Иерархическая кластеризация данных

Иерархическая кластеризация

- Пришло время изучить ещё один метод кластеризации данных!
- Иерархическая кластеризация часто применяется в биологии, позволяя визуализировать кластеры.
- Эта кластеризация помогает определить количество кластеров.

Иерархическая кластеризация

- Обзор раздела:
  - Теория иерархической кластеризации
  - Пример написания кода

- Замечание: В этом разделе не будет проверочных заданий – они будут позже, в разделе про кластеризацию DBSCAN.

Иерархическая кластеризация

Теория

Иерархическая кластеризация

- Как и большинство алгоритмов кластеризации, иерархическая кластеризация опирается на измерение расстояния между “похожими” точками.
- “Похожесть” определяется некоторой метрикой расстояния между точками.

Иерархическая кластеризация

- Затем нужна иерархическая кластеризация?
  - Её легко понять и визуализировать.
  - Она помогает решить, какое количество кластеров выбрать.
  - Можно не выбирать количество кластеров перед запуском алгоритма.

Иерархическая кластеризация

- Затем нужна иерархическая кластеризация?
  - Разделяет точки на **возможные** кластеры:
    - Агломеративный подход:
      - Каждая точка принадлежит кластеру, затем кластеры объединяются.
    - Разделяющий подход:
      - Все точки находятся в одном кластере, затем кластеры делятся на части.

Иерархическая кластеризация

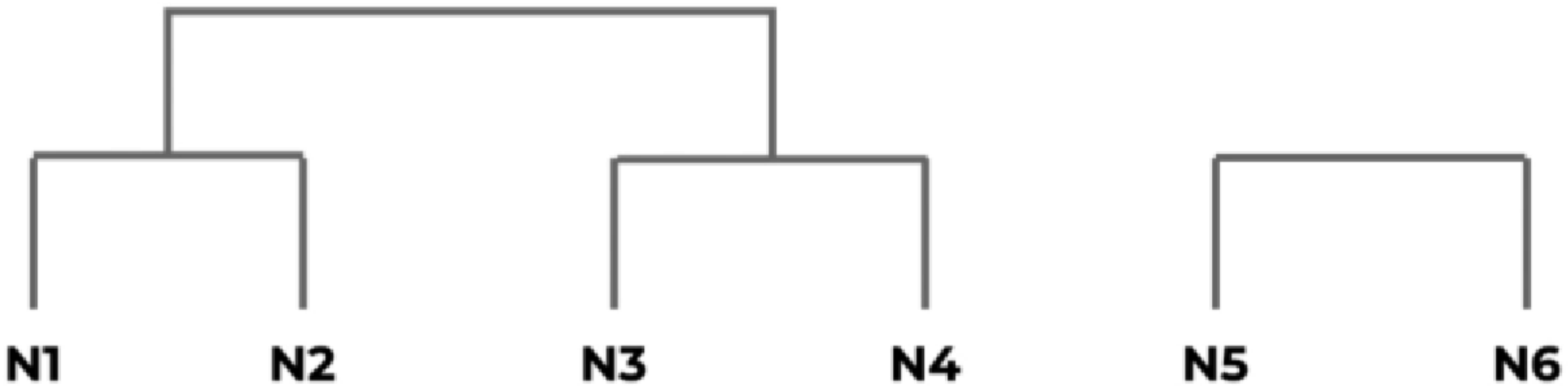
- Иерархическая кластеризация
  - Агломеративный подход:

Иерархическая кластеризация

- Иерархическая кластеризация
  - Агломеративный подход:

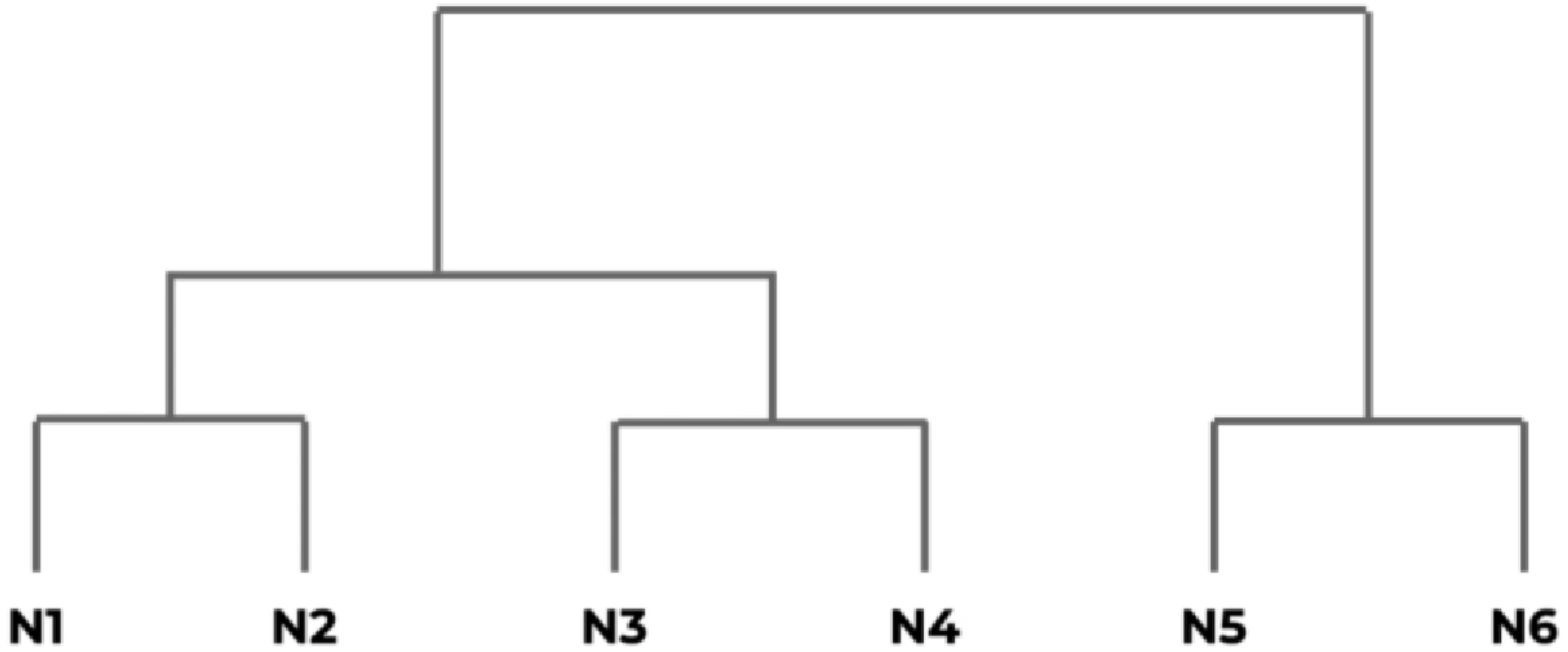
Иерархическая кластеризация

- Иерархическая кластеризация
  - Агломеративный подход:



Иерархическая кластеризация

- Иерархическая кластеризация
  - Агломеративный подход:

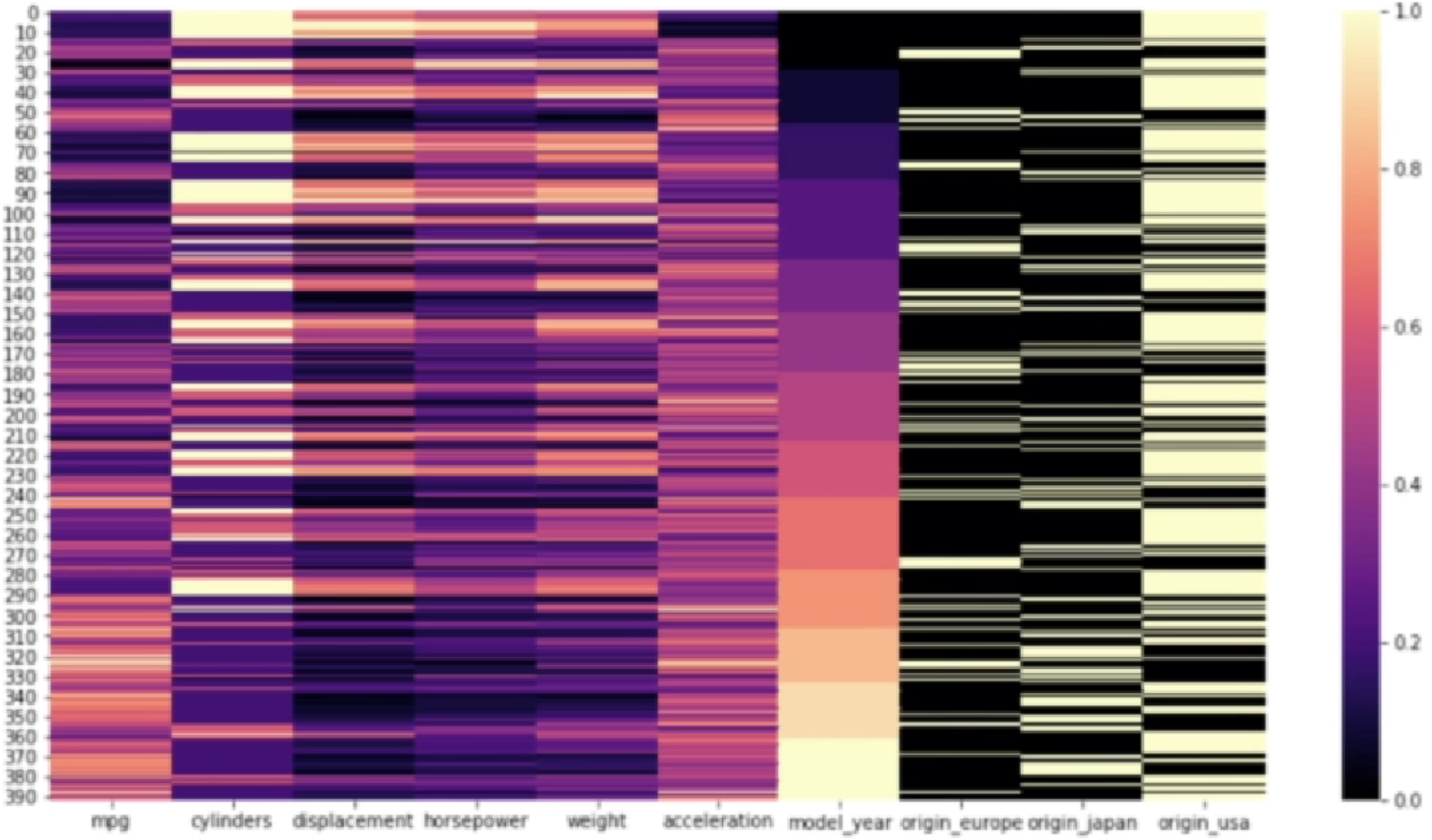


Иерархическая кластеризация

- **Процесс иерархической кластеризации:**
  - Сравниваем точки, находим наиболее похожие друг на друга точки.
  - Объединяем такие точки в кластеры.
  - Сравниваем наиболее похожие друг на друга кластеры, объединяем кластеры.
  - Повторяем шаги до тех пор, пока все точки не окажутся в одном кластере.

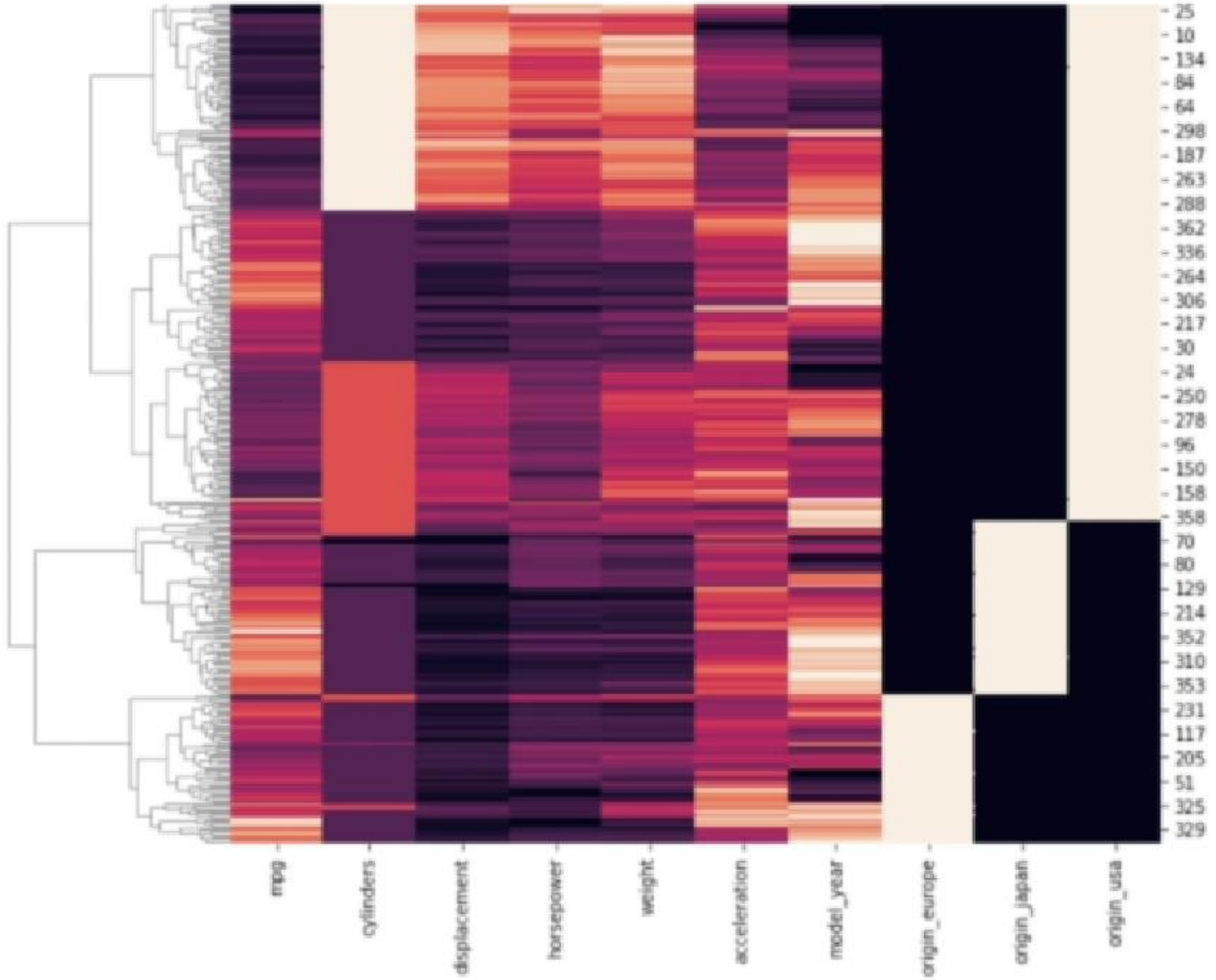
Иерархическая кластеризация

- **Процесс иерархической кластеризации**



Иерархическая кластеризация

- **Процесс иерархической кластеризации**



Иерархическая кластеризация

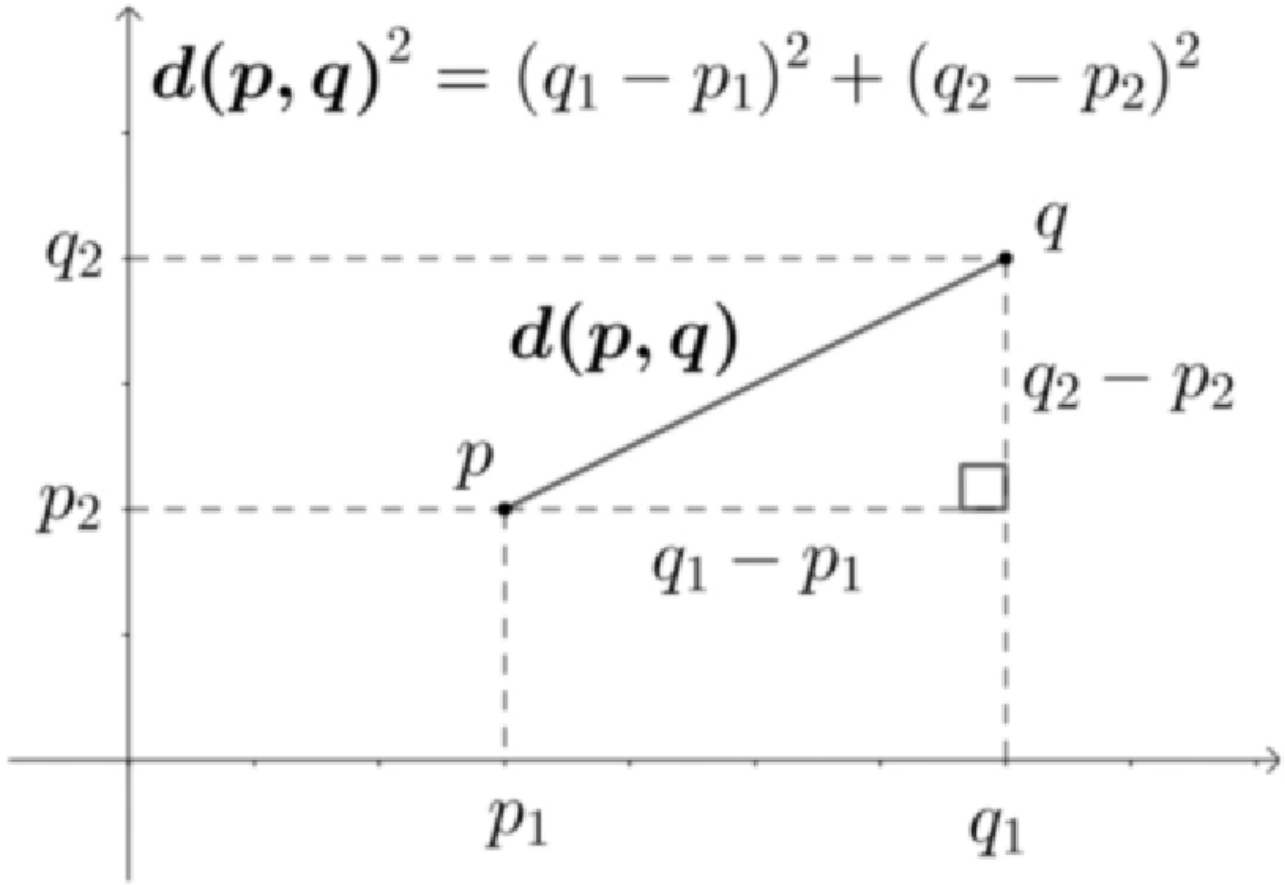
- В иерархической кластеризации участвуют следующие компоненты:
  - Метрика “похожести” (Similarity Metric)
  - Дендрограмма (Dendrogram)
  - Матрица связей (Linkage Matrix)

Иерархическая кластеризация

- Метрика “похожести” (Similarity Metric)
  - Измеряет расстояние между двумя точками.
  - Различные варианты:
    - Евклидово расстояние
    - Расстояние Манхэттена (расстояние городских кварталов)
    - Косинусное сходство
    - и много других метрик

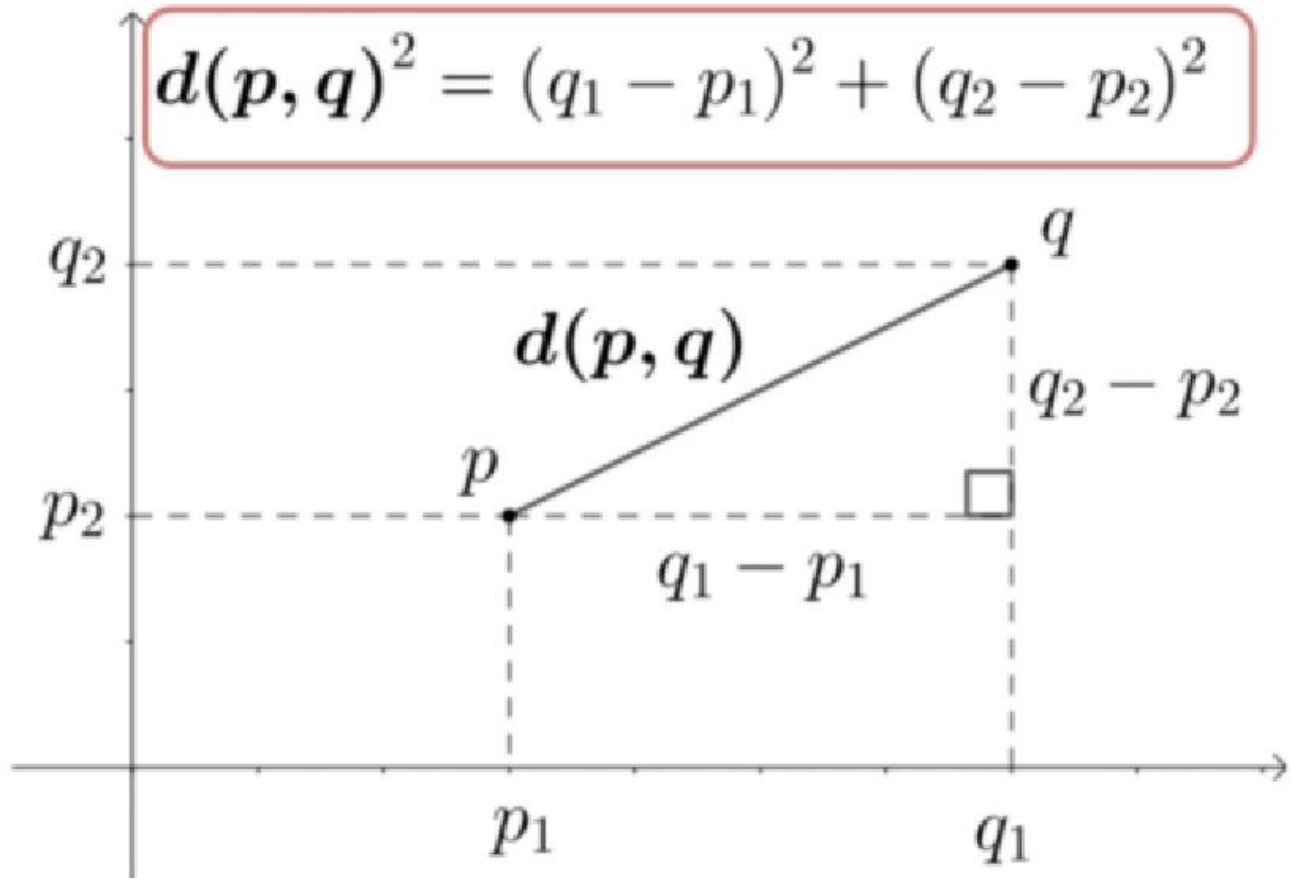
Иерархическая кластеризация

- Метрика “похожести” (Similarity Metric)
  - По умолчанию – Евклидово расстояние



Иерархическая кластеризация

- Метрика “похожести” (Similarity Metric)
  - По умолчанию – Евклидово расстояние



Иерархическая кластеризация

- Метрика “похожести” (Similarity Metric)
  - Каждое измерение – это признак (feature)
  - Для **n** точек и **p** признаков:
    - $D^2 = (x_{11} - x_{12})^2 + \ldots + (x_{n-1p-1} - x_{np})^2$

Иерархическая кластеризация

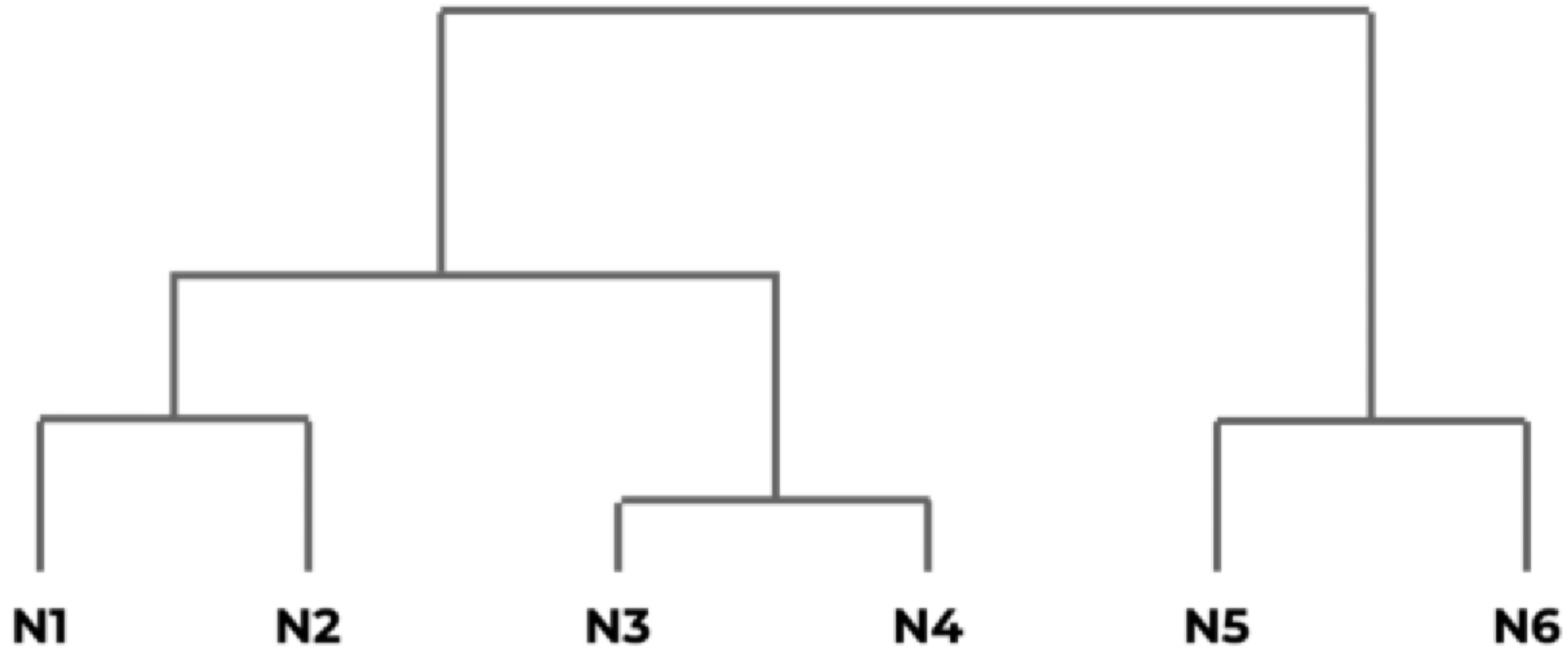
- Метрика “похожести” (Similarity Metric)
  - Каждое измерение – это признак (feature)
  - Для **n** точек и **p** признаков:
    - $D^2 = (x_{11} - x_{12})^2 + \ldots + (x_{n-1p-1} - x_{np})^2$
  - С помощью MinMaxScaler мы масштабируем все признаки, получаем значения от 0 до 1.
  - Максимальное расстояние по каждому признаку равно 1.

Иерархическая кластеризация

- Дендрограмма:
  - Визуальное отображение всех возможных кластеров.
  - Если данных много, то очень трудоёмко по вычислительным ресурсам.
  - Очень полезно для определения количества кластеров.

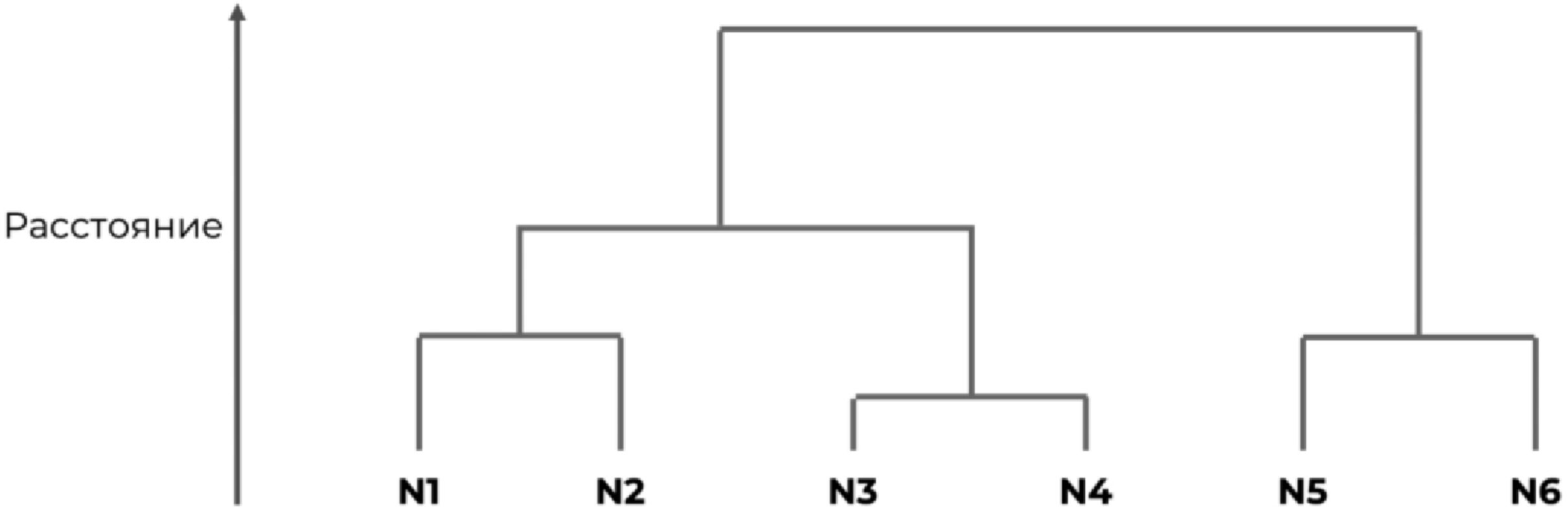
Иерархическая кластеризация

- Дендрограмма:



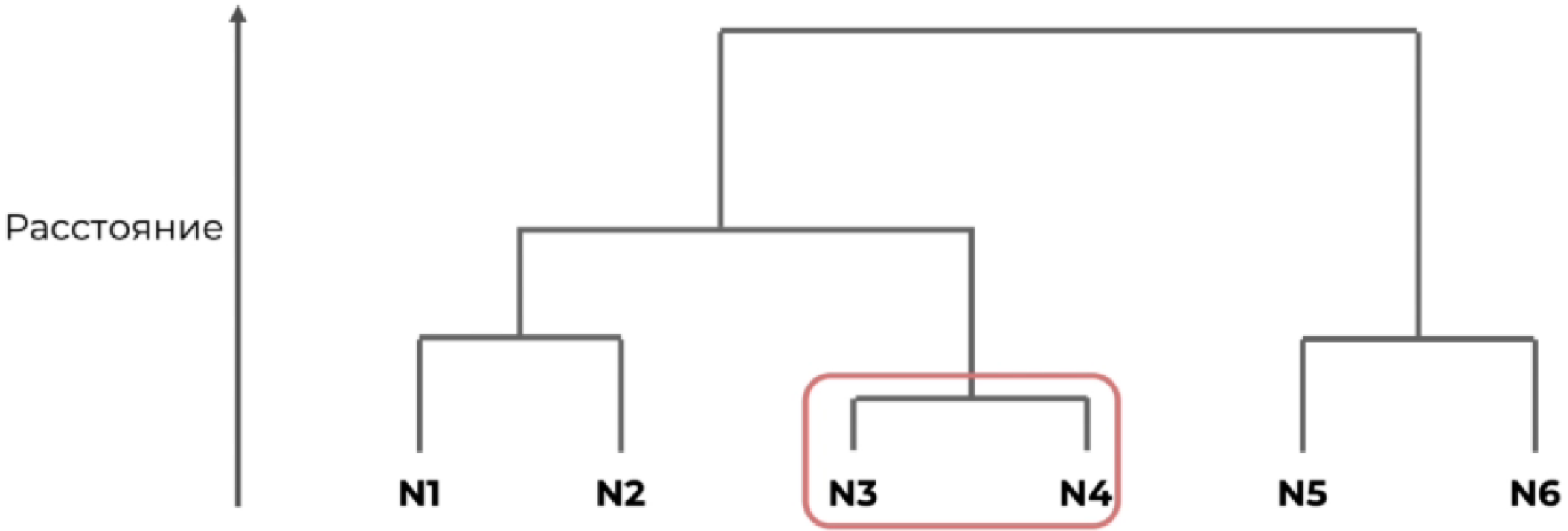
Иерархическая кластеризация

- Дендрограмма:



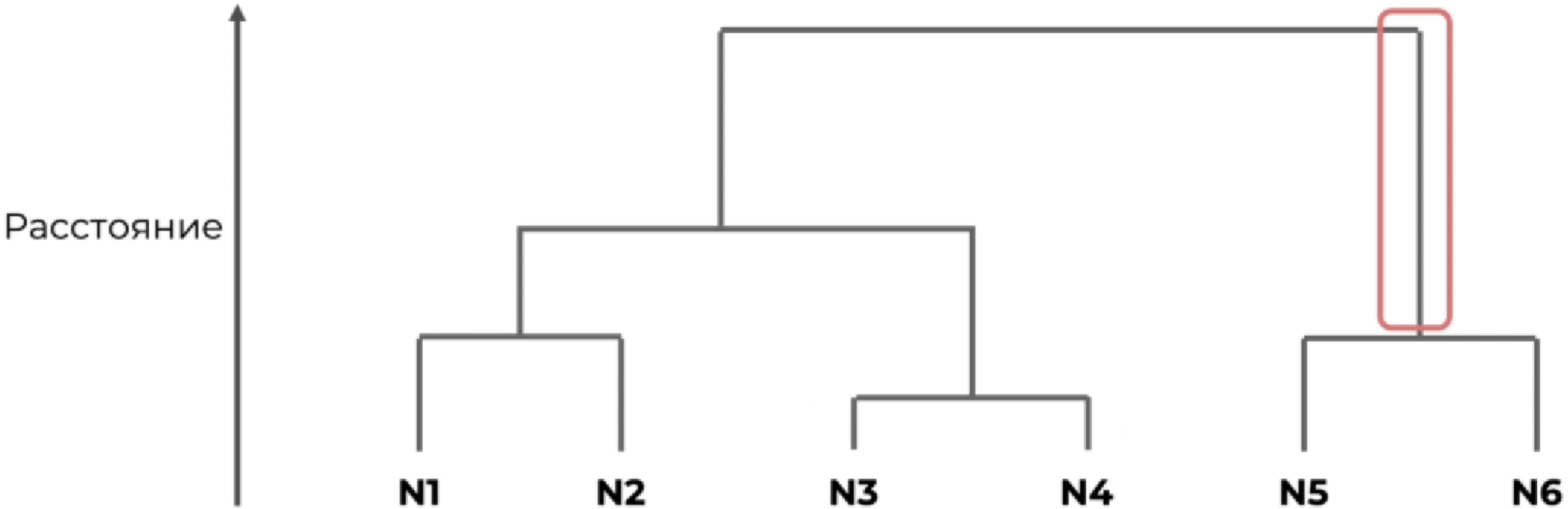
Иерархическая кластеризация

- Дендрограмма:



Иерархическая кластеризация

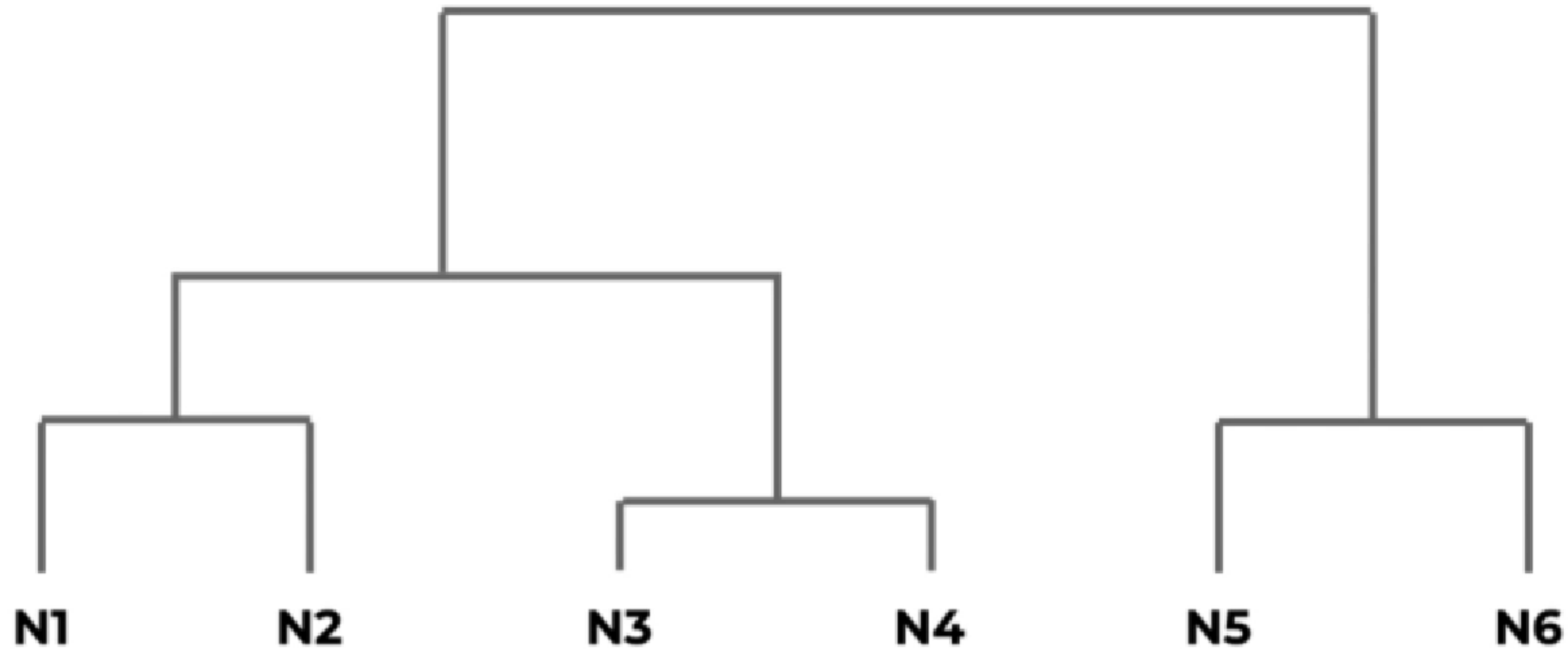
- Дендрограмма:



Иерархическая кластеризация

- Дендрограмма:

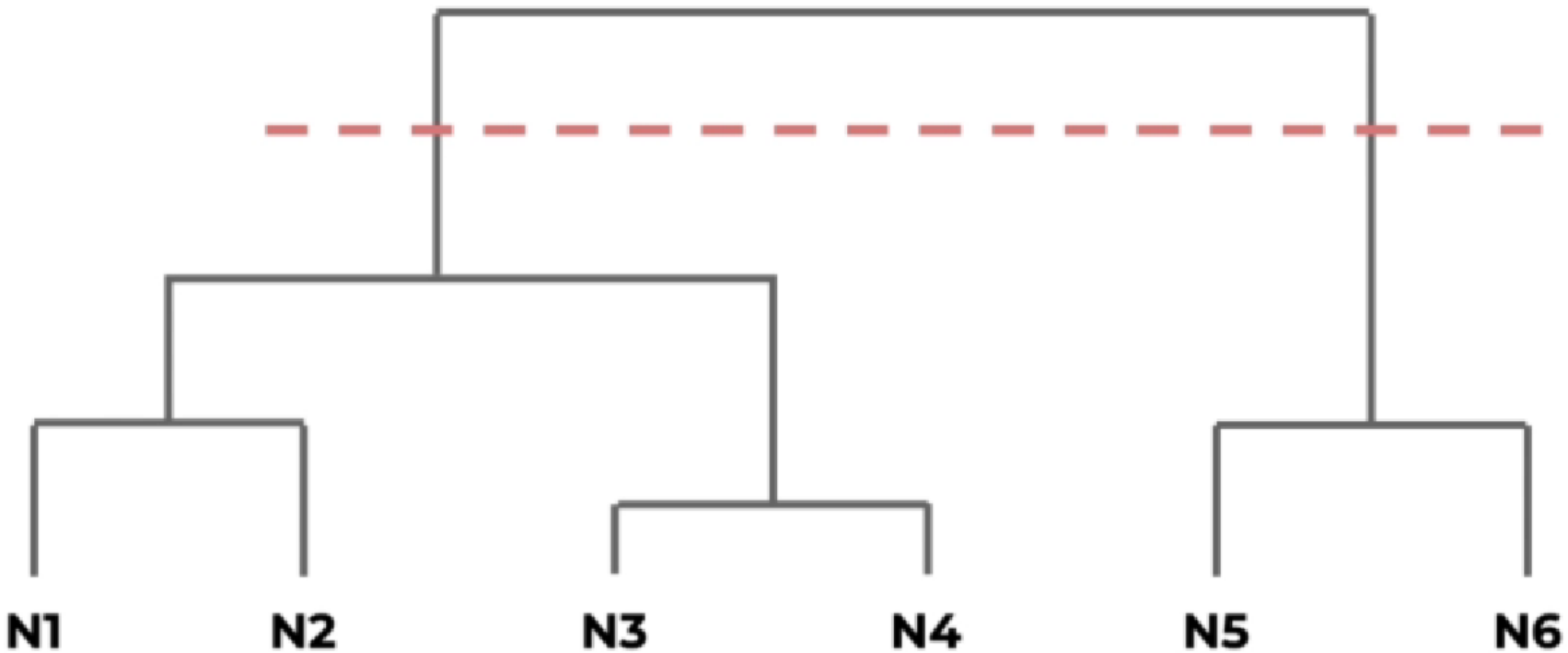
“Разделяем” дерево для выбора количества кластеров



Иерархическая кластеризация

- Дендрограмма:

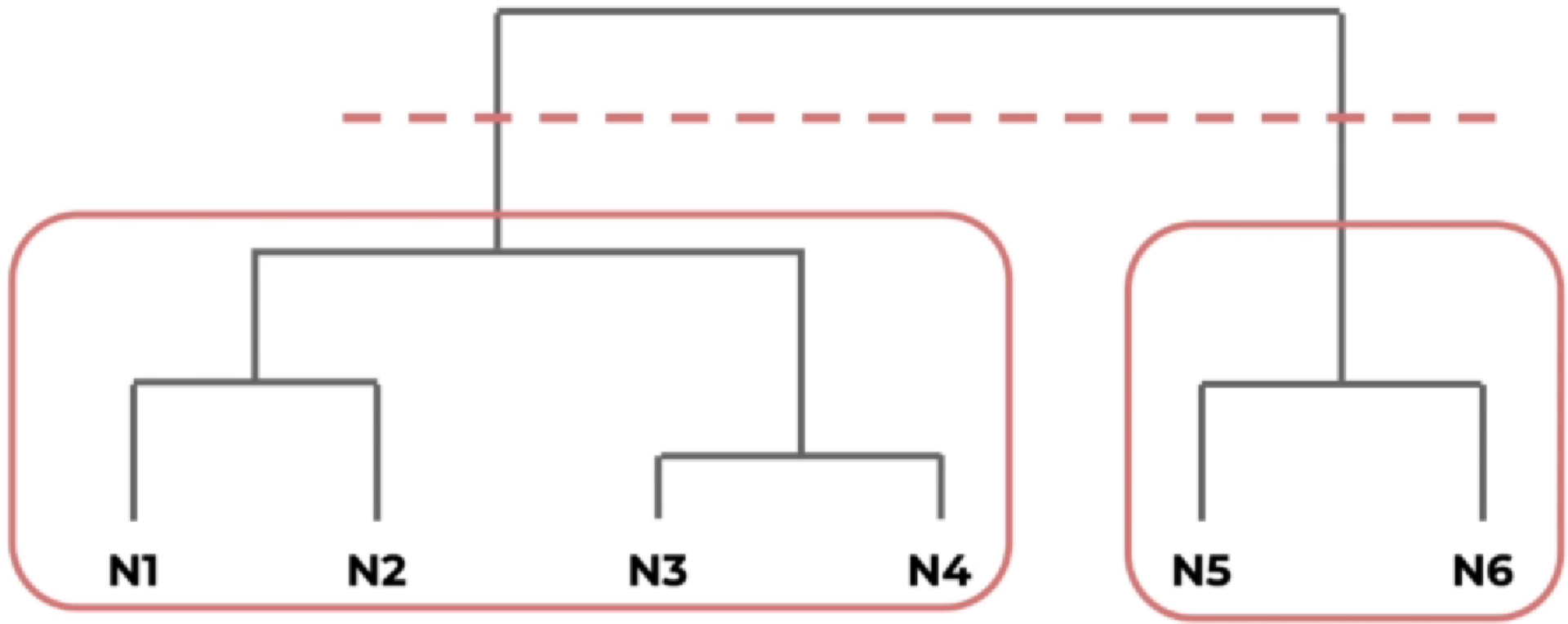
“Разделяем” дерево для выбора количества кластеров



Иерархическая кластеризация

- Дендрограмма:

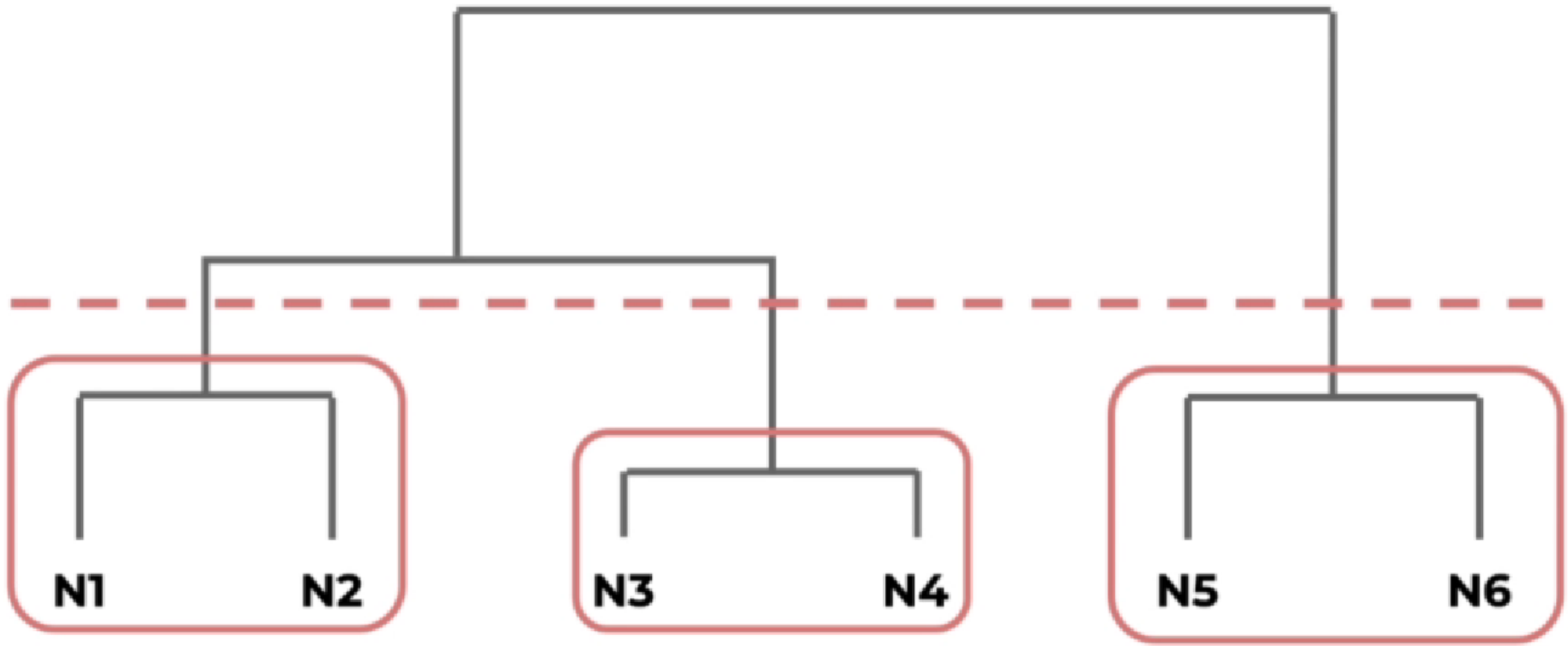
“Разделяем” дерево для выбора количества кластеров



Иерархическая кластеризация

- Дендрограмма:

“Разделяем” дерево для выбора количества кластеров



Иерархическая кластеризация

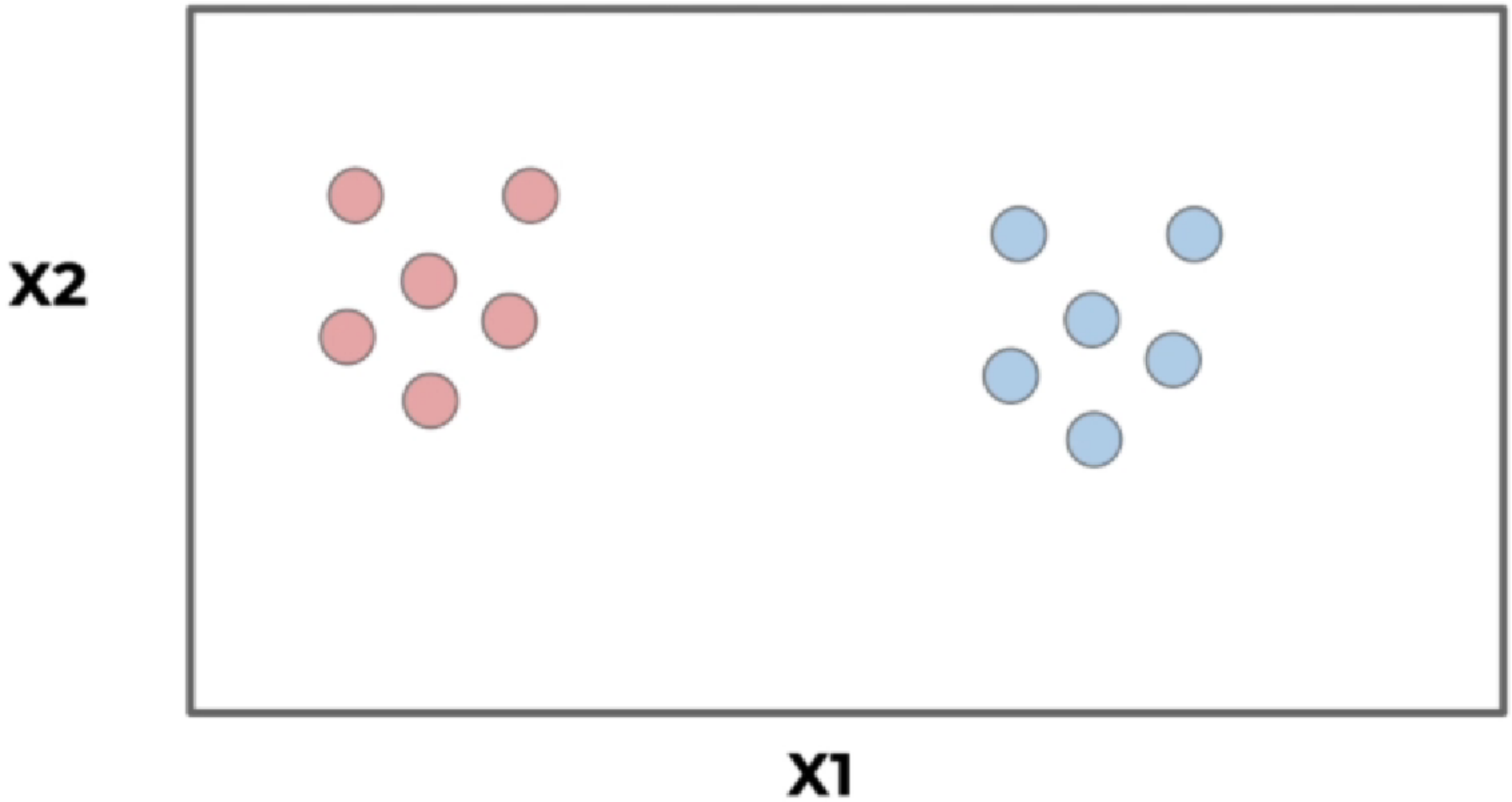
- Связи (Linkage)
  - Как мы измеряем расстояние между точкой и целым кластером?
  - Как мы измеряем расстояние между одним кластером и другим кластером?

Иерархическая кластеризация

- Связи (Linkage)
  - После объединения двух или нескольких точек в кластеры, при агломеративном подходе нам нужно объединять кластеры.
  - Для этого есть параметр **linkage**.

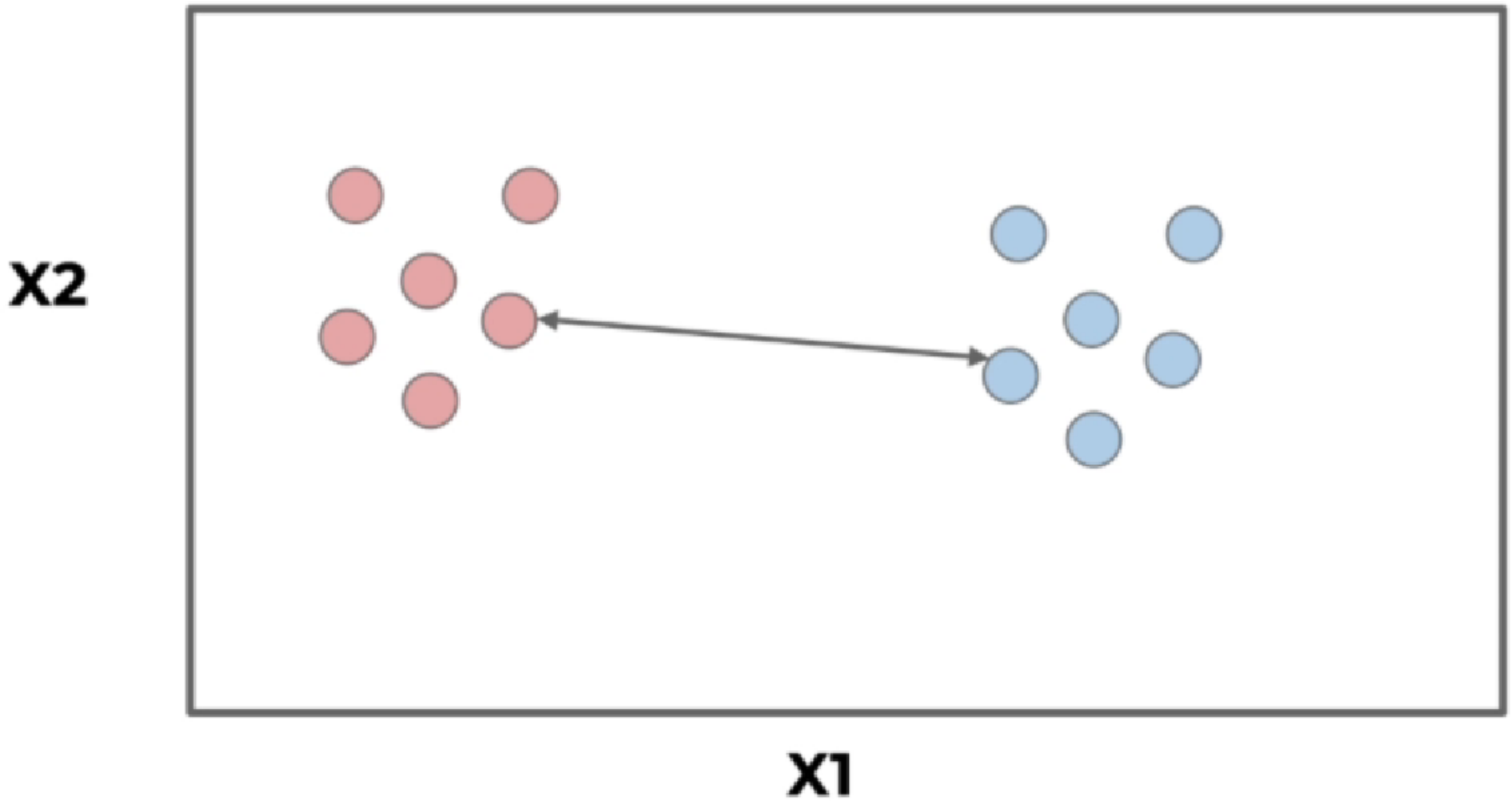
Иерархическая кластеризация

- Связи (Linkage)



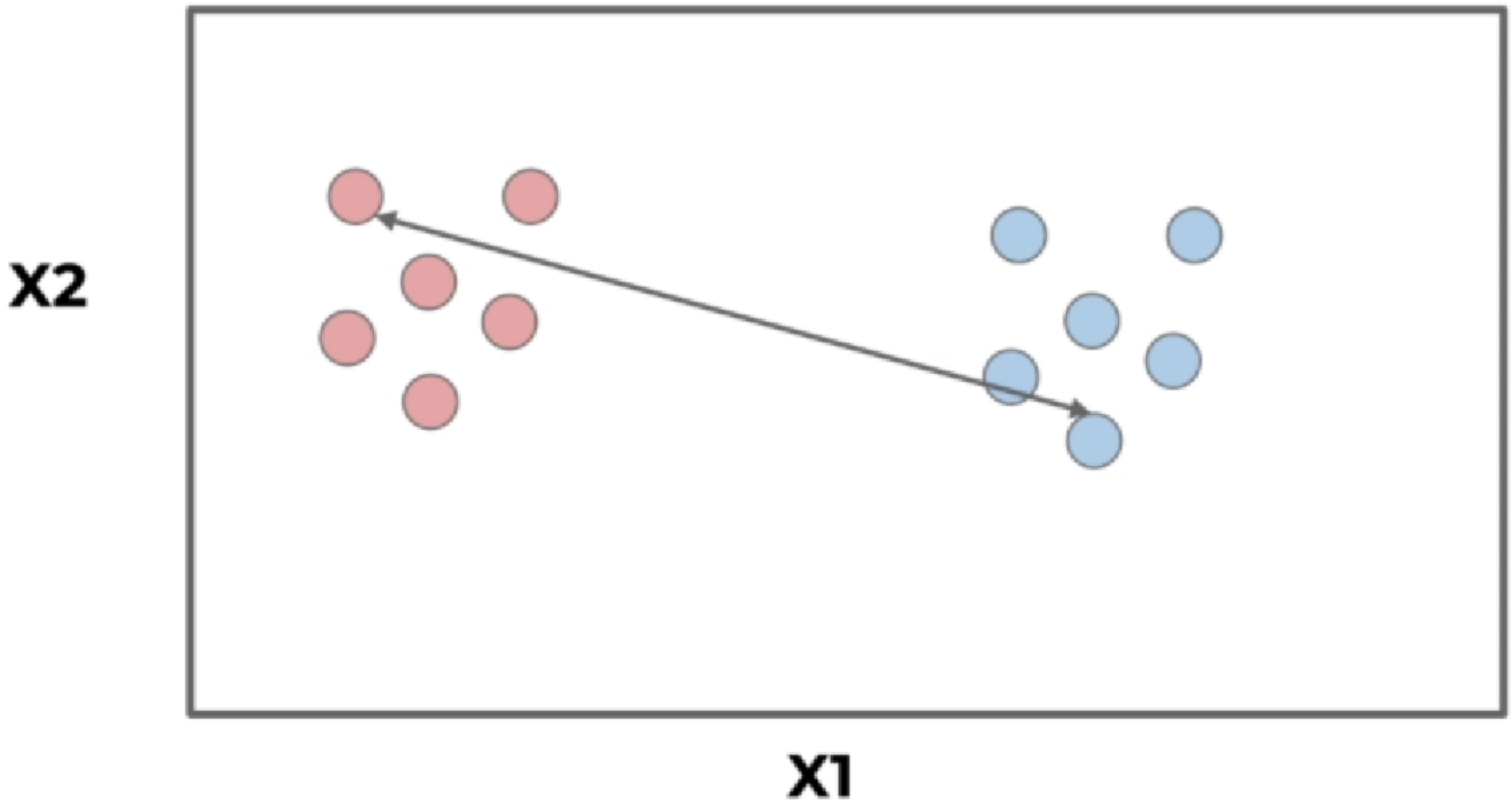
Иерархическая кластеризация

- Связи (Linkage)



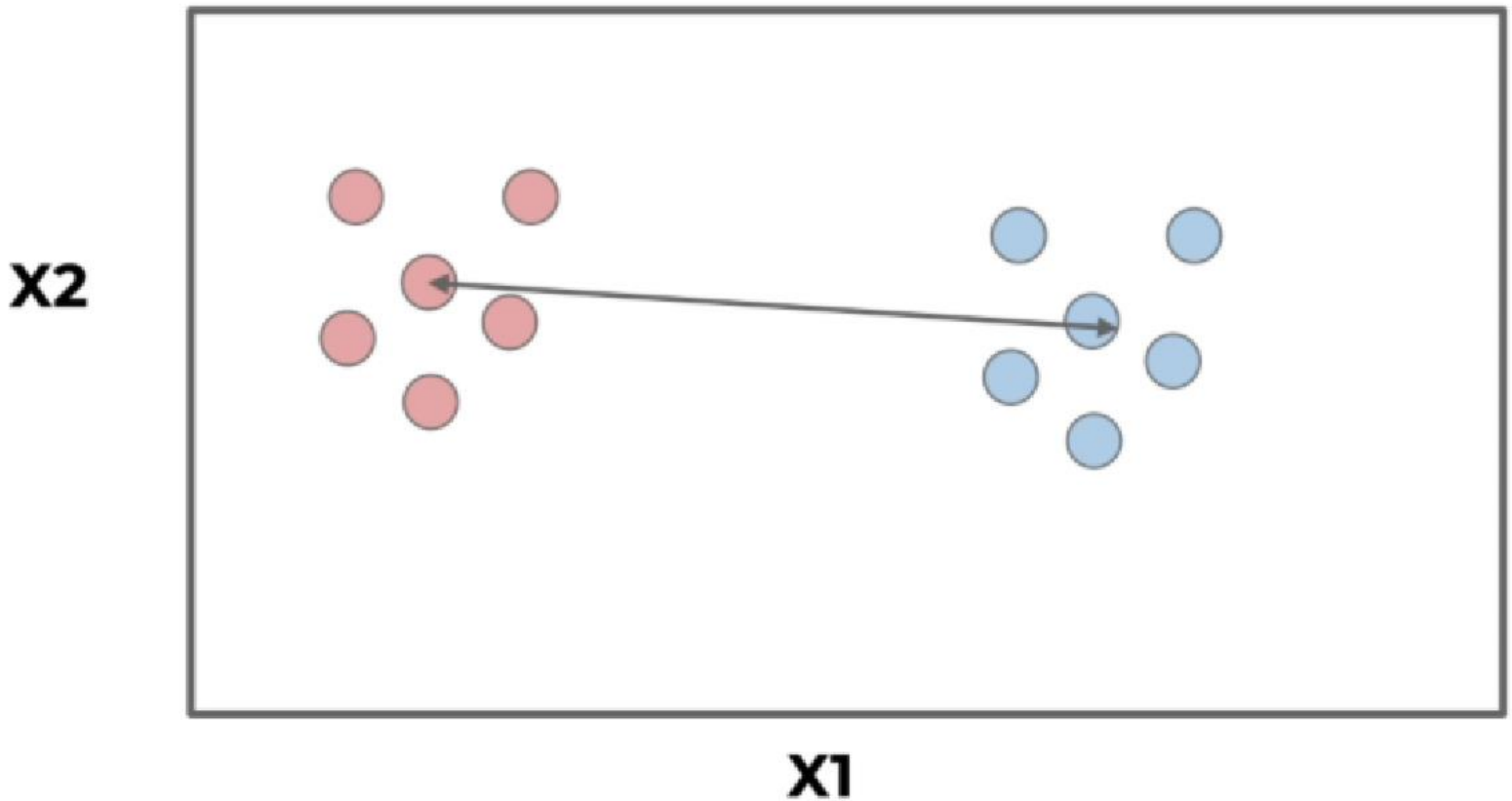
Иерархическая кластеризация

- Связи (Linkage)



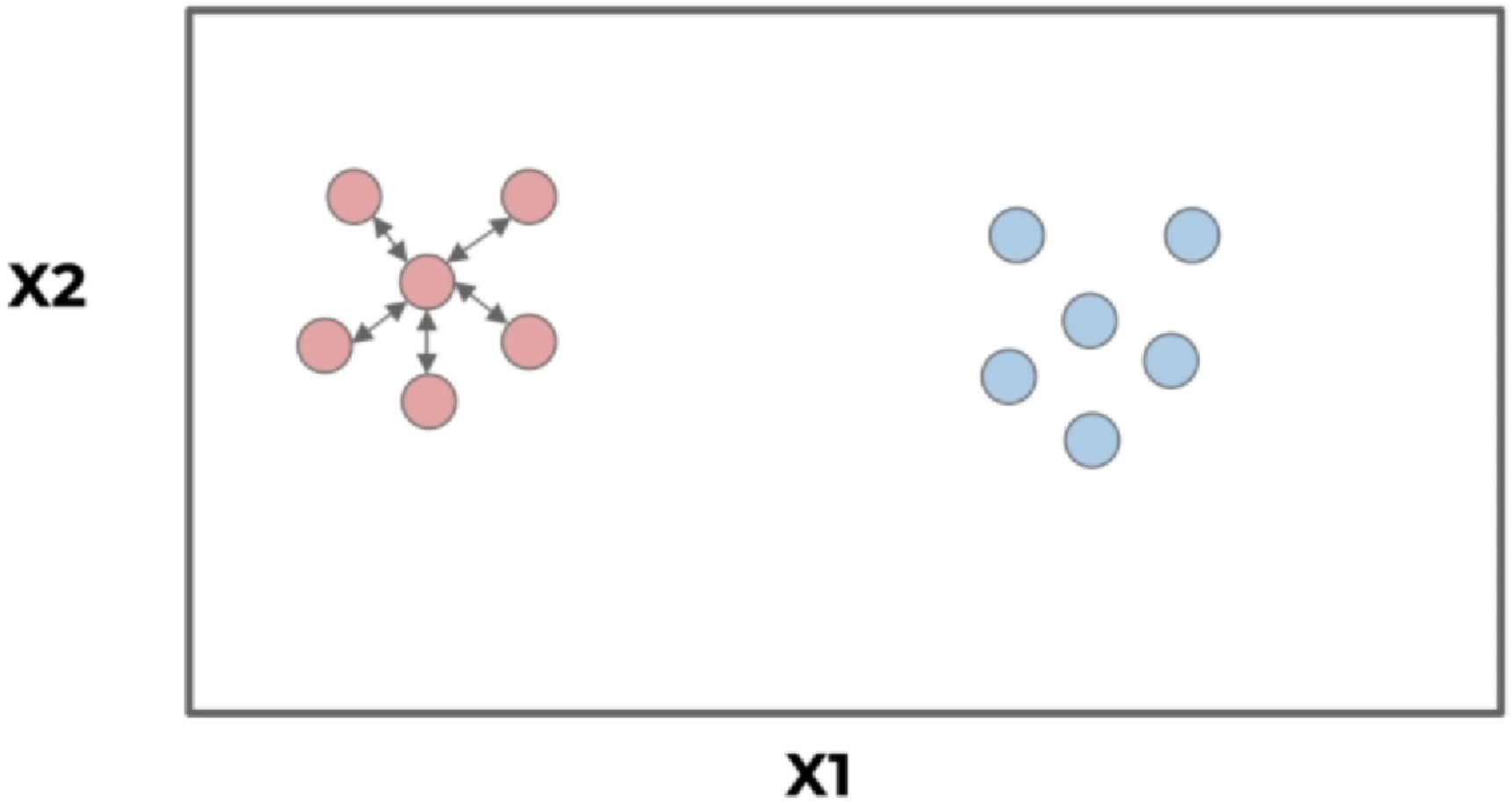
Иерархическая кластеризация

- Связи (Linkage)



Иерархическая кластеризация

- Связи (Linkage)



Иерархическая кластеризация

- Связи (Linkage)
  - Критерий для определения того, как измерять расстояние между наборами точек.
  - Алгоритм объединяет пары кластеров так, чтобы минимизировать этот критерий.

Иерархическая кластеризация

- Связи (Linkage)
  - **Ward:** минимизирует “variance” объединяемых кластеров.
  - **Average:** использует среднее расстояние между двумя наборами точек.
  - **Minimum** / **Maximum** – минимальное или максимальное расстояние между двумя наборами точек.