クラスタリングについて

2022年9月3日（土）

電気通信大学 特任教授　データサイエンティスト協会 事務局長

斉藤史朗

クラスタリングの特徴と課題

どういう時に使うのか

- ✓ 分析の初めの方で、データがどう分布しているのかを把握するため
- ✓ クラスタリングそのものを目的とする
  - 後述のようにデスマーチになりがちなので、時間のない時には勧めない

特徴

- ✓ 教師なし学習
- ✓ 計算量が必要
  - 安直なのはK-means:でも、それで良いのか？
  - 最近では計算スペックが上がったので、Ward法も実用的になった

課題

- ✓ クラスタ数も使う変数も、さらには手法もどうしたらよいのかわからない
- ✓ 何が「良い」結果なのかは、ビジネス的視点によるので、客観的基準がない
  - グリッドサーチ的なものを組み立てることができない
  - 良いものができるまでの「デスマーチ」になりがち

K-meansのクラスタ数

課題

- ✓ K-means法では、あらかじめ分割するクラスタ数=Kを決めておかなければならない。
  - Ward法であれば、出てきた結果を見て、よろしき数に決めればよい。

やるべきことは何か？

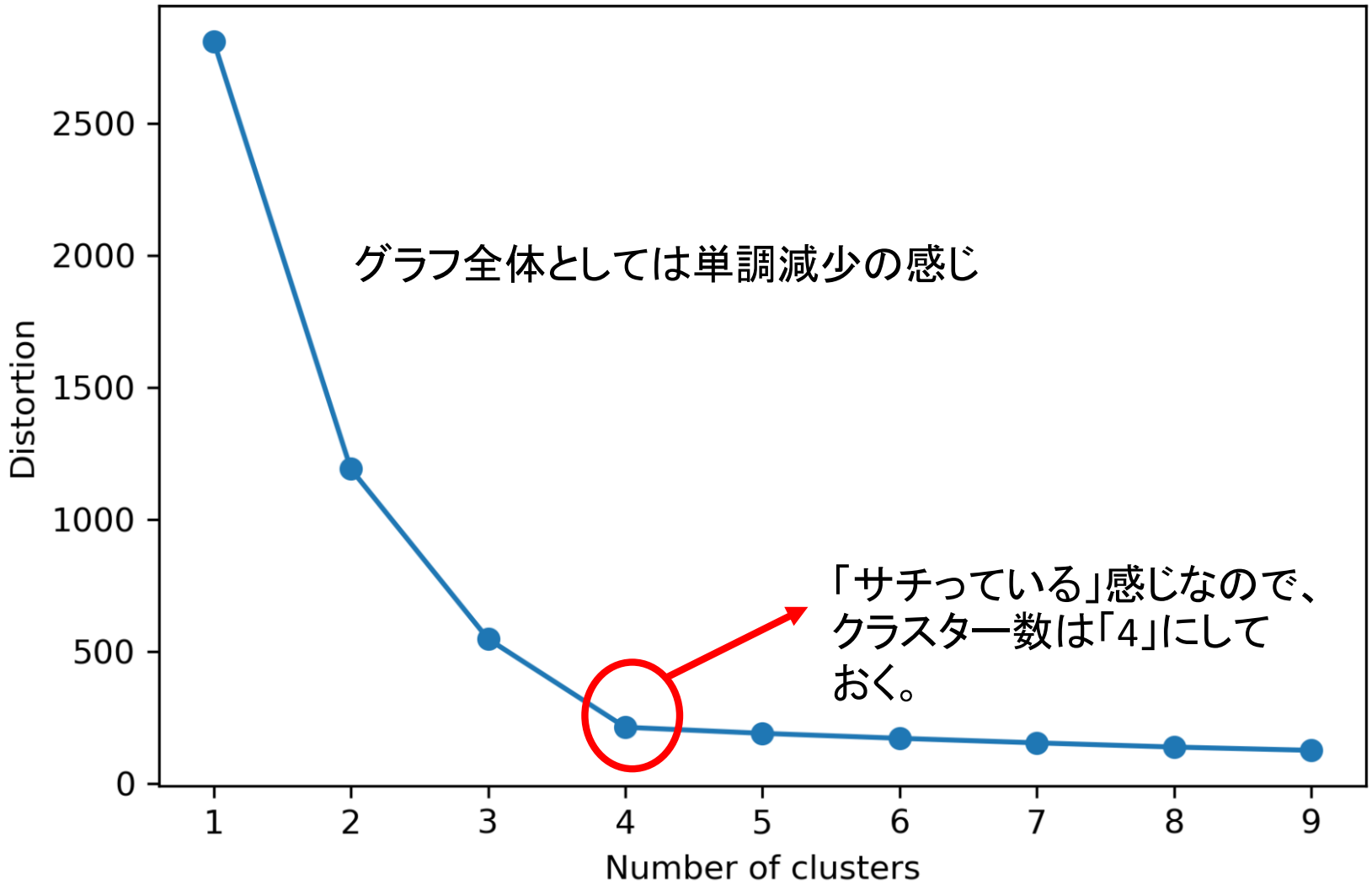
- ✓ Kの値をとりあえず決めて（例えば3から10とか）、それぞれの結果を見て、ビジネス的観点から判断する。
- ✓ グリッドサーチ的手法
  - そんなのめんどくさいから、プログラム書いて、よろしきところを「適当」に選んでくれないか。
  - 本当にそれで良いのか？

エルボー法

- ✓ クラスタ分類をしてできたクラスタごとに、ある評価基準(クラスタ内誤差平方和が多い）で評価して、その結果を可視化する
  - クラスタの中心とクラスタの要素の距離。（これが小さければよい）
  - ちょっとRMSEっぽい基準。
- ✓ なんとなく「サチったところ」を最適なK数とする。

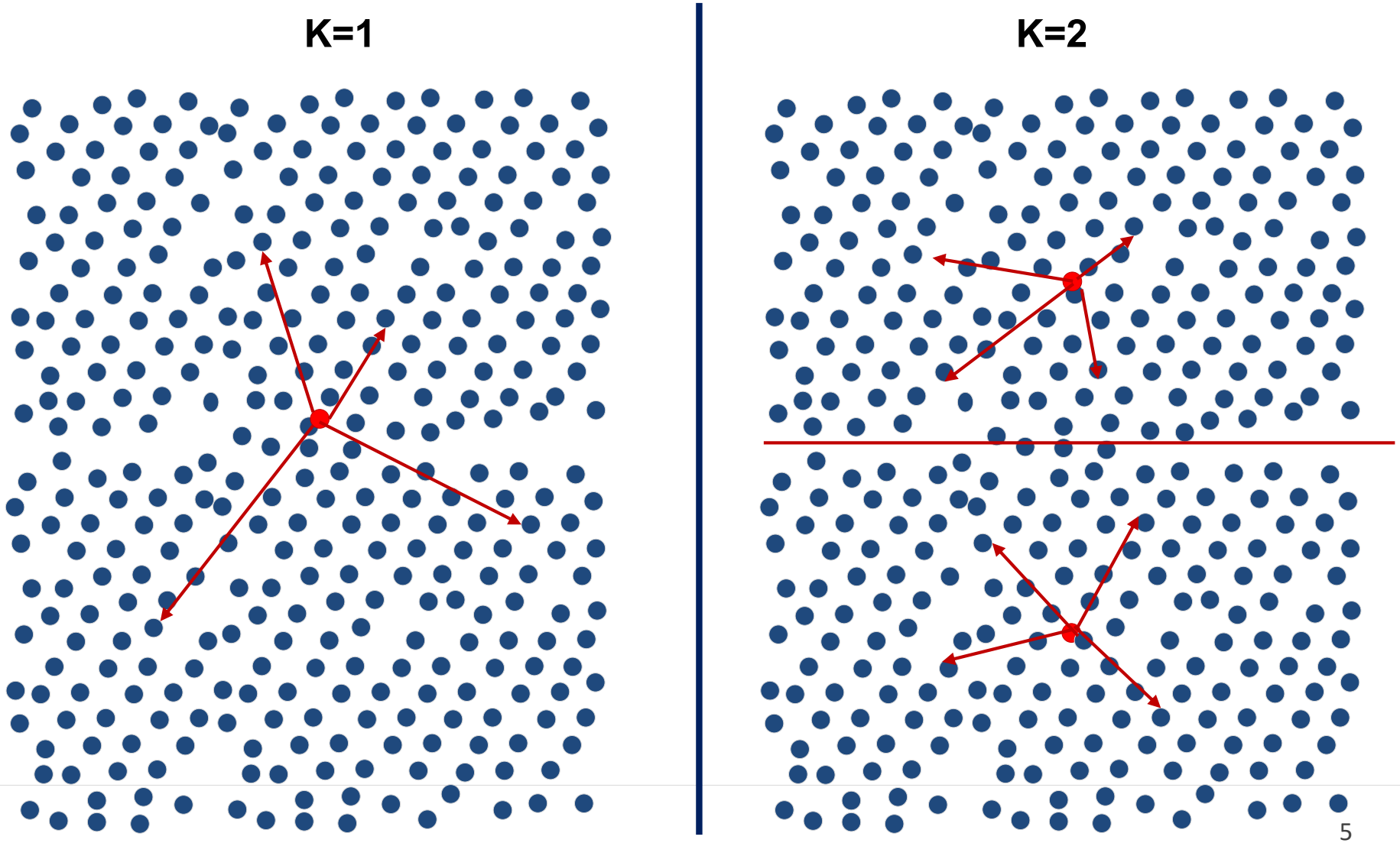
エルボー法

➢ Outputイメージ



世間一般のエルボー法のロジック

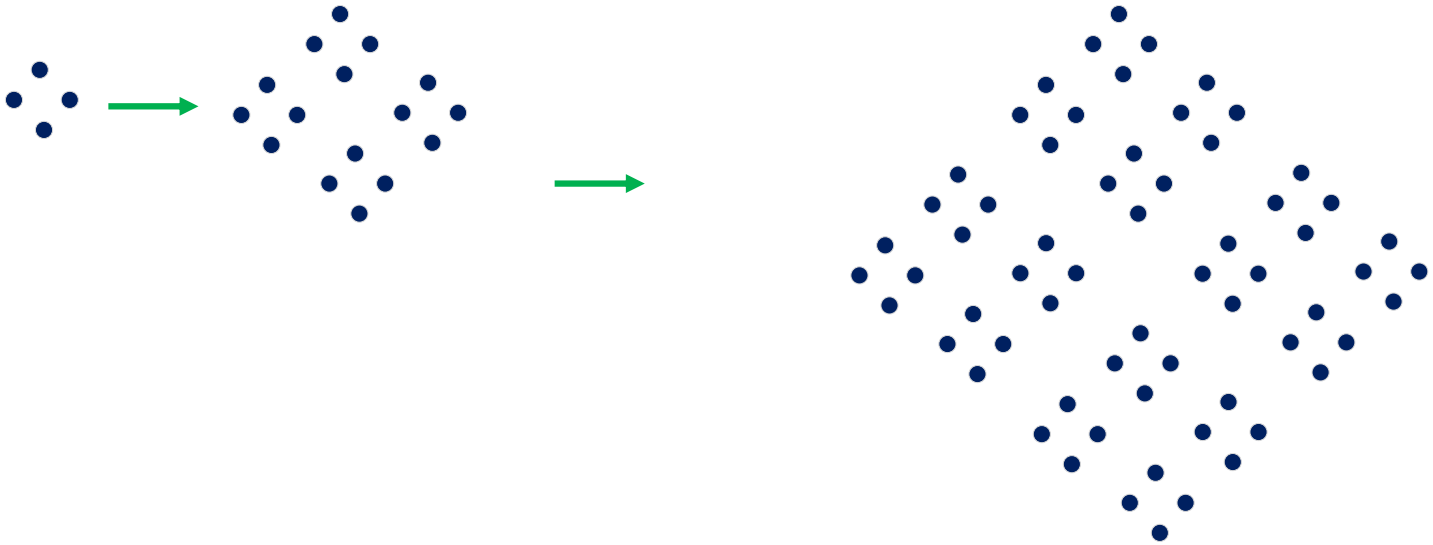
- **Distortion=クラスタ内誤差平方和(SSE)**
  - 「クラスの中でどれだけデータ同士が近いか」と言う指標



世間一般のエルボー法のロジック

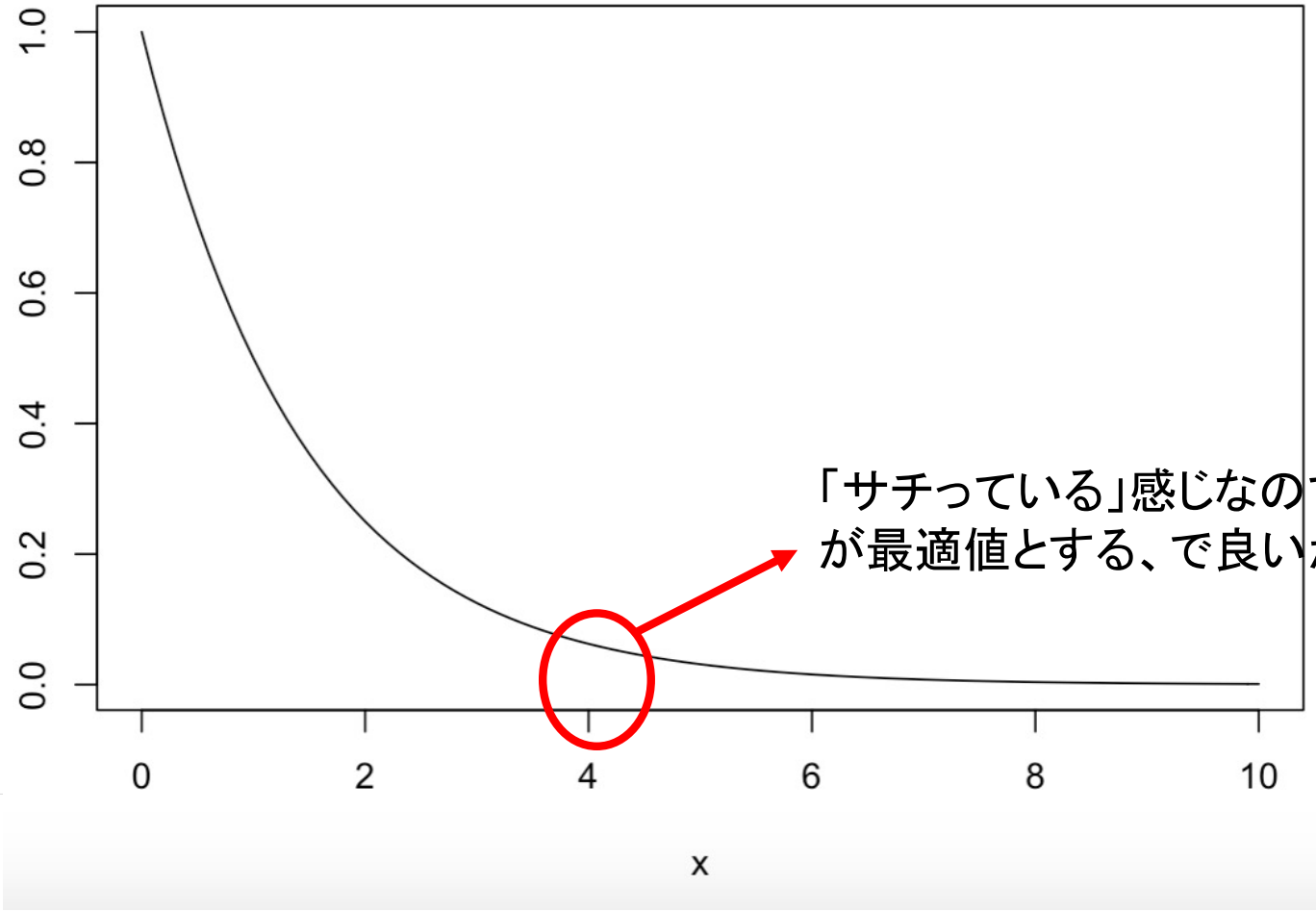
フラクタル図形みたいな分布

- これ、どのレベルでクラスタの数とするか、数値指標だけで決めることができるのか？



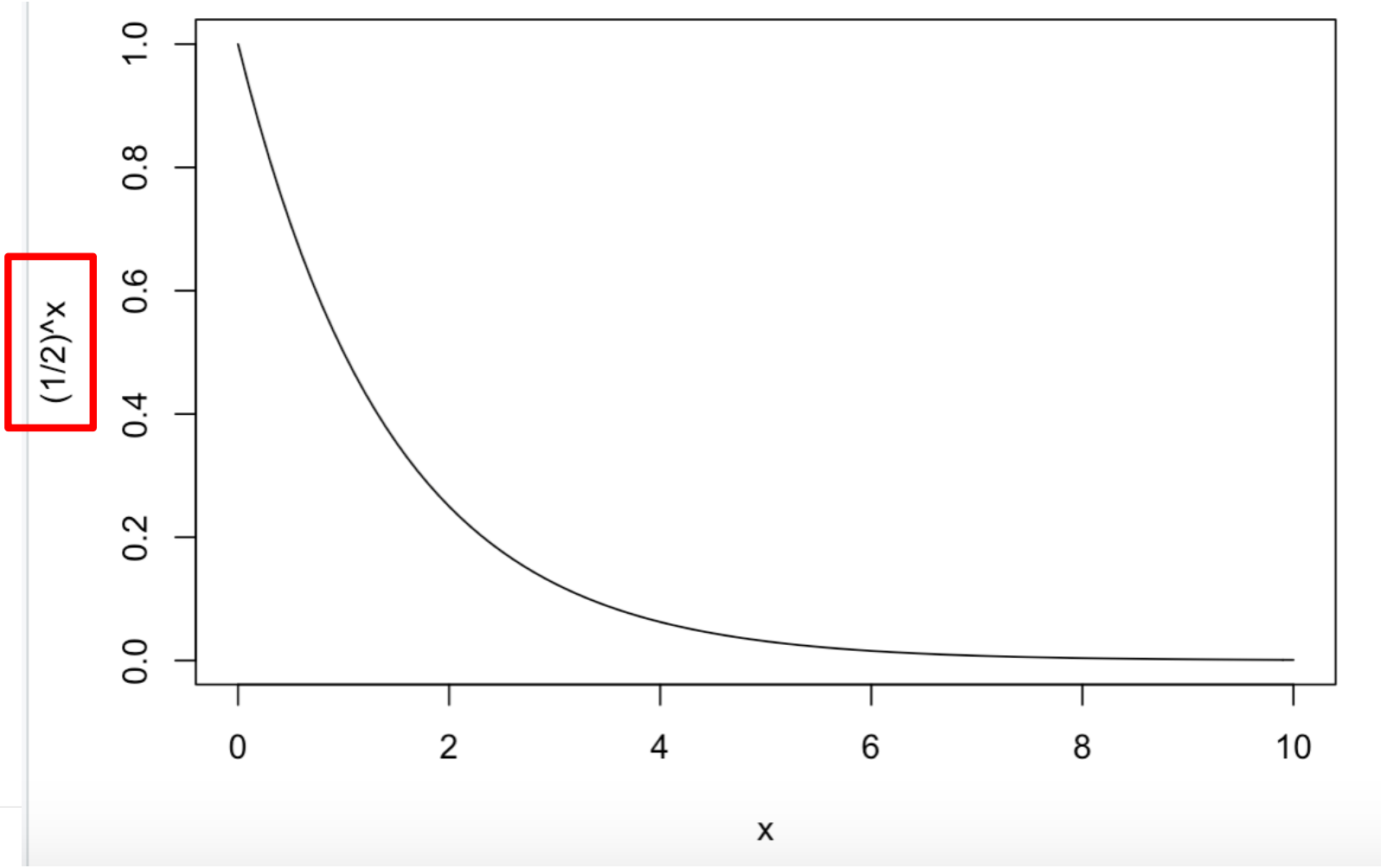
エルボー法で良いのか？

- クラスタ分割とSSE
  - ✓ 空間を分割していくのだから、SSEが分割数に応じて小さくなること＝単調減少は「当たり前」
  - ✓ 「サチっている」っぽい点は本当に最適値なのか？

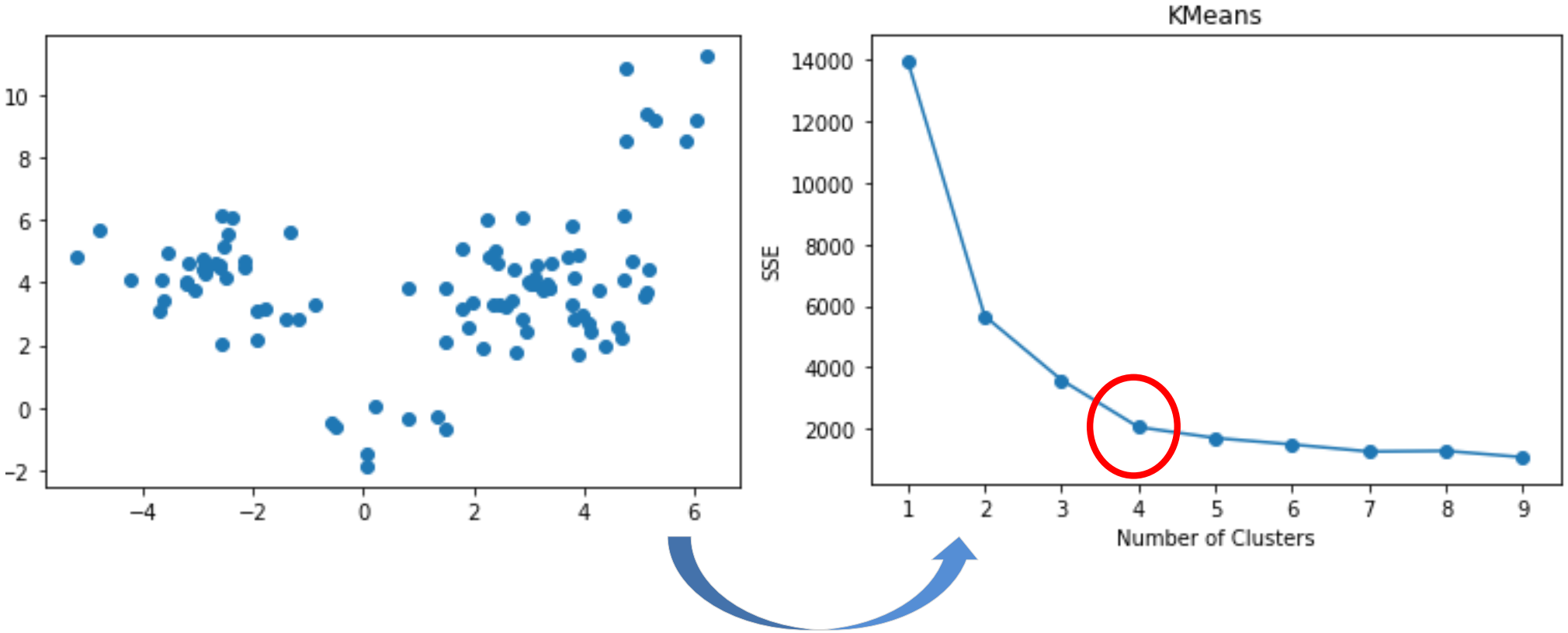


エルボー法で良いのか？

- クラスタ分割とSSE
  - ✓ これ、全然「サチって」**ません**
  - ✓ **減少率は常に「1/2」**



エルボー法を説明している例



※https://hkawabata.github.io/technical-note/note/ML/Evaluation/elbow-method.html より

- でも、これって結局、「見て」確認しているんじゃない？
- 矢印の逆をどうやって確かめるのか？

クラスタ数の決定のために考慮すべきこと

- **ビジネス的視点を除いた場合に、「良い」別れ方って何なのか？**
  - ✓ クラスタの中では「凝集している」＝「互いに近い」
  - ✓ 他のクラスタとは「離れている」＝「距離が遠い」
- **SSE**
  - ✓ 前者（クラスタ内の距離）を見ているだけ
  - ✓ 他のクラスタとの関係は見ていない
- **擬似F値（Pseudo F value)**
  - ✓ **クラスタ間分散と、クラスタ内分散の比を見るような指標**
  - ✓ **この値が大きい方がクラスタ内では凝集し、クラスタ間では分離していることになる**

$$PseudoF = \frac{(T - W_k)/(k-1)}{W_k/(n-k)}$$

T :全サンプルの距離二乗和

W_k :クラスタ内距離二乗和

k :クラスタ数

n : 全サンプル数

クラスタ数の決定のための条件

- **何らかの数値的指標を使ったとして**
  - ✓ 実際にn個のクラスタに分かれている⇒指標が「n」個あたりを表示
  - ✓ 指標が「n」個あたりを表示⇒実際にn個のクラスタに分かれる
  - ✓ 必要条件と十分条件あるいは、「逆は必ずしも真ならず」
- **指標の限界**
  - ✓ SSEはクラス内の距離しか見ていないのに対して、擬似F値はクラス間の距離も見ている
  - ✓ SSEより擬似F値の方がよく見えるが、別の観点からすれば、擬似F値でもまだ不足なところもある
    - 別の観点を見つけて、新しい評価指標を作るのは学問・研究の仕事
    - 学問・研究の進展に終わりがないように、「これで十分」と言う指標は多分ない。（学問に「定説」なし）
  - ✓ では、指標の永遠の進歩に対して、私たちは何を持って評価すれば良いのか？どういう観点で評価すれば十分と言えるのか？
    - 今、問題としているビジネス課題にとって、いくつくらいが妥当なのかを考えるしかない

実際にやるべきこと(k-meansの場合)

- **クラスタ分類に使う変数を決める**
  - ✓ これ、実は難しい
    - 変数を増やしたり減らしたりすると、当然、クラスタの別れ方は変わる
    - カテゴリー変数はどうするの？
    - 欠損値はどう処理すべきか？
- **k=3から8くらいにして、クラスタリングを実行**
  - ✓ 初期値をどう決めるかでも結果が変わる
    - 初期値を変えてやってみるのも良い
  - ✓ それぞれの場合について擬似F値を出してみる
    - プログラムはどっかに落ちているでしょう。
- **クラスタ数の選択**
  - ✓ やってみた中で一番F値が大きいクラスタ数(k)のまわりの3つか4つのクラスタ数を選択
  - ✓ それぞれのクラスタ数で出来上がっている分類を見て、よろしきものを選択する

原点に帰る

- **そもそも何のためにクラスタリングしていたのか？**
  - ✓ データの特徴、どんなグループに分けられるのかが知りたい
  - ✓ 分けられ方については、相手によって「良さ」が違う
    - まさに「ビジネス」的観点
- **エルボー法ってなんで導入したんだっけ？**
  - ✓ いちいち、結果を見て確認するのがめんどくさい
  - ✓ 何かの基準を作って、「自働的に」分割数を決定したい
  - ✓ 「計算量が多いので、あらかじめ分割数を決定して、それだけを実行することにして、計算時間を減らしたい」わけではない
    - あらかじめ全部計算しておいて、その中から選択する
- **クラスタ数の選択の要点**
  - ✓ 結局、結果を人間が吟味する必要がある
  - ✓ だったら、全部自働的になんて考えないで、腹をくくって、ある程度の数は内容を見て行くべき
    - エルボー法にしろ擬似F値にしろ最初にざっくりクラスタ数の範囲を決める時に使えば良い