Using Python and Data Science to tackle real-time transportation problems at Lyft

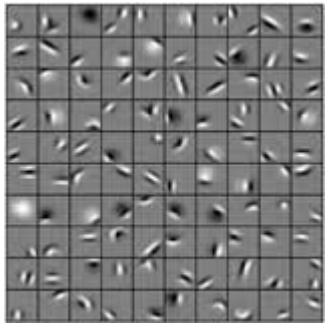
Clement Jambou

About me

Imperial College London

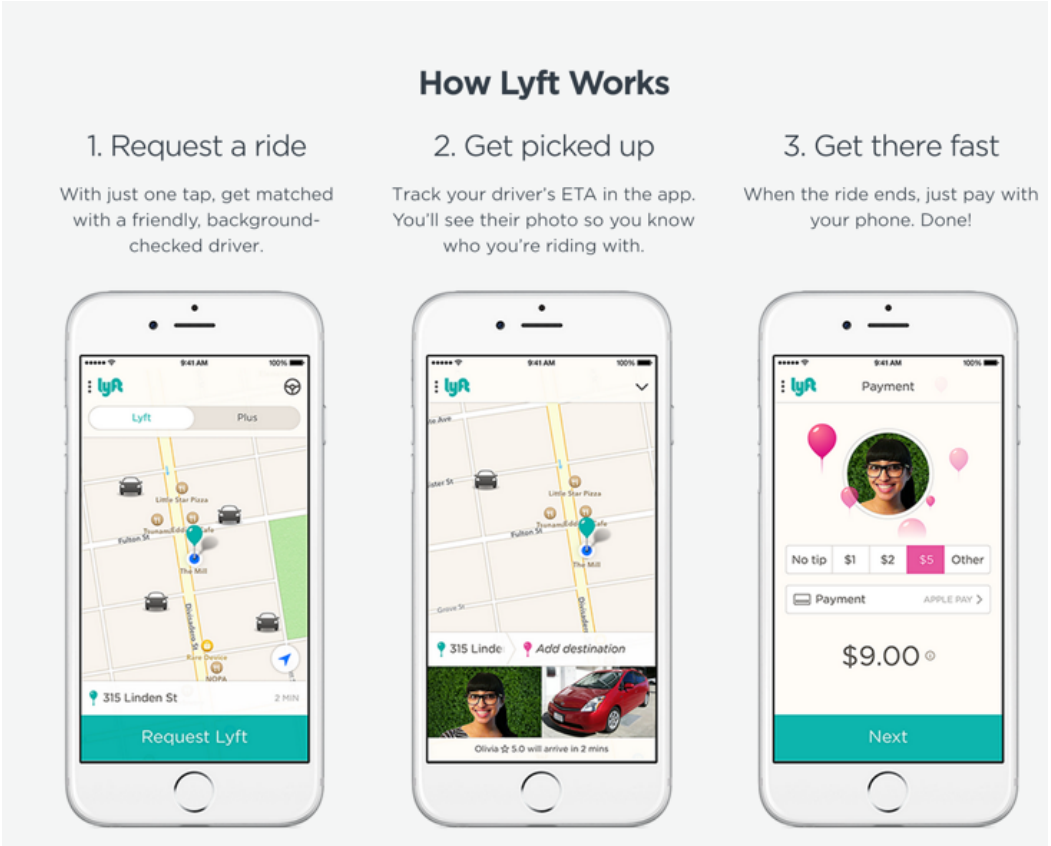
isae
Institut Supérieur de l'Aéronautique et de l'Espace
SUPAERO

Coupe de France de Robotique



What is Lyft

- over 60 cities in the US !
- 5X growth in 2014



Data Science at Lyft

Many complex problems require optimizations:

- Dynamic pricing
- Dispatching drivers efficiently
- Lyft Line Passenger matching
- Accurate ETAs
- Marketplace optimization : matching demand and supply through incentives and promotions
- Events detection
- Fraud ...

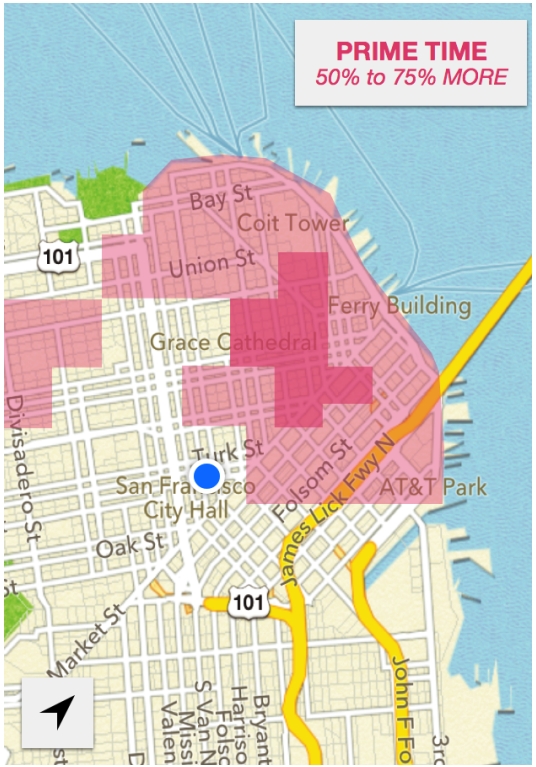
These problems becomes even more challenging with the geolocation component.

Dynamic Pricing: Prime Time

Event : A big concert is happening, demand >> supply

The prices will automatically increase in subregions to encourage drivers to drive in that direction:

Where/ When do we use Prime Time ?



Dispatching Drivers Efficiently

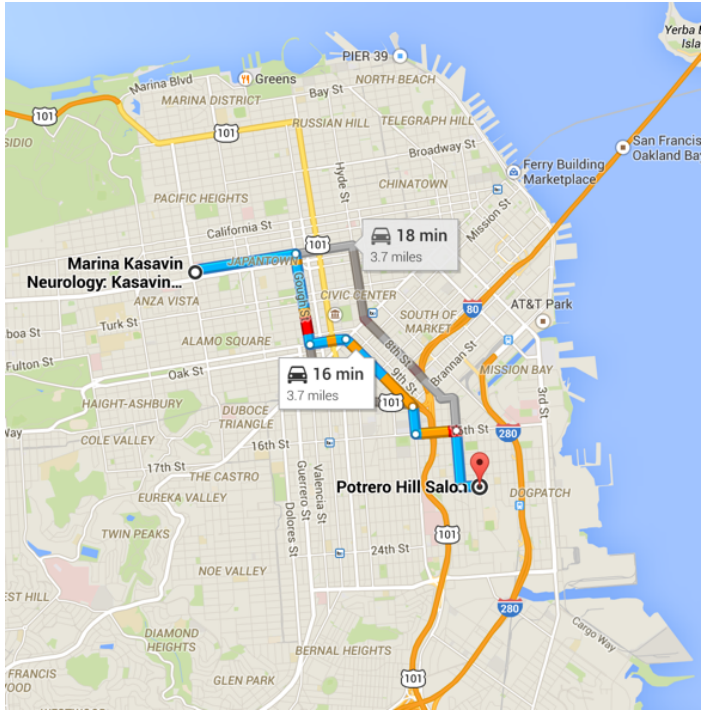
Request arrives in real-time:

Which driver should we dispatch ?

Should we decide not to dispatch a driver that is too far away ?

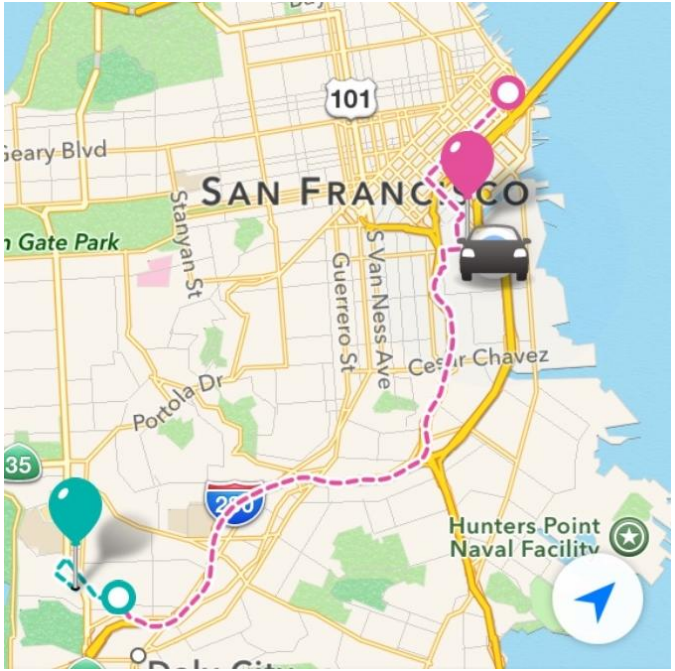
Should our decision depend on the time a driver has been waiting ?

How do we minimize pick up time and idle time of the drivers to optimize the pool of drivers ?



Lyft Line Matching

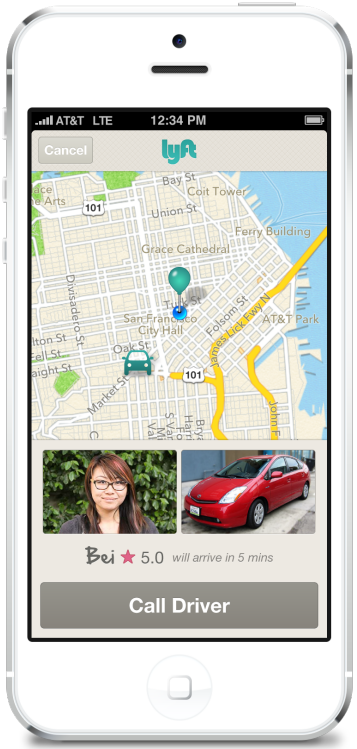
Matching Passenger going in the same direction.



Matching to optimize efficiency and reduce cost.

Accurate ETAs (Estimated Time of Arrival)

- Knowing when the driver will arrive or when you will reach your destination is useful information for the passenger.
- Order the drivers for dispatch based on ETAs
- We guarantee prices for Lyft Line, but drivers are paid depending on the distance and time of the route.



Python usage

We rely heavily on python in Data Science as well as for backend services at Lyft.

- numpy, scipy, matplotlib
- scikit-learn
- pandas, geopandas
- geohash, haversine, shapely

Geolocation Data

Many of our problems uses geolocation data. We use Open Source libraries and develop our own analysis and visualization tools.

- http://geojson.io/
- Demo Ipython Notebook

Data Pipeline

We use Amazon Redshift as our Data warehouse.

Pandas is directly integrated with redshift, so that we get our data in the right format from any SQL query.

Simulation

When the problem cannot be solved directly using historical data we often rely on simulations:

- replay a week of sessions data
- What would happen with more/less drivers
- How would our metrics change with a different pricing/dispatch algorithm ?

Our simulator is entirely written in python, runs locally or in parallel and uses the same algorithms used in production.

Questions

clement@lyft.com