



Civil Service  
Learning

# Open data driven policy making

## AN2: Data and analysis

# Welcome

Commissioned by



Civil Service  
Learning

Delivered by



Dr David Tarrant

The Open Data Institute

Lucy Knight

Local Government



## 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?



# 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 practical



Civil Service  
Learning

# Data driven policy making



# 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.



# 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.



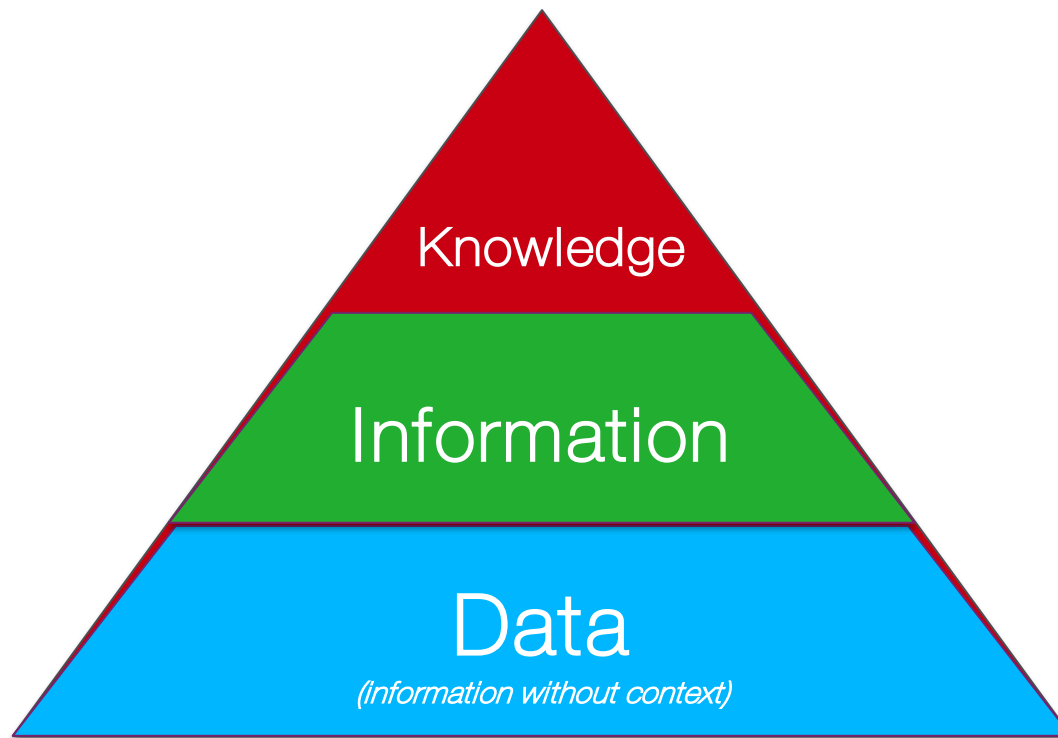


## What is data?

- 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.



# How is data different from information?





# Exercise

## What is open data?

In as few words as possible, what is 'open data'?



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.



Civil Service  
Learning

# Data that anyone can access, use and share.

The Open Data Institute



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.



# Open data

The important points are:


- licensed openly (e.g. Open Government License)
- 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



## Case study: LIDAR

**DATA.GOV.UK** Beta  
Opening up Government


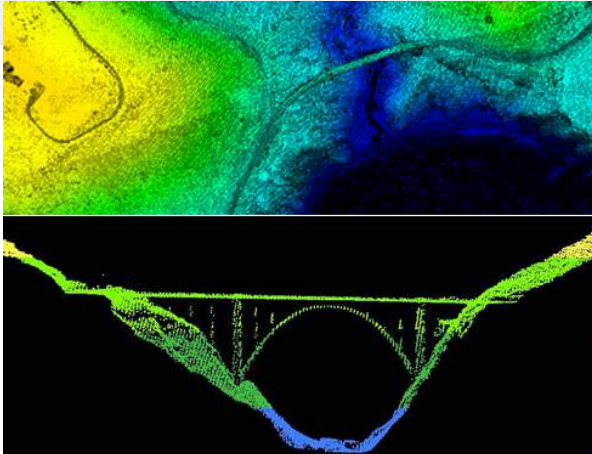
Home

[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: ★★★★★





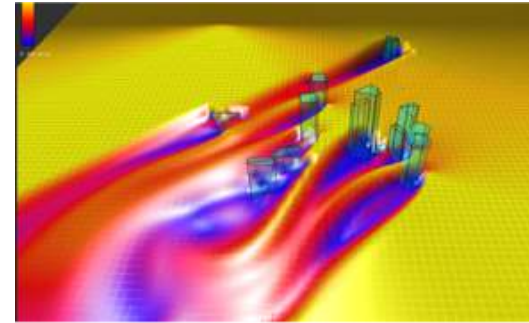


# Impact of LIDAR data

Wind modelling

Archology

Educational games (Minecraft)





Civil Service  
Learning

# Video

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



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.





# 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

<b>Volume</b> How much?	<b>Variety</b> How different?
<b>Veracity</b> How trusted?	<b>Velocity</b> How fast?



# 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<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup>

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<sup>1</sup>. 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



The screenshot shows a Google search interface. The search bar contains the text "fever aching joints headache". Below the search bar, the "All" tab is selected. The search results show "About 369,000 results (0.52 seconds)". The first result is from WebMD, titled "Chills, Fever, Headache and Joint aches: Common Related Medical ...". The snippet describes how these symptoms indicate conditions like Lyme disease, sinusitis, meningitis, or osteoarthritis.

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...](https://symptomchecker.webmd.com/multiple-symptoms?...chills%7Cfever%7Cheadache%7Cjo...)

About this result • Feedback





[Google.org home](#)

[Dengue Trends](#)

**Flu Trends**

[Home](#)

United States ▾

Washington ▾

[Download data](#)

[How does this work?](#)

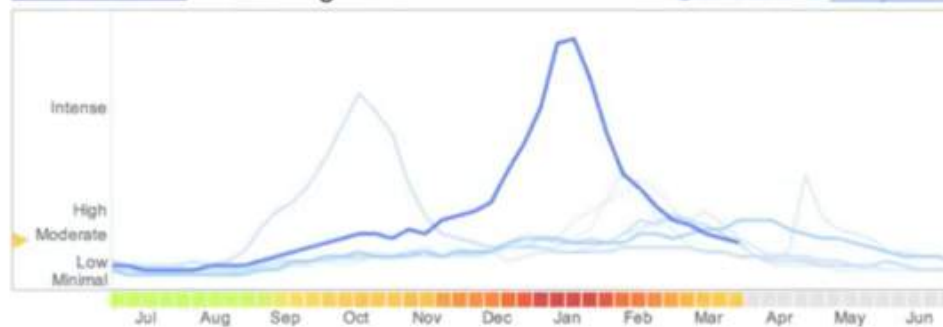
[FAQ](#)

## Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)

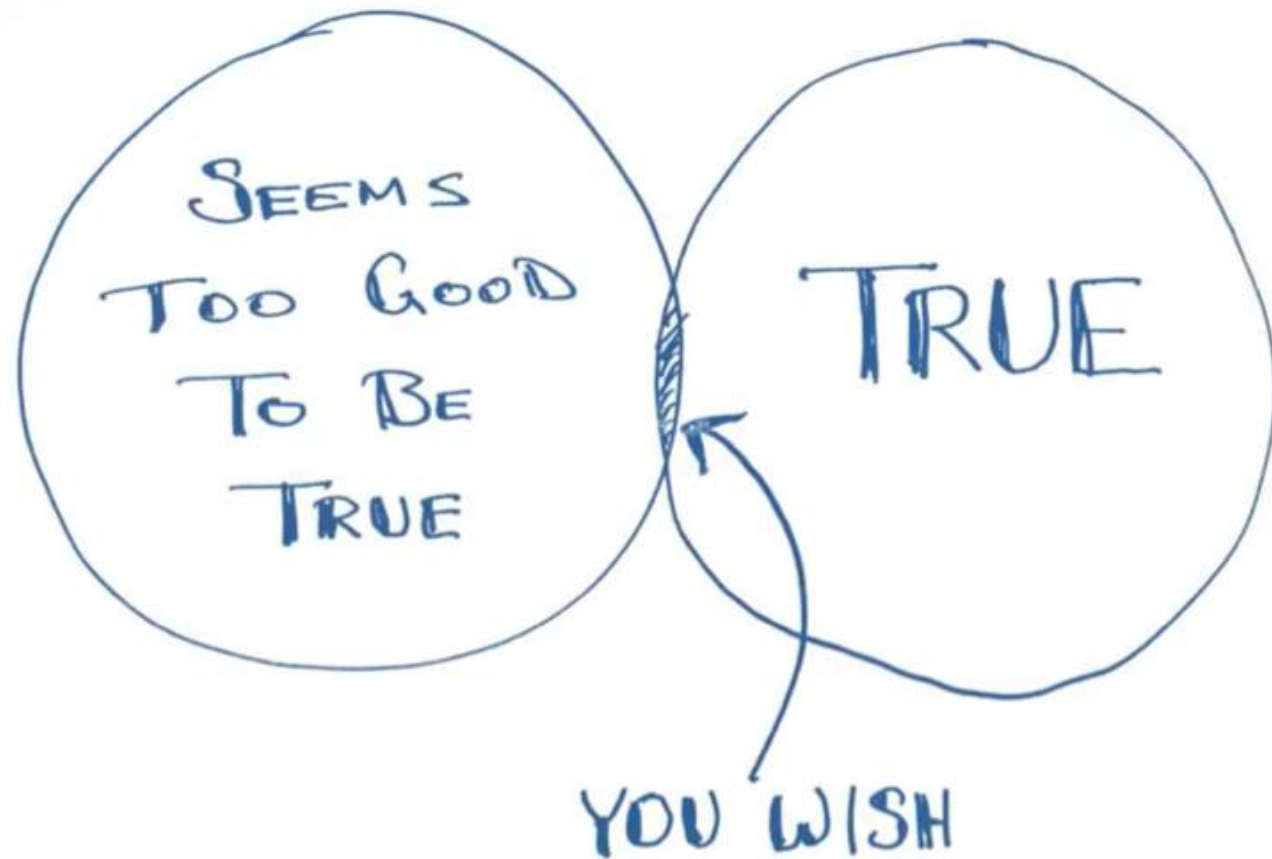
[United States](#) > Washington

● 2012-2013 ● Past years ▾



**States** | [Cities](#) (Experimental)







# Evaluation of Google Flu Trends

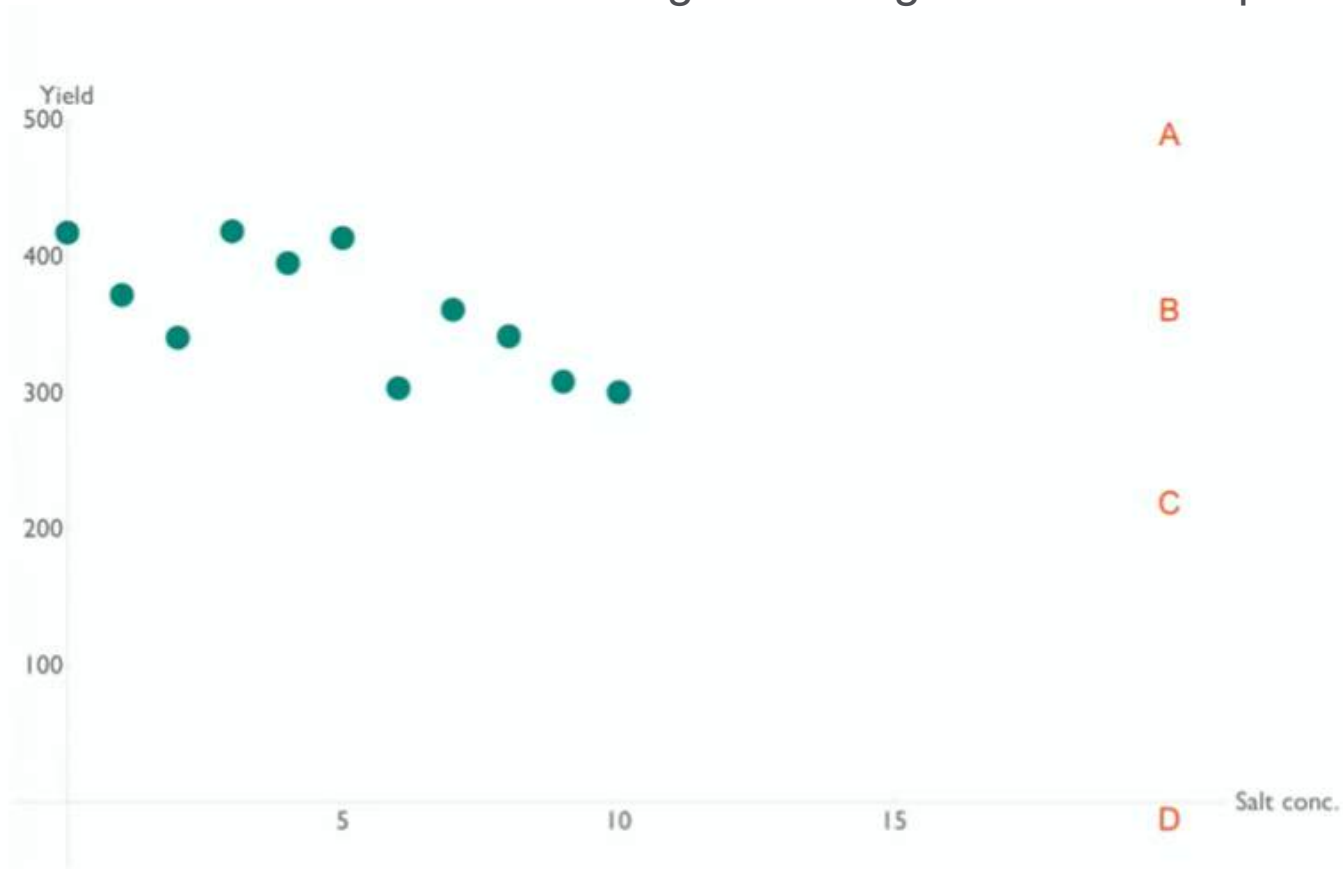
<b>Volume</b> How much?	<b>Variety</b> How different?
<b>Veracity</b> How trusted?	<b>Velocity</b> How fast?

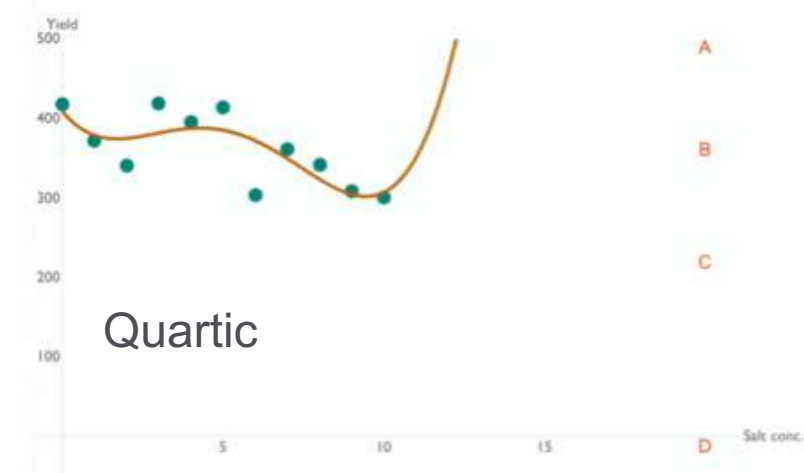
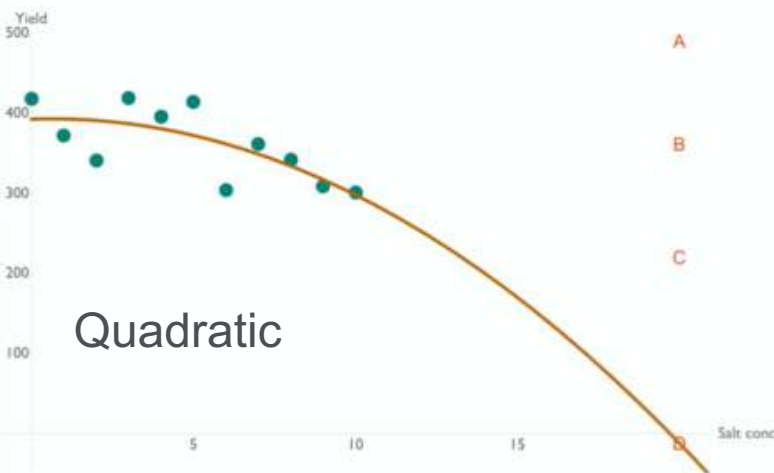
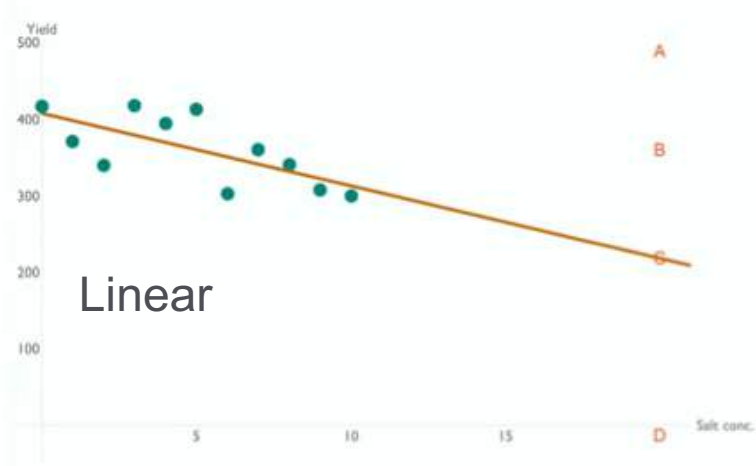
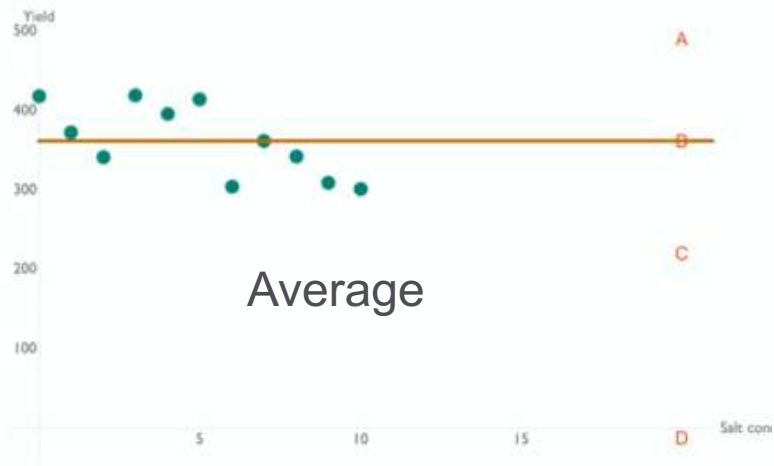


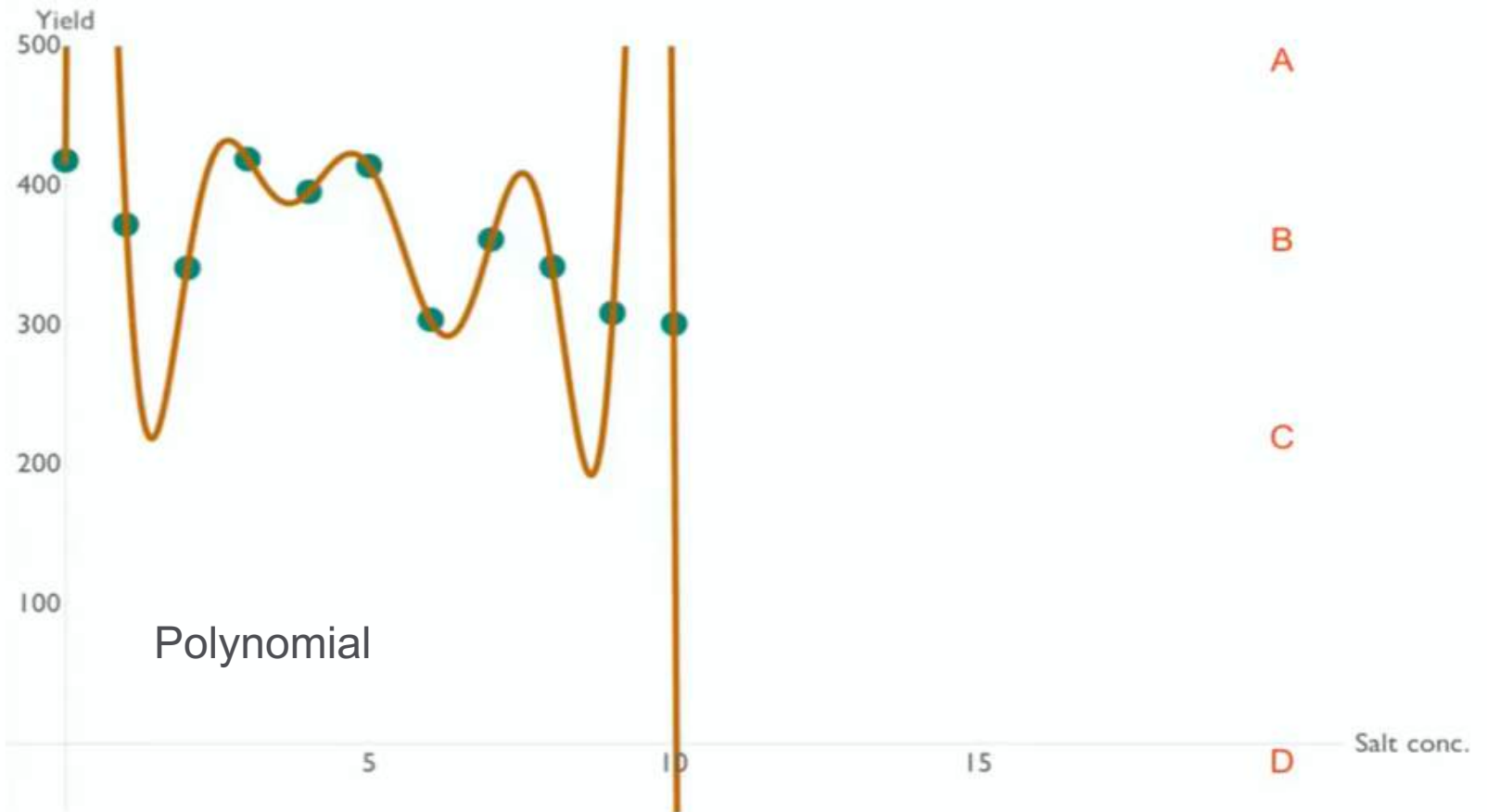
# Overfitting

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.

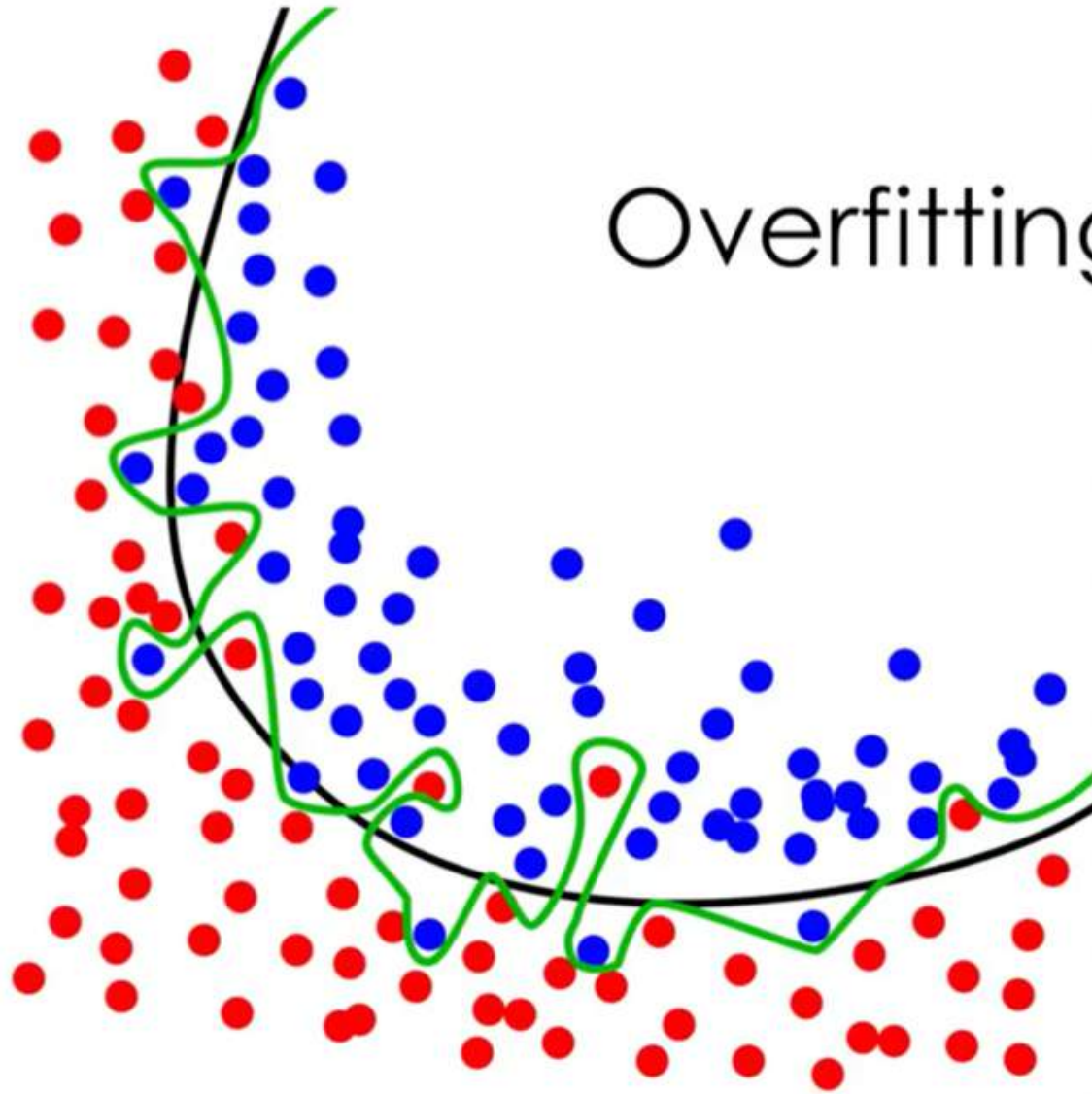








# Overfitting











## 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*



## Machine learning and prediction

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

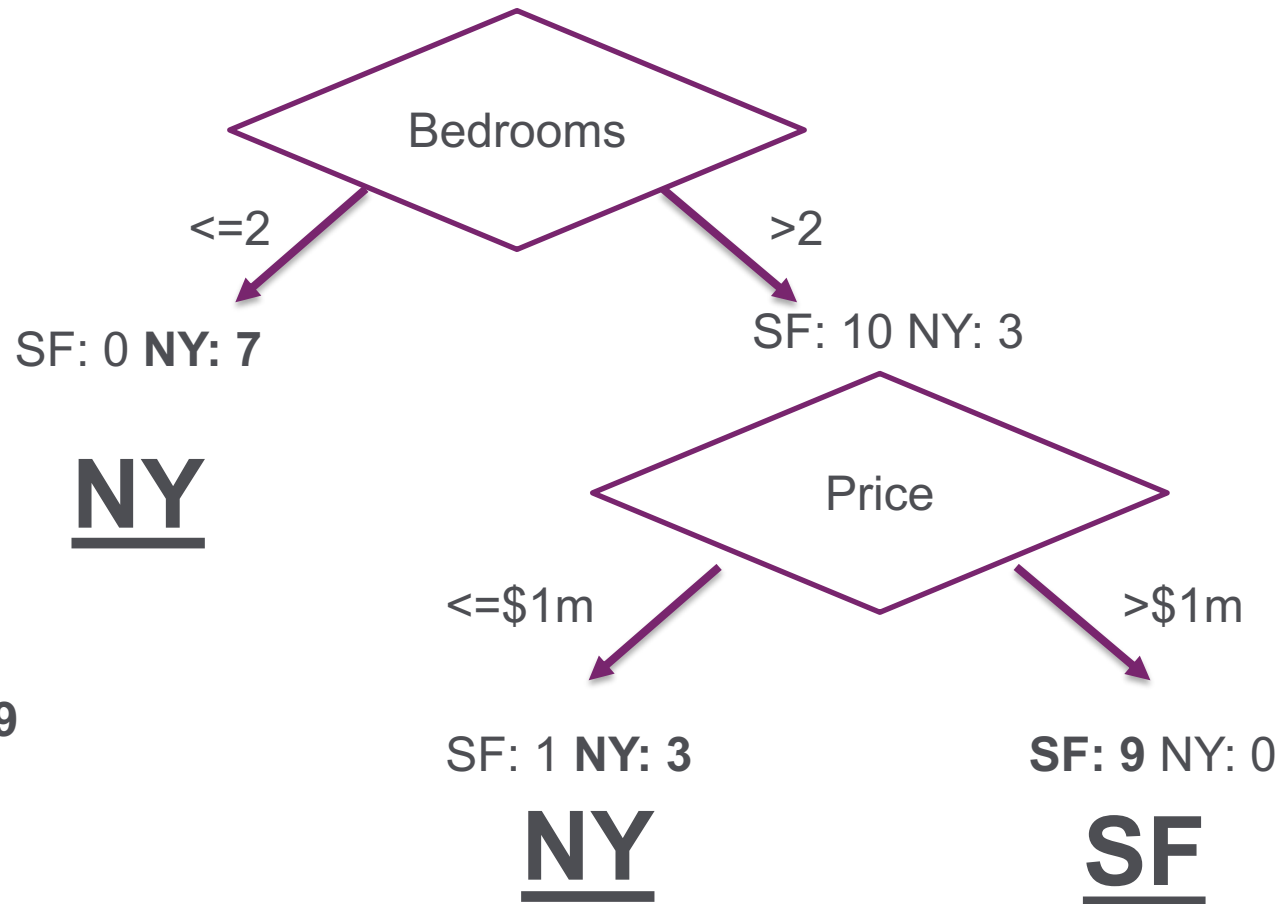
Build a decision tree to sort them into “New York” and “San Francisco”.

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



## Example decision tree

95%  
Confidence



**Correct: 19**  
Wrong: 1



# Machine techniques

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning



# 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?

Case study in action at [r2d3.us](http://r2d3.us)



## 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)





“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 preconception 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)



## 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



# Review

What is open data and what are the general benefits of open data?

What impact could data have on policy making?

In which situations should it be used/not used?

How to you spot red herrings in data analysis?



# Afternoon Session

## 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. Describe a number of ways to effectively present the results of data analysis.
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.
12. Review the role of open data in policy making



## 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)

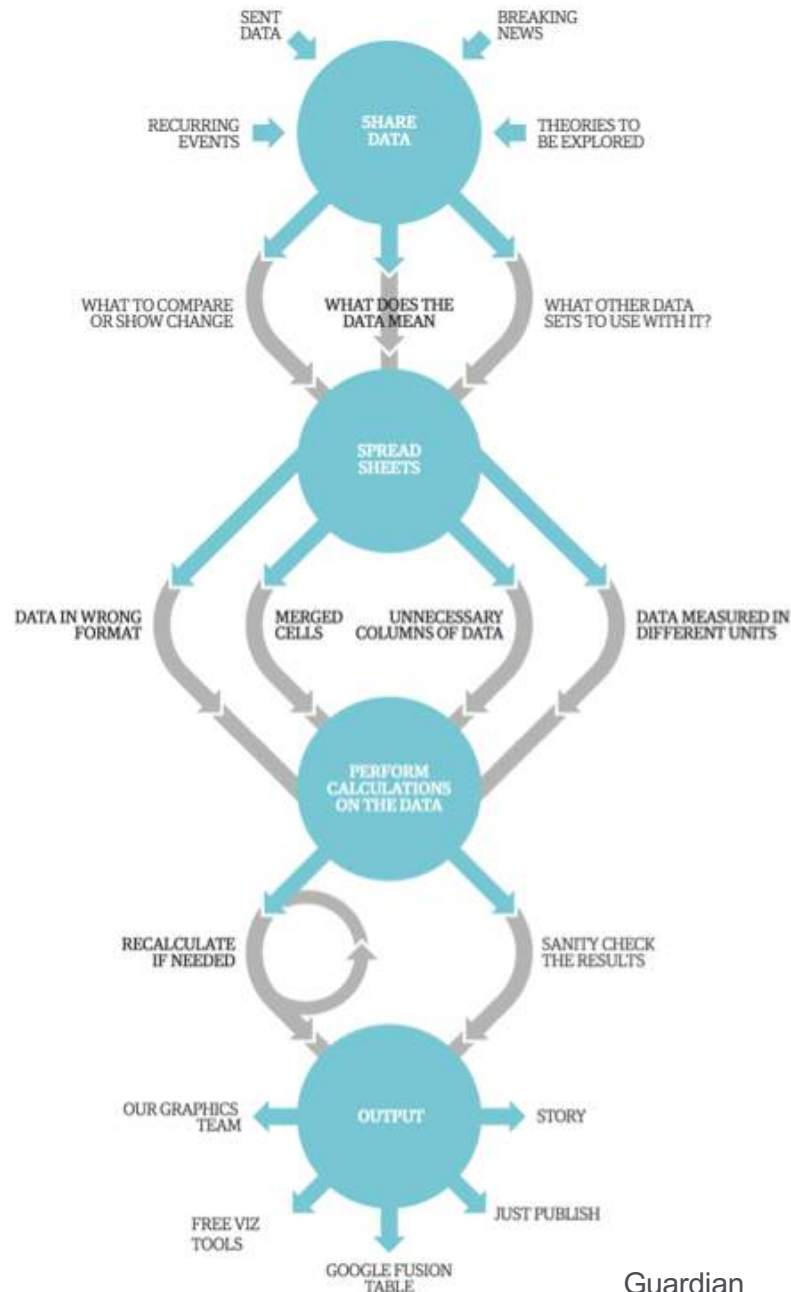


Paul Hudson



# 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.







# Data analysis practical

Modelling the impact of London Fire Station closures.

3. using performance data to advice on how to save money in the emergency services



Paul Hudson



## 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...



## 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 your design overfit (or use  $n$ th degree polynomial?)

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



Civil Service  
Learning

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



## Stage 1: collect data

<http://bit.ly/LFADData>

## Stage 2: establish context

- What does this dataset tell us?
- What doesn't it tell us?
- What more do we need to know to answer our question?
- Do we need more data?



## Stage 3: clean the data

Google refine government IT contracts [Permalink](#)

Facet / Filter Undo / Redo

Refresh Reset All Remove All

**Type of Contract** change insert reset  
815 choices Sort by: name count Cluster

FFAA: Fiscal/Financial Agent Agreement 3

FFIP 1

**FFP 512** edit statistics

FFP 1

FFP 1

FFP (OPS) 2

FFP (F&E) 1

FFP (Power Supply Retrofit) Old # OTFA01-82-D00004 1

FFP BPA 1

FFP CPAF CPiF 1

**512 matching rows (5200 total)**

Show as: rows records Show: 5 10 25 50 rows

	Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	
70.	2038	CGI FEDERAL INCORPORATED	FFP	10/03/2008	10/03/2008	1%
71.	2038	CGI FEDERAL INCORPORATED	FFP	01/08/2008	01/08/2008	0%
72.	2040	CGI FEDERAL INCORPORATED	FFP	01/08/2008	01/08/2008	0%
73.	2041	INTERNATIONAL BUSINESS MACHINES CORPORATION	FFP	03/17/2008	03/23/2009	1%
74.	2042	CGI FEDERAL INCORPORATED	FFP	04/21/2009	04/21/2009	0%
75.	2043	SOLUTIONS ENGINEERING CORP	FFP	11/01/2008	11/01/2008	10%
76.	2044	EVERGREEN INFORMATION TECHNOLO	FFP	11/20/2008	11/20/2008	0%
84.	7946	INTERNATIONAL BUSINESS MACHINES CORPORATION	FFP	10/01/2009	10/01/2009	0%
85.	7947	THE NEWBERRY	FFP	10/01/2009	12/01/2009	0%



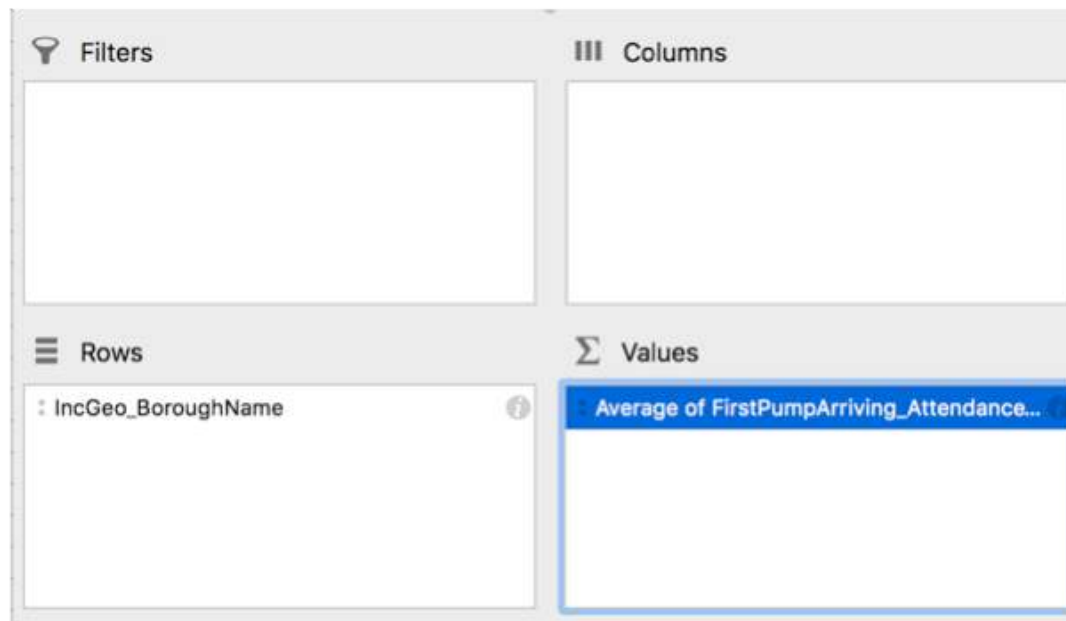
## Stage 4: translate, filter and merge data

- Remove records for before the closures (7 Jan 2013).
- Ensure that none of the closed stations have records entered after this date.
  - Belsize, Bow, Clerkenwell, Downham, Kingsland, Knightsbridge, Silvertown, Southwark, Westminster and Woolwich.



## Stage 5: initial analysis

- Work out the mean attendance time for the first appliance to arrive across London.
- Work out the mean attendance time per borough (using a pivot table).







## Stage 5: initial analysis (part 2)

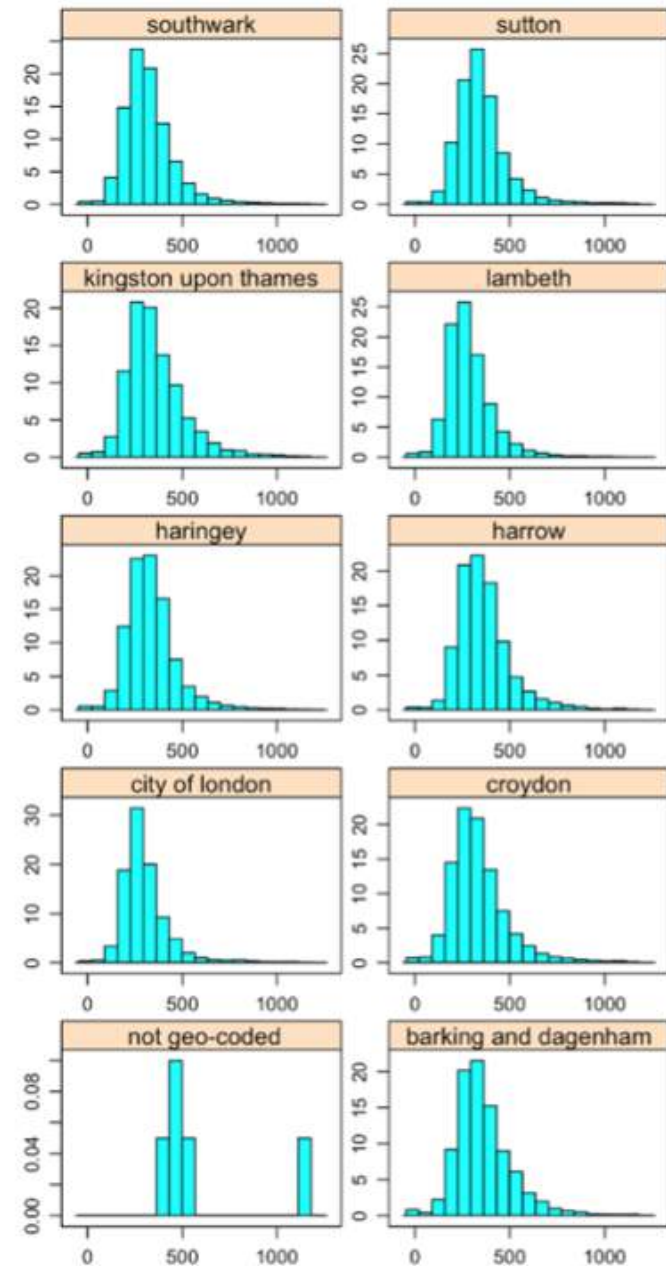
- Copy the average attendance time from Column B to Column C.
- Hide column B.
- Create column D from column C by dividing all values by 86400 (seconds in a year).
- Hide column C.
- Format the cells of column D and select mm:ss from custom data types.
- In a new cell at the top insert the value  $(=360/86400)$ .
- Format that as per column D to turn it to a time.
- Use conditional formatting to highlight rows where the response time is greater than this new cell value (6 minutes).



## Stage 6: secondary evaluation and analysis in R

R is an advanced statistical library that is much more powerful than excel and can handle vast quantities of data.

Follow the exercise at <http://bit.ly/LFA-R> to find out more.

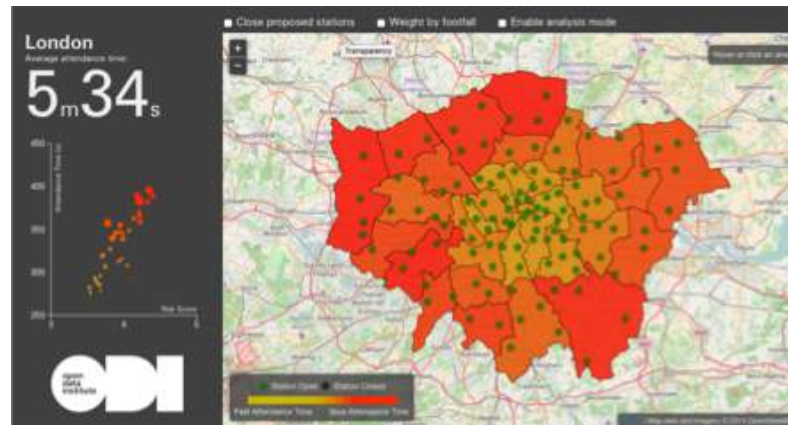




## Stage 7: sanity check

- Download the dataset prior to closures and look at differences in response times?
- Was the ODI impact analysis tool accurate?

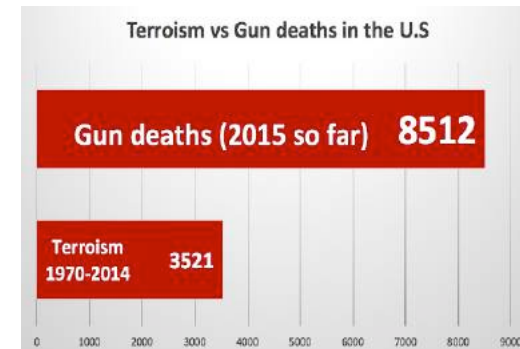
<http://london-fire.labs.theodi.org/explore/>





## Stage 8: creating output

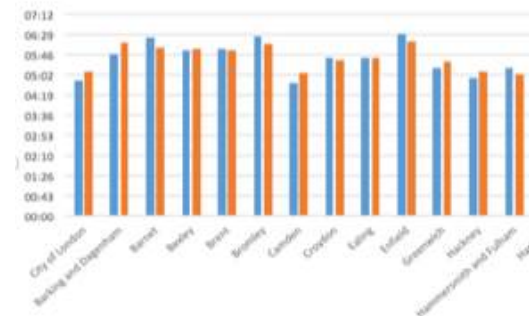
- What is your headline?
- If you want to communicate one message (or headline), keep any visualisations simple and effective.
- If you want people to be able to explore the data for themselves and create their own story, then it is possible to create something interactive.



## Stage 8: 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 and after statistics can be easily generated in excel.
- An highly interactive dashboard can be created with [dataseedapp.com](https://dataseedapp.com).



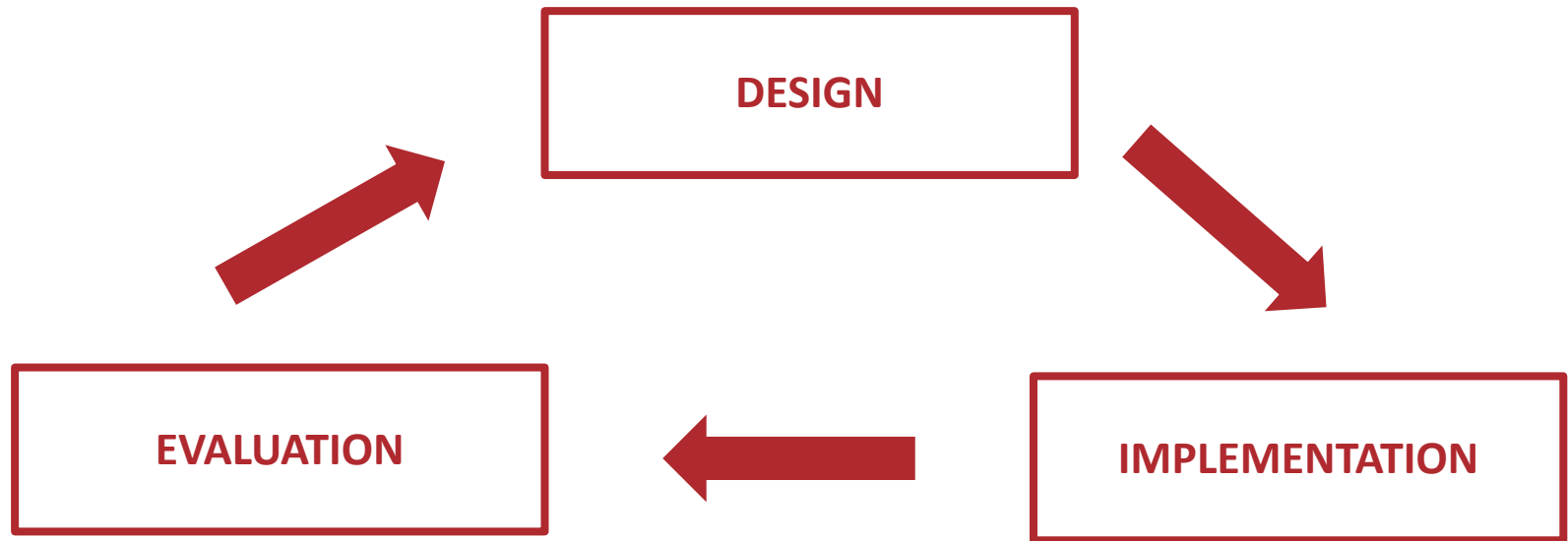


Civil Service  
Learning

# Open data in policy cycles



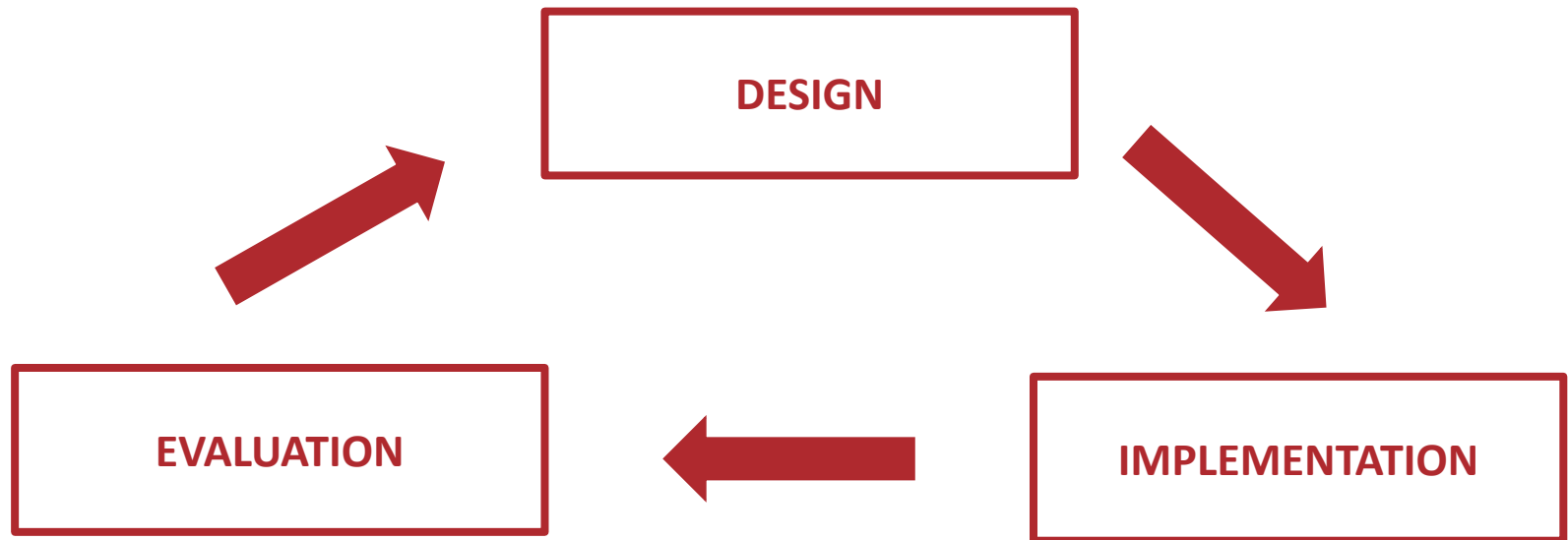
# Public policy cycle





## Question

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







# 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?





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**)





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?



# 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





## 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

Now we've used data as an input,  
what about output and as a tool?



### 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



# Review

In your group identify a policy problem (1 sentence)

Using what you have learnt today, create list of key considerations and actions you need to take to solve this problem

What is the main thing you have learnt today that changes your thinking on how to solve this problem?





Civil Service  
Learning

## Thank you

We hope you enjoyed this  
experience brought to you by

Delivered by



Civil Service  
Learning



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