СӘТБАЕВ УНИВЕРСИТЕТІ SATBAYEV UNIVERSITY

Основы искусственного интеллекта. Лекция 6. Кластеризация и уменьшение размерности. Методы кластеризации (K-means, DBSCAN), методы уменьшения размерности (PCA, t-SNE)

Преподаватель: **Қасенхан Арай Мейрамбайқызы, доктор PhD, ассоциированный профессор кафедры «Программная инженерия»**

a.kassenkhan@satbayev.university

Содержание

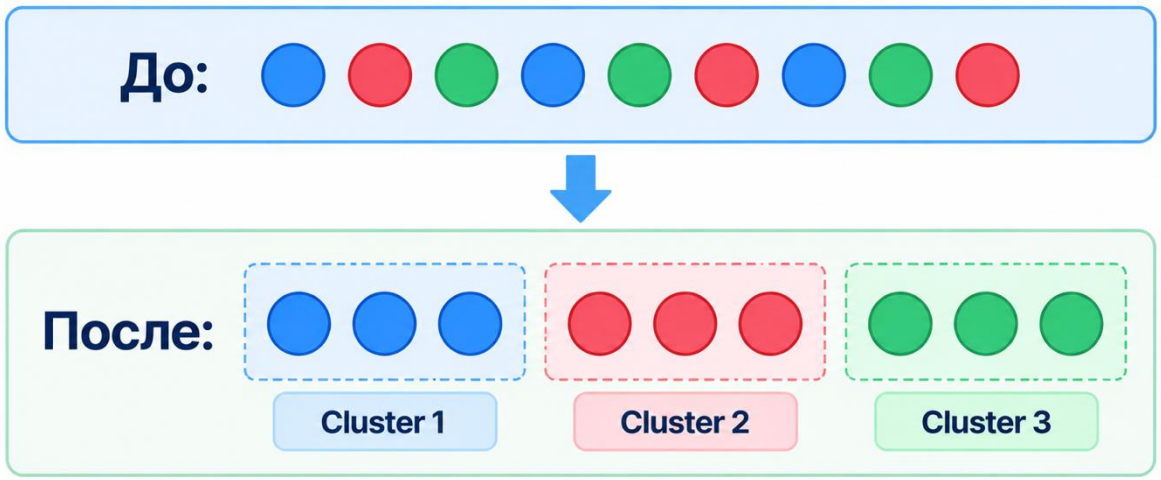
1. Что такое кластеризация
2. K-Means
3. DBSCAN
4. Сравнение методов кластеризации
5. Зачем уменьшать размерность
6. PCA
7. t-SNE
8. Сравнение PCA и t-SNE

По завершению урока Вы будете знать:

1. **Чем кластеризация отличается от классификации**
2. **Как работает алгоритм K-Means**
3. **Зачем нужен DBSCAN и чем он отличается от K-Means**
4. **Зачем уменьшать количество признаков**
5. **Как работает PCA**
6. **Для чего используется t-SNE**
7. **Как выбрать подходящий метод для конкретной задачи**

Что такое кластеризация?

Кластеризация — группировка похожих объектов без заранее известных меток.



Классификация
Есть метки классов заранее.

Кластеризация
Метки неизвестны — модель сама ищет группы.

Кластеризация = обучение без учителя

Где используется кластеризация?

Клиенты

Сегментация по поведению и интересам



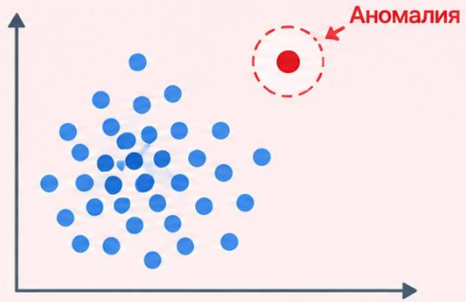
Изображения

Группировка похожих объектов



Аномалии

Обнаружение необычных объектов



Социальные сети

- Группировка пользователей
- Анализ интересов
- Рекомендации контента

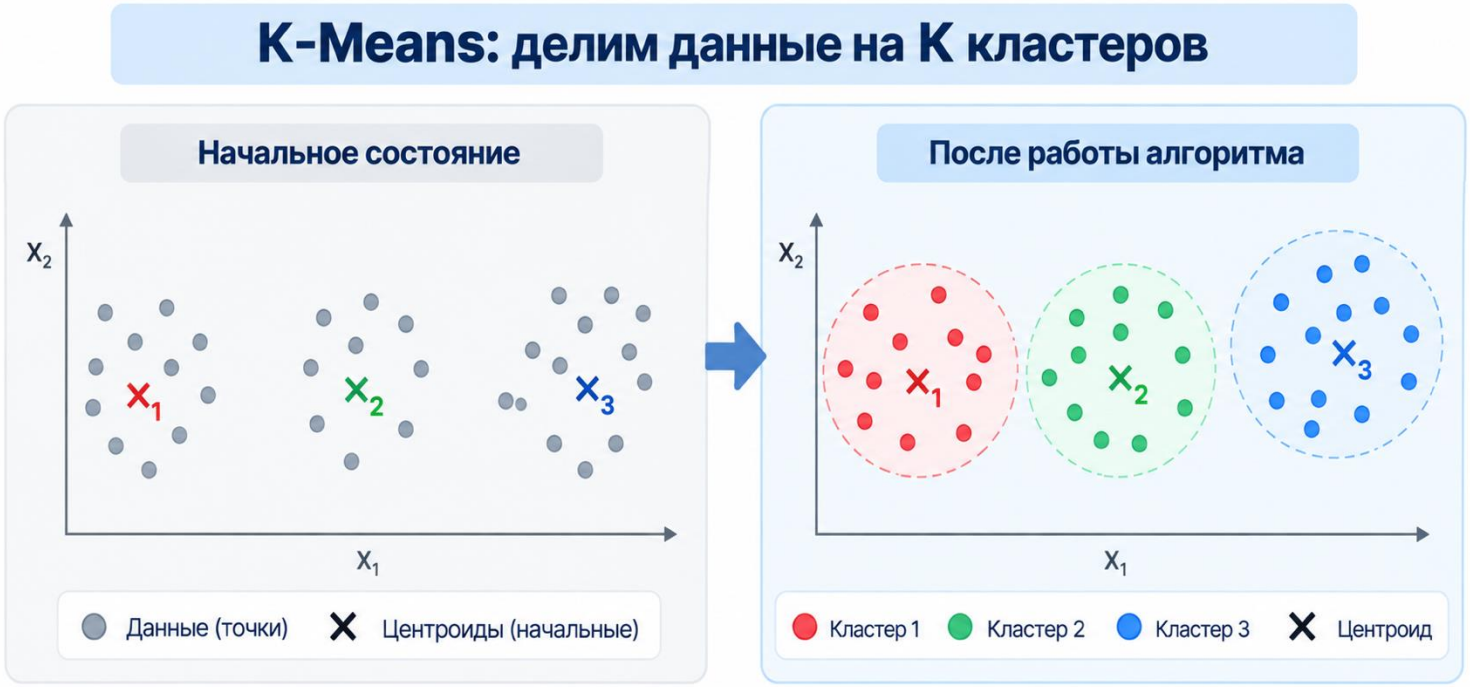
Медицина

- Группировка пациентов
- Выявление подгрупп заболеваний
- Персонализированное лечение

Маркетинг

- Сегментация клиентов
- Анализ поведения
- Персонализированные предложения

K-Means: основная идея



Идея:

1. выбираем K;
2. создаём K центроидов;
3. каждая точка идёт к ближайшему центроиду;
4. центроиды пересчитываются;
5. повторяем до стабилизации.

Цель — минимизировать расстояние точек до центров кластеров.

K-Means в Scikit-learn

```
from sklearn.cluster import KMeans

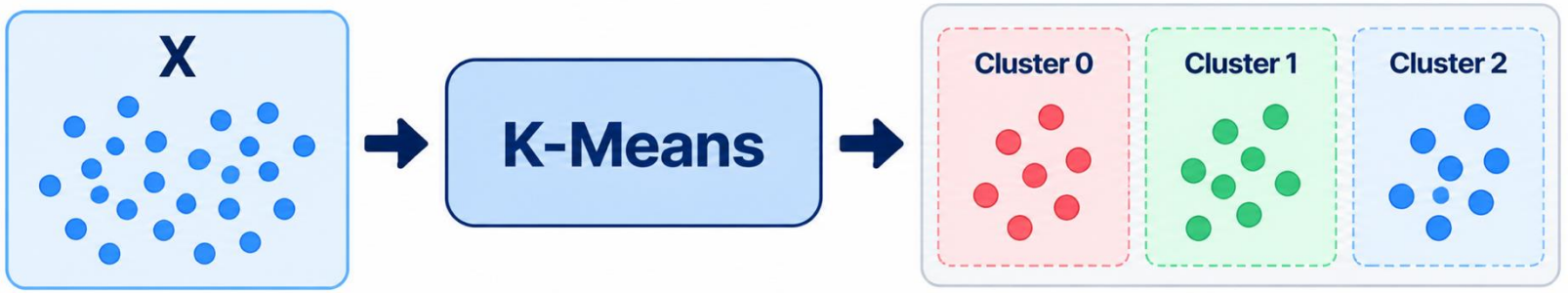
model = KMeans(
    n_clusters=3,
    random_state=42
)

labels = model.fit_predict(X)

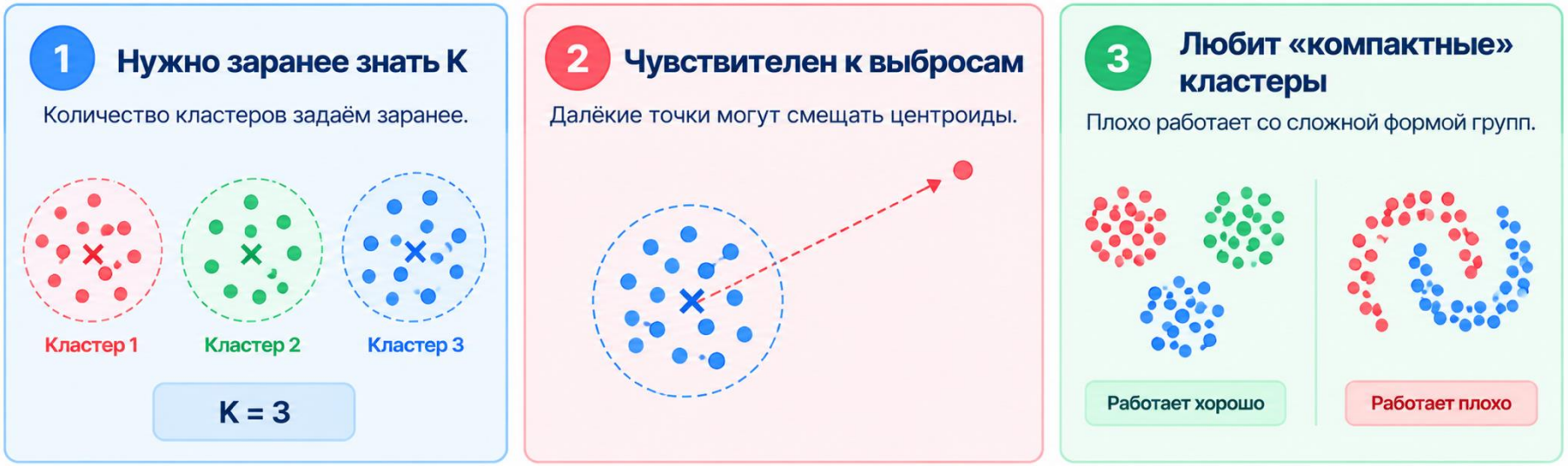
print(labels)
```

n_clusters=3 → хотим 3 кластера

fit_predict() → обучить + получить номер кластера

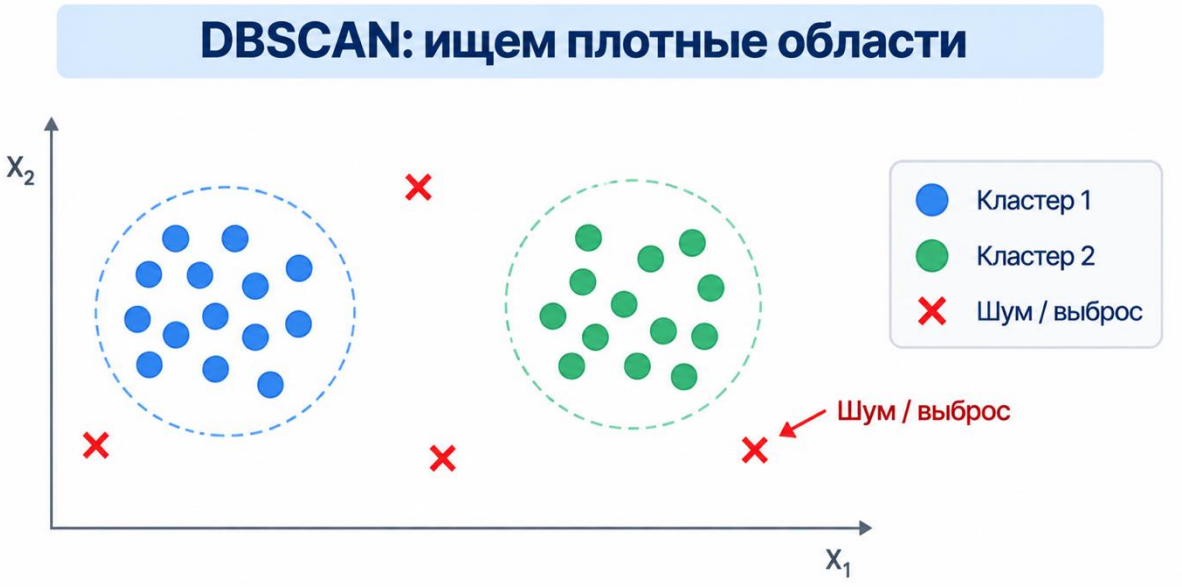


Ограничения K-Means



K-Means хорошо работает, когда кластеры компактны и хорошо разделены.

DBSCAN: кластеры по плотности



DBSCAN не требует заранее задавать число кластеров.

eps — радиус соседства;
min_samples — сколько точек нужно рядом.

Плотные области → кластеры, одиночные точки → шум

DBSCAN в Scikit-learn

```
from sklearn.cluster import DBSCAN

model = DBSCAN(
    eps=0.5,
    min_samples=5
)

labels = model.fit_predict(X)
```

eps → насколько близкими должны быть точки

min_samples → сколько соседей нужно для плотной области

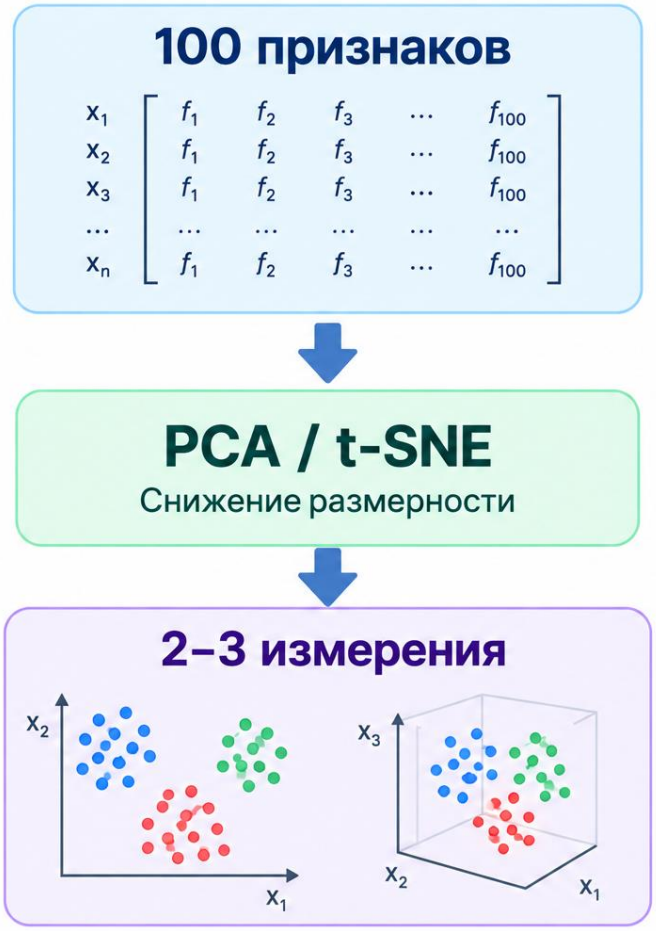
-1 → шум / выброс

K-Means или DBSCAN?

Характеристика	K-Means	DBSCAN
Нужно задать число кластеров	Да	Нет
Основа	расстояние до центроида	плотность
Выбросы	чувствителен	умеет выделять шум
Сложная форма кластеров	плохо	хорошо
Параметры	k	eps, min_samples

Нет лучшего алгоритма для всех данных — метод выбирается по структуре данных.

Зачем уменьшать размерность?

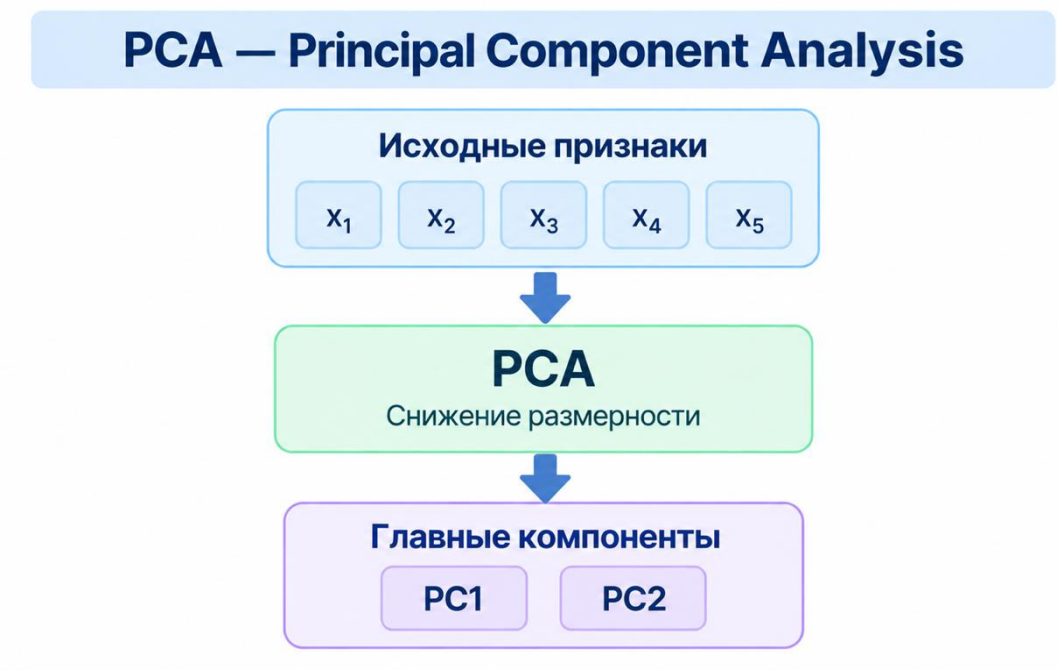


Зачем?

- **меньше вычислений;**
- **меньше шума;**
- **проще визуализация;**
- **проще анализ.**

Уменьшаем количество признаков, стараясь сохранить важную информацию.

PCA: главные компоненты



Идея

PCA ищет новые направления, которые объясняют наибольшую вариацию данных.

PC1 → максимальная дисперсия

PC2 → следующая по важности

PCA — линейное уменьшение размерности.

PCA в Scikit-learn

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)

X_pca = pca.fit_transform(X)

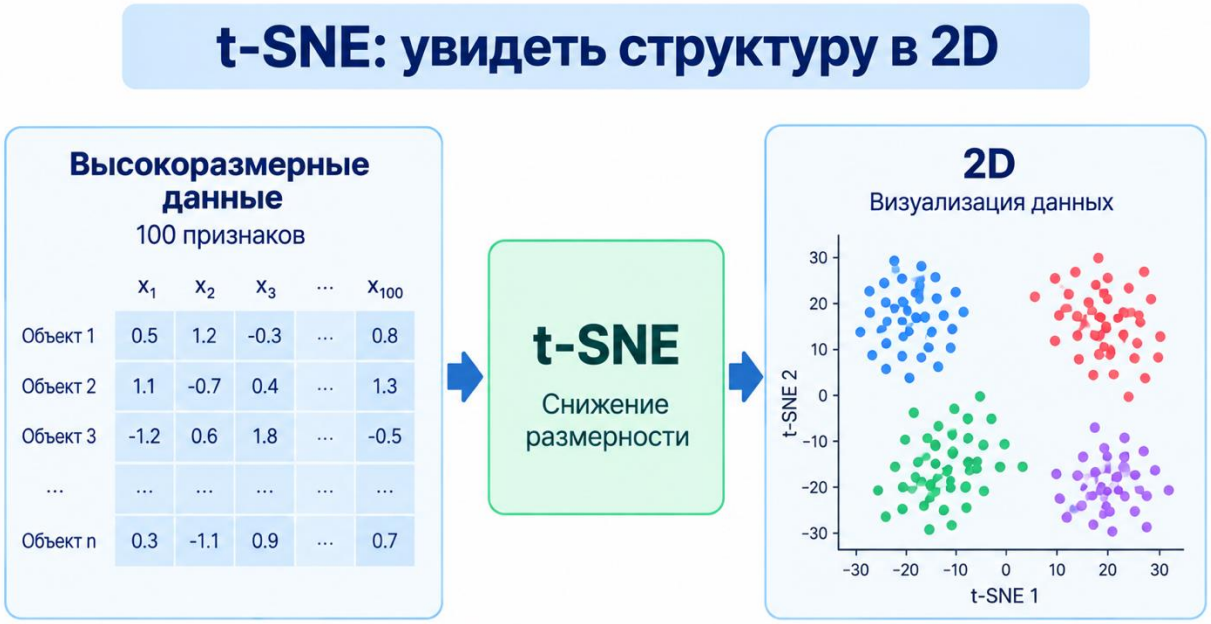
print(X.shape)
print(X_pca.shape)
```

(150, 4)
↓
(150, 2)

n_components=2 → оставляем 2 компоненты

4 признака → 2 главные компоненты

t-SNE: визуализация сложных данных

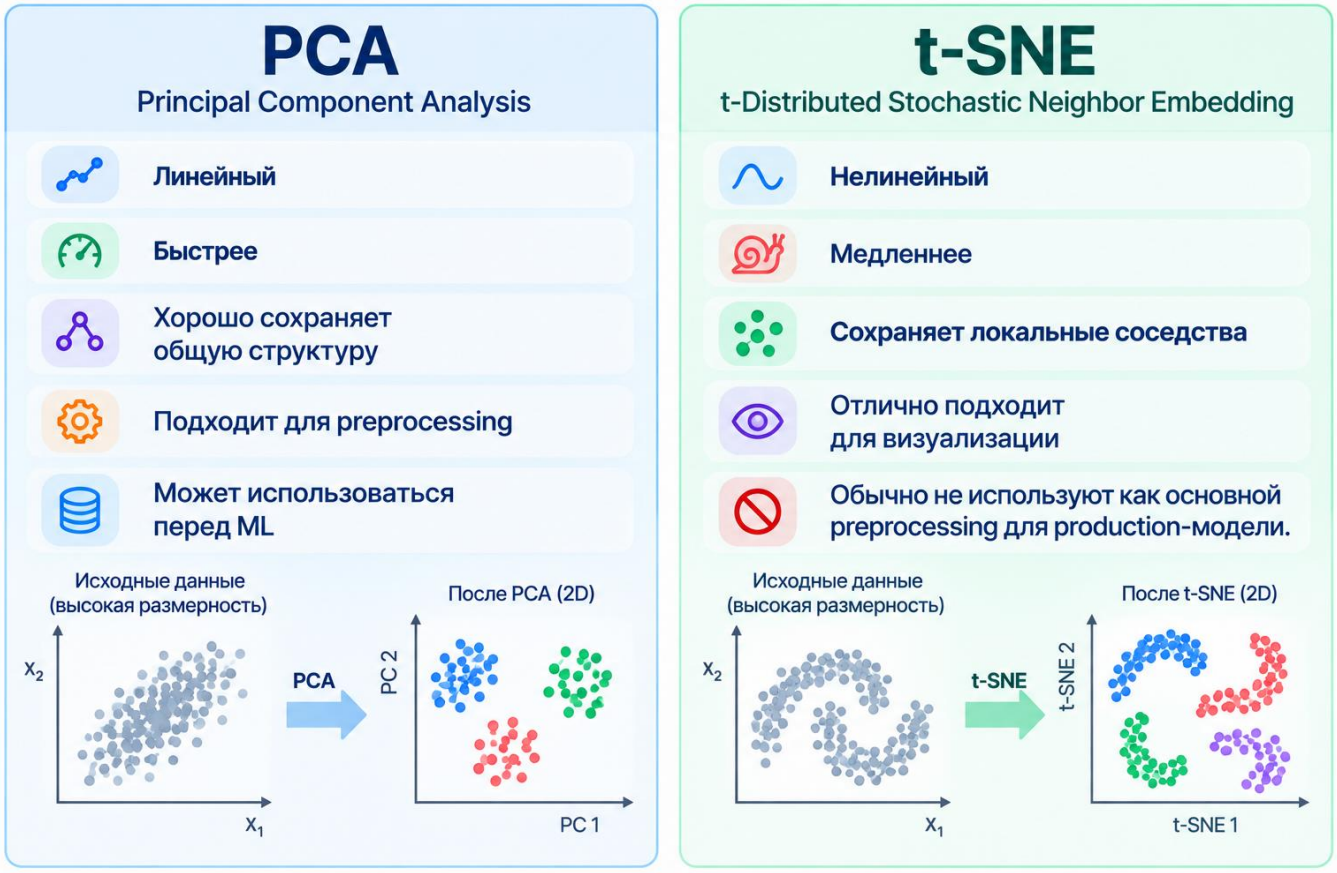


Идея

- сохраняет локальное соседство;
- похожие объекты остаются рядом;
- хорошо показывает группы и кластеры;
- используется главным образом для визуализации.

t-SNE помогает увидеть структуру сложных многомерных данных.

PCA или t-SNE?



PCA → сжатие и подготовка данных

t-SNE → визуализация структуры

Весь процесс



K-Means → центроиды

DBSCAN → плотность

PCA → главные компоненты

t-SNE → 2D-визуализация

Заключение

Сегодня мы рассмотрели:

- **кластеризацию как обучение без учителя;**
- **K-Means и работу с центроидами;**
- **DBSCAN и кластеризацию по плотности;**
- **уменьшение размерности;**
- **PCA;**
- **t-SNE.**



Кластеризация помогает найти структуру в неразмеченных данных, а уменьшение размерности — сделать сложные данные более понятными и управляемыми.