

Machine Learning with Python Scikit-Learn: Part I

Research Computing Services
Information Services & Technology



RCS Team and Expertise (help@scc.bu.edu)

What We Do?



Maintains and administers the
Shared Computing Cluster
(SCC)

Located in Holyoke, MA

29,000 cores running Linux

400 GPUs

14 petabytes of storage for
research data

Our Team



Scientific Programmers

Systems Administrators

Graphics/Visualization
Specialists

Account/Project Managers

Special Initiatives (Grants)

Our Consulting Focus



AI & Data Science

Bioinformatics

Data Analysis/Statistics

Molecular modeling

Geographic Information Systems

Scientific/Engineering Simulation

Visualization and more



After the tutorial



Navigate to <http://rcs.bu.edu/eval>



Tutorial feedback survey

Plan of the Presentation



What is Machine Learning?



What is Scikit-Learn?



ML Model building:

Linear-regression: curve fitting

Linear Regression: Predicting Bicycle Traffic

Classification: Iris dataset



Conclusion & Next Steps

What is Machine Learning?



Example 1: **Predicting house prices** based on features like square footage, number of bedrooms, and location.



Example 2: **Weather prediction** for the next day based on historical weather data.



Example 3: **Email spam detection**, where emails are classified as "Spam" or "Not Spam."



Example 4: **Diagnosing diseases** based on patient symptoms and medical test results.



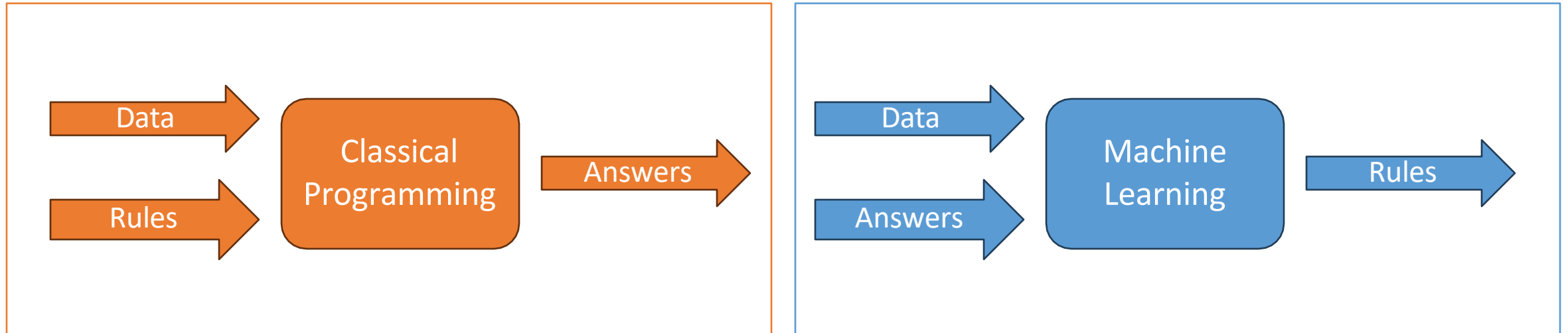
Example 5: **Grouping news articles by topic** using textual similarity (e.g., sports, politics, technology).



Example 6: **Compressing image datasets** for faster processing by retaining only the most important pixel information.

In machine learning, the model is not explicitly defined by predetermined rules or hardcoded logic. Instead, it learns patterns and relationships from data during the model training process.

What is Machine Learning?



What is Machine Learning?



Classical Programming



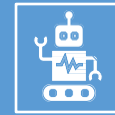
Offers complete control over logic and behavior



Defining, updating and maintaining explicit rules is difficult



Struggles with complex problems



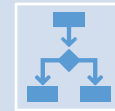
Machine Learning



Excels at handling complex tasks by learning patterns from data

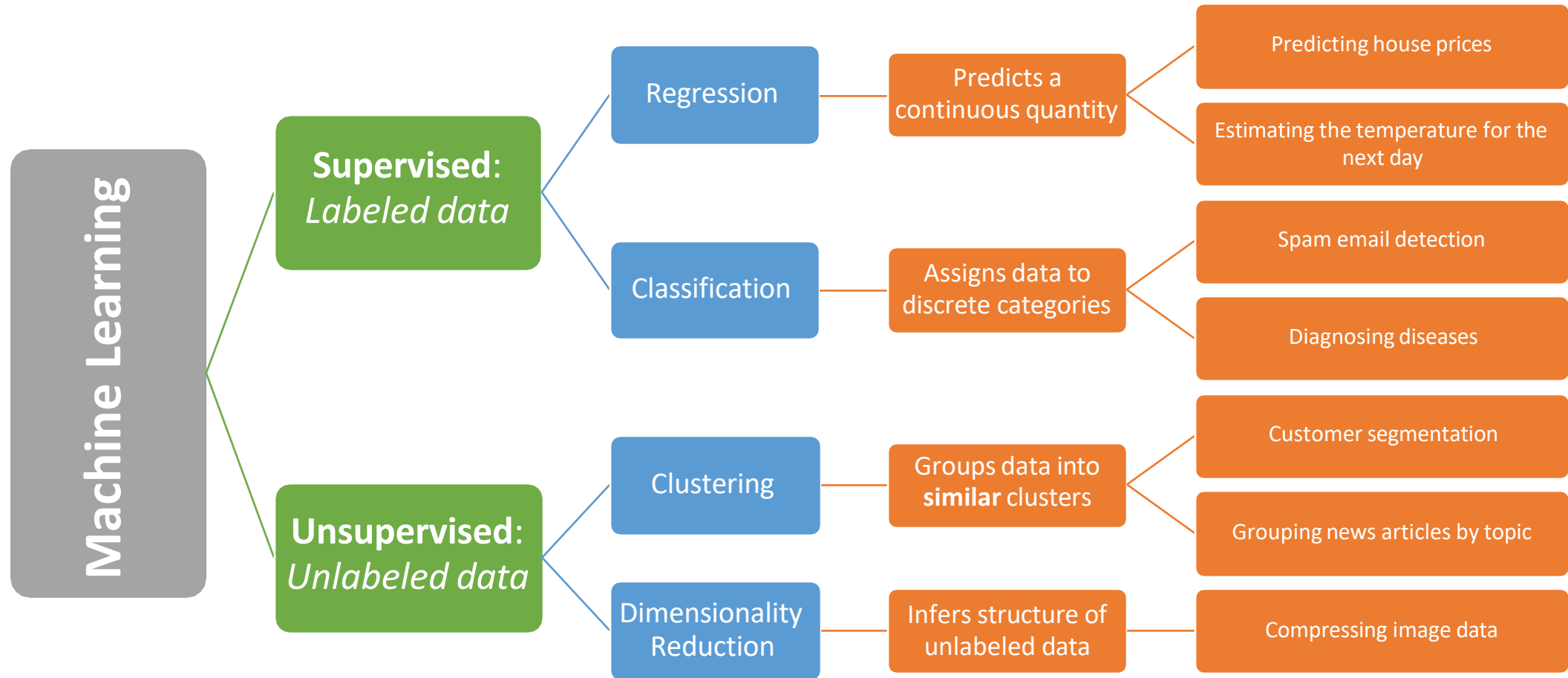


Requires large amounts of data and computational resources



Model's decision-making can be difficult to interpret. Black-box?

Branches of Machine Learning



What is Scikit-Learn?



Scikit-learn is an open-source Python library offering a wide range of ML models with simple interface and robust implementations.



Versatility: Supports classification, regression, clustering, dimensionality reduction, and more.



Easy-to-Use API: Uniform and intuitive API simplifies implementation and deployment.



Easy Integration: Works seamlessly with libraries like NumPy, pandas, and Matplotlib.

Advantages of Using Scikit-Learn



Streamlined and Uniform API

Consistent interface across all models simplifies learning and usage.



Cross-Platform

Supports macOS, Windows, and Linux.



Support for Installation

Easy installation on various platforms using pip or conda.



Detailed Online Documentation

Comprehensive guides, tutorials, and examples for every functionality.



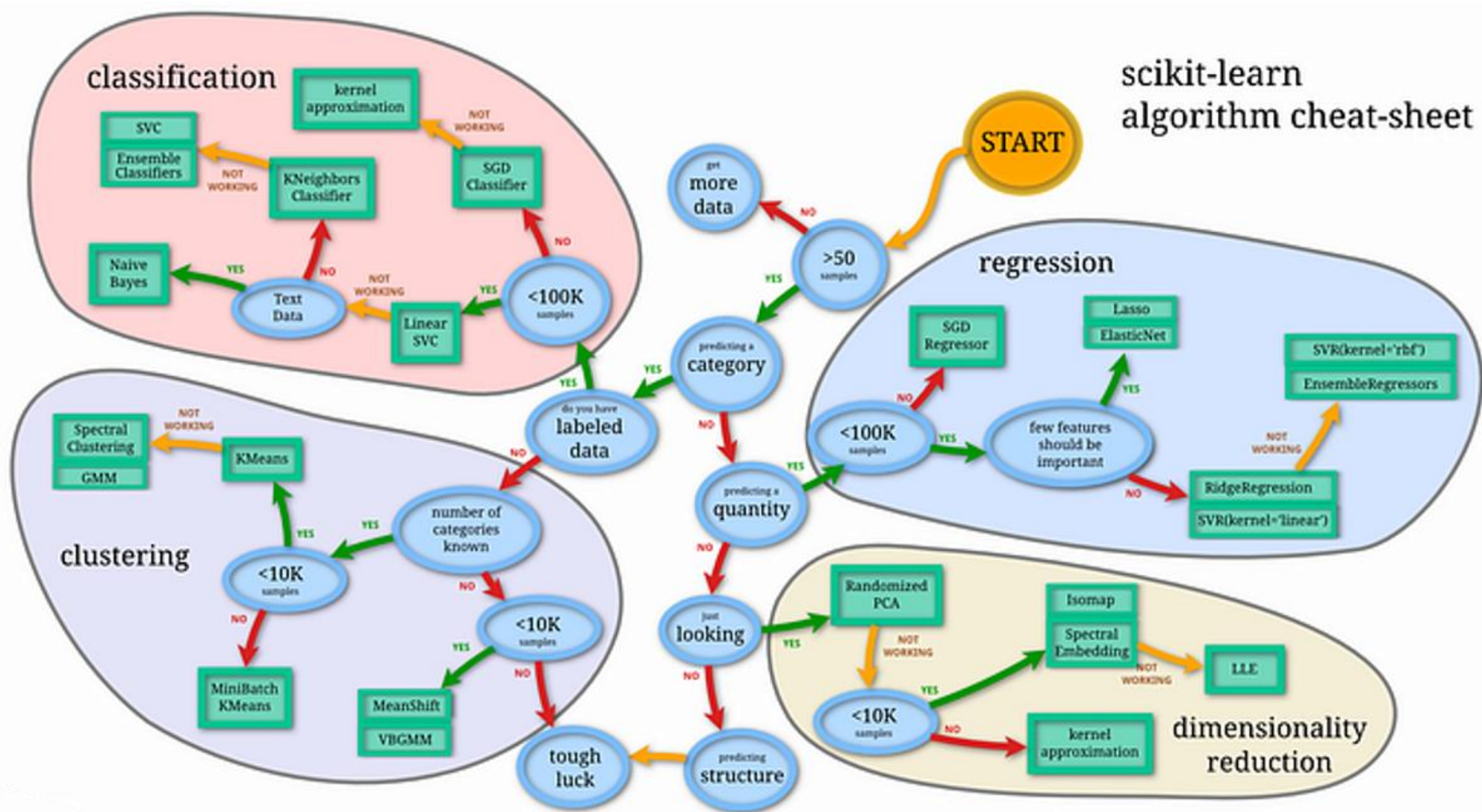
Active Development and Community

Frequent updates, new features, and a vibrant user community.



Scalable

Works with small datasets and scales to larger datasets when combined with tools like Dask.



Linear Regression

Linear Regression Example-1

Problem Statement

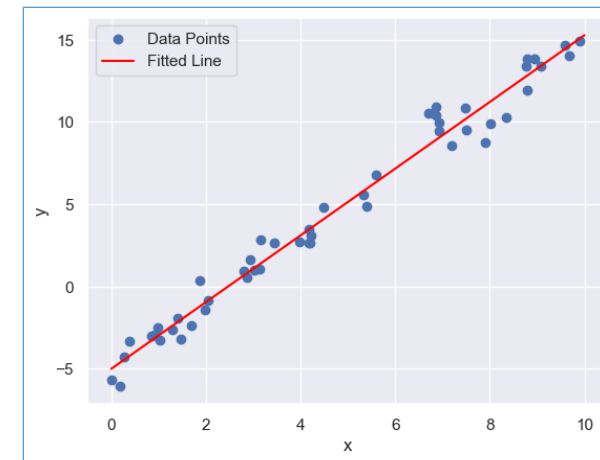
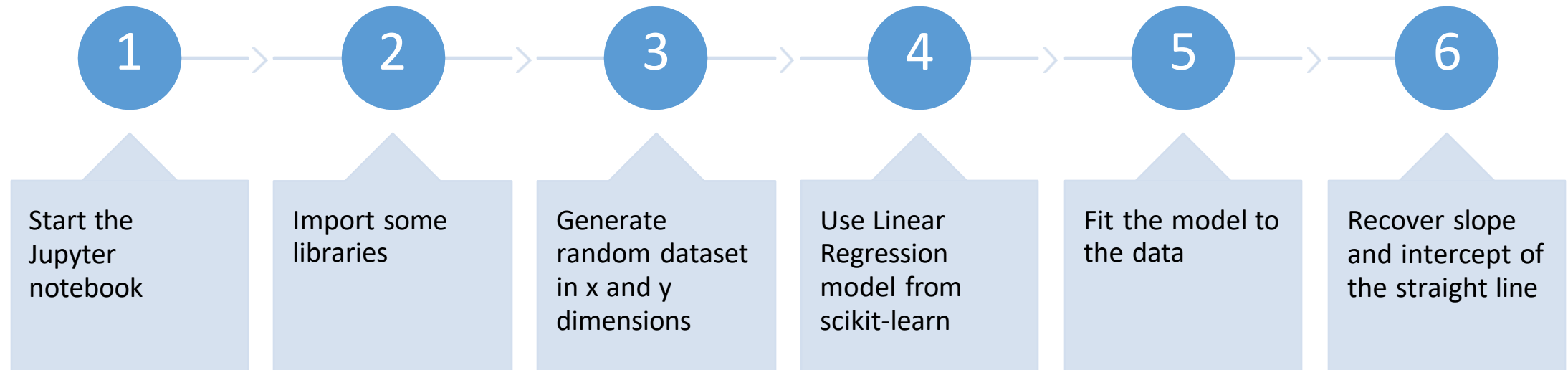
Fit a curve through data points to model the underlying pattern or trend.

Objectives

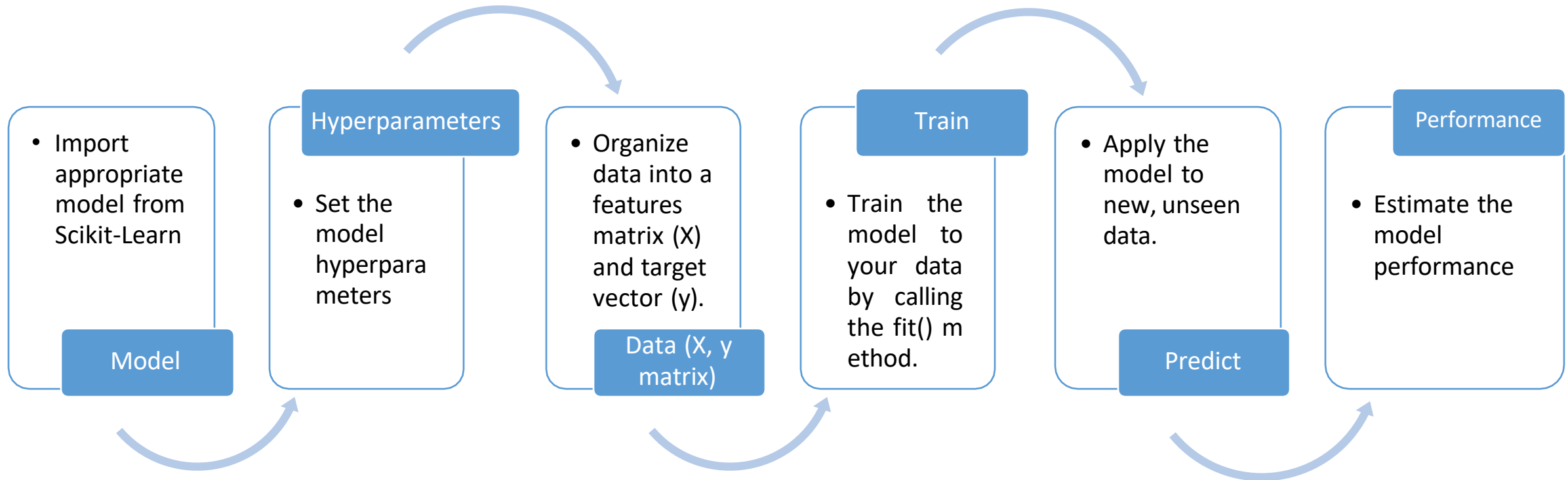
- Select and implement a ML model from scikit-learn
- Evaluate model performance
- Visualize results
- Make predictions



Plan: Linear Regression Example-1



Some Concepts in Machine Learning



Data Representation in Scikit-Learn

The easiest way to think about data within Scikit-Learn is in terms of tables.

Feature Matrix (X)
n_features $\rightarrow$

$\leftarrow$ n_samples					

Target Vector (y)

$\leftarrow$ n_samples	

Feature Matrix (X):

- 2D array, with shape $[n_samples, n_features]$
- NumPy array or a Pandas DataFrame or SciPy sparse matrices, Python lists
- Indexed: integers, dates, strings

Target Matrix (y):

- 1D or 2D array, with shape $[n_samples]$
- NumPy array or a Pandas DataFrame or Python lists.

Feature names and row indices are not used for computation.

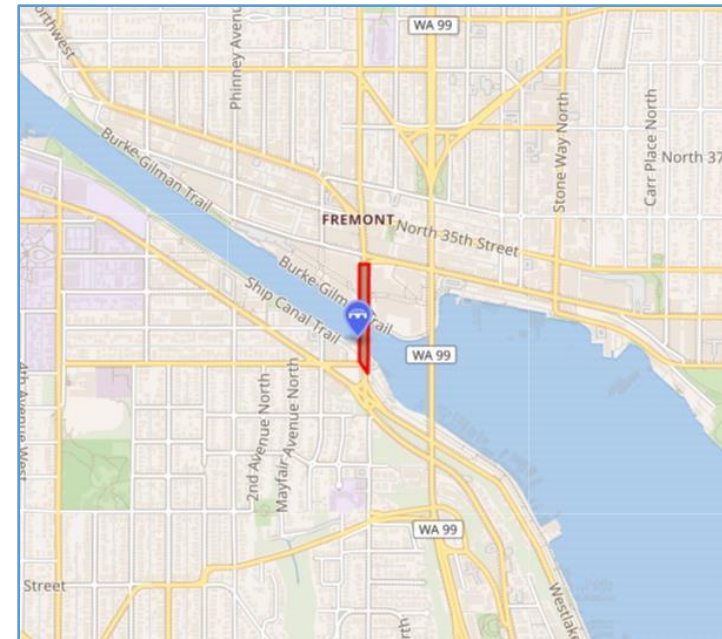
Linear regression: Predicting Bicycle Traffic

Problem Statement

Build a ML model to forecast bicycle traffic based on various influencing factors.

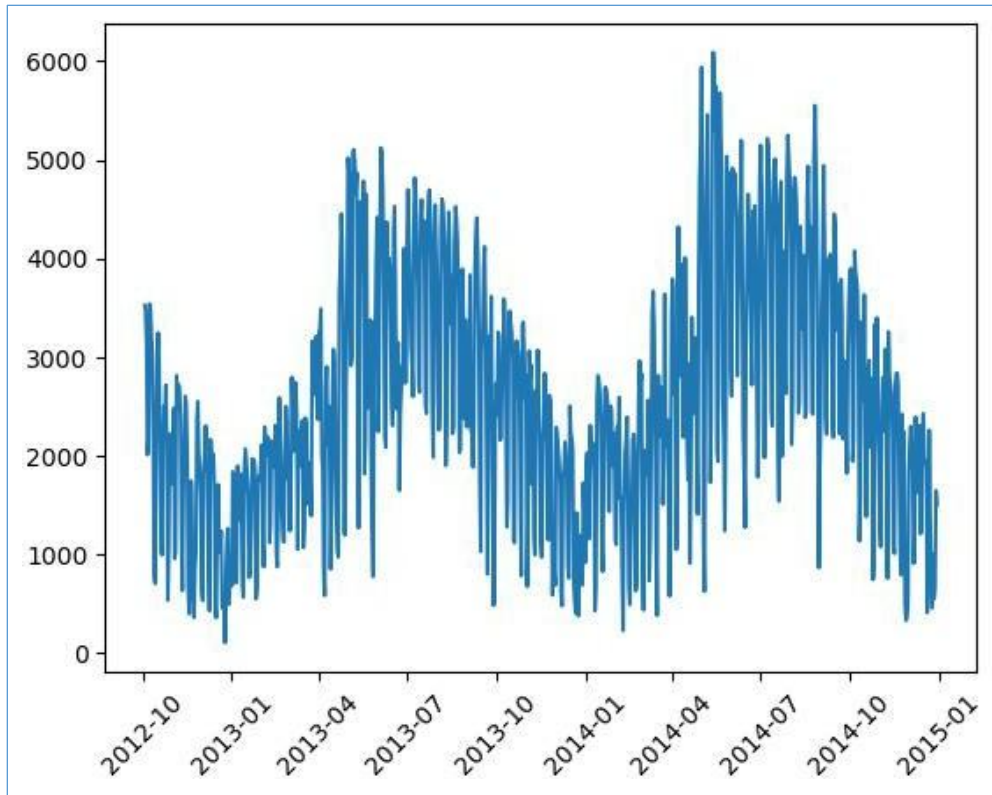
Objectives

- Visualize feature importance to understand the factors influencing bicycle traffic
- Select and implement a ML model from scikit-learn
- Evaluate model performance
- Visualize results
- Make predictions



The Fremont Bridge in Seattle was opened on Friday June 15, 1917. A bicycle counter was installed in 2012. Bike traffic reached 1.12 million rides in 2019

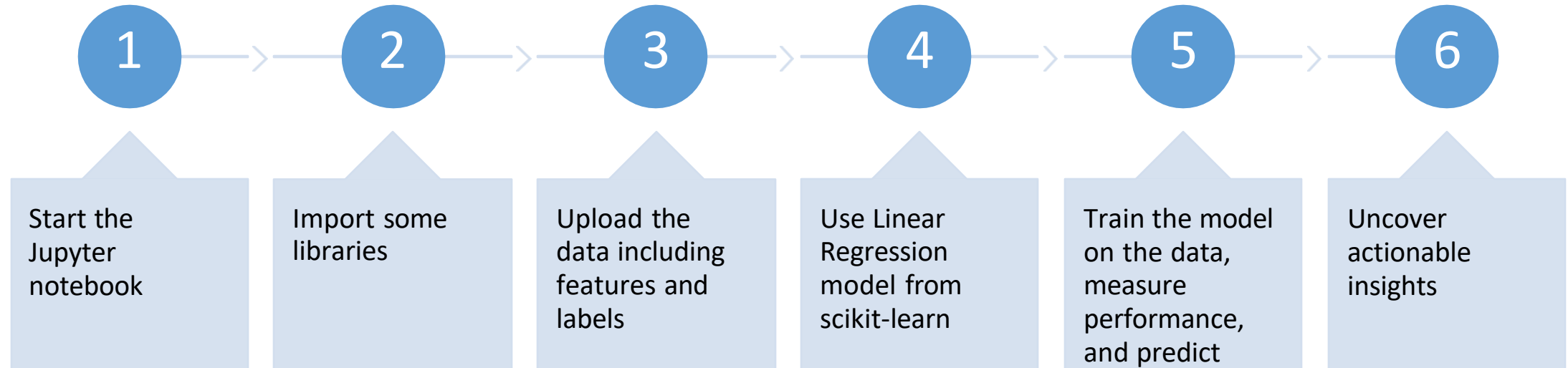
Linear regression: Predicting Bicycle Traffic



Feature Engineering: Food for Thought

What factors might be affecting the number of bike riders in the city of Seattle?

Plan: Linear regression: Predicting Bicycle Traffic



Conclusion and Next Steps

What is Machine Learning?

How is it different from classical programming?

What is Scikit-Learn?

Linear-regression: curve fitting

Linear Regression: Predicting Bicycle Traffic

Classification using Decision Trees and Random Forests

Dimensionality Reduction using Principal Component Analysis

Clustering using K-Means

Where to go next with Machine Learning?

Part II: Thursday, 1pm