CC BY

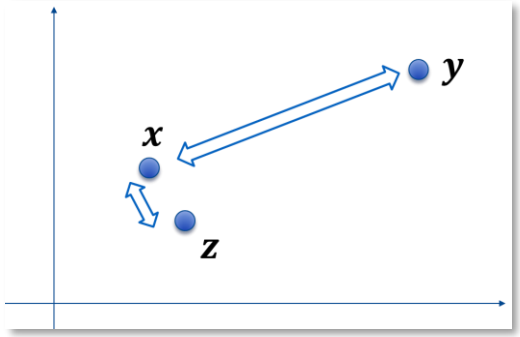
データサイエンス概論I＆II

クラスタリングと異常検出

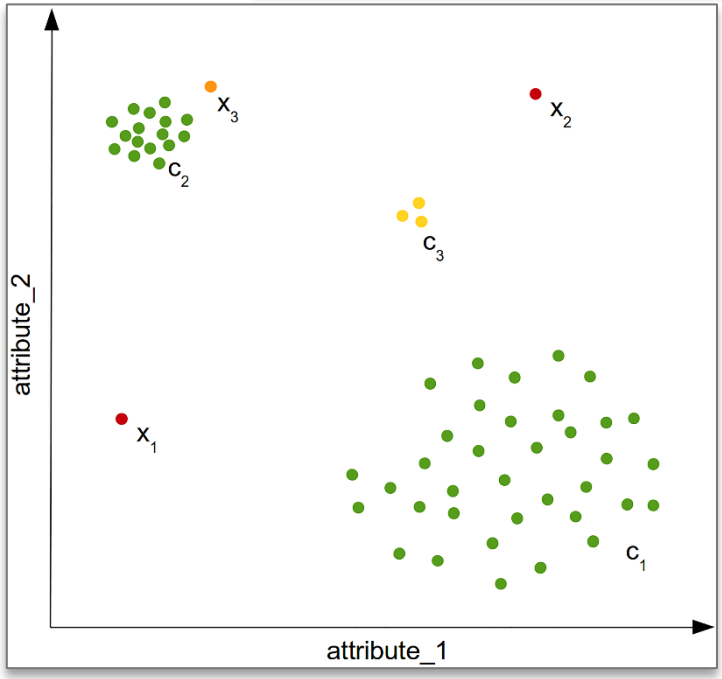
九州大学 数理・データサイエンス教育研究センター

（再掲）距離は超便利！

- データ間の比較が定量的にできる
  - 「xとyは全然違う/結構似ている」「xとyは28ぐらい違う」
  - 「xにとっては，yよりもzのほうが似ている」



- データ集合のグルーピングができる
  - 「近く」のデータどうしでグループを作る
  - 「クラスタリング」と呼ばれる



[Goldstein, Uchida, PLoSONE, 2016]

- データの異常度が測れる
  - 「近く」にデータがたくさんあれば正常，一つもなければ異常

データのクラスタリング①
クラスタリングの基本的考え方

クラスタリング＝“clustering”

データ集合をグルーピングする

- 我々は「距離」や「類似度」を手に入れた
- その結果，「与えられたデータ集合」を「それぞれ似たデータからなる幾つかのグループに分ける」ことが可能に！

- 要は「大量のデータを整理整頓してわかりやすく！」

グルーピング（クラスタリング）は何が便利？
全国ラーメン屋の「味」データを例に考える

- いくつのグループ（クラスタ）に分かれたか？
  - グループの分布の把握

主要なラーメンのタイプはいくつあるのか？

- 各クラスタのメンバ数はどうなっているか？
  - 各グループのメンバの多寡
  - 微小クラスタ（マイノリティ）や，巨大クラスタ（マジョリティ）

各タイプの店がどれぐらいあるのか？（どの味がマジョリティー？）

- 各クラスタの代表データ(代表ベクトル)は？
  - 主要例を理解
  - 膨大なデータを高々数個で概観

各タイプの最も代表的な店はどれか？（それだけ食べ歩けばラーメンの味はほぼわかる）

- 各クラスタの広がりは？
  - 各グループの変動傾向理解

各タイプで，味のバリエーションはどれぐらい？

グルーピング（クラスタリング）は何が便利？ アンケートデータを例に考える

- いくつのグループ（クラスタ）に分かれたか？
  - グループの分布の把握

典型的な回答のタイプは
いくつあるのか？

- 各クラスタのメンバ数はどうなっているか？
  - 各グループのメンバの多寡
  - 微小クラスタ（マイノリティ）や，巨大クラスタ（マジョリティ）

「典型的な回答のタイプ」
それぞれはどれぐらい
メジャーなのか？

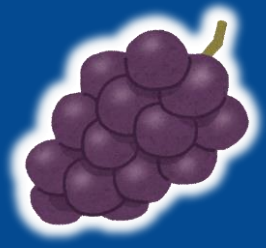
- 各クラスタの代表データ(代表ベクトル)は？
  - 主要例を理解
  - 膨大なデータを高々数個で概観

典型的な回答は，
どんな回答パターンか？
全回答をみなくても，
ざっくり傾向がわかる！

- 各クラスタの広がりは？
  - 各グループの変動傾向理解

各タイプに属する回答例の
バリエーションはどれぐらい？

クラスタ(cluster)とは？

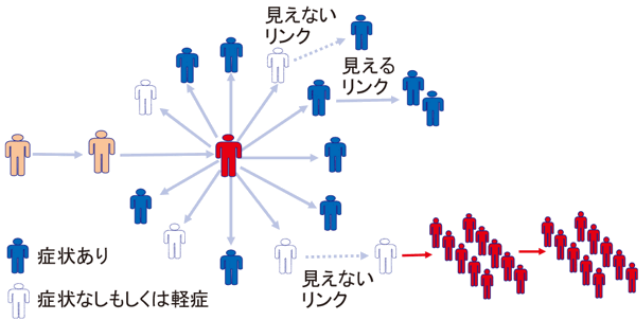


- 「房（ふさ）」,「集団」,「群」

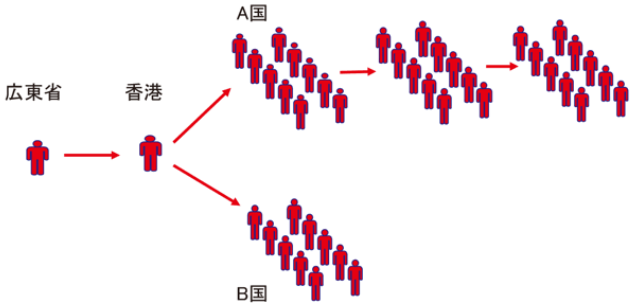
http://www.ims.cs.uec.ac.jp/~naoki

星団(star cluster)

クラスタ感染[国立感染症研究所]



A) COVID-19クラスター模式図：1人の感染者から多数の患者に感染させるイベントを考えたときに（図中央■）、症状の有無にかかわらずほとんどが他者に感染させない（■の周りの症例）。症状なしもしくは軽症な人から次のクラスターが形成（図右下の■集団）されるには、見えないリンクを介する場合もあり、感染連鎖を把握することは難しい。



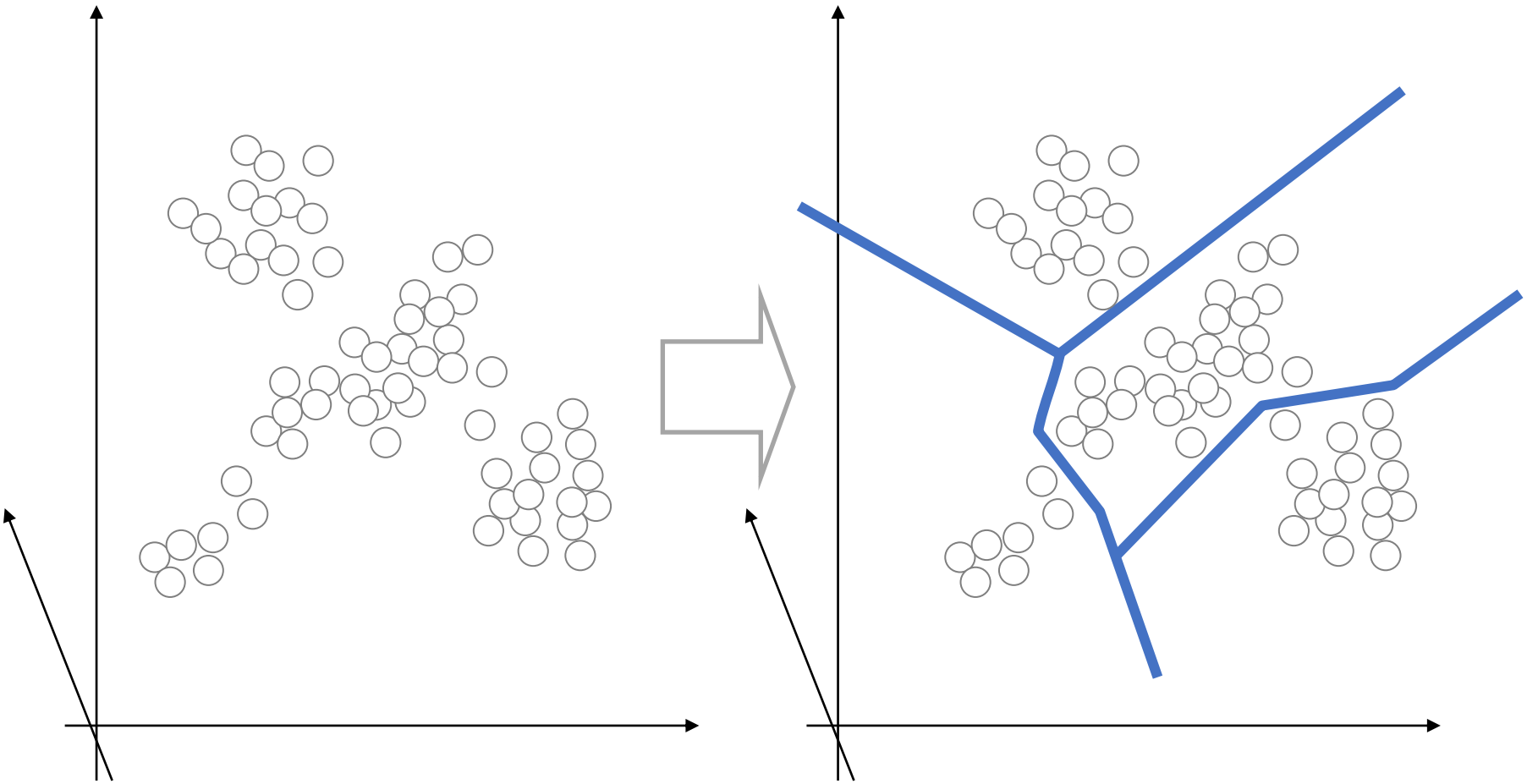
B) SARSクラスター模式図：同じコロナウイルスでも、重症急性呼吸器症候群（SARS）コロナウイルスによるアウトブレイクにおいては、感染者の多くが重い肺炎を起こすために患者の発見が比較的容易であったことも相まって、世界中の感染連鎖を初期の広東省まで辿ることができ、結果として感染を封じ込めることができた。

図. 新型コロナウイルス感染症と重症呼吸器症候群の感染の広がり方

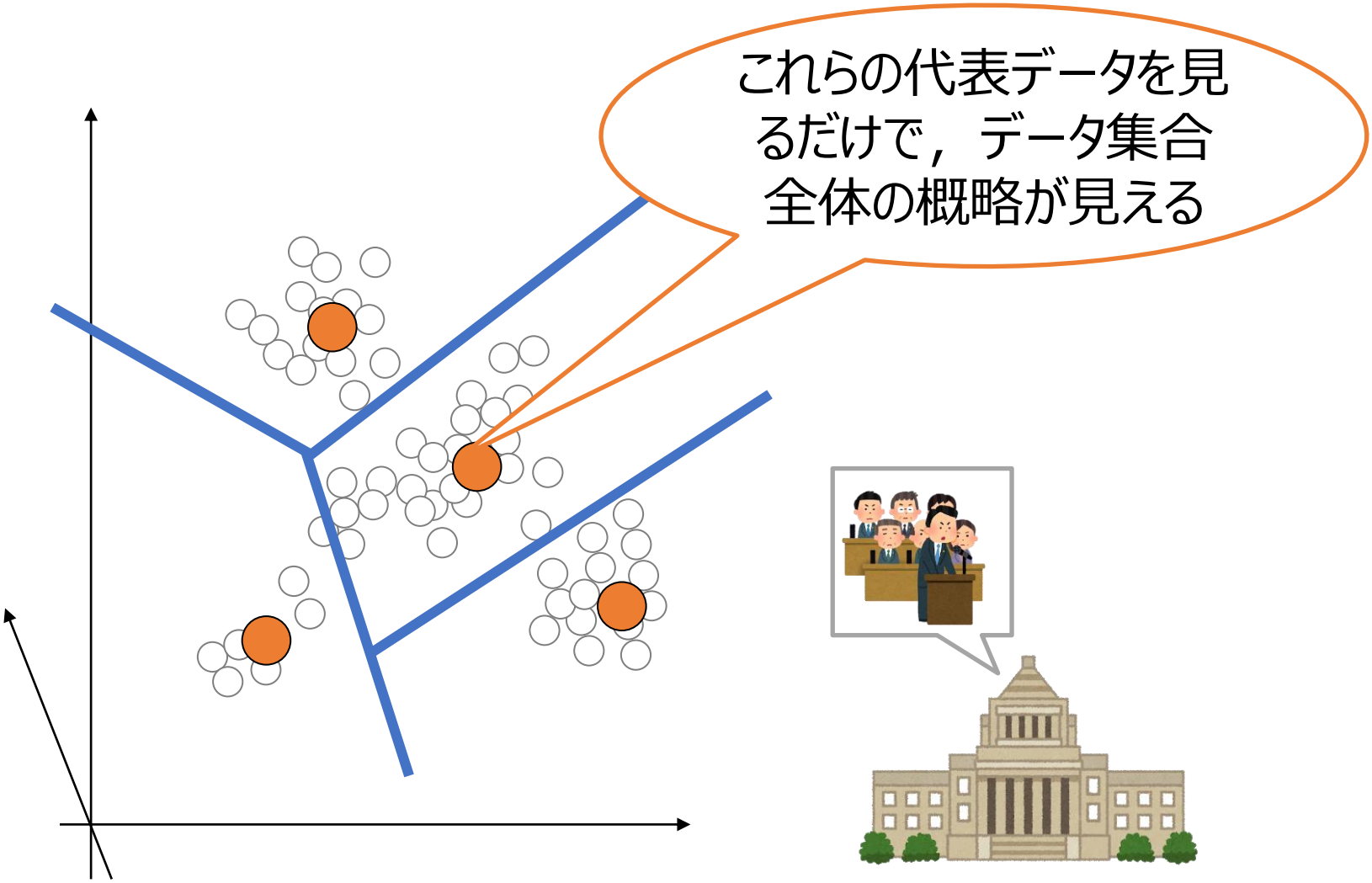
IASR

クラスタリング(clustering)＝
データの集合をいくつかの部分集合に分割する(グルーピング)

- 各部分集合＝「クラスタ」と呼ばれる

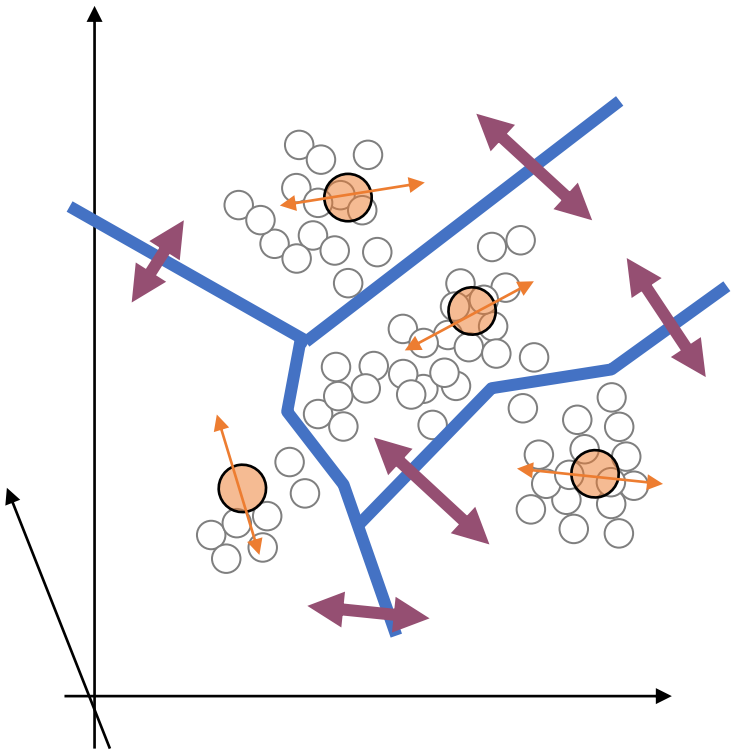


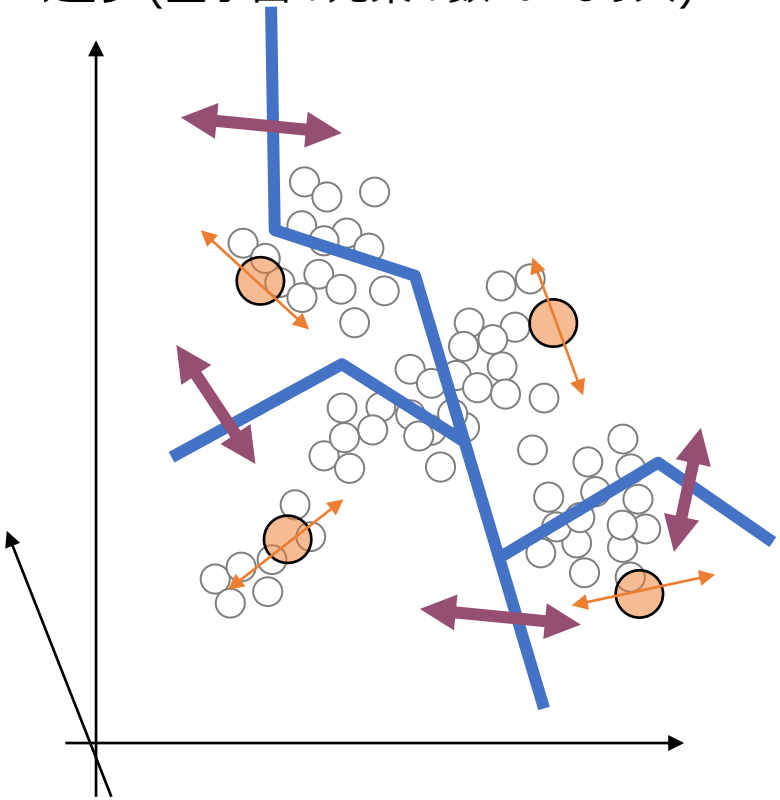
各クラスタから代表的なデータを選ぶと…



どういう分割がよいのか？

- N個のデータをK個に分割する方法はおよそK^N通り
  - 100個のデータを10分割→およそ10^{100}通り (全宇宙の元素の数10^{80}より大)





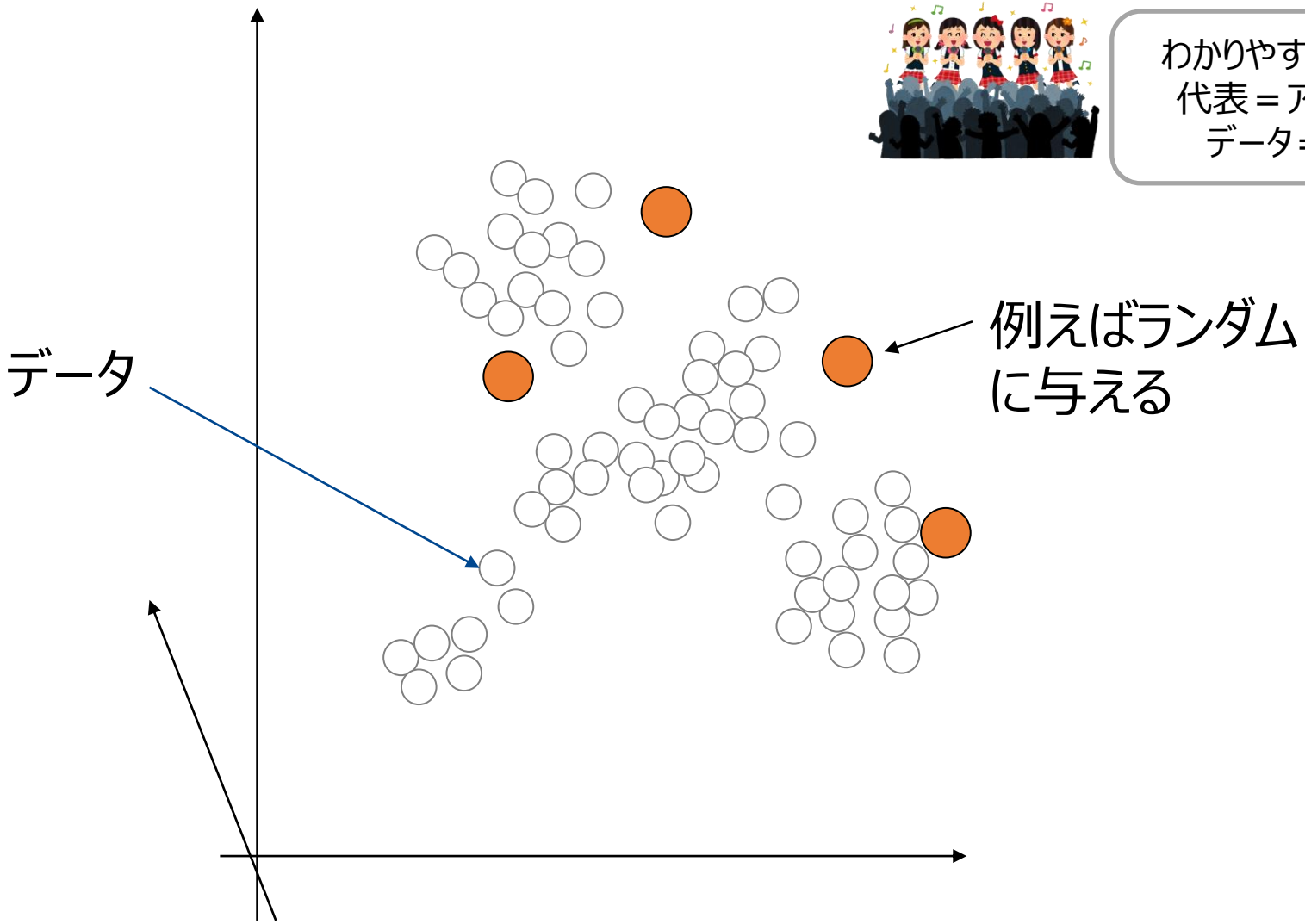
- 近くにあるデータが，なるべく同じグループになるように

データのクラスタリング②
k-means法

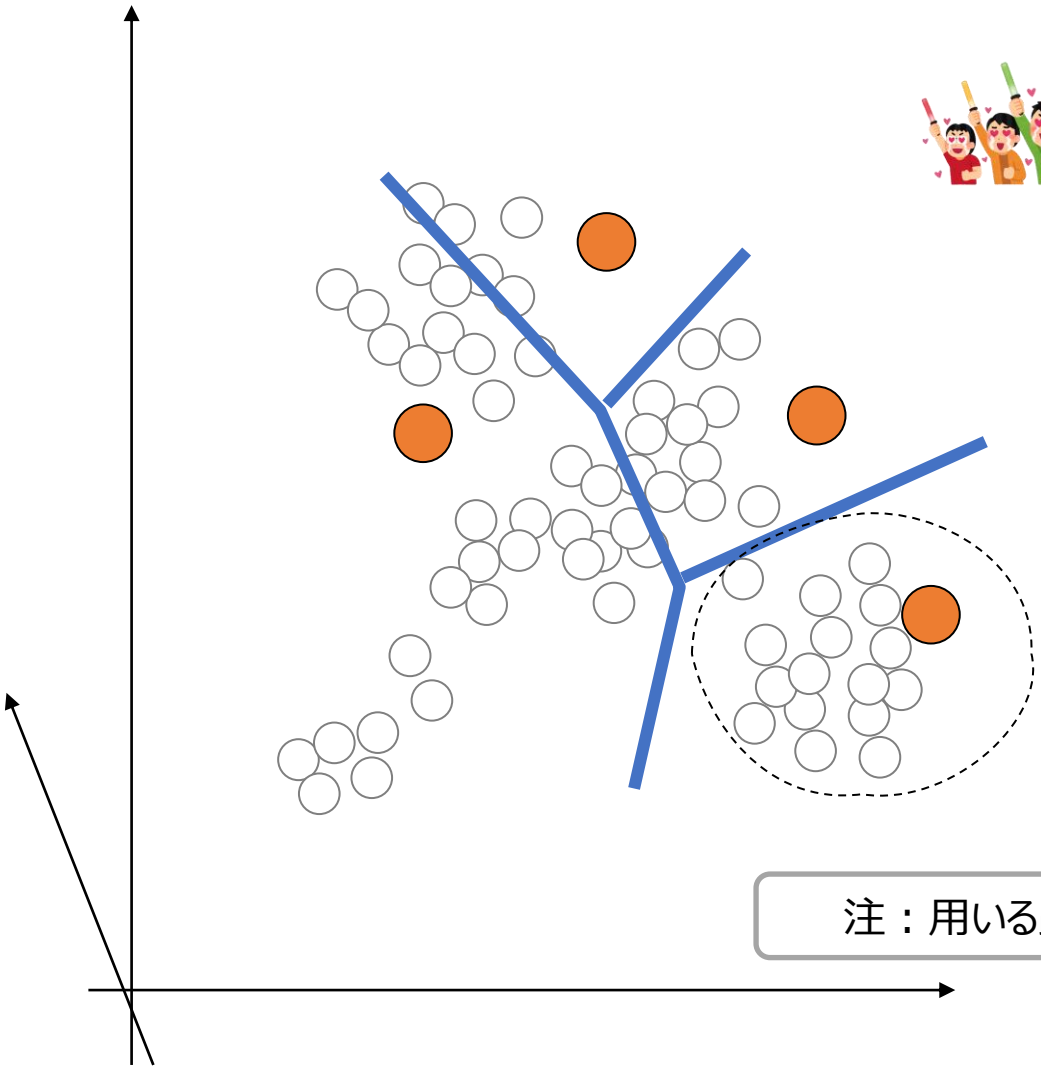
Mean = 平均．だから「平均をk個」求める方法．
古典的だが，いまでもクラスタリングの代表的方法（便利！）

k-means法（0）初期代表ベクトル

わかりやすい比喩：
代表＝アイドル,
データ＝民衆



k-means法（1）データの分割：
代表ベクトルとの距離でデータをわける

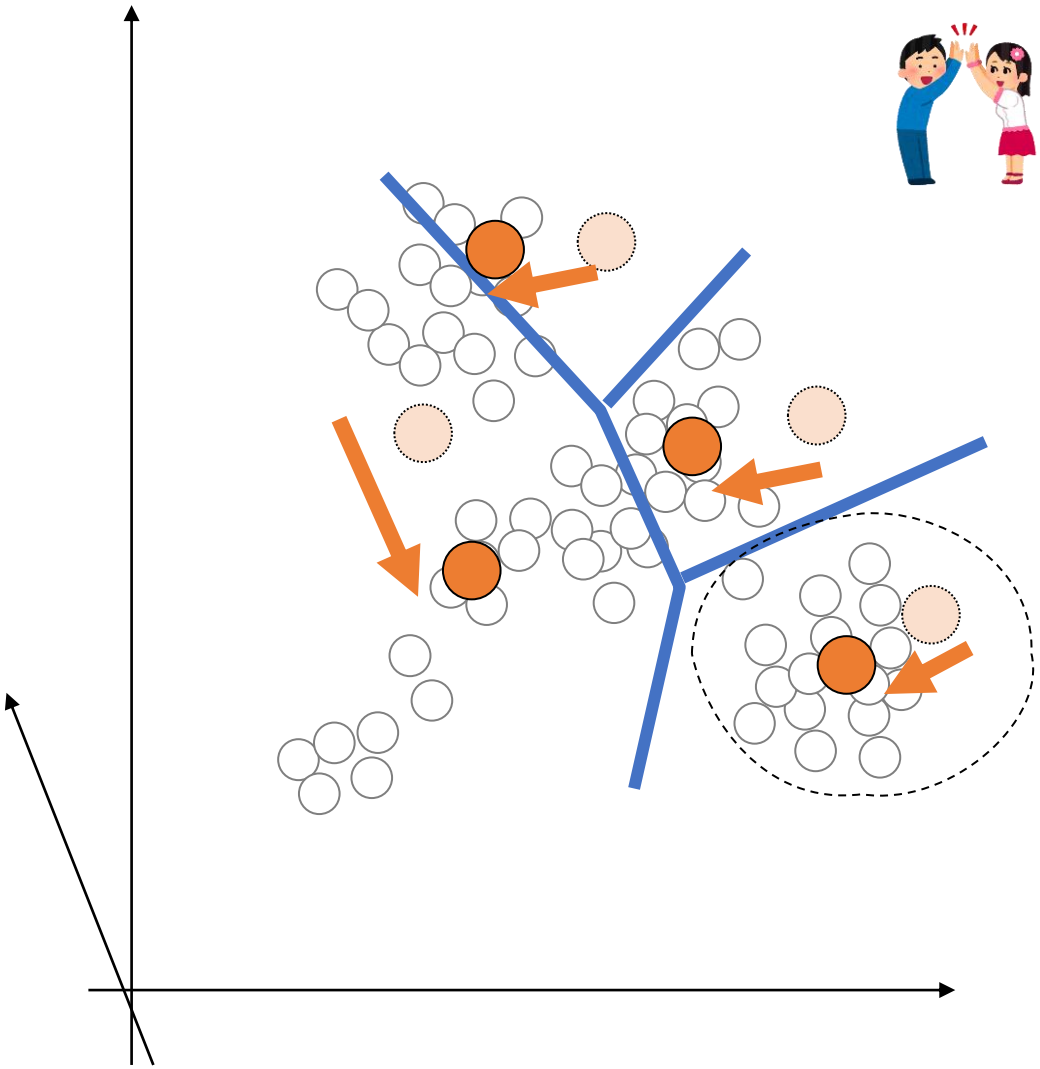


わかりやすい比喩：
民衆は自分に一番近い
アイドルのファンになる

代表ベクトルの
ファンクラブ
（支持者集合）

注：用いる距離によって分割は変わる

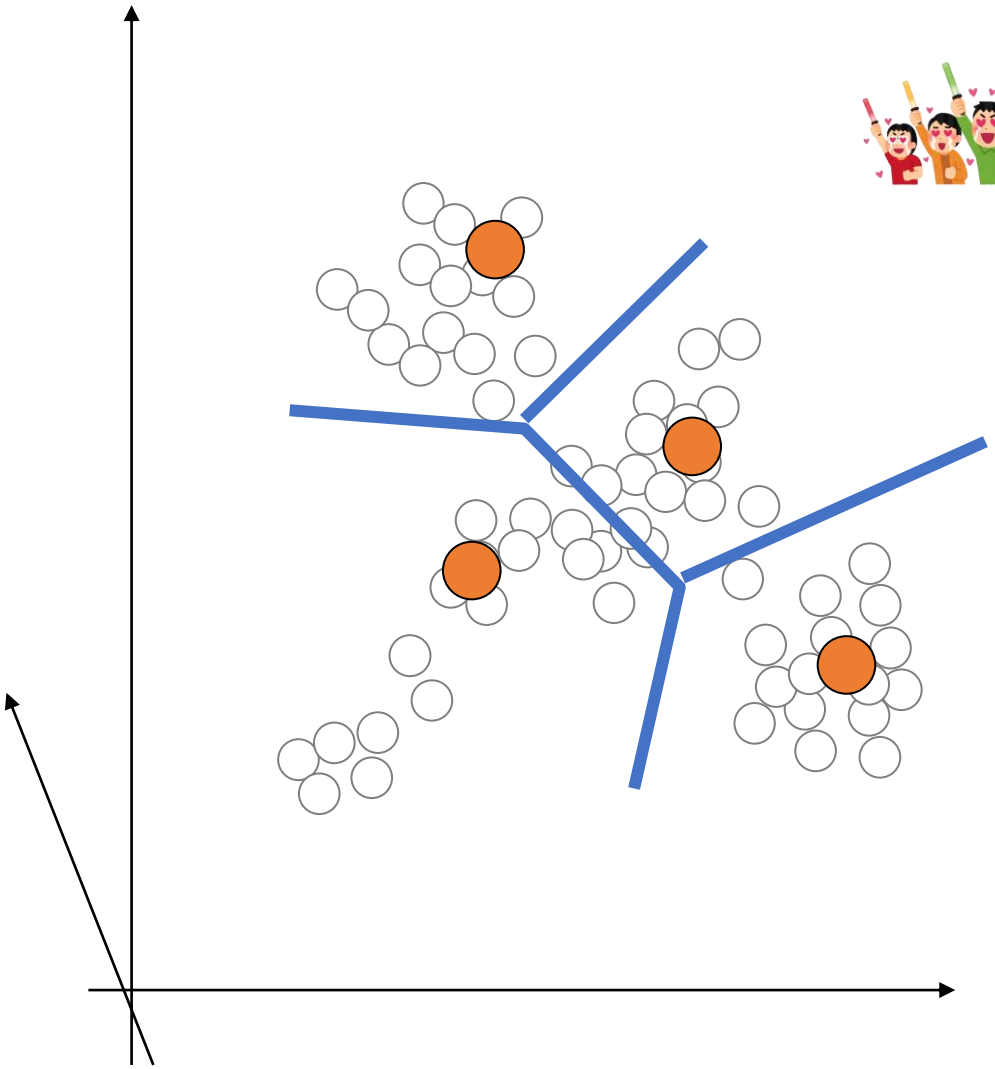
k-means法（2）代表ベクトル更新



わかりやすい比喩：
けなげなアイドルが
ファンクラブの真ん中に移動

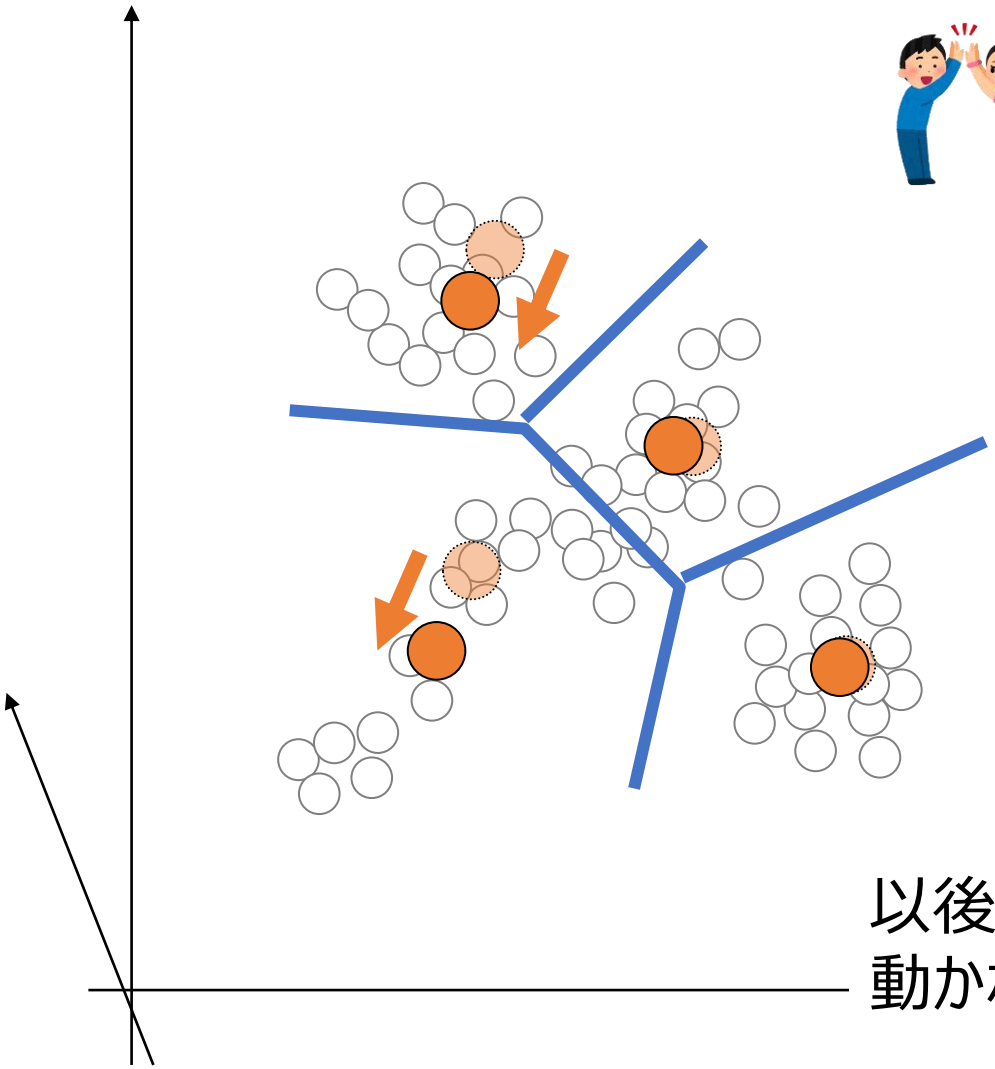
この中のデータの
平均に移動

k-means法（1）データ分割



わかりやすい比喩：
アイドルの移動により，
ファンクラブ構造が変ってしまう

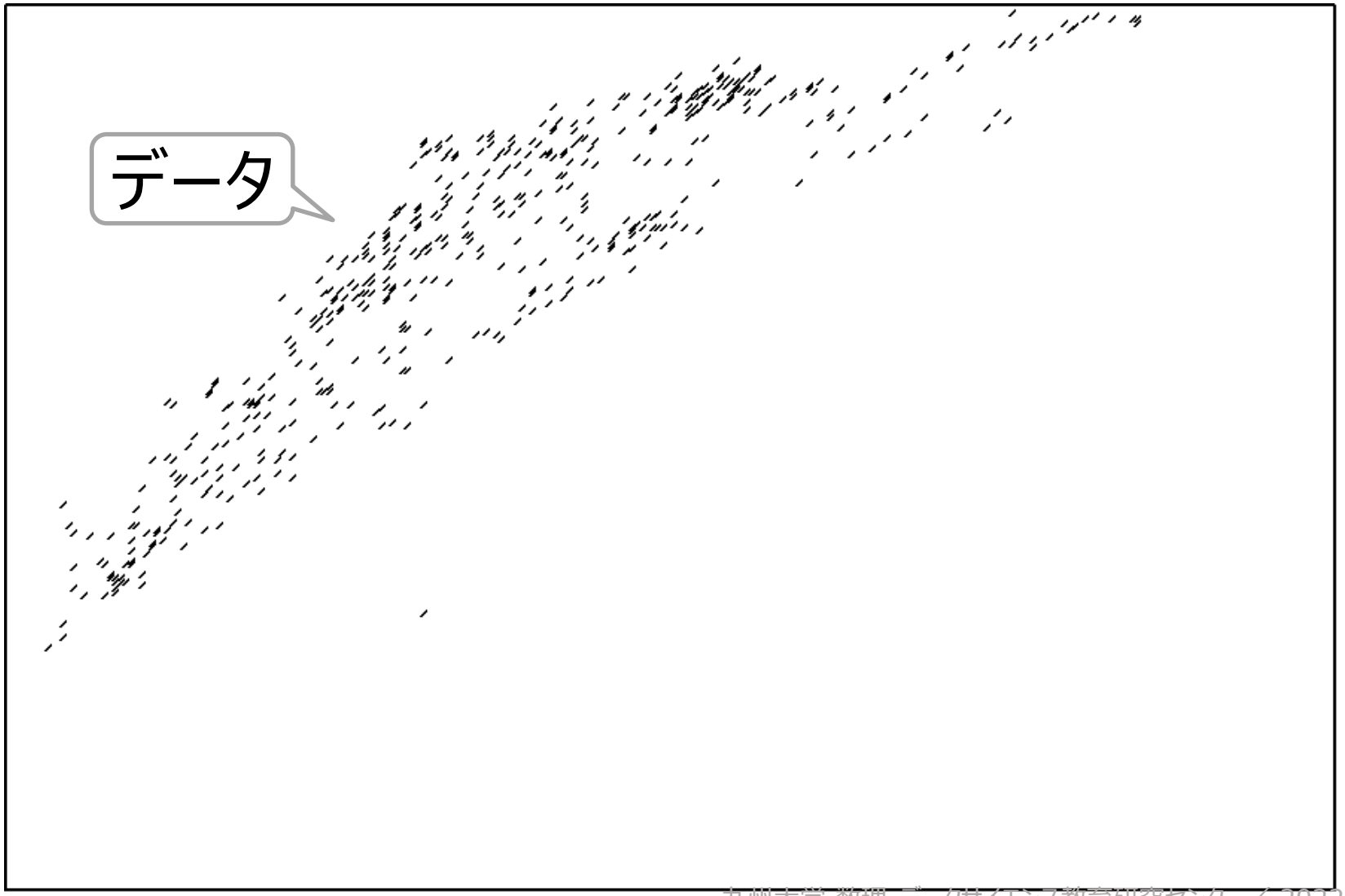
k-means法（2）代表ベクトル更新



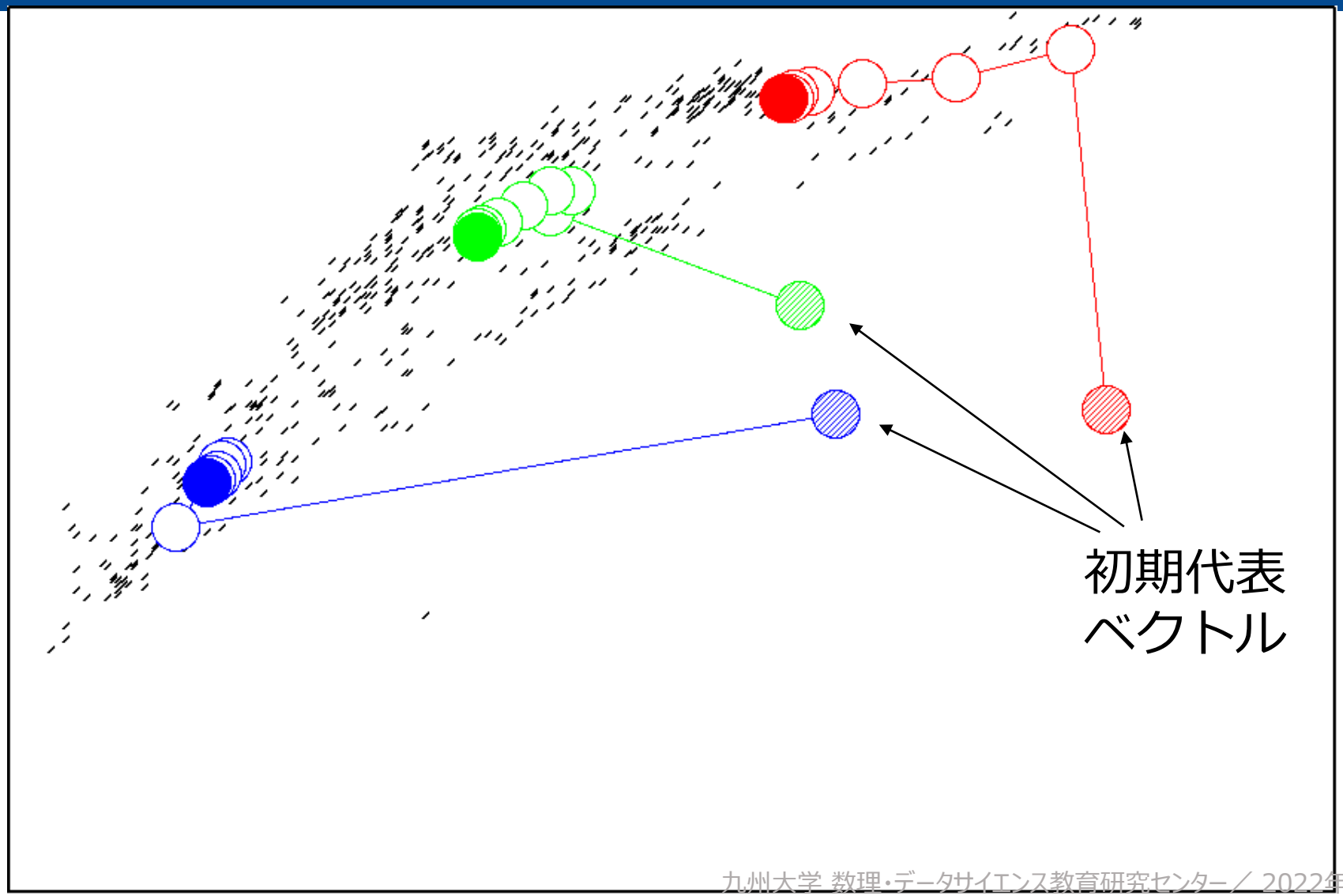
わかりやすい比喩：
けなげなアイドルは
新しいファンクラブの
真ん中に再び移動

以後，代表ベクトルが
動かなくなるまで反復

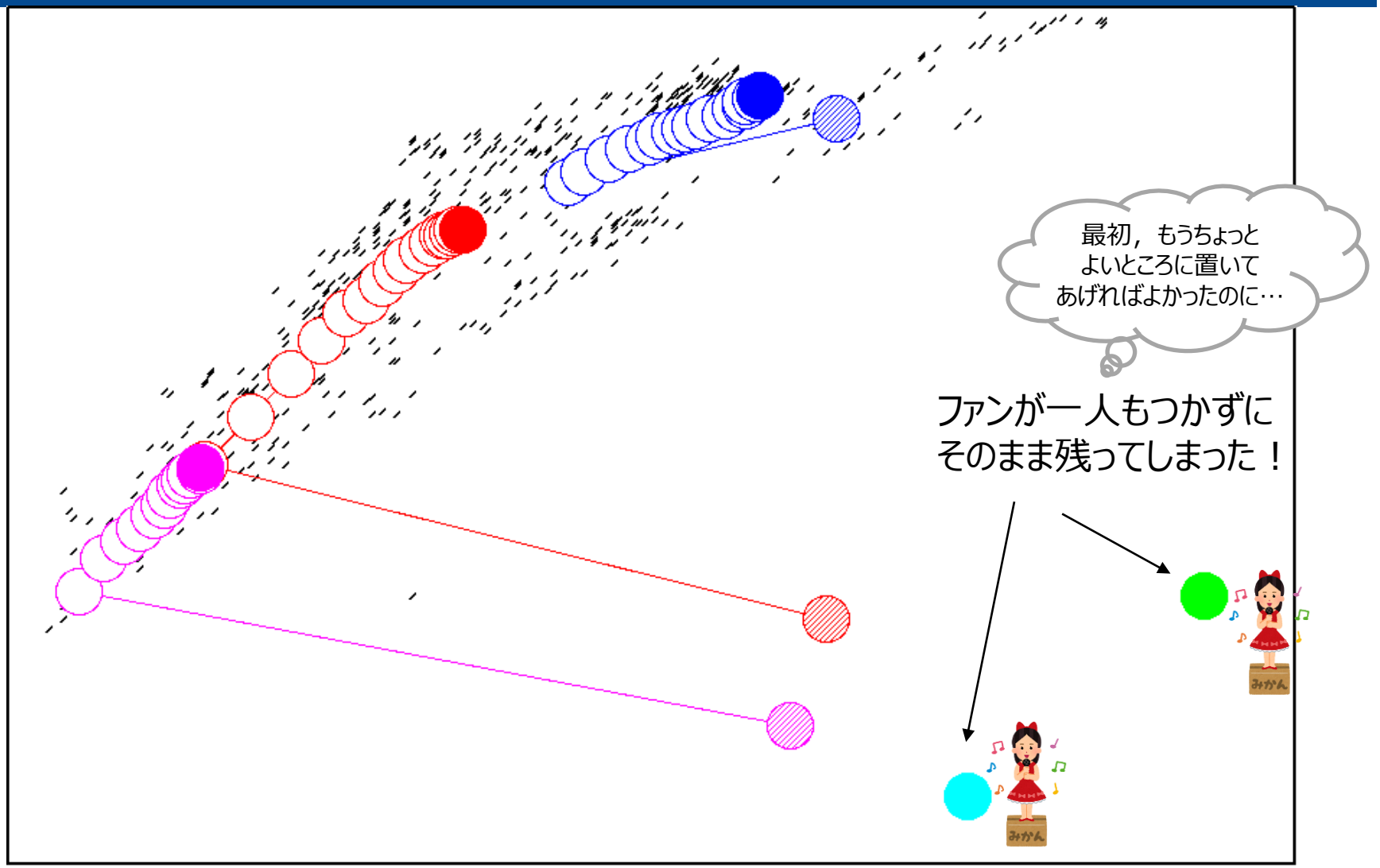
k-means法実例：データセット



k-means法実例：結果（$k = 3$）



k-means法実例：結果（$k = 5$）

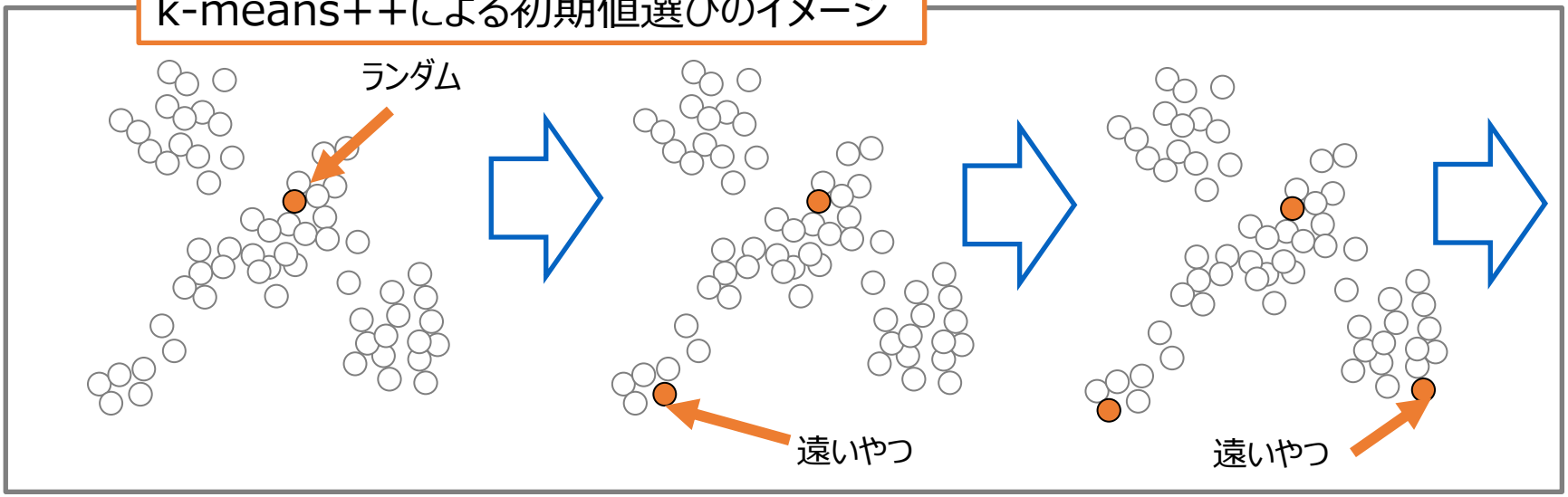


k-meansは「初期値が違うと結果が異なる」：それが困るなら…

- 「マルチスタート戦略」をやってみてもよい
  - 異なる初期値でP回k-meansを実施
  - P個の結果のなかで，最もよかった結果を選ぶ

- “k-means++”という，うまい初期値の決定法もあり

k-means++による初期値選びのイメージ



クラスタ数kの決め方も色々→付録

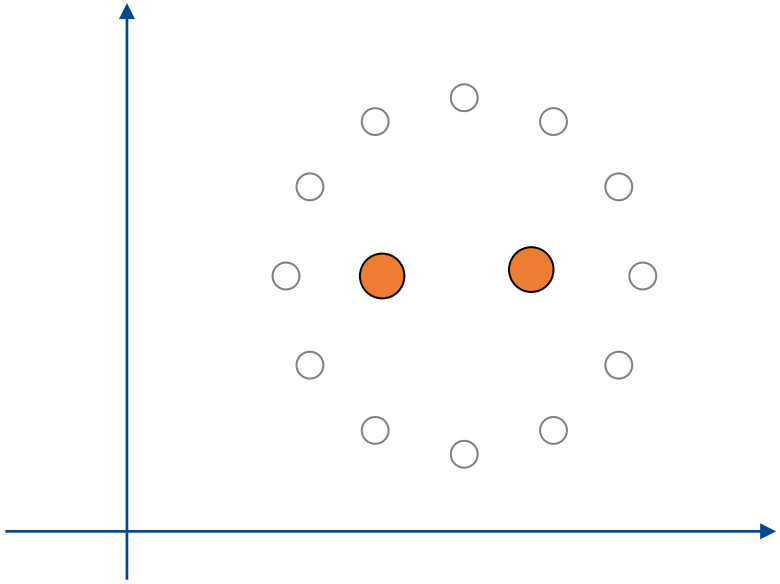
データのクラスタリング③

その他のクラスタリング法

まだ色々あります

k-means，もう一つの問題（？）

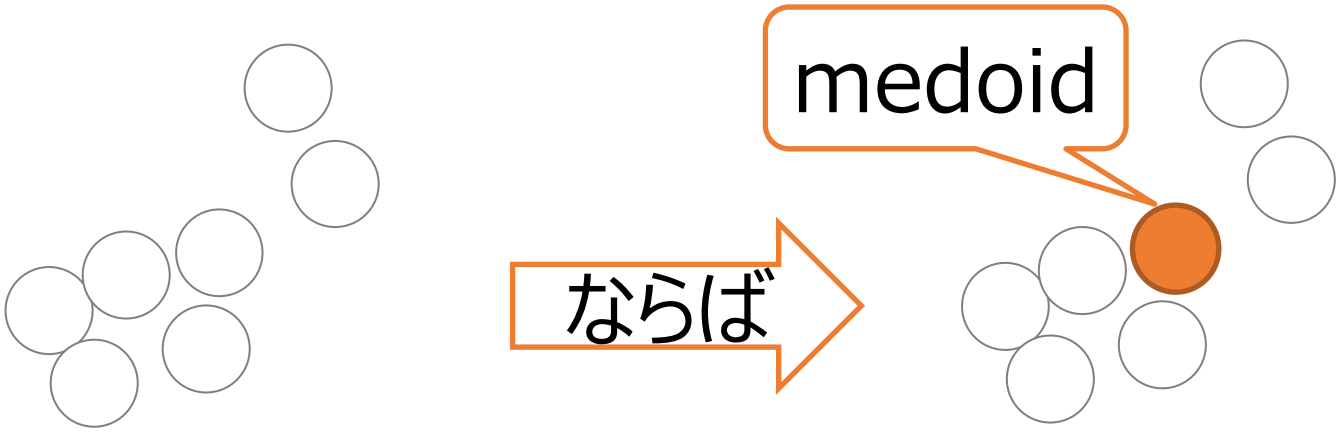
- 代表データは「平均」なので，原データの一つではない



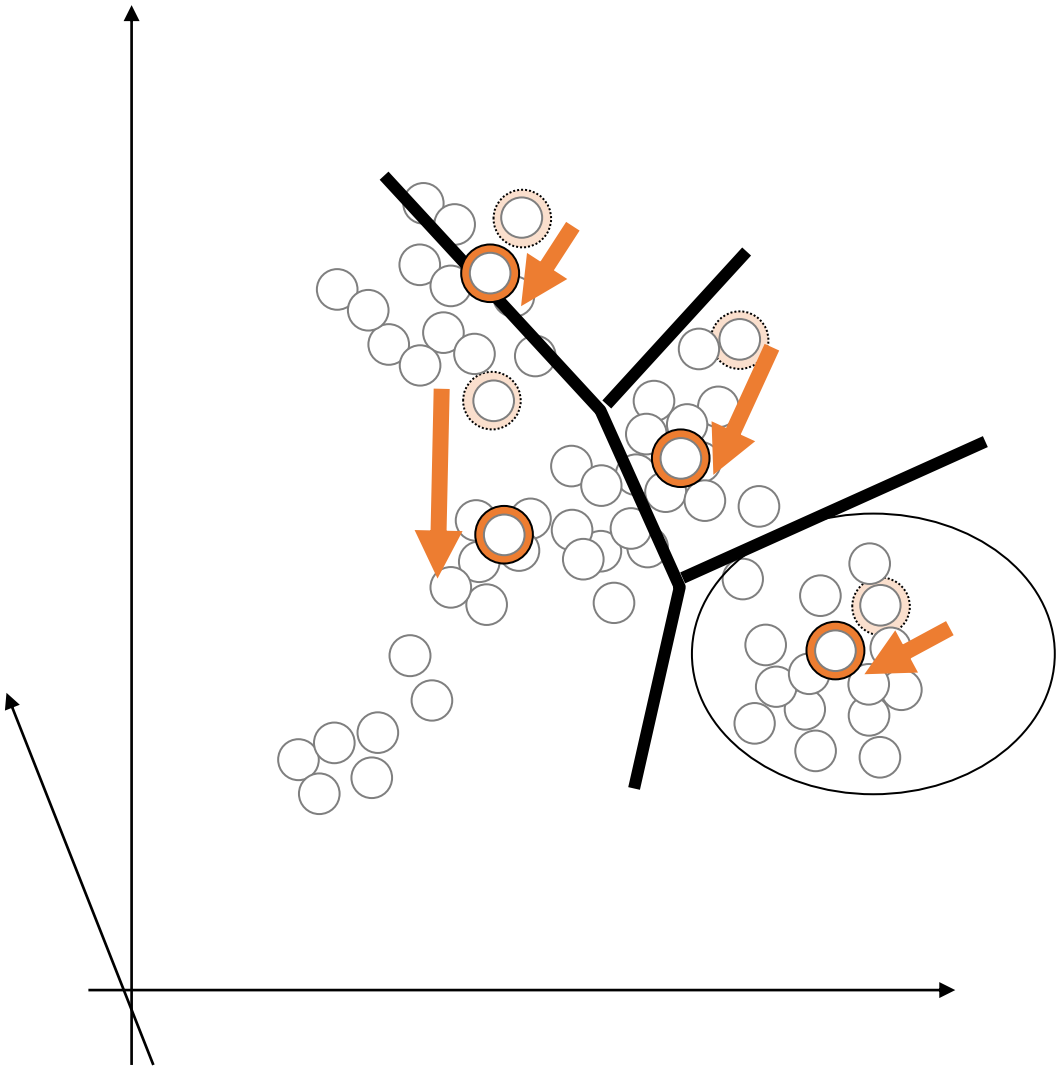
- 原データにあるものを「代表」としたい場合には…!?

k-medoid法

- Medoid ＝「クラスタ内のデータの中で他のデータとの距離を最も小さくできるもの」
  - 「平均」ではない.

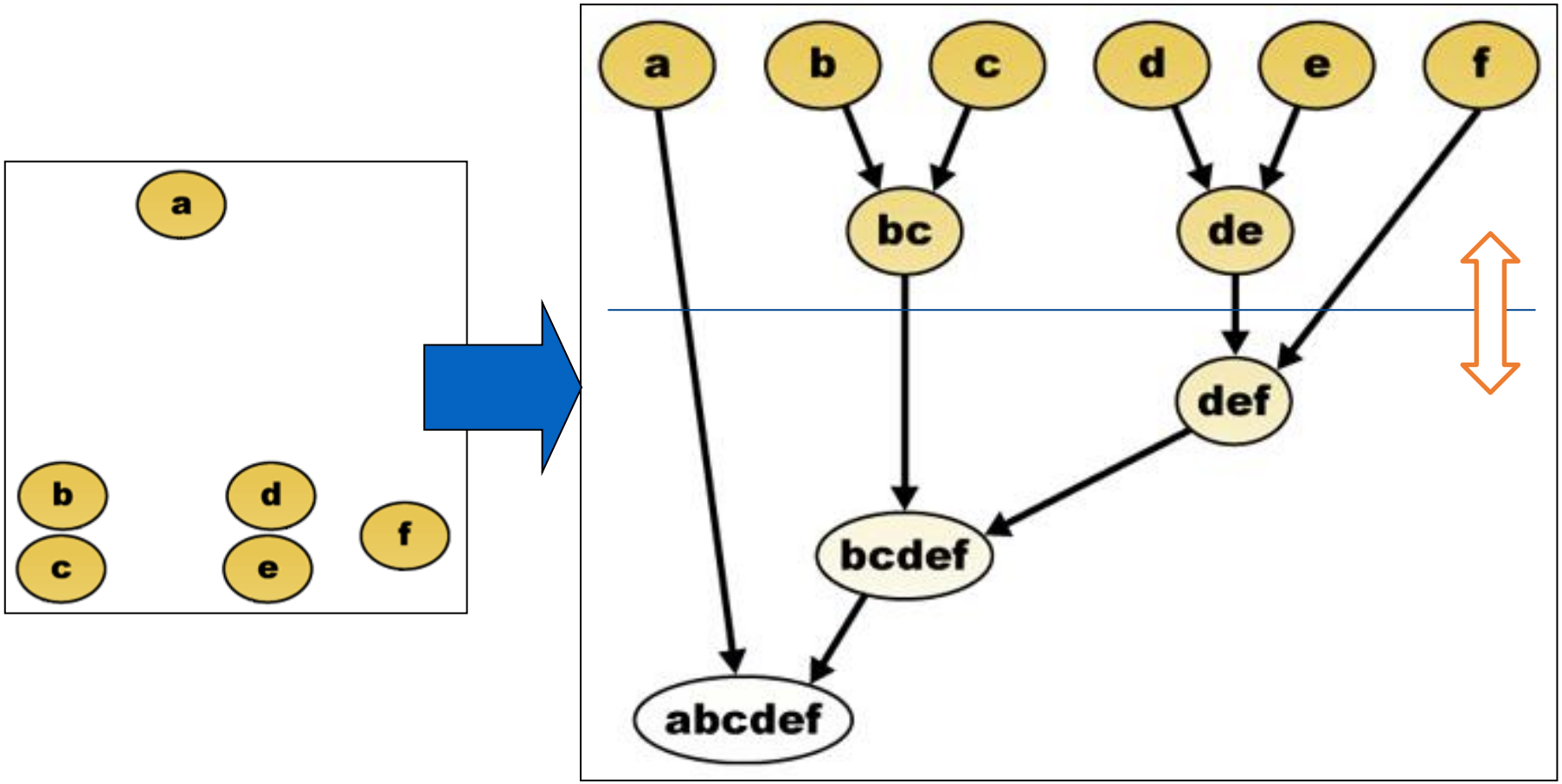


k-medoid法とk-meansの違いは一か所だけ！



この中のデータの
medoidに移動

もう一つの代表的クラスタリング法：階層的クラスタリング



wikipedia の図を一部引用

異常検出

距離さえあれば，異常なデータを見つけることができる！

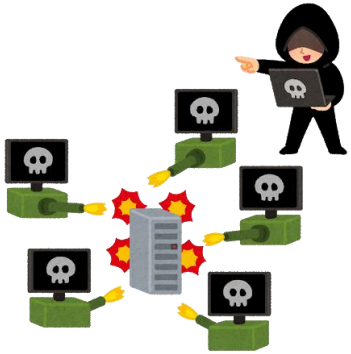
異常検出（異常検知）とは？

- 今与えられたデータが「一般的に期待していたデータ」とは異なるものであることを見出す手法

- なぜ異常は起こるか？
  - 機器やセンサの故障，身体の病気やケガ
  - うっかりや見落とし，事故や失敗など，人為的ミス
  - 侵入や破壊，悪用など，意図的な悪意のある行為
  - 甚大な自然災害など，想定外もしくは稀な現象の発生
  - Etc…

異常検出の応用例

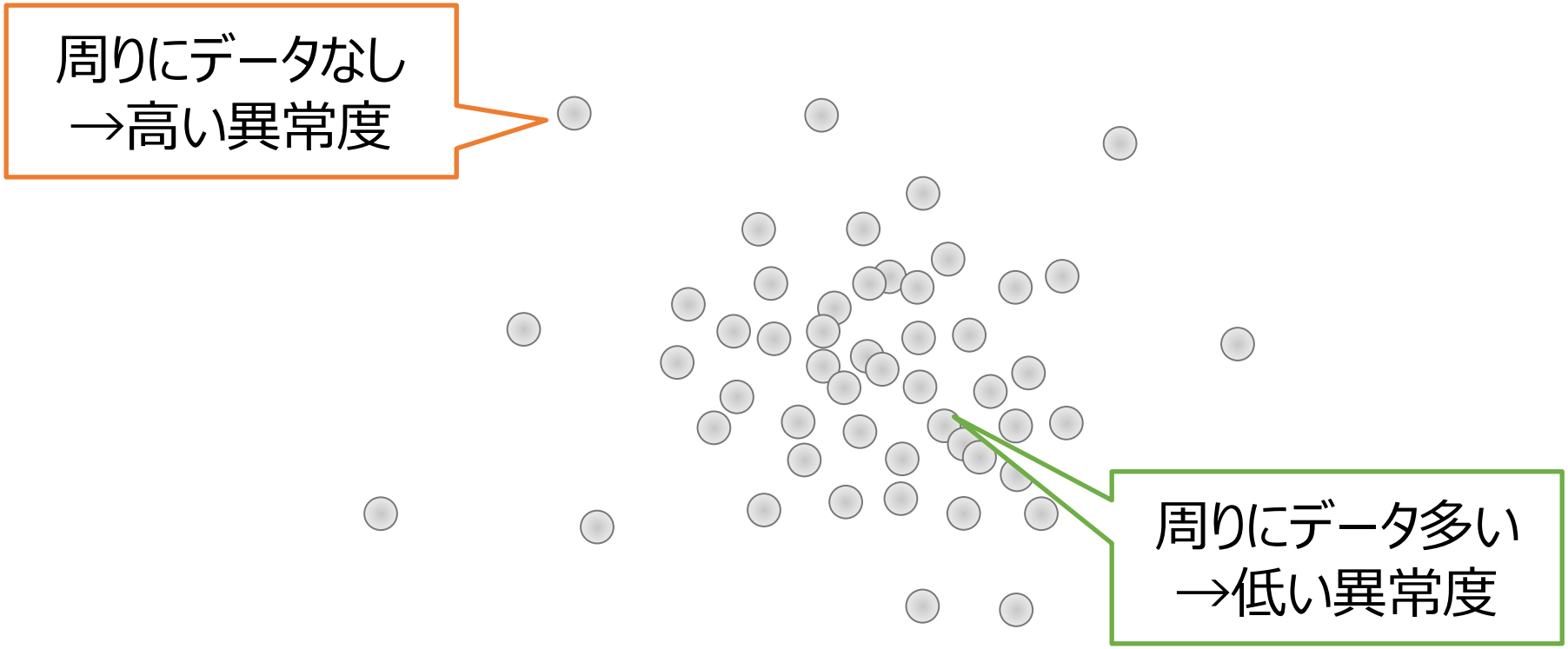
- 人々を対象とした異常検出
  - エレベータ内のカメラで人々の動きを認識し，さらに普通でない動きを判断
  - 独居者の異常検出：日常と異なる行動パターンを通知し通知
  - 病院内での患者の異常検出：特に気づきにくい早期の異常検出
- 食品や生産物の異常検出
  - 表面の傷の検出，異物の検出
- 機械・建造物・コンピュータシステムの異常検出
  - サイバー攻撃の検出
  - 機械の故障の検出
  - 橋や道の異常検出



※そのうちやる「分類・パターン認識」とも関係

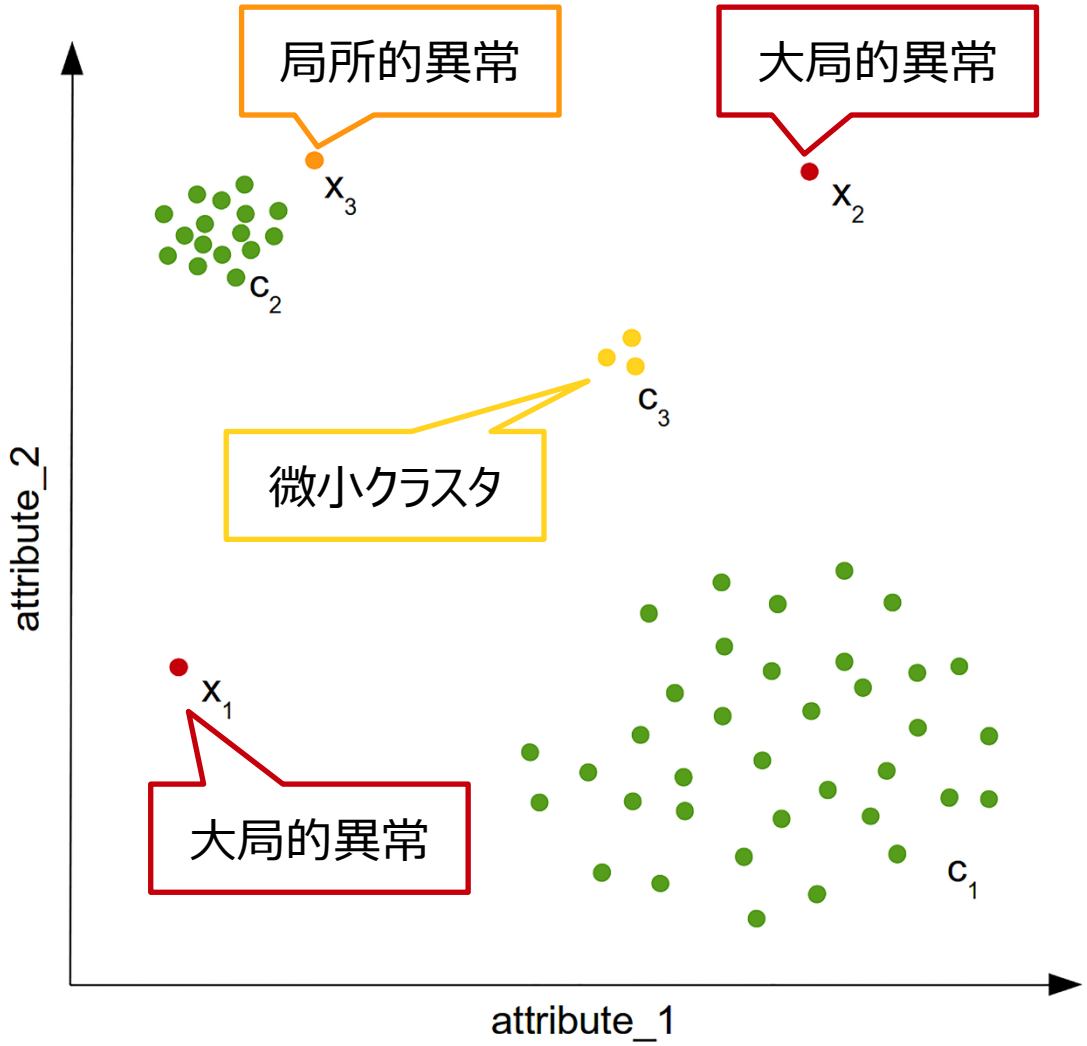
距離による異常検出の基本的な考え方

- 「注目しているデータが，他のデータから離れている(距離が遠い)」→異常度が高い



異常の種類(1/3)

- 大局的異常
  - まわりに「近い」データがない
    「似た」データがない
    ＝どう考えても異常
  - x_1, x_2

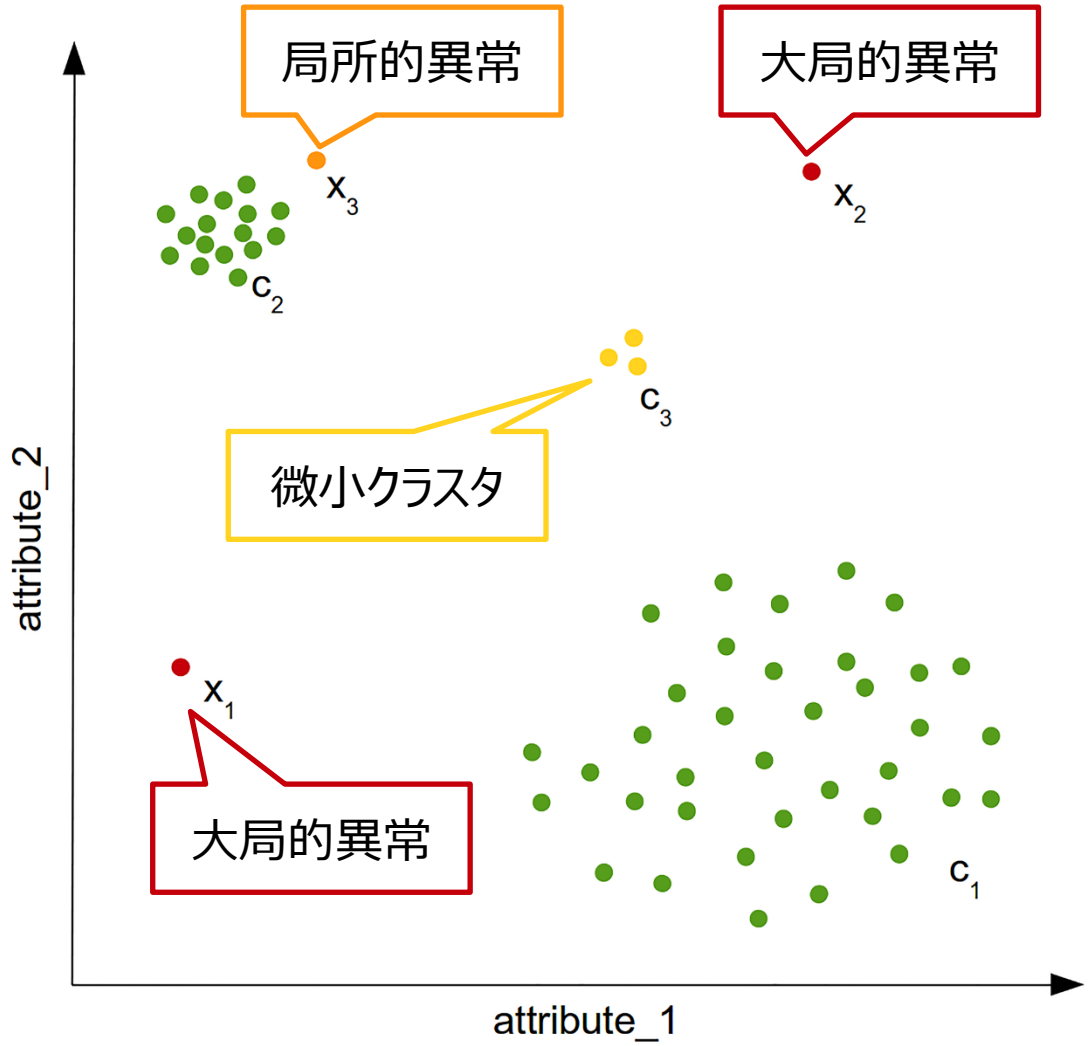


[Goldstein, Uchida, PLoSONE, 2016]

異常の種類(2/3)

- 局所的異常
  - そのデータ周辺の平均的な近さで考えると，「近くない」
  - x_3は局所的異常
    - 付近にあるc_2グループ基準でみると異常
    - ただしc_1グループ基準だとそこまで遠くはない

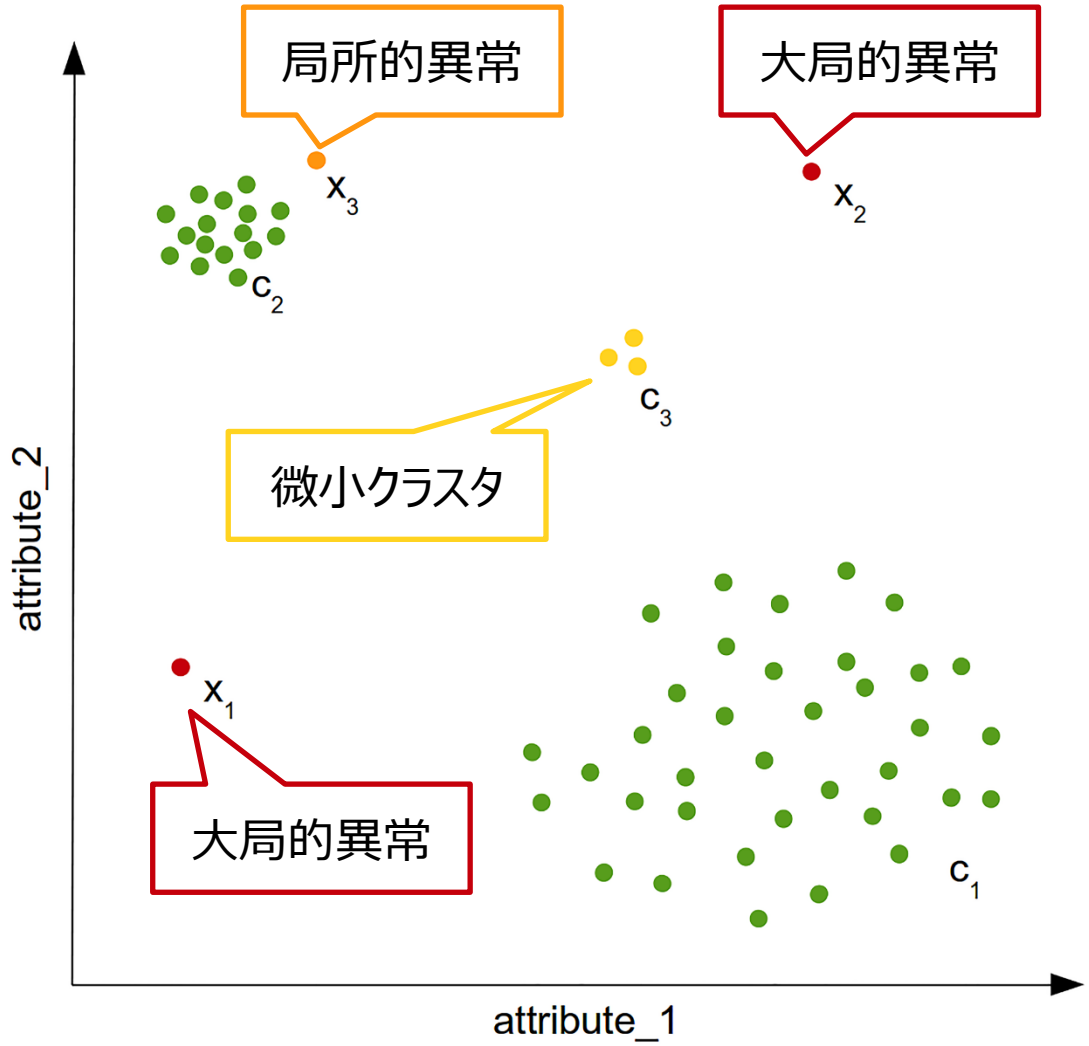
わかりやすい比喩：
その家がある都会では
「町はずれの一軒家」だが，
田舎視点だと十分ご近所



[Goldstein, Uchida, PLoSONE, 2016]

異常の種類(3/3)

- 微小クラスタ
  - 周りに近いデータはあるものの，その数は限定的
  - c_3 の3つのデータが相当



わかりやすい比喩：
要は「オタク」集団(?).
近くに同志は少数いるが，
世間全体からは浮いている

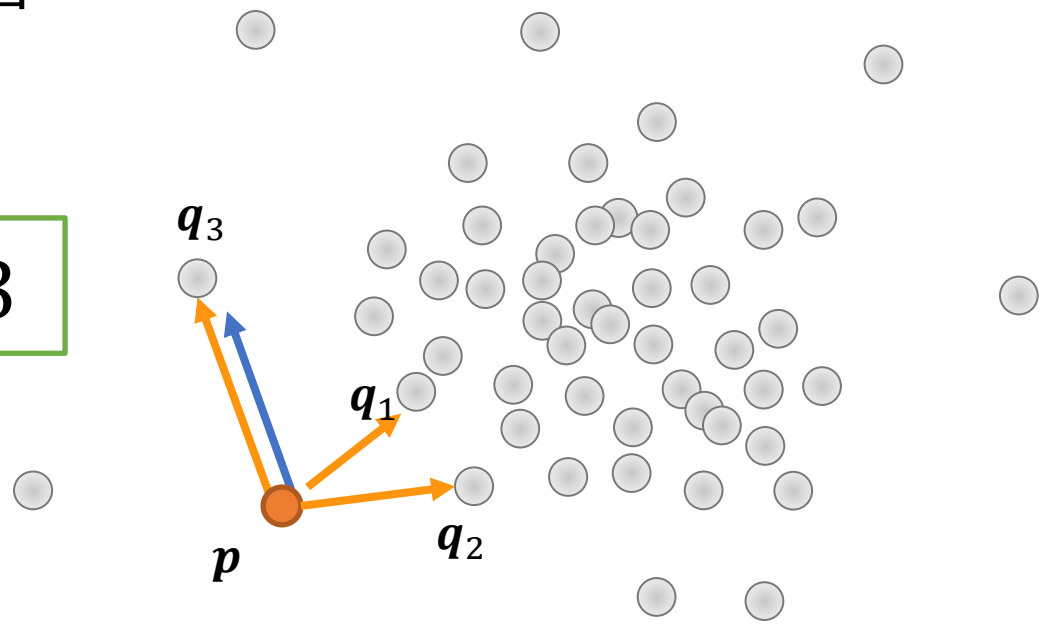
[Goldstein, Uchida, PLoSONE, 2016]

k近傍法による異常検出(1/3) 原理

k個の近いデータ

- 単一タイプ：k番目に近いデータへの距離 $\|\boldsymbol{p}-\boldsymbol{q}_k\|$ or
- 合計タイプ：$\|\boldsymbol{p}-\boldsymbol{q}_1\|+\|\boldsymbol{p}-\boldsymbol{q}_2\|+\cdots+\|\boldsymbol{p}-\boldsymbol{q}_k\|$
- 大局的異常検出

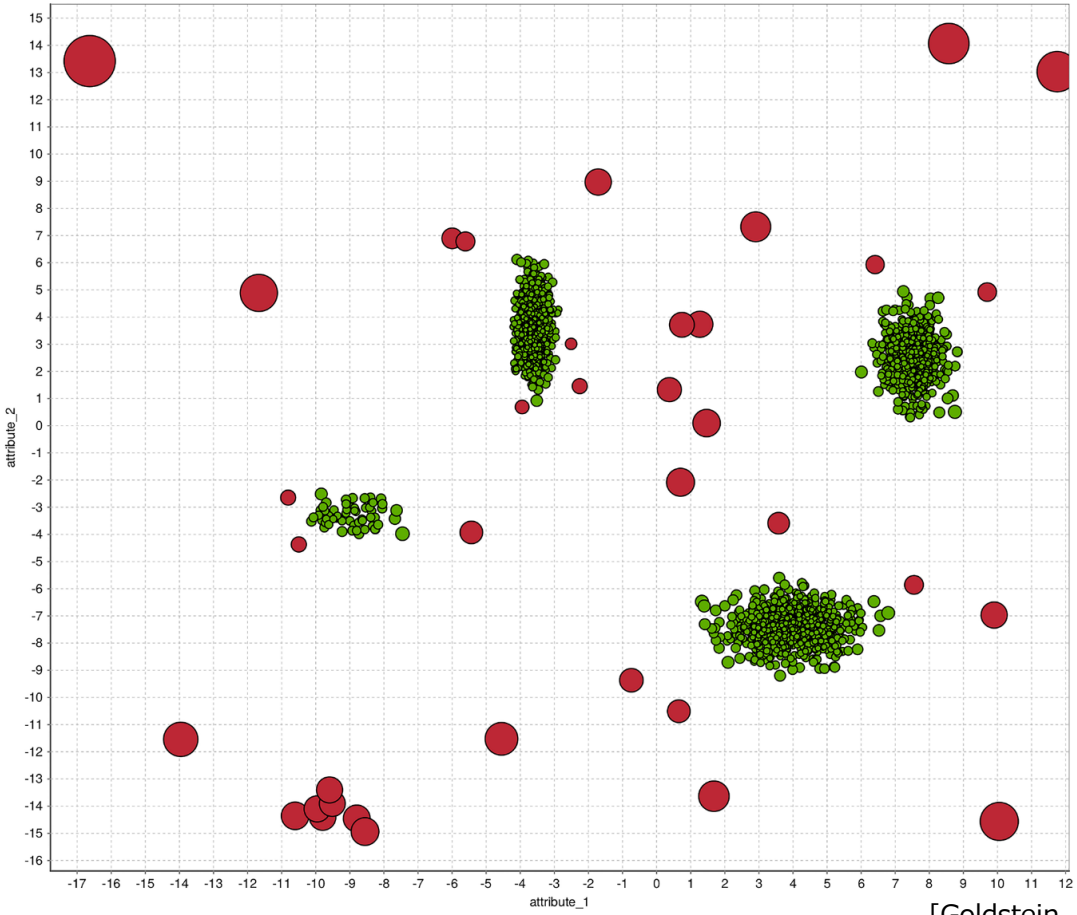
$k=3$



- $k-1$個のデータからなる微小クラスタも異常として検出可能

k近傍法による異常検出(2/3)
異常度計算結果例($k = 10$, 合計)

- 半径が大きな点(データ)ほど異常度が高いと計算されている



[Goldstein, Uchida, PLoSONE, 2016]

k近傍法による異常検出(3/3)
問題点

- 「ぎっしりクラスタ」と「すかすかクラスタ」が共存していると…

「都会基準で田舎を全部異常」と見ていいのか？
「都会は都会基準」「田舎は田舎基準」で見たいところ
(要は局所異常の検出が苦手→LOF法(付録)を使おう)

【付録１】
クラスタ数kはどうやって決めるか？

「絶対これ！」という正解はないのがポイント.
「このクラス全体をグループにわけろ」と言われても,
グループ数は色々考えられるのと同じ

k-means，代表パターン数 k は？

- これは中々難しい

Wikipedia “Determining the number of clusters in a data set”

- 「えいや！」と決める
- kを増やしながら誤差や認識率を測る
  - 飽和したらやめる

Contents [hide]

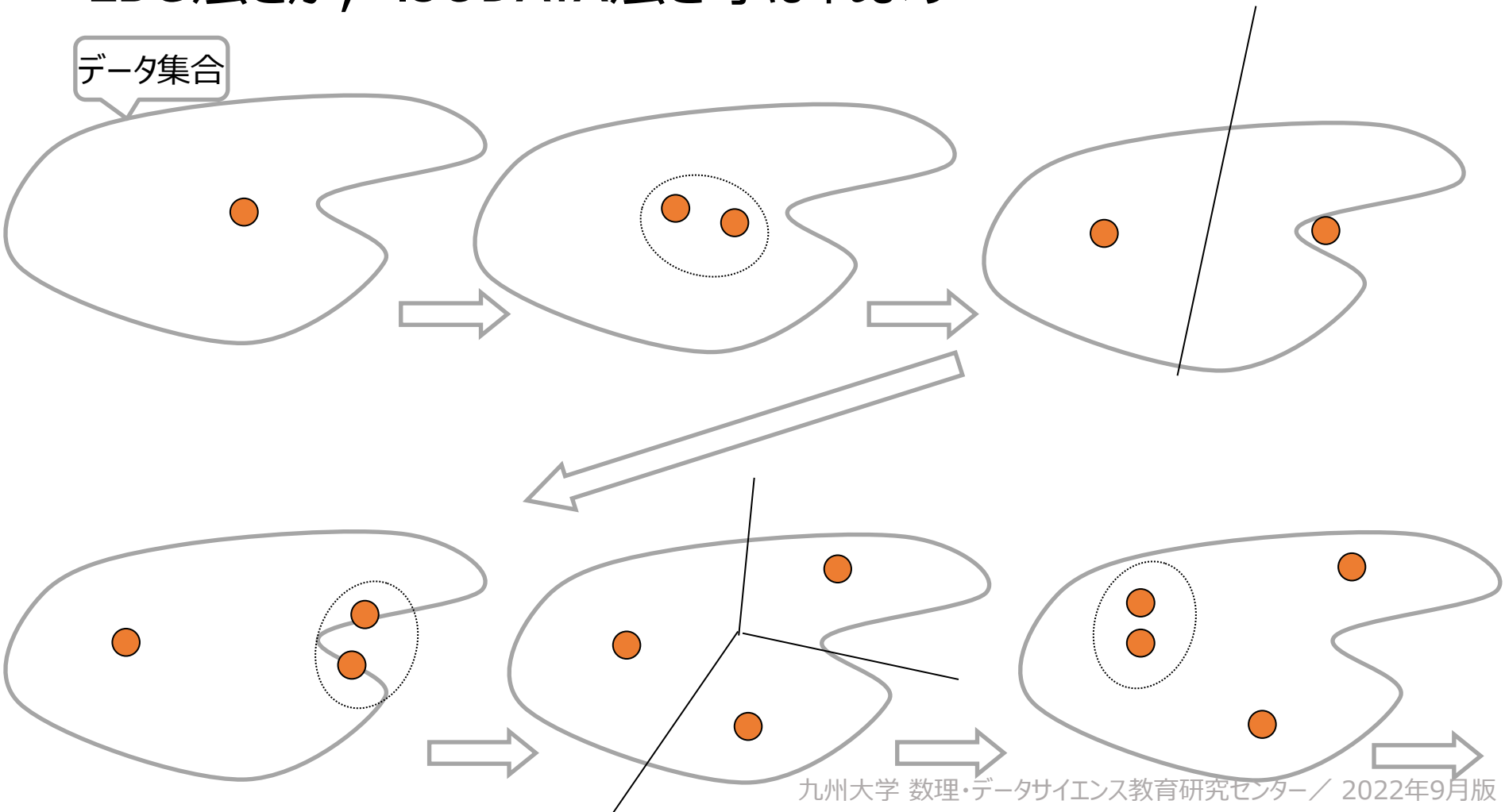
1 The elbow method
2 X-means clustering
3 Information criterion approach
4 An information–theoretic approach
5 The silhouette method
6 Cross-validation
7 Finding number of clusters in text databases
8 Analyzing the kernel matrix
9 Bibliography
10 External links

- LBG/ISODATAやx-meansを使う
  - 次スライド

- Calinski-Harabasz基準のような，クラスタリングの精度評価指標を利用する

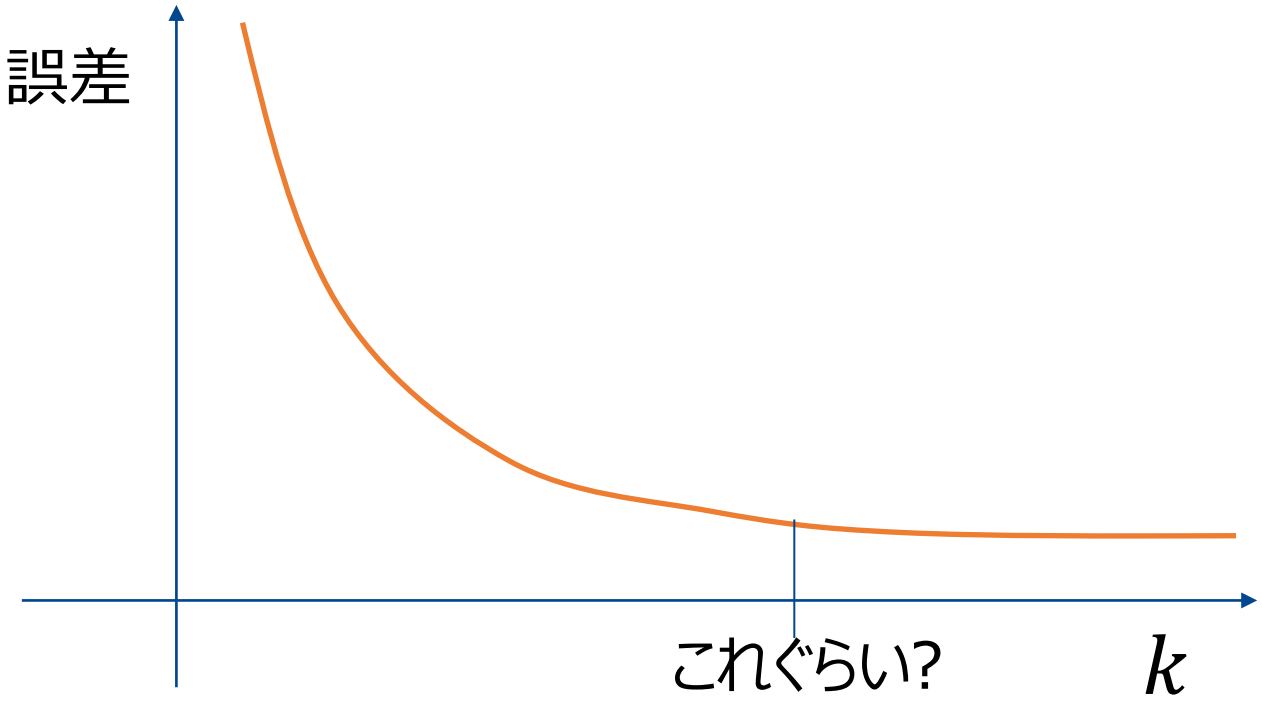
kをどうやって決めるか？(1/4)
徐々にkを増やすアルゴリズム

- LBG法とか，ISODATA法と呼ばれます



kをどうやって決めるか？(2/4)
とにかくkを変えながら，誤差を評価

- 例えば「誤差＝代表ベクトルとの平均的距離」と定義して，
- kを徐々に変えながらその誤差を測って，



- あまり減らなくなったところで「えいや！」と決める

kをどうやって決めるか？(3/4) x-means

k-meansを実行

各クラスタで
2-means

BICを測る
(ラフに言えばクラスタ数と
精度のバランスを評価)

2-meansにより
BICが改善されて
るのなら採用

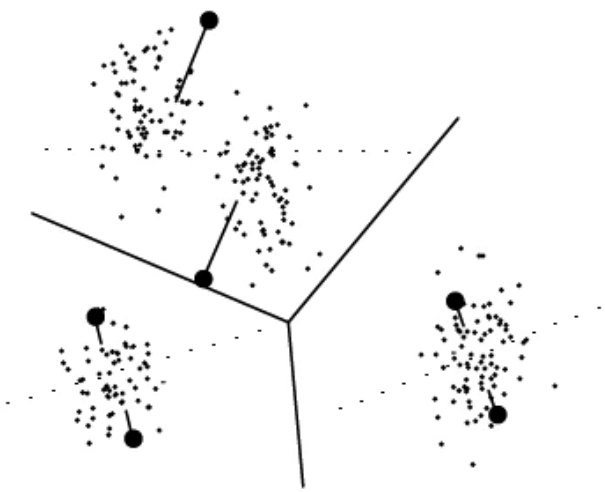


Figure 3: The first step of parallel local 2-means. The line coming out of each centroid shows where it moves to.

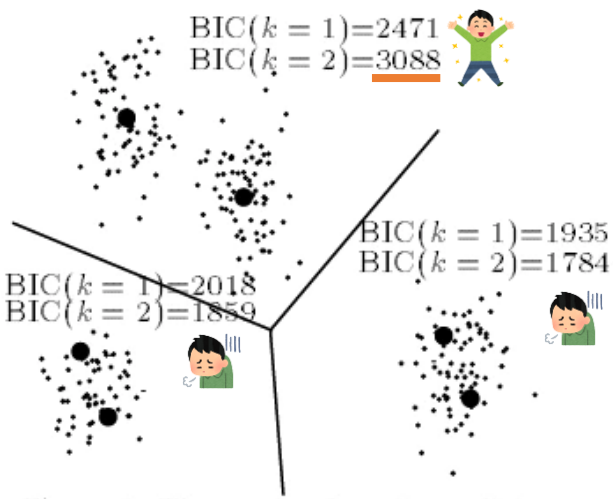


Figure 4: The result after all parallel 2-means have terminated.

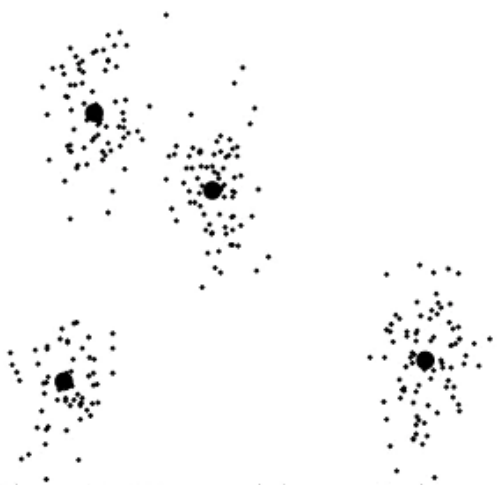


Figure 5: The surviving centroids after all the local model scoring tests.

kをどうやって決めるか？(4/4) Calinski-Harabasz基準（Pseudo Fとも）

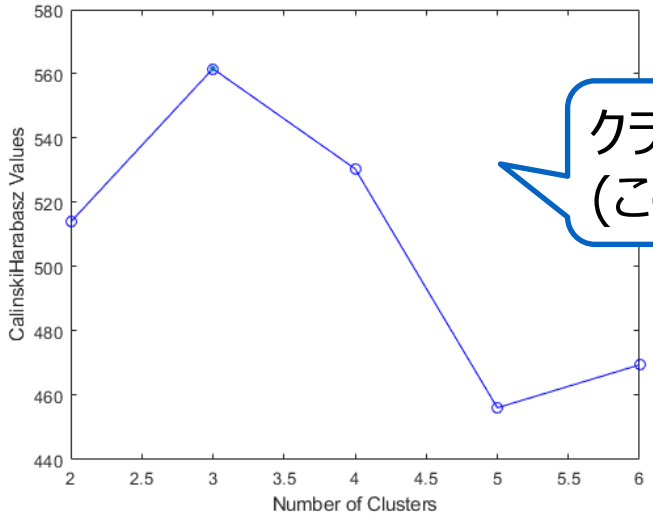
- 定義

代表パターンがどれぐらい散らばっているか

$$\frac{\text{クラスタ間分散}}{\text{クラスタ内分散}} \cdot \frac{N-k}{k-1}$$

各クラスタ内でサンプルがどれぐらいまとまっているか

- 大きいほどよい！



クラスタ数3の場合が(この基準では)最良

https://jp.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html

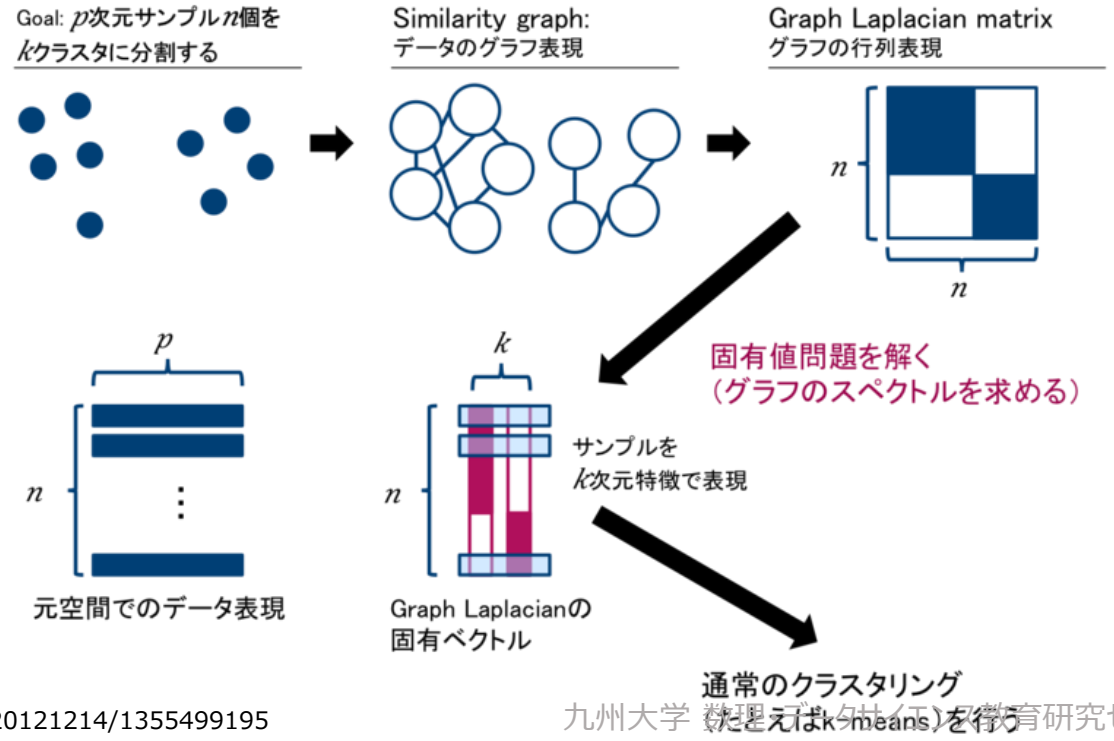
【付録２】
スペクトラルクラスタリング

グラフ＝近いものどうしが手を結んでできるネットワークのこと

スペクトラル・クラスタリング (1/2)

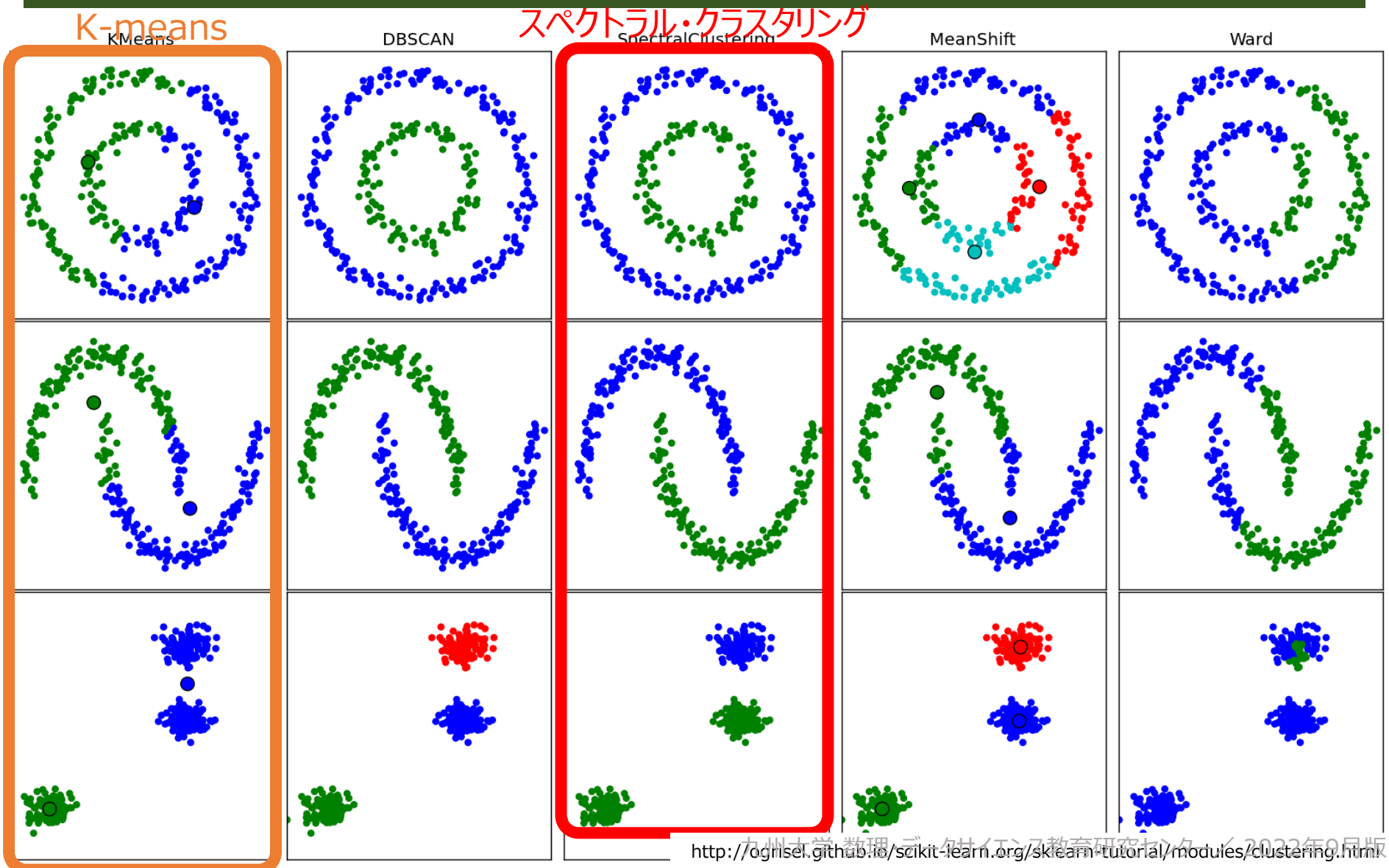
- グラフ表現されたデータ集合のクラスタリング
  - 近いデータ間をつないでグラフ構築
  - その近接性を極力保存したまま低次元空間に写像
  - 低次元空間でk-means

「近さ」には任意性あり



http://d.hatena.ne.jp/mr_r_i_c_e/20121214/1355499195

スペクトラル・クラスタリング (2/2)



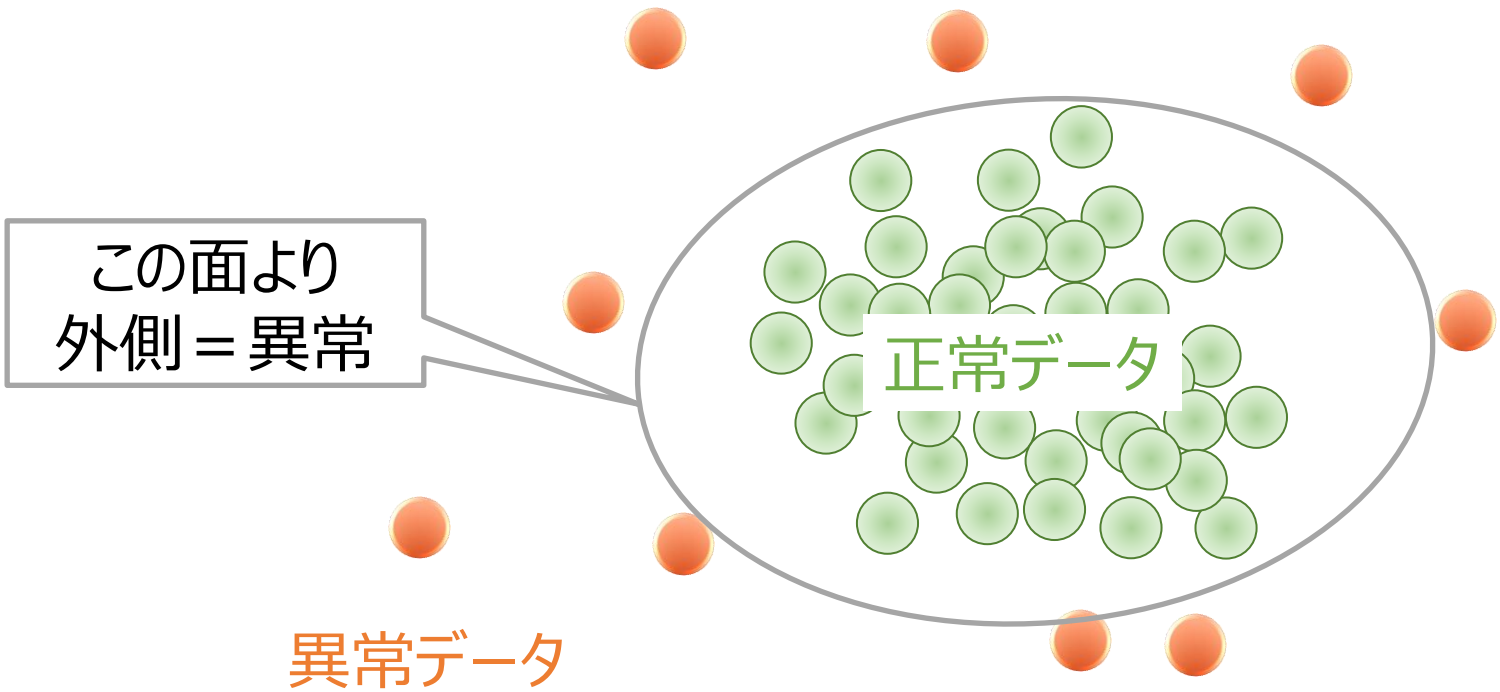
http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/modules/clustering.html

【付録３】
異常検出，３つのシナリオ

本編では，シナリオ３「単にデータの集合しか与えられない」状況を
説明しました

異常検出，３つのシナリオ(1/3)：「正常データ」と「異常データ」が共に準備できる場合

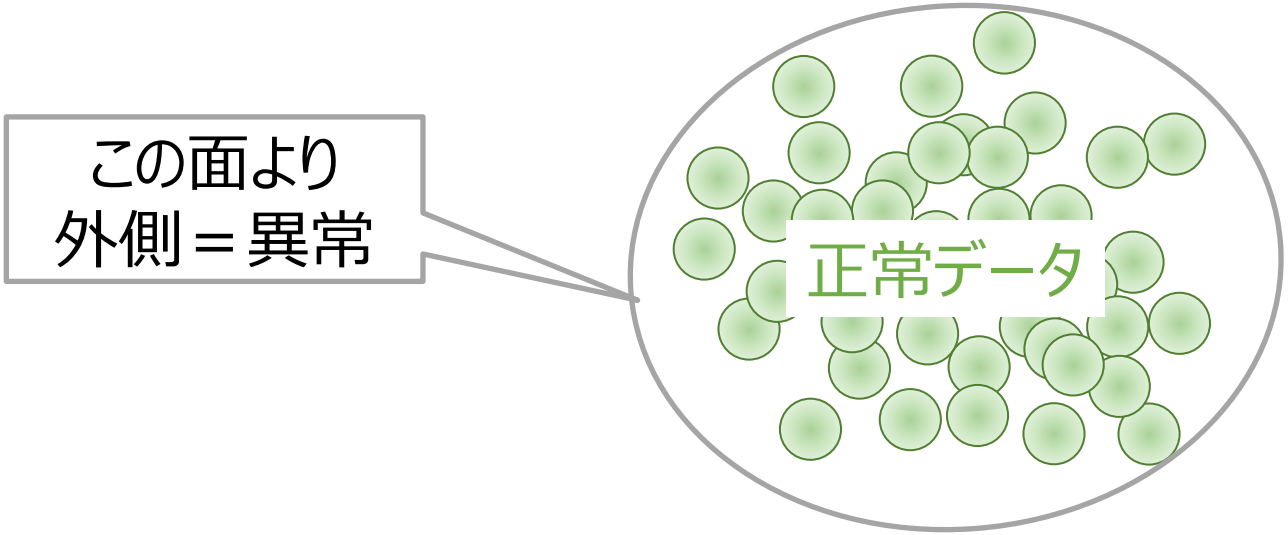
- 正常・異常を分離する「面」を求める問題



- 難点：「めったに発生しないのにバリエーションは膨大」な異常データを集めるのが大変！

異常検出，３つのシナリオ(2/3)：「正常データ」のみが準備できる場合

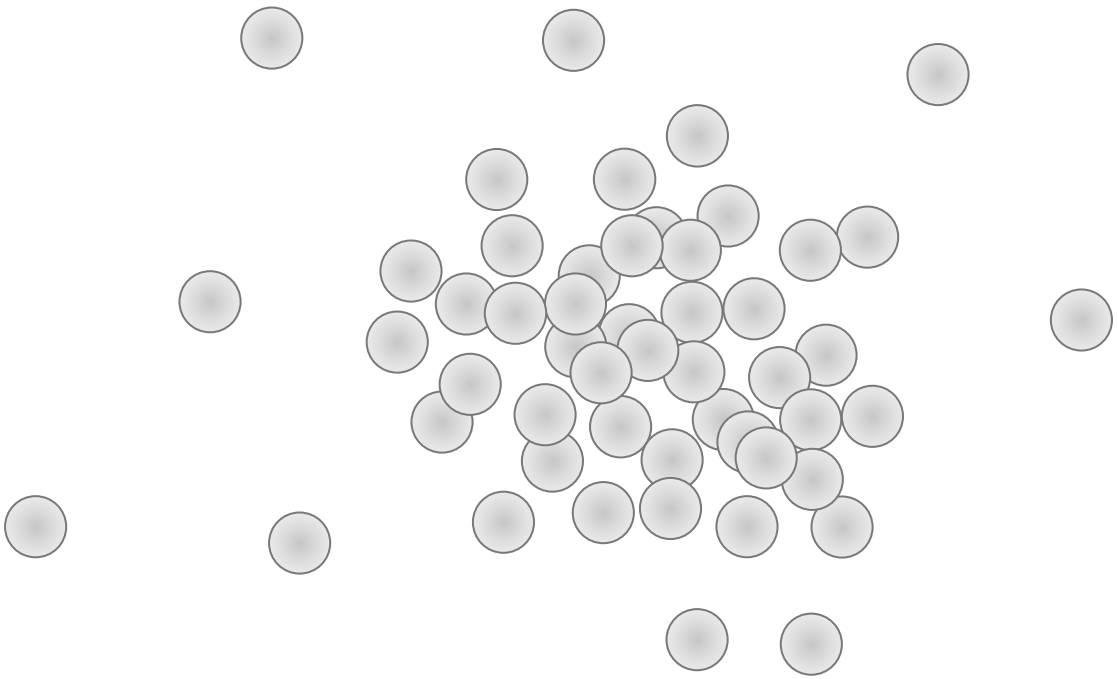
- 正常データを包含する「面」を求める問題



- 利点：異常データ収集の手間は不要
- 難点：しかし全データが正常であることを「担保」する必要は残る

異常検出，３つのシナリオ(3/3)：単にデータの集合しか与えられない

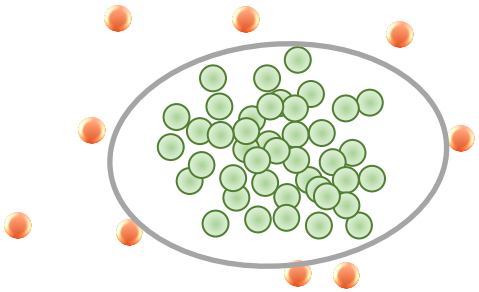
- データ自身に「自分が異常かどうかを判断させる」問題



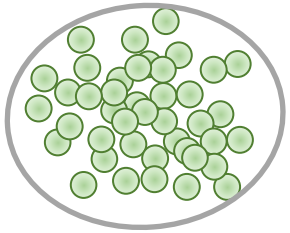
- 最も現実的なシナリオだが，「判断基準」をどうするか？

３つのシナリオ～それぞれの呼称

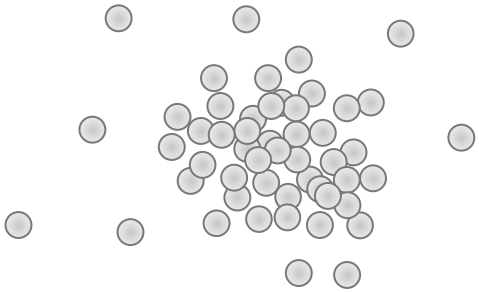
- 第１ ：「教師あり」異常検出
  - 結果は 正常 or 異常



- 第２ ：「半教師付き」異常検出
  - 結果は 正常 or 異常

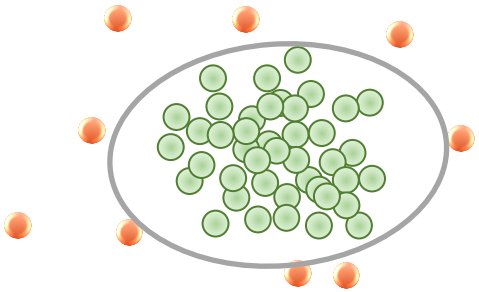


- 第3 ：「教師無し」異常検出
  - 結果は 異常の程度(異常度)

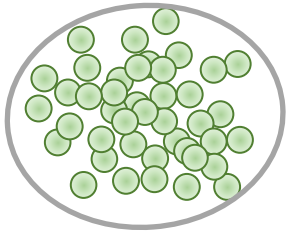


３つのシナリオ～それぞれの対応法

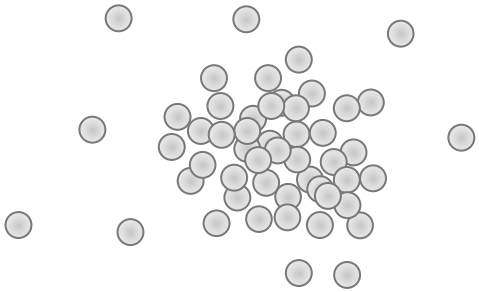
- 第１ ：「教師あり」異常検出
  - (2クラス)パターン認識 (後述)



- 第２ ：「半教師付き」異常検出
  - 1-class SVM



- 第3 ：「教師無し」異常検出
  - k近傍法, LOF



【付録４】
Local Outlier Factor (LOF)法

局所異常も検出できる手法！

Local Outlier Factor (LOF)法 (1/2) 原理

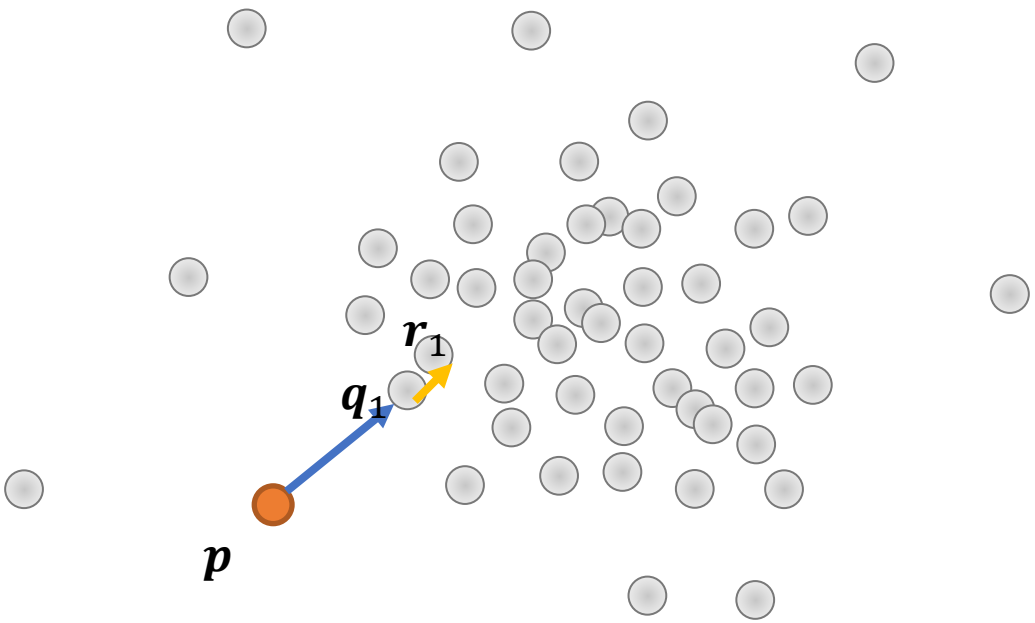
- 注目データ周囲の密度が他に比べ低い→LOF大
- 局所異常検出も可能

最もわかりやすい
$k = 1$の場合

$\boldsymbol{p}$の最近傍

$$\mathrm{LOF} = \frac{\|\boldsymbol{p} - \boldsymbol{q}_1\|}{\|\boldsymbol{q}_1 - \boldsymbol{r}_1\|}$$

$\boldsymbol{q}_1$の最近傍



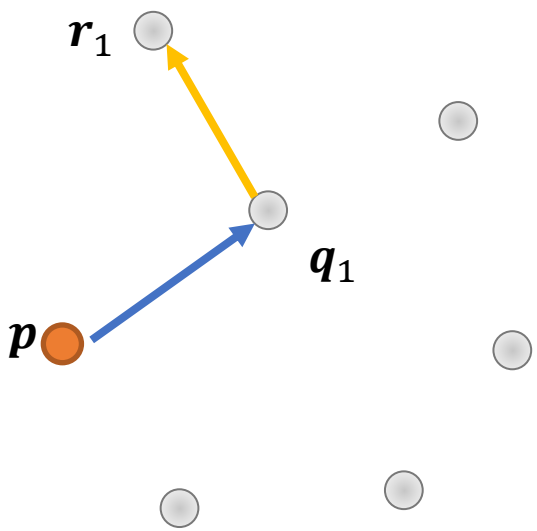
要するに，自分の周り(分子)と自分の近傍の周り(分母)を比較

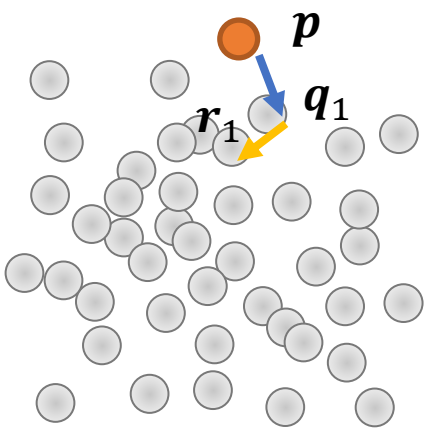
Local Outlier Factor (LOF)法 (2/2)
ぎっしり・スカスカどちらも大丈夫

$$\mathrm{LOF}=\frac{\|\boldsymbol{p}-\boldsymbol{q}_1\|}{\|\boldsymbol{q}_1-\boldsymbol{r}_1\|}\sim 1$$

同程度

$$\mathrm{LOF}=\frac{\|\boldsymbol{p}-\boldsymbol{q}_1\|}{\|\boldsymbol{q}_1-\boldsymbol{r}_1\|}\sim 1$$





$k>1$の場合はもう少し式も
ややこしくなります