クラスタ解析技術

奈良先端科学技術大学院大学
情報科学研究科
論理生命学分野研究室　助手　大羽　成征
http://hawaii.naist.jp/

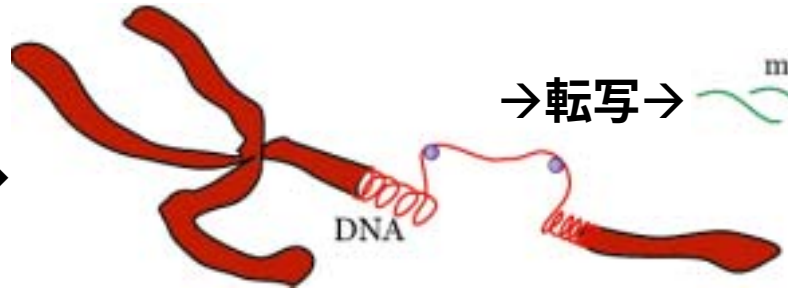
クラスタ解析の目的：
似たものをグループ化

- **例**
  - **Webページを内容で分類**
  - **ロボットが自分のいる環境を分類**
  - **金融商品の銘柄を価格変動パターンで分類**
  - **癌細胞を遺伝子発現パターンで分類**

- **教師付き分類←→教師無し分類**
  - **カテゴリがあらかじめ与えられている← 教師付き**
  - **新規カテゴリの自動発見← 教師無し**

主な研究対象：遺伝子発現プロファイル

環境刺激→

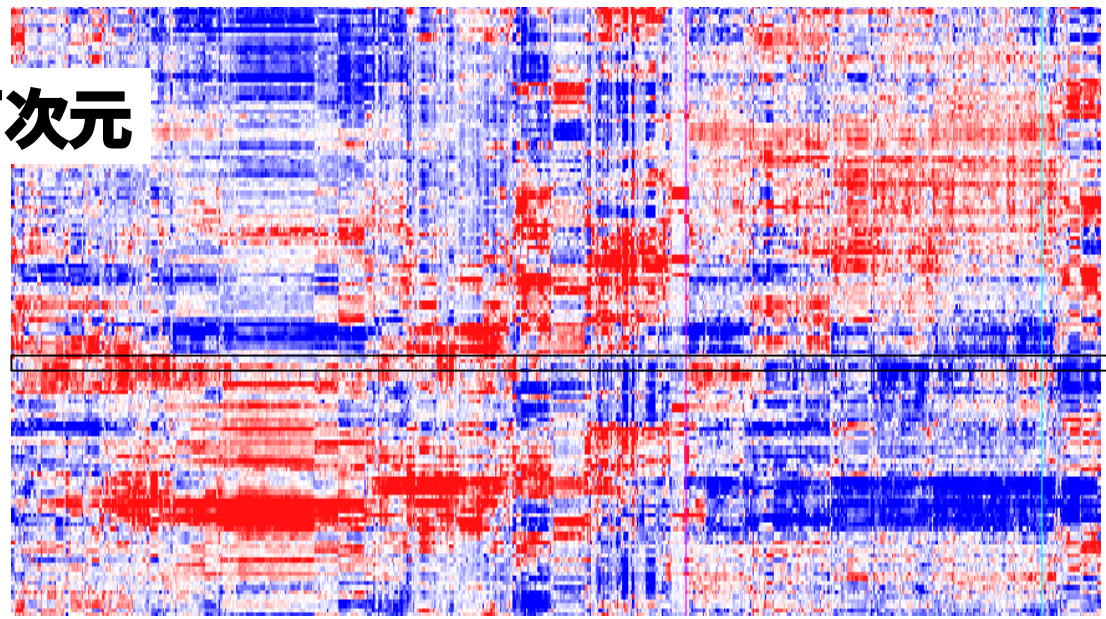


→転写→

→タンパク質生成→ 生命活動

数十～数百次元

症例



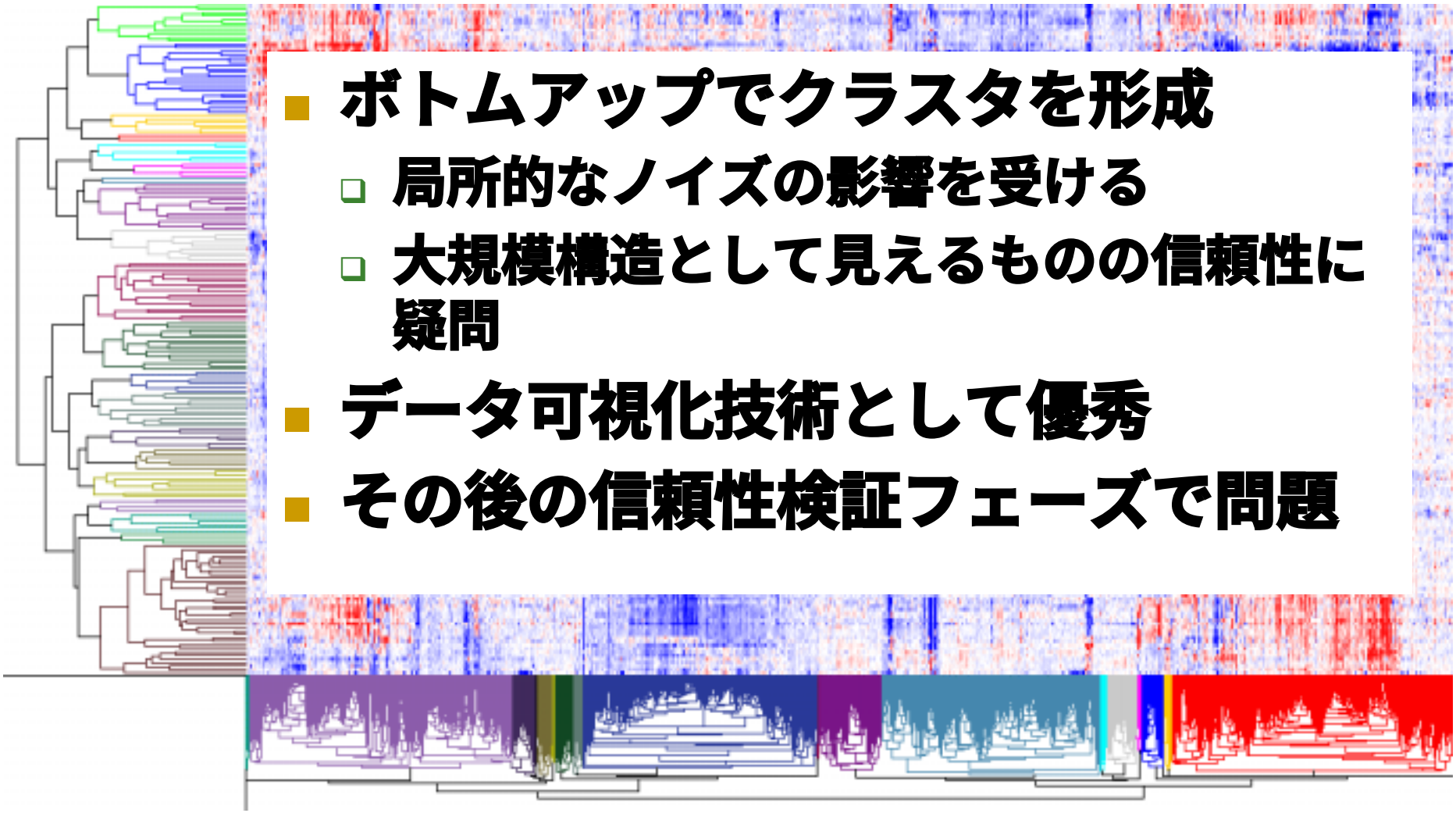
遺伝子　数百～数千次元

■：高発現　□：平均的　■：低発現

従来技術

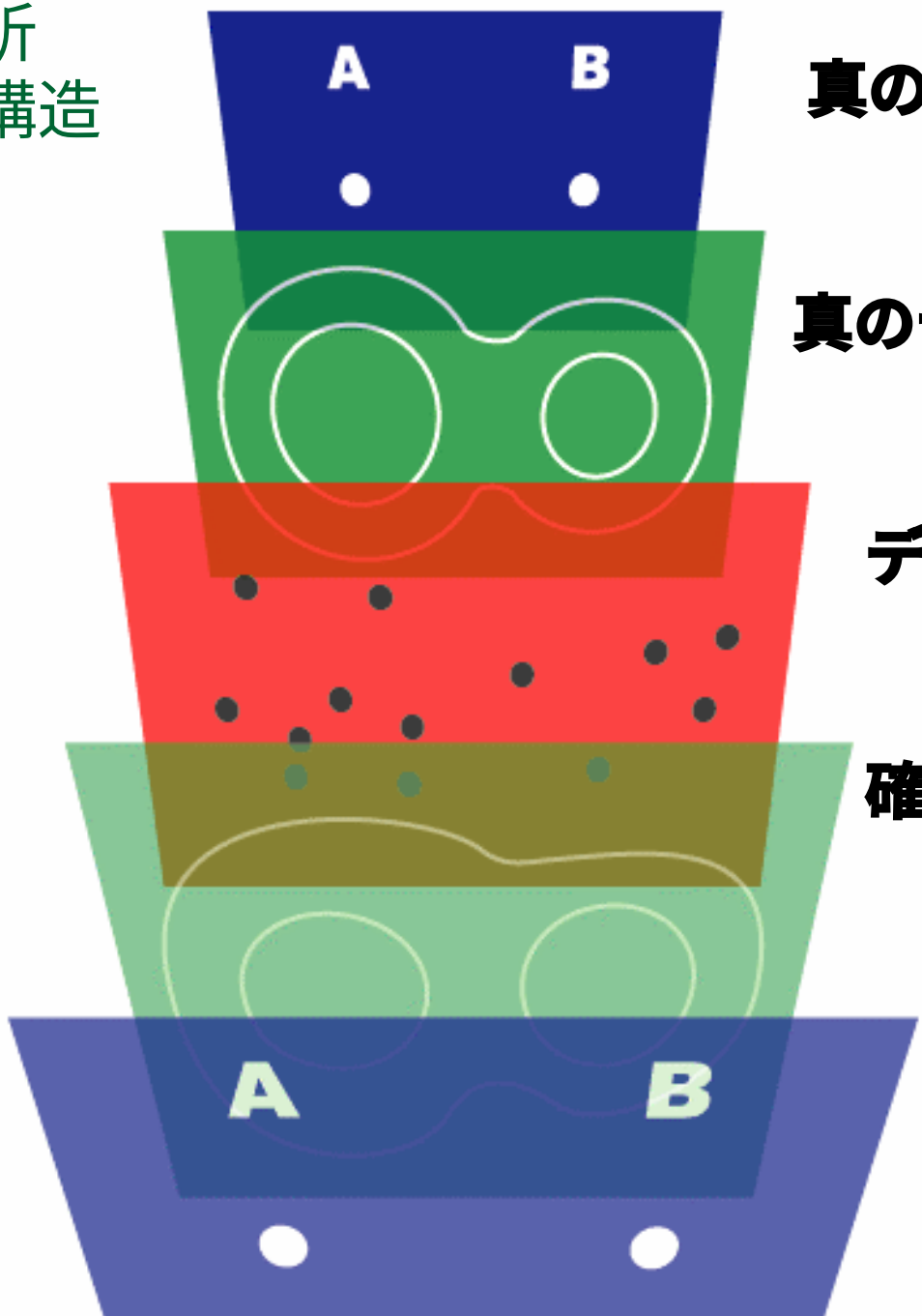
- **階層化クラスタリング**
- ***K*-平均法クラスタリング**
- **パラメトリッククラスタリング**

階層化クラスタリングと樹木図



- ボトムアップでクラスタを形成
  - 局所的なノイズの影響を受ける
  - 大規模構造として見えるものの信頼性に疑問
- データ可視化技術として優秀
- その後の信頼性検証フェーズで問題

クラスタ解析
プロセスの構造



真の名義尺度

真のデータ生成過程

データ空間

確率密度推定

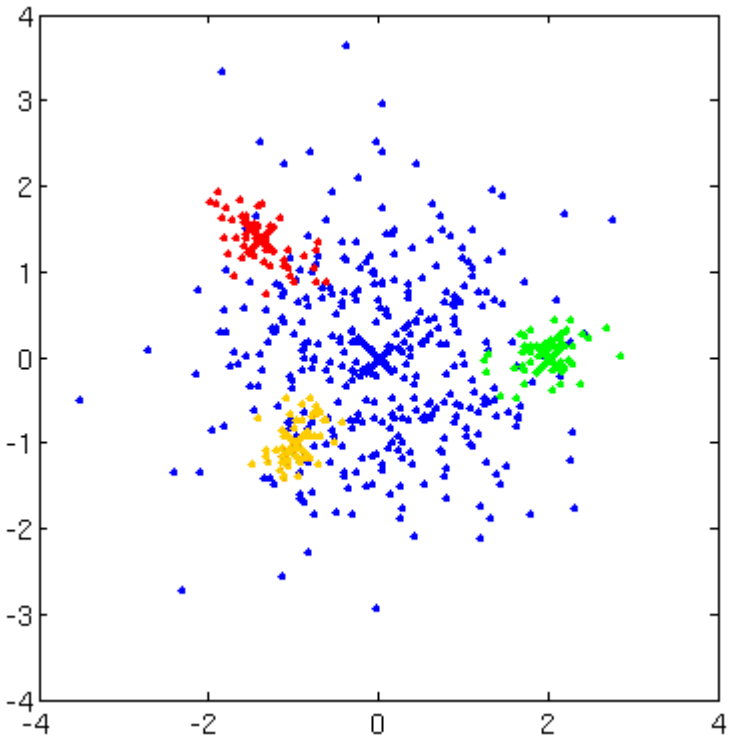
クラスタリングによる名義尺度推定

我々が提案するクラスタ解析技術

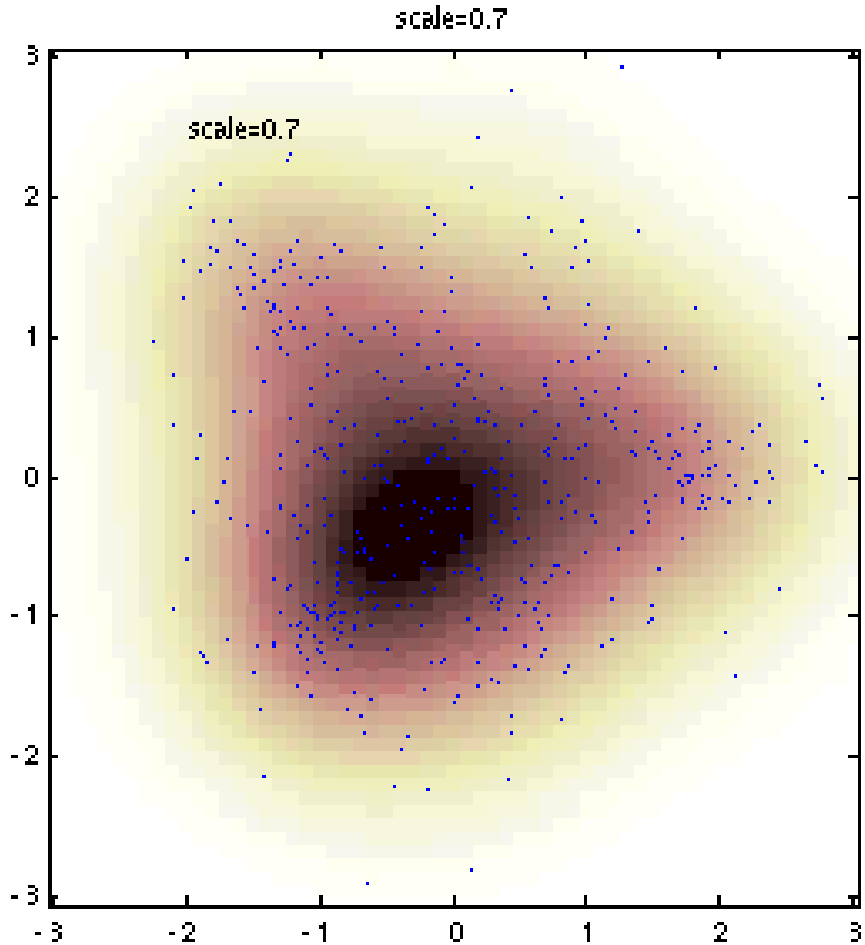
- **ミーンシフトクラスタリング**
  - **「確率分布推定」に基く手法**
  - **ピントをボかしてデータを見る**
  - **カタマリに見えたものをクラスタとして扱う**
- **マルチスケールクラスタリング**
  - **ピントを少しずつ合わせてゆく(マルチスケール)**
  - **さまざまなボケ具合で見えた光景を同時可視化する新規図法 (煉瓦図)**

例1：ノイズの多い２次元データ

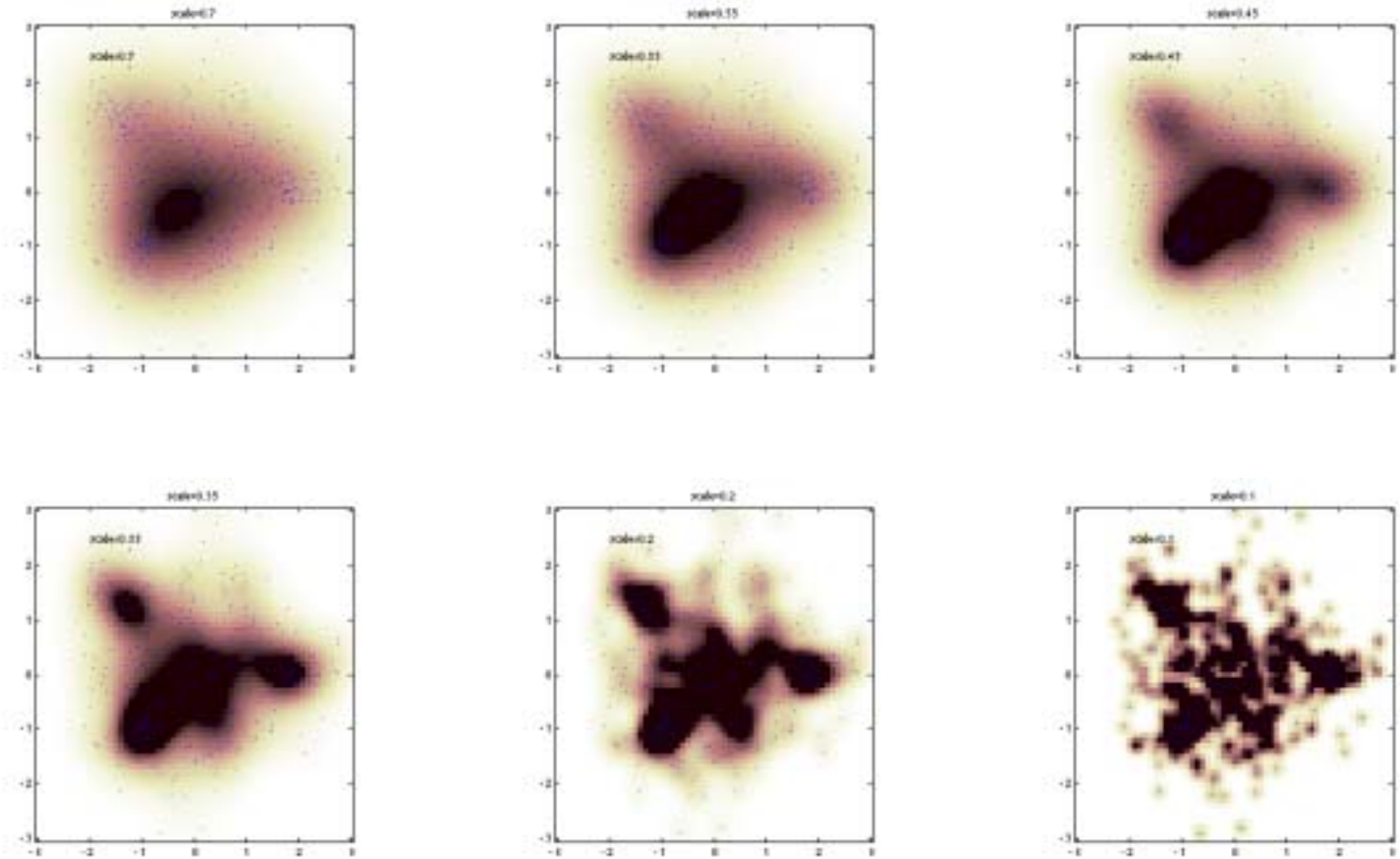
- **500点の二次元データの中に、3つのクラスタ(赤緑黄)と背景ノイズ(青)がある**



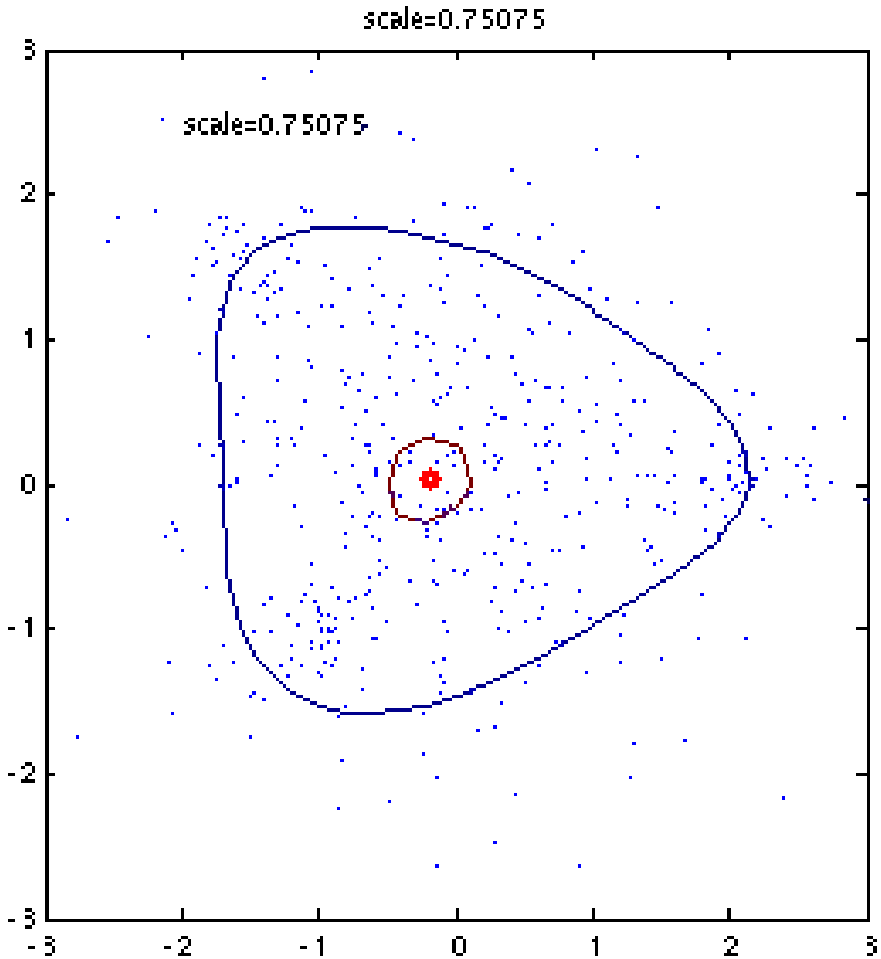
例1：ピントボカシのスケール



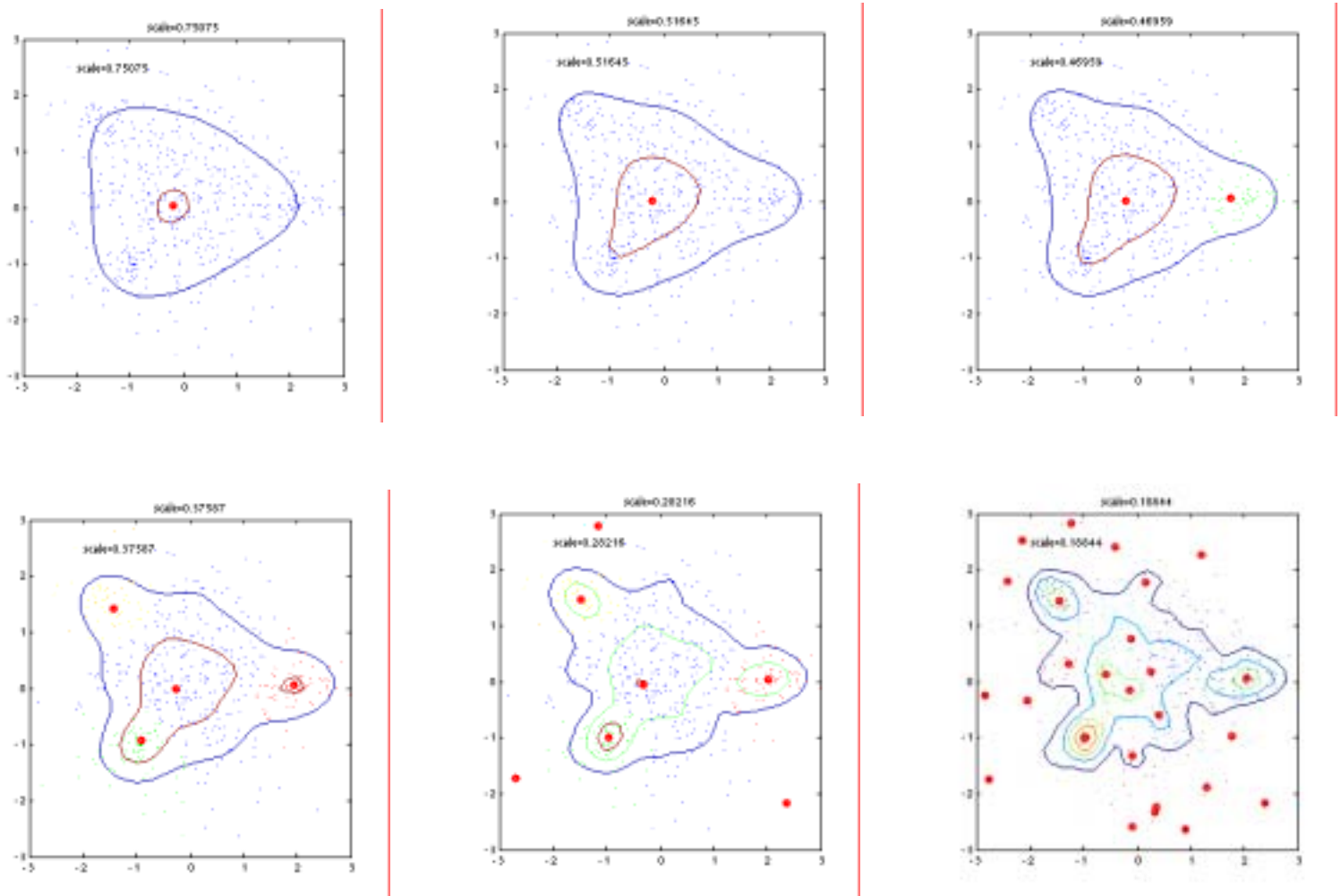
例1：ピントボカシのスケール



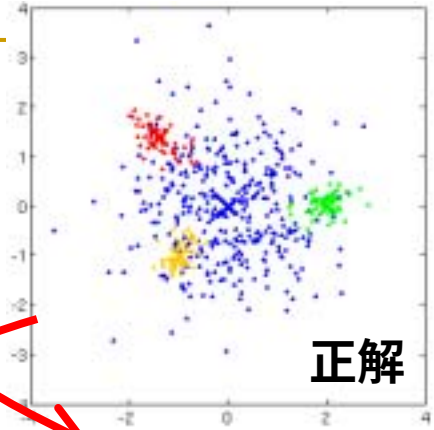
例1：各スケールでのクラスタ

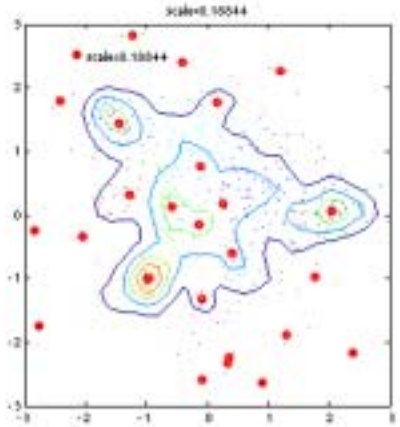


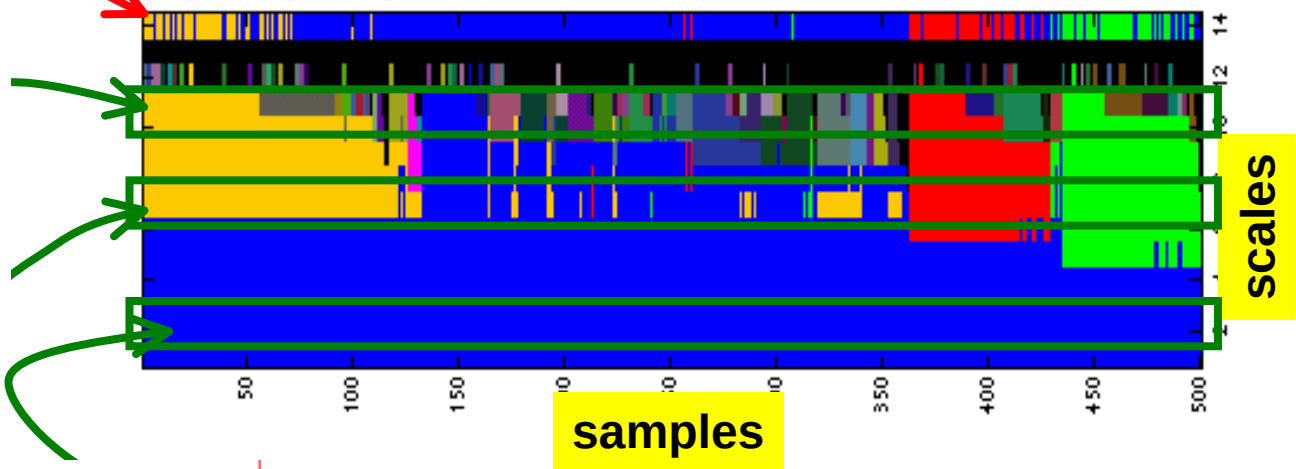
例1：各スケールでのクラスタ

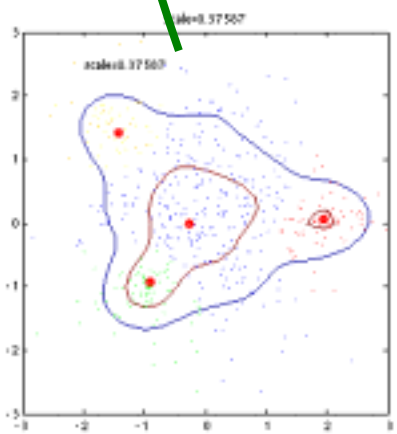


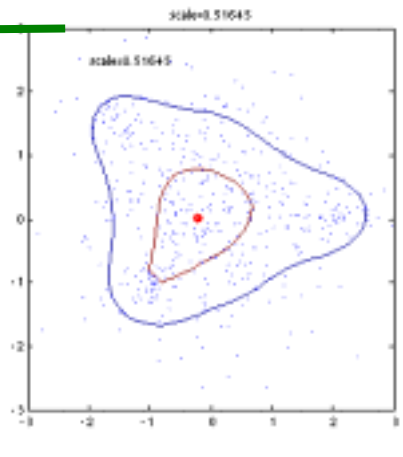
煉瓦図







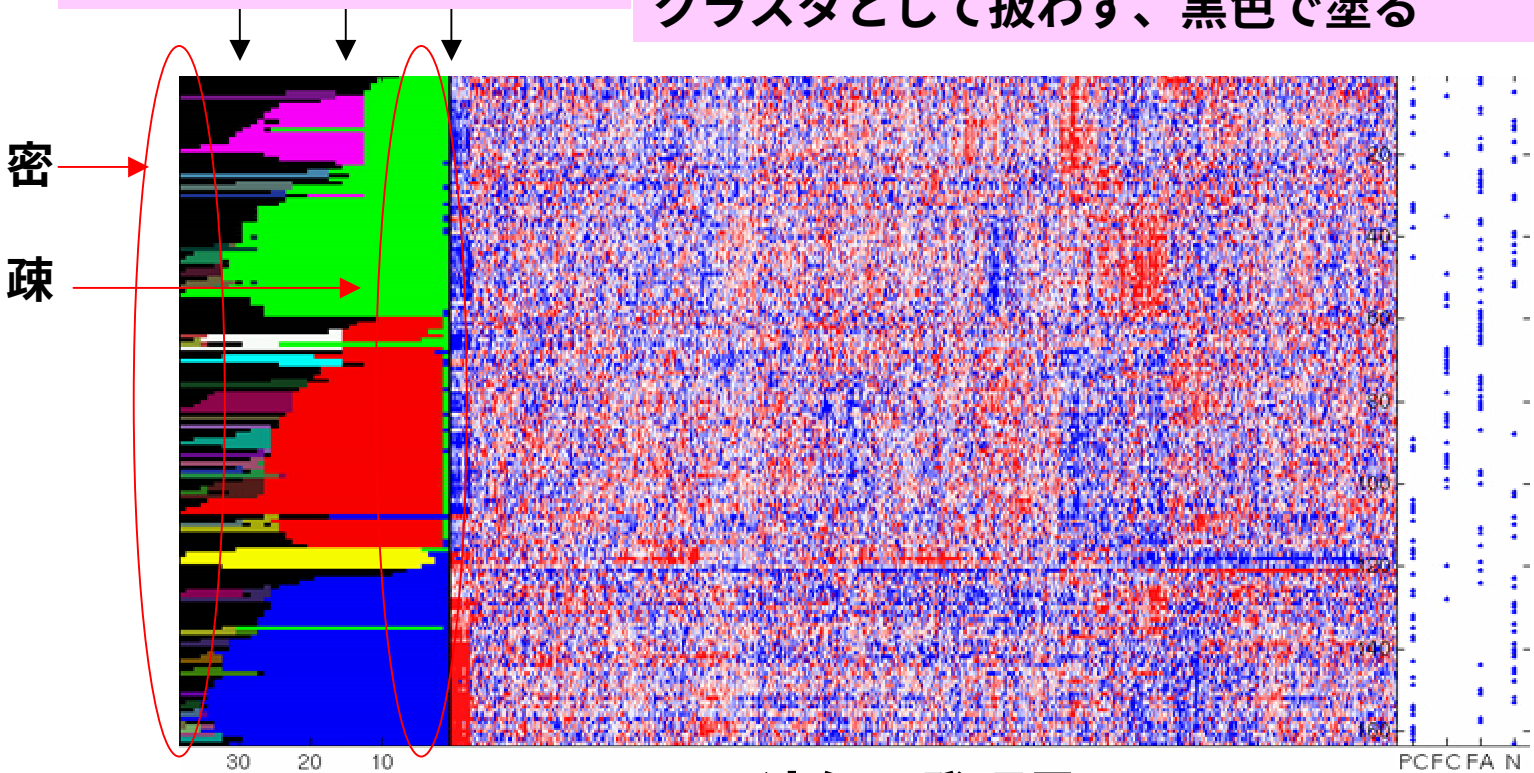




例2：提案手法による
甲状腺癌の遺伝子発現量ベクトル自動分類

各スケールにおける
クラスタ構造を色分け

ただし、
1データが1クラスタをなすようなものは
クラスタとして扱わず、黒色で塗る



煉瓦図による
マルチスケール
クラスタリング結果表示

遺伝子発現量

各症例の
臨床的分類

クラスタ解析に対する提案手法の効果

- **煉瓦図法**
  - **大規模構造から小規模構造までを一目で見ることのできる可視化手法**
  - **クラスタ個数などの、重要なスケールパラメータを手作業で決める必要がない**
- **煉瓦図とミーンシフトクラスタリングの組み合わせ**
  - **ノイズを含むデータに対して、安定なクラスタを検出**
  - **とくに非常に高次元のベクトルデータ (遺伝子発現プロファイル等) に対して有効**