

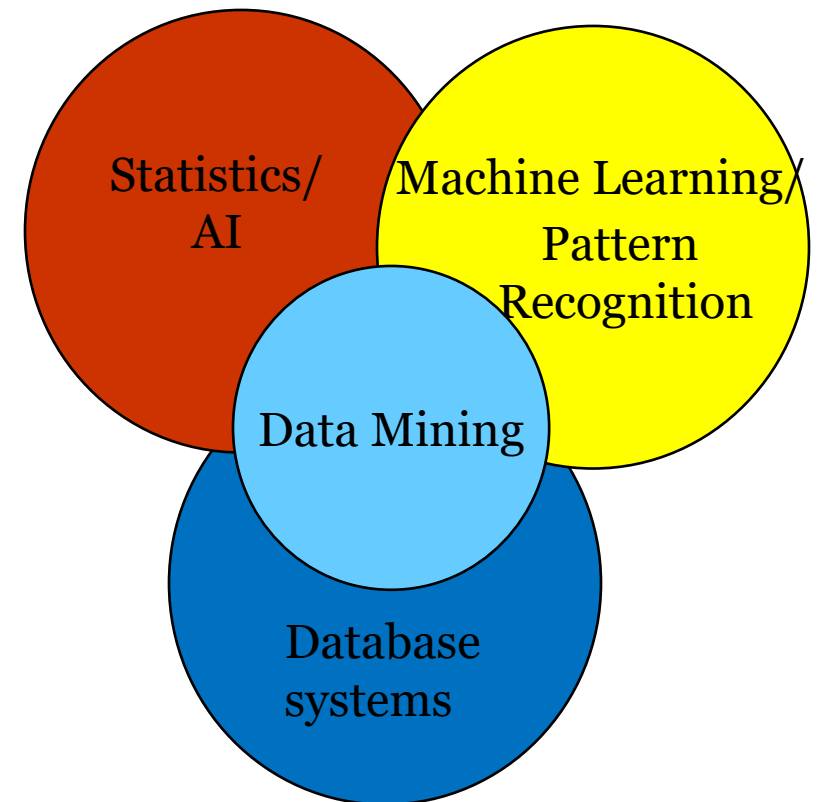
CISA 4313 Programming for Data Analytics

Review of Data Mining
Concepts and Models

Dr. Mohammad Abdel-Rahman

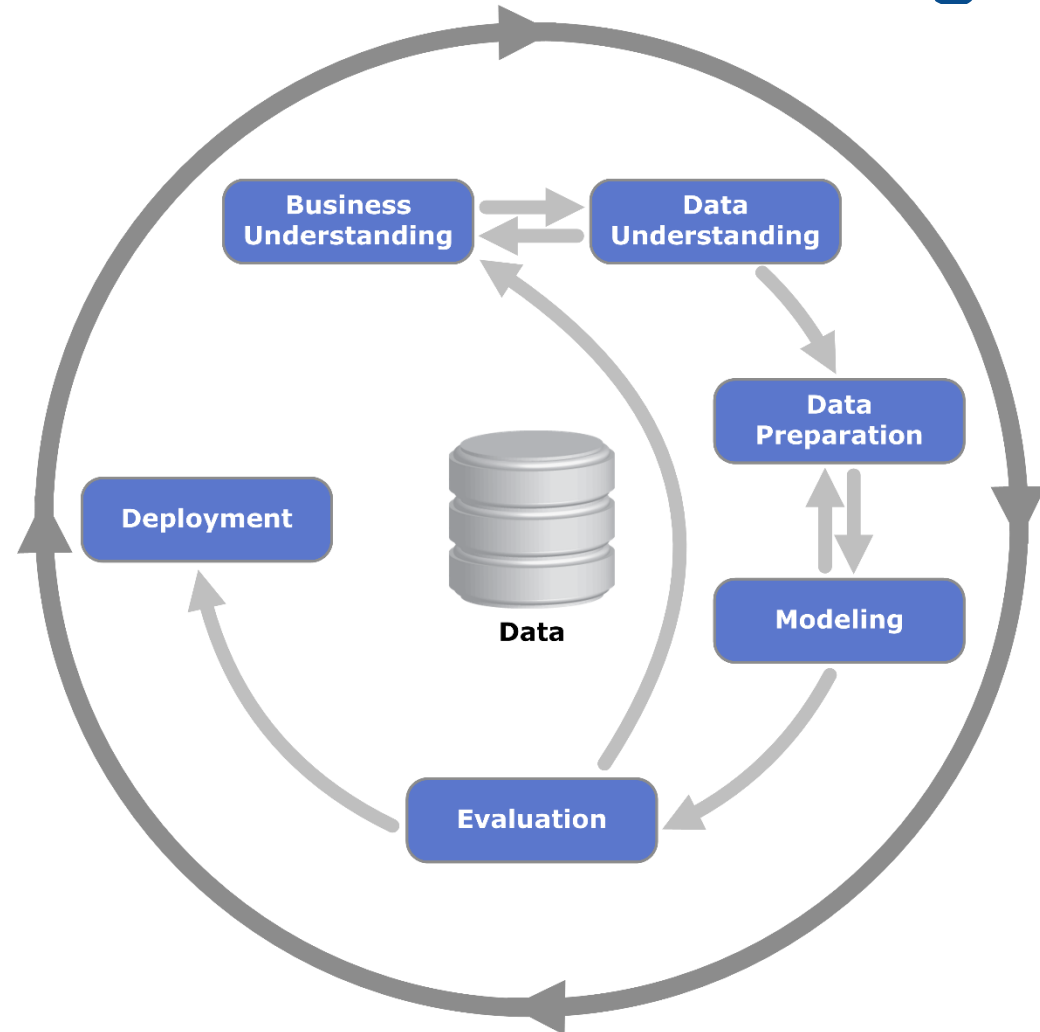
What is Data Mining?

- Discovering meaningful and non-trivial patterns/knowledge from data.
- Data mining vs. data science
 - [Google trends](#)



Cross-Industry Standard Process for Data Mining (CRISP-DM)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



[Kdnuggets poll](#)

Business Understanding

- Map a business problem to one or many data analysis tasks.
 - Determine the project objective and success criteria.
 - Assess the situation
 - Review of available resources, requirements, assumptions, and risks
 - Determine analysis goals
 - Translate the primary objective to technical data mining goals.
 - Model properties:
 - Interpretability
 - Robustness
 - Flexibility
 - Run time

Business Understanding: German Credit Data

- Primary objective:
 - Minimize credit risks by accurately determine whether a loan application should be approved based on the applicant's personal information.
- The situation:
 - Available resources: historical data.
 - Assumption: no significant change in the economic environment and the demographics.
- Data analysis goal:
 - Build a robust and interpretable model that can classify lending decision in reasonable time.

Data Understanding: Problems

- Gain general insights about the data that will be potentially helpful for further analysis.
 - Attribute types
 - Missing values
 - Sparsity
 - Accuracy
 - Heterogeneity
 - Outliers
 - Size
 - Granularity
 - Balance of classes
 - Timeliness

Data Understanding: Techniques

- Descriptive Statistics:
 - Mean, median, mode, variance, percentiles
 - Correlation
 - Frequency table
- Visualization
 - Box plot
 - Histogram
 - Scatter plot
 - Line plot

Data Preparation: 80% of the Efforts

- Transformation
 - Renaming
 - Derive new variables
 - Grouping
 - Partition
 - Sampling
- Missing values
 - Missing at random
 - Ignore/delete records with missing values
 - Imputation
 - Missing not at random
 - Create the Missing variable
 - Recollect data
- Tools:
 - SQL, R, Python

Modeling

Data Mining Concepts

Basic Concepts

- Instance, observation, record, row
- Attribute, feature, variable, column
- Training data, validating data, testing data
- Structured data, unstructured data, semi-structured data

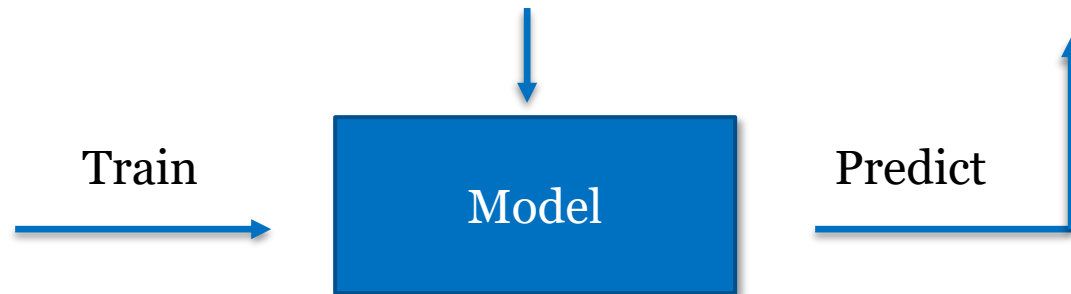
Supervised Learning

Training data: attributes + known outcomes

Credit History	Annual Income	Credit Amount	Approve
Good	50,000	12,000	No
Bad	60,000	5,000	No
Excellent	45,000	10,000	Yes
Good	85,000	15,000	Yes
Good	13,000	6,000	Yes
Bad	65,000	3,000	No
Good	74,000	14,000	Yes

Testing data: attributes + unknown outcomes

Credit History	Annual Income	Credit Amount	Approve
Good	70,000	22,000	?
Good	80,000	15,000	?
Bad	45,000	10,000	?
Bad	95,000	5,000	?



Unsupervised Learning

Credit History	Annual Income	Credit Amount
Good	50,000	12,000
Bad	60,000	5,000
Excellent	45,000	10,000
Good	85,000	15,000
Good	13,000	6,000
Bad	65,000	3,000
Good	74,000	14,000

Group similar customers



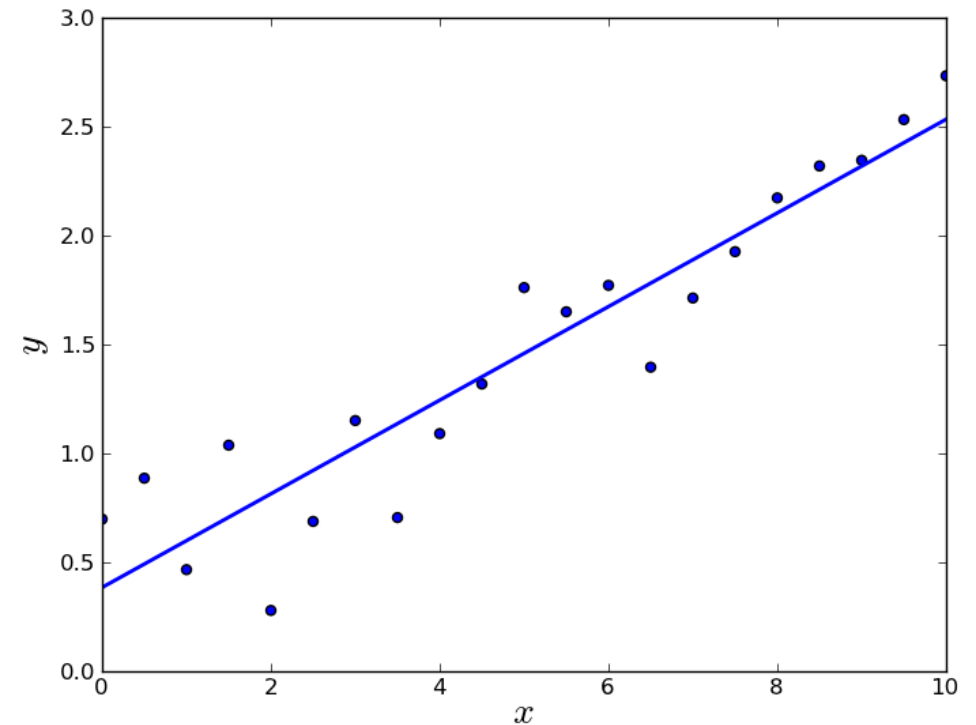
Credit History	Annual Income	Credit Amount
Good	50,000	12,000
Bad	60,000	5,000
Excellent	45,000	10,000
Good	85,000	15,000
Good	13,000	6,000
Bad	65,000	3,000
Good	74,000	14,000

Common Data Mining Tasks

- Regression
- Classification
- Clustering
- Association rule mining

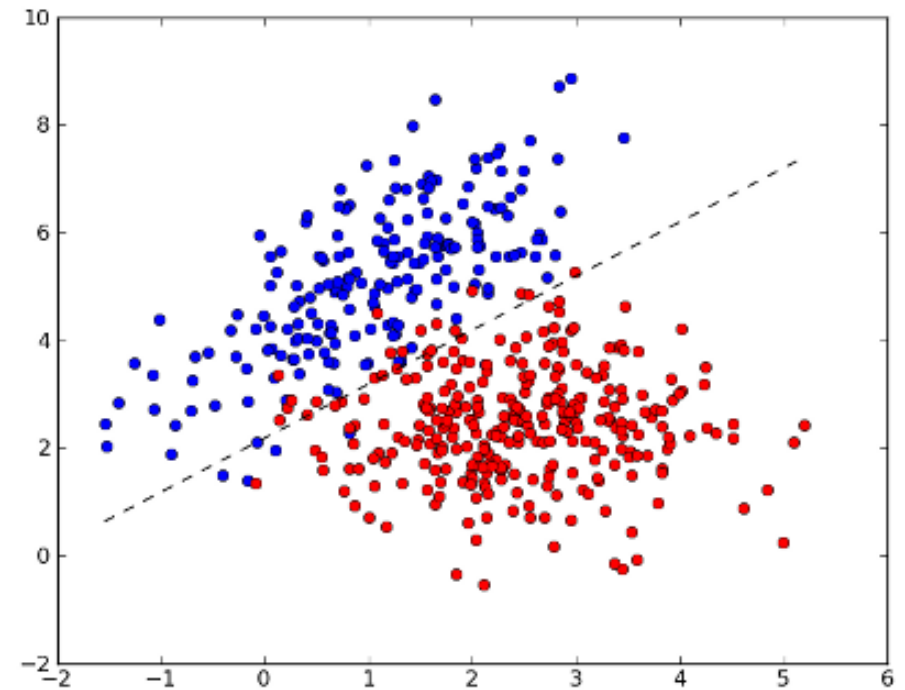
Regression

- The variable to be predicted is continuous/numerical.
- Supervised learning
- Examples:
 - Finance: predict stock price based on dividends ratio
 - Economics: predict unemployment rate based on national consumption
 - Marketing: predict sales based on advertising expense



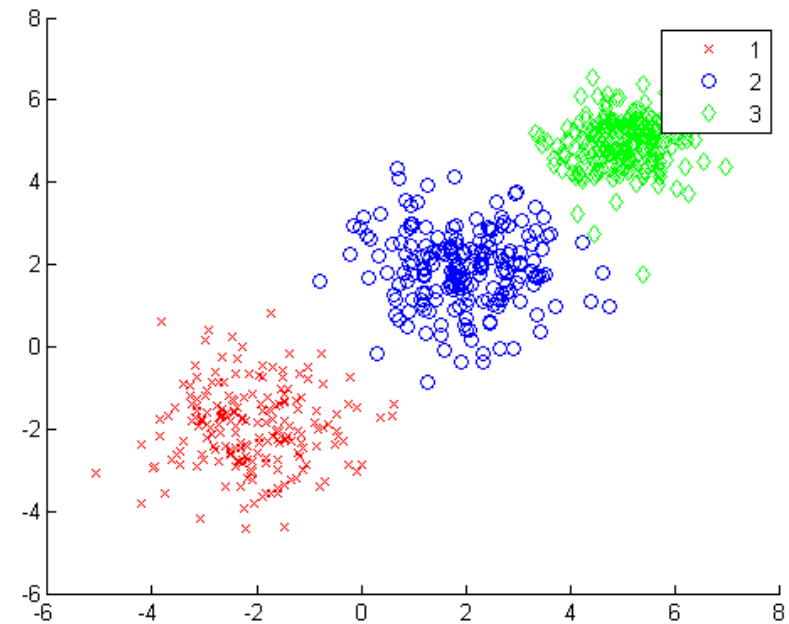
Classification

- The variable to be predicted is categorical/nominal/discrete.
- Supervised learning
- Examples:
 - Direct mail marketing: if you send a promotion mail to a customer, will the customer make a purchase?
 - Spam filter: given an incoming email, is it a spam?



Clustering

- The outcome variable is not designated or unknown.
- Unsupervised learning
- Instances in ones cluster are more similar to one another
- Examples:
 - Market segmentation: customers are grouped into subsets which are to be targeted with different marketing mix.



Association Rule Mining

- Finding frequent patterns of attributes
- Example:
 - Market basket analysis: are mice and keyboards often bought together?

Transaction ID	Items Bought
1	Monitor, Mouse
2	Mouse , Keyboard , Speaker
3	Monitor, Mouse , Keyboard
4	Keyboard , HDMI Cable

Modeling

Commonly Used Models

Linear Regression

- The outcome variable (continuous) is predicted using the linear combination of other attributes

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Logistic Regression

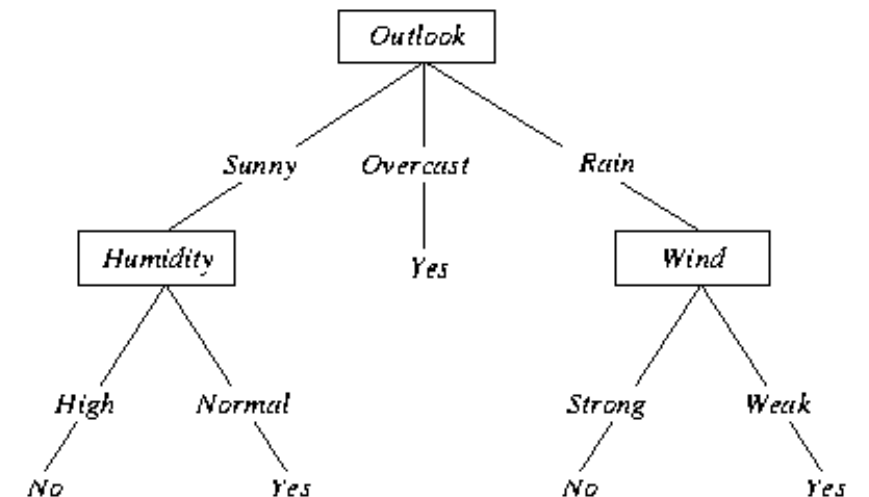
- The logit (log odds) of the outcome variable (discrete) is predicted by the linear combination of other attributes

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- Multinomial Logistic Regression – more than two discrete outcomes
- Ordinal Logistic Regression – ordinal outcomes

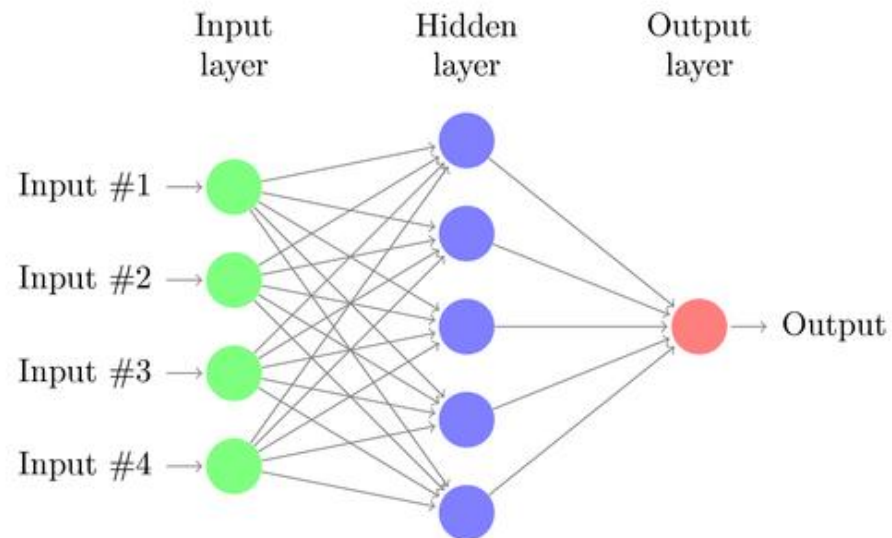
Decision Tree

- The outcome variable (discrete) is predicted using other discrete attributes.
- Continuous attributes need to be discretized.
- C4.5 uses information gain
- CHAID uses Chi-square test
- Random forests



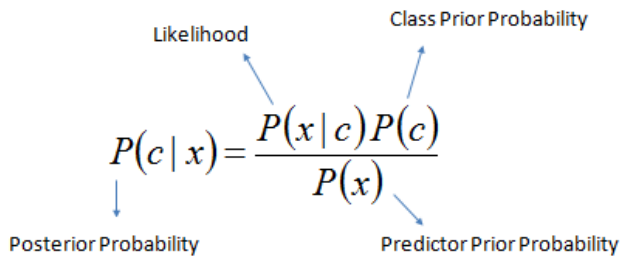
Neural Network

- Inspired by human central nervous system.
- Overfitting, black-box
- Deep learning: model high-level abstractions in data



Naïve Bayes

- The Bayes theorem
- Why naïve? The independence assumption.
- Violation of the independence assumption may yield to incorrect probability estimates by does not affect the classification as long as the violation is evenly distributed across classes.



A diagram showing the Bayes' theorem formula $P(c | x) = \frac{P(x | c)P(c)}{P(x)}$ with blue arrows pointing from labels to the corresponding parts of the formula. The label 'Likelihood' points to $P(x | c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c | x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

Likelihood

Class Prior Probability

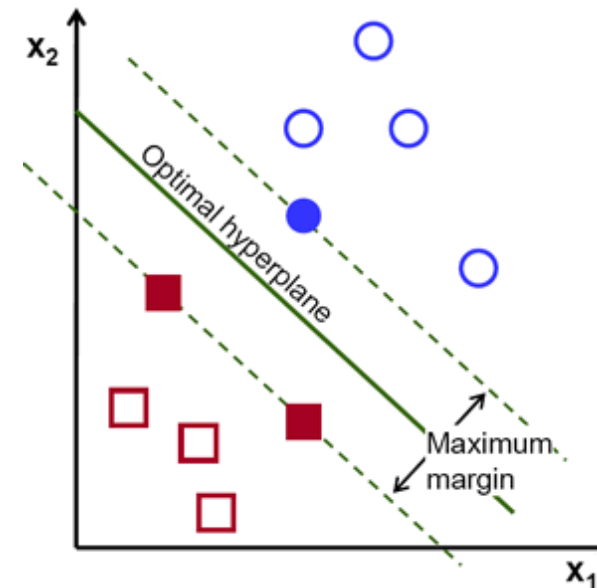
Posterior Probability

Predictor Prior Probability

$$P(c | \mathbf{X}) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)$$

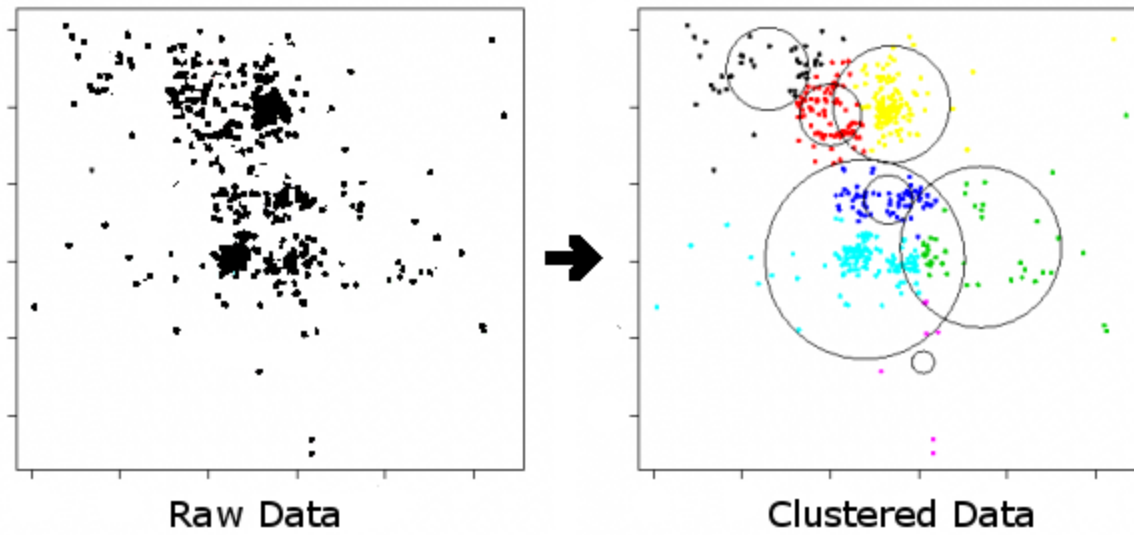
Support Vector Machine

- Search for the hyperplane that separates two classes with the maximum margin.
- Robust to high-dimensional data.
- Kernel trick.



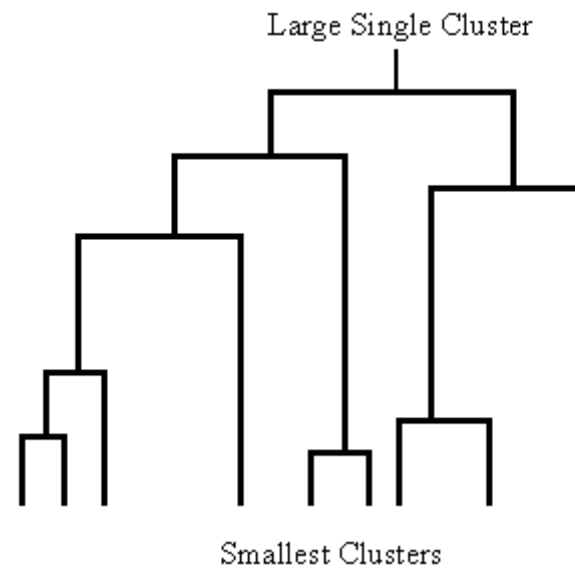
K-Means Clustering

- Iteratively search for the center of each cluster.



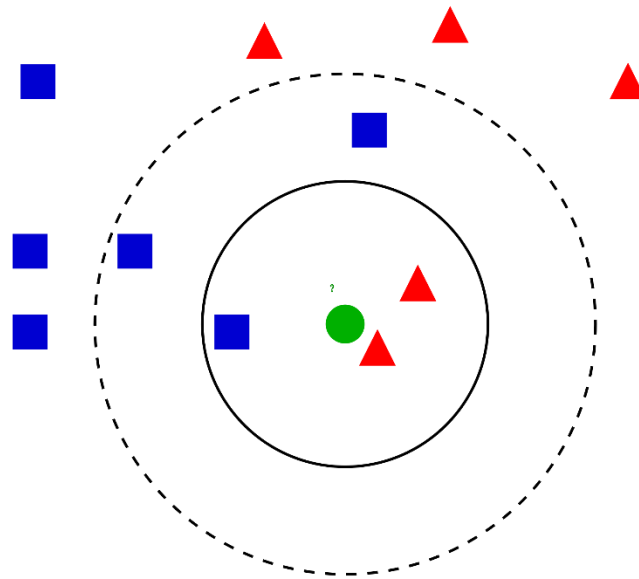
Hierarchical Clustering

- A clustering approach which builds a hierarchy of clusters.



Nearest Neighbor

- Lazy learning, instance-based learning
- A few closest training instances are used to predict the outcome of a given instance.



No Single Best Model

- How well a model suits the business needs:
 - Prediction correctness
 - Interpretability
 - Run time