



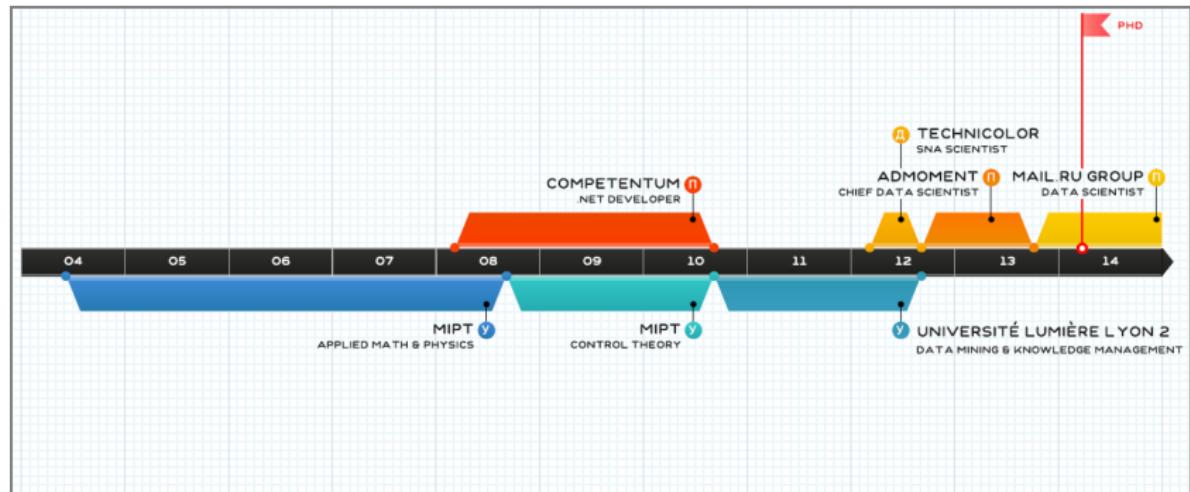
# ТЕХНОСФЕРА

## Лекция 1 Задачи Data Mining

Николай Анохин

28 февраля 2015 г.

# Николай Анохин



e-mail: n.anokhin@corp.mail.ru

тел.: +7 (903) 111-44-60

# План лекции

Структура курса

Что такое Data Mining

Унификация процесса Data Mining

Exploratory data analysis

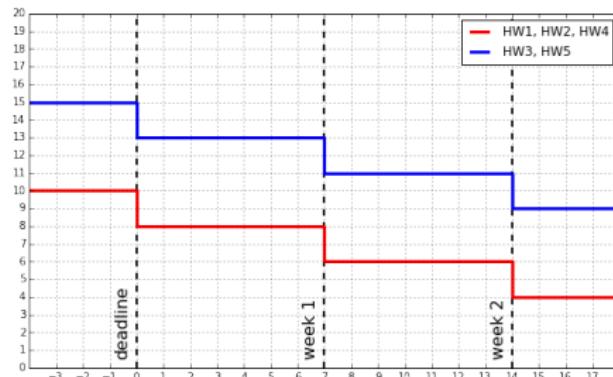
# Структура курса

1. Задачи Data Mining<sup>HW1</sup>
2. Задача кластеризации и EM-алгоритм<sup>HW2</sup>
3. Алгоритмы кластеризации<sup>HW3</sup>
4. Визуализация результатов кластеризации
5. Кластеризация: итоговое занятие<sup>K1</sup>
6. Задача классификации
7. Naïve Bayes и работа с текстом<sup>HW4</sup>
8. Решающие деревья
9. Линейные модели<sup>HW5</sup>
10. Метод опорных векторов
11. Data Mining в реальных системах
12. Классификация: итоговое занятие<sup>K2</sup>
13. Защита семестрового проекта

# Контроль знаний

## ДЗ

- ▶ ДЗ-1, ДЗ-2, ДЗ-4 : максимум 10 баллов за каждое
- ▶ ДЗ-3, ДЗ-5: максимум 15 баллов за каждое



## Теория

- ▶ К-1, К-2 : максимум 10 баллов за каждый
- ▶ Защита проекта: максимум 20 баллов

# Шкала оценок



(a) 0 — 49



(b) 50 — 79



(c) 80 — 94



(d) 95 — 100

# Правила

- + Можно задавать вопросы по ходу лекции
- + Можно входить и выходить, не мешая коллегам
- Нельзя нарушать порядок в аудитории
- Нельзя разговаривать по телефону
- ▶ Общение с преподавателем на “Вы”

Ваши правила?

# DM как KDD

## Data Mining

Процесс извлечения знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т.д. Полученные знания должны быть достоверными, полезными и интерпретируемыми.

# DM как моделирование

## Data Mining

Процесс построения модели, хорошо описывающей закономерности, которые порождают данные.

Подходы к построению моделей

- ✓ статистический
- ✓ на основании машинного обучения
- ✗ вычислительный

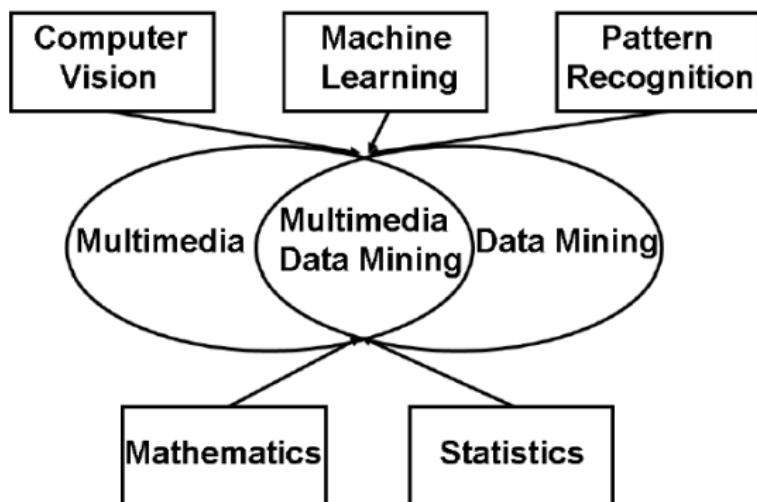
## Data Scientist

Person who is better at statistics than any software engineer and better at software engineering than any statistician  
(J. Wills, Data Scientist at Cloudera Inc.)

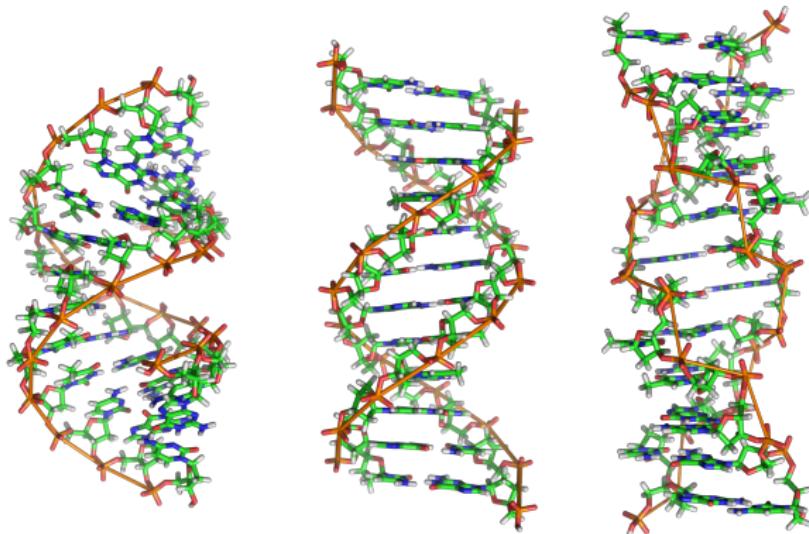
Data-

- ✖ -architecture
- ✖ -acquisition
- ✓ -analysis
- ✖ -archiving

# Multimedia Data Mining



# Data Mining в медицине и биологии



# Data Mining в финансовой сфере

- ▶ Предсказание биржевых цен
- ▶ Предсказание курса валют
- ▶ Работа с рисками и банкротством
- ▶ Кредитный scoring
- ▶ Выявление мошенников

# Data Mining для CRM и целевого маркетинга

CRM = Customer Relationship Management

- ▶ Как определить, кто собирается уйти?
- ▶ Какие продукты предложить клиенту?
- ▶ Как найти новых клиентов?
- ▶ Реклама!

# Data Mining для Высшего блага (на самом деле fail)

- ▶ Наблюдаем  $10^9$  человек
- ▶ Человек в среднем посещает отель раз в 100 дней
- ▶ Есть  $10^5$  отелей на 100 человек каждый
- ▶ Проверим посещения за 1000 дней

Вероятность для конкретной пары встретиться в отеле в конкретный день:

$$p_1 = \left( \frac{1}{100} \right)^2 \cdot 10^{-5} = 10^{-9}$$

Всего пар людей

$$n_{pp} = C_2^{10^9} \approx \frac{(10^9)^2}{2} = 5 \cdot 10^{17}$$

а пар дней

$$n_{pd} = C_2^{10^3} \approx \frac{(10^3)^2}{2} = 5 \cdot 10^5$$

Ожидаемое количество “подозрительных” встреч в отелях

$$N = p_1^2 n_{pp} n_{pd} = 250000 >> 10$$

## Принцип Бонферрони

Вычислить количество рассматриваемых событий при предположении их полной случайности. Если это количество намного превосходит количество событий, о котором идет речь в задаче, полученные результаты нельзя будет считать достоверными.

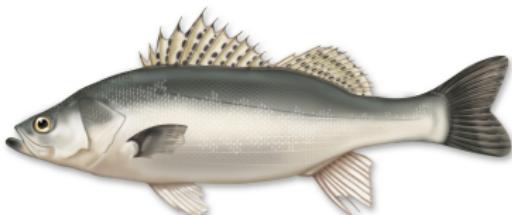
# Cross Industry Standard Process for Data Mining

## CRISP-DM

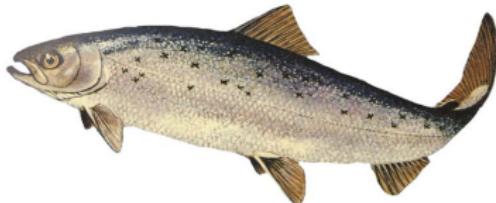
- ▶ IBM/SPSS
- ▶ Teradata
- ▶ Daimler AG
- ▶ NCR Corporation
- ▶ OHRA



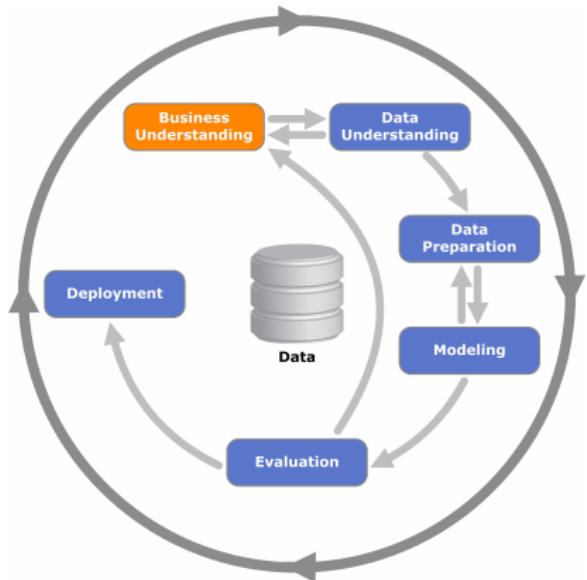
Задача: на рыболовном  
предприятии автоматизировать  
сортировку улова



(a) Сибас



(b) Лосось



# Признаки

$\mathcal{D}$  – множество объектов (data set)

$d \in \mathcal{D}$  – обучающий объект

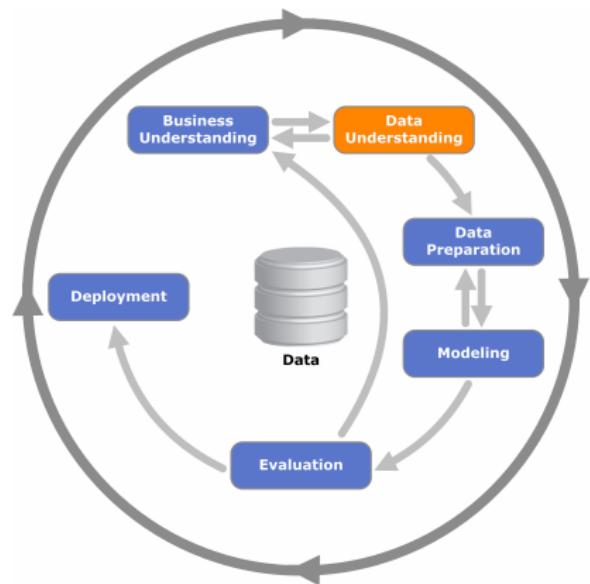
$\phi_j : D \rightarrow F_j$  – признак

Виды признаков

- ▶ Бинарные/Binary  
 $F_j = \{ \text{true}, \text{false} \}$
- ▶ Номинальные/Categorical  
 $F_j$  – конечно
- ▶ Порядковый/Ordinal  
 $F_j$  – конечно, упорядочено
- ▶ Количественный/Numerical  
 $F_j = \mathbb{R}$

Представление  $d_i$ :

$$\mathbf{x}_i = (\phi_1(d_i), \dots, \phi_n(d_i)) \in \mathcal{X}$$

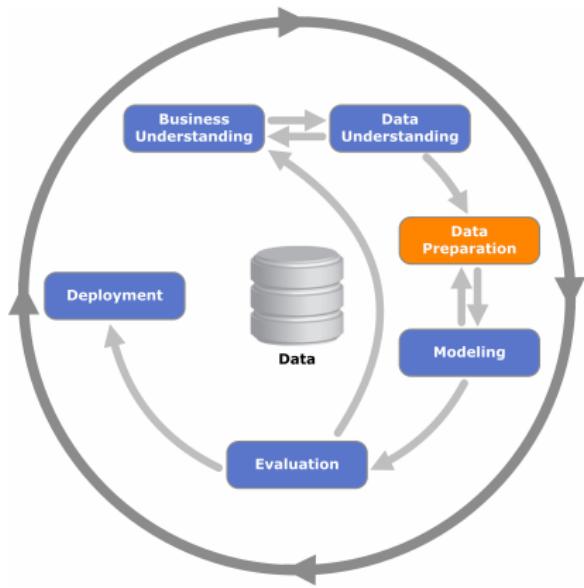


- ▶ Удаление шума
- ▶ Заполнение отсутствующих значений
- ▶ Трансформация факторов
- ▶ Выбор факторов
- ▶ Использование априорных знаний

В итоге:

$$X = \begin{pmatrix} x_1 \\ \dots \\ x_N \end{pmatrix}, \quad x_i \in \mathcal{X}$$

$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_N \end{pmatrix}, \quad y_i \in \mathcal{Y}$$



## Модель

семейство параметрических функций вида

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}\}$$

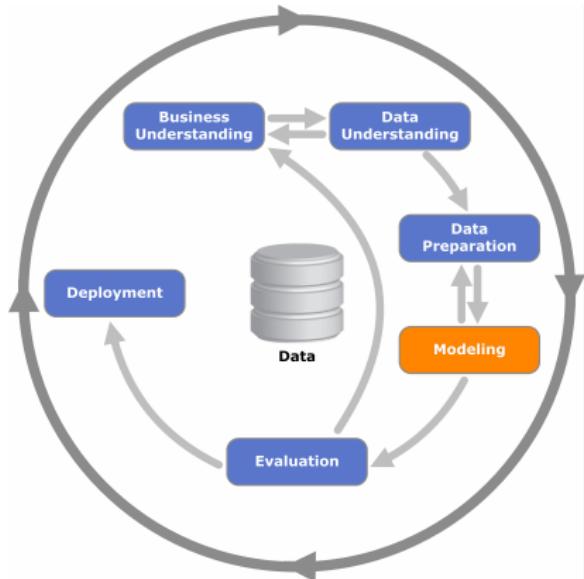
## Алгоритм обучения

выбор наилучших параметров  $\theta^*$

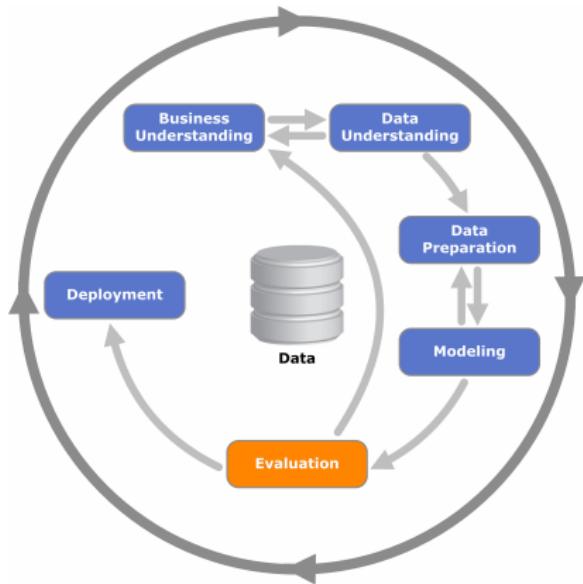
$$A(X, Y) : (\mathcal{X} \times \mathcal{Y})^N \rightarrow \Theta$$

В итоге:

$$h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$$

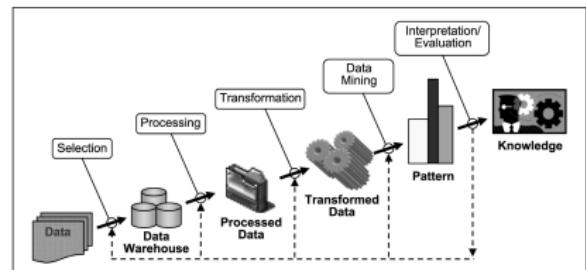
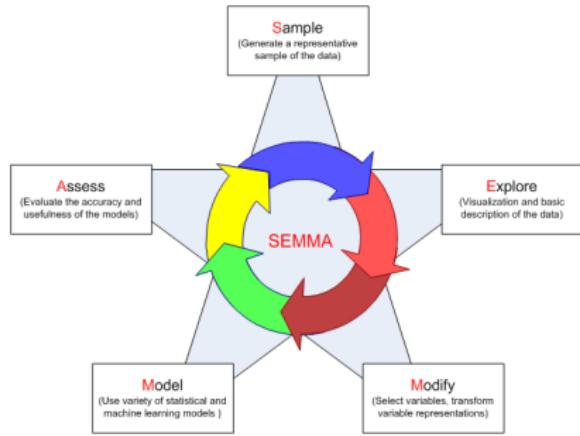


- ▶ точность или аппроксимация?
- ▶ bias или variance?
- ▶ интерпретируемость или качество?





# Другие процессы: SEMMA, KDD



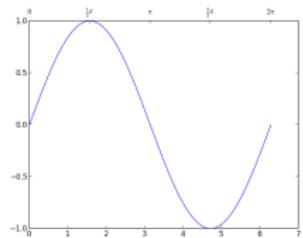
# Exploratory data analysis

EDA направлен на предварительное изучение данных

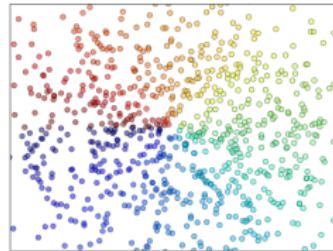
- ▶ формирование гипотез относительно структуры данных
- ▶ выбор необходимых инструментов анализа

Особенность метода состоит в визуализации и поиске важных характеристик и тенденций

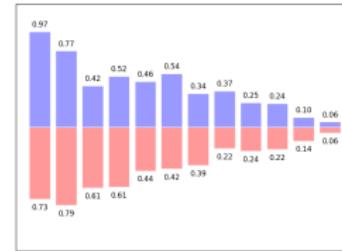
# Примеры



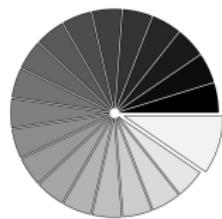
(a) Plot



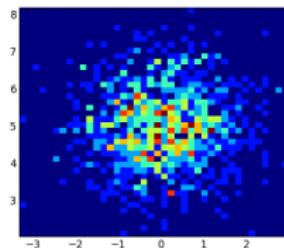
(b) Scatter



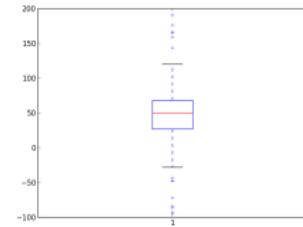
(c) Barplot



(d) Piechart



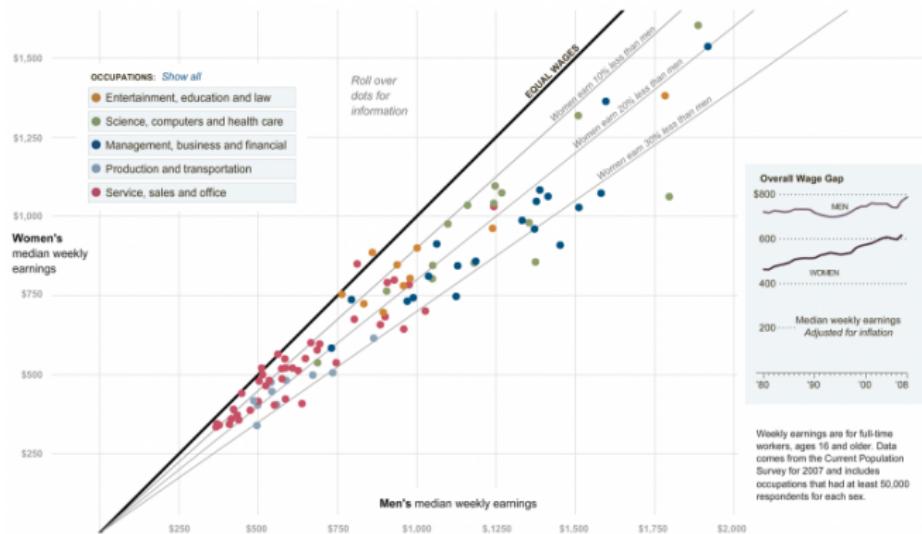
(e) Heatmap



(f) Boxplot

# Полезные советы

- ▶ Все познается в сравнении
  - ▶ Причинно-следственные связи
  - ▶ Размерность имеет значение  
(больше-лучше)
- ▶ Не избегать пояснений
  - ▶ Content is king



## Вопросы

