6CS030 Big Data

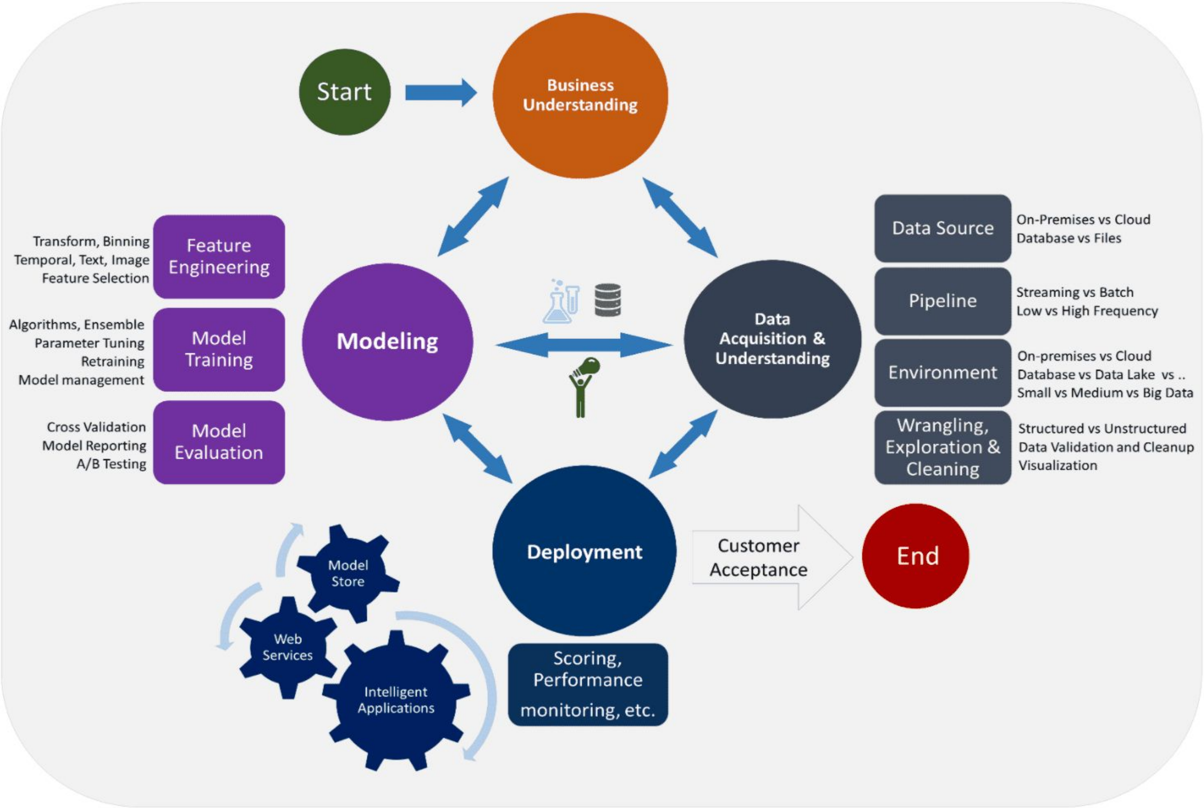
Steps in the Data Science Process

- Acquiring data
- Exploring and preprocessing data Analysing data
- Reporting insights and taking action

Data Science

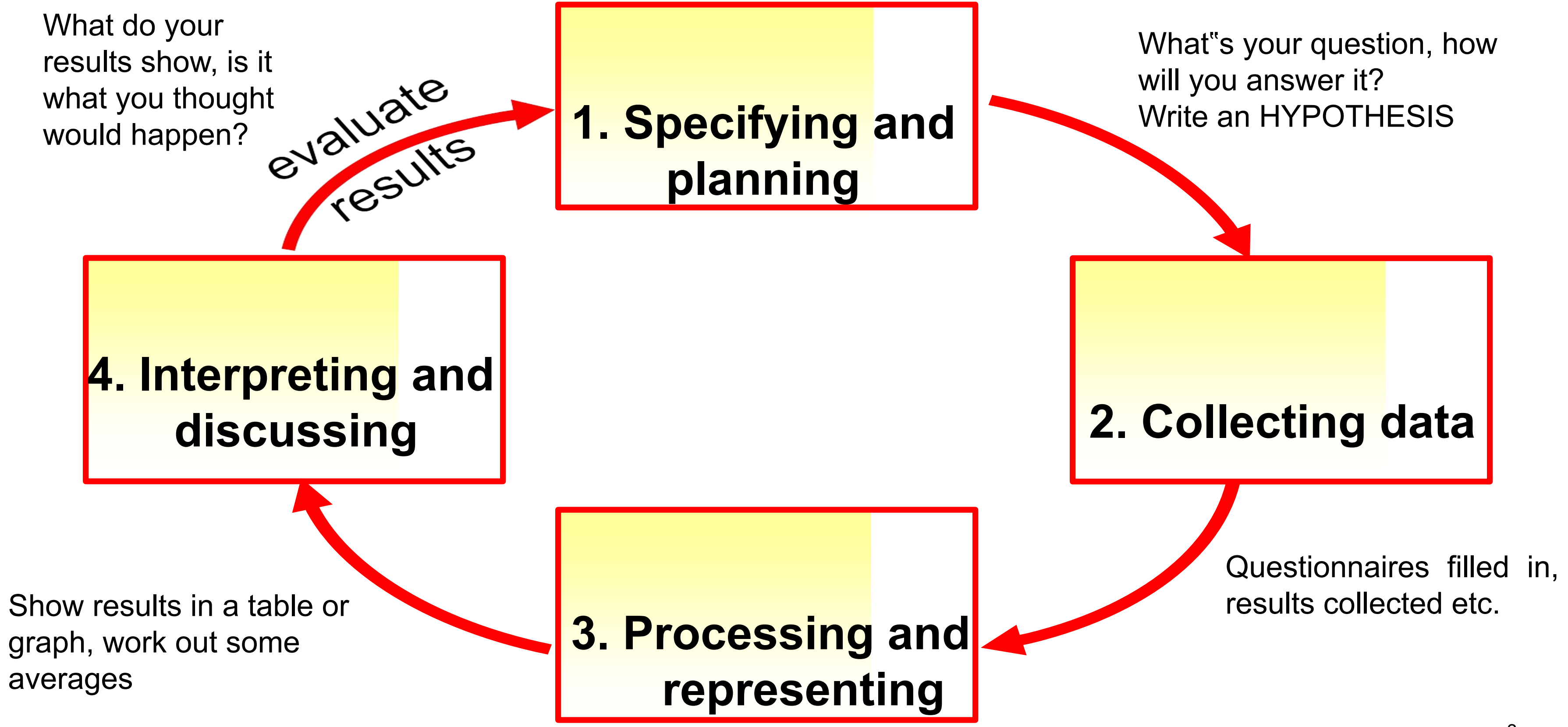
- When handling Big Data you need a process, or framework to work with
- A data-handling lifecycle
- Various approaches exist:
  - Team Data Science Lifecycle (TDSL)
  - Cross-industry standard process for data mining (CRISP-DM)
  - Knowledge Discovery in Databases (KDD)
  - AWS Data Pipeline

Team Data Science Lifecycle



https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

The Data Handling Cycle – TES



Data Handling Framework

- Many big data analytics lifecycles or workflows can be found
- The following steps are fairly typical of what is suggested:
  - Acquire data
  - Prepare or Process data
  - Analyse data
  - Report or visualise data
  - Act
- All require some sort of question/business case to be answered
- Important to track the provenance throughout the workflow
  - May have to justify decisions, so need to be able to reproduce the data processes undertaken.

Step 1: Acquiring the Data

- This involves:
  - Identifying suitable data sets
  - Where is the data?
    - Can come from many places, local and remote
    - Can be many varieties: structured and unstructured
    - Can have different velocities
  - Acquire all the available data
    - If some left out may lead to incorrect conclusions
  - Querying the data
    - SQL and query browsers help examine the data

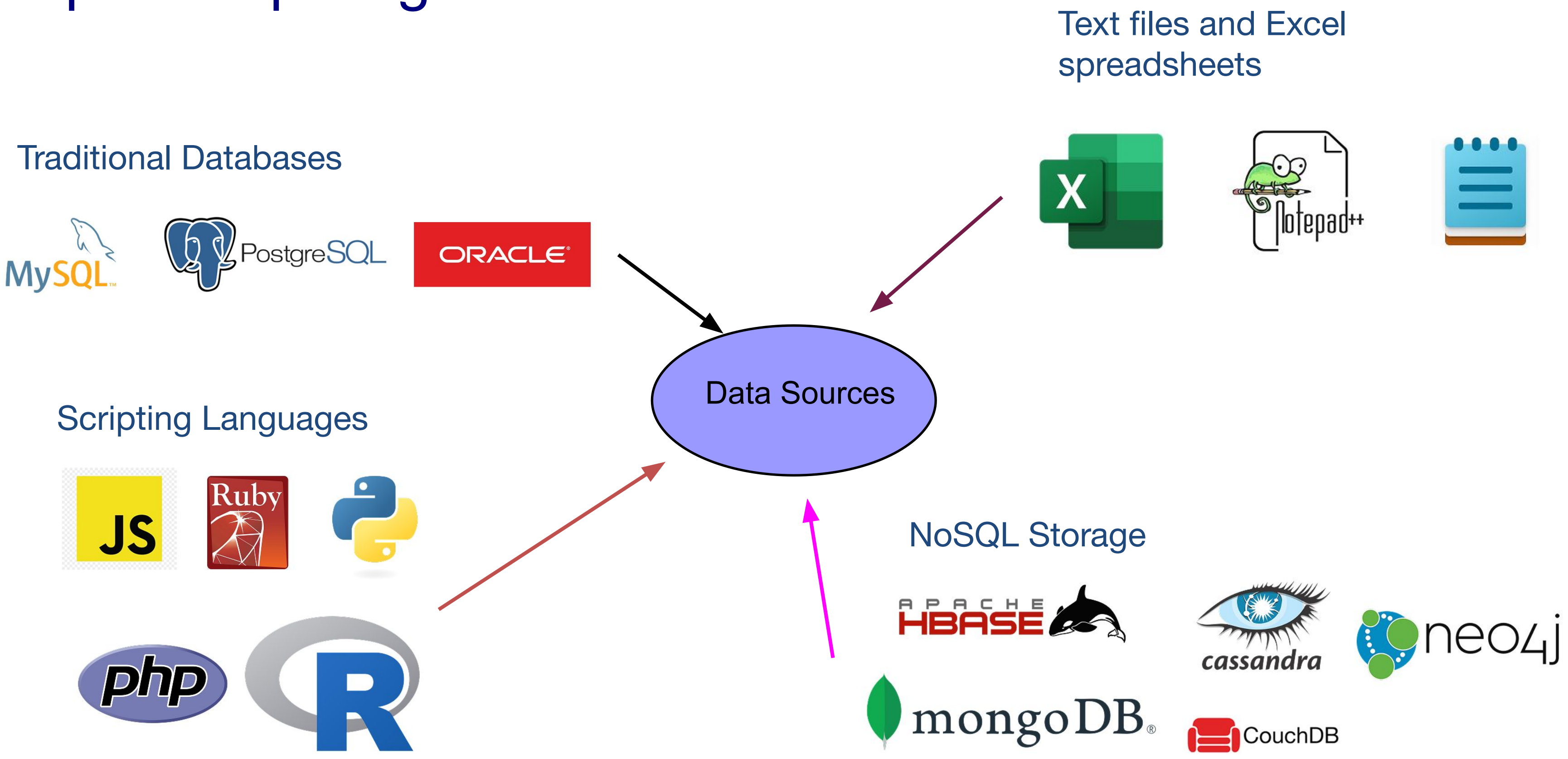
Step 1: Acquiring the Data

- Data comes from many places
  - Every minute:
    - 204 million emails are sent.
    - 200,000 photos are uploaded.
    - 1.8 million likes are generated on Facebook.
    - On YouTube, 1.3 million videos are viewed and 72 hours of video are uploaded.

24th Aug 2015 saw 1 billion users login on a single day!

- And many ways to access it!

Step 1: Acquiring the Data



Step 1: Found Data Examples

Name	URL
ONS	https://www.ons.gov.uk
EU Stats	http://ec.europa.eu/eurostat
European commission stats	http://ec.europa.eu/eurostat/data/statistics-a-z/abc
UK Government	https://www.gov.uk/government/statistics
US Government	https://www.usa.gov/statistics
Edinburgh University data share	http://datashare.is.ed.ac.uk/
List of high quality data sets	https://github.com/caesar0301/awesome-public-datasets
AW public data sets	https://aws.amazon.com/datasets/
Comparative political data set	www.cpds-data.org
Stanford – Computational Journalism lab	http://cjlab.stanford.edu/
KDNuggets – data sets	www.kdnuggets.com/datasets/mining/discovery
UK Healthcare	http://www.hscic.gov.uk/datasets
Halifax house prices	http://www.lloydsbankinggroup.com/media/economic-insight/halifax- house-price-index/
Nationwide house prices	http://www.nationwide.co.uk/about/house-price-index/headlines
Historical weather	http://www.wunderground.com/history

Step 1: Analysed Data Examples

Website	URL
Fact checking E.g., 16/11/18 NI and Brexit	https://fullfact.org/ https://fullfact.org/europe/brexit-agreement-northern-ireland/?utm_source=homepage&utm_medium=main_story
Mapping inequalities in England	https://theconversation.com/heres-what-we-learned-from-mapping-out-englands-inequalities-48562
How to know if where you live is "up and coming"	https://medium.com/@Sam_Floy/how-to-know-if-where-you-live-is-up-and-coming-fried-chicken-vs-coffee-shops-546080119f98
Find meaning in 40 years of UK political debate	https://thestack.com/iot/2015/10/14/big-data-40-years-uk-parliament-debate- complex-politics/
Evolution of US Girls Names over 100 years	https://youtu.be/qVh2Qw5KSFg
Evolution of US Boys names	https://www.youtube.com/watch?v=WQv99sEPDsw
Popular UK baby names	http://www.babycentre.co.uk/popular-baby-names
Nuclear Detonations from 1945	https://cdn.theguardian.tv/mainwebsite/2015/08/14/150813Detonations_From GAus-16x9.mp4
World's best footballers (2015)	http://www.theguardian.com/football/datablog/2015/dec/24/worlds-best-footballers-and-where-they-play-the-numbers-crunched

Step 1: Measurement Scales

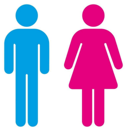
- Once the data is acquired you need to know what sort of data types it contains, since this will affect what analysis you can do

- For any statistical analysis it is important to know about the different scales of measurement:
  - **CONTINUOUS**
    Scale with a fixed and defined interval e.g. temperature or time.
  - **ORDINAL**
    Scale for ordering observations from low to high with any ties attributed to lack of measurement sensitivity, e.g. score from a questionnaire.
  - **NOMINAL with order**
    Scale for grouping into categories with order, e.g. mild, moderate or severe. This can be difficult to separate from ordinal.
  - **NOMINAL without order**
    Scale for grouping into unique categories, e.g. eye colour.
  - **DICHOTOMOUS**
    As for nominal but two categories only, e.g. male/female.

Step 1: Measurement Scales

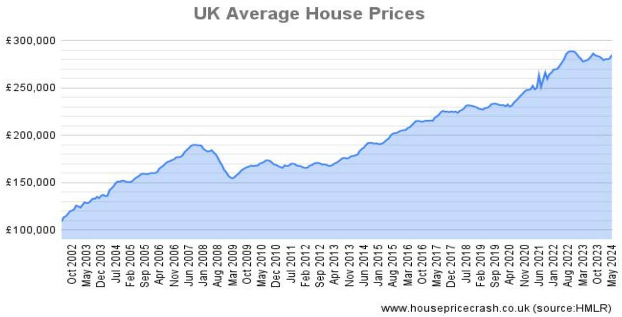
- Also important is whether the data is:
  - **CATEGORICAL (qualitative)**

    Data that represent categories, such as dichotomous (two categories) and nominal (more than two categories) observations, are collectively called categorical.



  - **NUMERICAL (quantitative)**

    Data that are counted or measured using a numerically defined method are called numerical. Examples: house prices, temperatures etc.



Step 2a: Prepare the Data

- Once you have the data, you need to understand it before building a model with it.
- Involves two sub-steps
  - Exploring the data
    - The Goal is to understand your data
      - What it means
      - Its quality and format
    - Carry out some preliminary analysis
      - Look at some samples of the data to try and understand it
      - Look for
        - Trends
        - Correlations
        - Outliers
      - Carry out some statistics

Step 2a: Prepare the Data

- Statistics include:
  - Mean: average score of the data
  - Mode: values that occur most frequently in the data set
  - Median: middle value in a data set
  - Range: measures the difference between the largest and smallest values
  - Standard deviation: a measure used to quantify the amount of variation in a set of data values
  - Count: count number of values
  - Sum: sum total of values in a dataset
  - Min and Max: minimum and maximum values

➢ These can help identify if there is something wrong in the data.
  - For example, negative numbers or percentages greater than 100 for exam scores.

➢ Will be used later too for more complex analysis.

➢ Initial visualisations can help.

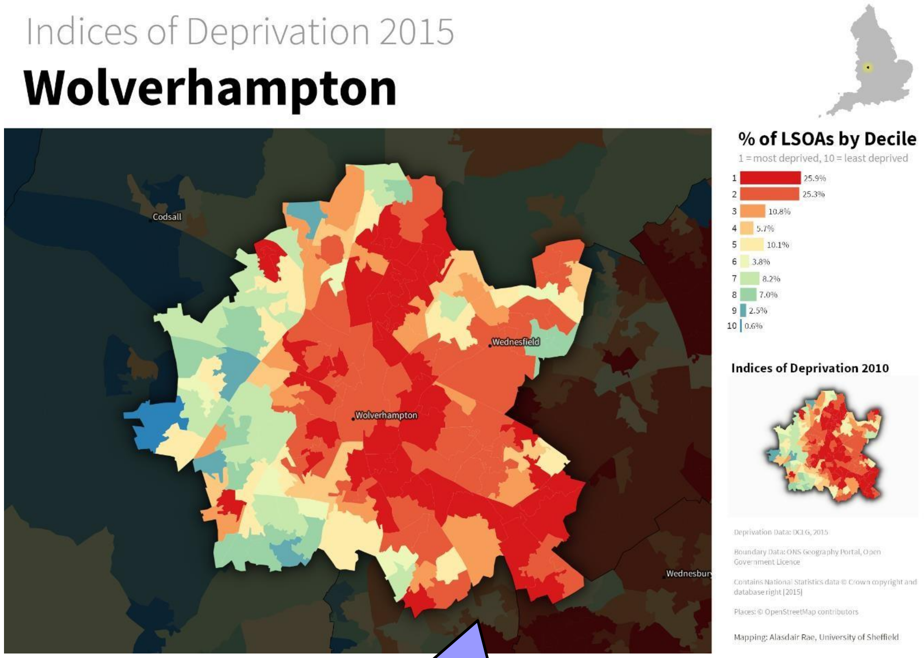
Step 2a: Visualisation Examples

Heat Maps/Infographics

- Can quickly show where the hotspots are
- E.g., English Index of Multiple Deprivation (IMD) 2015

	A	B	C	D	E	F
1	LSOA code (2011)	LSOA name (2011)	Local Authority District code (2013)	Local Authority District name (2013)	Index of Multiple Deprivation (IMD) Rank (where 1 is most deprived)	Index of Multiple Deprivation (IMD) Decile (where 1 is most deprived 10% of LSOAs)
32141	E01010556	Wolverhampton 005B	E08000031	Wolverhampton	10,619	4
32142	E01010557	Wolverhampton 005C	E08000031	Wolverhampton	18,798	6
32143	E01010559	Wolverhampton 005D	E08000031	Wolverhampton	5,245	2
32144	E01010560	Wolverhampton 005E	E08000031	Wolverhampton	6,063	2
32145	E01010435	Wolverhampton 006A	E08000031	Wolverhampton	1,820	1
32146	E01010501	Wolverhampton 006B	E08000031	Wolverhampton	9,345	3
32147	E01010502	Wolverhampton 006C	E08000031	Wolverhampton	18,976	6
32148	E01010503	Wolverhampton 006D	E08000031	Wolverhampton	4,628	2
32149	E01010482	Wolverhampton 007A	E08000031	Wolverhampton	3,255	1
32150	E01010483	Wolverhampton 007B	E08000031	Wolverhampton	1,374	1
32151	E01010484	Wolverhampton 007C	E08000031	Wolverhampton	2,087	1
32152	E01010485	Wolverhampton 007D	E08000031	Wolverhampton	1,044	1
32153	E01010458	Wolverhampton 008A	E08000031	Wolverhampton	8,494	3
32154	E01010459	Wolverhampton 008B	E08000031	Wolverhampton	10,852	4
32155	E01010460	Wolverhampton 008C	E08000031	Wolverhampton	14,203	5
32156	E01010461	Wolverhampton 008D	E08000031	Wolverhampton	22,120	7
32157	E01010566	Wolverhampton 008E	E08000031	Wolverhampton	8,875	3
32158	E01010539	Wolverhampton 009A	E08000031	Wolverhampton	16,371	5
32159	E01010540	Wolverhampton 009B	E08000031	Wolverhampton	25,622	8
32160	E01010542	Wolverhampton 009C	E08000031	Wolverhampton	14,928	5
32161	E01010544	Wolverhampton 009D	E08000031	Wolverhampton	21,448	7
32162	E01010545	Wolverhampton 009E	E08000031	Wolverhampton	19,896	7
32163	E01010554	Wolverhampton 010A	E08000031	Wolverhampton	16,411	5
32164	E01010558	Wolverhampton 010B	E08000031	Wolverhampton	14,894	5
32165	E01010562	olverhampton 010C	E08000031	Wolverhampton	6,120	2
32166	E01010563	olverhampton 010D	E08000031	Wolverhampton	5,624	2
32167	E01010567	olverhampton 010E	E08000031	Wolverhampton	5,550	2
32168	E01010478	lverhampton 011A	E08000031	Wolverhampton	860	1
32169	E01010479	lverhampton 011B	E08000031	Wolverhampton	6,696	3
32170	E01010480	verhampton 011C	E08000031	Wolverhampton	8,749	3
[illegible]	[illegible]	on 011D	E08000031	Wolverhampton	811	1
[illegible]	[illegible]	on 012A	E08000031	Wolverhampton	4,689	2
[illegible]	[illegible]	on 012B	E08000031	Wolverhampton	23,594	8
[illegible]	[illegible]	on 012C	E08000031	Wolverhampton	4,601	2
[illegible]	[illegible]	on 012D	E08000031	Wolverhampton	23,923	8

Sample raw data



More meaningful as an image?

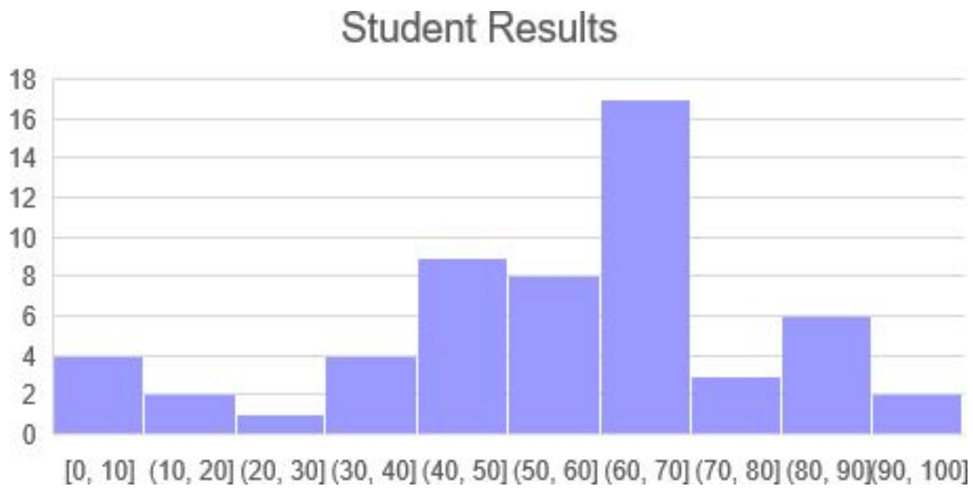
Step 2a: Visualisation Examples

Histogram

- Can show the distribution of the data and any skewness or unusual dispersion
- Given this set of student results, can you predict their overall performance?

	A	B
1	**Student Name**	**Result**
2	Student1	0
3	Student2	50
4	Student3	72
5	Student4	78
6	Student5	22
7	Student6	81
8	Student7	40
9	Student8	40
10	Student9	62
11	Student10	62
12	Student11	54
13	Student12	90
14	Student13	53
15	Student14	60
16	Student15	90
17	Student16	54
18	Student17	57
19	Student18	0
20	Student19	17
21	Student20	48
22	Student21	0
23	Student22	86
24	Student23	83
25	Student24	70
26	Student25	45
27	Student26	67
28	Student27	61
29	Student28	40
30	Student30	67
31	Student31	67
32	Student32	65

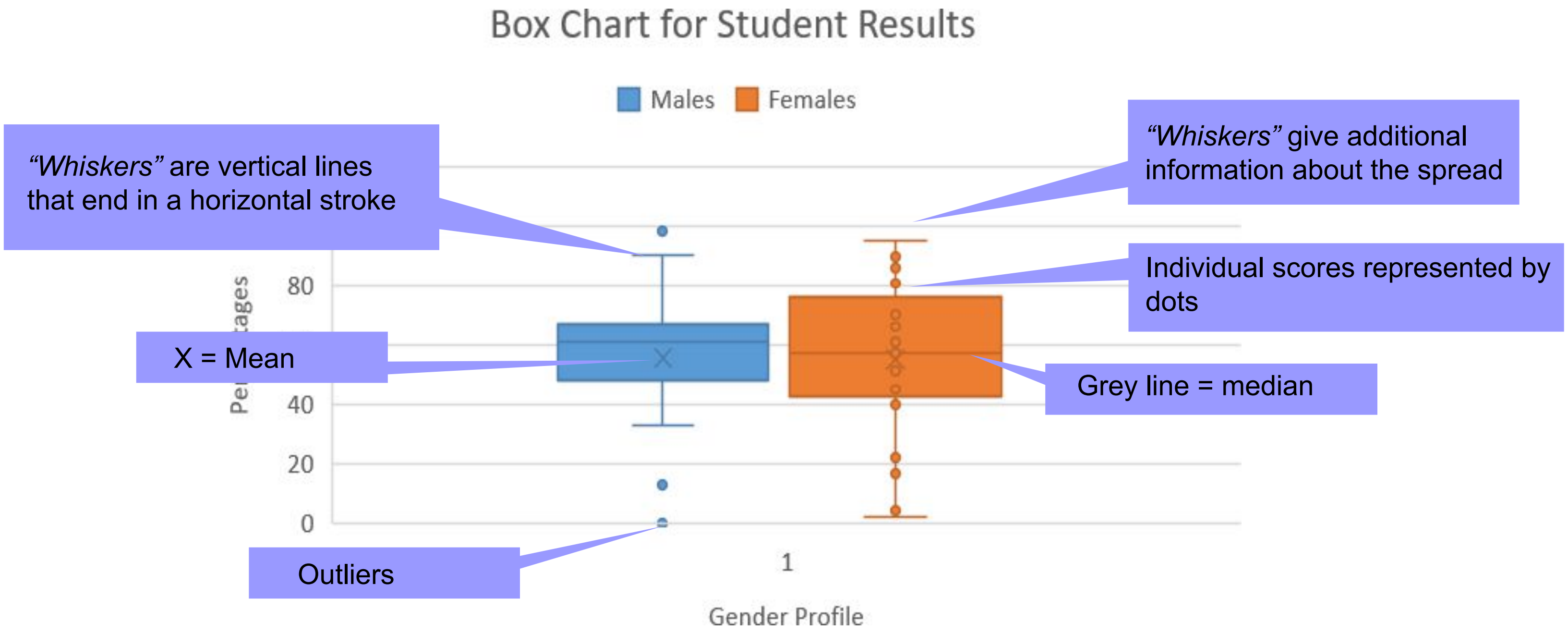
Does this make it easier?



Step 2a: Visualisation Examples

Box Plots

- Another type of plot for showing data distribution
- Useful for identifying outliers and comparing distributions
- Excel calls these *Box and Whisker* charts



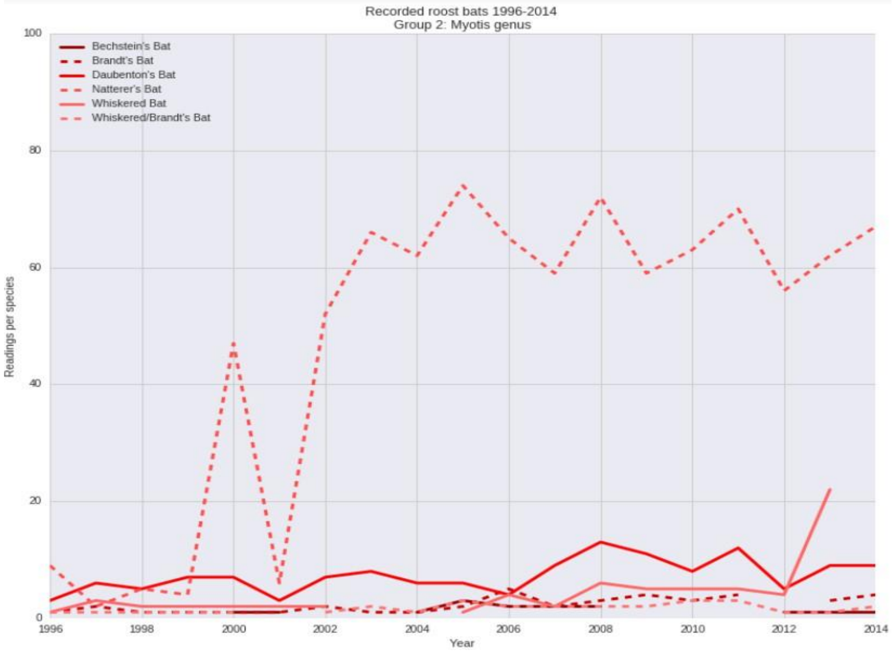
Step 2a: Visualisation Examples

Line graphs

- Useful for seeing how values in the data changes over time.
- Spikes in the data are also easy to spot
  - For example data on bats:

Species	Year	Bat	Bechstein's Bat	Brandt's Bat	Brown Long-eared Bat	Common Pipistrelle	Daubenton's Bat	Greater Horseshoe Bat	Grey Long-eared Bat	Lesser Horseshoe Bat	Lesser Noctule	Natterer's Bat	Noctule Bat	Pipistrelle
0	1996	NaN	NaN	1	10	28	3	NaN	NaN	149	NaN	9	3	58
1	1997	NaN	NaN	2	12	123	6	11	NaN	50	NaN	2	3	154
2	1998	1	NaN	1	9	195	5	10	NaN	63	NaN	5	2	190
3	1999	3	NaN	1	11	238	7	11	NaN	121	NaN	4	3	276
4	2000	1	1	1	14	210	7	12	NaN	105	NaN	47	4	258
5	2001	2	1	1	60	158	3	17	NaN	62	NaN	6	3	162
6	2002	3	NaN	2	85	257	7	20	NaN	126	NaN	52	4	191
7	2003	2	NaN	1	84	256	8	15	NaN	127	NaN	66	3	179
8	2004	2	1	1	94	309	6	16	NaN	140	NaN	62	5	142
9	2005	2	3	2	113	316	6	27	NaN	175	NaN	74	3	162
10	2006	2	2	5	110	392	4	31	NaN	160	NaN	65	3	152
11	2007	2	2	2	124	369	9	27	NaN	167	1	59	NaN	122
12	2008	1	2	3	146	365	13	26	NaN	151	3	72	2	154
13	2009	1	NaN	4	136	398	11	25	NaN	157	2	59	NaN	125
14	2010	2	NaN	3	142	394	8	31	NaN	185	2	63	2	93
15	2011	3	NaN	4	147	368	12	28	NaN	199	3	70	2	88
16	2012	6	1	NaN	109	347	5	36	NaN	175	NaN	56	NaN	68
17	2013	4	1	3	115	327	9	36	1	238	NaN	62	2	80
18	2014	12	1	4	107	285	9	38	3	233	NaN	67	4	62

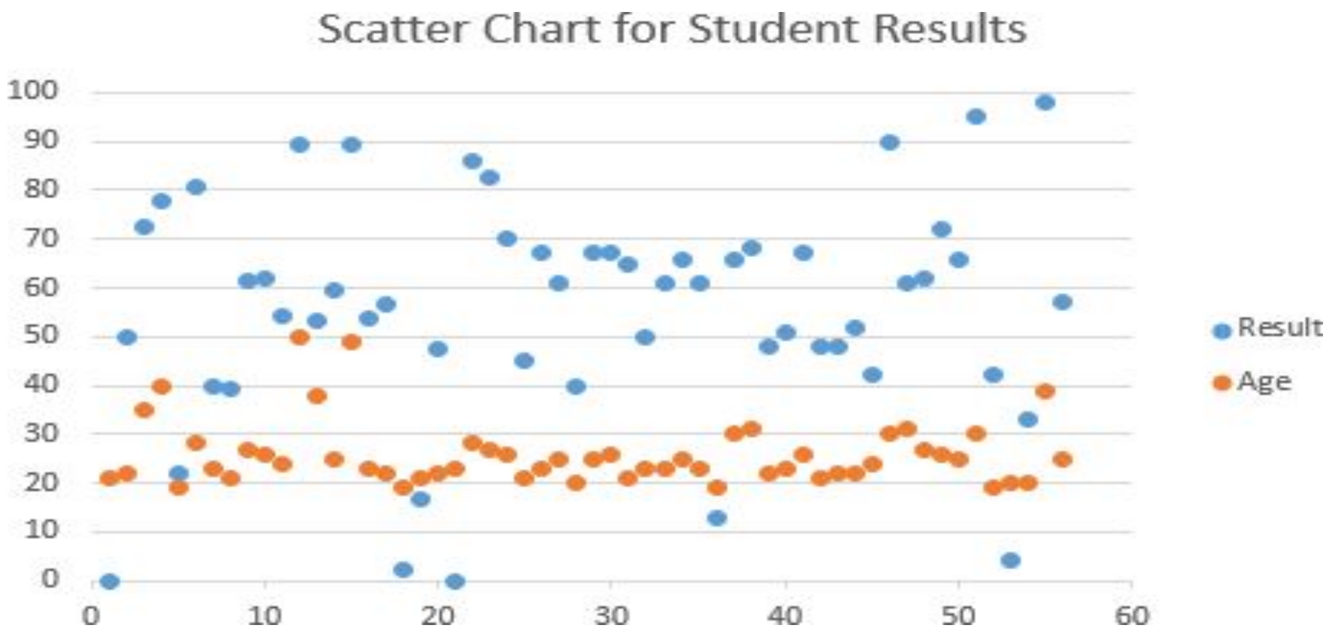
Easier to visualise using line chart:



Step 2a: Visualisation Examples

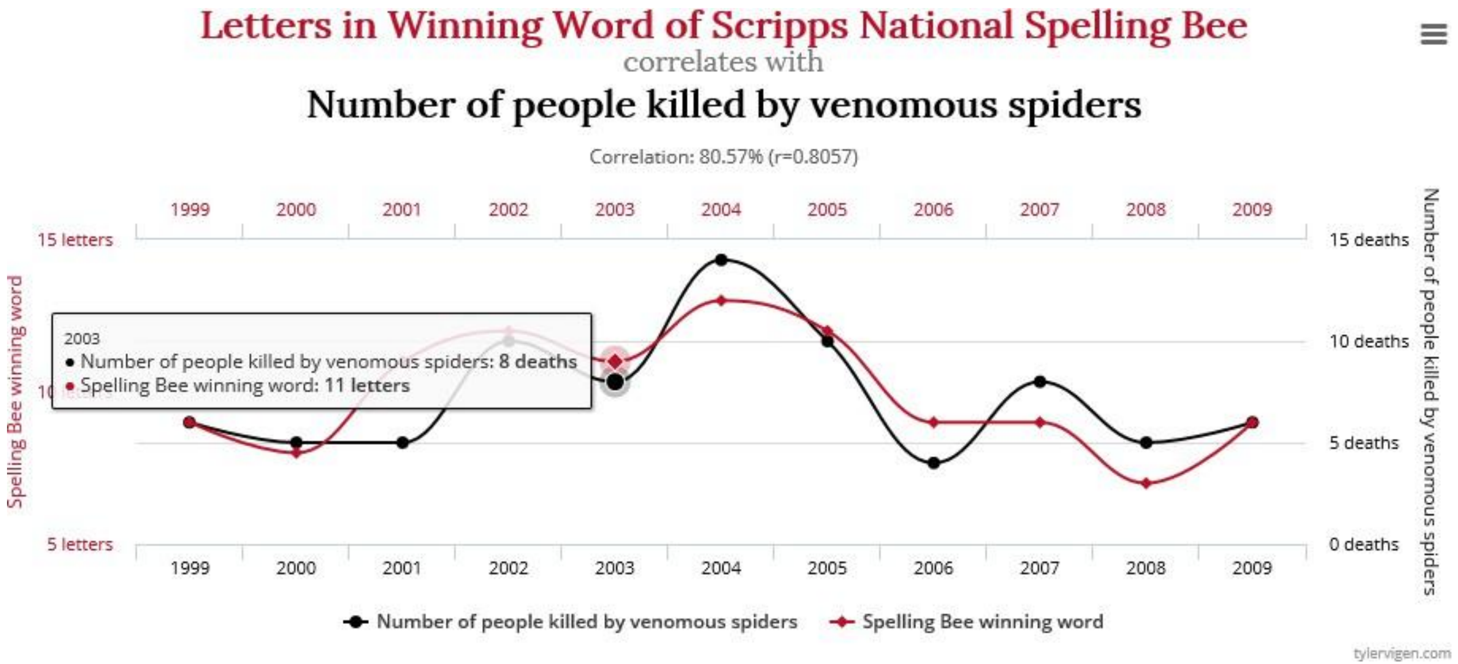
Scatter plots

- Can show correlation between two variables
- Is there any correlation between results and a student's age?



Correlation

- Beware: correlation does not always imply causation!



http://tylervigen.com/spurious-correlations

Step 2b: Prepare Data

- After the exploratory analysis you need to prepare the data
- The raw data acquired is not usually in the format you want
- Integration
  - You may need to merge data from multiple sources
    - Are they all using the same formats, naming conventions?
    - E.g., date formats can vary in DBMSs:

      Oracle"s *DD-MON-YY* v"s MySQL *YYYY-MM-DD*
- Two main goals in data pre-processing:
  - Clean the data to address data quality issues
  - Transform the raw data to make it suitable for analysis

Step 2b: Pre-process the Data

- Clean the data
  - Garbage In Garbage Out
- Real-world data is messy!
  - Inconsistent values
    - Customer with 2 different addresses
  - Duplicate records
    - Customer with more than one record
  - Missing values
    - E.g., missing a customer‟s age which is needed for a demographic study
  - Invalid data
    - Postcode in the wrong format
  - Outliers
    - Values that are much higher/lower than expected

Step 2b: Prepare Data

- You will have to decide and document whether to:
  - Remove data with missing values
  - Merge duplicate records
    - Need to decide what to do if they have conflicting values?
    - E.g., keep the latest value
  - Generate best estimates for invalid values
    - E.g., estimate a missing employee‟s age from their length of service
  - Remove outliers
    - Could be real values that were just extremes on occasions

Step 2b: Cleaning Data

Common types of data errors (Kim 2003):

Dirty data error	Description
Validity	Do values match constraints? Are values in range?
Accuracy	Are values accurate, e.g., compare to reference lookup? Correct spelling? Correct capitalisation?
Completeness	Are all mandatory fields present, that is, not null?
Consistency	Are the same type values in different cells in the same column, e.g., names, numbers?
Uniformity	Are formats the same for the same fields, e.g. dates? Is white space present? Are Units of Measurement the same?

Handling dirty data:

Approach	Outcome
Fix it	Replace incorrect value with correct value Insert missing values
Remove it	Delete value or group of values => impact?
Replace it	Put marker in dataset indicating it is an inappropriate value
Leave it	Note and accept any dirty data

Don‟t forget to document what has been added/deleted/changed

Step 2b: Cleaning Data

- Below is a sample student data set for the fictitious Borchester University.
- What issues are there with this data?

studentNo	studentName	Gender	DOB	avgMark
1234	Johnny Rick Philips	b	12-Jan-08	59
2345	Sheila Hebden Lloyd	f	22/04/1957	85
3355	Perks, Jamie	m	09/25/05	
4455	will grundy	m	23-Feb-81	105
6541	Madikane, Kate	F	02/08/1978	-55

Mixed case, inconsistent first and surname order

Inconsistent case, b = boy?

Inconsistent date formats

Percentages? Minus or >100 ok?

Step 2b: Prepare Data

- Knowledge about the application the data came from is important
  - How the data was collected; intended use
  - Called the *domain knowledge*

- The domain knowledge helps make informed decisions on how to handle incomplete or incorrect data
  - For example, if there were no integrity checks, there is more likely to be rogue data

- Getting data into shape is called many things:
  - Data munging
  - Data wrangling
  - Data pre-processing
  - Data manipulation

Step 2b: Prepare Data – Operations

Types of operations include:

- Dimensionality reduction
  - E.g., change 3D model to 2D
- Data manipulation
  - Shaping the data to fit new requirements
  - Filtering the data – may not need everything
- Transformation
  - To reduce noise and variability
  - One example is aggregation.
    - This generally results in data with less variability.
    - For example, daily sales figures may have many peaks and troughs. Aggregating values to weekly or monthly sales figures will result in similar data

Step 2b: Prepare Data – Operations

- Feature selection
  - Remove redundant or unnecessary features.
  - Combine features, such as adding salary and commission to create a total salary.
  - Creating new features, such as adding an applicant's education level to a loan approval.

- Scaling
  - Involves changing the range of values to be between a specified range.
  - Done to avoid large values dominating the results.

Step 3: Analyse the Data

- Involves several steps
  - Select analytical techniques
  - Build Models
    - This may need several iterations and involve going back to Steps 1 and 2.
      E.g., if need further data or need to package the data using a specific format
    - This involves taking the input data from the previous steps and generating an output model

- Validate Model

Step 3: Analyse the Data

- Categories of Analysis Techniques
  - Classification
    - Goal: predict category
    - E.g., predict weather as sunny, rainy or cloudy
  - Regression
    - Goal: predict numerical value
    - E.g., predict price of a stock
  - Clustering
    - Goal: organise similar items into groups
    - E.g., organise customers to seniors, adults and teenagers
  - Association Analysis
    - Goal: find rules to capture associations between items
    - E.g., if you buy one item, what else might you buy
  - Graph Analytics
    - Goal: Use graph structures to find connections between entities
    - E.g., explore spread of a disease by analysing hospital records

Step 3: Analyse the Data

- Before moving on you need to evaluate the results
- For example,
  - For classification and regression you could compare the predicted values against some correct values.
  - For clustering do the groups make sense for the application?
  - For association analysis and graph analysis, further investigation is needed to check whether the results are correct. E.g., does what your model predict actually happen?

Step 4: Communicate Results

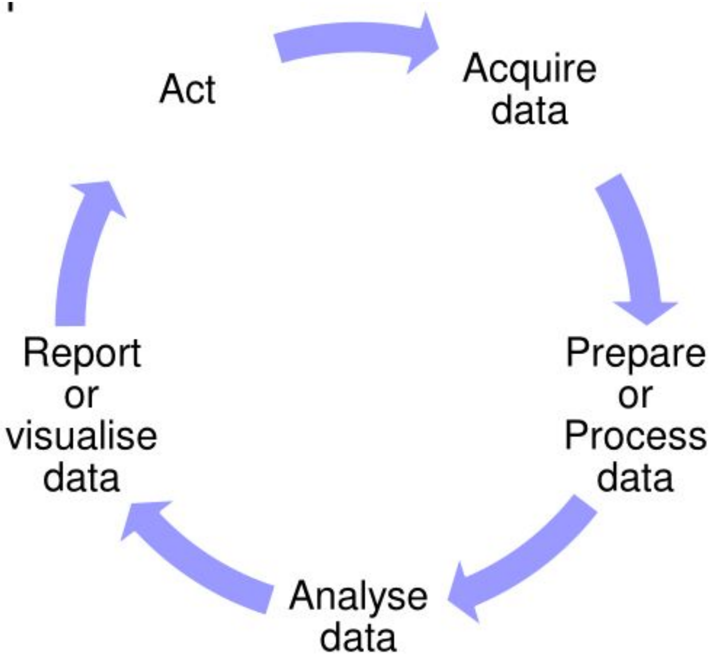
- Evaluation of the analytical results
- Usually involves Visualisation techniques
  - A picture is worth a 1000 words!
- Involves interpreting the results, summarising or visualising
- May need to acknowledge the source of data if the licence agreement requires this.
- Various tools exists to help with the visualisation:
  - Power BI
  - Tableau
  - Google Charts
  - R and Python
  - Many others…..

Step 5: Apply Results

- There should be some purpose to the exercise
  - The main reason why data science is needed
- Involves reporting insights from the analysis and determining actions
- May involve helping business needs
- Need to determine next steps:
  - Is extra analysis needed to yield better results?
  - Any data needs revisiting?
  - Any further opportunities to explore?
- Remember: big data and data science are only useful if the insights can be turned into actions and the actions are carefully defined and evaluated.

Summary

This lecture has looked at a variety of techniques in the data science process:



Note: these steps should be an iterative process!