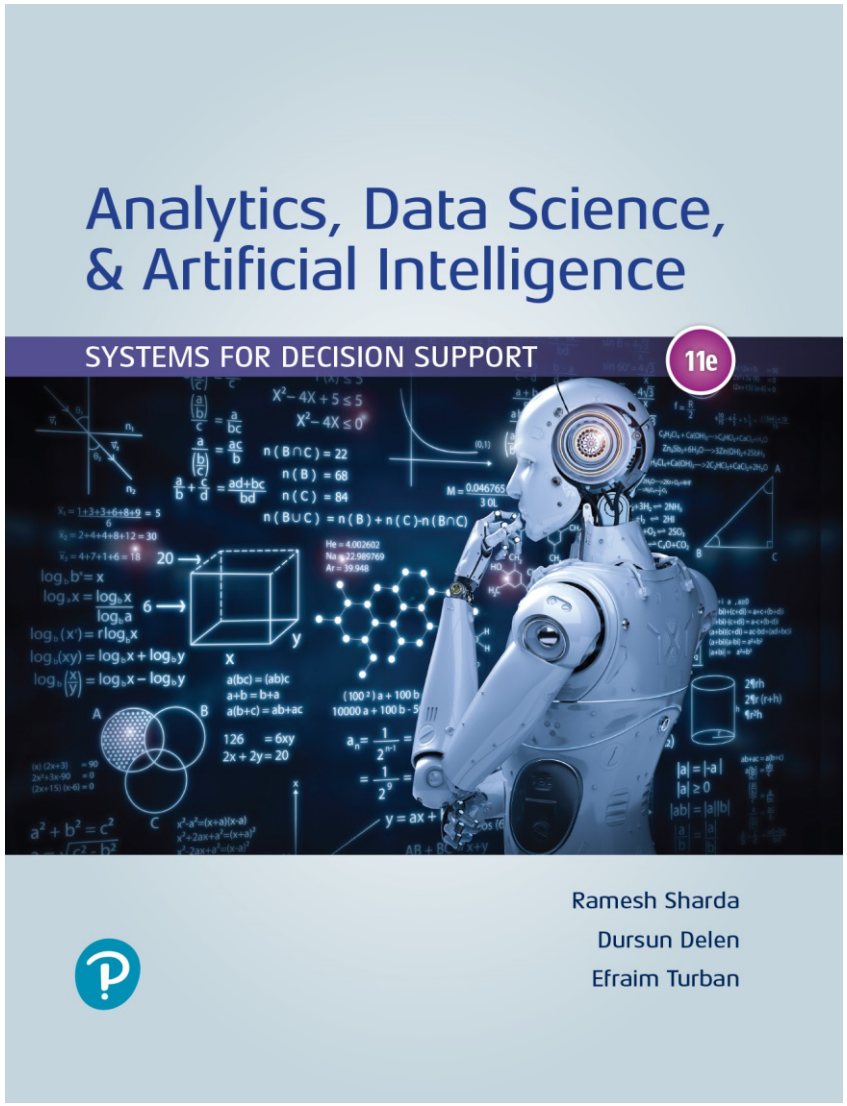
Analytics, Data Science and A I: Systems for Decision Support

Eleventh Edition



Chapter 4

Data Mining Process, Methods, and Algorithms

Learning Objectives

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.

4.6 Build awareness of the existing data mining software tools

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

What is Data Mining?

- Data mining is a process that uses statistics, mathematical, and AI technologies to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.

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

Opening Vignette

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

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.
- Decrease in hardware and software for data storage & processing costs.
- Movement toward conversion of information resources into nonphysical form.

Data Mining Is a Blend of Multiple Disciplines

Figure 4.1 Data Mining Is a Blend of Multiple Disciplines.

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

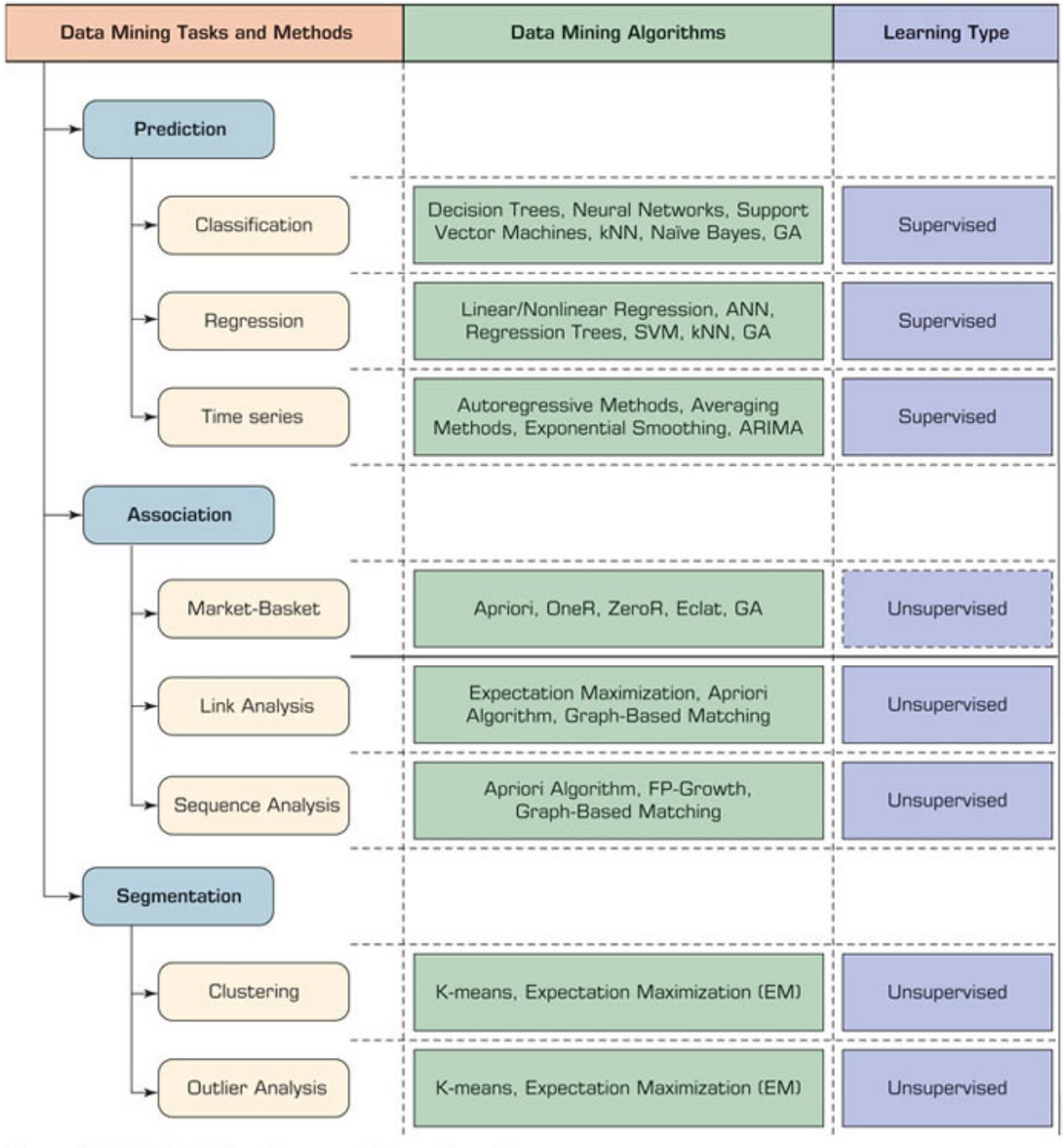
- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - There are four different types of patterns:
    - Prediction: tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
    - Association: find the commonly co-occurring groupings of things, such as baby formula and diapers going together in market-basket analysis.

How Data Mining Works

- Types of patterns (continued)
  - Clusters (or segmentation): identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.
  - Sequential relationships: discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

A Taxonomy for Data Mining

Figure 4.2 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…

Data Mining Process

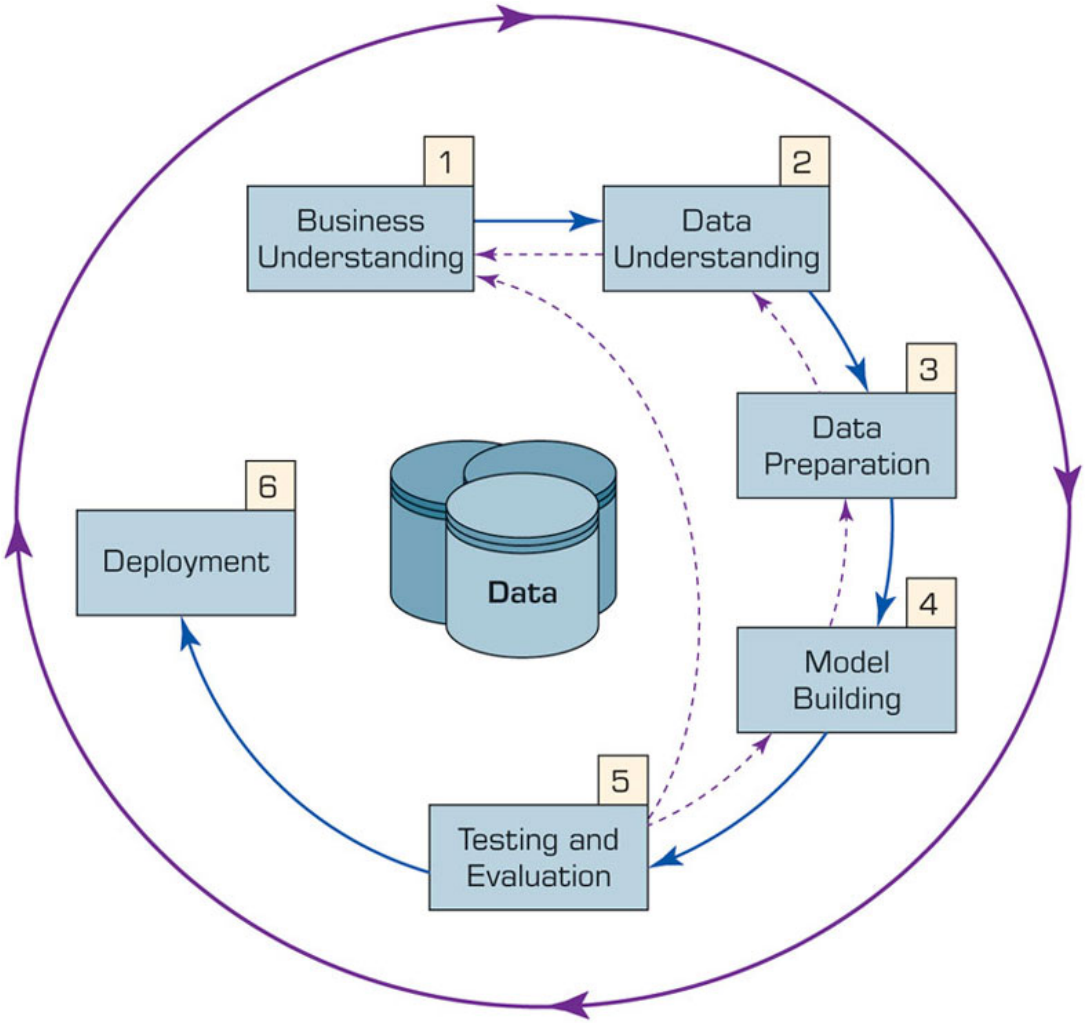
- A manifestation of the best practices
- A systematic way to conduct DM projects
- Moving from Art to Science for DM project
- Everybody has a different version
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

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

- Cross Industry Standard Process for Data Mining
- 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

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

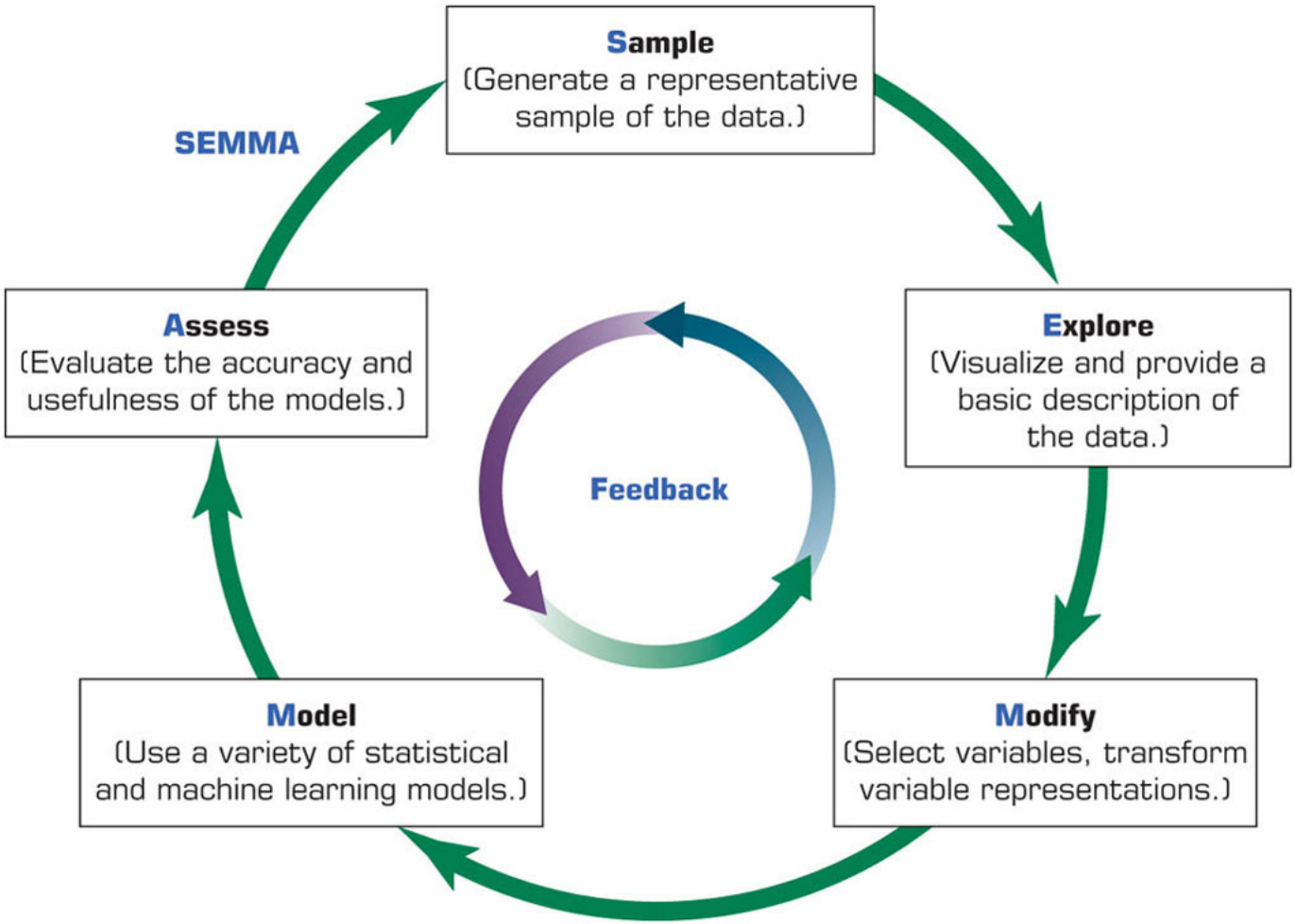
- **Figure 4.3** The Six-Step CRISP-DM Data Mining Process.
- The process is highly repetitive and experimental (DM: art versus science?)



Data Mining Process: SEMMA

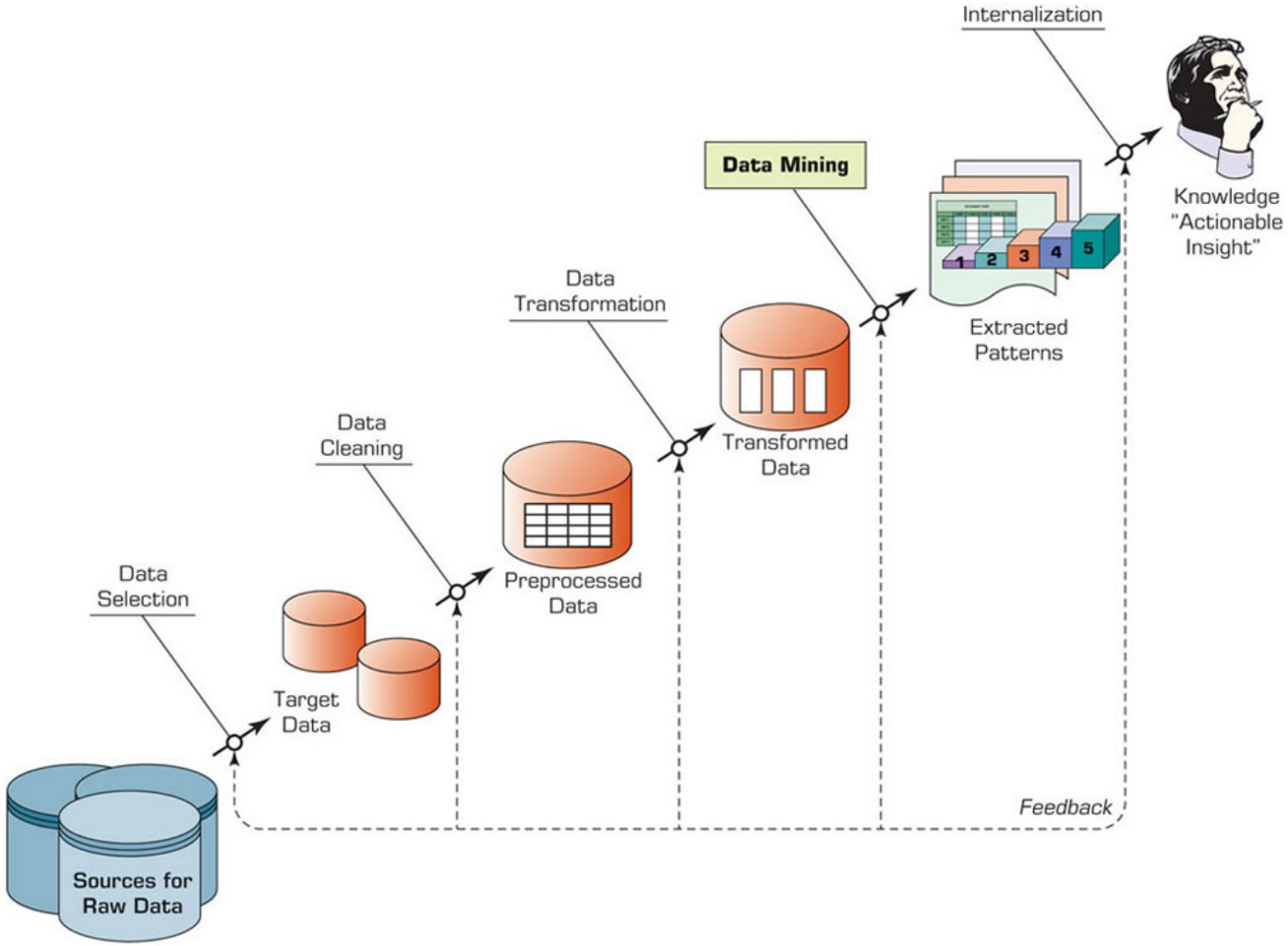
Figure 4.5 SEMMA Data Mining Process.

- Developed by SAS Institute



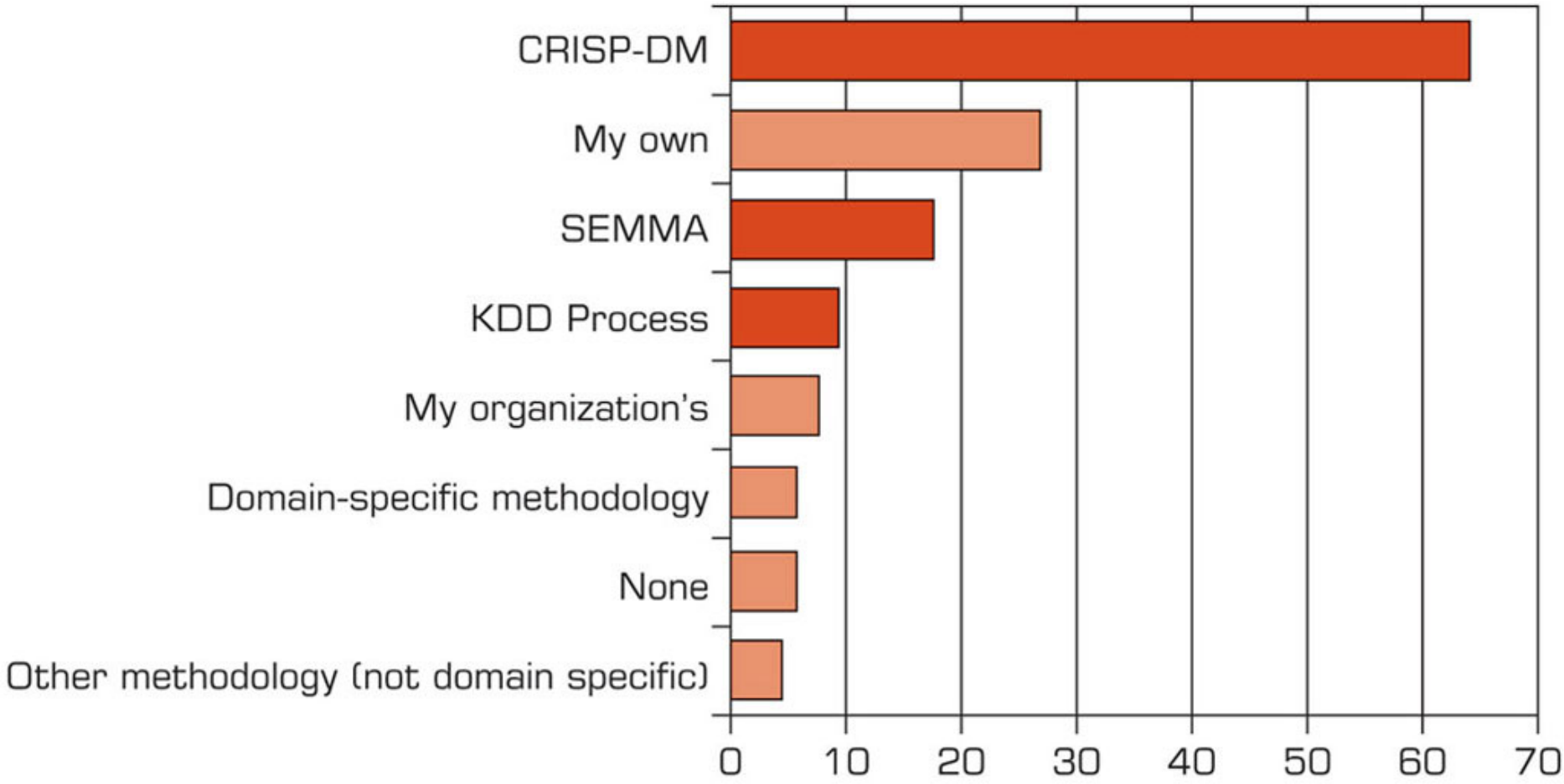
Data Mining Process: KDD

Figure 4.6 KDD (Knowledge Discovery in Databases) Process.



What Data Mining Methodology are you using?

Figure 4.7 Ranking of Data Mining Methodologies/Processes.



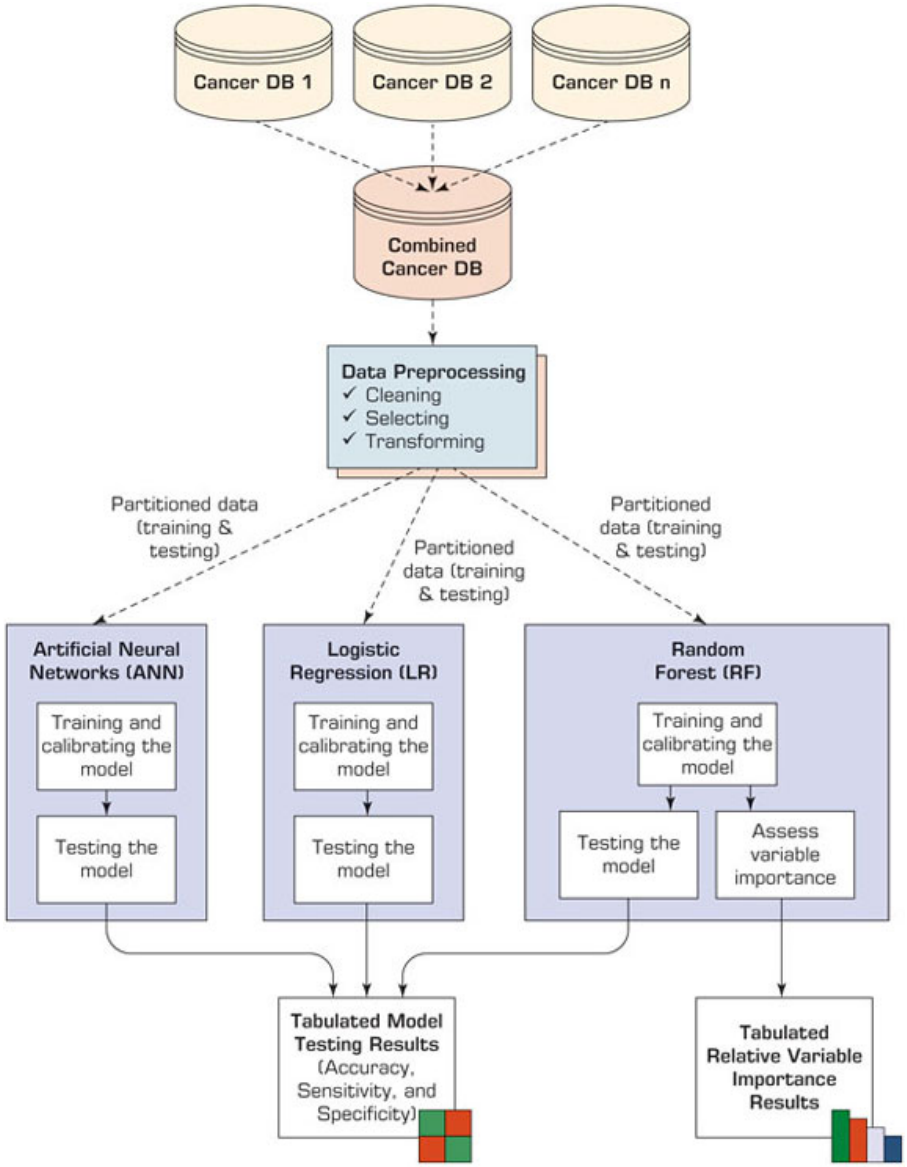
Source: Used with permission from KDnuggets.com.

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?



Best Algorithms based on type of DM Task

- Depending on the business need, different types of data mining tasks can be used: prediction, clustering, or association.

- Most popular algorithms to be used based on type of task:
  1. decision trees for classification (prediction),
  2. *k*-means for clustering (segmentation),
  3. Apriori algorithm for association rule mining.

Data Mining Methods for Prediction:

- Classification versus regression
  - Classification – what is being predicted is a class label
    - weather: sunny, cloudy, rainy
    - credit approval: good, bad credit risk
  - Regression: what is being predicted is a numeric value
    - Temperature: 31
    - Number of attendees: 100,000

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

Data Mining Methods: Classification

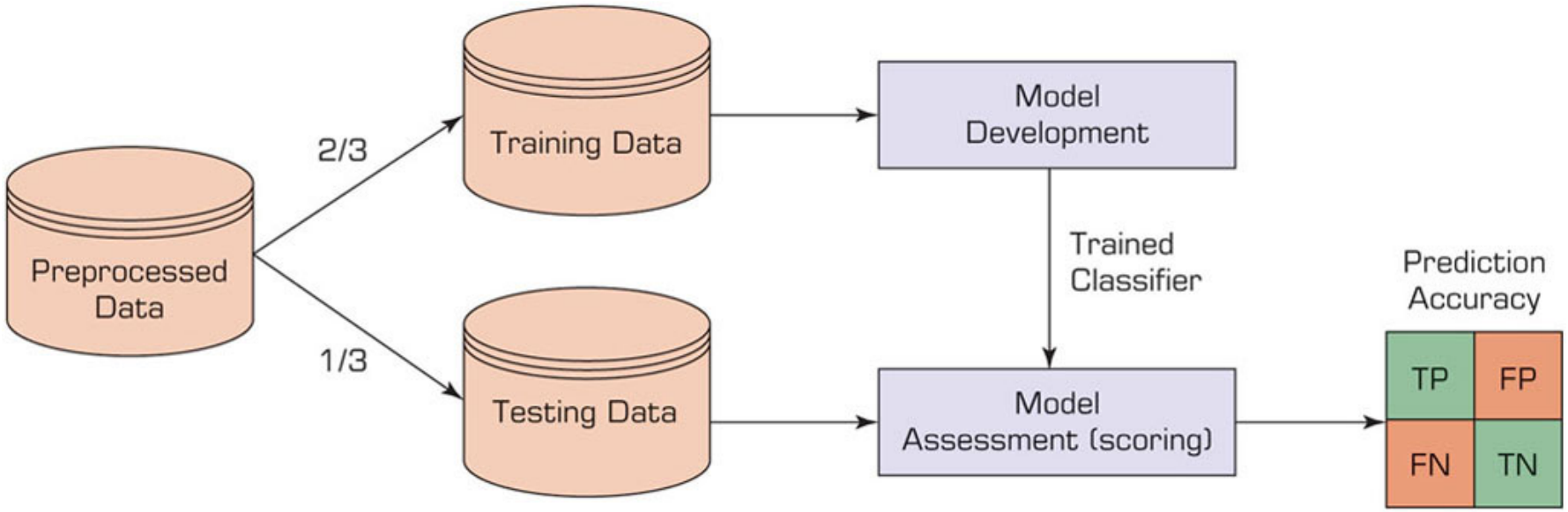
- The output variable is categorical (nominal or ordinal) in nature
  - Nominal data:
    - data that can be labelled or classified into mutually exclusive categories within a variable.
    - Categories cannot be ordered in a meaningful way.
    - Example, for the nominal variable of preferred mode of transportation, you may have the categories of car, bus, train, tram or bicycle.
  - Ordinal data:
    - statistical data type where the variables have natural, ordered categories
    - Example: For a grading system: excellent, very good, good, poor;
      - or for winner in a race: first, second, third.

Data Mining Methods: Classification

- Two-step Methodology of classification-type prediction involves:
  - Model development/training
    - A collection of input data (variables), including the predicted actual known class labels (for loans approval as an example: good, risky) is used for building and train the model.
  - Model testing/deployment
    - The model is tested against the holdout sample for accuracy assessment and eventually deployed for actual use where it is to predict classes of new data instances (where the class label is unknown).

Estimation Methodologies for Classification: Single/Simple Split

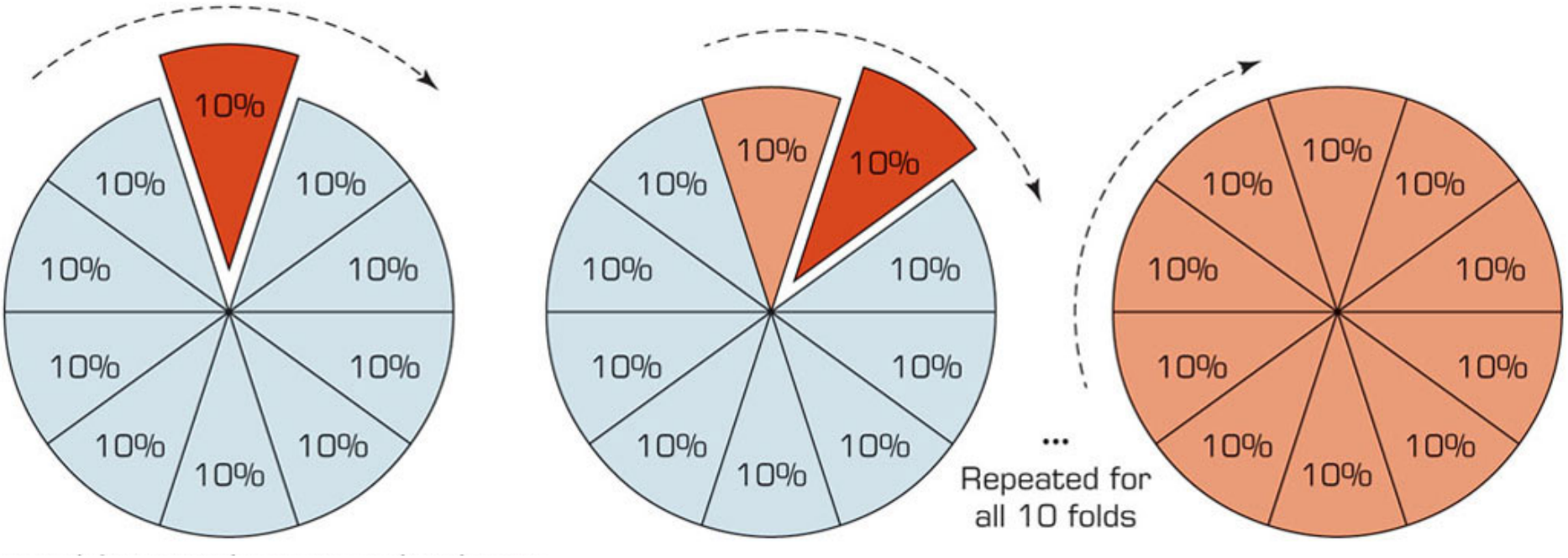
- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets: training (~70%) and testing (30%)



Estimation Methodologies for Classification: *k*-Fold Cross Validation

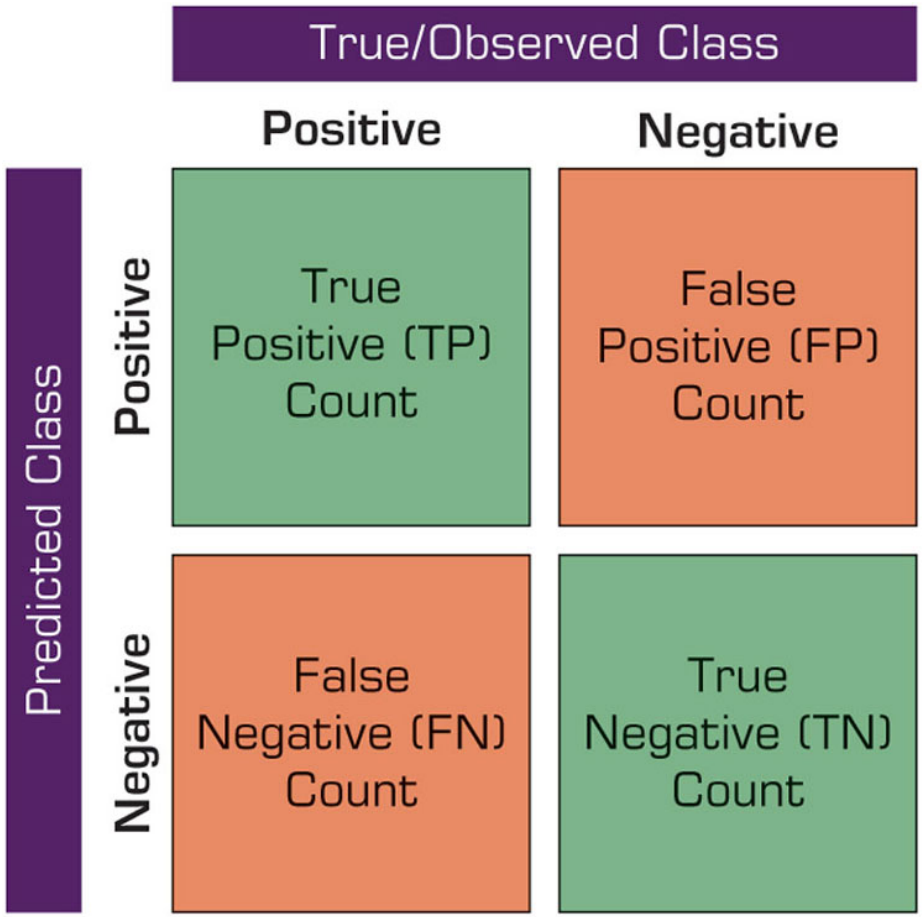
- Data is split into *k* mutual subsets and *k* number training/testing experiments are conducted

Figure 4.10 A Graphical Depiction of *k*-Fold Cross-Validation.



Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix (or, classification matrix)



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

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

$$Recall = \frac{TP}{TP + FN}$$

Figure 4.8 Matrix for tabulation of two-classification results

Classification Techniques

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

Decision Trees

- 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 the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

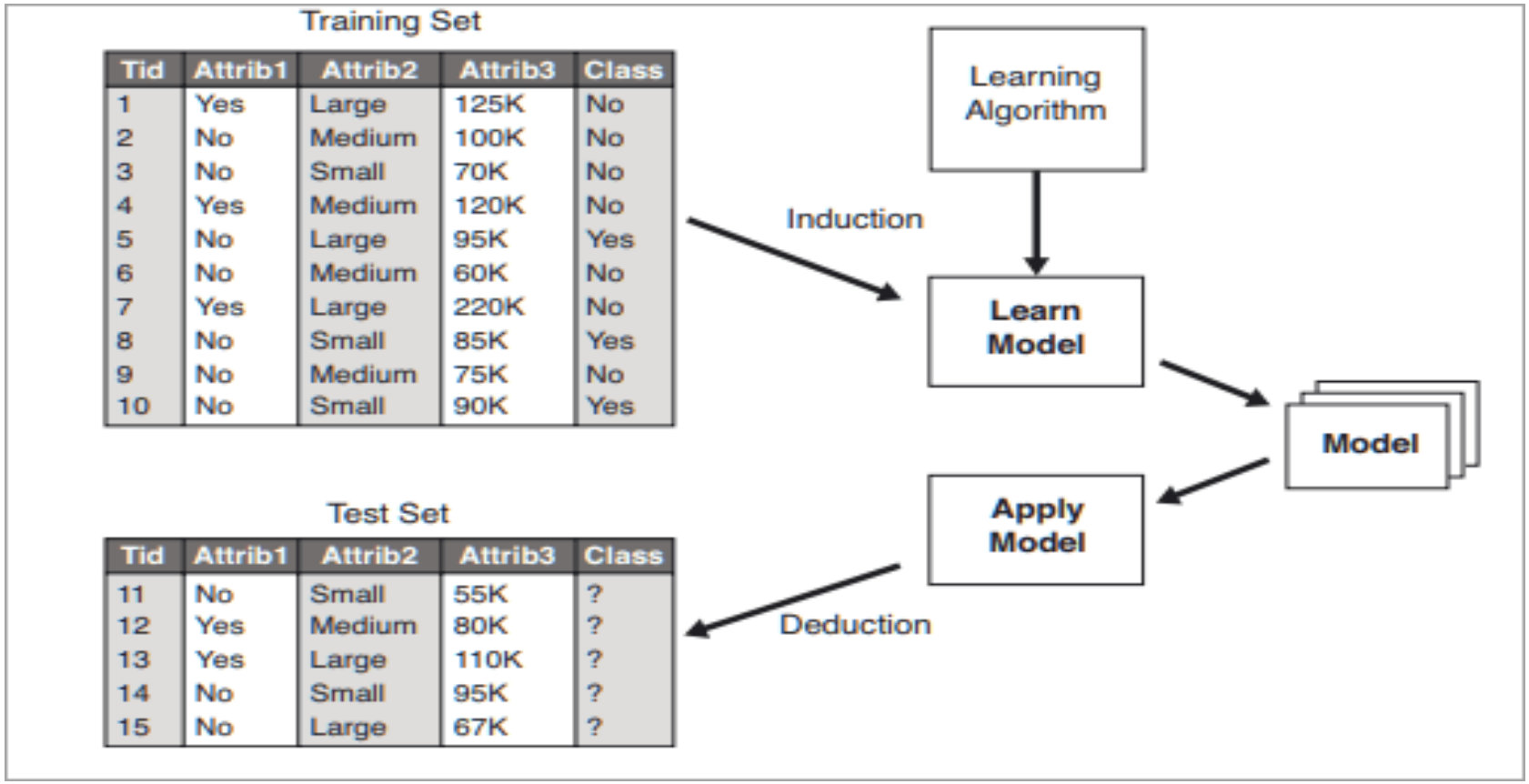
Decision Trees

- 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)
     - Which parts of the tree to remove
- Most popular DT algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

Decision Trees

Source:
https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/

Decision Trees



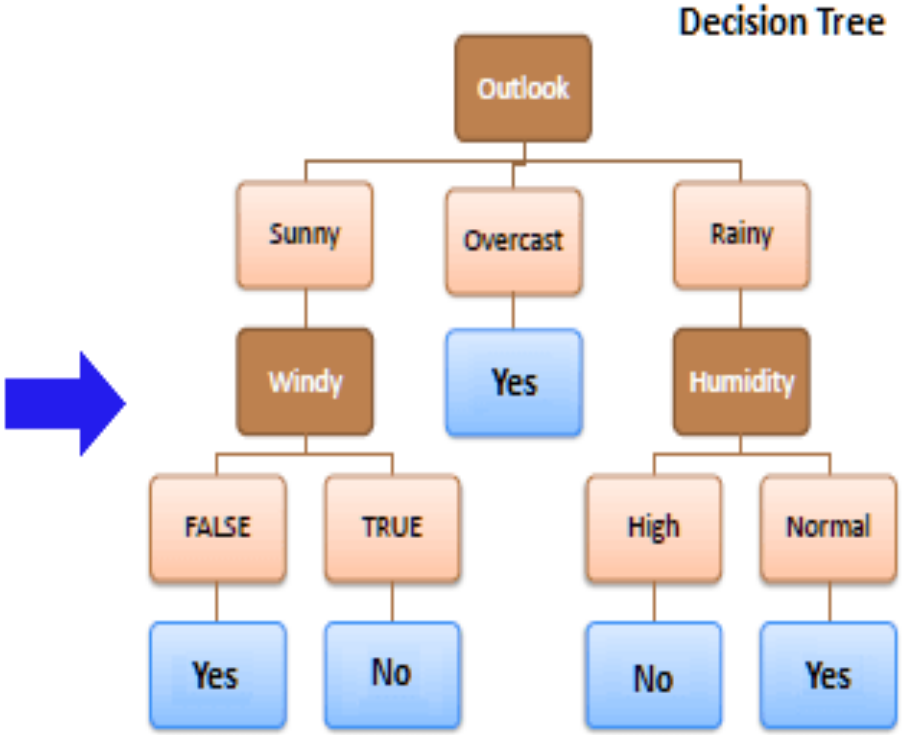
Source of image:
https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/

Decision Trees

- Example: Should I play golf or not?

Predictors | Target

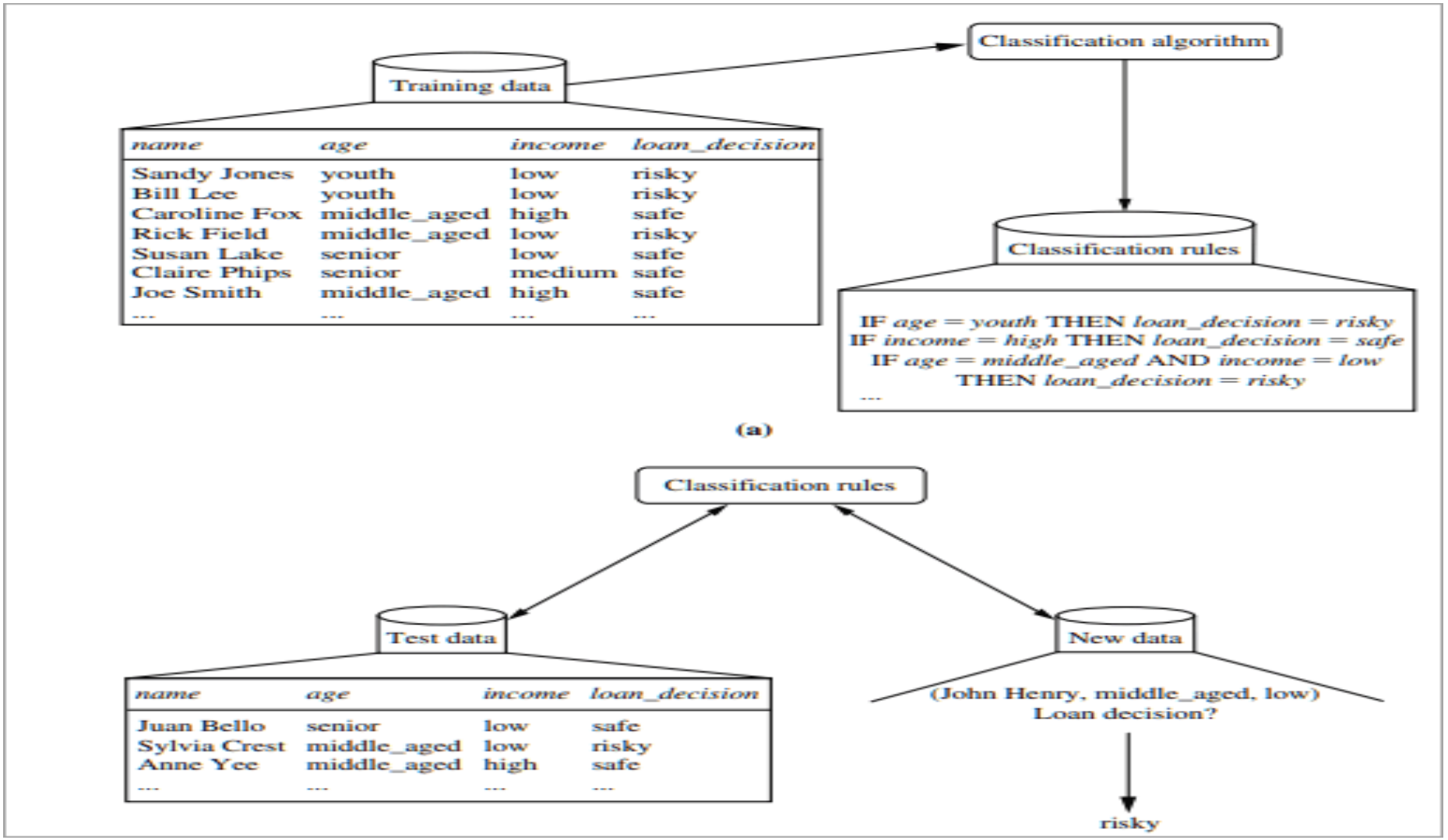
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Source of image:
https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/

Decision Trees

- Example: Should I give a loan or not?



Source of image:
https://www.softwaretestinghelp.com/decision-tree-algorithm-examples-data-mining/

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 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.

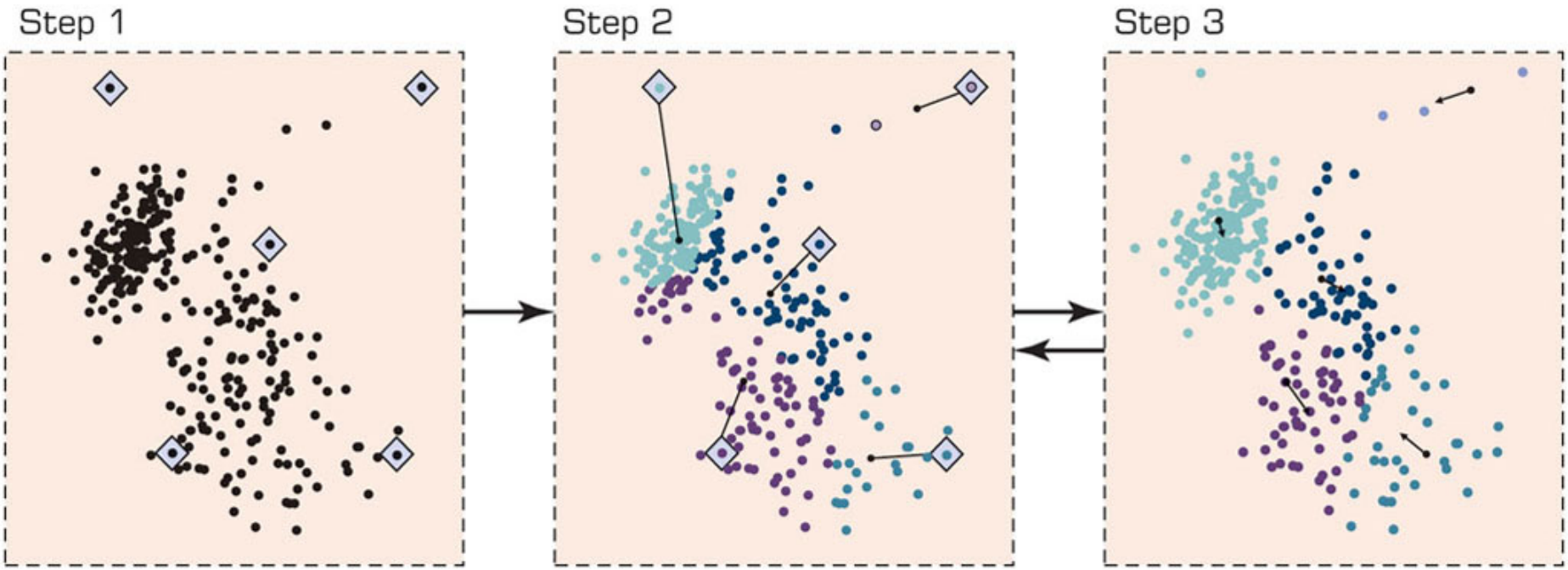
  Step 2: Assign each point to the nearest cluster center.

  Step 3: Re-compute the new cluster centers.

  Repetition step: Repeat steps 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

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



Association Rule Mining (1 of 7)

- A very popular D M 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 or affinity analysis

Association Rule Mining (2 of 7)

- 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 7)

- 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 (4 of 7)

- 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) ~ antecedent

Y: Right-hand-side (RHS) ~ consequent

S: Support: how often **X** and **Y** go together

$$\text{Supp}(X \rightarrow Y) = \frac{\text{(Number of baskets that contain both X \& Y)}}{\text{Total number of baskets}}$$

C: Confidence: how often **Y** go together with the **X**

$$\text{Confidence}(X \rightarrow Y) = \text{Supp}(X \rightarrow Y) / \text{Supp}(X)$$

Association Rule Mining (5 of 7)

Example:

In the total number of transactions data:

{Laptop Computer, Antivirus Software} ? {Extended Service Plan} [30%, 75%]

- laptops, antivirus software, and extended service plan were present in 30% of total transactions, and
- in cases where laptops and antivirus software were present also extended service plan was found 75% of the time.

If total transactions = 100,

- Number of times laptops, antivirus, and extended service plan were found together 30 times
- If number of laptops and antivirus (together) were found in 40 transactions, for these 40 times, it was found that extended service plan was ALSO present 30 times: 30/40 = 75%

Association Rule Mining (6 of 7)

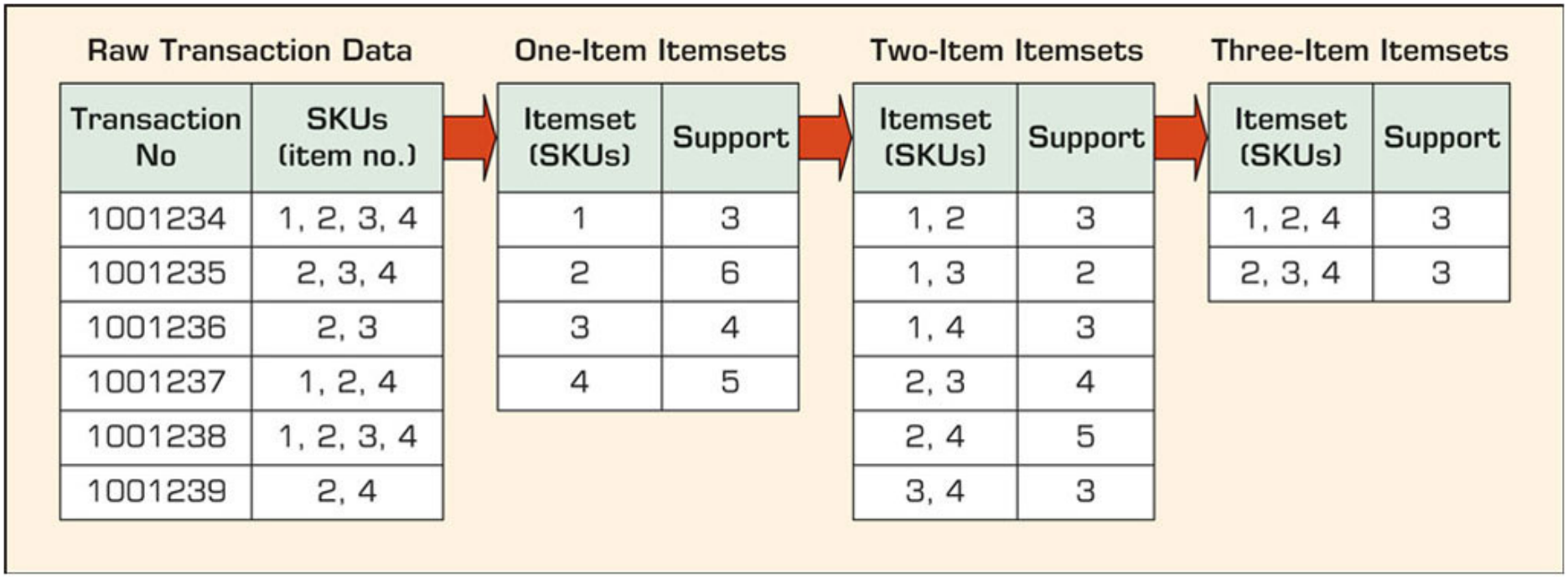
- Several algorithms are developed for discovering (identifying) association rules
  - Apriori
  - Eclat
  - F P-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules

Association Rule Mining (7 of 7)

- Apriori Algorithm
  - Finds subsets that are common to at least a minimum number of the itemsets (support value)
  - 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 value. (*see the figure next slide*)

Association Rule Mining Apriori Algorithm

Figure 4.14 A Graphical Illustration of the Steps in the *Apriori* Algorithm.



Raw Transaction Data

Transaction No	SKUs (item no.)
1001234	1, 2, 3, 4
1001235	2, 3, 4
1001236	2, 3
1001237	1, 2, 4
1001238	1, 2, 3, 4
1001239	2, 4

One-Item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-Item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-Item Itemsets

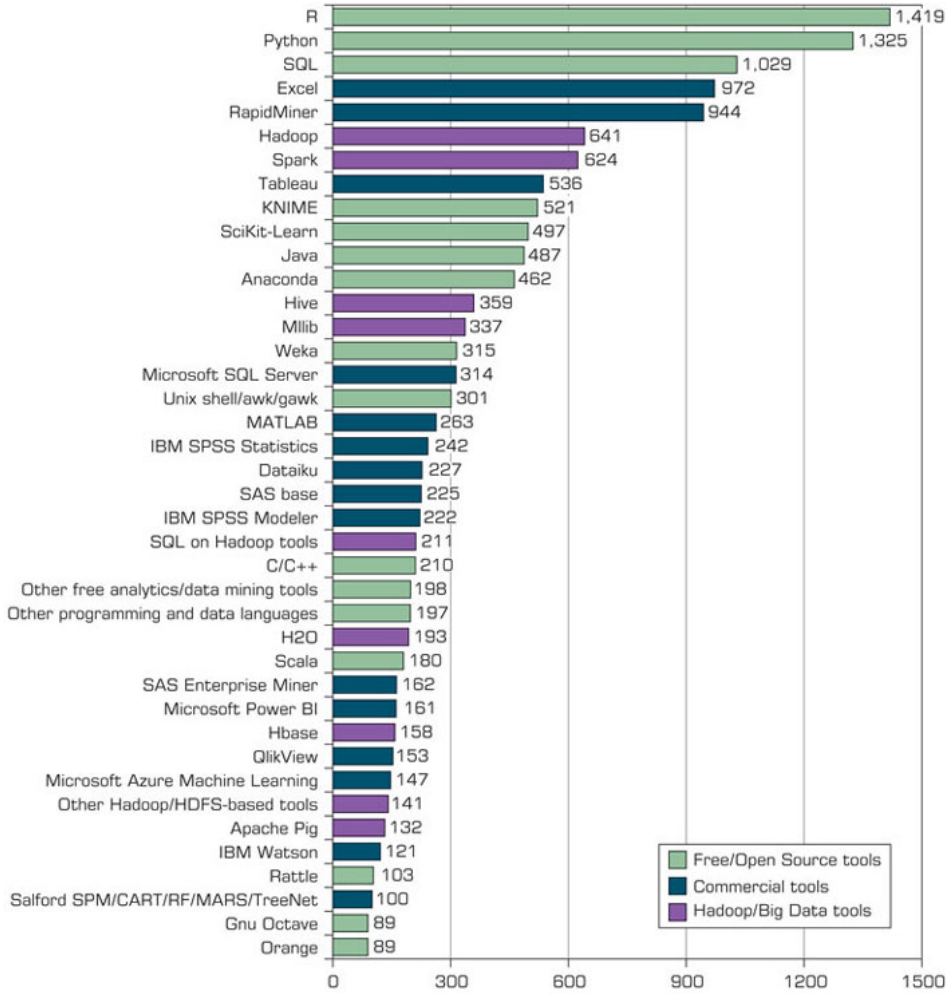
Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Set minimum support value in this example to be 3

Data Mining Software Tools

Figure 4.15 Popular Data Mining Software Tools (Poll Results).

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



Source: Used with permission from **KDnuggets.com**.

Application Case 4.6 (1 of 4)

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 4)

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

A Typical Classification Problem

Table 4.3 Movie Classification based on Receipts

Class No.	1	2	3	4	5	6	7	8	9
Range (in millions of dollars)	>1 (Flop)	>1 <610	>10 <20	>20 <640	>40 <665	>65 <6100	>100 <6150	>150 <6200	>200 (Blockbuster)

Table 4.4 Summary of Independent Variables

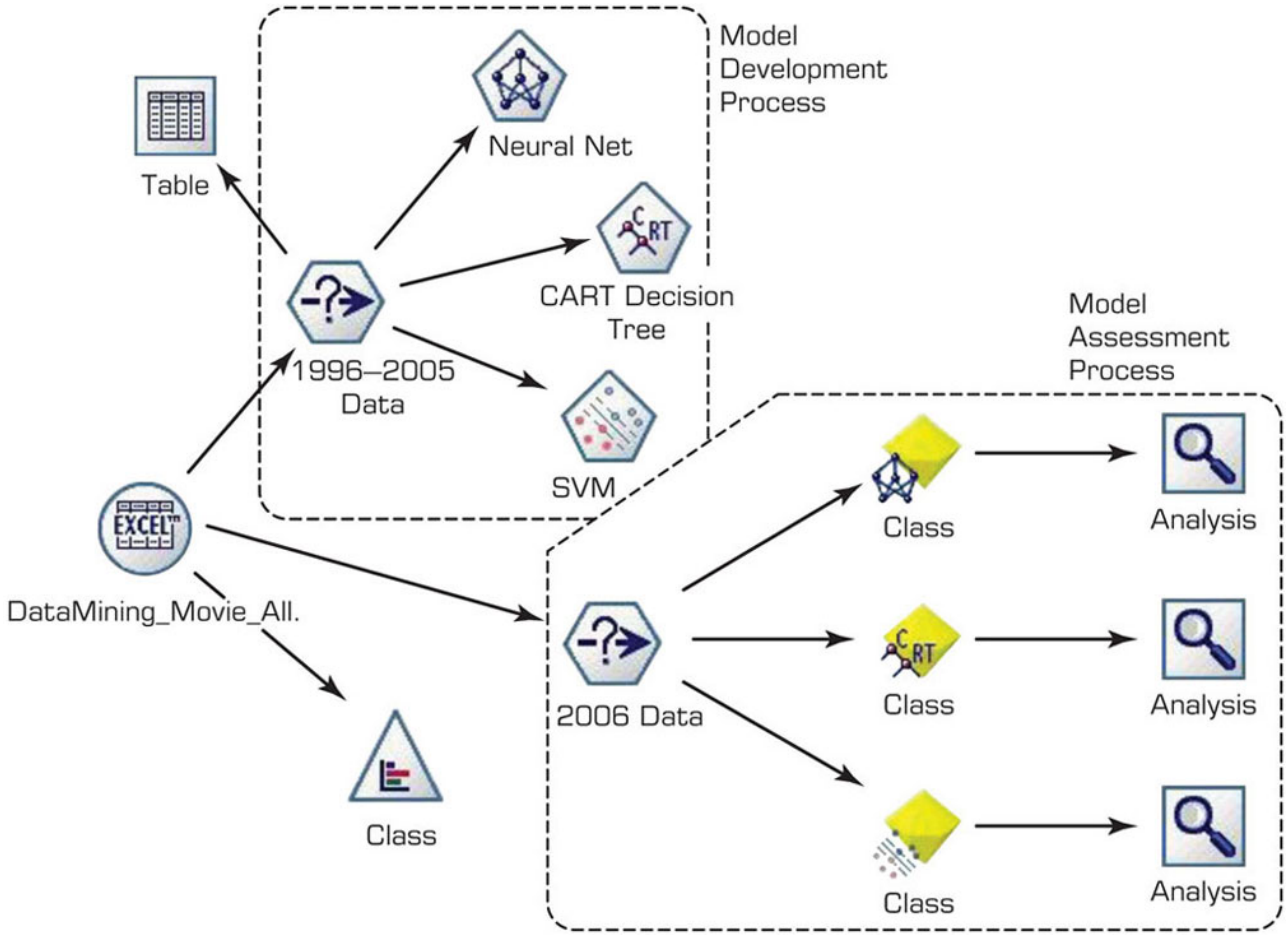
Independent Variable	Number of Values	Possible Values
MPA A Rating	5	G, PG, PG-13, R, NR
Competition	3	High, medium, low
Star value	3	High, medium, low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, medium, low
Sequel	2	Yes, no
Number of screens	1	A positive integer between 1 and 3,876

Application Case 4.6 (3 of 4)

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

FIGURE 4.16 Process Flow Screenshot for the Box-Office Prediction System.

The DM Process Map in IBM SPSS Modeler



Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

Application Case 4.6 (4 of 4)

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

TABLE 4.5 Tabulated Prediction Results for Individual and Ensemble Models

	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%	76.88%	89.60%	84.10%	**90.75%**
Standard deviation	0.93	0.87	1.05	0.76	0.84	**0.63**

Copyright