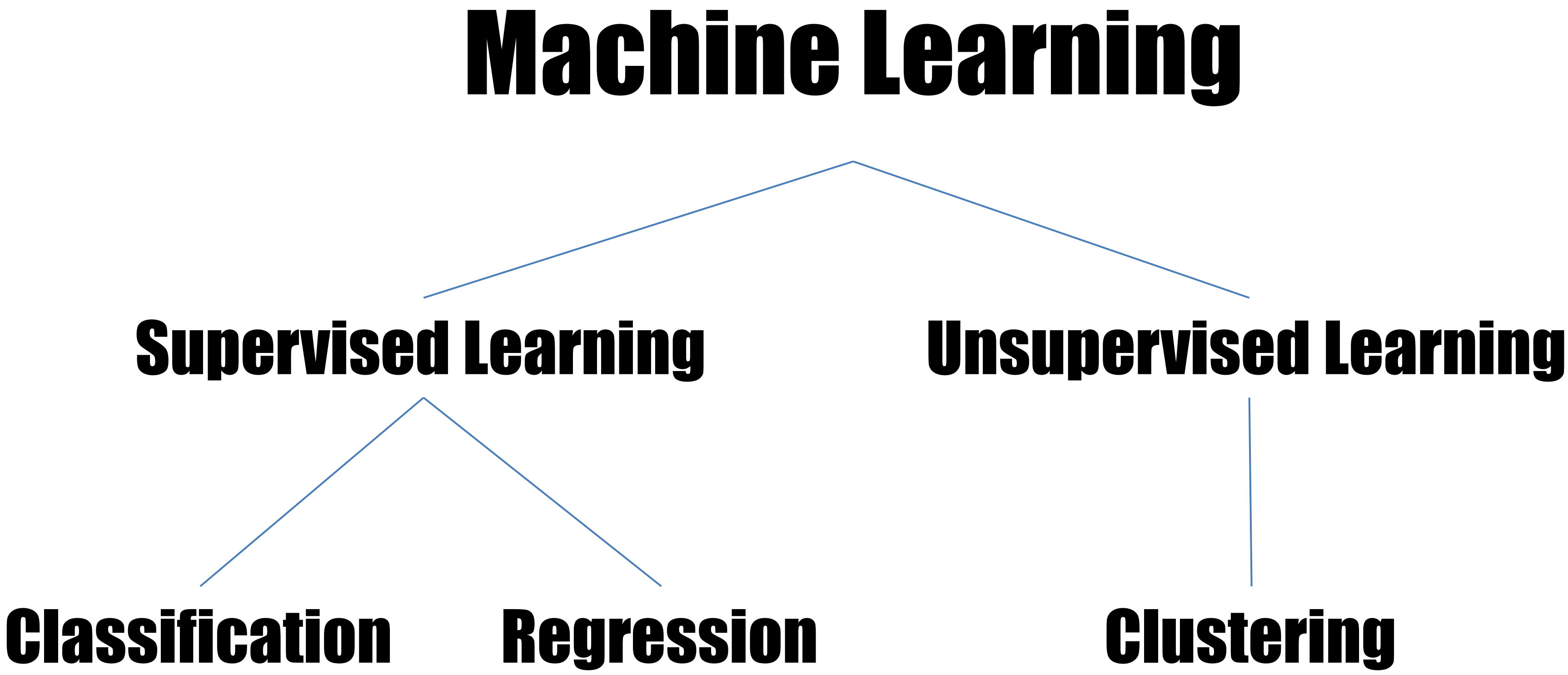
Machine Learning

Boris Velichkov

Machine Learning?

How do we learn?


Machine Learning
Supervised Learning
Unsupervised Learning
Classification
Regression
Clustering

Classification

	Attribute 1	Attribute 2	…	…	Attribute N	Class
X1						„Yes“
.						„No“
.						„Yes“
. Xm						„Yes“

Dataset

Xt = (a1, a2, … , aN)

Class(Xt) = ?

Classification

Dataset

	Humidity	Wind	Temperature	Class (Play Tennis)
X1	high	mid	35°	No
X2	mid	strong	5°	No
X3	mid	weak	20°	Yes
X4	high	weak	25°	Yes
X5	low	strong	18°	No

Xt = (low, weak, 21°)

Class(Xt) = ?

Regression

Dataset

	Attribute 1	Attribute 2	…	…	Attribute N	Class
X1						1.5
.						0.7
.						1.8
.						
Xm						1.2

Xt = (a1, a2, … , aN)

Class(Xt) = ?

Regression

Dataset

	Humidity	Wind	Outlook	Class (T°)
X1	high	mid	rainy	35
X2	mid	strong	overcast	5
X3	mid	weak	sunny	20
X4	high	weak	overcast	25
X5	low	strong	overcast	18

Xt = (low, weak, sunny)

Class(Xt) = ?

Data Preprocessing

Data Preprocessing

- „Garbage in, garbage out"
- Data-gathering methods are often loosely controlled
  - Out-of-range values (e.g., Income: −100)
  - Impossible data combinations (e.g., Sex: Male, Pregnant: Yes)
  - Missing values
- Misleading results

Why Preprocessing ?

- Real world data are generally
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - Noisy: containing errors or outliers
  - Inconsistent: containing discrepancies in codes or names
- Tasks in data preprocessing
  - Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
  - Data integration: using multiple databases, data cubes, or files.
  - Data transformation: normalization and aggregation.
  - Data reduction: reducing the volume but producing the same or similar analytical results.
  - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data Cleaning

- Fill in missing values (attribute or class value):
  - Ignore the tuple: usually done when class label is missing.
  - Use the attribute mean (or majority nominal value) to fill in the missing value.
  - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
  - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.
- Identify outliers and smooth out noisy data:
  - Binning
    - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
    - Then smooth by bin means, bin median, or bin boundaries.
  - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
  - Regression: smooth by fitting the data into regression functions.
- Correct inconsistent data: use domain knowledge or expert decision.

Data Transformation

- Normalization:
  - Scaling attribute values to fall within a specified range.
    - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
  - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev
- Aggregation: moving up in the concept hierarchy on numeric attributes.
- Generalization: moving up in the concept hierarchy on nominal attributes.
- Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data Reduction

- Reducing the number of attributes
  - Data cube aggregation: applying roll-up, slice or dice operations.
  - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).
  - Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..
- Reducing the number of attribute values
  - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
  - Clustering: grouping values in clusters.
  - Aggregation or generalization
- Reducing the number of tuples
  - Sampling

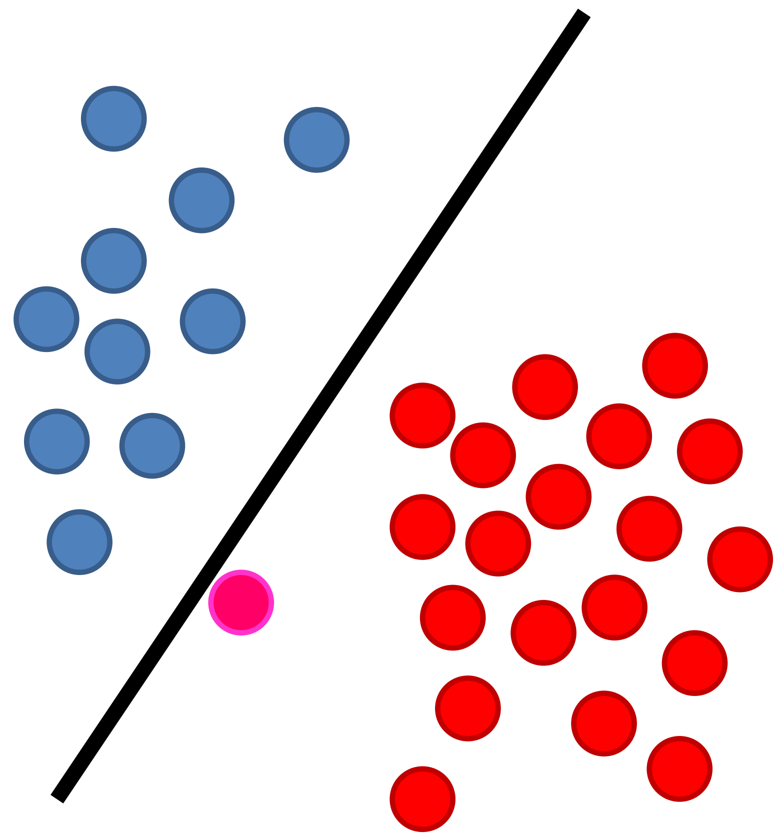
Data discretization

- Part of data reduction, replacing numerical attributes with nominal ones.
- Nominal to numerical?
- Binarization?

Classification Algorithms

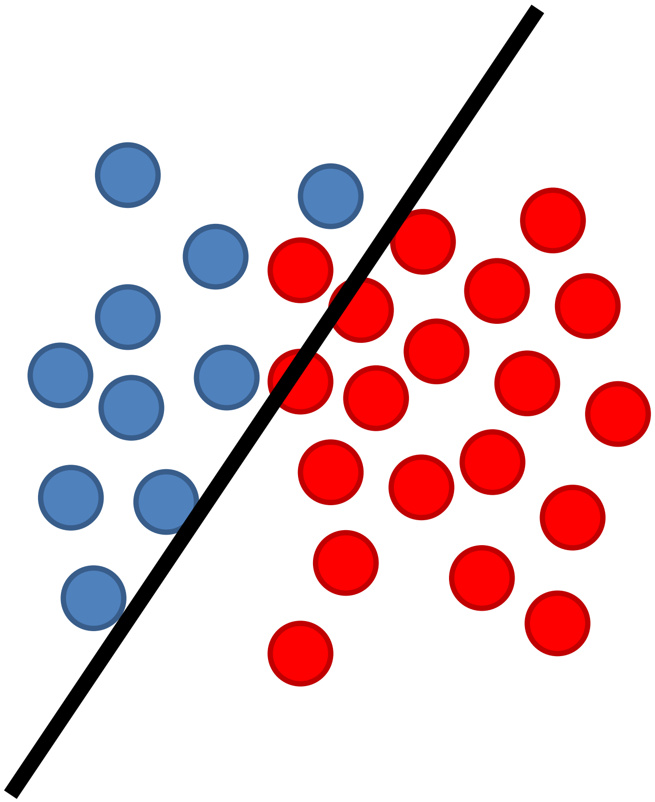
Classification Algorithms

- Linear classifiers
  - Fisher's linear discriminant
  - Logistic regression
  - Naive Bayes classifier
  - Perceptron



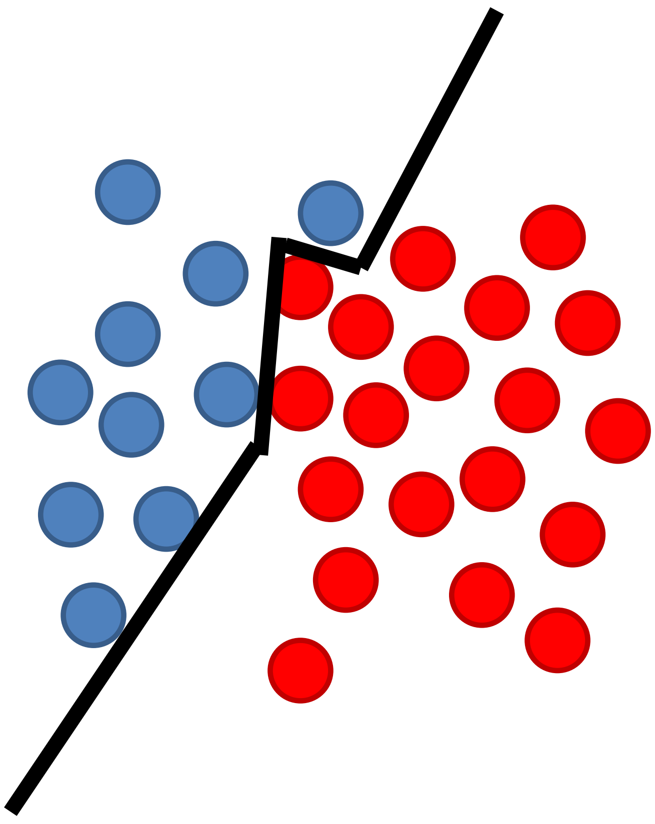
Classification Algorithms

- Linear classifiers
  - Fisher's linear discriminant
  - Logistic regression
  - Naive Bayes classifier
  - Perceptron



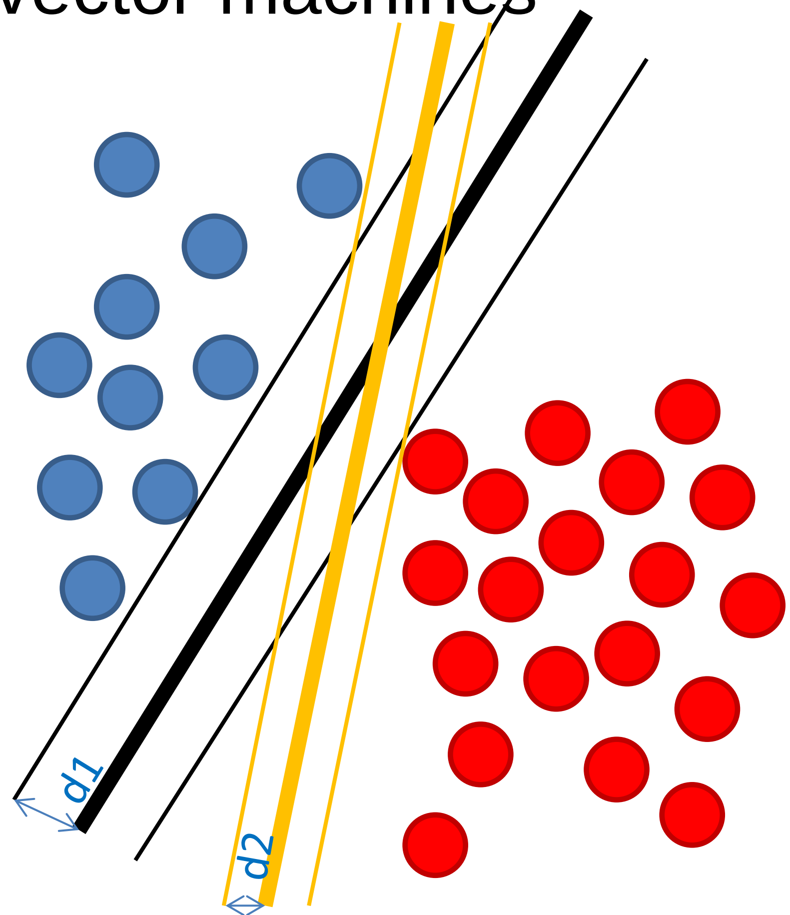
Classification Algorithms

- Linear classifiers
  - Fisher's linear discriminant
  - Logistic regression
  - Naive Bayes classifier
  - Perceptron



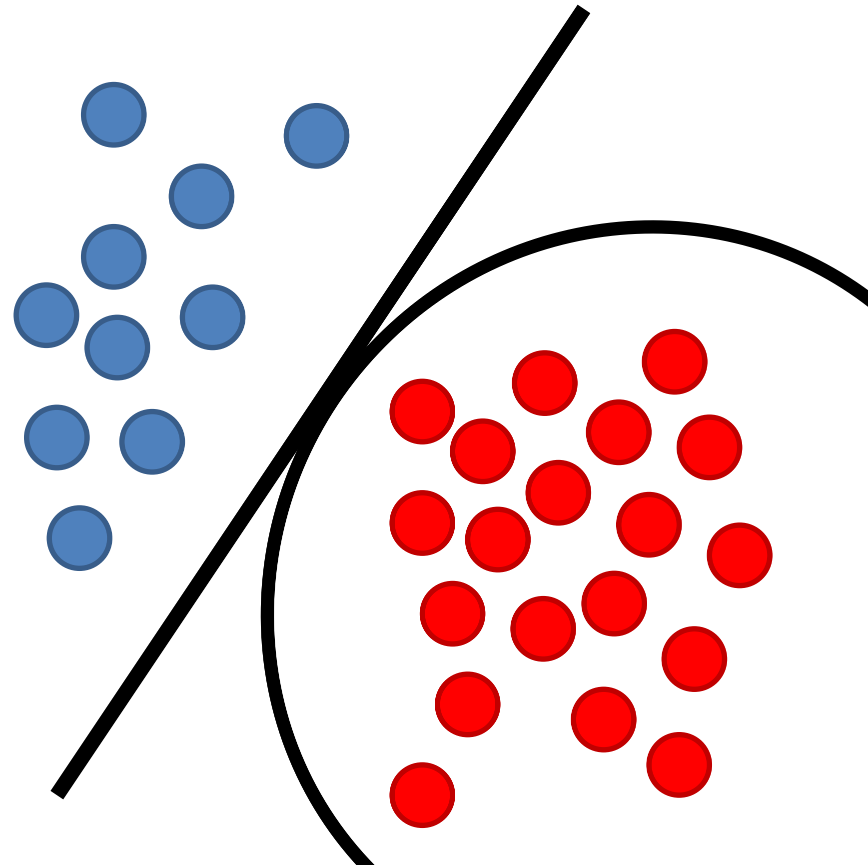
Classification Algorithms

- Support vector machines
  - Least squares support vector machines



Classification Algorithms

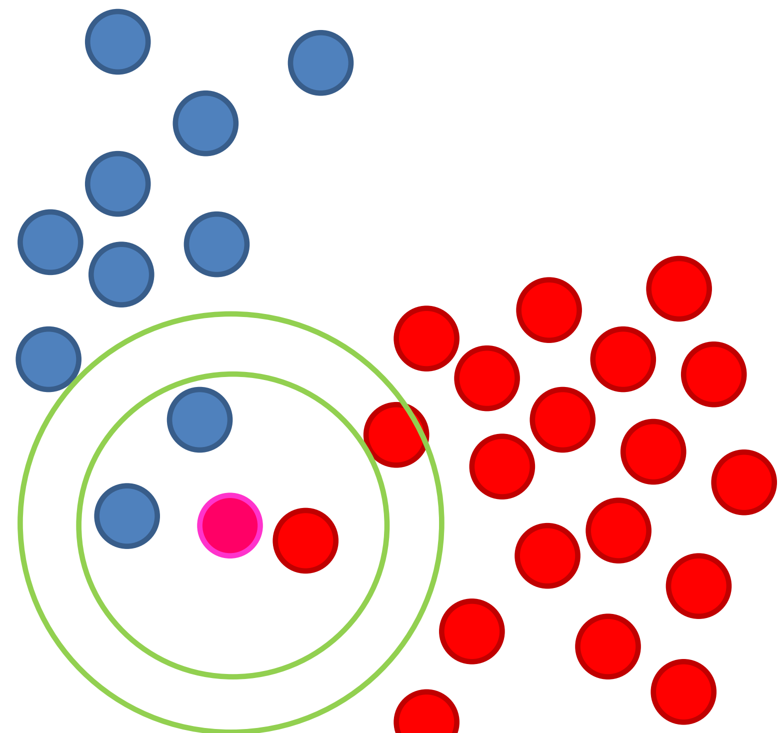
- Quadratic classifiers



Classification Algorithms

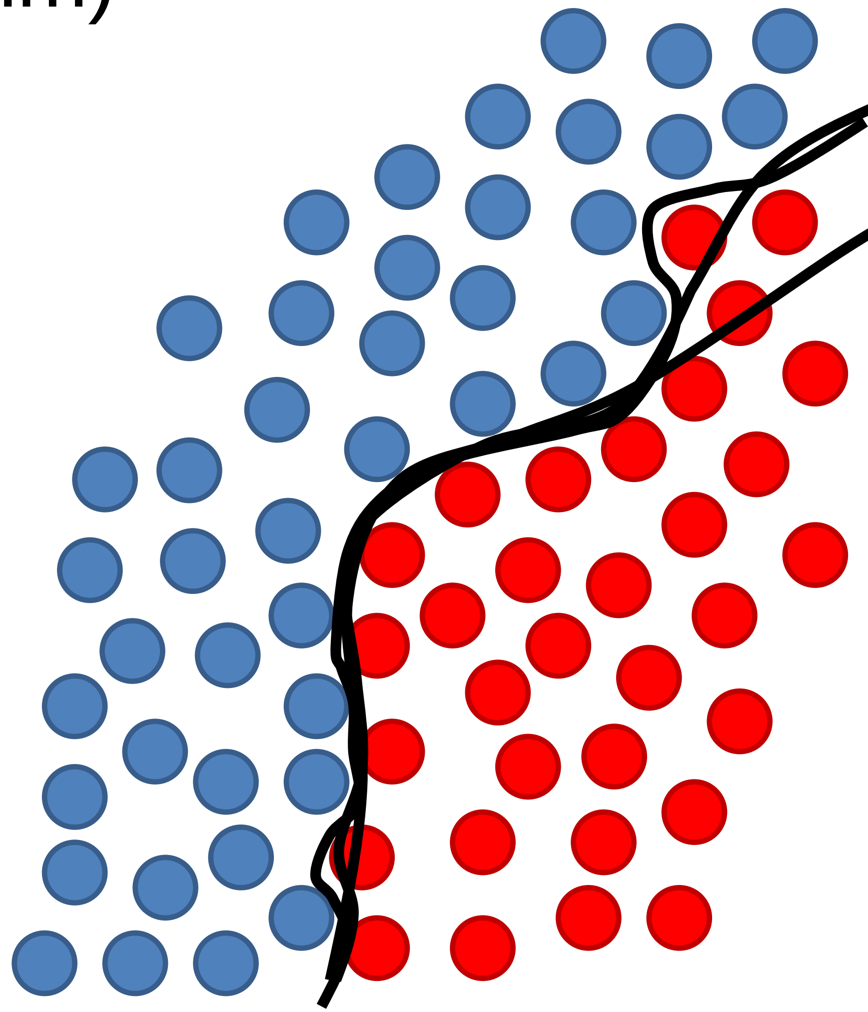
- Kernel estimation
  - k-nearest neighbor

K = 3



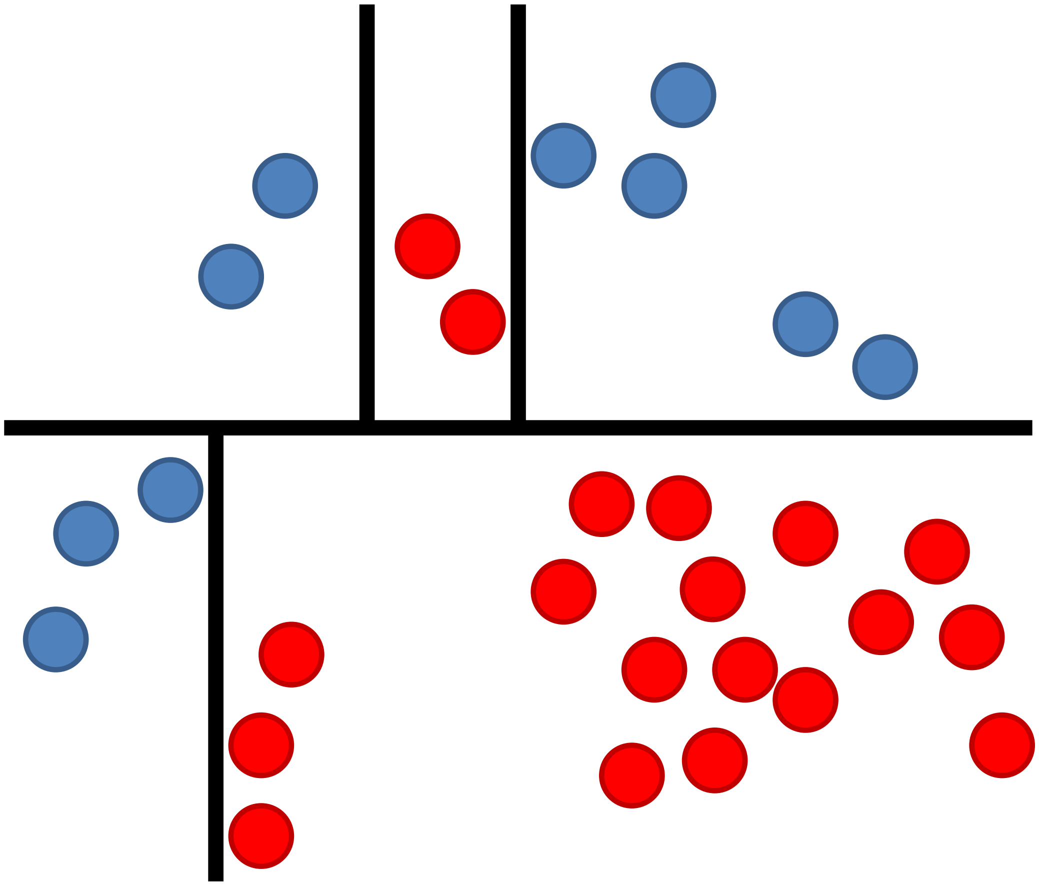
Classification Algorithms

- Boosting (meta-algorithm)
  - AdaBoost



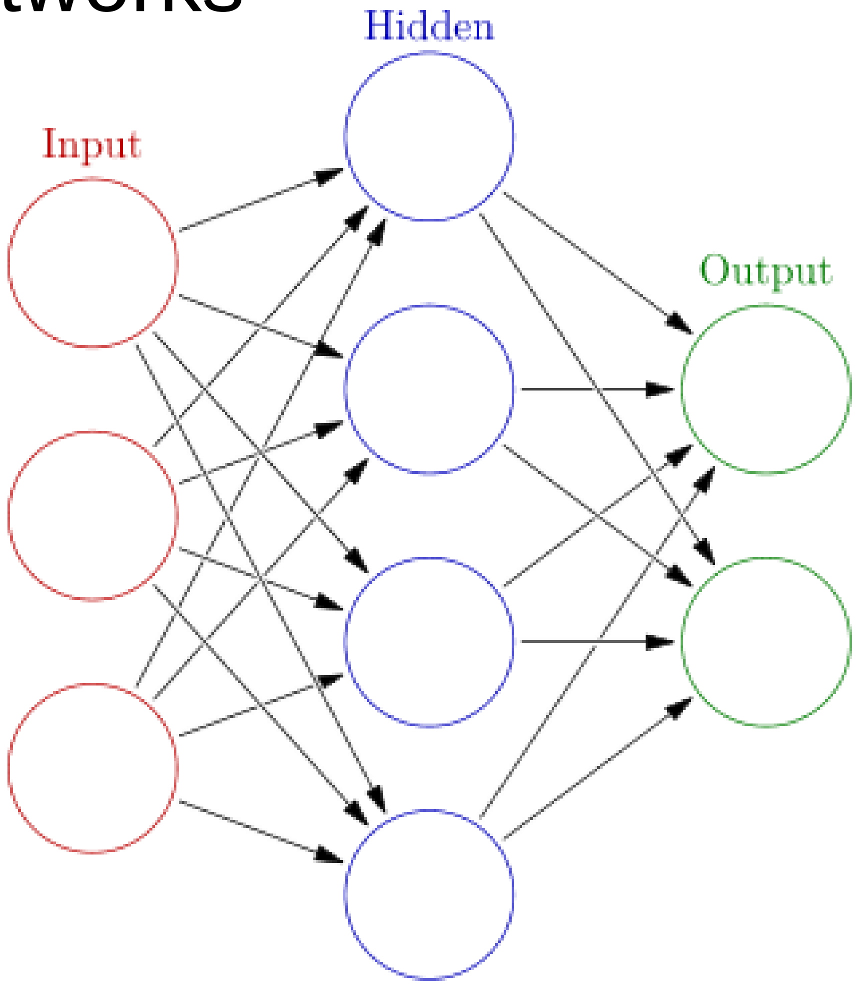
Classification Algorithms

- Decision trees
  - Random forests



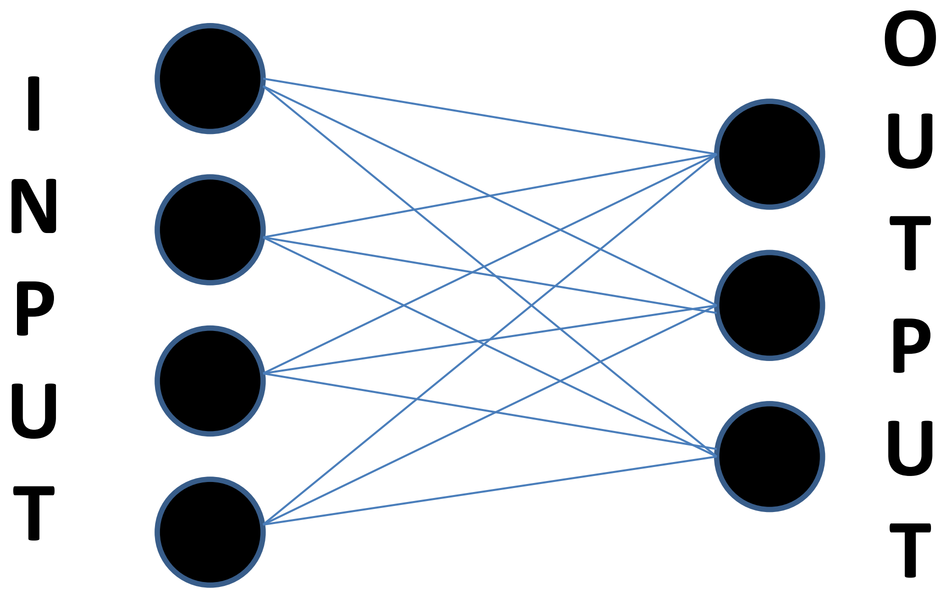
Classification Algorithms

- Neural networks



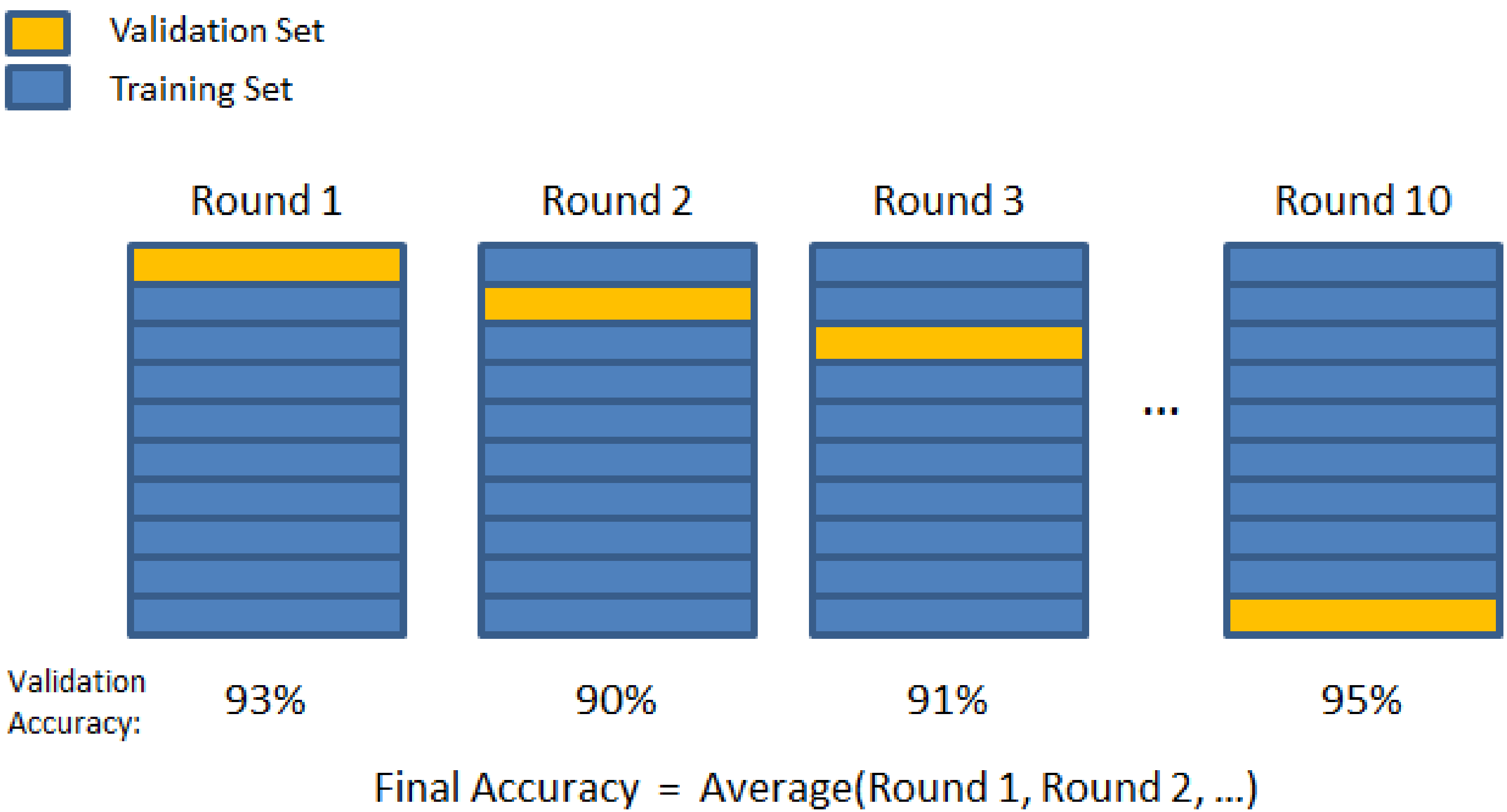
Classification Algorithms

- Learning vector quantization



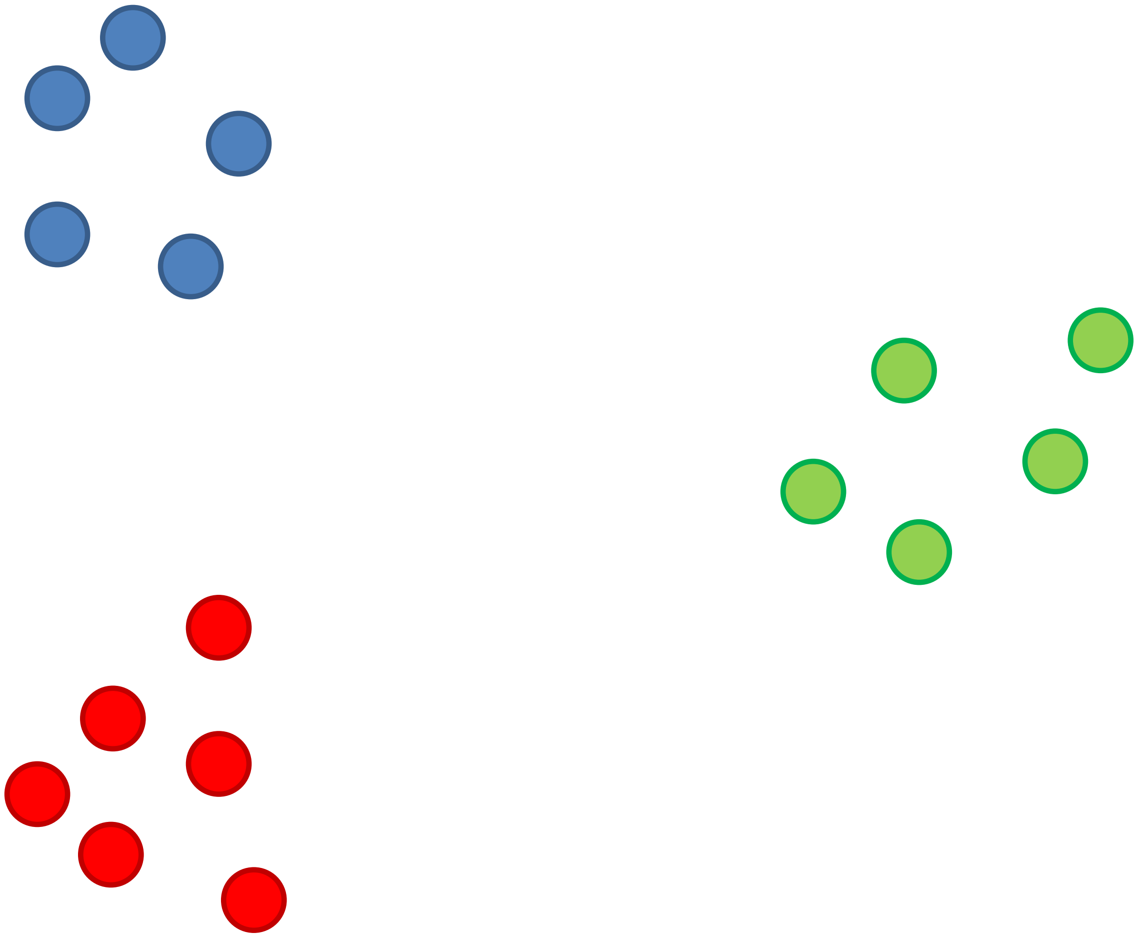
„winner-take-all“

Cross-validation

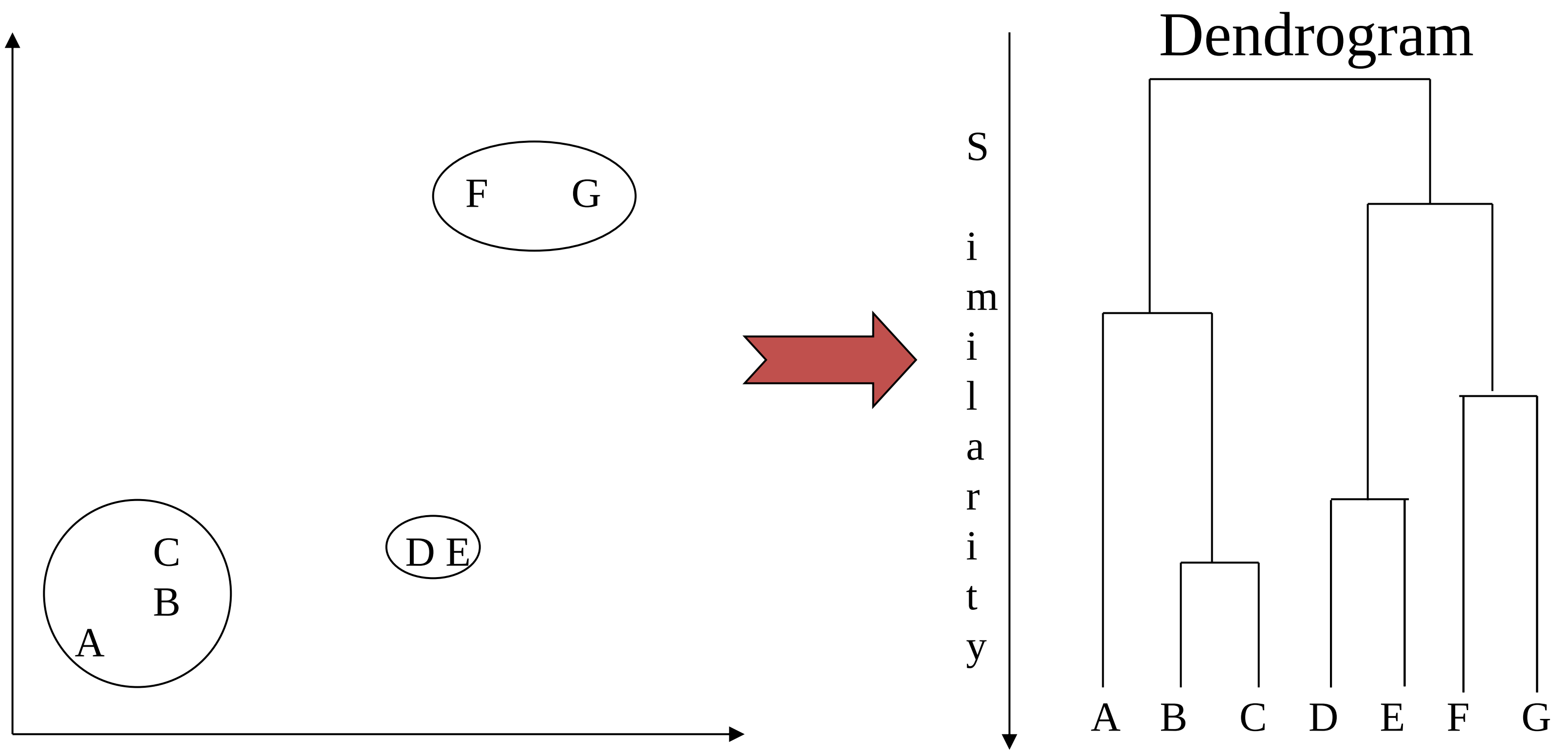


Clustering

Clustering Example



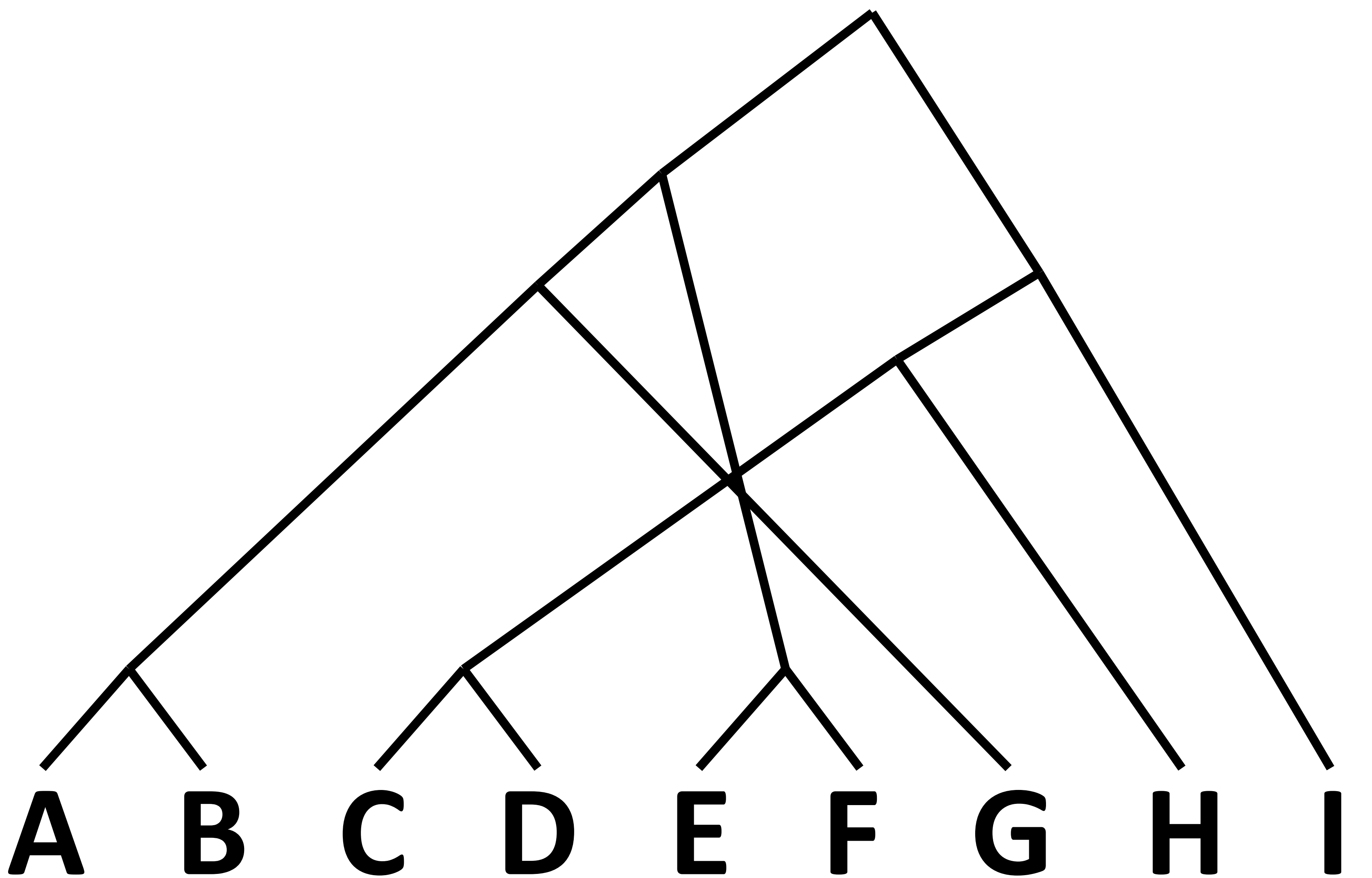
Hierarchical Clustering



Aglommerative vs. Divisive Clustering

- ✓ ***Aglommerative*** **(*bottom-up*)**
- ✓ ***Divisive*** **(*partitional, top-down*)**

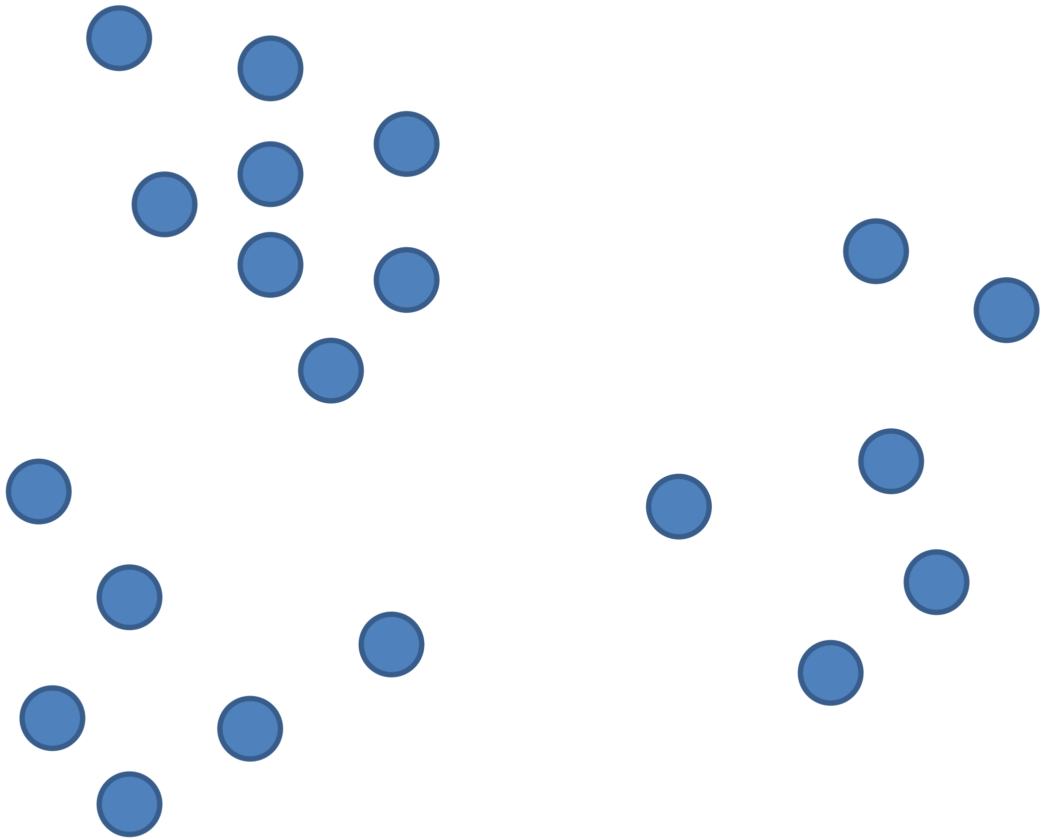
Agglomerative Clustering



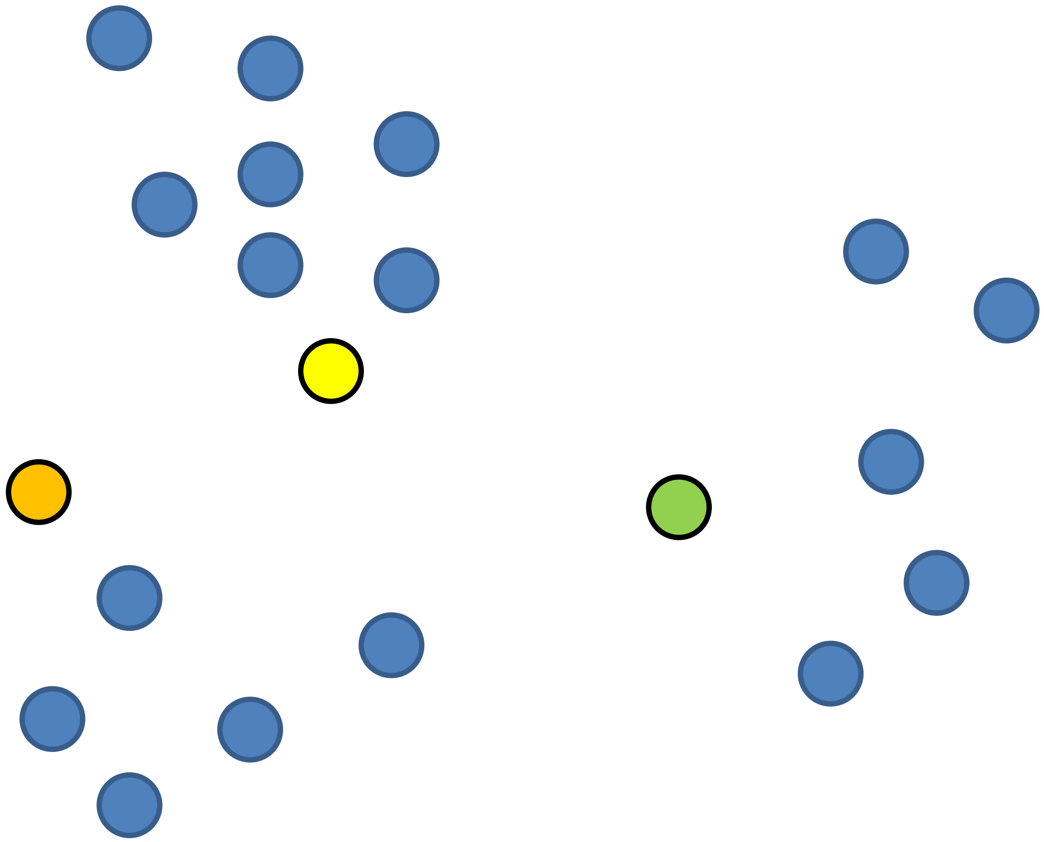
Cluster Similarity

- **Single Link**: Similarity of the two most similar members.
- **Complete Link**: Similarity of the two least similar members.
- **Group Average**: Average similarity between members.

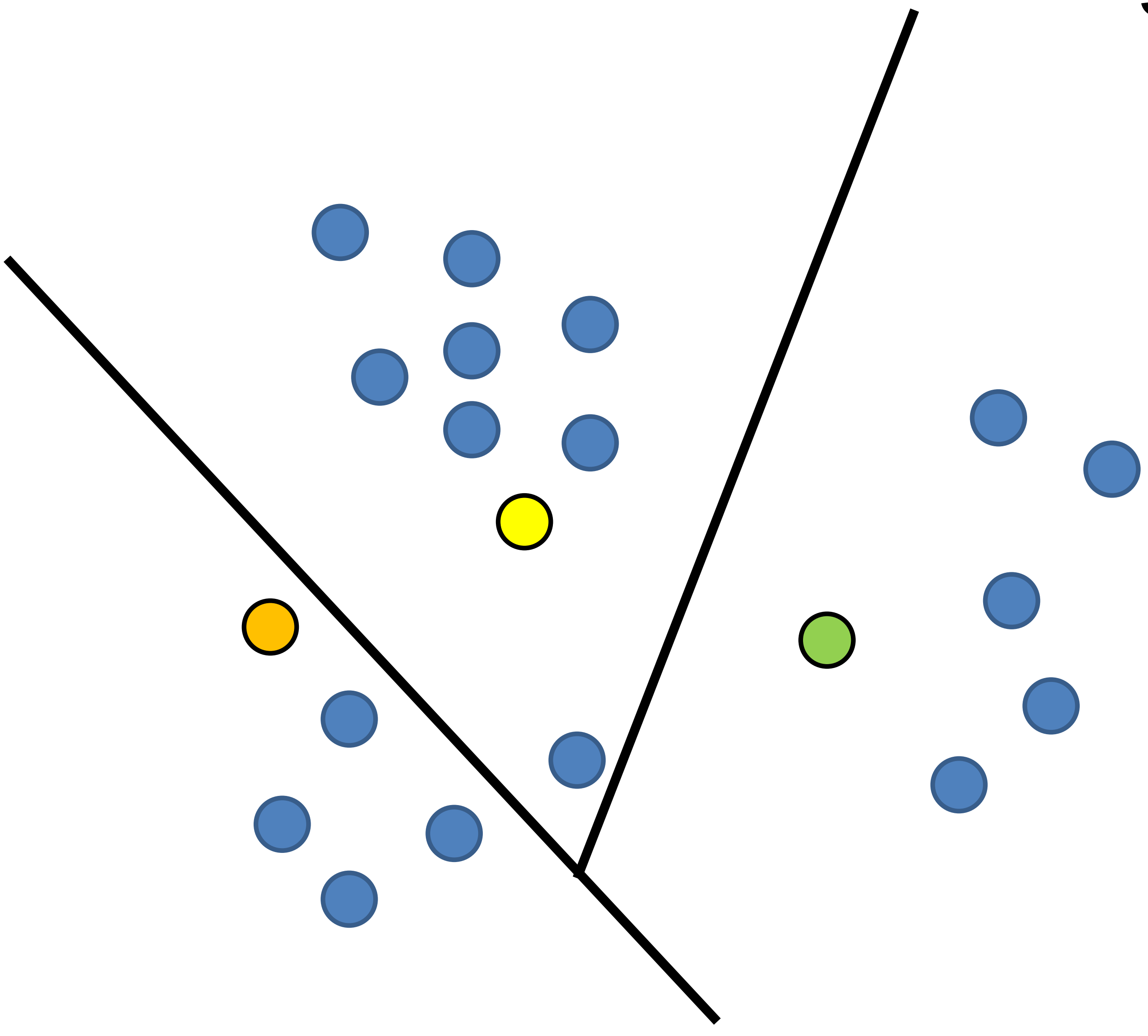
K-Means Clustering



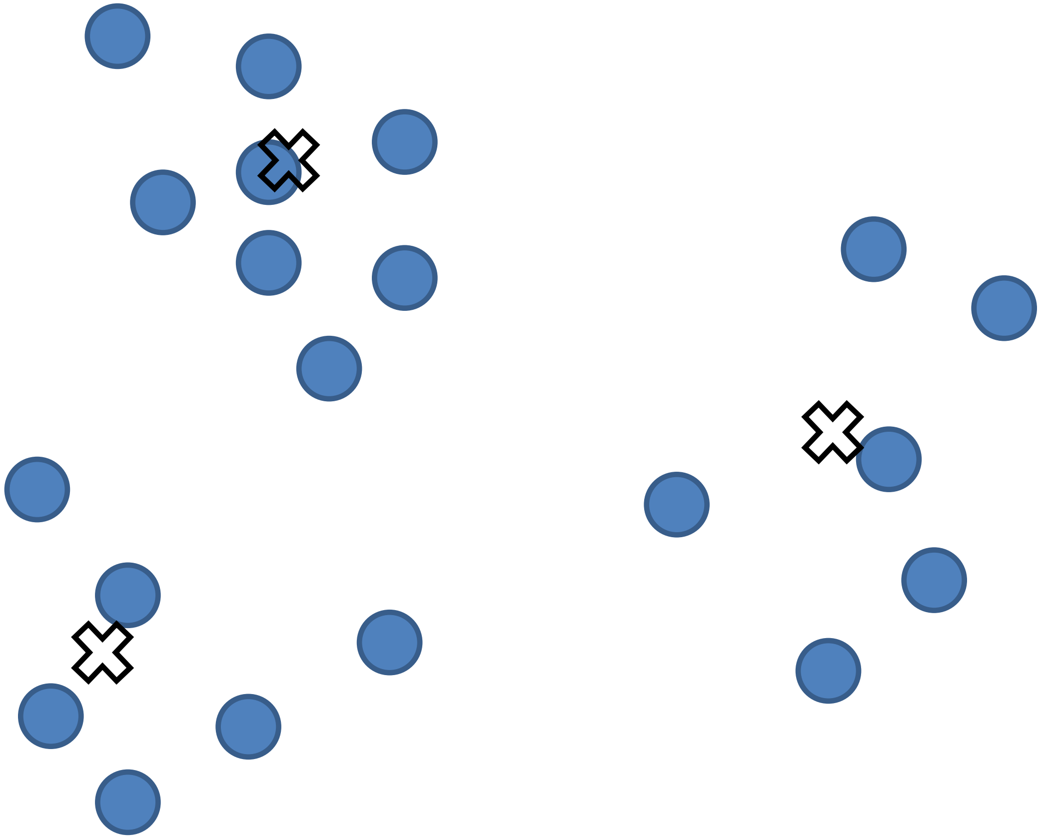
K-Means Clustering



K-Means Clustering



K-Means Clustering



K-Means Clustering