Advanced Data Analysis
（データ解析特論）

Masashi Sugiyama
(Department of Computer Science)
杉山　将（計算工学専攻）

Contact:　W8E-505
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

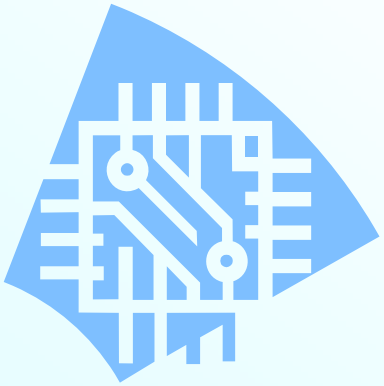
Contents of This Lecture (1)

- Syllabus (what I will provide in this course):

  The objective of this course is to introduce basic ideas and practical methods for discovering useful structure hidden in the data.

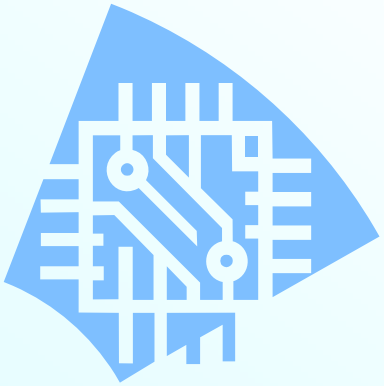
Statistical machine learning

and data mining



Contents of This Lecture (2)

- What you are expected to learn from this course:
  - How to use data analysis tools.
  - Conceptual ideas behind the methods.
  - Something useful in your own research/life.



Grading System

- Regular homework (every week):

$$s_H, \quad 0 \leq s_H \leq 80$$

- Final presentation & report:

$$s_P, \quad 0 \leq s_P \leq 80$$

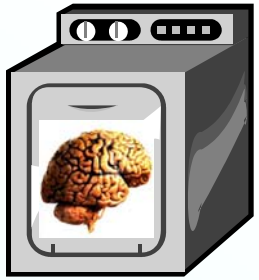
- Final score:

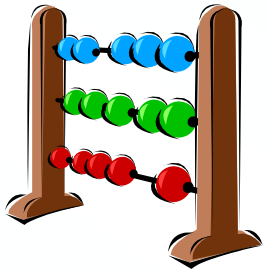
$$s_F = \min(100, s), \quad s = s_H + s_P$$

- Note: s may be non-linearly rescaled depending on the distribution of scores.

Brief Overview of the Course (1)

- 3 topics in the research of “learning”
  - Understanding human brains
  - Developing learning machines
  - Mathematically clarifying mechanism of learning

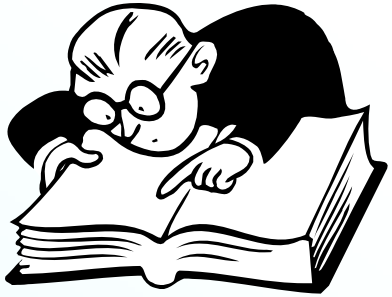


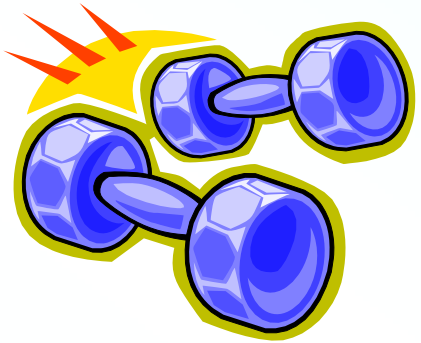


Brief Overview of the Course (2)

3 types of learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning





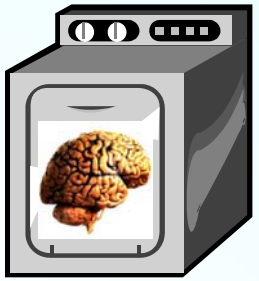
Brief Overview of the Course (3)

- Topics in unsupervised learning
  - Dimensionality reduction
  - Data clustering
  - Blind source separation
  - Outlier/novelty detection

Textbook

- Handouts are provided if necessary.
- Pointers to relevant articles will be provided.
- The following reference may be useful for learning elementally (and advanced) matrix calculations.
  - Matrix Cookbook: http://matrixcookbook.com/

3 Topics in Learning Research

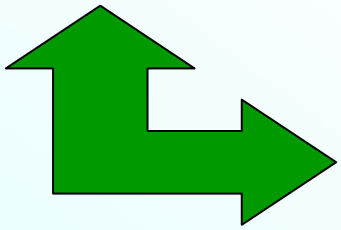


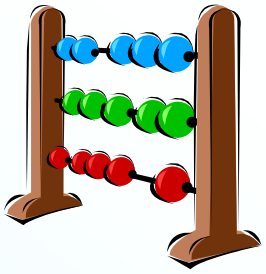
Understanding the brain

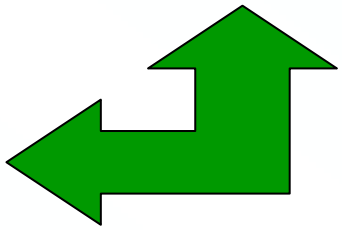
(Physiology, psychology, neuroscience)

Developing learning machines

(Computer and electronic engineering)







Mathematically clarifying mechanism of learning

(Computer and information science)

Understanding the Brain (1)

- Our brain consists of tens of billion neurons.
- Neurons are connected to each other like a network.

From http://www.si.hirosaki-u.ac.jp/~ei_sys/neuro/nnindex.html

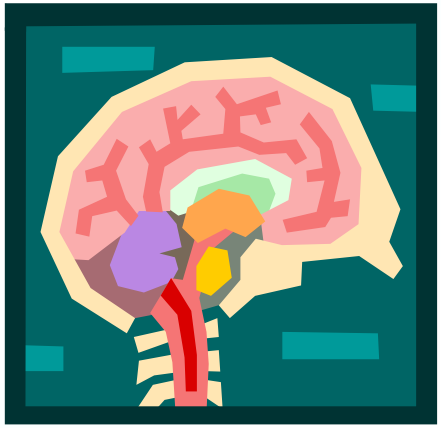
Understanding the Brain (2)

- Each neuron has dendrites and axons, and the axon is connected with other neurons via synapses.
- Neurons receive signals from other neurons through dendrites and send signals through axons.

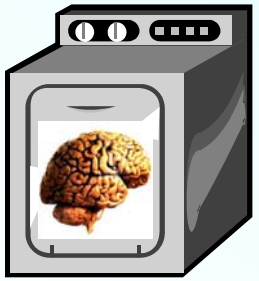
From http://www.si.hirosaki-u.ac.jp/~ei_sys/neuro/nnindex.html

Understanding the Brain (3)

- Structures and mechanisms of the brain have been clarified considerably.
- However, it is not still clear how learning is carried out using a number of neurons.



3 Topics in Learning Research

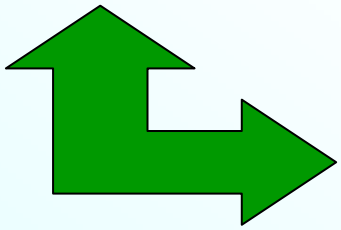


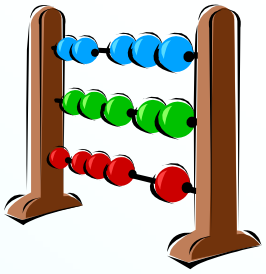
Understanding the brain

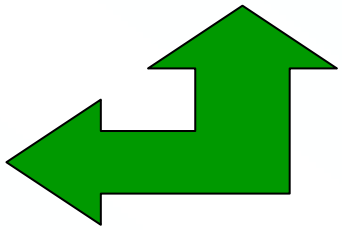
(Physiology, psychology, neuroscience)

Developing learning machines

(Computer and electronic engineering)





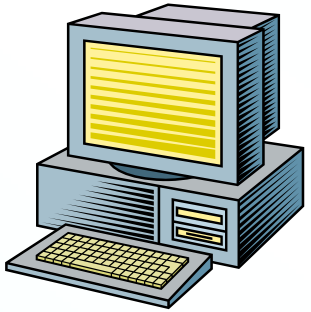


Mathematically clarifying mechanism of learning

(Computer and information science)

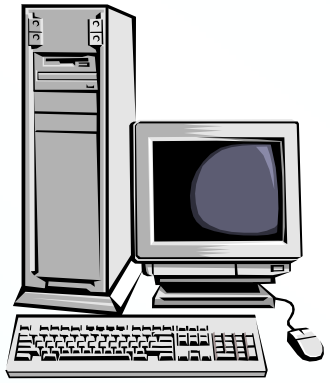
Developing Learning Machines (1)

- Computers we are usually using are called the von Neumann-type.
- Computing principles are based on logical computation and symbol processing.
- Computational theories of Turing machines play central roles.



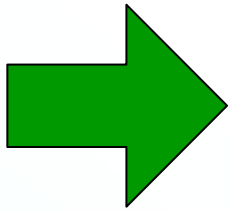
Developing Learning Machines (2)

- Suitable for repeating simple straightforward calculation or processing the data following prescribed procedures.
- However, even state-of-the-art computers are inferior to babies in performing complex tasks such as recognizing humans' faces.



Developing Learning Machines (3)

- A computer that imitates information processing carried out in our brains is being developed (neurocomputer).



Biological neurons

Artificial neurons

Developing Learning Machines (4)

We want neurocomputes to be equipped with the following functions:

- Adaptable to new environments, i.e., no need to prescribe responses for all possible situations.
- Handle vague, noisy, and contradictory information.
- A number of artificial neurons work independently.
- Robust against noise, especially, faults of other neurons.
- Small and efficient in electricity consumption.

Developing Learning Machines (6)

- Several realizations of neurocomputers with electronic or optical circuits have been proposed.

  Pulse Density Modulating Digital Neural Network System developed by University of Tsukuba

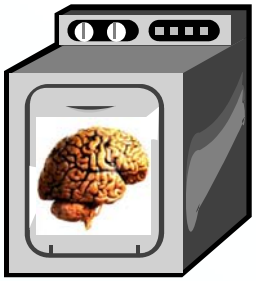
From http://www.viplab.is.tsukuba.ac.jp/~hirai/PDM/index.html

# of artificial neurons	1,008
# of synapses	1,028,160
Internal Potential	12bit
Output	11bit
Synapse weight	7bit
Time constants	From 516μs to 26.4 ms
Max output frequency	5MHz or 10 MHz
Main clock frequency	20MHz

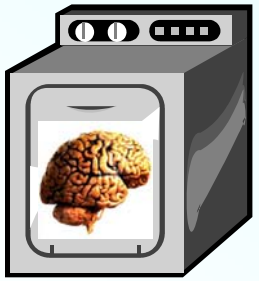
Developing Learning Machines (8)

However, current neurocomputers have the following problems:

- The number of neurons is not so large.
- Size is big.
- It is not clear how to train the computer!!



3 Topics in Learning Research

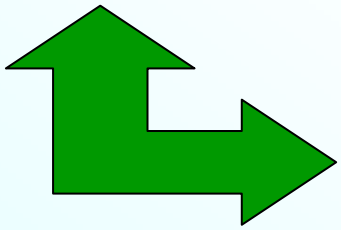


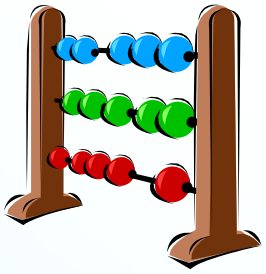
Understanding the brain

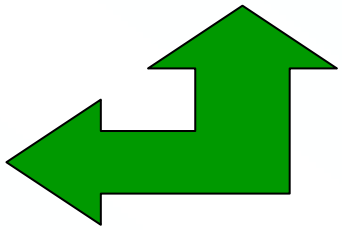
(Physiology, psychology, neuroscience)

Developing learning machines

(Computer and electronic engineering)





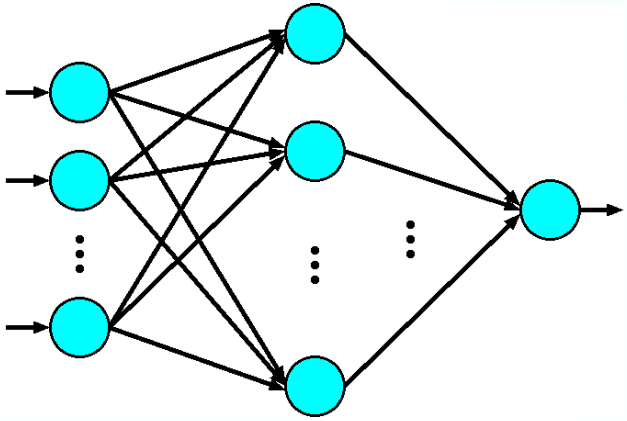


Mathematically clarifying mechanism of learning

(Computer and information science)

Clarifying Learning Mathematically (1)

- In order to understand our brains and develop neurocomputers, we need to clarify how information is processed using a number of neurons.



Clarifying Learning Mathematically (2)

- Our brains have been formed through longtime evolution, so they do not necessarily have the optimal structure.
- When developing learning machines, their architecture should be computer-scientifically suitable, rather than just imitating humans' brains.

Clarifying Learning Mathematically (3)

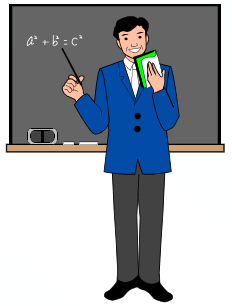
Mathematical tools for clarifying the essence of learning

- Mathematical statistics
- Algebraic geometry
- Functional analysis
- Information geometry
- Statistical physics
- etc.

A Little Break…

- There are 3 topics in learning research.
  - Understanding human brains
  - Developing learning machines
  - Mathematically clarifying mechanism of learning
- The third topic plays an important role for achieving the first two goals.
- We focus on the third topic:

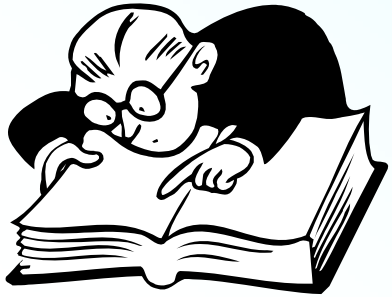
  “Theories of learning”



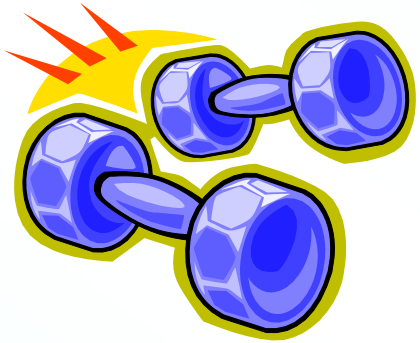
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2012 spring)

- Unsupervised learning

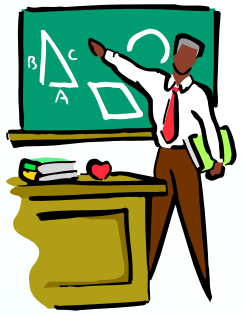


- Reinforcement learning



What Is Supervised Learning?

- The goal of supervised learning is to estimate an unknown input-output rule.
- You are allowed to ask questions to a supervisor (“oracle”) who knows the rule.
- The supervisor answers your questions using the rule.



What Is Supervised Learning?

- Pairs of questions and answers are called training examples.
- If the underlying rule can be successfully estimated, we can answer untaught questions.
- Such an ability is called the generalization capability.

Example 1

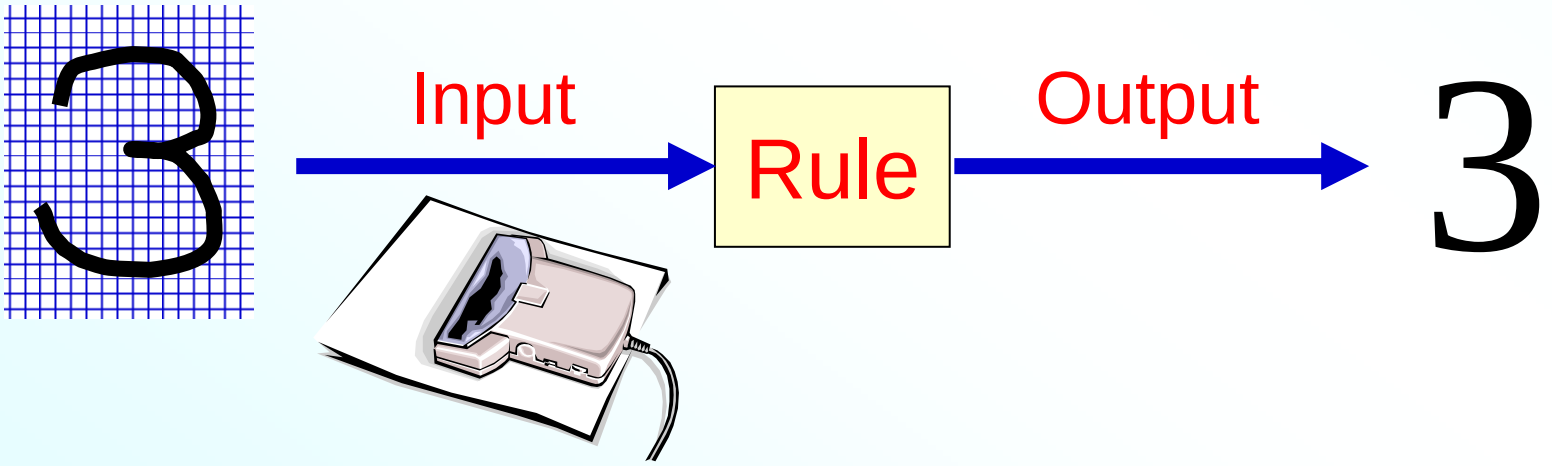
- Hand-written number recognition

  We want to recognize scanned hand-written characters.

MNIST data from LeCun et al., Proc. IEEE, 1998

Example 1

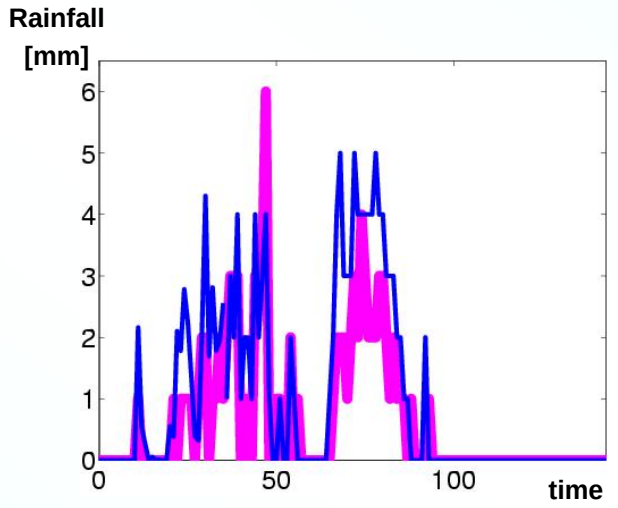
- Training examples consist of
  { (hand-written number, its recognition result) }.
- If underlying input-output rule is successfully learned, unlearned hand-written numbers can be recognized.



Example 2

- Rainfall Estimation

  Using the past rainfall and weather radar data, we want to estimate the rainfall tomorrow.



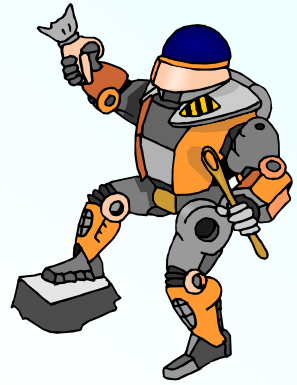
Weather radar data from IEICE Precipitation Estimation Contest 2001

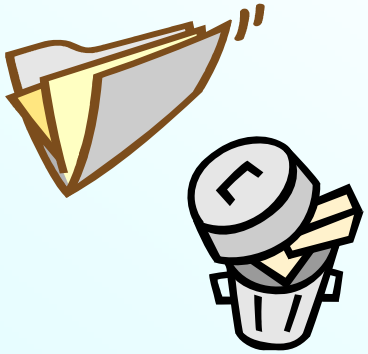
Example 2

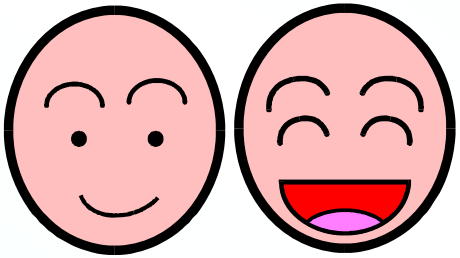
- Training examples are
  {(past rainfall and radar data,
  rainfall the next day)}
- If the rule is successfully learned, we can estimate the future rainfall by using the past rainfall and radar data.

Other Examples

- Other examples are…
  - Stock price estimation
  - Robot motor control
  - Computer vision
  - Spam filter
  - DNA classification



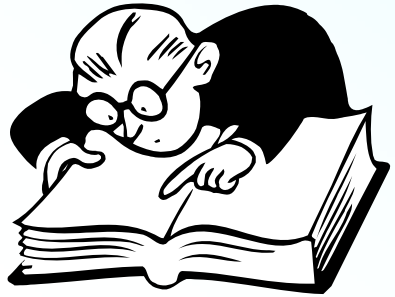




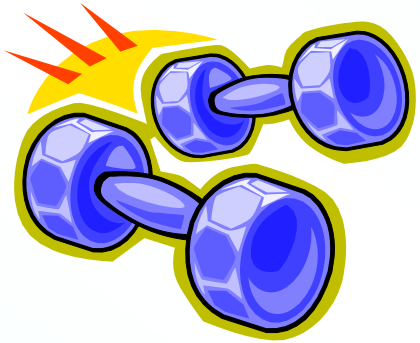
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2012 spring)

- Unsupervised learning
  (This course!)

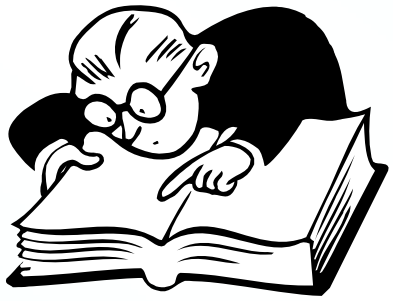


- Reinforcement learning



What Is Unsupervised Learning?

- You are given questions (input data) without answers (output data).
- The goal is to find an “interesting” structure in the data.



What Is Unsupervised Learning?

- The goal of unsupervised learning depends on the definition of "interestingness":
  - Dimensionality reduction
  - Clustering
  - Blind source separation
  - Outlier detection

Dimensionality Reduction

- Dimensionality reduction (embedding)
  - We are given high-dimensional data.
  - High-dimensional data is too complex to analyze: Even estimating the density is extremely difficult ("curse of dimensionality")
  - We want to have a low-dimensional expression of the data without losing intrinsic information.
  - Data visualization: Reduced data is less than equal to 3-dimensional.

Example 1

- “Swiss Roll”
- Data is 3D but it essentially lies on a 2D manifold.
- We want to “unfold” the roll.

2D

3D

3D

http://www.cs.toronto.edu/~roweis/lle/swissroll.html

Example 2

- Embedding face images into 2D space.
- Images of the same face from different angles (64x64=4096D).

http://isomap.stanford.edu/datasets.html

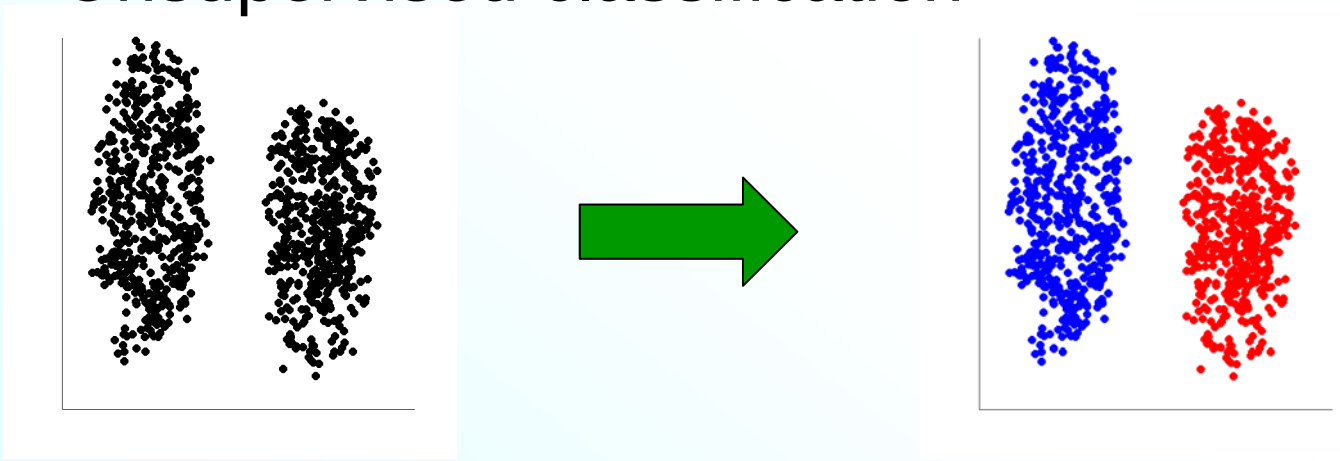
Example 3

- Embedding hand-written numbers into 2D space.
- Images of different “2”s (64x64=4096D).

http://isomap.stanford.edu/datasets.html

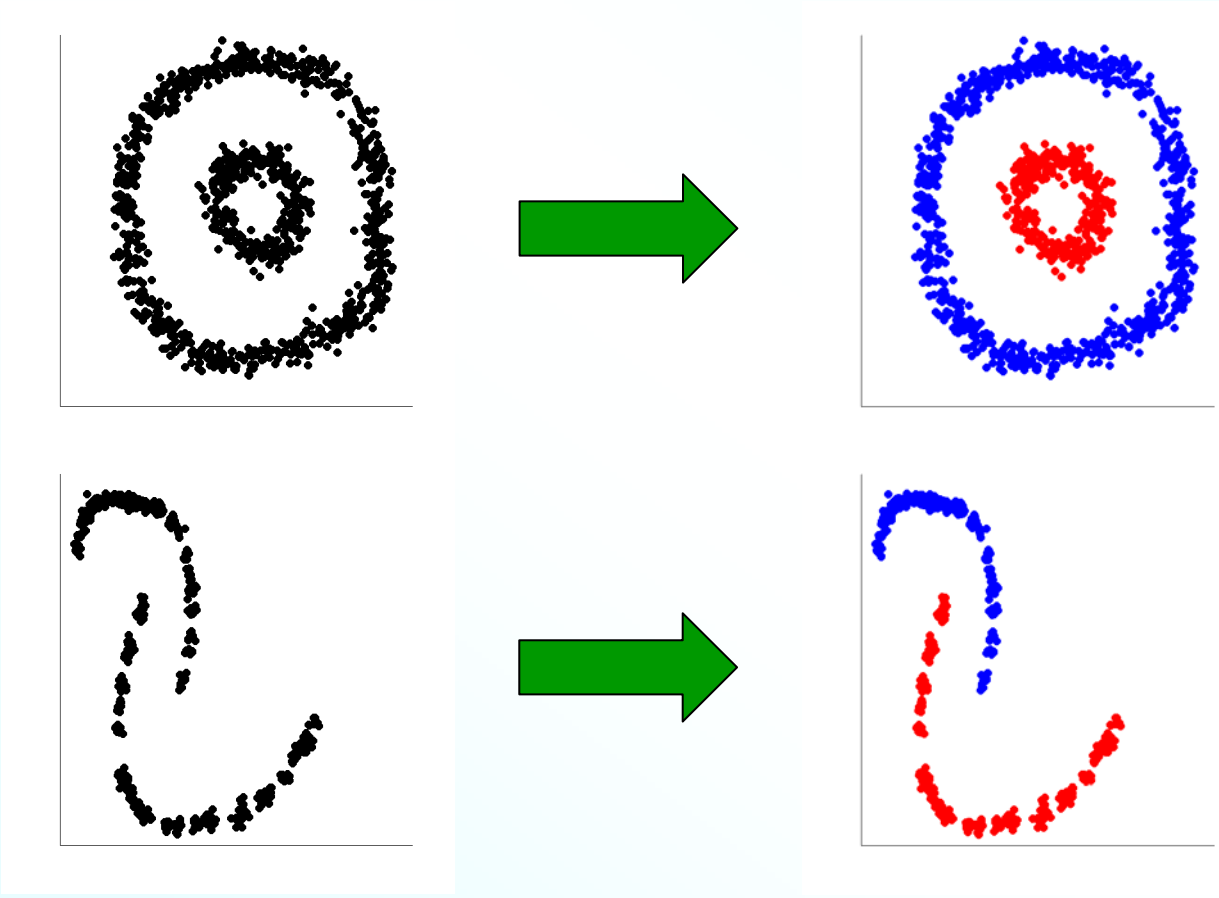
Data Clustering

- Clustering:
  - We want to divide the data into disjoint groups so that
    - Data in the same group have similar characteristics.
    - Data in different groups have different characteristics.
  - “Unsupervised classification”



Example 1

- “Connected” points seem to be in the same cluster, rather than “close” points.



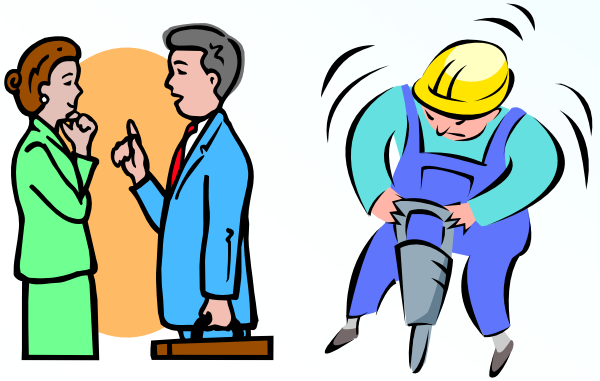
Example 2

- Image segmentation

http://www.cis.upenn.edu/~jshi/software/demo2.html

Blind Source Separation

- We can extract what a person is speaking in a noisy environment.

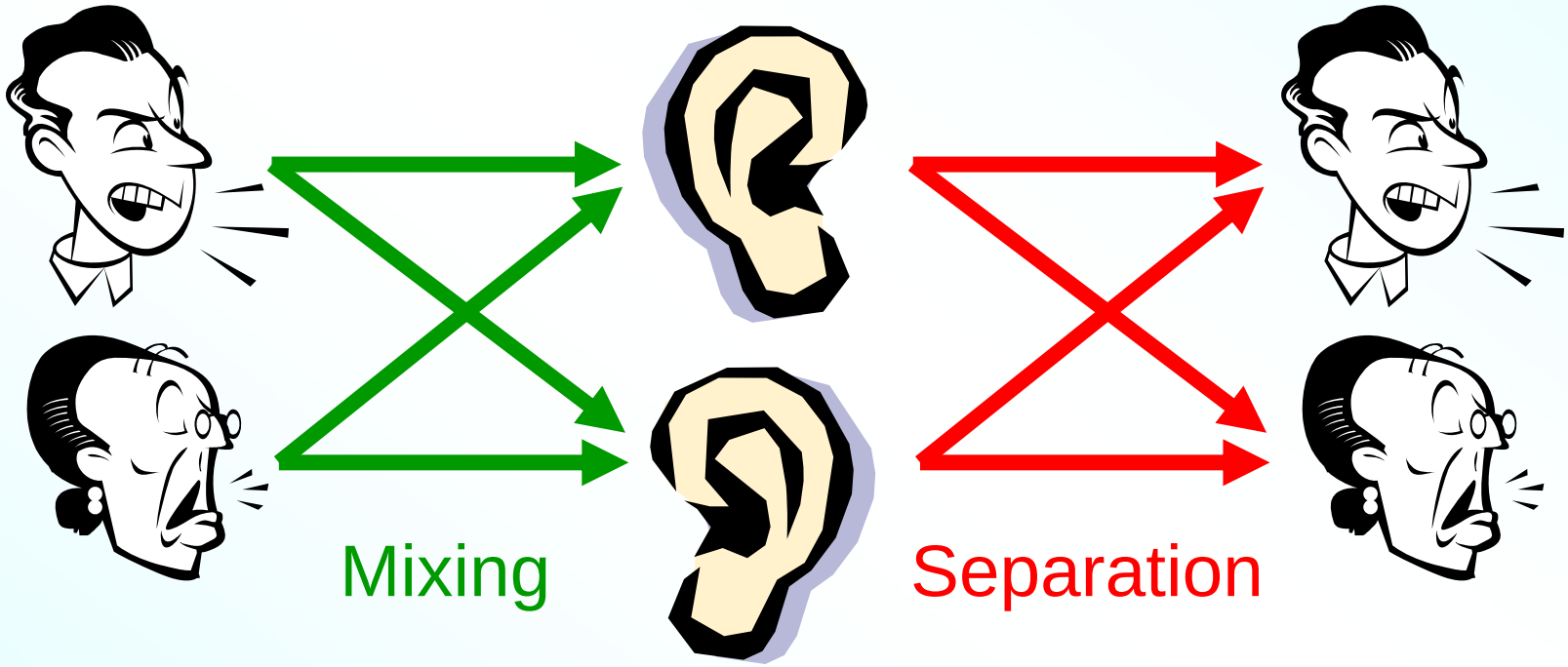


- Syotoku-taishi can distinguish 10 conversations?

From http://www.boj.or.jp/money/now/1man.htm

Blind Source Separation

- Cocktail-party problem:

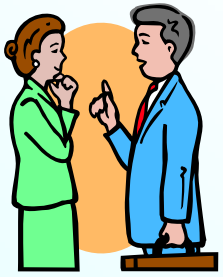


- We want to separate mixed signals into original ones.

Example

	Mixed signal	Separated signal 1	Separated signal 2
Conversation + Conversation			
Conversation + Instrument			

From http://www.brain.kyutech.ac.jp/~shiro/research/blindsep.html



Outlier Detection

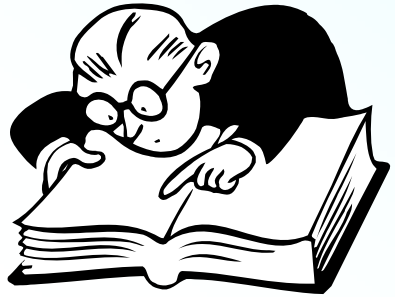
- When a new data sample is given, we want to know whether it is different from the samples collected so far.
- Also referred to as novelty detection or one-class classification.

B. Schölkopf et al (Neural Comp. 2001)

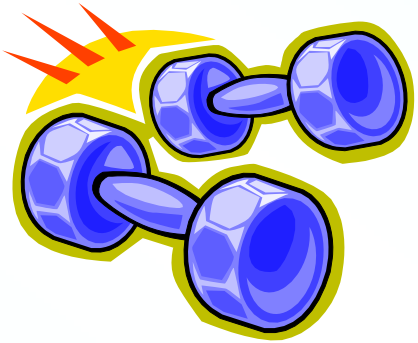
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2012 spring)

- Unsupervised learning
  (This course!)

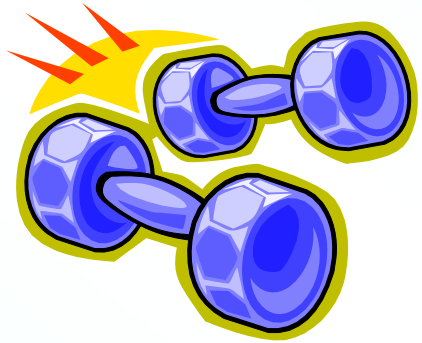


- Reinforcement learning



What Is Reinforcement Learning?

- The goal of reinforcement learning is same as supervised learning, i.e., to estimate an unknown underlying rule.
- However, different from supervised learning, we are not allowed to ask questions to the teacher.
- Instead, we can get rewards (reinforcement signals) for our estimated answer.



What Is Reinforcement Learning?

- Practically, we assume that the rule that maximizes the rewards is the underlying rule.
- Under this assumption, the rule is learned so that the rewards are maximized.
- Reinforcement learning can be regarded as being placed between supervised learning and unsupervised learning.

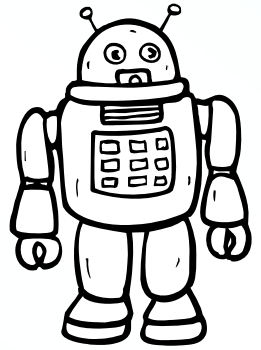
Example

- Learning stand-up motion:
  - The robot consists of 3 links connected by 2 joints.
  - Robot can control it's joint angles by itself.
  - The goal is to learn the control rule for stand up.
  - Control rule: mapping from inner states to control signal.

From IEICE Trans. Vol. J82-D-II, pp.2118-2131, 1999

Example

- Positive reward is given when stand-up motion has been succeeded; otherwise reward is zero.
- However, this does not work well in practice.
- Continuous reward is preferred.
  - Standing-up is equivalent to lifting the head.
  - The reward is designed so that the higher the head is, the more the reward is.



Example

- Before learning

From http://www.kawato.jst.go.jp/xmorimo/stup_movie.html

Example

- After 750 trials

From http://www.kawato.jst.go.jp/xmorimo/stup_movie.html

Example

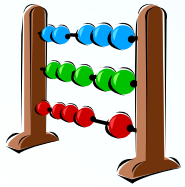
- After 920 trials

From http://www.kawato.jst.go.jp/xmorimo/stup_movie.html

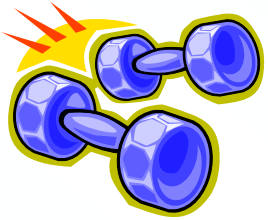
Conclusions

- There are 3 topics in learning research.
  - Understanding human brains
  - Developing learning machines
  - Mathematically clarifying mechanism of learning

- There are 3 types of learning.
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning



- Topics of unsupervised learning:
  - Dimensionality reduction
  - Data clustering
  - Blind source separation
  - Outlier/novelty detection



Homework

1. Prepare a high dimensional data set and explain the specification of the data.
   - Samples should be real-valued vectors!
   - Better if samples are from your own research area.
   - Better if dimensionality is not so small but not too large (say 10 to 100).
   - Better if the number of samples is large (say >>100).
   - If your data samples are not vectors (say sequences, images, graphs, texts, etc.), you may use some feature extraction software (developed in your research area) for converting them into vectors.
   - You do not have to finalize your data set now; it will be used in the final assignment. Use this opportunity to start searching good data sets.

Homework (cont.)

2. Prepare a computer environment where you can solve eigenproblems:

   e.g., MATLAB, octave, scilab, R, etc.

- Deadline: next class (or e-mail me before the class)