



Machine Learning Overview



Machine Learning

- It is finally time to dive deep into Machine Learning!
- This Machine Learning Overview section is designed to help get us in the correct frame of mind for the paradigm shift to Machine Learning.
- First, let's quickly review where we are in the Machine Learning Pathway....



ML Pathway



**Real
World**

**Problem
to Solve**

**Question
to
Answer**



ML Pathway



**Real
World**

**Problem
to Solve**

How to fix or change X?

**Question
to
Answer**

How does a change in X affect Y?



ML Pathway



**Real
World**

**Problem
to Solve**

How to fix or change X?

**Question
to
Answer**

How does a change in X affect Y?



ML Pathway



**Real
World**

**Data
Product**

**Data
Analysis**



ML Pathway



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

**Exploratory
Data
Analysis**



ML Pathway



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

**Exploratory
Data
Analysis**

**Machine
Learning
Models**

Supervised Learning:

Predict an Outcome

Unsupervised Learning:

Discover Patterns in Data



ML Pathway



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

**Exploratory
Data
Analysis**

**Machine
Learning
Models**

Supervised Learning:

Predict an Outcome

Unsupervised Learning:

Discover Patterns in Data



Machine Learning

- Our main goals in ML Overview section:
 - Problems solved by Machine Learning
 - Types of Machine Learning
 - Supervised Learning
 - Unsupervised Learning
 - ML Process for Supervised Learning
 - Discussion on Companion Book



Machine Learning

- Our main goals in ML Overview section:
 - No coding in this section!
 - Purely a discussion on critically important ideas applied to ML problems.



Machine Learning

- Many other relevant topics will be discussed later in the course as we “discover” them, including:
 - Bias-Variance Trade-off
 - Cross-validation
 - Feature Engineering
 - Scikit-learn
 - Performance Metrics and much



Machine Learning

- Machine Learning Sections
 - Section for Type of Algorithm
 - Intuition and Mathematical Theory
 - Example code-along of application of Algorithm
 - Expansion of Algorithm
 - Project Exercise
 - Project Exercise Solution



Machine Learning

- Machine Learning Sections
 - Exception for Linear Regression
 - Intuition and Mathematical Theory
 - Simple Linear Regression
 - Scikit-learn and Linear Regression
 - Regularization
 - “Discovering” additional ML topics



Machine Learning

- Machine Learning Sections
 - “Discovering” additional ML topics
 - Performance Metrics
 - Feature Engineering
 - Cross-validation
 - Revisit Linear Regression to combine discovered ML ideas for Project Exercise.



Machine Learning

- Let's continue by starting to understand why we use machine learning and the use cases for it!



Why Machine Learning?



Machine Learning

- Machine learning in general is the study of statistical computer algorithms that improve automatically through data.
- This means unlike typical computer algorithms that rely on human input for what approach to take, ML algorithms infer best approach from the data itself.



Machine Learning

- Machine learning is a subset of Artificial Intelligence.
- ML algorithms are not explicitly programmed on which decisions to make.
- Instead the algorithm is designed to infer from the data the most optimal choices to make.



Machine Learning

- What kinds of problems can ML solve?
 - Credit Scoring
 - Insurance Risk
 - Price Forecasting
 - Spam Filtering
 - Customer Segmentation
 - Much more!



Machine Learning

- Structure of ML Problem framing:
 - Given **features** from a data set **obtain** a desired **label**.
 - ML algorithms are often called “estimators” since they are estimating the desired **label** or output.



Machine Learning

- How can ML be so robust in solving all sorts of problems?
- Machine learning algorithms rely on data and a set of statistical methods to learn what features are important in data.



Machine Learning

- Simple Example:
 - Predict the price a house should sell at given its current features (Area,Bedrooms,Bathrooms,etc...)



Machine Learning

- House Price Prediction
 - Typical Algorithm
 - Human user defines an algorithm to manually set values of importance for each feature.



Machine Learning

- House Price Prediction
 - ML Algorithm
 - Algorithm automatically determines importance of each feature from existing data



Machine Learning

- Why machine learning?
 - Many complex problems are only solvable with machine learning techniques.
 - Problems such as spam email or handwriting identification require ML for an effective solution.



Machine Learning

- Why not just use machine learning for everything?
 - Major caveat to effective ML is good data.
 - Majority of development time is spent cleaning and organizing data, **not** implementing ML algorithms.



Machine Learning

- Do we develop our own ML algorithms?
 - Rare to have a need to manually develop and implement a new ML algorithm, since these techniques are well documented and developed.



Machine Learning

- Let's continue this discussion by exploring the types of machine learning algorithms!



Types of Machine Learning



Machine Learning

- There are two main types of Machine Learning we will cover in upcoming sections:
 - Supervised Learning
 - Unsupervised Learning



Machine Learning

- Supervised Learning
 - Using **historical** and **labeled** data, the machine learning model predicts a value.
- Unsupervised Learning
 - Applied to **unlabeled** data, the machine learning model discovers possible patterns in the data.



Machine Learning

- Supervised Learning
 - Requires **historical labeled** data:
 - Historical
 - Known results and data from the past.
 - Labeled
 - The desired output is known.



Machine Learning

- Supervised Learning
 - Two main label types
 - Categorical Value to Predict
 - Classification Task
 - Continuous Value to Predict
 - Regression Task



Machine Learning

- Supervised Learning
 - Classification Tasks
 - Predict an assigned category
 - Cancerous vs. Benign Tumor
 - Fulfillment vs. Credit Default
 - Assigning Image Category
 - Handwriting Recognition



Machine Learning

- Supervised Learning
 - Regression Tasks
 - Predict a continuous value
 - Future prices
 - Electricity loads
 - Test scores



Machine Learning

- Unsupervised Learning
 - Group and interpret data without a label.
 - Example:
 - Clustering customers into separate groups based off their behaviour features.



Machine Learning

- Unsupervised Learning
 - Major downside is because there was no historical “correct” label, it is much harder to evaluate performance of an unsupervised learning algorithm.



Machine Learning

- Machine Learning Sections
 - We first focus on supervised learning to build an understanding of machine learning capabilities.
 - Then shift focus to unsupervised learning for clustering and dimensionality reduction.



Machine Learning

- Finally, before we dive into coding and linear regression in the next section, let's have a deep dive into the entire Supervised Machine Learning process to set ourselves up for success!



Supervised Machine Learning Process



Machine Learning

- Machine Learning Pathway



**Real
World**



**Collect &
Store
Data**



Machine Learning

- Machine Learning Pathway



**Real
World**



**Collect &
Store
Data**



**Clean &
Organize
Data**



Machine Learning

- Machine Learning Pathway



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

**Exploratory
Data
Analysis**

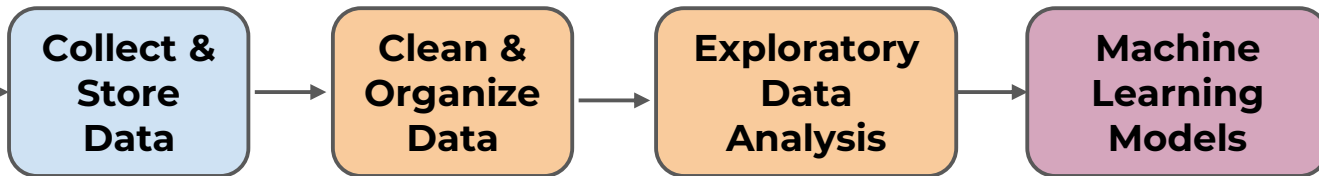


Machine Learning

- Machine Learning Pathway



**Real
World**



Supervised Learning:

Predict an Outcome

Unsupervised Learning:

Discover Patterns in Data



Machine Learning

- Machine Learning Pathway



Real
World

Collect &
Store
Data

Clean &
Organize
Data

Exploratory
Data
Analysis

Machine
Learning
Models

Jupyter, NumPy, Pandas, Matplotlib, Seaborn

Supervised Learning:

Predict an Outcome

Unsupervised Learning:

Discover Patterns in Data

Scikit-learn



Machine Learning

- Machine Learning Pathway



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

**Exploratory
Data
Analysis**

**Machine
Learning
Models**

Supervised Learning:
Predict an Outcome

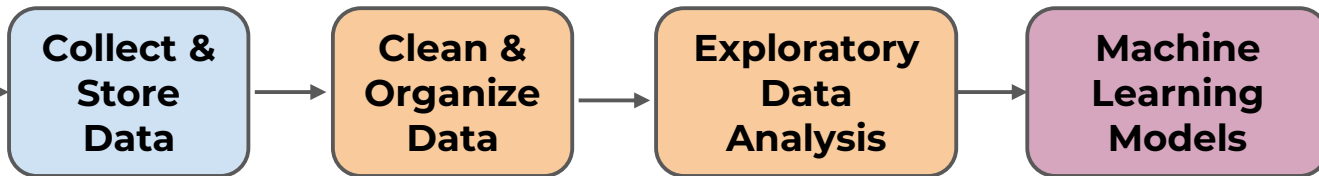


Machine Learning

- ML Process : Supervised Learning Tasks



**Real
World**



Supervised Learning:
Predict an Outcome

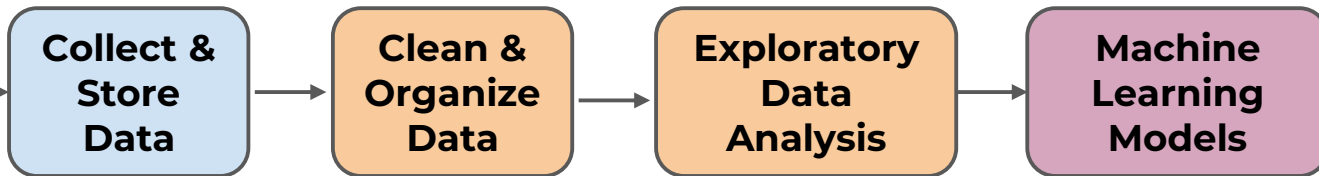


Machine Learning

- **Predict price a house should sell at.**



**Real
World**



Supervised Learning:
Predict an Outcome



Machine Learning

- **Supervised** Machine Learning Process
- Start with collecting and organizing a data set based on history:

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Historical labeled** data on previously sold houses.

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- If a new house comes on the market with a known Area, Bedrooms, and Bathrooms: *Predict what price should it sell at.*

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- Data Product:
 - Input house features
 - Output predicted selling price

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- Using **historical, labeled** data predict a future outcome or result.

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000

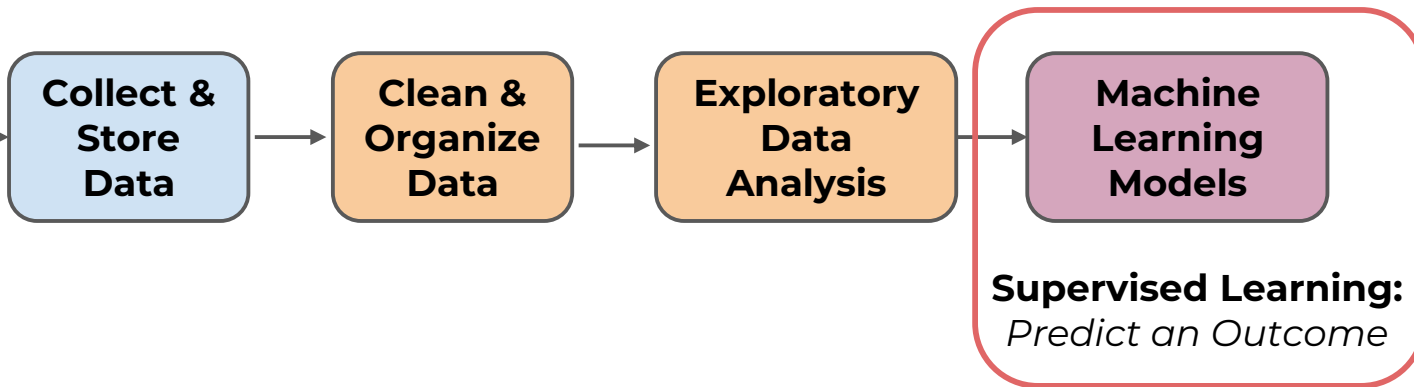


Machine Learning

- **Predict price a house should sell at.**



**Real
World**





Machine Learning

- **Predict price a house should sell at.**

**Machine
Learning
Models**

Supervised Learning:
Predict an Outcome



Machine Learning

- **Predict price a house should sell at.**

Machine Learning Models

Supervised Learning:
Predict an Outcome



Machine Learning

- **Predict price a house should sell at.**

Machine Learning Models

Supervised Learning:
Predict an Outcome

Data



Machine Learning

- **Supervised** Machine Learning Process

Data



Machine Learning

- **Supervised** Machine Learning Process

Data



X: Features
y: Label



Machine Learning

- **Supervised Machine Learning Process**

Data

X: Features
y: Label

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Label** is what we are trying to predict

Data

X: Features
y: Label

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Label** is what we are trying to predict

Data

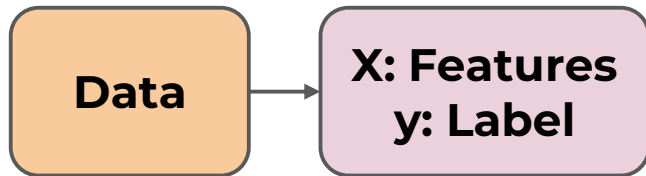
X: Features
y: Label

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Features** are known characteristics or components in the data

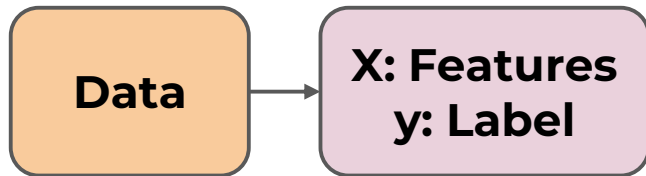


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Features** are known characteristics or components in the data

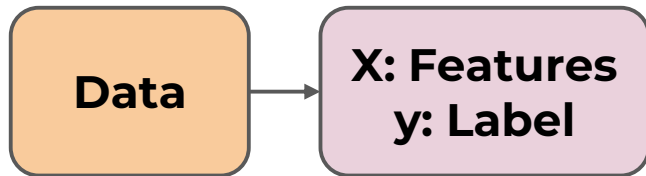


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Machine Learning

- **Features** and **Label** are identified according to the problem being solved.

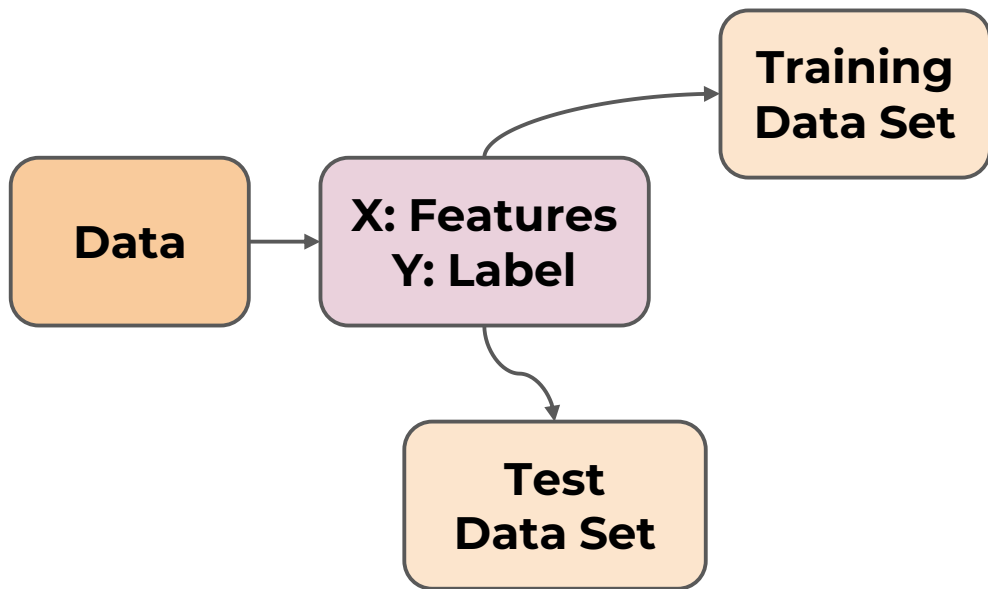


Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

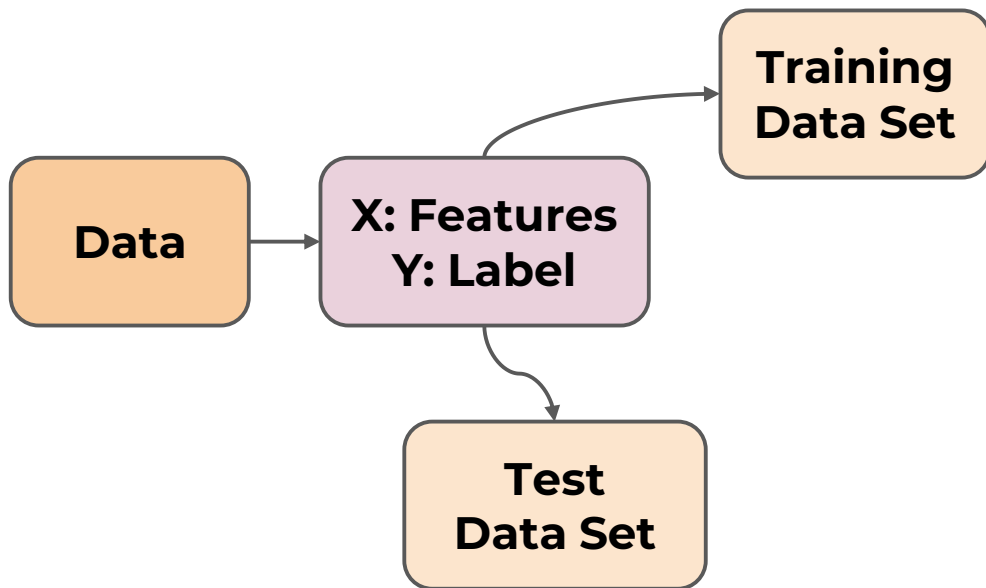
- Split data into training set and test set





Supervised Machine Learning Process

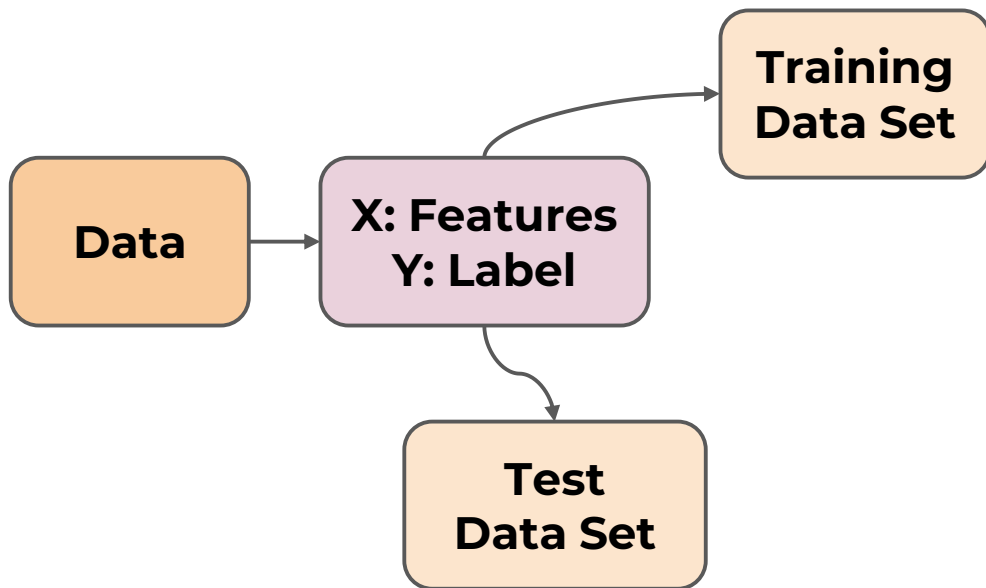
- Later on we will discuss cross-validation





Supervised Machine Learning Process

- Why perform this split? How to split?





Supervised Machine Learning Process

- Why perform this split? How to split?

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- How would you judge a human realtor's performance?



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Ask a human realtor to take a look at historical data...



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Then give her the features of a house and ask her to predict a selling price.



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- But how would you measure how accurate her prediction is? What house should you choose to test on?



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- You can't judge her based on a new house that hasn't sold yet, you don't know it's true selling price!



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- You shouldn't judge her on data she's already seen, she could have **memorized** it!



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Thus the need for a Train/Test split of the data, let's explore further...



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- We already organized the data into **Features (X)** and a **Label (y)**

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Now we will split this into a training set and a test set:

TRAIN

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Now we will split this into a training set and a test set:

	Area m ²	Bedrooms	Bathrooms	Price
TRAIN	200	3	2	\$500,000
	190	2	1	\$450,000
	230	3	3	\$650,000
TEST	180	1	1	\$400,000
	210	2	2	\$550,000



Supervised Machine Learning Process

- Notice how we have 4 components

		Area m ²	Bedrooms	Bathrooms	Price		
X TRAIN	Y TRAIN	200	3	2	\$500,000		
		190	2	1	\$450,000		
		230	3	3	\$650,000		
X TEST	Y TEST	180	1	1	\$400,000		
		210	2	2	\$550,000		



Supervised Machine Learning Process

- Let's go back to fairly testing our human realtor....



Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Let's go back to fairly testing our human realtor....



TRAIN

TEST

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000
180	1	1	\$400,000
210	2	2	\$550,000



Supervised Machine Learning Process

- Let her study and learn on the training set getting access to both X and y.



TRAIN

Area m ²	Bedrooms	Bathrooms	Price
200	3	2	\$500,000
190	2	1	\$450,000
230	3	3	\$650,000



Supervised Machine Learning Process

- After she has “learned” about the data, we can test her skill on the test set.



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Supervised Machine Learning Process

- Provide only the X test data and ask for her predictions for the sell price.



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Supervised Machine Learning Process

- This is new data she has never seen before! She has also never seen the real sold price.



TEST

Area m ²	Bedrooms	Bathrooms
180	1	1
210	2	2



Supervised Machine Learning Process

- Ask for predictions per data point.



Predictions	Area m ²	Bedrooms	Bathrooms
\$410,000	180	1	1
\$540,000	210	2	2



Supervised Machine Learning Process

- Then bring back the original prices.



Predictions	Area m ²	Bedrooms	Bathrooms	Price
\$410,000	180	1	1	\$400,000
\$540,000	210	2	2	\$550,000



Supervised Machine Learning Process

- Finally compare predictions against true test price.



Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Supervised Machine Learning Process

- This is often labeled as \hat{y} compared against y



\hat{y}	y
Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Supervised Machine Learning Process

- Later on we will discuss the many methods of evaluating this performance!

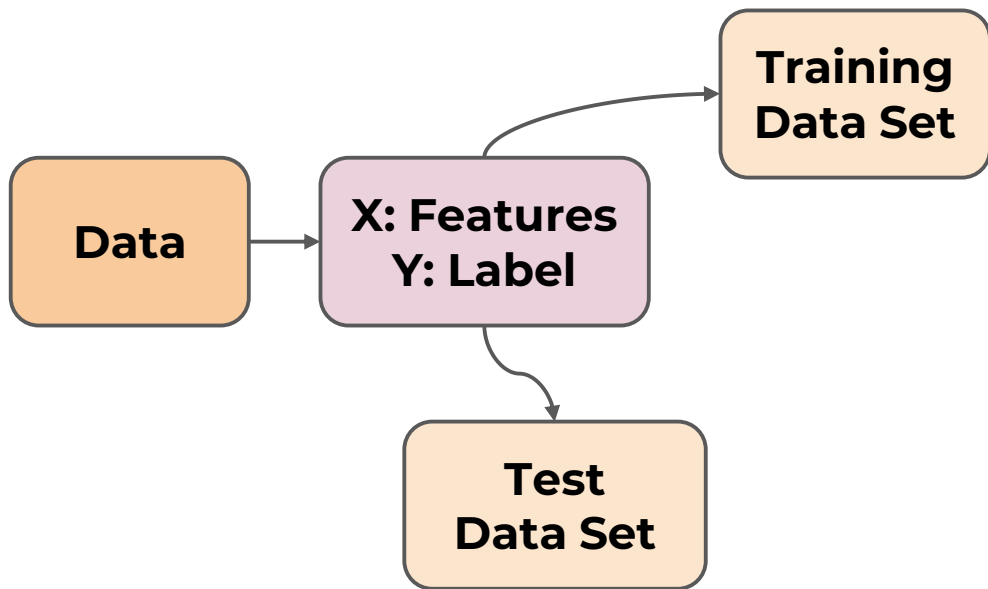


Predictions	Price
\$410,000	\$400,000
\$540,000	\$550,000



Supervised Machine Learning Process

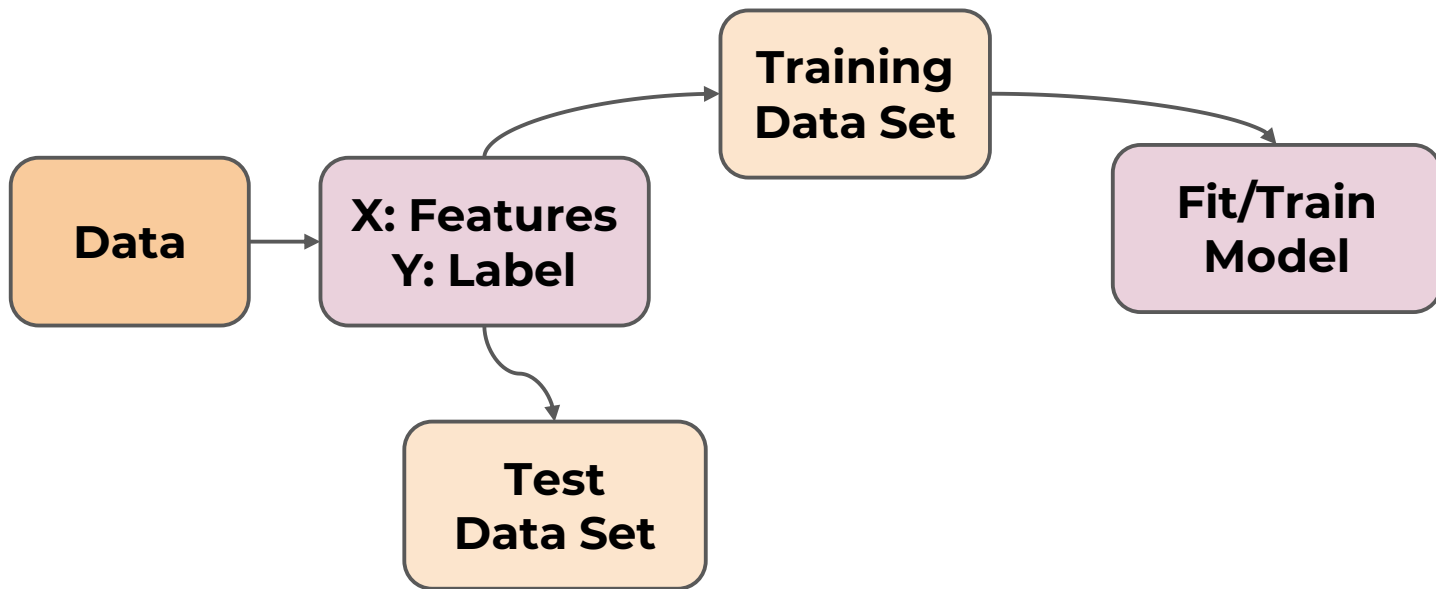
- Split Data





Supervised Machine Learning Process

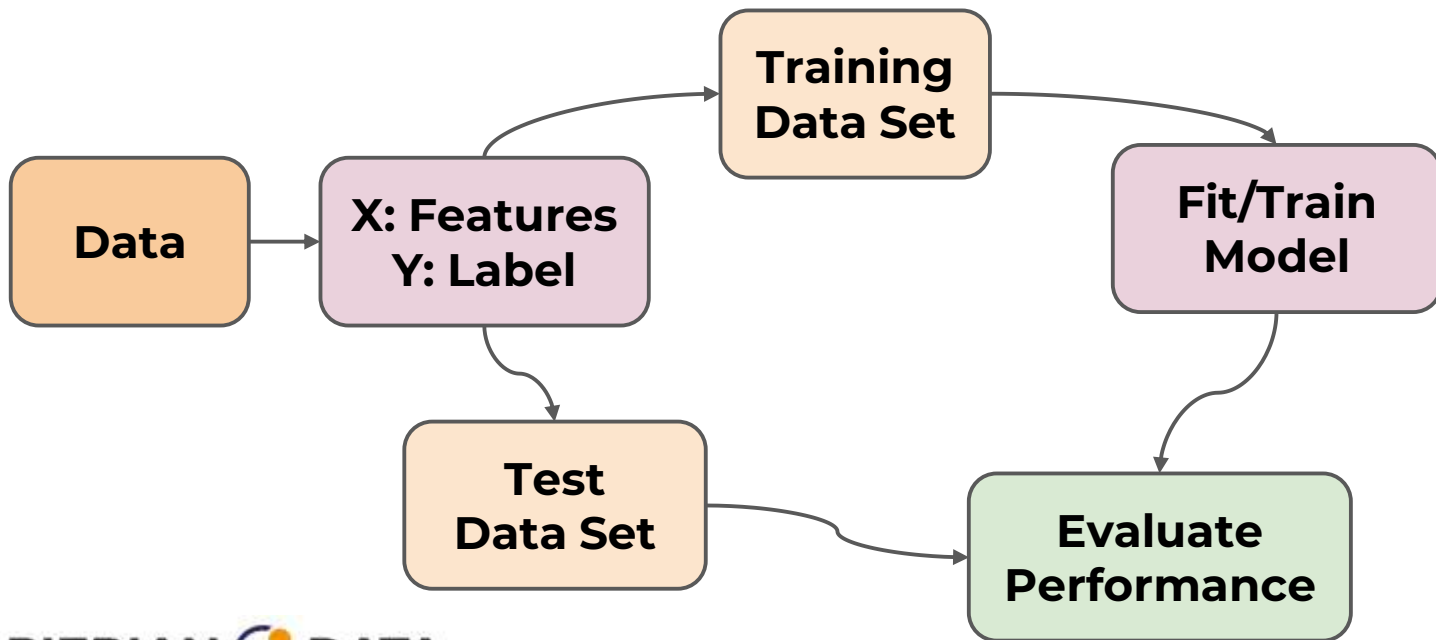
- Split Data, Fit on Train Data





Supervised Machine Learning Process

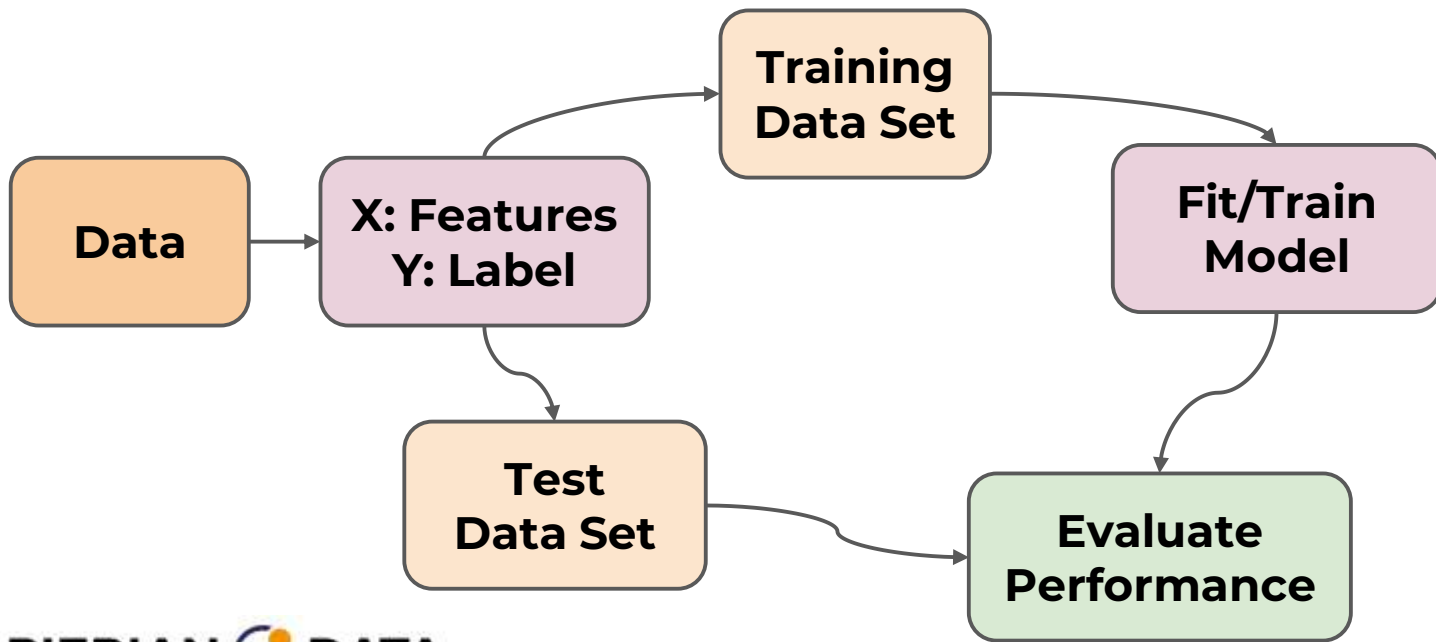
- Split Data, Fit on Train Data, Evaluate Model





Supervised Machine Learning Process

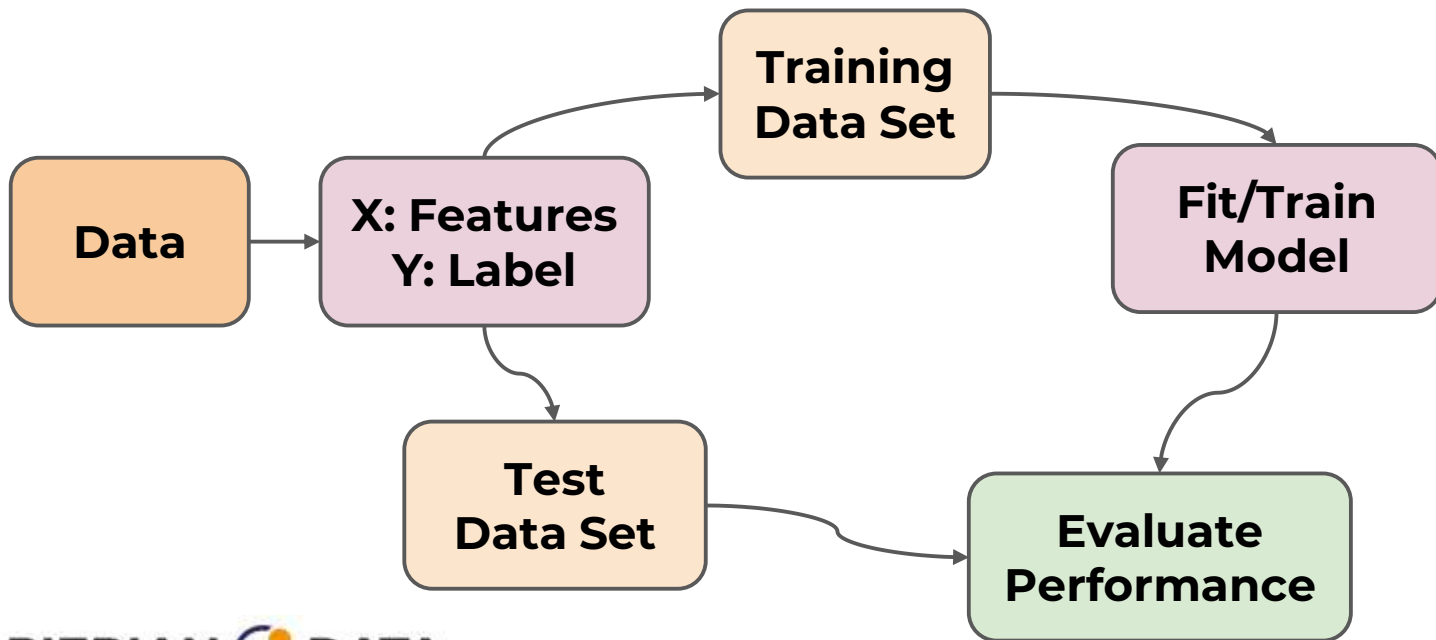
- What happens if performance isn't great?





Supervised Machine Learning Process

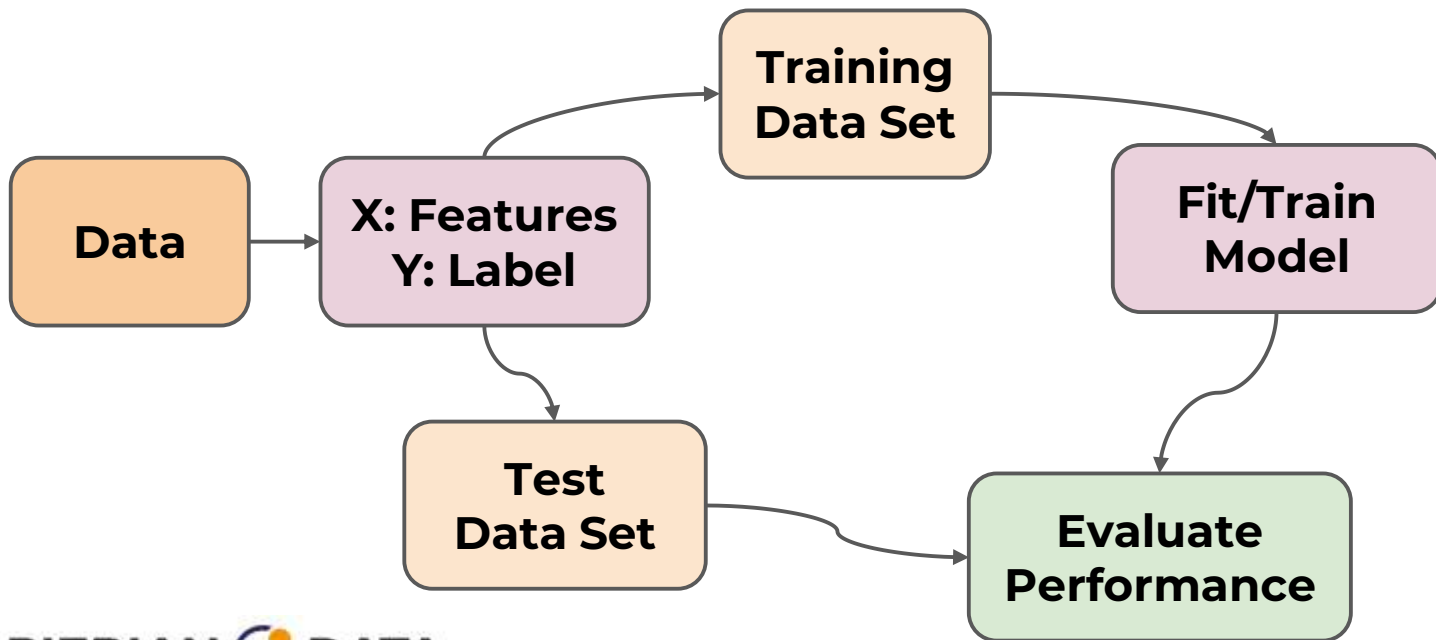
- We can adjust model **hyperparameters**





Supervised Machine Learning Process

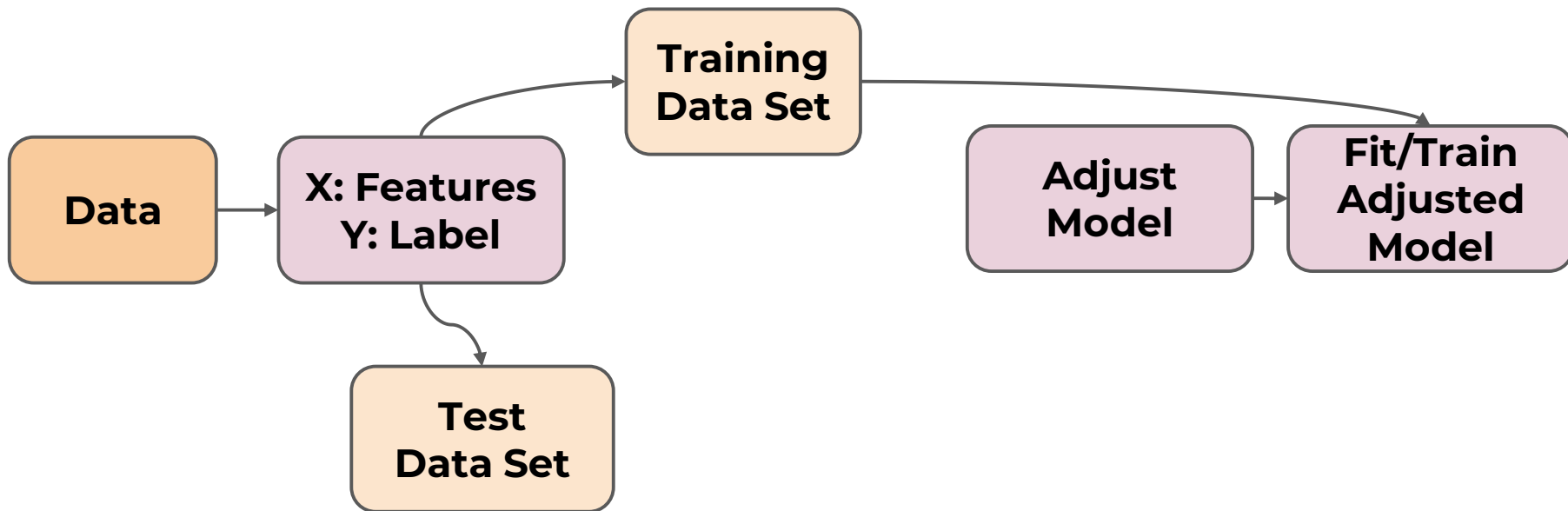
- Many algorithms have adjustable values





Supervised Machine Learning Process

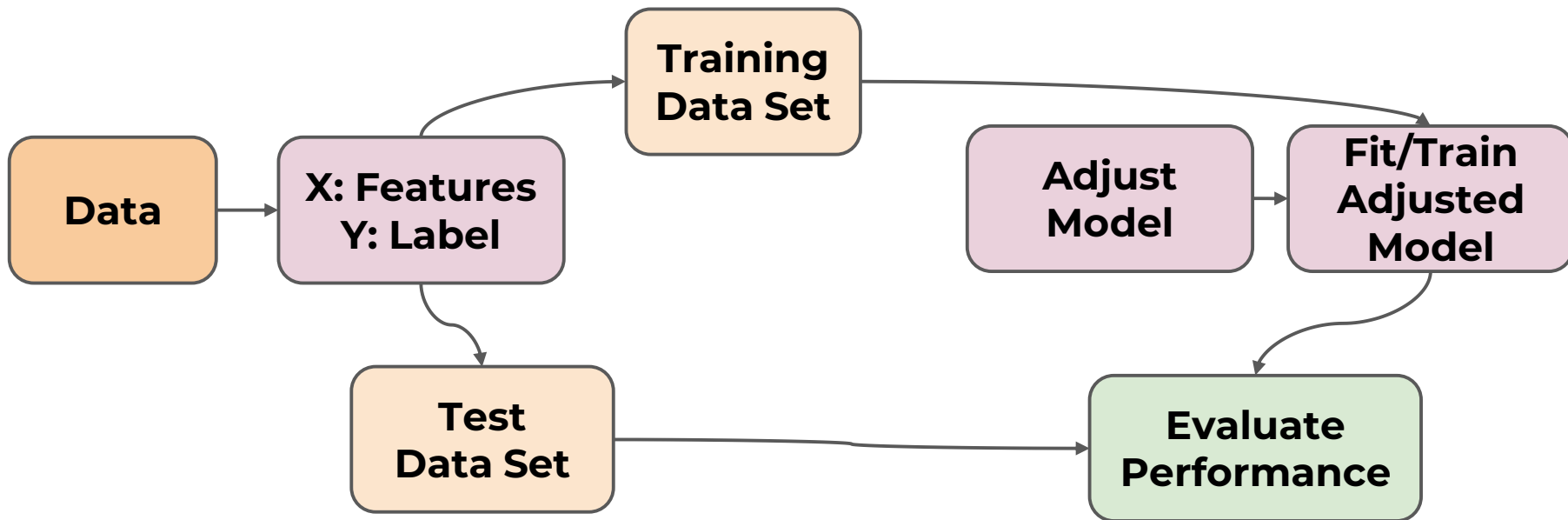
- Many algorithms have adjustable values





Supervised Machine Learning Process

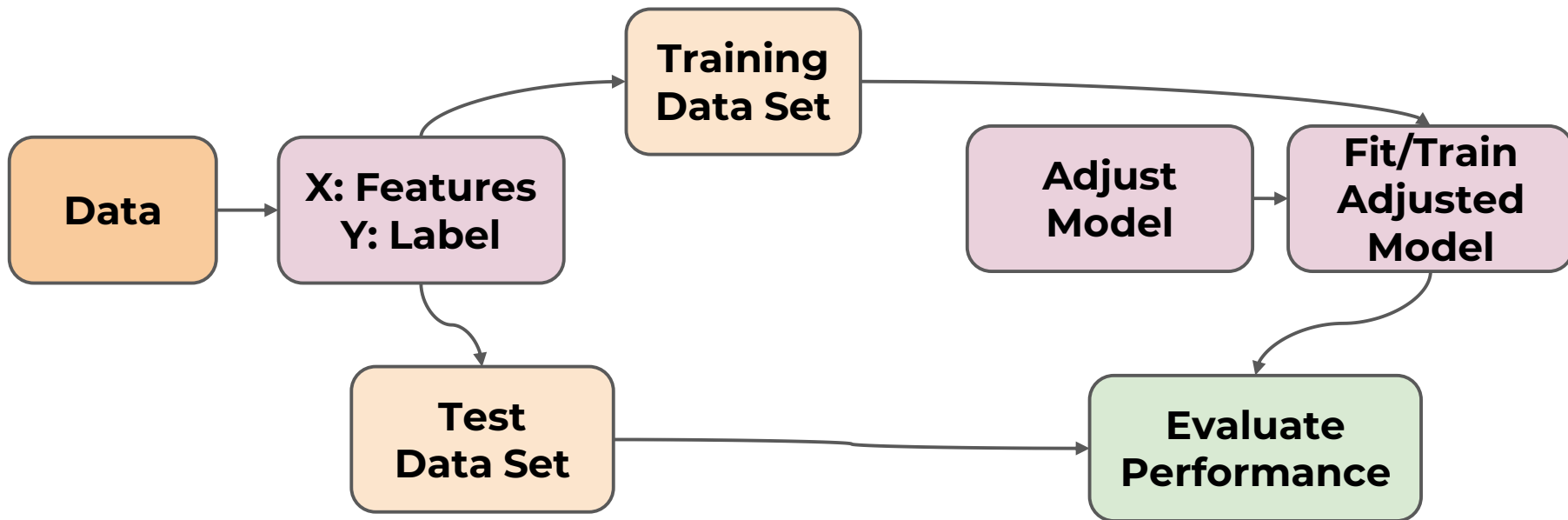
- Evaluate adjusted model





Supervised Machine Learning Process

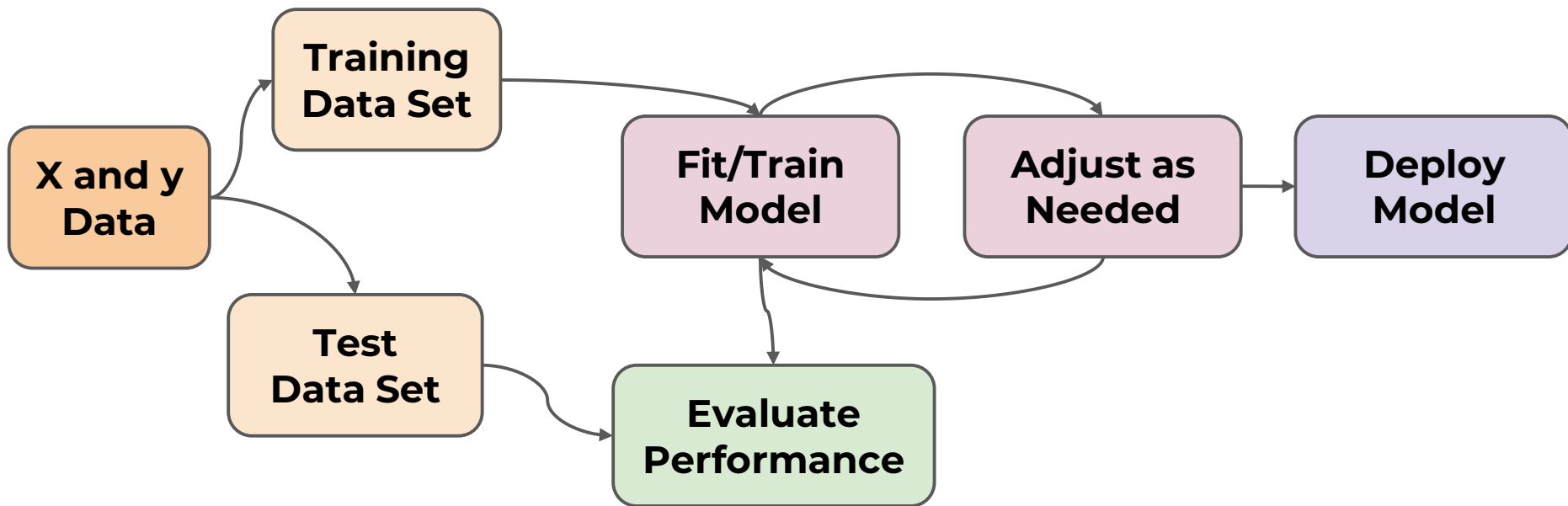
- Can repeat this process as necessary





Supervised Machine Learning Process

- Full and Simplified Process





Supervised Machine Learning Process

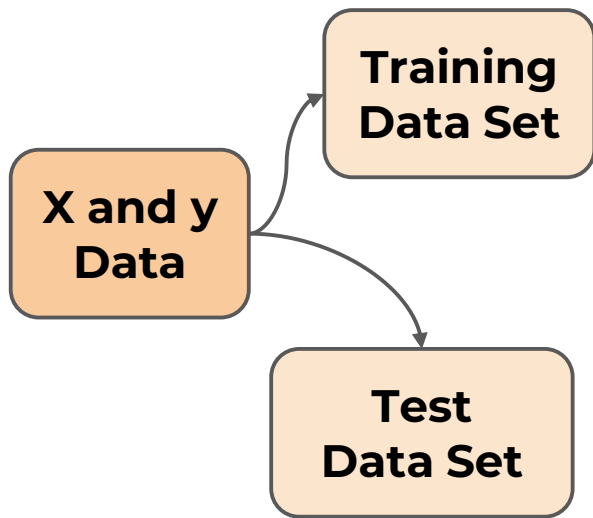
- Get X and y data

**X and y
Data**



Supervised Machine Learning Process

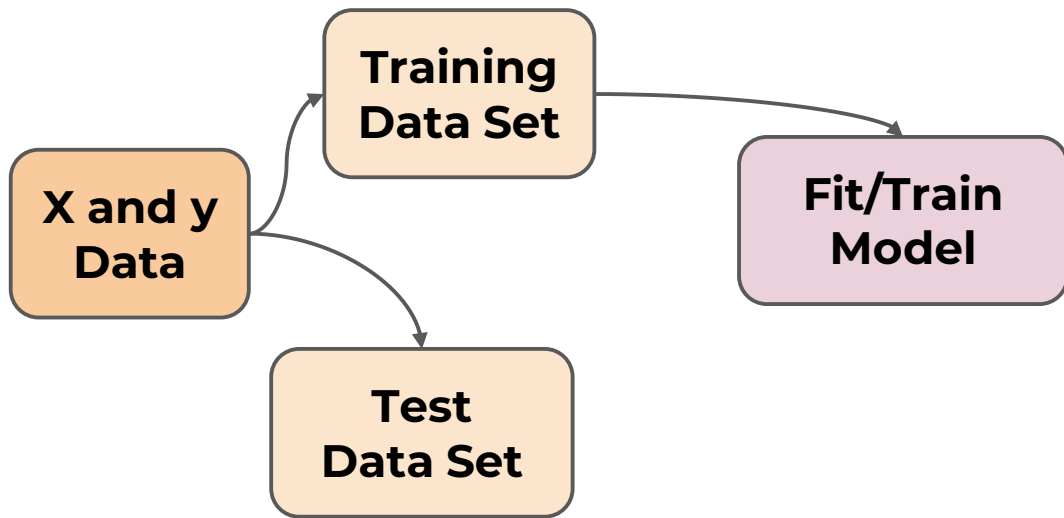
- Split data for evaluation purposes





Supervised Machine Learning Process

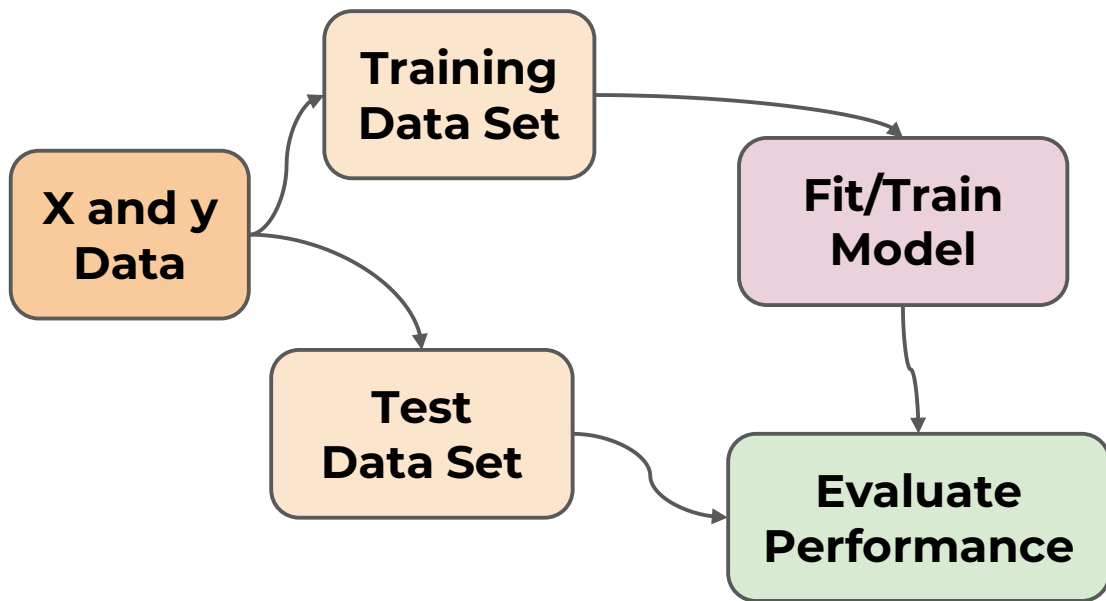
- Fit ML Model on Training Data Set





Supervised Machine Learning Process

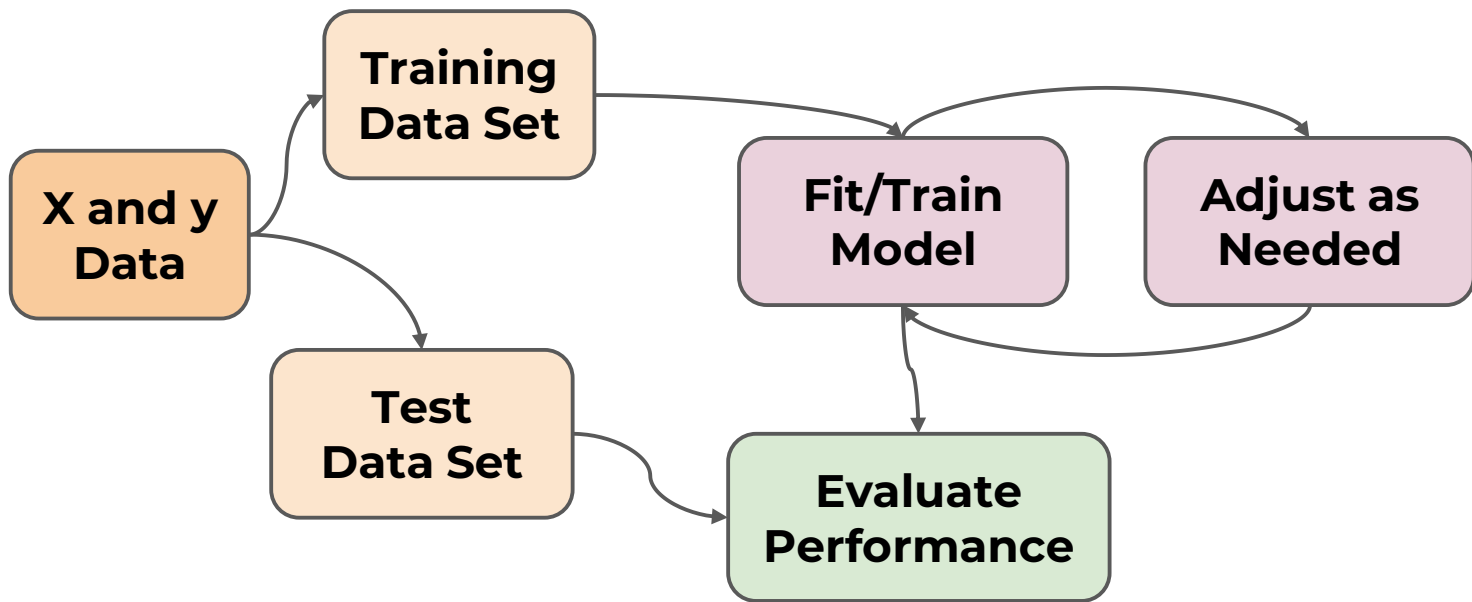
- Evaluate Model Performance





Supervised Machine Learning Process

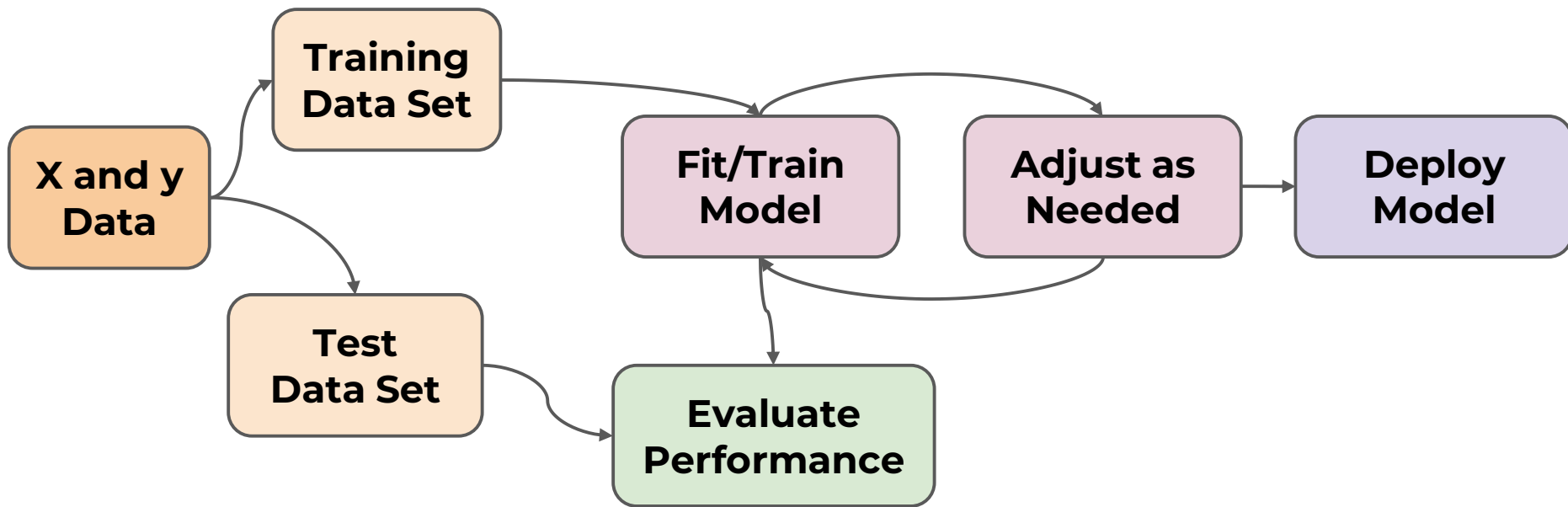
- Adjust model hyperparameters as needed





Supervised Machine Learning Process

- Deploy model to real world





Machine Learning

- ML Process : Supervised Learning Tasks



**Real
World**

**Collect &
Store
Data**

**Clean &
Organize
Data**

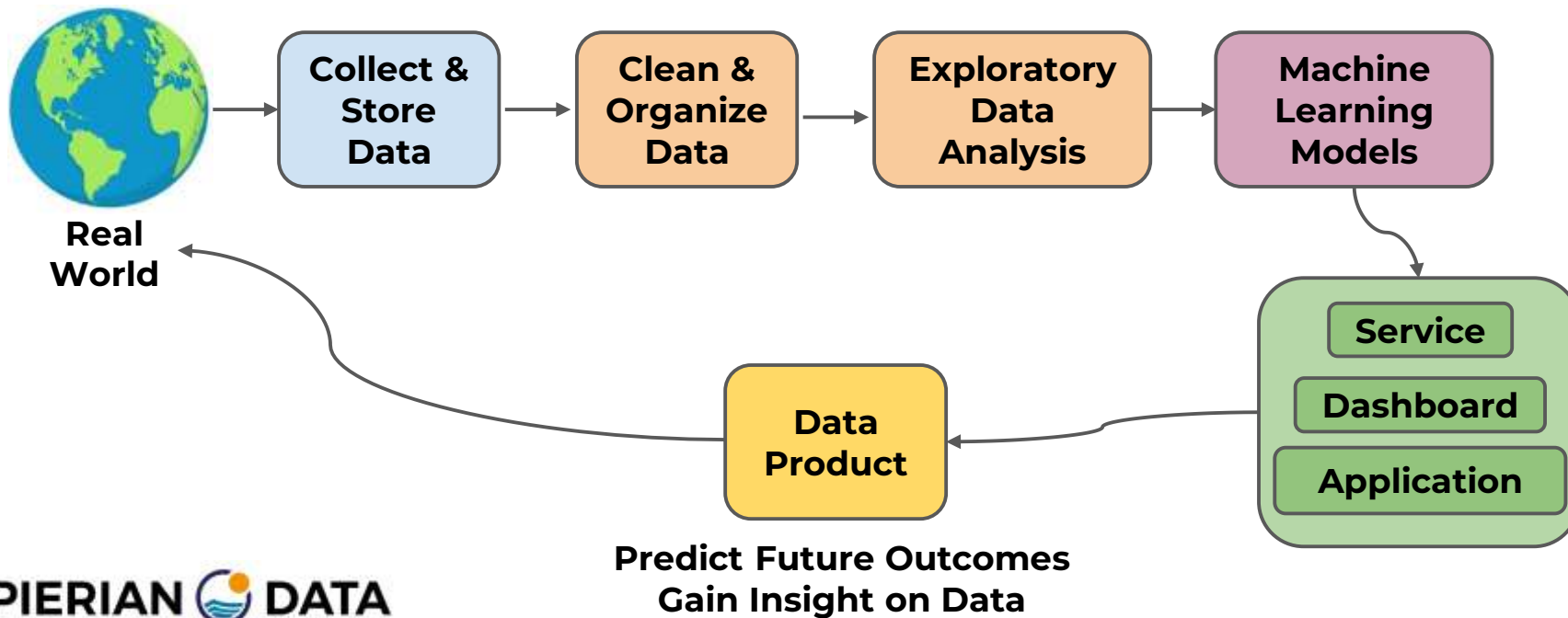
**Exploratory
Data
Analysis**

**Machine
Learning
Models**

Supervised Learning:
Predict an Outcome



ML Pathway





Companion Book



Machine Learning

- ISLR - Introduction to Statistical Learning
 - Freely available book that gives a fantastic overview of many of the ML algorithms we discuss in the course.
 - Quick note, it's code is for R users, but the math behind algorithms is the same regardless of programming language used in development.



Machine Learning

- We will refer to the book for optional reading assignments.
- A few examples will line up nicely with the book content.
- Book is freely available, simply google search for relevant links:
 - ISLR + Pdf