Society of Actuaries in Ireland

Demystifying Data Science
Part II

23rd October 2018

Disclaimer

The views expressed in these presentations are those of the presenter(s) and not necessarily of the Society of Actuaries in Ireland

Welcome

- Pedro Ecija Serrano

  Chair, Data Analytics Subcommittee

- Second of a series of three presentations

Disclaimer:

The material, content and views in the following presentation are those of the presenter(s).

Demystifying Data Science II - Agenda

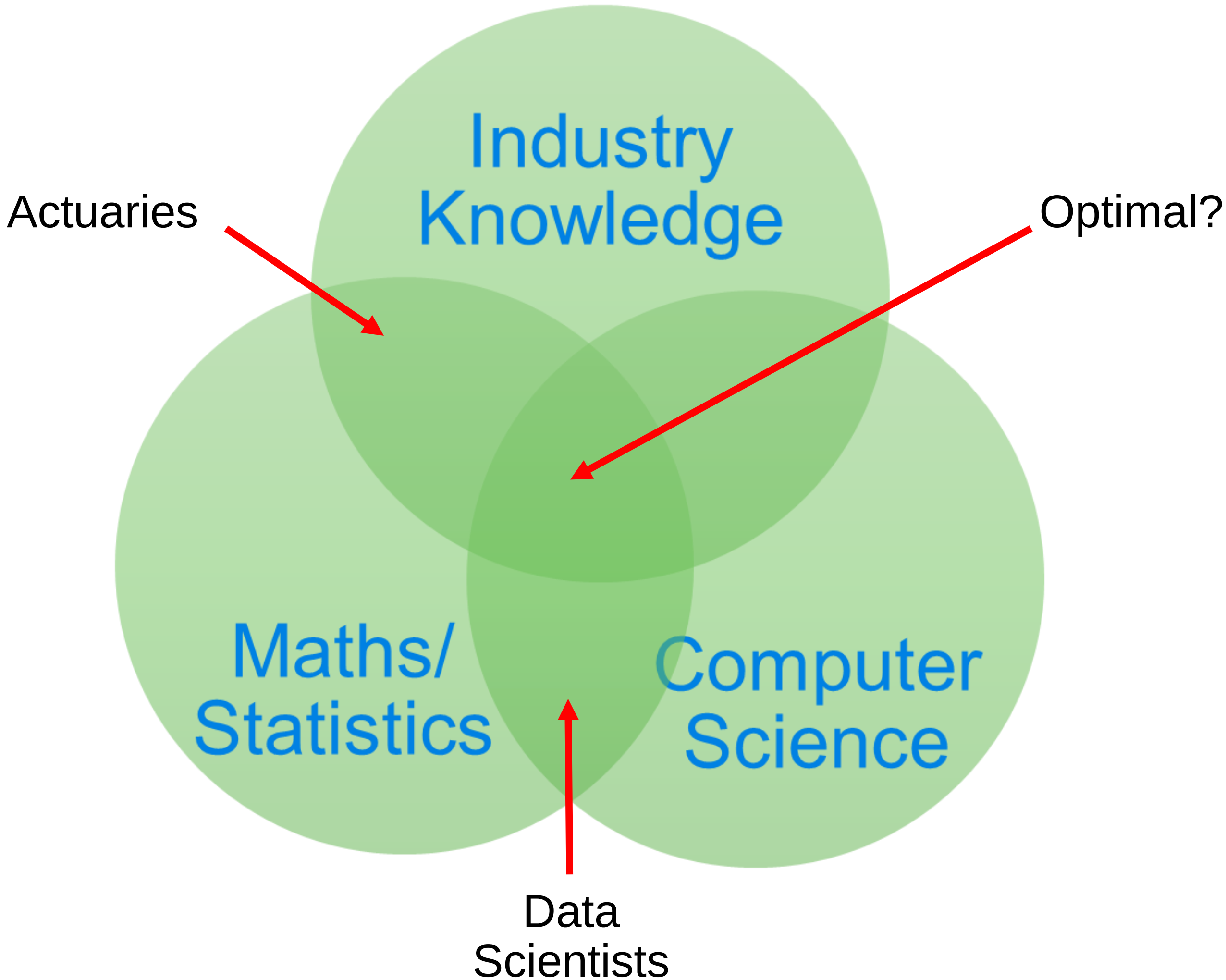
- What is Data Science?
- Why has it Grown So Quickly?
- Opportunities and Threats
- Open Source vs Closed Source
- Practical Examples – Unsupervised Learning
- Modelling Disciplines
- Practical Examples – Supervised Learning
- Honourable Mentions
- Wrap up
- Questions

What is Data Science?

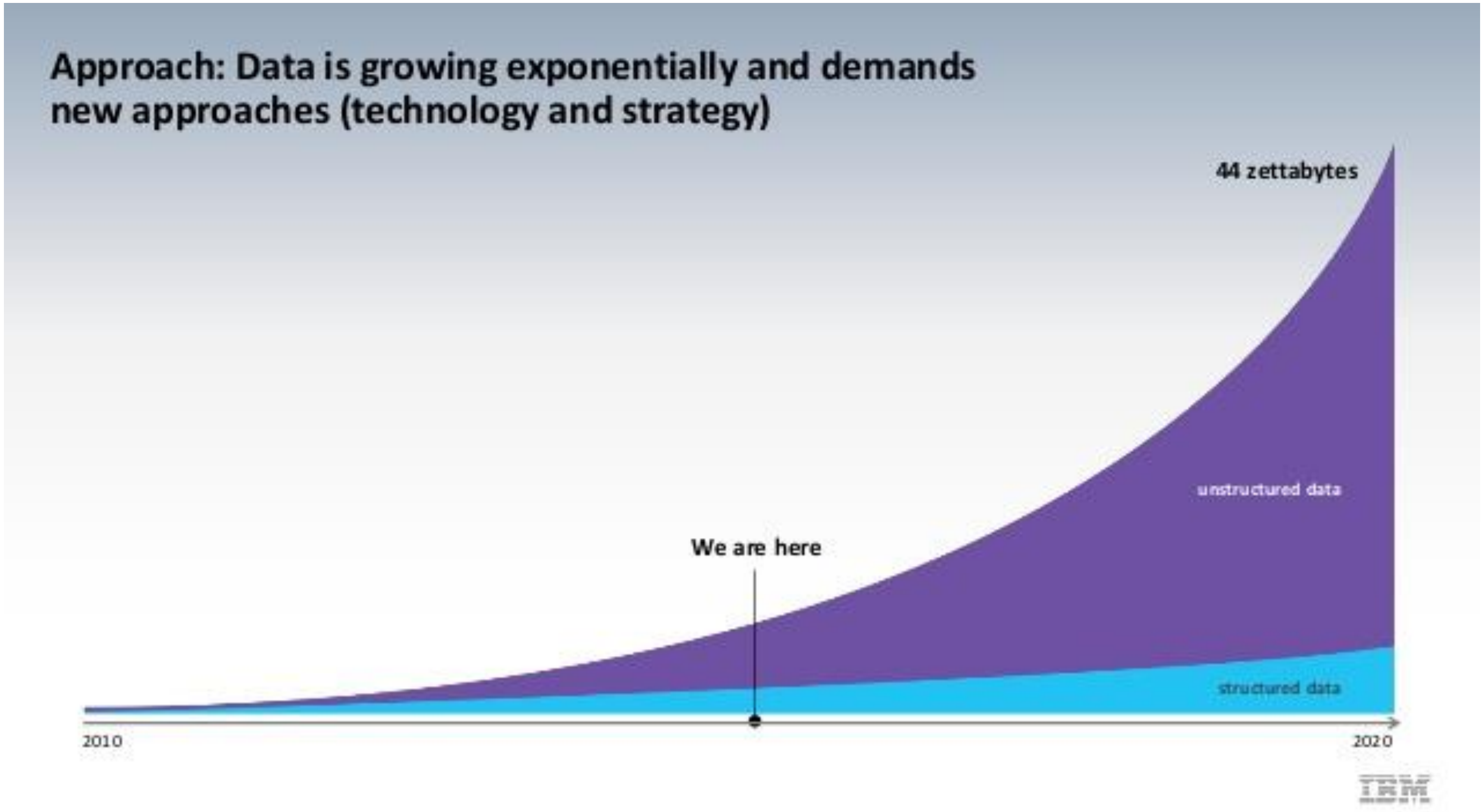
‘Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights from data** in various forms’

—Wikipedia

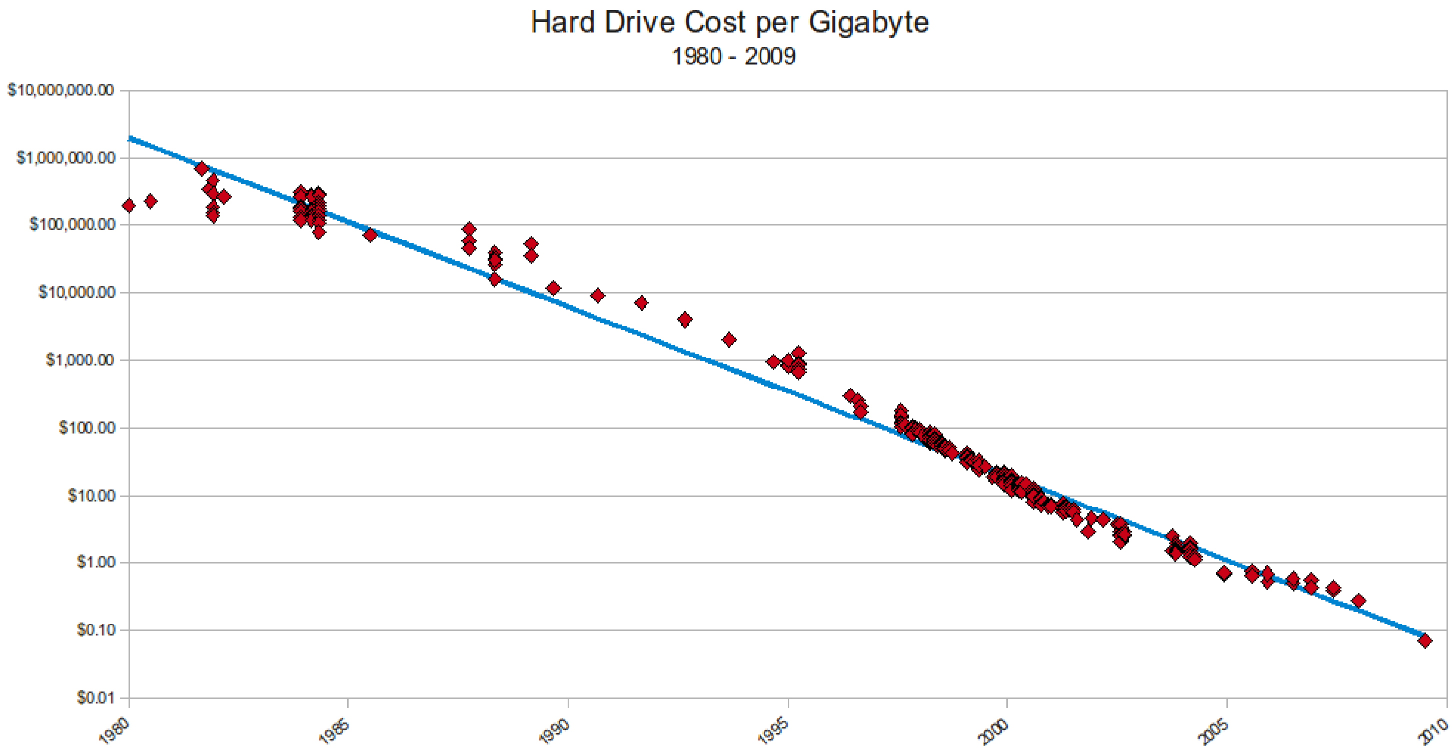
Data Science Map: Insurance Industry



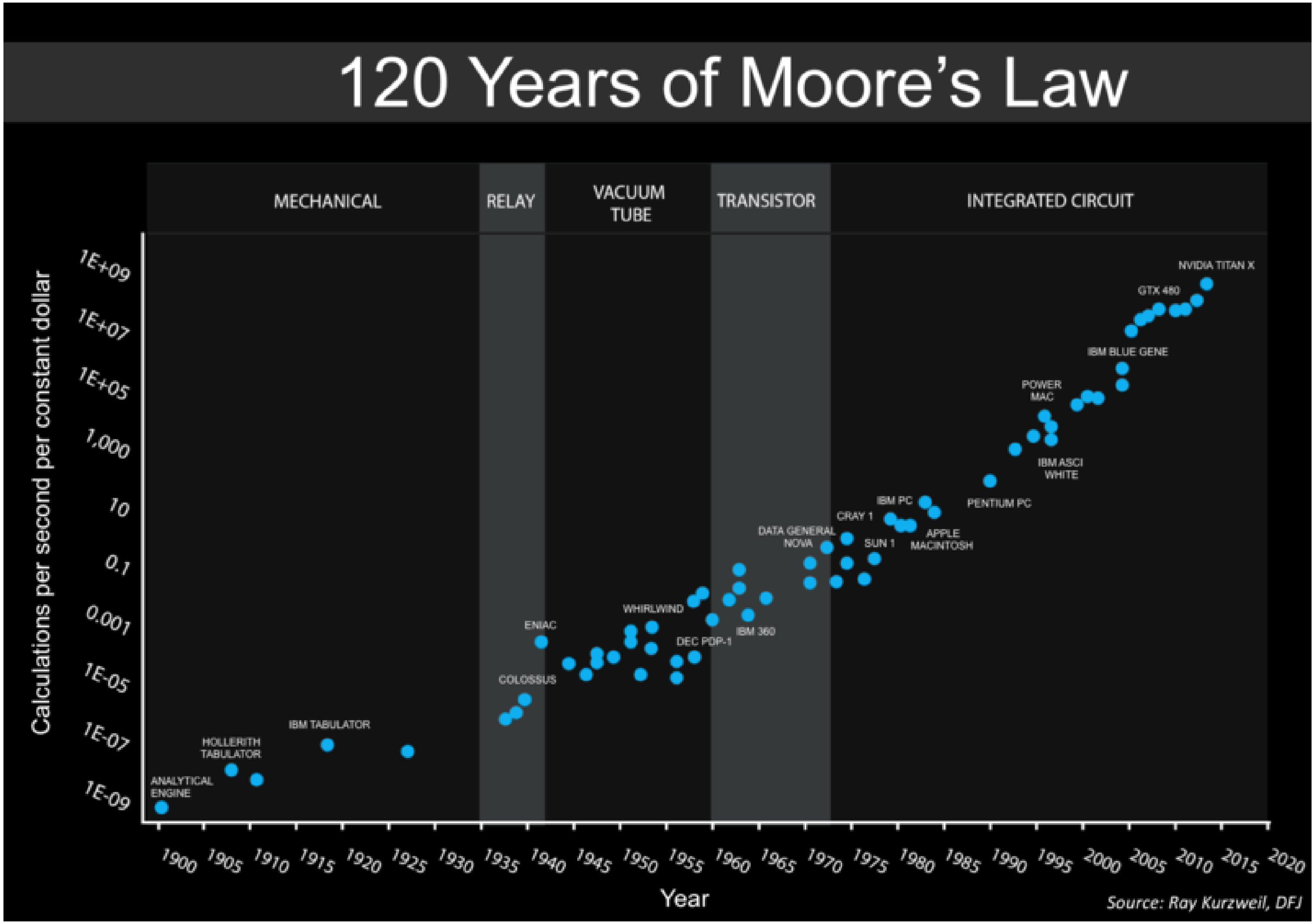
Volume of Data



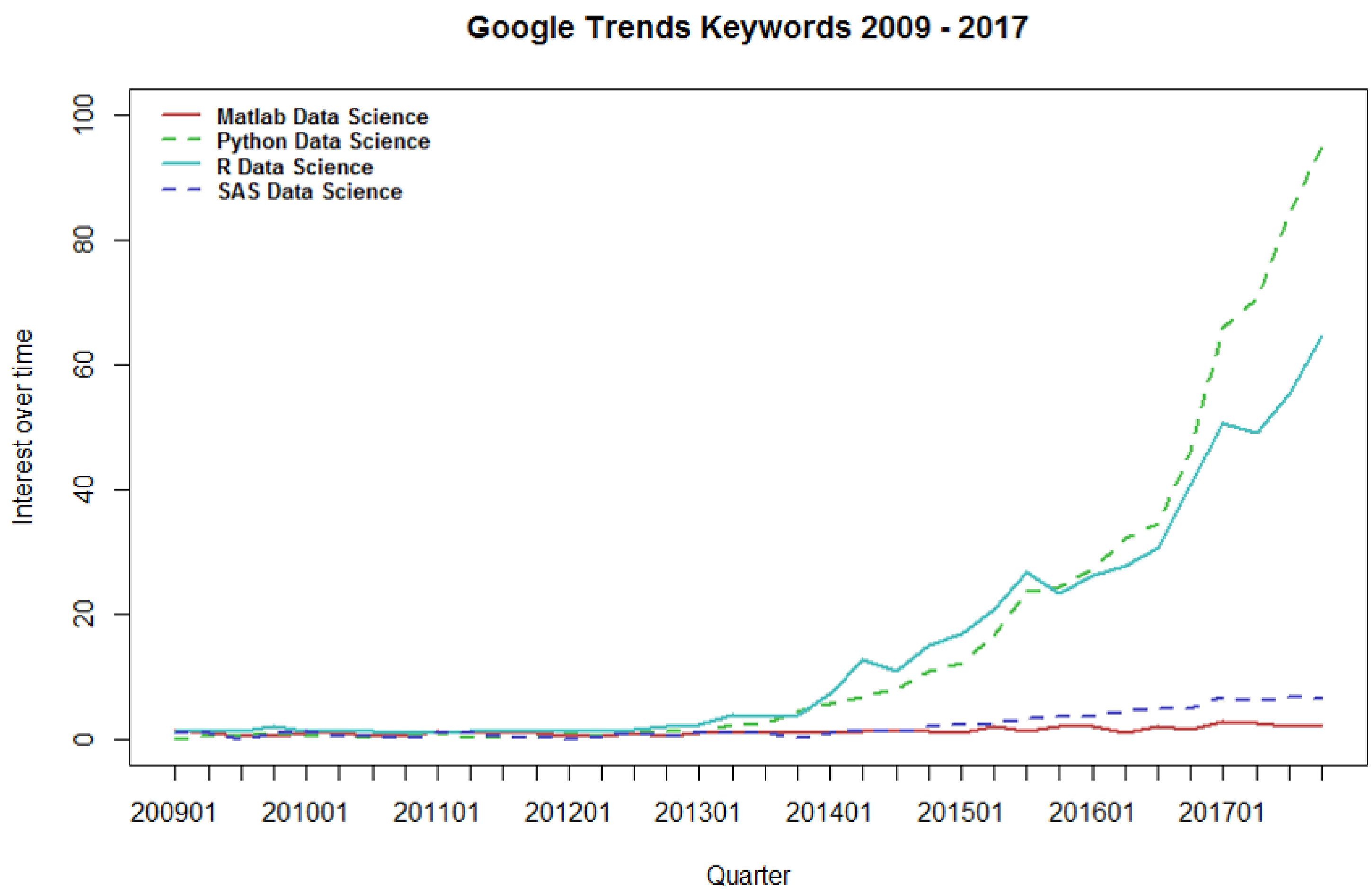
Storage Costs



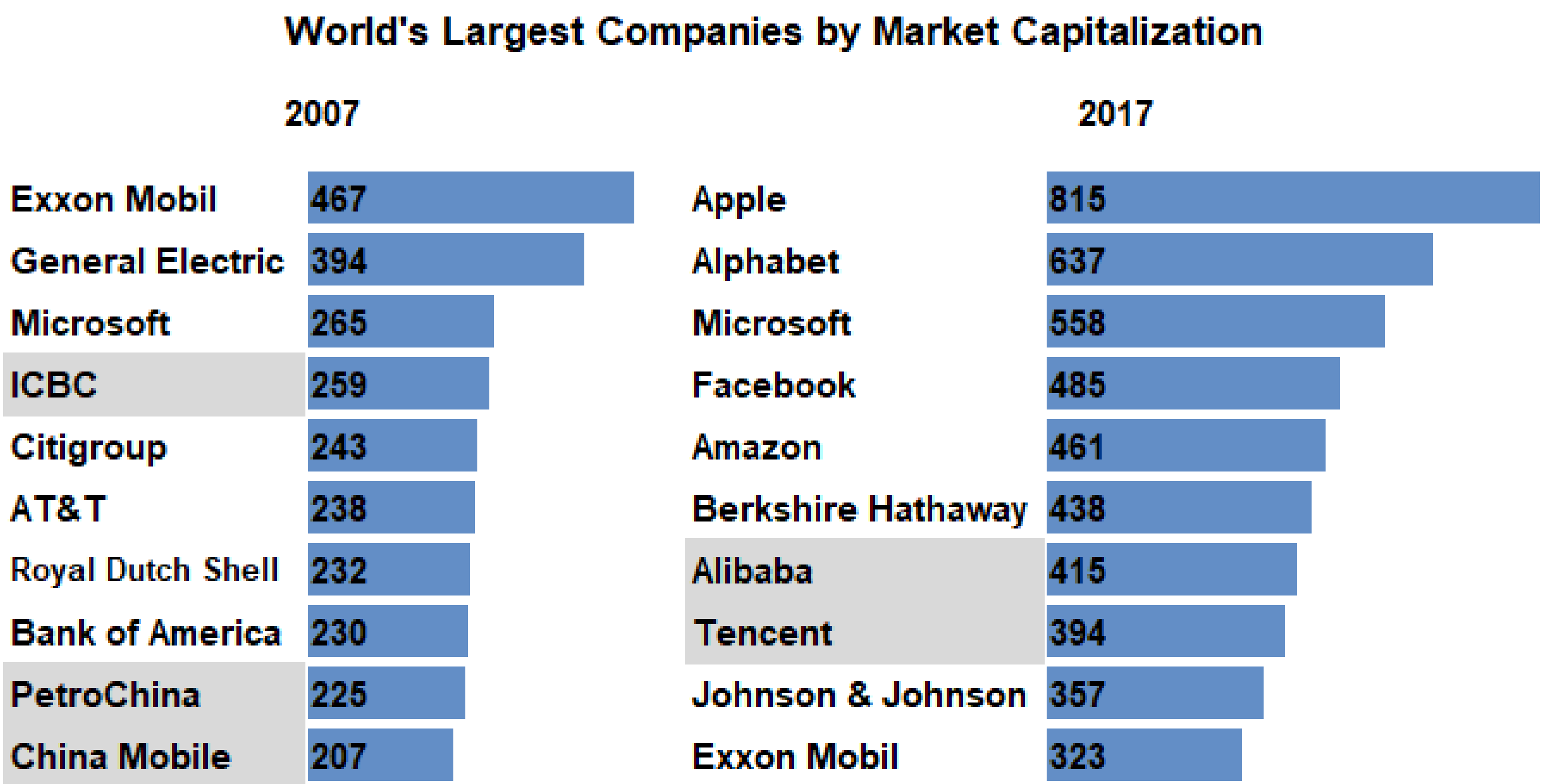
Computer Speeds



Data Science Tools



Is Data an Asset?



Is Data an Asset?

- The ultimate wider field?
- Opportunity to drive revenue growth
  - (e.g. using policyholder-level predictive modelling)
- Opportunity to work in different industries
- Powerful new tools to solve real-world problems
- Already familiar with handling data and building complex models
- CDO Roles
- Superstar salaries for top researchers

Jobs that pay over $100k

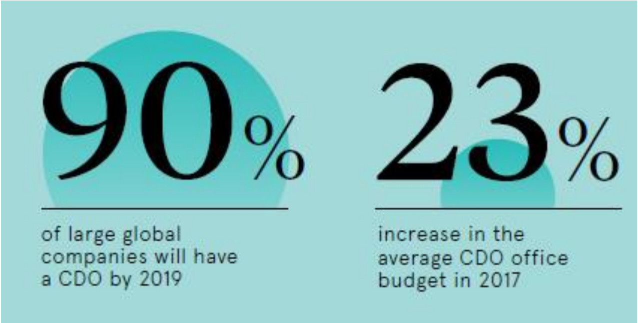
Job title	Average annual salary	Job title	Average annual salary
Neurologist	$217,837	Data scientist	$135,315
Psychiatrist	$194,563	Chief financial officer	$127,887
Anesthesiologist	$173,694	Android developer	$120,971
Radiologist	$168,706	Senior software engineer	$119,791
Physician	$165,391	Full stack developer	$111,709
Dentist	$157,250	Actuary	$111,474
Director of product management	$147,363	Tax manager	$108,515
Surgeon	$140,892	Director of business development	$107,789
Machine learning engineer	$137,332	Architect	$104,080
Vice president of sales	$136,071	Nurse practitioner	$103,233

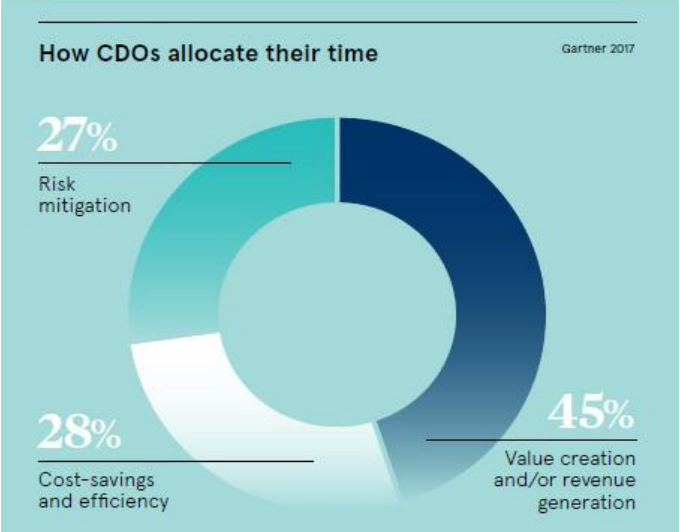
Source: Indeed

indeed

Source: Indeed.com, November 2017

Opportunities for Actuaries: Chief Data Officers





Source: VisualCapitalist.com: The Rise of the Chief Data Officer

Threats for Actuaries

- Increased competition from data scientists
  - Who have strong computer skills
  - Who have powerful predictive models
  - Strong ability to handle data and extract information from the Company's data
  - Particularly for younger actuaries

Threat Mitigation

- Improve data science skills within each actuarial team
  - Mainly by improving computer skills and learning about machine learning models

- Gain access to open-source data science tools at work
  - Overcome internal challenges to open-source software
  - e.g. the IT department might be reluctant to use new software

Opportunities for Companies

- Extract value from their data asset
- Make better data-driven decisions
- Better understanding of risks and opportunities by doing quick, novel analyses of the data
- Streamline operations

Threats for Companies

- New companies could develop massive structural advantages over incumbents?
- E.g. Amazon have massive structural advantages over traditional retailers

Open Source Languages – Python and R

- Python is a high level, general purpose programming language with readable syntax
- R is a statistical programming language designed by statisticians for statisticians
- Both are widely used for data science
- Both have similar market-leading functionality

Open Source vs Closed Source

	Open Source	Closed Source
Source Code	Open	Hidden
Redistributable?	Yes	No
Modifiable?	Yes	No
Licence and Subscription Fees?	No	Yes
Documentation, Helpdesk and Tutorials	Online (Google / Stackoverflow)	Provided by Provider (for a fee)
Responsiveness to bugs and market	Quick to respond	Depends on Provider
Version Control Systems	Available	Depends on Provider

Open-Source Advantages

- Fast
- Scalable
- Capable of full automation
- No licencing fees
- Auditability
- Flexibility
- Sustainability
- Easy to find or train developers
- Fast Learning Curve

Open-Source Misconceptions

- Not secure
- Too hard to learn
- No documentation / bad documentation
- Not as good as proprietary software

Closed Source Advantages

- It's the standard / well known
- Easier for unskilled users
- Guaranteed support (for a fee)
- Managers prefer buying Software as a Service rather than building own systems?
- Warranties and Indemnity Liability
- Unlikely to become obsolete?

Closed Source Risks

- Expensive
- Restrictive licences
- Lock-in / Capture
- Time-consuming / Hard to learn
- Management Incentives (Planned obsolescence / cash cow)
- Bankruptcy
- Unknown code quality
- Unknown level of security
- No incentive to provide good documentation

Big Data

Big Data: Datasets that are too big and complex for traditional data processing software

- Need to use new software which can distribute the storage and calculations across different machines

Data Mining

Data Mining is the process of finding patterns and relationships in large datasets

- Goal is to extract valuable understandable information from data

Predictive Analytics

Predictive Analytics is a set of statistical techniques that make predictions about future unknown events

Predictive Modelling

Predictive Models are models which make predictions about future unknown events.

- Using current and historical data
- Allowing for relationships among many factors
- Make predictions about every example in the dataset
- These predictions can be used to guide decision making

Predictive Modelling

Two main types:

- Traditional predictive models
- Machine learning models

Traditional Predictive Models

Characteristics of traditional predictive models:

- Explainable and interpretable
- Grounded in maths and statistics
- All parameters derived manually using closed form mathematical solutions or simple algorithms
- Lots of manual effort required to build high accuracy models

Machine Learning Models

Machine learning models are predictive models which have the ability to learn from data without being explicitly programmed

Learning = progressively improving performance on a specific task

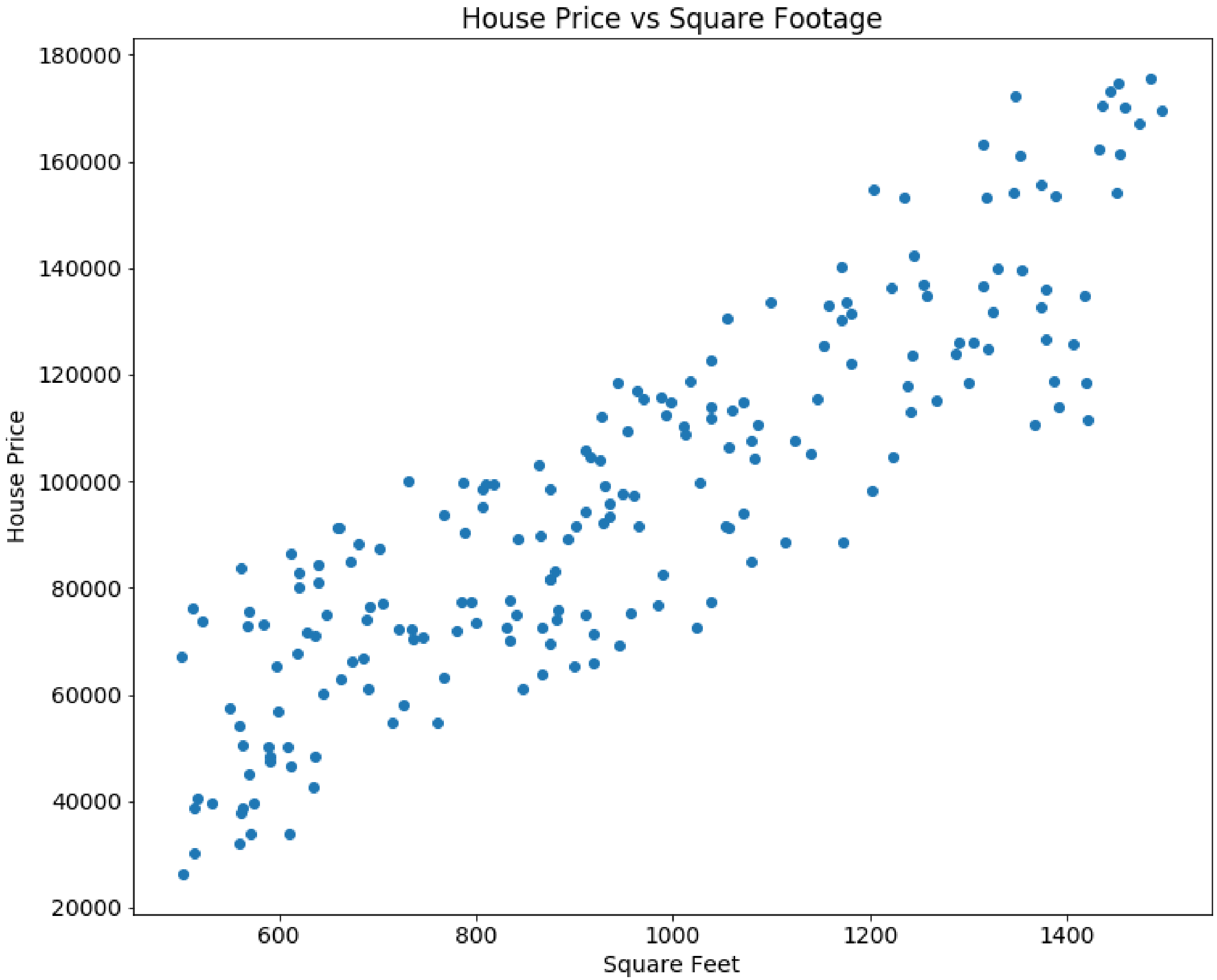
Machine Learning Models

Characteristics of machine-learning models:

- May be explainable or a black box
- Grounded in computer science
- Most parameters derived automatically using a machine learning algorithm
- Little manual effort required to build high accuracy models

Practical Example: Traditional Predictive Modelling and Machine Learning

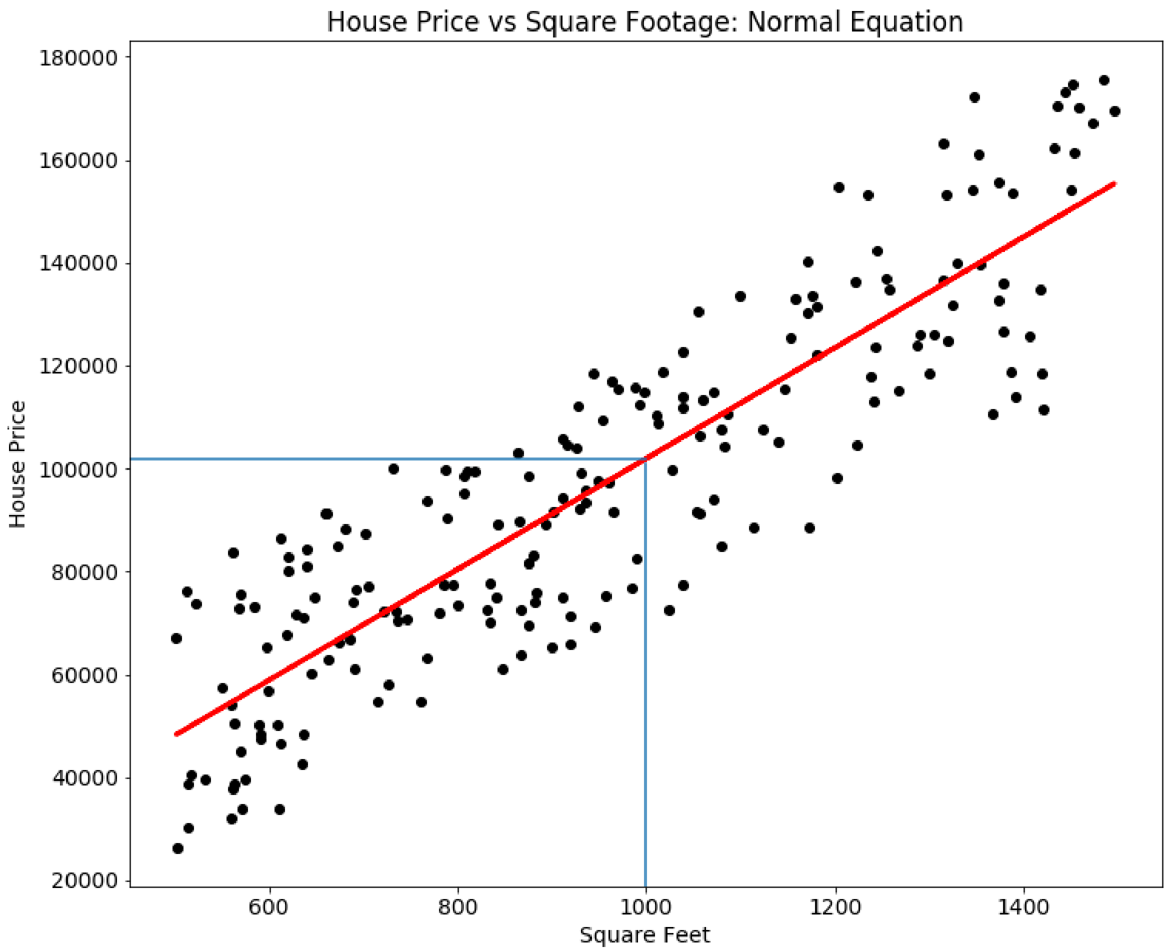
How much is a 1000 square foot house?



Eyeball approach:

Around €90k

Linear Regression Predictive Model



- Linear Regression Model:
  - Price = €101,955
  - Slope = 108
  - Intercept = -5,700
  - MSE = 258 million

- But how do you find the slope and intercept?

Approach 1: Normal Equation

Linear Regression Model:

$$\hat{y} = ax + b = \theta X$$

where:

- $\theta = [a \;\; b]$
- $X = [x \;\; 1]$

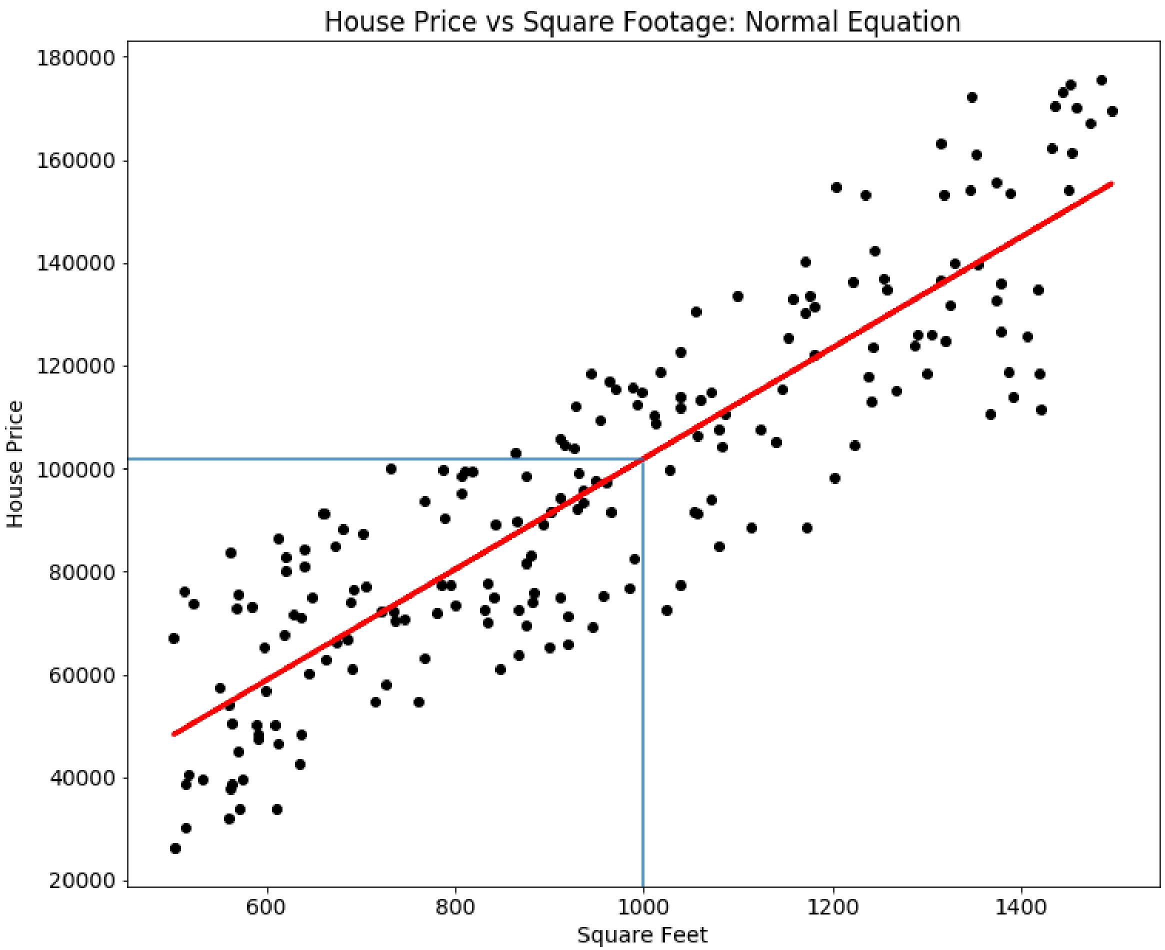
```
theta = (np.linalg.pinv(X.T * X) * X.T) * Y
y_hat = X * theta
```

Choose Loss Function, such as Mean Squared Error

Calculate parameters theta using formula:

$$\theta = (X^T X)^{-1} X^T y$$

Approach 1: Linear Regression Predictive Model



Linear Regression Model:

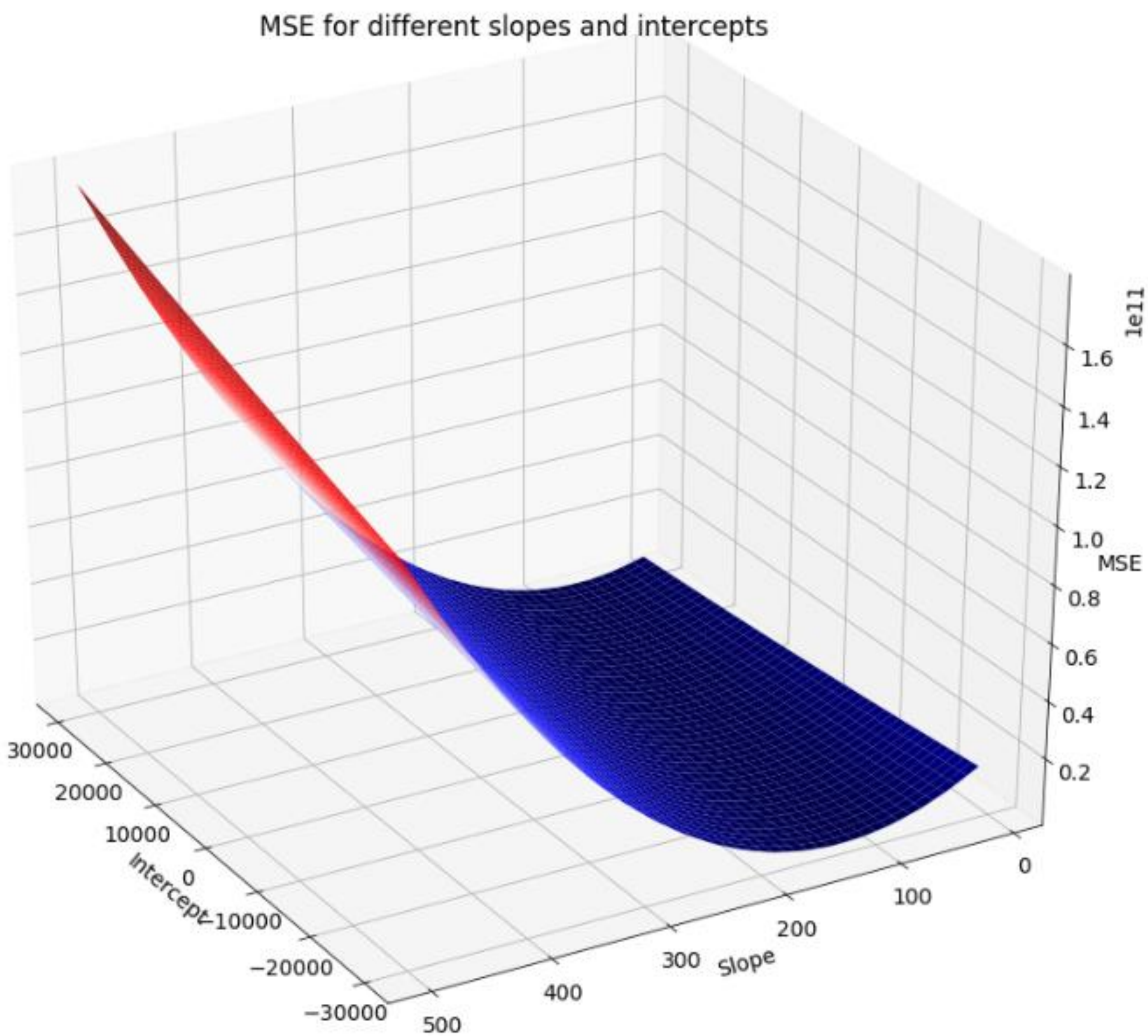
- Price = €101,955
- Slope = 108
- Intercept = -5,700
- MSE = 258 million

Approach 1: Normal Equation

Problem with normal equation:

- Only works if matrix is invertible
- Doesn’t work on other models
- Doesn’t work well on large datasets

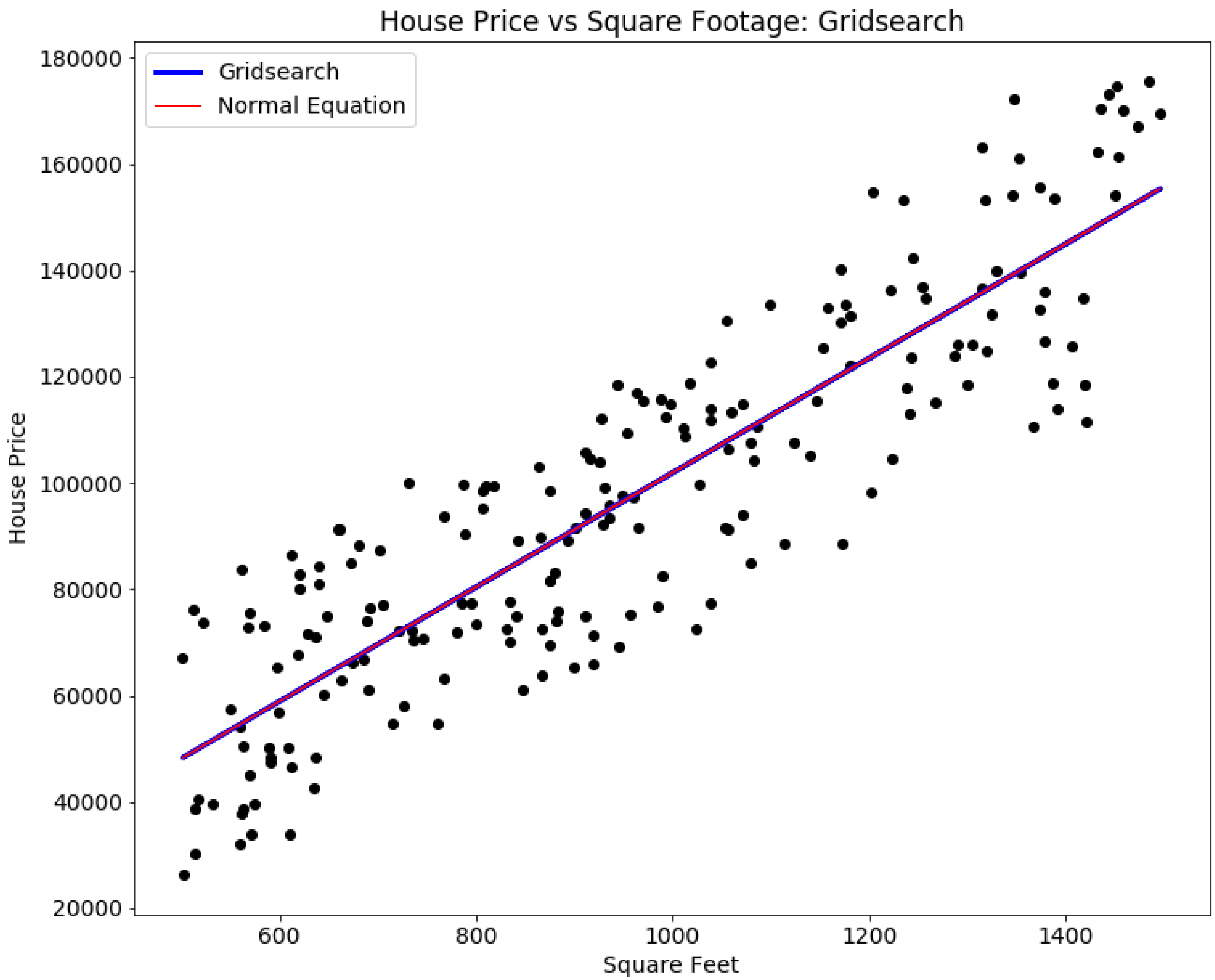
Approach 2: Gridsearch



Point with minimum MSE:

	86
Slopes	113.16
Intercepts	-11,052.63
MSE	261,059,459.22

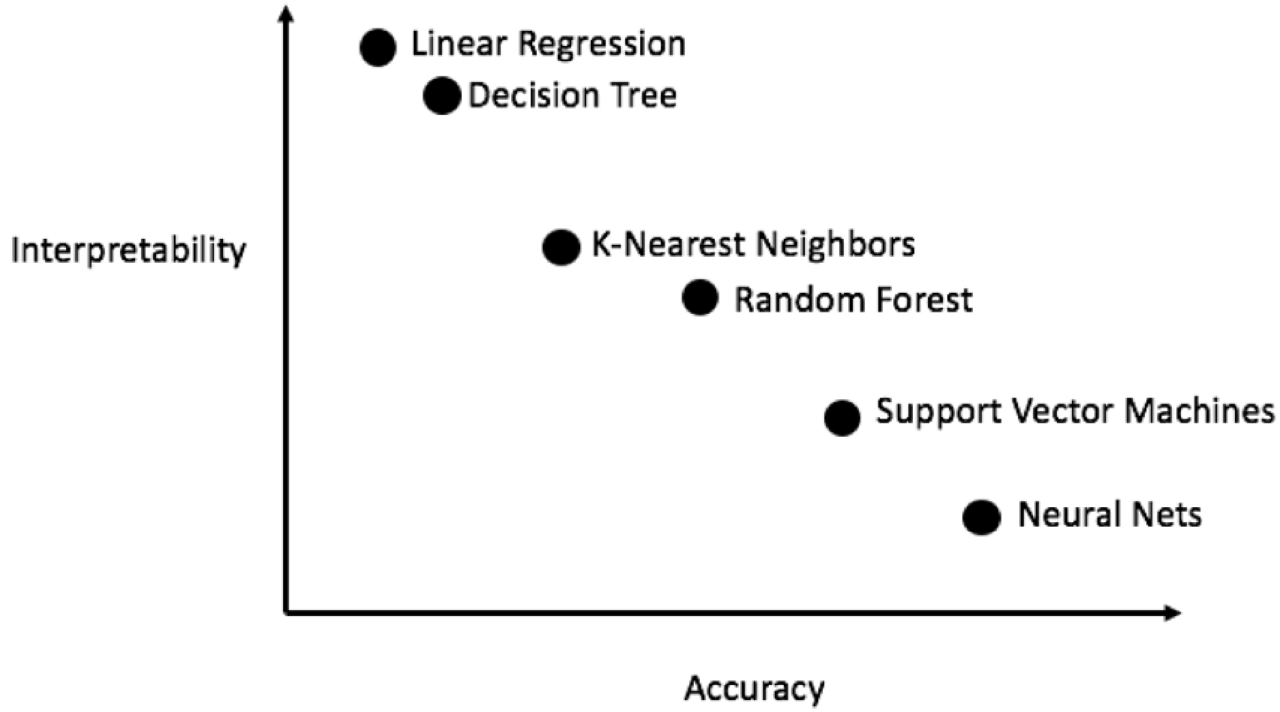
Approach 2: Gridsearch



Point with minimum MSE:

	1708
Slopes	107.68
Intercepts	-5,743.72
MSE	258,689,013.27

Machine Learning Models



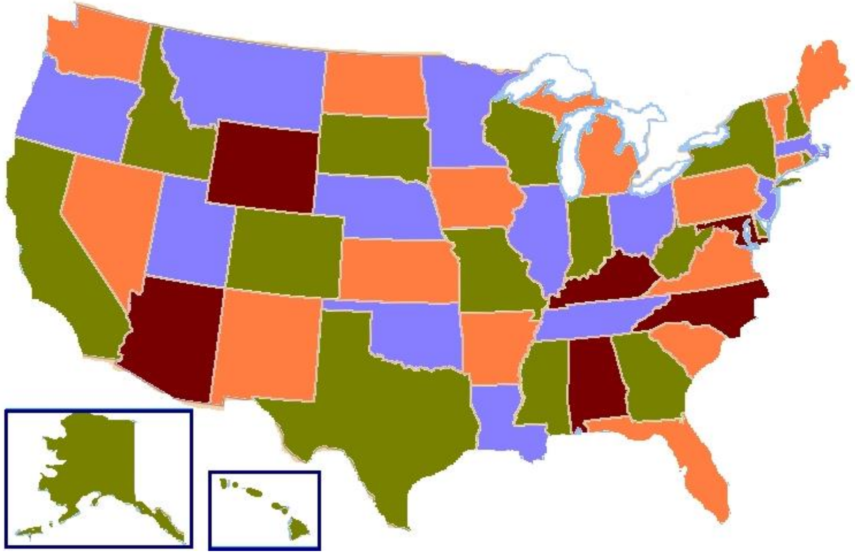
Practical Examples – Getting started

- Big Data
  - More Data
  - More Computing Power
  - More Analysis
- Computers in Actuarial Work
- A Word on Terminology
- Practical Examples

The role of Computers in Actuarial Work

- Mainframe Systems
- Valuation Software
- Spreadsheets
- A precise answer…
- ...given assumptions
- Computers may be able to ‘solve’ problems
- Or at least give valuable insights

Example 1 - Four Colour problem solved



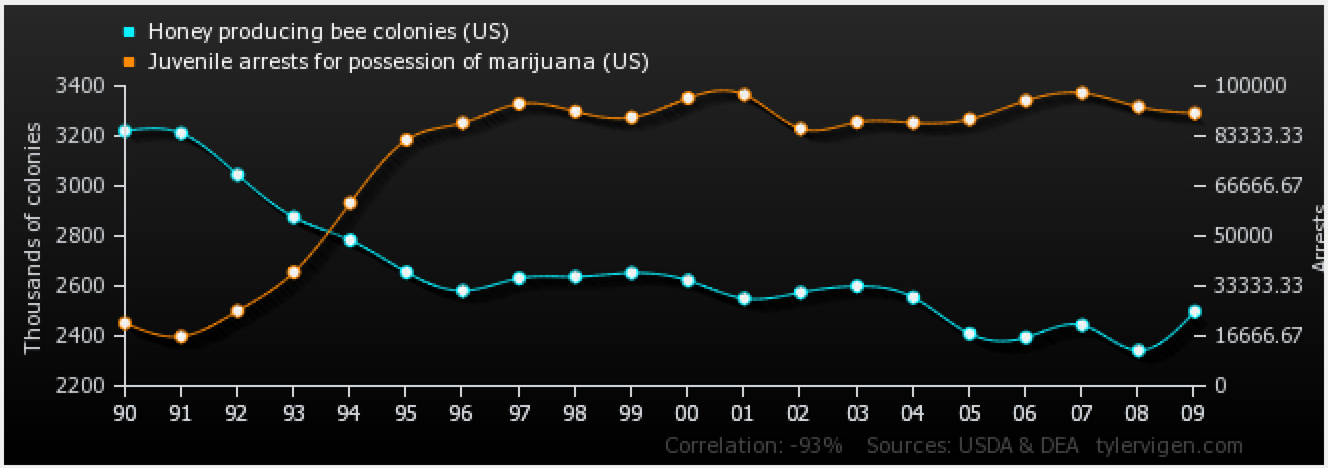
- Proved in 1976
- First major theorem proved by computer

Example 2 - Fermat's Last Theorem solved (almost)

- $x^n+y^n = z^n$
- Solved by computer for all primes up to 4,000,000

Correlation and Causation!

Honey producing bee colonies (US)
inversely correlates with
Juvenile arrests for possession of marijuana (US)



Honey producing bee colonies (US) Thousands of colonies (USDA)	'90: 3,220; '91: 3,211; '92: 3,045; '93: 2,875; '94: 2,783; '95: 2,655; '96: 2,581; '97: 2,631; '98: 2,637; '99: 2,652; '00: 2,622; '01: 2,550; '02: 2,574; '03: 2,599; '04: 2,554; '05: 2,409; '06: 2,394; '07: 2,443; '08: 2,342; '09: 2,498
Juvenile arrests for possession of marijuana (US) Arrests (DEA)	'90: 20,940; '91: 16,490; '92: 25,004; '93: 37,915; '94: 61,003; '95: 82,015; '96: 87,712; '97: 94,046; '98: 91,467; '99: 89,523; '00: 95,962; '01: 97,088; '02: 85,769; '03: 87,909; '04: 87,717; '05: 88,909; '06: 95,120; '07: 97,671; '08: 93,042; '09: 90,927
Correlation: -0.933389	

- Results always need to be interpreted!

http://tylervigen.com/spurious-correlations

A word on Terminology

- Actuaries didn't get here first!
- P = A / ä

Periodic Policy Amount =

Bounded Risk Benefit /

Contribution Vector

- Terminology not intuitive...
- ...concepts are

Association Rule Mining 1

- Purchasing datasets

	Bread	Milk	Eggs	…	Yoghurt	Tuna	Fruit
Customer 1	x						
Customer 2	x	x					x
Customer 3			x			x	
: :		x					
Customer n					x		

- Very very sparse
- Think of Amazon

Association Rule Mining 2

- Of interest, what items occur together?
- As a purchasing dataset will have very sparse data, ideas will be illustrated by a medical dataset
- 240 Patients
- 6 Symptoms

Association Rule Mining Dataset

- Illustrative dataset

	Symptoms					
	1	2	3	4	5	6
Patient 1					x	
Patient 2		x		x		
Patient 3			x	x		x
: :	: :	: :	: :	: :	: :	: :
Patient 240				x	x	
Total	**19**	**157**	**55**	**85**	**58**	**181**

- Less sparse

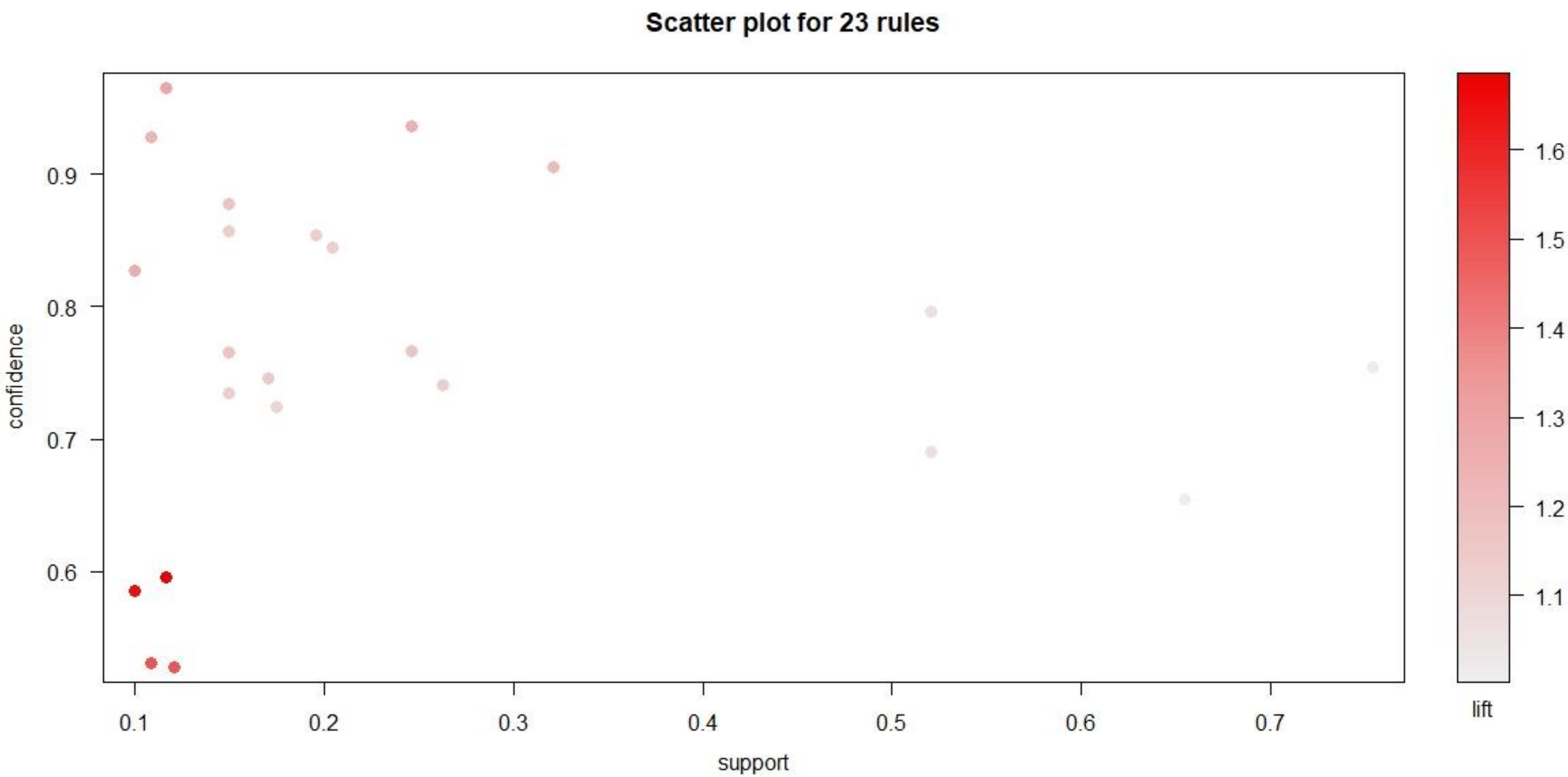
Association Rule Mining Investigation

- Which symptoms occur together?
- Three key concepts...

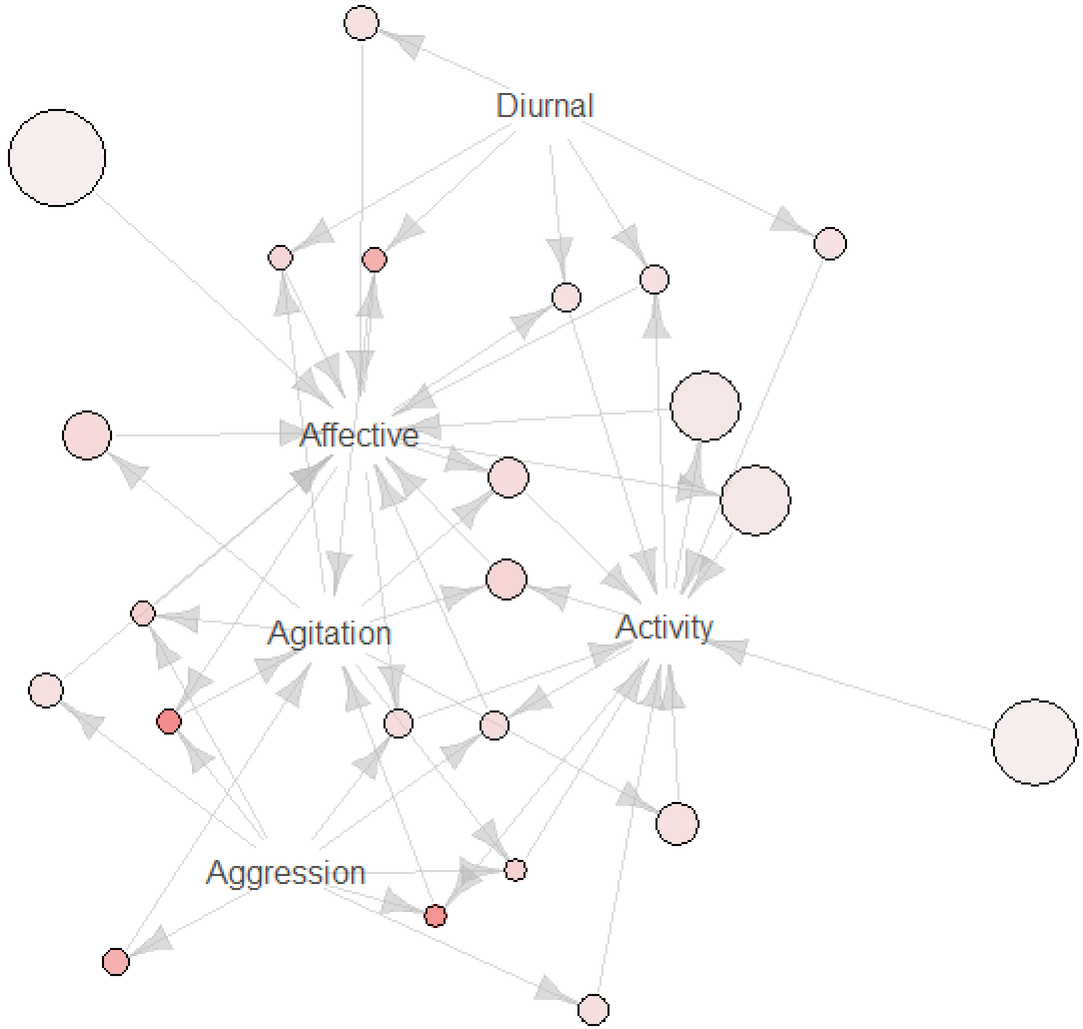
For symptoms A & B

1) **Support** = $P(A \cap B) = P(A,B)$
2) **Confidence** = $P(B|A) = P(A,B) / P(A)$
3) **Lift** = $P(A,B) / [P(A).P(B)]$

Association Rule Mining Result 1



Association Rule Mining Result 2



Association Rule Summary

- Concepts are not difficult
- Terminology and visualisation can be confusing at first
- Basic analysis can be enhanced by adding bounds and standardising results
- Very sophisticated algorithms can be developed but speed is an issue

What we’re looking to cover, a reminder

Unsupervised Learning

- No y value, Multiple x values

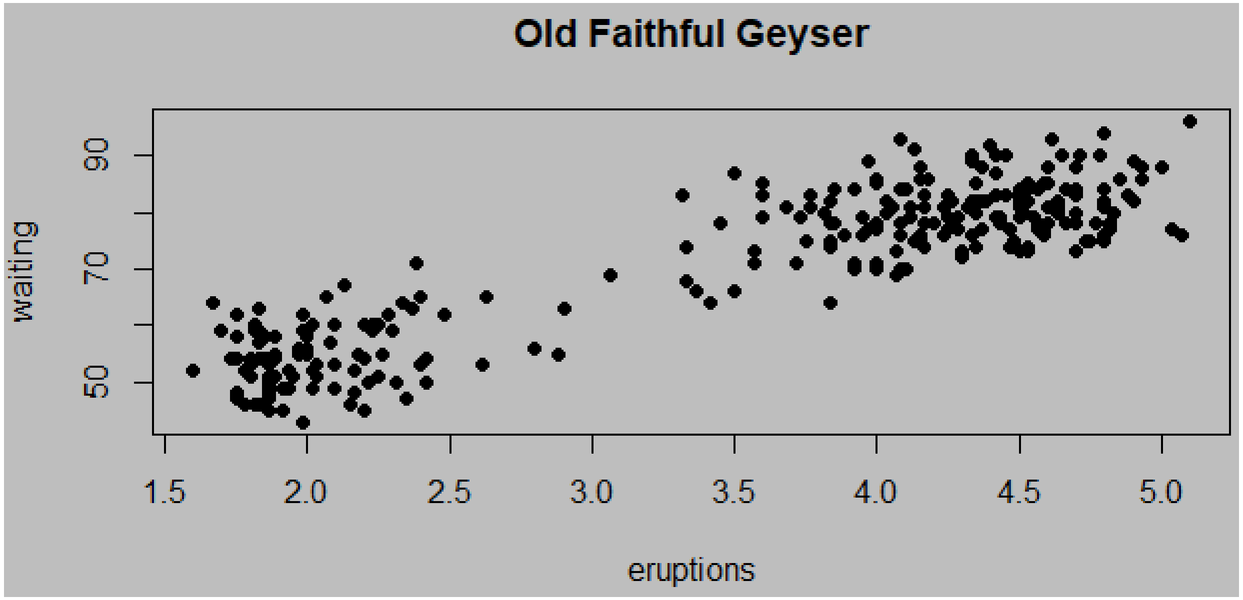
Supervised Learning

- We do have a y value & multiple x values

Unsupervised Learning 1

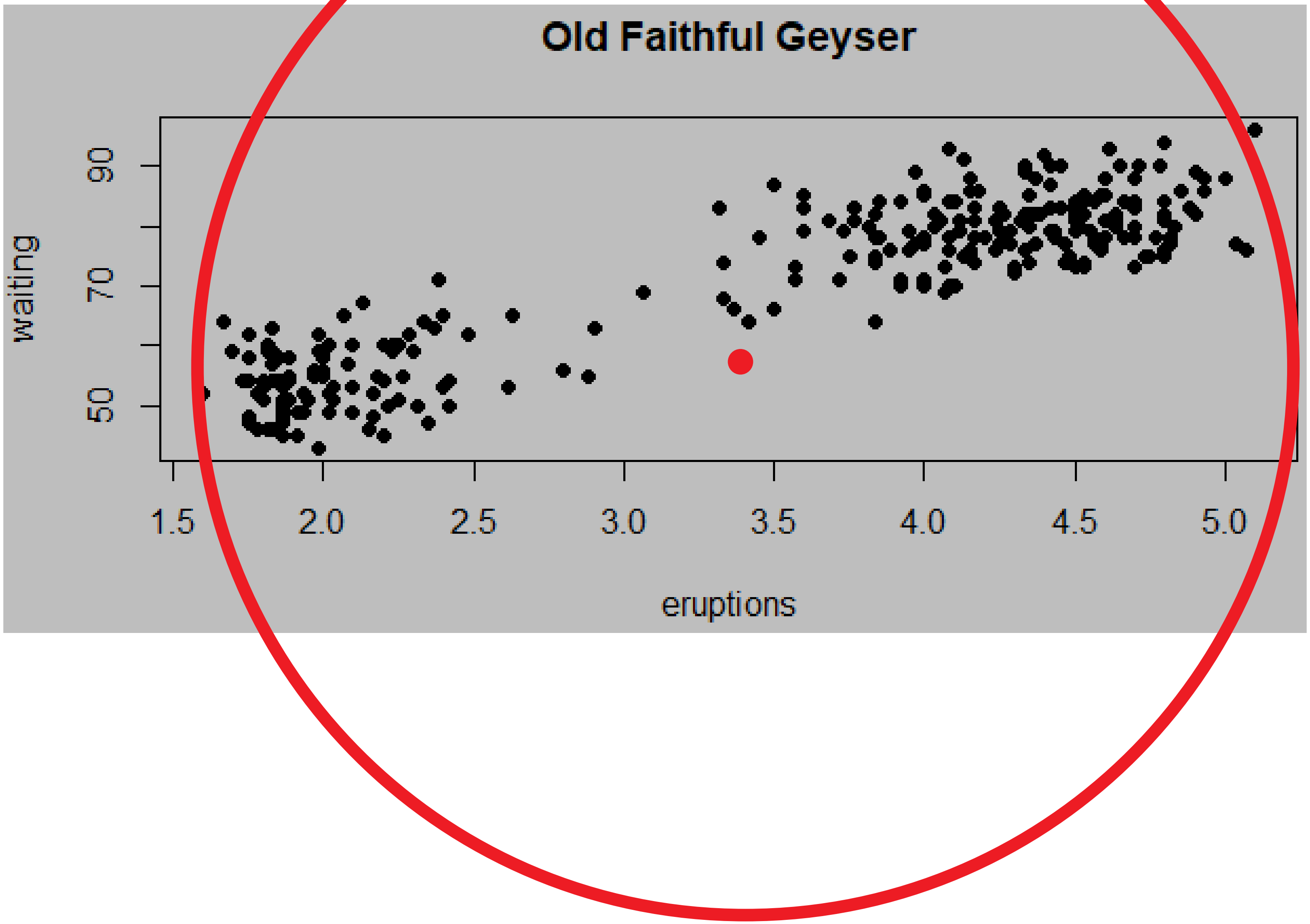
- Old Faithful Geyser
- 272 data points on Waiting & Eruption Times

Unsupervised Learning 2

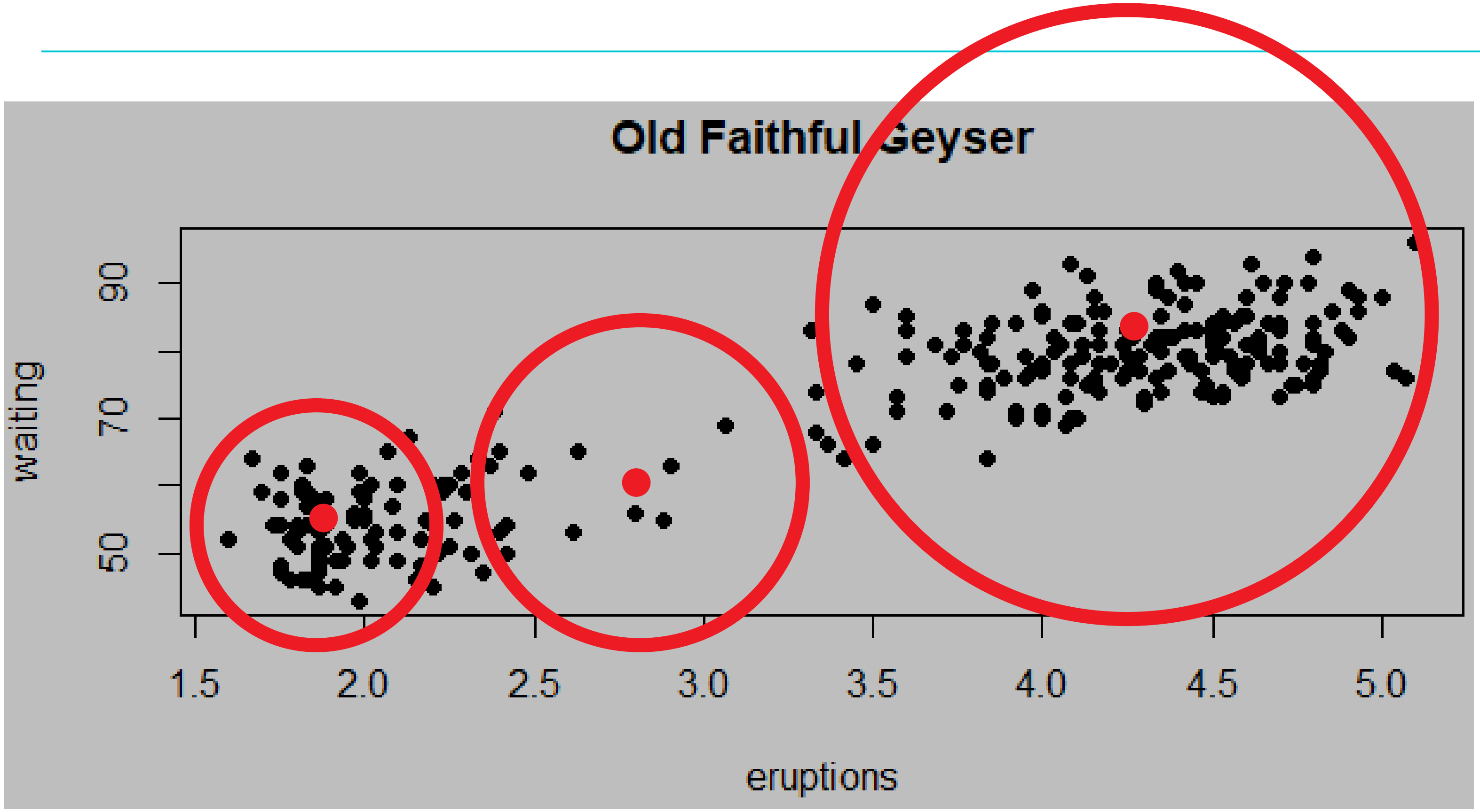


- Old Faithful Geyser
- 272 data points on Waiting & Eruption Times

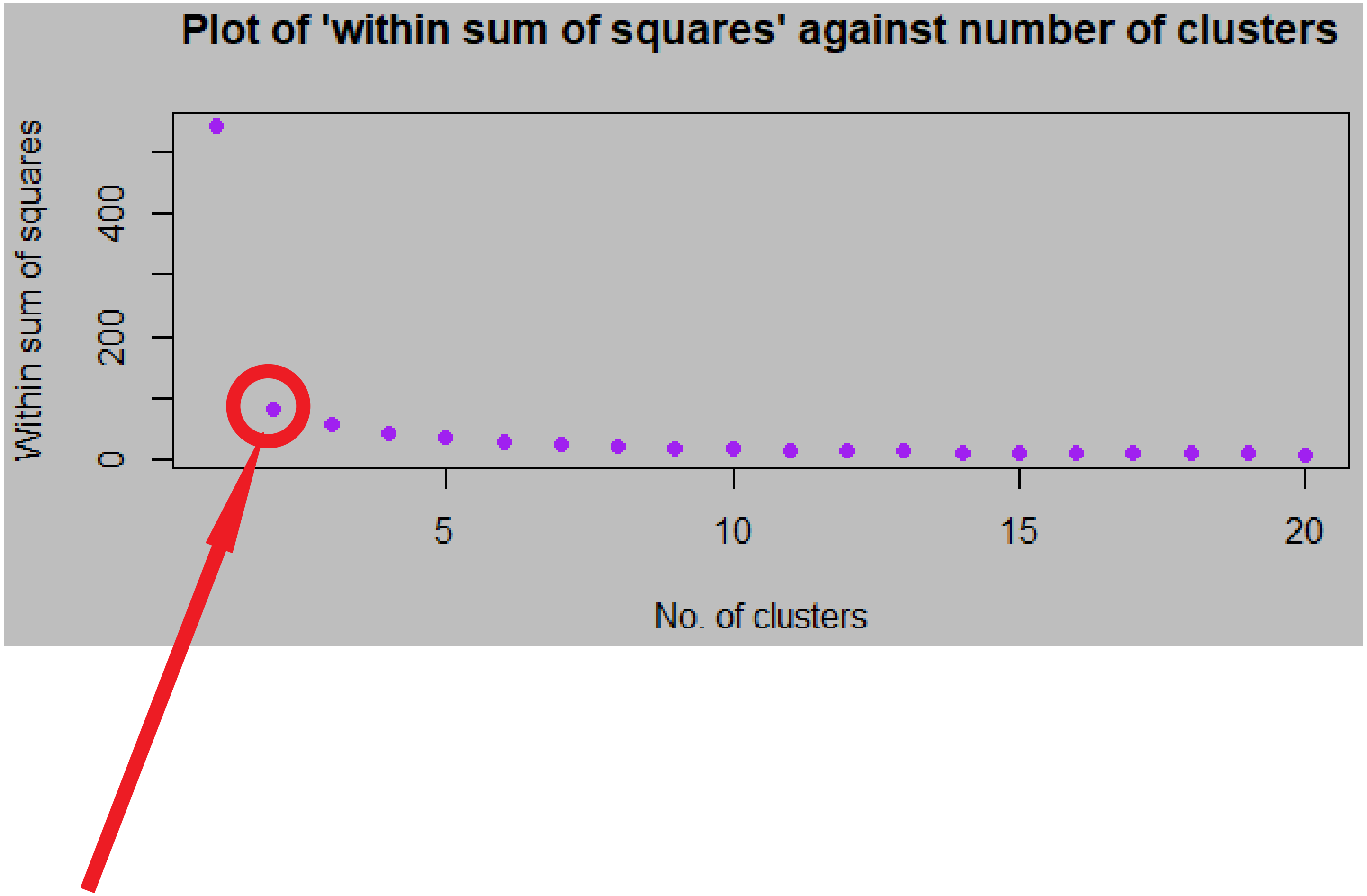
Unsupervised Learning 3



Unsupervised Learning 4



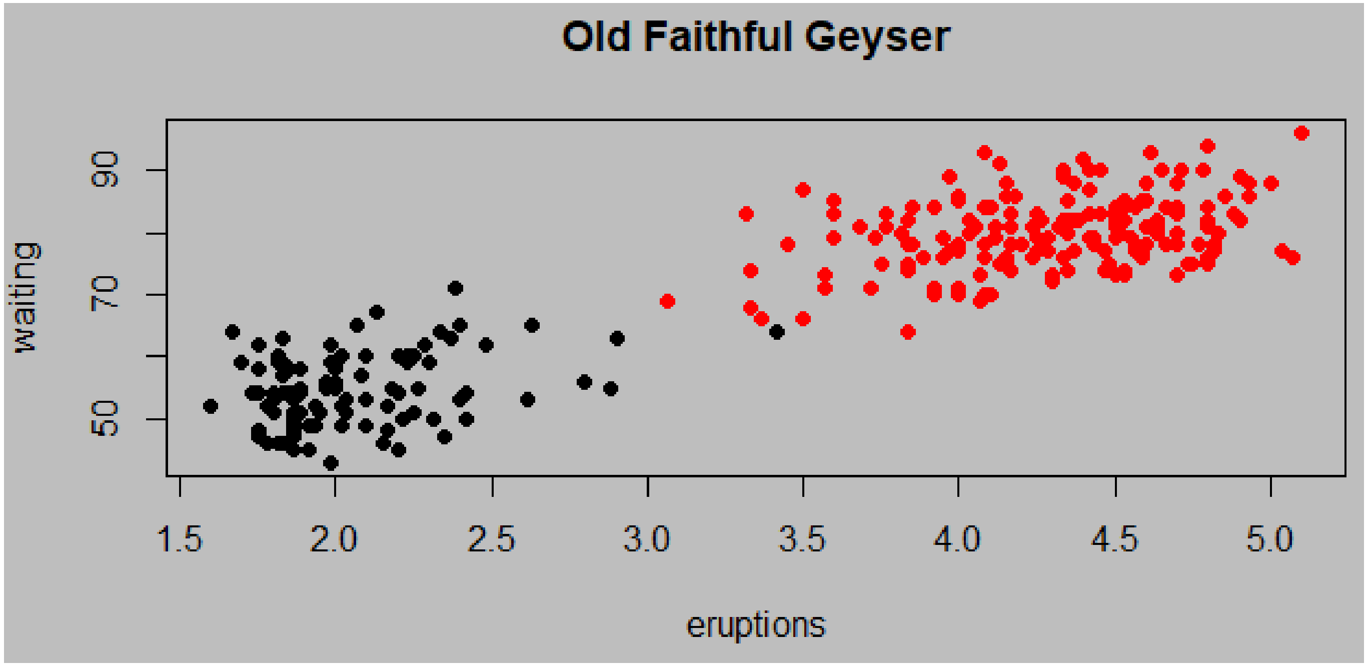
Unsupervised Learning 5



‘Elbow’

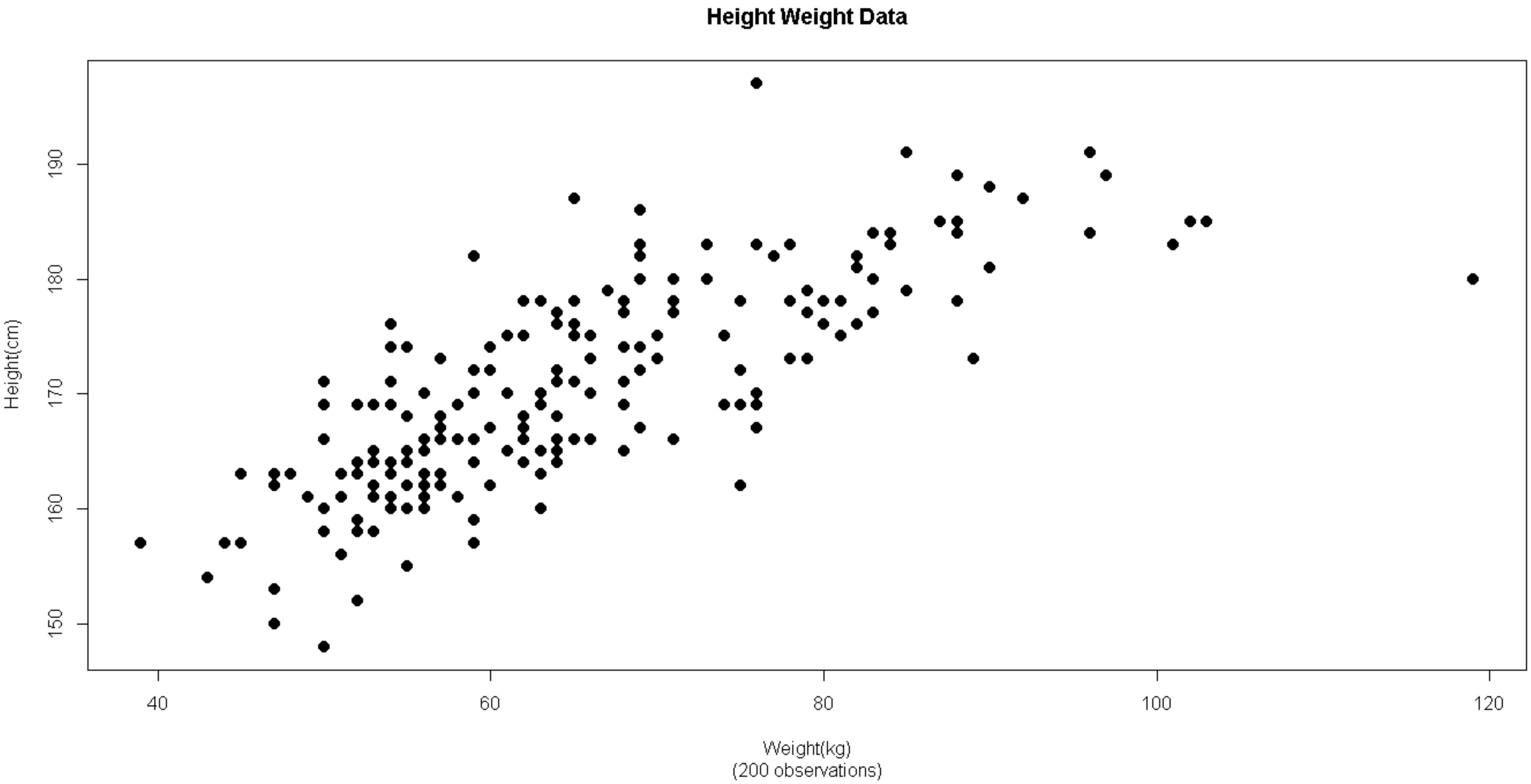
Unsupervised Learning 6

- Resulting Segmentation

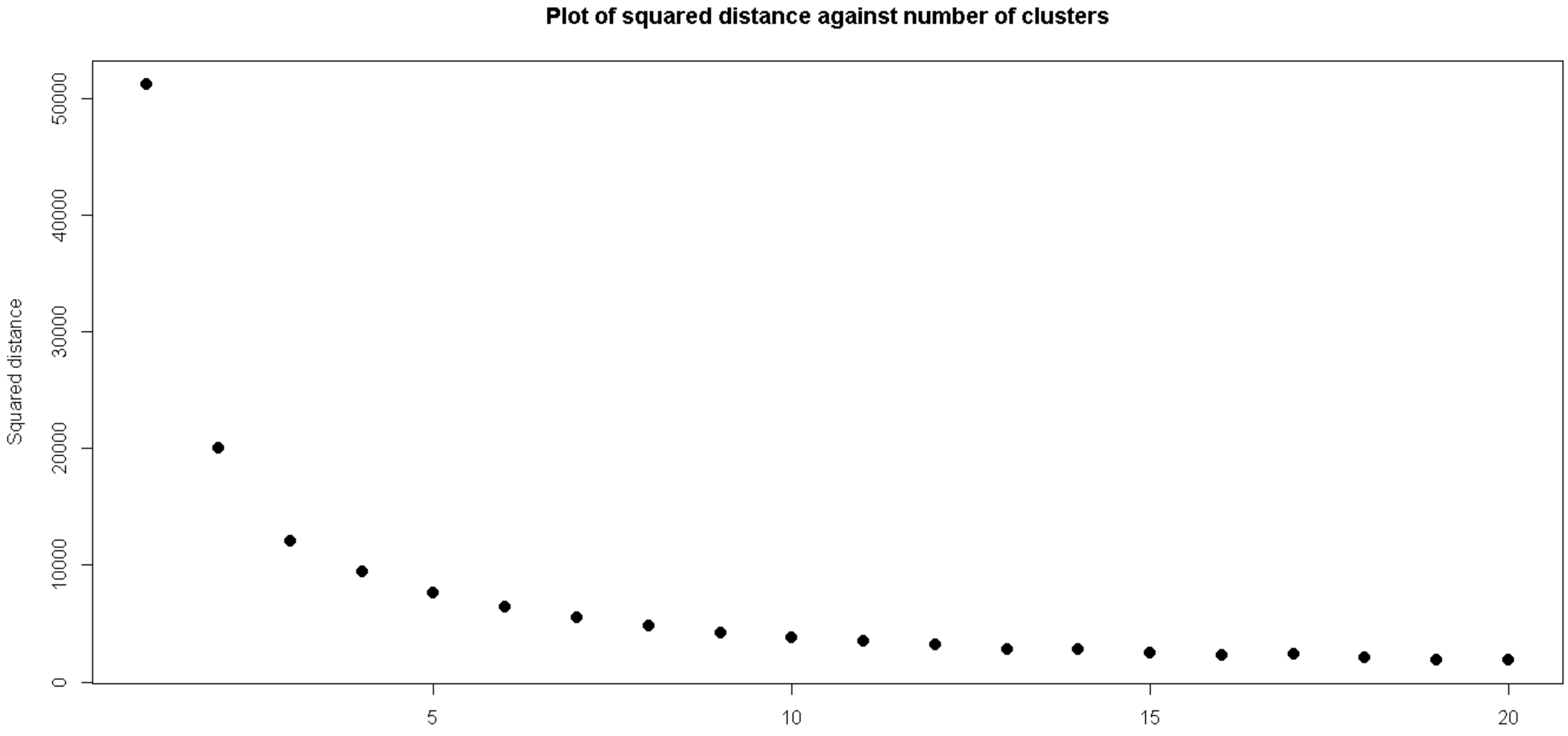


- Can be exploratory or detective

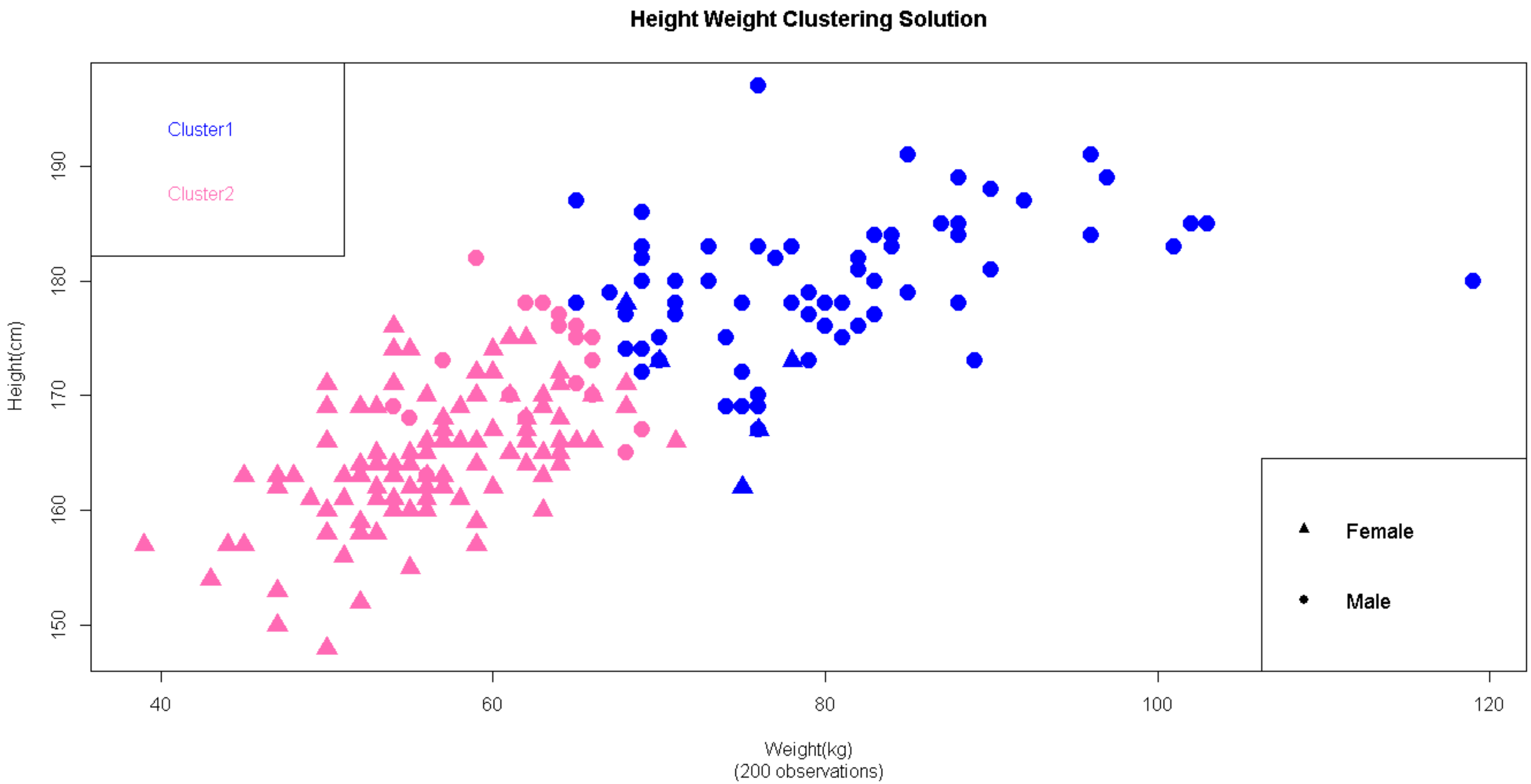
Another Grouping (Clustering) Example 1



Another Grouping (Clustering) Example 2



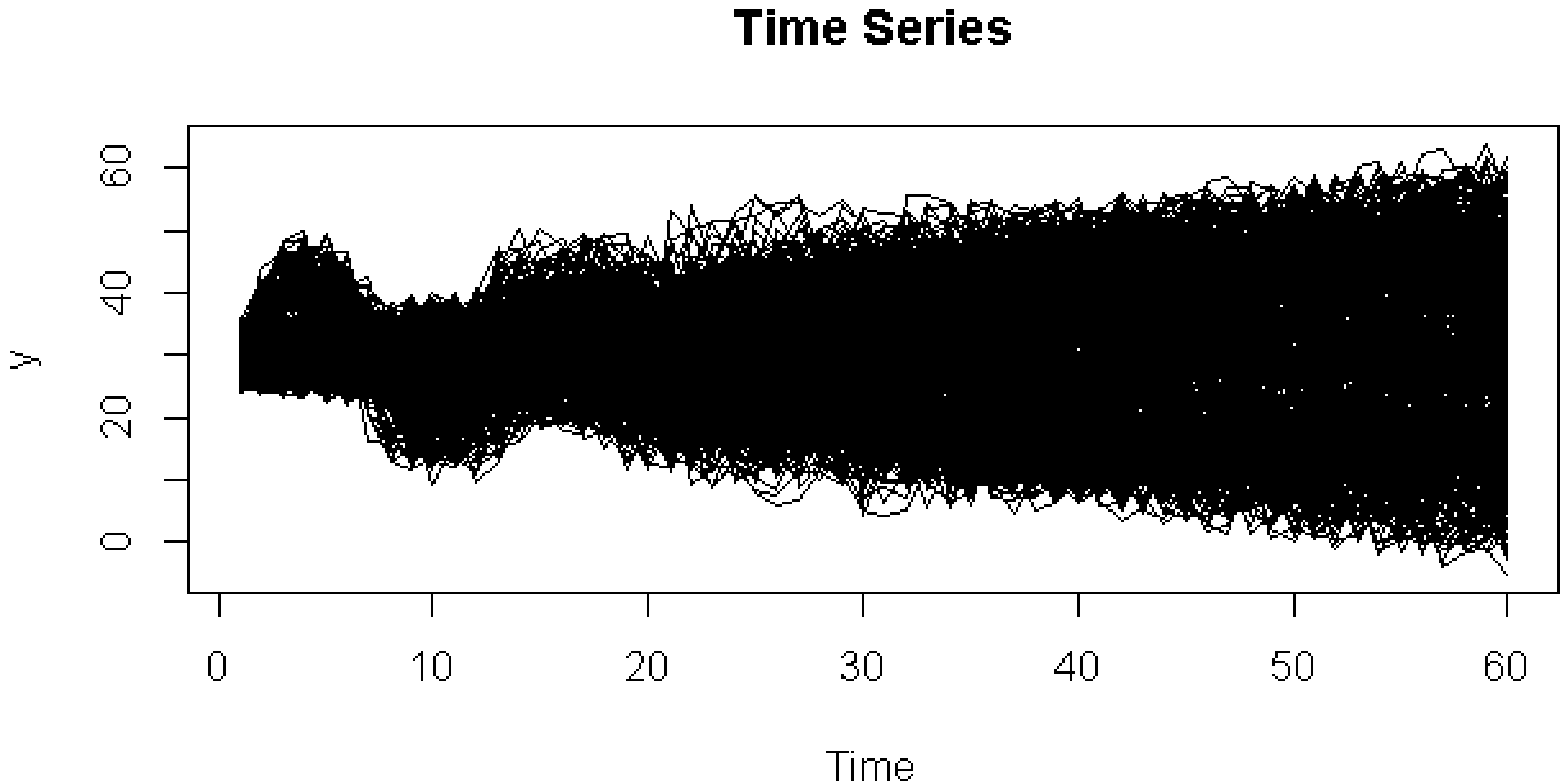
Another Grouping (Clustering) Example 3



Another Grouping (Clustering) Example 4

- Accuracy 88%
- ‘First pass’ result
- Readily implementable
- Methodology generalisable to n dimensions
- Where could this give more insight?
  - Segmentation (Distribution Channel)
  - Any homogeneous group selection
  - Deconstructing portfolios
  - Model point building
  - Outlier identification (Fraud etc.)
  - Trend analysis

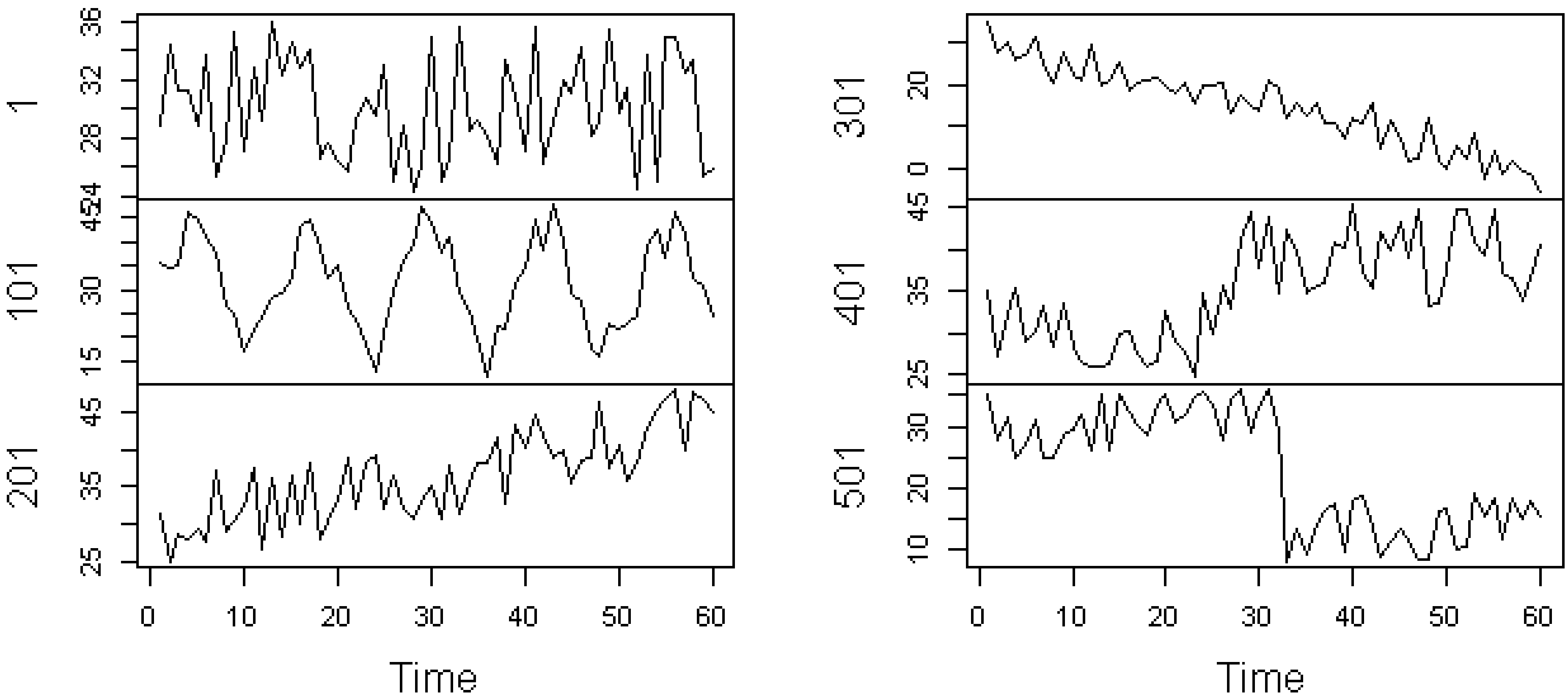
Deconstructing Trend Analysis 1



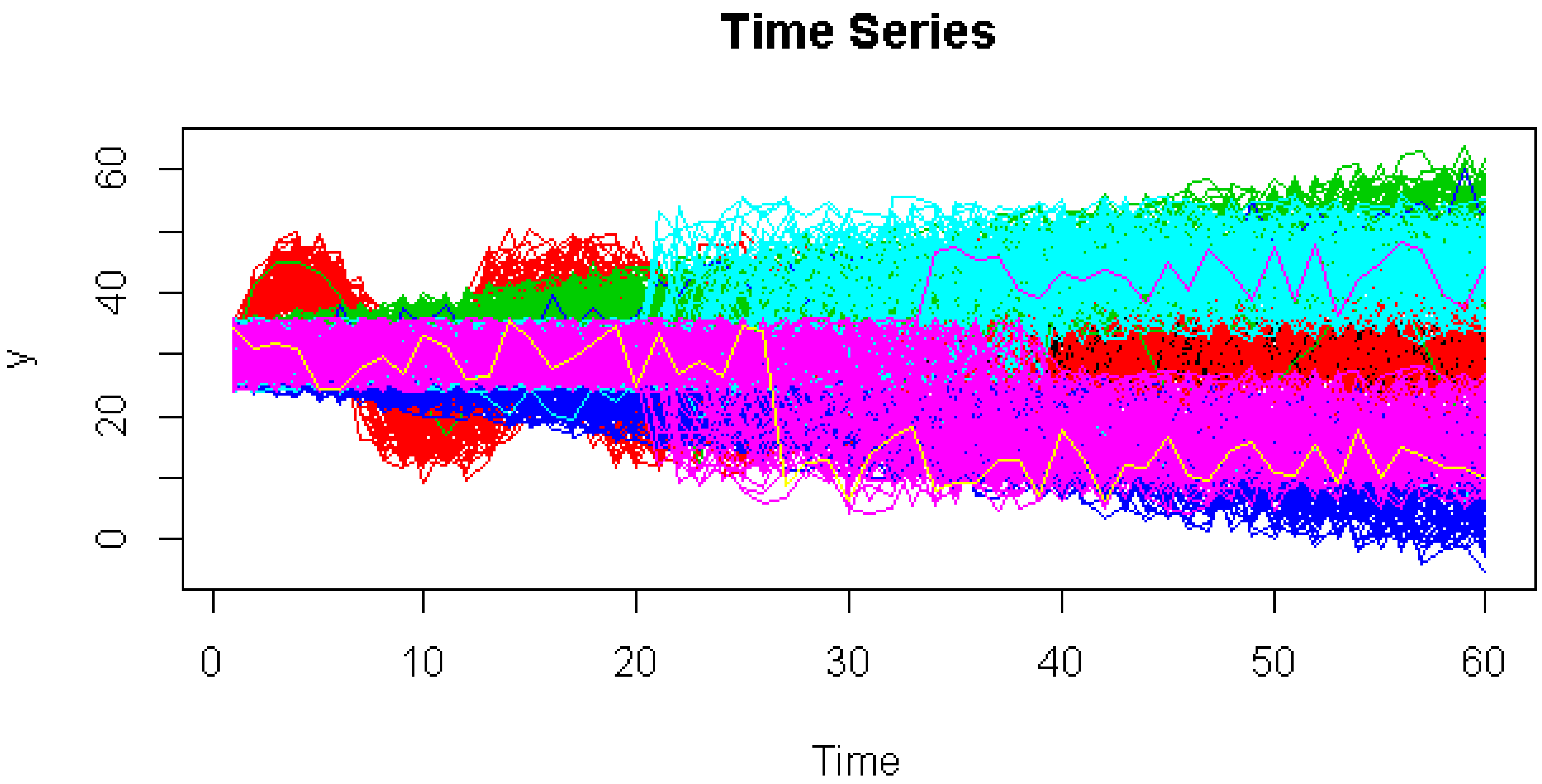
http://www.rdatamining.com/

Deconstructing Trend Analysis 2

- Constructed dataset
- 6 x 100 sub-series



Deconstructing Trend Analysis 3



Deconstructing Trend Analysis 4

Predicted Group / **Actual Group**

	1	2	3	4	5	6
1	97	3	0	0	0	0
2	1	99	0	0	0	0
3	0	0	81	0	19	0
4	0	0	0	63	0	37
5	0	0	16	0	84	0
6	0	0	0	1	0	99

- Accuracy 87%!

Deconstructing Trend Analysis 5

- Accuracy 87%!!!
- Where could this give more insight?
  - Claim rates
  - Seasonal / Selection Effects
  - Investment performance analysis
  - Stochastic model analysis
  - Trend analysis

Unsupervised Learning Summary

- Can help identify patterns in data
- Can help identify homogeneous groups
- Using computer power
- Relatively unsophisticated
- Possible to get answers quickly
- Perfect insight not possible
- Improved understanding may result

Cross Validation

- Should models be built on all data?
- Building and fitting on the same data, a good idea?
- But should we fit on a **Training** subset and **Test** on the remainder?
- This is **Cross Validation**

➔With a proposed model, exists in different forms

1. 75% **Training,** 25% **Test**
2. 10-fold validation – 90% **Training,** 10% **Test** repeated x 10
3. Leave one out validation – All bar one **Training,** one **Test** x n

➔With competing models

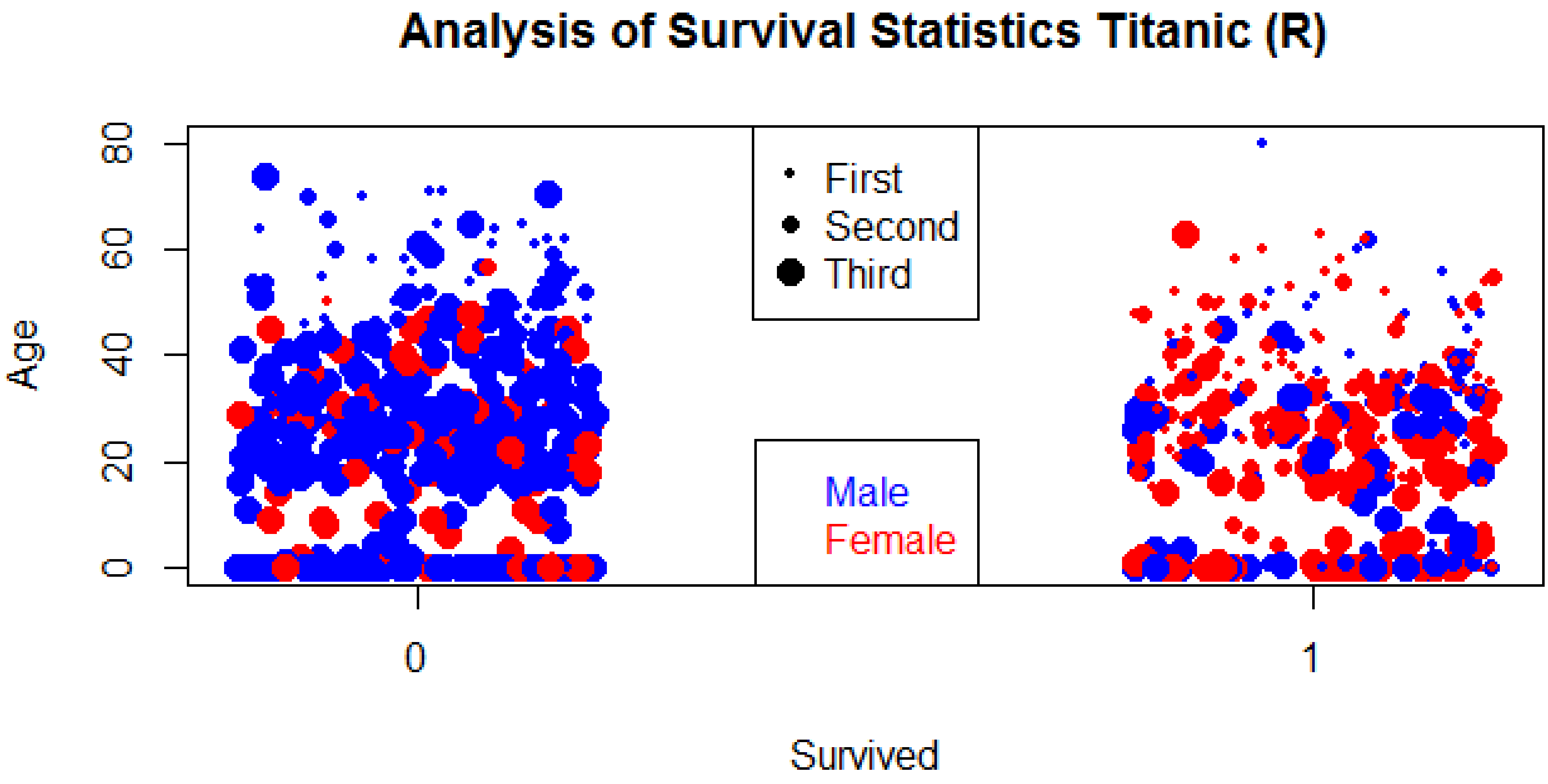
1. 50% **Training**, 25% **Validation**, 25% **Test**

Supervised Learning

- We propose a model...
- ...as opposed to just looking for patterns in the data
- Simple linear regression is a supervised learning method
- Many different models exist
- Can all be used for prediction
- Data accuracy as much an issue in analytics as in other actuarial work

Note:Examples are illustrative. They should not be taken to imply that any one technique is preferable to another or suitable for a particular situation

Titanic - Data Overview



Logistic Regression 1 – Model

- Model to be fitted
- P(Survived) = $\dfrac{\exp(\alpha + \beta_P PClass + \beta_S Sex + \ldots.)}{1 + exp(\alpha + \beta_P PClass + \beta_S Sex + \ldots.)}$
- Ensures P(Survived) falls between 0 and 1

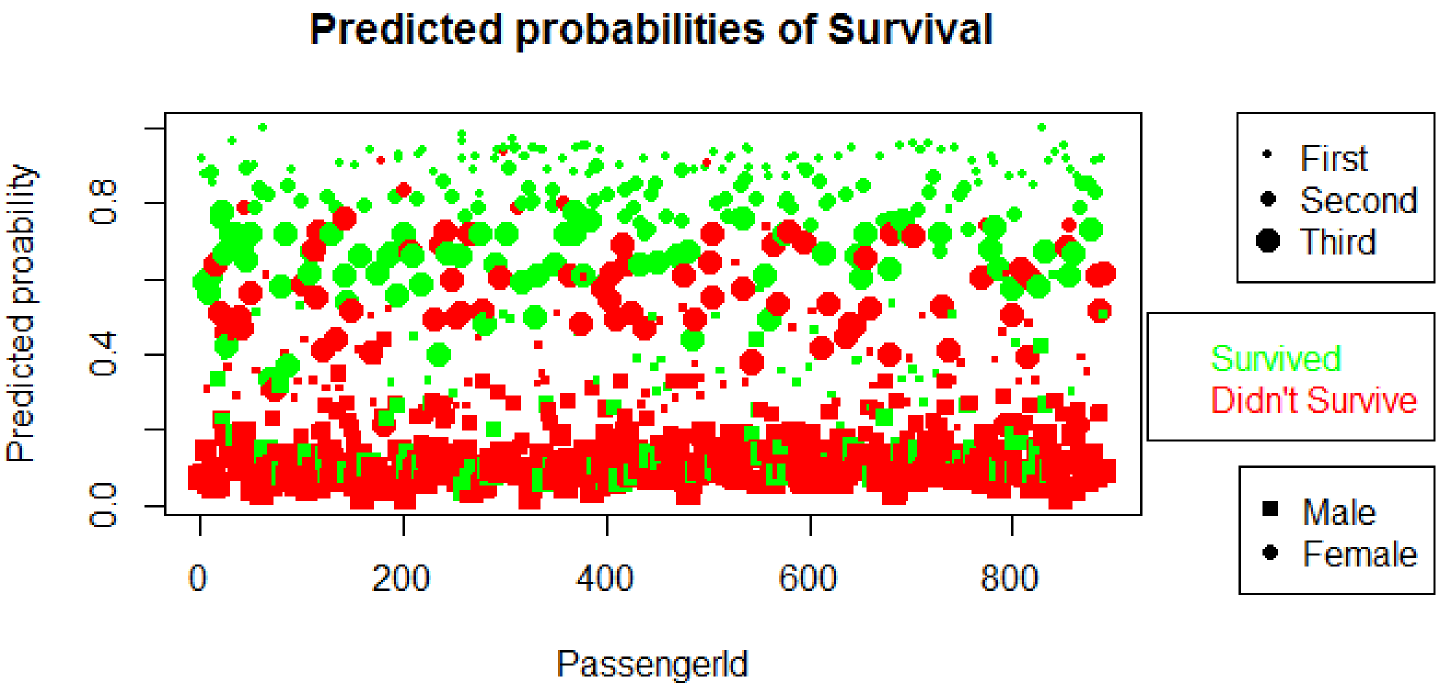
Logistic Regression 2 - Output

- Sample output

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   15.177420 621.502737   0.024  0.98052    
Pclass2       -0.655760   0.287176  -2.283  0.02240 *  
Pclass3       -1.865384   0.287024  -6.499 8.08e-11 ***
Sexmale       -2.719417   0.199488 -13.632  < 2e-16 ***
Age           -0.016693   0.005530  -3.019  0.00254 ** 
SibSp         -0.273558   0.102416  -2.671  0.00756 ** 
Parch         -0.056490   0.115601  -0.489  0.62508    
Fare           0.002932   0.002474   1.185  0.23591    
EmbarkedC    -12.091459 621.502712  -0.019  0.98448    
EmbarkedQ    -12.395355 621.502774  -0.020  0.98409    
EmbarkedS    -12.523889 621.502698  -0.020  0.98392    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression 3

- We now have a series of probabilities for each individual



- But we need an absolute prediction (1/0)
- Find value τ such that
  - prob > τ, predict survive
  - prob < τ, predict not survive

Logistic Regression 4

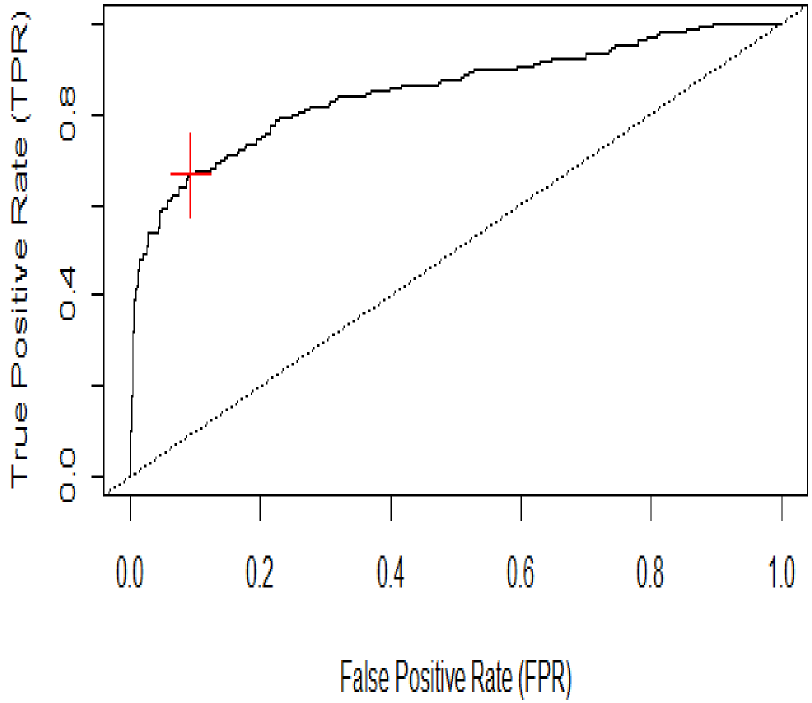
- Consider table

		Prediction Not Survive	Prediction Survive
Truth	Not Survive	True Negative (TN)	False Positive (FP)
Truth	Survive	False Negative (FN)	True Positive (TP)

- **True Positive Rate(TPR)**: TP / (TP + FN)
  – Of those that did survive, how many are classified correctly?
- **False Positive Rate(FPR)**: FP / (FP + TN)
  – Of those that did not survive, how many are classified wrongly?

Logistic Regression 5

- Plot TPR against FPR for various values of τ
- ROC (Receiver Operator Characteristic) graph shown below



- Select τ = 0.553 where TPR + (1-FPR) is maximised

Logistic Regression 6

- This then gives us the following summary

		Prediction Not Survive	Prediction Survive
Truth	Not Survive	498	114
Truth	Survive	51	228

- And an accuracy of 81.5% (498+228) / (498+228+51+114)

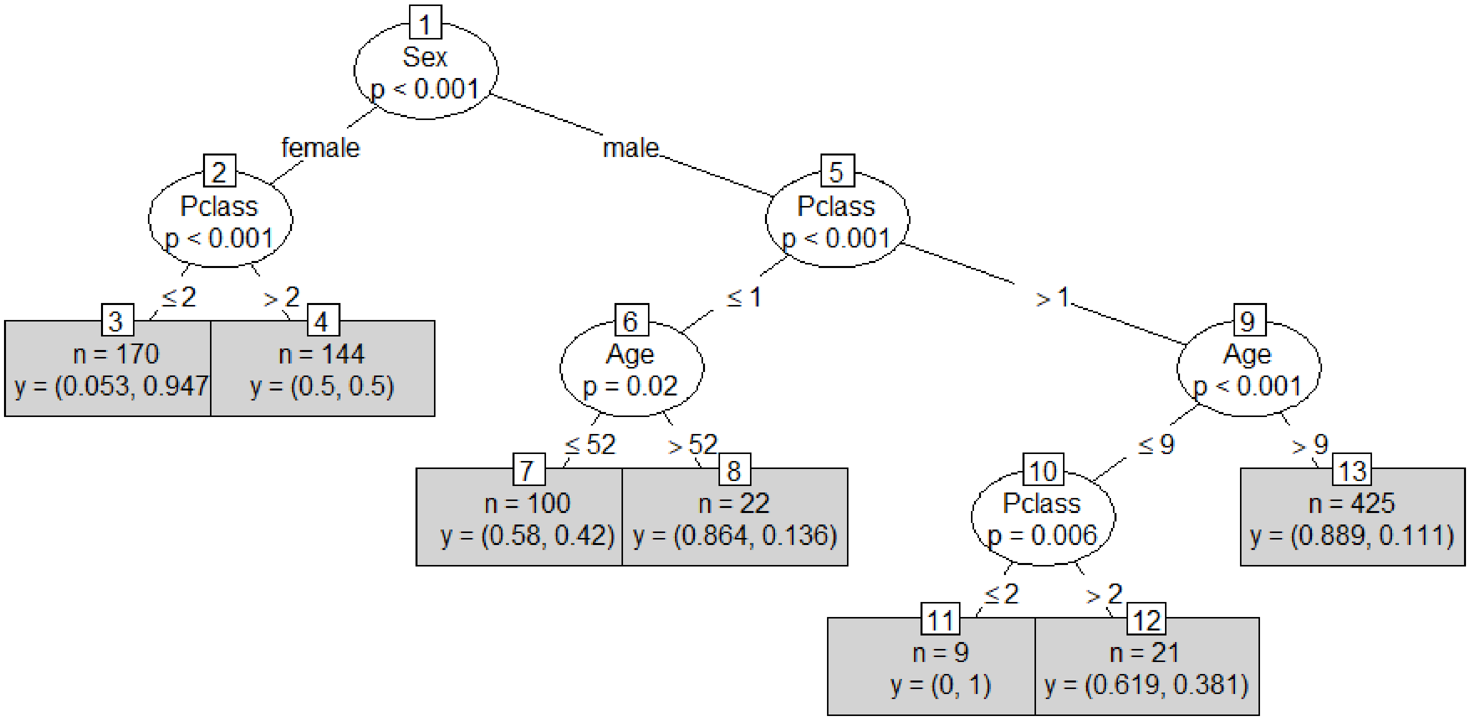
Logistic Regression 7

- We've fitted a model to give probabilities of survival

- P(Leonardo survives) = 0.1
- P(Kate survives) = 0.9
- We've then looked to find a single value above which we predict survival and below which we don't predict survival

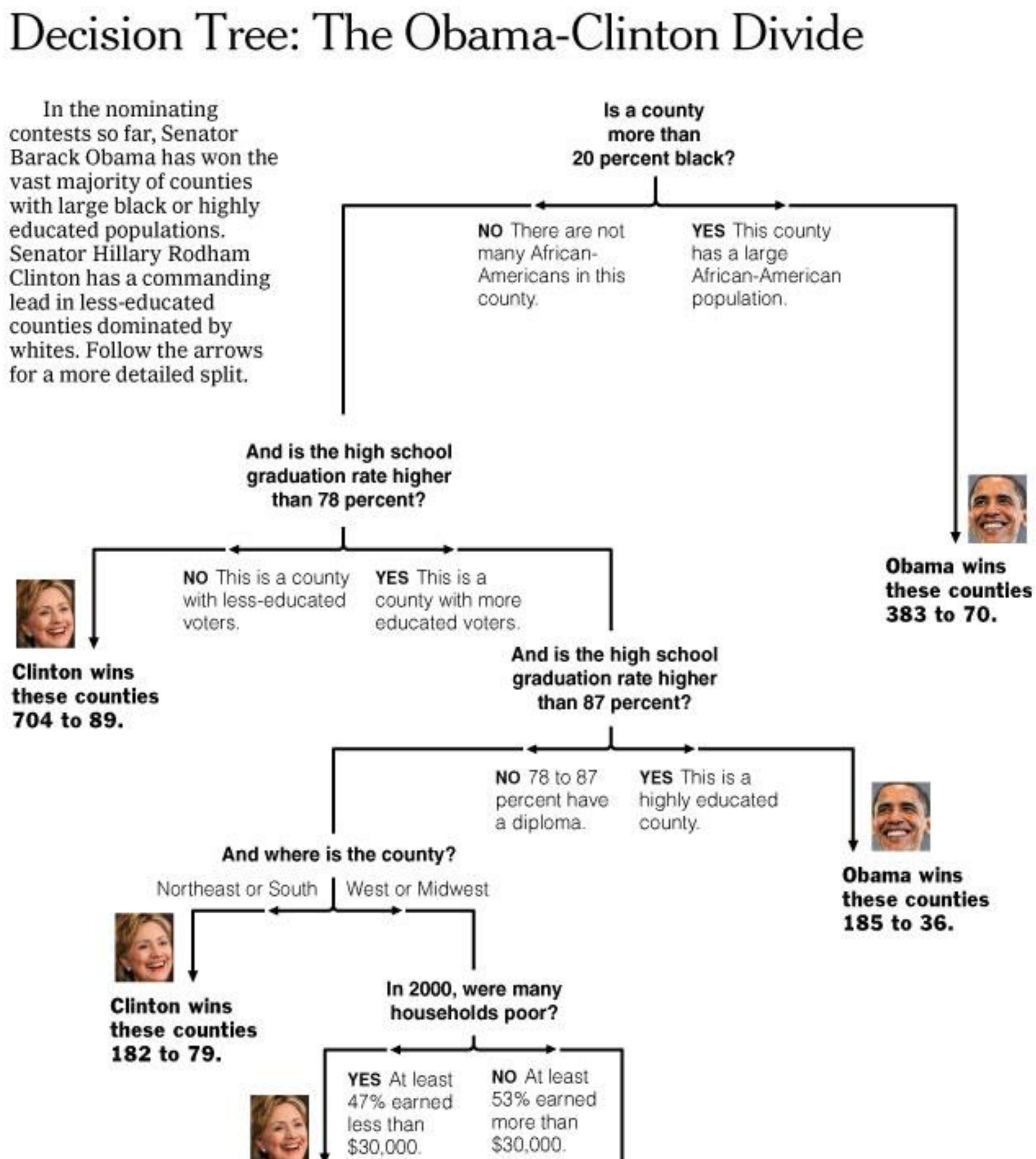
Classification (Decision) Trees 1 – Titanic

- Recursive partitioning



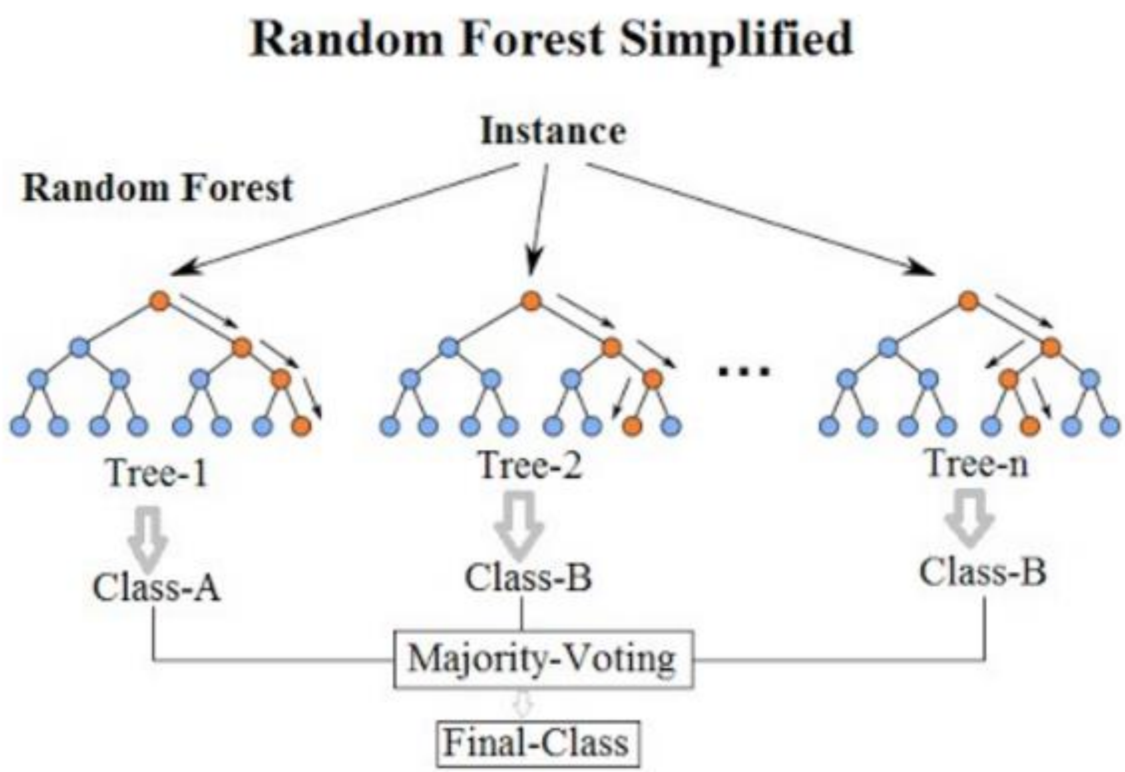
- Accuracy = 79.7%
- Greedy algorithm
- Can overfit

Classification (Decision) Trees 2 – US Elections



Random Forests 1

- Essentially multiple decision trees
- Analogy is between a single decision tree (~Dictator)
- Multiple decision trees (~Democracy)
- *Wisdom of crowds*



Random Forests 2

- Random Forests build on subsets of the data
- This introduces greater diversity
- This seems counter-intuitive
- But it’s compensated for building multiple branches
- Generally greatly increases performance but at the expense of interpretability
- Require some caution in practical use

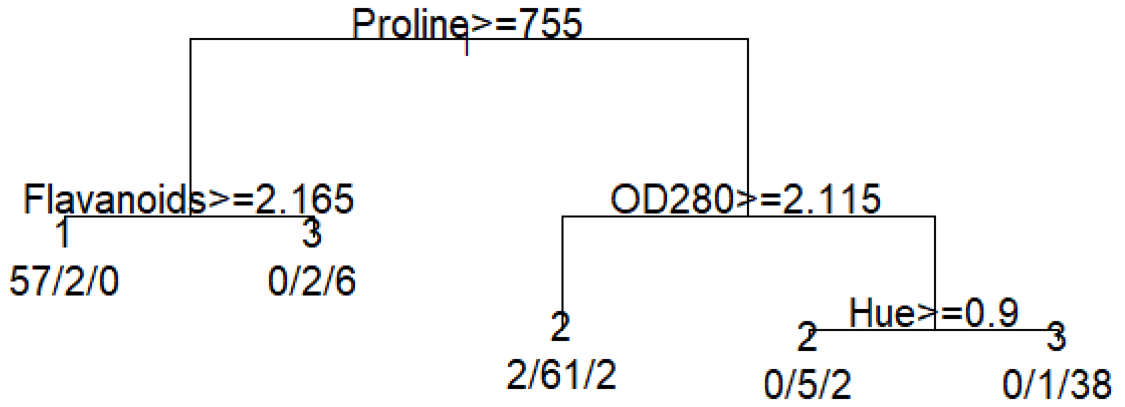
Random Forests 3 – Example

	Class	Alcohol	Alcalinity	...	Phenols	Proline
Sample 1	1					
Sample 2	2					
Sample 3	1					
: :						
Sample 178						

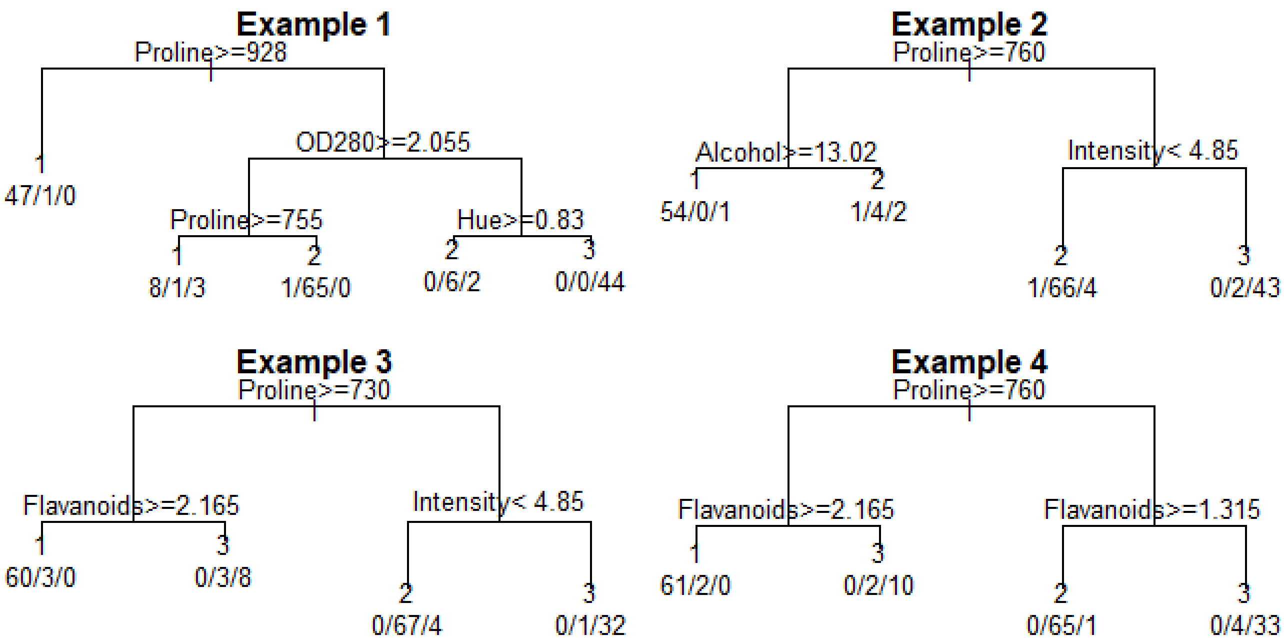
- Three classes of wine
- 13 measurements on each wine subsets of the data
- Let’s compare a single classification tree against a random forest

Random Forests 4

- **Single Decision Tree**



- **Random Forest Sample Trees**

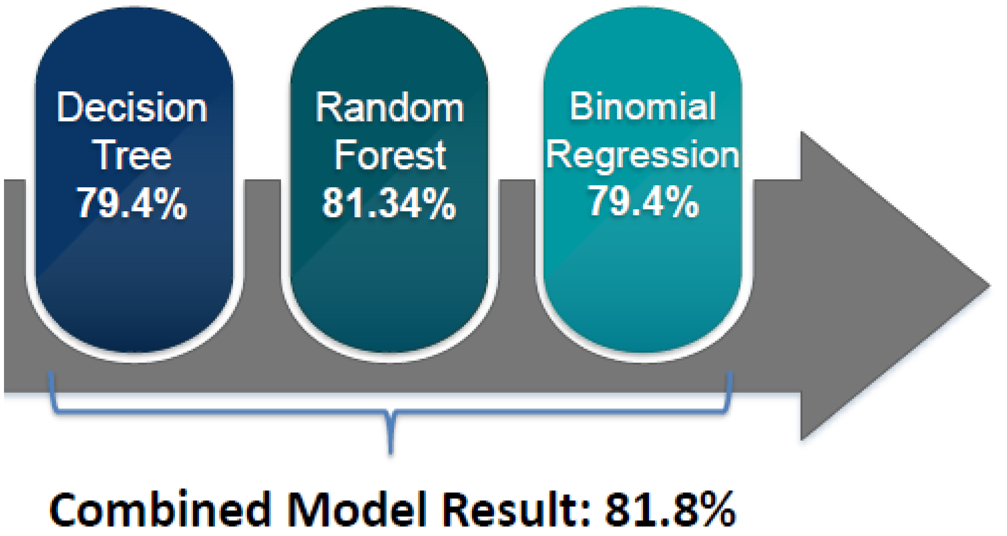


Random Forests 5 – Accuracy

- Single Tree – 93.8%
- Random Forest – 97.8%
- Generally get a significant boost to performance
- ...but at the expense of interpretability

Combining Models

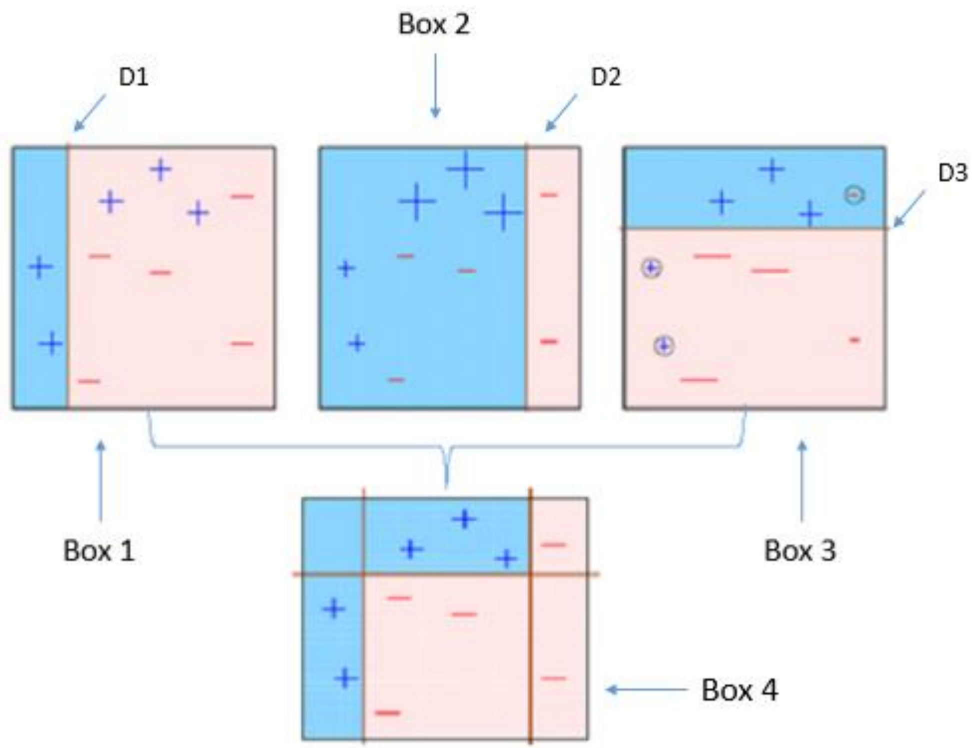
- Results from different (weak) models can be polled to combine into a stronger model



Source:Deloitte Team Presentation for SAI Titanic Competition

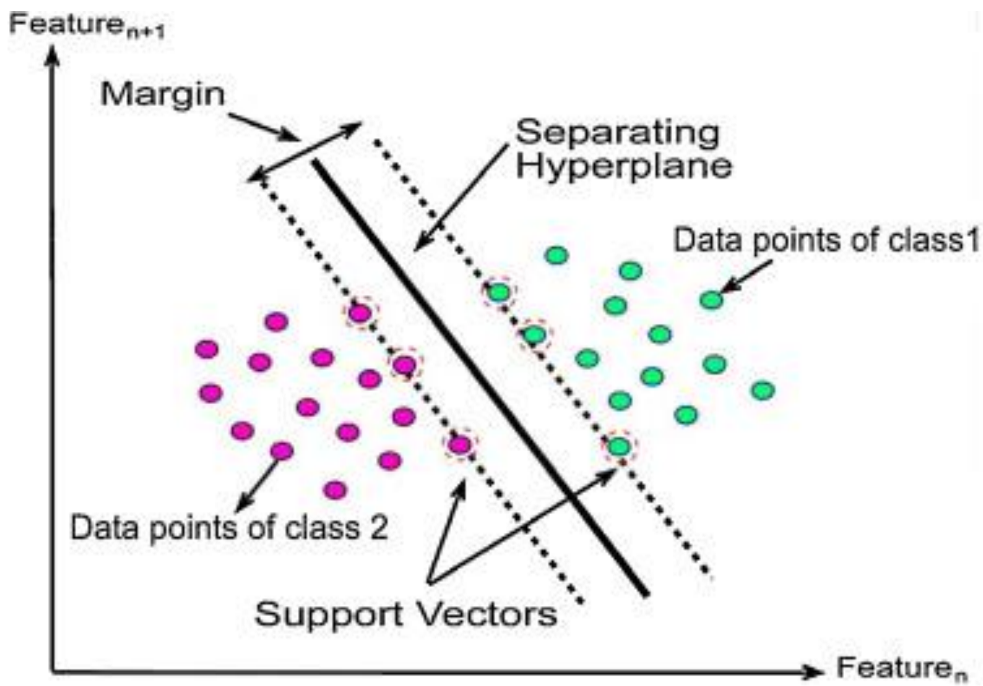
Improving Decision Trees – Boosting

- Boosting works to give misclassified observations greater weight
- Analogous to weighted regression



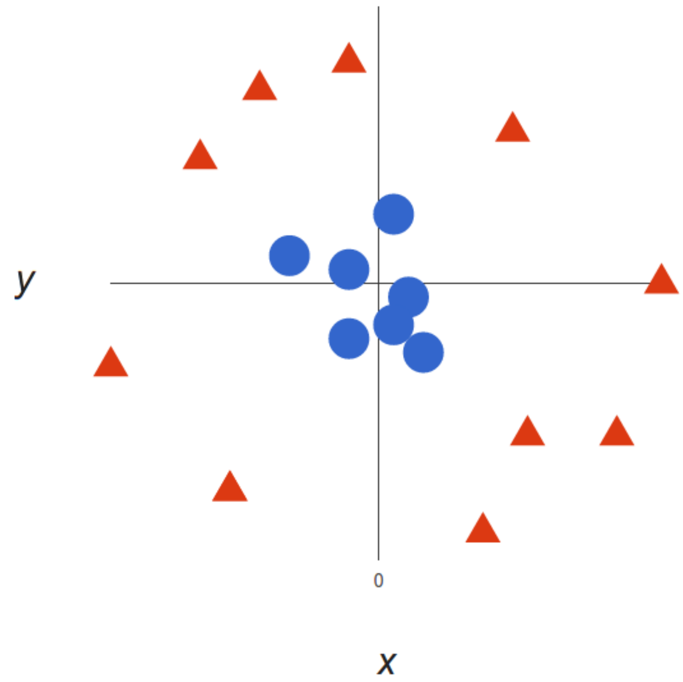
- Can lead to overfitting particularly if there is bad data
- Generally get a significant boost to performance

Support Vector Machines 1



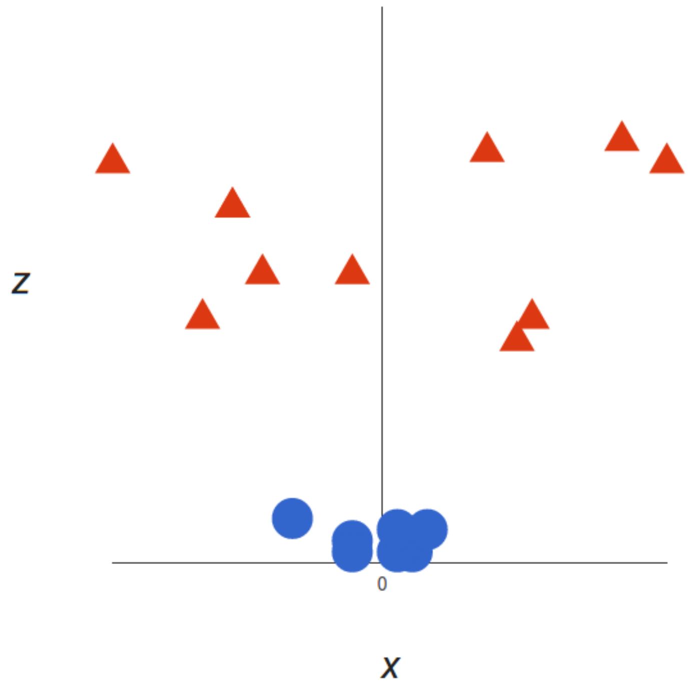
- With clearly separated data, support vectors are clear
- Any classification method would work well
- Life is rarely this simple!

Support Vector Machines 2



- How would we proceed here?
- Data is clearly separable but not linearly
- Move to a higher plane

Support Vector Machines 3



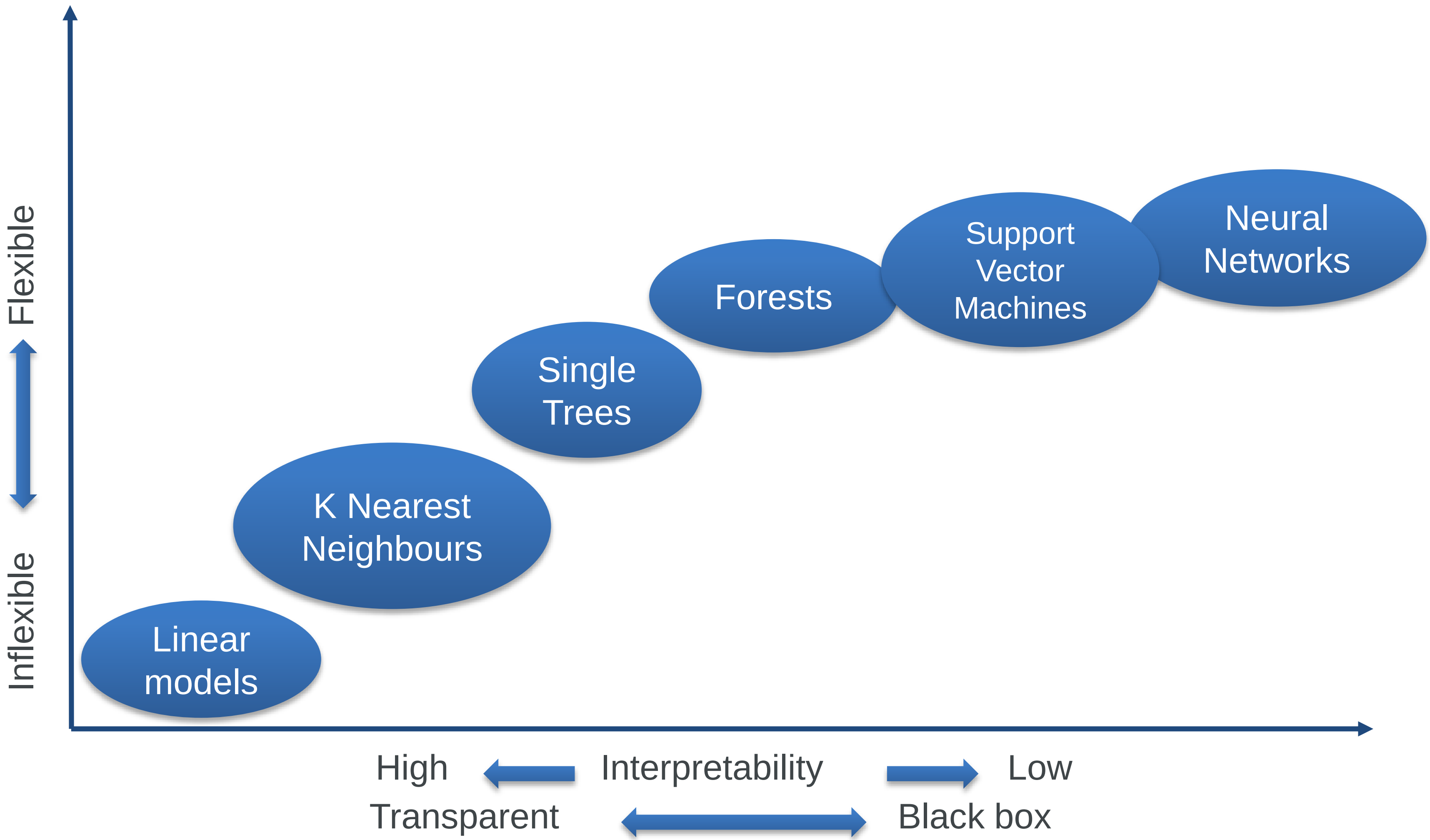
- Data is now linearly separable

Support Vector Machines 4 (Example)

- Spam dataset from HP
- 4,601 e-mails with 57 variables giving frequency of certain words and characters
- Build (train) competing models on a random 50% of the data
- Measure performance on remaining 50%

- Decision Tree – 89.0% accuracy
- Support VM – 93.0% accuracy

Comparing Algorithms



Supervised Learning - Summary

- Discipline of **Test** and **Training** data
- Logistic Regression – Suitable for binary classification
- Decision Trees – Can be improved with boosting
- Random Forests can improve performance but can be opaque
- Support Vector Machines can help with data which is not easily separable
- All models are readily implementable in most data science packages
- Neural Networks for another day

Honourable Mentions

- Primarily Numeric
- Visualisation
- Principal Component Analysis / Factor Analysis
- Twitter / Sentiment Analysis
- Outlier Detection
- Text mining / Word clouds
- Social Network Analysis
- Bayesian approaches
- Neural networks

Questions

- Thank you!