MIT Big Data Challenge

Transportation in the City of Boston

Team SACLARKE:
Sean Aidan Clarke
Christoph Hafemeister

Thursday, February 27, 2014

Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Predictors used:
- pickup counts in spatial bins

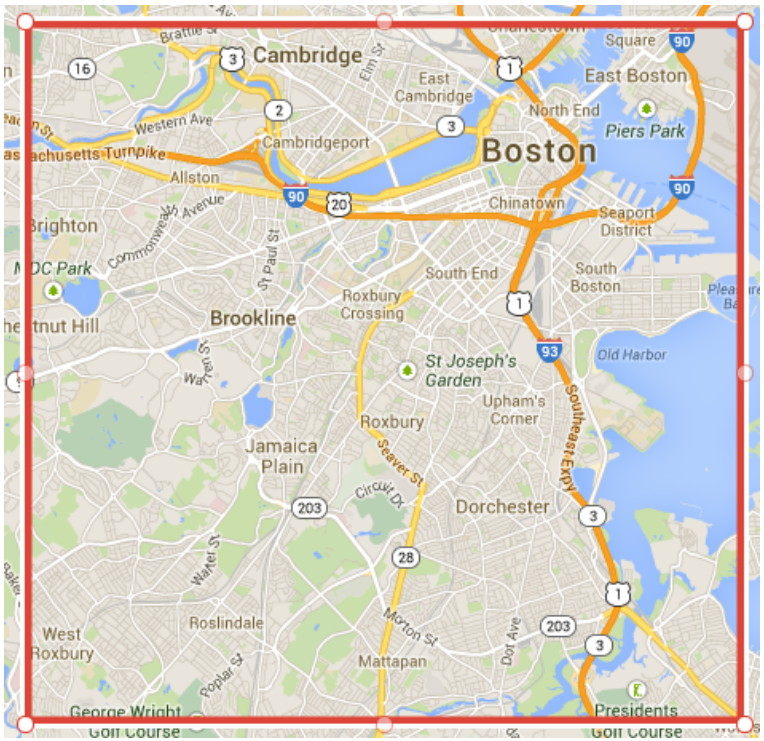
Query: test location and two hour time window

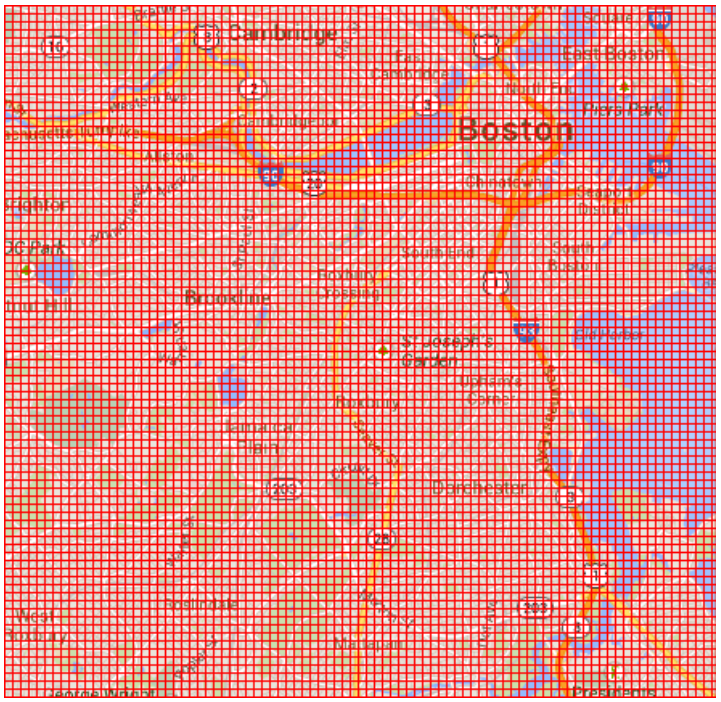
Response variable: pickups in location binned into two hour windows

Predictors used:

- pickup counts in spatial bins

1023





Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Predictors used:

- pickup counts in spatial bins 1023
- AFC station/route trip counts divided by total 282
- ODrail station trip counts divided by total 80

Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Predictors used:

• pickup counts in spatial bins	1023
• AFC station/route trip counts divided by total	282
• ODrail station trip counts divided by total	80
• number of tweets in location / number of tweets outside of location	1
• number of dropoffs in location	1
• number of dropoffs in location / number of dropoffs outside of location	1

Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Predictors used:

- pickup
- AFC
- ODrail

} keep only the 32 variables most correlated with response

- number of tweets in location / number of tweets outside of location
- number of dropoffs in location
- number of dropoffs in location / number of dropoffs outside of location

1023
282
80
1
1
1

Query: test location and two hour time window

Response variable: pickups in location binned into two hour windows

Predictors used:

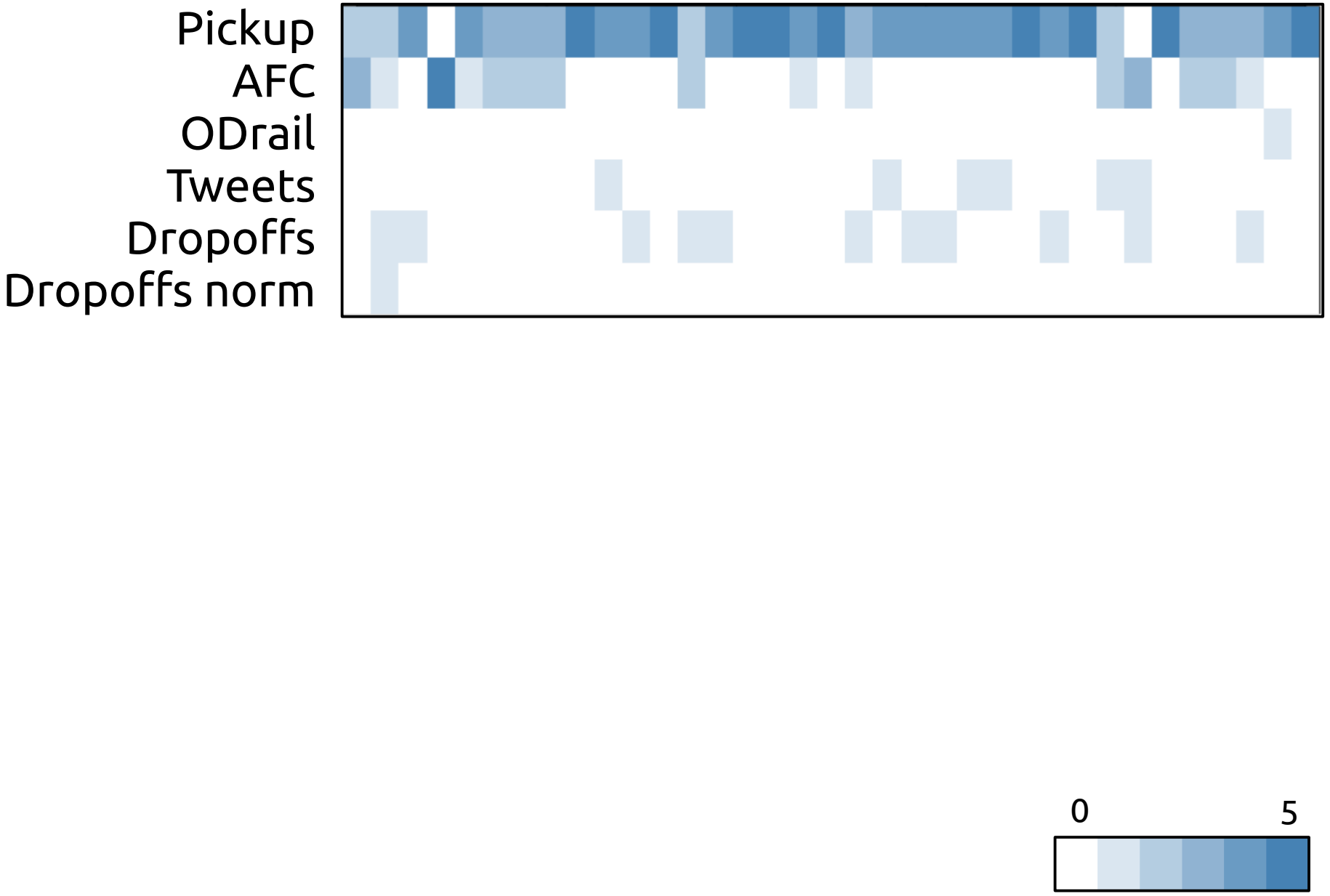
- pickup — 1023
- AFC — 282
- ODrail — 80

keep only the 32 variables most correlated with response

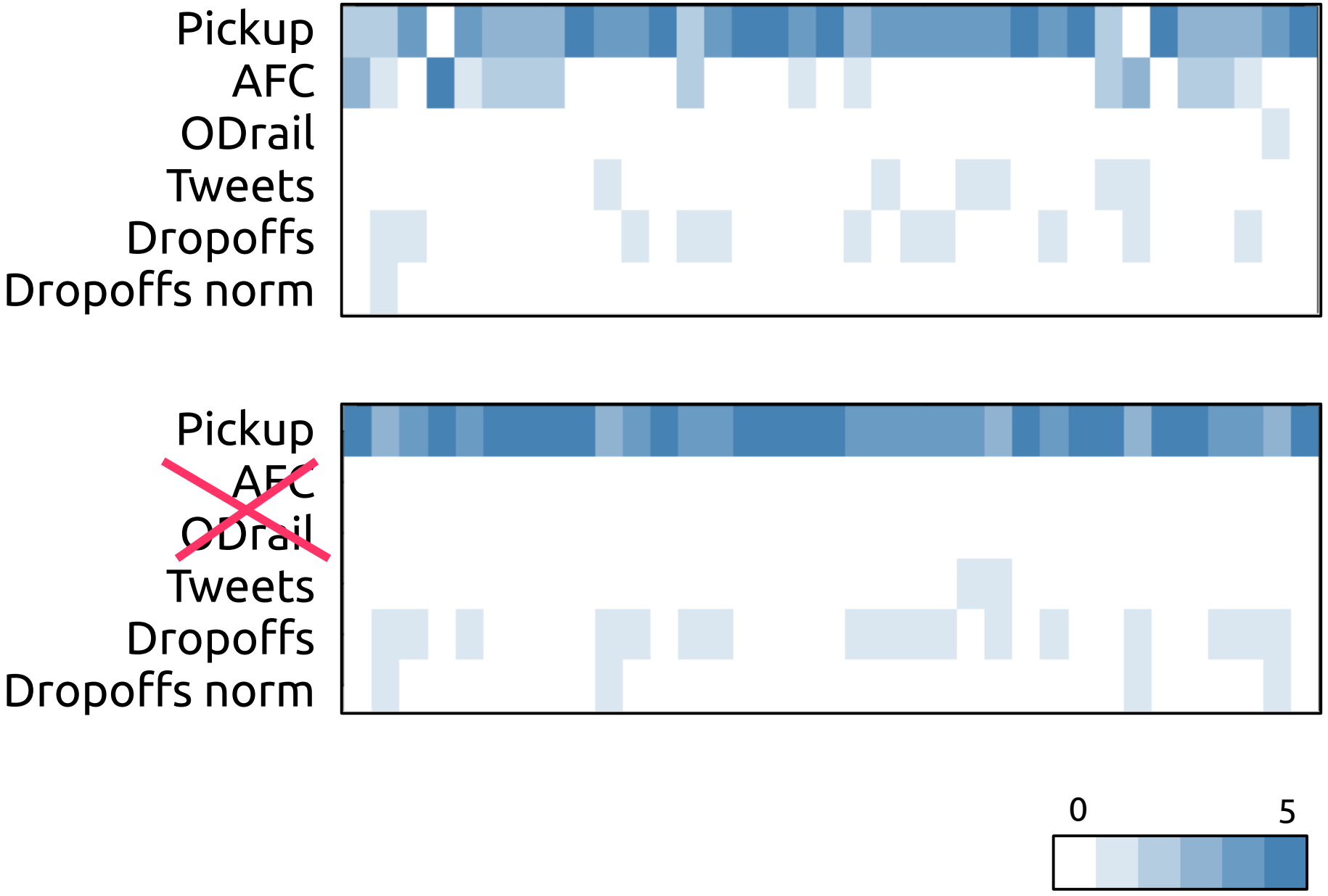
- number of tweets in location / number of tweets outside of location — 1
- number of dropoffs in location — 1
- number of dropoffs in location / number of dropoffs outside of location — 1

Use random forest regressor

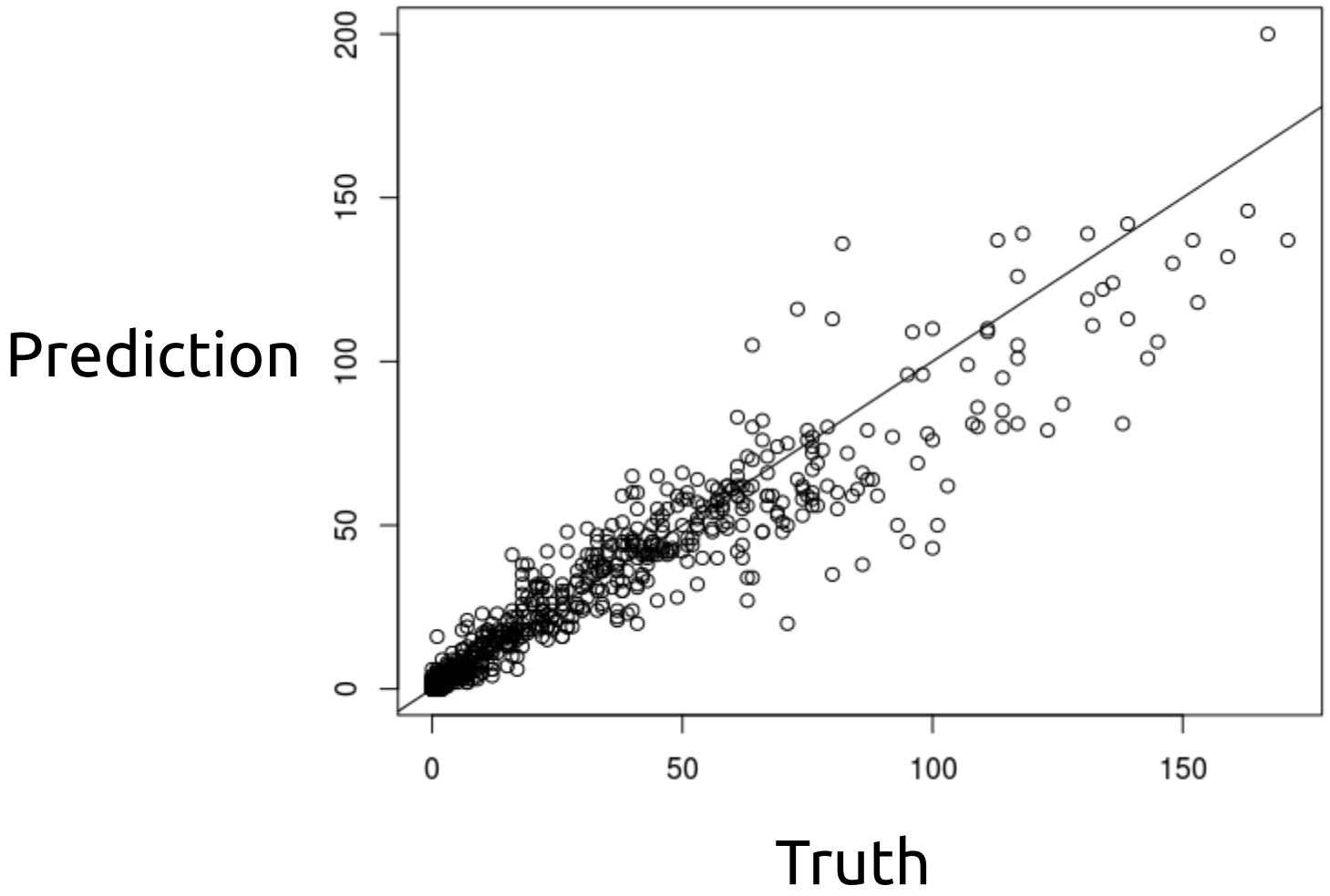
Top five predictors for each location



Top five predictors for each location

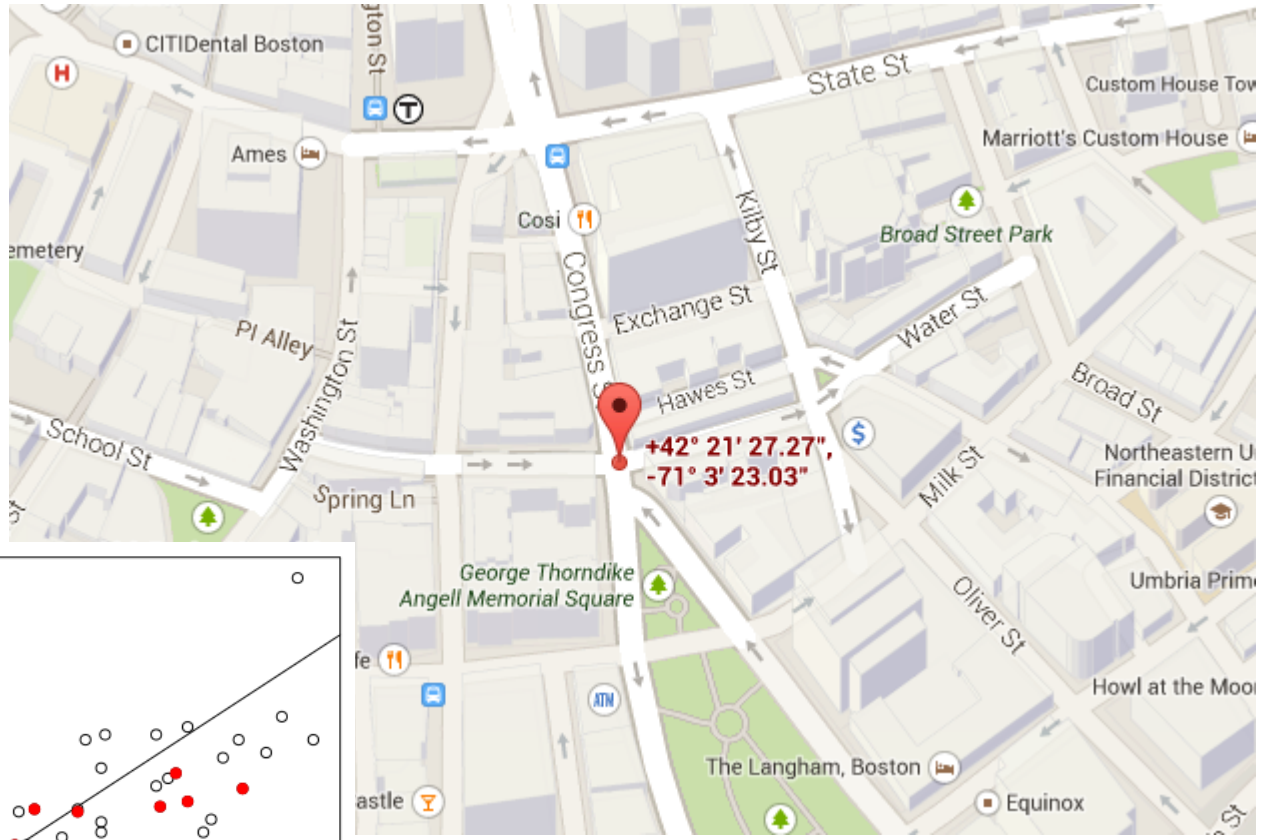


Predictions

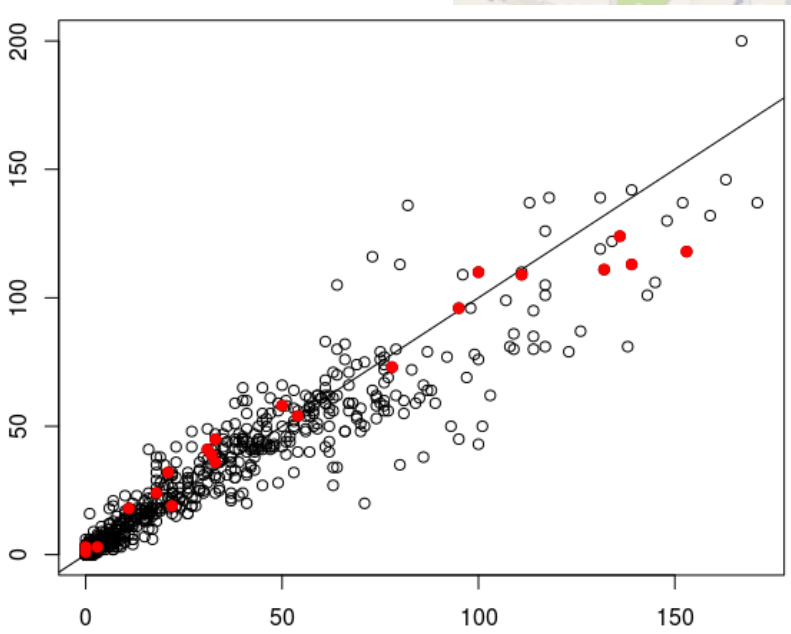


Predictions

Good location



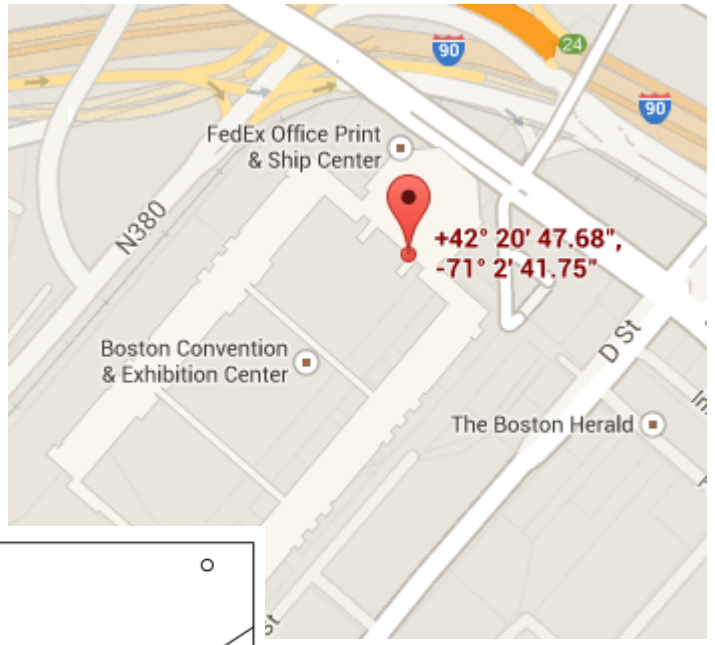
Prediction



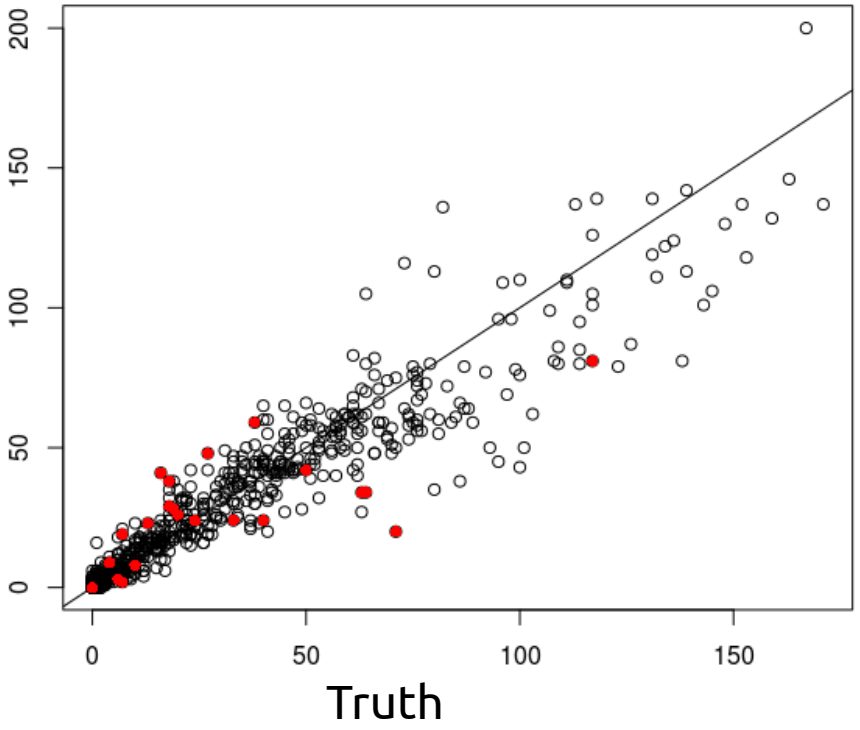
Truth

Predictions

Bad location

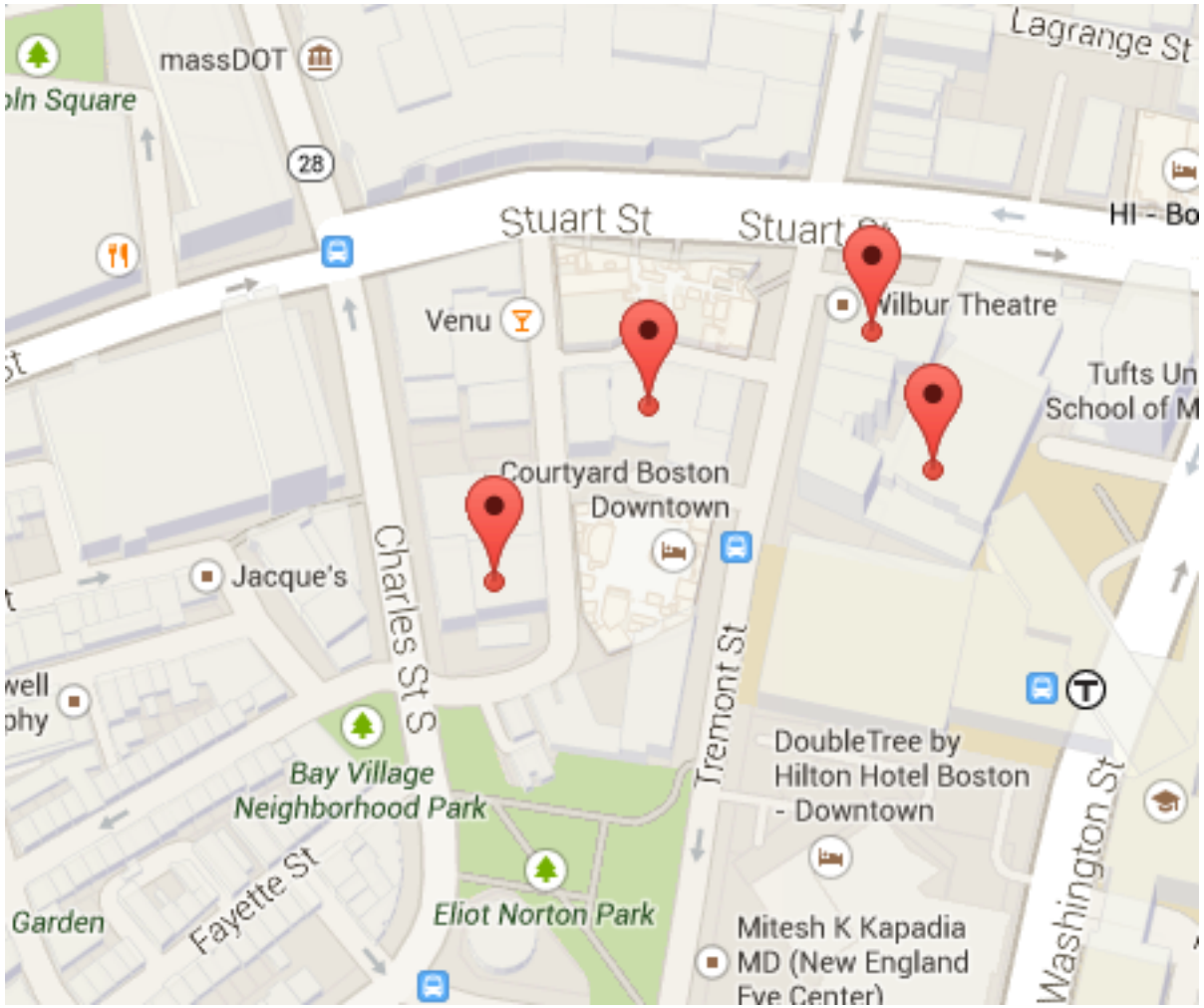


Prediction



Predictions

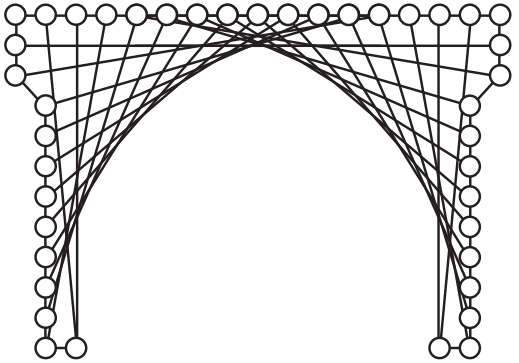
More bad locations



Acknowledgments

My mentor: Rich Bonneau

NEW YORK UNIVERSITY



CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY