



# データマイニング

## Data Mining

### 15:ニューラルネットワーク③ Neural Network

#### 講義のまとめ Conclusion

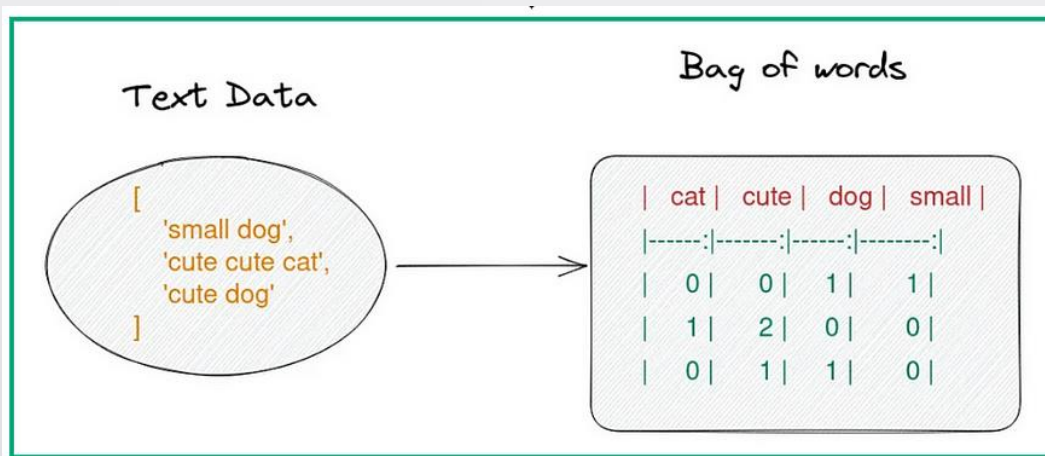
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

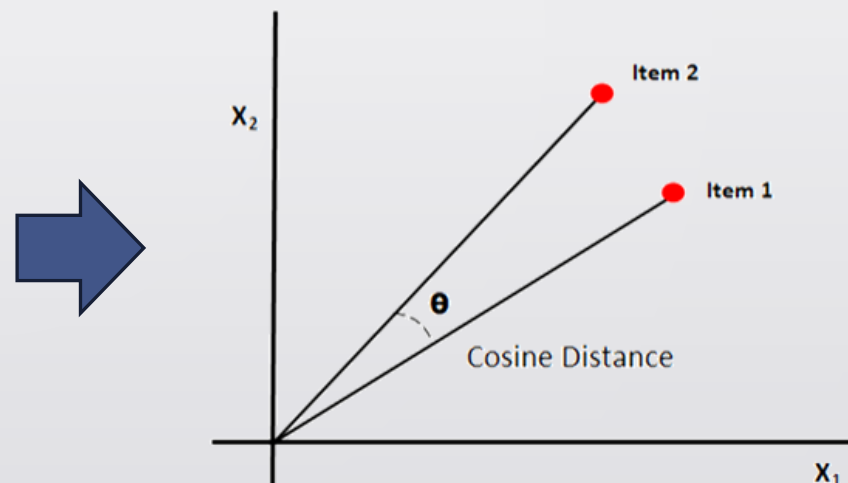
# Bag of Words

形態素解析をした後、各単語の出現頻度によりテキストの特徴量ベクトルを生成する

Generate feature vector of a text by counting frequency of each morpheme after morphological analysis



<https://ayselaydin.medium.com/4-bag-of-words-model-in-nlp-434cb38cdd1b>



## Term Frequency-Inverse Document Frequency (TF-IDF)

ある文書における特定の単語の重要度を評価する  
Measure of importance of a certain word in document

$$TF - IDF(t, d) = TF(t, d) * IDF(t, d)$$

$$TF(t, d) = \frac{\text{number of } t \text{ in document } d}{\text{total number of words in document } d}$$

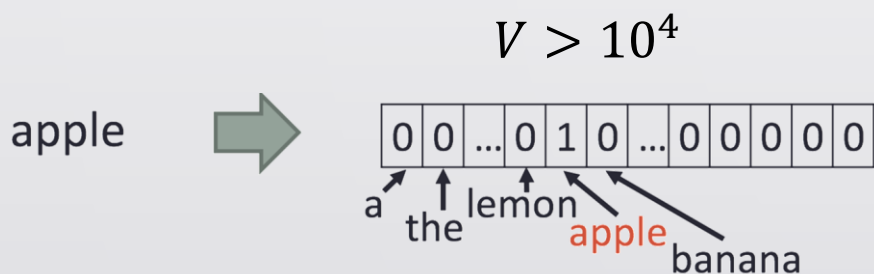
$$IDF(t, d) = \log \left( \frac{N}{1 + df} \right) \quad \begin{array}{l} N: \text{Total number of documents} \\ df: \text{Number of documents containing word } t \end{array}$$

# 分布仮説と分散表現

## Distributional Hypothesis and Distributed Representation

### One-Hot Encodingによる 単語表現

Representation of words by one-hot encoding



<https://qiita.com/kouhara/items/e895f6350aa1ebe77133>

### 分布仮説に基づく低次元ベクトルでの単語 の意味表現

Representation of word meaning by low dimensional vector based on "Distributional Hypothesis"

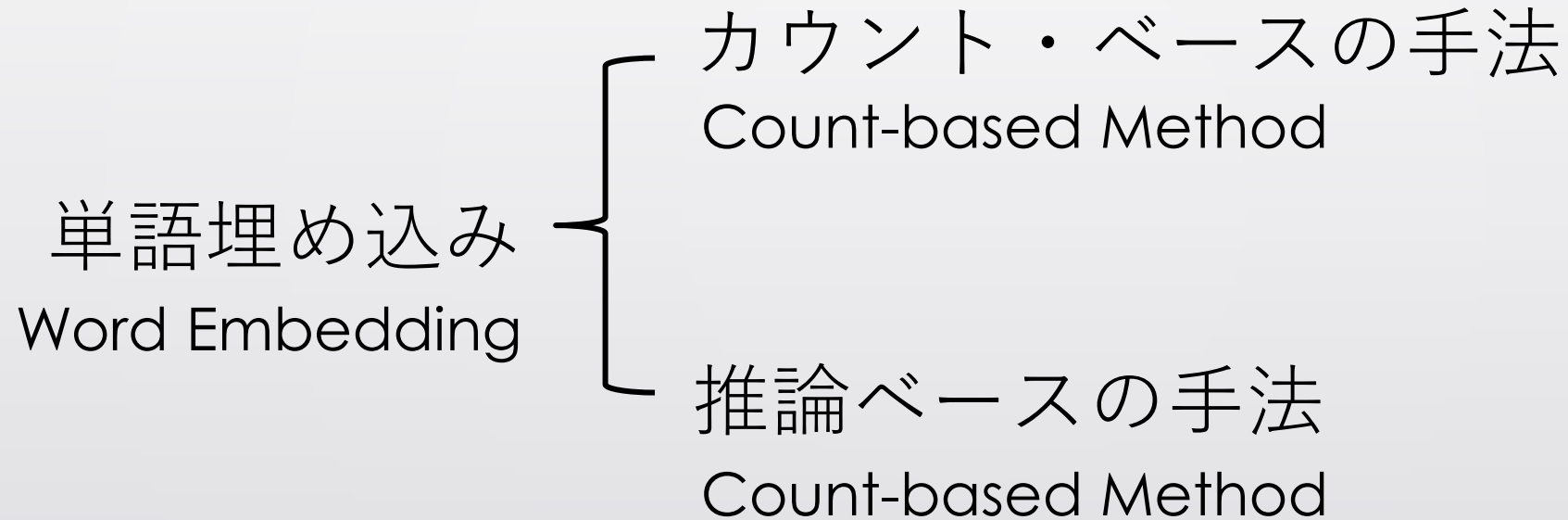
$$apple = [ \text{circle} \text{ circle } \dots \text{ circle} ]$$

### ※分布仮説 Distributional Hypothesis

類似の文脈に登場する単語は似た意味を持つ

Words occurring in similar contexts have similar meanings

# 分散表現の獲得 Acquisition of Distributed Representation





# 潜在意味解析 Latent Semantic Analysis

## 共起行列 Co-occurrence Matrix



<https://medium.com/@imamitsehgai/nlp-series-distributional-semantics-co-occurrence-matrix-31283629951e>

## 相互情報量行列

Point-wise Mutual Information(PMI) Matrix

$$PMI(w_1, w_2) = \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

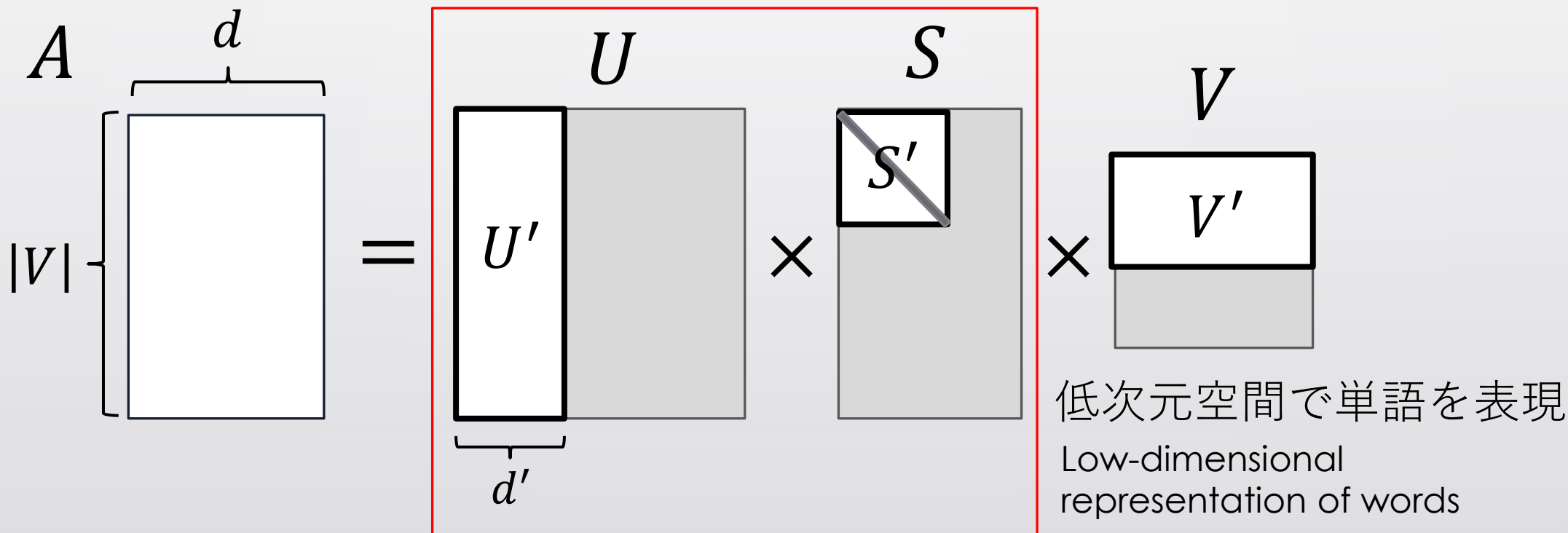
$P(w_i)$ : 文書に単語 $w_i$ が登場する確率

$P(w_i, w_j)$ : 文書に単語 $w_i$ と $w_j (i \neq j)$ が同時に登場する確率

# 潜在意味解析 Latent Semantic Analysis

PMI行列を特異値分解で次元削減

Dimension reduction by singular value decomposition of PMI matrix

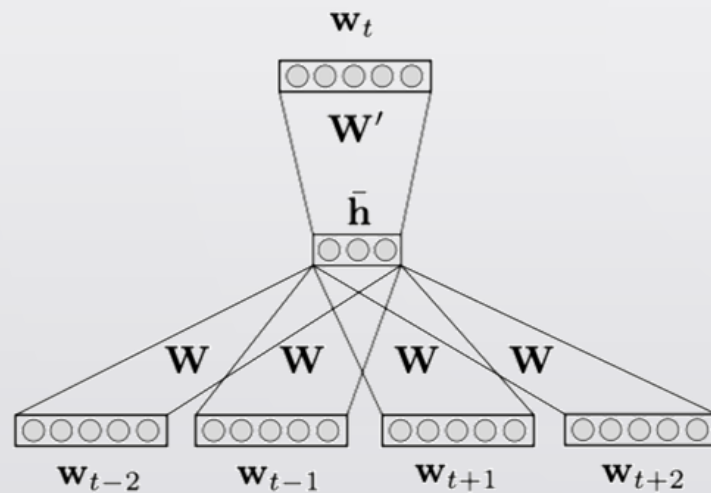


# Word2vec

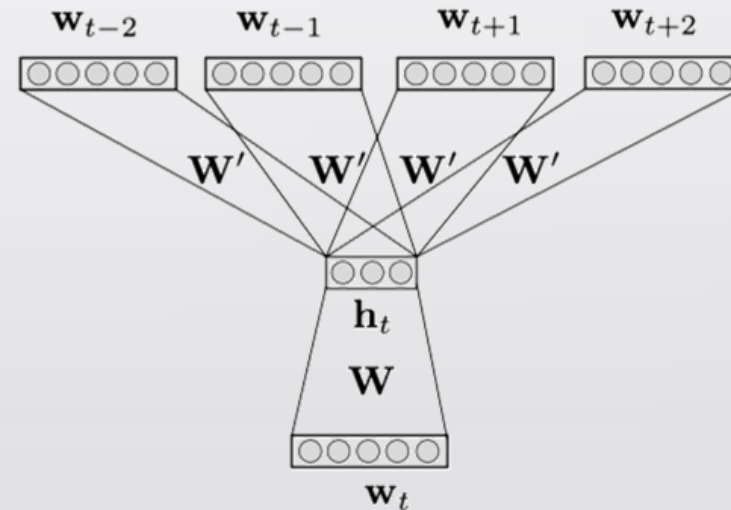
分布仮説に基づき単語埋め込みを行うニューラルネットワークモデル

Neural network models that conduct word embedding based on distributional hypothesis

CBOw



Skip-gram

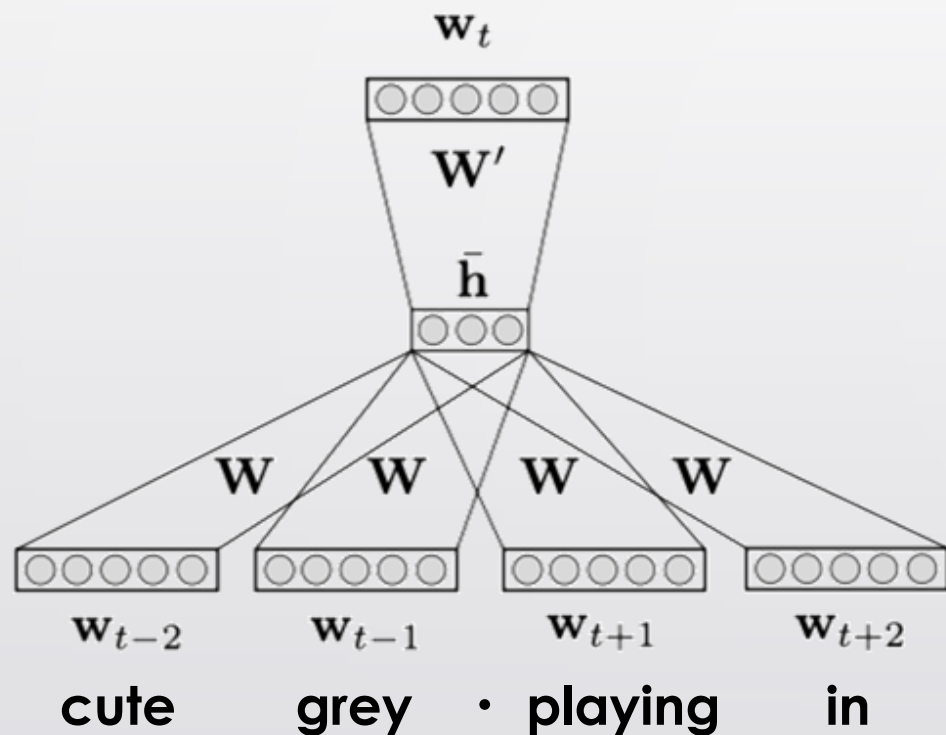


<https://cvml-expertguide.net/terms/nlp/word2vec/>



# Continuous Bag of Words (CBOW)

“**cat**”を予測 Predict the word “cat”



前後の単語から中心にくる単語を推測する

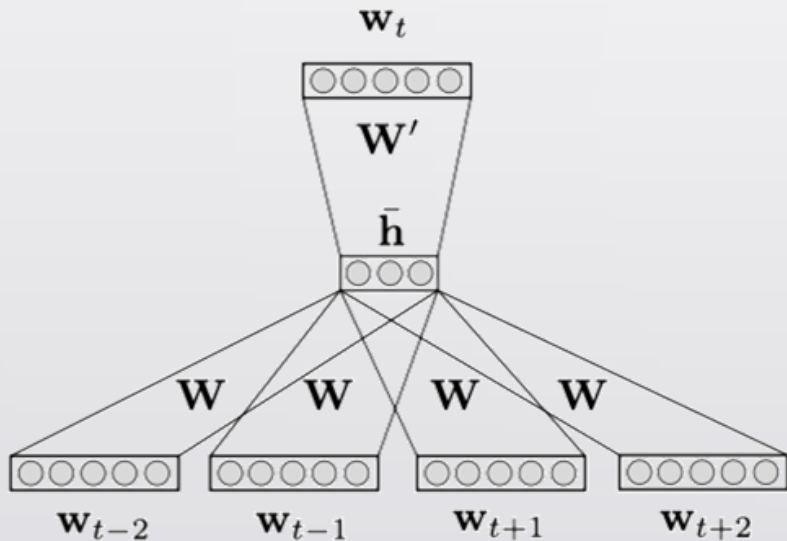
Train the network so that it can predict a target word flanked by context words preceding or following the target word

One-hot Encodingされた文脈単語の入力

# Continuous Bag of Words (CBOW)

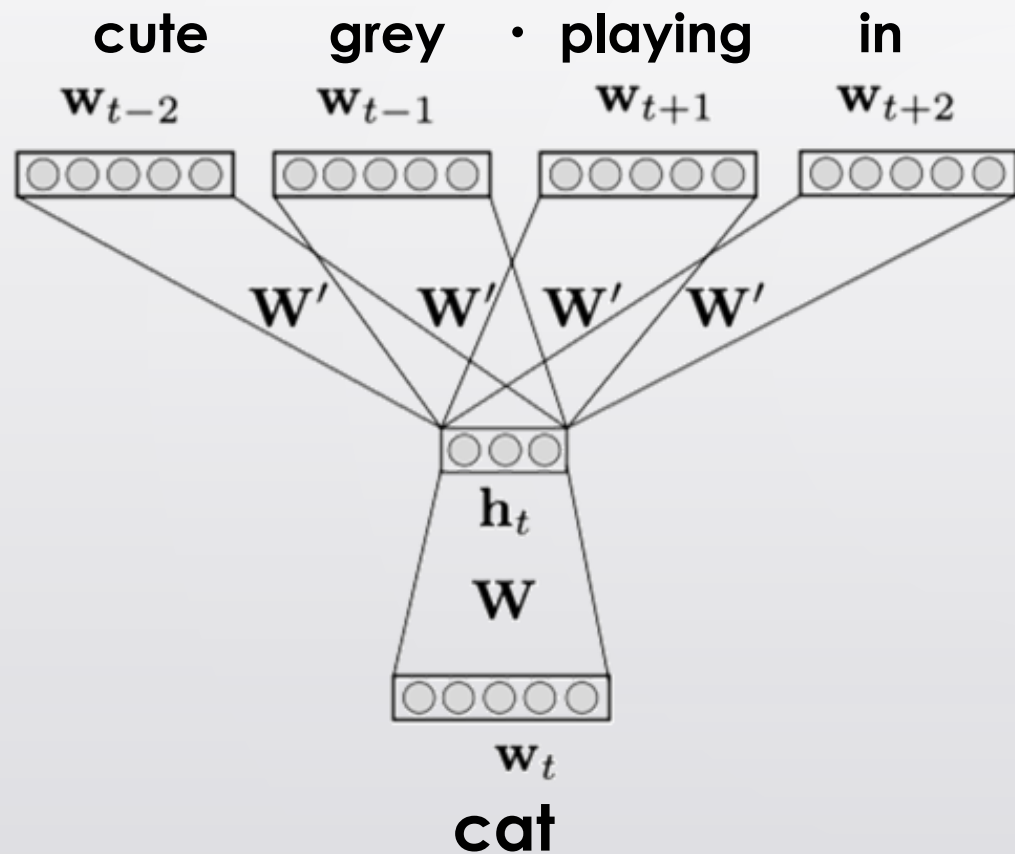
重み行列 $W$ に単語の意味が低次元で表現される

Low-dimensional representation of word meaning is stored in weight matrix  $W$



$$\begin{matrix} w_{t-2} & [0, 0, 0, 1, 0, \dots 0, 0, 0, 0] \\ w_{t-1} & [0, 0, 0, 0, 0, \dots 1, 0, 0, 0] \\ w_{t+1} & [1, 0, 0, 0, 0, \dots 0, 0, 0, 0] \\ w_{t+2} & [0, 0, 1, 0, 0, \dots 0, 0, 0, 0] \end{matrix} \cdot \begin{matrix} W \\ (|V|, 3) \end{matrix} = \begin{matrix} h \\ (1, 3) \end{matrix}$$

# Skip-gram



CBOWとは逆に、入力後の前後の文脈単語を予測する

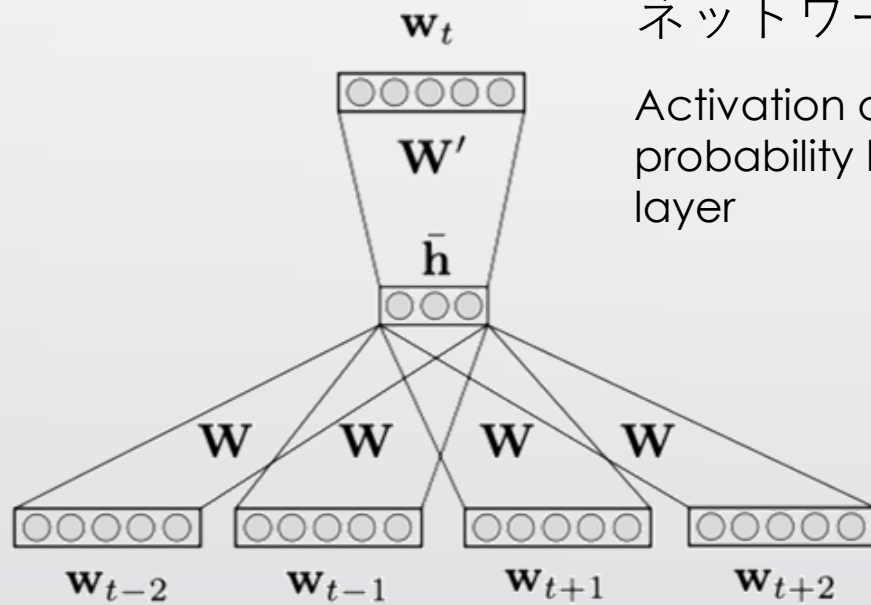
In contrast to CBOW, a network is trained so that it predicts context words flanking the input word

# Word2Vecの損失関数 Loss Function of Word2Vec

出力層ではソフトマックス関数により、  
ネットワークの活性が確率に変換される

Activation of network is converted into  
probability by softmax function in the output  
layer

$$P(w_k) = \frac{\exp(y(w_k))}{\sum_{|V|} \exp(y(w_i))}$$



損失関数は交差エントロピー  
Loss function is cross entropy

$$L = - \sum_{i=1}^{i=|V|} t_i \log[P(w_i)]$$

$$\mathbf{t} = (t_1, t_2 \cdots t_{|V|})$$
$$t_k \in \{0, 1\}$$

## 負例サンプリング Negative Sampling

コーパスデータを用いた学習では $|V|$ が巨大な数値になる


The value of  $|V|$  is huge in training based on corpus data

$$P(w_k) = \frac{\exp(y(w_k))}{\sum_{|V|} \exp(y(w_i))} \quad L = - \sum_{i=1}^{i=|V|} t_i \log[P(w_i)]$$

二値分類問題に置き換えることで計算を高速化

Accelerate computation by replacing the multiclass classification with binary classification





# データマイニングの流れ Steps in Data Mining

1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis ・ モデリング Modeling
6. 性能評価 Performance Evaluation
7. (ディプロイメント Deployment)

# データ収集 Data collection



Confidentiality

機密性



Integrity

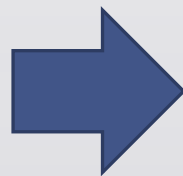
完全性



Availability

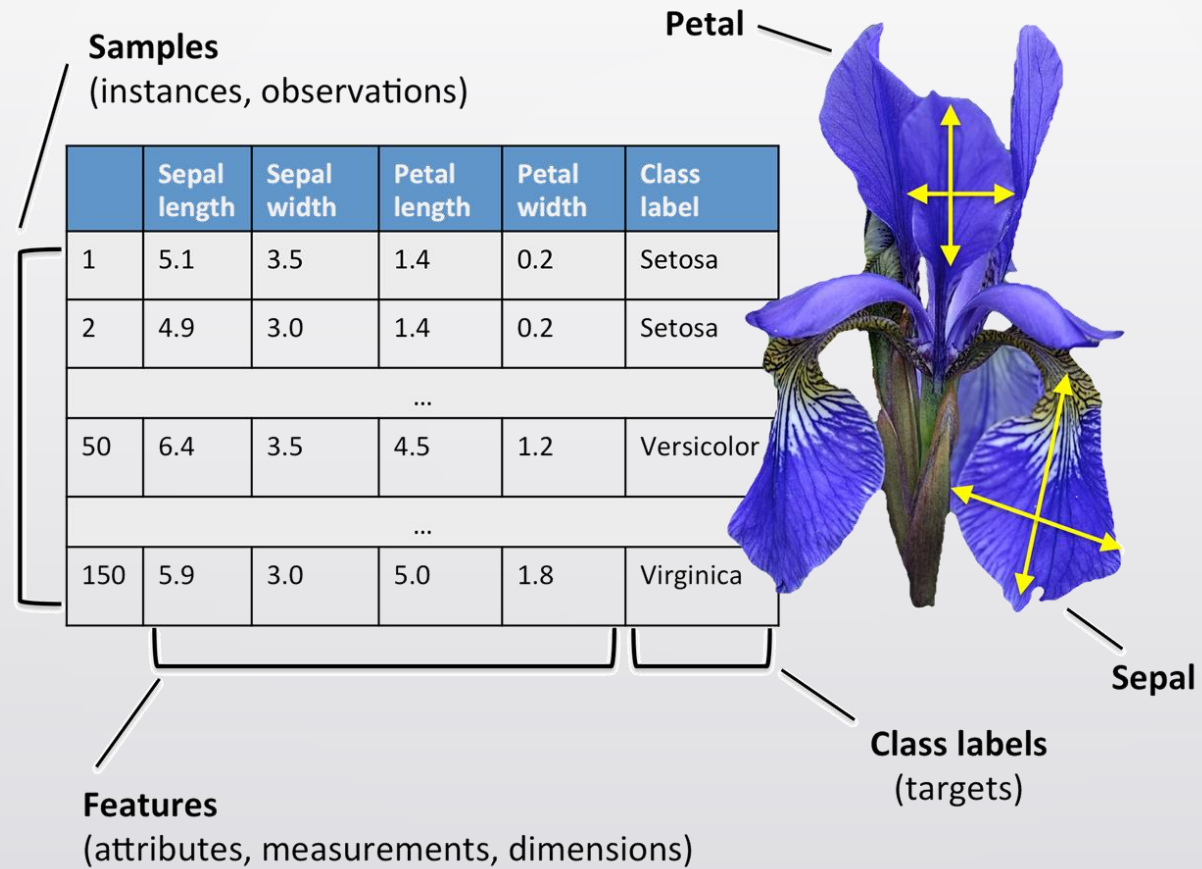
可用性

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円



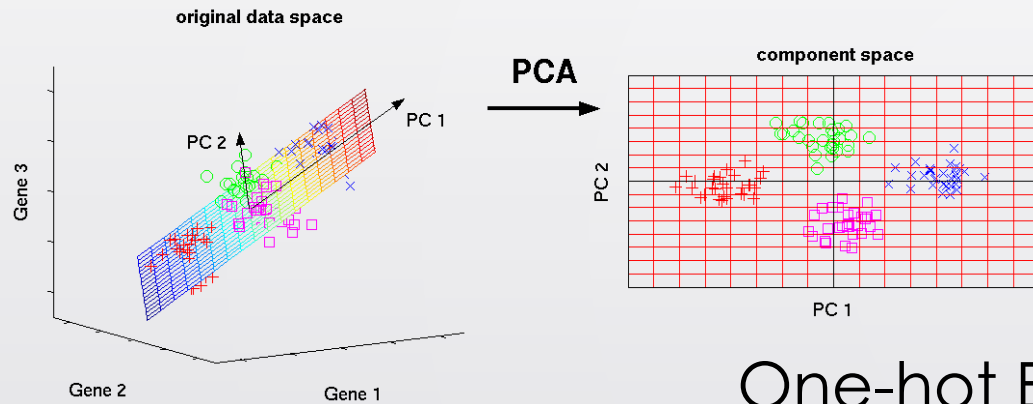
性別	年齢	年収	
男	[20-29]	[300-499]	万円
女	[30-39]	[500-699]	万円
女	[30-39]	[500-699]	万円
男	[20-29]	[300-499]	万円

# 特徴量選択 Feature Selection



# 前処理 Pre-processing

## 次元削減 Dimension Reduction



## スケーリング Scaling

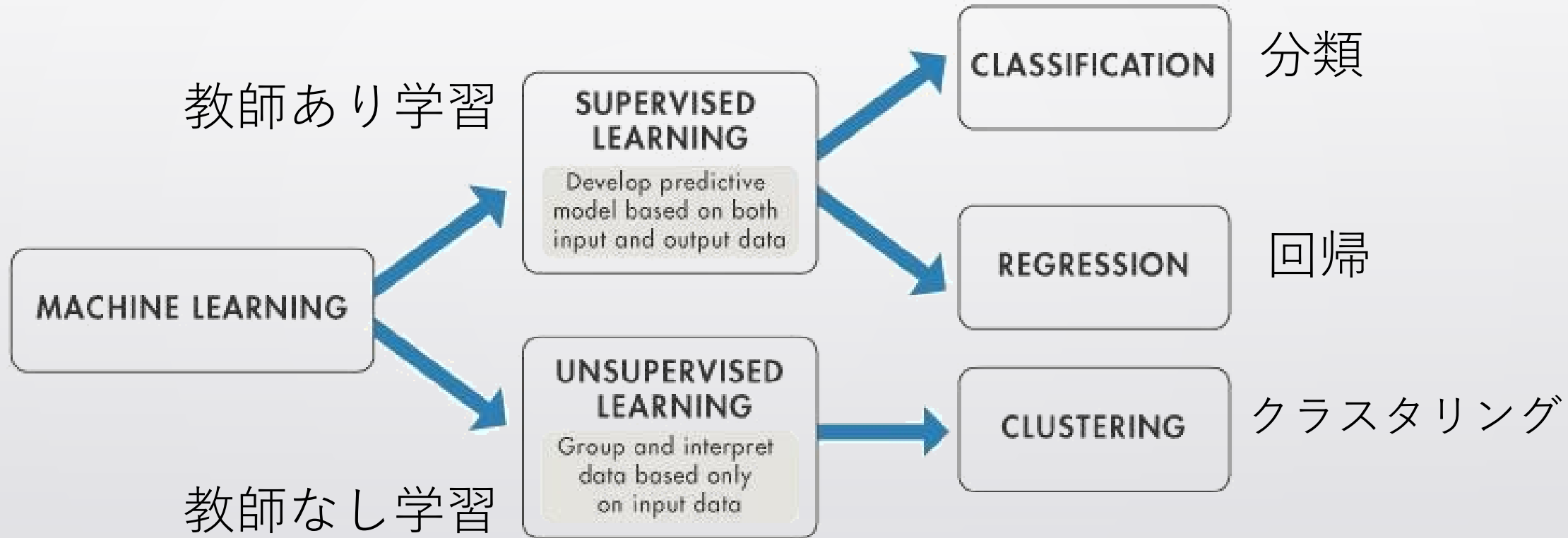
$$x' = \frac{x - \mu}{\sigma}$$

$$x' = \frac{x - \text{median}}{NIQR}$$

## One-hot Encoding

Color			
Red	Red	Yellow	Green
Red	1	0	0
Yellow	1	0	0
Green	0	1	0
Yellow	0	0	1

# データ分析・モデリング



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)



# 多項式回帰 Polynomial Regression

Simple  
Linear  
Regression

$$y = b_0 + b_1 x_1$$

Multiple  
Linear  
Regression

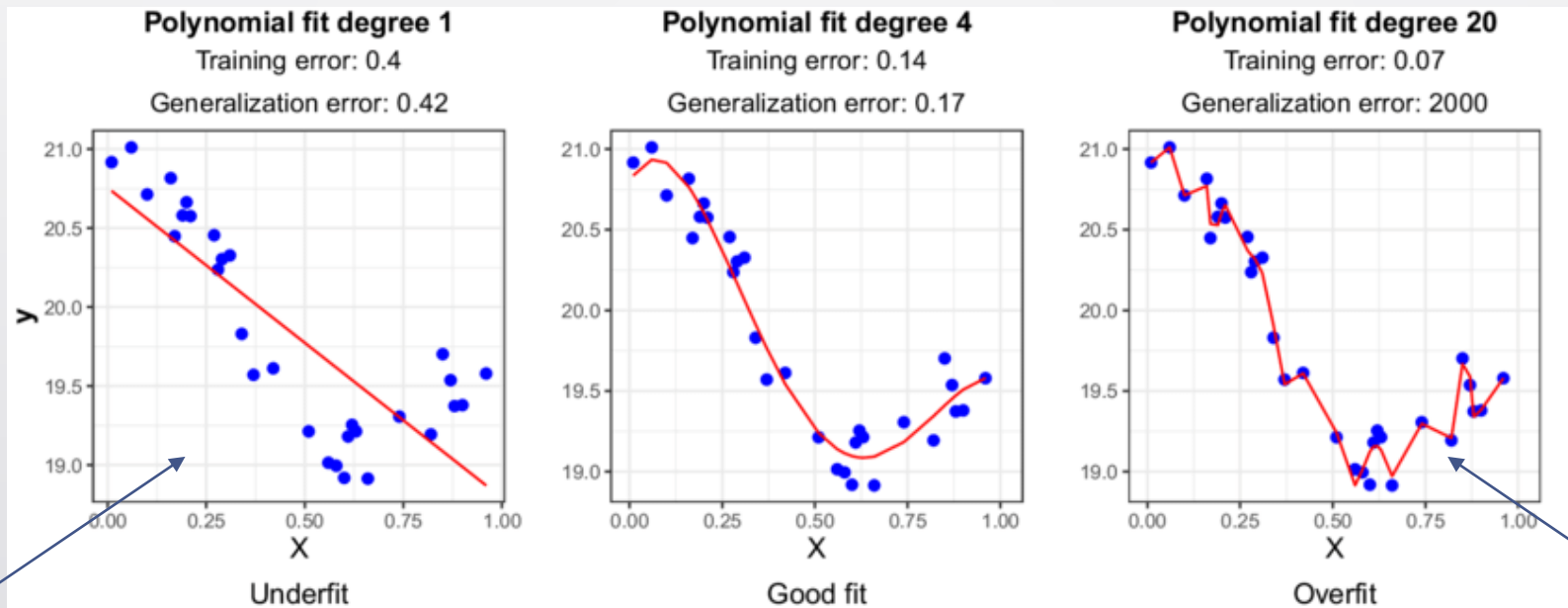
$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Polynomial  
Linear  
Regression

$$y = b_0 + b_1 x_1 + \underline{b_2 x_1^2} + \dots + b_n x_1^n$$

# 性能評価 Performance Evaluation

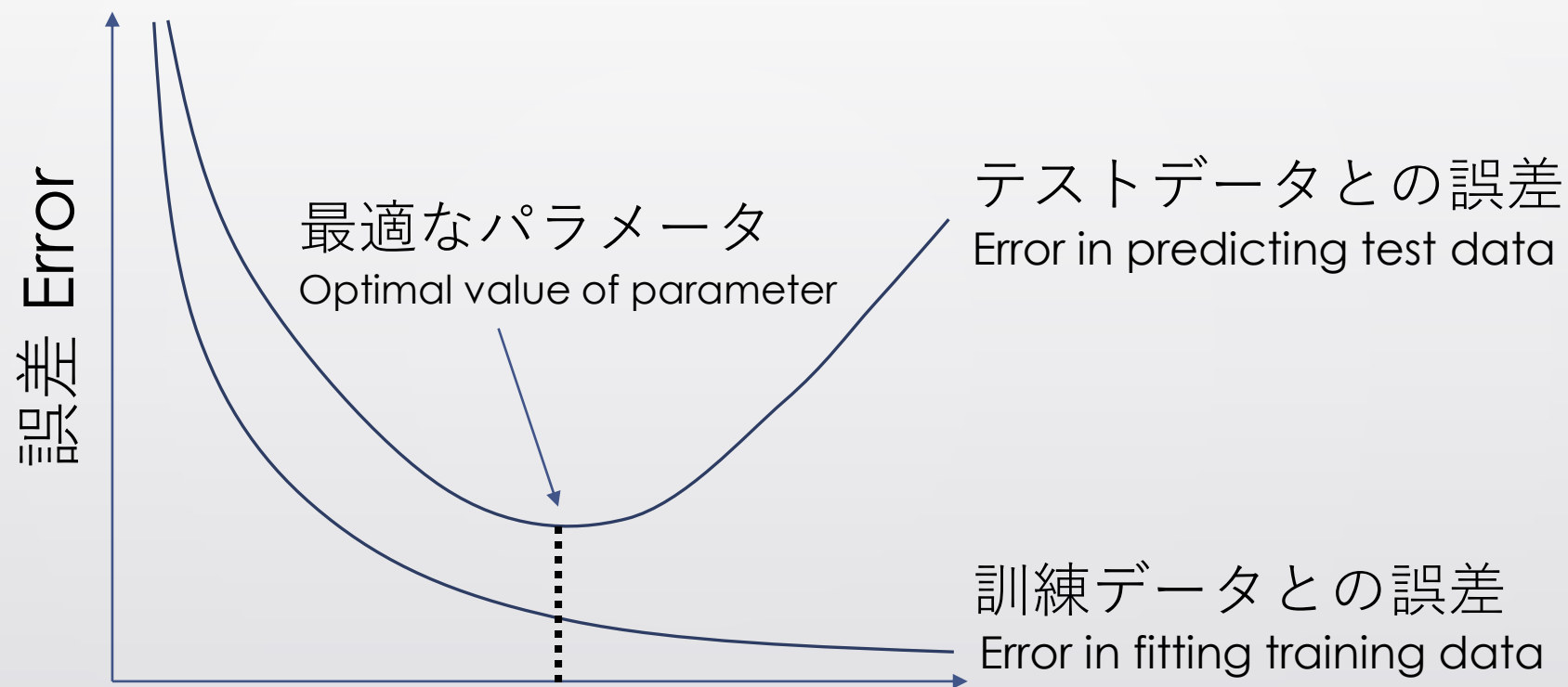
観測されたデータはノイズを含む Observed data contains random noise



単純すぎてはダメ  
Should not be too simplistic

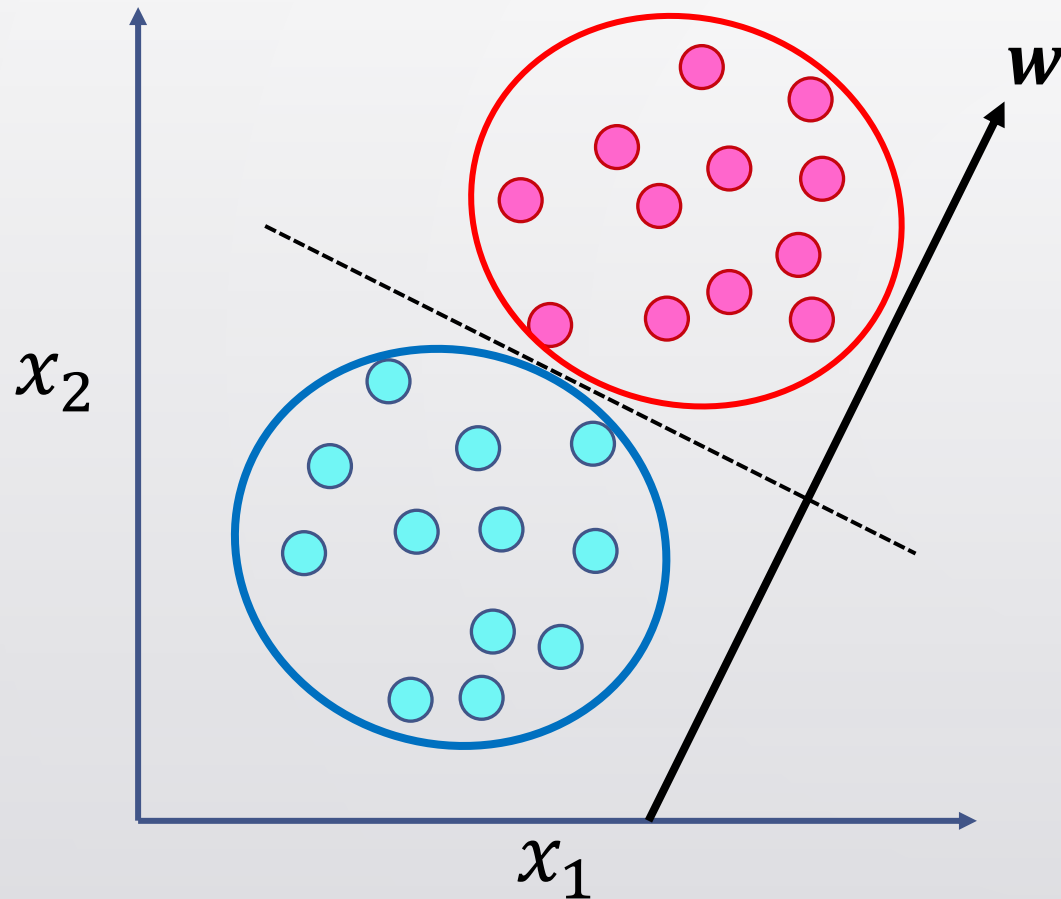
複雑すぎてはダメ  
Should not be too complex

# 性能評価 Performance Evaluation



調整できるパラメータ : 選択された変数、ハイパーパラメータ 等  
Adjustable Parameter      Selected variable, hyperparameter etc

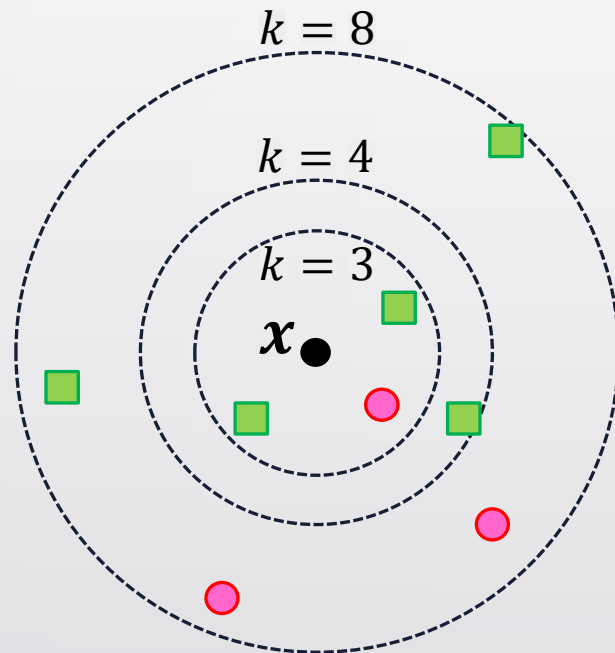
# 線型判別分析 Linear Discriminant Analysis (LDA)



よい決定境界は、下の二つの条件を満たす  
A good decision boundary meets the two conditions below

1. 2 クラスの中心が離れている  
Centers of the two classes are distant from each other
2. 各クラスのクラス内分散が小さい  
Within-class variance of each class is small

# $k$ 最近傍法 $k$ Nearest Neighbor Method



データ  $x$  のクラスを最近傍にある  $k$  個のデータの多数決投票により決定する

Class of data  $x$  is determined by majority voting of  $k$  data points closest to  $x$



# ナイーブベイズ Naïve Bayes

メールに“秘密”“技術”“大当たり”という 3 つの単語が含まれていた。

An e-mail contains three words, “Secret”, “Technology” and “Jackpot”

$$\frac{P_3(H_s|W_3)}{P_3(H_a|W_3)} = \frac{P(W_3|H_s)P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_3|H_a)P(W_2|H_a)P(W_1|H_a)P_1(H_a)} = \frac{628 \times 10^{-4}}{128 \times 10^{-4}}$$

$$\textit{Probability of being a spam} = \frac{628}{628 + 128} = 0.83$$

スパムメールと判定するかどうかは閾値による

It depends on the threshold whether the e-mail is judged to be a spam or not

## 閾値と偽陽性 Threshold and False Positives

正解 Answer

判定

Judgment

	スパムメール Spam Mail	普通のメール Authentic Mail
スパムメール Spam Mail	真陽性 True Positive	偽陽性 False Positive
普通のメール Authentic Mail	偽陰性 False Negative	真陰性 True Negative

スパムと判定する閾値を下げる

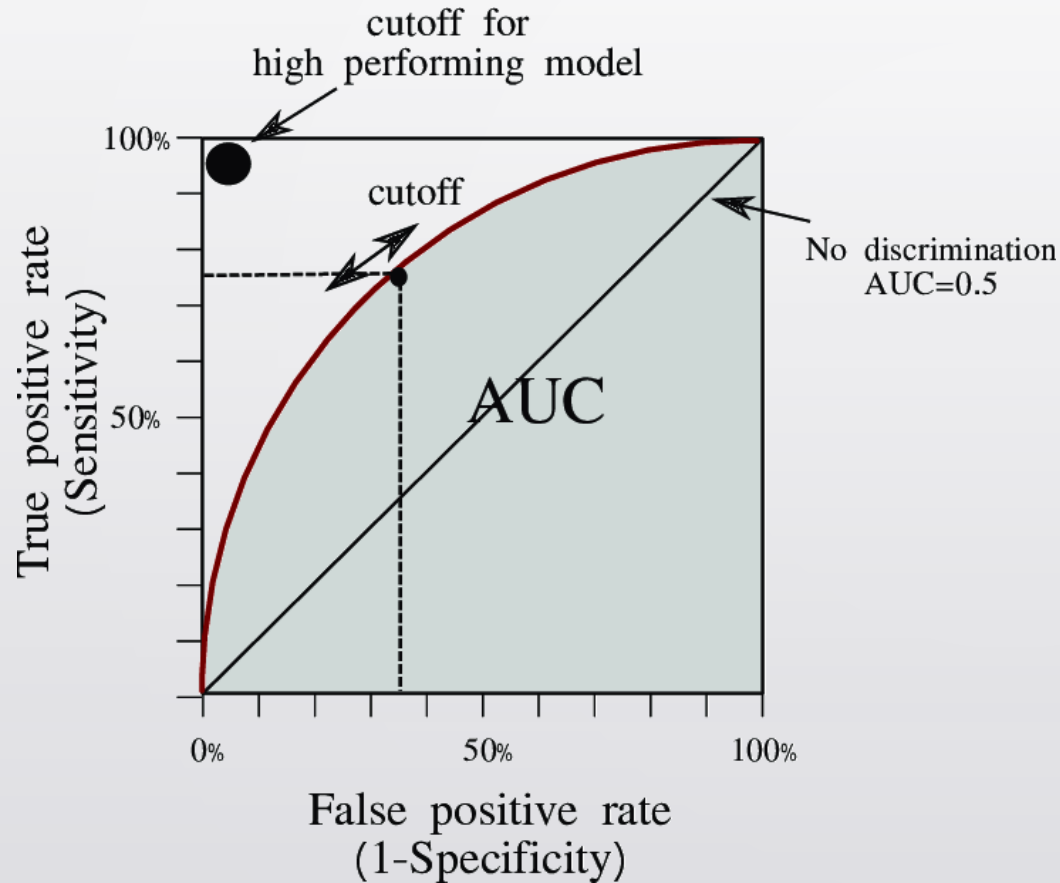


偽陽性率が上がる

Lowering threshold for judging to be a spam

Higher false positive rate

# ROC曲線 ROC(Receiver-Operator Characteristics) Curve



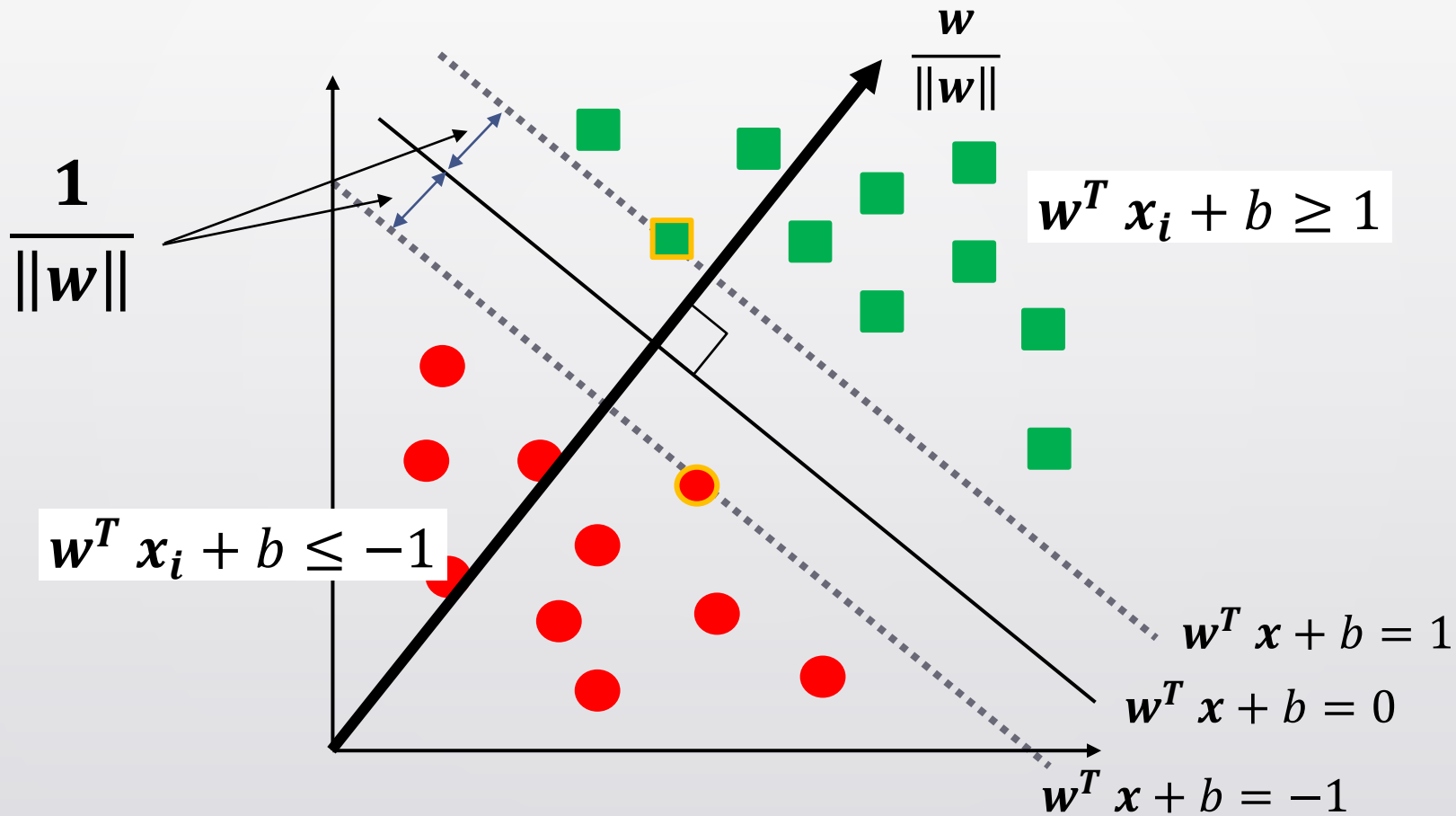
AUC: Area Under Curve

AUCが大きいほど、分類器の性能が良い

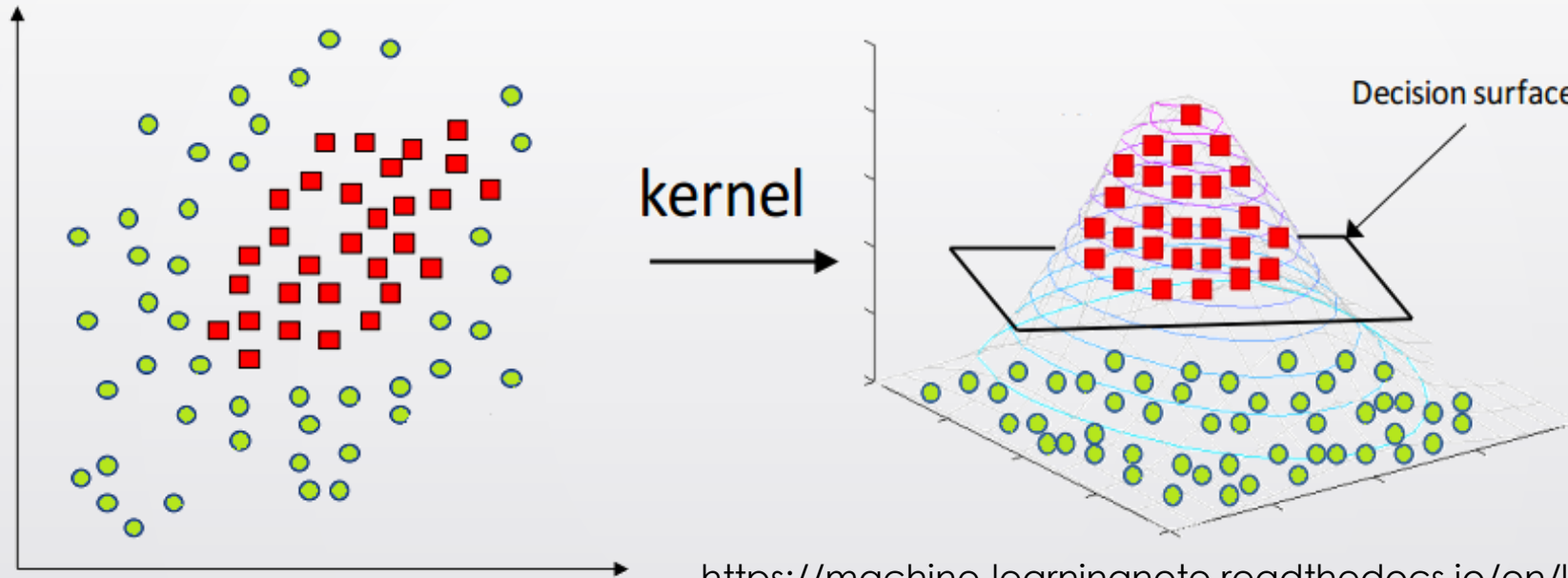
Larger AUC indicates better performance of classifier

AUC	
0.9 - 1.0	High accuracy
0.7 - 0.9	Moderate accuracy
0.5 - 0.7	Low accuracy

# サポートベクターマシン Support Vector Machine



# カーネル法 Kernel Methods



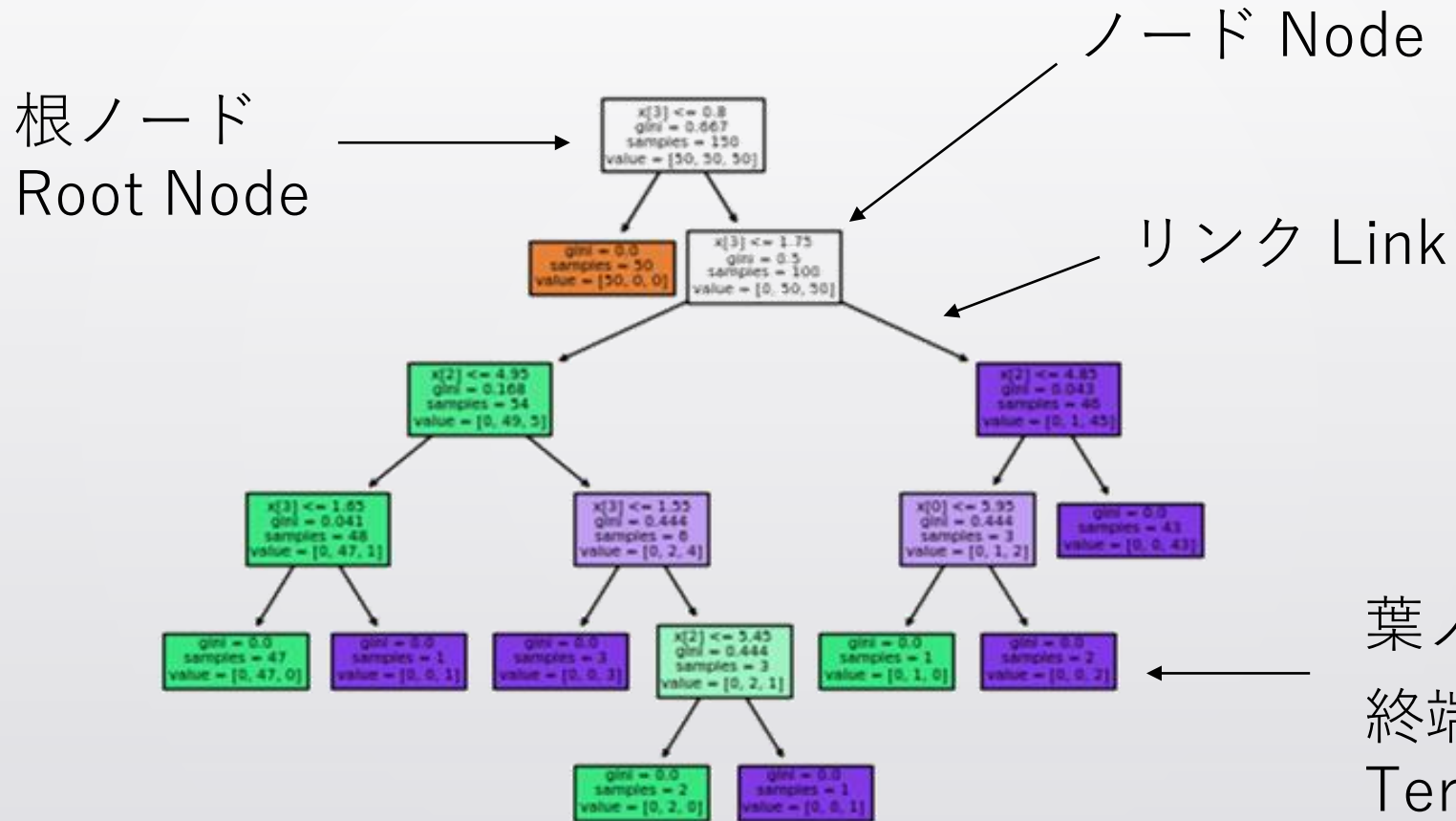
<https://machine-learningnote.readthedocs.io/en/latest/algorithm/svm.html>

データを高次元空間に写像することで、線型分離不可能な問題を線型分離可能にする

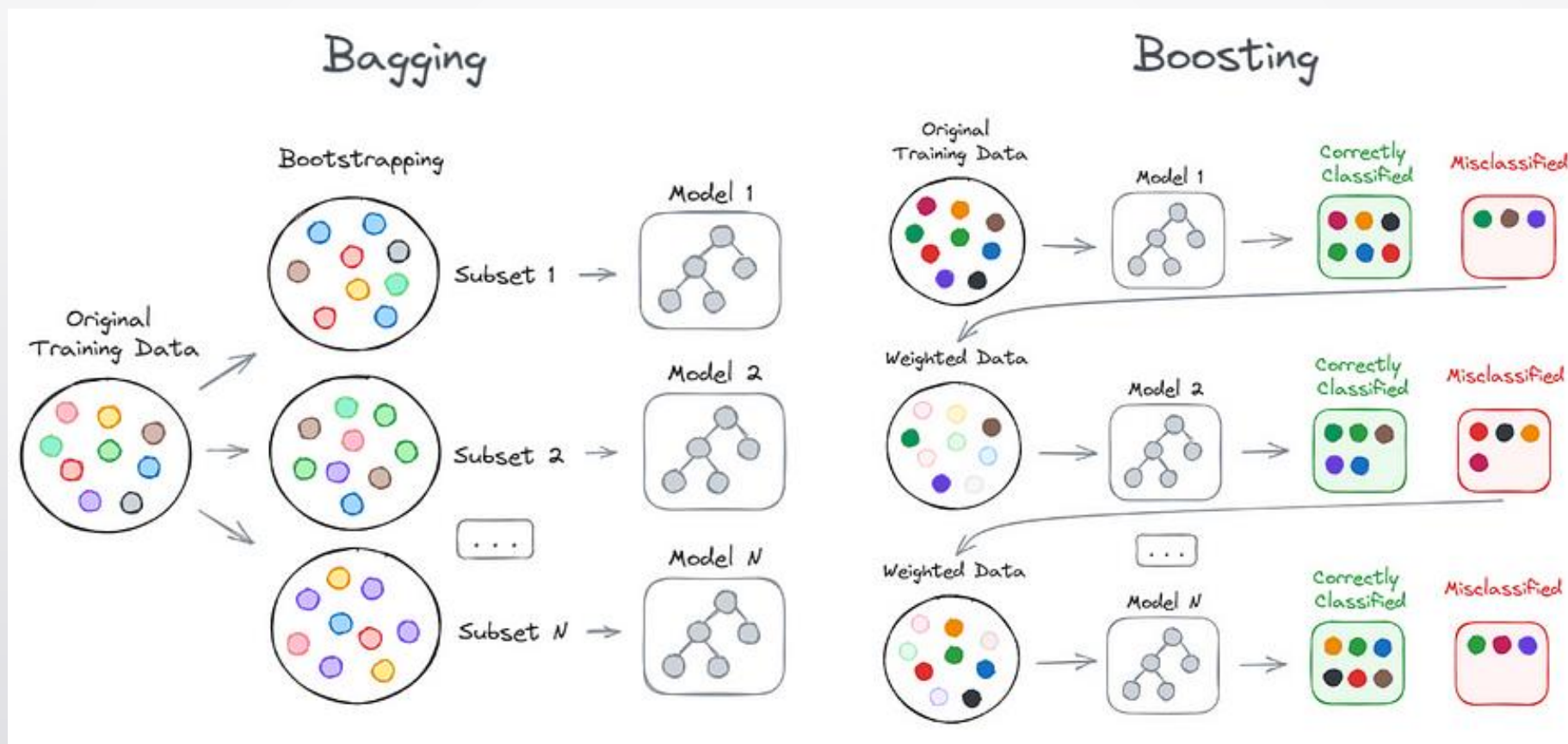
Transforming linearly inseparable problem to linearly separable one by mapping data to higher-dimensional space



# 決定木 Decision Tree

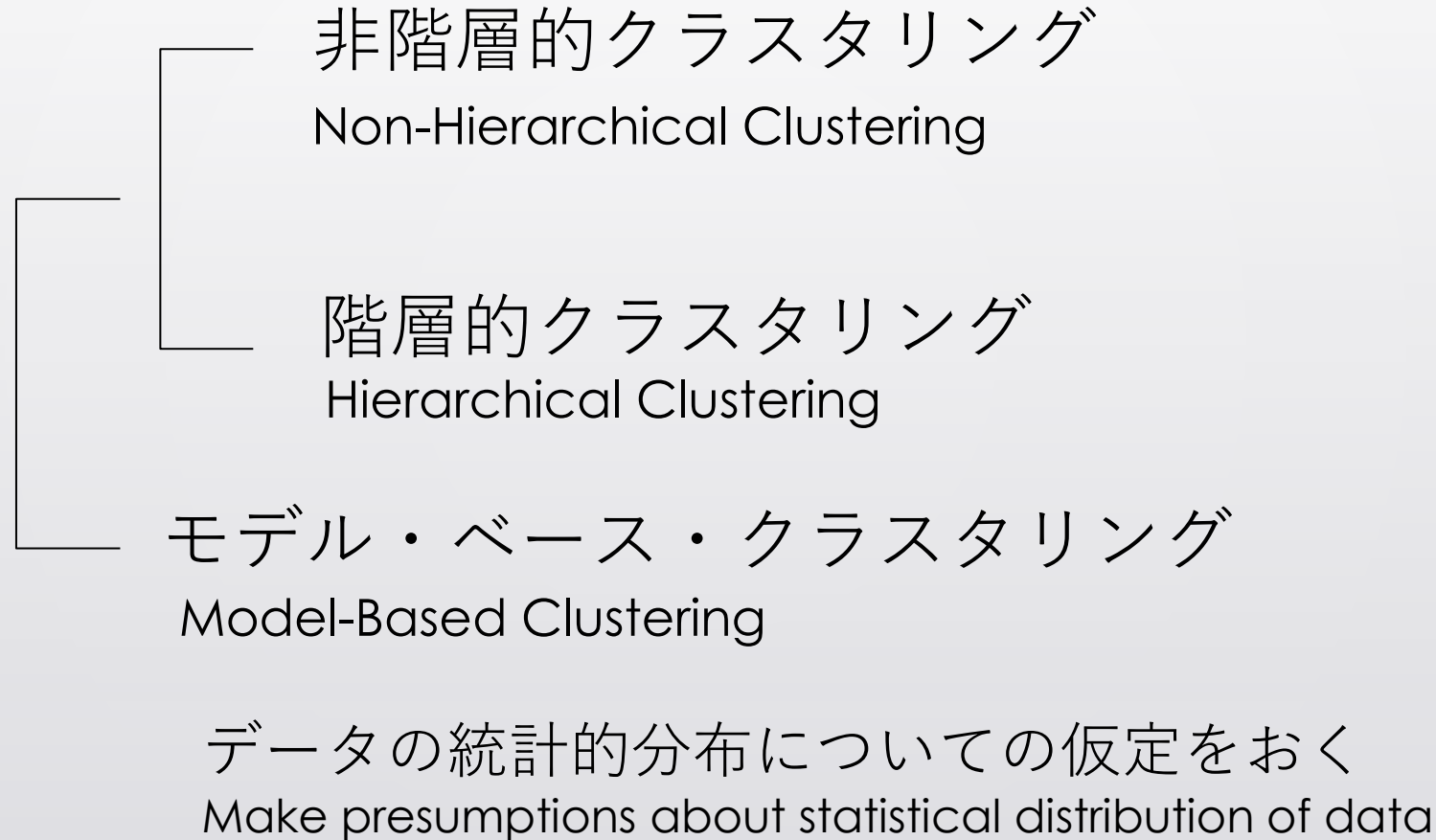


# バギングとブースティング Bagging and Boosting

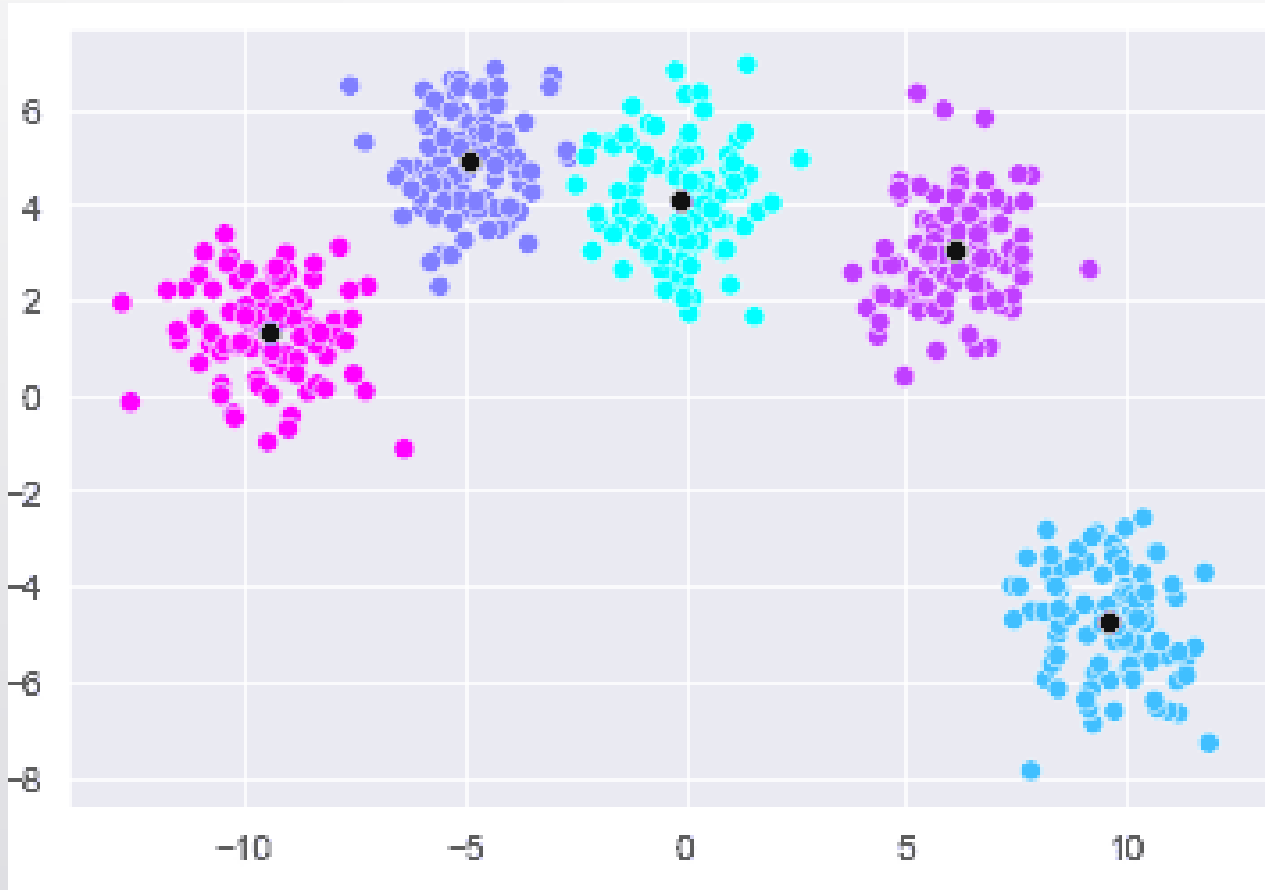


<https://pub.towardsai.net/bagging-vs-boosting-the-power-of-ensemble-methods-in-machine-learning-6404e33524e6>

# クラスタリングの種類 Types of Clustering



# K平均クラスタリング k-means clustering



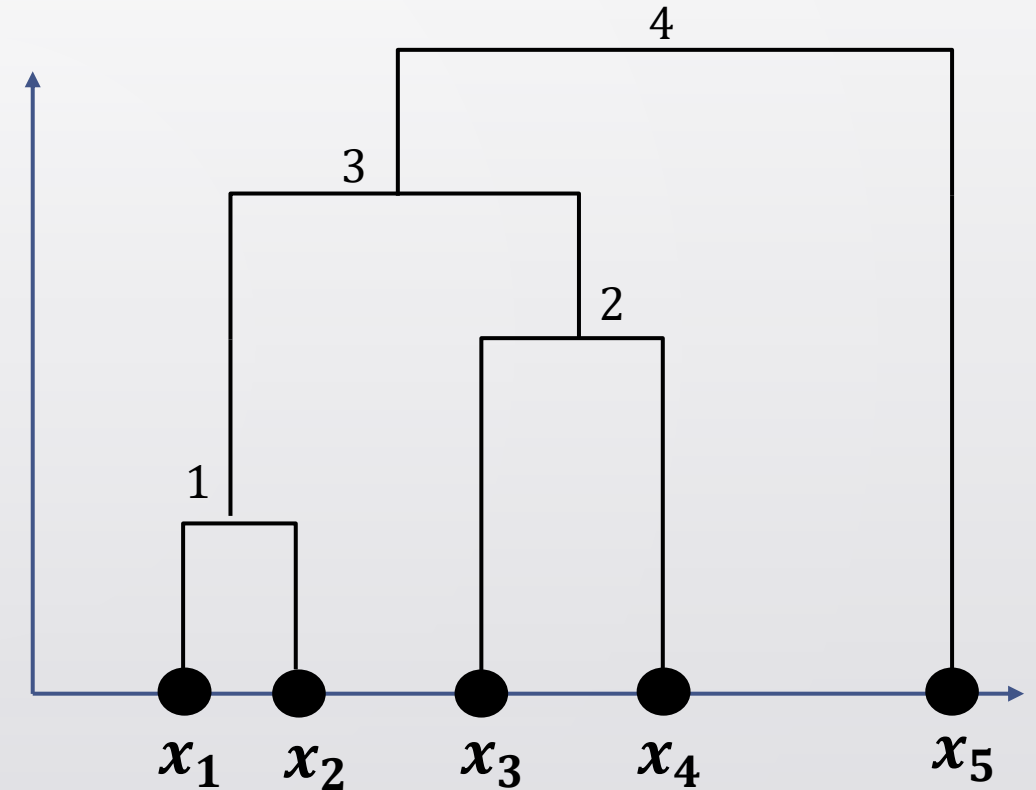
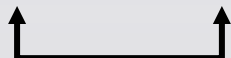
クラスターの数指定しなくては  
いけない

You have to specify the number of  
clusters,  $k$ .

# 凝集性階層的クラスタリング Agglomerative Hierarchical Clustering

	$\{x_1, x_2, x_3, x_4\}$	$x_5$
$\{x_1, x_2, x_3, x_4\}$	0	
$x_5$	4	0

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
1	2	5	7	11



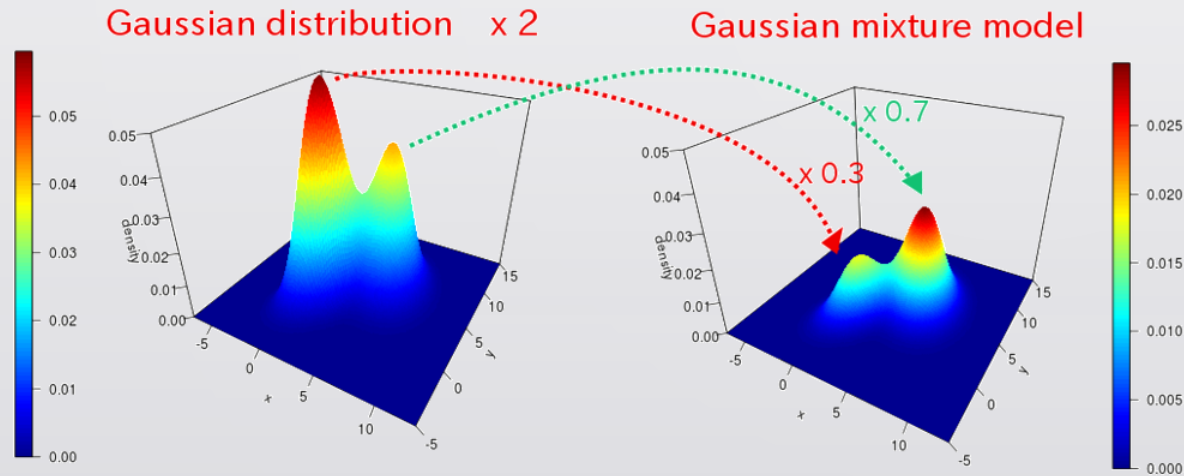


# 混合ガウス分布 Gaussian Mixture Distribution

$M$ 個の正規分布の重ね合わせにより確率分布を表現する

Represent probability distribution as weighted mixture of  $M$  normal distributions

$$p(\mathbf{x}) = \sum_{m=1}^M \pi_m N(\mathbf{x} | \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \quad 0 \leq \pi_m \leq 1 \quad \sum_{m=1}^M \pi_m = 1 \quad \pi_m : \text{混合比 Mixing Ratio}$$

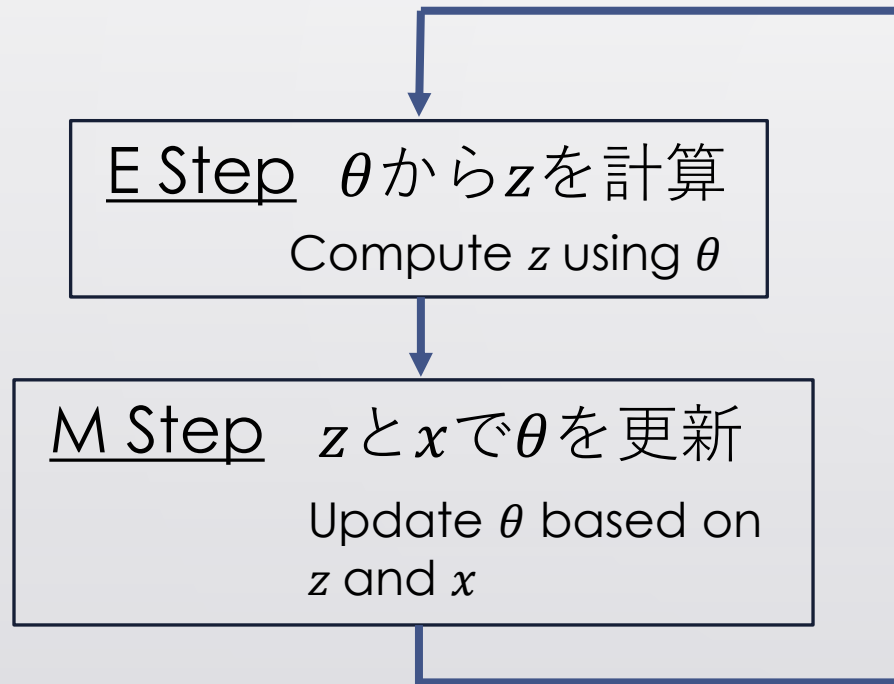


<https://work-in-progress.hatenablog.com/entry/2018/11/08/224826>

# EM アルゴリズム Expectation-Maximizing Algorithm

潜在変数を含むモデルの代表的なパラメータ推定法

Algorithm for parameter estimation of models including latent variables



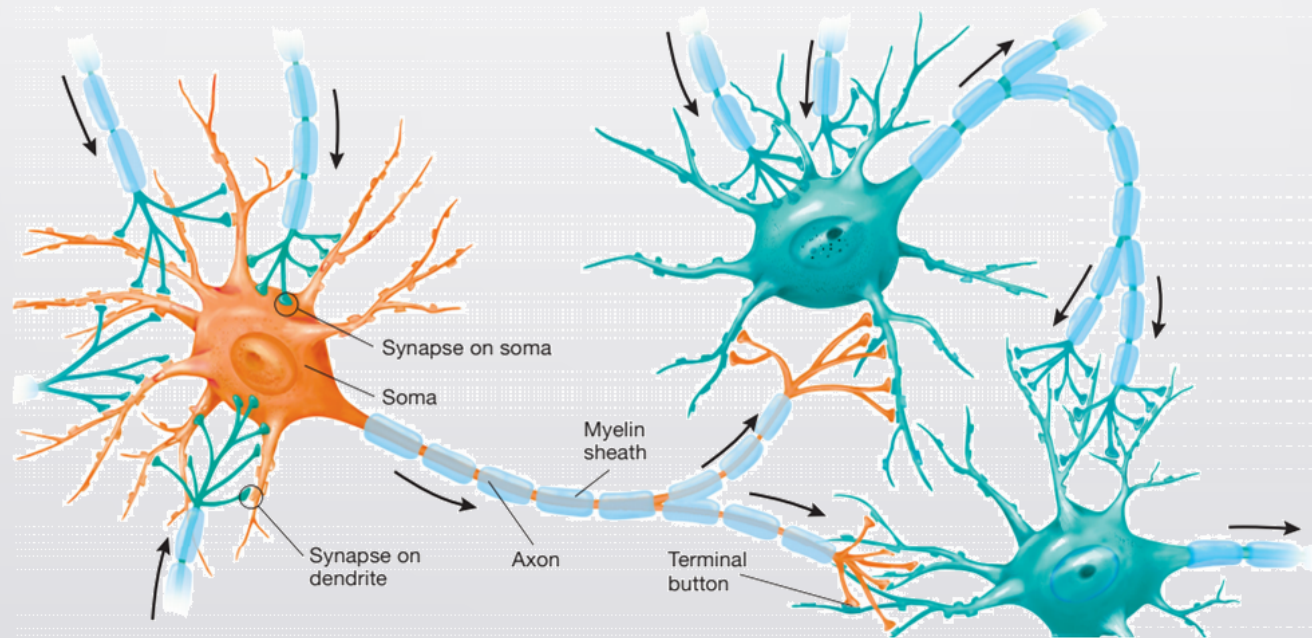
$\mathbf{x}$ : 観測 Observations

$\theta$ : 確率密度関数のパラメータセット  
Parameter set of probability distribution functions

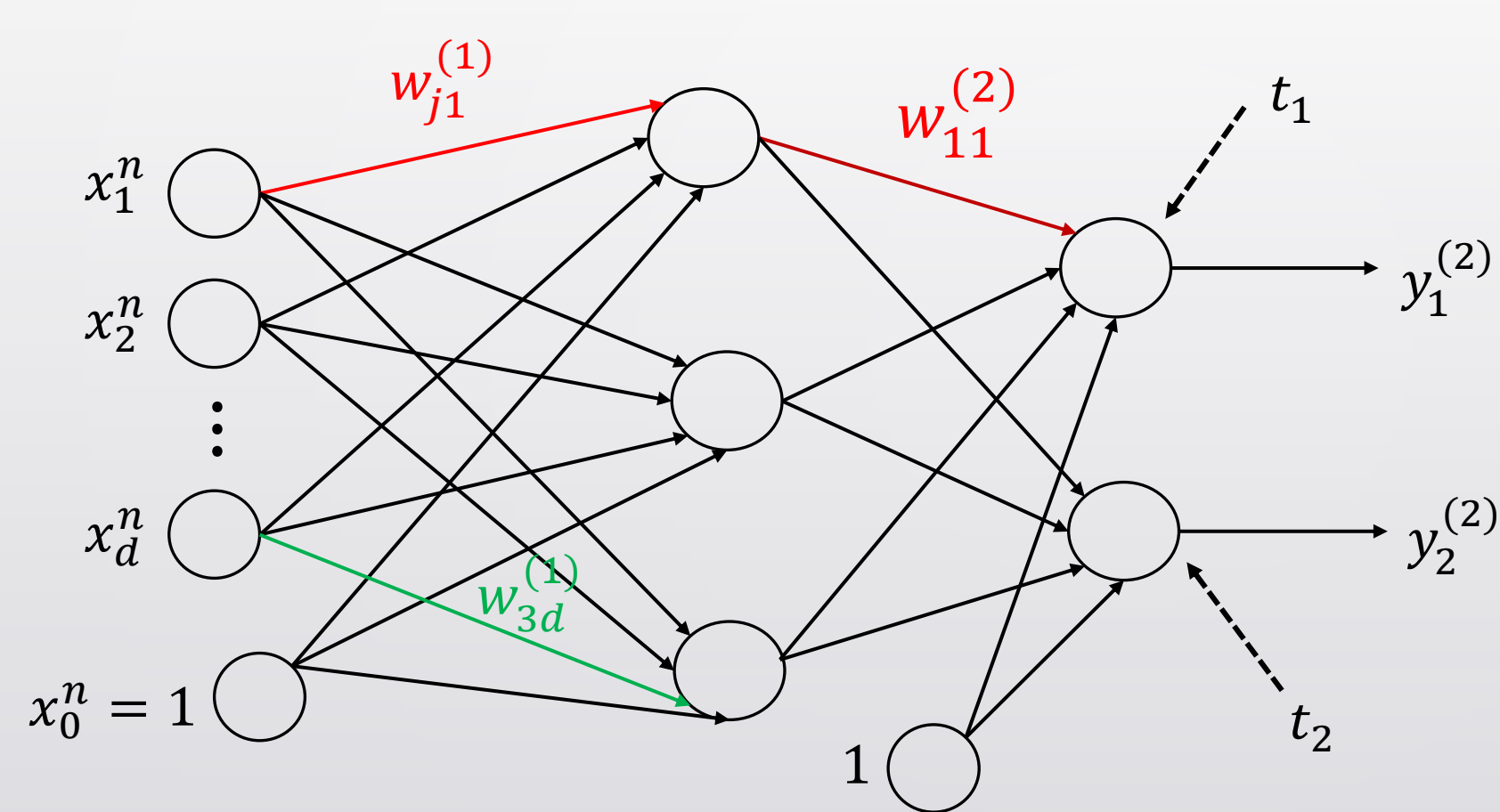
$\mathbf{z}$ : 潜在変数 Latent Variables

# 神経細胞の興奮 Neuronal Excitation

神経系の活動 = 神経細胞が電気活動が発生させ、神経細胞間で伝えていくこと  
Inter-neuronal transmission of electrical activity



# 多層パーセプトロン Multi-layered Perceptron



$w_{ji}^{(1)}$ : 入力層から隠れ層への重み  
Weights from input to hidden layer

$w_{kj}^{(2)}$ : 隠れ層から出力層への重み  
Weights from hidden to output layer

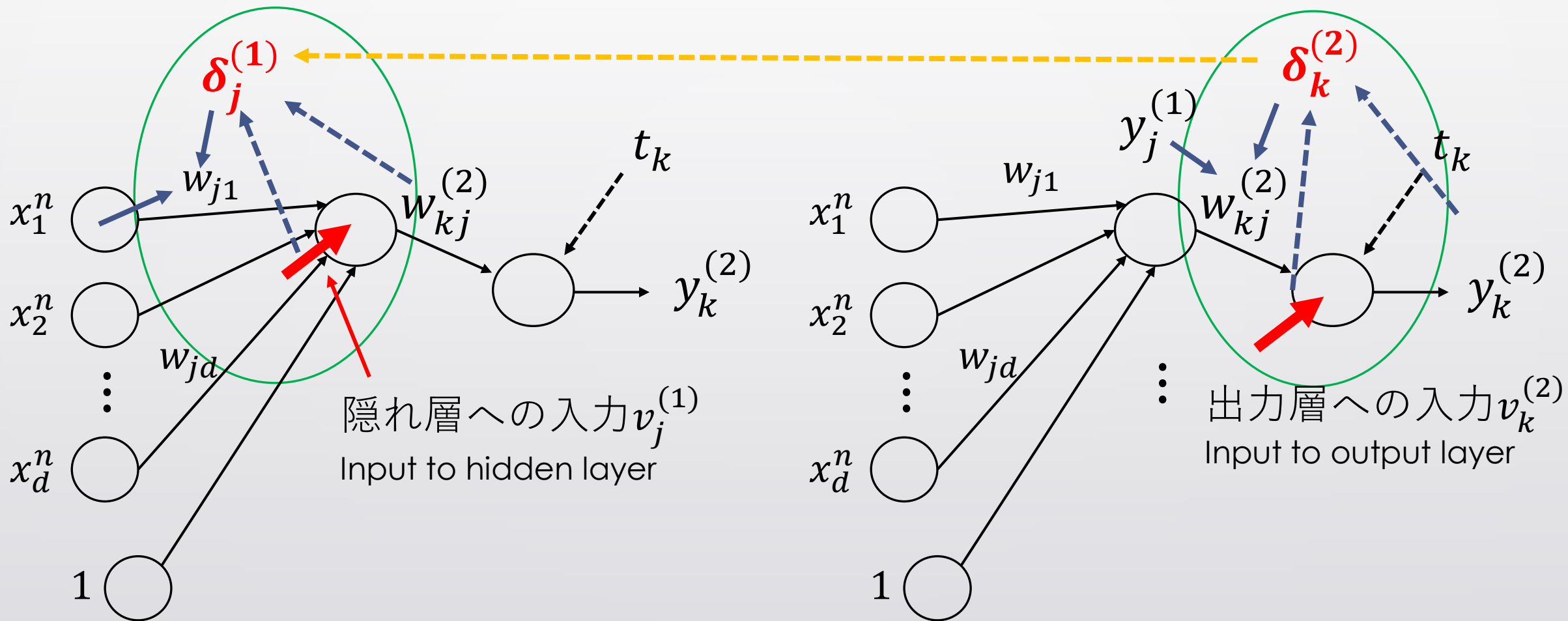
$i = 0, 1, 2 \dots d$

$j = 0, 1, 2 \dots M$

$k = 1, 2 \dots C$



## 重みの更新 Weight Updating

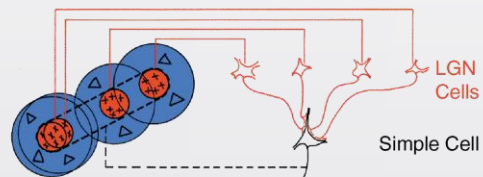




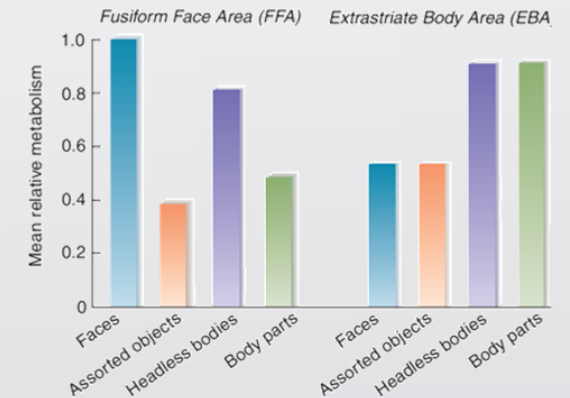
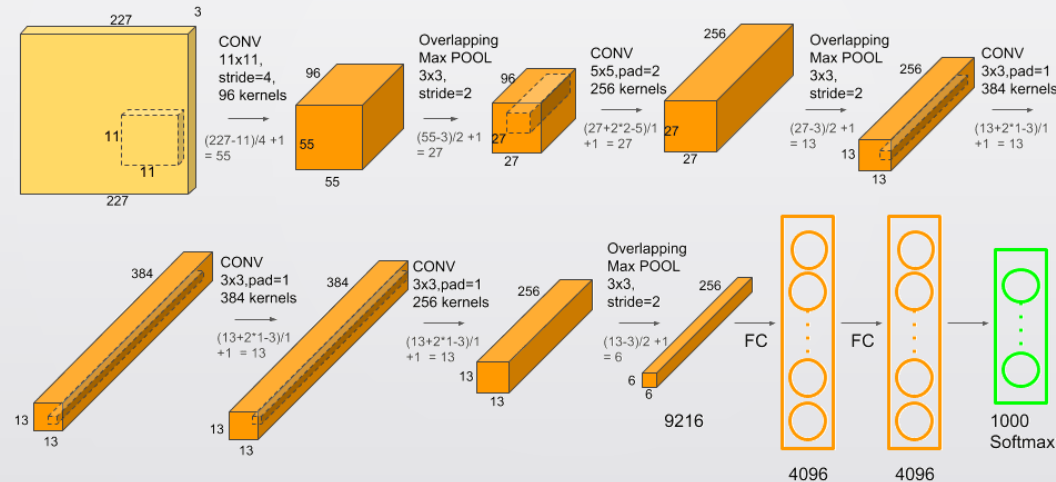
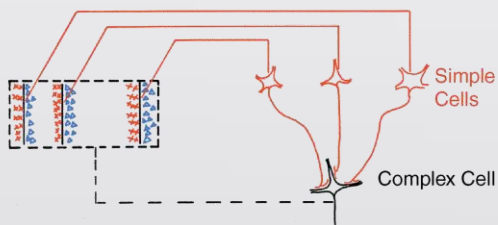
# CNNと脳の類似性 Similarity between CNN and Brain

プーリングの効果で受容野が広くなる  
Receptive field gets broader as a result of pooling

Circuit Building a Simple Cell from LGN Cells



Building a Complex Cell from Simple Cells

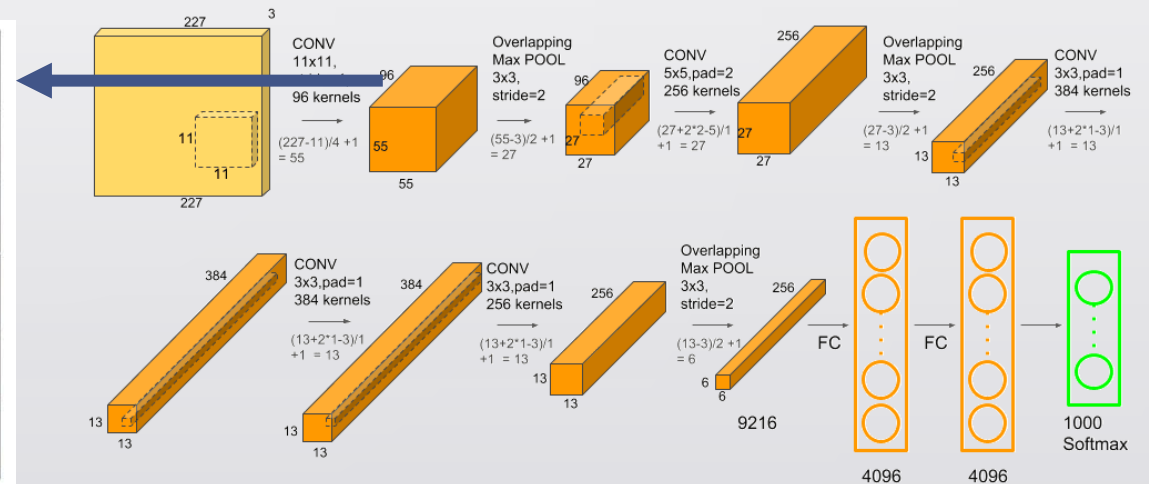
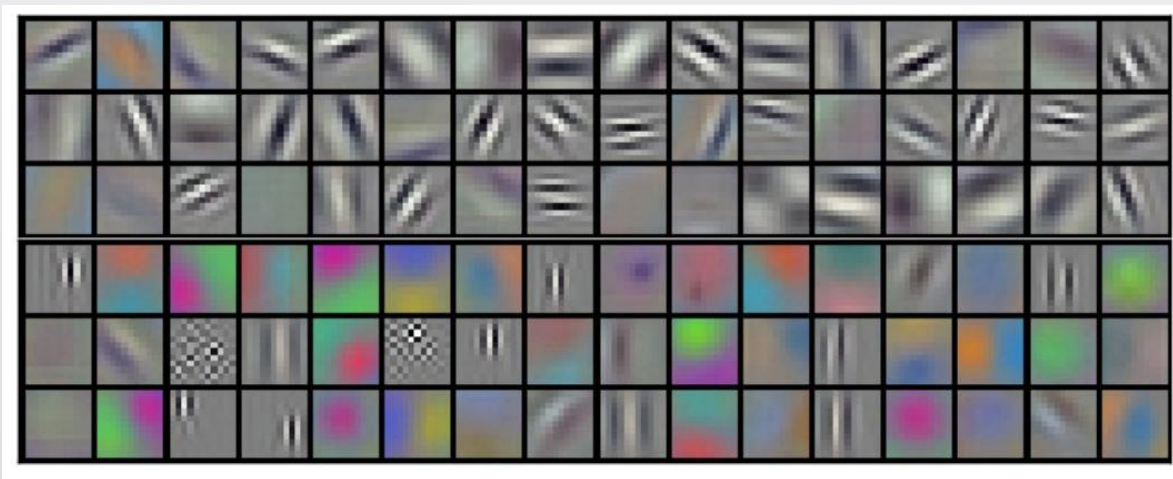


畳み込み層で複雑な情報を表現する  
More complex information is represented at convolution layers

# フィルターの学習 Acquisition of Filters

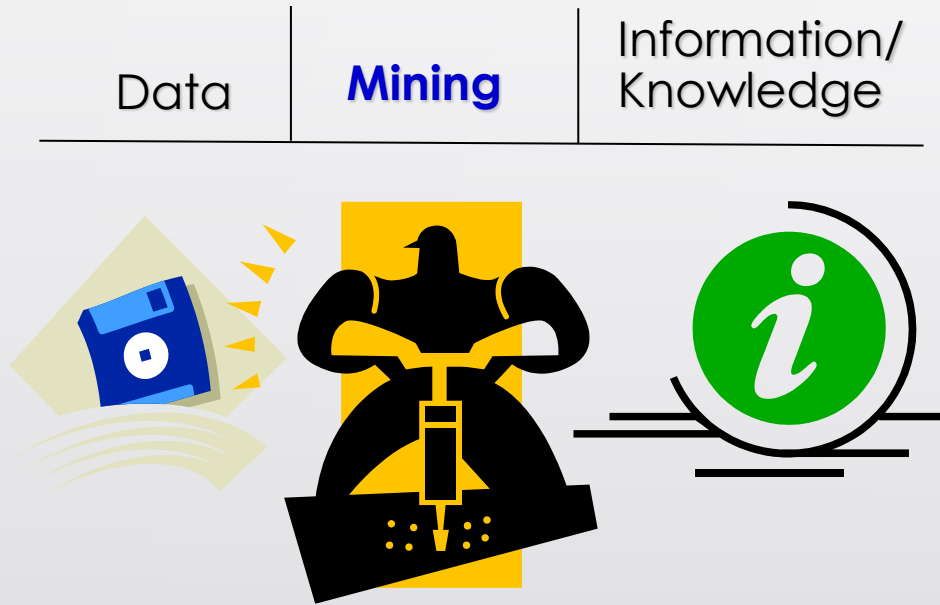
誤差逆伝搬法による学習の結果、Alex Netの第一の畳み込み層でガボールフィルターが獲得された

Weight updating by backpropagation led to acquisition of Gabor filter at the first convolution layer of Alex Net



////////////////////

データマイニング ≡ 金鉱の採掘  
Data Mining ≡ Gold Mining



<https://www.legendsofamerica.com/mining/>