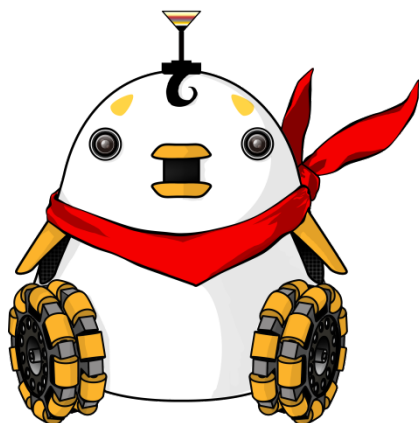


# 人工知能

## 第10回 学習と認識(1) クラスタリング

---

立命館大学 情報理工学部 知能情報学科  
萩原良信



# STORY 学習と認識(1)

- さて，迷路を探索し，通り抜ける方法もわかった．自分の位置を見失っても自己位置推定で思い出すことができる．ホイールダック2号はこれで大丈夫だと思った．
- 「さあ，お宝にとってゴールに向かうぞ！」
- しかし，ちょっと待てよ．「お宝」や「ゴール」って何だろう．「お宝」とはどんなもので「ゴール」ってどんな見た目なんだろう．ホイールダック2号は地図はわかるが，目の前に「お宝」や「ゴール」があったとしても，それが「お宝」や「ゴール」であることを認識することができない．まずは，「お宝」や「ゴール」とはどんなものなのか，学習していないと話にならない．



# 仮定学習と認識(1)

- ホイールダック2号は適切な画像特徴量を有限次元ベクトルとして取得できるものとする.



情報取得！



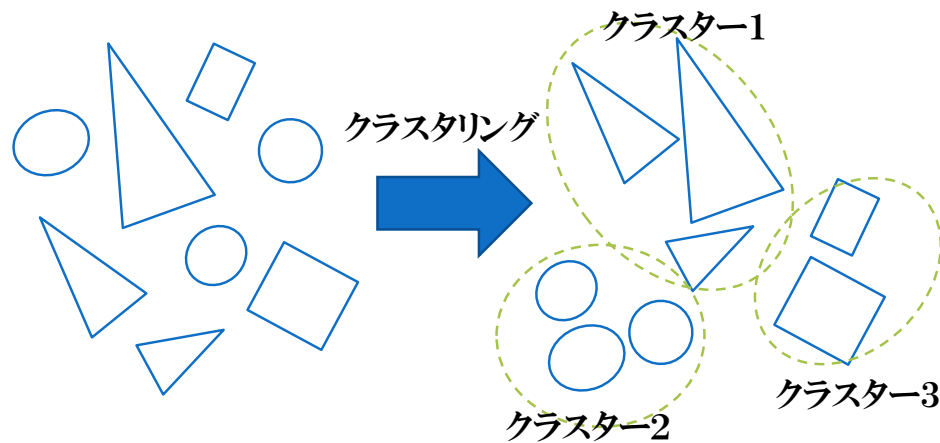
エンコーディング！  
(encoding)

# Contents

- 10.1 クラスタリング
- 10.2 K-means法
- 10.3 混合ガウス分布
- 10.5 低次元化

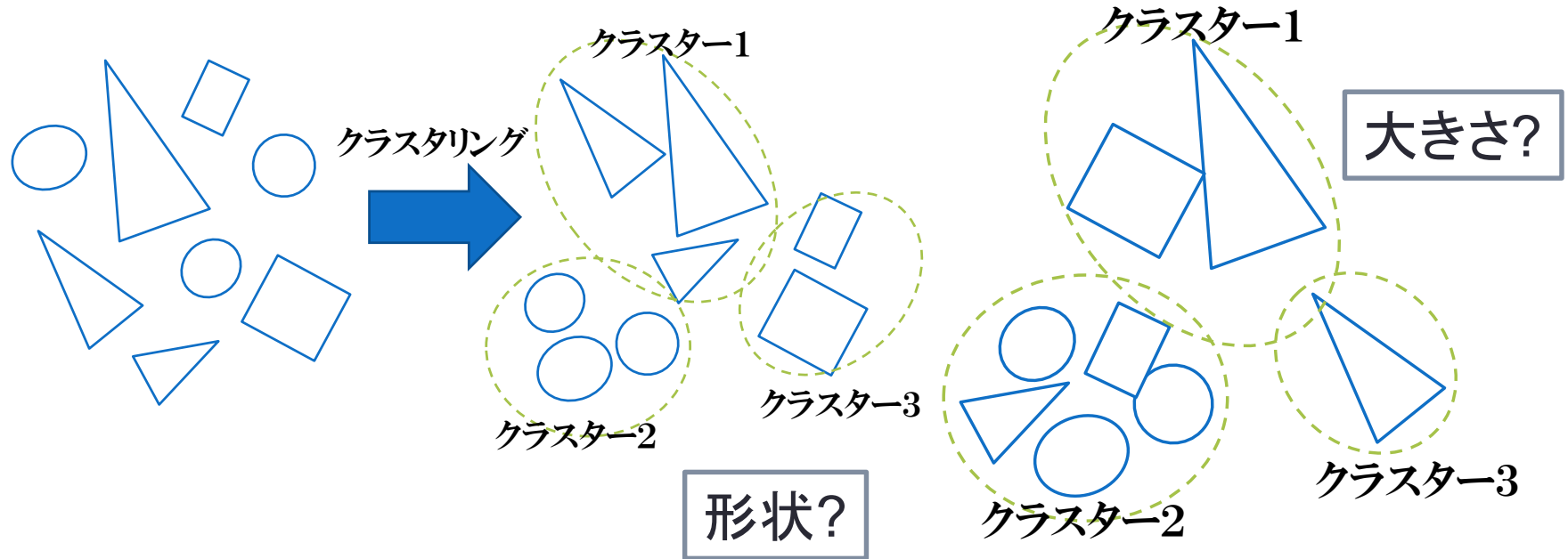
## 10.1.1 クラスタリングとは何か？

- データの集まりをデータ間の類似度にしたがっていくつかのグループに分類することをクラスタリング (clustering)という.
- この作業を自動化するのが機械学習におけるクラスタリングという種類に属する手法
- 自ら概念を獲得するロボットをつくらうとする場合にはクラスタリングは重要な要素技術になる.



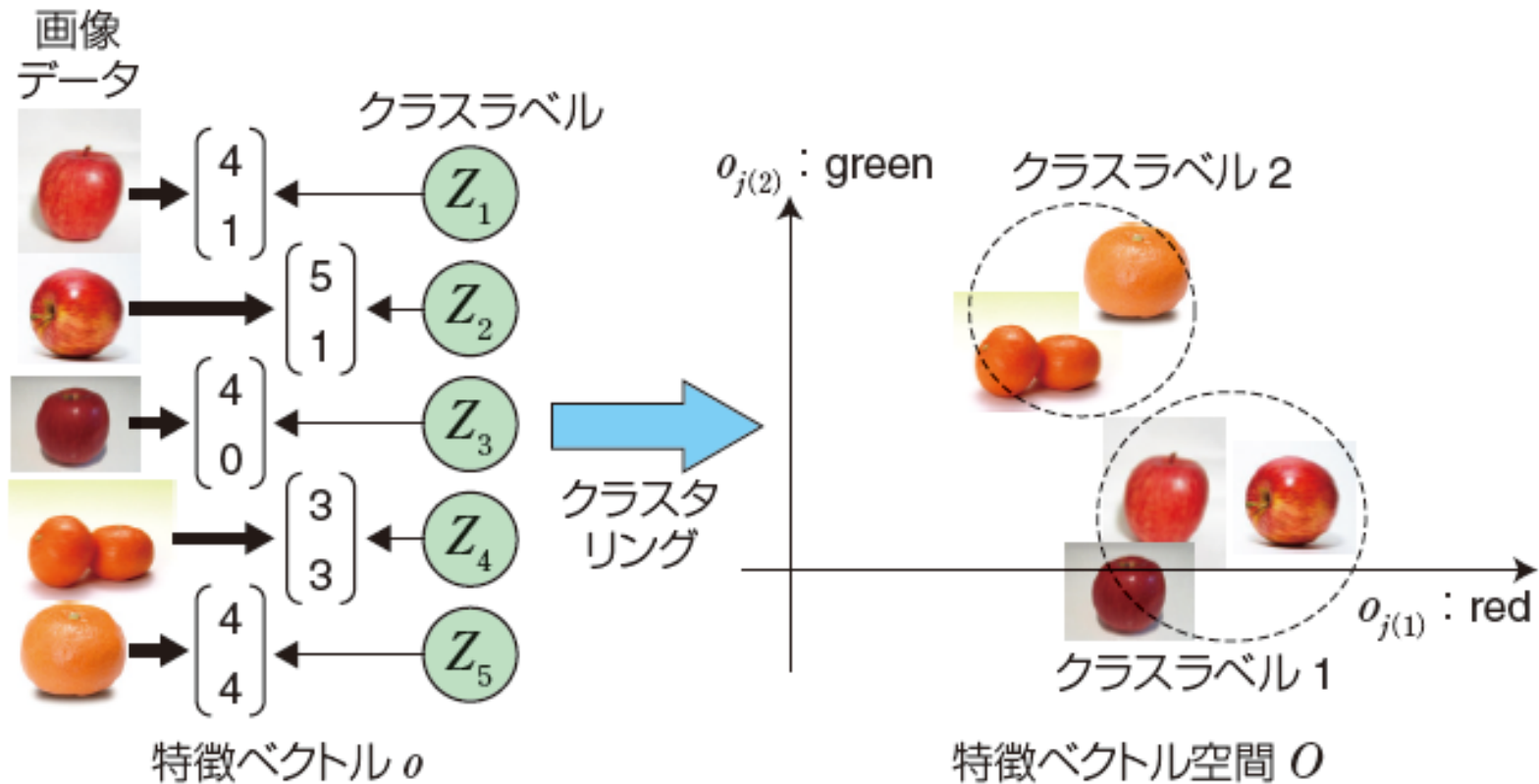
## 10.1.2 特徴抽出

### 「自然な」クラスタリングとは？



- ロボットにとってこのグループ分けが「自然な」ものであるかどうかは、ロボットにどのような基準を与えるかに依存する。
- そのような類似性を定義するために、特徴量や特徴ベクトルによって張られる特徴空間の設計が重要になる。

# 特徴量抽出とクラスタリング



対象が特徴空間上の点として表されると、クラスタリングは特徴空間上の点をグループ分けする数学的な問題になる。

# 教師なし学習

- 入力として与えられたデータに潜む知識を発見する方法
- クラスタリング
  - 大量のデータを幾つかのグループに自動的に分類する.
  - 分類問題を教師データを用いずに行う.
- 低次元化
  - 高次元のデータをより低次元な空間に写像することで, データを説明する少数のパラメータを発見する. または, 可視化する.



# Contents

- 10.1 クラスタリング
- 10.2 K-means法
- 10.3 混合ガウス分布
- 10.5 低次元化

## 10.2.1 K-means法のアルゴリズム

### Algorithm 10.1 K-means 法

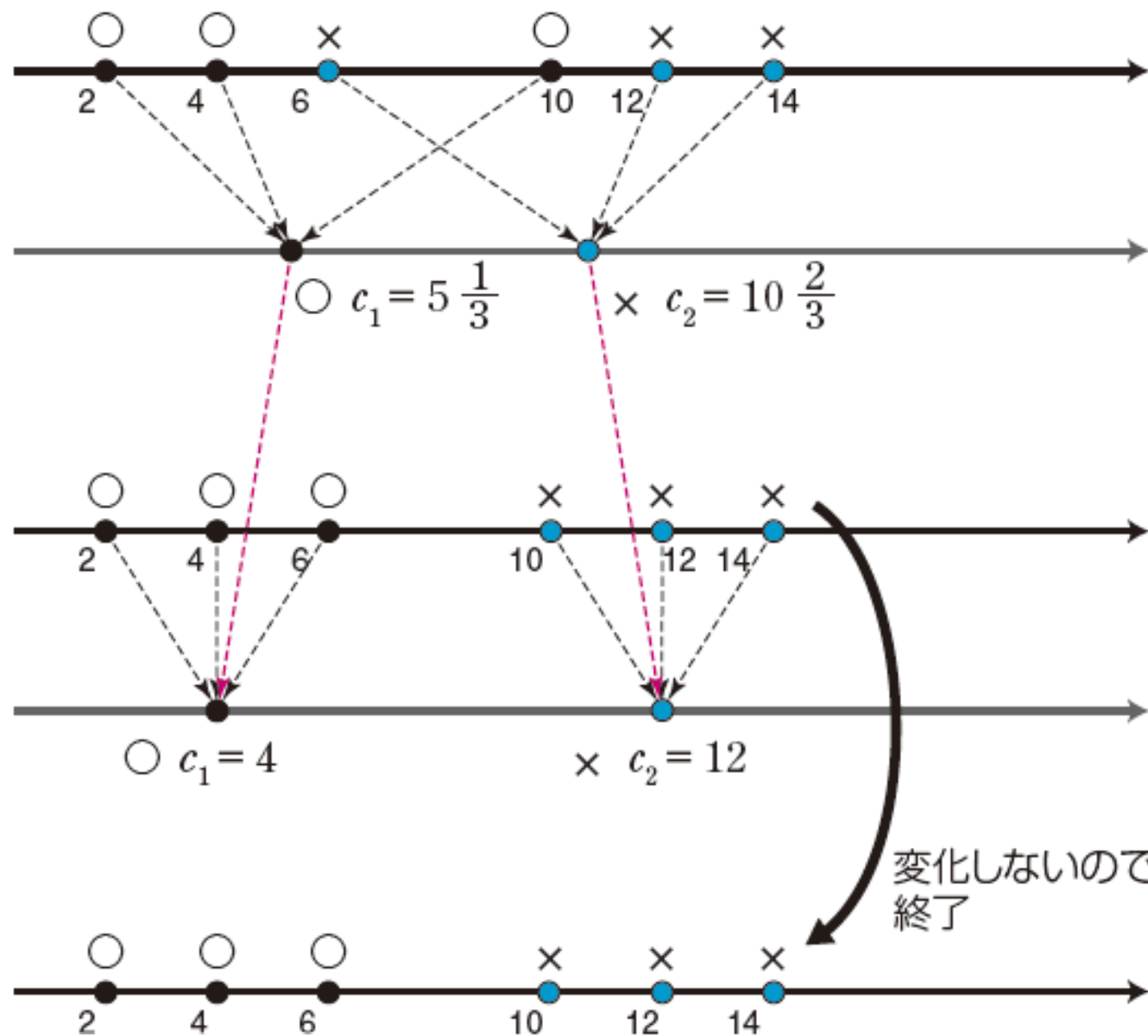
- ①  $K$  個のクラスタの代表点  $(c_1, c_2, \dots, c_K)$  を初期化する.
- ② repeat
- ③ 各データ  $o_i$  ( $i \in \{1, 2, \dots, N\}$ ) について,  $o_i$  と  $c_j$  の距離を  $d(x, y) = \|x - y\|^2$  で測り,  $o_i$  のクラスタラベルである  $z_i$  を  $o_i$  と最も近いクラスタ代表点  $c_j$  の添字  $j$  に更新する.

$$z_i \leftarrow \arg_j \min d(o_i, c_j) \quad (10.1)$$

- ④  $c_j$  を各クラスタに含まれるデータの重心値で更新する.
- ⑤ until すべてのクラスタの割り当て  $z_i$  が変化しなくなる.

## 10.2.2 K-means法の例

- $S=\{2,4,6,10,12,14\}$ という6個の一次元データがあったとする. これをk-means法を用いてクラスタリングする.
- 初期クラスターを $S_1=\{2,4,10\}$ ,  $S_2=\{6,12,14\}$ とした際に, k-means法のアルゴリズムを実行する.
- まず, 初めのステップで, 各クラスターの重心値は
  - $c_1 = (2 + 4 + 10)/3 = 16/3 = 5+1/3$
  - $c_2 = (6 + 12 + 14)/3 = 32/3 = 10+2/3$



変化しないので  
終了

## K-means法の実行例

○はクラスタ 1    × はクラスタ 2 を表す

# 演習10-1 K-means法とは？

- K-means 法の説明として最も不適切なものを選べ.
  - ① データを最も近いクラスタに帰属させ、その後にクラスタの代表点を更新する.
  - ② クラスタ内のデータとクラスタの代表点の距離の和を減少させる.
  - ③ クラスタの代表点を更新する際にはデータの重心値をとるのであって中央値をとるのではない.
  - ④ K 個の方法を組み合わせて学習を進行させる.

# Contents

- 10.1 クラスタリング
- 10.2 K-means法
- 10.3 混合ガウス分布
- 10.5 低次元化

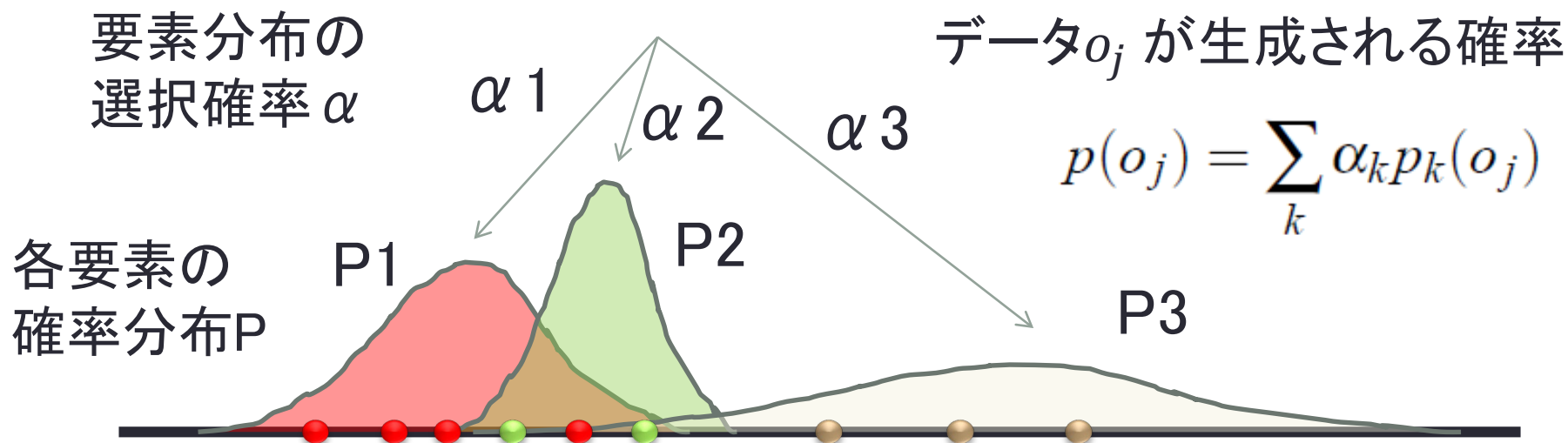
## 10.3.1 確率モデルに基づくクラスタリング

- K-meansでは境界が確定的なので、クラスタへの帰属度合いなどが議論しにくい.
- また、データがどのクラスタに属するかの判定が距離のみで判断されるために、クラスタごとにデータ分布の広がりが異なるようなデータを適切に分けることができない.
- 確率モデルに基づいたクラスタリングとして混合分布モデルに拠るアプローチがある.

⇒データが生成される確率を考える

## 10.3.2 混合分布モデルのデータ生成過程

- 混合分布モデルでは、データが、元々どのようなようにして生成されたデータであるか、というモデルを考えて、その生成過程をベイズの定理を用いて逆方向に推定することでクラスタリングを行う。
- 確率分布(クラスタに相当)が $K$ 個あり、 $k$ 番目の確率分布のもとで観測データ $o_j$ が生成される確率を $p_k(o_j)$ とする





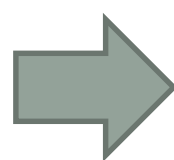
# ベイズ定理を用いた解釈

$$p(o_j) = \sum_k \alpha_k p_k(o_j)$$

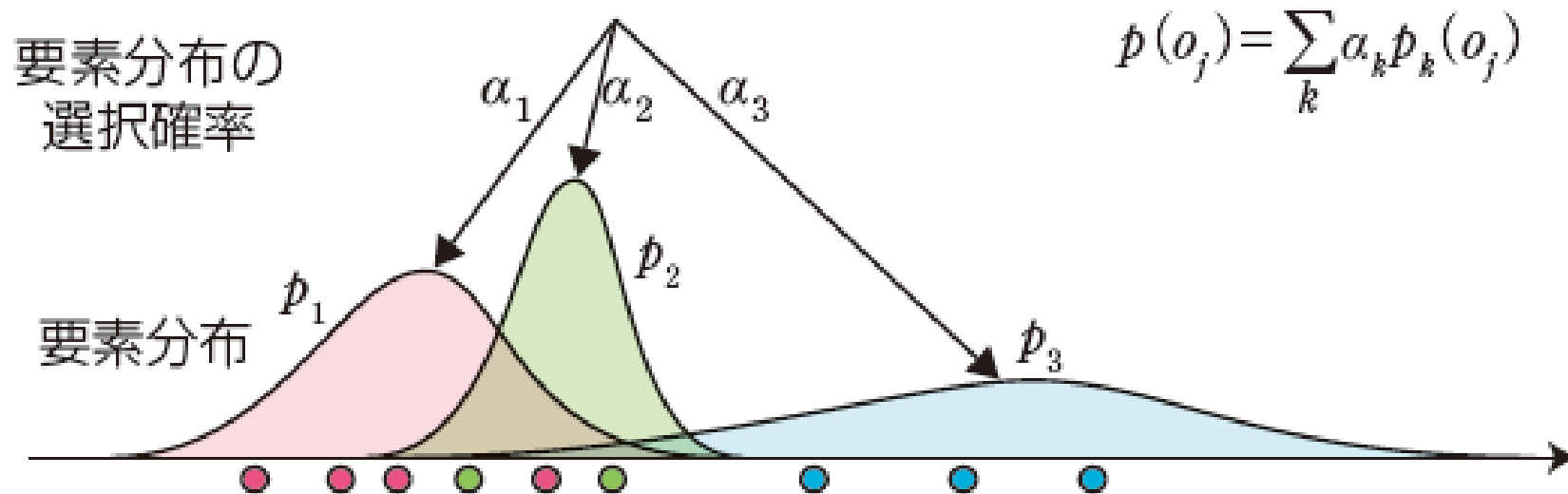
この時に $\alpha_k = P(k)$ であり, 条件付き確率の視点から書き換えれば, 上式は

$$p(o_j) = \sum_k P(o_j, k) = \sum_k P(k)P(o_j|k)$$

観測データ $o_j$ に対して $P(k|o_j)$ を求めるのがクラスタリングなのでベイズの定理より


$$P(k|o_j) = \frac{P(o_j|k)P(k)}{P(o_j)} = \frac{P(o_j|k)P(k)}{\sum_{k'} P(o_j|k')P(k')}$$

# 混合ガウス分布



- 混合ガウス分布

- 混合分布モデルで要素分布がガウス分布であるもの.
- 各要素分布が平均パラメータと分散パラメータを持つ.
- パラメータ更新がk-means法の重心の更新に相当する.

## 演習10-2 混合分布モデル

- 混合分布モデルでは、各データの各クラスタへの割り当ては何によって決定されるか。適切なものを選び。
  - ① 再帰的な最適化計算により漸近的にしか決定されえない。
  - ② ベイズの定理により事後確率を計算する事で決定される。
  - ③ クラスタの代表点への距離のみで決定される。
  - ④ マルコフ決定過程を用いて決定される。

## 演習10-3 確率的クラスタリング

$$p(o_j) = \sum_k \alpha_k p_k(o_j)$$

- 上の混合モデルが与えられた時に
- 観測 $o$ が与えられた際にこれがクラスター $k$ に属する確率 $p(k|o)$ を上にも用いた記号を使って示せ.

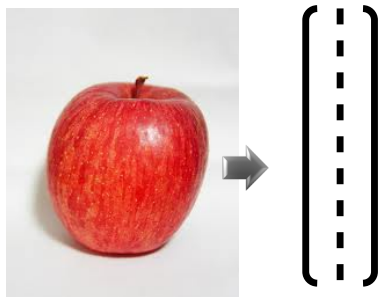
# Contents

- 10.1 クラスタリング
- 10.2 K-means法
- 10.3 混合ガウス分布
- 10.5 低次元化

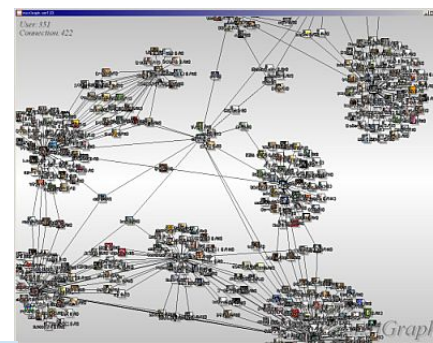
## 10.5.1 クラスタリングと低次元化

- クラスタリングと並ぶ教師なし学習の手法
- 高次元のデータをより低次元のベクトルで表現するのが低次元化の手法である。

特徴ベクトル抽出



可視化



データ圧縮

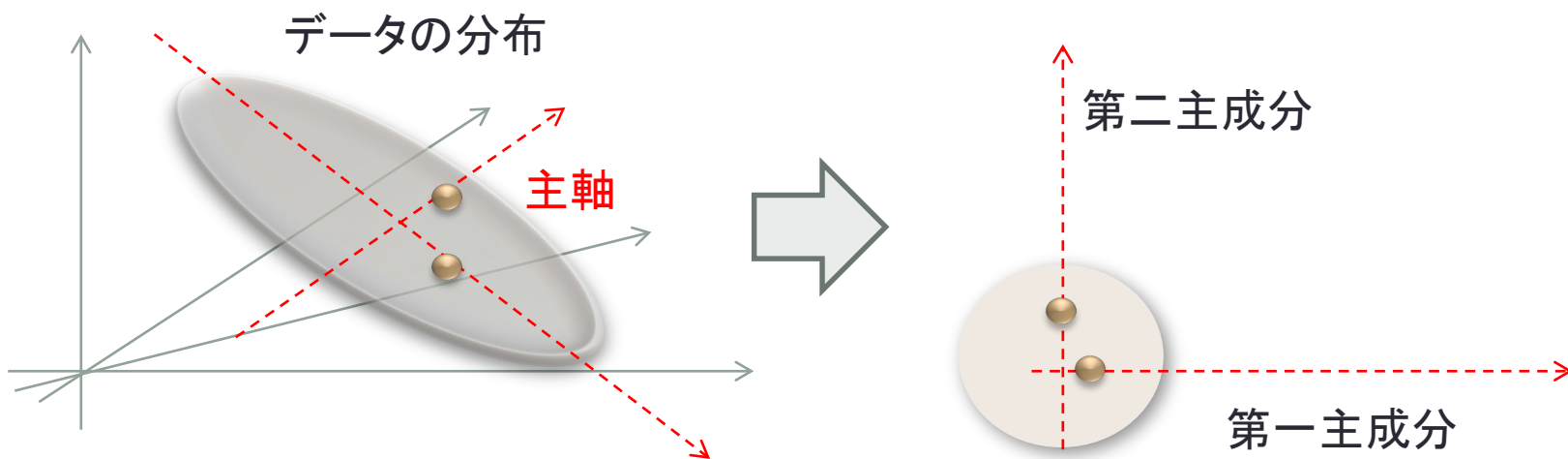


ソーシャルネットワークグラフ

[twitter mention map](#)

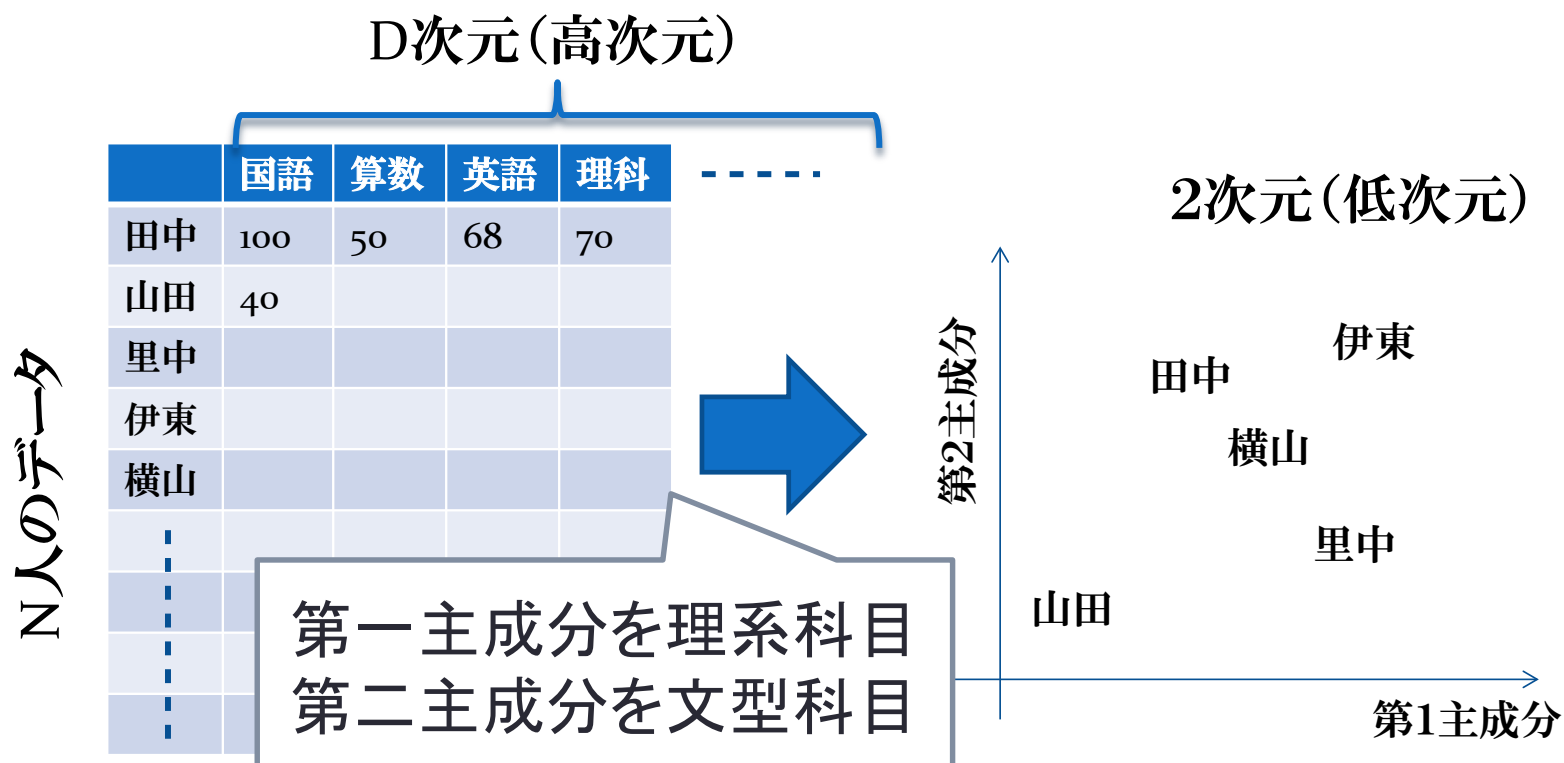
# 主成分分析

- 主成分分析は具体的にはデータが高次元空間上でガウス分布をしていると仮定して、その分布の主軸方向（最も分散の大きい方向）を発見し、それを第1主成分とする。その後、その次に分散の大きい軸をとるというように、順次、軸をとっていくことで、低次元空間を得ていく。



# 主成分分析の例

- $N = 1000$  人の学生が  $D = 30$  科目の授業の履修を終えて、それぞれに100点満点の成績を得たとする.
- 30次元のデータを最も上手く表現できるような低次元の表現を得る.



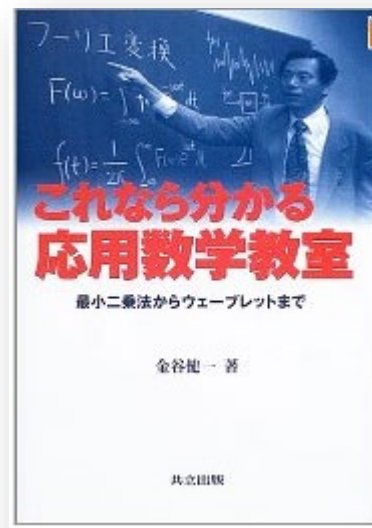


# 様々な低次元化手法

- 主成分分析
- 独立成分分析
- カーネル主成分分析
- MDS (多次元尺度法)
- 自己組織化マップ(SOM)
- GPLVM
- Deep Belief Network

Deep Learning が2011年ごろから  
音声認識, 画像認識で圧倒的最高性能を叩  
きだして, 現在, Deep Learning ブーム

主成分分析を学ぶなら  
とりあえず, これなど...



これなら分かる応用数学教室  
—最小二乗法からウェーブレッ  
トまで, 金谷 健一

## 10.5.5 深層学習(deep learning)

- 深層学習(deep learning) は2010 年代に入ってから急速に注目されている低次元化手法であり, 主にパターン認識のための特徴ベクトル抽出に用いられている. 音声認識や画像認識で非常に高い性能を出すことに貢献している.

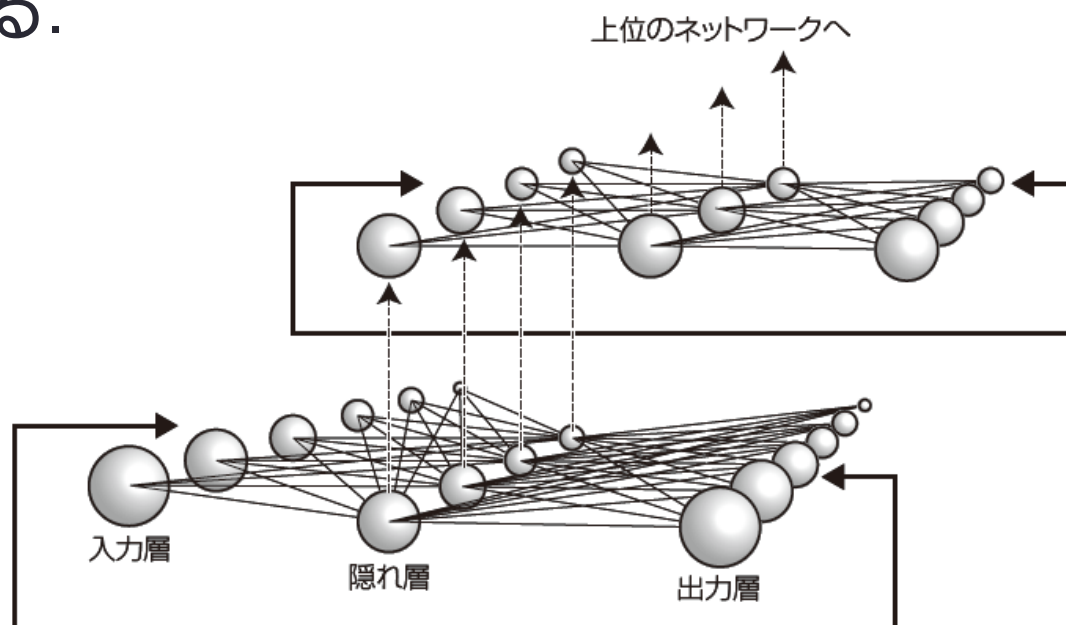
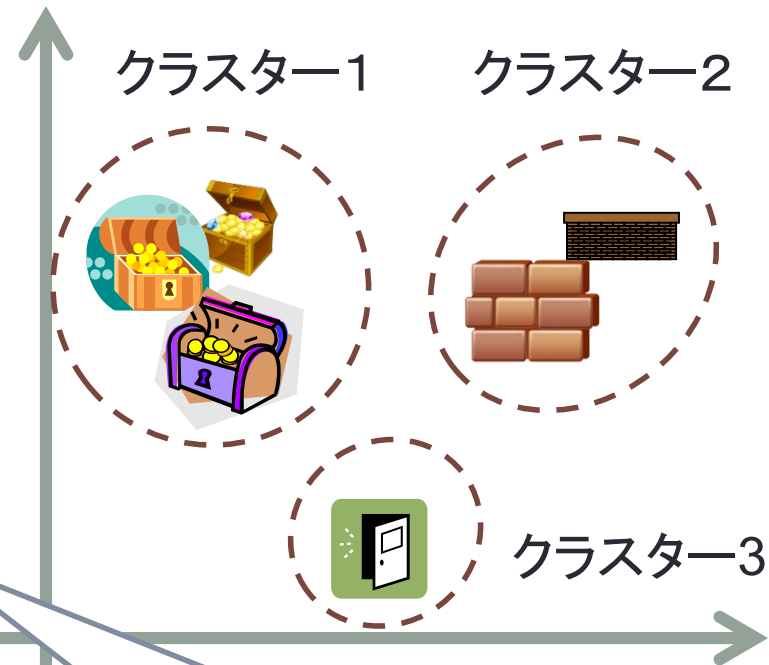
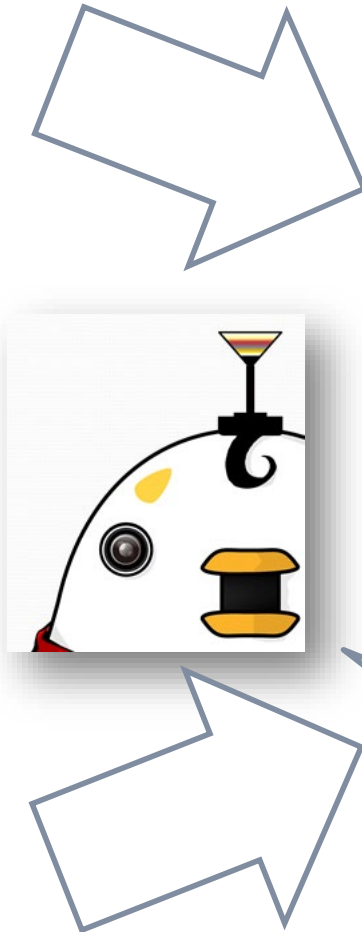
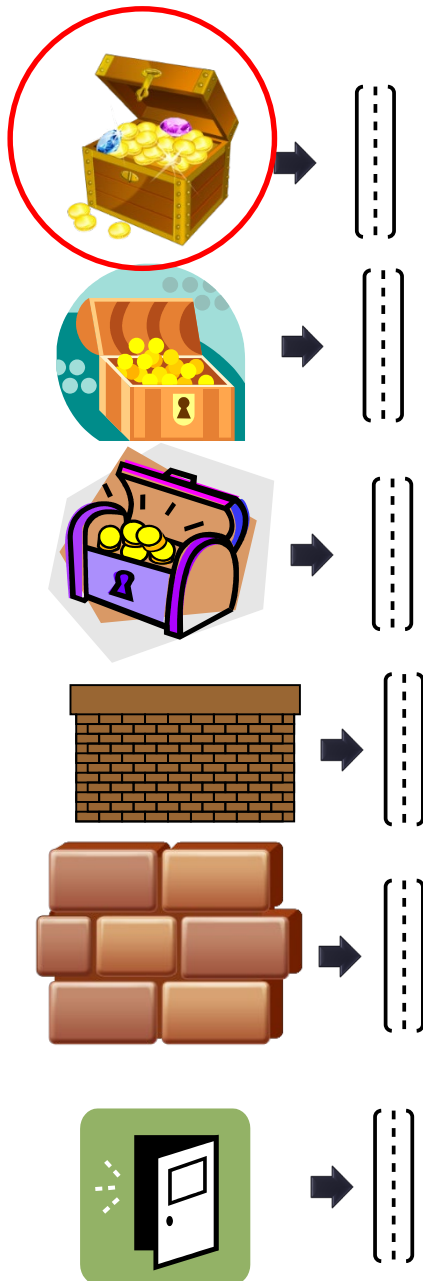


図 10.10 AutoEncoder による深層学習

# 宝箱という 知識を得る



”name[クラスター1]”に  
分類されるから宝箱だ！

# 演習10-4 低次元化手法

- 低次元化手法が用いられる目的として最も一般的では無いものを選べ。
  - ① データ圧縮
  - ② 可視化
  - ③ 特徴量抽出
  - ④ 強化学習

# まとめ

- クラスタリングの基礎について学んだ.
- K-means 法のアルゴリズムを学び, 簡単な数値例を通じてその動作を確認した.
- 混合ガウス分布によるクラスタリングの概略について学んだ.
- 低次元化手法の概要について学び, その代表的な手法である主成分分析, 深層学習の概要を知った.