

# **Big Data Analytics**

Session 1a **Big Data and Data Analytics** 

#### **Contact details**



Teacher: Tingting Han

- Email: tingting@dcs.bbk.ac.uk

- Room: MAL155

Office hours: By appointment

Teaching Assistant: Muawya Eldaw

#### Moodle

- Course materials
- Coursework submission dropboxes
- Announcements
- Discussions

## To pass module



- Coursework (20%)
  - 4 coursework with 5% each
  - Deadlines see next slides
  - Don't wait till the last minute
- In-class test (80%)
  - In term 3
  - 3 hours
- Pass mark: 50%

## **Tentative Schedule**

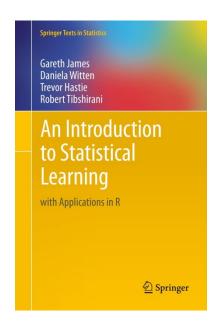


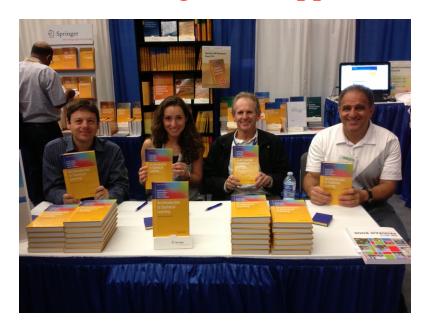
Session	Content	Coursework	
1 (29/9)	Introduction	1st CW due: 25/10 Sunday 9pm	
2 (6/10)	Basic Statistics + linear Regression	Getting to know R  Basic Statistics  Linear Regression	
3 (13/10)	Linear Regression	Linear Regression	
4 (20/10)	Logistic Regression		
5 (27/10)	Cross Validation	2 <sup>nd</sup> CW due: 15/11 Sunday 9pm	
6 (3/11)	Decision Trees	Logistic Regression Cross Validation	
7 (10/11)	Ensemble Methods	3 <sup>rd</sup> CW due: 6/12 Sunday 9pm	
8 (17/11)	SVM	Decision trees Ensemble Methods	
9 (24/11)	Clustering	4 <sup>th</sup> CW due: 3/1 Sunday 9pm	
10 (1/12)	Dimension Reduction	SVM Clustering	
11 (8/12)	Applications	Model Evaluation	

#### **Textbook**



• An Introduction to Statistical Learning: with Applications in R





- Available at <a href="http://www-bcf.usc.edu/~gareth/ISL/">http://www-bcf.usc.edu/~gareth/ISL/</a>
  - Book, datasets and R code
- A previous more advanced book:
  - The Elements of Statistical Learning (Hastie, Tibshirani and Friedman, 2<sup>nd</sup> edition)

## **Outline**



- Big Data
  - Big
  - Data
- Data Analytics

## **Outline**



- Big Data
  - Big
  - Data
- Data Analytics

## **Big Data Everywhere!**



• The world is creating ever more data, and it's a mainstream problem.

Where does the data come from?

## **Big Data Everywhere!**



- The world is creating ever more data, and it's a mainstream problem.
  - Science
    - Databases from astronomy, genomics, environmental data, transportation data, ...
  - Humanities and Social Sciences
    - Scanned books, historical documents, social interactions data, new technology like GPS, ...
  - Business & Commerce
    - Corporate sales, stock market transactions, airline traffic, amazon, ebay...
  - Entertainment
    - Internet images, Hollywood movies, MP3 files, ...
  - Medicine
    - MRI & CT scans, patient records, ...

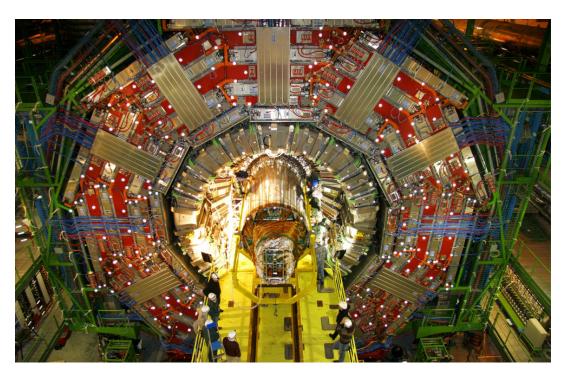


• US drone aircraft sent back 24 years worth of video footage in 2009





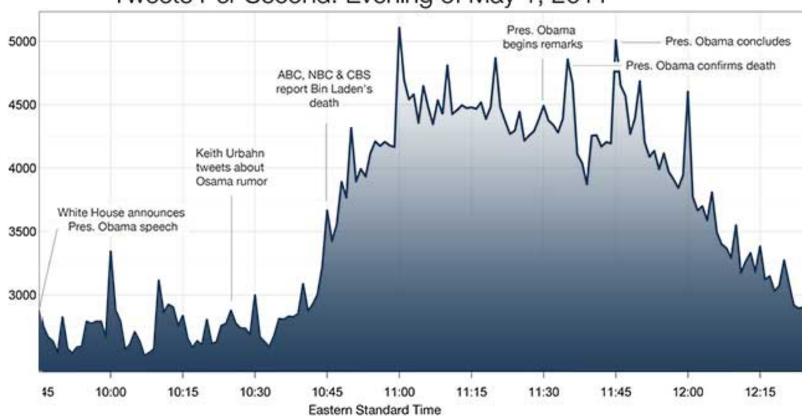
- CERN's Large Hadron Collider (LHC) generates 40 terabytes/second
  - CERN: The European Organization for Nuclear Research (Switzerland)
  - LHC: World's largest and most powerful particle accelerator
  - Terabytes (TB): all the catalogued books in America's library of Congress total 15TB





- Bin Laden's death: 5106 tweets/second
  - That's the highest rate of "sustained" tweeting ever, Twitter says.

Tweets Per Second: Evening of May 1, 2011





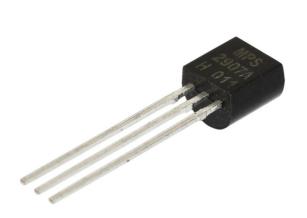
- Around 30 billion RFID tags produced per year
  - RFID: radio frequency identification

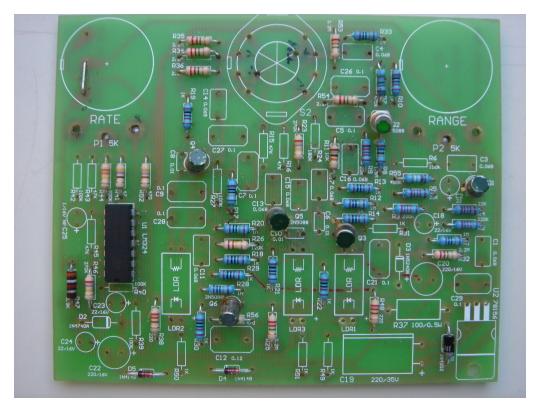






- Our world has 1 billion transistors per person
  - 7.125 billion population in 2013
  - 7.125\*10<sup>18</sup> transistors





## **A Quick Primer on Data Sizes**



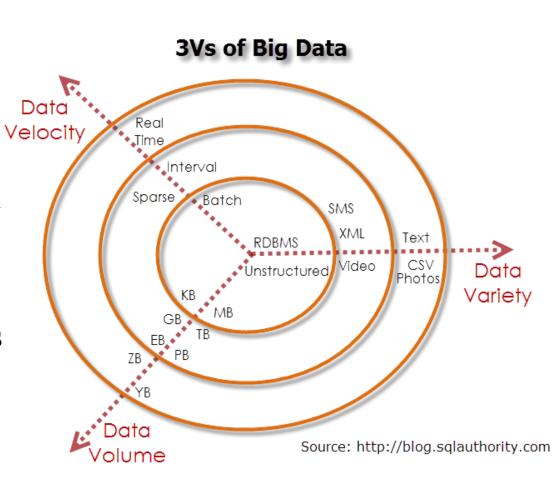
Data inflation		
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2 <sup>10</sup> , bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2 <sup>20</sup> bytes	From "large" in Greek. The complete works of Shakespeare total 5MB.  A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2 <sup>30</sup> bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2 <sup>40</sup> bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2 <sup>50</sup> bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2 <sup>60</sup> bytes	Equivalent to 10 billion copies of The Economist
Zettabyte (ZB)	1,000EB; 2 <sup>70</sup> bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2 <sup>80</sup> bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

## **Characteristics of Big Data**



- Big data spans four dimensions:
  - Volume, Velocity,Variety, and Veracity
- The first 3 Vs definition is widely used in much of the industry.
- The new V 'Veracity' is introduced by some organisations



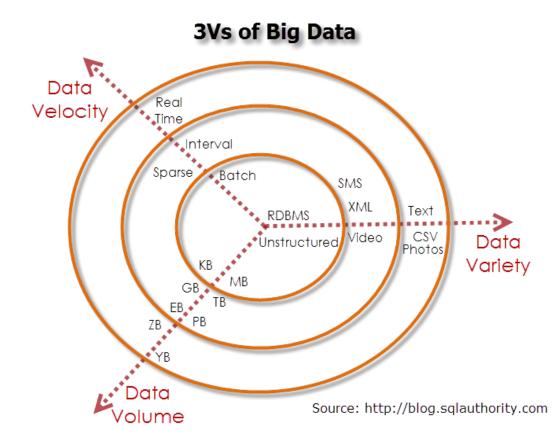
#### **Volume**



#### Volume:

#### "Data size."

- Enormous volumes of data
- The volume of data is not as much the problem as other V's.



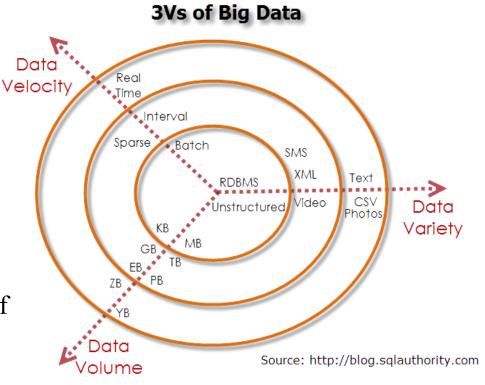
## **Velocity**



#### Velocity:

#### "Speed of change."

- The flow of data is massive and continuous.
- There are time-sensitive processes such as
  - Stock trading
  - Fraud catching
- Make valuable real-time decision if you are able to handle the velocity.
- Sampling data can help deal with velocity



## **Variety**

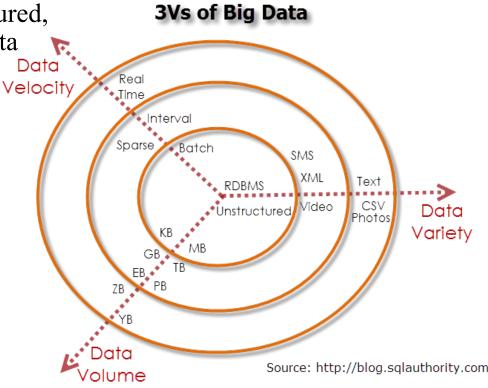


#### Variety:

#### "Different forms of data sources"

Big data is any type of data – structured,
 semi-structured and unstructured data
 such as

- Relational Data
  - Tables/Transaction/Legacy Data
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network
  - Semantic Web (RDF)
  - ...
- Streaming Data
  - You can only scan the data once



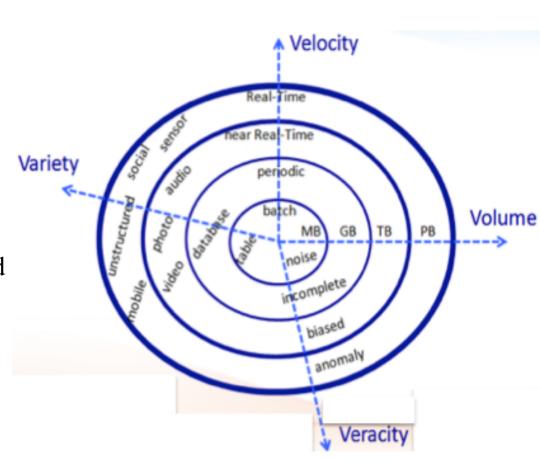
## **Veracity**



#### Veracity:

#### "Uncertainty of data"

- -Refers to the biases, noise, uncertainty, incompleteness and abnormality in data.
- —Is the data that is being stored, and mined meaningful to the problem being analysed?
- -The biggest challenge in data analysis: to keep data clean



## **Outline**



- Big Data
  - Big
  - Data

Data Analytics

#### **Data**



- Application Databases
- Wireless Sensor Data, Seismic, Astronomy Data
- Text Data
  - Webpages, Wikipedia, Emails, Enterprise Documents
- Social Media Data
  - Twitter, Blogs, Social Networks

Data is getting larger and more diverse

- Software Log Data
  - Server, API, Database Logs, Click Streams
- Images, Videos, Music
- Scientific Data, Medical, Microarray, Genome Data

## **Goal of Data Science**



Turn data into data products

#### **Data Products -- Twitter**



- Text Analysis Spam Filter/Similarity Search
- User Sentiment/Satisfaction/Feedback
- News Breakout
- Trend and Topics



200 million users as of 2011, generating over 200 million tweets and handling over 1.6 billion search queries per day

#### **Data Product -- Netflix**



- Personalised movie ratings
- Movie recommendations
- Similar movies



- Movie categories
  - e.g., 80's movie with a strong female lead, Kung Fu movies

BlockBuster is out of the business ...

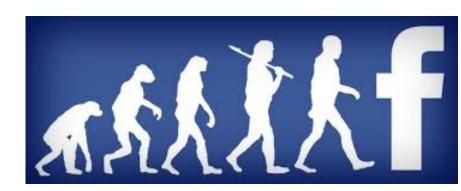
## **Data Products - LinkedIn/Facebook**



- People you may know
- Applications you may like
- Jobs/Events you might be interested
- Classifier for bad users and bad content
- With high accuracy, Facebook can guess whether you are single or married

Who does not have LinkedIn or Facebook Account?





## **Data Products -- Google**



- Web search
- News recommendation engine
- Google map
- Google Ads
- Google Analytics



Still the hottest IT company to work for now – Microsoft of the 90's, IBM of the 70's

# The Sexiest Job of the Century? Birkbeck



Data Scientist!



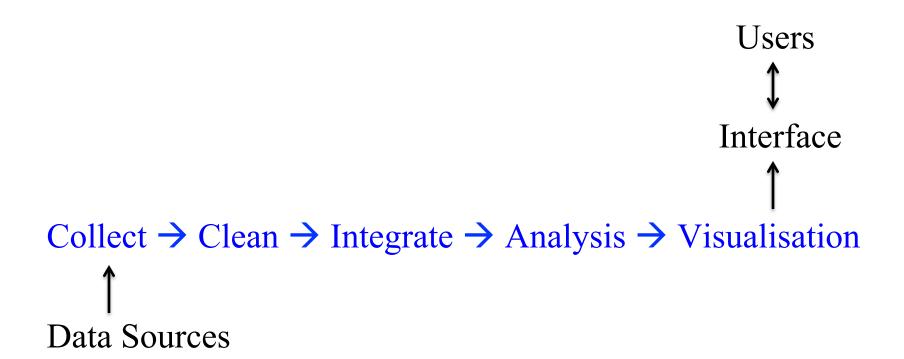
## But...



• This job title has almost as much ambiguity as the term "big data".

#### The Life of Data





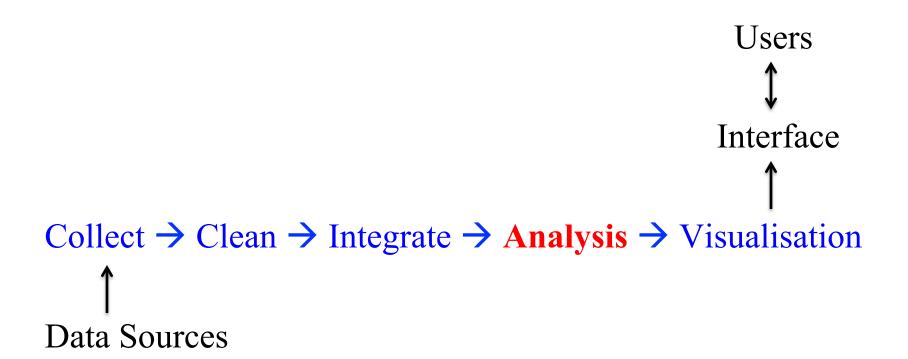
#### But...



- This job title has almost as much ambiguity as the term "big data".
- In job descriptions, a "data scientist" might do
  - Statistical analysis
  - Query and reporting
  - Database administration
  - Data warehouse management
  - Data integration
- The kind of tools a "data scientist" might need to know:
  - Hadoop, Pig, Hive, Python (typical big data manipulation tools);
  - SAS, SPSS, R (typical statistical analysis tools);
  - SQL, Business Objects, Cognos (typical query and reporting tools);
  - Excel (capable of a lot, but typically used for small-scale reporting & financial analysis);
  - Teradata, Informatica (data warehouse and loading tools).

#### The Life of Data





#### But...

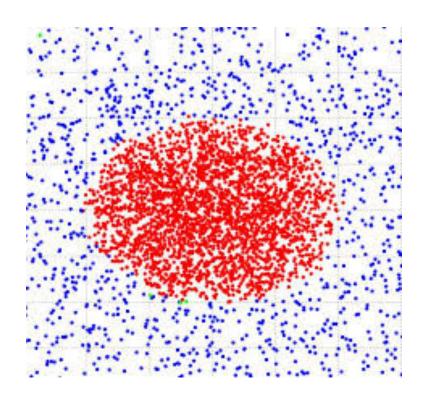


- This job title has almost as much ambiguity as the term "big data".
- In a job description, a "data scientist" might do
  - Statistical analysis
  - Query and reporting
  - Database administration
  - Data warehouse management
  - Data integration
- The kind of tools the "data scientist" might need to know:
  - Hadoop, Pig, Hive, Python (typical big data manipulation tools);
  - SAS, SPSS, R (typical statistical analysis tools);
  - SQL, Business Objects, Cognos (typical query and reporting tools)
  - Excel (capable of a lot, but typically used for small-scale reporting & financial analysis)
  - Teradata, Informatica (data warehouse and loading tools).

# **Techniques in Data Analysis**



- Statistics and Machine Learning
  - Data modeling
  - Inference
  - Prediction
  - Pattern recognition
  - **—** ...
- In this module, we will focus on statistical learning.



## **Outline**



- Big Data
  - Big
  - Data
- Data Analytics
  - Statistical learning

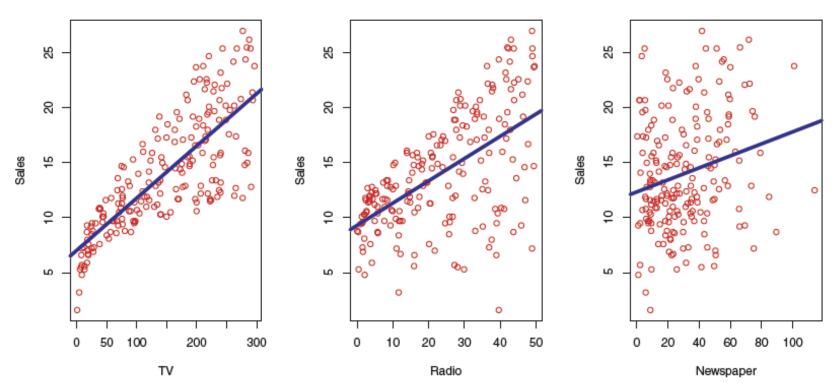
#### **Outline**



- What is Statistical Learning?
  - Why estimate *f*?
  - How do we estimate *f*?
  - The trade-off between prediction accuracy and model interpretability
  - Supervised vs. unsupervised learning
  - Regression vs. classification problems

# Advertising vs. Sales





- Is there an association between advertising and sales so that the company can adjust advertising budgets to increase sales?
  - Dataset:
    - sales in 200 markets
    - advertising budgets for 3 media in each market

# What is Statistical Learning?



- Input variables  $\mathbf{X} = (X_1, X_2, ..., X_p)$ 
  - Examples: TV budget, radio budget, newspaper budget
  - Predictors, independent variables, features, variables
- Output variable Y
  - Example: Sales
  - Response, dependent variables
- We believe there is a relationship between Y and at least one of the X's.
- We can model the relationship as

$$Y = f(\mathbf{X}) + \varepsilon$$

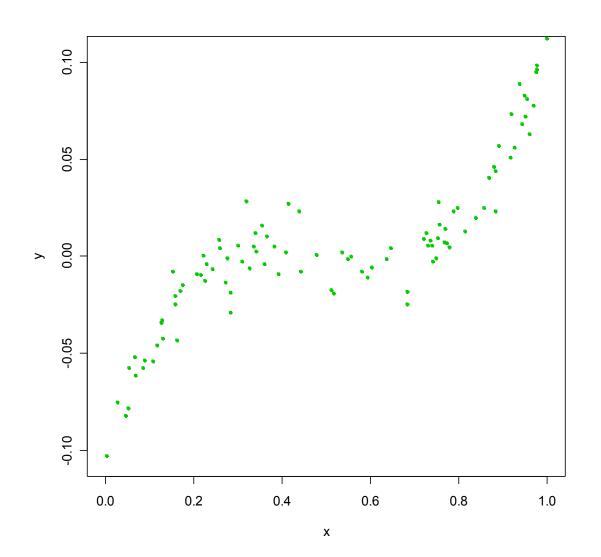
where f is an unknown function, and

 $\varepsilon$  is a random error that cannot be measured and is independent of **X**.

Irreducible error: no matter how well we estimate f, we cannot reduce the error introduced by  $\varepsilon$ .

# **A Simple Example**





#### Green dots:

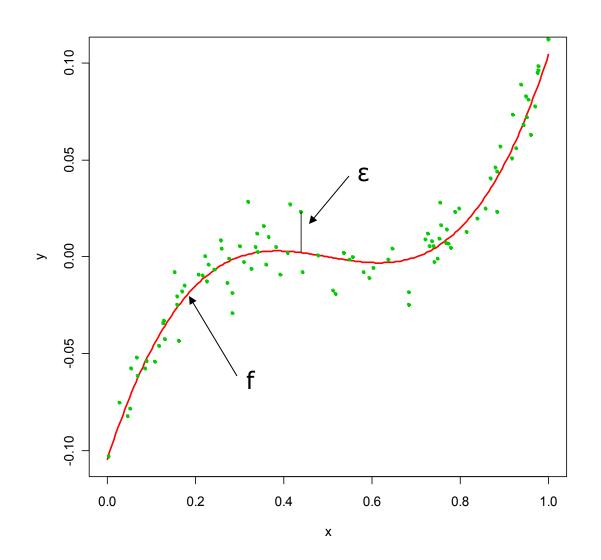
Observed values of x and y

#### Red curve:

True underlying relationship between x and y

# **A Simple Example**





#### Green dots:

Observed values of x and y

#### Red curve:

True underlying relationship between x and y

#### Black line:

The error  $\varepsilon$  associated with each observation

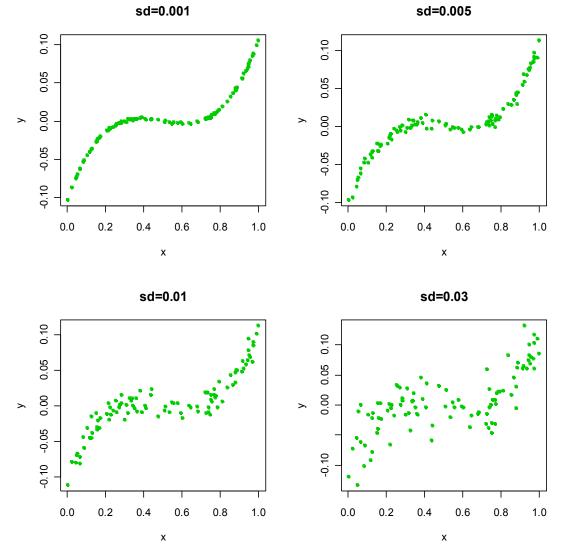
ε has a mean 0!

# **Different Standard Deviations**



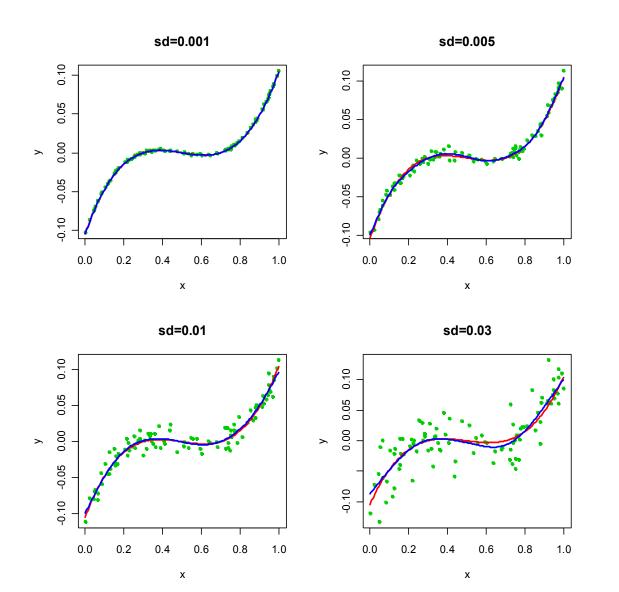
• Standard deviation measures the amount of variation from the average.

• The difficulty of estimating f will depend on the standard deviation of the  $\varepsilon$ 's.



# Different Estimates For f





#### Green dots: Observed values of x and y

#### Red curve:

True underlying relationship between x and y

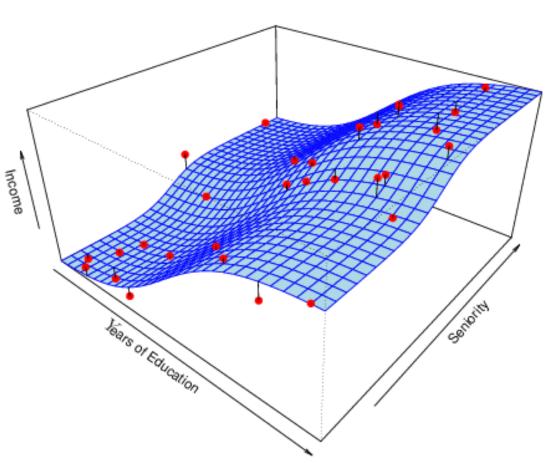
# Blue curve: Estimated relation

Estimated relationship between x and y

# Income, Education, Seniority



- Y: Income
- X: Year of Education, Seniority
- #observations: 30 (red points)
- f: the blue surface
- $\varepsilon$ : the vertical lines



# Why Do We Estimate f?



- Statistical learning is all about how to estimate *f*.
  - The term refers to using the data to "learn" f.
- Why do we care about estimating *f*?
  - There are 2 reasons for estimating *f*
  - Prediction
  - Inference

## **Prediction**



- If we can produce a good estimate for f (and the variance of  $\varepsilon$  is not too large) we can make accurate predictions for the response, Y, based on a new value of X.
- Example: Direct Mailing Prediction
  - Interested in predicting how much money an individual will donate when receiving a mailing based on observations from 90,000 people on which we have recorded over 400 different characteristics.
  - Don't care too much about each individual characteristic.
  - Just want to know: For a given individual should I send out a mailing?

### **Inference**



- Alternatively, we may also be interested in the type of relationship between *Y* and the *X*'s.
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?
- Example: Housing inference
  - Wish to predict median house price based on 14 variables.
  - Probably want to understand which factors have the biggest effect on the response and how big the effect is.
  - For example how much impact does a river view have on the house value etc.

# How Do We Estimate f?



• We will assume we have observed a set of training data

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$$

- We must then use the training data and a statistical learning method to estimate *f*.
- Statistical Learning Methods:
  - Parametric Methods
  - Non-parametric Methods

### **Parametric Methods**



- It reduces the problem of estimating f down to one of estimating a set of parameters.
- They involve a two-step model based approach
  - <u>STEP 1:</u>

Make some assumption about the functional form of f, i.e. come up with a model. The most common example is a linear model i.e.

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

where  $\beta_0, \beta_1, ..., \beta_p$  are parameters to be fitted.

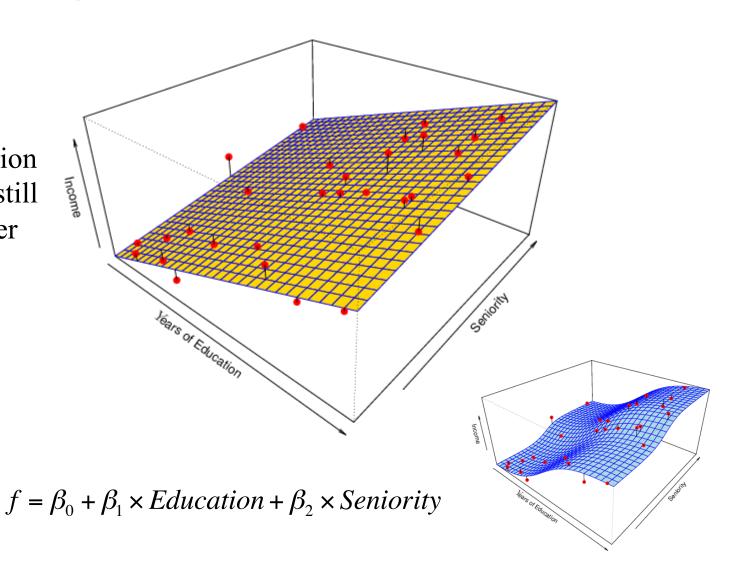
#### – <u>STEP 2:</u>

Use the training data to fit the model i.e. estimate f or equivalently the unknown parameters such as  $\beta_0, \beta_1, \beta_2, ..., \beta_p$ .

# **Example: A Linear Regression Estimate**



• Even if the standard deviation is low we will still get a bad answer if we use the wrong model.



# **Non-parametric Methods**



• They do not make explicit assumptions about the functional form of *f*.

### Advantages:

- They accurately fit a wider range of possible shapes of f.

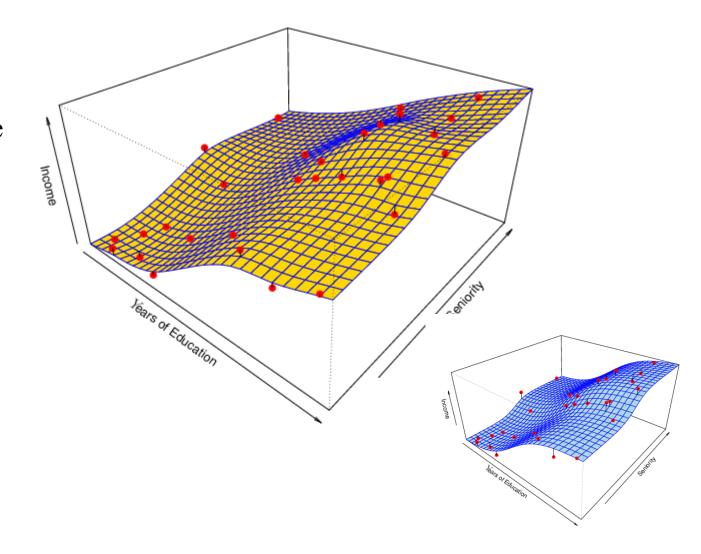
## • <u>Disadvantages:</u>

A very large number of observations is required to obtain an accurate estimate of f.

# **Example: A Thin-Plate Spline Estimate**



 Non-linear regression methods are more flexible and can potentially provide more accurate estimates.



# Prediction Accuracy vs. Model Interpretability

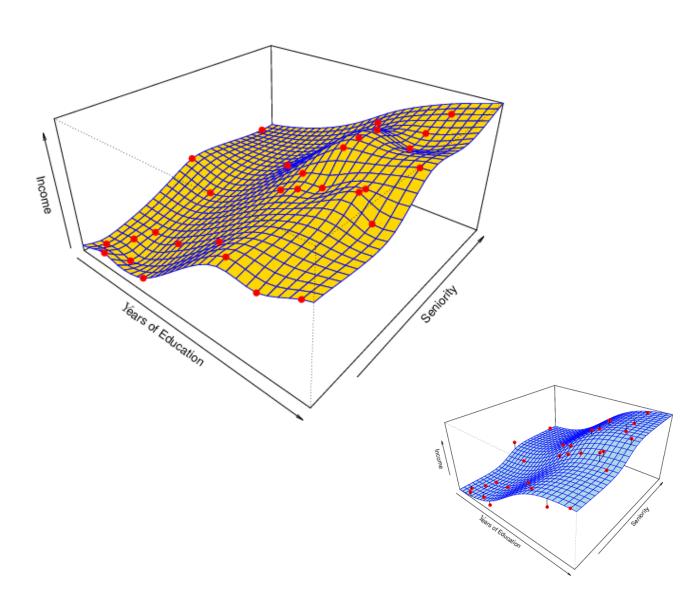


- Why not just use a more flexible method if it is more realistic?
- There are two reasons
  - Reason 1: (interpretability)
    - A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better).
  - Reason 2: (overfitting)
    - Even if you are only interested in prediction, so the first reason is not relevant, it is often possible to get more accurate predictions with a simple, instead of a complicated, model.
    - This seems counter intuitive but has to do with the potential for overfitting in highly flexible methods.

# **A Poor Estimate**



• Non-linear regression methods can also be too flexible and produce poor estimates for *f*.



# Supervised vs. Unsupervised Learning

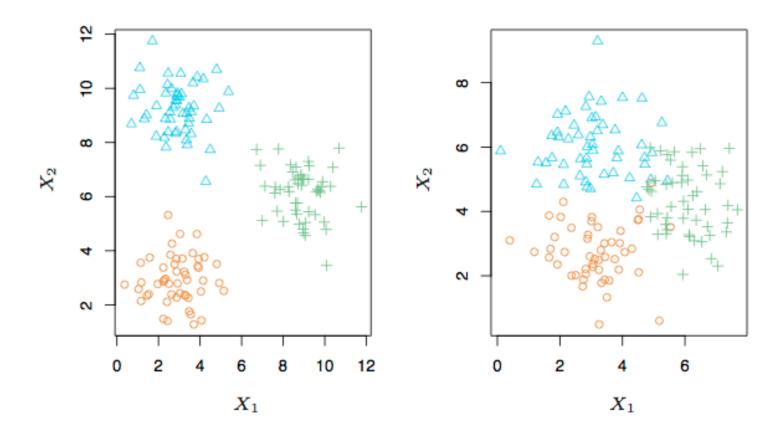


- We can divide all learning problems into Supervised and Unsupervised situations
  - Supervised Learning:
    - Supervised Learning is where both the predictors, **X**, and the response, *Y*, are observed.
    - Example: linear regression, logistic regression, boosting, SVM
  - Unsupervised Learning:
    - In this situation only the X's are observed.
    - We lack a response variable that can supervise our analysis. We need to use the **X**'s to guess what *Y* would have been and build a model from there.
    - Example: Clustering

# **A Simple Clustering Example**



• 150 observations, 2 variables:  $X_1, X_2$ , 3 distinct groups



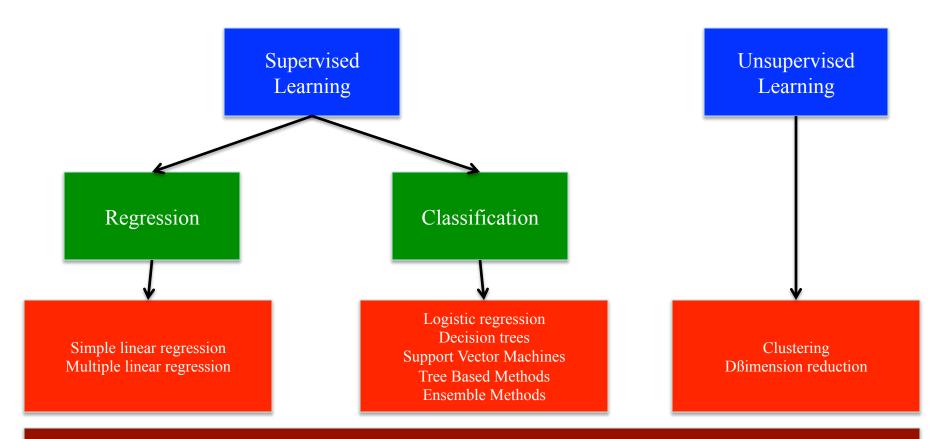
# Regression vs. Classification



- Supervised learning problems can be further divided into regression and classification problems.
- Regression covers situations where Y is continuous/numerical.
  - Examples
    - Predicting the value of the Dow in 6 months.
    - Predicting the value of a given house based on various inputs.
- Classification covers situations where Y is categorical
  - Examples
    - Will the Dow Jones index be up (U) or down (D) in 6 months?
    - Is this email a SPAM or not?

# **Course Syllabus Overview**





Choosing the best methods for a given application: Cross-validation