



«Анализ транскриптомных данных»

Лекция #12. **Кластеризация**

Серёжа Исаев

аспирант **MedUni Vienna**

Содержание курса

1. Bulk RNA-Seq:

- a. экспериментальные подходы,
- b. выравнивания и псевдовыравнивания,
- c. анализ дифференциальной экспрессии,
- d. функциональный анализ;

2. Single-cell RNA-Seq:

- a. экспериментальные подходы,
- b. отличия от процессинга bulk RNA-Seq,
- c. методы снижения размерности,
- d. кластера и траектории,**
- e. мультимодальные омики одиночных клеток.

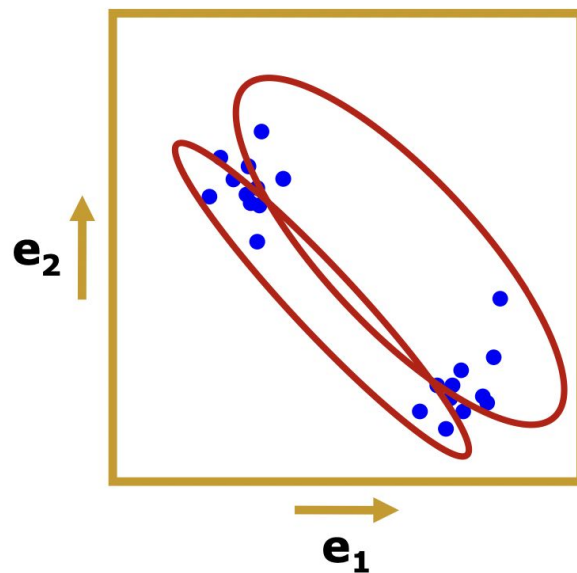
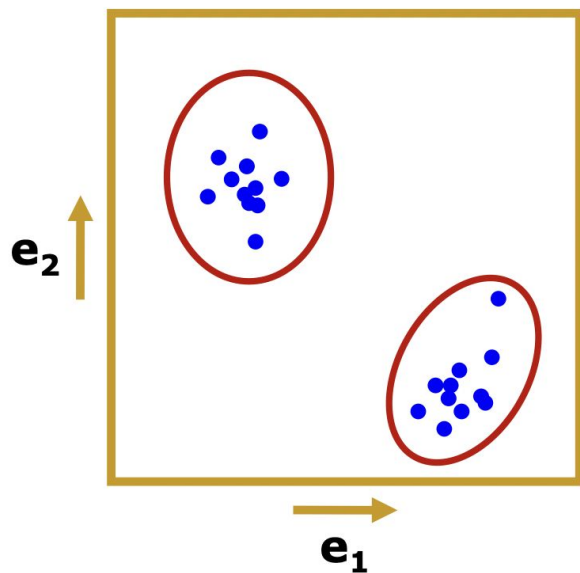
Основные подходы к кластеризации

После того, как мы снизили размерность мы можем успешно применять различные подходы, необходимые для интерпретации данных. В первую очередь это кластеризация, которая позволит нам ответить на вопрос, что за типы клеток были в исследуемом образце

1. Иерархическая кластеризация
2. K-Means
3. Графовые подходы

Что такое кластер?

Что из группировок ниже мы можем назвать кластером?

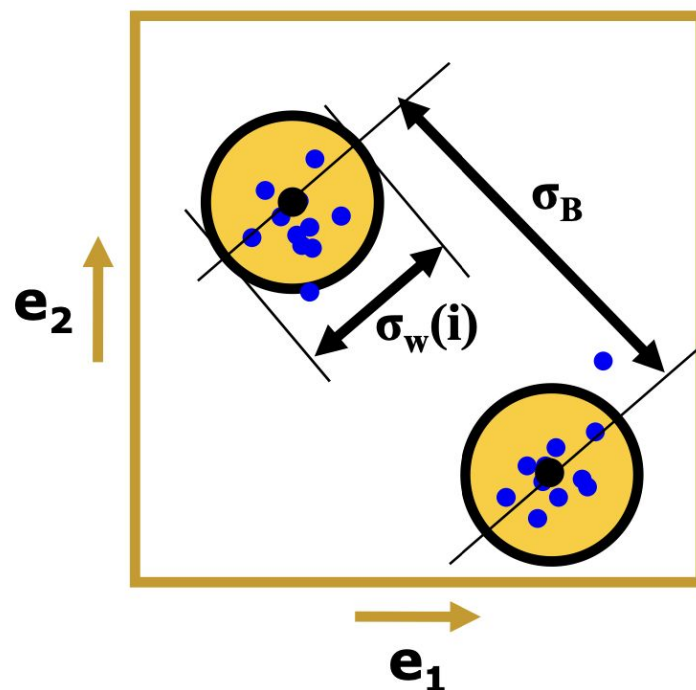


Что такое кластер?

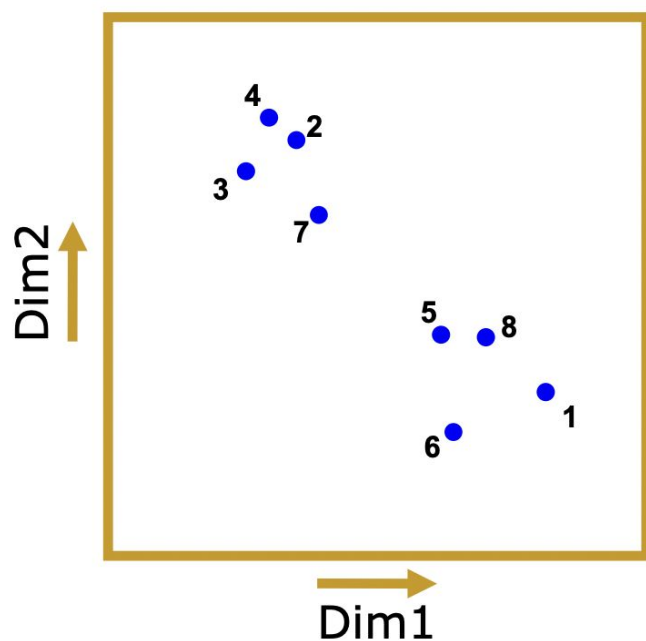
Наивное ожидание от кластеров формулируется следующим образом:

1. дисперсия внутри кластера не очень большая (σ_w),
2. дисперсия между кластерами (σ_B) большая

В целом требования к кластерам могут быть разными, от того существует множество алгоритмов кластеризации

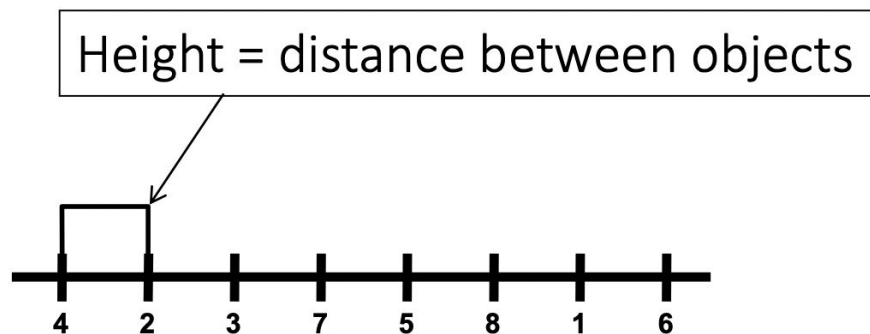
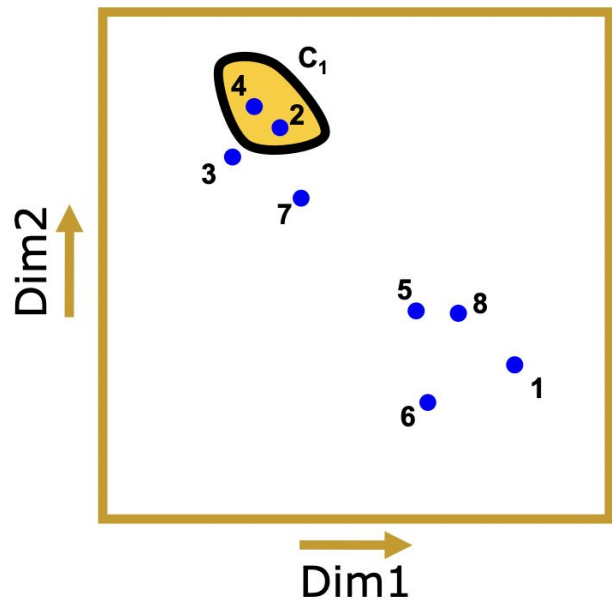


Иерархическая кластеризация



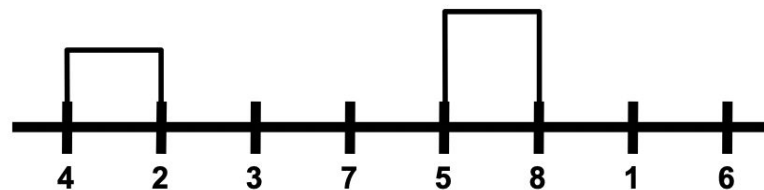
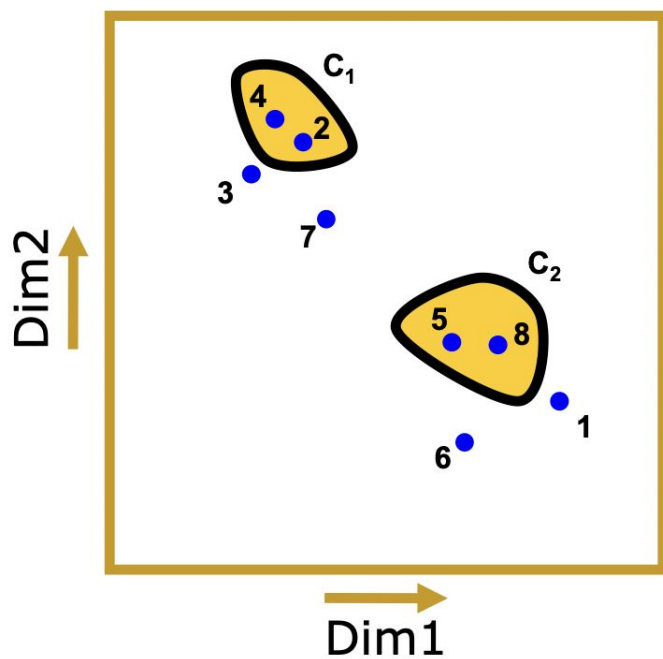
Иерархическая кластеризация

dendrogram



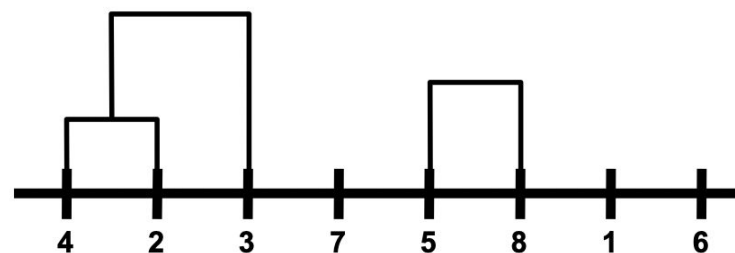
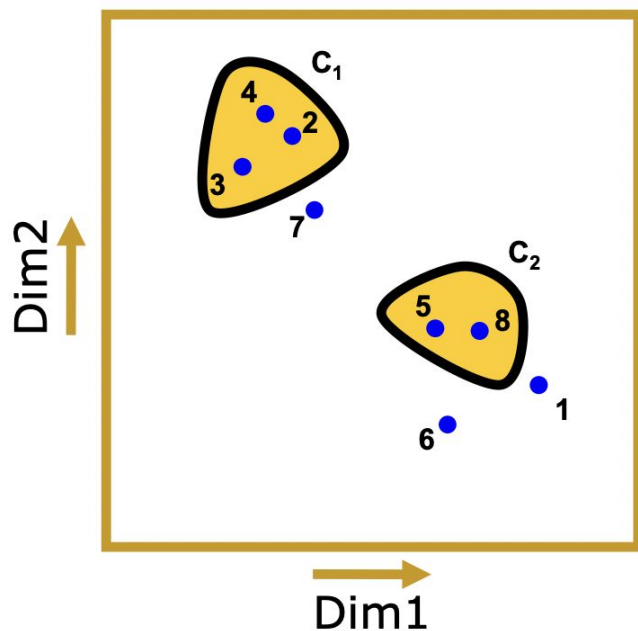
Иерархическая кластеризация

dendrogram

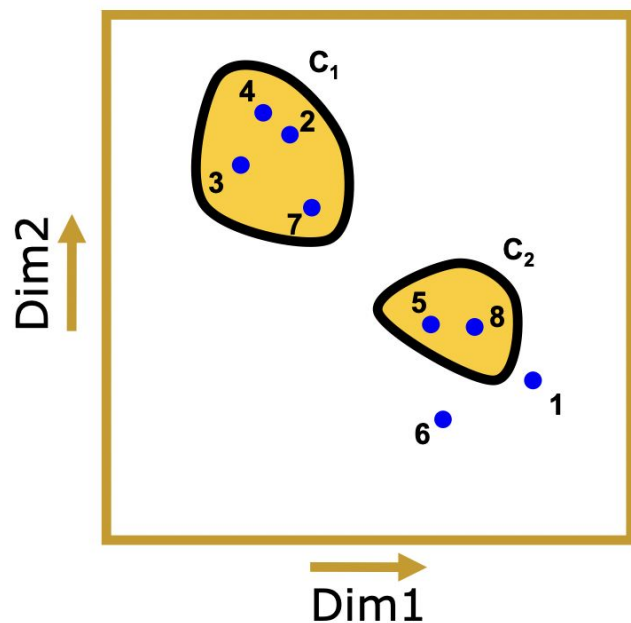


Иерархическая кластеризация

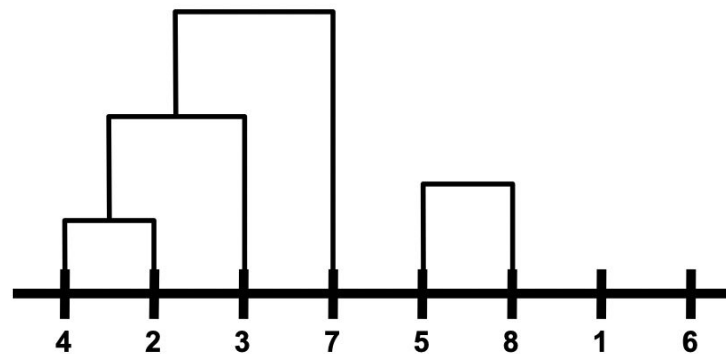
dendrogram



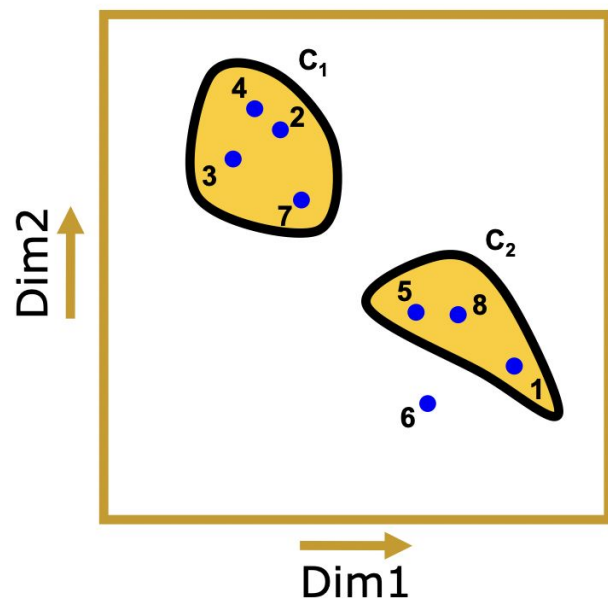
Иерархическая кластеризация



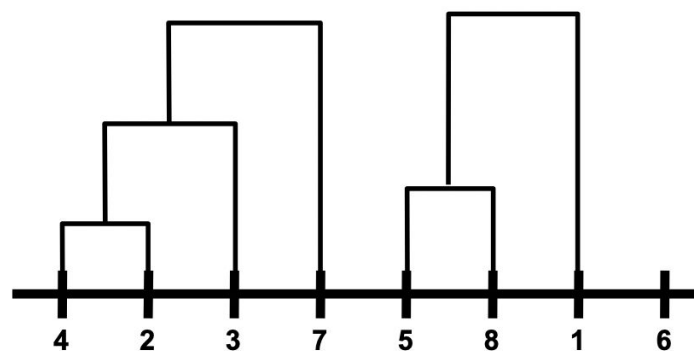
dendrogram



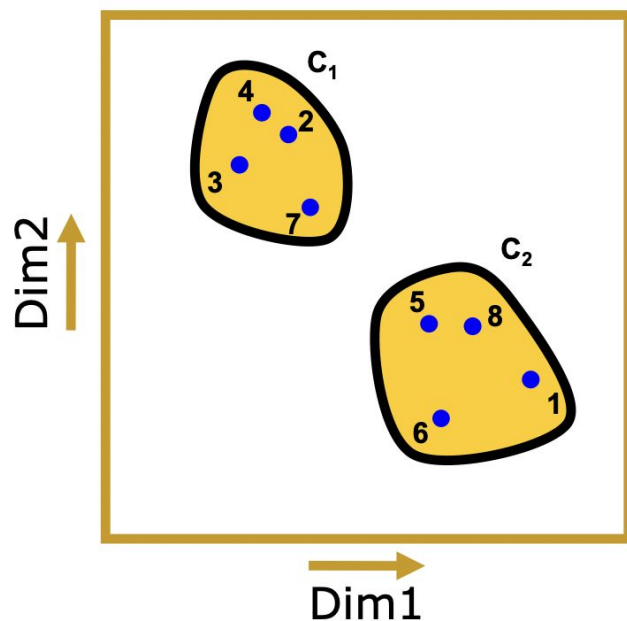
Иерархическая кластеризация



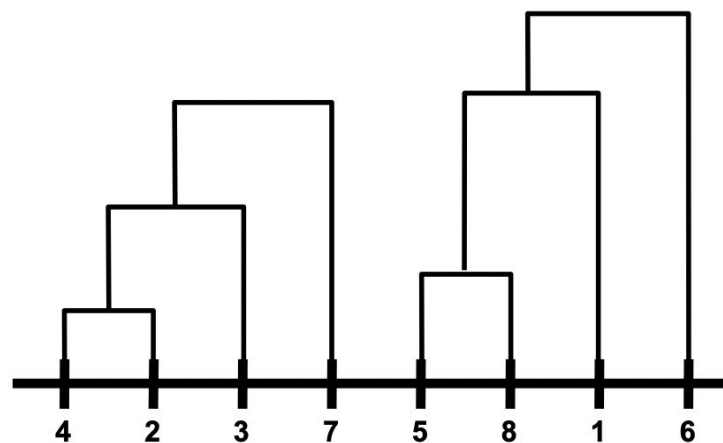
dendrogram



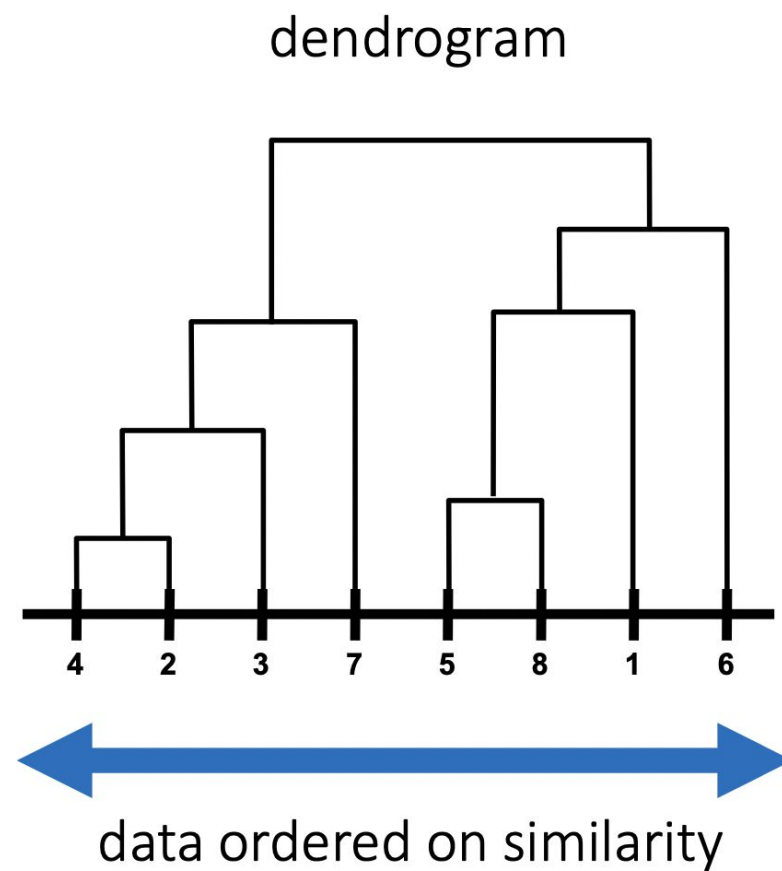
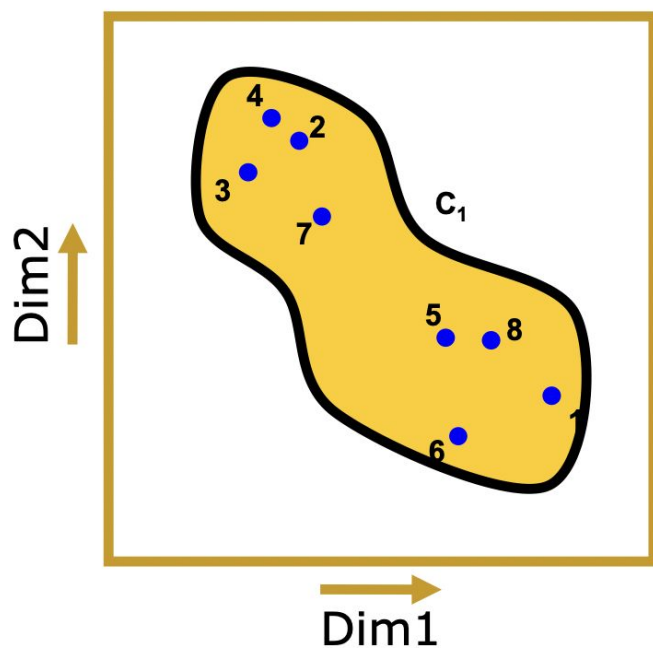
Иерархическая кластеризация



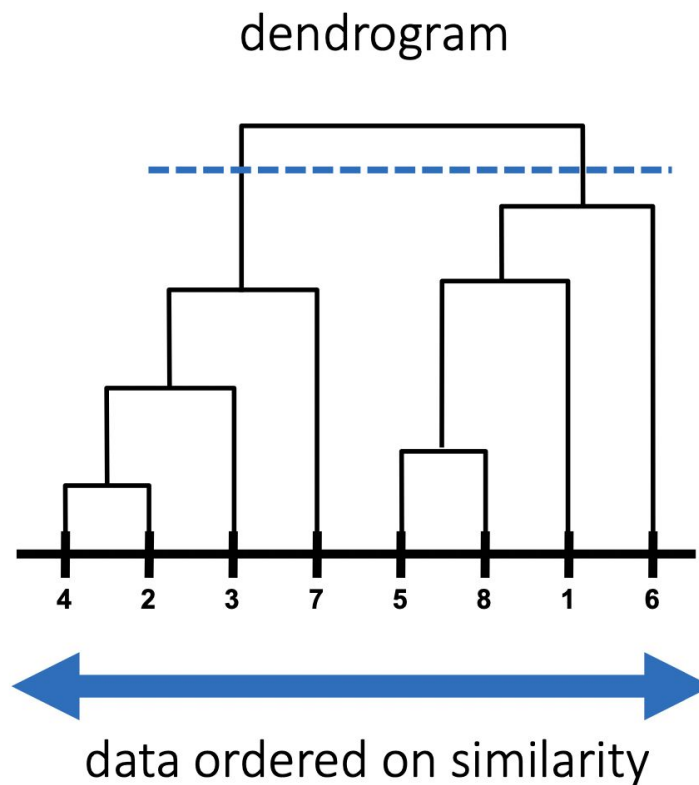
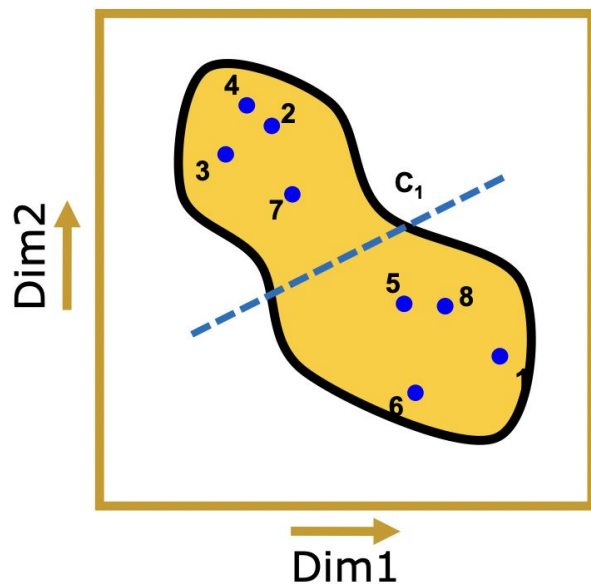
dendrogram



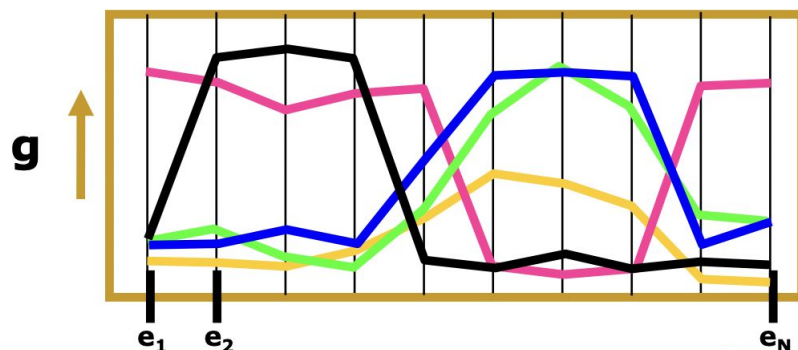
Иерархическая кластеризация



Иерархическая кластеризация



Метрики схожести



Euclidean distance

$$d(g_i, g_j) = \sqrt{\sum ((x_i - x_j)^2)}$$

$d(\bullet, \text{green}) < d(\bullet, \text{yellow})$
 $d(\bullet, \text{green}) << d(\bullet, \text{pink})$
 $d(\bullet, \text{green}) << d(\bullet, \text{black})$

Pearson correlation

$$1 - \rho_{ij}$$

$d(\bullet, \text{green}) \approx d(\bullet, \text{yellow})$
 $d(\bullet, \text{green}) << d(\bullet, \text{pink})$
 $d(\bullet, \text{green}) << d(\bullet, \text{black})$

Mixed Pearson correlation

$$1 - |\rho_{ij}|$$

$d(\bullet, \text{green}) \approx d(\bullet, \text{yellow})$
 $d(\bullet, \text{green}) \approx d(\bullet, \text{pink})$
 $d(\bullet, \text{green}) << d(\bullet, \text{black})$

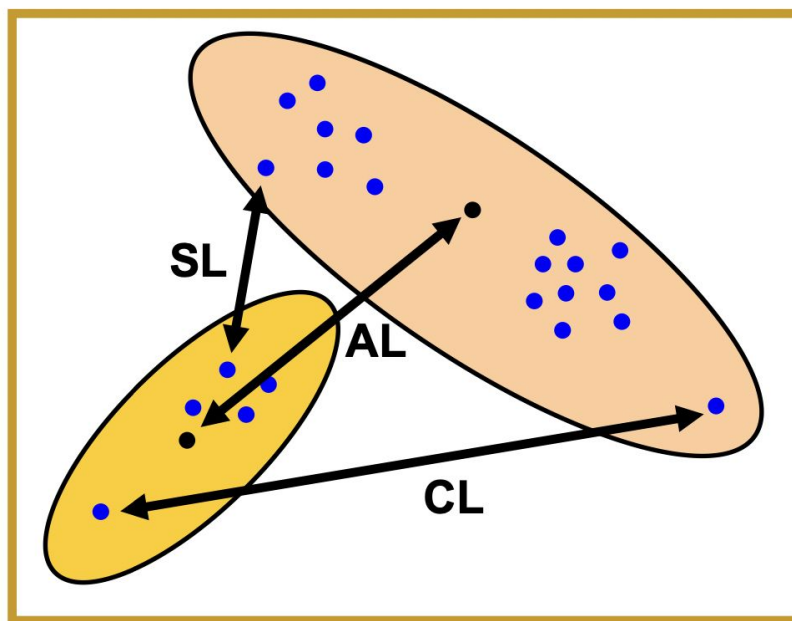
Расстояние между кластерами

Меры расстояния между кластерами:

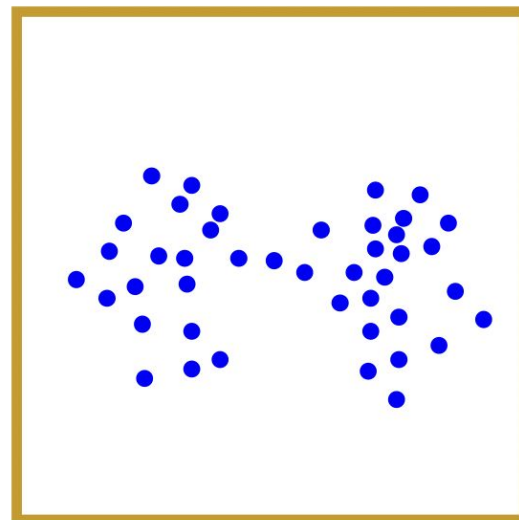
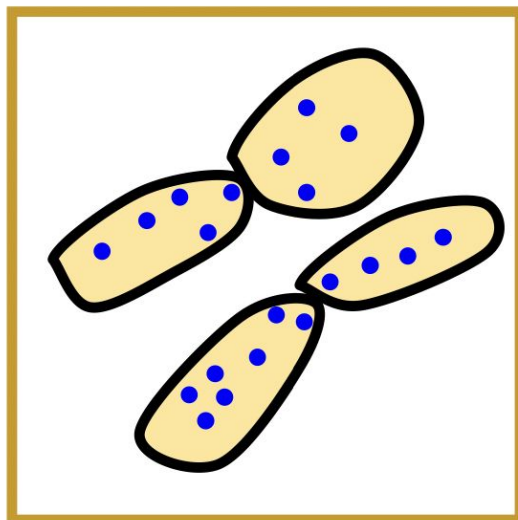
SL = Single Linkage — расстояние между ближайшими точками

AL = Average Linkage — расстояние между центрами масс

CL = Complete Linkage — расстояние между самыми далёкими точками



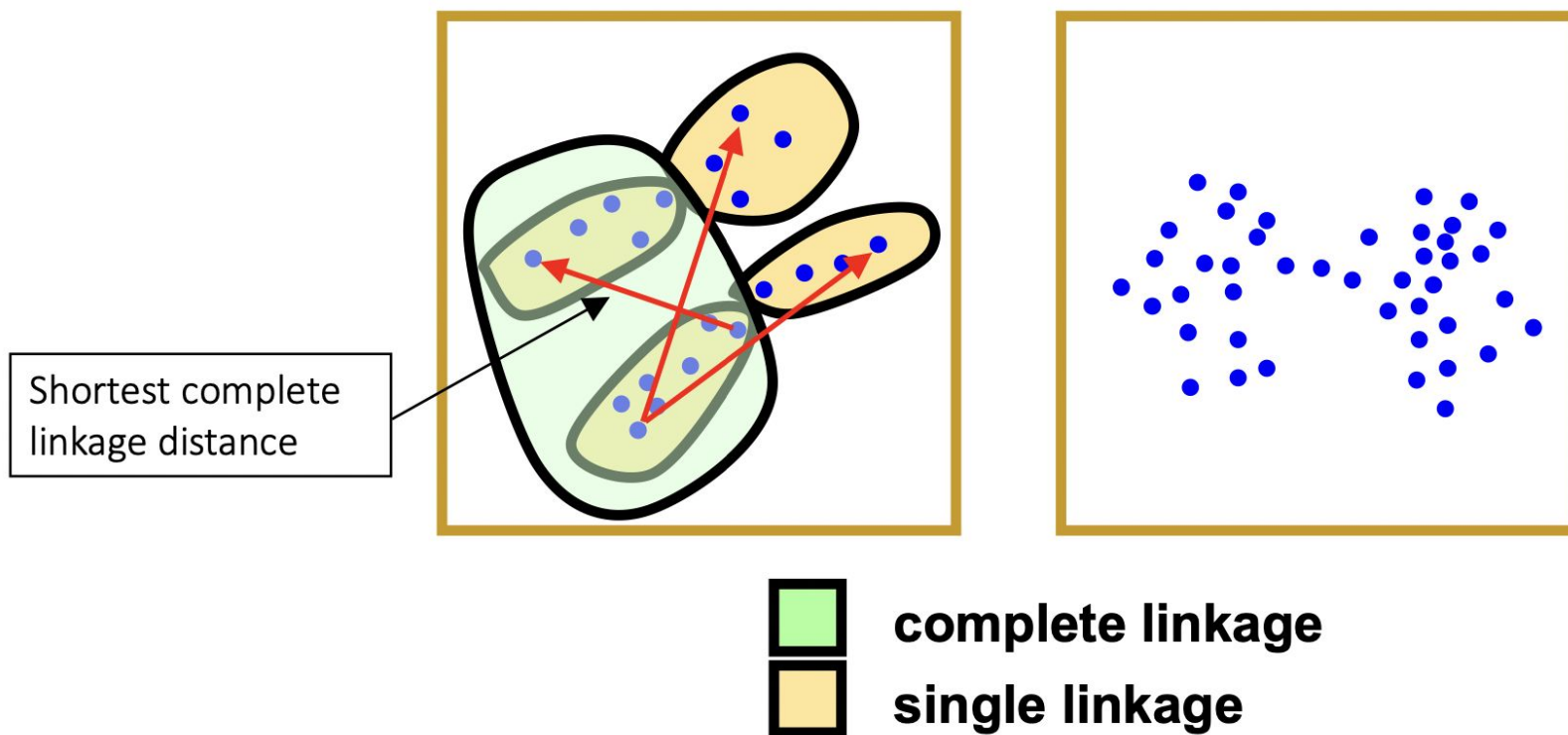
Расстояние между кластерами



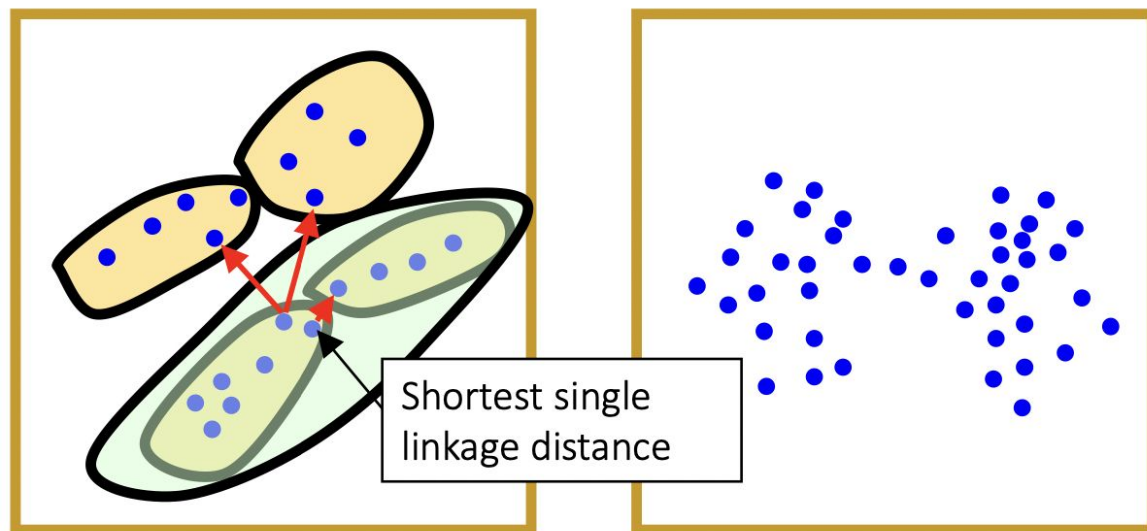
complete linkage

single linkage

Расстояние между кластерами



Расстояние между кластерами



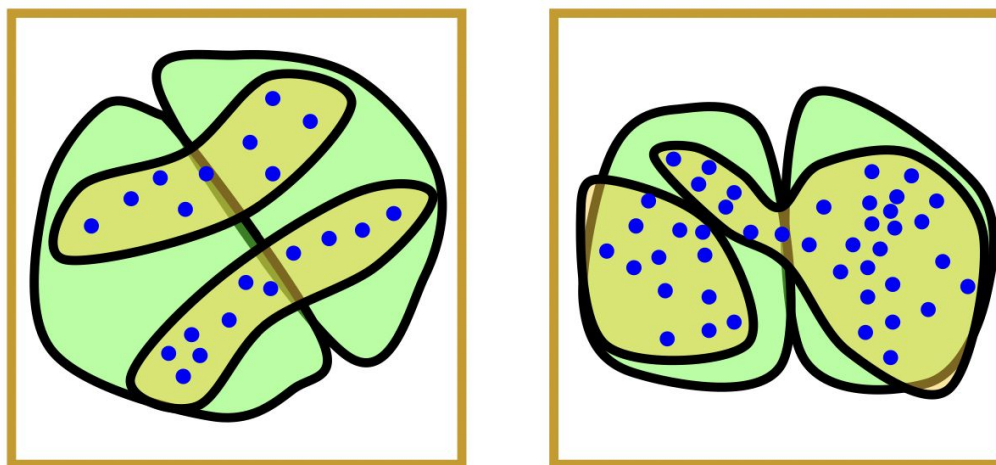
complete linkage

single linkage

Расстояние между кластерами

В итоге если в качестве расстояния между кластерами мы выберём single linkage, то получим “длинные” и “рыхлые” кластера

В случае с complete linkage мы получим маленькие и компактные кластера



Иерархическая кластеризация

Можно делать на данных экспрессии, но они очень шумные, поэтому обычно делают на пространстве сниженной размерности — например, PCA или Harmony

Не используется как основной метод кластеризации, однако может помочь сгруппировать кластера, которые у вас получаются при помощи других методов

K-Means

K-Means — это метод кластеризации, который основан на минимизации функции потерь, описанной снизу

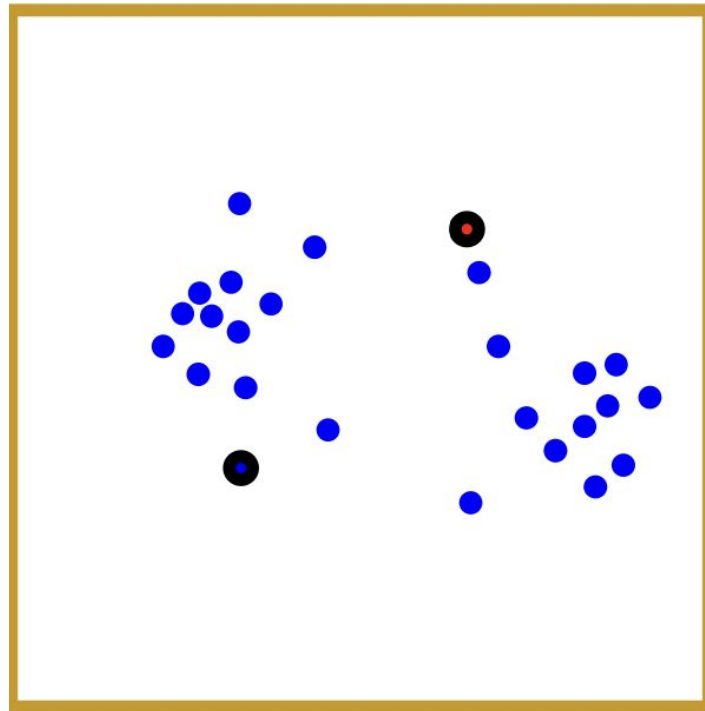
Для K-Means необходимо заранее задать количество кластеров

В целом метод похож на EM, рассмотренный в начале курса, однако по факту он гораздо проще

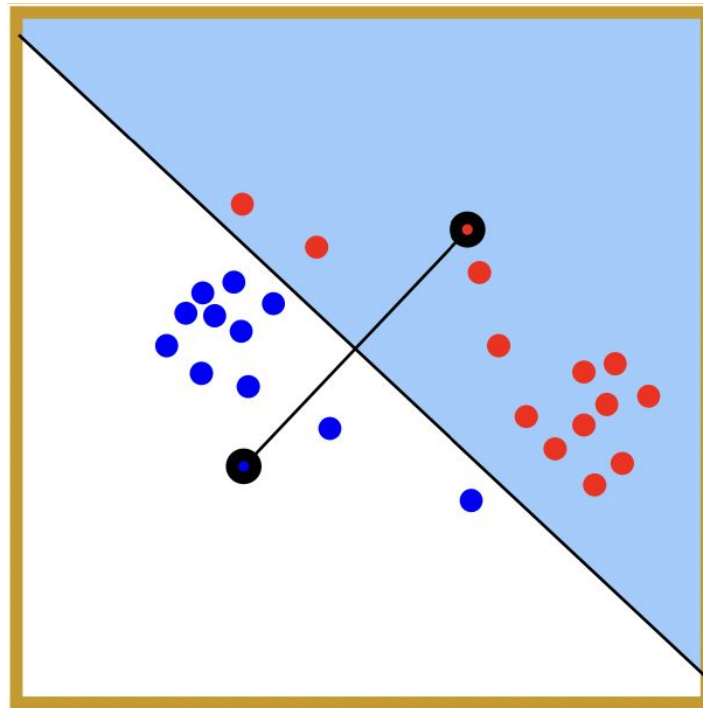
Имеет множество минусов, поэтому используется в scRNA-Seq анализе редко

$$\mathcal{L} = \sum_k \sum_{i \in S_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 = \sum_{k=1}^K \sum_{i=1}^n r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

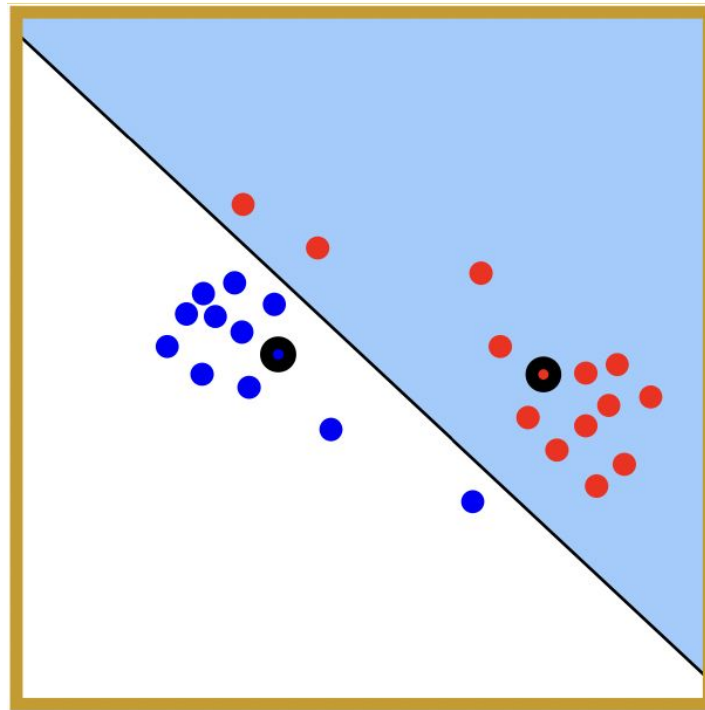
K-Means



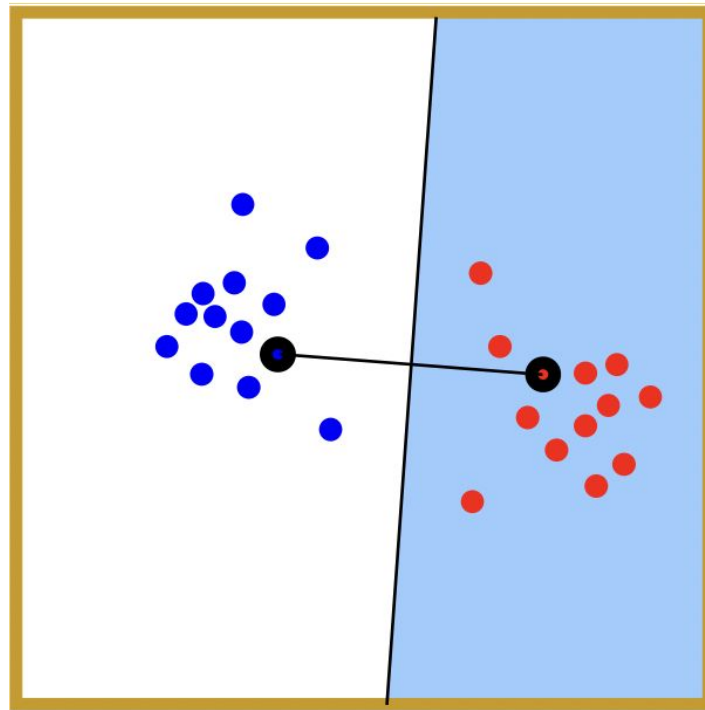
K-Means



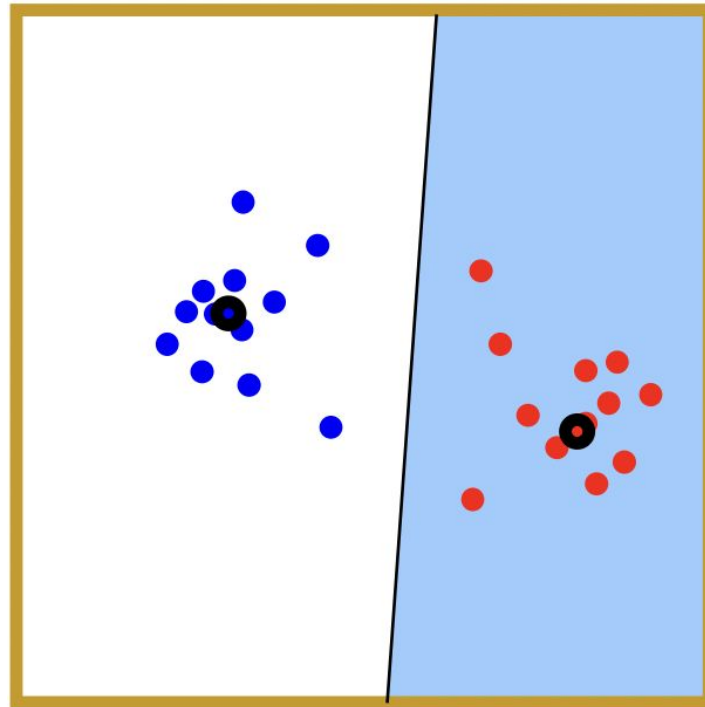
K-Means



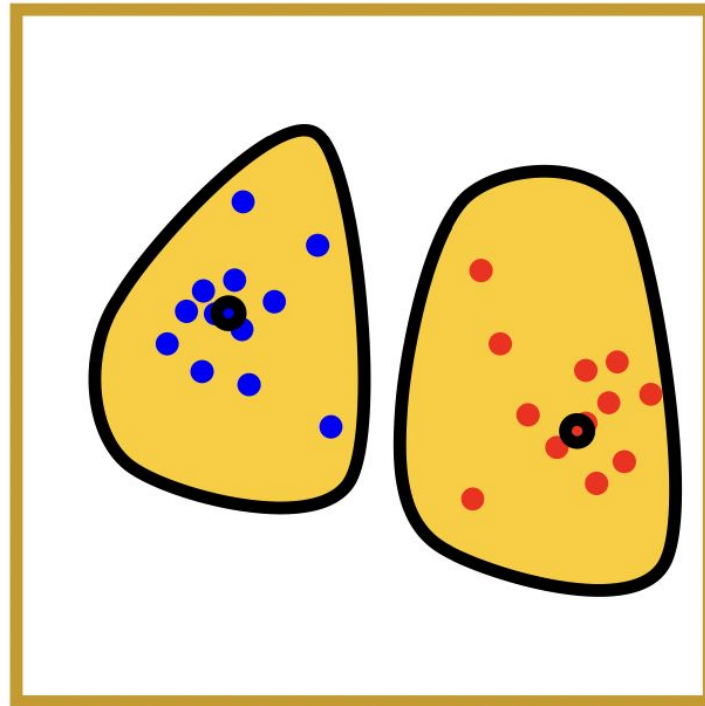
K-Means



K-Means

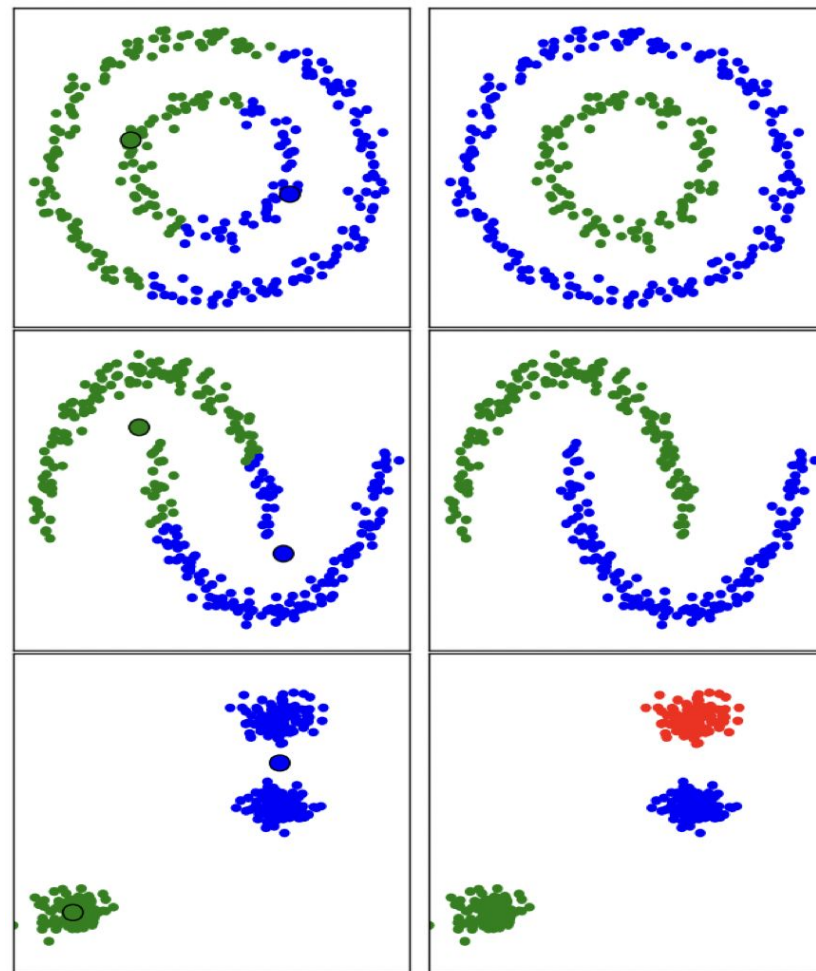


K-Means



Ограничения К-Means

1. В реальной жизни кластера могут описываться не только N -мерными шарами
2. Может сходиться очень долго для данных высокой размерности со сложной структурой
3. Нам необходимо заранее задавать количество кластеров K

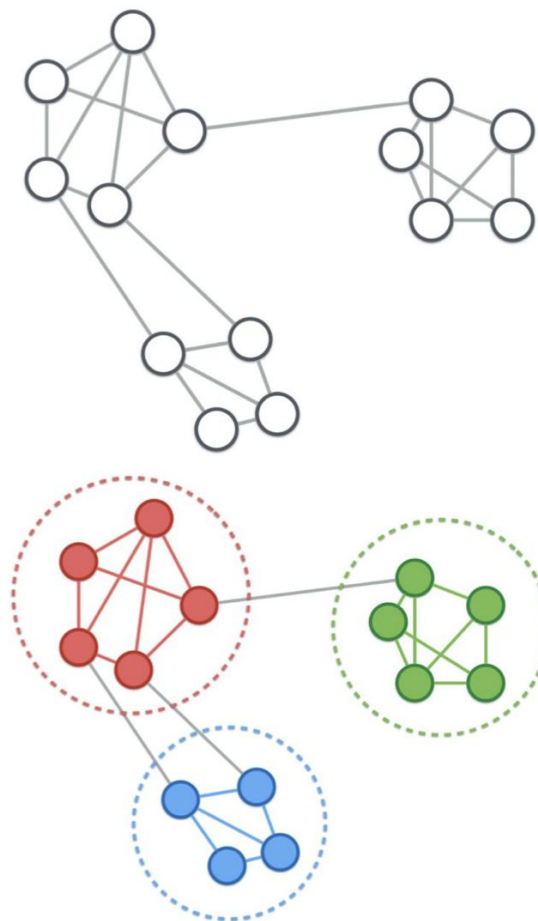


Графовые методы кластеризации

Графовые методы кластеризации завязаны на определении сообществ в графах

Сообщество — это такая группа вершин, каждая из которых имеет большую вероятность быть соединённой с другой вершиной из этого сообщества, чем из другого

Поиск сообществ сводится к определению таких групп клеток, у которых внутри группы плотность рёбер гораздо больше, чем между разными группами

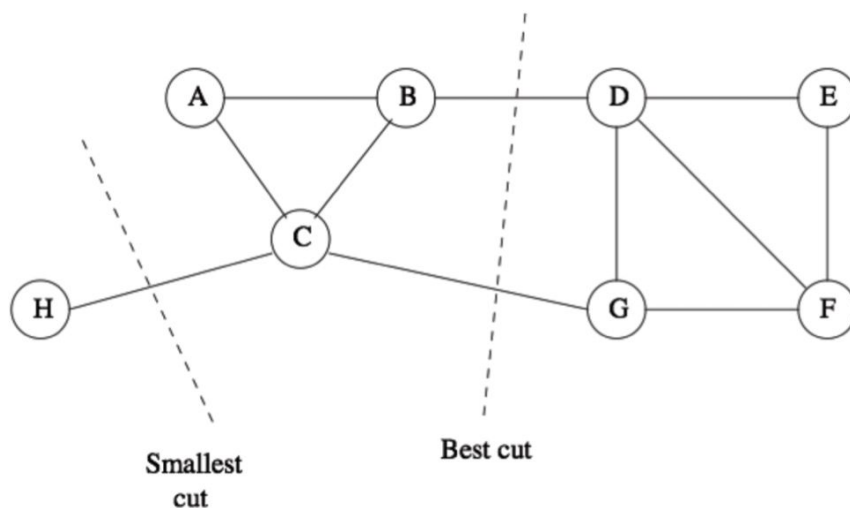


Разрезание графа

Разрезание графа разбивает граф на подграфы

Размер разреки (cut size) — это количество рёбер, которые необходимо устранить для разбиения графа

Задача сводится к поиску наименьшего размера разрезки графа, которая бы разбила граф на два подграфа. Однако не всегда это хорошо!



Нормализованное разрезание

Добавим некоторые метрики качества нашей разрезки:

1. $\text{vol}(S)$ — это сумма весов рёбер в (под)графе S ,
2. $\text{cut}(S, T)$ — это число рёбер, которые соединяют вершины в (под)графе S с вершинами в (под)графе T

Таким образом, можно искать такое разбиение на два подграфа, которое будет минимизировать не только $\text{cut}(S, T)$, но следующую функцию:

$$Ncut(S, T) = \frac{\text{cut}(S, T)}{\text{vol}(S)} + \frac{\text{cut}(S, T)}{\text{vol}(T)}$$

Нормализованное разрезание (примеры)

$$\text{cut}(S,T) = 0.1 + 0.2 = 0.3$$

$$\text{vol}(S) = 0.3 + 0.6 + 0.8 + 0.8 = 2.5$$

$$\text{vol}(T) = 0.3 + 0.8 + 0.8 + 0.6 = 2.5$$

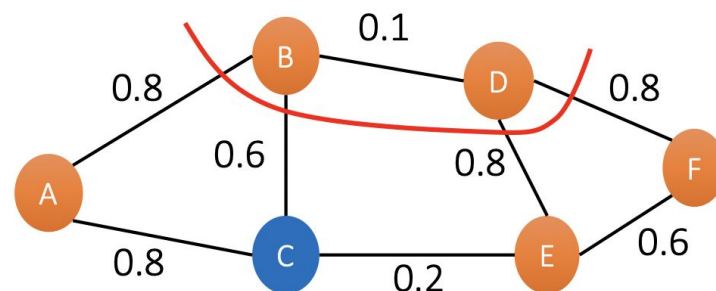
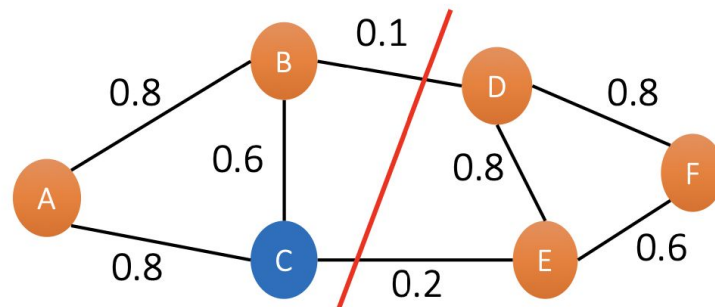
$$\text{Ncut}(S,T) = 0.3/2.5 + 0.3/2.5 = 0.24$$

$$\text{cut}(S,T) = 0.8 + 0.6 + 0.8 + 0.8 = 3.0$$

$$\text{vol}(S) = 3.0 + 0.1 = 3.1$$

$$\text{vol}(T) = 3.0 + 0.8 + 0.2 + 0.6 = 4.6$$

$$\text{Ncut}(S,T) = 3.0/3.1 + 3.0/4.6 = 1.62$$



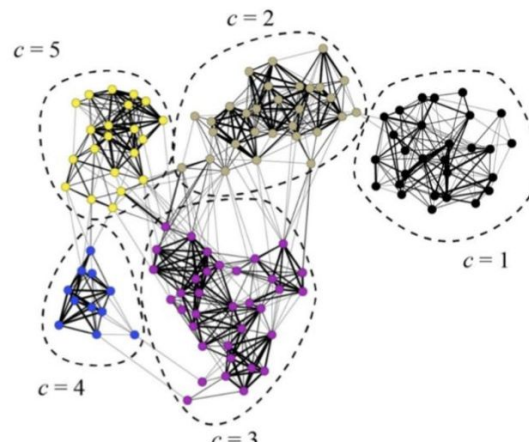
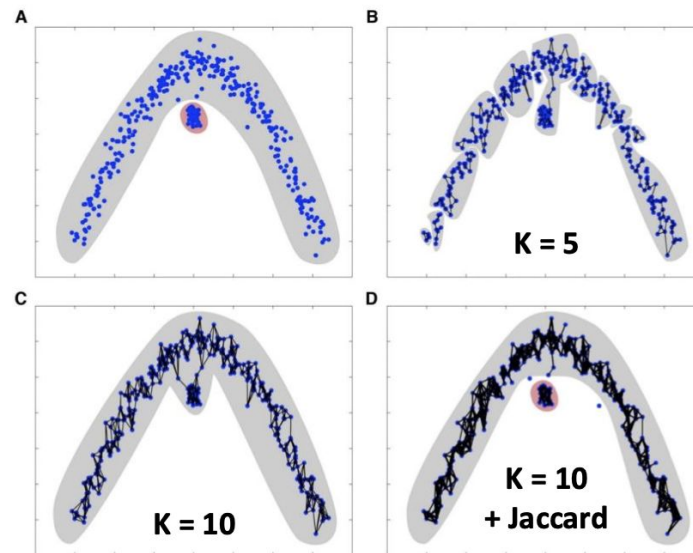
Нормализованное разрезание

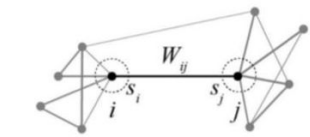
Поиск лучшего нормализованного разрезания — это NP-сложная задача, поэтому для её решения используется несколько эвристик:

1. спектральная кластеризация,
2. кластеризация Маркова,
3. алгоритм Louvain,
4. **алгоритм Leiden**,
5. ...

Алгоритм Shared Nearest Neighbors (SNN)

1. Сначала строится kNN-граф на пространстве PCA,
2. Теперь строится новый граф, вершины в котором клетки, а рёбра — это индекс сходства Джаккарда в соседстве клеток
3. Полученный граф кластеризуется при помощи метода Louvain




$$Q = \frac{1}{2m} \sum_{i,j} \left[W_{ij} - \frac{s_i s_j}{2m} \right] \delta(c_i, c_j)$$

Кластеризация субъективна!

В зависимости от выбора алгоритма, а также его параметров, кластера могут получиться разными

Более того, вы можете выбрать алгоритм или параметры в зависимости от того, что вы ожидаете от данных

Главная задача — это интерпретировать кластера, а не выполнить кластеризацию саму по себе

