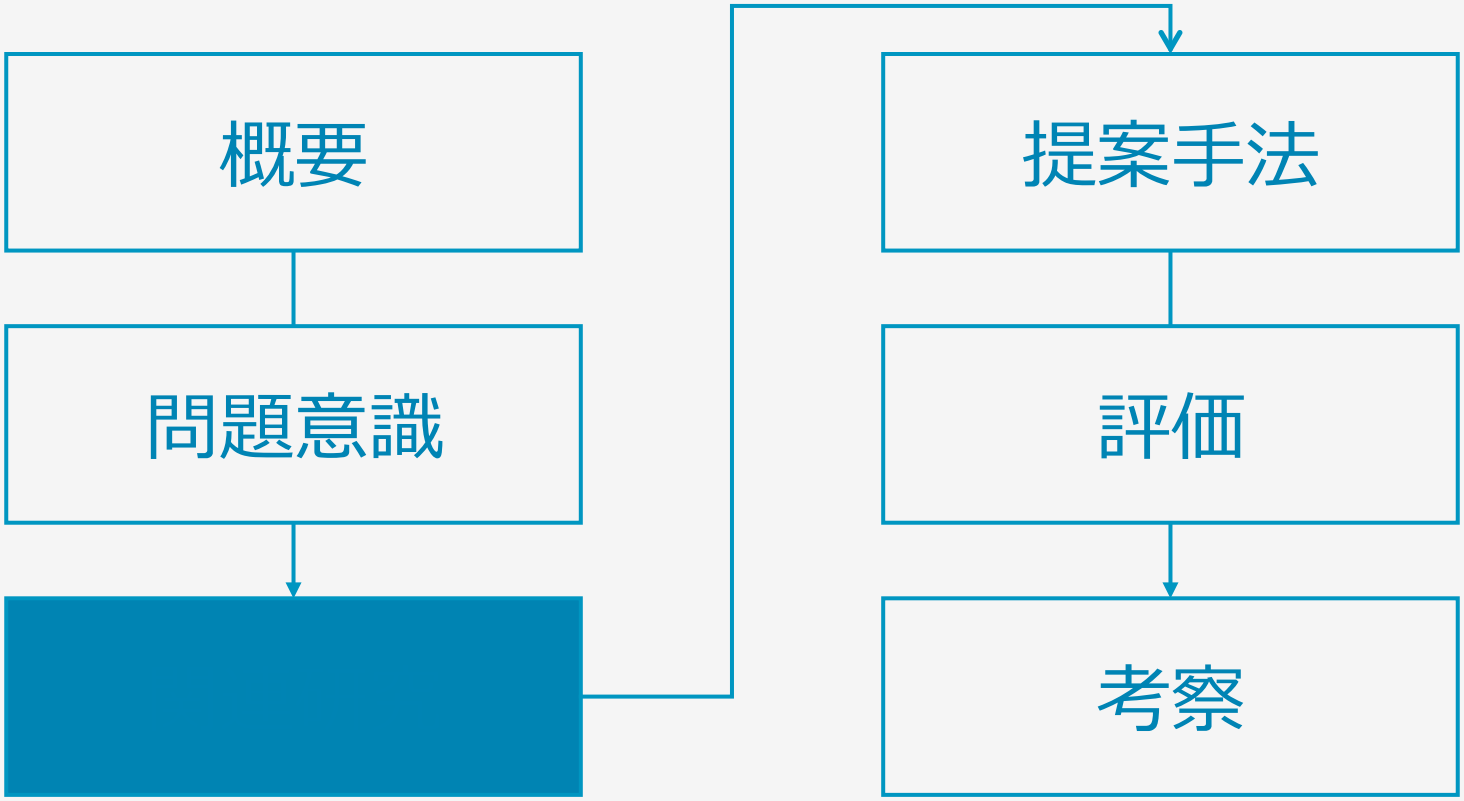
画像データの学習クラスタリング

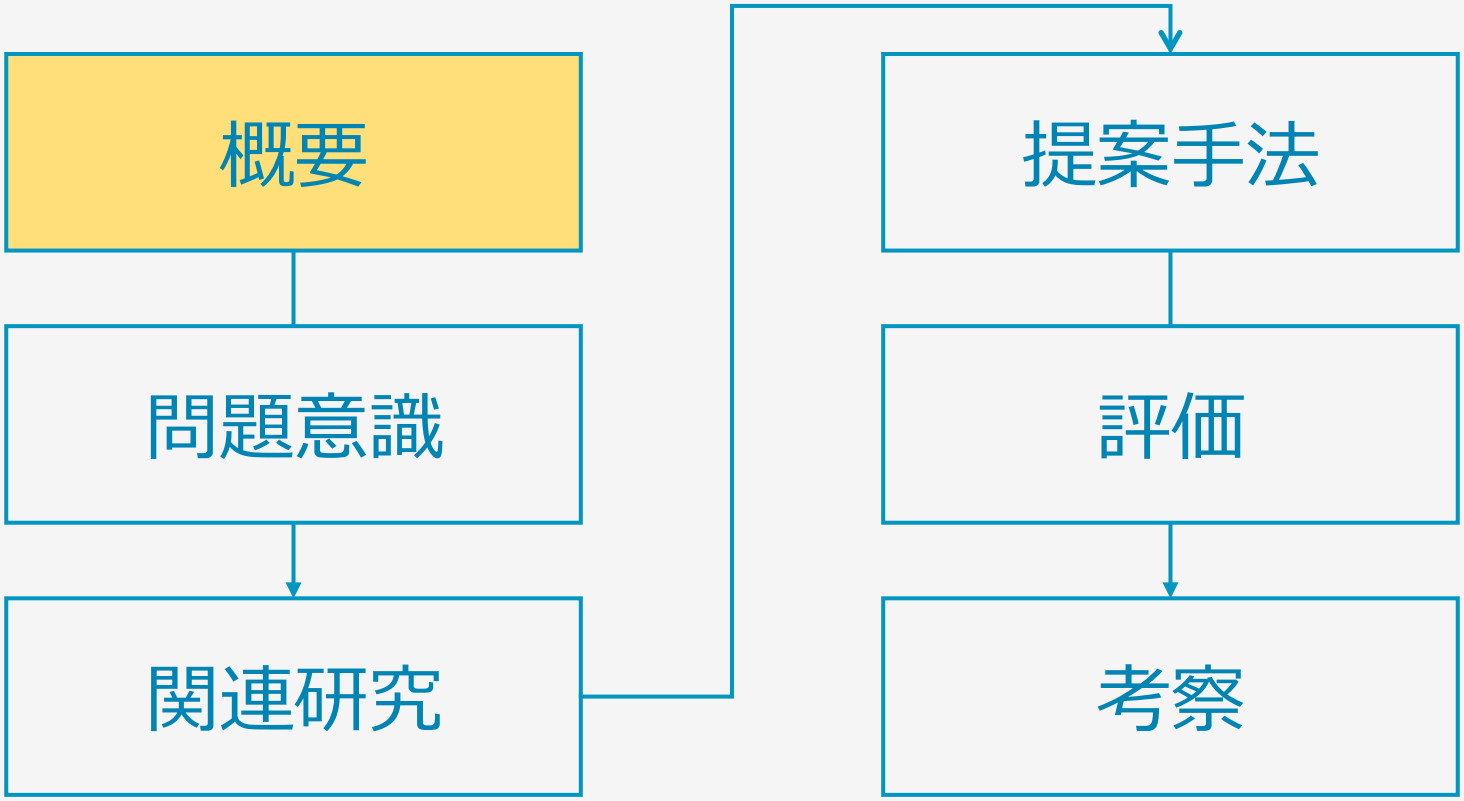
ITシステムプロジェクト

政策・メディア研究科 修士1年
笹本 将平

目次



目次



概要

一言で言うと、

ユーザーが意図するクラスターを得るための
クラスタリングパラメーター調節法を提案する。

クラスタリングとは

クラスタリング

あるデータの集合に対して分類をし、
似ているもの同士をグルーピングする手法。

- 一般的には、教師なしデータを用いる
- 人間が分類できないほどの大量のデータを分類する場合に有効
- 代表的なものにK平均法（k-means）や、Ward法などがある

階層的クラスタリング

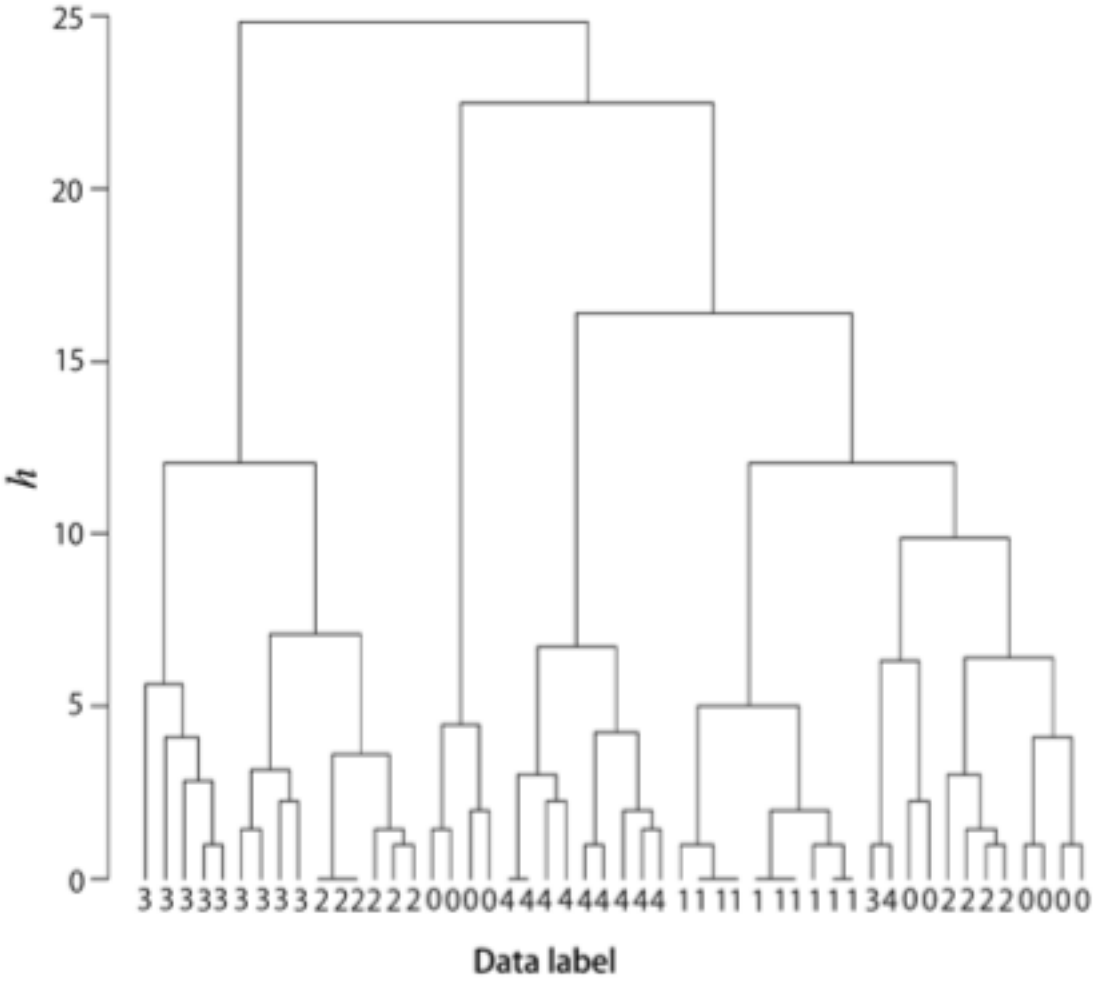
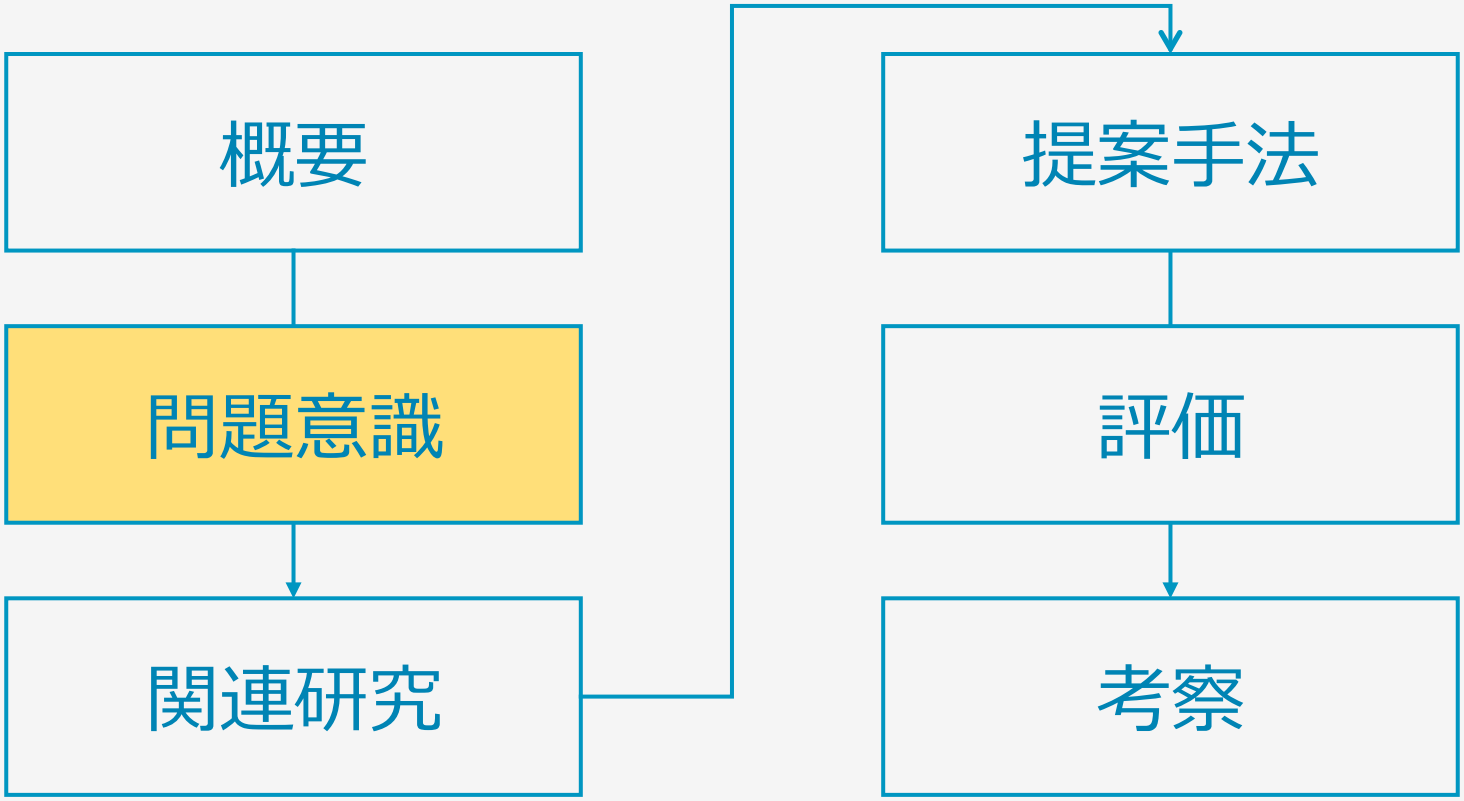


図 1　樹形図の例.

Fig. 1　Example of tree.

- 対象間の非類似度を手がかりとして、樹状の分類構造をつくる
- N個のデータを入力すると、1~N個のクラスタを得る
- 色々な手法がある
  - Ward法
  - 群平均法
  - McQuitty法

目次



問題意識

要するに

クラスタリングをしても意図するクラスターが得られない。

- k-means → 毎回結果が変わるし、
  クラスタ数が未知だと適用不可
- 階層的 → 樹形図をどの高さで切れば
  求めている結果が得られるのか不明

階層的クラスタリング

- 得られた樹形図を切る高さによって、得られるクラスター数が異なる
- 手法と非類似度の定義の組み合わせによって、樹形図の高さや形成されるクラスターが異なる

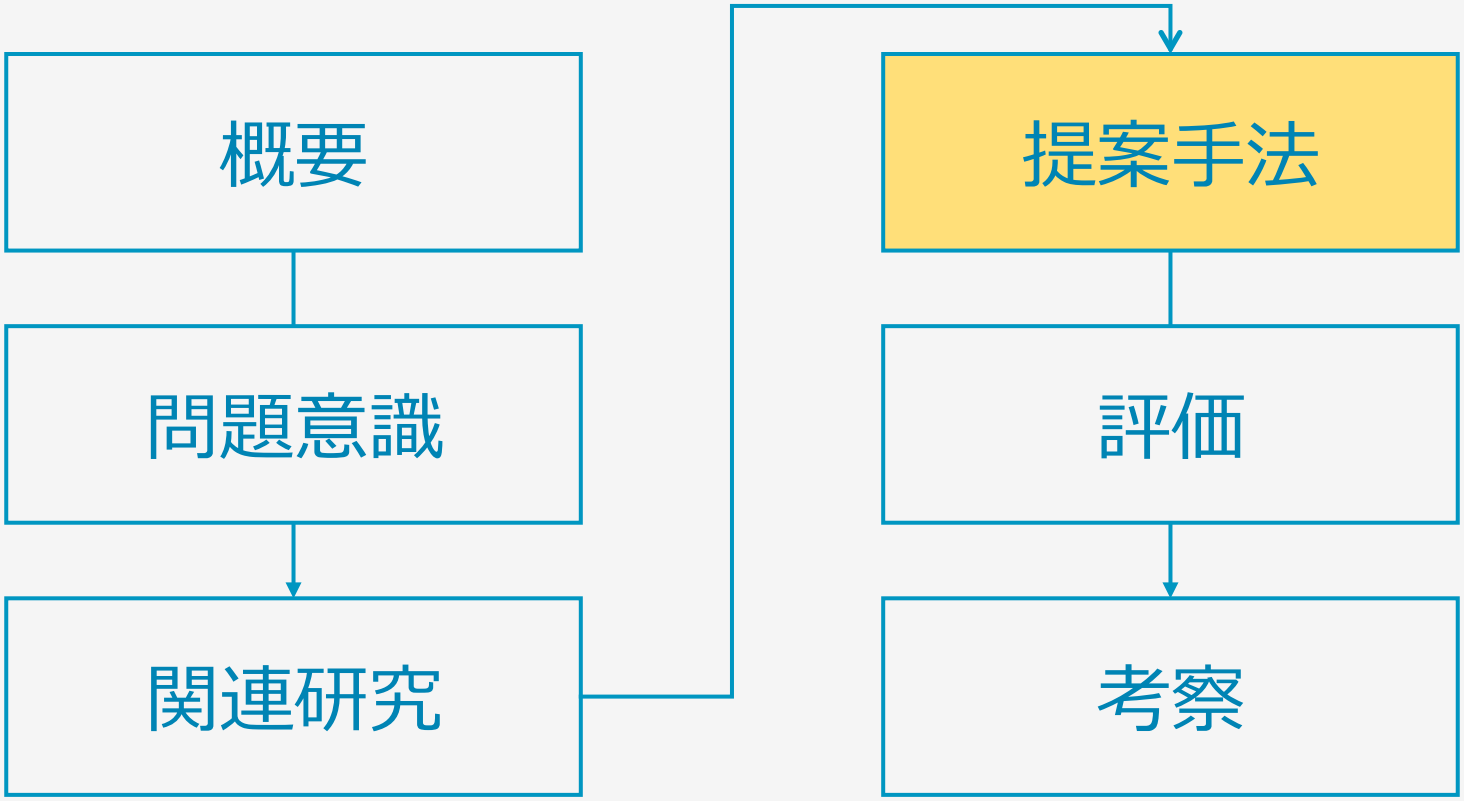
正しいクラスター数を得るには
クラスター数パラメーターを適切に決める必要がある。

階層的クラスタリング

- 自動でクラスター数を決める手法だと
  - 期待するクラスターと一致しない
  - クラスター数が一致しても、クラスタリング精度が悪い

クラスタ数精度とクラスタリング精度を改善するために、データを適切に変換する必要がある。

目次



登場する単語

- クラスター数精度: P_N
- クラスタリング精度: P_C
- クラスタリングパラメーター
  - データ変換パラメータ: σ_c, s_c
  - クラスター数パラメータ: h_c

提案手法

- 階層的クラスタリング手法に、以下の2つを導入
  - クラスター数パラメータ
    - 正しいクラスタ数を得るため
  - データ変換パラメータ
    - クラスタ数精度とクラスタリング精度を改善するため
- この2つのパラメータを調整する手法

アルゴリズム

1. 学習データの作成
    1. データの一部を抽出
    2. 人手でクラスタリング
2. 学習ステップ
3. 評価ステップ

学習ステップ

1. データ変換パラメータの調整
2. 学習用データの変換
3. 学習用データのクラスタリング
4. P_Nの計算 & h_cの調整
5. P_Cの計算
6. 最適なパラメータの探索が終了したらexit
   それ以外なら1に戻る

学習ステップ

1. データ変換パラメータの調整
2. 学習用データの変換
3. 学習用データのクラスタリング
4. P_Nの計算 → h_cの調整
5. P_Cの計算
6. 最適なパラメータの探索が終了したらexit
   それ以外なら1に戻る

クラスター数精度: P_N

$$P_N = \left(1 - \min\left(\frac{|K' - K|}{K}, 1\right)\right) \times 100$$

K：本来のクラスタ数
K’：クラスタリングの結果得られたクラスタ数

クラスター数パラメータ: h_c

$$h_c = \left(h_{\max} + h_{\min} \right) / 2$$

h_{max} : P_N=100%となる分類木の高さの最大値
h_{min} : P_N=100%となる分類木の高さの最小値

学習ステップ

1. データ変換パラメータの調整
2. 学習用データの変換
3. 学習用データのクラスタリング
4. P_Nの計算 & h_cの調整
5. P_Cの計算
6. 最適なパラメータの探索が終了したらexit
   それ以外なら1に戻る

クラスタリング精度: P_C

$$P_C = \frac{\left(\sum_{i=1}^{K} P_i \right) \times 100}{K}$$

K：クラスタリングの結果得られたクラスタ数
P_i：クラスタC_i（i=1～K）の分類精度(0～1)

学習ステップ

1. データ変換パラメータの調整
2. 学習用データの変換
3. 学習用データのクラスタリング
4. P_Nの計算 & h_cの調整
5. P_Cの計算
6. 最適なパラメータの探索が終了したらexit
   それ以外なら1に戻る

データ変換パラメータ:σ_c, s_c

画像の変換手法は既存研究を参考…

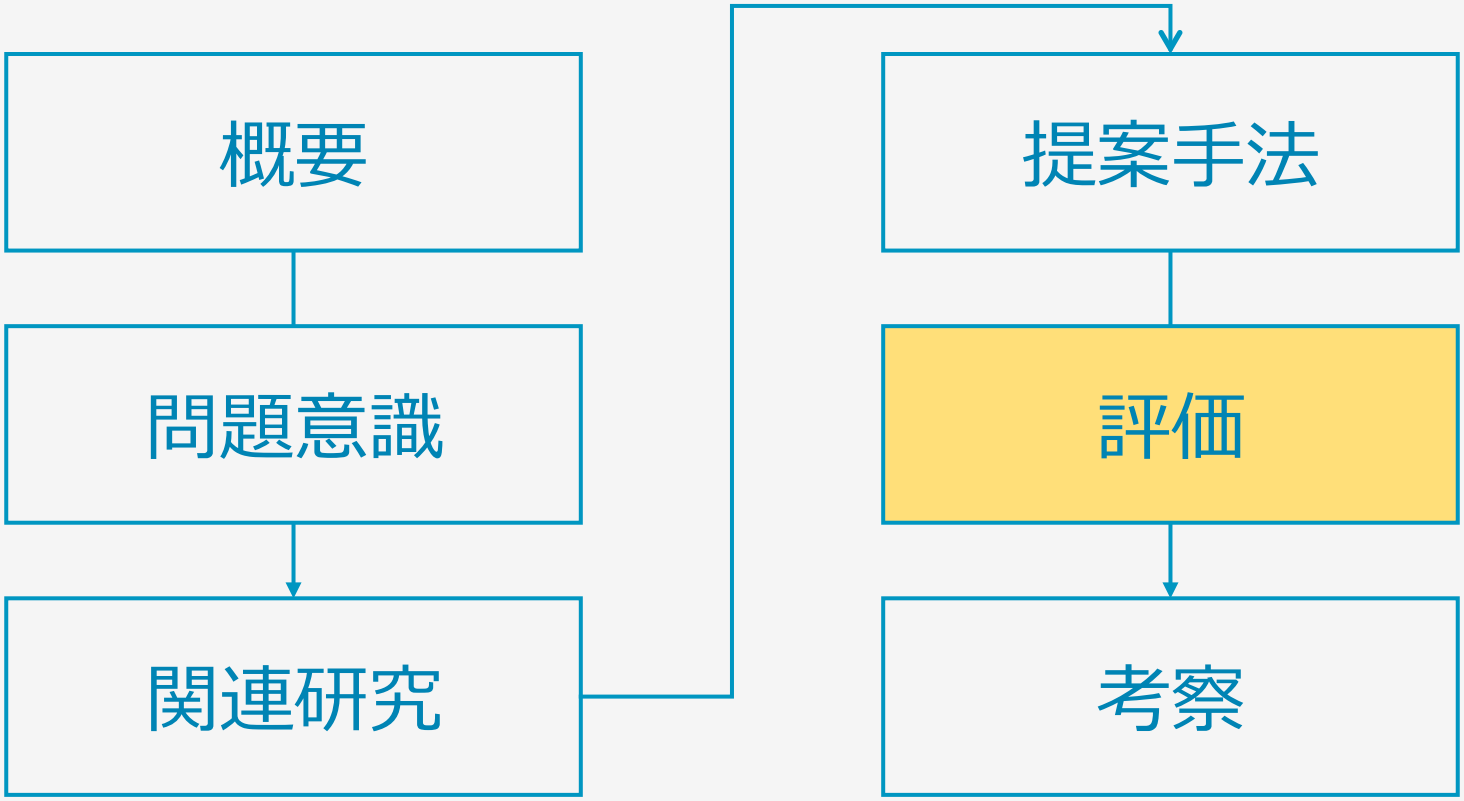
クラスタリング精度P_Cとクラスター数精度P_Nが最も高くなるσとsを求める。

σ：ガウスフィルタのパラメータ
s：SOMの学習ステップ数

評価ステップ

1. 得られたデータ変換パラメータを用いて評価用データの変換
2. 得られたクラスタ数パラメータを用いて評価用データのクラスタリング
3. P_NおよびP_Cの計算

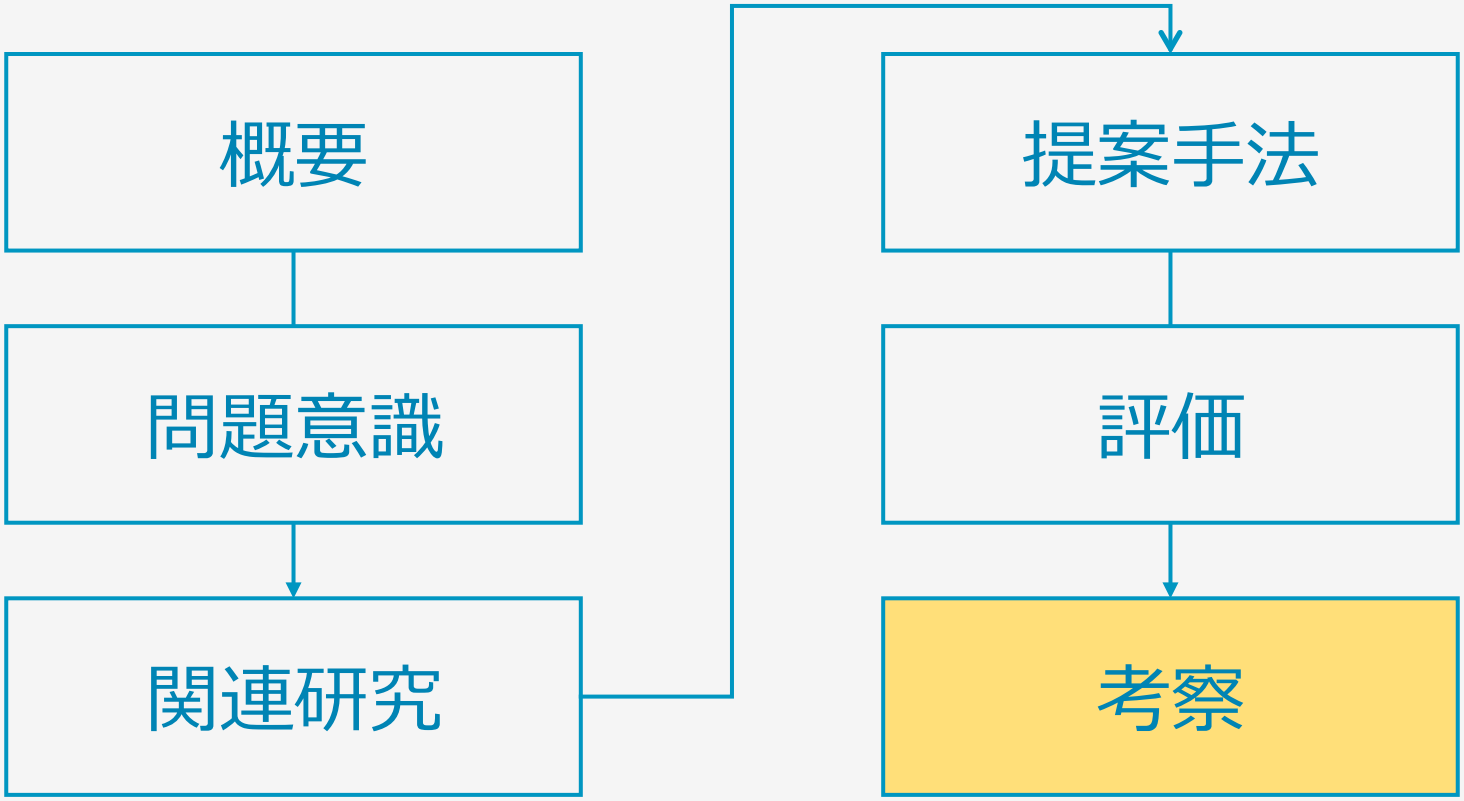
目次



性能評価

- 0~9の手書き文字の認識
  - 0~3を学習データ、4~9を評価用データとした

目次



FIN.

ご静聴ありがとうございました

SAMPLE SLIDE

- 発表用に調整されたシンプルデザイン

1. **フォント**

   和文フォントをメイリオ、欧文フォントを
   Segoe UI にデフォルト設定。

2. **カラー**

   黄色と水色の彩度を落とし、明るい印象を
   残したまま、落ち着いた配色設定。

3. **見本付き**

   オブジェクトやテキストの実際の配置や
   デザインの見本用スライドがあります。

比較ボックス

Aについて

- SAMPLE
  - A
  - A

Bについて

- SAMPLE
  - B
  - B

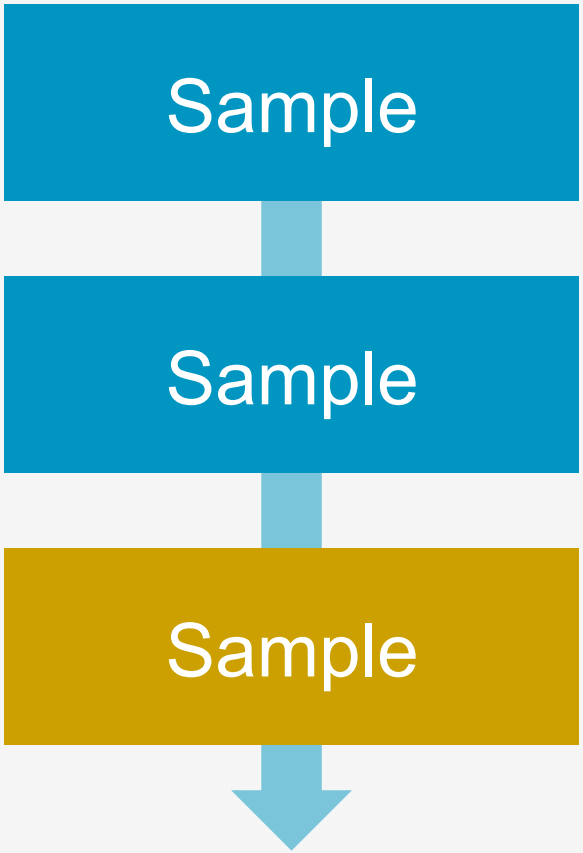
中表紙

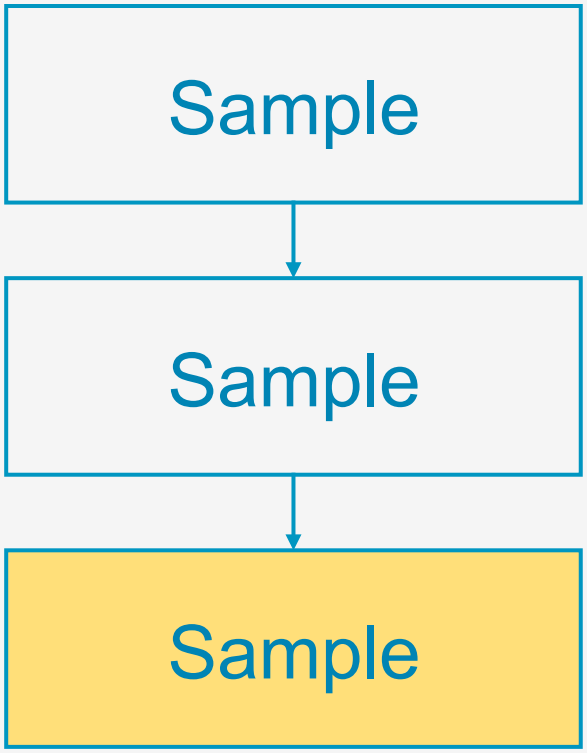
セクションの区切りなどに

見本｜通常の箇条書きスタイル

- 第 1 項目のサンプル
  - 第 2 項目のサンプル
    - 第 3 項目のサンプル
    - 第 3 項目のサンプル
- **第 1 項目のサンプル**
  - 第 2 項目のサンプル
    - 第 3 項目のサンプル
    - 第 3 項目のサンプル

見本｜オブジェクト

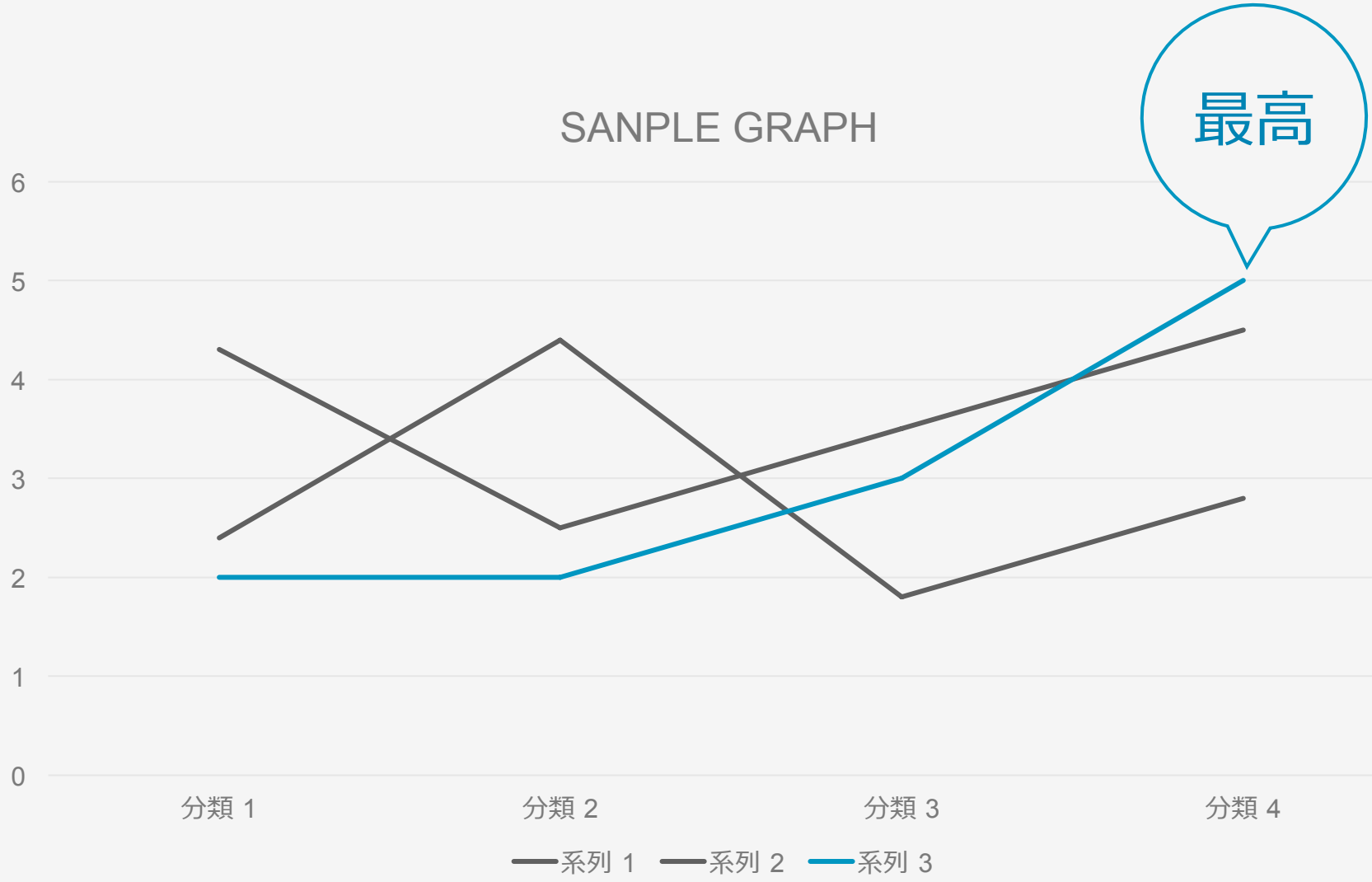




見本 | 表

素 材	器 具	結 果
素材01	器具X	0.01%
素材01	器具Y	0.33%
素材01	器具Z	0.03%
素材02	器具X	0.95%
素材02	器具Y	0.22%

見本 | グラフ



見本｜文章の見せ方の例

見出しはこのような感じで

本文はこのような感じで書き、強調する時は
太字にするか、<u>下線を引く</u>ようにする

- 箇条書きも同様にする
- 特に重要な用語には色もつける
- 文章はなるべく位置をそろえ、
  色は水色と黄土色を使いまわすように

まとめはこのような感じで堂々と