# Review Outline

## W1

### What is data science

The study of data and extract the knowledge of data(extract meaningful insights from data.)

### Drew Conway’s Venn diagram



### Usefulness of machine learning

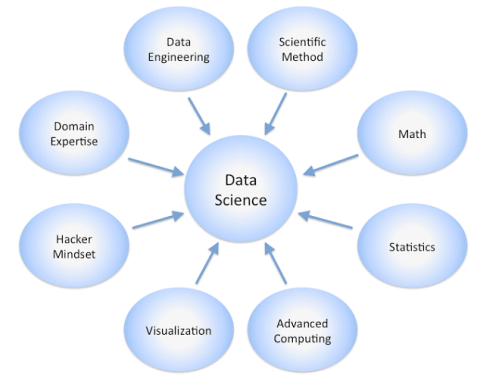
* Human expertise is not available
* Many solutions need to be adapted automatically
* Humans are expensive for the work
* Situation changes overtime
* large amounts of data

### Different components of a data science process

Standard Value Chain:

* Collection: getting the data
* Engineering: storage and computational resources across full lifecycle
* Governance: overall management of data across full lifecycle
* Wrangling: data preprocessing, cleaning
* Analysis: discovery (learning, visualisation, etc.)
* Presentation: arguing the case that the results are significant and useful
* Operationalisation: putting the results to work, so as to gain benefits or value

### Differentiate data science from other related disciplines



## W2

### Essentials for coding in Python for data science

### Interpret given Python codes

### Why we study Python and its importance for data science

## W3

### The importance/power of data visualization

### Approaches for data visualisation, explain where each approach is appropriate to be used

### Concepts in descriptive statistics

### More sophisticated group-by operations in Python

* Categorical-Nominal:
  + Discrete numbers of values, no inherent ordering
  + E.g., country of birth, sex
* Categorical-Ordinal:
  + Discrete number of states, but with an ordering
  + E.g., Education status, State of disease progression
* Numeric-Discrete:
  + Numeric, but the values are enumerable
  + E.g., Number of live births, Age (in whole years)
* Numeric-Continuous:
  + Numeric, not enumerable (i.e., real numbers)
  + E.g., Weight, Height, Distance from CBD

## W4

### Open data and linked open data

### How to access to new data sources through APIs

### How different APIs work

### Data quality problems in datasets

### Data wrangling commands in Python

## W5

### What are models and predictive models

### Analyse predictive models in different examples

### How to evaluate predictive models

### How to estimate linear regression model

### linear regression and polynomial regression in Python

## W6

### Overfitting and underfitting of different models

### Bias and variance trade-off

### “No Free Lunch Theorem”

### What are ensemble models

## W7

### Differentiate between classification and regression models

### How decision trees and regression trees work

### How random forest works

### How k-means clustering works

### Confusion matrix and prediction accuracy

### Different classification metrics

## W8

### Essentials for coding in R for data science

### Explain and interpret given R commands

### Apply R commands for data wrangling, visualisation, exploration and analysis

## W9

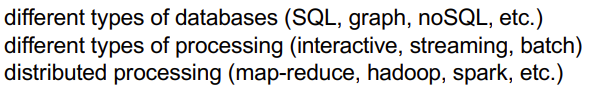
### Characterising big data: Volume, Velocity, Variety, Veracity

### What is metadata? different types of metadata

### Growth laws related to big data: Moore’s law, Koomey’s law, Bell’s Law and Zimmerman’s Law

## W10

### Processing big data



### What is deep learning

## W12

### Confidentiality and privacy

### Regulatory compliance

### Data management