

多変量解析

(特徴空間分割・次元圧縮)

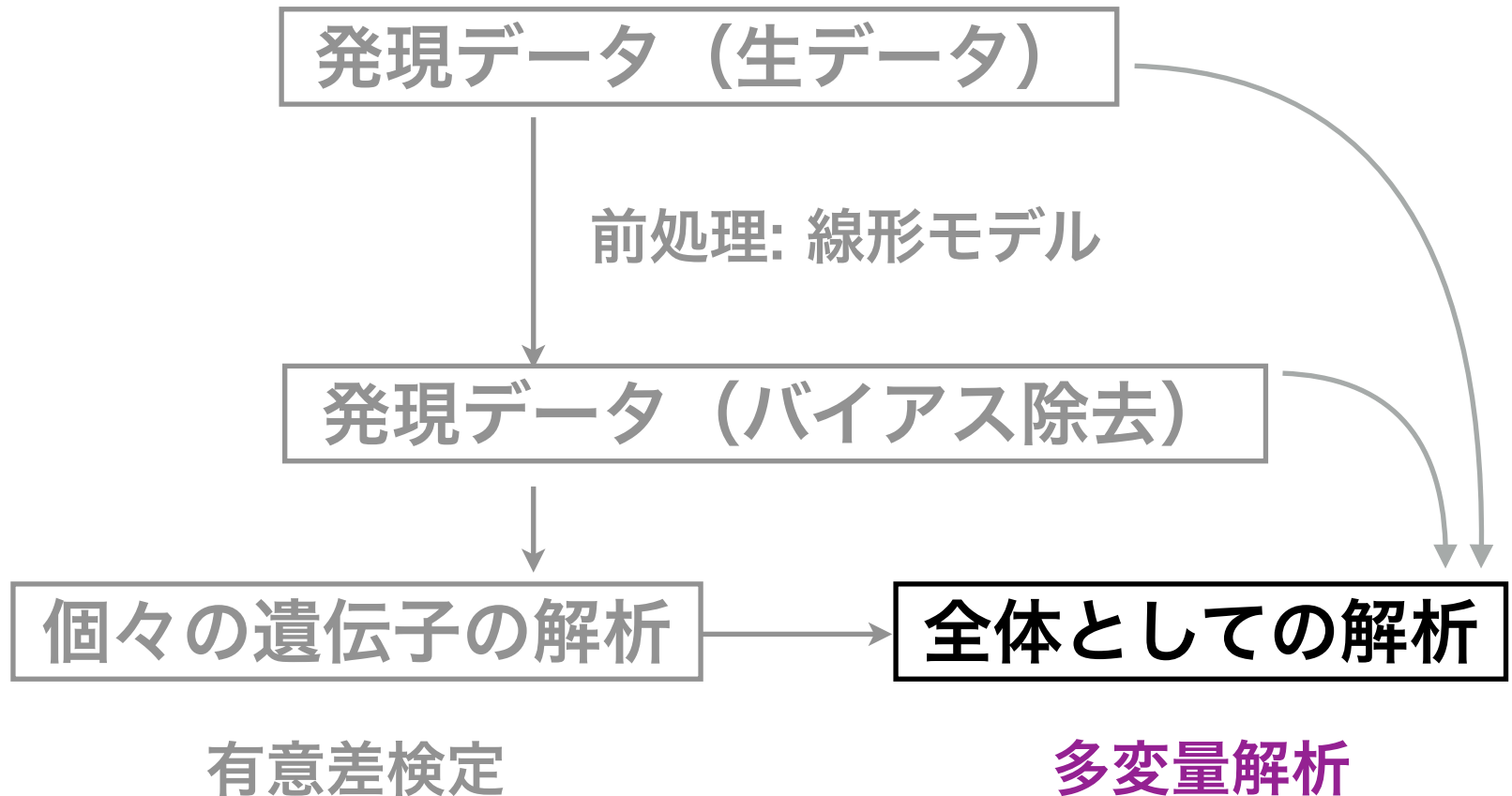
北海道大学大学院農学研究院

(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（例: 多パラメーター）を
より少ない指標を使って理解する

→ 高次元データを低次元で表現し

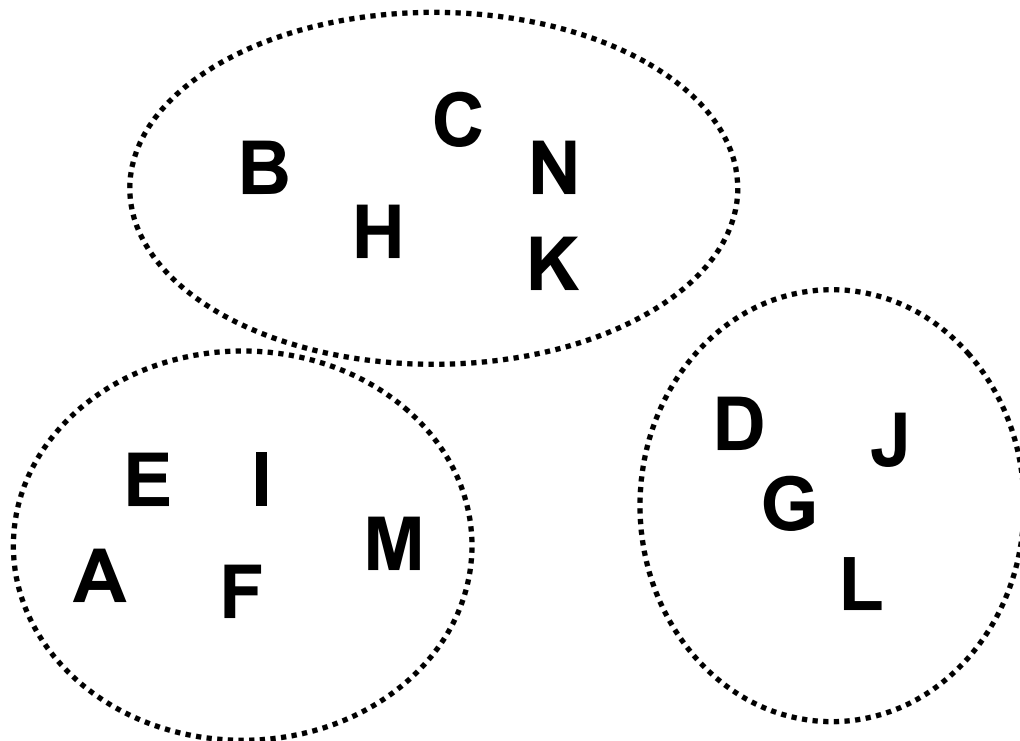
- 可視化する

- 特徴量を選択する

→ 人間が新たな解釈を与える

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

~/gitc/data/MS/Sato_A_thaliana-P_syringae_avrRpt2_6h_expRatio_small.txt

を眺めてみましょう

データの特徴は読み取れますか？

Rで可視化してみましょう：

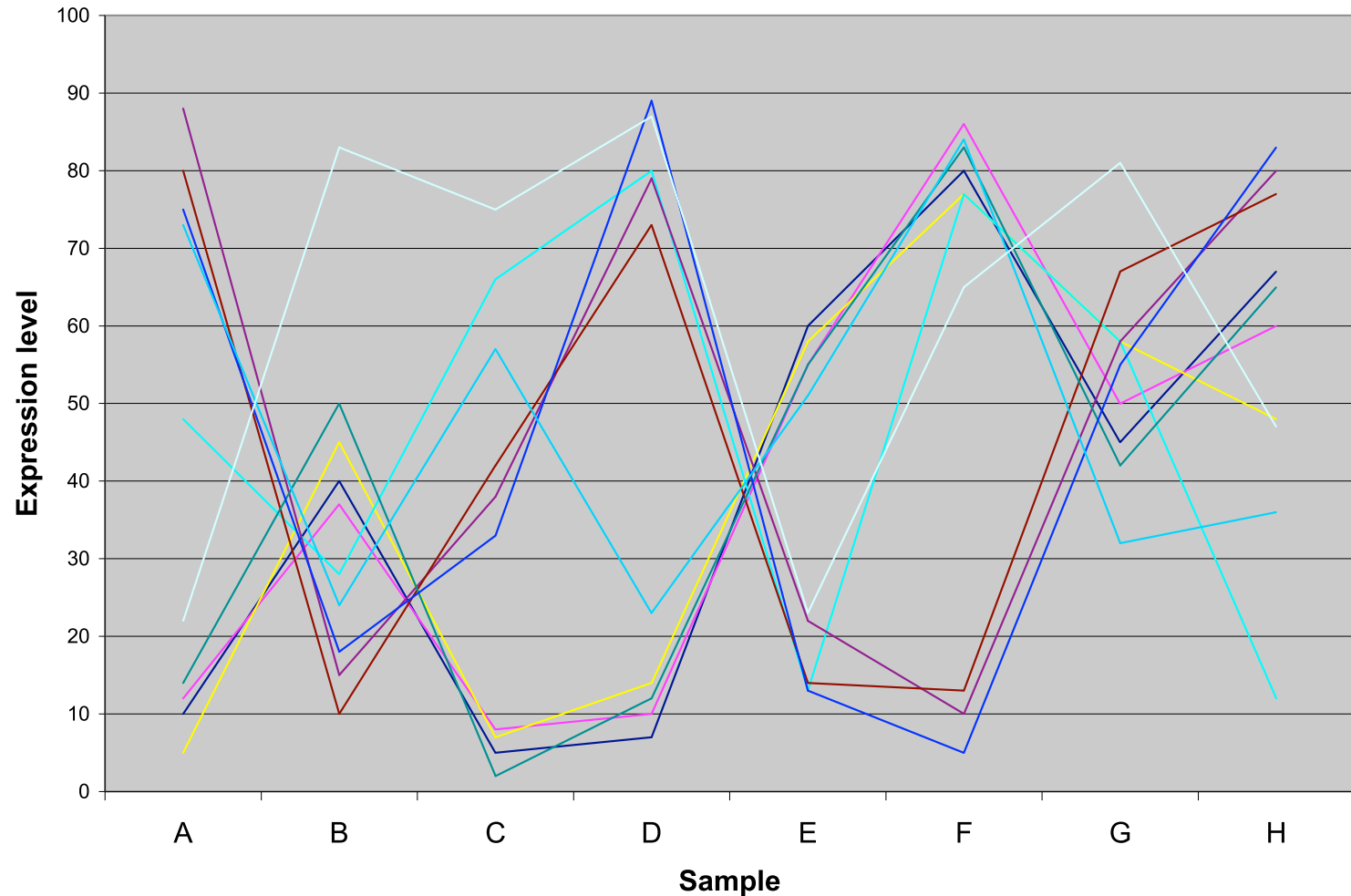
```
inputMatrix <- read.delim(  
  "~/gitc/data/MS/Sato_A_thaliana-P_syringae_avrRpt2_6h_expRatio_small.txt",  
  header=TRUE,  
  row.names=1  
)  
str(inputMatrix)      # データの構造を確認する  
image(t(inputMatrix)) # カラーコードで行列データをそのまま可視化  
  
heatmap(as.matrix(inputMatrix)) # 階層クラスタリングとヒートマップ
```

Rは簡単に「何か」を出力してくれる！

```
inputMatrix <- read.delim(  
  "~/gitc/data/MS/Sato_A_thaliana-P_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
```

統計の基礎知識とRへの正しい命令が必要

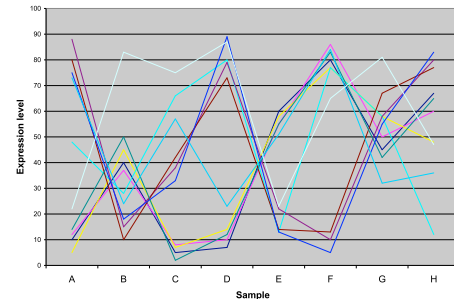
トランスクリプトームデータのある一部について可視化してみる



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

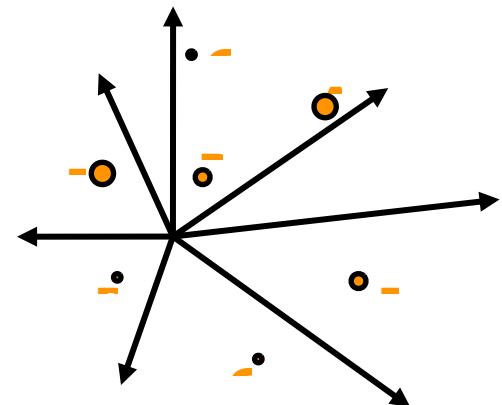


問題定義の変換

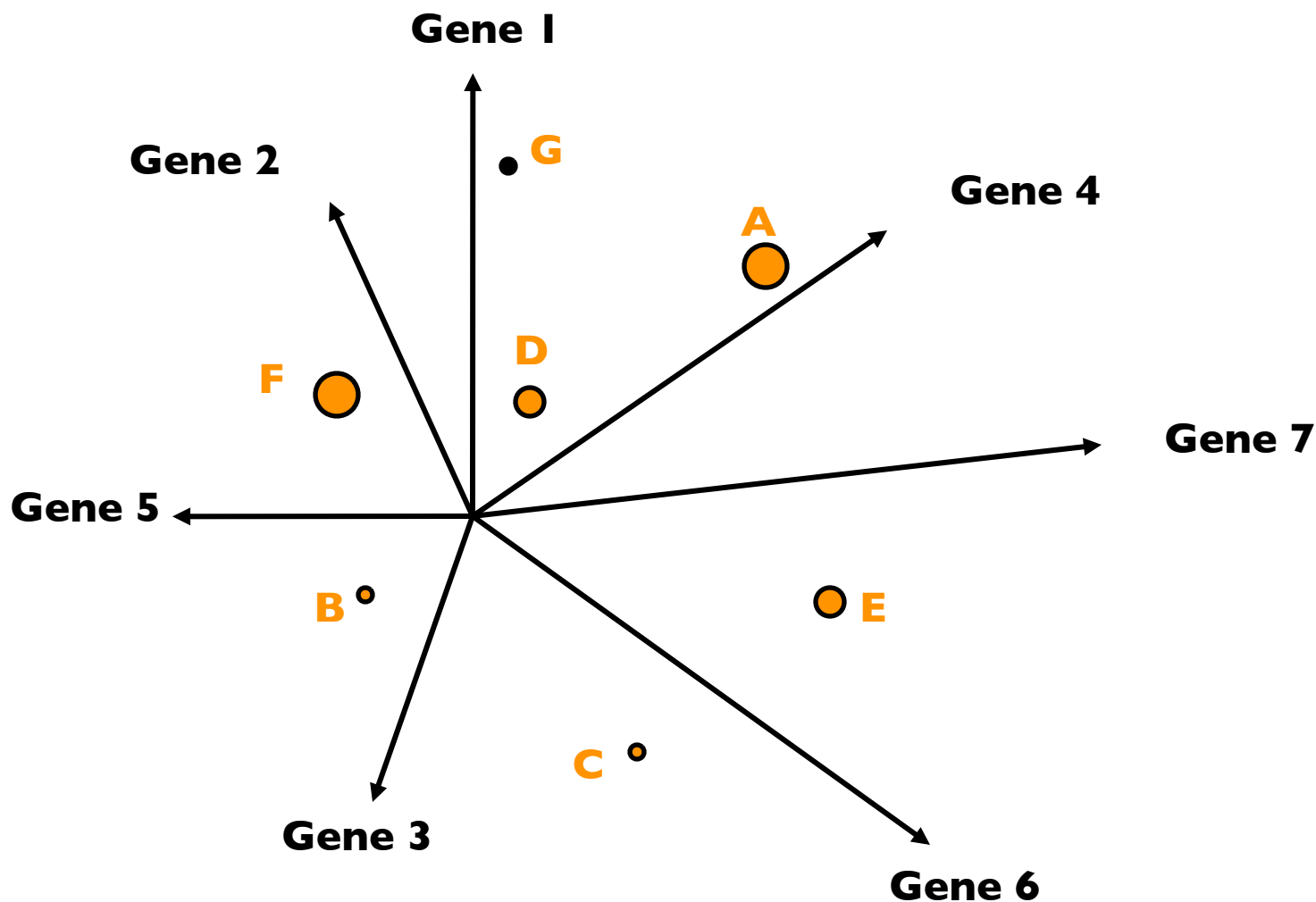
(生物学の問題を数学の問題に置き換える)

計算機

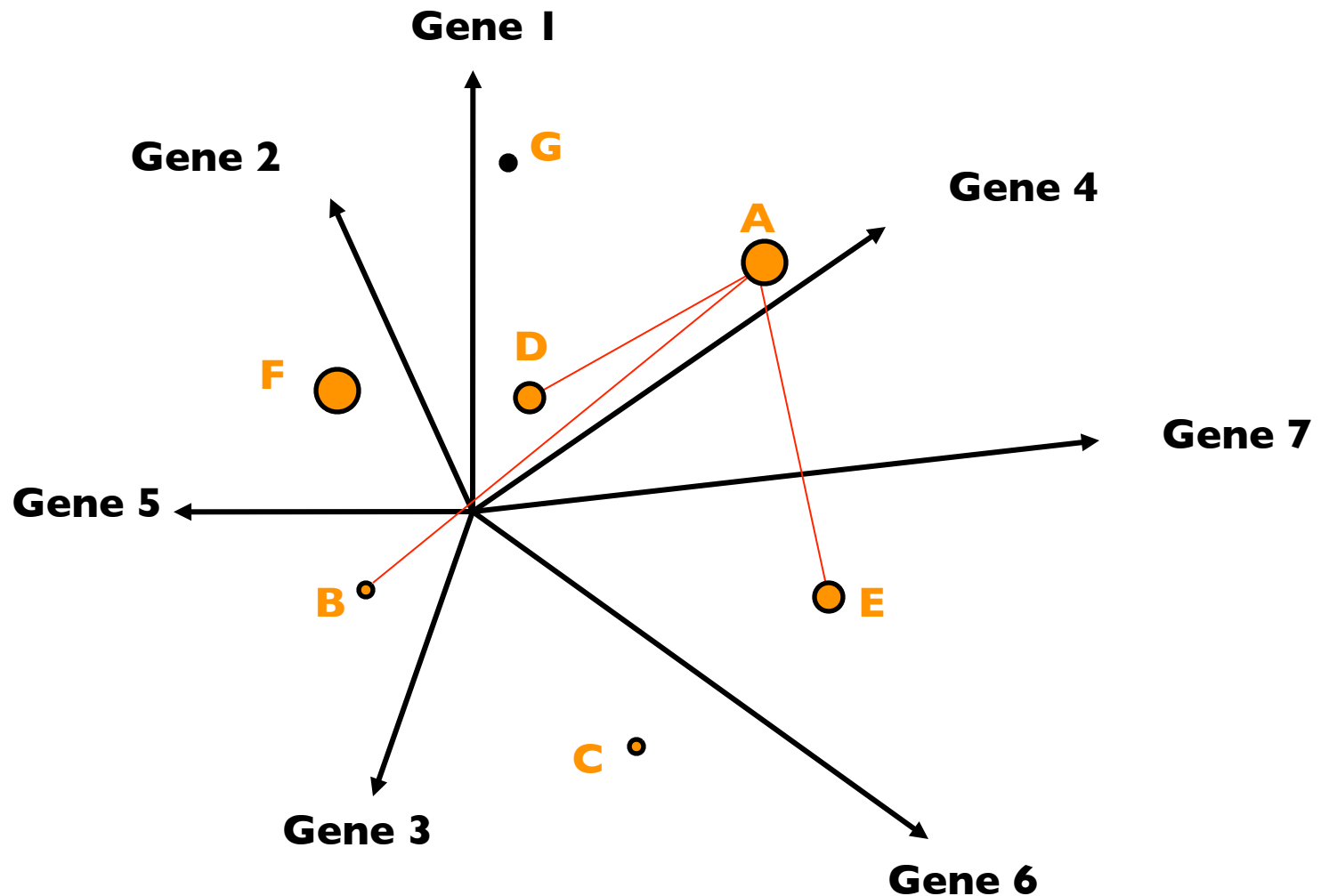
ベクトル等の
数学で扱える
特徴量を
用いた計算



7次元の遺伝子発現データセット

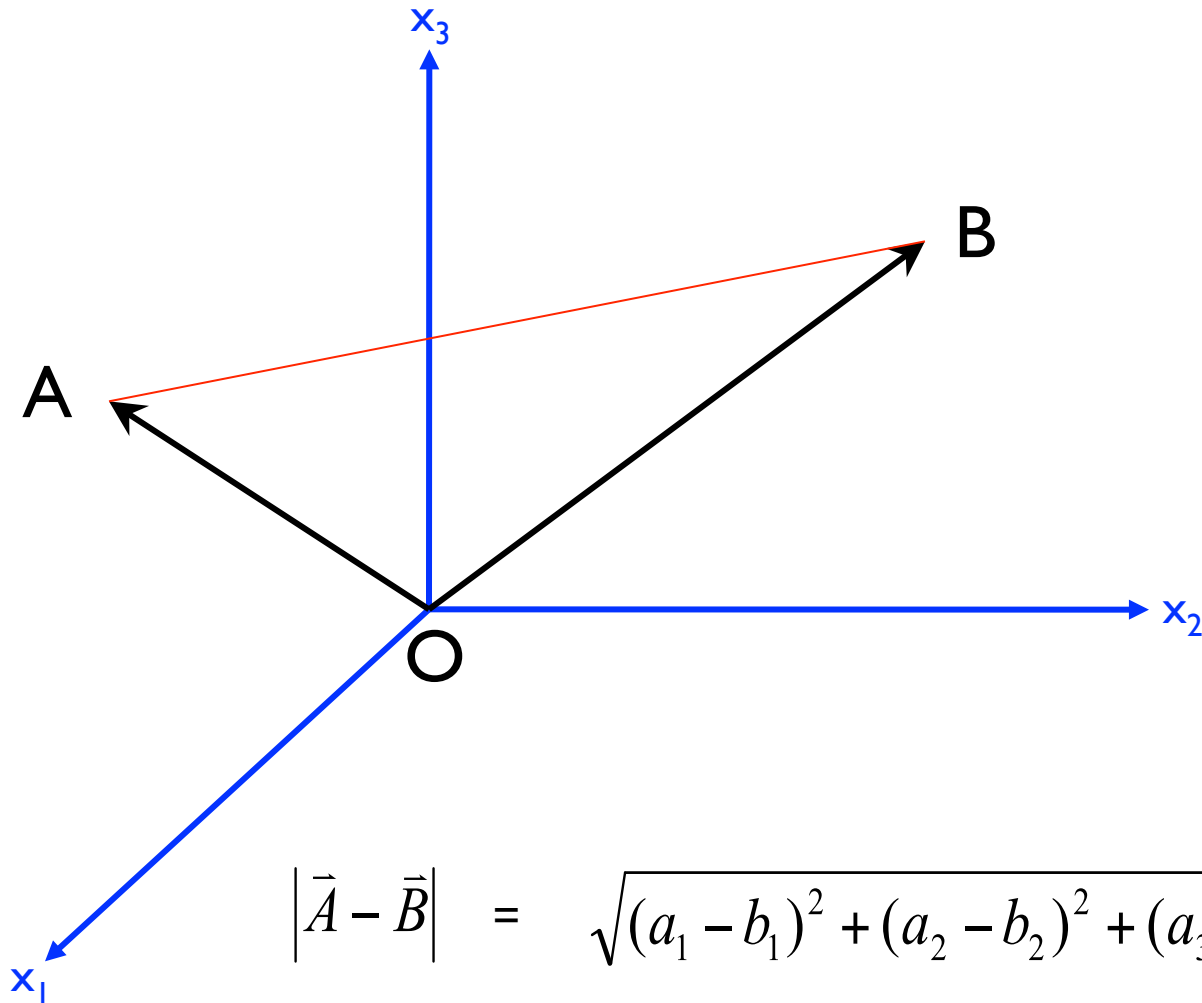


7遺伝子の発現プロファイル間の類似性は 7次元空間での距離によって決まる

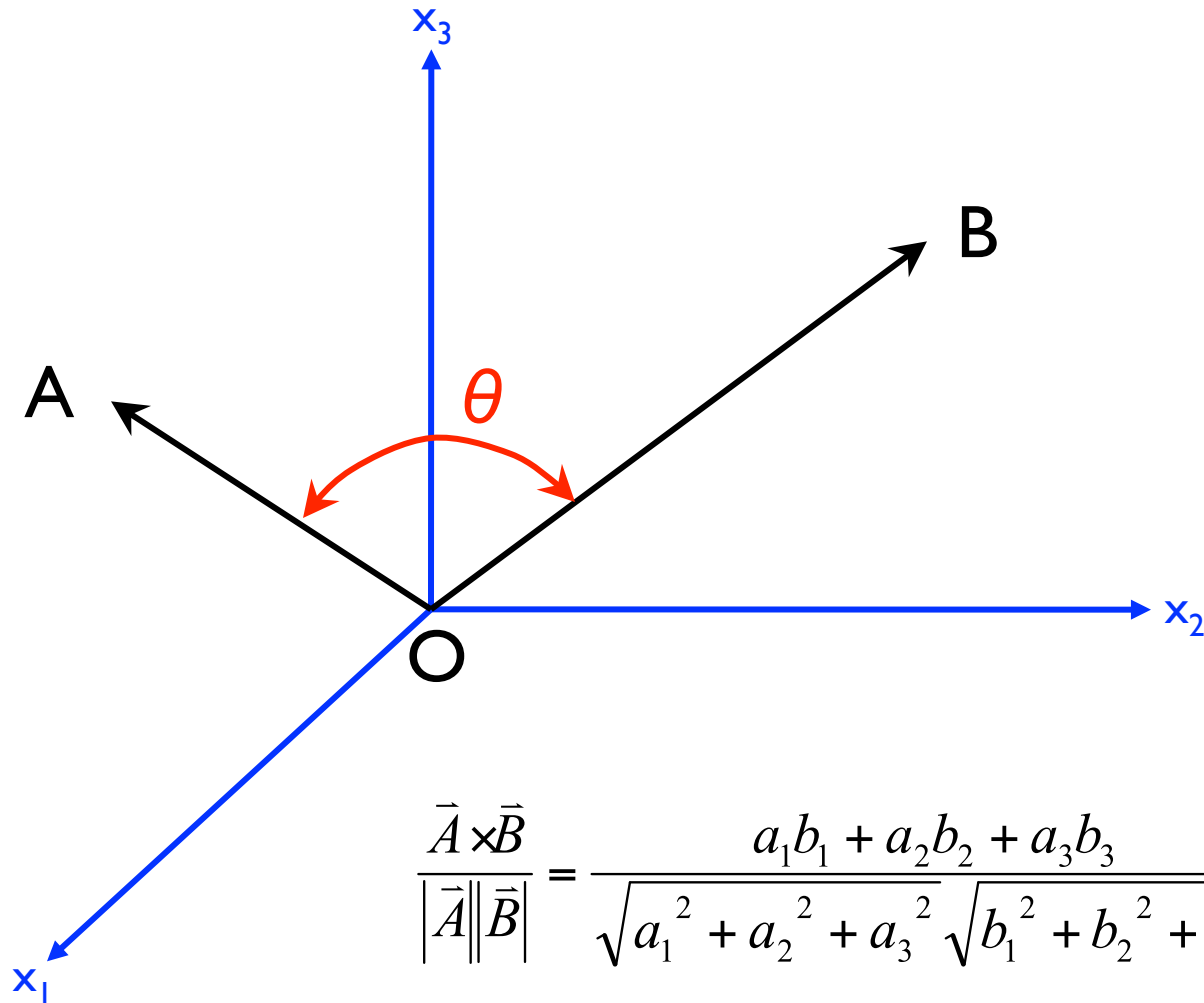


距離の基準を何にするか
距離尺度

ユークリッド距離

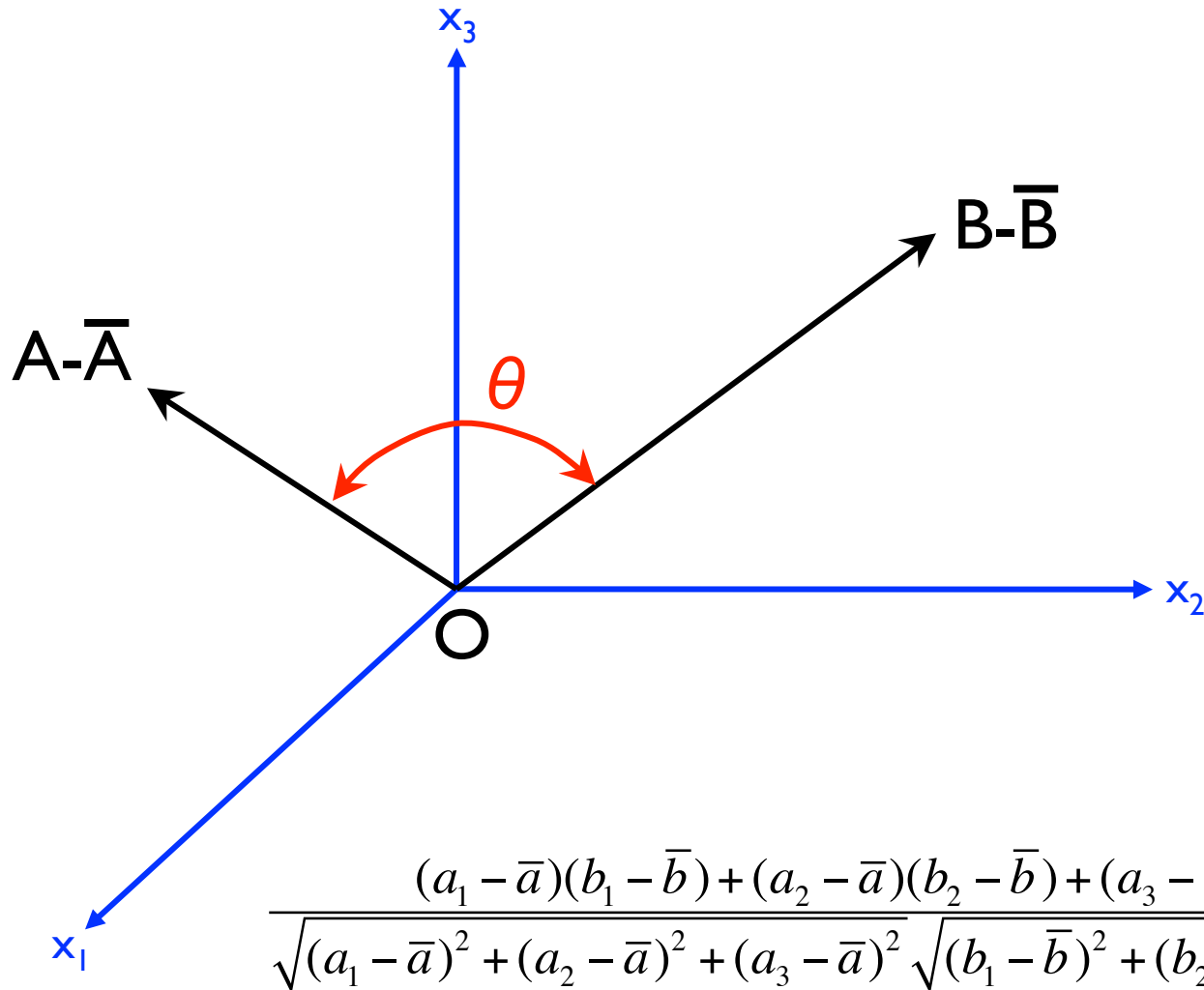


Uncentered Pearson correlation coefficient = $\cos\theta$



相関係数

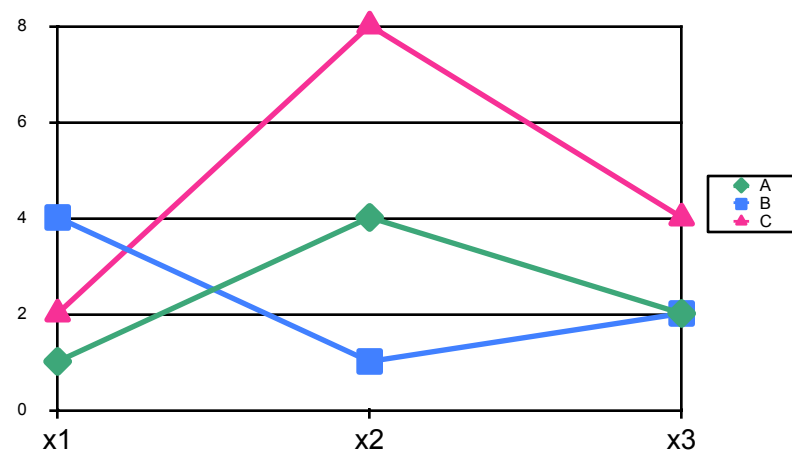
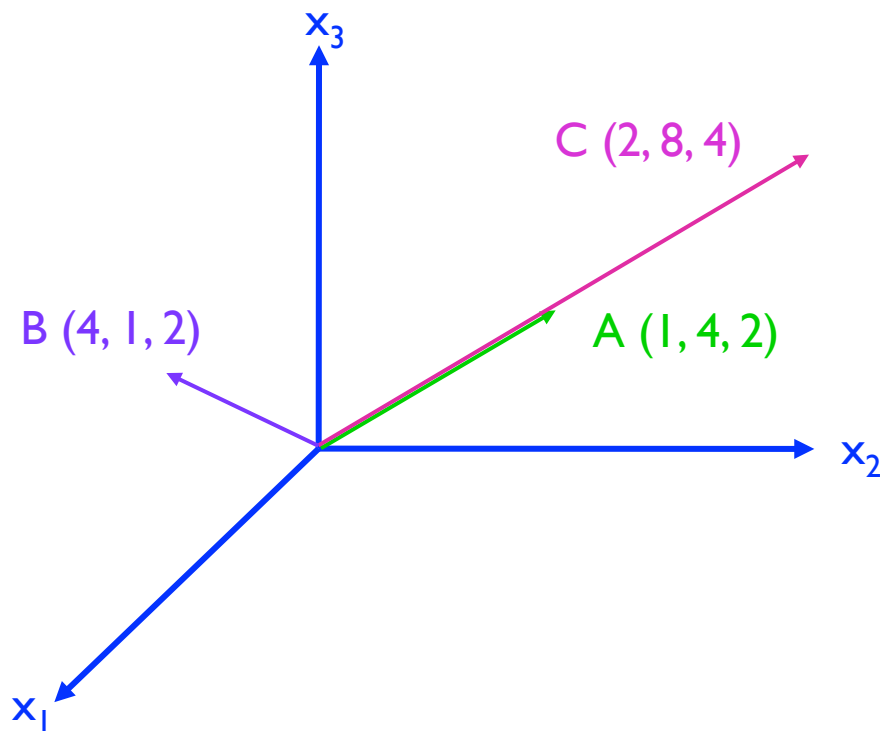
Pearson correlation coefficient



距離尺度の違い→解析対象の違い:

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



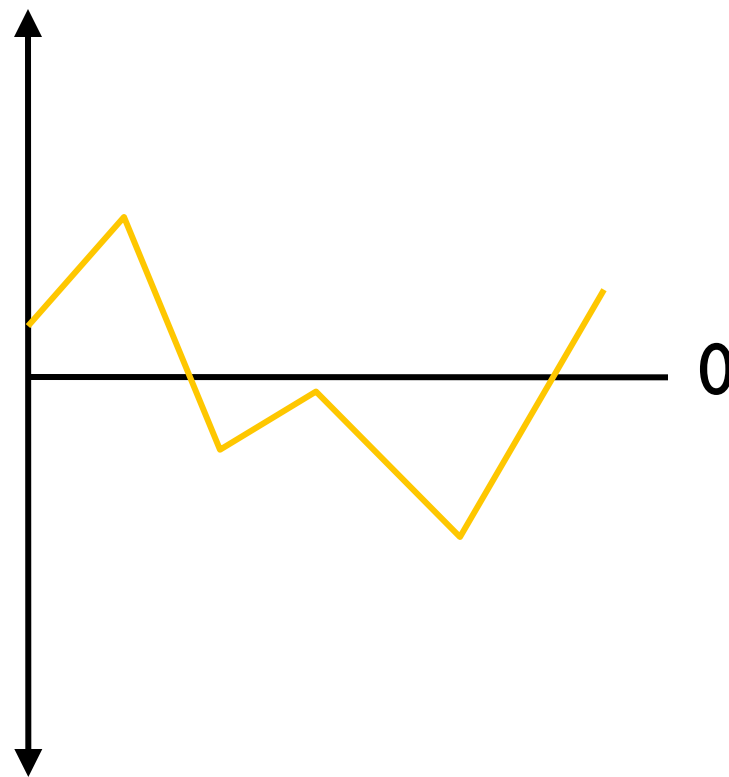
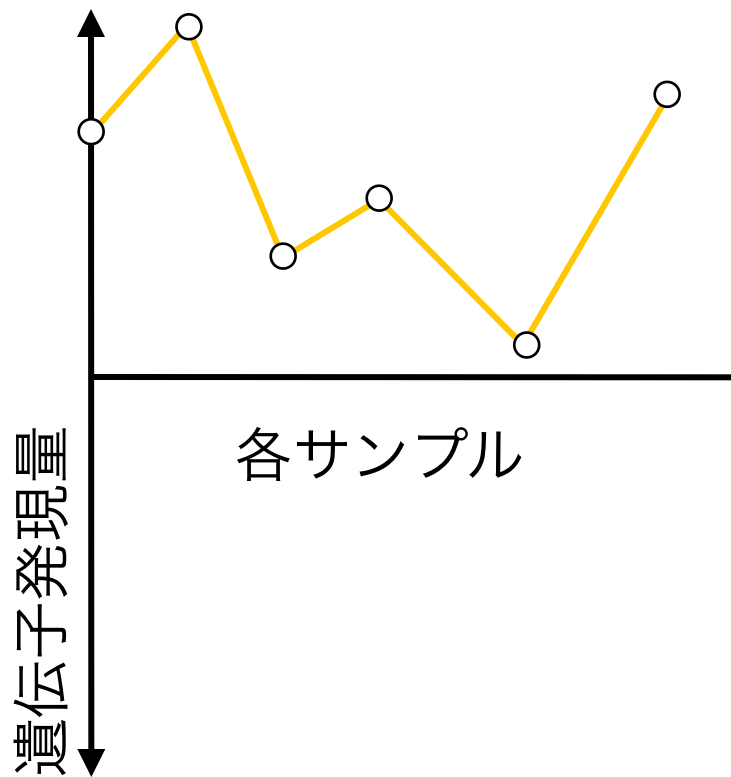
どの距離尺度を使うか？

- どんなプロファイルを
同じプロファイルと定義するか？
- 距離尺度計算の背後にあるものを
意識して選択する。

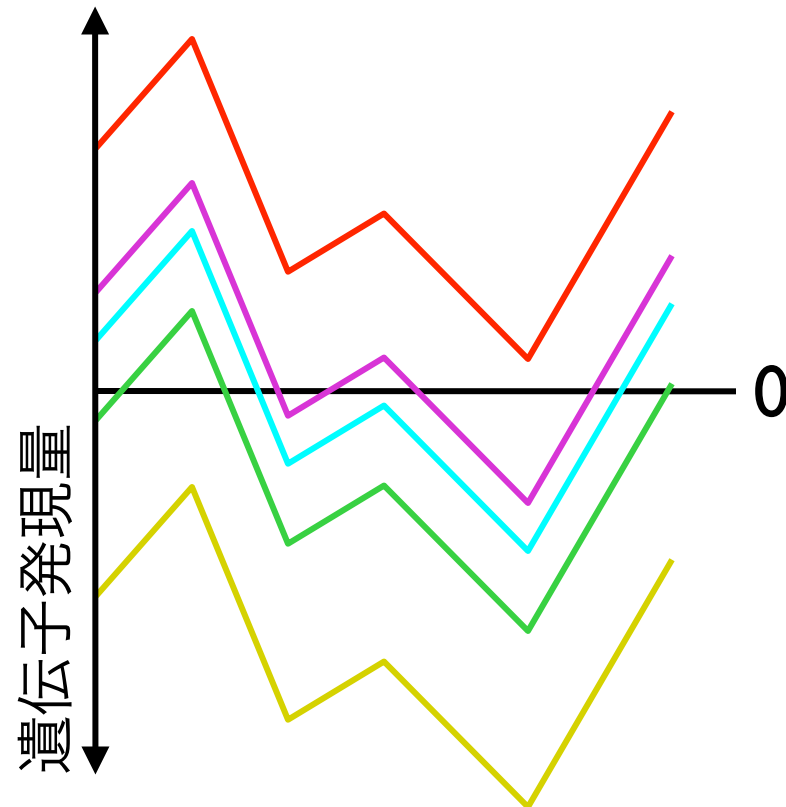
距離尺度計算の過程には

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

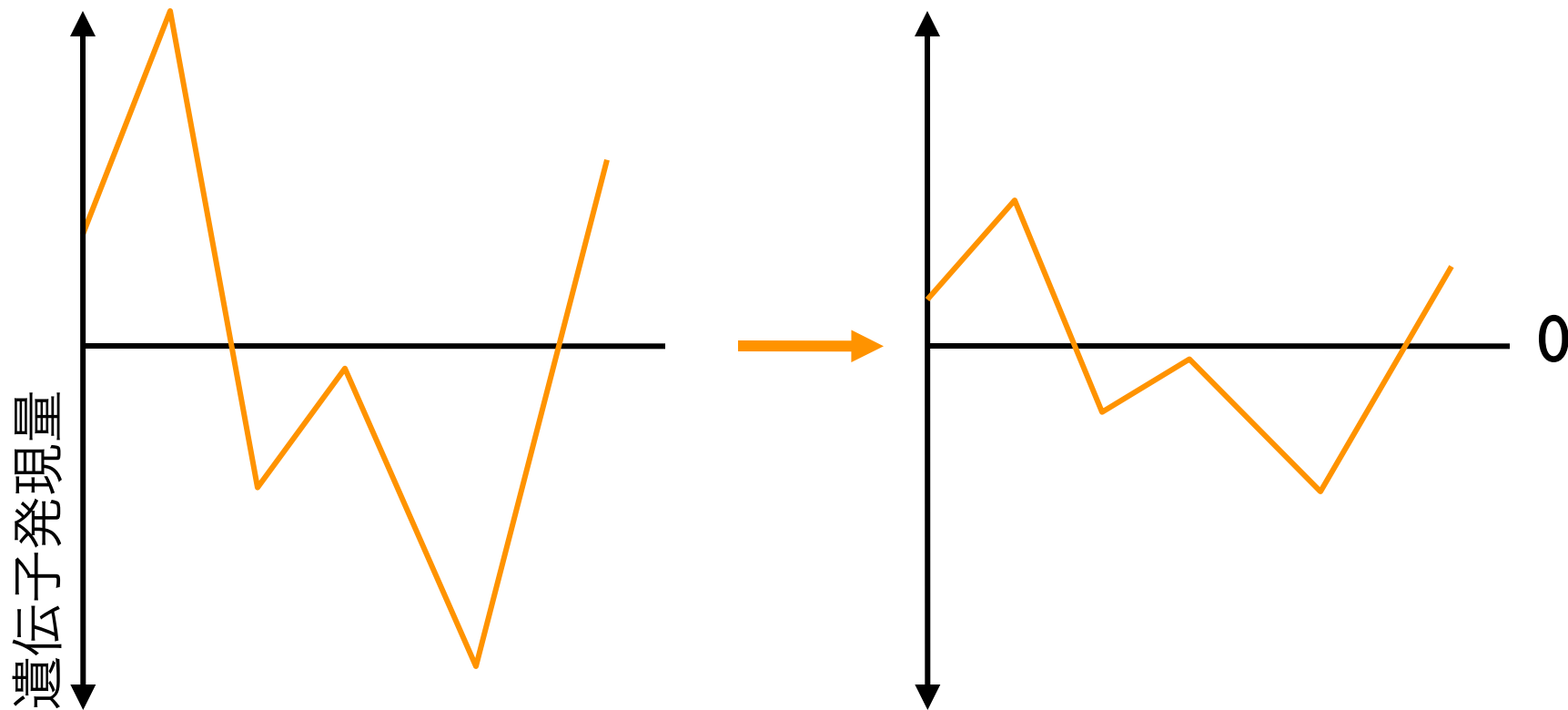
Centering



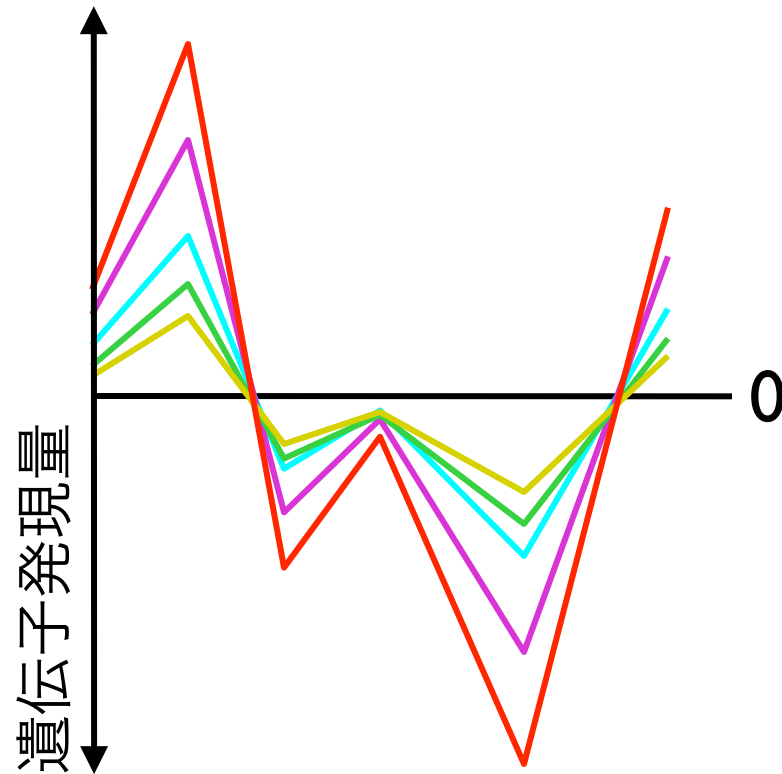
これらはcentering後は
全く同じプロファイルになる



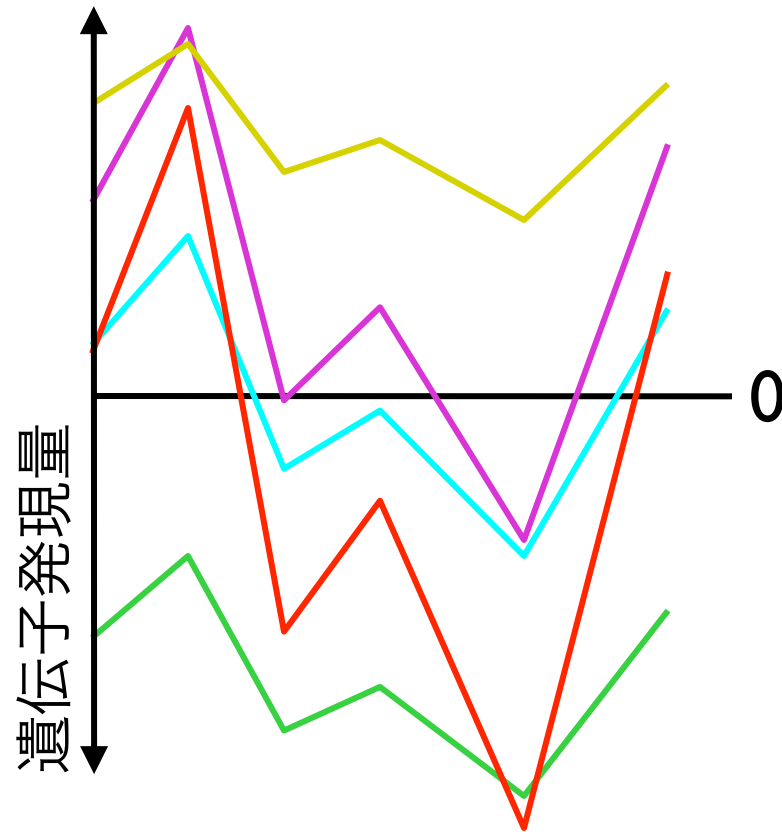
Scaling



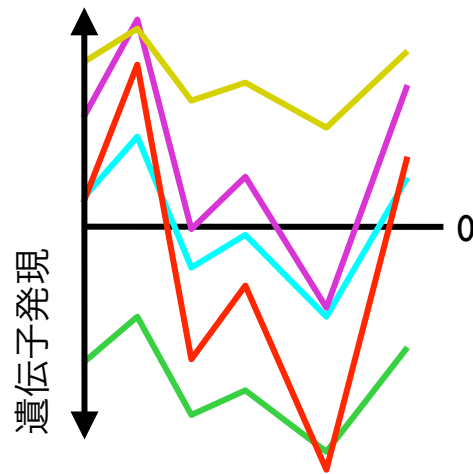
これらはscaling後は
全く同じプロファイルになる



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



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

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

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
- 基準サンプルがなく、パターンを比較: 相関係数

多変量解析の実際

教師有りか無しか

(supervised or unsupervised) ?

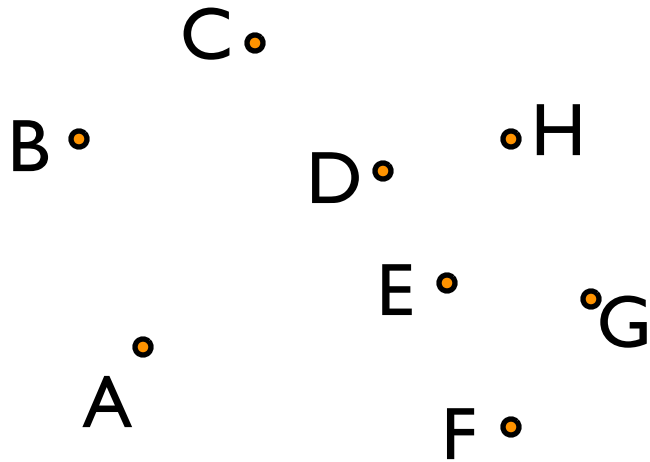
- 事前情報、前提はあるか？
- ある場合はk-means法などの利用を検討

どのような距離行列を使うか？

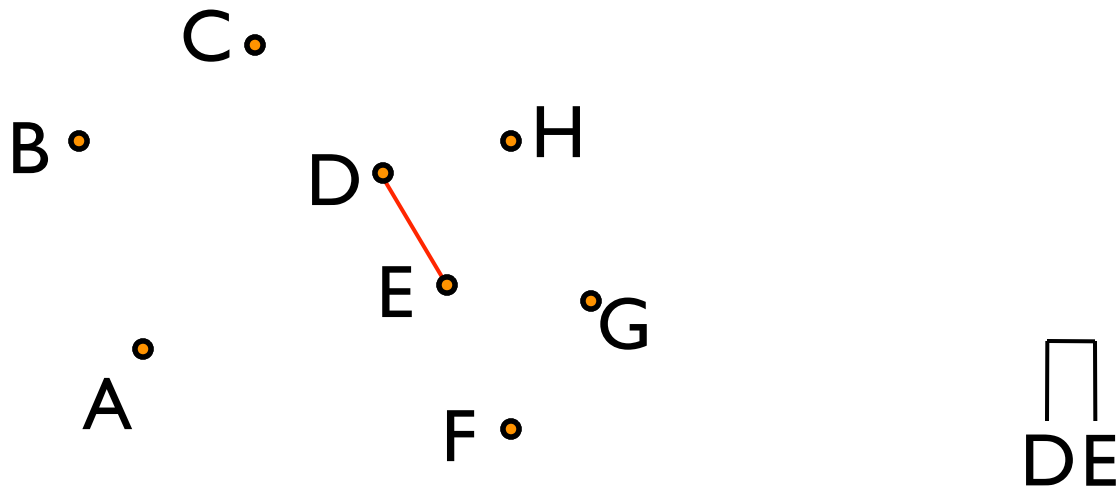
多変量解析の実際

階層クラスタリング

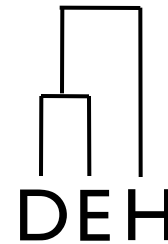
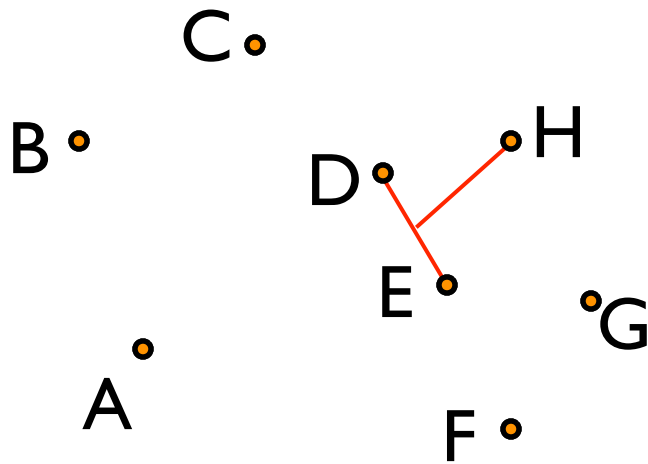
Agglomerative hierarchical clustering



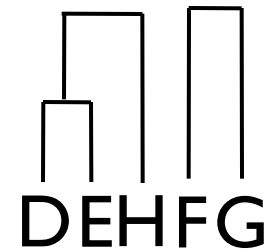
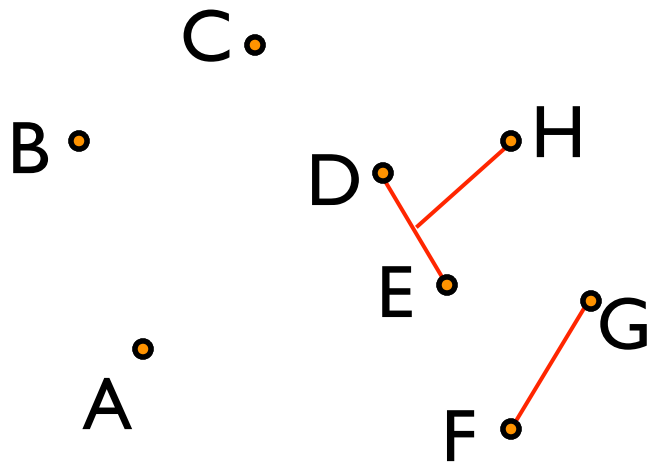
Agglomerative hierarchical clustering



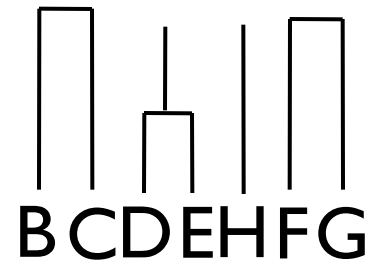
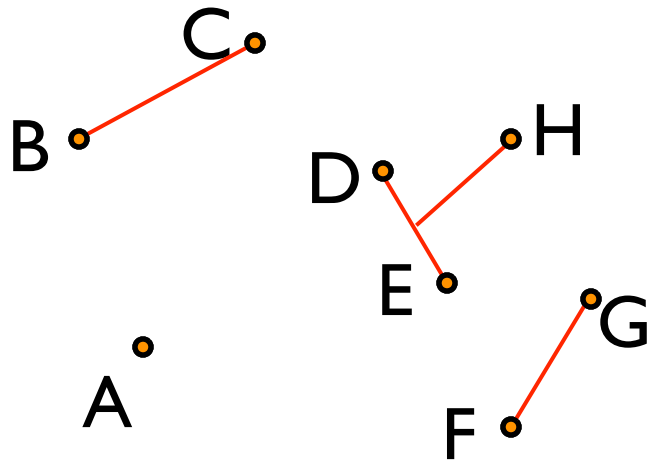
Agglomerative hierarchical clustering



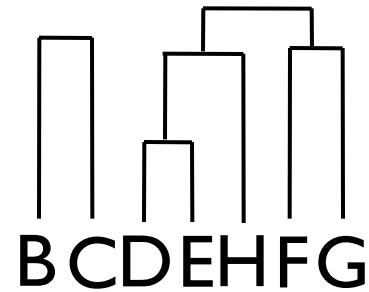
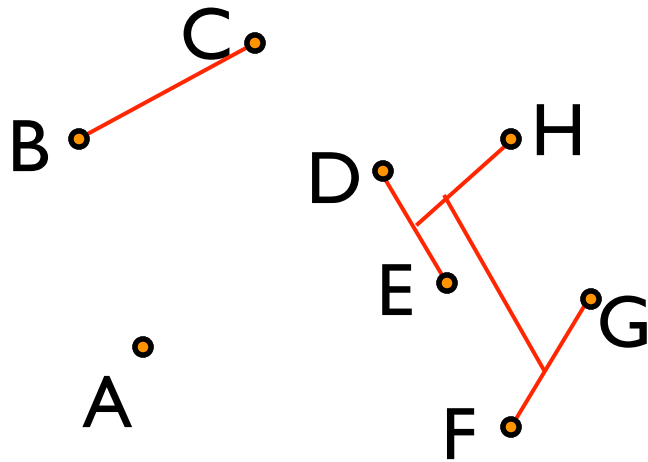
Agglomerative hierarchical clustering



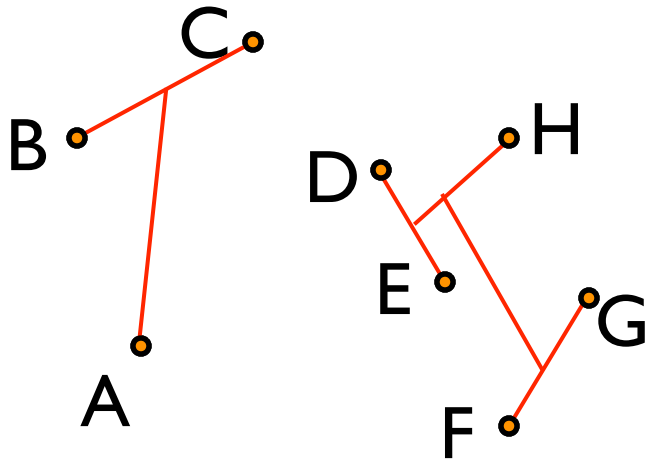
Agglomerative hierarchical clustering



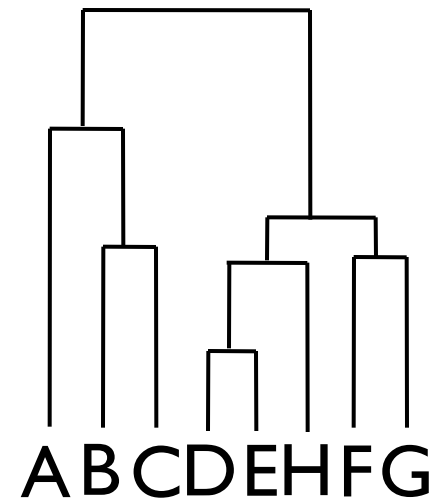
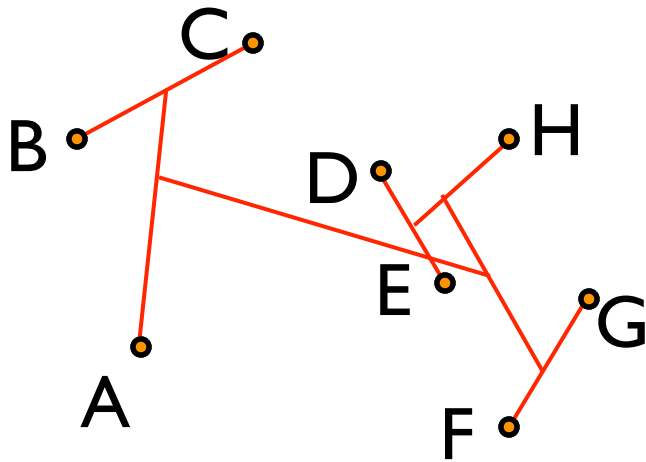
Agglomerative hierarchical clustering



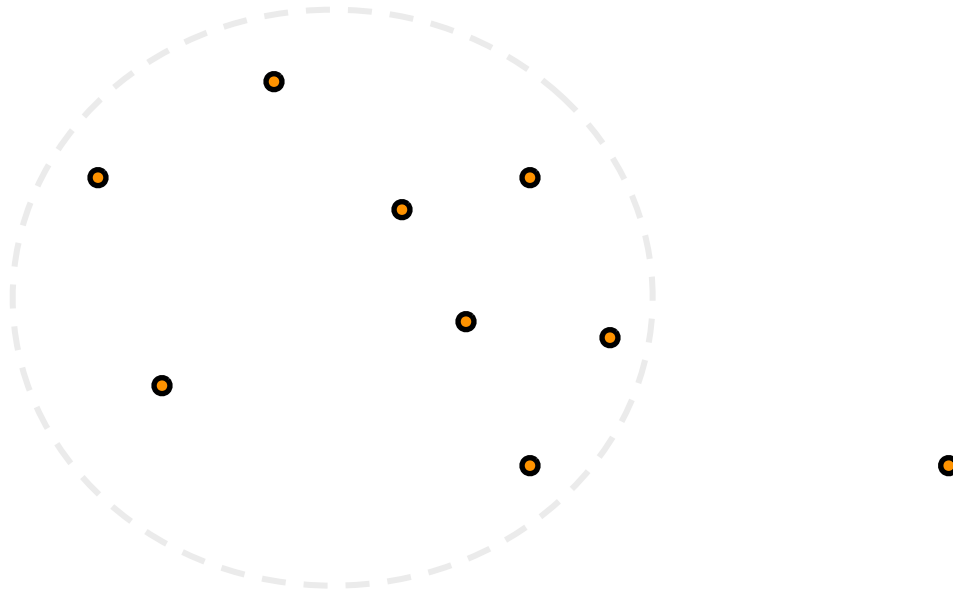
Agglomerative hierarchical clustering



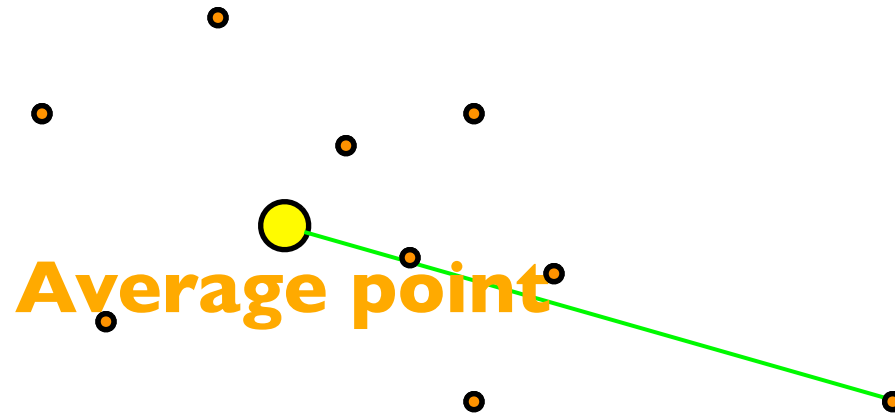
Agglomerative hierarchical clustering



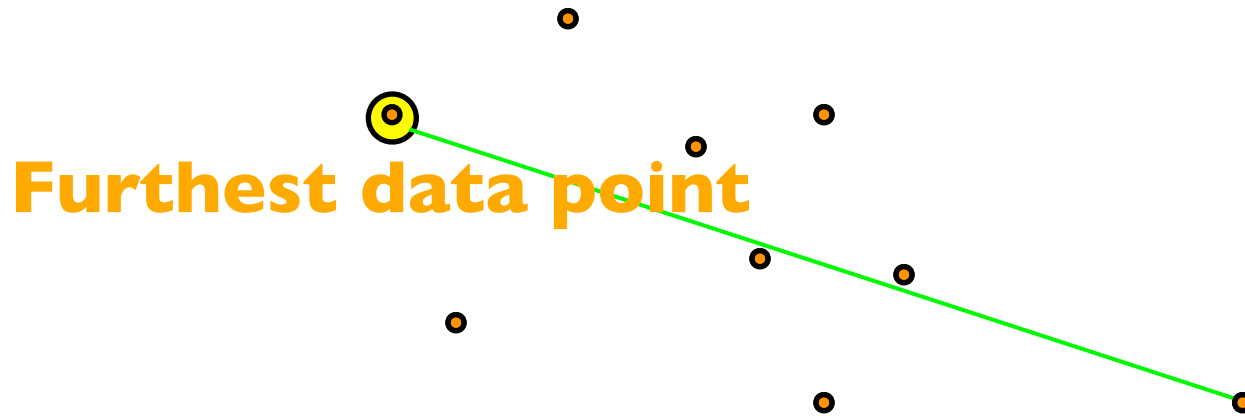
クラスター定義手法



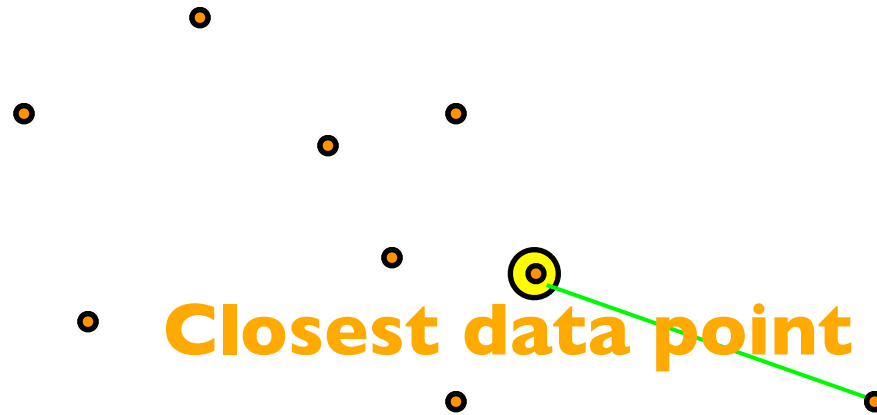
Average linkage



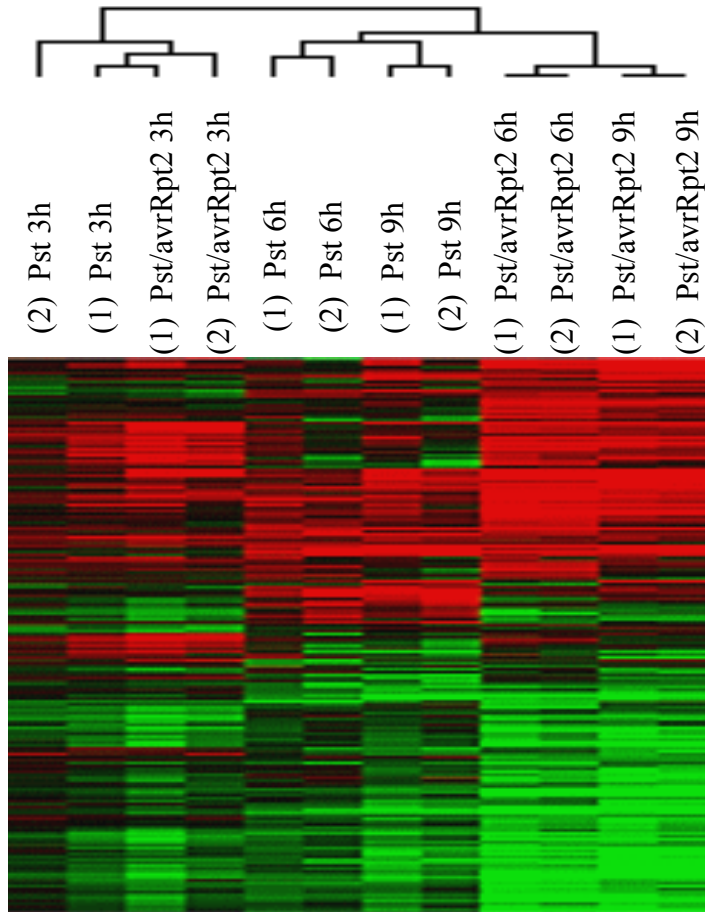
Complete linkage



Single linkage



階層クラスタリングの利点



- クラスタ化してより少数のカテゴリーを示す
- 人間が認識可能なパターンを示す

階層クラスタリングの欠点

- Bottom-up: 非常に「手順」依存性
- 一つの距離のみを指標としたクラスタリング

主成分分析

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

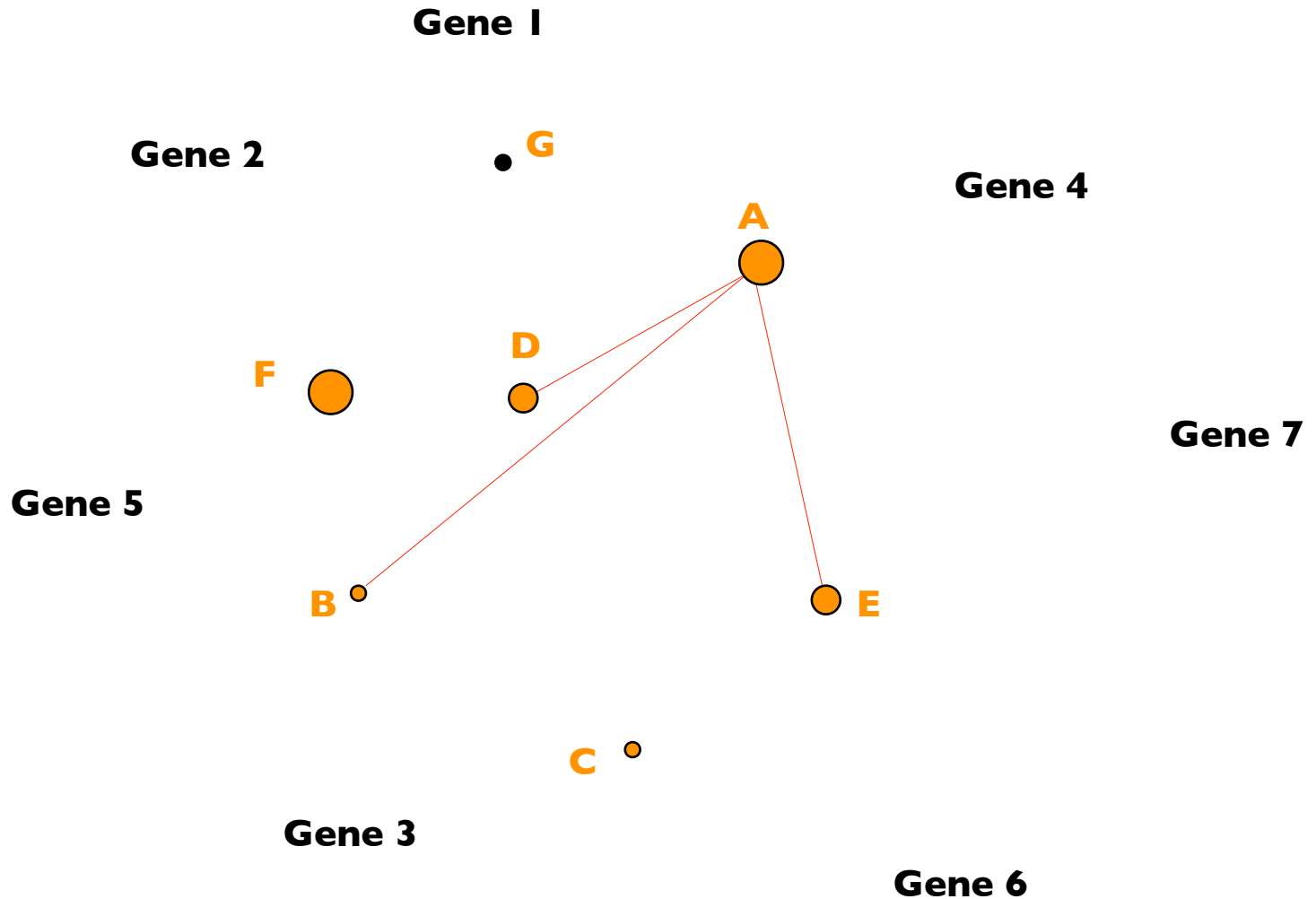
- ・ 大きなものから抽出し、
- ・ それら特徴を表現する**新たな軸**を作って

低次元でデータを可視化する

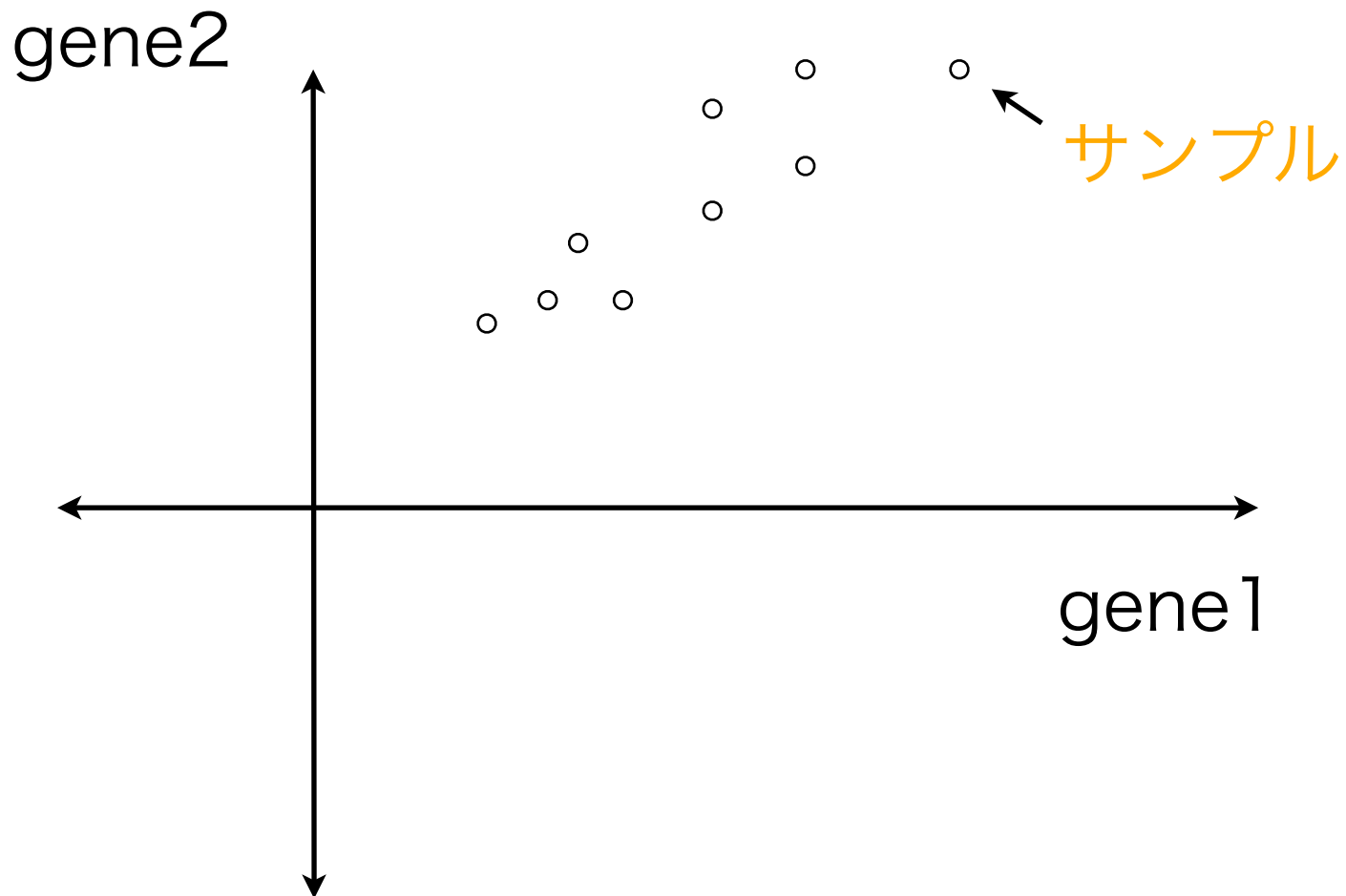
→ **人間が**新たな解釈を与える

階層クラスタリング:

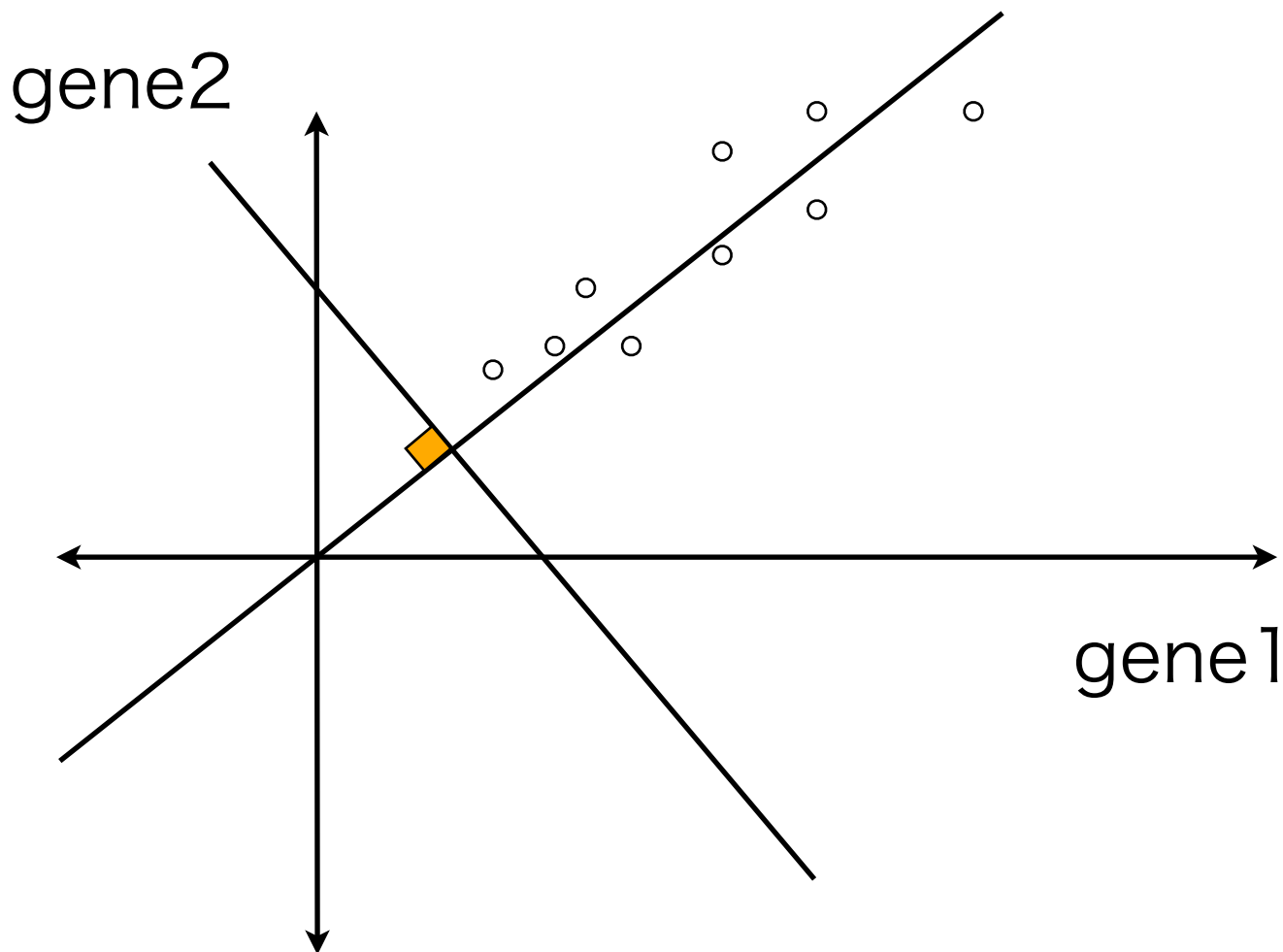
プロファイル間の類似性は空間での**1つの距離**によって決まる



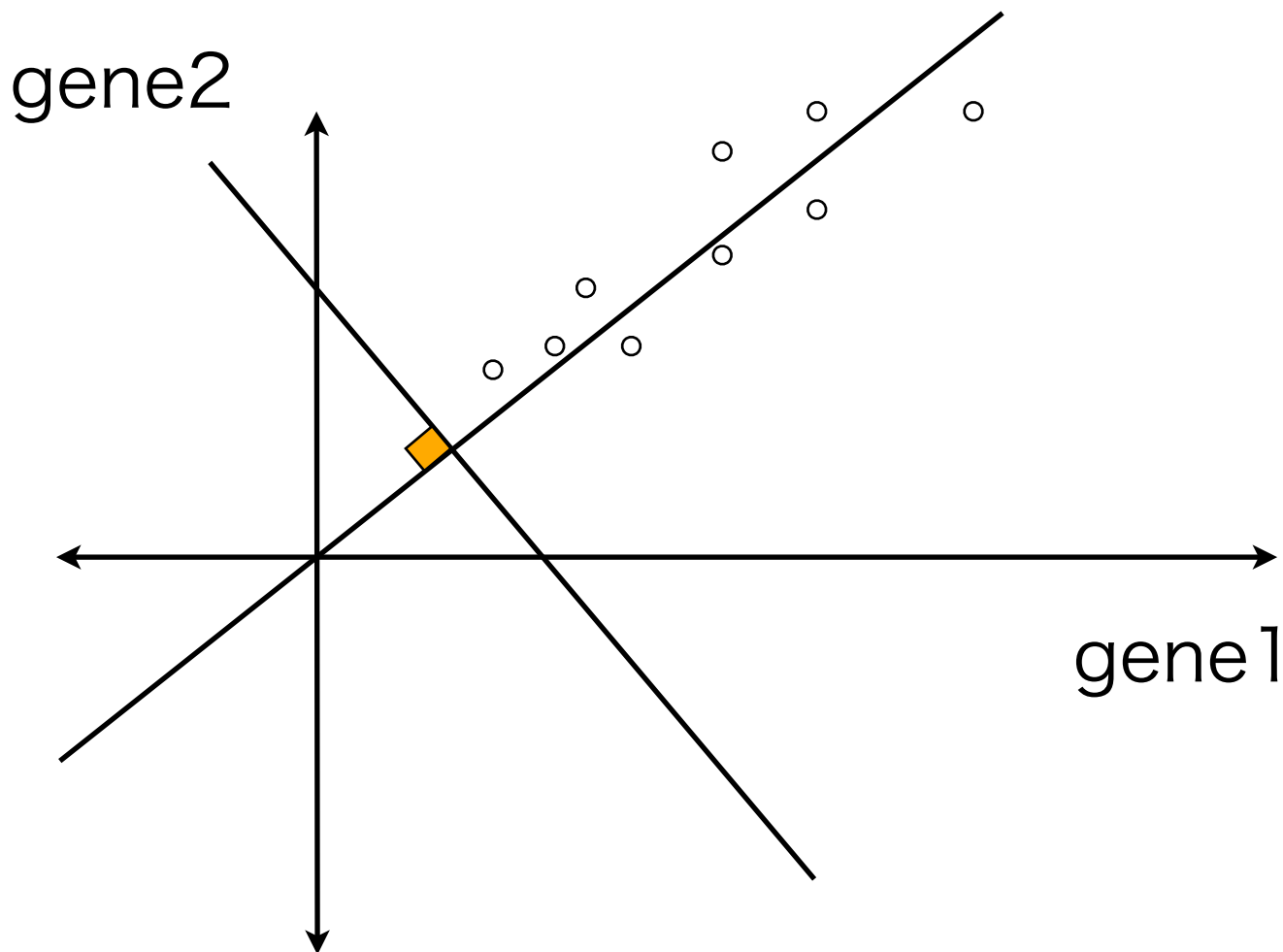
PCAは何をするのか？



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

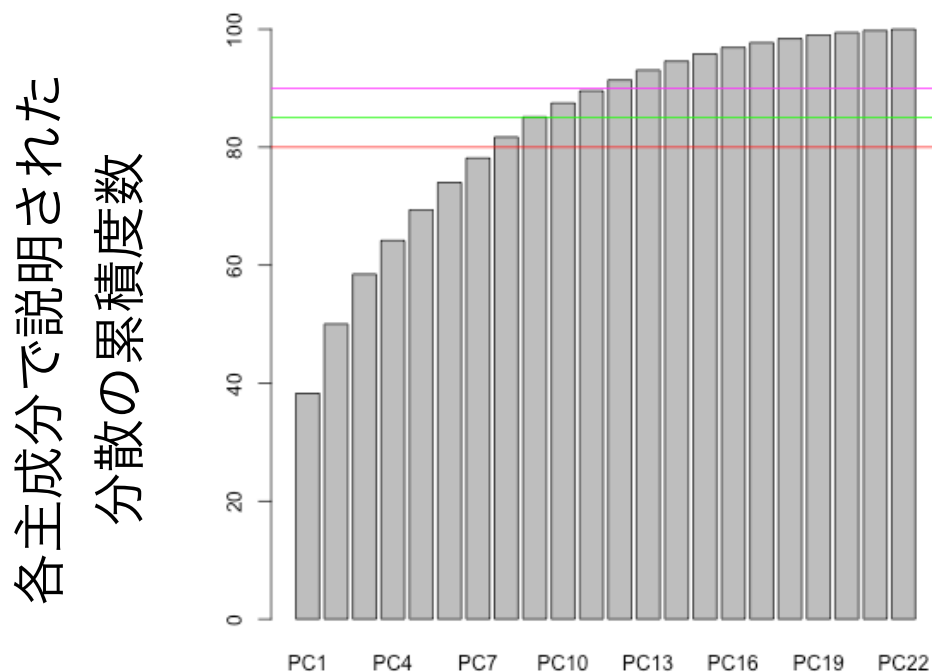
2. $a^2 + b^2 = 1$, u と v の相関係数0という制約の下でこれを解いて a_n, b_n を求める。

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

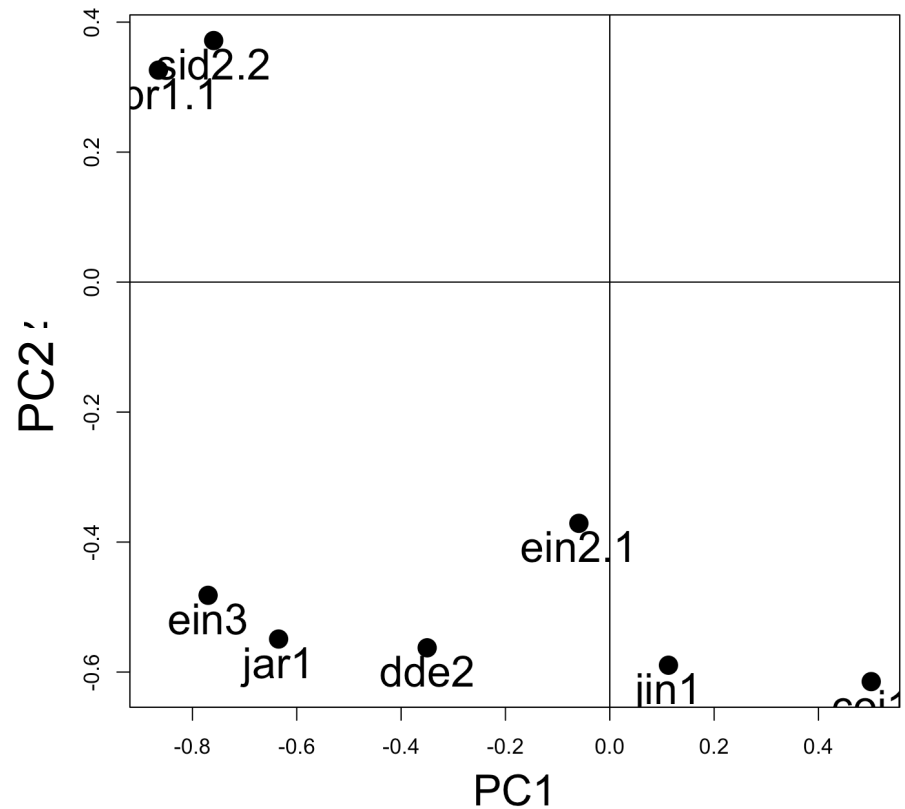
寄与率

- 各主成分が説明する分散の割合



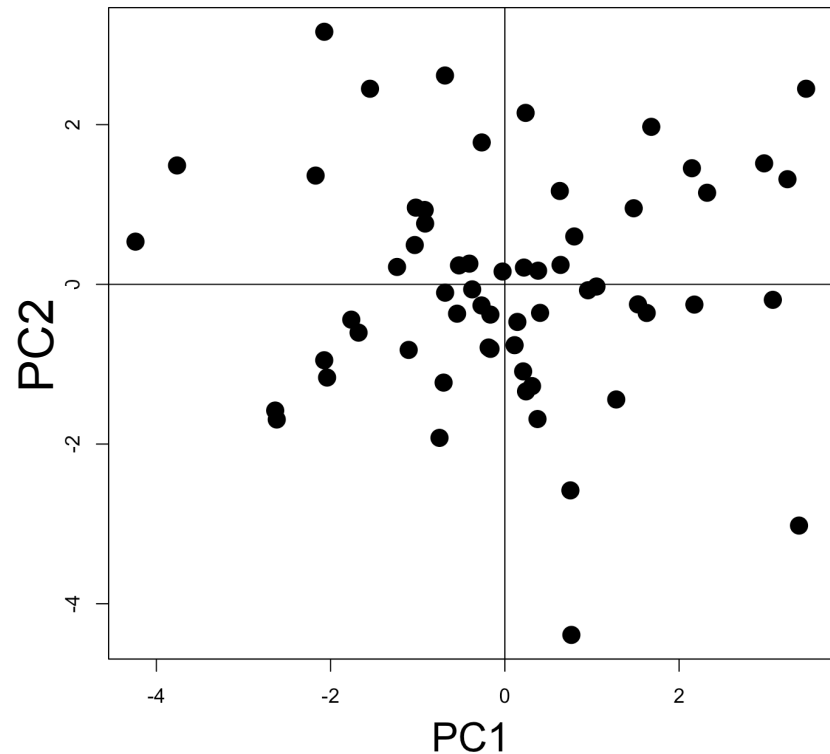
負荷量 loadings

- 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

- 各パラメーターの値を各主成分について標準化したもの



標準化: 平均0, SD=1

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),
Principle coordinate analysis

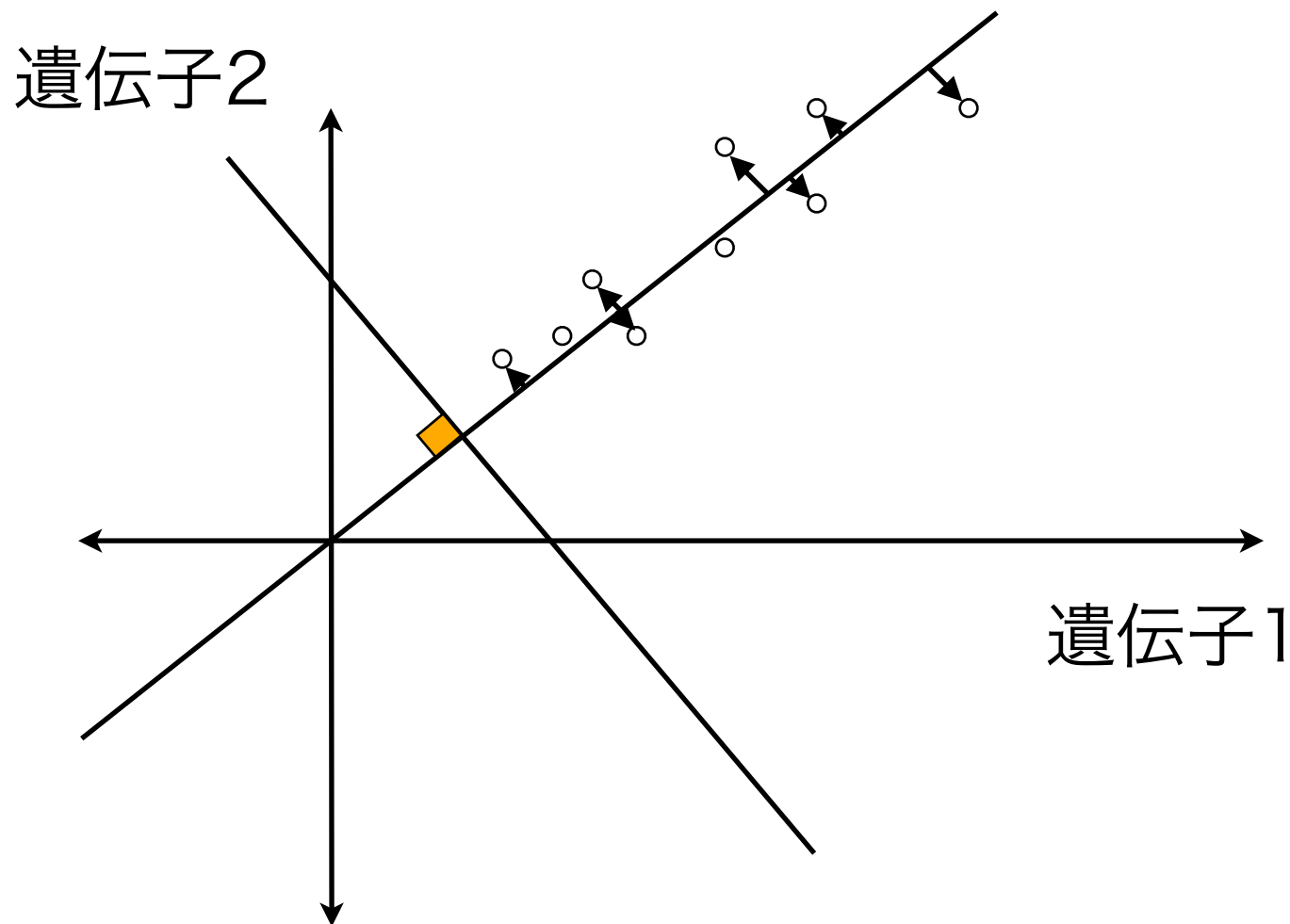
多次元尺度構成法とは？

モチベーション:

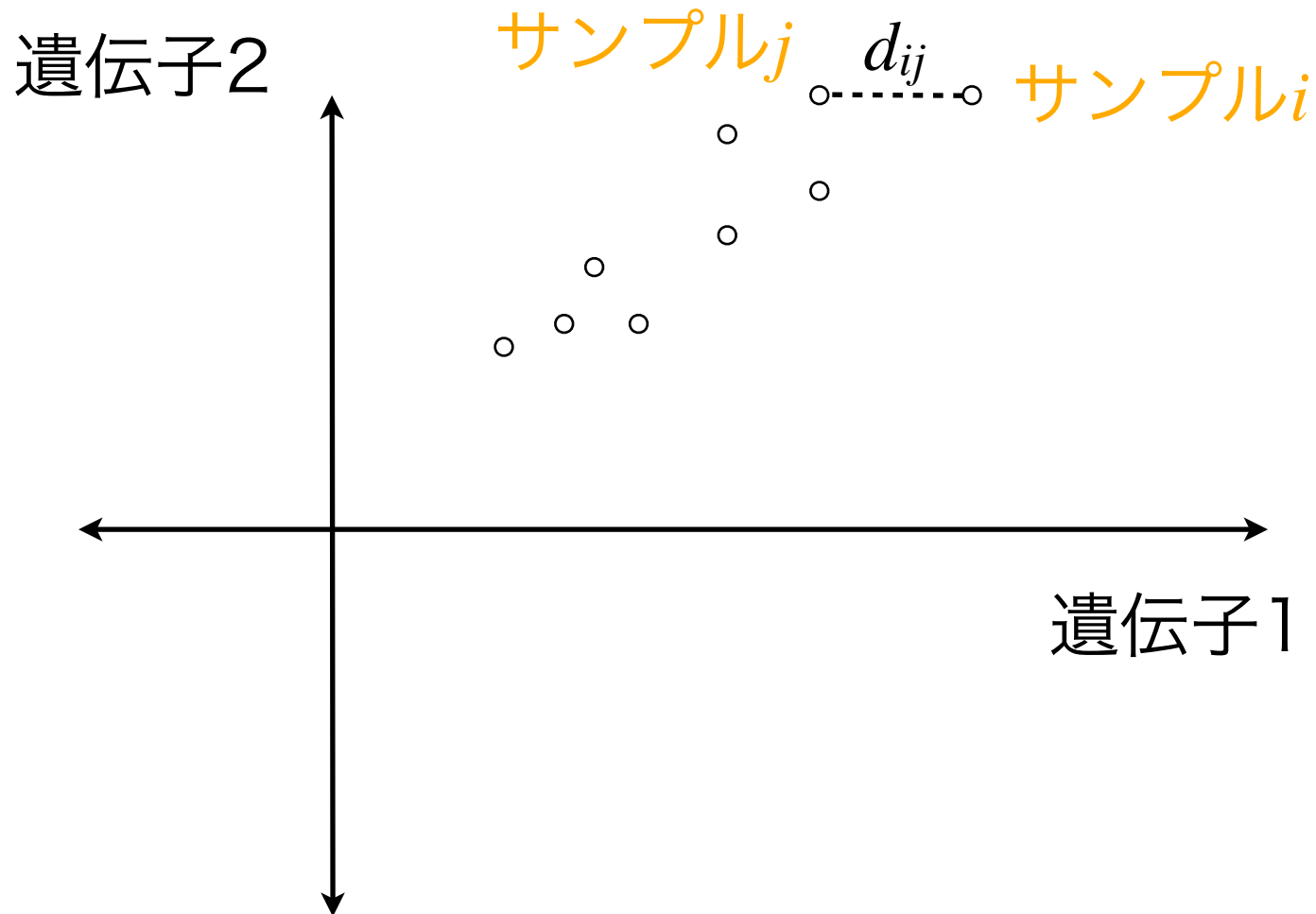
**多次元での各サンプル間の距離を保持して
低次元で表現する**

距離係数を元に次元圧縮するため、非線形
の関係にも対応 (PCA: 分散を使う [線形])

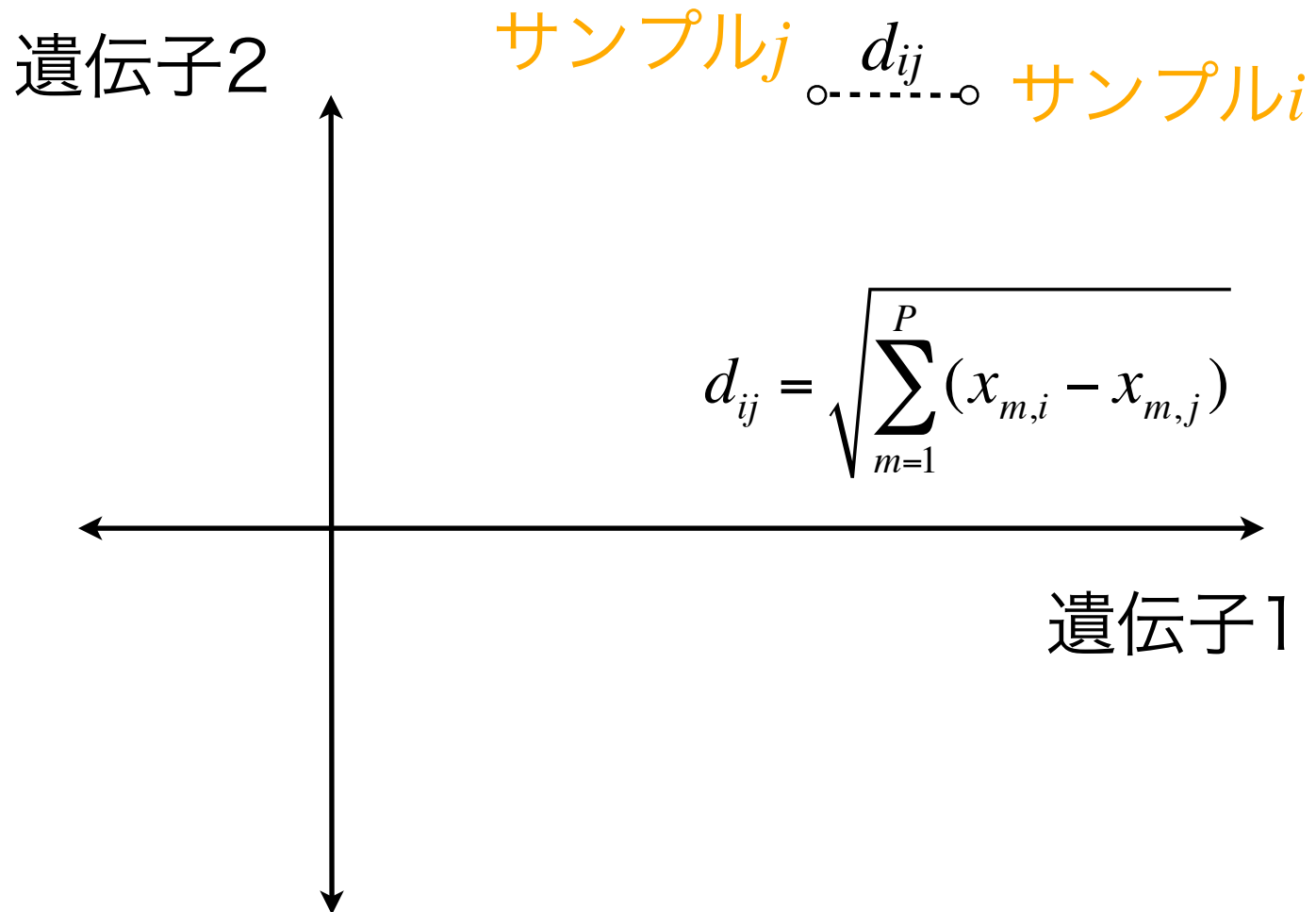
PCAが考慮する距離



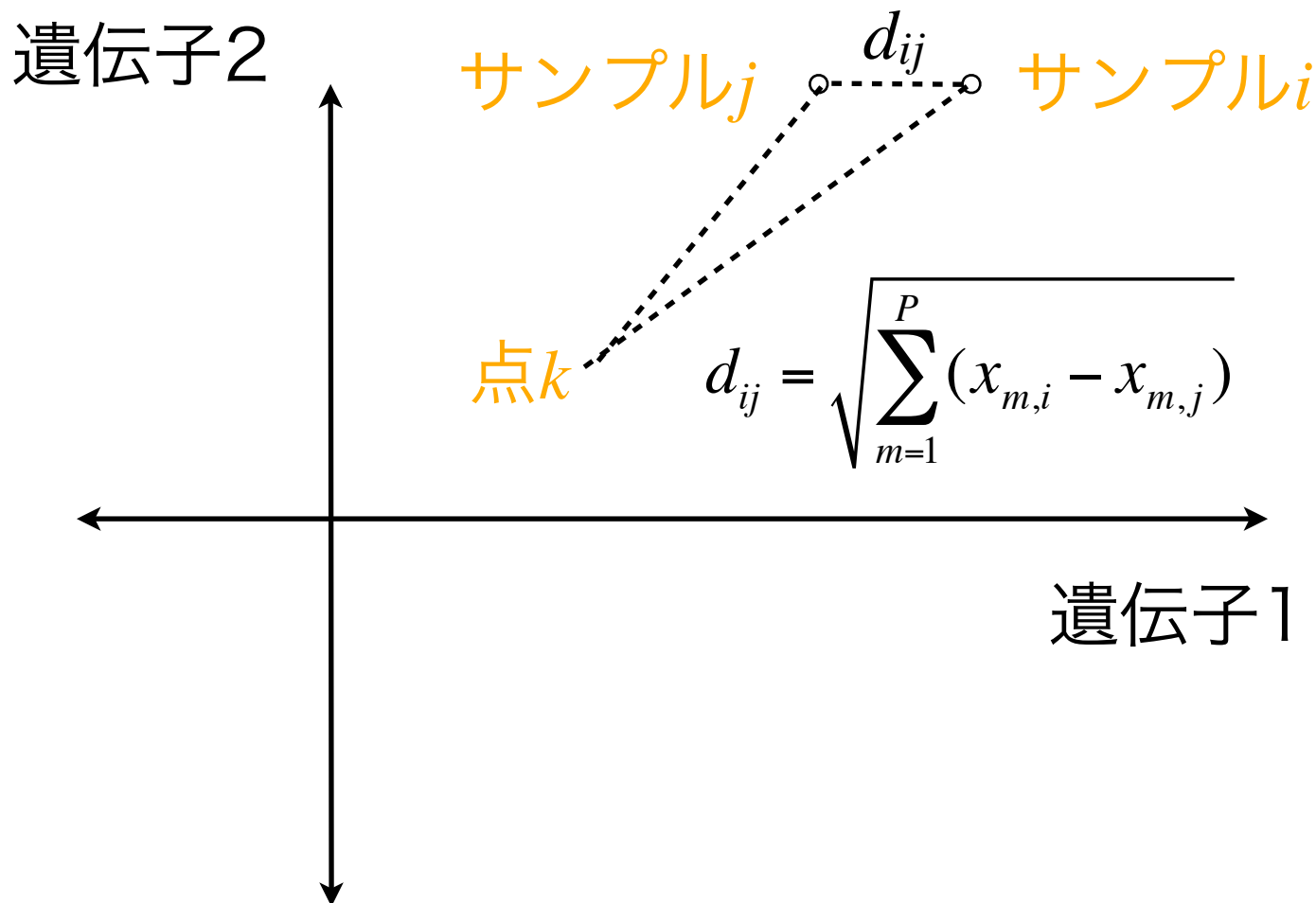
MDSが考慮する距離



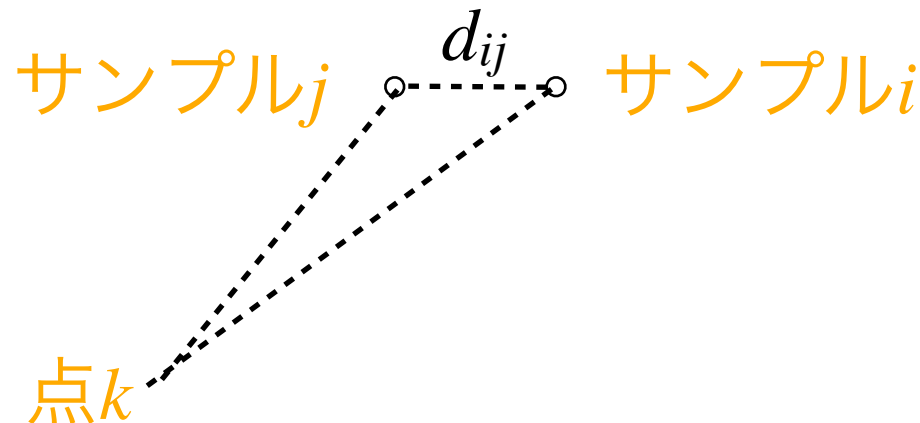
サンプル間の距離をまず計算する



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

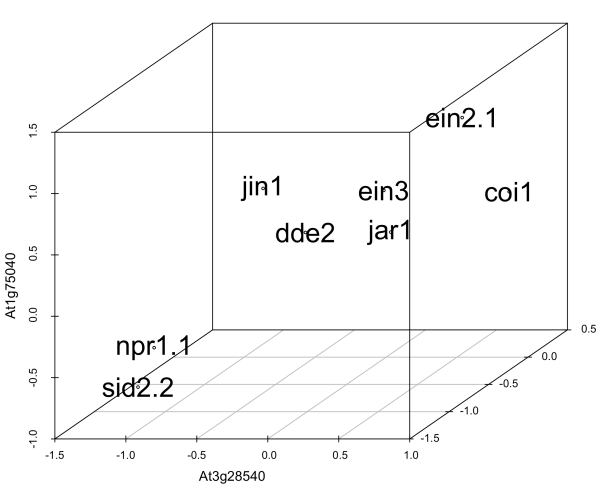


この定理はサンプル*i,j*に対し、どこを原点（点*k*）としても成り立つ

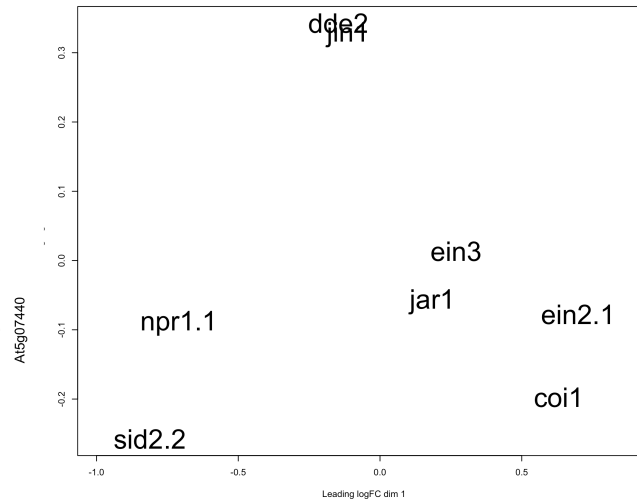


$$d_{ij}^2 = d_{ik}^2 + d_{jk}^2 - 2d_{ik}d_{jk}\cos\theta$$

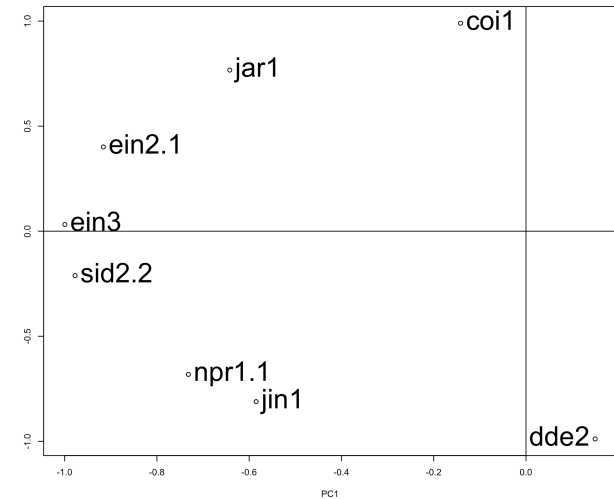
MDSとPCAの違い



Raw data (3 genes)



MDS (2D)



PCA (2 PCs)

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
- **軸**に意味がある（PCA） / ない（MDS）
- **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
- データの全体的な構造を保持しているか? PCA: yes, MDS: no.

(少しだけ発展) 多様体学習

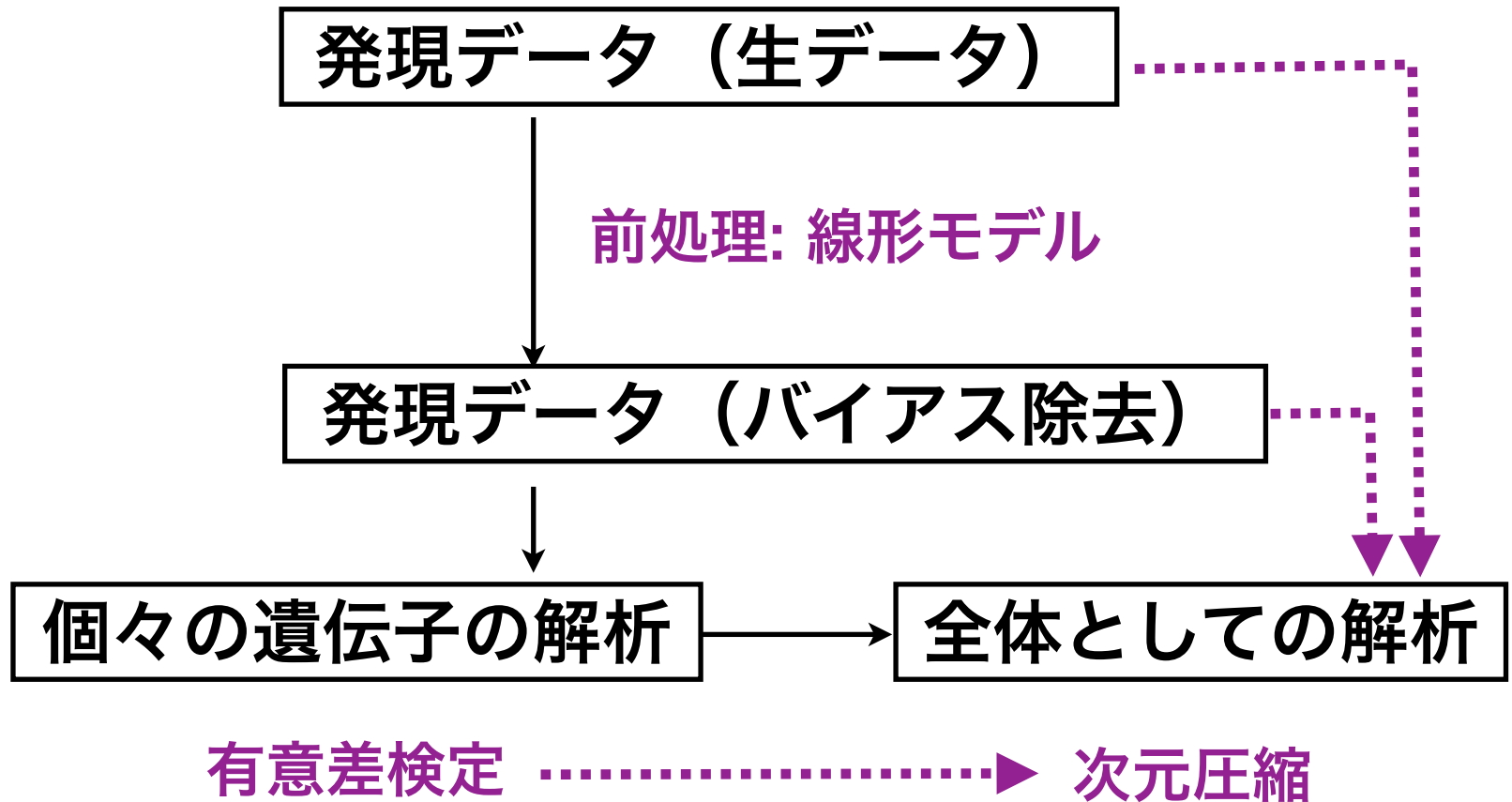
モチベーション：

非線形のデータ構造を低次元に圧縮して
表現する

- MDS
- LLE (locally linear embedding)
- Isomap
- t-SNE

注意点: 局所（サンプル間）の関係性を表現し、全体の構造は犠牲にしている

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
- 非線形クラスタリング・次元圧縮
 - ISOMAP, LLE (多様体学習)
 - self-organization map

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@abs.agr.hokudai.ac.jp