

# Кластеризация

# Повторение

# Задача понижения размерности

- Дано: матрица «объекты-признаки»  $X$  размера  $\ell \times D$
- Найти: новую матрицу «объекты-признаки»  $Z$  размера  $\ell \times d$
- $d < D$

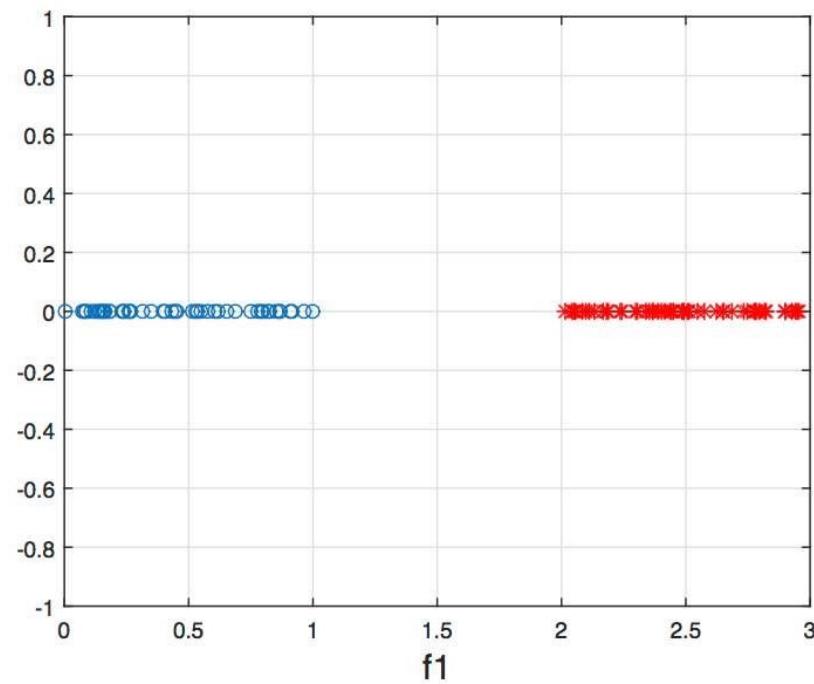
# Проклятие размерности

- Задача: классификация пончиков на вкусные и невкусные
- 100 объектов
- Цвет: 10 вариантов
- Цвет + размер:  $10 * 4 = 40$  вариантов
- Цвет + размер + форма:  $10 * 4 * 4 = 160$  вариантов



# Плохие признаки

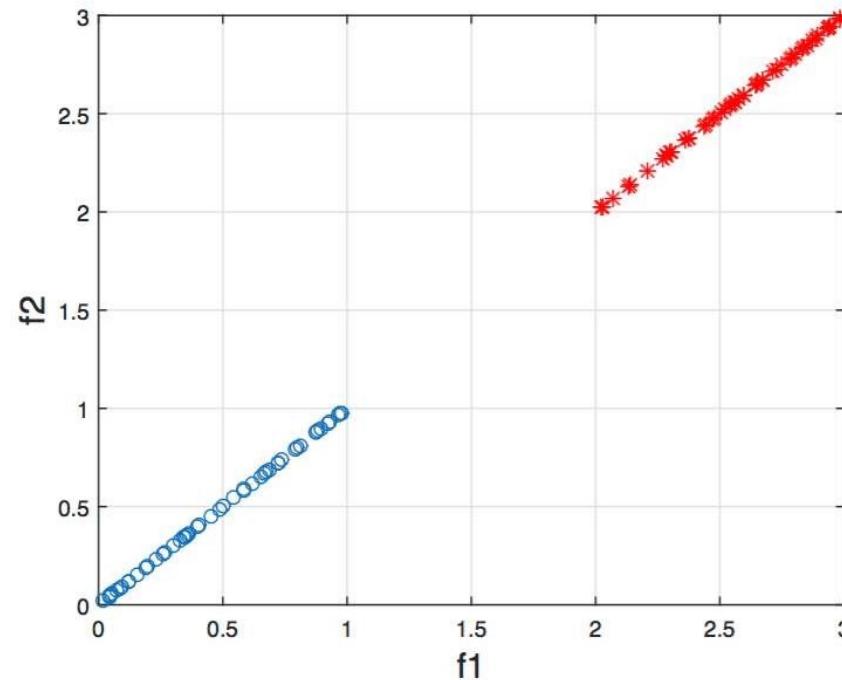
Информативный  
признак



# Плохие признаки

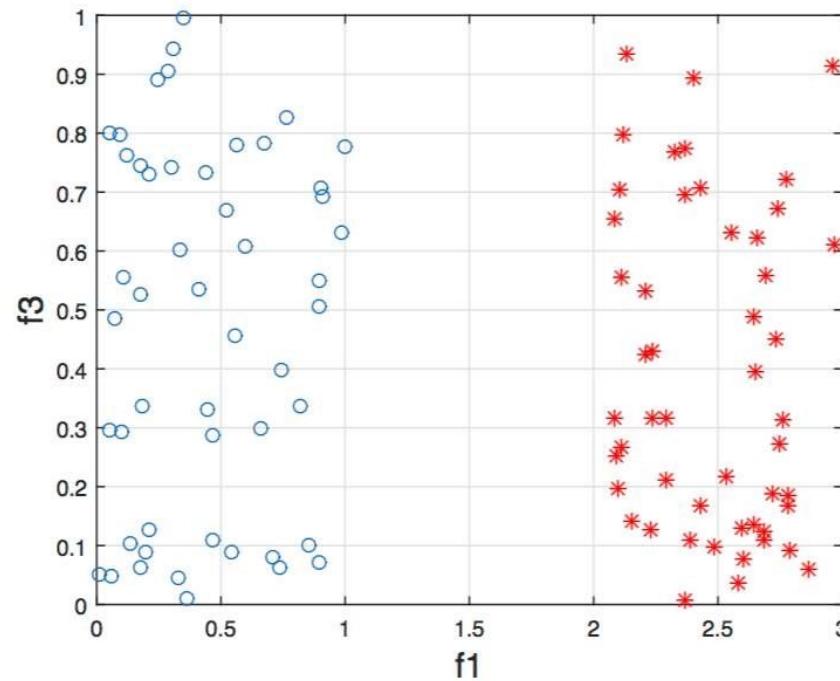
Коррелирующие  
признаки

$f_2$  — избыточный  
признак



# Плохие признаки

f3 — шумовой  
признак

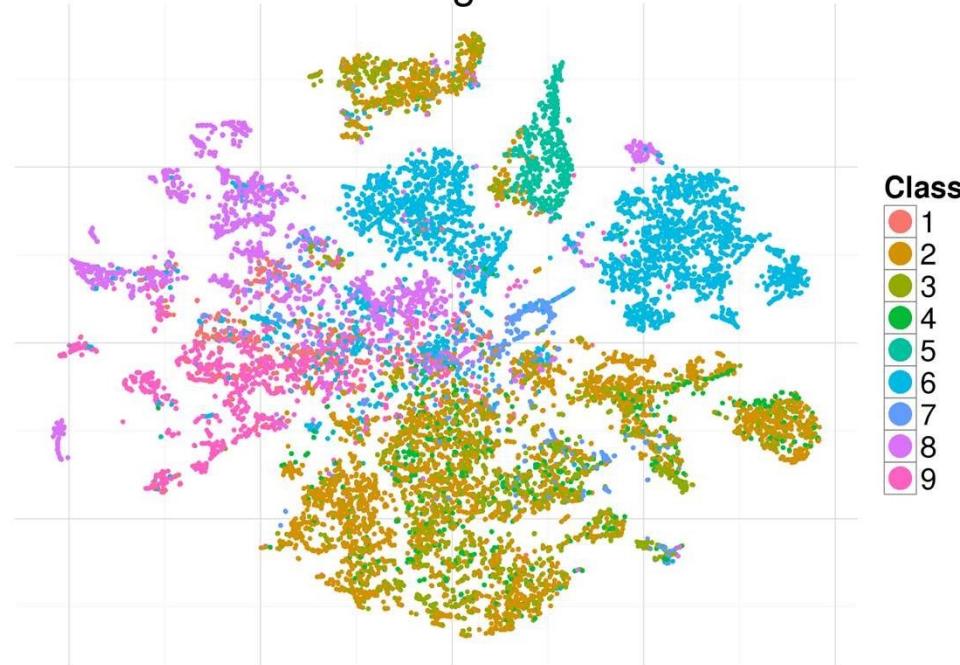


# Ускорение моделей

- Чем больше признаков, тем дольше обучаются модели
- Чем дольше обучаются модели, тем меньше экспериментов удаётся провести
- Чем сложнее модели, тем дольше они вычисляют прогнозы
- Могут быть жёсткие ограничения на скорость
- Пример: рекомендательные системы

# Визуализация

t-SNE 2D Embedding of Products Data



# Методы понижения размерности

- Отбор признаков (feature selection)
  - Выбрать  $d$  самых важных признаков
- Извлечение признаков (feature extraction)
  - Найти  $d$  новых признаков, выражающихся через исходные

# Методы понижения размерности

- Фильтрация (filter methods)
  - Понижение размерности без учёта модели
- Методы-обёртки (wrapper methods)
  - Выбор признаков, дающих лучшее качество для модели
- Понижение с помощью моделей (embedded methods)
  - Использование свойств моделей для оценивания важности признаков

# Одномерные методы

- Оценивают важность каждого признака по отдельности
- Относятся к **методам фильтрации**
- Относятся к **методам отбора признаков**

# Дисперсия признаков

$$R_j = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)^2$$

- Чем больше  $R_j$ , тем информативнее признак
- Никак не учитываются ответы
- Подходит для фильтрации константных и близких к ним признаков

# Корреляция

$$R_j = \frac{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{\ell} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{\ell} (y_i - \bar{y})^2}}$$

- Чем больше  $|R_j|$ , тем информативнее признак
- Учитывает только линейную связь

# T-score

$$R_j = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Для задач бинарной классификации
- Чем больше  $R_j$ , тем информативнее признак
- $\mu_1, \mu_2$  — средние значения признаков в первом и втором классах
- $\sigma_1^2, \sigma_2^2$  — дисперсии
- $n_1, n_2$  — число объектов в первом и втором классах

# F-score

$$R_j = \frac{\sum_{k=1}^K \frac{n_j}{K-1} (\mu_j - \mu)^2}{\frac{1}{\ell-K} \sum_{k=1}^K (n_j - 1) \sigma_j^2}$$

- Для задач многоклассовой классификации
- Чем больше  $R_j$ , тем информативнее признак
- $\mu_1, \dots, \mu_K$  — средние значения признаков в классах
- $\mu$  — среднее значение признака по всей выборке
- $\sigma_1^2, \dots, \sigma_K^2$  — дисперсии
- $n_1, \dots, n_K$  — число объектов в первом и втором классах

# Отбор с помощью моделей

- Оценивают важность признаков, используя модели машинного обучения
- Относятся к **методам отбора признаков**

# Линейные модели

$$a(x) = \sum_{j=1}^d w_j x_j$$

- Если признаки масштабированы, то веса можно использовать как показатели информативности
- Для повышения числа нулевых весов —  $L_1$ -регуляризация

# Регуляризация

$$Q(a, X) + \lambda \sum_{j=1}^d |w_j| \rightarrow \min_w$$

- Чем выше  $\lambda$ , тем больше весов зануляется
- Позволяет построить модель, использующую только самые важные признаки

# Решающие деревья

- Чем сильнее уменьшили  $H(X)$ , тем лучше признак
- Уменьшение критерия:

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

- Важность признака  $R_j$ : просуммируем уменьшения по всем вершинам, где разбиение делалось по признаку  $j$

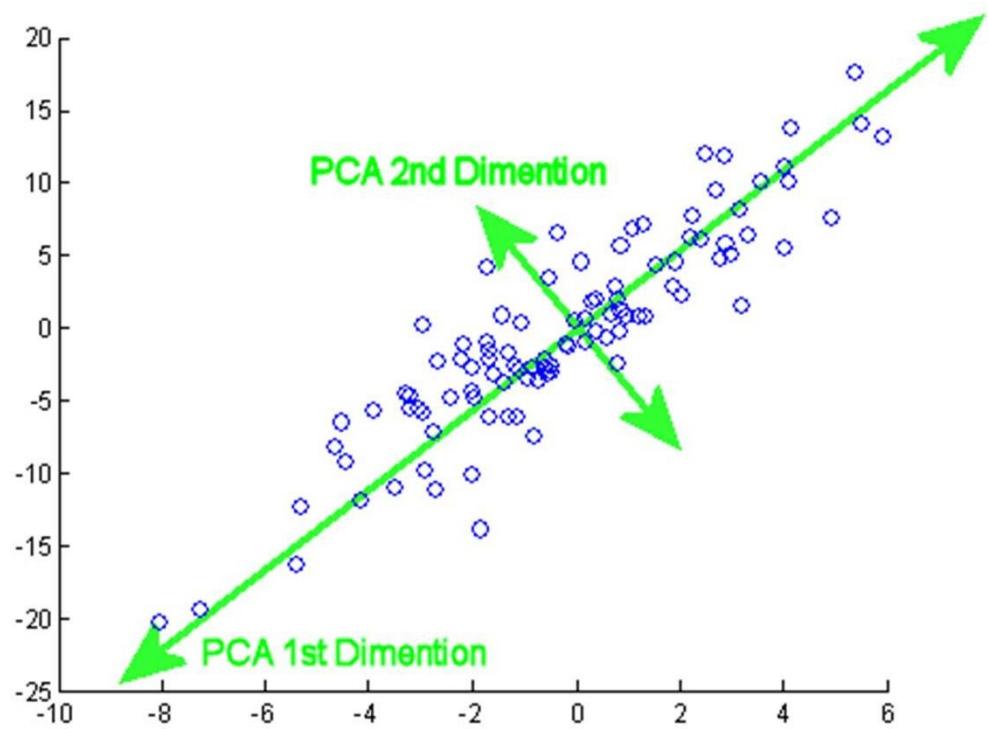
# Случайный лес

- Сумма важностей  $R_j$  по всем деревьям
- Чем больше, тем важнее признак
- Учитывается важность признаков в совокупности

# Метод главных компонент

- Principal component analysis (PCA)
- Проецирует данные в пространство меньшей размерности
- Относится к **методам фильтрации**
- Относится к **методам извлечения признаков**

# Извлечение признаков



# Извлечение признаков

- Исходные признаки:  $x_{ik}, D$  штук
- Новые признаки:  $z_{ij}, d$  штук
- Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

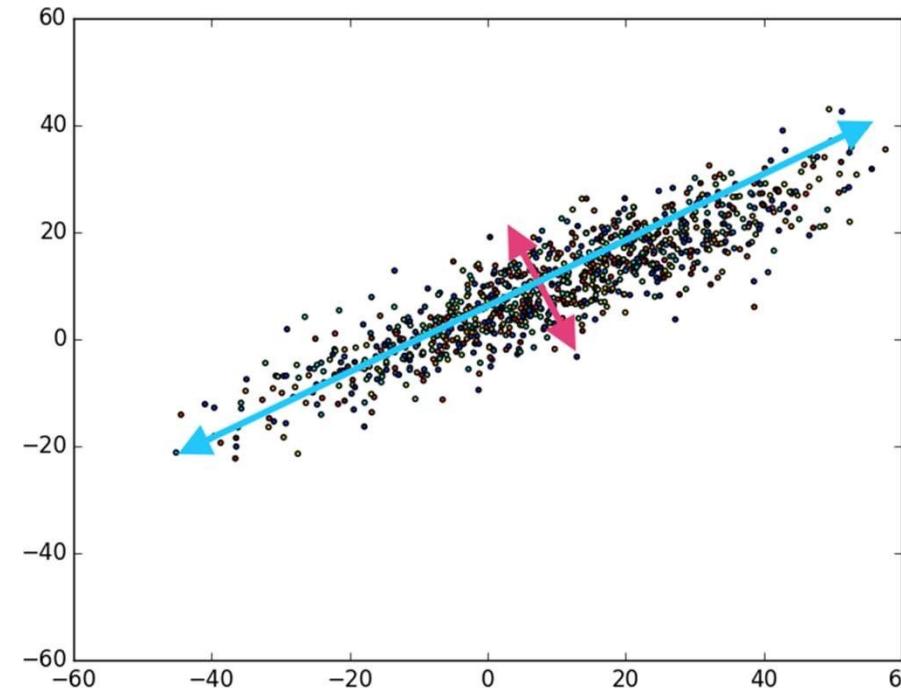
Новые признаки

Вклад исходного  $k$ -го  
признака в новый  $j$ -й

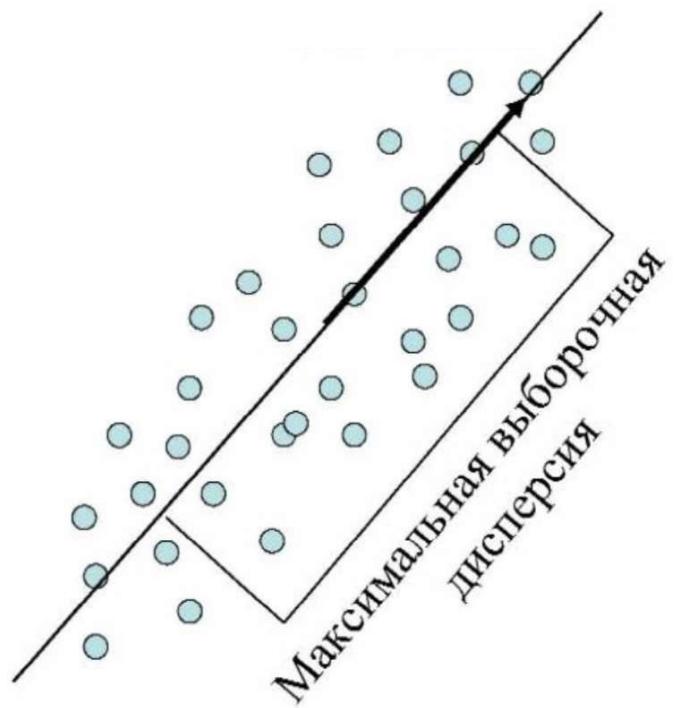
Исходные признаки

# Метод главных компонент

- Геометрический смысл — поиск гиперплоскости для проецирования выборки
- Как выбирать гиперплоскость?
- Чем выше дисперсия выборки после проецирования, тем лучше
- Дисперсия — мера количества информации



# Метод главных компонент



# Максимизация дисперсии

$$\left\{ \sum_{j=1}^d w_j^T X^T X w_j \rightarrow \max_W \right.$$

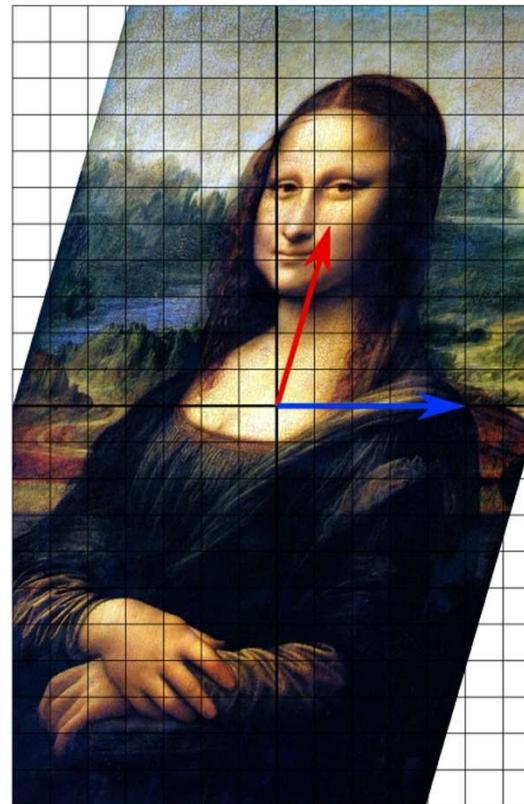
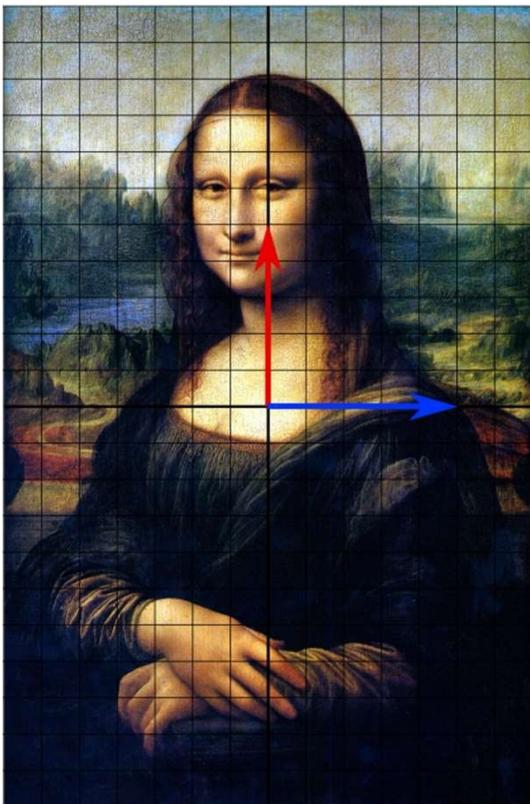
$$W^T W = I$$

Дисперсия выборки

# Собственные векторы

- $A$  — матрица размера  $n \times n$
- Пусть  $Ax = \lambda x$
- Тогда  $x$  — собственный вектор,  $\lambda$  — собственное значение
- $x$  — вектор, который не меняет направление под действием матрицы

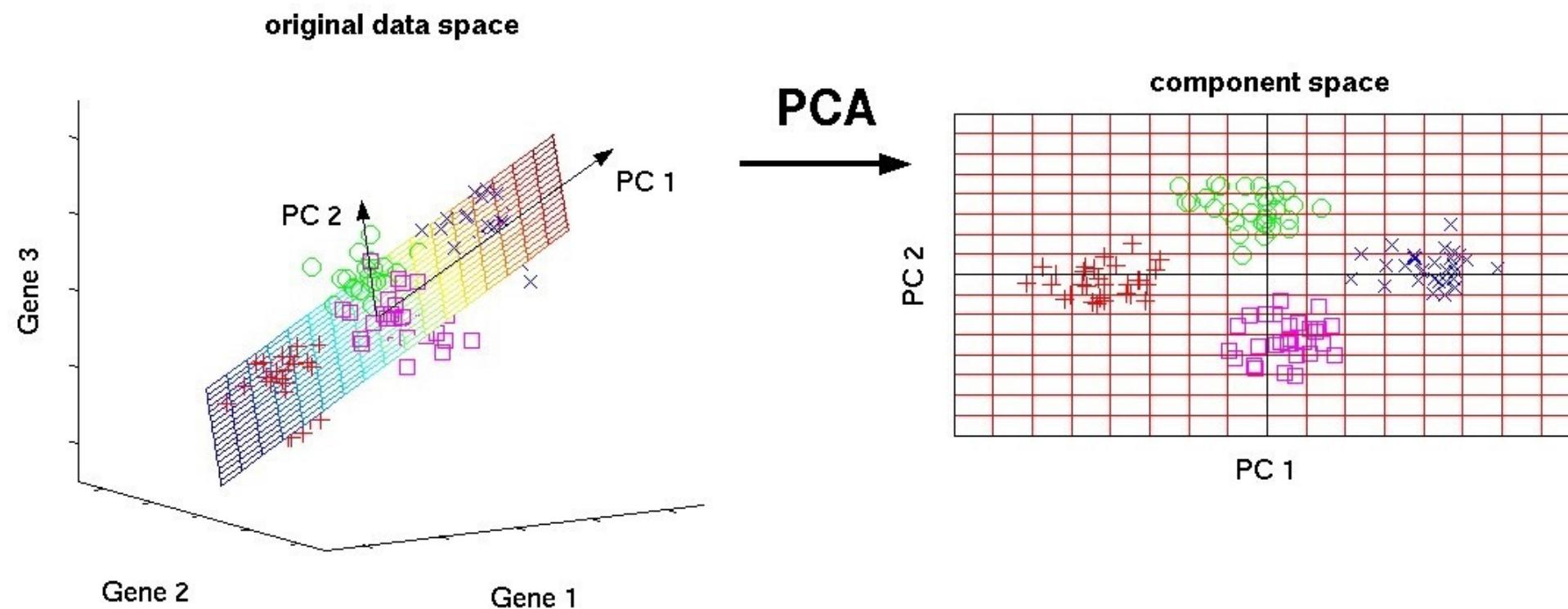
# Собственные векторы



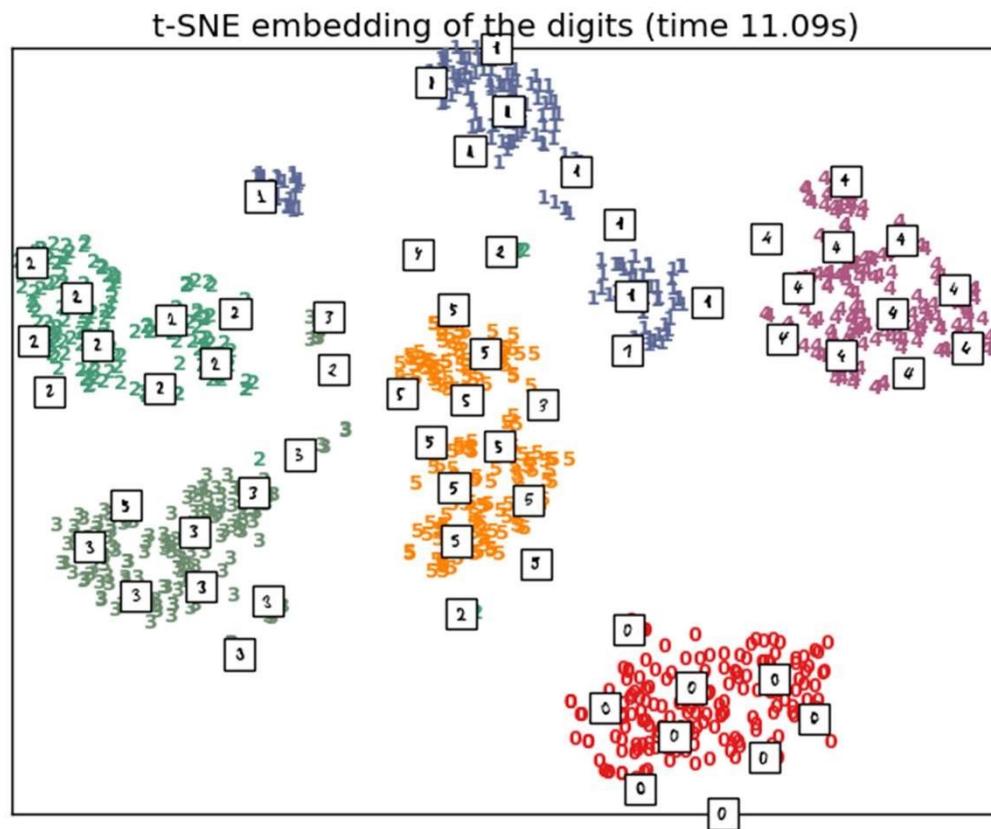
# Решение

- Столбцы  $W$  — собственные векторы матрицы  $X^T X$ , соответствующие наибольшим собственным значениям  $\lambda_1, \lambda_2, \dots, \lambda_d$
- $\frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i}$  — доля дисперсии, сохранённой при понижении размерности

# Метод главных компонент



# MNIST



# Кластеризация

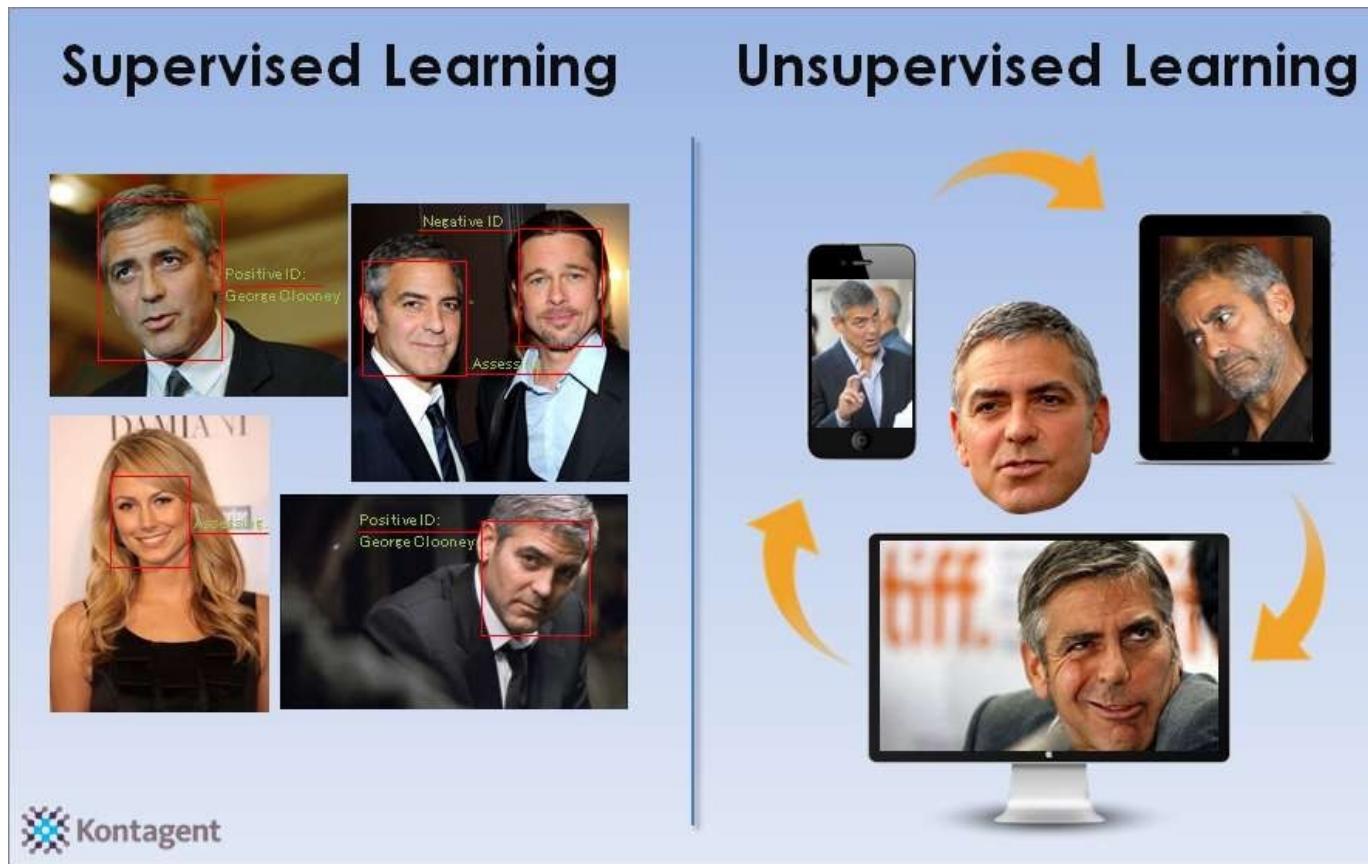
# Обучение с учителем (supervised learning)

- Для каждого объекта известен ответ (класс или число)
- Даны примеры объектов с ответами
- Нужно построить модель, которая будет предсказывать ответы для новых объектов

# Обучение без учителя (unsupervised learning)

- Даны объекты
- Нужно найти в них внутреннюю структуру
- Примеры:
  - Кластеризация
  - Обнаружение аномалий
  - Тематическое моделирование
  - Визуализация
  - Предсказание следующего кадра видео
  - ...
- Ближе к обучению в реальной жизни

# Обучение с учителем и без учителя



# Кластеризация

- Дано: матрица «объекты-признаки»  $X$
- Найти:
  1. Множество кластеров  $Y$
  2. Алгоритм кластеризации  $a(x)$ , который приписывает каждый объект к одному из кластеров
- Каждый кластер состоит из похожих объектов
- Объекты из разных кластеров существенно отличаются

# Отличия

## Обучение с учителем

- Цель: минимизация функционала ошибки
- Множество ответов известно заранее
- Конкретные способы измерения качества

## Кластеризация

- Нет строгой постановки
- Множество кластеров неизвестно
- Правильные ответы отсутствуют (в большинстве случаев) — нельзя измерить качество

# Зачем кластеризовать?

- Маркетинг: искать похожих клиентов
  - Модерация: проверять только одно сообщение из кластера
  - Соц. опросы: выделять группы схожих анкет
  - Соц. сети: искать сообщества
- 
- Выявлять типы людей и формировать поведенческие паттерны для каждого типа

# Виды кластеризации

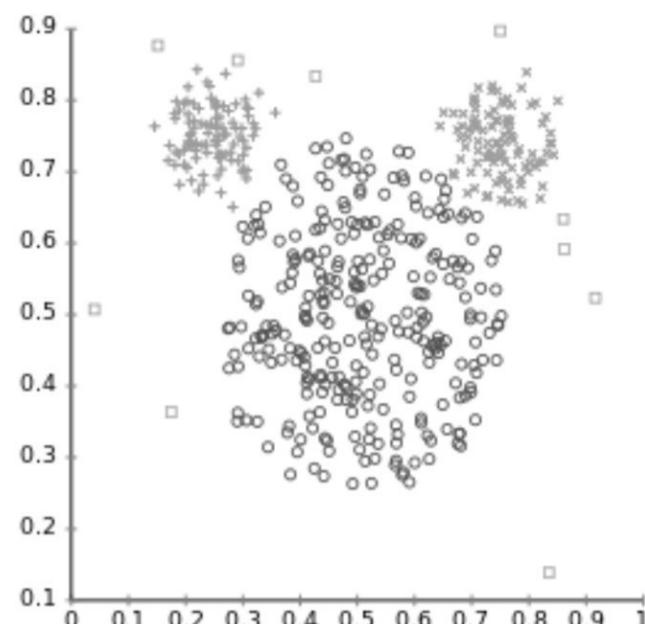
# Форма кластеров



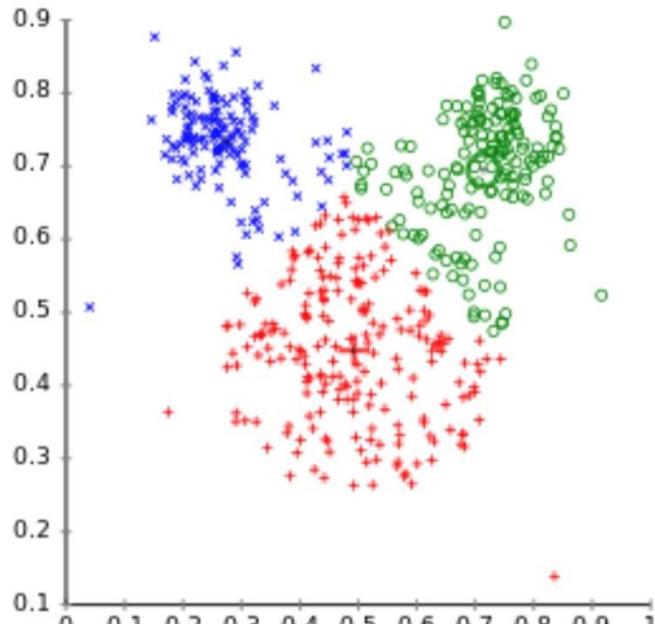
# Форма кластеров



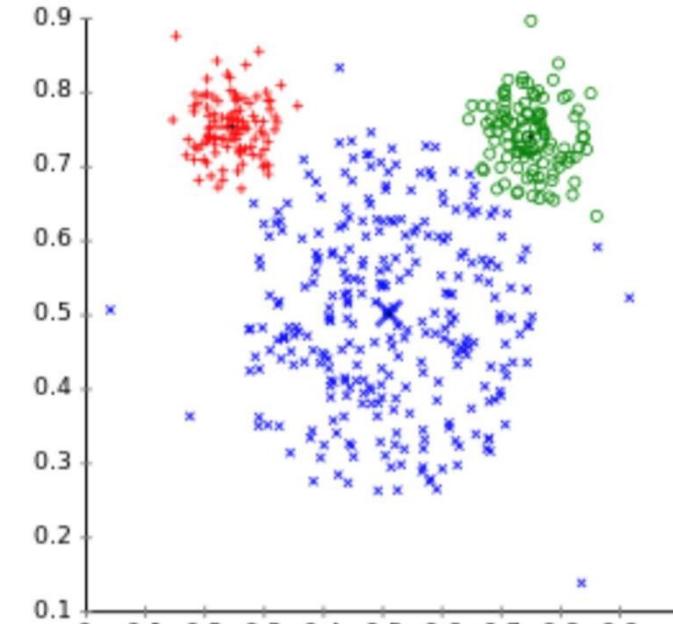
# Различия в результатах работы



Исходная выборка  
("Mouse" dataset)

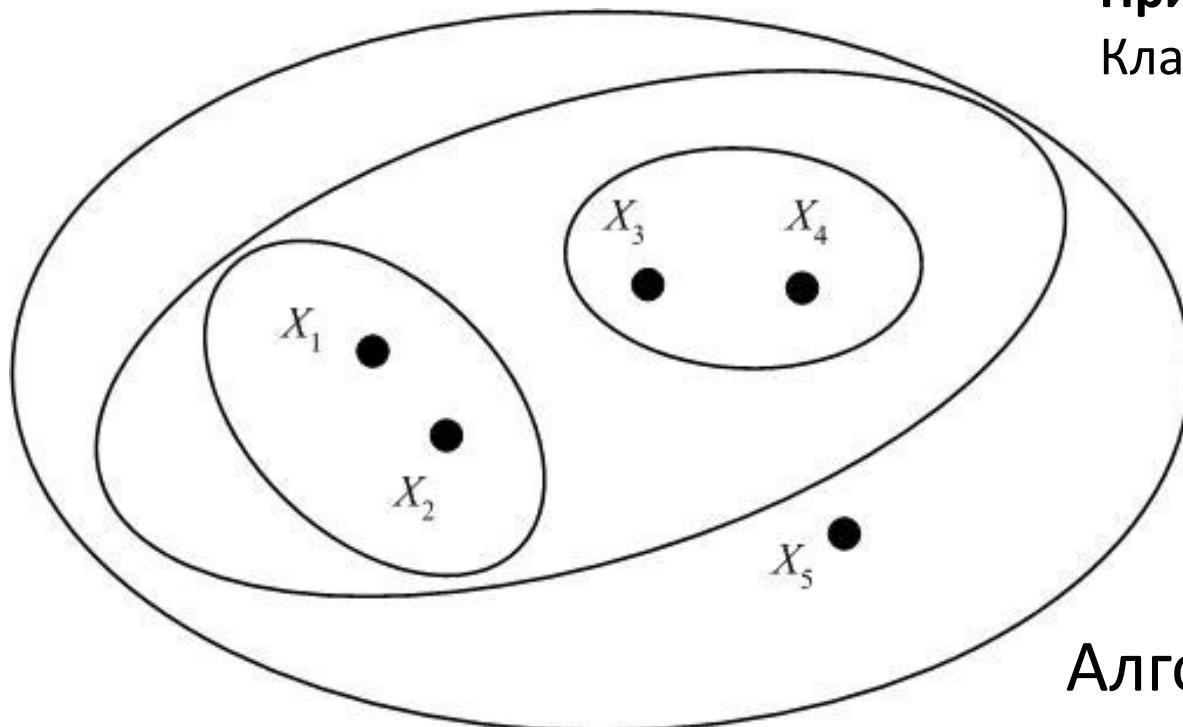


Метод 1



Метод 2

# Иерархическая кластеризация



Пример:

Кластеризация статей на Хабре

IT

Алгоритмы

Алгоритмы  
и структуры  
данных

Методы  
машинного  
обучения

# Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 2: в один кластер должны попадать новости об одном «большом» событии



Керлингистки сборной РФ сделали  
правильные выводы после ОИ -  
Сидорова

10:38 26.03.2014



Путин призвал МВД использовать в  
Крыму опыт работы на Олимпиаде

14:13 21.03.2014



Два "олимпийских" спецавтопарка  
останутся в Сочи как наследие Игр

11:50 26.03.2014

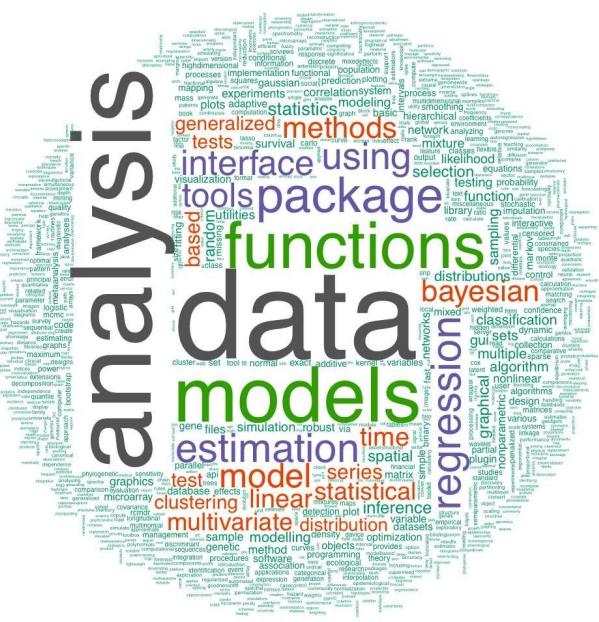
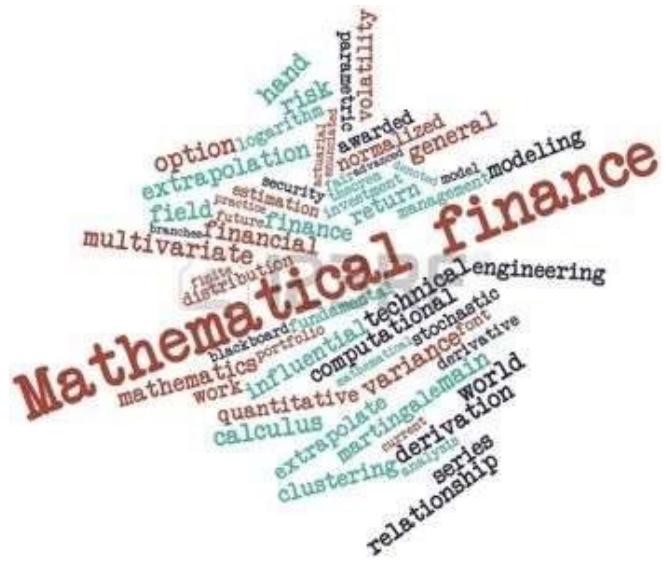
Скриншот с сайта РИА Новости (ria.ru)

# Требования к кластерам

- Чтобы проверить, выполняются ли требования, нужно делать разметку данных
- Для новостей: показывать асессору пары документов и спрашивать, относятся ли они к одному кластеру

# «Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



# Типы задач кластеризации

- Форма кластеров, которые нужно выделять
- Плоская или древовидная структура
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

# K-Means

# K-Means

- Дано: выборка  $x_1, \dots, x_\ell$
- Параметр: число кластеров  $K$
- Начало: случайно выбрать  $K$  центров кластеров  $c_1, \dots, c_K$
- Повторять по очереди до сходимости:

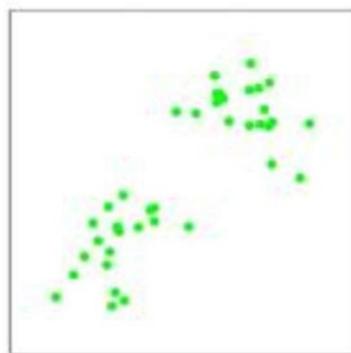
- Шаг А: отнести каждый объект к ближайшему центру

$$y_i = \arg \min_{j=1, \dots, K} \rho(x_i, c_j)$$

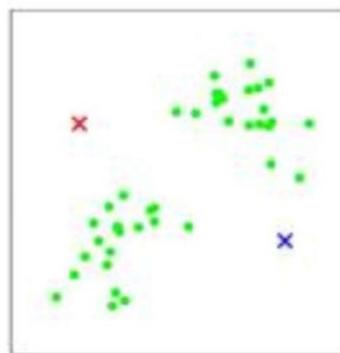
- Шаг Б: переместить центр каждого кластера в центр тяжести

$$c_j = \frac{\sum_{i=1}^{\ell} x_i [y_i = j]}{\sum_{i=1}^{\ell} [y_i = j]}$$

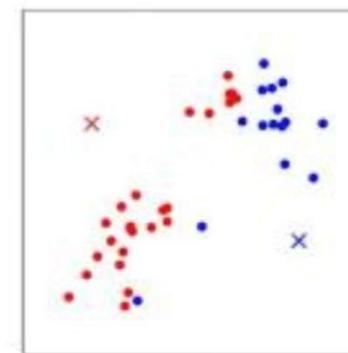
# K-Means



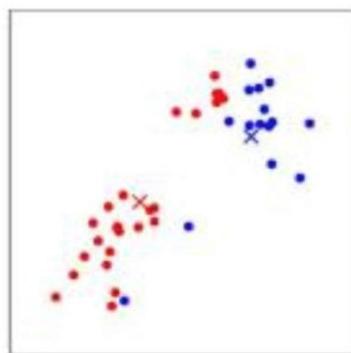
(a)



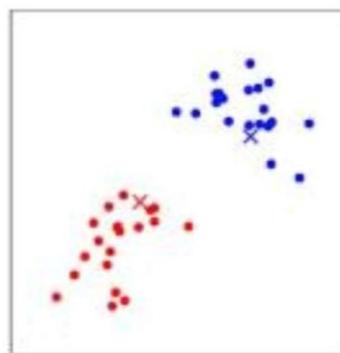
(b)



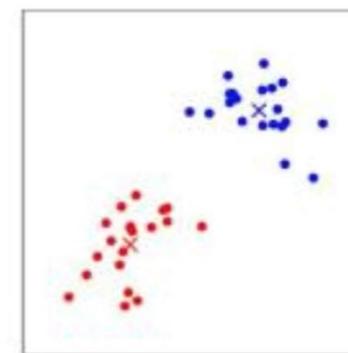
(c)



(d)

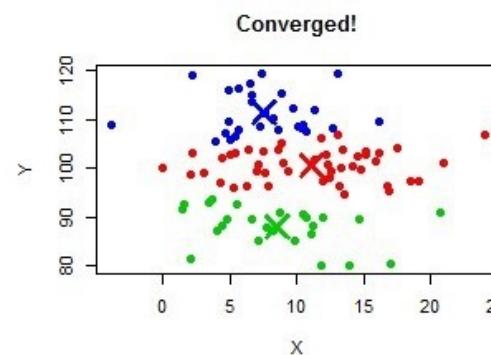
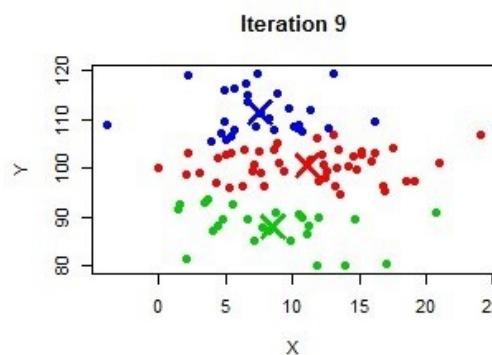
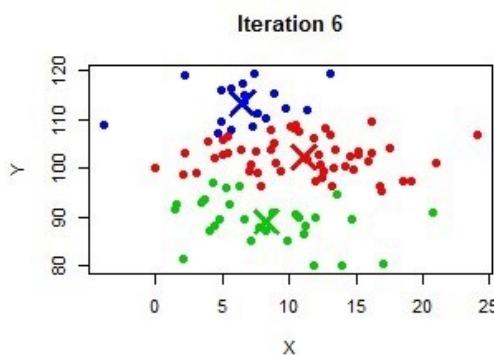
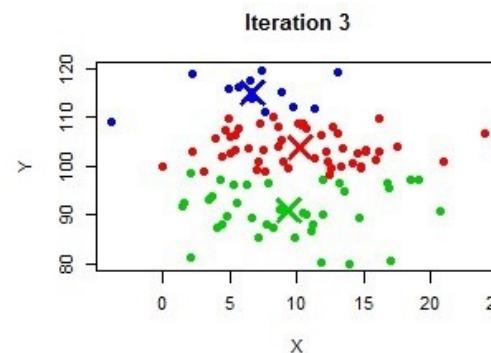
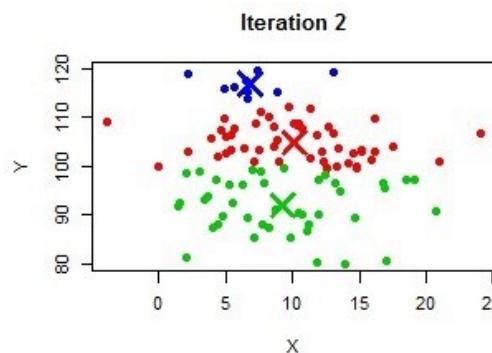
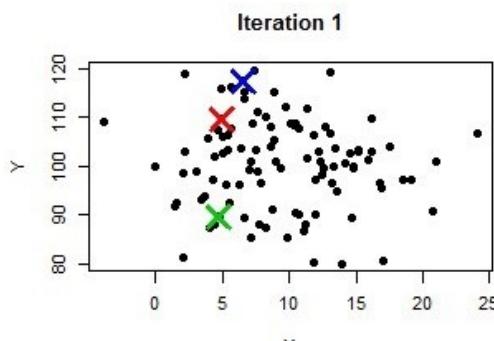


(e)



(f)

# K-Means



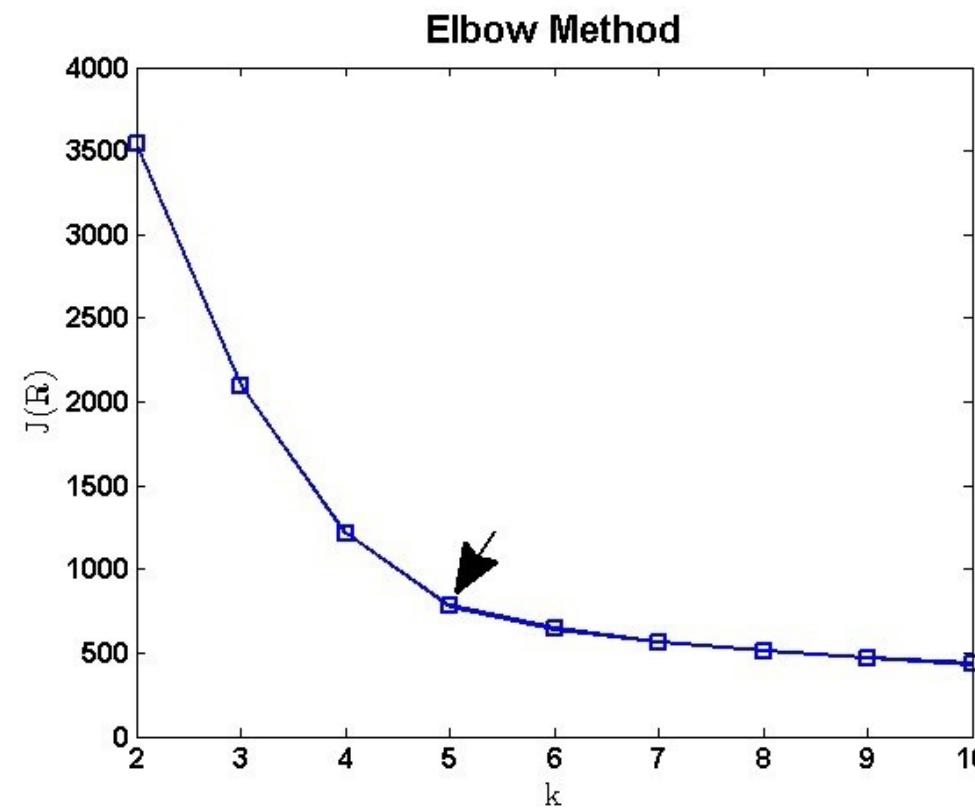
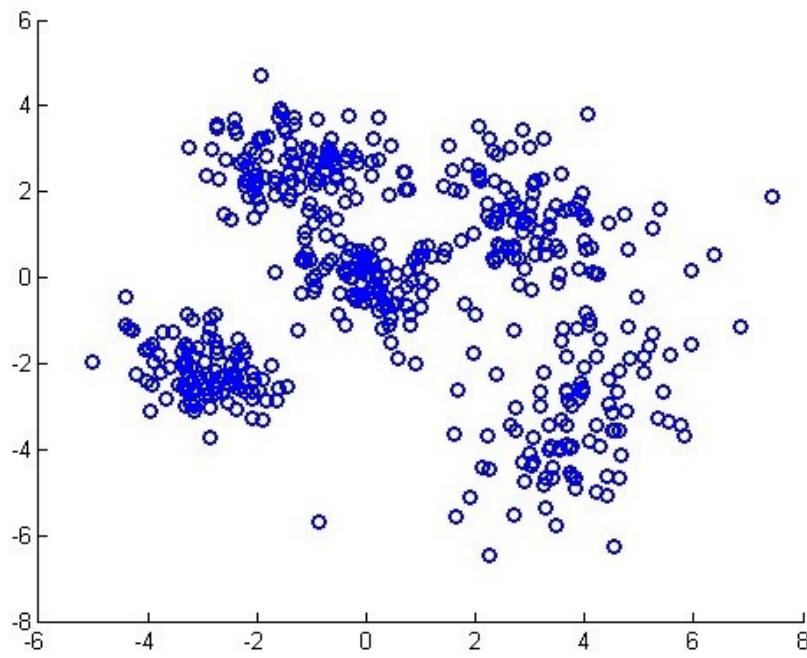
# Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

$$J(C) = \sum_{i=1}^{\ell} \rho(x_i, c_{y_i})$$

- Зависит от  $K$
- Нужно подобрать такое  $K$ , после которого качество меняется не слишком сильно

# Выбор числа кластеров

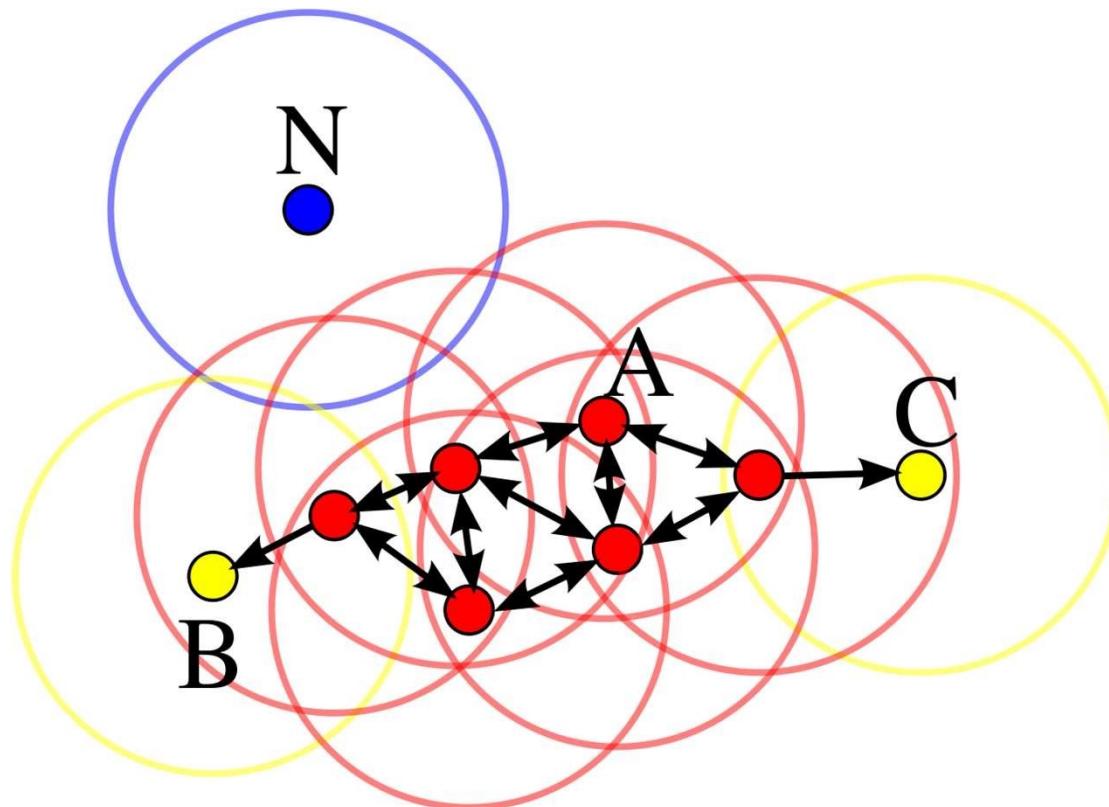


# Особенности K-Means

- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требует выбора числа кластеров

# Density-based clustering

# Основные, граничные и шумовые точки



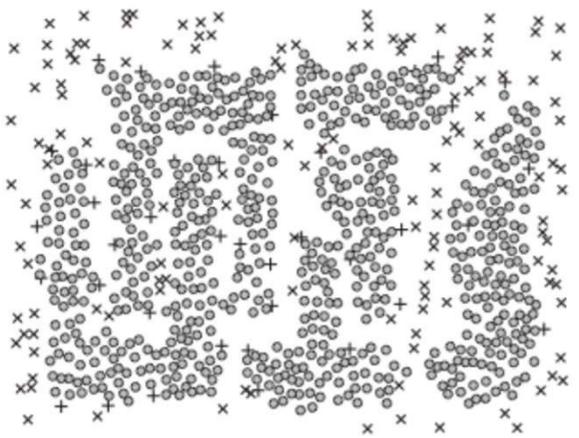
# Параметры DBSCAN

- Размер окрестности (eps)
- Минимальное число объектов в окрестности — для определения основных точек

# DBSCAN



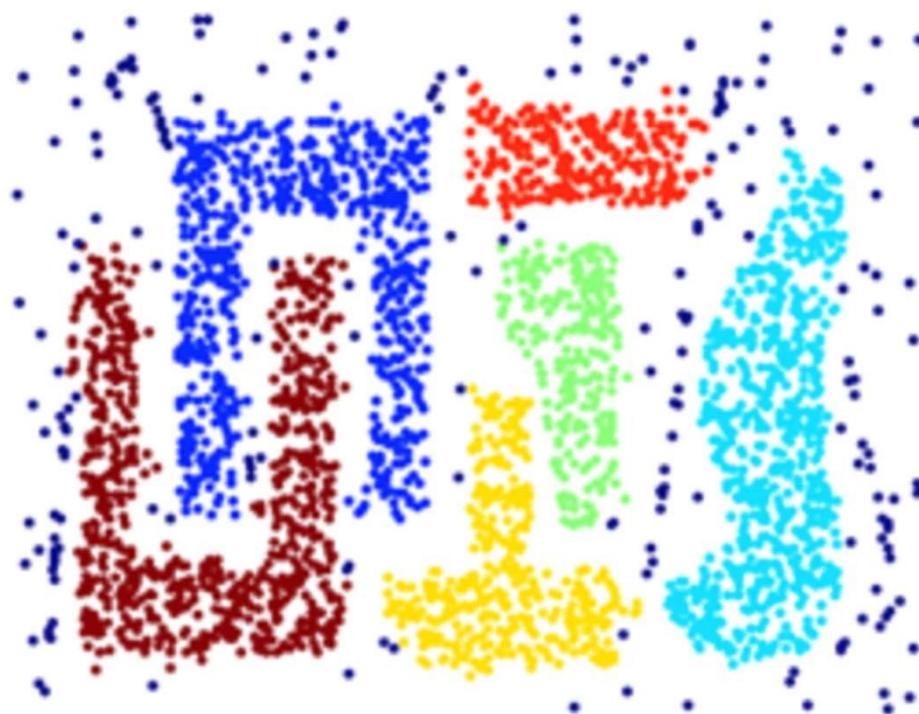
(a) Clusters found by DBSCAN.



(b) Core, border, and noise points.

1. Выбрать точку без метки
2. Если в окрестности меньше  $N$  точек, то пометить как шумовую
3. Создать новый кластер, поместить в него текущую точку
4. Для всех точек из окрестности  $S$ : (а) если точка шумовая, то отнести к данному кластеру, но не использовать для расширения; (б) если точка основная, то отнести к данному кластеру, а её окрестность добавить к  $S$
5. Перейти к шагу 1

# DBSCAN: результаты работы



# Особенности DBSCAN

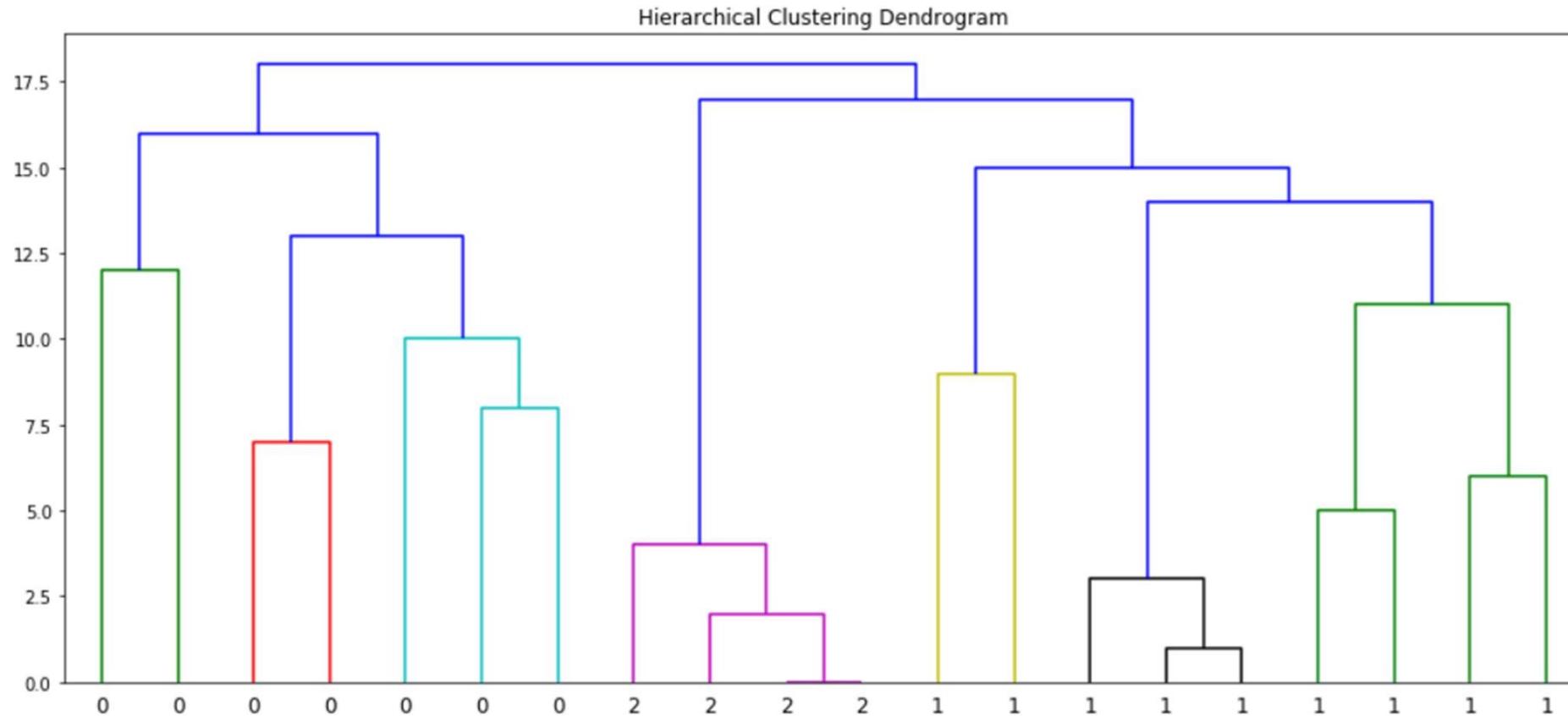
- Находит кластеры произвольной формы
- Может работать с большими объёмами данных
- Нужно подбирать размер окрестности ( $\text{eps}$ ) и минимальное число объектов в окрестности

# Иерархическая кластеризация

# Виды иерархической кластеризации

- Аггломеративная – на каждой итерации объединяем два меньших кластера в один побольше
- Дивизивная – на каждой итерации делим один большой кластер на два поменьше

# Виды иерархической кластеризации



# Агglomerативная кластеризация

1. Инициализация – каждая точка = кластер
2. Самые близкие (относительно какой-то метрики) кластеры объединяются
3. Повторяем до того момента, когда все точки будут в одном кластере
4. Останавливаемся, когда достигаем фиксированного числа кластеров, либо когда расстояние между кластерами больше заданного порога

# Обучение без учителя и текстовые данные

# Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?
- На основе данных!
- Слова со схожим смыслом часто идут в паре с одними и теми же словами
- У них похожие контексты

# Дистрибутивная семантика

- У похожих по смыслу слов похожие контексты
- Контекст — окрестность слова

...an efficient method for learning high quality distributed vector ...

The diagram shows the phrase "...an efficient method for learning high quality distributed vector ..." in green. A blue bracket labeled "Context" spans the first two words, "an efficient". Another blue bracket labeled "Context" spans the last three words, "high quality distributed vector". In the center, the word "learning" is written vertically above a blue arrow pointing upwards, with the label "focus word" written below it.

# Векторные представления слов

Хотим представить каждое слово в виде вещественного вектора:

$$w \rightarrow \vec{w} \in \mathbb{R}^d$$

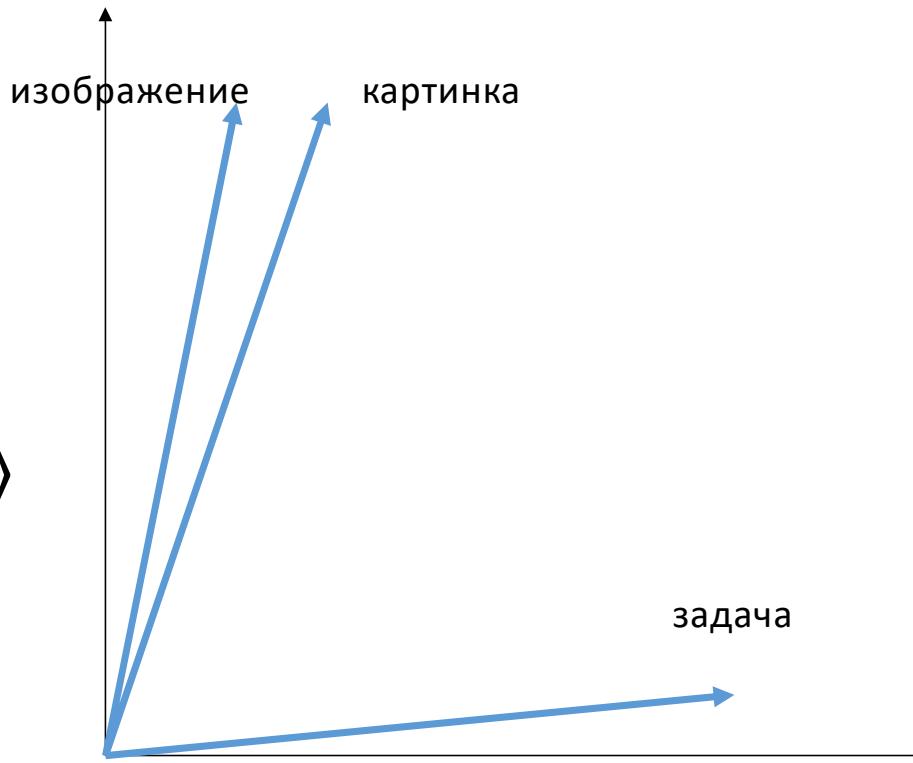
Требования к представлениям (embeddings):

- Размерность  $d$  должна быть не очень большой
- Похожие слова должны иметь близкие векторы
- Арифметические операции над векторами должны иметь смысл

# word2vec

Задача:

- Для каждого слова  $w$  построить вектор  $\vec{w}$
- Если два слова  $w_1$  и  $w_2$  идут рядом, то скалярное произведение  $\langle \vec{w}_1, \vec{w}_2 \rangle$  должно быть большим



## word2vec

Если два слова  $w_1$  и  $w_2$  идут рядом, то скалярное произведение  $\langle \vec{w}_1, \vec{w}_2 \rangle$  должно быть большим:

$$p(w_i | w_j) = \frac{\exp(\langle \vec{w}_i, \vec{w}_j \rangle)}{\sum_{w \in W} \exp(\langle \vec{w}, \vec{w}_j \rangle)}$$

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \sum_{\substack{k=-K \\ k \neq 0}}^K \log p(\vec{w}_{j+k} | \vec{w}_j) \rightarrow \max_{\{\vec{w}\}_{w \in W}}$$

# word2vec

Векторы можно прибавлять и вычитать:

- $\overrightarrow{\text{король}} - \overrightarrow{\text{мужчина}} + \overrightarrow{\text{женщина}} \approx \overrightarrow{\text{королева}}$
- $\overrightarrow{\text{медведь}} - \overrightarrow{\text{Россия}} + \overrightarrow{\text{Австралия}} \approx \overrightarrow{\text{кенгуру}}$

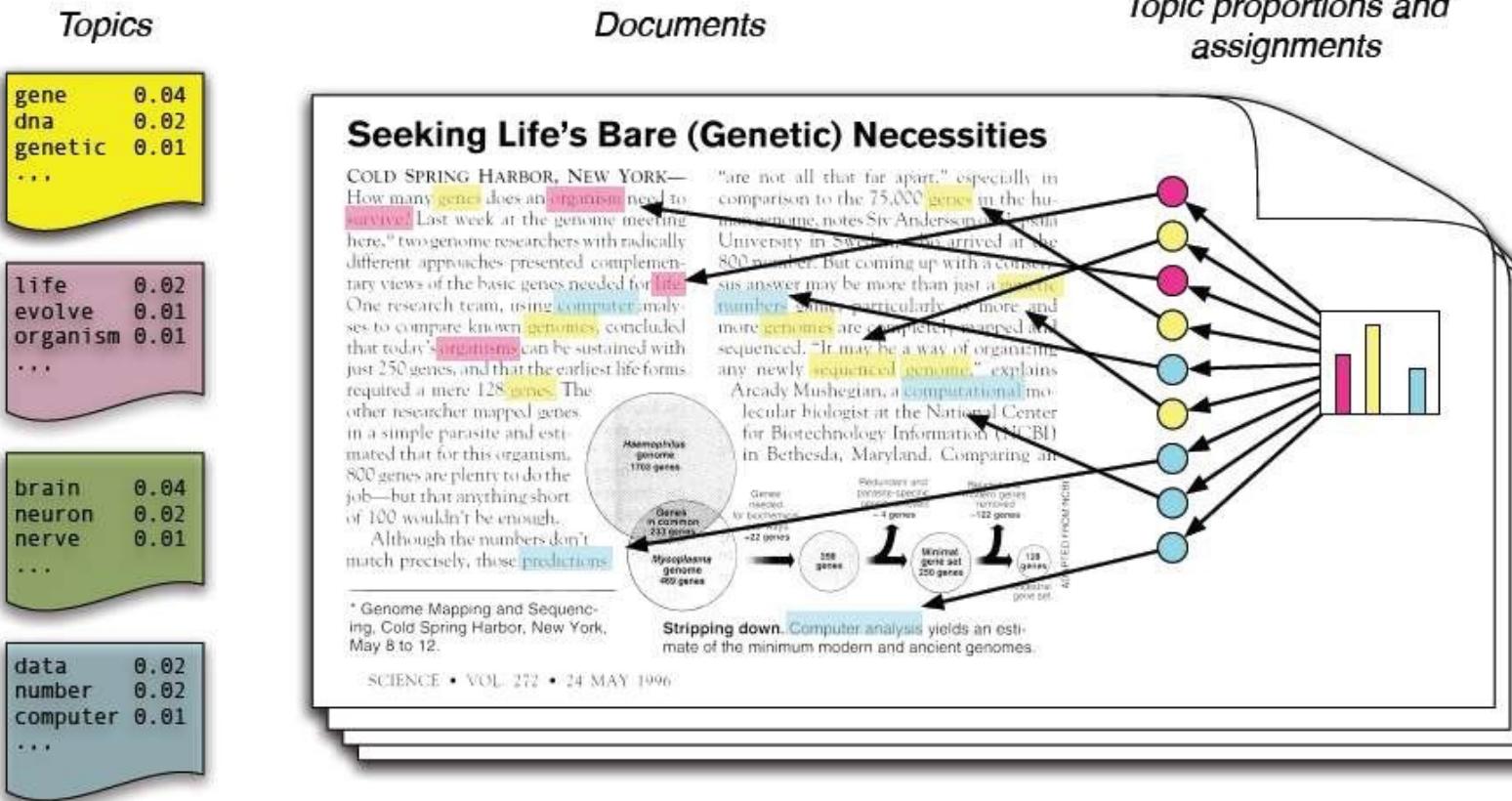
Можно переводить слова:

- $\overrightarrow{\text{математика}} + (\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}) \approx \overrightarrow{\text{math}}$
- $\overrightarrow{\text{король}} + (\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}) \approx \overrightarrow{\text{king}}$
- $\overrightarrow{\text{корова}} + (\overrightarrow{\text{word}} - \overrightarrow{\text{слово}}) \approx \overrightarrow{\text{cow}}$

# Тематическое моделирование

- Рассматриваем каждый документ как мешок слов
- Всего  $K$  тем
- Тема — распределение на словах
- Документ — распределение на темах

# Тематическое моделирование



# Модель PLSA

- Probabilistic Latent Semantic Analysis

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}$$

- $T$  — множество тем
- $p(w|t) = \varphi_{wt}$  — распределение слов в теме  $t$
- $p(t|d) = \theta_{td}$  — распределение тем в документе  $d$

# Модель PLSA

- Probabilistic Latent Semantic Analysis

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d) \rightarrow \max_{\varphi_{wt}, \theta_{td}}$$

Ограничения:  $\varphi_{wt} \geq 0, \theta_{td} \geq 0, \sum_{w \in W} \varphi_{wt} = 1, \sum_{t \in T} \theta_{td} = 1$

- $D$  — множество документов
- $W$  — множество слов

# Пример

- Данные: новостные заголовки

```
Topic: 0 Word: 0.008*"octob" + 0.006*"search" + 0.006*"miss" + 0.006*"inquest" + 0.005*"stori" + 0.005*"jam" + 0.004*"john" + 0.004*"harvest" + 0.004*"australia" + 0.004*"world"
Topic: 1 Word: 0.006*"action" + 0.006*"violenc" + 0.006*"thursday" + 0.005*"domest" + 0.005*"cancer" + 0.005*"legal" + 0.005*"u
nion" + 0.005*"breakfast" + 0.005*"school" + 0.004*"student"
Topic: 2 Word: 0.023*"rural" + 0.018*"govern" + 0.013*"news" + 0.012*"podcast" + 0.008*"grandstand" + 0.008*"health" + 0.007*"b
udget" + 0.007*"busi" + 0.007*"nation" + 0.007*"fund"
Topic: 3 Word: 0.030*"countri" + 0.028*"hour" + 0.009*"sport" + 0.008*"septemb" + 0.008*"wednesday" + 0.007*"commiss" + 0.006
*"royal" + 0.006*"updat" + 0.006*"station" + 0.005*"bendigo"
Topic: 4 Word: 0.014*"south" + 0.009*"weather" + 0.009*"north" + 0.008*"west" + 0.008*"coast" + 0.008*"australia" + 0.006*"eas
t" + 0.006*"queensland" + 0.006*"storm" + 0.005*"season"
Topic: 5 Word: 0.008*"monday" + 0.008*"august" + 0.006*"babii" + 0.005*"shorten" + 0.005*"hobart" + 0.004*"victorian" + 0.004*"d
onald" + 0.004*"safe" + 0.004*"scott" + 0.004*"donat"
Topic: 6 Word: 0.022*"interview" + 0.013*"market" + 0.009*"share" + 0.008*"cattl" + 0.008*"trump" + 0.008*"turnbul" + 0.007*"no
vemb" + 0.007*"michael" + 0.006*"australian" + 0.006*"export"
Topic: 7 Word: 0.019*"crash" + 0.014*"kill" + 0.009*"fatal" + 0.009*"dead" + 0.007*"die" + 0.007*"truck" + 0.007*"polic" + 0.00
6*"attack" + 0.006*"injur" + 0.006*"bomb"
Topic: 8 Word: 0.008*"drum" + 0.007*"abbott" + 0.007*"farm" + 0.006*"dairi" + 0.006*"asylum" + 0.006*"tuesday" + 0.006*"water"
+ 0.006*"labor" + 0.006*"say" + 0.005*"plan"
Topic: 9 Word: 0.017*"charg" + 0.014*"murder" + 0.011*"court" + 0.011*"polic" + 0.009*"woman" + 0.008*"assault" + 0.008*"jail"
+ 0.008*"alleg" + 0.007*"accus" + 0.007*"guilty"
```

# Резюме

- Кластеризация — задача без строгой постановки и без строгих критериев качества
- Много разновидностей в подходах
- Методы: K-Means, DBSCAN, иерархическая кластеризация и т.д.
- Обучение без учителя — гораздо более широкая область

# Спасибо за внимание!



Ildar Safilo

@Ildar\_Saf

[irsafilo@gmail.com](mailto:irsafilo@gmail.com)

<https://www.linkedin.com/in/isafilo/>