ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 1

Introduction to Data Analytics

Professor: Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Data Analytics

- Definition: Data Analytics is the process of aggregating large data sets in order to detect underlying patterns that might not be visible by just looking at raw data.
- These patterns give insight to maximize profits, improve health, lower electricity usage, etc.

Mountains of Data

- We now have more data being gathered/collected. Governments are starting to adopt openness policies of making public data freely available on the internet.
- Canada Open Government: http://open.canada.ca/en
- Seattle Open Data https://data.seattle.gov/
- Ontario Open Data https://www.ontario.ca/search/data-catalogue
- Ottawa Open Data: http://data.ottawa.ca/

Seattle Bicycle Traffic

- https://data.seattle.gov/Transportation/daily-bike-traffic/d4dx-u56x
- A traffic counter counts number of bicycles on the East and West sidewalks.
- There are traffic spikes from 7-9 am on the West side, and from 5-6pm, but only 5 days a week.

Profit

- As an entrepreneur, where would you sell hot dogs, or advertise?

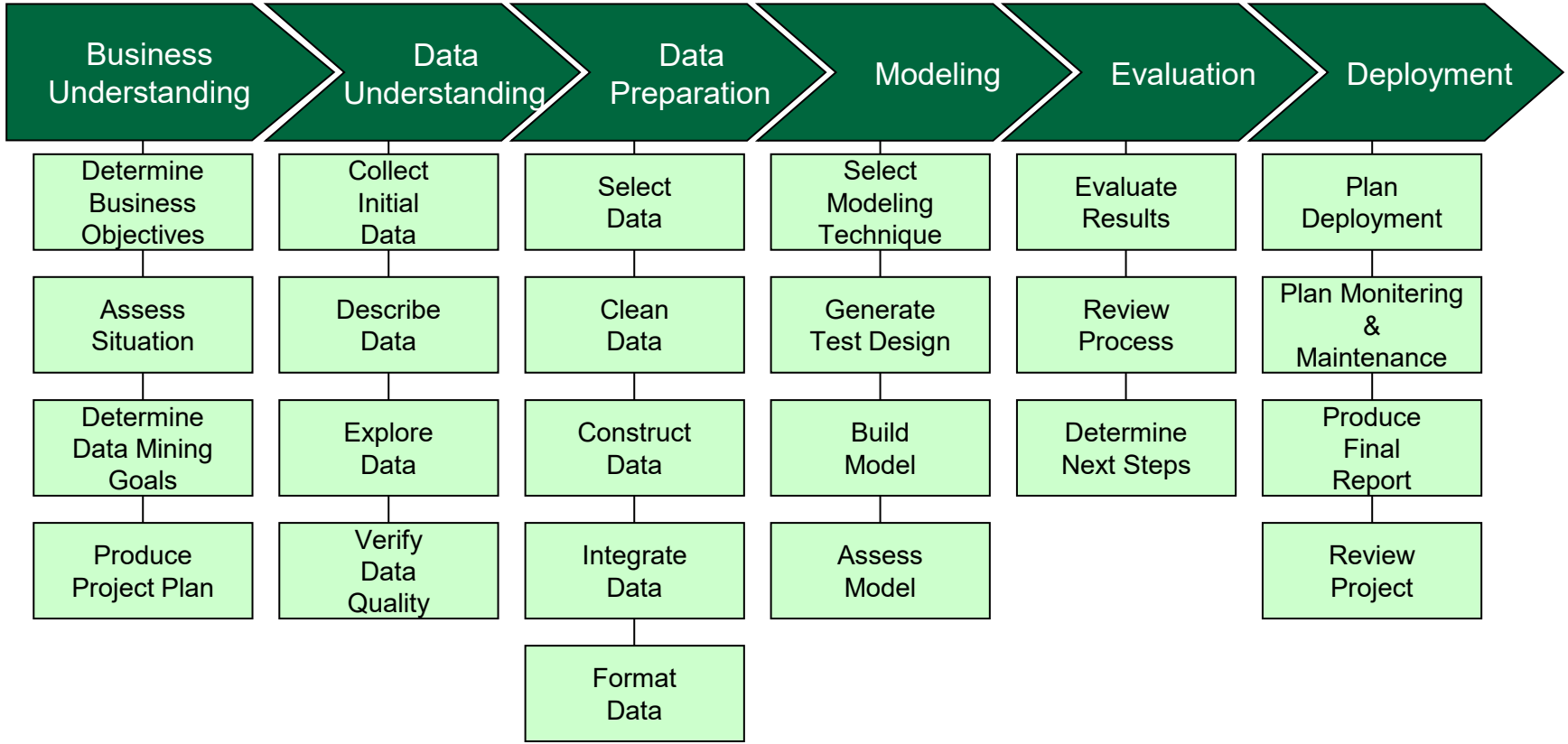
http://www.blogto.com/eat_drink/2015/07/everything_to_know_about_hot_dog_stands_in_toronto/

Healthcare

- The Real-World Benefits of Machine Learning in Healthcare
- Machine Learning Healthcare Applications – 2018 and Beyond

CRISP-DM

- Cross-industry Standard Process for Data Mining
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Format Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plan Monitering & Maintenance
Produce Final Report
Review Project

Understanding the business

- Identify what you want to accomplish from a business perspective
- Assess current situation
- Determine goals
- Produce project plan

Understanding data

- Describe data
- Explore data
- Verify data quality

Preparing data

- Select your data
- Clean your data
- Construct required data
- Integrate data

Types of Data

- Typically data fall in to one of 3 categories:
- Structured – The data are highly structured, where every element has the same fields: age, first name, last name, address, etc. Databases, objects are good examples
- Unstructured – The data have no common structure. News articles, websites, video, audio and photographs.
- Semi-structured – The data use some structure, but it is not common. This includes tree-type data like XML and JSON.

Data Preparation

- When getting data from different sources, some work is needed when putting it together:
  - Cleaning and filtering: Remove duplicate data, missing data, resolve incomplete data. Something like: ***Woodroffe Ave, Woodroffe, Woodroffe Avenue*** should all be the same.
  - Remove outliers: (data that is far outside the average). Every semester, some students register for a course but don't drop it. This means they get 0 for everything and lowers the class average. Another example is that sales for a store are $0 for regional some regional holidays.
  - Variable transformations. Changing how variables are represented (metric / imperial)

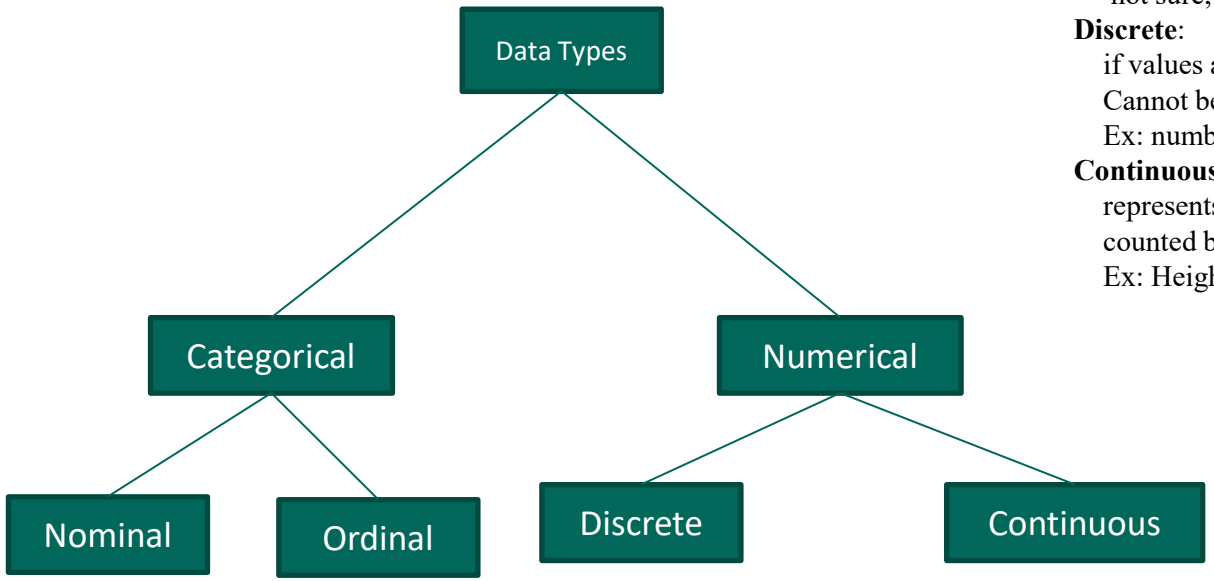
Data Preparation

- You get a data matrix with variables, attributes or features (columns)
- Instances are the rows (N).

Instance	Date	Price	Quantity	Label
1	May 2, 2016	5.50	1	Regular
2	May 28, 2016	3.79	2	Sale

- The data are typically brought together from various parts of the organization. They must be transformed into a single data format, for instance, price must all be $USD, or temperatures must all be Celsius instead of Fahrenheit.

Data Types



Nominal:
Gender: male, female

Ordinal:
survey questions: Strongly agree, agree, not sure, disagree, strongly disagree

Discrete:
if values are distinct and separate.
Cannot be measured but can be counted
Ex: number of heads in 100 coin flips

Continuous:
represents measurements. Values cannot be counted but can be measured
Ex: Height, salary

Data cleansing

- How do you detect outliers?
  - One method is to sort the data. The outliers will be at either end of the sorted sequence.
- For tagged data, make all similar tags the same: Woodroffe Ave.
- What about missing data? Replace with random numbers from average and standard deviation?
- Replace with “Missing” or “Unknown” tag.

Real Data from UCI Machine Learning Repository

Wine Dataset

```
1,13.9,1.68,2.12,16,101,3.1,3.39,.21,2.14,6.1,.91,3.33,985
1,14.1,2.02,2.4,18.8,103,2.75,2.92,.32,2.38,6.2,1.07,2.75,1060
1,13.94,1.73,2.27,17.4,108,2.88,3.54,.32,2.08,8.90,1.12,3.1,1260
1,13.05,1.73,2.04,12.4,92,2.72,3.27,.17,2.91,7.2,1.12,2.91,1150
1,13.83,1.65,2.6,17.2,94,2.45,2.99,.22,2.29,5.6,1.24,3.37,1265
1,13.82,1.75,2.42,14,111,3.88,3.74,.32,1.87,7.05,1.01,3.26,1190
1,13.77,1.9,2.68,17.1,115,3,2.79,.39,1.68,6.3,1.13,2.93,1375
1,13.74,1.67,2.25,16.4,118,2.6,2.9,.21,1.62,5.85,.92,3.2,1060
1,13.56,1.73,2.46,20.5,116,2.96,2.78,.2,2.45,6.25,.98,3.03,1120
1,14.22,1.7,2.3,16.3,118,3.2,3,.26,2.03,6.38,.94,3.31,970
1,13.29,1.97,2.68,16.8,102,3,3.23,.31,1.66,6,1.07,2.84,1270
1,13.72,1.43,2.5,16.7,108,3.4,3.67,.19,2.04,6.8,.89,2.87,1285
2,12.37,.94,1.36,10.6,88,1.98,.57,.28,.42,1.95,1.05,1.82,520
2,12.33,1.1,2.28,16,101,2.05,1.09,.63,.41,3.27,1.25,1.67,680
2,12.64,1.36,2.02,16.8,100,2.02,1.41,.53,.62,5.75,.98,1.59,450
```

Breast Cancer Dataset

```
842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189
842517,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.003532,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,152.5,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758
84348301,M,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956,1.156,3.445,27.23,0.00911,0.07458,0.05661,0.01867,0.05963,0.009208,14.91,26.5,98.87,567.7,0.2098,0.8663,0.6869,0.2575,0.6638,0.173
84358402,M,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.7813,5.438,94.44,0.01149,0.02461,0.05688,0.01885,0.01756,0.005115,22.54,16.67,152.2,1575,0.1374,0.205,0.4,0.1625,0.2364,0.07678
843786,M,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08089,0.2087,0.07613,0.3345,0.8902,2.217,27.19,0.00751,0.03345,0.03672,0.01137,0.02165,0.005082,15.47,23.75,103.4,741.6,0.1791,0.5249,0.5355,0.1741,0.3985,0.1244
844359,M,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.074,0.1794,0.05742,0.4467,0.7732,3.18,53.91,0.004314,0.01382,0.02254,0.01039,0.01369,0.002179,22.88,27.66,153.2,1606,0.1442,0.2576,0.3784,0.1932,0.3063,0.08368
84458202,M,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1.377,3.856,50.96,0.008805,0.03029,0.02488,0.01448,0.01486,0.005412,17.06,28.14,110.6,897,0.1654,0.3682,0.2678,0.1556,0.3196,0.1151
844981,M,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.406,24.32,0.005731,0.03502,0.03553,0.01226,0.02143,0.003749,15.49,30.73,106.2,739.3,0.1703,0.5401,0.539,0.206,0.4378,0.1072
84501001,M,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.599,2.039,23.94,0.007149,0.07217,0.07743,0.01432,0.01789,0.01008,15.09,40.68,97.65,711.4,0.1853,1.058,1.105,0.221,0.4366,0.2075
845636,M,16.02,23.24,102.7,797.8,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,1.187,2.466,40.51,0.004029,0.009269,0.01101,0.007591,0.0146,0.003042,19.19,33.88,123.8,1150,0.1181,0.1551,0.1459,0.09975,0.2948,0.08452
84610002,M,15.78,17.89,103.6,781,0.0971,0.1292,0.09954,0.06606,0.1842,0.06082,0.5058,0.9849,3.564,54.16,0.005771,0.04061,0.02791,0.01282,0.02008,0.004144,20.42,27.28,136.5,1299,0.1396,0.5609,0.3965,0.181,0.3792,0.1048
846226,M,19.17,24.8,132.4,1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9555,3.568,11.07,116.2,0.003139,0.08297,0.0889,0.0409,0.04484,0.01284,20.96,29.94,151.7,1332,0.1037,0.3903,0.3639,0.1767,0.3176,0.1023
846381,M,15.85,23.95,103.7,782.7,0.08401,0.1002,0.09938,0.05364,0.1847,0.05338,0.4033,1.078,2.903,36.58,0.009769,0.03126,0.05051,0.01992,0.02981,0.003002,16.84,27.66,112,876.5,0.1131,0.1924,0.2322,0.1119,0.2809,0.06287
```

Statistics

Random Number Generators

- The numbers seem random but they're not. They're pseudo-random
- The sequence of numbers generated depends on the starting "seed"
- ***Two number generators will produce the same numbers if they have the same seed.***
- Create two Random Objects with the constructor: Random(int seed)
- call nextInt() on both. They will produce the same sequence.
- To create a new sequence every time, use: System.currentTimeMillis() as the seed

Statistics

- Given a set of data (Numbers), there are several things we can compute:
- Mean: What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$
- Median: What is the middle item? Array[size/2]
- Mode: What item appears the most often:
- 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5 ?
- Order Statistics: What is the 3rd largest number? What is (n-2)nd largest number?

Computing Mean

- Computing the mean tells the average value, but sometimes you want to use different formulae:
- Moving average: What is the average of values in the last 5 days?
- If N is the total sum, and the day changes, you don't need to recompute the mean. Instead, add the new values, subtract the expired values, and divide by the new number of items, N.

Computing Rolling Mean

- If the mean, N, represents the average of 5 days:

  [1, 2],[3, 4], [4, 6], [5, 7], [9, 11]

- $N = (1 + 2 + 3 + 4 + 4 + 6 + 5 + 7 + 9 + 11)/10 = 5.2$
- Now new data are ready to be added: 10, 12
- Imagine that [1, 2] are now out of date and no longer part of your computation. Just add the new data, subtract the old data, and divide by the number of data:
- $N' = N + [(10+12) - (1+2)]/10$

  $= 5.2 + 1.9$

  $= 7.1$

- Verify: $(3 + 4 + 4 + 6 + 5 + 7 + 9 + 11 + 10 + 12)/10 = 7.1$

Weighted Average

- Suppose you want more recent data to be worth more than older data, like for predicting gas prices.
- Decide the comparative weights of the components:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}})$$

- Now divide by the weighted number of elements:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}}) \\ / (5 + 3 + 2 + 1)$$

Sample vs Population

- A population data set contains all members of a specified group (the entire list of possible data values)
  - Example: all people in Ottawa

- A sample data set contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken.
  - Example: some people in Ottawa

Standard Deviation

- Let X be an array: X = {21, 37, 13, 25, 32, 8 }
- What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$ = 22.6667
- **Standard deviation**: N is the number of elements
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/N}$. This formula is used when you have measured the entire population. In Excel, this is stdev.p() = 10.07748
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/(N-1)}$. This formula is used when you have only part of the data. In Excel, this is stdev.s() = 11.03932

Standard Deviation

	x	x-Mean	$(x\text{-Mean})^2$
	21	-1.6667	2.7778
	37	14.3333	205.4444
	13	-9.6667	93.4444
	25	2.3333	5.4444
	32	9.3333	87.1111
	8	-14.6667	215.1111
Mean	22.6667		

Sum of $(x\text{-Mean})^2$	609.3333
Sum / 6	101.5556
sqrt (sum/6)	10.0775

Mean and Variance

- Difference of means is the difference between the averages of two samples
- Variance is just standard deviation squared.

- https://www.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/v/range-variance-and-standard-deviation-as-measures-of-dispersion

- https://www.khanacademy.org/math/statistics-probability/displaying-describing-data/sample-standard-deviation/v/statistics-sample-variance

Distributions

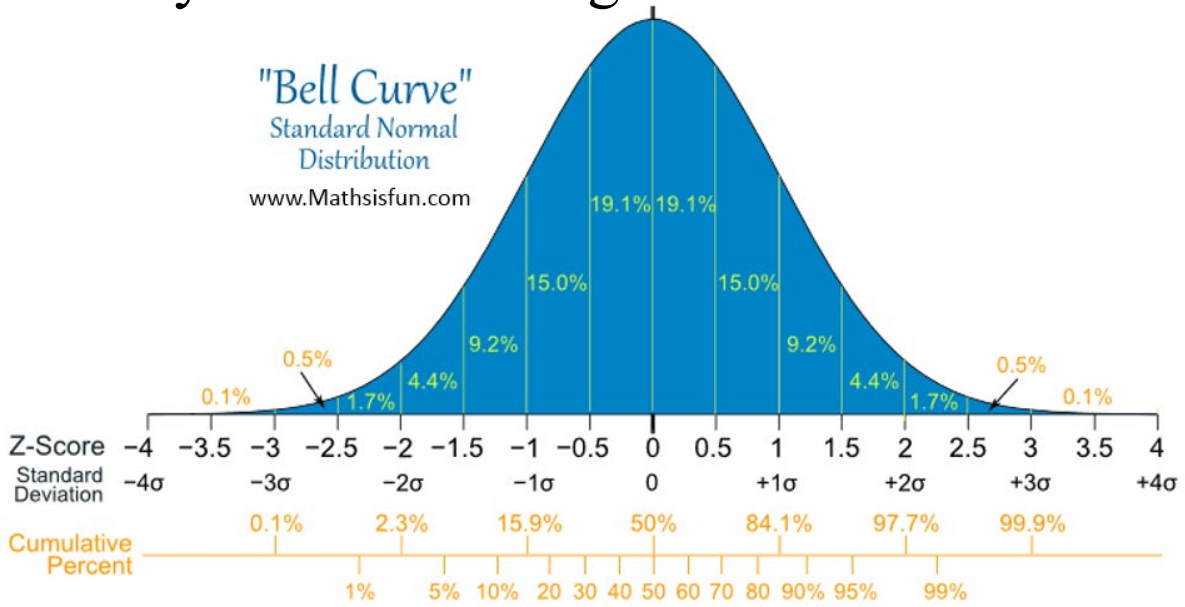
- Uniform – The probability of an event is equal (uniform). The probability of getting tails for flipping a coin, or rolling a 1 with a die.
- Gaussian (Normal) – The values are centered around a midpoint (mean), but decrease as you get farther from the mean: grades on a test.
- Geometric, Poisson, Exponential – These are other distributions that exist, but we don't have time to cover.

Normal / Gaussian

- Also called a bell curve, with midpoint of μ (pronounced me-you), and standard deviation of σ (pronounced sigma).
- The density of an event $(x \mid \mu, \sigma^2)$ is: $\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\sigma^2\pi}}$
- The variance is σ^2
- If $\mu = 0$ and $\sigma^2 = 1$ then this is called Standard Normal Distribution

Normal / Gaussian

- The area under the curve must add up to 1. Probabilities are calculated by a number being less than a number.



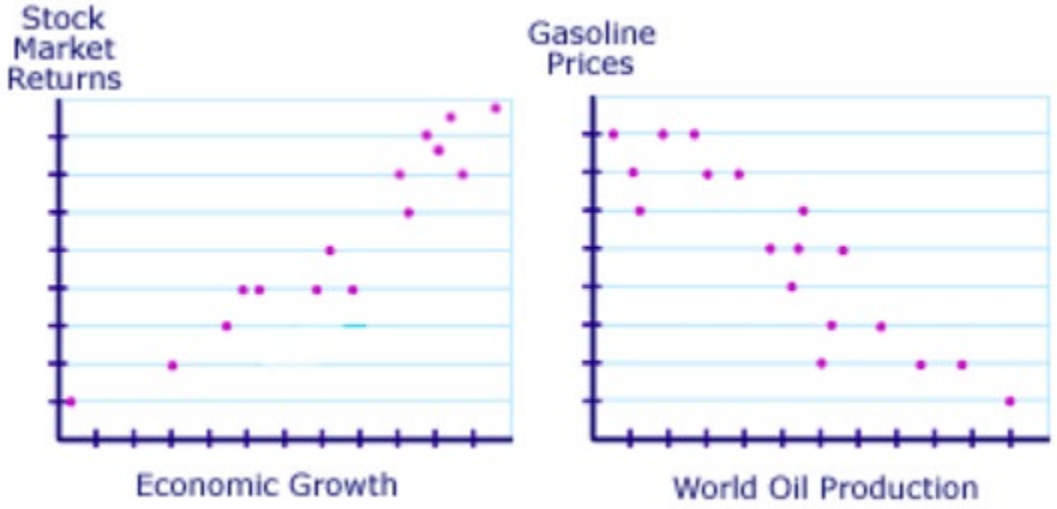
Rank Statistics

- Rank statistics compute where a number compares to the rest of the data, for instance to 5%, bottom 15%, etc.
- They are described in percentiles, meaning how much of the data is less than the number. 95% percentile means that 95% of the numbers are less. The median is the 50% percentile.

Covariance and correlation

- The covariance computes the strength and direction of the relationship of two sets of values. Do they both get bigger together? Does one get bigger as the other gets smaller? Is there no relationship?
- Covariance is calculated by: $$S_{xy} = \sum_{i=1}^{n} \frac{(X_i - \mu_x) * (Y_i - \mu_y)}{(n-1)}$$
- If S_{xy} is positive (large) then X and Y increase together. If it is negative, then X and Y are inversely related. If it is 0, then there is no relationship.

Covariance



http://ci.columbia.edu/ci/premba_test/c0331/s7/s7_5.html

Covariance and correlation

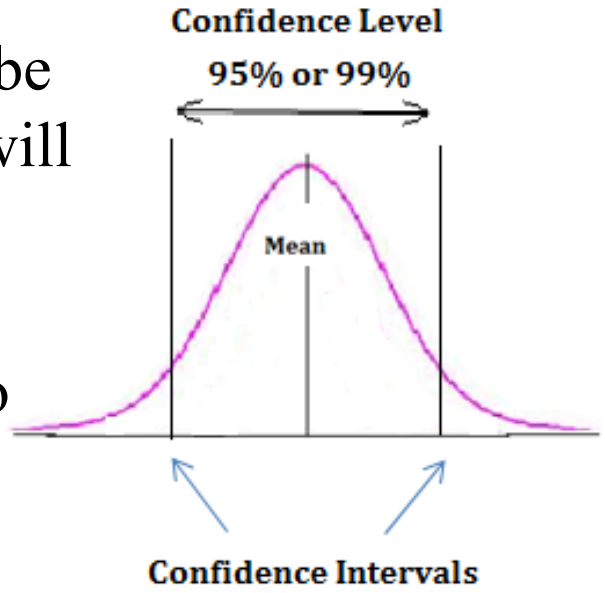
- Correlation also tells you the degree to which the variables tend to move together
- Strength of the relationship
- Correlation R_{xy} is calculated by:

$$R_{xy} = \frac{S_{xy}}{\sigma_x \sigma_y}$$

- It turns the Correlation into a number between -1 and +1

Confidence Intervals

- Confidence intervals describe the uncertainty of a parameter. If you repeatedly take a small sample to measure, your values will always be different. The mean you repeatedly sample will also follow a normal distribution.
- The confidence intervals calculates the probability that the actual mean falls close to your measured number.



Confidence Intervals

- The formula for calculating the confidence interval is: $\mu \pm z_c \left(\frac{\sigma}{\sqrt{n}}\right)$
- Z_c is the "critical value" for difference confidence levels: 90% is 1.645, 95% is 1.96, 99% is 2.575.
- This computes the limits for 90, 95 or 99% of the data.

- The 90% confidence interval says there is a 90% chance that the true mean falls within the range you have measured +/- some error
- The 95% confidence interval says there is 95% chance that it falls within a larger range.

Review

- Don’t memorize the formulas for probabilities! Just be familiar with the names: Uniform, Gaussian/Normal.
- What are rank statistics and percentiles?
- Know the terms: mean median, mode and variance
- Learn the formula for calculating mean and standard deviation for a data set.

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 2

Learning

Classification by kNN

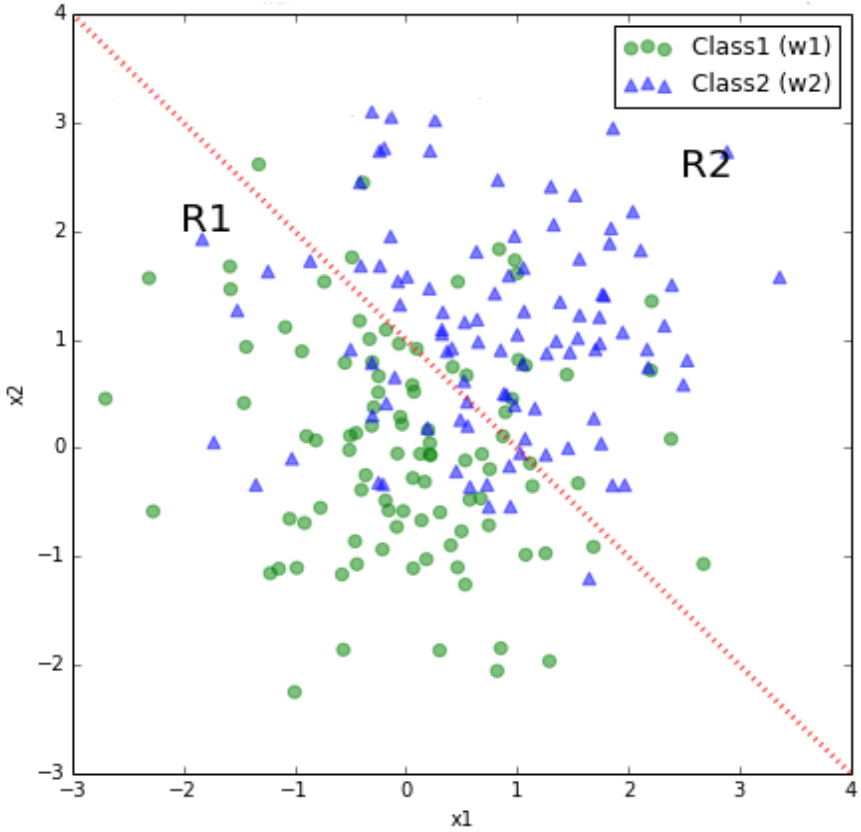
Professor: Dr. Anu Thomas

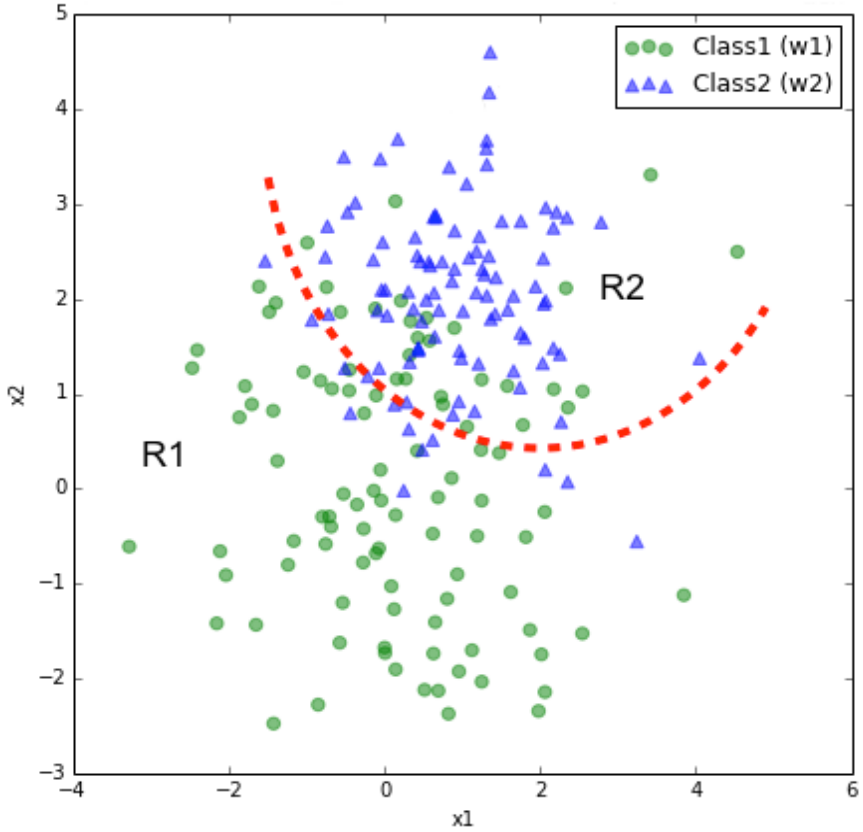
Learning

- Supervised learning – classification, regression
- Unsupervised learning – clustering, outlier detection
- Semi-supervised

Supervised learning: Classification

- Data has class labels
- Based on the labels, classifiers are generated
- New data will be classified based on the generated classifier
- Predicts a **discrete** class label
- Example 1: Cancer dataset – Malignant and benign labels are present for each instance.
- Example 2: Iris dataset – data from 3 types of flowers – every instance has a class label





https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Supervised learning: Regression

Regression predicts **continuous** values (numbers) as the output.

Example, housing prices for various houses: 2 bedroom, 3 bedroom, garage size, property size, and the computer must interpolate predictions.

Unsupervised learning

- data has no class labels

- The algorithm tries to identify the objects as being part of some group using a clustering algorithm. Similar instances grouped together to form clusters. (Ex. Insurance: Identifying groups of motor insurance policy holders with a high average claim cost)

- Anomaly detection tries to find those instances which are distinct from the nature of the majority of instances. (Ex. Financial fraud detection)

Semi-supervised learning

- Typically a small amount of labeled data with a large amount of unlabeled data

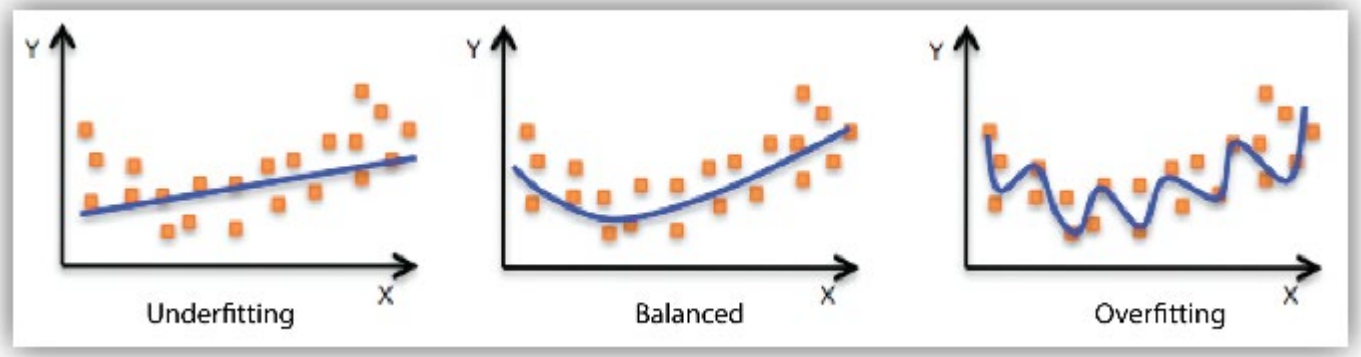
Training & Test set

- To perform learning, you need data.
- Learning will generate a classifier that can perform classification
- In order to test your classifier, you need data which is not used in learning process
- To test the effectiveness of your algorithm, you can split your data into two parts: a training set and a test set.
- The test set should be independent of the training set. It is required to verify the error rate of your algorithm.

Model Overfitting

- Be careful not to over fit. Overfitting is when you are trying to achieve 100% accuracy, even learning from the examples that are wrong. Instead, you want to generalize the data to find the underlying trends.
- Your model is overfitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.

Underfitting vs Overfitting



Your model is *underfitting* the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input and the target values (often called Y).

Refer to: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Error Estimation

- Random sampling with repeated holdout. Run once with a random 2/3 training data, 1/3 test data. Then re-run it with a different 1/3 of the data. Continue this process until your error rate stabilizes.
- K-fold cross-validation – partition into K equal groups. K-1 groups are training data, and test on the last remaining group. Repeat this K times, where K is usually 10. Take the average accuracy rate as the overall accuracy.

Accuracy

- A confusion matrix is defined as the possible outcomes:

	Predicted +	Predicted -
Actually +	a	b
Actually -	c	d

Terms

- The ***accuracy*** of your model is the cases you got right: (a+d) / (a+b+c+d)
- The ***precision*** is defined as: a / (a+c)
- The ***recall***, and ***Sensitivity***, both mean: a / (a+b). These are the number of true cases you got right.
- The ***specificity*** is d / (b+d). These are the number of false cases you got right.

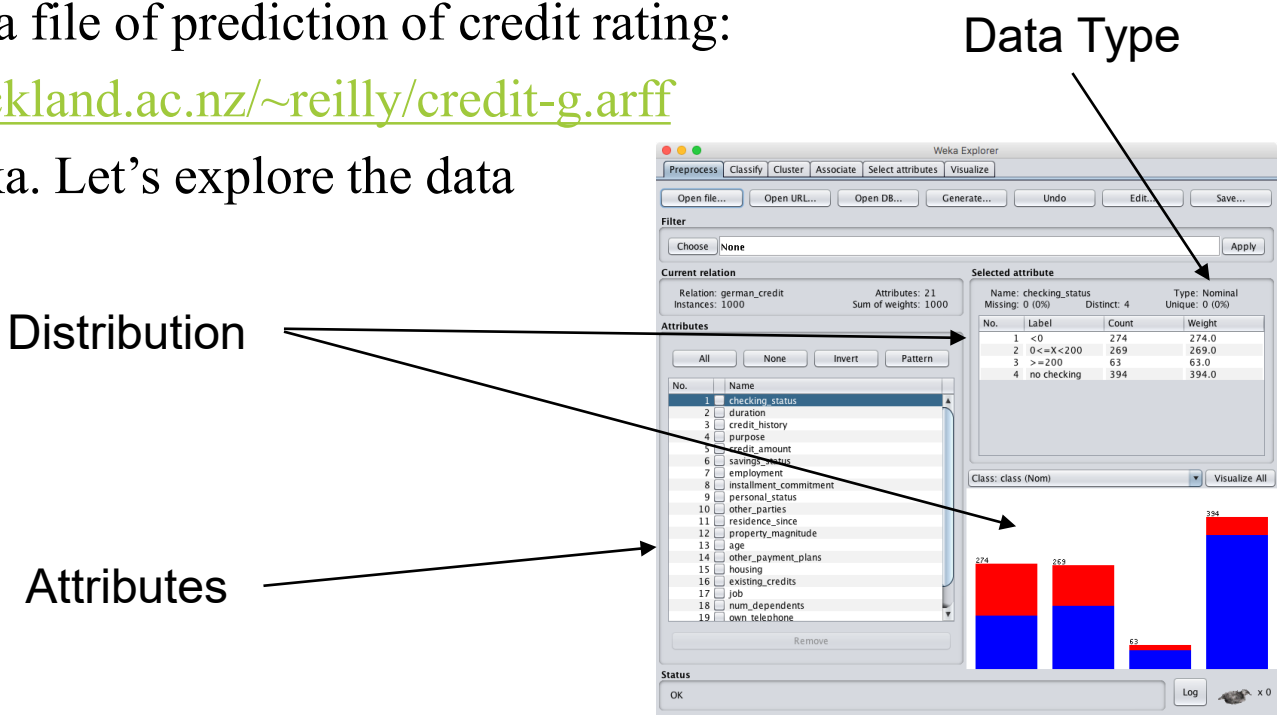
K-Nearest neighbors

- One of the easiest classification algorithms
- Create a plot of the data, and compute which are the K nearest items for your unknown sample.
- From the K-nearest, calculate a simple majority wins estimate for the value you want to predict.
- For predicting final grades, find students with similar final numeric grades, and pick the most popular letter grade.
- For predicting weather, look at previous data for date, temperature, etc. and pick the most popular classification.

Demo in Excel

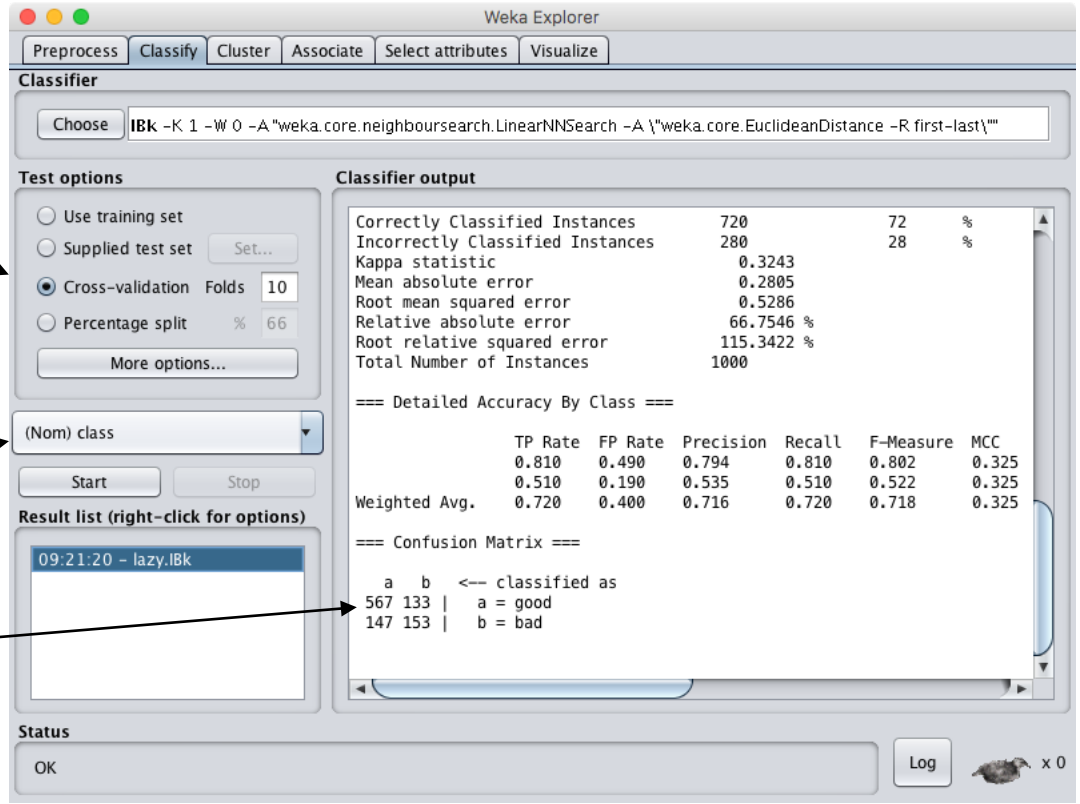
Testing in Weka

- Have a look at a data file of prediction of credit rating:
- https://www.stat.auckland.ac.nz/~reilly/credit-g.arff
- Load the file in Weka. Let's explore the data



Testing

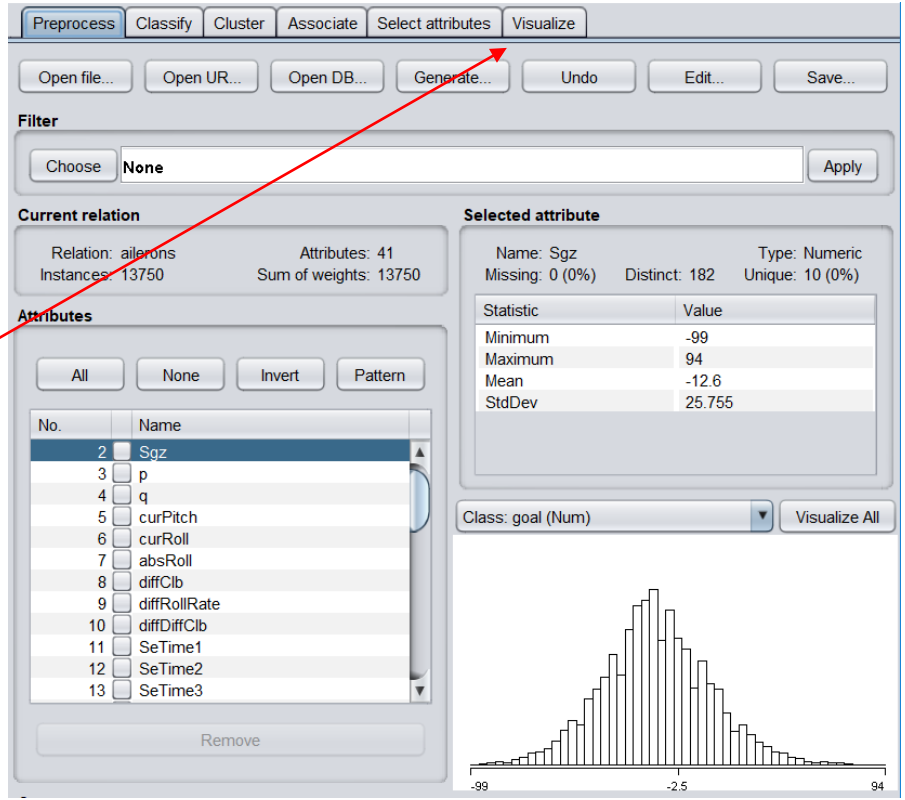
- Validation method
- What are you predicting?
- Confusion matrix



Data Visualization

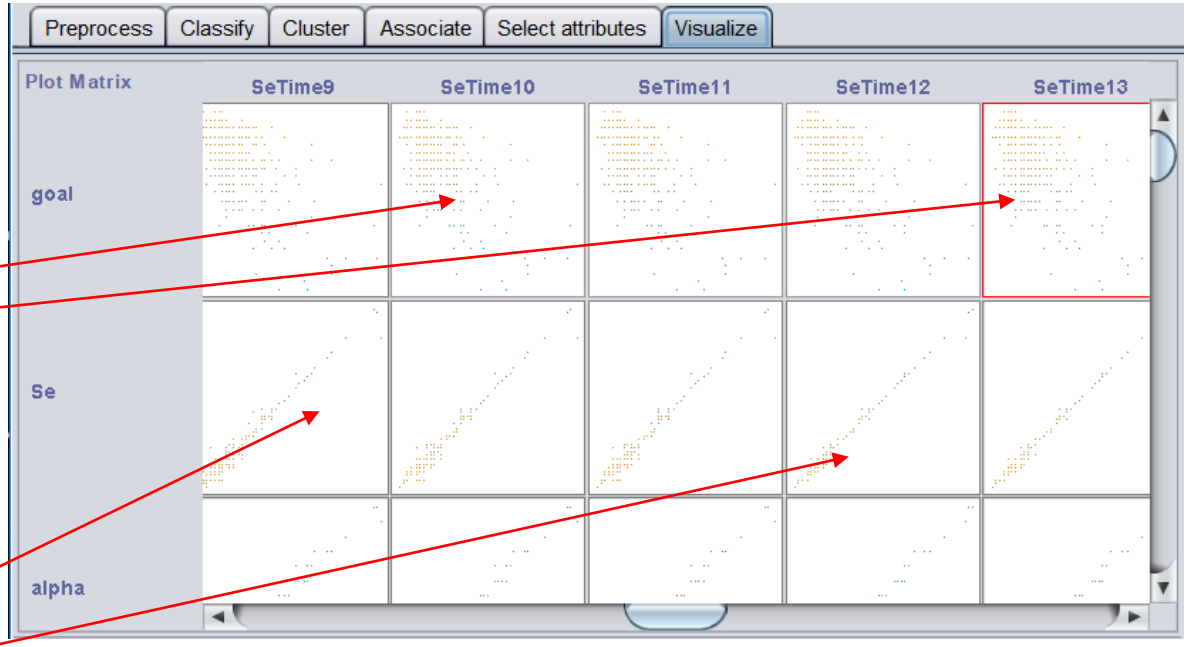
- Looking at regression/ailerons.arff. If the distribution is numeric, it shows the histogram:

- Next click on the “Visualize” tab



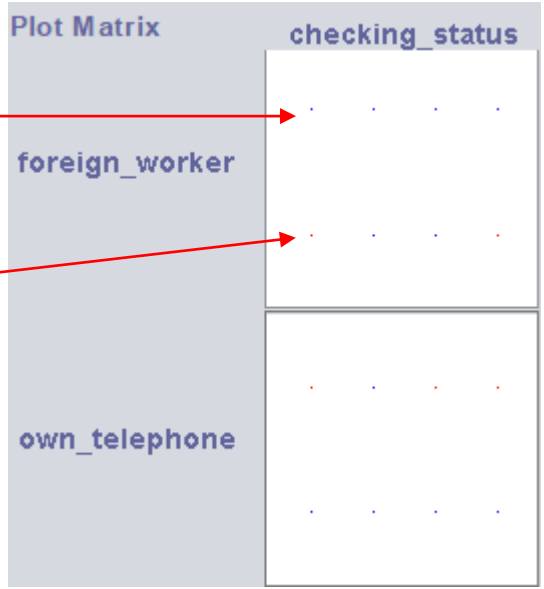
Data Visualization

- You can see the relationship of every variable with each other.
- If there is no relationship, you see a cloud.(No Correlation)
- If there is a relationship, you see a linear pattern. (Correlation)



Data Visualization

- If the data are Categorical, you see clear separation. Here we see yes/no values.
- This is the credit_g.arff file
- Increase the jitter to add some noise. This will give you an idea on how many points are represented by each point.



References

K-Nearest Neighbour:

- http://sens.tistory.com/277
- http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- https://www.youtube.com/watch?v=SQOdBjjA2y8
- https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/

Crisp-DM:

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- http://www.sv-europe.com/crisp-dm-methodology/

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

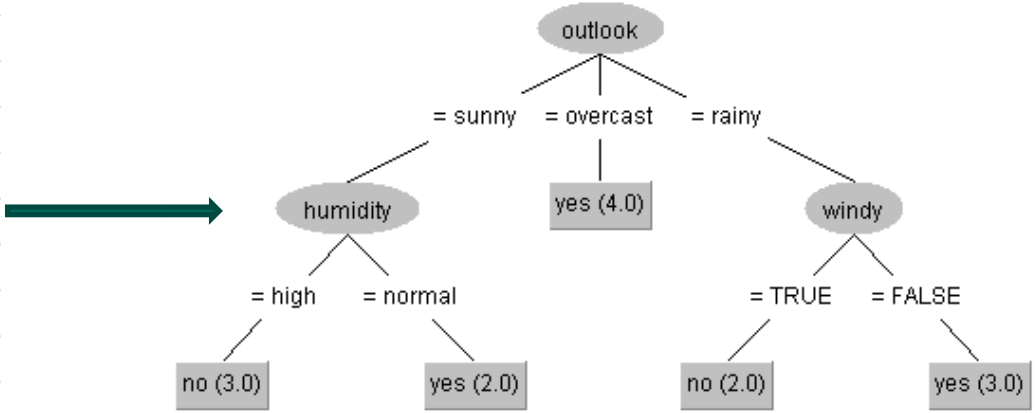
Week 3

Classification – Decision Trees

Professor : Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Decision Trees

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no



Decision Trees

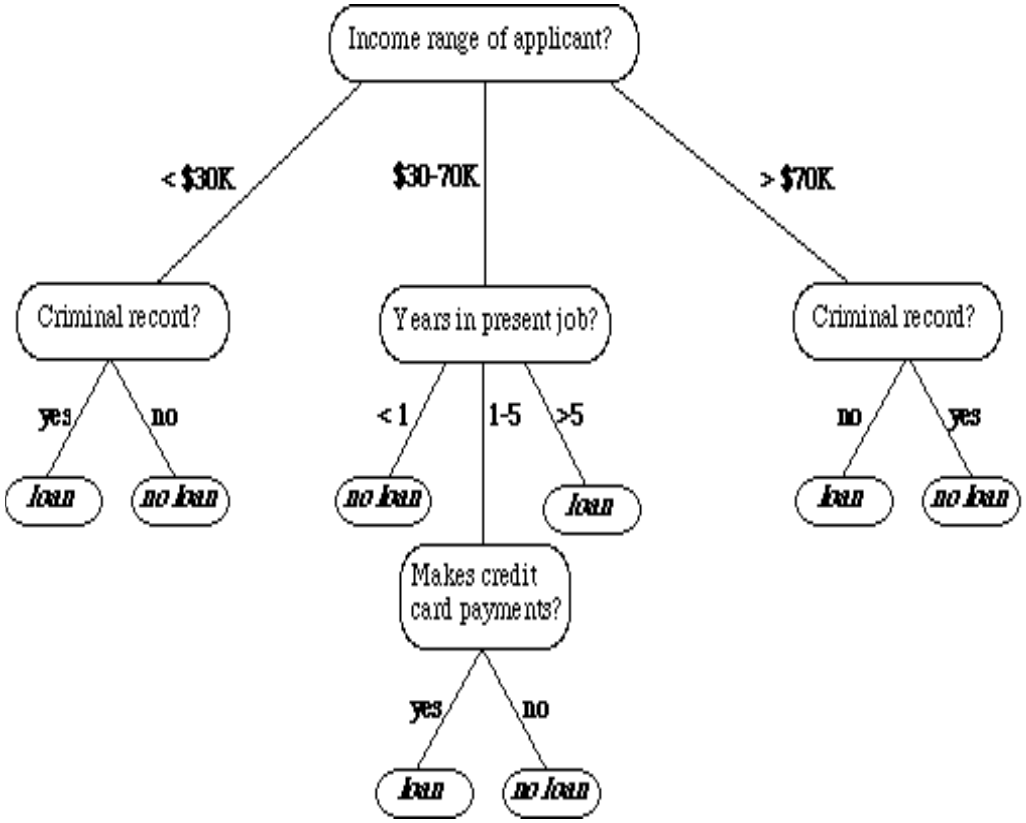
- How to construct decision trees?
- How to avoid overfitting?

Decision Trees

- Decision tree is a tree where:
  - each node represents a feature (attribute)
  - each branch represents a decision (rule)
  - each leaf represents an outcome (categorical or continuous values)

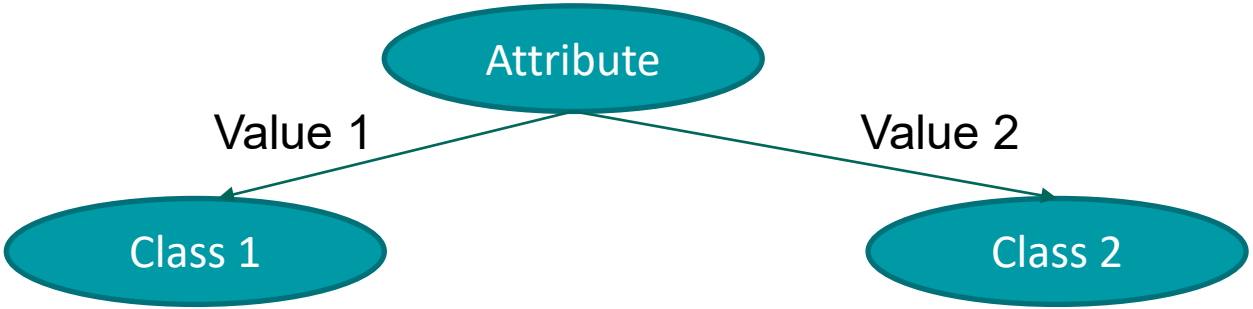
Decision Tree

- One of the most popular ML algorithms
- Used for both classification and regression
- You use data that you know is correct and the tree is created to repeat the decisions in the data.



Decision Tree Construction

- It is a method for approximating discrete-valued functions (Labels).
- It is a divide and conquer approach. Each leaf of the tree is the classification group. Each internal node tests an attribute, and a branch is the value.



How to build

- Start with the root of the tree. Pick an attribute to divide the instances into different groups. Then for each group, repeat the process until the groups are all the same.
- We want the smallest tree so that there are less things to compare. We will use the impurity criterion, or information gain (highest entropy)
- If an event is highly predictable, then it has low entropy or low uncertainty
- Random probabilities have higher entropy, or higher uncertainty.

$$\text{entropy}(p_1, p_2, p_3, \ldots, p_n) = -p_1\log(p_1) - p_2\log(p_2) - p_3\log(p_3) .. -p_n\log(p_n)$$

Decision Tree Algorithms

- ID3 (Iterative Dichotomiser 3)
  - Uses Entropy function and Information gain as metrics
- CART (Classification and Regression Trees)
  - Uses Gini Index as metric

Classification using ID3 Algorithm

Weather Dataset

Based on weather conditions, predict Y or N for “Play”.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Entropy

- Measure of the amount of impurity or uncertainty in the dataset

$$H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

Where S – current dataset for which entropy is being calculated

C – set of classes in S Example: $C = \{yes, no\}$

$p(c)$ – The proportion of the number of elements in class c to the number of elements in S

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S on the current iteration.

Information Gain

- Measure of the difference in entropy from before to after the set S is split on an attribute A.
- Measure on how much uncertainty in S was reduced after splitting S on attribute A

Information Gain

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where $H(S)$ - *Entropy of set S*

T – Subset created by splitting S by attribute A

$p(t)$ – The proportion of the number of elements in t to the number of elements in S

$H(t)$ – Entropy of subset t

Metrics for Weather dataset

Steps

1. Compute the entropy for the dataset
2. For every attribute:
    i. Calculate entropy for all categorical values
    ii. Take average for the current attribute
    iii. Calculate gain for the current attribute
3. Pick the attribute with highest gain
4. Repeat until we get the tree we desired

Entropy for Weather dataset

$$H(S) = \sum_{c \in C} -p(c) \log_2 p(c)$$

Out of 14 instances, 9 are classified as Yes and 5 as No

$$P_{Yes} = -\frac{9}{14} * \log_2 \frac{9}{14} = 0.41$$

$$P_{No} = -\frac{5}{14} * \log_2 \frac{5}{14} = 0.53$$

$$H(S) = P_{Yes} + P_{No} = 0.94$$

Entropy of Outlook feature of Weather dataset

- $H(Outlook = Sunny) = -\frac{2}{5} * \log_2\frac{2}{5} - \frac{3}{5} * \log_2\frac{3}{5} = 0.5288 + 0.4422 = 0.971$
- $H(Outlook = Overcast) = -\frac{4}{4} * \log_2\frac{4}{4} - \frac{0}{4} * \log_2\frac{0}{4} = 0$
- $H(Outlook = Rainy) = -\frac{3}{5} * \log_2\frac{3}{5} - \frac{2}{5} * \log_2\frac{2}{5} = 0.4422 + 0.5288 = 0.971$
- $Average\ Entropy\ for\ Outlook$
- $M(Outlook) = \frac{5}{14} * 0.971 + \frac{4}{14} * 0 + \frac{5}{14} * 0.971 = 0.6936$

- $\text{Gain}(Outlook) = H(S) - M(Outlook) = 0.94 - 0.6936 = 0.2464$

Entropy of Windy feature of Weather dataset

- $H(Windy = False) = -\frac{6}{8} * \log_2\frac{6}{8} - \frac{2}{8} * \log_2\frac{2}{8} = 0.3113 + 0.5 = 0.8113$
- $H(Windy = True) = -\frac{3}{6} * \log_2\frac{3}{6} - \frac{3}{6} * \log_2\frac{3}{6} = 0.5+0.5=1$

- $Average\ Entropy\ for\ Windy$
- $M(Windy) = \frac{8}{14} * 0.8113 + \frac{6}{14} * 1 = 0.4636 + 0.4286 = 0.8922$

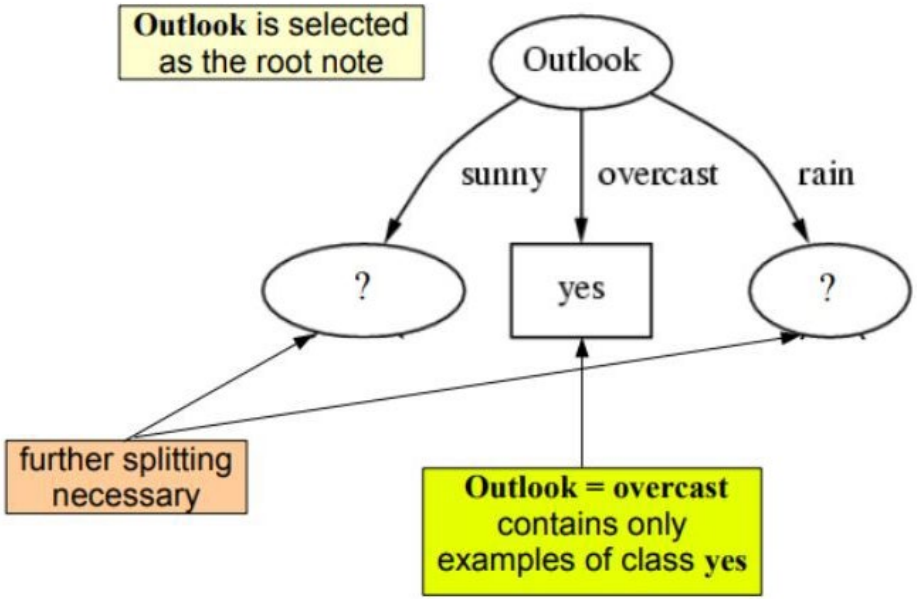
- $\text{Gain}(Windy) = H(S) - M(Windy) = 0.94 - 0.8922 = 0.0478$

Metrics Summary

Outlook		**Temperature**	
Average Entropy:	0.693	Average Entropy:	0.911
Information Gain:	**0.247**	Information Gain:	0.029
Humidity		Windy	
Average Entropy:	0.788	Average Entropy:	0.892
Information Gain:	0.152	Information Gain:	0.048

As Outlook has the highest Information Gain, our root node is **Outlook**

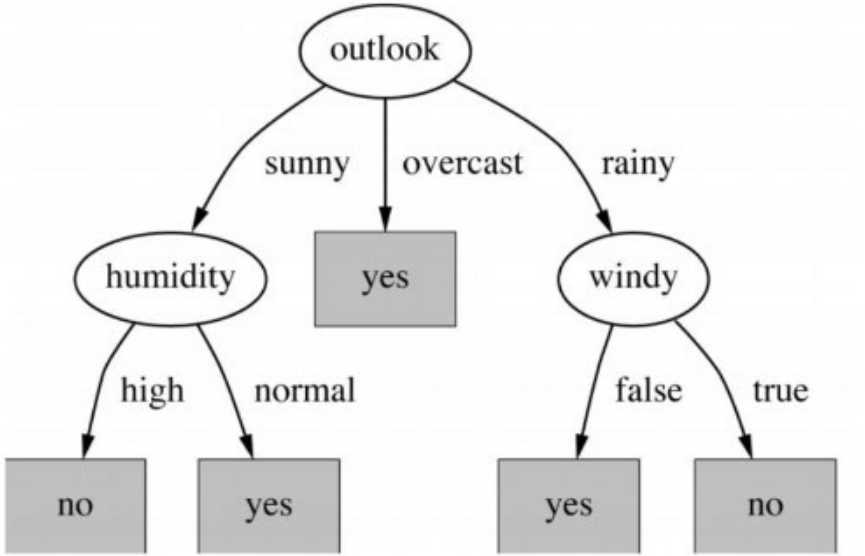
Initial Tree for Weather Dataset



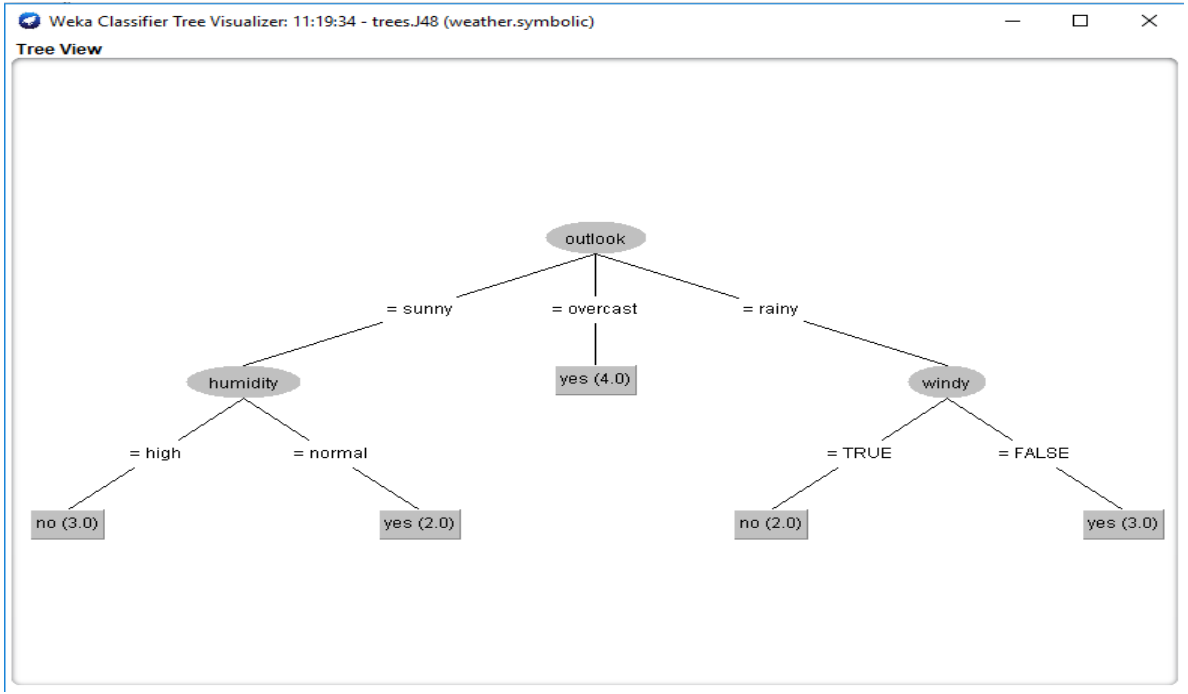
Developing Tree

- Repeat the same step for subtrees

Final decision tree



Weka Demo


Weka Classifier Tree Visualizer: 11:19:34 - trees.J48 (weather.symbolic)
Tree View
outlook
= sunny
= overcast
= rainy
humidity
yes (4.0)
windy
= high
= normal
= TRUE
= FALSE
no (3.0)
yes (2.0)
no (2.0)
yes (3.0)

References

Decision Tree:

- http://www.saedsayad.com/decision_tree.htm
- http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class3-DataMiningWithWeka-2013.pdf

Covariance and correlation:

- http://www.dummies.com/education/math/business-statistics/how-to-measure-the-covariance-and-correlation-of-data-samples/

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 4

Clustering

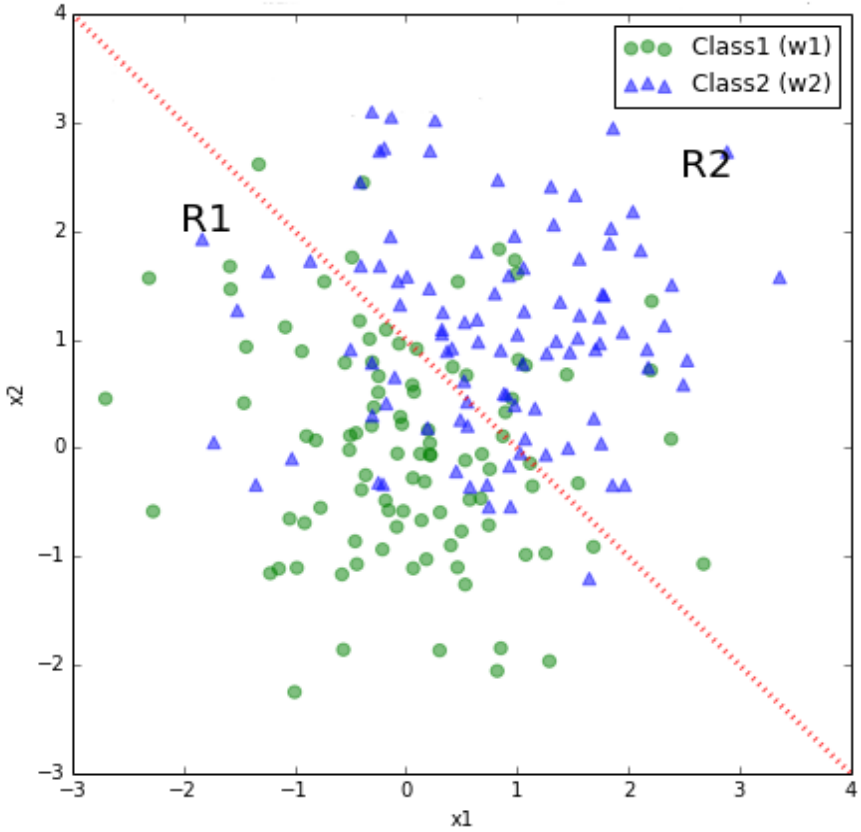
Professor: Dr. Anu Thomas

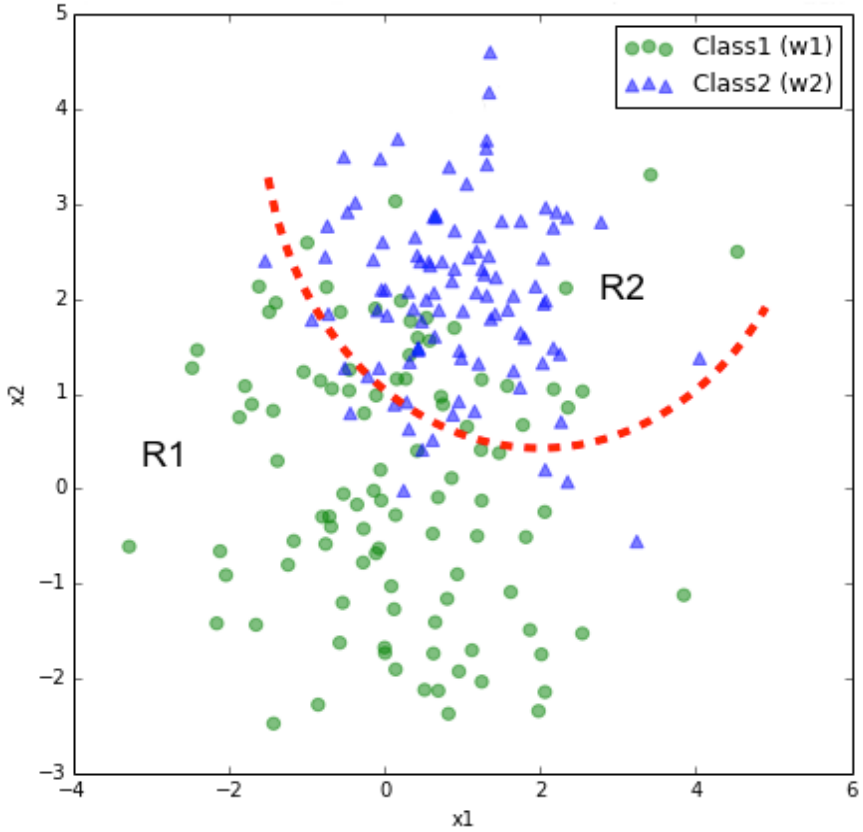
Learning

- Supervised learning – classification, regression
- Unsupervised learning – clustering, outlier detection
- Semi-supervised

Supervised learning: Classification

- Data has class labels
- Based on the labels, classifiers are generated
- New data will be classified based on the generated classifier
- Predicts a **discrete** class label
- Example 1: Cancer dataset – Malignant and benign labels are present for each instance.
- Example 2: Iris dataset – data from 3 types of flowers – every instance has a class label





https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html

Unsupervised learning

- data has no class labels

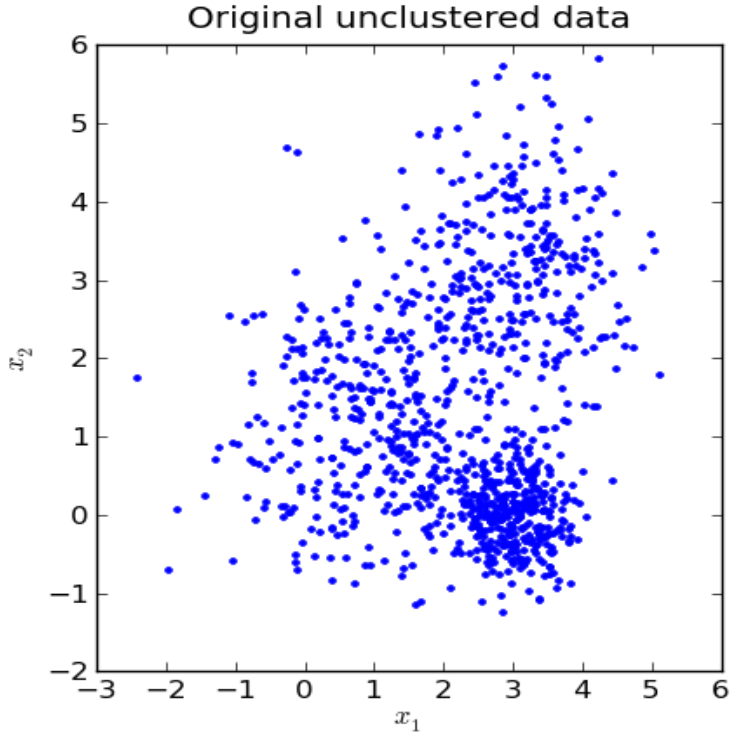
- The algorithm tries to identify the objects as being part of some group using a clustering algorithm. Similar instances grouped together to form clusters. (Ex. Insurance: Identifying groups of motor insurance policy holders with a high average claim cost)

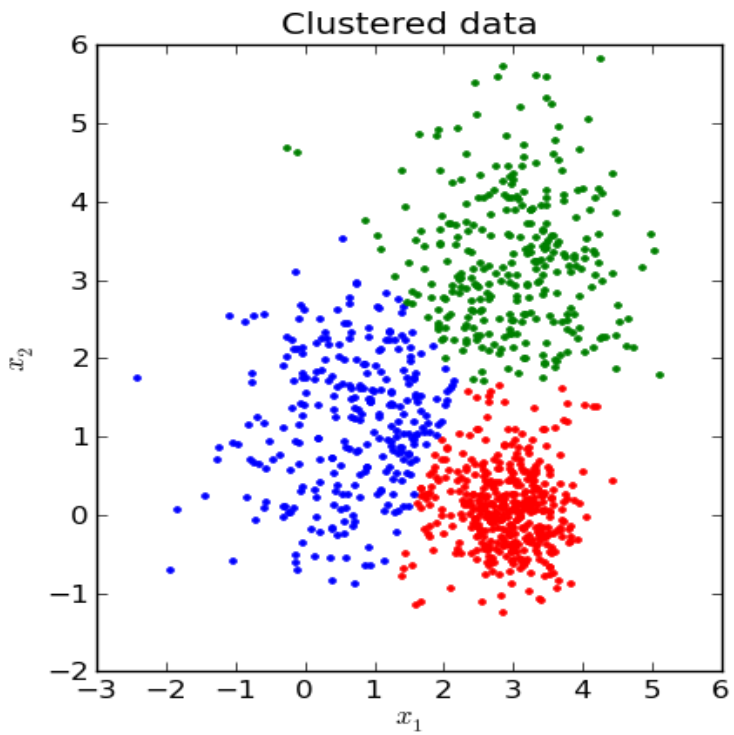
- Anomaly detection tries to find those instances which are distinct from the nature of the majority of instances. (Ex. Financial fraud detection)

Clustering

- Cluster – a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering is an unsupervised classification method

Clustering - example





Clustering Algorithms

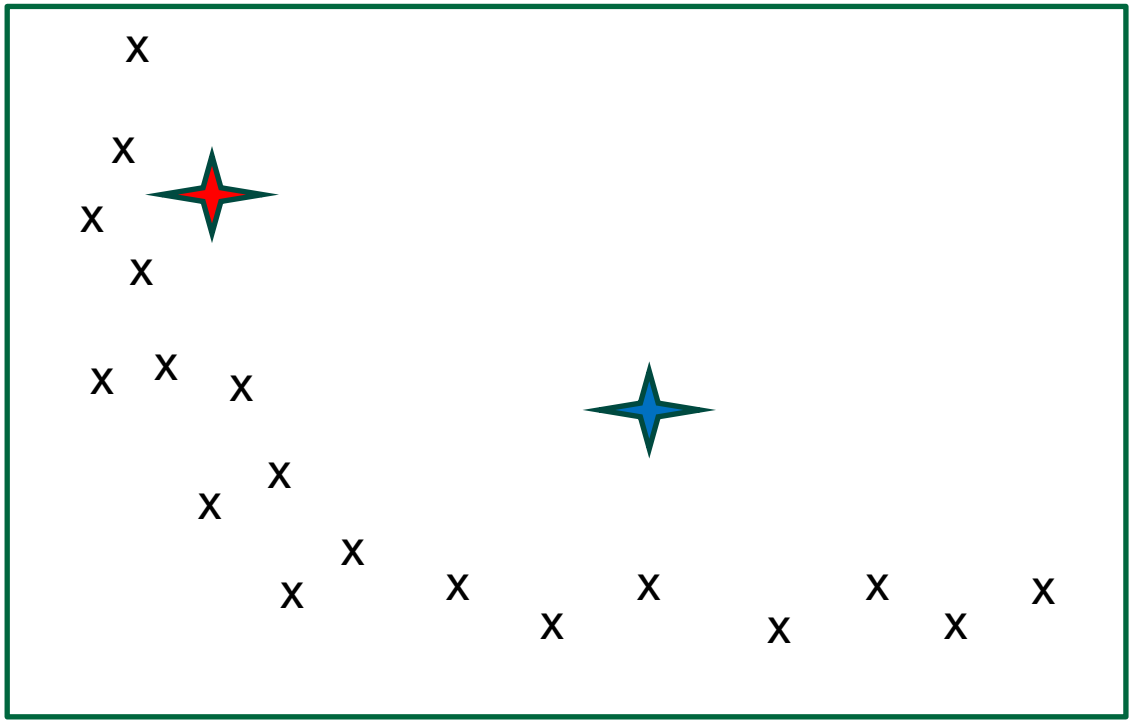
- K-Means
- Mean-shift
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Expectation-Maximization

K-Means Clustering Algorithm

- K-Means is an unsupervised learning algorithm. It uses unlabeled numeric data. It automatically groups data elements into different groups.
- The parameter K refers to how many groups for the data.
- The data must be numeric because it calculates distance, using square root. The square root of labels, like Hot and Cold, doesn't make sense.

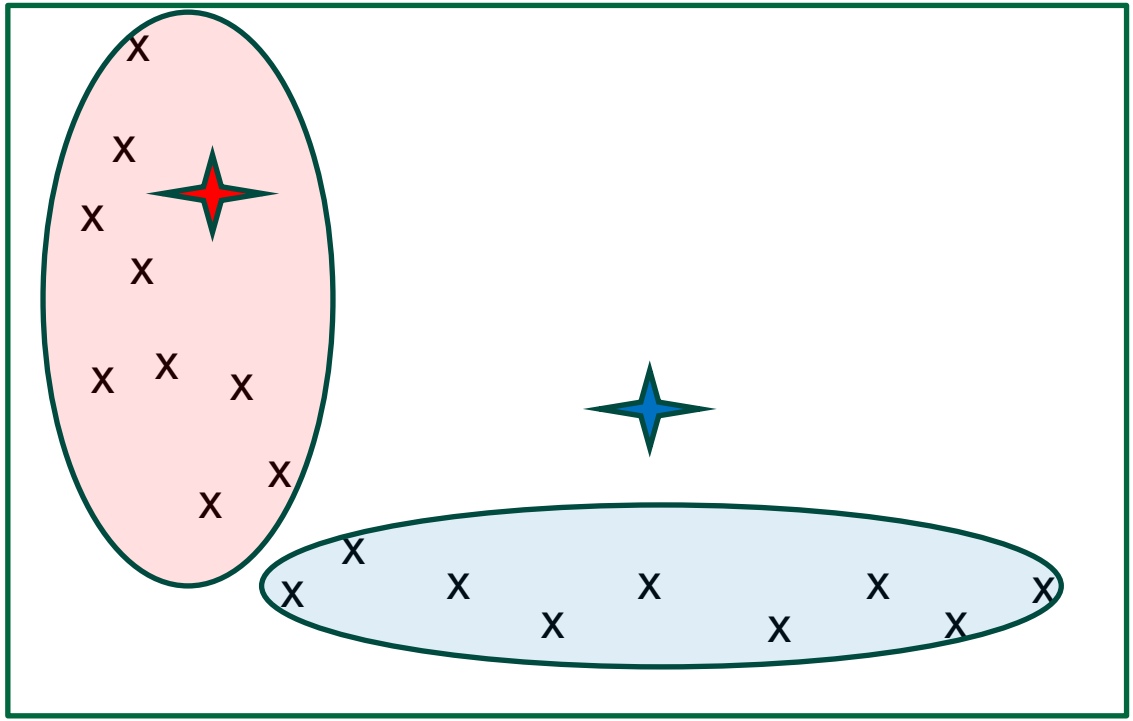
K-Means Clustering Algorithm

- Given a data set:
- Decide how many groups you want to calculate.
- You must know K in advance.
- Give each group a starting point (Centroid).

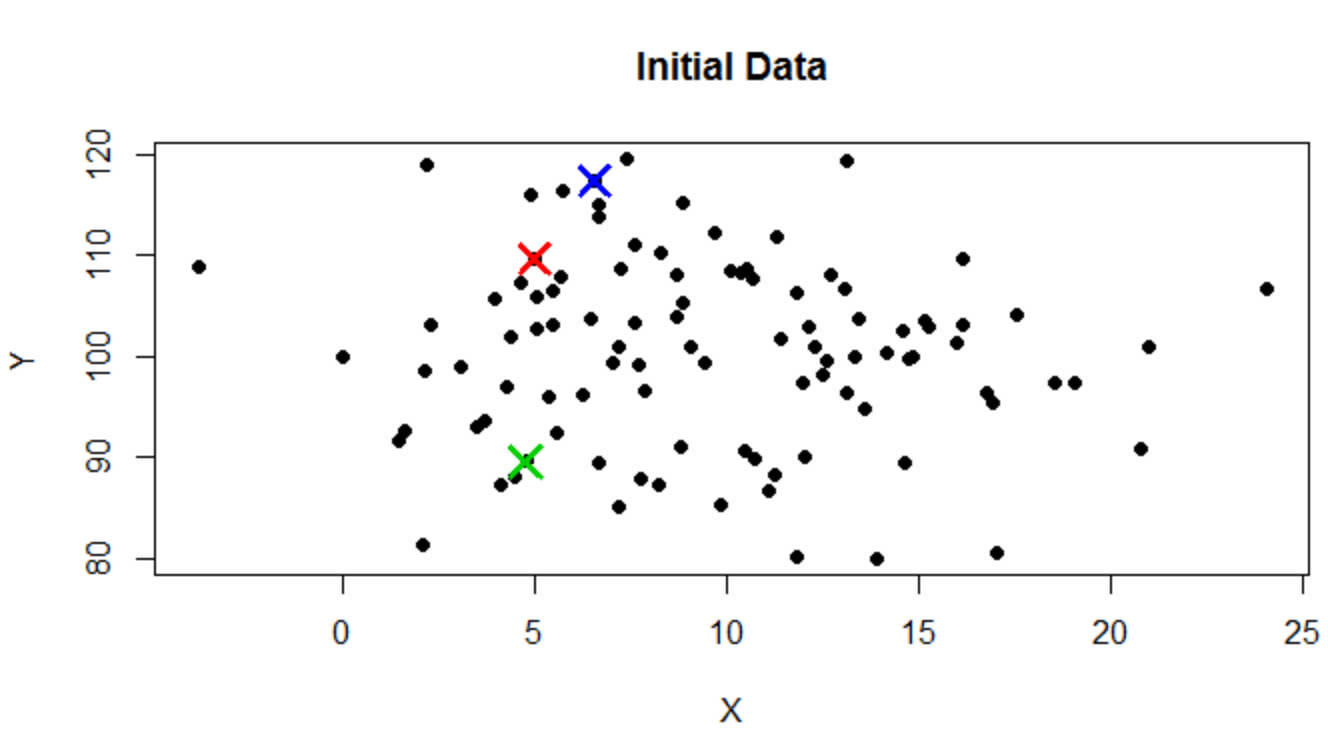


Repeat:

- For each data point, which is the nearest Centroid?
- Cluster items with the same Centroids.
- Re-compute Centroid location to middle of cluster until they stop moving.

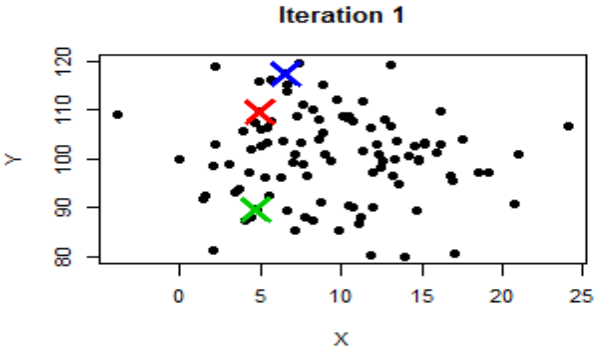


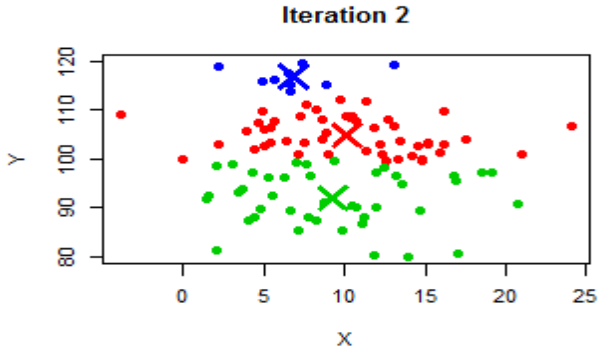
Example

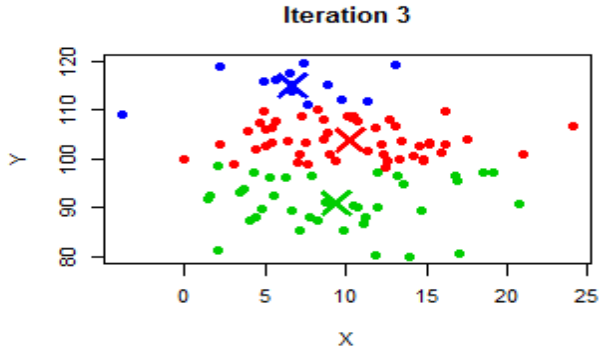


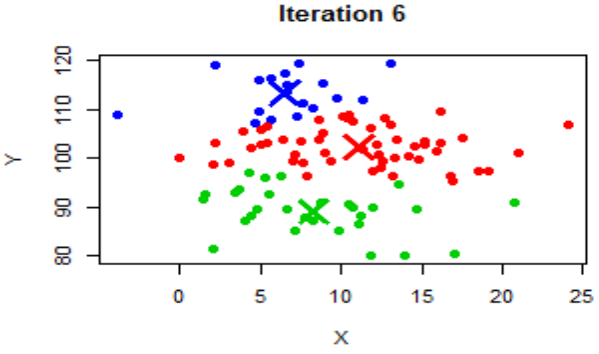
Taken from: http://www.learnbymarketing.com/methods/k-means-clustering/

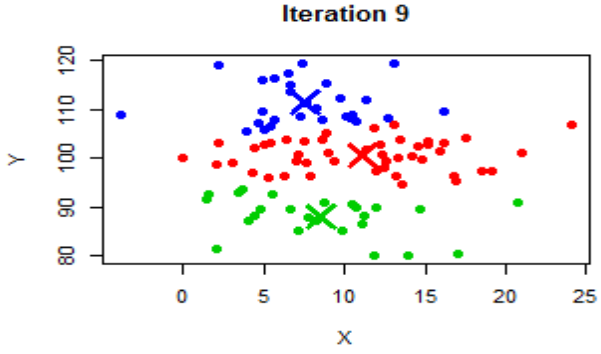
K-means clustering

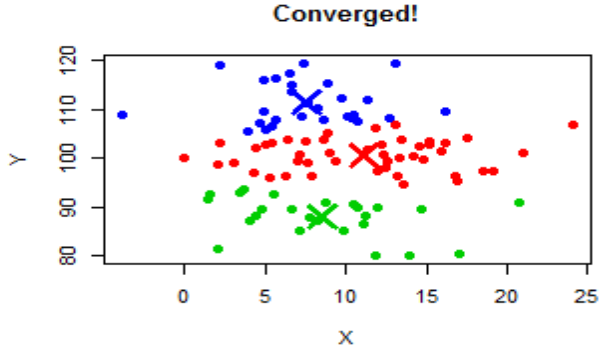








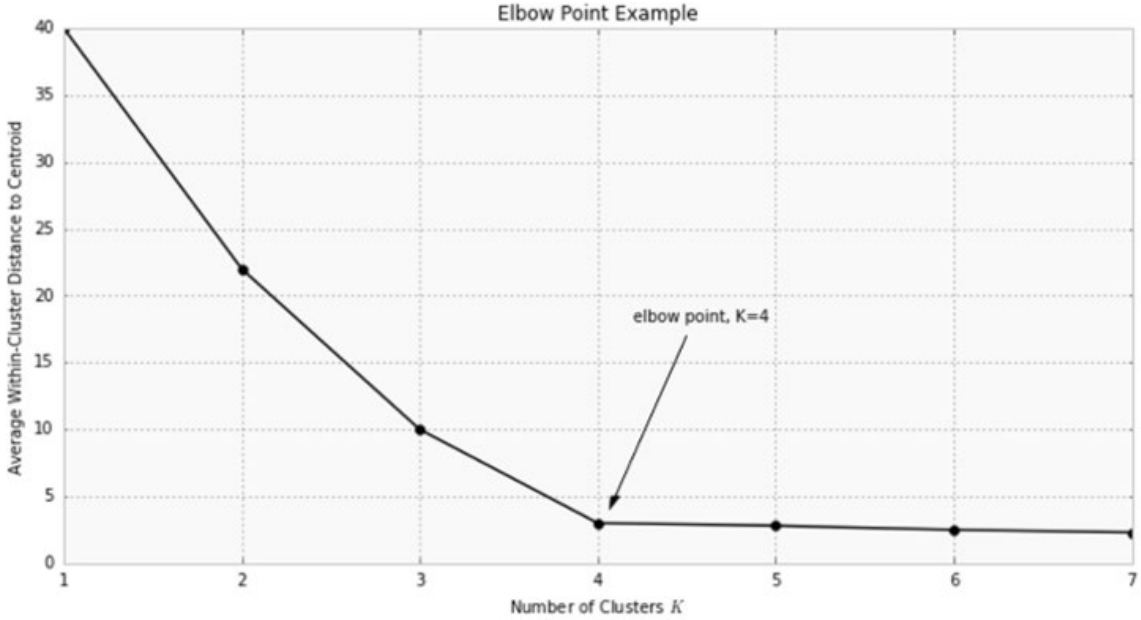




How do you choose K?

- Run the algorithm with 2 centroids. Then calculate the average distance from each point to its nearest centroid.
- Repeat the steps with 3, 4, 5, … n centroids. If you plot the average within-cluster distance to the nearest centroid, you will see an “elbow point”. That value should be the value of K, the number of groups in your data.
- The centroids are the average value for each cluster.

How do you choose K?



- https://www.datascience.com/blog/k-means-clustering

Demo – Calculations in Excel

Weka Demo

References

- https://www.datascience.com/blog/k-means-clustering
- Play with the examples:
- http://www.onmyphd.com/?p=k-means.clustering
- The 5 Clustering Algorithms Data Scientists Need to Know

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 5

Outlier Detection

Assignment 2

Final Project

- Project selections must be submitted by Sunday June 20 at 11:59 PM.

Decision Trees - Recap

- Type of attributes
- Parameters
  - Minimum number of objects
  - Pruning
- Numbers at leaf: The first number is the total number of instances reaching the leaf. The second number is the number of misclassified instances

Introduction

- What is outlier detection?
- Outlier Detection using Statistical Methods
- Outlier Detection using Machine Learning techniques

What is an outlier?

- A data object that deviates significantly from the Majority of normal objects
- Ex.: unusual credit card purchase

Applications of Outlier Detection

- Financial fraud detection (banking, credit card etc.)
- Telecom fraud detection
- Medical Diagnosis
- Web Analytics

Types of Outliers

- Three types:
  - Global Outlier (point anomalies)
  - Contextual outlier (conditional outlier)
  - Collective Outliers

Global Outlier

- A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found

100144	Yulma	Peyntue	ypeyntue26@mayoclinic.com	'3257 American Crossing'	China	3	CHY	150000
100145	Reade	McCumesky	rmccumesky2y@list-manage.com	'6766 Schmedeman Road'	China	3	CHY	150000
100146	Maximilian	Camies	mcamiesv@so-net.ne.jp	'7201 Cambridge Park'	U.S.A.	1	USD	4000
100147	Sloane	Andrzejak	sandrzejak3t@netlog.com	'44 Troy Crossing'	Mexico	4	MXD	40500
100148	Carlye	Blunsen	cblunsen1o@admin.ch	'8131 Stephen Park'	Germany	2	EUR	59500
100149	Darcy	Addie	daddie1k@jalbum.net	'836 Marquette Pass'	Germany	2	EUR	60500999
100150	Cissy	Duley	cduley38@fotki.com	'198 Westerfield Way'	Mexico	4	MXD	18000
100151	Ingmar	Durward	idurwardd@jimdo.com	'38 Badeau Road'	U.S.A.	1	USD	30000
100152	Brittan	Timson	btimson32@yellowbook.com	'9 Crownhardt Way'	China	3	CHY	150000
100153	Malvin	Houdmont	mhoudmont2k@google.it	'654 7th Drive'	China	3	CHY	190000

Contextual Outlier

- If the value deviates significantly based on a selected context
- Ex: a temp of -30.7 degree Celsius during the month of June in Ottawa.

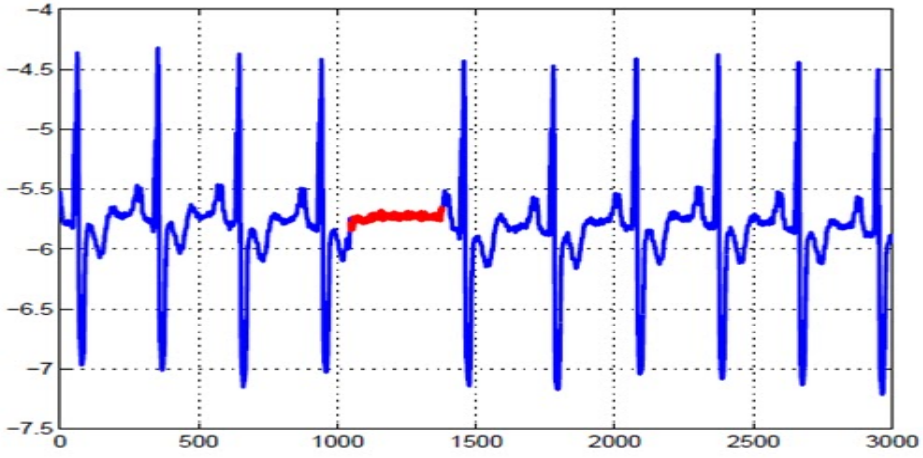
Year	Month	Day	Max Temp (°C)	Min Temp (°C)	Mean Temp (°C)
2018	6	12	27.7	8.8	18.3
2018	6	13	20.7	13.6	17.2
2018	6	14	17.6	1137	577.3
2018	6	15	25.4	8.2	16.8
2018	6	16	28.1	10.4	19.3
2018	6	17	-30.7	13.6	22.2
2018	6	18	30.4	16.6	23.5
2018	6	19	24.5	11.5	18
2018	6	20	28.8	9.7	19.3
2018	6	21	20.9	9.2	15.1

Contextual Outlier - Example

Id	first_name	last_name	email	Address	Country	Branch	Currency	Salary
100230	Nissie	Burney	nburneyr@paginegialle.it	34 Dovetail Point	U.S.A.	1	USD	26500
100231	Darby	Mandell	dmandell1z@ovh.net	922 Sachs Avenue	Germany	2	EUR	38000
100232	Fonzie	Rasell	frasell44@eepurl.com	991 Scoville Trail	Mexico	4	MXD	46888
100233	Bel	Hodgin	bhodgin2g@msu.edu	60 Bellgrove Court	**Japan**	3	CHY	600000
100234	Sylvia	Holborn	sholborn13@paypal.com	83094 Packers Alley	Germany	2	EUR	69000
100235	Dur	Atlee	datlee3k@hugedomains.com	39084 Thackeray Center	Mexico	4	MXD	46000
100236	Cesaro	Kinnock	ckinnock18@liveinternet.ru	518 Center Way	Germany	2	EUR	50000
100237	Clarette	Headford	cheadford23@flickr.com	674 International Plaza	Germany	2	EUR	70000
100238	Wittie	Guarin	wguarint@vkontakte.ru	3 Graceland Hill	U.S.A.	1	USD	39200
100239	Lavinia	Thorneloe	lthorneloe1f@ameblo.jp	09 Huxley Pass	Germany	2	EUR	95000
100240	Katina	Borel	kborelo@github.io	629 Hansons Terrace	U.S.A.	1	USD	68000
100241	Stuart	Dello	sdeldello3u@msu.edu	8669 Warner Park	Mexico	4	MXD	31000
100242	Rosalia	Boseley	rboseleyi@sfgate.com	97917 Brentwood Alley	U.S.A.	1	USD	60000
100243	Feodor	Tine	ftine1e@flickr.com	04 Moland Point	Germany	2	EUR	32000
100244	Olivie	Knightly	oknightly34@godaddy.com	04375 Bunting Pass	China	3	CHY	150000
100245	Saundra	Morphey	smorphey43@diigo.com	63 Red Cloud Parkway	Mexico	4	MXD	28000
100246	Nettle	Gleadhall	ngleadhall3@umn.edu	511 Loftsgordon Plaza	U.S.A.	1	USD	29000
100247	Nelson	McRinn	nmcrinn3p@economist.com	56053 Buell Terrace	Mexico	**2**	MXD	19999
100248	Georgine	Racher	gracherf@webeden.co.uk	68311 Lake View Park	U.S.A.	1	USD	42500
100249	Aurore	Grece	agrece24@technorati.com	093 Stuart Place	China	3	CHY	180000
100250	Briana	Catchpole	bcatchpole2c@over-blog.com	19005 Bluejay Park	China	3	CHY	900000

Collective Outliers

- A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers



The highlighted region denotes an outlier because the same low value exists for an abnormally long time. The low value by itself is not an outlier but its successive occurrence for long time is an outlier.

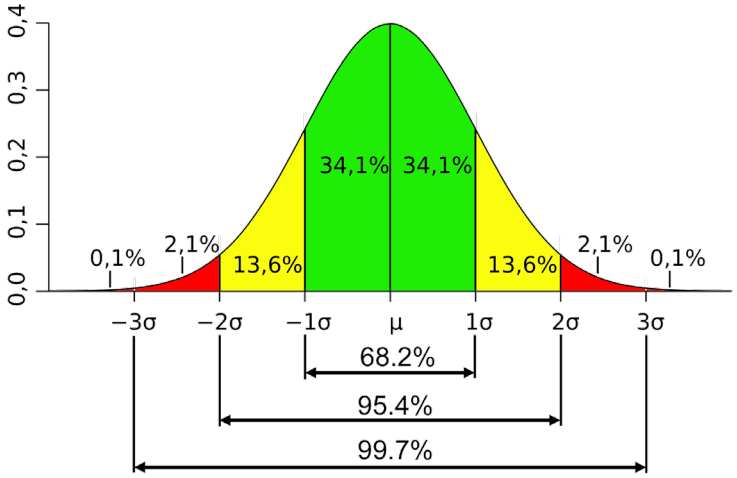
Methods for Outlier Detection

- Statistical Methods
- Proximity-based methods
  - Distance-based
  - Density-based (Ex. Local Outlier Factor - LOF)
- Clustering-based methods

Gaussian distribution

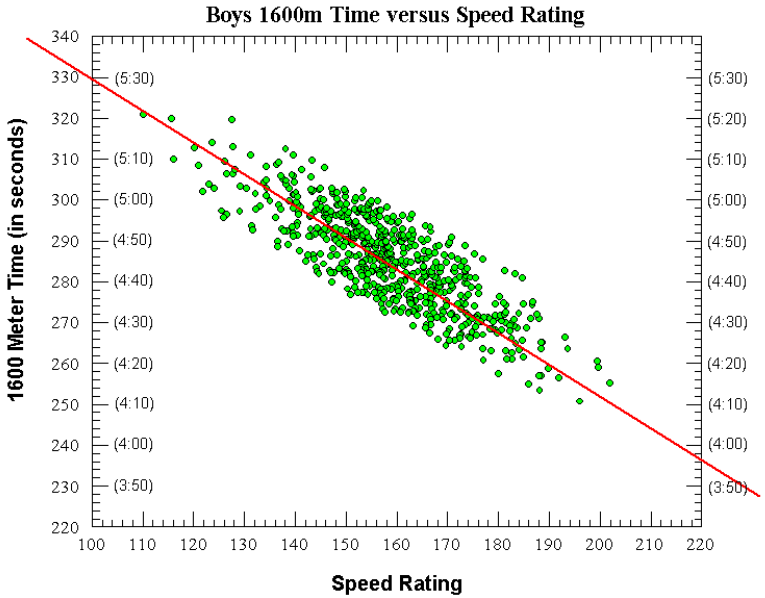
- Use an error margin “ε” to set the limit of what is an outlier. It is a probability at which everything beyond will be categorized as an outlier.

For instance, use 1% as the limit. This means that everything that has a less than 1% chance of happening is an outlier.



Outlier detection

- Calculate the mean and standard deviation. Then calculate the 99% limit. Then use the range as your classification.
- This works for each attribute independently but not when the data have a correlation.

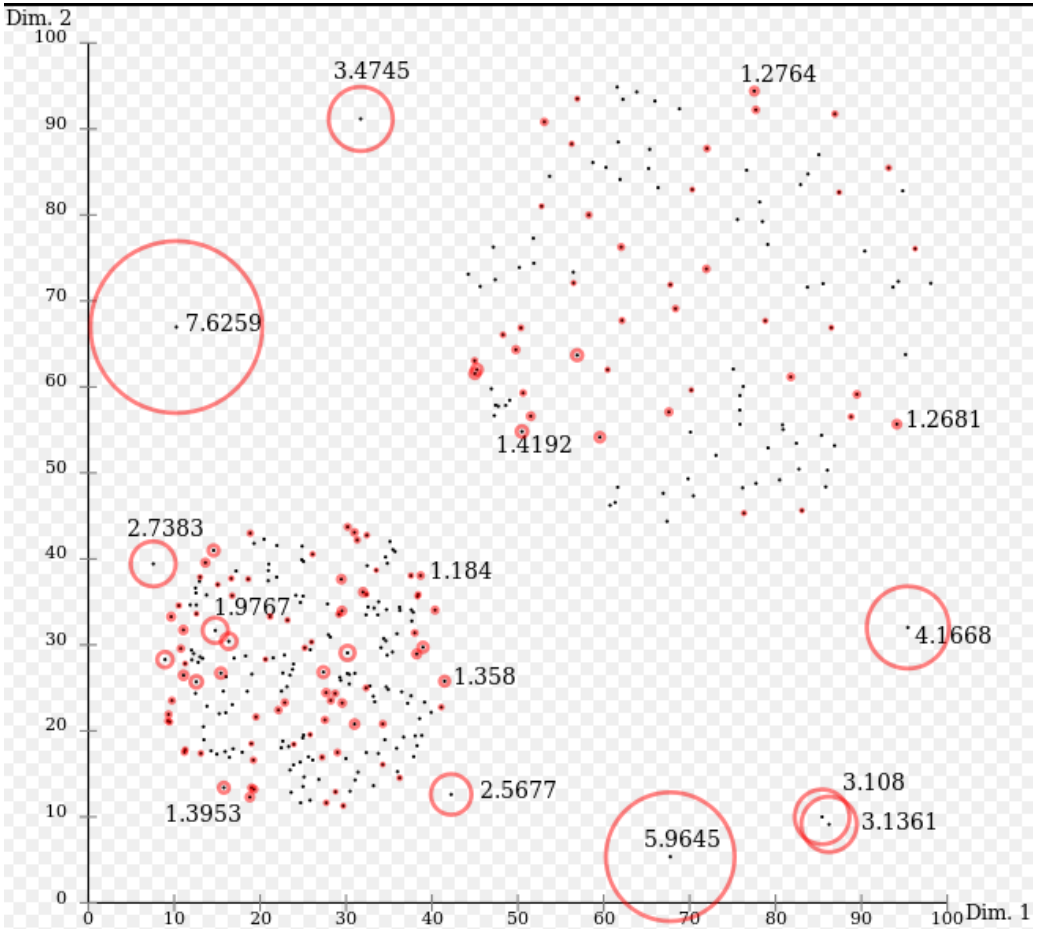


Local Outlier Factor - LOF

- Local outliers: Outliers comparing to their local neighborhood, instead of the global data distribution
- LOF: finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors

LOF

Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbors.



Weka Demo

- Interquartile range
- LOF

Isolation Forest

- explicitly identifies anomalies instead of profiling normal data points
- built on the basis of decision trees
- partitions are created by
  - first randomly selecting a feature and
  - then selecting a random split value between the minimum and maximum value of the selected feature.
- should be identified closer to the root of the tree with fewer splits necessary.

Excel Demo

Weka Demo

References

- http://researchmining.blogspot.com/2012/10/types-of-outliers.html
- http://scikit-learn.org/stable/modules/outlier_detection.html

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 6

Regression

Agenda

- Linear regression
  - Simple linear regression
  - Multiple linear regression
- Multivariate regression
- Logistic regression

Types of Relationships

- Deterministic (or functional) relationship
- Statistical relationship

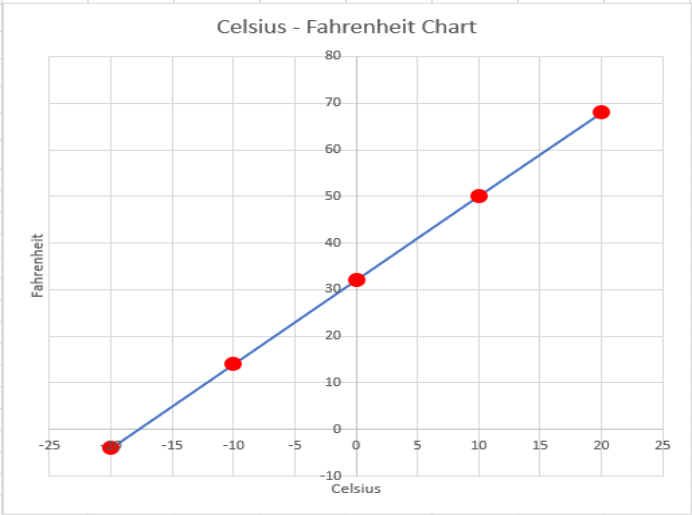
Deterministic (or functional) Relationship

- Ex. Relationship between Celsius and Fahrenheit

➢ $F = \frac{9}{5} * C + 32$

The observed (x, y) data points fall directly on the line.

m	C	b	F
1.8	-20	32	-4
1.8	-10	32	14
1.8	0	32	32
1.8	10	32	50
1.8	20	32	68



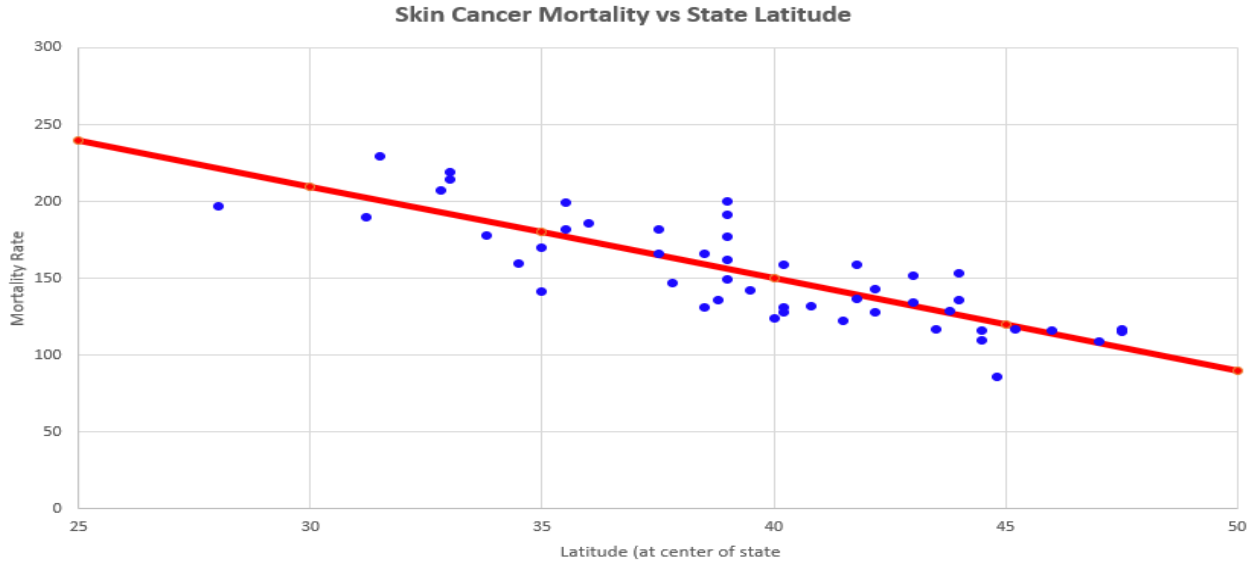
For deterministic relationship, the equation exactly describes the relationship between the two variables.

Statistical Relationship

- Examples
  - Height and weight — as height increases, you'd expect weight to increase, but not perfectly.
  - Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
  - Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Statistical Relationship

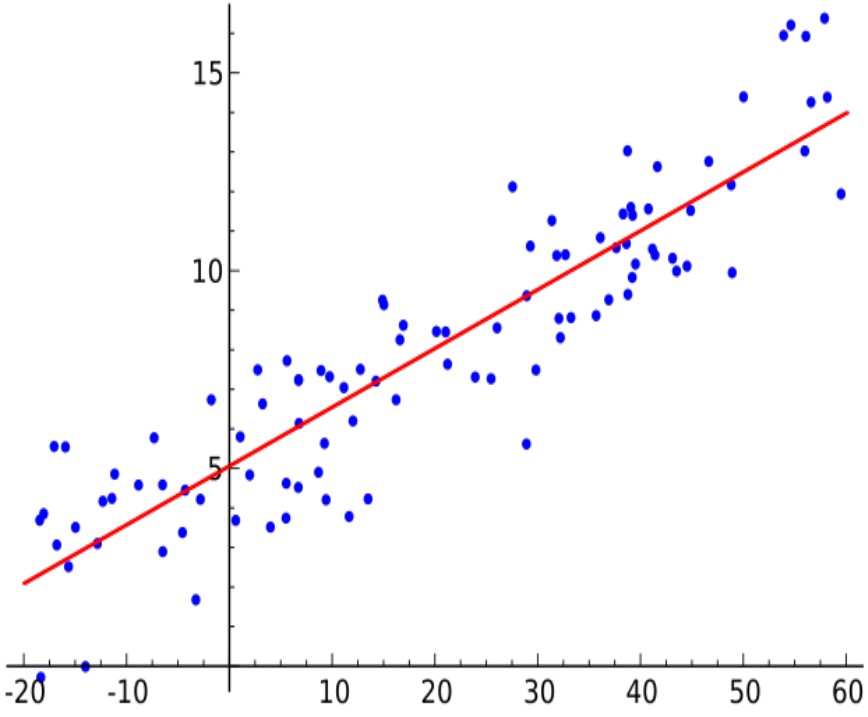
Example: The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of 49 states in the U.S.



State	Lat	Mort	Ocean	Long
Alabama	33	219	1	87
Arizona	34.5	160	0	112
Arkansas	35	170	0	92.5
California	37.5	182	1	119.5
Colorado	39	149	0	105.5
Connecti	41.8	159	1	72.8
Delaware	39	200	1	75.5
Wash,D.C	39	177	0	77
Florida	28	197	1	82
Georgia	33	214	1	83.5
Idaho	44.5	116	0	114
Illinois	40	124	0	89.5
Indiana	40.2	128	0	86.2
Iowa	42.2	128	0	93.8
Kansas	38.5	166	0	98.5
Kentucky	37.8	147	0	85
Louisiana	31.2	190	1	91.8
Maine	45.2	117	1	69
Maryland	39	162	1	76.5
Massachu	42.2	143	1	71.8
Michigan	43.5	117	0	84.5
Minnesot	46	116	0	94.5
Mississipp	32.8	207	1	90
Missouri	38.5	131	0	92
Montana	47	109	0	110.5
Nebraska	41.5	122	0	99.5
Nevada	39	191	0	117
NewHamp	43.8	129	1	71.5
NewJerse	40.2	159	1	74.5
NewMexic	35	141	0	106
MewYork	43	152	1	75.5
NorthCarc	35.5	199	1	79.5
NorthDak	47.5	115	0	100.5
Ohio	40.2	131	0	82.8
Oklahoma	35.5	182	0	97.2
Oregon	44	136	1	120.5
Pennsylv	40.8	132	0	77.8
Rhodelsla	41.8	137	1	71.5
SouthCar	33.8	178	1	81
SouthDak	44.8	86	0	100
Tennesse	36	186	0	86.2
Texas	31.5	229	1	98
Utah	39.5	142	0	111.5
Vermont	44	153	1	72.5
Virginia	37.5	166	1	78.5
Washingt	47.5	117	1	121
WestVirgi	38.8	136	0	80.8
Wisconsir	44.5	110	0	90.2
Wyoming	43	134	0	107.5

Regression

- When you have a series of continuous data that follow some sort of pattern.
- determines the strength of the relationship between dependent variable and a series of other changing variables (known as independent variables).



Simple Linear Regression

- Statistical method that allows us to summarize and study relationships between two continuous variables
  - One variable, denoted as x, as the independent (predictor) variable
  - The other variable, denoted as y, as the dependent (response) variable

Parameters for line:

- In mathematics, a line needs two parameters:

$$y = mx + b$$

- m is the slope, b is the y-intercept
- In regression, the parameters take different names:
- $h(x) = \Theta_0 + \Theta_1 x$
- $h(x)$ is the predicted value for x
- Θ_0 , Θ_1 are the coefficients.

Linear Regression with one variable

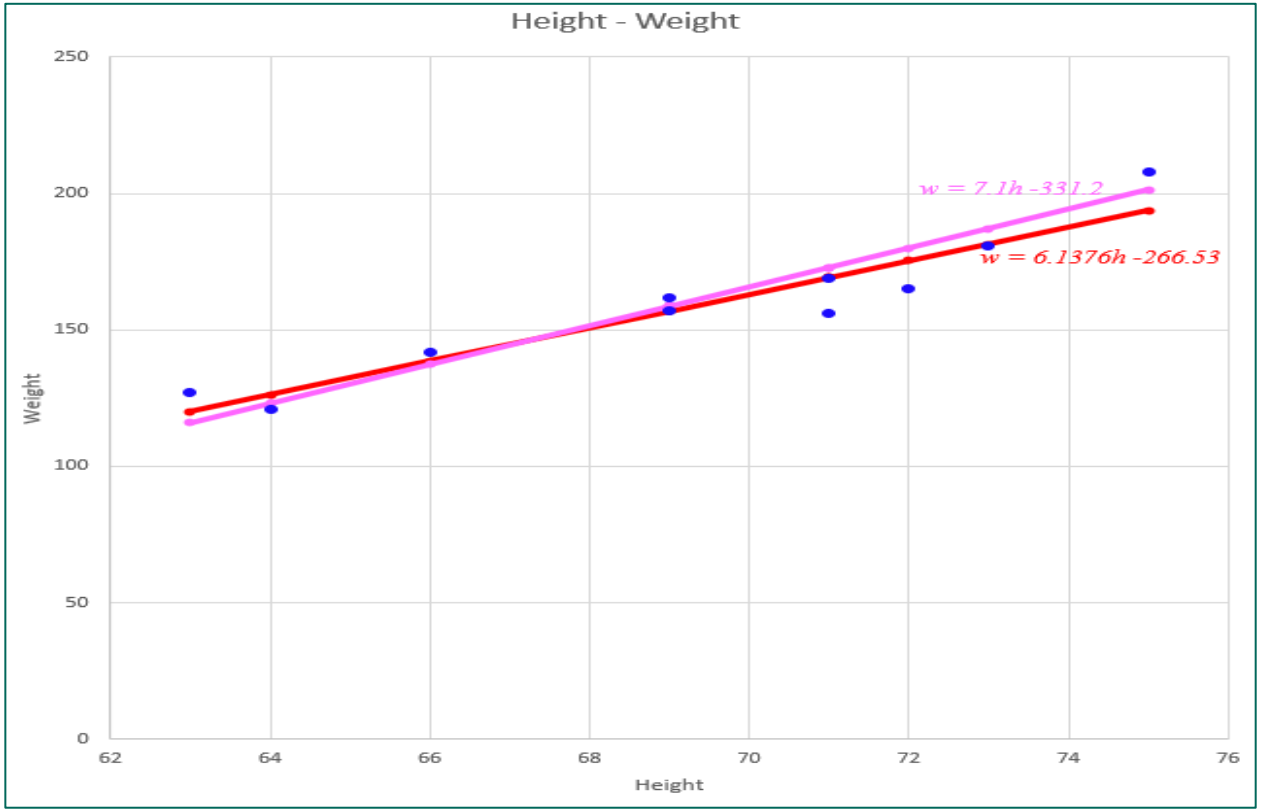
- Try to fit a best-fit line to a data set. This line is then used to predict real values for continuous output.
- Need a training set:
  - x – an input variable
  - y – The output variable.
  - h is a function that maps x -> y
  - $h(x) = \Theta_0 + \Theta_1 x$, or $y = mx + b$
- Also called Univariate linear regression.

Linear Regression with one variable

- To choose the best values of Θ_0 and Θ_1, we use a cost function.
- This calculates the total error between your predicted value, and the actual values. We continue to change the values until we find the minimum error.
- The h function deals with x, where the cost function deals with Θ_1.

Linear Regression - Example

Height	Weight
63	127
64	121
66	142
69	157
69	162
71	156
71	169
72	165
73	181
75	208



Which line (red or pink) is the best fit?

Red line: w = -266.53 + 6.1376h *Pink line: w = -331.2 + 7.1h*

For the student with the height 63 inches, actual weight is 127 pounds.
Based on the red fitted line, weight is -266.53 + 6.1376 * 63 = **120.1**
Prediction Error = 127 – 120.1 = 6.9

Based on the pink fitted line, weight is -331.2 + 7.1 * 63 = **116.1**
Prediction Error = 127 – 116.1 = 10.9

A line that fits the data "best" will be the one with overall minimal prediction errors.

In order to find the overall prediction error, "least squares criterion" can be used.

Least Squares Criterion

$w = -266.53 + 6.1376h$

		x	y_i	y_i'	$y_i - y_i'$	$(y_i - y_i')^2$
-266.53	6.1376	63	127	120.1388	6.8612	47.07607
-266.53	6.1376	64	121	126.2764	-5.2764	27.8404
-266.53	6.1376	66	142	138.5516	3.4484	11.89146
-266.53	6.1376	69	157	156.9644	0.0356	0.001267
-266.53	6.1376	69	162	156.9644	5.0356	25.35727
-266.53	6.1376	71	156	169.2396	-13.2396	175.287
-266.53	6.1376	71	169	169.2396	-0.2396	0.057408
-266.53	6.1376	72	165	175.3772	-10.3772	107.6863
-266.53	6.1376	73	181	181.5148	-0.5148	0.265019
-266.53	6.1376	75	208	193.79	14.21	201.9241
					Total	597.3863

$w = -331.2 + 7.1h$

		x	y_i	y_i'	$y_i - y_i'$	$(y_i - y_i')^2$
-331.2	7.1	63	127	116.1	10.9	118.81
-331.2	7.1	64	121	123.2	-2.2	4.84
-331.2	7.1	66	142	137.4	4.6	21.16
-331.2	7.1	69	157	158.7	-1.7	2.89
-331.2	7.1	69	162	158.7	3.3	10.89
-331.2	7.1	71	156	172.9	-16.9	285.61
-331.2	7.1	71	169	172.9	-3.9	15.21
-331.2	7.1	72	165	180	-15	225
-331.2	7.1	73	181	187.1	-6.1	37.21
-331.2	7.1	75	208	201.3	6.7	44.89
					Total	766.51

$y_i - y_i'$: Prediction error

$(y_i - y_i')^2$: Squared prediction error

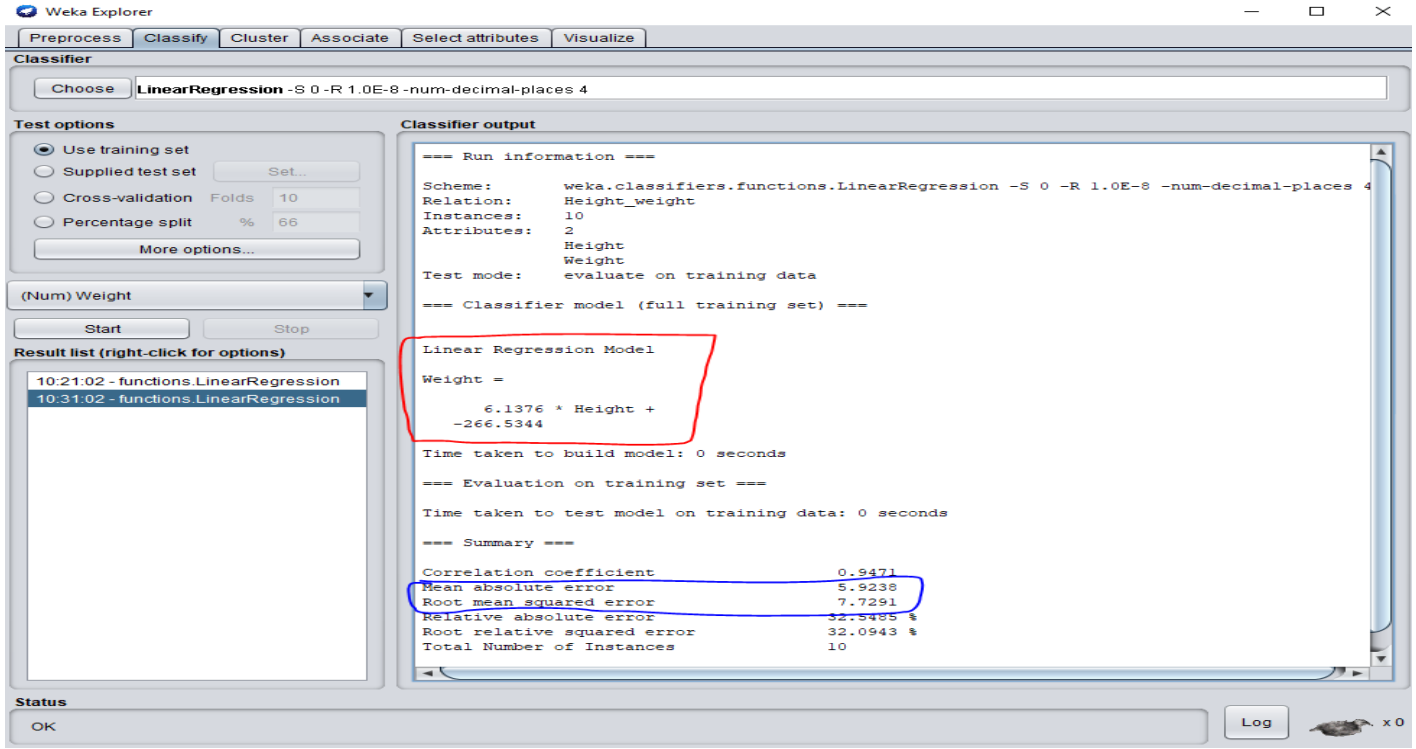
$$Overall\ squared\ prediciton\ error = \sum_{i=1}^{n} (yi - yi')2$$

Finding m and b

	x	y_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$x_i - \bar{x}$ * $y_i - \bar{y}$
	63	127	69.3	-6.3	39.69	158.8	-31.8	1011.24	200.34
	64	121	69.3	-5.3	28.09	158.8	-37.8	1428.84	200.34
	66	142	69.3	-3.3	10.89	158.8	-16.8	282.24	55.44
	69	157	69.3	-0.3	0.09	158.8	-1.8	3.24	0.54
	69	162	69.3	-0.3	0.09	158.8	3.2	10.24	-0.96
	71	156	69.3	1.7	2.89	158.8	-2.8	7.84	-4.76
	71	169	69.3	1.7	2.89	158.8	10.2	104.04	17.34
	72	165	69.3	2.7	7.29	158.8	6.2	38.44	16.74
	73	181	69.3	3.7	13.69	158.8	22.2	492.84	82.14
	75	208	69.3	5.7	32.49	158.8	49.2	2420.64	280.44
Sqrt(Sum)					11.7516			76.15510488	847.6
								$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	
									6.137581463

	m	6.1375815	
	SD	3.7161808	24.08236
	Mean	69.3	158.8
$b = \bar{y} - m\bar{x}$		-266.5344	

Weka Demo for Height-Weight file



Measuring accuracy - How can you tell if your regression line is a good fit?

- Calculate the "Coefficient of determination", the residual, or also called r^2, where r is the correlation coefficient.

- This is a number between 0 and 1, which normally means how close your data is to the line. If your data is always on the line, then $R^2 = 1$. If your data is far away from the line then R^2 will be low.

Measuring accuracy

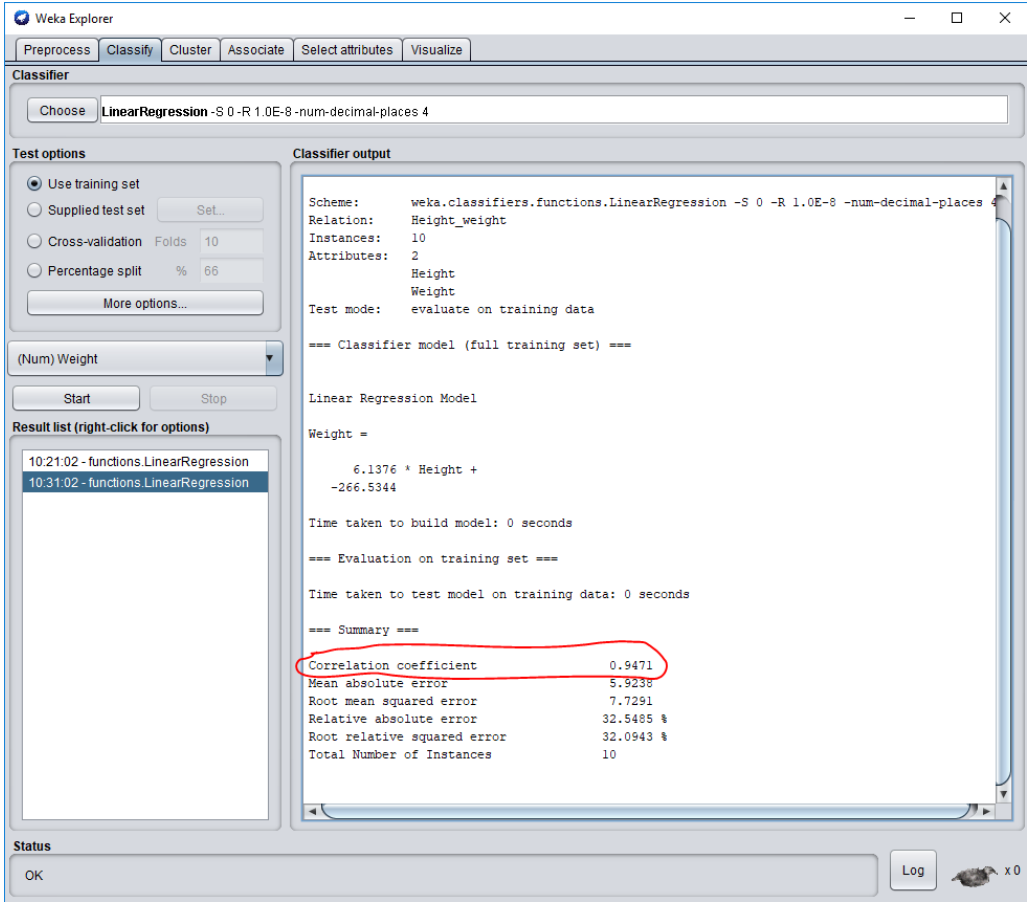
Correlation Coefficient

$$r = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} * b_1$$

where b_1 is the slope in the equation $y = b_0 + b_1 x$

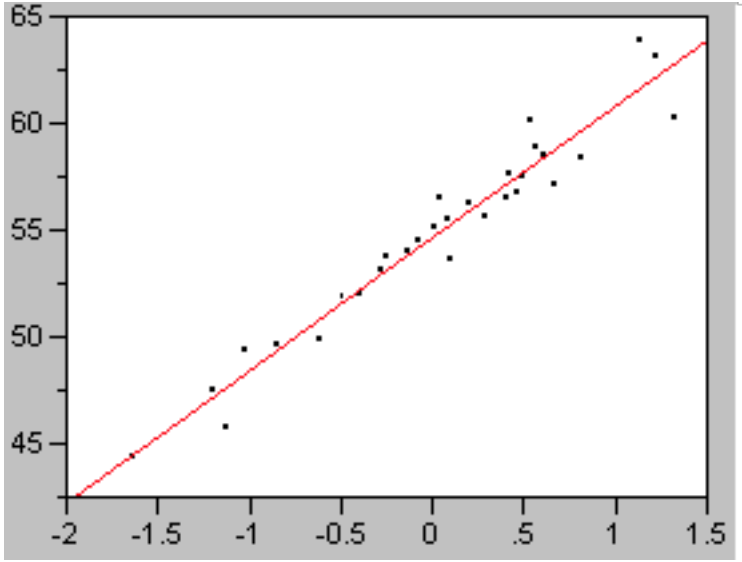
	x	y_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
	63	127	69.3	-6.3	39.69	158.8	-31.8	1011.24
	64	121	69.3	-5.3	28.09	158.8	-37.8	1428.84
	66	142	69.3	-3.3	10.89	158.8	-16.8	282.24
	69	157	69.3	-0.3	0.09	158.8	-1.8	3.24
	69	162	69.3	-0.3	0.09	158.8	3.2	10.24
	71	156	69.3	1.7	2.89	158.8	-2.8	7.84
	71	169	69.3	1.7	2.89	158.8	10.2	104.04
	72	165	69.3	2.7	7.29	158.8	6.2	38.44
	73	181	69.3	3.7	13.69	158.8	22.2	492.84
	75	208	69.3	5.7	32.49	158.8	49.2	2420.64
Sqrt(Sum)					11.7516			76.1551049
				Correlation Coefficient	$\frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} * b_1$		0.947101228	

Weka

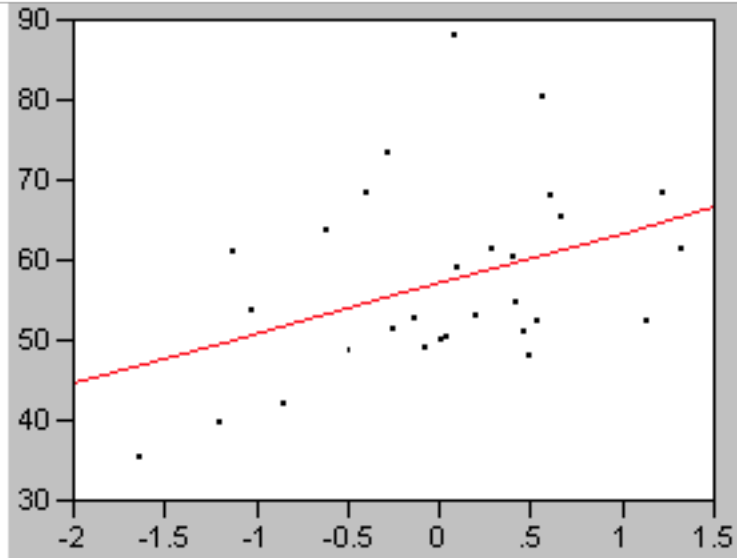

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) Weight
Start
Stop
Result list (right-click for options)
10:21:02 - functions.LinearRegression
10:31:02 - functions.LinearRegression
Classifier output
Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places
Relation: Height_weight
Instances: 10
Attributes: 2
Height
Weight
Test mode: evaluate on training data
=== Classifier model (full training set) ===
Linear Regression Model
Weight =
6.1376 * Height +
-266.5344
Time taken to build model: 0 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correlation coefficient 0.9471
Mean absolute error 5.9238
Root mean squared error 7.7291
Relative absolute error 32.5485 %
Root relative squared error 32.0943 %
Total Number of Instances 10
Status
OK
Log
x 0

Measuring accuracy

High R^2, data is close to line



Lower R^2, data is far from line



Multiple Regression Model

Linear Regression Model for cpu.arff:

class = 0.0491 * MYCT +
0.0152 * MMIN +
0.0056 * MMAX +
0.6298 * CACH +
1.4599 * CHMAX +
-56.075

The weights tells the relationship of each variable to the outcome, whether they are positive or negative.

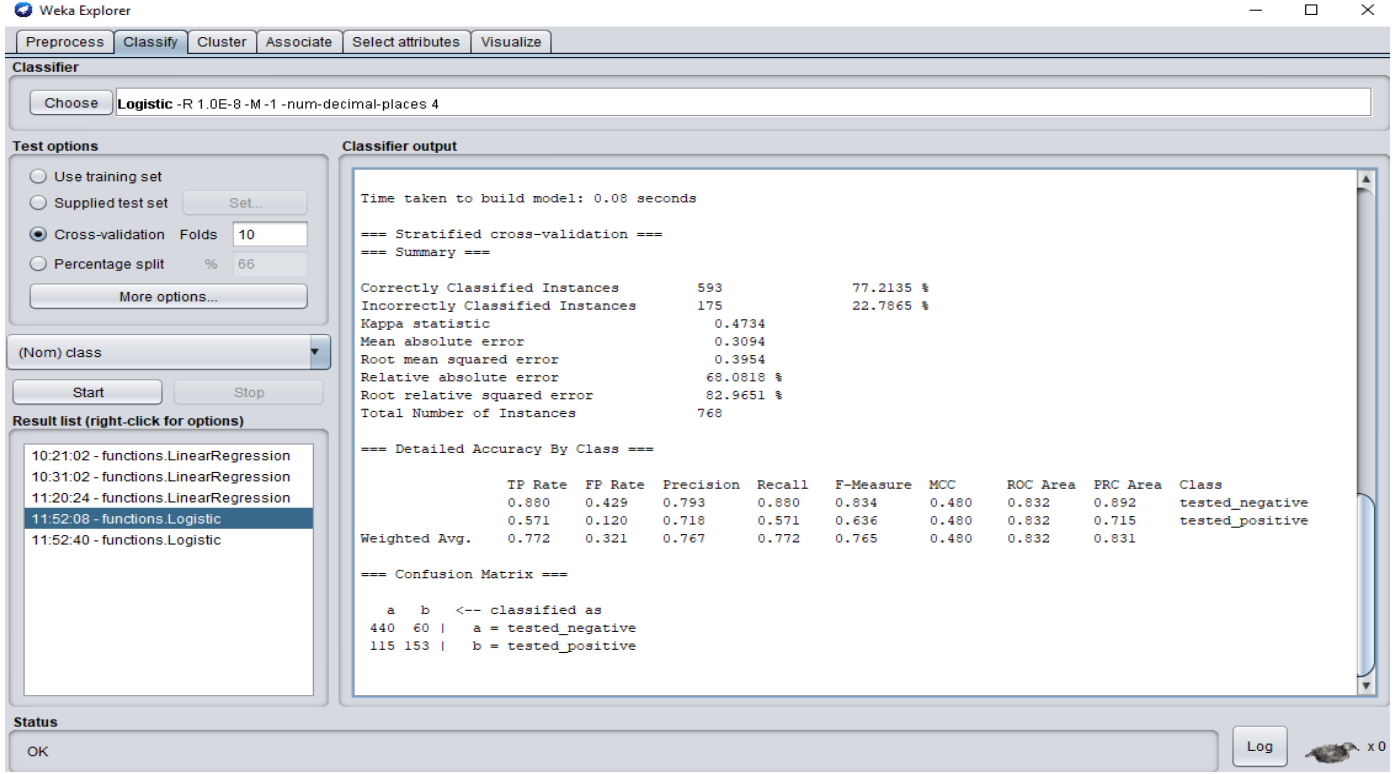
Multivariate Regression

- a technique that estimates a single regression model with more than one outcome variable.
- Example: A doctor has collected data on cholesterol, blood pressure, and weight. She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week). She wants to investigate the relationship between the three measures of health and eating habits.

Logistic Regression

- Models a relationship between independent (predictor) variable and a categorical response variable.
- Helps us to estimate a probability of falling into a certain level of the categorical response given a set of predictors

Weka Demo with Diabetes dataset



References

- https://www.youtube.com/watch?v=6tDnNyNZDF0
- https://www.youtube.com/watch?v=YIxoyiN8lxo
- https://www.youtube.com/watch?v=ThmZU3dTIDo
- https://onlinecourses.science.psu.edu/stat501/lesson/1

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 7

Association Rule

Association Rule Mining

- a **rule**-based machine learning method
- Objective is to discover interesting relations between variables in large databases.
- intended to identify strong **rules** discovered in databases using some measures of interestingness.

Association Rule Mining (Cont'd….)

- important data mining model studied extensively by the database and data mining community.
- Assume all data are categorical.
- No good algorithm for numeric data.
- Initially used for Market Analysis to find how items purchased by customers are related.

Example

- analyzes and predicts customer behavior
- if a customer buys bread, there is an 80% chance to buy butter too. bread → butter
- buys {onions, potatoes} → buys{tomatoes} (likely to buy tomatoes also)
- useful for marketing activities like product promotion or product pricing

Example (Cont'd…)

Bread → Butter[20%, 45%]

If then statement ... if Bread is purchased, then Butter is purchased

- LHS is antecedent
- RHS is consequent

- Bread : Antecedent
- Butter : Consequent
- 20% : Support
- 45% : Confidence

Example (Cont'd…)

A → B[Support, Confidence]

- Support denotes probability that contains A
- Confidence denotes probability that a transaction containing A also contains B

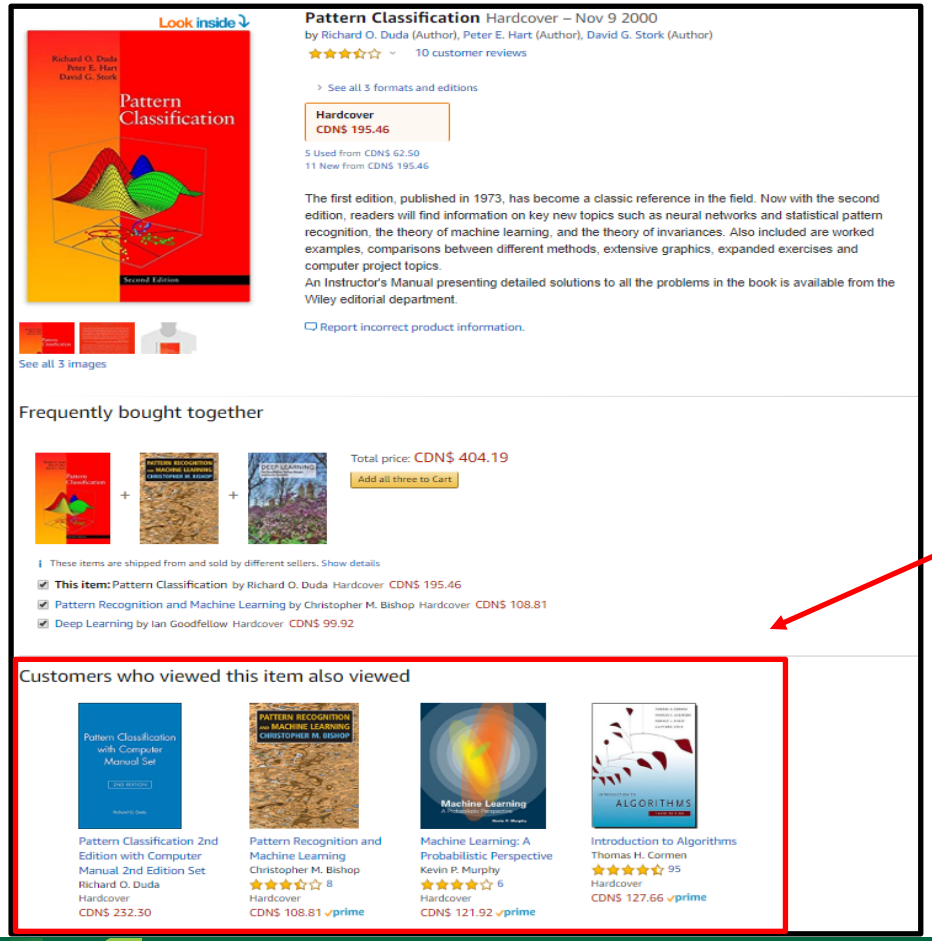
- Total Transactions in a store: 100
- Out of this 100, in 20 transactions, bread is present

- so 20/100 = 0.2 which is 20% is the **Support**

- Out of this 20 transactions, 9 has butter too.
- So, 9/20 = 0.45 = 45% which is the **Confidence**

Applications

- Web usage
- Banking
- Bioinformatics
- Market based analysis
- Credit/debit card analysis
- Product clustering
- Catalog design



Used in many recommender systems

Example 2

Transaction	Onion	Potato	Burger	Milk	Pepsi
T1	1	1	1	0	0
T2	0	1	1	1	0
T3	0	0	0	1	1
T4	1	1	0	1	0
T5	1	1	1	0	1
T6	1	1	1	1	0

Itemset I = {Onion, Potato, Burger, Milk, Pepsi}

Dataset has 6 transactions

Example rule:
{Onion, Potato} → {Burger}

Frequent Item Sets

- Ideally, we want to create all possible combinations of items
- Problem: When number of items increases, computational time increases exponentially
- Solution: Consider only “frequent item sets”
- Criteria to identify frequent items: Support

Algorithms for Association Rule

- Apriori Algorithm (the best)
- Elcat Algorithm
- F.P. Growth Algorithm

Apriori algorithm

Generating Frequent Item Sets:

- For k products:
  - User sets a minimum support criterion
  - Generate a list of one-item sets that meet the support criterion
  - Use the list of one-item sets to generate list of two-item sets that meet the support criterion
  - Use the list of two-item sets to generate list of three-item sets that meet the support criterion
  - Continue up through k-item sets

Support

$$\text{Support}(X) = \frac{\textit{Number of transactions in which X appears}}{\textit{Total number of transactions}}$$

Item	Frequency of transactions
Onion	4
Potato	5
Burger	4
Milk	4
Pepsi	2

Less than 50% which is 3.

Only 4 items are significant, which occurred greater than 50%

$$\text{Support(Onion)} = \frac{4}{6} = 0.6667$$

$$\text{Support(Potato)} = \frac{5}{6} = 0.8333$$

$$\text{Support(Burger)} = \frac{4}{6} = 0.6667$$

$$\text{Support(Milk)} = \frac{4}{6} = 0.6667$$

Support Threshold is set as 50%

Item	Frequency of transactions	Support
Onion	4	0.666666667
Potato	5	0.833333333
Burger	4	0.666666667
Milk	4	0.666666667
Pepsi	2	

Creating 2-itemset

Itemset	Frequency of transactions	Support
Onion, Potato	4	0.666666667
Onion, Burger	3	0.5
Onion, Milk	2	
Potato, Burger	4	0.666666667
Potato, Milk	3	0.5
Burger, Milk	2	

Creating 3-itemset

Itemset	Frequency of transactions	Support
Onion, Potato, Burger	3	0.5
Potato, Burger, Milk	2	

Apply self-join to create 3-itemset

→ from the 2-itemset, find two pairs with the same first item, so we get

[Onion, Potato] & [Onion Burger]. From this, we make [Onion, Potato, Burger]

Confidence

$$Conf(X \rightarrow Y) \frac{Support(X \cup Y)}{Support(X)}$$

- The confidence is 1 (maximal) for a rule X->Y if the consequent and antecedent always occur together.
- Range: [0, 1]

Lift

- $Lift(X \rightarrow Y) \frac{Support(X \cup Y)}{Support(X) * Support(Y)}$
- It measures how often X and Y happen together versus happen independently.
- If the lift is greater than 1, then there is a positive correlation.
- If the lift is less than 1, then it implies there is a negative correlation.
- If the lift is 1 then there is no relationship.
- Range: $[0, \infty]$

Results (Partial)

	Confidence	Lift
Onion --> Potato	1.00	1.2
Onion --> Burger	0.75	1.125
Potato --> Burger	0.80	1.2
Potato --> Milk	0.60	0.9
Potato --> Onion	0.80	1.2
Burger --> Onion	0.75	1.125
Burger --> Potato	1	1.2
Milk --> Potato	0.75	0.9

Demo in Weka to get full results!!!

Summary

- Association rules (or *affinity analysis,* or *market basket analysis*) produce rules on associations between items from a database of transactions
- Widely used in **recommender systems**
- Most popular method is **Apriori algorithm**
- To reduce computation, we consider only “frequent” item sets (=support)
- Performance is measured by *confidence* and *lift*

References

- http://www.saedsayad.com/association_rules.htm
- https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Introduction to BI

BI Components & Architecture

Professor: Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Agenda

- Assignment 3
- Final Project
- BI, Components & Architecture

Business Intelligence

- Business Intelligence systems have been around for almost 20 years now. They try to summarize large datasets and try to understand what is happening.
- Definition: Business Intelligence is the process of collecting data to put together a picture of what is going on in a business.
- These data are then analyzed to detect trends (good or bad) so that management staff can make informed decisions.

Business Intelligence

- It is meant to help management to make decisions:
  - Strategic – use information to make better decisions
  - Tactical – reaching short term goals
  - Operational – helps identify processes to optimize
- http://www.managementstudyguide.com/strategic-decisions.htm

Strategic vs Administrative vs Operational

Strategic Decisions	Administrative Decisions	Operational Decisions
Strategic decisions are long-term decisions.	Administrative decisions are taken daily.	Operational decisions are not frequently taken.
These are considered where The future planning is concerned.	These are short-term based Decisions.	These are medium-period based decisions.
Strategic decisions are taken in Accordance with organizational mission and vision.	These are taken according to strategic and operational Decisions.	These are taken in accordance with strategic and administrative decision.
These are related to overall Counter planning of all Organization.	These are related to working of employees in an Organization.	These are related to production.
These deal with organizational Growth.	These are in welfare of employees working in an organization.	These are related to production and factory growth.

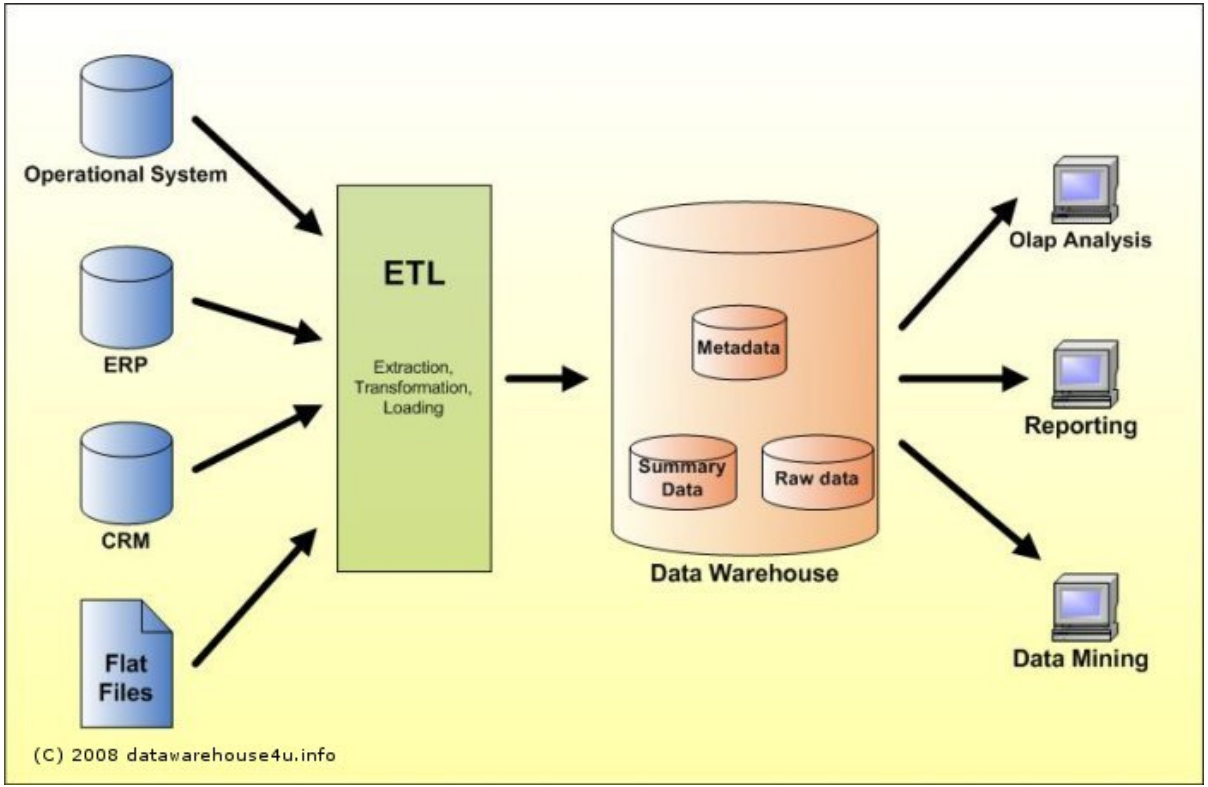
Business Intelligence

- BI systems have 4 main components:
  - Data Warehouses
  - Business Analytics tools – manipulating and analyzing data
  - Business Performance tools – monitoring and analyzing performance
  - Visualization– portals, dashboards and scorecards.

Data Warehouse

- A Data Warehouse is a large store of data that is accumulated from different sources within a company.
- The Data Warehouse pulls data from inventory, sales, etc. so that an overall picture can be created to support decision making. A Data Warehouse is typically updated on a daily basis.

Data warehouse

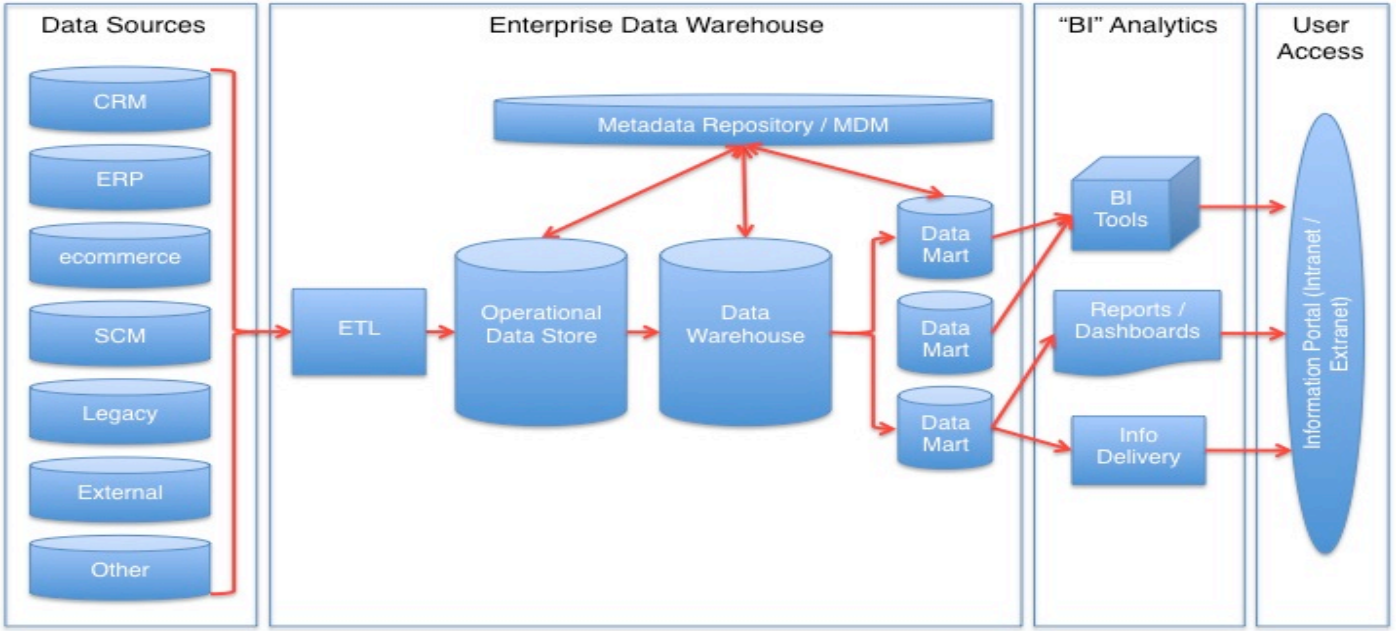


Picture taken from www.datawarehouse4u.info

ETL

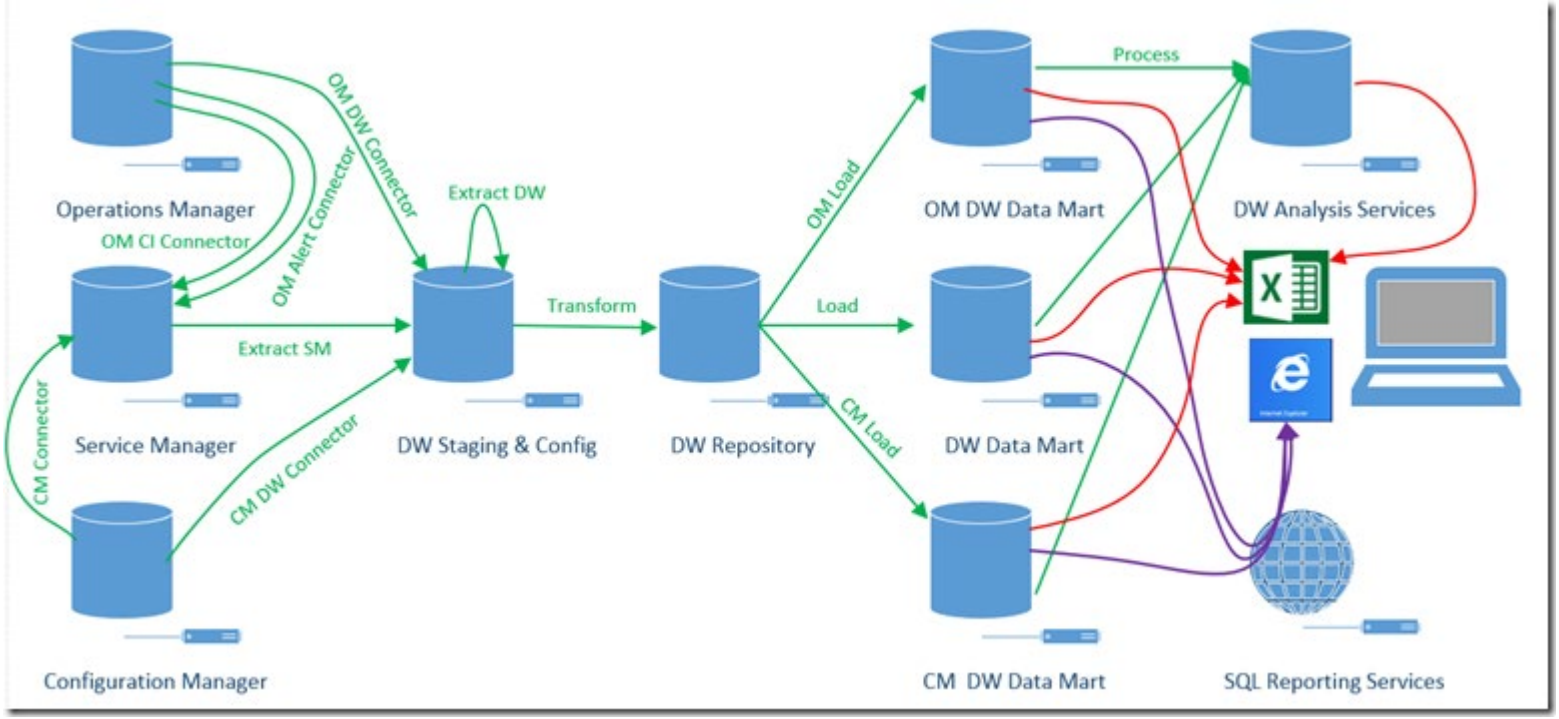
- The process of gathering data is referred to as Extract, Transform, Load (ETL)
  - Extraction pulls data from its raw source.
  - Transform manipulates the data to convert it to the format you want.
  - Load stores the data in the final target (Data Warehouse or Data Mart).
- If the data won't fit into memory, then you must store it first at the target database and then transform it after. This process is referred to as ELT, where it is loaded before being transformed.

Data Warehouse



Picture taken from https://en.wikipedia.org/wiki/Data_warehouse

Data warehouse (Contd.)



Picture taken from https://blogs.technet.microsoft.com/servicemanager/2013/02/16/new-data-warehouse-architecture-diagram/

Data Warehouse vs Database

	Data Warehouse	Database
Definition	Pulls together data from different sources for reporting and analysis	Normalized data organized in columns, rows and tables
Purpose	To store large quantities of historical data and enable fast, complex queries across all data	To store current transactions and enable fast access to specific transactions for ongoing business processes
Data structure	Use fewer tables Doesn't exclude data redundancies	Related data separated into multiple tables. Data is organized to make sure no redundant data
Data	Denormalized data Offers better performance when reading data for analytics purposes	Normalized data More complex queries required to read the data as a single data combines data from many tables

OnLine Transaction Processing System - OLTP

- These are systems that process any type of transaction (Banking, Ticket Reservation, Point of Sale, etc.)
- Online means that it is always running, waiting for input.
- They ensure speed of transactions and maintain data integrity.
- Mainly INSERT, UPDATE, DELETE operations

OnLine Analytical Processing - OLAP

- OLAP systems operate long-running queries through the data looking for patterns. These are the systems that generate business reports and mines the data.
- They work with data from the Data Warehouse to provide summaries.
- http://www.solverglobal.com/blog/2014/04/data-warehouse-vs-olap-cube/

OLTP Query Example

```
SELECT PASSWORD FROM ACCOUNTS
WHERE USERID = 1131234567
```

OLAP Query Example

```
SELECT dealer, year, SUM(price)
FROM (Sales NATURAL JOIN Autos) JOIN Days ON date = day
WHERE model = 'Gobi' AND
      color = 'red' AND
      (year = 2001 OR year = 2002)
GROUP BY year, dealer;
```

OLTP vs. OLAP

OLAP	OLTP
Works on historical data	Works on current data
Helps analyze the business	Helps run the business
Provides summarized data	Provides raw data
Use long-running queries to refresh summaries	Focus on fast & secure queries
Used by specialized users for decision support	Used by normal company staff

Data Marts

- A Data Mart is a mini warehouse, that typically addresses one portion of the warehouse: sales, marketing, finance, etc. Each part of the organization would have their own specialized data mart.
- It is extracted from a Data Warehouse and can store summarized data instead of raw data.

Decision Making

- In Decision Making, it is important to understand what is happening, so you can do something about it:
  - Define the problem.
  - Constructing a Model for computer simulations.
  - Identify and Evaluating possible solutions
  - Recommending potential solutions.

Decision Support Systems

- There are several kinds of decisions that can be faced:
  - Structured decisions – choosing whether to buy a new product or service, what inventory levels are needed, etc.
  - Semi-structured decisions – setting budgets, pricing
  - Unstructured decisions – no clear solutions: advertising, designs
- Read: http://www.inc.com/encyclopedia/decision-support-systems.html

Types of DSS

- Data Driven – Process numerical data from data warehouse. Produces dashboards and scorecards.
- Document Driven – Works on documents, videos, transcripts, media (Wikipedia is an example).
- Knowledge Driven – Uses knowledge and past experiences to support decision making. (Neural networks)
- Model Driven – Uses models and simulation to support decision making.
- Communication Driven – Group support systems for collaboration (Facebook, Skype)

Dashboards

- Once the data has been gathered, loaded and transformed, the results can be presented.
- Dashboards are a graphical presentation showing whatever data that a decision maker needs to know.
- This process is automated on a regular basis. For reports that are not needed on a regular basis, there are ad-hoc reports which are executed when needed.

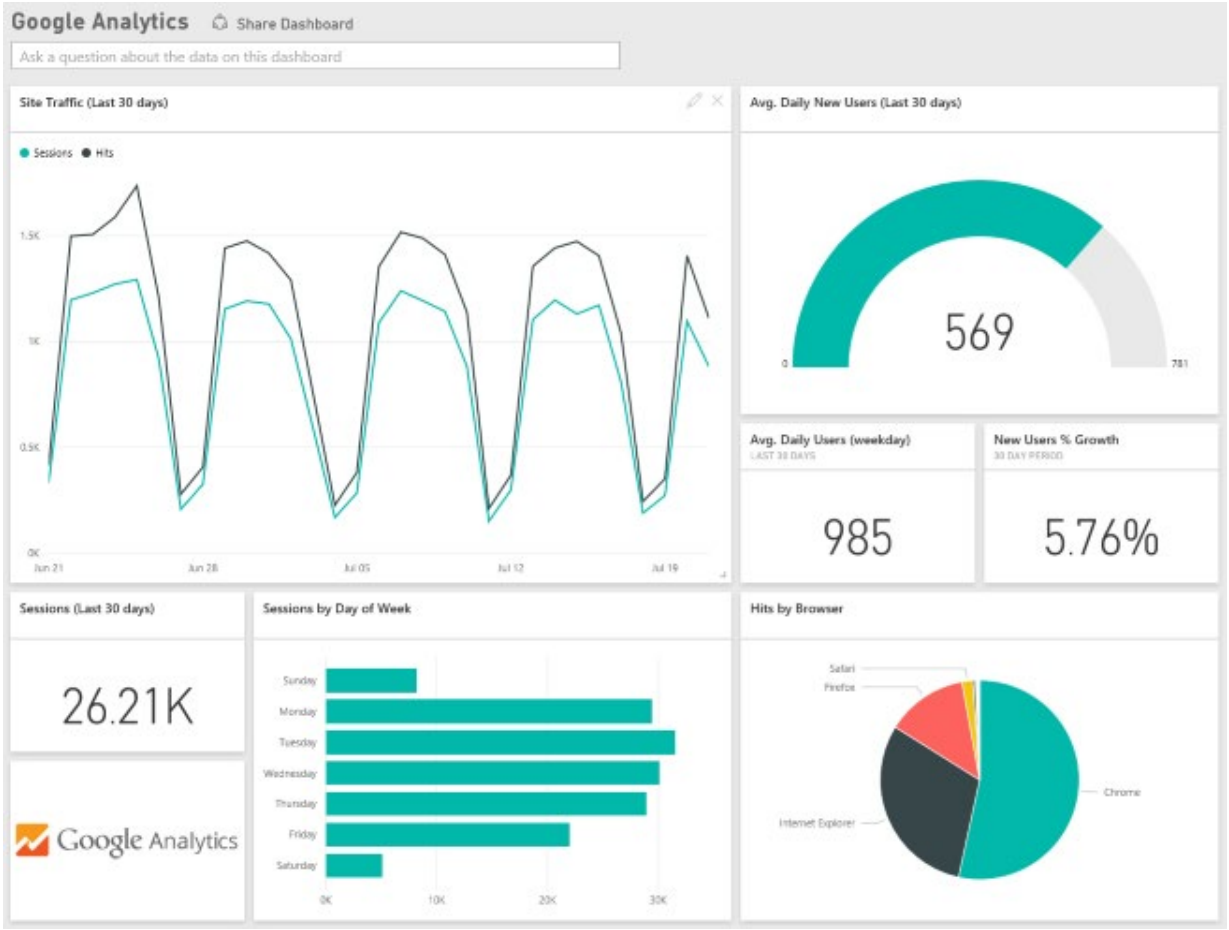
Dashboards

- Dashboards are graphical interfaces that display Key Performance Indicators (KPIs). It gives managers a quick overview of the KPIs.
- A dashboard should draw your attention to items that are not within expected ranges (sales are down, accidents are high etc.)
- Dashboards should be hierarchical, meaning that a user should be able to “drill down” to view more of the data.

Dashboard Example

http://enterprise-dashboard.com/sas-dashboards-to-get-slick-kpi-gauges-out-of-the-box-in-next-release/

Dashboard Example



Drilling down

- Clicking on a dashboard item (for example: total sales) should show another screen showing the data that was used in the calculation.
- Drilling down on total sales should show a detailed screen showing sales in all provinces, or states. Clicking on one of the provinces should then show sales by city in the province. Clicking on a city should then show sales per store within that city.

Key Performance Indicators KPI

- Key performance indicators are a set of indicators to measure data against a goal, or success metric.
- Did the company meet their sales goal for the month?
- Were products built and delivered on time?
- Was quality (defects per 1000 units) maintained?
- https://www.klipfolio.com/resources/kpi-examples
- http://www.pnmsoft.com/resources/bpm-tutorial/key-performance-indicators/
- https://www.collegesontario.org/en/resources/2019-20-key-performance-indicators

Review

- Data is retrieved from the organization (Extraction)
- It is collected, cleaned converted and processed (Transformed)
- The results are stored in databases (Loaded)
- Decision makers can then view the data in graphical form using Dashboards.
- There are different data types (Structured, Unstructured, Semi-structured).

Review

- Watch videos on Business Intelligence:
  - https://www.youtube.com/watch?v=LFnewuBsYiY
  - https://www.youtube.com/watch?v=KGHbY_Sales
- Read Chapter 1 of textbook (skip case study)
- Some additional references:
  - https://www.promptcloud.com/blog/business-intelligence-Vs-data-analytics/
- https://www.inetsoft.com/evaluate/demo/flashdemo.jsp
- https://www.youtube.com/watch?v=eiY8GEF_0-U

ALGONQUIN COLLEGE

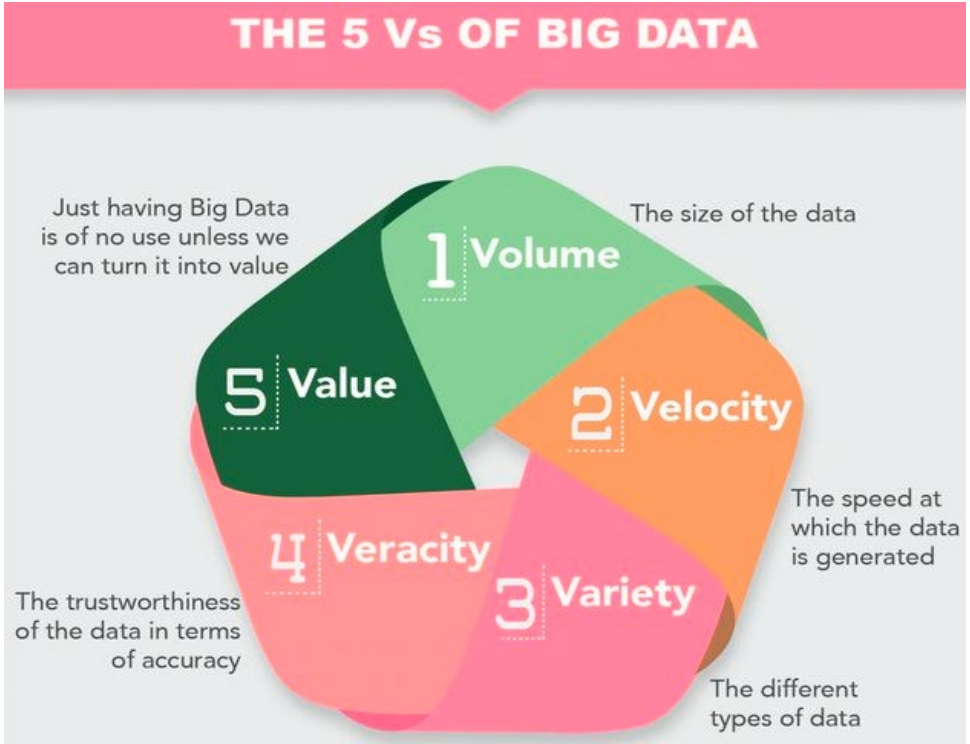
CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 10

Big Data

Characteristics of Big Data (Five Vs)

- Volume
- Velocity
- Variety
- Veracity
- Value



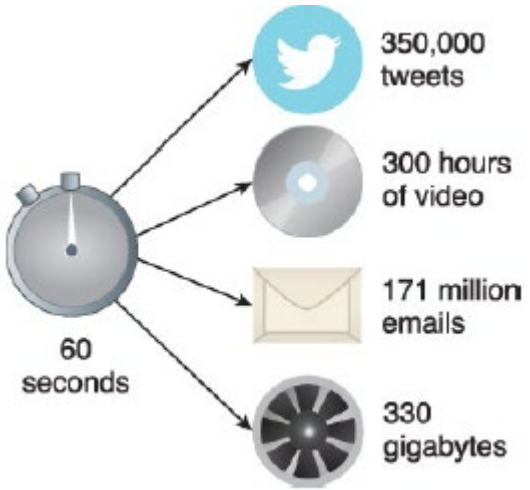
Taken from: http://bigdata.black/featured/what-is-big-data/

Volume

- refers to the vast amount of data that is generated every second/minute/hour/day in the digitized world
- Examples of data sources:
  - Online transactions such as point-of-sale and banking
  - Sensors such as GPS sensors, accelerometer, gyroscope etc.
  - Social media such as Facebook and Twitter

Velocity

- refers to the speed at which data is being generated and the pace at which data moves from one point to the next



Variety

- refers to the ever-increasing different forms of data that can come in
- Brings challenges in terms of data integration, transformation, processing and storage

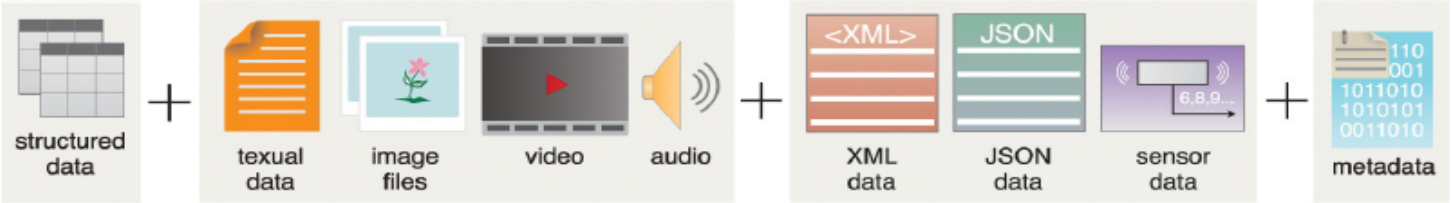


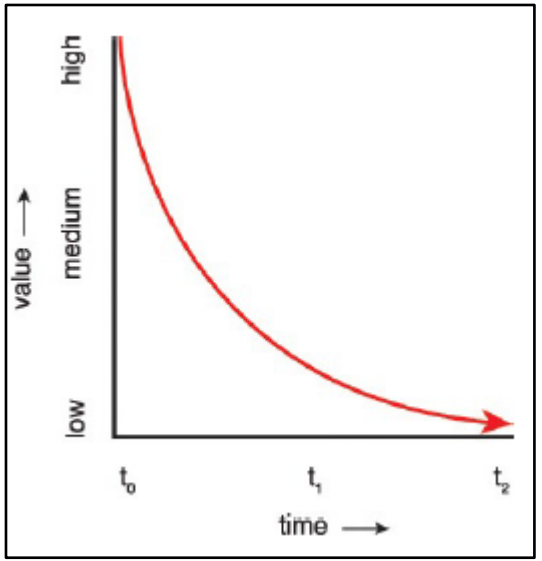
Figure 1.14 Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data and metadata.

Veracity

- refers to the quality of the data, which can vary greatly.
- Noise from the data to be removed

Value

- Refers to the usefulness of data for an enterprise
- Value and time are inversely related. The longer it takes for data to be turned into meaningful information, the less value it has for a business.



Big Data Analytics Lifecycle

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Big Data Storage Concepts

- Clusters
- Distributed File Systems
- Sharding
- Replication
  - Master-Slave
  - Peer-to-Peer

Data Wrangling

- Data acquired from external sources is often not in a format or structure that can be directly processed. To overcome these incompatibilities and prepare data for storage and processing, data wrangling is necessary.

- **Data wrangling includes steps to filter, cleanse and otherwise prepare the data for downstream analysis.**

- From a storage perspective, a copy of the data is first stored in its acquired format, and, after wrangling, the prepared data needs to be stored again.

https://tdwi.org/articles/2017/02/10/data-wrangling-and-etl-differences.aspx

https://www.talend.com/resources/data-wrangling-vs-etl/

Clusters

- A cluster is a tightly coupled collection of servers, or nodes. These servers usually have the same hardware specifications and are connected together via a network to work as a single unit.
- Each node in the cluster has its own dedicated resources, such as memory, a processor, and a hard drive.
- A cluster can execute a task by splitting it into small pieces and distributing their execution onto different computers that belong to the cluster.

Distributed File Systems

- A file system is the method of storing and organizing data on a storage device, such as flash drives, DVDs and hard drives.
- A file system provides a logical view of the data stored on the storage device and presents it as a tree structure of directories and files
- A distributed file system is a file system that can store large files spread across the nodes of a cluster. To the client, files appear to be local; however, this is only a logical view as physically the files are distributed throughout the cluster. This local view is presented via the distributed file system and it enables the files to be accessed from multiple locations.

Sharding

- Process of horizontally partitioning a large dataset into a collection of smaller, more manageable datasets called *shards*. The shards are distributed across multiple nodes, where a node is a server or a machine.
- Each shard is stored on a separate node and each node is responsible for only the data stored on it.
- Each shard shares the same schema, and all shards collectively represent the complete dataset.

Sharding - Example

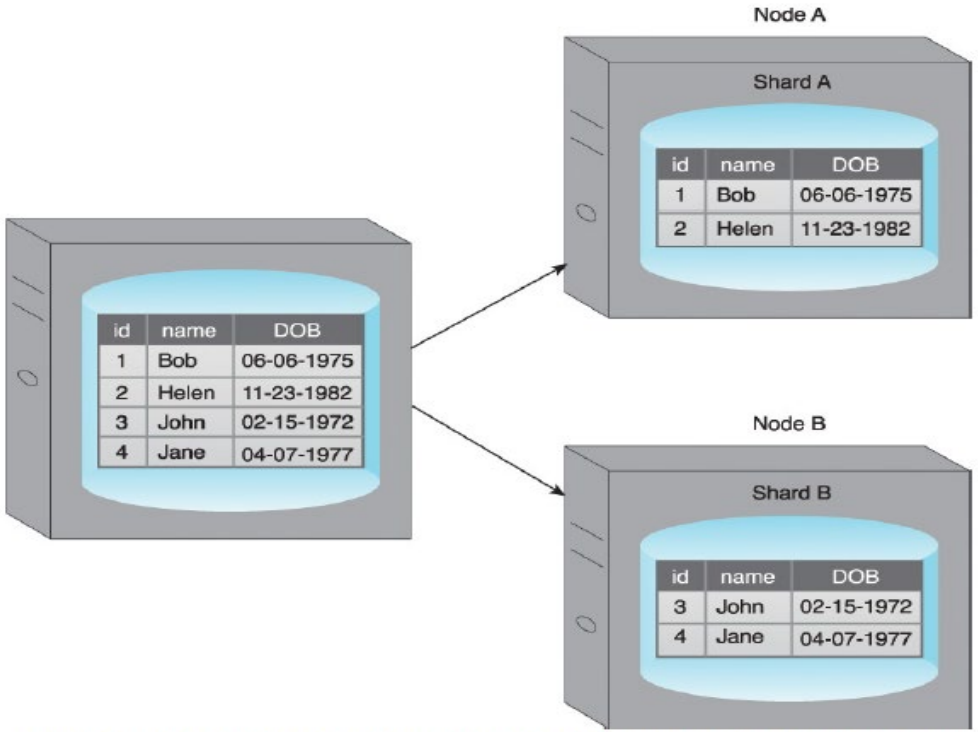


Figure 5.5 An example of sharding where a dataset is spread across Node A and Node B, resulting in Shard A and Shard B, respectively.

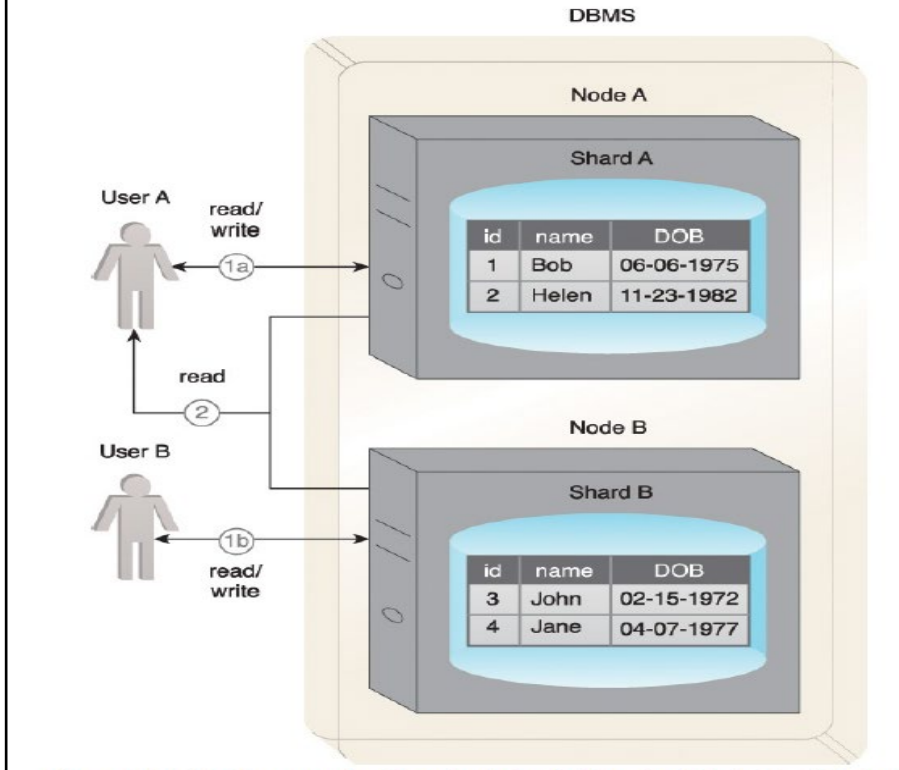


Figure 5.6 A sharding example where data is fetched from both Node A and Node B.

Sharding

- Sharding allows the distribution of processing loads across multiple nodes to achieve horizontal scalability.
- Horizontal scaling is a method for increasing a system's capacity by adding similar or higher capacity resources alongside existing resources.
- Since each node is responsible for only a part of the whole dataset, read/write times are greatly improved.

Replication

- Stores multiple copies of a dataset, known as *replicas*, on multiple nodes
- Replication provides scalability and availability due to the fact that the same data is replicated on various nodes.
- Fault tolerance is achieved since data redundancy ensures that data is not lost when an individual node fails.
- There are two different methods that are used to implement replication:
  - master-slave
  - peer-to-peer

Replication

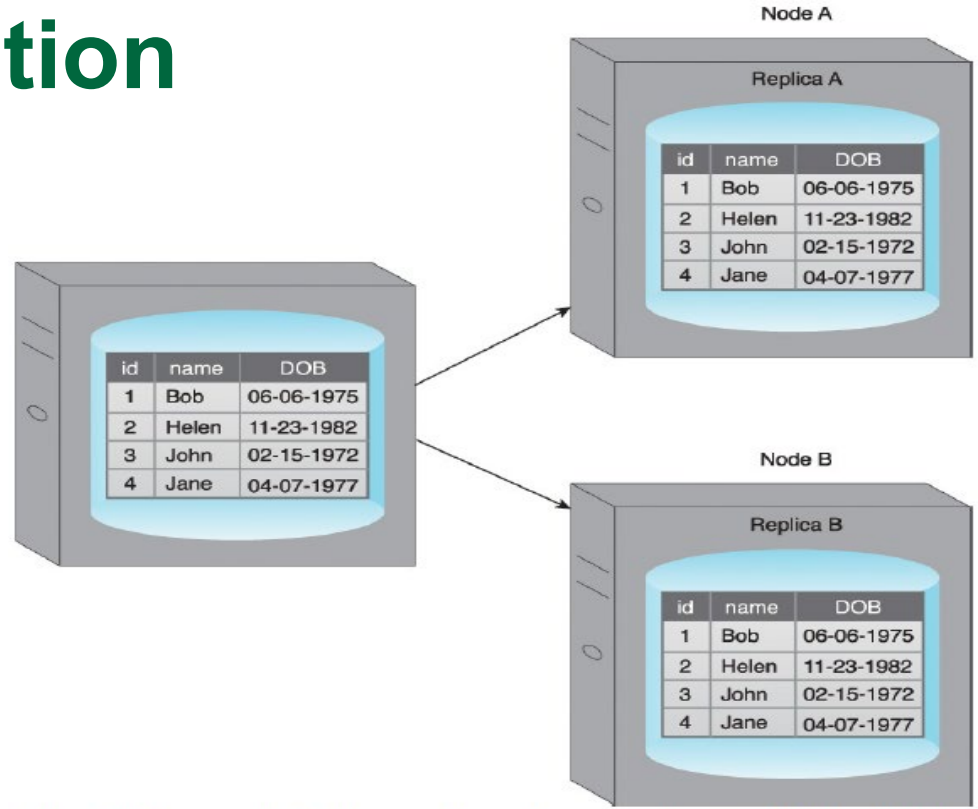


Figure 5.7 An example of replication where a dataset is replicated to Node A and Node B, resulting in Replica A and Replica B.

Master-Slave Replication

- nodes are arranged in a master-slave configuration, and all data is written to a master node. Once saved, the data is replicated over to multiple slave nodes.

- All external write requests, including insert, update and delete, occur on the master node, whereas read requests can be fulfilled by any slave node.

Master-Slave Replication

- Ideal for read intensive loads rather than write intensive loads since growing read demands can be managed by horizontal scaling to add more slave nodes.
- Writes are consistent, as all writes are coordinated by the master node. The implication is that write performance will suffer as the amount of writes increases.
- If the master node fails, reads are still possible via any of the slave nodes.
- A slave node can be configured as a backup node for the master node. In the event that the master node fails, writes are not supported until a master node is reestablished.
- The master node is either resurrected from a backup of the master node, or a new master node is chosen from the slave nodes.

Peer-to-Peer

- With peer-to-peer replication, all nodes operate at the same level.
- Each node (peer) is equally capable of handling reads and writes.
- Each write is copied to all peers

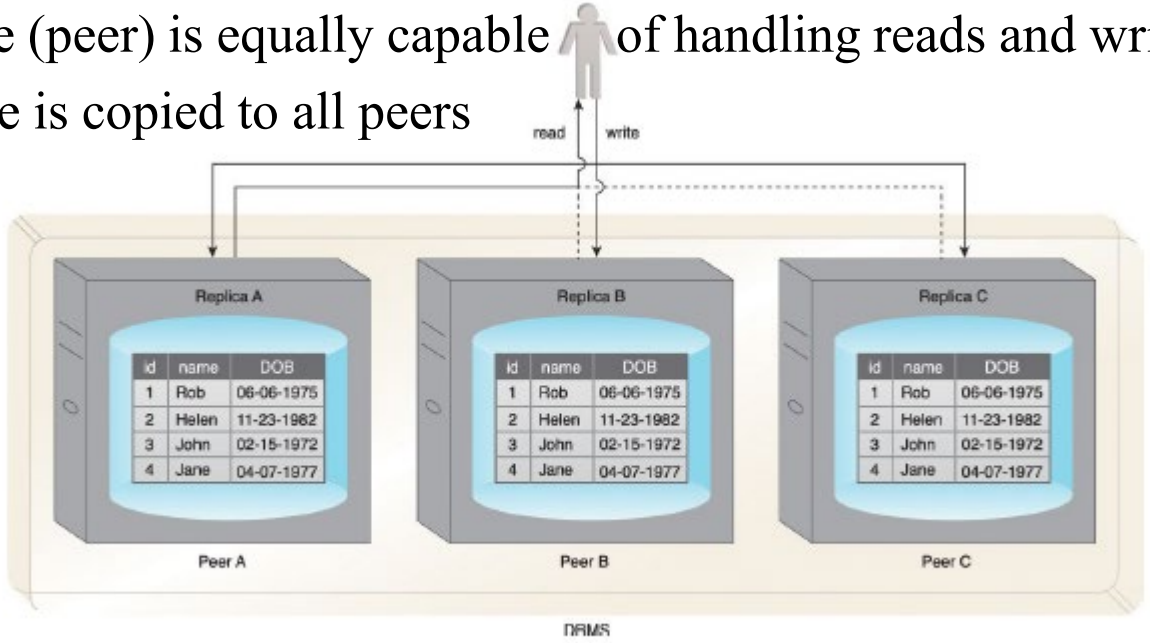


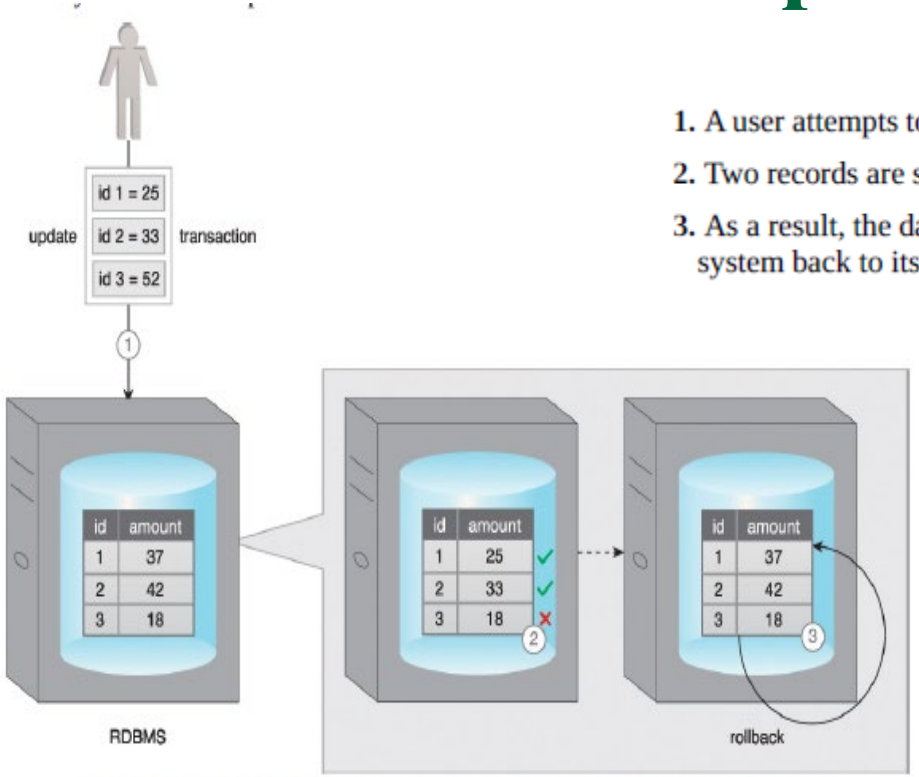
Figure 5.10 Writes are copied to Peers A, B and C simultaneously. Data is read from Peer A, but it can also be read from Peers B or C.

ACID

Database design principle related to transaction management

- Atomicity – ensures that all operations will always succeed or fail completely (no partial transactions)
- Consistency - ensures that the database will always remain in a consistent state by ensuring that only data that conforms to the constraints of the database schema can be written to the database.
- Isolation - ensures that the results of a transaction are not visible to other operations until it is complete.
- Durability - ensures that the results of an operation are permanent. In other words, once a transaction has been committed, it cannot be rolled back. This is irrespective of any system failure.

ACID – Example of Atomicity



1. A user attempts to update three records as a part of a transaction.
2. Two records are successfully updated before the occurrence of an error.
3. As a result, the database roll backs any partial effects of the transaction and puts the system back to its prior state.

Figure 5.18 An example of the atomicity property of ACID is evident here.

ACID – Example of Consistency

1. A user attempts to update the amount column of the table that is of type float with a varchar value.
2. The database applies its validation check and rejects this update because the value violates the constraint checks for the amount column.

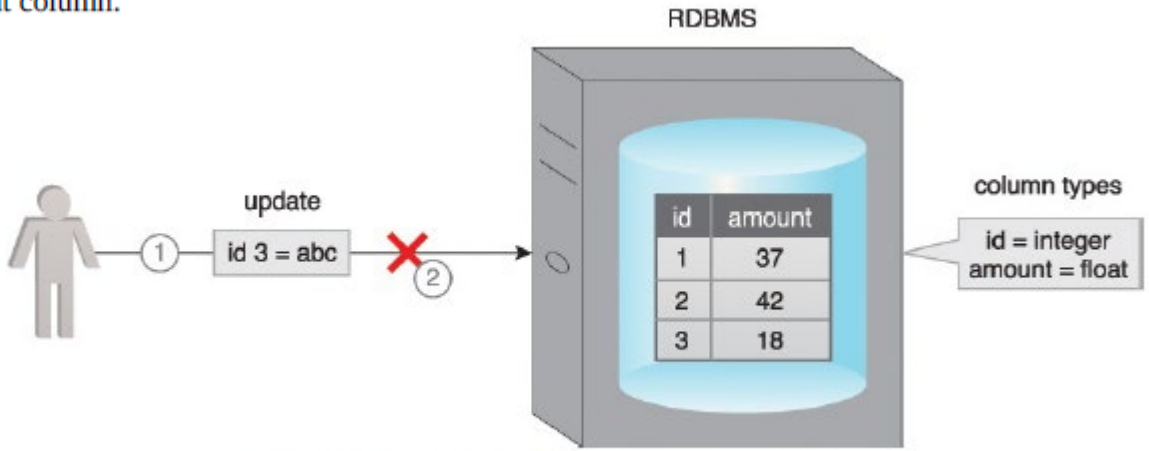
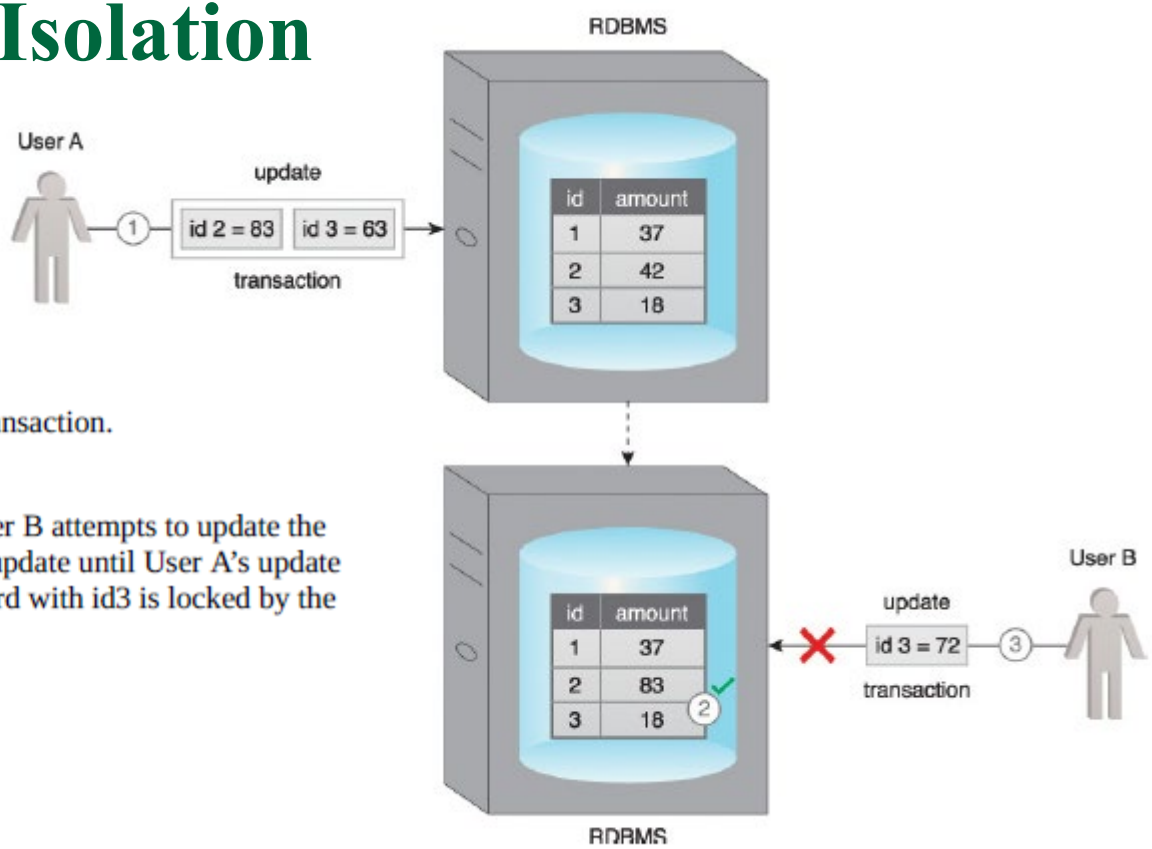


Figure 5.19 An example of the consistency of ACID.

ACID – Example of Isolation



1. User A attempts to update two records as part of a transaction.
2. The database successfully updates the first record.
3. However, before it can update the second record, User B attempts to update the same record. The database does not permit User B's update until User A's update succeeds or fails in full. This occurs because the record with id3 is locked by the database until the transaction is complete.

Figure 5.20 An example of the isolation property of ACID.

ACID – Example of Durability

1. A user updates a record as part of a transaction.
2. The database successfully updates the record.
3. Right after this update, a power failure occurs. The database maintains its state while there is no power.
4. The power is resumed.
5. The database serves the record as per last update when requested by the user.

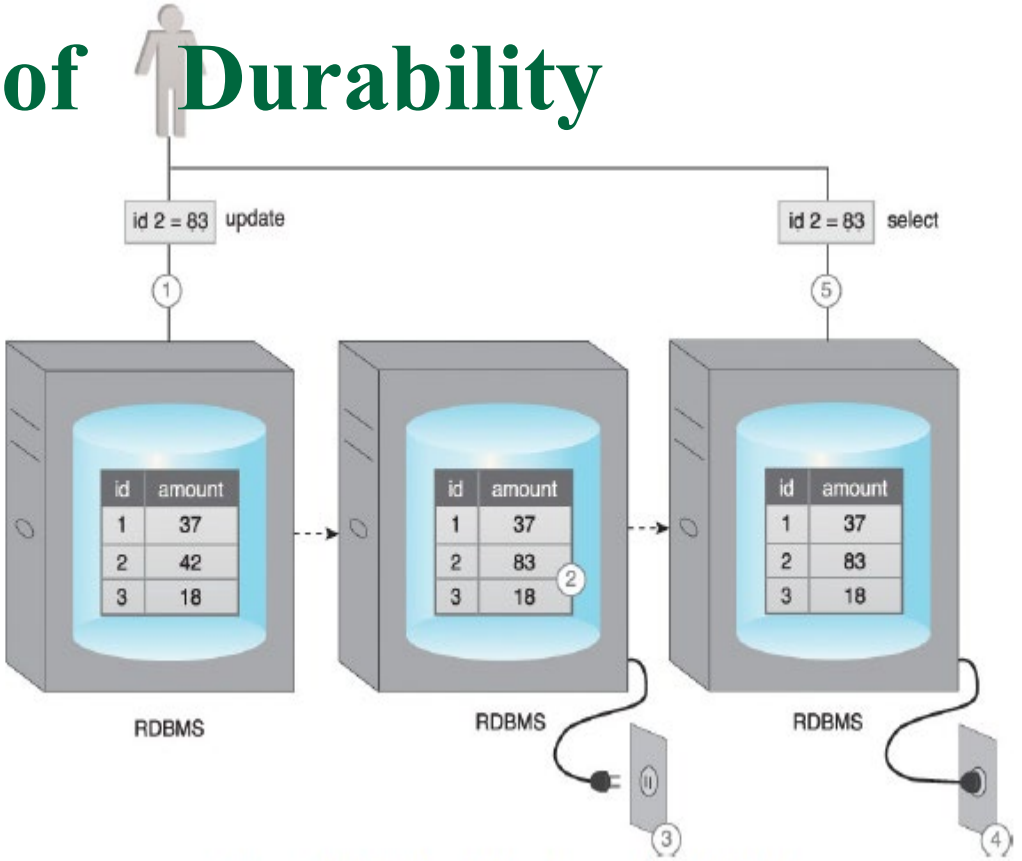


Figure 5.21 The durability characteristic of ACID.

ACID Principle - Example

1. User A attempts to update a record as part of a transaction.
2. The database validates the value and the update is successfully applied.
3. After the successful completion of the transaction, when Users B and C request the same record, the database provides the updated value to both the users.

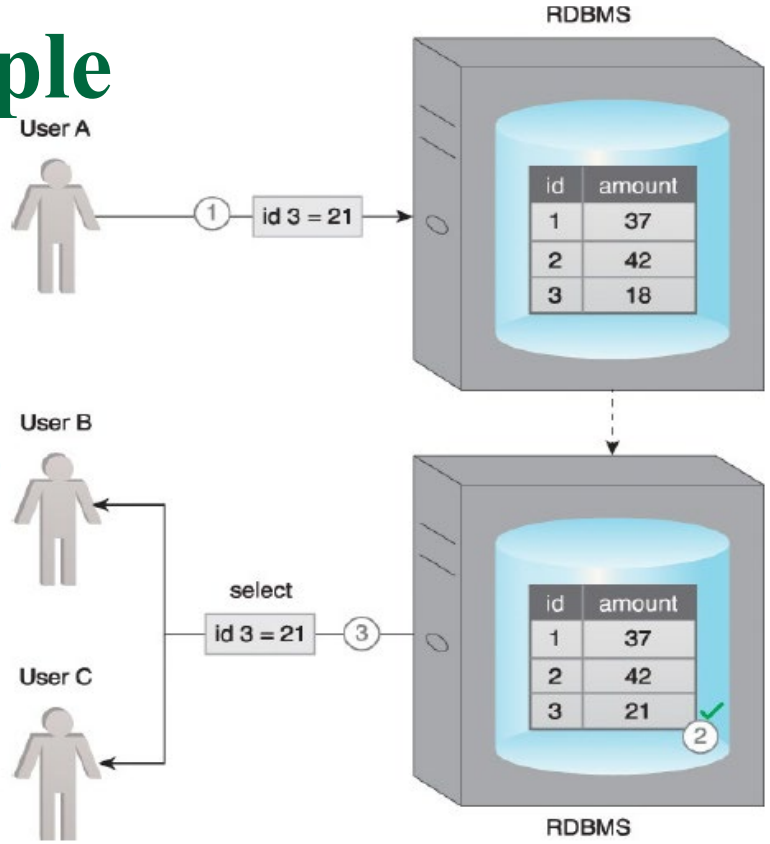


Figure 5.22 The ACID principle results in consistent database behavior.

ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

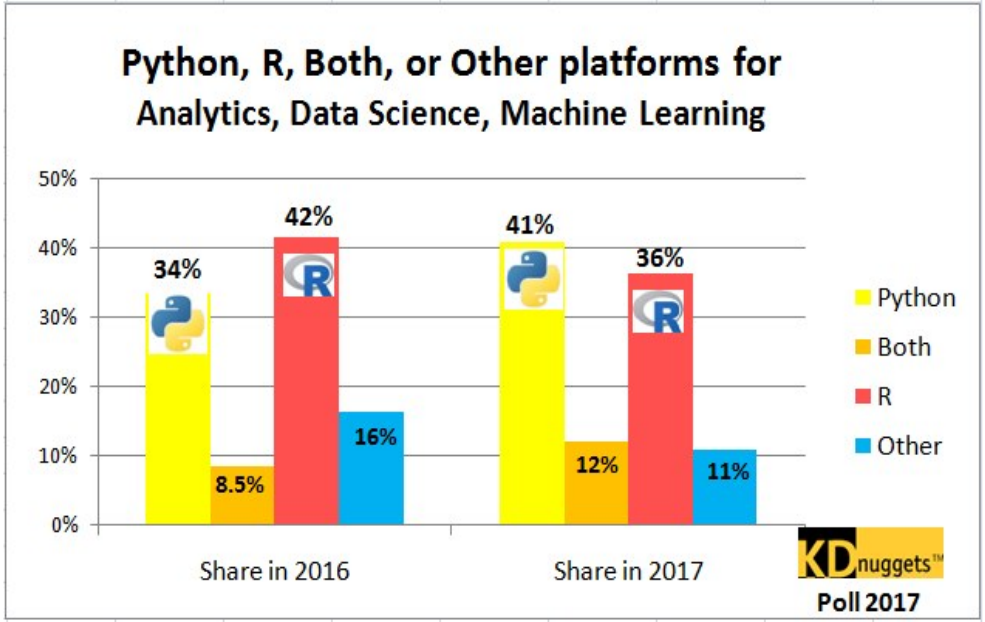
Data Science Tools

Professor : Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Tools for Data Science

Data science tool	% of respondents using the tool
Python	76.3
R	59.2
SQL	53.6
Jupyter notebooks	40.3
TensorFlow	28.4
Amazon Web services	23.5
Unix shell / awk	23.3
Tableau	20.4
C/C++	19.2
NoSQL	19.2

Showing 1 to 10 of 49 entries Previous Next



Taken from https://blog.appliedai.com/data-science-tools/

Plug & Play Data Science Tools

- RapidMiner
- DataRobot
- BigML
- Google Cloud AutoML
- Paxata
- Trifacta
- MLBase
- Weka
- Driverless AI
- MS Azure ML Studio

Top Analytics/Data Science Tools

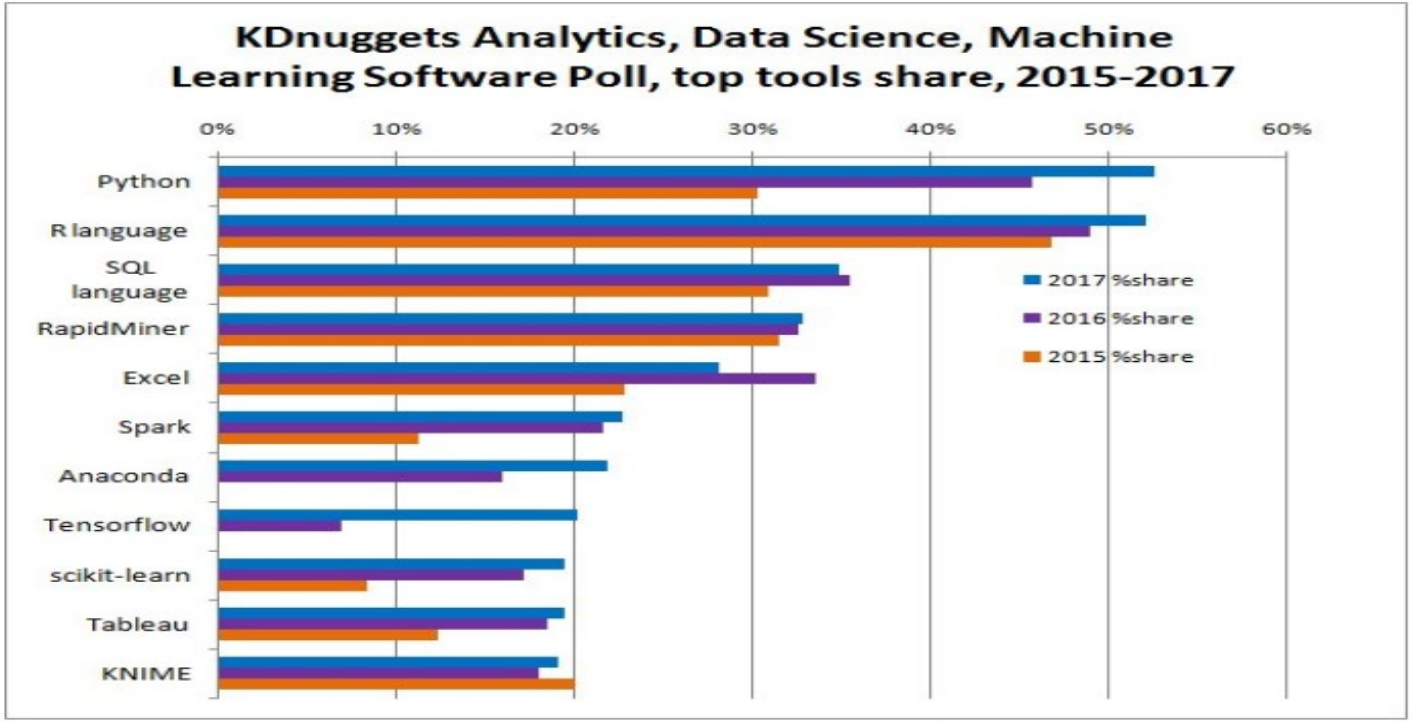


Fig 1: KDnuggets Analytics/Data Science 2017 Software Poll: top tools in 2017, and their usage in the 2015-6 polls

Taken from https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Python - Introduction

- Python is an interpreted high-level programming language for general-purpose programming.
- Created by Guido van Rossum and first released in 1991
- Meant to be simple for non-programmers.
- Can be downloaded from https://www.python.org/

Python

- Python 3.7 is the latest version
- A lot of libraries available for various functionalities
- PyCharm – pretty good IDE for python (can be downloaded from https://www.jetbrains.com/pycharm/)

Python Libraries

Core Libraries	Visualization	Machine Learning	Statistics
NumPy	Matplotlib	Scikit-Learn (sklearn)	Statsmodels
SciPy	Seaborn		
Pandas	Bokeh		
	Plotly		

Python Libraries

- NumPy – fundamental package for scientific computing
  - a powerful homogeneous N-dimensional array object
  - sophisticated (broadcasting) functions
  - tools for integrating C/C++ and Fortran code
  - useful linear algebra, Fourier transform, and random number capabilities
- SciPy – library of software for engineering and science
  - SciPy contains modules for linear algebra, optimization, integration, and statistics.
  - The main functionality of SciPy library is built upon NumPy, and thus, its arrays make substantial use of NumPy.

Python Libraries

- Pandas - open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools
  - Two main data structures – Series & DataFrames
  - Easily add and delete columns from a DataFrame
  - Convert data structures to DataFrame objects
  - Handle missing data, represented as NaNs
  - Group by functionality
- Scikit-Learn - a concise and consistent interface to common machine learning algorithms, making it simple to bring ML into production systems.
  - combines quality code and good documentation, ease of use and high performance, and is de-facto industry standard for machine learning with Python.

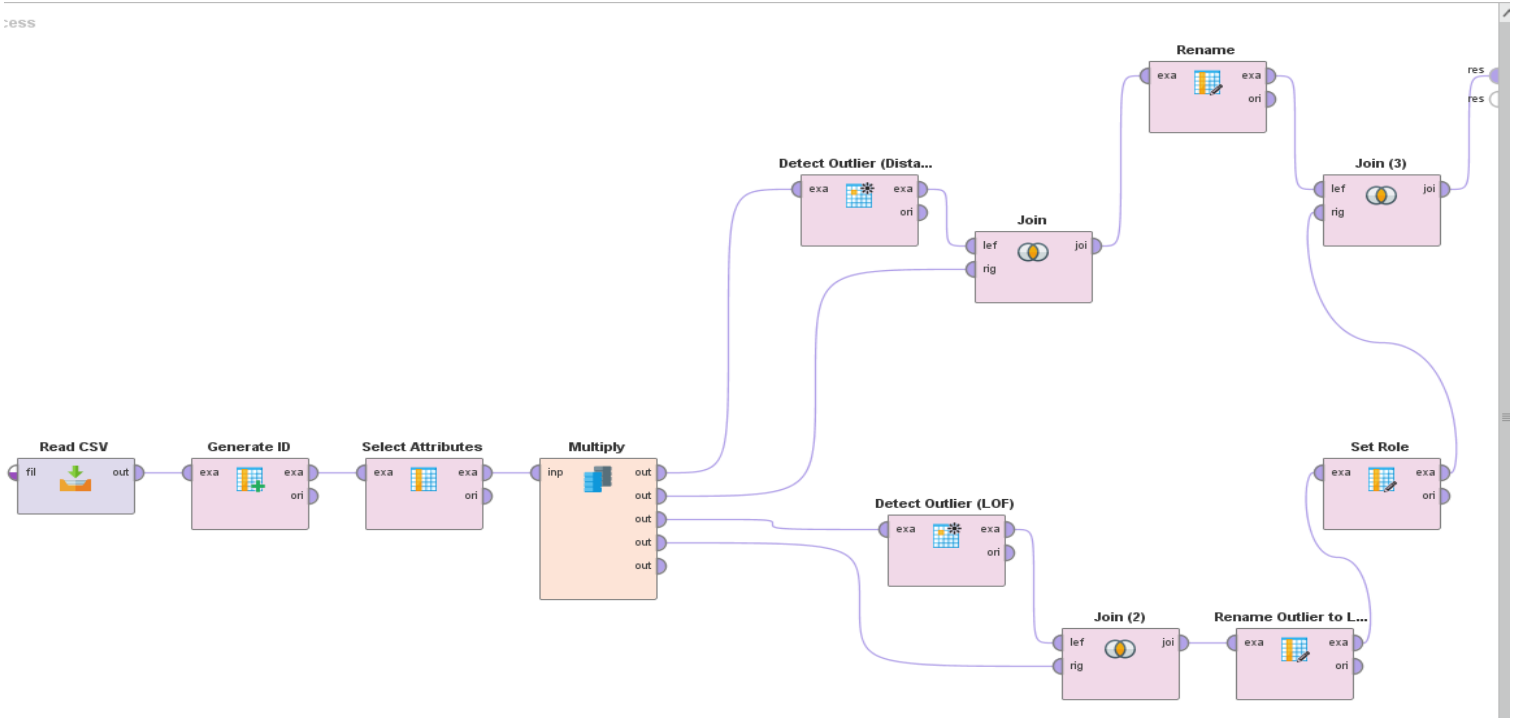
Python Demo

RapidMiner

- data science software platform that provides an integrated environment for
  - data preparation
  - machine learning
  - deep learning
  - text mining
  - predictive analytics

Can download from https://rapidminer.com/get-started/

RapidMiner Demo



References

- https://www.datascienceweekly.org/articles/what-tools-do-employers-want-data-scientists-to-know
- https://www.analyticsvidhya.com/blog/2018/05/19-data-science-tools-for-people-dont-understand-coding/
- https://www.kdnuggets.com/tag/data-science-tools
- https://www.python.org/
- https://www.jetbrains.com/pycharm/
- https://www.datascience.com/blog/top-python-libraries-for-data-science-in-2017

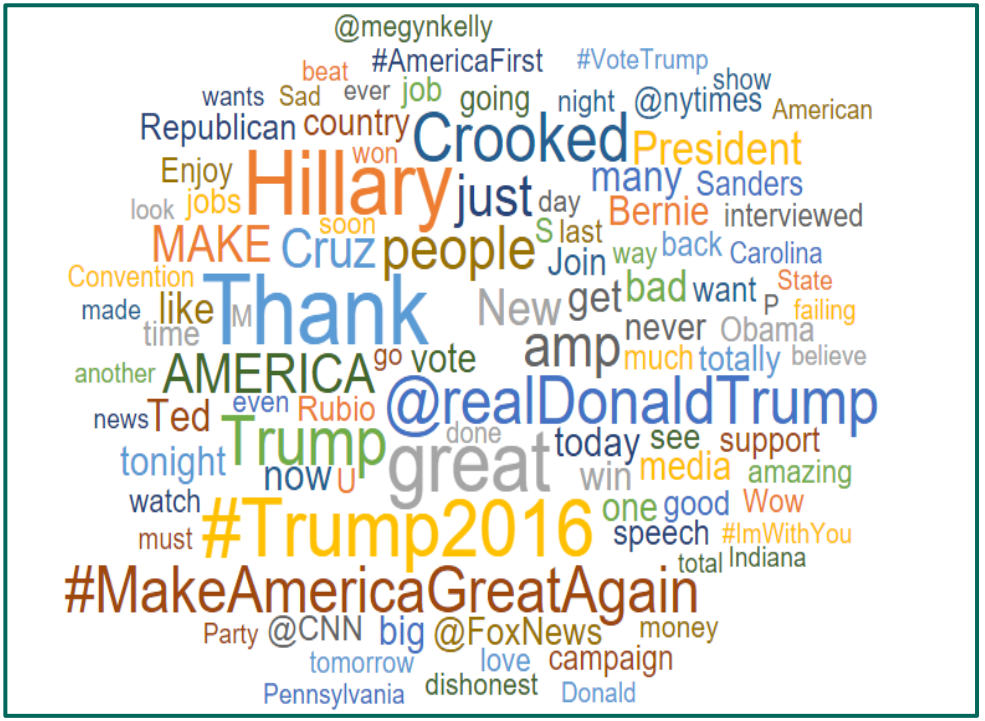
ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

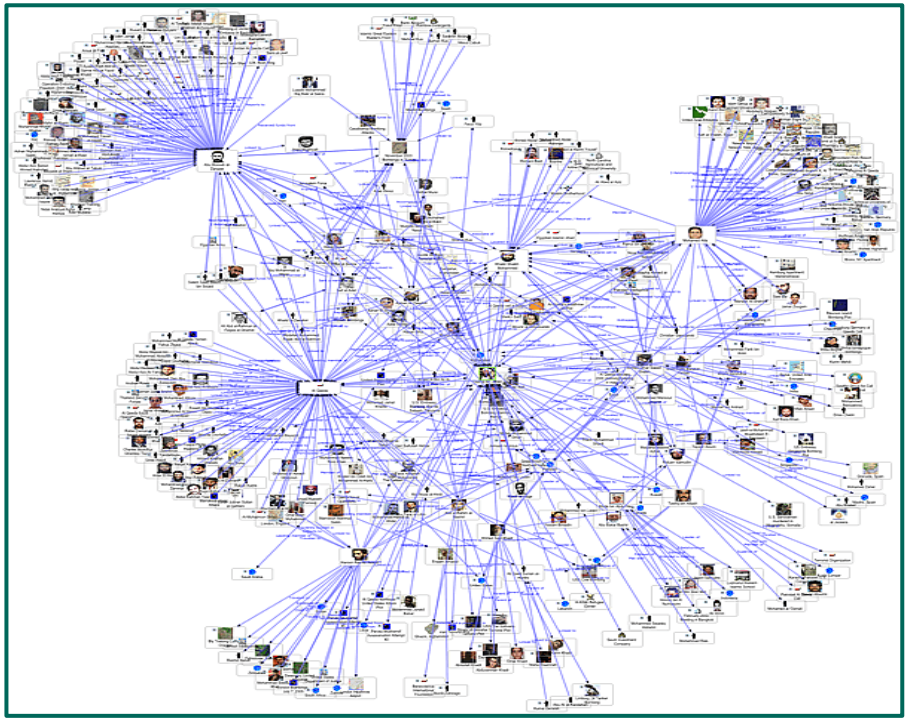
Week 12

Trending Topics in Industry

Sentiment Analysis



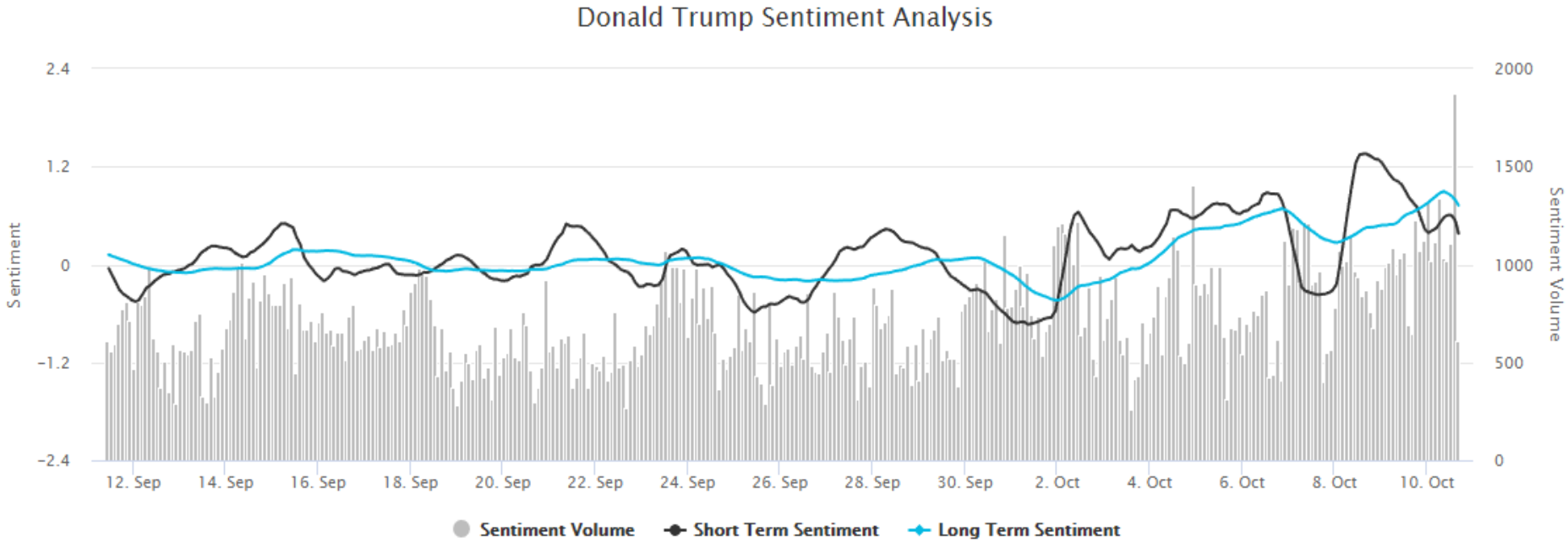
Link Analysis



Sentiment Analysis

- Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.
- Comes under Natural Language Processing (NLP)

Example - Politicians



Taken from http://sentdex.com/political-analysis/us-politicians/

Other names for Sentiment Analysis

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

Examples of Sentiment Analysis

- Movie: is this review positive or negative?
- Products: what do people think about the new iPhone?
- Public sentiment: how is consumer confidence?
- Politics: what do people think about this candidate or issue?
- Prediction: predict election outcomes or market trends from sentiment

Data Sources - Examples

- Review sites
- Blogs
- News
- Social media

Affective states

- Emotion: happy, sad, angry, proud, ashamed etc.
- Mood: cheerful, gloomy, irritable, depressed etc.
- Interpersonal stances: friendly, warm, supportive etc.
- Attitudes: liking, loving, hating, desiring etc.
- Personality traits: nervous, anxious, jealous, hostile etc

Sentiment Analysis

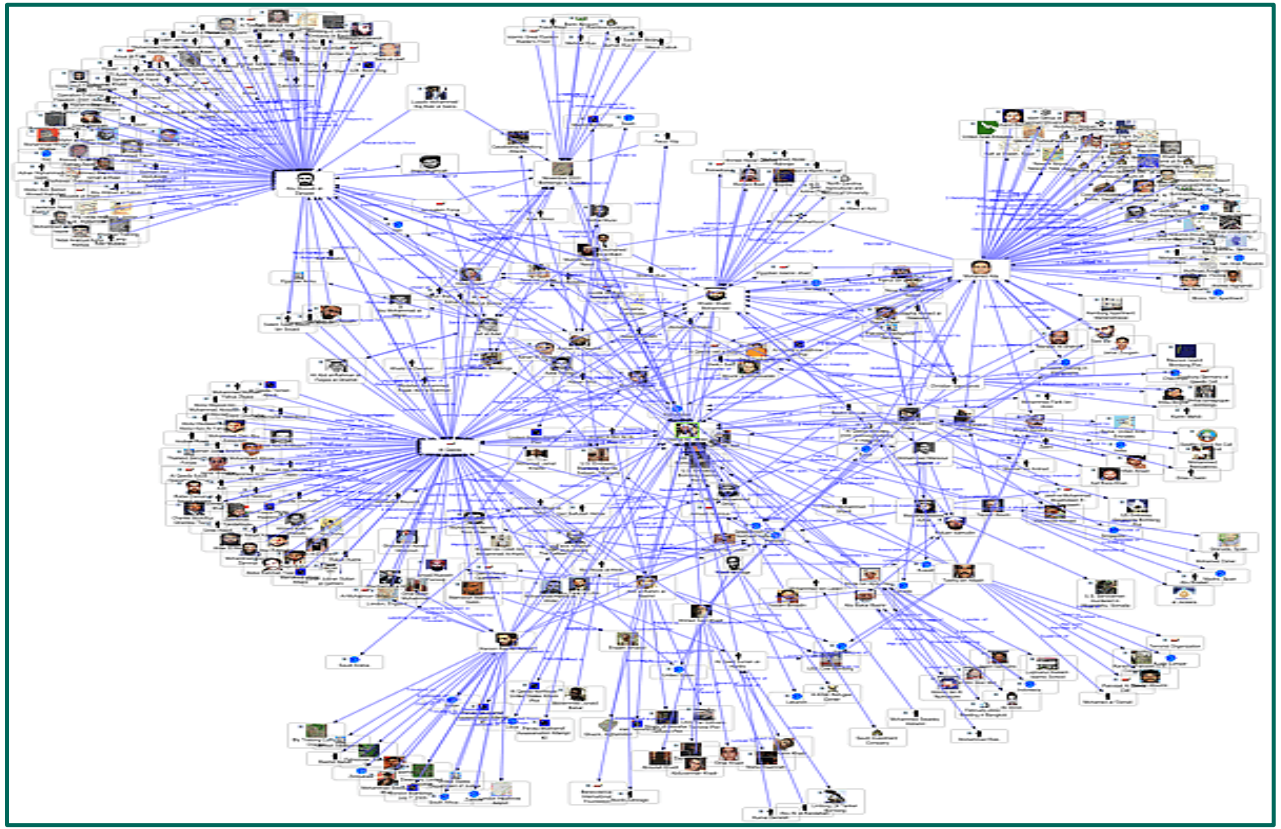
- Is the detection of attitudes
  1. Holder (source) of attitude
  2. Target (aspect) of attitude
  3. Type of attitude
  4. Text containing the attitude
- Positive, negative, or neutral together with the strength
- Task is to identify whether the text is positive or negative

Analysis

- Machine learning
  - Supervised
  - Unsupervised
- Lexicon-based
  - Dictionary
- Discourse analysis

Demo

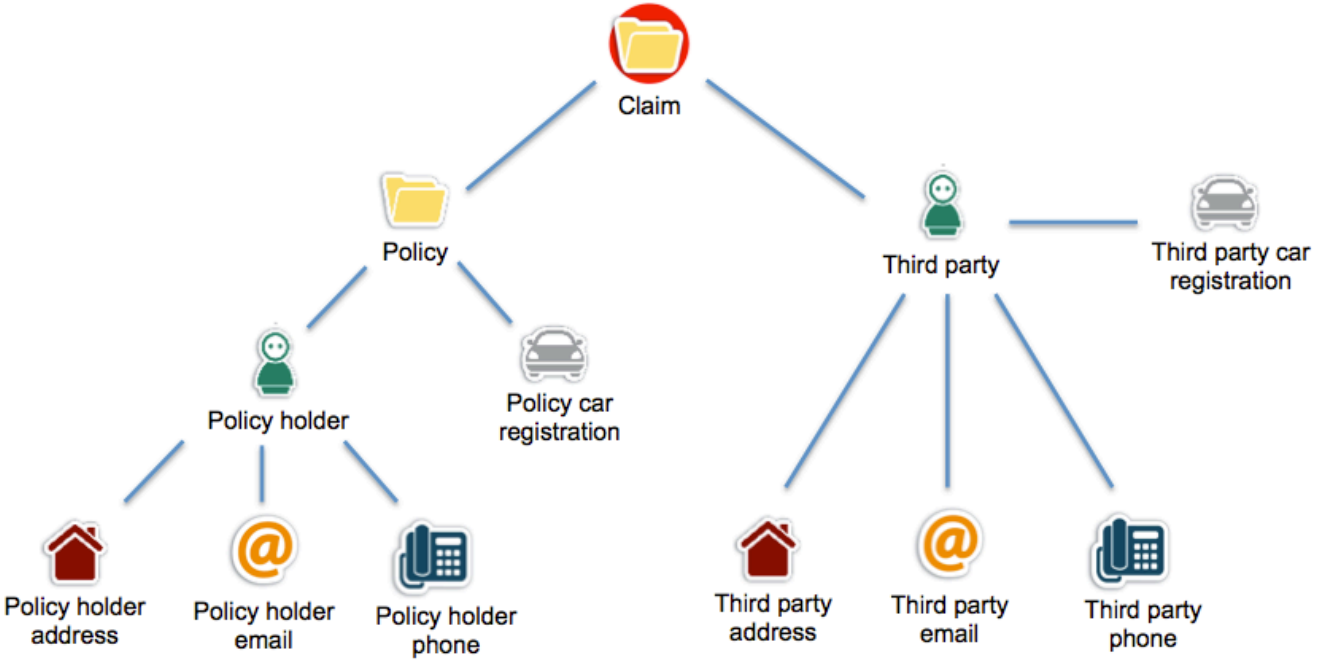
Link Analysis



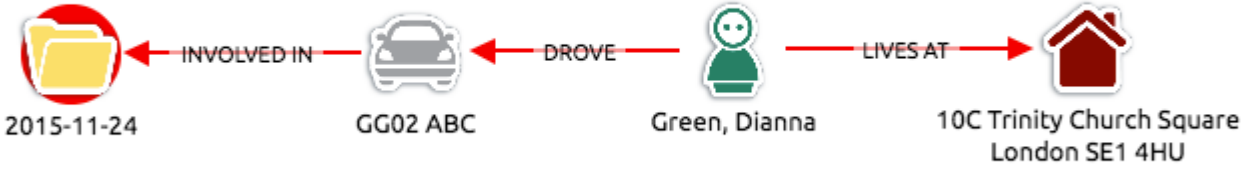
Link Analysis

- Data-analysis technique used to evaluate relationships (connections) between nodes.
- Relationships may be identified among various types of nodes (objects), including organizations, people and transactions.
- Used for investigation of criminal activity (fraud detection, counterterrorism, and intelligence), computer security analysis, search engine optimization, market research, medical research, and art.

Example – Investigation on Insurance Claims



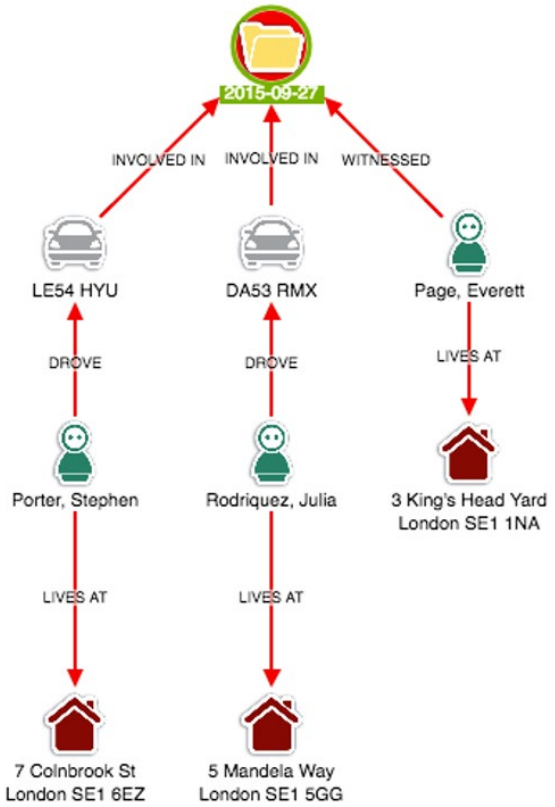
Example: Visual Data Model



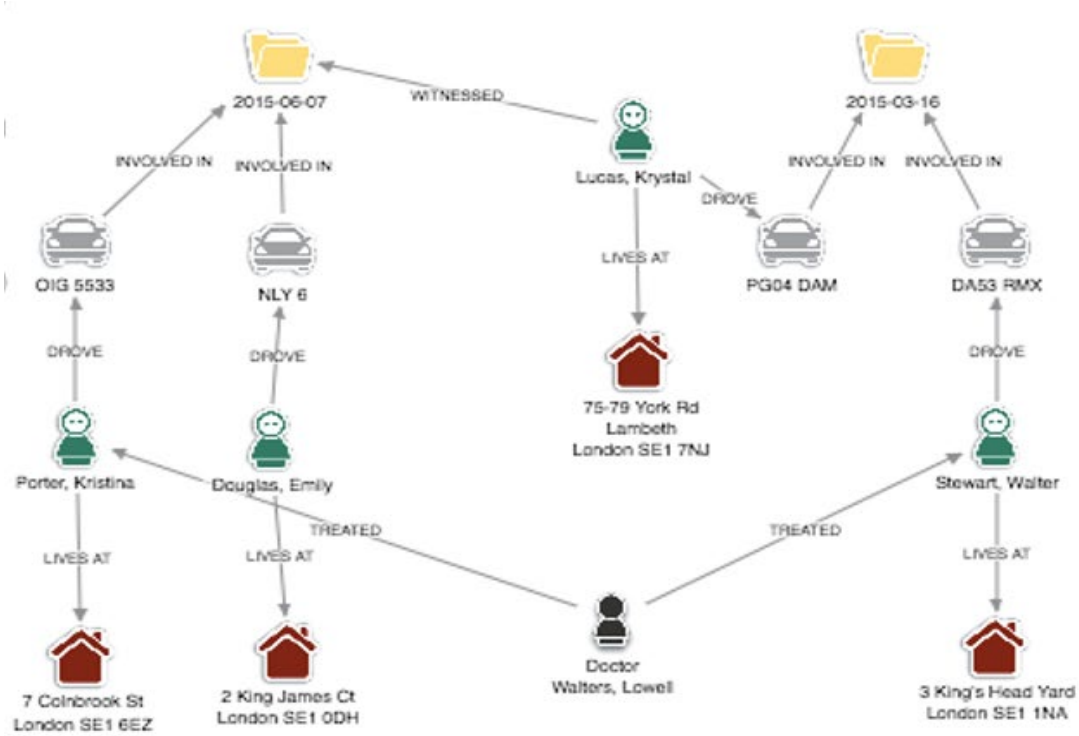
- Claim – being investigated
- Vehicle – involved in the claim
- Claimant – associated with the vehicle
- Address – at which the claimant lives

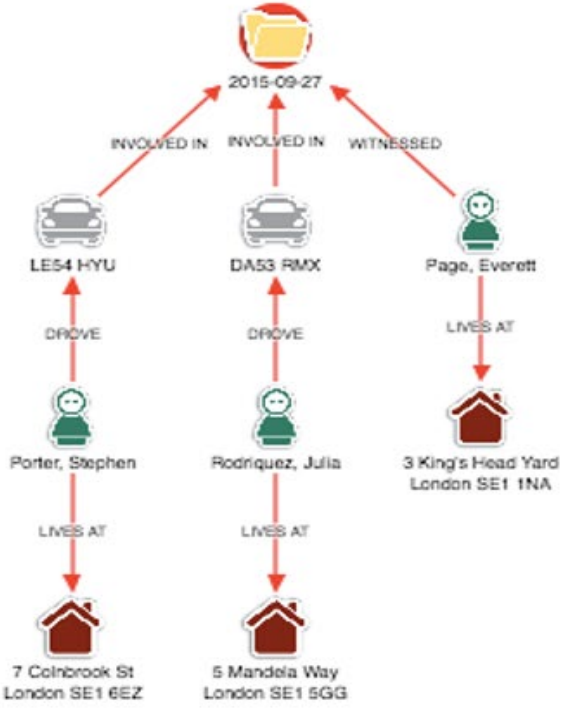
1. Load a claim

Involves two vehicles and three claimants, associated with three separate addresses



2. Find matches





Anything suspicious???

3. Combine Matches

References

- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
- https://monkeylearn.com/sentiment-analysis/
- https://en.wikipedia.org/wiki/Link_analysis
- https://cambridge-intelligence.com/link-analysis-fraud-detection/