Tamkang University
淡江大學

Big Data Mining
巨量資料探勘

Tamkang University

關連分析
(Association Analysis)

1042DM03
MI4 (M2244) (3094)
Tue, 3, 4 (10:10-12:00) (B216)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/

2016-03-01

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2016/02/16	巨量資料探勘課程介紹 (Course Orientation for Big Data Mining)
2	2016/02/23	巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統 (Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem)
3	2016/03/01	關連分析 (Association Analysis)
4	2016/03/08	分類與預測 (Classification and Prediction)
5	2016/03/15	分群分析 (Cluster Analysis)
6	2016/03/22	個案分析與實作一 (SAS EM 分群分析)：Case Study 1 (Cluster Analysis – K-Means using SAS EM)
7	2016/03/29	個案分析與實作二 (SAS EM 關連分析)：Case Study 2 (Association Analysis using SAS EM)

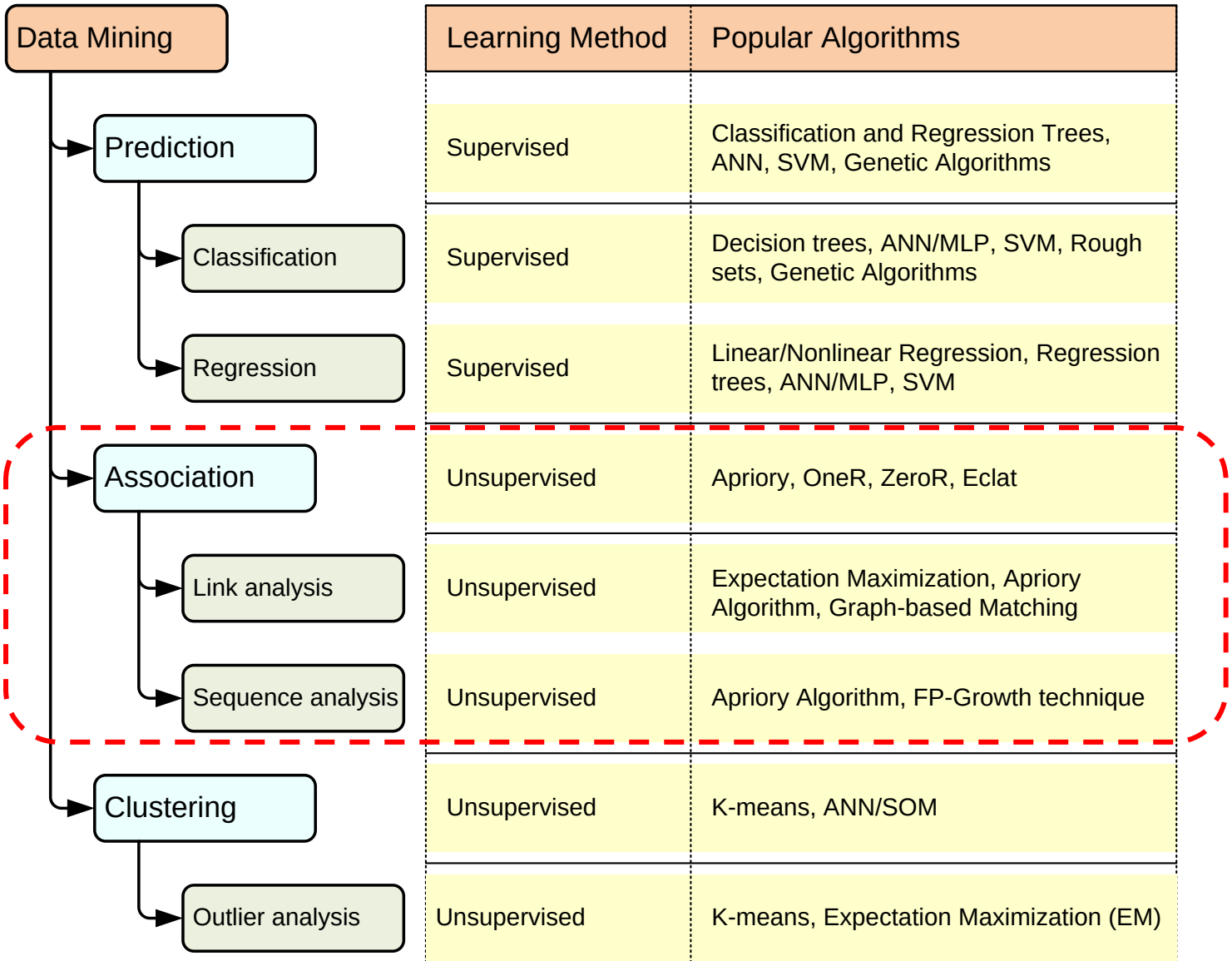
課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
8	2016/04/05	教學行政觀摩日 (Off-campus study)
9	2016/04/12	期中報告 (Midterm Project Presentation)
10	2016/04/19	期中考試週 (Midterm Exam)
11	2016/04/26	個案分析與實作三 (SAS EM 決策樹、模型評估)：Case Study 3 (Decision Tree, Model Evaluation using SAS EM)
12	2016/05/03	個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)
13	2016/05/10	Google TensorFlow 深度學習 (Deep Learning with Google TensorFlow)
14	2016/05/17	期末報告 (Final Project Presentation)
15	2016/05/24	畢業班考試 (Final Exam)

Outline

- Big Data Analytics Lifecycle
- Data Mining Process
- Data Mining
- Association Analysis
- Apriori algorithm
  - Frequent Itemsets
  - Association Rules

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)

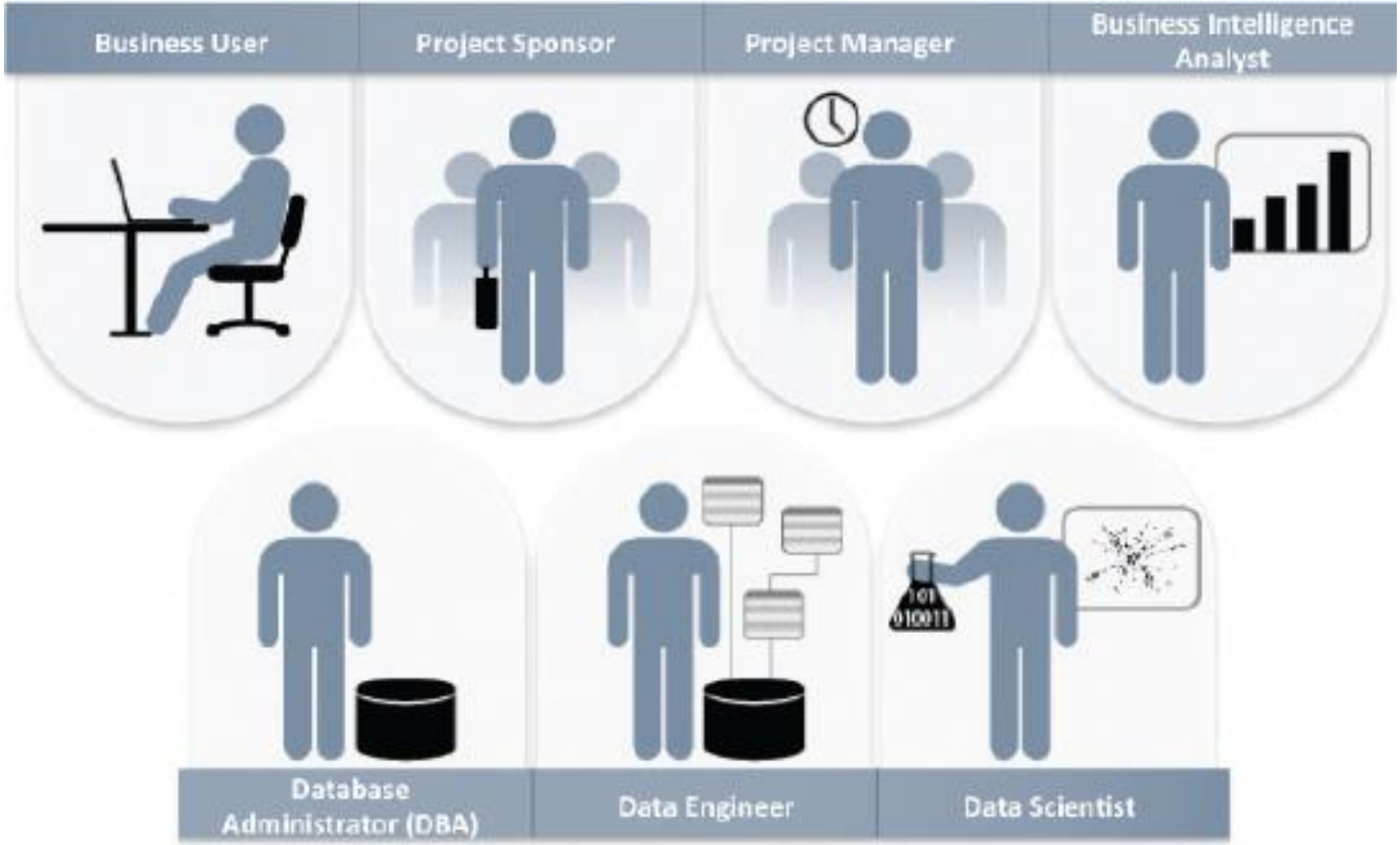
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Transaction Database

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

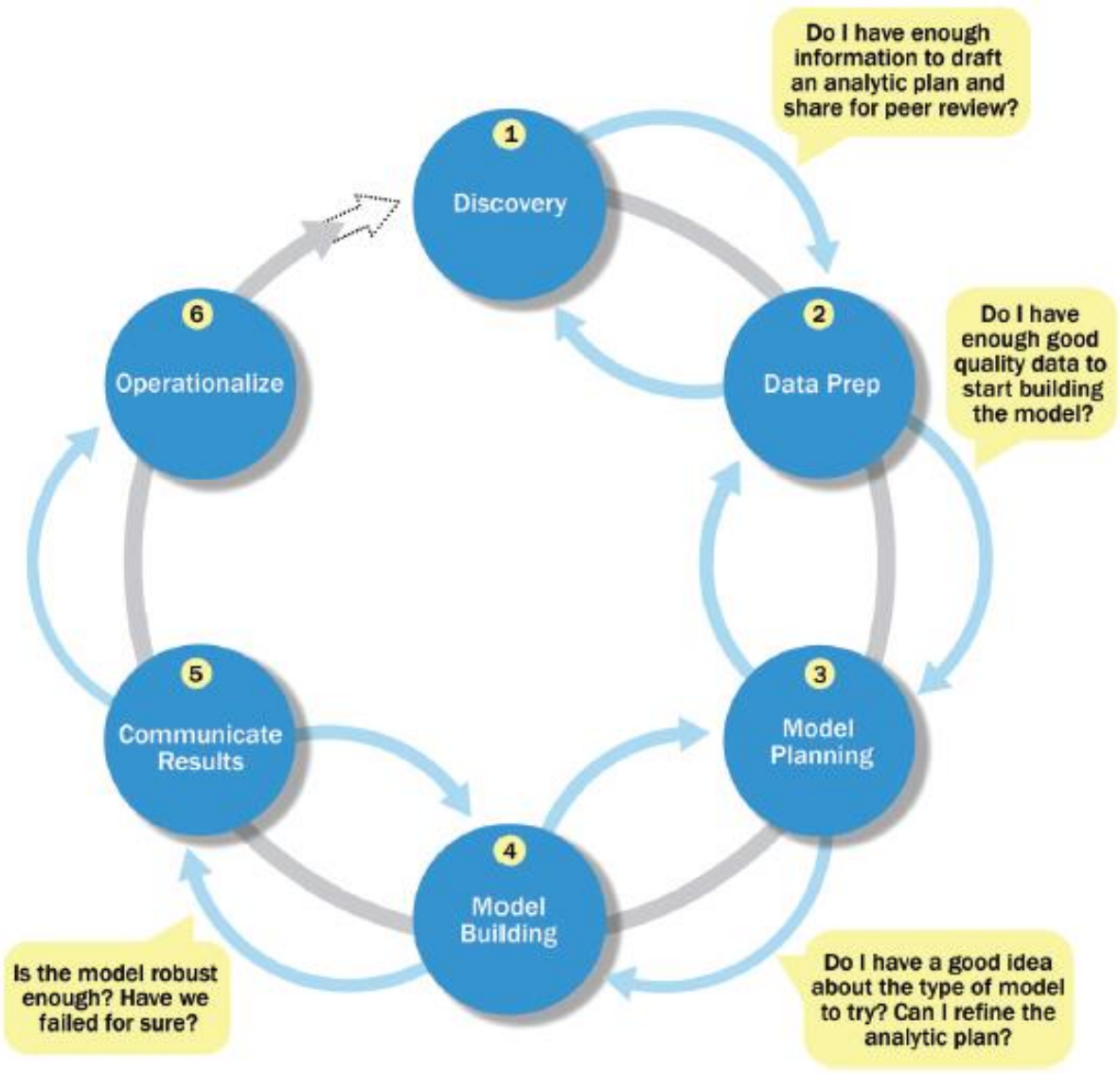
Big Data Analytics Lifecycle

Key Roles for a Successful Analytics Project



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle



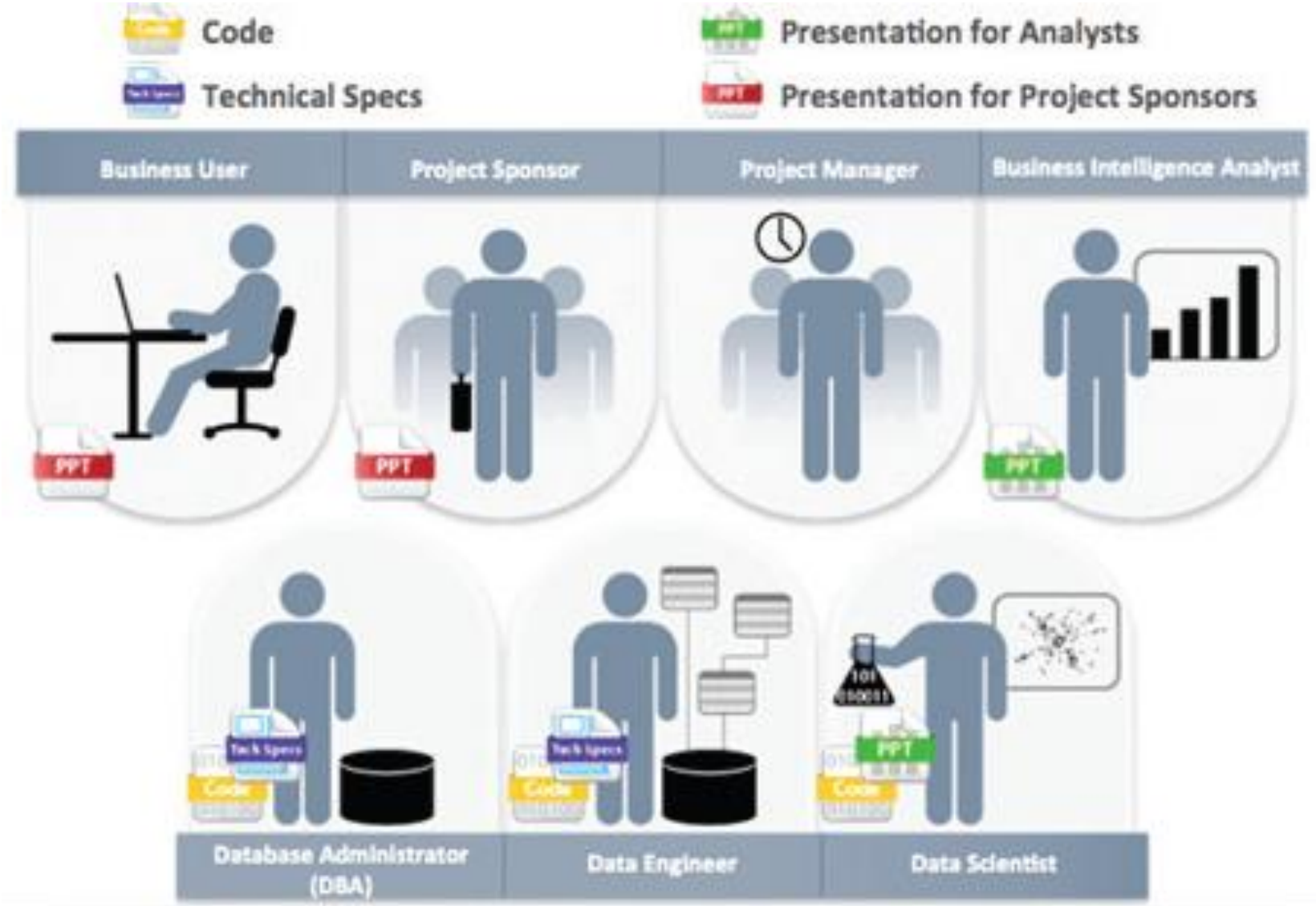
Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Overview of Data Analytics Lifecycle

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Key Outputs from a Successful Analytics Project



Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Data Mining Process

Data Mining Process

- 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)

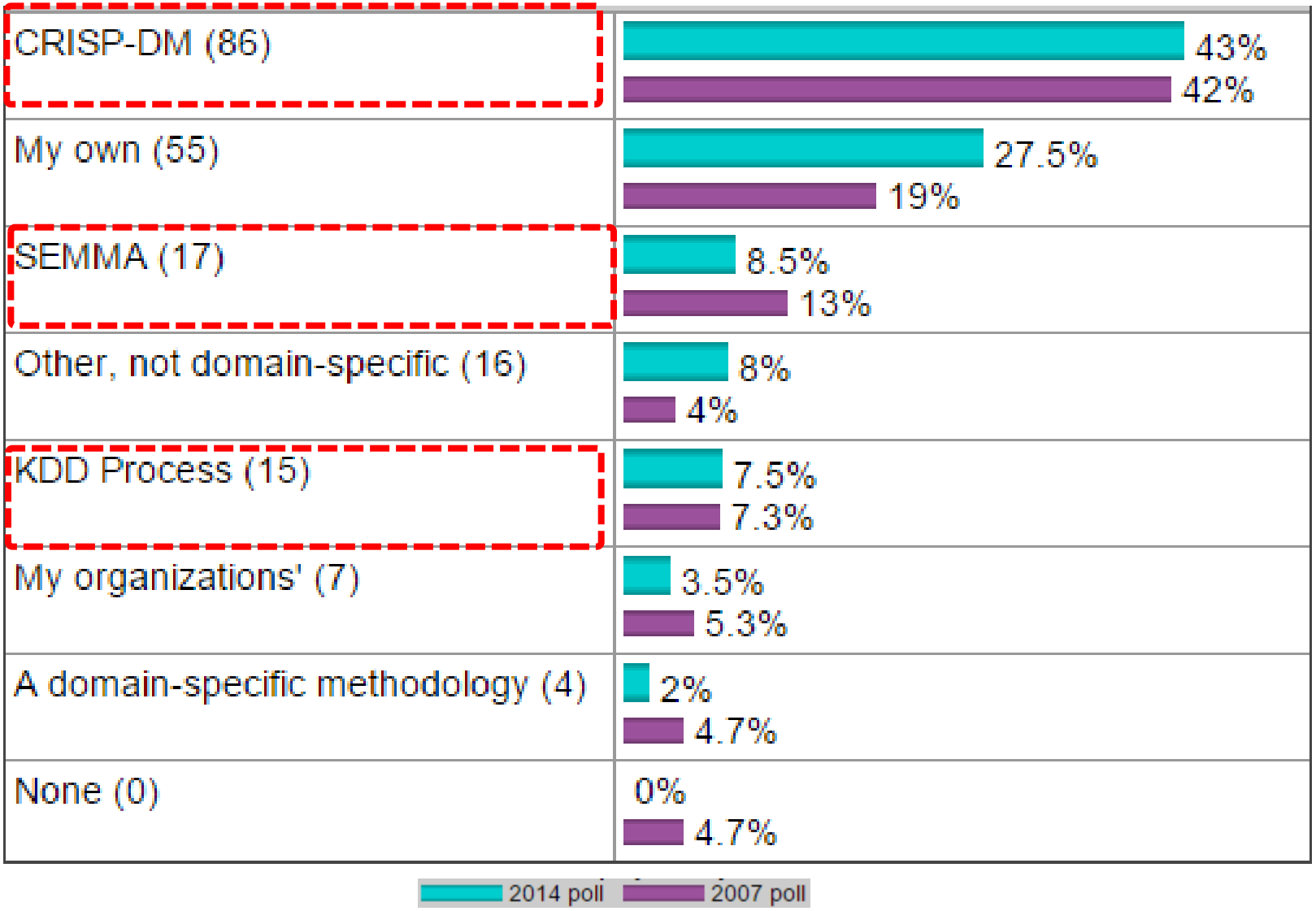
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Process (SOP of DM)

What main methodology are you using for your **analytics**, **data mining**, or **data science** projects ?

Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining Process



Source: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Data Mining:

Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Source: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39(11), 27-34.

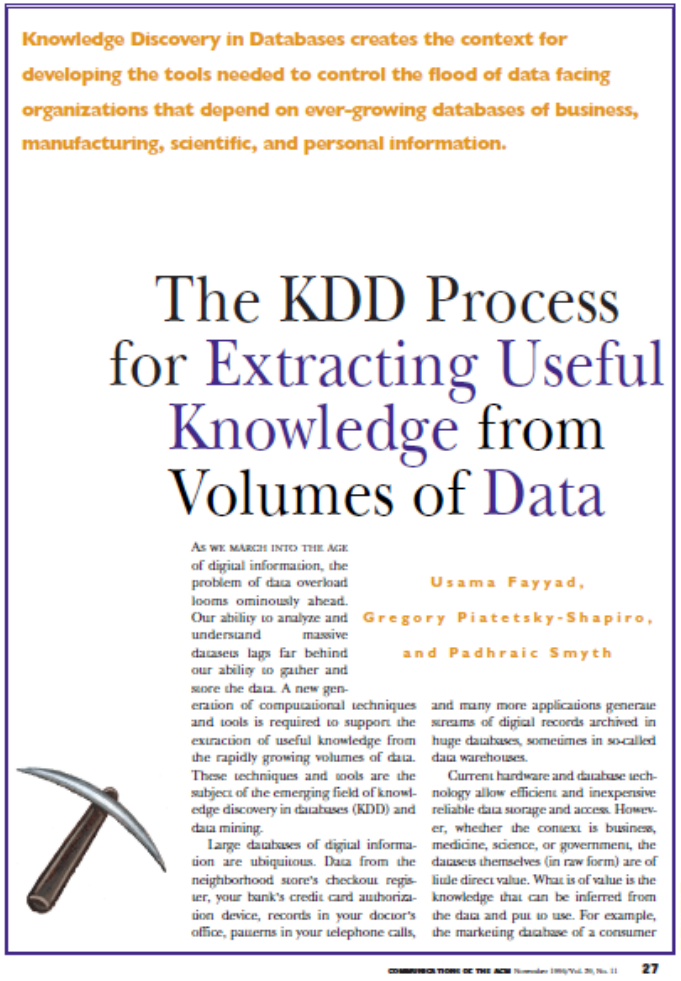
Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).

The KDD Process for

Extracting Useful Knowledge

from Volumes of Data.

Communications of the ACM, 39(11), 27-34.



Knowledge Discovery in Databases creates the context for developing the tools needed to control the flood of data facing organizations that depend on ever-growing databases of business, manufacturing, scientific, and personal information.

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Usama Fayyad,
Gregory Piatetsky-Shapiro,
and Padhraic Smyth

As we march into the age of digital information, the problem of data overload looms ominously ahead. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data. A new generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining.

Large databases of digital information are ubiquitous. Data from the neighborhood store's checkout register, your bank's credit card authorization device, records in your doctor's office, patterns in your telephone calls, and many more applications generate streams of digital records archived in huge databases, sometimes in so-called data warehouses.

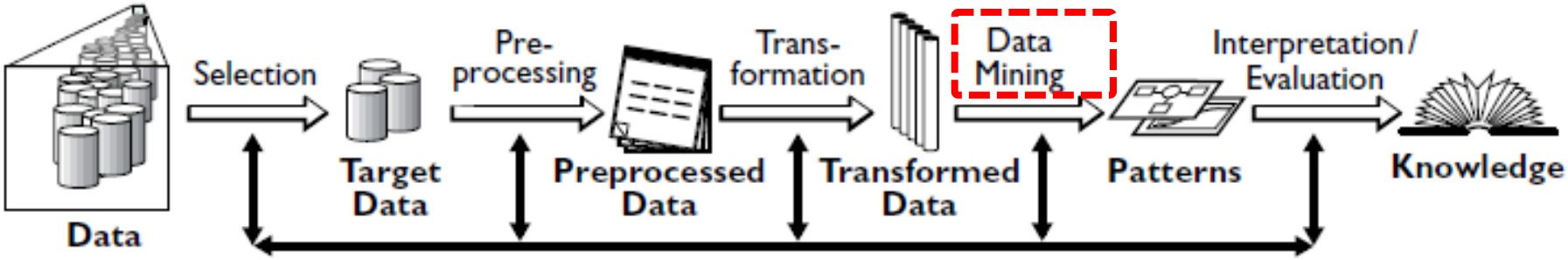
Current hardware and database technology allow efficient and inexpensive reliable data storage and access. However, whether the context is business, medicine, science, or government, the datasets themselves (in raw form) are of little direct value. What is of value is the knowledge that can be inferred from the data and put to use. For example, the marketing database of a consumer

TERRY WEEHER

Data Mining

Knowledge Discovery in Databases (KDD) Process

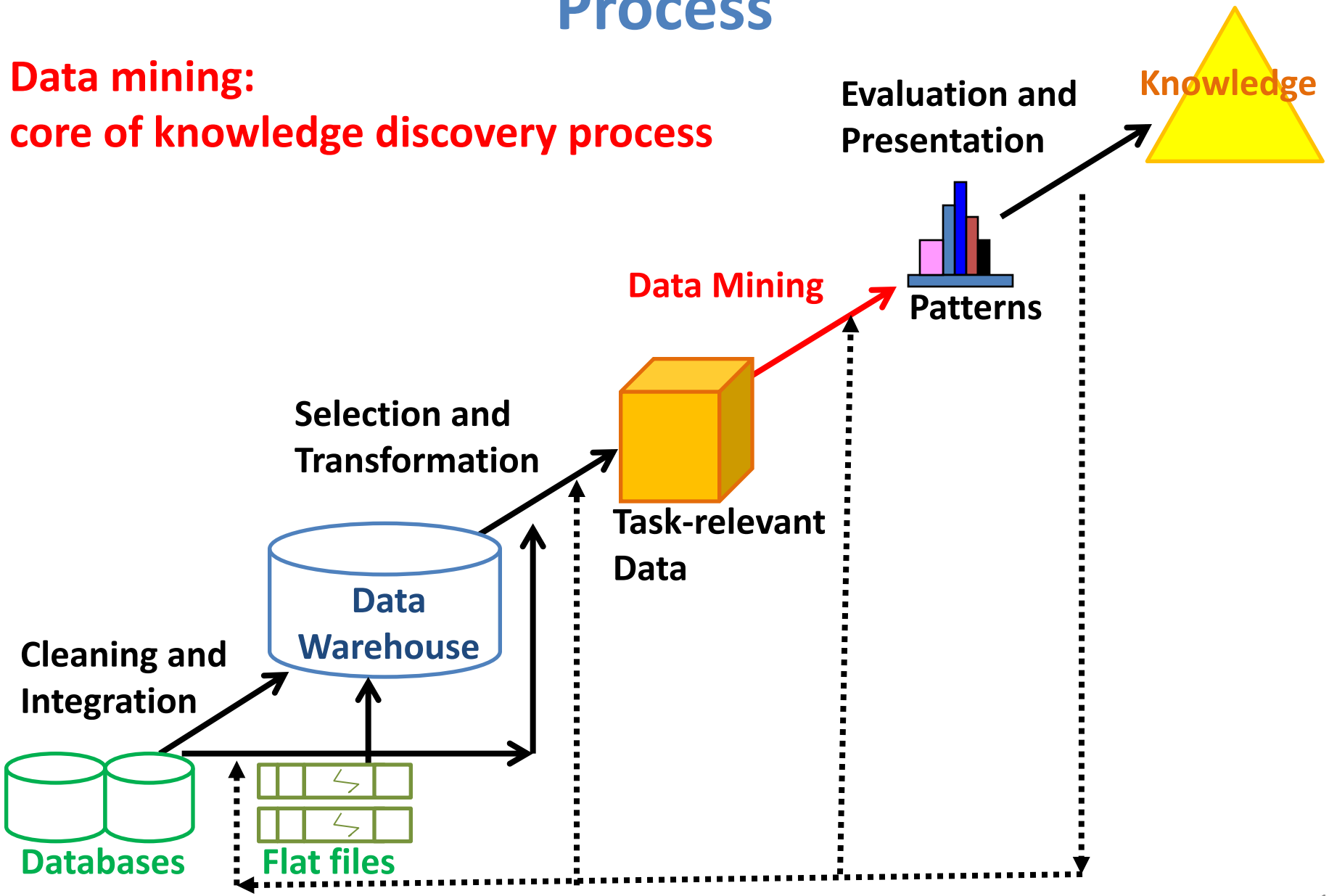
(Fayyad et al., 1996)



Source: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 39(11), 27-34.

Knowledge Discovery in Databases (KDD) Process

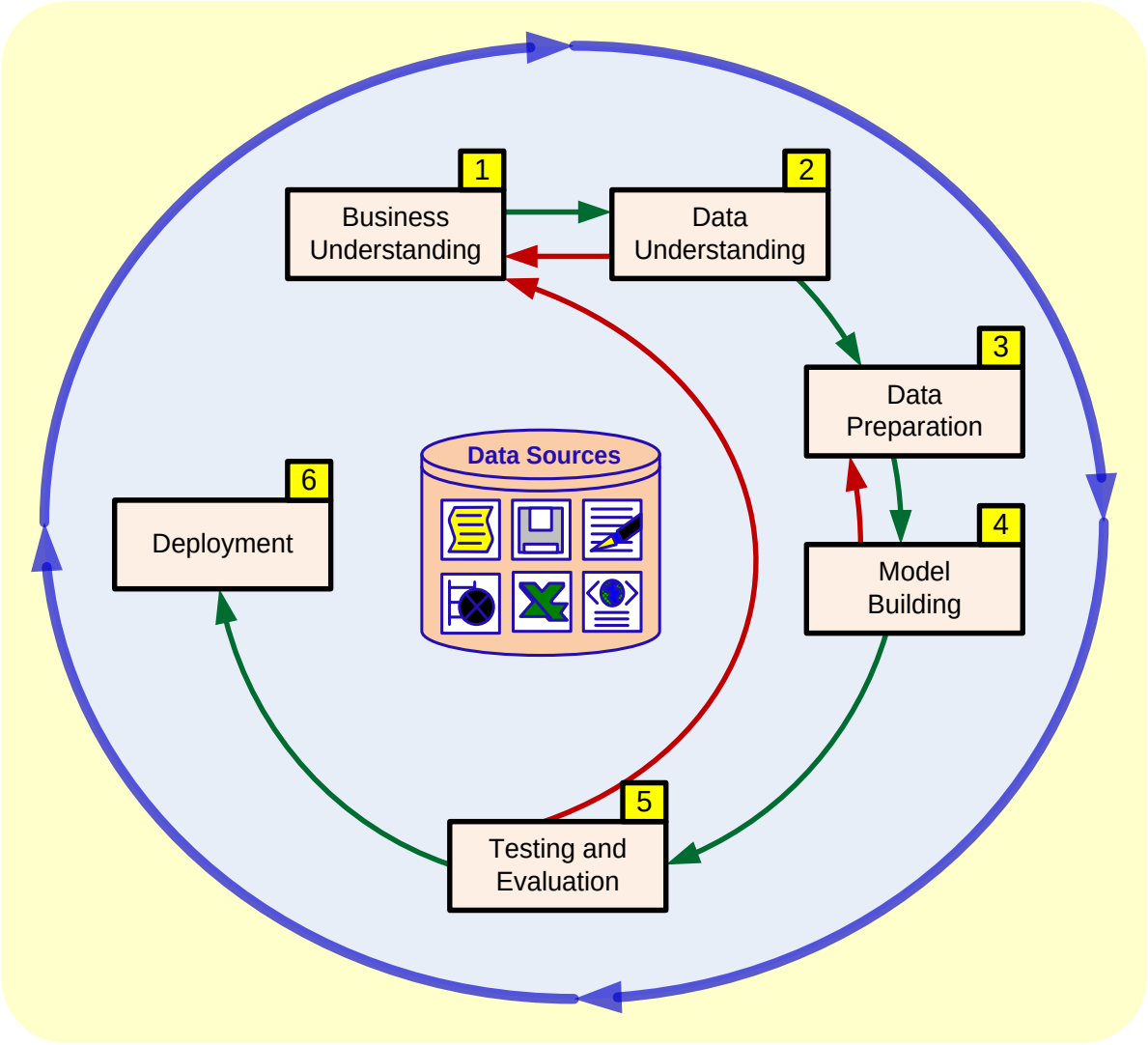
Data mining:
core of knowledge discovery process



Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier

Data Mining Process:

CRISP-DM



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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

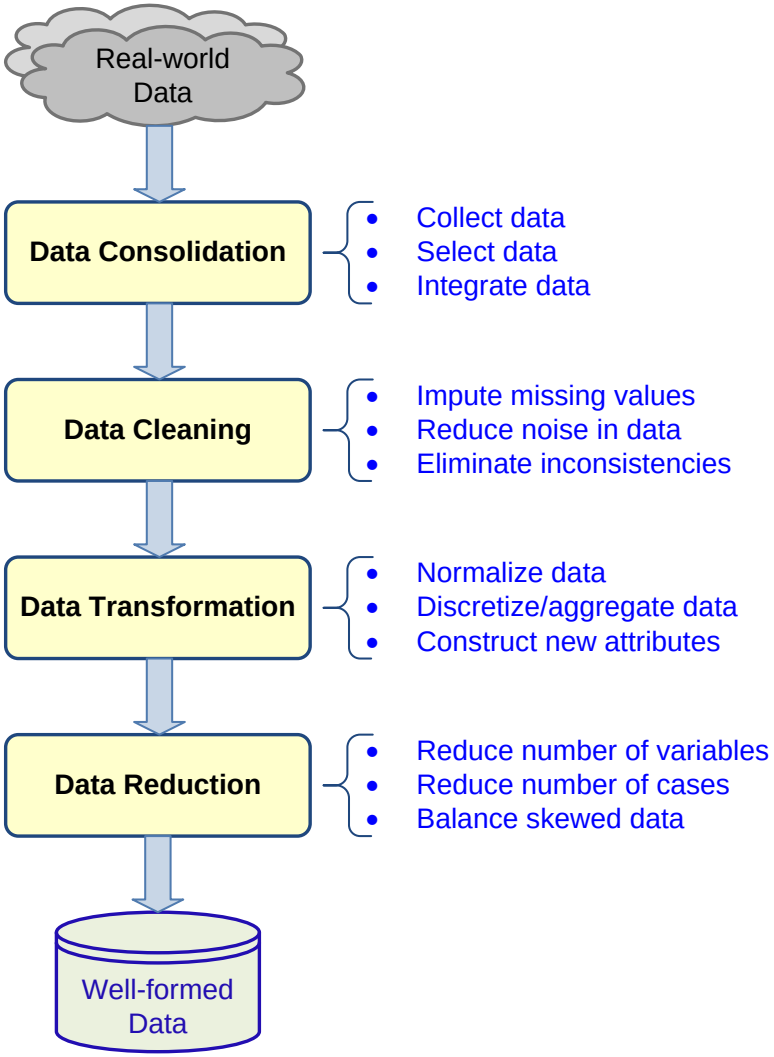
Step 6: Deployment

Accounts for ~85% of total project time

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

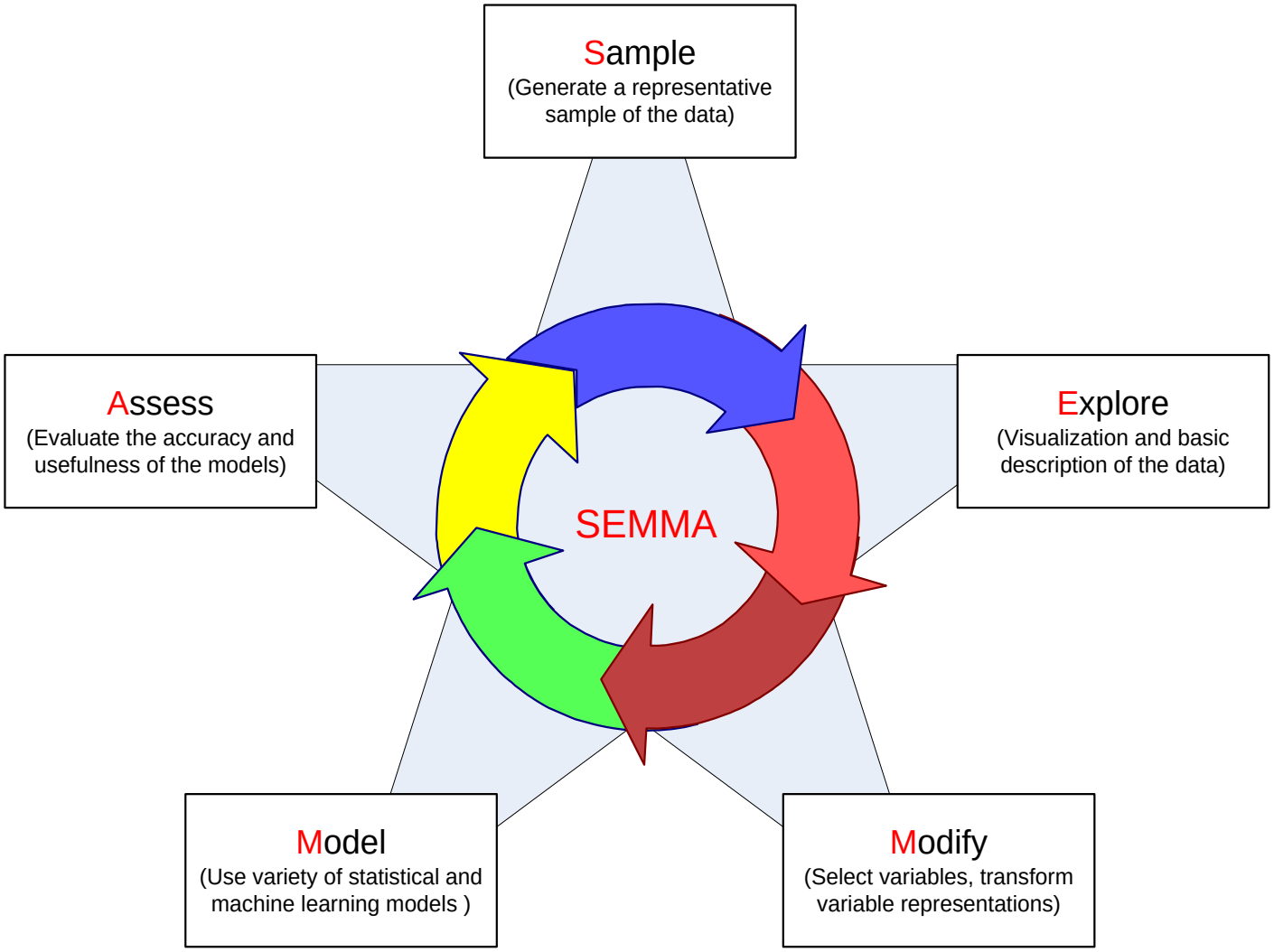
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Preparation – A Critical DM Task



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

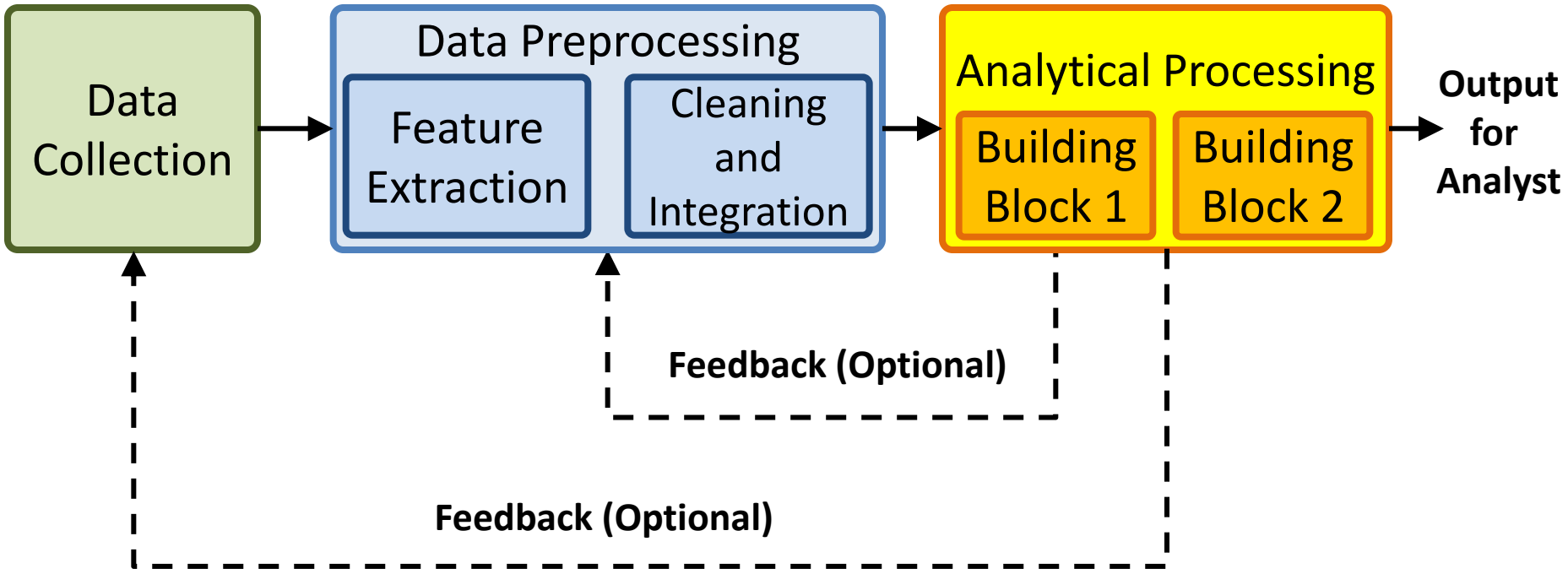
Data Mining Process: SEMMA



Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Processing Pipeline

(Charu Aggarwal, 2015)



Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer

Data Mining

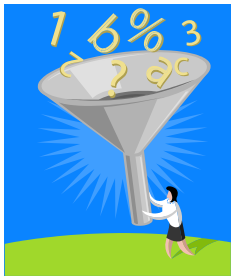
Why Data Mining?

- 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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

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,…

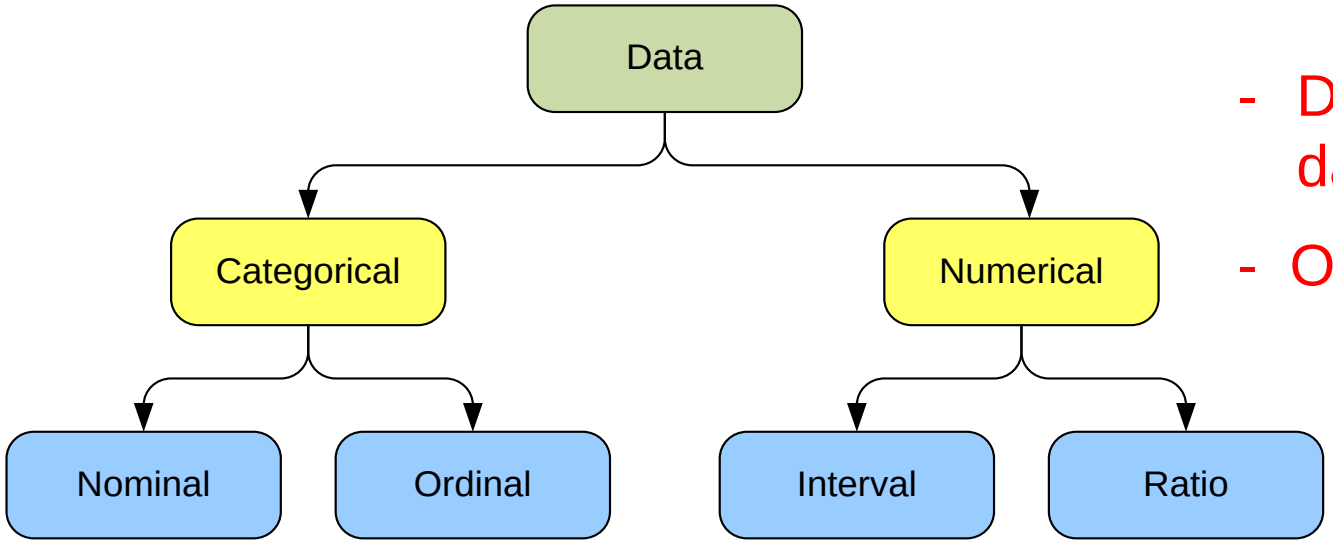


Data Mining Characteristics/Objectives

- 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.)

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?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

What Does DM Do?

- 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

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
  - 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

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Data Mining Applications (cont.)

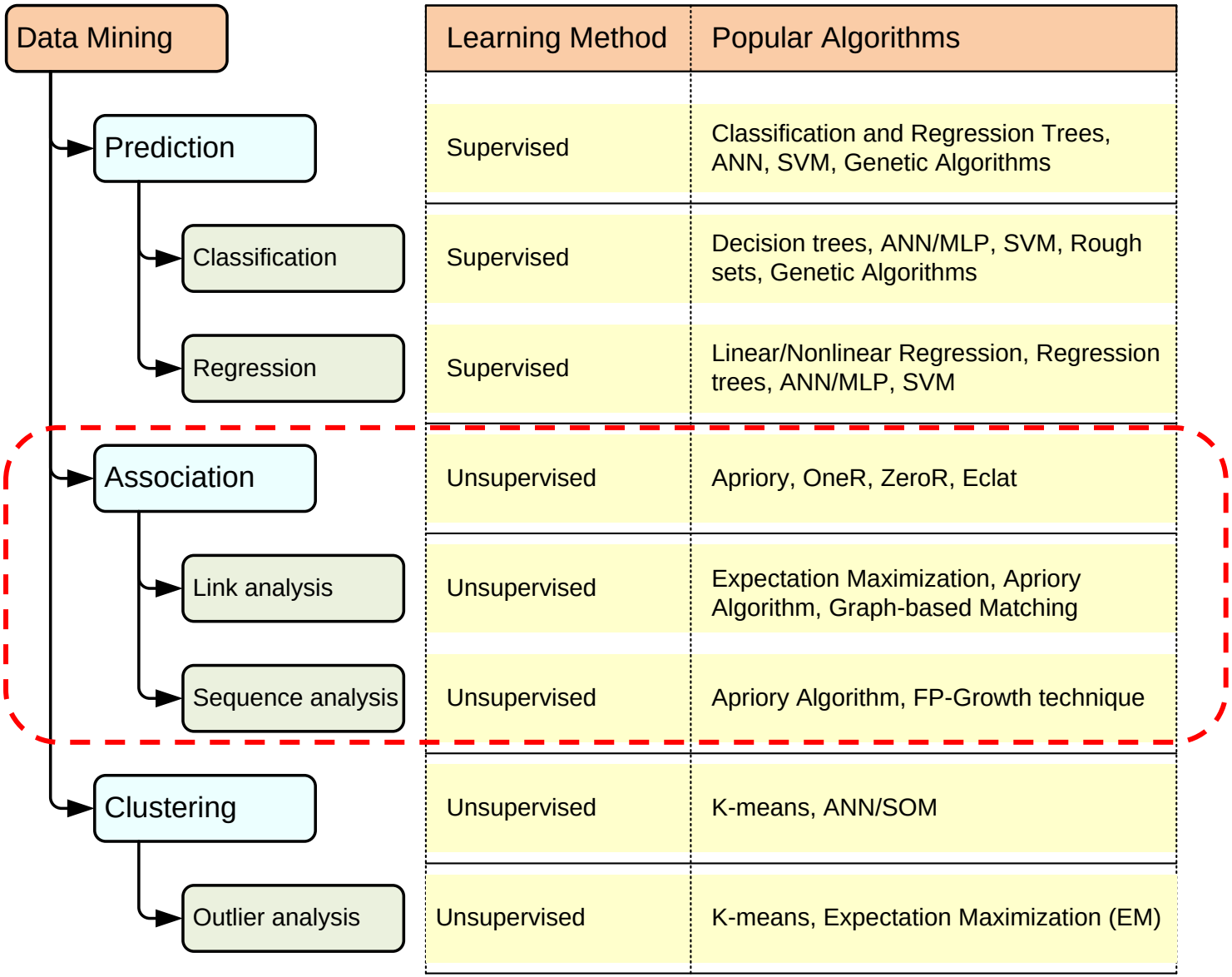
- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine

} Highly popular application areas for data mining

- Entertainment industry
- Sports
- Etc.

Association Analysis

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)

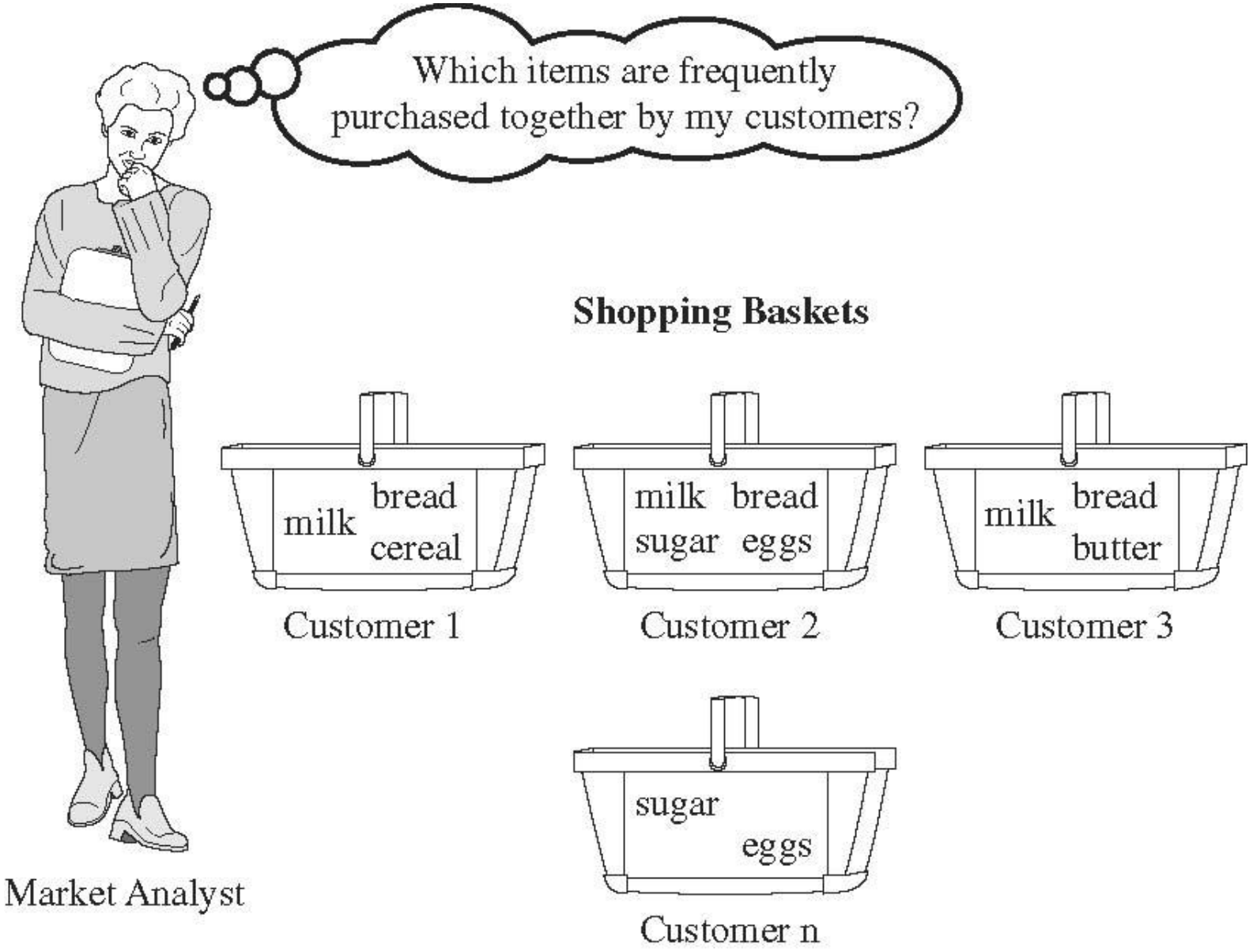
Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

Association Analysis: Mining Frequent Patterns, Association and Correlations

- Association Analysis
- Mining Frequent Patterns
- Association and Correlations
- Apriori Algorithm

Source: Han & Kamber (2006)

Market Basket Analysis



Source: Han & Kamber (2006)

Association Rule Mining

- Apriori Algorithm

Raw Transaction Data

Transaction No	SKUs (Item No)
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	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

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

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

  "Customer who bought a laptop 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 for ease of finding
  - Promote the items as a package (do not put one on sale if the other(s) are on sale)
  - Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially seeing and buying other items

Association Rule Mining

- 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

- 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

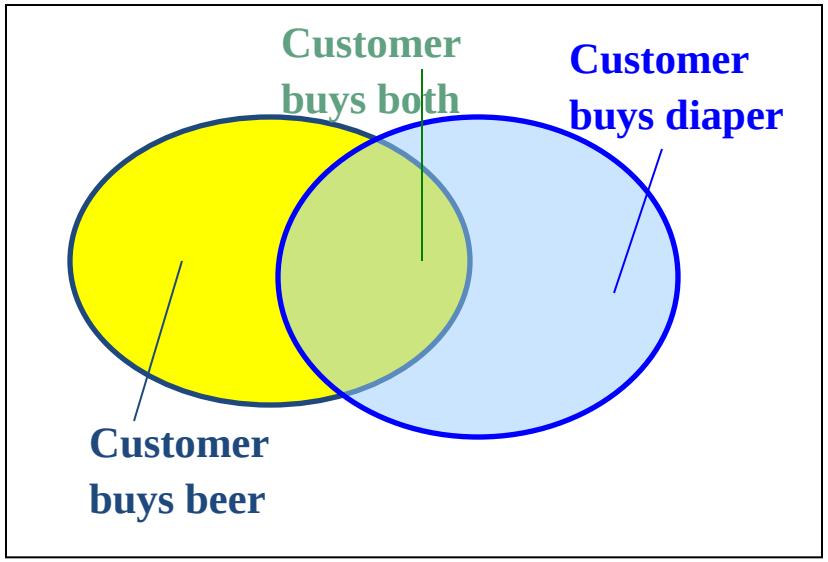
- Algorithms are available for generating association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules

Association Rule Mining

- 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

Basic Concepts: Frequent Patterns and Association Rules

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F



- Itemset X = $\{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, *s*, probability that a transaction contains $X \cup Y$
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

Let $sup_{min} = 50\%$, $conf_{min} = 50\%$

Freq. Pat.: {A:3, B:3, D:4, E:3, AD:3}

Association rules:

$A \rightarrow D$ (60%, 100%)

$D \rightarrow A$ (60%, 75%)

$A \rightarrow D$ (support = 3/5 = 60%, confidence = 3/3 =100%)

$D \rightarrow A$ (support = 3/5 = 60%, confidence = 3/4 = 75%)

Source: Han & Kamber (2006)

Market basket analysis

- Example
  - Which groups or sets of items are customers likely to purchase on a given trip to the store?
- Association Rule
  - *Computer → antivirus_software*
    [support = 2%; confidence = 60%]
    - A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
    - A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Source: Han & Kamber (2006)

Association rules

- Association rules are considered interesting if they satisfy both
  - a minimum support threshold and
  - a minimum confidence threshold.

Source: Han & Kamber (2006)

Frequent Itemsets, Closed Itemsets, and Association Rules

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with **support** s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the *union* of sets A and B, or say, both A and B). This is taken to be the probability, $P(A \cup B)$.[1] The rule $A \Rightarrow B$ has **confidence** c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is,

$$\textit{Support } (A \rightarrow B) \quad = P(A \cup B)$$

$$\textit{Confidence } (A \rightarrow B) = P(B|A)$$

Source: Han & Kamber (2006)

Support ($A \rightarrow B$) = $P(A \cup B)$
Confidence ($A \rightarrow B$) = $P(B|A)$

- The notation *$P(A \cup B)$ indicates the probability that a transaction contains the union of set A and set B*
  - *(i.e., it contains every item in A and in B).*
- *This should not be confused with P(A or B),* which indicates the probability that a transaction contains either *A or B.*

Source: Han & Kamber (2006)

Does diaper purchase predict beer purchase?

- Contingency tables

	Beer: Yes	Beer: No	
No diapers	6	94	100
diapers	40	60	100

DEPENDENT (yes)

	Beer: Yes	Beer: No
	23	77
	23	77

INDEPENDENT (no predictability)

Source: Dickey (2012) http://www4.stat.ncsu.edu/~dickey/SAScode/Encore_2012.ppt

$$Support\ (A \rightarrow B) = P(A \cup B)$$

$$Confidence\ (A \rightarrow B) = P(B|A)$$

$$Conf\ (A \rightarrow B) = Supp\ (A \cup B) / Supp\ (A)$$

$$Lift\ (A \rightarrow B) = Supp\ (A \cup B) / (Supp\ (A) \times Supp\ (B))$$

Lift (Correlation)

Lift (A → B) = Confidence (A → B) / Support(B)

Source: Dickey (2012) http://www4.stat.ncsu.edu/~dickey/SAScode/Encore_2012.ppt

Lift

Lift = Confidence / Expected Confidence if Independent

Checking → / Saving ↓	No (1500)	Yes (8500)	(10000)
No	500	3500	4000
Yes	1000	5000	6000

SVG=>CHKG Expect 8500/10000 = 85% if independent
Observed Confidence is 5000/6000 = 83%
Lift = 83/85 < 1.
Savings account holders actually LESS likely than others to have checking account !!!

Minimum Support and Minimum Confidence

- Rules that satisfy both a minimum support threshold (*min_sup*) and a minimum confidence threshold (*min_conf*) are called **strong**.
- By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

Source: Han & Kamber (2006)

K-itemset

- itemset
  - A set of items is referred to as an itemset.
- K-itemset
  - An itemset that contains *k items is a k*-itemset.
- Example:
  - The set {*computer, antivirus software*} is a 2-itemset.

Source: Han & Kamber (2006)

Absolute Support and Relative Support

- Absolute Support
  - The occurrence frequency of an itemset is the number of transactions that contain the itemset
    - frequency, support count, or count of the itemset
  - Ex: 3
- Relative support
  - Ex: 60%

Source: Han & Kamber (2006)

Frequent Itemset

- If the relative support of an itemset *I satisfies a prespecified minimum support threshold, then I is a* frequent itemset.
  - *i.e., the absolute* support of *I satisfies the corresponding minimum support count threshold*
- The set of frequent *k-itemsets is commonly denoted by* L_K

Source: Han & Kamber (2006)

Confidence

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$$

- the confidence of rule *A → B can be easily derived from the* support counts of *A and A ∪ B.*
- once the support counts of *A, B, and A ∪ B are* found, it is straightforward to derive the corresponding association rules *A →B and B →A* and check whether they are strong.
- Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

Source: Han & Kamber (2006)

Association rule mining: Two-step process

1. Find all frequent itemsets
   - By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, *min_sup.*
2. Generate strong association rules from the frequent itemsets
   - By definition, these rules must satisfy minimum support and minimum confidence.

Source: Han & Kamber (2006)

Efficient and Scalable Frequent Itemset Mining Methods

- The Apriori Algorithm
  - Finding Frequent Itemsets Using Candidate Generation

Source: Han & Kamber (2006)

Apriori Algorithm

- **Apriori** is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties, as we shall* see following.

Source: Han & Kamber (2006)

Apriori Algorithm

- Apriori employs an iterative approach known as a *level-wise search, where k-itemsets are used to explore (k+1)-itemsets.*
- *First, the set of frequent 1-itemsets is found* by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1.
- *Next, L_1 is used to find L_2,* the set of frequent 2-itemsets, which is used to find L_3, *and so on, until no more frequent k-itemsets can be found.*
- *The finding of each L_k requires one full scan of the database.*

Source: Han & Kamber (2006)

Apriori Algorithm

- To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property.
- Apriori property
  - *All nonempty subsets of a frequent itemset must also be frequent.*

Source: Han & Kamber (2006)

Apriori algorithm
(1) Frequent Itemsets
(2) Association Rules

Transaction Database

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Table 1 shows a database with 10 transactions.
Let *minimum support = 20%* and *minimum confidence = 80%*.
Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

Table 1: Transaction Database

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Apriori Algorithm
$C_1 \rightarrow L_1$

Step 1-1

C_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

minimum support = 20% = 2 / 10
Min. Support Count = 2

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

Apriori Algorithm
$C_2 \rightarrow L_2$

Step 1-2

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

C_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3
D, E	1

minimum support = 20% = 2 / 10
Min. Support Count = 2

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Apriori Algorithm
$C_3 \rightarrow L_3$

Step 1-3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

C_3

Itemset	Support Count
A, B, C	1
A, B, D	2
A, B, E	1
A, C, D	1
A, C, E	2
B, C, D	2
B, C, E	2

minimum support = 20% = 2 / 10
Min. Support Count = 2

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Generating Association Rules

Step 2-1

minimum confidence = 80%

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

Association Rules Generated from L_2

A→B: 3/6	B→A: 3/7
A→C: 4/6	C→A: 4/6
A→D: 3/6	D→A: 3/7
A→E: 2/6	E→A: 2/3
B→C: 3/7	C→B: 3/6
B→D: 6/7=85.7% *	D→B: 6/7=85.7% *
B→E: 2/7	E→B: 2/3
C→D: 3/6	D→C: 2/7
C→E: 3/6	E→C: 3/3=100% *

Generating Association Rules

Step 2-2

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

minimum confidence = 80%

Association Rules Generated from L_3

A→BD: 2/6	B→CD: 2/7
B→AD: 2/7	C→BD: 2/6
D→AB: 2/7	D→BC: 2/7
AB→D: 2/3	BC→D: 2/3
AD→B: 2/3	BD→C: 2/6
BD→A: 2/6	CD→B: 2/3
A→CE: 2/6	B→CE: 2/7
C→AE: 2/6	C→BE: 2/6
E→AC: 2/3	E→BC: 2/3
AC→E: 2/4	BC→E: 2/3
AE→C: 2/2=100%*	BE→C: 2/2=100%*
CE→A: 2/3	CE→B: 2/3

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

Frequent Itemsets and Association Rules

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

L_1

Itemset	Support Count
A	6
B	7
C	6
D	7
E	3

L_2

Itemset	Support Count
A, B	3
A, C	4
A, D	3
A, E	2
B, C	3
B, D	6
B, E	2
C, D	3
C, E	3

L_3

Itemset	Support Count
A, B, D	2
A, C, E	2
B, C, D	2
B, C, E	2

minimum support = 20%
minimum confidence = 80%

Association Rules:

B→D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D→B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E→C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)

Table 1 shows a database with 10 transactions.
Let *minimum support = 20%* and *minimum confidence = 80%*.
Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

Transaction ID	Items bought
T01	A, B, D
T02	A, C, D
T03	B, C, D, E
T04	A, B, D
T05	A, B, C, E
T06	A, C
T07	B, C, D
T08	B, D
T09	A, C, E
T10	B, D

Association Rules:

B→D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D→B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E→C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)

Summary

- Big Data Analytics Lifecycle
- Data Mining Process
- Data Mining
- Association Analysis
- Apriori algorithm
  - Frequent Itemsets
  - Association Rules

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Elsevier, 2006.
- Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann 2011.
- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, Pearson, 2011.
- EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015