Civil Service
Learning

Open data driven policy making

AN2: Data and analysis

Please fill in the register
bit.ly/csl-register

Welcome

Commissioned by

Civil Service Learning

Delivered by

open data institute

Dr David Tarrant

The Open Data Institute

Lucy Knight

Local Government

Civil Service
Learning

Aim

To equip policy makers with the knowledge and skills they need to effectively use open data in policy making

Civil Service Learning

Introductions

Who are you?

What is your role in relation to data driven policy making?

What would you like to get out of this training?

Civil Service Learning

Today

Introduction to open data driven policy making

Benefits, caveats and risks of using data in policy making

Planning a data driven policy process

Data driven policy making in practice

Civil Service
Learning

Data driven policy making

Civil Service Learning

Introduction to data driven policy making

Outcomes

1. Define data, big data and open data.
2. Describe how data is used to inform policy making in different fields.
3. Identify the benefits of using data in policy making.
4. Assess the risks, caveats and limitations of using data in policy making.

Civil Service Learning

Exercise

What is data?

In as few words as possible, define ‘data’.

You can use an example as an answer if you wish.

One twist: you cannot use the word **information** in your answer.

Civil Service Learning

What is data?

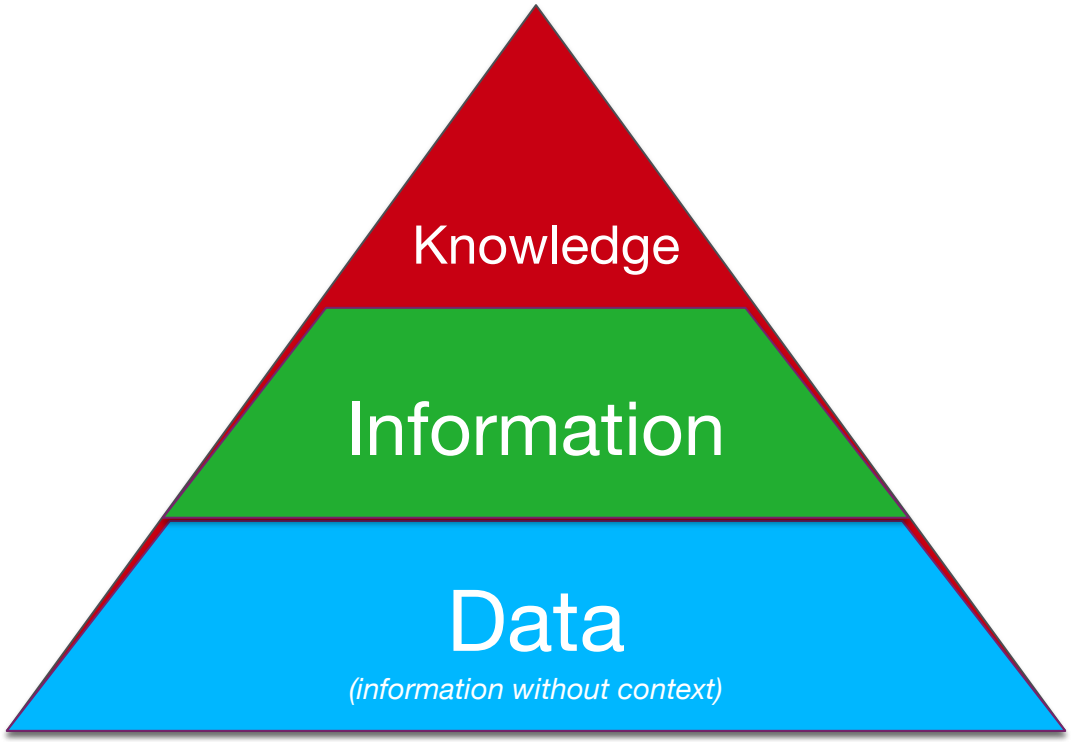
- A collection of facts, information and statistics that can be analysed to develop new knowledge.

- A collection of numbers assigned as values to quantitative variables and / or characters assigned as values to qualitative variables.

- The lowest level of abstraction from which information and then knowledge are derived.

Civil Service Learning

How is data different from information?



Civil Service
Learning

Civil Service
Learning

What did you see?

Data: What did you see?

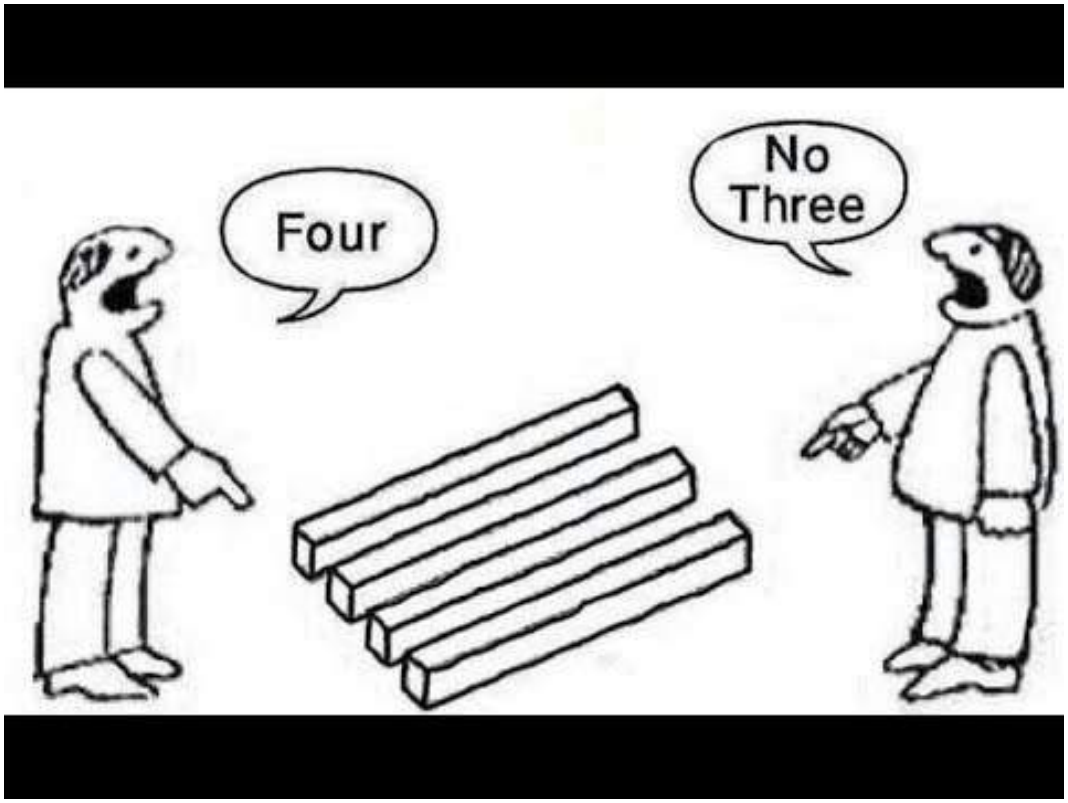
Information: How did you interpret it?

Knowledge: Did it mean (or relate to) anything for you?

Civil Service
Learning

Seeing what isn't there

Even counting is difficult (more on that later)



Civil Service
Learning

Exercise

What is open data?

In as few words as possible, what is ‘open data’?

Civil Service Learning

A piece of data or content is open if anyone is free to use, reuse, and redistribute it - subject only, at most, to the requirement to attribute and/or share-alike.

Open Knowledge International

Civil Service Learning

Data that anyone can access, use and share.

The Open Data Institute

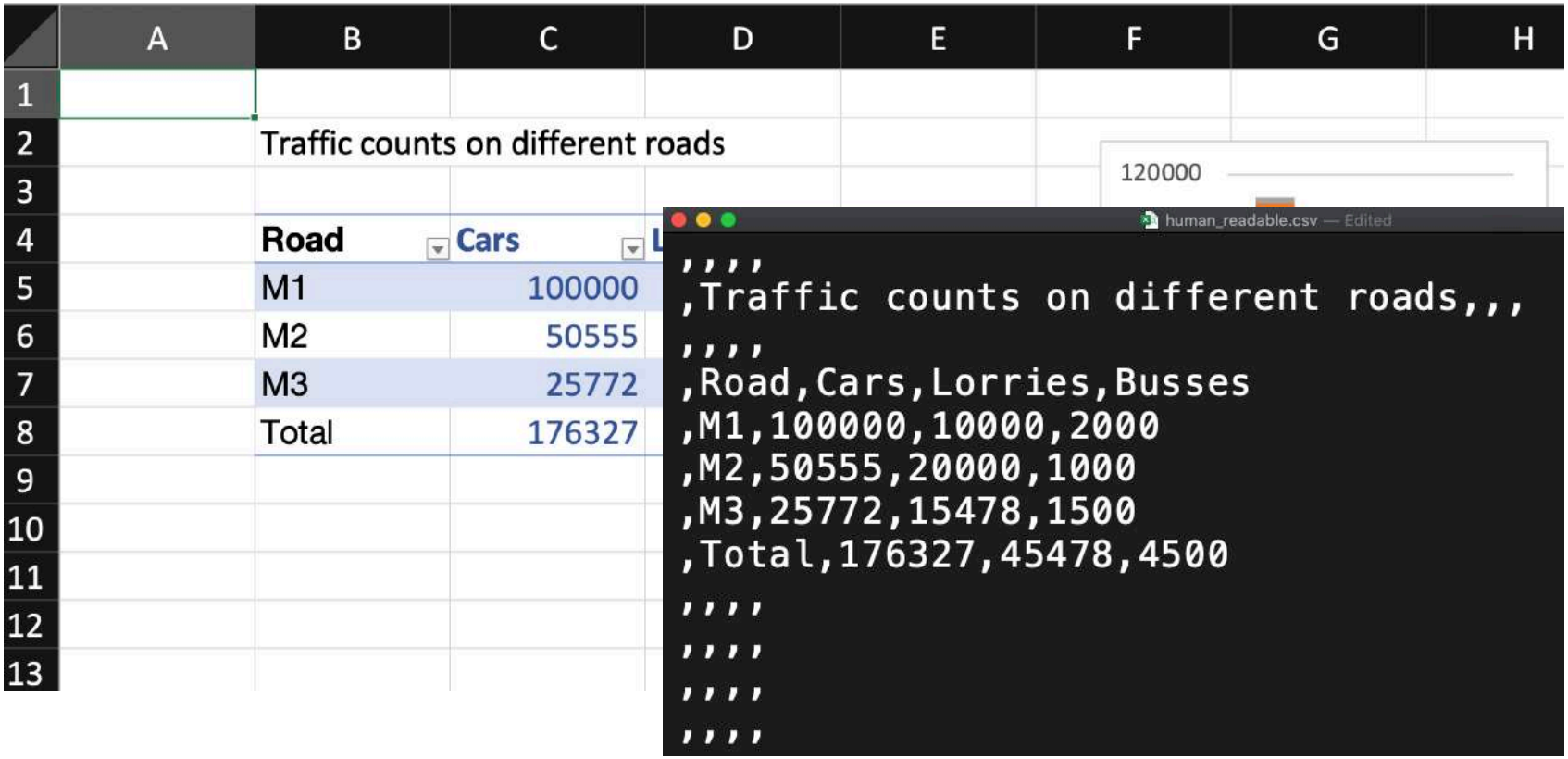
Civil Service Learning

Open data is data that is published in an open format, is machine readable and is published under a license that allows for free reuse.

data.gov.uk

Civil Service
Learning

Machine readable data?



	A	B	C	D	E	F	G	H
1								
2		Traffic counts on different roads						
3								
4		Road	Cars					
5		M1	100000					
6		M2	50555					
7		M3	25772					
8		Total	176327					
9								
10								
11								
12								
13								

Machine readable data?

	A	B	C	D
1	Road	Cars	Lorries	Busses
2	M1	100000	10000	2000
3	M2	50555	20000	1000
4	M3	25772	15478	1500
5				
6				
7				
8				

machine_readable.csv

```
Road,Cars,Lorries,Busses
M1,100000,10000,2000
M2,50555,20000,1000
M3,25772,15478,1500
```

Civil Service
Learning

Open data

The important points are:

- licensed openly (e.g. Open Government Licence)
- free to use, not always free to access (currently government open data must be available at no cost)
- users must be able to modify and redistribute the data

Data.gov.uk includes:

- published in open format
- machine readable

Civil Service Learning

Case study: LIDAR

DATA.GOV.UK Beta
Opening up Government

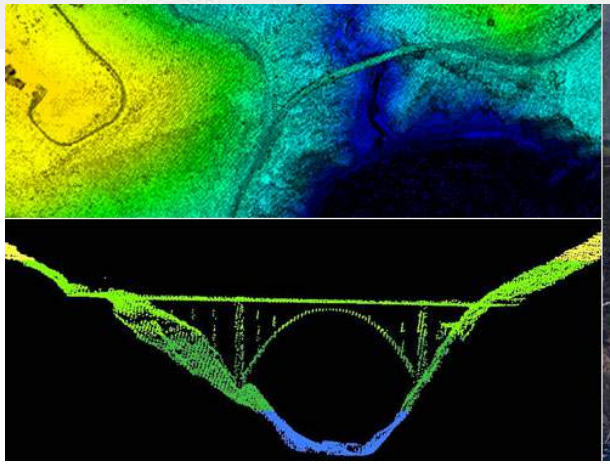
Home

Datasets | Map Search | Data Requests | Publishers | Data API | Organograms | Site Analytics

Home / Datasets / LIDAR Composite DTM - 50cm

LIDAR Composite DTM - 50cm

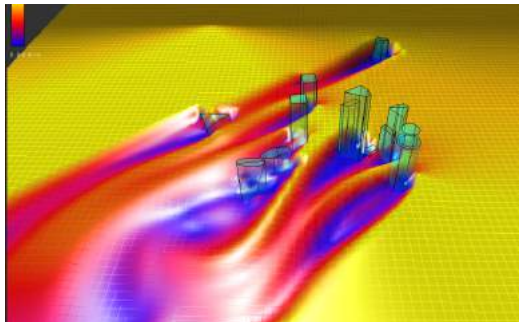
Published by Environment Agency. Licensed under OGL Open Government Licence.
Openness rating: ★★★☆☆



Civil Service Learning

Impact of LIDAR data

Wind modelling



Archology

Educational games (Minecraft)

Civil Service Learning

Video

Q: What types and uses of data do Emily and David focus on?

Civil Service
Learning

Civil Service Learning

Review

What types of data do David and Emily focus on?

EMILY focusses on open data and data that is already showing benefit regardless of how big or small it is. Such as LIDAR, Transport, Environmental, Energy and other non-personal data.

DAVID focusses on Big Data and Exhaust Data that come from social interactions and from people. This brings into question lots of ethical and data protection issues and the Big Data Hubris.

Civil Service Learning

Big data is changing the world

But what is big data?

Collect some ideas together on some post-it notes. Do you have any categories emerging?

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.

Civil Service
Learning

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]

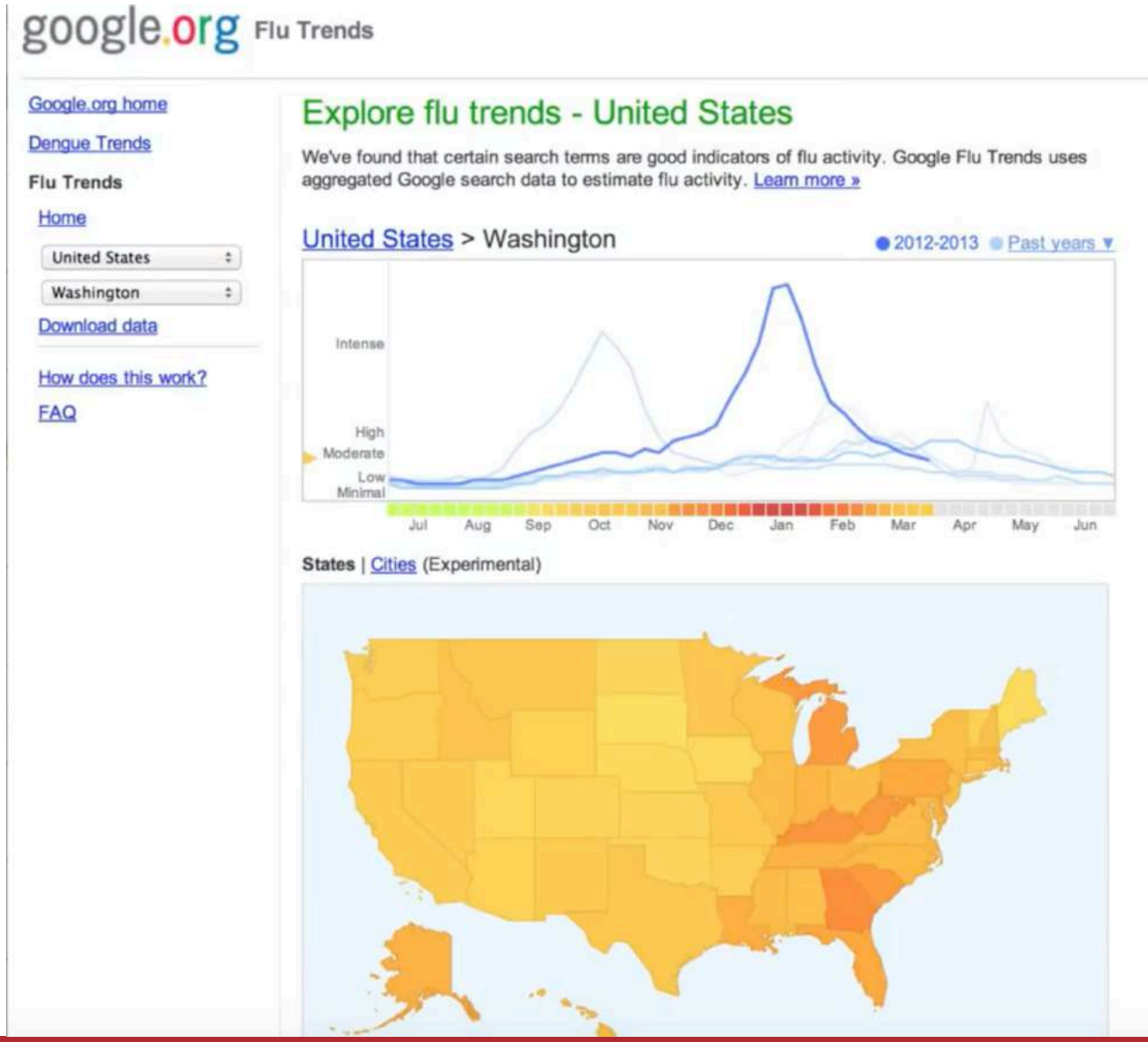
Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human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

Civil Service Learning

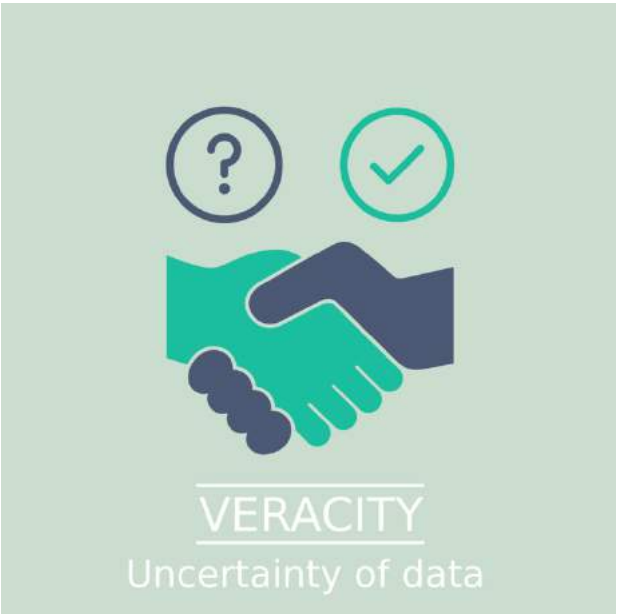
97%

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



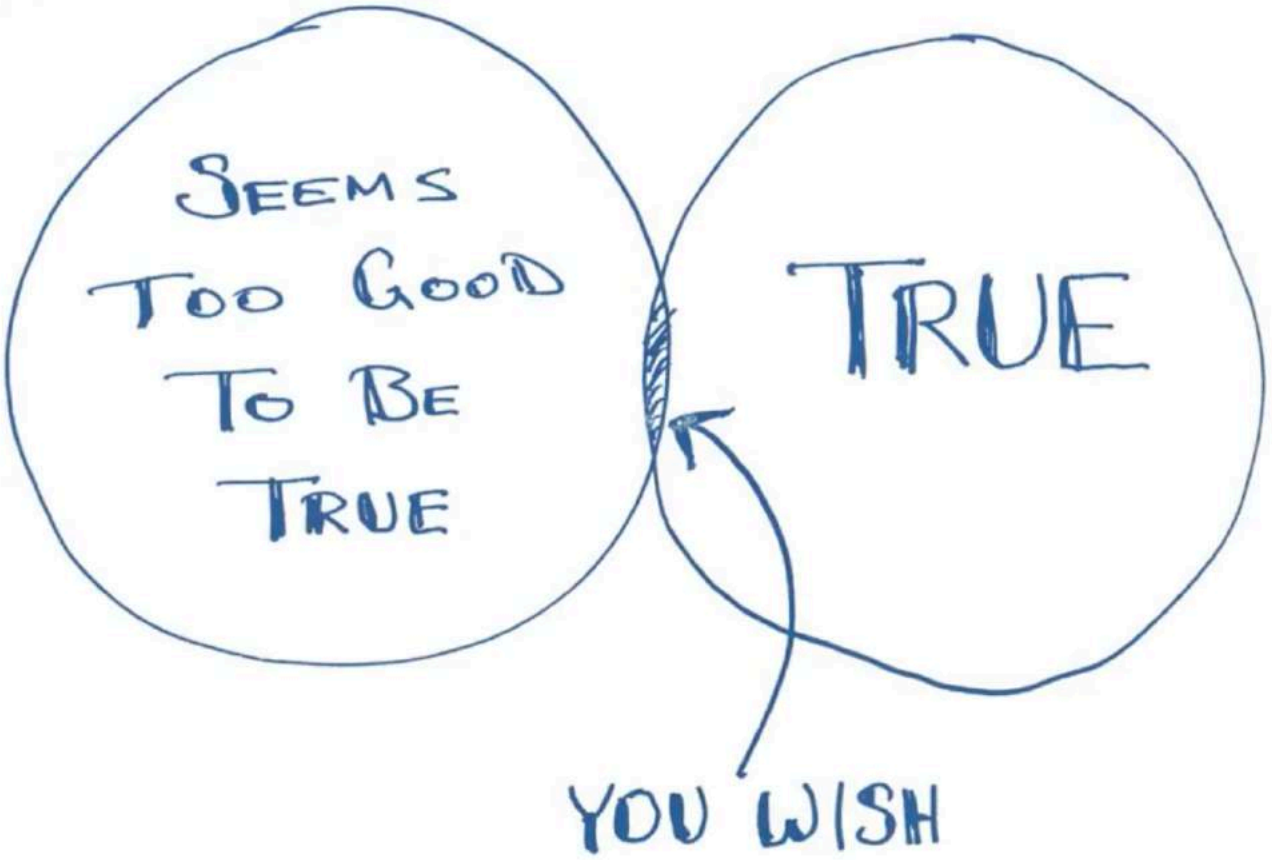

VARIETY
More data sources
VOLUME
More data collected
BIG
DATA
VERACITY
Uncertainty of data
VELOCITY
Faster than ever

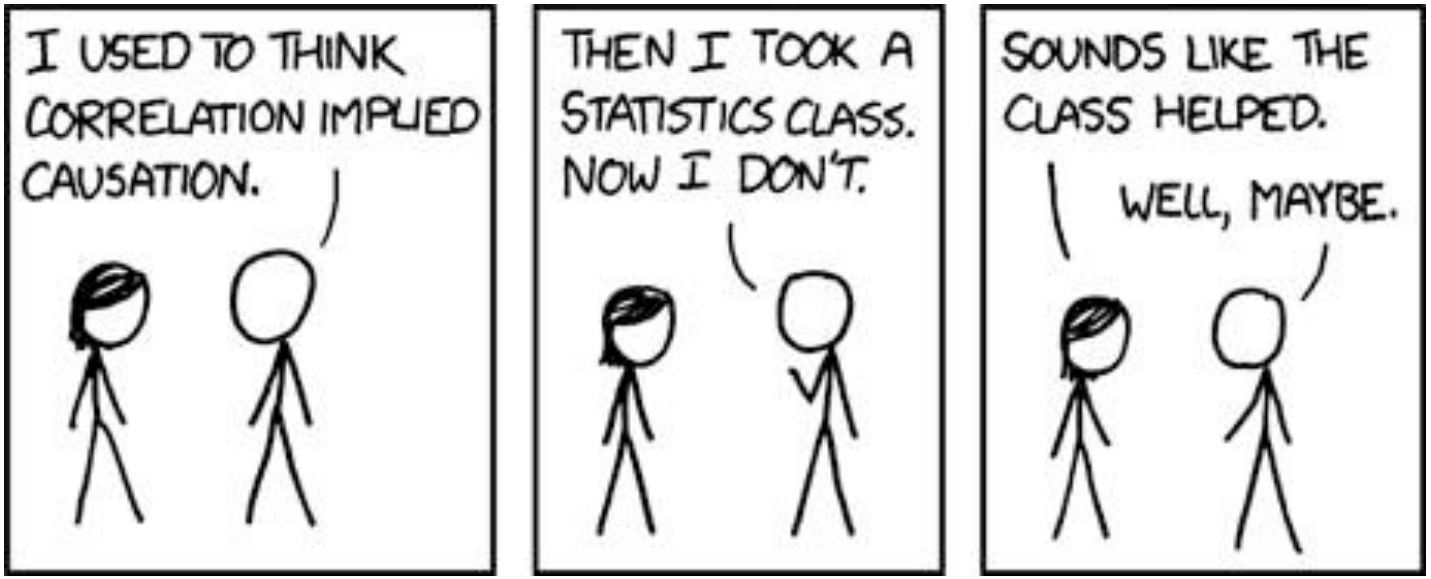

Google
fever aching joints headache
All
Shopping
Images
News
Videos
More
Settings
Tools
About 369,000 results (0.52 seconds)
WebMD Symptom Checker helps you find the most common medical conditions indicated by the symptoms chills, fever, headache and joint aches including Lyme disease, Acute sinusitis, 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



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?


SEEMS
TOO GOOD
TO BE
TRUE
TRUE
YOU WISH



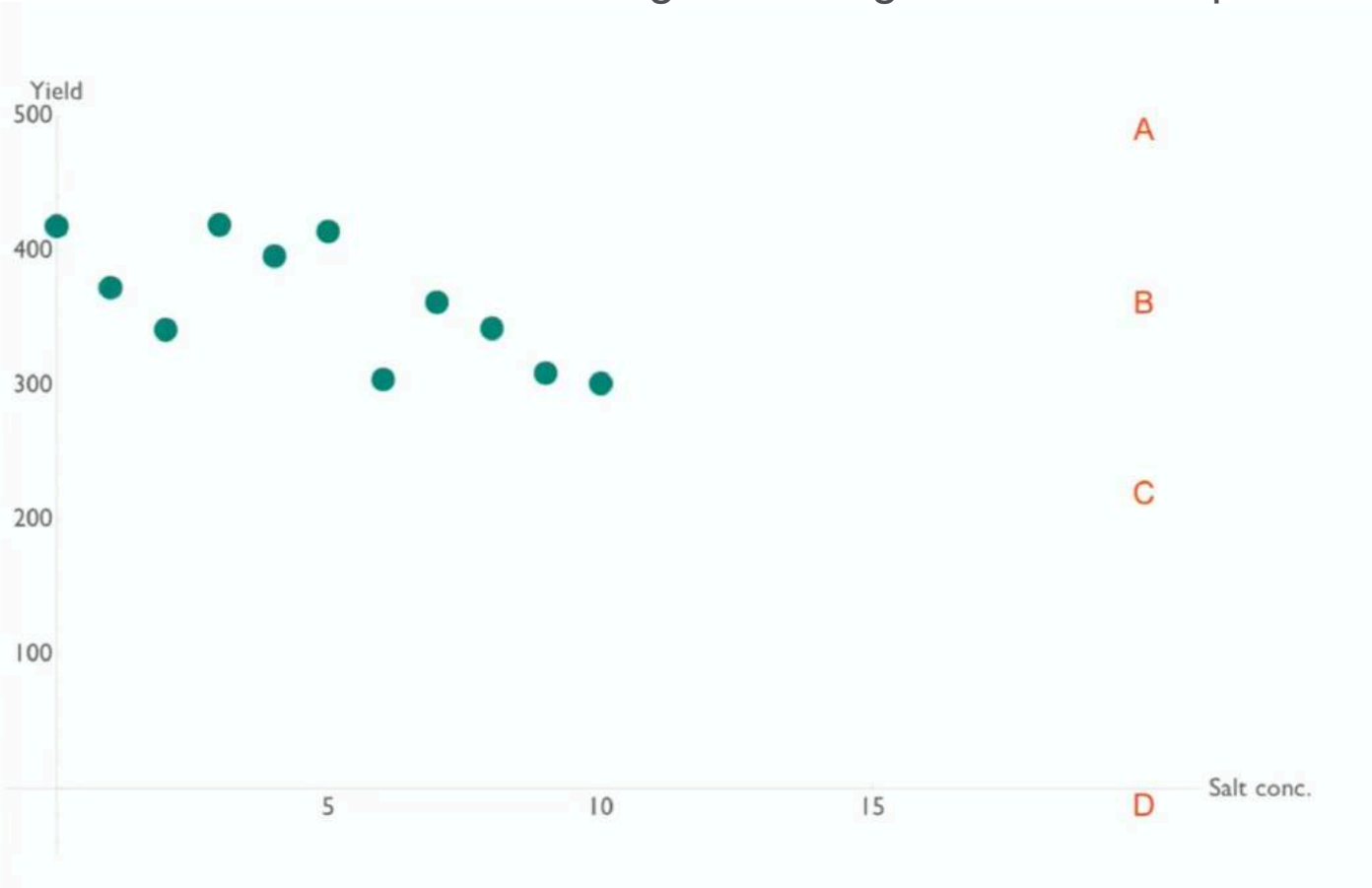
https://xkcd.com/552/

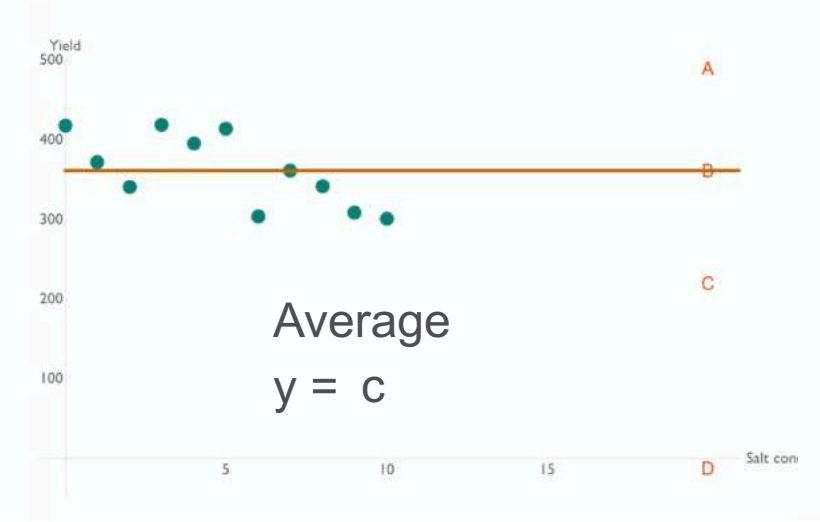
Civil Service Learning

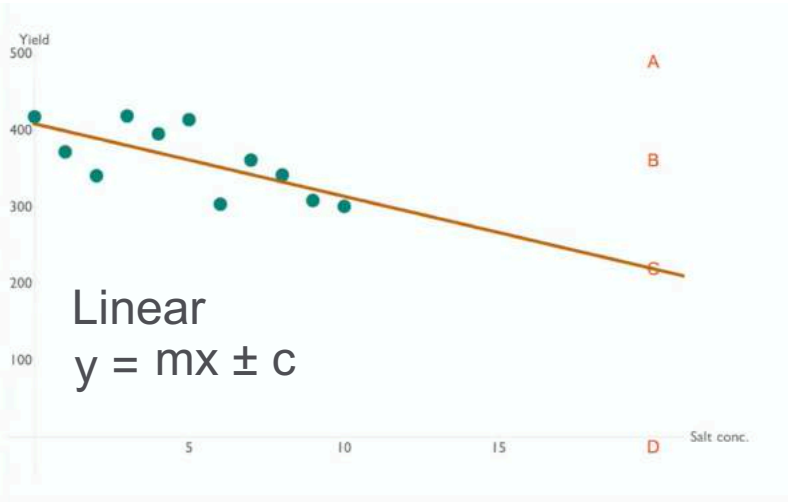
Overfitting

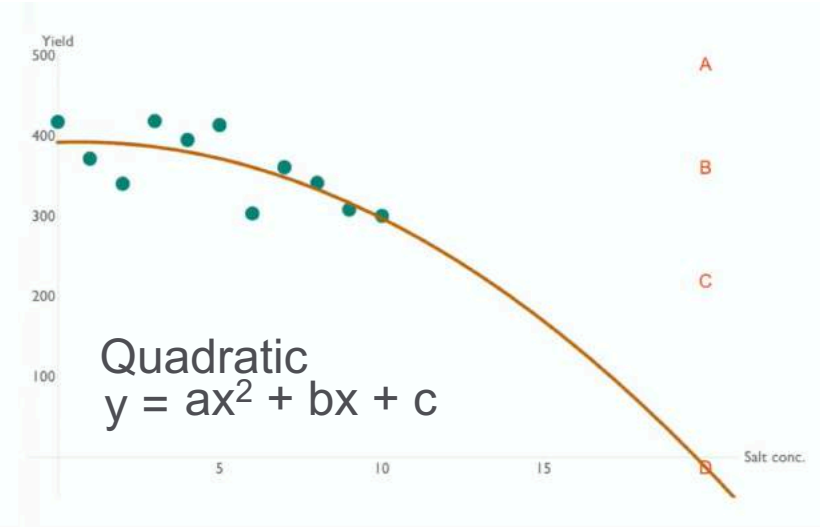
Plotted here are 10 points in a series that have been generated using a function.

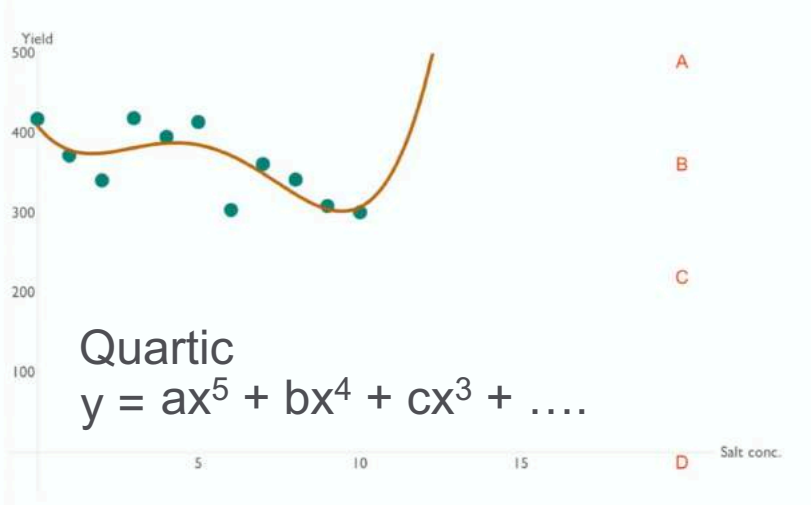
If another 10 points were plotted of the same function which point (A,B,C or D) would they tend towards. Add a line of best fit that goes through one of these points.

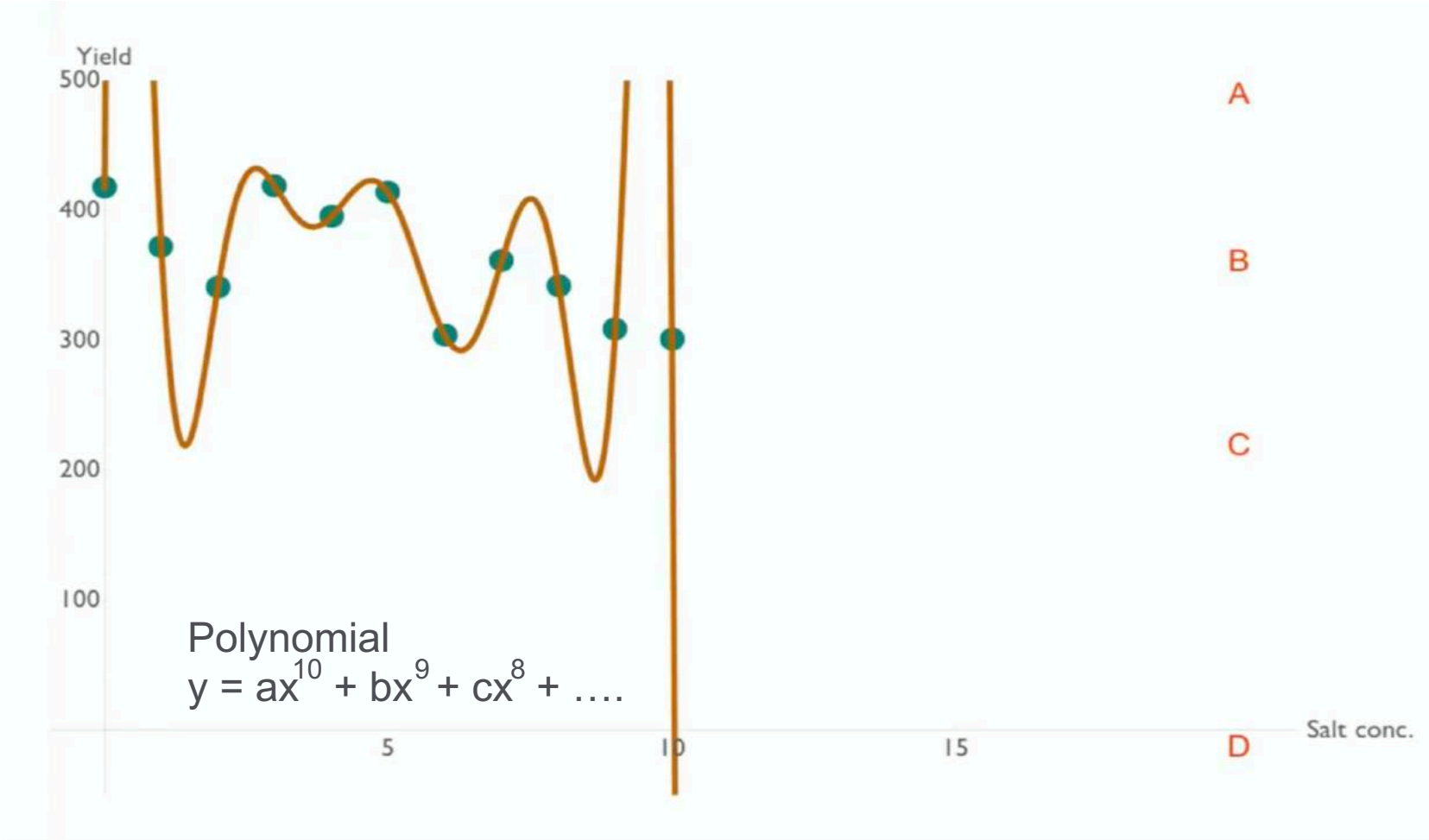



Yield
500
400
300
200
100
A
B
C
D
5
10
15
Salt conc.
Average
y = c


Yield
500
400
300
200
100
A
B
C
D
5
10
15
Salt conc.
Linear
y = mx ± c


Yield
500
400
300
200
100
A
B
C
D
5
10
15
Salt conc.
Quadratic
y = ax2 + bx + c


Yield
500
400
300
200
100
A
B
C
D
5
10
15
Salt conc.
Quartic
y = ax5 + bx4 + cx3 + ….

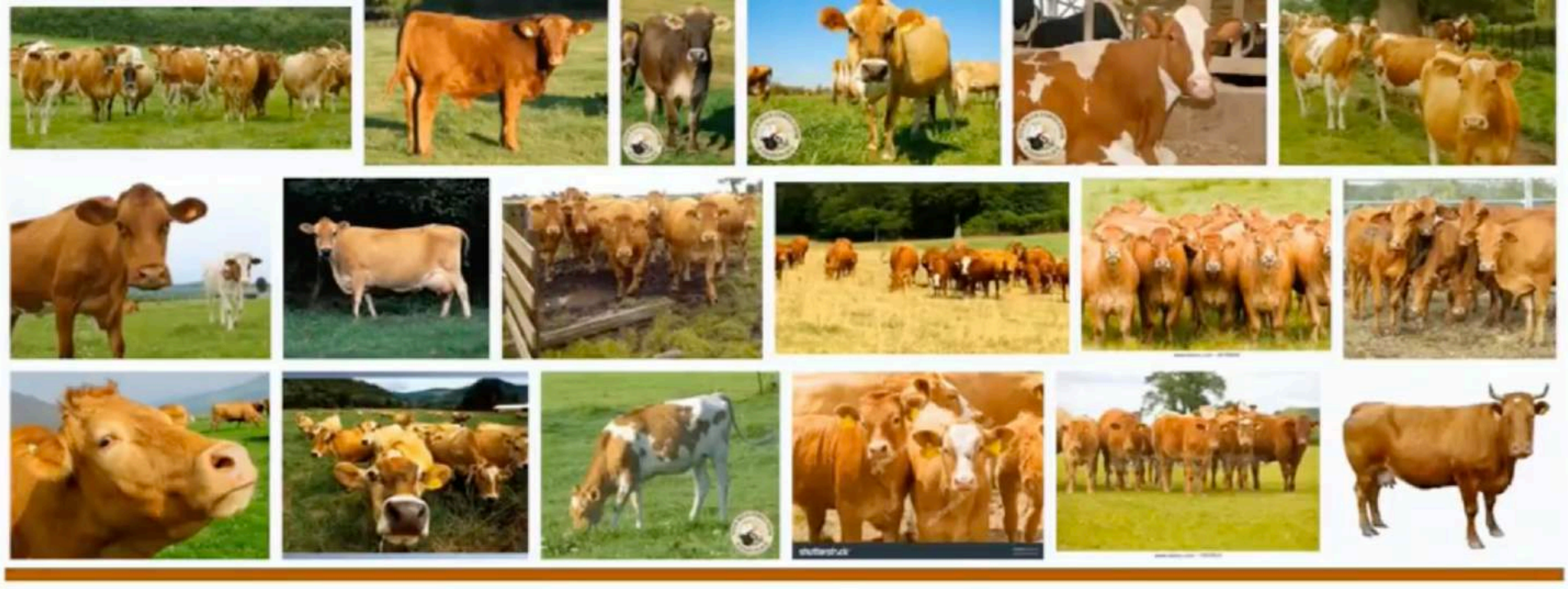

Yield
500
400
300
200
100
5
10
15
Salt conc.
A
B
C
D
Polynomial
y = ax10 + bx9 + cx8 + ….

Why did the flu trends example fail?

- Big Data Hubris (complement not supplement)
- Huge veracity 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




New Glarus
Brewing Co.

Civil Service
Learning

Data analysis 101

You each have a set of cards containing numerical values representing something. Can you work out what the dataset represents?

Data analysis stages:

Data: What do you see?

Information: What can you calculate?

Knowledge: What does it mean to you?

Civil Service
Learning

Averages

- 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 mean?

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: ~125k

16% of 25 = 4

16% trimmed mean: 128,109.09

=trimmean(RANGE,0.16)

Civil Service
Learning

The distribution of data

Divide the data into 6 evenly distributed buckets of values and stack the cards as per the diagram here.

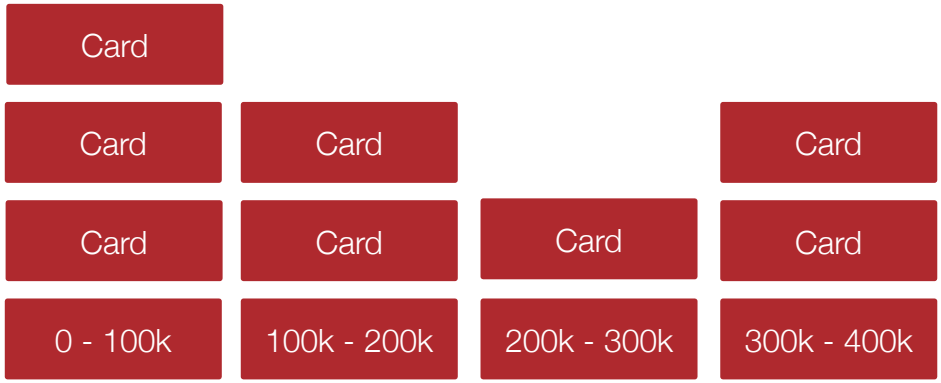
Sugegsted buckets:

0 – 100,000

100,000 – 200,000

200,000 - 300,000

…

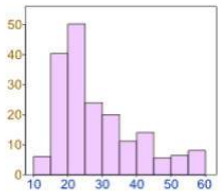


Civil Service
Learning

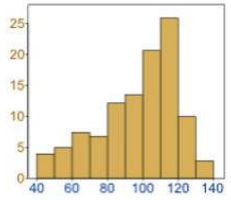
The distribution of data

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.

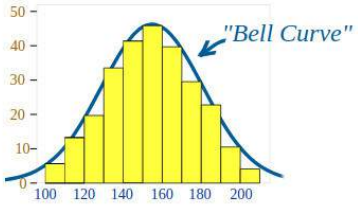
The examples here show different types of distributions of data.



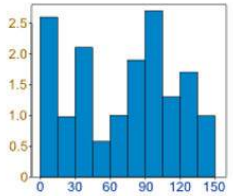
Positively skewed distribution



Negatively skewed distribution



Normal distribution



Non-normal distribution

Civil Service
Learning

5 number summary (of the trimmed dataset)

Plot 3 more lines

1) Median
2) Value ¼ through dataset
3) Value ¾ through dataset

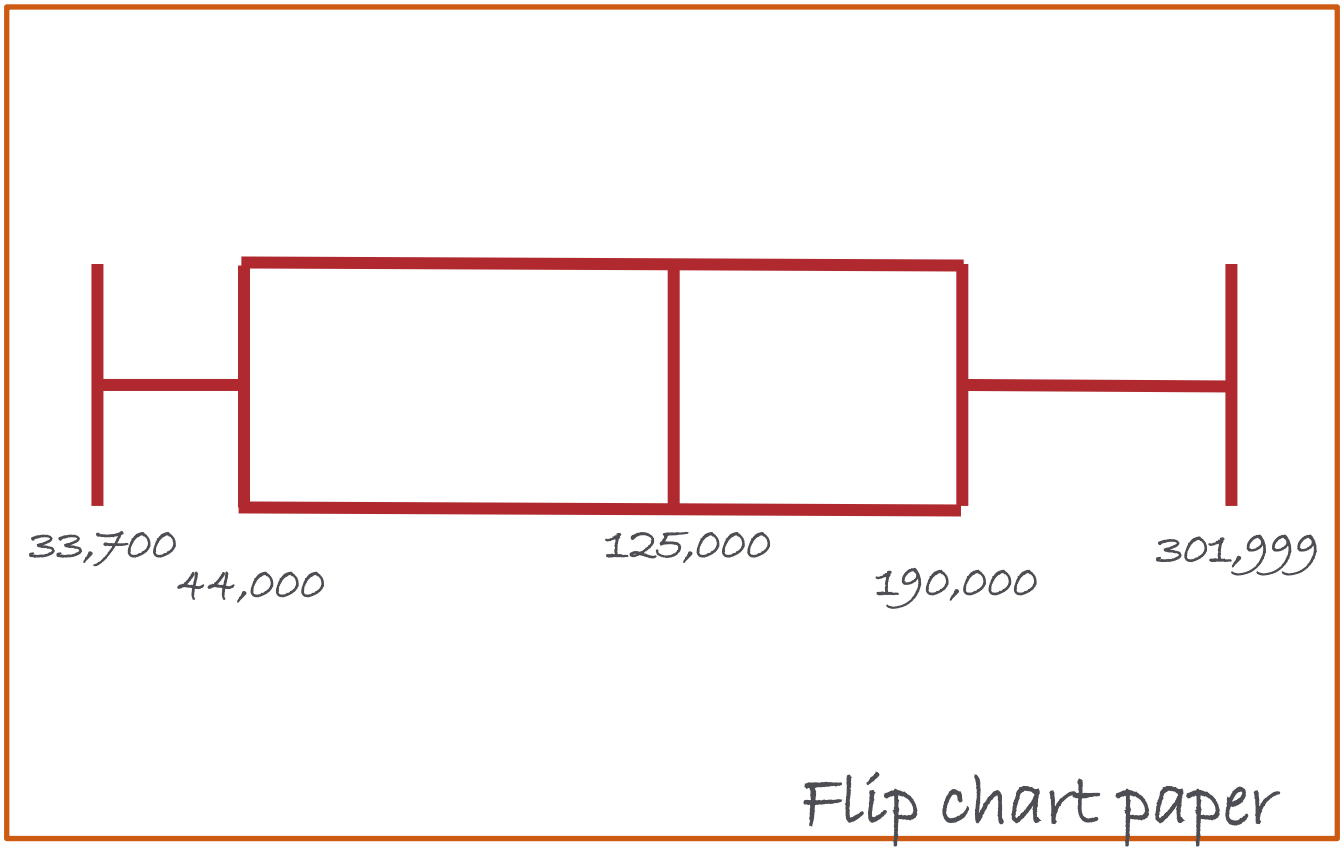
33,700

301,999

Flip chart paper

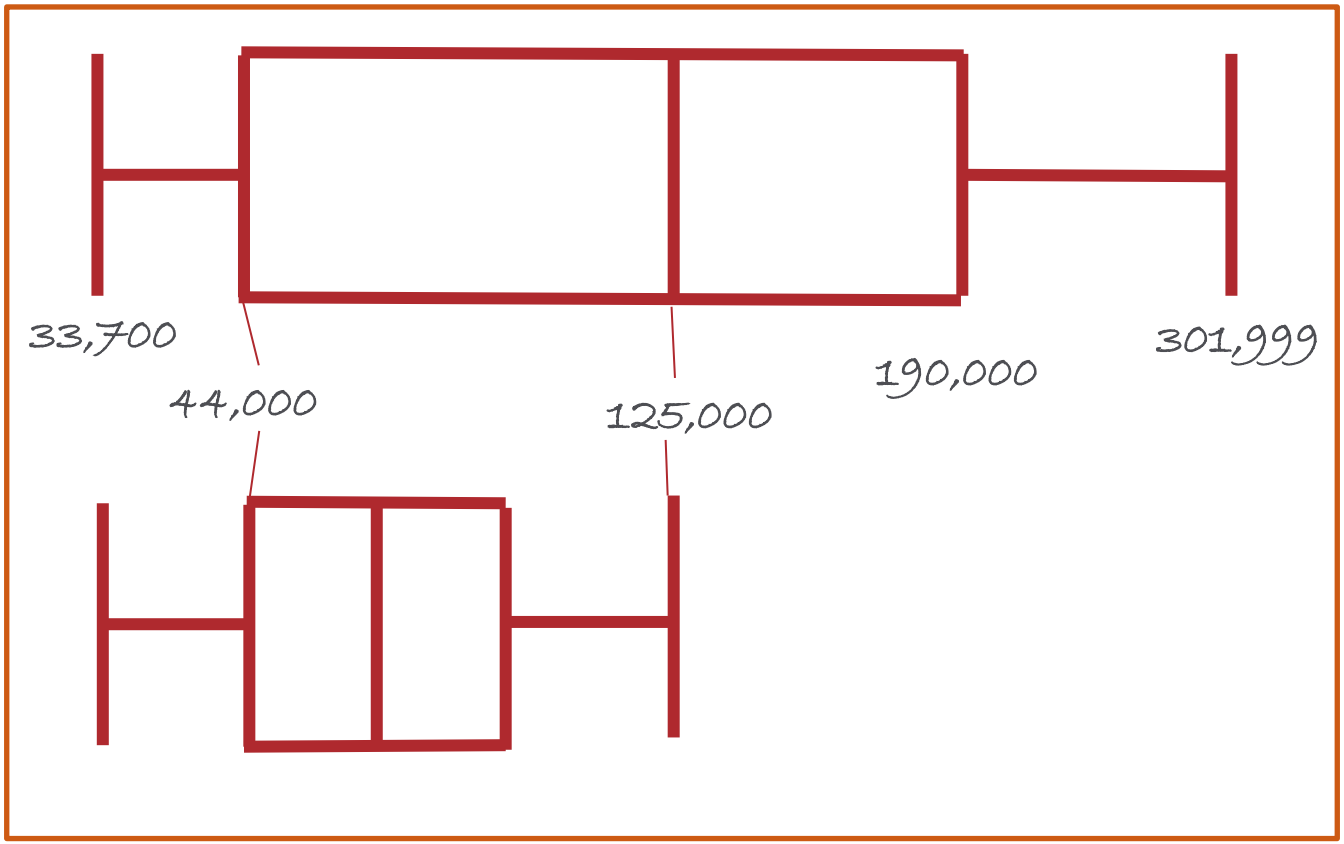
Civil Service
Learning

5 number summary (of the trimmed dataset)



Civil Service
Learning

5 number summary (of the trimmed dataset)

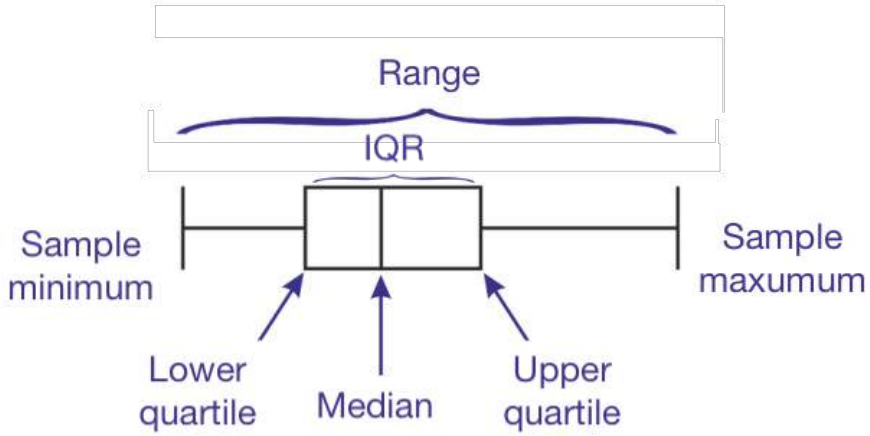


Civil Service
Learning

Five-number summary

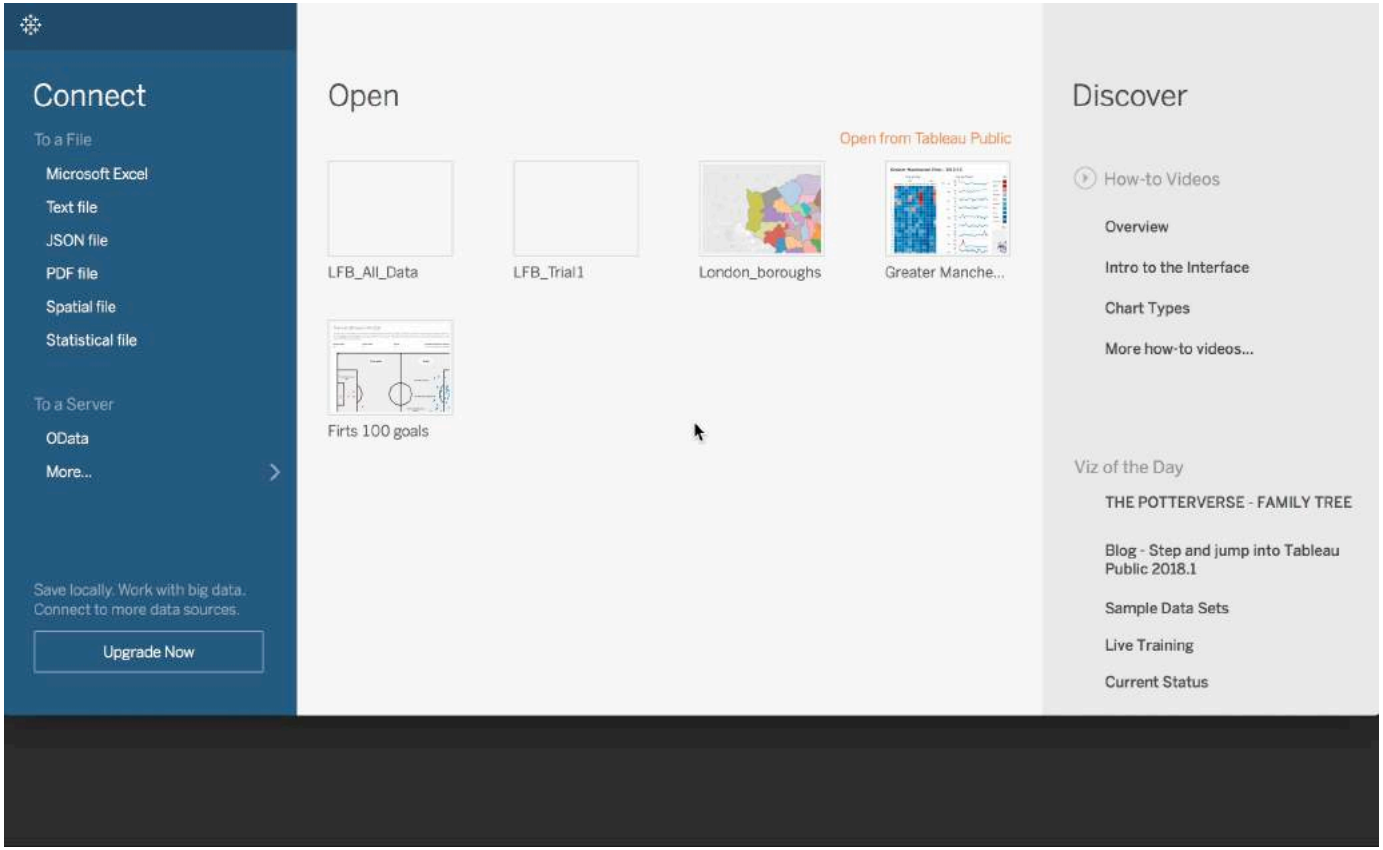
The five-number summary

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)



Civil Service
Learning

Five-number summary in Tableau

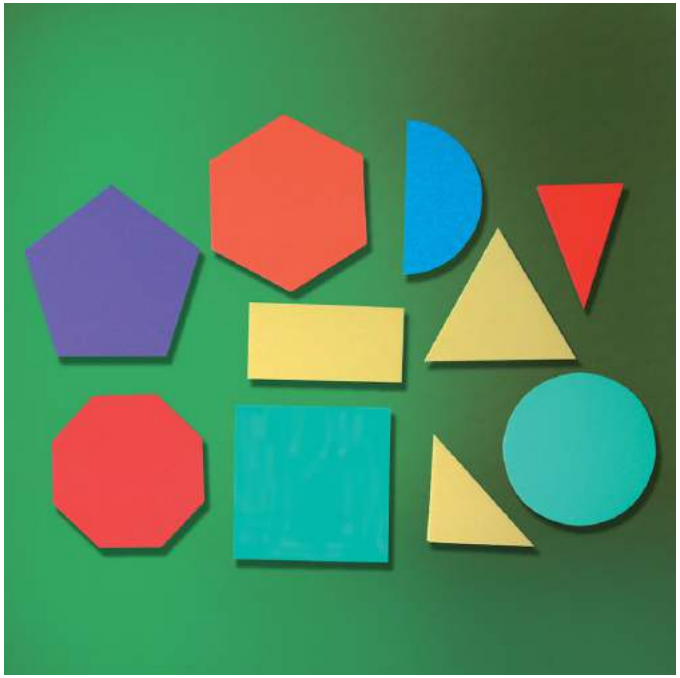


Civil Service
Learning

Summary

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.



Civil Service
Learning

Machine learning and prediction

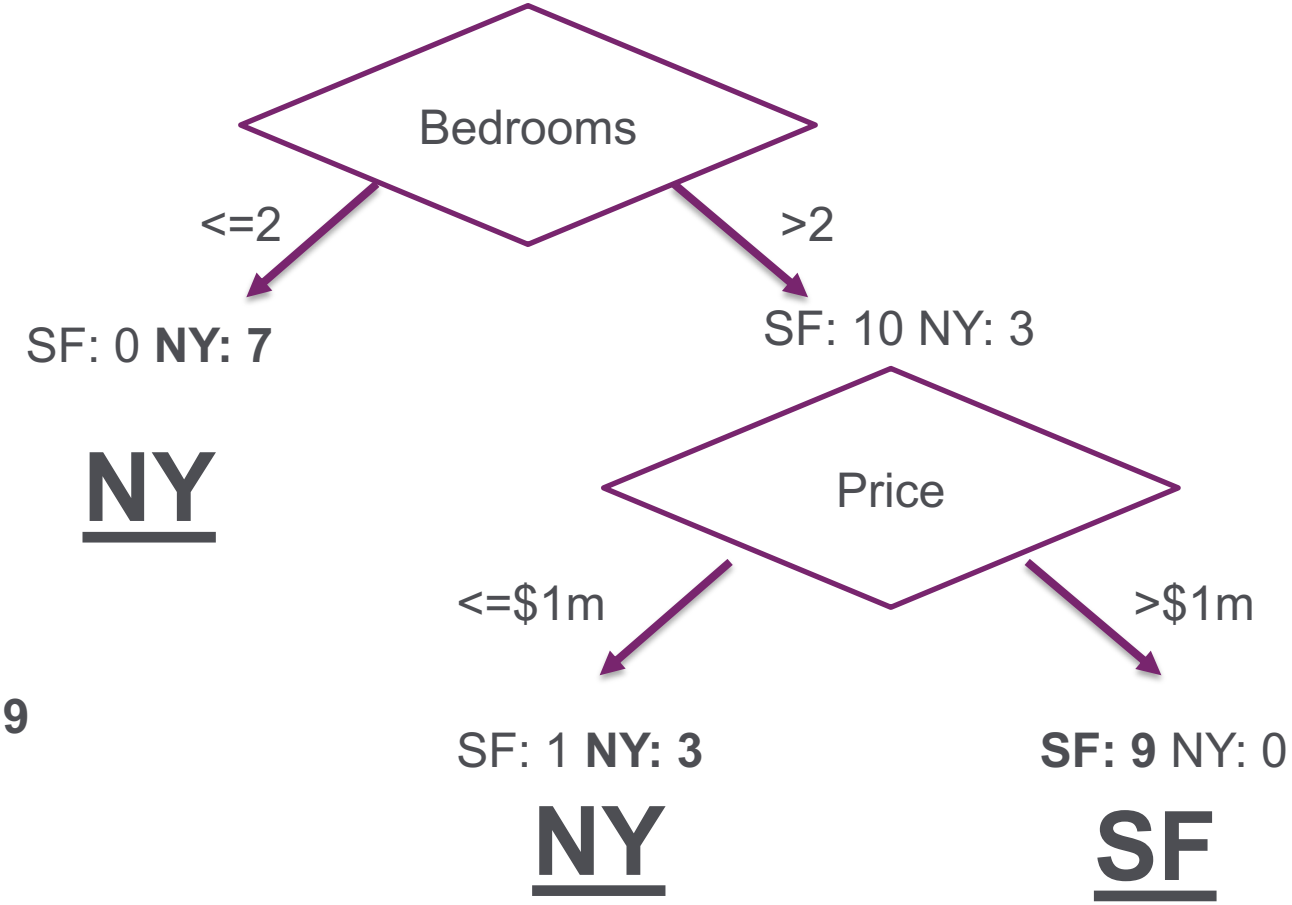
Each table has a set of “Top Trump” cards relating to properties in two cities.

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

Civil Service
Learning

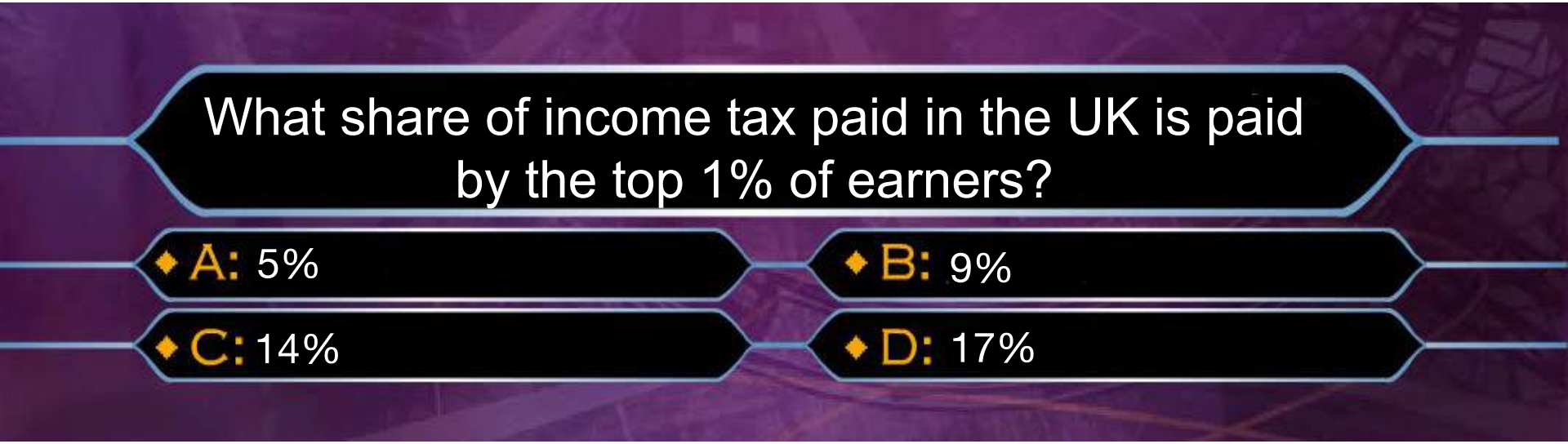
Approaches

Data first

Knowledge first

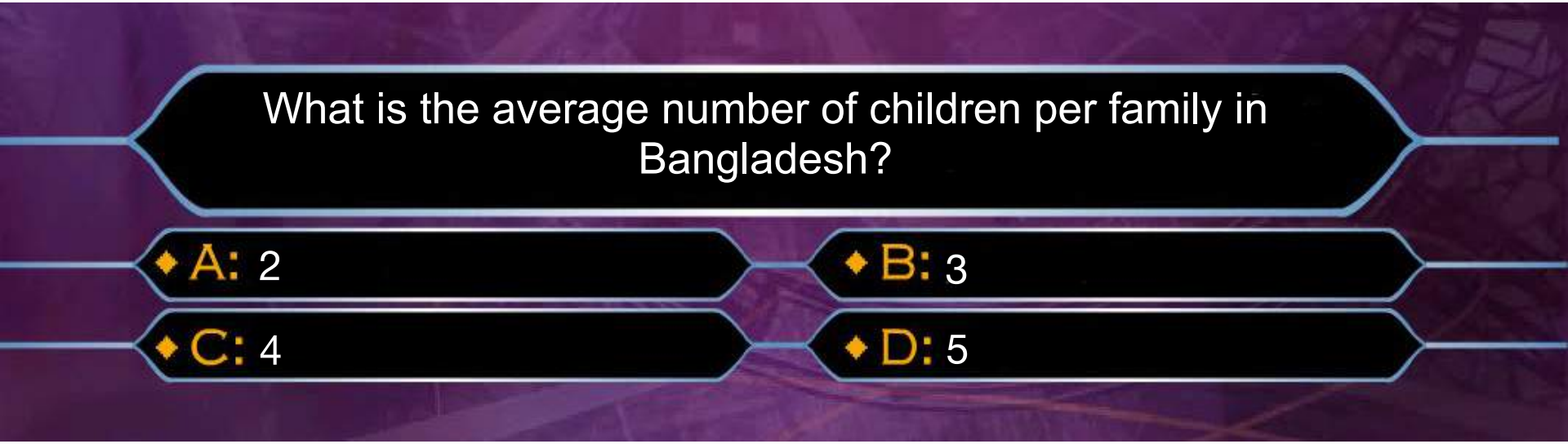
Civil Service Learning

Assumption

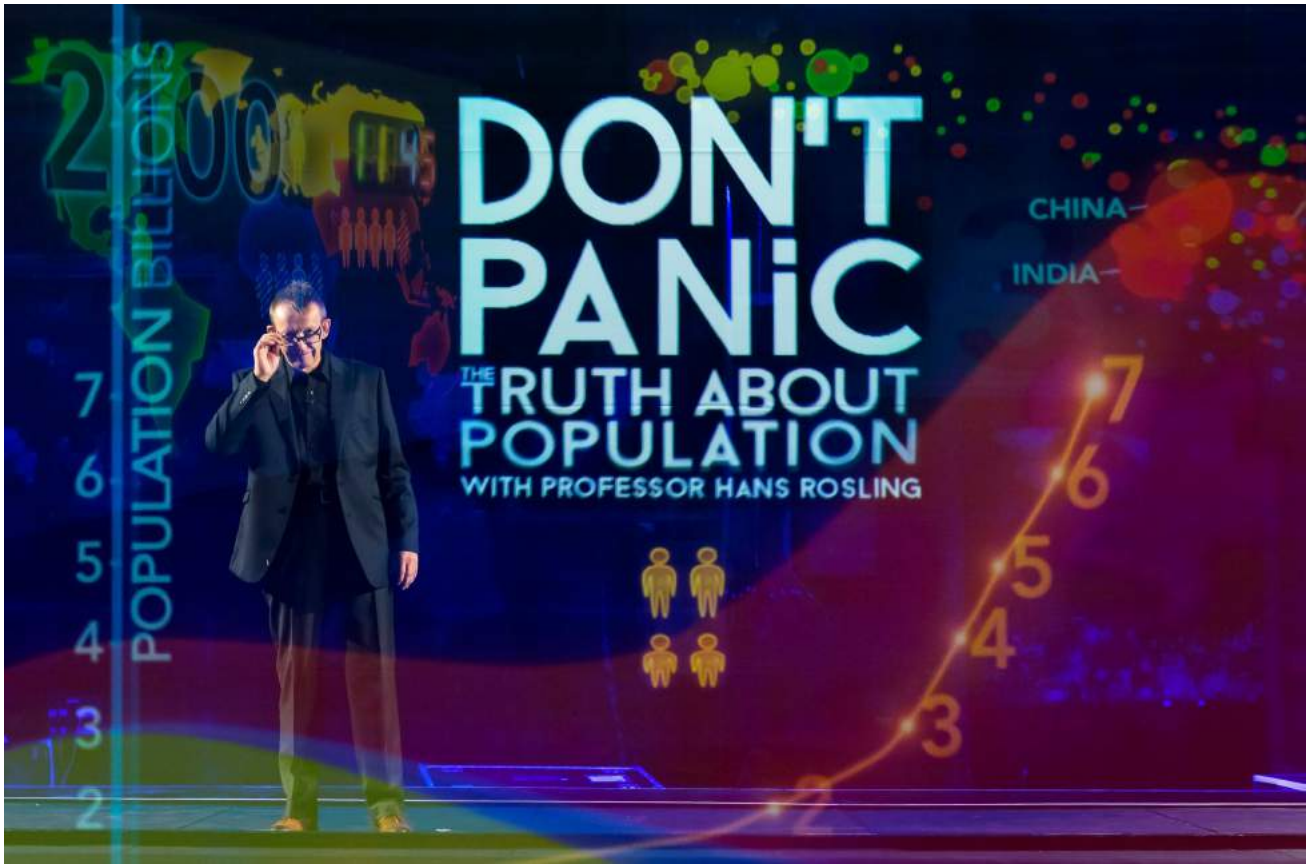


Civil Service Learning

Assumption



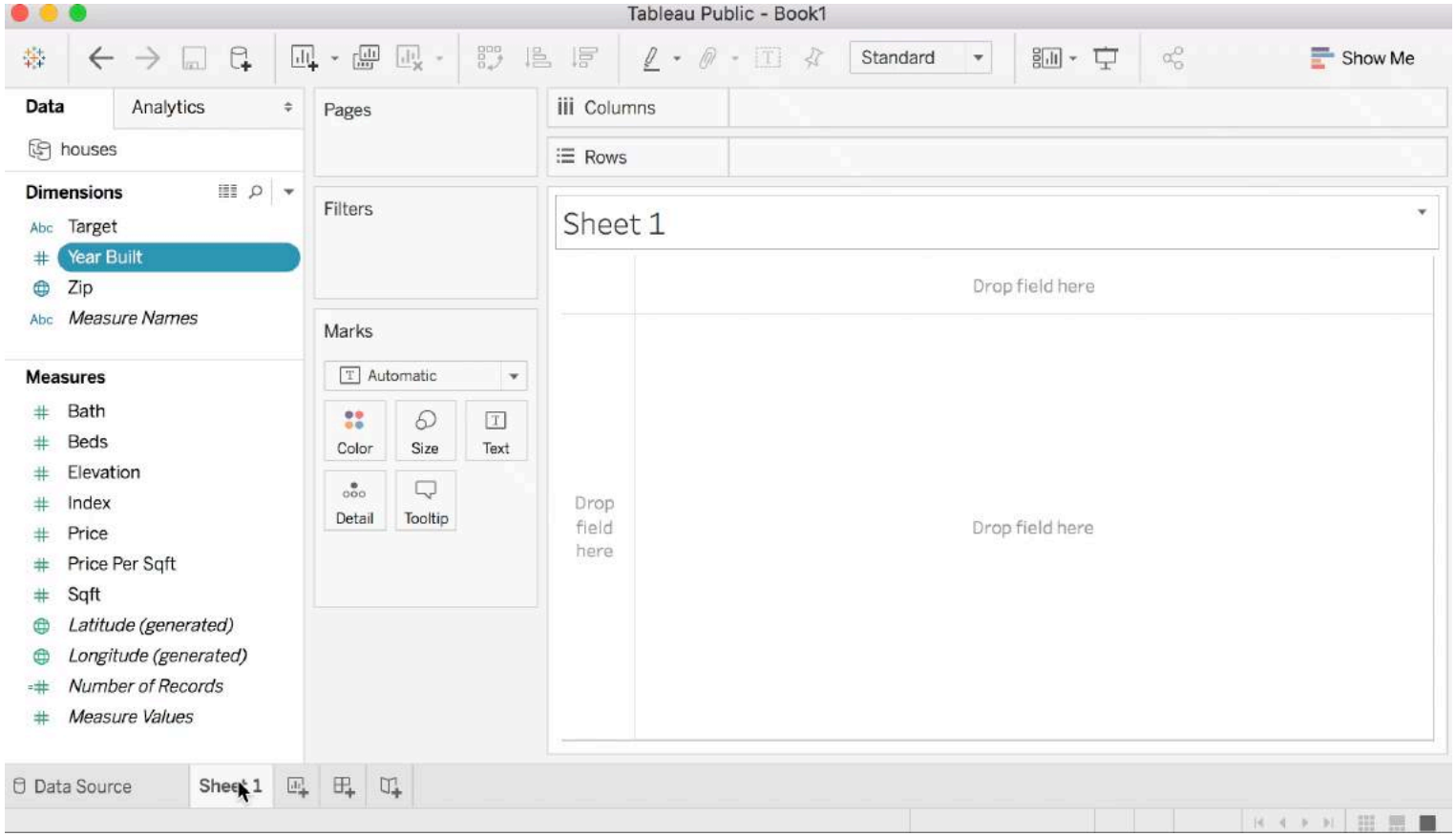
Civil Service
Learning



gapminder.org

Civil Service
Learning

Data analysis of 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?

Civil Service
Learning

A visual introduction to machine learning

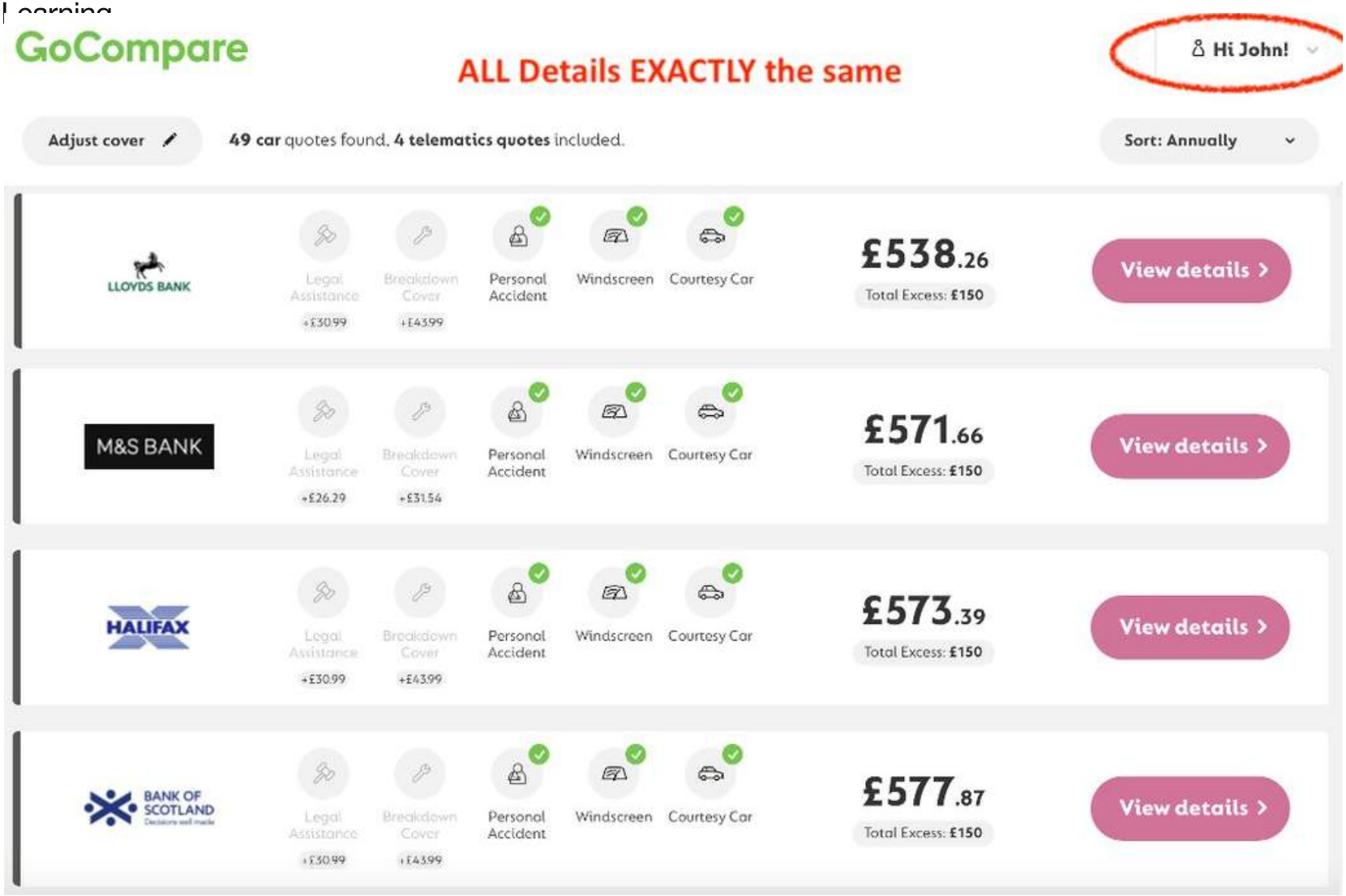
r2d3.us

Considerations

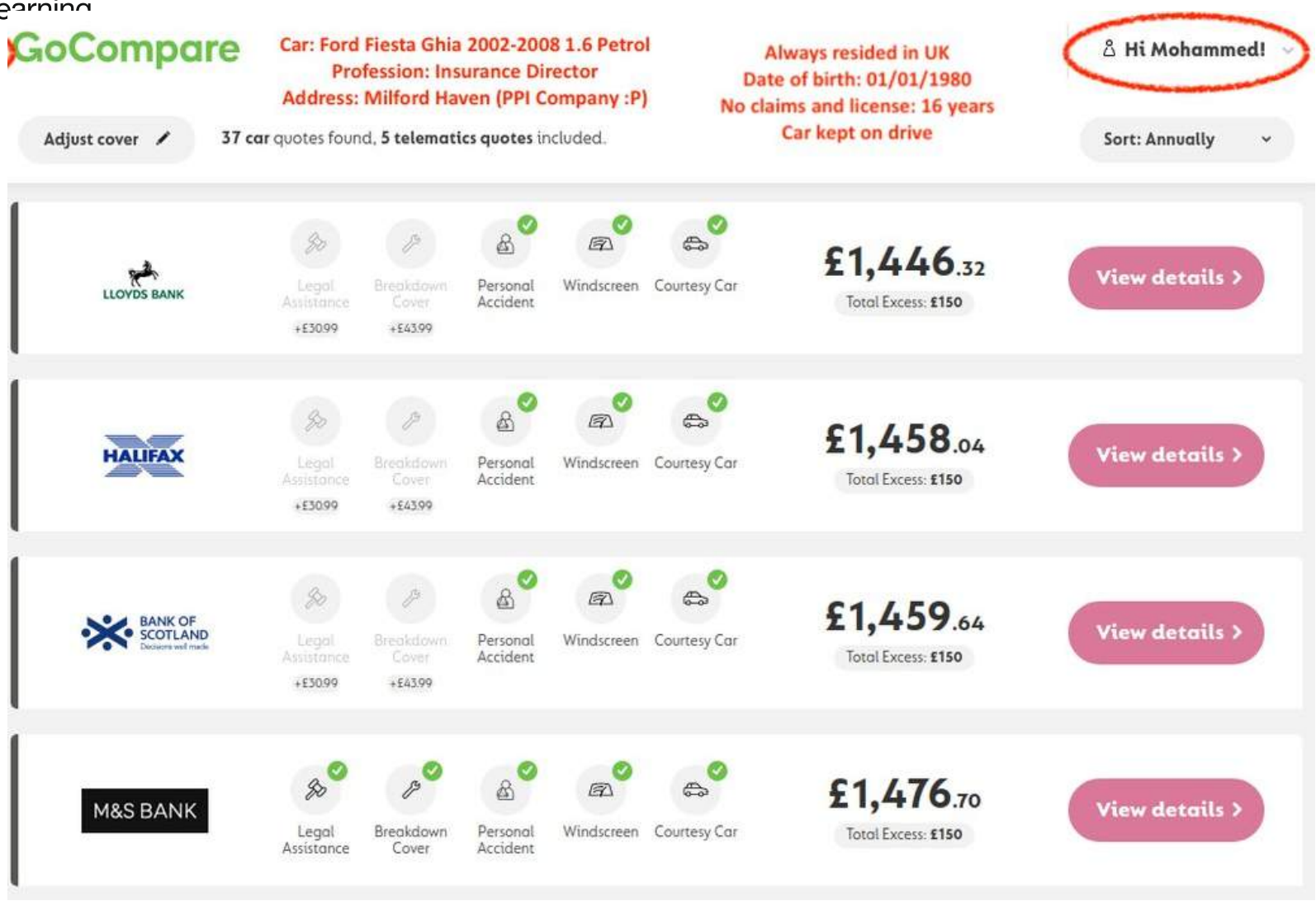
Combine the data with existing knowledge but make sure the existing knowledge is correct!

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

Civil Service
Learning



Civil Service Learning



Civil Service Learning

Good or bad ideas?

1. a tool that analyses the sentiment of a user's tweets, assesses whether they are suicidal and alerts friends	4. a risk-assessment tool that uses AI to advise on prison sentences based upon criminal profile analysis
2. automatic pricing algorithm for taxi firm which responds to surges in demand	5. using energy efficiency data and winter fuel allowance data to target efficiency advice
3. using performance data to advise on how to save money in the emergency services	6. publishing genomes of 100,000 individuals for use in public health

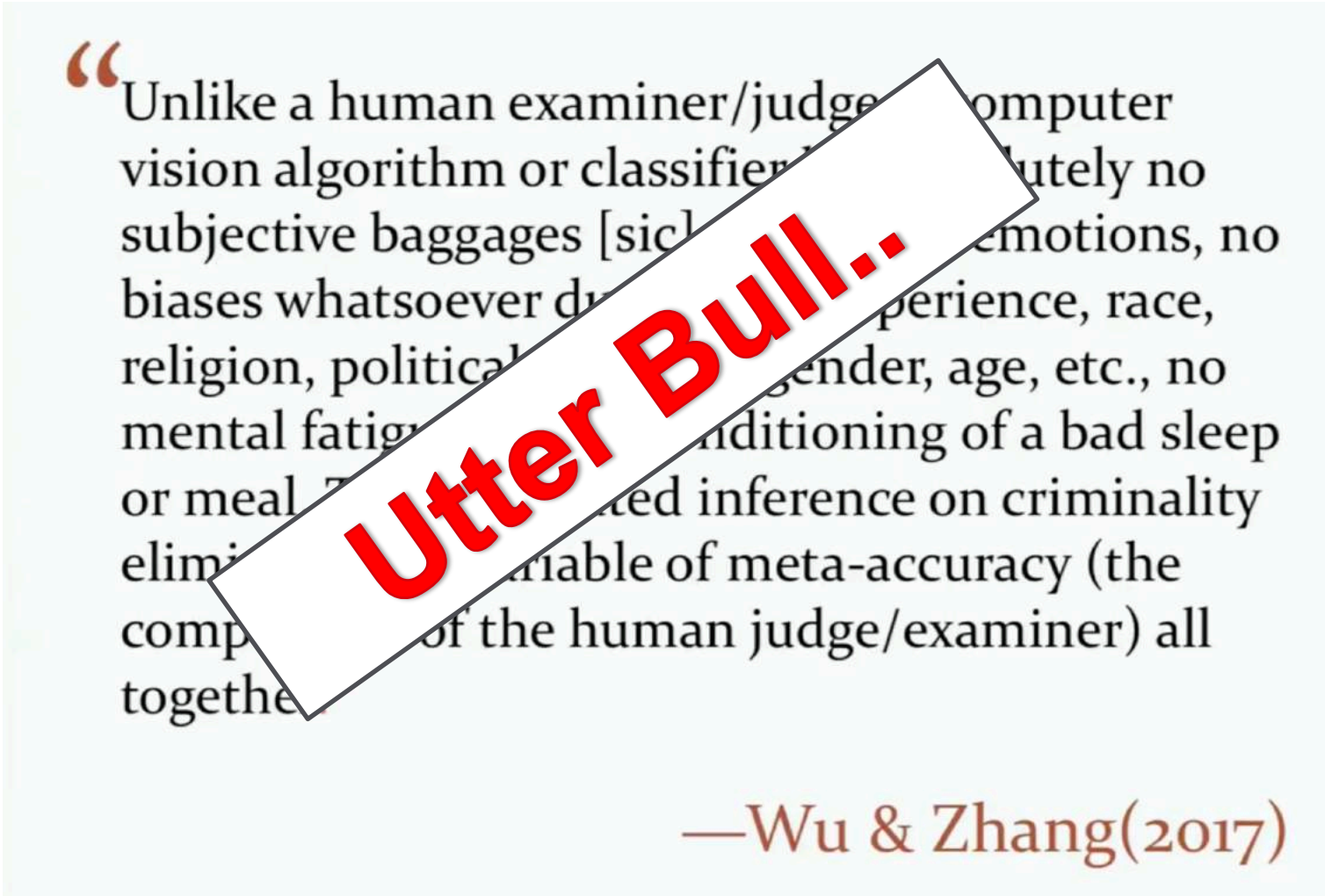
Automated Inference on Criminality using Face Images

Xiaolin Wu
McMaster University
Shanghai Jiao Tong University
xwu510@gmail.com

Xi Zhang
Shanghai Jiao Tong University
zhangxi_19930818@sjtu.edu.cn

"Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages [sic], having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together."

—Wu & Zhang(2017)



Example from "calling bullshit" lecture series from the University of Washington

Civil Service Learning

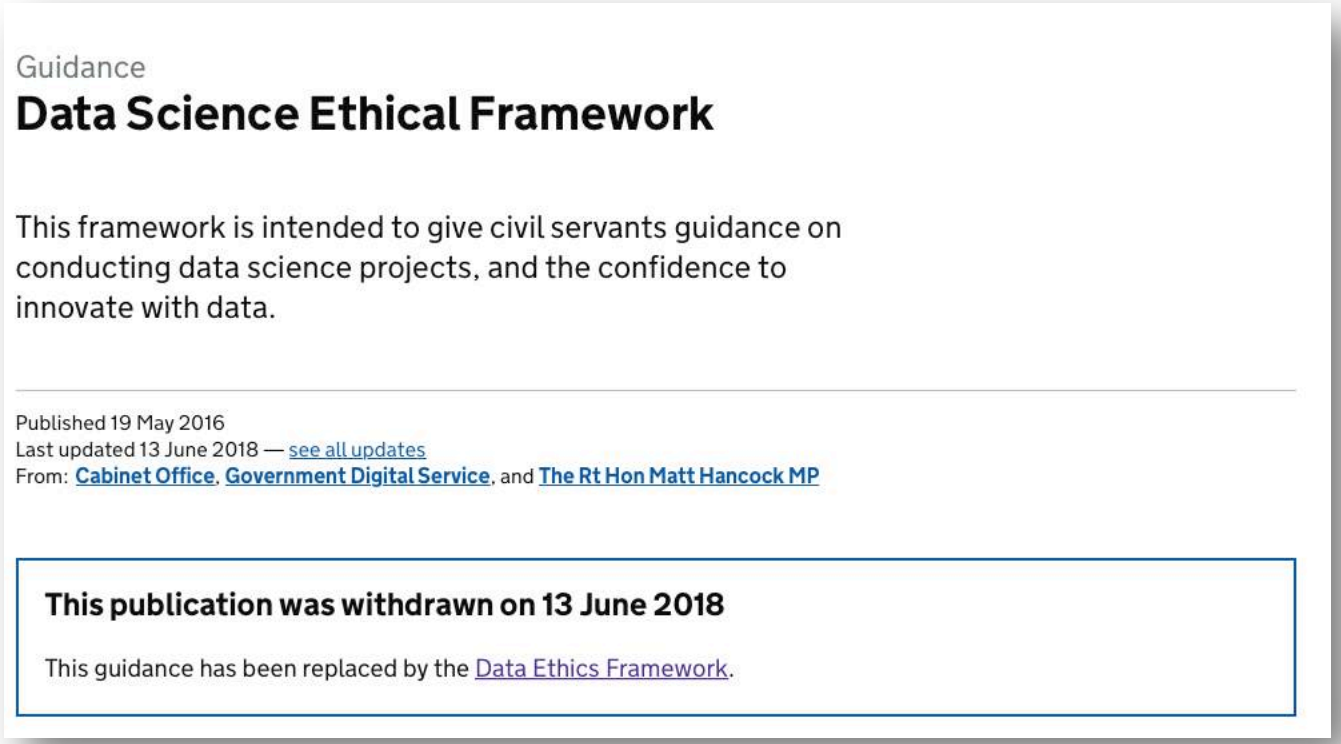
Good or bad ideas?

1. a tool that analyses the sentiment of a user's tweets, assesses whether they are suicidal and alerts friends	4. a risk-assessment tool that uses AI to advice on prison sentences based upon criminal profile analysis
2. automatic pricing algorithm for taxi firm which responds to surges in demand	5. using energy efficiency data and winter fuel allowance data to target efficiency advice
3. using performance data to advice on how to save money in the emergency services	6. publishing genomes of 100,000 individuals for use in public health

Civil Service
Learning

Data ethics

More on the examples used in this section (as well as more examples) -



Guidance

Data Science Ethical Framework

This framework is intended to give civil servants guidance on conducting data science projects, and the confidence to innovate with data.

Published 19 May 2016
Last updated 13 June 2018 — see all updates
From: Cabinet Office, Government Digital Service, and The Rt Hon Matt Hancock MP

This publication was withdrawn on 13 June 2018

This guidance has been replaced by the Data Ethics Framework.

https://www.gov.uk/government/publications/data-science-ethical-framework

Civil Service Learning

Data ethics canvas

https://theodi.org/article/data-ethics-canvas/

Civil Service
Learning

Course materials

http://training.theodi.org/csl-an2/

AN2 - Data and analysis

Welcome. This page accompanies the **AN2 - Data and analysis** course and provides access to slides exercises and related case studies.

The main aim of the course is to *equip policy makers with the knowledge and skills they need to effectively use open data in policy making*.

Key outcomes of the course include:

- Define data, open data and big data
- Describe how data is used to inform policy making
- Identify the benefits of open data to policy making
- Perform some basic statistical analysis on data to identify the shape and trends in data.
- Access the ethical risks of using data in policy making
- Perform a practical piece of policy making with real data

Civil Service Learning

Data analysis in practice

Outcomes

6. List the stages in carrying out data analysis for policy making.
7. Create a plan for carrying out data analysis for policy making.
8. Review the role of open data in policy making

Homework

9. Carry out a simple data analysis using a number of tools.
10. Create an interactive data visualisation.
11. Communicate the results of a data analysis to decision makers.

Civil Service
Learning

What has been the impact of closing 10 fire stations in London?

Civil Service
Learning

Data analysis practical

You have been tasked with saving money in the local fire service.

You must propose a solution that saves money while minimising the impact on service delivery.

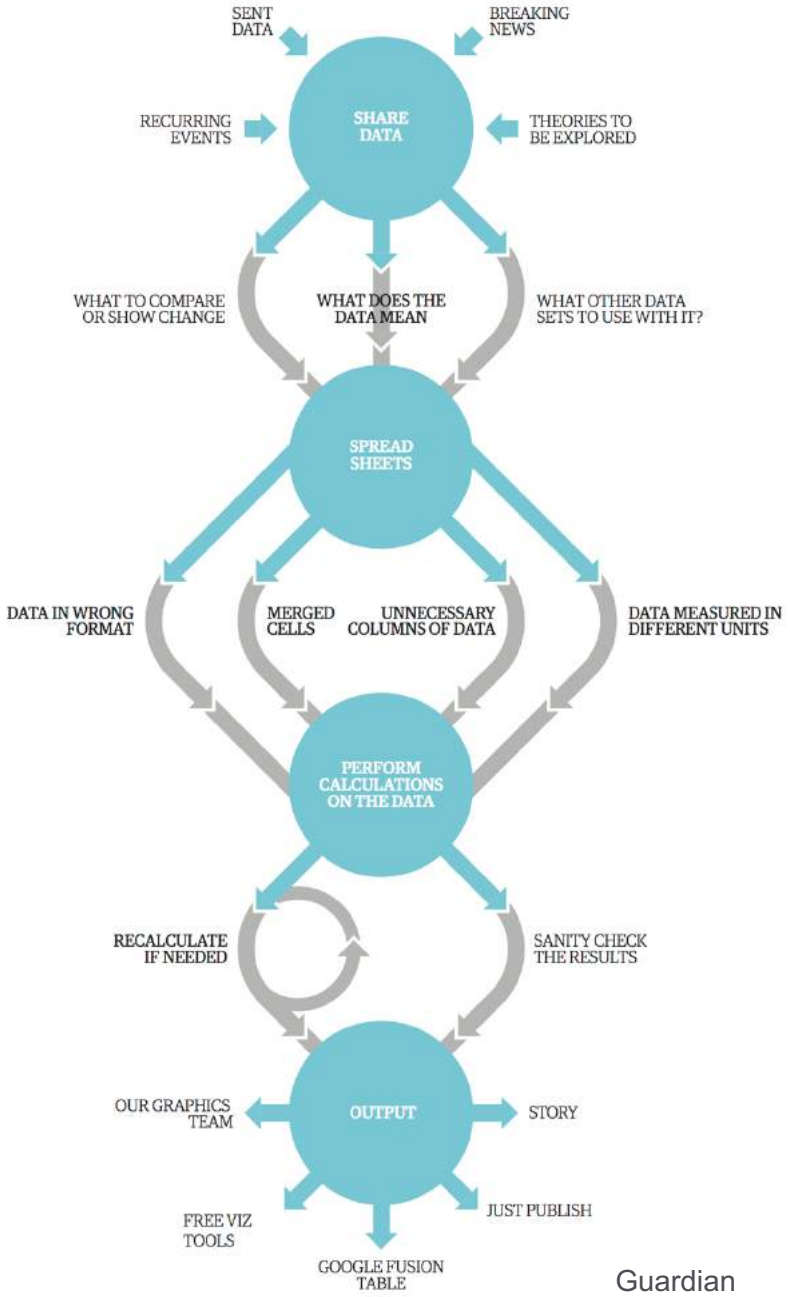
Which factor do you think is the most significant in analysing performance to identify savings (top of the decision tree)

Paul Hudson

Civil Service Learning

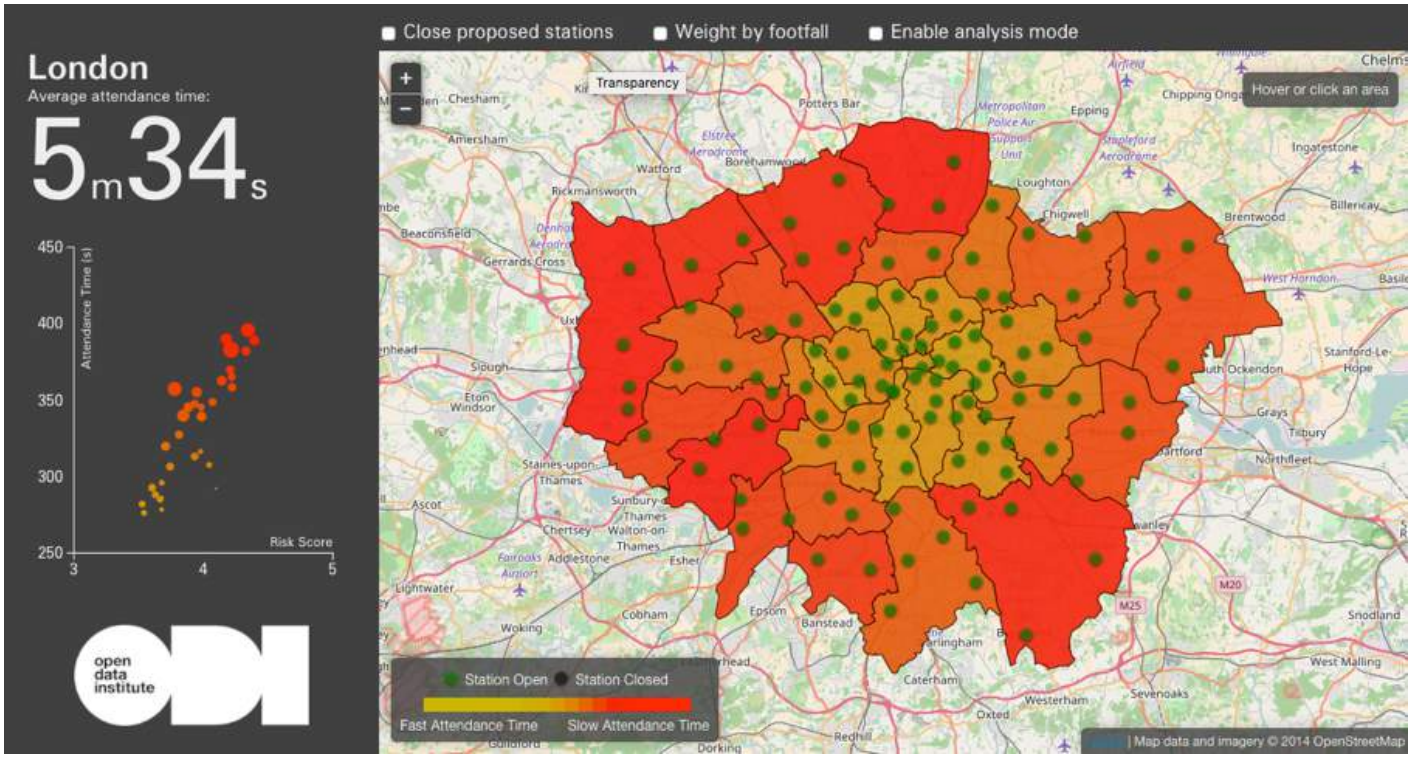
Using data in analysis

1. Collect data.
2. Establish context of data.
3. Clean the data.
4. Translate, filter and merge data.
5. Initial evaluation.
6. Go back to start?
7. Secondary evaluation and analysis.
8. Sanity check.
9. Create output.



Guardian

http://london-fire.labs.theodi.org/explore/



Civil Service
Learning

Stage 1: collect data

Make a space on your group desk

Put a yellow post-it note in the middle and draw a fire on it. Write a time on it between 4 minutes and 7 minutes.

Scatter a series of all different colour (including yellow) post-it around this one to keep it in the middle. Write a time between 4 and 7 minutes on each one.

Stage 2: establish context

Each post-it note represents an incident that has been responded to by the London fire brigade.

Each colour represents a different fire stations from which the appliance (fire engine) was sent

Stage 3: clean the data

Given the context does your data make sense?

Stage 4: Translate, filter and merge data

Nothing to do…

Civil Service
Learning

Stage 5: initial evaluation

How do we calculate the impact of closing the YELLOW fire station (and specifically the middle incident)?

What is your design?

Stage 8: Sanity check

Does you design overfit (or use nth degree polynomial?)

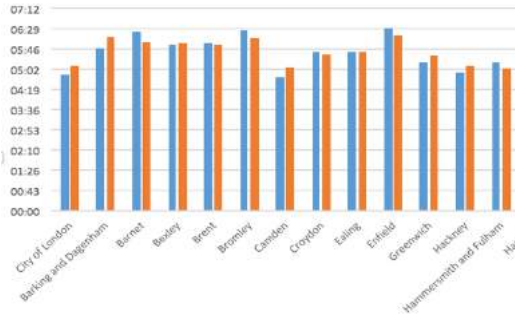
Can your design really model the future or just show how the past was?

Civil Service Learning

Stage 9: Creating output

There are many options for output. Consider the best one for your purpose.

- A map can be ideal for such geographic datasets (search for London Data Store borough Excel KML).
- A chart showing a comparison of before an after statistics can be easily generated in excel.
- An highly interactive dashboard can be created with dataseedapp.com.



Civil Service
Learning

Review

London Fire Stations analysis was a story started by a participant in an ODI Training course like this one. What is your story? Who and what would you need to help you?

If I'm a policy maker, what do I now need to go and do?

What do I need to go and do next?

What is your main takeaway from the course overall?

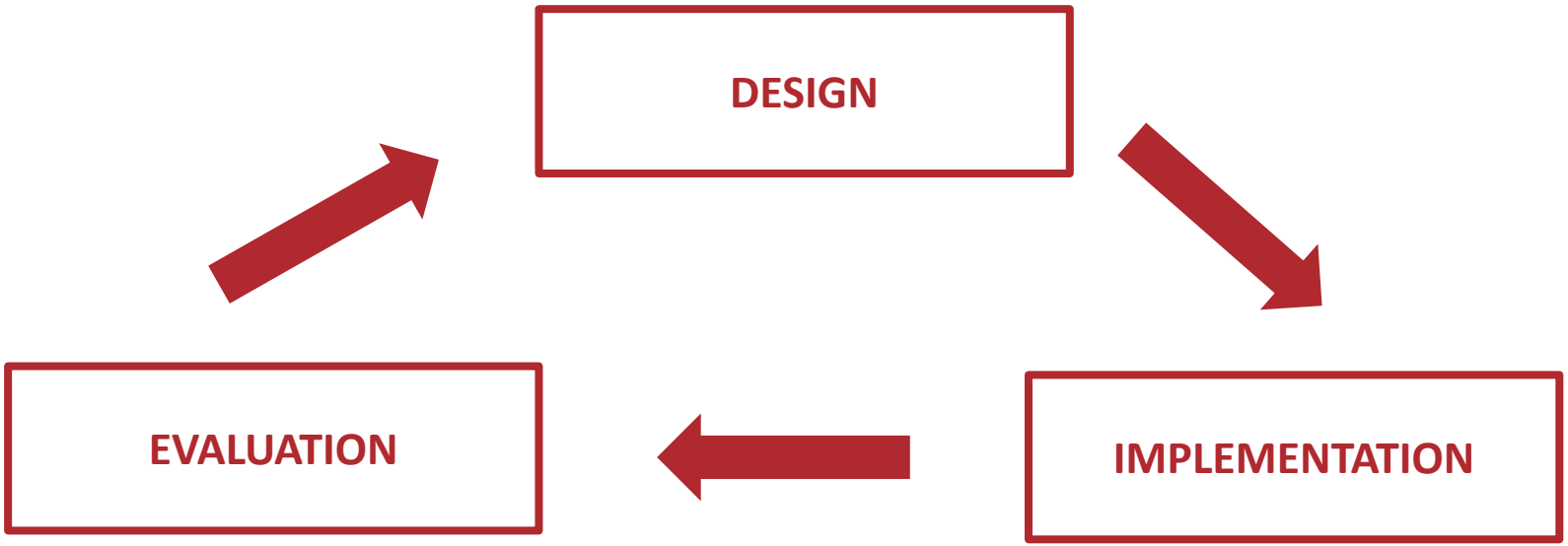
What are you going to apply back in the workplace?

Civil Service
Learning

Open data in policy cycles

Civil Service Learning

Public policy cycle

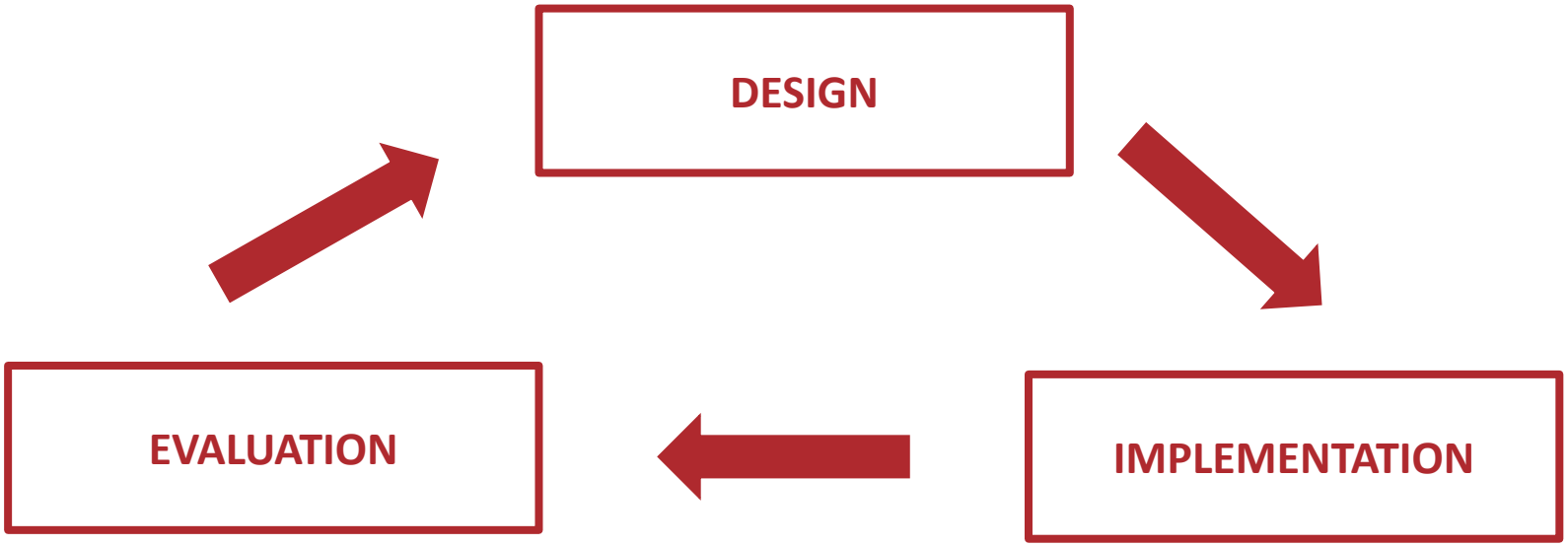


Slide courtesy of Andrea Barenque

Civil Service
Learning

Question

What role can open data play in each stage of the policy cycle?



Civil Service Learning

The city of Xalapa's waste management problem

Rubbish collection was frustrating the public as there was no clarity how the service operated.

The main problems:

- Where is the rubbish collected from?
- When is the rubbish collection?
- What day is the rubbish collected?
- What time is the rubbish collected?

Create a policy cycle/s

What role can open data play?

Slide courtesy of Andrea Barenque

Civil Service Learning

1. Put a GPS on every garbage truck (design)
2. Collected the data (implementation)
3. Opened the data (implementation)
4. Engaged community in design (design)
5. Analyzed the maps for new routes (evaluation & design)
6. Implemented new routes (implementation)
7. Evaluated the impact, repeating 1-3 (evaluation)
8. Civil society organizations organized a Hackathon (design)
9. Developed an app (exact location, timetables, citizens reports) (implementation)
10. Evaluated the solution (evaluation)

Slide courtesy of Andrea Barenque

1. Put a GPS on every garbage truck (design)
2. Collected the data (implementation)
3. Opened the data (implementation)
4. Engaged community in design (design)
5. Analyzed the maps for new routes (evaluation & design)
6. Implemented new routes (implementation)
7. Evaluated the impact, repeating 1-3 (evaluation)
8. Civil society organizations organized a Hackathon (design)
9. Developed an app (exact location, timetables, citizens reports) (implementation)
10. Evaluated the solution (evaluation)

Is open data just an input in this cycle?

If not what does this change?

Slide courtesy of Andrea Barenque

Civil Service Learning

1. Input data for evidence based policy making

Data is drawn from a number of sources and is analysed to inform policy making.

Examples include:

- environmental impact of third runway at Heathrow
- impact of London fire station closures
- how to regulate peer to peer lending

flickr: lucianf

Civil Service
Learning

2. Output data for transparency and to encourage citizen interaction

Data is published as part of a transparency or other open government agenda. There is no immediate desire for the data to have any other impact.

Examples include:

- government spending data
- planning application data
- LIDAR data

Civil Service Learning

3. Tool to change behaviour

Where the data is the catalyst for change required by the policy.

Examples include:

- plastic bag usage data
- waste and emissions data
- pay gap data
- mobile coverage data
- broadband speed data

US Fish and Wildlife Service

Civil Service Learning

Thank you

We hope you enjoyed this experience brought to you by

Civil Service Learning

Delivered by

open data institute

Now over to you! What are you going to do differently?