

Введение в анализ данных

Лекция 16

Кластеризация

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2020

На прошлых лекциях

- Дано: матрица «объекты-признаки» X и, возможно, ответы y
- Найти: подмножество признаков или новые признаки

На прошлых лекциях

- Методы обучения с учителем: линейные модели, решающие деревья, случайные леса, ...
- Дано: матрица «объекты-признаки» X и ответы y
- Найти: модель $a(x)$

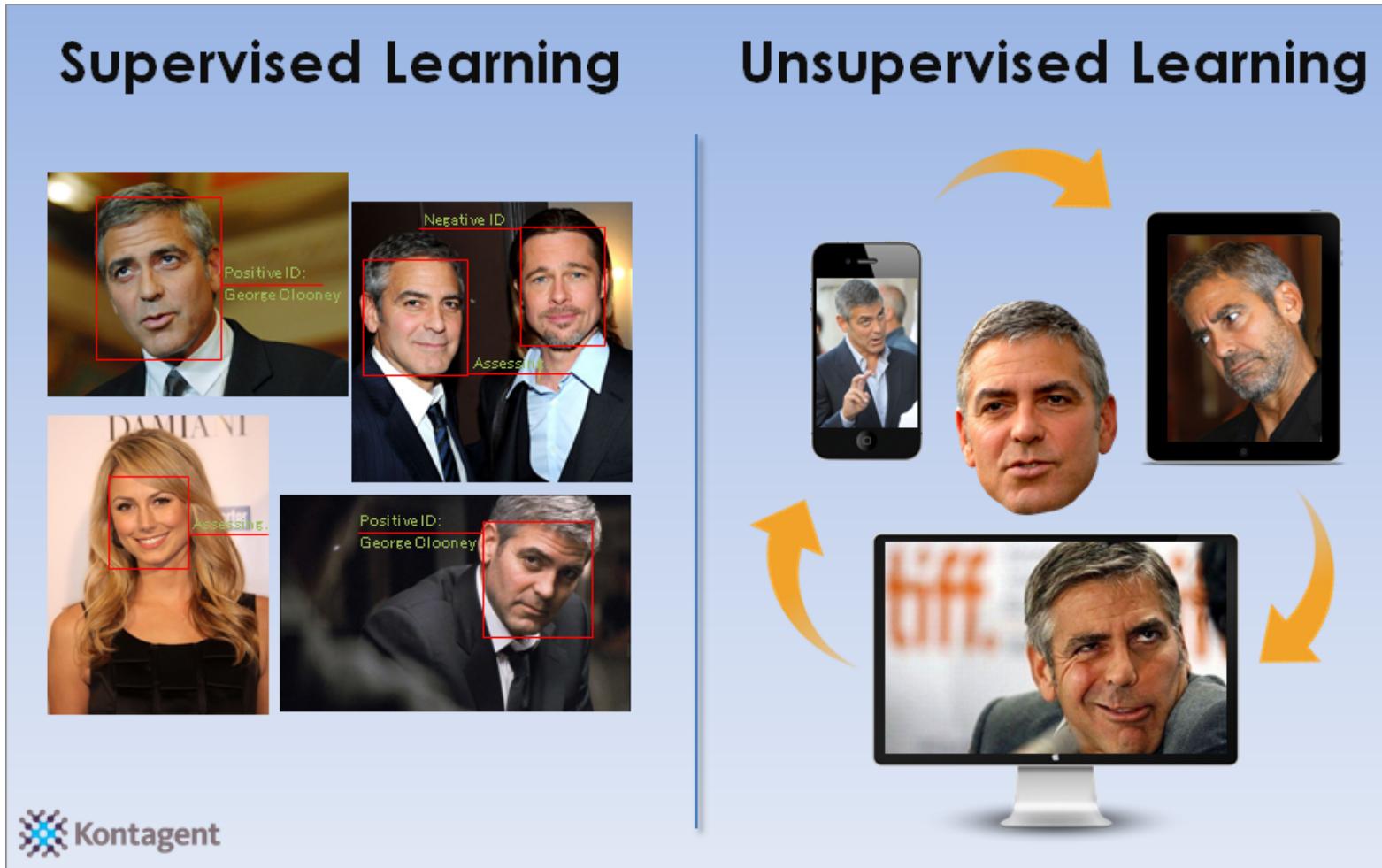
Обучение с учителем (supervised learning)

- 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

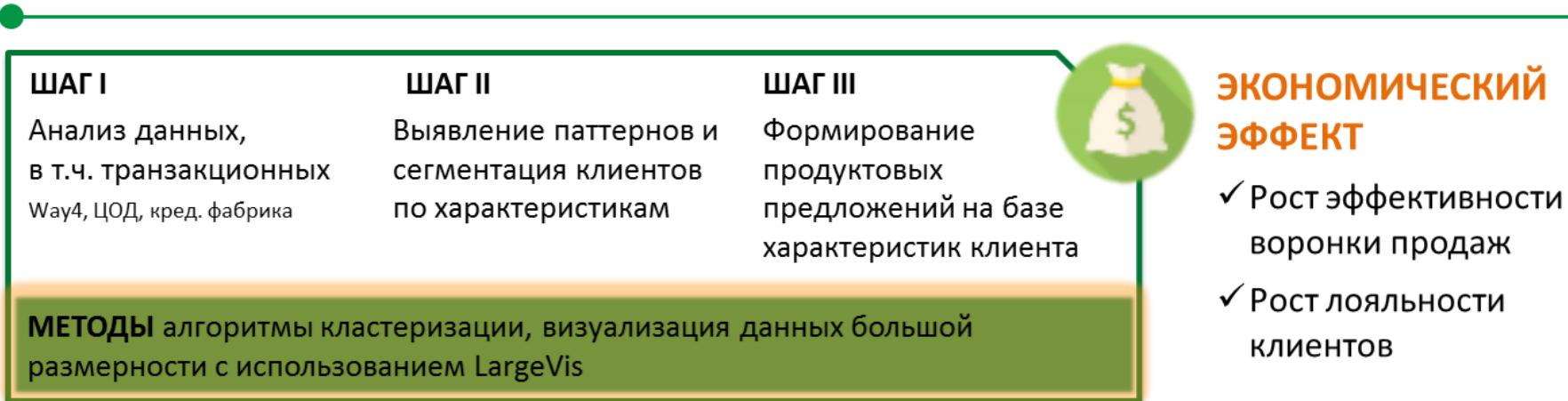


Обучение без учителя: предсказание кадра

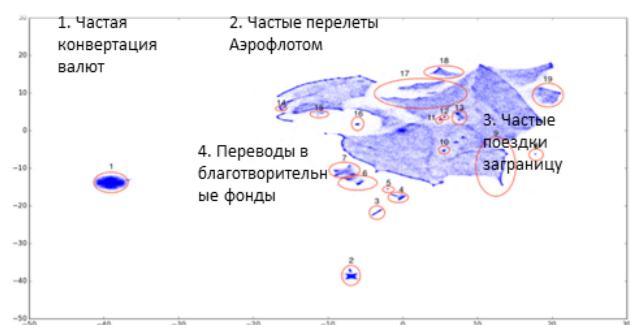


Обучение без учителя: кластеризация

Case 2. Оптимизация воронки продаж



КЛАСТЕРИЗАЦИЯ КЛИЕНТОВ ПО ХАРАКТЕРУ ТРАНЗАКЦИЙ



В ЗАВИСИМОСТИ ОТ КЛАСТЕРА
КЛИЕНТА ПРЕДЛОЖИТЬ
РЕЛЕВАНТНЫЙ ПРОДУКТ



Паттерн	Продукт
1. Частая конвертация валют	Мультивалютный счет
2. Частые перелеты Аэрофлотом	Карта «Аэрофлот Бонус»
3. Частые поездки заграницу	Страховка для выезжающих за рубеж
4. Переводы в благотворительные фонды	Карта «Подари жизнь»

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

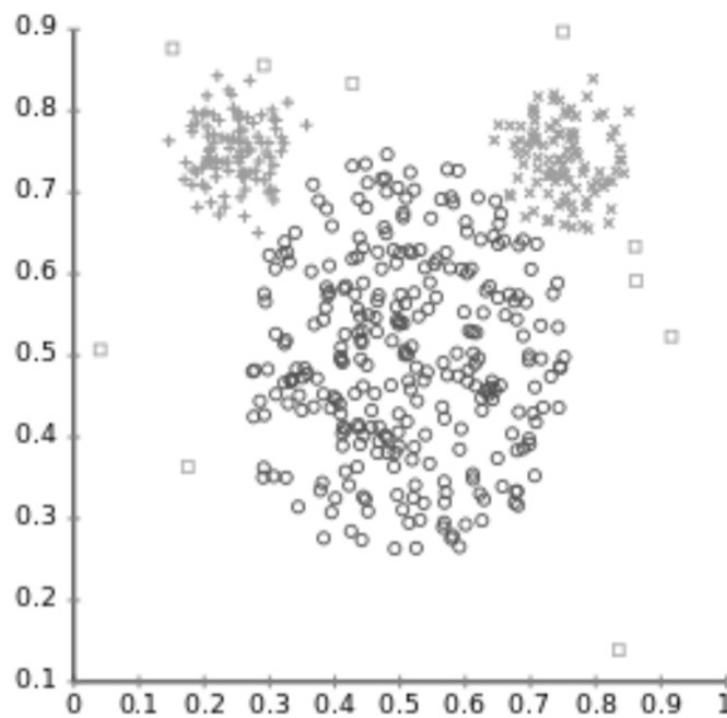
Форма кластеров



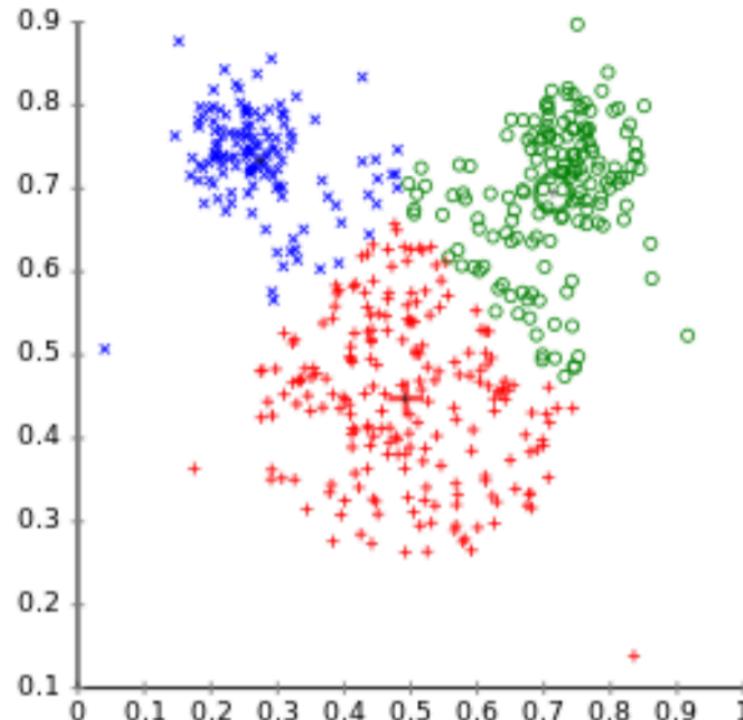
Форма кластеров



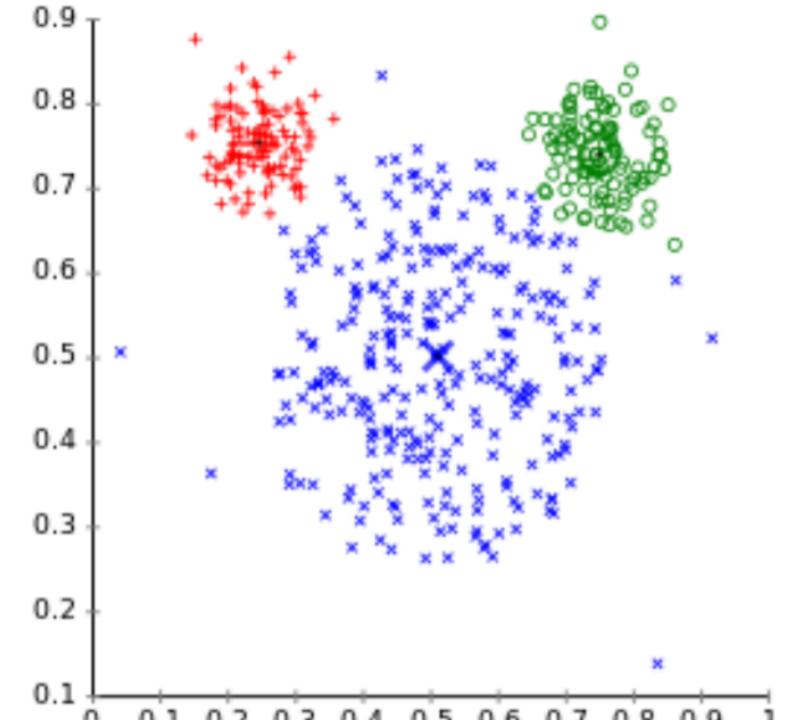
Различия в результатах работы



Исходная выборка
("Mouse" dataset)

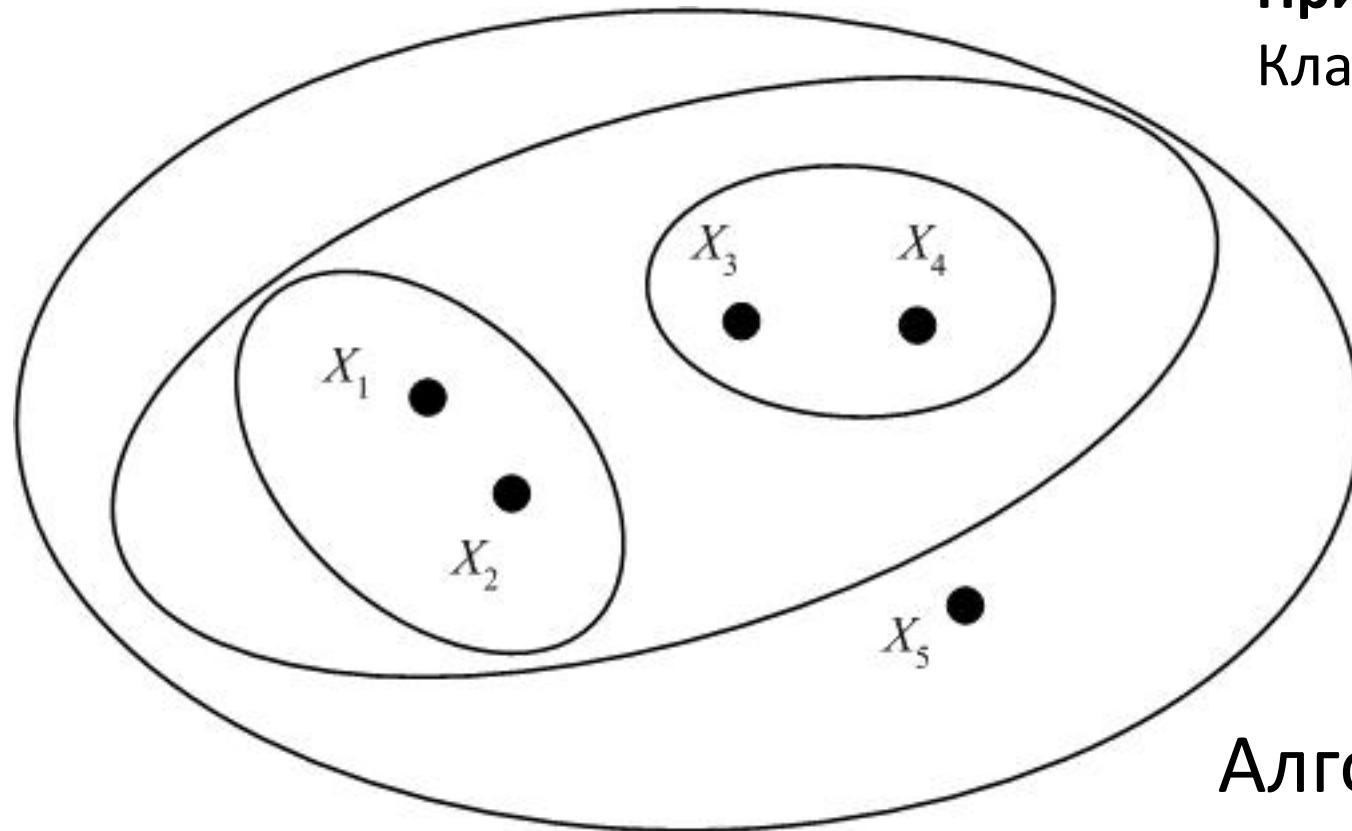


Метод 1



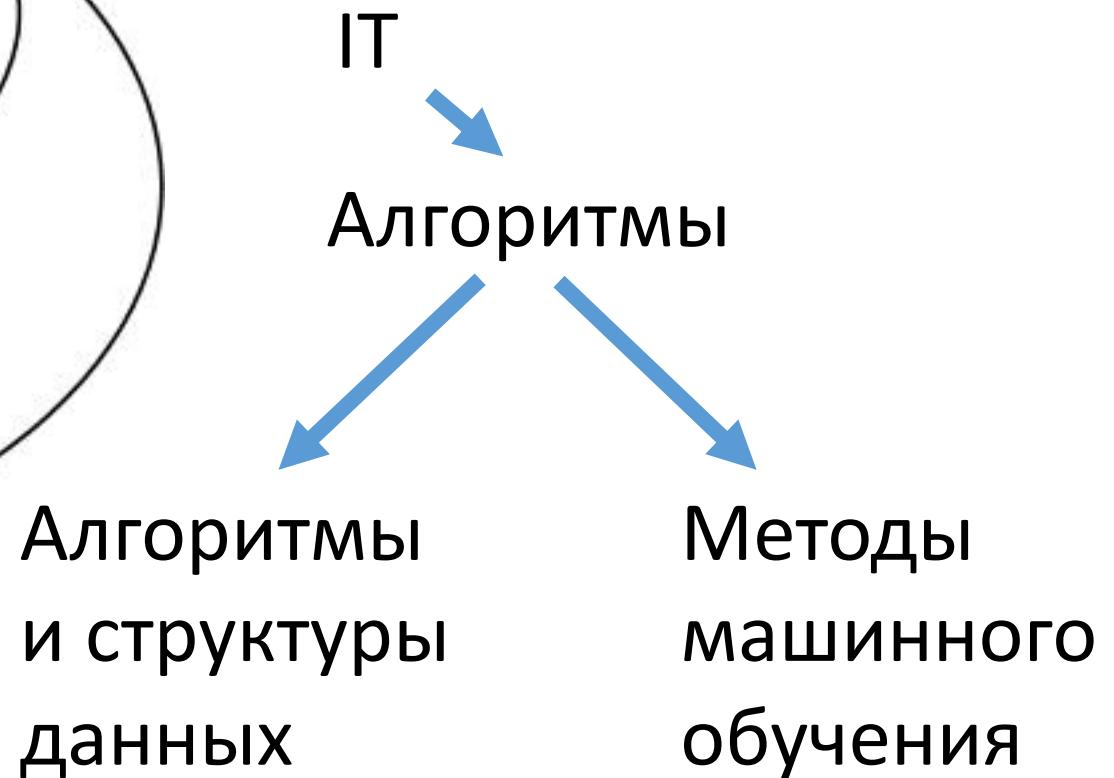
Метод 2

Иерархическая кластеризация



Пример:

Кластеризация статей с Хабра



Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 1: в один кластер должны попадать новости на одну тему



[Батыршин сыграет вместо Хабарова у «Магнитки» в матче с «Салаватом»](#)

Место в третьей паре защиты «Магнитки» на третью встречу плей-офф Кубка Гагарина с «Салаватом Юлаевым» занял защитник Рафаэль Батыршин, сообщает из Уфы корреспондент «Чемпионата» Павел Панышев. Травмированный Ярослав Хабаров выбыл на неопределённый срок. Для форварда Оскара Осалы сезон закончен.



[Футболисты ЦСКА проиграли «Долгопрудному» в товарищеском матче](#)

Футболисты московского ЦСКА со счетом 2:3 проиграли клубу второго дивизиона "Долгопрудный" в товарищеском матче, который состоялся в Москве на стадионе "Октябрь". У армейцев забитыми мячами отличились Александр Цауня (15-я минута) и Сергей Ткачев (54).

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

- Задача кластеризации новостей по содержанию.
- Постановка 3: в один кластер должны попадать тексты об одной и той же новости

11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи
посмотрели несколько миллиардов
человек

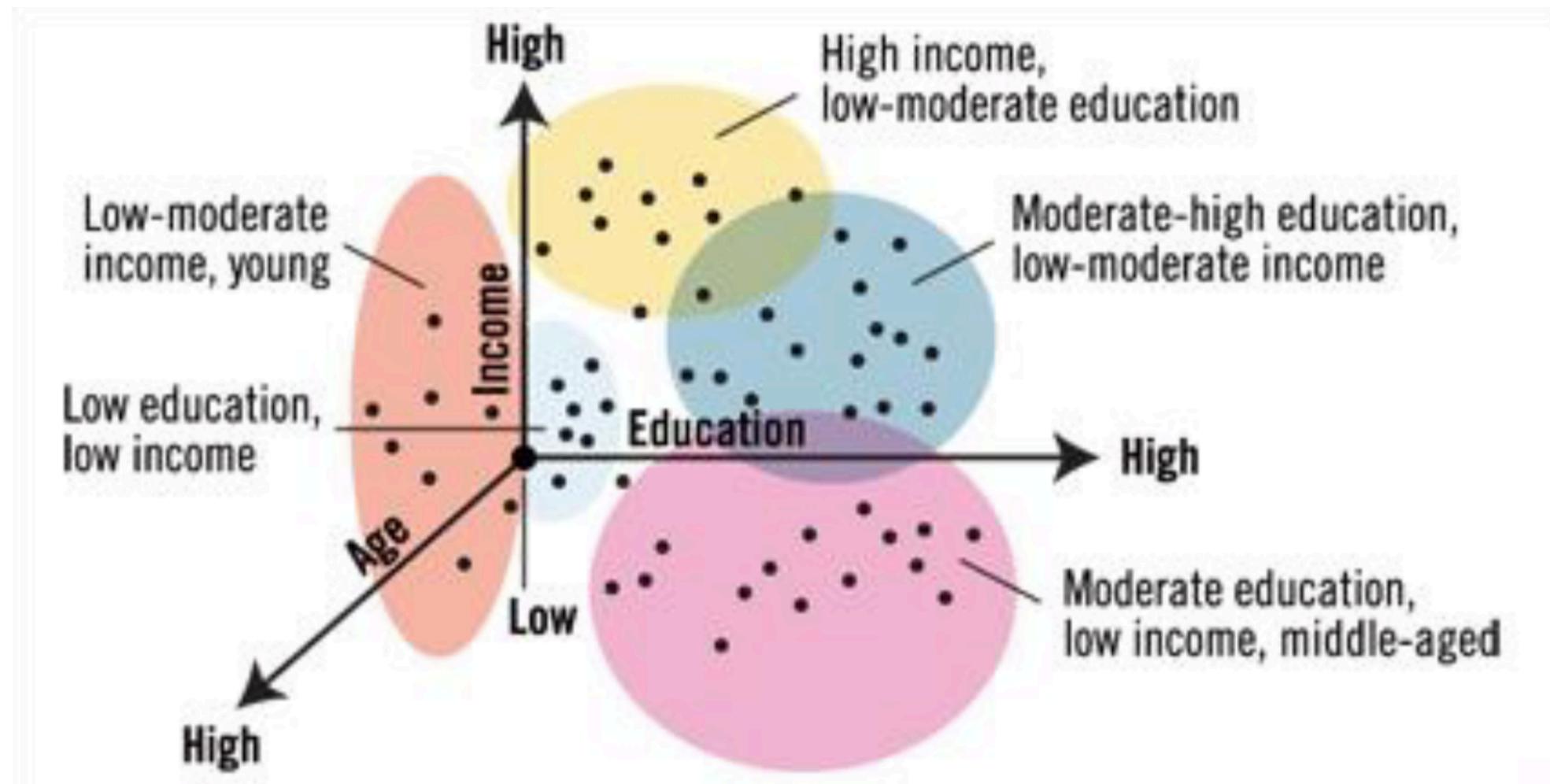
Олимпиада в Сочи открыта

**Церемония открытия Олимпиады в
Сочи. Онлайн-репортаж**

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

Кластеризация как основная задача



Кластеризация как вспомогательная задача

Цель: улучшение распознавания

5

5

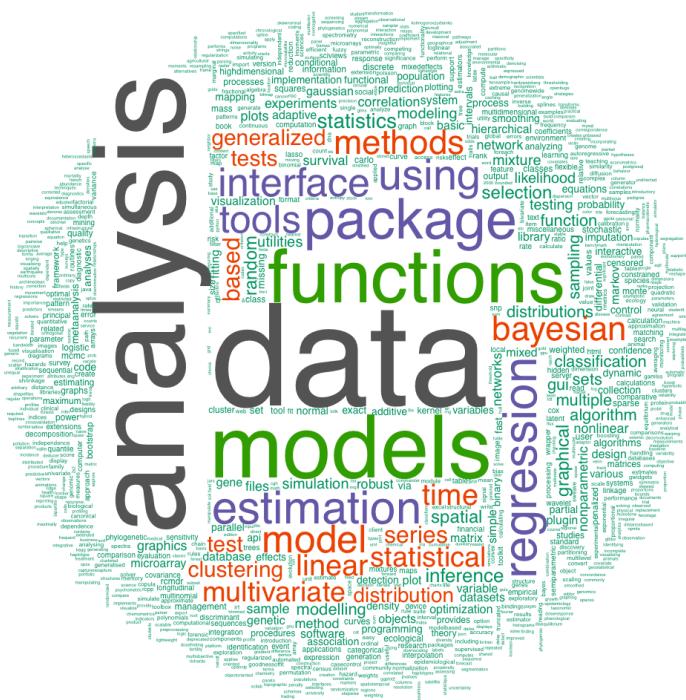
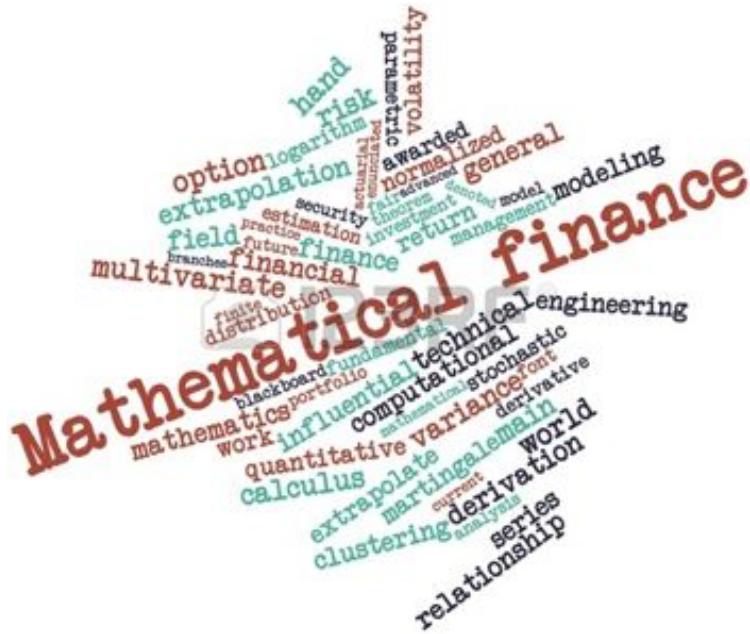
5

5

5

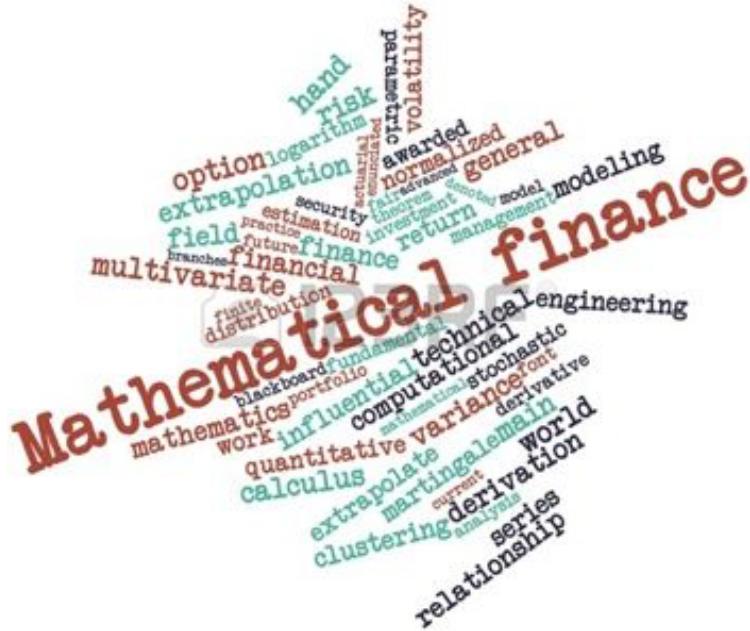
«Жёсткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

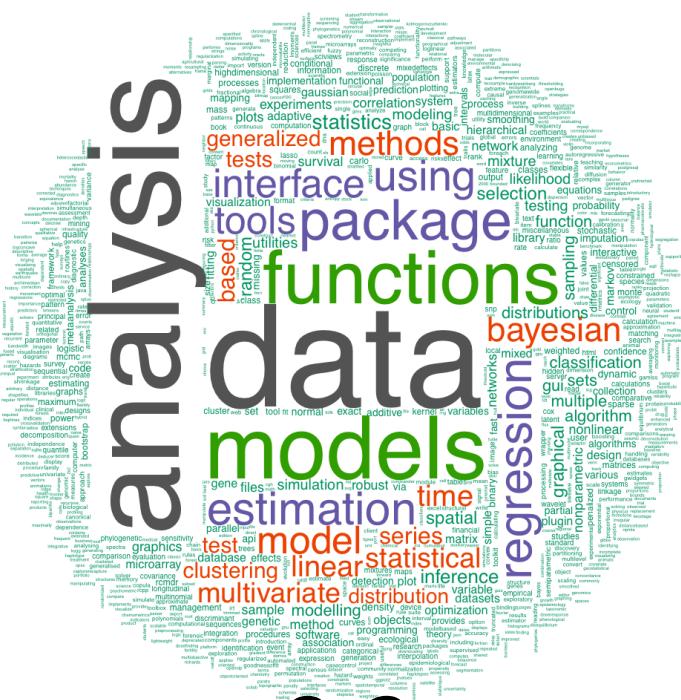


«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»



0.2



0.3



0.5

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_ℓ
- 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

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

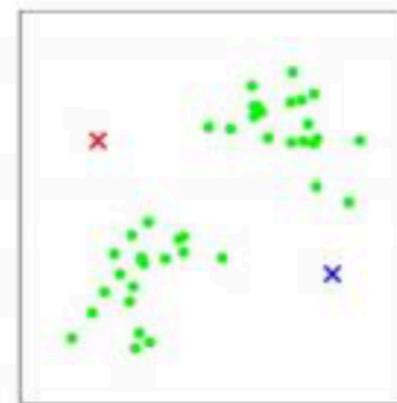
- Шаг Б: переместить центр каждого кластера в центр тяжести

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

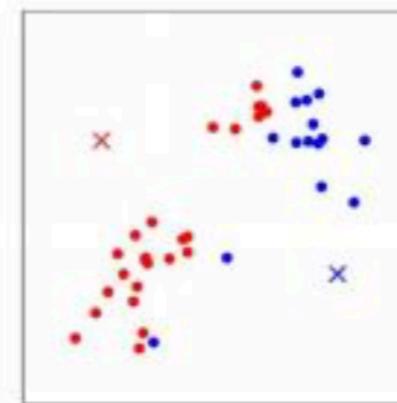
K-Means



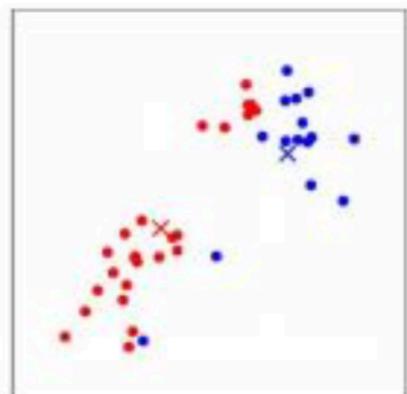
(a)



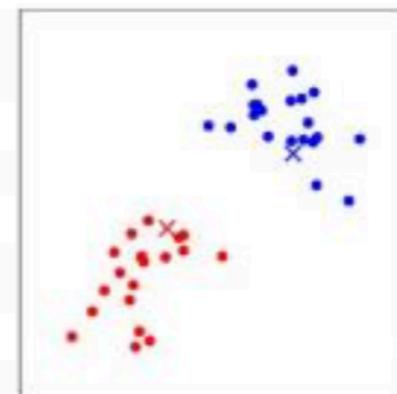
(b)



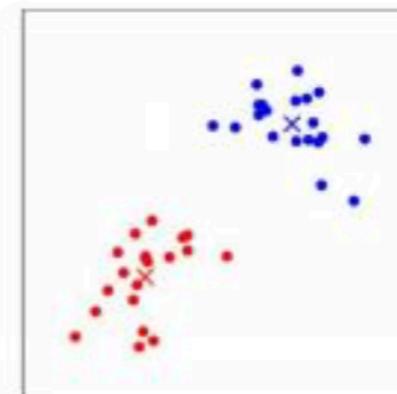
(c)



(d)

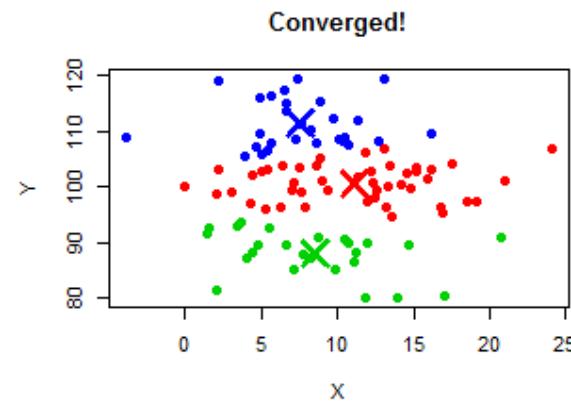
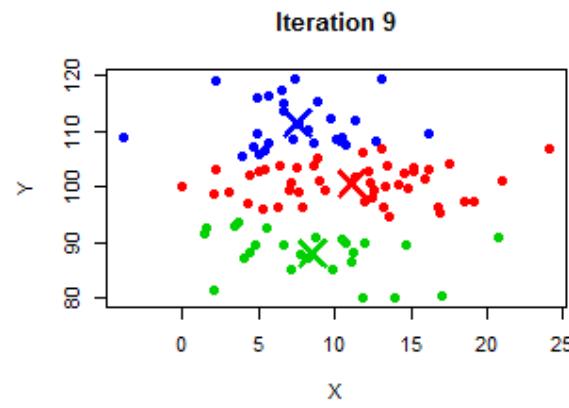
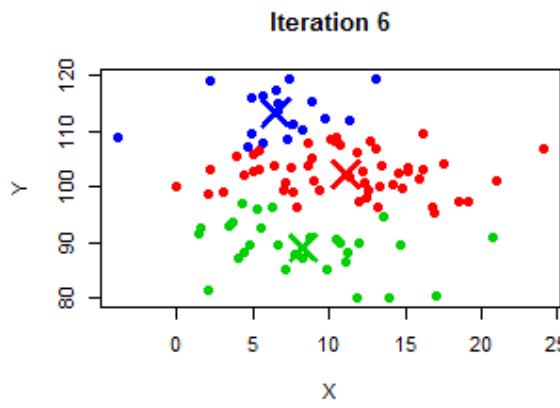
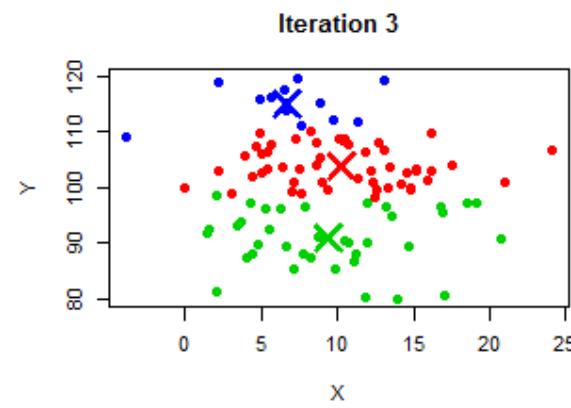
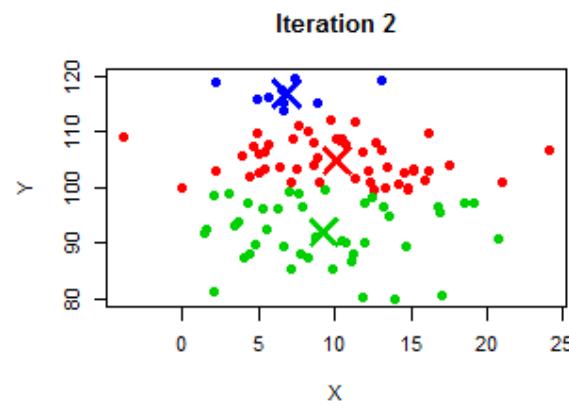
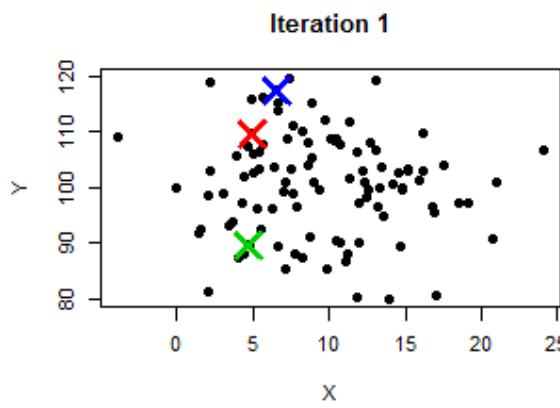


(e)



(f)

K-Means



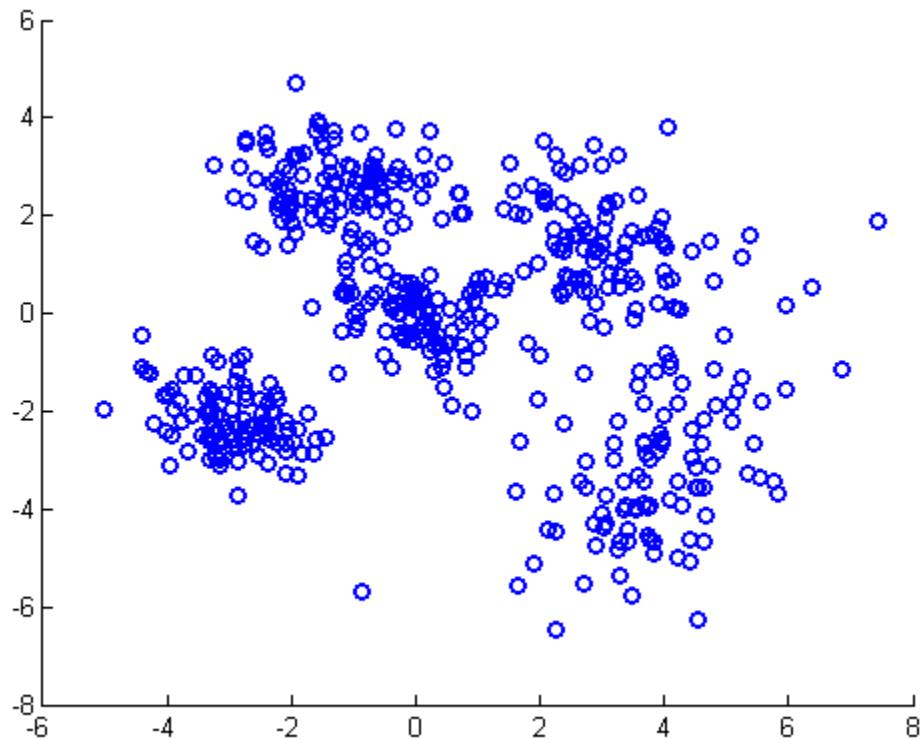
Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

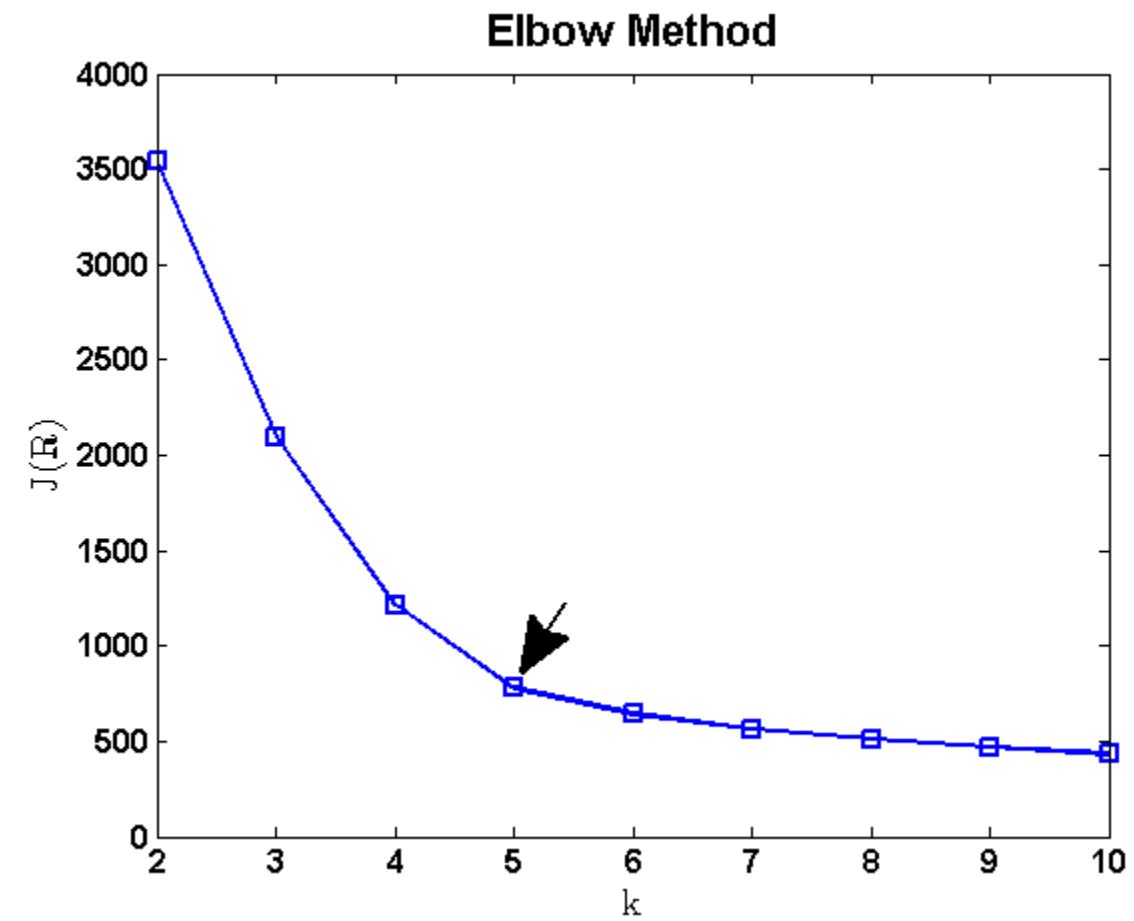
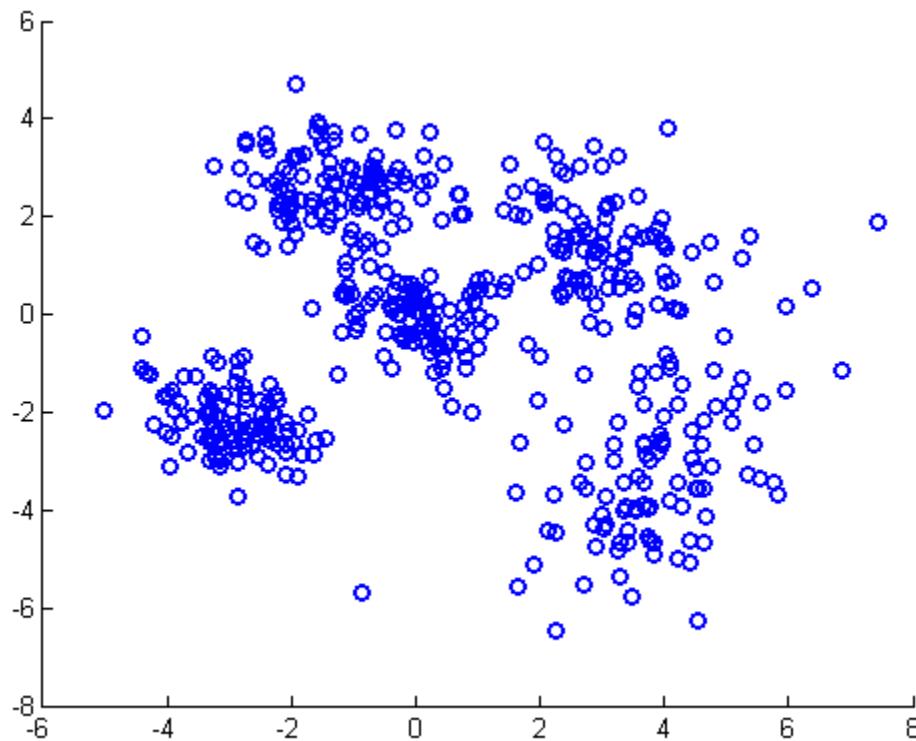
$$J(C) = \sum_{i=1}^{\ell} \rho(x_i, c_{y_i})$$

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

Выбор числа кластеров



Выбор числа кластеров

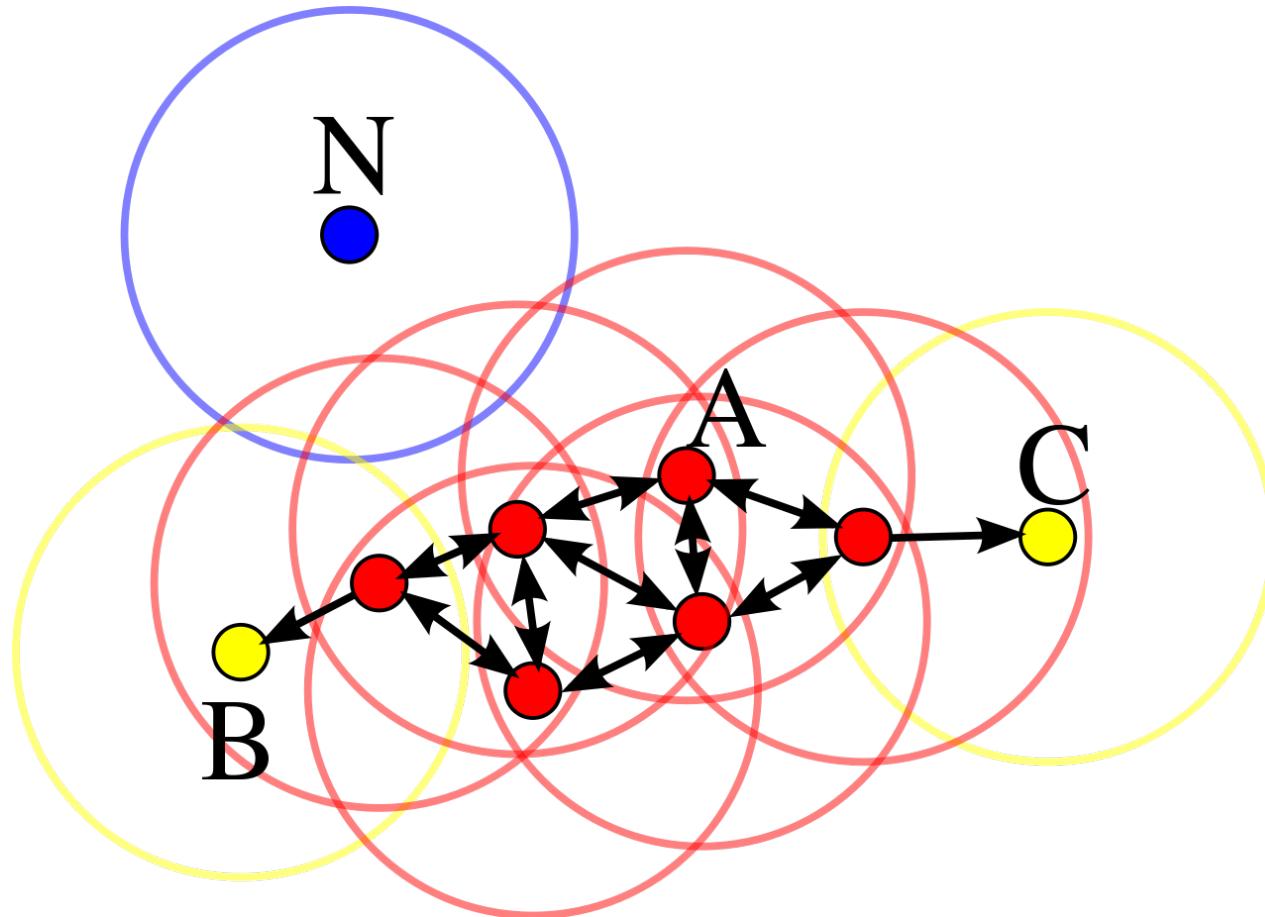


Особенности K-Means

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

Density-based clustering

Основные, граничные и шумовые точки



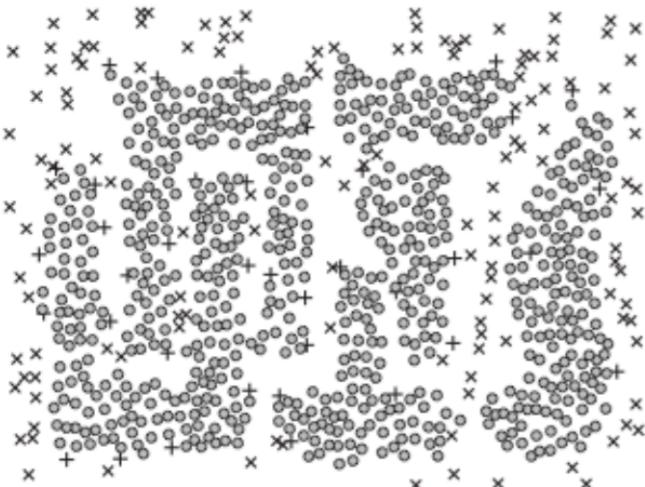
Параметры DBSCAN

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

DBSCAN



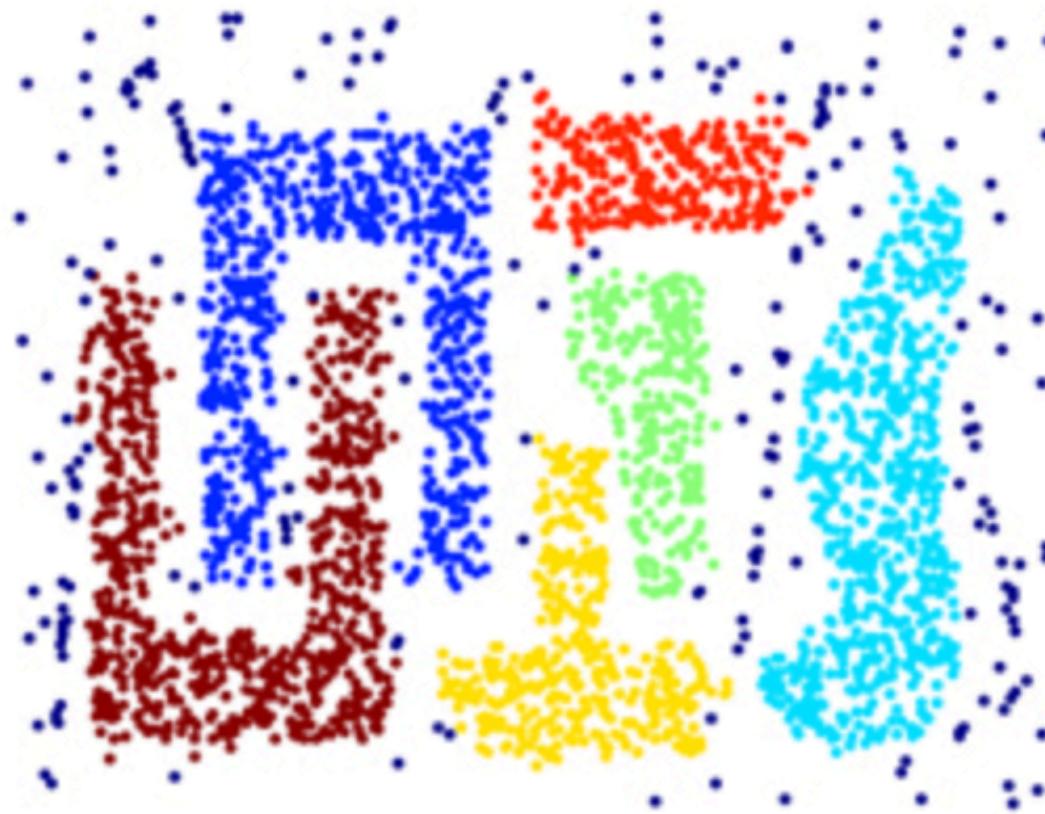
(a) Clusters found by DBSCAN.



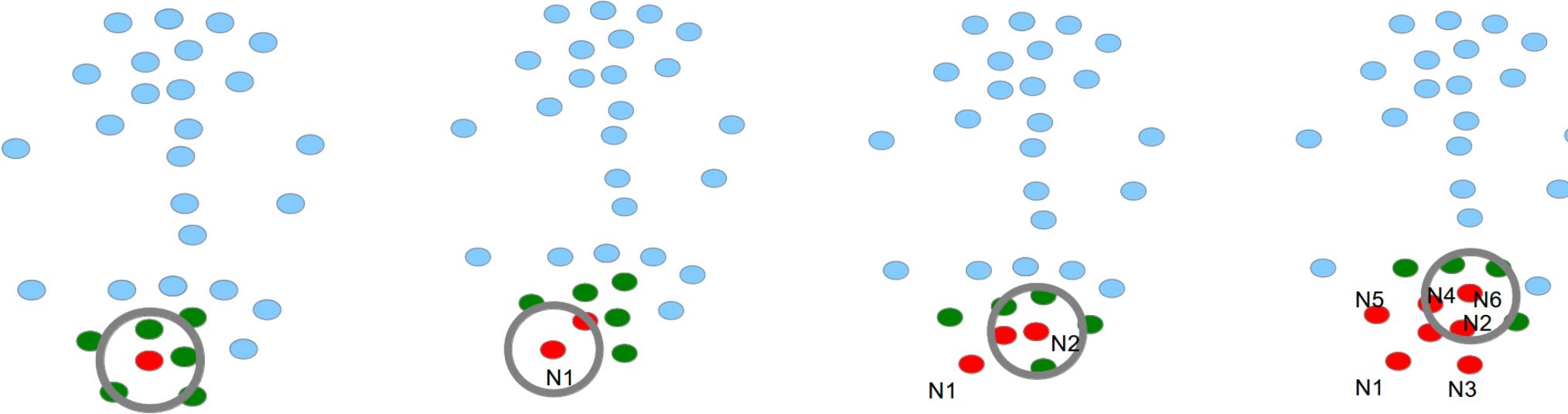
(b) Core, border, and noise points.

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

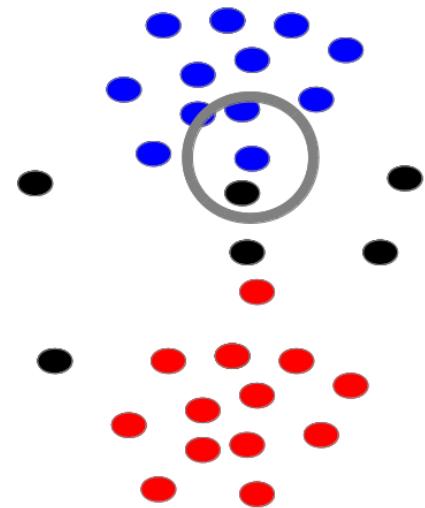
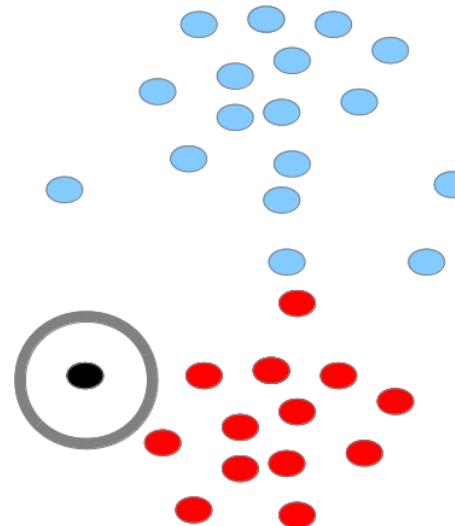
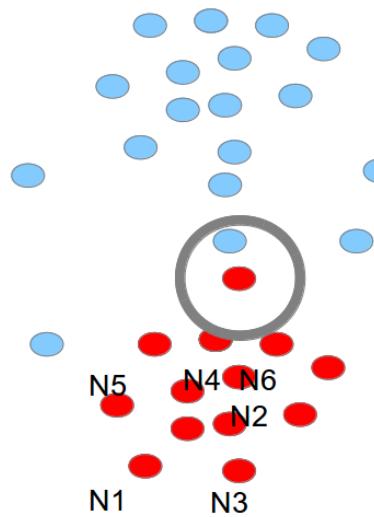
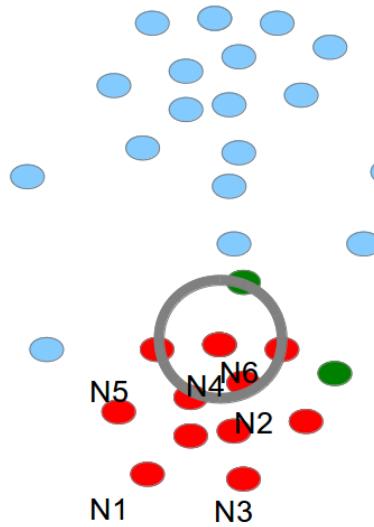
DBSCAN: результаты работы



Пример



Пример



Особенности DBSCAN

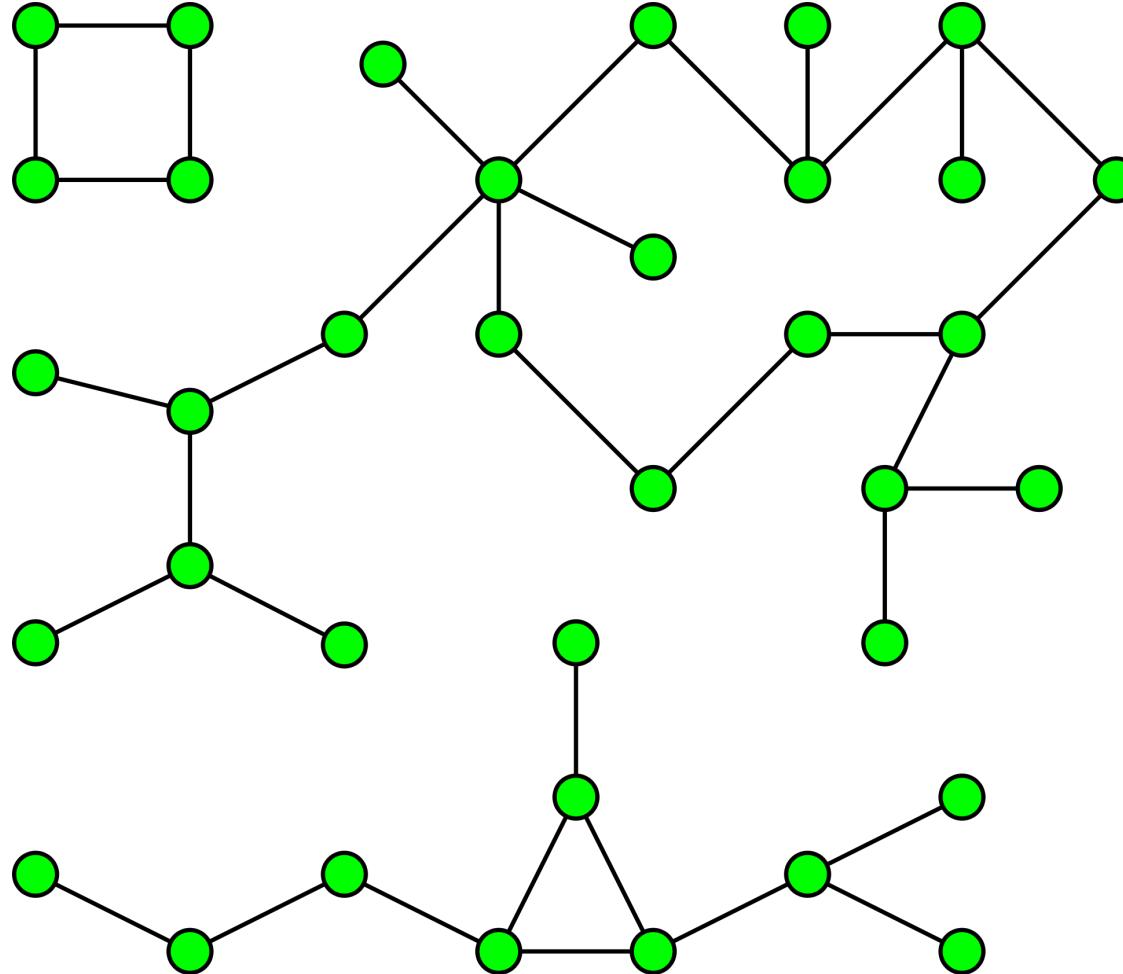
- Находит кластеры произвольной формы
- Может работать с большими объёмами данных
- Нужно подбирать размер окрестности (eps) и минимальное число объектов в окрестности

Графовые методы

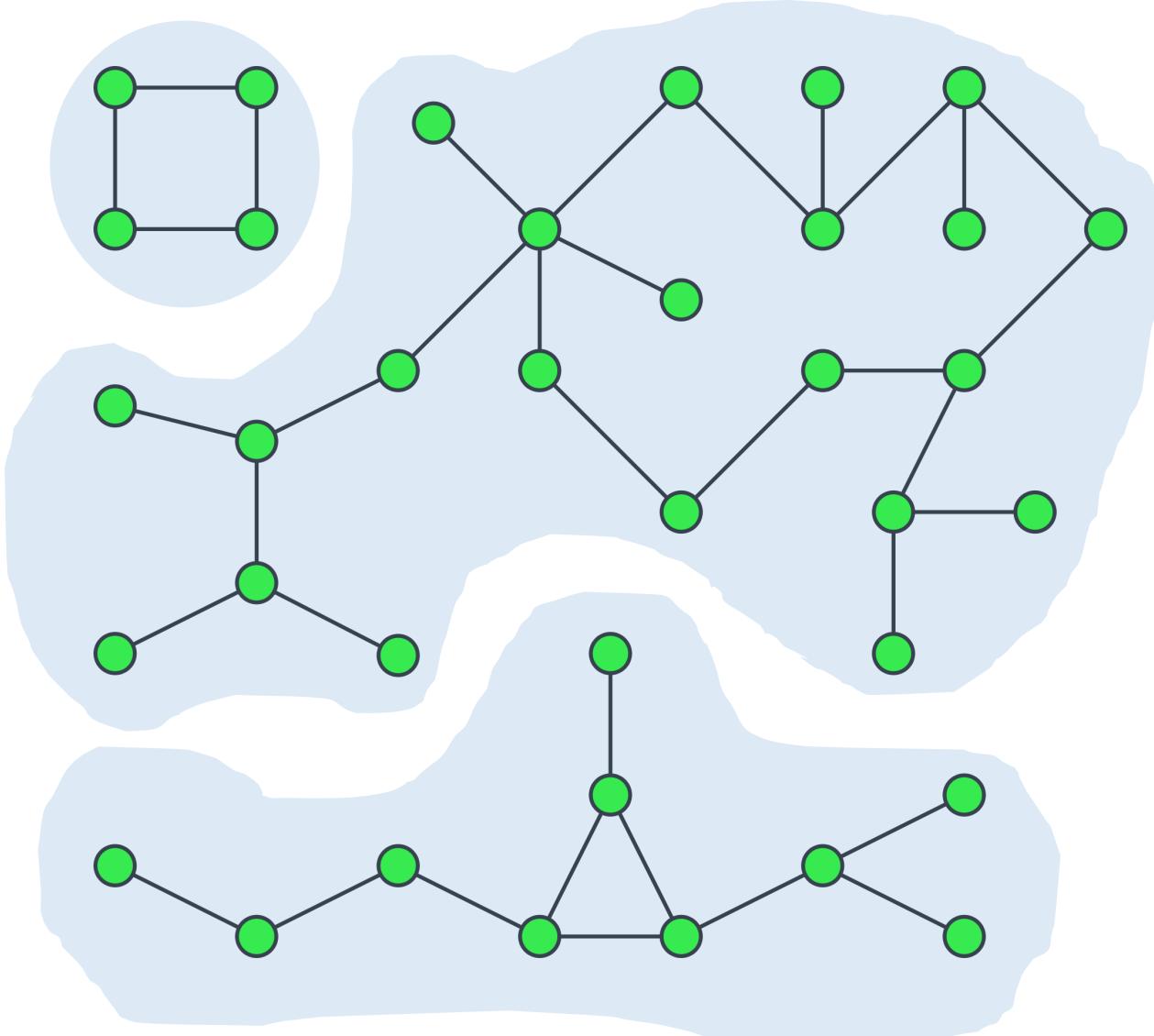
Кластеризация по компонентам связности

- Граф: вершины соответствуют объектам
- Соединяем ребром объекты, расстояние между которыми меньше R
- Выделяем компоненты связности

Выделение связных компонент



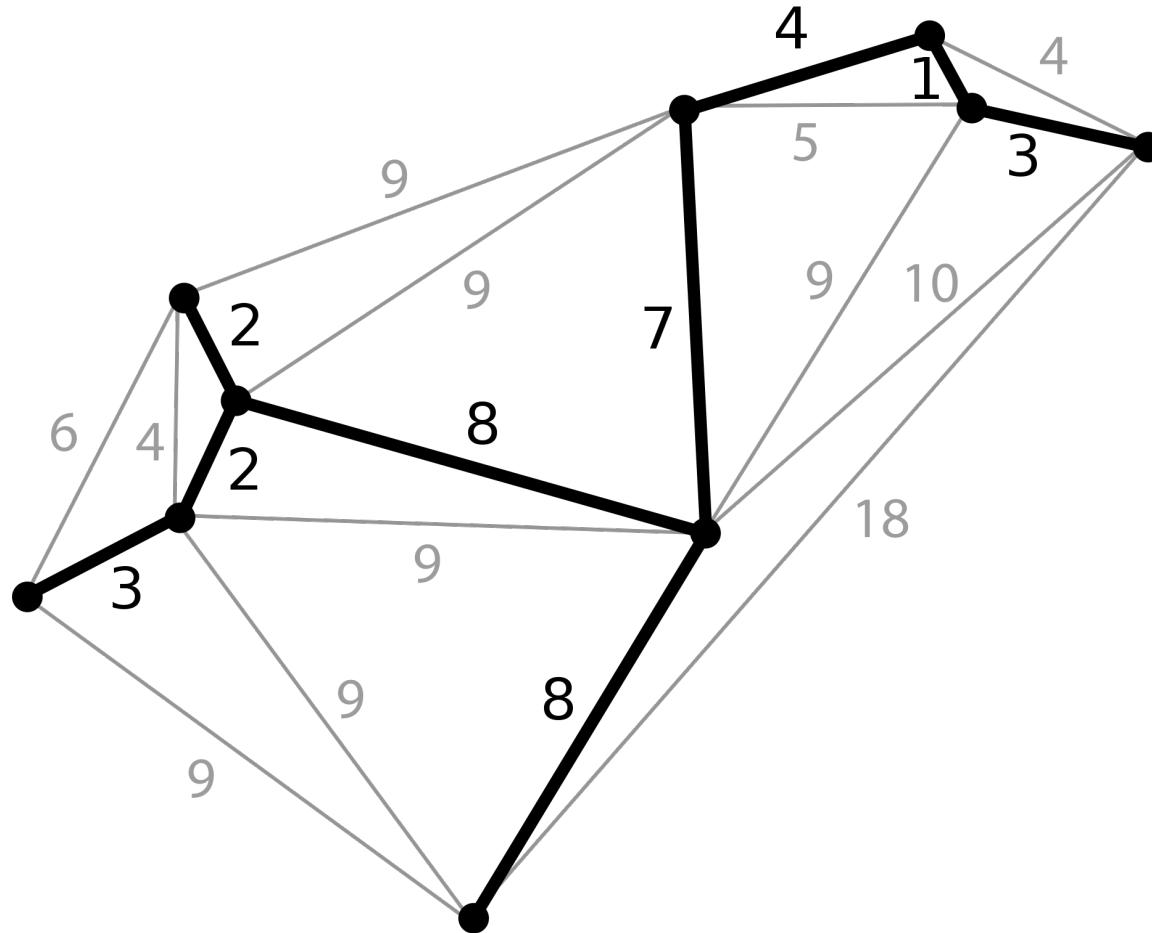
Выделение связных компонент



Кластеризация по компонентам связности

- Быстрая и простая
- Параметр — минимальное расстояние R
- Проблема: непонятно, как выбрать R , если нужно получить K кластеров

Минимальное остовное дерево



Минимальное оставное дерево

- Строим взвешенный граф, где веса ребер – расстояния между объектами
- Строим минимальное оставное дерево для этого графа
- Дерево имеет $\ell - 1$ ребро
- Удаляем $K - 1$ ребро с максимальным весом
- Получаем K компонент связности, которые интерпретируем как кластеры

Обучение без учителя и
текстовые данные

Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

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 sentence "...an efficient method for learning high quality distributed vector ..." with two green brackets underneath. The first bracket covers the words "an efficient method for learning". The second bracket covers the words "high quality distributed vector". A blue arrow points upwards from the center of the second bracket to the word "vector". Below the first bracket is the word "context" and below the second bracket is the word "context". Between the two brackets is the word "focus word" with a blue arrow pointing upwards to the word "vector".

Векторные представления слов

Хотим представить каждое слово в виде вещественного вектора:

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

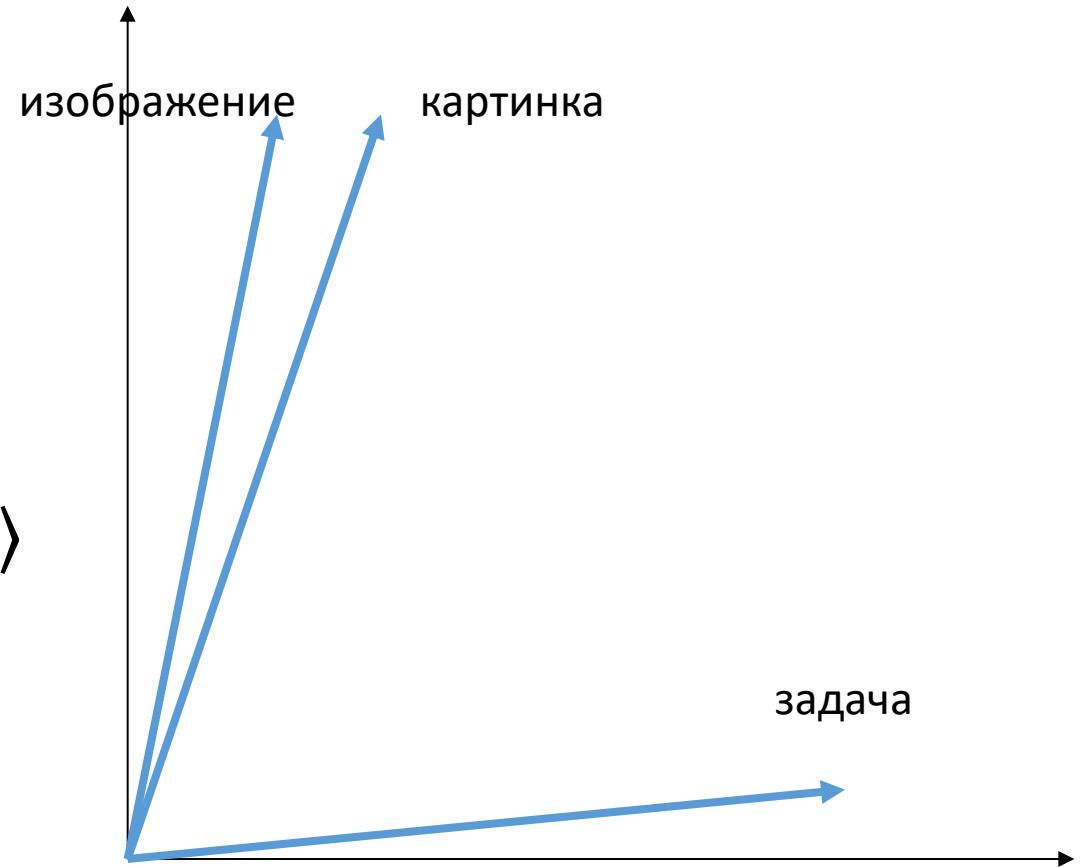
Требования к представлениям (embeddings):

- 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:

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

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

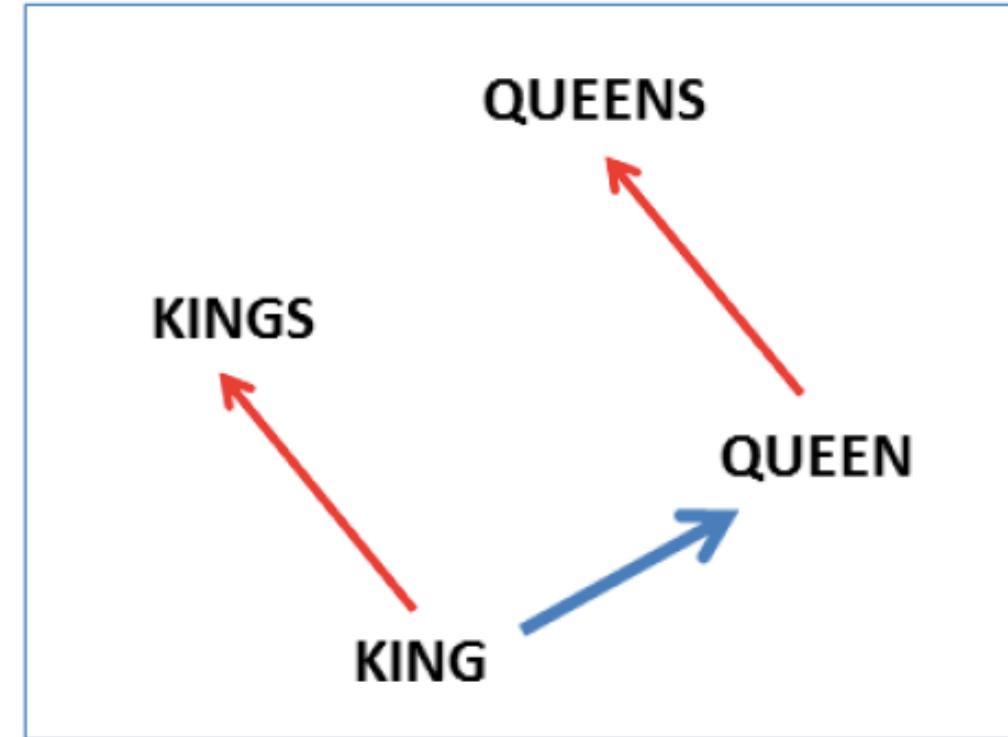
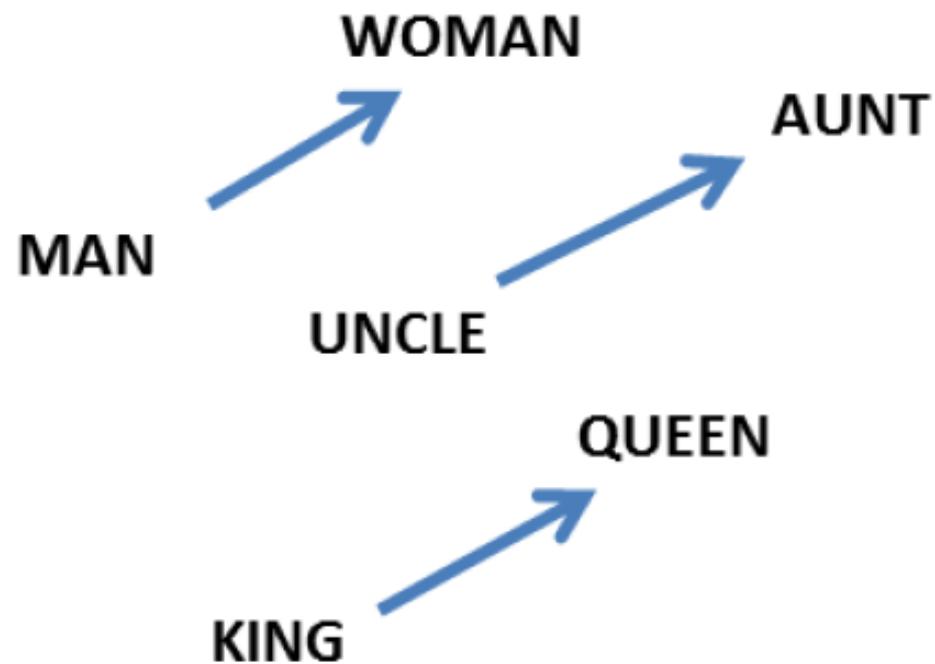
word2vec

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:

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



Примеры представлений

- food2vec
 - Документ — названия ингредиентов из рецепта
 - Слово — один ингредиент

food2vec

Tools for cooking with machine intelligence

Food similarity tool

Type in an ingredient to see which foods are similar:

Type in a food (e.g. Apple, Peanut, Bread)

Almonds



LOOK UP SIMILAR FOODS!

Whole almonds	0.708
Slivered almonds	0.695
Chopped almonds	0.657
Blanched almonds	0.651
Chocolate	0.616
Prunes	0.602
Raisins	0.597
Toasted almonds	0.589
Dried dates	0.586

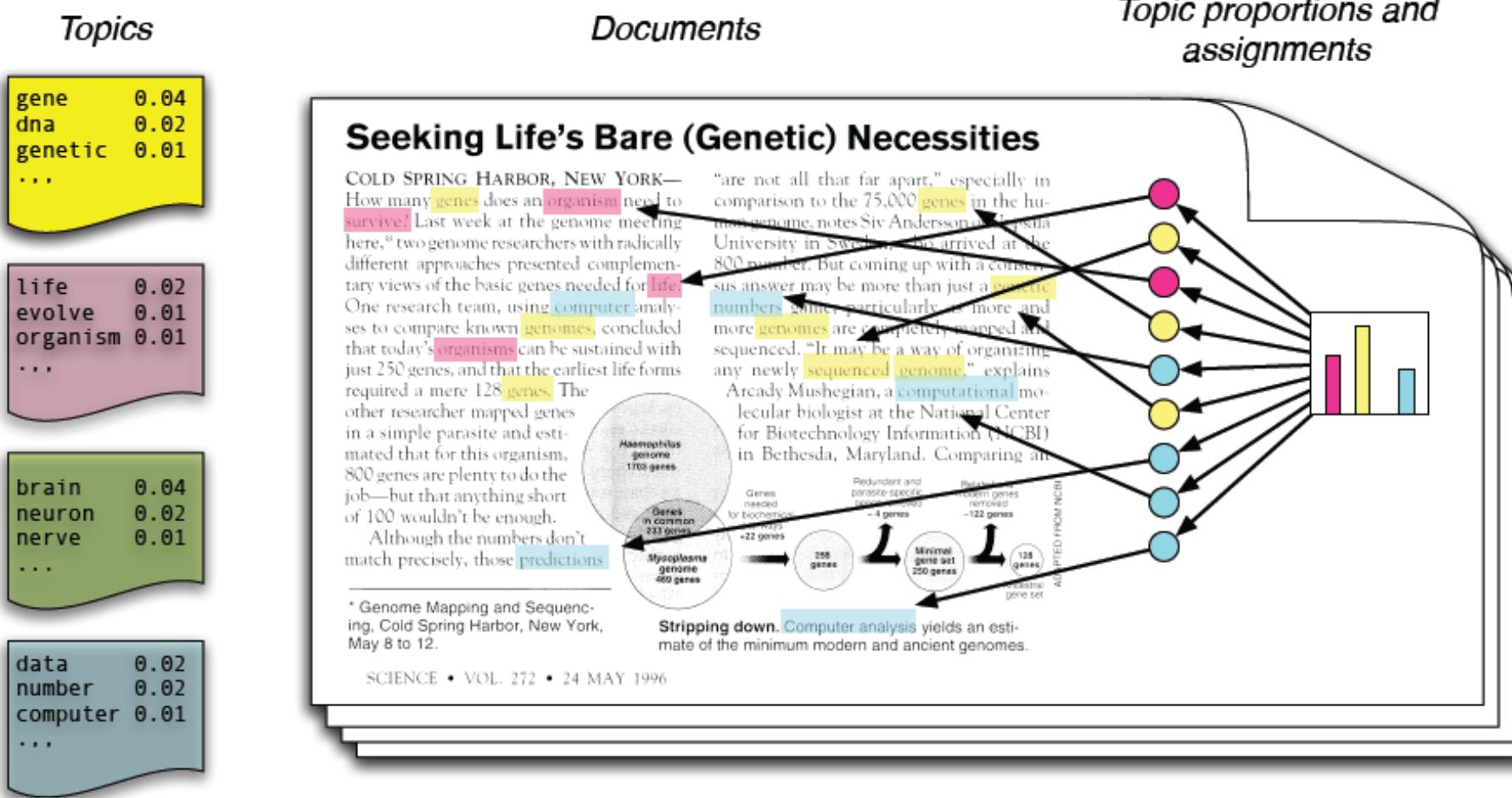
Примеры представлений

- Общее семантическое пространство в Яндексе
 - В векторы переводятся страницы сайтов, изображения, видео, ...

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

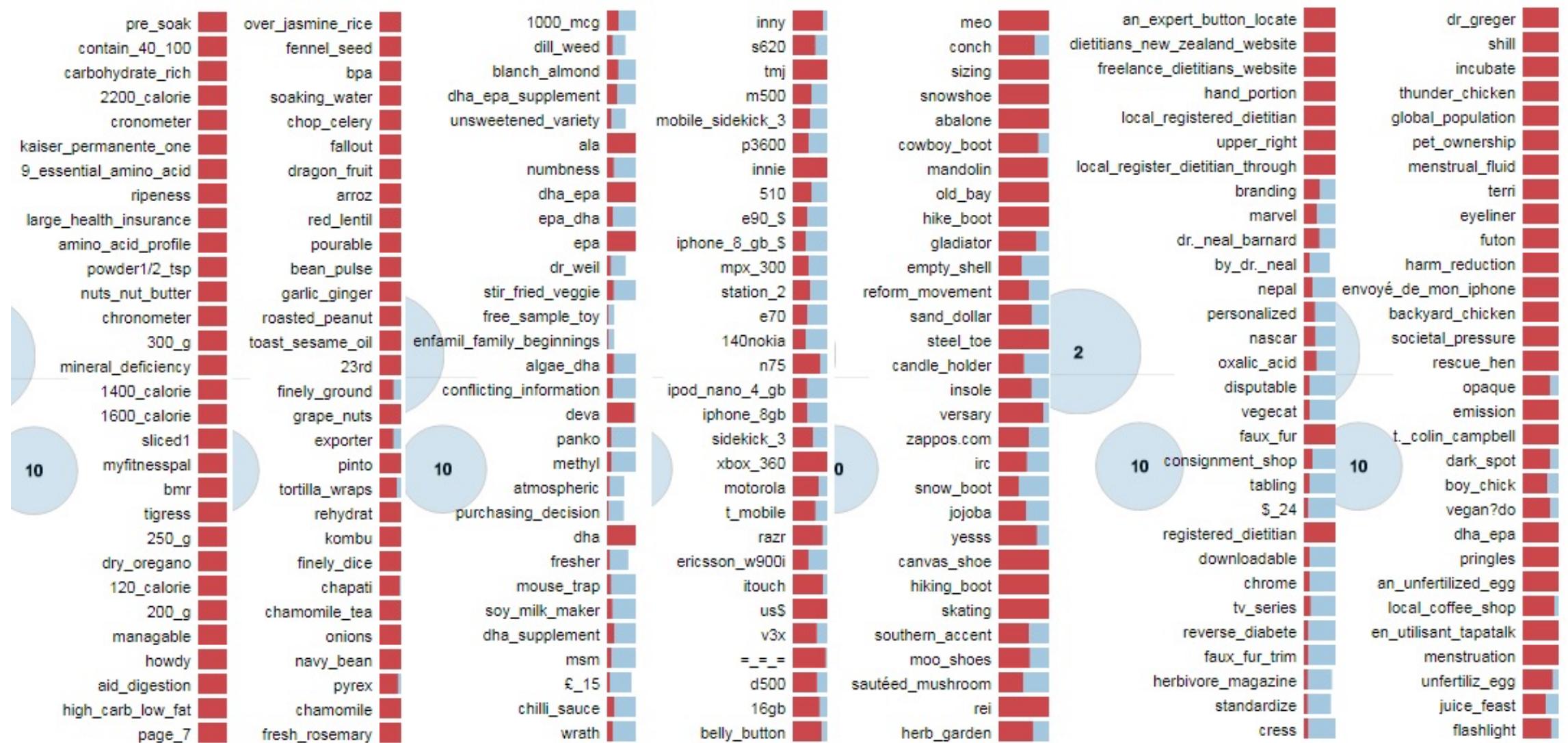
$$\sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w|d) \rightarrow \max_{\varphi_{wt}, \theta_{td}}$$

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

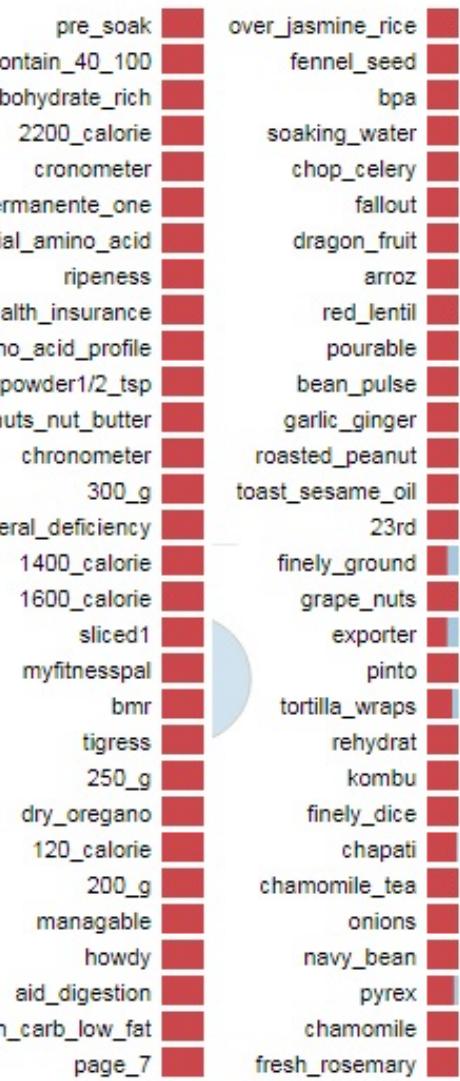
- D — множество документов
- W — множество слов

Примеры

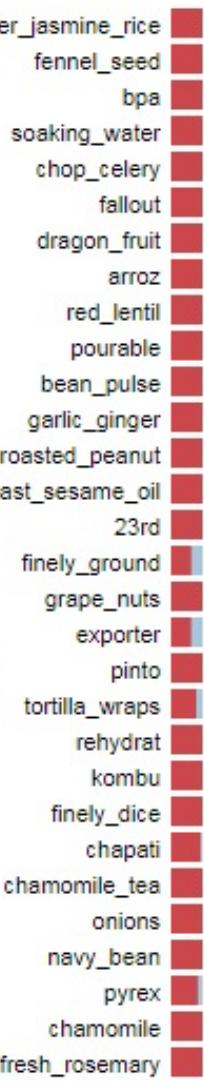
- Данные: сообщения с форума вегетарианцев



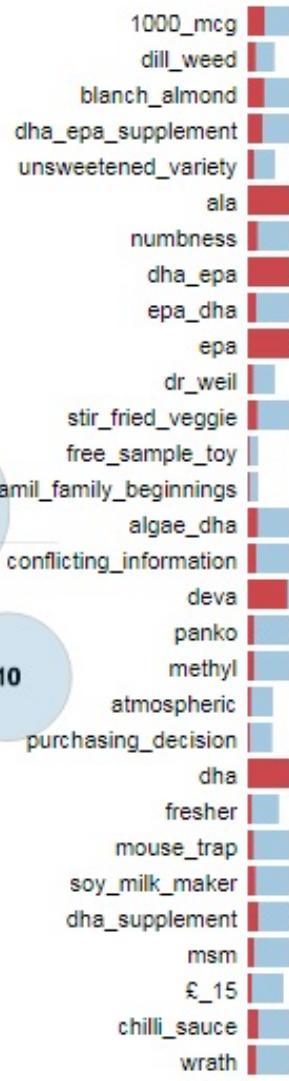
11 Nutrition



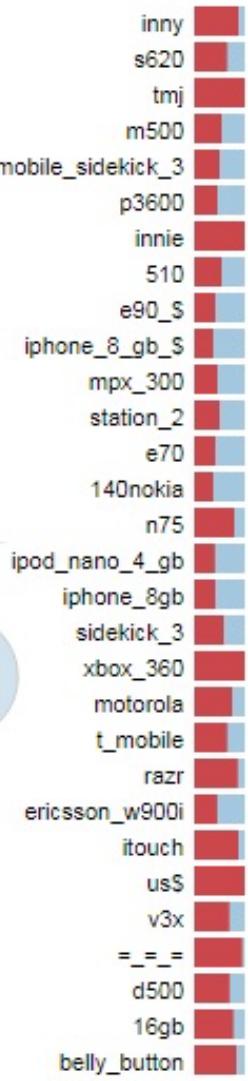
12 Ingredients



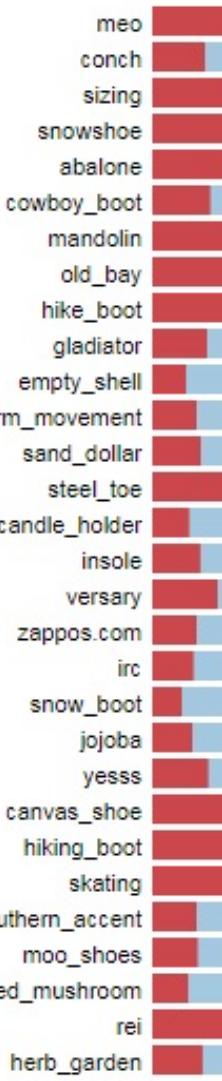
26 Supplements



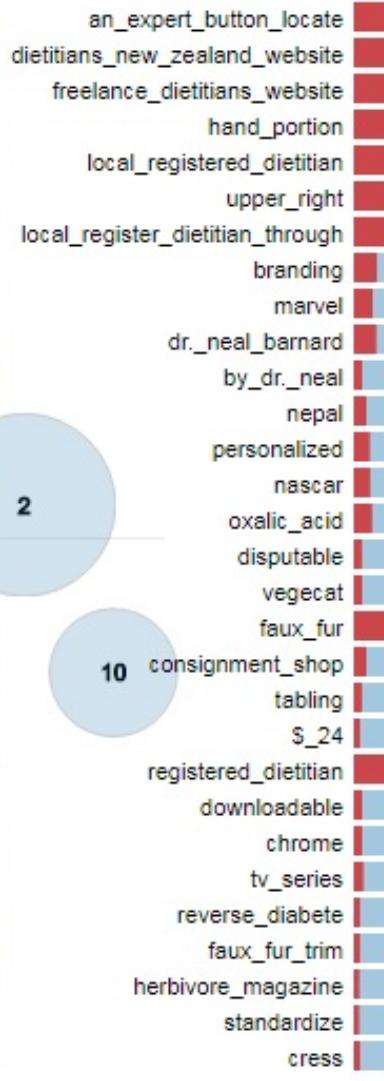
34 Electronics



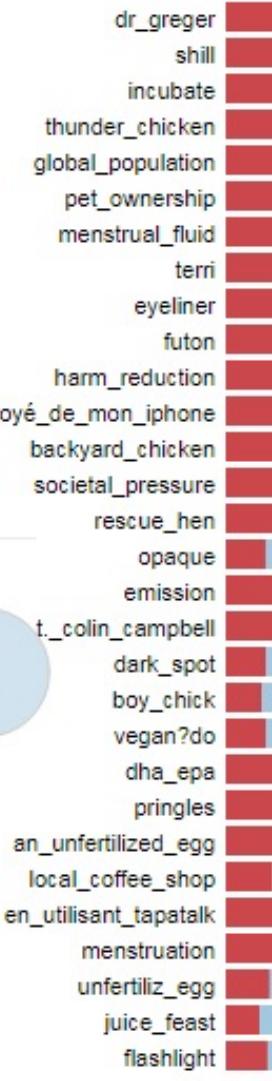
22 Shoes



21 Dieticians



19 Eggs



Примеры

- 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*"babí" + 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, графовые и т.д.
- Обучение без учителя — гораздо более широкая область