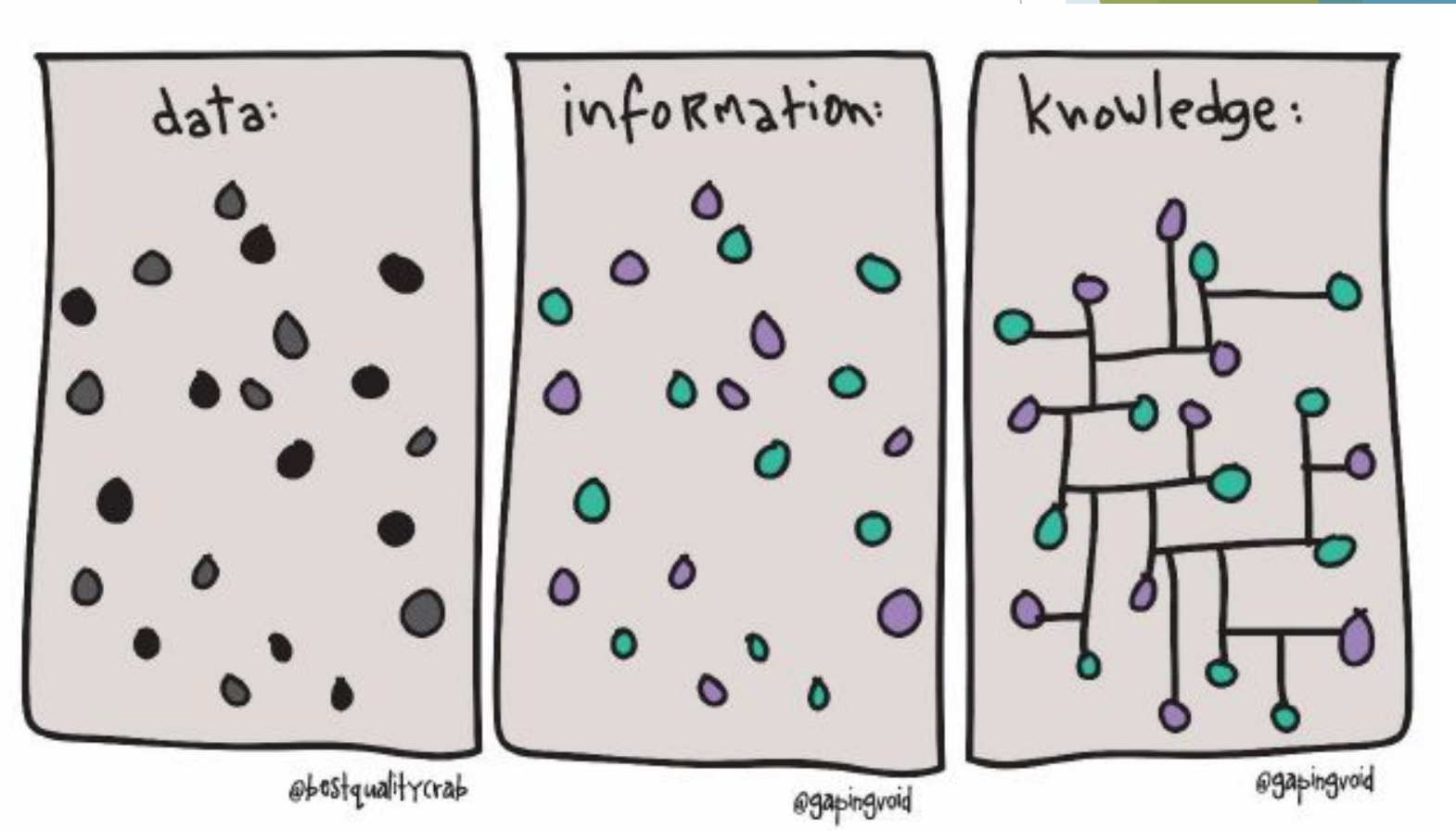
# Основы анализа больших данных

# Задачи

# Основная задача

- нахождение полезных закономерностей в массиве данных

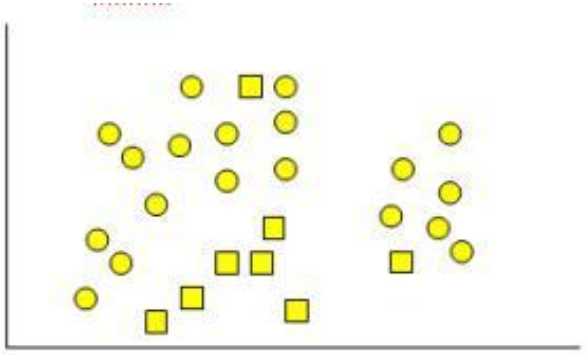


# Задачи Data Mining

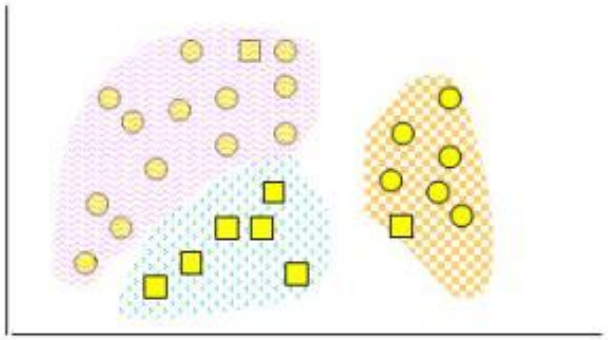
- по виду искомых закономерностей
  - Классификация
  - Кластеризация
  - Прогнозирование
  - Ассоциация
  - Визуализация
  - др.

- Единого мнения относительно того, какие задачи следует относить к *Data* Mining, нет

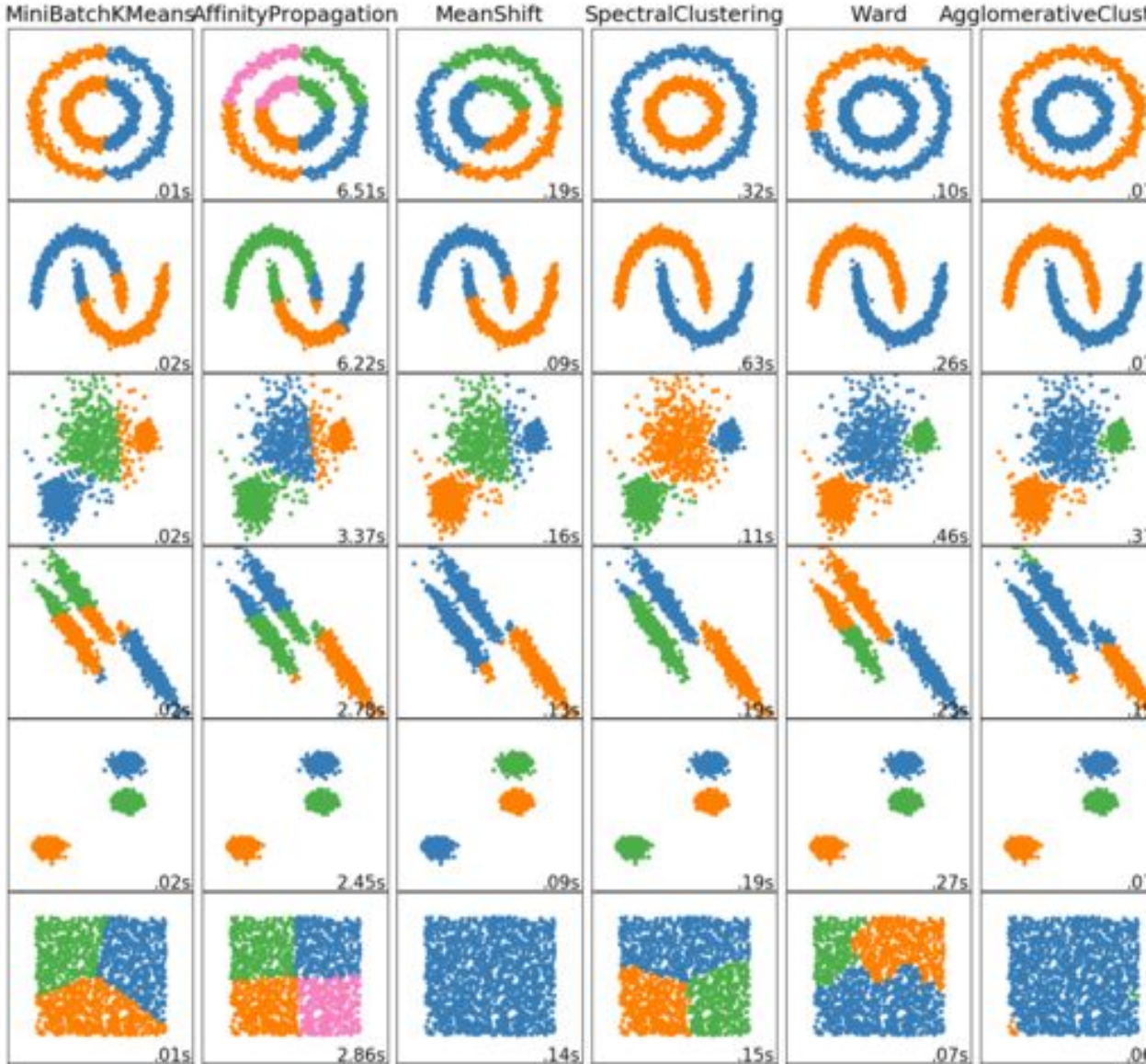
# Классификация и кластеризация



*Классификация: классы предопределены изначально*



*Кластеризация: классы не предопределены, осуществляется поиск наиболее похожих, однородных групп*



# Ассоциация (Association) поиск ассоциативных правил

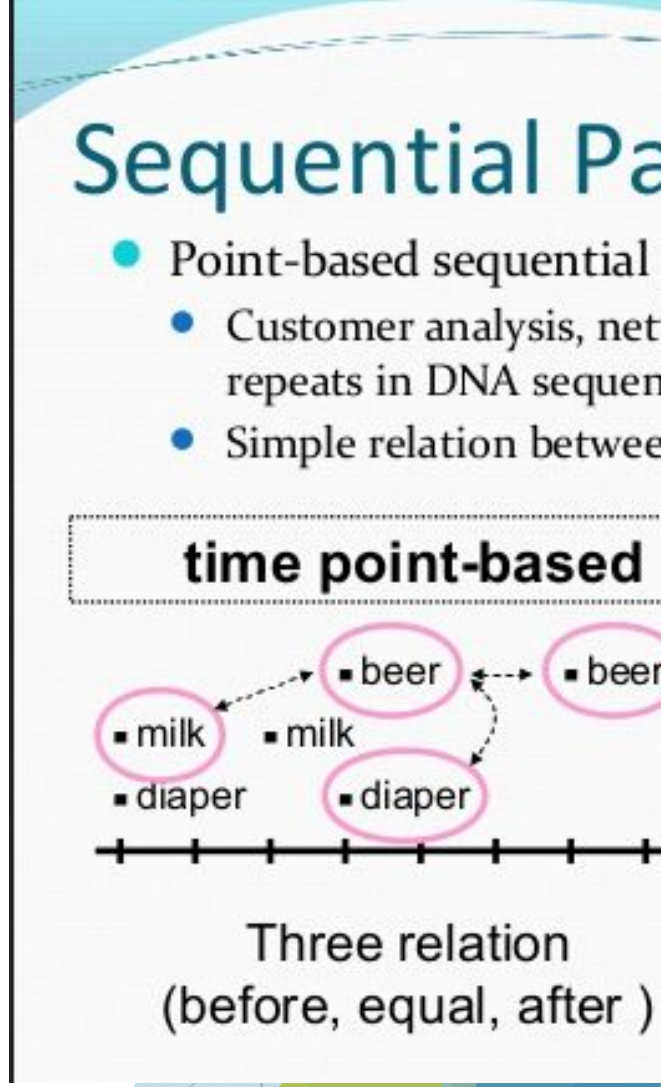
- нахождение закономерностей между связанными одновременными событиями в наборе данных без учета свойств самих объектов

- Пример
  - На основе анализа поведения пользователя в сети интернет можно предсказать степень его интереса к определённой тематике

# Последовательность (Sequence)
## последовательная ассоциация (sequential association)

- нахождение закономерностей между связанными неодновременными событиями в наборе данных без учета свойств самих объектов
- Ищется наибольшая вероятность цепочки связанных во времени событий

- Пример
  - На основе анализа последовательности просмотренных пользователем сайтов в сети интернет можно предсказать вероятность выбора следующего сайта

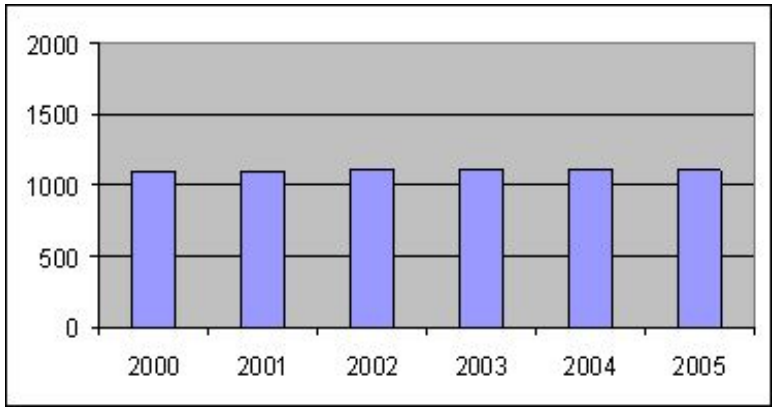


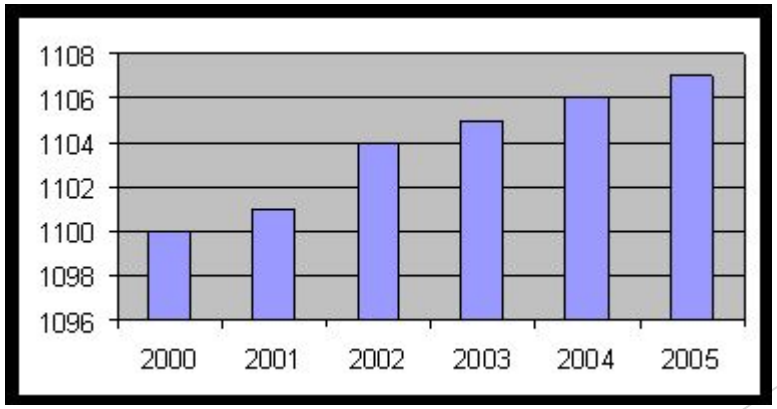
# Прогнозирование

- ***Прогнозирование*** (от греческого Prognosis), в широком понимании этого слова, определяется как опережающее отражение будущего.
  Целью *прогнозирования* является предсказание будущих событий.
- *Прогнозирование* направлено на *определение* тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем.

# Визуализация

- Позволяет перейти от символов к образам
  - линия *тренда* или скопления точек на диаграмме рассеивания позволяет аналитику намного быстрее определить закономерности и прийти к нужному решению
- Может ввести в заблуждение
  - Хорошая визуализация
  - Плохая визуализация



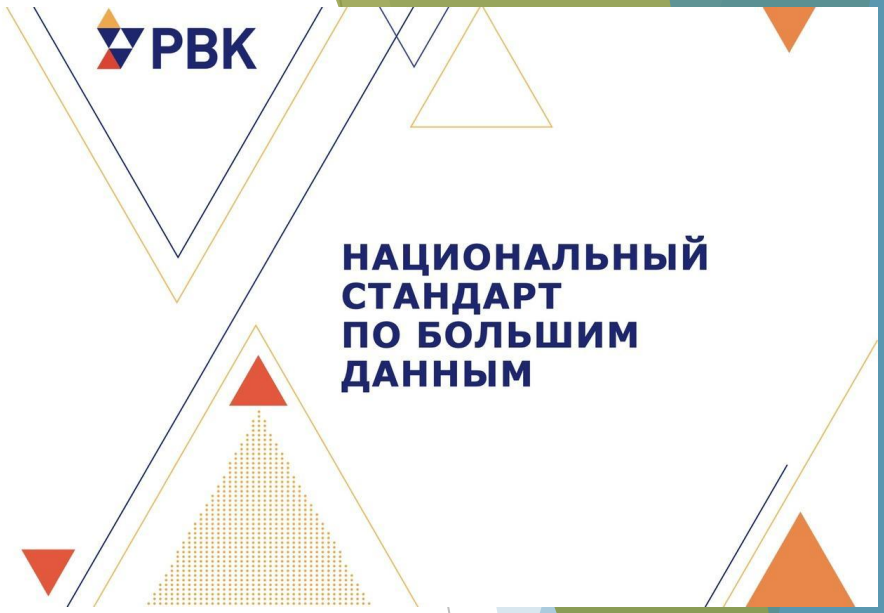
# Стандарты в области больших данных

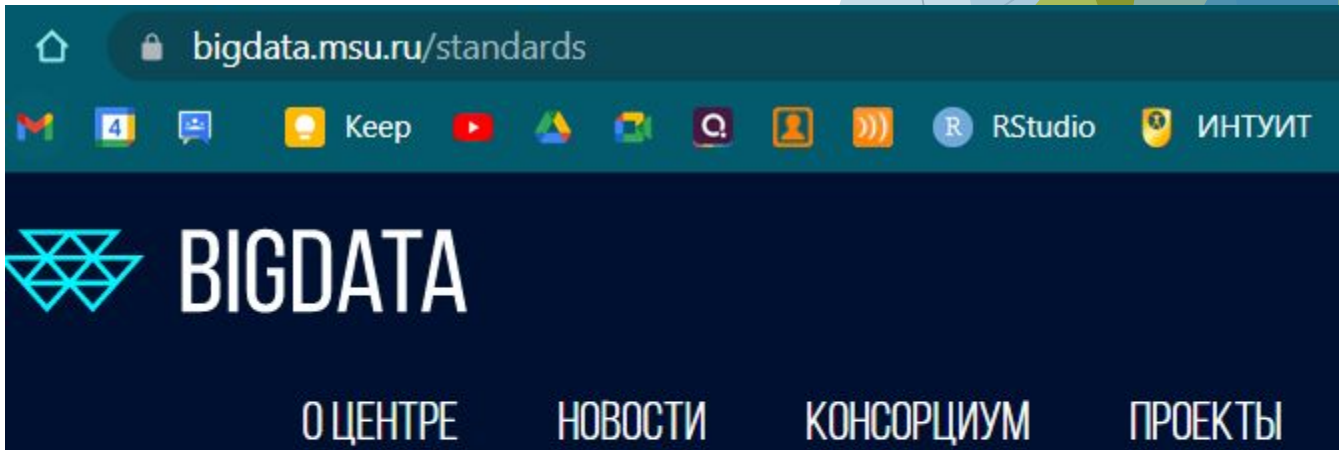
# Международные стандарты ИСО/МЭК

- ISO/IEC 20546:2019 Information technology – Big data – Overview and vocabulary
- ISO/IEC TR 20547-1:2020 Information technology – Big data reference architecture – Part 1: Framework and application process
- ISO/IEC WD 5259-1 Data quality for analytics and ML – Part 1: Overview, terminology, and examples

# Национальные стандарты

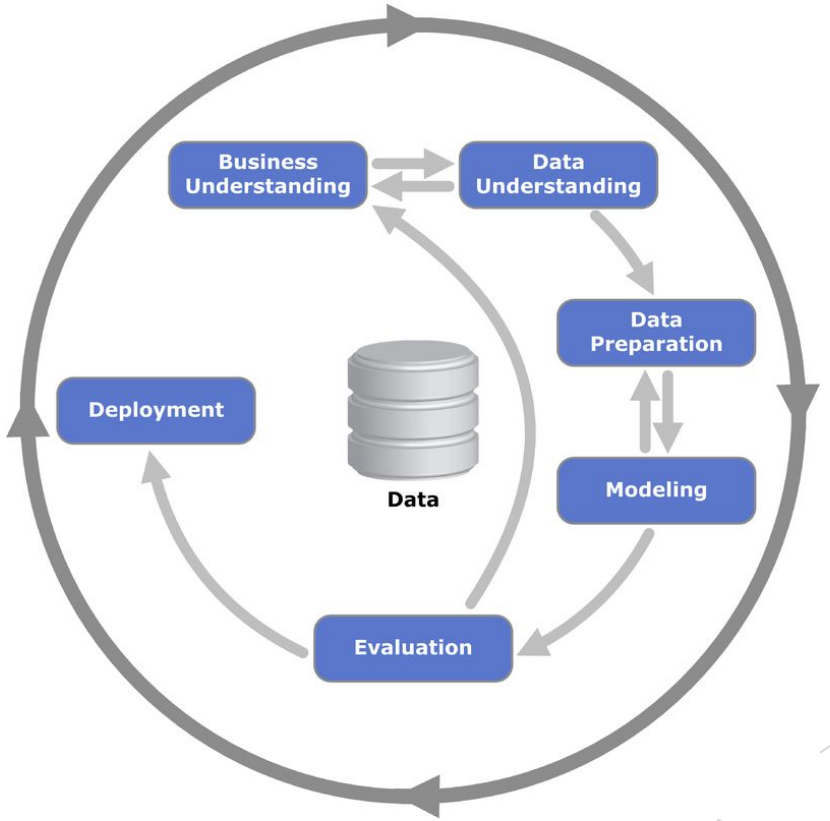
- Публичное обсуждение стандарта о направлениях стандартизации больших данных
-



# Межотраслевые / корпоративные стандарты

- CRISP-DM (*Cross-Industry Standard Process for Data Mining*) — наиболее распространённая методология по исследованию данных.



# A Guide On How To Become A Data Scientist (Step By Step Approach)


- STEP 1: Choose A Programming Language (Python / R)
- STEP 2. Statistics
- STEP 3: Learn SQL
- STEP 4. Data Cleaning
- STEP 5: Exploratory Data Analysis
- STEP 6: Learn Machine Learning Algorithms

# CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

## SUPERVISED

Predict a category

### CLASSIFICATION

«Divide the socks by color»

Predict a number

### REGRESSION

«Divide the ties by length»

Data is not labeled in any way

## UNSUPERVISED

Divide by similarity

### CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

### DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Identify sequences

### ASSOCIATION

«Find what clothes I often wear together»