

# データサイエンスへの誘い

第11回: 統計的学習

瓜生真也 (デザイン型AI教育研究センター・助教)

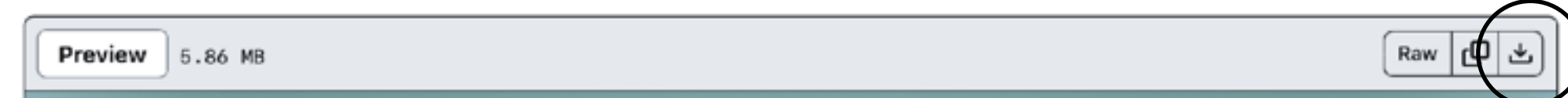
# 講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



1. ガイダンス、データサイエンスとは何か
2. 現代社会におけるデータサイエンスの活用事例
3. データ処理の手法
4. データの要約
5. データの可視化
6. データと確率
7. データからの推論
8. 複数のデータを比較する

9. 統計のウソ
10. 統計モデリング
11. 統計的学習
12. ゲスト講師による特別講演
13. 機械学習とAI
- ~~14. コンピューターを用いた分析~~
15. 期末試験（7月25日）
16. 振り返り＆統括（8月1日）

# 今日の目標

統計的学習で扱う問題と

---

枠組みを理解する

---

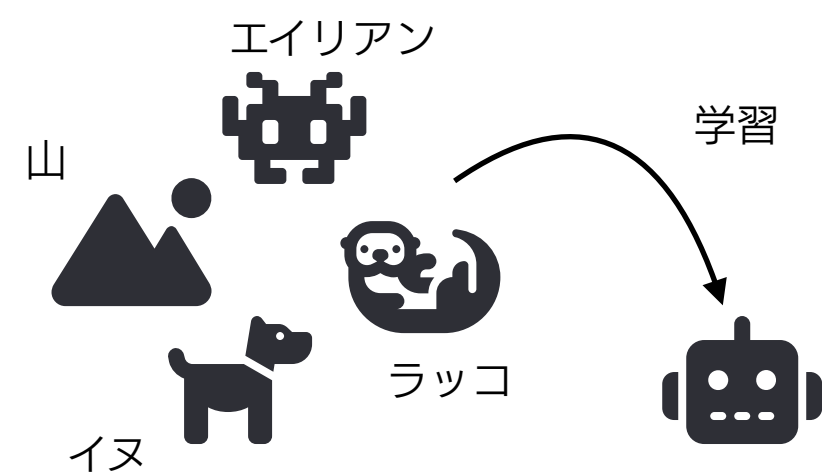
# 統計的学習の概要

# 機械学習モデルの学習手法の違い

目的や問題設定、条件に応じて異なる学習手法が存在する

## 教師あり学習

問題と答えの組み合わせから傾向を学習、  
新しいデータ（答えは不明）が与えられた時にデータの予測を行う

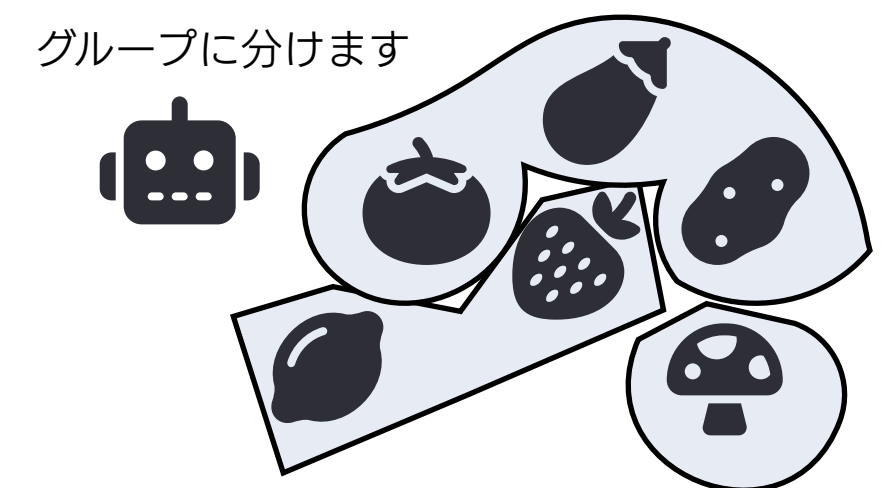


分類問題→離散値の予測… ロジスティック回帰モデルやサポートベクターマシンなど

回帰問題→連続値の予測… 線形回帰モデル、ランダムフォレストなど

## 教師なし学習

答えのない状態でデータの特徴（構造やパターン）を学習、データの特徴を抽出する  
→クラスタリングや次元削減などデータの潜在的な構造を抽出する



# 教師あり学習

# 教師あり学習の流れ

入力から出力  $y$  を予測する関数  $y = f(x; \theta)$  を学習する


$\theta$  は入力に対する重み、パラメータ

 訓練データ（学習データ）  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

学習 訓練データを使ってモデルを学習  $f(x; \theta)$   
→パラメータを調整

どうやって  $x$  から  $y$  を表現できる？

$\theta$  を最適な値とするには？

推論 学習モデルを使って、与えられたデータ（ テストデータ）から出力  $y$  を予測する

## 回帰

例) 住宅の特徴量（変数）から住宅価格を予測  
→出力が連続値

## 分類

例) 画像の特徴量（ピクセル情報）から対象物（ラベル）を予測  
→出力が離散値



# 教師あり学習の流れ

取得済み  
データ

探索的データ分析

データ分割

訓練データ

前処理・  
特徴量エンジニアリング

モデルの学習  
→パラメータの決定

検証データ

前処理・  
特徴量エンジニアリング

モデルの推論・性能評価・選択  
→モデルの決定

テストデータ

前処理・  
特徴量エンジニアリング

モデルの推論・性能評価



# 汎化性能

訓練データに対する、未知のデータへの対応能力、予測精度

モデルの学習に用いるデータとは別に、汎化性能を調べるためのデータを用意する  
→訓練データとテストデータ

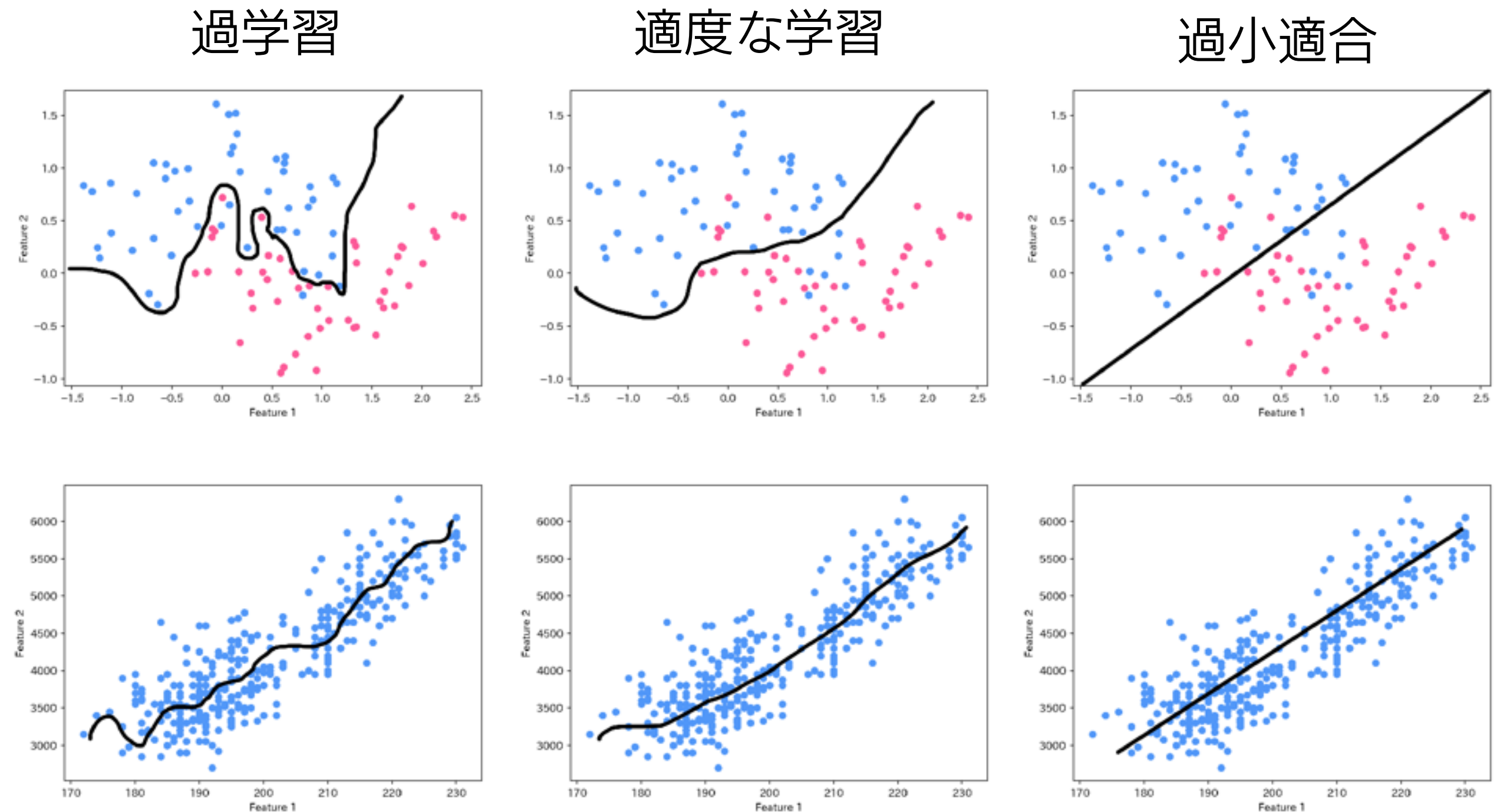
## 過学習、モデルの過剰適合

訓練データに過度に依存したモデルを構築したことにより、未知のデータへの予測精度が低下する

対策

- 交差検証法の採用
- 正則化
- データ増強
- モデルの簡略化

非線形



# データ分割

# ランダムに訓練データとテストデータに分割

## ホールドアウト法

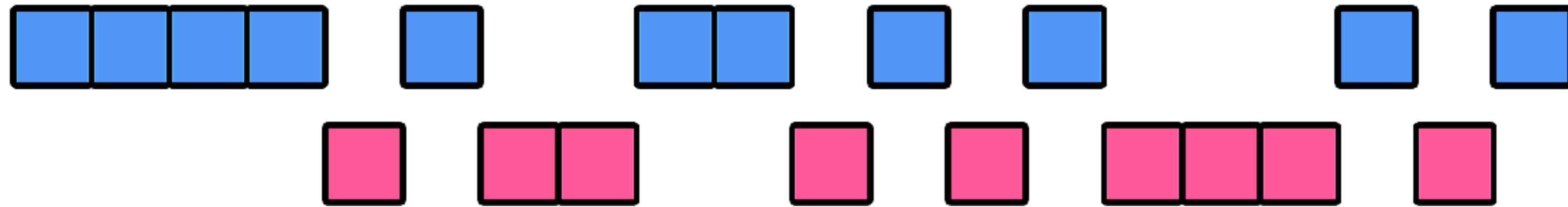


訓練データ



テストデータ

元データ



分割の方法によってはデータに偏りが生じ、過学習につながるおそれがある

例) 時系列データでのランダムな分割はNG (訓練データに未来のデータが含まれる)

# 交差検証法

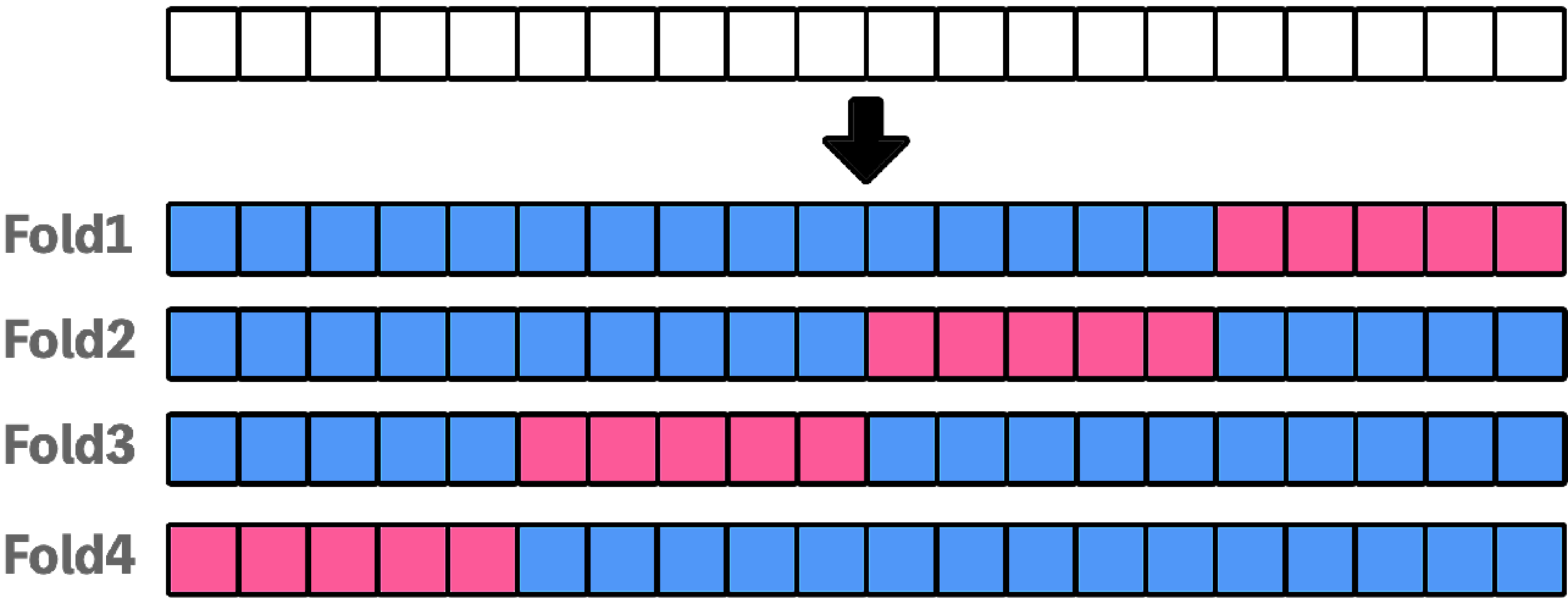
複数個の訓練データとテストデータの組み合わせを用意

過学習の影響を排除して、予測モデルの正確な精度が測定できる（分割時の偶然性による影響を軽減）

適切なハイパーパラメータの選択のためにも使われる

分割方法によっていくつかのバリエーションがある

## *k*分割交差検証



# 前処理の必要性はモデルによって異なる

多くのモデルは入力データのスケールに敏感

例) 数字2桁の変数と7桁の変数がある場合、変数間の効果に差が生じる

→線形回帰、k-means、主成分分析などのモデルは変数間のスケールを揃える操作（スケーリング）が必要

→木ベースのモデル（決定木、ランダムフォレスト）は変数のスケールの影響を受けないため、スケーリングは不要

欠損値や外れ値への対応はモデルや利用するライブラリによって対応が異なることが多い

例) ライブラリ側で自動的に欠損値を含むデータを削除

→k近傍法やサポートベクターマシンは外れ値の影響を受けやすい

## 数値データに潜む問題

例えば… スケールが大きくことなる  
歪んだ分布  
外れ値を含む

変数間で複雑な関係をもつ  
冗長な情報を含む

→適切な前処理・特徴量エンジニアリング、  
適切なモデルの選択が求められる

# ロジスティック回帰

# ロジスティック回帰

第10回参照

リンク関数がロジットで、誤差構造が二項分布の場合の一般化線形モデル

目的変数の値… 0または1（二値変数）を予測する

入力変数とそれらの重みを組み合わせた線形関数を利用（線形回帰と同じ）

$$z = w_0 + w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n$$

その結果(ここではz)をロジスティック関数に入力として与える

→0から1の範囲からなる値を出力。データがあるカテゴリに属する確率として解釈できる



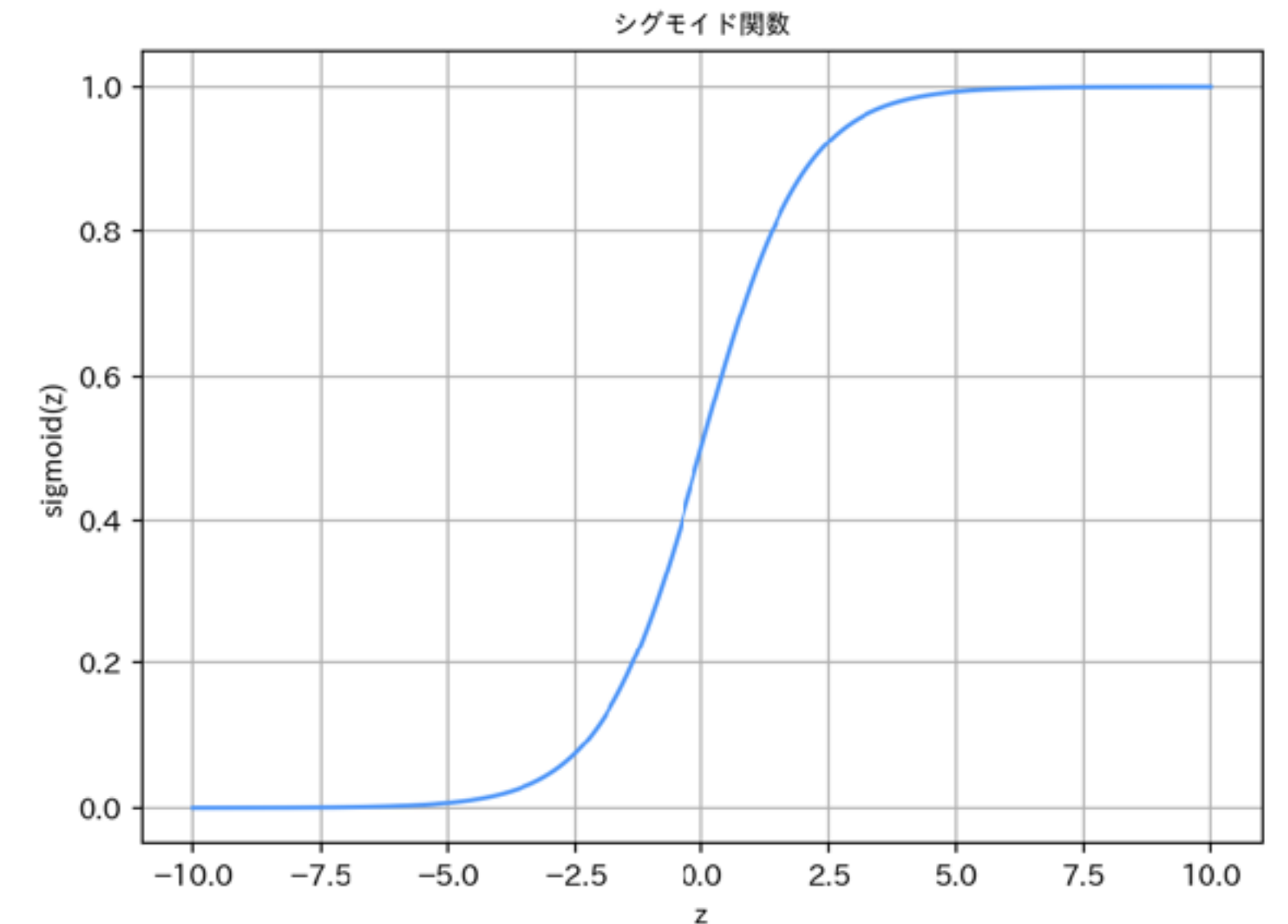
# ロジスティック関数とロジット関数

## ロジスティック関数（シグモイド関数）

$$p = \frac{1}{1 + \exp^{-z}}$$

ロジット関数… ロジスティック関数の逆関数

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



## オッズ比

$$\frac{p}{1-p}$$

事象の起こりやすさを2つの群で比較して示す

例) 男性が商品を購入する確率 ( $p_1$ ) が0.8、女性が購入する確率( $p_2$ )が0.2のとき、

オッズは次のように求められる

$$\frac{p_1}{1-p_1} = 4 \quad \frac{p_2}{1-p_2} = 0.25$$

比較する2群のオッズから比（オッズ比）を求める。オッズ比が1より大きい場合、男性が女性に比べて商品を購入する確率が高いことを示す。1より小さい場合は男性よりも女性が商品を購入する確率が高いことを示す。

この場合、女性よりも男性が商品を購入する確率が16倍高いことを示す。



# モデルの評価

# 分類問題におけるモデルの評価指標の例

真の値とモデルの予測結果を比較する

- 正解率(accuracy): モデルが正しく予測したデータの割合
- 適合率 (precision): 正と予測したデータのうち、実際に正である割合。
- 再現率(recall): 実際に正であるもののうち、正であると予測された割合。
- F1スコア(f1-score): 適合率と再現率の調和平均により得られた値。適合率と再現率のバランスを考慮した評価指標。この値が高いほど、適合率と再現率の両方が高いことを示す。

# 教師なし学習: k平均法

# クラスタリング

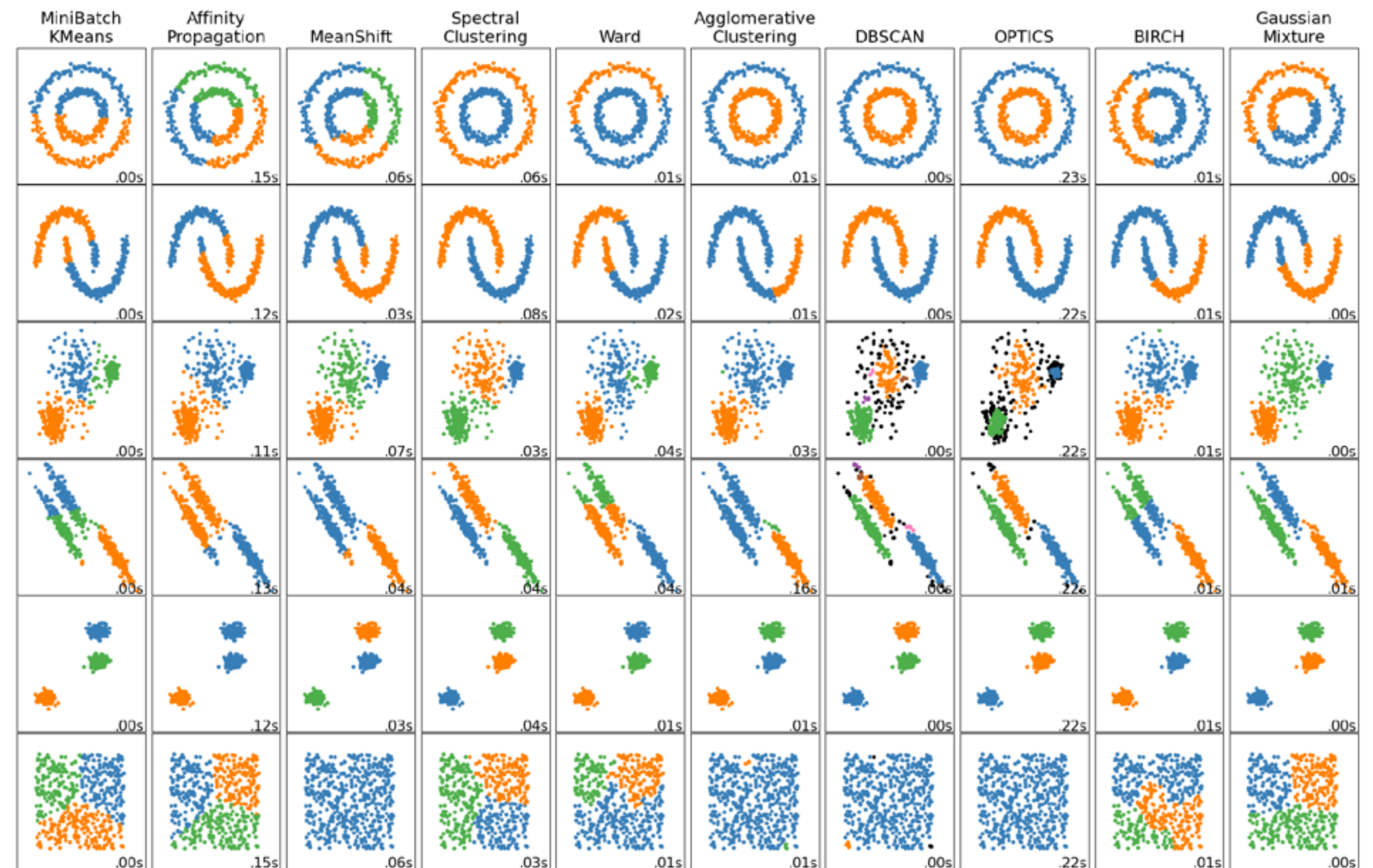
データ間の類似度・距離をもとに、データを未知のグループ（クラスター）に分割する

さまざまなアルゴリズム

k平均法

階層クラスタリング

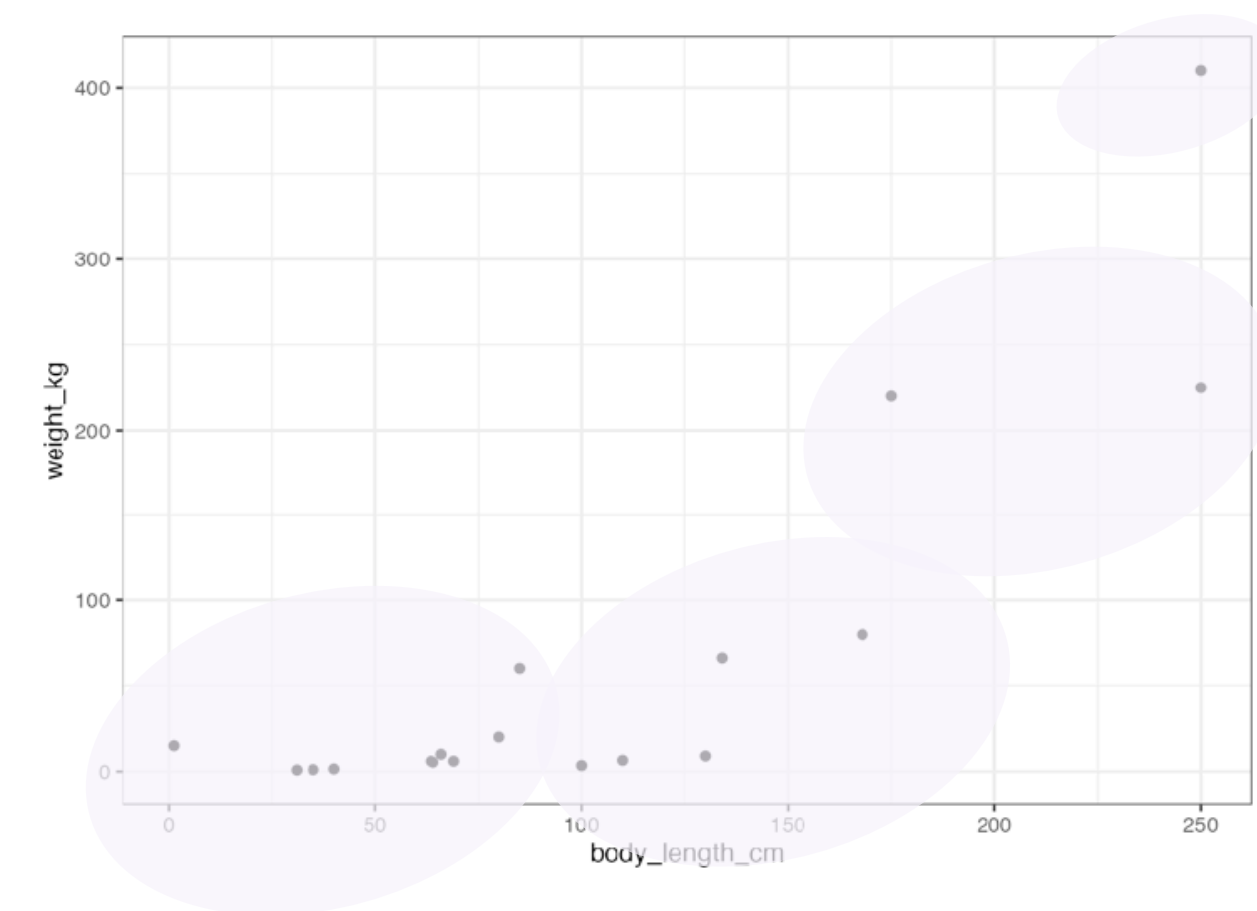
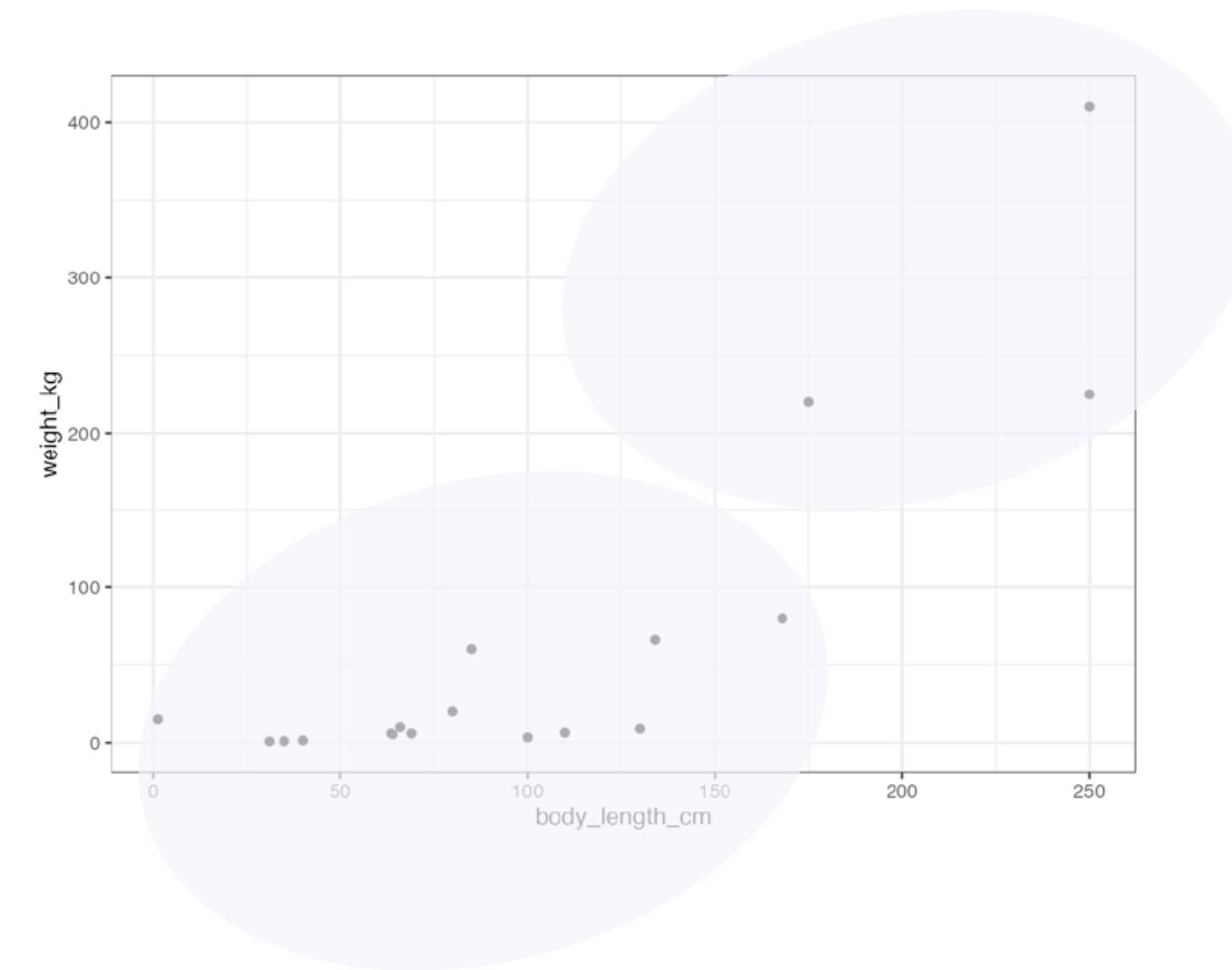
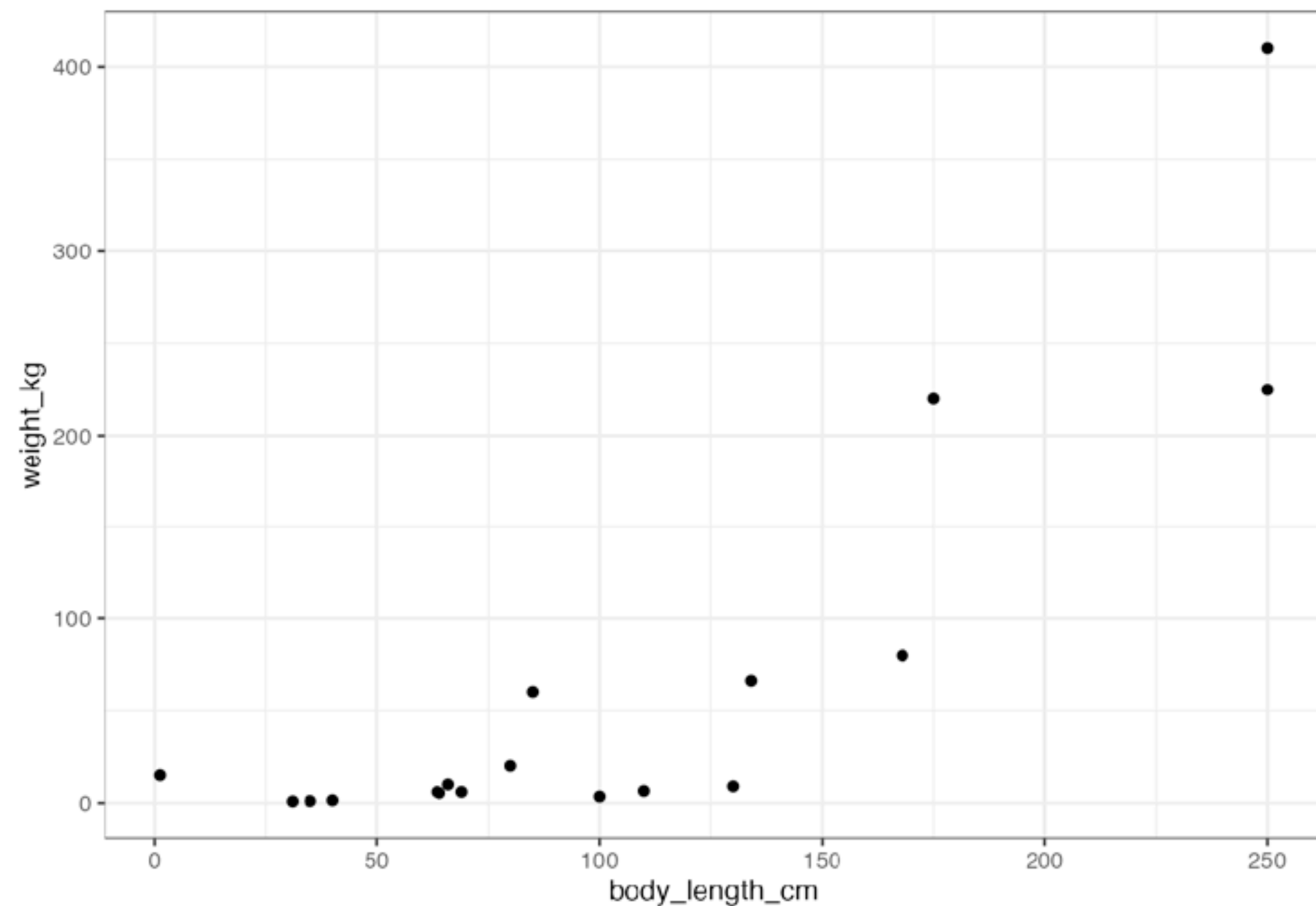
DBSCAN



# k平均法

クラスタの数  $k$  をあらかじめ決めておき、  
クラスタ内の平均からの距離の二乗和が最小となるよう、入力を $k$ 個に分類する

動物の体長と体重データをいくつかのグループに分けるとしたら？





# k平均法の手順

- ① クラスタの数  $k$  を決める
- ② ランダムに各データをクラスタに割り振る
- ③-1 各クラスターの重心（平均値）を求め、各データからの距離（ユークリッド距離など）を求める
- ③-2 各データを最も近い重心に対応するクラスタに振り分け直す

手順②のランダムな割り振りのために、K平均法の結果は実行のたびにわずかに異なることがある

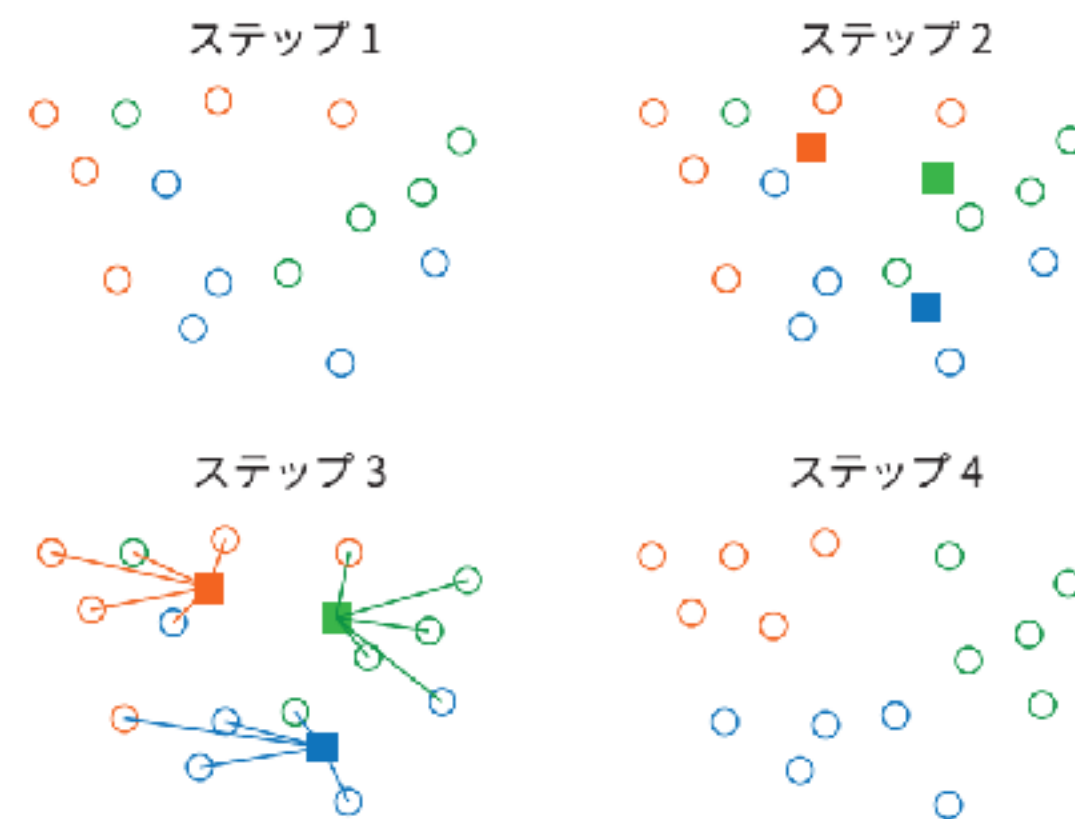
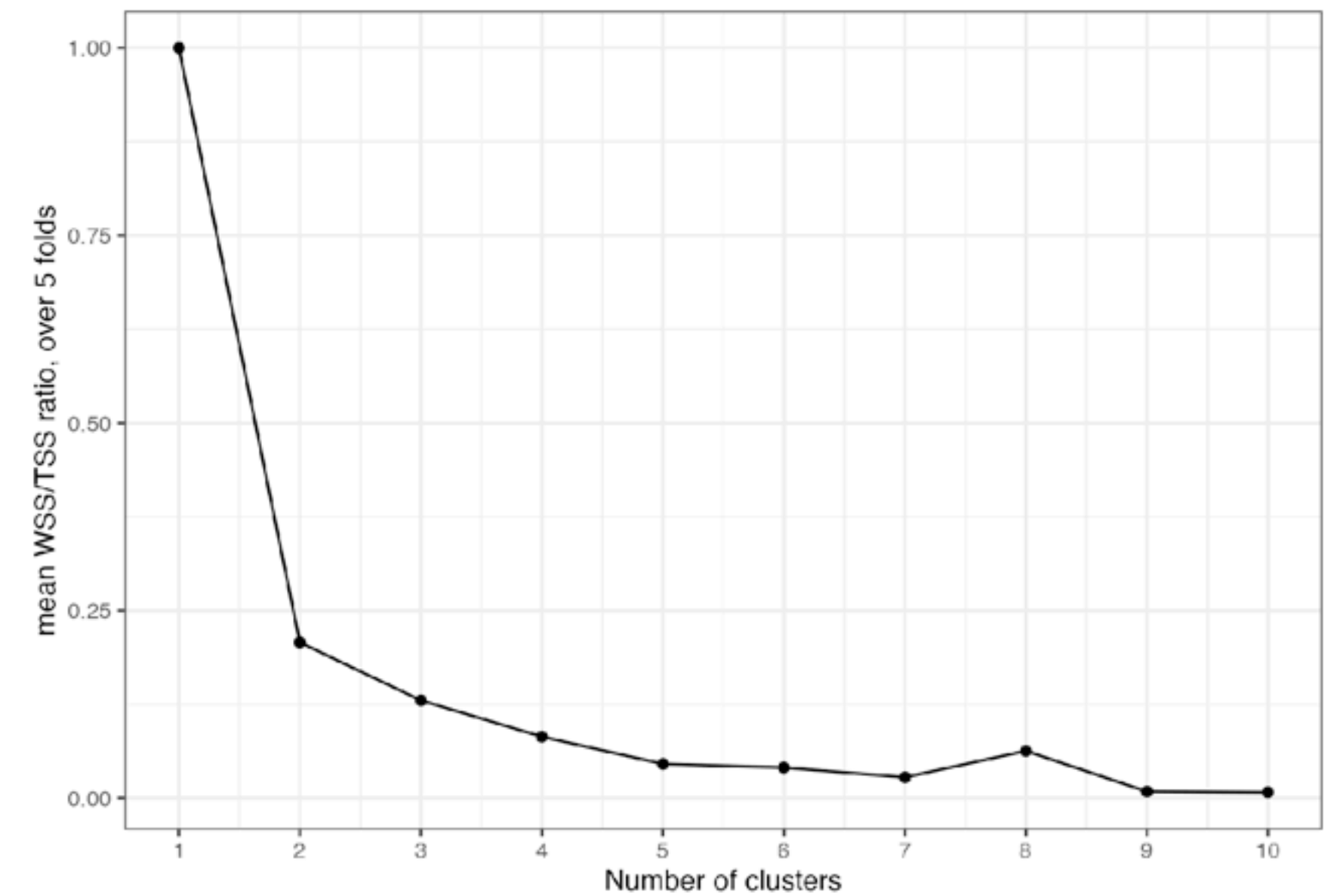
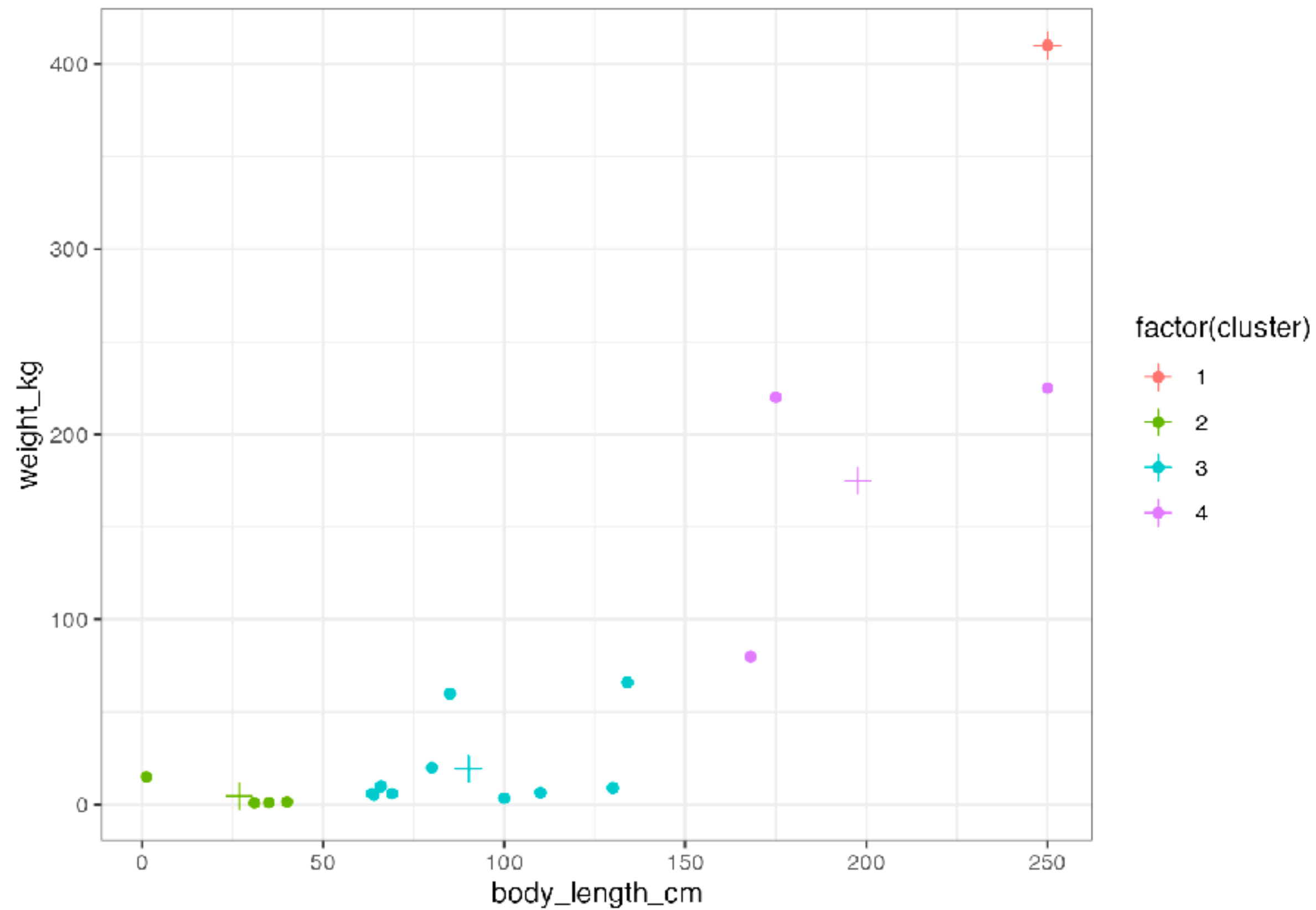


図 4.12 k-平均法の手順。まず、それぞれのサンプルにランダムにクラスタを割り当てて（ステップ1）、次に四角で示されるクラスタ平均を求めます（ステップ2）。そして、クラスタ平均から距離の近いサンプルを再度、クラスタに割り当てなおします（ステップ3, 4）。この手順を繰り返しながら、サンプルのクラスタ割り当てを調整していきます。

# kの数をどうやって決める？

クラスタ内でのバラツキの和を最小にすることが最適化

k = 4の場合の動物データのクラスタリング結果



# 参考資料・URL

📖 Trevor Hastie, Robert Tibshirani, Jerome Friedman 著 and 杉山将, 井手剛, 神寫敏弘, 栗田多喜夫, 前田英作 監訳 and 井尻善久 他訳  
『統計的学習の基礎：データマイニング・推論・予測』（2014）  
共立出版. ISBN: 978-4-320-12362-5

📖 藤原幸一『スモールデータ解析と機械学習』  
（2022）オーム社. ISBN: 978-4-274-22778-3

📖 松村優哉, 瓜生真也, 吉村広志『Rユーザのためのtidymodels〈実践〉入門：モダンな統計・機械学習モデリングの世界』（2023）  
技術評論社. ISBN: 978-4-297-13236-1

📖 G.James, D.Witten, T.Hastie, R.Tibshirani [著] and 落海浩, 首藤信通 訳  
『Rによる統計的学習入門』  
（2018）朝倉書店. ISBN: 978-4-254-12224-4

📖 金森敬文『Rによる機械学習入門』（2017）オーム社.  
ISBN: 978-4-274-22112-5