

Rotman

INTRODUCTION TO SCIKIT-LEARN

A Python Package for Machine Learning

August 17, 2021 Prepared by Niti Mishra



Rotman School of Management
UNIVERSITY OF TORONTO

Agenda

1. What is Scikit-Learn?
2. Data Modelling
3. Machine Learning
4. Installation
5. Hands-on Implementation

What is Scikit-Learn?

What is Scikit-Learn?

1/5



- **Scikit-Learn (Sklearn) is a powerful and robust open-source machine learning library for Python.**
- **Sklearn provides tools for efficient implement of classification, regression, clustering and dimensionality reduction techniques.**
- **Sklearn has a clean and uniform API as well as complete online documentation.**
- **Basic knowledge of NumPy, Pandas, SciPy and Matplotlib is required to successfully use Sklearn for machine learning.**

What is Scikit-Learn?

1/5



- **2007: Sklearn was initially developed by David Cournapeau as a Google summer code project.**
- **2010: Developers from French Institute for Research in Computer Science and Automation took sklearn to another level and made its first public release (v0.1)**
- **Since then there have been 12+ versions of iterations and improvements. The latest version is 0.24.0.**

What is Scikit-Learn?

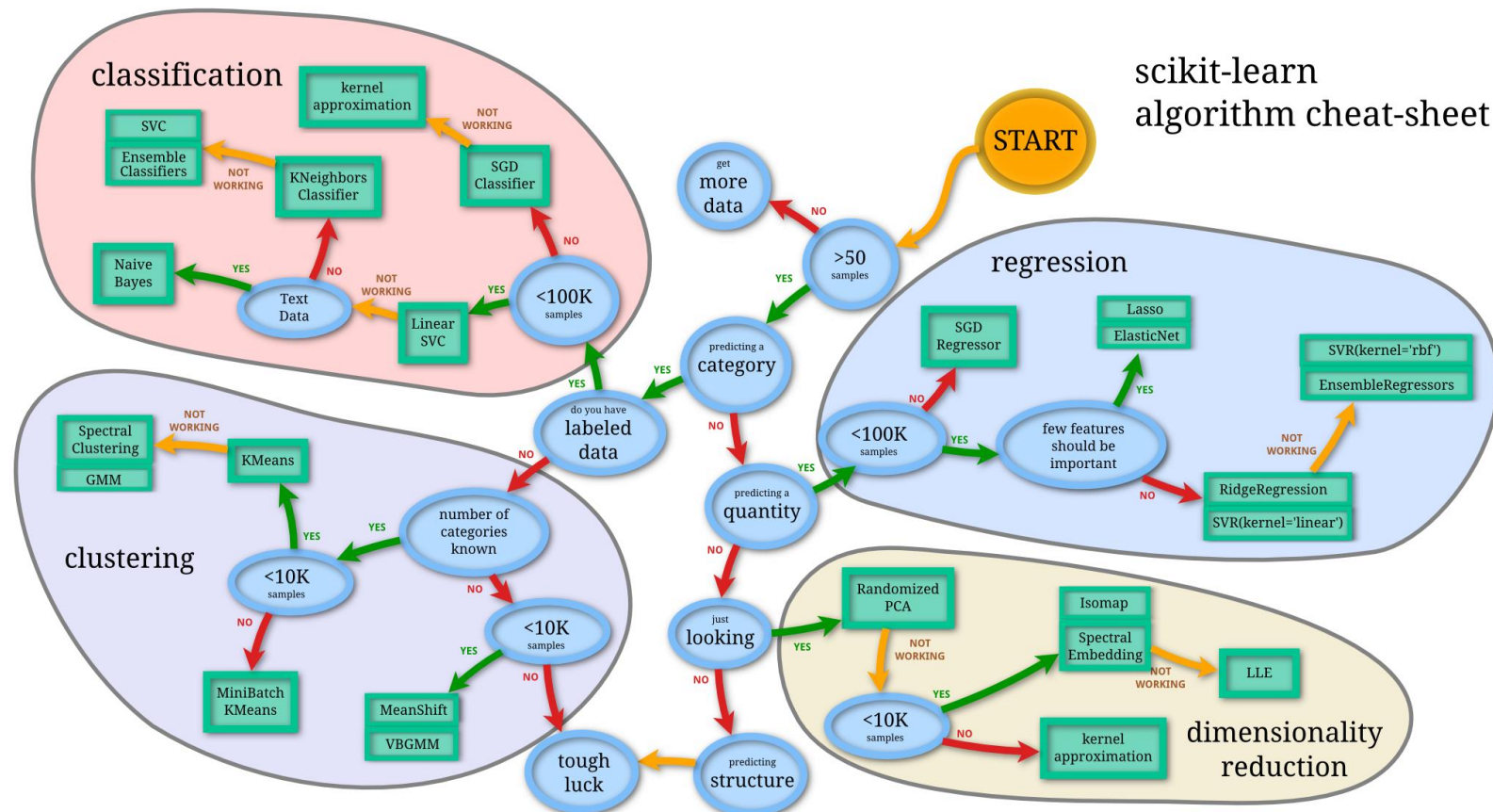
1/5



- Sklearn is an community project and anyone can contribute to it.
- Currently, there are more than 2058 contributors on its [github repository](#).
- Various organizations including booking.com, JP Morgan, Evernote, Spotify use Sklearn.

Data Modelling

- Sklearn is focused on modelling the data rather than reading or writing data.



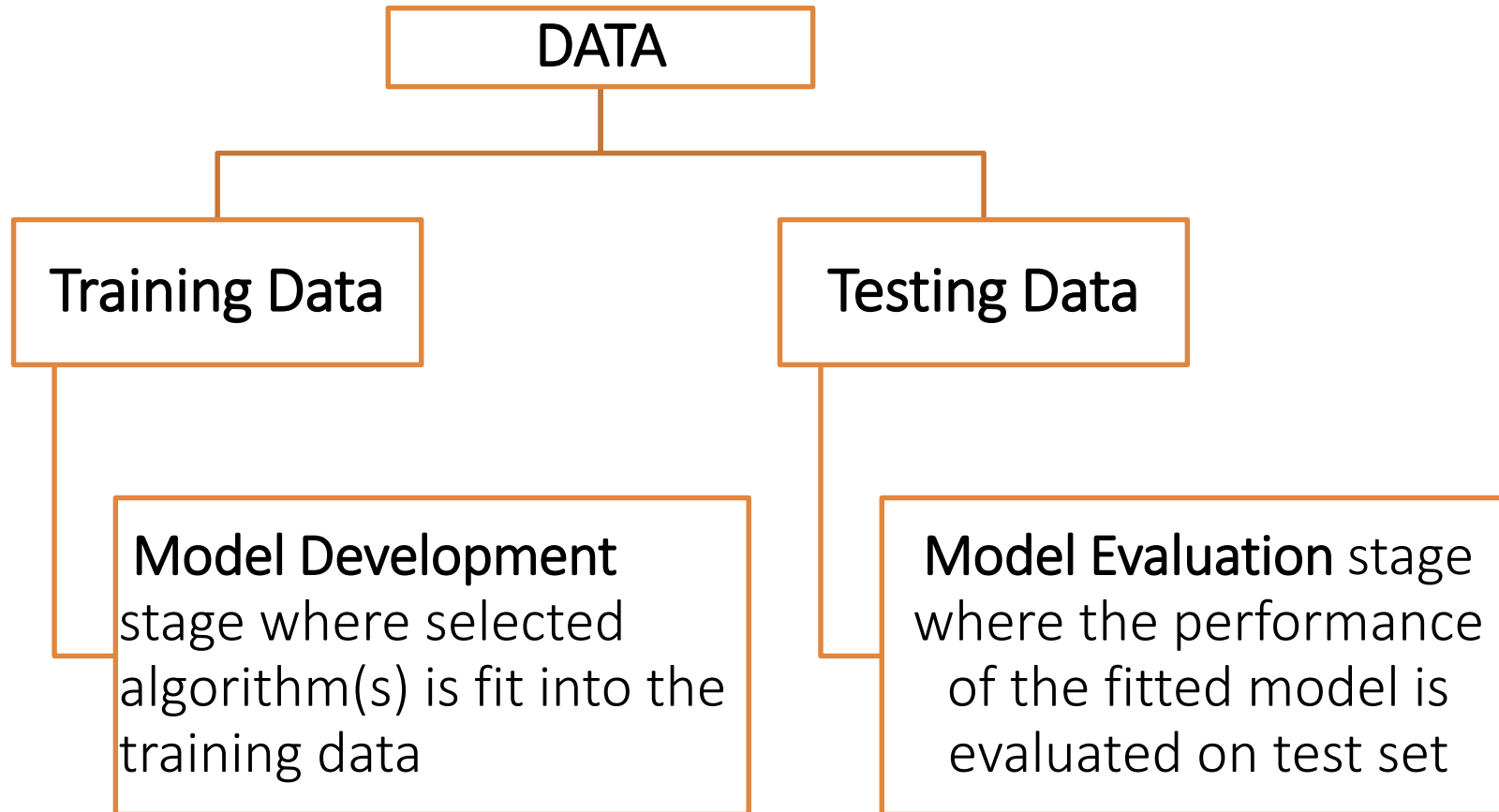


- **Sklearn offers numerous tools for**
 - **efficient data modelling**
 - **preprocessing support such as data encoding**
 - **feature selection & extraction**
 - **hyper-parameter search tools**
 - **end to end data modelling pipeline**

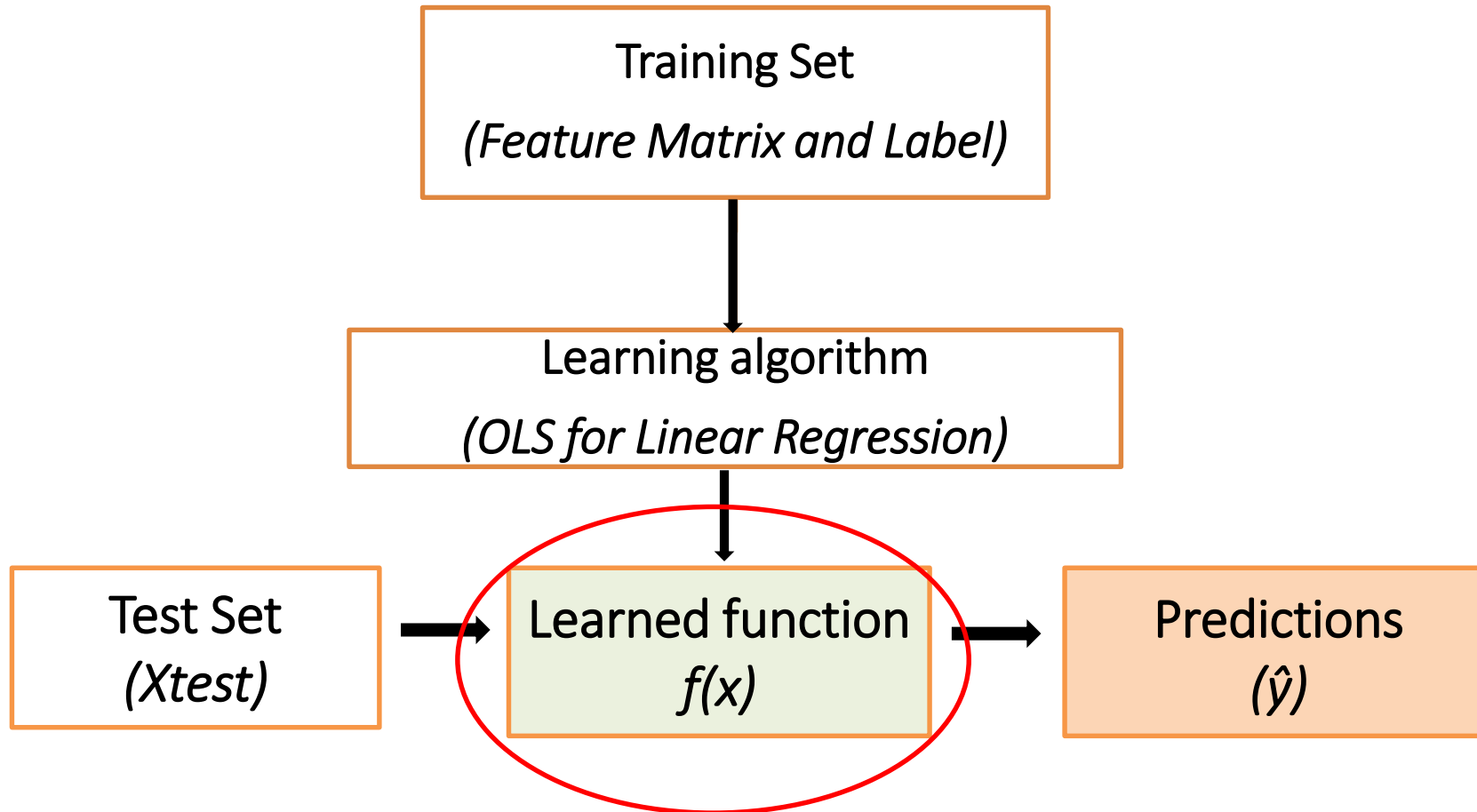
Machine Learning

- **Machine Learning (ML) is a study of algorithms that can learn to solve a specified task using data.**
- **ML models are trained using a sample of historical data called the training data and the model itself is evaluated based on its performance on an unseen data called the test data.**
- **ML has wide variety of application from research to health to finance to speech recognition and language translation.**

- **There are two main types of ML models:**
 1. **Supervised:**
 - **Model learns to identify pattern in data using inputs and desired outputs called labels.**
 - **Each training example has an array of properties, known as feature vector or input vector and a label, known as output.**
 - **Examples: Linear Regression, Logistic Regression, Random Forest Classifier, Decision Trees**
 2. **Unsupervised**
 - **Model learns to identify pattern and structure in the data without any labels**
 - **Examples: K-means Clustering, Principal Component Analysis, etc.**



- Both model development and model evaluation stage comprises additional steps. For example:
 - Crossvalidation
 - Hyperparameter search
- All these steps can be neatly packed into a pipeline object.



- Goal is to learn this function $f(x)$ that is a “good” predictor for y i.e. minimizes the error
- To do so, it starts with some initial guess and then makes changes to make the error smaller
- It iterates until hopefully it has converged to a value that is minimum

Installation

To install sklearn:

```
conda install -c conda-forge scikit-learn
```

Type and enter on
your Anaconda
prompt application

Prerequisite packages
will also be installed

To check sklearn version installed:

`conda list scikit-learn` } Type and enter on
your Anaconda
prompt application

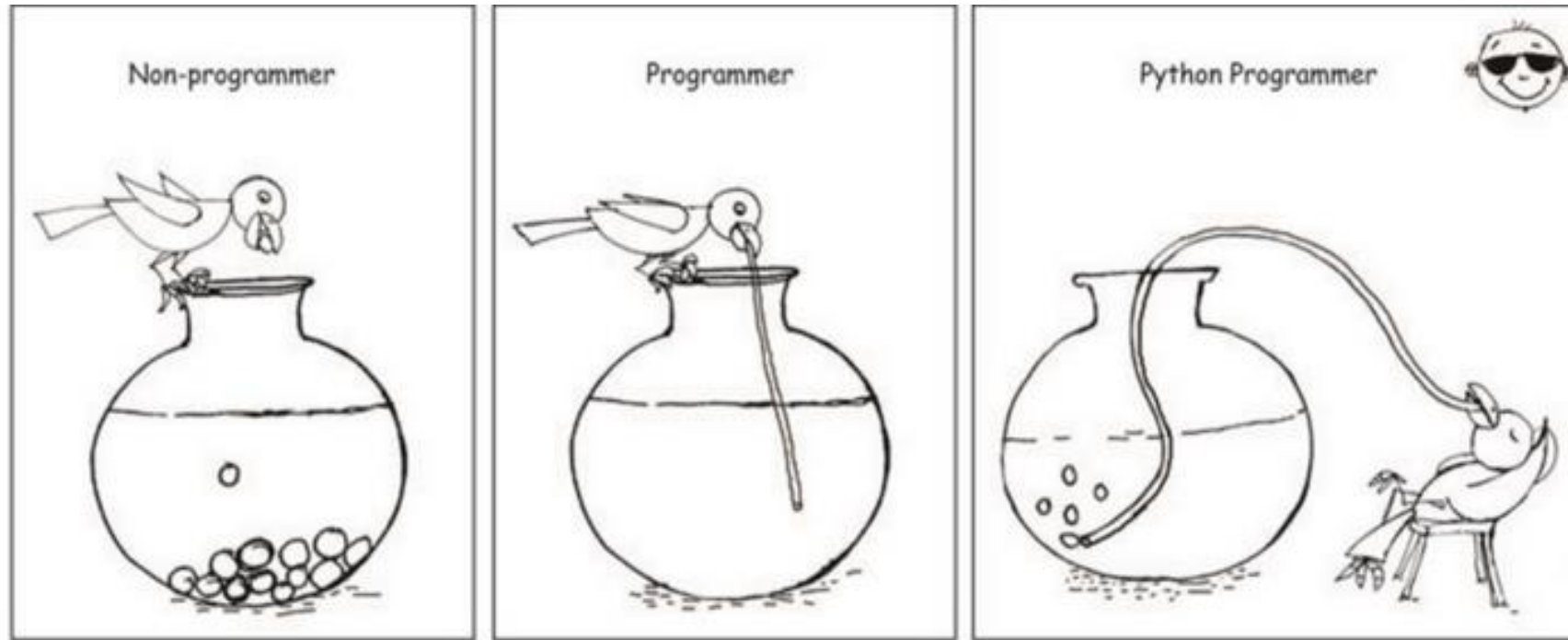
Or to a see list of installed packages:

`conda list`

Hands-on Implementation

- Go to <https://tdmdal.github.io/sklearn-workshop/>
- To open notebook on your local jupyter notebook
 - Download “Introduction to Scikit-learn” under Python Notebook Notebooks
 - Download “Advertising and Sales” under Data
 - Save both file in one folder
 - Open jupyter notebook from that folder
- To open notebook on google drive
 - Go to “Introduction to Scikit-learn” under Python Notebook Notebooks
 - Click on “Open in Colab”
 - Download “Advertising and Sales” under Data and upload on google drive
 - Mount your google drive to the drive folder where the data is uploaded

- Tutorials:
 - Linear Regression in Python <https://realpython.com/linear-regression-in-python/#simple-linear-regression-with-scikit-learn>
 - Sklearn Quick Start Tutorial <http://scikit-learn.org/stable/tutorial/basic/tutorial.html>
 - Sklearn User Guide http://scikit-learn.org/stable/user_guide.html
 - Sklearn API Reference <http://scikit-learn.org/stable/modules/classes.html>
 - [PyCon 2014 Scikit-learn Tutorial](https://github.com/ogrisel/sklearn_pycon2014) by Jake VanderPlas (https://github.com/ogrisel/sklearn_pycon2014)
 - Introducing Sklearn <https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html>
- Books:
 - [Python Data Science Handbook](#) (2016)
 - [Learning scikit-learn: Machine Learning in Python](#) (2013)
 - [Hands-On Machine learning with Scikit-Learn and Tensorflow](#) (2017)



Who wants to become a Python Programmer?

Questions?

Thank you