Data Mining

An Introduction

Terri Hoare - September 2018

Learning Objectives

- Define Data Mining
- Explore data mining as an enabling technology for business intelligence
- Understand the objectives and benefits of business analytics and data mining
- Recognize the wide range of applications of data mining
- Learn the standardized data mining processes
  - CRISP-DM

(Continued…)

Learning Objectives

- Understand the steps involved in data preprocessing for data mining
- Learn different methods and algorithms of data mining
- Build awareness of the existing data mining software tools
  - Commercial versus free/open source
- Understand the pitfalls and myths of data mining

Data Mining Concepts and Definitions. Why Data Mining?

- Exponential increase of data available
- Increase in computing power and data storage capacity at reduced cost
- Advancement in machine learning methods to analyse complex datasets
- More intense competition at the global scale
- Recognition of the value in data sources
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses.
- Movement toward conversion of information resources into nonphysical form.

Definition of Data Mining

- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases. *Fayyad et al., (1996)*

- **Keywords in this definition**: **Process, nontrivial, valid, novel, potentially useful, understandable, patterns**.

- Data mining: a misnomer?

- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,…

Definition of Data Mining cont.

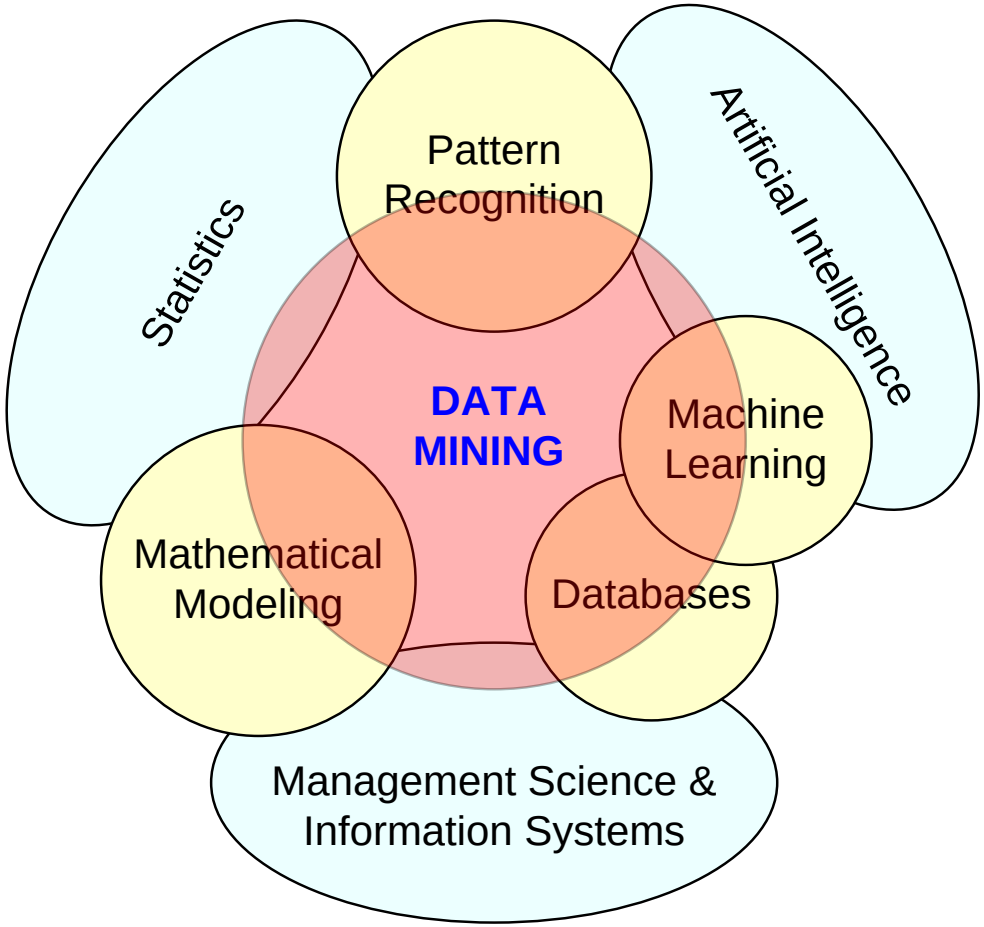
- "…hybrid of artificial intelligence, statistics, database research, and machine learning. The actual process entails the automatic or semi-automatic analysis of large datasets to extract previously unknown yet interesting patterns, anomalies or dependencies that could be exploited."

Bernard Marr (Data Strategy - How to Profit from a World of Big Data, Analytics and the Internet of Things - 2017)

Definition of Data Mining cont.

Manual extraction of patterns from data has occurred for centuries. Bayes Theorem (1700s) and Regression Analysis (1800s). Inncreasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s) and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets. *https://en.wikipedia.org/wiki/Data_mining*

Data Mining – At the Intersection of Many Disciplines

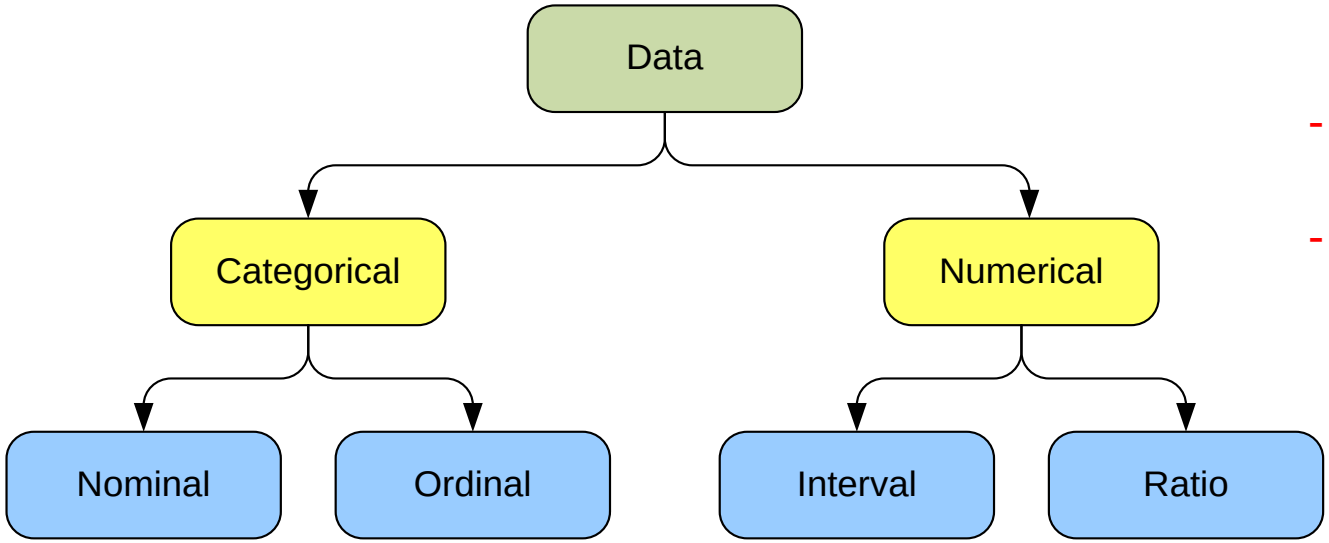


Data Mining Characteristics / Objectives

- Source of data for DM can be but is not always the consolidated data warehouse
- DM environment is usually a client-server or a Web-based information systems architecture.
- Data is the most critical ingredient for DM which may include soft/unstructured data.
- The miner is often an end user.
- Striking it rich requires creative thinking.
- Data mining tools' capabilities and ease of use are essential (Web, Parallel processing, etc.).

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments.
- Data may consist of numbers, words, images, …
- Data: lowest level of abstraction (from which information and knowledge are derived).



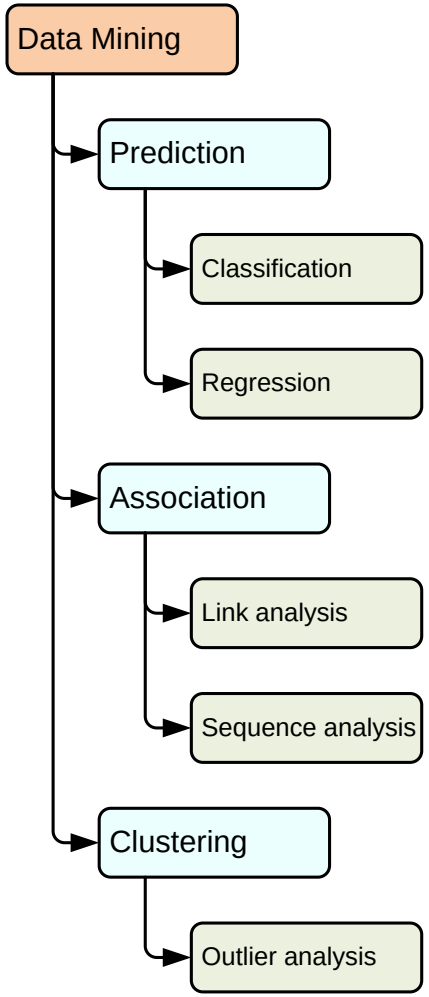
- DM with different data types?
- Other data types?

What Does DM Do? How Does it Work?

- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items

- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

A Taxonomy for Data Mining Tasks



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Data Mining Tasks (cont.)

- Feature selection
- Text Mining
- Anomaly detection
- Time-series forecasting
- Visualization
- Types of DM
  - Hypothesis-driven data mining (predictive models)
  - Discovery-driven data mining (descriptive patterns)

Data Mining Applications

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross-, up-selling)
  - Identify and treat most valued customers

- Banking & Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Maximize customer value (cross-, up-selling)
  - Optimizing cash reserves with forecasting

Data Mining Applications (cont.)

- Retailing and Logistics
  - Optimize inventory levels at different locations
  - Improve the store layout and sales promotions
  - Optimize logistics by predicting seasonal effects
  - Minimize losses due to limited shelf life

- Manufacturing and Maintenance
  - Predict/prevent machinery failures
  - Identify anomalies in production systems to optimize the use manufacturing capacity
  - Discover novel patterns to improve product quality

Data Mining Applications (cont.)

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading

- Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

Data Mining Applications (cont.)

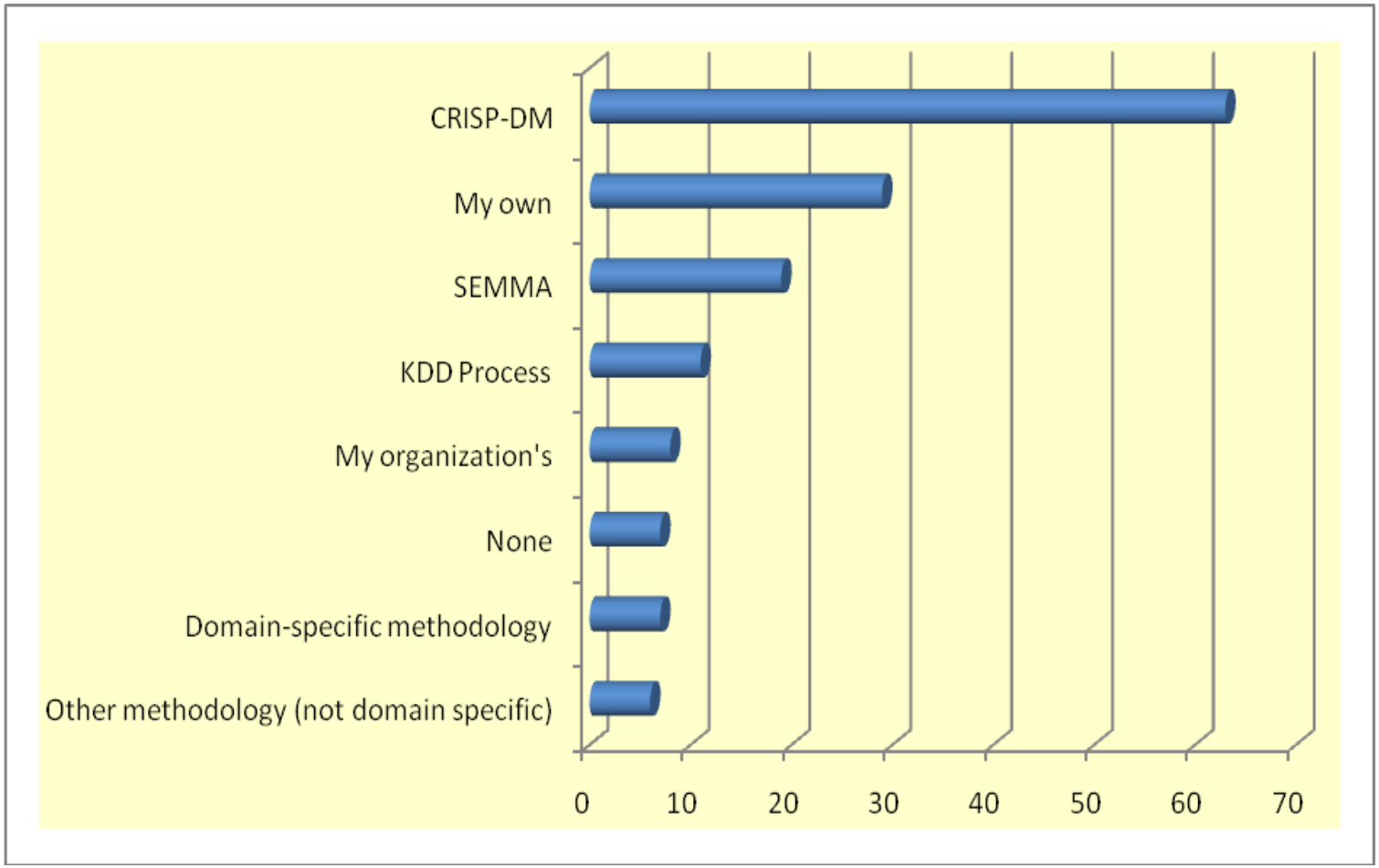
- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine
- Entertainment industry
- Sports
- Etc.

Very popular application areas for data mining

Data Mining Process - Methodology

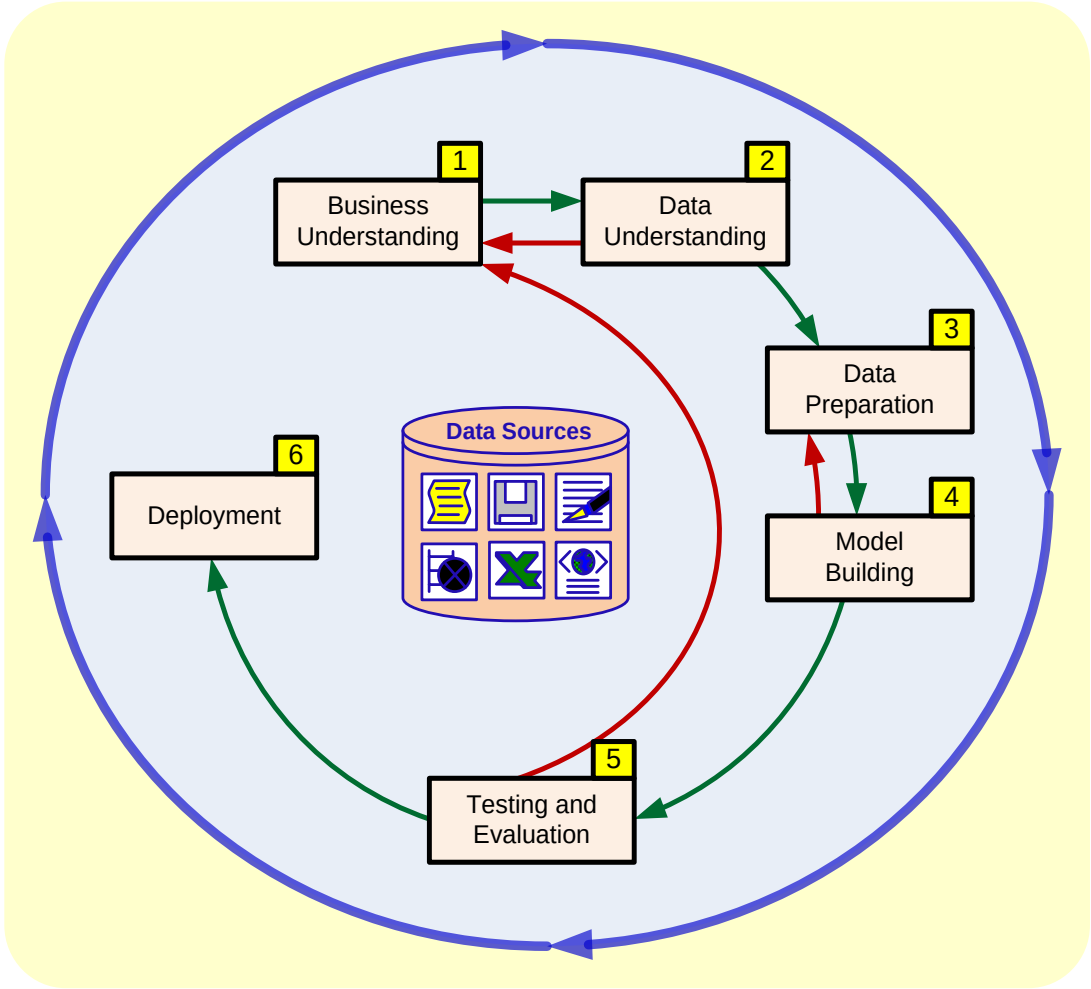
- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Data Mining Process



Source: KDNuggets.com

Data Mining Process: CRISP-DM



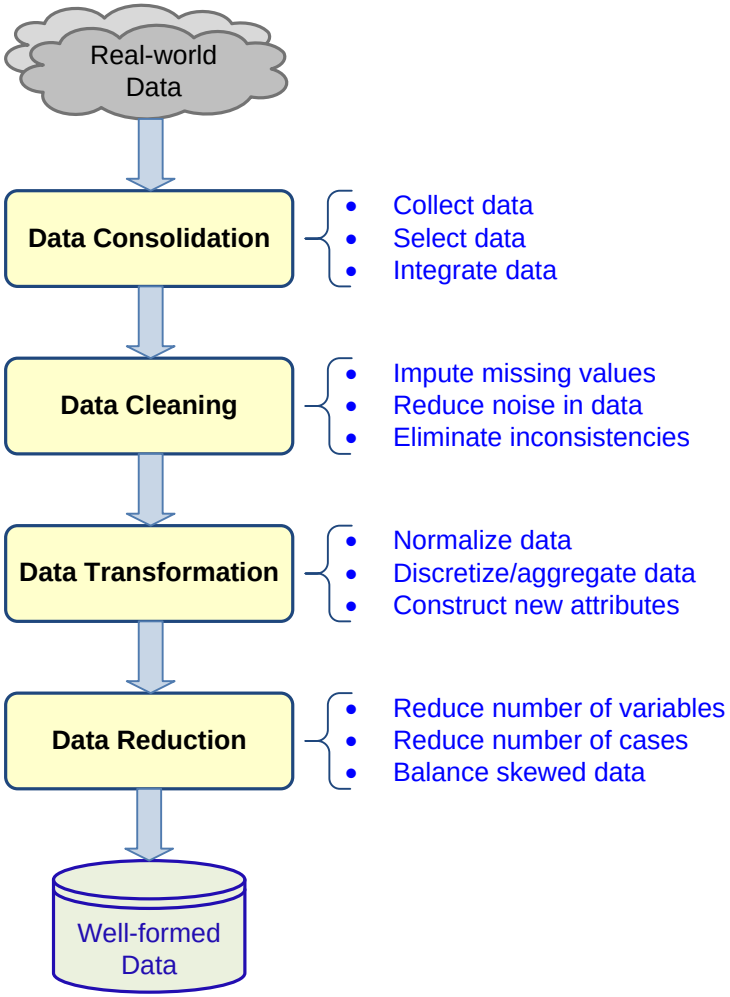
Data Mining Process: CRISP-DM

Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

Accounts for ~85% of total project time

- The process is highly repetitive and experimental (DM: Art versus Science?)

Data Preparation – A Critical DM Task



Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- **Classification versus regression?**
- **Classification versus clustering?**

Assessment Methods for Classification

- Predictive accuracy
  - Hit rate
- Speed
  - Model building; predicting
- Robustness
  - producing a result that takes the uncertainty or potentially corrupted data (outliers) into account
- Scalability
  - Ability to handle increased data volumes, hardware additions
- Interpretability
  - Transparency, explainability

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the **confusion matrix**

		True Class: Positive	True Class: Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
Predicted Class	Negative	False Negative Count (FN)	True Negative Count (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

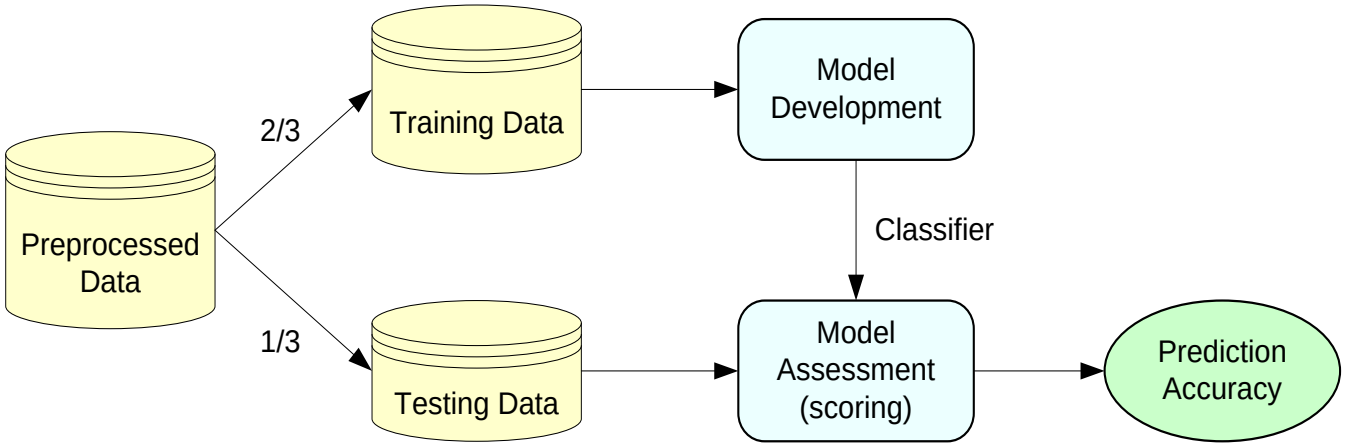
$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)



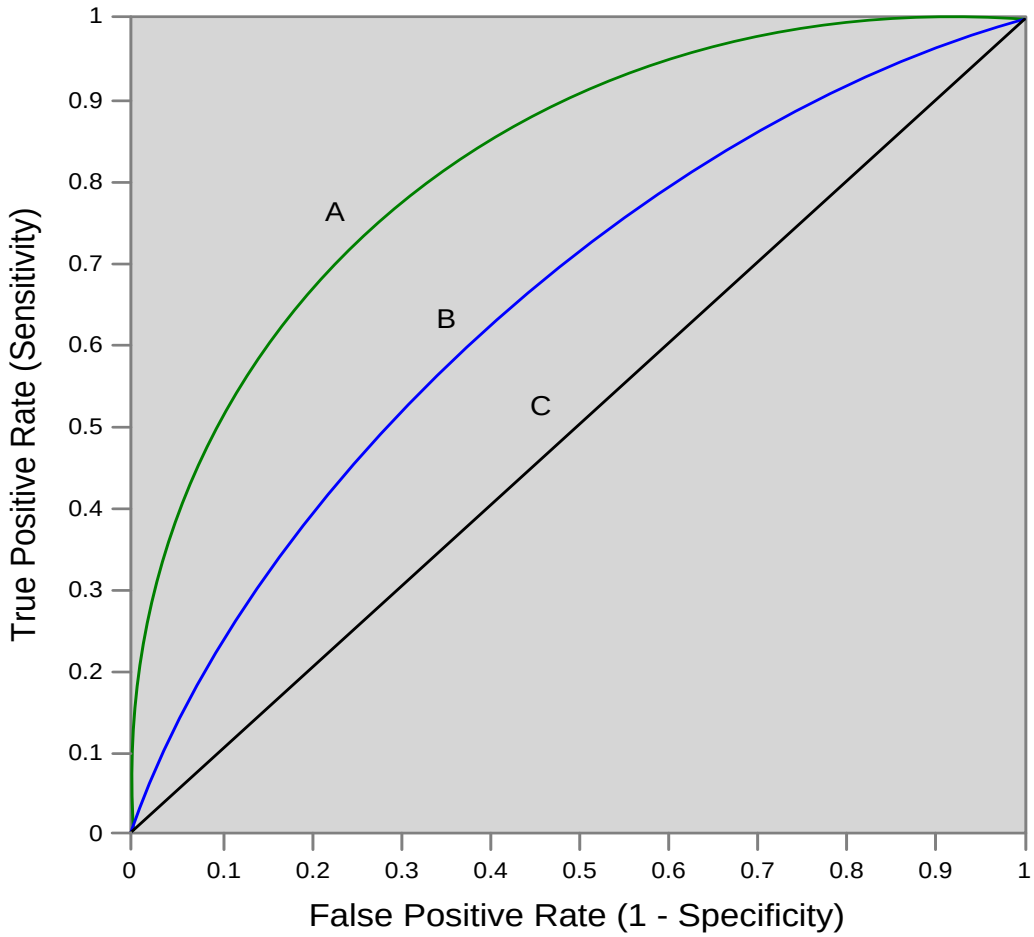
  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

Estimation Methodologies for Classification (contd.)

- ***k*-Fold Cross Validation** (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training

- Other estimation methodologies
  - **Leave-one-out, bootstrapping, jackknifing**
  - **Area under the ROC curve**

Estimation Methodologies for Classification – ROC Curve



Classification Techniques

- Decision tree analysis
- Neural networks
- Support vector machines
- Bayesian classifiers
- Genetic algorithms

  (Based on natural selection, the process that drives biological evolution. The **genetic algorithm** repeatedly modifies a population of individual solutions).

Decision Trees

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class

A general algorithm for decision tree building

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

Decision Trees

- DT algorithms mainly differ on
  1. **Splitting criteria**
     - Which variable, what value, etc.
  2. **Stopping criteria**
     - When to stop building the tree
  3. **Pruning (generalization method)**
     - Pre-pruning versus post-pruning
- Most popular DT algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

Decision Trees

- Alternative splitting criteria
  - **Gini index** determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    - Used in CART
  - **Information gain** uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    - Used in ID3, C4.5, C5
  - **Chi-square statistics** (used in CHAID)

Cluster Analysis for Data Mining

- Used for automatic identification of natural groups in data
- Part of the machine-learning family
- Employs unsupervised learning
- Also known as segmentation

Cluster Analysis for Data Mining

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Cluster Analysis for Data Mining

- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on.
  - Neural networks (for example self-organizing map [SOM])
  - Genetic (evolutionary) algorithms
  - Fuzzy logic (e.g., fuzzy c-means algorithm)

    (A form of many-valued logic in which the truth values of variables may be any real number between 0 and 1. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false)

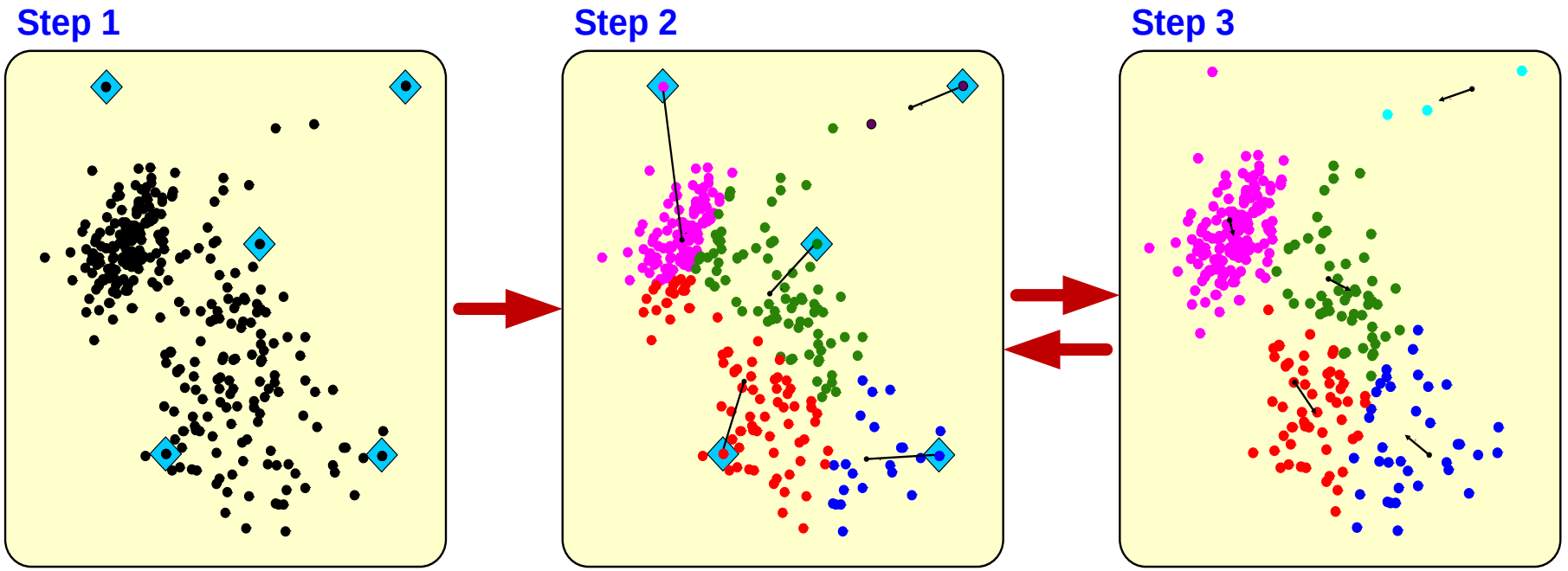
Cluster Analysis for Data Mining

- How many clusters?
  - There is not a “truly optimal” way to calculate
  - Heuristics are often used
- Most cluster analysis methods involve the use of a **distance measure** to calculate the closeness between pairs of items.
  - Euclidian versus Manhattan (taxi-cab)
- Graphs (networks) use geodeisic distance (number of edges of shortest path between nodes in the graph)

Cluster Analysis for Data Mining

- ***k*-Means Clustering Algorithm**
  - *k* : pre-determined number of clusters
  - Algorithm
    (Step 0: Determine value of *k*)
    Step 1: Randomly generate *k* random points as initial cluster centers.
    Step 2: Assign each point to the nearest cluster center.
    Step 3: Re-compute the new cluster centers.
    Repetition Step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm



Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as **market basket analysis**
- Often used as an example to describe DM to ordinary people, such as the famous “relationship between diapers and beers!”

Association Rule Mining (contd.)

- **Input**: the simple point-of-sale transaction data
- **Output**: Most frequent affinities among items
- Example: according to the transaction data…

  "Customers who bought a lap-top computer and virus protection software, also bought an extended service plan 70 percent of the time."
- How do you use such a pattern/knowledge?
  - Put the items next to each other
  - Promote the items as a package
  - Place items far apart from each other!

Association Rule Mining (contd.)

- A representative applications of association rule mining include
  - **In business**: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - **In medicine**: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)

Association Rule Mining (contd.)

- Are all association rules interesting and useful?

  A Generic Rule: **X ⇒ Y [S%, C%]**

  X, Y: products and/or services
  X: Left-hand-side (LHS)
  Y: Right-hand-side (RHS)
  S: Support: how often **X** and **Y** go together
  C: Confidence: how often **Y** go together with the **X**

  Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

Association Rule Mining (contd.)

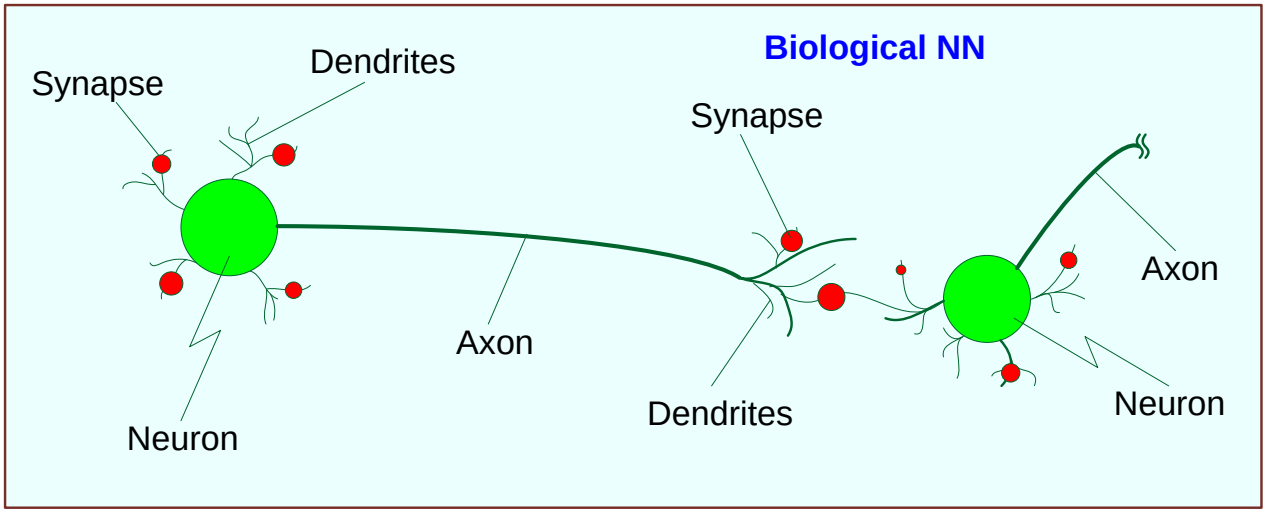
- Algorithms are available for generating association rules
  - Apriori
  - FP-Growth
- The algorithms help identify the **frequent item sets**, which are, then converted to association rules

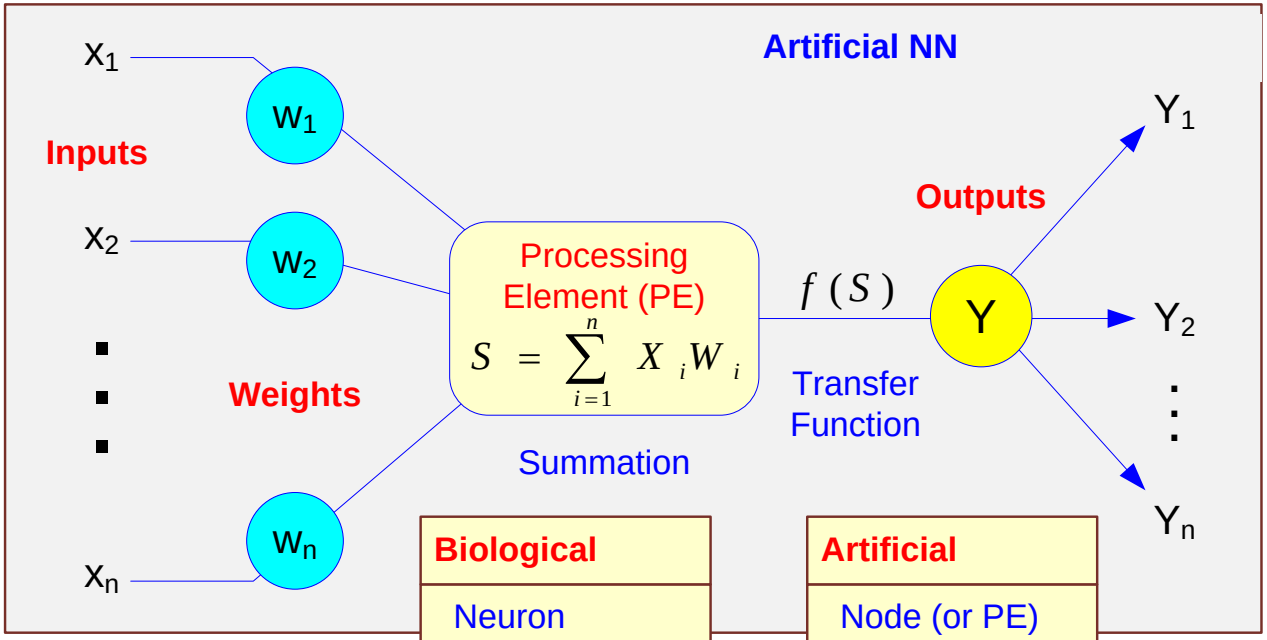
  (Intuitively, a set of items that appears in many baskets is said to be “frequent”)

Artificial Neural Networks for Data Mining

- Artificial neural networks (ANN or NN) is a brain metaphor for information processing
- a.k.a. Neural Computing
- Very good at capturing highly complex, non-linear functions
  Many uses – prediction (regression, classification), clustering / segmentation
- Many application areas including finance, medicine, marketing, manufacturing, service operations, information systems

Biological versus Artificial Neural Networks





Biological	Artificial
Neuron	Node (or PE)
Dendrites	Input
Axon	Output
Synapse	Weight
Slow	Fast
Many (10^9)	Few (10^2)

Data Mining Software

- **Commercial; Free and/or Open Source**
  - SAS - Enterprise Miner;
  - IBM SPSS Modeler, Intelligent Miner
  - RapidMiner
  - Alteryx
  - Databricks Spark
  - Knime
  - H_2O.ai
  - R; Python; Anaconda
  - Keras (high level neural network API supporting popular deep learning libraries like Tensorflow and Microsoft Cognitive Toolkit)
  - Weka
  - Neo4j Graph Platform

Data Mining Myths

- Data mining …
  - provides instant solutions/predictions
  - is not yet viable for business applications
  - requires a separate, dedicated database
  - can only be done by those with advanced degrees
  - is only for large firms that have lots of customer data
  - is another name for statistics
  - Uses OLAP (dimensional slice and dice)

Source: KDNuggets.com

Common Data Mining Blunders

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Leaving insufficient time for data acquisition, selection and preparation
4. Looking only at aggregated results and not at individual records / predictions
5. Not keeping track of the data mining procedure and results

Top 11 Data Mining Mistakes

1. Lack Data
2. Focus on Training
3. Rely on One Technique
4. Ask the Wrong Question
5. Listen (Only) to the Data
6. Accept Leaks from the Future
7. Discount Pesky Cases
8. Extrapolate (difficult to unlearn suppositions)
9. Answer Every Inquiry
10. Sample Casually
11. Believe the Best Model

Statistical Analysis and Data Mining Applications') Nisbet, Elder, Miner (2009)

Recommended Reading

Predictive Analytics and Data Mining – Vijay Kotu and Bala Deshpande (Chapters 1& 2)