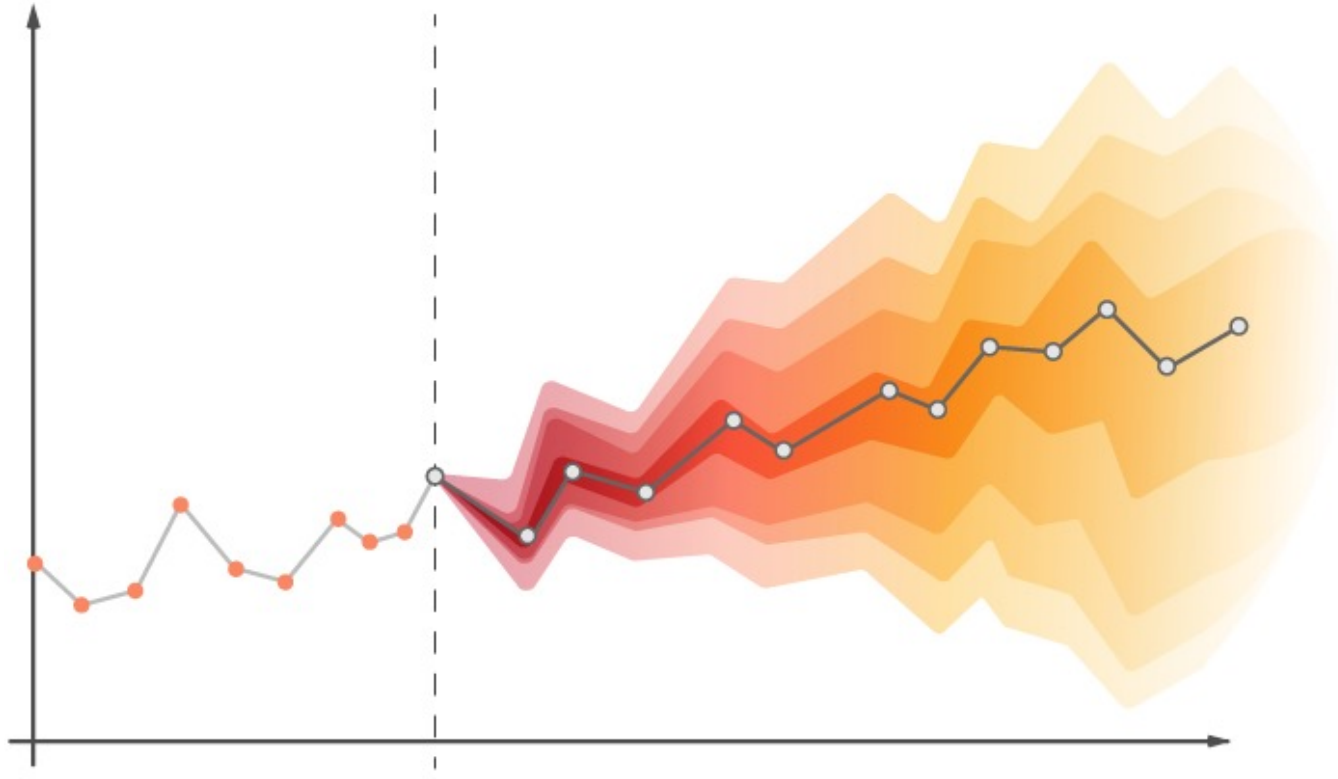
Final project

Time-series data and application to stock markets

Goal of the final project

The project aims at getting students familiar with **time-series data** and its applications by analyzing and deriving practical solutions using predictive analytics for stock markets.

But what is time-series data?



Introduction to time-series data

Time Series

['tīm 'sir-(,)ēz]

A sequence of data points that occur in successive order over some period of time.

Investopedia

Introduction to time-series data

Social and economic environment are constantly **changing** over time, data analysts must be able to assess and predict the effects of these changes, in order to suggest the most appropriate actions to take.

Time-series data is made up by dynamic data collected over time. Thus, it requires to have appropriate forecasting techniques to support business, operations, technology, research, etc.

Example: quarterly GDP of the last 10 years, weekly supermarket sales of the previous year, yesterday's hourly temperature measurements, etc.

What is the difference between **time-series data** and **sequential data**?

Introduction to time-series data

Several time-series analysis applications:

Logistics & Transportation

Forecasting of **shipped packages** (workforce planning).

Retail grocery

Forecasting of **sales** during promotions (optimizing warehouses).

Insurance

Claims prediction (determining insurance policies).

Manufacturing

Predictive Maintenance (improving operational efficiency).

Energy & Utilities

Energy load forecasting (better planning and trading strategies).

Objectives of time-series analysis

Time-series analysis is necessary for:

- **Summary** description (graphical and numerical) of data point versus time.
- **Interpretation** of specific series features (e.g., seasonality, trend, relationship with other series).
- **Forecasting** (e.g., predict the series values in $t + 1,\ t + 2, \ldots,\ t + k$).
- Hypothesis **testing and simulation** (i.e., comparing different scenarios).

Final project - Description

Time series data analysis and prediction: stock markets.

- International stock market: Nasdaq.
- Vietnam stock market: HOSE, HNX and UPCOM.

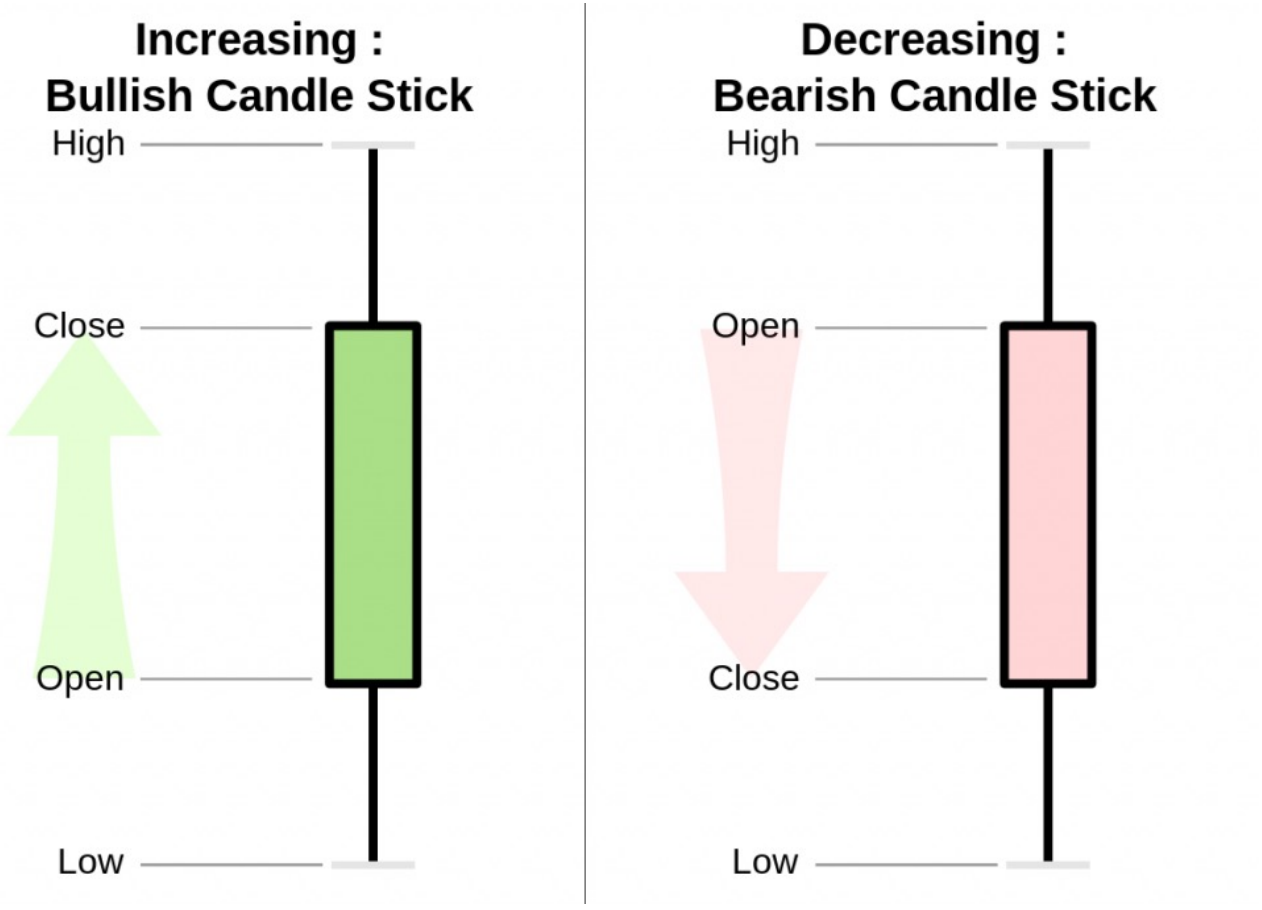
Nasdaq dataset

Data features:

- **Date**: data collection date.
- **Low**: refers to the lowest price of the period.
- **Open**: refers to the price at which a stock started trading when the opening bell rang.
- **Volume**: refers to the total number of shares traded in a day.
- **High**: refers to the highest price at which a stock is traded during a period.
- **Close**: refers to cost of shares at the end of the day.
- **Adjusted close**: considers other factors like dividends, stock splits, and new stock offerings. Since the adjusted closing price begins where the closing price ends, it can be called a more accurate measure of stocks' value.

As of Dec. 12, 2022, the Nasdaq dataset contains 1,564 companies and takes 841 MBs of csv text data after extraction.

Nasdaq dataset



Nasdaq dataset

1	Date	Low	Open	Volume	High	Close	Adjusted Close
2	19-05-1994	5.06584406	5.06584406	134232	5.36979389	5.26847696	3.038053274
3	20-05-1994	5.06584406	5.16716099	54285	5.36979389	5.26847696	3.038053274
4	23-05-1994	5.06584406	5.26847696	16532	5.26847696	5.06584406	2.92120409
5	24-05-1994	5.06584406	5.26847696	7649	5.26847696	5.06584406	2.92120409
6	25-05-1994	5.06584406	5.16716099	13325	5.16716099	5.06584406	2.92120409
7	26-05-1994	5.06584406	5.06584406	19987	5.21781921	5.06584406	2.92120409
8	27-05-1994	4.96452713	5.16716099	12338	5.16716099	4.96452713	2.862779856
9	31-05-1994	5.06584406	5.06584406	1481	5.26847696	5.26847696	3.038053274
10	1/6/94	5.06584406	5.06584406	52311	5.26847696	5.26847696	3.038053274
11	2/6/94	5.11650181	5.26847696	5182	5.26847696	5.26847696	3.038053274
12	3/6/94	5.26847696	5.26847696	2468	5.26847696	5.26847696	3.038053274
13	6/6/94	5.06584406	5.26847696	5675	5.26847696	5.06584406	2.92120409
14	7/6/94	5.06584406	5.26847696	15545	5.26847696	5.26847696	3.038053274
15	8/6/94	5.06584406	5.06584406	19740	5.26847696	5.06584406	2.92120409
16	9/6/94	5.06584406	5.06584406	2714	5.26847696	5.26847696	3.038053274
17	10/6/94	5.21781921	5.26847696	987	5.26847696	5.21781921	3.008842945
18	13-06-1994	5.26847696	5.26847696	494	5.26847696	5.26847696	3.038053274
19	14-06-1994	5.06584406	5.26847696	52558	5.31913614	5.16716099	2.979628801

Vietnam stock market dataset

Data features:

- **TradingDate**: data collection date.
- **Low**: refers to the lowest price of the period.
- **Open**: refers to the price at which a stock started trading when the opening bell rang.
- **Volume**: refers to the total number of shares traded in a day.
- **High**: refers to the highest price at which a stock is traded during a period.
- **Close**: refers to cost of shares at the end of the day.

The dataset contains three Vietnam stock exchanges:

- HoSE (VNIndex).
- HNX (HNXIndex).
- Upcom (UpcomIndex).

The Vietnam stock dataset is frequently updated. Refer to data folder for more details.

Vietnam stock market dataset

1		Open	High	Low	Close	Volume	TradingDate
2	0	9758	9758	7319	7904	140500	12/30/09
3	1	8441	8441	8441	8441	48900	12/31/09
4	2	9027	9027	8783	9027	36500	1/4/10
5	3	9612	9612	9027	9514	75600	1/5/10
6	4	9856	9856	8880	8880	65300	1/6/10
7	5	8539	9466	8295	8880	151000	1/7/10
8	6	9466	9466	9466	9466	38000	1/8/10
9	7	10100	10100	10100	10100	23400	1/11/10
10	8	10783	10783	10783	10783	123800	1/12/10
11	9	11466	11515	10051	10734	235500	1/13/10
12	10	10490	10929	10002	10246	32400	1/14/10
13	11	9661	9758	9563	9563	68300	1/15/10
14	12	9027	9027	8929	8929	24400	1/18/10
15	13	8587	8783	8343	8539	46100	1/19/10
16	14	8539	8587	7953	8051	87100	1/20/10
17	15	8197	8197	7514	7514	104300	1/21/10
18	16	7075	7953	7075	7221	61700	1/22/10
19	17	7319	7709	7172	7514	32800	1/25/10
20	18	7807	8002	7807	8002	50400	1/26/10
21	19	8295	8295	7514	7660	19700	1/27/10
22	20	7319	7319	7172	7270	34100	1/28/10

Vietnam stock market dataset

In addition, there are other datasets for Vietnam stock exchanges:

- dividend-history: historical dividends of companies.
- financial-ratio: financial health of companies.
- industry-analysis: information about companies in the same industry.

		exerciseDate	cashYear	cashDividendPercentage	issueMethod
1					
2	0	26/02/20	2019	0.05	cash
3	1	10/7/19	2019	0.05	cash
4	2	21/02/19	2018	0.03	cash
5	3	7/11/18	2018	0.05	cash
6	4	21/02/18	2017	0.05	cash
7	5	8/2/17	2016	0.05	cash
8	6	25/12/15	2015	0.06	cash
9	7	5/2/15	2014	0.1	cash
10	8	25/02/14	2013	0.1	cash

		ticker	quarter	year	priceToEarning	priceToBook	valueBeforeEbitda	dividend	roe	roa
1										
2	0	ABS	3	2022	7.2	0.5	17.8		0.077	0.042
3	1	ABS	2	2022	11.9	0.9	23.9		0.075	0.042
4	2	ABS	1	2022	27.8	2	31		0.073	0.04
5	3	ABS	4	2021	31.8	2.2	39		0.096	0.047
6	4	ABS	3	2021	32.4	1.8	27.8		0.079	0.046
7	5	ABS	2	2021	46.2	2.6	23.3		0.078	0.045
8	6	ABS	1	2021	70.6	4	26.4		0.078	0.045
9	7	ABS	4	2020	31.6	3.6	87.5		0.121	0.056
10	8	ABS	3	2020	32.9	2.3	77		0.07	0.044
11	9	ABS	2	2020	27.5	1.8	15.8		0.065	0.036

		ticker	marcap	price	numberOfDays	priceToEarning	peg	priceToBook	valueBeforeEbitda
1									
2	0	ABR	174	8720	-1	6.8	0.2	0.6	4.2
3	1	IPA	2823	13200	0	5.3	-0.1	0.8	44.2
4	2	TV2	1526	22600	3	13.6	-0.2	1.1	8.5
5	3	HSA	407	47000	0	12.4	0.1	2.4	
6	4	VNC	359	34200	0	10.8	-1.7	1.4	4.1
7	5	TV1	302	11300	1	24.3	-0.3	1	6.9
8	6	TV4	255	12900	-1	6.9	-1	1.1	5.1
9	7	PPS	161	10700	0	10	-0.6	0.9	5.7
10	8	VQC	75	5900	1	3.2	0.2	0.2	6.4
11	9	PPE	31	15300	0	17.2	0.3	3	
12	10	SDC	22	8300	0	9.5	0.2	0.4	14

Project evaluation

The project evaluation is based on the completion of these tasks:

1. (20%) Nasdaq stock price prediction.
2. (20%) Vietnam stock price prediction.
3. (20%) Vietnam/Nasdaq trading point identification.
4. (20%) Vietnam/Nasdaq portfolio/risk management.
5. (20%) Report.

Extra credit (25%): industry standard for deployment and ease of use.

Total: up to 125% of the ordinary project grading schema.

Project requirements - (20%) Nasdaq stock price prediction

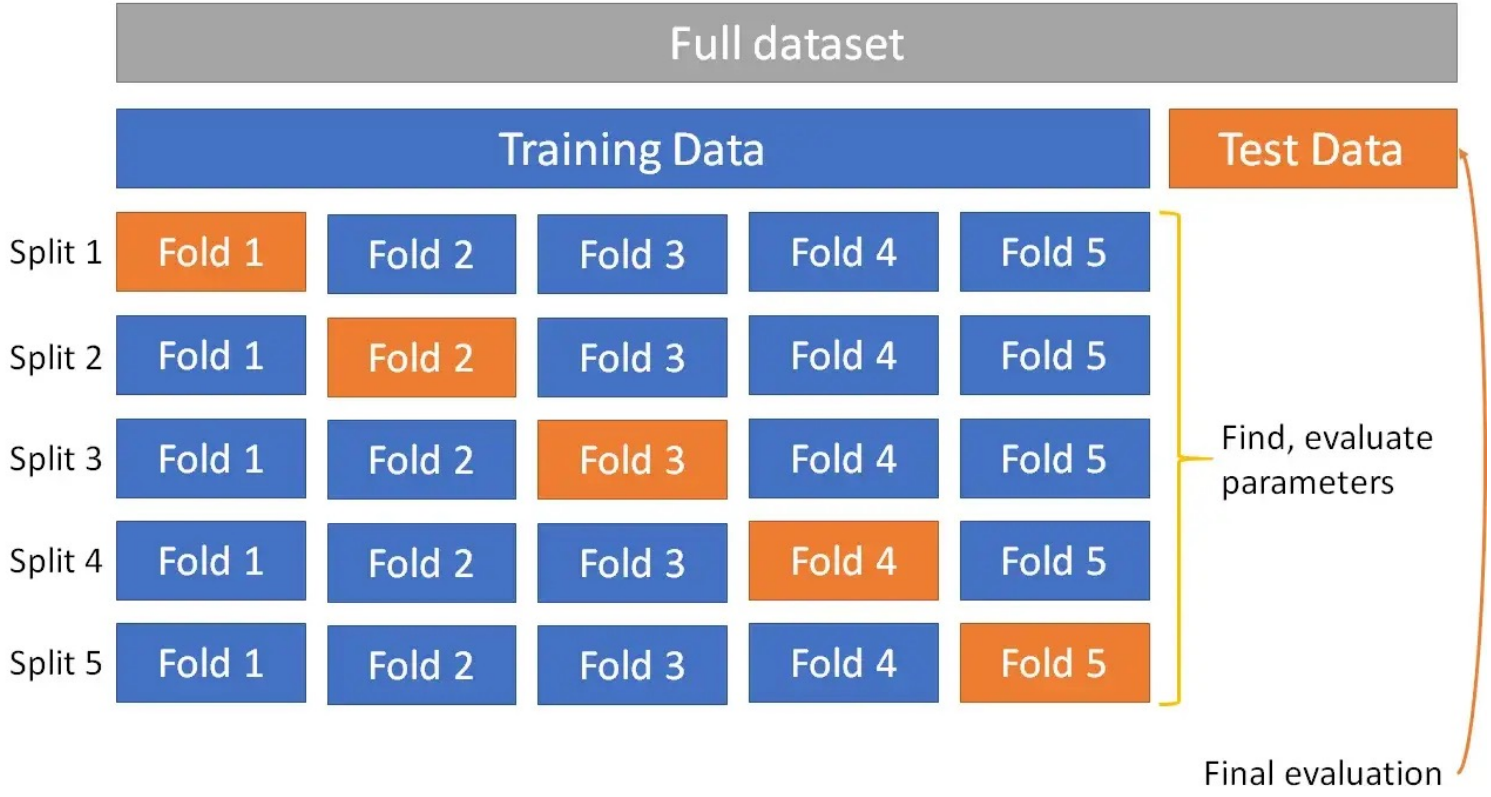
1. (20%) Nasdaq stock price prediction. Basic questions to answer:
    - Nasdaq stock price prediction, i.e., how much the price increases/decreases in a certain time window.
    - How does the price change in different time windows, e.g., one day, one week/seven days in a row, one month/thirty days in a row, etc.?

You have to figure out the following things:

- Training / Validation / Test split conforming to time-series data.
- Cross-validation conforming to time-series data.
- Time window, e.g., one-month training and one-week testing.
- Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc.
- ...

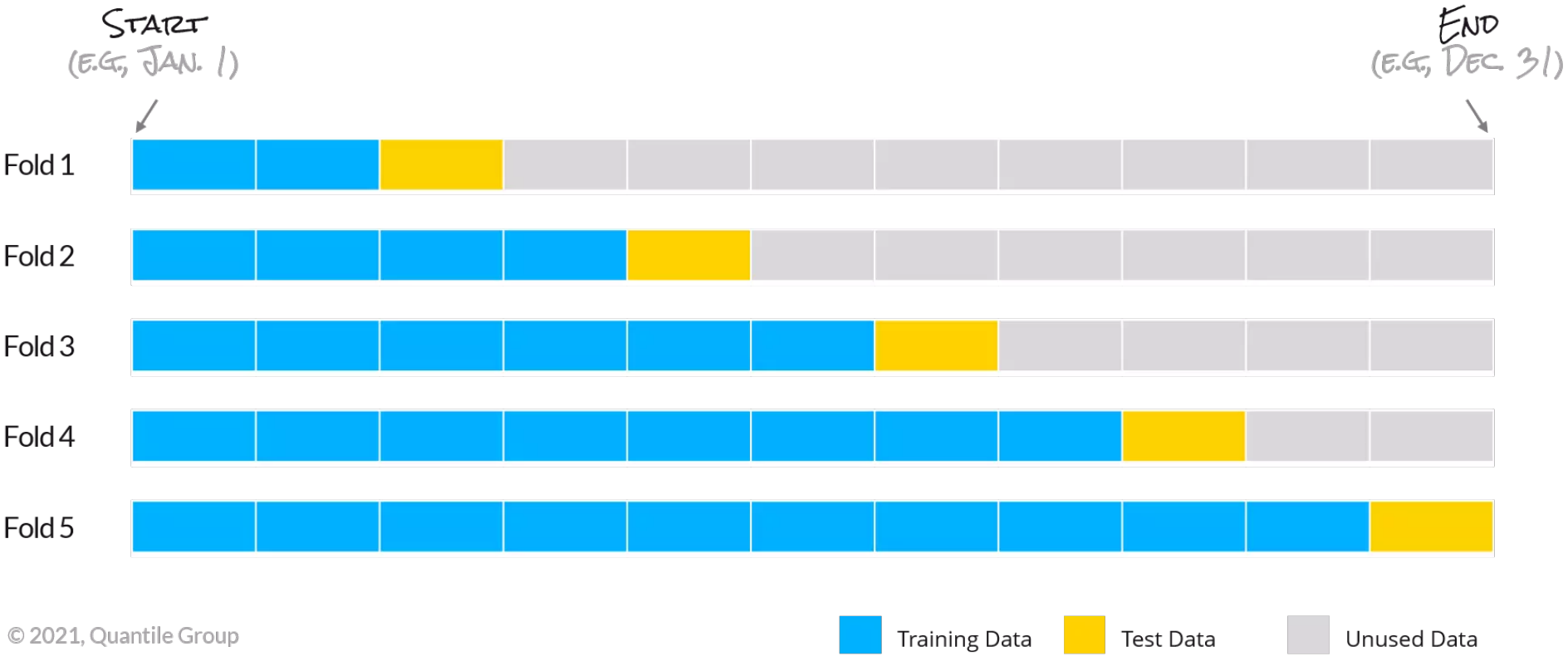
Cross-validation in time-series data

Cross validation in non time-series data



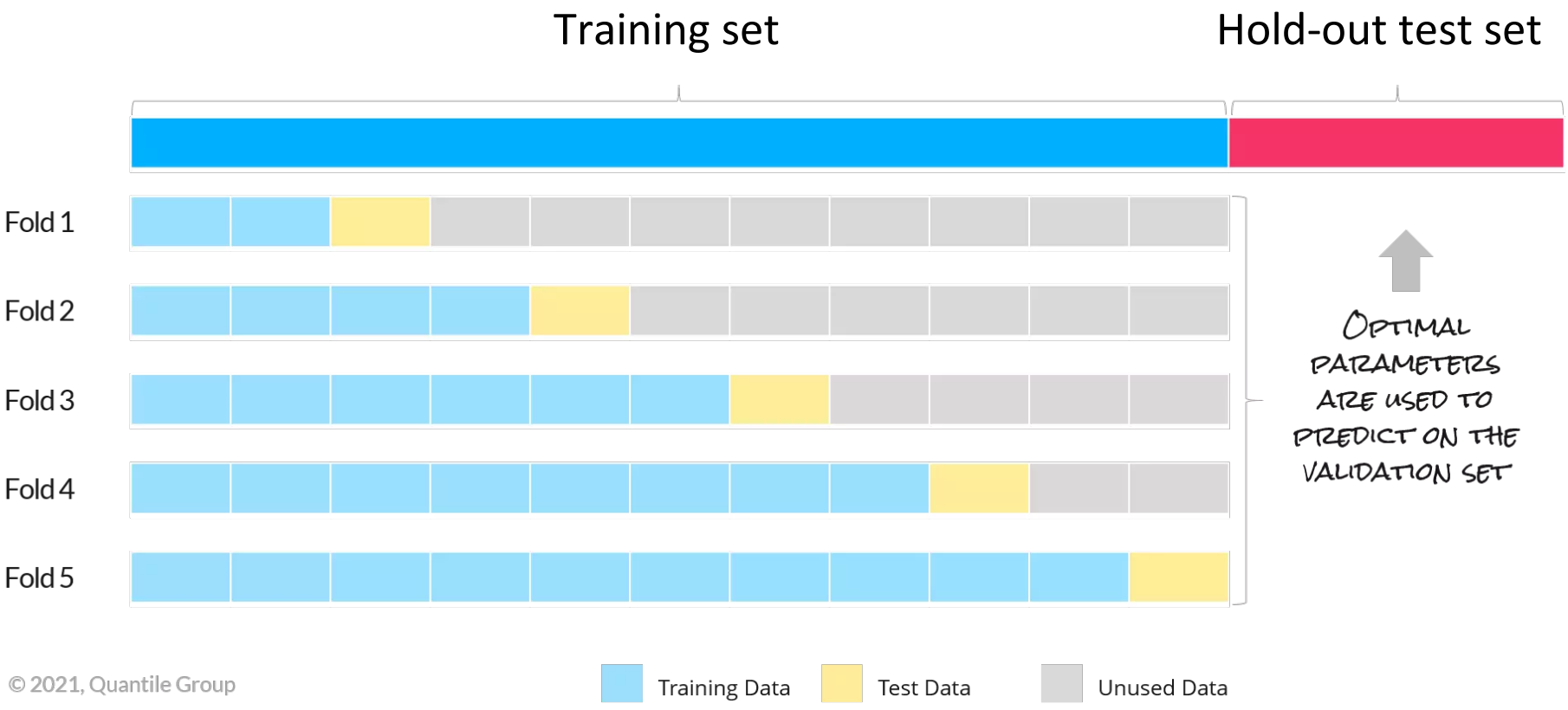
Cross-validation in time-series data

Illustration: Expanding Window CV



Cross-validation in time-series data

Cross validation conforming to the time-series aspect with a hold-out test set.



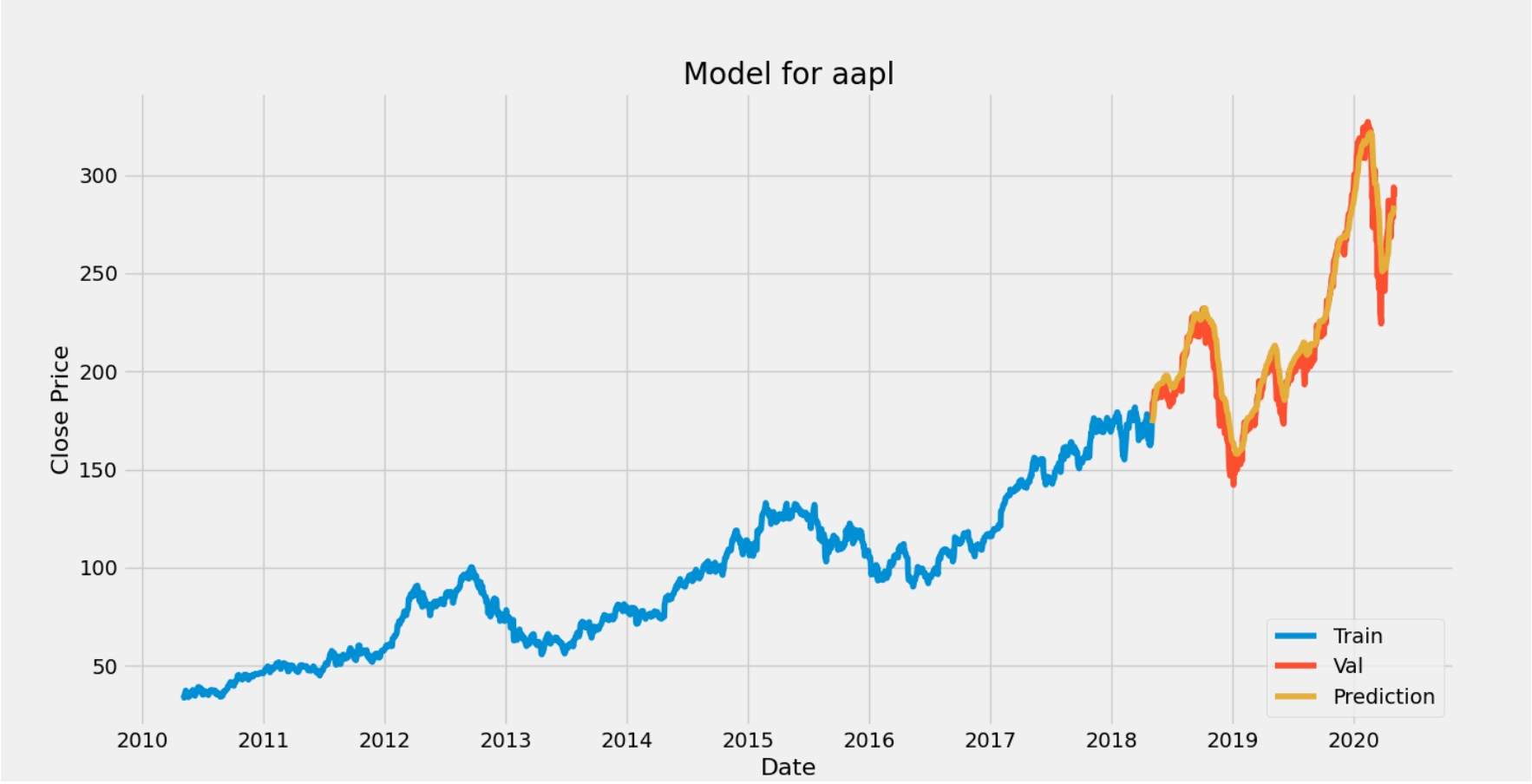
Project requirements - (20%) Vietnam stock price prediction

1. (20%) Vietnam stock price prediction. Basic questions to answer:
    - Vietnam stock price prediction, i.e., how much the price increases/decreases in a certain time window.
    - How does the price change in different time windows, e.g., one day, one week/seven days in a row, one month/thirty days in a row, etc.?

You have to figure out the following things:

- Training / Validation / Test split conforming to time-series data.
- Cross-validation conforming to time-series data.
- Time window for training and testing.
- Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc.
- Is it good to make use of additional Vietnam data such as dividend history, industry analysis, financial ratio.

Project requirements - (20%) Vietnam stock price prediction



Project requirements - (20%) Trading point prediction

1. (20%) Vietnam/Nasdaq trading point prediction. Basic questions to answer:
    - What is a good signal for buying stock of certain company?
    - What is a good signal for selling stock of certain company?

You have to figure out the following things:

- Training / Validation / test split conforming to time-series data.
- Cross-validation conforming to time-series data.
- Time window for training and testing.
- Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries, etc.
- Is it good to do the manual feature engineering such as Simple Moving Average (SMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), etc., to determine the training points?

Note: Building model for Vietnam market is more challenging, therefore more preferable.

Project requirements - (20%) Trading point prediction

Project requirements - (20%) Portfolio/risk mngt

1. (20%) Vietnam/Nasdaq portfolio/risk management: basic questions to answer:
    - What is the list of companies to hold? What is the profit within a certain period?
    - What is the list of companies to get rid of? Why?
    - How to combine potential scores and risk scores into a portfolio to optimize investment strategy?

You have to figure out the following things:

- Training / Validation split conforming to time-series data.
- Cross-validation conforming to time-series data.
- Time window for training and testing.
- Company filtering, e.g., those with at least 120 historical data points, companies in certain stock exchanges, companies in certain industries.
- What should be the list of companies to hold if investors are risk-taking or prudent?

Note: Building model for Vietnam market is more challenging, therefore more preferable.

Project requirements - (20%) Portfolio/risk mngt

Portfolio Management

[pȯrt-ˈfō-lē-ˌō ˈma-nij-mənt]

The art and science of selecting and overseeing a group of investments that meet the long-term financial objectives and risk tolerance of a client, a company, or an institution.

Investopedia

Project requirements - (20%) Report

1. (20%) Report:
    - Describing the journey about your experiments, observations, findings and conclusions.
    - Minimum six-page report.
    - Using AI-powered report writing tools is strictly prohibited.

Note:

- Experiment failures are as valuable as experiment successes.
- The report must include instruction, if necessary, of how to run your code such as external libraires to be installed, etc.

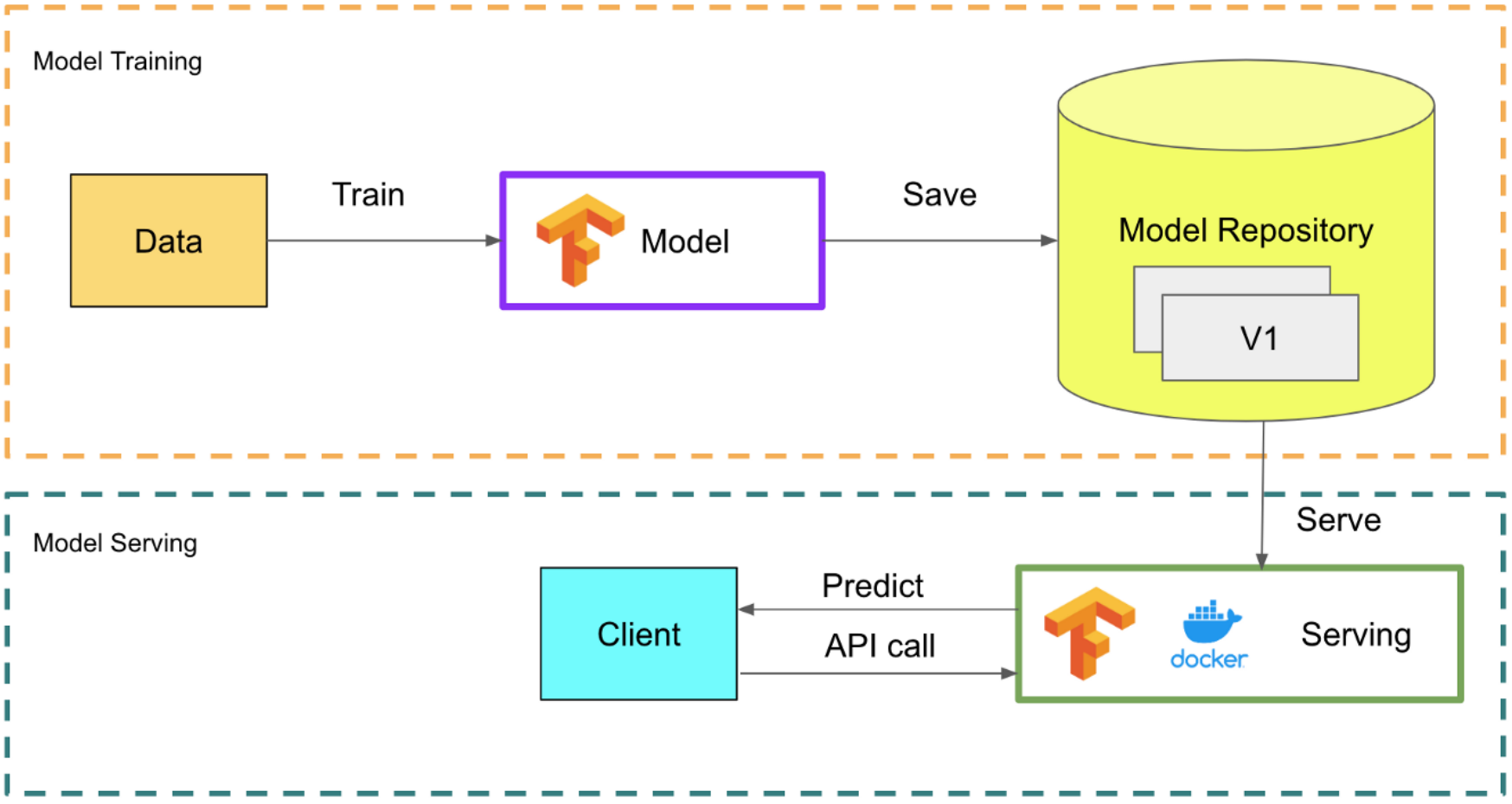
Project requirements - (25%) Extra credit

1. (25%) Extra credit: industry standard for deployment and ease of use.
    - Deploy the prediction models as API services.
    - Deploy the prediction models as a web-based Software-as-a-Service (SaaS).
    - Design an engineering flow to automate the tasks.

Some keywords to research:

- API services: Tensorflow Serving (TFServing), REST APIs, gRPC.
- Web-based SaaS: TensorflowJS, Superset / Tableau / PowerBI.
- Engineering and automation flow: SQL, MongoDB, Airflow, Airbyte, dbt.

Project requirements - (25%) Extra credit



Final project data

Data folder is organized as follows:

- ./stock-historical-data: folder containing historical stock price data.
- ./dividend-history: folder containing historical dividends.
- ./financial-ratio: folder containing financial health of companies.
- ./industry-analysis: folder containing analysis of companies in the same industries.
- companies.csv: list of companies.
- ticker-overview.csv: overview of companies.
- crawl-vn-data.py: script for crawling Vietnam stock data.

Technology stack

Technology stack: a priori, the project is not limited to any tools, libraries and programming languages. Here follows some examples:

Main requirements:

- Python (for programming language).
- Tensorflow (for deep learning and model serving).
- Scikit-learn (for machine learning and data analysis).

Extra credit:

- SQL/MongoDB (for database).
- Airflow (for task orchestration).
- Airbyte (as DB connector).
- dbt (for data transformation).
- Superset (for dashboard).

Final project submission

The structure of submission folder should be organized as follows:

./<StudentID>-project-notebook.ipynb: Jupyter notebook containing source code.
./<StudentID>-project-report.pdf: project report.

The submission folder is named DL4AI-<StudentID>-project (e.g., DL4AI-2012345-project) and then compressed with the same name.

Final project evaluation

This is a free-style project so that it is up to you to decide how to **programmatically** (and **reasonably**) formulate problems and come up with solutions. You will have to decide on your own (with **justification** or experimental result) what problems are feasible, i.e., acceptable model accuracy. The evaluation is based on how accurate and reasonable your solution and justification are.

Final project deadline

Please visit Canvas for details.

This project scope is large enough to be in your résumé for job application.

Thank you