Topics in Data Engineering

Session 2

Masaomi Kimura

Topics in this session

- What is data mining? (again)
- Models
- Major analysis methods

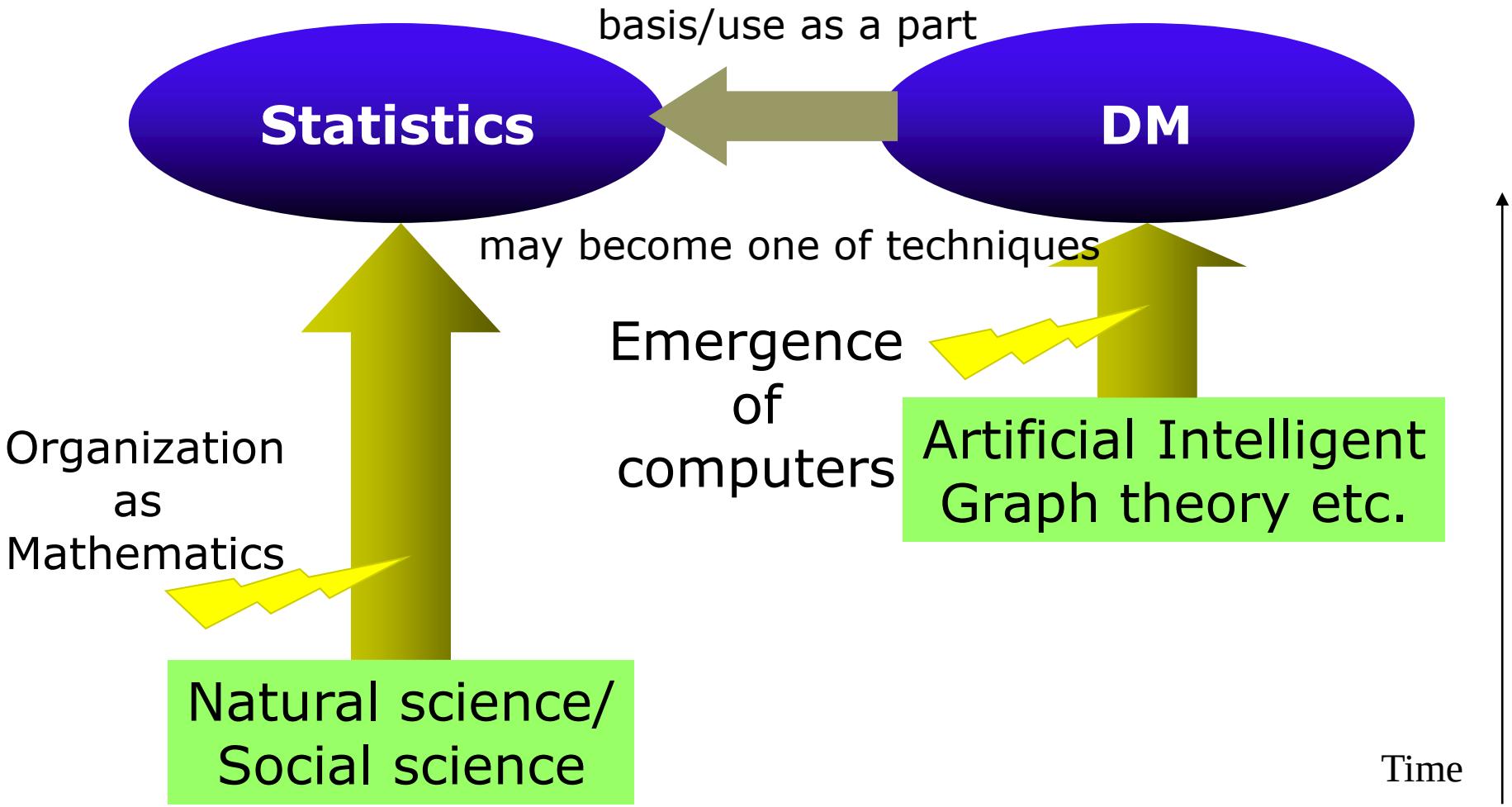
What is data mining (DM)?

1. Store huge amount of data
2. Select method suitable to the objective of the analysis
3. Modify the data format to the one suitable the method (data cleaning)

After applying the method and obtaining results,

4. interpret the results based on business knowledge
5. and improve the business

Relation between statistics and DM



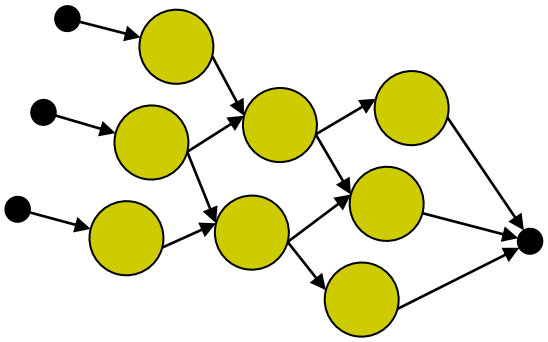
Difference of statistics and DM

- Statistics
  - describes data distribution by small number of parameters (means, variants)
  - requires a null hypothesis for a test
    - target is small size of data out of a population
- Data mining
  - is used to categorize data, to generate rules, to predict, etc.
  - uses all data
  - assumes no null hypothesis

Points to learn data mining technique

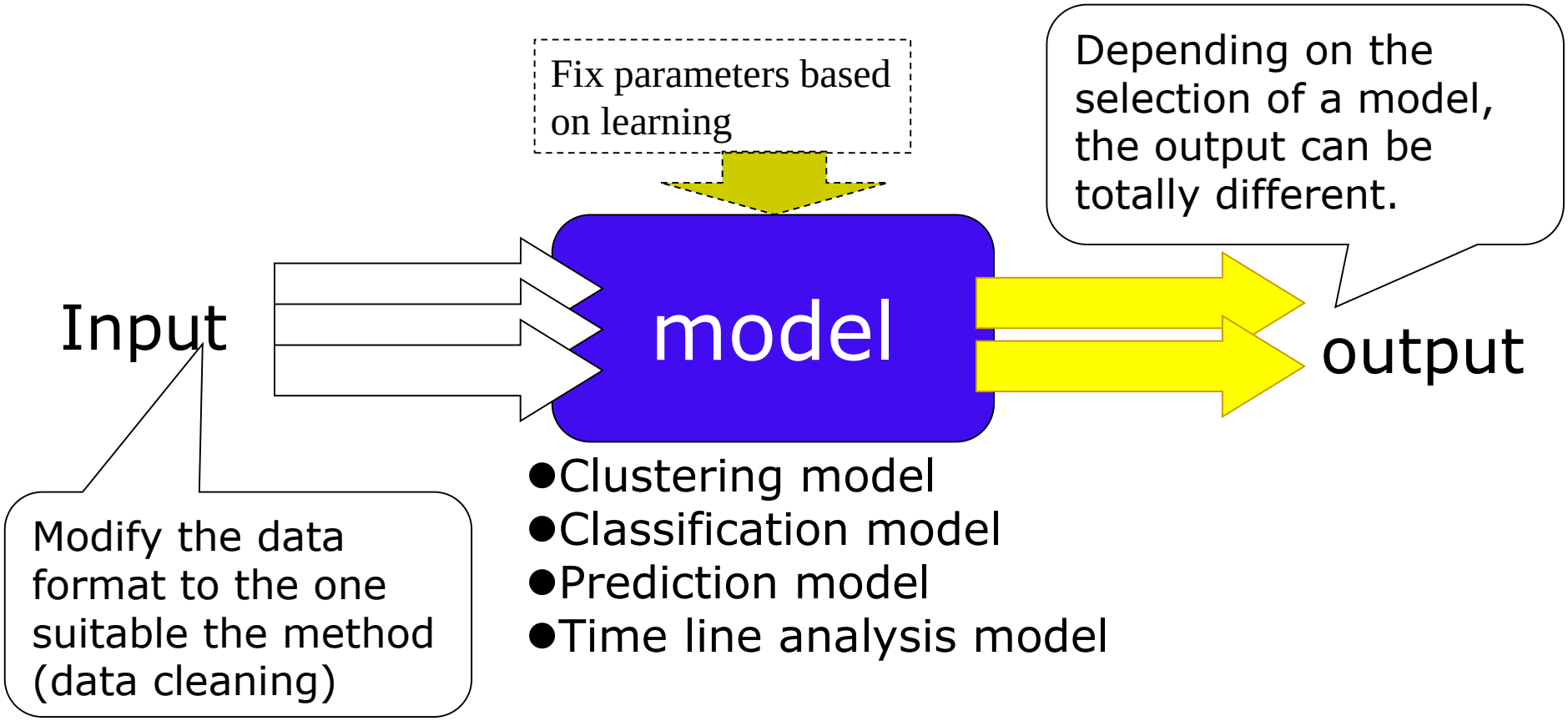
- Understand the merit/demerit of each of DM methods
- Investigate the previous cases of applications of DM methods
  - as a reference to judge whether information enough to take an action can be obtained or not

P(X|Y)/P(Y)>1?



0 or *1*?

Model


Fix parameters based on learning
Depending on the selection of a model, the output can be totally different.
Input
model
output
Modify the data format to the one suitable the method (data cleaning)
Clustering model
Classification model
Prediction model
Time line analysis model

Types of models

- Verification-oriented models
  - require an analyst's assumption of results
  - are, mainly, used to classify data or predict based on model parameters learnt in advance
- Discovery-oriented models
  - do not require assumption of results
  - are used to understand data
  - are, mainly, used to find clusters, etc.

Overfit/underfit

- Overfitting
  - a model having too many learning parameters and having learnt characteristics only existing in training data
  - the model does not give good result for generic data
- Underfitting
  - a model having insufficient parameters to extract meaningful information that can be applicable to generic data

Explainability

- Models whose results are easy to be explained
  - clustering, association analysis, decision tree, etc.
- Models whose results are difficult to be explained
  - neural network

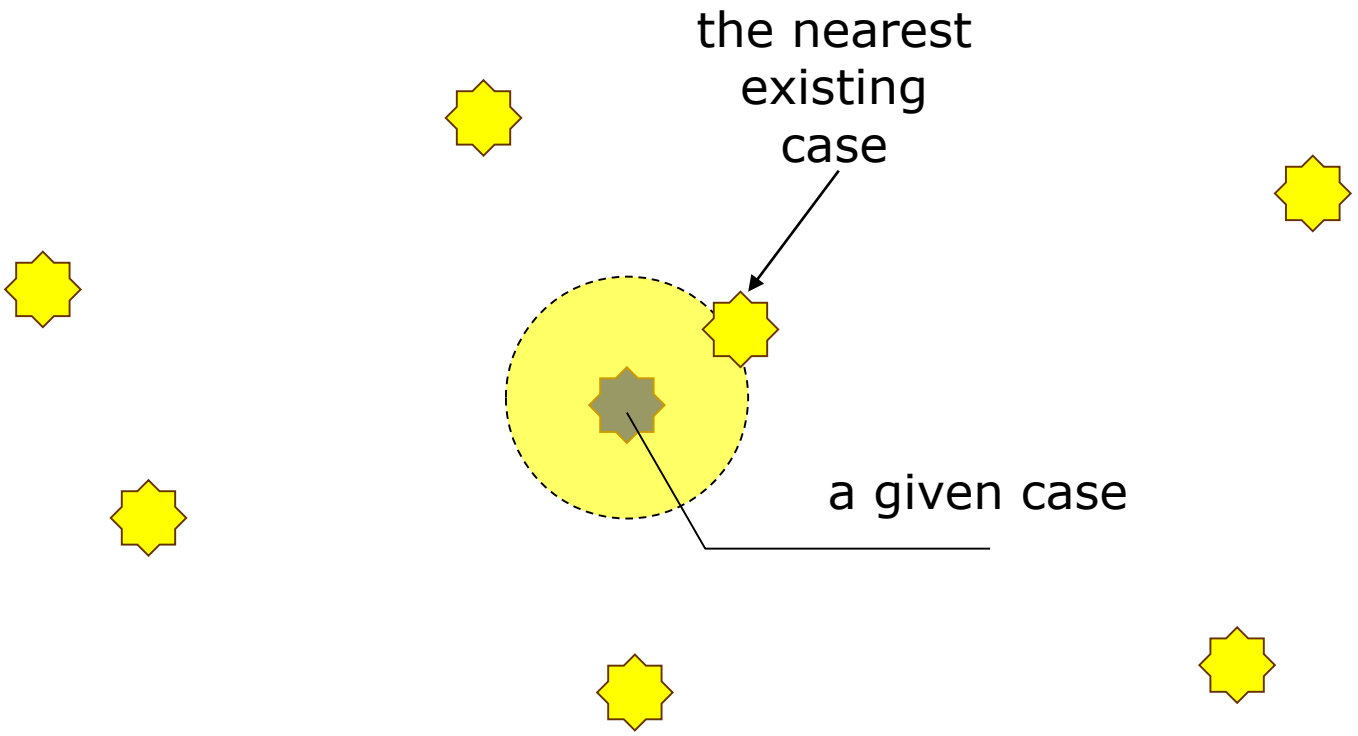
1.Association analysis

diapers

Beer

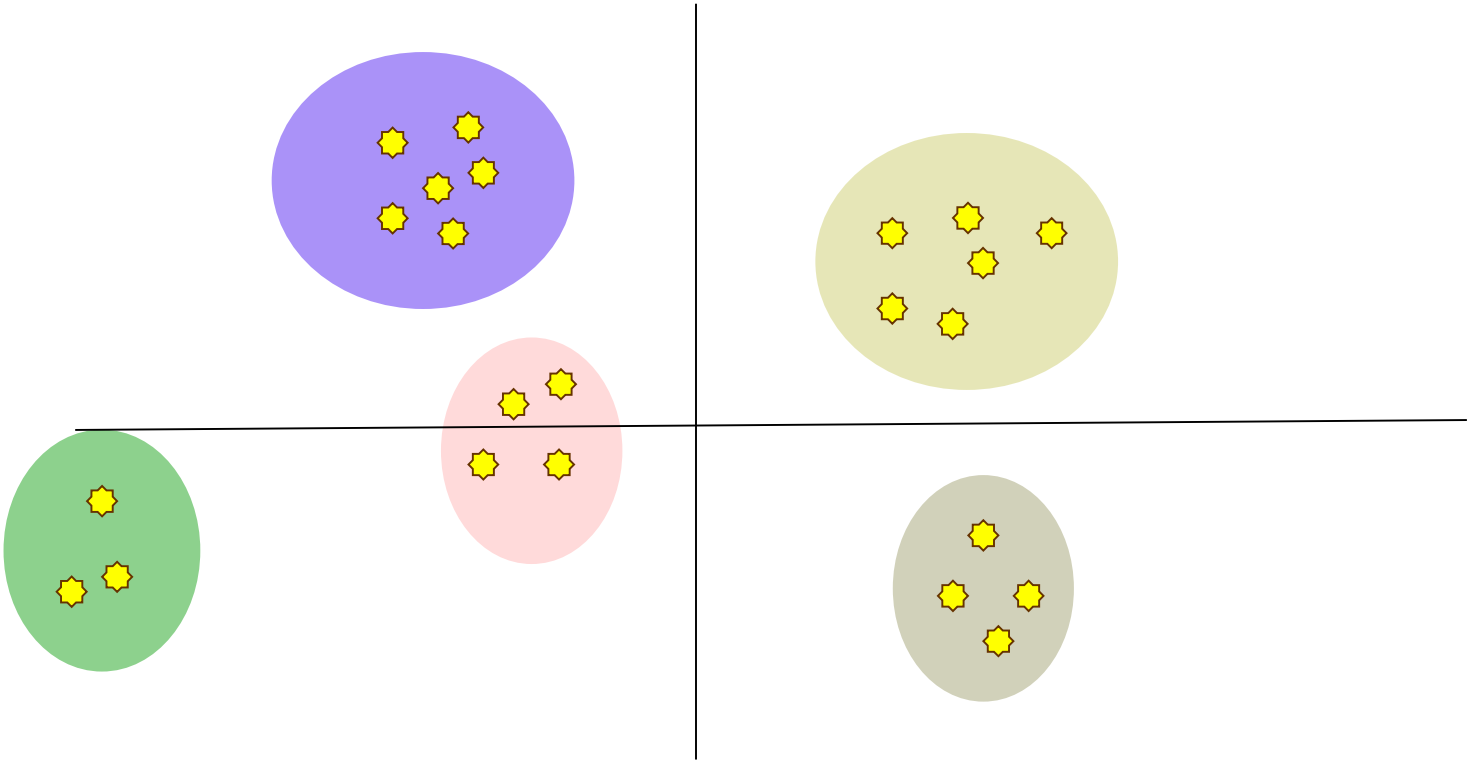
snacks

2.Memory based reasoning



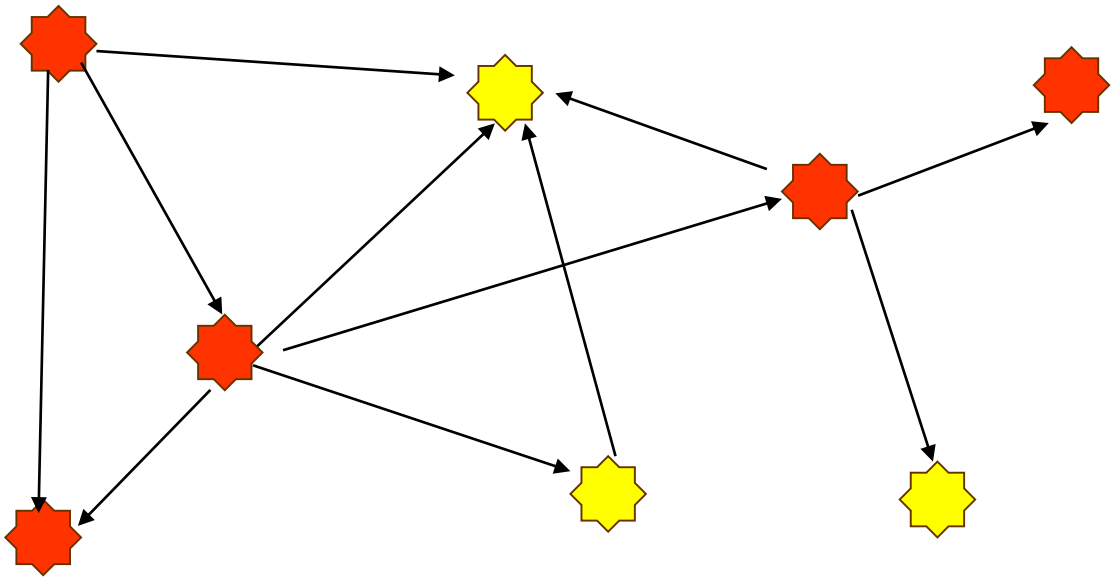
- A distance measure and a combination function are necessary

3.Clustering



- Finds clusters, within whom data are similar

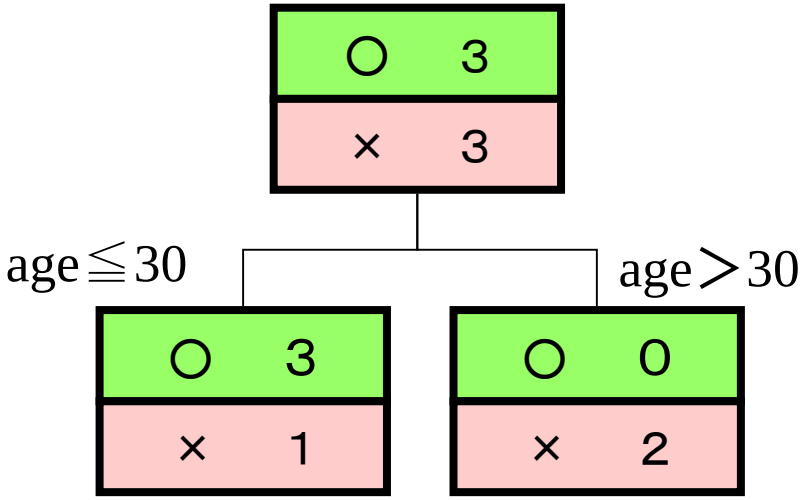
4. Link analysis



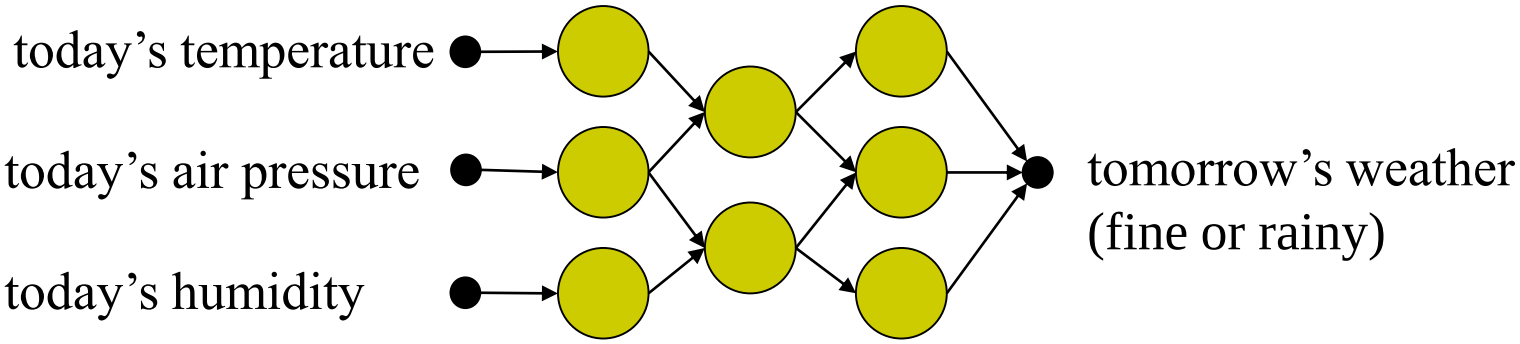
- Finds patterns in links of data based on graph theory or network theory

5. Decision trees

Want to buy?	ages
○	21
○	25
○	30
×	29
×	50
×	60

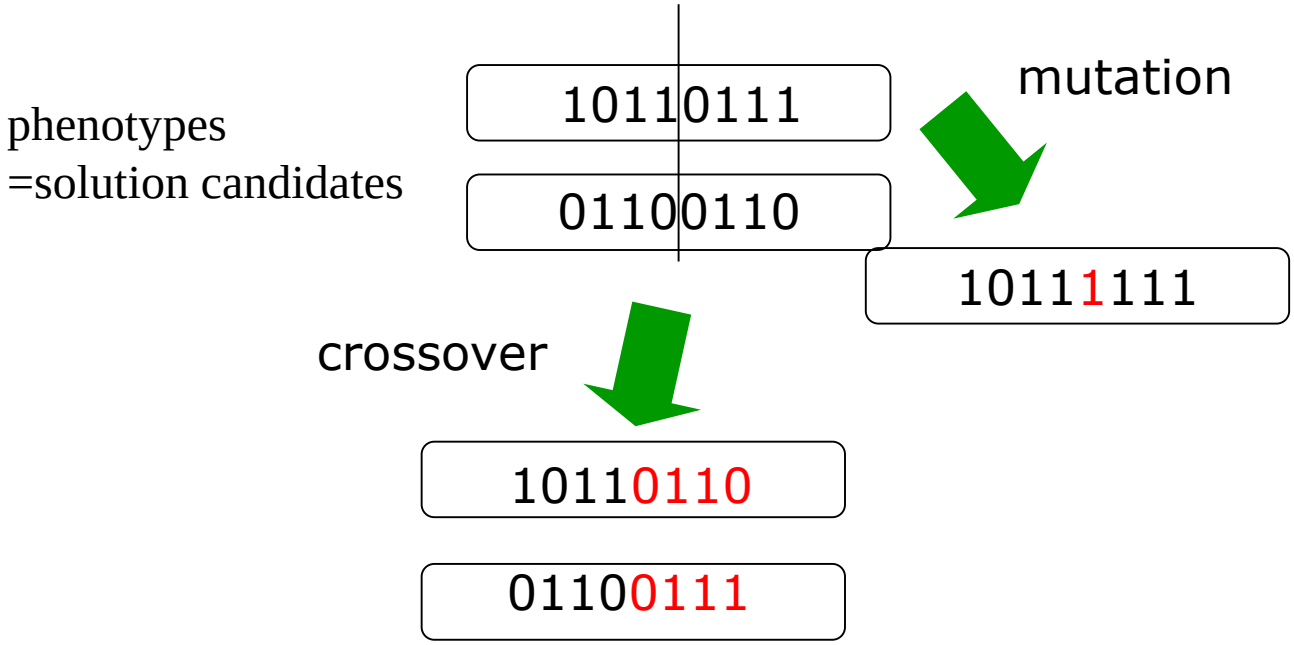


6. (Artificial) neural networks



- is a model of neuron networks in a brain
- learns patterns between input and corresponding output in training data

7. Genetic algorithm(GA)



- is a model to find a solution that maximizes a given function
- is based on the idea of evolution of species (animals)