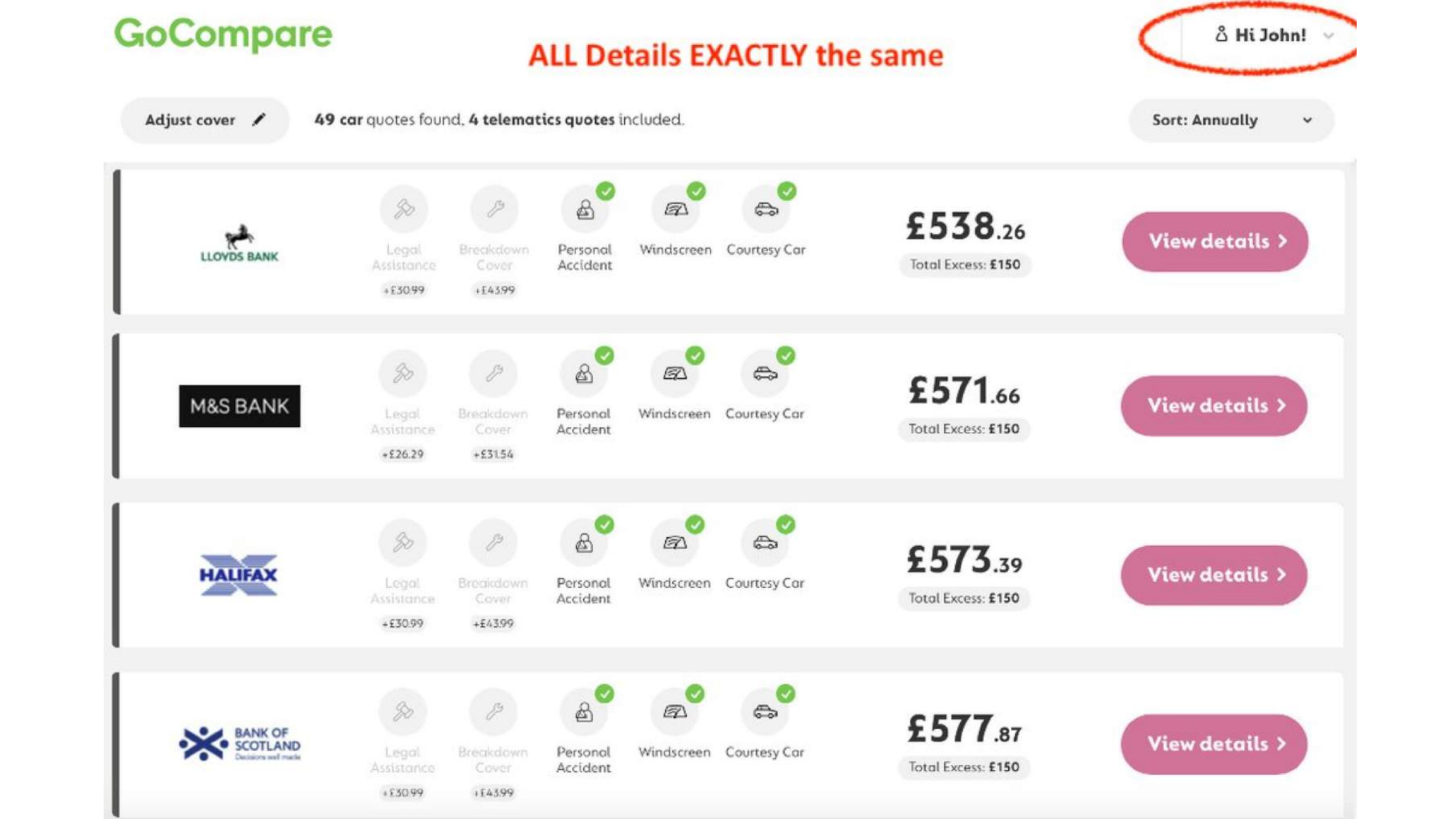
nile by Pixabay

GoodGuyPaul by Pixabay

Uber

Uber's automated pricing algorithm responds to surges in demand.

During the hostage crisis in Sydney in December 2014, the algorithm raised the prices by up to four times the normal rate in the crisis area.


GoCompare
ALL Details EXACTLY the same
Hi John!
Adjust cover
49 car quotes found, 4 telematics quotes included.
Sort: Annually
LLOYDS BANK
Legal Assistance +£30.99
Breakdown Cover +£43.99
Personal Accident
Windscreen
Courtesy Car
£538.26
Total Excess: £150
View details
M&S BANK
Legal Assistance +£26.29
Breakdown Cover +£31.54
Personal Accident
Windscreen
Courtesy Car
£571.66
Total Excess: £150
View details
HALIFAX
Legal Assistance +£30.99
Breakdown Cover +£43.99
Personal Accident
Windscreen
Courtesy Car
£573.39
Total Excess: £150
View details
BANK OF SCOTLAND
Legal Assistance +£30.99
Breakdown Cover +£43.99
Personal Accident
Windscreen
Courtesy Car
£577.87
Total Excess: £150
View details

GoCompare

Car: Ford Fiesta Ghia 2002-2008 1.6 Petrol
Profession: Insurance Director
Address: Milford Haven (PPI Company :P)

Always resided in UK
Date of birth: 01/01/1980
No claims and license: 16 years
Car kept on drive

Hi Mohammed!

Adjust cover

37 car quotes found, 5 telematics quotes included.

Sort: Annually

	Legal Assistance	Breakdown Cover	Personal Accident	Windscreen	Courtesy Car		
LLOYDS BANK	+£30.99	+£43.99	✓	✓	✓	£1,446.32 Total Excess: £150	View details >
HALIFAX	+£30.99	+£43.99	✓	✓	✓	£1,458.04 Total Excess: £150	View details >
BANK OF SCOTLAND Decisions well made	+£30.99	+£43.99	✓	✓	✓	£1,459.64 Total Excess: £150	View details >
M&S BANK	✓	✓	✓	✓	✓	£1,476.70 Total Excess: £150	View details >

Image citation: stevepb by Pixabay

nile by Pixabay

GoodGuyPaul by Pixabay

Car insurance

At the beginning of 2018, large global firms like Admiral and Marks & Spencers faced public backlash when the Sun newspaper found that insurance quotes for drivers with the traditional English name ‘John’ were far lower than quotes of the same for drivers named ‘Mohammed’.

Image citation: DarkoStojanovic by Pixabay

Counselling by Pixabay

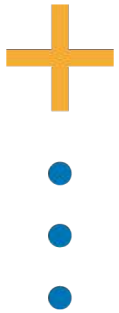
wp paarz by Flickr

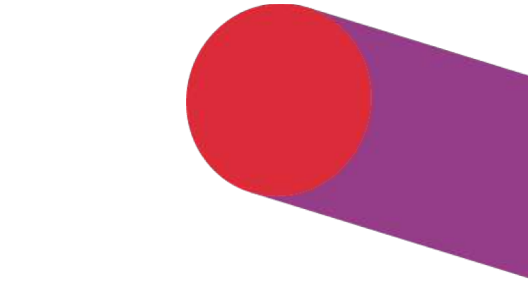
Migrant data

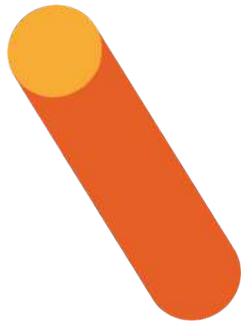
In 2017, NHS Digital signed an MOU with the Home Office to share non-clinical data about migrants e.g addresses, contact details, GP office. The MOU was cancelled in May 2018 amide discussions on the UK Data Protection Bill.

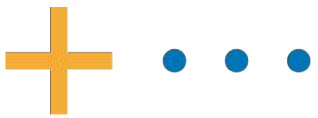
It’s not that easy.

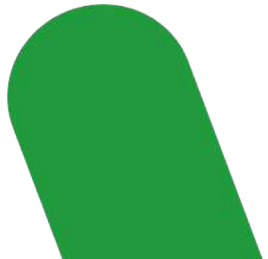
There are cultural and societal implications and evolving social norms. Let’s consider this...

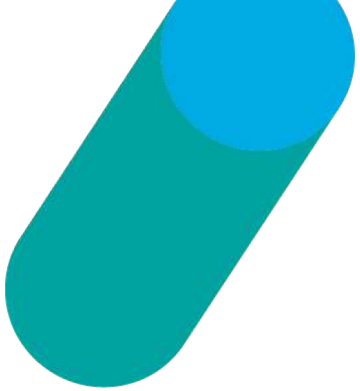












The trolley problem

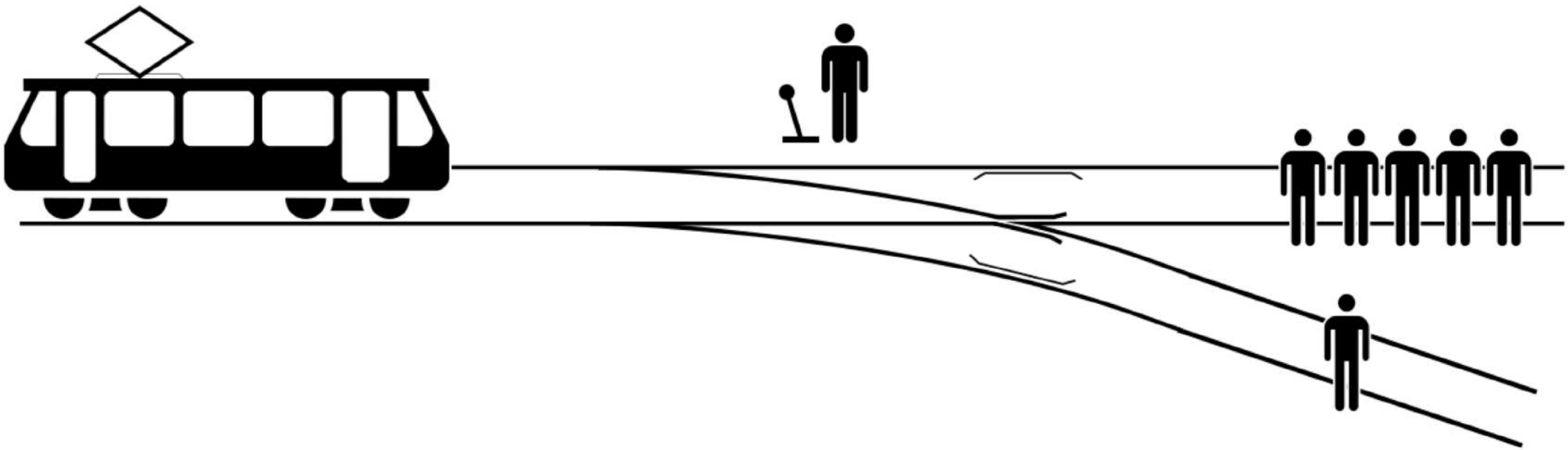


Image citation: Zapyon by Wikipedia

The Trolley Problem

You are taking your daily walk near the train tracks when you notice that the train that is approaching is out of control. You see what has happened: the driver of the train saw five men working on the track ahead and slammed on the brakes, but the brakes failed and the driver fainted.

The train is moving so fast that anyone it hits will die immediately.

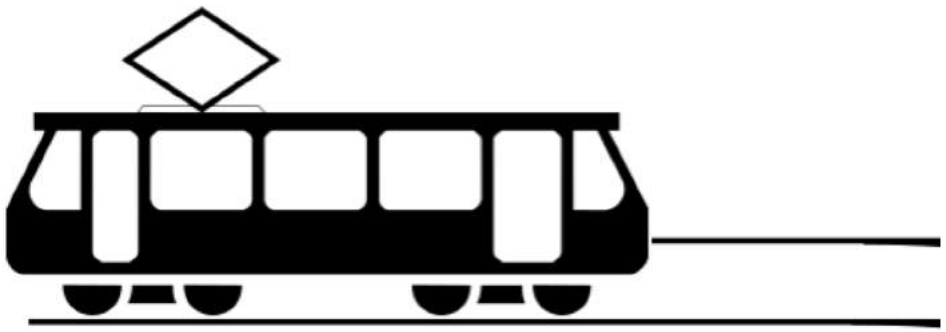
There are five people working on the main track. It is obvious that they will not be able to get off the track in time and, if nothing is done, they will be killed.

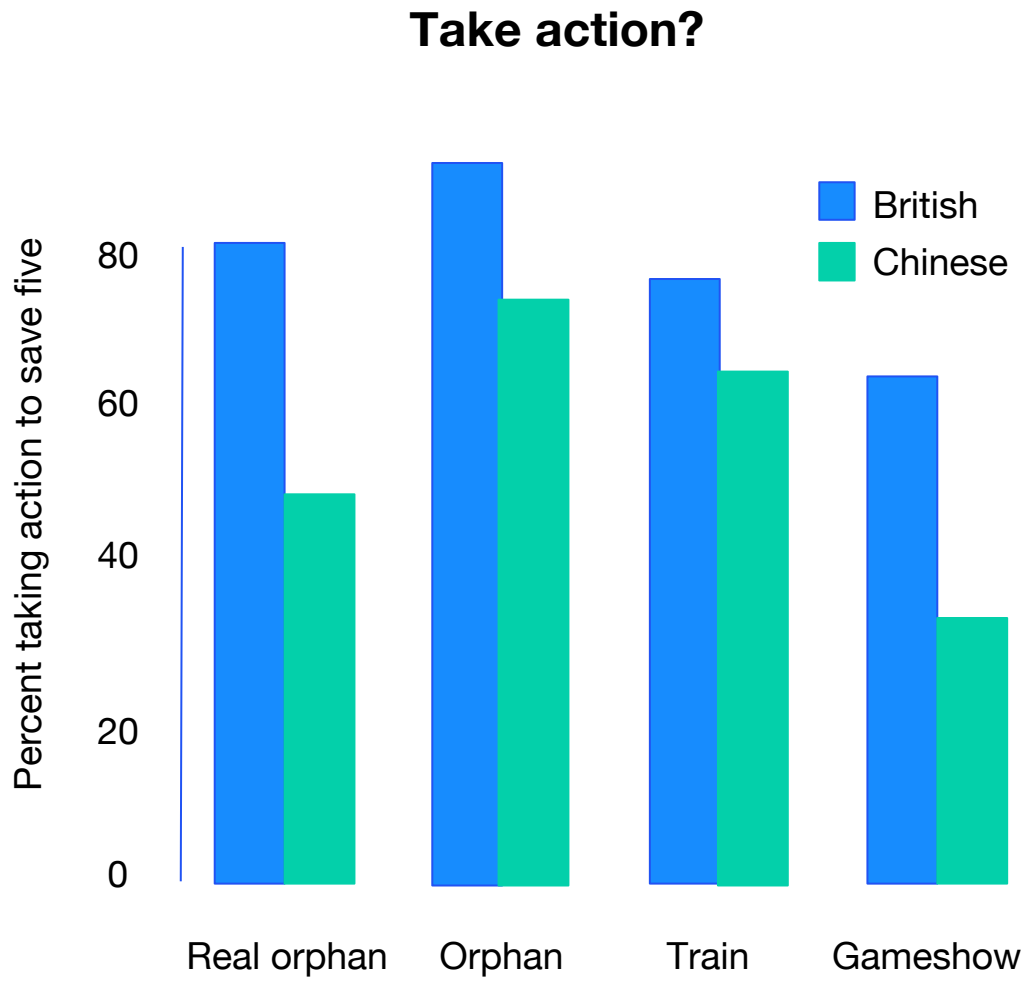
The track has a side-track leading off to the left. You are standing next to a lever. If you pull the lever, that will turn the train onto the side track and the five people on the main track will not die. But a person is working on the side track. If the train goes onto the side track, then the person on the side track will die. You are aware of all these facts.

Thus, you can pull the lever, in which case the one person will die but the five people will not; or you can refrain from pulling the lever, in which case the five people will die but the one person will not.

Do you pull the lever? (Y or N)

The trolley problem





A branch of ethics that evaluates data practises with the potential to adversely impact on people and society - in data collection, sharing and use.

The ODI

Data sources

Name and describe key data sources used in your project, whether you're collecting them yourself or getting access from third parties.

1.

Limitations in your data sources

Are there any limitations that might influence the outcomes of your project? Consider:

- bias in data collection, inclusion, algorithm
- gaps, omissions
- other sensitivities such as data categorisation

2.

Sharing this data with other organisations

Are you going to be sharing data with other organisations? If so, who?

Under what conditions?

3.

Relevant legislation and policies

What legislation or policies shape your use of this data? Eg data protection legislation, IP and database rights legislation, anti-discrimination laws, sector-specific data sharing policies/regulation (eg health, employment, taxation), sector-specific ethics legislation

4.

Rights over data sources

Where did you get the data from? Is it data produced by an organisation or data collected directly from individuals?

Do you have permission or another basis on which you're allowed to use this data? What ongoing rights will the data source have?

5.

Existing ethical frameworks

Countries, sectors and communities have existing ethical codes and frameworks. Which ones are relevant to this project?

6.

Your reason for using this data

What is your primary purpose for using data in this project?

What are you attempting to do?

What is your primary use case and your business model?

Are you replacing another project or service? How ethical was that?

Are you making things better? How? For whom?

7.

Communicating your purpose

Do people, especially those the data is about or who are impacted by its use, understand your purpose?

Who has been told about your purpose? Has this communication been clear?

8.

Positive effects on people

Which individuals, demographics or organisations will be positively affected by this project?

How will they be positively affected?

How could you increase the positive impact of this project?

How are you measuring positive impact?

9.

Negative effects on people

Who could be negatively affected by this project?

Could the manner in which this data is collected, shared and used:

- cause harm?
- be used to target, profile or prejudice people?
- unfairly restrict access (eg exclusive arrangements)?

Could people perceive it to be harmful?

10.

Minimising negative impact

What steps can you take to minimise harm?

Are there measures you could take to reduce limitations in your data sources?

Could you monitor potential negative impact to support mitigating activities?

What benefits will these actions add to your project?

How are you measuring negative impact?

11.

Engaging with people

How can people engage with you?

Can people affected appeal or request changes to the service?

To what extent?

Are the appeal mechanisms reasonable?

12.

Communicating risks and issues

Are you building into the project the thoughts, ideas and considerations of people affected by your project? How?

Are you communicating potential risks or issues?

How are limitations and risks being communicated to people? Consider:

- those the data is about
- those impacted by its use or affected by your project
- organisations using data

What methods are you using?

13.

Reviews and iterations

How will ongoing issues related to data ethics be monitored and discussed?

When will your responses to the canvas be reviewed or updated?

14.

Your actions

What actions are you going to take before moving forward with this project?

Which of them should take priority?

Who will be responsible for these actions and who needs to be involved in them?

Will you publish your actions and answers to this canvas openly?

15.

Pokemon Go

The game mixes up data about the real-world such as maps, streets and popular landmarks such as museums, art galleries, restaurants and statues with fictional data about wild animals.

The role of data in Artificial Intelligence

Machine learning, AI & ethics

training.theodi.org/ML101

Dr David Tarrant
@davetaz
theODI.org

Any questions?

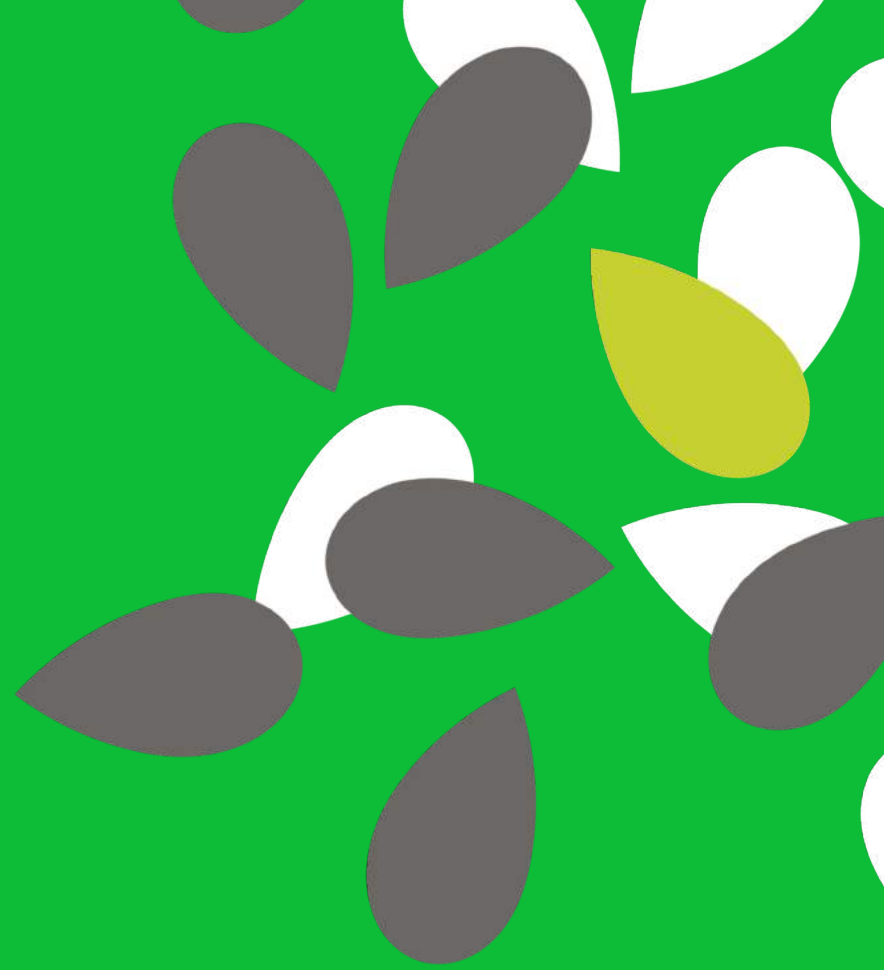
Get in touch

If you would like to talk to us about collaborating, partnering, supporting our work, or anything else, we'd love you to get in touch.

info@theodi.org

+44 (0)20 3598 9395

@ODIHQ



open data institute