ДЗ обучение с учителем

r2_score(y_test, y_pred)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

ДЗ обучение без учителя

model.fit() - определяет координаты центроидов

model.predict() - предсказывает к какому кластеру относится объект

model.fit_predict() - делает сразу то и другое

Обучение без учителя

План вебинара

- Кластеризация
  - K-means
  - Другие методы
-
- Понижение размерности
  - PCA
  - t-SNE

Обучение без учителя

Есть только описания объектов

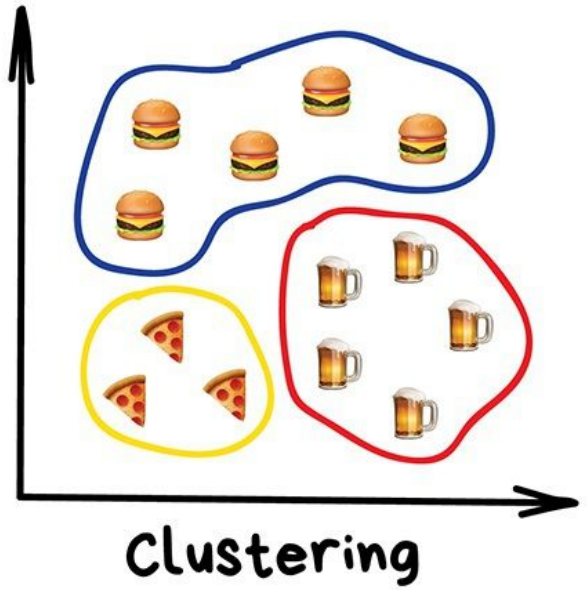
Задача обнаружить внутренние закономерности и взаимосвязи между признаками

Обучение без учителя

- Разметка данных дорого или трудно получить
- Поиск ошибок и скрытых зависимостей
- Способ убрать сильно коррелирующие признаки
- Исключение малоинформативных признаков
- Визуализация

Применение кластеризации

1. Сегментация пользователей, рекомендательные системы
2. Объединение объектов на карте
3. Сжатие изображений
4. Определение аномального поведения



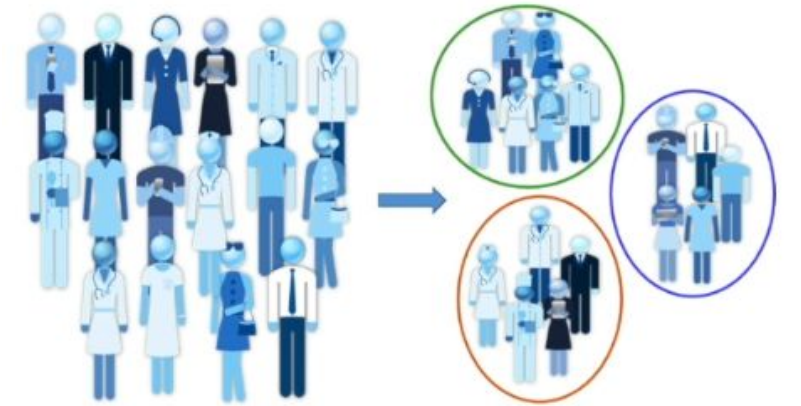
Сегментация пользователей

Собираем данные о пользователях: пол, возраст, что покупал, заходил ли на страницу “page N” сайта, потратил там X секунд, открыл 50% рекламных писем, email на yandex итд.

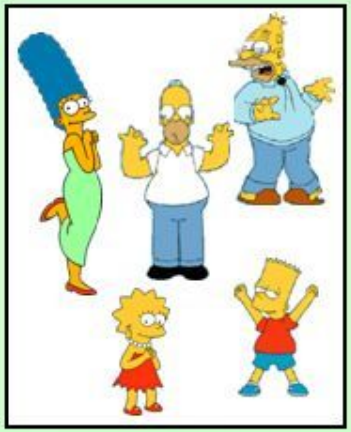
Кластеризуем

Смотрим средние значения признаков в кластерах

Понимаем, что кластер из 13-летних девочек заходит на страницы и ничего не покупает, не включаем их в показ рекламы

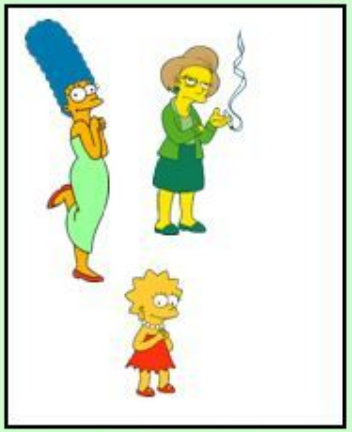


Кластеризация субъективна



Simpson's Family

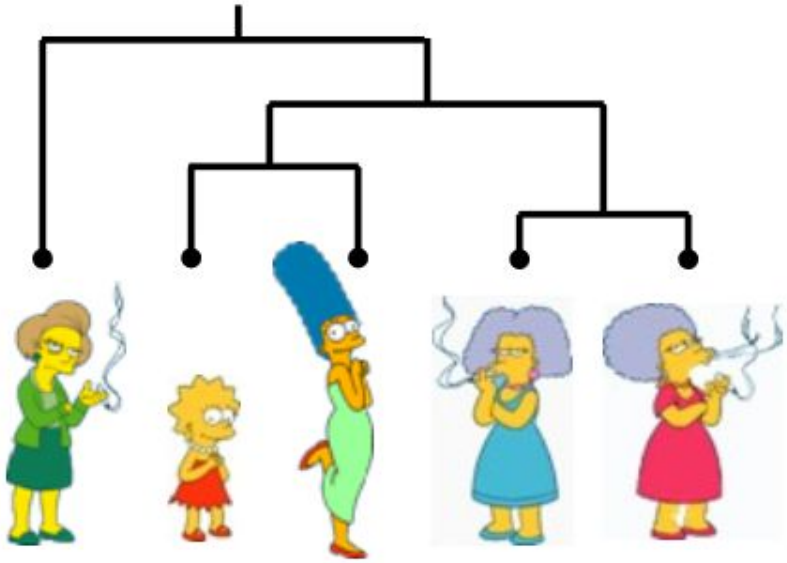
School Employees

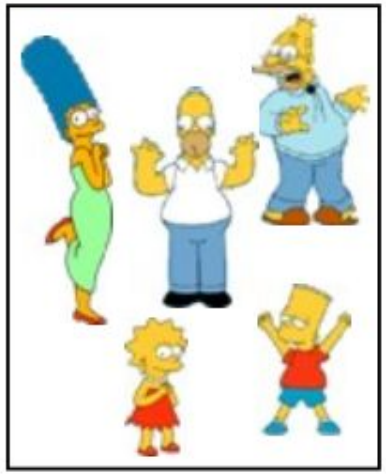


Females

Males

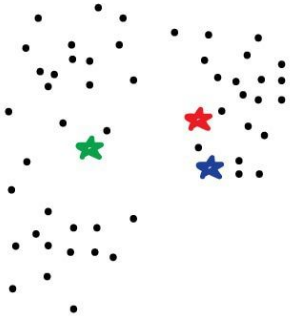
Иерархическая и плоская кластеризация



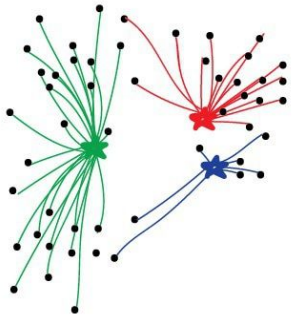


Ставим три ларька с шаурмой оптимальным образом

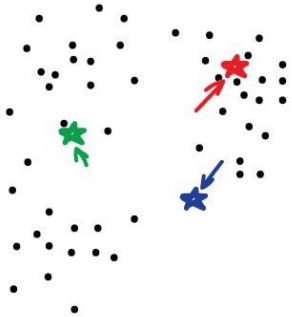
(иллюстрируя метод K-средних)



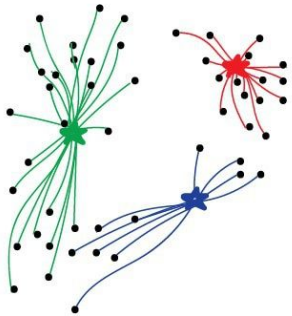
1. Ставим ларьки с шаурмой в случайных местах



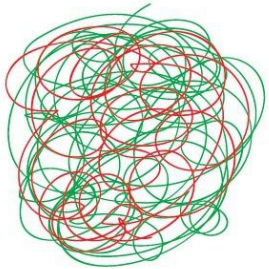
2. Смотрим в какой кому ближе идти



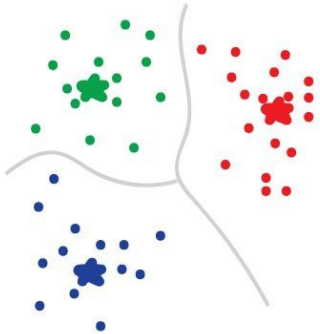
3. Двигаем ларьки ближе к центрам их популярности



4. Снова смотрим и двигаем



5. Повторяем много раз



6. Готово, вы великолепны!

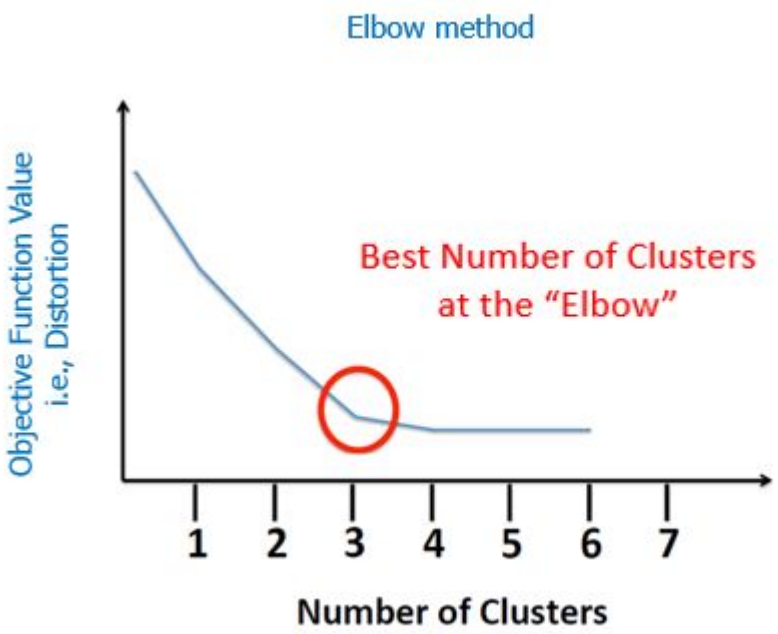
K-means

Кластеры сферические

Хорошо отделимы

Примерно одинаковое число точек в каждом

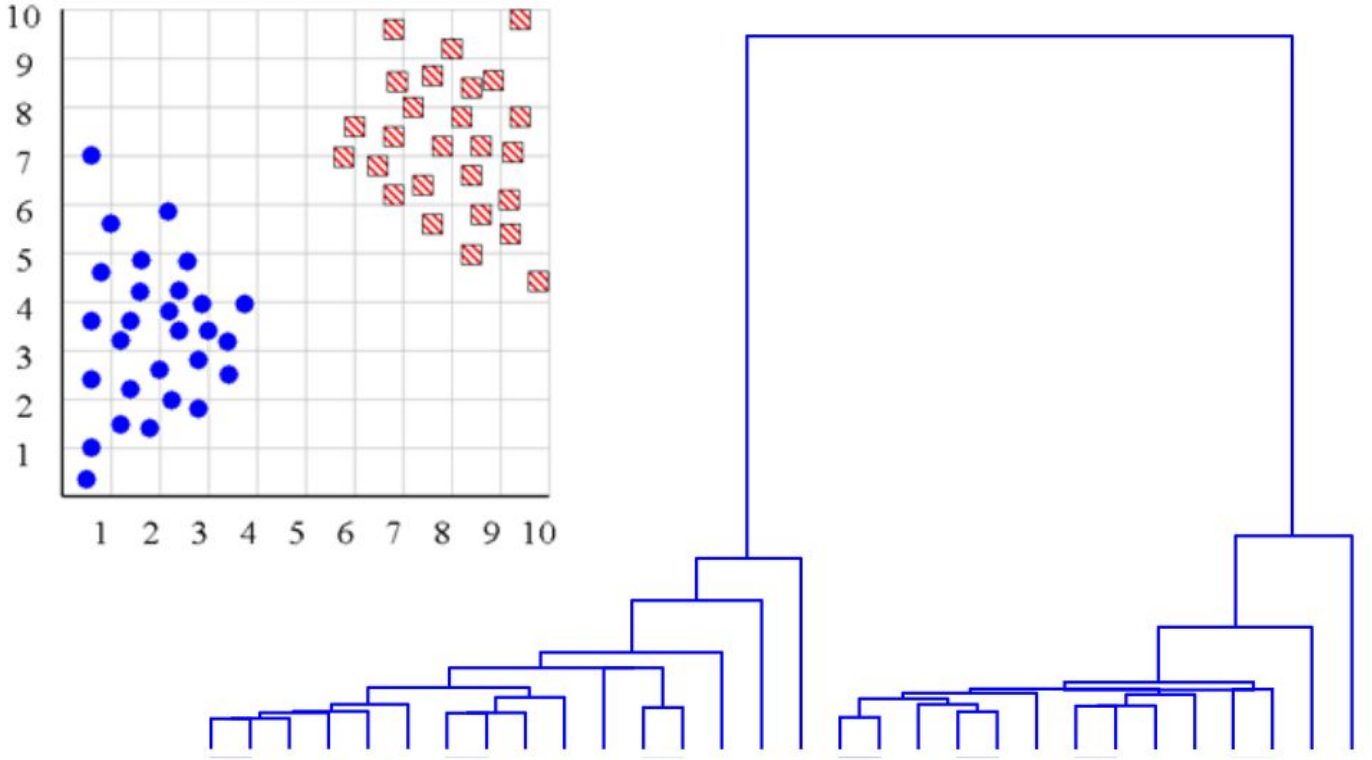
Как определить кол-во кластеров



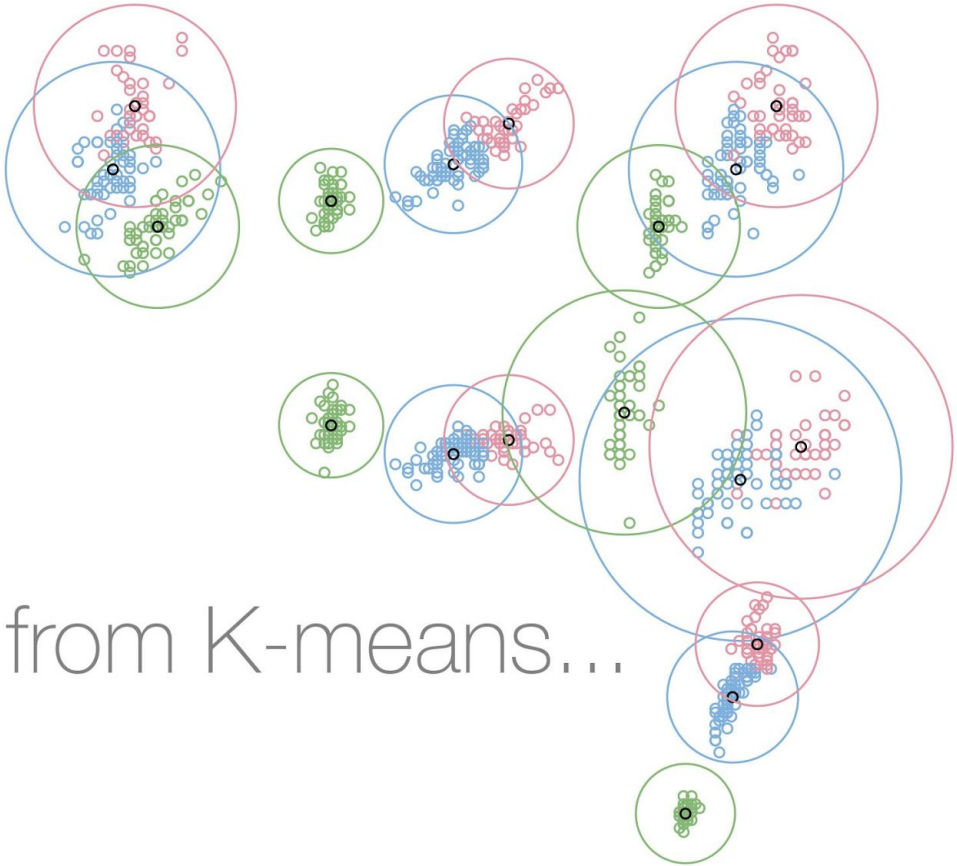
Метод локтя - строим для всех кластеры и смотрим когда происходит перелом с резкого падения на почти плато

tSNE

Метод дерева

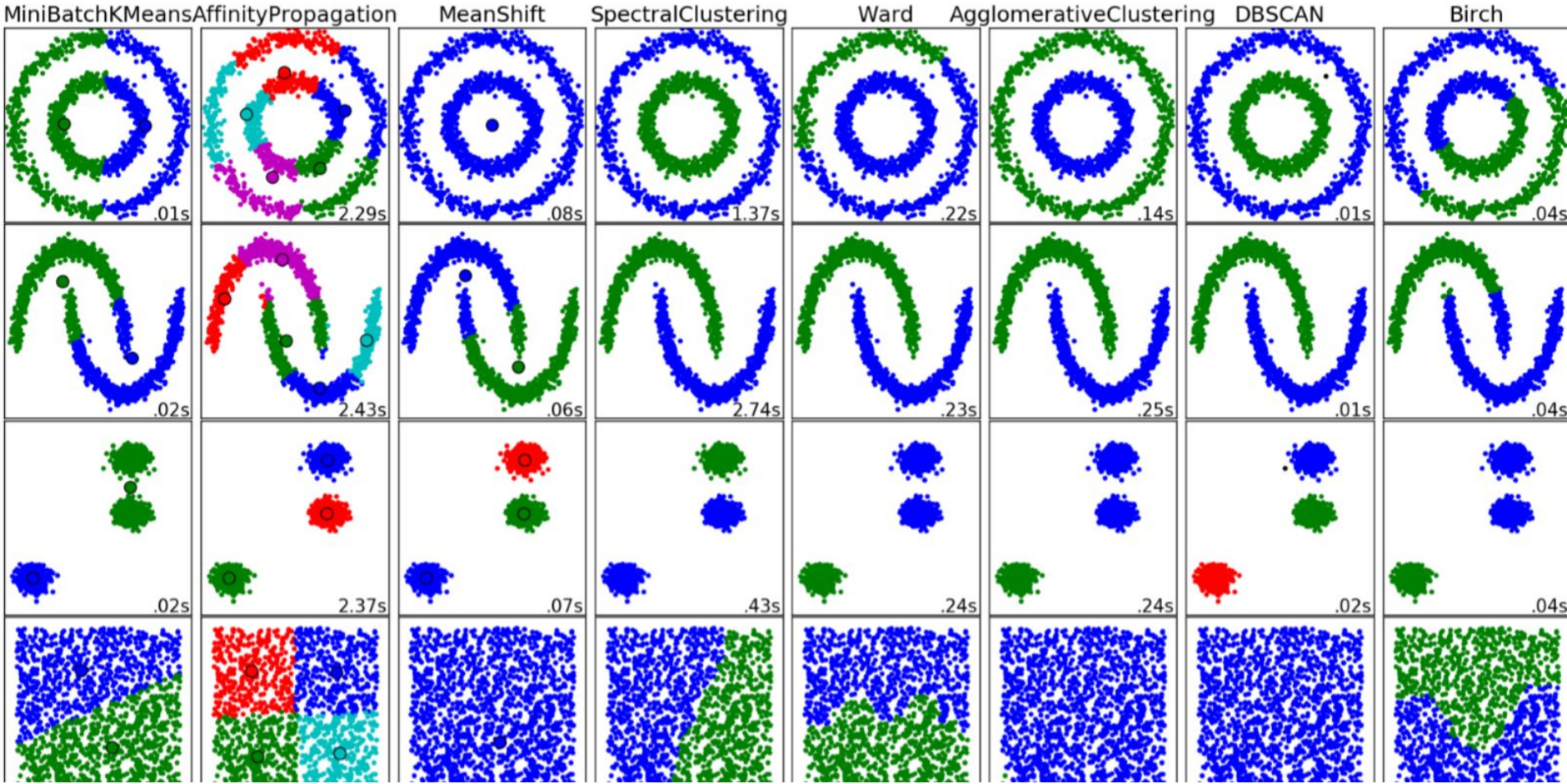


Смешанные Гауссовские модели

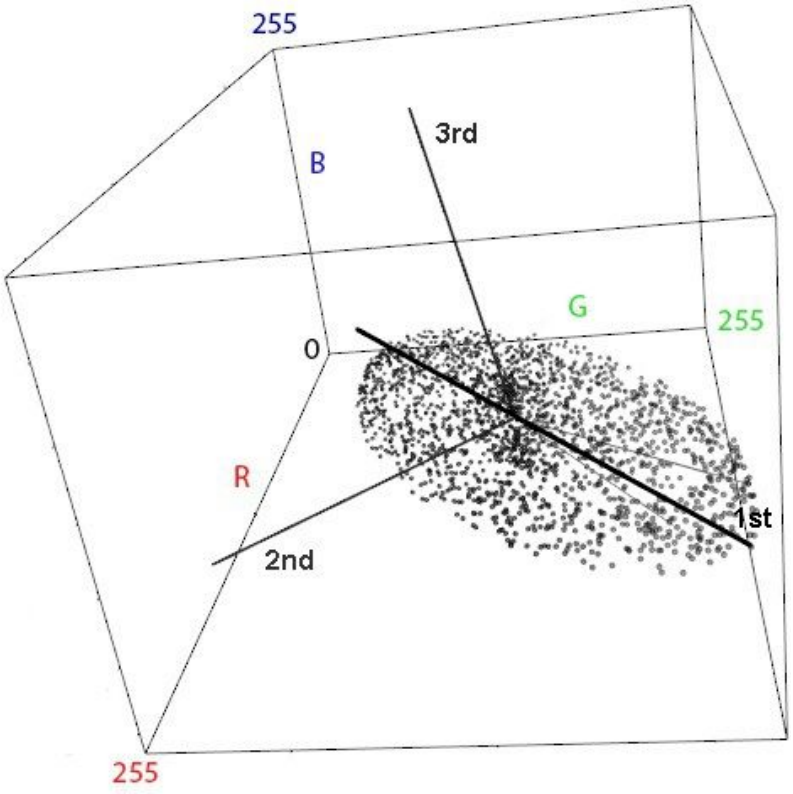


from K-means...

... to GMM


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
.01s
2.29s
.08s
1.37s
.22s
.14s
.01s
.04s
.02s
2.43s
.06s
2.74s
.23s
.25s
.01s
.04s
.02s
2.37s
.07s
.43s
.24s
.24s
.02s
.04s

Снижение размерности

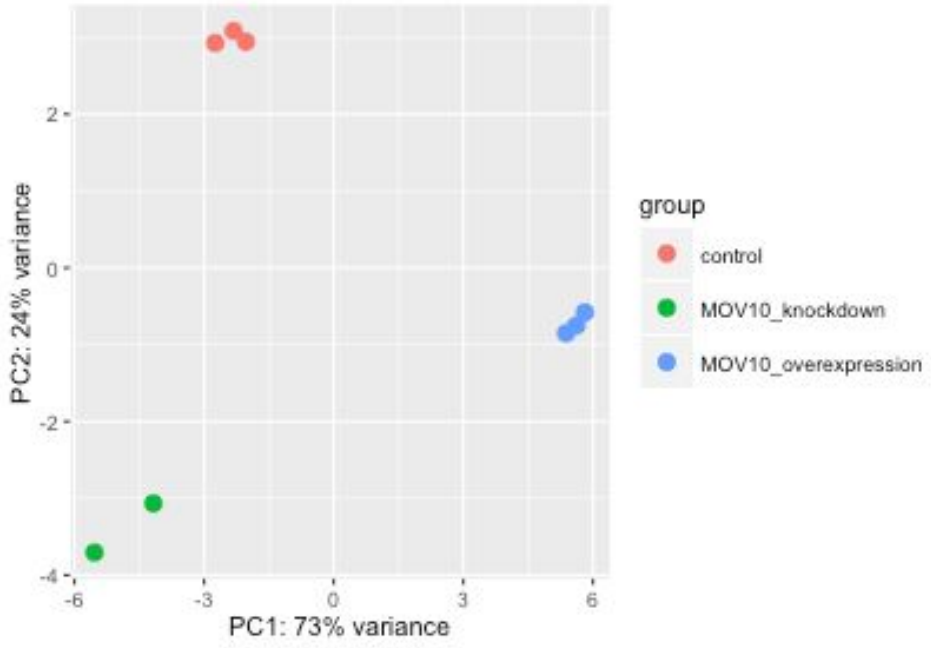


PCA - Principal Component Analysis

tSNE -
t-distributed Stochastic Neighbor Embedding

Зачем снижать размерность данных

- Исходные данные избыточны или сильно разряжены
- Оптимизация вычислительных затрат
- Подготовка данных для дальнейшего анализа
- Отбор признаков
- Визуализация данных



Применение уменьшения размерности

Поиск отфотошопленных изображений

Сжатие изображений

Распознавание лиц

Анализ и прогнозирование финансов (оптимизация портфеля)

Поиск аномалий

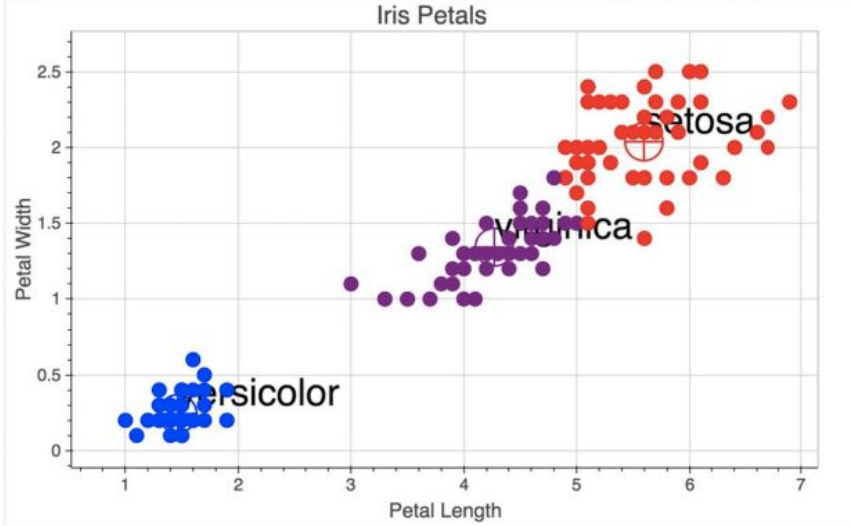
Анализ тематики текста

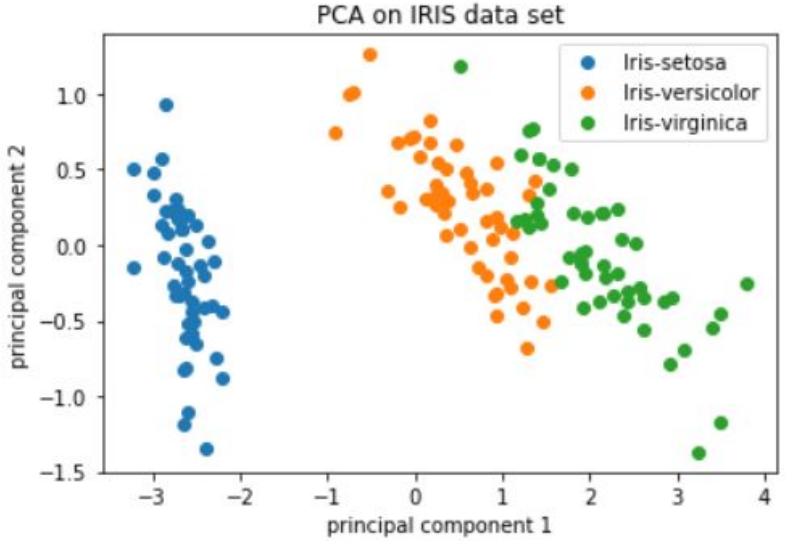
Поиск фотошопа

Разница между кластеризацией и уменьшением размерности

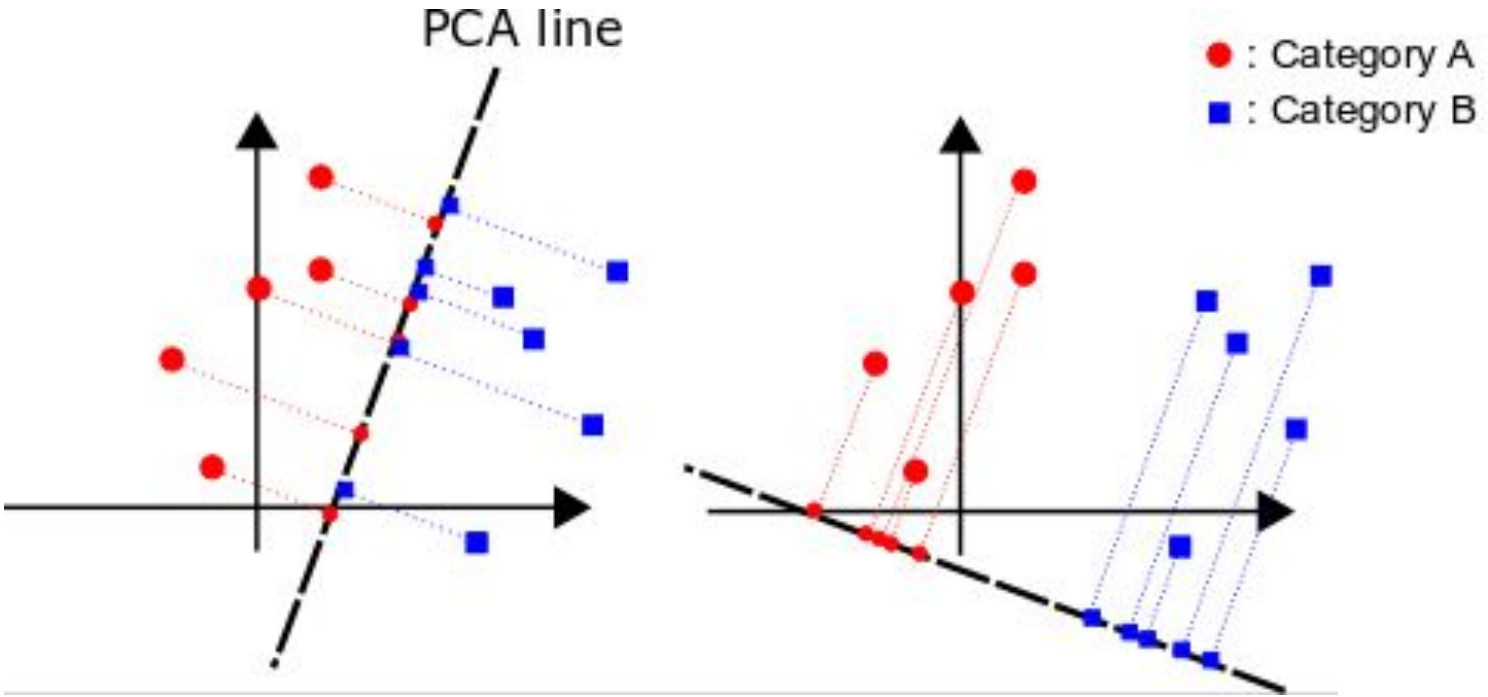
Кластеризация работает с расстояниями между точками в многомерном пространстве, как и tSNE

PCA (SVD, ISA) - с проекциями точек в многомерном пространстве





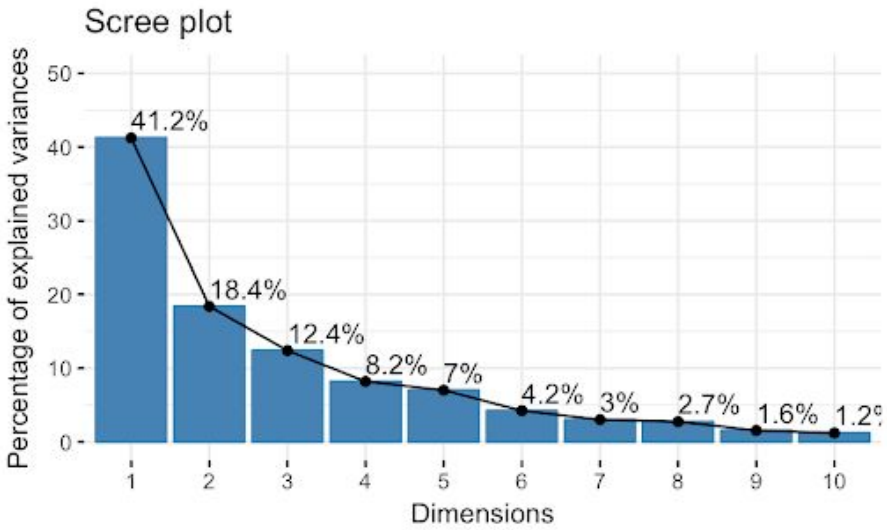
PCA

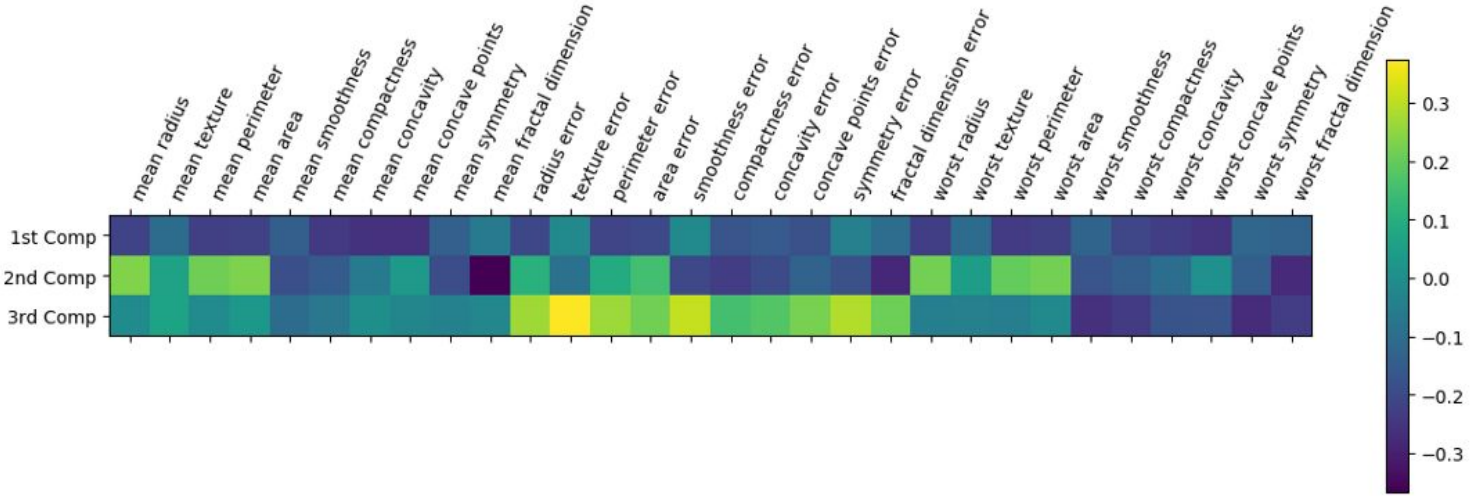


PCA компоненты

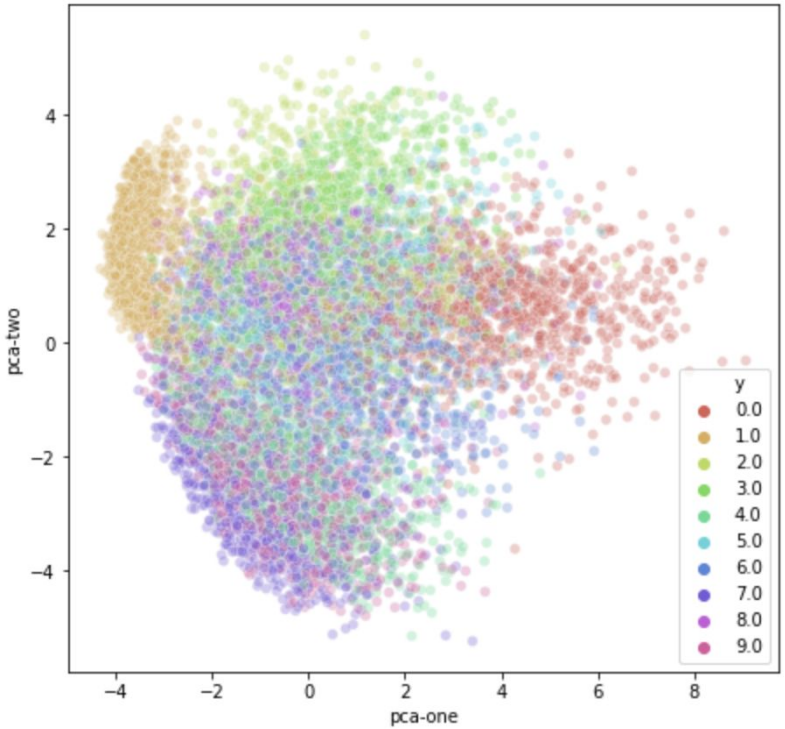
Можно узнать вклад каждой компоненты в объясненную дисперсию (explained_variance_ratio_)

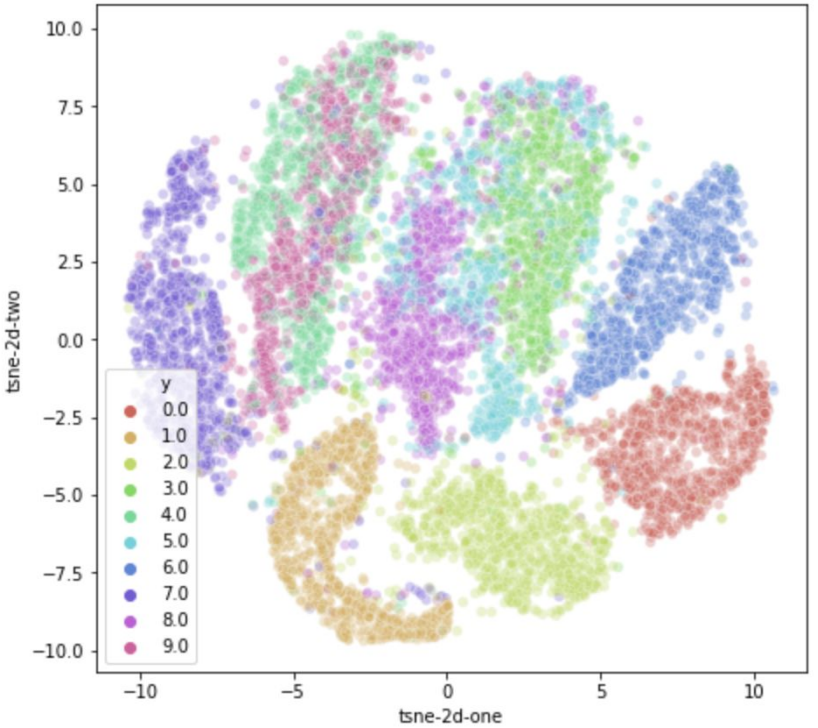
Из каких признаков “состоит” компонента (corr)





t-SNE метод для красивой визуализации

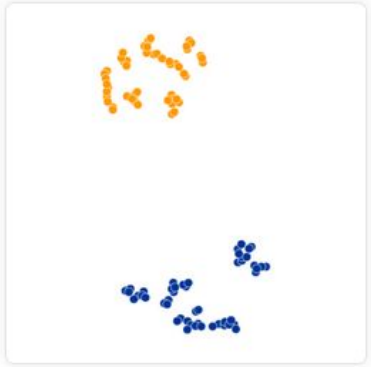




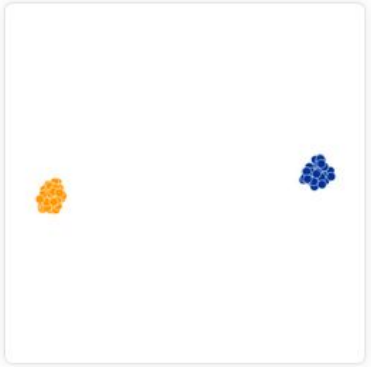
tSNE параметры: Perplexity

Отражает баланс между глобальными и локальными аспектами данных ~ сколько в среднем должно быть соседей у каждой точки

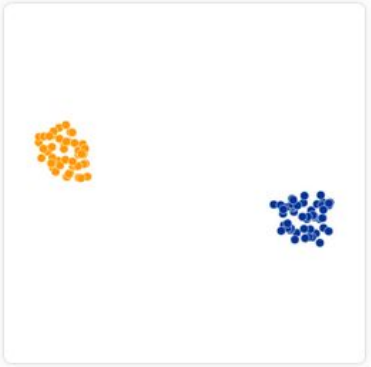
Perplexity: 2
Step: 5,000



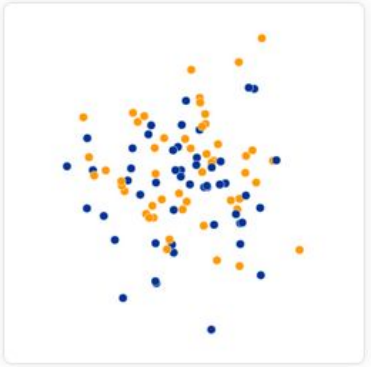
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000



Perplexity: 50
Step: 5,000

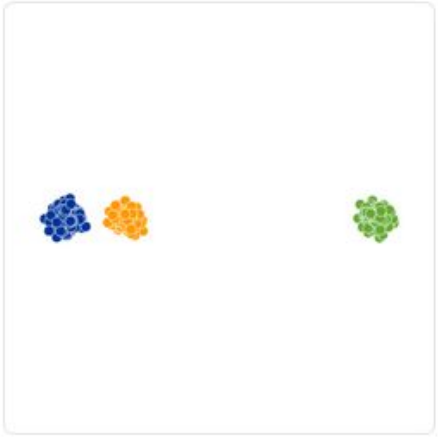


Perplexity: 100
Step: 5,000

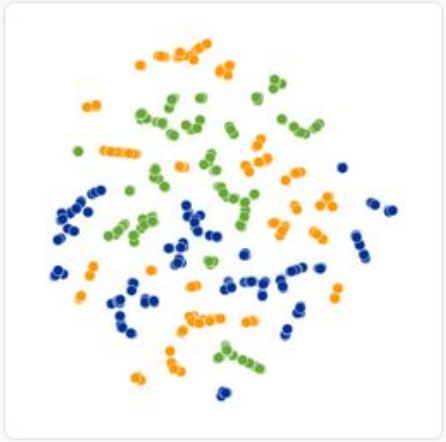
tSNE осторожно!

https://distill.pub/2016/misread-tsne/

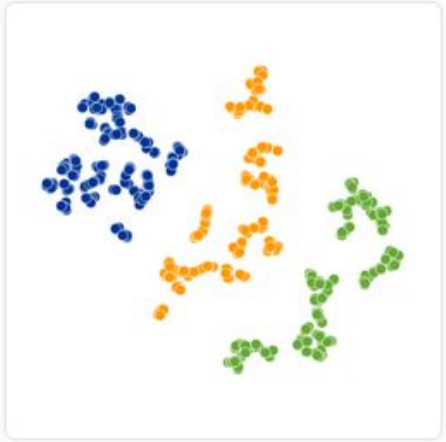
- Расстояние между кластерами не отражает различие между данными



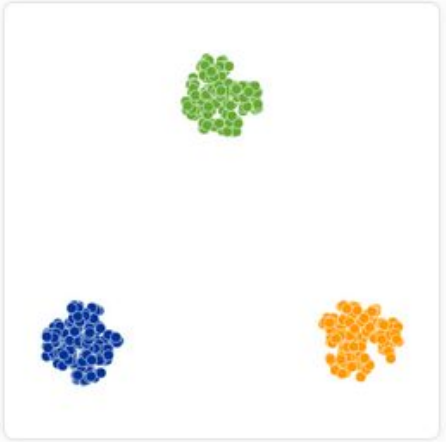
Original



Perplexity: 2
Step: 5,000

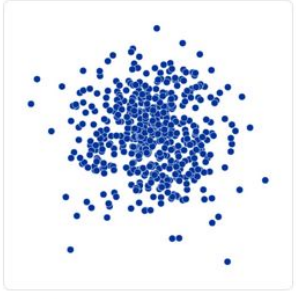


Perplexity: 5
Step: 5,000

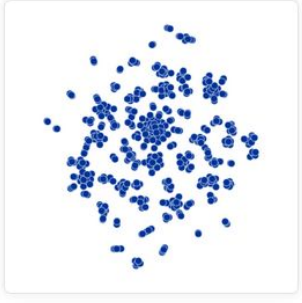


Perplexity: 30
Step: 5,000

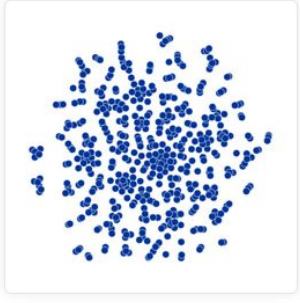
- Может показывать кластеры там где их нет



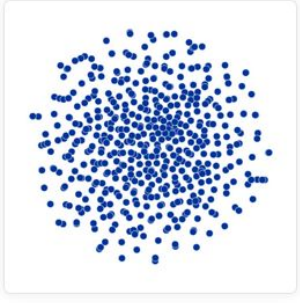
Original



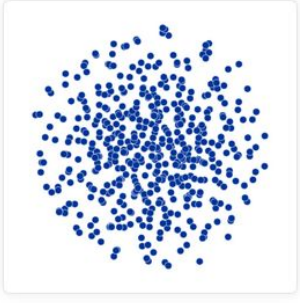
Perplexity: 2
Step: 5,000



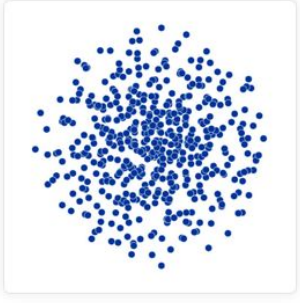
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

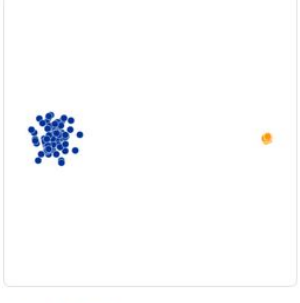


Perplexity: 50
Step: 5,000

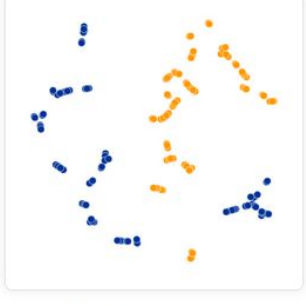


Perplexity: 100
Step: 5,000

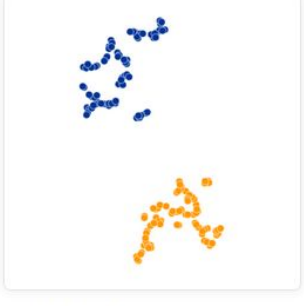
- Размер кластера не отражает расстояния внутри кластера



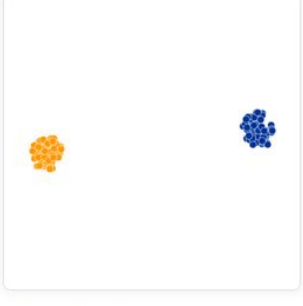
Original



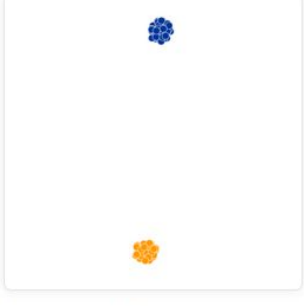
Perplexity: 2
Step: 5,000



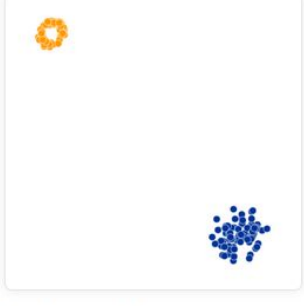
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

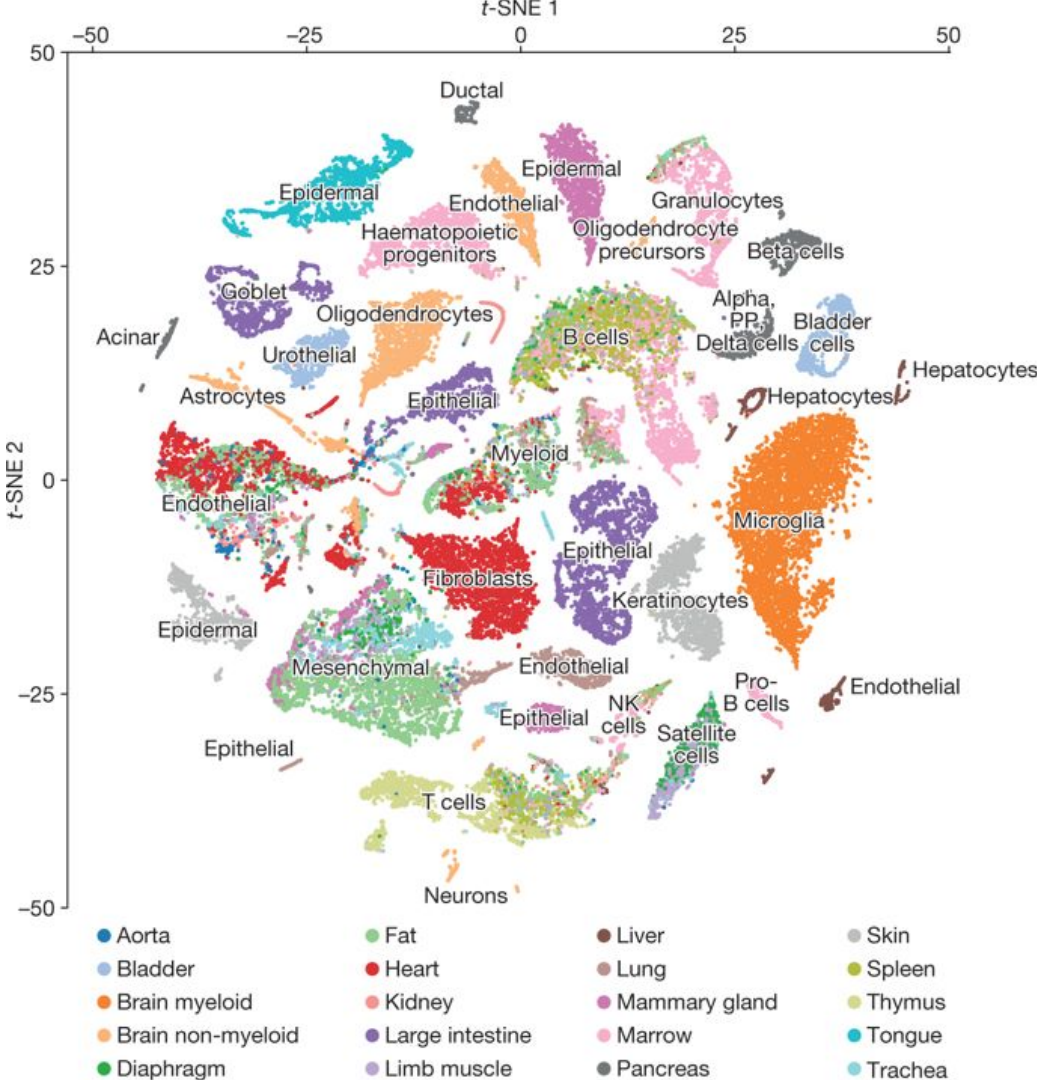


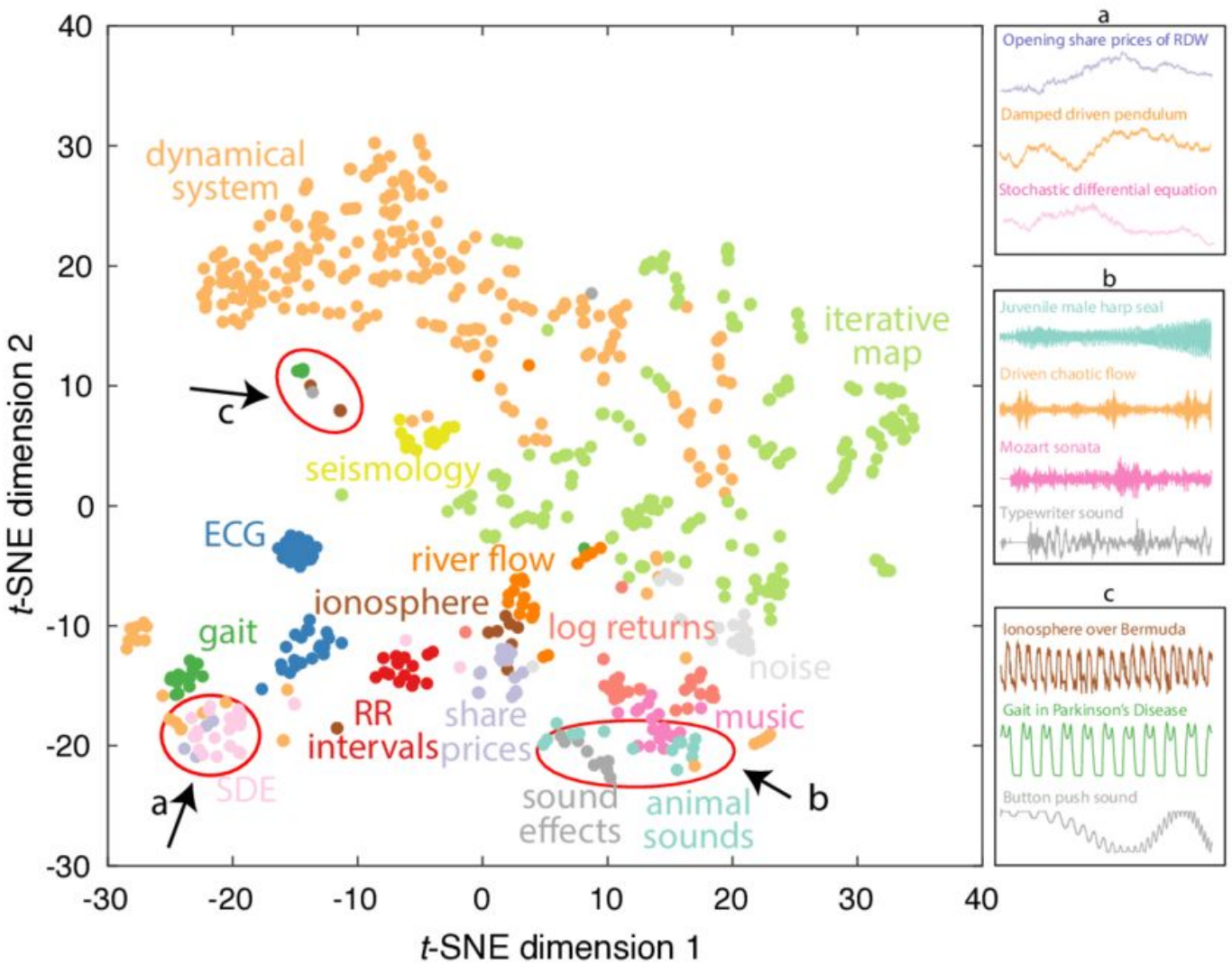
Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

t-SNE примеры из жизни




dynamical system
iterative map
seismology
ECG
river flow
ionosphere
gait
log returns
noise
RR intervals
share prices
music
SDE
sound effects
animal sounds
a
b
c
t-SNE dimension 1
t-SNE dimension 2
40
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
40
a
Opening share prices of RDW
Damped driven pendulum
Stochastic differential equation
b
Juvenile male harp seal
Driven chaotic flow
Mozart sonata
Typewriter sound
c
Ionosphere over Bermuda
Gait in Parkinson's Disease
Button push sound