

多変量解析

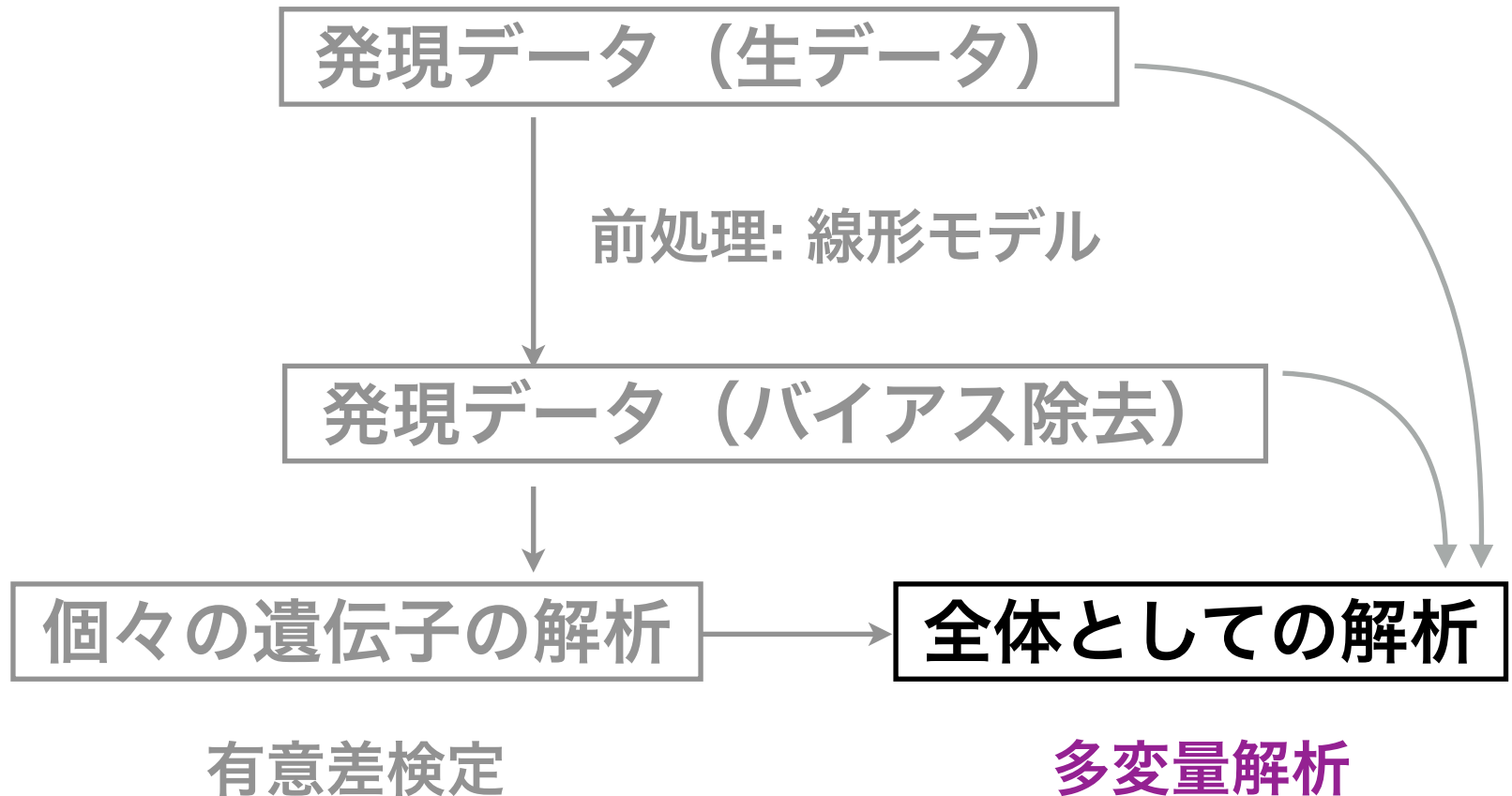
北海道大学大学院農学研究院

(兼) 数理・データサイエンス

教育研究センター

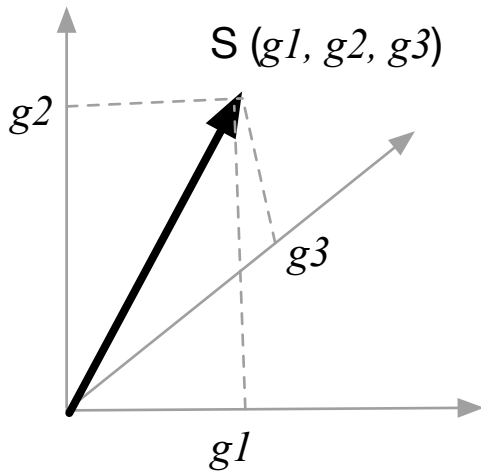
佐藤昌直

RNA-seq解析における 多変量解析の位置付け



モチベーション:

多次元データを人間が解釈できるように補助する



3遺伝子測定データ → RNA-seq: 数千-数万次元

モチベーション:

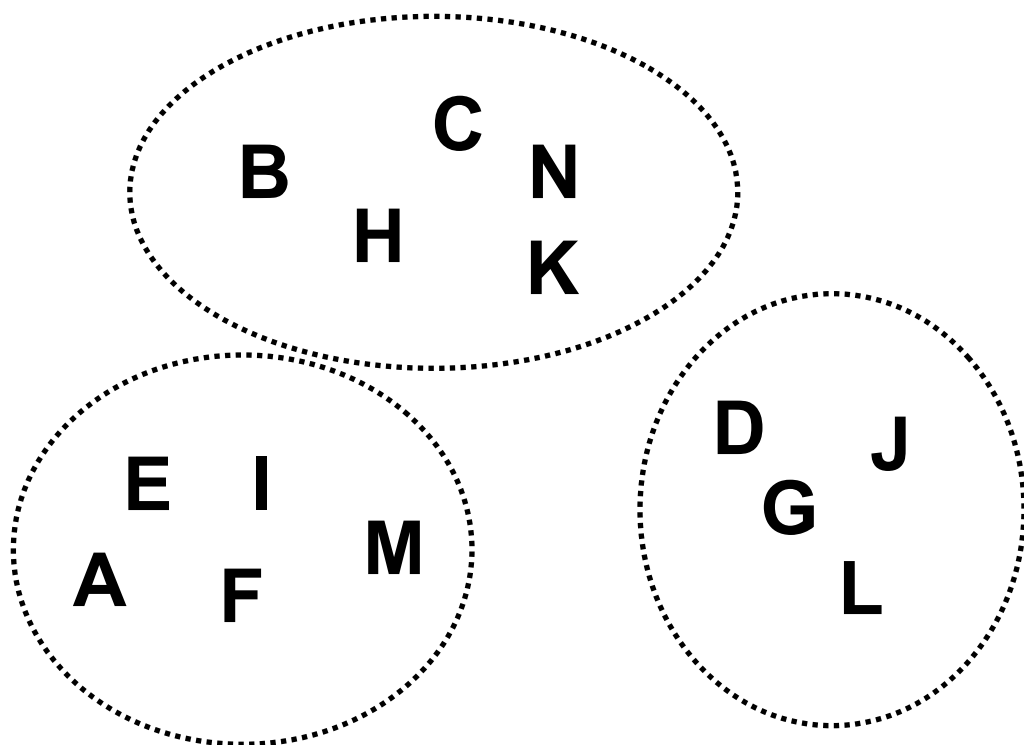
多次元データを人間が解釈できるように補助する

→ 高次元データを低次元で表現し
可視化する

→ 高次元データを統計量で表現し
特徴を選択する/
優先順位をつける

高次元（多パラメーター）データの
認識における問題をどう扱うか？

クラスタリングによる分類



14 プロファイル
→ 3 クラスター

データの確認：目視

Sato_A_thaliana-P_syringae_avrRpt2_6h_expRatio_small.txt

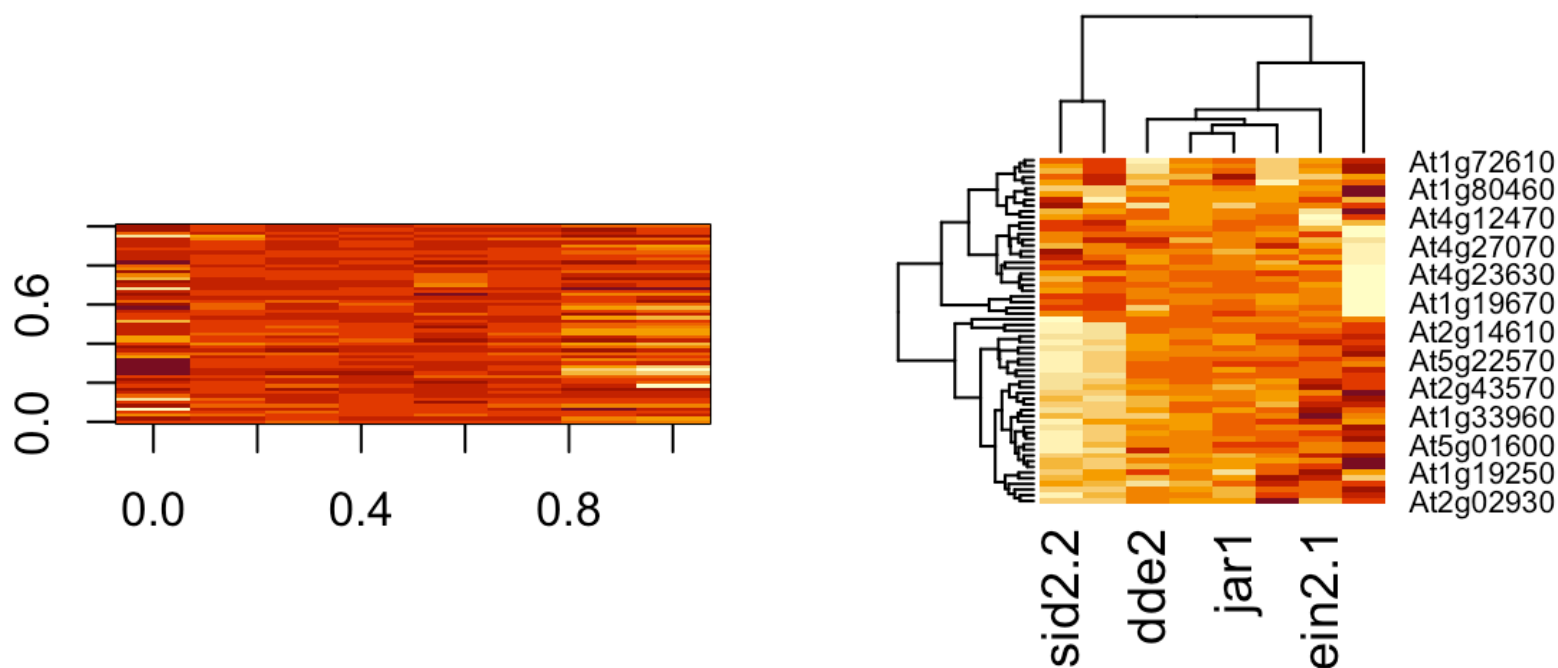
をエクセルで開いて見えます。

データの確認：Rでの可視化

```
inputMatrix <- read.delim(  
  file.path("~/Desktop/gitc/data/MS/dataset/",  
            "Sato_A_thalianaP_syringae_avrRpt2_6h_expRatio_small.txt"  
),  
  header=TRUE,  
  row.names=1  
)  
str(inputMatrix)      # データの構造を確認する  
image(t(inputMatrix)) # カラーコードで行列データをそのまま可視化  
  
heatmap(as.matrix(inputMatrix)) # 階層クラスタリングとヒートマップ
```

ex601-0

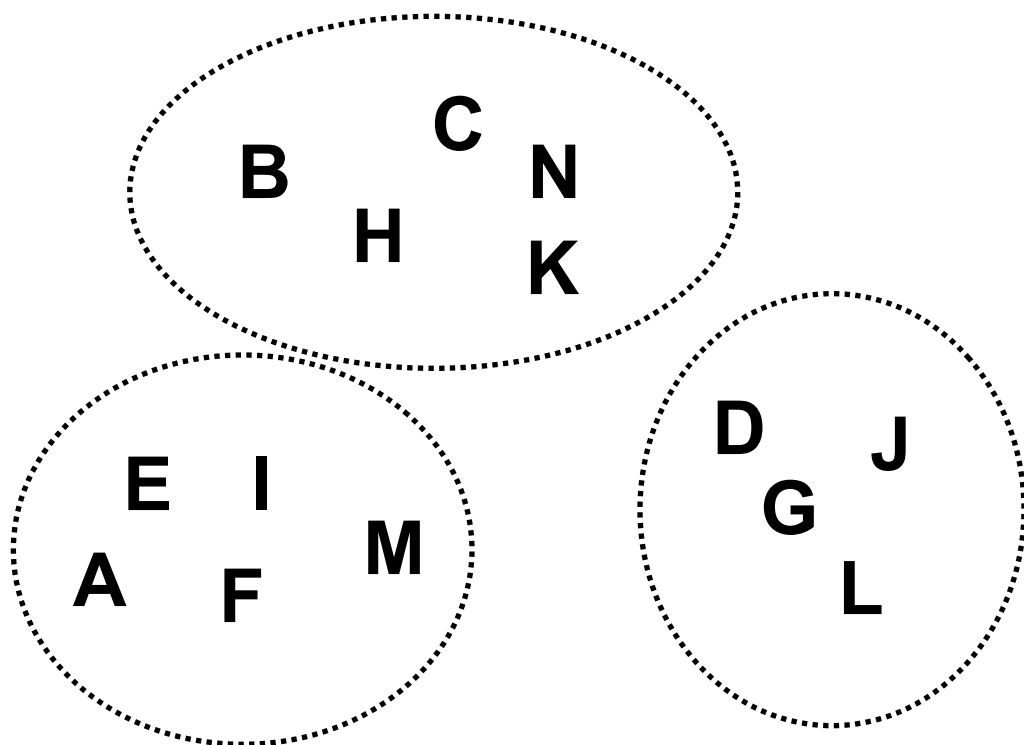
Rは簡単に「何か」を出力してくれる！



統計の基礎知識とRへの正しい命令が必要

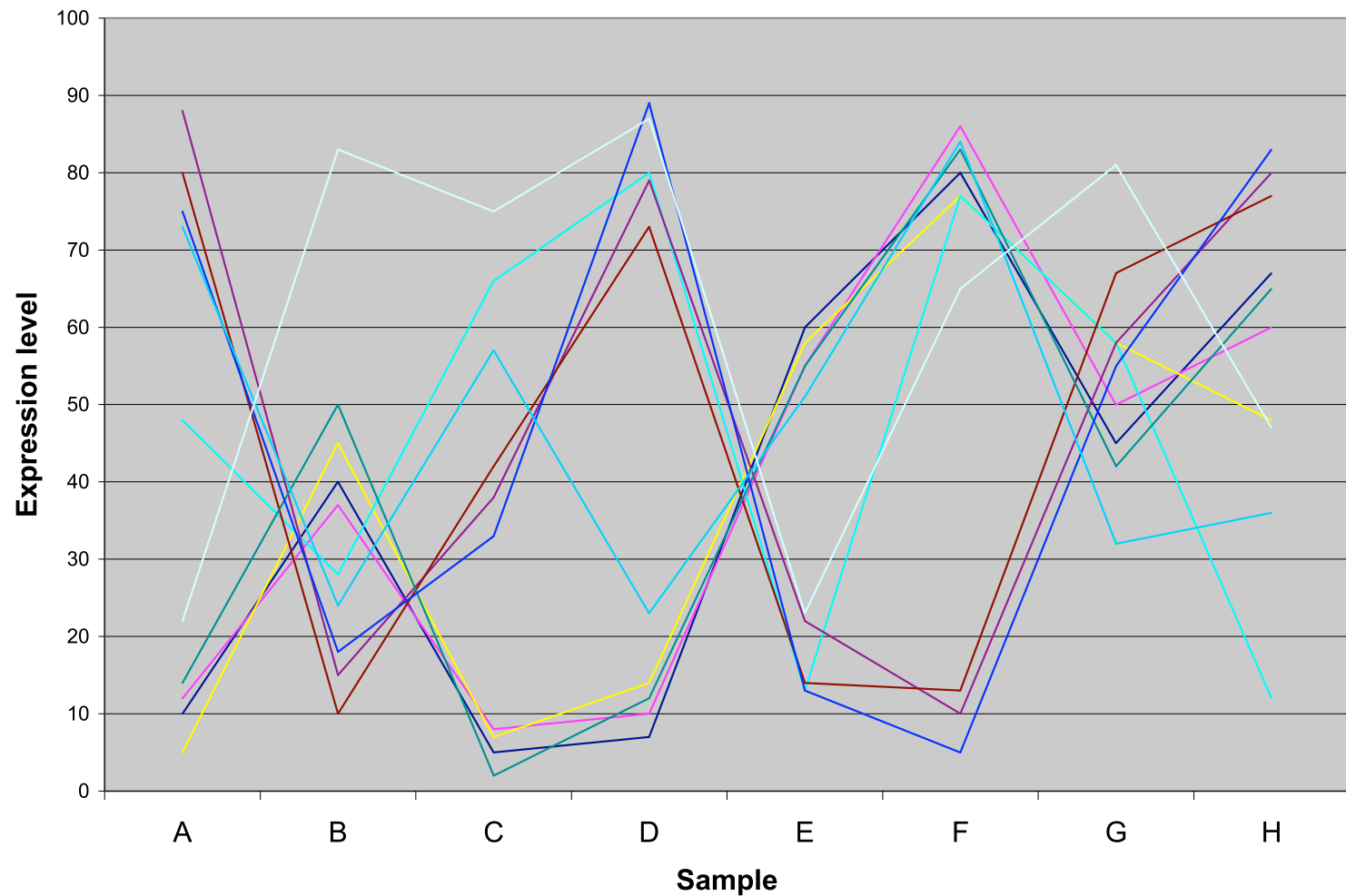
高次元（多パラメーター）データの
認識における問題をどう扱うか？

クラスタリングによる分類



14 プロファイル
→ 3 クラスター

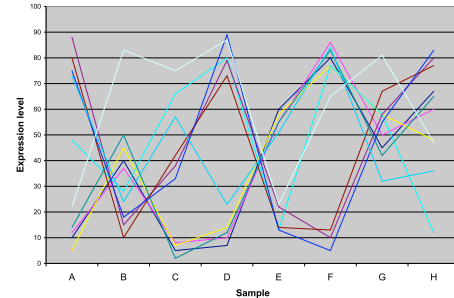
人間は低次元データだと パターン認識するのは得意



コンピューターにどうデータを渡せば この問題をどう扱えるか？

人間

遺伝子発現プロ
ファイルの比較

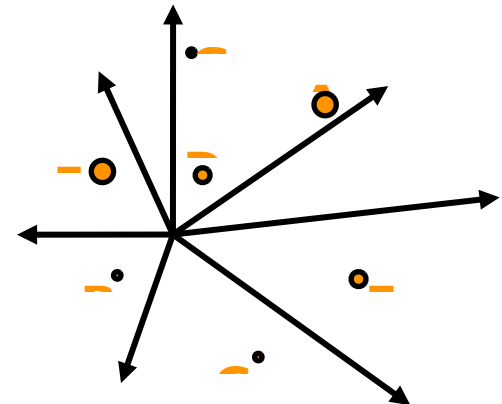


問題定義の変換

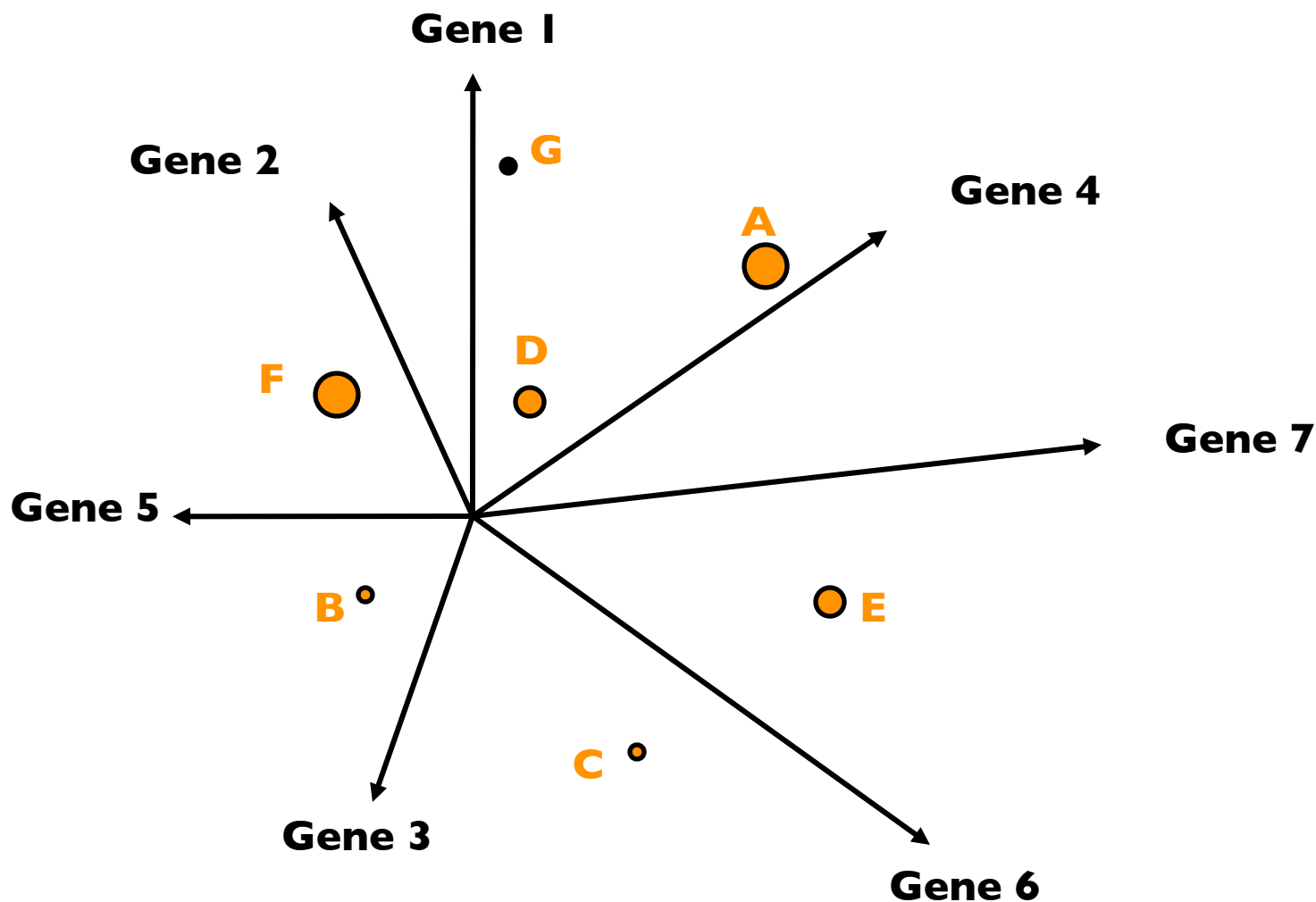
(生物学の問題を数学の問題に置き換える)

計算機

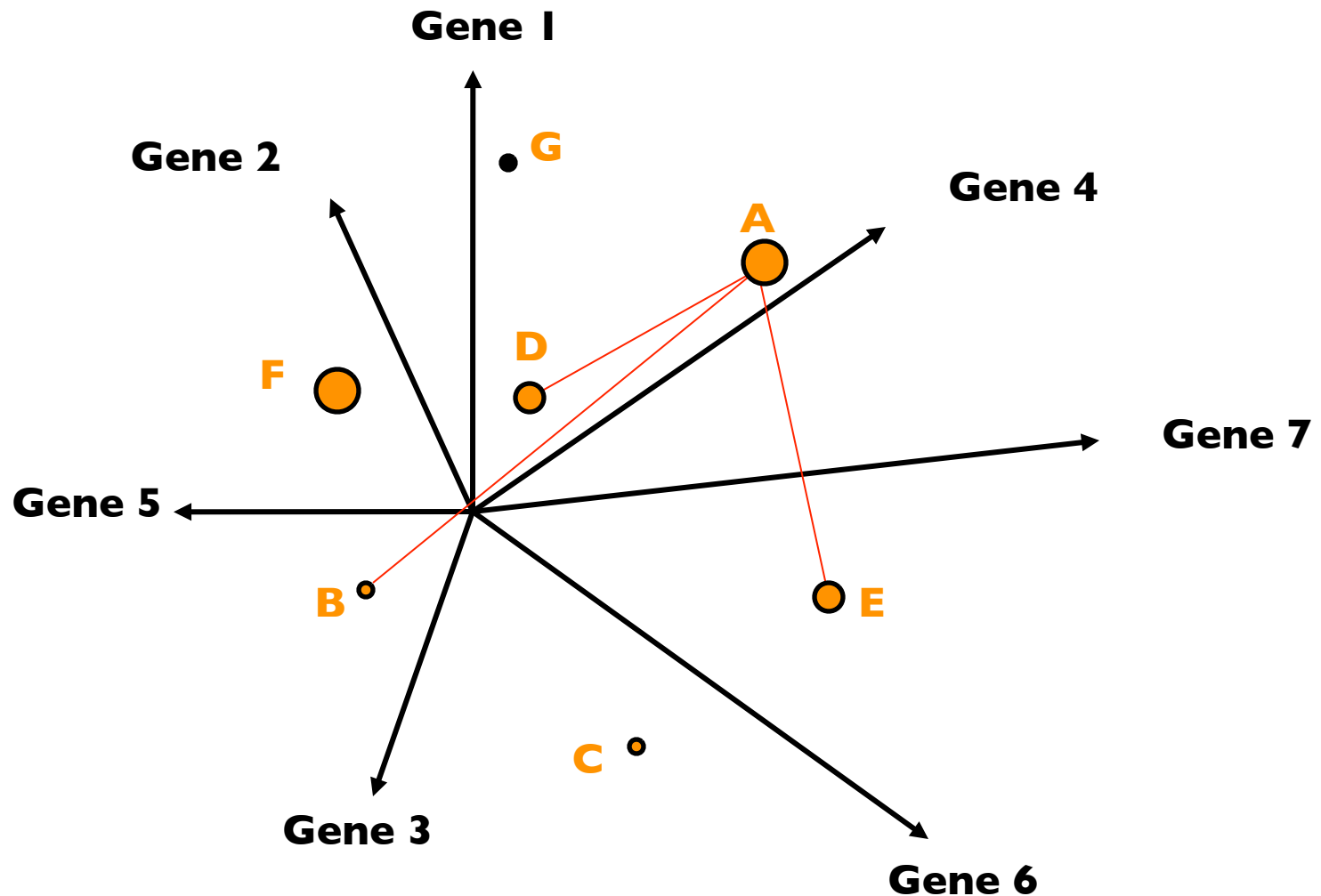
ベクトル等の
数学で扱える
特徴量を
用いた計算



7次元の遺伝子発現データセット

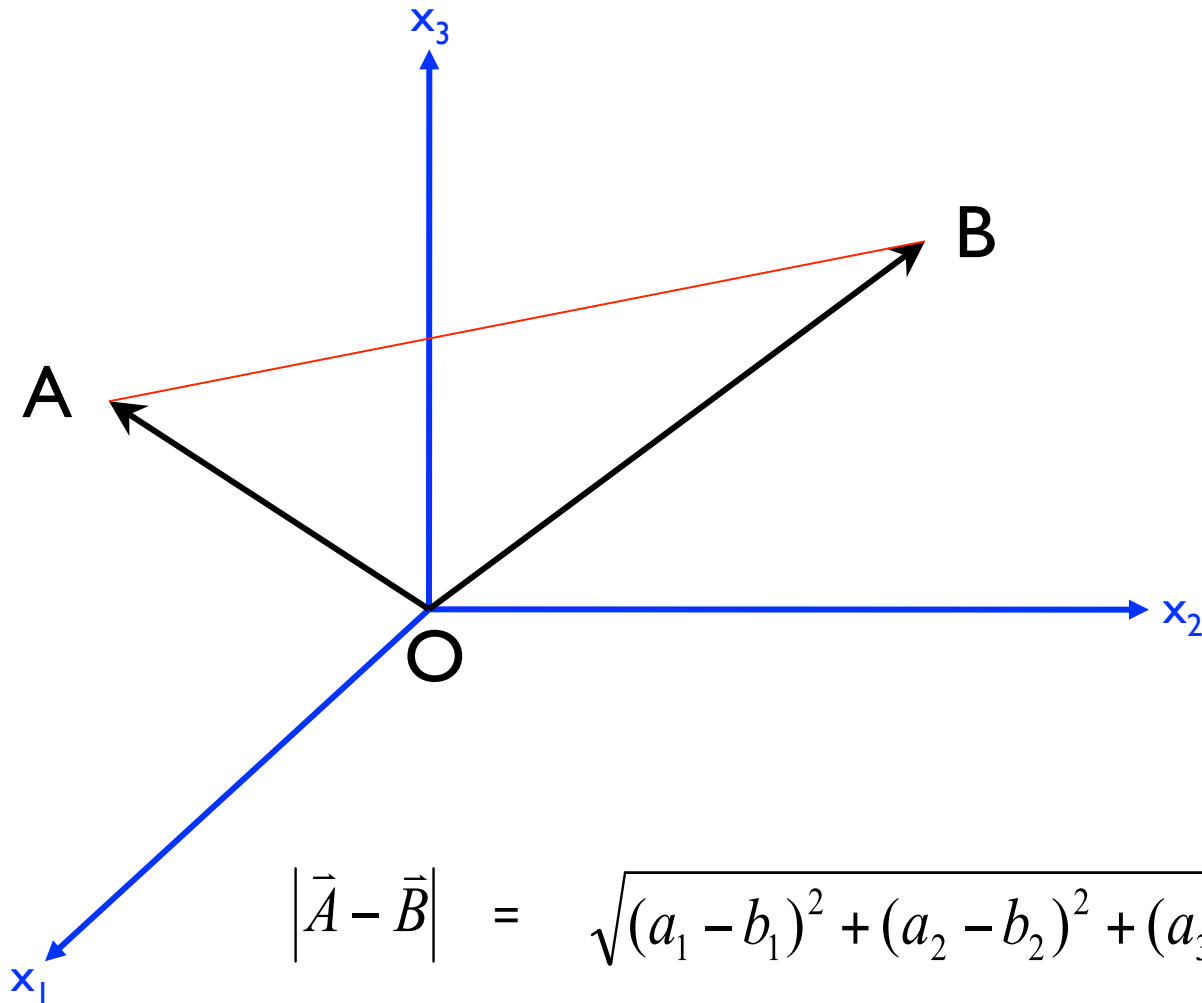


7遺伝子の発現プロファイル間の類似性は 7次元空間での距離によって決まる

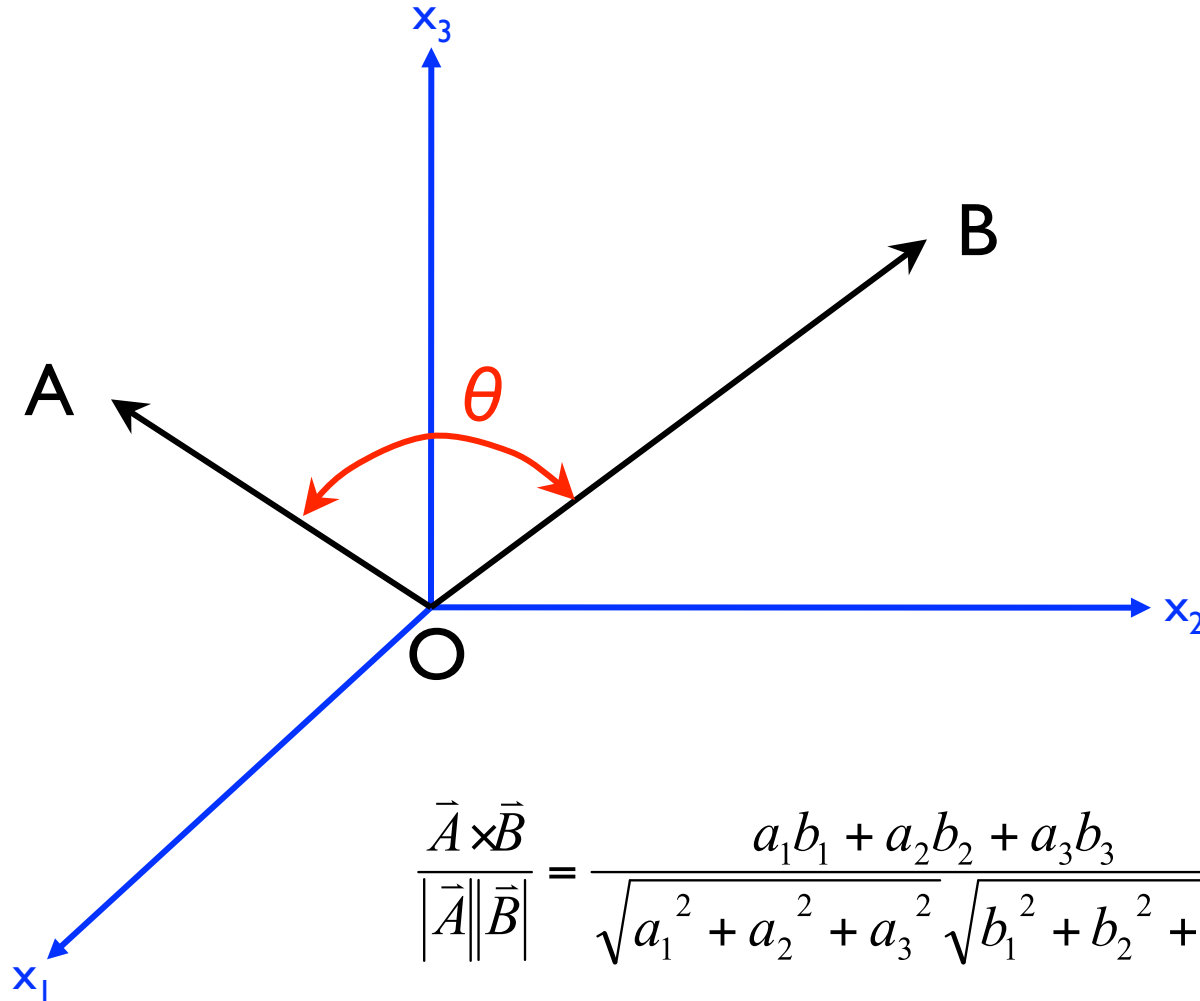


距離の基準を何にするか
距離尺度

ユークリッド距離

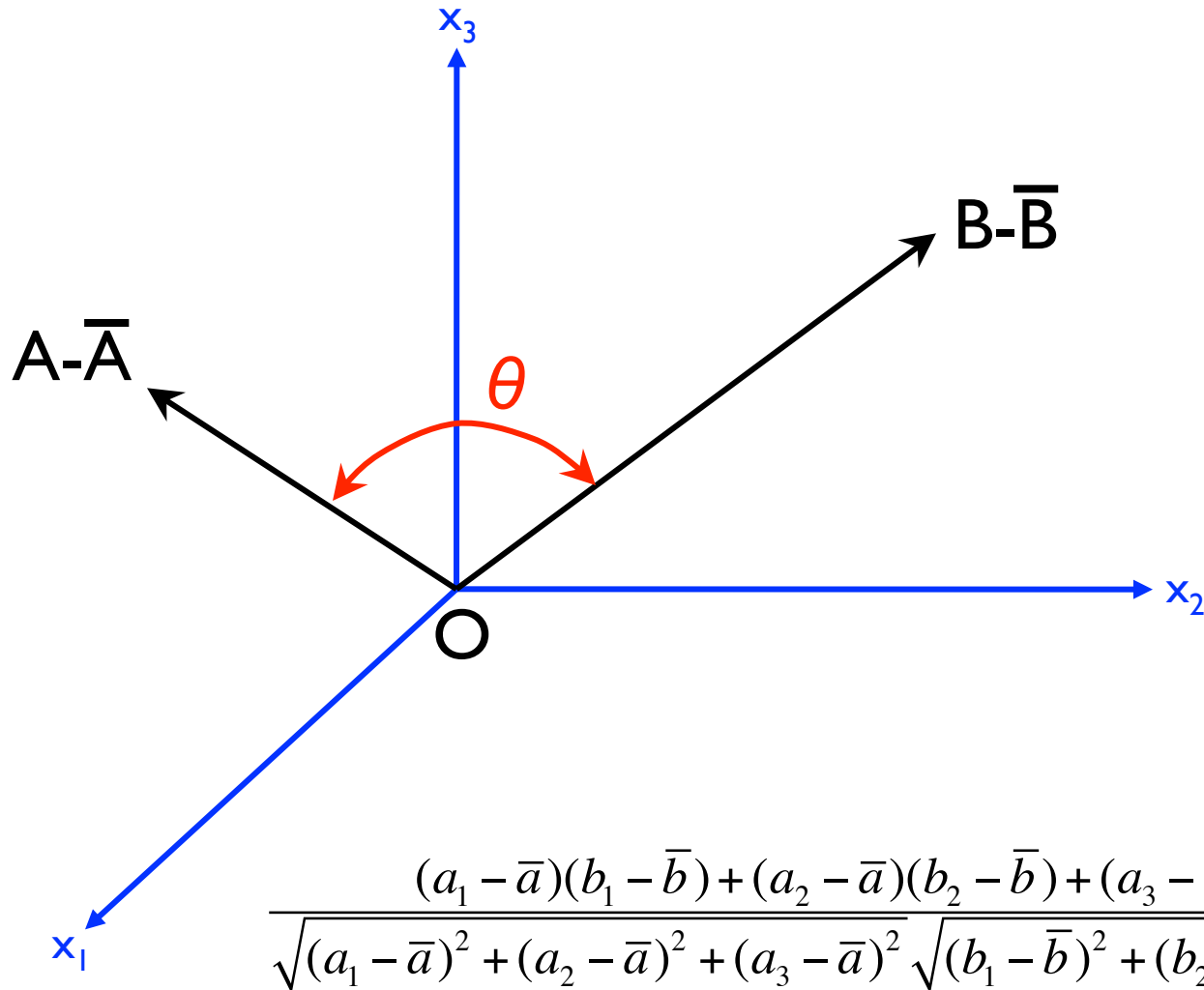


Uncentered Pearson correlation coefficient = $\cos\theta$



相関係数

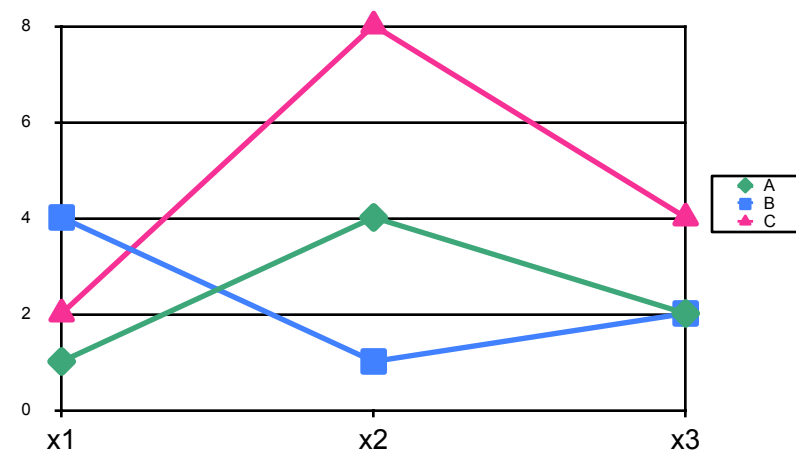
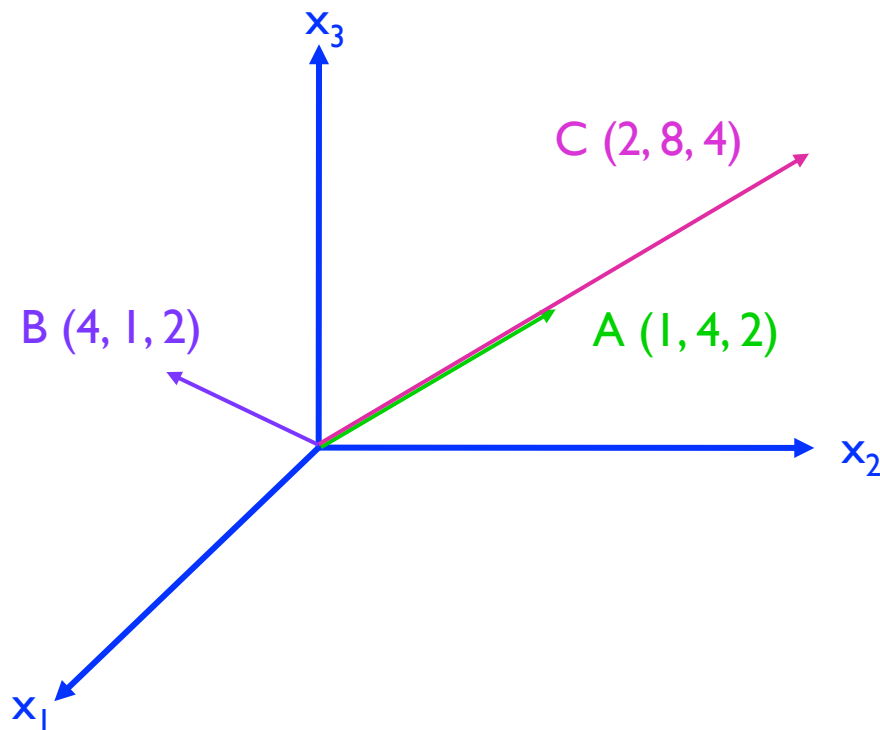
Pearson correlation coefficient



距離尺度の違い→解析視点の違い:

遺伝子発現プロファイルの形と大きさ

- 形: ベクトルの方向
- 大きさ: ベクトルのサイズ



解析視点の決め方:

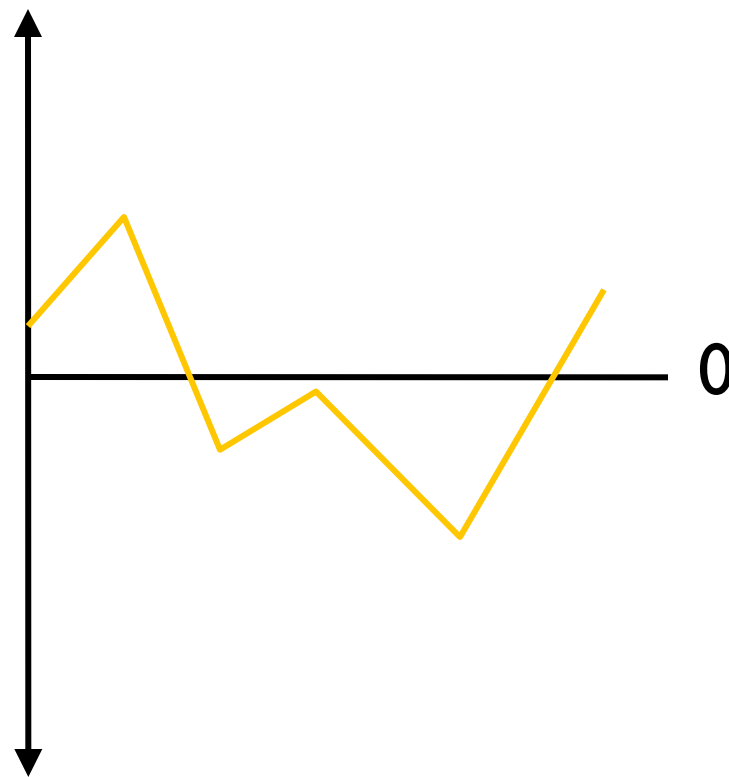
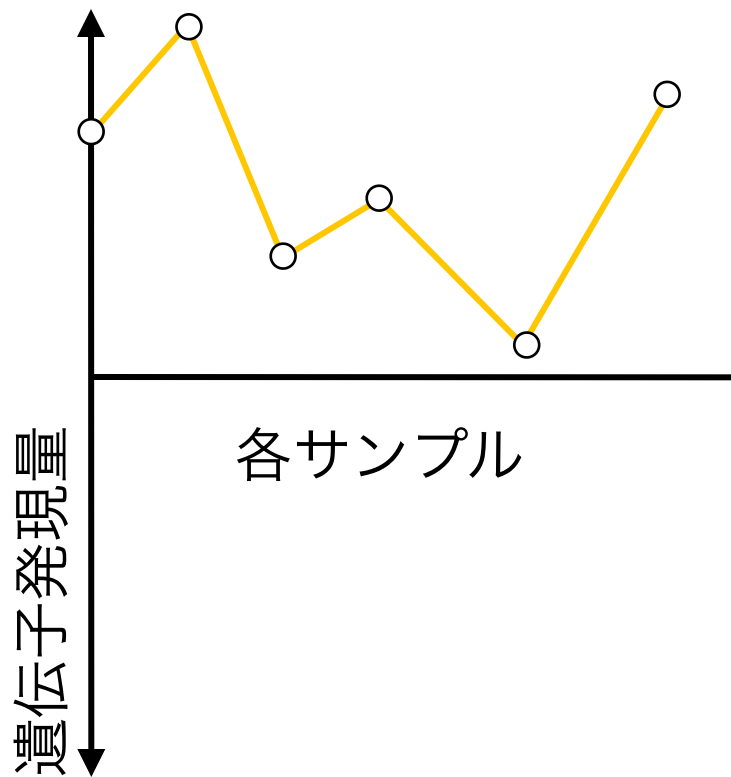
どの距離尺度を使うか？

- どんなプロファイルを同じプロファイルと定義するか？比較したいのは
 - 発現変動の大きさ → ユークリッド距離
 - 発現パターン → コサイン係・相関係数
- 👉 これらを理解するために距離尺度計算の内容を次に確認します

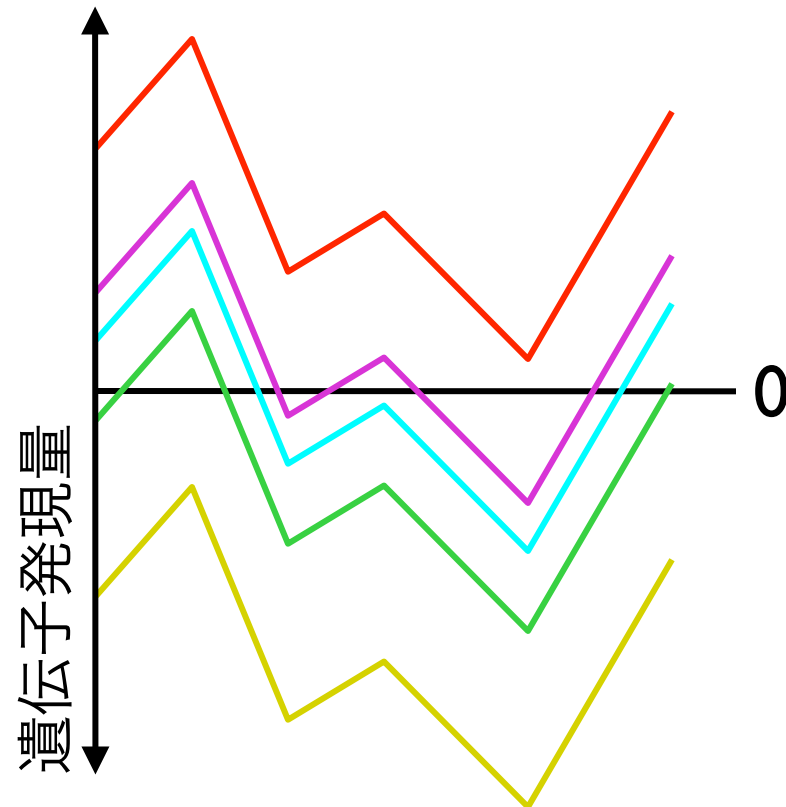
距離尺度計算の基本要素:

- **Centering:** 平均値をゼロにする
- **Scaling:** ベクトルの大きさを1にする

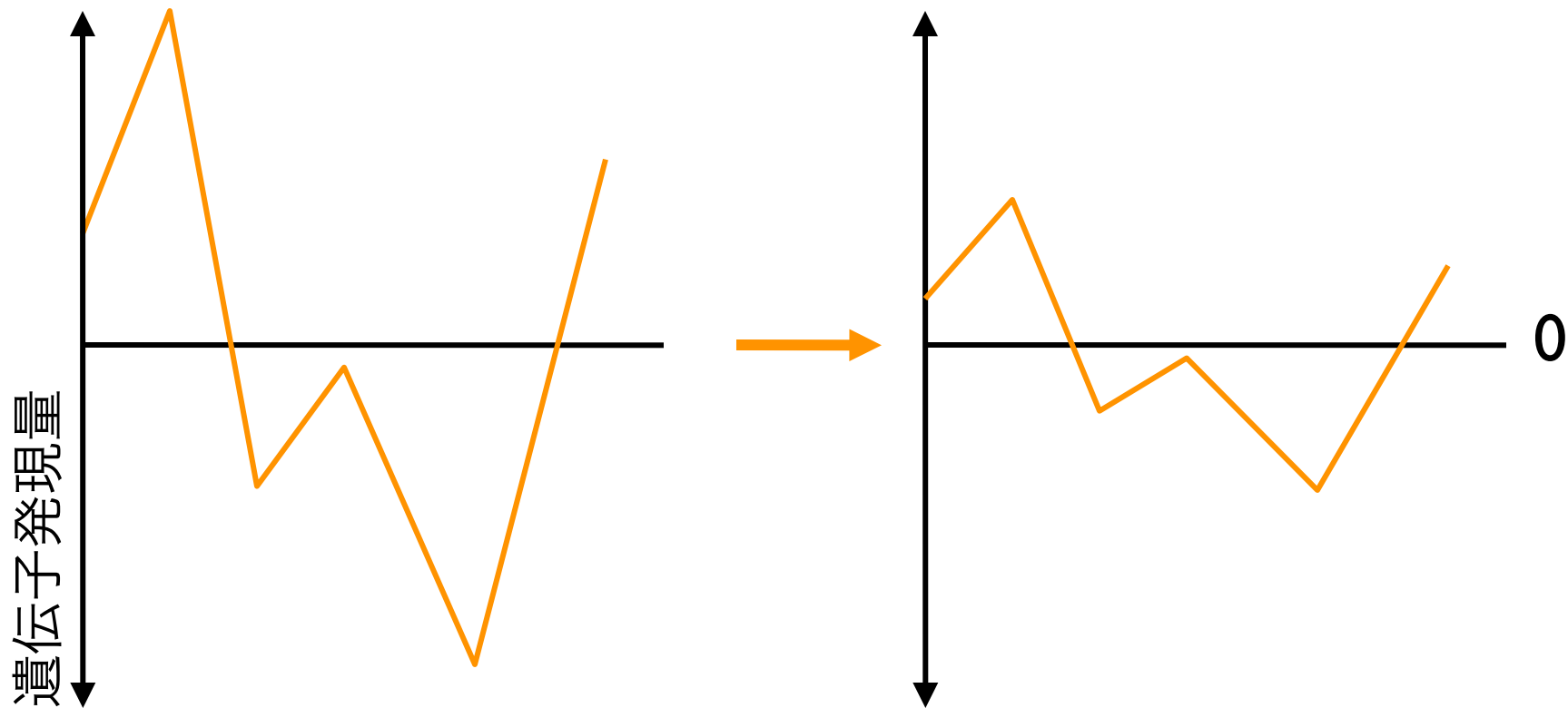
Centering



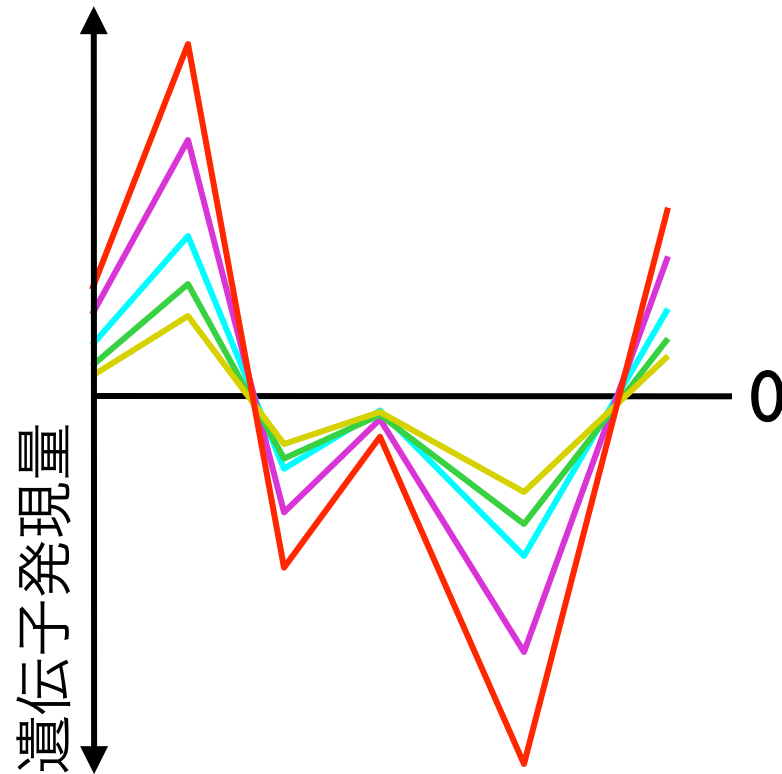
これらはcentering後は
全く同じプロファイルになる



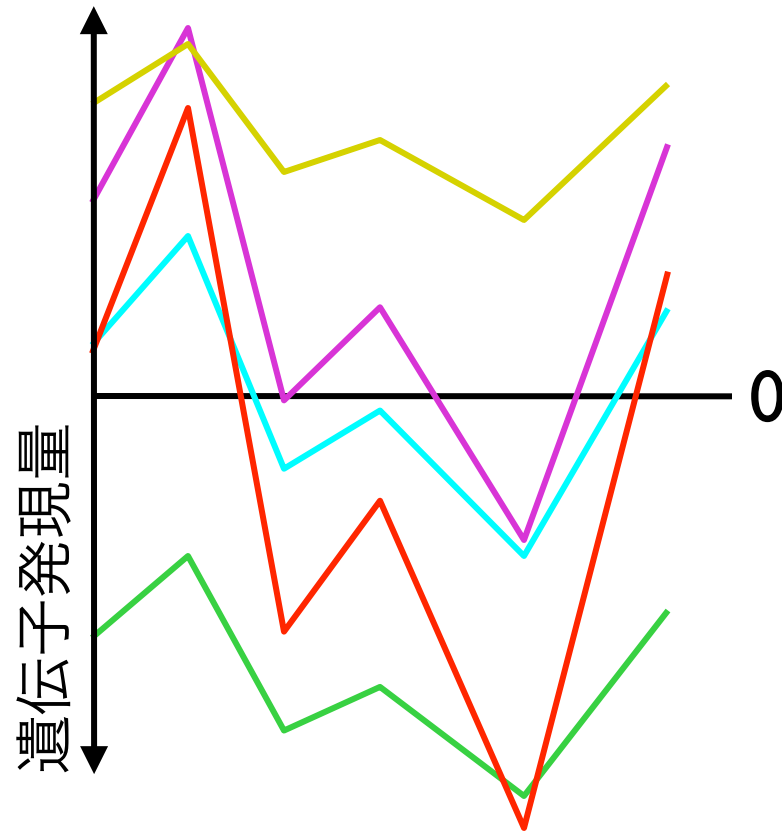
Scaling



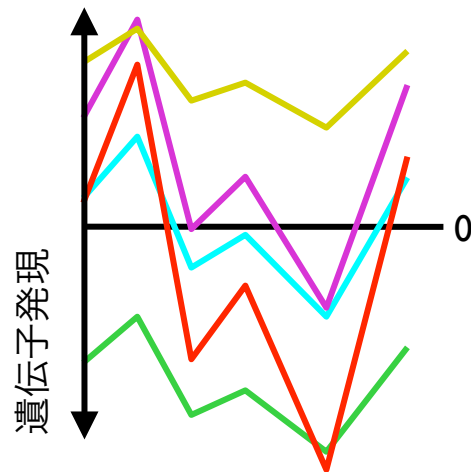
これらはscaling後は
全く同じプロファイルになる



これらはcentering, scaling後は
全く同じプロファイルになる



アルゴリズムに注目: 相関係数の場合



$$(a_1 - \bar{a})(b_1 - \bar{b}) + (a_2 - \bar{a})(b_2 - \bar{b}) + (a_3 - \bar{a})(b_3 - \bar{b})$$

$$\frac{\sqrt{(a_1 - \bar{a})^2 + (a_2 - \bar{a})^2 + (a_3 - \bar{a})^2} \sqrt{(b_1 - \bar{b})^2 + (b_2 - \bar{b})^2 + (b_3 - \bar{b})^2}}{\sqrt{(a_1 - \bar{a})^2 + (a_2 - \bar{a})^2 + (a_3 - \bar{a})^2} \sqrt{(b_1 - \bar{b})^2 + (b_2 - \bar{b})^2 + (b_3 - \bar{b})^2}}$$

センタリング

スケーリング

距離尺度選択における注意点

方法依存的に抽出される特徴:

どのような特徴を認識したいのか/
しているのか意識すること

- 処理間の変動の大きさ: ユークリッド距離
- 処理間のパターンの違い: コサイン係数
- パターンを比較
 - 基準サンプルあり: コサイン係数
 - 基準サンプルなし: 相関係数

多変量解析の実際

教師有るか無しか

(supervised or unsupervised) ?

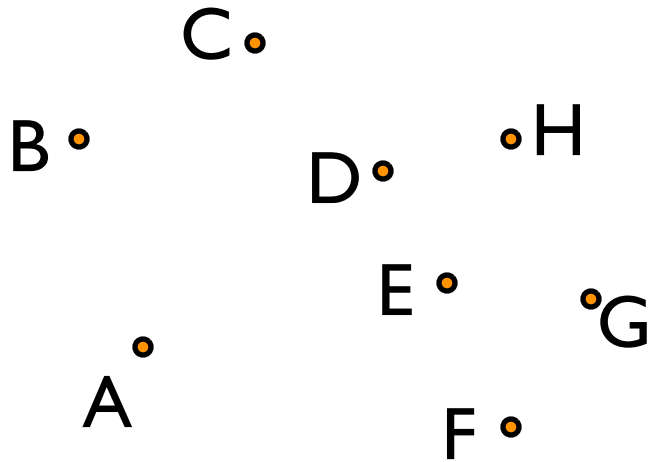
- 事前情報、前提はあるか？
- ある場合はk-means法などの利用を検討

どのような距離行列を使うか？

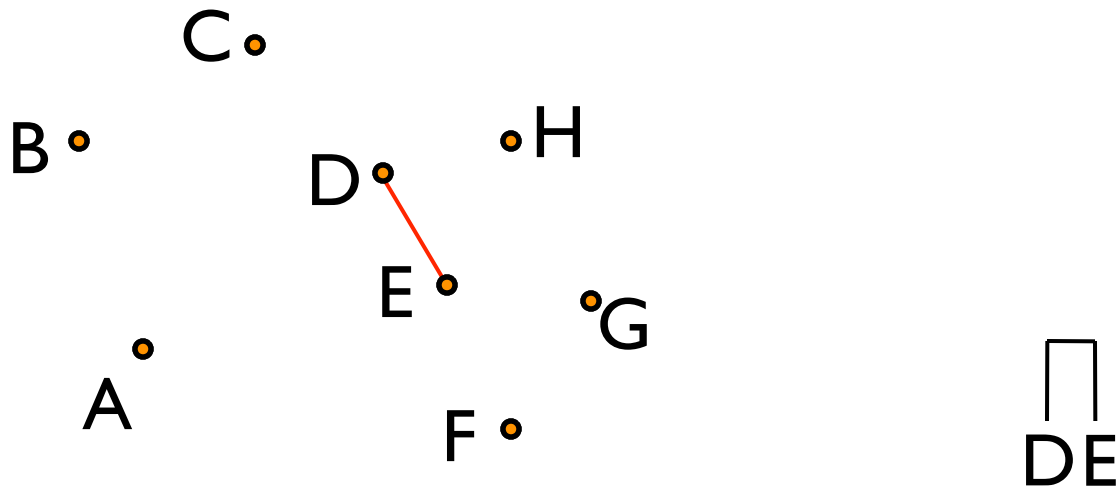
多変量解析の実際

階層クラスタリング

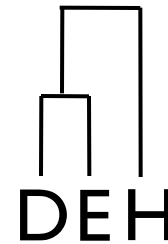
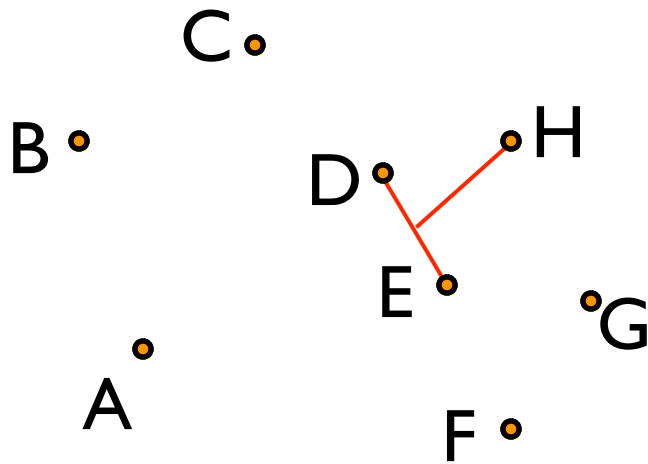
Agglomerative hierarchical clustering



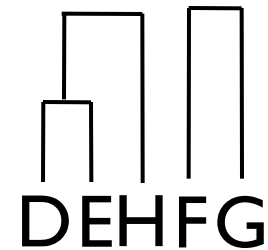
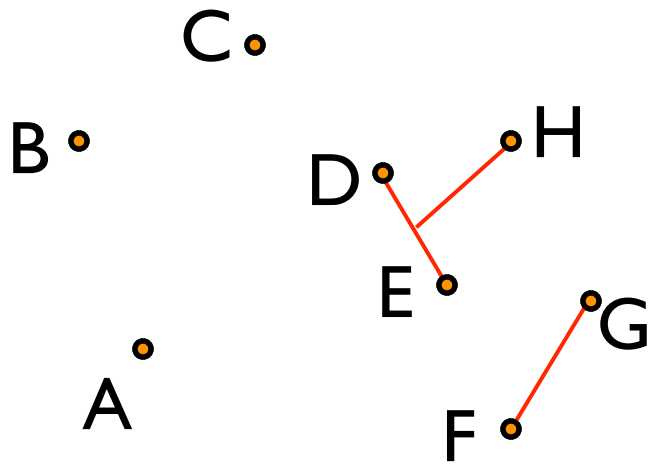
Agglomerative hierarchical clustering



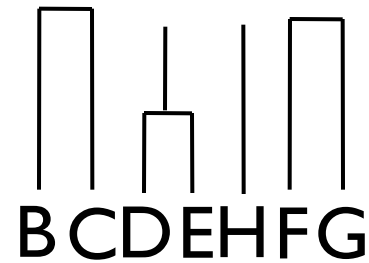
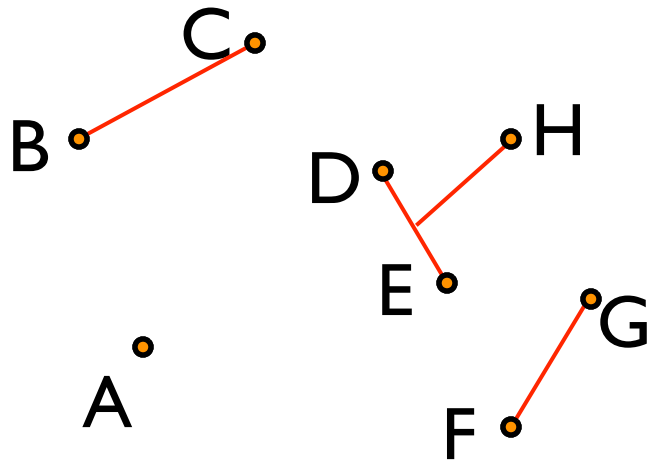
Agglomerative hierarchical clustering



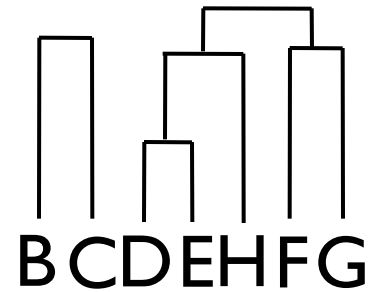
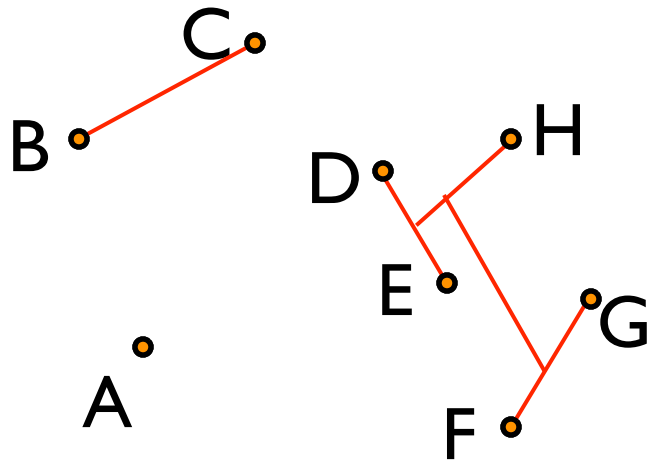
Agglomerative hierarchical clustering



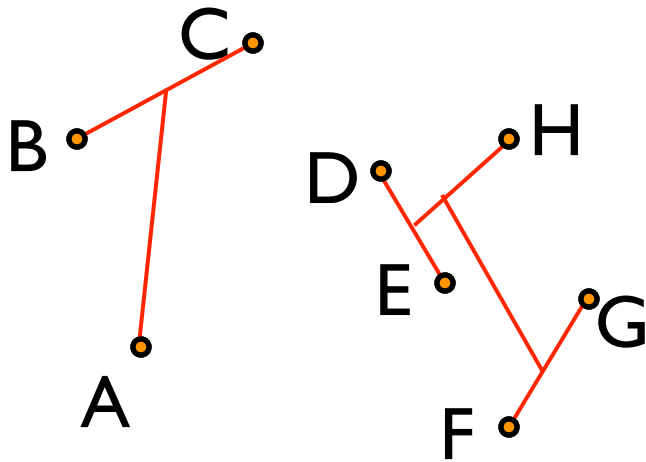
Agglomerative hierarchical clustering



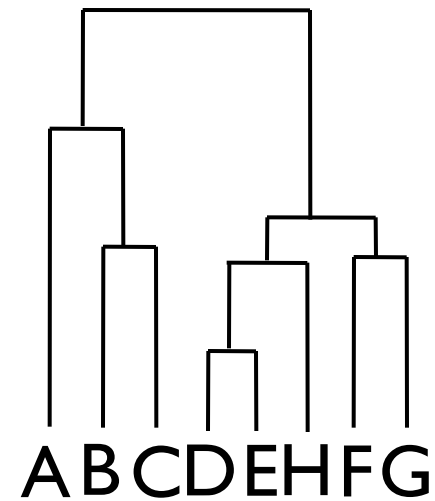
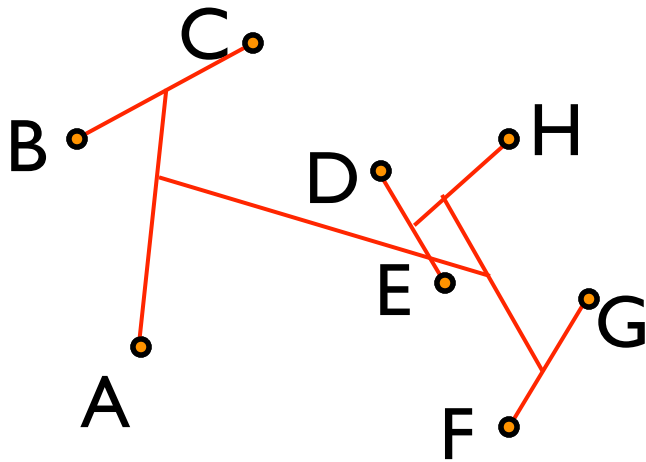
Agglomerative hierarchical clustering



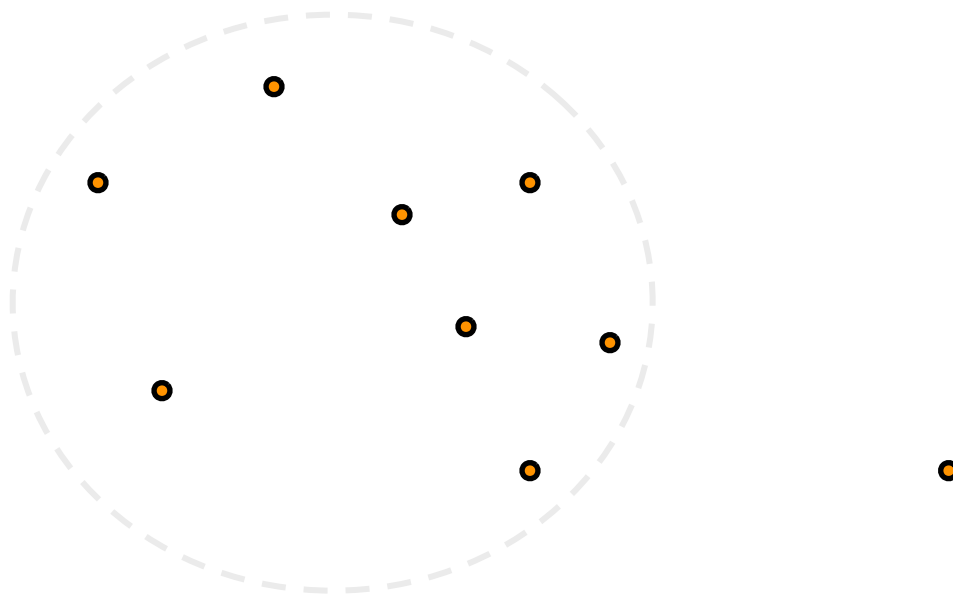
Agglomerative hierarchical clustering



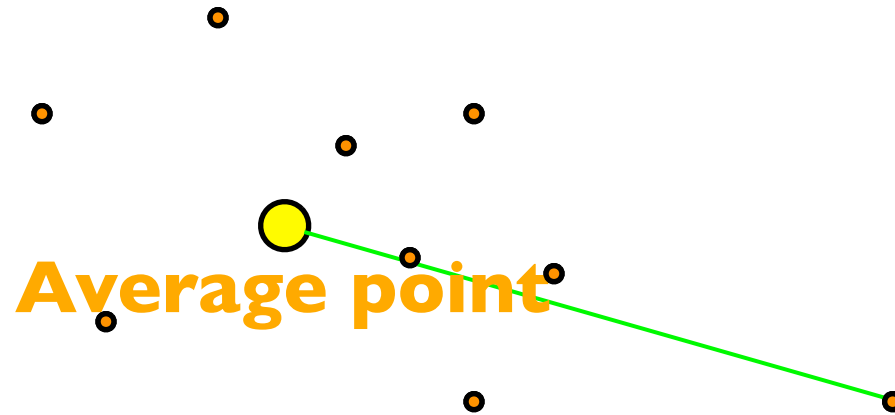
Agglomerative hierarchical clustering



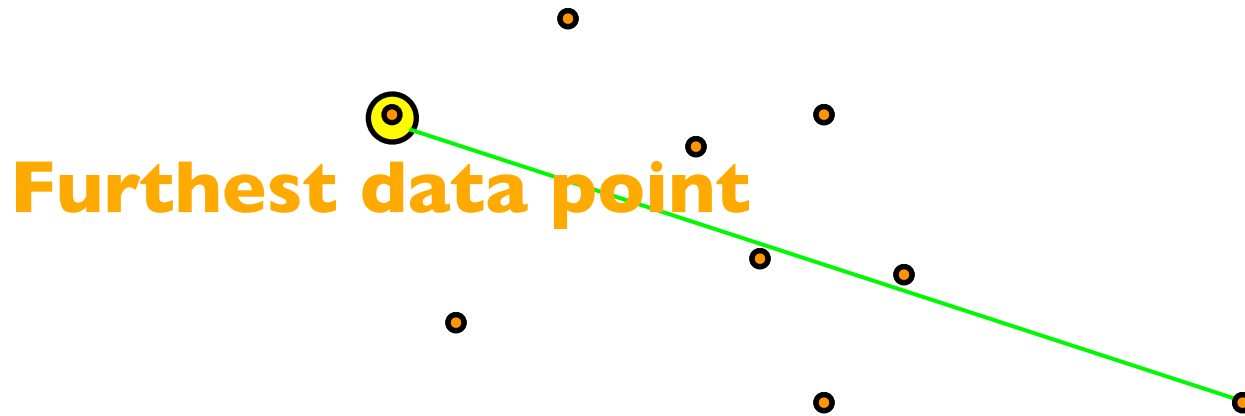
クラスター定義手法



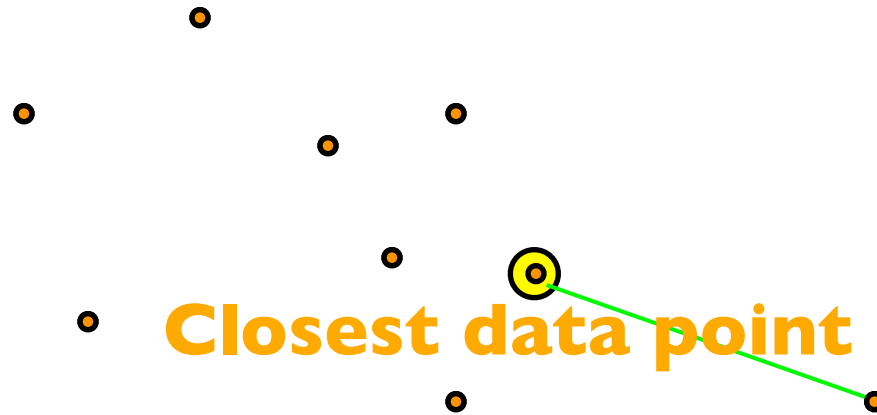
Average linkage



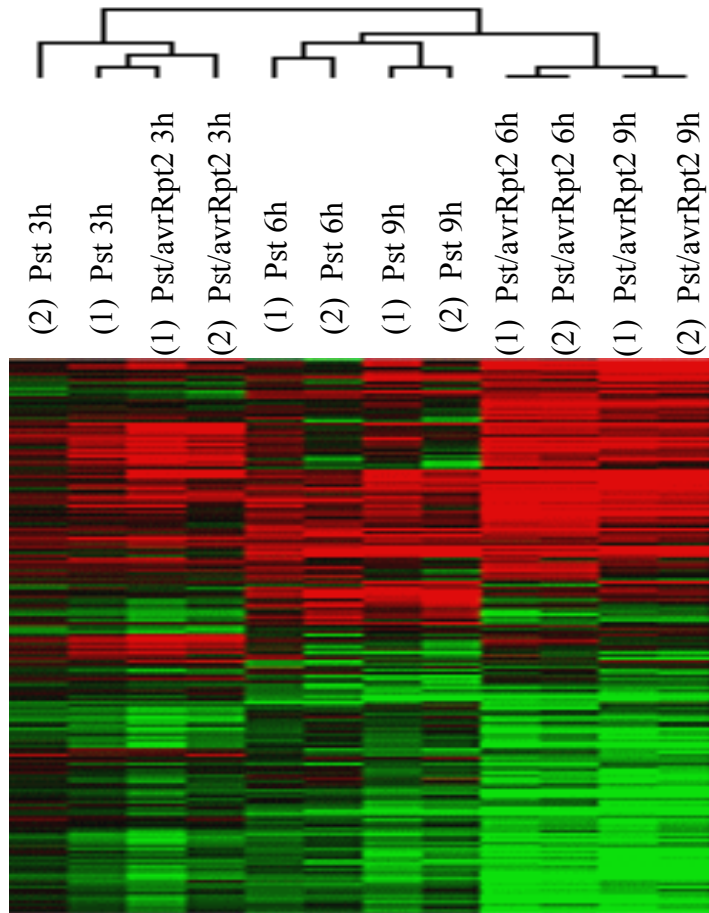
Complete linkage



Single linkage



階層クラスタリングの利点・欠点



利点

- 簡便・直感的な可視化
- デンドログラム長でのクラスター比較

欠点

- サンプル・手順依存的
- 1つの距離を指標

主成分分析

Principal component analysis

主成分分析とは？

モチベーション:

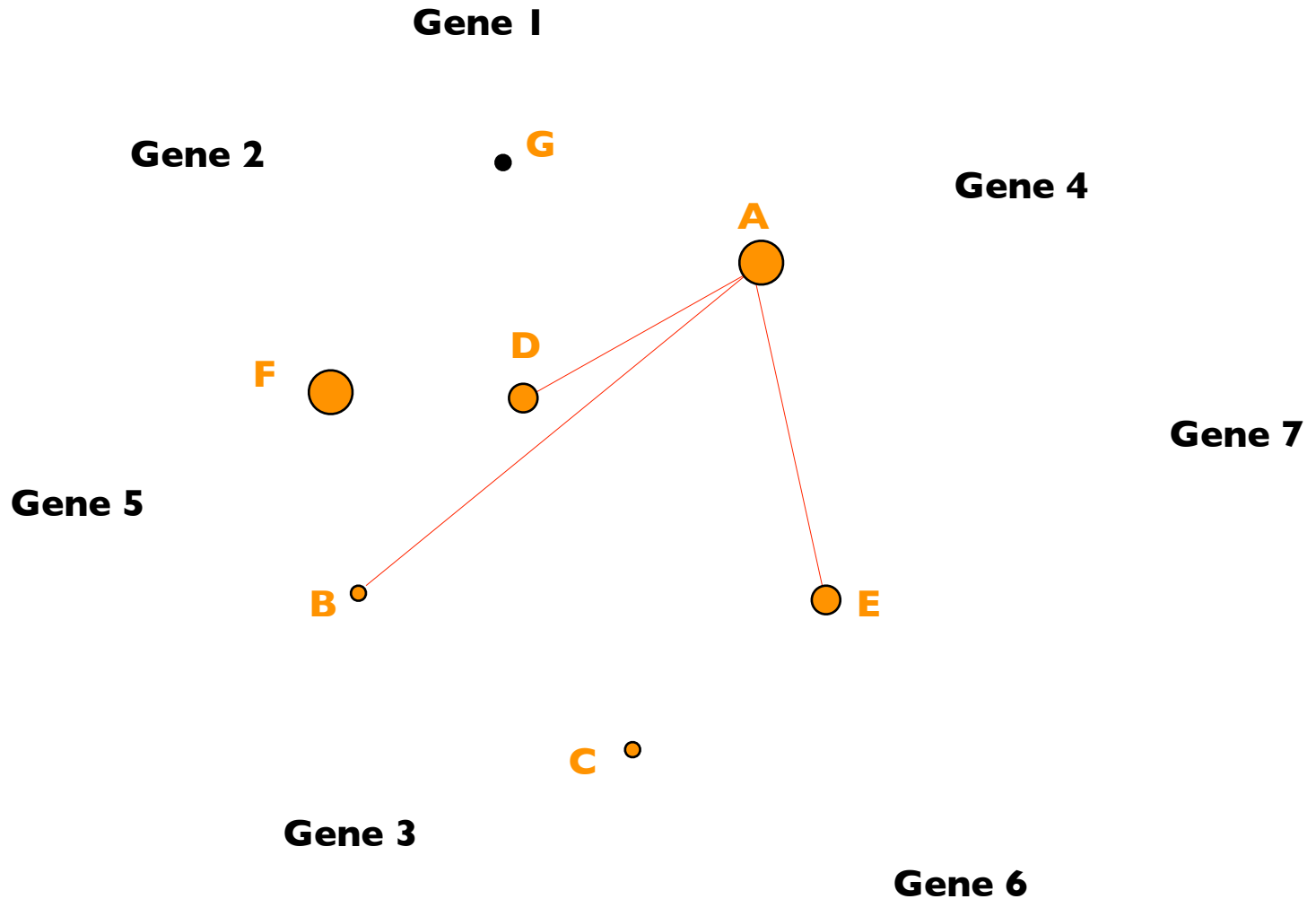
多次元データ（多数の遺伝子もしくは多数のサンプル）に含まれる特徴を

- ・ 大きなものから抽出して**新たな軸**を作り
- ・ **情報量の大きな低次元でデータを可視化する**

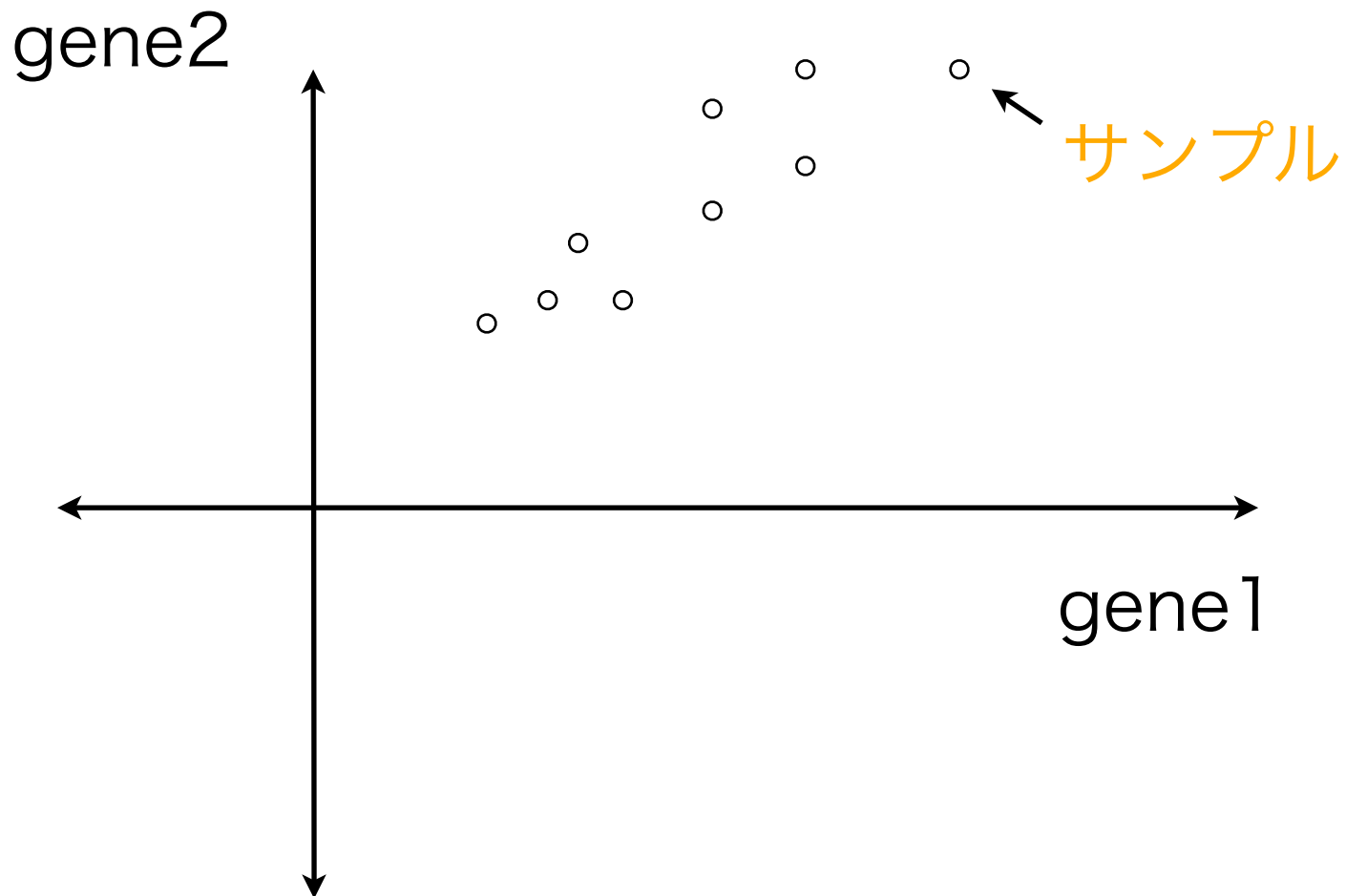
→ **人間が**新たな解釈を与える

階層クラスタリング:

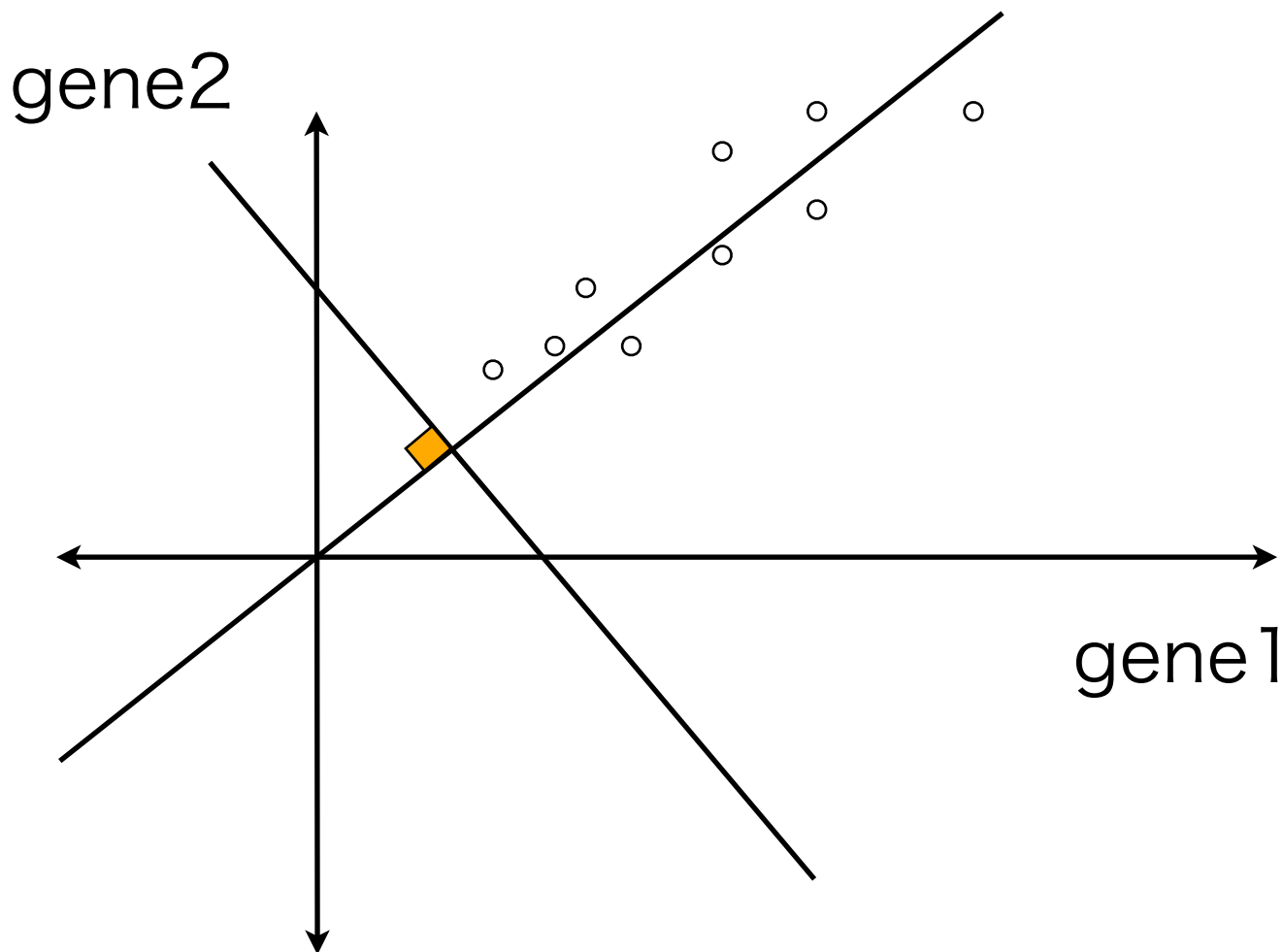
プロファイル間の類似性は空間での**1つの距離**によって決まる



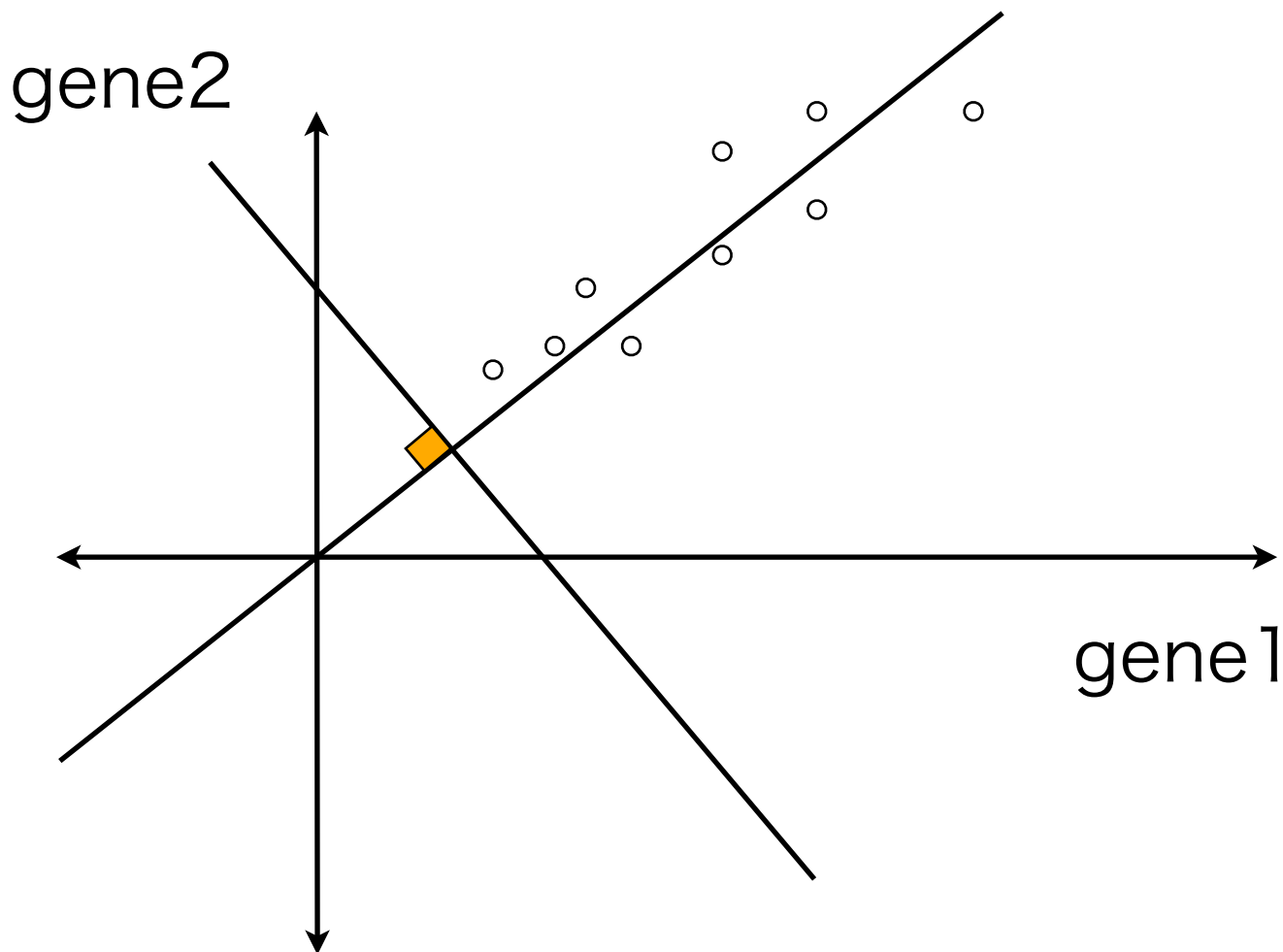
PCAは何をするのか？



PCAは何をするのか？



PCAは何をするのか？



PCAの概略(2次元)

1. 各サンプル (1.. n) の観察値(x_n, y_n)を

$$\begin{aligned} u_n &= a_1 x_n + b_1 y_n \\ v_n &= a_2 x_n + b_2 y_n \end{aligned}$$

とおく

足し算→線形

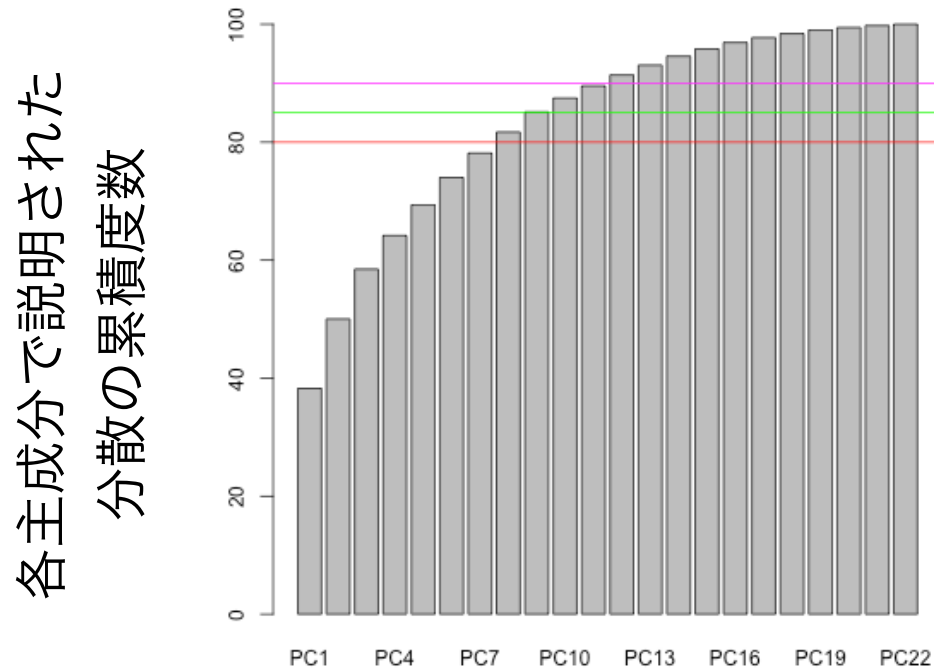
2. $a^2 + b^2 = 1$, u と v の相関係数0という制約の下でこれを解いて a_n, b_n を求める。

PCAで得られる重要な統計量

- 寄与率
- 因子負荷量
- 主成分得点

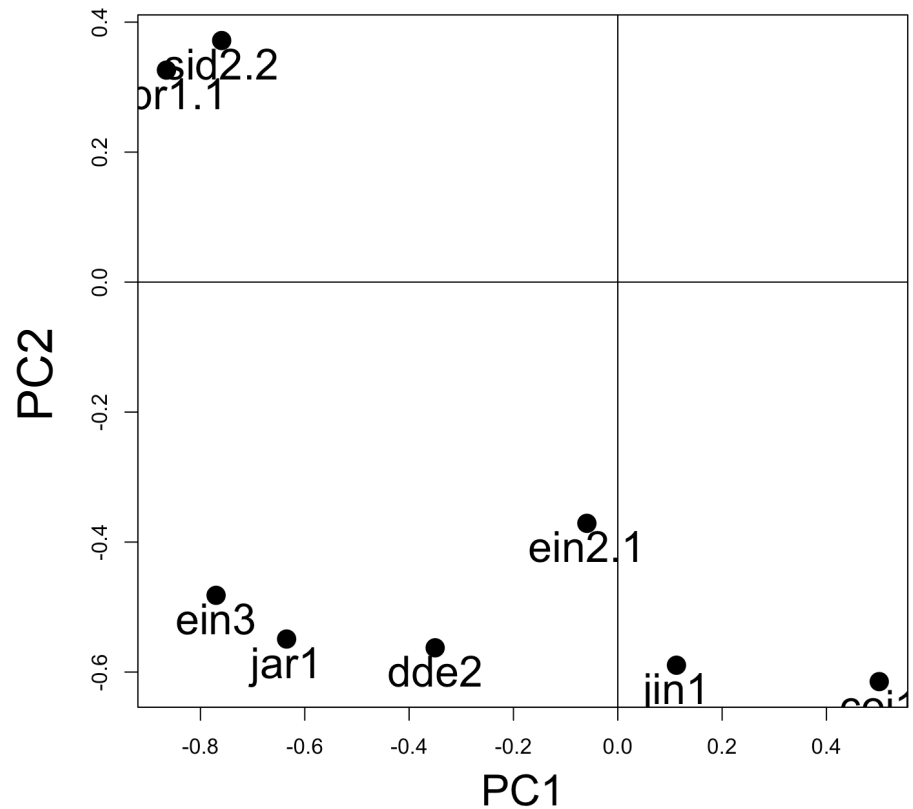
寄与率

- 各主成分が説明する分散の割合



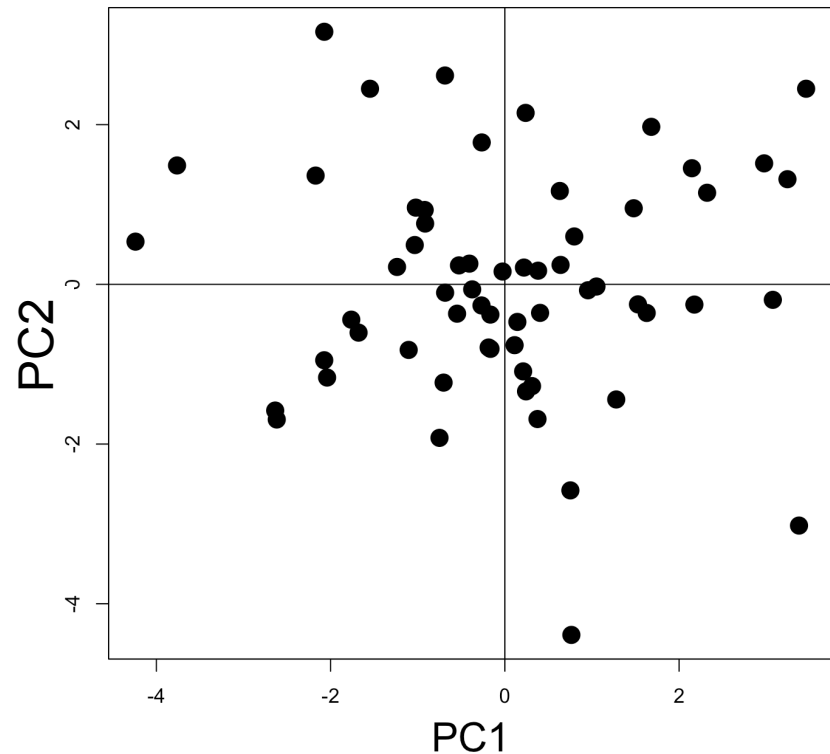
負荷量 loadings

- 得られた主成分と元データのパラメーターの相関
- 各パラメーターがもとのデータの情報をどれだけ有するか



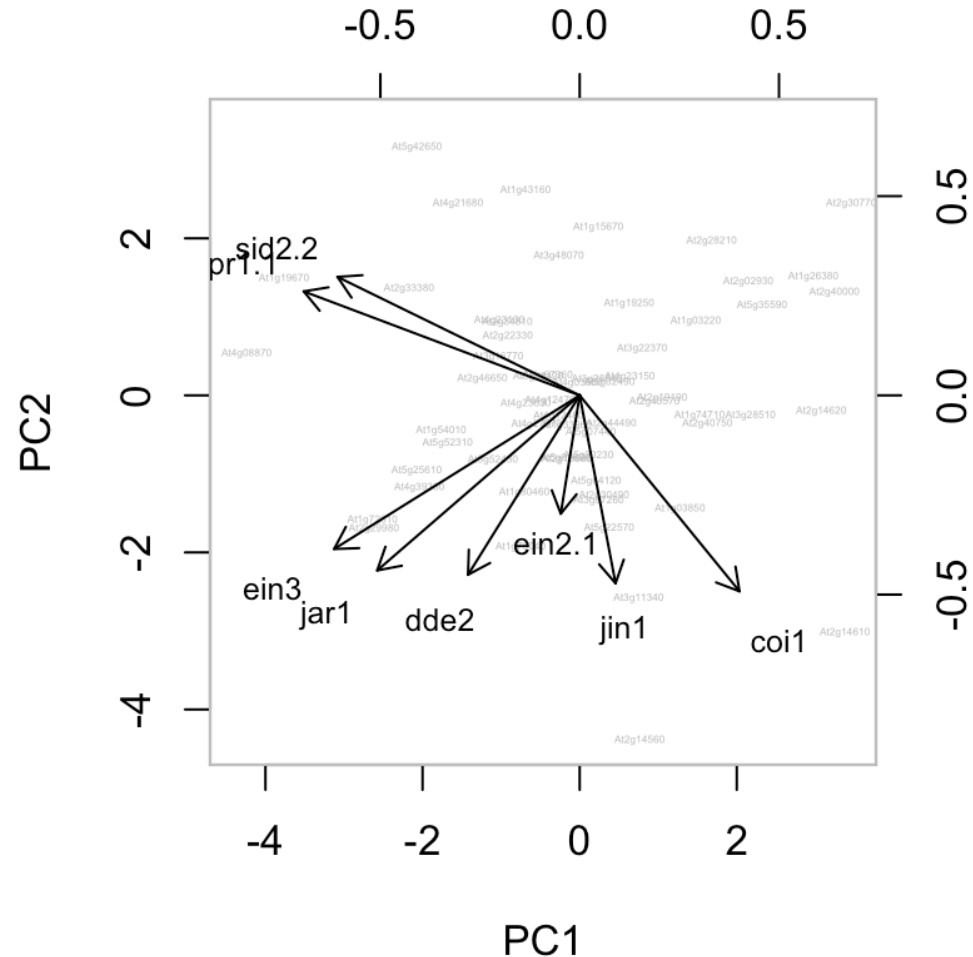
主成分得点 scores

- 各パラメーター
(遺伝子) の値
を各主成分につ
いて標準化した
統計量



標準化: 平均0, SD=1

因子負荷量と 主成分得点を 同時に可視化



ex601-3で着
目するPCを決めて可
視化してください

主成分分析(まとめ)

- 主成分分析はデータの分散を説明する新たな軸を計算する方法
 - 寄与率
 - 因子負荷量
 - 主成分得点

多次元尺度構成法

Multi-dimensional scaling(MDS),
Principal coordinate analysis

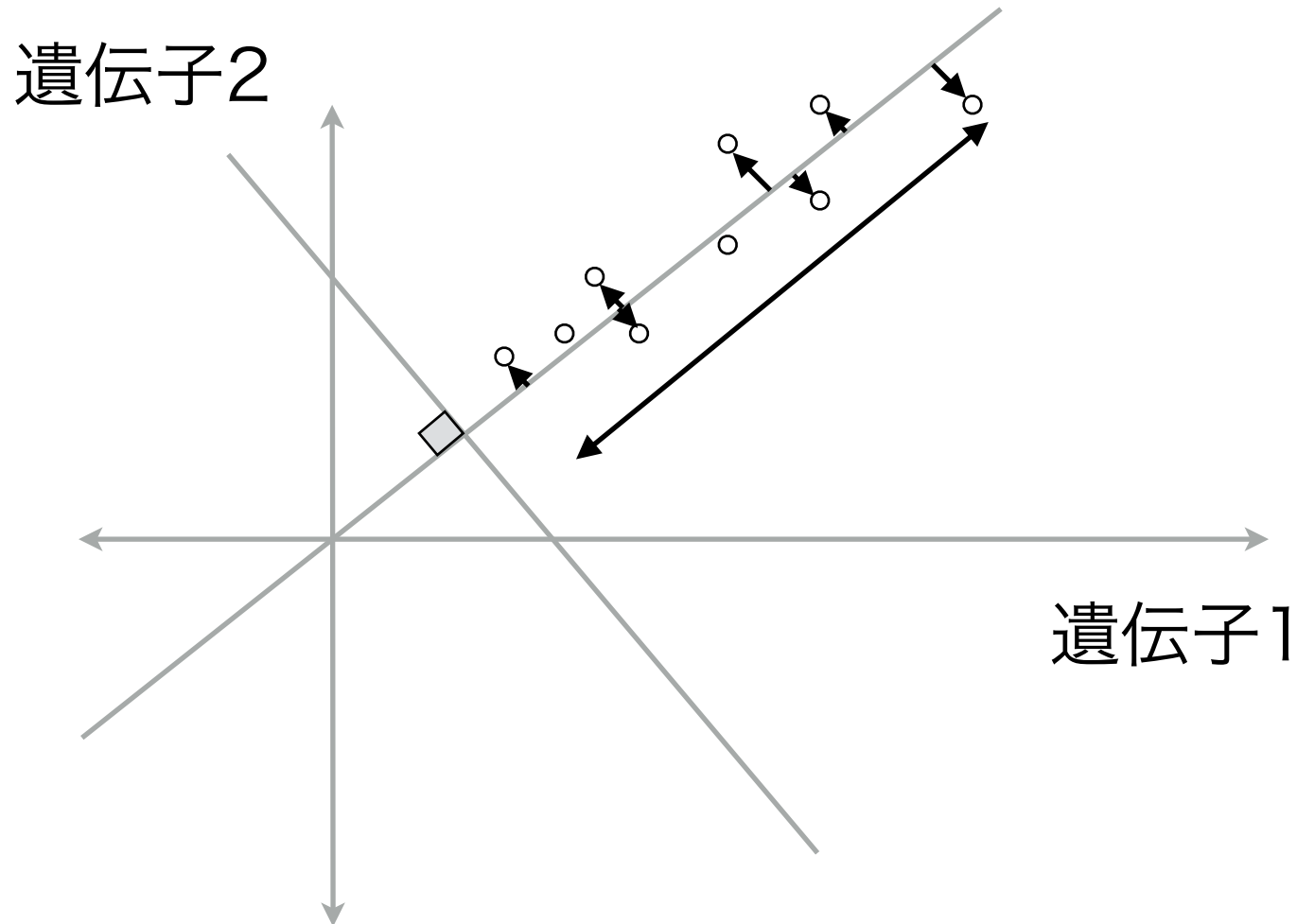
多次元尺度構成法とは？

モチベーション:

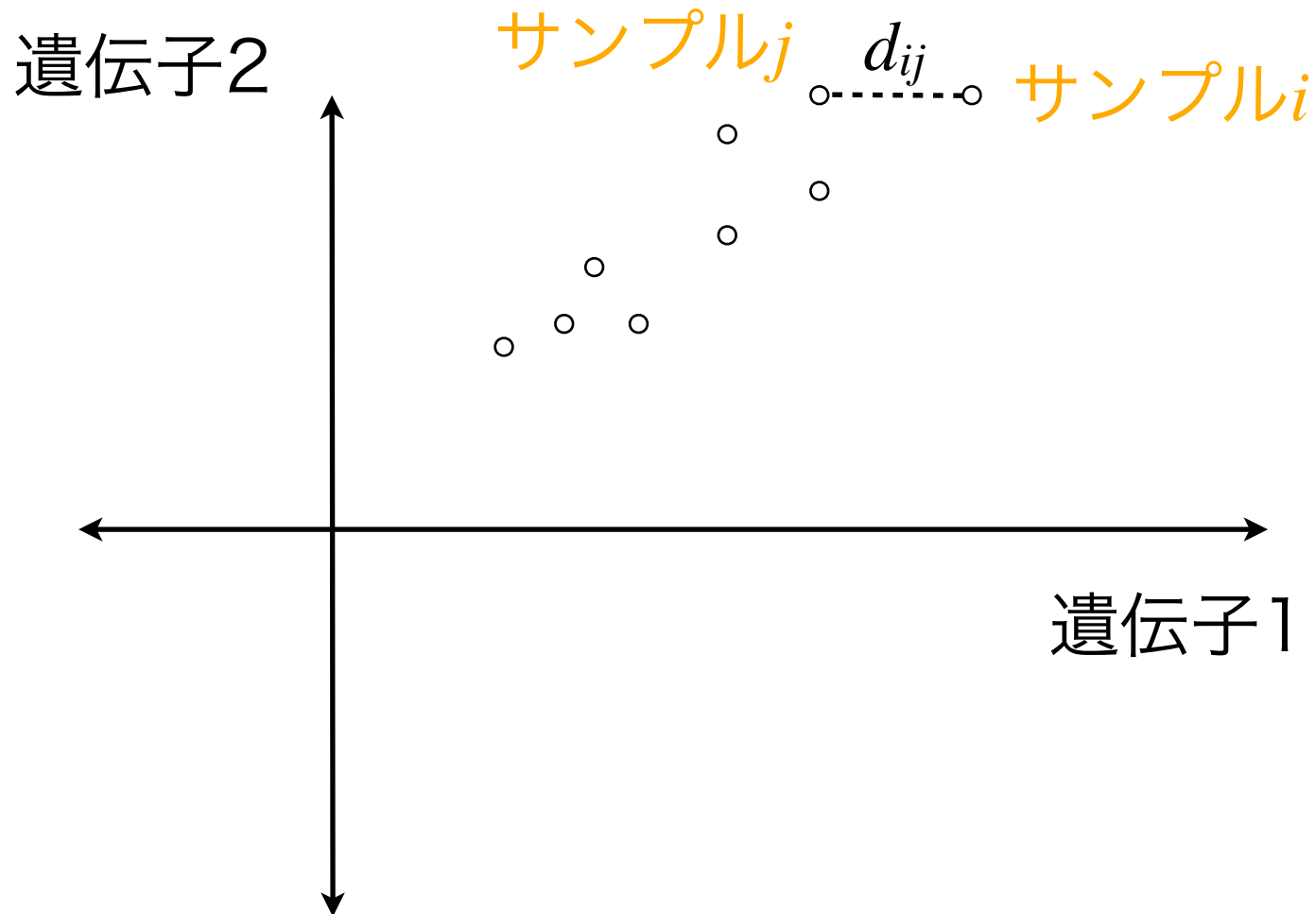
**多次元での各サンプル間の距離を保持して
低次元で表現する (多様体学習の1つ)**

⇒ 高次元の距離を低次元に圧縮するため
軸に意味がない

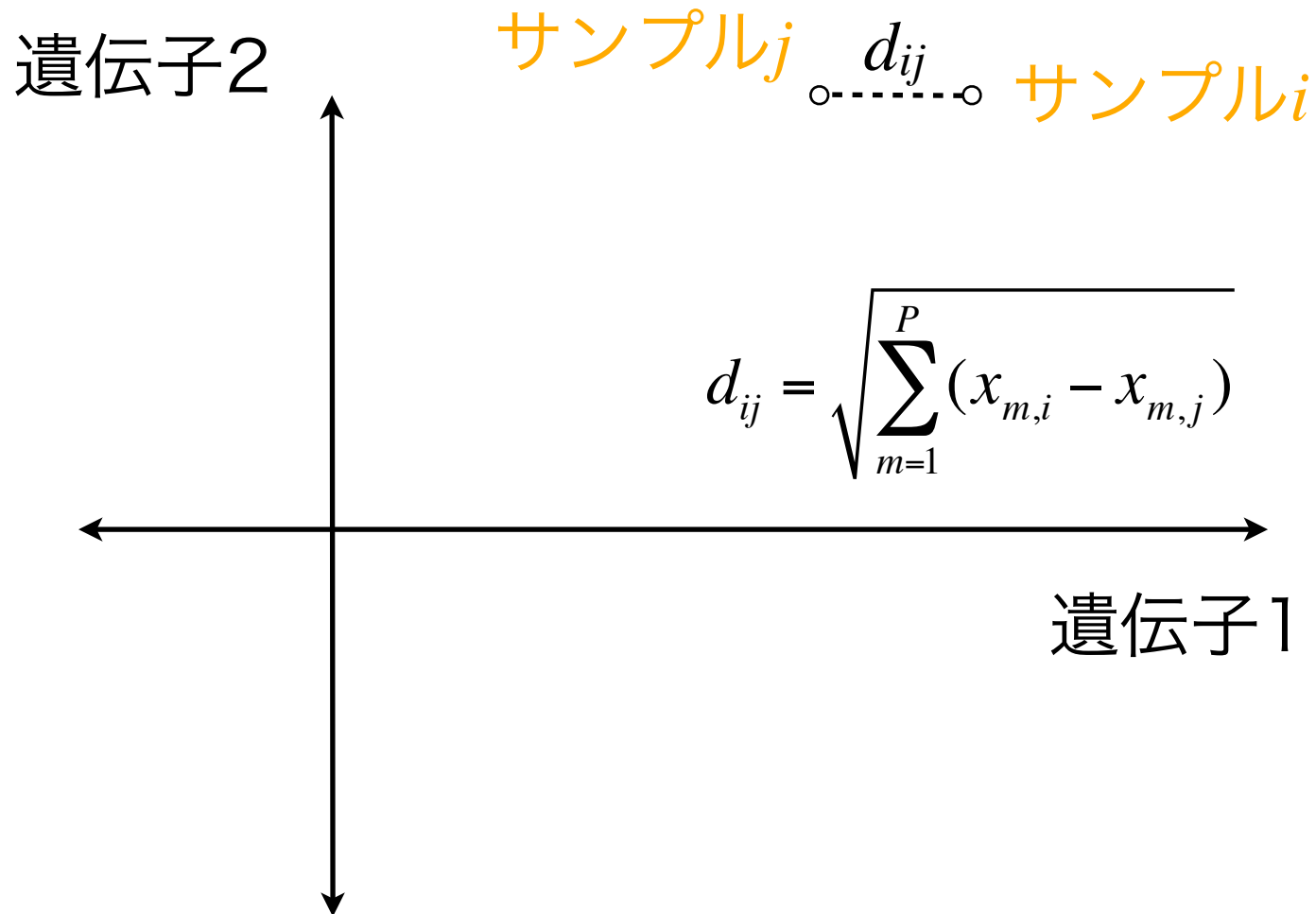
PCAが考慮する距離



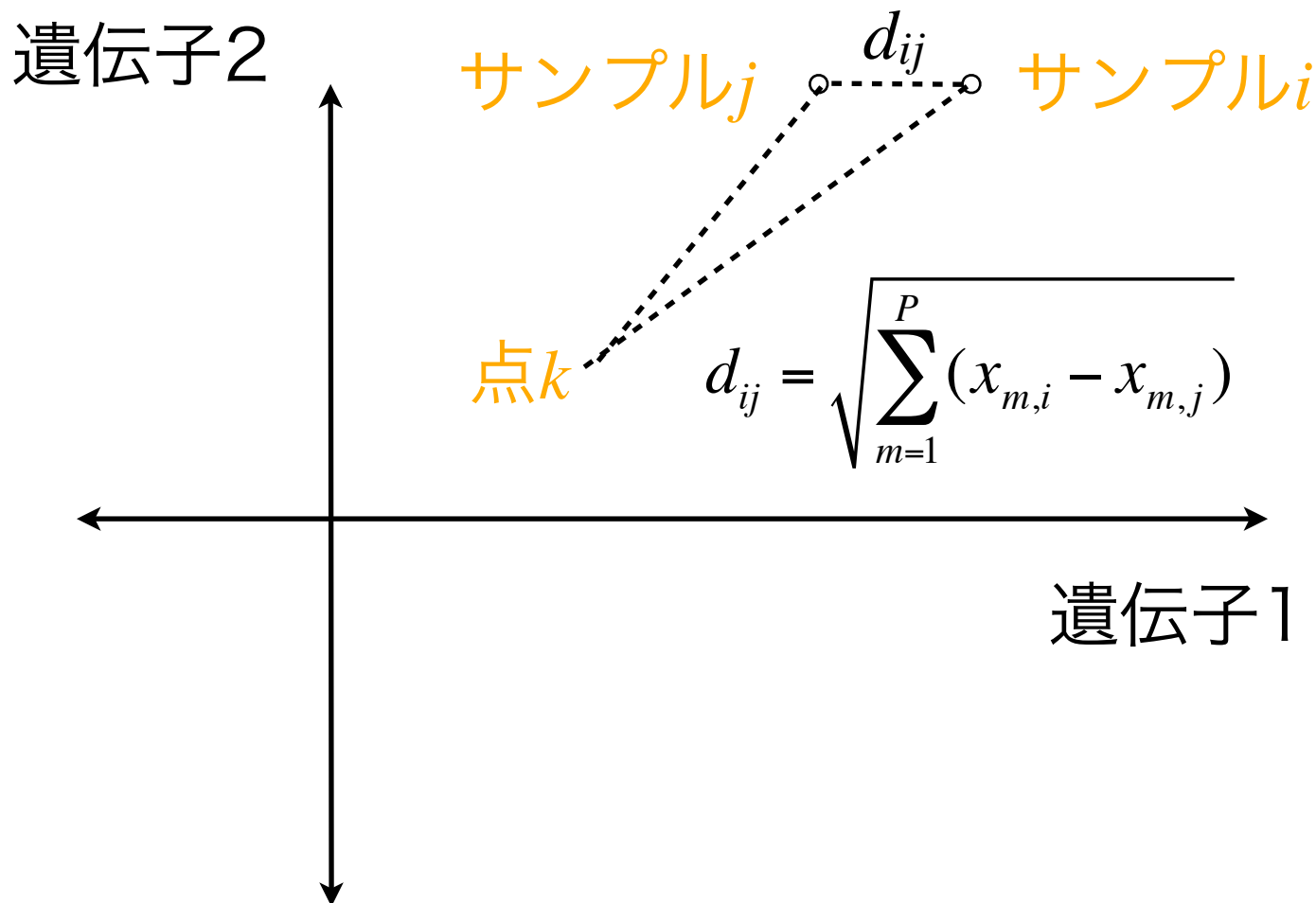
MDSが考慮する距離



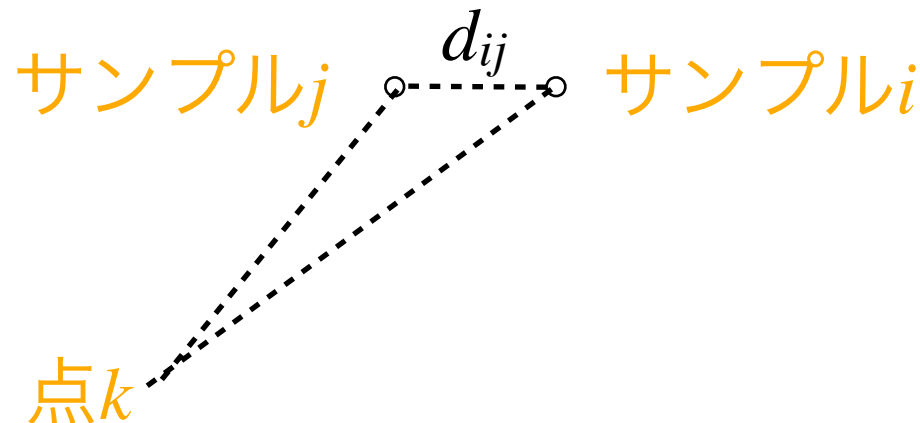
サンプル間の距離をまず計算する



この定理はサンプル*i,j*に対し、どこを原点
点（点*k*）としても成り立つ

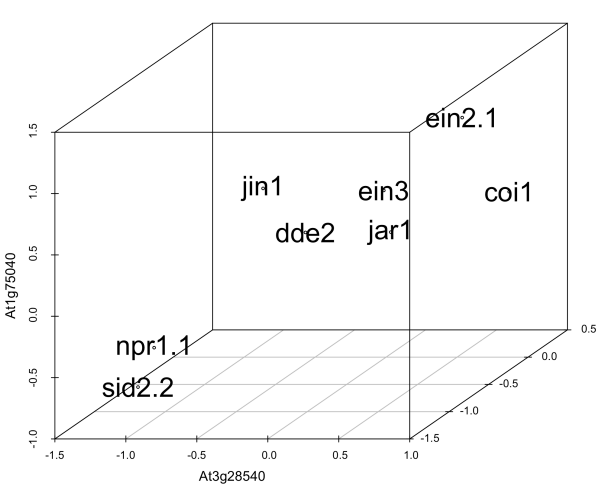


この定理はサンプル*i,j*に対し、どこを原点（点*k*）としても成り立つ

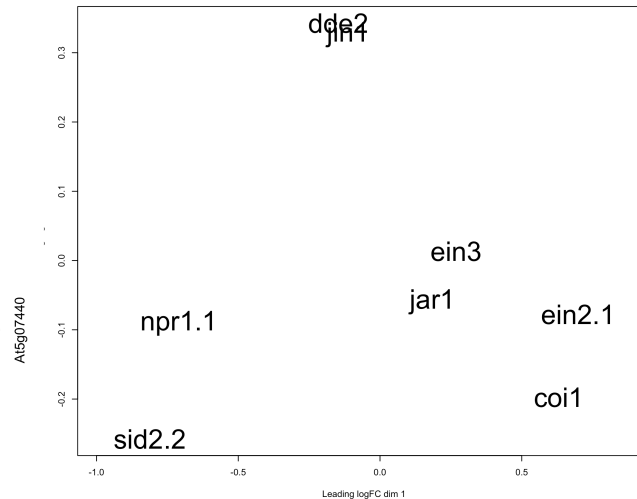


$$d_{ij}^2 = d_{ik}^2 + d_{jk}^2 - 2d_{ik}d_{jk}\cos\theta$$

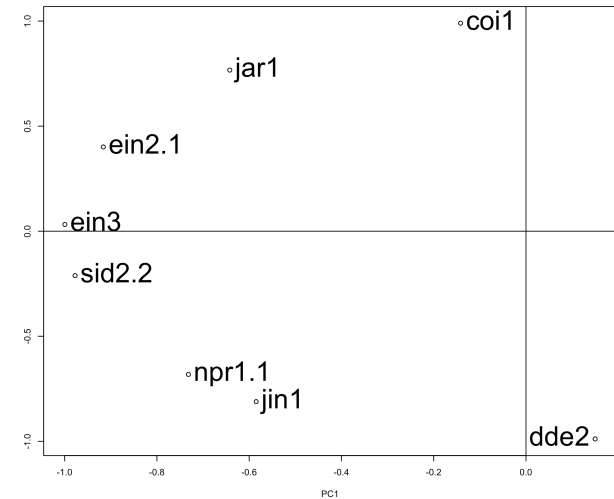
MDSとPCAの違い



Raw data (3 genes)



MDS (2D)



PCA (2 PCs)

多次元尺度構成法とは？

モチベーション:

**多次元での各サンプル間の距離を保持して
低次元で表現する**

⇔ 高次元の距離を低次元に圧縮するため
軸に意味がない

PCA/MDSのまとめ

データがもつ類似性を低次元で表現し、評価・可視化する

	PCA	MDS
軸 に意味がある	Yes	No
データ全体におけるサンプルの総体的な位置関係を保持するか？		
データが線形である場合	Yes	Yes/No
データが非線形である場合	No	Yes/No データが非線形であるときに局所的な類似関係をうまく表現しうる

- **重心の置き方に違い:** 入力データをどのように前処理するか

多様体学習: 非線形データの多変量解析

モチベーション:

非線形のデータ構造を低次元に圧縮して表現する

- Locally linear embedding
 - Isomap
 - t-SNE
 - UMAP
- } NGS解析にあまり使われないので
割愛するが、IsomapはMDSの
延長として勉強するのに有用

注意点: 局所 (サンプル間) の距離・関係を重視し、
全体の距離は犠牲にしている

t-SNE, UMAP

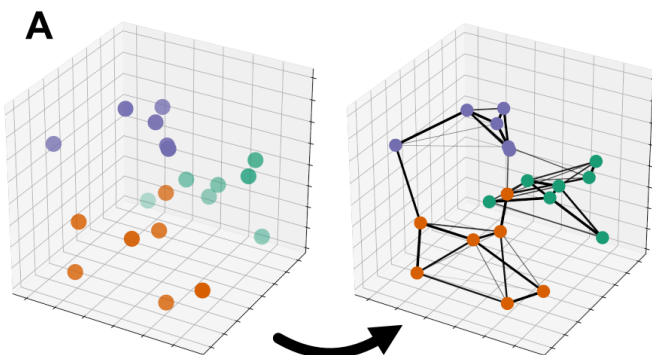
モチベーション：

非線形のデータ構造を低次元に圧縮して表現する

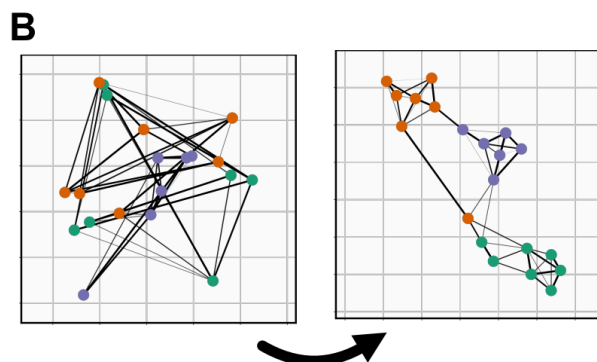
- 多様な状態のサンプル間の関係性、クラスターを特定する
- single cell omicsデータ解析におけるMDSの立ち位置

t-SNE, UMAPアルゴリズム概略

1. サンプル間の距離を計算
2. 高次元でのサンプル間距離が低次元でも同様になるよう調整 (embedding, 埋め込み)
 1. t-SNE: t-分布を使って距離を調整
 2. UMAP: グラフ解析を利用



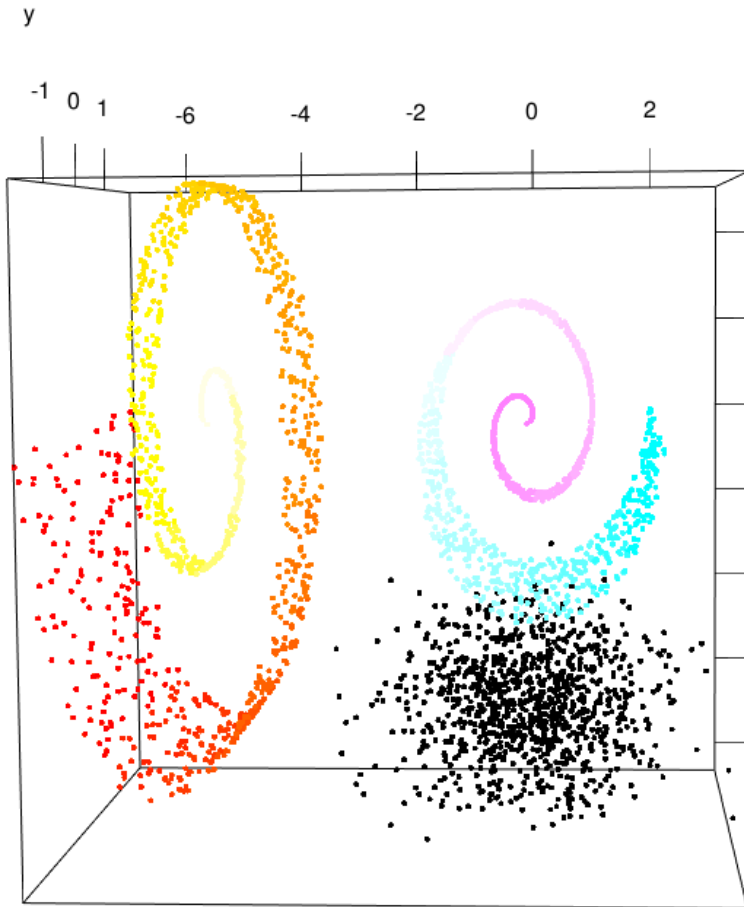
Step 1: Compute a graphical representation of the dataset



Step 2 (non-parametric): Learn an embedding that preserves the structure of the graph

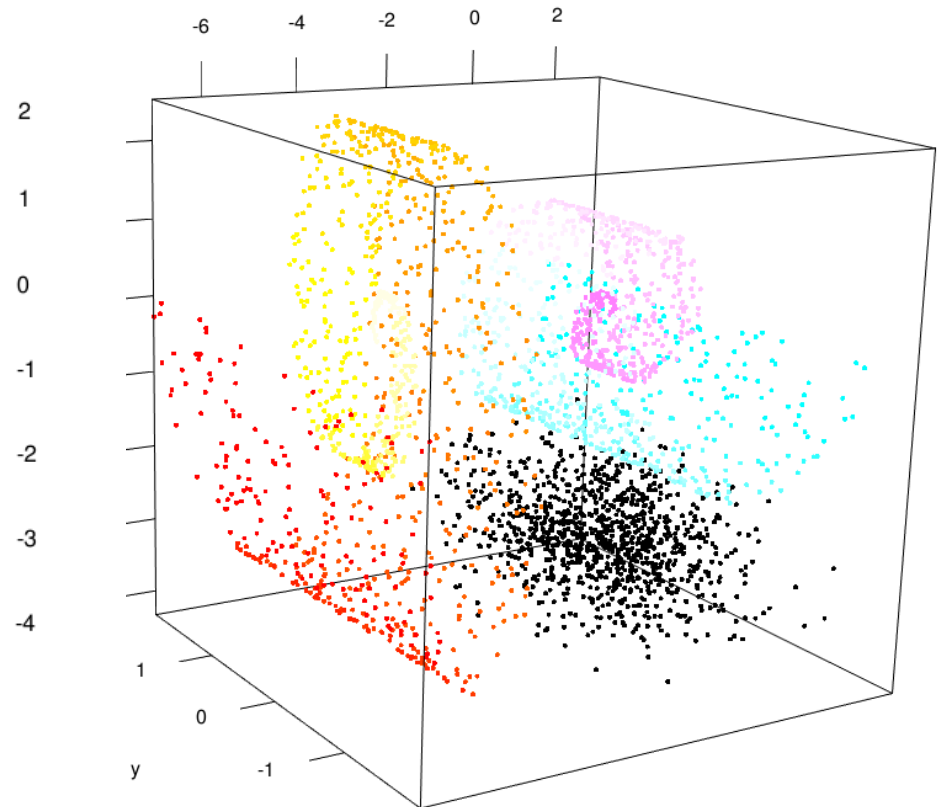
Sainberg et al. (2009)
[arXiv:2009.12981](https://arxiv.org/abs/2009.12981)

人工非線形データの多変量解析



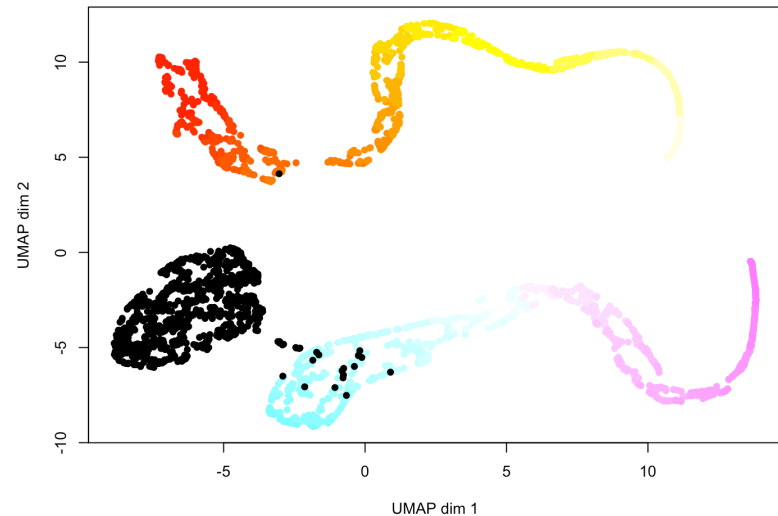
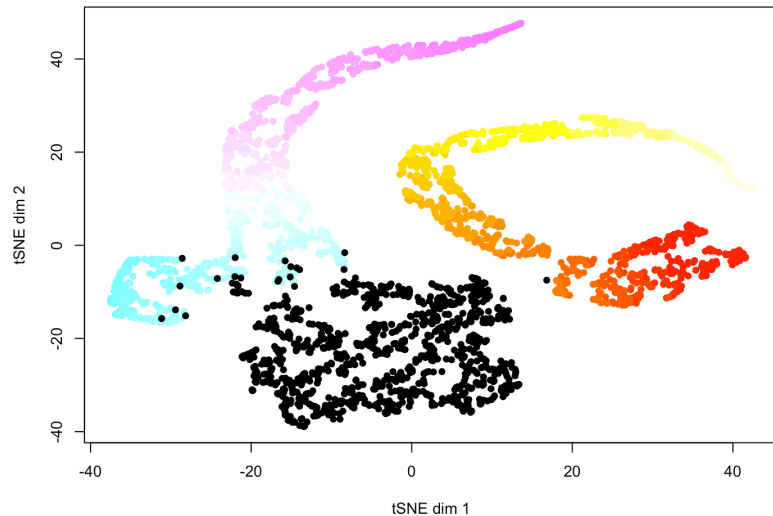
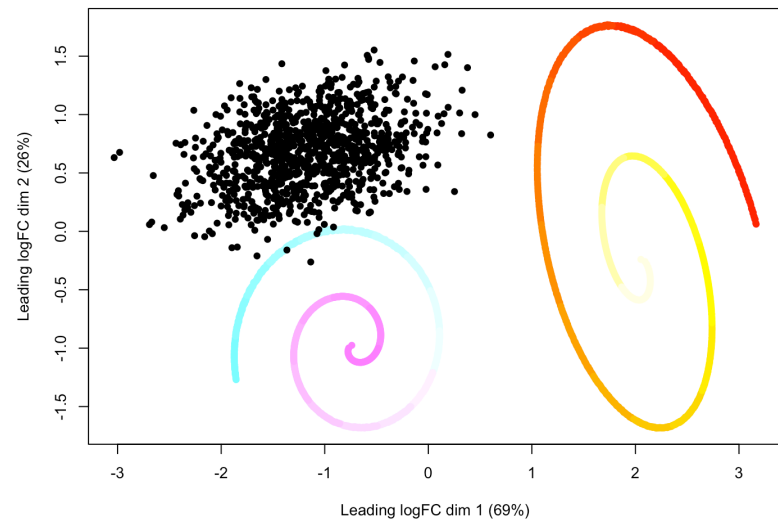
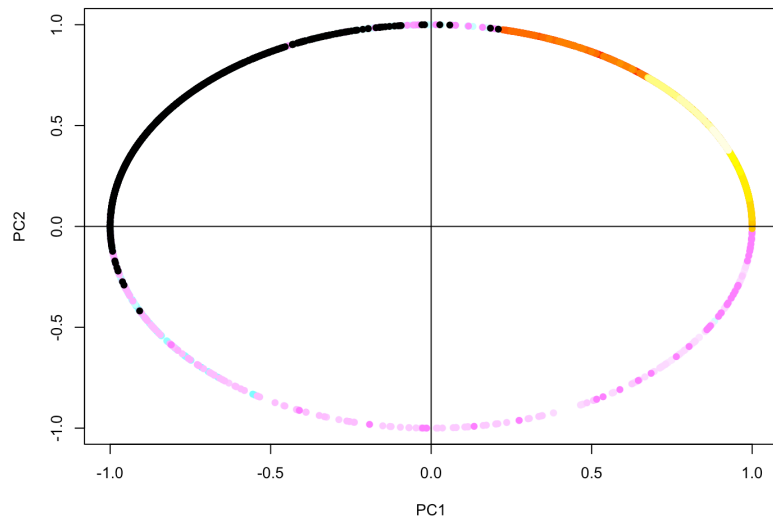
仮定：3遺伝子（次元）データ

- 時間に変化する細胞集団2つ
- ランダムな状態の細胞集団1つ

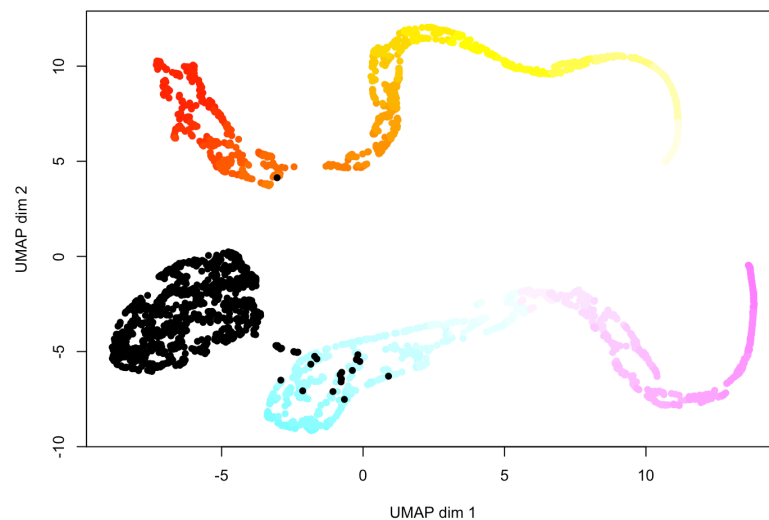
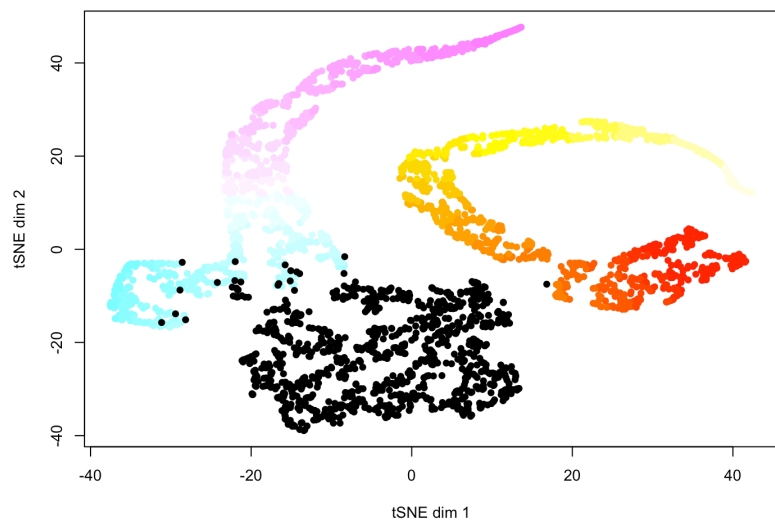
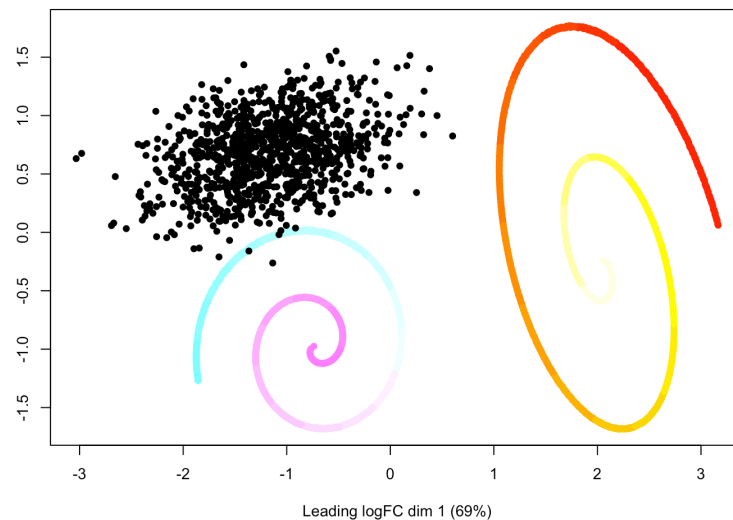
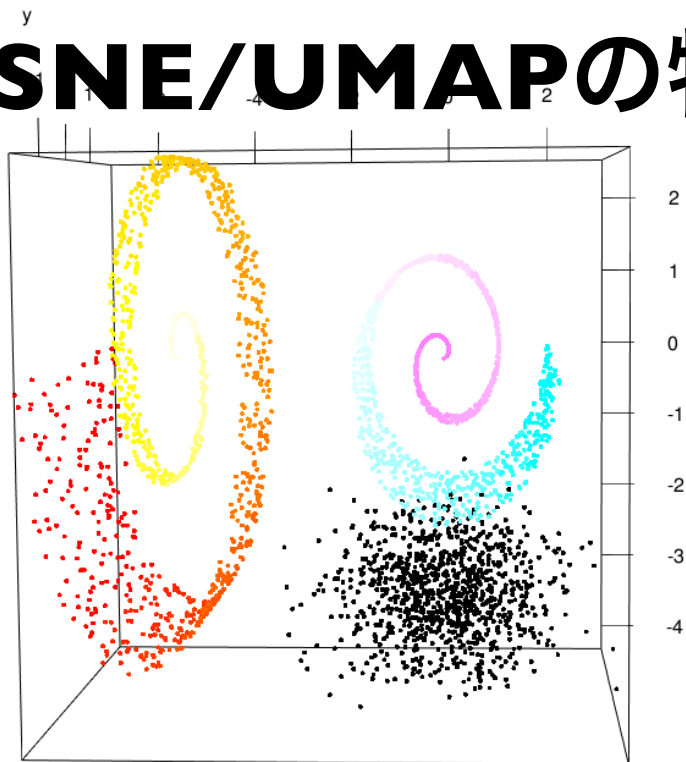


- 分布：2次元ではスイスロール
- 残りの1次元: 初めは多様性が低い

PCA, MDS, t-SNE, UMAP

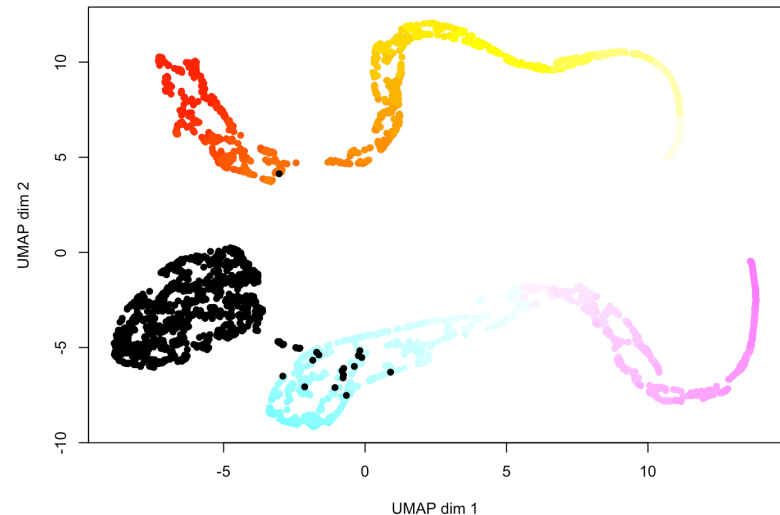
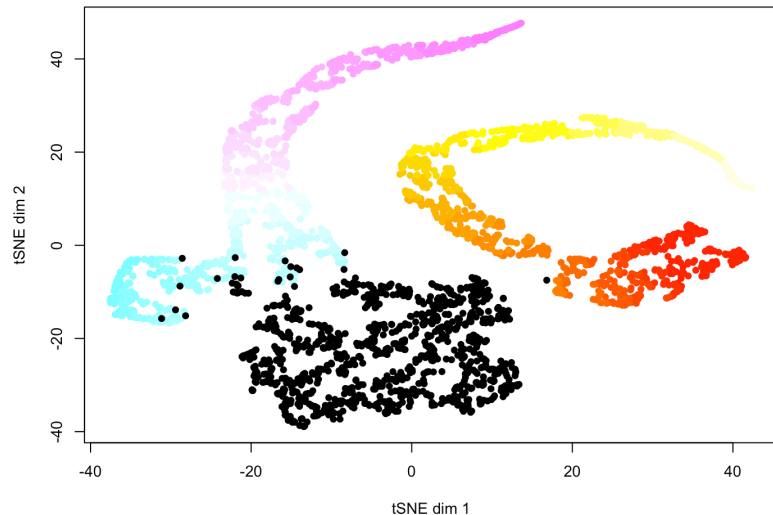
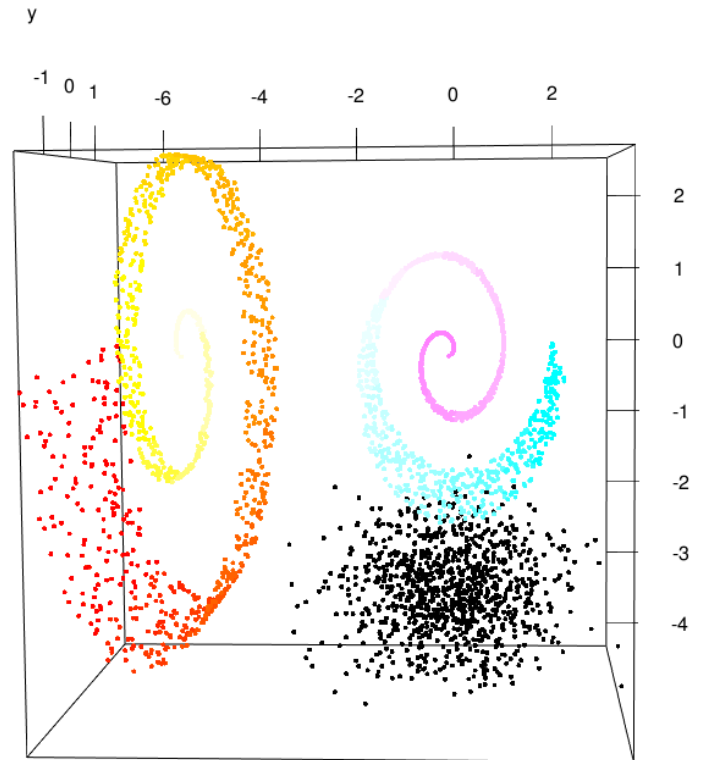


t-SNE/UMAPの特徴



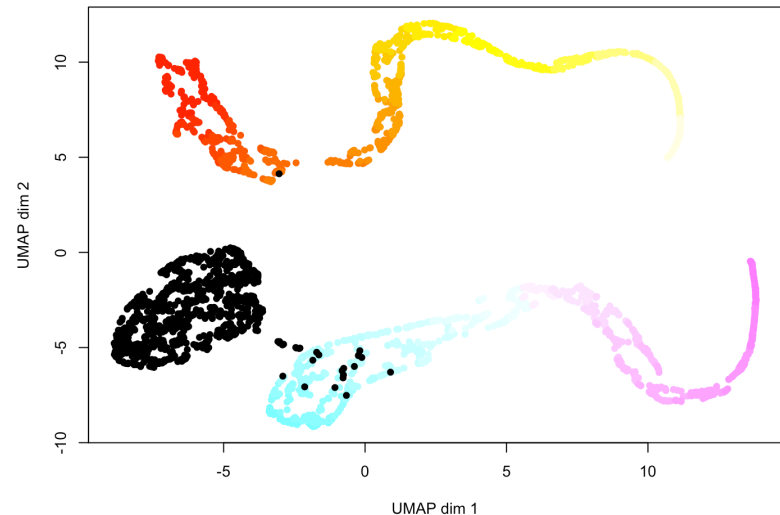
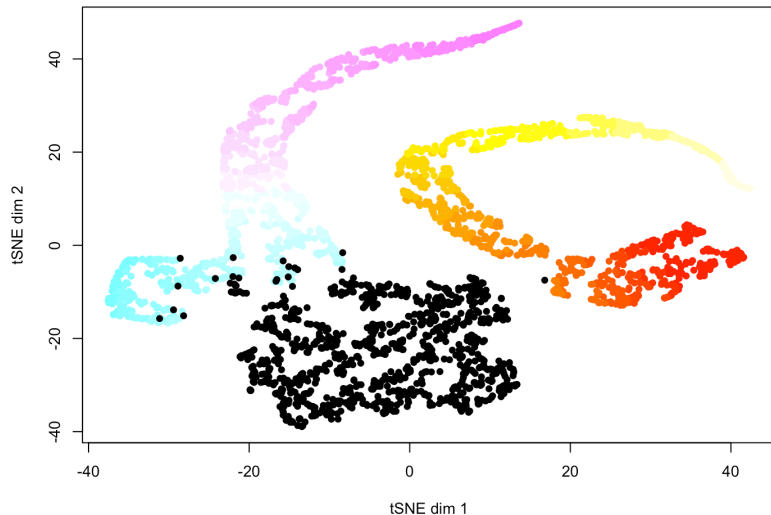
t-SNE/UMAPの特徴

- 近傍のサンプルのクラスター化
- 実空間での距離・サンプル分布は反映しない
 - クラスターサイズは関係ない



t-SNE/UMAPの違い

- クラスターの分離: UMAP > t-SNE
- 速度: UMAP < t-SNE
- t-SNEは結果が必ず同じ結果になるとは限らない（収束していない可能性がある）

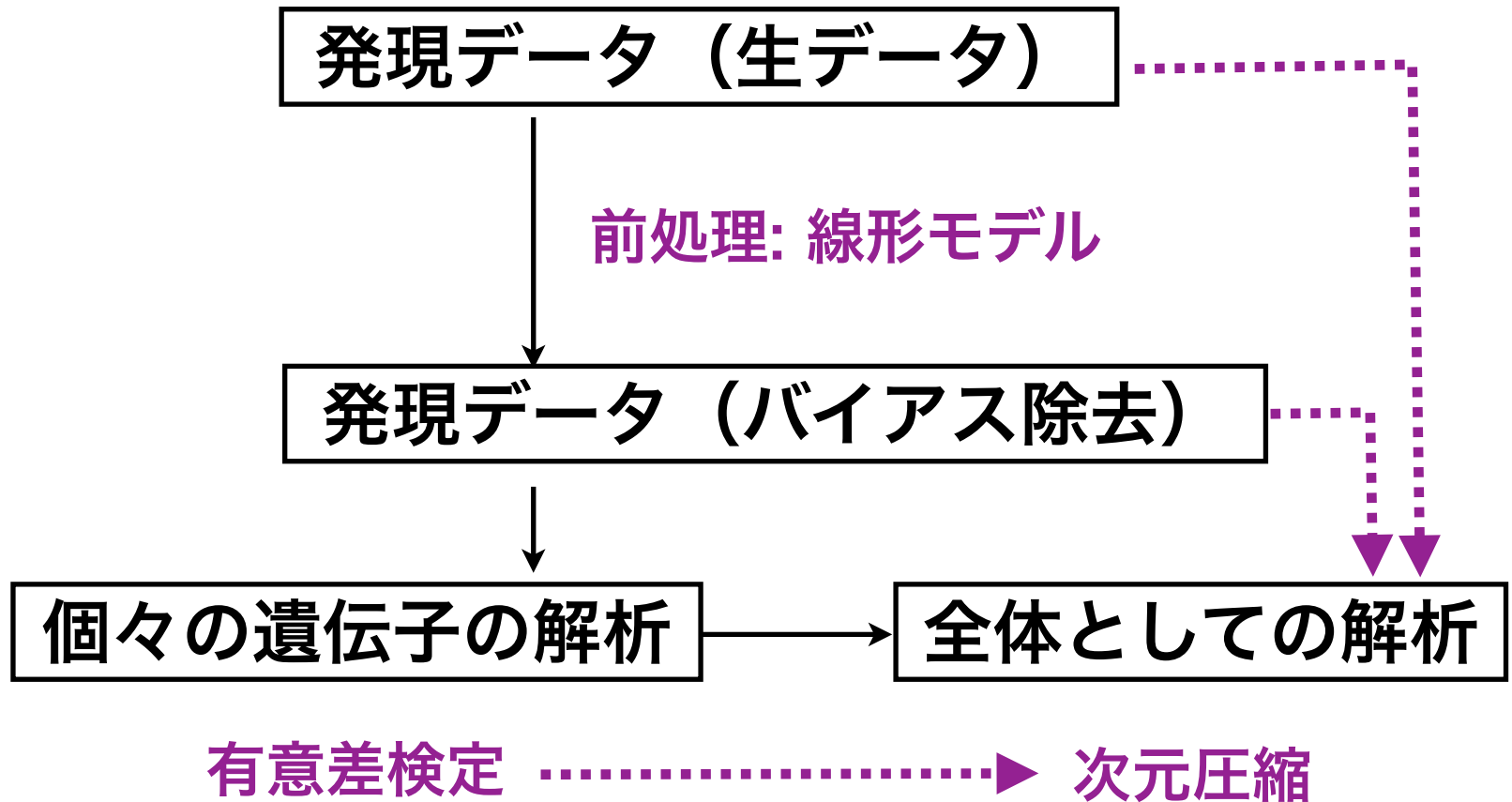


多様体学習: 非線形データの多変量解析

- t-SNE
- UMAP
 - : 非線形高次元データのクラスタリングが可能

注意点: 局所（サンプル間）の距離・関係を重視し、
全体の距離は犠牲にしていることを念頭に
入れて結果を解釈する必要性あり

多変量解析をもう一歩進めて: 入力データは何を使うか？



多変量解析をもう一步進めて:

人間の解釈をアシストするデータ取得を心がける

多変量解析の枠組み

多次元（例: 多パラメーター）を
より少ない指標を使って理解する



N個のサンプルをM個 ($M < N$)の
グループに分類する

→ 人間が新たな解釈を与える

コントロール、
指標サンプルは
含められるか？

今回の内容で扱わなかった重要項目

- **教師あり多変量解析**
 - k-means法
- **非線形多変量解析・次元圧縮の詳細**

連絡：コピーライト

コピーライトは佐藤にあります。
資料内容の使用については下記連絡先までご
連絡ください。

- satox@agr.hokudai.ac.jp