НЕТОЛОГИЯ
групп

ОБЗОР БИБЛИОТЕКИ МАШИННОГО ОБУЧЕНИЯ SCIKIT-LEARN

ВЯЧЕСЛАВ МУРАШКИН

ВЯЧЕСЛАВ МУРАШКИН

Разработчик-

исследователь, Яндекс

mvjacheslav@gmail.com

a4tunado_

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ:

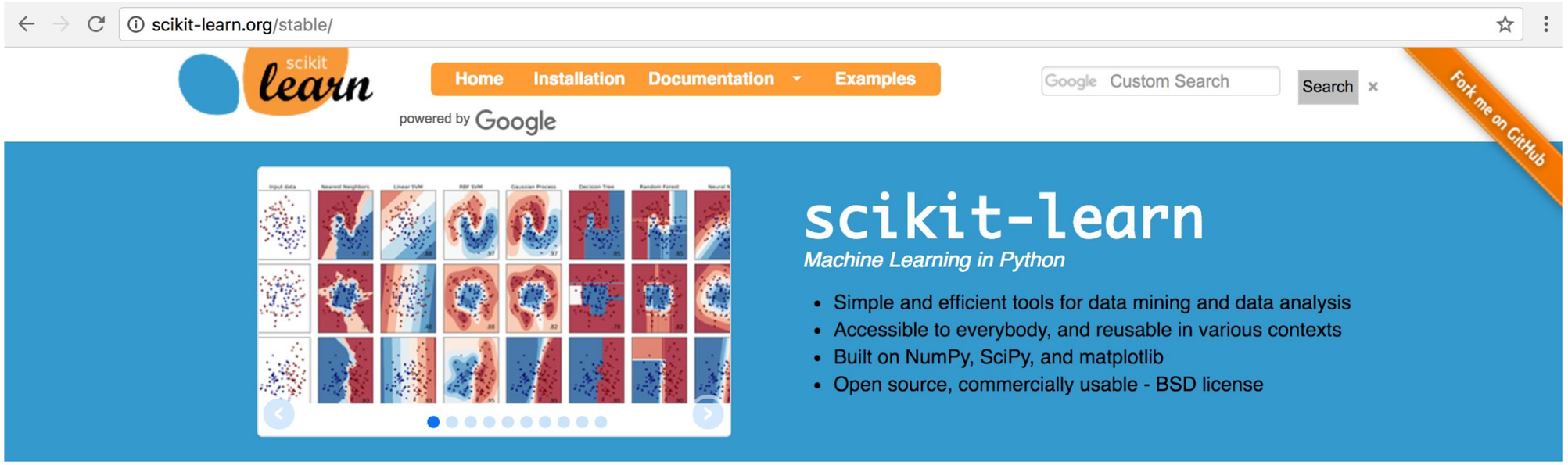
– изучите API библиотеки scikit-learn
– познакомитесь с основными модулями библиотеки scikit-learn
– научитесь применять библиотеку на реальных данных

О ЧЁМ ПОГОВОРИМ И ЧТО СДЕЛАЕМ

1. Возможности и ограничения scikit-learn
2. Модули библиотеки scikit-learn
3. Модель линейной регрессии
4. Кросс-валидация и подбор гиперпараметров модели
5. Пример применения библиотеки на реальных данных

1. БИБЛИОТЕКА SCIKIT-LEARN

БИБЛИОТЕКА SCIKIT-LEARN



ВОЗМОЖНОСТИ SCIKIT-LEARN

- **Построение моделей** машинного обучения
- Реализованы модели **классификации**, **регрессии** и **кластеризации**
- Для оценки качества моделей и подбора гипер-параметров реализованы алгоритмы **предобработки данных**, **алгоритмы кросс-валидации**
- Предоставляет API интерфейс для **Python**

ОСОБЕННОСТИ SCIKIT-LEARN

- Использует **методы оптимизации** из библиотеки SciPy
- Для большинства алгоритмов реализована возможность **параллельного обучения** с использованием библиотеки Joblib
- Для загрузки данных необходимо использовать **внешние библиотеки**, например NumPy или Pandas
- Данные для обучения модели должны быть загружены в **оперативную память**

2. МОДУЛИ SCIKIT-LEARN

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

- sklearn.**linear_model** - линейные модели классификации и регрессии
  - LinearRegression
  - Ridge
  - LogisticRegression
- sklearn.**tree** - дерево решений
  - DecisionTreeClassifier

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

- sklearn.**ensemble** - ансамбли решений: бустинг, лес
  - RandomForestClassifier
  - AdaBoostClassifier
  - GradientBoostingClassifier
- sklearn.**cluster** - обучение без учителя
  - KMeans
  - DBSCAN

ОЦЕНКА КАЧЕСТВА

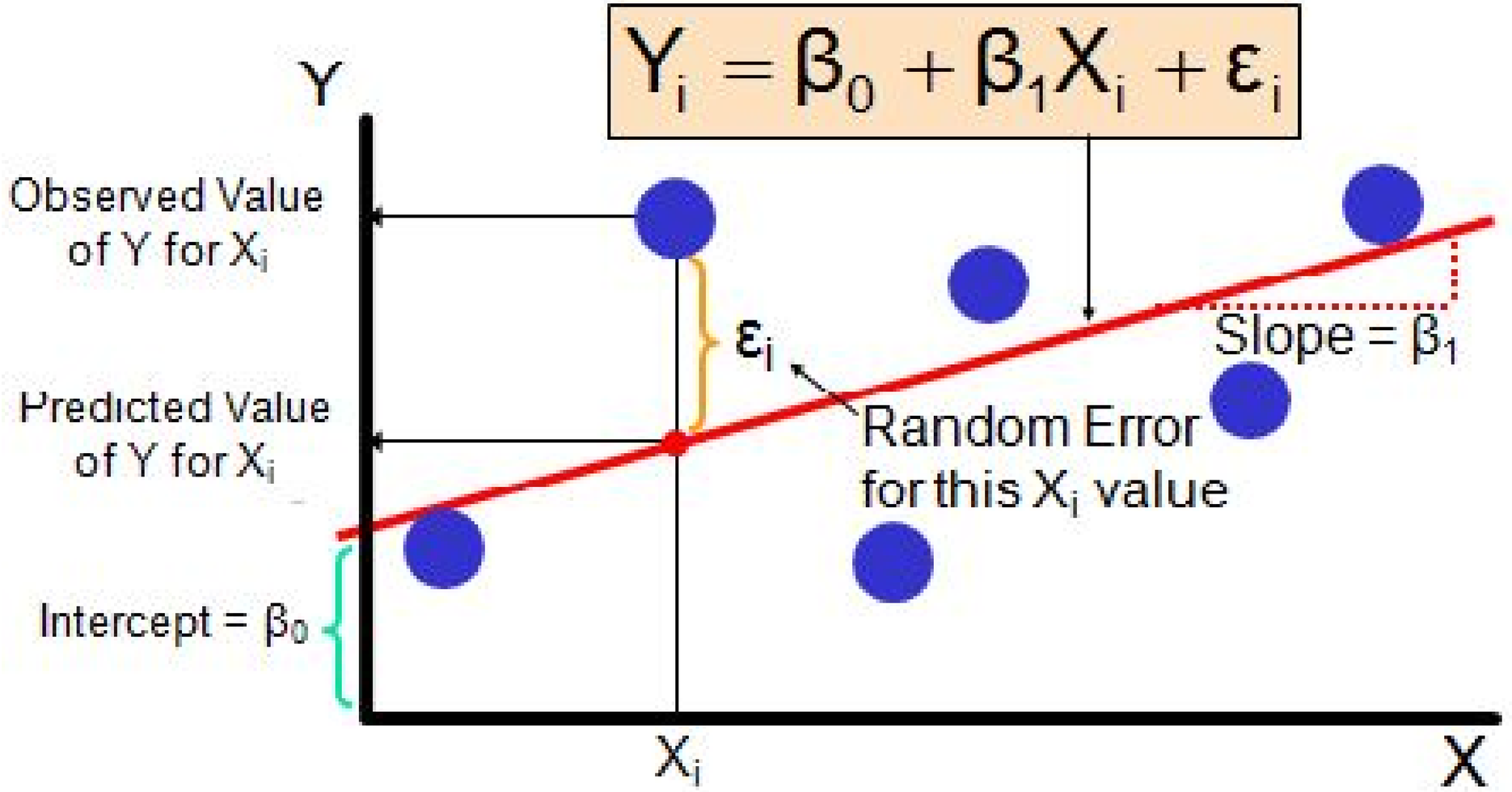
- sklearn.**metrics** - метрики качества алгоритмов
  - classification_report
  - mean_squared_error
- sklearn.**feature_selection** - оценка важности признаков
  - RFE
- sklearn.**model_selection** - оценка качества и подбор гипер-параметров
  - cross_val_score
  - GridSearchCV

ПРЕДОБРАБОТКА ДАННЫХ

- sklearn.**preprocessing** - нормализация, центрирование, бинаризация
  - StandardScaler
- sklearn.**feature_extraction** - предобработка сырых данных
  - HashingVectorizer
  - TfidfTransformer
- sklearn.**decomposition** - разложение матриц и снижение размерности
  - PCA
  - TruncatedSVD

3. ЛИНЕЙНАЯ РЕГРЕССИЯ

ЗАДАЧА


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

ОЦЕНКА ПАРАМЕТРОВ

$$Y_i = \beta_0 + \beta_1 X_i \qquad SSE = \sum_i^n (Y_i - \hat{Y}_i)^2$$

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \beta_i \overline{x}$$

РЕГУЛЯРИЗАЦИЯ (RIDGE)

$$SSE + \lambda \sum_{j=1}^{k} \beta_j^2$$

$$\sum_{i}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

МЕТРИКИ КАЧЕСТВА

1. Среднеквадратичная ошибка

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

2. Коэффициент детерминации

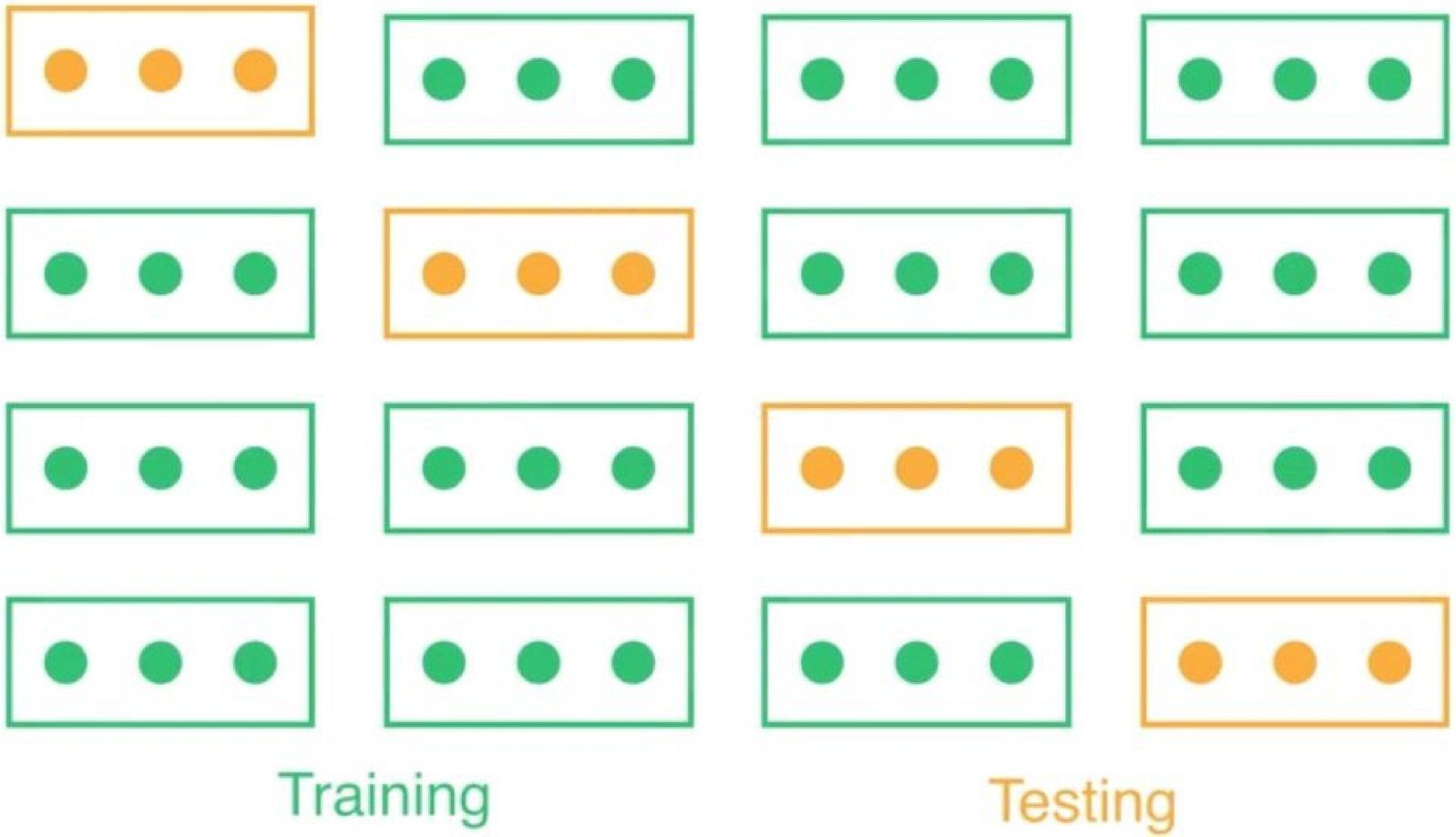
$$R2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \overline{y})^2}$$

4. КРОСС-ВАЛИДАЦИЯ

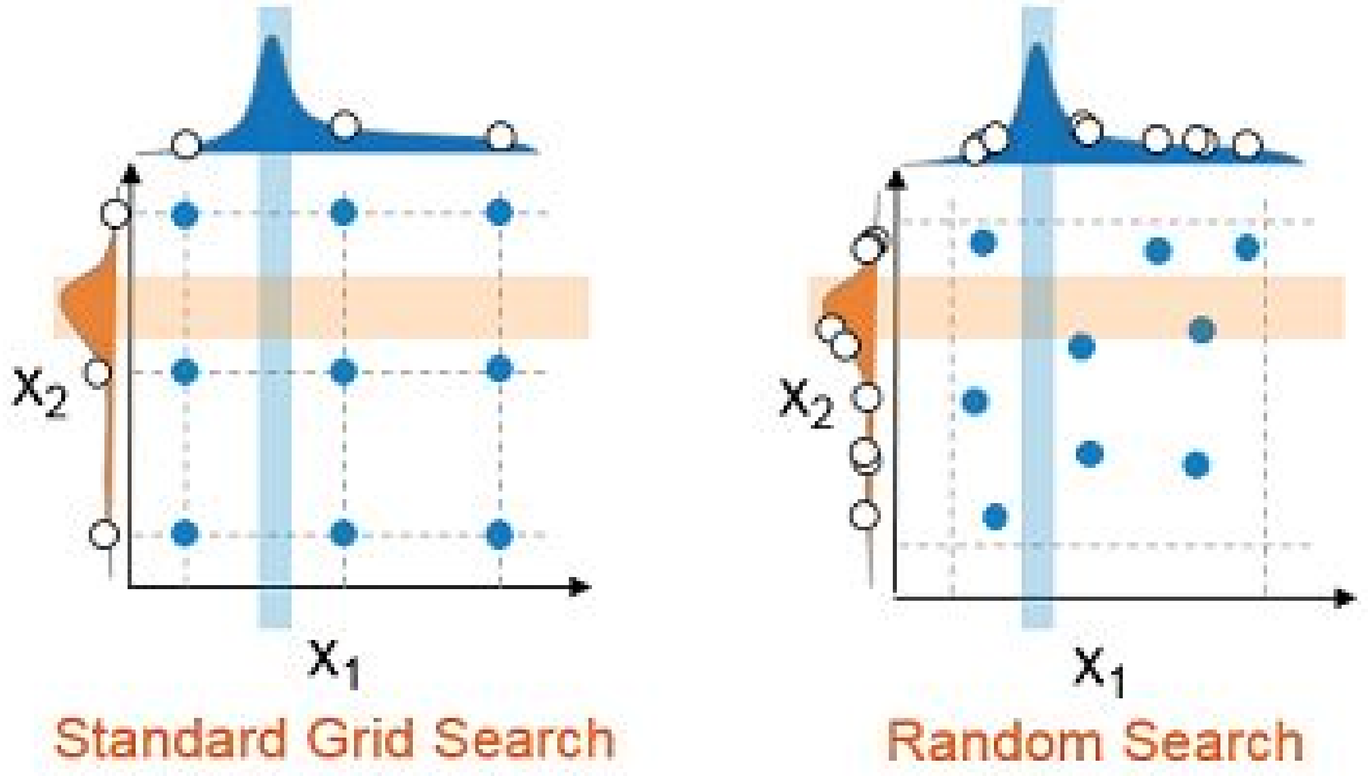
ОБОБЩАЮЩАЯ СПОСОБНОСТЬ

- Алгоритм обучения обладает **обобщающей способностью**, если вероятность ошибки на тестовой выборке не сильно отличается от ошибки на обучающей выборке
- **Переобучение** - нежелательное явление, при котором вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке
- Переобучение связано с **избыточной сложностью** используемой модели

K-FOLD КРОСС-ВАЛИДАЦИЯ



ПОДБОР ГИПЕР-ПАРАМЕТРОВ



ПРАКТИЧЕСКОЕ ЗАДАНИЕ

ЗАДАНИЕ

1. Построить регрессионную модель на данных Boston Housing Dataset
2. Оценить качество модели на кросс-валидации
3. Подобрать гипер-параметры модели
4. Сравнить различные алгоритмы регрессии для данной задачи

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

1. Познакомились с API библиотеки scikit-learn
2. Изучили возможности и особенности библиотеки
3. Научились применять scikit-learn для решения практических задач

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

1. An introduction to machine learning with scikit-learn
2. Scikit_Learn_Cheat_Sheet_Python.pdf
3. MachineLearning.ru

НЕТОЛОГИЯ
групп

Спасибо за внимание!

МКРАШКИН ВЯЧЕСЛАВ

mvjacheslav@gmail.com

a4tunado_