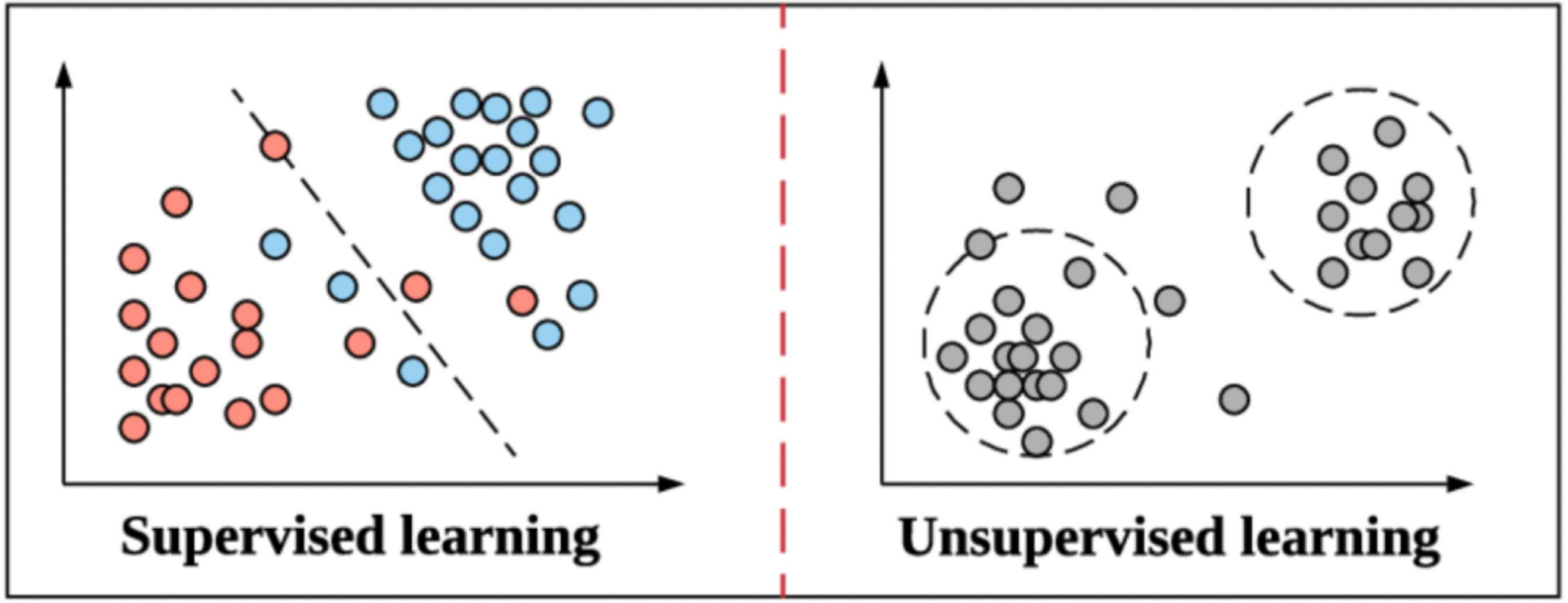
ТИНЬКОФФ

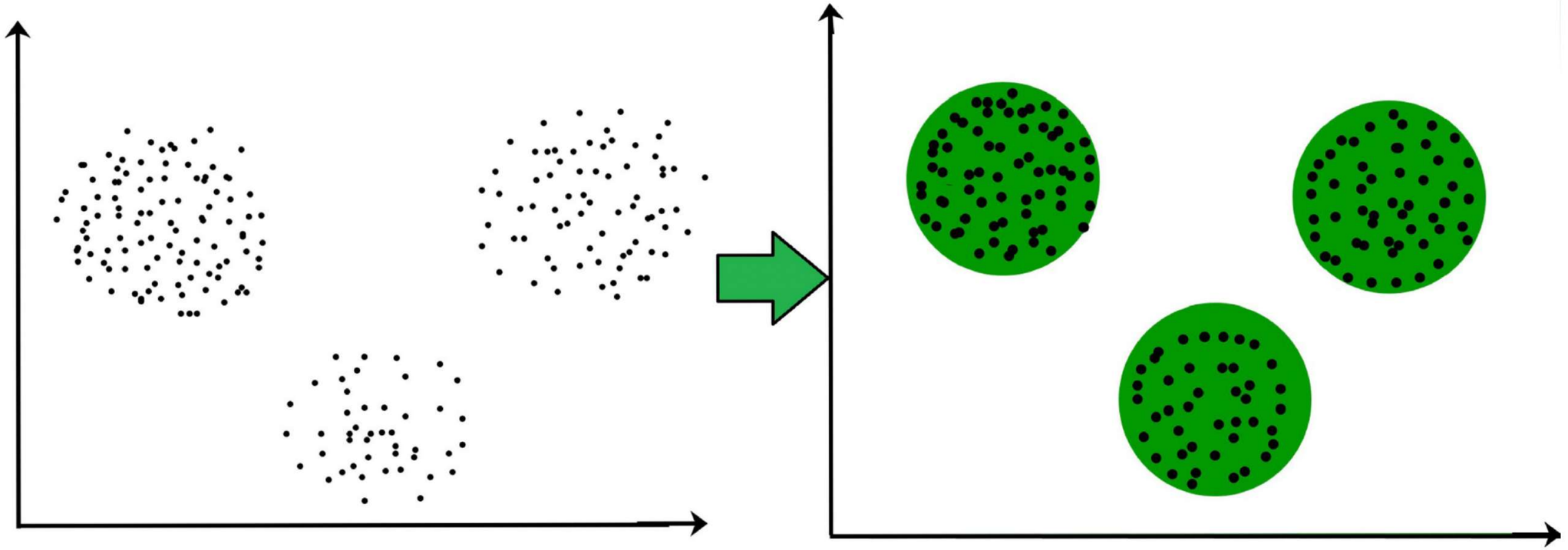
Лекция 8

Unsupervised learning

Supervised VS Unsupervised



Supervised VS Unsupervised



K-Means

Визуализация алгоритма

Алгоритм:

Выбираем K — количество кластеров

Далее итеративно выполняем шаги, пока происходят изменения:

- Обновить кластеры, приписав каждой точке кластер ближайшего центра
- Обновляем центр кластера. Центр = центр масс класстера

Плюсы

Просто и понятно

Минусы

Нужно знать K, слишком простая модель (является выпуклой оболочкой), если плохо выбрать центры, то можем не сойтись

DBSCAN

[Визуализация алгоритма](#)

Параметры:

- eps — расстояние, на котором две вершины считаются соседями
- min_samples — сколько нужно соседей из кластера, чтобы считать вершину коренной вершиной кластера

Алгоритм:

Выбираем случайную точку, которую ещё не просматривали

Проверяем, что в радиусе eps с центром в точке X находится min_samples точек. Если в радиусе меньше точек, то **X — шум**.
Если >= min_samples точек, то:

- Область покрываемая радиусом eps содержится в кластере => X принадлежит данному кластеру
- Область не находится ни в каком кластере => X и другие точки внутри радиуса образуют кластер

Плюсы

- Сам подбирает число кластеров
- Опирается на плотность точек, кластеры могут вытянутыми или даже не выпуклыми

Минусы

- Требуется подбор двух параметров
- Предполагается, что в разных частях данных одинаковая плотность точек

Агломеративная Кластеризация

Алгоритм:

Присваиваем каждой точке свой кластер

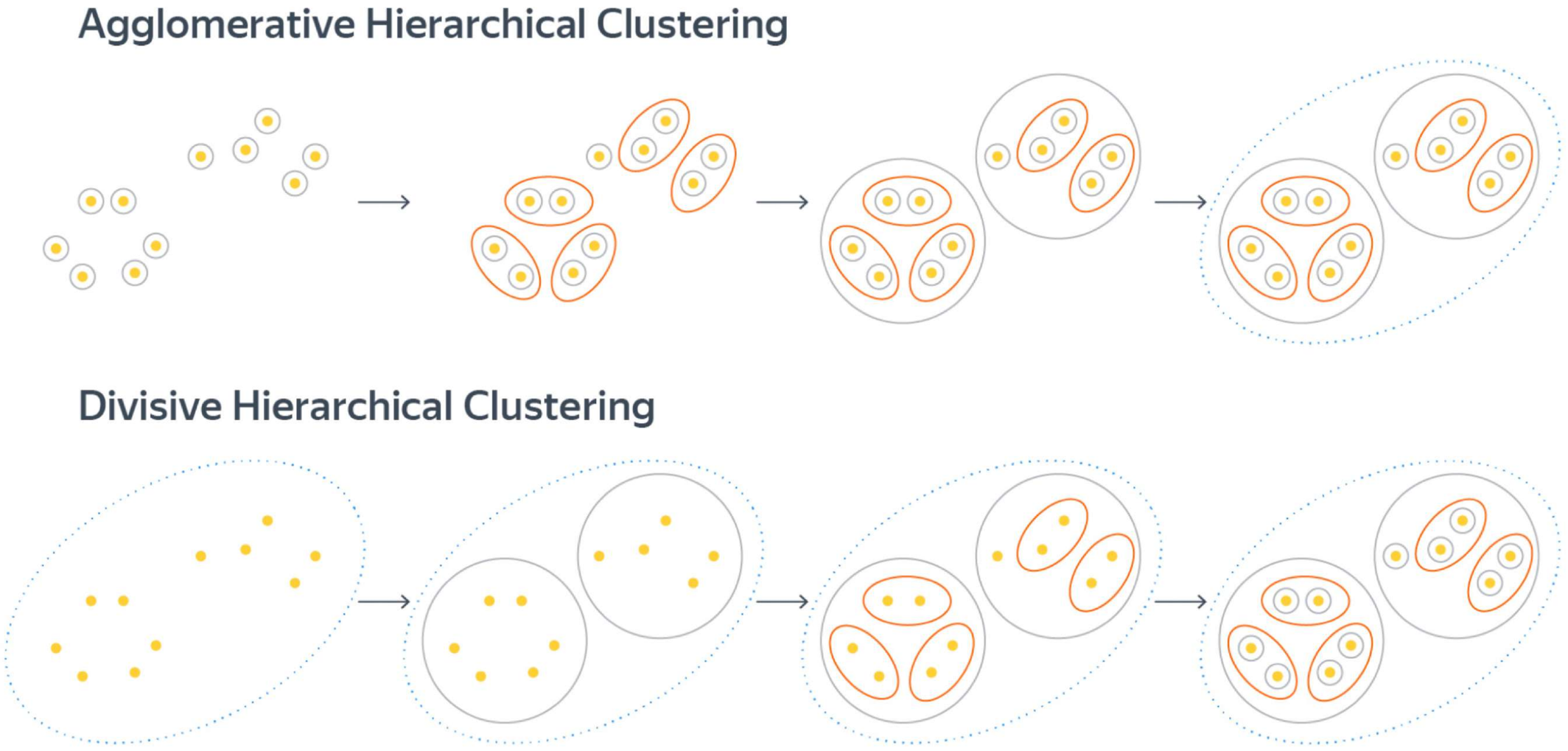
Повторяем итеративно, пока не склеим
все в один кластер:

- Сортируем попарные расстояние между центрами кластеров по возрастанию
- Берем пару ближайших кластеров, склеиваем их в один и пересчитываем центры

Итог:

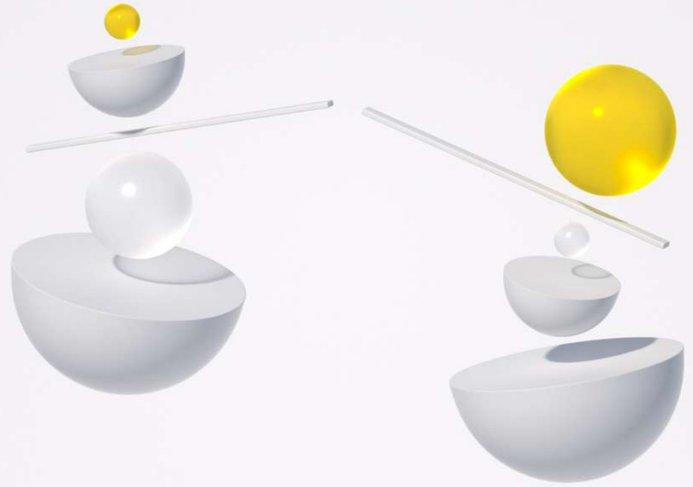
Получаем дерево склеиваний. Выбираем шаг,
на котором получили требуемое число кластеров

Агломеративная Кластеризация



Агломеративная Кластеризация

Как выбрать ближайшие кластеры?



- Расстояния между центрами
- Минимум попарных расстояний между точками из двух кластеров
- Максимум попарных расстояний между точками из двух кластеров
- Среднее попарных расстояний между точками из двух кластеров

Оценка качества. Простые метрики

- Среднее внутрикластерное расстояние
- Среднее межкластерное расстояние

Оценка качества

Коэффициент силуэта

$$S(x_i) = \frac{B(x_i) - A(x_i)}{\max(B(x_i), A(x_i))}$$

- $\boldsymbol{B(xi)}$

  среднее расстояние между и объектами следующего ближайшего кластера

- $\boldsymbol{A(xi)}$

  среднее расстояние между $\boldsymbol{x_i}$ и объектами того же кластера

ТИНЬКОФФ

Он такой один