Introduction to Data Mining

- What is data mining?
- Data mining process

Image sources: https://www.pikist.com/
https://howtofindgoldnuggets.com/metal−detect−fringe−spots−find−gold−nuggets/

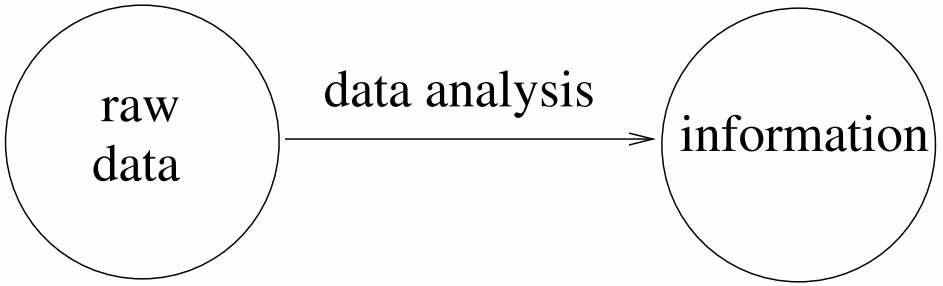
What is Data mining (DM)?

- no definite and clear answer
- computationally nontrivial data analysis for finding new useful **information** from large collections of **data**
  - interesting patterns like relationships and groupings
- Challenge: data volumes are all the time inreasing!

  $\Rightarrow$ more efficient algorithms needed

  $\Rightarrow$ number of patterns and spurious discoveries increases $\Rightarrow$ How to find interesting and reliable patterns?

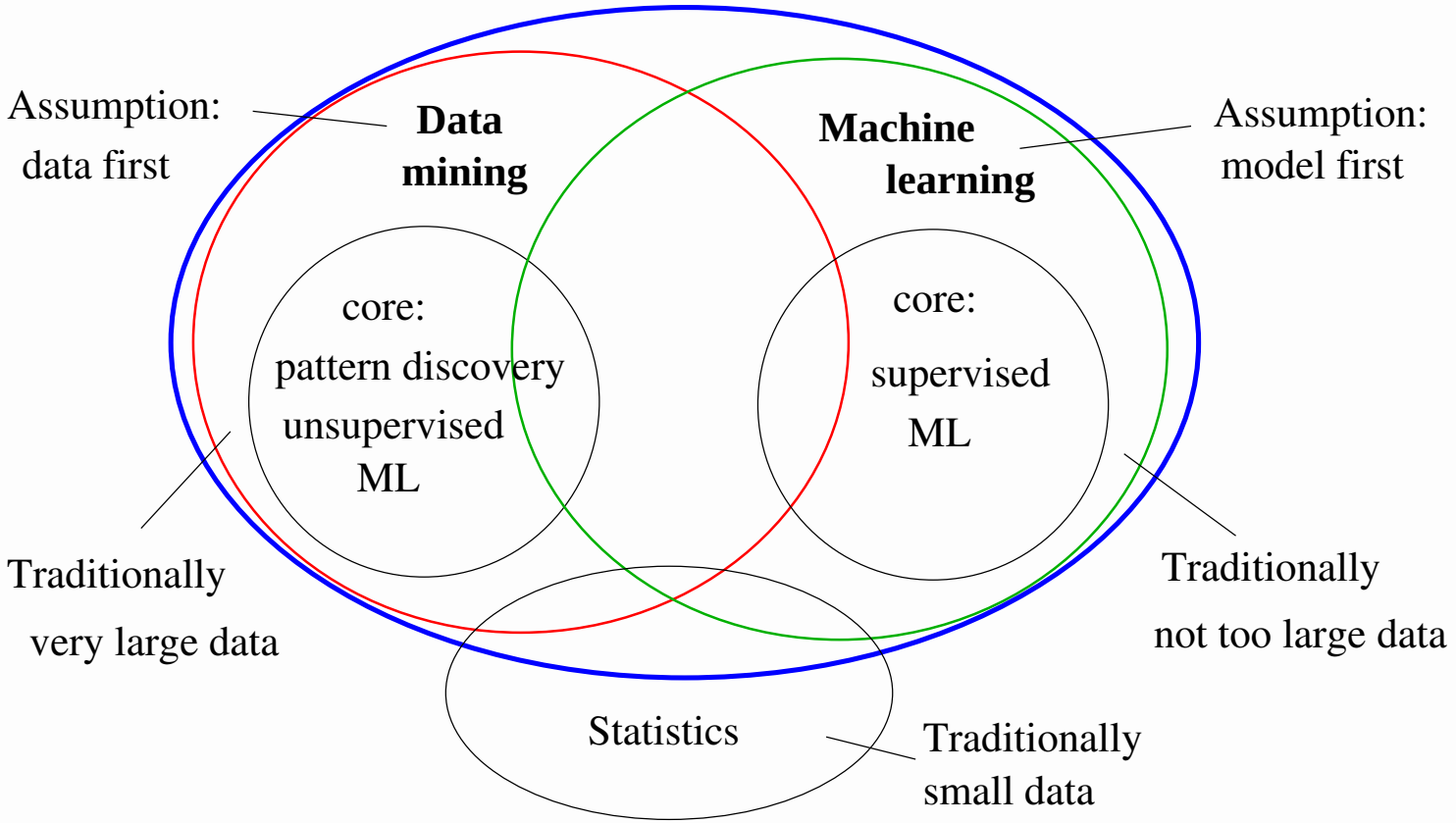
Data vs. Information?



- raw data = unprocessed, uninterpreted facts (e.g, measurements)
- information = knoweldge that has meaning, “interpreted data”
- relative terms: the resulting information from one process may be source data for another process

Relationship to closest neighbouring fields

DM ~ **knowledge discovery (from databases)** (KDD)
Machine learning strongly overlapping/synonymous!



Model vs. pattern?

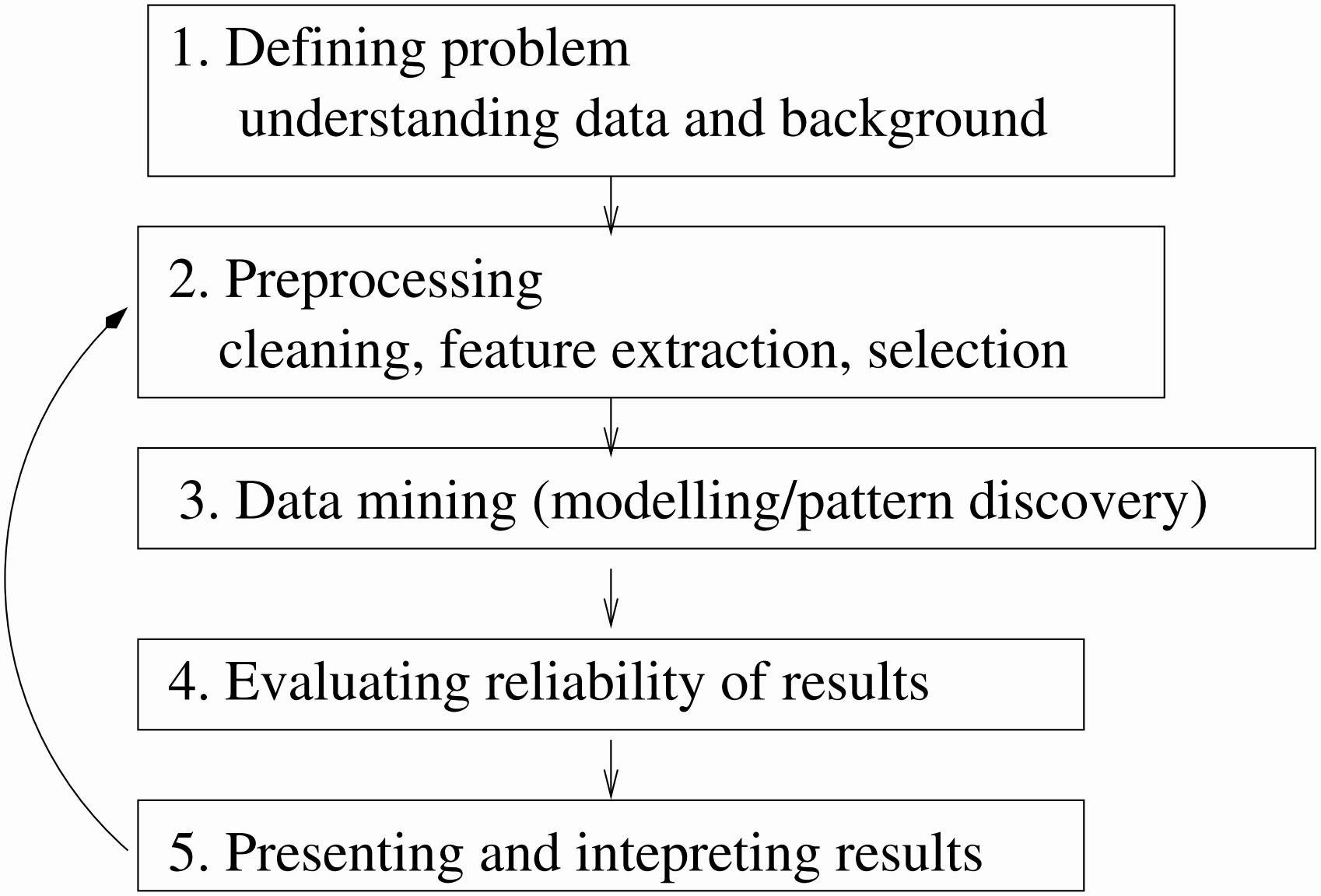
Model

- global (fits entire data)
- e.g., course success (passing the course) can be predicted from exercise points, time spent on course and participation in exercise groups

Pattern

- local model (describes some part of data)
- e.g., if students obtain high points in assignment 2 they tend to obtain high points also in the exam task 3

DM process


1. Defining problem
understanding data and background
2. Preprocessing
cleaning, feature extraction, selection
3. Data mining (modelling/pattern discovery)
4. Evaluating reliability of results
5. Presenting and intepreting results

1. Defining the problem

- Understanding data: what variables measure/describe?
- What are data types? How much there is data?
- What kinds of patterns would be interesting or useful to find?
- What is already known?
- It is worth studying some background theory!
- Difficulty: How people from different fields find the same language?

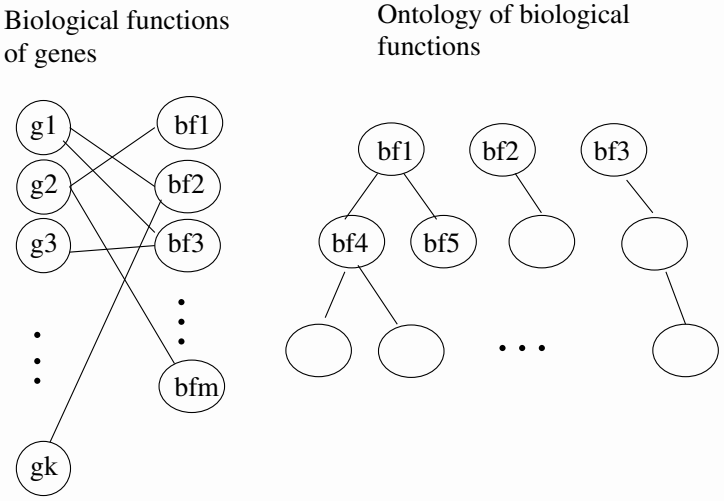
Example: defining problem

Medical scientist: How TNF-α stimulation affects gene regulation in prostate cancer cells and which biological functions are involved?

Computer scientist: First, explain the data

Data matrix: expression of genes g1,...,gk and class

id	expression values of g1	g2	g3	...	gk	class
s1	7.1	2.3	4.6		3.1	cancer
s2	6.5	8.0	4.9		5.4	healthy
s3						cancer
						healthy
						healthy
						cancer



So, I should find g_is that differ significantly in two groups and corresponding bfs?

2. Preprocessing

- Combining data from different sources (may require transformations)
- Preliminary analysis: means, standard deviations and distributions of variables, correlations, ... (e.g., with statistical tools)
- Data cleaning: handling missing values, detecting and correcting errors
- Feature selection and extraction
- Possibly dimension reduction (combines feature extraction and selection)

3. Data mining

- Typical building blocks dependency analysis, classification, clustering, outlier detection
- Always good to begin from dependency analysis! → choosing features and modelling methods
- Usually descriptive modelling helps in building a predictive model
  - e.g., gene–habit–disease data
  - Descriptive: Find 100 most significant association rules related to variable Diabetes
  - Predictive: Learn (from selected data) the best model that predicts diabetes

4. Evaluating reliability of results

- Are discovered patterns or models sensible?
  - it is possible there are no models or patterns in the data – but **the methods tend to return something even from random data!**
- validating predictive models easy (test set, cross validation)
- evaluating reliability of descriptive models more difficult
- Goal: Some guarantees that the **discovered pattern is not due to chance**
- tools: statistical significance testing, use of validation data

5. Presenting and interpreting results

- present results illustratively so that essential things are emphasized
- domain experts have an important role!
- Did you find something new? Could you formulate a hypothesis based on results? What should be studied further?
- leads often to a new DM round; try new variables and possibly other methods
- finish the iterative DM process when you are satisfied or nothing new seems to be discovered