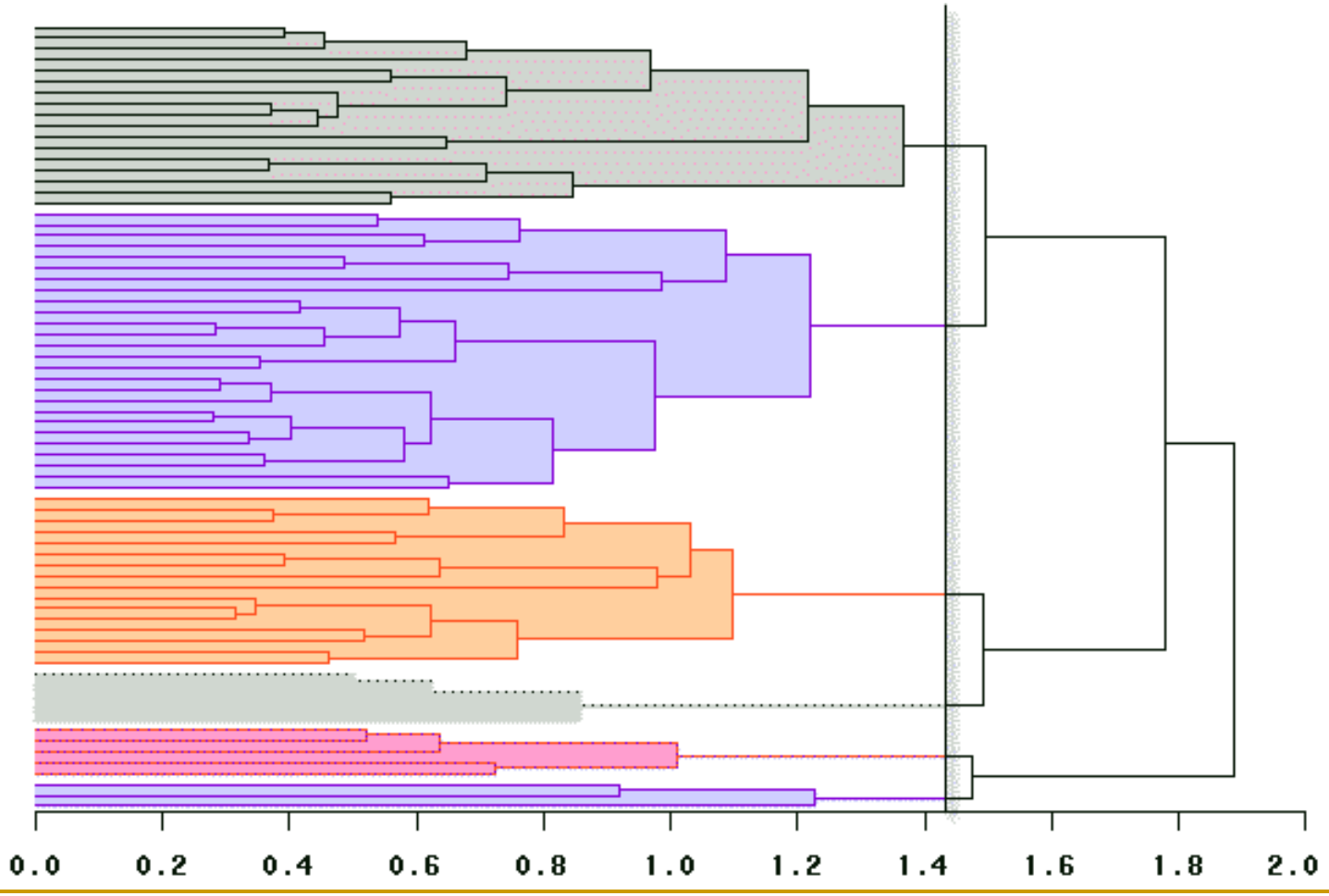
Автоматизация определения числа кластеров

Аббакумов
Вадим Леонардович

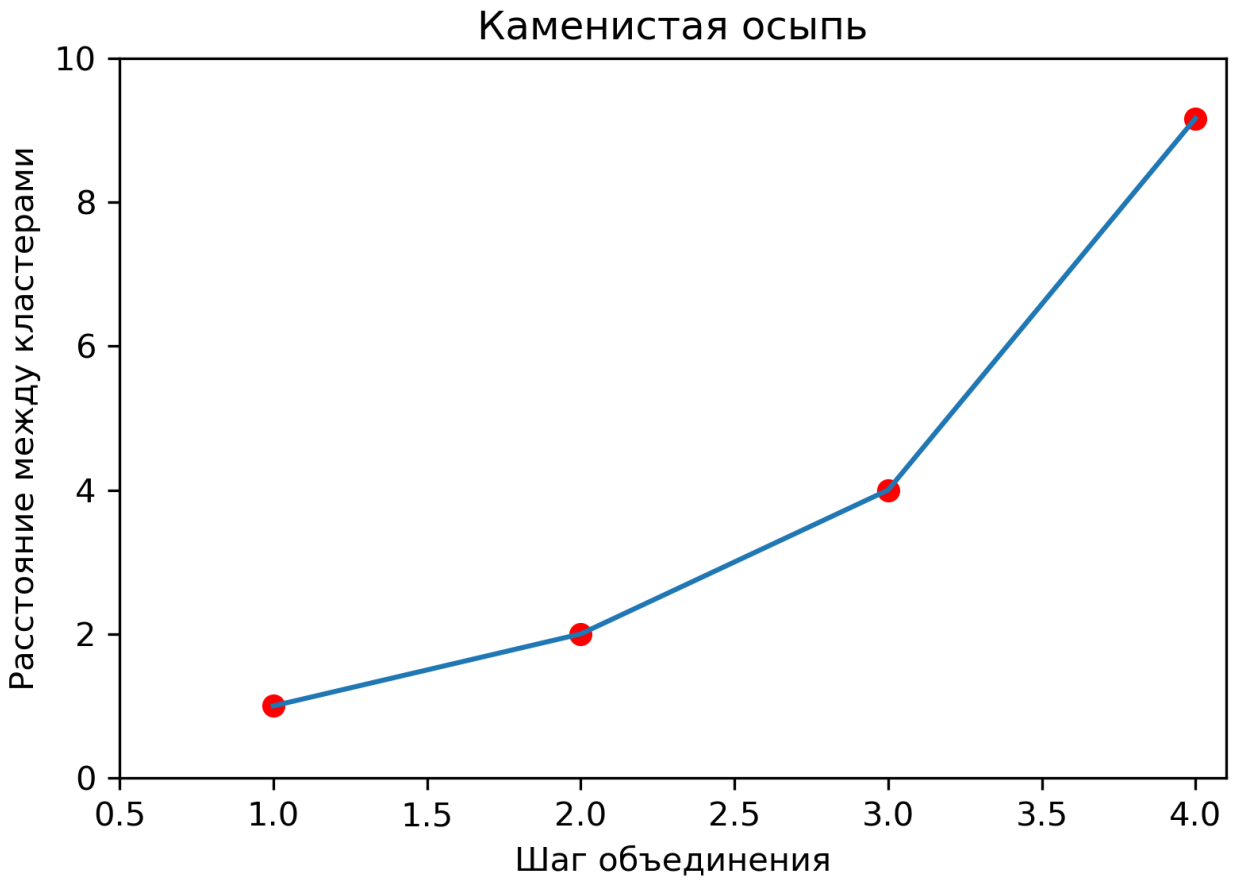
Версия 01

Методы универсальные

Пример дендрограммы



Пример каменистая осыпь / локоть



Силуэт


График значений силуэта
Силуэт
0.48
0.46
0.44
0.42
0.40
0.38
0.36
2
3
4
5
6
7
8
9
Число кластеров

Критерий качества Silhouette

- $Dist(x_i, c_k)$ = среднее расстояние от $x_i \in c_k$ до других объектов из кластера c_k (компактность),
- Dist(x_i, c_l) =среднее расстояние от $x_i \in c_k$ до объектов из ближайшего другого кластера c_l: $k \neq l$ (отделимость).
- $$Silhouette(x_i) = \frac{Dist(x_i, c_l) - Dist(x_i, c_k)}{max(Dist(x_i, c_k), Dist(x_i, c_l))}$$
- Среднее по кластеру, по всей выборке

Вопрос

Если у кластеризации А значение силуэта больше,

чем у кластеризации Б, то какая из них лучше?

Три вопроса

- Может ли силуэт быть больше единицы?
- Может ли силуэт быть отрицательным?
- В каком интервале заключен силуэт точки?

Дано $x > 0, \quad y > 0$

$$\frac{x-y}{max(x,y)} = \frac{x}{max(x,y)} - \frac{y}{max(x,y)}$$

Вопрос

Очевидно,

что если силуэт отрицательный,

то кластеризация ОЧЕНЬ плохая

Или не очевидно?

Вопрос

Что измеряет силуэт,

если даже у отличной кластеризации

значение силуэта может быть отрицательным

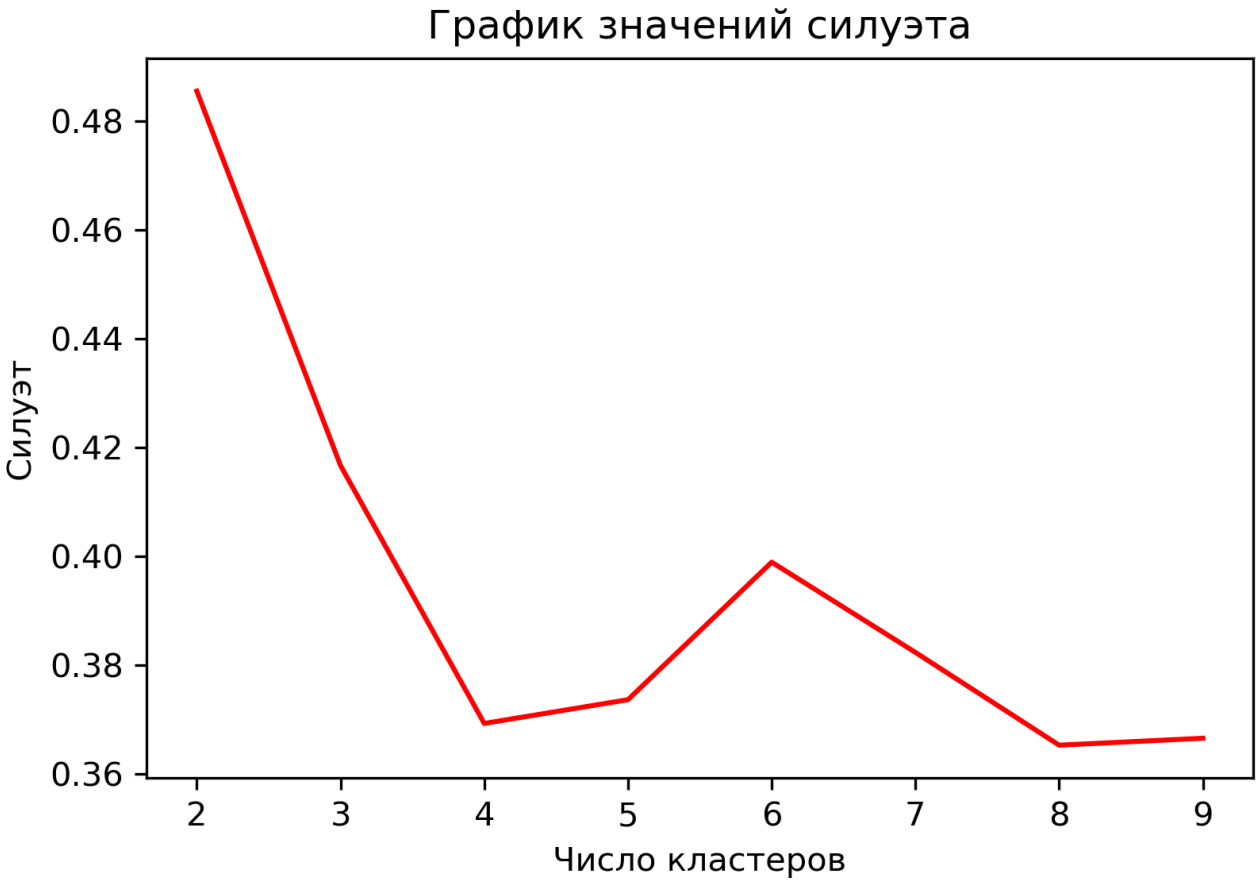
иногда даже близким к -1 ?

метод силуэт - эмпирический

А если компактность измерять
как среднее расстояние до центра кластера?

Почему считаем средние, а не медианы?

Силуэт для определения числа кластеров



Вопрос

Почему выбираем решение с шестью кластерами?

У решения с двумя кластерами значение силуэта больше!

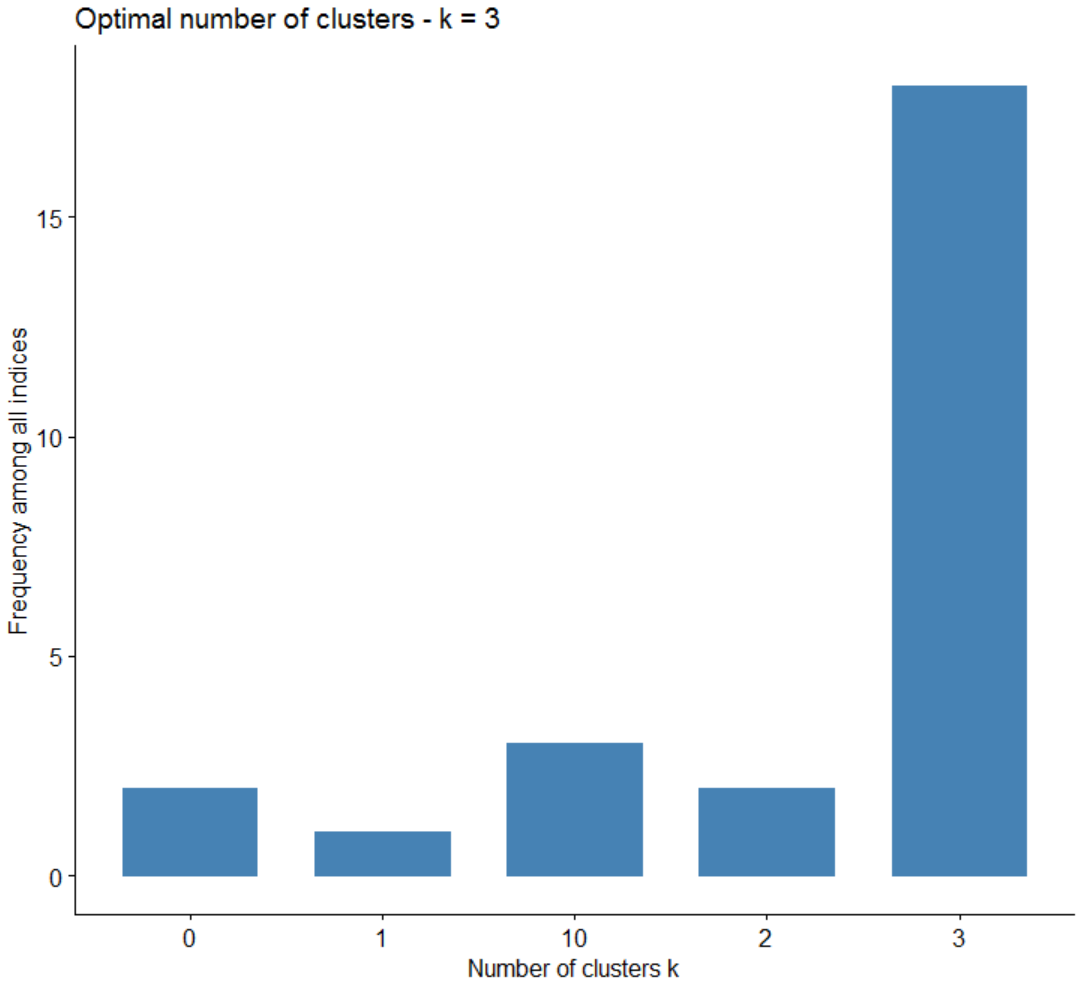
Другие критерии качества кластеризации

- Gordon Classification 2ed

- https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B5_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8#.D0.A1.D0.B8.D0.BB.D1.83.D1.8D.D1.82_.28.D0.B0.D0.BD.D0.B3.D0.BB._Silhouette.29

Авторы NbClust

- Malika Charrad
- Nadia Ghazzali
- Veronique Boiteau
- Azam Niknafs

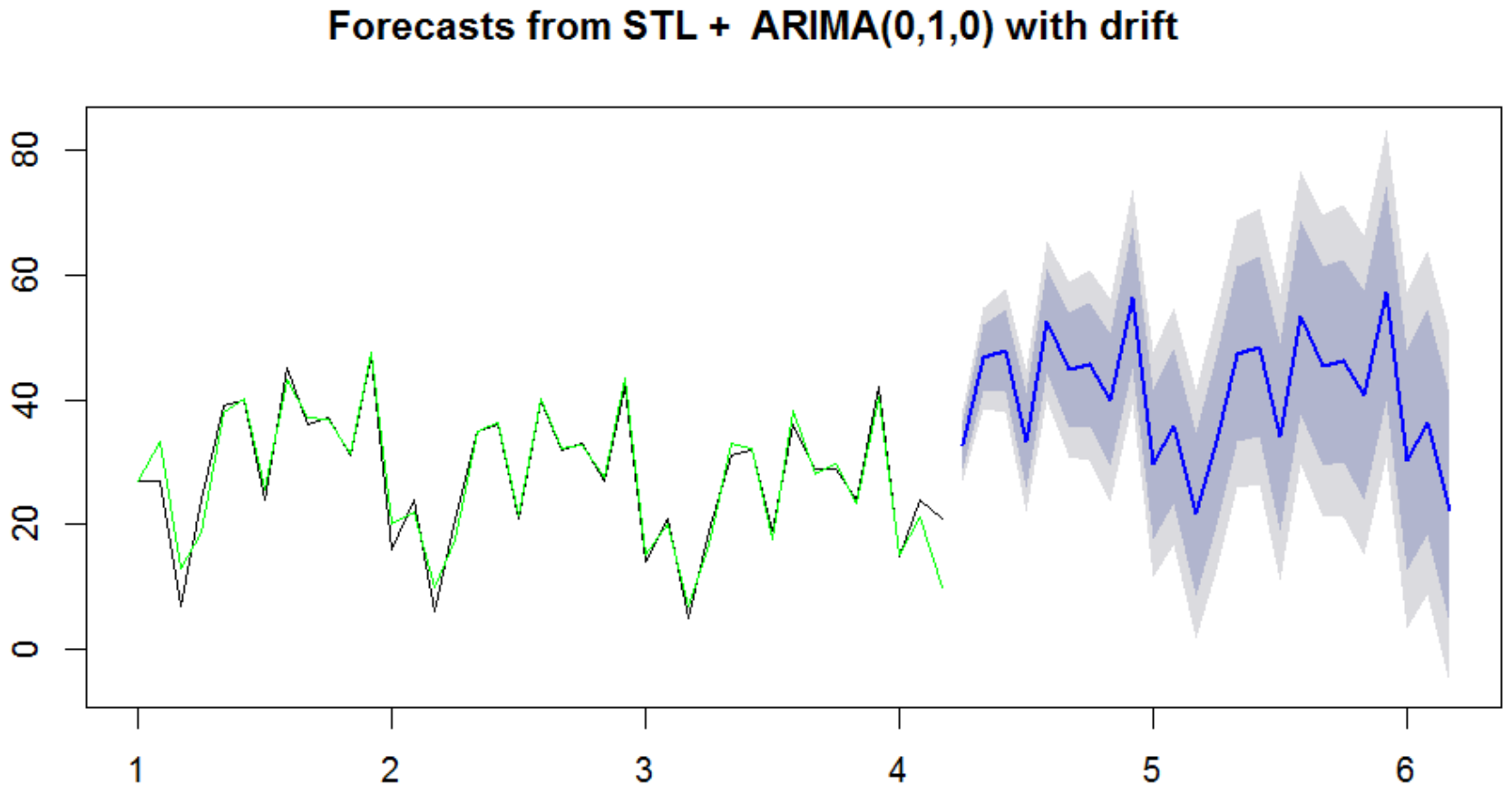

Optimal number of clusters - k = 3
Frequency among all indices
0
5
10
15
0
1
10
2
3
Number of clusters k

- rpy2
- rpy2 is an interface to R running embedded in a Python process
- https://rpy2.github.io/

- Джеймс Шуровьески

- Мудрость толпы

  Почему вместе мы умнее, чем поодиночке, и как коллективный разум влияет на бизнес, экономику, общество и государство
- The wisdom of crowds. Why the Many Are Smarter Than the Few


Forecasts from STL + ARIMA(0,1,0) with drift
0
20
40
60
80
1
2
3
4
5
6

- Кластеризовали 100000 объектов, получили 21 кластер

- Получим 20 кластеров, состоящих из одного наблюдения каждый. И один кластер, содержащий все остальные (100000-20)=99980 наблюдений.

Очень трудно искать черную кошку в темной комнате, особенно, если там ее нет (Может это сказал Конфуций, может нет)

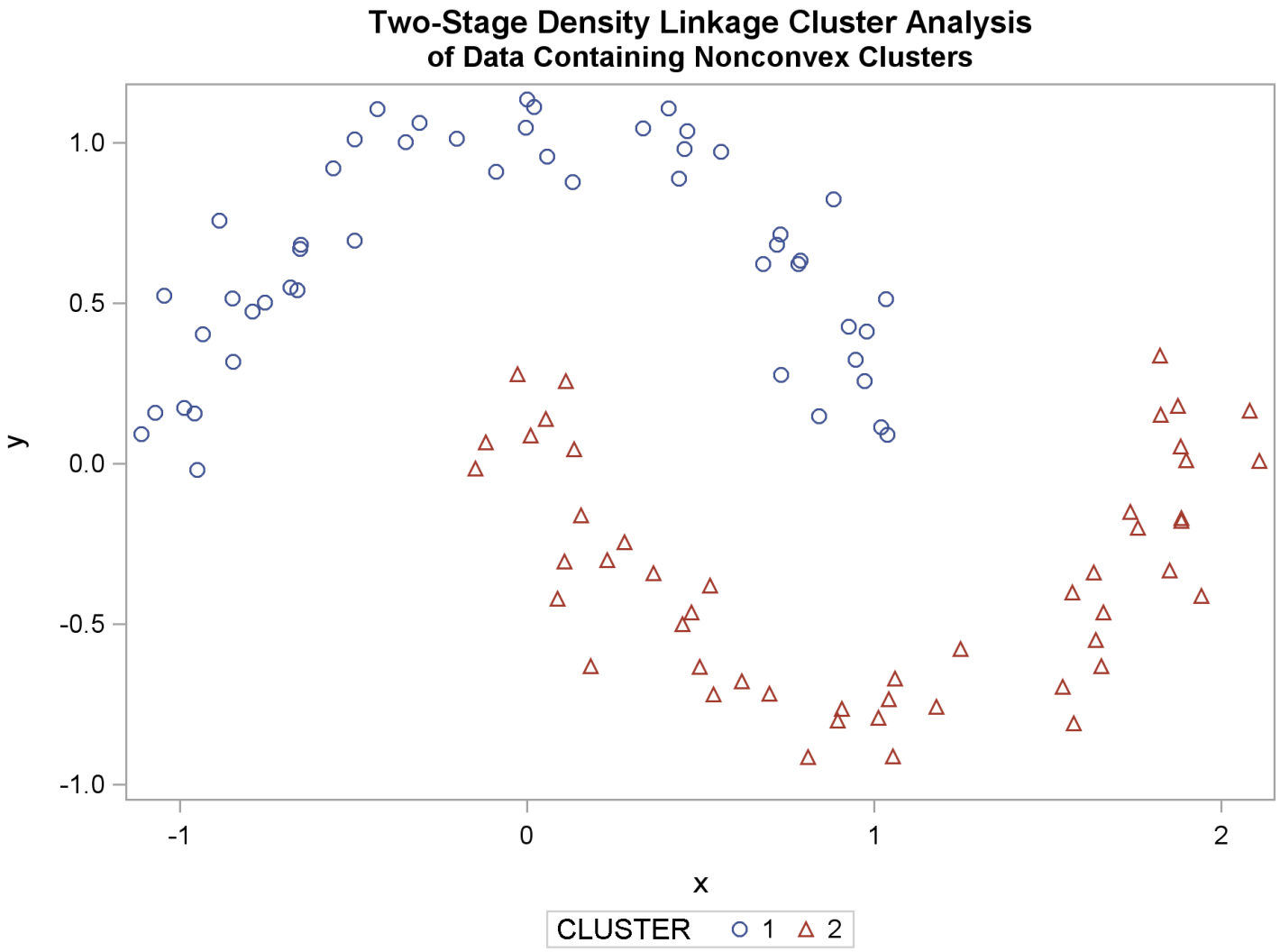
Если кластеров нет, они все равно будут найдены

Автоматическое определение числа кластеров

- Разные методы дают разное число кластеров (один 2 кластера, другой 19)
- Мудрость толпы
- В R пакет/процедура Nbclust

- В Питоне аналогов пока нет

Метод ближайшего соседа



- Если кластер содержит единственную точку, то эта точка — выброс

- распознаем породы собак
- фотография слоненка