

<https://shorturl.at/hG7iP>

KYOTO UNIVERSITY

# Statistical Learning Theory

## - Introduction -

Hisashi Kashima / Koh Takeuchi / Kyohei Atarashi

GRADUATE SCHOOL OF INFORMATICS

# Statistical learning theory:

## Foundations of recent data analysis technologies

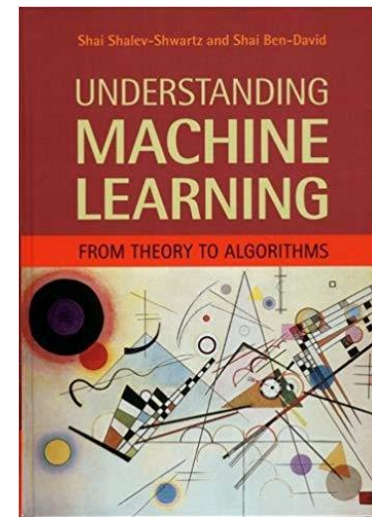
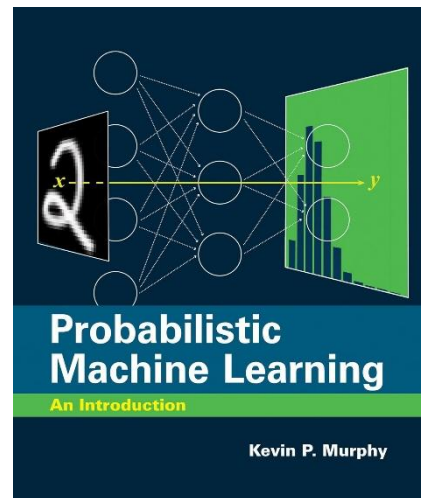
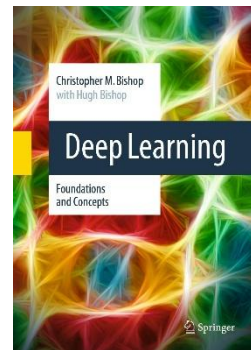
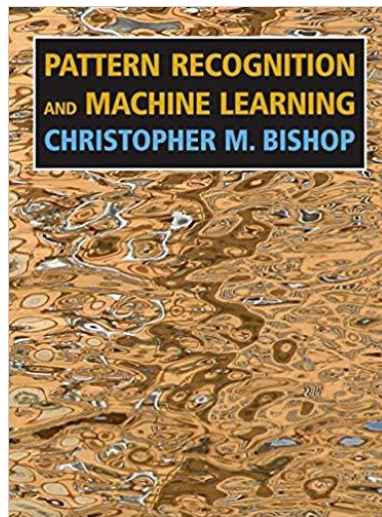
---

- This lecture will cover:
  - Basic ideas, problems, solutions, and applications of statistical machine learning
    - Supervised & unsupervised learning
    - Models & algorithms: linear regression, SVM, neural nets, ...
  - Statistical learning theory
    - Theoretical foundation of statistical machine learning
  - Hands-on practice (using Python)
- Advanced topics: Graph neural networks, attentions, self-supervised learning, ...

# Textbooks?:

Most of the topics can be found in...

- Pattern recognition and machine learning / Bishop
  - Or, Deep Learning
- Probabilistic Machine Learning / Murphy
- Understanding machine learning / Shalev-Shwartz & Ben-David

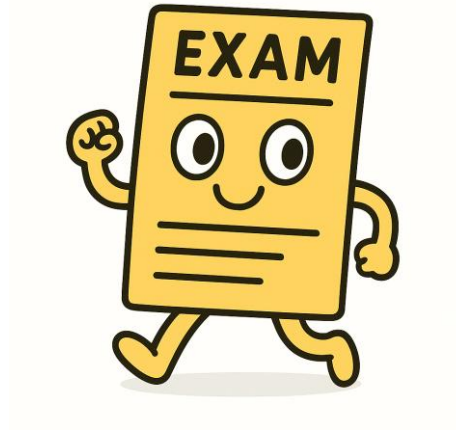


# Evaluations:

## Final Exam is All You Need

---

- Evaluation is solely based on the final exam
  - The examination is a standard written exam
  - You are NOT allowed to use any reference materials
- No attendance checks. No homework.



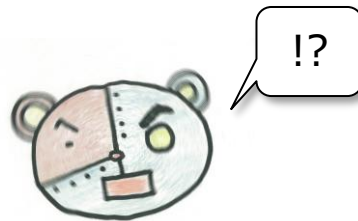
# Contents:

## Basic ideas of machine learning and applications

---

1. What is machine learning?
2. Learning machines
3. Machine learning applications:
  1. Applications of supervised learning: text classification
  2. Applications of unsupervised learning: Anomaly detection

# What is machine learning?



# “The third A.I. boom”:

## Machine learning is a core technology behind the boom

- You can see many successes of “Artificial Intelligence”:
  - Q.A. machine beating quiz champions and Go program surpassing top players
  - Protein folding, that was thought to be unsolvable, was solved
  - Are large language models (LLMs) the realization of general-purpose artificial intelligence?
- Current A.I. boom owes machine learning
  - Especially, deep learning



# What is machine learning? :

## A branch of artificial intelligence

---

- Originally started as a branch of artificial intelligence
  - has its more-than-50-years history
  - Computer programs that “learns” from experience
  - Based on logical inference
- Pioneers who invented the computer also already dreamed of realizing artificial intelligence



# What is machine learning? :

## A data analytics technology

---

- Rise of “statistical” machine learning
  - Successes in bioinformatics, natural language processing, and other business areas
- Technically, it can be considered as a data analysis technology
  - Buzzwords: “big data” and “data scientist”
    - Data scientist is “the sexiest job in the 21st century” (?)
- Led the success of deep learning
  - The 3rd AI boom

# What can machine learning do?: Prediction, discovery, ... and generation

---

## 1. Prediction

- “What will happen in future data?”
- Given past data, predict about future data

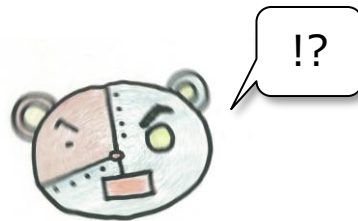
## 2. Discovery

- “What is happening in data in hand?”
- Given past data, find insights in them

## 3. Data generation

- “Generate new data satisfying certain properties”
- Given past data, generate similar data

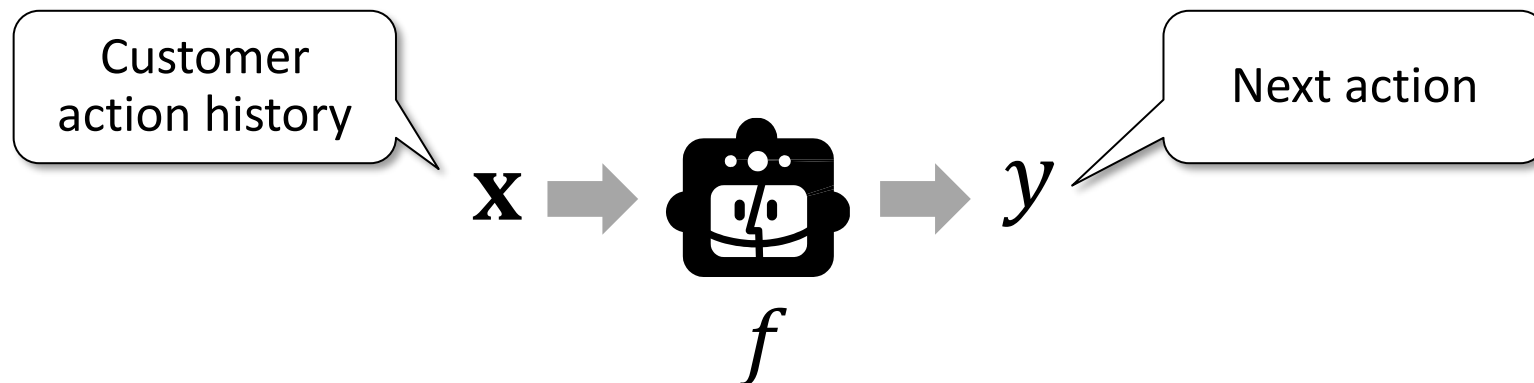
# Learning Machines



# Prediction machine:

## A function from a vector to a scalar

- We model our “prediction machine” as a math function:
  - Relationship of input and output  $f: \mathbf{x} \rightarrow y$ 
    - Input  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top \in \mathbb{R}^D$  is a  $D$ -dimensional vector
    - Output  $y$  is one dimensional
  - Regression: real-valued output  $y \in \mathbb{R}$
  - Classification: discrete output  $y \in \{C_1, C_2, \dots, C_M\}$

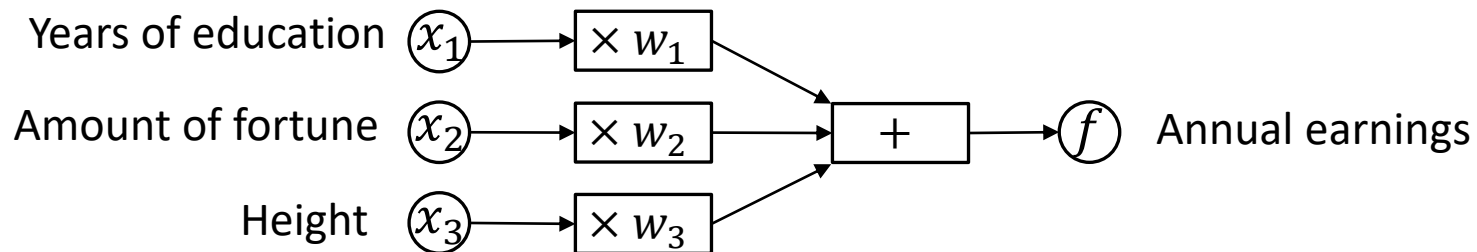


# A model for regression: Linear regression model

- Model  $f$  takes an input  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$  and outputs a real value

$$f(\mathbf{x}) = w_1x_1 + w_2x_2 + \dots + w_Dx_D$$

- Model parameter  $\mathbf{w} = (w_1, w_2, \dots, w_D)^\top \in \mathbb{R}^D$



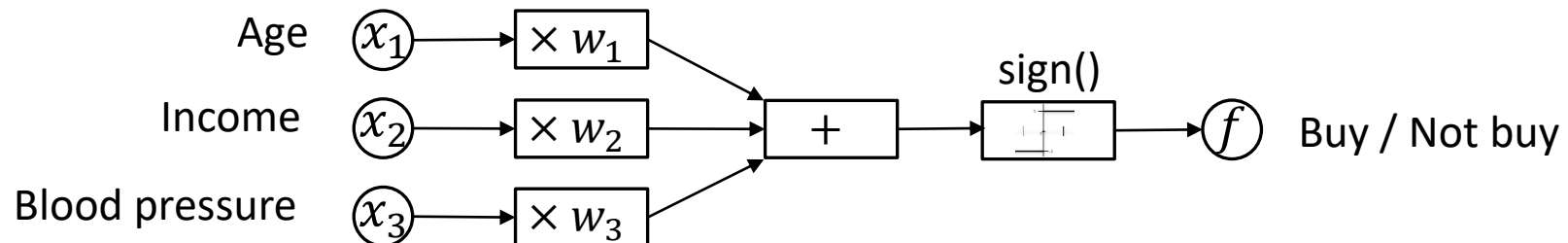
# A model for classification:

## Linear classification model

- Model  $f$  takes an input  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$  and outputs a value from  $\{+1, -1\}$  (class label)

$$f(\mathbf{x}) = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_Dx_D)$$

- Model parameter  $\mathbf{w} = (w_1, w_2, \dots, w_D)^\top \in \mathbb{R}^D$  :
  - $w_d$  : contribution of  $x_d$  to the output (if  $w_d > 0$ ,  $x_d > 0$  contributes to  $+1$ ,  $x_d < 0$  contributes to  $-1$ )

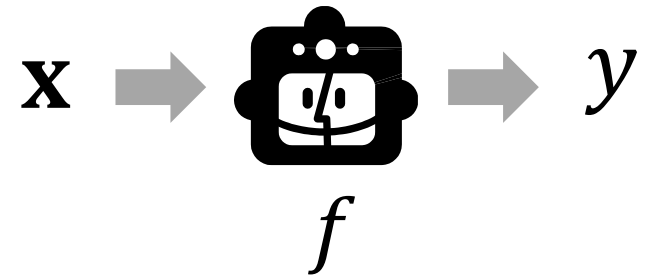


# Formulations of machine learning problems:

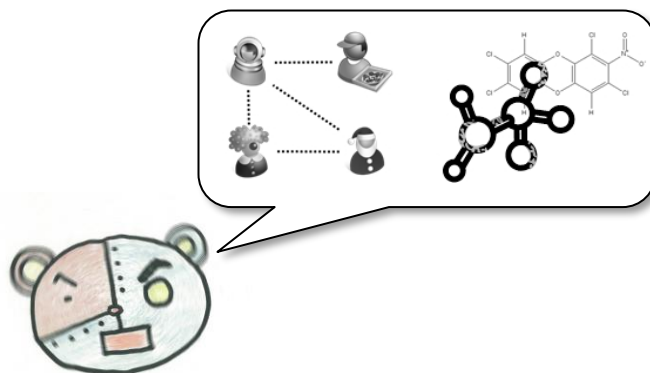
## Supervised learning and unsupervised learning

---

- What we want is the function  $f$  (or its parameters  $\mathbf{w}$ )
  - We estimate  $f$  (or  $\mathbf{w}$ ) from data
- Two learning problem settings: supervised and unsupervised
  - Supervised learning: input-output pairs are given
    - $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\} : N$  pairs
  - Unsupervised learning: only inputs are given
    - $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\} : N$  inputs



# Machine learning applications

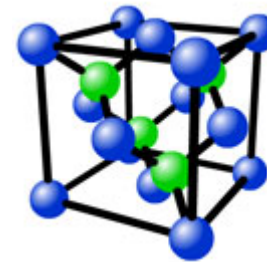


# Growing ML applications:

## Emerging applications from IT areas to non-IT areas

---

- Recent advances in ML offer:
  - Methodologies to handle uncertain and enormous data
  - Black-box tools
- Not limited to IT-related areas, ML is wide-spreading over non-IT areas
  - Healthcare, airline, automobile, material science, education,  
...



# Various applications of machine learning: From on-line shopping to system monitoring

## ■ Marketing

- Recommendation
- Sentiment analysis
- Web ads optimization



## ■ Finance

- Credit risk estimation
- Fraud detection



## ■ Science

- Biology
- Material science



## ■ Web

- Search
- Spam filtering
- Social media



## ■ Healthcare

- Medical diagnosis



## ■ Multimedia

- Image/voice understanding

## ■ System monitoring

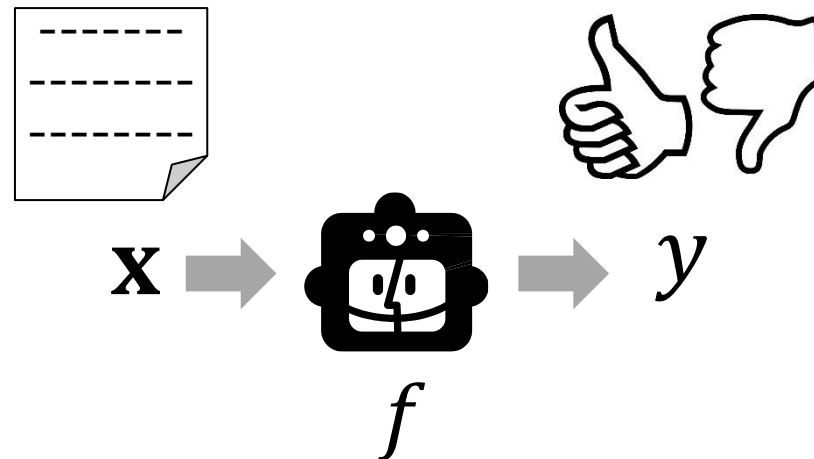
- Fault detection



# An application of supervised classification learning: Sentiment analysis

---




- Judge if a document ( $\mathbf{x}$ ) is positive or not ( $y \in \{+1, -1\}$ ) toward a particular product or service
- For example, we want to know reputation of our new product  $S$ , and gain marketing insights
- Collect tweets/posts by searching the word “ $S$ ”, and analyze them



# An application of supervised learning:

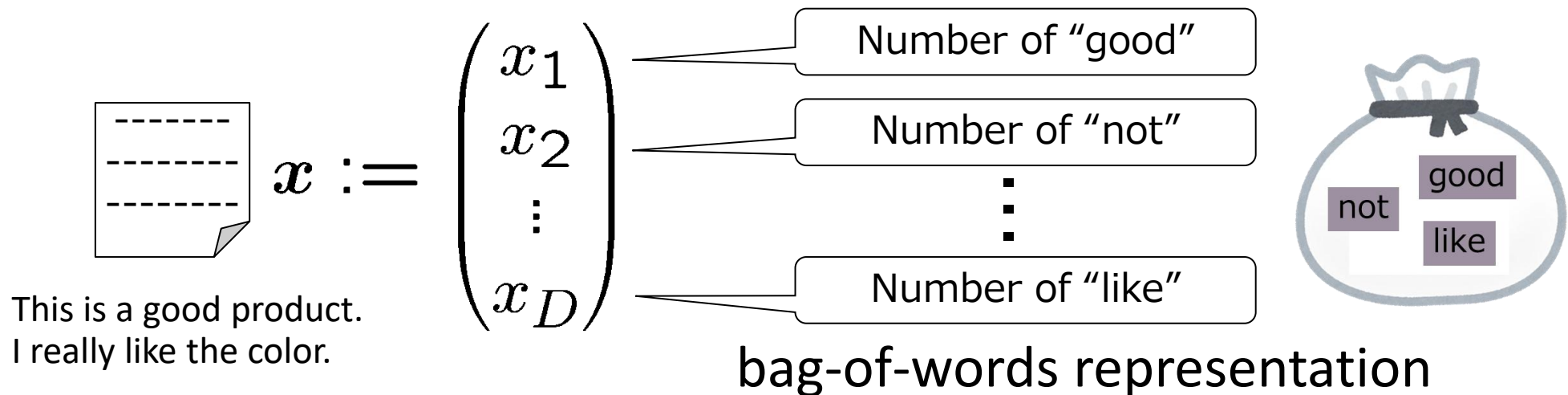
## Some hand labeling followed by supervised learning

---

- We need a training dataset
- First, give labels to some of the collected documents
  - 10,000 posts hit the word “S”
  - Manually read 300 of them and give sentiment labels
    - “I used S, and found it not bad.” → 
    - “I gave up S. The power was not on.” → 
    - “I like S.” → 
- Use the collected 300 labels to train a predictor. Then apply the predictor to the rest 9,700 posts

# How to represent a document as a vector: bag-of-words representation

- Represent a post  $\mathbf{x}$  using words appearing in it



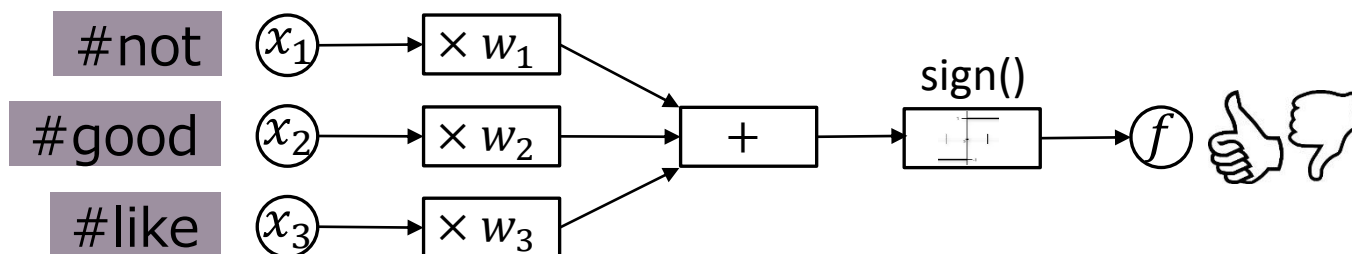
- Usually, the feature vector is very high-dimensional and sparse
- Note: design of the feature vector is left to users

# A simple model for sentiment analysis: Linear binary classification model

- Model  $f$  takes a BoW vector  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$  and outputs a sentiment label from  $\{+1, -1\}$ :

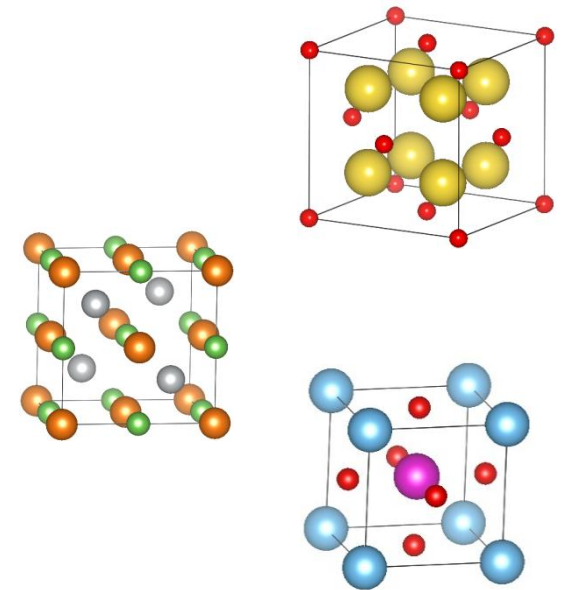
$$f(\mathbf{x}) = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_Dx_D)$$

- Model parameter  $\mathbf{w} = (w_1, w_2, \dots, w_D)^\top \in \mathbb{R}^D$  :
  - $w_d$  : contribution of the  $d$ -th word (e.g. “good”) to the sentiment label



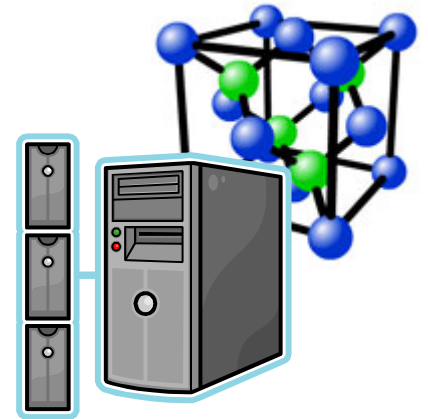
# An application of supervised *regression* learning: Discovering new materials

- Material science aims at discovering and designing new materials with desired properties
  - Volume, density, elastic coefficient, thermal conductivity, ...
- Traditional approach (try-and-error):
  1. Determine chemical structure
  2. Synthesize the chemical compounds
  3. Measure their physical properties
  4. If not satisfied, return to 1



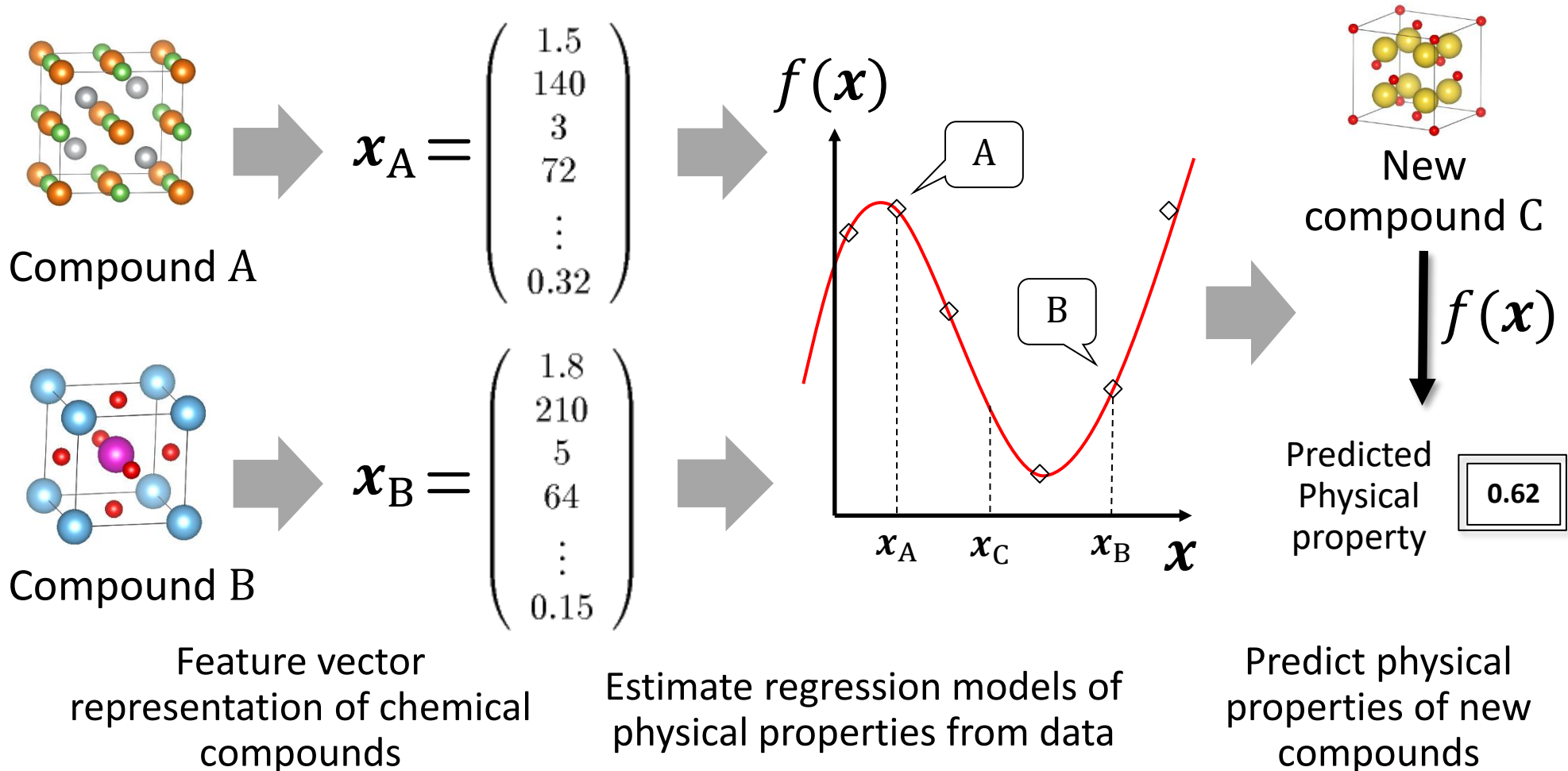
# Computational approach to material discovery: Still needs high computational costs

- Computational approach: First-principle calculations based on quantum physics to run simulation to estimate physical properties
- First-order principle calculation still requires high computational costs
  - Proportional to the cubic number of atoms
  - Sometimes more than a week or a month...



# Data driven approach to material discovery: Regression to predict physical properties

- Predict the result of first-order principle calculation from data



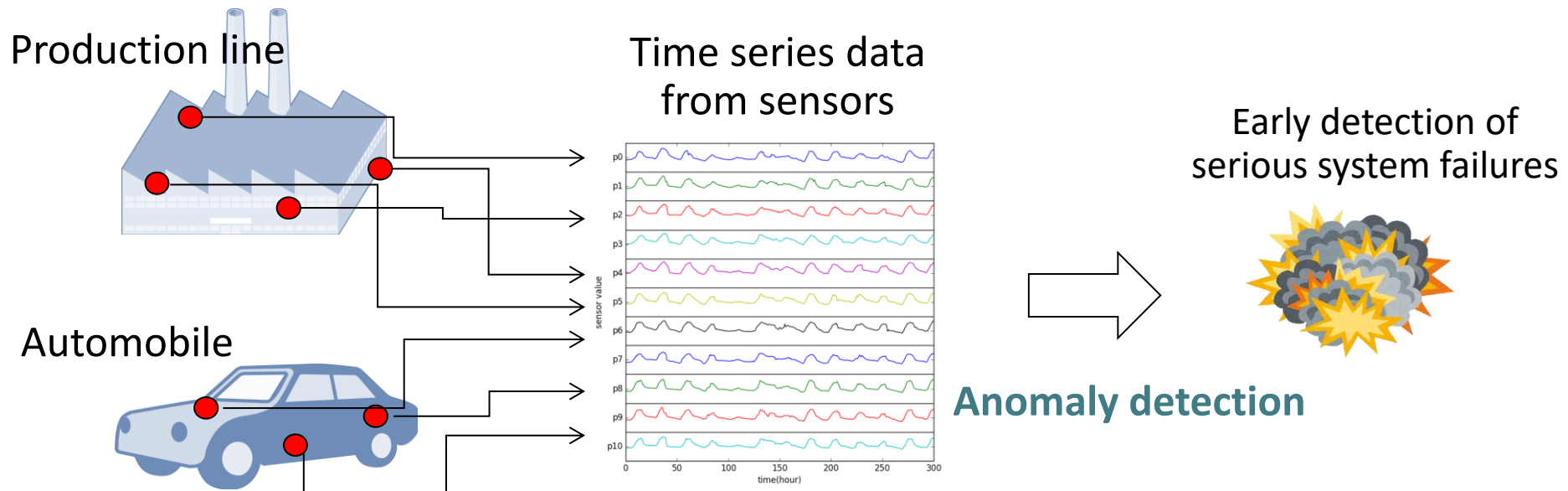
# Anomaly detection



# Anomaly detection:

## Early warning for system failures reduces costs

- A failure of a large system can cause a huge loss
  - Breakdown of production lines in a factory, infection of computer virus/intrusion to computer systems, credit card fraud, terrorism, ...
- Modern systems have many sensors to collect data
- Early detection of failures from data collected from sensors



# Anomaly detection techniques:

## Find “abnormal” behaviors in data

---

- We want to find precursors of failures in data
  - Assumption: Precursors of failures are hiding in data
- Anomaly: An “abnormal” patterns appearing in data
  - In a broad sense, state changes are also included:  
appearance of news topics, configuration changes, ...
- Anomaly detection techniques find such patterns from data and report them to system administrators

# Difficulty in anomaly detection:

## Failures are rare events

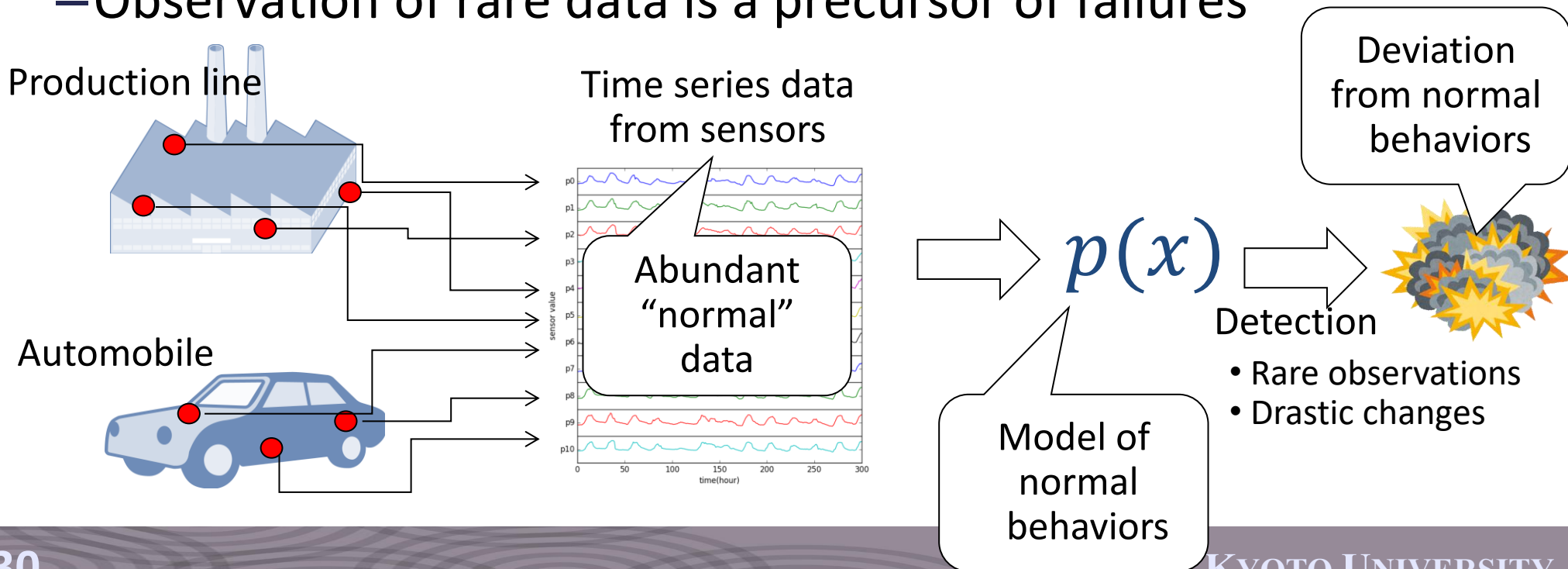
---

- If target failures are known ones and many data are available, they are detected by using supervised learning:
  1. Construct a predictive model from past failure data
  2. Apply the model to system monitoring
- However, serious failures are usually rare, and often new ones  
→ (Almost) no past data are available
- Supervised learning is not applicable in most cases

# An alternative idea:

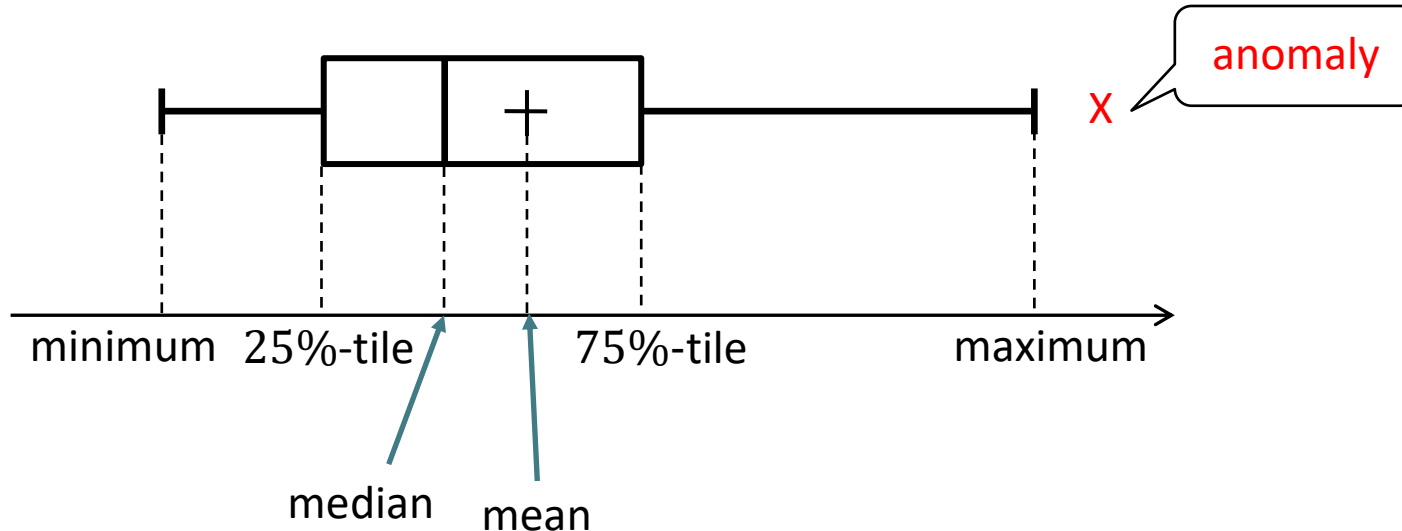
## Model the normal times, detect deviations from them

- Difficult to model anomalies → Model normal times
  - Data at normal times are abundant
- Report “strange” data according to the normal time model
  - Observation of rare data is a precursor of failures



# A simple unsupervised approach: Anomaly detection using thresholds

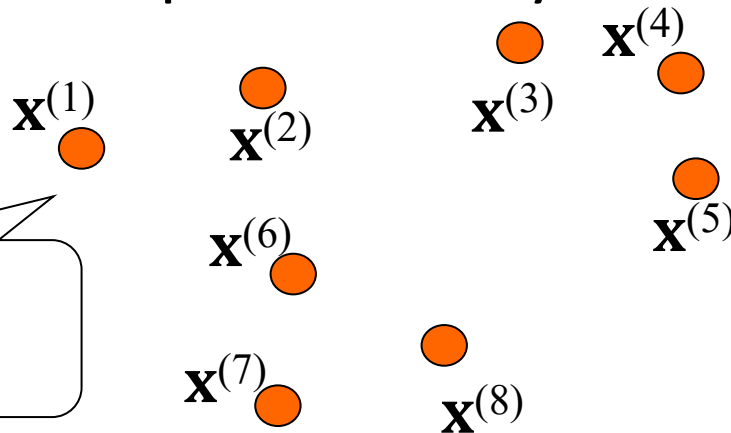
- Suppose a 1-dimensional case (e.g. temperature)
- Find the value range of the normal data (e.g. 20-50 °C)
- Detect values deviates from the range, and report them as anomalies (e.g. 80°C is not in the normal range)



Box plot showing the five-number summary, with the mean indicated by X

# Clustering for high-dimensional anomaly detection: Model the normal times by grouping the data

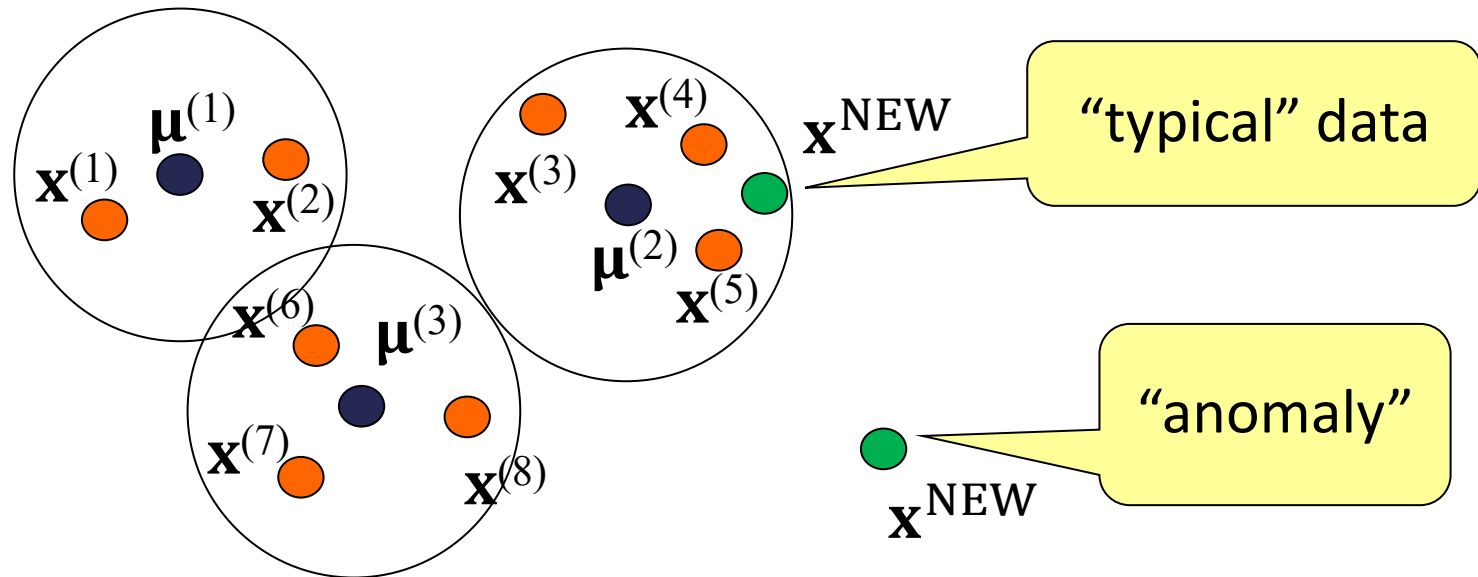
- More complex cases:
  - Multi-dimensional data
  - Several operation modes in the systems
- Divide normal time data  $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$  into  $K$  groups
  - Groups are represented by centers  $\{\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \dots, \boldsymbol{\mu}^{(N)}\}$



Temperature,  
Flow speed,  
Acceleration, ...

# Clustering for high-dimensional anomaly detection: Find anomalies not belonging to the groups

- Divide normal data  $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$  into  $K$  groups
  - Groups are represented by centers  $\{\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \dots, \boldsymbol{\mu}^{(K)}\}$
- Data  $\mathbf{x}$  is an “anomaly” if it lies far from all of the centers  
= system failures, illegal operations, instrument faults



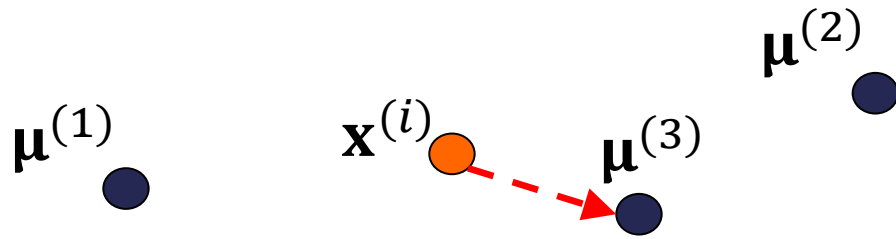
# $K$ -means algorithm:

## Iterative refinement of groups

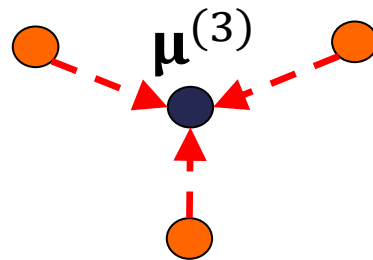
---

- Repeat until convergence:

1. Assign each data  $\mathbf{x}^{(i)}$  to its nearest center  $\boldsymbol{\mu}^{(k)}$



2. Update each center to the mean of the assigned data



# Anomaly detection in time series:

## On-line anomaly detection

---

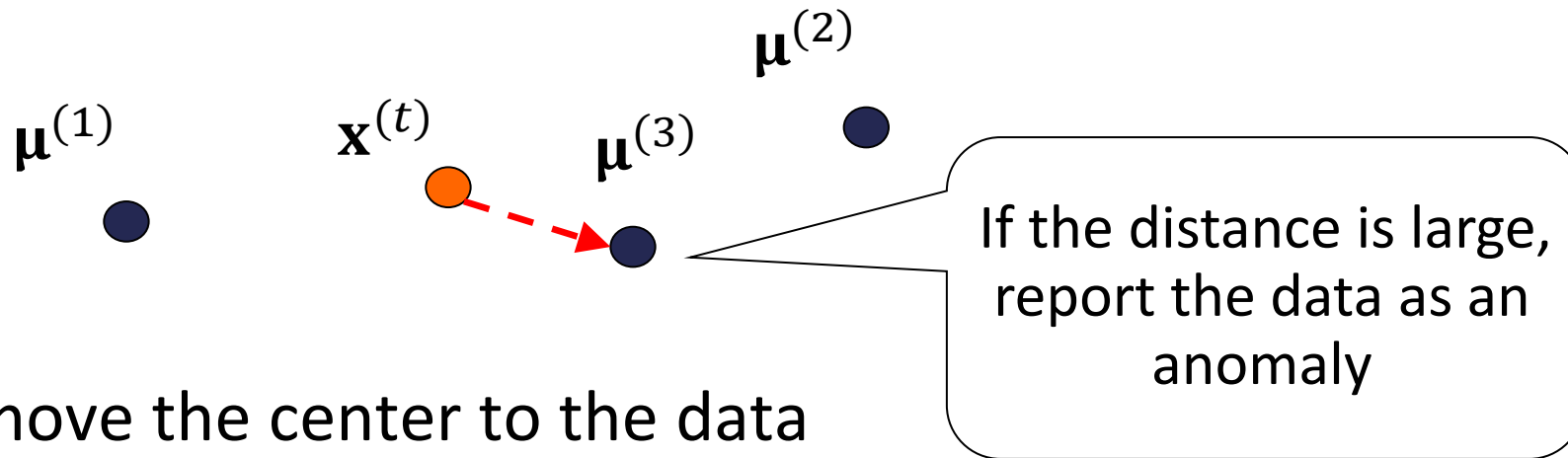
- Most anomaly detection applications require real-time system monitoring
- Data instances arrive in a streaming manner:
  - $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(t)}, \dots$  : at each time  $t$ , new data  $\mathbf{x}^{(t)}$  arrives
- Each time a new data arrives, evaluate its anomaly
- Also, models are updated in on-line manners:
  - In the one dimensional case, the threshold is sequentially updated
  - In clustering, groups (clusters) are sequentially updated

# Sequential $K$ -means:

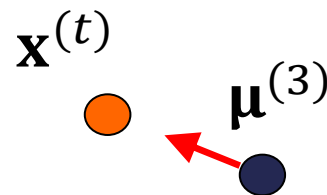
## Simultaneous estimation of clusters and outliers

- Data arrives in a streaming manner, and apply clustering and anomaly detection at the same time

1. Assign each data  $\mathbf{x}^{(t)}$  to its nearest center  $\boldsymbol{\mu}^{(k)}$



2. Slightly move the center to the data

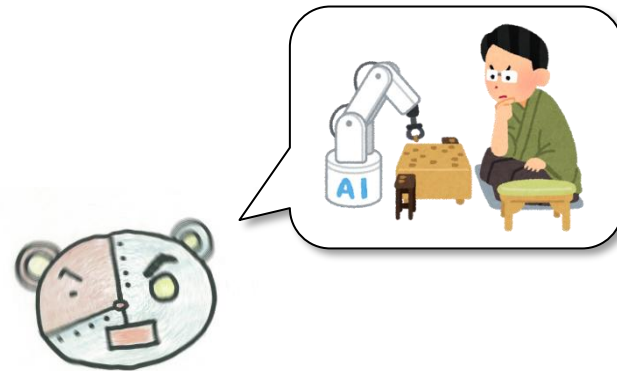


# Limitation of unsupervised anomaly detection: Details of failures are unknown

---

- In supervised anomaly detection, we know what the failures are
- In unsupervised anomaly detection, we can know something is happening in the data, but cannot know what it is
  - Failures are not defined in advance
- Based on the reports to system administrators, they have to investigate what is happening, what are the reasons, and what they should do

# Recent topics: Deep Learning



# Emergence of deep learning:

## Significant improvement of prediction accuracy

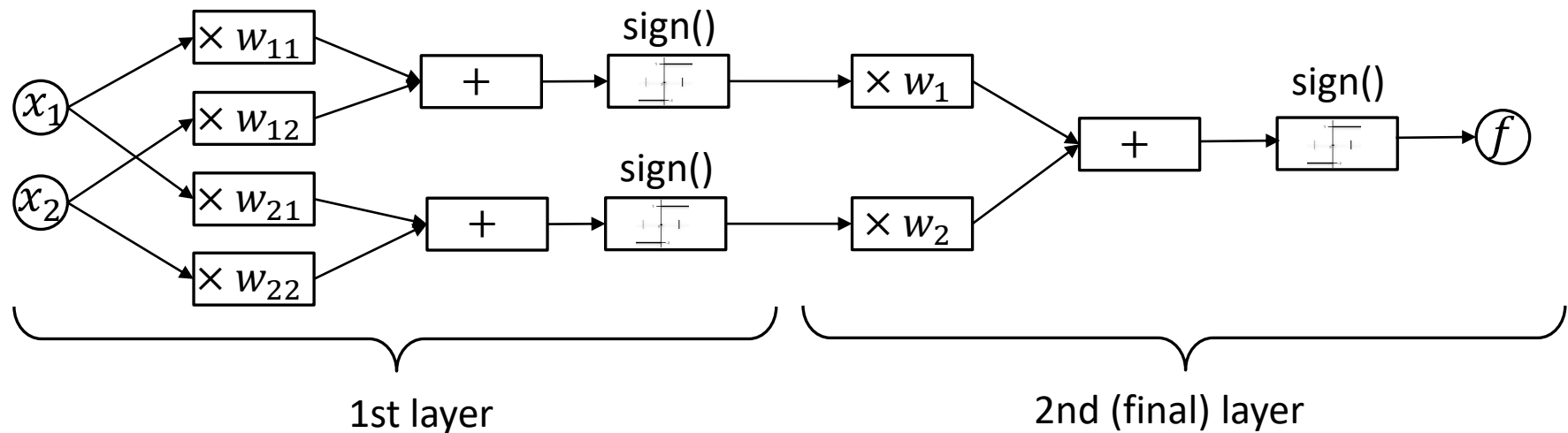
---

- Artificial neural networks were hot in 1980s, but burnt low after that...
- In 2012, a deep NN system won in the ILSVRC image recognition competition with 10% improvement
- Major IT companies invest much in deep learning technologies
- A big trend in machine learning research

# Deep neural network:

## Deeply stacked NN for high representational power

- Essentially, multi-layer neural networks
  - Regarded as stacked linear classification models
    - First to semi-final layers bear feature extraction
    - Final layer makes predictions
- Stacking many layers makes the model highly non-linear, and gives it strong power to represent complex patterns.



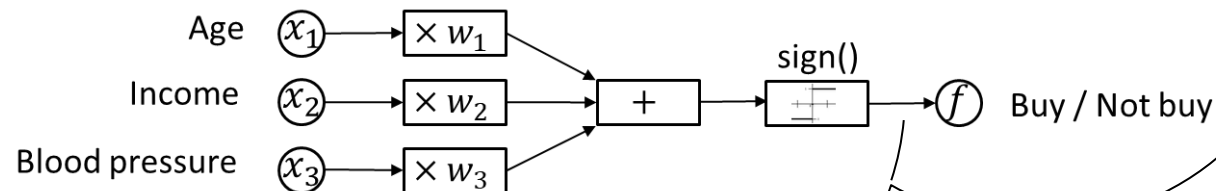
## A model for classification: Linear classification model

- Model  $f$  takes an input  $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$  and outputs a value from  $\{+1, -1\}$

$$f(\mathbf{x}) = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_Dx_D)$$

–Model parameter  $\mathbf{w} = (w_1, w_2, \dots, w_D)^\top \in \mathbb{R}^D$  :

- $w_d$  : contribution of  $x_d$  to the output  
( $x_d > 0$  contributes to  $+1$ ,  $x_d < 0$  contributes to  $-1$ )



# What is the difference from the past NN?:

## Deep structures and new techniques with modern flavors

---

- Differences from the ancient NNs:
  - Far more computational resources are available now
  - Deep network structure: from wide-and-shallow to narrow-and-deep
  - New techniques and model architectures: Dropout, batch normalization, adversarial learning, ReLU, graph neural networks, attention, ...
- We will look at some of the key ideas in this lecture

## Summary:

# Introduction to statistical learning theory

---

- Machine learning is a data-driven approach to prediction, discovery, and generation
- Applications are expanding from IT to various fields: healthcare, finance, materials, and more
- Learning = estimating a model (function) from data: supervised & unsupervised learning
- Anomaly detection finds irregular patterns by modeling normal behavior
- Deep learning leverages stacked neural networks with modern techniques and large-scale data