
BAY AREA BikeShare

# Predicting Station Status For Load Balancing

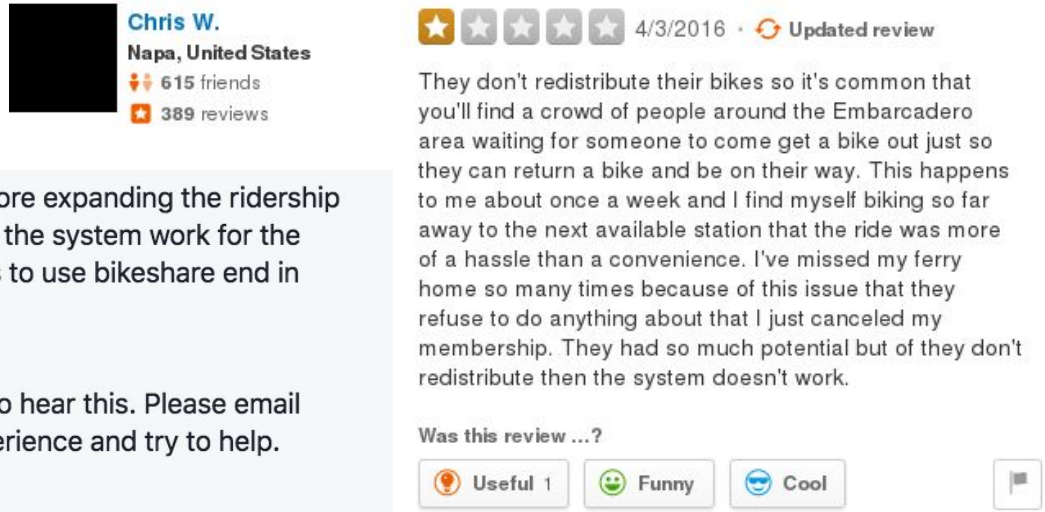
Cole McGinn Jakob Roberts Jordan Vlieg

Ramunas Wierzbicki James Woo

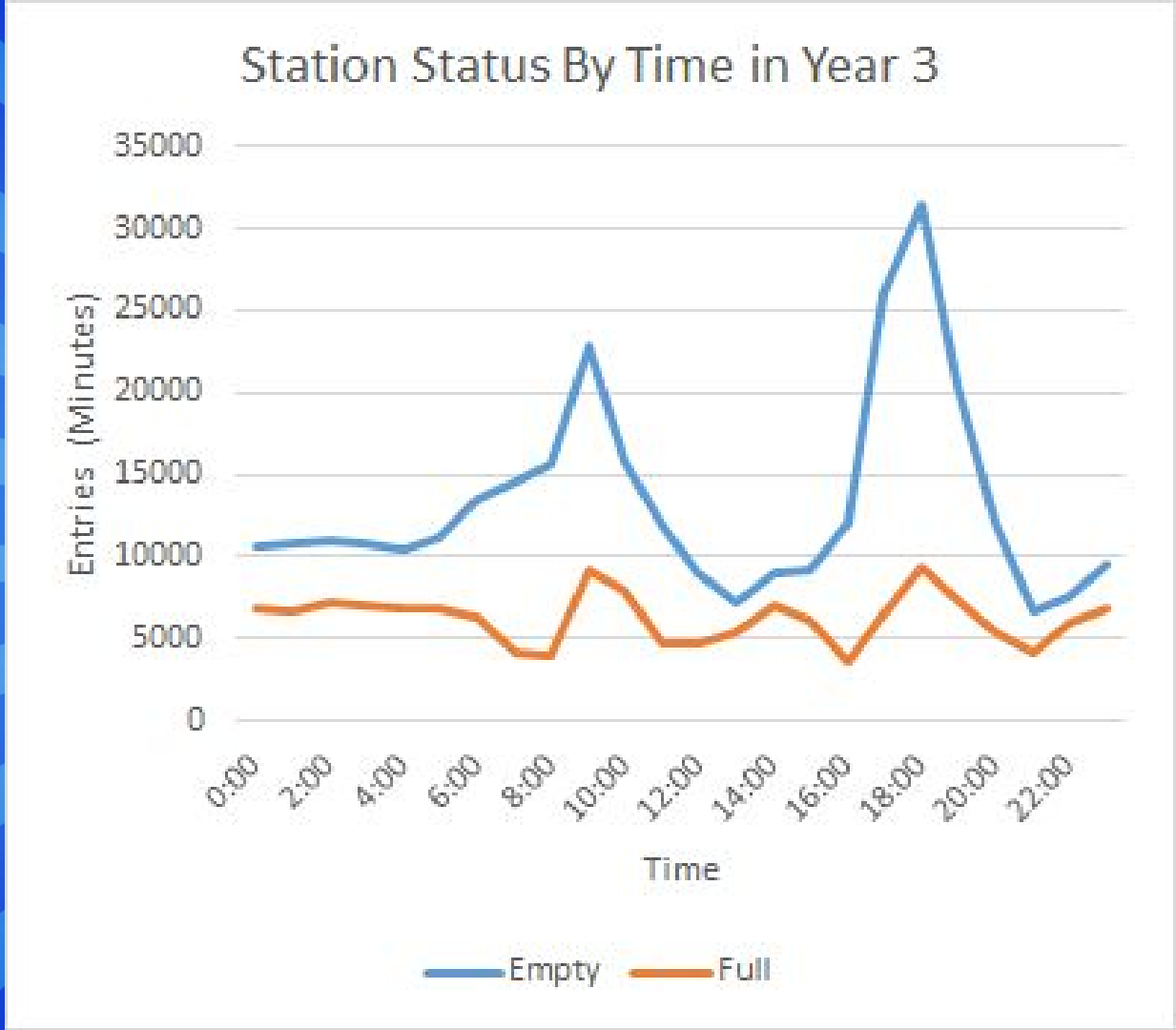
# Outline

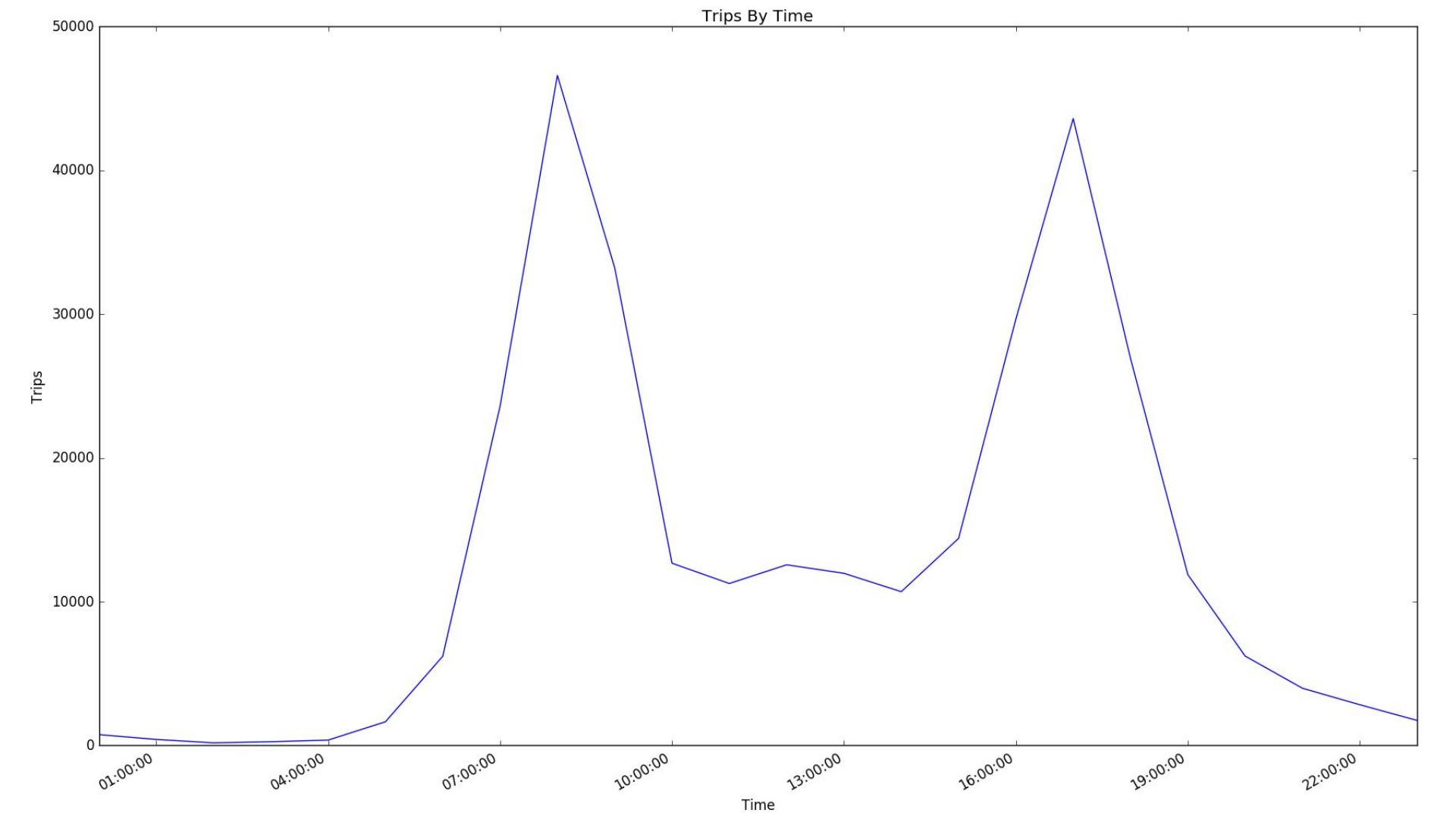
- Project Motivation
- Data Source Acquisition
- Analysis Approach
- Results

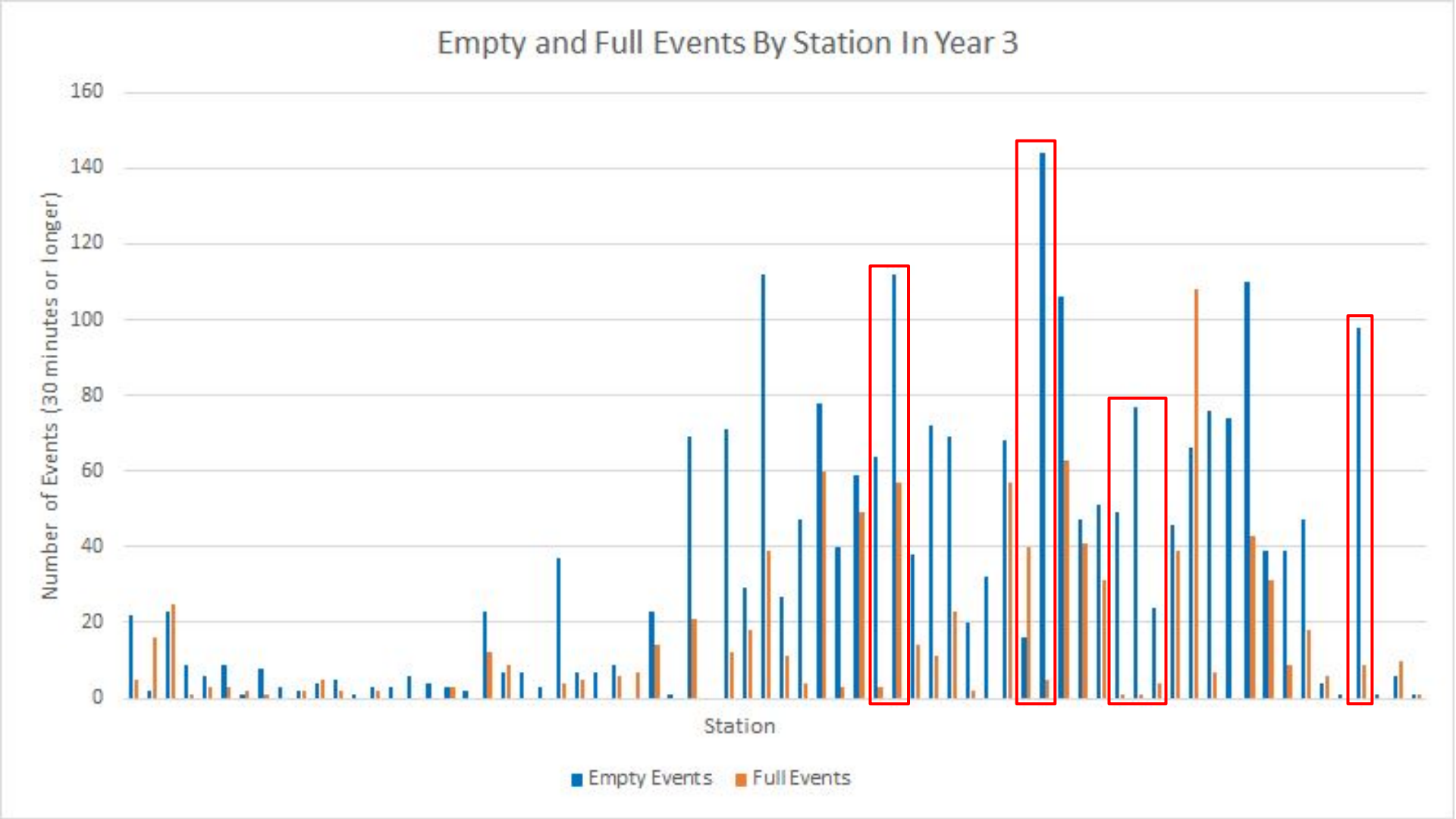
# Motivation

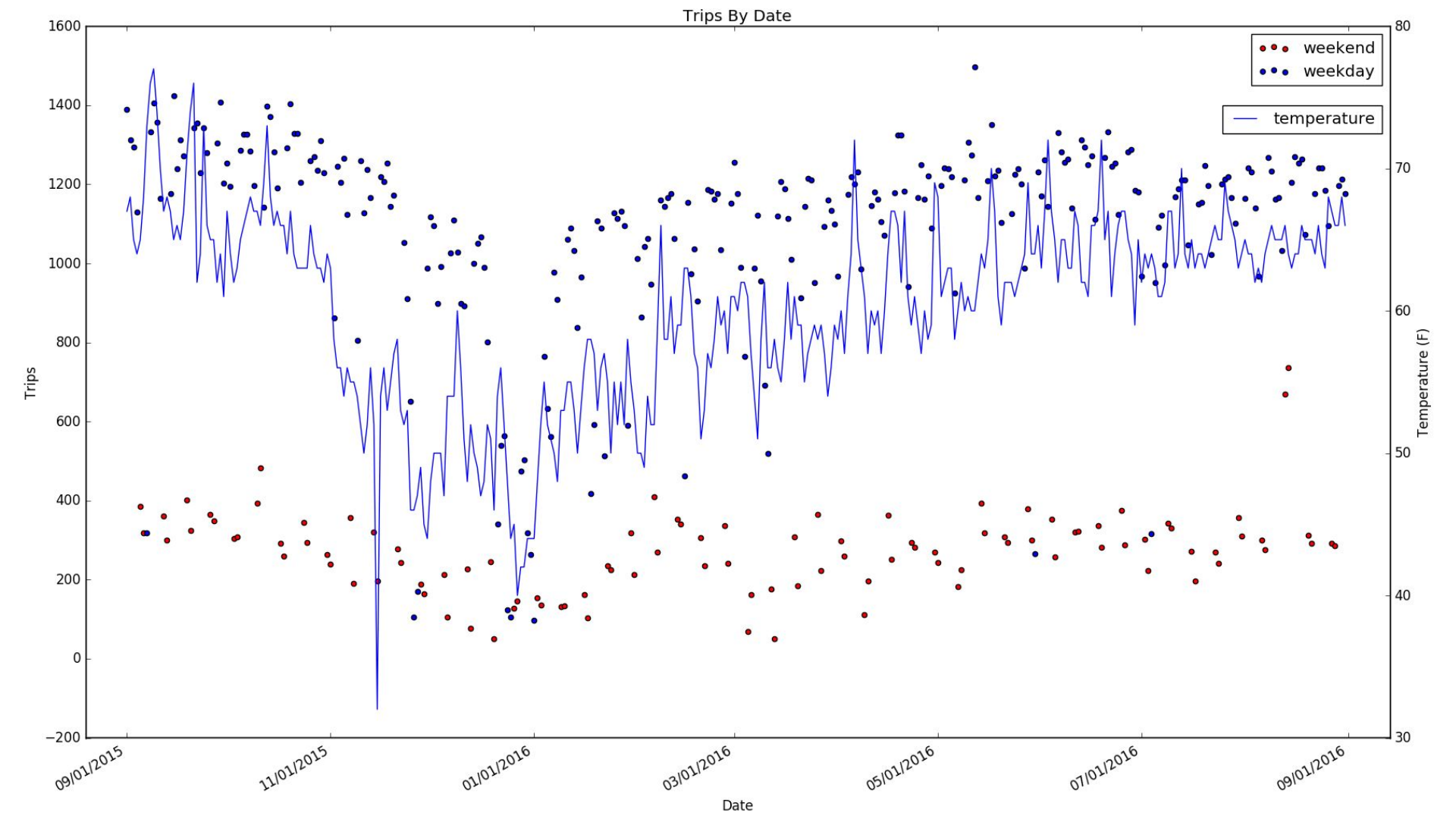




- Real customer complaints
- Ensure customers can dock or rent bikes
- Need to Predict load balancing of bikes between stations


Station Status By Time in Year 3
35000
30000
25000
20000
15000
10000
5000
0
Entries (Minutes)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Empty
Full

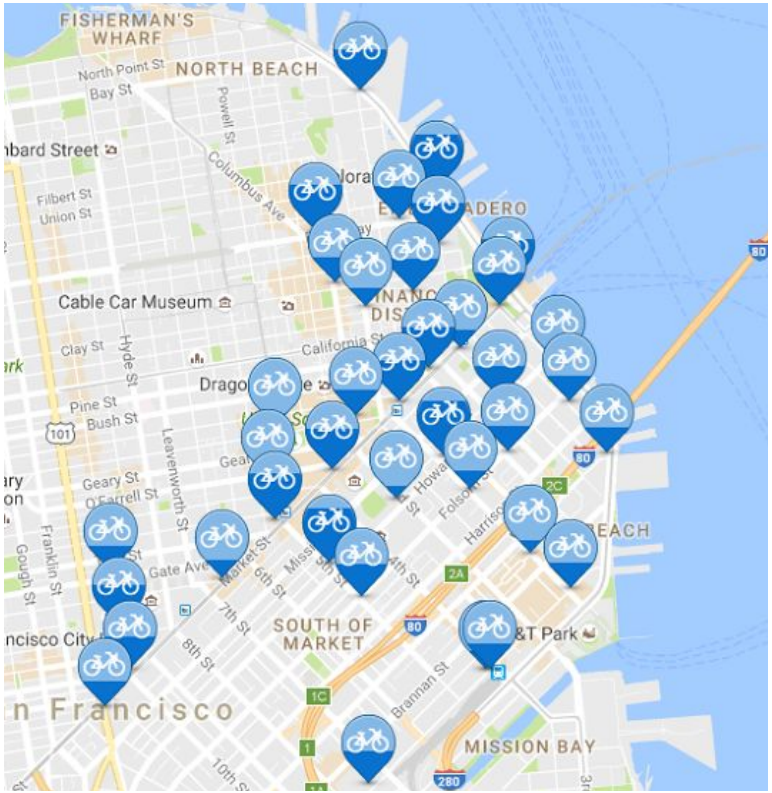

Trips By Time
50000
40000
30000
20000
10000
0
Trips
01:00:00
04:00:00
07:00:00
10:00:00
13:00:00
16:00:00
19:00:00
22:00:00
Time


Empty and Full Events By Station In Year 3
160
140
120
100
80
60
40
20
0
Number of Events (30 minutes or longer)
Station
Empty Events
Full Events


Trips By Date
weekend
weekday
temperature
Trips
Temperature (F)
Date
1600
1400
1200
1000
800
600
400
200
0
−200
80
70
60
50
40
30
09/01/2015
11/01/2015
01/01/2016
03/01/2016
05/01/2016
07/01/2016
09/01/2016

# Data Source

- http://www.bayareabikeshare.com/open-data
- 3 years of data
- Status data
- Station information
- Trip data
- Weather data



# Status Data

| station_id | bikes_available | docks_available | time |
|---|---|---|---|

station_id: station ID number

bikes_available: number of bikes currently in a station

docks_available: number of open docks in a station

time: date & time, PST

>35,000,000 entries; 1 entry per minute per station

# Station Information

| station_id | name | lat | long | dockcount | landmark | installation |
|---|---|---|---|---|---|---|

station_id: station ID number

name: name of station

lat: latitude

long: longitude

dockcount: number of total docks at station (Min = 11, Max = 35)

landmark: city

installation: original date that station was installed.

# Trip Data

| trip_id | start_date | start_ terminal | end_date | end_ terminal | bike_id |
|---|---|---|---|---|---|

trip_id: numeric ID of bike trip
start_date: start date of trip with date and time, in PST
start_terminal: numeric reference for start station
end_date: end date of trip with date and time, in PST
end_terminal: numeric reference for end station
bike_id: ID of bike used

# Weather Data

| date | mean_temperature | events |
| --- | --- | --- |

date: date of measurement
mean_temperature: mean temperature of day in Bay Area
events: nothing, rain, or fog

# Approach: Predicting Days in Advance

| Status | Measurement Criteria |
|---|---|
| Rideability | Time of day and weather<br>- Month, day of week, hour, mean temperature, rain |
| Threshold for count | T = 4, based on station capacities |
| Station status classification (based on threshold) | "**Empty**" indicates # of bikes < T |
| | "**Full**" indicates # of docks available < T |
| | **None** |
| ***Predict*** station status given rideability | **Empty** indicates need for more bikes |
| | **Full** indicates need for removal of bikes |

# Results (All Stations)

| **Algorithm** | Station Status (T=3) | | | Station Status (T=4) | | | Station Status (T=5) | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1 Empty | F1 Full | Accuracy (%) | F1 Empty | F1 Full | Accuracy (%) | F1 Empty | F1 Full | Accuracy (%) |
| Gaussian Naive Bayes | N/A | N/A | N/A | 0.023 | 0.008 | 79.8 | N/A | N/A | N/A |
| Decision Tree | 0.133 | 0.110 | 86.5 | 0.157 | 0.212 | 75.3 | 0.301 | 0.228 | 61.2 |
| Random Forest | 0.090 | 0.109 | 88.0 | 0.126 | 0.115 | 78.1 | 0.226 | 0.170 | 62.8 |
| Random Classifier | 0.175 | 0.096 | 33.3 | 0.175 | 0.096 | 33.3 | 0.175 | 0.096 | 33.3 |

# Results (Individual Stations)

| **Algorithm** | Station 45 (T=4) | | | | | Station 11 (T=4) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P Empty | R Empty | P Full | R Full | Accuracy (%) | P Empty | R Empty | P Full | R Full | Accuracy (%) |
| Decision Tree | 0.457 | 0.338 | 0.203 | 0.156 | 49.2 | 0.003 | 0.014 | 0.003 | 0.007 | 82.9 |
| Random Forest | 0.506 | 0.326 | 0.201 | 0.081 | 52.7 | 0.006 | 0.009 | 0.14 | 0.007 | 91.2 |
| Random Classifier | 0.119 | 0.33 | 0.056 | 0.33 | 33.3 | 0.119 | 0.33 | 0.056 | 0.33 | 33.3 |

P = Precision, R = Recall


Live
Demo

# Future Work

- Once docks are rebalanced, can’t use data for future training
- Model that will work with live data for load balancing
- Take large events into account for rebalancing
- Simplifying attributes
- Short term prediction