Inspiring the extraordinary

Machine Learning

COMP2261 Artificial Intelligence

Lecturer: **Dr. Yang Long**

Associate Professor, MRC Innovation Fellow, SMIEEE
Director of Hybrid Intelligence Lab, VIViD Group
Department of Computer Science,
Durham University

The sub-module **Machine Learning** is taught by: Dr Yang Long in Term 1
Email: yang.long@durham.ac.uk
Office: Room 2009, Mathematical Sciences and Computer Science Building (MCS)

About me
Dr. Yang Long is currently an Associate Professor and Ph.D. supervisor at the Department of Computer Science, Durham University, UK. He is also the Director of the Hybrid Intelligence Lab and an MRC Innovation Fellow at the UK Medical Research Council. He is a Senior Member of IEEE (Institute of Electrical and Electronics Engineers). His primary research focuses on zero-shot vision, semantic multimodal hybrid intelligence systems, and their applications in interdisciplinary fields. Core technologies include artificial intelligence, machine learning, deep learning, computer vision, and natural language processing. His research spans theory, paradigms, and applications, with interdisciplinary applications in digital healthcare, neuroscience, cognitive science, computational social sciences, and engineering. He has published over 100 papers in top-tier conferences and journals, including IEEE TPAMI, TIP, CVPR, and AAAI.

Dr. Long has served as an Area Chair for BMVC since 2017 and as a reviewer for top academic journals and conferences, including IEEE TPAMI, TIP, CVPR, and ICML. He has received funding from the MRC Innovation Fellow program and participated as the primary applicant for Durham University in projects funded by the UK EPSRC, such as energy grids and smart transportation systems, securing over £14 million in funding. He has also participated in the review of several national key projects. On the industrial side, he has helped local internet industries empower AI through the UK Knowledge Transfer Partner (KTP) program, receiving funding of over £1.8 million. Dr. Long established the Hybrid Intelligence Lab in 2019 and currently leads a team of 20 full-time postdoctoral researchers and Ph.D. students. He also teaches advanced courses in computer vision and machine learning. From 2018 to 2019, he delivered AI courses to local UK companies and established strategic interdisciplinary AI research labs with several UK universities, promoting academic exchange. During the pandemic, he built strategic partnerships with multiple Chinese universities for research collaboration.

Lectures
In Term 1, Lectures are at **5 pm every Tuesday** and **9 am every Friday** in D/CLC203 in the Calman Learning Centre.

Office hour
There will be an **office hour on Tuesday at 4 pm** during each week of term where students can drop into my office. This provides an opportunity for students to ask me questions about the course or assignment. Alternatively, if you would like to chat via Zoom then email me and I'll set up a session at a mutually agreeable time. Of course, I will respond to questions by email too, whenever you send them.

Chapter 0: Introduction and Overview (2 hours)

- **Hour 1: Machine Learning System Demo and Course Logistics**
  - Overview of a complete ML system pipeline (data exploration, model building, evaluation).
  - Course logistics and feedback from last year.
- **Hour 2: Machine Learning in the Broader AI Landscape**
  - ML's place in AI, key applications such as NLP, computer vision, symbolic AI.

Chapter 1: Data Exploration, Preprocessing, and Problem Framing (4 hours)

- **Hour 3: Understanding Real-World Objectives and Problem Framing**
  - Defining real-world problems and framing them as ML tasks.
  - Translating business objectives into ML problems.
- **Hour 4: Mathematical Thinking in Machine Learning**
  - Importance of linear algebra, matrix operations, and vector thinking in ML.
  - Applications of these concepts in regression, PCA, etc.
- **Hour 5: Data Availability and Quality**
  - Assessing data quantity and quality.
  - Handling missing data, noise, and imbalanced datasets.
- **Hour 6: Data Preprocessing Techniques and Exploratory Data Analysis (EDA)**
  - Techniques for data preprocessing: scaling, normalisation, and encoding.
  - Conducting EDA using statistical methods and visualisation tools.

Chapter 2: Settings and Paradigms (4 hours)

- **Hour 7: Traditional Core ML Paradigms**
  - Supervised, unsupervised, semi-supervised learning, clustering, and basic reinforcement learning concepts.
- **Hour 8: Learning Paradigms for Complex Data**
  - Few-shot learning, zero-shot learning, representation learning, and meta-learning for complex and limited data.
- **Hour 9: Learning Paradigms for Complex Environments**
  - Federated learning, multi-agent learning, continual learning, and lifelong learning for complex, distributed environments.
- **Hour 10: Advanced Hybrid Learning Paradigms**
  - GANs, neuro-symbolic learning, adversarial learning, and hybrid models combining multiple approaches.

Chapter 3: Model Development and Training (4 hours)

- **Hour 11: Discriminative Models**
  - Linear regression, logistic regression, SVM, decision trees, random forests, KNN, and gradient boosting.
- **Hour 12: Generative Models**
  - Naive Bayes, GMM, HMM, GANs, and VAEs (with a focus on their role in traditional ML).
- **Hour 13: Representation Learning Models**
  - PCA, t-SNE, SOM, autoencoders, ICA, and LDA for dimensionality reduction and feature extraction.
- **Hour 14: Clustering and Unsupervised Learning Models**
  - K-means, hierarchical clustering, DBSCAN, agglomerative clustering, and affinity propagation.

Chapter 4: Evaluation and Analysis (4 hours)

- **Hour 15: Evaluation Metrics and Model Performance**
    - Accuracy, precision, recall, F1-score, AUC, and confusion matrix.
- **Hour 16: Error Analysis and Bias-Variance Tradeoff**
    - Techniques for error analysis, understanding bias-variance tradeoff, and overfitting/underfitting.
- **Hour 17: Cross-Validation and Data Splitting**
    - Cross-validation techniques and data splitting to combat overfitting.
    - Special validation methods (e.g., Zero-shot Learning).
- **Hour 18: Visualisation and Qualitative Analysis for Model Evaluation**
    - Visualisation tools (e.g., ROC curves, confusion matrices, t-SNE, PCA).
    - Qualitative analysis using tools like SHAP and LIME.

Chapter 5: Conclusion and Coursework Guidance (2 hours)

- **Hour 19: Cutting-edge Research and Future Directions**
    - Live demo of cutting-edge ML systems, discussions on state-of-the-art models, and ethical issues in AI.
- **Hour 20: Coursework Guidance**
    - Specific guidance on completing the coursework and final projects.
    - Review of expectations, milestones, and best practices for success.

Chapter 0: Introduction and Overview (2 hours)

Hour 1: Machine Learning System Demo and Course Logistics

- Overview of a complete ML system pipeline (data, paradigms, model, and evaluation).
- Course logistics and feedback from last year.

What's this sub-module all about?

Machine learning (ML) represents a core sub-discipline within the broader field of artificial intelligence (AI). At its essence, ML focuses on the design, development, and deployment of algorithms that enable computers to infer patterns and make decisions based on data, rather than relying on explicit programming. These algorithms adjust and improve autonomously over time as they are exposed to more data. The applicability of ML is vast, spanning from natural language processing and computer vision to predictive analytics in finance and healthcare. This sub-module delves into the foundational theories of machine learning, its various paradigms such as supervised, unsupervised, and reinforcement learning, and the mathematical underpinnings that drive these technologies. By understanding the intricacies of ML, students will be better equipped to harness its potential in addressing complex, data-driven challenges across various domains.

A Complete ML system in a nutshell:

Stock Price Prediction and Visualization Using Machine Learning
https://colab.research.google.com/drive/1kE5O6rJI0zZXklGFUJ9sps-R78xUBA6j?usp=sharing

In this project, we use historical stock data from three major companies: Apple (AAPL), Google (GOOGL), and Amazon (AMZN). The goal is to:

- Visualize the stock price trends from 2020-01-01 to 2022-01-01.
- Train machine learning models to predict the stock prices from 2022-01-01 to 2023-01-01.
- Compare the final predicted stock prices at 2023-01-01 with the actual price trends by plotting them on the same chart.

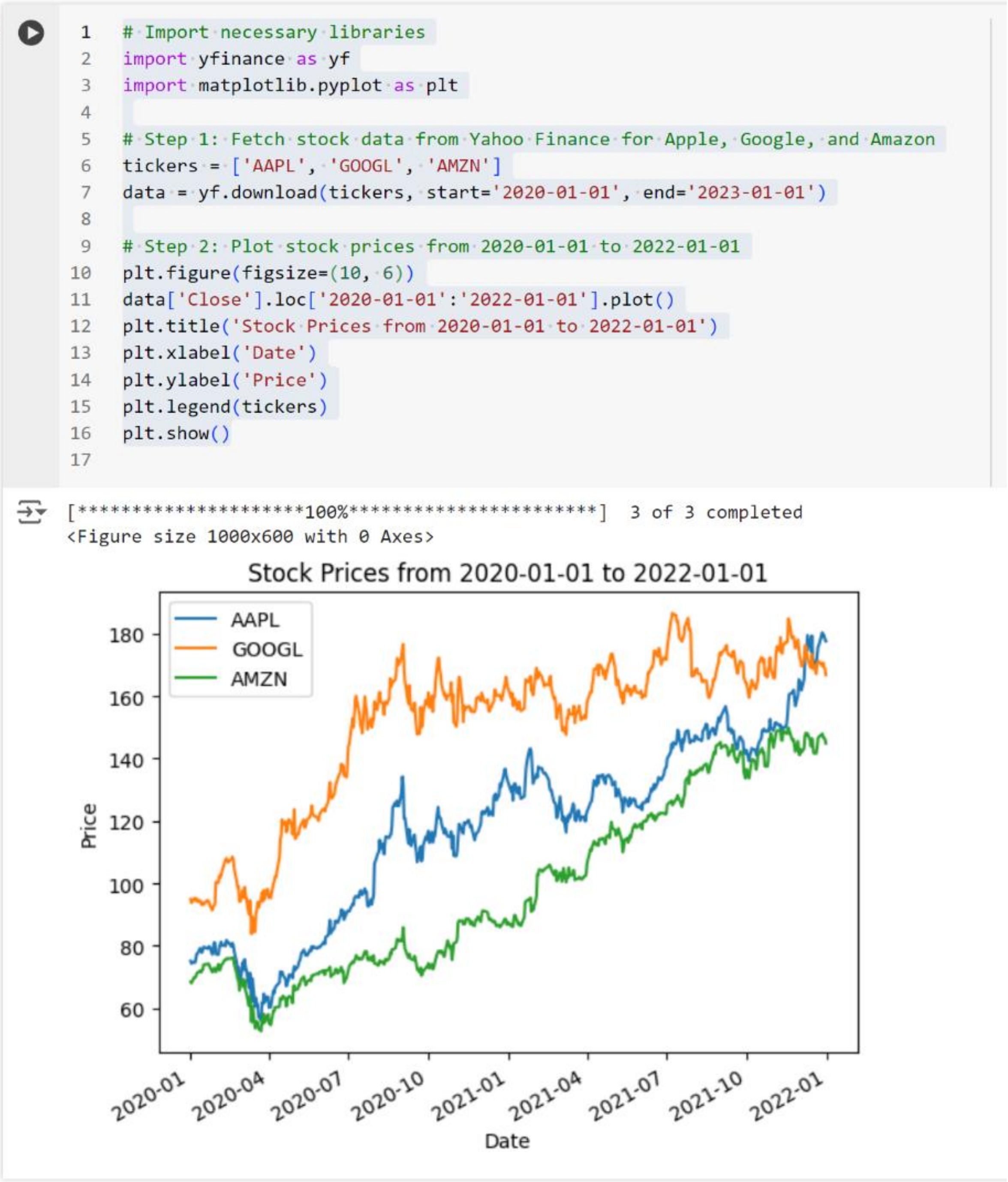
Step 1: Visualizing Stock Prices from 2020-01-01 to 2022-01-01
Objective:
In this step, we aim to visualize the historical trends of the three stocks, showing how the stock prices evolved over the period from 2020-01-01 to 2022-01-01. This visualization gives insights into the performance of each stock and helps students understand the market trends before any predictions are made.

Code Summary:
We fetch stock data for Apple (AAPL), Google (GOOGL), and Amazon (AMZN) using the yfinance library, and plot the closing prices for the specified date range.

Key Learning:
Students can analyse the upward or downward trends of each stock and compare them visually. This step sets up the context for predicting the stock prices for the following year (2022-2023).

Discussion:

- By 2023-01, can you predict the stock prices for the three companies?
- Which company could get the highest price among the three?
- What features can you see are useful for the prediction?



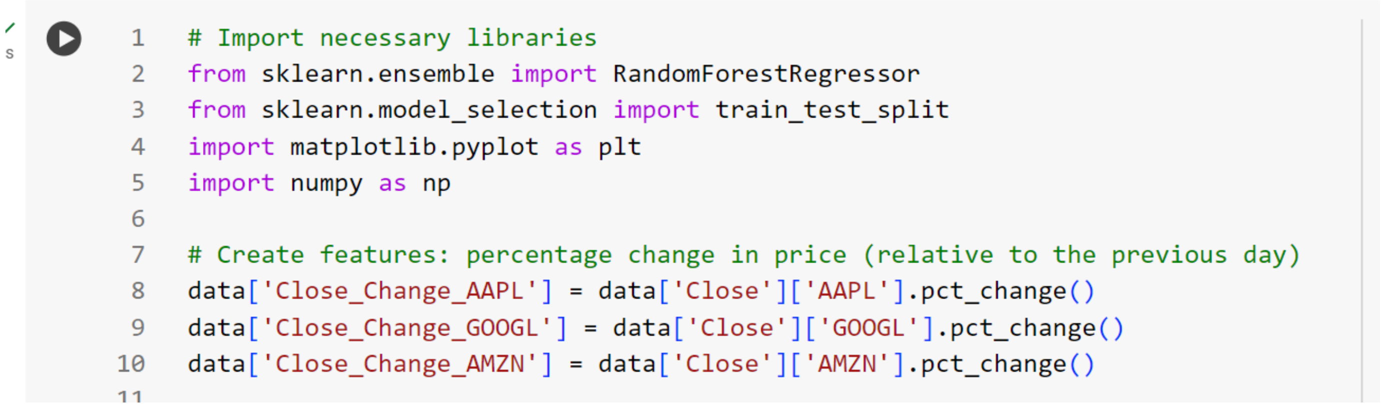
```python
# Import necessary libraries
import yfinance as yf
import matplotlib.pyplot as plt

# Step 1: Fetch stock data from Yahoo Finance for Apple, Google, and Amazon
tickers = ['AAPL', 'GOOGL', 'AMZN']
data = yf.download(tickers, start='2020-01-01', end='2023-01-01')

# Step 2: Plot stock prices from 2020-01-01 to 2022-01-01
plt.figure(figsize=(10, 6))
data['Close'].loc['2020-01-01':'2022-01-01'].plot()
plt.title('Stock Prices from 2020-01-01 to 2022-01-01')
plt.xlabel('Date')
plt.ylabel('Price')
plt.legend(tickers)
plt.show()

```

```
[*********************100%***********************]  3 of 3 completed
<Figure size 1000x600 with 0 Axes>
```

Step 2: Training the Model and Predicting Stock Prices from 2022-01-01 to 2023-01-01

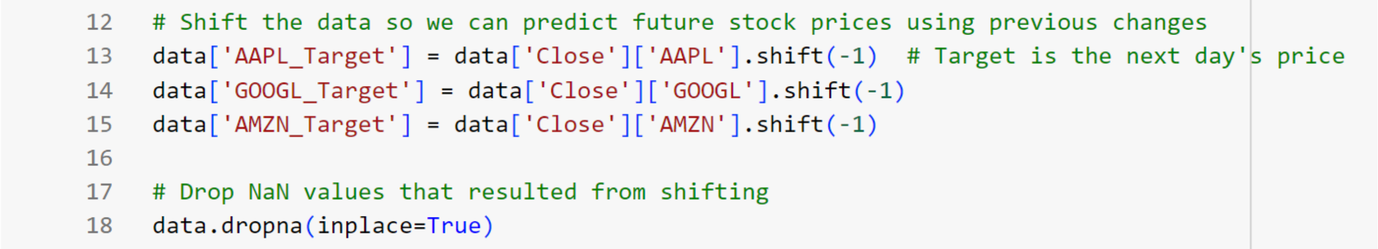
Objective:
In this step, we use historical stock data (from 2020-01-01 to 2022-01-01) to train a Random Forest Regressor for each stock. The model uses the percentage change in stock prices as features and predicts the stock prices for the period 2022-01-01 to 2023-01-01.

Machine Learning Process:
Feature Engineering: We calculate the percentage change in the stock prices for each company (AAPL, GOOGL, AMZN).



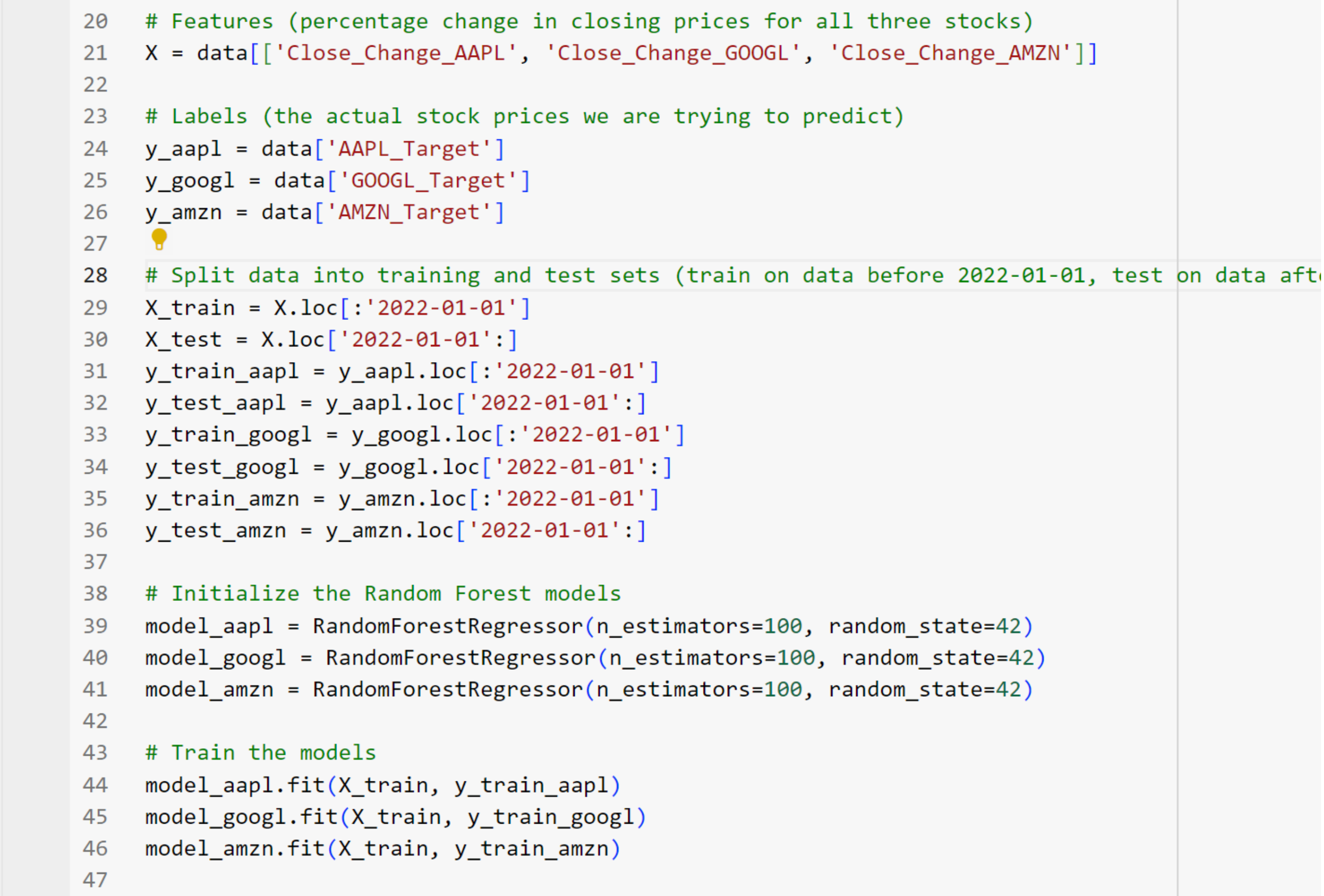
```
# Import necessary libraries
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import numpy as np

# Create features: percentage change in price (relative to the previous day)
data['Close_Change_AAPL'] = data['Close']['AAPL'].pct_change()
data['Close_Change_GOOGL'] = data['Close']['GOOGL'].pct_change()
data['Close_Change_AMZN'] = data['Close']['AMZN'].pct_change()
```

Target Variable: The target variable is the stock price for the next day, which allows the model to predict future prices based on past data.



```
# Shift the data so we can predict future stock prices using previous changes
data['AAPL_Target'] = data['Close']['AAPL'].shift(-1)  # Target is the next day's price
data['GOOGL_Target'] = data['Close']['GOOGL'].shift(-1)
data['AMZN_Target'] = data['Close']['AMZN'].shift(-1)

# Drop NaN values that resulted from shifting
data.dropna(inplace=True)
```

Training: We train three Random Forest Regressor models—one for each stock—using the historical data up to 2022-01-01.



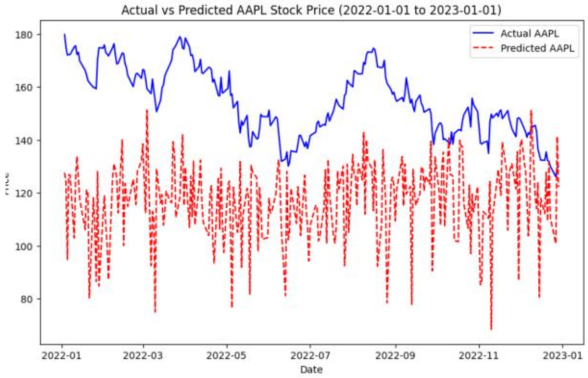
```
# Features (percentage change in closing prices for all three stocks)
X = data[['Close_Change_AAPL', 'Close_Change_GOOGL', 'Close_Change_AMZN']]

# Labels (the actual stock prices we are trying to predict)
y_aapl = data['AAPL_Target']
y_googl = data['GOOGL_Target']
y_amzn = data['AMZN_Target']

# Split data into training and test sets (train on data before 2022-01-01, test on data aft
X_train = X.loc[:'2022-01-01']
X_test = X.loc['2022-01-01':]
y_train_aapl = y_aapl.loc[:'2022-01-01']
y_test_aapl = y_aapl.loc['2022-01-01':]
y_train_googl = y_googl.loc[:'2022-01-01']
y_test_googl = y_googl.loc['2022-01-01':]
y_train_amzn = y_amzn.loc[:'2022-01-01']
y_test_amzn = y_amzn.loc['2022-01-01':]

# Initialize the Random Forest models
model_aapl = RandomForestRegressor(n_estimators=100, random_state=42)
model_googl = RandomForestRegressor(n_estimators=100, random_state=42)
model_amzn = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the models
model_aapl.fit(X_train, y_train_aapl)
model_googl.fit(X_train, y_train_googl)
model_amzn.fit(X_train, y_train_amzn)

```

Prediction: The trained models predict the stock prices from 2022-01-01 to 2023-01-01.

Key Learning:

Students gain insights into how machine learning models can be used to predict stock prices based on past trends.

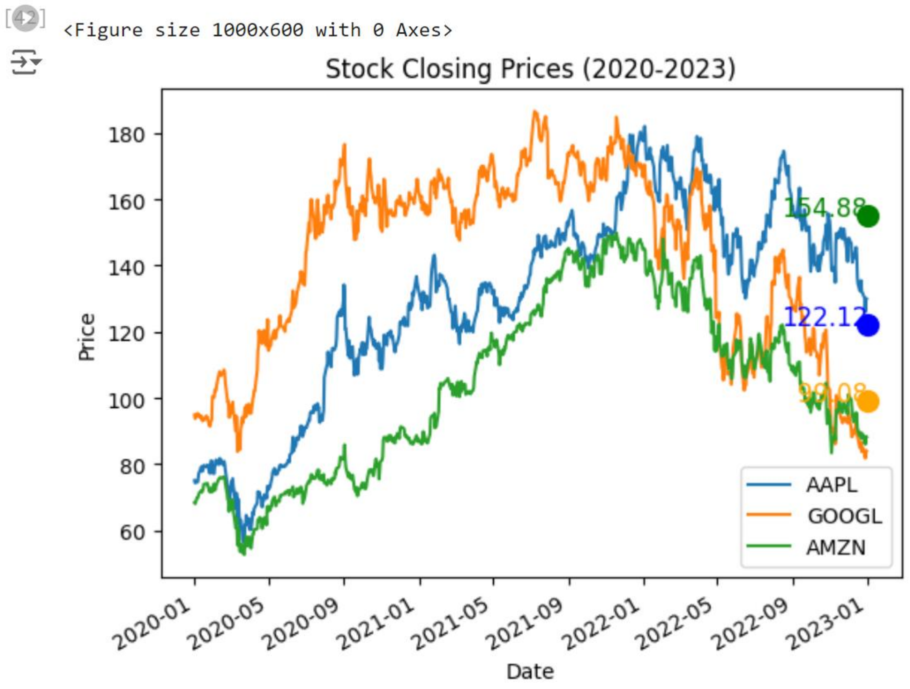
They can compare the predicted stock price trends to the actual trends once the model has been trained.



Step 3: Plotting the Final Predicted Prices at 2023-01-01
Objective:
The final step involves comparing the predicted stock prices for each company at 2023-01-01 with the actual stock price trends. The goal is to highlight the final predicted prices on the same chart that shows the full price trends from 2020-01-01 to 2023-01-01.

Visualization:
We plot the actual stock price trends for the full range (2020-2023).
We overlay the predicted prices at 2023-01-01 (from the model) as scatter points on the chart.
The final predicted prices are labeled on the plot for clarity.
Key Learning:
Students can visualize how close the predicted final prices are to the actual stock price trends.
They can evaluate the effectiveness of the model by comparing predicted prices with real market data.

Conclusion:
This three-step process provides a comprehensive workflow for using machine learning to predict stock prices and visually compare the predictions with real market data. By analyzing historical trends, training a predictive model, and overlaying predictions on actual data, students can understand both the power and limitations of machine learning in financial markets.



```
<Figure size 1000x600 with 0 Axes>
```

Logistics and Expectations

Weekly Structure:
Each week will involve:

- Lectures: Covering key theoretical concepts and examples.
- Practice Sessions: Build essential ML models and develop coursework.

Assessments:
Coursework: Focused on building a complete ML system with four focuses: *Data analysis, Paradigm, Model, and Evaluation.*

Will be released by the **end of week four.**

Feedback from Previous Years:
In previous years, students provided valuable feedback that highlighted both strengths and areas for improvement. While the overall course was well-received, some students expressed a desire for clearer marking criteria and more alignment between lectures and coursework. In response to the feedback, I have made the following changes:

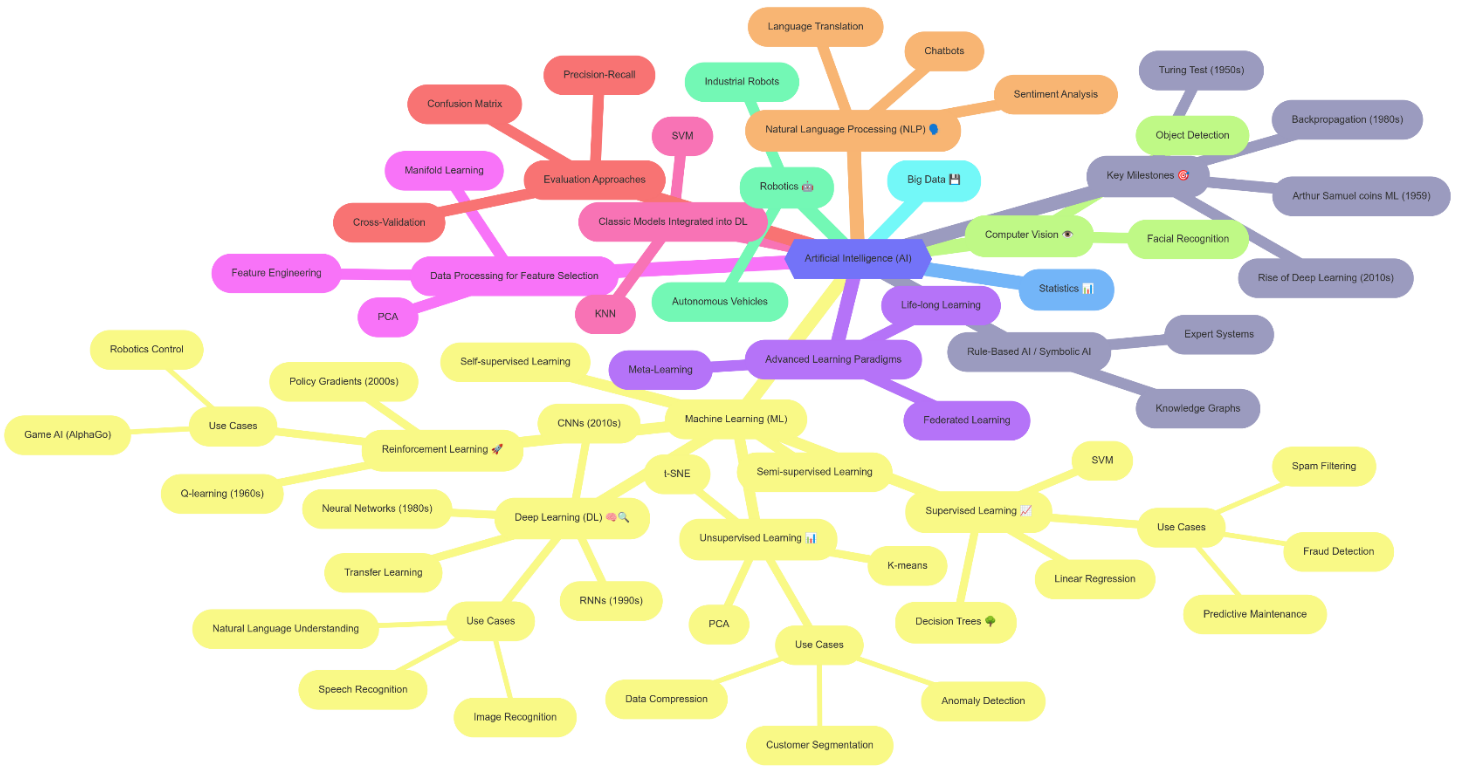
- Redesigned the entire module to better align with student feedback.
- New textbook now cover advanced models and paradigms, giving deeper insights into machine learning topics.
- Directly linked to coursework, focusing on key aspects:
  - Data exploration
  - Learning paradigms
  - Model development
  - Evaluation techniques

Practice:
Each week we have practical sessions with the assistant of demonstrators. Reading materials with codes of e.g. classic ML algorithms are provided. Students are encouraged to complete coursework during these practical sessions.

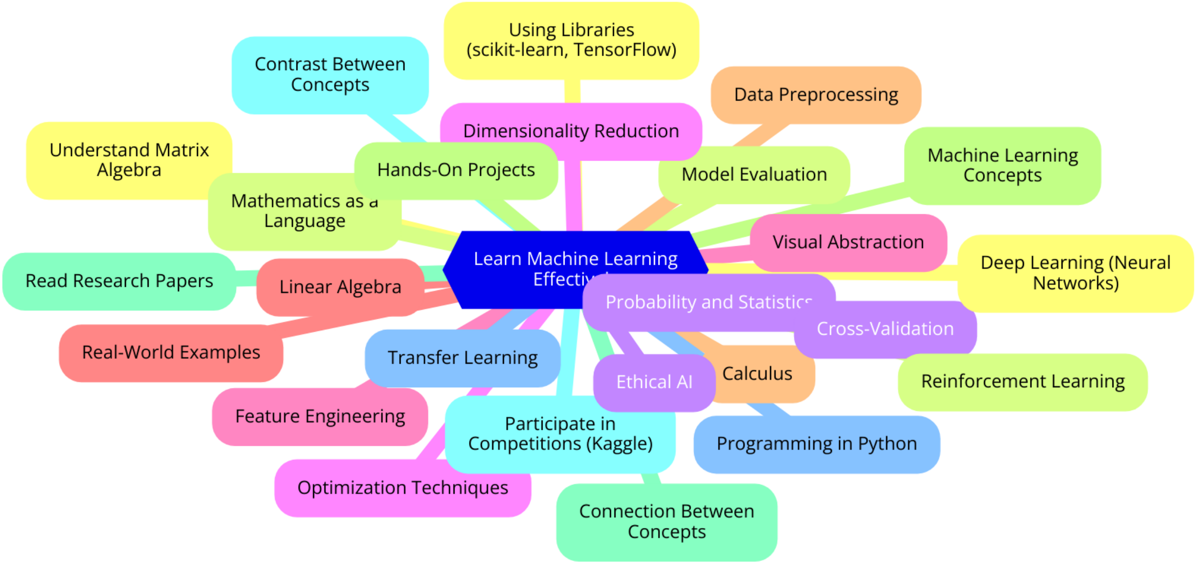
Mini conference
In addition, at the end of the course (week 9-10), we will hold a mini-conference, where students will present their coursework as if submitting to a real academic conference. You will be guided on how to upload your work to ArXiv in weeks 9-10 practice. The conference will simulate the academic review process, with myself acting as the program chair and volunteer PhD students serving as reviewers. You will receive feedback on your work by the end of December, allowing you to make improvements before the final submission in **28th January**. This is a unique opportunity to experience the academic publishing process while refining your machine learning project.

Hour 2: Machine Learning in the Broader AI Landscape

Discussion：What do you know about ML?



Discussion：What is essential to learn ML efficiently?


Learn Machine Learning Effectiv
Using Libraries (scikit-learn, TensorFlow)
Contrast Between Concepts
Data Preprocessing
Dimensionality Reduction
Understand Matrix Algebra
Hands-On Projects
Model Evaluation
Machine Learning Concepts
Mathematics as a Language
Visual Abstraction
Deep Learning (Neural Networks)
Read Research Papers
Linear Algebra
Probability and Statistic
Cross-Validation
Real-World Examples
Transfer Learning
Ethical AI
Calculus
Reinforcement Learning
Feature Engineering
Participate in Competitions (Kaggle)
Programming in Python
Optimization Techniques
Connection Between Concepts