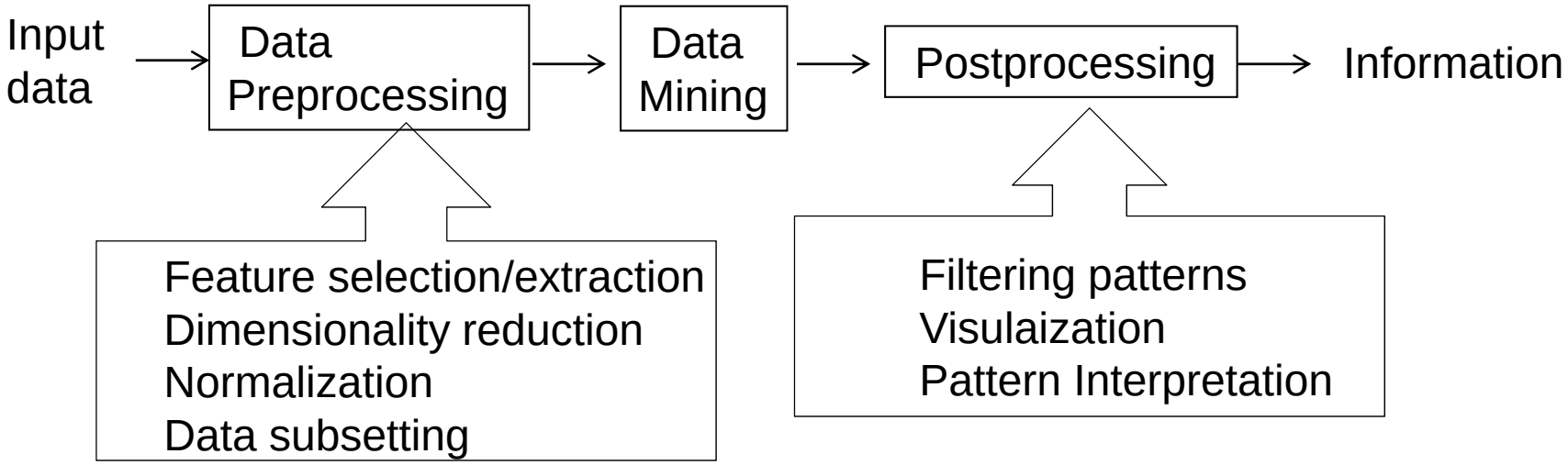
Data Mining
Machine Learning
Pattern Recognition

데이타마이닝 연구실
박정희

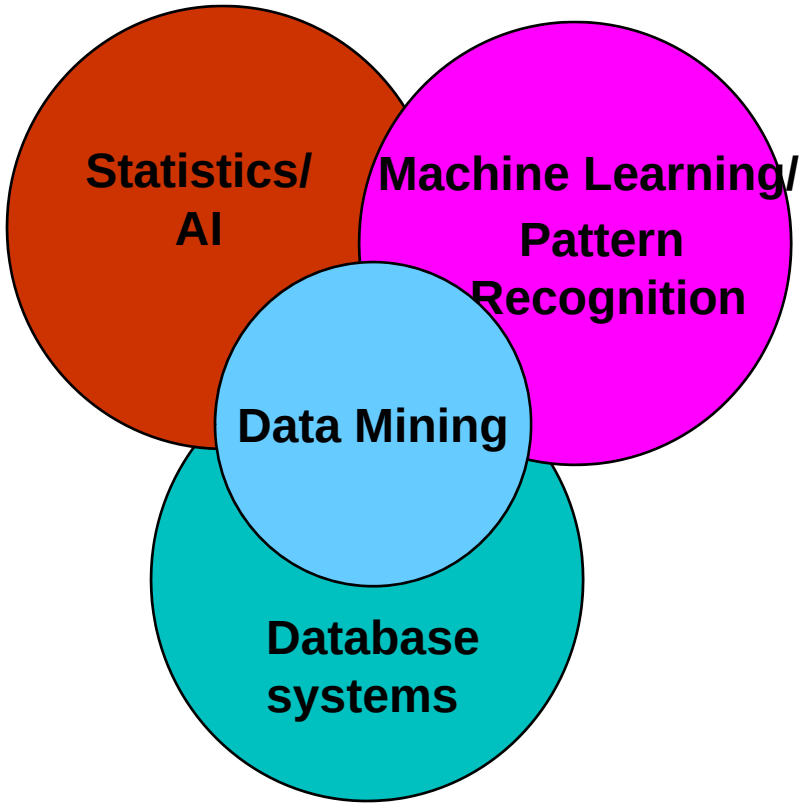
Knowledge Discovery process

- Lots of data is being collected and stored at enormous speeds
- Computers have become cheaper and more powerful
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data



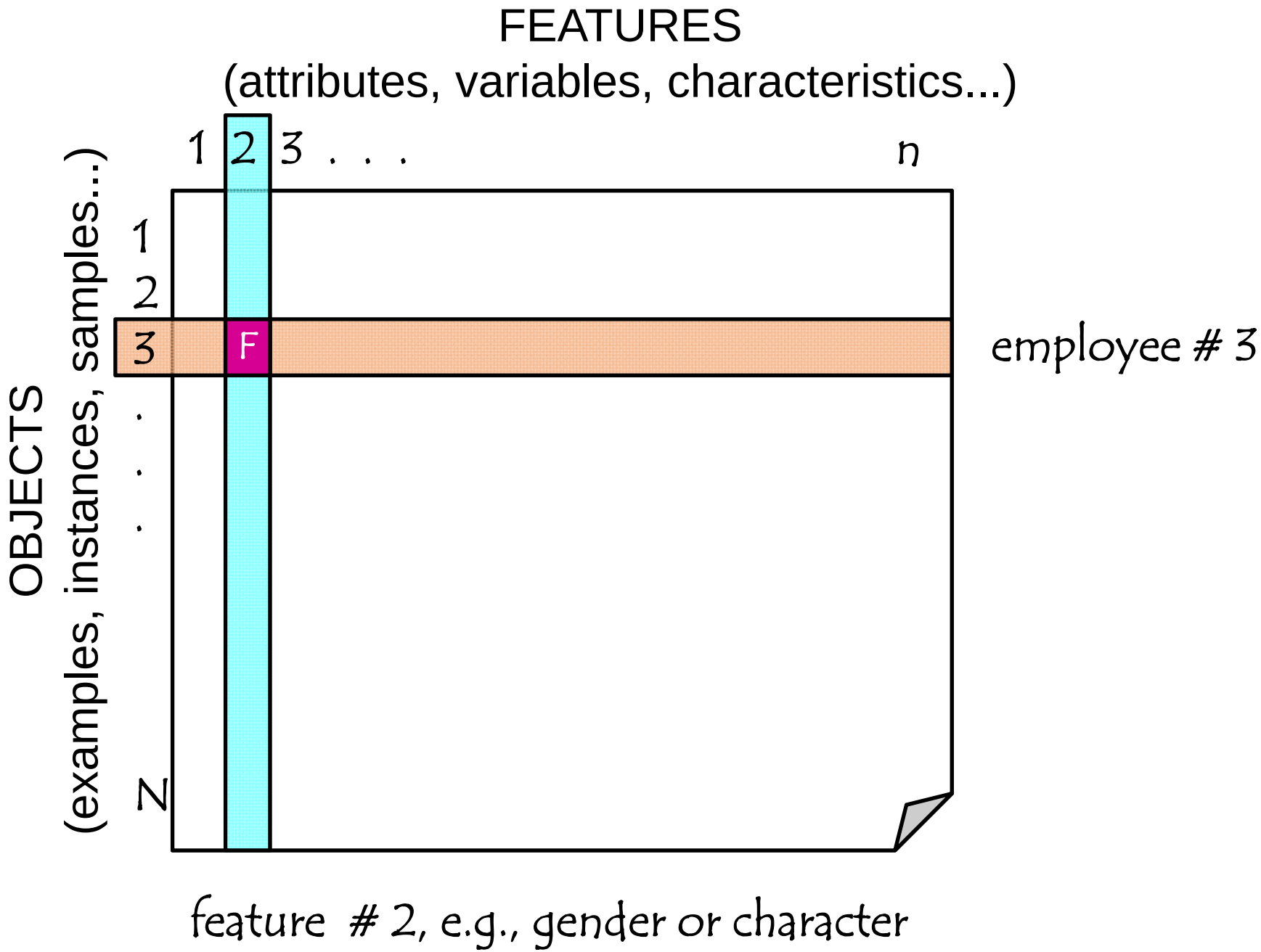
Origins of Data Mining

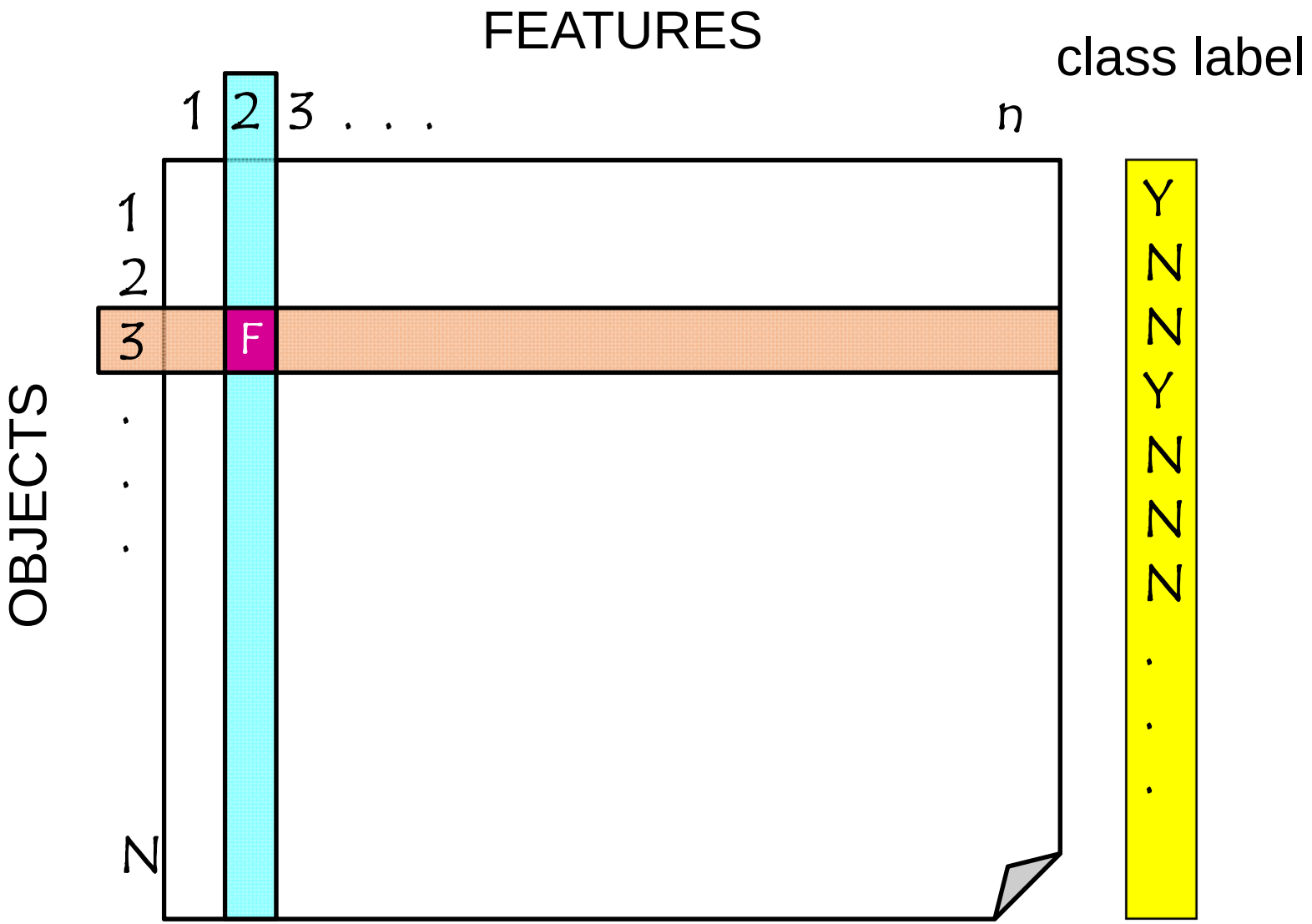
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

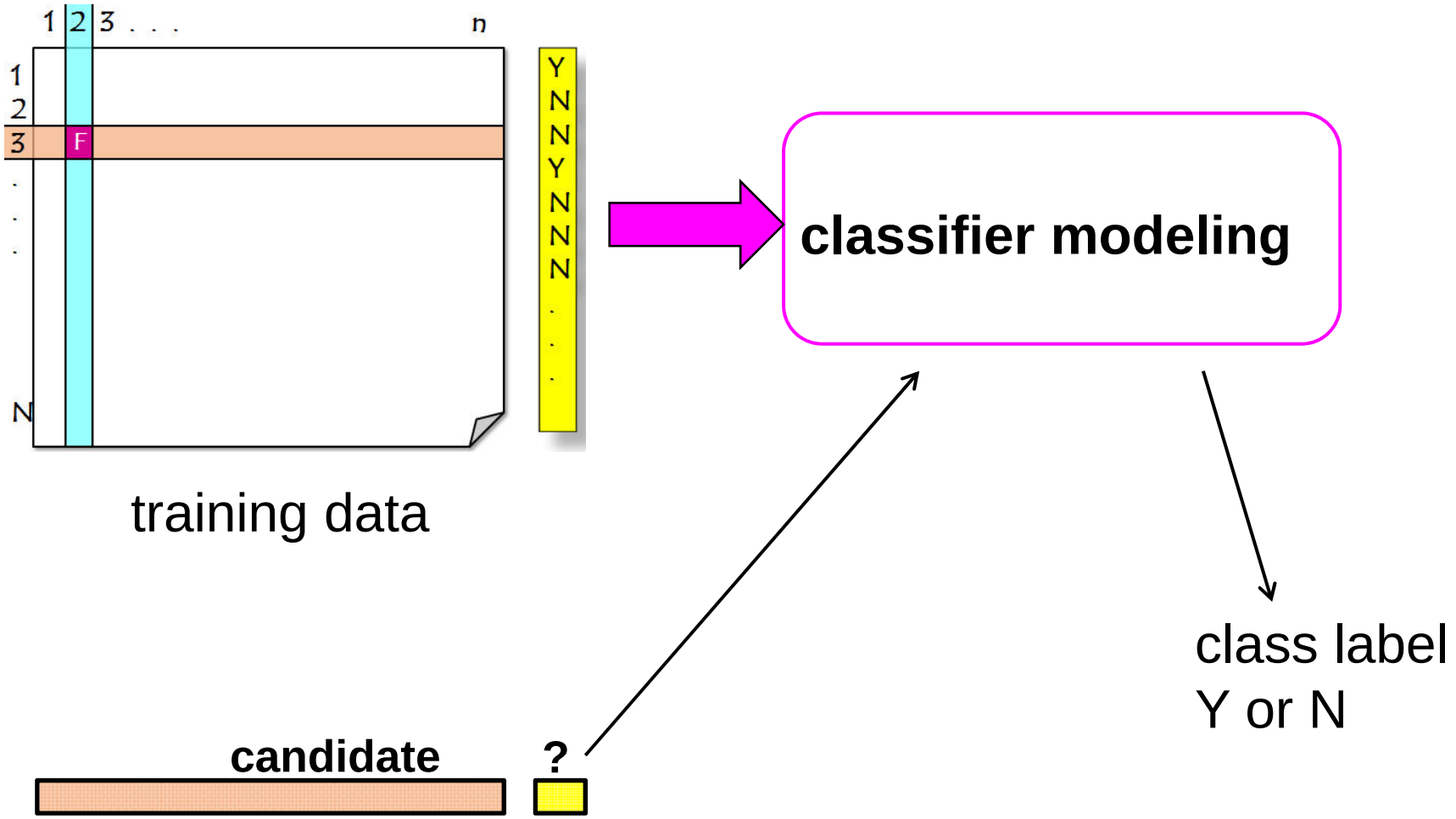


Example: Hiring new employees

- Problem: Who would leave within a year?
- Data collection
  - All employee information
- Classifier modeling
- Perform prediction for a new data sample


FEATURES
(attributes, variables, characteristics...)
1 2 3 . . . n
OBJECTS
(examples, instances, samples...)
1
2
3
. . .
N
F
employee # 3
feature # 2, e.g., gender or character


FEATURES
class label
1 2 3 . . . n
OBJECTS
1
2
3
F
. . .
N
Y N N Y N N N . . .


1 2 3 . . . n
1
2
3
F
N
Y
N
N
Y
N
N
N
N
training data
classifier modeling
class label
Y or N
candidate
?

Major Tasks

- Classification
- Clustering
- Association Rule Discovery
- Regression
- Outlier Detection
- Sequential Pattern Discovery
- ...

Classification

- Given a collection of records,
  - Each record contains a set of *attributes*
  - one of the attributes is the *class*.

- Find a *model* for class attribute as a function of the values of other attributes.

- Goal: Previously unseen records should be assigned a class as accurately as possible.

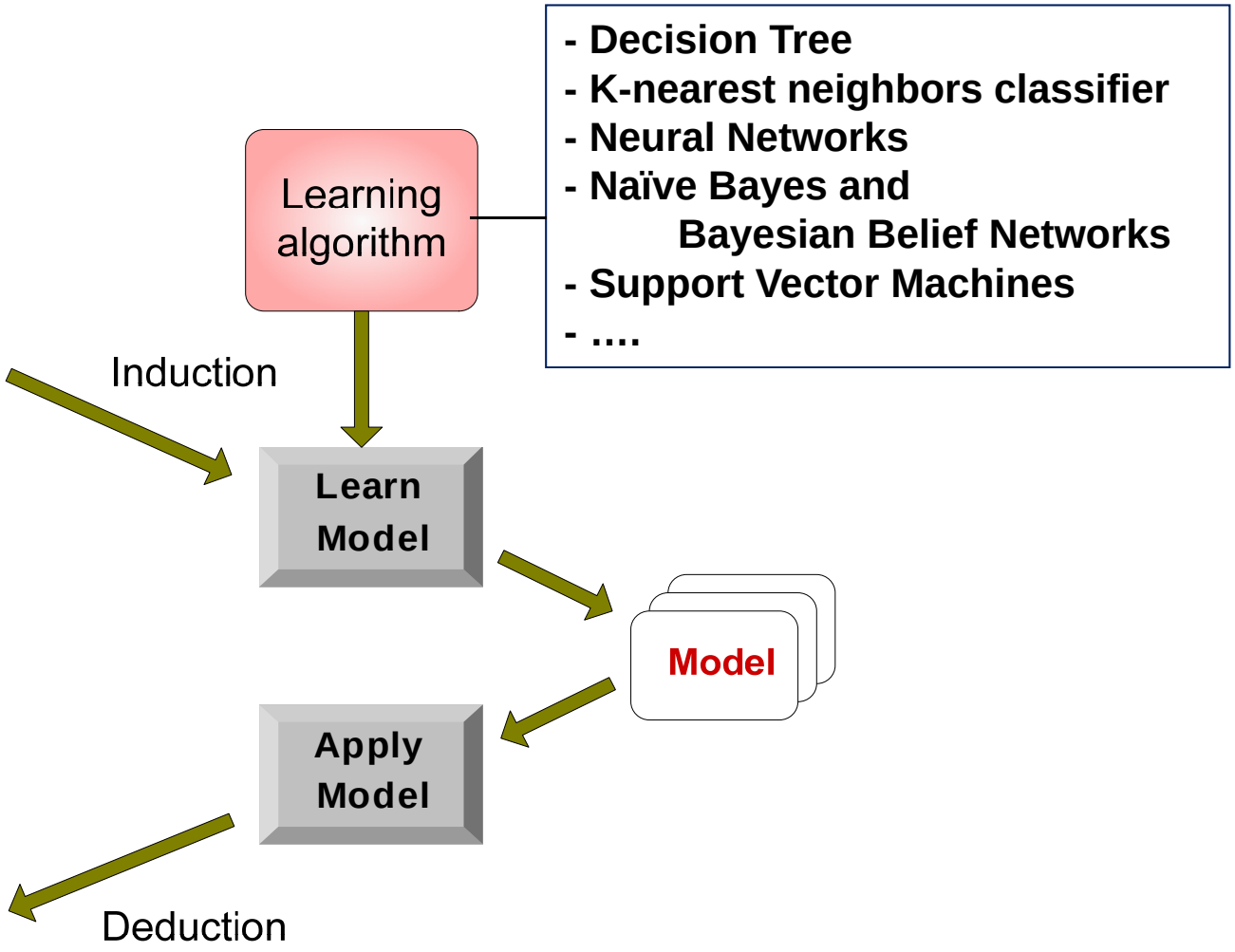
Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

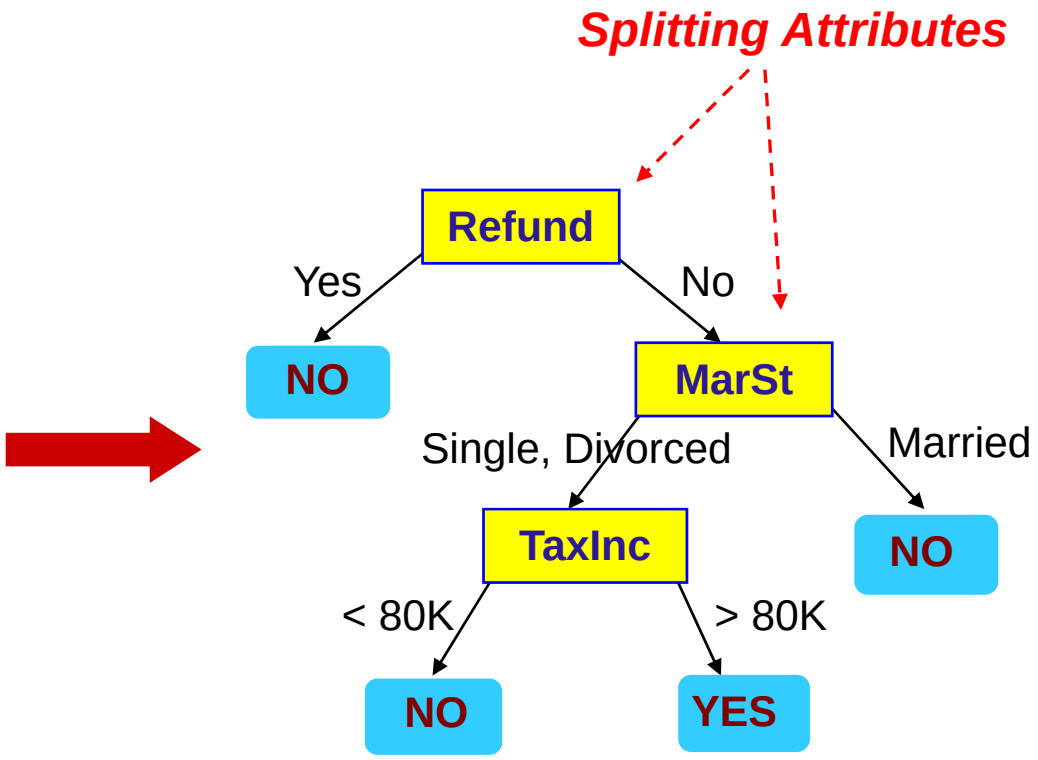
Test Set



Decision Tree

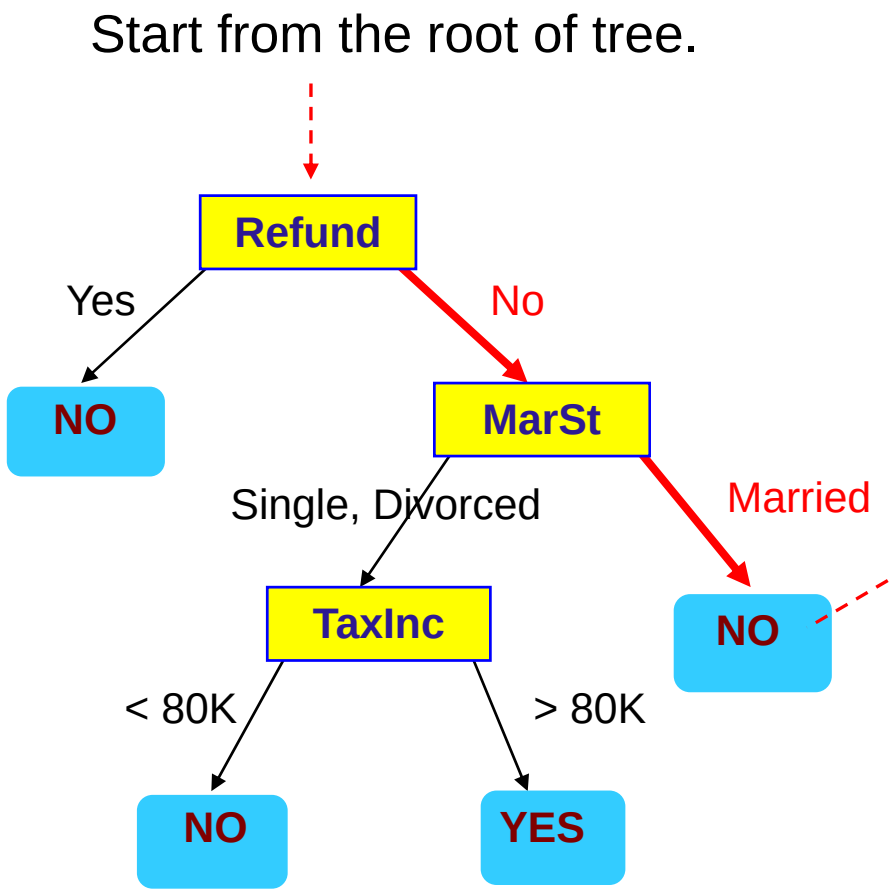
Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Apply Model to Test Data



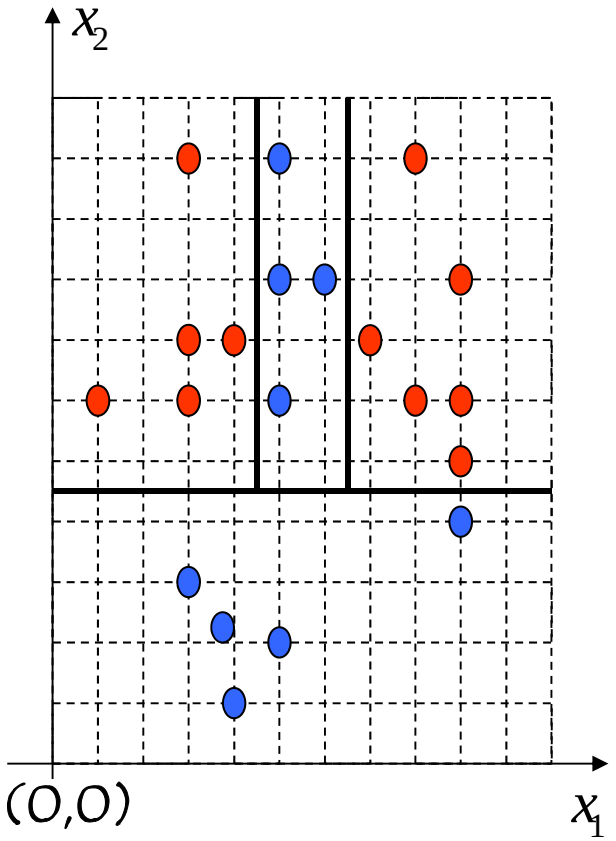
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Assign “No” to Cheat

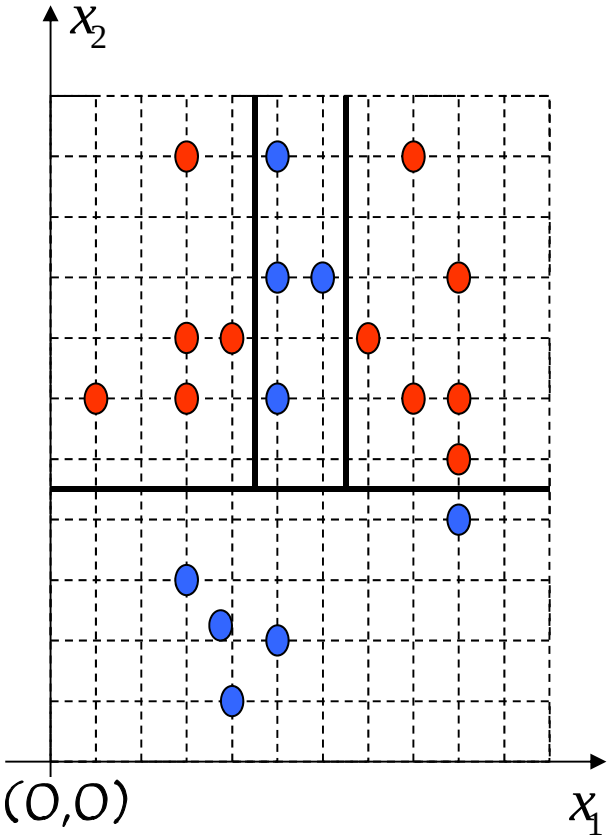
Construction of decision tree

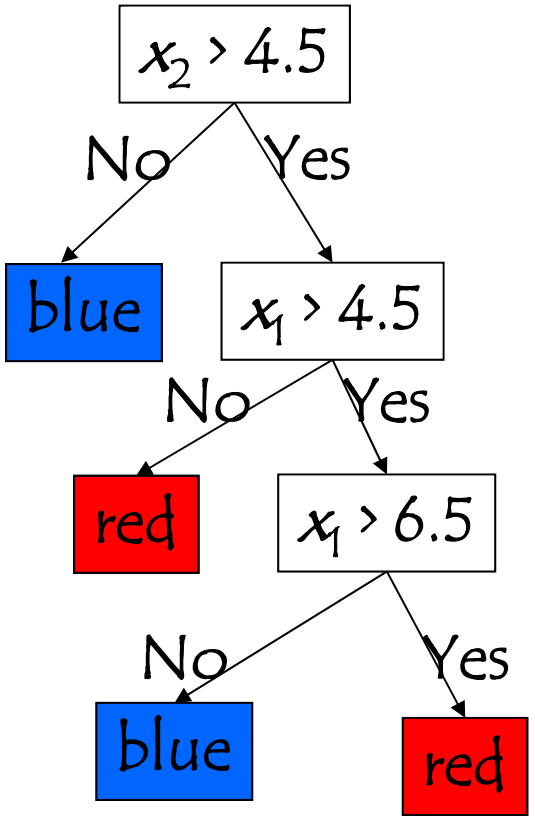
An example: Class 1 = red, Class 2 = blue.



Construction of decision tree

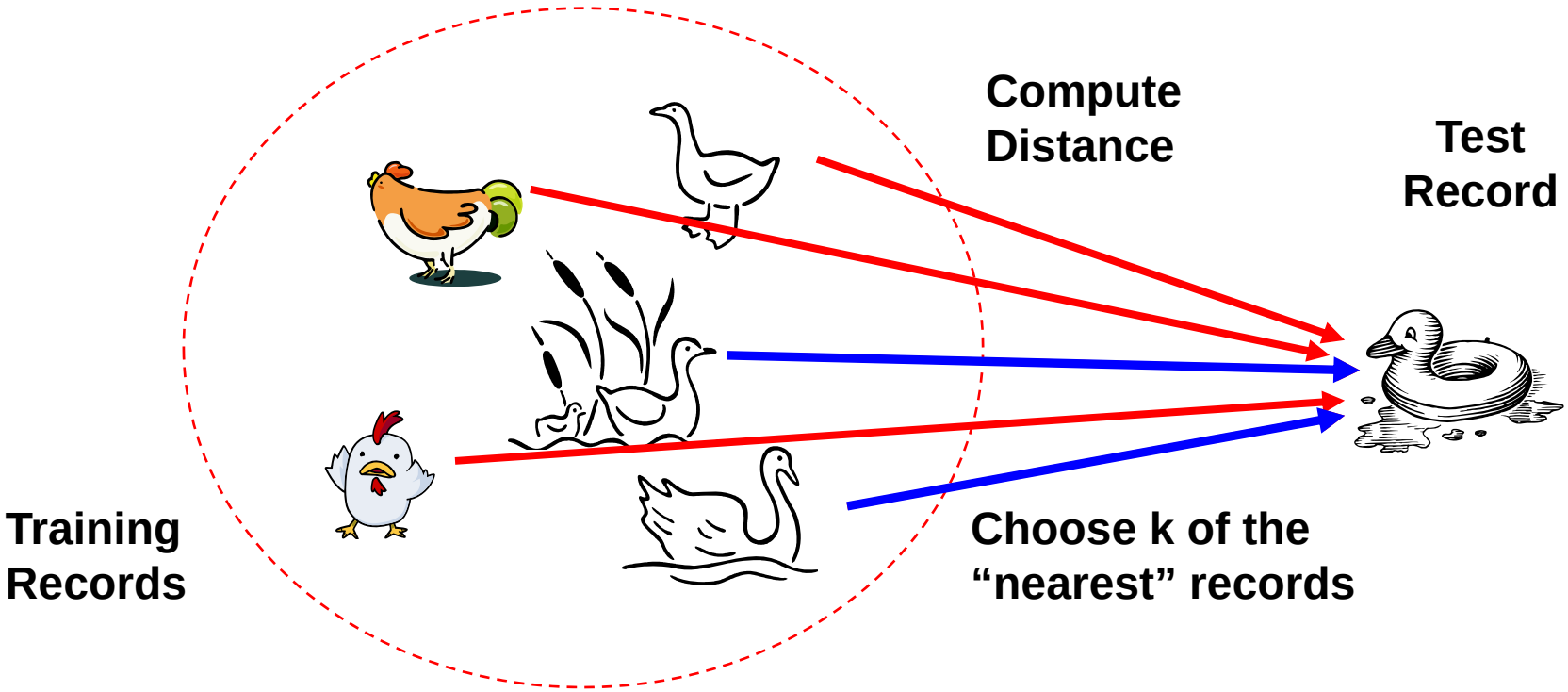
An example: Class 1 = red, Class 2 = blue.



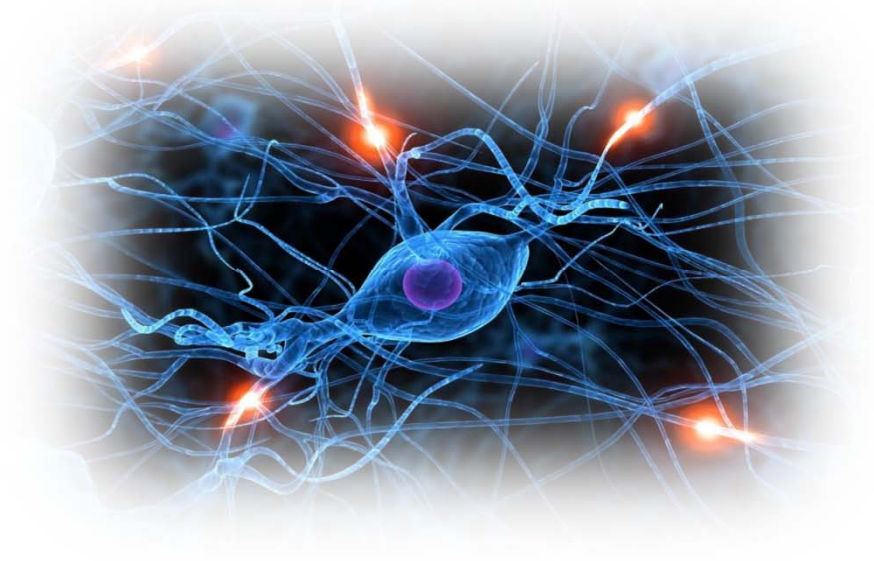


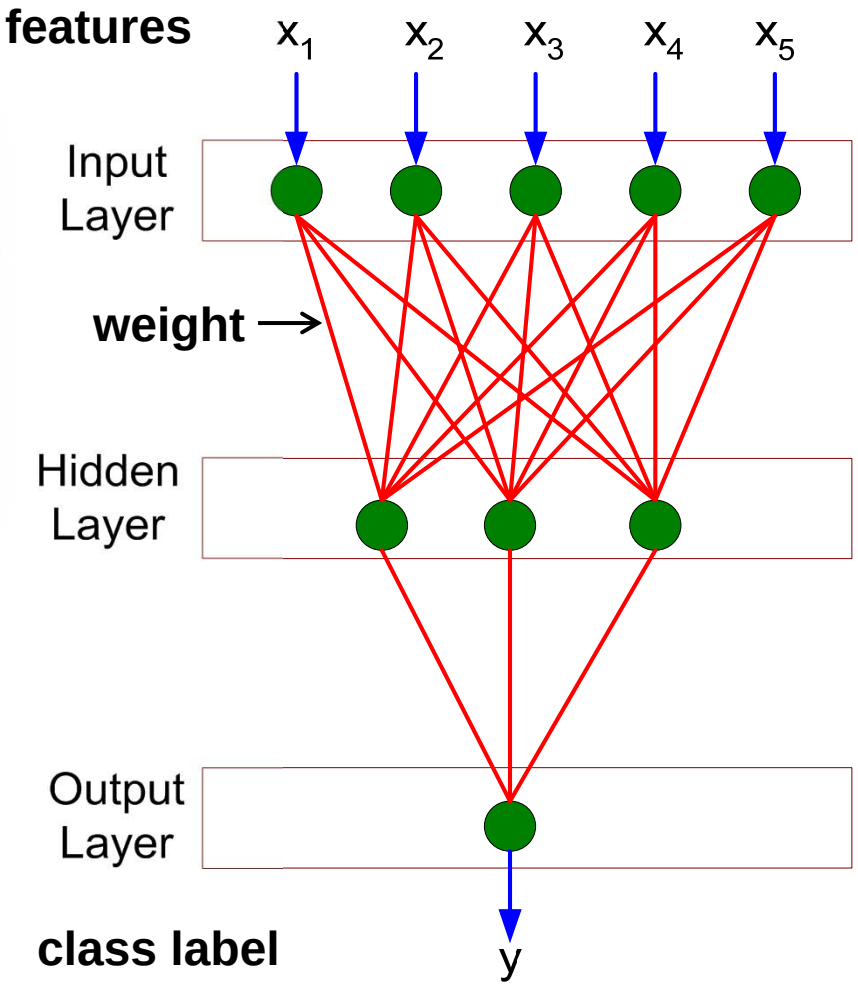
k-nearest neighbors classifier

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck

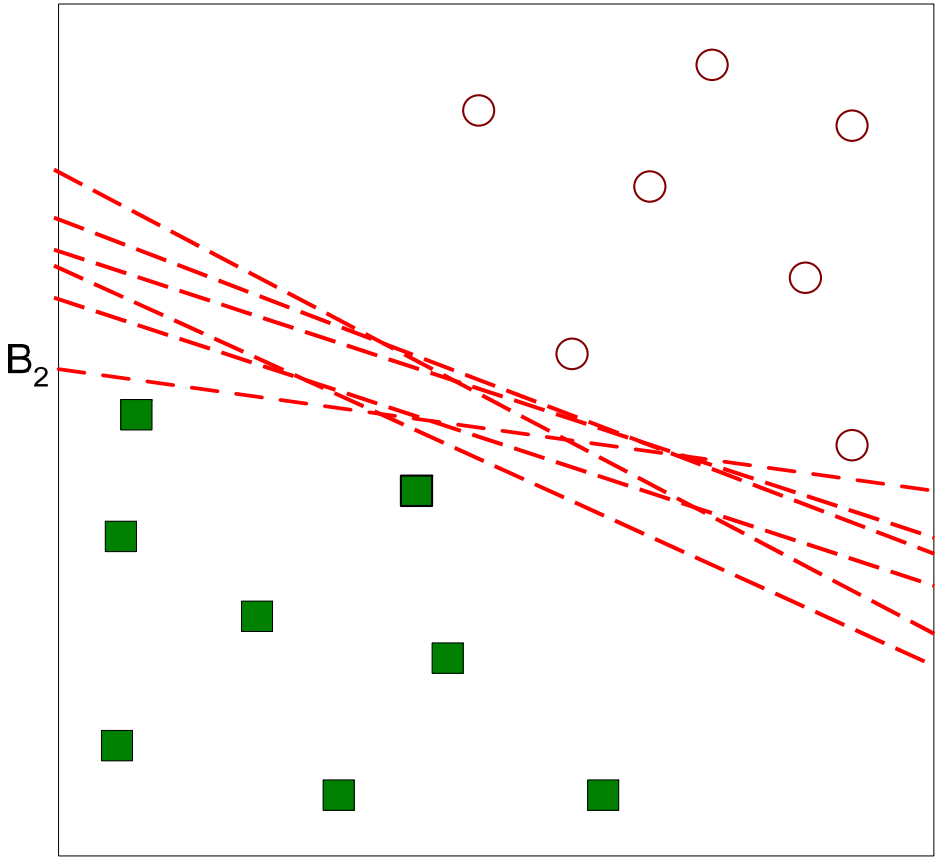


Artificial Neural Networks



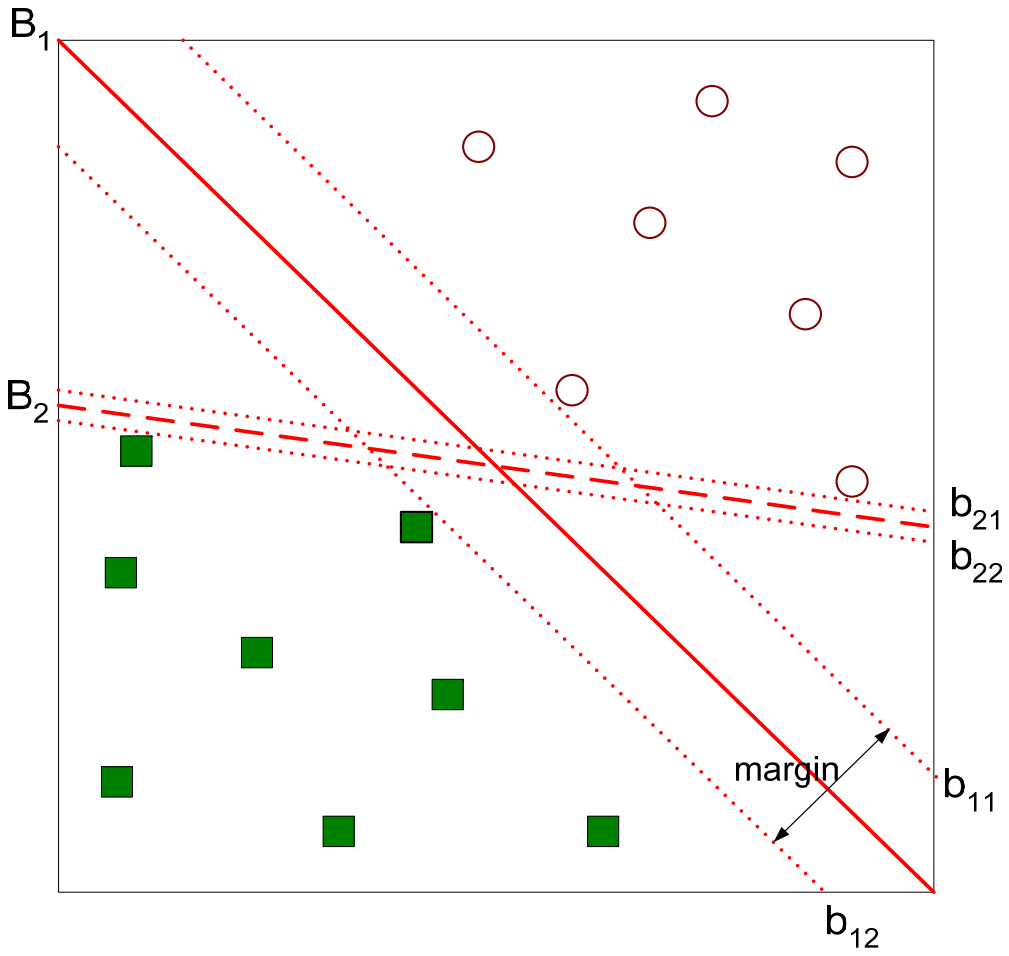


Support Vector Machines



- Other possible solutions

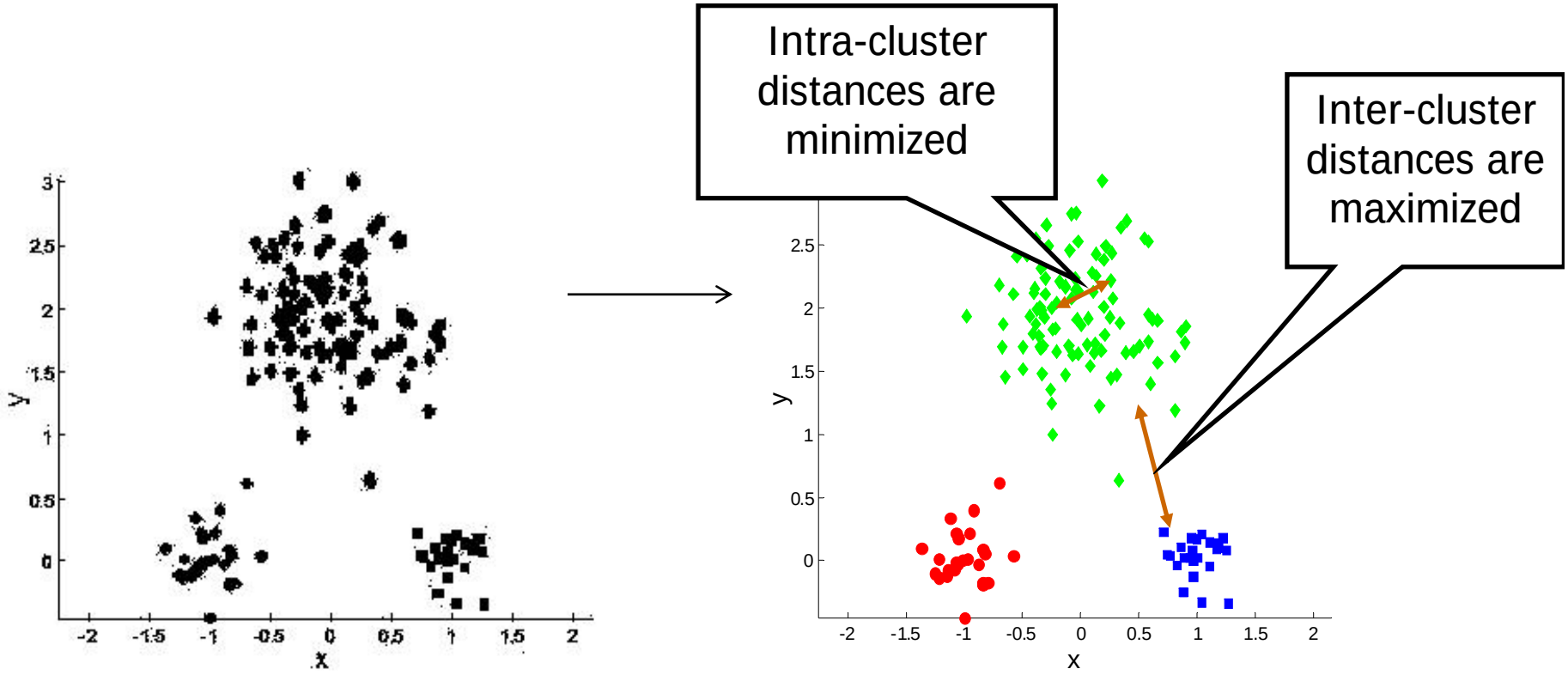
Support Vector Machines



- Find hyperplane maximizes the margin => B1 is better than B2

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

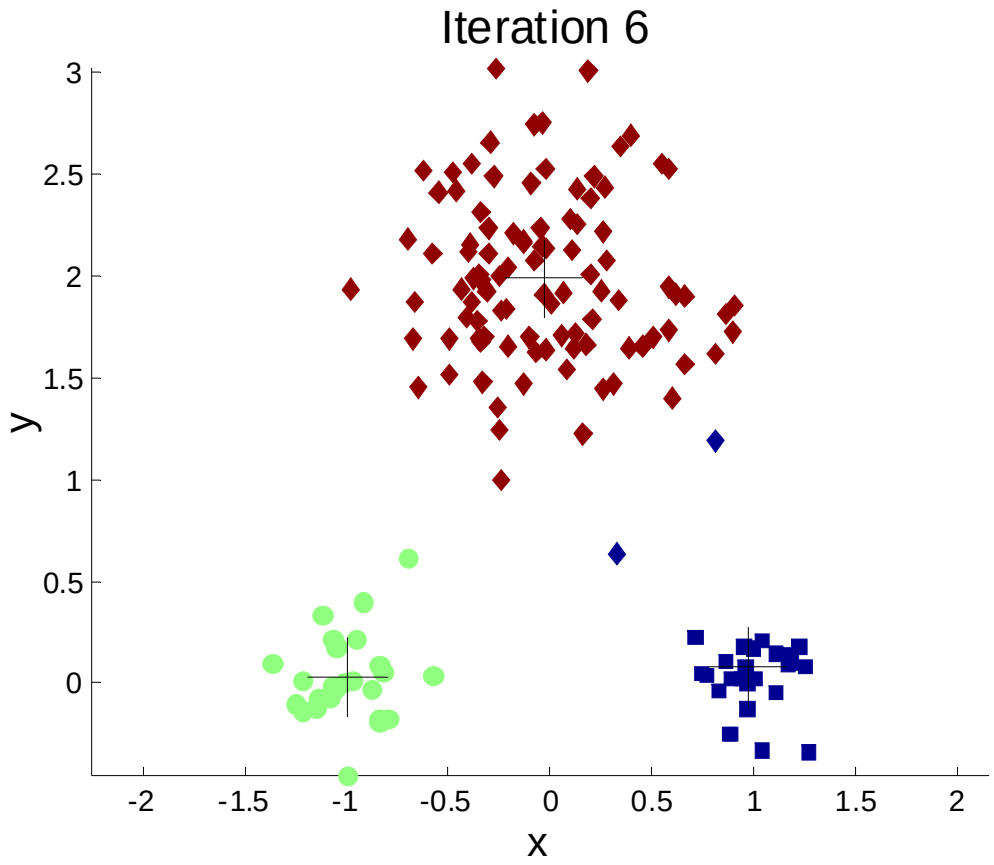


K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (or center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Example of K-means clustering



Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Association Rule

An implication expression of the form X $\rightarrow$ Y, where X and Y are itemsets

Example of Association Rules

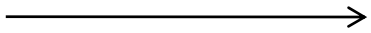
{Diaper} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs,Coke},
{Beer, Bread} $\rightarrow$ {Milk},

Dimension reduction

features

A1	A2	A3	…	…	…	C
10						Y
21						N
13						Y
14						N
15						Y
10						Y
:						
:						
:						

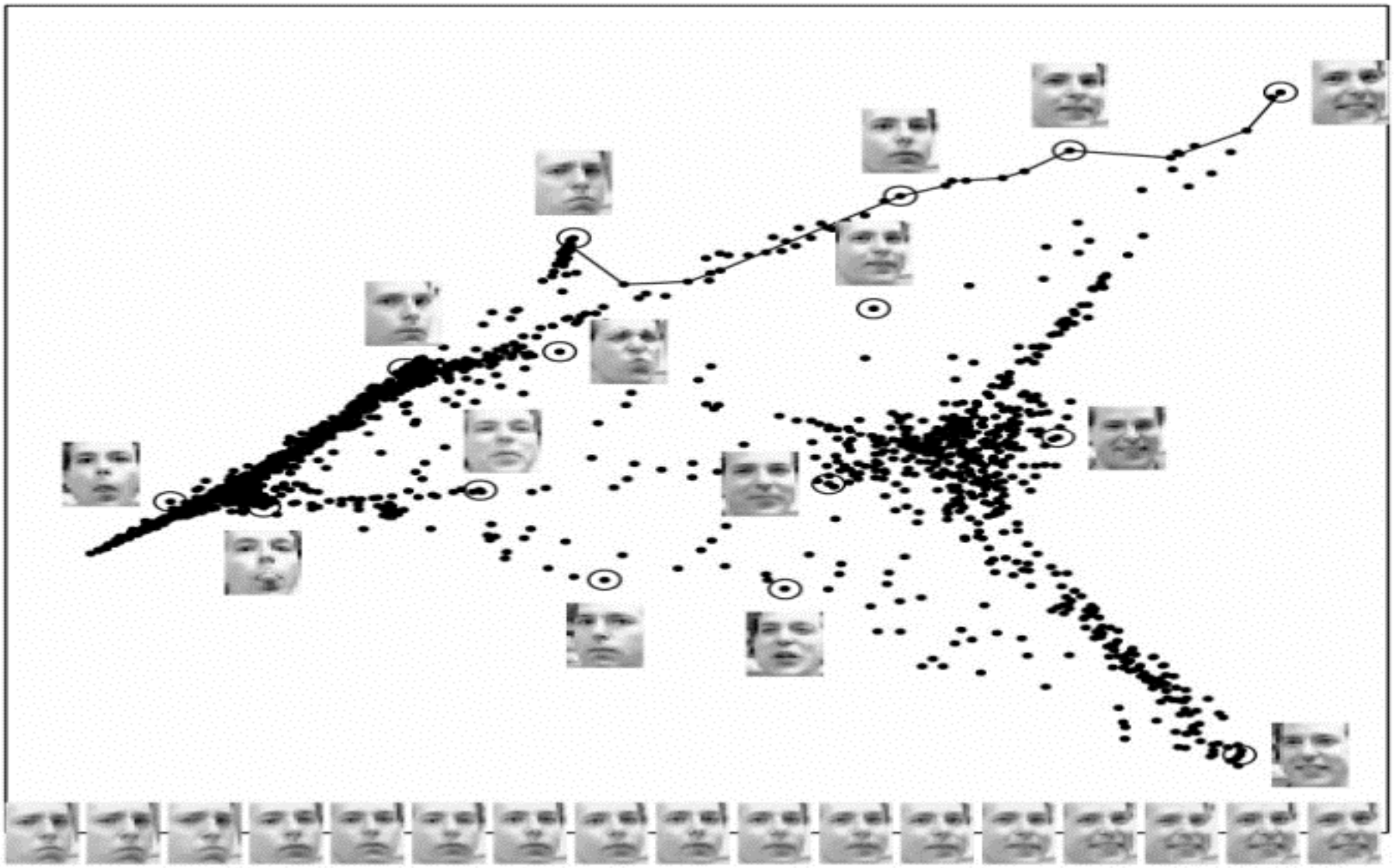
dimenion reduction
(차원감소)



B1	B2	…	C
:			Y
:			N
:			Y
:			N
:			Y
:			Y

- **dimension: the number of features**

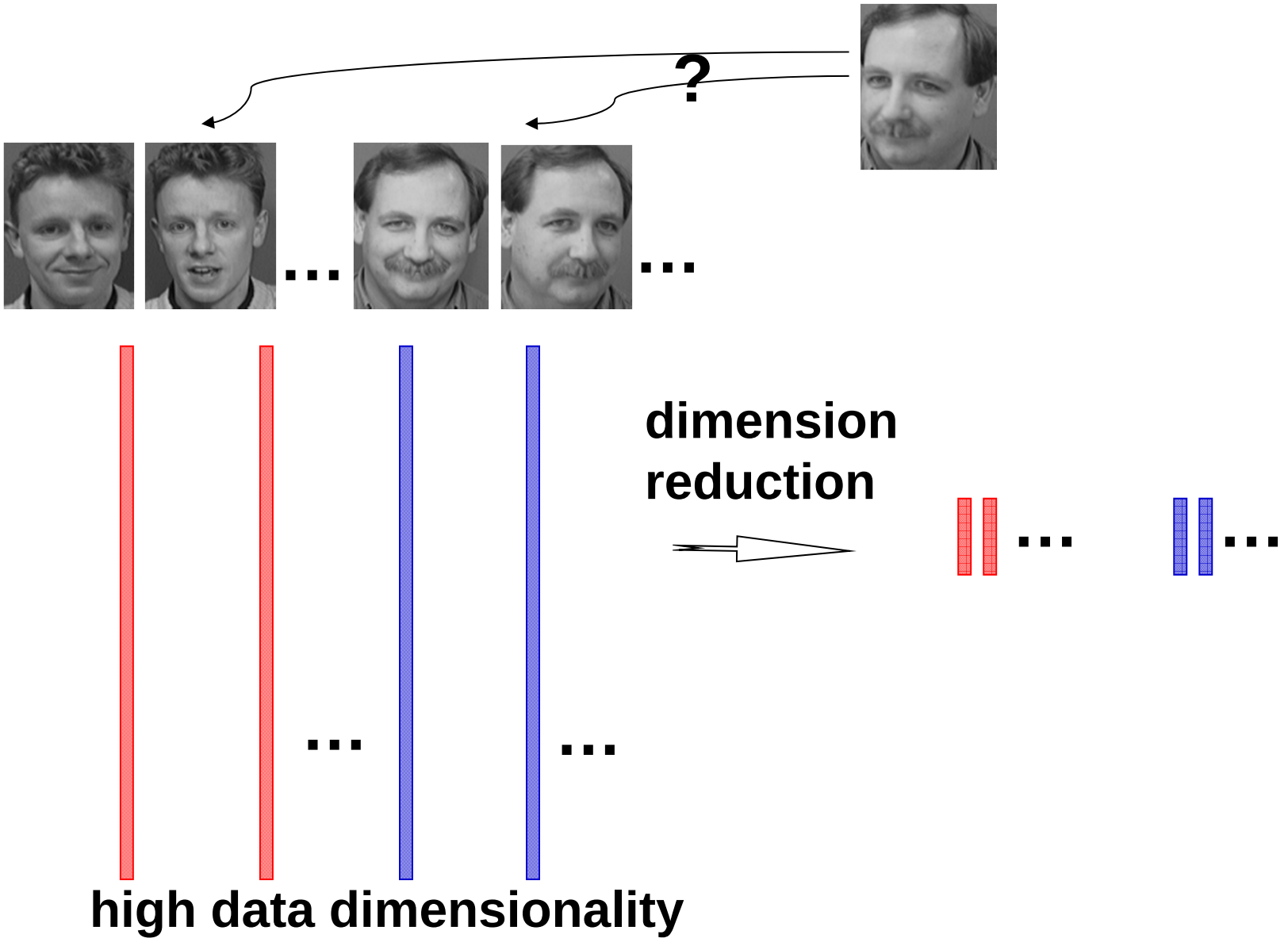
Low Dimensional Embedding of High-Dimensional Data



Why Dimension reduction?

- Reduce the dimensionality of high dimensional data
- Identify new meaningful underlying features
- The computational overhead of the subsequent processing stages is reduced
- Reduce noise effects
- Visualization of the data

Applications: Face recognition



Streaming data mining

- Data stream is a sequence of data samples which is continuously generated over time
- Concept drift
  - underlying data distribution may be changed
  - the concept of interest can be moved
- How to detect concept drift?
- How to adapt classification models incrementally?

reference

- Introduction to data mining, P.Tan and M. Steinbach and V. Kumar, Addison wesley, 2006
- Pattern Recognition and Machine Learning, Lucy Kuncheva, School of Computer Science, Bangor University