Chapter 4

Predictive Analytics I: Data Mining Process, Methods, and Algorithms

Learning Objectives (1 of 2)

4.1 Define data mining as an enabling technology for business analytics

4.2 Understand the objectives and benefits of data mining

4.3 Become familiar with the wide range of applications of data mining

4.4 Learn the standardized data mining processes

4.5 Learn different methods and algorithms of data mining

Learning Objectives (2 of 2)

4.6 Build awareness of the existing data mining software tools

4.7 Understand the privacy issues, pitfalls, and myths of data mining

Opening Vignette (1 of 3)

Miami-Dade Police Department Is Using Predictive Analytics to Foresee and Fight Crime

- Predictive analytics in law enforcement
  - Policing with less
  - New thinking on cold cases
  - The big picture starts small
  - Success brings credibility
  - Just for the facts
  - Safer streets for smarter cities

Opening Vignette (2 of 3)

Discussion Questions

1. Why do law enforcement agencies and departments like Miami-Dade Police Department embrace advanced analytics and data mining?
2. What are the top challenges for law enforcement agencies and departments like Miami-Dade Police Department? Can you think of other challenges (not mentioned in this case) that can benefit from data mining?

Opening Vignette (3 of 3)

3. What are the sources of data that law enforcement agencies and departments like Miami-Dade Police Department use for their predictive modeling and data mining projects?
4. What type of analytics do law enforcement agencies and departments like Miami-Dade Police Department use to fight crime?
5. What does "the big picture starts small" mean in this case? Explain.

What is data mining?

Data Mining Concepts and Definitions
Why Data Mining?

- 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.
- The exponential increase in data processing and storage capabilities; and decrease in cost.
- 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.
- Data mining: a misnomer?
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,…

Figure 4.1 Data Mining is a Blend of Multiple Disciplines

Application Case 4.1

Visa Is Enhancing the Customer Experience While Reducing Fraud with Predictive Analytics and Data Mining

Questions for Discussion

1. What challenges were Visa and the rest of the credit card industry facing?
2. How did Visa improve customer service while also improving retention of fraud?
3. What is in-memory analytics, and why was it necessary?

Data Mining Characteristics & Objectives

- Source of data for DM is often a consolidated data warehouse (not always!).
- 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.).

How Data Mining Works

- 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

Application Case 4.2

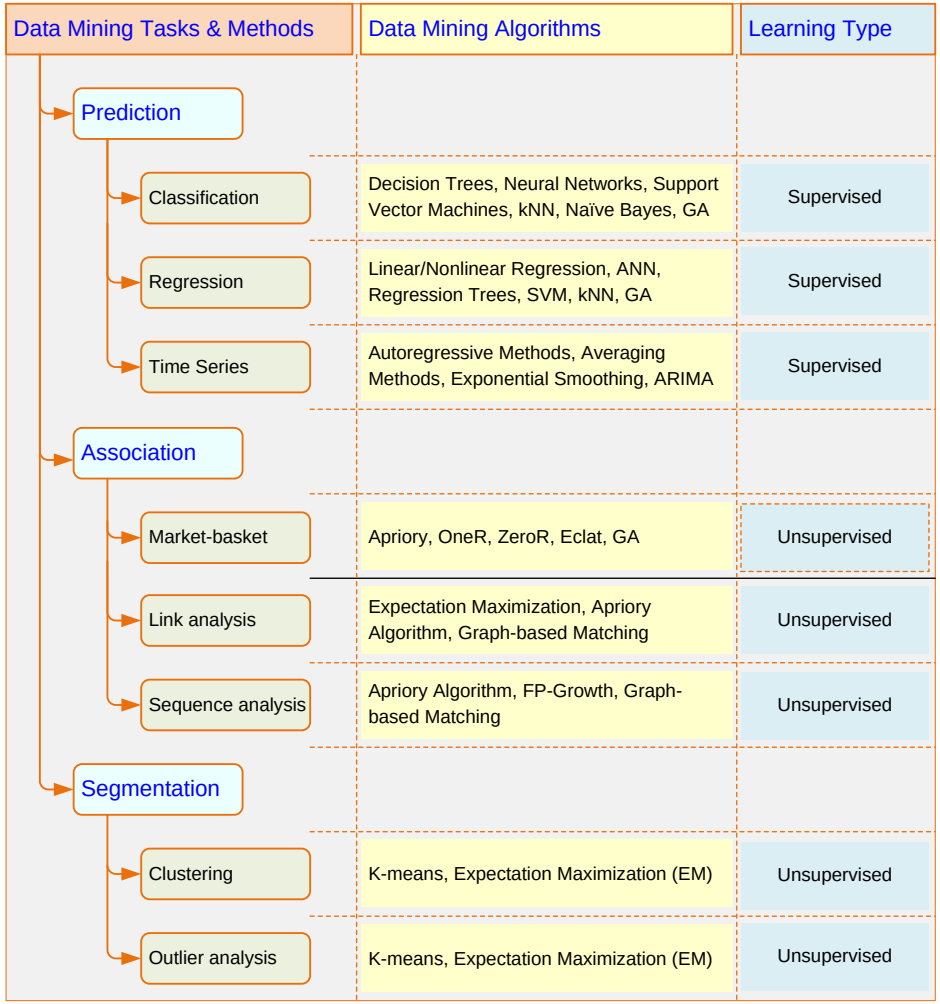
Dell Is Staying Agile and Effective with Analytics in the 21st Century

Questions for Discussion

1. What was the challenge Dell was facing that led to their analytics journey?
2. What solution did Dell develop and implement? What were the results?
3. As an analytics company itself, Dell has used its service offerings for its own business. Do you think it is easier or harder for a company to taste its own medicine? Explain.

A Taxonomy for Data Mining

- A Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms



Other Data Mining Patterns/Tasks

- Time-series forecasting
  - Part of the sequence or link analysis?
- Visualization
  - Another data mining task?
- **Data Mining versus Statistics**
  - Are they the same?
  - What is the relationship between the two?

Data Mining Applications (1 of 4)

- 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 (2 of 4)

- 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 (3 of 4)

- 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 (4 of 4)

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel, entertainment, sports
- Healthcare and medicine
- Sports,… virtually everywhere…

Application Case 4.3

Predictive Analytic and Data Mining Help Stop Terrorist Funding

Questions for Discussion

1. How can data mining be used to fight terrorism? Comment on what else can be done beyond what is covered in this short application case.
2. Do you think data mining, although essential for fighting terrorist cells, also jeopardizes individuals' rights of privacy?

Data Mining Process

- A manifestation of the best practices
- A systematic way to conduct DM projects
- Moving from **Art to Science** for DM project
- Everybody has a different version
- 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)

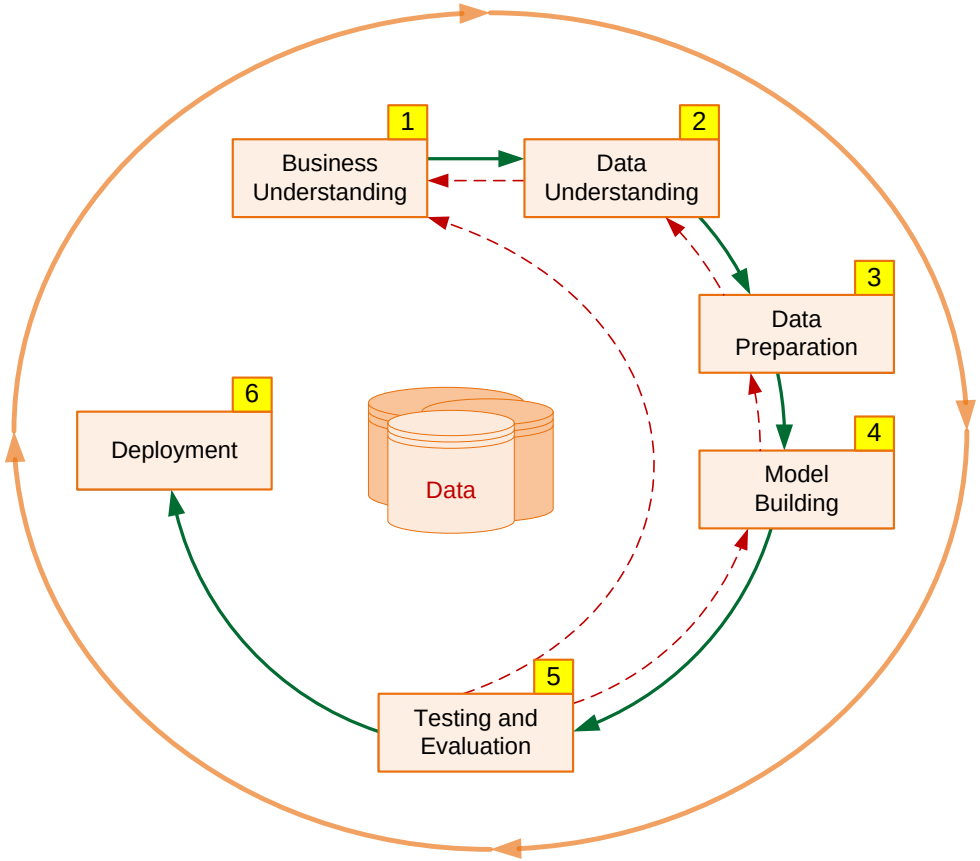
Data Mining Process: CRISP-DM (1 of 2)

- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- Proposed in 1990s by a European consortium
- Composed of six consecutive phases
  - **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

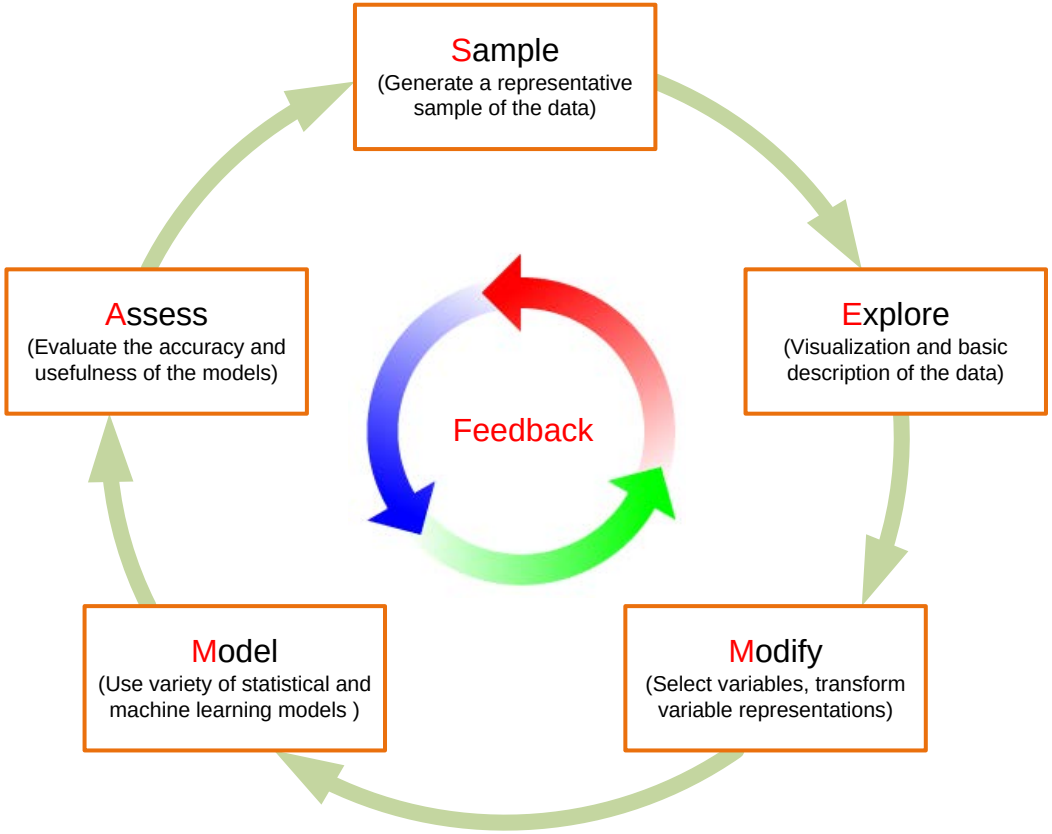
Data Mining Process: CRISP-DM (2 of 2)

- **Figure 4.3** The Six-Step CRISP-DM Data Mining Process →
- The process is highly repetitive and experimental (DM: art versus science?)



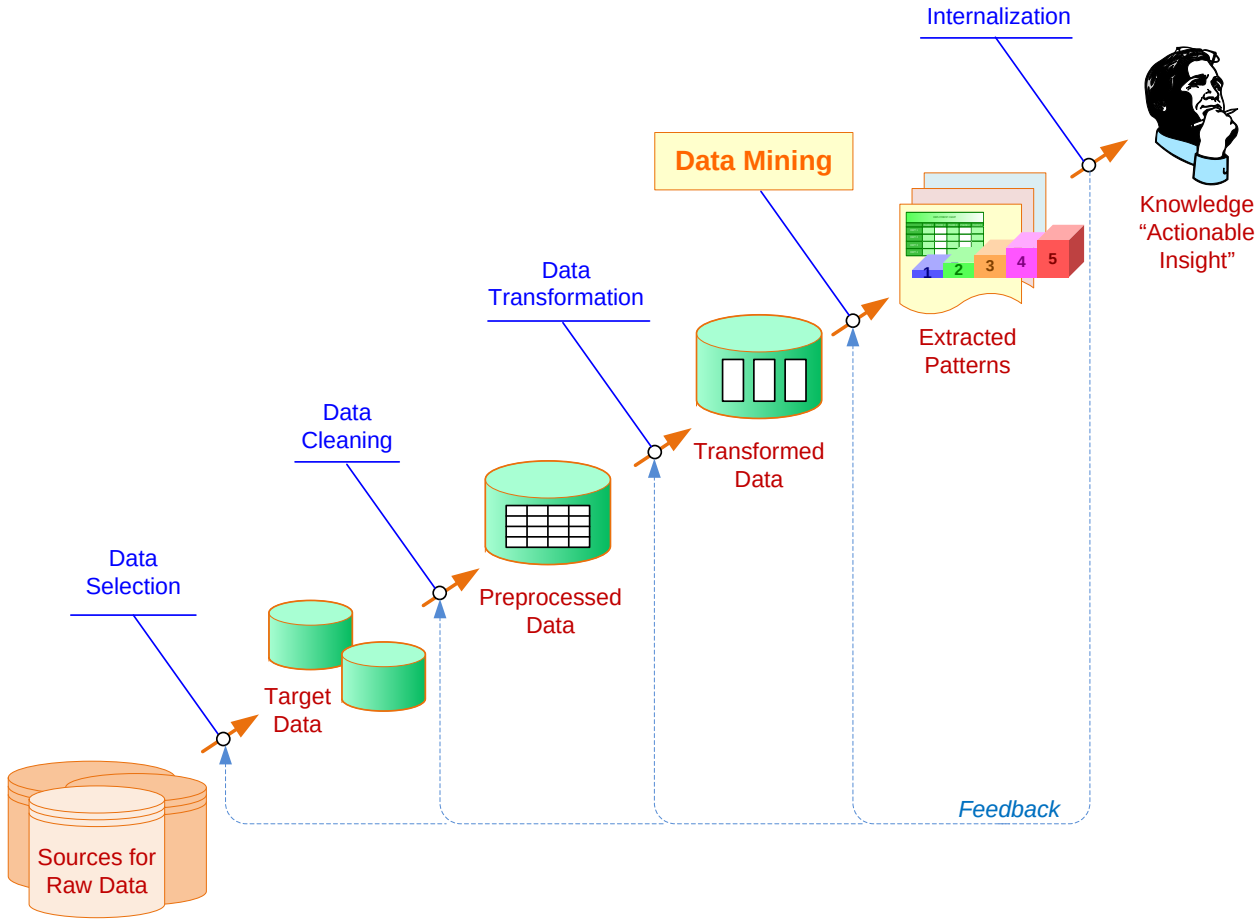
Data Mining Process: SEMMA

- **Figure 4.5** SEMMA Data Mining Process
- Developed by SAS Institute



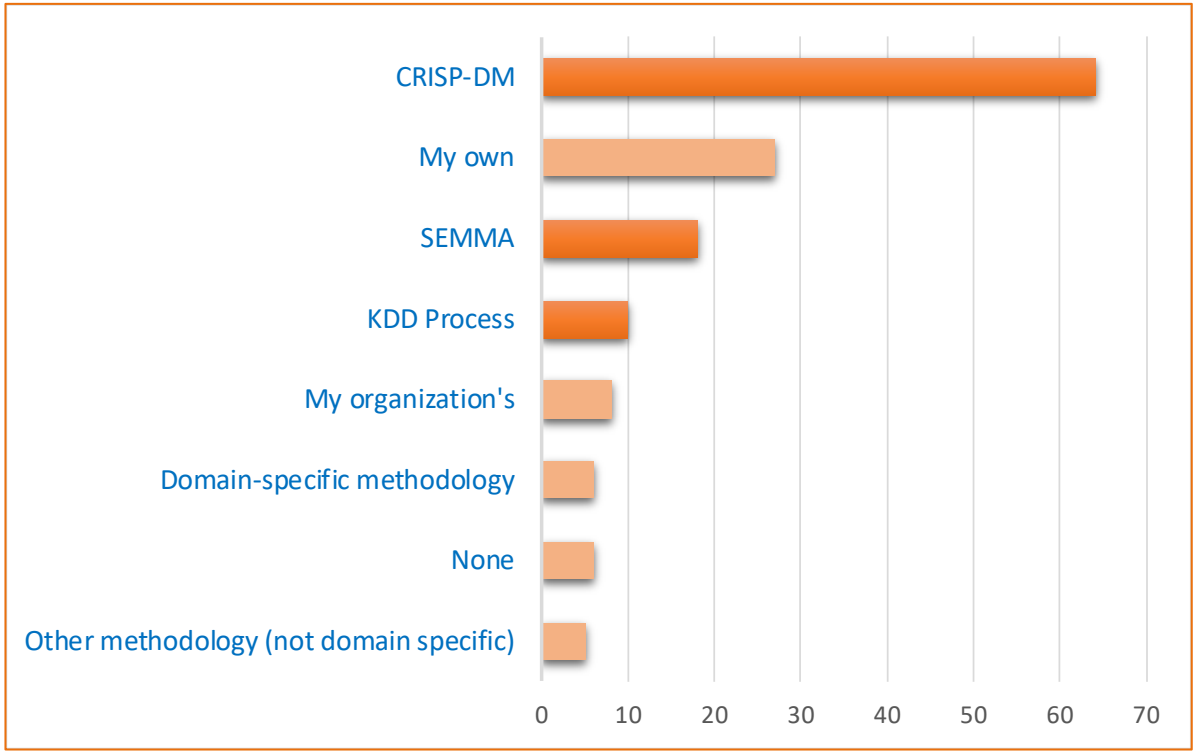
Data Mining Process: KDD

- **Figure 4.6** KDD (Knowledge Discovery in Databases) Process



Which Data Mining Process is the Best?

- **Figure 4.7** Ranking of Data Mining Methodologies/Processes.



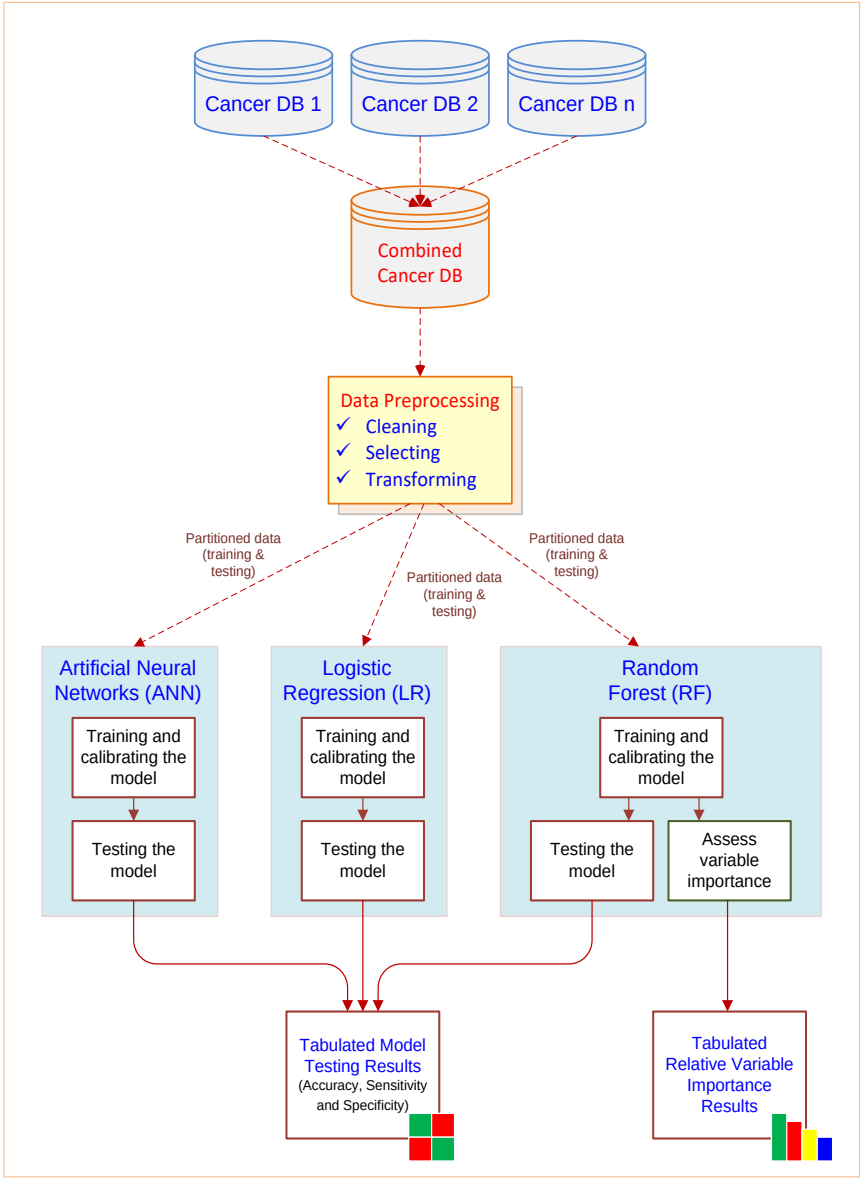
Source: Used with permission from KDnuggets.com.

Application Case 4.4

Data Mining Helps in Cancer Research

Questions for Discussion

1. How can data mining be used for ultimately curing illnesses like cancer?
2. What do you think are the promises and major challenges for data miners in contributing to medical and biological research endeavors?



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 versus predicting/usage speed
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

Accuracy of Classification Models

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

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

$$True\ PositiveRate = \frac{TP}{TP + FN}$$

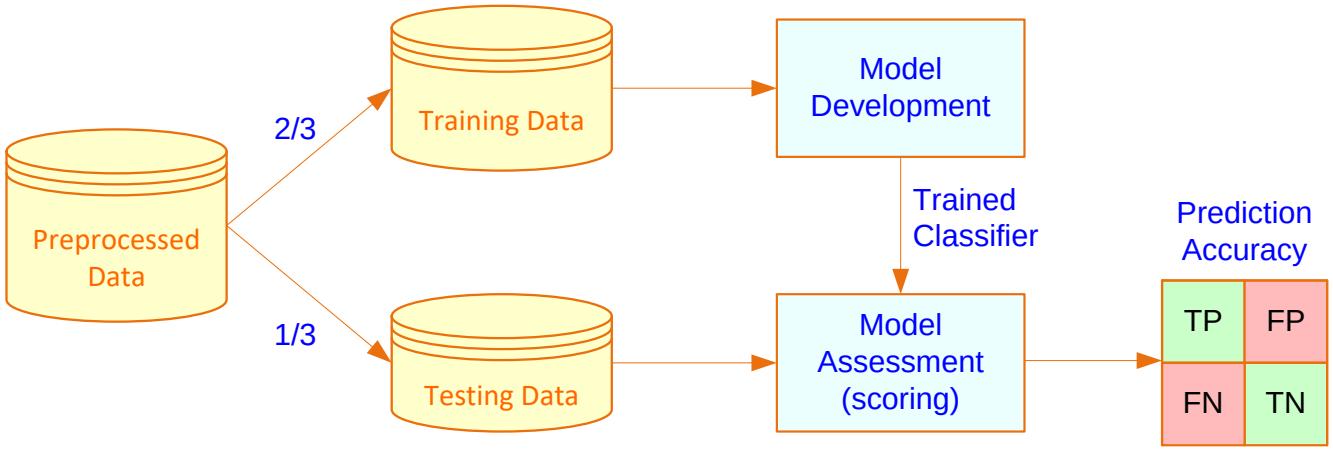
$$True\ NegativeRate = \frac{TN}{TN + FP}$$

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

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

Estimation Methodologies for Classification: Single/Simple Split

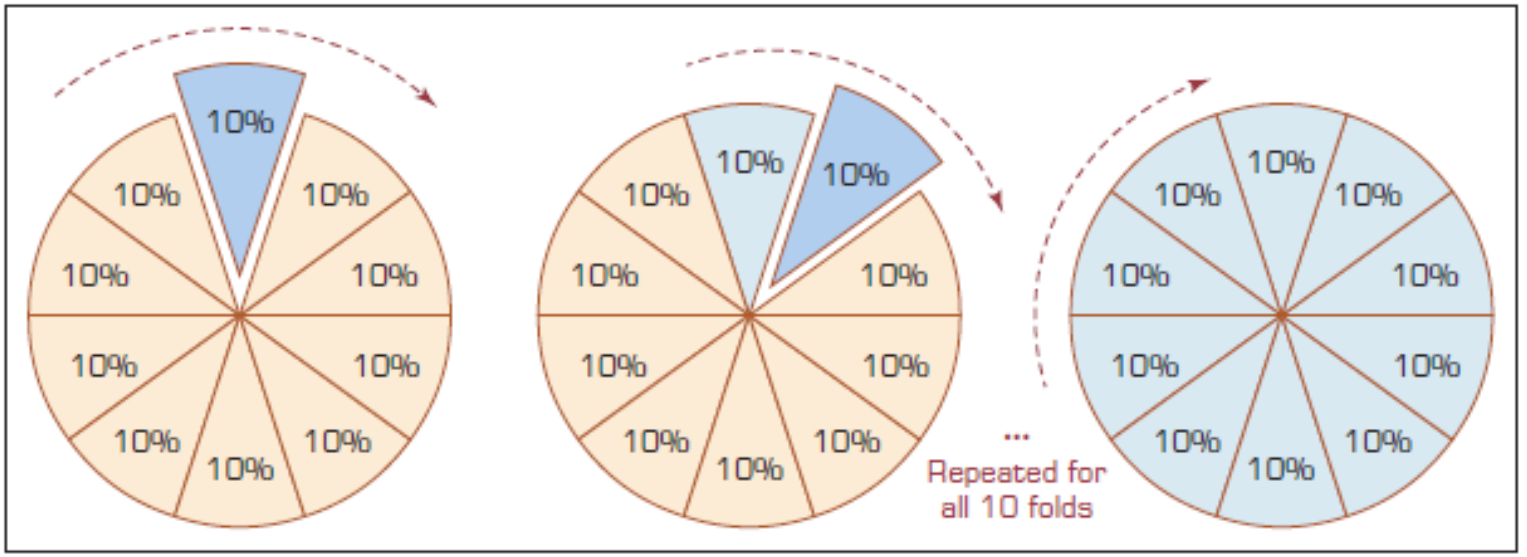
- **Simple split** (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets: training (~70%) and testing (30%)



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

Estimation Methodologies for Classification: *k*-Fold Cross Validation (rotation estimation)

- Data is split into *k* mutual subsets and *k* number training/testing experiments are conducted
- **Figure 4.10** A Graphical Depiction of k-Fold Cross-Validation

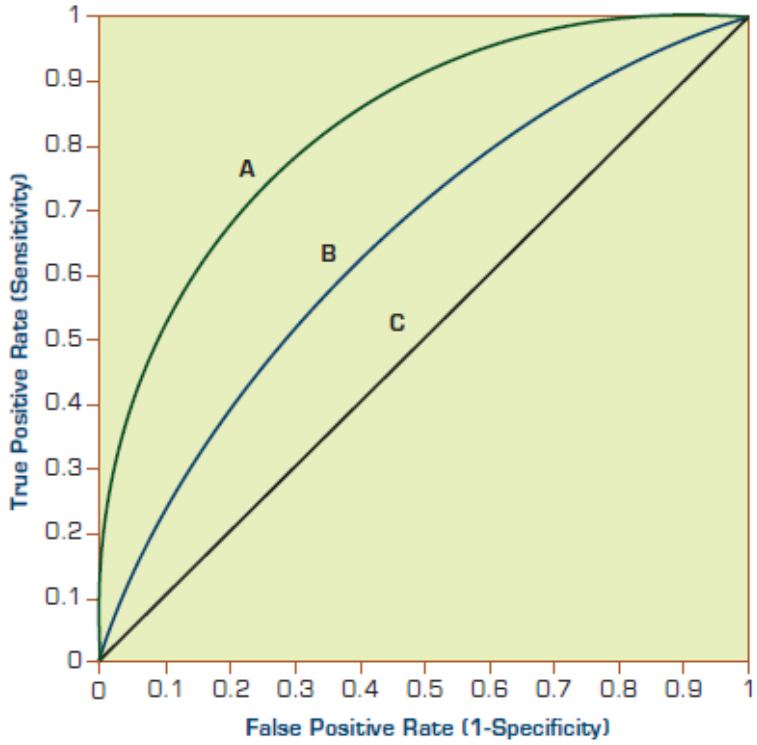


Additional Estimation Methodologies for Classification

- Leave-one-out
  - Similar to *k*-fold where *k* = number of samples
- Bootstrapping
  - Random sampling with replacement
- Jackknifing
  - Similar to leave-one-out
- Area Under the ROC Curve (AUC)
  - ROC: Receiver Operating Characteristics (a term borrowed from radar image processing)

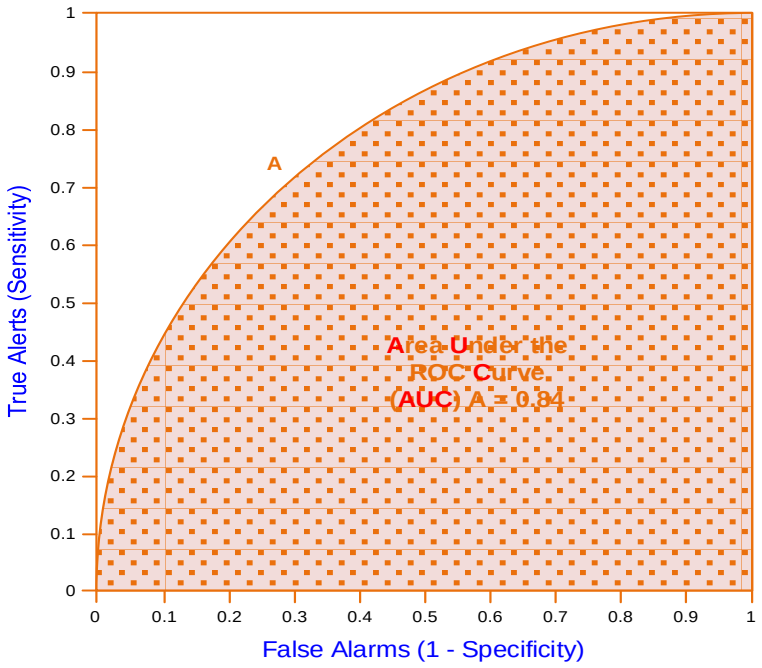
Area Under the ROC Curve (AUC) (1 of 2)

- Works with binary classification
- **Figure 4.11** A Sample ROC Curve



Area Under the ROC Curve (AUC) (2 of 2)

- Produces values from 0 to 1.0
- Random chance is 0.5 and perfect classification is 1.0
- Produces a good assessment for skewed class distributions too!



Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Decision Trees (1 of 2)

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

A general algorithm (steps) for building a decision tree

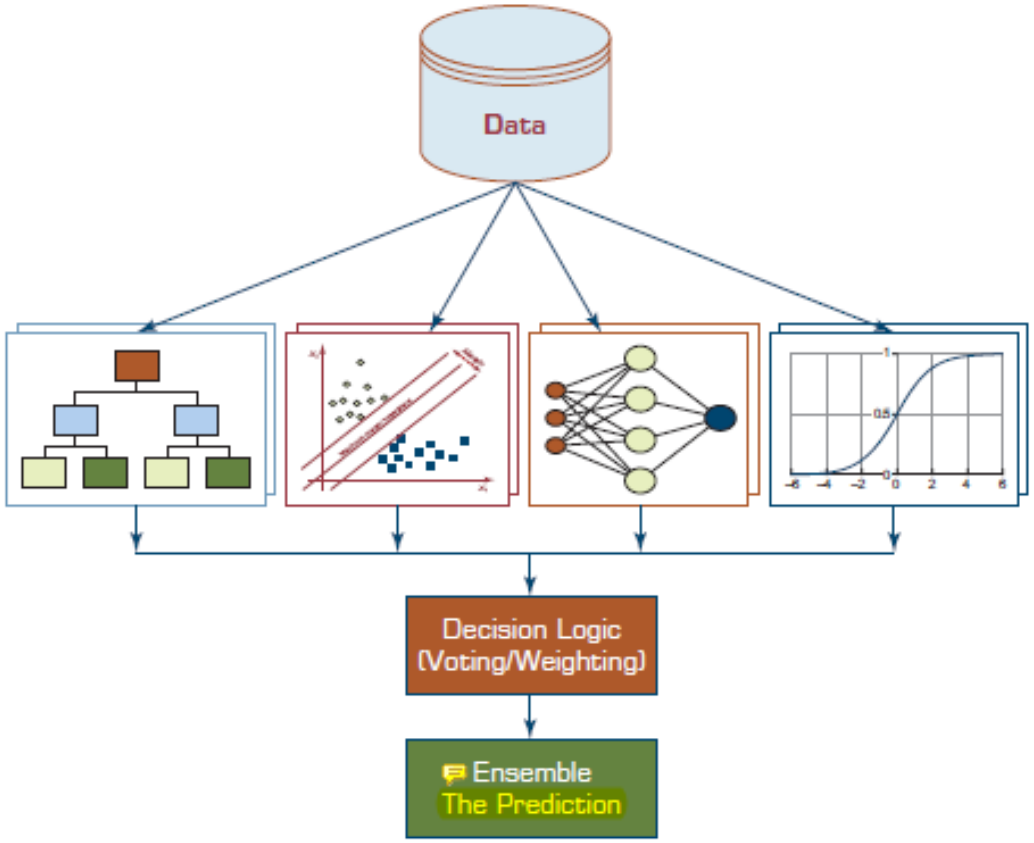
1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
4. Repeat steps 2 and 3 for each and every leaf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

Decision Trees (2 of 2)

- 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

Ensemble Models for Predictive Analytics

- Produces more robust and reliable prediction models
- **Figure 4.12** Graphical Illustration of a Heterogeneous Ensemble



Application Case 4.5

Influence Health Uses Advanced Predictive Analytics to Focus on the Factors That Really Influence People's Healthcare Decisions

Questions for Discussion

1. What did Influence Health do?
2. What were the challenges, the proposed solutions, and the obtained results?
3. How can data mining help companies in the healthcare industry (in ways other than the ones mentioned in this case)?

Cluster Analysis for Data Mining (1 of 4)

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output/target variable
- In marketing, it is also known as segmentation

Cluster Analysis for Data Mining (2 of 4)

- 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 (3 of 4)

- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on.
  - Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms
- **How many clusters?**

Cluster Analysis for Data Mining (4 of 4)

- ***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.

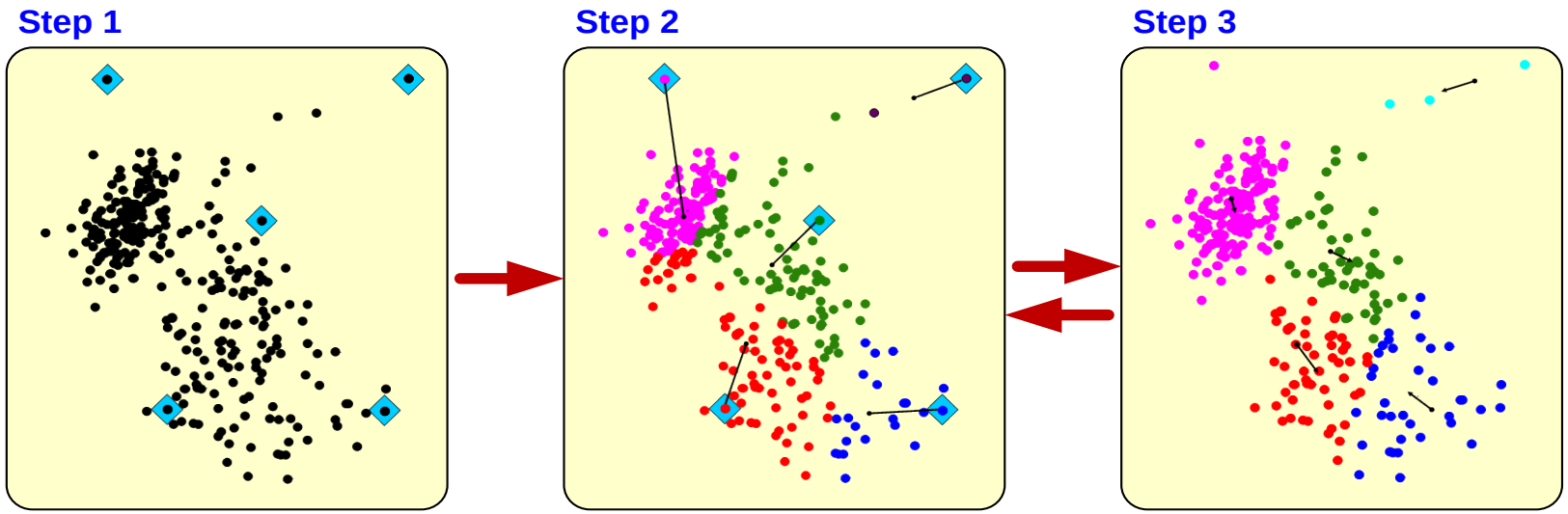
  Step 2: Assign each point to the nearest cluster center.

  Step 3: Re-compute the new cluster centers.

  Repetition step: Repeat steps 3 and 4 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

- **Figure 4.13** A Graphical Illustration of the Steps in the *k*-Means Algorithm



Association Rule Mining (1 of 6)

- 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 (2 of 6)

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

  "Customer who bought a lap-top computer and a virus protection software, also bought 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 (3 of 6)

- A representative application of association rule mining includes
  - **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 (4 of 6)

- Are all association rules interesting and useful?

  A Generic Rule: X $\Rightarrow$ 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} $\Rightarrow$

{Extended Service Plan} [30%, 70%]

Association Rule Mining (5 of 6)

- Several algorithms are developed for discovering (identifying) association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the **frequent itemsets**, which are then converted to association rules

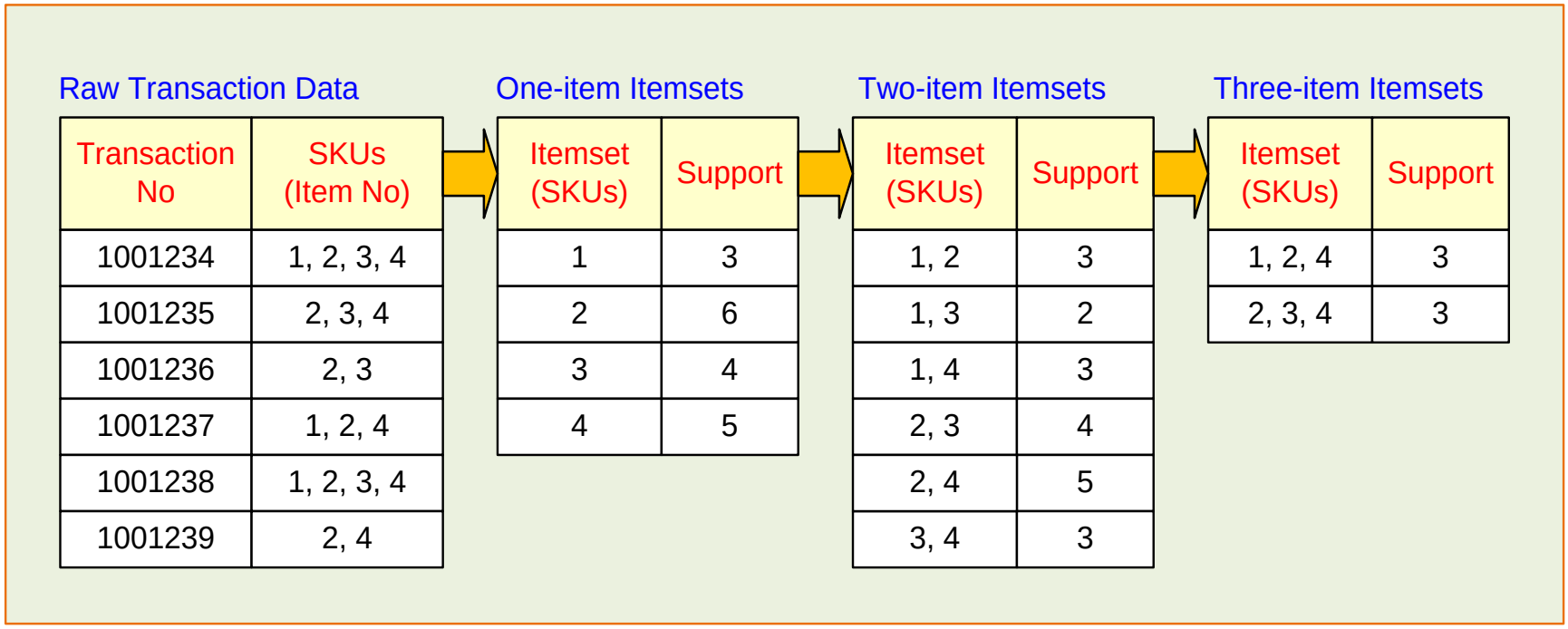
Association Rule Mining (6 of 6)

- **Apriori Algorithm**
  - Finds subsets that are common to at least a minimum number of the itemsets
  - Uses a bottom-up approach
    - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    - groups of candidates at each level are tested against the data for minimum support

    (see the figure) → --

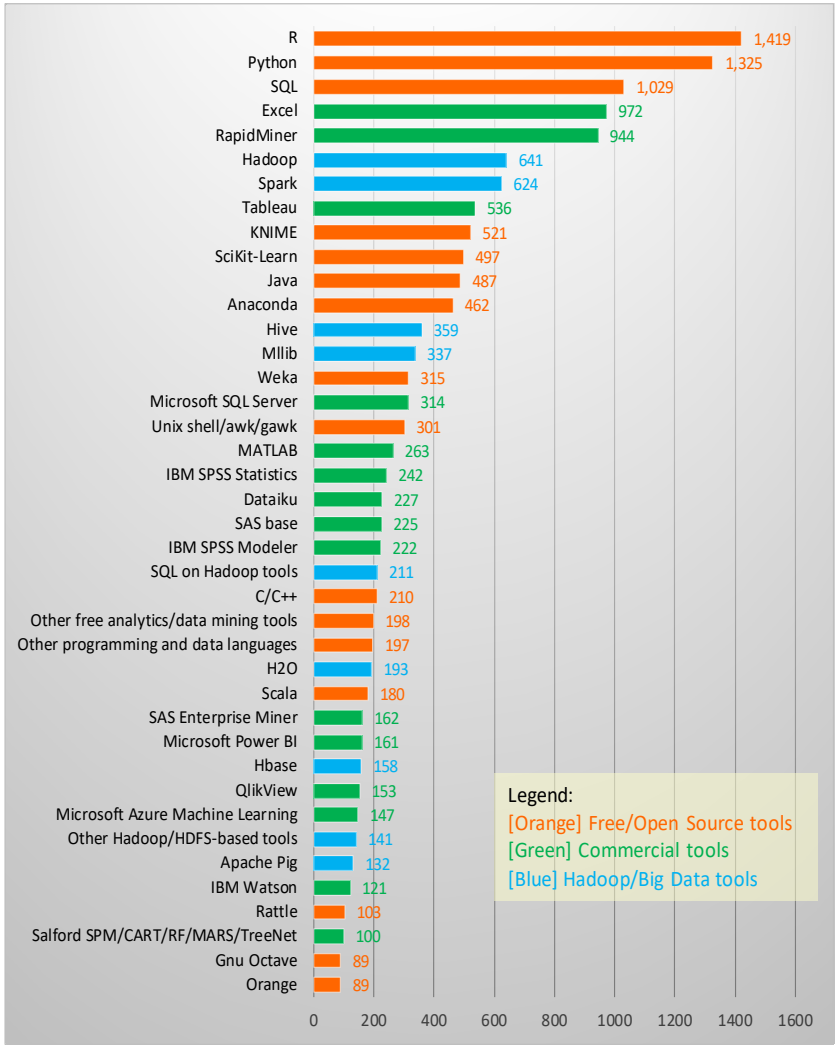
Association Rule Mining Apriori Algorithm

- **Figure 4.13** A Graphical Illustration of the Steps in the *k*-Means Algorithm



Data Mining Software Tools

- **Commercial**
  - IBM SPSS Modeler (formerly Clementine)
  - SAS Enterprise Miner
  - Statistica - Dell/Statsoft
  - … many more
- **Free and/or Open Source**
  - KNIME
  - RapidMiner
  - Weka
  - R, …



Application Case 4.6 (1 of 5)

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

- Goal: Predicting financial success of Hollywood movies before the start of their production process
- How: Use of advanced predictive analytics methods
- Results: promising

Application Case 4.6 (2 of 5)

A Typical Classification Problem

Dependent Variable

Class No.	1	2	3	4	5	6	7	8	9
Range (in $Millions)	> 1 (Flop)	> 1 > 10	> 10 < 20	> 20 < 40	> 40 < 65	> 65 < 100	> 100 < 150	> 150 < 200	> 200 (Blockbuster)

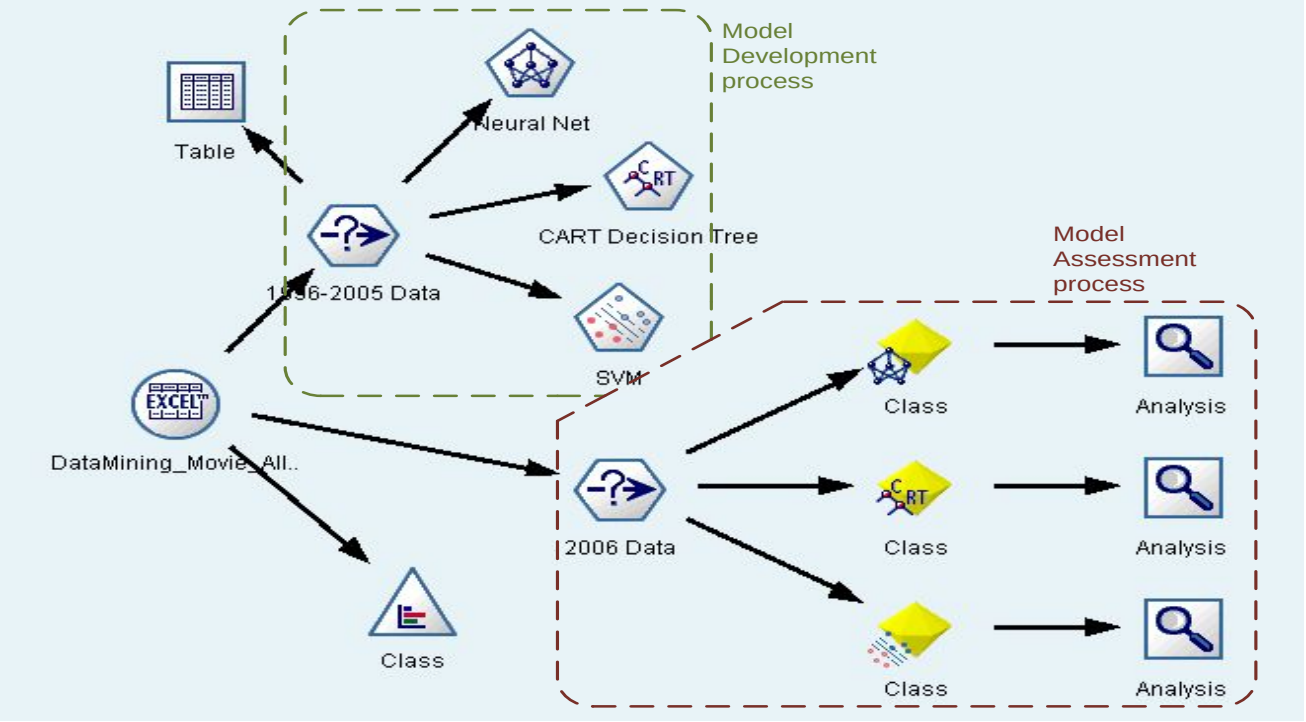
Application Case 4.6 (3 of 5)

Independent Variables

Independent Variable	Number of Values	Possible Values
MPAA Rating	5	G, PG, PG-13, R, NR
Competition	3	High, Medium, Low
Star value	3	High, Medium, Low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, Medium, Low
Sequel	2	Yes, No
Number of screens	1	Positive integer

Application Case 4.6 (4 of 5)

The DM Process Map in IBM SPSS Modeler



Application Case 4.6 (5 of 5)

	Prediction Models					
	Individual Models			Ensemble Models		
Performance Measure	SVM	ANN	CART	Random Forest	Boosted Tree	Fusion (Average)
Count (Bingo)	192	182	140	189	187	**194**
Count (1-Away)	104	120	126	121	104	**120**
Accuracy (% Bingo)	55.49%	52.60%	40.46%	54.62%	54.05%	**56.07%**
Accuracy (% 1-Away)	85.55%	87.28%	87.28%	89.60%	84.10%	**90.75%**
Standard deviation	0.93	0.87	1.05	0.76	0.84	**0.63**

*Training set 1998 – 2005 movies; Test set : 2006 Movies

Table 4.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Data Mining Mistakes

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. Beginning without the end in mind
4. Not leaving sufficient time for data acquisition, selection, and preparation
5. Looking only at aggregated results and not at individual records/predictions
6. … 10 more mistakes… in your book