КЛАСТЕРИЗАЦИЯ

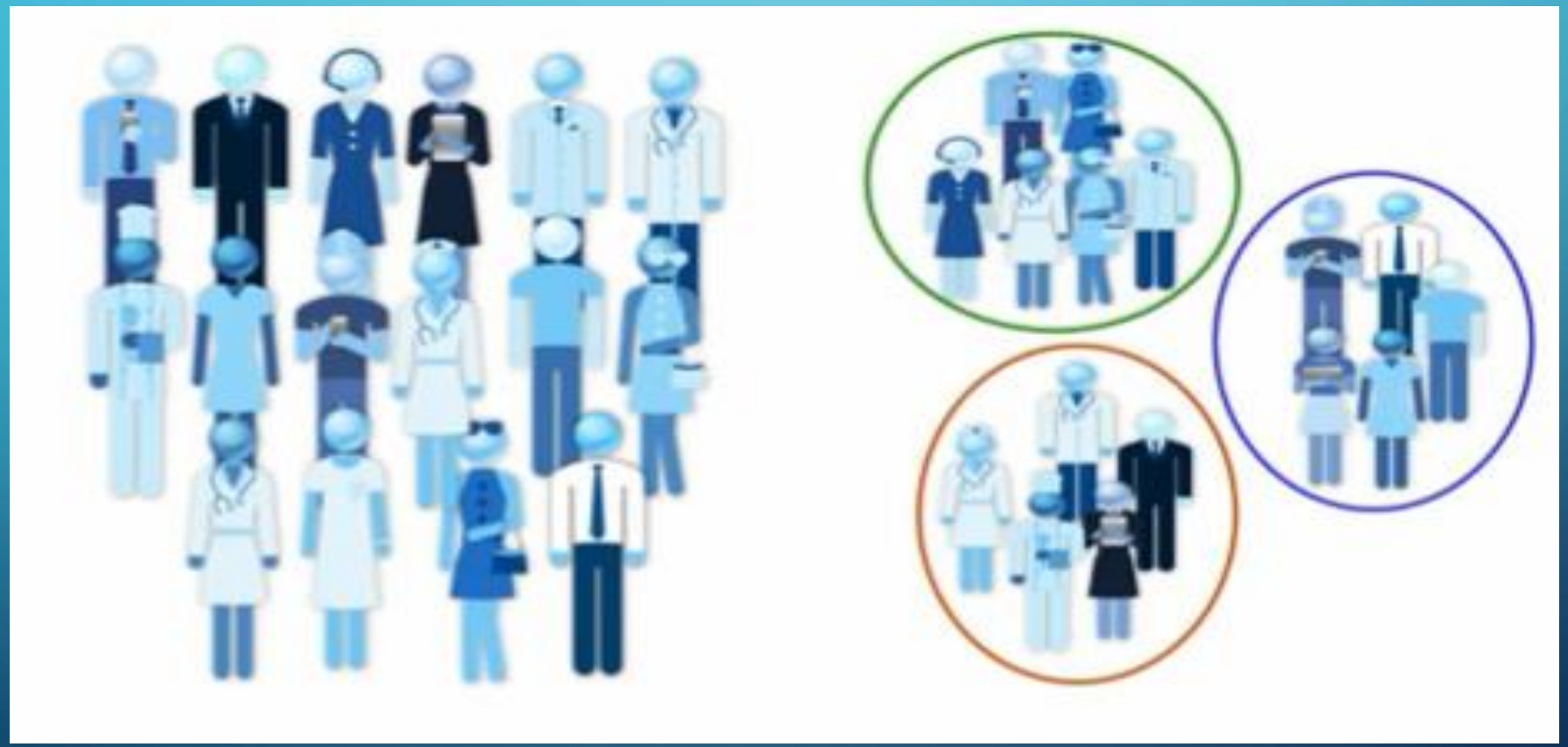
СЕМИНАР *5*

КЛАСТЕРИЗАЦИЯ

- Кластеризация – процесс разделения объектов по определенным признакам (обучение без учителя).
- Схожие объекты мы причисляем к одному классу.
- Проблема в том, что мы заранее не знаем какие объекты к какому классу будут отнесены.

КАК ЭТО РАБОТАЕТ?

Определение людей на группы/категории



ОБЛАСТИ ПРИМЕНЕНИЯ КЛАСТЕРИЗАЦИИ

- Ритейл/маркетинг
- Страховки
- Банки
- Медицина
- Публикации

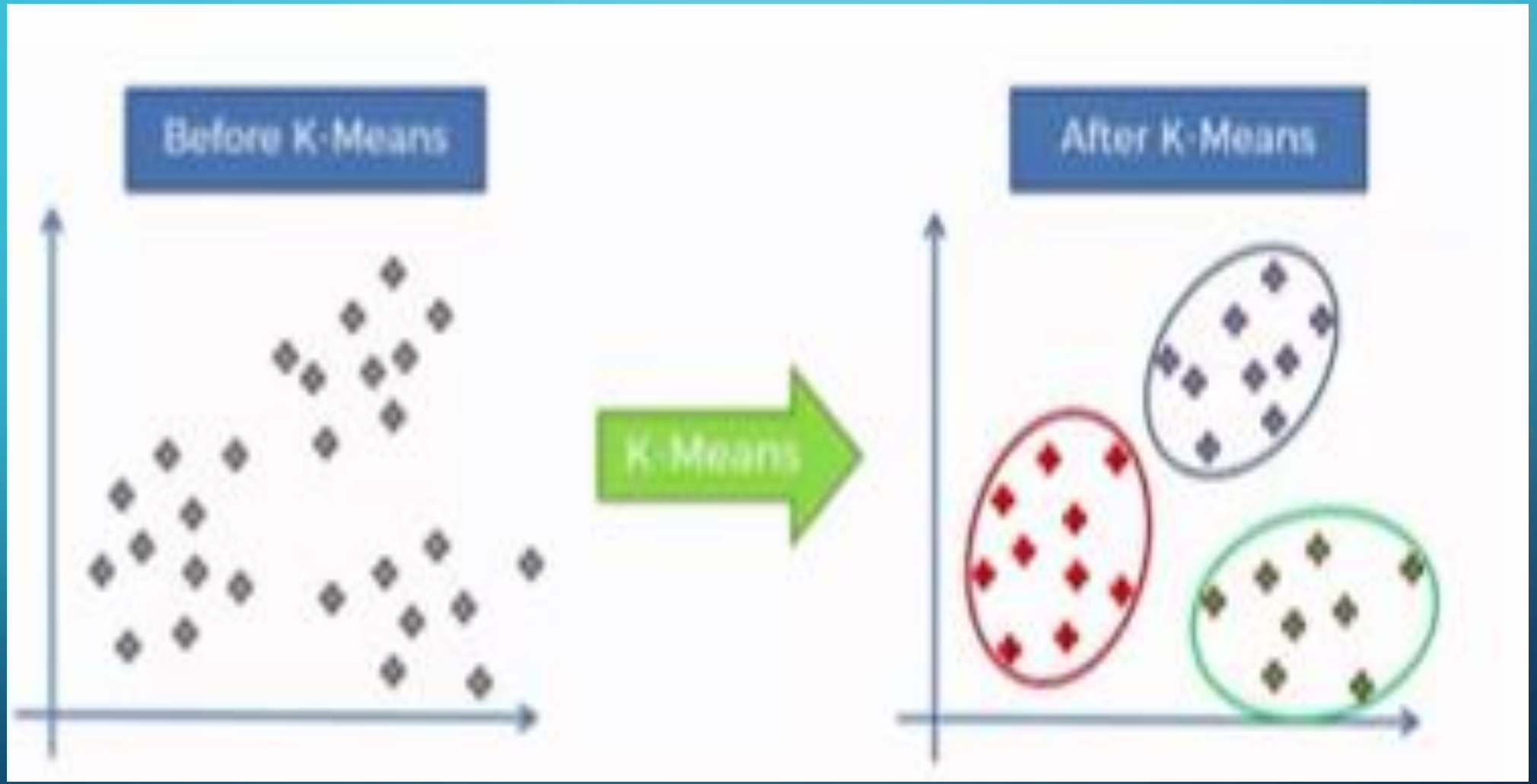
КАК МЫ БУДЕМ КЛАСТЕРИЗОВАТЬ ДАННЫЕ?

- Размещение данных в пространстве
- Подсчет расстояния между данными (точками)

КАК ОПРЕДЕЛИТЬ СХОЖЕСТЬ ДАННЫХ?

- Теорема Пифагора в случае двумерной плоскости
- Евклидово пространство в многомерном случае

K-MEANS



K-MEANS. АЛГОРИТМ

- Определение k центроида
- Подсчет расстояния данных до центроида
- Поиск для центроида ближайших точек
- Расчет нового центроида

КОРРЕКТНОЕ КОЛИЧЕСТВО КЛАСТЕРОВ

- Кластер строится на основе средних расстояний от центроидов до точек
- Количество кластеров = количество центроидов
- Если видно некорректное деление на кластеры, то включаем, например, метод локтя

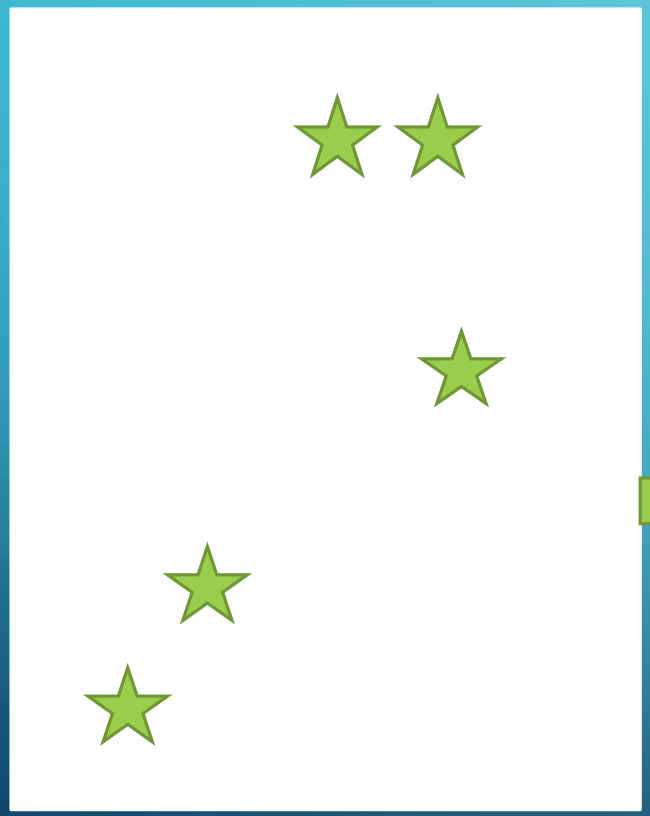
ПРИМЕР

- colab.research.google.com/drive/1xqHgQSd29rzISLaQLWfKpSc7rT3M9KMi#scrollTo=EfCDsBHOoZr_

ЗАДАНИЕ

- Проанализировать файл csv
- Методом локтя определить оптимальное количество кластеров
- Кластеризовать данные
- Визуализировать кластеризацию

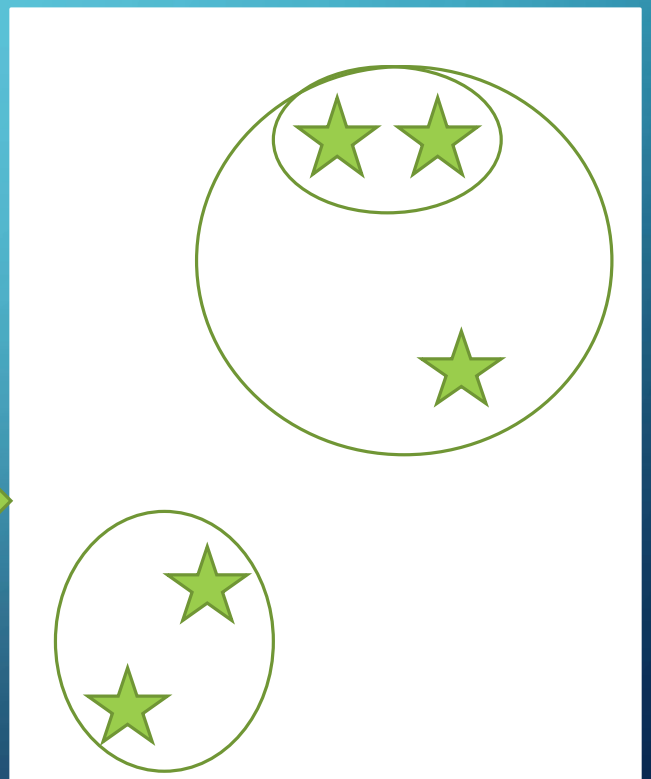
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ



Объединение в кластеры

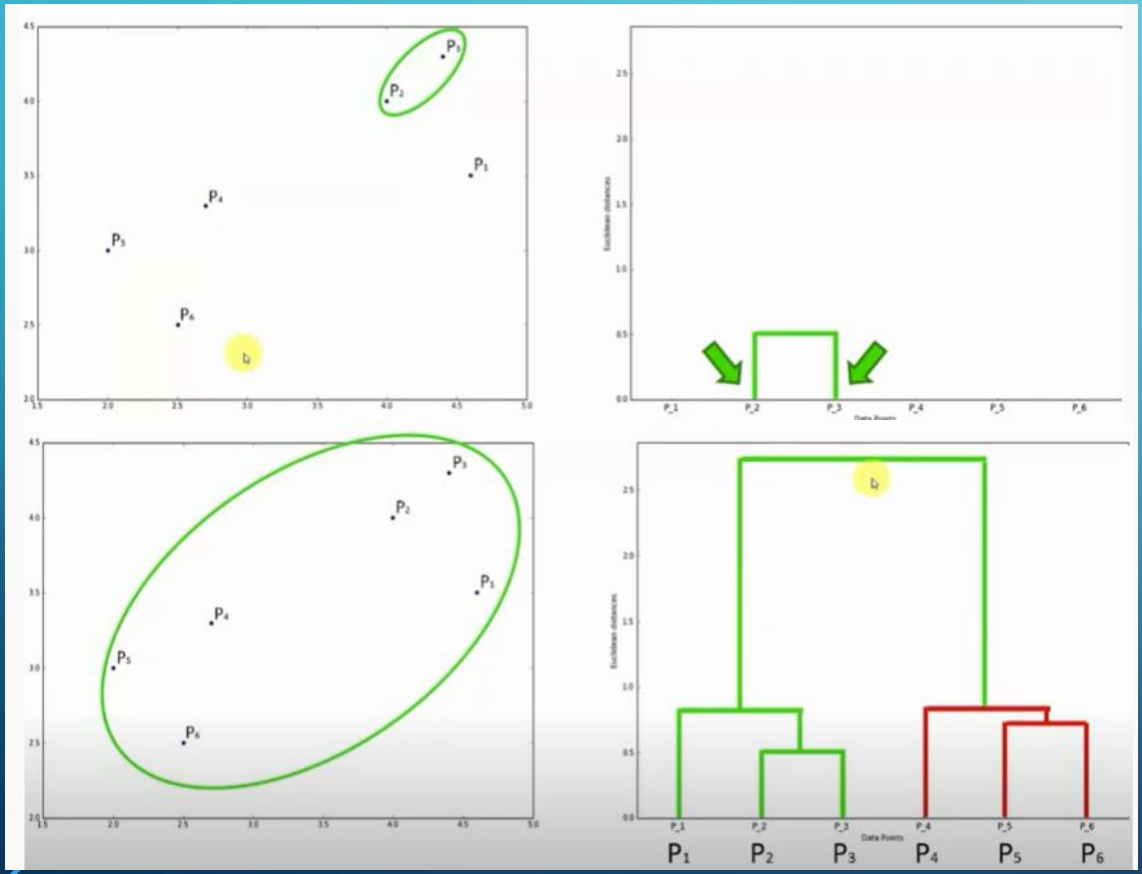
Расстояние между точками = евклидово расстояние

Расстояние между кластерами = по ближайшему соседу, по среднему, ...



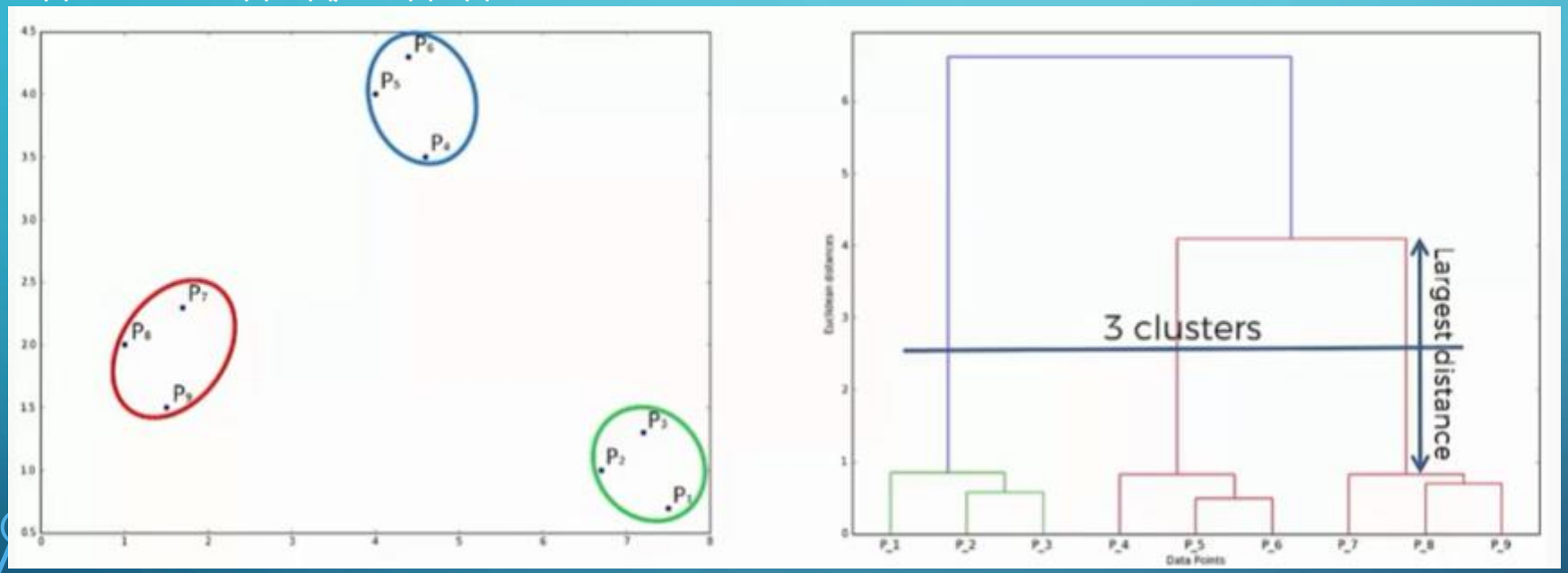
КАК ВЫБРАТЬ КОЛИЧЕСТВО КЛАСТЕРОВ?

- Дендрограмма

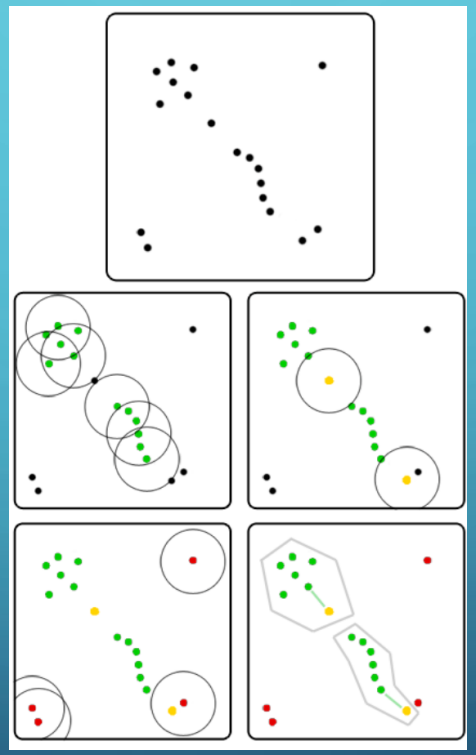


РАЗУМНЫЙ ВЫБОР КОЛИЧЕСТВА КЛАСТЕРОВ

Задача: определить в какой момент произошло объединение далеких друг от друга кластеров?
Удобный подход, когда данных немного.



DBSCAN. ВИЗУАЛИЗАЦИЯ



DBSCAN. АЛГОРИТМ

1. Подходим к случайному человеку из толпы.

2. Если рядом с ним меньше трёх человек, переносим его в список возможных отшельников и выбираем кого-нибудь другого.

3. Иначе:

- Исключаем его из списка людей, которых надо обойти.
- Вручаем этому человеку зелёный флажок и создаём новую группу, в которой он пока что единственный обитатель.
- Обходим всех его соседей. Если его сосед уже в списке потенциальных одиночек или рядом с ним мало других людей, то перед нами край толпы. Для простоты можно сразу пометить его жёлтым флагом, присоединить к группе и продолжить обход. Если сосед тоже оказывается «зелёным», то он не создает новую группу, а присоединяется к уже созданной; кроме того мы добавляем в список обхода соседей соседа. Повторяем этот пункт, пока список обхода не окажется пуст.

4. Повторяем шаги 1-3, пока так или иначе не обойдём всех людей.

5. Разбираемся со списком отшельников. Если на шаге 3 мы уже раскидали всех краевых, то в нём остались только выбросы-одиночки — можно сразу закончить. Если нет, то нужно как-нибудь распределить людей, оставшихся в списке.

ПРАКТИКА

- colab.research.google.com/drive/1e_O_CvQC4K47aZ2x6m0QeACU_JaJYgeM#scrollTo=UlenNZWfzfGY

ЗАДАНИЕ

- Проанализировать файл xls/csv
- Кластеризовать данные методом k-means
- Визуализировать кластеризацию
- Кластеризовать данные DBScan
- Визуализировать кластеризацию
- Зафиксировать выбросы после кластеризаций
- Сделать выводы