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）

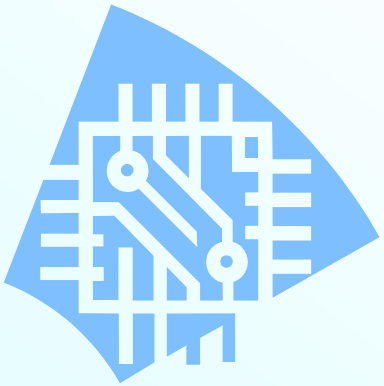
Contact:　W8E-505
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi

Contents of This Lecture (1)

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

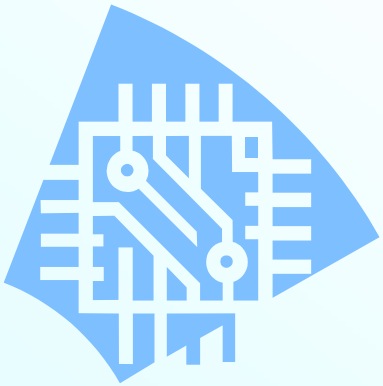
  The objective of this course is to introduce basic ideas and practical methods of discovering useful structure hidden in the data.

Statistical machine learning

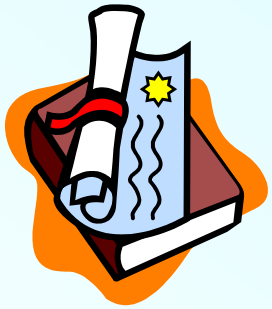


Contents of This Lecture (2)

- What you are expected to learn in this course:
  - How to use data analysis methods
  - Ideas behind the methods
  - Novel research topics in data analysis
  - Something useful in your own research/life



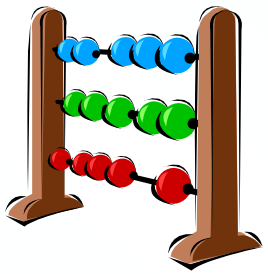
Grading



- Small reports
  - Several times
  - Deadline: 1 week
- Final reports
  - Topics not determined yet
  - Deadline: Mid August

Brief Overview of the Course (1)

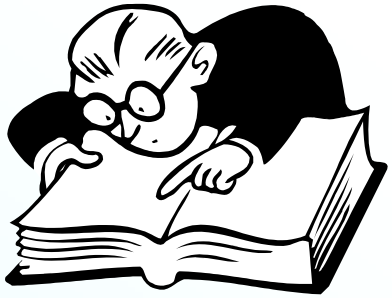
- 3 topics in the research of “learning”
  - Understanding human brains
  - Developing learning machines
  - Clarifying essence of learning mathematically

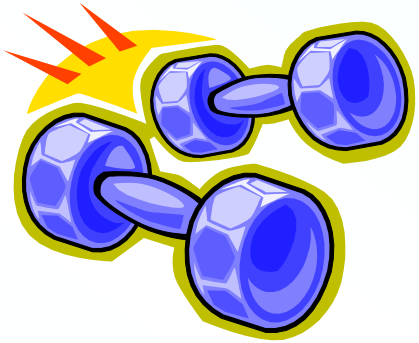


Brief Overview of the Course (2)

3 types of learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning





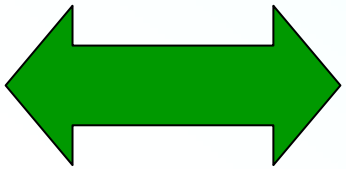
Brief Overview of the Course (3)

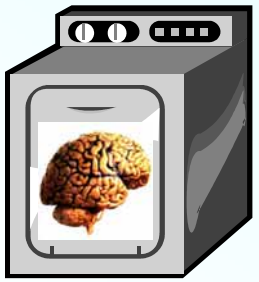
- Topics in unsupervised learning
  - Dimensionality reduction
  - Data clustering
  - Blind source separation
  - Outlier/novelty detection

Textbook

- Handouts are provided if necessary.
- Pointers to related articles will be provided.

3 Topics in Learning Research



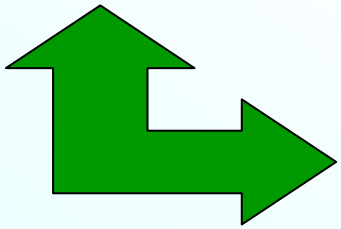


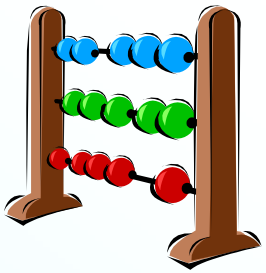
Understanding the brain

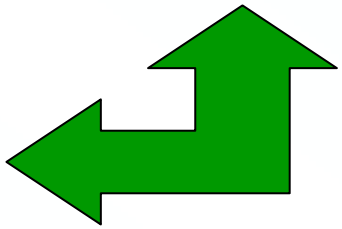
(physiology, psychology, neuroscience)

Developing learning machines

(computer and electronic engineering)







Clarifying learning mathematically

(computer and information science)

Understanding the Brain (1)

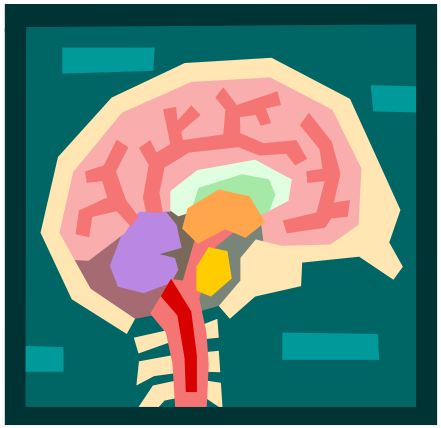
- Our brain consists of tens of billion neurons.
- Neurons are connected each other like a network.

Understanding the Brain (2)

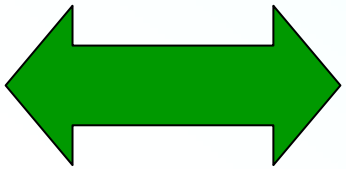
- Each neuron has dendrites and axons, and the axon connects to other neurons via synapses.
- Neurons receive signals from other neurons through dendrites and send signals through axons.

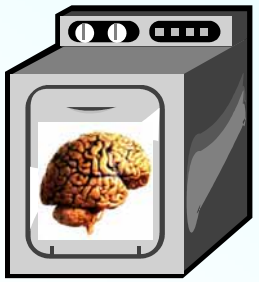
Understanding the Brain (3)

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



3 Topics in Learning Research



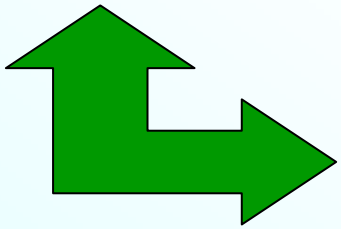


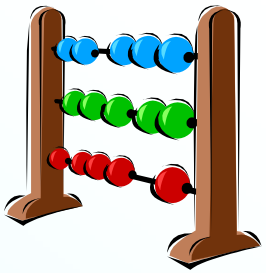
Understanding the brain

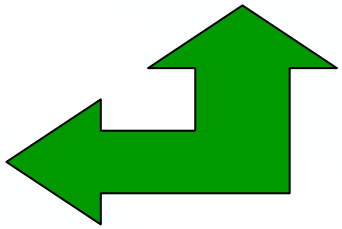
(physiology, psychology, neuroscience)

Developing learning machines

(computer and electronic engineering)





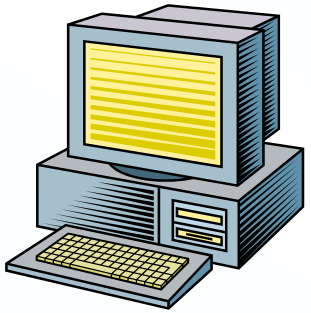


Clarifying learning mathematically

(computer and information science)

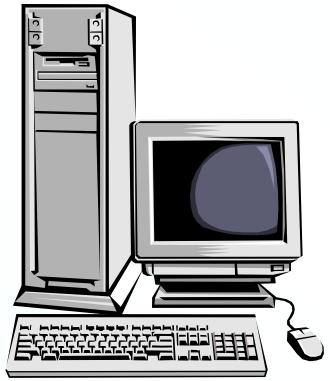
Developing Learning Machines (1)

- 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 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).

Developing Learning Machines (4)

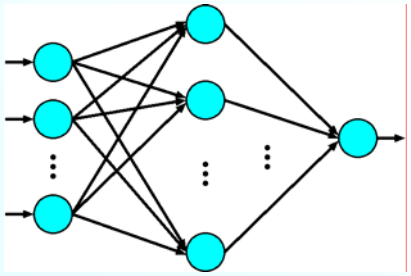
We want neurocomputes to equip the following functions:

- They are adaptable to new environments, i.e., we do not have to prescribe responses for all possible situations.
- They can process vague, noisy, and contradictory information.

Developing Learning Machines (5)

We want neurocomputes to equip the following functions:

- They consist of a number of artificial neurons and each neuron works independently.
- They are robust against noise, especially, faults of other neurons.
- They are small and efficient in electricity consumption.



Developing Learning Machines (6)

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

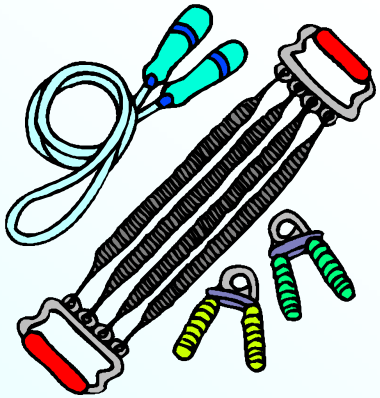
  Pulse Density Modulating Digital Neural Network System developed by University of Tsukuba

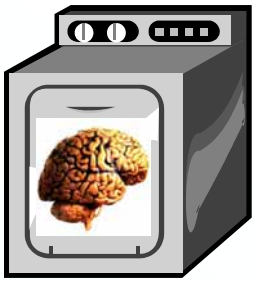
See http://www.viplab.is.tsukuba.ac.jp/~hirai/PDM/index.html

Developing Learning Machines (8)

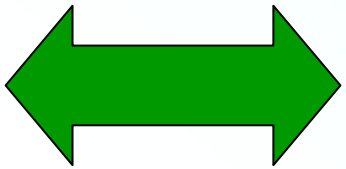
However, current neurocomputers have the following problems:

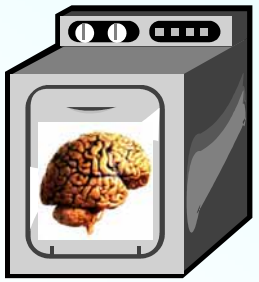
- The number of neurons are not so large.
- Size is big.
- It is not clear how to train the computer!!





3 Topics in Learning Research



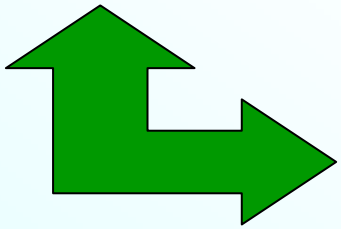


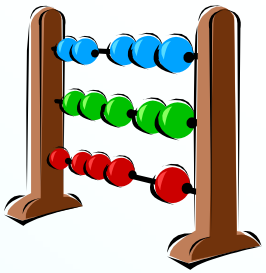
Understanding the brain

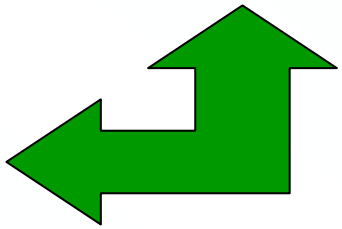
(physiology, psychology, neuroscience)

Developing learning machines

(computer and electronic engineering)





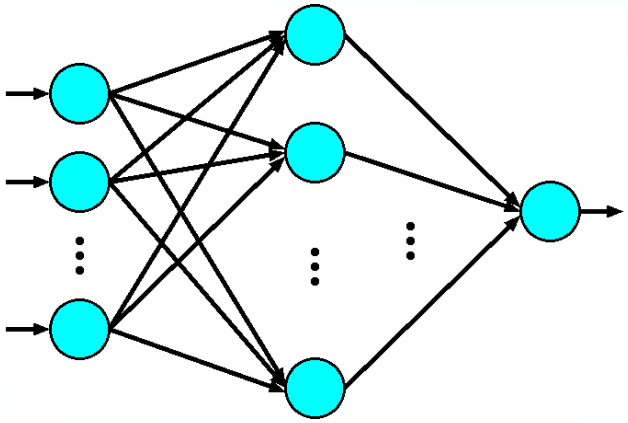


Clarifying learning mathematically

(computer and information science)

Clarifying Learning Mathematically (1)

- In order to understand our brains and develop neurocomputers, we have to clarify how information is processed with 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' brain.

Clarifying Learning Mathematically (3)

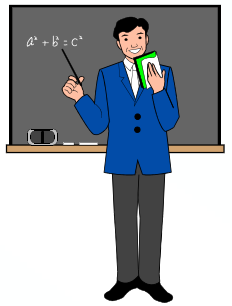
Mathematical tools for clarifying essence of learning

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

A Little Break…

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

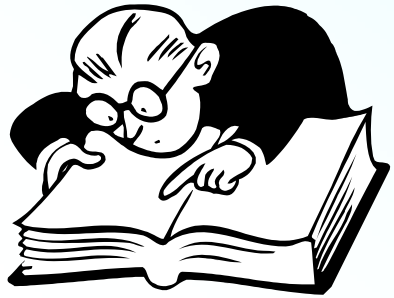
  “Theories of learning”



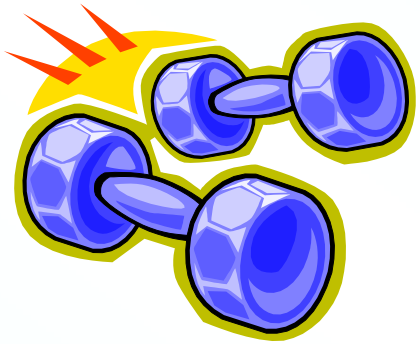
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2006 spring)

- Unsupervised learning

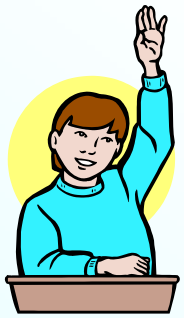


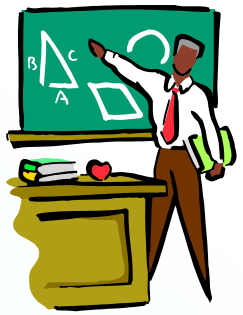
- Reinforcement learning



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 the training examples.
- If the underlying rule can be successfully estimated, we can answer to the questions that we have never taught.
- Such an ability is called the generalization capability.

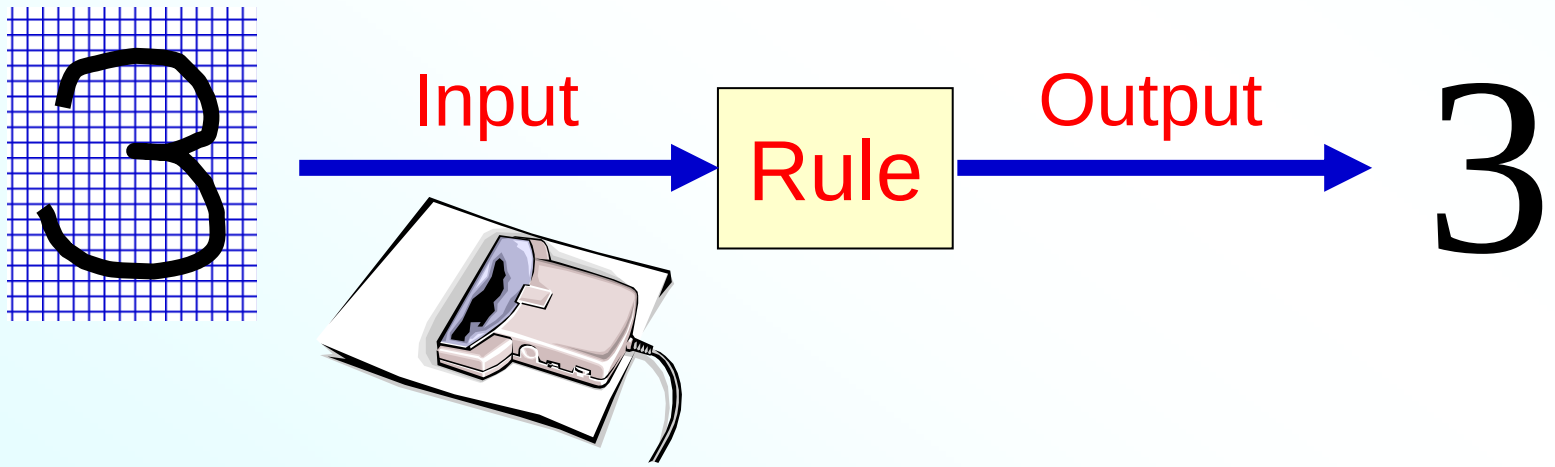
Example 1

- Hand-written number recognition

  We want to recognize the scanned hand-written characters.

Example 1

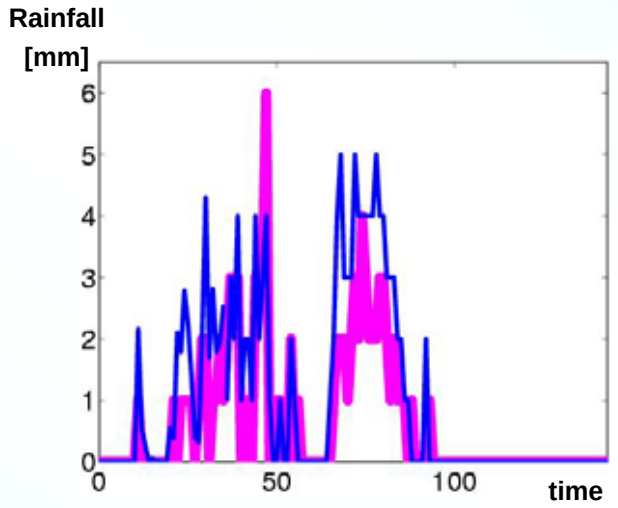
- Training examples consist of
  { (hand-written number, its recognition result) }.
- If underlying input-output rule is successfully learned, unlearned hand-written numbers can be recognized.



Example 2

- Rainfall Estimation

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

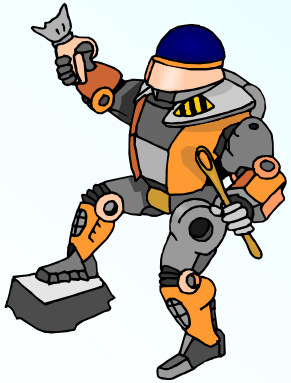


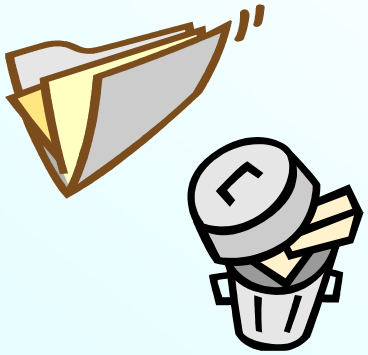
Example 2

- 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

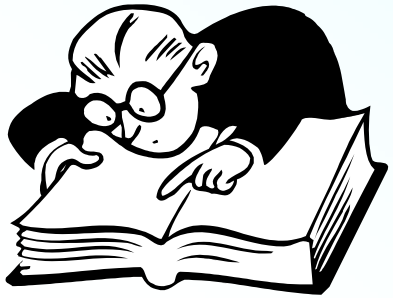




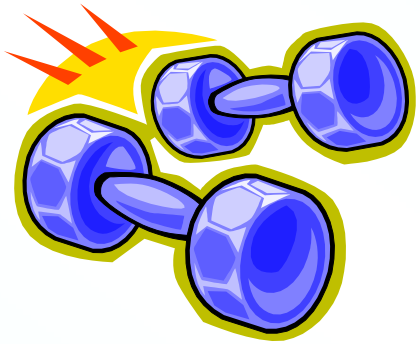
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2006 spring)

- Unsupervised learning
  (This course!)

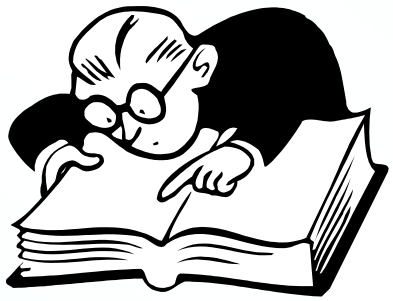


- Reinforcement learning



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

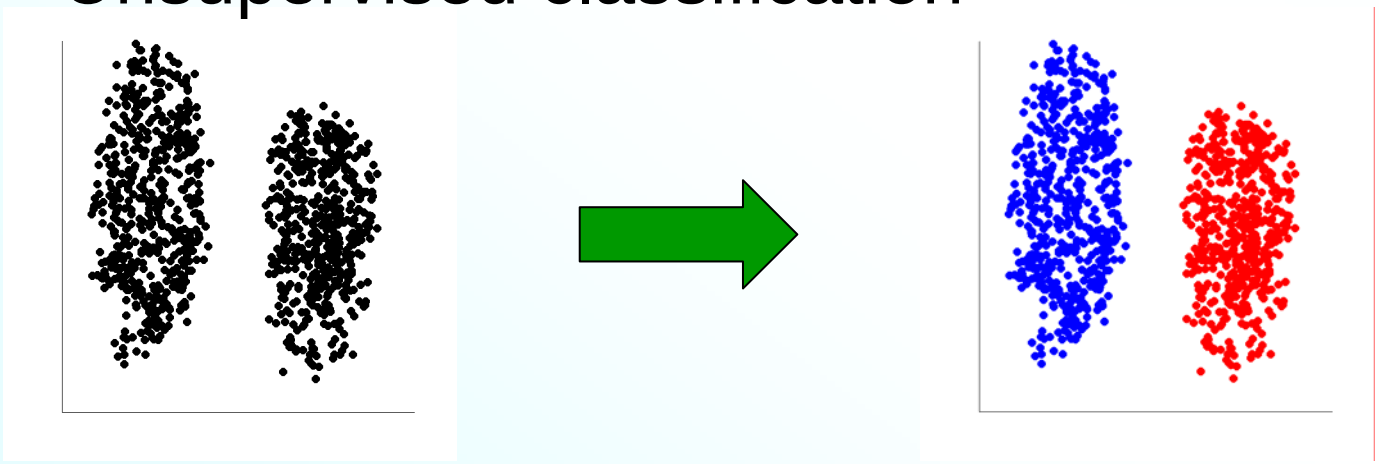
- Dimensionality reduction (Embedding)
  - We are given high-dimensional data.
  - High-dimensional data is too complex to analyze: Even estimating the density is extremely difficult ("curse of dimensionality")
  - We want to have a low-dimensional expression of the data without losing intrinsic information.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

Example

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

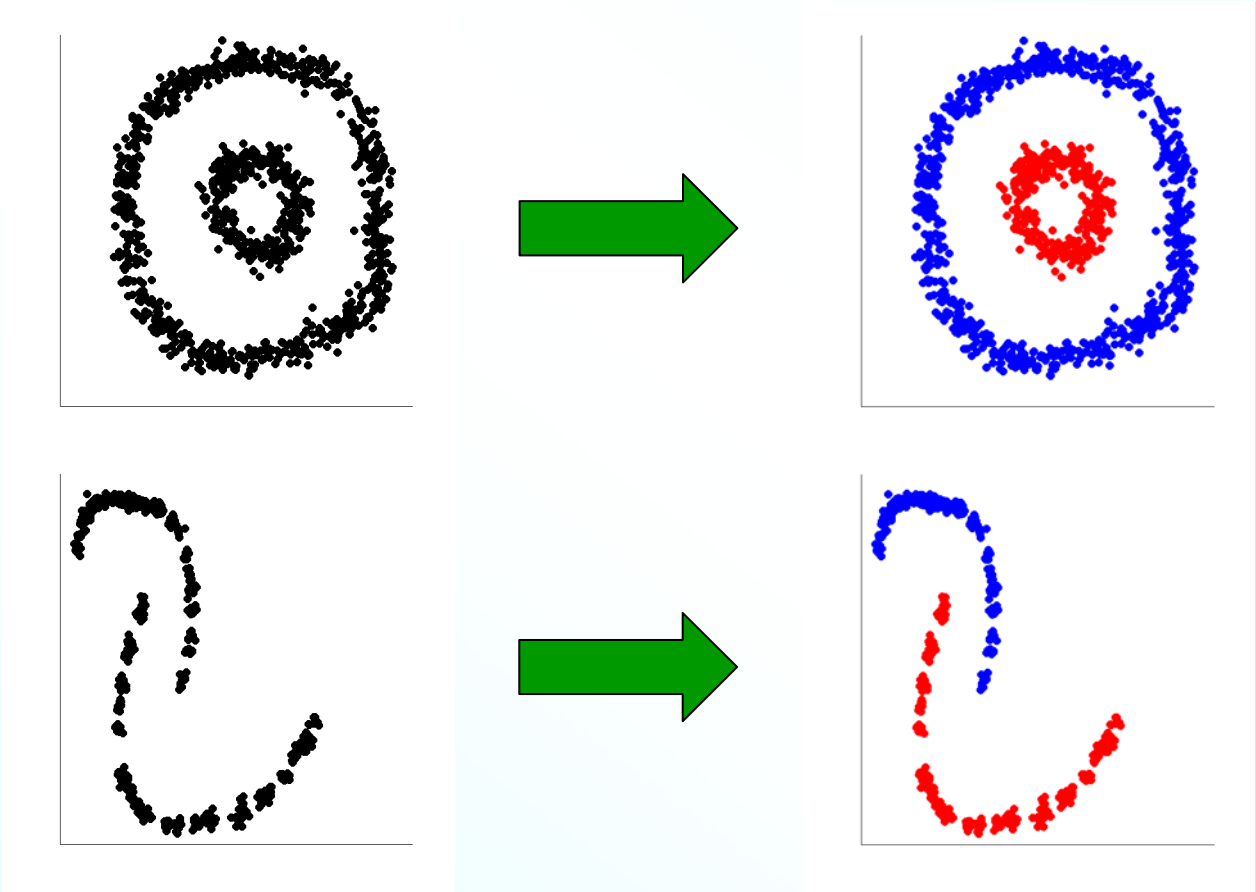
Data Clustering

- 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”



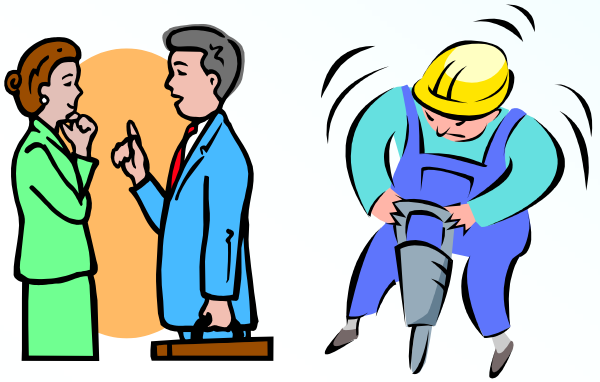
Example 1

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



Blind Source Separation

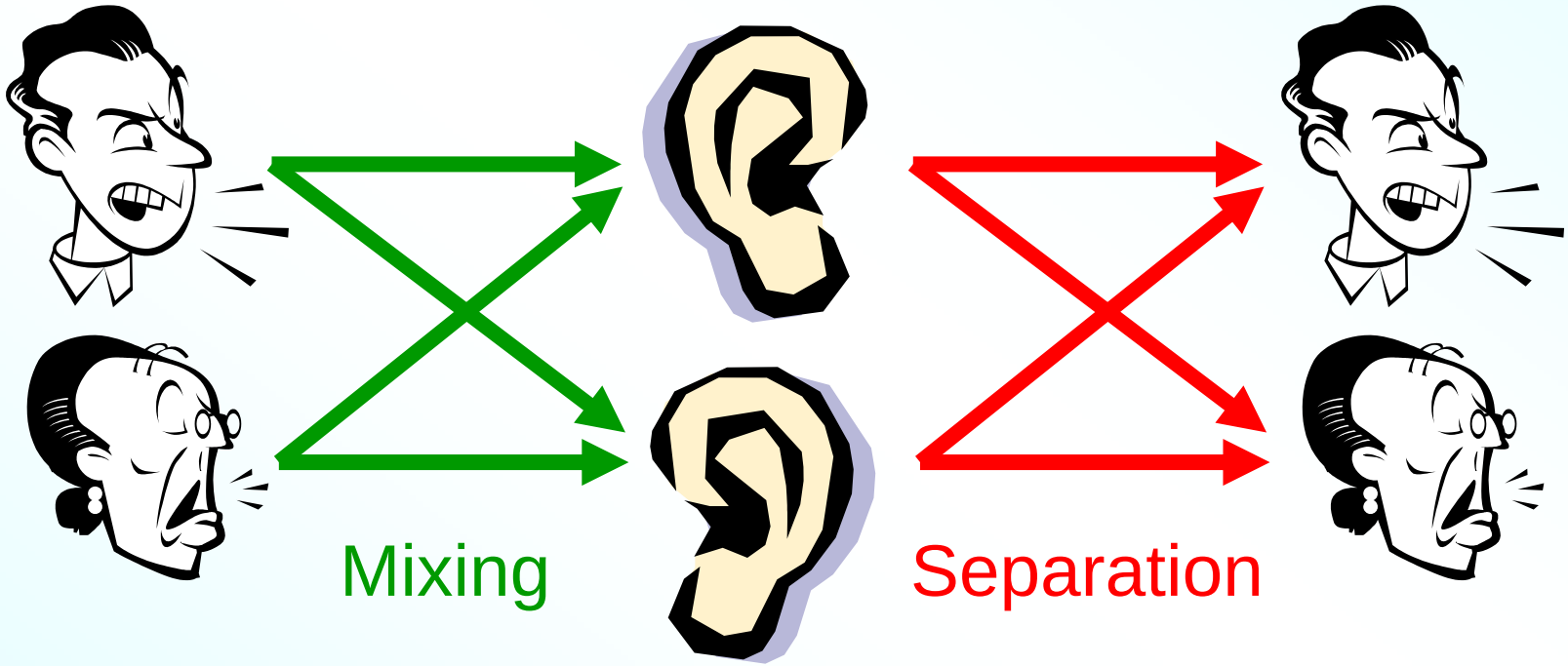
- We can extract what a person is speaking in a noisy environment.



- Syotoku-taishi can distinguish 10 conversations?

Blind Source Separation

- Cocktail-party problem



- We want to separate mixed signals into original ones.

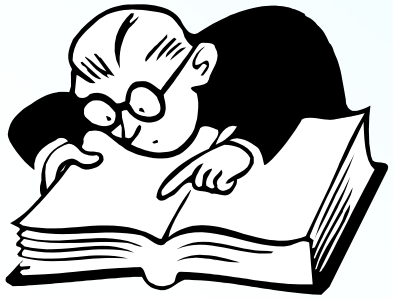
Outlier Detection

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

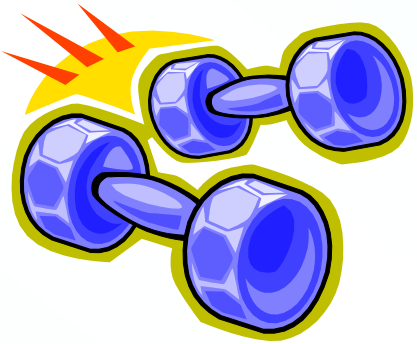
Three Types of Learning

- Supervised learning
  ("Pattern information processing", 2006 spring)

- Unsupervised learning
  (This course!)

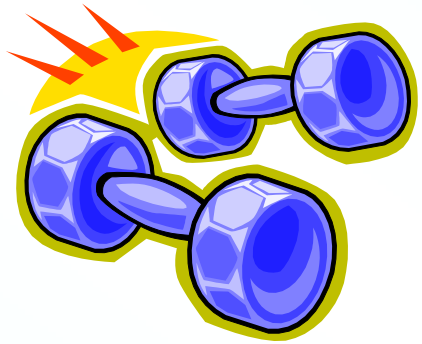


- Reinforcement learning
  (Prof. Shigenobu Kobayashi,
  Dept. of Computational intelligence
  and Systems Science)



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?

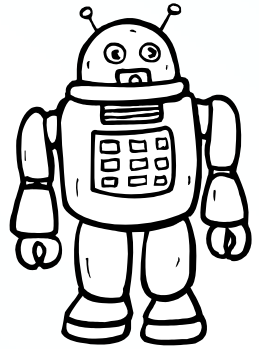
- Practically, we assume that the rule that maximizes the rewards is the underlying rule.
- Under this assumption, the rule is learned so that the rewards is maximized.
- Reinforcement learning can be regarded as being placed between supervised learning and unsupervised learning.

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.

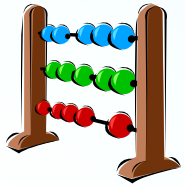
Example

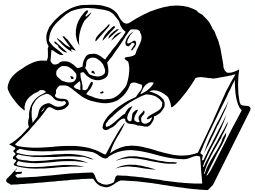
- Essentially, reward is given when stand-up motion has been succeeded, otherwise reward is zero.
- However, this does not work well in practice.
- Continuous reward is preferred.
- For example, stand-up is equivalent to lifting the head, the reward is designed such that the higher the head is, the more the reward is.

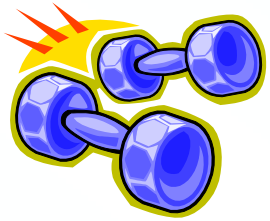


Conclusions

- There are 3 topics in learning research.
  - Understanding human brains
  - Developing learning machines
  - Clarifying essence of learning mathematically
- 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







Matrix Formulas

- In this course, we use some formulas of matrix calculations.
- “Matrix Cookbook”:

  http://2302.dk/uni/matrixcookbook.html

Homework

- Find some high dimensional data sets (either from your research domain or from the web), and explain the specification of the data.
- Prepare a computer environment where you can solve eigenproblems for the next homework

  e.g., MATLAB, octave, scilab, R…
- Deadline: April, 19 (Tue)

Important Notice!

- There are two invited lectures today at W8E-10F
- 14:30-15:30
  Dr. Julian Laub (Fraunhofer FIRST, Germany)
  Analyzing Non-Euclidean Pairwise Data
  (related to data clustering)
- 15:30-16:30
  Dr. Simone Fiori (Univ. Perugia, Italy)
  Formulation and Integration of Learning Differential Equations on the Stiefel Manifold
  (related to blind source separation)