Introduction to Data Science

Lecture 1. General Introduction

30.09.2021

Maxim Panov, Mikhail Belyaev

Skoltech
Skolkovo Institute of Science and Technology

Outline

- **Info about the course**
- Examples of data science problems from everyday life
- Major data analysis problem statements

Learning Outcomes

Know

1. Statements of all major machine learning problems.
2. Some mathematical details of the most important data analysis methods and algorithms.

Be able

1. Select an appropriate method for solving particular data analysis problems.
2. Perform basic data processing and visual analysis, generate features for subsequent machine learning.
3. Apply machine learning libraries, select algorithm's hyperparameters.
4. Critically evaluate the obtained results and redesign data-processing pipelines.
5. Solve real-world data science problems using modern machine learning techniques.

Course Structure

Lecture 1 (MP). General Introduction
Lecture 2 (MP). Regression, Quality Metrics, Cross-validation
Lecture 3 (MB). Model Selection
Lecture 4 (MB). Classification
Lecture 5 (MP). Decision Trees & Ensembling
Lecture 6 (MB). Features Engineering & Selection
Lecture 7 (MB). Dimensionality Reduction
Lecture 8 (MP). Clustering

Assignements and Project

Assignments

Homework 1 (2nd week, 20%)

Exploratory analysis, regression and cross-validation:

- basic data manipulations;
- visual analysis of the dataset;
- basic machine learning experiments.

Homework 2 (3rd week, 30%)

Feature engineering and classification:

- machine learning pipeline setup;
- basic machine learning experiments (i.e. selection of algorithms parameters);
- advanced machine learning experiments (i.e. generation of features).

Final project (50%)

A real-life problem from https://www.kaggle.com/competitions
More details will be available soon via Canvas

Course Logistics

- **Lectures**
- Pre-recorded;
- Discussed in class in large groups (40-70 people).

- **Practical seminars**
- Pre-recorded;
- Discussed in class in smaller groups (up to 30 people).

- **Contact us**
- Piazza for discussions https://piazza.com/skoltech.ru/fall2021/ma030111/home;
- Canvas for main announcements;
- Telegram channel for rapid information https://t.me/ds_intro;
- Office hours with TAs (online via Telegram)

Course Instructors

- **Mikhail Belyaev, m.belyaev@skoltech.ru**
- **Maxim Panov, m.panov@skoltech.ru**

Who we are (both MB & MP)

- **Assistant Professors at Skoltech**
- **PhD in Data Science (Candidate of Science in Math)**
- **Have 5+ years experience of work as Data Scientists at Datadvance company:**
  - **Developed machine learning algorithms in the context of an industrial data analysis library intended mainly for aerospace and automotive**
  - **Solved a set of data analysis problem from Airbus, Astrium, Areva, Eurocopter, Force India F1 and many others**

Skoltech
Skolkovo Institute of Science and Technology

Teaching Assistants

Alexander Artemenkov

Kirill Fedyanin

Bogdan Kirillov

Nikita Kotelevskii

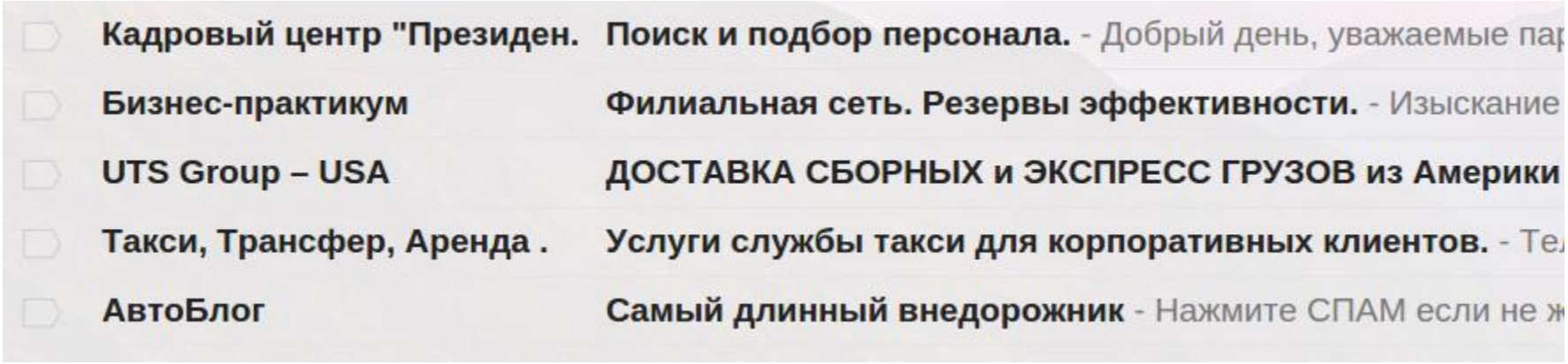
Alexander Rubashevsky

Anvar Kurmukov

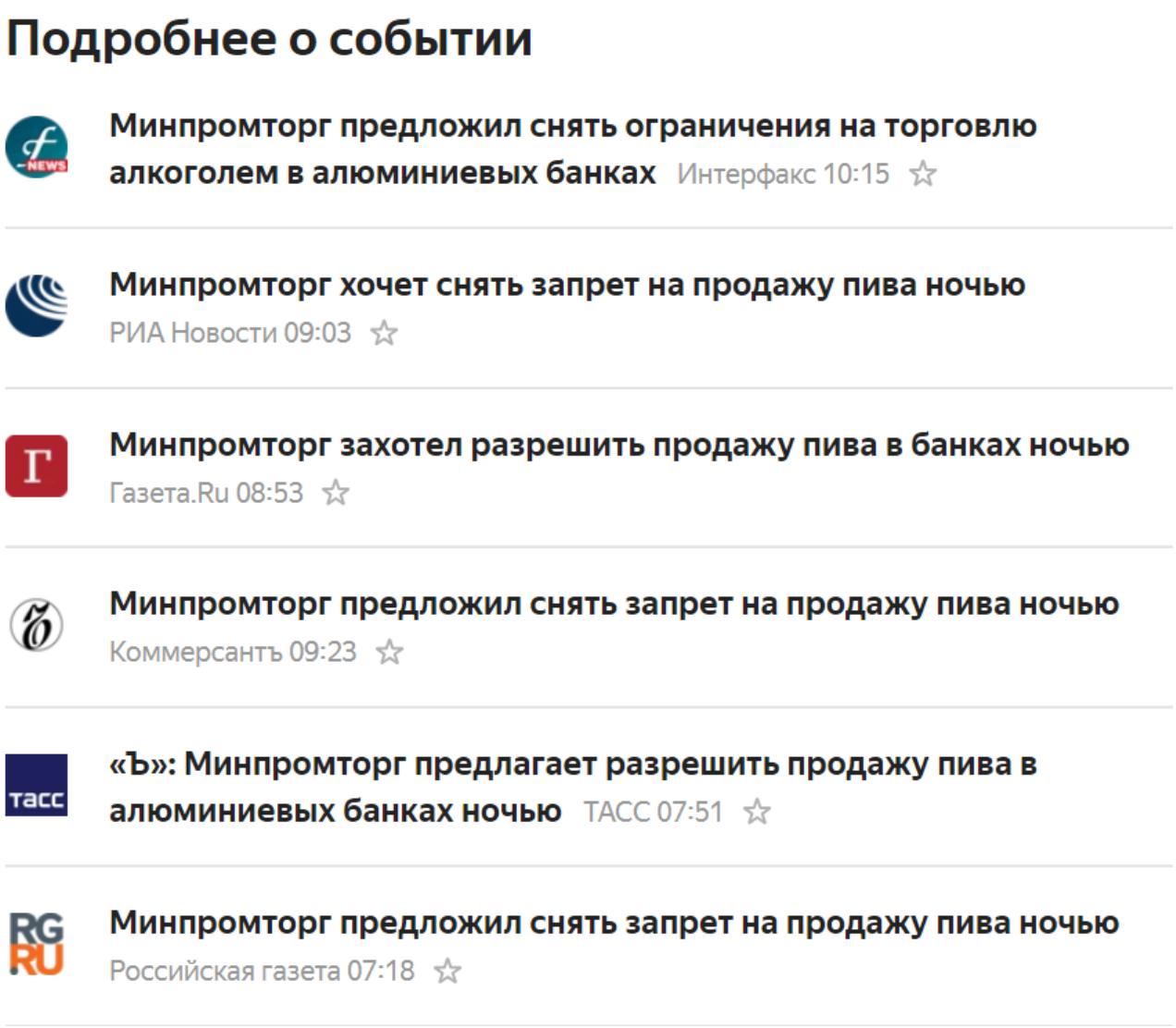
Outline

- Info about the course
- **Examples of data science problems from everyday life**
- Major data analysis problem statements

Spam filtering

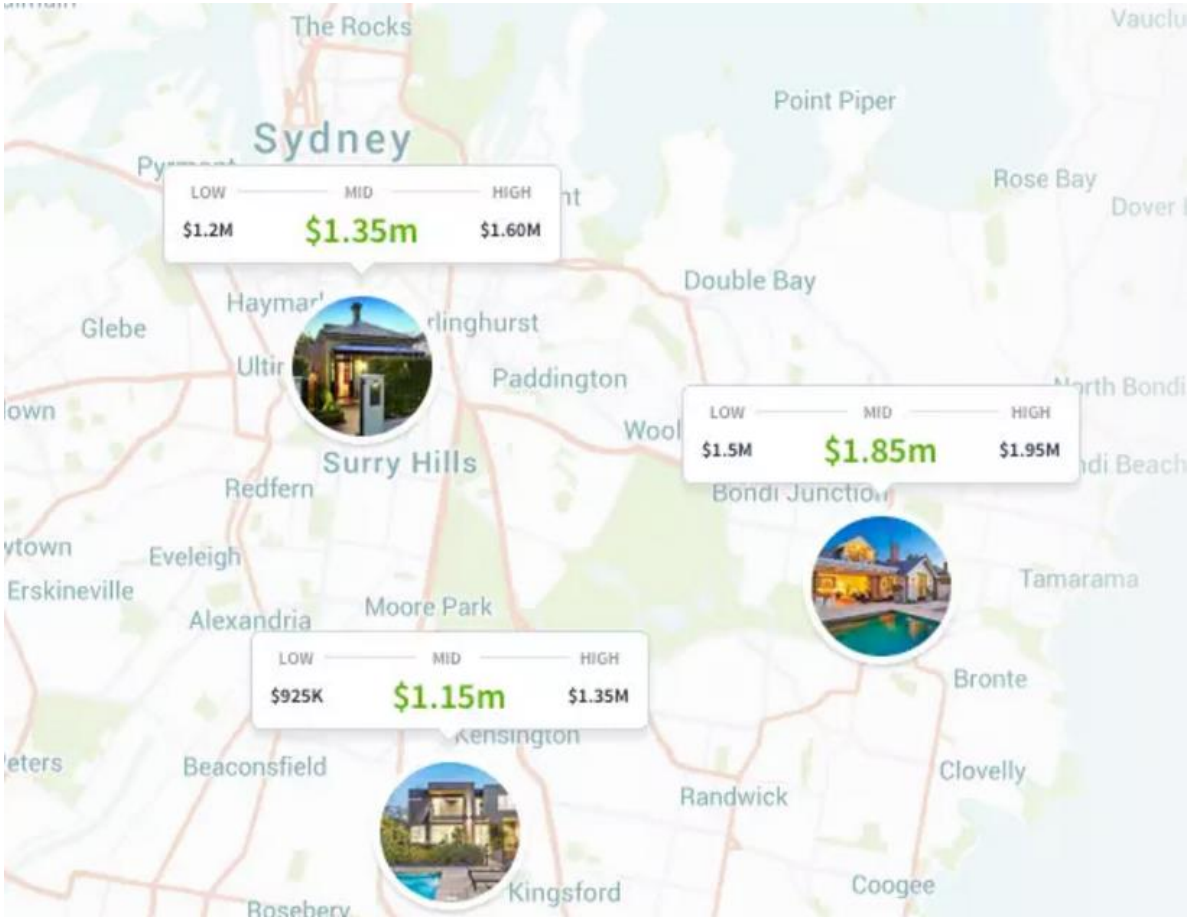


Aggregation of news from different sites



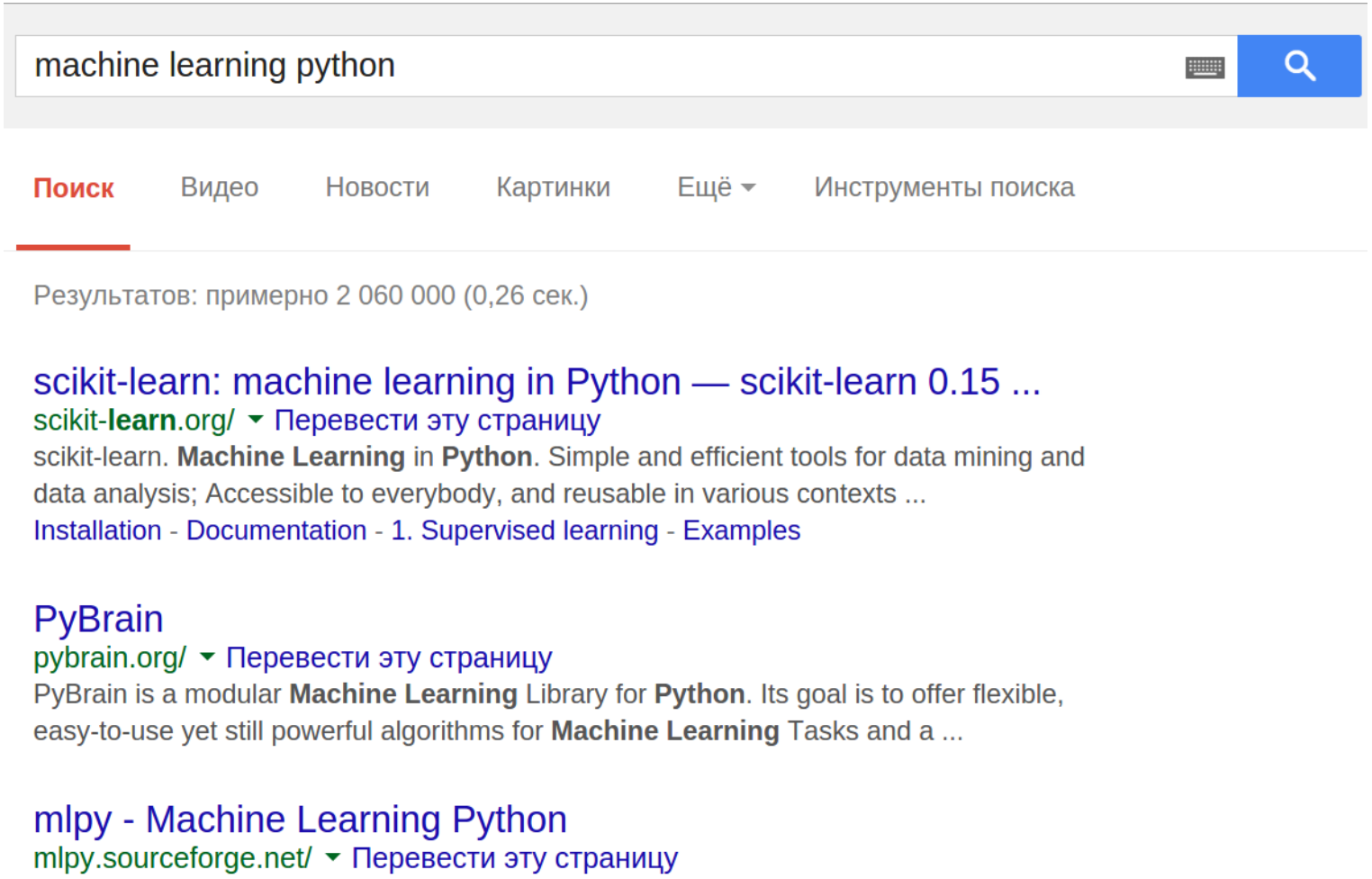
Skoltech
Skolkovo Institute of Science and Technology

Real estate price estimation

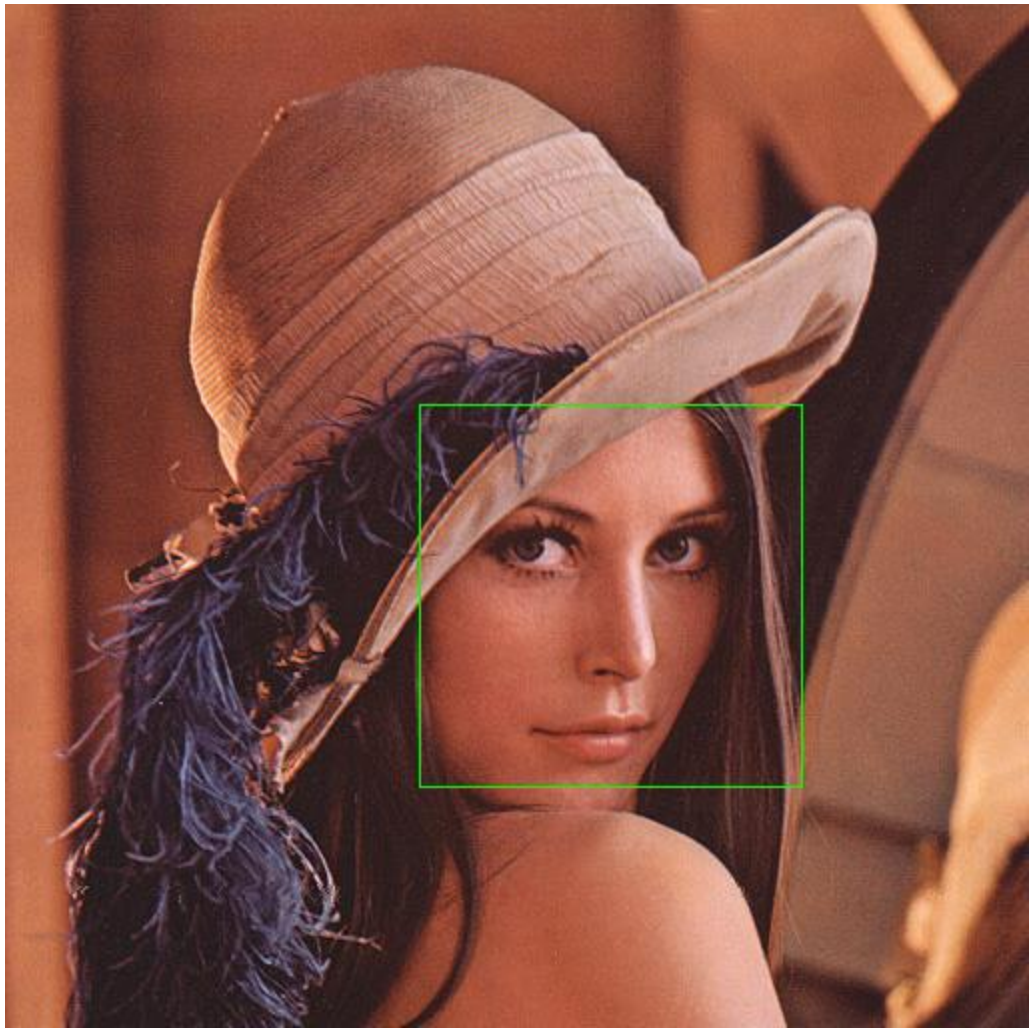


Recommender systems

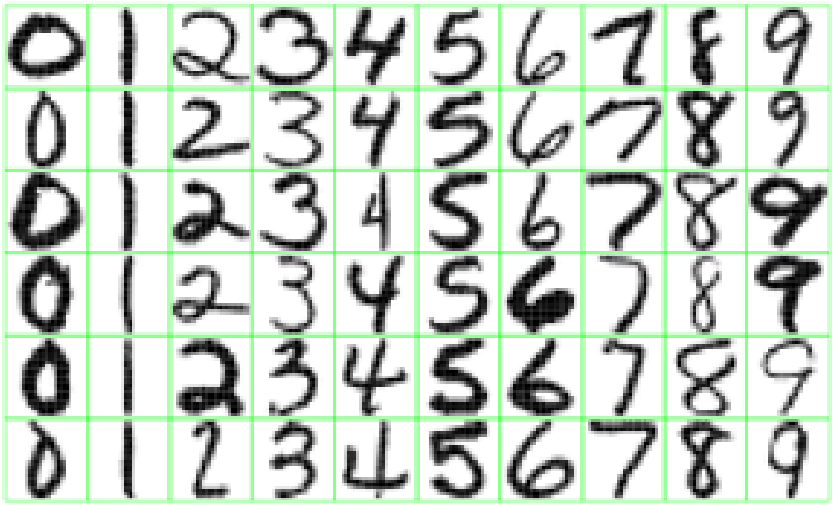
Web page ranking



Face detection



Digits recognition



Automatic sorting of letters based on handwritten ZIP code

Address recognition

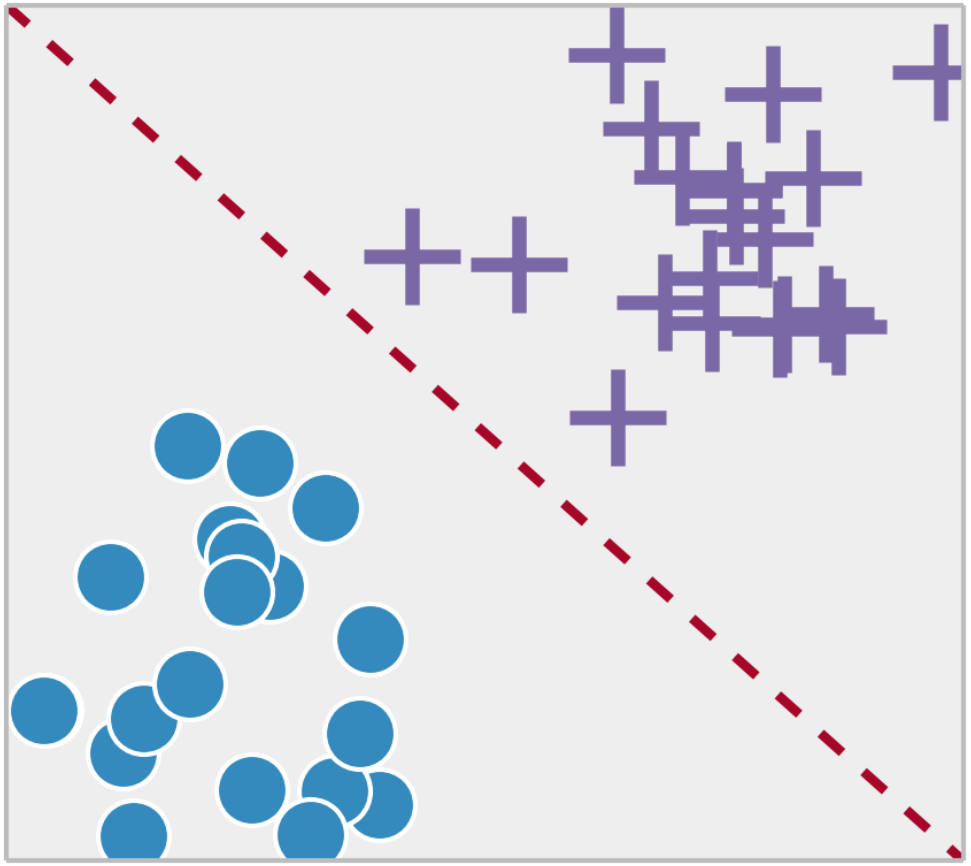
Outline

- Info about the course
- Examples of data science problems from everyday life
- **Major data analysis problem statements**

Supervised learning - classification

$$\{x_i \in R^d, y_i\}_{i=1}^n \rightarrow \hat{f}(x)$$
$$y_i \in \{c_1, \ldots, c_k\}, k < \infty$$

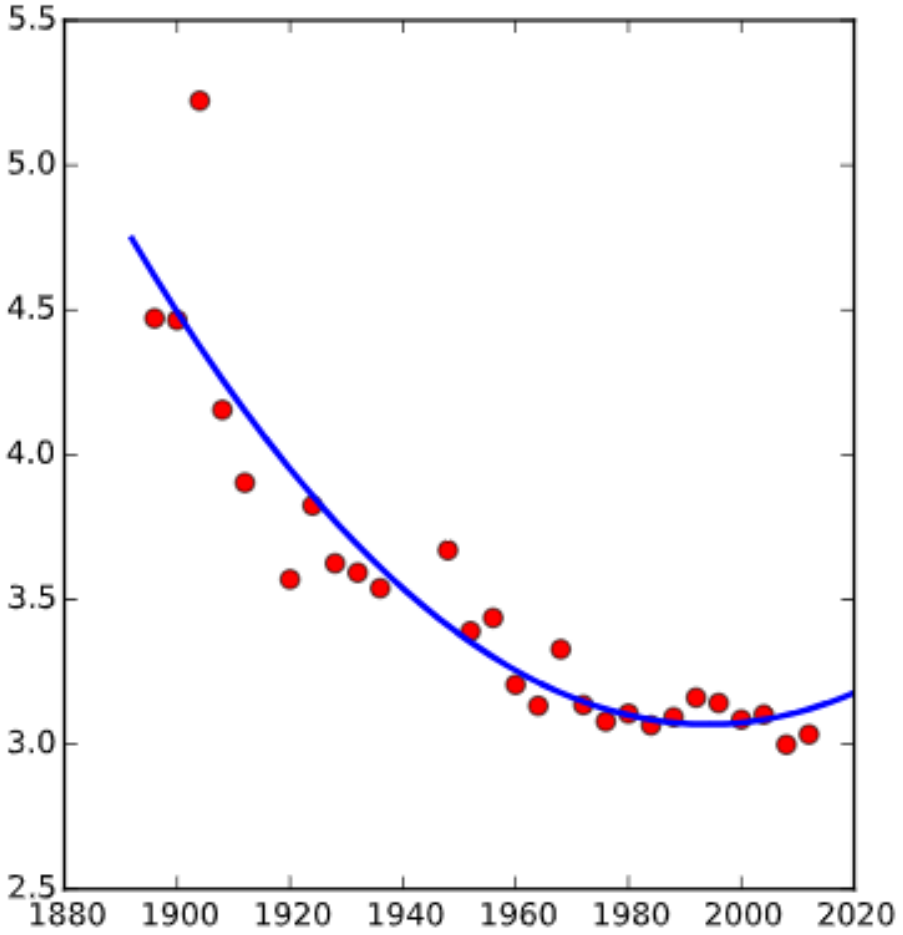
An example: spam filtering



Supervised learning - regression

$$\{x_i \in R^d, y_i\}_{i=1}^n \rightarrow \hat{f}(x) \qquad y_i \in R$$

An example: predict price of a house

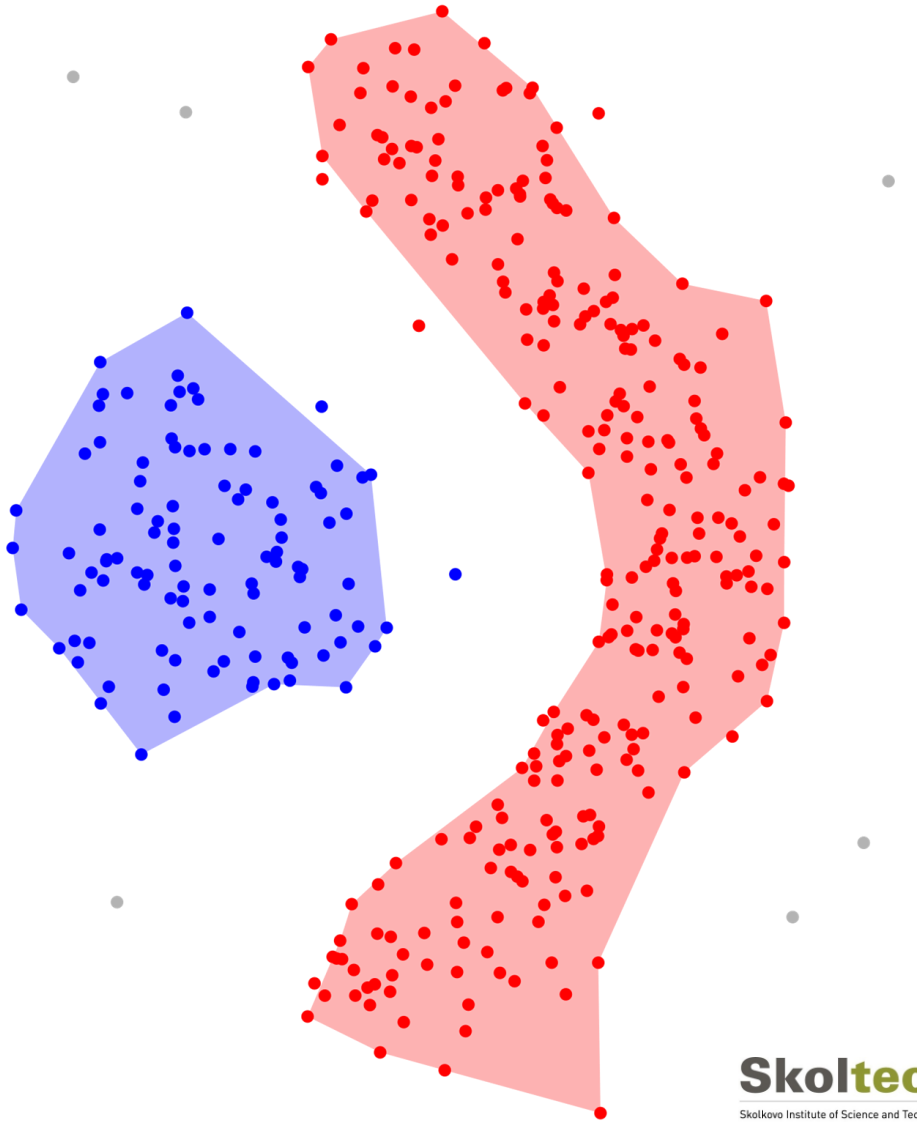


Unsupervised learning - clustering

$$\{x_i \in R^d\}_{i=1}^n$$

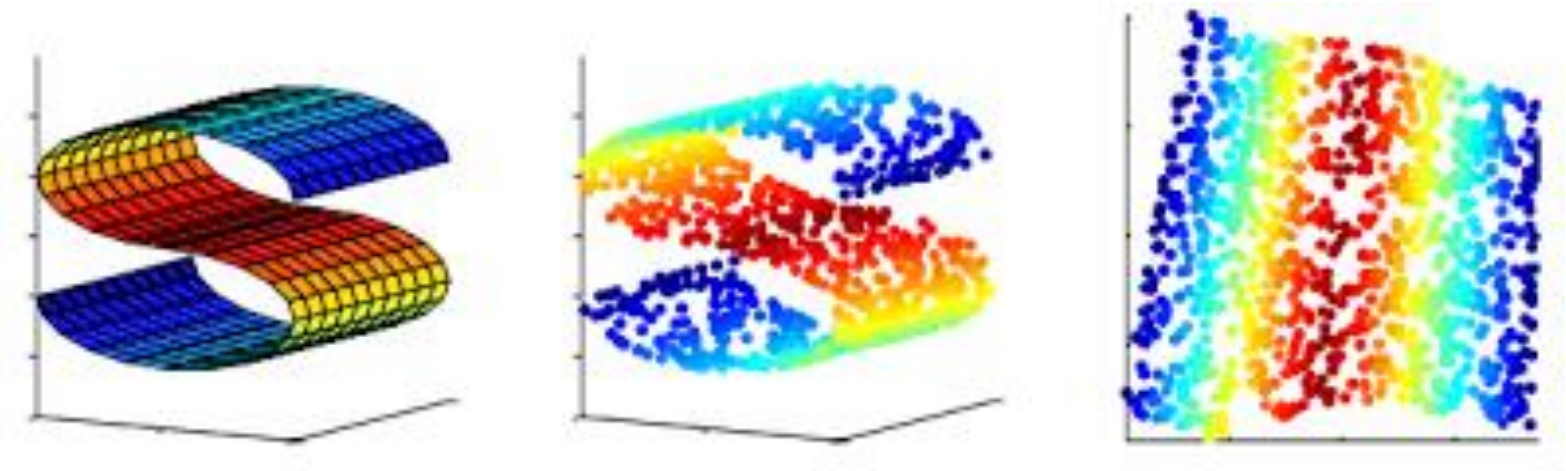
An example:

aggregation of news from different sites



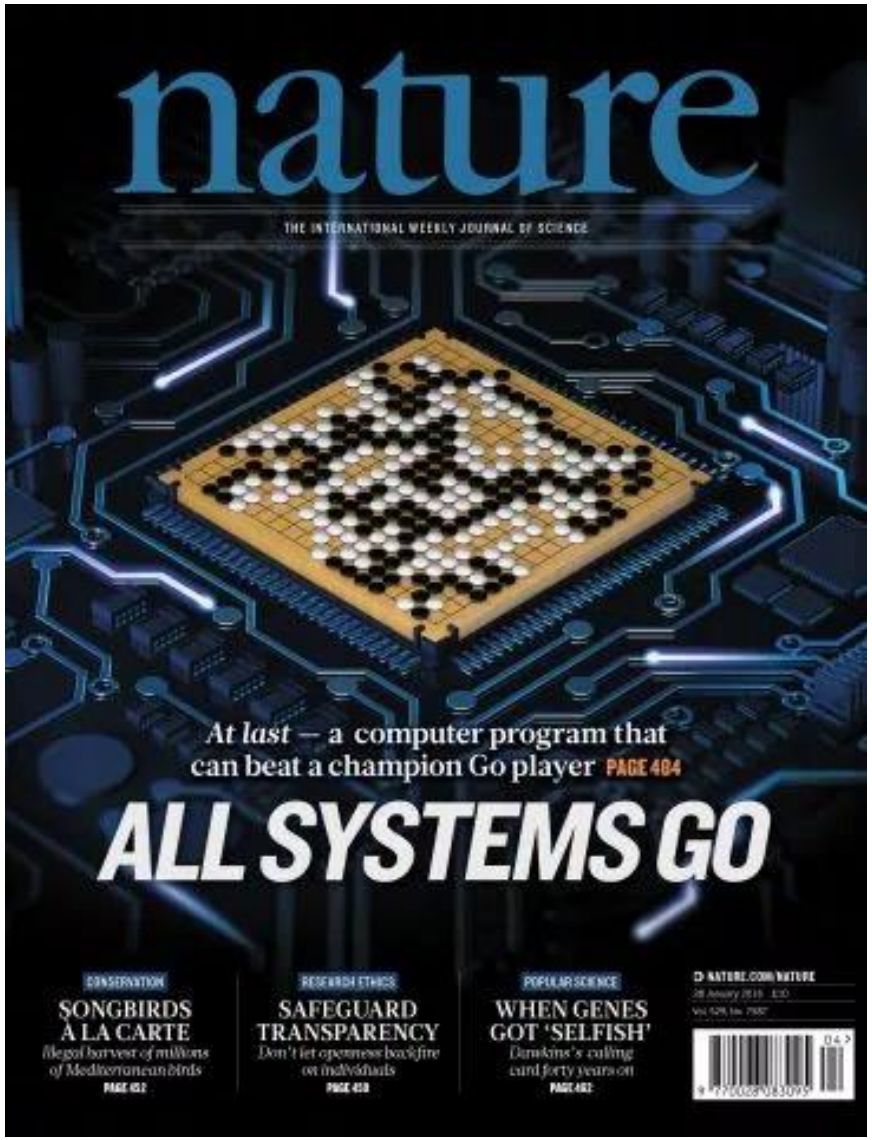
Unsupervised learning - dimensionality reduction

$$\{x_i \in R^d\}_{i=1}^n$$

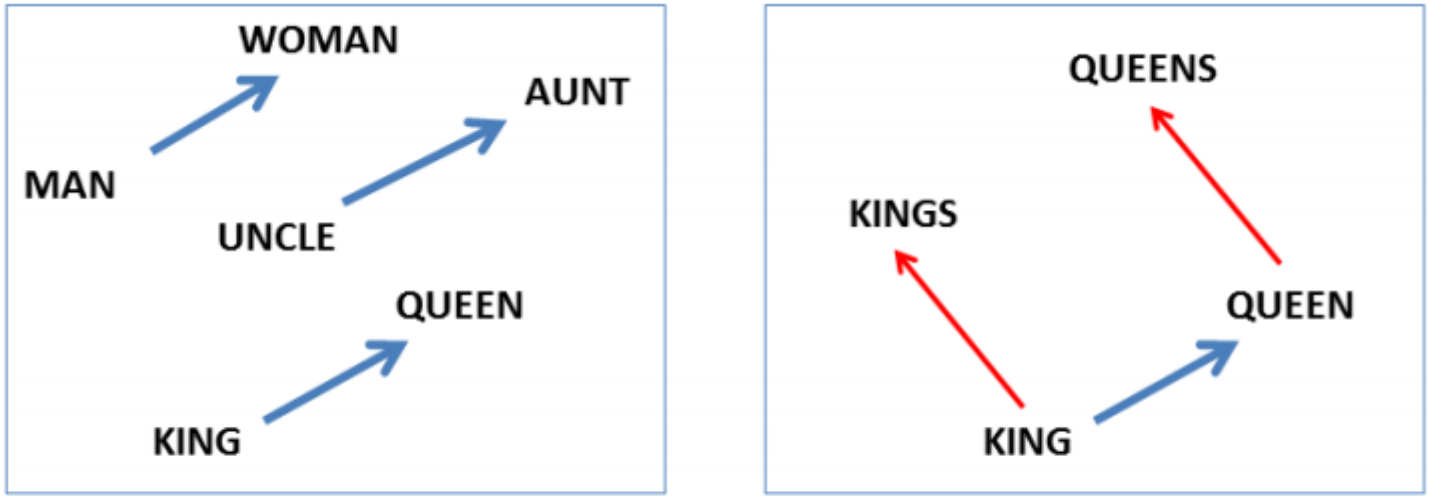


An example: generation of new airfoils (to be discussed in lecture)

Reinforcement learning



Representation learning



(Mikolov et al., NAACL HLT, 2013)

Examples - texts classification, sentiment analysis

Python libraries & links

To reproduce computational experiments you'll need the following libraries:

- Python 3.9
- Jupyter
- Open source machine learning libraries: Scikit learn; Pandas; Matplotlib; Seaborn.

Useful links (clickable):

- A Crash Course in Python for Scientists
- Scientific Computing with Python (the first notebook contains instructions for python libraries installation)
- Exploring the Titanic dataset with seaborn