



# Finding Stories in Data

---

David Tarrant · [@davetaz](https://twitter.com/davetaz)

Misleading statistics  
Enriching data and using pivot tables

# Session 3

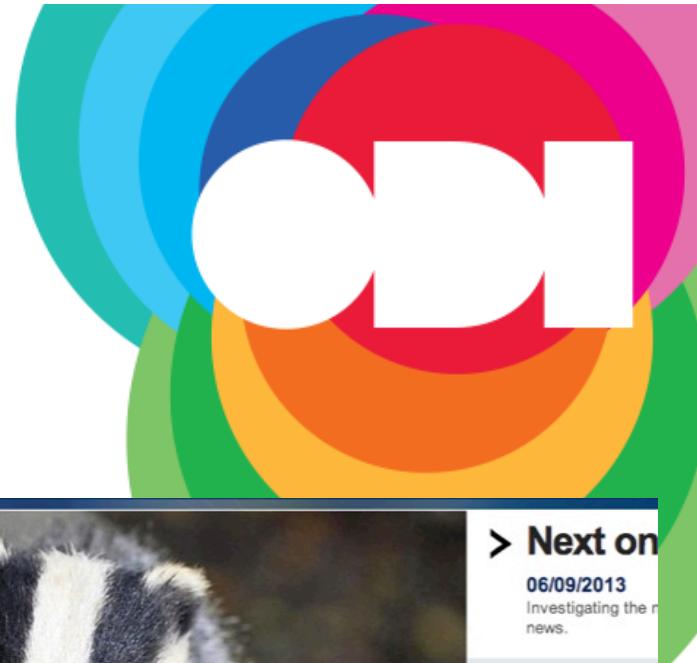
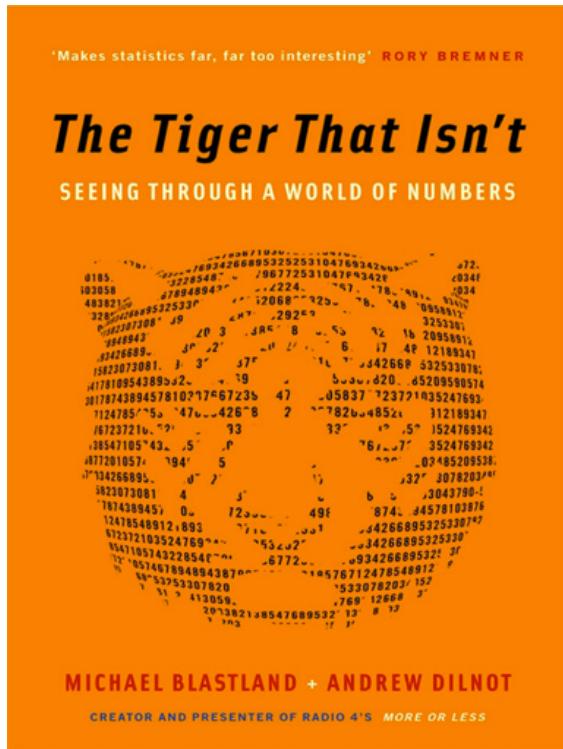
# Statistical analysis of data



# Misleading statistics



# 12 problems with numbers



The image shows a BBC Radio 4 program page for 'More or Less'. It features a close-up photograph of a badger's face. To the right of the photo, there is text: '> Next on 06/09/2013 Investigating the news... news.' Below that, it says 'BBC RADIO 4 Friday 16:00 BBC 1 FM 106.3 See all upcoming More or Less (2)'.

**Latest episode**  
**What price the life of a badger?**  
Tim Harford queries the numbers of the badger cull, plus NHS deaths and climate migrants.

**Listen now**

**Free downloads**



# 1. Counting



Flickr: mattbrittain

## 2. Big numbers

£300m

boost for childcare

1,000,000

new places



£1.15  
per week  
per child

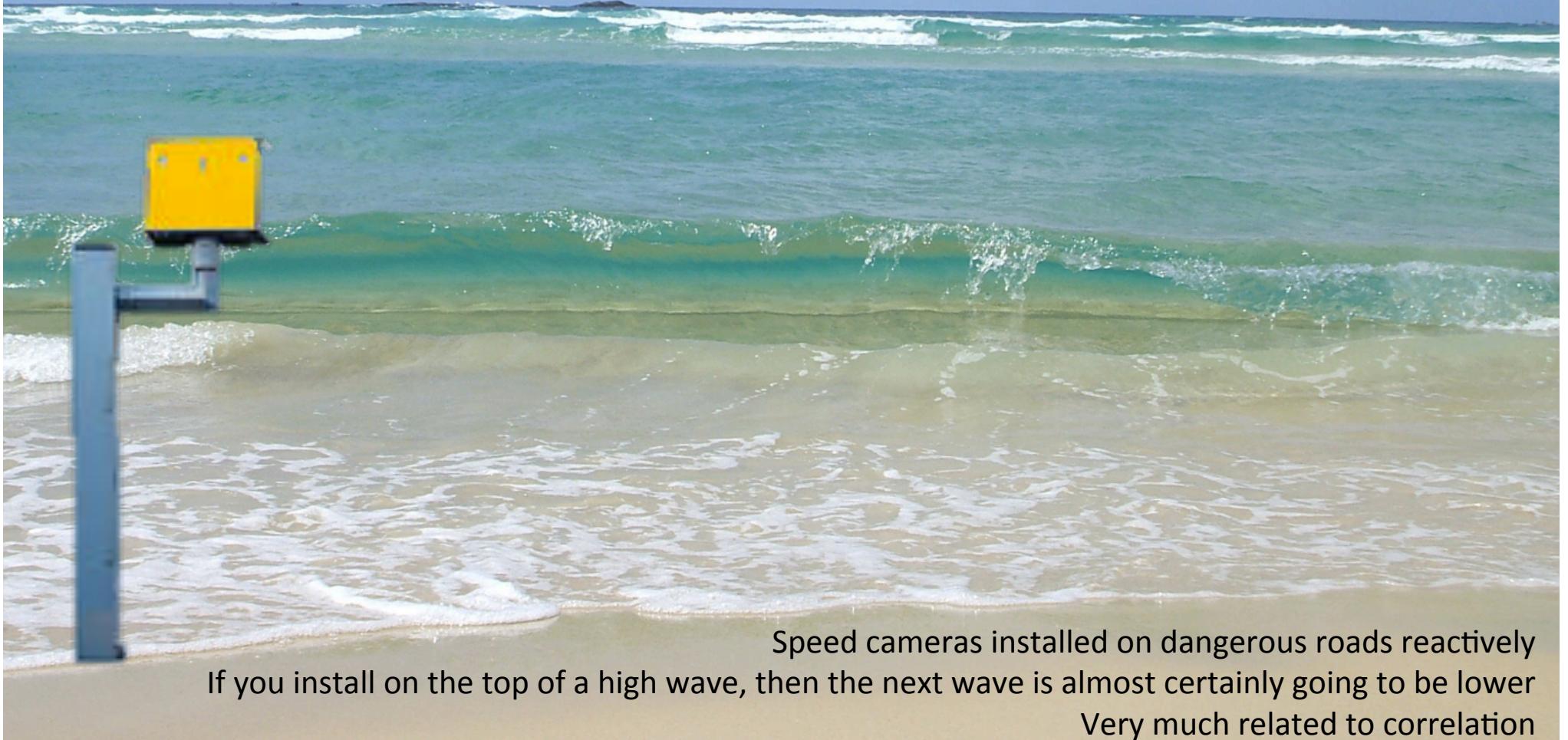
### 3. Chance



Random events cluster

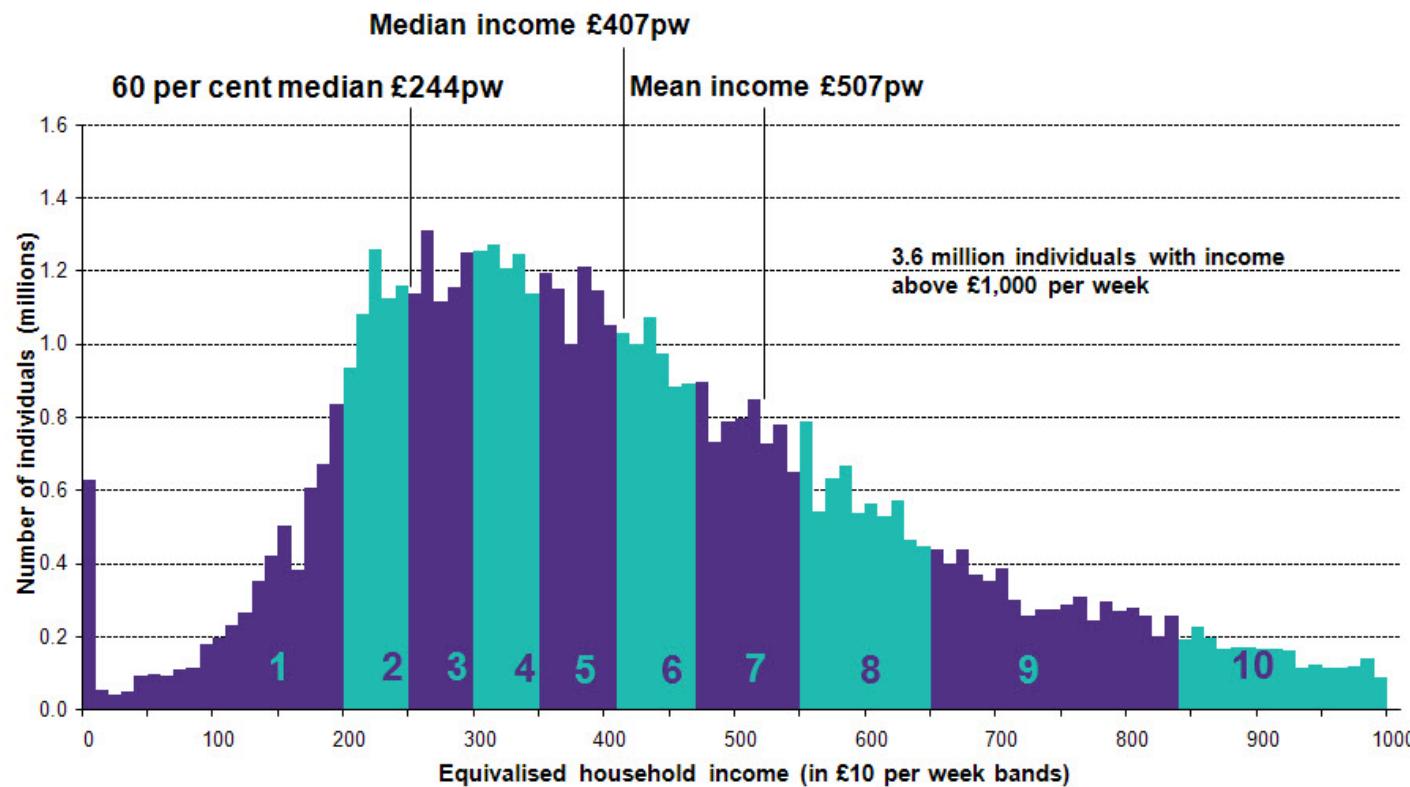
They do not evenly distribute

## 4. Fluctuation



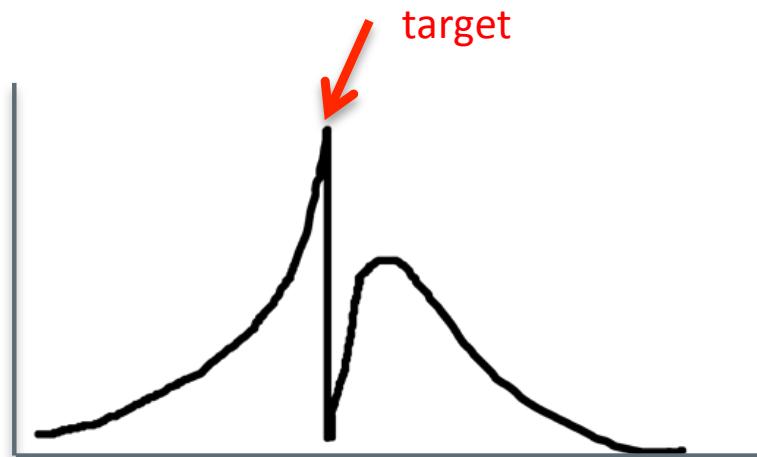
Speed cameras installed on dangerous roads reactively  
If you install on the top of a high wave, then the next wave is almost certainly going to be lower  
Very much related to correlation

# 5. Averages



## 6. Targets

Look at one aspect to measure an entire service.



Encourage gamification

# 7. Risk

 SECTIONS OTTAWA CITIZEN

HOME NEWS NATIONAL LOCAL NEWS

# Combined vac seizure risk in

 ELIZABETH PAYNE More from... Published on: June 9, 2014 | Last Updated: June 9, 2014

The number of worldwide cancer cases is set increase by 75 per cent in the next two decades, according to researchers in France.

The scientists predict cancer cases will increase from 12.7 million in 2008 to 22.2 million by 2030.



CANCER  
RESEARCH  
UK

HOME MENU ▾ SEARCH ▾

[Home](#) > [About us](#) > [Cancer News](#) > [News report](#) > Global cancer incidence predicted to increase by 75 per cent by 2030

## Global cancer incidence predicted to increase by 75 per cent by 2030

 31 May 2012  In collaboration with the Press Association

**Recent news**

[Investing in cancer research boosts economy as well as](#)



## 8. Sampling

Between 2,000 and 5 million cases of norovirus in winter 2007/08.

Based upon 2,000 confirmed cases extrapolated from BMJ report based on sample size of .... 1

## The Telegraph

[Home](#) [News](#) [World](#) [Sport](#) [World Cup](#) [Finance](#) [Comment](#) [Culture](#)

[Politics](#) [Investigations](#) [Obits](#) [Education](#) [Earth](#) [Science](#) [Defence](#)

[HOME](#) » [NEWS](#) » [UK NEWS](#)

GPs urge millions hit by bug to stay at home



The NHS advises symptoms

[News Front Page](#)  
[World](#)  
[UK](#)  
[England](#)  
[Northern Ireland](#)  
[Scotland](#)  
[Wales](#)  
[Business](#)  
[Politics](#)  
[Health](#)  
[Medical notes](#)  
[Education](#)  
[Science & Environment](#)  
[Technology](#)

[LIVE](#) [BBC NEWS CHANNEL](#)

Last Updated: Friday, 11 January 2008, 12:02 GMT

[E-mail this to a friend](#)

[Printable version](#)

### Vomiting bug 'hits three million'

Almost three million people have been affected by the norovirus stomach bug so far this winter, figures suggest.

Surveillance from the Health Protection Agency shows cases in England and Wales are double those seen last year.

Doctors advise people to stay at home for 48 hours after



Norovirus causes sudden vomiting and diarrhoea



## 9. Data (known unknowns)



What share of income tax paid in the  
Singapore is paid by the top 20% of earners?

♦ A: 20%

♦ C: 60%

♦ B: 40%

♦ D: 80%



## 9. Data (known unknowns)



How much bigger is the Singapore economy now (inflation adjusted) than in 1945?

♦ A: 300%

♦ B: 600%

♦ C: 900%

♦ D: 1200%

## 9. Data (known unknowns)

What is the average number of children per family in Bangladesh?

♦ A: 2

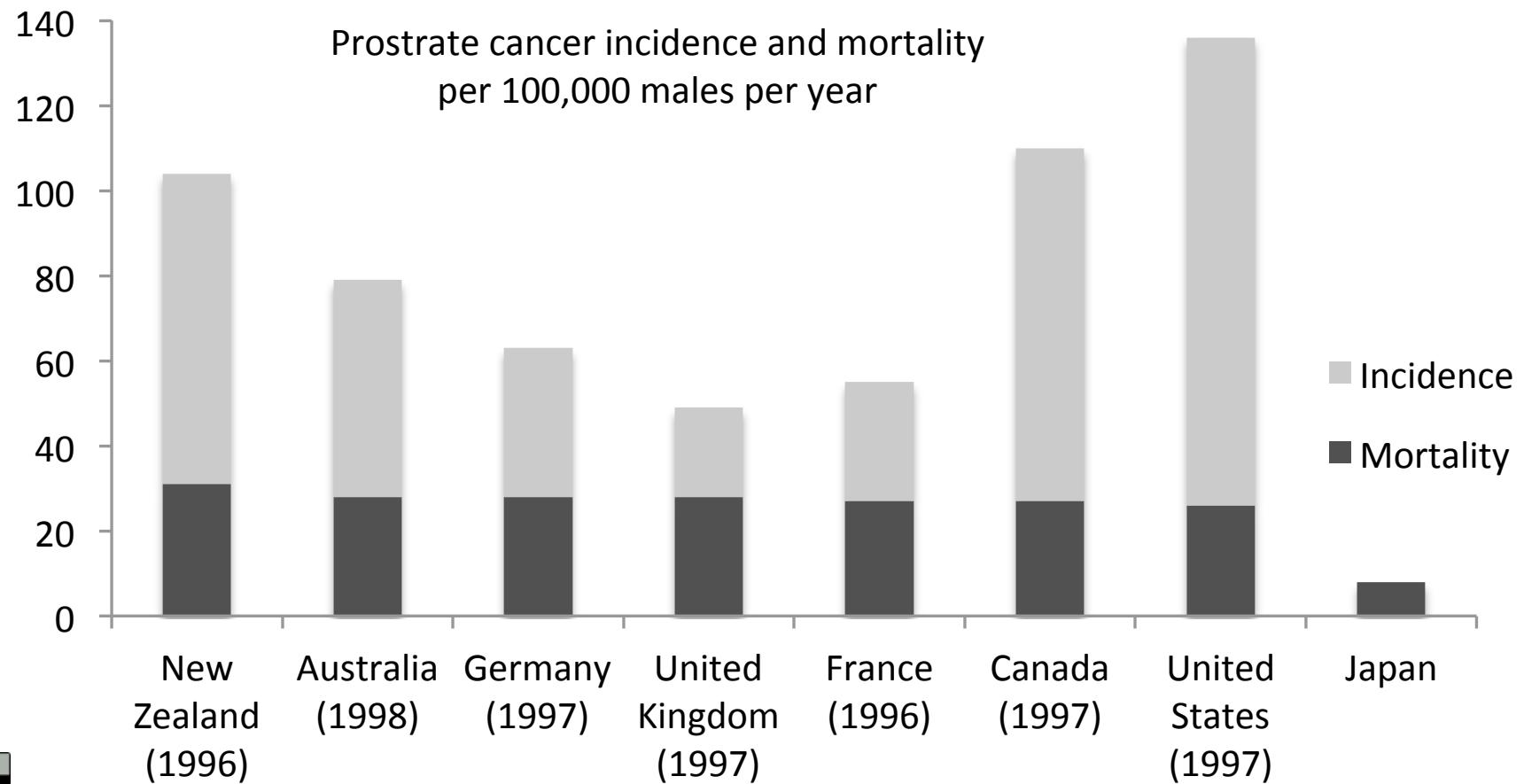
♦ C: 4

♦ B: 3

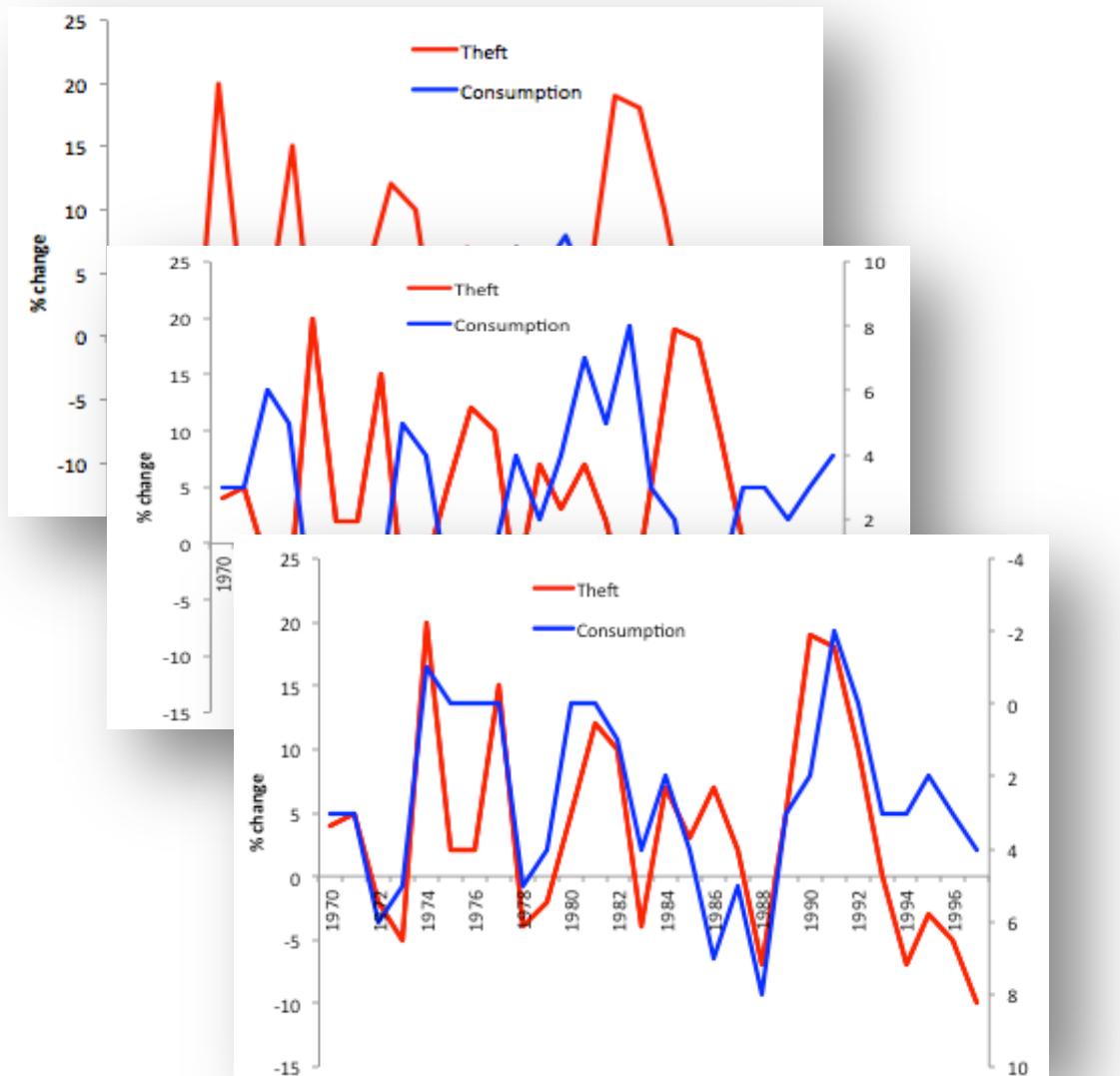
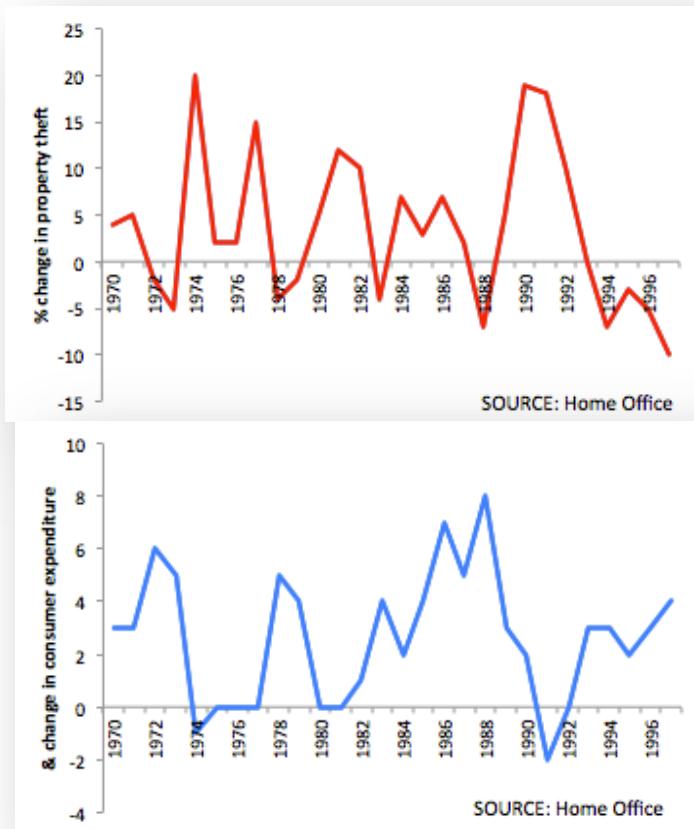
♦ D: 5



# 10: Comparison



# 11. Correlation



## 12. Percentages

Know the difference between a **percentage** and a **percentage point**.

VAT increased from 17.5% to 20% on January 2011.

This is a rise of 2.5 percentage points not a rise of 2.5%.

How much would a rise in 2.5% actually be?



$$17.5 * 1.025 = 17.9375$$



From: Stories and Statistics by Frank Swain

# 5 ideas for finding a story

## **1. Create a ranking**

Sort columns to explore top and bottom property prices

## **2. Compare groups**

Newly built properties to established residential building

## **3. Use sum/average/min/max functions**

Find the median price of a property

## **4. Use pivot tables to compare groups**

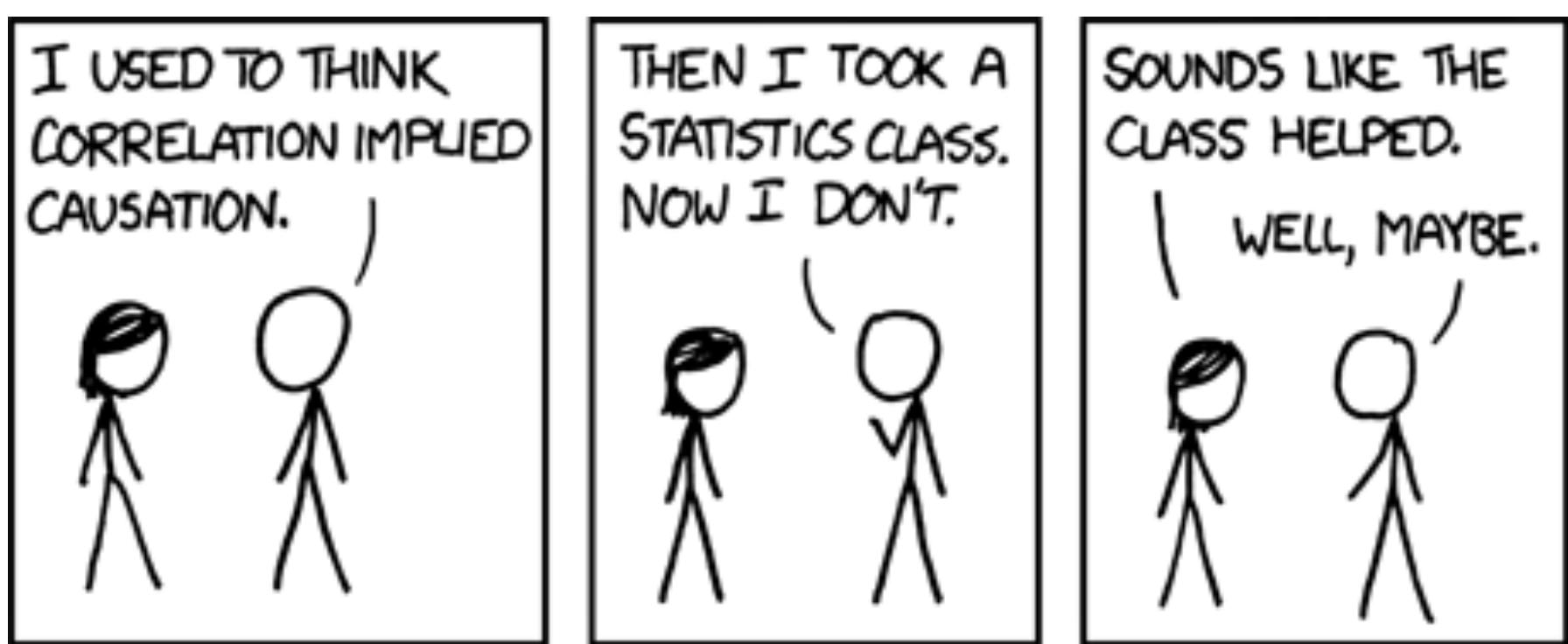
Compare the prices of different geographies in the UK

## **5. Look for anomalies in the data**

e.g. in the “date” or “postcode” columns



# XKCD



Creative Commons Attribution Non-Commercial <http://xkcd.com/552/>

# Prepare

Prepare

**2.1 CLEAN**

**2.2 TRANSFORM**

**2.3 COMBINE**

**2.4 ENRICH**

**2.5 ANALYSE**



# Enriching data and using pivot tables





The Open Database Of The Corporate World

We have information on  
**70,597,888** companies

# Aggregator/Enabler

search companies  search officers

SEARCH

Filter by jurisdiction

**1,298** Abu Dhabi (UAE)

**144,755** Alaska (US)

**40,157** Albania

**899,455** Arizona (US)

**46,537** Aruba

**165,582** Bahamas

**99,185** Bahrain

**88,563** Bangladesh

**22,140** Belarus

Just released:  
OpenCorporates API v0.3

Corporate network data,  
financial accounts, complex  
filters, and more. [Read more](#)

Get data access to over  
60 million companies

Open data

- All the data on the world's largest  
open database of companies
- Available as either raw or  
structured open data or  
commercially
- All data includes sources, allowing  
tracking of data

Quality data

- Data from primary public sources
- All data is rigorously  
checked and validated
- Many users on OpenCorporates  
with quality control for free

Unique data

- Open, transparent and highly  
granular model
- Many unique datasets and tools
- Enrich your existing data, or build  
new services

Announcing Open LEIs

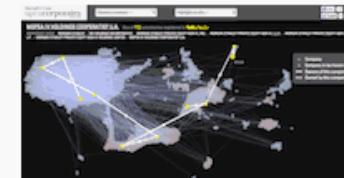
Today, OpenCorporates  
announces a new sister website,  
[Open LEIs](#), a user-friendly  
interface on the emerging Global  
Legal Entity Identifier System.  
[Read more](#)

**OPENLEIs**

A BETA VIEW ON THE LEI SYSTEM

New! Just added: Open  
corporate network data

[Read more](#) about this important  
new feature



# Exercise

Enriching a dataset containing  
company names (e.g. transactions)  
with company data from  
OpenCorporates



# Sense-checking

The best way to sense-check is to get a second pair of eyes to help you.

Any stories of common mistakes you'd like to share?





G'DAY

Thank-you