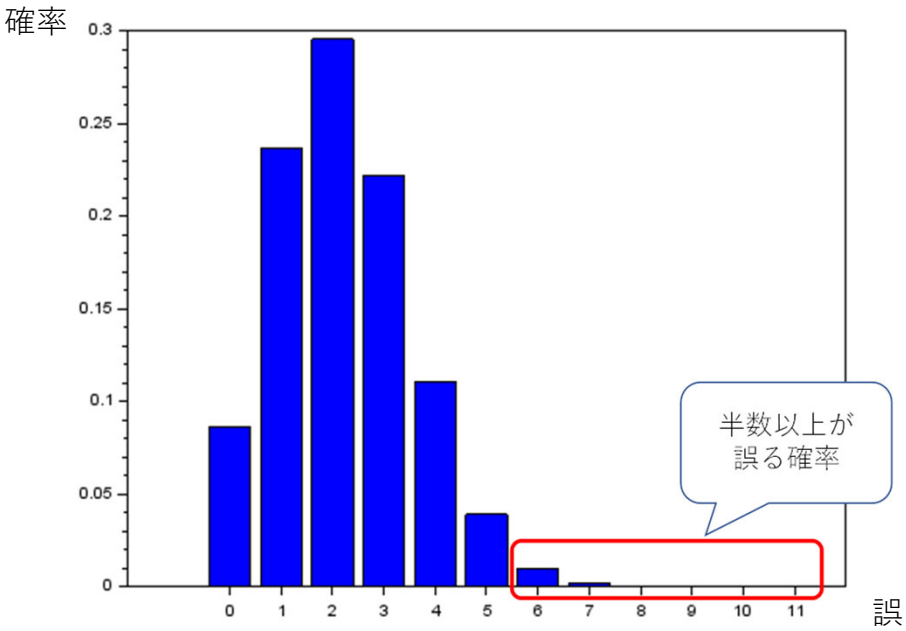
アンサンブル学習（5章）

p.146　2コマ目

アンサンブル学習

- アンサンブル学習とは
  - 識別器を複数組み合わせ、それらの結果を統合することで、個々の識別器よりも性能を向上させる方法
- なぜ複数の識別器で性能が向上するのか



- 誤り率20%の識別器を11個作成
- それぞれが誤るのは独立

多数決を取った場合

⇒ 半数以上が誤るのは1.2%

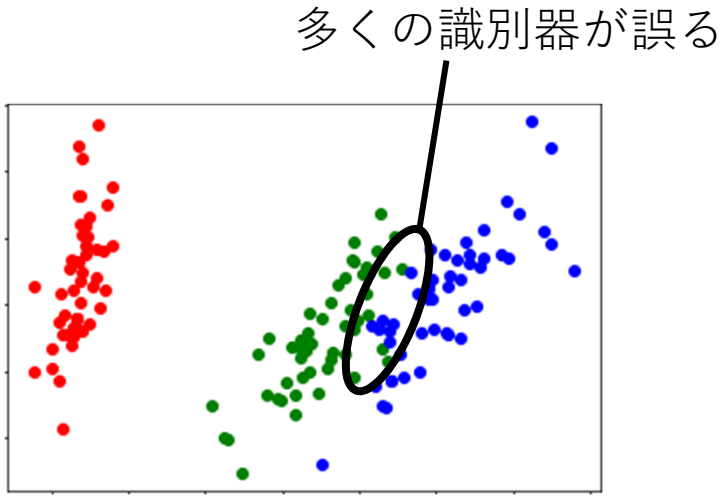
アンサンブル学習

- ここまでの議論の非現実的なところ

  「それぞれの識別器の誤りが独立」

  ⇒　データの誤りやすさに差はない ×

識別面付近のデータなど、
普通は成立しない



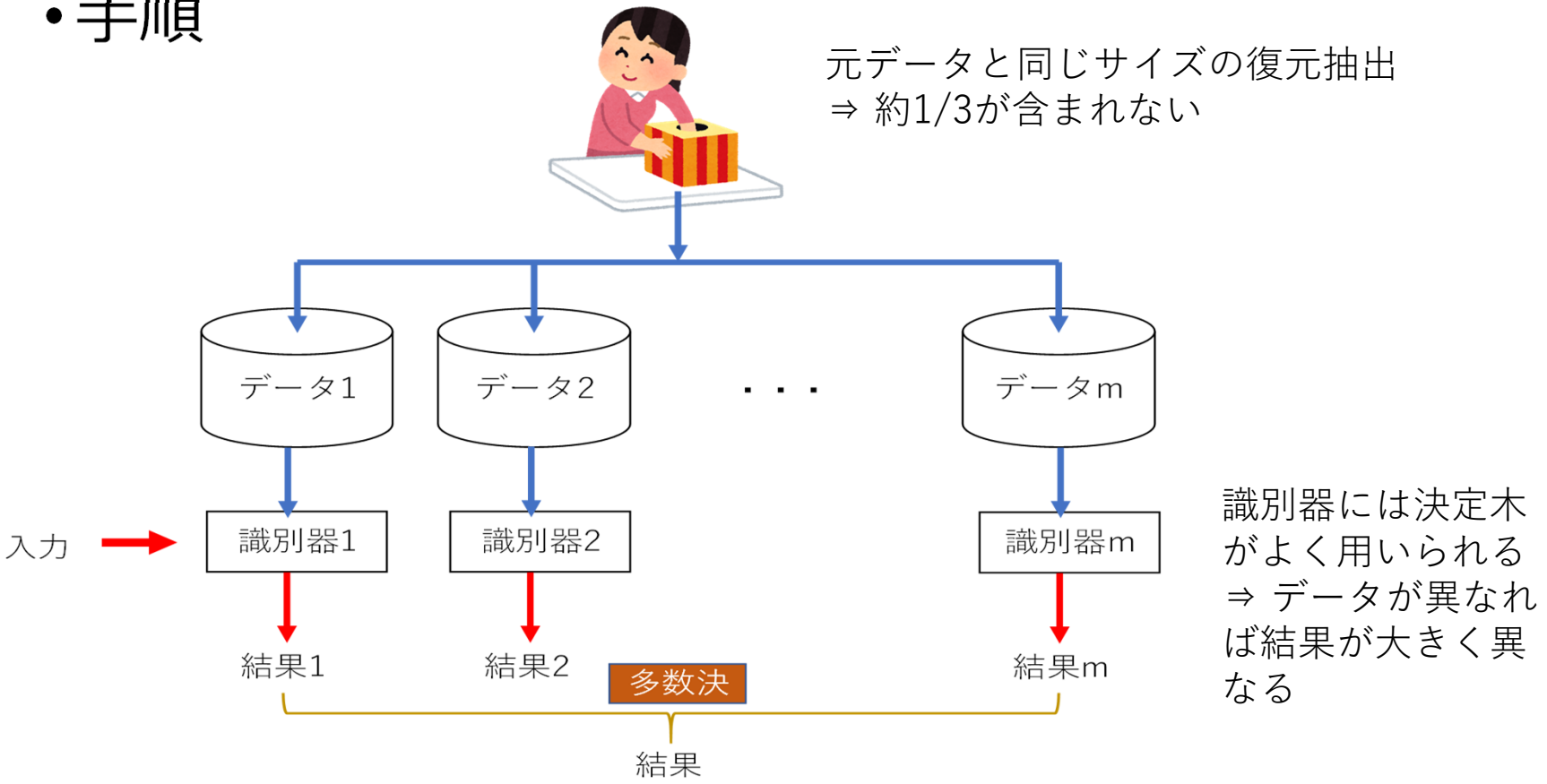
- アンサンブル学習の目標
  - なるべく異なる振る舞いをする識別器を作成する

アンサンブル学習

- アンサンブル学習の手法
  - バギング
  - ランダムフォレスト
  - ブースティング

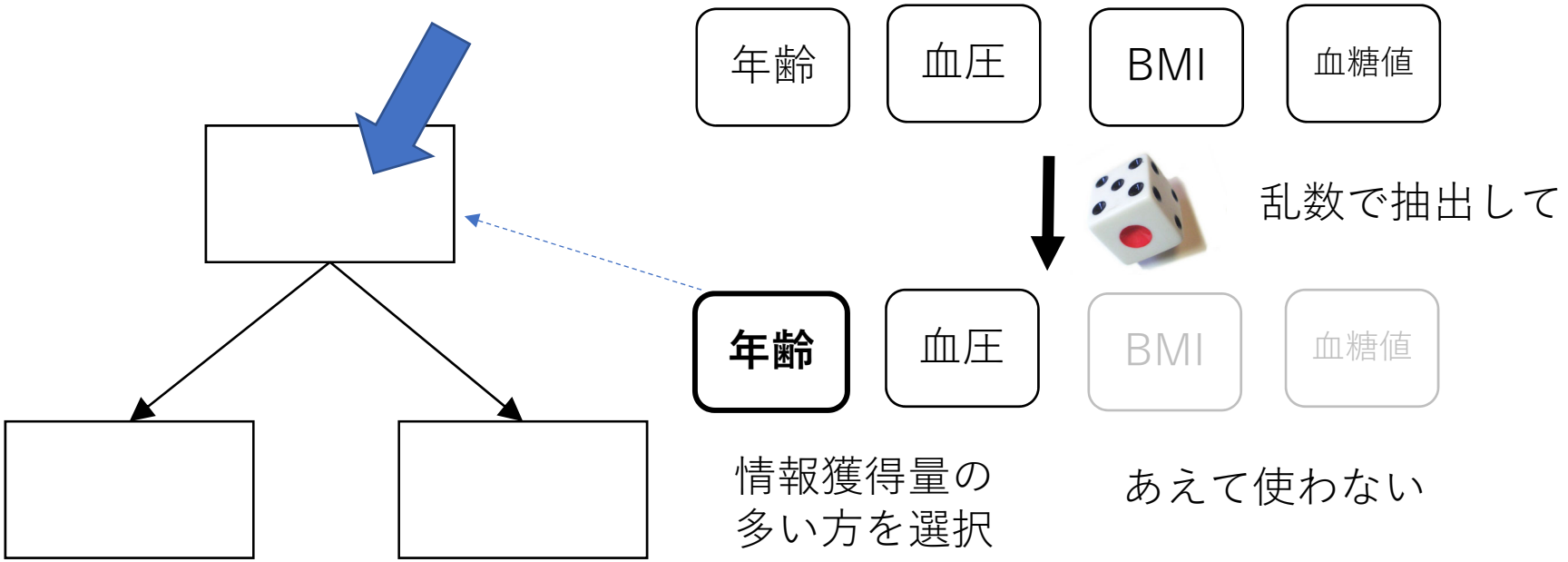
バギング

- バギングのアイディア
  - 異なる学習データから作成された識別器は異なる
- 手順



ランダムフォレスト

- ランダムフォレストのアイディア
  - バギング＋識別器を作成する毎に異なる特徴を用いることで異なった識別器を複数作成する
- 手順

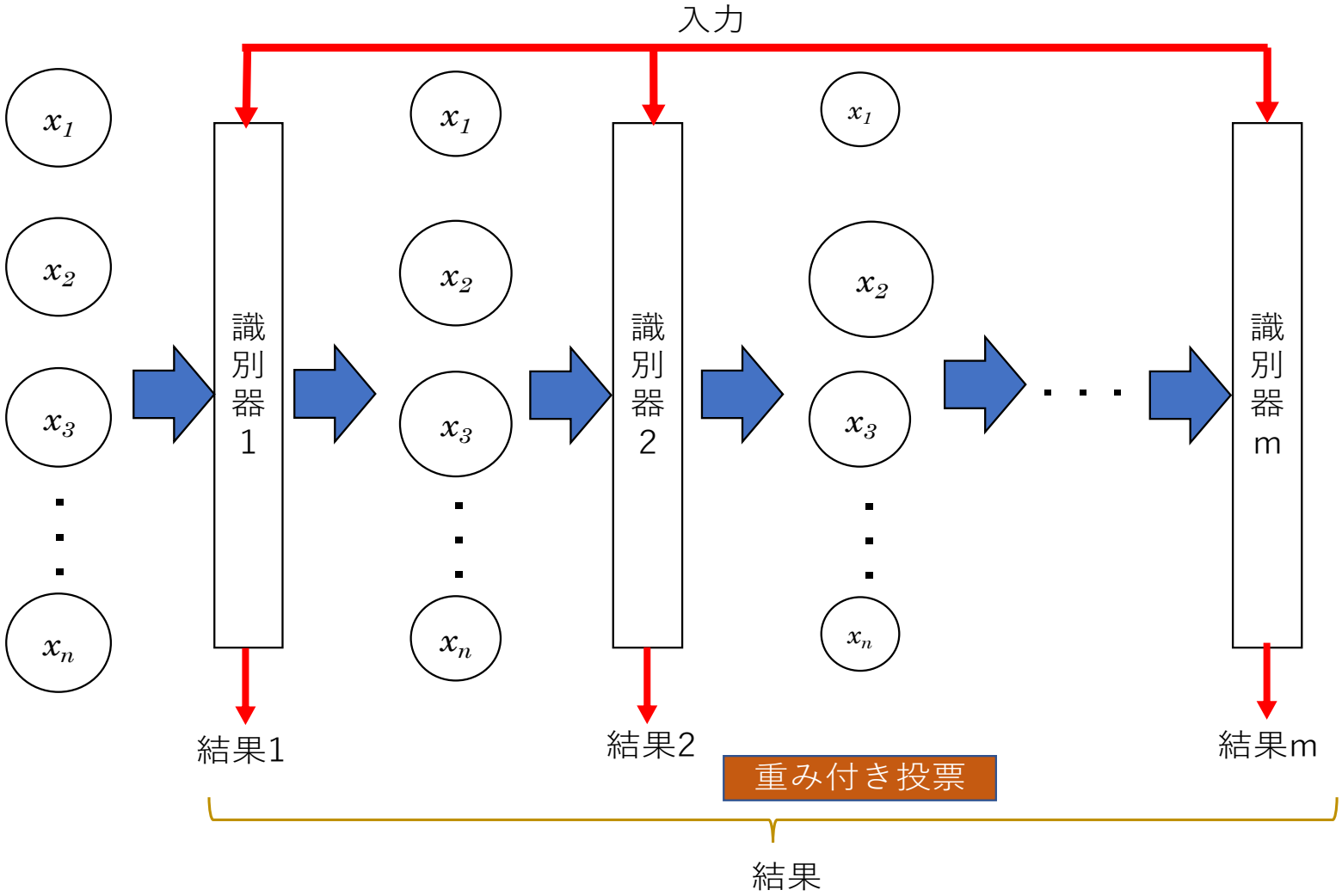


ブースティング

- ブースティングのアイディア
  - 現在の識別器が誤識別を起こすデータを正しく識別する識別器を逐次的に追加
  - 過学習とならないように、識別器として浅い決定木を用いることが多い

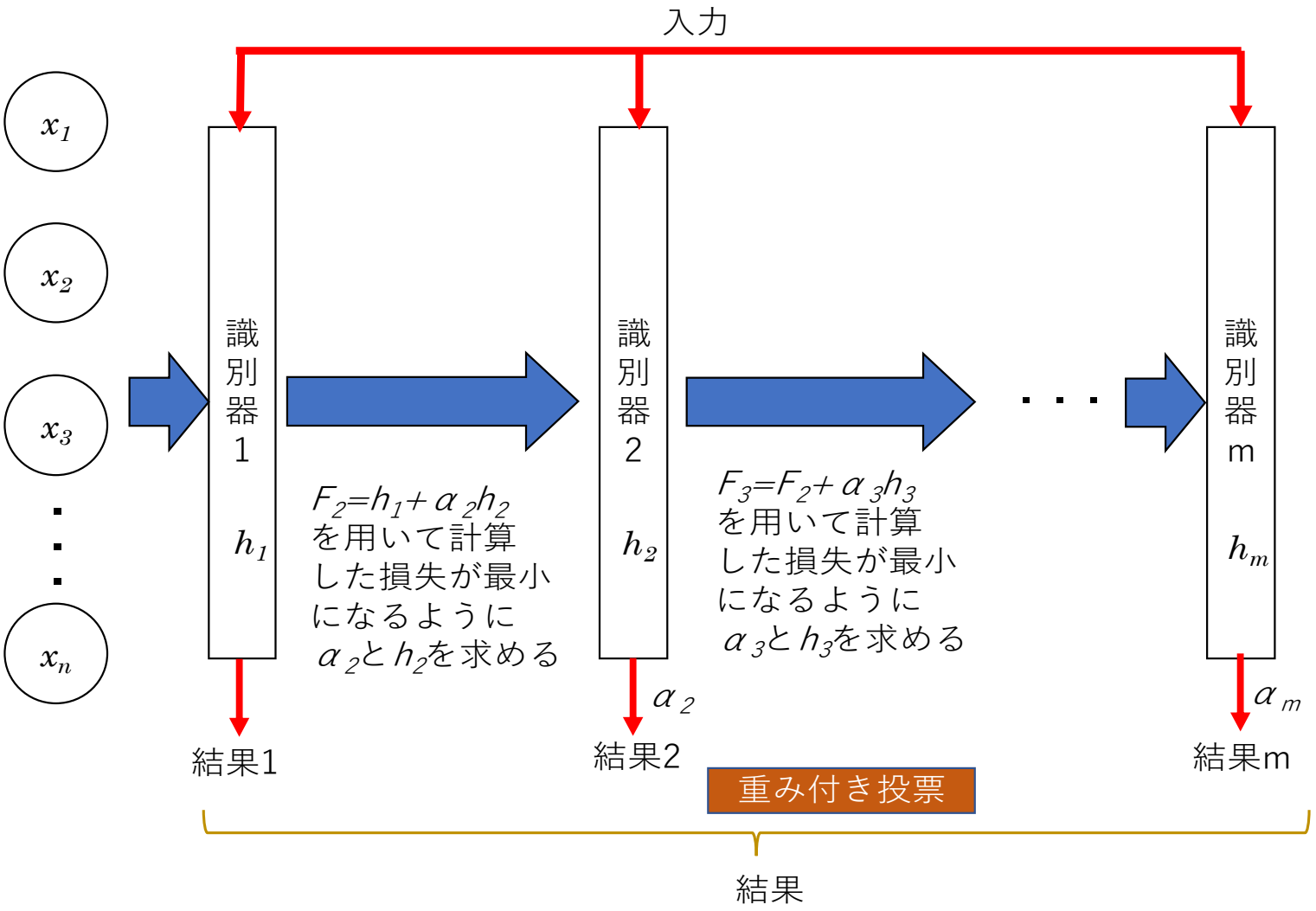
AdaBoost

- 前段の識別器が誤ったデータの重みを重くする



勾配ブースティング

- 損失が最小となるような識別器を逐次加える



教師なし学習（6章）

p.169　3コマ目

教師なし学習

- 教師なし学習とは
  - 正解情報が付いていないデータに対して、何らかの規則性を発見する手法
  - 規則がカバーする範囲によって問題が分かれる
    - データ全体をカバー：クラスタリング
    - データの部分集合をカバー：行列分解

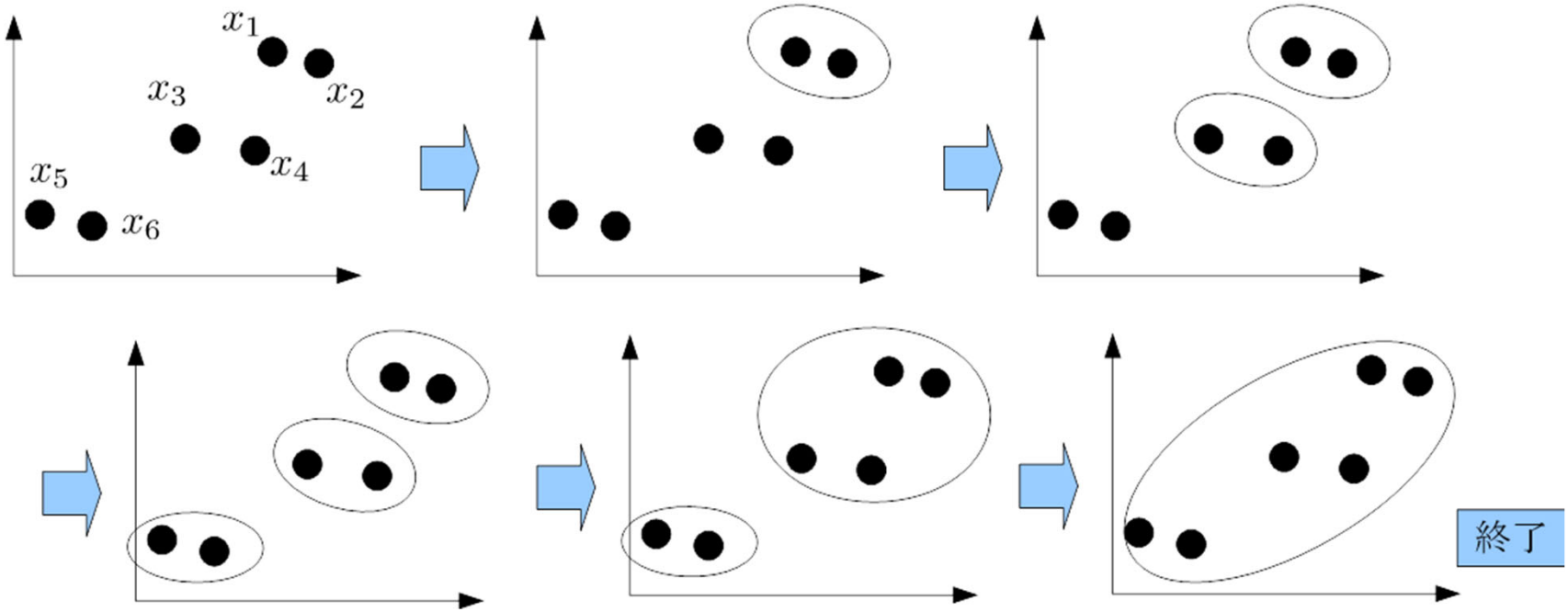
クラスタリング

- クラスタリングとは
  - 同一の性質を持つと見なされるデータのまとまりを見つけること

  例）マーケティングでのユーザグループ発見
- クラスタリングの手法
  - 階層的手法
    - ボトムアップ的にデータをまとめてゆく
  - 分割最適化手法
    - トップダウン的にデータ集合を分割してゆく

階層的クラスタリング

- 1データ1クラスタから始めて、近いクラスタを合併してゆく
- 近さの基準の選択によって、結果が異なる



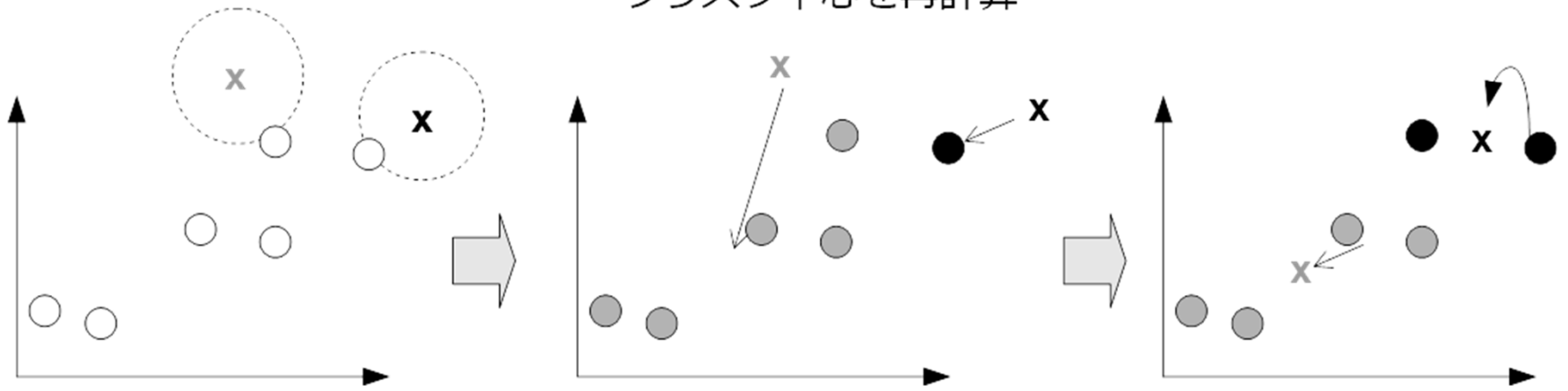
分割最適化クラスタリング

- k-means法
  - k個の平均値をランダムに決めるところから始めて、所属するデータを基準に適切な位置を決める

① 初期値として乱数で
クラスタ中心を配置

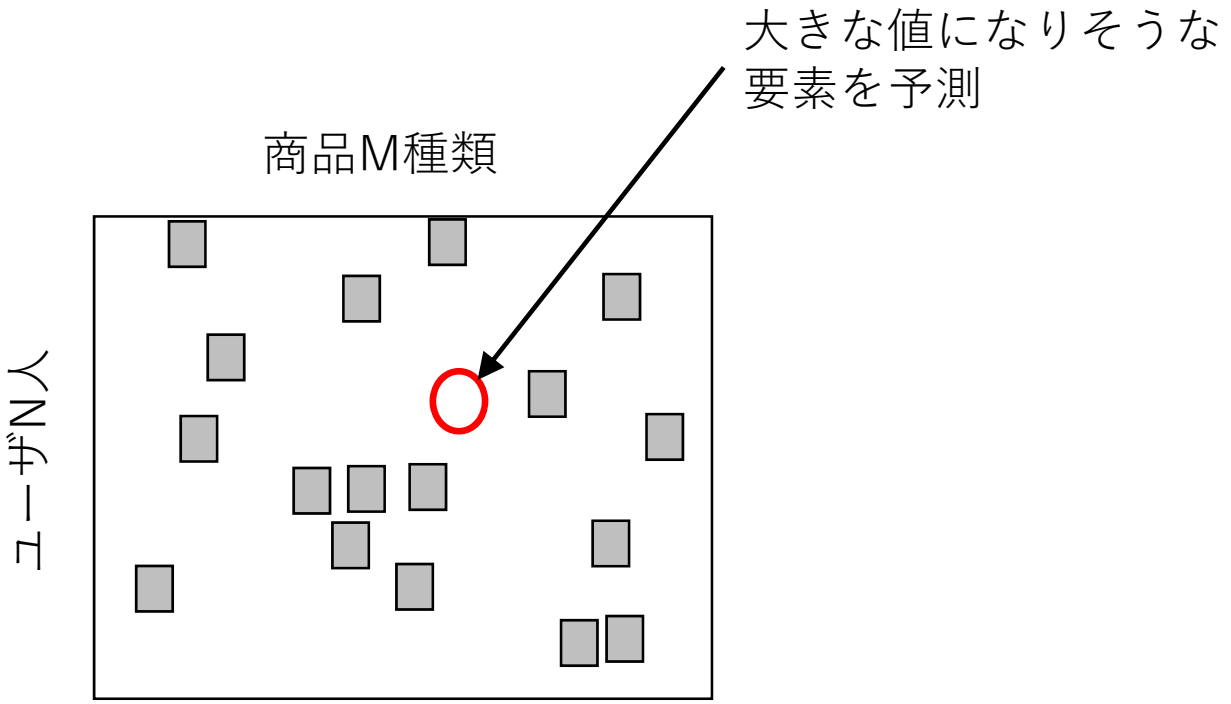
②各データを、最も近い
クラスタ中心に配属
③所属しているデータから
クラスタ中心を再計算

④ ②, ③の処理を繰り返す



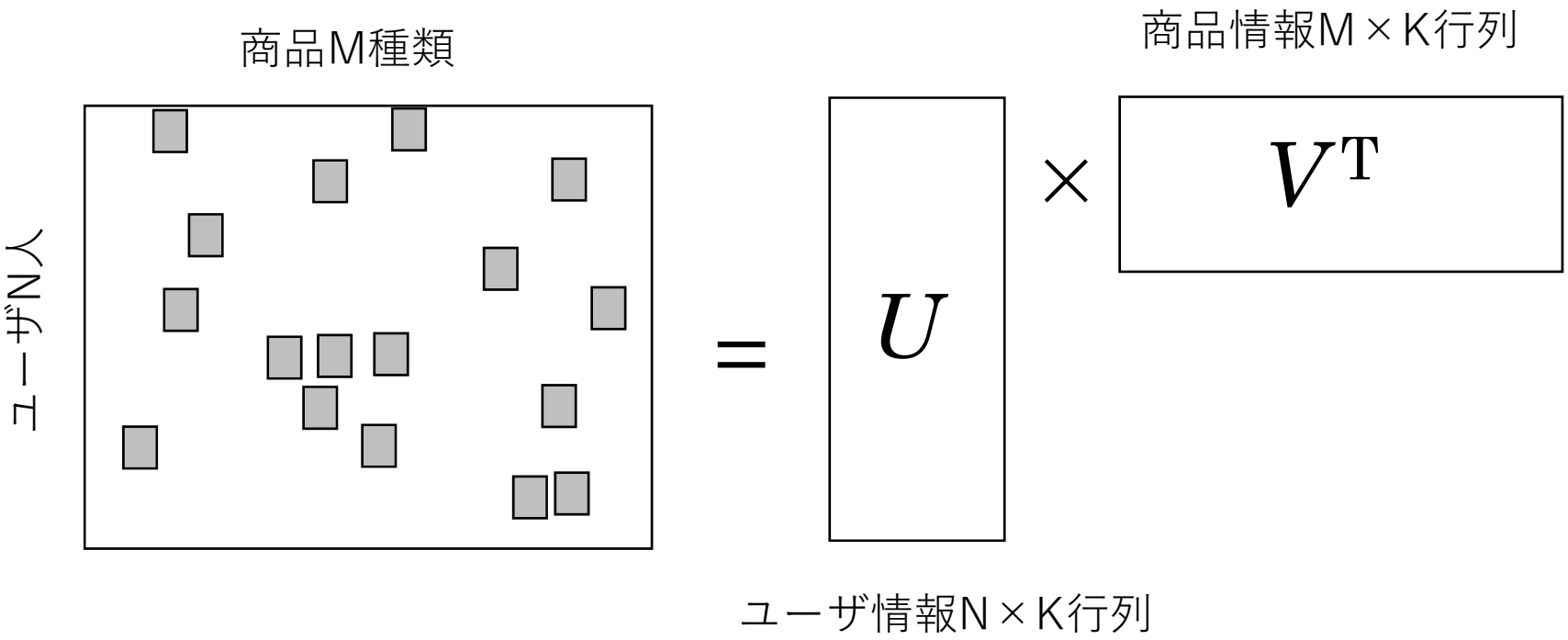
行列分解

- 推薦システムの基本手法
  - サイズが大きく、ほとんど値が埋まっていないデータを対象



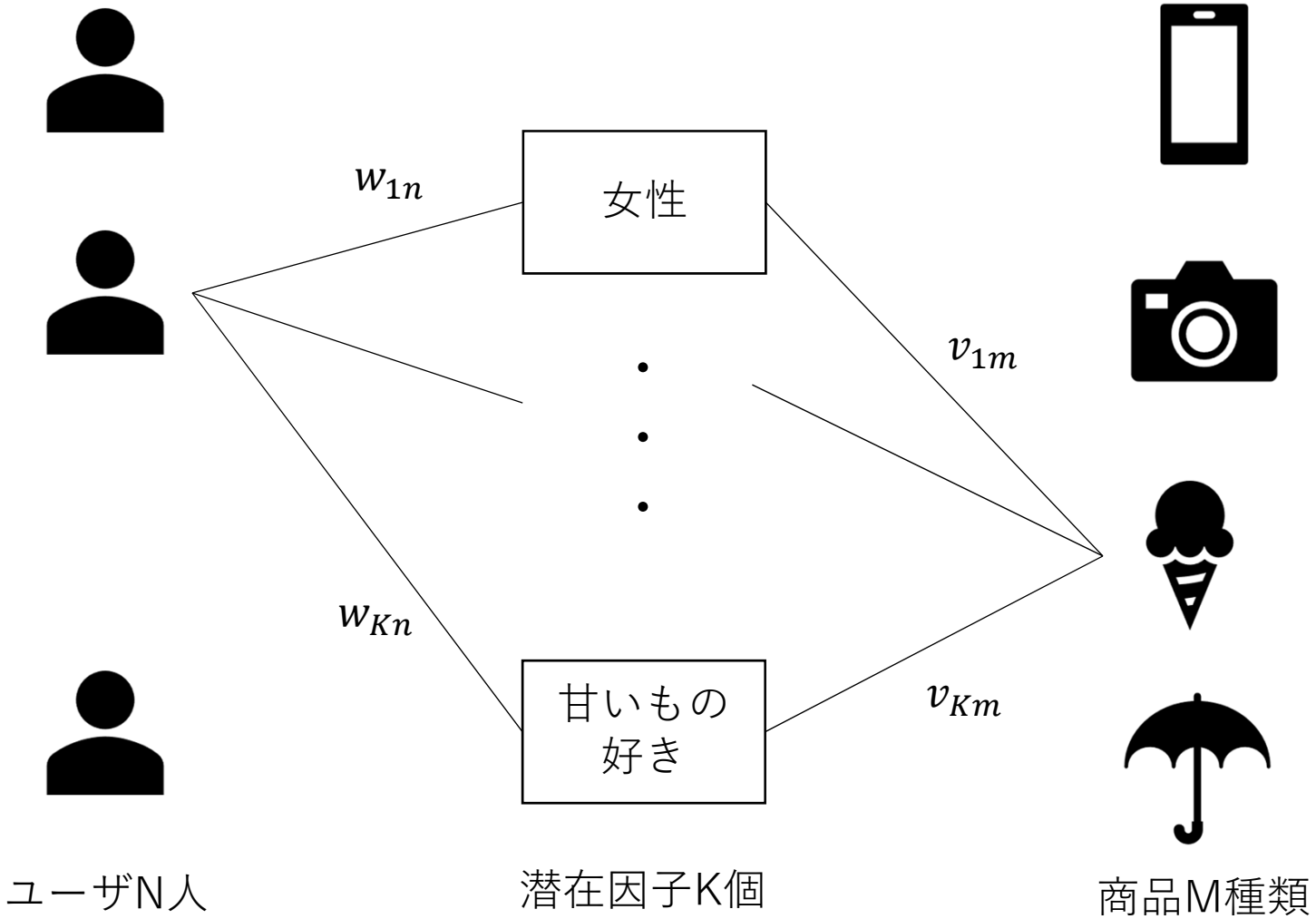
行列分解

- 行列の低次元分解



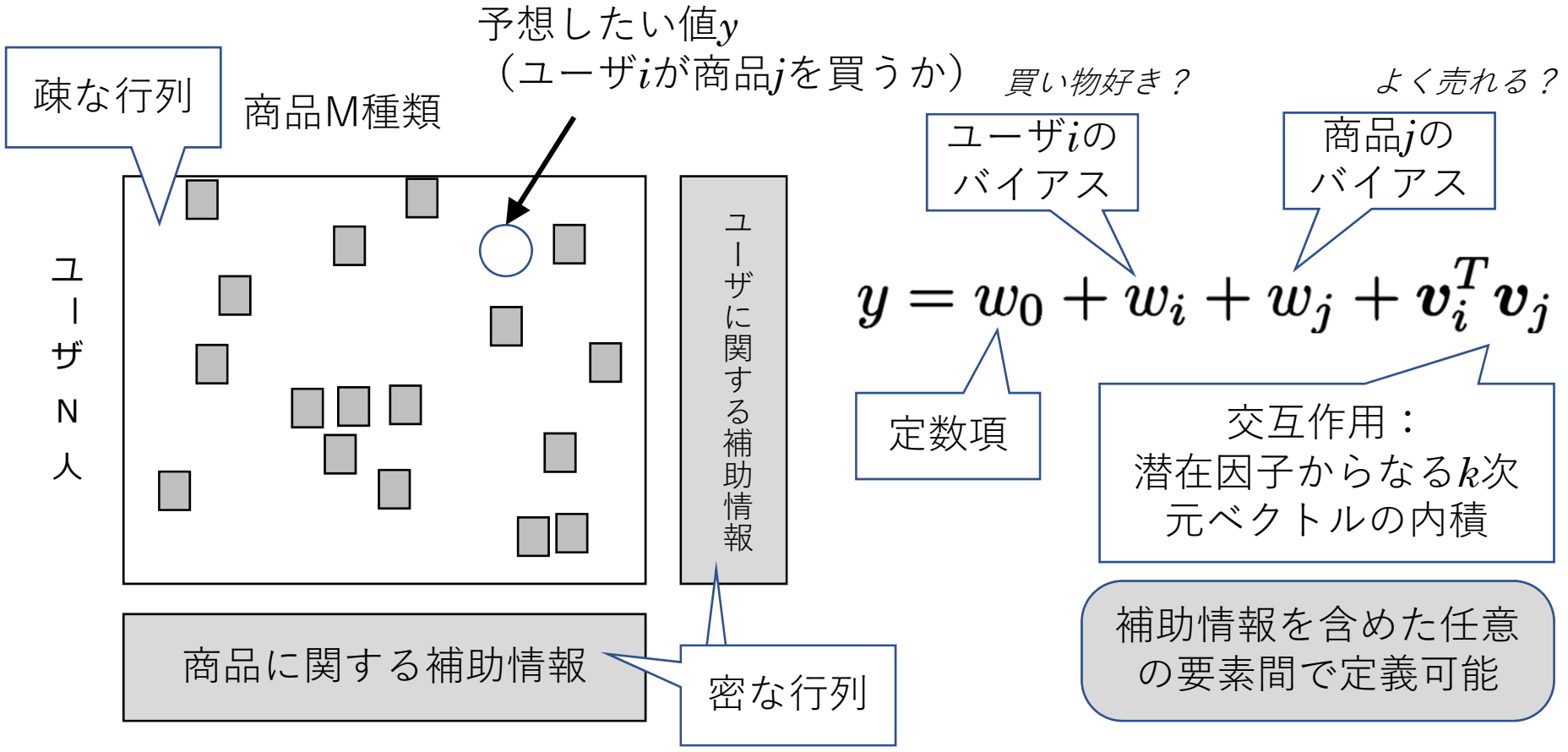
行列分解

- 低次元分解の解釈



Factorization Machine

- 別途入手可能な補助情報を用いることができる



教師なし学習の応用事例

- 日販、富士通
  - 書店に並べる本のレコメンド
    - 結果に対する書店員からの評価も利用

    https://tech.nikkeibp.co.jp/atcl/nxt/news/18/01209/
- NTT
  - Convex Factorization Machine(CFM)

  http://www.kecl.ntt.co.jp/openhouse/2016/exhibition/1/poster.pdf