

---

# Introduction to Data Science

## Lecture 1. General Introduction

**30.09.2021**

**Maxim Panov, Mikhail Belyaev**

# 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

# Assignments and Project

---

## Assignments

### Homework 1 (2<sup>nd</sup> week, 20%)

Exploratory analysis, regression and cross-validation:

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

### Homework 2 (3<sup>rd</sup> 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](https://t.me/ds_intro);
- Office hours with TAs (online via Telegram)

# Course Instructors

---

- Mikhail Belyaev, [m.belyaev@skoltech.ru](mailto:m.belyaev@skoltech.ru)
- Maxim Panov, [m.panov@skoltech.ru](mailto: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

# 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

---

- |   |  |
|---|--|
| <input type="checkbox"/> <b>Кадровый центр "Президен.</b> | <b>Поиск и подбор персонала.</b> - Добрый день, уважаемые па |
| <input type="checkbox"/> <b>Бизнес-практикум</b>          | <b>Филиальная сеть. Резервы эффективности.</b> - Изыскание   |
| <input type="checkbox"/> <b>UTS Group – USA</b>           | <b>ДОСТАВКА СБОРНЫХ и ЭКСПРЕСС ГРУЗОВ из Америки</b>         |
| <input type="checkbox"/> <b>Такси, Трансфер, Аренда .</b> | <b>Услуги службы такси для корпоративных клиентов.</b> - Те  |
| <input type="checkbox"/> <b>АвтоБлог</b>                  | <b>Самый длинный внедорожник</b> - Нажмите СПАМ если не ж    |

# Aggregation of news from different sites

## Подробнее о событии



**Минпромторг предложил снять ограничения на торговлю  
алкоголем в алюминиевых банках** Интерфакс 10:15 ☆

---



**Минпромторг хочет снять запрет на продажу пива ночью**  
РИА Новости 09:03 ☆

---



**Минпромторг захотел разрешить продажу пива в банках ночью**  
Газета.Ru 08:53 ☆

---



**Минпромторг предложил снять запрет на продажу пива ночью**  
Коммерсантъ 09:23 ☆

---



**«Ъ»: Минпромторг предлагает разрешить продажу пива в  
алюминиевых банках ночью** ТАСС 07:51 ☆

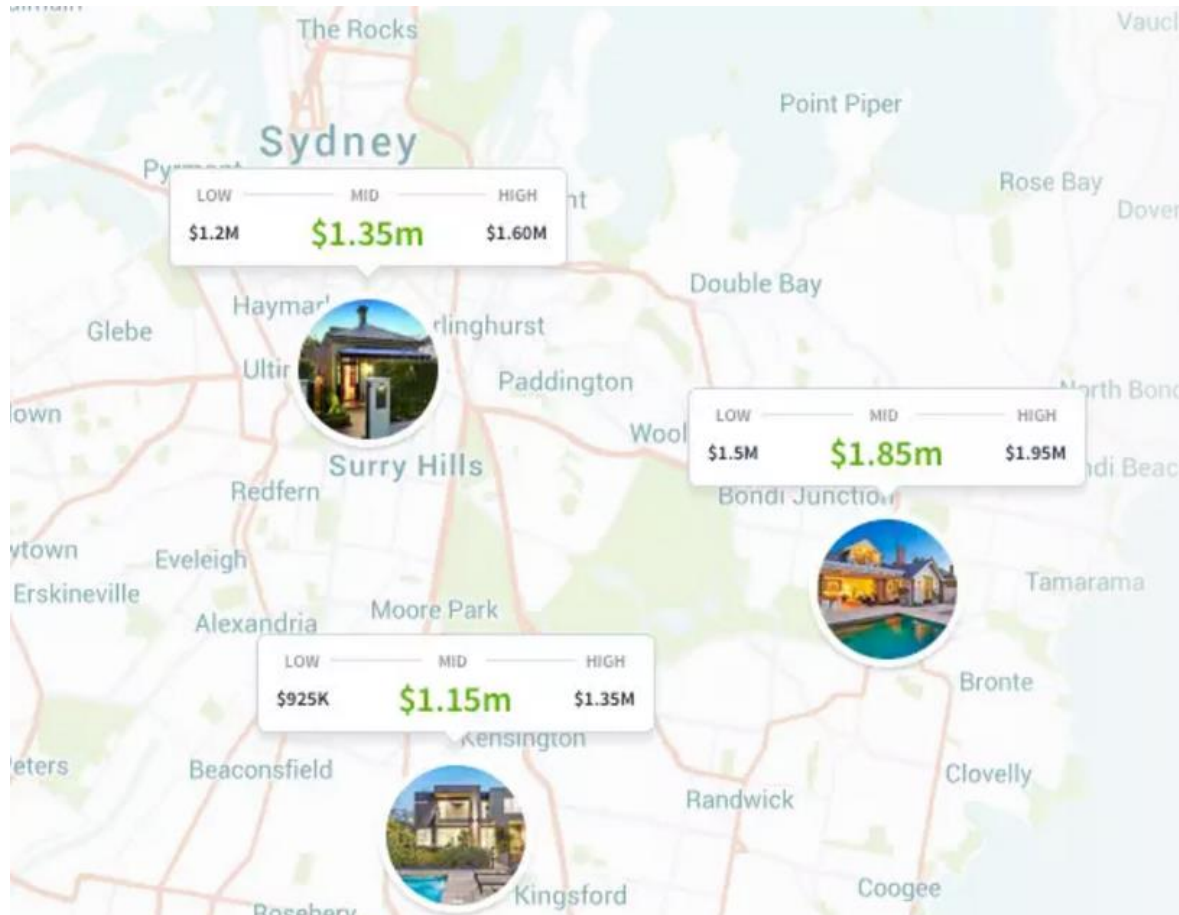
---



**Минпромторг предложил снять запрет на продажу пива ночью**  
Российская газета 07:18 ☆

---

# Real estate price estimation



# Recommender systems

## Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).



[Principles of Data Mining \(A...](#) ☐  
by David J. Hand

★★★★☆ (17) \$52.00



[Python in a Nutshell, Secon...](#) ☐  
by Alex Martelli

★★★★☆ (40) \$26.39



[Introductory Statistics wit...](#) ☐  
by Peter Dalgaard

★★★★☆ (20) \$48.56

# Web page ranking



Поиск

Видео

Новости

Картинки

Ещё ▾

Инструменты поиска

Результатов: примерно 2 060 000 (0,26 сек.)

[scikit-learn: machine learning in Python — scikit-learn 0.15 ...](#)

[scikit-learn.org/](#) ▾ [Перевести эту страницу](#)

scikit-learn. **Machine Learning** in **Python**. Simple and efficient tools for data mining and data analysis; Accessible to everybody, and reusable in various contexts ...

[Installation](#) - [Documentation](#) - [1. Supervised learning](#) - [Examples](#)

[PyBrain](#)

[pybrain.org/](#) ▾ [Перевести эту страницу](#)

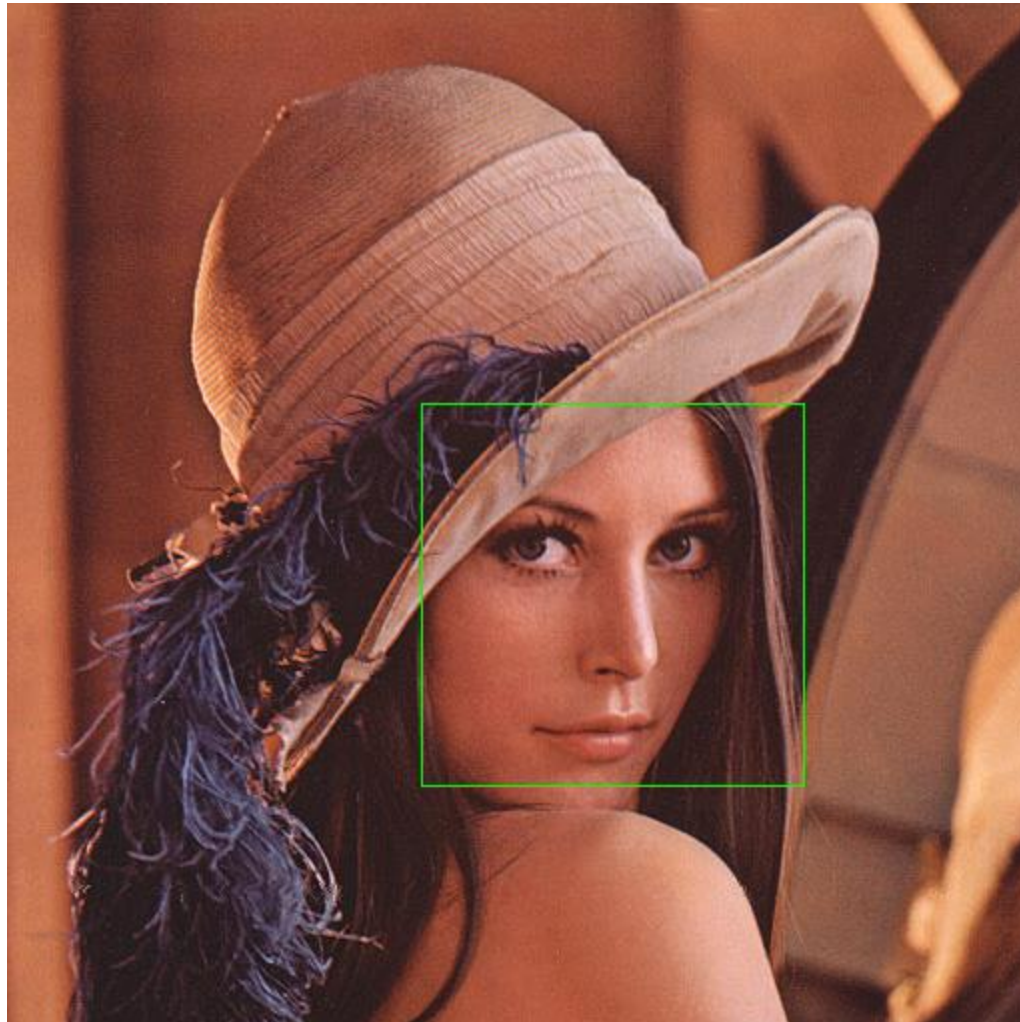
PyBrain is a modular **Machine Learning** Library for **Python**. Its goal is to offer flexible, easy-to-use yet still powerful algorithms for **Machine Learning** Tasks and a ...

[mlpy - Machine Learning Python](#)

[mlpy.sourceforge.net/](#) ▾ [Перевести эту страницу](#)

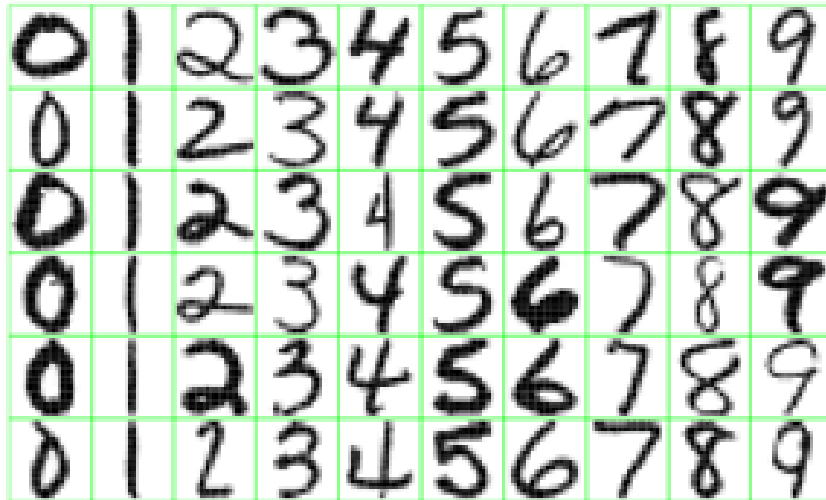
# Face detection

---





# Digits recognition



Automatic sorting of  
letters based on  
handwritten ZIP code

Address recognition





# Outline

---

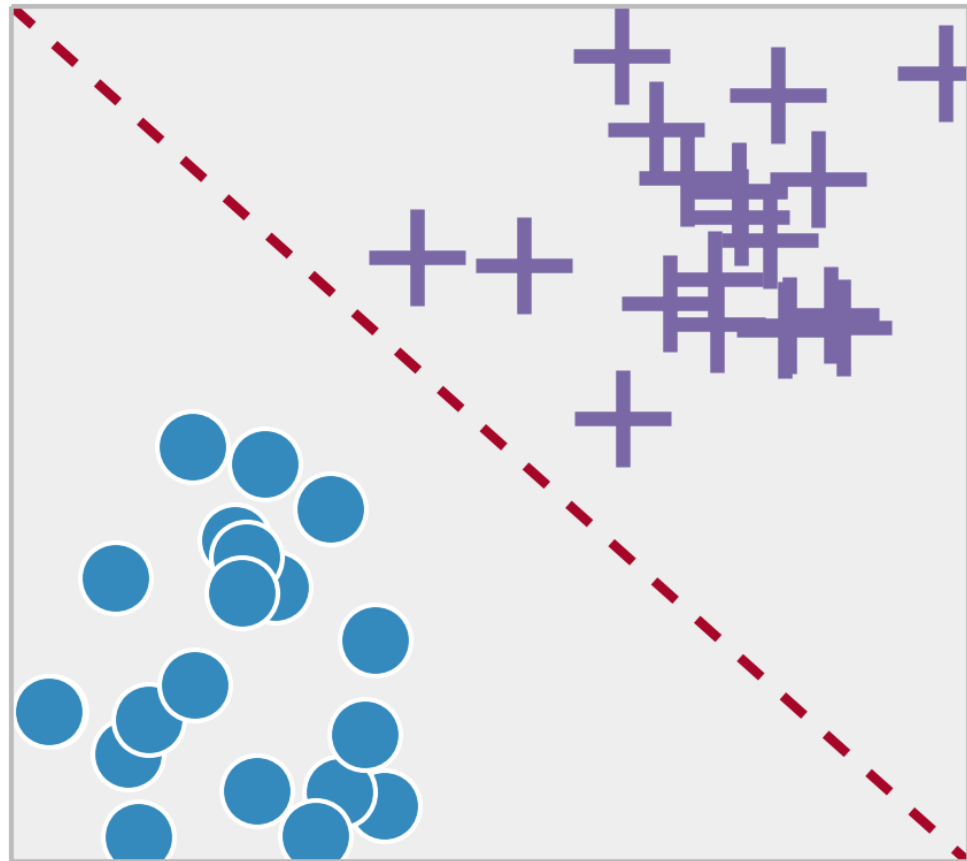
- Info about the course
- Examples of data science problems from everyday life
- **Major data analysis problem statements**

# Supervised learning - classification

---

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

An example: spam filtering

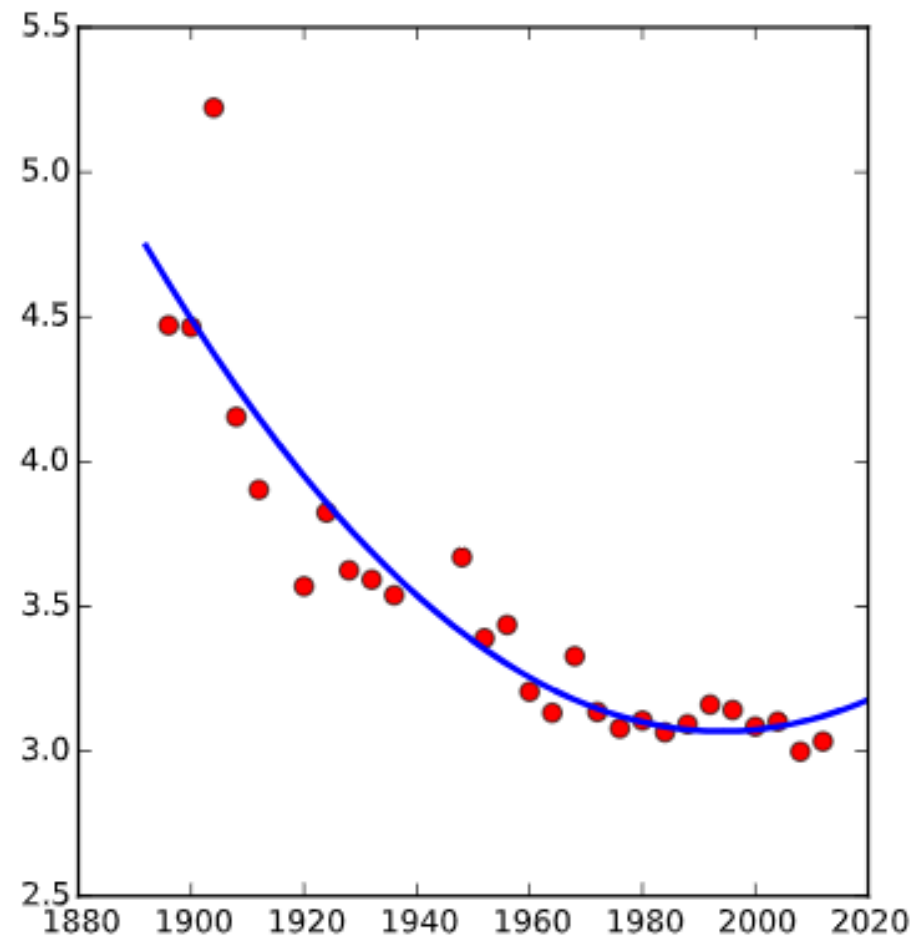


# Supervised learning - regression

---

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

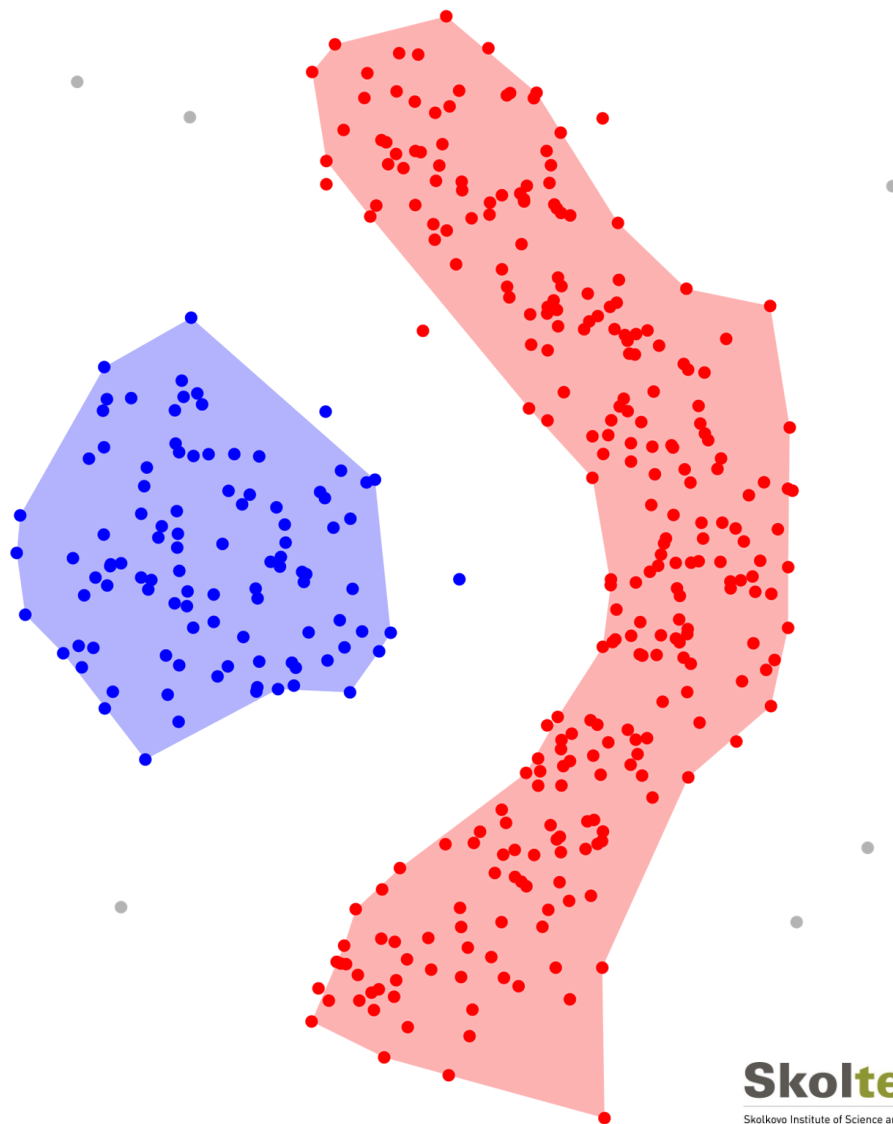
An example: predict price  
of a house



# Unsupervised learning - clustering

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

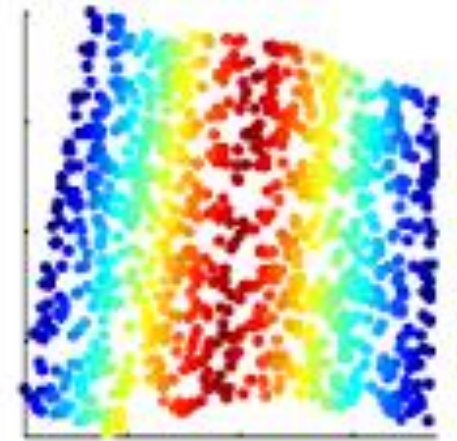
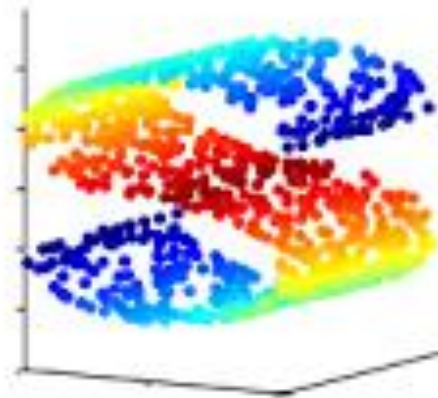
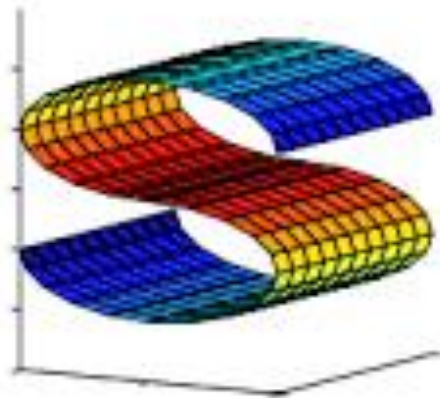
An example:  
aggregation of news  
from different sites



# Unsupervised learning - dimensionality reduction

---

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

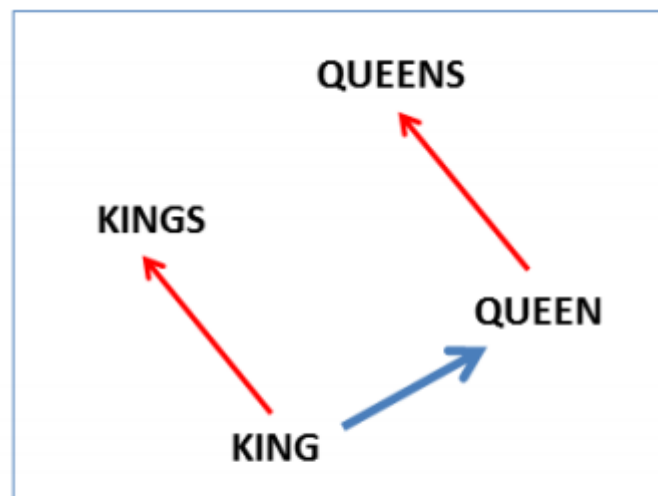
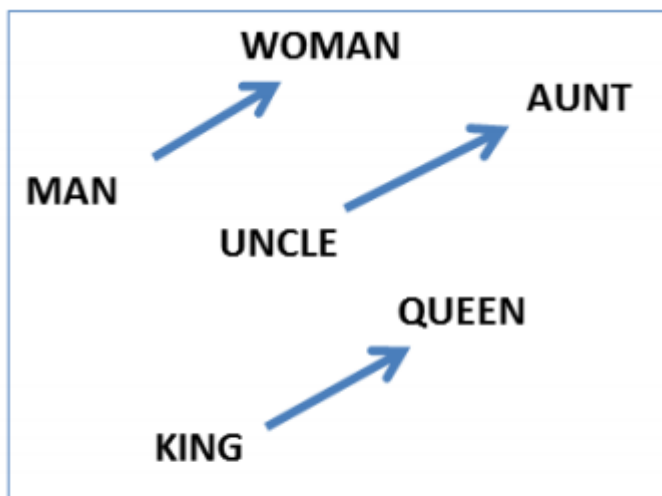


An example: generation of new airfoils (to be discussed in lecture)

# 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](#)