# Week 1 Lecture 1

Unit Coordinator - Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

WESTERN SYDNEY UNIVERSITY

# What is Data Science?

Data Science is;

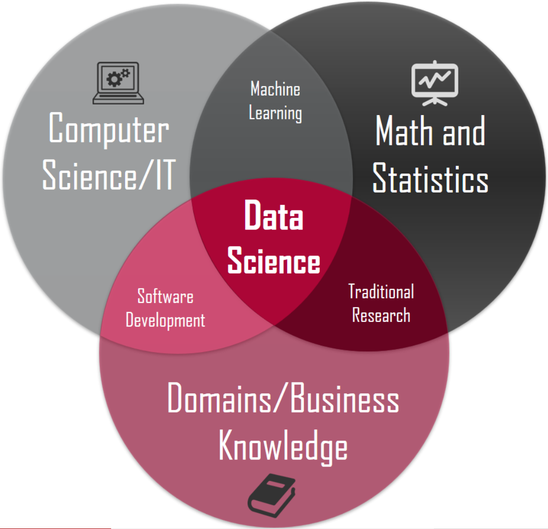
- Statistics?
- Machine Learning?
- Big Data?

Data Science is the extraction of knowledge from large volumes of ***structured*** or ***unstructured data***. It has application in Science, Business, Social Science, wherever data is collected. . .

We will look at some areas of Data Science.

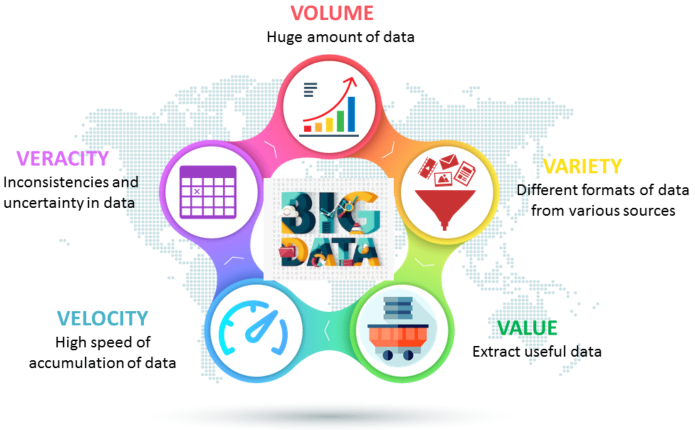
# Data Science

From *towardsdatascience.com*



# Big data

From *eureka.co*



VOLUME
Huge amount of data

VARIETY
Different formats of data
from various sources

VALUE
Extract useful data

VELOCITY
High speed of
accumulation of data

VERACITY
Inconsistencies and
uncertainty in data

# Jobs in Data Science

From *indeed.com*

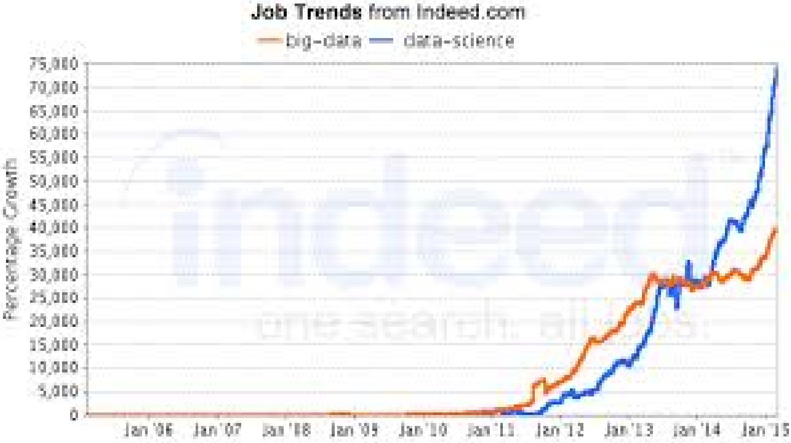


Figure 3:

# Jobs in Big Data Analytics

From *indeed.com*

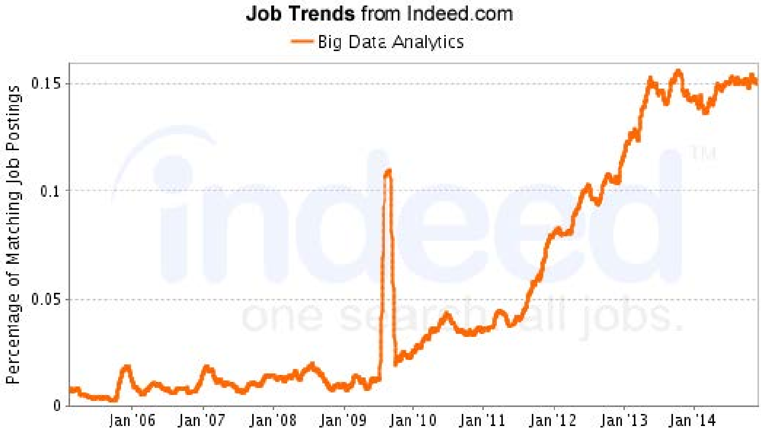


Figure 4:

# Data Science or Statistics

***Data Science*** uses a blend of methods from ***statistics, computing*** and ***machine learning*** to extract information from data. It is **LESS** concerned about ***p-values*** and ***hypothesis testing***. But these have uses.

Typical data has multiple measurements (***variables***) on several ***observations***.

For example, Fisher's Iris data - measured the species, sepal lengths and widths and petal lengths and widths for 150 iris flowers.

WESTERN SYDNEY
UNIVERSITY

# Types of Data

**Structured data**

- ***Quantitative*** or ***Numeric data*** - height, weight, salary, sales dollars
- ***Qualitative*** or ***Factor/Categorical data*** - Ethnicity, product code, hair colour

Structured data is usually a series of measurements on distinct observational units, and **can be arranged as a table**.

**Unstructured data**

- Images - Flickr
- Videos - Youtube
- Text - eg. Twitter

Unstructured data usually **DOES NOT** look like a nice table, but contains information non-the-less. Often the first step in analysing unstructured data is to extract information to make structured data.

# Big Data

You’ve no doubt heard the hype around ***“Big Data”***.

Data from multiple sources that can be linked to form a complete (or extensive) picture of an individual?

This is probably harder than it looks.

However, many areas now have access to large amounts of data or linked “big data”.

# Examples of Data Science problems

- Predict the outcome of marketing campaigns
- Model (anticipate) demand for a product or service
- Model (understand) the relationship between stress and working conditions
- Recommend similar products to previous purchases
- Find likely fraudulent insurance claims
- Understand structures in groups or networks

# Supervised versus Unsupervised

Data Science problems generally split into ***supervised*** and ***unsupervised*** problems.

***Supervised learning*** involves data where each observational unit has one special variable - the ***output/outcome***.

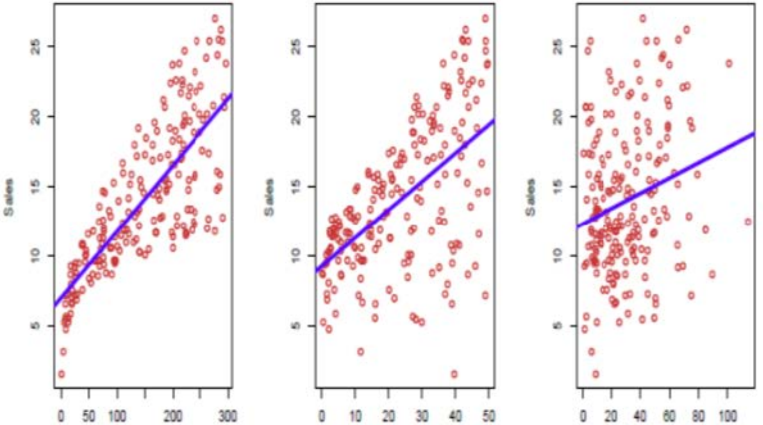
- Patients survive a treatment or not
- Customer spend on a certain product

***Unsupervised learning*** **DOES NOT** have a special variable, we are interested in discovering patterns.

- **Fraud** - finding observations that don't fit the usual pattern
- **Segmentation** - grouping a market into more homogeneous groups

# Supervised Learning

- In supervised learning, we have a ***response*** or ***outcome***.
- We are interested in understanding or predicting the relationship between the output and several inputs.



# Supervised Learning

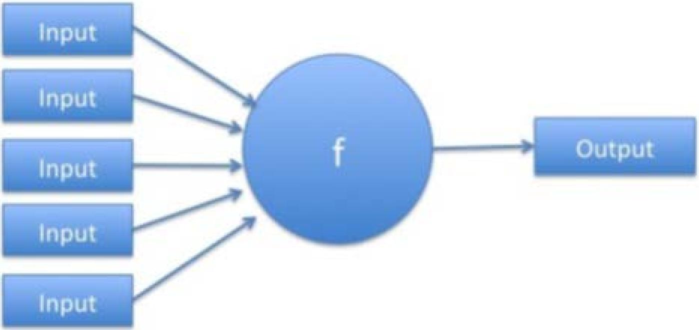


Figure 6:

We want to learn about $f$ from a sample of inputs and outputs

# Supervised Learning

Starting point:

- Outcome measurement $Y$ (also called ***dependent variable***, ***response***, ***target***).
- Vector of $p$ predictor measurements $X$ (also called ***inputs***, ***regressors***, ***covariates***, ***features***, ***independent variables***).
- In the ***regression problem***, $Y$ is ***quantitative*** (e.g price, blood pressure).
- In the ***classification problem***, $Y$ takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have ***training data*** $(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \ldots, (\mathrm{x}_n, \mathrm{y}_n)$. These are ***observations*** ( ***examples***, ***instances***) of these measurements.

# Supervised Learning

- In Mathematical terms we would write;

$$E(Y) = f(X_1, X_2, ..., X_p)$$

- $y$ is the ***output*** and the $x'$s are the ***inputs***.
- We **DO NOT** model $y$ but its ***expected value***.
- Even for the same set of inputs, the output may vary; ***measurement error***, ***random variation***, etc.
- So, $E(y)$ is the expected value or average value, for a given set of inputs.
- Any difference between the expected value and an observed value is called ***noise***.

# Supervised Learning

The simplest form of supervised learning is ***simple linear regression***

$$E(Y) = a + bX$$

There is one input $(X)$ and one output, and two parameters, $a$ and $b$. A sample of data; input/output pairs, would be used to estimate the parameters.

The results could then be used to;

- Make inferences about the relationship between input and output.
- Make prediction of future outputs for a given input value.

# Bias and Variance

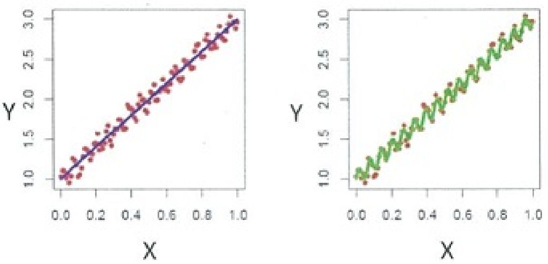


Figure 7:

Both graphs have the same $Y$ and $X$. The left graph has a ***very simple smooth*** $f$ , the right a ***more complex*** $f$ . Modelling must choose the right form of $f$ as well as fitting the actual $f$ (***estimating parameters***).

# Bias and Variance

Generally, fitting complex functions results in ***more variance*** - there is more uncertainty around the fitted parameters.

Fitting simple functions can result in ***more bias*** - there is systematic differences between the fitted and true functions.

Both bias and variance contribute to ***prediction accuracy***.

# Prediction Accuracy and Interpretation

***Prediction Accuracy*** refers to how closely we can predict a future observation.

Often it must be estimated from the same sample that was used to fit the function (**Not good practice**).

- ***Training data set*** is used to build the model
- ***Validation data set*** is used to validate the model accuracy
- ***Testing data set*** is used to measure prediction accuracy.

More complex functions sometimes have better prediction accuracy but the results can be hard to interpret.

# Regression vs. Classification

When the output is a ***numeric variable***, supervised learning is sometimes referred to as ***regression***. Some examples of regression methods are;

- (Simple) Linear Regression
- Generalised Linear Models
- Neural Networks

When the output is a ***class*** or ***factor variable***, supervised learning is ***classification***. Some examples of classification methods are;

- Nearest Neighbours
- Generalised Linear Models (Logistic Regression)
- Support Vector Machines

## Unsupervised Learning

When there is **NO** variable that can be considered an output or special, then ***unsupervised learning*** may be appropriate. Unsupervised learning looks for patterns amongst the input variables.

Methods can range from;

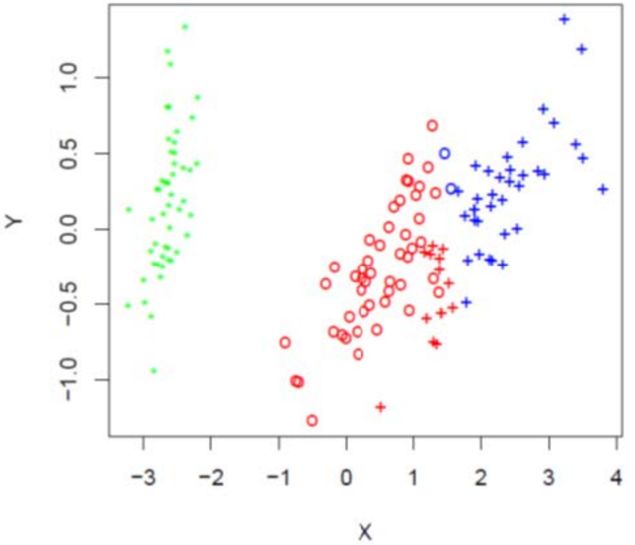
- **Visualisation and dimension reduction** (eg: PCA)

for a large number of inputs finding combinations of variables that can be plotted to display features in the data.

- **Clustering** (eg: kmeans, hierarchical clustering)

using automated techniques to find groups in the data.

# Clustering the Iris data



RN SYDNEY
VERSITY

# Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy and groups of samples that behave similarly, and features that behave similarly, and linear combinations of features with the most variation.
- Difficult to know how well you are doing.
- Different from supervised learning, but can be useful as a ***pre-processing step*** for supervised learning.

# This Unit

Supervised Learning:

- Linear models: Simple Linear Regression and Multiple Linear Regresion
- Classification: Logistic Regresion, Discrimination and kNN
- Classification and Regression Trees (Decision Trees)
- Support Vector Machines

Unsupervised Learning:

- Dimension reduction: Principal Component Analysis
- Clustering: K Means and Hierarchical

Unstructured Data:

- Text Mining (NOT COVERED)

Resampling and Error estimation

Visualisation

# Objectives

On the basis of the training data, we would like to:

- Accurately ***predict*** unseen test cases.
- ***Understand*** which inputs affect the outcome, and how.
- ***Assess*** the quality of our predictions and inferences.

# Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [**simpler methods often perform as well as fancier ones!**]
- This is an exciting research area, having important applications in science, industry and finance.
- ***Statistical learning*** is a fundamental ingredient in the training of a modern data scientist.

WESTERN SYDNEY
UNIVERSITY

# TEXTBOOK

Lecture notes are based on the textbook,

for further reference refer;

Prescribed Textbook

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R Springer.