

Pattern Recognition & Machine Learning

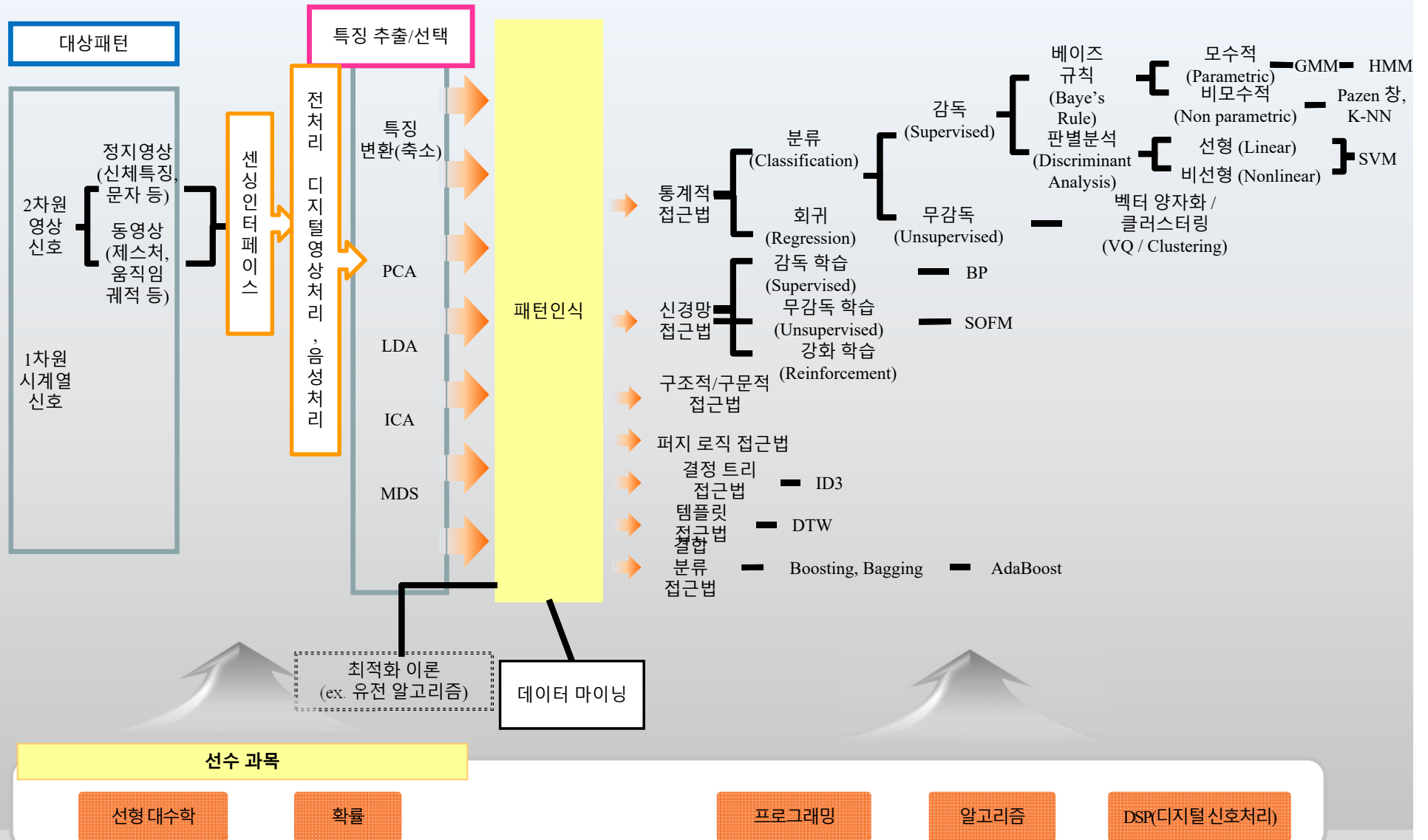
머신러닝 기반 빅데이터 엔지니어링 과정
빅데이터 X Campus (단국대학교)
2018.08
컴퓨터공학과 최상일 교수

01

Pattern Recognition



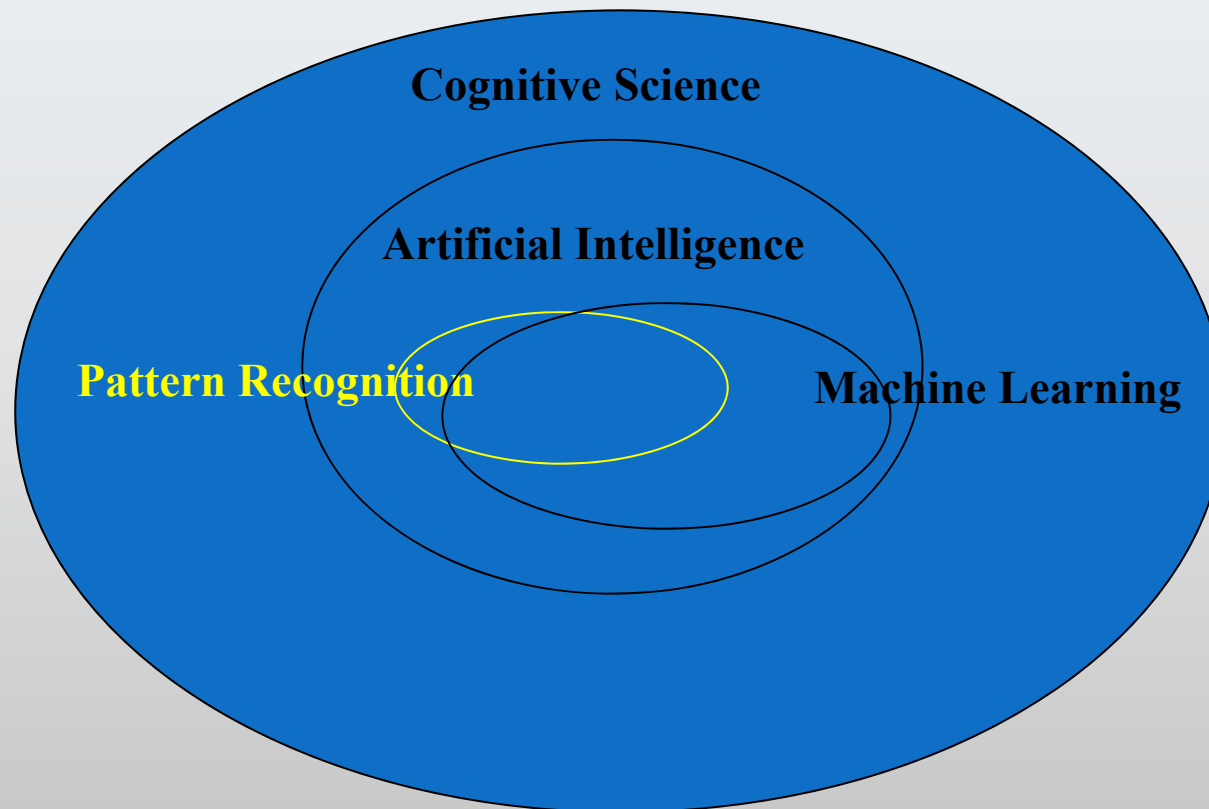
Pattern Recognition



❖ Background

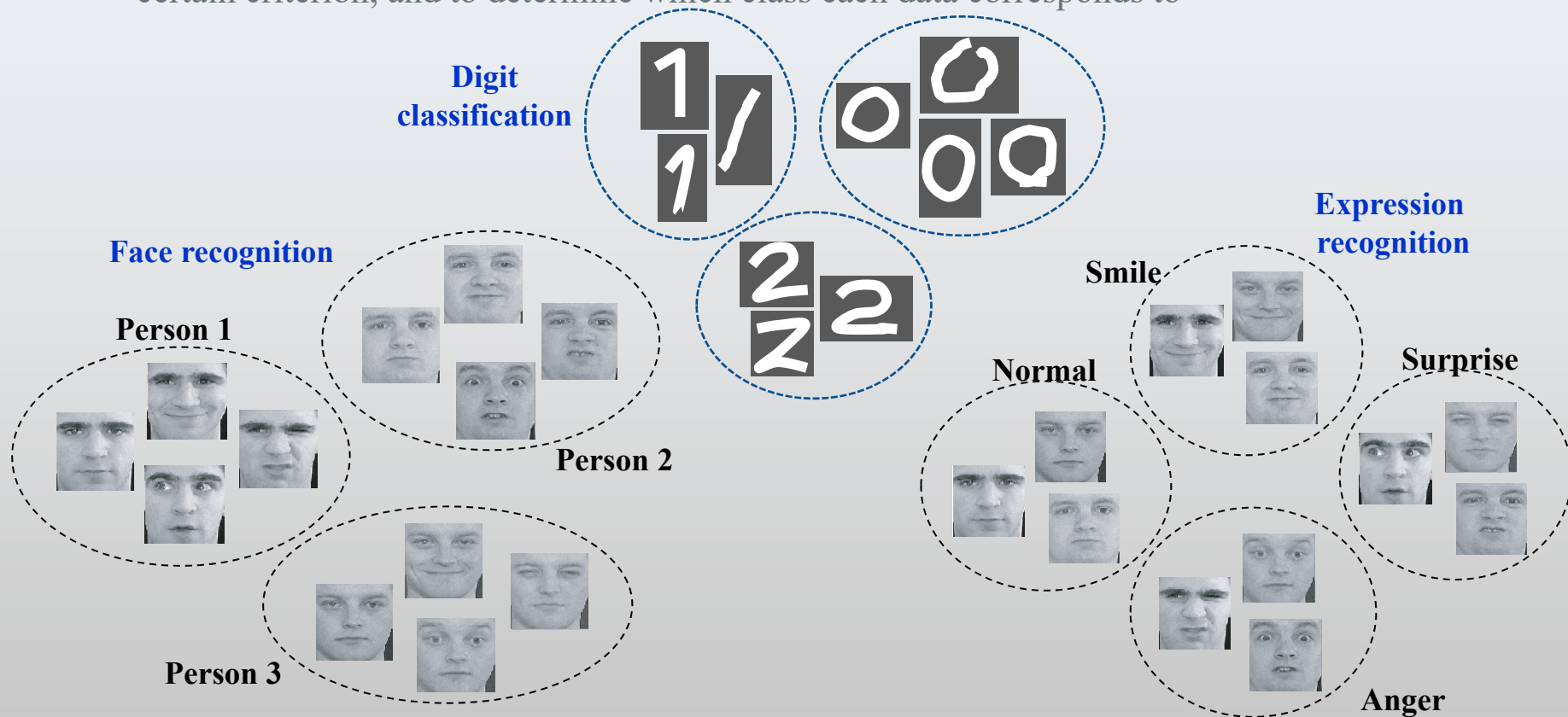
- The difference between man and machine is the attractiveness of attracting researchers
 - Recognition is extremely easy for human.
 - Recognition is extremely difficult for machine
- Scientific approach
 - Based on some understanding of the brain's information processing
 - The desire for a computer that imitates a brains
 - Neural Network
 - Use statistical analysis of data
 - Probability based recognition

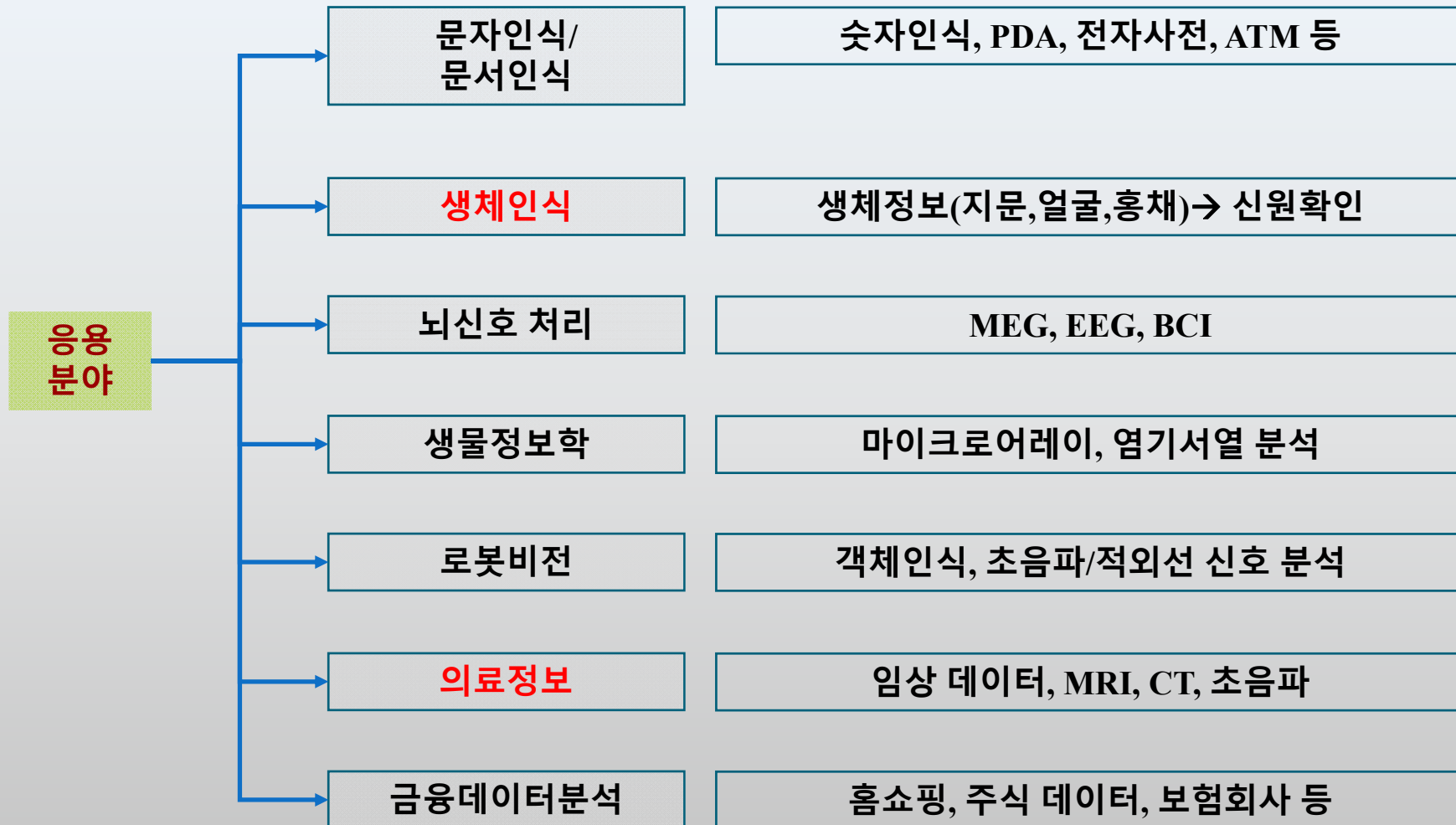
- A field of artificial intelligence that deals with the problem of recognizing objects that can be computed by a mechanical device (computer)



❖ Basic Goal

- Given input data (pattern) is divided into a group of several patterns (classes) according to a certain criterion, and to determine which class each data corresponds to





⋮

Table 1: Example pattern recognition applications.

Problem Domain	Application	Input Pattern	Pattern Classes
Document image analysis	Optical character recognition	Document image	Characters, words
Document classification	Internet search	Text document	Semantic categories
Document classification	Junk mail filtering	Email	Junk/non-junk
Multimedia database retrieval	Internet search	Video clip	Video genres
Speech recognition	Telephone directory assistance	Speech waveform	Spoken words
Natural language processing	Information extraction	Sentences	Parts of speech
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Medical	Diagnosis	Microscopic image	Cancerous/healthy cell
Military	Automatic target recognition	Optical or infrared image	Target type
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective/non-defective product
Industrial automation	Fruit sorting	Images taken on a conveyor belt	Grade of quality
Remote sensing	Forecasting crop yield	Multispectral image	Land use categories
Bioinformatics	Sequence analysis	DNA sequence	Known types of genes
Data mining	Searching for meaningful patterns	Points in multidimensional space	Compact and well-separated clusters

Pattern Recognition Applications



DANKOOK UNIVERSITY
Machine Learning &
Pattern Analysis lab.

From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

Nov 10, 1999

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

From
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

Nov 10, 1999

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

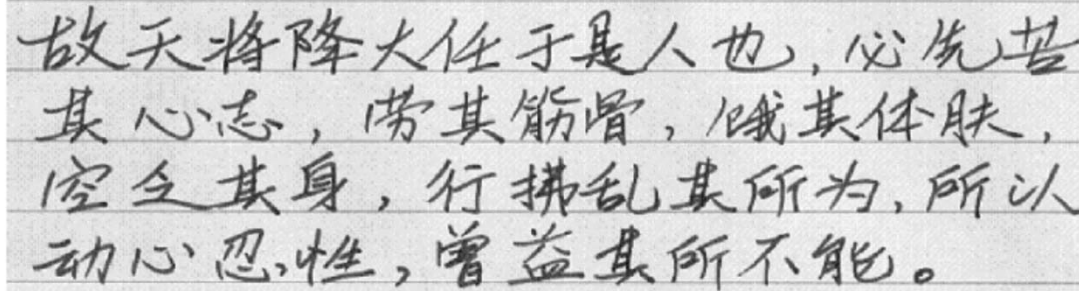
It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

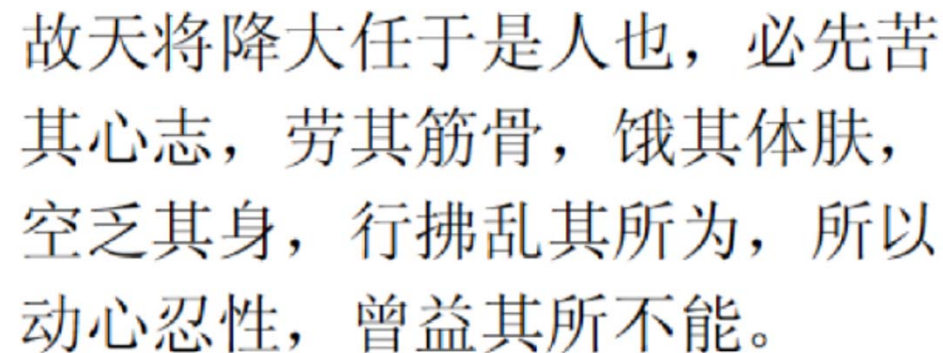
Thank you!
Jim

Figure 1: English handwriting recognition.

A photograph of a piece of paper with Chinese text written in black ink. The handwriting is cursive and fluid, typical of traditional Chinese calligraphy. The text is arranged in five horizontal lines.

故天将降大任于是人也，必先苦
其心志，劳其筋骨，饿其体肤，
空乏其身，行拂乱其所为，所以
动心忍性，曾益其所不能。

(a) Handwriting

A photograph of a piece of paper with the same Chinese text as in (a), but printed in a clean, standard serif font. The text is arranged in five horizontal lines.

故天将降大任于是人也，必先苦
其心志，劳其筋骨，饿其体肤，
空乏其身，行拂乱其所为，所以
动心忍性，曾益其所不能。

(b) Corresponding Machine Print

Figure 2: Chinese handwriting recognition.

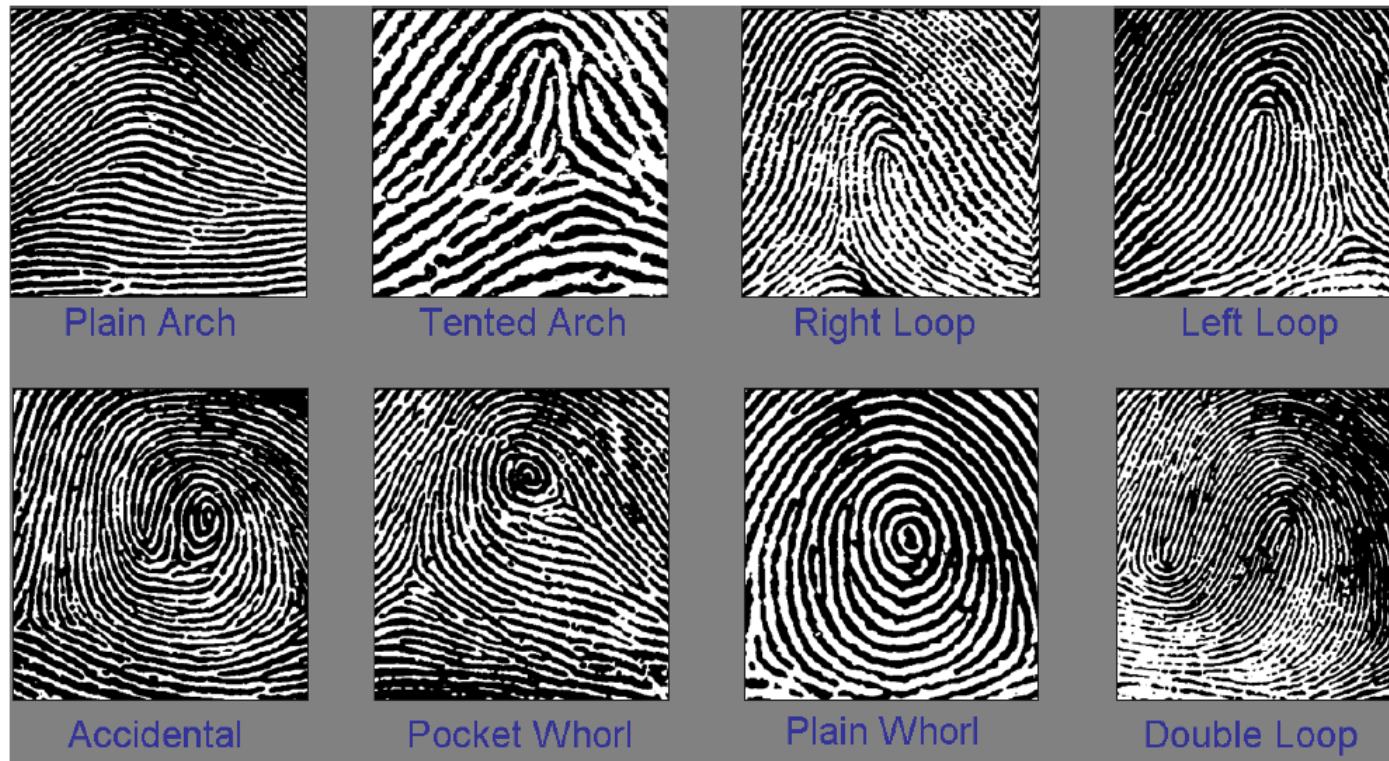


Figure 3: Fingerprint recognition.

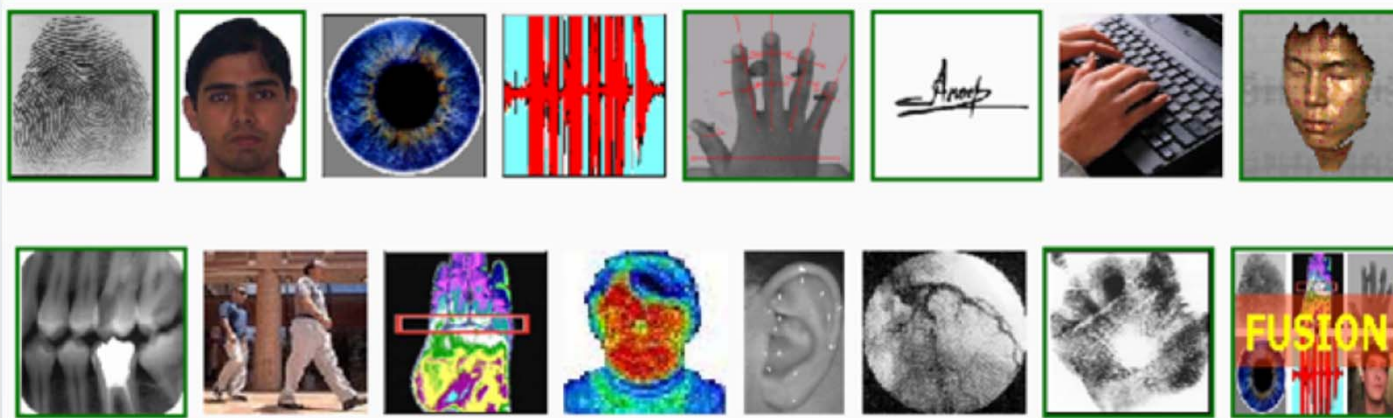


Figure 4: Biometric recognition.

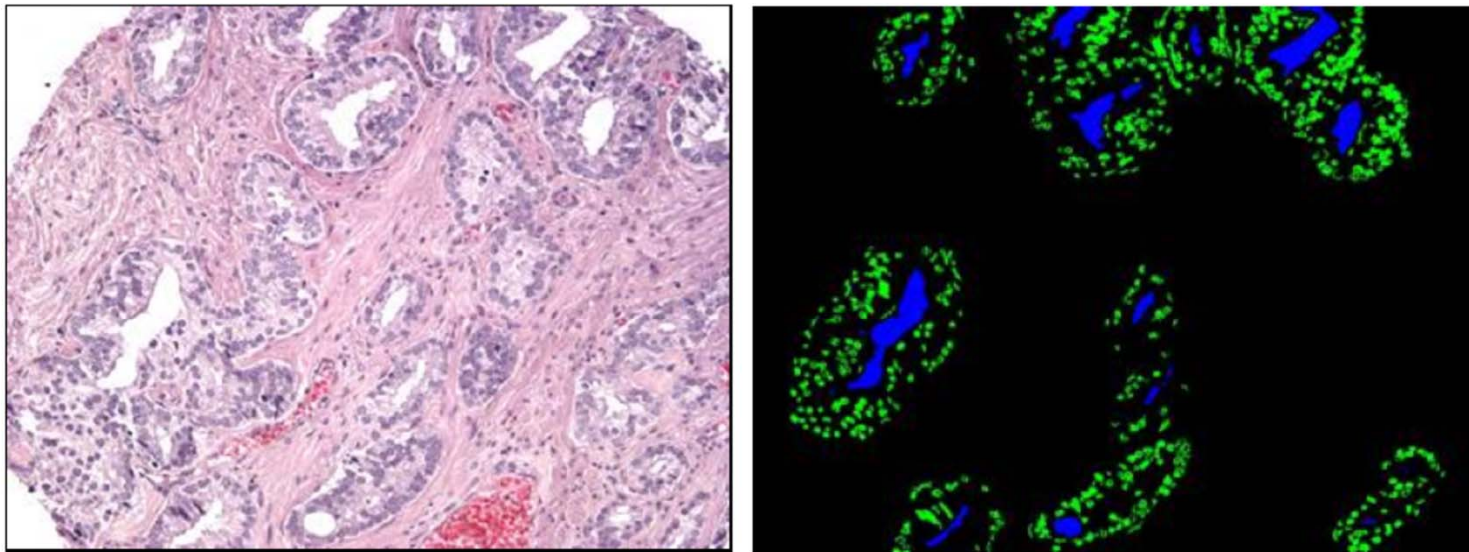


Figure 5: Cancer detection and grading using microscopic tissue data.



Figure 7: Land cover classification using satellite data.



Figure 9: License plate recognition: US license plates.

❖ Basic Approach of Pattern Recognition

- (Method1) Pattern definition and recognition by structural features



**What structural features are needed
for face/expression recognition?**



Limitation of structural method

■ (Method2) Template Matching

원형 영상
(template)



d_1



d_2



d_3



d_4



Input data

$d_3 = \min\{d_i\}$

Class

1

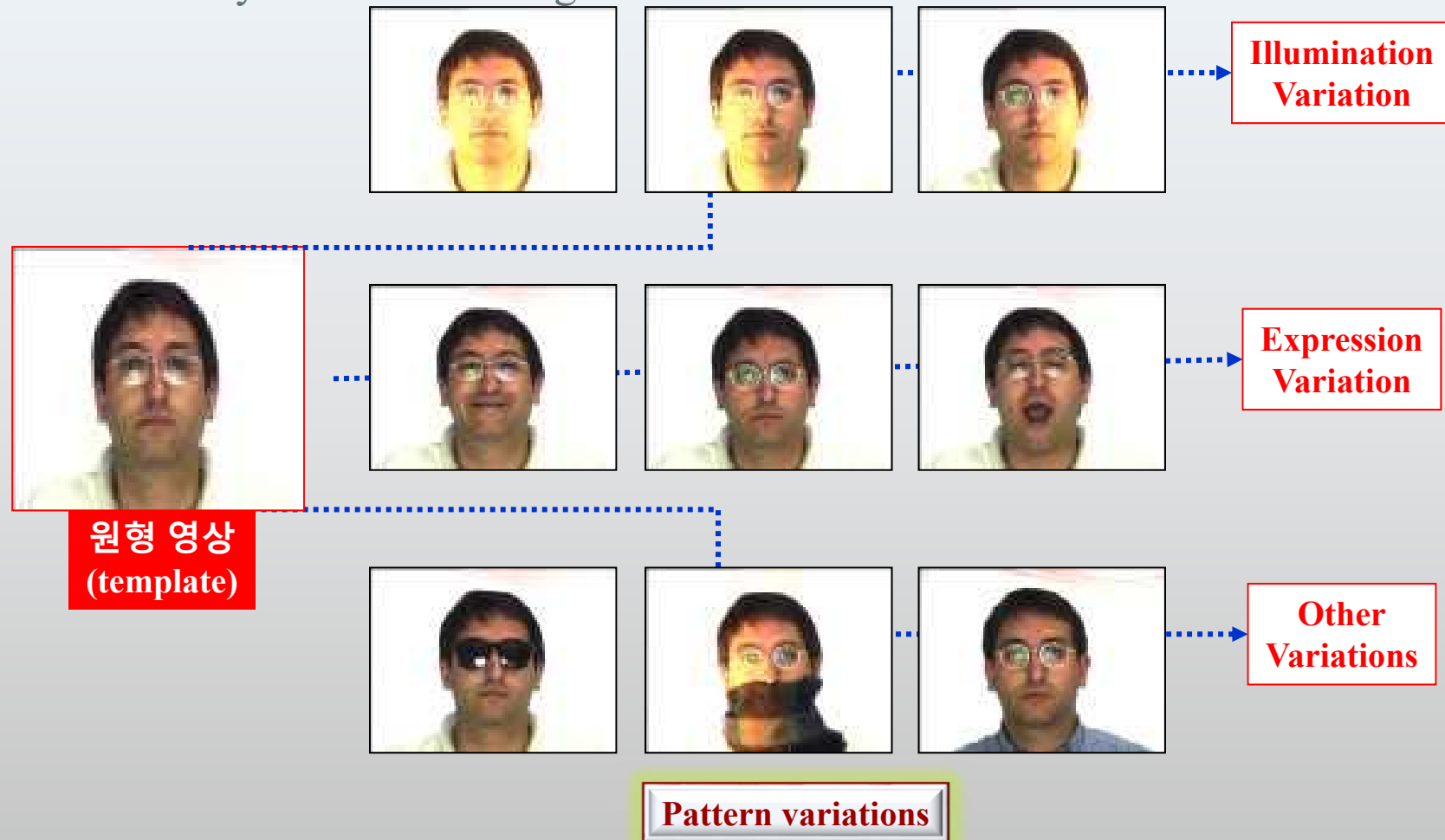
2

3

4

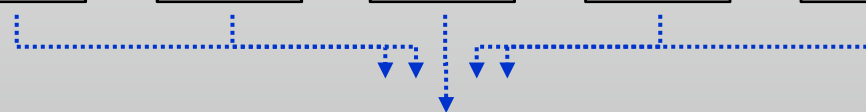


❖ Difficulty in Pattern Recognition



❖ Need for Machine Learning

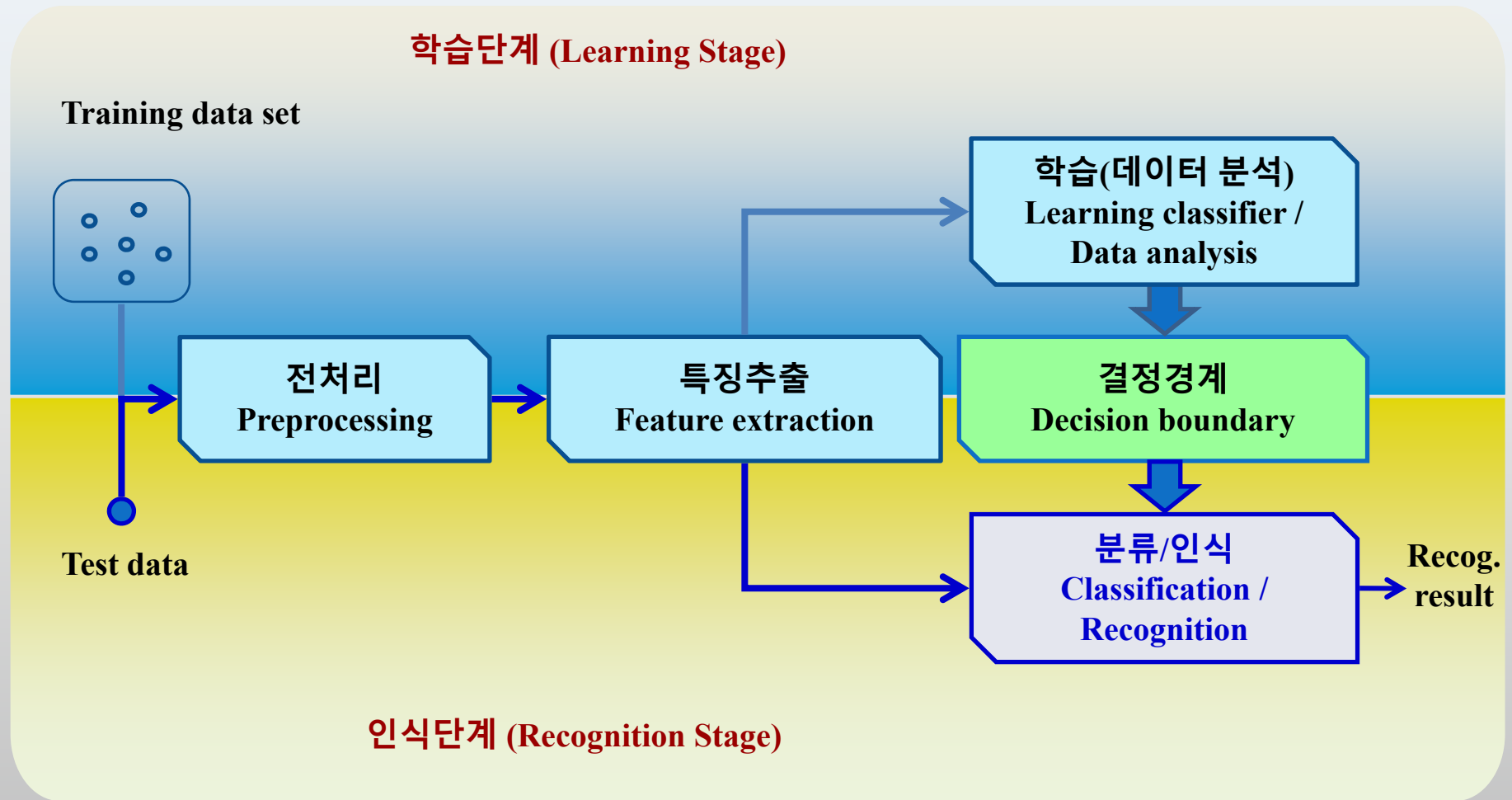
- Various types of variations of the pattern
 - Major cause of difficulty of pattern recognition
 - Need a more elaborate way
- “Machine Learning”
 - How to implement learning ability that is human's own intelligent function through machine
 - Develop a methodology that analyzes given data and **automatically** extracts **general rules** or new knowledge from it
 - Ex. Once you learn how to ride a bicycle, you can ride any bike
 - Ex. Variations of number ‘5’



Mean image

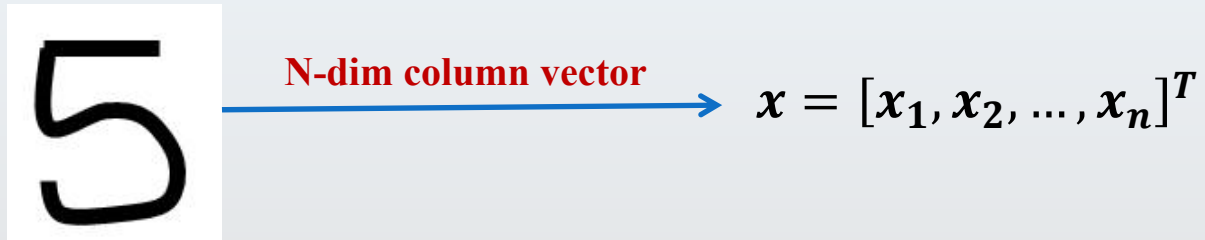
**Using statistical information
(mean)**

❖ Process of Pattern Recognition



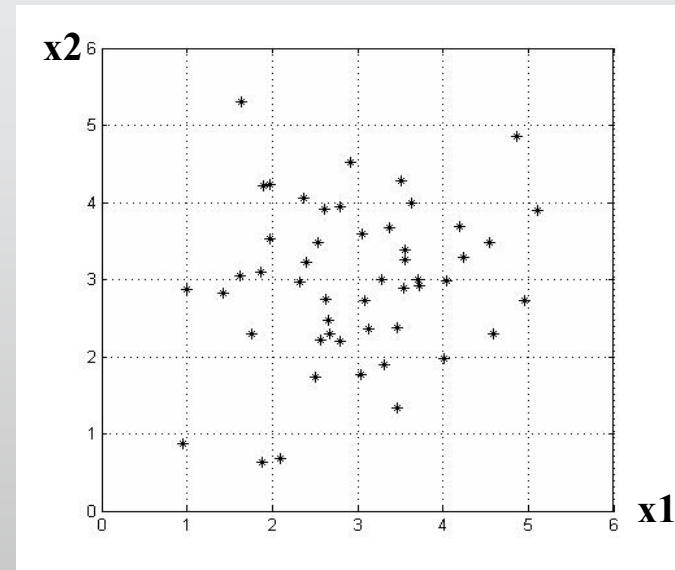
❖ Data Representation

■ Vector form

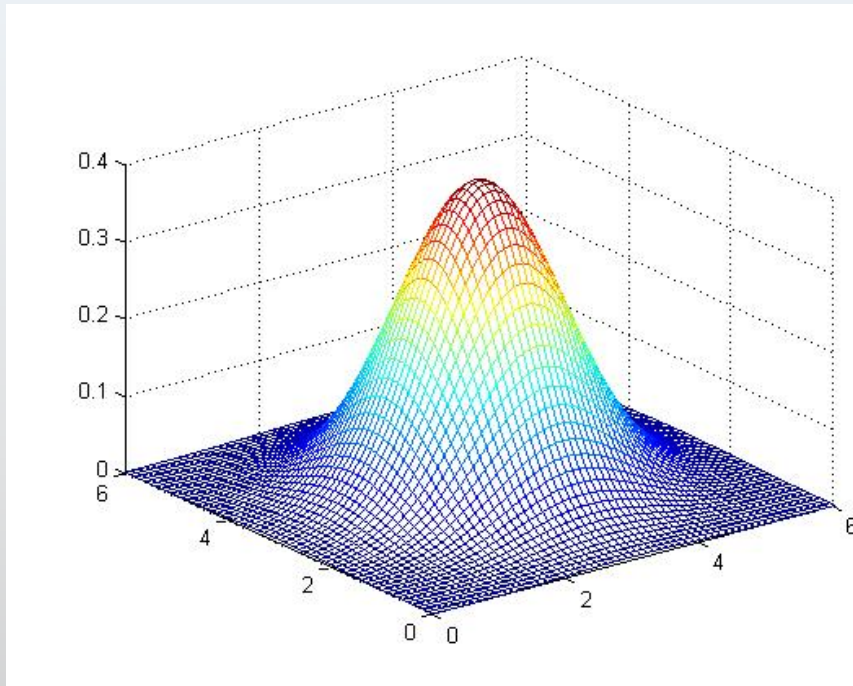


■ Characteristics of data distribution

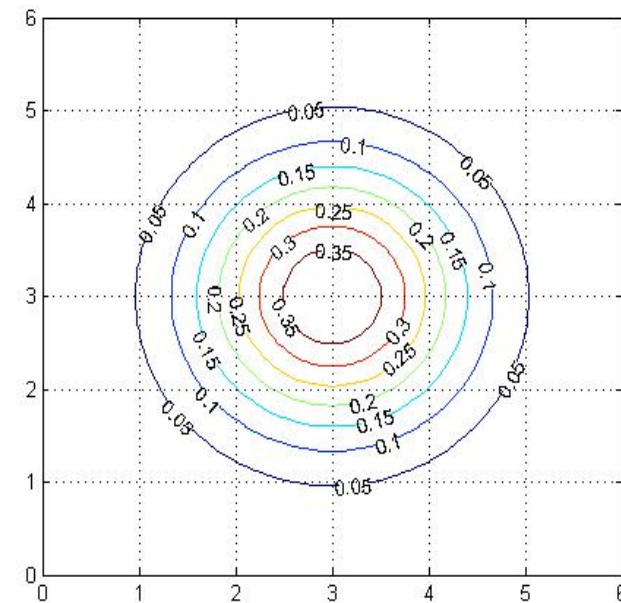
- Ex. Scatter plot of a two-dimensional dataset :
 - following the Gaussian distribution
 - mean $[3,3]$
 - covariance $[[1,0]^T, [0,1]^T]$
 - number of data sample: 50



❖ Plot of Data Distribution

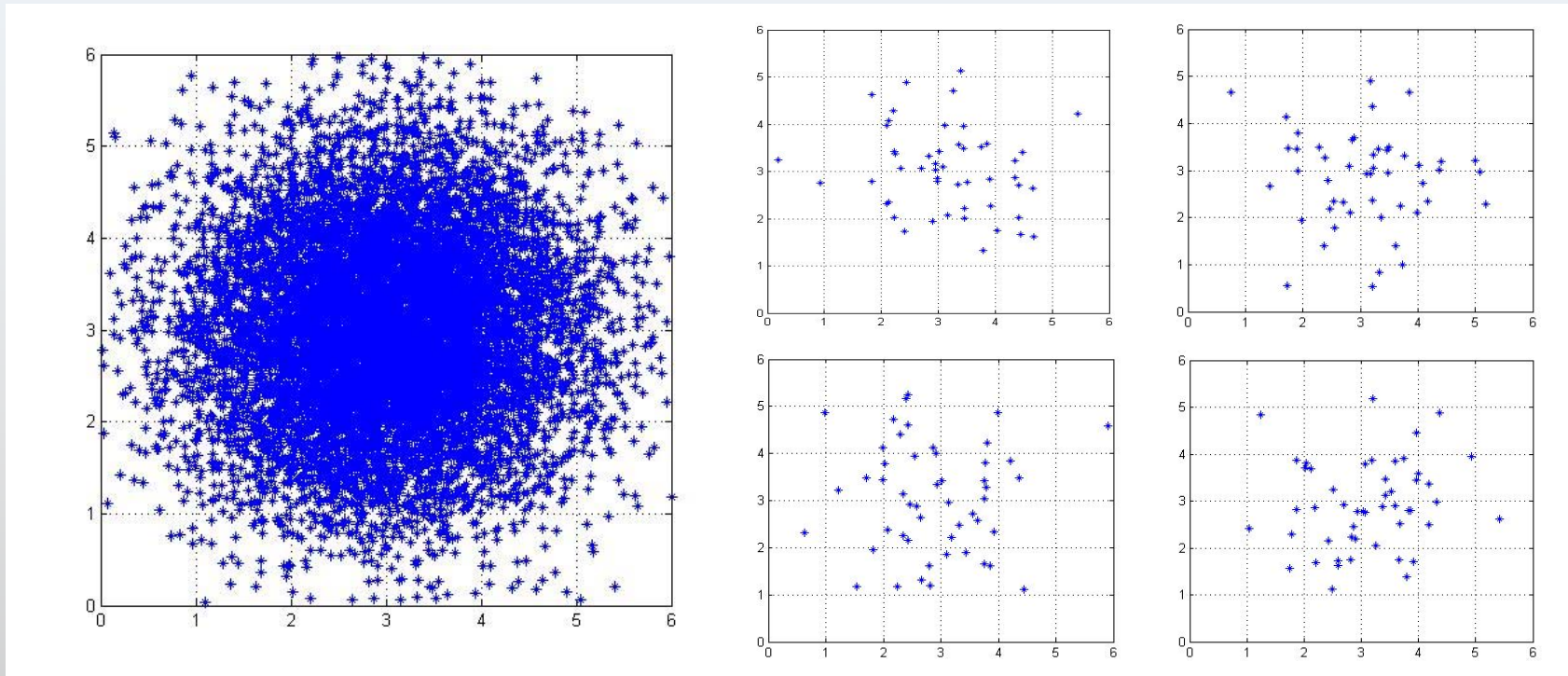


Probability density function of population



Density function denoted by contour lines

❖ Data Distribution (Training / Test data)



Probability density function of population

4 kinds of sample sets

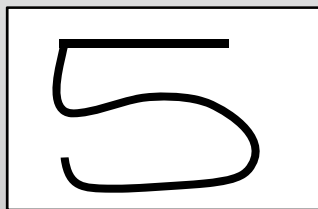
❖ Feature & Pattern

■ Feature

- The distinguishing aspect, quality, or characteristic of an object that an object has
- Instead of using the input data as it is, it extracts only key information (**features**) that can express the characteristics of each pattern
- Reducing the difficulty of problem solving due to noise, increasing cost (computation, memory)

■ Pattern

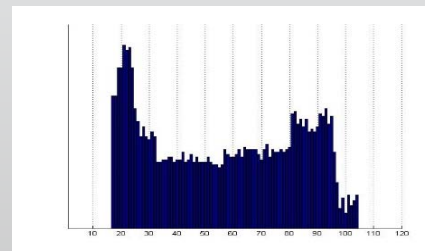
- A set of traits or features of an individual object



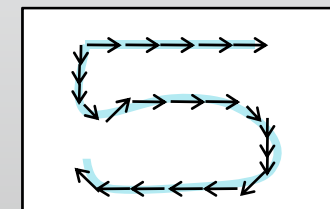
(a) Original Image



(b) Lattice Feature 12x12

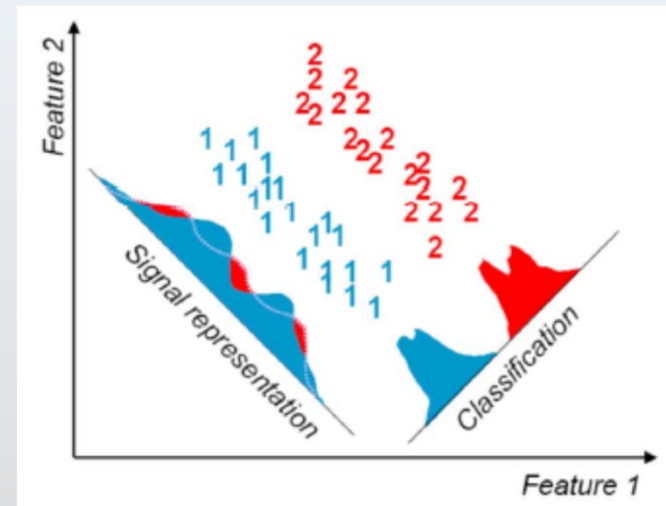
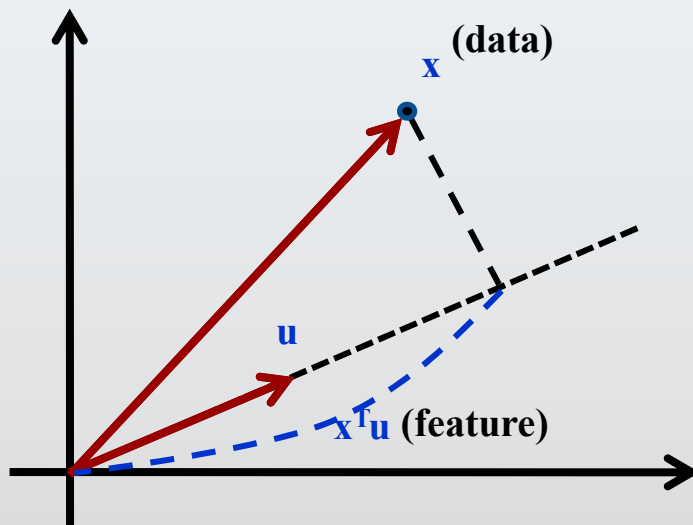


(c) Vertical histogram



(d) Direction feature

❖ Feature Extraction by Projection



■ What direction is it better to project?

- It is important to **extract key information for recognition**, not the purpose of dimension reduction itself
- Choose the direction that best represents the distribution characteristics of the given data
- Invariant features with respect to translation, rotation and scale

❖ Types of pattern recognition

분류(classification)

Classifying a given set of data into several already **defined classes**

Provided with input data and class labels for each data → $\{x_i, y(x_i)\}$

Digit recognition, face recognition, etc.

Bayes classifier
K-Nearest Neighbor method
Multilayer perceptrons
Support Vector Machine

회귀(regression)

Outputs the estimated result of the **real valued label** by regression

Provided with input data and target values for each data → $\{x_i, y(x_i)\}$

Weight prediction, stock price forecast, etc.

Least squares
Bayesian linear regression
Mixed models
Principal component regression

군집화(clustering)

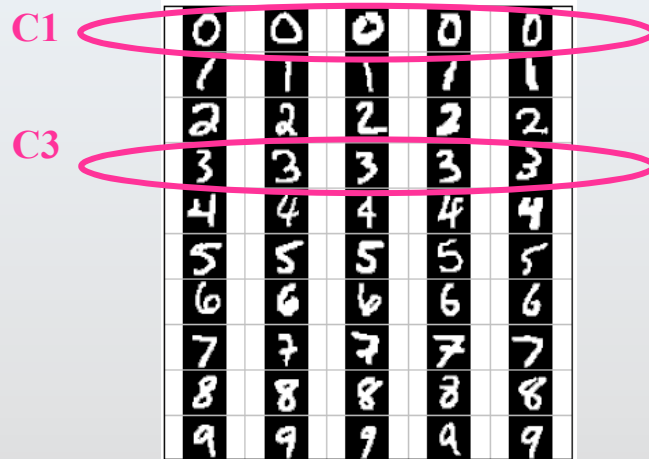
Analyzing the distribution characteristics (similarity of input values) of input data and **dividing them into arbitrary plural groups**

Provide only input values without information about classes → $\{x_i\}$

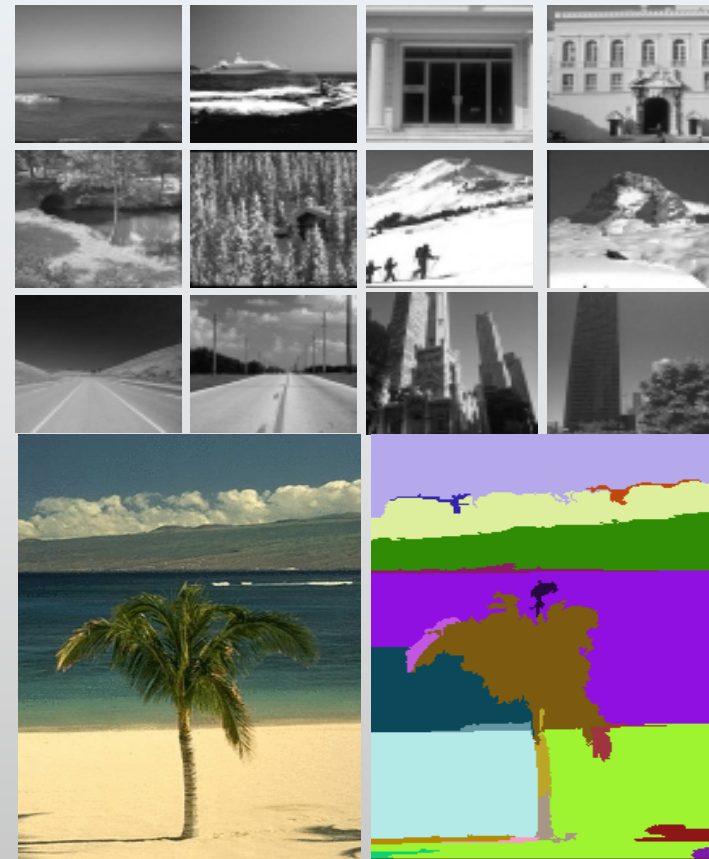
Image segmentation

K-means clustering
Learning Vector Quantization
Hierarchical clustering
Self Organizing feature Map

❖ Types of Pattern Recognition



Classification



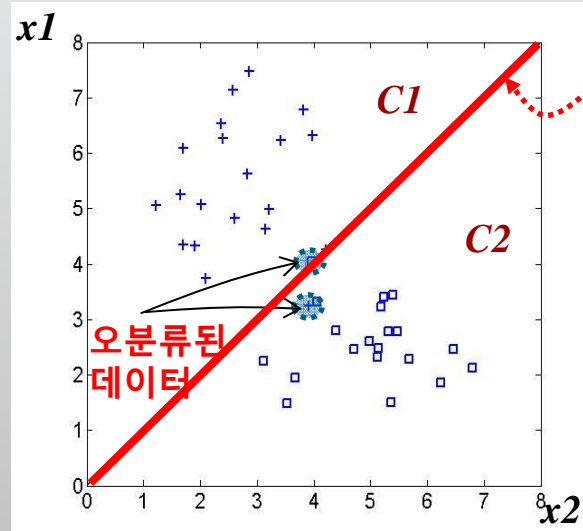
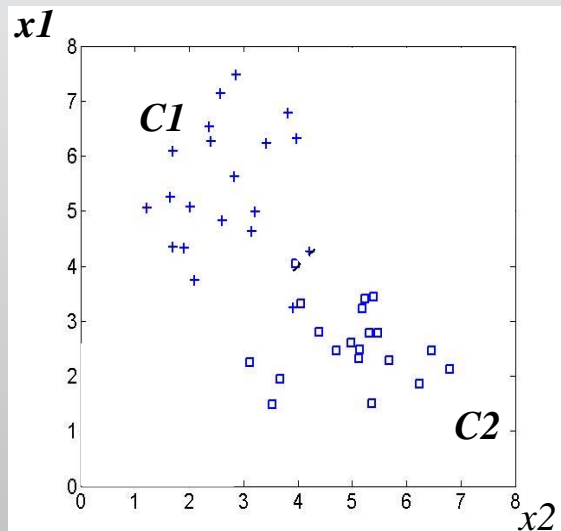
Clustering

❖ Classification

- The data set $X = \{x_1, x_2, \dots, x_N\}$ is given, the class label $y(x_i)$ corresponding to each data x_i is determined

❖ Decision Boundary

- Line / Curve / Plane to **distinguish the class**
- A boundary on the input/feature space defined by a function expression such as a discriminant $g(x) = 0$



결정경계

$$g(x_1, x_2) = x_1 - x_2 = 0$$

$$y(x) = \begin{cases} +1 & \text{if } g(x) \geq 0 \ (x \in C_1) \\ -1 & \text{if } g(x) < 0 \ (x \in C_2) \end{cases}$$

결정규칙

$$\text{sign}(x_2 - x_1)$$

❖ Classification Rates & Error

- Evaluation measurement
 - Classification rate

$$\text{분류율}(\%) = \frac{\text{Number of correctly classified samples}}{\text{Number of total samples}} \times 100$$

- Classification error rate

$$\text{분류오차}(\%) = \frac{\text{Number of incorrectly classified samples}}{\text{Number of total samples}} \times 100$$

- Training error

$$E_{train} = \frac{1}{N_{train}} \sum_{x \in X_{train}} \delta[t(x) - y(x)]$$

- Test error

$$E_{test} = \frac{1}{N_{test}} \sum_{x \in X_{test}} \delta[t(x) - y(x)]$$

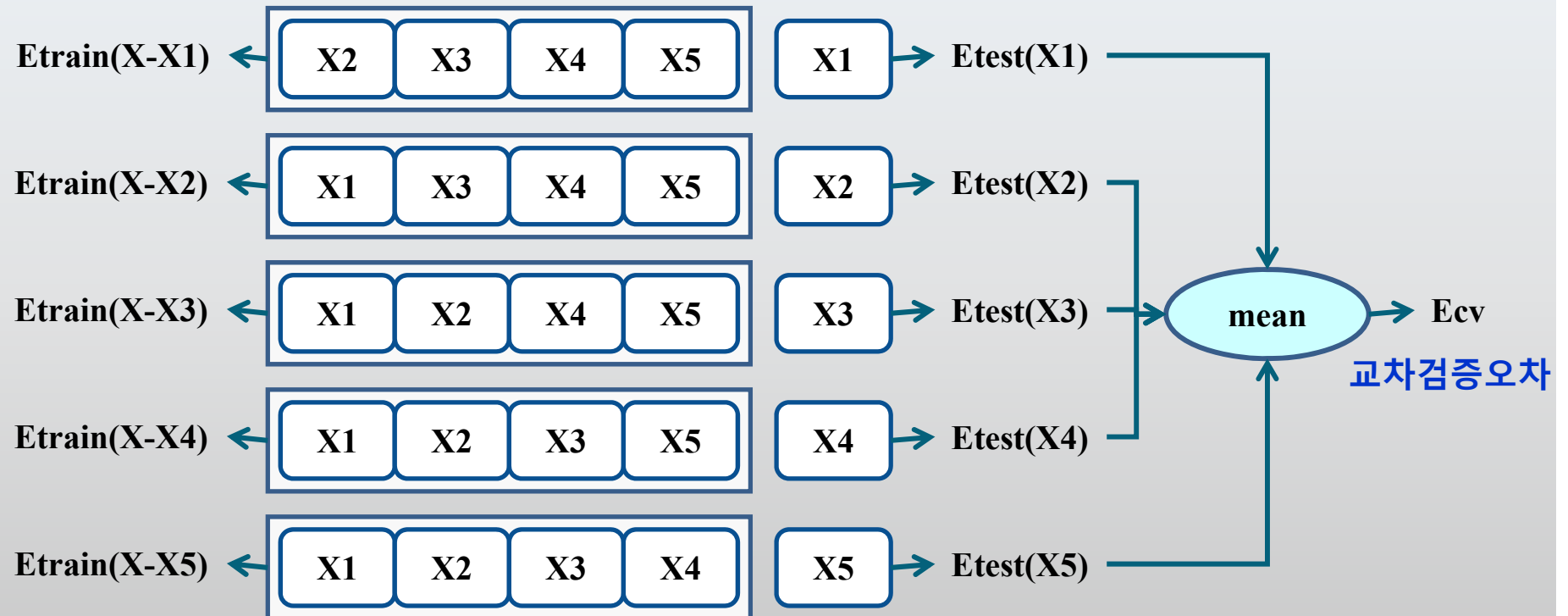
- Generalized error

$$\begin{aligned} E_{\text{gen}} &= \mathbb{E} [\delta[t(x) - y(x)]] \\ &= \int_{-\infty}^{\infty} \delta[t(x) - y(x)] p(x) dx \end{aligned}$$

- For theoretical performance analysis
- Can not be calculated in real applications
 - ∴ We do not know the probability density function of the whole data set
- Test error → ‘empirical error’
 - » Test errors are only an empirical approximation of generalized error

❖ Cross Validation

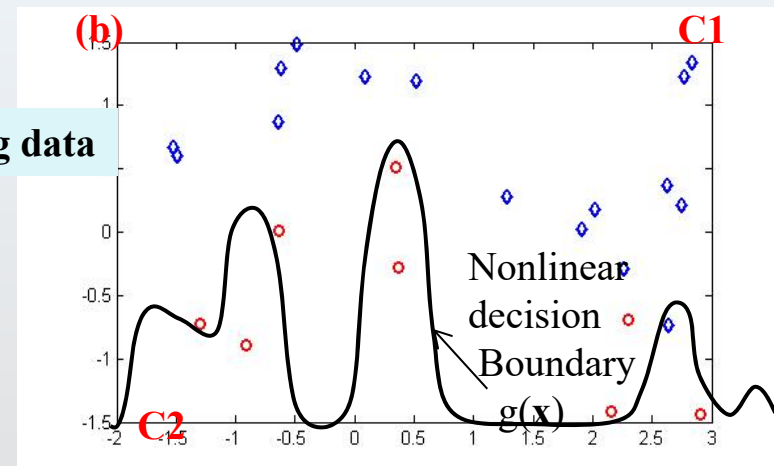
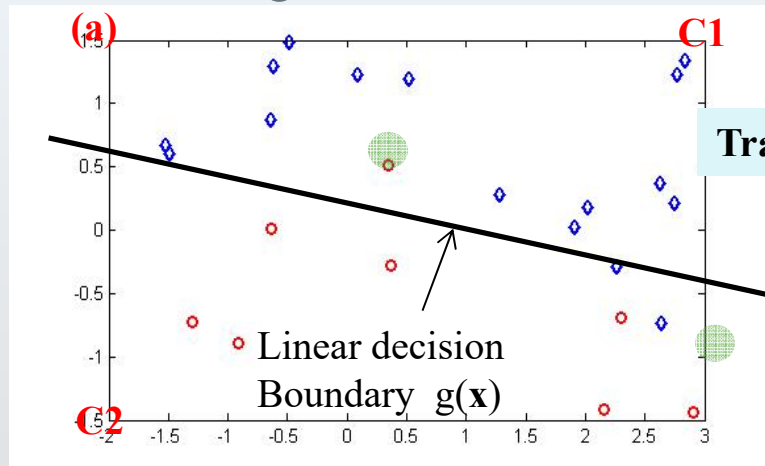
- Evaluation method when the number of data is small



❖ Overfitting

- The undesirable phenomenon that the classifier creates a decision boundary in an **overly suitable form only for the training data**
- Due to the lack of probabilistic noise of training data and the number of training data
- Need a way to adjust the complexity of classifier properly
 - Early termination of learning
 - Using error function with normalization term
 - How to choose a model

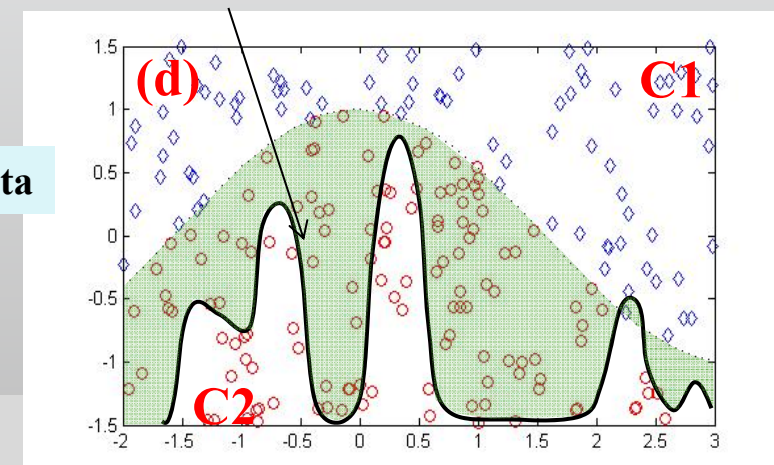
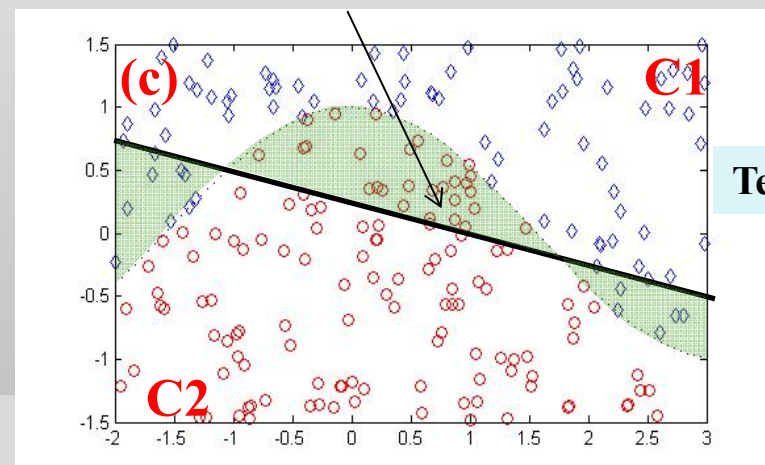
❖ Overfitting



Mis-classified area

Linear decision boundary

Non-Linear decision boundary



■ 교사학습(supervised learning)

- There is a teacher who pre-informs the desired output value of the recognizer when selecting the model
 - Least squares
 - perceptron learning
 - Error propagation learning algorithm for multi-layer perceptron
- Ex. Classification

■ 비교사학습(unsupervised learning)

- The type of learning without information about the desired output value of the recognizer during learning
 - EM method for Guassian mixture model
 - Self organized map
- Ex. Clustering

■ 강화학습(reinforcement learning)

- A learning method in which an agent (ex. Robot) receives a reward for behaviors performed by itself and acts in a better direction
 - The agent recognizes the current state and acts accordingly.
 - Finding a policy that is defined by a set of behaviors that maximizes the rewards that the agent will accumulate in the future

❖ Simple Classifier of Two-dimensional Data

- Generate probabilistic patterns along a specific probability distribution
 - Understand the overall development process of pattern recognizers
 - Ability to visually identify the data distribution characteristics and the decision boundaries of the recognizer
- Definition of data distribution
 - Uniformly distribution
 - » `rand()`
 - Gaussian distribution (Standard normal distribution)
 - » `randn()`

Example of Pattern Recognition

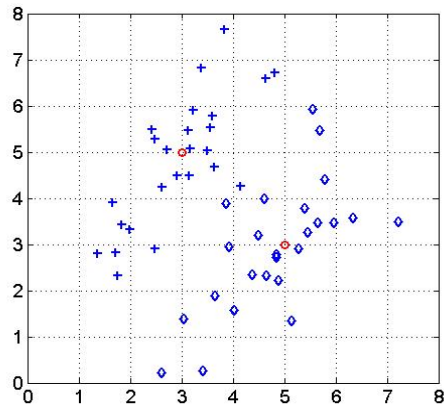
❖ Data Generation

$$p(x|C_1) \sim G(\mu_1, \Sigma_1)$$

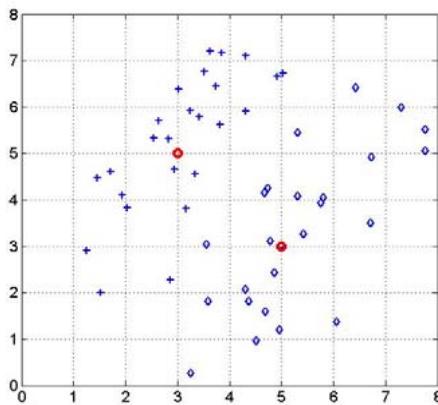
$$\mu_1 = \begin{pmatrix} 3 \\ 5 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

$$p(x|C_2) \sim G(\mu_2, \Sigma_2)$$

$$\mu_2 = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$



Training data



Test data

```
N=25;
```

```
m1= repmat([3,5], N,1);
```

```
m2= repmat([5,3], N,1);
```

```
s1=[1 1; 1 2];
```

```
s2=[1 1; 1 2];
```

```
X1=randn(N,2)*sqrtm(s1)+m1;
```

```
X2=randn(N,2)*sqrtm(s2)+m2;
```

```
plot(X1(:,1), X1(:,2), '+');
```

```
hold on;
```

```
plot(X2(:,1), X2(:,2), 'd');
```

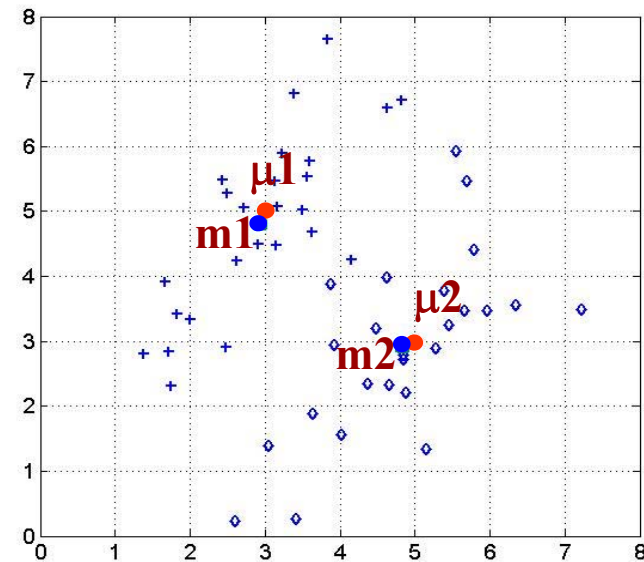
```
save data2_1 X1 X2;
```


❖ Training : Analysis of Data Distribution

- Find decision boundary
 - Estimate mean and covariance from training data

$$M_1 \begin{pmatrix} 2.94 \\ 4.80 \end{pmatrix}, S_1 = \begin{pmatrix} 0.86 & 0.99 \\ 0.99 & 1.93 \end{pmatrix}, M_2 = \begin{pmatrix} 4.83 \\ 2.91 \end{pmatrix}, S_2 = \begin{pmatrix} 1.14 & 0.97 \\ 0.97 & 1.89 \end{pmatrix}$$

```
load data2_1;  
m1 = mean(X1);  
m2 = mean(X2);  
s1 = cov(X1);  
s2 = cov(X2);  
save mean2_1 m1 m2 s1 s2;
```



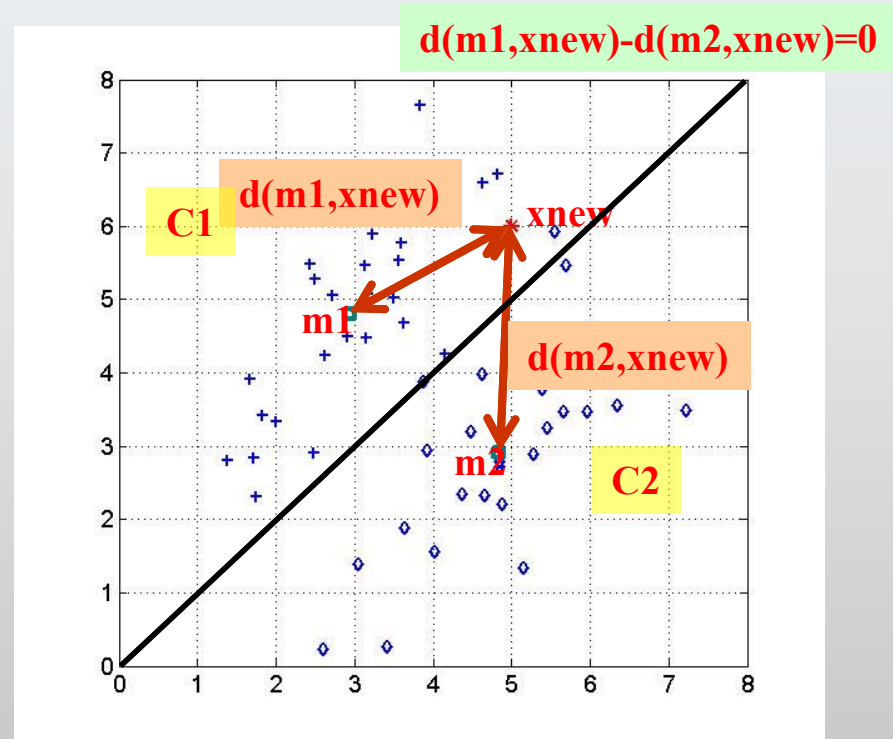
❖ Classification : Find Decision Boundary

- Discriminant function

- $G(x) = d(x, m_2) - d(x, m_1) = 0$

- Class label

- $y(x) = f(x) = \begin{cases} 1, & \text{if } g(x) > 0 \\ -1, & \text{if } g(x) < 0 \end{cases}$



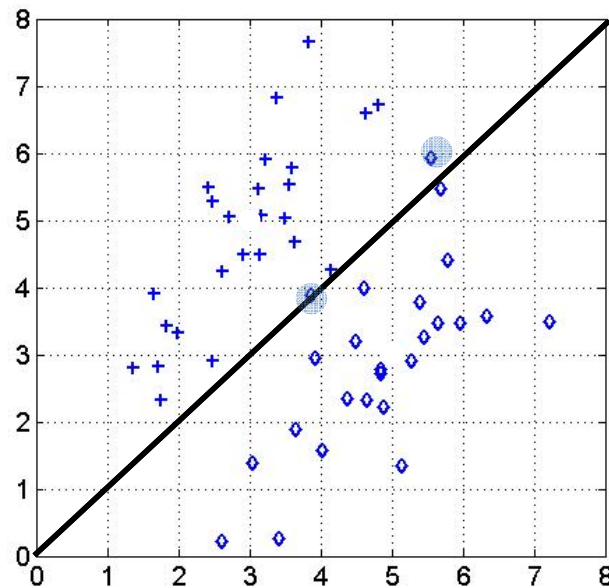
❖ Evaluation

■ Computation of training error

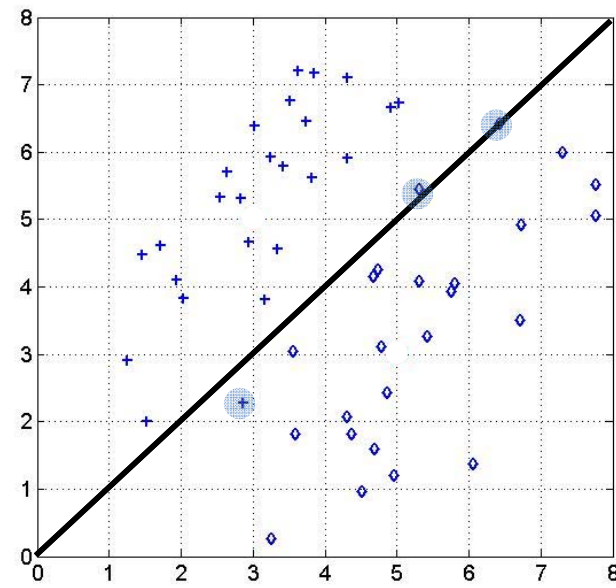
```
load data2_1;
load mean2_1
Etrain=0;
N=size(X1,1)
for i=1:N
    d1=norm(X1(i,:)-m1);
    d2=norm(X1(i,:)-m2);
    if (d1-d2) > 0 Etrain = Etrain+1; end
    d1=norm(X2(i,:)-m1);
    d2=norm(X2(i,:)-m2);
    if (d1-d2) < 0 Etrain = Etrain+1; end
end
fprintf(1,'Training Error = %.3f\n', Etrain/50);
```

Example of Pattern Recognition

❖ Evaluation



Training data 2/50



Test data 3/50

Statistical analysis

→ Probability distribution of data, pattern recognition

→ Probability density estimation

→ Parametric method

Maximum likelihood estimation

Maximum likelihood estimation
of Gaussian probability density function

→ Non-Parametric method

→ Histogram method

→ Generalization of Histogram method

→ Kernel density estimation method

Parzen window

Gaussian kernel

→ k-nearest neighbor rule

❖ Statistical Decision Theory

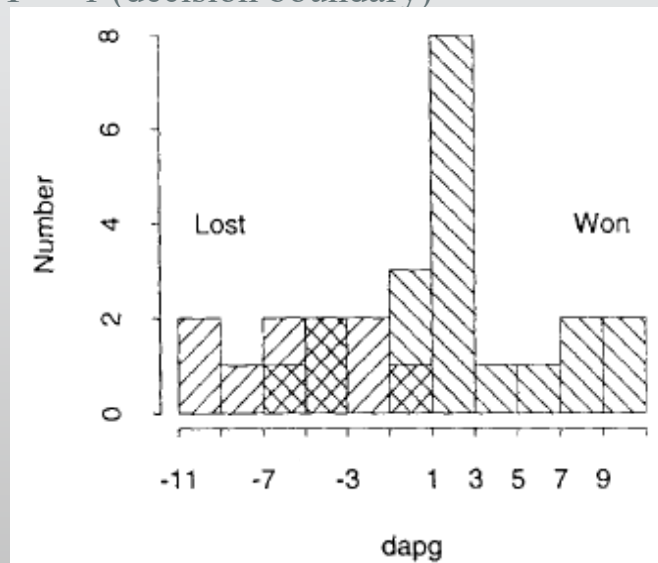
- Classification system can be designed on the basis of statistical or other decision theoretical techniques

ex) Bayes' theorem, nearest neighbor rule, etc.

- Example : problem of predicting the winner of a game

Game	<i>dapg</i>	Home Team		Game	<i>dapg</i>	Home Team
1	1.3	Won		16	-3.1	Won
2	-2.7	Lost		17	1.7	Won
3	-0.5	Won		18	2.8	Won
4	-3.2	Lost		19	4.6	Won
5	2.3	Won		20	3.0	Won
6	5.1	Won		21	0.7	Lost
7	-5.4	Lost		22	10.1	Won
8	8.2	Won		23	2.5	Won
9	-10.8	Lost		24	0.8	Won
10	-0.4	Won		25	-5.0	Lost
11	10.5	Won		26	8.1	Won
12	-1.1	Lost		27	-7.1	Lost
13	2.5	Won		28	2.7	Won
14	-4.2	Won		29	-10.0	Lost
15	-3.4	Lost		30	-6.5	Won

- Training set : scores of previously played games
- Problem : Given a game to be played, predict the result
- Feature : home team avg. point/game – visiting team avg. point/game
- Histogram
 - Convenient way to describe the data
 - Can predict the result by using threshold
 - Ex. set the threshold $T = -1$ (decision boundary)

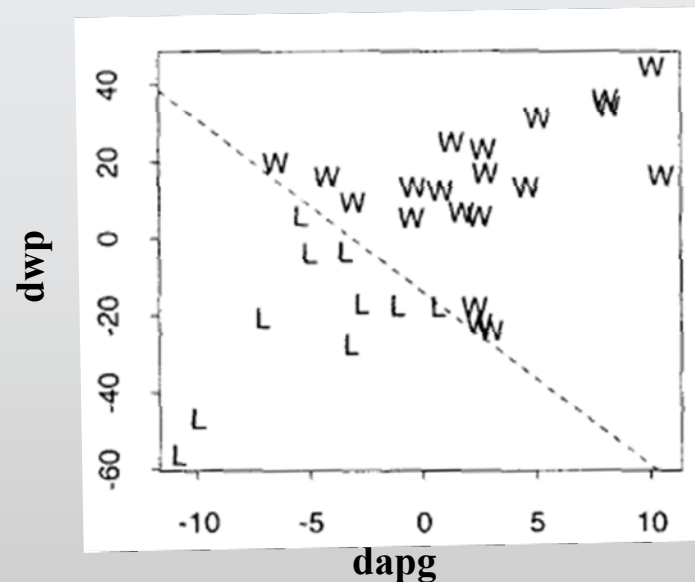


- Additional features often increases the accuracy of classification
 - dwp = Home team winning percent. – visiting team wp

Game	dapg	dwp	Home Team	Game	dapg	dwp	Home Team
1	1.3	25.0	Won	16	-3.1	9.4	Won
2	-2.7	-16.9	Lost	17	1.7	6.8	Won
3	-0.5	5.3	Won	18	2.8	17.0	Won
4	-3.2	-27.5	Lost	19	4.6	13.3	Won
5	2.3	-18.0	Won	20	3.0	-24.0	Won
6	5.1	31.2	Won	21	0.7	-17.8	Lost
7	-5.4	5.8	Lost	22	10.1	44.6	Won
8	8.2	34.3	Won	23	2.5	-22.4	Won
9	-10.8	-56.3	Lost	24	0.8	12.3	Won
10	-0.4	13.3	Won	25	-5.0	-3.8	Lost
11	10.5	16.3	Won	26	8.1	36.0	Won
12	-1.1	-17.6	Lost	27	-7.1	-20.6	Lost
13	2.5	5.7	Won	28	2.7	23.2	Won
14	-4.2	16.0	Won	29	-10.0	-46.9	Lost
15	-3.4	-3.4	Lost	30	-6.5	19.7	Won

■ Scatterplot

- Feature vector
 - (dapg, dwp)
- Feature space can be divided into two decision regions by straight line, called a linear decision boundary



■ Classifier

- A function $g : x \rightarrow \{1, 2, \dots, M\}$ represents one's guess of y given x .
- The mapping g is called a classifier.