

---

# Иерархический кластерный анализ

---

Аббакумов

Вадим Леонардович

---

# Происхождение термина

- Кластер – калька слова «cluster»,
  - «сгусток», «гроздь (винограда)», «скопление (звезд)» и т.п.
-

---

# Ранее использовались другие термины

- распознавание образов без учителя,
  - стратификация,
  - таксономия,
  - автоматическая классификация.
-

---

# Задача

- Кластерный анализ разбивает набор объектов на группы
  - Попутно определяется число групп
-

---

# Определение

- Группы, на которые разбита выборка, называются кластерами.
-

---

## Еще раз:

- при иерархическом кластерном анализе заранее неизвестно число кластеров (групп, на которые разбивается набор объектов).
-

---

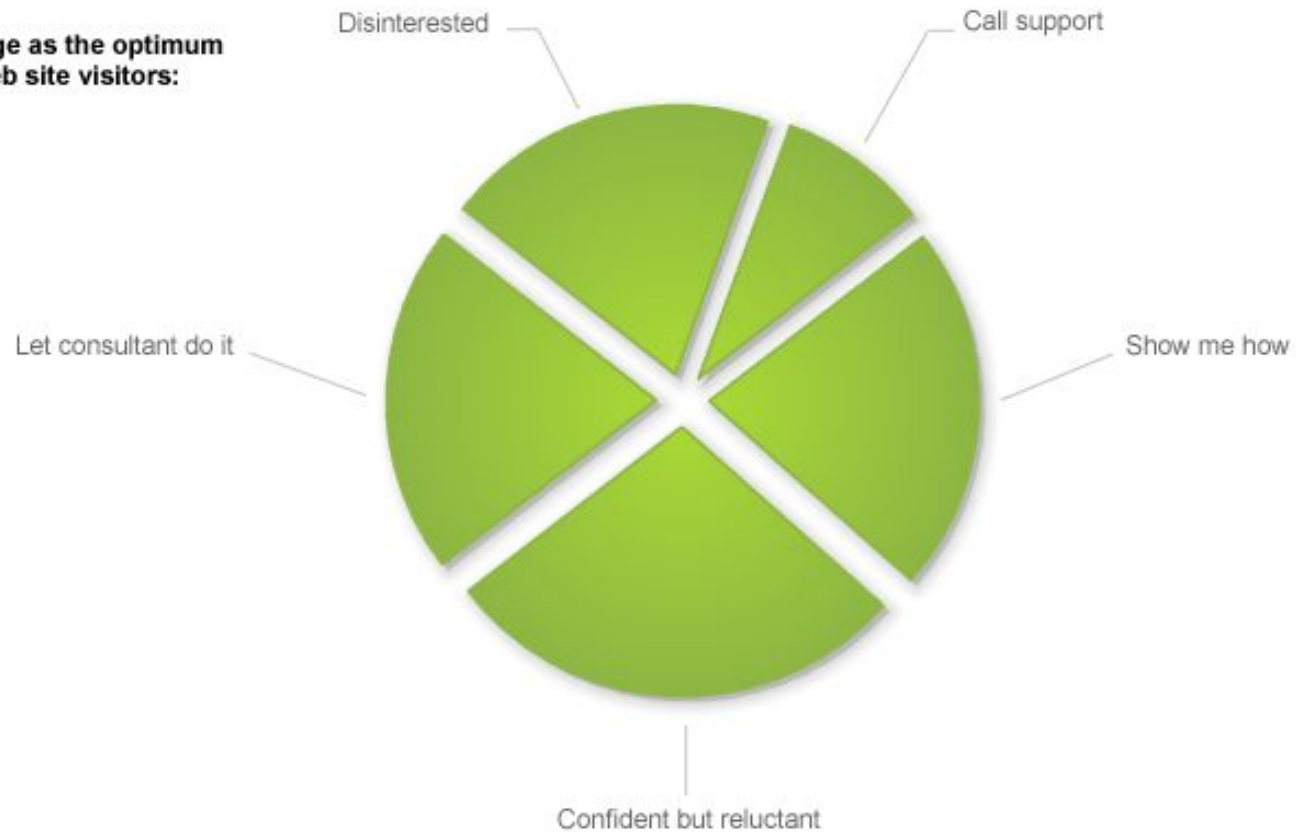
# Другие методы кластеризации

- Метод k-средних
  - Самоорганизующиеся карты Кохонена (SOM)
  - Смесь (нормальных) распределений
  - ...
-

# В маркетинге: Сегментирование рынка

## Persona Segments

5 "Persona" segments emerge as the optimum solution representative of web site visitors:





---

# Пример:

## Определение групп потребителей

- – По данным о покупателях (результаты опроса, поведение на сайте) выявить и описать/понять рыночные сегменты.
  - – Прежде, чем фирма определится, какие сегменты рынка создают для нее наибольшие возможности, надо решить, какие сегменты уже существуют.
-

- 
- Страховая компания интересуется группами, на которые разделяются потенциальные клиенты.
  - Результаты классификации используются, чтобы для разных групп определять оптимальные цены на услуги, оптимальные тарифы
-

---

# Пример:

## Определение групп потребителей

– Для разбиения потребителей на группы можно выбирать разные наборы характеристики объектов, например возраст, образование, место жительства, тип личности, и так далее.

Несложно разделить покупателей на сегменты по **одной** (или по каждой) характеристике.

Кластерный анализ может помочь выявить уже сложившееся разбиение потребителей на *«группы со схожими потребностями в отношении конкретного товара или услуги, достаточными ресурсами, а также готовностью и возможностью покупать»* учитывая **все** выбранные показатели одновременно.

---

---

# Пример: товарные группы для рекомендательной системы

На рынке присутствует большой выбор товаров схожего назначения под разными торговыми марками. Надо разбить товары на группы.

Иногда такое разбиение известно и получается без применения статистической техники. Например, компьютеры бывают «для дома», «для офиса», «серверы» и «специализированные».

Кластерный анализ применяется, если нет классификации, признанной всеми.

Важно! Результат будет зависеть от выбора набора показателей.

---

---

# Пример

- Определение целевой аудитории баннерной рекламной компании в интернете.
  - 100000 сайтов
  - Каждый из них указывает на интересы куки, на текущее настроение куки...
  - Надо отождествить схожие сайты
-

---

# Другие задачи классификации

- Machine Learning
  - Классификация с учителем
  - Распознавание образов
-

---

# Отличие

- Заранее известно, к какому классу принадлежит каждое из наблюдений.
  - Технологически - среди переменных присутствует так называемая группирующая переменная.
-

---

# Что тогда классифицировать?

- Надо придумать правило.
  - Для классификации новых наблюдений.
-



---

# Другие задачи классификации

- Классификация с обучающей выборкой
    - наивный байесовский классификатор
    - дискриминантный анализ
    - деревья классификации
    - K-го ближайшего соседа
    - Нейронная сеть прямого распространения
    - SVM
    - Случайный лес
    - Gradient boosting machine
-

- 
- Вернемся к кластерному анализу
-

---

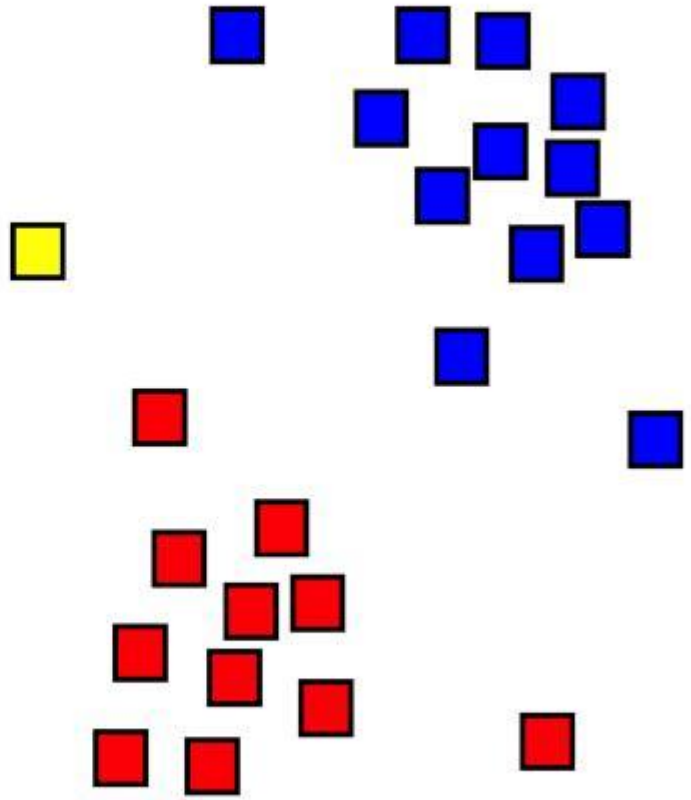
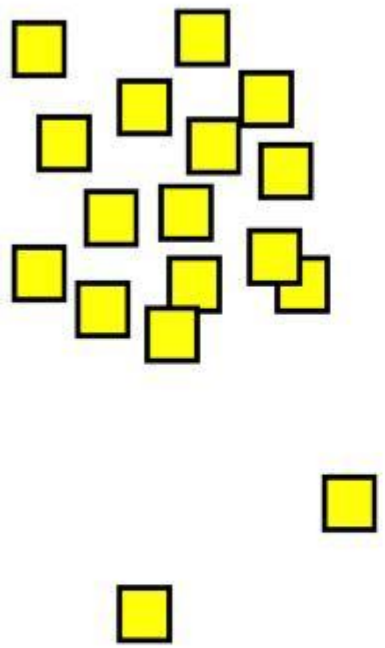
# Идея метода

- Сведем задачу к геометрической
-

---

# Сведем задачу к геометрической

- Каждый объект – точка.
  - Похожие объекты расположены «близко» друг к другу
  - Различающиеся объекты расположены «далеко»
  - Скопления точек – кластер.
-



---

# Расстояние между объектами

- Евклидово расстояние
  - Квадрат Евклидова расстояния
  - Блок (Манхеттен, сити-блок)
  - и так далее...
-

---

# Расстояние Евклида

- *Две точки*

$(x_1, x_2, x_3)$

$(y_1, y_2, y_3)$

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$



---

Квадрат евклидова расстояния  
не является расстоянием...





# *Расстояние Block*

(Manhattan, таксиста).



## *Расстояние Block*

(Манхаттан, таксиста, Минковского при  $p=1$ ).

$$X = (x_1, x_2, \dots, x_k)$$

$$Y = (y_1, y_2, \dots, y_k)$$

$$d_{XY} = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_k - y_k|$$

---

# Расстояние Хэмминга

- Число позиций, в которых соответствующие символы двух слов одинаковой длины различны
  - 
  - $D(1011101, 1001001) =$
  - $D(2173896, 2233796) =$
  - $D(\text{toned}, \text{roses})$
-

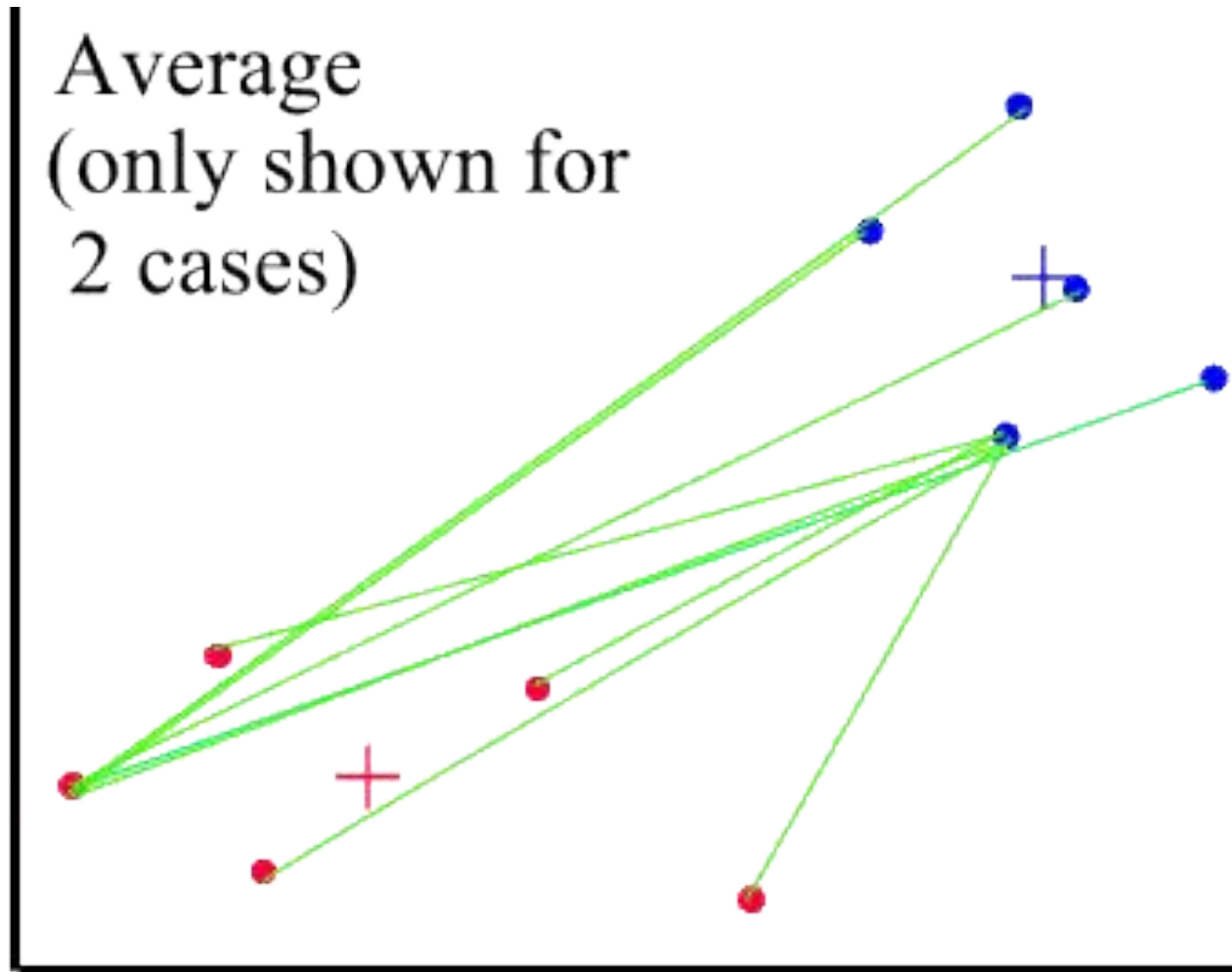
- 
- Вопрос:
  - Когда выбирать евклидово расстояние, а когда расстояние Манхэттен?
-

---

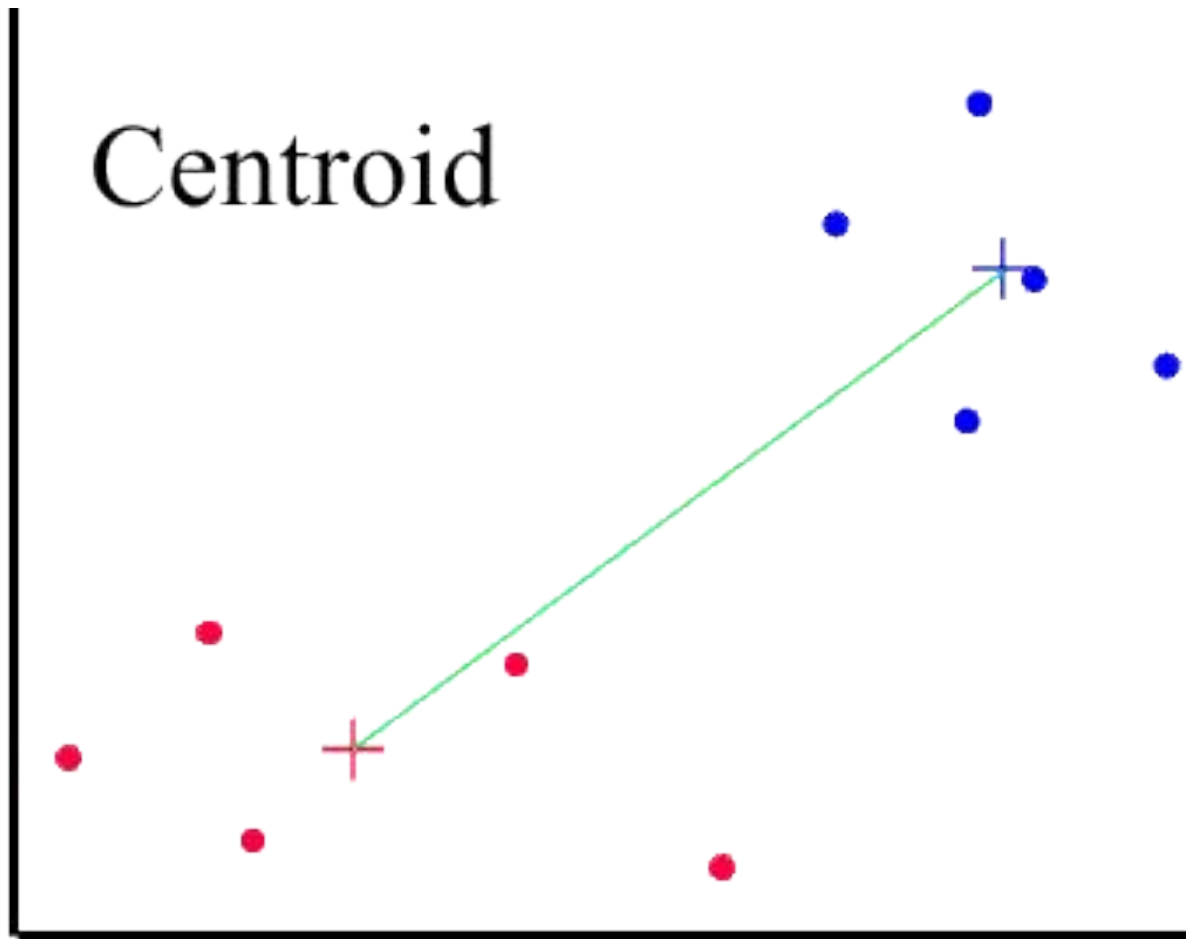
# Расстояние между кластерами

- Среднее невзвешенное расстояние (Average linkage clustering).
  - Центроидный метод (Centroid Method).
  - Метод дальнего соседа, максимального расстояния (Complete linkage clustering).
  - Метод ближайшего соседа (Single linkage clustering).
  - Метод Варда (Ward's method).
-

# Среднее невзвешенное расстояние



# Центроидный метод



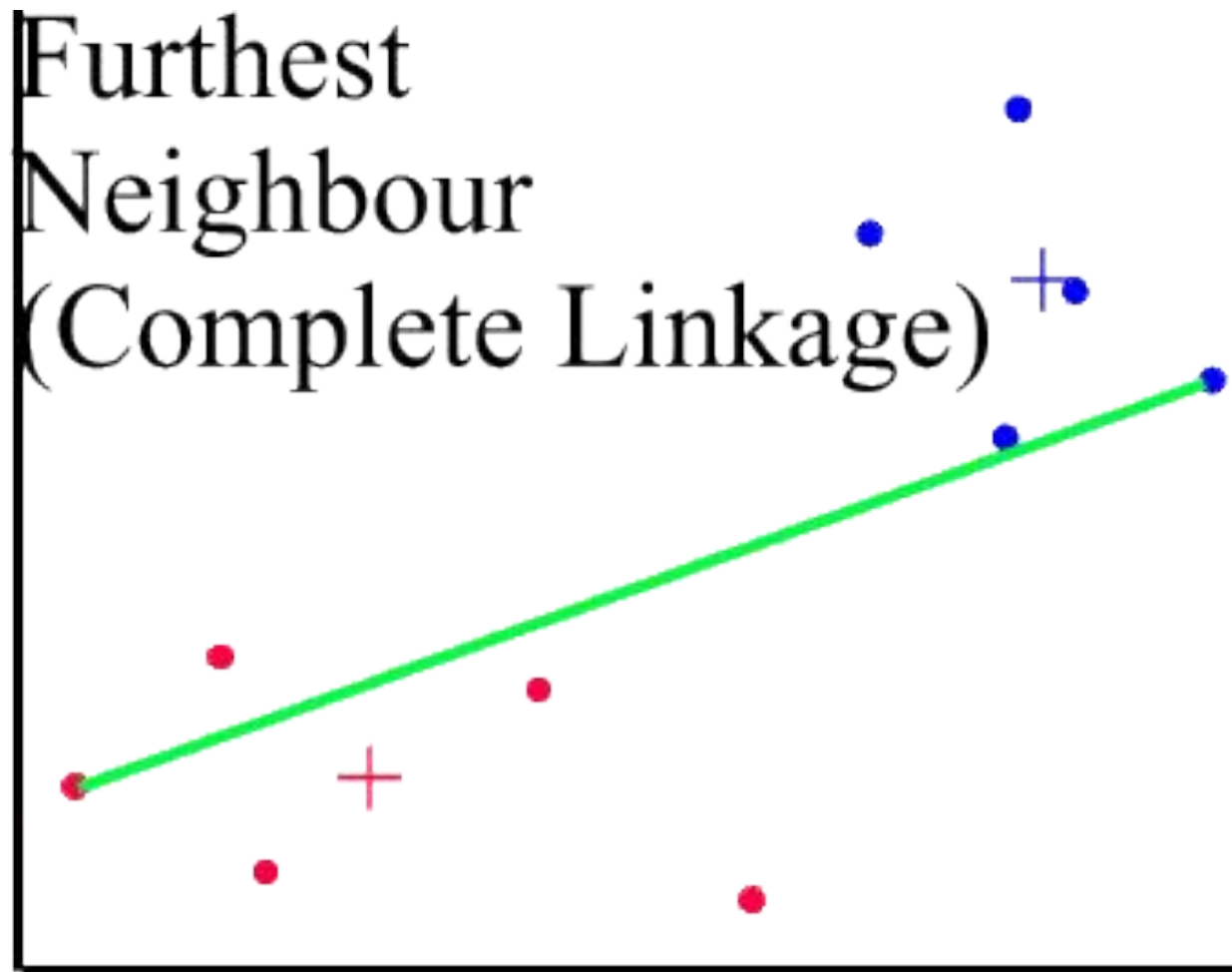
---

# Центроидный метод

- Вычислительная простота
  - Объем кластера не влияет.
  - Дендрограмма может иметь самопересечения
  - Выходит из употребления
-

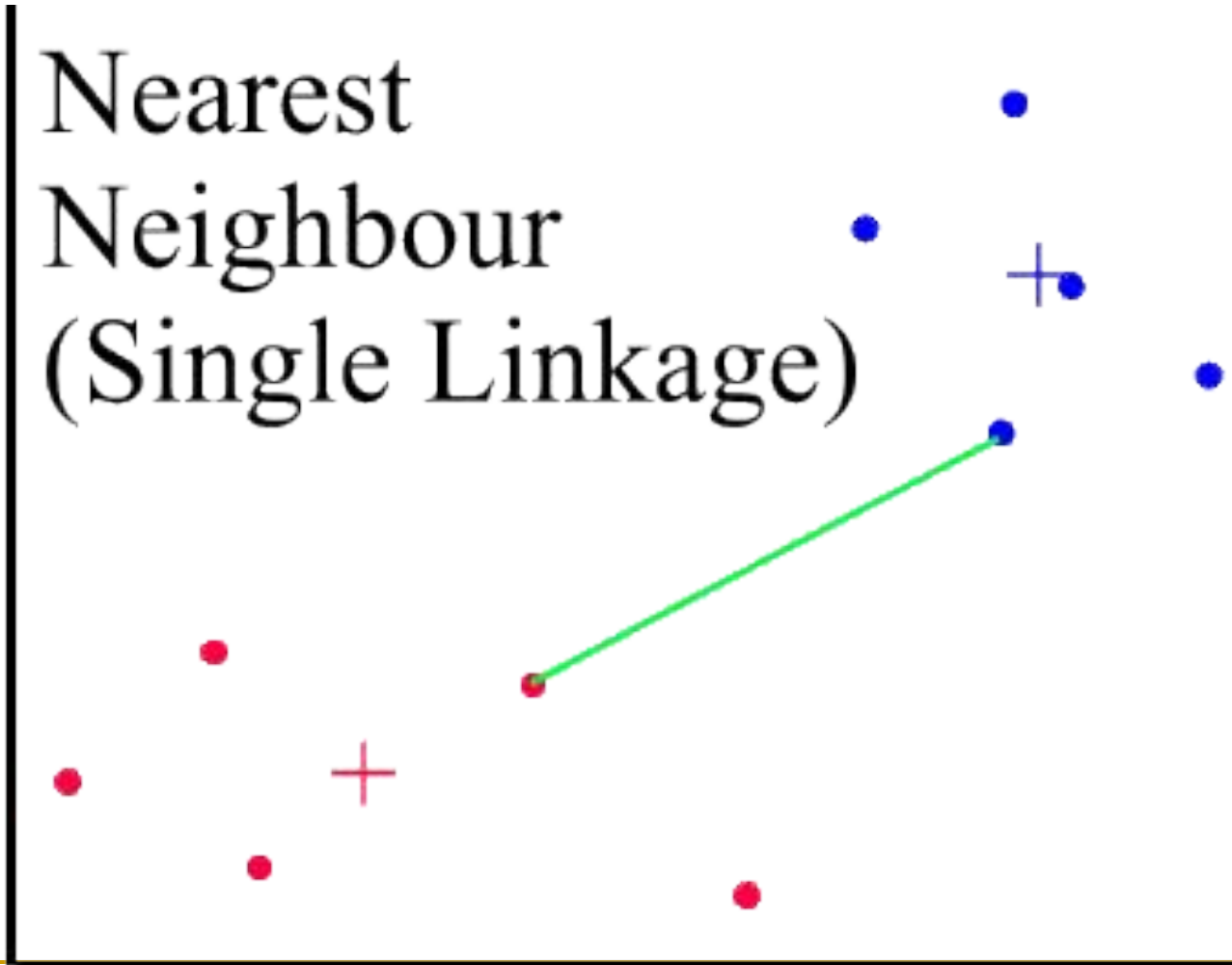


# Метод дальнего соседа



# Метод ближайшего соседа

Nearest  
Neighbour  
(Single Linkage)



---

# Расстояние Sørensen–Dice

---

- 
- Метод Варда (WARD).
  - Предполагается использование квадрата евклидова расстояния
-

---

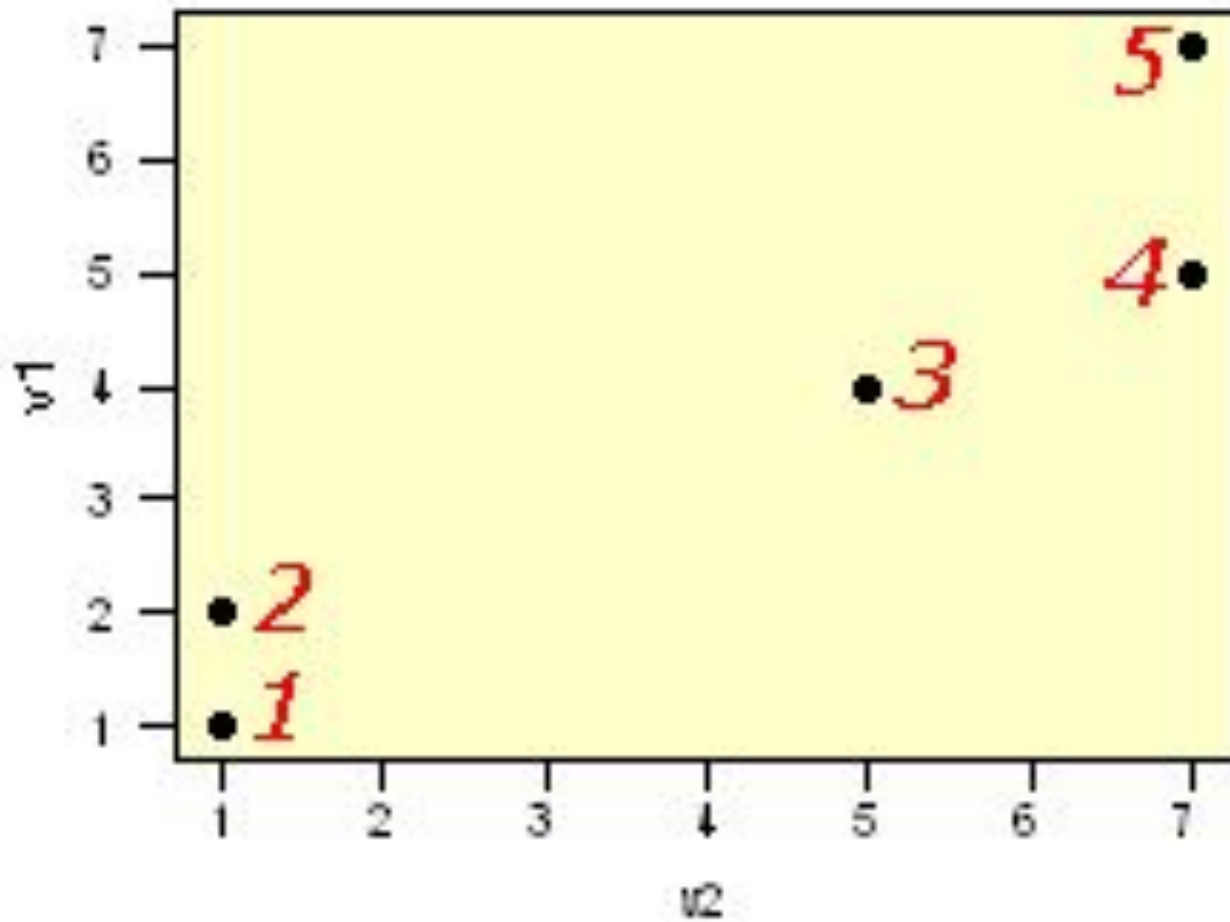
# Начинающим рекомендуем

- - – метод Варда;
  - – метод ближнего соседа (Complete linkage clustering);
  - – среднее невзвешенное расстояние (Average linkage clustering).
-

---

# Алгоритм кластерного анализа

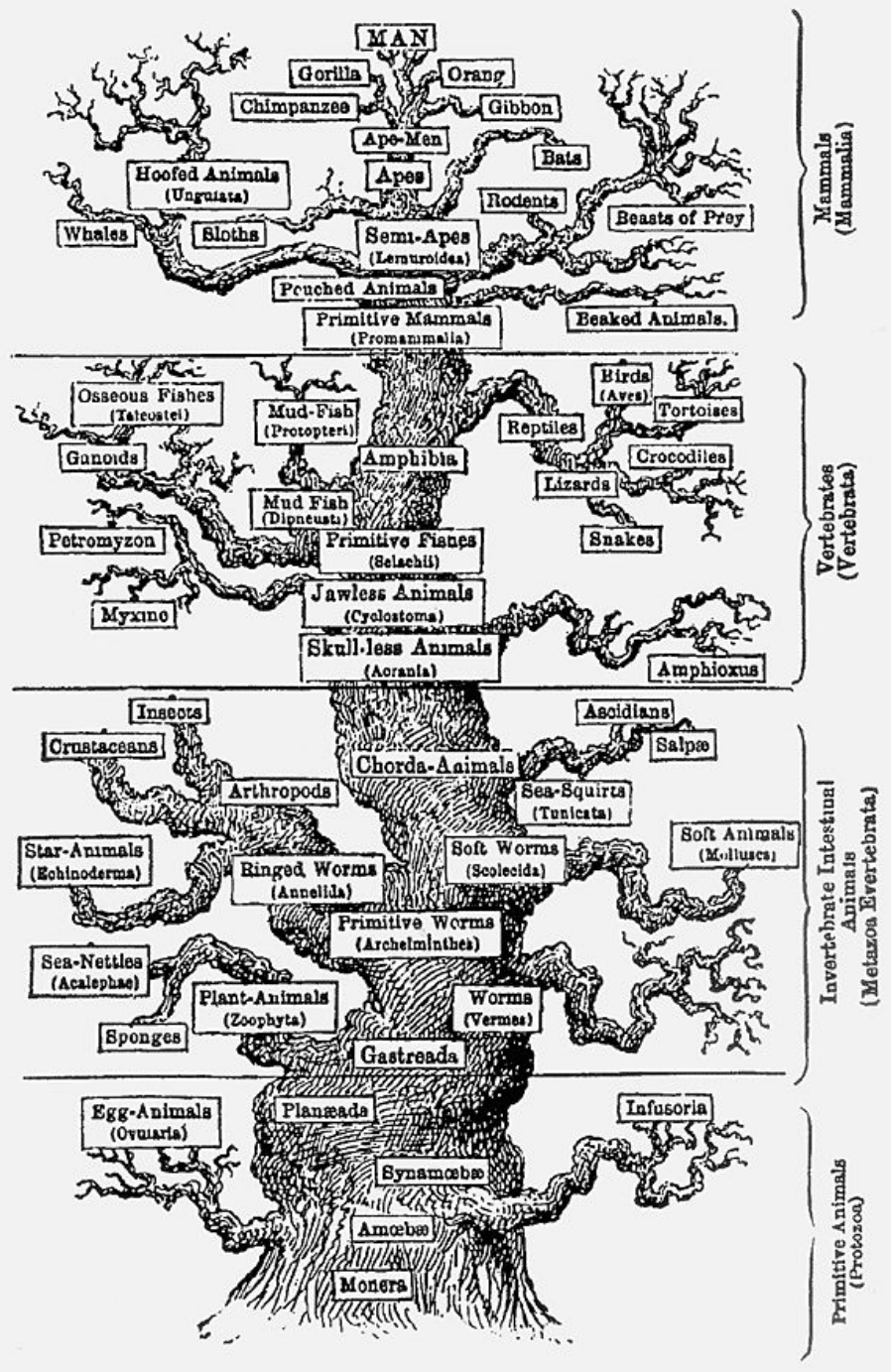
- Разберемся с процедурой иерархического кластерного анализа на примере
-



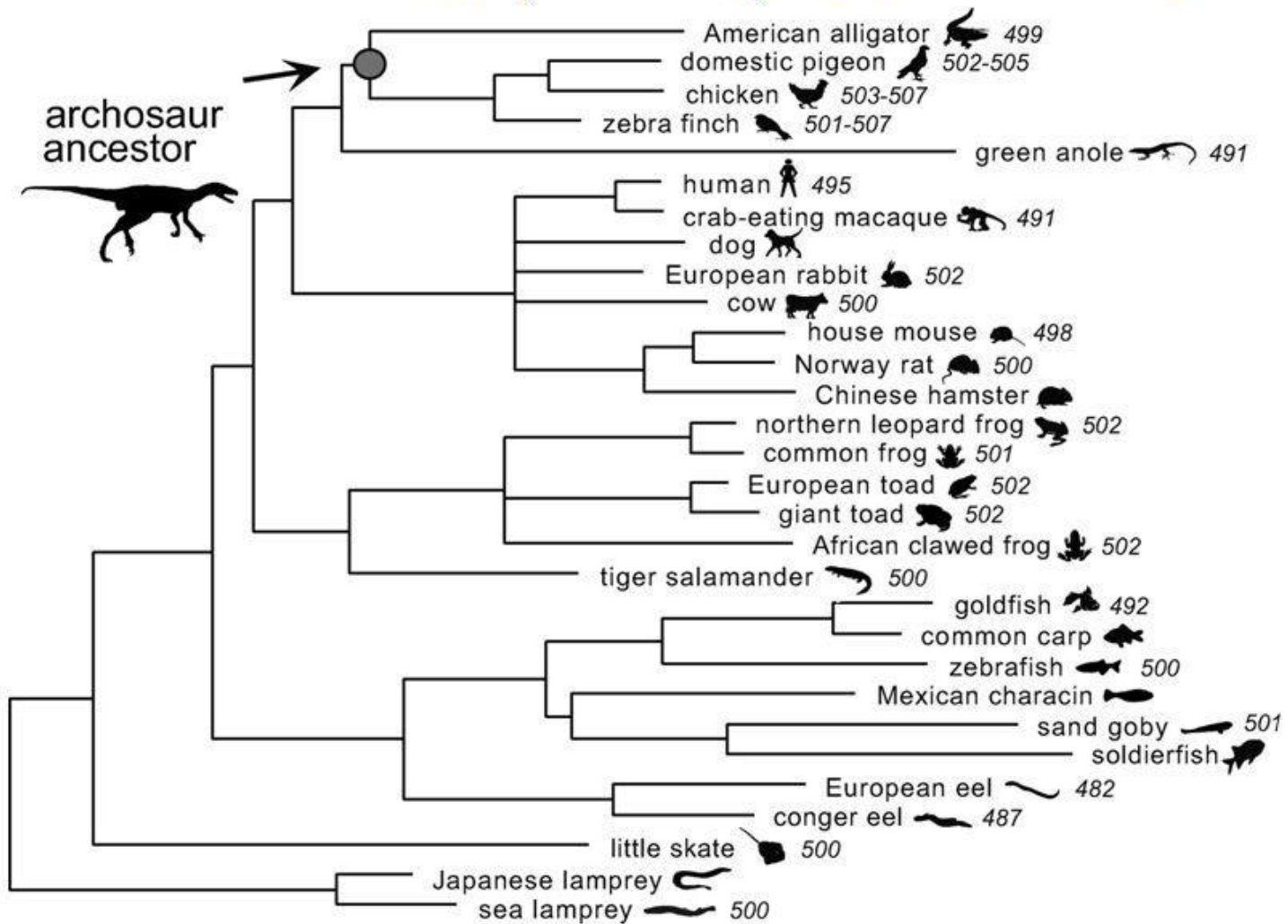
- 
- Алгоритм построения дендрограммы
-



- 
- Ernst Haeckel
  - Tree of Life
  - The Evolution of Man (1879)
  - 
  - Но он не был первым...
  - Древо Порфирия (300+ год)
-



# Типичное дерево (родопсины)

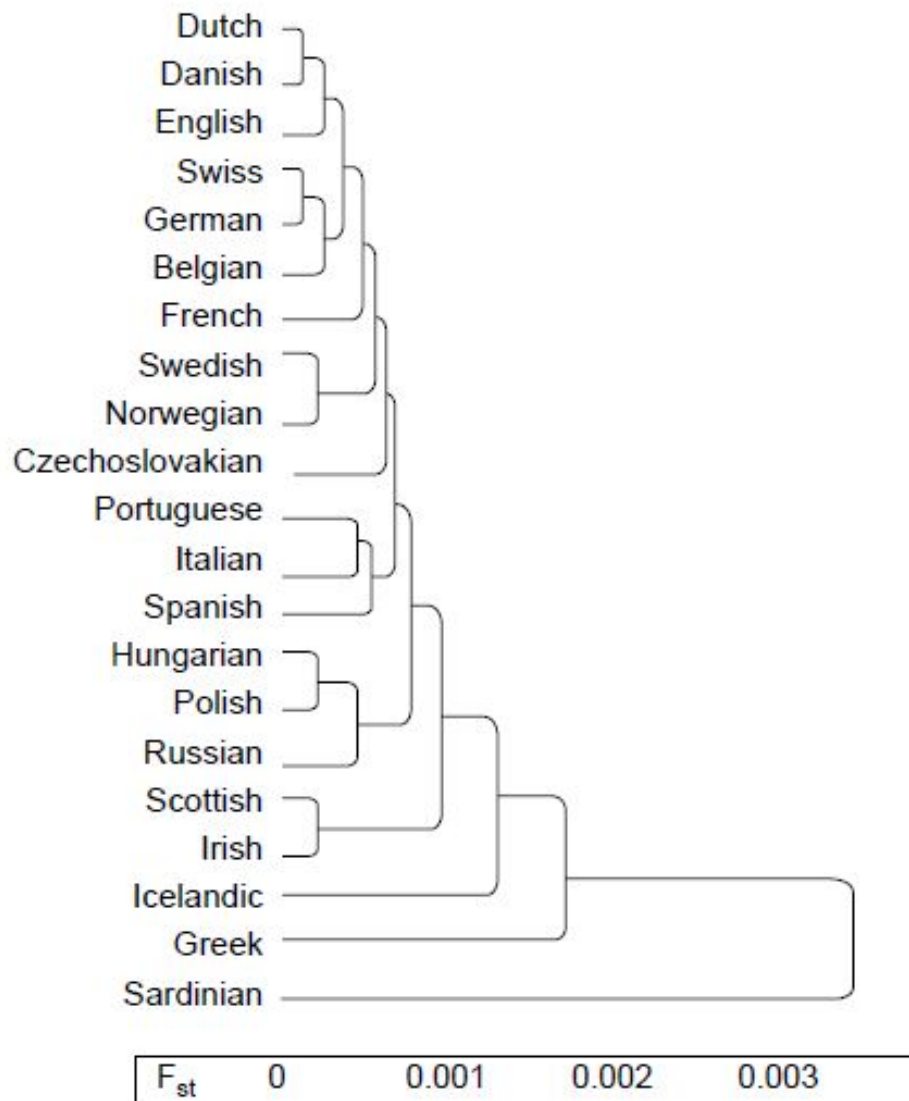


Recreating a Functional Ancestral Archosaur Visual Pigment

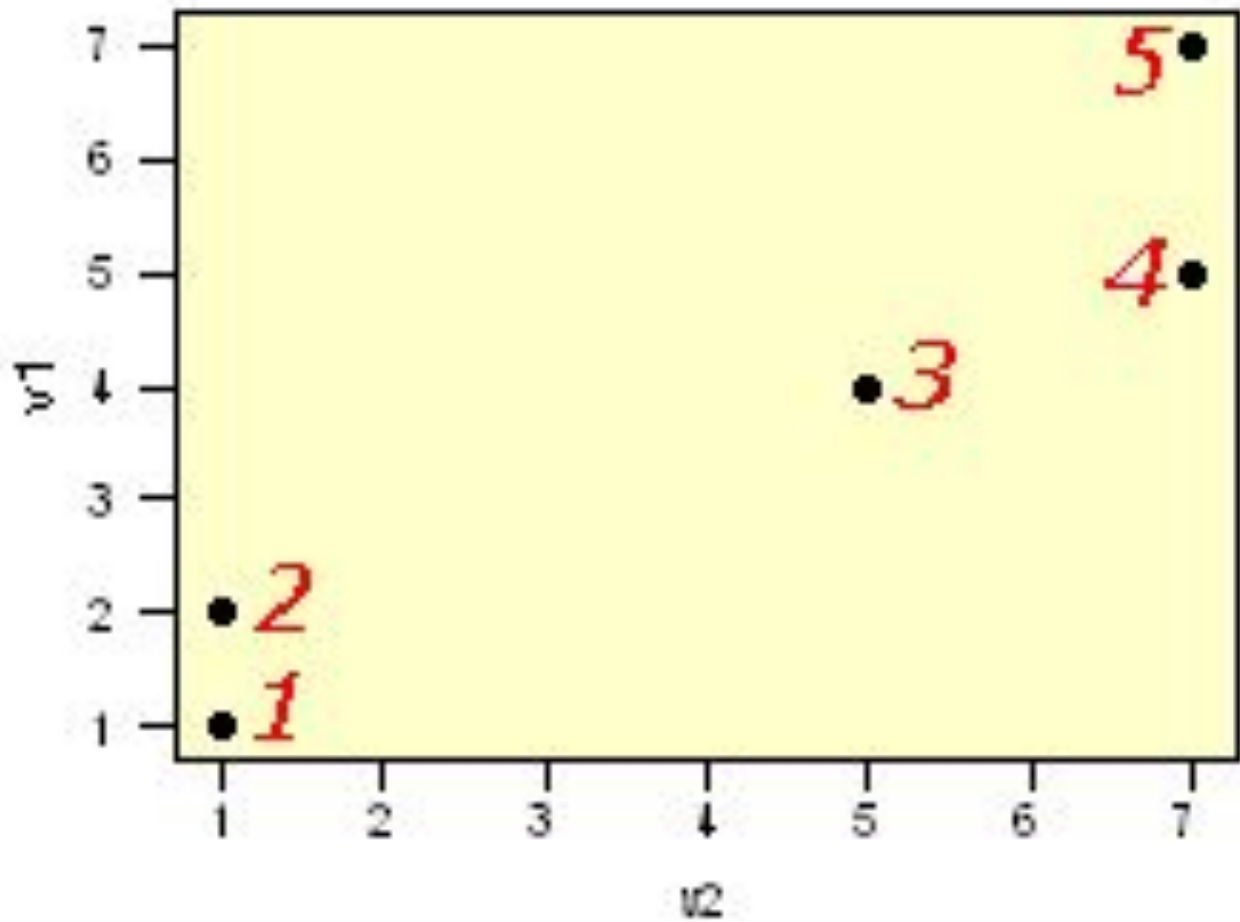
Belinda S. W. Chang,\* Karolina Jönsson,\* Manjia A. Kazmi,\* Michael J. Donoghue,† and Thomas P. Sakmar\*

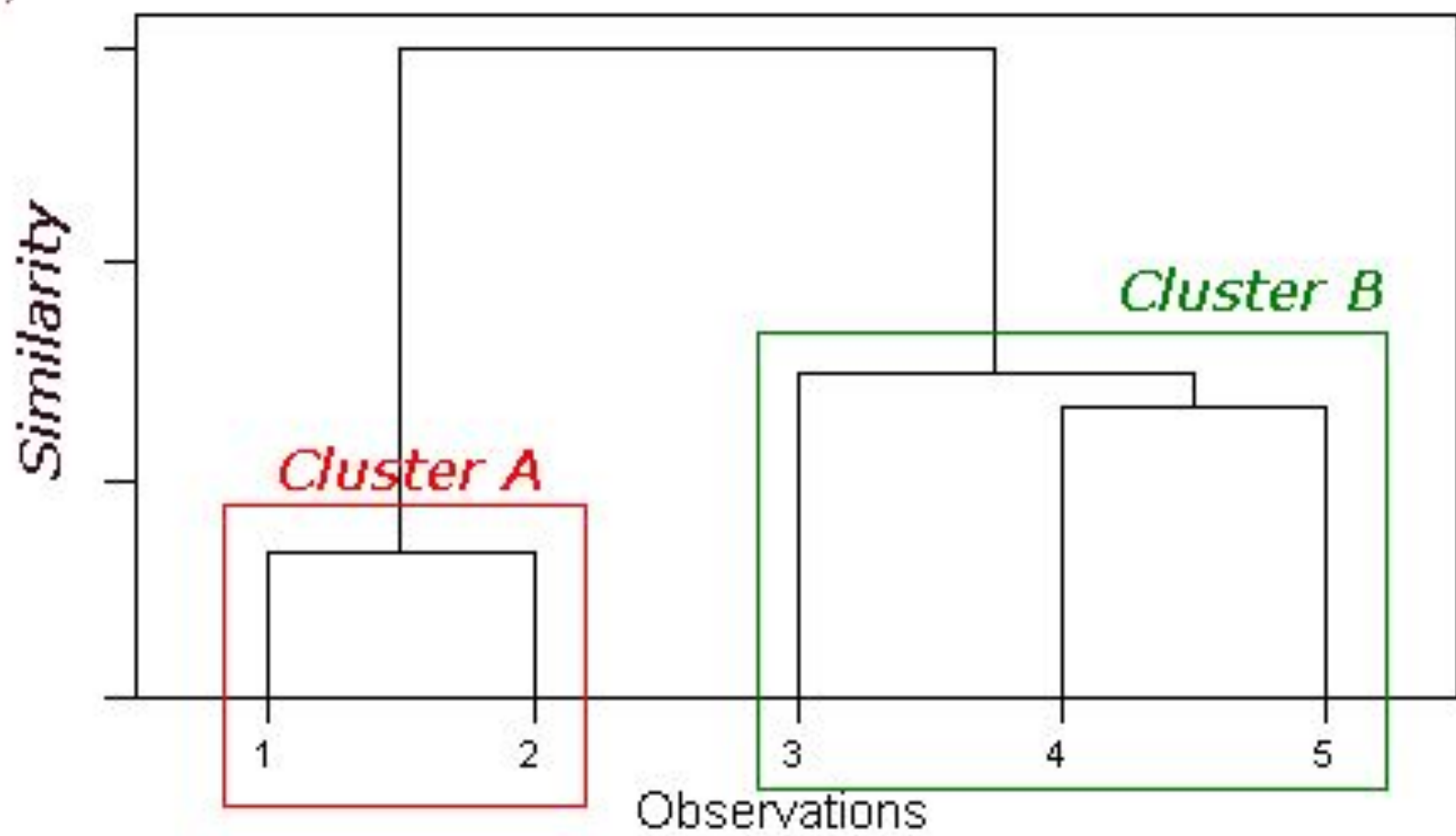
*Mol. Biol. Evol.* 19(9):1483–1489, 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

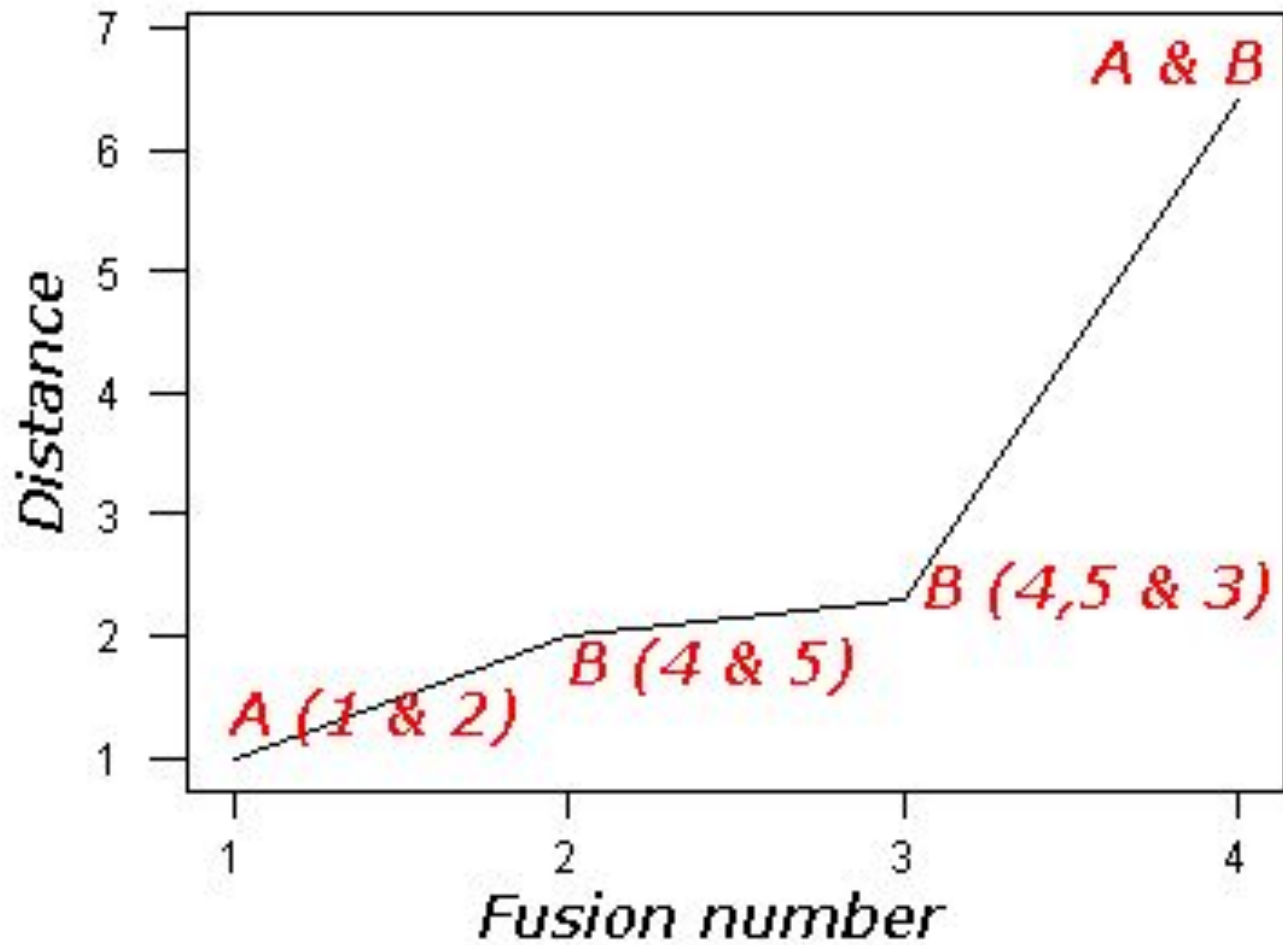


**Fig. 5.2** Genetic tree of European populations from genetic distances ( $= F_{st}$ ) between populations, based on 88 genetic polymorphisms from data in Cavalli-Sforza et al. (1994)



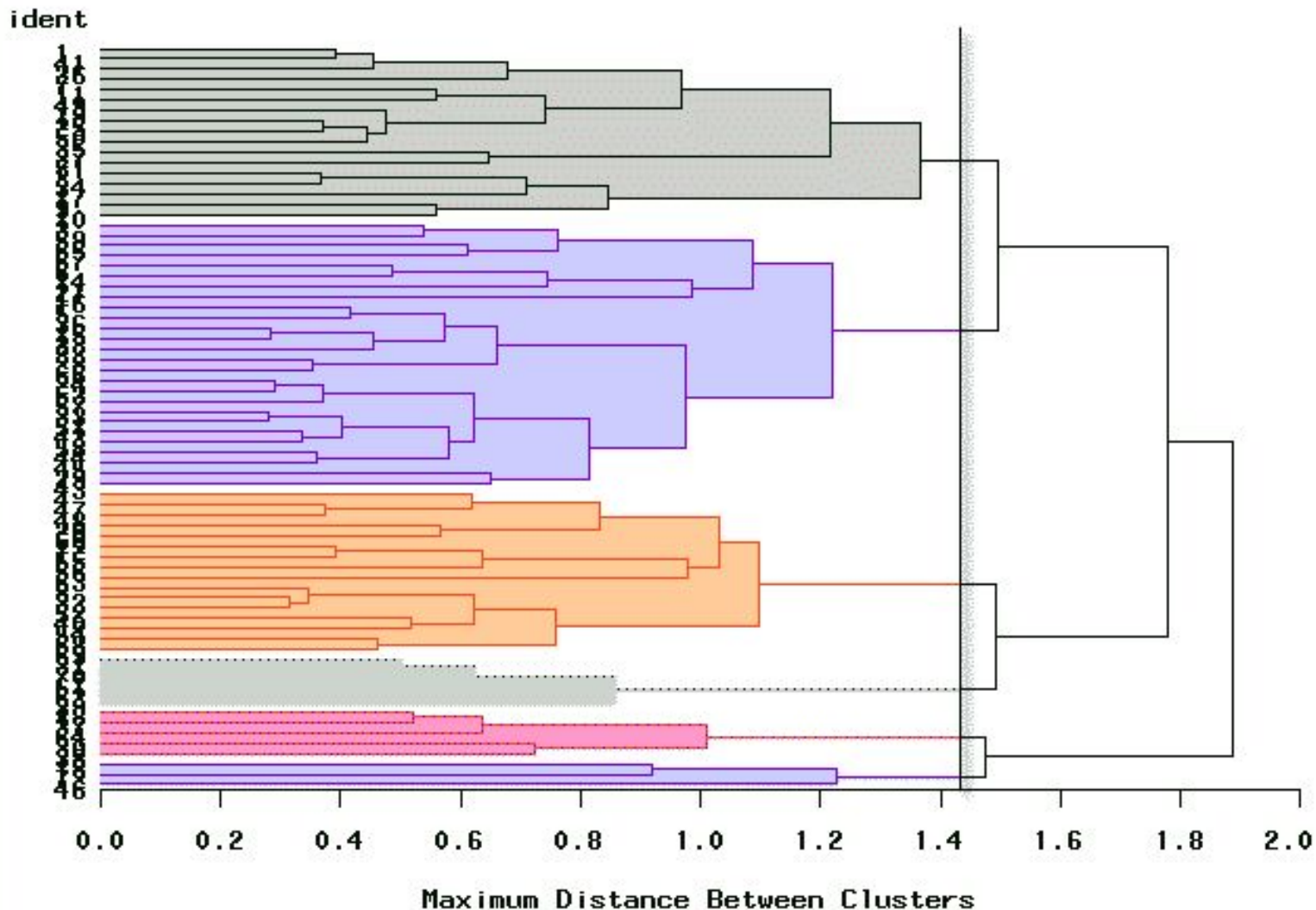


# каменистая осыпь / ЛОКОТЬ



# Где на дендрограмме кластеры?

Cluster Analysis — Woodyard Hammock — Complete Linkage

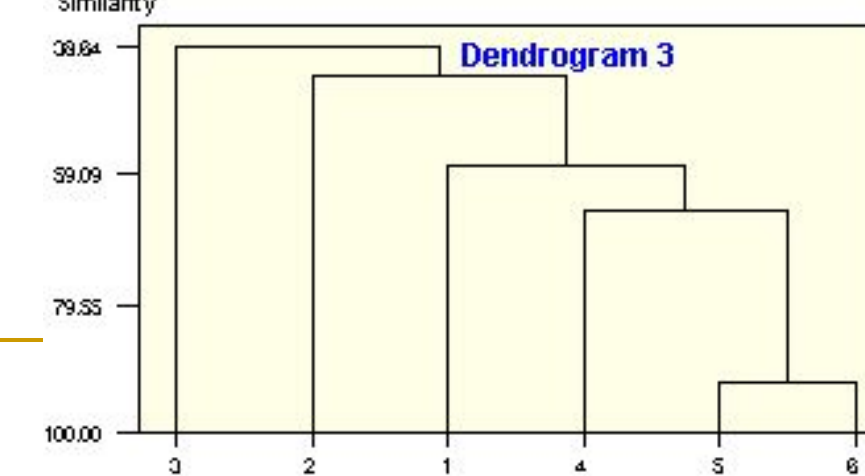
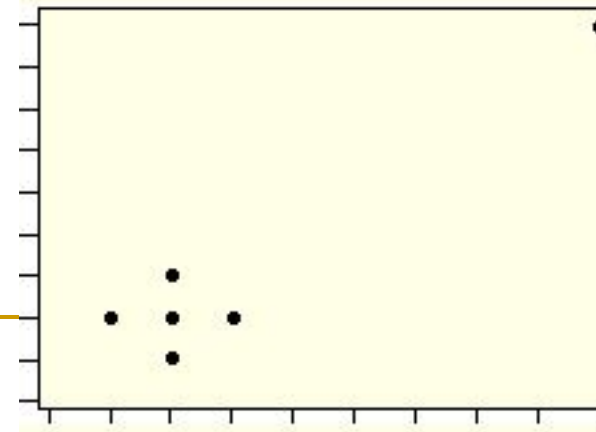
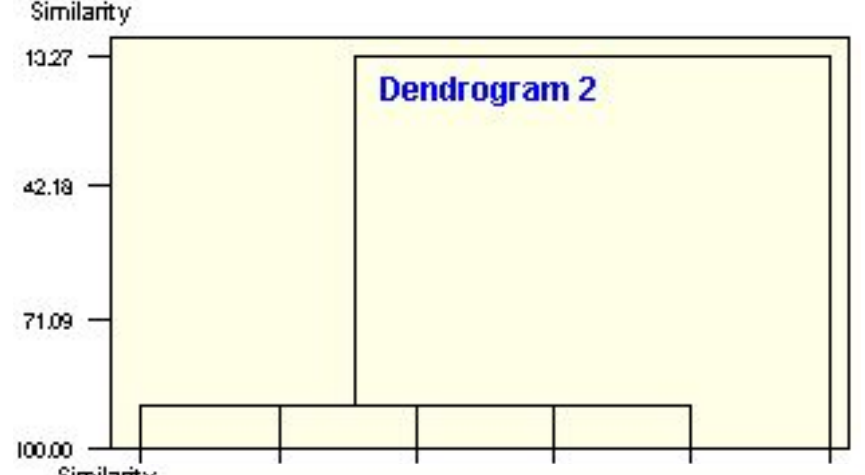
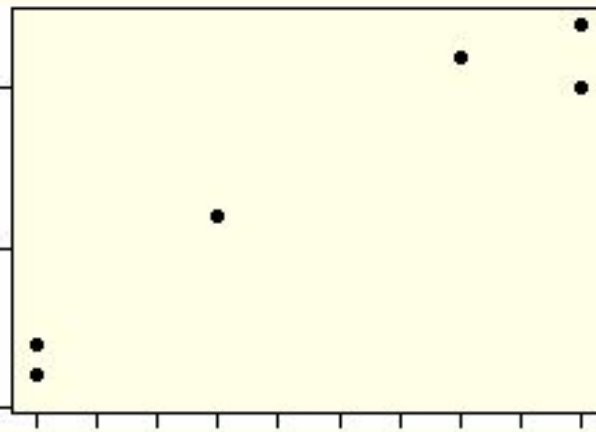
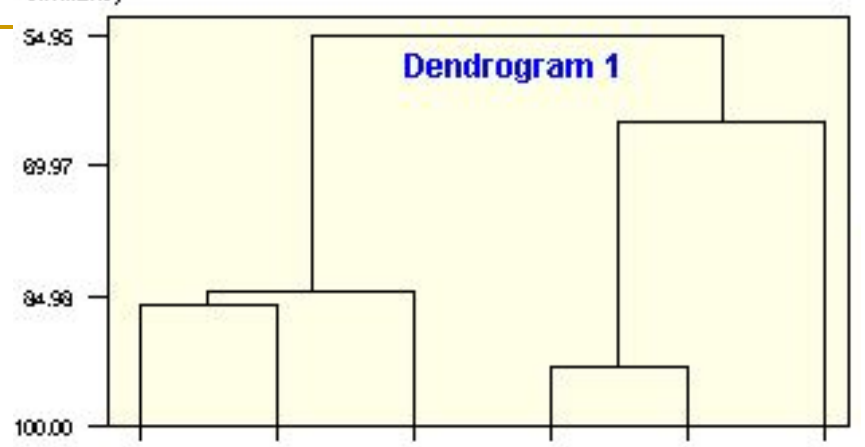
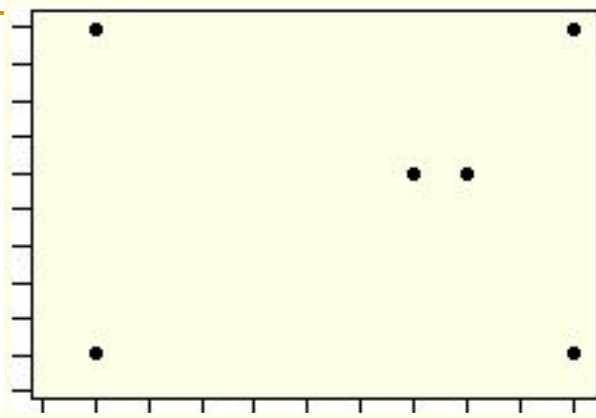




---

# Упражнение

- Разбить на пары:
  - Каждой диаграмме рассеивания поставить в соответствие дендрограмму
-



---

# Участие аналитика

1. Отбор переменных
  2. Метод стандартизации
  3. Расстояние между кластерами
  4. Расстояние между объектами
-

---

# Отбор переменных

- 1. Какие переменные будут использоваться при анализе?
  - Все?
  - Как влияет цвет глаз покупателя на средний объем выпиваемого пива?
  - Распознавание танков
-

---

# С другой стороны

- если нам неизвестны зарплаты/доходы покупателей, но для каждого из них известны профессия, образование и стаж работы, исключение этих трех переменных влечет за собой исключение из рассмотрения платежеспособность покупателей.
  - Если классифицируются школы, и не включены ни переменная «число школьников», ни переменная «число учителей», то кластеры будут формироваться без учета размера школ.
-

---

# Вывод

- Правильный выбор переменных очень важен.
  - Критерием при отборе переменных для анализа является в первую очередь ясность интерпретации полученного результата, во вторую – интуиция исследователя.
-

---

# Надо ли стандартизировать переменные?

- Правило для новичка:
  - если Вы не знаете, стандартизировать или нет, стандартизируйте.
-

# Надо стандартизировать

5296782.7	0.5	1
7400381.4	0.7	0
9362870.2	0.1	0
7594038.5	0.4	0
6455034.1	0.4	1



---

# Стандартизация

- Для каждого столбца.
  - Линейное преобразование
    1. Максимальное значение = 1, минимальное = 0 (-1)
    2. z-метки. Среднее равно 0, выборочная дисперсия равна 1.
-

- 
- Иногда решением будет преобразование данных



- 
- Если кластеров нет
  - Они все равно будут найдены
-



---

# Результаты кластерного анализа нуждаются в интерпретации

- какой вариант кластеризации даст лучшие результаты?
  - тот, который вы смогли понять и проинтерпретировать
-

---

## Еще раз об участии аналитика

Иерархический кластерный анализ требует вдохновенного выбора способа подсчета расстояния между объектами и расстояния между кластерами. Кроме того, надо угадать число кластеров. Потом останется неясной геометрия кластеров. Таким образом, многое надо угадать и осмыслить. Не всегда это удастся.

---

---

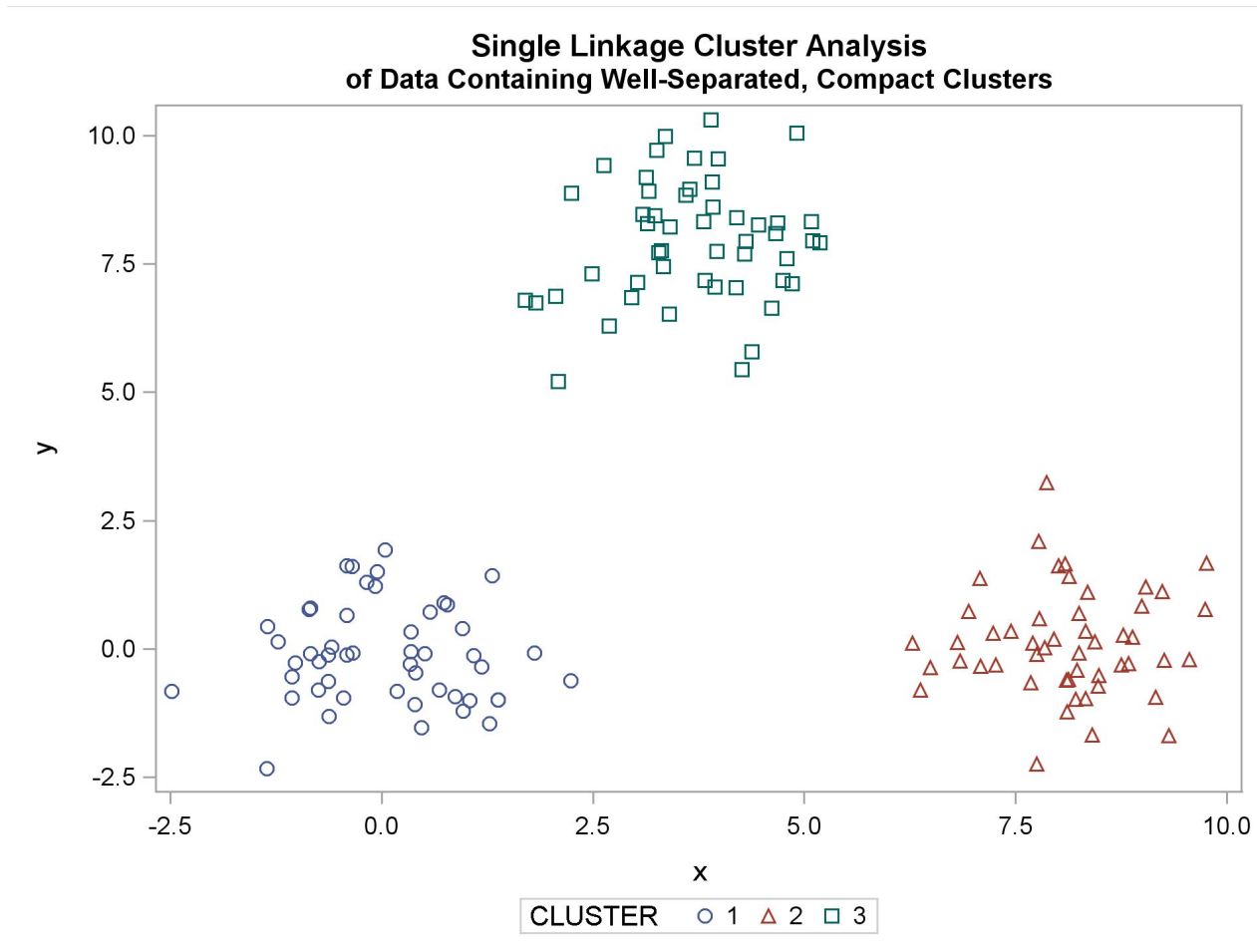
# Типы кластеров

- Шаровые
  - Ленточные
  - ...
-

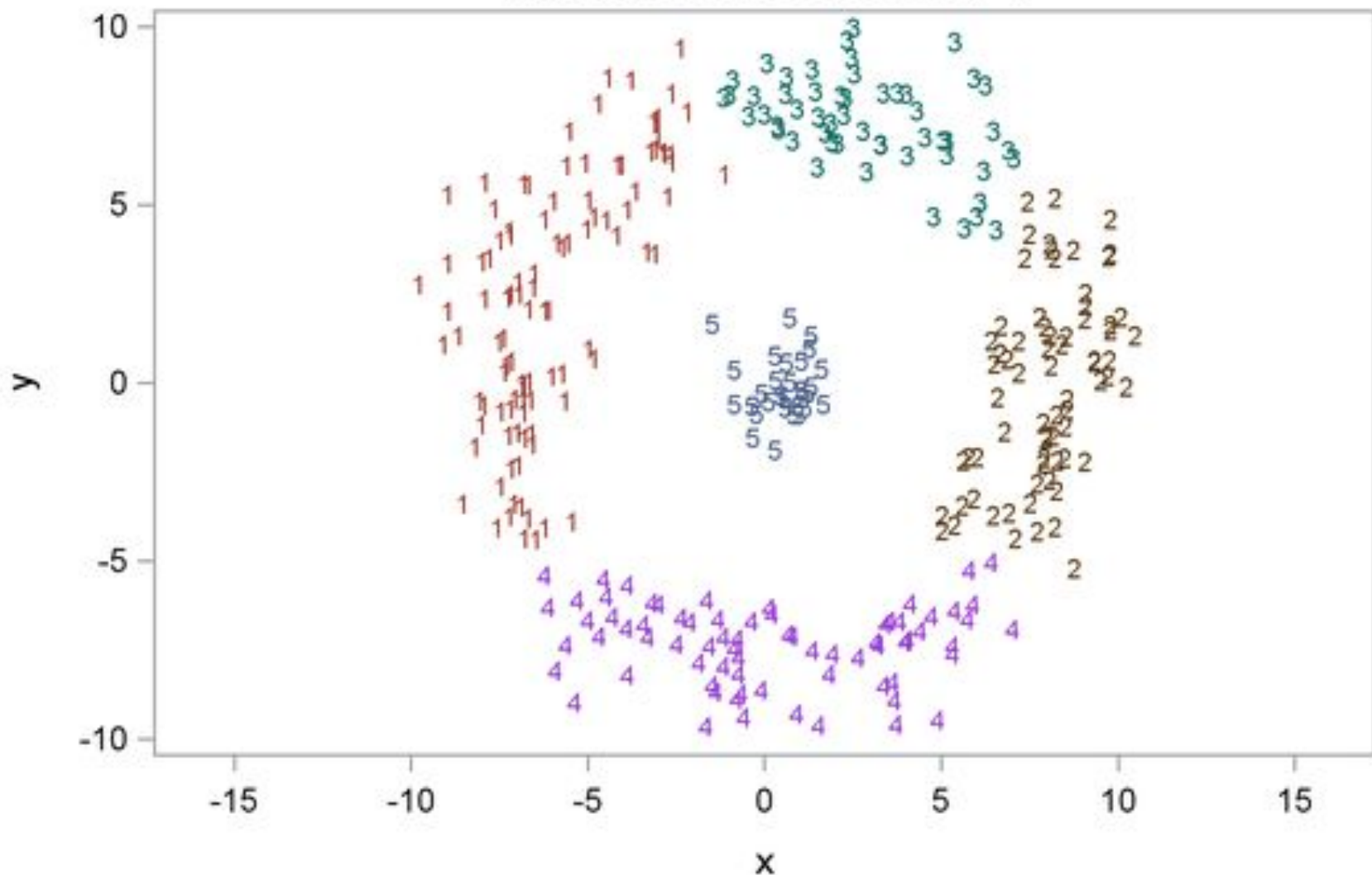
- 
- Выбор расстояния между кластерами
-



# Выраженные кластеры – все равно какой метод

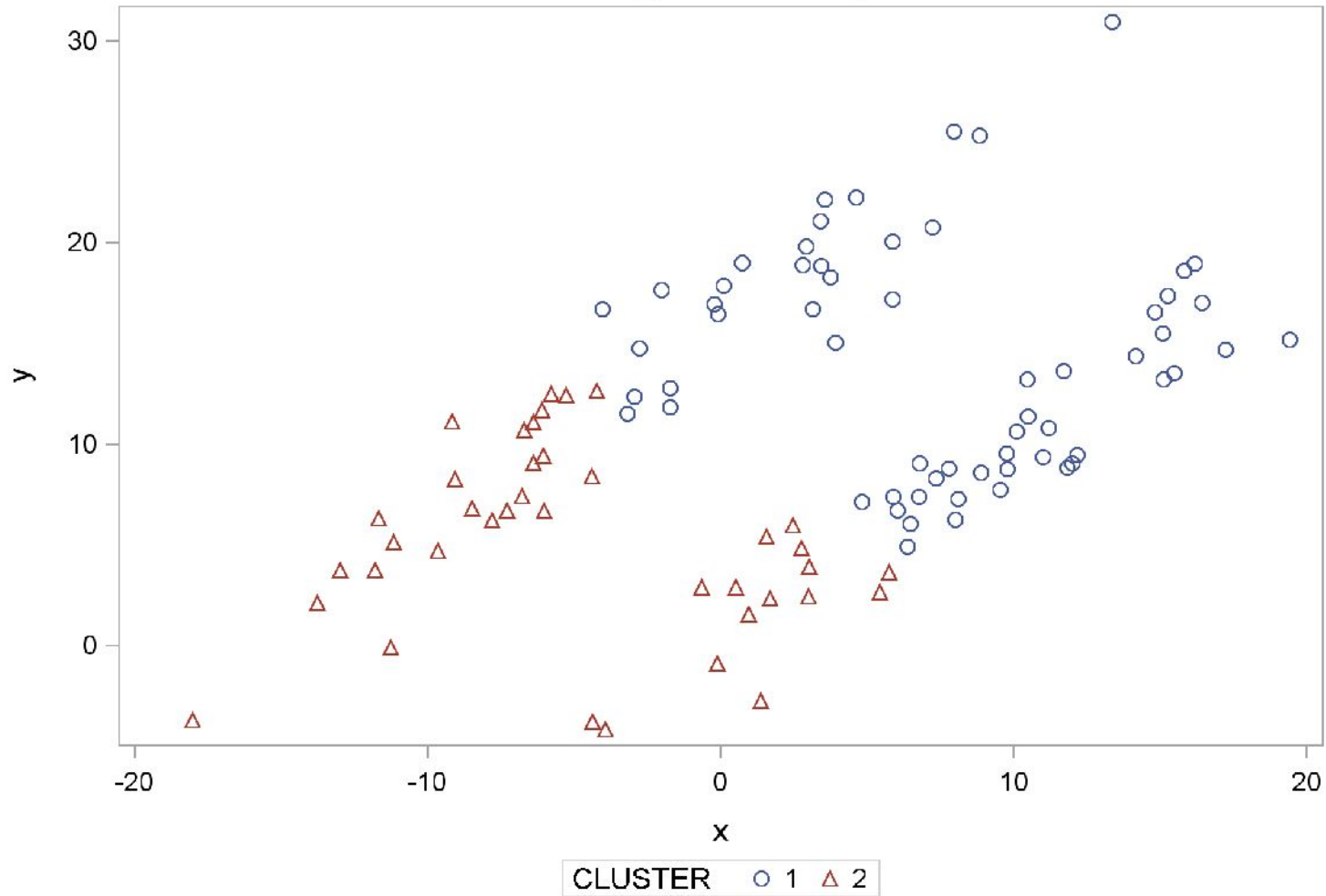


**Modeclus Analysis with the JOIN= option**  
**A Normal Cluster Surrounded by a Ring Cluster**  
**Number of Clusters Joined=1**



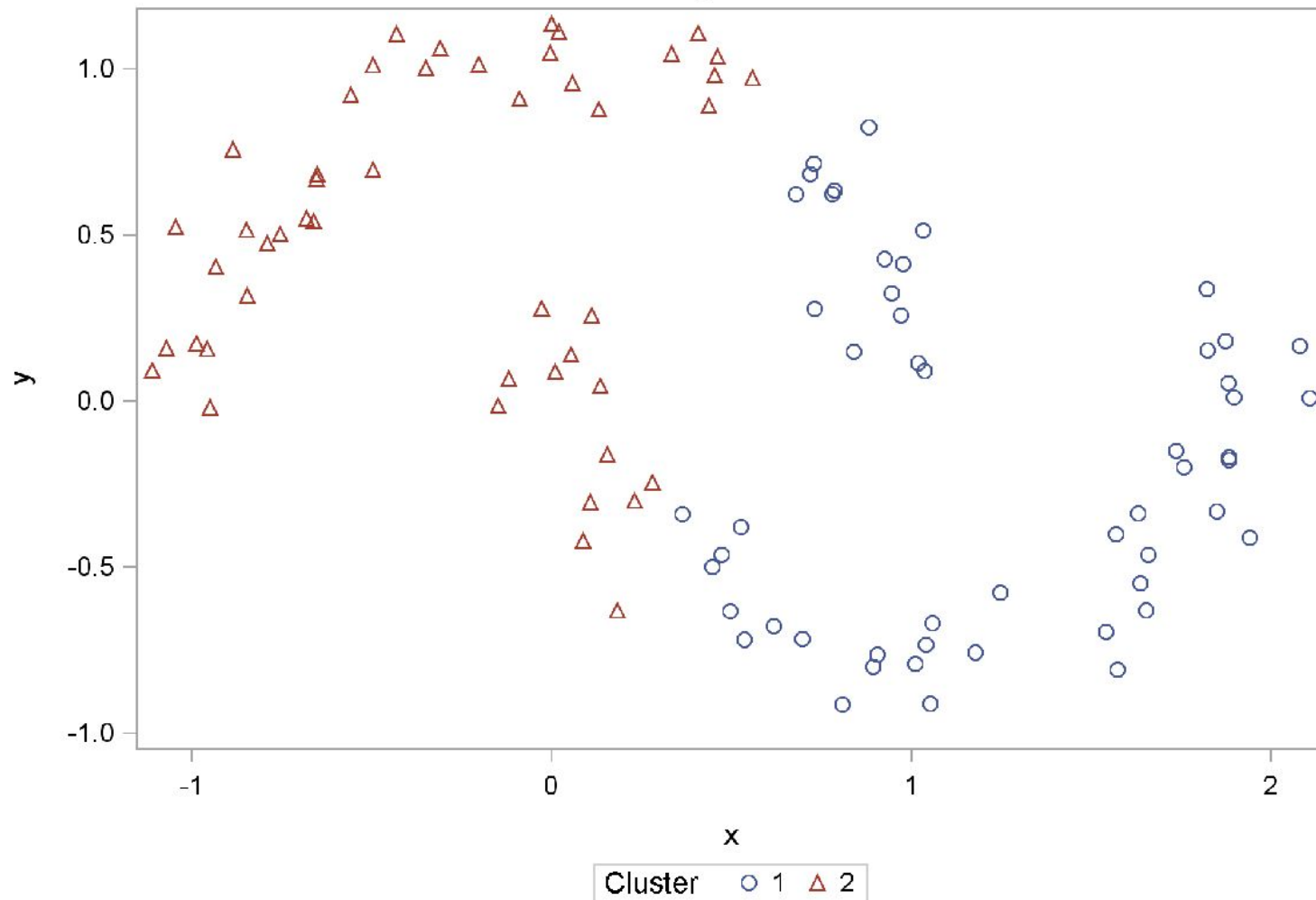
# Какой метод будет лучше?

**Average Linkage Cluster Analysis  
of Data Containing Parallel Elongated Clusters**



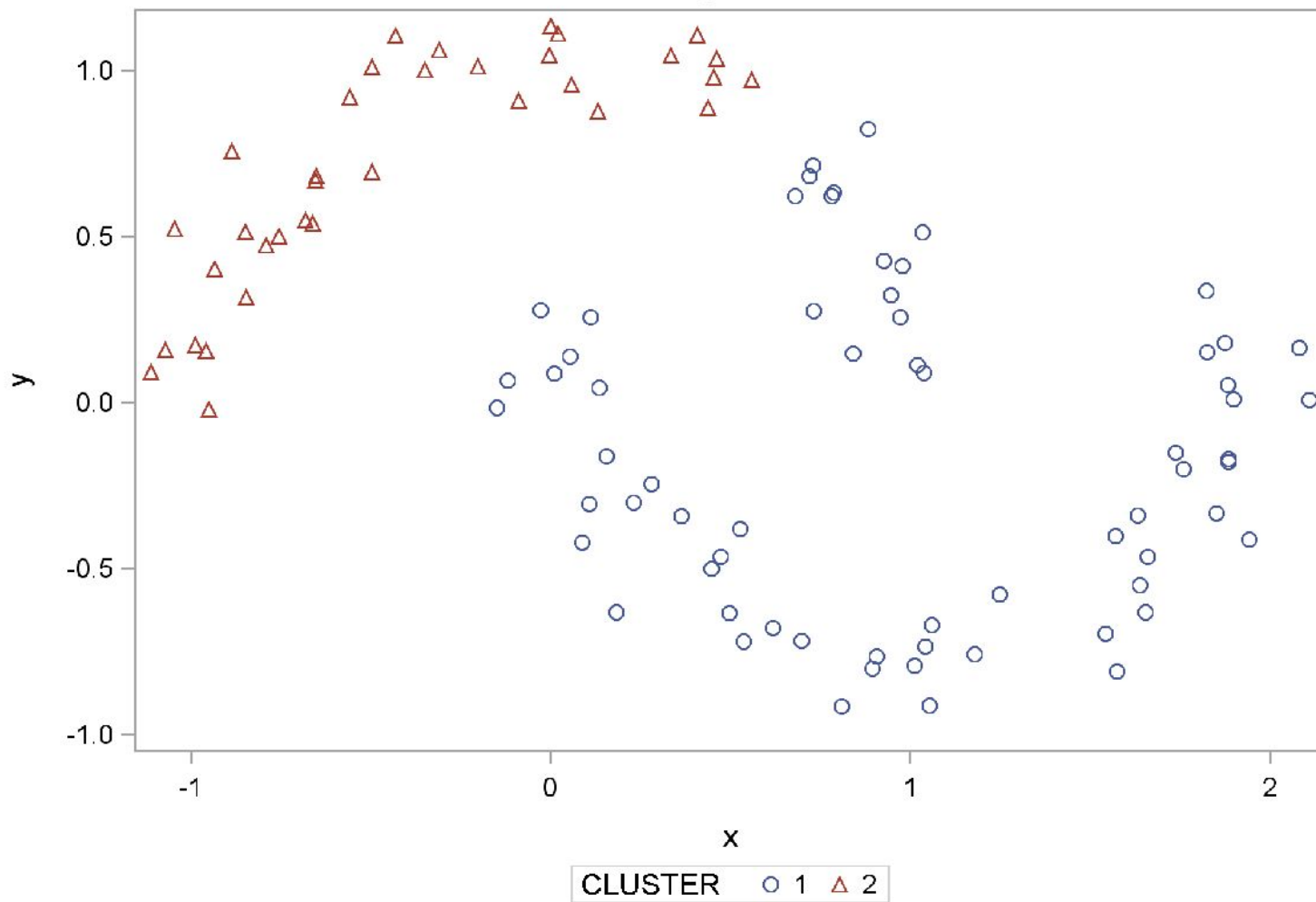
# Неудача

## FASTCLUS Analysis of Data Containing Nonconvex Clusters



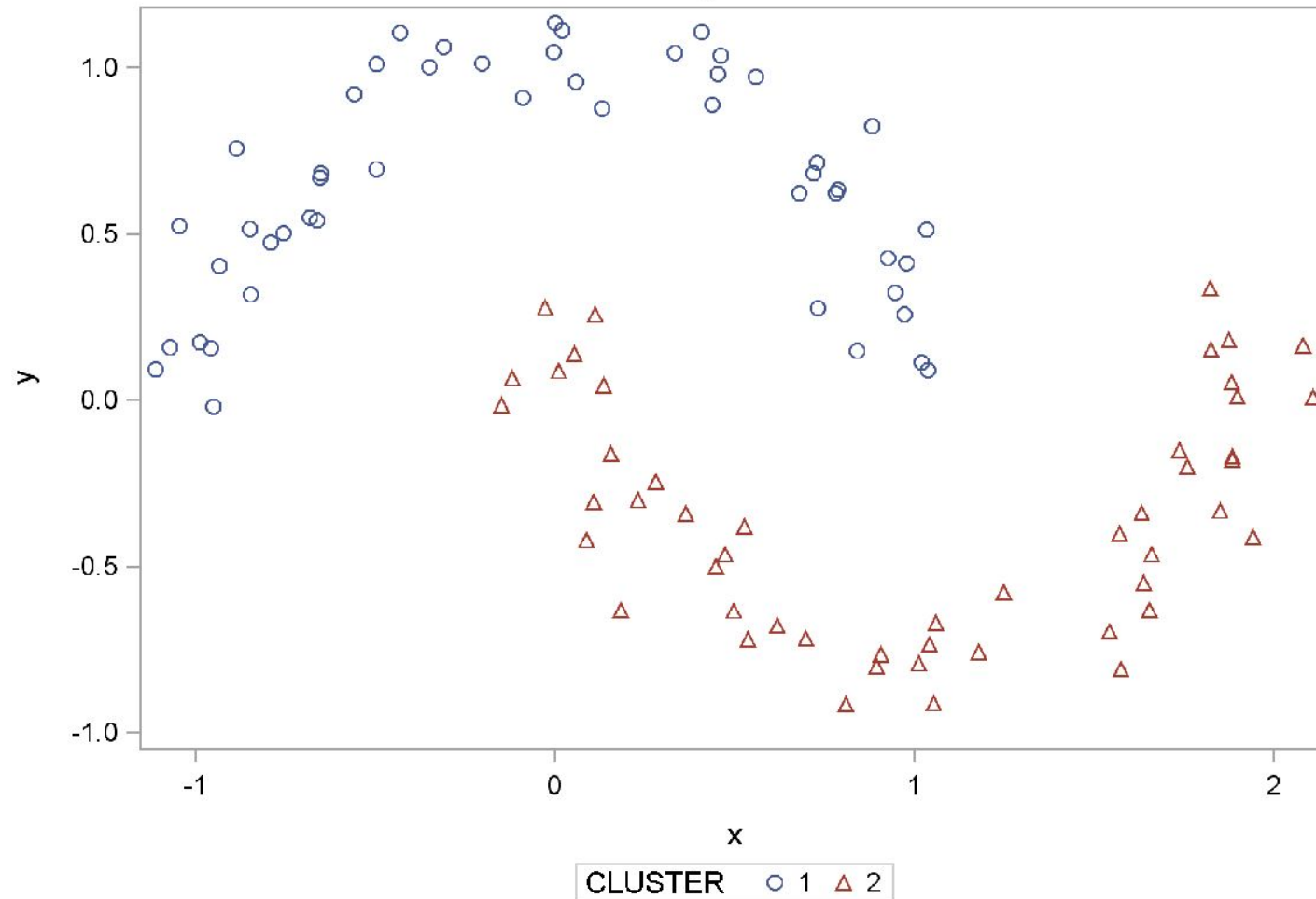
# Неудача

Centroid Cluster Analysis  
of Data Containing Nonconvex Clusters



# Метод ближайшего соседа

Two-Stage Density Linkage Cluster Analysis  
of Data Containing Nonconvex Clusters



---

# Пример

- Сегментация потребителей безалкогольных напитков

- 
- Компания провела опрос с целью выявить, какие напитки предпочитают респонденты. Опрошенные указывали, какие напитки из предложенного списка они пьют регулярно.
-

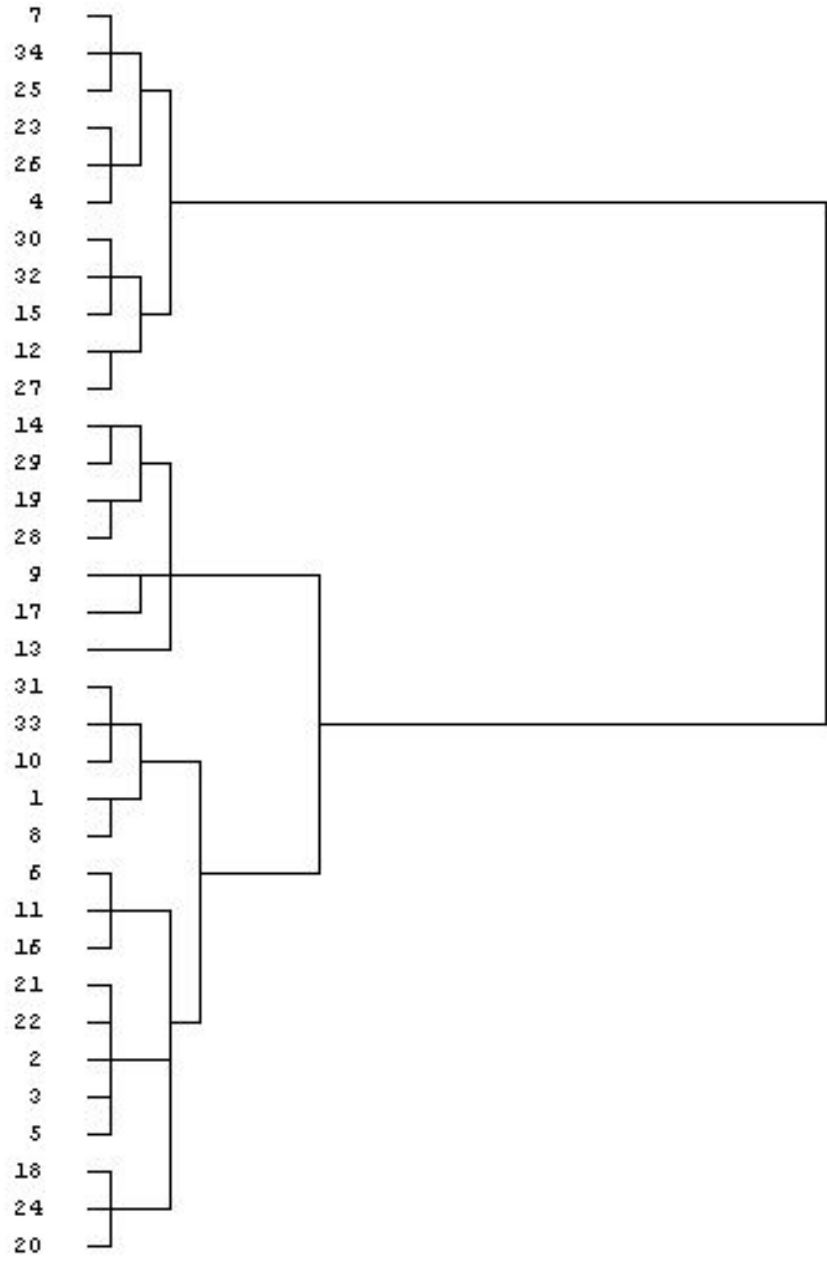


---

# В списке присутствовали

- Соса-Cola,
  - диетическая Соса-Cola,
  - Pepsi-Cola,
  - диетическая Pepsi-Cola,
  - 7-Up
  - диетический 7-Up,
  - Спрайт,
  - минеральная вода
-

C A S E 0 5 10 15 20 25  
Label Num +-----+-----+-----+-----+



---

# Решение для трех кластеров

- перечисляя сверху вниз на дендрограмме,
  - В верхний кластер войдут респонденты с номерами от 7-го до 27-го,
  - в средней группе – от 14-го до 13-го,
  - в нижний – от 31-го до 20-го.
-

- 
- R нумерует кластеры не сверху вниз!
  - Как ему захочется!
-

# 1 кластер 16 наблюдений

COKE	15
D_COKE	4
D_PEPSI	1
D_7UP	0
PEPSI	16
SPRITE	5
TAB	0
SEVENUP	5

## 2 кластер 11 наблюдений

COKE	0
D_COKE	11
D_PEPSI	6
D_7UP	6
PEPSI	0
SPRITE	0
TAB	10
SEVENUP	0

## 3 кластер 7 наблюдений

COKE	5
D_COKE	2
D_PEPSI	1
D_7UP	1
PEPSI	0
SPRITE	6
TAB	1
SEVENUP	4

---

# Потребление протеинов в Европе

- Переменные
  - redmeat Мясо
  - whitemeat Птица
  - eggs Яйца
  - milk Молоко
  - fish Рыба
  - cereals Хлебо-булочные
  - starch Крахмал: картофель, макароны
  - nuts Орехи
  - fruits\_v Фрукты и овощи
-

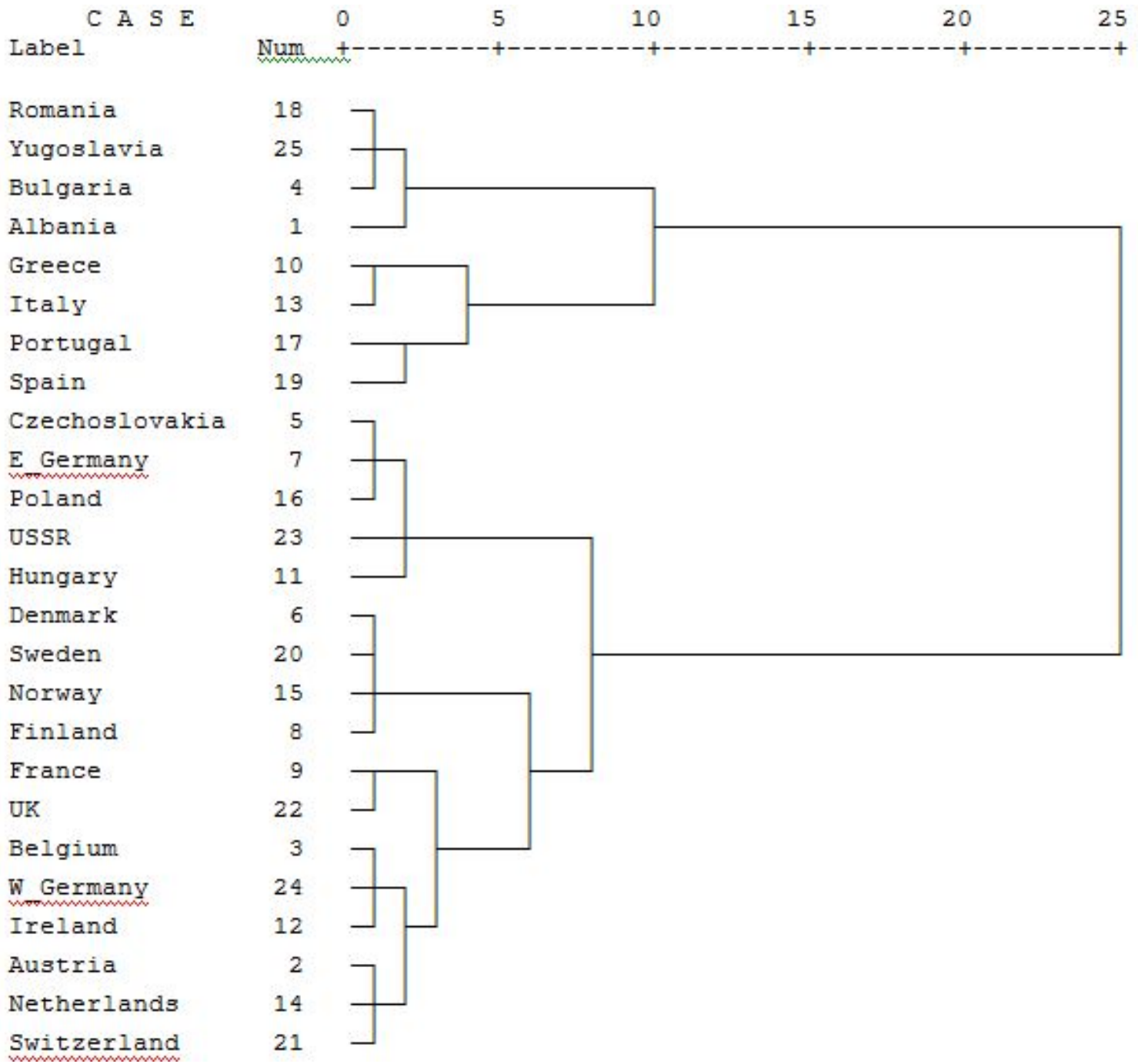


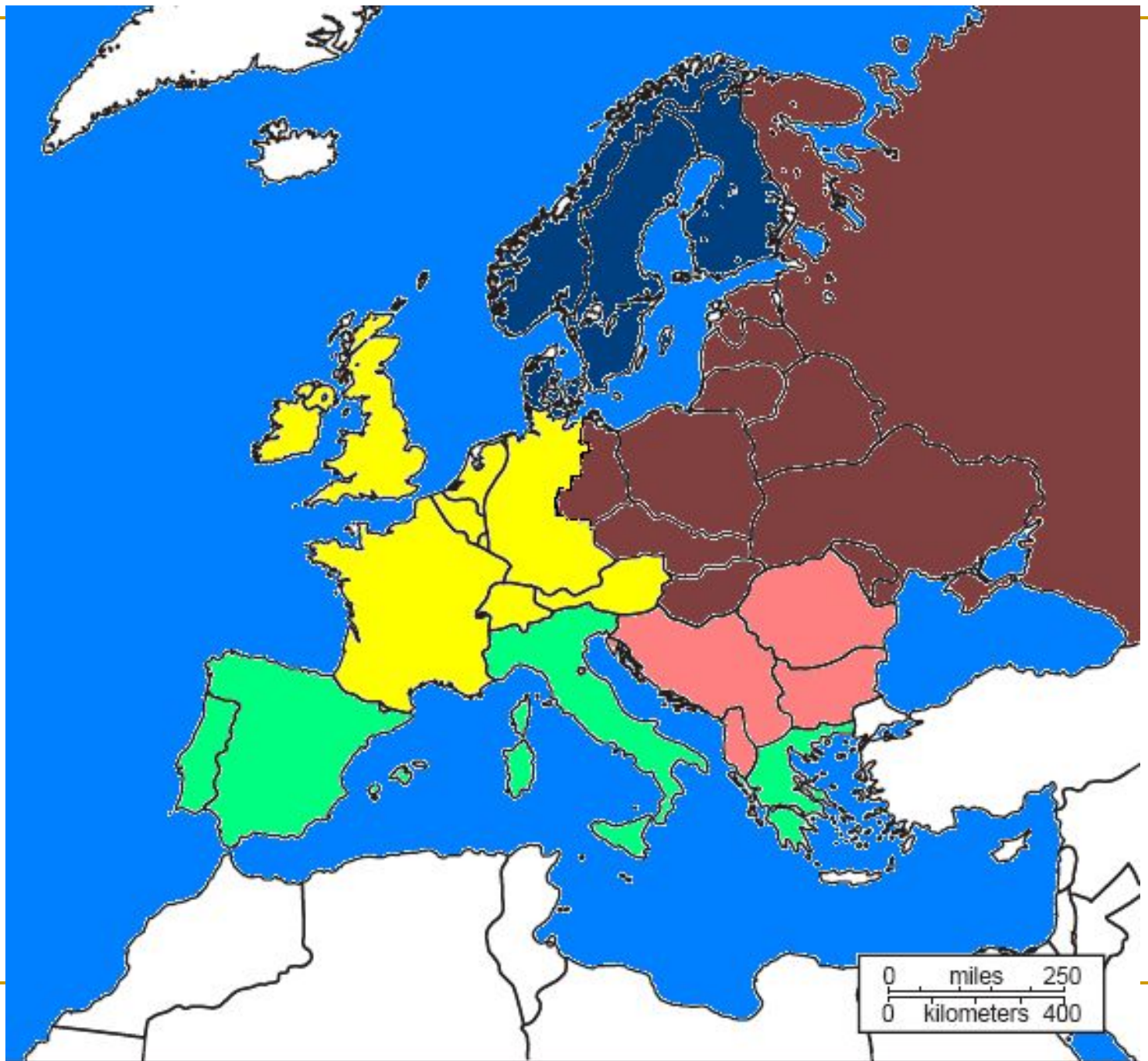
---






## Задача:

- Разбить страны на группы.
  - Надо ли проводить стандартизацию?
  - Как отличаются кластеры?
  - (Использовалось решение Тропинина.)
-

- 
- Стандартизация обязательна, так как средние значения некоторых переменных отличаются в десятки раз.
  - Из всех методов иерархического кластерного анализа наиболее понятную картину дал
  - метод Варда + стандартизация  $[0, 1]$
-





- 
1.  Страны Балканского полуострова (кроме Греции), социализм
  2.  Западная Европа
  3.  Восточная Европа (остальные страны соц.лагаря)
  4.  Северная Европа
  5.  Страны Южной Европы, капитализм
-

- 
- особенности питания зависят от
  - географического положения и от
  - экономического строя,
  
  - что вполне естественно
-

# Далее, сравниваем потребление в разных кластерах

- 1 кластер: большое потребление злаков и орехов (Pulses, nuts, and oil-seeds);
- маленькое потребление мяса (Red meat, White meat), рыбы, крахмалистых продуктов (Starchy foods) и яиц.
- 2 кластер: большое потребление мяса, яиц, молока; небольшое потребление злаков и орехов.
- 3 кластер: большое потребление птицы (White meat), крахмалистых продуктов; небольшое потребление орехов.
- 4 кластер: большое потребление яиц, молока, рыбы; маленькое потребление злаков, орехов, фруктов и овощей.
- 5 кластер: большое потребление рыбы, орехов, фруктов и овощей; маленькое потребление птицы.

---

Мы узнали что-то новое?

Или результат тривиальный?

- Англия и Франция
  - Социализм или капитализм: две германии в разных кластерах
  - Два социализма
-