

---

---

# **Extended MongoDB for Machine Learning: Data (Exploration) Analysis, Data Processing and Model Training**

Group 9

林天行 黃千睿 黃品翰 王睿謙 王雅茵

---

---

# Outline

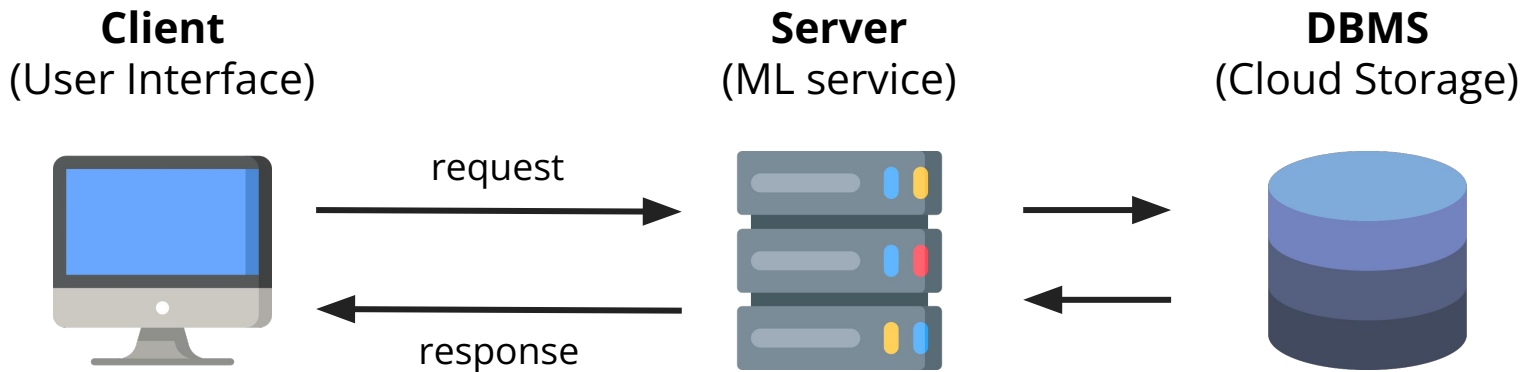
- Problem
- Solution
- Implementation
- Tools
- Objectives & Results (Phase I , Phase II)
- Limitation
- Possible extensions

# Problem

- 機器學習(ML)的需求日益增長，處理龐大的資料集經常是使用者會面對的挑戰
  - 一般使用情境：
    1. 先將遠端的訓練資料下載到本地端
    2. 於本地端進行資料探索→ 資料前處理 → 訓練模型
  - Challenges:
    1. 資料量太過龐大導致耗時、佔用電腦資源 & 儲存空間：升級硬體規格才能改善
    2. 環境架設難度高：ERROR一堆，處處不相容，總之套件就是裝不上去 ==
- ⇒ 困難重重

# Solution

- Architecture design:  
使用者先傳送請求到遠端server，並透過遠端server連結雲端DB的資料來運行ML，再傳輸結果至client端。



# Solution

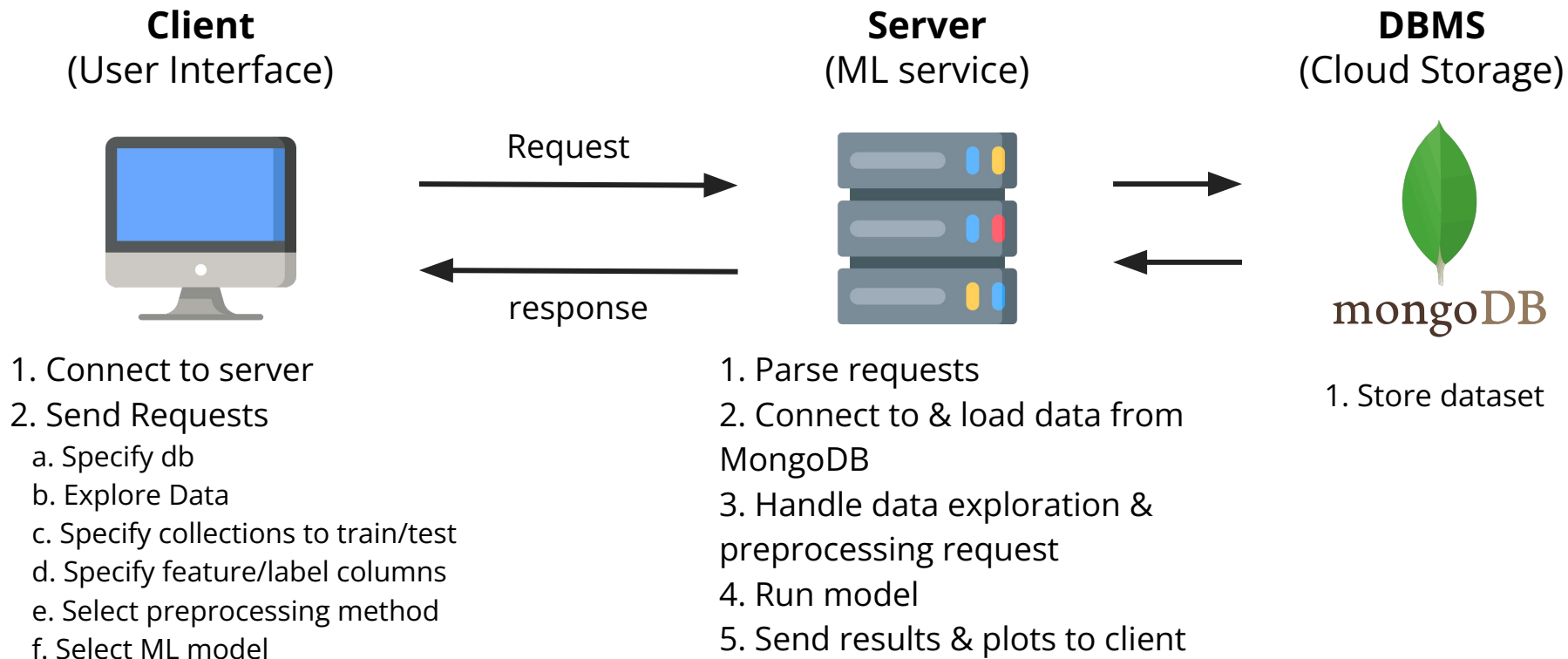
- 優點:

1. **遠端運行、本地接收:** 將機器學習模型部署在遠端server, 使用雲端或專門的計算資源來處理資料集和訓練模型, client端只需要下指令與接收結果。  
→ 大幅節省本地資源的使用, 並享受遠端服務提供的高效率計算能力。
2. **資料前處理:** 提供基本的前處理功能(scaling, duplicates, missing value, etc.)  
→ 由server負責與雲端DB互動, 不需自行撰寫函數或上傳/下載資料。
3. **使用者介面:** 使用者可直接選擇資料前處理方式、模型、資料視覺化的方式。  
→ 降低使用門檻, 增進人機互動的使用者體驗。

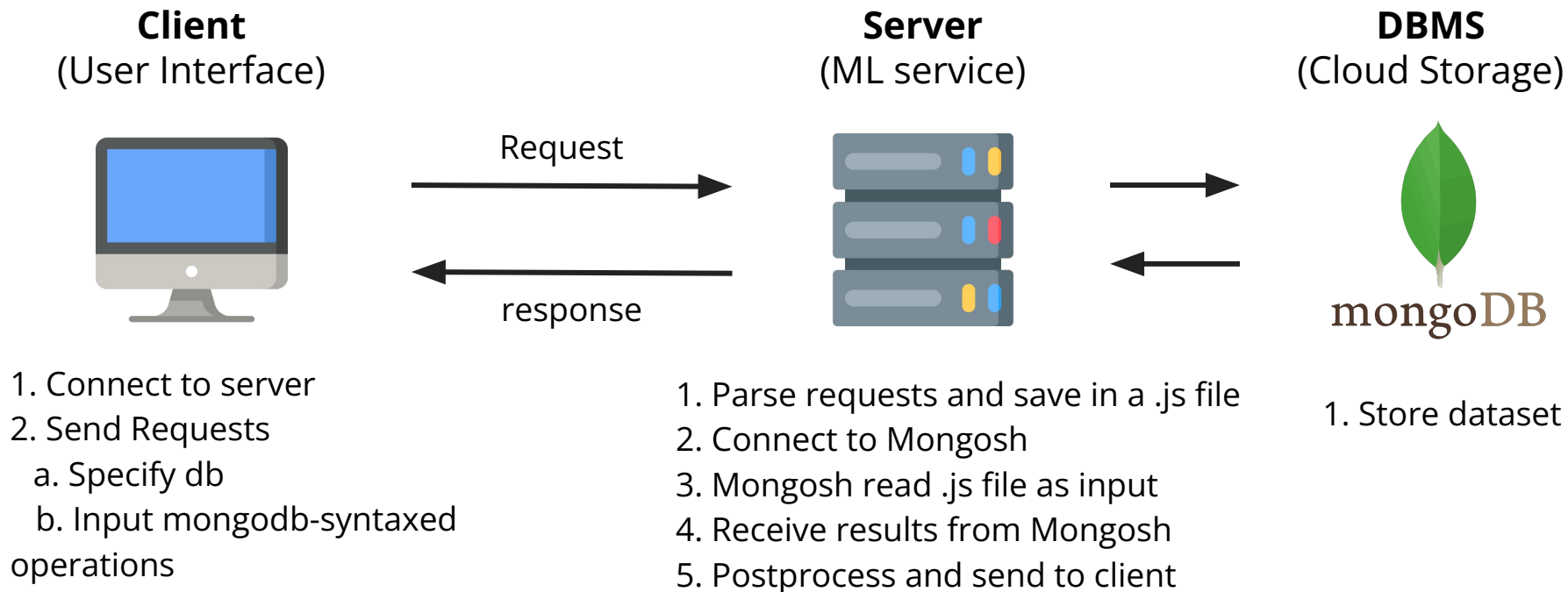
# Implementation

- 整合兩種介面：
  - Original MongoDB operations
  - Extended MongoDB for data exploration & preprocessing and model training
- 提供之功能
  - **Original MongoDB operations**
  - **Data Exploration & Analysis (e.g. feature box plot)**
  - **Data Preprocessing: Various re-scaling, imputing, remove outliers strategies**
  - **Model Selection & Model Training**
  - **Analyze prediction results (e.g. Heatmap...)**

# Implementation (Extended MongoDB)



# Implementation (Original MongoDB)





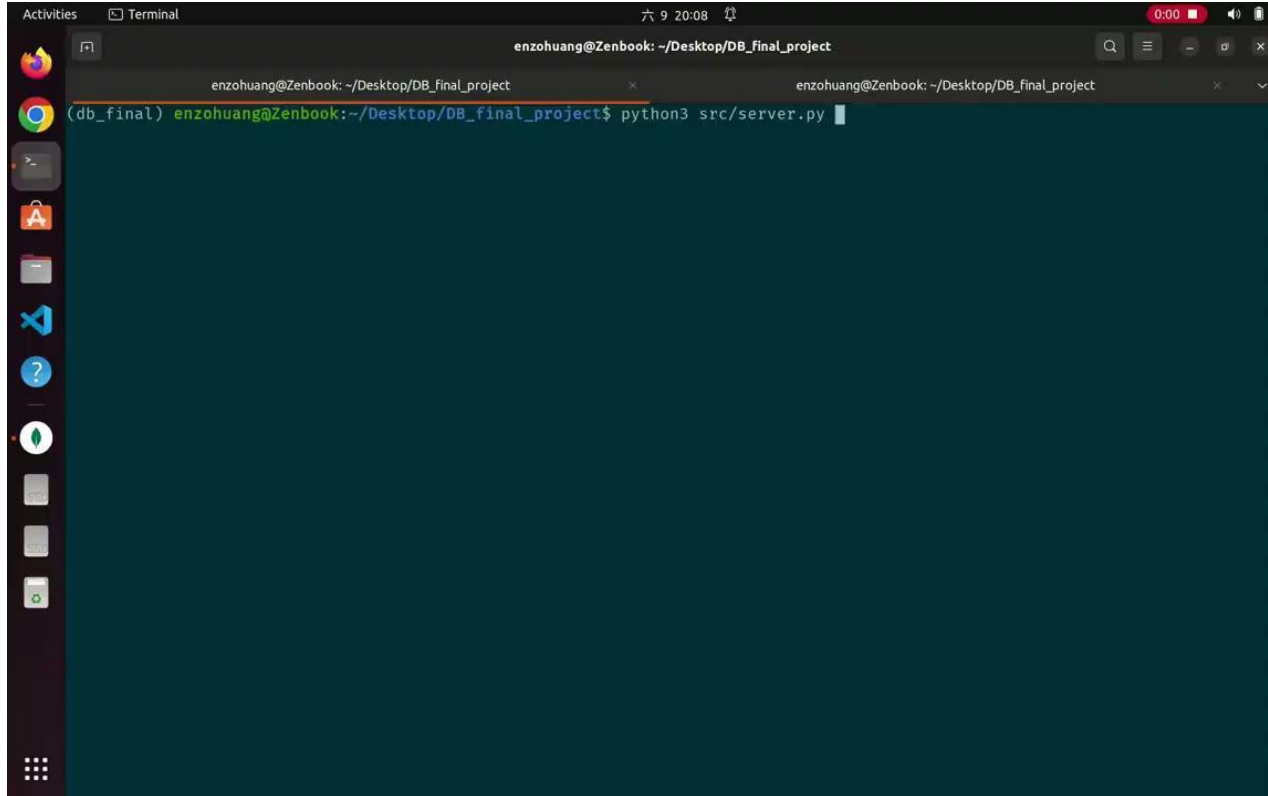
# Tools

- 使用之第三方套件、框架、程式碼
  - Programming: Python
  - Client: UI (not implemented yet)
  - Server: pymongo, socket...
  - Data preprocessing: sklearn, pandas, numpy
  - Data Exploration: seaborn, missingno, matplotlib
  - DBMS: MongoDB
- Data preprocessing functions:
  - Scaling, Removing duplicates, Handling missing value
- Data exploration functions:
  - Missing value indicator, numeric distribution

# Phase I Objectives

- 確定可以使用 PyMongo & other packages 完成對數據的分析和處理
- Data Exploration → Data Preprocessing → Model selection  
→ Model Training & Prediction 流程
- 使用 socket 實現 Client-Server 傳輸
- 支援原本的 MongoDB 操作

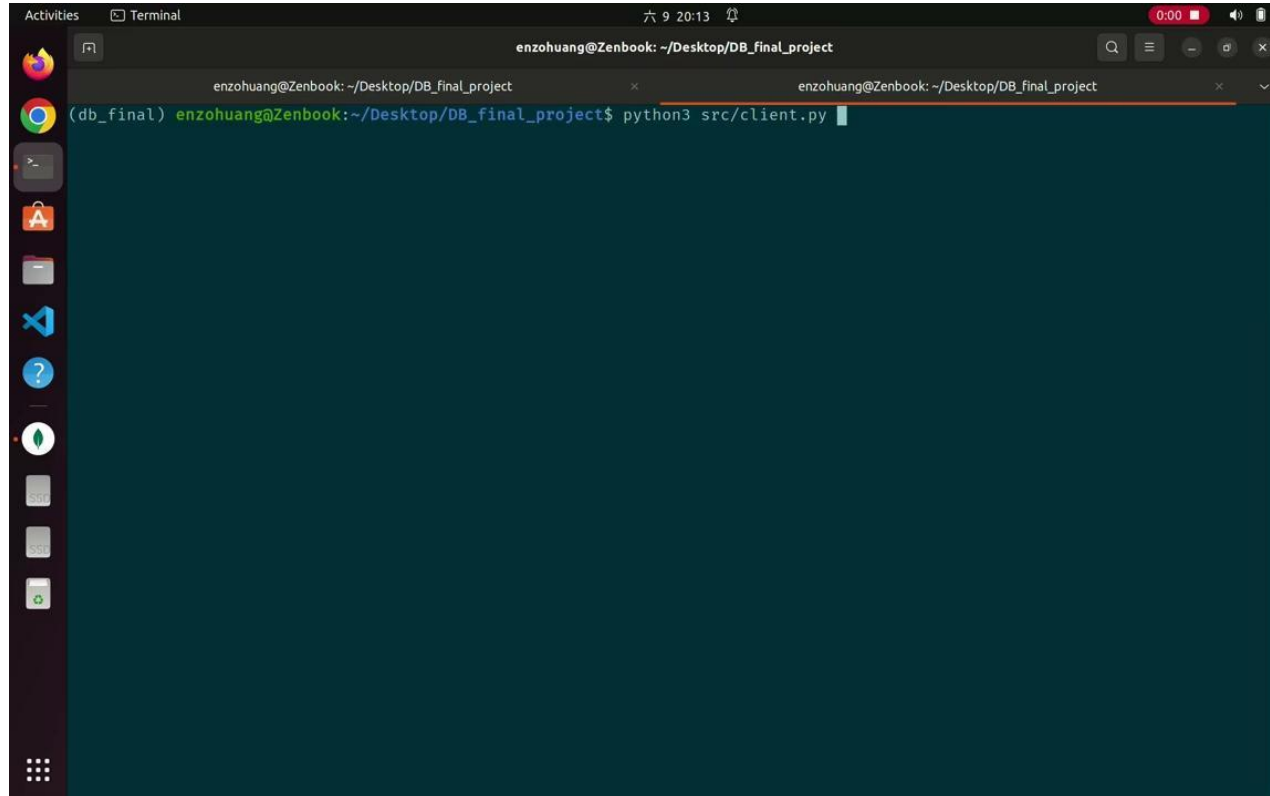
# Phase I Result- Data exploration 1



A screenshot of a Linux terminal window. The window title is "Terminal" and the current directory is "~/Desktop/DB\_final\_project". The prompt is "enzohuang@Zenbook: ~/Desktop/DB\_final\_project\$". The command "python3 src/server.py" has been entered and is being executed. The terminal output is currently blank. The window is part of a desktop environment with a sidebar on the left containing icons for various applications like Firefox, Chrome, and VS Code. The top status bar shows the time as 20:08 and a battery level indicator.

```
enzohuang@Zenbook: ~/Desktop/DB_final_project$ python3 src/server.py
```

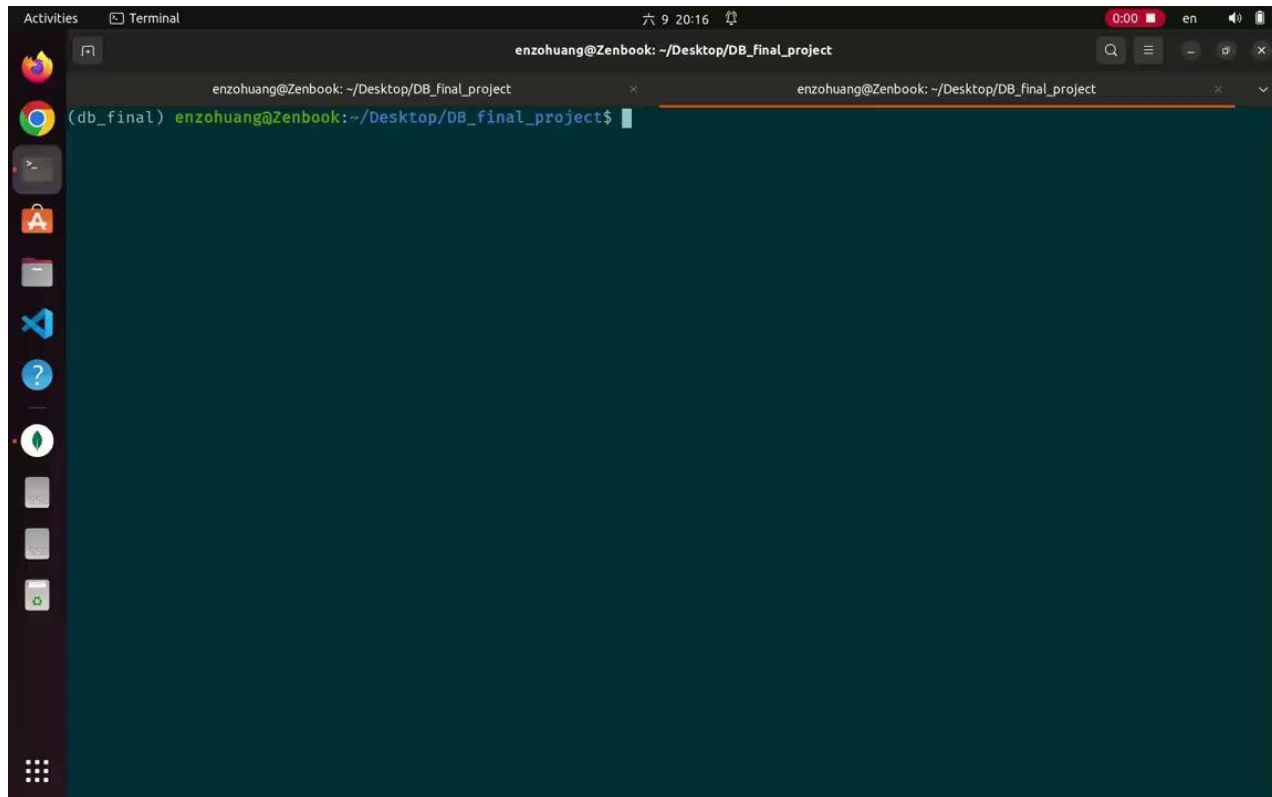
# Phase I Result- Data exploration 2



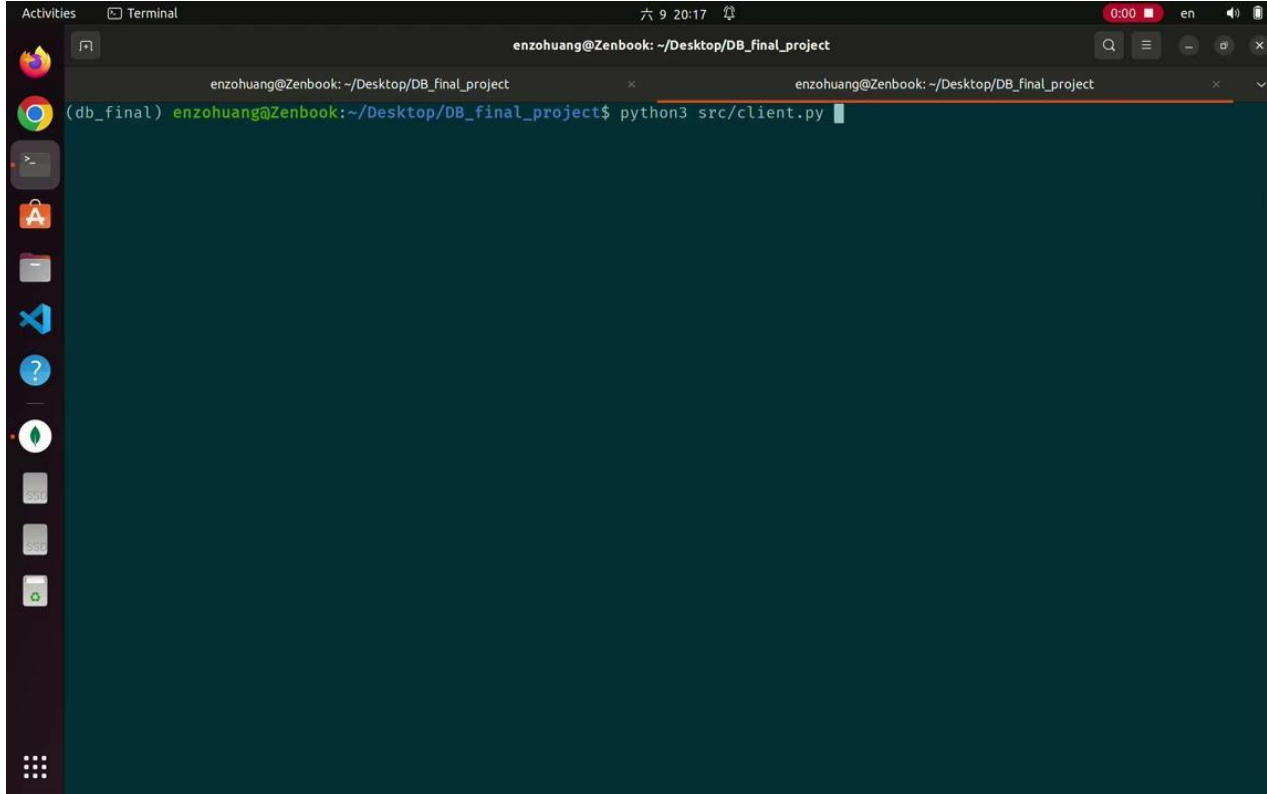
A screenshot of a Linux terminal window. The window title is "Terminal" and the current directory is "~/Desktop/DB\_final\_project". The prompt is "enzohuang@Zenbook: ~/Desktop/DB\_final\_project". The command "python3 src/client.py" has been entered and is being executed. The terminal output is currently empty. The window also shows a sidebar with various application icons and a top bar with system status indicators.

```
enzohuang@Zenbook: ~/Desktop/DB_final_project  
(db_final) enzohuang@Zenbook:~/Desktop/DB_final_project$ python3 src/client.py
```

# Phase I Result- ML Prediction



# Phase I Result- MongoDB operation

A screenshot of a Linux terminal window. The window title is "Terminal" and it shows the user "enzohuang@Zenbook" in the directory "~/Desktop/DB\_final\_project". The prompt is "(db\_final) enzohuang@Zenbook: ~/Desktop/DB\_final\_project\$". The command "python3 src/client.py" has been entered, and the cursor is at the end of the line. The terminal has a dark blue background. The window is part of a desktop environment with a sidebar on the left containing icons for various applications like Firefox, Chrome, and VS Code. The top of the window shows system status including the time "六 9 20:17" and language "en".

```
enzohuang@Zenbook: ~/Desktop/DB_final_project
(db_final) enzohuang@Zenbook: ~/Desktop/DB_final_project$ python3 src/client.py
```

# Phase II Objectives

1. 製作前端介面
  - 1.1. 提供 editor 讓使用者輸入 MongoDB 的指令
  - 1.2. 提供資料前處理、ML模型種類的選單, 取代CMD操作
  - 1.3. 讓使用者可以更輕易的操作, 有更好的使用體驗
2. 增加更多的資料探索、資料前處理、訓練模型的種類

# Limitation

1. 不夠客製化: 目前的 data 相關的 method 或是 model 都是 pre-defined 的, 使用者無法選擇沒有提供的 model 或是 method
2. 缺少身份驗證的機制 (e.g. 誰可以連到 mongoDB)
3. 無法偵測惡意的 MongoDB operation: 使用者可能會試圖執行惡意或不符合權限的指令 (如: 刪掉 db 裡面的所有資料), 但是 server 目前沒有檢查的機制



# Possible Extension

1. 增加更多的資料前處理、訓練模型的種類
2. 提供使用者能夠 extend 功能的格式化介面, 能夠輕鬆擴展並客製化自己常用的 Data methods 和 models
3. 權限控制: server 要可以辨別該使用者可以執行哪些操作

**THANK YOU**

# Demo Video Link

Video link: <https://youtu.be/31xufIYUrtU>

Github link: [https://github.com/a113062130630210/DB\\_final\\_project](https://github.com/a113062130630210/DB_final_project)