ML410C

Projects in health informatics – Project and information management

Data Mining

Course logistics

- Instructor: Panagiotis Papapetrou
- Contact: panagiotis@dsv.su.se
- Office: 7511
- Office hours: by appointment only

Course logistics

- Three lectures
- Project

Schedule

DATE	TIME	ROOM	TOPIC
MONDAY 2013-09-09	10:00-11:45	502	Introduction to data mining
WEDNESDAY 2013-09-11	09:00-10:45	501	Decision trees, rules and forests
FRIDAY 2013-09-13	10:00-11:45	Sal C	Evaluating predictive models and tools for data mining

Project

- Will involve some data mining task on medical data
- Data will be provided to you
- Some pre-processing may be required
- Readily available GUI Data Mining tools shall be used
- A short report (3-4 pages) with the results should be submitted
- More details on Friday…

Textbooks

- D. Hand, H. Mannila and P. Smyth: Principles of Data Mining. MIT Press, 2001

- Jiawei Han and Micheline Kamber: Data Mining: Concepts and Techniques. Second Edition. Morgan Kaufmann Publishers, March 2006

- Research papers (pointers will be provided)

Above all

- The goal of the course is to learn and enjoy

- The basic principle is to ask questions when you don’t understand

- Say when things are unclear; not everything can be clear from the beginning

- Participate in the class as much as possible

Introduction to data mining

- Why do we need data analysis?
- What is data mining?
- Examples where data mining has been useful
- Data mining and other areas of computer science and statistics
- Some (basic) data-mining tasks

Why do we need data analysis

- Really really lots of raw data data!!
  - Moore’s law: more efficient processors, larger memories
  - Communications have improved too
  - Measurement technologies have improved dramatically
  - It is possible to store and collect lots of raw data
  - The data-analysis methods are lagging behind
- Need to analyze the raw data to extract knowledge

The data is also very complex

- Multiple types of data: tables, time series, images, graphs, etc
- Spatial and temporal aspects
- Large number of different variables
- Lots of observations → large datasets

Example: transaction data

- Billions of real-life customers: e.g., supermarkets
- Billions of online customers: e.g., amazon, expedia, etc.
- Critical areas: e.g., patient records

Example: document data

- Web as a document repository: 50 billion of web pages
- Wikipedia: 4 million articles (and counting)
- Online collections of scientific articles

Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 200 million users
- MySpace: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

Example: genomic sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- 310^9 nucleotides per person → 310^12 nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Example: environmental data

- Climate data (just an example)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”

- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”

We have large datasets…so what?

- Goal: obtain useful knowledge from large masses of data

- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst"

- Tell me something interesting about the data; describe the data

What can data-mining methods do?

- Extract frequent patterns
  - There are lots of documents that contain the phrases "association rules", "data mining" and "efficient algorithm"
- Extract association rules
  - 80% of the ICA customers that buy beer and sausage also buy mustard
- Extract rules
  - If occupation = PhD student then income < 30,000 SEK

What can data-mining methods do?

- Rank web-query results
  - What are the most relevant web-pages to the query: “Student housing Stockholm University”?
- Find good recommendations for users
  - Recommend amazon customers new books
  - Recommend facebook users new friends/groups
- Find groups of entities that are similar (clustering)
  - Find groups of facebook users that have similar friends/interests
  - Find groups amazon users that buy similar products
  - Find groups of ICA customers that buy similar products

Goal of this course

- Describe some problems that can be solved using data-mining methods
- Discuss the intuition behind data mining methods that solve these problems
- Illustrate the theoretical underpinnings of these methods
- Show how these methods can be useful in health informatics

Data mining and related areas

- How does data mining relate to machine learning?
- How does data mining relate to statistics?
- Other related areas?

Data mining vs. machine learning

- Machine learning methods are used for data mining
  - Classification, clustering
- Amount of data makes the difference
  - Data mining deals with much larger datasets and scalability becomes an issue
- Data mining has more modest goals
  - Automating tedious discovery tasks
  - Helping users, not replacing them

Data mining vs. statistics

- "tell me something interesting about this data" – what else is this than statistics?
  - The goal is similar
  - Different types of methods
  - In data mining one investigates lots of possible hypotheses
  - Data mining is more exploratory data analysis
  - In data mining there are much larger datasets → algorithmics/scalability is an issue

Data mining and databases

- Ordinary database usage: deductive
- Knowledge discovery: inductive
- New requirements for database management systems
- Novel data structures, algorithms and architectures are needed

Machine learning

The *machine learning* area deals with artificial systems that are able to improve their performance with experience.

Supervised learning
Experience: objects that have been assigned class labels
Performance: typically concerns the ability to classify new (previously unseen) objects

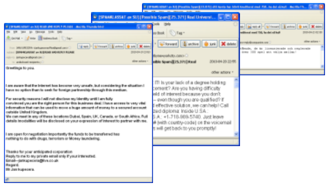
Predictive data mining

Unsupervised learning
Experience: objects for which no class labels have been given
Performance: typically concerns the ability to output useful characterizations (or groupings) of objects

Descriptive data mining

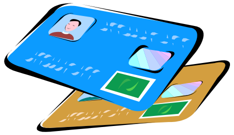
Examples of supervised learning

- Email classification (spam or not)

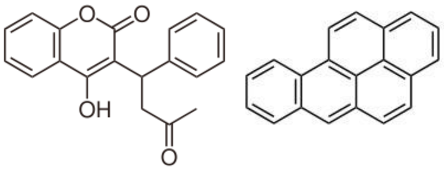


- Customer classification (will leave or not)

- Credit card transactions (fraud or not)

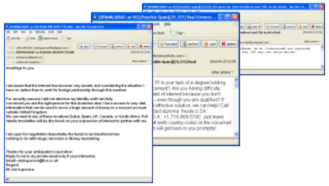


- Molecular properties (toxic or not)



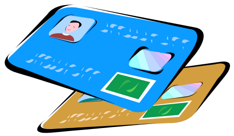
Examples of unsupervised learning

- find useful email categories

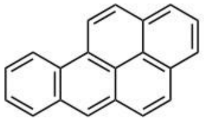


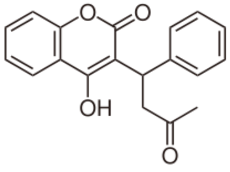
- find interesting purchase patterns

- describe normal credit card transactions



- find groups of molecules with similar properties





Data mining: input

- Standard requirement: each case is represented by <u>one</u> row in <u>one</u> table
- Possible additional requirements

  - only numerical variables

  - all variables have to be normalized

  - only categorical variables

  - no missing values
- Possible generalizations

  - multiple tables

  - recursive data types (sequences, trees, etc.)

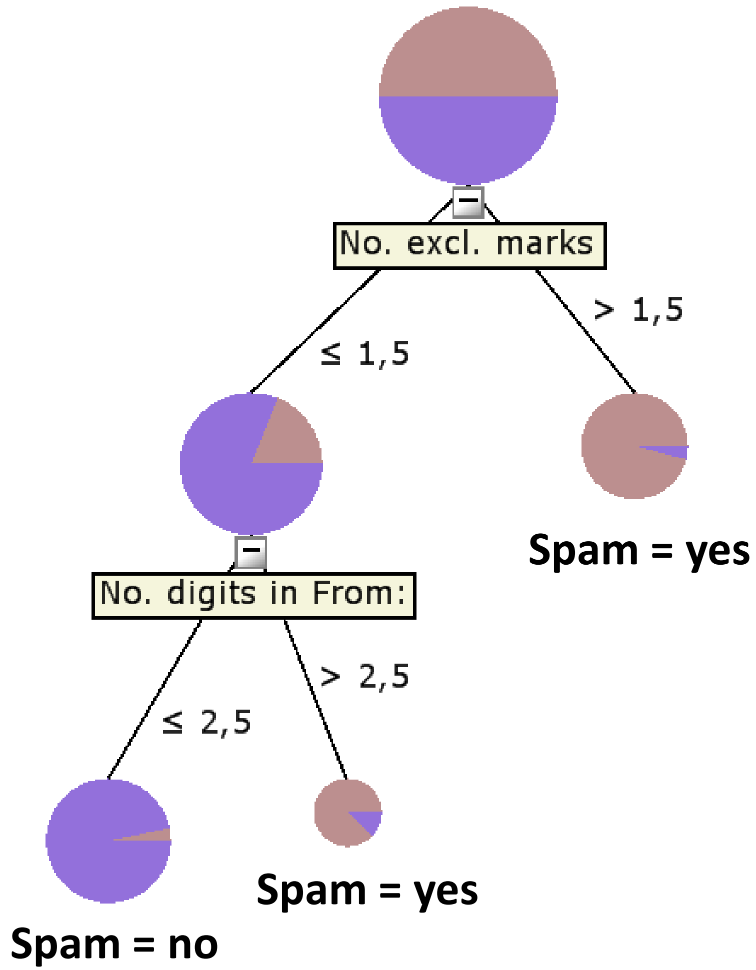
An example: email classification

Features (attributes)

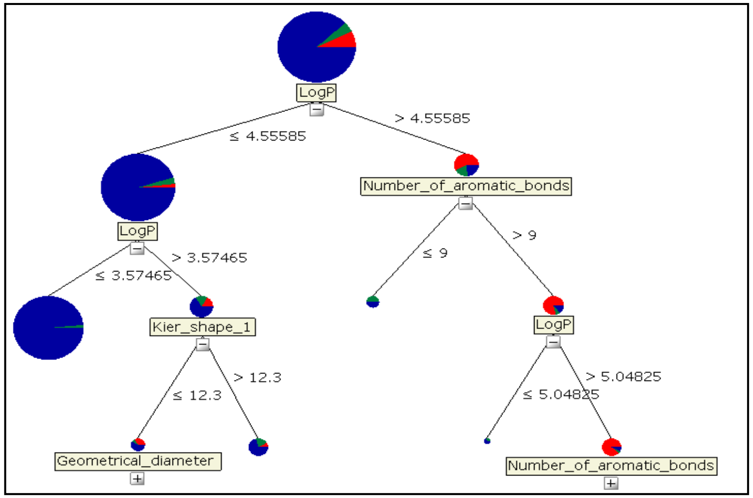
Examples (observations)

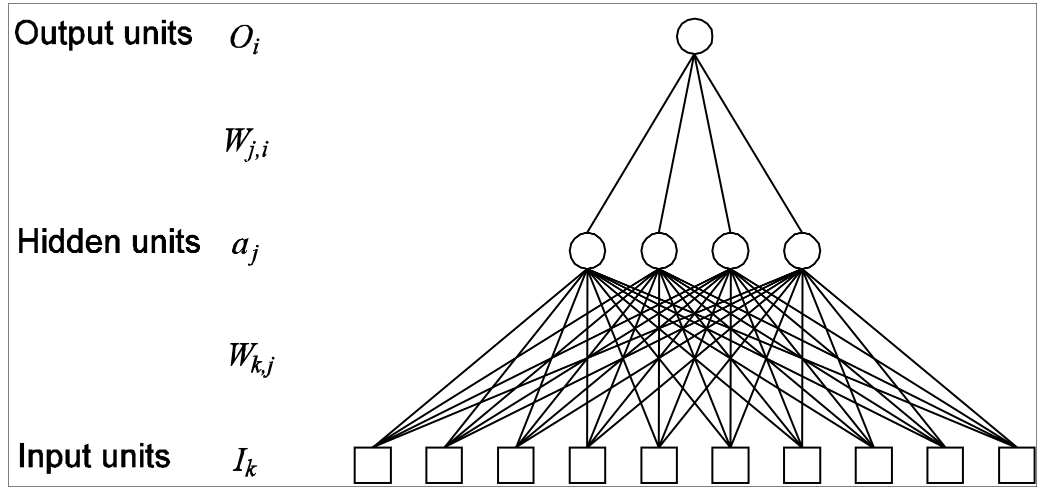
Ex.	All caps	No. excl. marks	Missing date	No. digits in From:	Image fraction	Spam
e1	yes	0	no	3	0	yes
e2	yes	3	no	0	0.2	yes
e3	no	0	no	0	1	no
e4	no	4	yes	4	0.5	yes
e5	yes	0	yes	2	0	no
e6	no	0	no	0	0	no

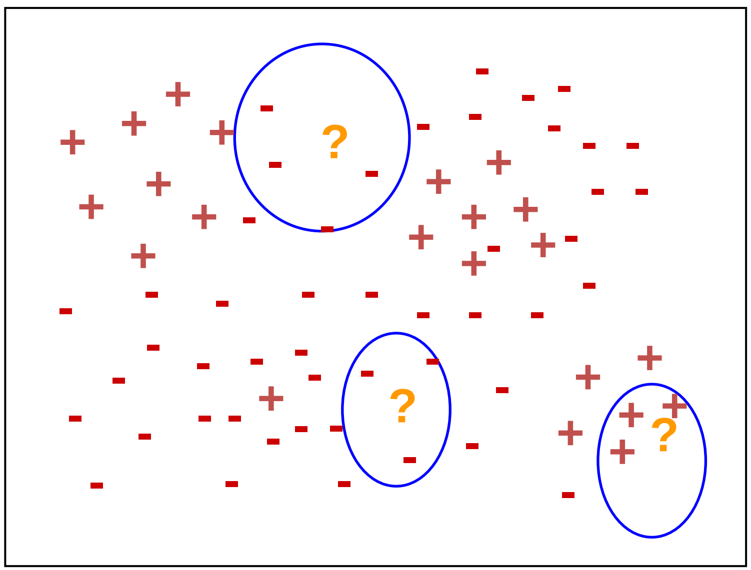
Data mining: output

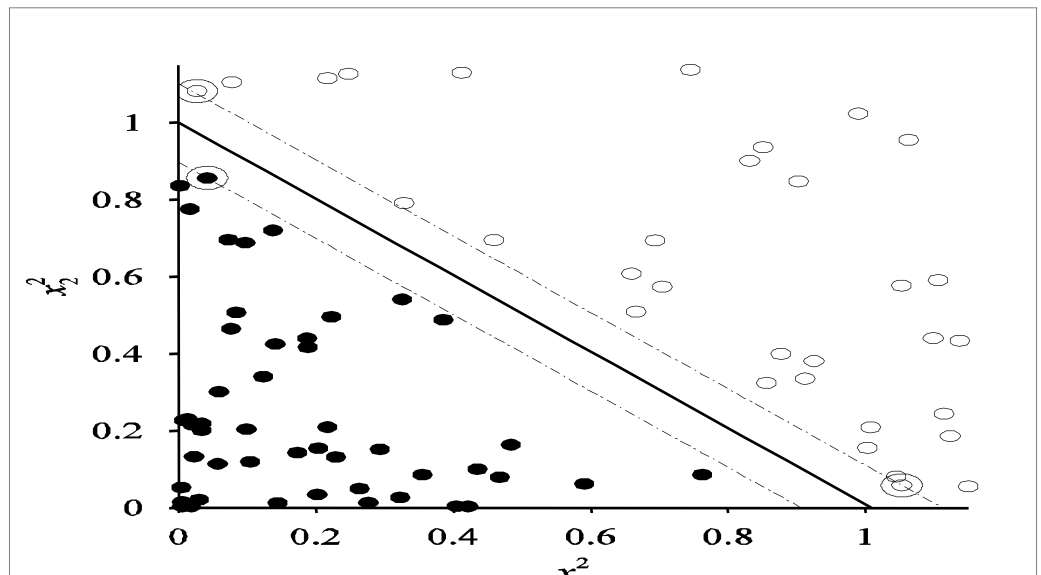


Data mining: output









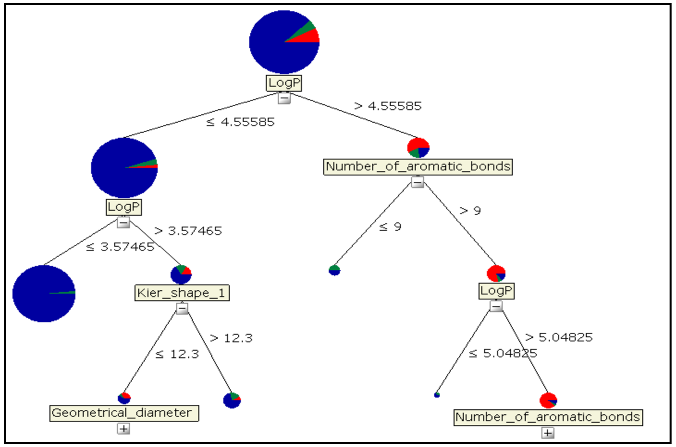
Data mining: output

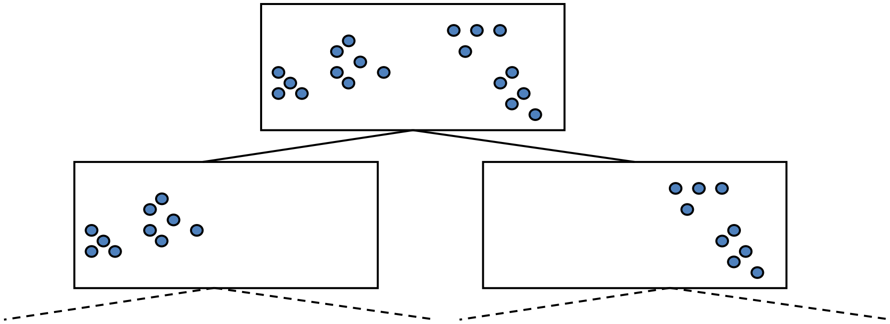
- Interpretable representation of findings
  - equations, rules, decision trees, clusters

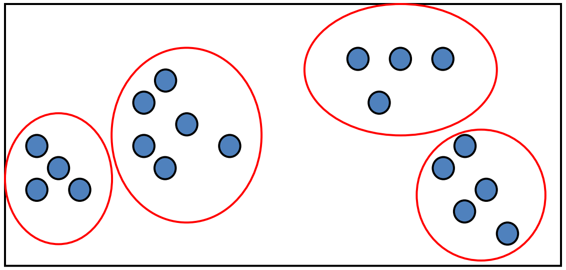
$$y = 0.25 + 4.5x_1 - 2.2x_2 + 3.1x_3$$

$$\textbf{if } x_1 > 3.0 \,\&\, x_2 \le 1.8 \textbf{ then } y = 1.0$$

BuysMilk & BuysCereals → BuysJuice
[Support : 0.05, Confidence : 0.85]

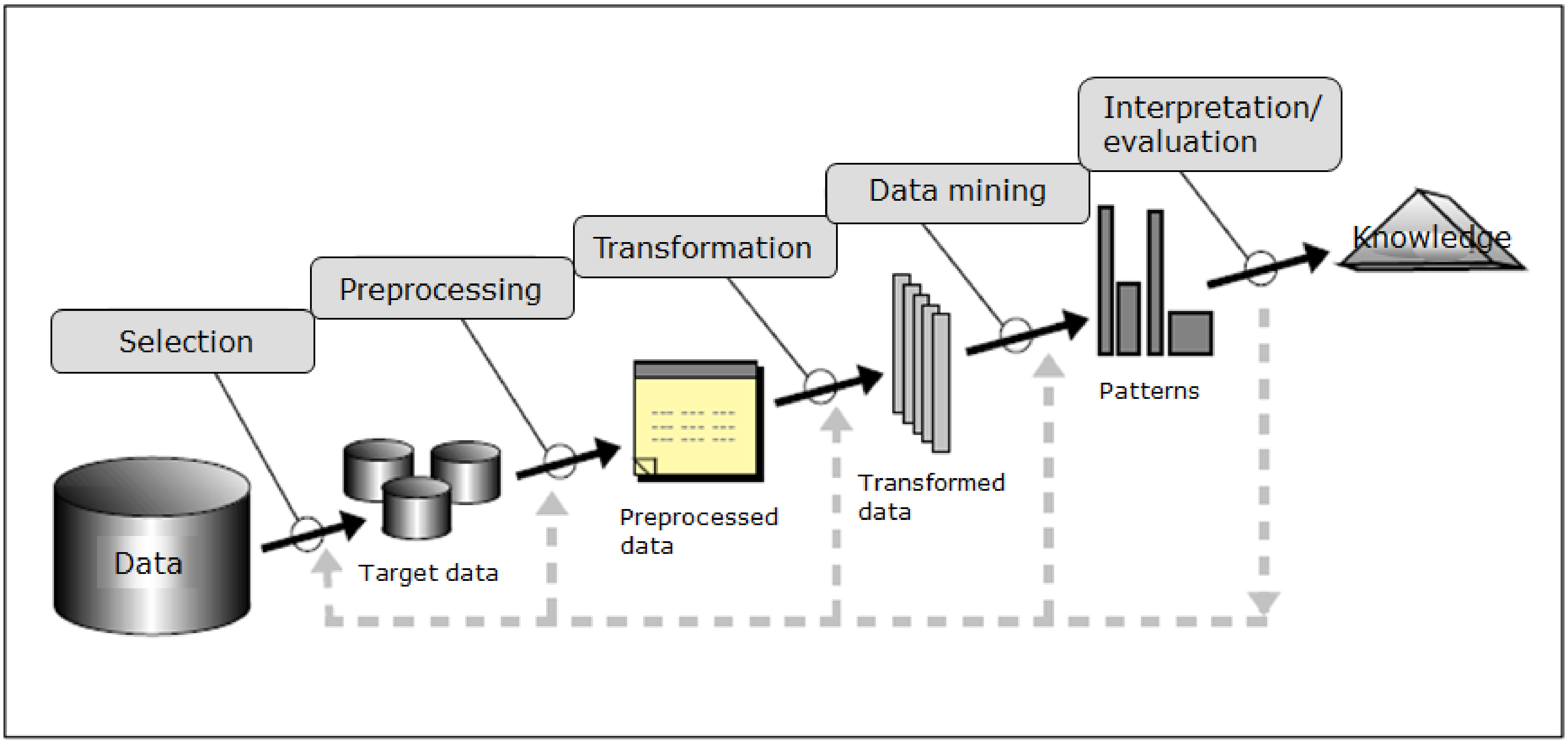






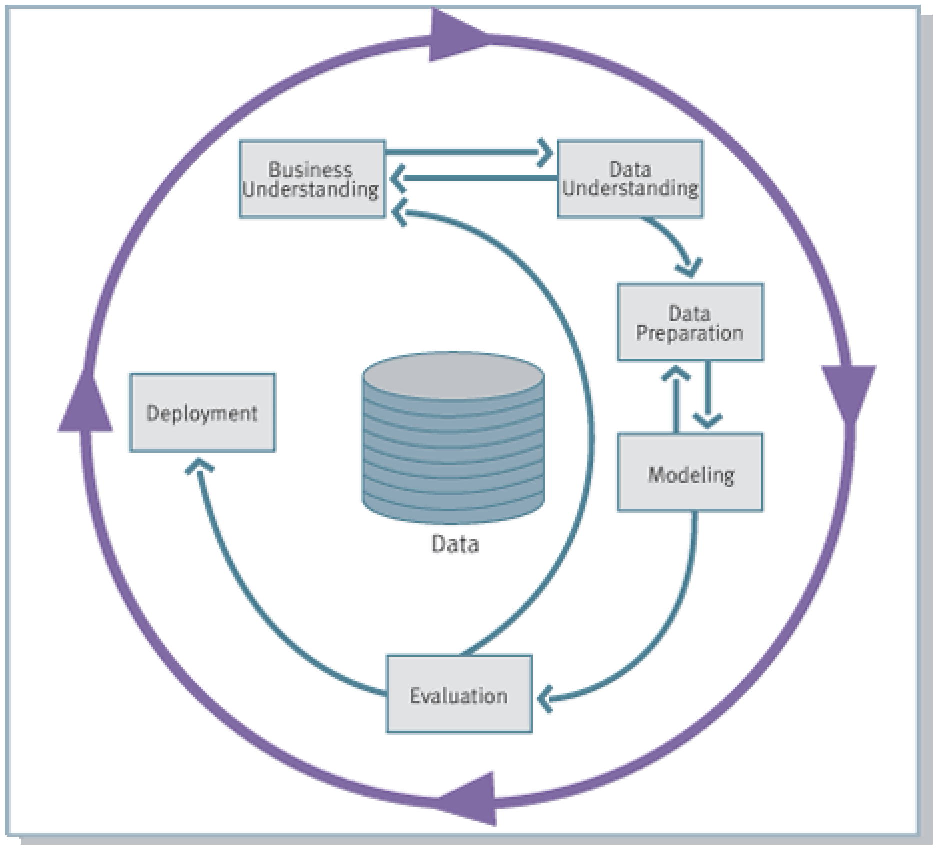
The Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



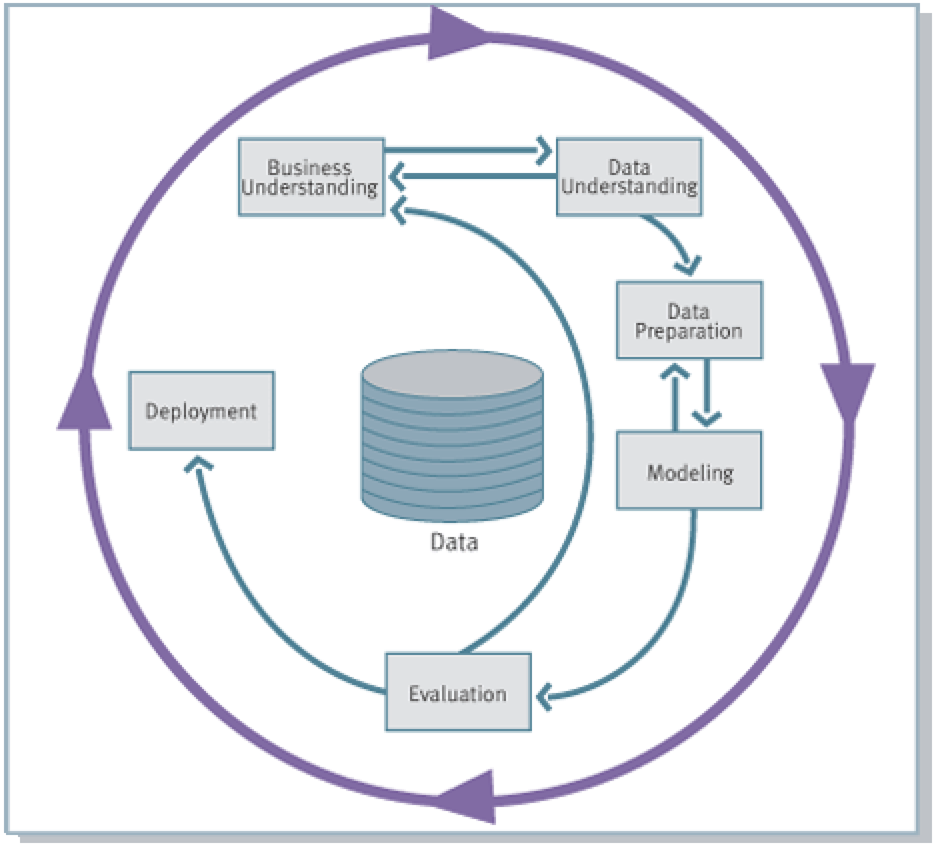
U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine 17(3): 37-54 (1996)

CRISP-DM: CRoss Industry Standard Process for Data Mining



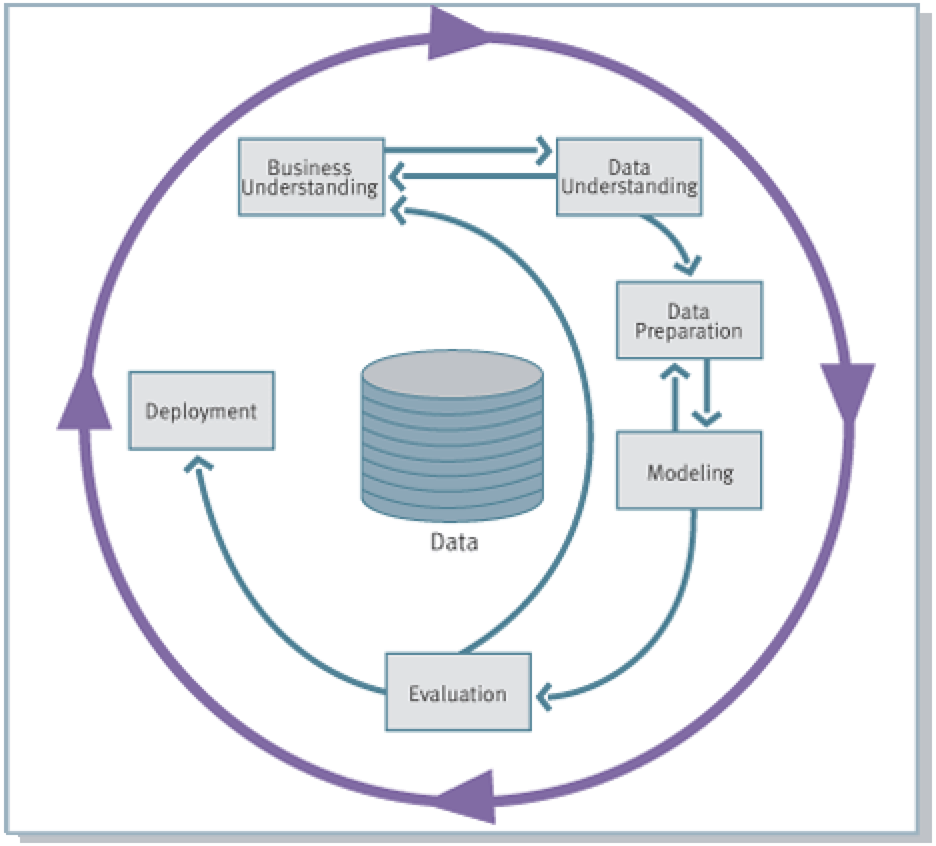
Shearer C., “The CRISP-DM model: the new blueprint for data mining”, Journal of Data Warehousing 5 (2000) 13-22 (see also www.crisp-dm.org)

CRISP-DM



- **Business Understanding**
  - understand the project objectives and requirements from a business perspective
  - convert this knowledge into a data mining problem definition
  - create a preliminary plan to achieve the objectives

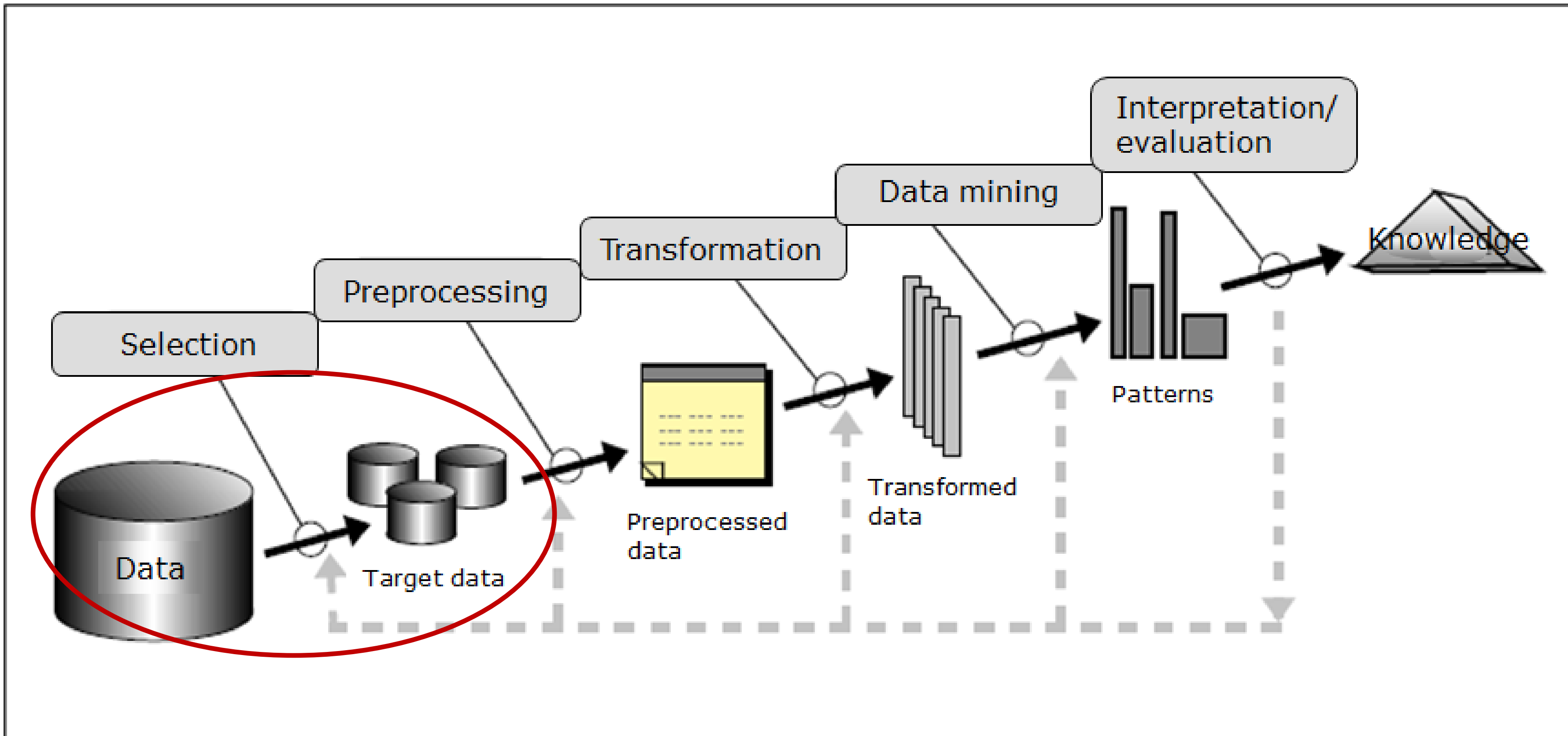
CRISP-DM



- **Data Understanding**
  - initial data collection
  - get familiar with the data
  - identify data quality problems
  - discover first insights
  - detect interesting subsets
  - form hypotheses for hidden information

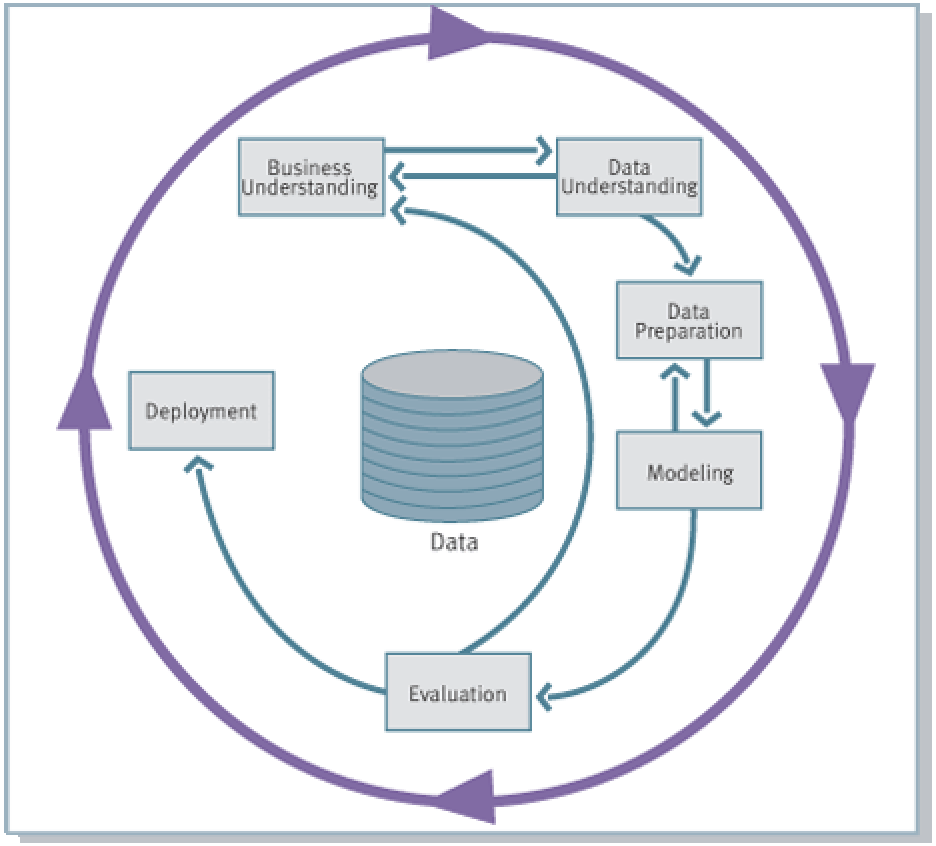
The Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine 17(3): 37-54 (1996)

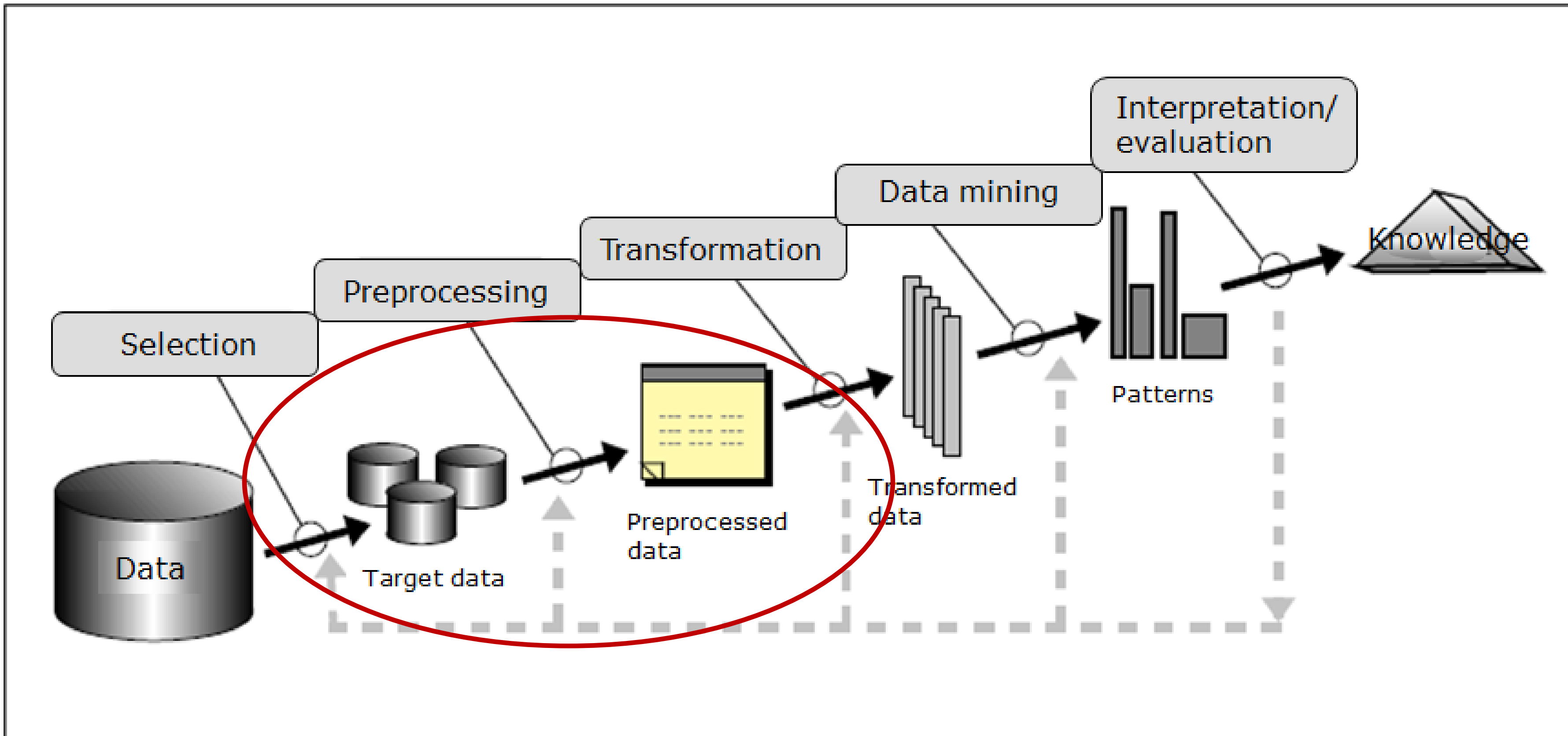
CRISP-DM



- **Data Preparation**
  - construct the final dataset to be fed into the machine learning algorithm
  - tasks here include: table, record, and attribute selection, data transformation and cleaning

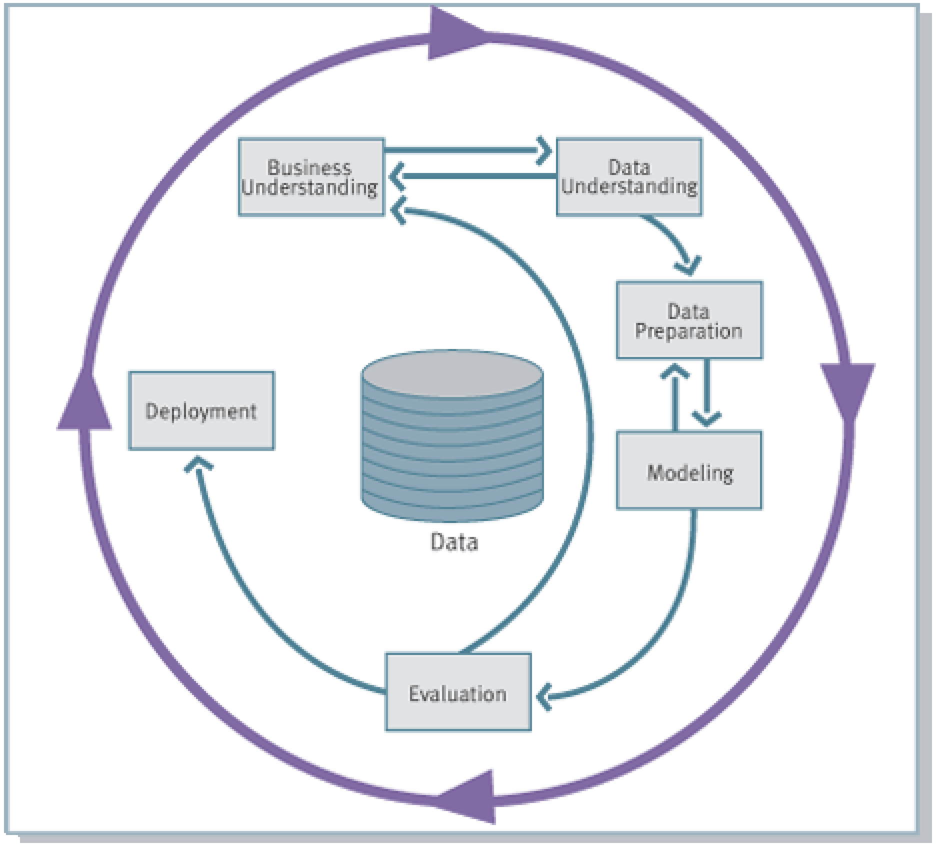
The Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine 17(3): 37-54 (1996)

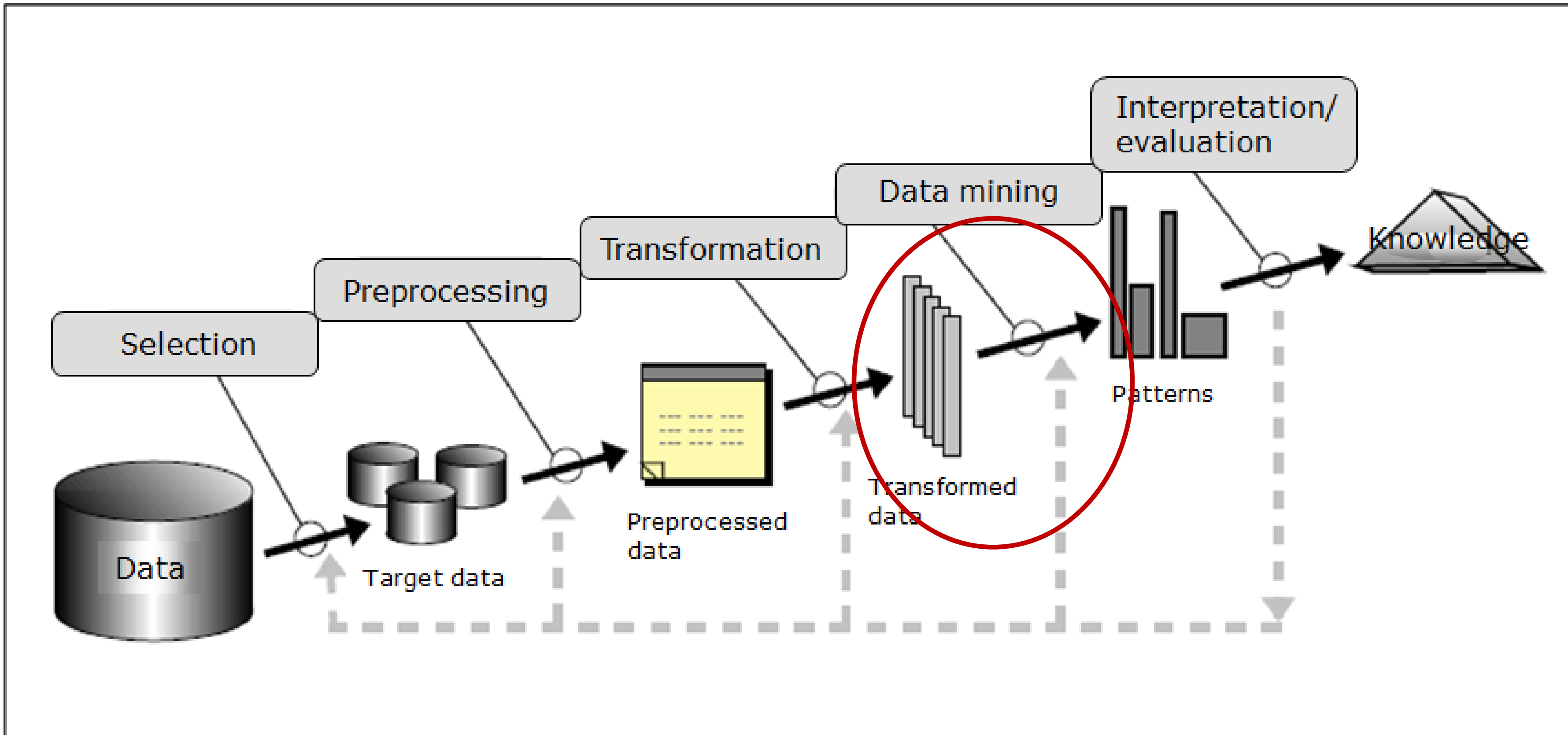
CRISP-DM



- **Modeling**
  - various data mining techniques are selected and applied
  - parameters are learned
  - some methods may have specific requirements on the form of input data
  - going back to the data preparation phase may be needed

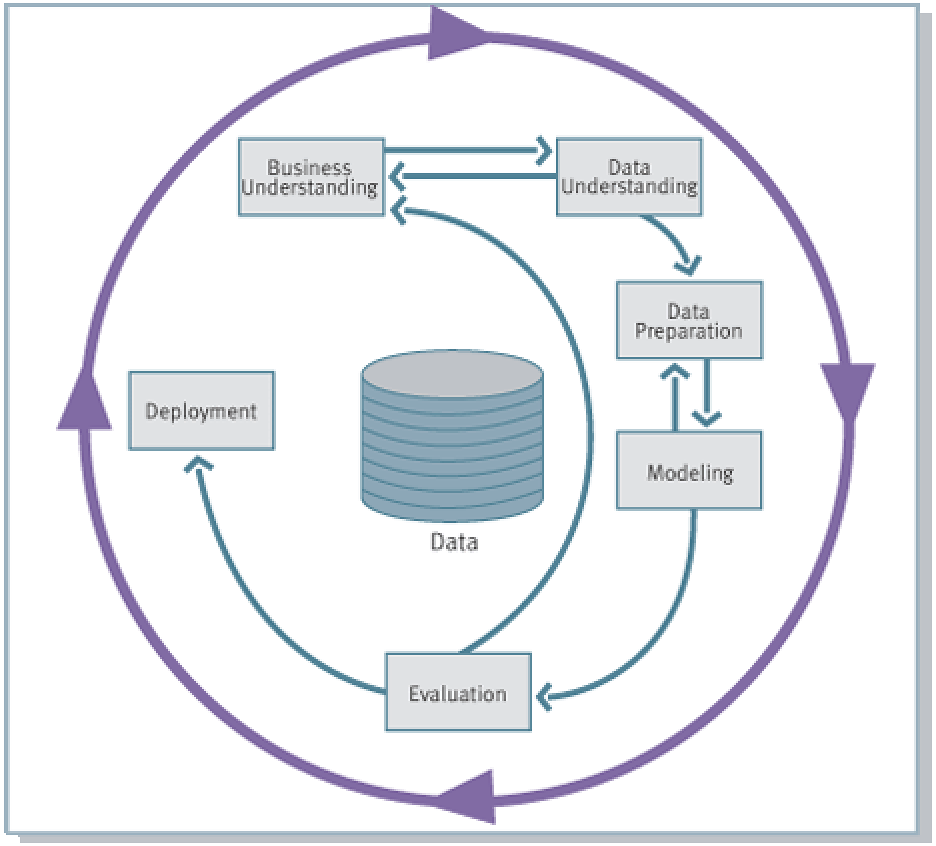
The Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



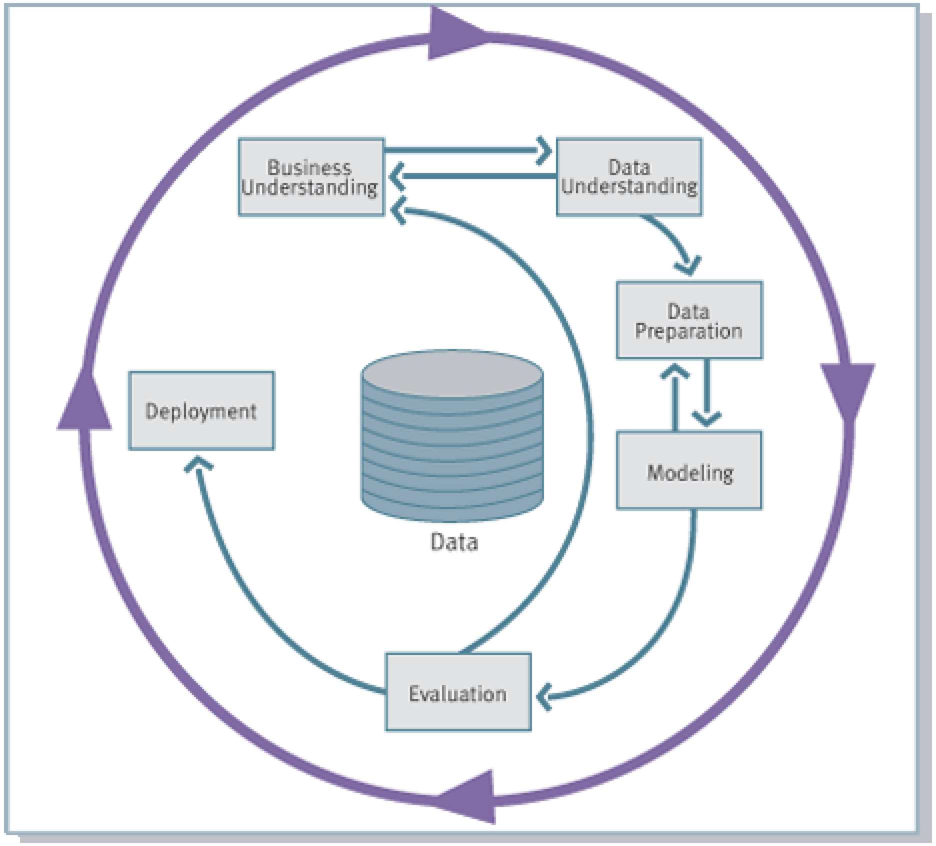
U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine 17(3): 37-54 (1996)

CRISP-DM



- **Evaluation**
  - current model should have high quality from a data mining perspective
  - before final deployment, it is important to test whether the model achieves all business objectives

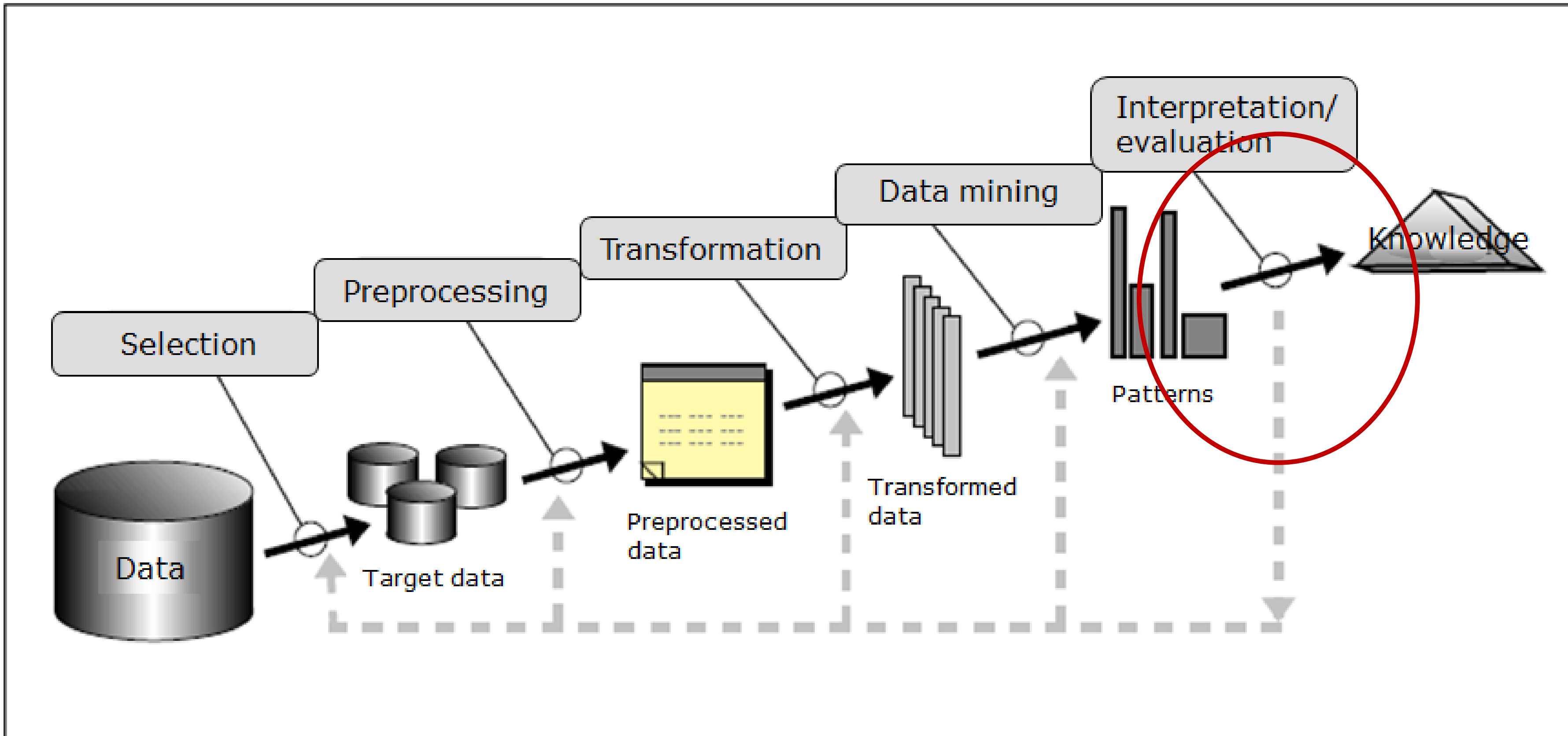
CRISP-DM



- **Deployment**
  - just creating the model is not enough
  - the new knowledge should be organized and presented in a usable way
  - generate a report
  - implement a repeatable data mining process for the user or the analyst

The Knowledge Discovery Process

Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.



U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine 17(3): 37-54 (1996)

Tools

- Many data mining tools are freely available
- Some options are:

Tool	URL
WEKA	www.cs.waikato.ac.nz/ml/weka/
Rule Discovery System	www.compumine.com
R	www.r-project.org/
RapidMiner	rapid-i.com/

More options can be found at www.kdnuggets.com

A Simple Problem

- Given a stream of labeled elements, e.g.,

{C, B, C, C, A, C, C, A, B, C}

- Identify the majority element: element that occurs > 50% of the time

- Suggestions?

Naïve Solution

- Identify the corresponding “alphabet”

{A, B ,C}

- Allocate one memory slot for each element
- Set all slots to 0
- Scan the set and count

Naïve Solution

- Stream of elements

{C, B, C, C, A, C, C, A, B, C}

- Counter_A = 2
- Counter_B = 2
- Counter_C = 6

Naïve Solution

- Stream of elements

{C, B, C, C, A, C, C, A, B, C}

- Counter_A = 2
- Counter_B = 2
- $\boxed{\text{Counter\_C = 6}}$

Efficient Solution

- X = first item you see; count = 1
- **for** each subsequent item Y

  if (X==Y) count = count + 1

  else {

  count = count - 1

  if (count == 0) {X=Y; count = 1}

  }

  endfor

- Why does this work correctly?

Efficient Solution

- Stream of elements

{**C**, B, C, C, A, C, C, A, B, C}

- Counter = 1
- X = C

Efficient Solution

- Stream of elements

{C, **B**, C, C, A, C, C, A, B, C}

- Counter = 1
- X = C
- Y = B

Efficient Solution

- Stream of elements

{C, **B**, C, C, A, C, C, A, B, C}

- Counter = 1
- X = C
- Y = B
- Y != X

Efficient Solution

- Stream of elements

{C, **B**, C, C, A, C, C, A, B, C}

- Counter = 0
- X = C
- Y = B
- Y != X

Efficient Solution

- Stream of elements

{C, **B**, C, C, A, C, C, A, B, C}

- Counter = 1
- X = B

Efficient Solution

- Stream of elements

{C, B, **C**, C, A, C, C, A, B, C}

- Counter = 1
- X = C

Efficient Solution

- Stream of elements

{C, B, C, **C**, A, C, C, A, B, C}

- Counter = 2
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, **A**, C, C, A, B, C}

- Counter = 1
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, A, **C**, C, A, B, C}

- Counter = 2
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, A, C, **C**, A, B, C}

- Counter = 3
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, A, C, C, **A**, B, C}

- Counter = 2
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, A, C, C, A, **B**, C}

- Counter = 1
- X = C

Efficient Solution

- Stream of elements

{C, B, C, C, A, C, C, A, B, **C**}

- Counter = 2
- X = C

Why does this work?

- Stream: n elements
- M: majority element that occurs x times
- Counter is set to 1
- If M occurs first:
  - Counter will increase $x - 1$ times and decrease $n - x$ times
- If M does not occur first:
  - Counter will increase x times and decrease $n - x - 1$ times
- Hence, eventually:

$$\text{Counter} = 1 + x - (n - x) - 1$$

$$\Rightarrow \quad \text{Counter} = 2x - n$$

Why does this work?

So far, we have:

$$\text{Counter} = 2x - n$$

If n is even, then

$$x > n/2$$

Hence,

$$\text{Counter} > 2(n/2) - n$$
$$\Rightarrow \quad \text{Counter} > 0$$

If n is odd, then

$$x > n/2 + 1$$

Hence,

$$\text{Counter} > 2(n/2 + 1) - n$$
$$\Rightarrow \quad \text{Counter} > 2$$

Why does this work?

So far, we have:

$$\text{Counter} \quad = \; 2x - n$$

Hence, in both cases,

if the majority element exists:

Counter > 0

$$\text{Counter} \quad > \quad 2(n/2) - n$$
$$\Rightarrow \quad \text{Counter} \quad > \quad 0$$

$$\text{Counter} \quad > \quad 2(n/2 + 1) - n$$
$$\Rightarrow \quad \text{Counter} \quad > \quad 2$$

Today

- Why do we need data analysis?
- What is data mining?
- Examples where data mining has been useful
- Data mining and other areas of computer science and statistics
- Some (basic) data-mining tasks

Next time

DATE	TIME	ROOM	TOPIC
MONDAY 2013-09-09	10:00-11:45	502	Introduction to data mining
WEDNESDAY 2013-09-11	09:00-10:45	501	Decision trees, rules and forests
FRIDAY 2013-09-13	10:00-11:45	Sal C	Evaluating predictive models and tools for data mining