MASTER OF SCIENCE
IN ENGINEERING

New Taxi Agencies

HES-SO, Big Data Analytics

Romain Claret, Jämes Ménétrey and Damien Rochat

Dataset

Provided by **Chris Whong**, Urbanist, New York City

Trips and fares taxis in New York City (2013)

2x 12 CSV files (one per month), total of **173’179’759** records

Trips dataset: Taxi identification, start and end time, GPS coordinates

Fares dataset: Taxi identification, start time, amount, payment type, tip

Goals

In order to open new taxi agencies and increase customer satisfaction (and profit):

(1) Find the best spots to find customers
(2) Identify the most profitable spots

Preprocessing

Removed out of bound GPS coordinates (1.75% of the records)

Removed invalid fares (1% of the records)

Machine Learning

PySpark

K-means, centroids for clusters of best locations for new agencies

Customers satisfaction:

- Using pickups coordinates

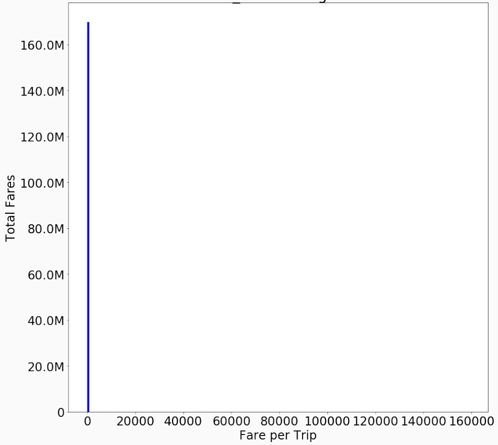
Profit:

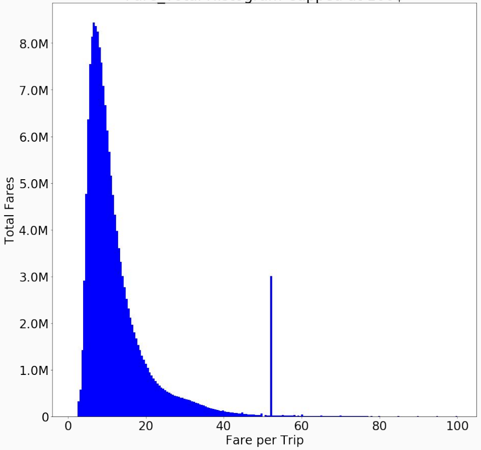
- Using pickups coordinates
- Weighted with fare

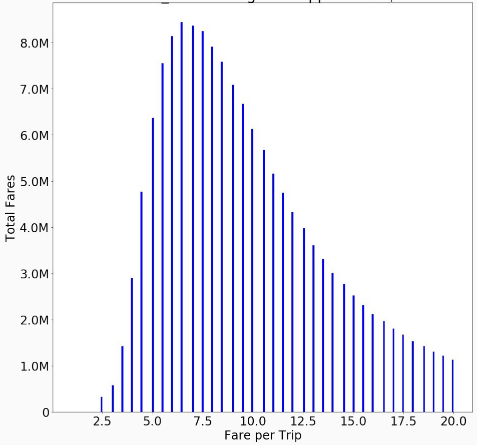
Machine Learning

total fare = amount of the fare + surcharge

maximum fare: 158’995.8125 $


Fare_Total Histogram


Fare_Total Histogram Capped at 100$


Fare_Total Histogram Capped at 20$

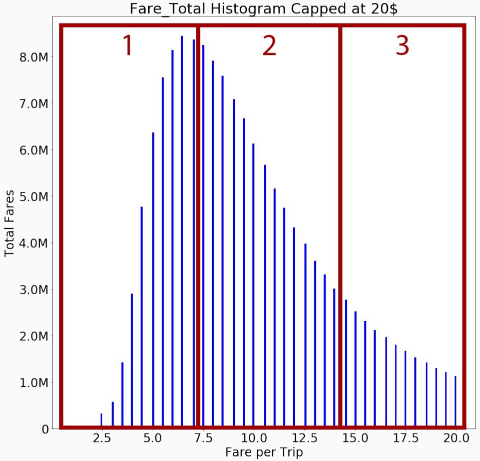
Optimization

Fare weight: proportionally duplicated records to add attraction

- Increased the dataset size of 2’575 %

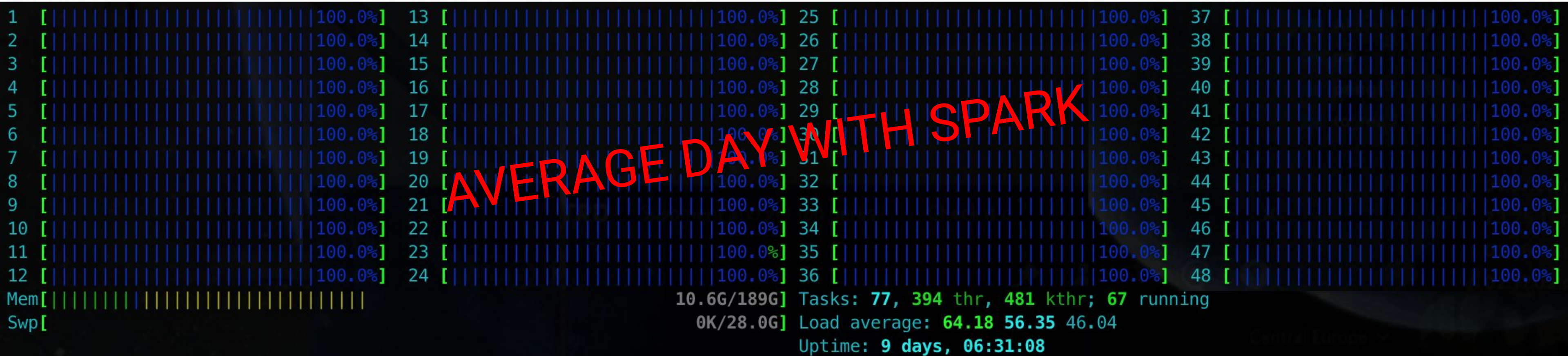
Sampling fares in groups, example:

- 05.0$ → not cloned
- 11.5$ → cloned once
- 20.0$ → 3 cloned twice



Testing and evaluation

- Custom softwares for data transformations
- Our own Spark cluster
- Try, Fails and Meta Parameterizing

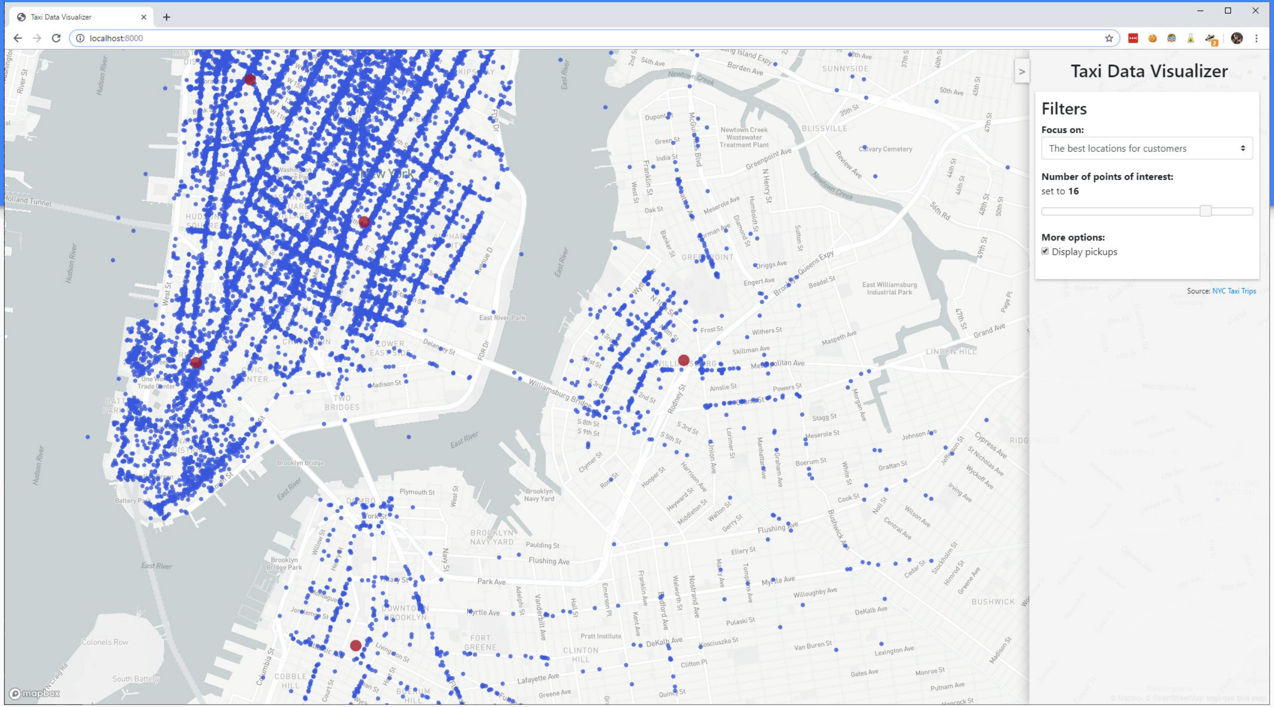


```
1  [||||||||||||||||||||||||100.0%]  13 [||||||||||||||||||||||||100.0%]  25 [||||||||||||||||||||||||100.0%]  37 [||||||||||||||||||||||||100.0%]
2  [||||||||||||||||||||||||100.0%]  14 [||||||||||||||||||||||||100.0%]  26 [||||||||||||||||||||||||100.0%]  38 [||||||||||||||||||||||||100.0%]
3  [||||||||||||||||||||||||100.0%]  15 [||||||||||||||||||||||||100.0%]  27 [||||||||||||||||||||||||100.0%]  39 [||||||||||||||||||||||||100.0%]
4  [||||||||||||||||||||||||100.0%]  16 [||||||||||||||||||||||||100.0%]  28 [||||||||||||||||||||||||100.0%]  40 [||||||||||||||||||||||||100.0%]
5  [||||||||||||||||||||||||100.0%]  17 [||||||||||||||||||||||||100.0%]  29 [||||||||||||||||||||||||100.0%]  41 [||||||||||||||||||||||||100.0%]
6  [||||||||||||||||||||||||100.0%]  18 [||||||||||||||||||||||||100.0%]  30 [||||||||||||||||||||||||100.0%]  42 [||||||||||||||||||||||||100.0%]
7  [||||||||||||||||||||||||100.0%]  19 [||||||||||||||||||||||||100.0%]  31 [||||||||||||||||||||||||100.0%]  43 [||||||||||||||||||||||||100.0%]
8  [||||||||||||||||||||||||100.0%]  20 [||||||||||||||||||||||||100.0%]  32 [||||||||||||||||||||||||100.0%]  44 [||||||||||||||||||||||||100.0%]
9  [||||||||||||||||||||||||100.0%]  21 [||||||||||||||||||||||||100.0%]  33 [||||||||||||||||||||||||100.0%]  45 [||||||||||||||||||||||||100.0%]
10 [||||||||||||||||||||||||100.0%]  22 [||||||||||||||||||||||||100.0%]  34 [||||||||||||||||||||||||100.0%]  46 [||||||||||||||||||||||||100.0%]
11 [||||||||||||||||||||||||100.0%]  23 [||||||||||||||||||||||||100.0%]  35 [||||||||||||||||||||||||100.0%]  47 [||||||||||||||||||||||||100.0%]
12 [||||||||||||||||||||||||100.0%]  24 [||||||||||||||||||||||||100.0%]  36 [||||||||||||||||||||||||100.0%]  48 [||||||||||||||||||||||||100.0%]
Mem[||||||||||||||||||||||||||||||||                   10.6G/189G]  Tasks: 77, 394 thr, 481 kthr; 67 running
Swp[                                                      0K/28.0G]  Load average: 64.18 56.35 46.04
                                                                     Uptime: 9 days, 06:31:08
```

AVERAGE DAY WITH SPARK

Results

Demo time



Taxi Data Visualizer

Filters

Focus on:

The best locations for customers

Number of points of interest:

set to **16**

More options:

☑ Display pickups

Source: NYC Taxi Trips

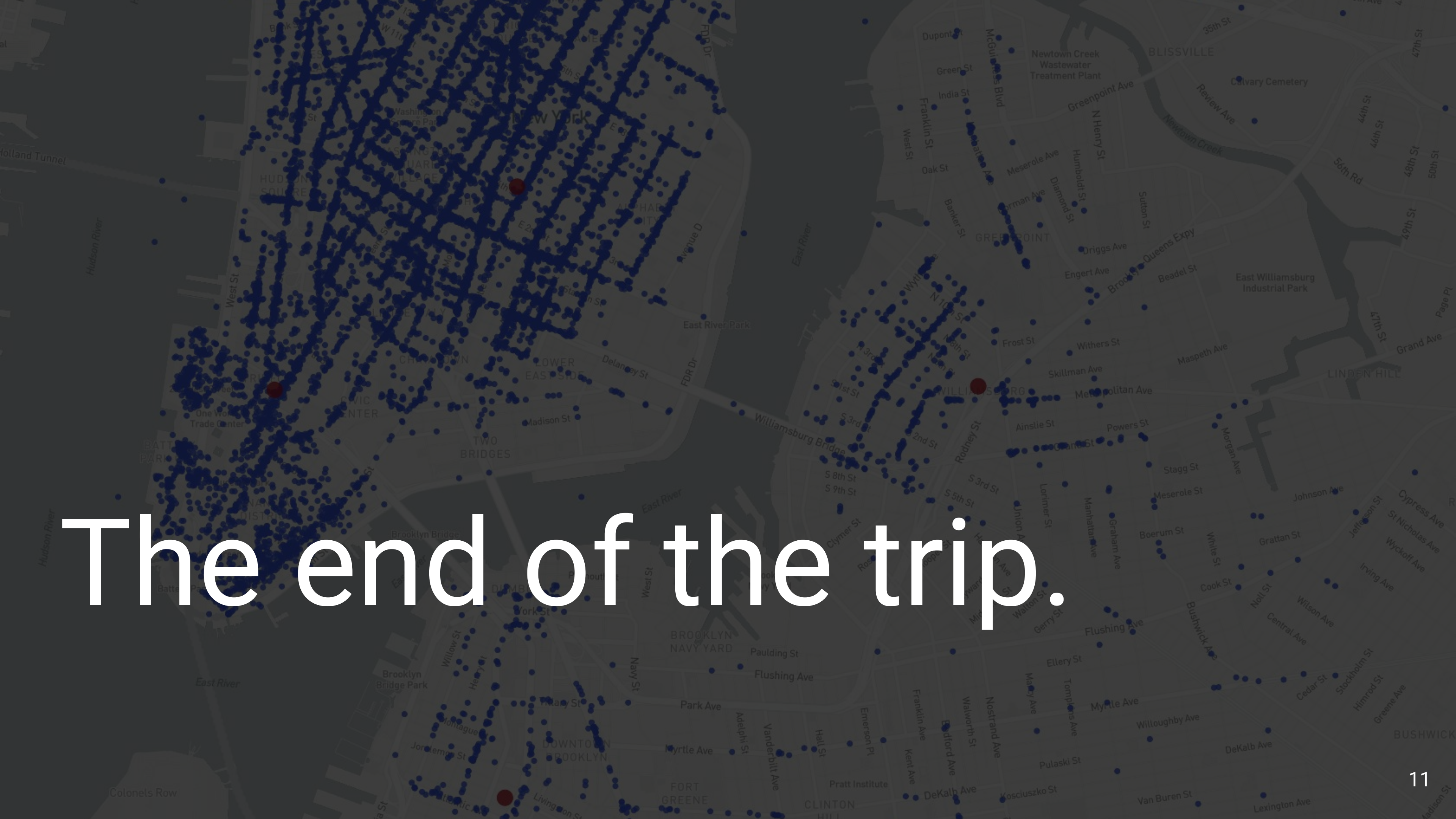
http://bit.do/nyc-taxis

Conclusion

Best spots for customers are also the most profitable ones :)

Next steps:

- More processing time
- Zoom in data
- Prediction
- More data



The end of the trip.

https://github.com/ZenLulz/hesso-bigdata-analytics