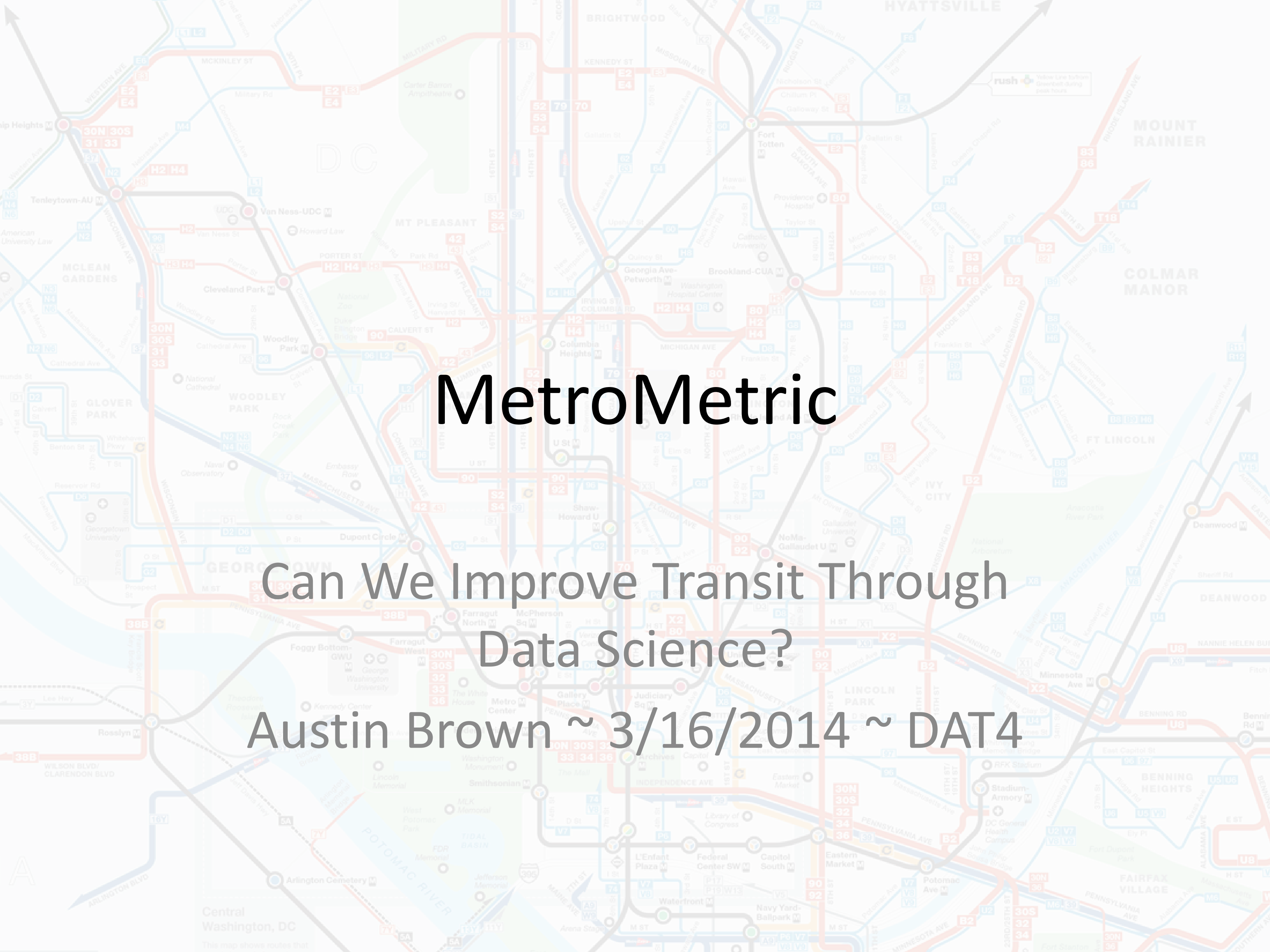
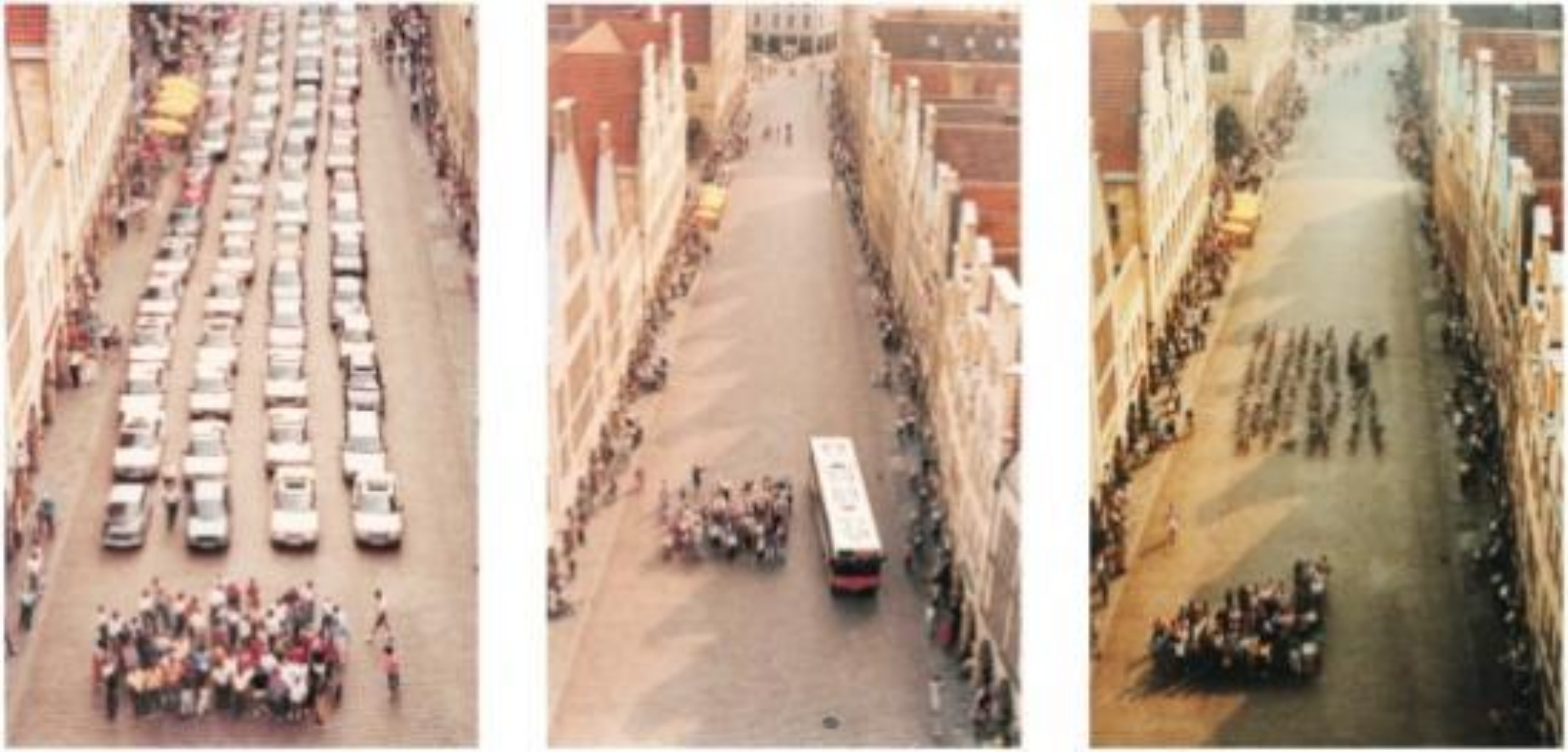


# MetroMetric

## Can We Improve Transit Through Data Science?

Austin Brown ~ 3/16/2014 ~ DAT4

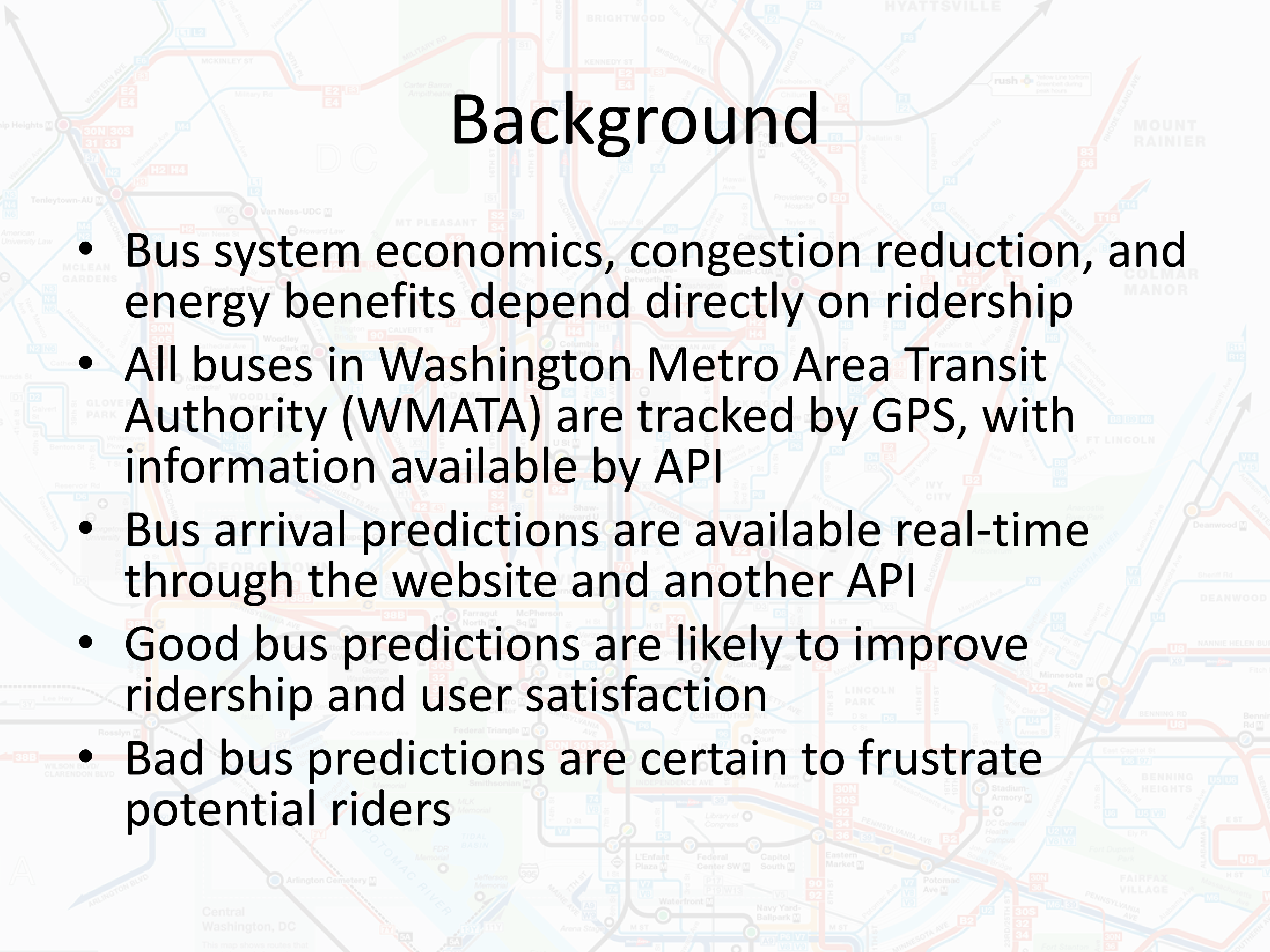
# Buses Are Great



(But a pain to use)

http://www.howwedrive.com/2010/06/28/german-efficiency/

Münster



# Background

- Bus system economics, congestion reduction, and energy benefits depend directly on ridership
- All buses in Washington Metro Area Transit Authority (WMATA) are tracked by GPS, with information available by API
- Bus arrival predictions are available real-time through the website and another API
- Good bus predictions are likely to improve ridership and user satisfaction
- Bad bus predictions are certain to frustrate potential riders

# Open Transit Data

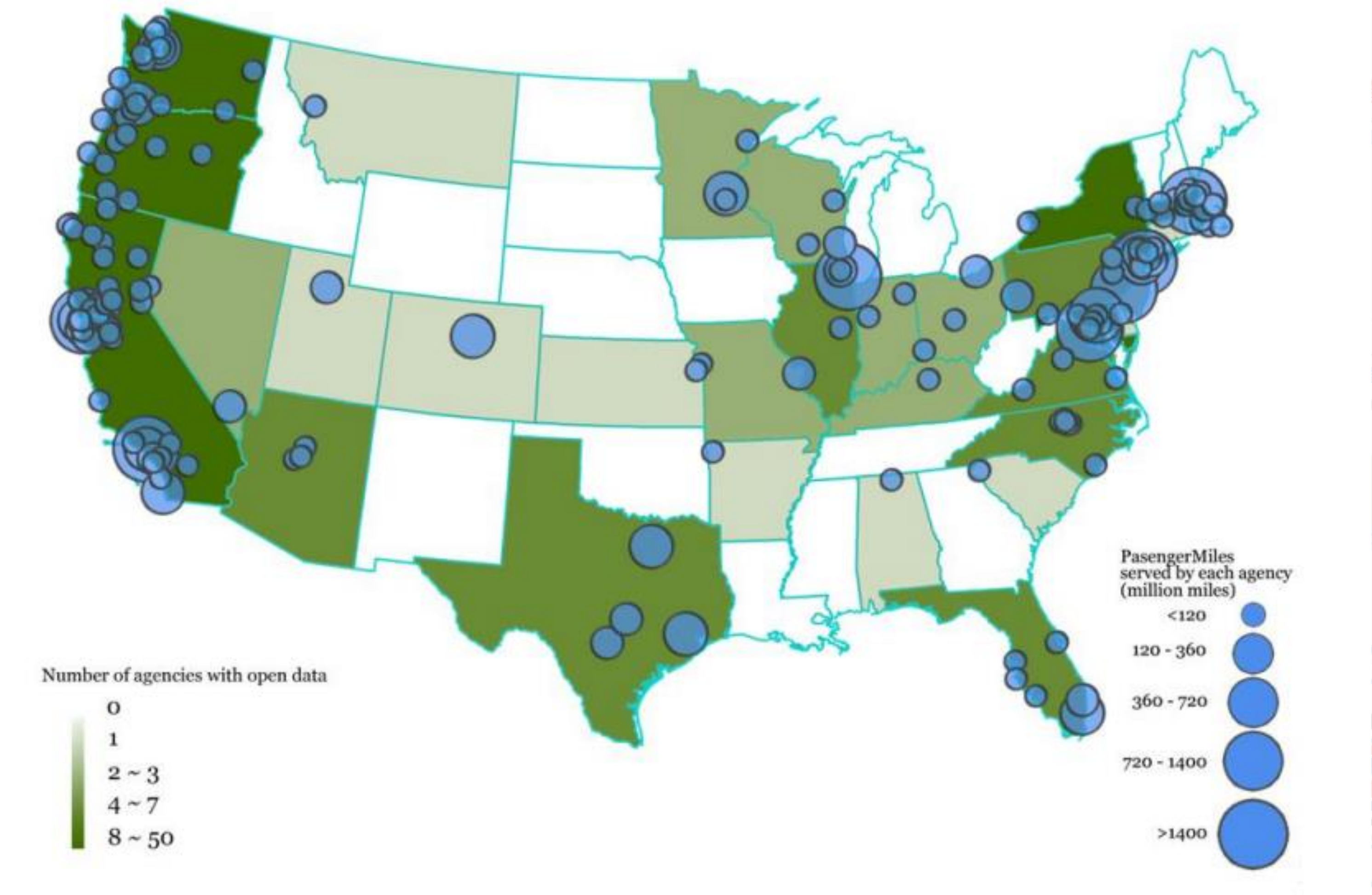
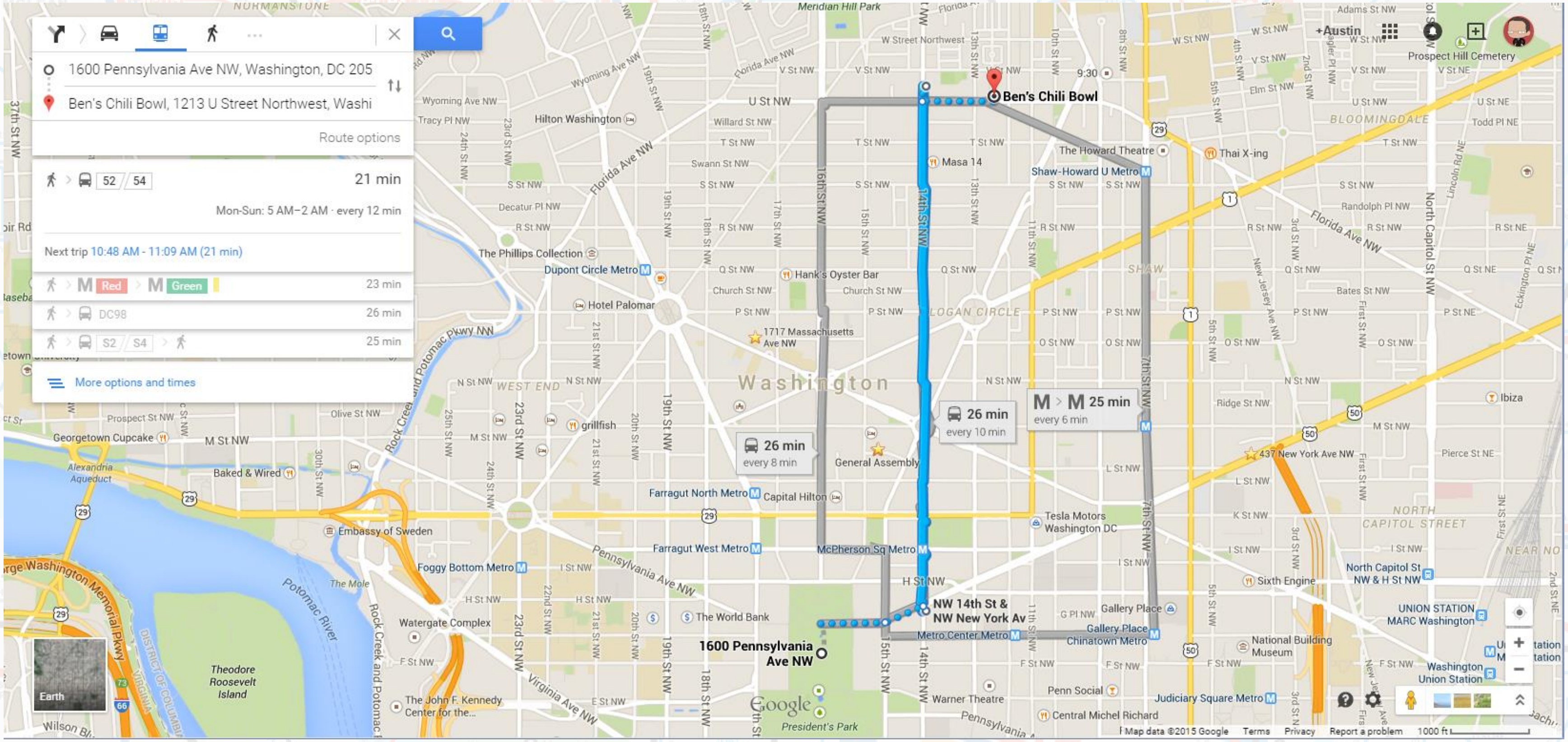

PasengerMiles served by each agency (million miles)
<120
120 - 360
360 - 720
720 - 1400
>1400
Number of agencies with open data
0
1
2 ~ 3
4 ~ 7
8 ~ 50

**Figure 1**- Number and size of agencies with open data by state.

Illustration from http://www.prism.gatech.edu/~lreed3/opendata.pdf

# Example Application



1600 Pennsylvania Ave NW, Washington, DC 205

Ben's Chili Bowl, 1213 U Street Northwest, Washi

Route options

52 / 54 — 21 min

Mon-Sun: 5 AM–2 AM · every 12 min

Next trip 10:48 AM - 11:09 AM (21 min)

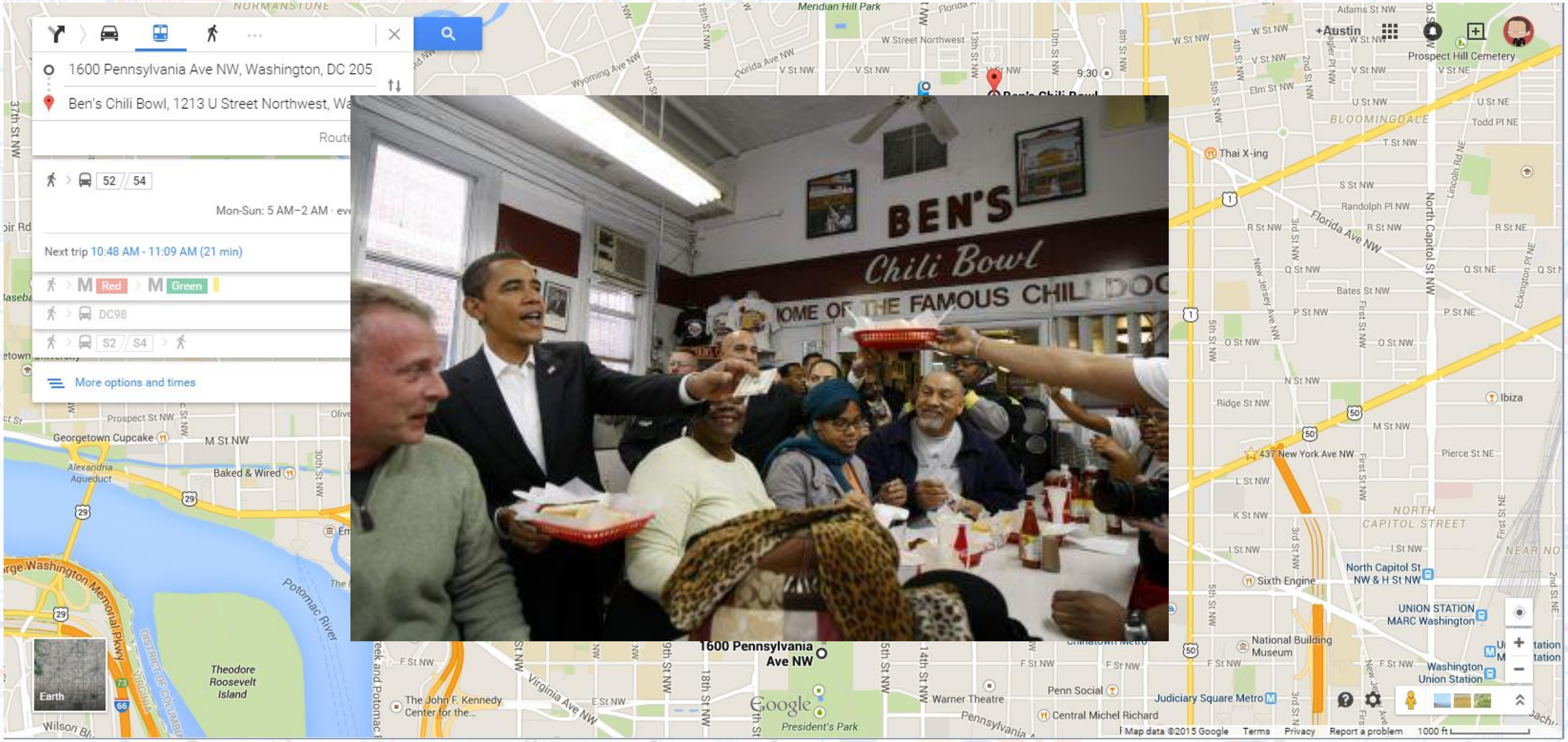
M Red > M Green — 23 min

DC98 — 26 min

S2 / S4 — 25 min

More options and times

# Example Application



1600 Pennsylvania Ave NW, Washington, DC 205

Ben's Chili Bowl, 1213 U Street Northwest, Wa

Route

52 / 54

Mon-Sun: 5 AM–2 AM · eve

Next trip 10:48 AM - 11:09 AM (21 min)

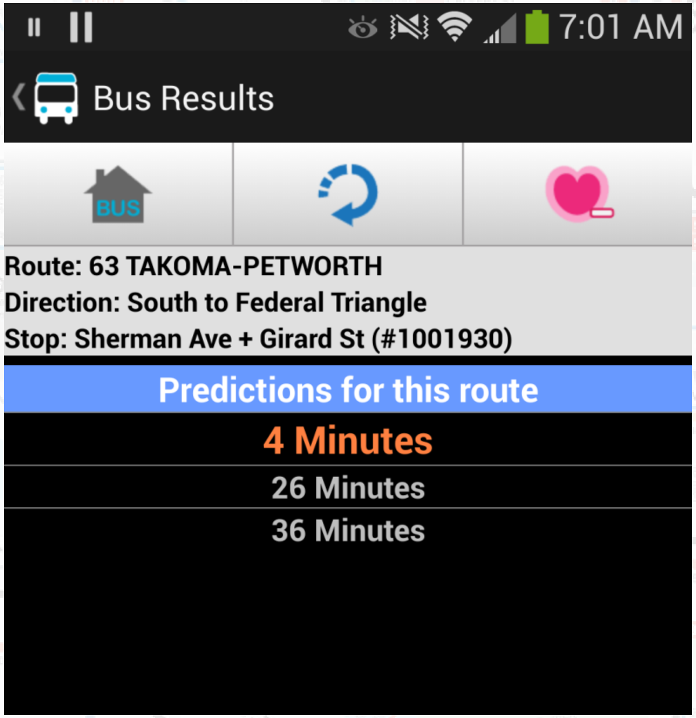
M Red > M Green

DC98

S2 / S4

More options and times

# Personal Example


7:01 AM

Bus Results

Route: 63 TAKOMA-PETWORTH
Direction: South to Federal Triangle
Stop: Sherman Ave + Girard St (#1001930)

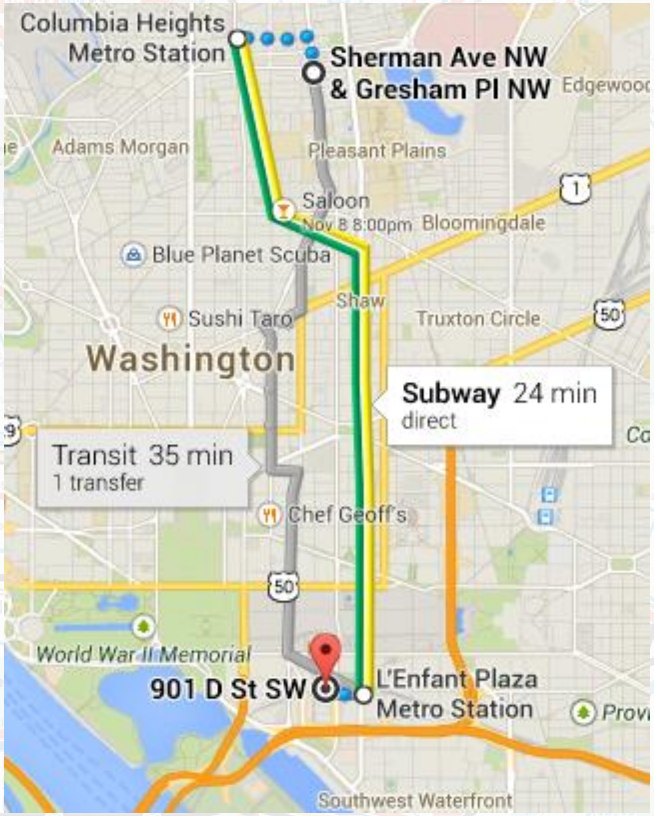
Predictions for this route

4 Minutes

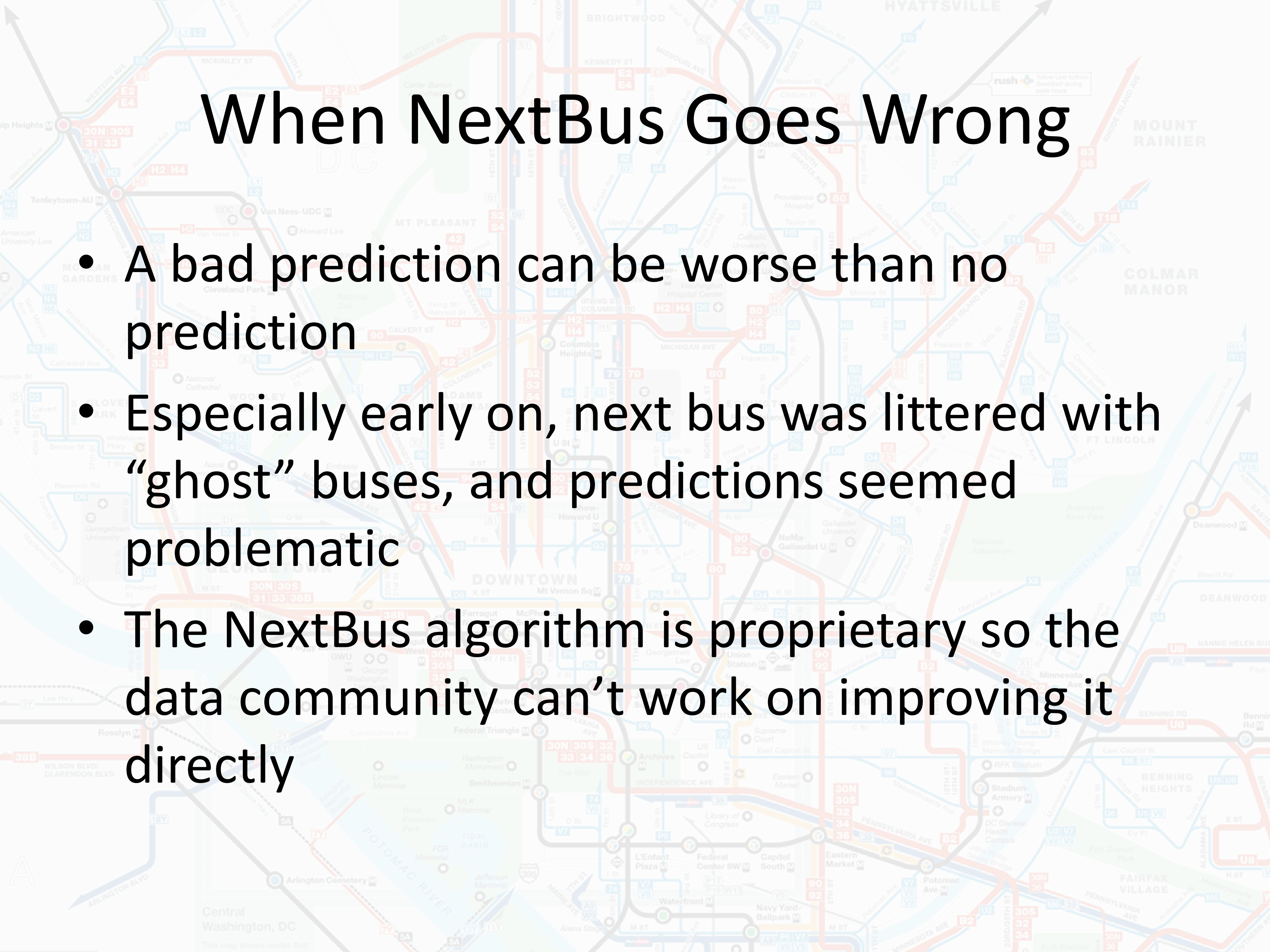
26 Minutes

36 Minutes

DC Metro Transit ap



Google Maps

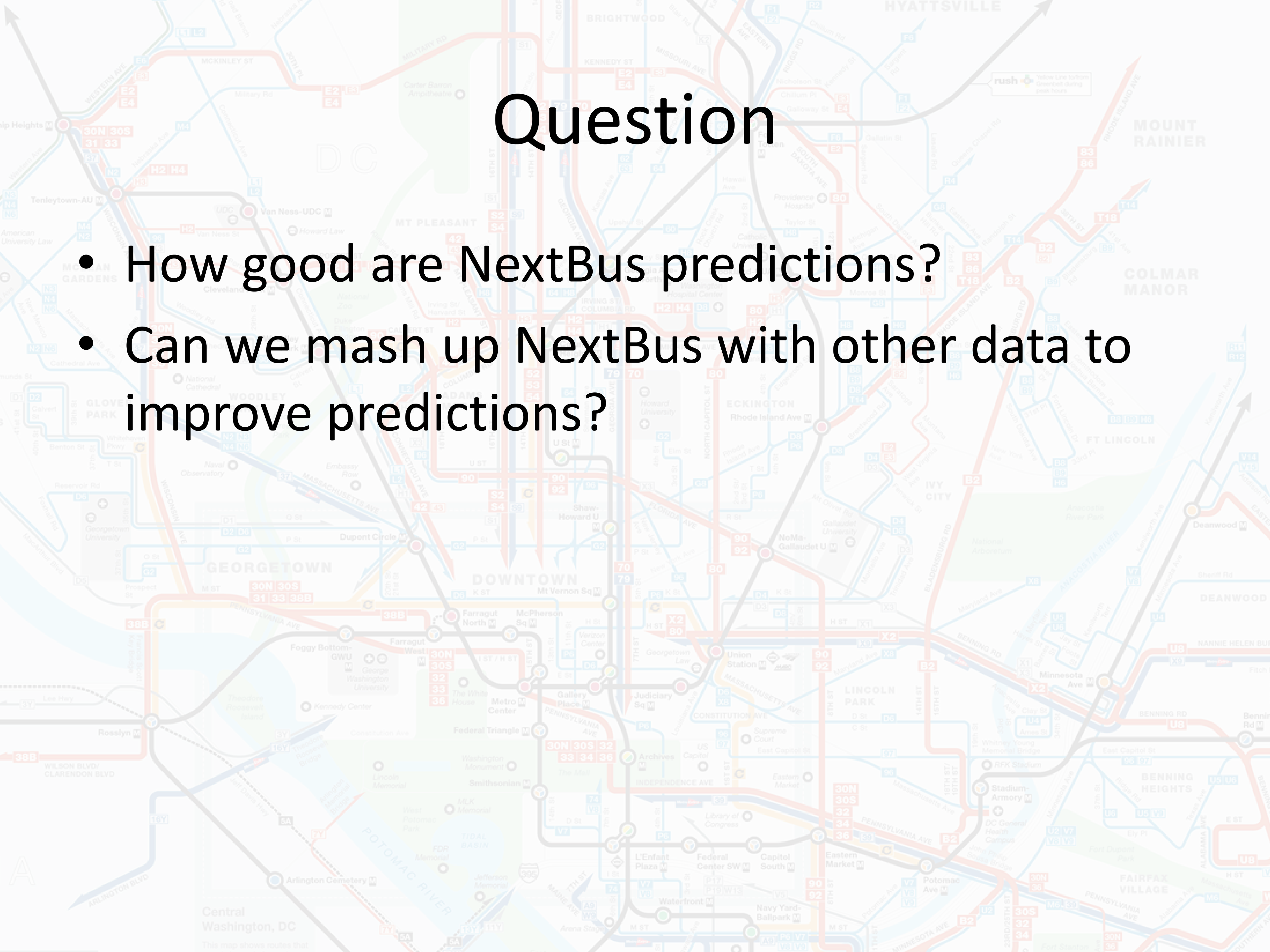


# When NextBus Goes Wrong

- A bad prediction can be worse than no prediction
- Especially early on, next bus was littered with “ghost” buses, and predictions seemed problematic
- The NextBus algorithm is proprietary so the data community can’t work on improving it directly

# Question

- How good are NextBus predictions?
- Can we mash up NextBus with other data to improve predictions?

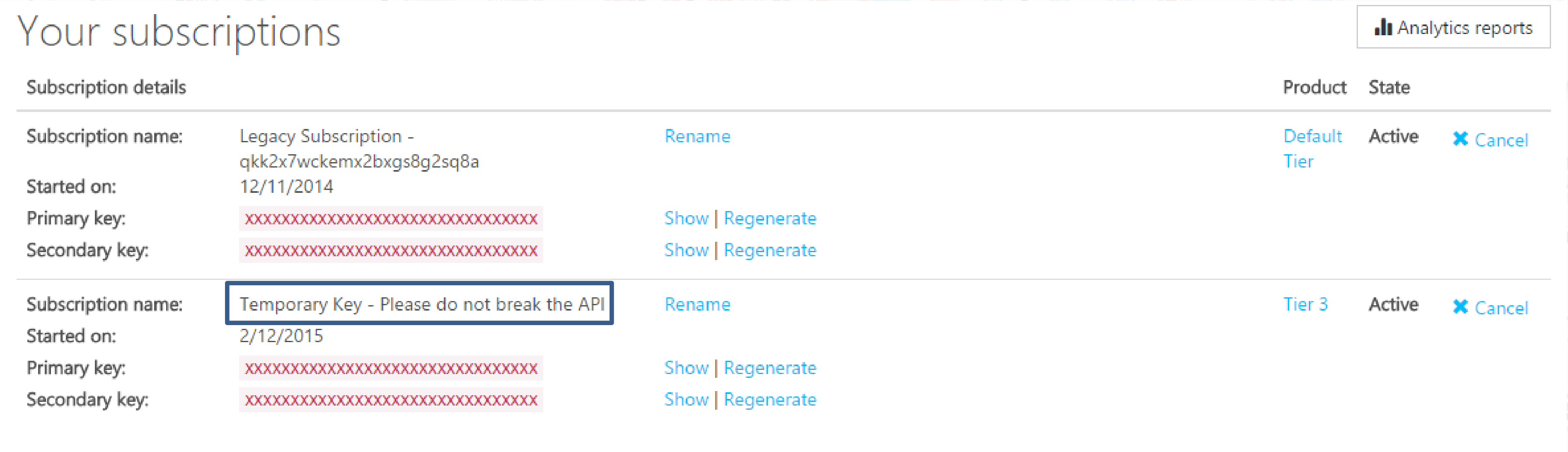


# Data Structure Concept

| Field(s) | Description | Source |
|---|---|---|
| Route, Direction | e.g. 70 | Route API |
| StopID, Lat, Lon | Stop Itentified and location | Route API |
| BusID, Lat, Lon | Bus Identified and location | BusLocation API |
| Deviation | How far behind the bus says it is | BusLocation API |
| TripID | Unique trip identifier | BusLocation / Nextbus API |
| Time / Date | Time the prediction was made | Datetime package |
| Weather / Temp | Reported weather | Openweathermap API |
| Predicted Arrival | In minutes, when the bus should arrive | NextBus API |
| Actual Arrival | In minutes, how long from the prediction the bus arrived | Calculated |

The basic data unit is one prediction. Data collection is done from four APIs in real time and stored in a data frame / csv

# WMATA API Key

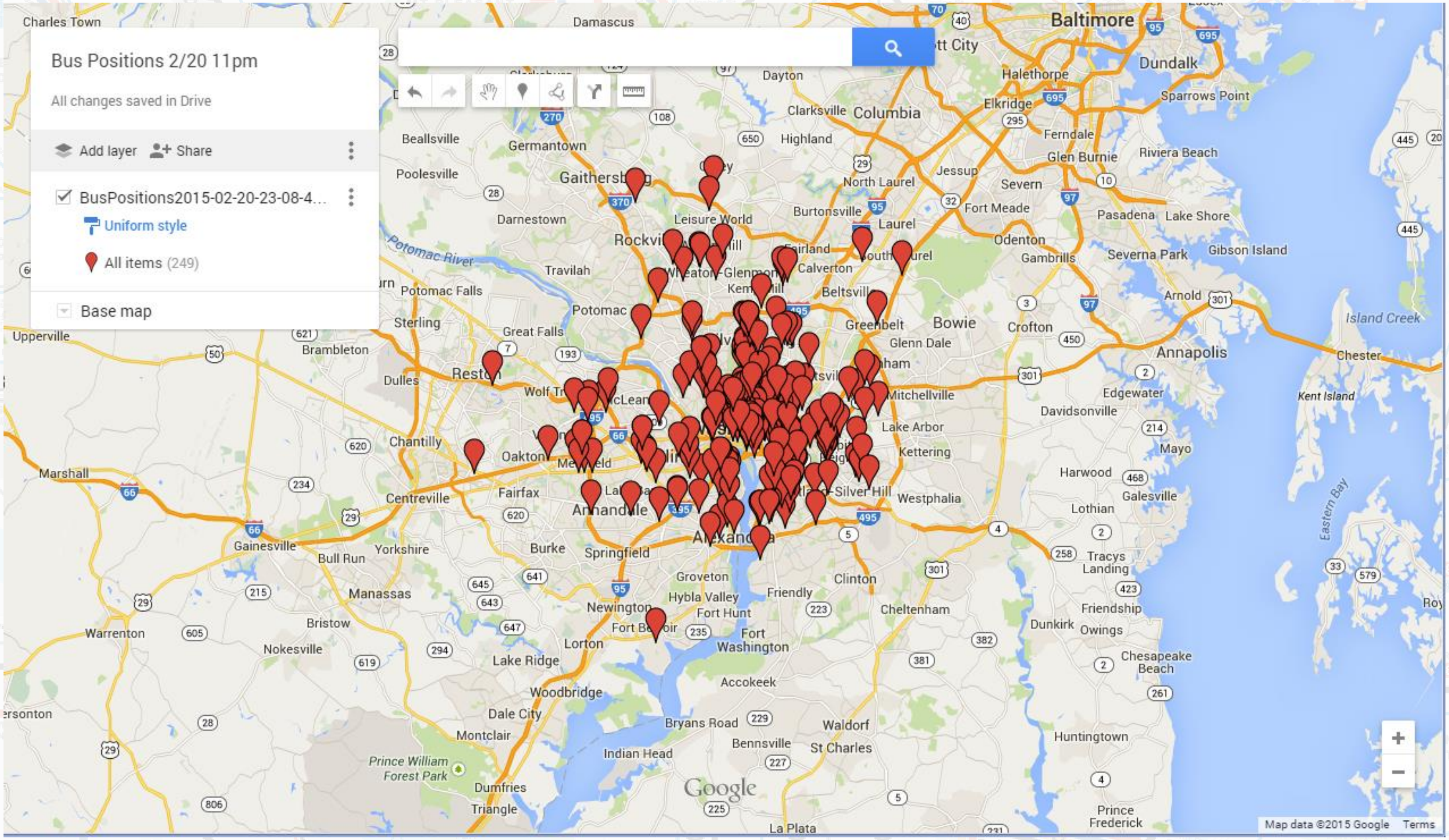


## Your subscriptions

| Subscription details | | | Product | State | |
|---|---|---|---|---|---|
| Subscription name: | Legacy Subscription - qkk2x7wckemx2bxgs8g2sq8a | Rename | Default Tier | Active | ✖ Cancel |
| Started on: | 12/11/2014 | | | | |
| Primary key: | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx | Show \| Regenerate | | | |
| Secondary key: | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx | Show \| Regenerate | | | |
| Subscription name: | Temporary Key - Please do not break the API | Rename | Tier 3 | Active | ✖ Cancel |
| Started on: | 2/12/2015 | | | | |
| Primary key: | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx | Show \| Regenerate | | | |
| Secondary key: | xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx | Show \| Regenerate | | | |

I need a lot of calls… WMATA to the rescue

# WMATA APIs – Bus Positions



Bus Positions 2/20 11pm

All changes saved in Drive

Add layer Share

☑ BusPositions2015-02-20-23-08-4...

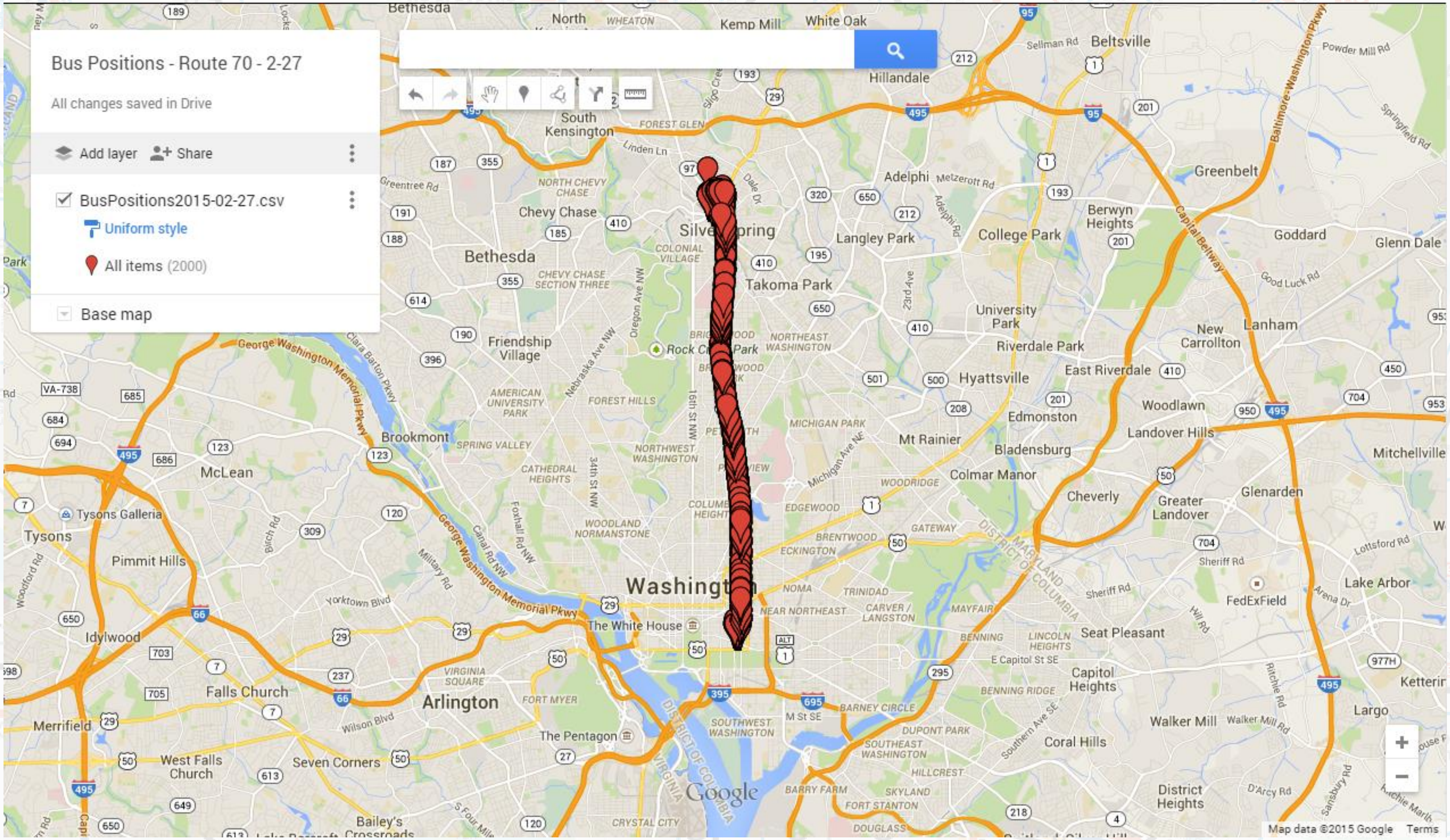
Uniform style

All items (249)

Base map

One data call for all bus positions. 249 buses in system (at 11 pm)

# WMATA APIs – Bus Positions



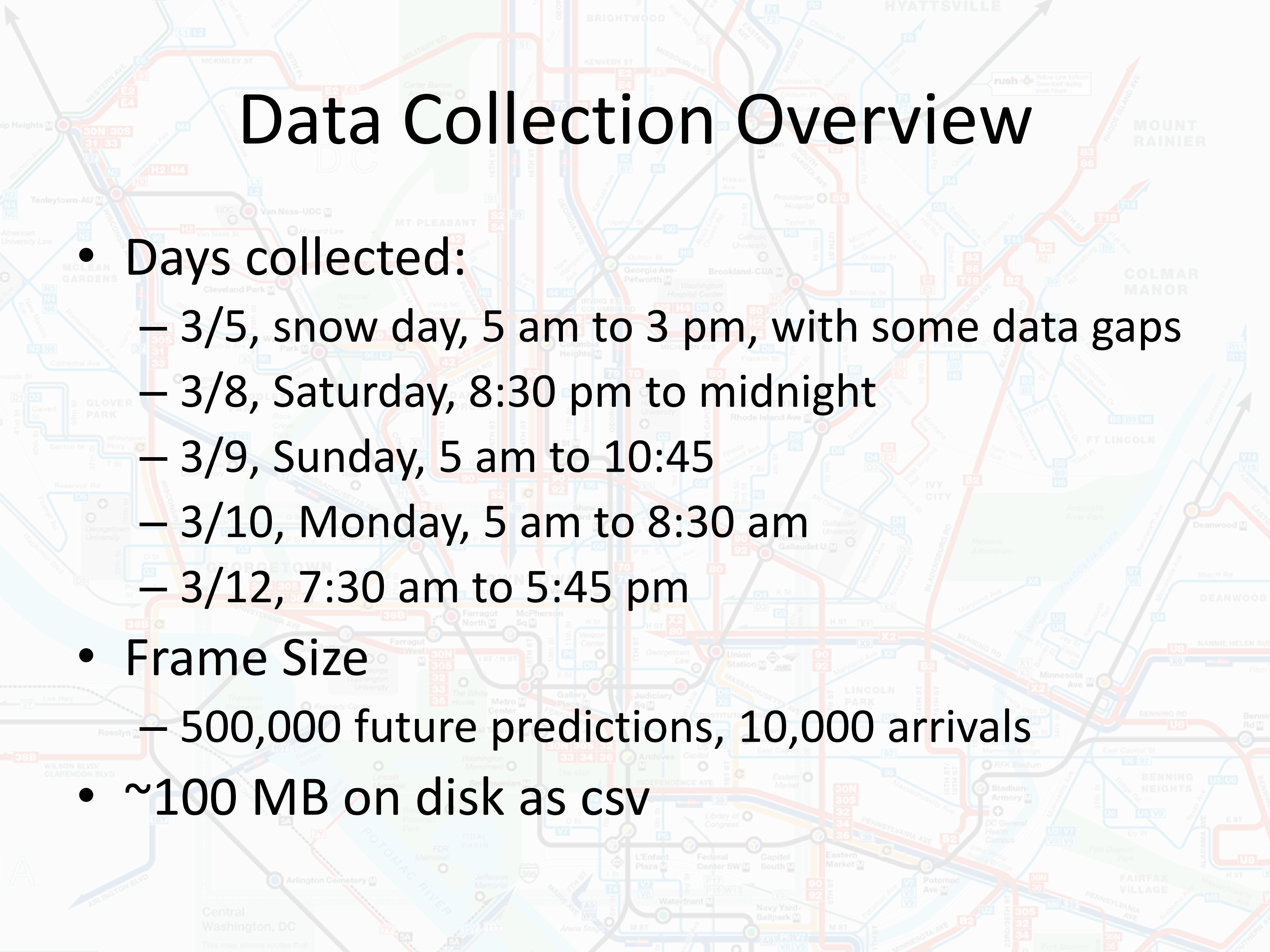
Bus Positions - Route 70 - 2-27

All changes saved in Drive

☑ BusPositions2015-02-27.csv

All items (2000)

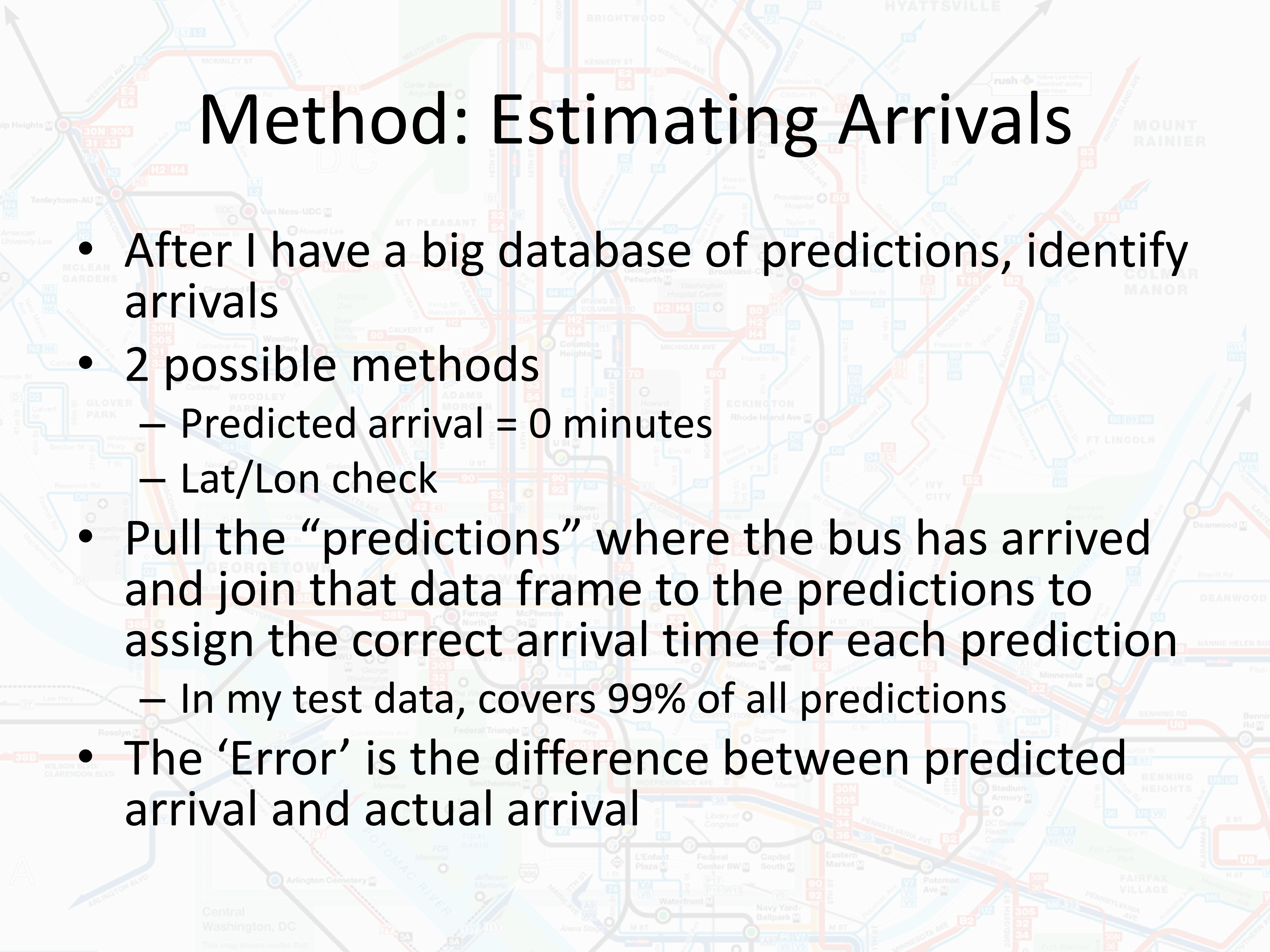
Many data calls for route 70 only



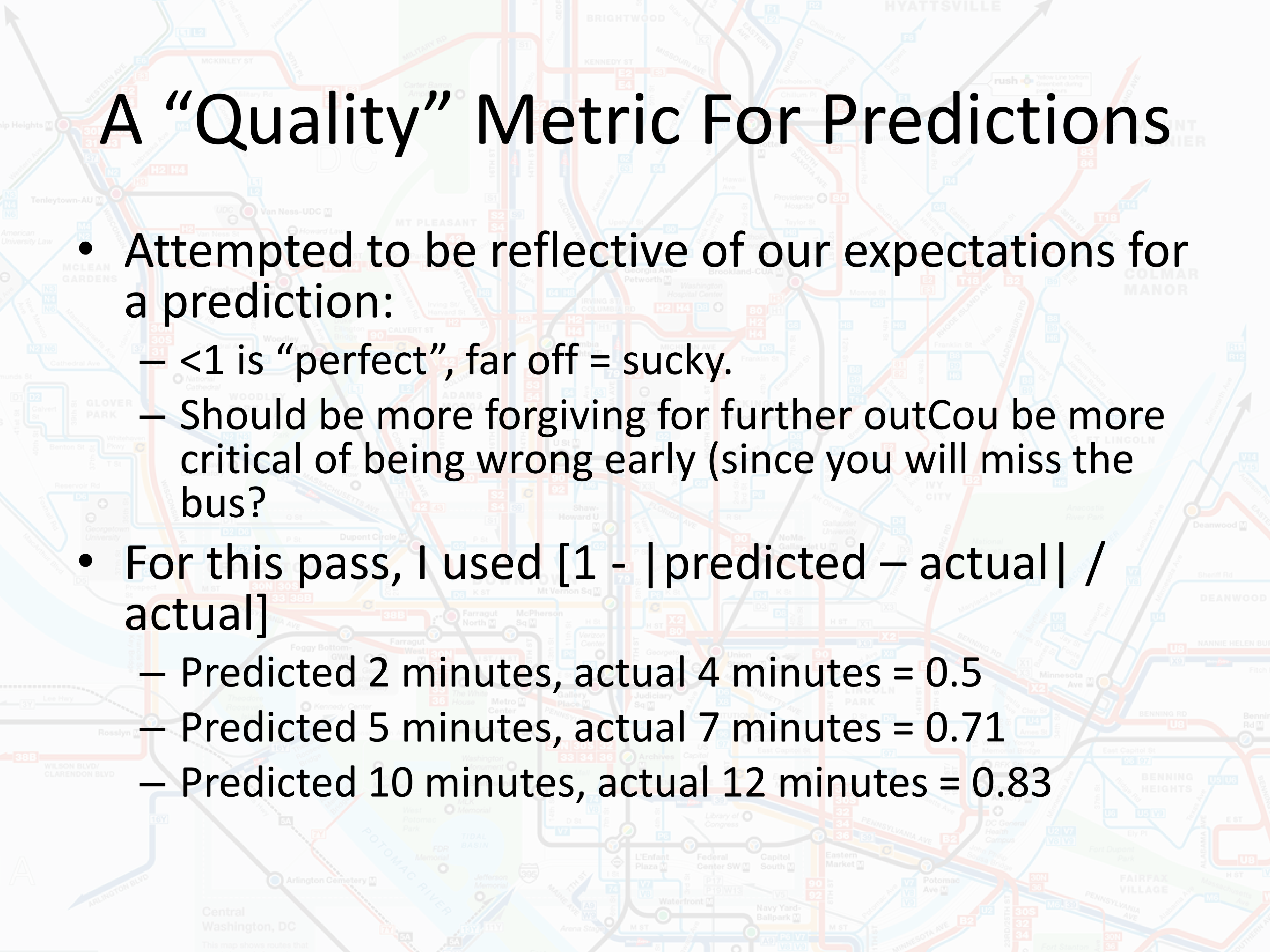
# Data Collection Overview

- Days collected:
  - – 3/5, snow day, 5 am to 3 pm, with some data gaps
  - – 3/8, Saturday, 8:30 pm to midnight
  - – 3/9, Sunday, 5 am to 10:45
  - – 3/10, Monday, 5 am to 8:30 am
  - – 3/12, 7:30 am to 5:45 pm
- Frame Size
  - – 500,000 future predictions, 10,000 arrivals
- ~100 MB on disk as csv

# Method: Estimating Arrivals



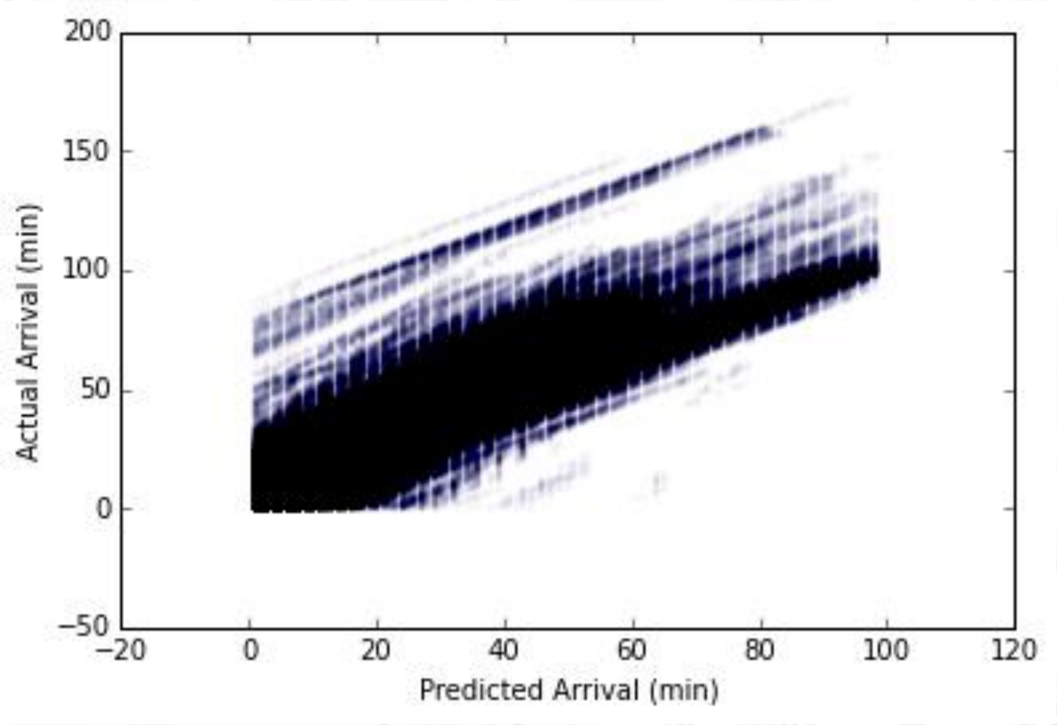
- After I have a big database of predictions, identify arrivals
- 2 possible methods
  - Predicted arrival = 0 minutes
  - Lat/Lon check
- Pull the "predictions" where the bus has arrived and join that data frame to the predictions to assign the correct arrival time for each prediction
  - In my test data, covers 99% of all predictions
- The 'Error' is the difference between predicted arrival and actual arrival



# A “Quality” Metric For Predictions

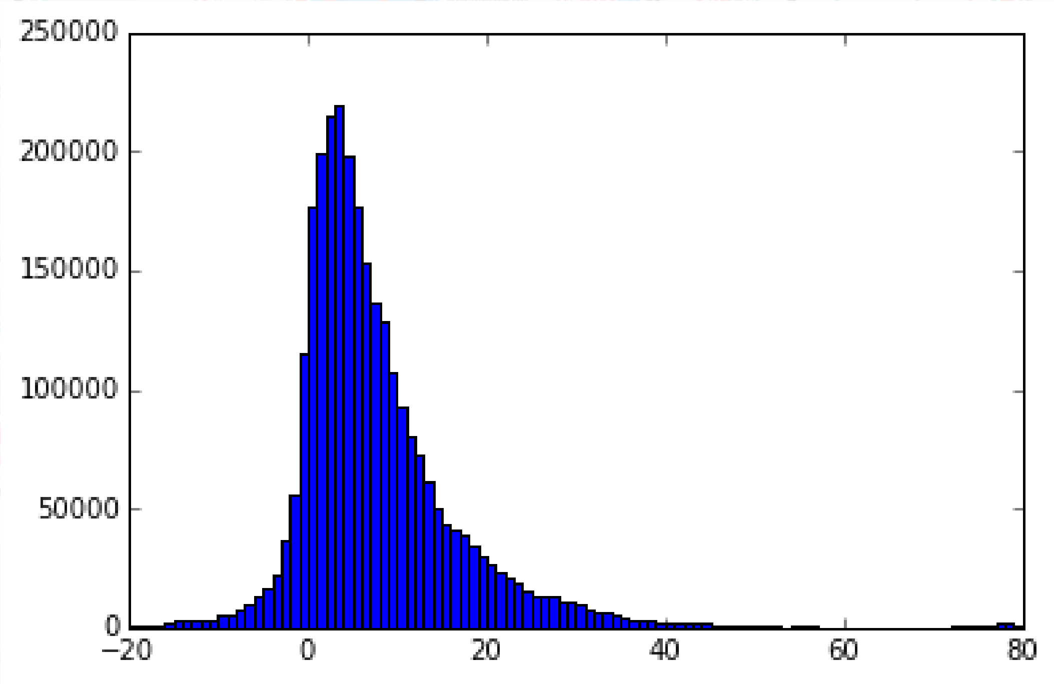
- Attempted to be reflective of our expectations for a prediction:
  - <1 is “perfect”, far off = sucky.
  - Should be more forgiving for further outCou be more critical of being wrong early (since you will miss the bus?
- For this pass, I used [1 - |predicted – actual| / actual]
  - Predicted 2 minutes, actual 4 minutes = 0.5
  - Predicted 5 minutes, actual 7 minutes = 0.71
  - Predicted 10 minutes, actual 12 minutes = 0.83

# Results: NextBus Performance



# Results: NextBus Performance

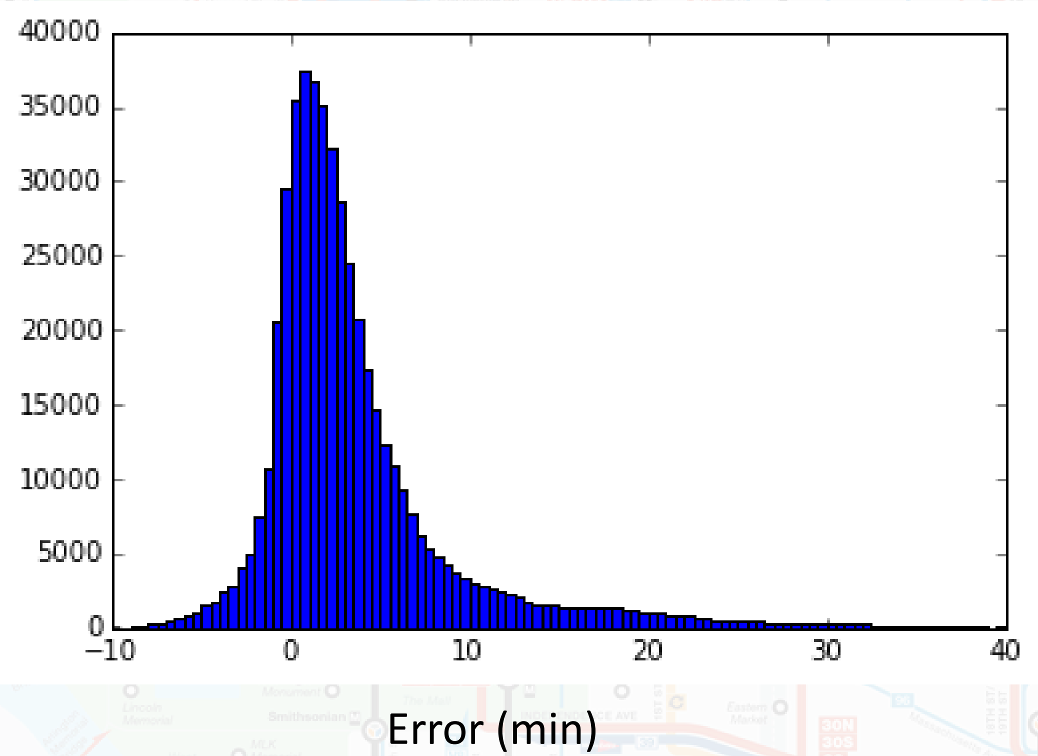
All Predictions



Actual Arrival – Predicted Arrival (positive means the bus was later than predicted)

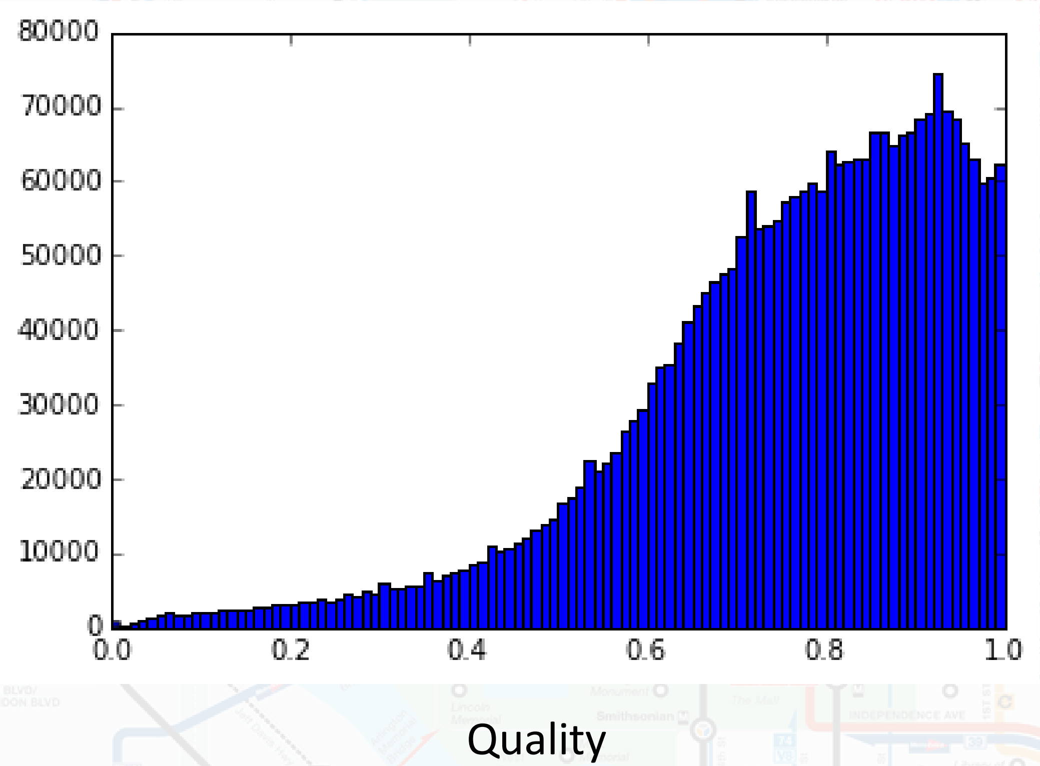
# Results: NextBus Performance

Only predictions < 10 mins



Actual Arrival – Predicted Arrival (positive means the bus was later than predicted)

# “Quality” Metric

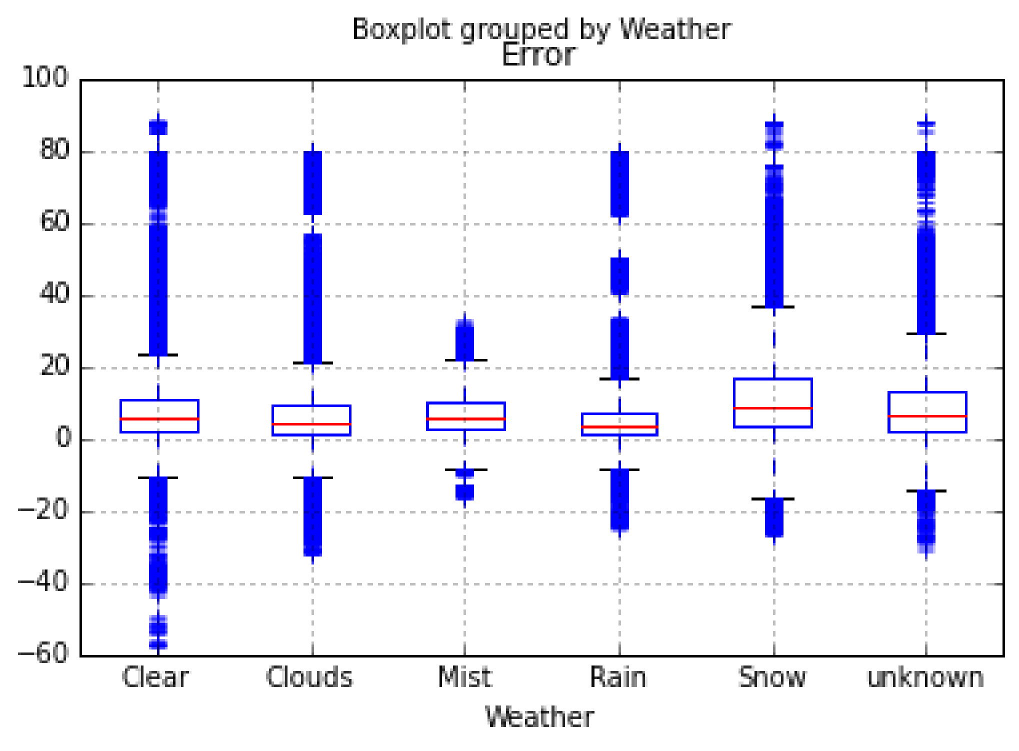


```
mean        0.641955
std         3.706173
min     -1198.000000
25%         0.647815
50%         0.782870
75%         0.893502
max         1.000000
```

Recall 0.5 means it was as far off as the total time the bus took to arrive.
About half of predictions are better than 0.8, meaning ~ within 20%

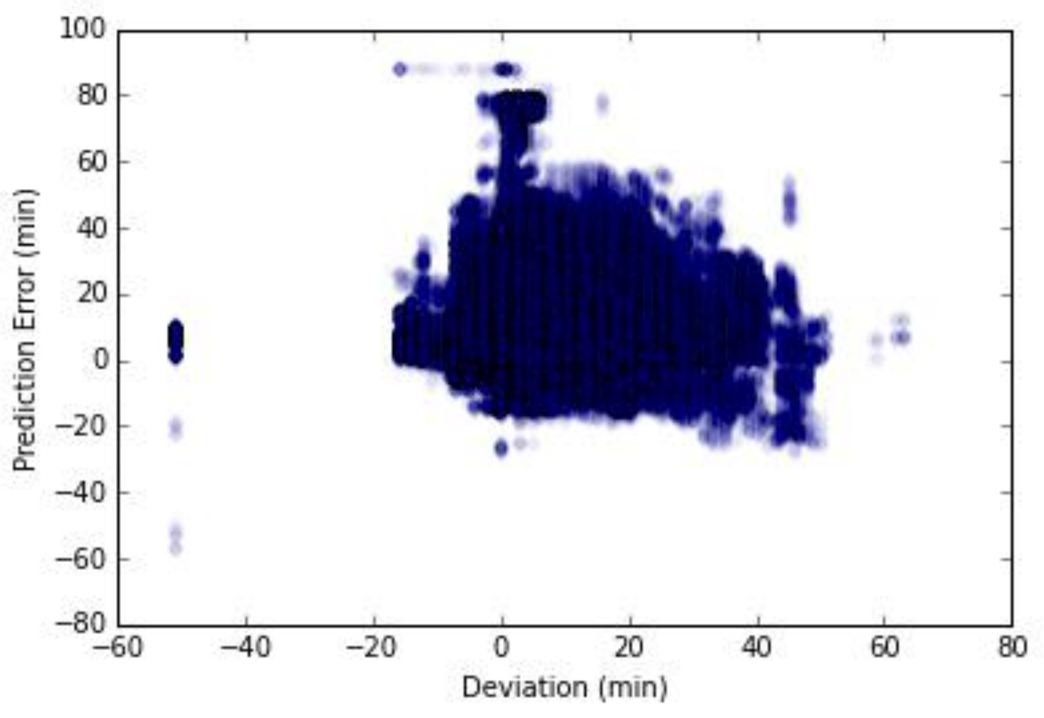
# Results: NextBus Performance

- Some evidence (from one day!) that snow increases error

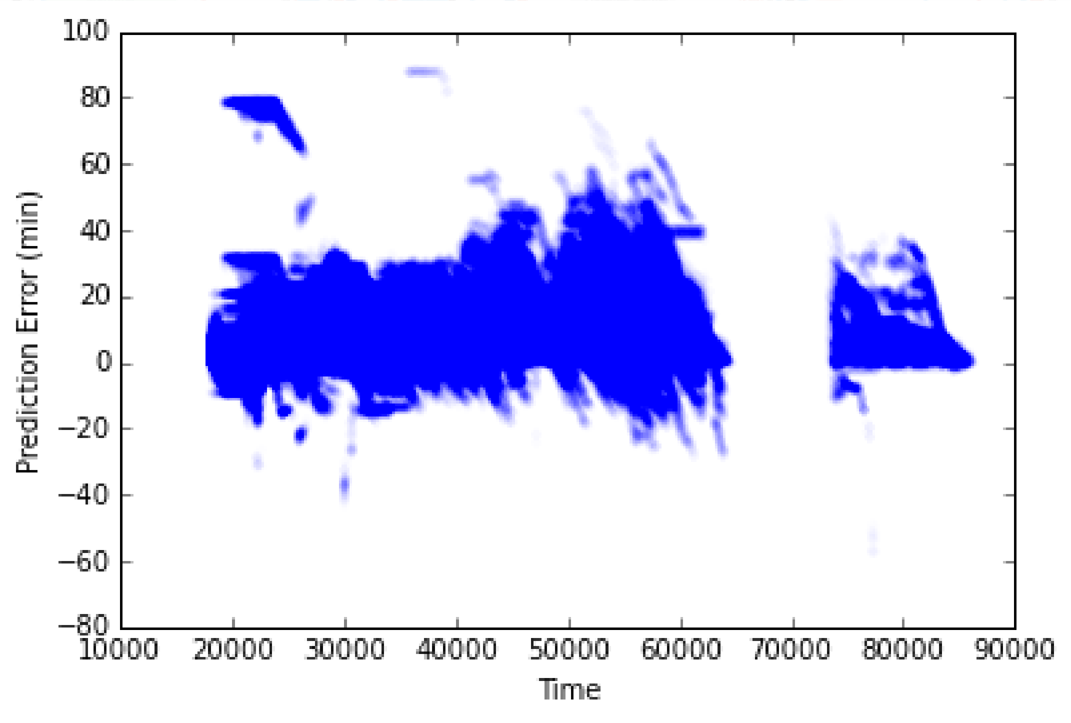


# Results: NextBus Performance

- Lateness looks like somewhat of a predictor of more lateness



# Results: NextBus Performance



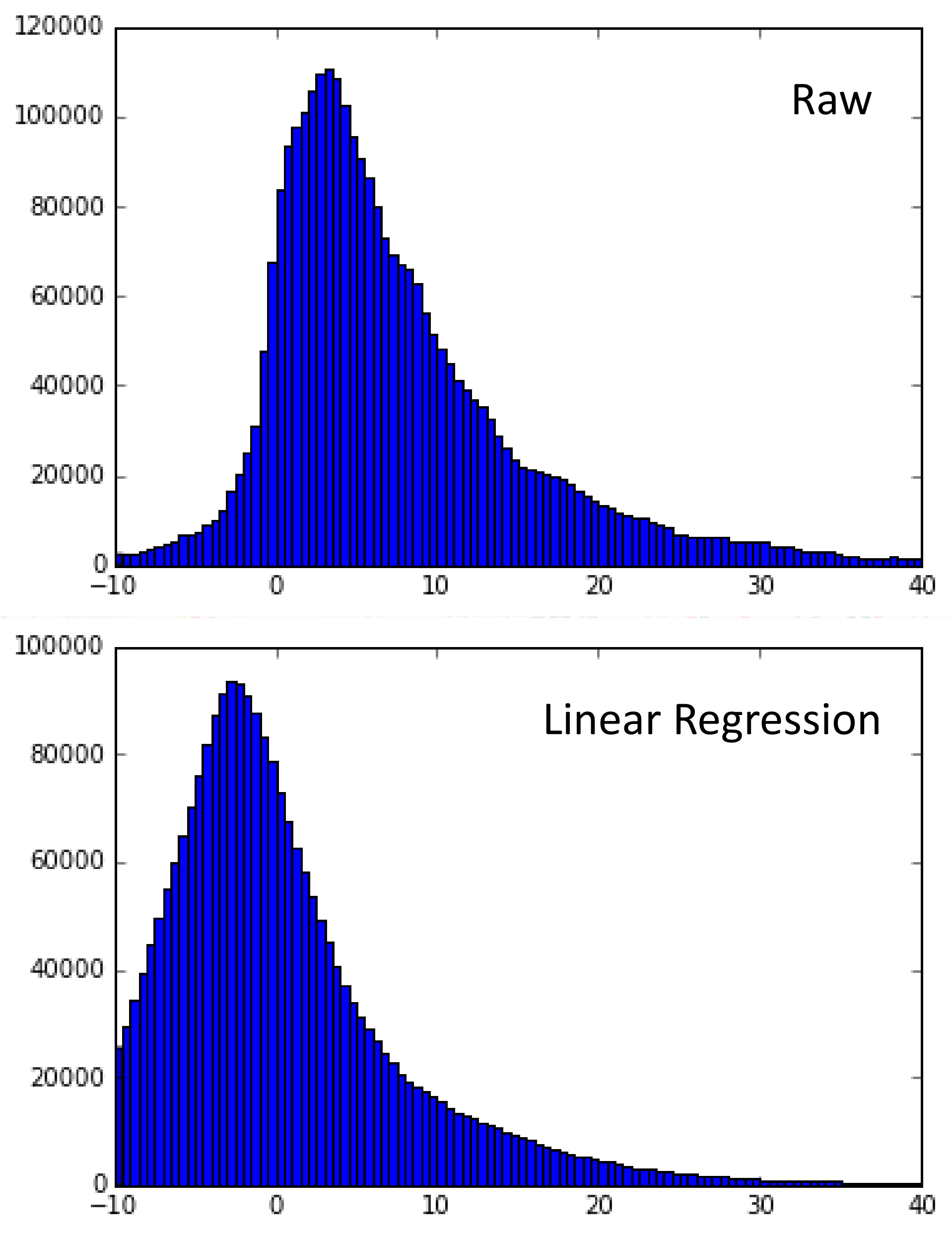
# Regression To Improve Predictions

- Variables considered
  - Predicted Arrival (hopefully this will be the major predictor!)
  - Deviation (how late the bus is reporting itself to be)
  - Weather (text descriptor, created dummy variable)
  - Time (for now, as a linear variable… non-ideal)

# Regression Results

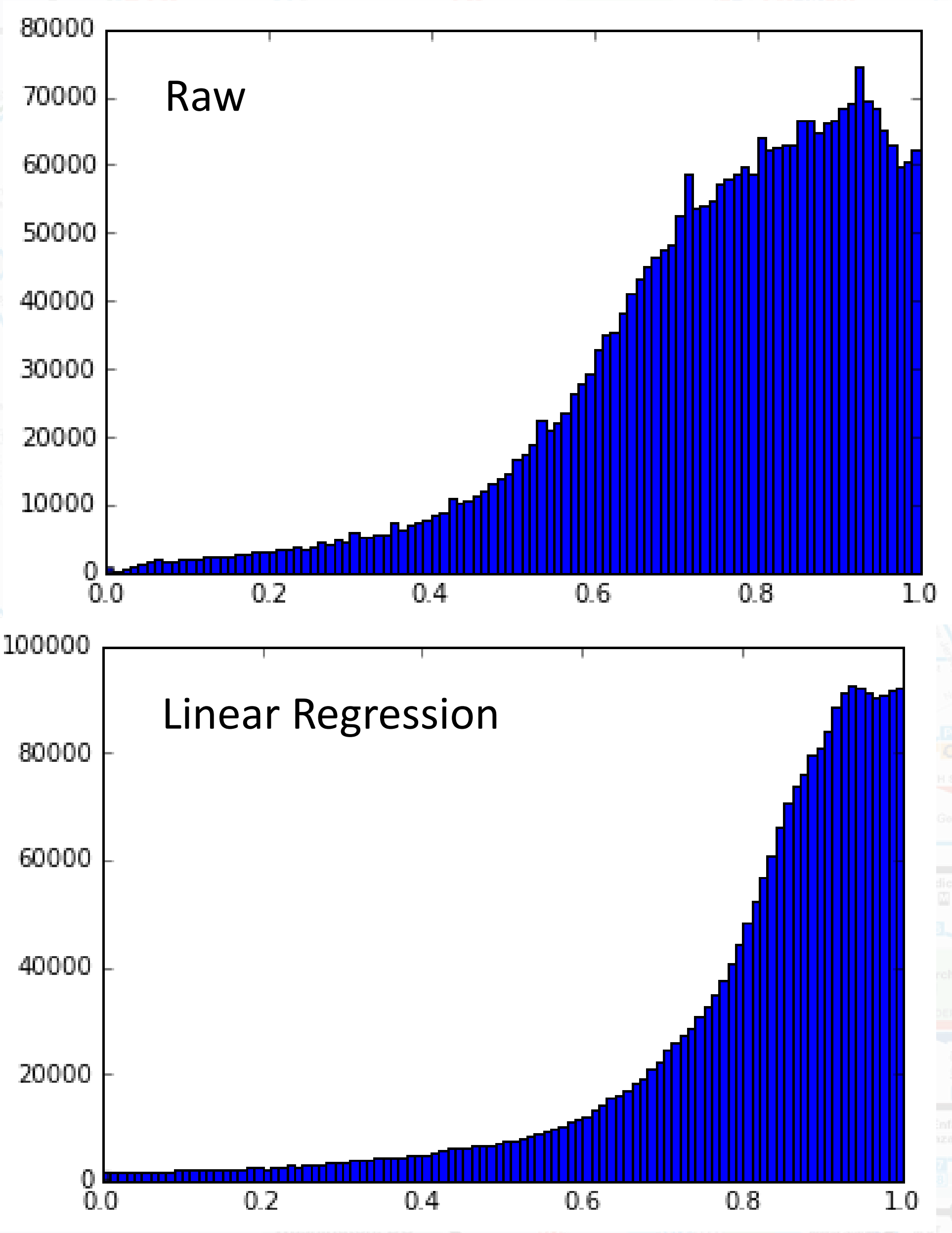
- On all data:
  - Predicted arrival: 1.107 (predictions high by 10% on average)
  - Deviation: -0.0007
  - BadWeather: 1.028 (on average rain / snow increase error)
  - Time: 0.0002
  - $R^2$ = 0.836 (same when done w/ train-test-split)

# Regression Results



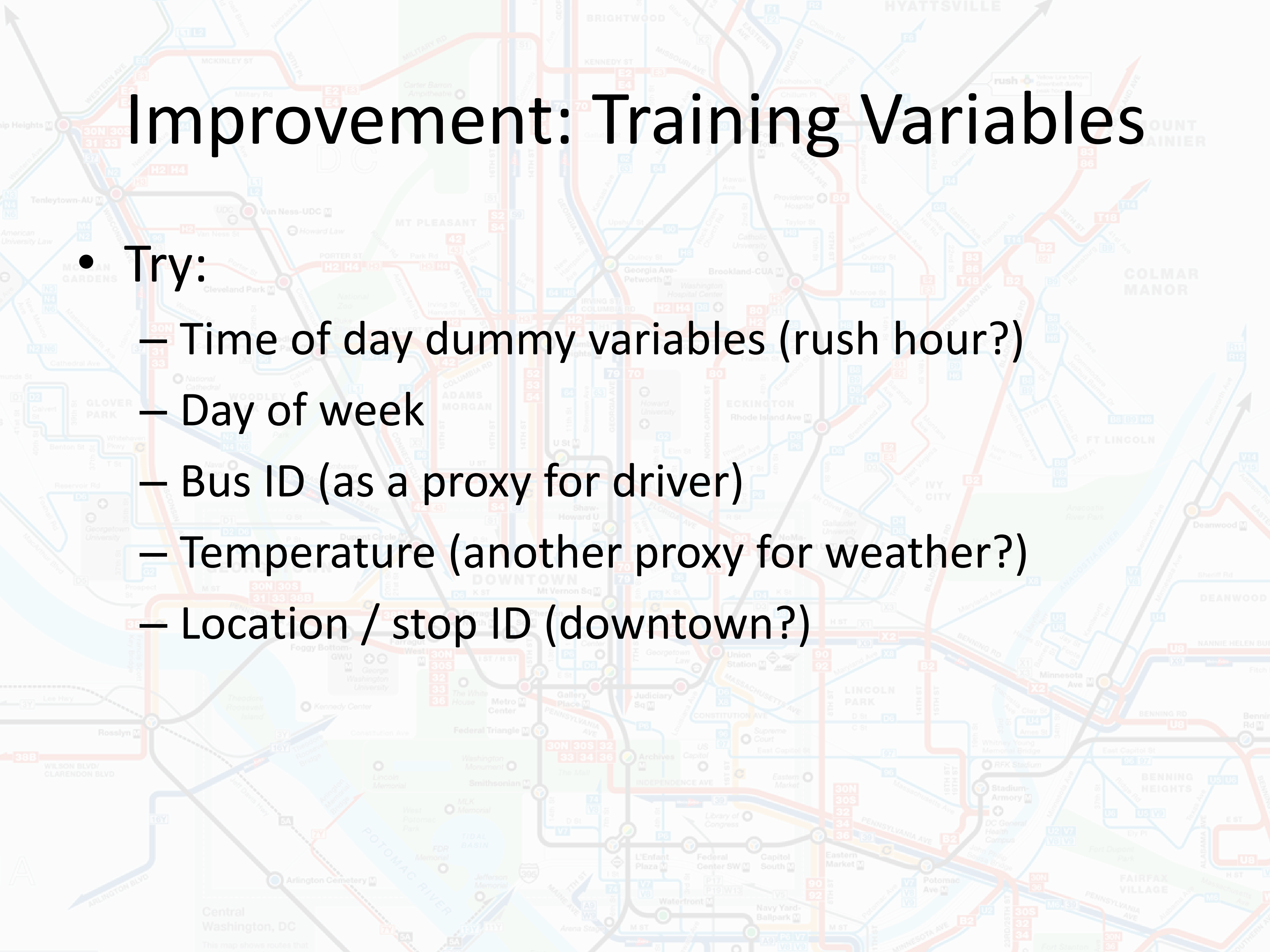
- Not really an improvement, seems to fix one problem and cause another

# Regression Results



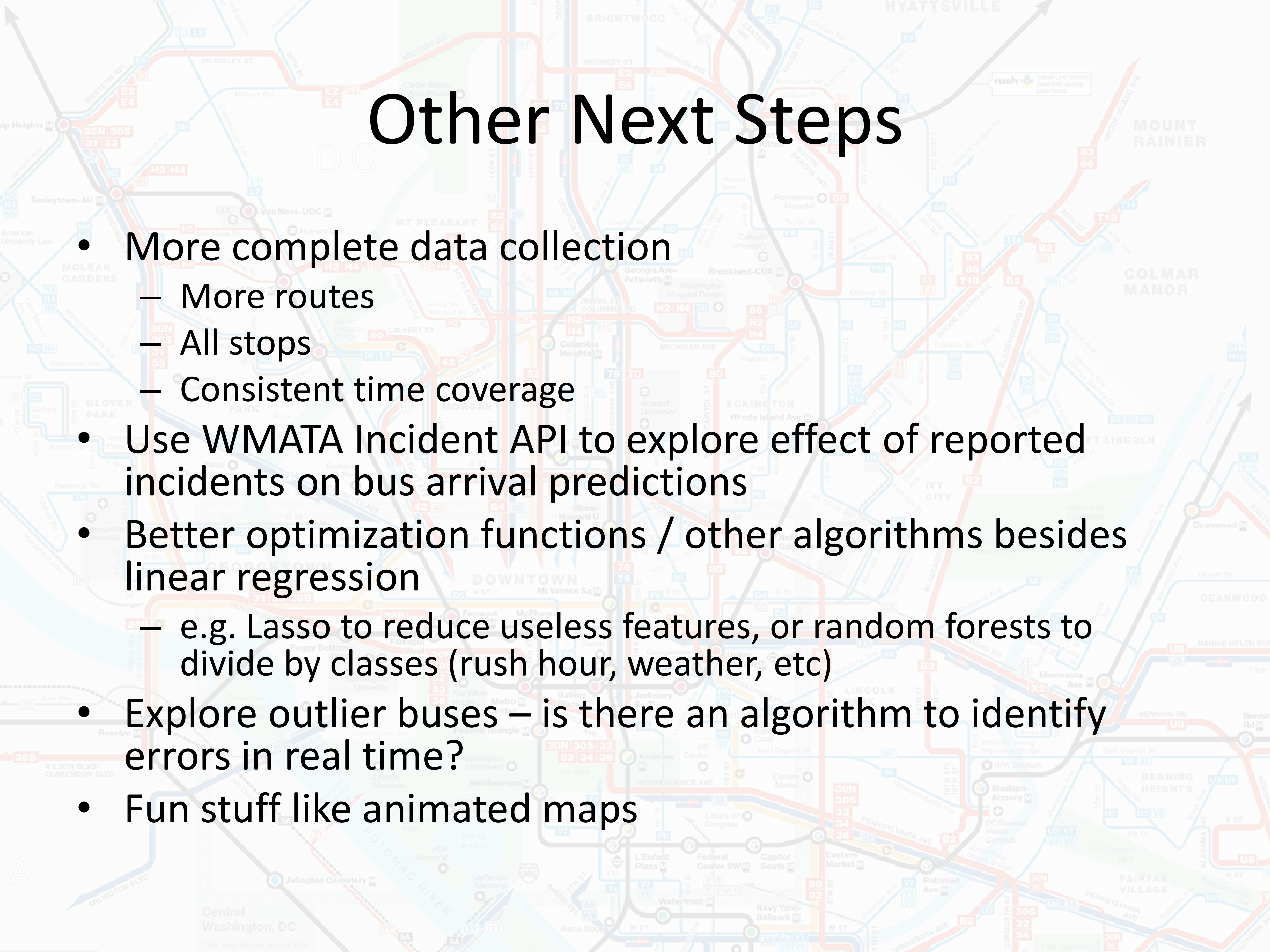
- On quality metric, looks better – but I think this means my quality metric isn't great...

# Improvement: Training Variables



- Try:
  - Time of day dummy variables (rush hour?)
  - Day of week
  - Bus ID (as a proxy for driver)
  - Temperature (another proxy for weather?)
  - Location / stop ID (downtown?)

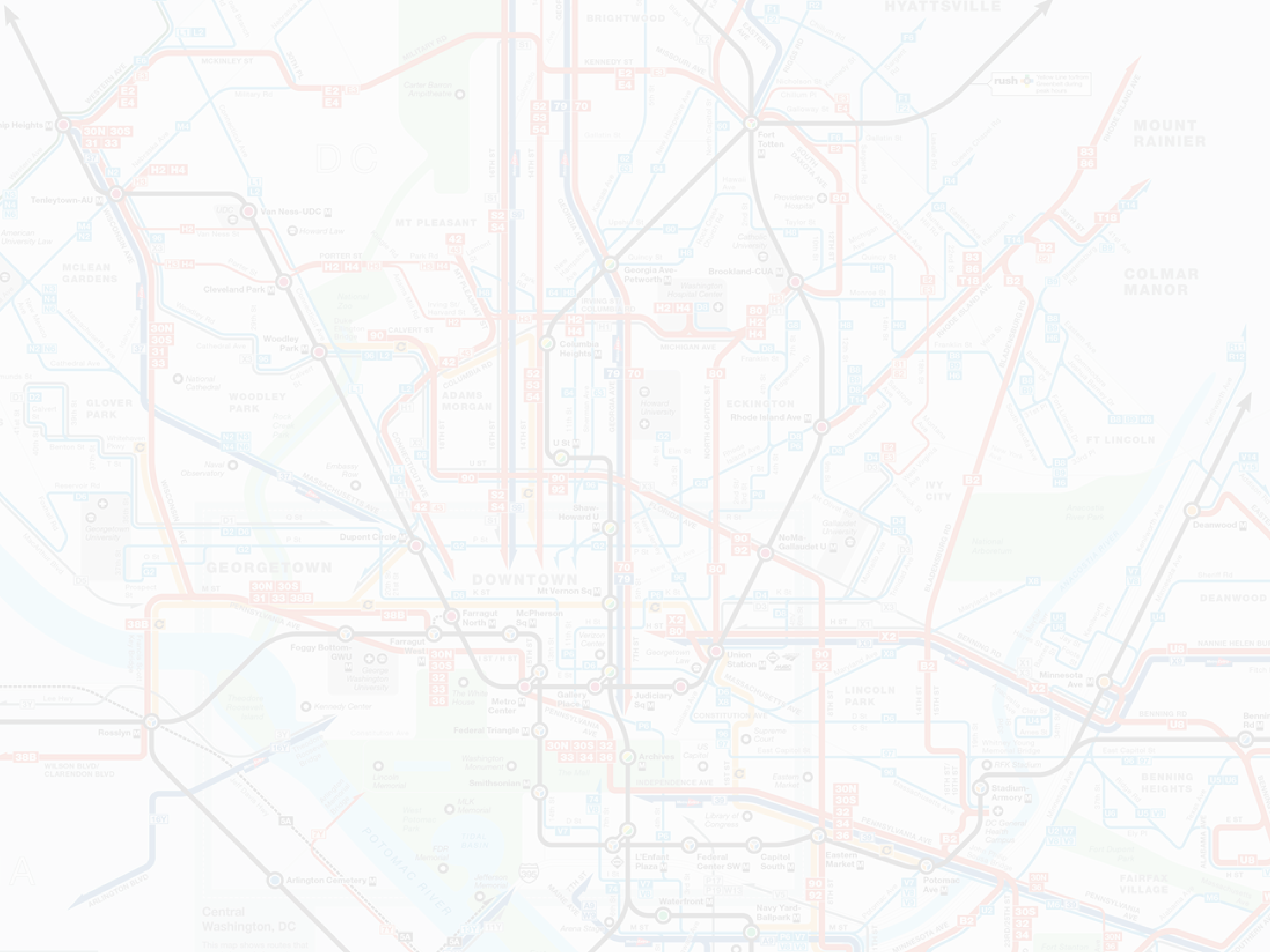
# Other Next Steps



- More complete data collection
  - More routes
  - All stops
  - Consistent time coverage
- Use WMATA Incident API to explore effect of reported incidents on bus arrival predictions
- Better optimization functions / other algorithms besides linear regression
  - e.g. Lasso to reduce useless features, or random forests to divide by classes (rush hour, weather, etc)
- Explore outlier buses – is there an algorithm to identify errors in real time?
- Fun stuff like animated maps

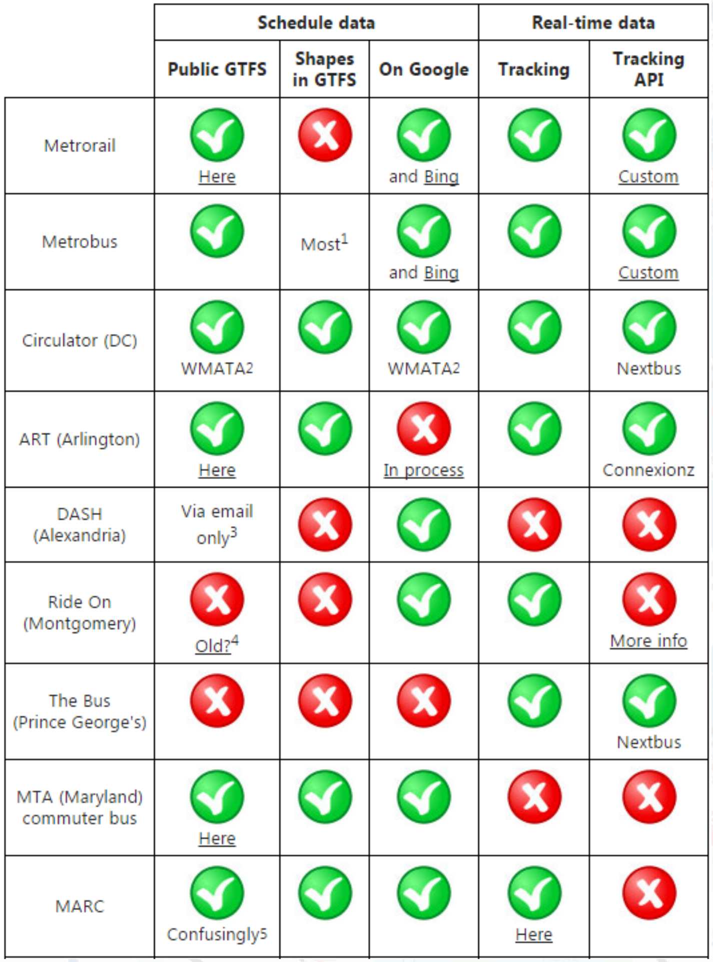
# Thank You!

When you ride ALONE you ride with Hitler!
Join a Car-Sharing Club TODAY!

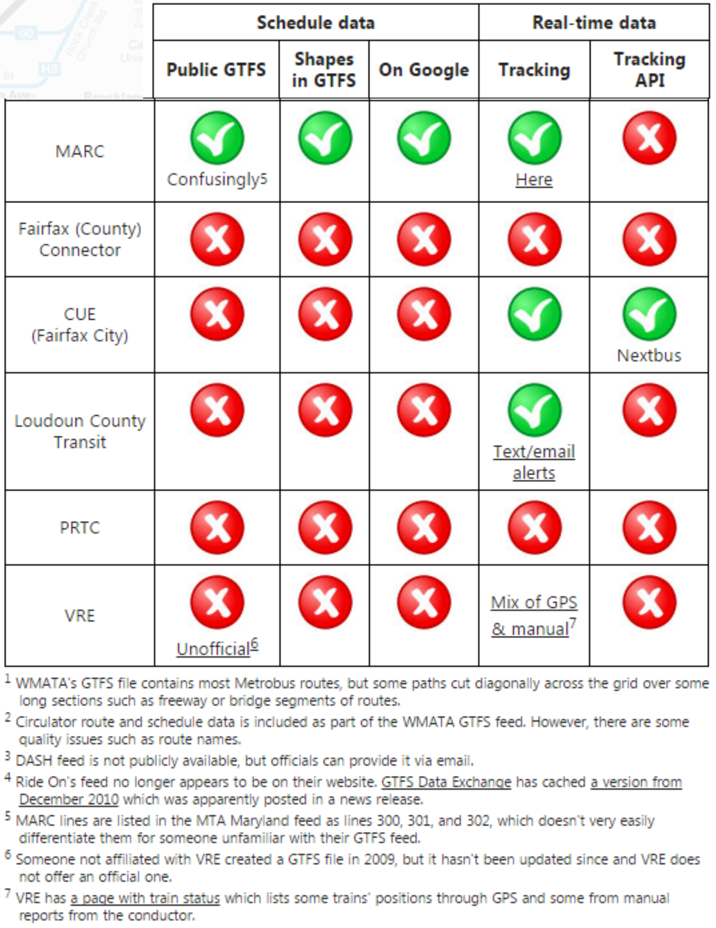


HYATTSVILLE
BRIGHTWOOD
MOUNT RAINIER
COLMAR MANOR
DC
MT PLEASANT
MCLEAN GARDENS
GLOVER PARK
WOODLEY PARK
ADAMS MORGAN
ECKINGTON
FT LINCOLN
IVY CITY
GEORGETOWN
DOWNTOWN
DEANWOOD
LINCOLN PARK
BENNING HEIGHTS
FAIRFAX VILLAGE
Tenleytown-AU
Van Ness-UDC
Cleveland Park
Woodley Park
Dupont Circle
Fort Totten
Georgia Ave-Petworth
Brookland-CUA
Columbia Heights
Rhode Island Ave
U St
Shaw-Howard U
NoMa-Gallaudet U
Deanwood
Farragut North
Farragut West
McPherson Sq
Mt Vernon Sq
Foggy Bottom-GWU
Metro Center
Gallery Place
Judiciary Sq
Union Station
Minnesota Ave
Benning Rd
Rosslyn
Federal Triangle
Smithsonian
Archives
Stadium-Armory
L'Enfant Plaza
Federal Center SW
Capitol South
Eastern Market
Potomac Ave
Waterfront
Navy Yard-Ballpark
Arlington Cemetery
POTOMAC RIVER
rush
Yellow Line to/from Greenbelt during peak hours
Central Washington, DC
This map shows routes that

# DC Transit Agencies Status (2012)



| | Schedule data | | | Real-time data | |
|---|---|---|---|---|---|
| | Public GTFS | Shapes in GTFS | On Google | Tracking | Tracking API |
| Metrorail | ✓<br>Here | ✗ | ✓<br>and Bing | ✓ | ✓<br>Custom |
| Metrobus | ✓ | Most[1] | ✓<br>and Bing | ✓ | ✓<br>Custom |
| Circulator (DC) | ✓<br>WMATA[2] | ✓ | ✓<br>WMATA[2] | ✓ | ✓<br>Nextbus |
| ART (Arlington) | ✓<br>Here | ✓ | ✗<br>In process | ✓ | ✓<br>Connexionz |
| DASH (Alexandria) | Via email only[3] | ✗ | ✓ | ✗ | ✗ |
| Ride On (Montgomery) | ✗<br>Old?[4] | ✗ | ✓ | ✓ | ✗<br>More info |
| The Bus (Prince George's) | ✗ | ✗ | ✗ | ✓ | ✓<br>Nextbus |
| MTA (Maryland) commuter bus | ✓<br>Here | ✓ | ✓ | ✗ | ✗ |
| MARC | ✓<br>Confusingly[5] | ✓ | ✓ | ✓<br>Here | ✗ |



| | Schedule data | | | Real-time data | |
|---|---|---|---|---|---|
| | Public GTFS | Shapes in GTFS | On Google | Tracking | Tracking API |
| MARC | ✓<br>Confusingly[5] | ✓ | ✓ | ✓<br>Here | ✗ |
| Fairfax (County) Connector | ✗ | ✗ | ✗ | ✗ | ✗ |
| CUE (Fairfax City) | ✗ | ✗ | ✗ | ✓ | ✓<br>Nextbus |
| Loudoun County Transit | ✗ | ✗ | ✗ | ✓<br>Text/email alerts | ✗ |
| PRTC | ✗ | ✗ | ✗ | ✗ | ✗ |
| VRE | ✗<br>Unofficial[6] | ✗ | ✗ | Mix of GPS & manual[7] | ✗ |

[1] WMATA's GTFS file contains most Metrobus routes, but some paths cut diagonally across the grid over some long sections such as freeway or bridge segments of routes.

[2] Circulator route and schedule data is included as part of the WMATA GTFS feed. However, there are some quality issues such as route names.

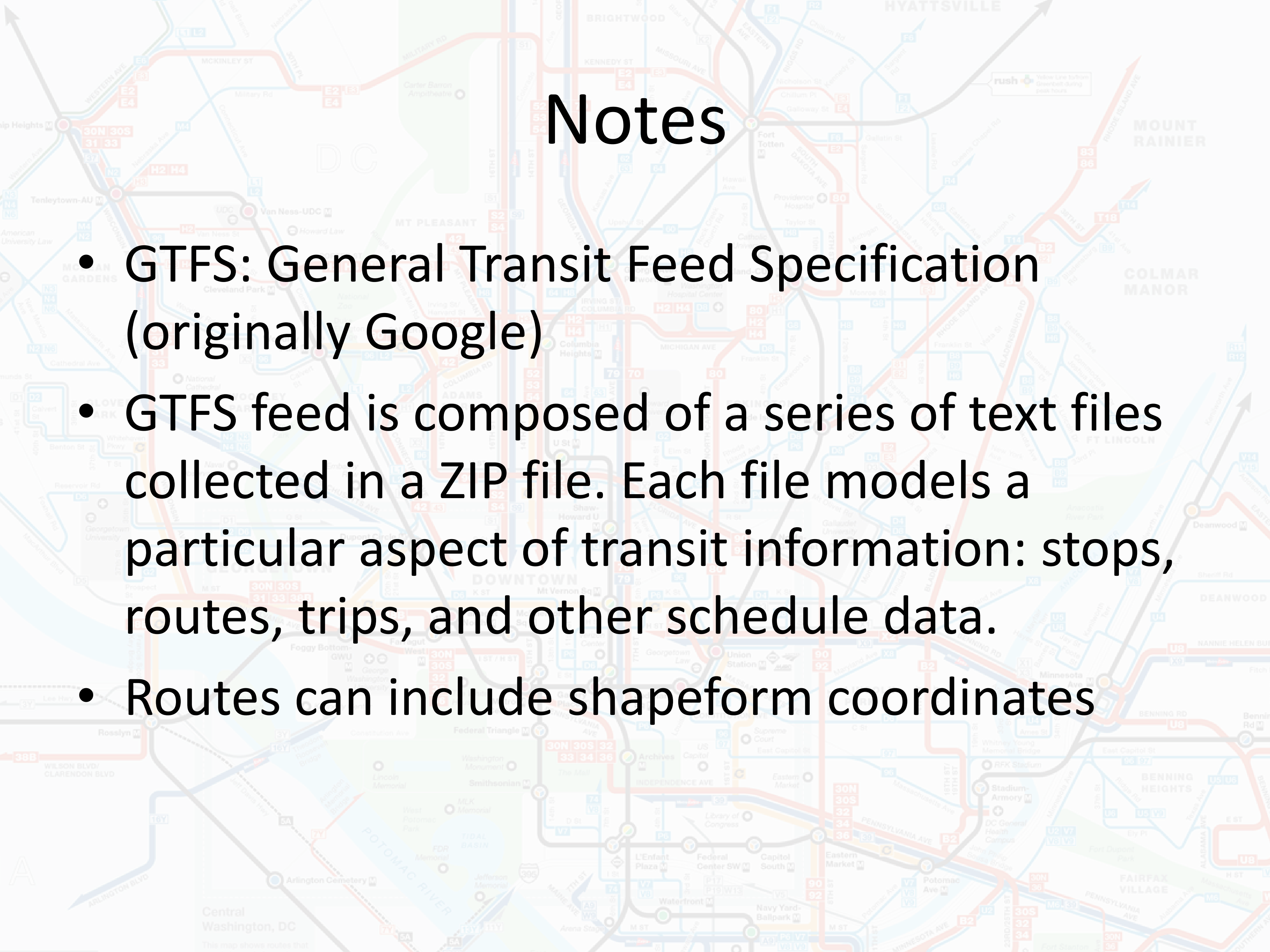
[3] DASH feed is not publicly available, but officials can provide it via email.

[4] Ride On's feed no longer appears to be on their website. GTFS Data Exchange has cached a version from December 2010 which was apparently posted in a news release.

[5] MARC lines are listed in the MTA Maryland feed as lines 300, 301, and 302, which doesn't very easily differentiate them for someone unfamiliar with their GTFS feed.

[6] Someone not affiliated with VRE created a GTFS file in 2009, but it hasn't been updated since and VRE does not offer an official one.

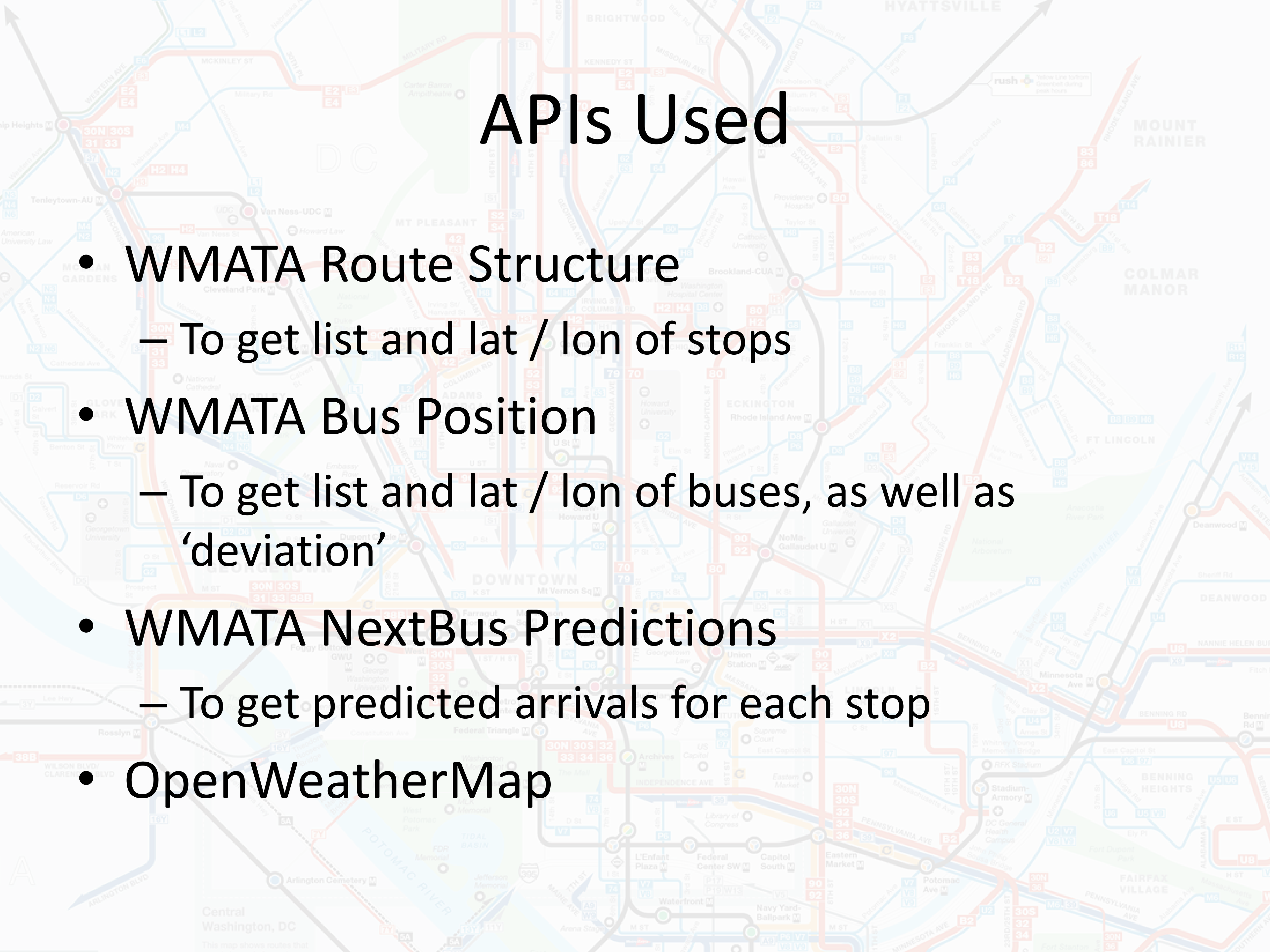
[7] VRE has a page with train status which lists some trains' positions through GPS and some from manual reports from the conductor.



# Notes

- GTFS: General Transit Feed Specification (originally Google)
- GTFS feed is composed of a series of text files collected in a ZIP file. Each file models a particular aspect of transit information: stops, routes, trips, and other schedule data.
- Routes can include shapeform coordinates

# APIs Used



- WMATA Route Structure
  - To get list and lat / lon of stops
- WMATA Bus Position
  - To get list and lat / lon of buses, as well as ‘deviation’
- WMATA NextBus Predictions
  - To get predicted arrivals for each stop
- OpenWeatherMap