Трек Data Science

- Анализ данных (5 семестр)
  Вычислительная статистика (6 семестр)

- Можно было назвать Data Science I и Data Science II.

- В этом семестре: учимся, работая с «хорошими» наборами данных
  В следующем семестре: учимся работать с произвольными наборами данных

Data Science – это статистика?

- ДА: Статистика – это сбор данных, выявлением природы и составление прогнозов по этим данным. «Data Science» занимается тем же самым (с применением вычислительной техники).

- НЕТ: В статистике небольшие выборки (≤ 100) Data Science работает с мегабайтами и гигабайтами данных => другие возможности и другие методы.

- Классическую мат. статистику нужно знать.

Карьера в Data Science

- Примеры требований: IBM Yandex
- Требуются как навыки программирования и обработки данных, так и навыки статистики и машинного обучения.
- Как правило, требуется уровень образования не ниже магистра.
- Спрос большой, но подходящих кандидиатов мало.

Ландшафт ПО для анализа данных

- Язык программирования: в принципе, подходит любой, но предпочтительнее тот, на котором легче выражать идеи – R, **Python**, Scala, Julia
- Библиотеки для Python: scikit-learn, matplotlib, numpy, scipy, pandas
- Big Data: Hadoop, Spark
- Билиотеки/standalone решения: Vowpal Wabbit, XGBoost, CatBoost

«Хорошие» наборы данных

- scikit-learn: boston, iris, diabetes, digits, linnerud, wine, breast_cancer

- Kaggle: `https://www.kaggle.com/competitions`
  Категории: Getting started, Playground, InClass
  Titanic, House Prices, Digit Recognizer, New York Taxi Trip Duration, etc.

- R datasets: package 'datasets'

- Используйте эти наборы в своих проектах

Подготовка рабочей среды

- Установить Python
- Установить библиотеки scikit-learn, matplotlib, numpy, scipy, pandas
- Если все установлено правильно, код на следующий слайдах должен исполняться

Загрузка набора данных

```
from sklearn.datasets import load_boston
import matplotlib.pyplot as plt

boston = load_boston()
boston['data']
boston['target']
boston.feature_names

plt.scatter([x[1] for x in boston['data']],
  boston['target'])  # smt. crazy, right?
plt.show()
```

Не очень удобно...

pandas.dataframe

```
import pandas as pd
from sklearn.datasets import load_boston
import matplotlib.pyplot as plt

dataset = load_boston()
df = pd.DataFrame(dataset.data, columns=dataset.feature_names)
df['target'] = dataset.target
df
df.describe()

plt.scatter(df['CRIM'], df['target'])
```

Домашнее задание

1. Установите на свой компьютер Python и необходимые библиотеки

2. Загрузите набор данных load_iris в dataframe.

3. Выведите наименования признаков (feature names)

4. Постройте график зависимости зависимой переменной (target) от каждого признака

5. Напишите (от руки) получившиеся программу и сдайте на следующем занятии