ISA 419: Data-Driven Security

12: Introduction to Machine Learning

Fadel M. Megahed, PhD

Endres Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Spring 2024

Quick Refresher of Last Class

- ✅ Utilize standalone data viz packages to construct and tailor your graphs.
- ✅ Examine the use of network and/or spatial plots in the context of network data.

Learning Objectives for Today's Class

- Explain the difference between supervised and unsupervised learning.
- Understand the importance of different preprocessing steps and when they should be used in ML.
- Explain typical error measures used for supervised and unsupervised learning tasks.

What is Machine Learning?

Definition

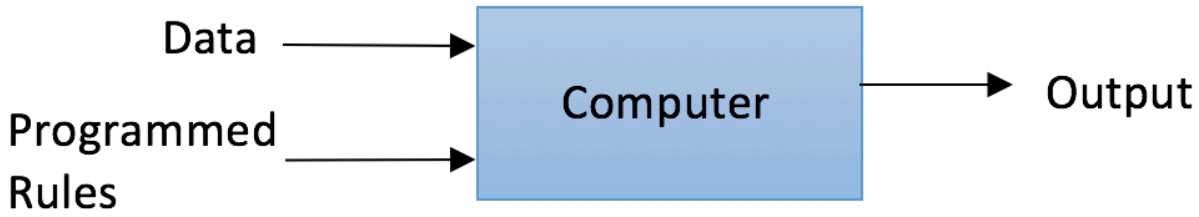
- Mitchell 2006 has elegantly defined the scientific field of machine learning to be centered around answering the following question:

> How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

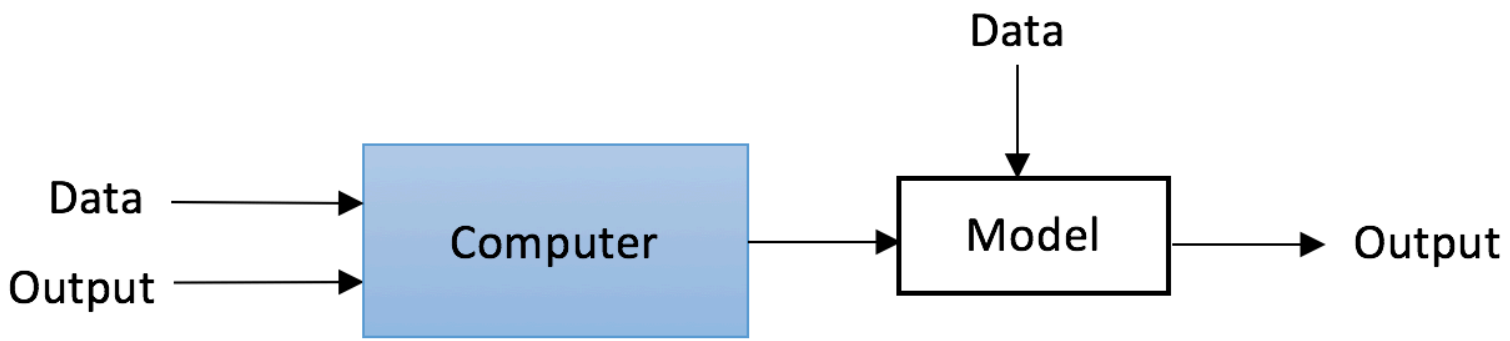
In his view, machine learning is the study of algorithms that:

- Improve their performance P
- at some task T
- with experience E.

A Paradigm Shift in Programming



Traditional Programming



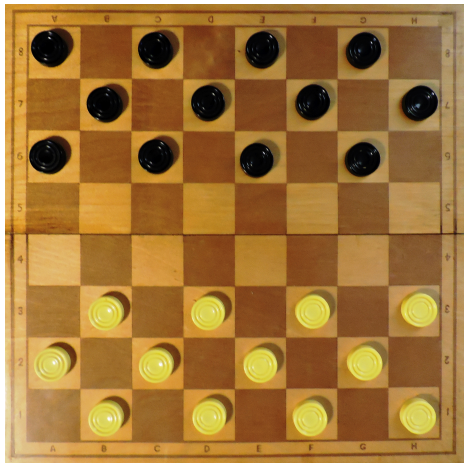
Machine Learning

Image Source: Yuxi Lui (2019). Python Machine Learning by Example. Packt Publishers.

Examples of a Task, Performance, and Experience

Improve on task `(T)`, with respect to performance metric `(P)`, based on experience `(E)`.

- T: Playing checkers
- P: Winning percentage against an arbitrary opponent
- E: Playing practice games against itself



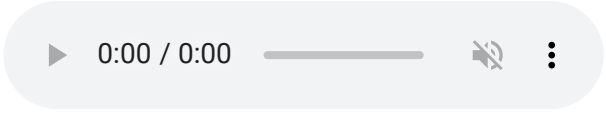
Note: This idea in Samuel (1959) led to the popularization of machine learning.

Image Source: Wikipedia

Examples of a Task, Performance, and Experience

Improve on task (T), with respect to performance metric (P), based on experience (E).

- T: Recognizing spoken words
- P: Accuracy of recognizing words
- E: Listening to a large number of words and their pronunciations



This is the Micro Machine Man presenting the most midget miniature motorcade of Micro Machines. Each one has dramatic details, terrific trim, precision paint jobs, plus incredible Micro Machine Pocket Play Sets. There's a police station, fire station, restaurant, service station, and more. Perfect pocket portables to take any place. And there are many miniature play sets to play with, and each one comes with its own special edition Micro Machine vehicle and fun, fantastic features that miraculously move. Raise the boatlift at the airport marina. Man the gun turret at the army base. Clean your car at the car wash. Raise the toll bridge. And these play sets fit together to form a Micro Machine world. Micro Machine Pocket Play Sets, so tremendously tiny, so perfectly precise, so dazzlingly detailed, you'll want to pocket them all. Micro Machines are Micro Machine Pocket Play Sets sold separately from Galoob. The smaller they are, the better they are.

Audio Source: OpenAI

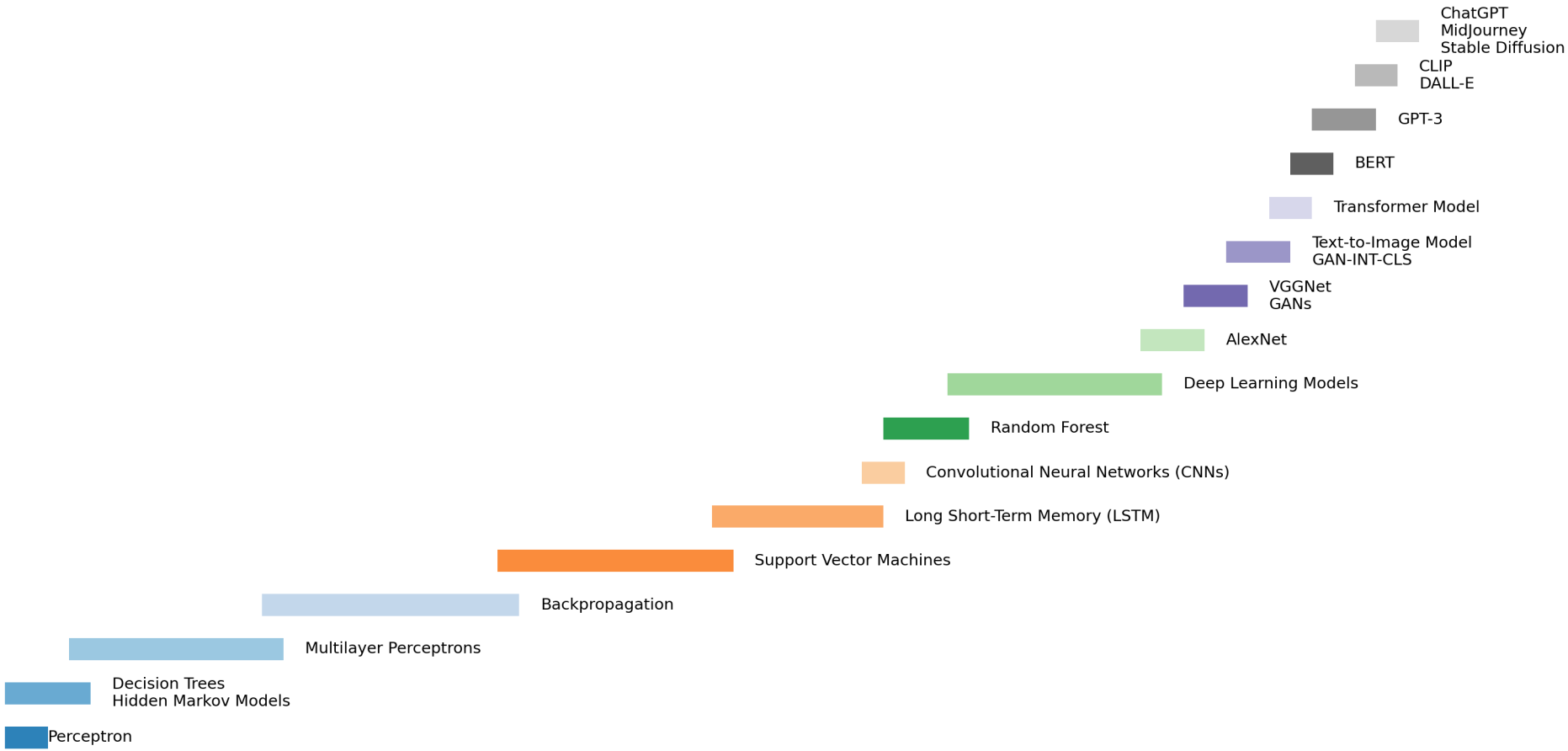
Examples of a Task, Performance, and Experience

Improve on task `(T)`, with respect to performance metric `(P)`, based on experience `(E)`.

- T: Classifying emails as spam or not spam
- P: Accuracy of classifying emails
- E: Reading a large number of emails and identifying which ones are spam

Image Source: Sarah Mestiri (2017). Applied Text-Classification on Email Spam Filtering - Part 1. Medium..

History of Machine Learning



Types of Machine Learning

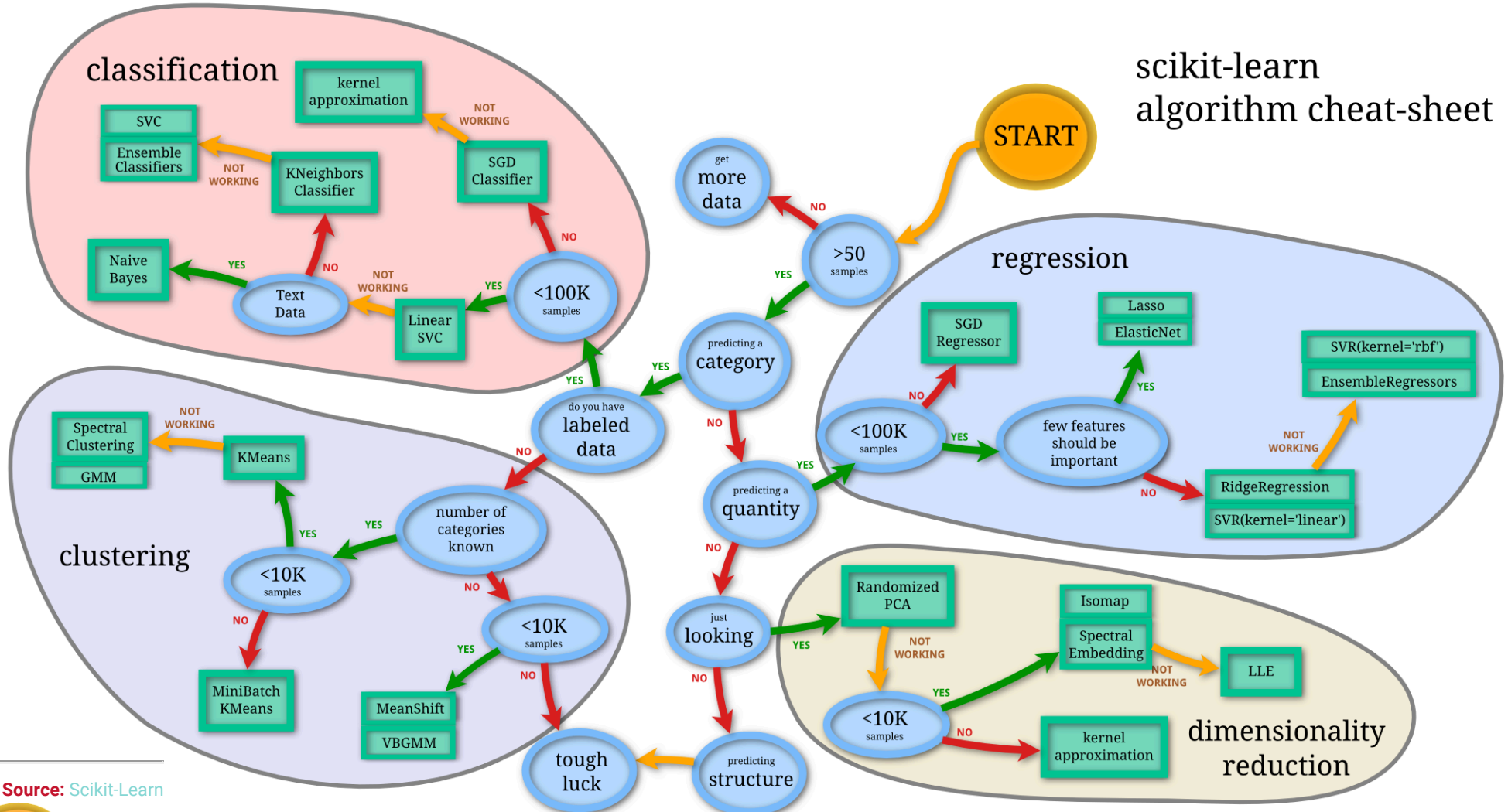
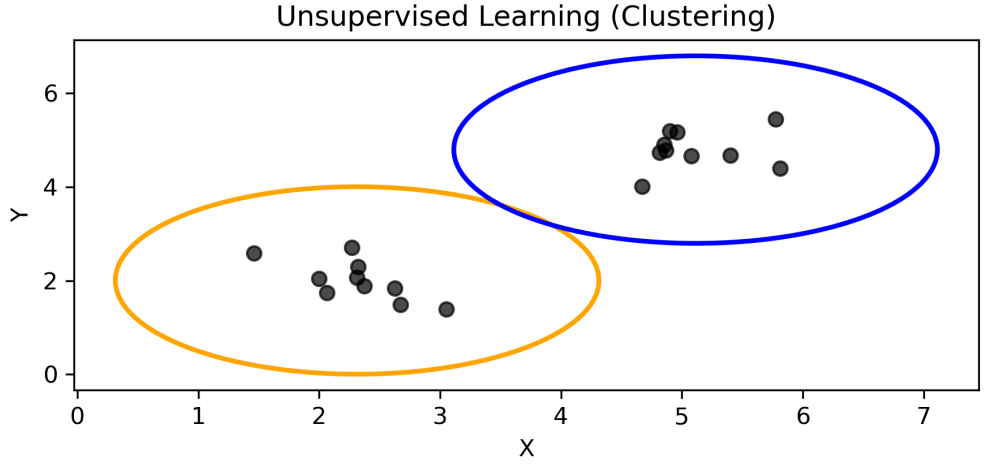


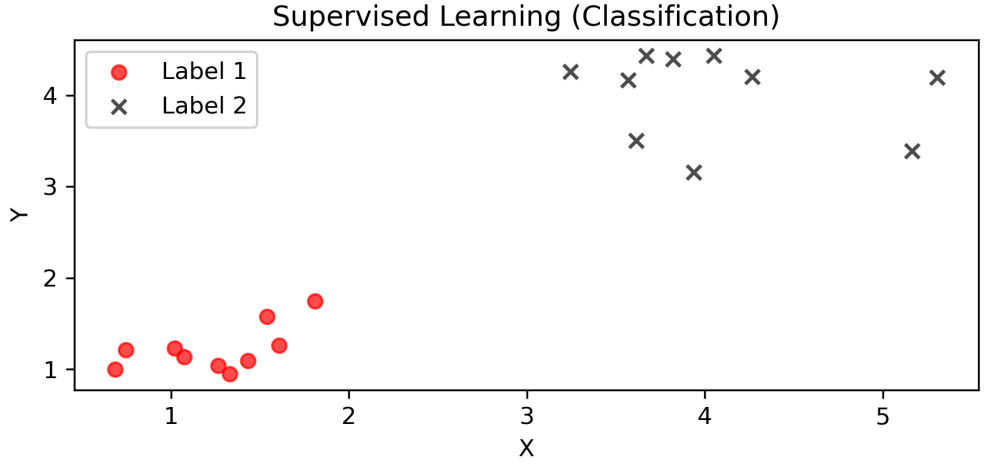
Image Source: Scikit-Learn

Explain the Difference Between Unsupervised and Supervised Learning.

Unsupervised Vs. Supervised Learning

Aspect	Unsupervised Learning	Supervised Learning
Labels	Labels are **NOT** known	Labels are known
Learning Objective	The algorithm tries to **learn the underlying structure** of the data	The algorithm tries to **learn the mapping** from input to output
Examples	**Clustering**, **Dimensionality Reduction**	**Regression**, **Classification**
Evaluation	Cannot be evaluated easily	Can be evaluated using **error measures**





Unsupervised Learning

- We retroactively try to understand the structure of the data.
- We do not have any labels to guide us.
- So we try to find patterns in the data (and then, we try to make sense of these patterns).
- Once we make sense of these patterns, we can use them to guide our decision-making process.
- **Cybersecurity** applications include:
  - **Anomaly Detection**, where we try to find patterns that are different from the norm.
  - **Clustering**, where we try to group similar entities together (e.g., similar network traffic).
  - **Dimensionality Reduction**, where we try to reduce the number of features in our data so that we can better understand/visualize/model it.

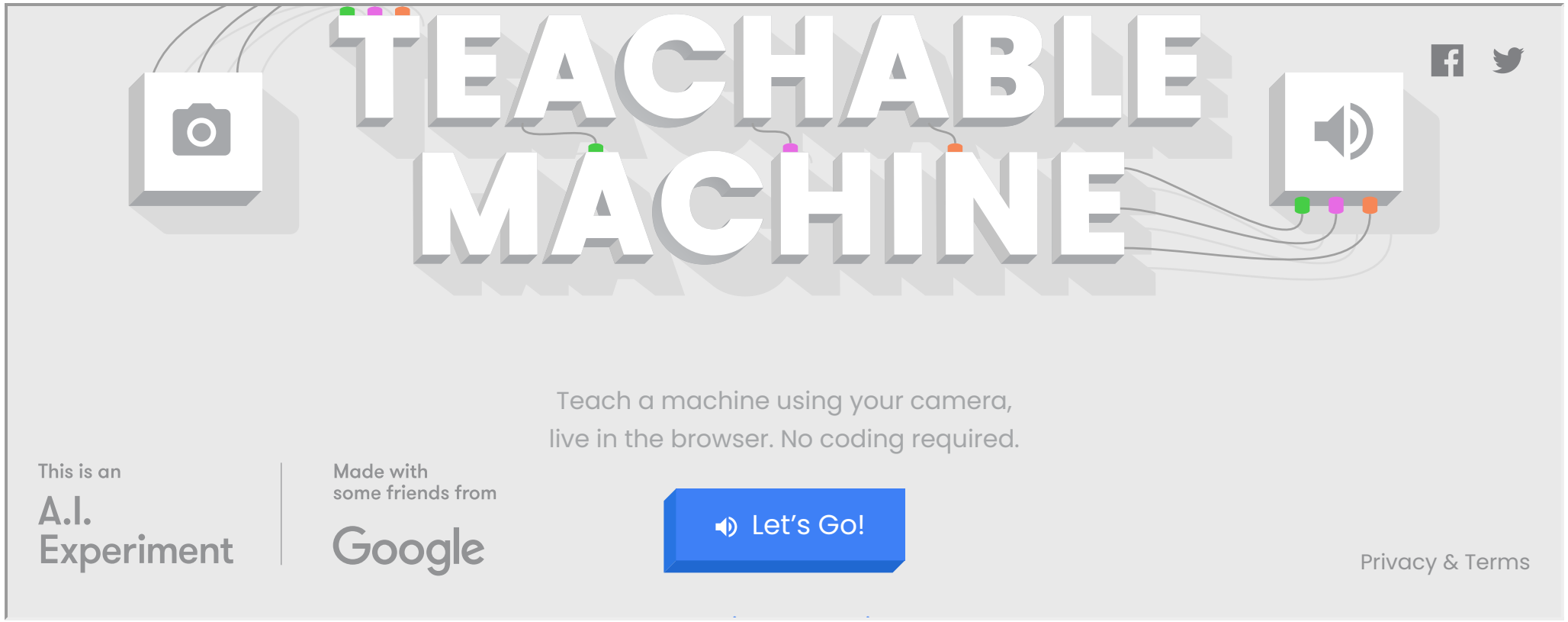
Supervised Learning

- We have labels to guide us.
- We build a model that allows us to see:
  - How the input data is related to the output data.
  - How we can use this relationship to make predictions.
- **Cybersecurity** applications include:
  - Classification of network traffic as benign or malicious.
  - Regression to predict the number of attacks on a network.
  - Time series forecasting to predict future network traffic (technically not a ML problem, but it can be modeled sometimes using ML-type models).

Class Activity: Build a Fun Model with No Code

05:00

Go to https://teachablemachine.withgoogle.com/v1/ and build a simple model to classify images by following the instructions on the page.



Preprocessing Steps in Machine Learning

Preprocessing Steps in Machine Learning

- There are several preprocessing steps that are typically used in machine learning.
- These steps are used to prepare the data for the model.
- These include:

Step	Description
Z-Transformation	Scaling the continuous data so that they have a mean of 0 and a standard deviation of 1.
Data Scaling	Scaling continuous variables so that they have a min of 0 and a max of 1.
Data Encoding	Encoding categorical data using label or one-hot encoding.
Data Imputation	Filling in missing data (apply with caution, only after understanding the type of missingness).
Data Reduction	Reducing the number of features in the data.

Preprocessing Steps in Machine Learning

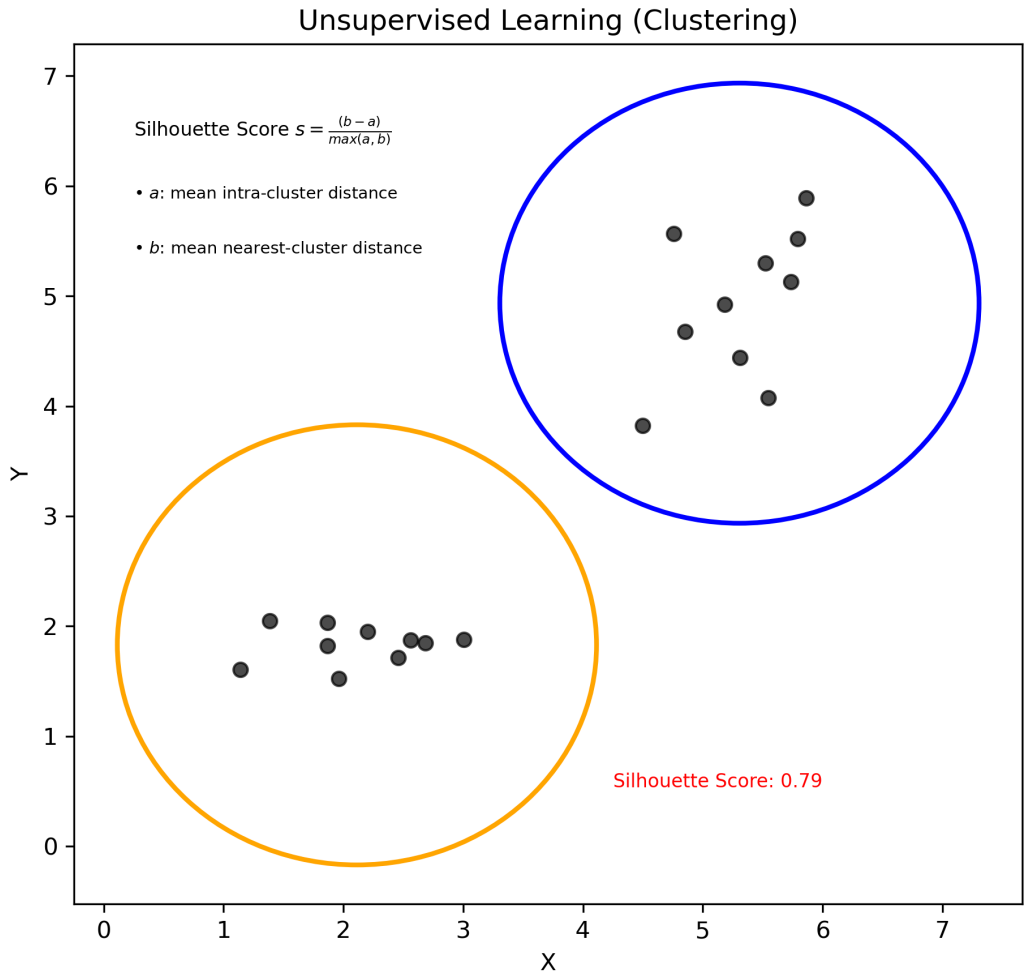
- **Z-Transformation**:
  - This is done to ensure that the data have a mean of 0 and a standard deviation of 1.
  - This is important for algorithms that are sensitive to the scale of the data (e.g., K-Means clustering).
  - It is also important for algorithms that use gradient descent (e.g., neural networks).
- **Data Scaling**:
  - This is done to ensure that the data have a minimum of 0 and a maximum of 1.
  - This has a similar use case to Z-Transformation.
- **Data Encoding**:
  - This is done to convert categorical data into a format that can be used by the model.
  - This is important because most models (e.g., neural networks) cannot handle categorical data.

Preprocessing Steps in Machine Learning

- **Data Imputation**:
  - This is done to fill in missing data.
  - This is important because most models cannot handle missing data.
  - However, this should be done with caution, as it can introduce bias into the model.
- **Data Reduction**:
  - This is done to reduce the number of features in the data.
  - This is important because it can help to reduce the complexity of the model, which can help to reduce overfitting.

Error Measures in Machine Learning

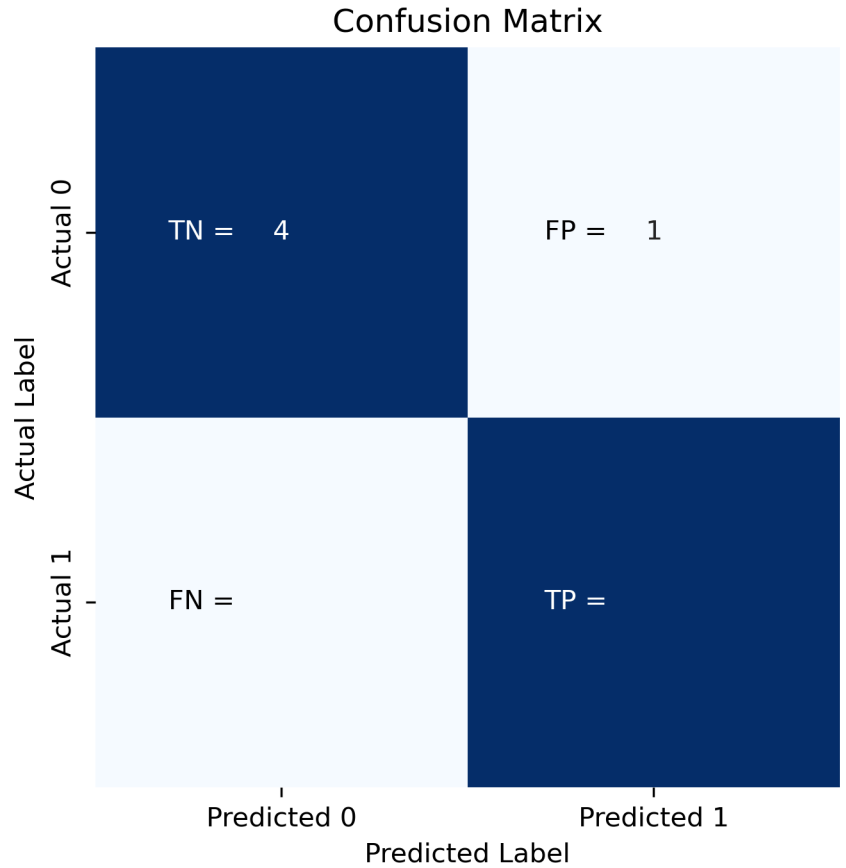
Supervised Learning Model Evaluation: Clustering



Supervised Learning Model Evaluation: Clustering

- **Silhouette Score**:
  - This is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
  - The silhouette score ranges from -1 to 1.
  - A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Supervised Learning Model Evaluation: Classification



Supervised Learning Model Evaluation: Classification

- **Accuracy**:
  - This is the proportion of correctly classified instances (i.e., $\frac{TP+TN}{TP+TN+FP+FN}$)
  - It is **not** a good measure when the classes are imbalanced or when the cost of misclassification is high.
- **Precision**:
  - This is the proportion of true positive predictions out of all positive predictions (i.e., $\frac{TP}{TP+FP}$).
  - It is a good measure when the cost of false positives is high.
  - It is not a good measure when the cost of false negatives is high.

Supervised Learning Model Evaluation: Classification

- **Recall**:
  - This is the proportion of true positive predictions out of all actual positive instances (i.e., $\frac{TP}{TP+FN}$).
  - It is a good measure when the cost of false negatives is high.
  - It is not a good measure when the cost of false positives is high.
- **F1-Score**:
  - This is the harmonic mean of precision and recall (i.e., $2 \times \frac{precision \times recall}{precision + recall}$).
  - It is a good measure when the cost of false positives and false negatives is high.
  - It is a good measure when the classes are imbalanced.

Supervised Learning Model Evaluation: Classification

- **ROC Curve**:
  - This is a plot of the true positive rate against the false positive rate (i.e., $\frac{TP}{TP+FN}$ vs. $\frac{FP}{FP+TN}$).
- **AUROC**:
  - This is the area under the ROC curve.
  - It is a good measure when the cost of false positives and false negatives is high.
  - It is a good measure when the classes are imbalanced.

Supervised Learning Model Evaluation: Regression

- **Mean Error (ME)**:
  - This is the average of the differences between predictions and actual values.
  - It is a measure of bias.
- **Mean Absolute Error (MAE)**:
  - This is the average of the absolute differences between predictions and actual values.
  - It is a good measure when the cost of large errors is high.
  - It is not a good measure when the cost of small errors is high.
- **Mean Squared Error (MSE)**:
  - This is the average of the squared differences between predictions and actual values.
  - It is a good measure when the cost of large errors is high.
  - It is not a good measure when the cost of small errors is high.

Recap

Summary of Main Points

By now, you should be able to do the following:

- Explain the difference between supervised and unsupervised learning.
- Understand the importance of different preprocessing steps and when they should be used in ML.
- Explain typical error measures used for supervised and unsupervised learning tasks.

 Review and Clarification 

- **Class Notes**: Take some time to revisit your class notes for key insights and concepts.
- **Zoom Recording**: The recording of today's class will be made available on Canvas approximately 3-4 hours after the end of class.
- **Questions**: Please don't hesitate to ask for clarification on any topics discussed in class. It's crucial not to let questions accumulate.