Christian Granados
H195A Senior Thesis
Data Source Documentation
October 27, 2023

Preliminary Datasets
These datasets are purposefully simple to serve as a sanity check. Simple datasets and models allow better interpretability and ability to follow the flow of execution; this will allow an easier time finding bugs. These datasets are purposely diverse as to not build for one specific field or type of data.

PC1 Software Defect Prediction
Higgs Dataset
MNIST Dataset
Fashion MNIST
Musk V2

Training Datasets
These datasets are more akin to real-world tasks, and pose more challenges in prediction/classification. MLPs can likely still be used in this step, but here is where I will start looking into CNN elements for better performance.

SIFT1M
CIFAR-100
Caltech 256

Industry Benchmarks
These datasets are used to holistically evaluate current architectures and networks.

ImageNet
MLCommons Dollar Street

All of the datasets are free for research use, but need proper citation. This data is available from many online portals due to the original author's distribution terms. There has been an effort to have a diverse set of data — for the MLCommons Dollar Street, this was an express goal for the collected images — and most, if not all, of these datasets have stood the test of time.