# CROWD FLOW ESTIMATION AND PREDICTION USING PASSIVE WIFI DATA

PROJECT ID: 2324ENG1001

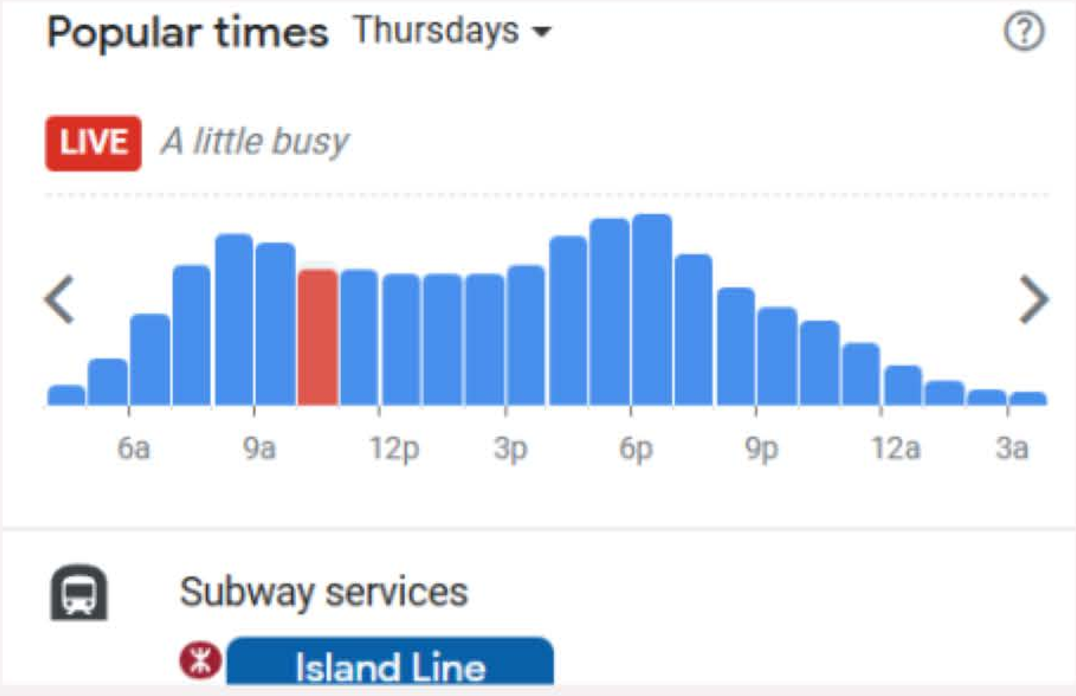
## INTRODUCTION

Our campus is getting more and more crowded. But can we know how crowded it is now in the library, a particular classroom, or a favorite canteen? In this project, we aim to build a technological solution to crowd flow estimation and prediction by leveraging the privacy-insensitive log data of the campus-wide WiFi network.

WiFi networks record client activities, e.g., the time when a client is connected/disconnected from an Access Point (AP). These data allow us to infer the crowd density at different places with advanced data analytics. From the WiFi logs, we capture key information regarding the positions and movement behaviors of devices, and can use it to infer the positions of the person carrying the device; on the other hand, the user id of the connected devices are masked in the WiFi logs we use, preserving the anonymity and privacy of the WiFi users. And, the whole project will use passive data, not requiring any assistance from the users.


## OBJECTIVES

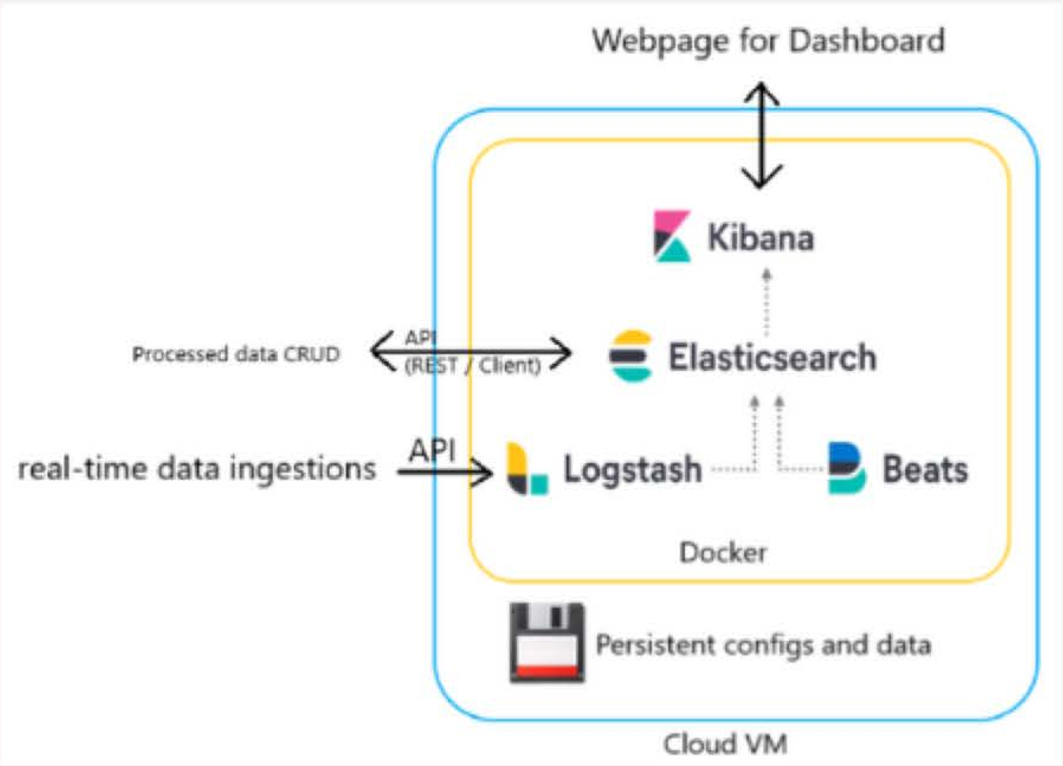
(1) Estimate the real-time crowd density at a particular place from the WiFi AP data,
(2) Analyze the historical crowd flow for different places,
(3) Predict the crowd flow variations, which together deliver a campus version of the "Popular times" feature in Google Maps.




## METHODOLOGY

### 1. OVERVIEW OF ARCHITECTURE

The major source of data comes from the WiFi logs, which consists of many different types of messages. However, the log file generated by 1 AC in 1 day can easily take ~3.5GiB of space and occupy millions of lines. This makes efficient processing of the data challenging. We built the following architecture to resolve the challenge:




### CHOICES OF TECH STACK USED

Scalability and Efficiency
ElasticSearch is a highly scalable and efficient NoSQL database which has inverted indexing scheme for the text fields in a document, which makes analysis of population at any selected location over a time period very efficient, since it does not need to query all documents to look for which records correspond to the chosen location. And, Elastic Stack provides powerful extensions that can be used in our pipeline.

Ease of Deployment
We deployed the ElasticSearch in Docker container on a cloud VM, so as to reduce the complexity of configurations and ease of access.

### 2. PREPROCESSING OF DATA

To efficiently use the data to estimate the location of a user, our main focus is on the following types of message:
(1) Authentication messages arising from new connections made between AP and user device. This indicates a new person has entered the campus / a specific location;
(2) Handover of a device from an AP to the next AP. This shows how a user moves within campus, which determines the population density of selected locations;
(3) Disconnection, which indicates the user leaves the campus.

```
Sep 25 09:40:52 2023 notice 147.8.123.13 147.8.123.13 09:40:52.560.2+08:00 AC6800V-HKU-5 %%01CM/5/
USER_ACCESSRESULT(s)[4254]:[WLAN_STA_INFO_AUTHENTICATION]ACMAC:30-e9-8e-95-f8-78;ACNAME:AC6800V-HKU-5;
APMAC:30-c5-0f-68-91-e0;APNAME:TW_CCB_LG1STAIR_5760_51;SSID:HKU;RADIOID:0;USER:XXXXX;MAC:7e-ed-a6-5d-e0-c5;
IPADDRESS:-;IPV6ADDRESS:-;TIME:1695606052;ZONE:UTC+0800;DAYLIGHT:false;ERRCODE:0;RESULT:success;USERGROUP:NULL;
AUTHENPLACE:RADIUS;CIB ID:131598;INTERFACE:Wlan-Dbss35109;ACCESS TYPE:802.1x;VLAN:268;SelectRDSIP:-;
SelectERRCODE:-;RDSIP:147.8.235.45;Portal TYPE:-;AUTHID=2336582085;AuthfailType:None;AUTHPROTOCOL:-;
```

We built parsers to parse the logs before ingesting them into ElasticSearch, which reduces the overhead when analyzing the data.
And, we parsed actual locations of the access points according to their naming rules, so we can later analyze population density of specific locations.
As every device has their unique MAC address, we can use the address as an identifier of a person in campus, so as to estimate the crowd.

### 3. VISUALIZATION OF DATA

We used Kibana as the visualization tool, as it is integrated in Elastic Stack, and has powerful functions for data visualization. One can easily access the Kibana page at http://35.241.125.13:5601/.
A dashboard was built on Kibana, allowing the users to view the overall view of the population density, the number of people visiting any specific location, and also the 2D heat map for an area.

### 4. ANALYSIS OF HISTORICAL DATA

We analyzed the historical data with the WiFi logs from semester 1, 2023-2024. Assumption that each person carries 1 device connected to the WiFi was made.
Then, at a given location, to estimate the number of people currently inside the area, we can count the number of distincet devices inside the area over time.
On the other hand, analyzing devices may lead to double-counting (one person may carry >1 device), which can lead to over-estimation.

### 5. REAL-TIME POPULATION DENSITY

We allow real-time data ingestion and analysis by allowing data ingestion via the APIs of ElasticSearch, so the server which generate these logs can first process them and then use the API to upload to the database. Kibana can thus visualize real-time data.

### 6. POPULATION FLOW PREDICTION

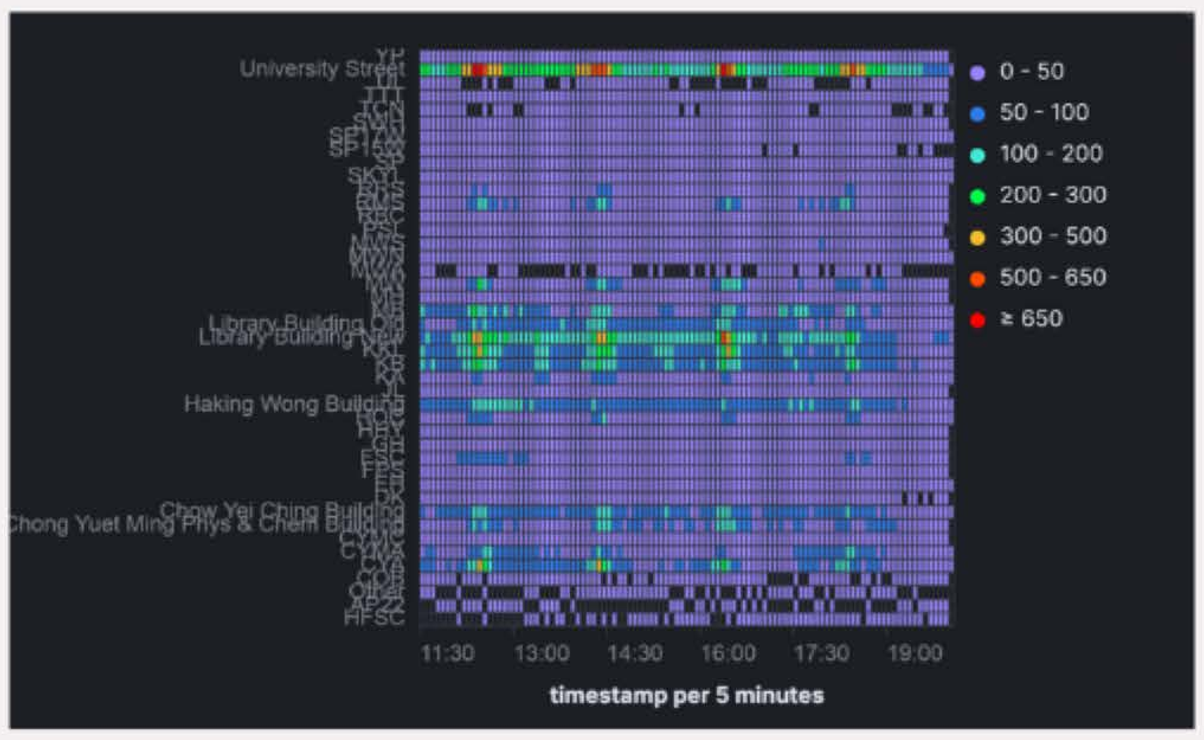
With the historical trends of the data, we can effectively find the patterns of the historical population flow and use it to predict the future flow.
We are currently trying to apply deep learning methods on the graph (GNN) to make predictions.

## RESULTS

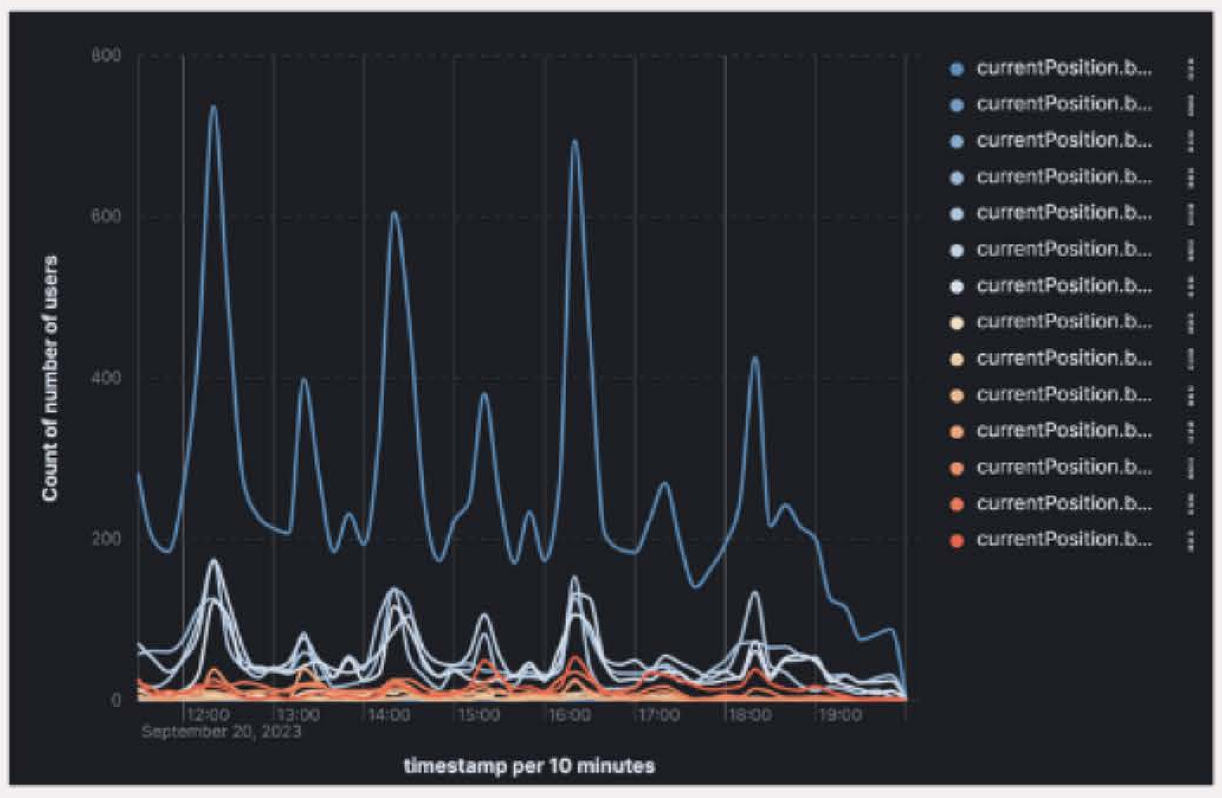
### CROWD FLOW ANALYSIS

An estimate of population densities for all major campus facilities was successfully visualized, from various levels including the heat map of an entire campus, then down to individual buildings/places, then, when applicable, an estimate of individual floors within that building.

Below is a heat map for the entire campus on Septemer 20, 2023, from 11:30 a.m. to 7 p.m.:



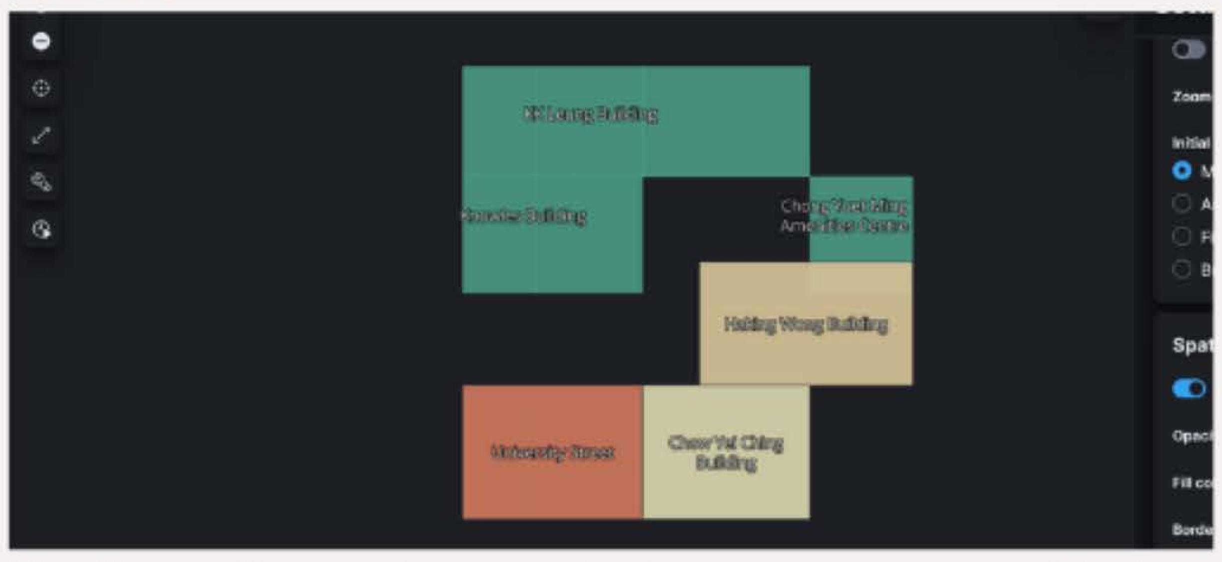

Consider the lower level analysis of KK Leung Building, Including a graph for the entire building, the top blue line, combined with the number of users for individual levels in the building, from LG3 (light blue) through Level 9 (red).




The lines can be hidden if one wishes to focus on a specific level, and a detailed observation of the numbers can easily be made:


## CONCLUSION

Lastly, these visualization maps can be customized into many other formats. A more user-friendly approach, for example, would be to use region maps that implement custom-written GeoJSON files to represent crowd densities and flow:




The illustration can be modeled to represent a specific floor of a building, where different color gradients depict different magnitudes of crowd densities in a given room, space, etc.

STUDENTS
Jiang Zhejun, Year 3, Department of Computer Science, Faculty of Engineering
Abraham Chandafa, Year 2, Department of Computer Science, Faculty of Engineering

EUREKA SUPERVISOR
Dr. Wu Chenshu, Department of Computer Science, Faculty of Engineering