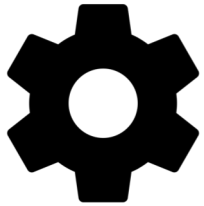
Determining Home and Work from X-Mode Data



Use Cases

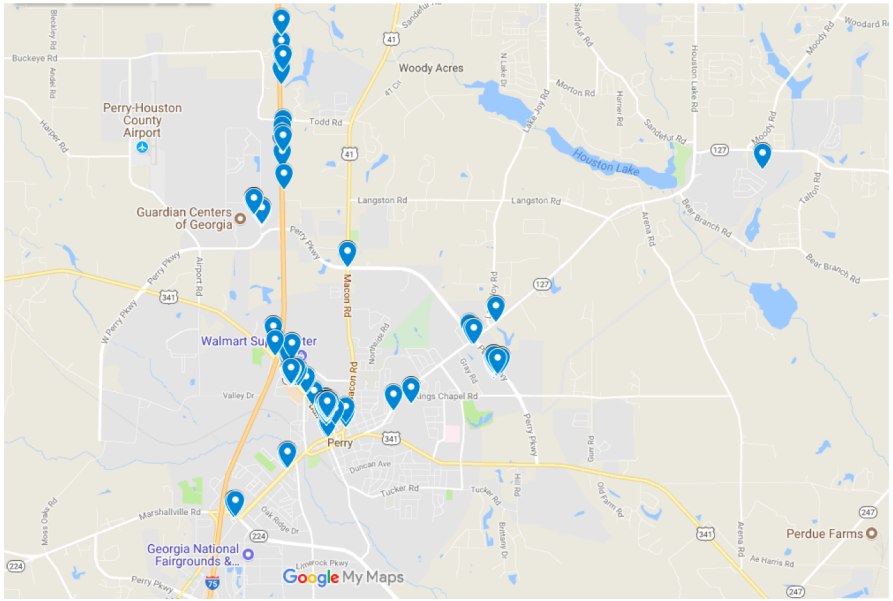
- Take home data out of datasets to protect consumers
- Determine if a user is at or leaving home/work, and change SDK collection methods
  - increase radius for iOS at home/work
  - turn on driving settings after determining commute
- Mailing campaigns after a person has visited a location

1.

Raw Data Mapped

Google My Maps

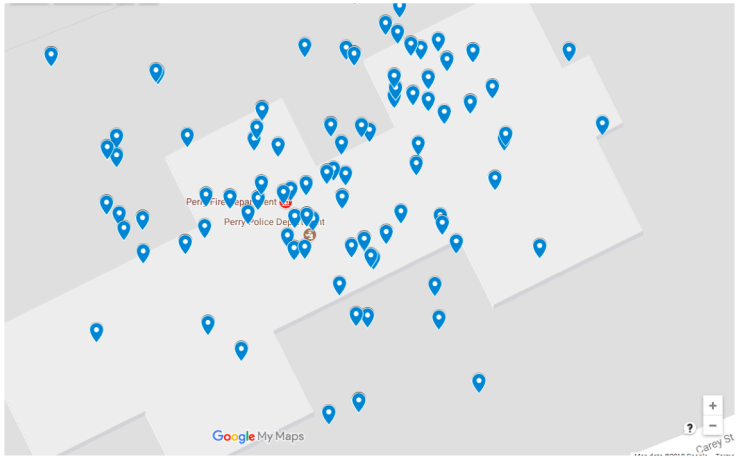
From a human perspective, it’s easy to see clusters of points and identify the building as a workplace or neighborhood



Google My Maps

Work

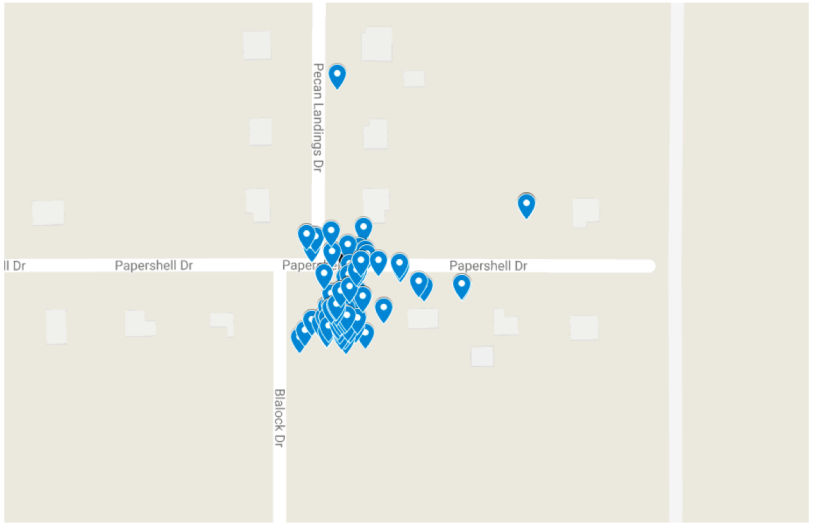
This person either works at the police dept or fire dept



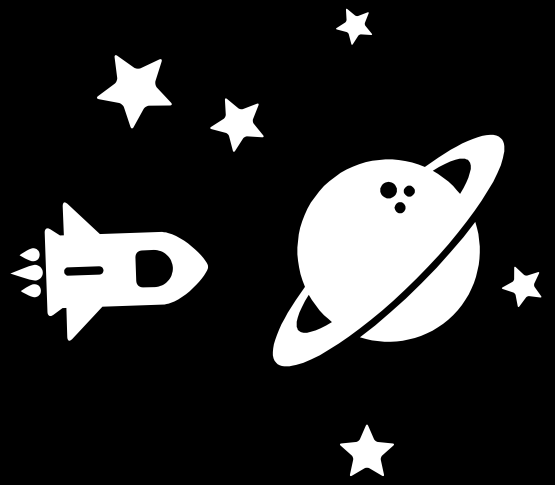
Google My Maps

Home

And lives on Papershell Dr.



How can we figure the same thing out with machine learning?



2.

Data Analysis

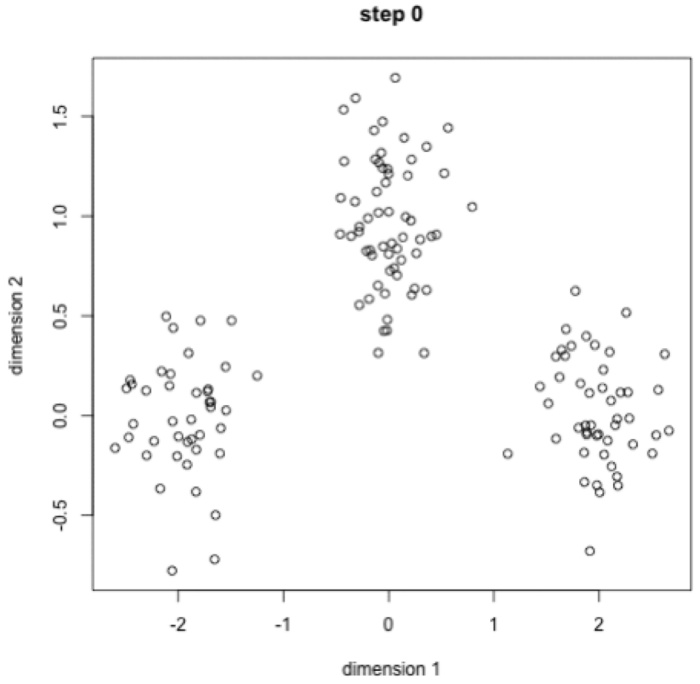
Filtering Data - We Want:

- Weekdays
- Time of day: 10 a.m-4 p.m & 9 p.m-5 a.m
- Speed under 3 meters/second (around 6.7 mph)
- Horizontal accuracy under 200 meters

These criteria reduce a user's data points from thousands to hundreds and makes it easier to identify clusters

K-Means Clustering

This algorithm puts down “centroids”, or center points, and adjusts them until they correctly fit the groups. Each cluster has a centroid.



Our Algorithm

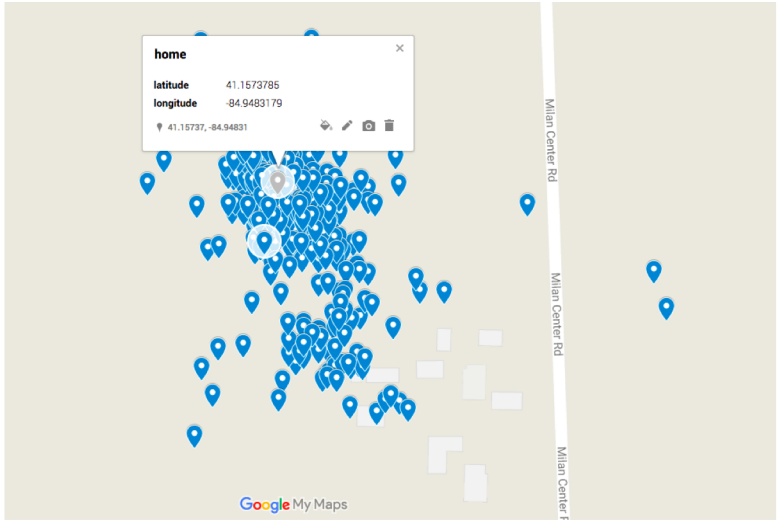
1. Run K-Means clustering
2. Check if: in the two largest clusters, the distance between the centroid and any point in that cluster exceeds 200 meters
3. If it does, run K-Means clustering again on a larger number of clusters, hoping to increase accuracy
4. If it does not, then the two largest groups are assigned home and work (largest = home, second largest = work)

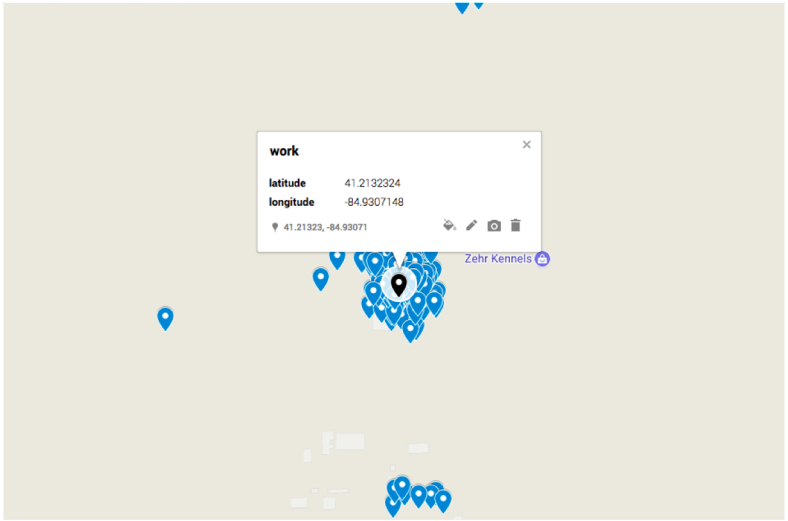
3.

Findings

Results

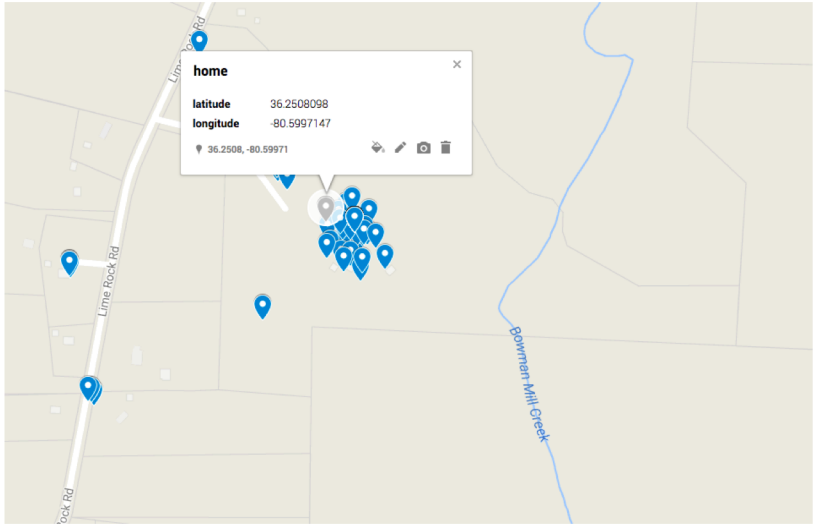
Sometimes, the algorithm works really well...

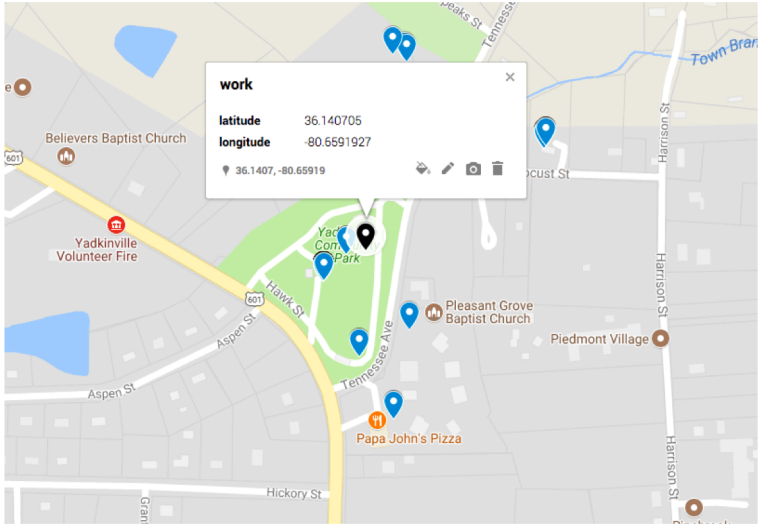




Results

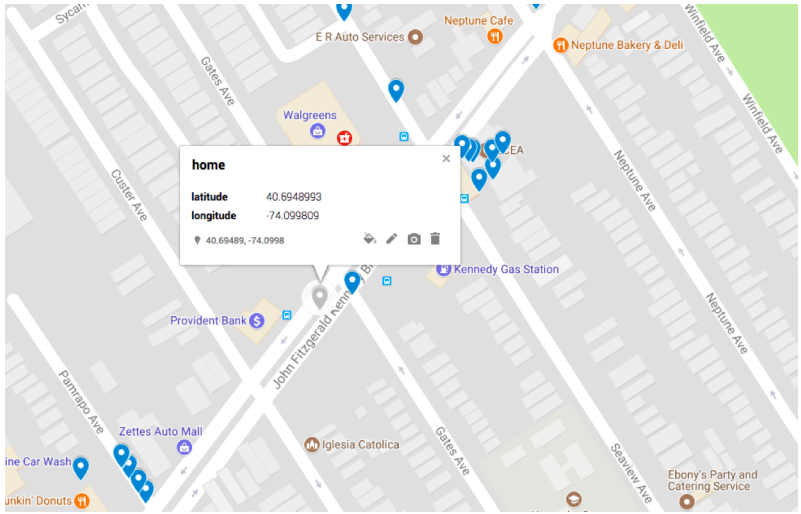
Sometimes, less well...

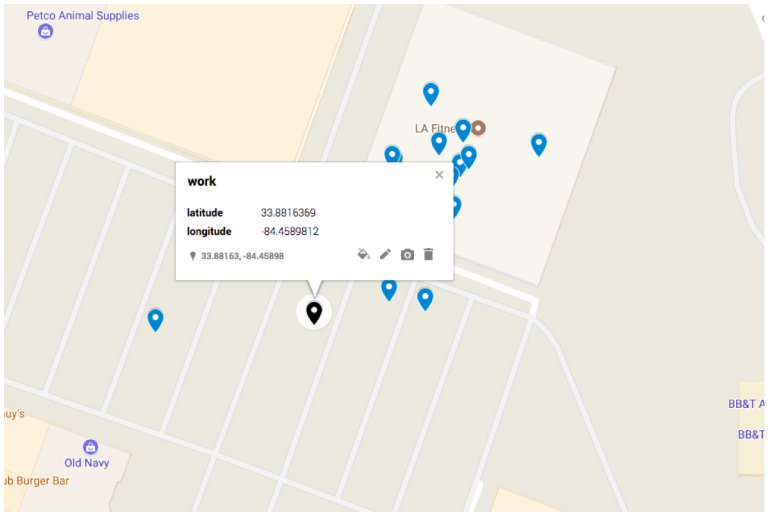




Results

And other times, not so well...





Analysis

- Users with less points (< 500 points) after filtering had worse home and work predictions
- Clusters are sometimes off center
- Home and work are sometimes swapped
- Had to create 15 clusters, on average, before getting a prediction
- In order to validate the algorithm, we need a test set that has home and work appended to it

Confidence Score

Since we noticed that users with less than 500 points had lower accuracy, we came up with a logarithmic confidence score for our predictions:

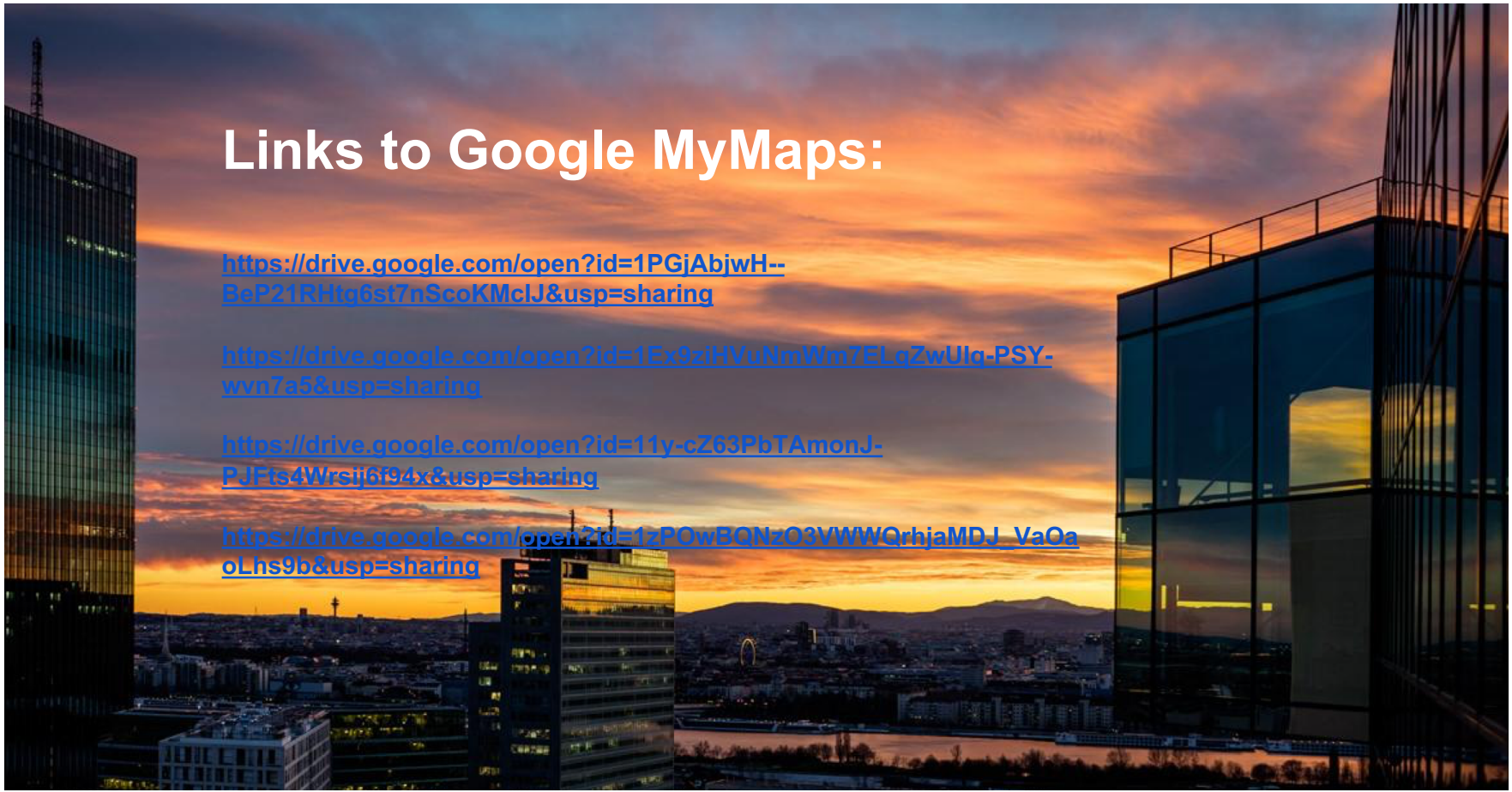
$$confidence = \frac{1}{1 + e^{-\left(\frac{x}{70} - 5\right)}}$$

where x is the number of points

Basically, users with above 500 points will have very high confidence, and users with below 500 points will have very low confidence

Conclusions

- Given enough points (~500-1000), K-Means clustering is generally at finding the two most visited place by a user, but might mix up home and work
- If SDK collection methods change, might have to depend on dwell time in addition to clustering
- Current algorithm needs to be adjusted to scale for millions of users
- Determining home and work might be difficult for people who have abnormal schedules, but is definitely achievable for everyone else with high accuracy (especially home)


Links to Google MyMaps:
https://drive.google.com/open?id=1PGjAbjwH--BeP21RHtg6st7nScoKMcIJ&usp=sharing
https://drive.google.com/open?id=1Ex9ziHVuNmWm7ELoZwUlq-PSY-wvn7a5&usp=sharing
https://drive.google.com/open?id=11y-cZ63PbTAmonJ-PJFts4Wrsij6f94x&usp=sharing
https://drive.google.com/open?id=1zPOwBQNzO3VWWQrhjaMDJ_VaOaoLhs9b&usp=sharing