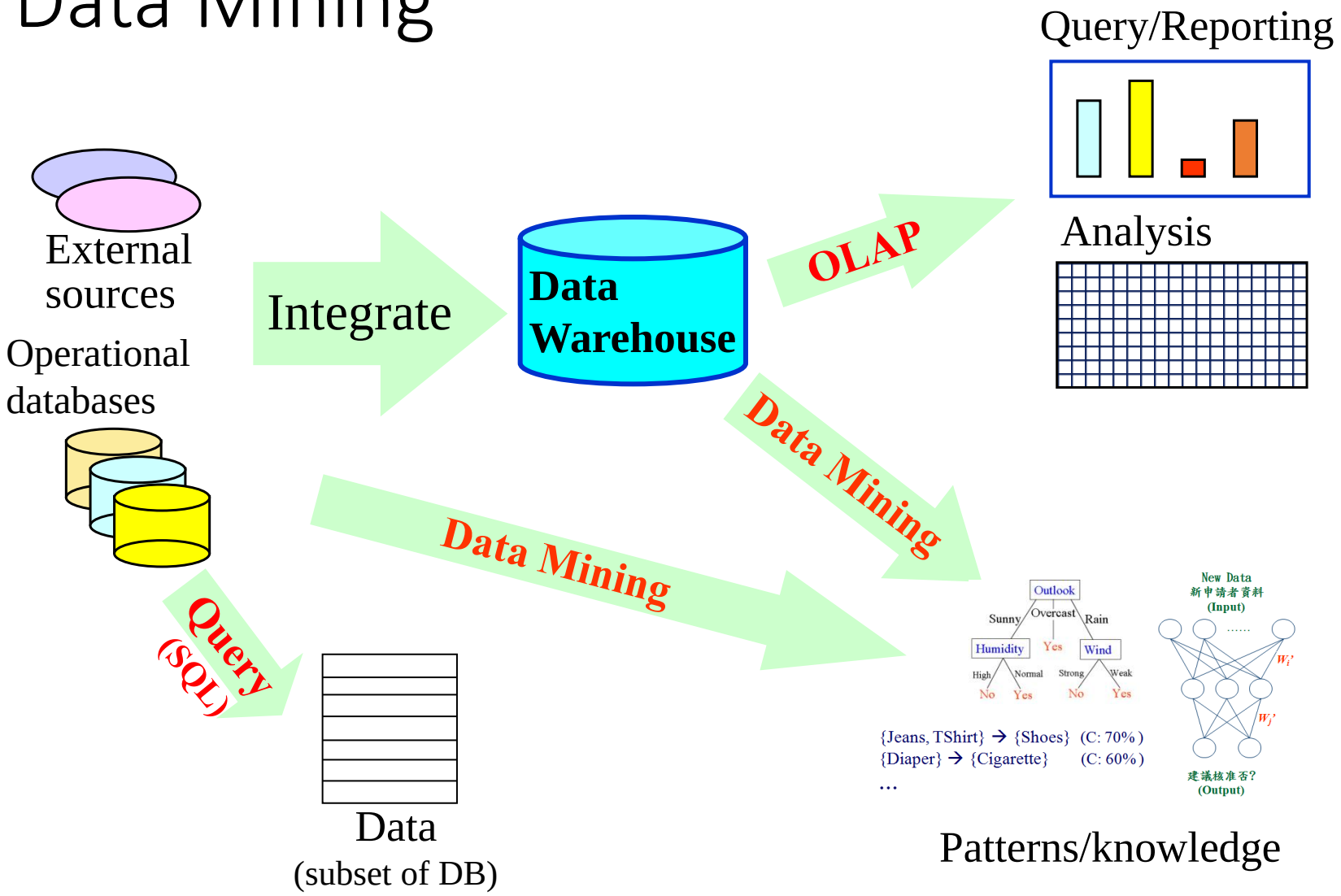
Introduction to Data Mining

- Database, Data Warehouse, Data Mining
- Knowledge Discovery in Data
- Functionality and Tasks of Data Mining
- Popular Data Mining Methods
- Potential Applications
- Challenges

Database, Data Warehouse, Data Mining



Database, Data Warehouse, Data Mining

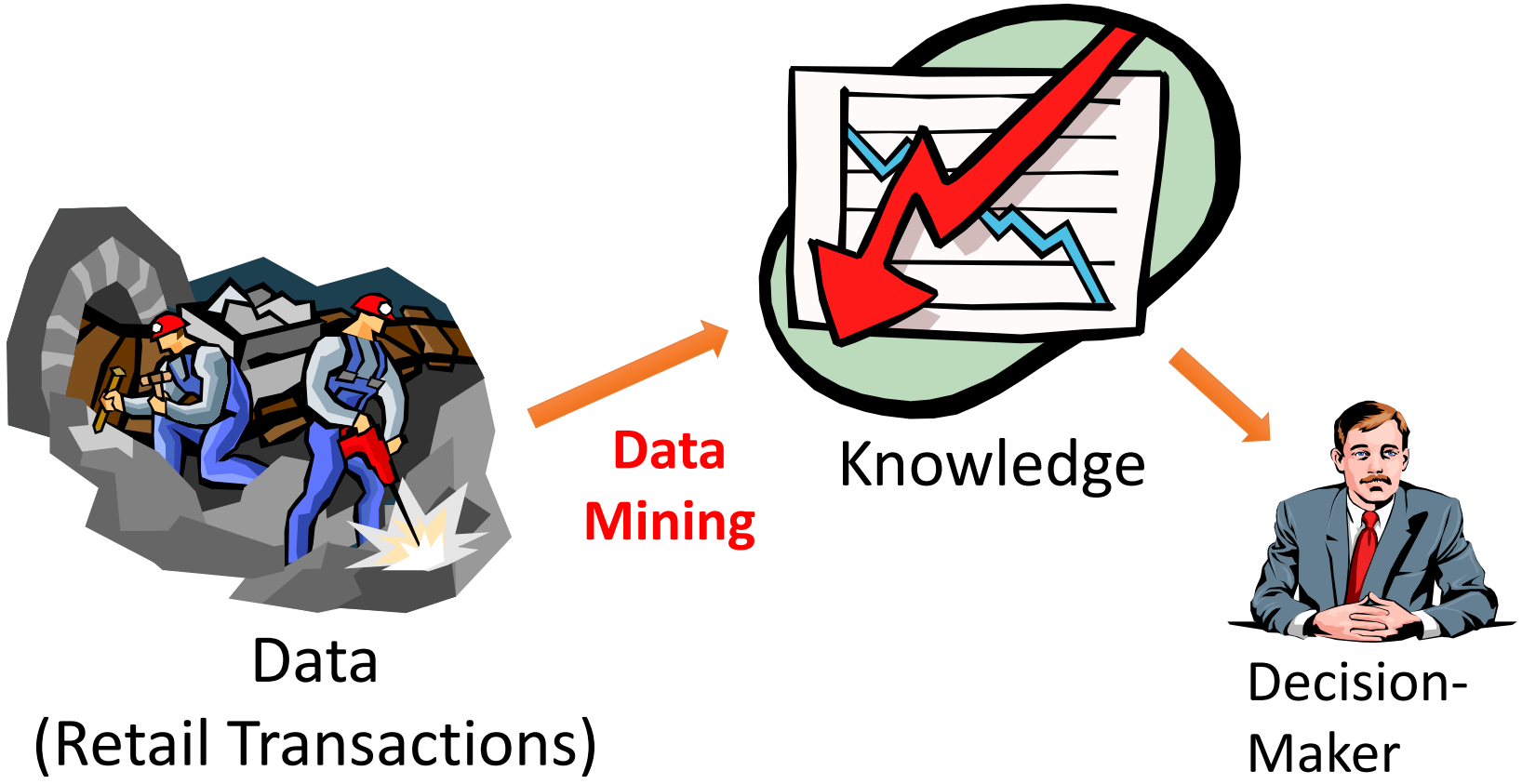
- Identify customers who have purchased more than $10,000 today.
  - DB Query: well defined; SQL
  - DB Output: precise; subset of database
- Analyze last-year's purchase amount of customers on drink in each region of North, Central, South, and East Taiwan.
  - DW Query: well defined; OLAP
  - DW Output: precise; aggregation of subset of database
- Identify customers with similar buying habits. (Clustering)
  - DM Query: poorly defined; no precise query language
  - DM Output: fuzzy; not a subset of database

More Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Find all customers who have purchased milk.
- Data Warehouse
  - Analyze the last-month's sales of drink in each county.
  - Rank the last-year's sales of drink among regions.
- Data Mining
  - Find all credit applicants who are poor credit risks. (Classification)
  - Find all items which are frequently purchased with milk. (Association rules)

Data Mining

To discover useful knowledge from massive data for supporting decision-making



Data: Retail Transactions

Transaction	Items
t_1	Blouse
t_2	Shoes,Skirt,TShirt
t_3	Jeans,TShirt
t_4	Jeans,Shoes,TShirt
t_5	Jeans,Shorts
t_6	Shoes,TShirt
t_7	Jeans,Skirt
t_8	Jeans,Shoes,Shorts,TShirt
t_9	Jeans
t_{10}	Jeans,Shoes,TShirt
t_{11}	TShirt
t_{12}	Blouse,Jeans,Shoes,Skirt,TShirt
t_{13}	Jeans,Shoes,Shorts,TShirt
t_{14}	Shoes,Skirt,TShirt
t_{15}	Jeans,TShirt
t_{16}	Skirt,TShirt
t_{17}	Blouse,Jeans,Skirt
t_{18}	Jeans,Shoes,Shorts,TShirt
t_{19}	Jeans
t_{20}	Jeans,Shoes,Shorts,TShirt

Data Mining:
Association Analysis

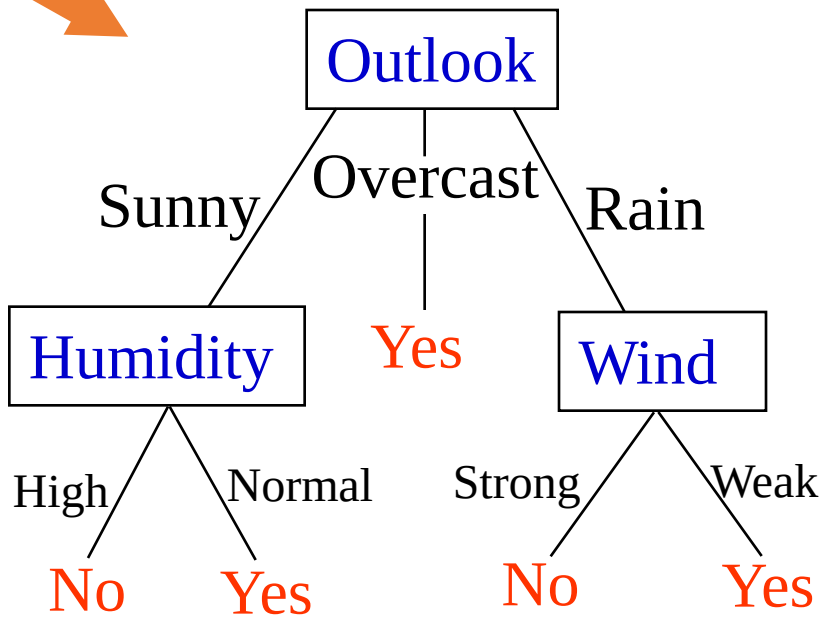
{Jeans, TShirt} → {Shoes} (*c*: 70%)
{Diaper} → {Beer} (*c*: 60%)

…

Decision Tree and Rules

No	Outlook	Temp.	Humidity	Wind	P.Tennis
01	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	**Overcast**	**Hot**	**High**	**Weak**	**Yes**
04	Rain	Mild	High	Weak	Yes
05	Rain	Cool	Normal	Weak	Yes
06	Rain	Cool	Normal	Strong	No
07	**Overcast**	**Cool**	**Normal**	**Strong**	**Yes**
08	Sunny	Mild	High	Weak	No
09	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	**Overcast**	**Mild**	**High**	**Strong**	**Yes**
13	**Overcast**	**Hot**	**Normal**	**Weak**	**Yes**
14	Rain	Mild	High	Strong	No

mining



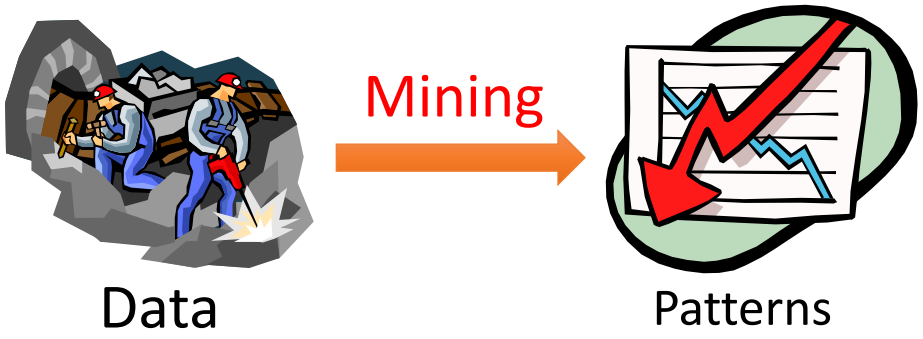
- If **Outlook** = Overcast, then **PlayTennis** = **Yes**
- If **Outlook** = Rain, …

Application: Classification
(Rain Cool High Weak) **?**

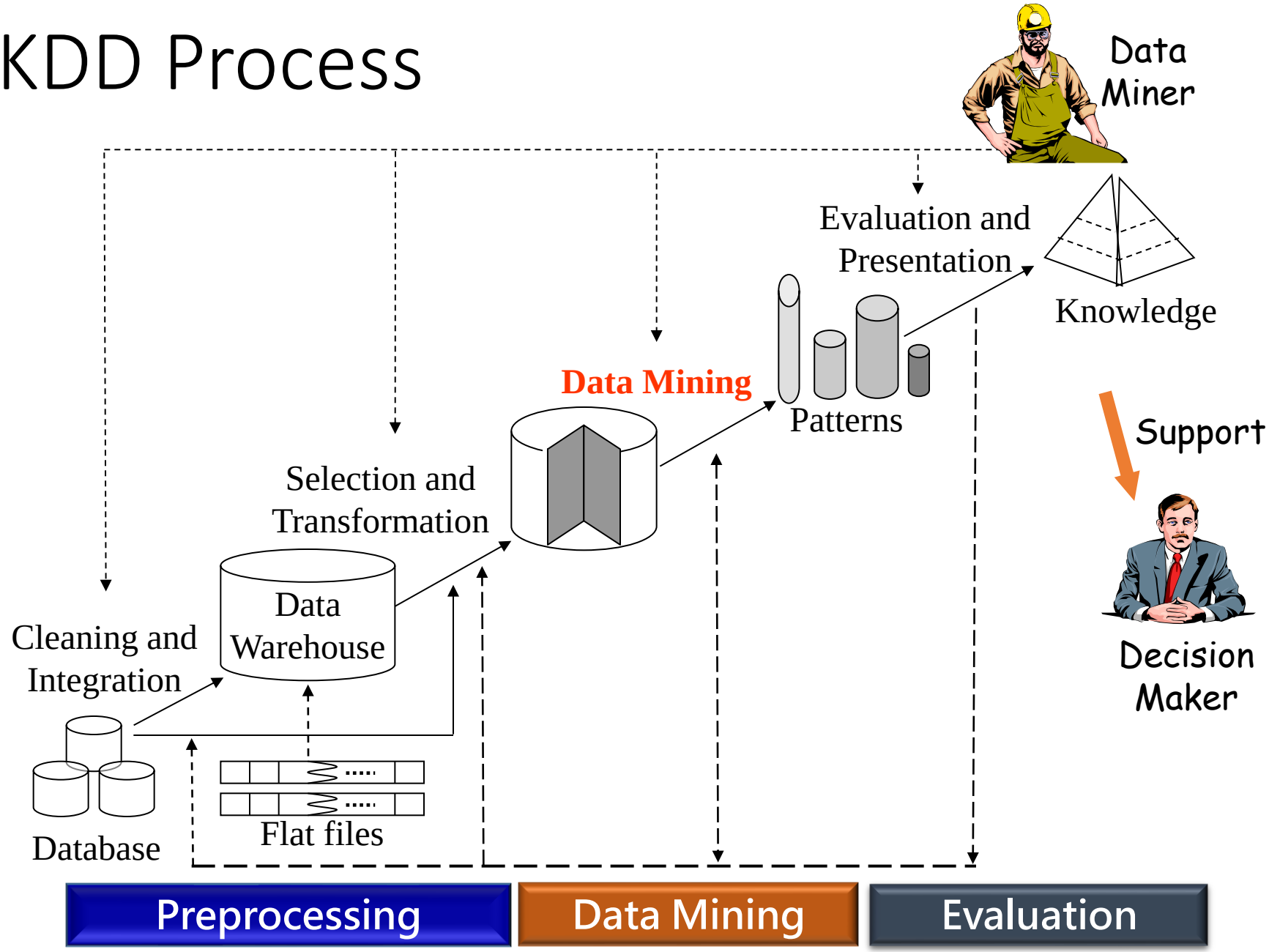
Knowledge Discovery in Data

- ***Non-trivial*** process of identifying
  valid,
  novel,
  potentially ***useful***, and
  ultimately ***understandable***

  patterns in data



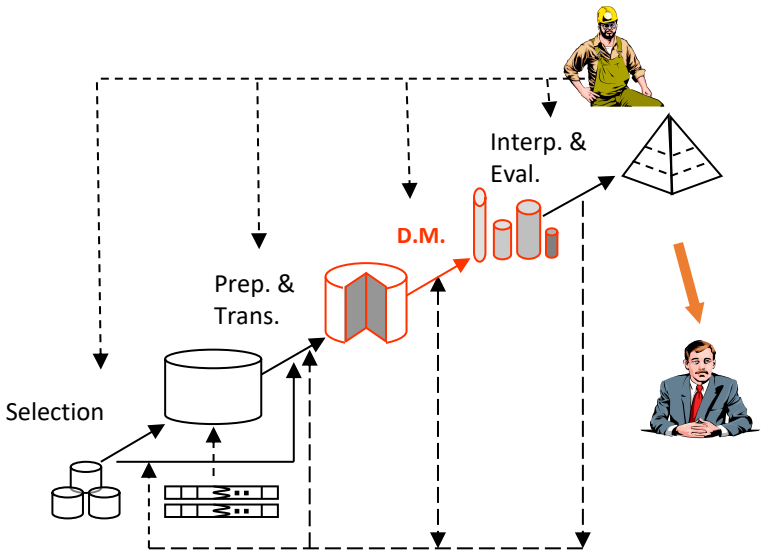
KDD Process


Data Miner
Evaluation and Presentation
Knowledge
Data Mining
Patterns
Support
Selection and Transformation
Data Warehouse
Cleaning and Integration
Decision Maker
Database
Flat files
Preprocessing
Data Mining
Evaluation

Example: Web Log Mining

- ***Selection:***
  - Select log data (dates and locations) to use
- ***Preprocessing:***
  - Remove identifying URLs
  - Remove error logs
- ***Transformation:***
  - Sessionize (sort and group)
- ***Data Mining:***
  - Identify frequently accessed sequences
- ***Interpretation/Evaluation:***
  - Display frequently accessed sequences.
- ***Potential User Applications:***
  - Cache prediction
  - Personalization

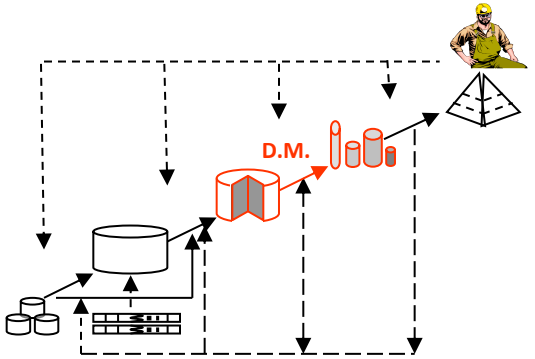
03/01 13:20 www.acm.org/index.html
03/01 13:22 www.acm.org/papers.html
03/01 13:22 www.acm.org/dl.html
…
…



A→C→E→H s = 50%
A→B→D→H s = 20%

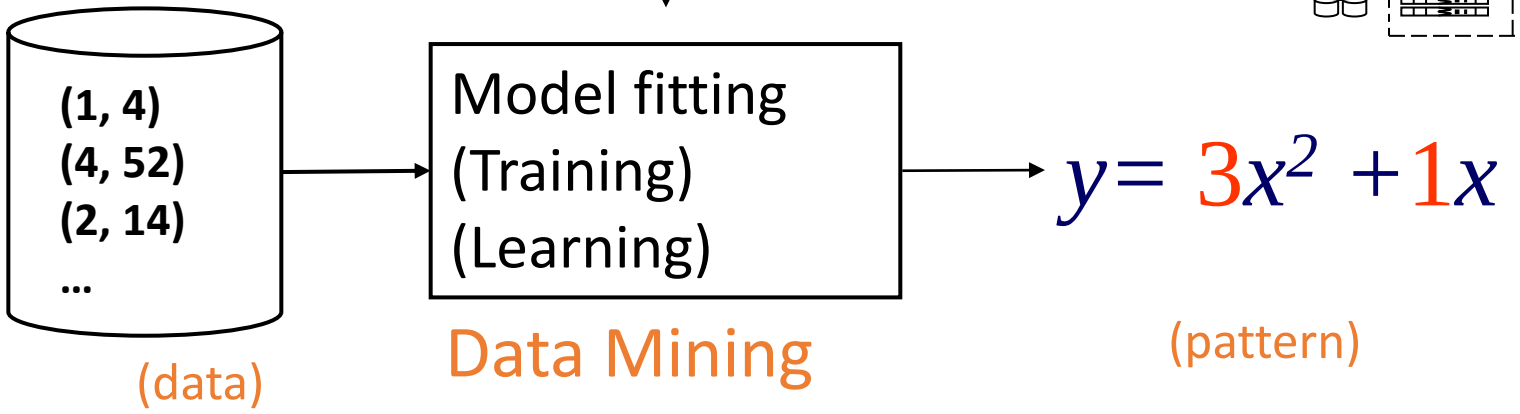
Data Mining

- a step in KDD process
  - involves fitting models to, or determining patterns from observed data
  - consisting of particular algorithms that, under some acceptable computational efficiency limitations, produces a particular enumeration of patterns



$$y = \alpha x^2 + \beta x$$

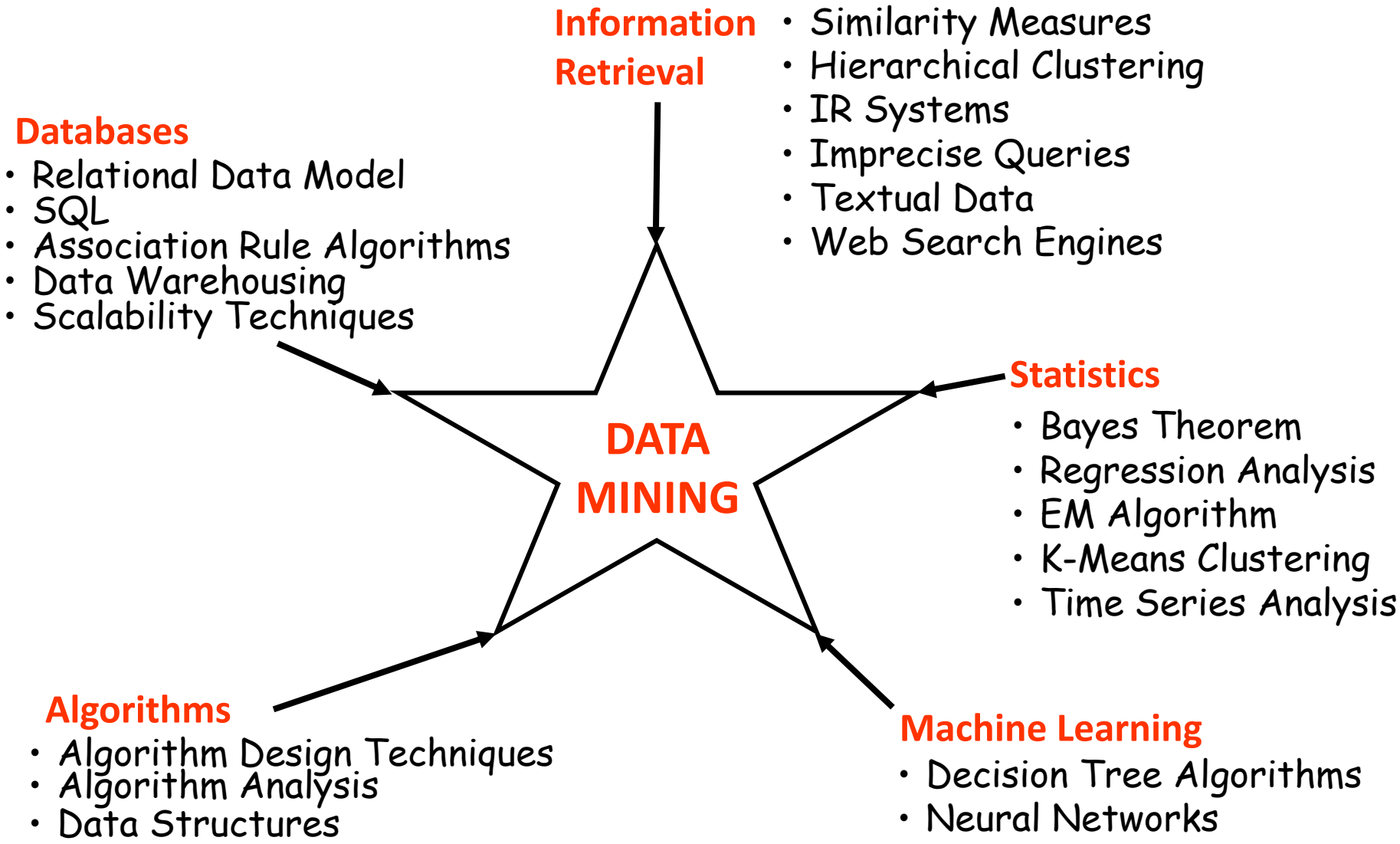
(model)



$$y = 3x^2 + 1x$$

(pattern)

Data Mining: Confluence of Multiple Disciplines

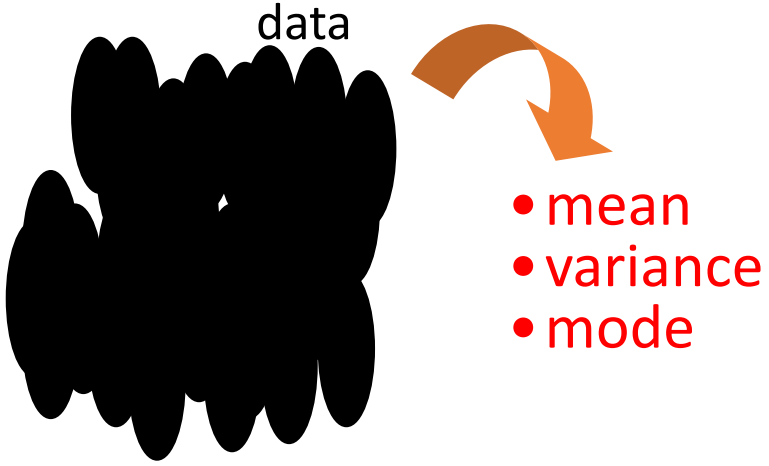


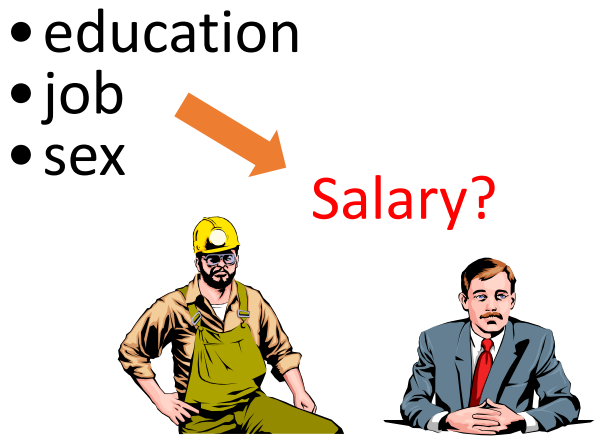
Functionality of Data Mining

- Description
  - Finding human-interpretable, compact patterns describing the data

    e.g. mean, variance, mode, distribution
- Prediction
  - Using some variables or fields in database to predict unknown or future values of other variables of interest

    e.g. education, job, sex ⇒ salary?





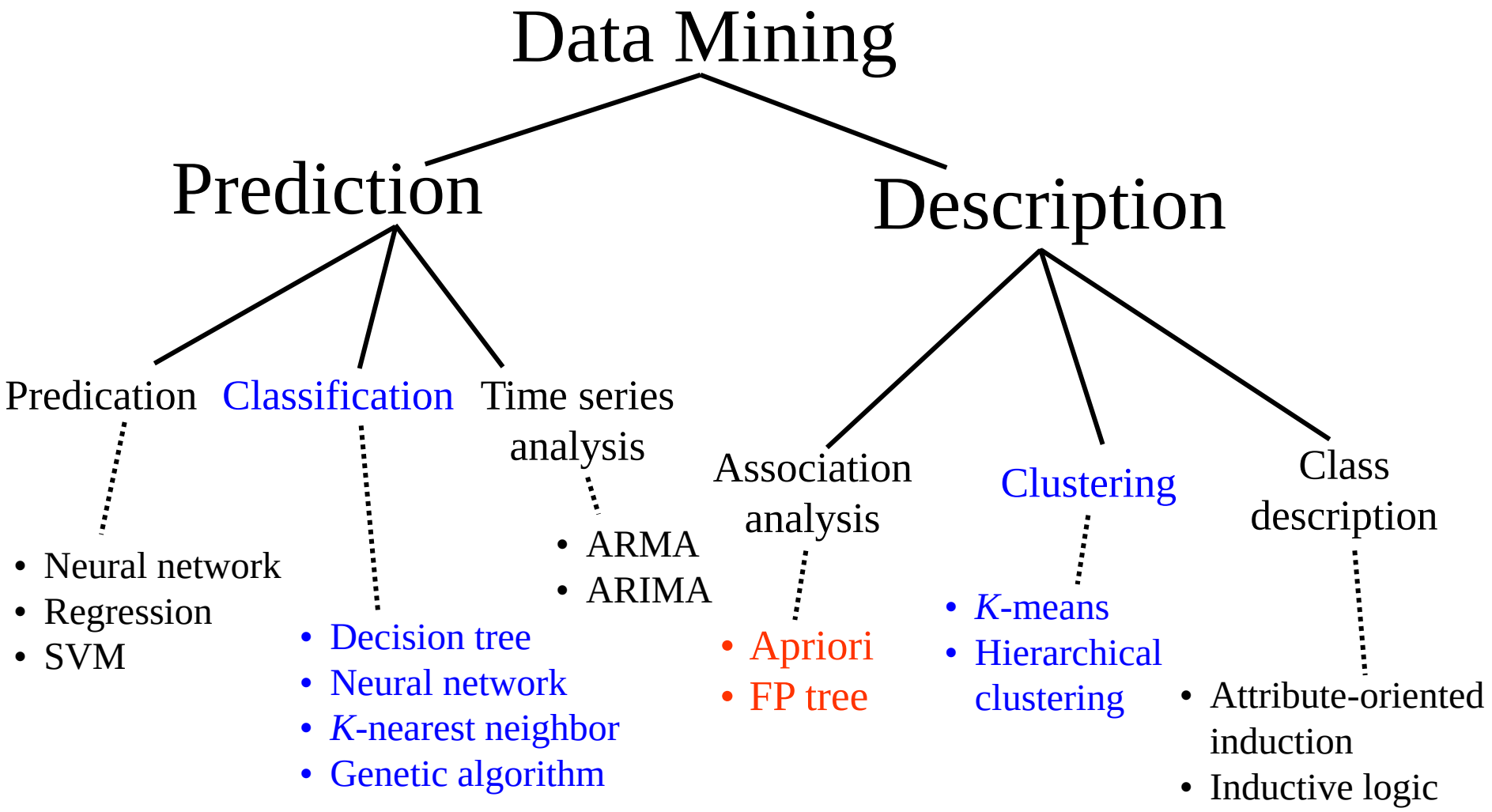
Primary Tasks of Data Mining

- Description
  - Cluster Analysis
  - Class Description
  - Association Analysis
- Prediction
  - Classification and Prediction
  - Outlier Analysis
  - Time Series Analysis

Popular Data Mining Methods

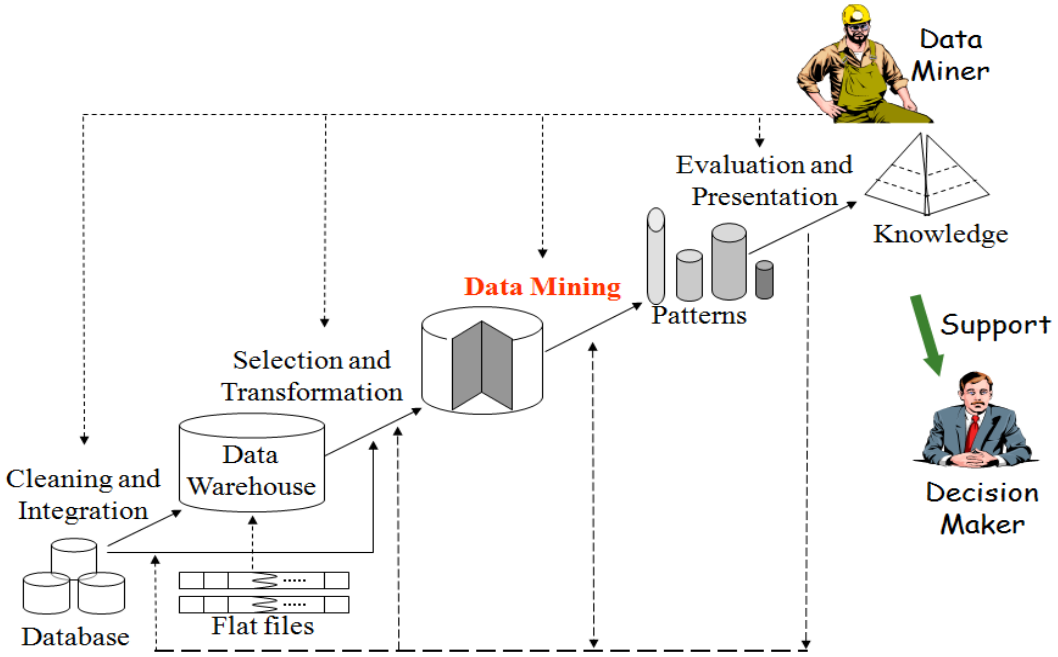
- Description
  - Cluster Analysis
    - Hierarchical clustering, *k*-means clustering
  - Class Description
    - Attribute-oriented induction, inductive logic
  - Association Analysis
    - Apriori, FP tree
- Prediction
  - Classification and Prediction
    - *K*-nearest neighbors, decision tree and rules, neural networks, genetic algorithm, regression
  - Outlier Analysis
  - Time Series Analysis

Tasks and Methods of Data Mining



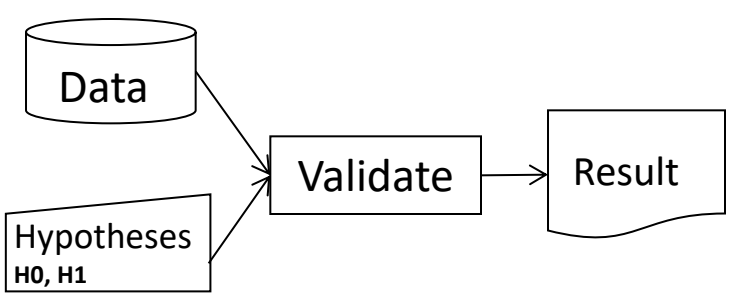
Potential for KDD Applications

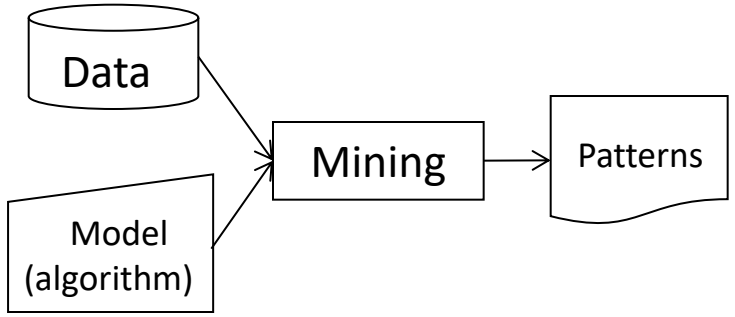
- Suitable Domains
  - information-rich
  - have a changing environment
  - not already have existing models
  - require knowledge-based decisions
  - high payoff for the right decisions



KDD and Statistics

- Traditional Data Analysis
  - Assumption driven
  - A hypothesis is manually formed and validated (by statistical means) against the data
  - E.g., H0: high-income family has higher probability of owning a BMW car
- Knowledge Discovery
  - Discovery driven
  - Useful patterns are automatically extracted from the data
  - E.g., What products are usually bought together? (association analysis)





Challenges for KDD -1

- Huge Databases
  - can't fit in main memory at one time, e.g., POS data of 7-11
  - solutions: sampling, approximation methods, parallel processing, …
- High Dimensionality
  - increase size of search space for model induction in a combinatorially explosive manner
  - increase chances that learner will find spurious patterns that are not valid in general
  - solutions: use prior knowledge to identify irrelevant variables, ...

Name	Age	Height	Hobby	Job	…
Tom	14	100	...	...	...
Mary	24	160	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

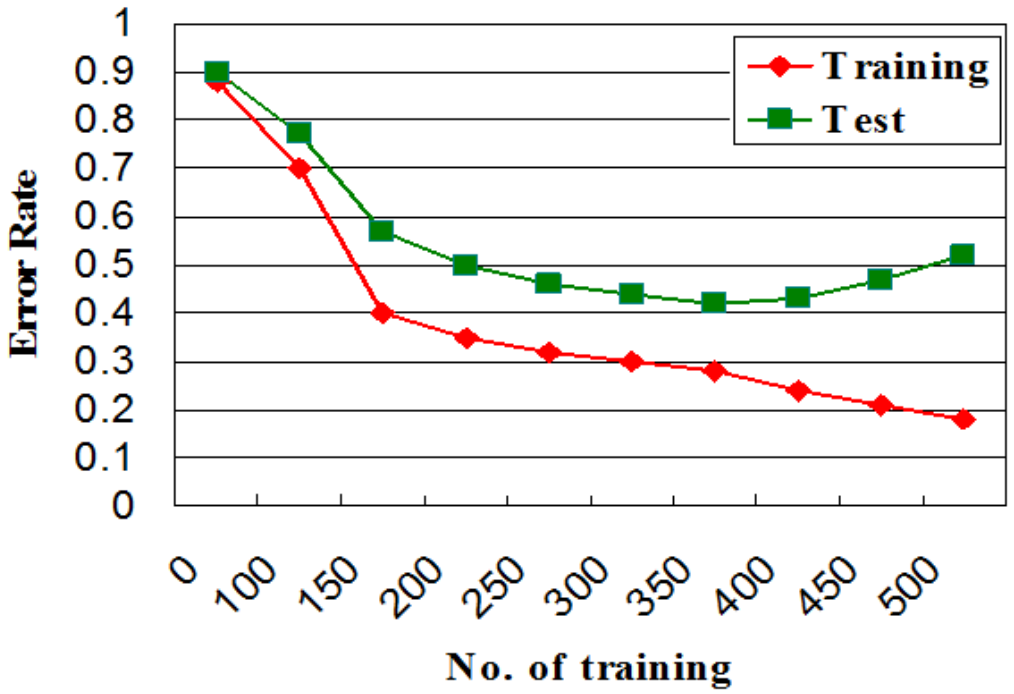
Challenges for KDD -2

- Changing Data and Knowledge
  - changing may make previously discovered patterns invalid
  - solutions: incremental methods for updating the patterns, …
- Missing and Noisy Data
  - solutions: statistical strategies to identify hidden variables and dependencies

Name	Age	Height	Hobby	⋯
Tom	14	100	Video Game	⋯
Mary	24		Golf	⋯
John	**200**	180	Play Wii	⋯
…	…	…	…	⋯

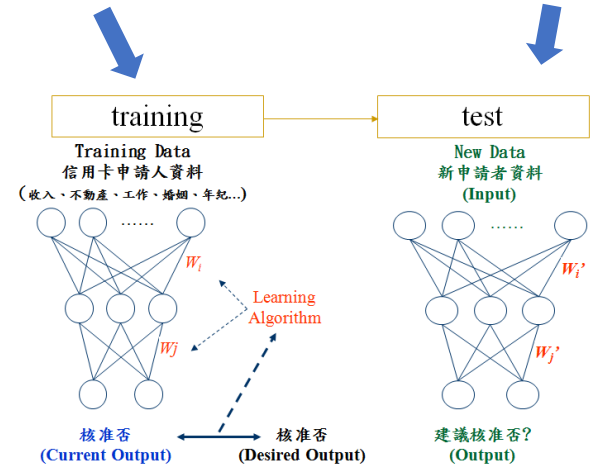
Challenges for KDD-3: Overfitting

- good performance on training data, but poor performance on real data
- solutions: cross-validation, regularization, other statistical strategies



No	Outlook	Temp.	Humidity	Wind	P.Tennis
01	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	Overcast	Hot	High	Weak	Yes
04	Rain	Mild	High	Weak	Yes
05	Rain	Cool	Normal	Weak	Yes
06	Rain	Cool	Normal	Strong	No
07	Overcast	Cool	Normal	Strong	Yes
08	Sunny	Mild	High	Weak	No
09	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

(Rain Hot High Strong) ?
(Rain Cool High Weak) ?
(Sunny Hot Low Weak) ?
(Rain Hot Low Strong) ?
…



References

- Books
  - Dunham, M., *Data mining: introduction and advanced topics*, Prentice Hall, 2003.
  - Kantardzic, M., *Data mining: concepts, models, methods, and algorithms*, John Wiley & Sons, 2003.
  - Han, J. and M. Kamber, *Data mining: concepts and techniques*, 2nd ed., Morgan Kaufmann, 2006.
- Journals
  - ACM Transactions on Knowledge Discovery from Data
  - IEEE Transactions on Knowledge and Data Engineering
  - IEEE Transactions on Neural Networks and Learning Systems
  - Data Mining and Knowledge Discovery
  - Neural Networks
  - Pattern Recognition
- Conferences
  - ACM International Conference on Knowledge Discovery in Data
  - IEEE International Conference on Data Mining