

***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 0.21.0.**

# What is Scikit-Learn?

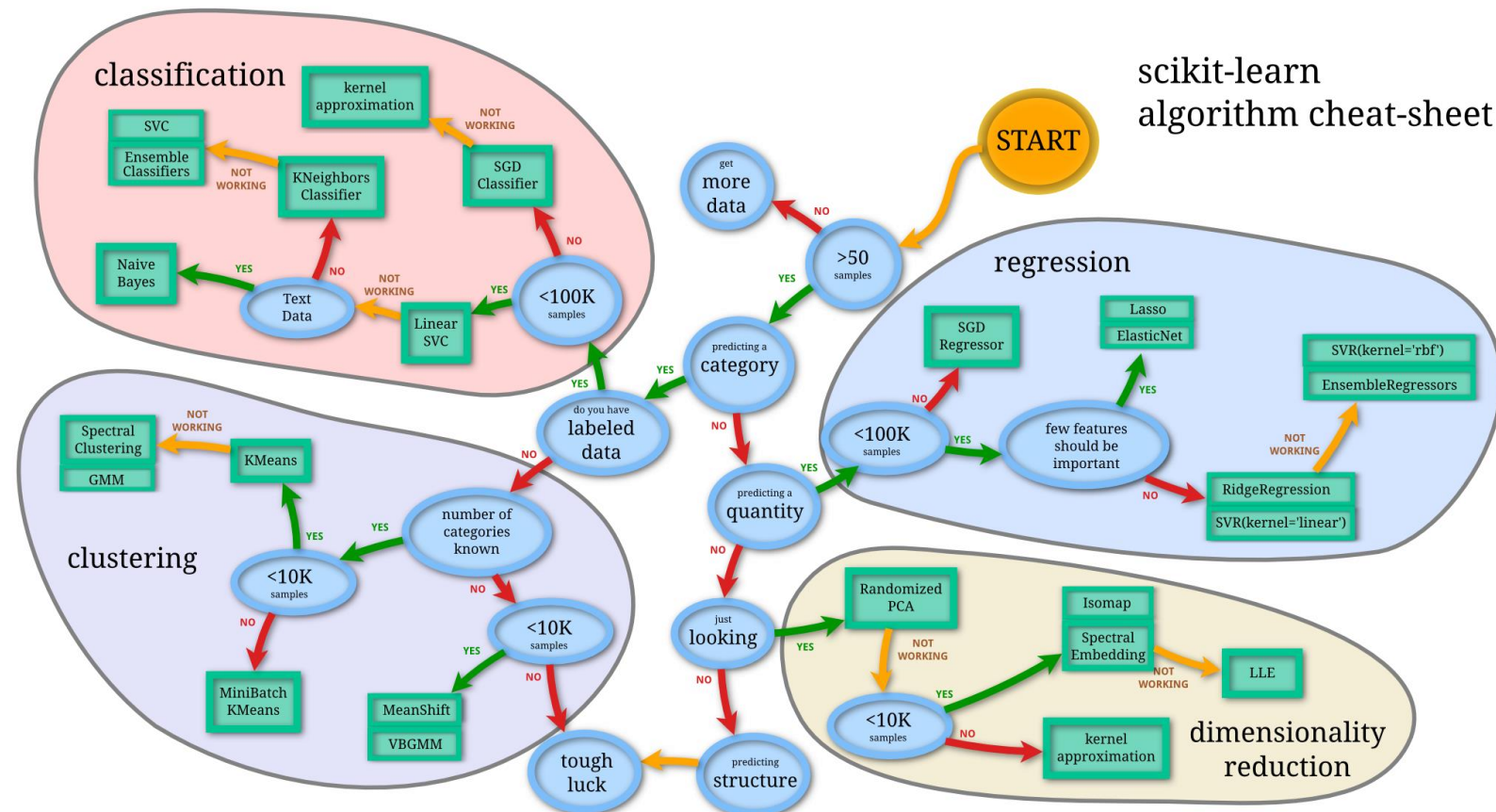
1/5



- 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

- Rotman***





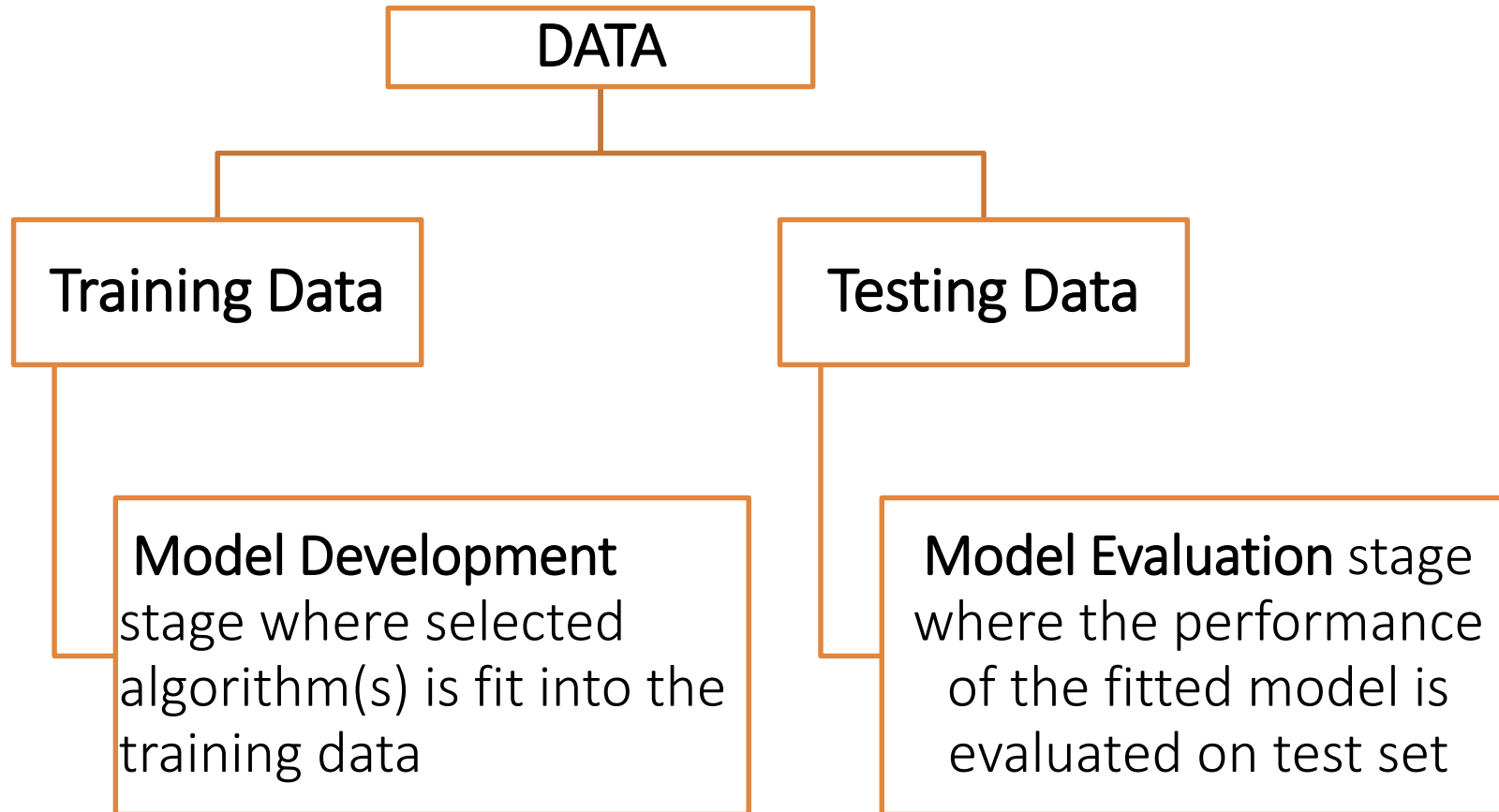


- **Sklearn offers numerous tools for**
  - **efficient data modelling**
  - **preprocessing support such as data encoding**
  - **feature selection / extraction**
  - **hyper-parameter search tools**
  - **end to end data modelling pipeline**

# 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.**

- **Machine Learning (ML) is a study of algorithms that can learn to solve a specified task using data.**
- **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.

# Installation

To install sklearn:

```
conda install -c anaconda 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:

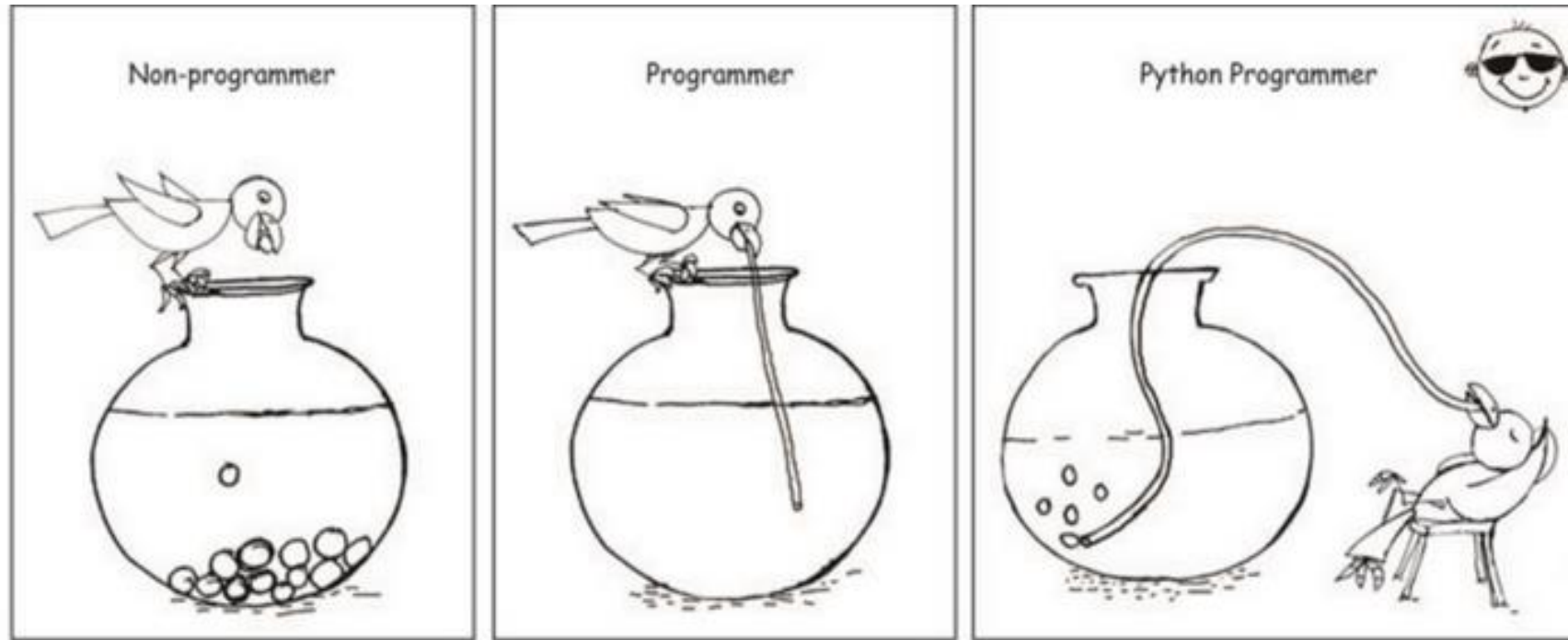
`conda list`



# Hands-on Implementation

- Go to [link]
- To open on google drive
  - Click on introtsklearn.ipynb to
  - Download data file and upload on google drive
  - Mount your google drive to the folder where the data is uploaded
- To open on your local jupyter notebook
  - Download introtsklearn.ipynb file
  - Download data file
  - Save both file in one folder
  - Open jupyter notebook

- Tutorials:
  - Quick Start Tutorial <http://scikit-learn.org/stable/tutorial/basic/tutorial.html>
  - User Guide [http://scikit-learn.org/stable/user\\_guide.html](http://scikit-learn.org/stable/user_guide.html)
  - API Reference <http://scikit-learn.org/stable/modules/classes.html>
  - Example Gallery [http://scikit-learn.org/stable/auto\\_examples/index.html](http://scikit-learn.org/stable/auto_examples/index.html)
  - [PyCon 2014 Scikit-learn Tutorial](#) by Jake VanderPlas
  - [Parallel Machine Learning with scikit-learn and IPython](#) by Olivier Grisel (also offered at Strata 2014)
- Books:
  - [Learning scikit-learn: Machine Learning in Python](#) (2013)
  - [Building Machine Learning Systems with Python](#) (2013)
  - [Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data](#) (2014).



Who wants to become a Python Programmer?

# Questions?

Thank you