



# 情報アクセス論 第8回

## 「分類・クラスタリング」



# 分類とクラスタリング

(Search Engines: Information Retrieval in Practice, Pearson Education, 2010)

- 情報をカテゴリ別に分類する技術
- パターン認識や機械学習の古典的な問題
- **(自動)分類** (classification)
  - 「ある文書がどのカテゴリ(クラス)に属するか」
  - **教師あり学習**のタスク
- **クラスタリング** (clustering)
  - 「多数の文書群をどうグループ化するか」
  - **教師なし学習**のタスク
- 「文書」以外にもさまざまな対象に適用可能

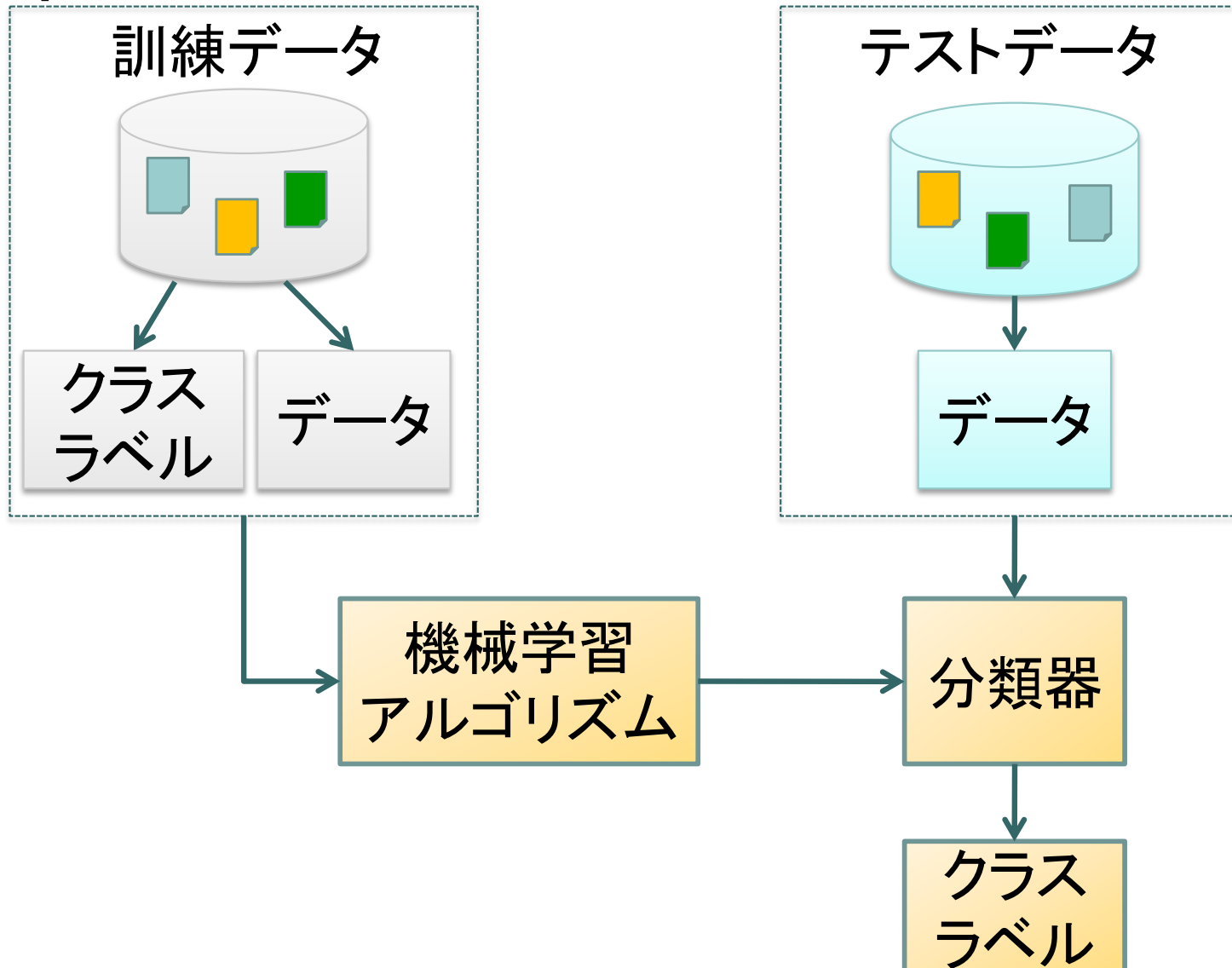
# 人間による分類の手順

- 食品を「健康性」によって分類することを考える
  - 健康性の指標となる「特徴」を見つける
    - ・ 脂肪分, コレステロール, 糖分, 塩分, etc.
  - 食品から特徴を抽出する
    - ・ 栄養成分表を見る, 化学分析を行う, etc.
  - 特徴から得られる「証拠」から仮説を立てる
    - ・ 健康に関する特徴から「健康性の指標」を得る

例: 
$$H(f) \approx w_{fat} fat(f) + w_{chol} chol(f) + w_{sugar} sugar(f) + w_{sodium} sodium(f)$$

- 最終的に, 証拠に基づいて食品进行分类する
  - ・ 「健康性の指標」が一定の値以上であれば, その食品は健康に良いとみなす

# 機械学習による分類の手順



# 単純ベイズ分類器

(Naïve Bayes Claasifier)

- ベイズの定理に基づく確率的分類器

$$P(C | D) = \frac{P(C)P(D | C)}{P(D)}$$

- $C$ はクラスに対応する確率変数
- $D$ は入力(たとえば文書)に対応する確率変数
- 右辺の分母はクラスに依存しないため、無視できる

# ● ● ● | 単純ベイズ分類器による分類

- 文書は以下のように分類される

$$\begin{aligned}\text{Class}(d) &= \arg \max_{c \in C} P(c | d) \\ &= \arg \max_{c \in C} P(c)P(d | c)\end{aligned}$$

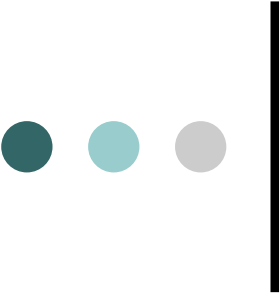
- $P(c)$  と  $P(d | c)$  を推定する必要
  - $P(c)$  はクラス  $c$  を観測する確率
  - $P(d | c)$  はクラスが  $c$  である時に  $d$  を観測する確率

## ● ● ● | $P(c)$ の推定

- $P(c)$ はクラス  $c$  を観測する確率
- **学習データ**のうちクラス  $c$  に属する文書数の割合で推定できる

$$P(c) = \frac{N_c}{N}$$

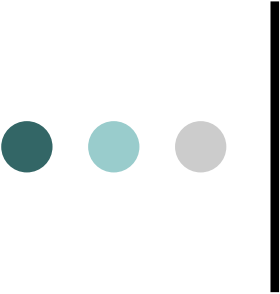
- $N_c$  は学習データのうちクラス  $c$  に属する文書数
- $N$  は学習データに含まれる全文書数



# $P(d \mid c)$ の推定

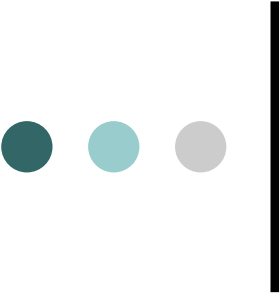
- $P(d \mid c)$ はクラスが  $c$  である時に  $d$  を観測する確率
- 推定は, 文書の表現に用いる**事象空間**に依存
- 事象空間とは?
  - ある確率変数に対して起こりうるすべての事象
    - ・ たとえば, コイン投げの確率変数の場合, 事象空間は  $S=\{\text{表}, \text{裏}\}$
  - 文書の場合, 文書中に現れる各単語が確率変数と考えられる





# 多変数ベルヌーイ事象空間

- 2値ベクトルで文書を表現
  - 索引語中の各語について1つの次元
  - 文書中に単語  $i$  が出現すれば  $i=1$ , しなければ  $i=0$ 
    - 出現回数は考慮しない
- 多変数ベルヌーイ事象空間は, 2値ベクトル上の分布をモデル化するのに適する
- 古典的な確率モデルによる検索(第6回で説明)に用いられるのと同じ事象空間



# 2値ベクトルによる文書の表現 の例

## ○ 英語のスパムメールの判定の例

document <i>id</i>	cheap	buy	banking	dinner	the	<i>class</i>
1	0	0	0	0	1	not spam
2	1	0	1	0	1	spam
3	0	0	0	0	1	not spam
4	1	0	1	0	1	spam
5	1	1	0	0	1	spam
6	0	0	1	0	1	not spam
7	0	1	1	0	1	not spam
8	0	0	0	0	1	not spam
9	0	0	0	0	1	not spam
10	1	1	0	1	1	not spam

# ● ● ● $P(d | c)$ の推定

- $P(d | c)$ は以下のように計算される

$$P(d | c) = \prod_{w \in V} P(w | c)^{\delta(w, d)} (1 - P(w | c))^{1 - \delta(w, d)}$$

文書集合中の  
すべての単語

デルタ関数(単語  $w$  が文書  $d$  に  
出現したとき1, それ以外は0)

- $P(w | c)$ は以下のように推定できる

$$P(w | c) = \frac{df_{w, c} + 1}{N_c + 1}$$

クラス  $c$  に属する文書  
のうち  $w$  を含む文書数

クラス  $c$  に属する文書数



# クラスタリング

- 文書集合の潜在的な構造を見つけるための教師なし学習アルゴリズム
- 類似文書のグループ(クラスタ)を見つけるのが目的
- たとえば, たくさんの果物の形状, 色, ビタミンC含有量, 価格の情報が与えられたとき, これらの果物をどのようにグループ化するか?
  - どのような基準を用いるか?
  - 類似度をどう定義するか?
- クラスタリングは, 対象の表現や類似度の定義に大きく影響を受ける

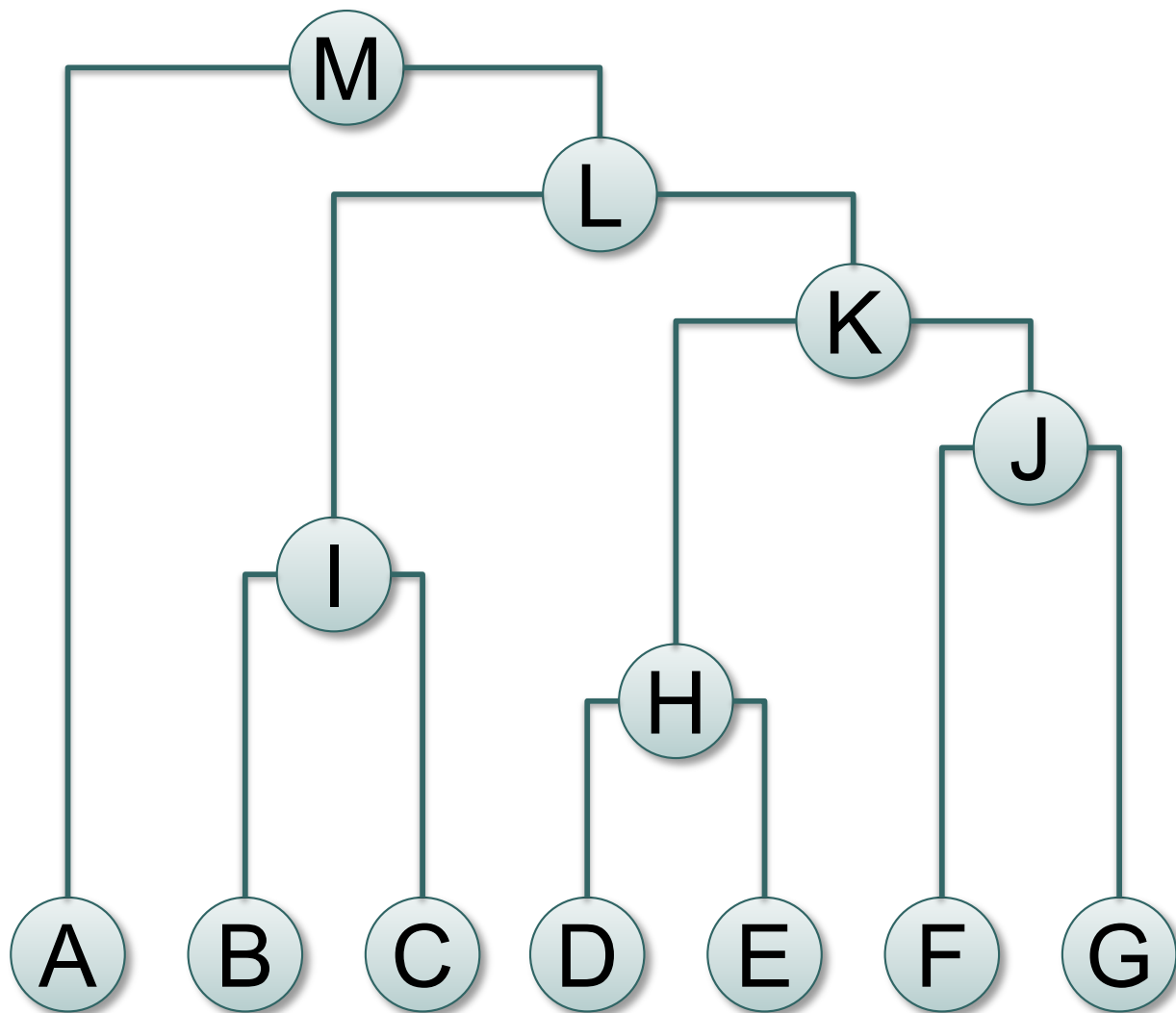


# 階層的クラスタリング

- クラスターの階層を作成する
  - 階層の最上位はすべてを含む一つのクラスター
  - 階層の最下位は個々の対象からなるN個（対象の個数）のクラスター
- 階層的クラスタリングの方法は2種類ある
  - 分割型（トップダウン）
  - 凝集型（ボトムアップ）
- 階層構造は**デンドログラム**（樹形図）で可視化できる



# デンドログラムの例





# 分割型と凝集型

## ○ 分割型

- すべての対象を含む一つのクラスタから開始
- 単一の対象からなるクラスタだけになるまで...
  - 既存のクラスタを二つの新しいクラスタに**分割**

## ○ 凝集型

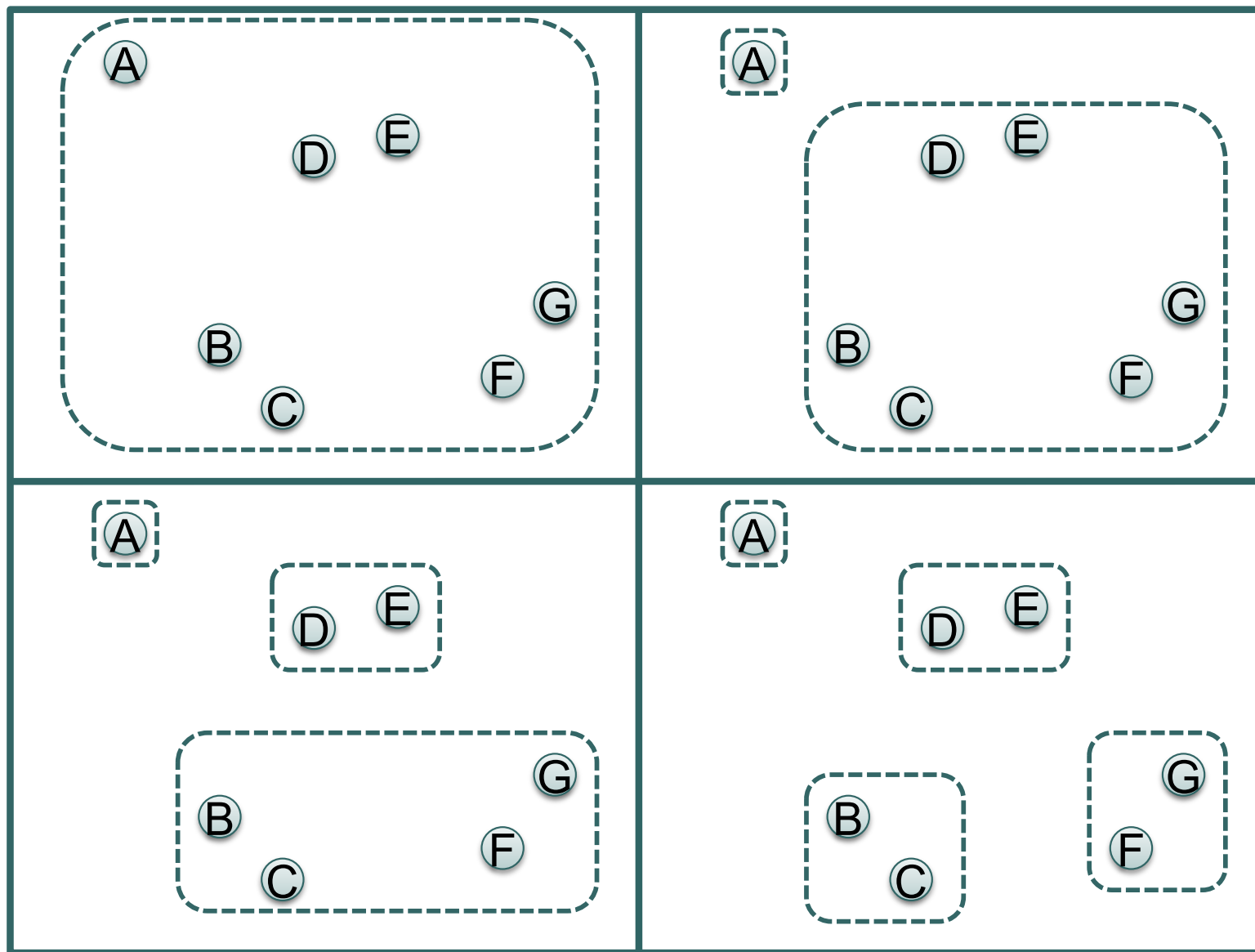
- 個々の対象からなるN個のクラスタから開始
- クラスタが一つになるまで...
  - 二つの既存のクラスタを一つに**結合**

## ○ どのようにクラスタを分割／結合すれば良いか？

- 分割／結合の**コスト**を定義
- 最も低コストな分割／結合を行う

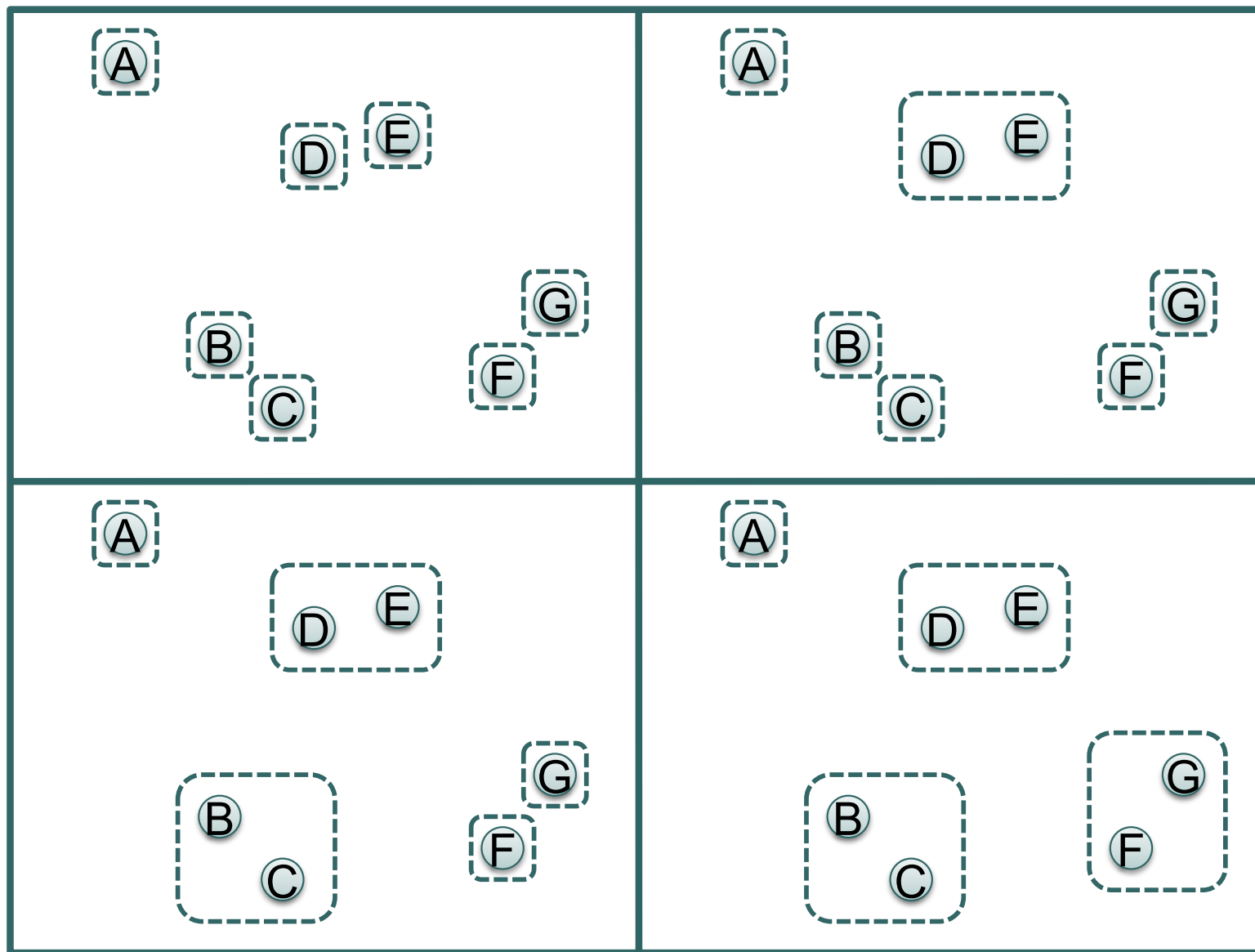


# 分割型の階層的クラスタリング





# 凝集型の階層的クラスタリング



# ● ● ● クラスタリングのコスト関数

- **単連結法** (2つのクラスタ中で最も近い対象の間の距離)

$$COST(C_i, C_j) = \min\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

- **完全連結法** (2つのクラスタ中で最も遠い対象の間の距離)

$$COST(C_i, C_j) = \max\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

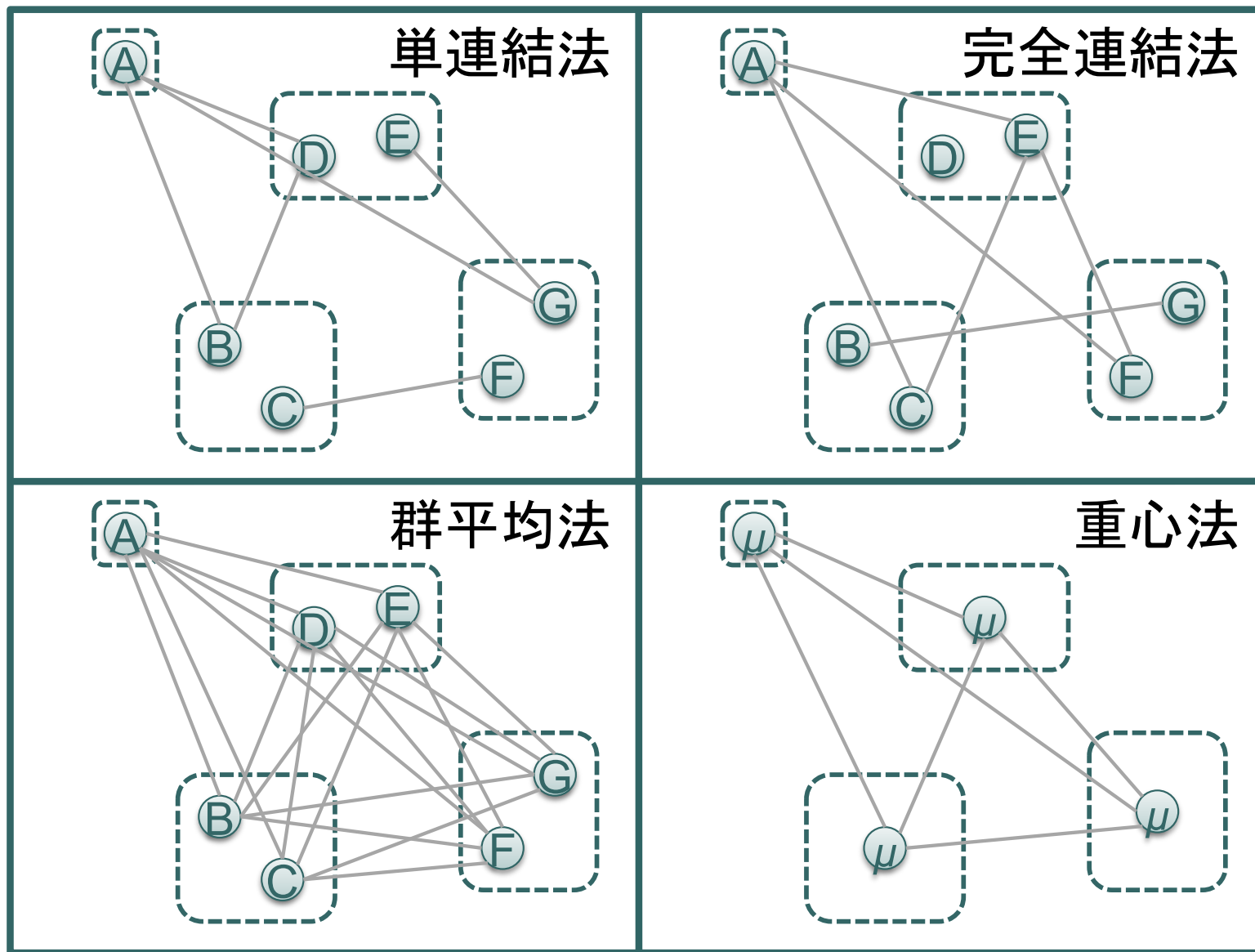
- **群平均法** (2つのクラスタ中の全対象の間の距離の平均)

$$COST(C_i, C_j) = \frac{\sum_{X_i \in C_i, X_j \in C_j} dist(X_i, X_j)}{|C_i||C_j|}$$

- **重心法** (2つのクラスタの重心間の距離)

$$COST(C_i, C_j) = dist(\mu_{C_i}, \mu_{C_j})$$

# クラスタリングのコスト関数の比較

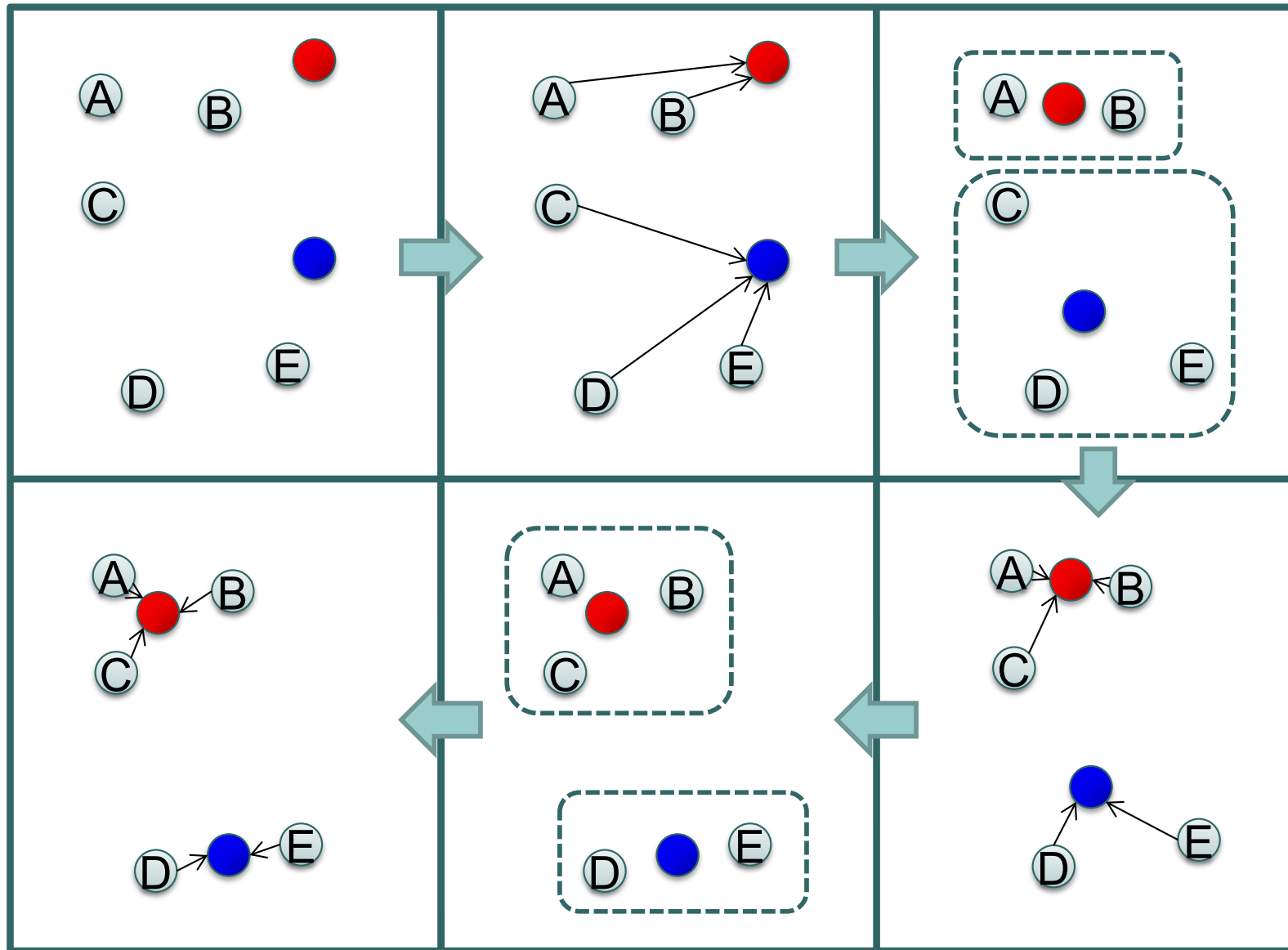


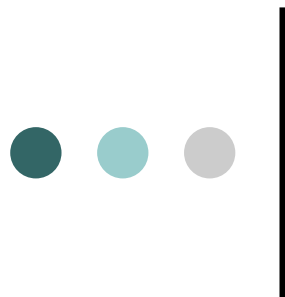


# K平均法(K-means)

- 非階層型クラスタリングの代表的な手法
- 常にK個のクラスタを維持する
  - クラスタは「重心(centroid)」で表される
- 基本的なアルゴリズム：
  - Step 0: K個の重心をランダムに選択
  - Step 1: 各点を最も近い重心に割り当てる
  - Step 2: 各クラスタの重心を再計算
  - Step 3: Step 1に戻る
- 比較的早く収束する
- 最初に選択する重心によって結果が変わる
- クラスタ数(K)を最初に決めなければならない

# K平均法によるクラスタリング の例 (K=2)





# まとめ

- 情報をカテゴリ別に分類する技術として, 自動分類およびクラスタリングの手法を説明した
- 自動分類の確率的手法の一つである単純ベイズ分類器について説明した
- クラスタリングの代表的な手法として, 階層的クラスタリングとK平均法について説明した