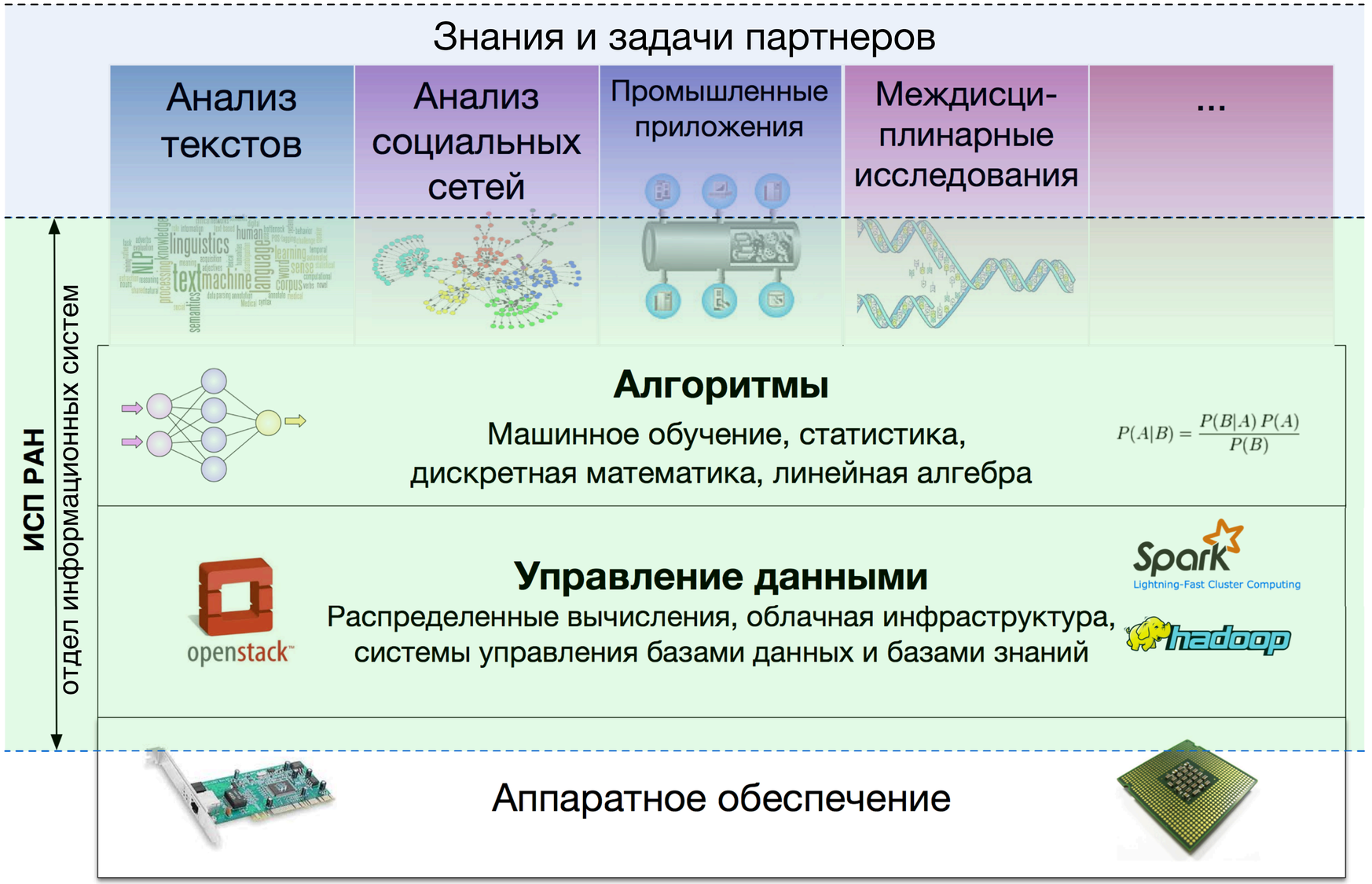
Управление данными и информационные системы

Просеминар ВМК 2016

Денис Юрьевич Турдаков

Отдел информационных систем



Обработка текстов

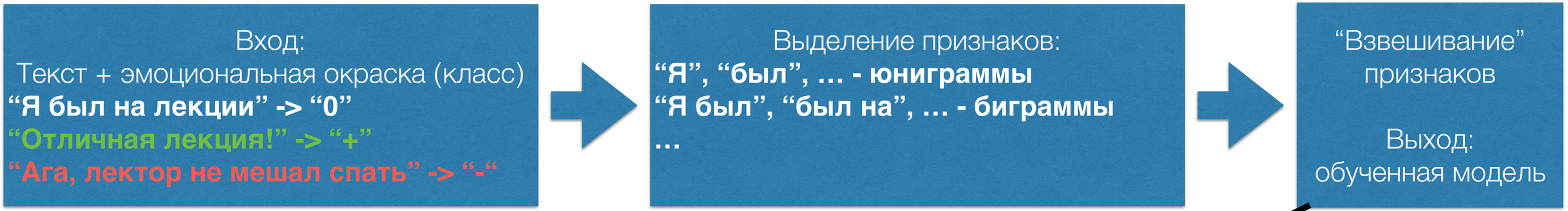
- Демонстрация: https://api.ispras.ru/demo/texterra
- Как это сделано на примере инструмента анализа эмоциональной окраски
- Где узнать подробнее про область

Анализ эмоциональной окраски

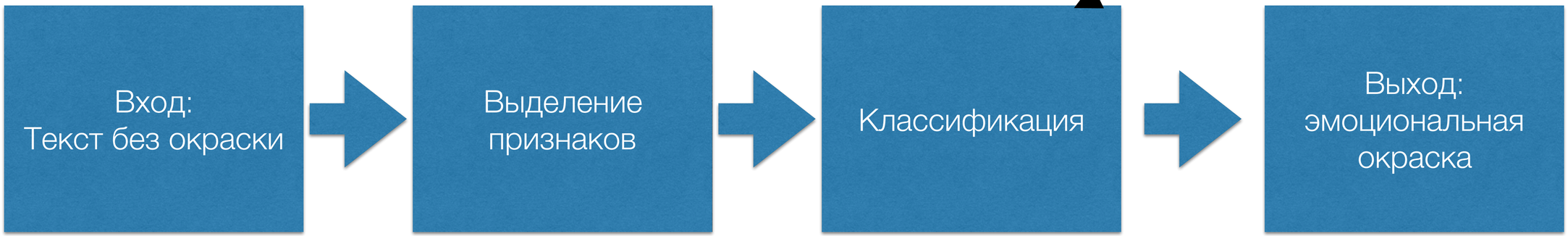
- Англ.: Sentiment analysis, Opinion mining
- Лингвистический подход:
  - описать какие эмоции несут слова,
  - написать правила определения эмоции по комбинации слов
- *Статистический подход*:
  - вывод “правил” на основе примеров
  - машинное обучение (англ. machine learning)

Машинное обучение (классификация)

Этап обучения:



Этап работы:



Наивный байесовский классификатор

- Выбор наиболее вероятного значения

$$\hat{s} = \arg\max_{s \in S} P(s|f)$$

- По правилу Байеса

$$\hat{s} = \arg\max_{s \in S} \frac{P(s)P(f|s)}{P(f)} = \arg\max_{s \in S} P(s)P(f|s)$$

- Наивное предположение об условной независимости признаков

$$\hat{s} = \arg\max_{s \in S} P(s) \prod_{j=1}^{n} P(f_i|s)$$

- Метод максимального правдоподобия
- Другими словам, просто считаем

$$P(s) = \frac{count(s)}{\sum_{s \in S} count(s_i)} \quad P(f_j|s) = \frac{count(f_j, s)}{count(s)}$$

- Алгоритм прост в реализации, но
  - Исчезновение значащих цифр → использовать сумму логарифмов вместо произведения
  - Нулевые вероятность → сглаживание или предположение о распределении $P(f_j|s)$

Где узнать подробнее

- https://www.coursera.org/learn/machine-learning
- https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie
- http://scikit-learn.org/
- Спецкурс “Основы обработки текстов”: http://tpc.at.ispras.ru

Анализ социальных сетей

- Демонстрация: http://egozoom.at.ispras.ru
- Как это сделано: алгоритм кластеризации
- Где узнать подробнее

K-средних

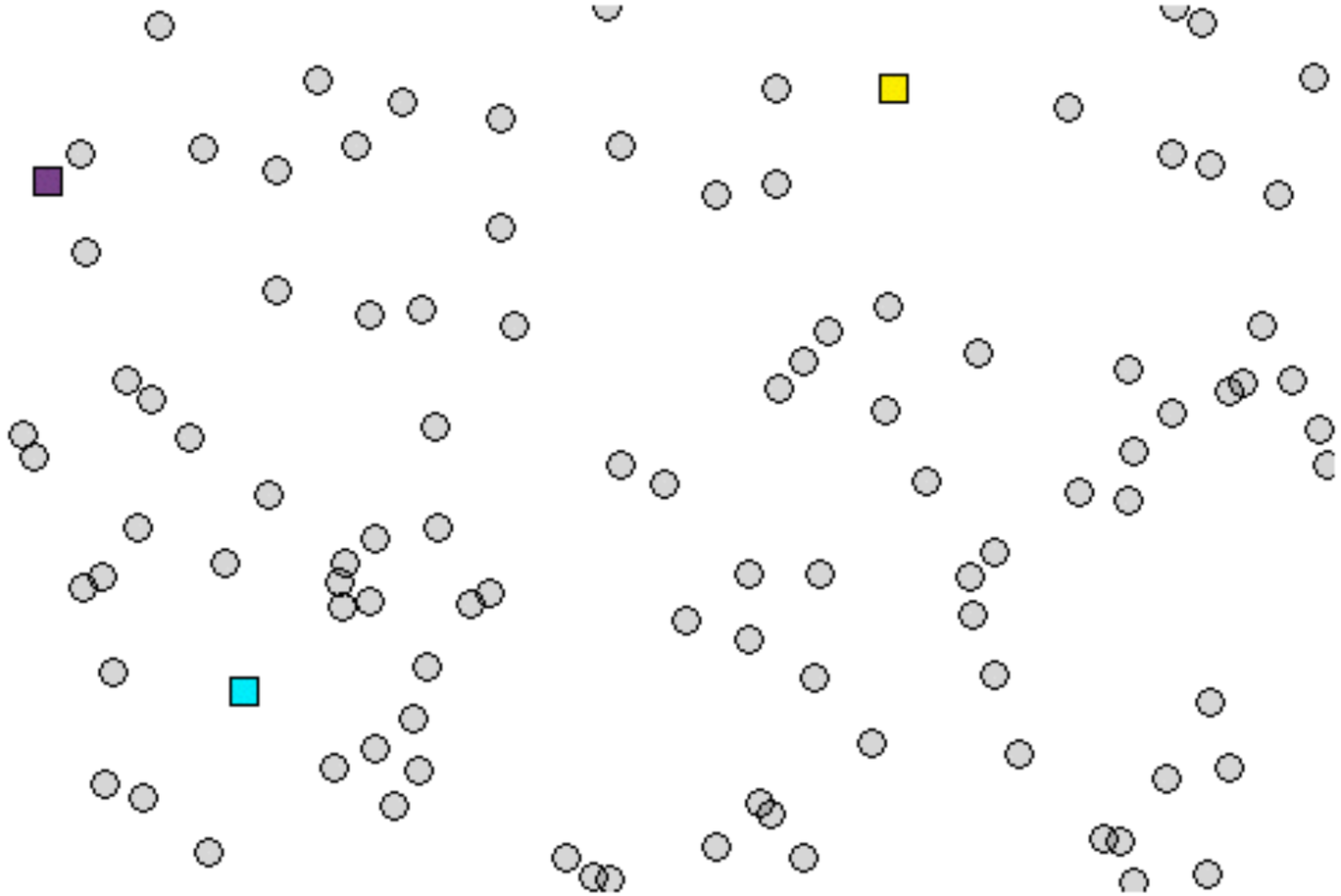
- Алгоритм машинного обучения без учителя
- Алгоритм k-means разбивает данные на k кластеров
  - –Каждый кластер имеет центр - центроид
  - –Параметр k - задается вручную
- Алгоритм
  1. Выбираются k точек в качестве начальных центроидов
  2. Сопоставить каждой точке ближайший центроид
  3. Пересчитать центроиды
  4. Если алгоритм не сошелся перейти на шаг 2

Критерий останова

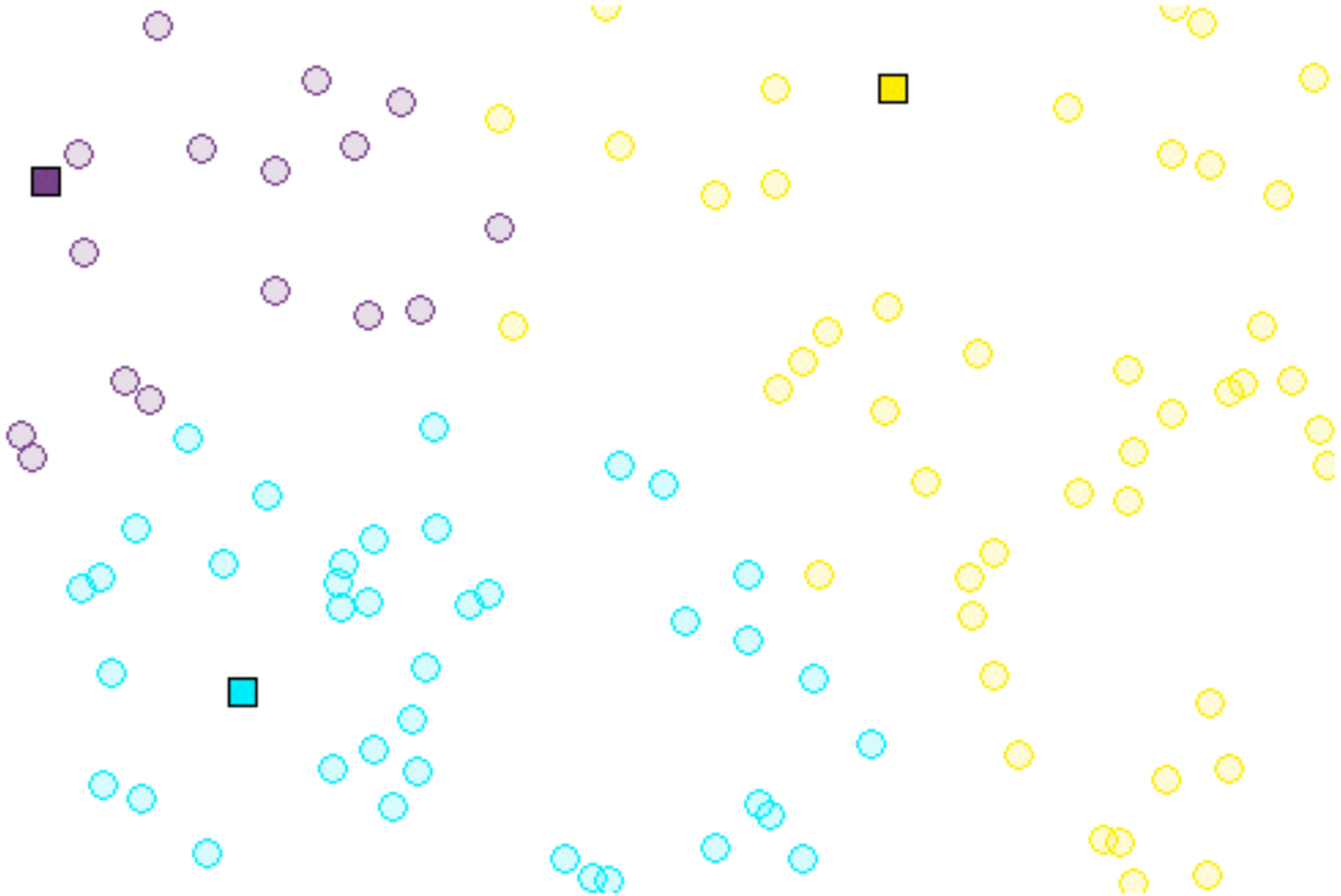
- Нет перехода точек в другой кластер
- Нет (незначительно) изменение центроидов
- Мало убывает погрешность (sum of squared error)

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} dist(x, m_j)^2$$

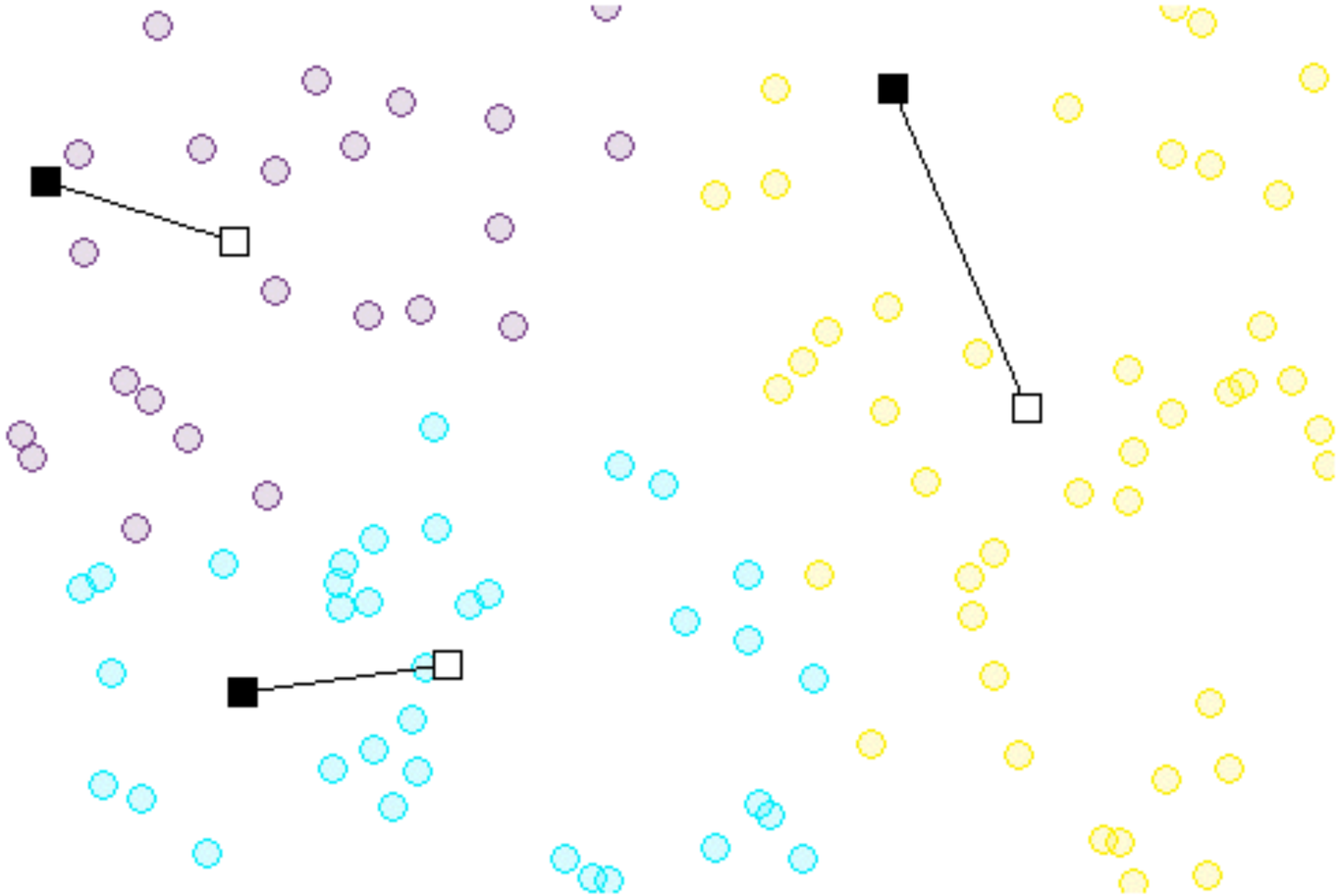
K-средних. Пример



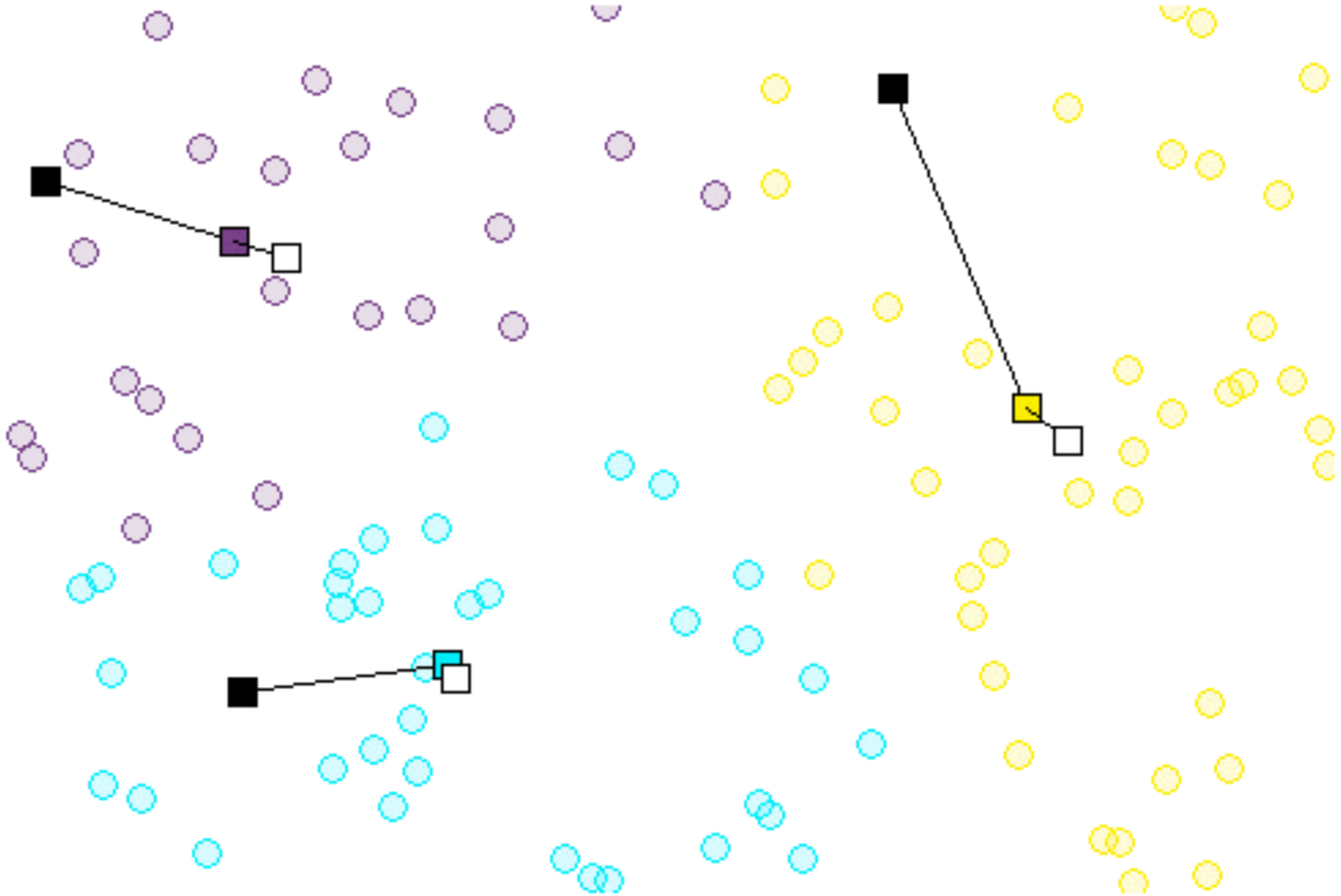
K-средних. Пример



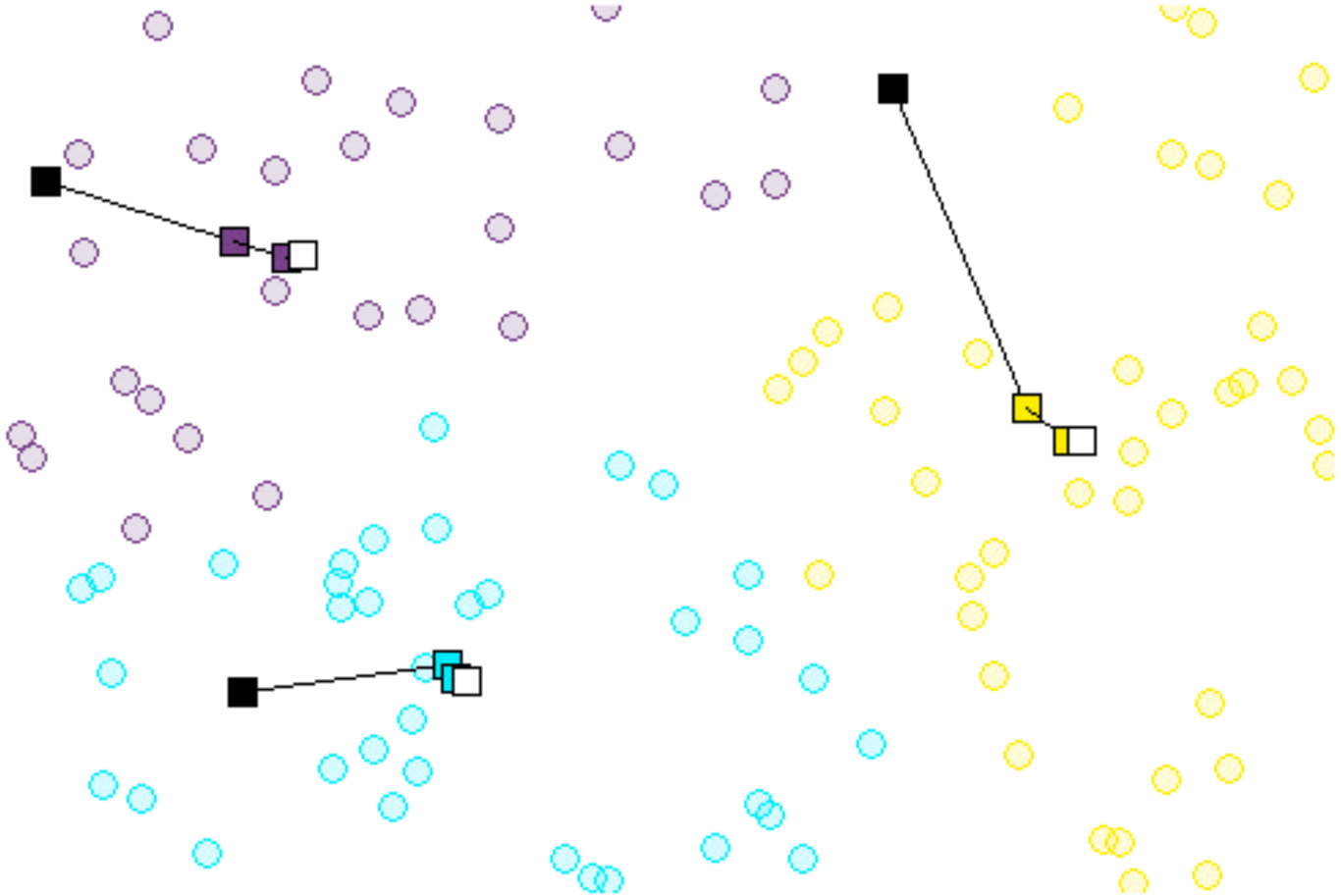
K-средних. Пример



K-средних. Пример



K-средних. Пример



Проблемы

- Алгоритм чувствителен к начальному выбору центроидов
  - запуск с различной начальной инициализацией и выбор варианта с наиболее плотными кластерами
- Чувствителен к выбросам
  - можно фильтровать выбросы
- Не подходит для нахождения кластеров, не являющихся элипсоидами
  - преобразование пространства

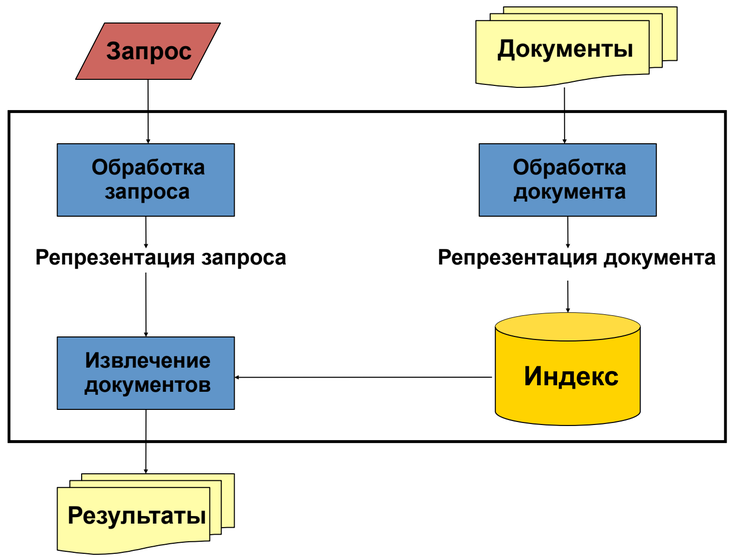
Где узнать подробнее

- Charu C. Aggarwal. 2011. Social Network Data Analytics (1st ed.). Springer Publishing Company, Incorporated.
- http://scikit-learn.org/
- Barabási, Albert-László. Linked: The New Science of Networks. Perseus Books Group.

Информационный поиск

- Демонстрация: http://blognoon.com
- Как это работает?
- Где узнать подробнее

Как это работает?



Инвертированный индекс

Doc 1

This is a sample document with one sample sentence

Doc 2

This is another sample document

Term	# docs	Total freq
This	2	2
is	2	2
sample	2	3
another	1	1
…	…	…

Doc id	Freq
1	1
2	1
1	1
2	1
1	2
2	1
2	1
…	…
…	…

Векторная модель

Пространство документов	t_1	t_2	t_3	…	t_n
D_1	a_{11}	a_{12}	a_{13}	…	a_{1n}
D_2	a_{21}	a_{22}	a_{23}	…	a_{2n}
D_3	a_{31}	a_{32}	a_{33}	…	a_{3n}
…					
D_m	a_{m1}	a_{m2}	a_{m3}	…	a_{mn}
Q	b_1	b_2	b_3	…	b_n

Пространство терминов

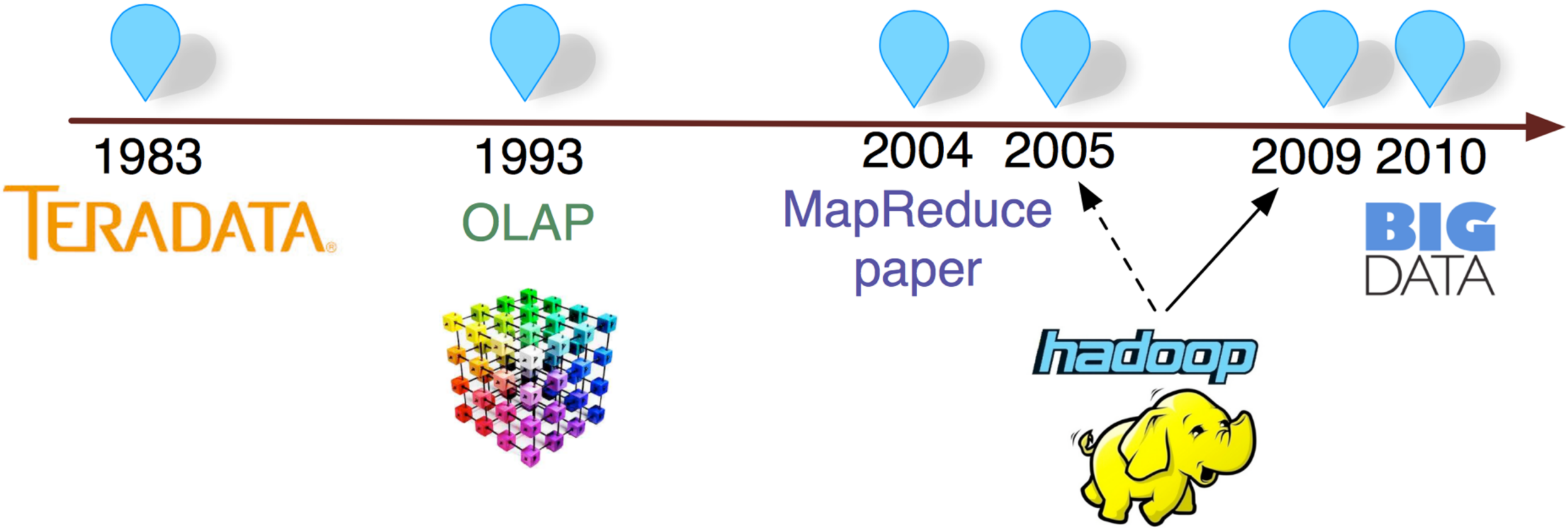
$$Sim(D,Q) = \sum (a_i * b_i)$$

Где узнать подробнее

- Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
- ElasticSearch (+Kibana): https://www.elastic.co
- Стюарт Рассел, Питер Норвиг, ”Искусственный интеллект: современный подход”. (вообще про современное состояние работ в области искусственного интеллекта)

Обработка больших объемов данных

- Парадигма MapReduce



Что изучить

- Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

- Apache Spark
  - http://spark.apache.org
  - https://databricks.com/spark/developer-resources

- Matei Zaharia, et.al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. (NSDI'12).

План действий

- Чтобы попасть на кафедру:
  - Написать кафедру первым приоритетом при выборе кафедры
  - На обратной стороне написать, что вы ходили на просеминар и на какой спец. семинар вы хотите попасть

- Чтобы попасть на спец. семинар:
  - http://seminar.at.ispras.ru
  - Посмотреть курсовые работы предыдущих лет
  - Определиться с направлением
  - Написать письмо с пожеланиями на turdakov@ispras.ru
  - Изучить информацию по ссылкам в этой презентации