# L02: End-to-End Machine Learning Project

## CSci 574 Machine Learning (Géron) Ch. 2

Derek Harter

Department of Computer Science
East Texas A&M University

Summer 2025

EAST TEXAS A&M

# Starting a Machine Learning Project: The Big Picture

- Frame the problem
  - Know the objective (business, academic).
  - Determine the training supervision needed: supervised, unsupervised, semi-supervised, etc.
  - Is it a classificatin task, regression or something else?
  - Use batch learning or online techniques.
- Select a performance measure
  - For regression, might use root mean square error (RMSE) or mean absolute error (MAE).
  - For classification might use overall accuracy if appropriate
- Check the assumptions

# Get the Data

- Running the class notebooks using class DevContainer environments
- Using Jupyter Notebooks
  - Interactive code and documentation
  - Run all cells cleanly from first to last cell
- Download the data
- Loading the data for a first look

# Create a Test Set

- The goal of ML is to build a model that generalizes well to data that has not been seen before.
- ML models can (often easily) overfit a training set of data, which means they essentially remember particular instances, or patterns that are only coincidental noise present in the training data.
- So pick some instances randomly that will not be used to train the model, only to evaluate the models ability to generalize.
  - Typically pick 20% of the dataset, or less if the dataset is very large.
- Careful if you always shuffle randomly and train many times, you may be inadvertantly biasing your model
  - Thus for a project should split only 1 time, or split with a known seed so you always get the same train/test data split.
  - Even that can be problematic, what if the dataset changes?
  - If a unique id is present, can hash the id and select those in bottom 20% for example

# Sampling Bias

# Bibliography