

Decision Support and Business Intelligence Systems

(9th Ed., Prentice Hall)



Chapter 5: Data Mining for Business Intelligence



Why Data Mining?

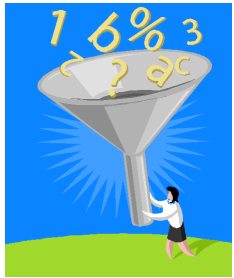
- More intense competition at the global scale
- Recognition of the value in data sources
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses
- The exponential increase in data processing and storage capabilities; and decrease in cost

Definition of Data Mining



- The **nontrivial** (meaning involved) **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,...





Data Mining at the Intersection of Many Disciplines



Data Mining

Characteristics/Objectives

- Source of data for DM is often (but not always) a 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.

Data Types in Data Mining

- **Categorical Data** (Specific grouping : Categorical Variables: Discrete, not calculable, no fraction but sub groups)

(Examples: race, sex, age group, education levels)

- **Nominal:**

- (*marital status*: 1. single, 2. married, 3. Widowed, 4. divorced)
- *Performance rating*: 1. poor, 2. acceptable, 3. good, 4. Excellent, 5. Exemplary)

- **Ordinal:**

- (*credit*: high, medium, low,
- *Age*: child, young, middle age, old
- *Education*: high school, JC, undergrad, graduate)

- **Numerical Data** (numeric, can be continuous, can have fractions)

(**Credit score,**

(**Age: in yeas**

- **Interval (scale) data**

- (temperature: 0-100 Celsius ~ 32-212 Fahrenheit)
- (Customer inter-arrival time)

- **Ratio data**

- (mass, angle, energy – relative to a non-arbitrary base: absolute zero -273.15 Celsius)

– -

Time /date

Text

Image

audio



What Does DM Do?

- DM extract patterns from data
 - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
 - Association (**dipper & baby food**)
 - Prediction (**weather forecasting**)
 - Cluster (segmentation) [**age-group behavior, certain crime location and demographic**]
 - Sequential (or time series) relationships [**does drug use leads to steeling?**]



Data Mining Tasks (cont.)

- Time-series forecasting
 - Part of sequence or link analysis?
- Visualization
 - In connection to any data mining task
- Types of DM
 - **OLD:** Hypothesis-driven data mining
 - **New:** Discovery-driven data mining
(the foundation of machine-learning)



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 and 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

- 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.

Highly popular application
area for data mining



Data Mining Process

- 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

Data Mining Process: CRISP-DM

Step 1: Business Understanding

(the goal of the mining? Customer attrition, why?)

Step 2: Data Understanding

(* What data are valuable -- 'abstraction'? “)

(* Recall: Ayati: barrier of observability and measurability”)

(*Retail: female-summer clothing line:

“female customer data: e.g. zip-code,
credit card, age group, etc.?)

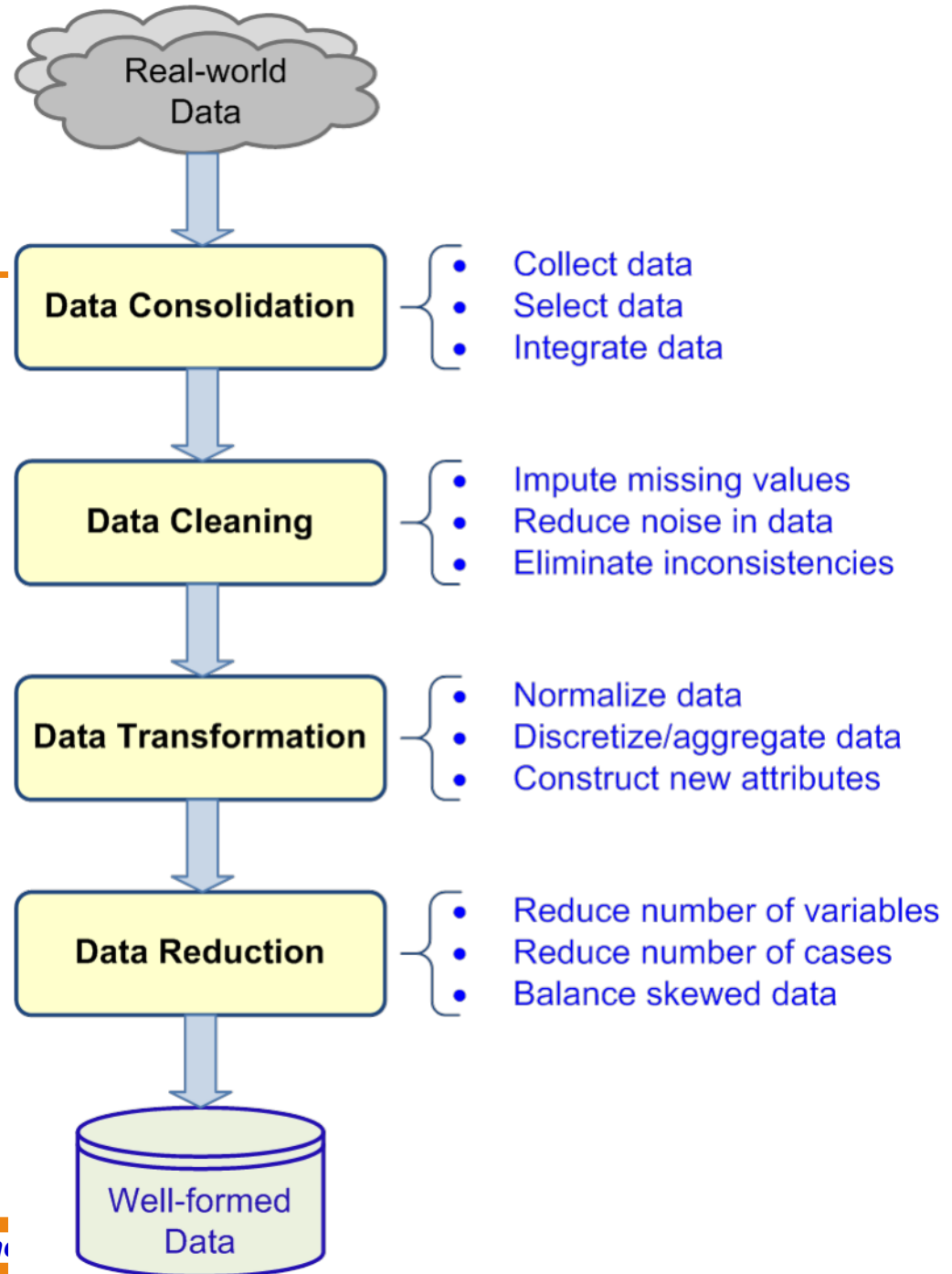
Step 3: Data Preparation (!)

See next slide

Accounts for ~85% of total project time

Data Preparation

A Critical DM Task



Data Mining Process: CRISP-DM (con't)

Step 4: Model Building (means using a variety of methods)

Step 5: Testing and Evaluation

Step 6: Deployment

- The process is highly repetitive and experimental (DM: art versus science?)





Data Mining Process: SEMMA



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

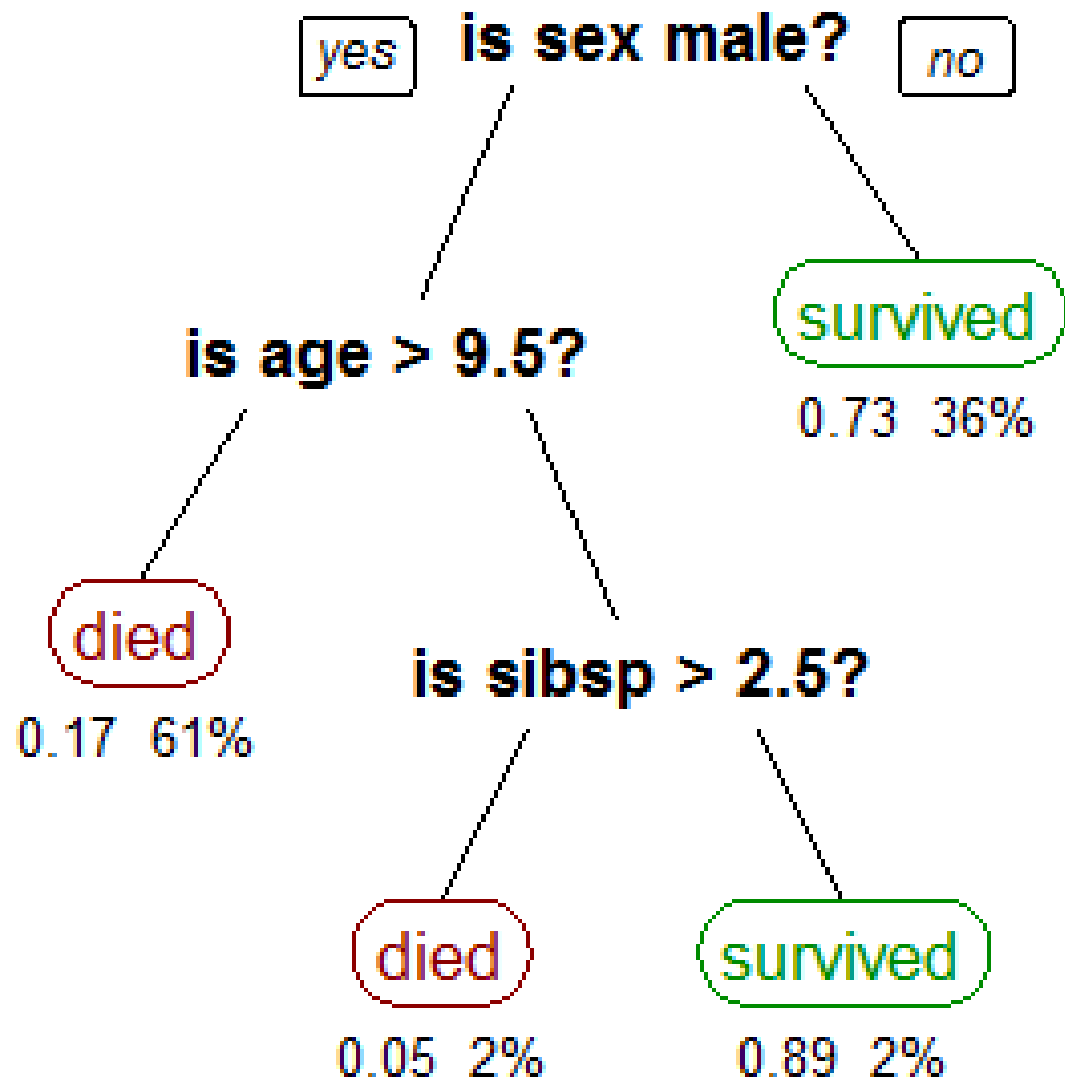
Predictive: Decision Tree*

- Identify the factors driving customer behavior and predict future behavior

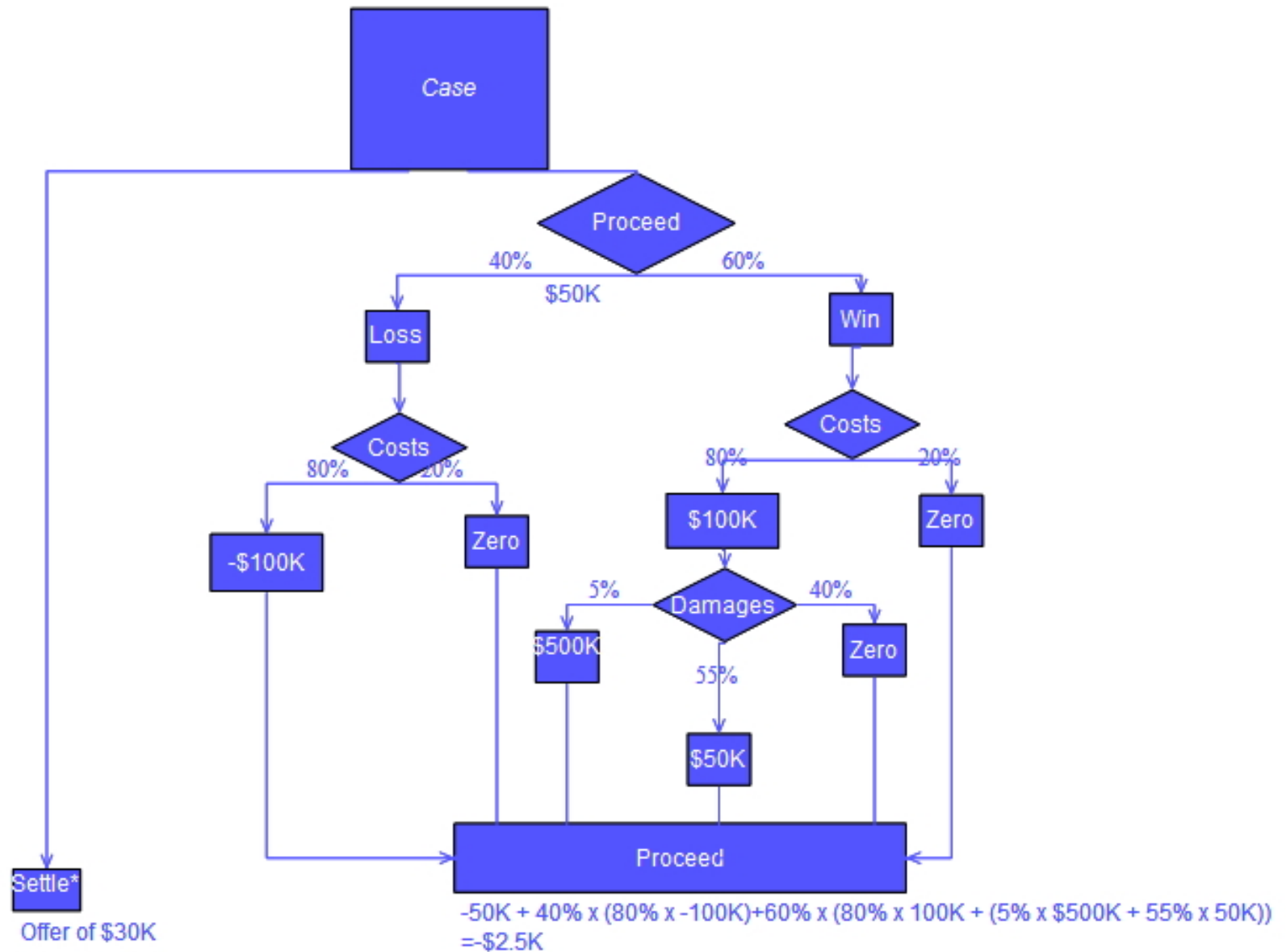
	Customer	Income	Age	Credit Rating	Etc.	Buying Behavior
Customers - Historical Data (query)	Mick Jones	\$ 100000	48	Excellent	...	Yes
	Elton Brown	\$ 130000	22	Fair	...	No
	Jack Turner	\$ 118000	36	Excellent	...	Yes
	Etc.
How will other Customers behave? New Data (query)	Willie Nelson	\$ 165000	34	Fair	...	
	Carol Lee	\$ 80000	63	Excellent	...	
	Etc.	
						

*Ayati: This example shows the common features of Decision Tree and Decision Table, which is the underlying principle of **Expert Systems**

A tree showing survival of passengers on the [Titanic](#) ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf.



Source:



Source:



Cluster Analysis for Data Mining

- 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 variable
- 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 - *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!”



Data Mining Software

■ Commercial

- SPSS - PASW (formerly Clementine)
- SAS - Enterprise Miner
- IBM - Intelligent Miner
- StatSoft – Statistical Data Miner
- ... many more

■ Free and/or Open Source

- Weka
- RapidMiner...

Source: KDNuggets.com, May



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 the good-old statistics



onomy for Mining Tasks



End of the Chapter

- Questions / Comments...