

# 509 Project

Stephanie Daniella Hernandez Prado

Nanzhu Li

Serena Zhou

March 26, 2025

## 3. Data

### 3.1 Data Collection

We collected historical stock data for **Apple Inc. (AAPL)** using the **yfinance**. The dataset spans from **January 1, 2012 to March 15, 2025**, with data sampled at a daily interval. The raw dataset includes the following features:

- Open, High, Low, Close prices
- Adjusted Close price
- Daily trading volume

This financial time series serves as the basis for return calculations and trend prediction.

### Data Processing

#### a) Return Calculation

Daily returns were computed using the percentage change in adjusted closing prices:

$$\text{Return}_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \text{pct\_change}()$$

#### b) Features

To capture short-term return dynamics, we created sliding windows of **15 consecutive daily returns**. Each window forms a single input sample to the neural network:

$$\begin{aligned}\text{Sample}_1 &= [r_1, r_2, \dots, r_{15}] \\ \text{Sample}_2 &= [r_2, r_3, \dots, r_{16}] \\ &\vdots \\ \text{Sample}_{3304} &= [r_{3302}, r_{3303}, \dots, r_{3304}]\end{aligned}$$

This resulted in a dataset of **3,304 samples**, each consisting of **15 features**.

**c) Label**

For each sample, we calculated the sum of returns over the following 5 days and labeled the trend based on a threshold defined as  $0.3 \times$  standard deviation of future returns. The target variable was then classified into three categories:

- **0:** Downtrend (future return  $< -\text{threshold}$ )
- **1:** Neutral (within  $\pm\text{threshold}$ )
- **2:** Uptrend (future return  $> \text{threshold}$ )

Thus, each row in the final dataset consists of a 15-day return window (input) and a 3-class trend label (output).