Ciro Donalek (Caltech)

Introduction to Machine Learning

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Objective

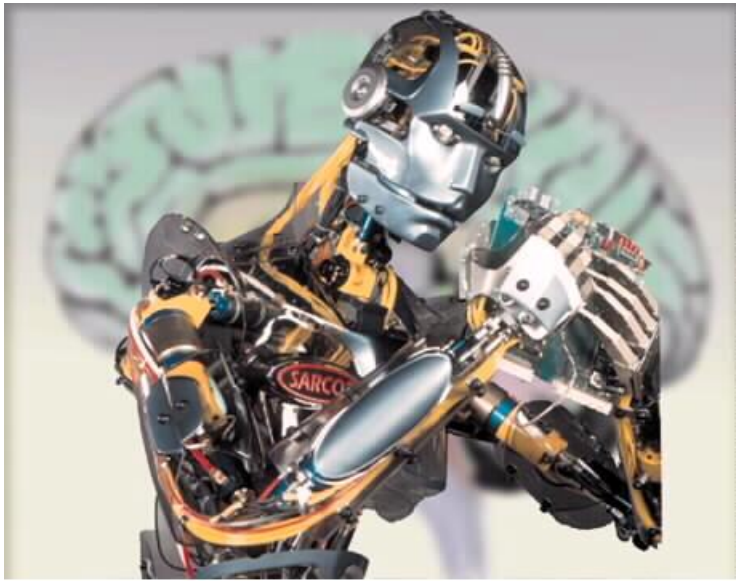
- What is Machine Learning (ML) and why we need it
- Types of learning
- Data Mining
- Knowledge Discovery in Databases

ML: historic definition

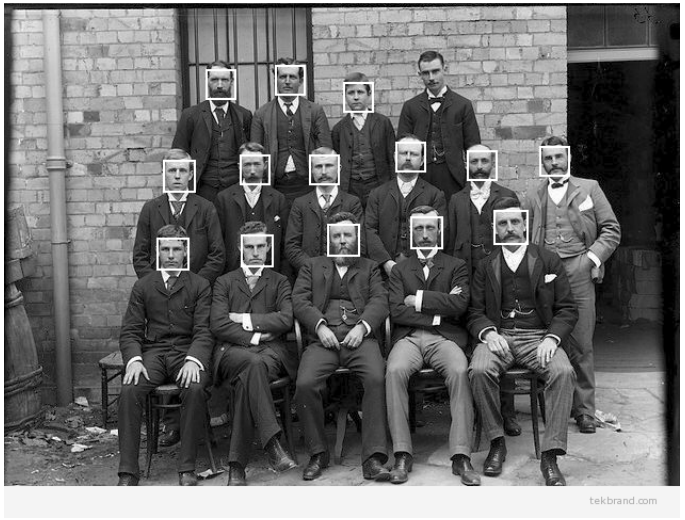
- Machine Learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (Arthur Samuel – 1959).

ML: a more formal definition

- "A computer program is said to learn from experience ***E*** with respect to some class of tasks ***T*** and performance measure ***P***, if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E***" (T. Mitchell, '98).
- Example: SPAM filter
  - ***T***: recognize spam
  - ***E***: user flagging mail as spam
  - ***P***: percentage of spam correctly classified as such



Applications



Face detection / automatic tagging

Medical Diagnosis

Advertisement, product engagement

Science

Why we need a ML approach

- We are in an era dominated by large, multi-dimensional and heterogeneous datasets (**most of which will never be seen by humans**).

- Big Data is often described using five **V***s*:
  1. **V**olume: large amount of data;
  2. **V**elocity: data generated and moved around at great speed;
  3. **V**ariety: different sources;

Why we need a ML approach

- We are in an era dominated by large, multi-dimensional and heterogeneous datasets (**most of which will never be seen by humans**).

- Big Data is often described using five **V***s*:
  4. **V**eracity: quality and accuracy are less controllable;
  5. **V**alue: **extract knowledge from data!**

When a ML approach can be applied?

- We need to extract knowledge
- We know that a pattern exists
  - if there is no pattern, no harm done!
- We cannot pin it down mathematically
  - If we can, may be *just* not the recommended technique…
- We have (lots of) data
  - data driven approach

Types of Learning

- Different types of Learning:
  - Supervised Learning
  - Unsupervised Learning
  - Semi-Supervised Learning
  - Other types:
    - Reinforcement Learning
    - Transduction
    - Learning to Learn

Quick definitions

- **Supervised Learning**: for some of the samples, the desired output is known and it is used during the training process.
- **Unsupervised Learning**: the model is not provided with the correct results during the training; can be used to cluster the input data in classes on the basis of their statistical properties only.
- **Semi-Supervised Learning**: combines both labeled and unlabeled examples to generate an appropriate function or classifier.

Algorithms

- There are many good tools out there, but you need to choose the right ones for your needs.
- No “one size fits all” solution.

Supervised Algorithms
Neural Networks (MLP)
Boltzmann Machines
RBM
Decision Trees
Nearest Neighbor
Naive Bayes Classifiers
Bayesian Networks
Gaussian Processes
Regression
…

Unsupervised Algorithms
K-Means
Self-Organizing Maps
RDF
Fuzzy Clustering
CURE
ROCK
Vector Quantization
Probabilistic Principal Surfaces
…

Choose the right approach

- What is being learned?
- How data is generated?
- Is there any missing / incomplete data?
- How the data is presented (one at a time, all in once)?
- What is the goal?

Data Mining

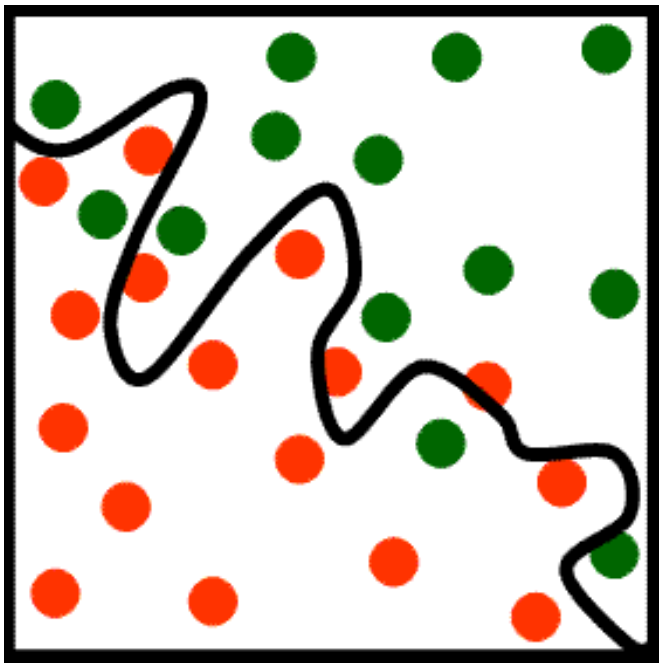
- Machine Learning and Data Mining are often confused.
- Data Mining can be defined as the process of extracting ("*mining*") knowledge and **interesting** data patterns hidden in (*large*) datasets, often using Machine Learning techniques.
- **Interesting** = "non-trivial, previously unknown and potentially useful".
- Data Mining is the analysis step of Knowledge Discovery in Databases (KDD).

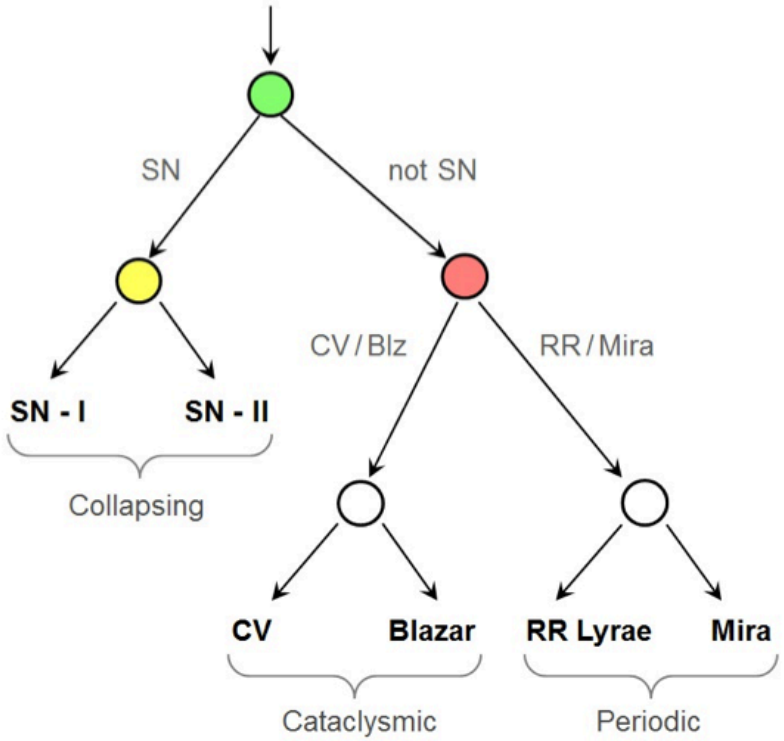
Data Mining Tasks

- Classification
- Regression
- Forecasting
- Clustering
- Search for Outliers (Deviation Analysis)
- Path Analysis (Sequence Analysis)
- Association (Market Basket Analysis)

DM tasks: Classification

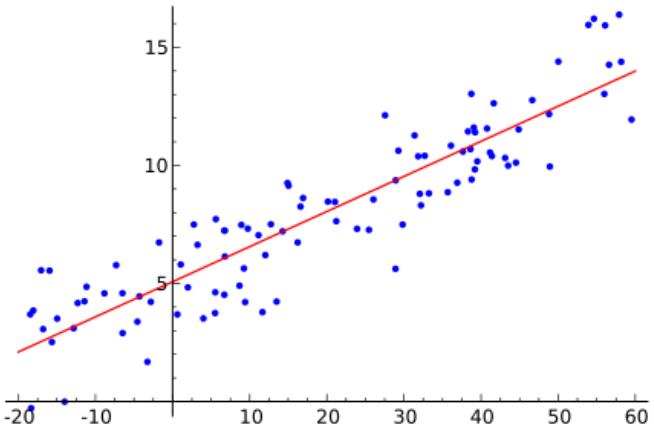
- Assign samples into categories (classes) based on a predictable attribute.
- The goal of classification is to accurately predict the target class for each case in the data set.

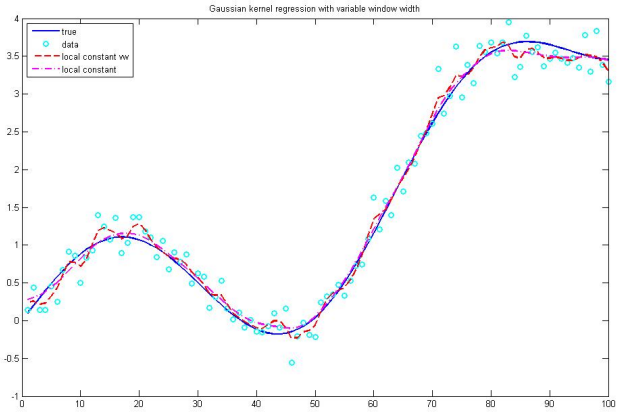




DM tasks: Regression

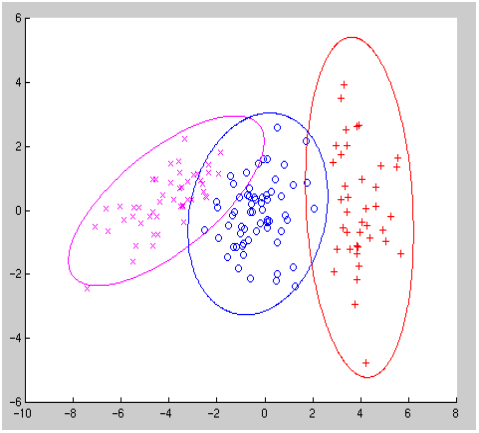
- Compute the new values for a dependent variable based on the values of one or more measured attributes.
- Examples:
  - predict wind velocities based on temperature, air pressure and humidity;
  - predict coupon redemption rate based on the face value, distribution method and distribution volume



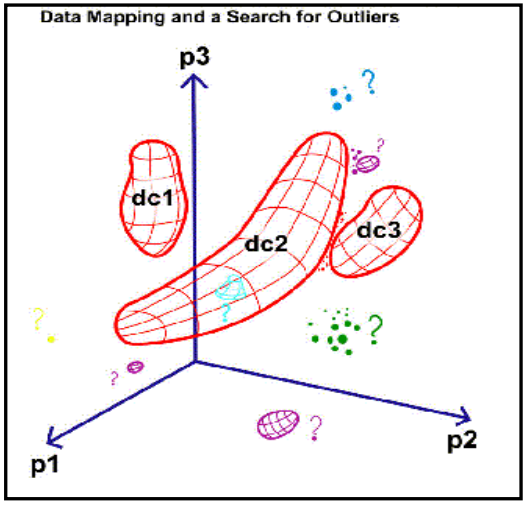


DM tasks: clustering

- Clustering
  - partitioning of a data set into subsets (clusters) so that data in each subset ideally share some common characteristics.



- Deviation Analysis (search for outliers)
  - anomalies;
  - peculiar objects.



DM tasks: Path Analysis

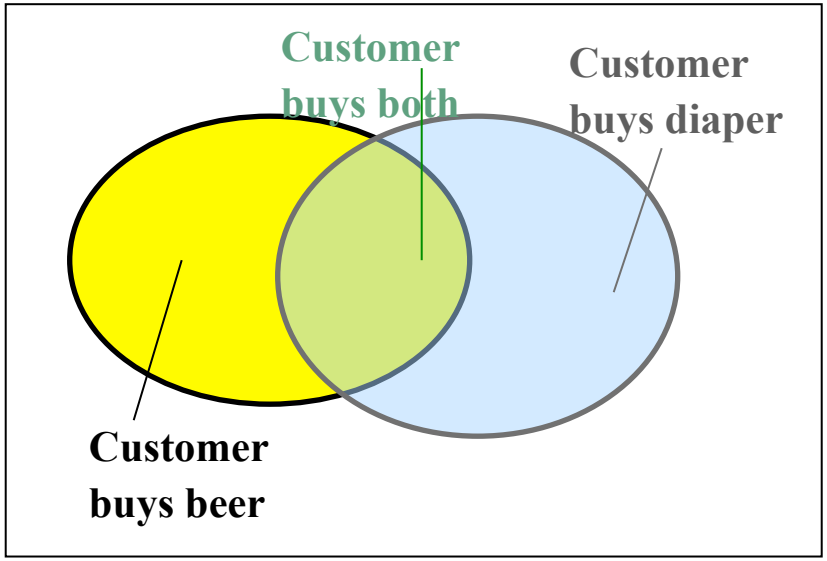
- Path Analysis (or Sequence Analysis) is used to find patterns in discrete series.
- Looking for patterns in which one event leads to a later event.
- In the sequence analysis also the order is important.
  - buying item A before buying item B is different from buying item B before A.

Web path analysis: mine log file, excellent way to get knowledge about how visitors use a site.
Generating set of rules that can predict which users are likely to click on a particular link or make a purchase.

Eg, Google AdSense/Analytics

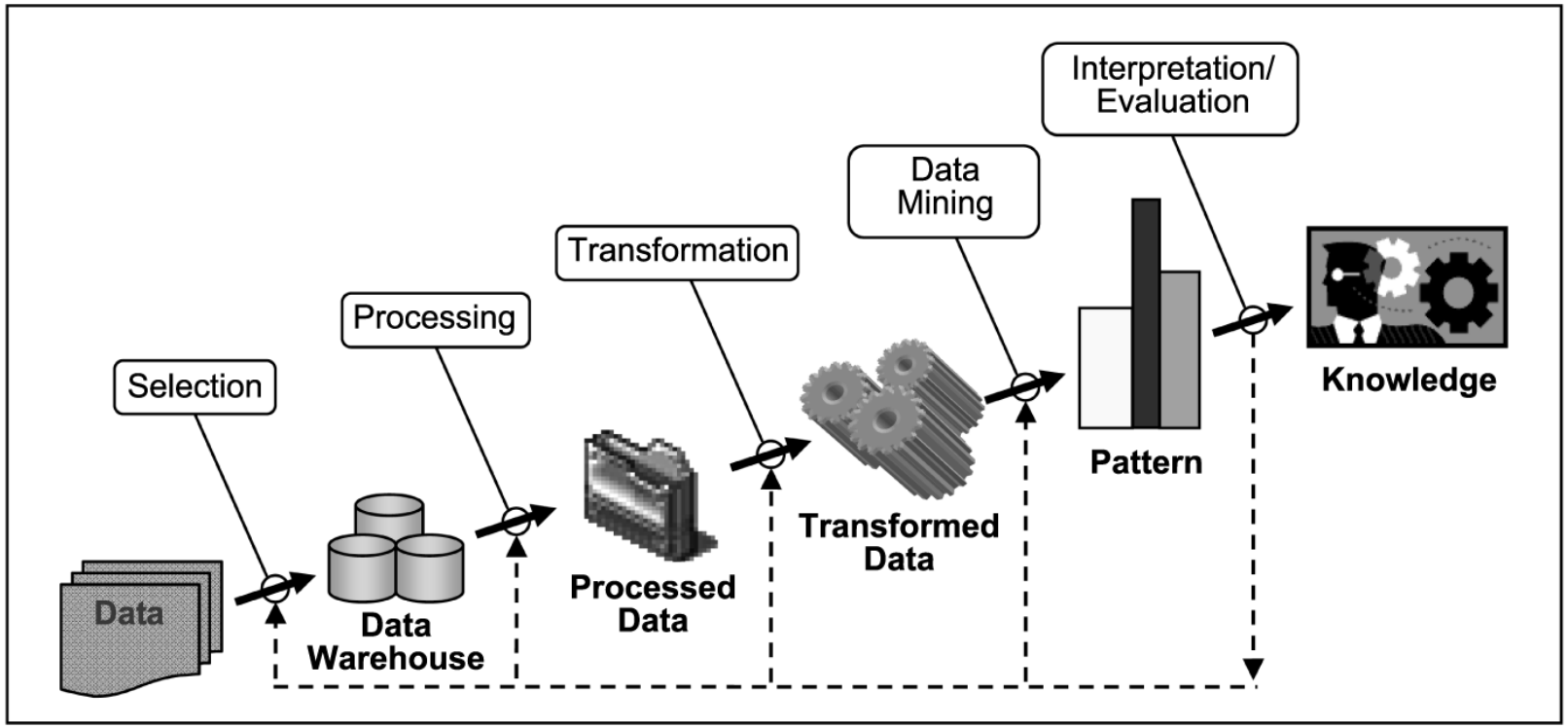
DM tasks: Association

- Association: finding inherent regularities in data.
- Looking for frequent patterns (eg. "*What items are frequently purchased together?*", cross-marketing).
- The order is not important:
  - buying item A before buying item B is no different of buying item B before A (all the items are equal and independent).



Knowledge Discovery in Databases

- "KDD is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al, 1996).



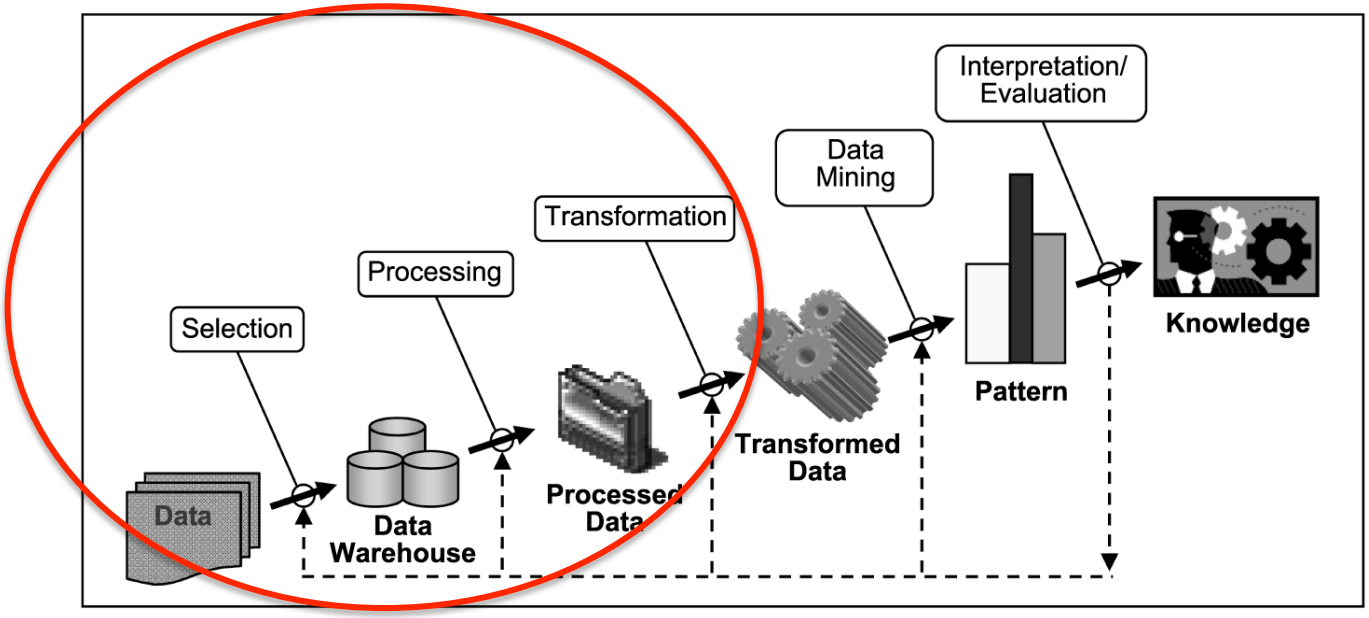
Picture credit: An-Pin Chen, Chia-Chen Chen

KDD steps: pre-processing

① Creating the data set: data integration (i.e., merging data from different sources), a priori knowledge, etc.

② Data cleaning: noise removal, handling missing data, data cleaning, put the data in the right format.

③ Data reduction and transformation: feature selection, feature extraction.

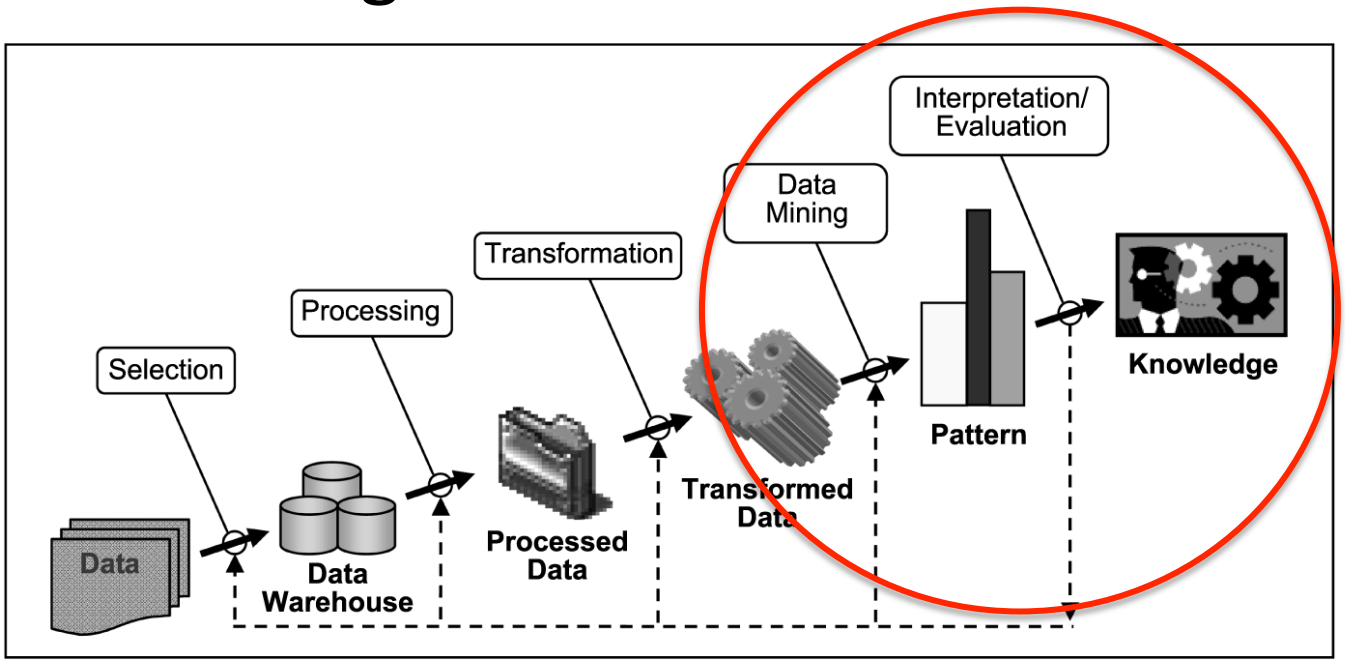


KDD steps: mining and evaluation

④ Data mining: choose the right task, choose the machine learning model, perform the analysis.

⑤ Interpretation and Evaluation: evaluate and interpret the discovered patterns.

⑥ Use the knowledge!



Summary

- Introduction to Machine Learning
- Different types of Learning
- Data Mining
- Knowledge Discovery in Databases