Dr Eddy Campbell, a little about me …

- PhD in Statistics (1993).
- Consultant Statistician with Shell Research for ~ 3 years
- Research Statistician with CSIRO for ~ 20 years.
  - Held project and other leadership roles, along with being a Principal Research Statistician.
- I do some teaching with Curtin, as well as working with Fremantle Hospital's Mental Health Service.
- Research interests include advanced Baseball analytics.

Unit outline

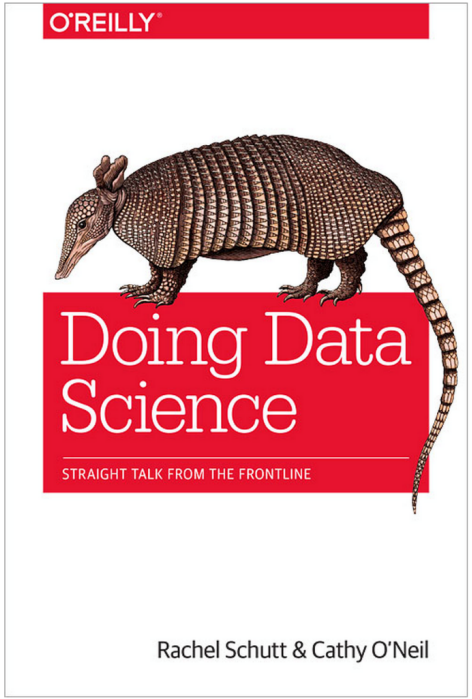
- See Unit Outline on BB for more details
- Staff
  - Unit co-ordinator, lecturer: **Eddy Campbell** (Edward.Campbell@curtin.edu.au)
  - Tutors: Shi Ching Fu, Casey Josman
- Classes
  - Lecture: **Tuesday** 16:00 – 17:00 (401.002);
    Workshop 17:00 – 18:00 (When there is a speaker).
  - Computing Labs: A number of options, so check schedule and Bb.
    - **Will begin Week 2**

Outline

- What is data science?
- Unit objectives and how it’s all going to work
- Contemporary data science

WHAT IS DATA SCIENCE?

What's the hype?



- What does DS mean?
  - Science of 'big data'?
  - Whatever Facebook and Google do?
- What's the hype?
  - 'Masters of the universe'!
- Is it really new?
  - Machine learning algorithms weren't just invented last week!

Is there really a neat definition of data science?

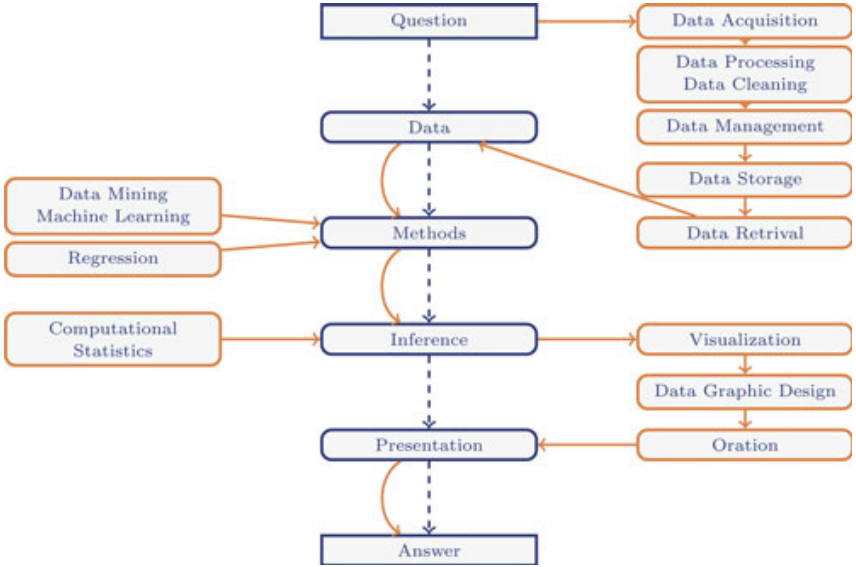
So what is data science anyway?

- "… the latest tech marketing buzz word ….? (Ray Garcia)
- Applied statistics by a different name? (Nate Silver)
- Statistics with a deep understanding of computer science? (Justin Megahan)
- What's a data scientist?

  "*Someone who knows more about statistics than a computer scientist, more computer science than a statistician, and who can communicate better than either of them*" (anon.)
- Data science is just what data scientists do! (anon.)

Defining data science by its constituent steps?

- Obtaining data, e.g., web scraping
- Cleaning and then visualizing the data
- Statistical modelling or machine learning
- Presentation and communication



http://dx.doi.org/10.1080/00031305.2015.1081105

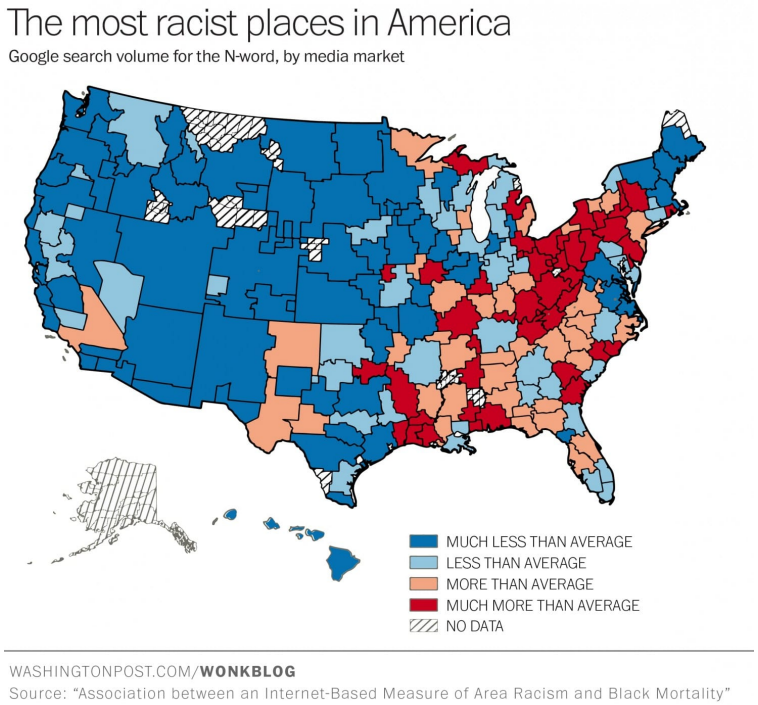
Data science – why now?

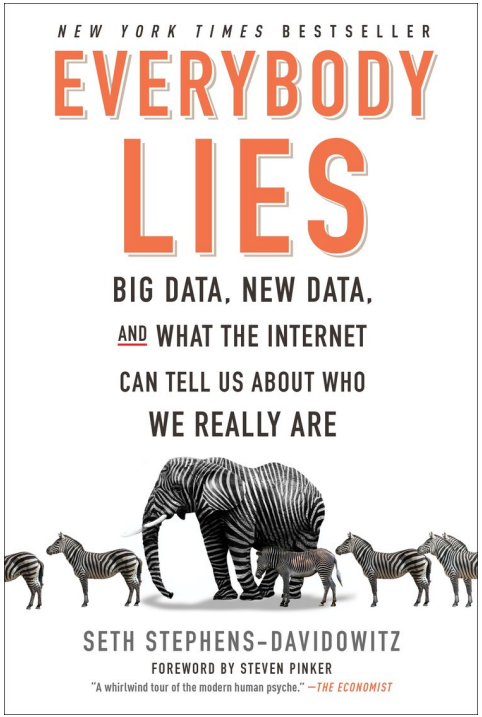
- Since at least the time of Galileo, we've been using data to answers all sorts of questions; but …
- We now have the capacity to measure and store massive amounts of data, and we have the computing power to analyze it all
  - Data gathered from online tracking: websites visited, clicks, …
  - Data in other fields: finance, medical records, pharmaceuticals, bioinformatics, social welfare, retail, ...
  - Data on when you log into BB, connect to wifi on campus ...
- Challenge is to extract useful *information*
  - Data becomes the building blocks of products, e.g., recommendation systems
- Cannot lose sight of the ethical implications of what we do with all this data

Interest in ‘data science’

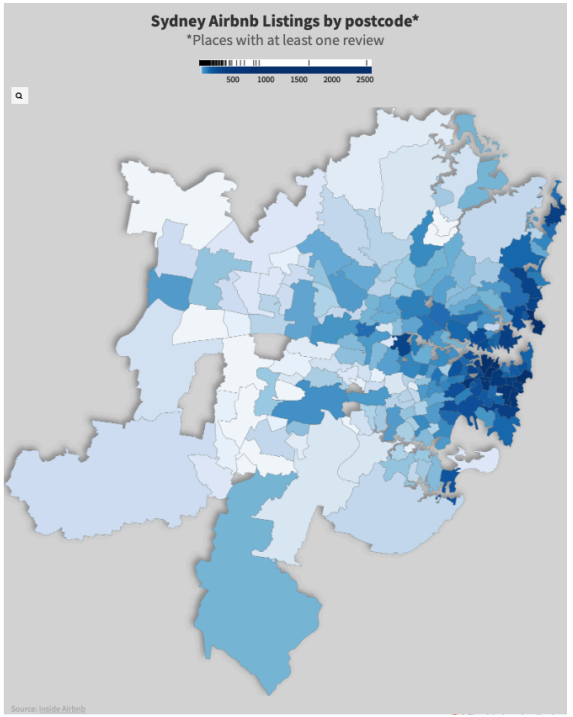
https://trends.google.com/trends/explore?date=all&geo=AU&q=data%20science

Google searches as data





Airbnb listings as data



Example: unintended consequences – Netflix Prize (2009)

- Can we predict what rating a user is going to give a movie that s/he hasn't watched yet?
- Used to make "You may like to watch …" suggestions to the subscriber
- Grand prize of $1M

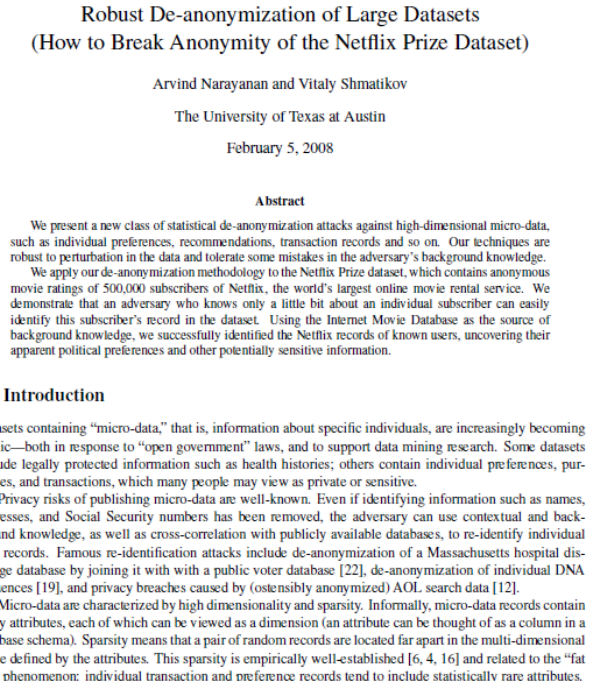
Netflix Prize

- A *training* dataset consisted of quadruplets of the form

  `<user, movie, date of grade, grade>`

  for over 100M ratings from 480K users to 18K movies
- Models were evaluated on a *test* set consisting of 1.4M ratings
- The winning team obtained the smallest RMSE (`0.8567`) on the test set that was also less than the RMSE (`0.9525`) of Netflix's own algorithm
- However, …

Netflix prize and privacy



Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

1 Introduction

Datasets containing "micro-data," that is, information about specific individuals, are increasingly becoming public—both in response to "open government" laws, and to support data mining research. Some datasets include legally protected information such as health histories; others contain individual preferences, purchases, and transactions, which many people may view as private or sensitive.

Privacy risks of publishing micro-data are well-known. Even if identifying information such as names, addresses, and Social Security numbers has been removed, the adversary can use contextual and background knowledge, as well as cross-correlation with publicly available databases, to re-identify individual data records. Famous re-identification attacks include de-anonymization of a Massachusetts hospital discharge database by joining it with with a public voter database [22], de-anonymization of individual DNA sequences [19], and privacy breaches caused by (ostensibly anonymized) AOL search data [12].

Micro-data are characterized by high dimensionality and sparsity. Informally, micro-data records contain many attributes, each of which can be viewed as a dimension (an attribute can be thought of as a column in a database schema). Sparsity means that a pair of random records are located far apart in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [6, 4, 16] and related to the "fat tail" phenomenon: individual transaction and preference records tend to include statistically rare attributes.

- Researchers were able to identify some individual users who had also rated movies on publicly-available sites such as IMDB!

Any other examples?

Unit outline details …

Unit objectives

- Describe and explain the importance of each of the four key aspects of data science to extracting information from data
- Manipulate and visualise data for effective presentation
- Apply statistical and machine learning methods to real datasets
- Present and describe results of data analysis

Why use *R*?

- ***R* is central to data science, as is Python, which you're learning in COMP1005**
- Free, open-source, multi-platform statistics computing environment that has become the standard platform for statistical analysis
- (Relatively) easy to learn to do simple things
- Lots of additional packages, online resources available

Rstudio and *knitr*

- *Rstudio* is a very useful and powerful environment that makes *R* a lot easier to use and manage
  - Think of it as an IDE (integrated development environment) for *R*
- *Knitr* is integrated into *Rstudio* and is an engine for dynamic report generation with *R*
  - Allows for easy integration of *R* code into different text processing environments
  - We'll be using an extension known as *Rmarkdown*, which allows a number of output formats to be produced, e.g., PDF, html, MS Word
  - One of the main advantages of using *knitr* is that it allows the analyst to provide **reproducible analyses and/or research**

Reproducible research

- *"Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations. Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available."*
  - https://www.coursera.org/learn/reproducible-research

knitr / Rmarkdown example

Example.Rmd

Preview | Insert | Run

```
1  ---
2  title: "R Notebook"
3  author: "Aloke Phatak"
4  date: "1 March 2017"
5  output: html_notebook
6  ---
7
8  This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code
   within the notebook, the results appear beneath the code.
9
10 Try executing this chunk by clicking the *Run* button within the chunk or by placing
   your cursor inside it and pressing *Cmd+Shift+Enter*.
11
12 ```{r}
13 summary(cars)
14 ```
15
16 Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing
   *Cmd+Option+I*.
17
18 ```{r}
19 plot(cars)
20 ```
21
22
23 When you save the notebook, an HTML file containing the code and output will be saved
   alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML
   file).
24
```

Files | Plots | Packages | Help | Viewer — Publish

Code

R Notebook

Aloke Phatak

1 March 2017

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

Hide

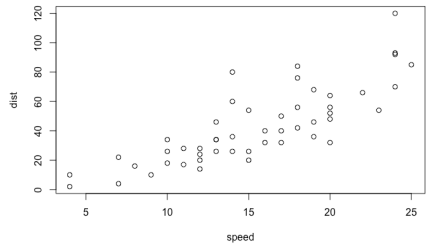
```
summary(cars)
```

```
     speed           dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

Hide

```
plot(cars)
```



When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

Other tools

- R statistical language environment will be the principal tool, but you are encouraged to use others, e.g., visualization software, alongside it:
  - Tableau
  - PowerBI

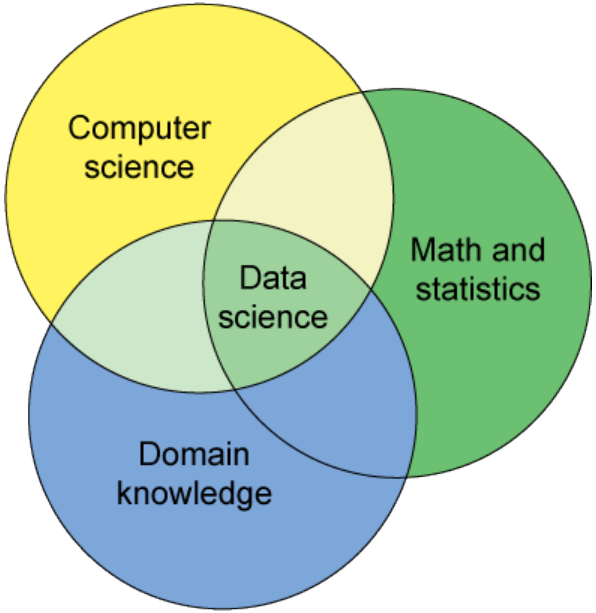
Project

- Think of a topic of interest to *you* – and then start looking for data!
- Where to start?
  - The web!
  - State and Federal governments
  - News websites, e.g., ABC
- What is the story you're going to tell?

CONTEMPORARY DATA SCIENCE

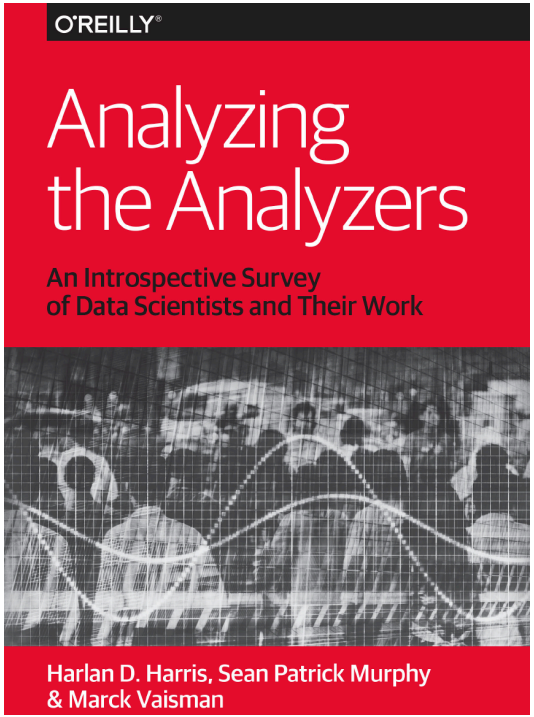
Data science

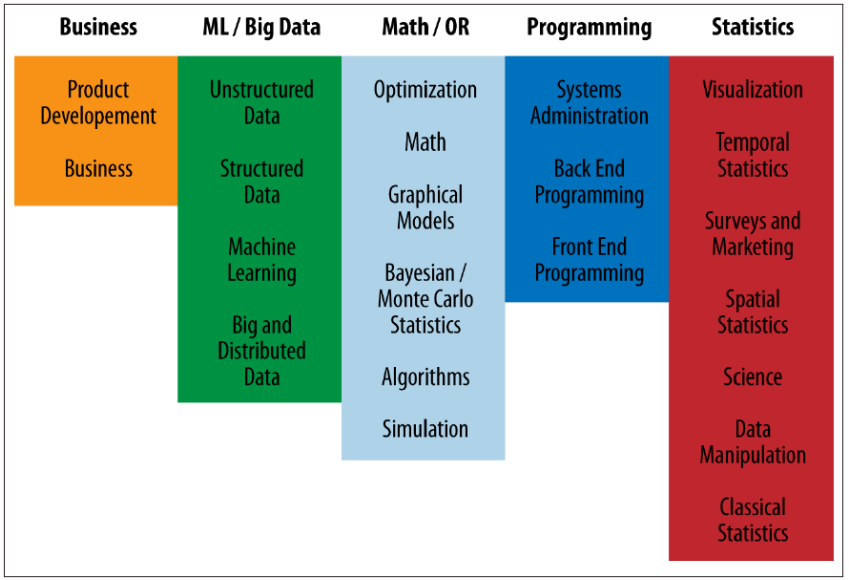
- Data science is not just another word for ‘statistics’!
- It sits at the intersection – or should that be union? – of several domains:
  - Statistics
  - Computer science
  - Visualization
  - Communication



http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png

Analyzing the analyzers

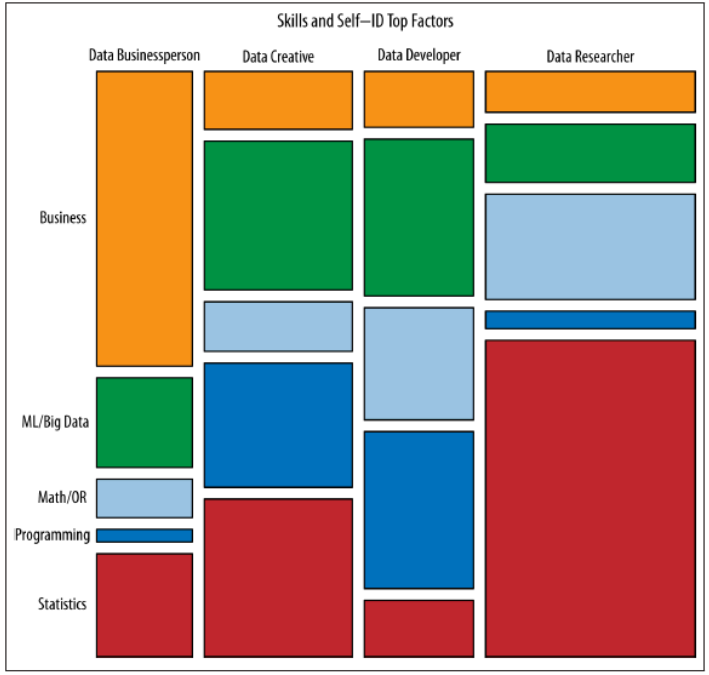




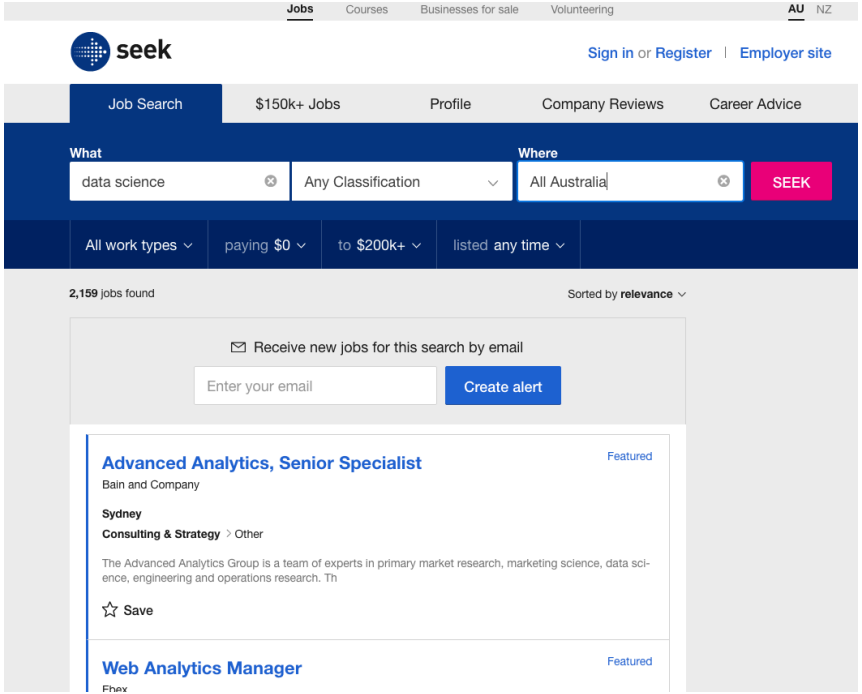
Different kinds of data scientists

- Data businesspeople
  - Focused on organizations and how data projects yield profits
  - Self-rated as leaders and entrepreneurs
- Data Creatives
  - Broad skill-set, from extracting data, to integrating it, to performing statistical or other advanced machine learning analyses, to creating compelling visualizations, to making analyses scalable and dynamic
- Data Developers
  - Focus on the problems of managing data – how to get it, store it, and learn from it
- Data Researchers
  - Academic researchers who have started in one field and have migrated to data science; deep knowledge of statistics and machine learning

Skills and roles



What's the demand?



Data sources

www.kofax.com

Obtaining data

Drawing a Map from Pub Locations with the Matplotlib Basemap Toolkit

Based on notebook from Ramiro Gómez

In this notebook we will draw a map of Britain and Ireland from location data using the Matplotlib Basemap Toolkit. The data points that will be drawn are pub locations extracted from OpenStreetMap.

You need to download and extract the Points Of Interest (POI) for pubs in the world as a CSV file called *pubs.csv*.

The map was chosen to limit it to Britain and Ireland where POI coverage seems quite comprehensive and there are a lot of pubs. Who could have thought?

Next we load the required libraries and define a function that checks whether a given location tuple is within the given bounding box.

In [2]:
```
%matplotlib inline

import pandas as pd
import matplotlib.pyplot as plt

from mpl_toolkits.basemap import Basemap


def within_bbox(bbox, loc):
    """Determine whether given location is within given bounding box.

    Bounding box is a dict with ll_lon, ll_lat, ur_lon and ur_lat keys
    that locate the lower left and upper right corners.

    The location argument is a tuple of longitude and latitude values.
    """

    return bbox['ll_lon'] < loc[0] < bbox['ur_lon'] and bbox['ll_lat'] < loc[1] < bbox['ur_lat']
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call last)
<ipython-input-2-b8b3cb284875> in <module>()
      4 import matplotlib.pyplot as plt
      5
----> 6 from mpl_toolkits.basemap import Basemap
      7
      8

ImportError: No module named 'mpl_toolkits.basemap'
```

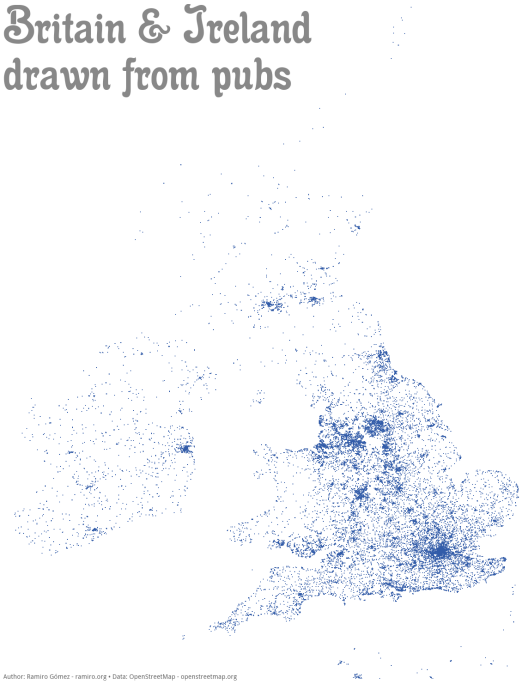
The next statements load the pub dataset into a Pandas DataFrame, and print the length and the first few entries.

In [3]:
```
df = pd.read_csv('world_pubs.csv')
print(len(df))
df.head()
```

```
100441
```

Out[3]:

	105	-0.1432091	51.5236810	pub	The Green Man
0	108042	-0.135599	51.523544	pub	The Jeremy Bentham
1	282706	-0.725178	51.035030	pub	Kings Arms



Visualization – from tables …

http://www.global-migration.info

Immigration (in), emigration (out) and net migration flows for 196 countries in 2005–10 (in 1,000s)

The estimates capture the number of people who permanently changed their country of residence over the five-year period 2005 to 2010 and thus reflect movements over a longer time period than currently published statistics.

Country	In	Out	Net
EUROPE			
Albania	31	79	-48
Austria	214	54	160
Belarus	60	110	-51
Belgium	215	15	200
Bosnia & Herzegovina	20	30	-10
Bulgaria	34	84	-50
Croatia	37	27	10
Cyprus	45	1	44
Czech Republic	241	0	240
Denmark	109	19	90
Estonia	4	4	0
Finland	73	0	72
France	752	251	500
Germany	1330	780	550
Greece	212	58	154
Hungary	84	9	75
Iceland	13	2	10
Ireland	167	67	100
Italy	2007	8	1999
Latvia	0	10	-10
Lithuania	0	36	-36
Luxembourg	43	0	42
Macedonia	18	16	2
Malta	5	0	5
Moldova	7	179	-172
Montenegro	18	20	-3
Netherlands	297	247	50
Norway	171	0	171
Poland	93	38	55

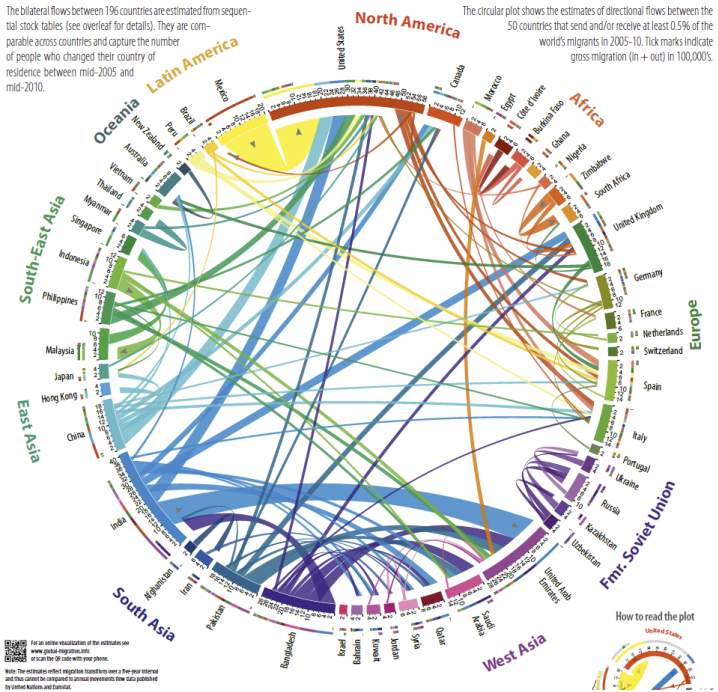
Country	In	Out	Net
Ukraine	386	426	-41
United Kingdom	1722	700	1021
AMERICA			
Argentina	74	273	-200
Aruba	4	0	4
Bahamas	6	0	6
Barbados	2	2	-1
Belize	6	7	-1
Bolivia	28	193	-165
Brazil	5	506	-502
Canada	1392	293	1098
Chile	101	71	30
Colombia	20	139	-120
Costa Rica	119	43	75
Cuba	0	190	-191
Dominican Republic	65	205	-140
Ecuador	139	259	-120
El Salvador	3	295	-292
French Guiana	9	3	6
Grenada	0	5	-5
Guadeloupe	2	5	-4
Guatemala	5	205	-200
Guyana	3	43	-40
Haiti	1	241	-240
Honduras	1	101	-100
Jamaica	2	102	-100
Martinique	2	4	-2
Mexico	123	1926	-1803
Netherlands Antilles	11	3	8
Nicaragua	0	200	-200

Country	In	Out	Net
Venezuela	111	71	40
Virgin Islands	0	3	-4
AFRICA			
Algeria	55	195	-140
Angola	83	0	82
Benin	79	28	50
Botswana	38	19	18
Burkina Faso	263	387	-124
Burundi	370	0	370
Cameroon	35	53	-18
Cape Verde	3	20	-18
Central African Republic	39	34	5
Chad	74	149	-75
Comoros	0	10	-10
Côte d'Ivoire	206	565	-359
Congo DR	72	94	-22
Djibouti	2	2	0
Egypt	50	393	-343
Equatorial Guinea	20	0	20
Eritrea	56	0	55
Ethiopia	0	296	-297
Gabon	35	30	5
Gambia	25	38	-14
Ghana	263	312	-50
Guinea	3	302	-300
Guinea-Bissau	8	18	-10
Kenya	80	268	-188
Lesotho	1	21	-20
Liberia	322	21	300
Libya	32	52	-21

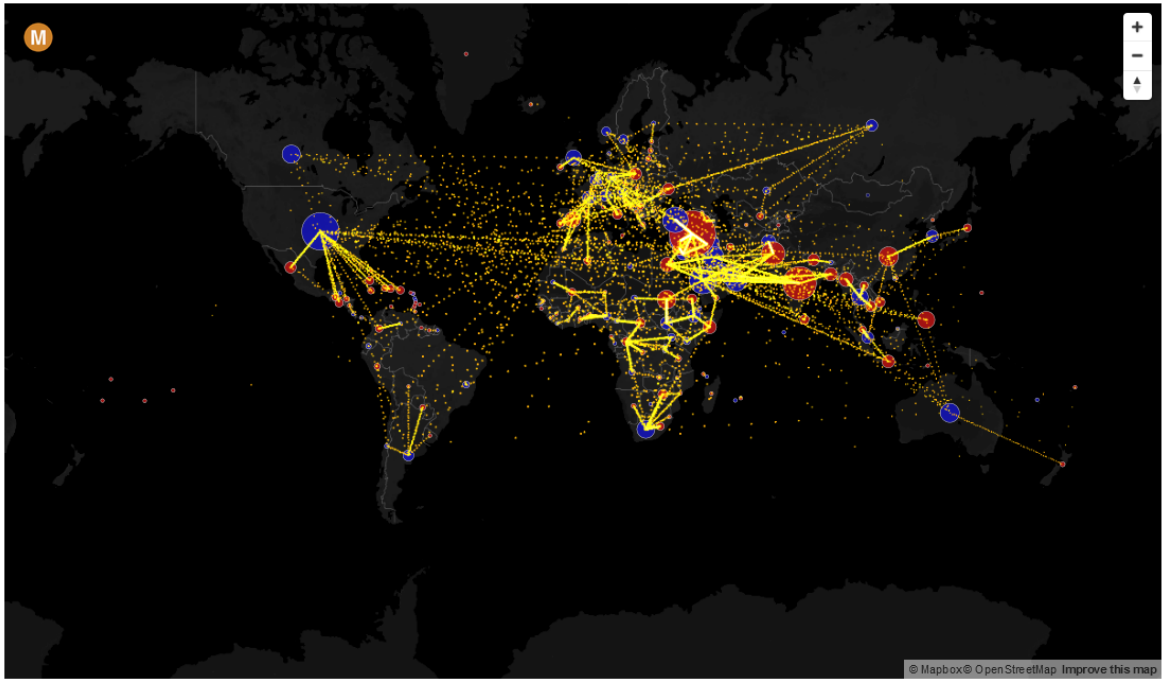
Country	In	Out	Net
Nigeria	150	435	-286
Republic of Congo	50	0	50
Réunion	3	3	0
Rwanda	62	47	15
Sao Tome & Principe	0	7	-7
Senegal	19	151	-133
Sierra Leone	75	14	60
Somalia	0	299	-300
South Africa	799	98	701
Sudan	199	62	137
Swaziland	11	17	-6
Tanzania	67	366	-299
Togo	12	17	-5
Tunisia	9	28	-20
Uganda	12	146	-134
Western Sahara	47	0	47
Zambia	42	126	-85
Zimbabwe	0	899	-900
ASIA			
Afghanistan	13	392	-379
Armenia	19	94	-75
Azerbaijan	67	13	53
Bahrain	447	0	447
Bangladesh	18	2918	-2900
Bhutan	19	2	16
Brunei	49	46	3
Cambodia	0	254	-255
China	127	2021	-1895
East Timor	0	49	-50
Georgia	1	151	-150

Country	In	Out	Net
Kyrgyzstan	0	132	-132
Laos	0	75	-75
Lebanon	87	99	-13
Macao SAR	55	4	50
Malaysia	696	610	85
Maldives	0	0	-1
Mongolia	0	15	-15
Myanmar	0	498	-499
Nepal	81	179	-99
North Korea	19	22	-3
Oman	184	31	153
Pakistan	33	2022	-1990
Palestine	0	89	-90
Philippines	30	1260	-1230
Qatar	862	5	857
Saudi Arabia	1287	230	1056
Singapore	721	0	721
South Korea	80	110	-30
Sri Lanka	1	250	-250
Syria	397	452	-55
Tajikistan	0	296	-296
Thailand	508	15	493
Turkmenistan	2	56	-55
United Arab Emirates	3077	0	3076
Uzbekistan	7	525	-519
Vietnam	19	448	-430
Yemen	77	211	-134
OCEANIA			
Australia	1164	39	1125
Fiji	2	31	-29

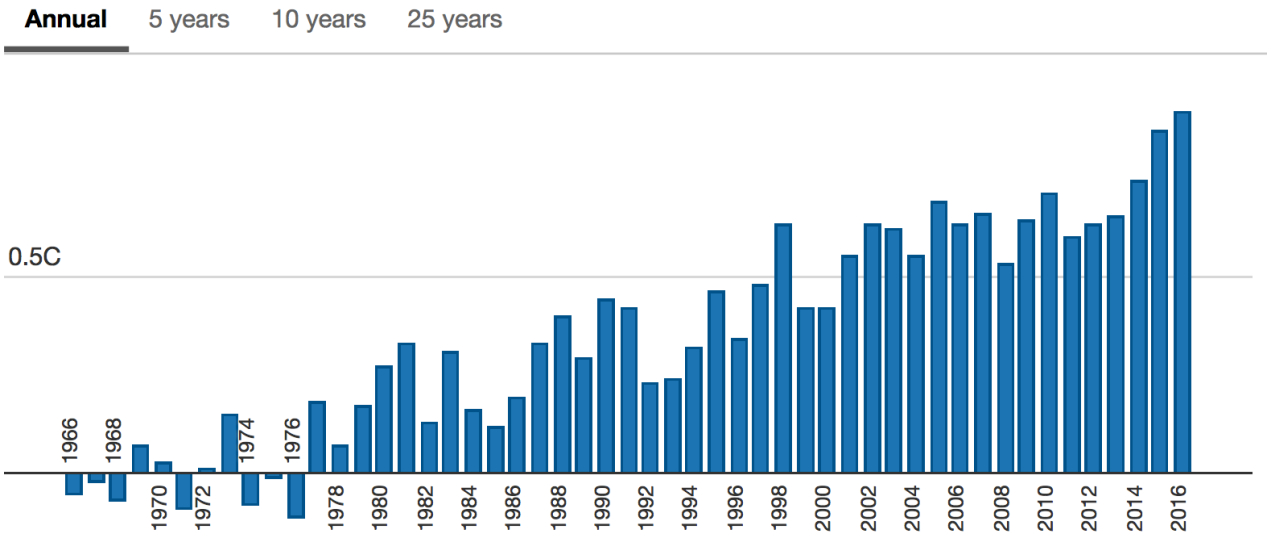
Visualization – … to figures …



Visualization – … to dynamic graphics

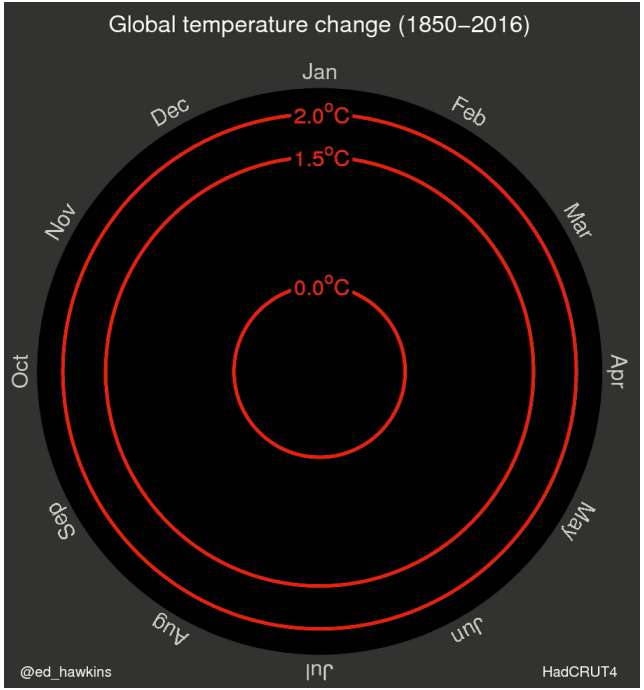


Visualization – static …

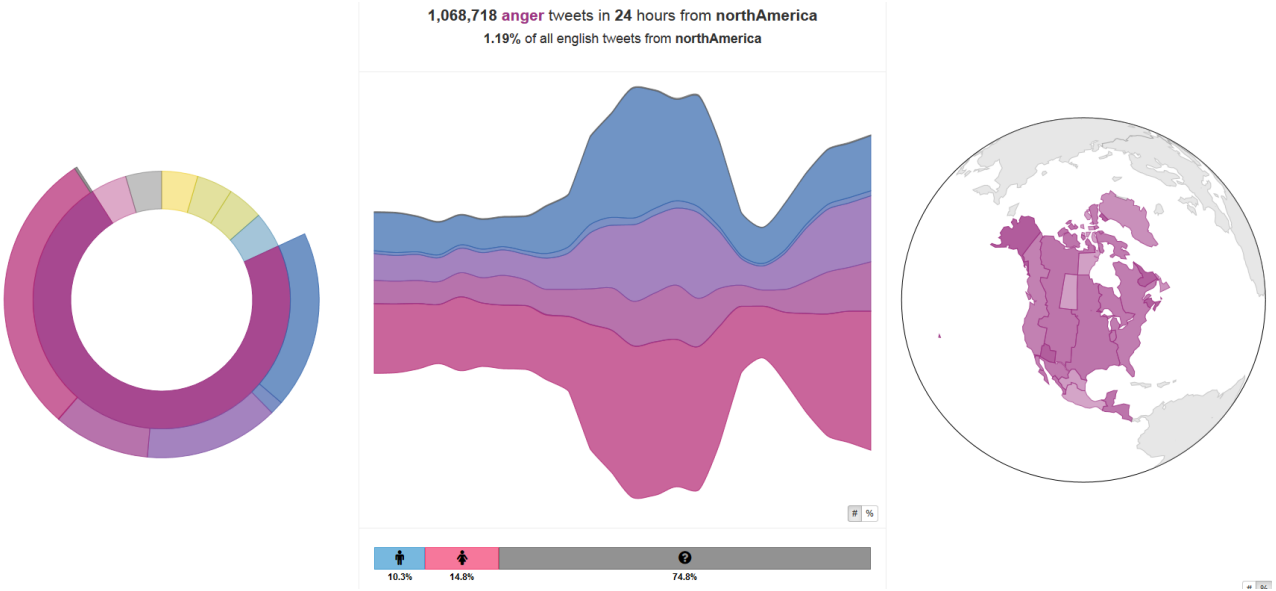


https://www.theguardian.com/commentisfree/2016/jul/28/i-spend-my-time-writing-about-the-economy-but-the-climate-data-hits-me-right-in-the-gut

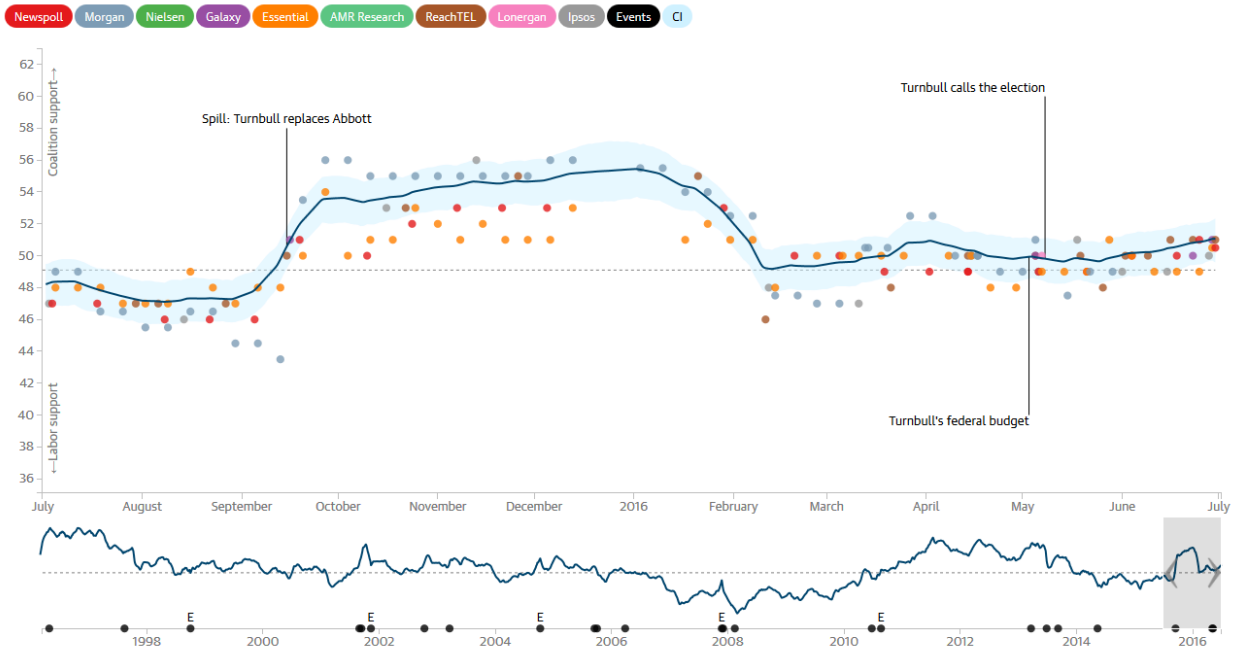
Visualization – … to dynamic



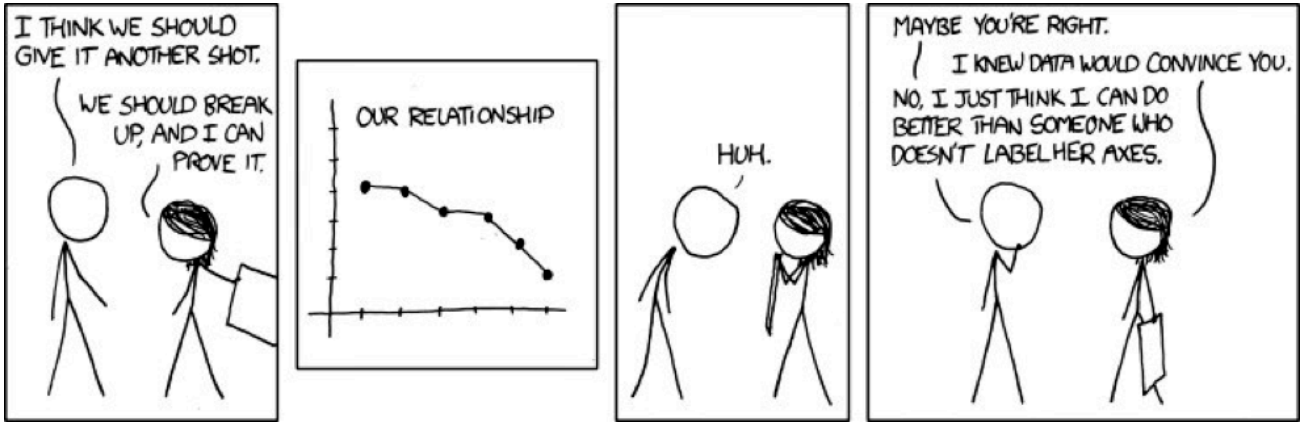
Scraping, analysis & visualization



Scraping, statistical modelling, visualization, and storytelling



Presentation and communication



Curtin University

Task for next week

- Identify a striking visualization/analysis; explain what aspect of it appeals to you

Next week

- All about data