Jacob Crisan
Shengjia He
Alex Bailey
Aditya Sharoff

DataEng: Project Assignment 3

Data Integration

Assignment date: February 16
Due date: February 28, 2021 @10pm PT
Submit: assignment submission form

Congratulations! By now you have a working, end-to-end data pipeline. Unfortunately, it does not have enough data to properly implement our Data Scientist's visualization. To fill out information such as "route ID" you need to access another source of data and build a new pipeline to integrate it with your initial pipeline. Here are your steps:

A. access the stop event data
B. build a new pipeline for the stop event data
C. integrate the stop event data with the bread crumb data
D. testing

A. Stop Event Data

Access C-Tran "Stop Event" data at this URL: http://rbi.ddns.net/getStopEvents As with the previous data source, this data set gives all C-Tran vehicle stop events for a single day of operation.

B. New Pipeline

Your job is to build a new pipeline that operates just like the previous one, including use of Kafka, automation, validation and loading.

C. Integrate Stop Events with Bread Crumbs

The two pipelines (BreadCrumb pipeline and StopEvent pipeline) must update the values in the Trip table such that all of the columns of both tables are filled correctly.

D. Visualization

Aman developed a new visualization tool that allows you to view your bread crumb data and display it on a map. See Aman's descHeatMapription here. Your job is to integrate this tool with

your database tables so that you can query the breadcrumb and trip data in your database server, transform to geoJSON format and display the resulting map visualization.

Submission

Make a copy of this document and update it to include the following visualizations. **For each visualization extract from your database a list of {latitude, longitude, speed} tuples** and then **use the provided visualization code (see Section D above) to display bus speeds at all of the corresponding geographic coordinates**. So, for example, if you are asked to visualize a “trip”, then you must query your database to find all of the {latitude, longitude, speed} tuples for that trip, and then display a map showing the recorded/calculated bus speed at each {latitude,longitude} location.

No need to produce software that neatly displays trips, routes, dates, times, etc. onto the visualization itself. Instead, **just paste a screen capture of the map-based speed visualization** into your submission document and then include a text description of the contents of the visualization. For example, text like this: “Bus Speeds for all outbound trips of route 65 between 9am and 11am on Sunday October 32, 2020.”

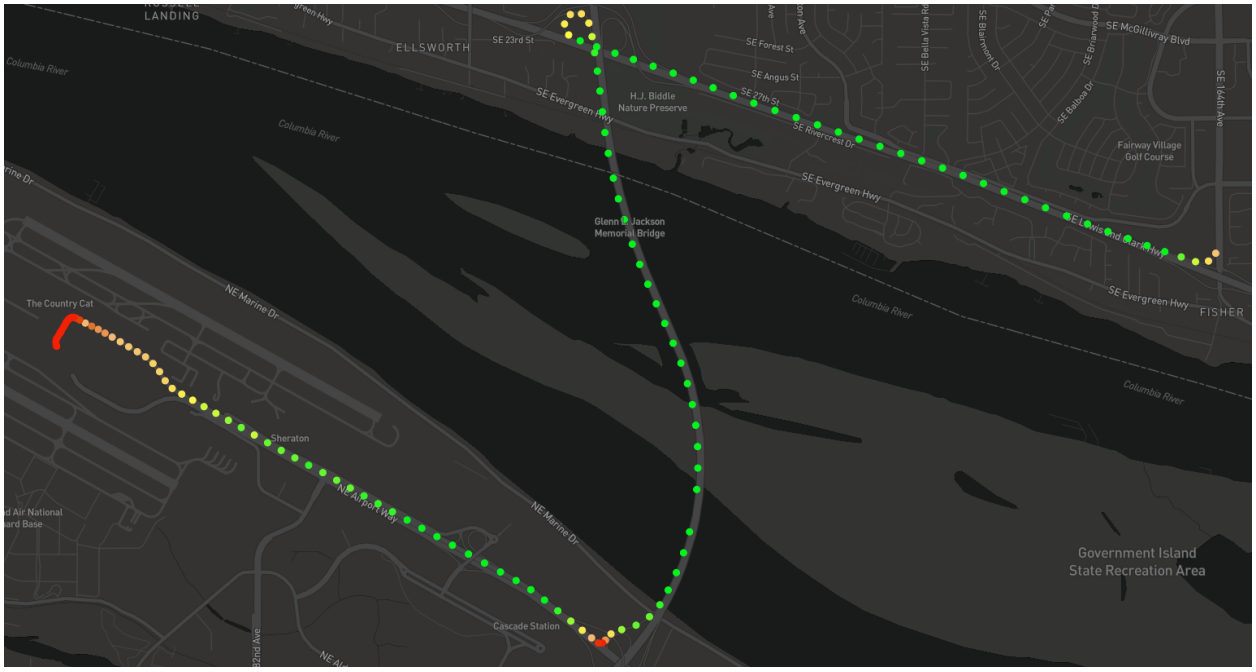
Visualization 1. A visualization of speeds for a single trip for any bus route that crosses the **Glenn Jackson I-205 bridge**. You choose the day, time and route for your selected trip. To find a trip that traverses this bridge, consider finding a trip that includes breadcrumb sensor points within this bounding box: [45.592404, -122.550711, 45.586158, -122.541270]. Any bus trip that includes breadcrumb points within that box either crosses the bridge or goes swimming in the Columbia river!

Database Query:

```
SELECT latitude, longitude, speed
FROM BreadCrumb
WHERE trip_id = 169124234;
```

*We have confirmed that the trip with **169124234** has crossed the Glenn Jackson bridge*

Answer:



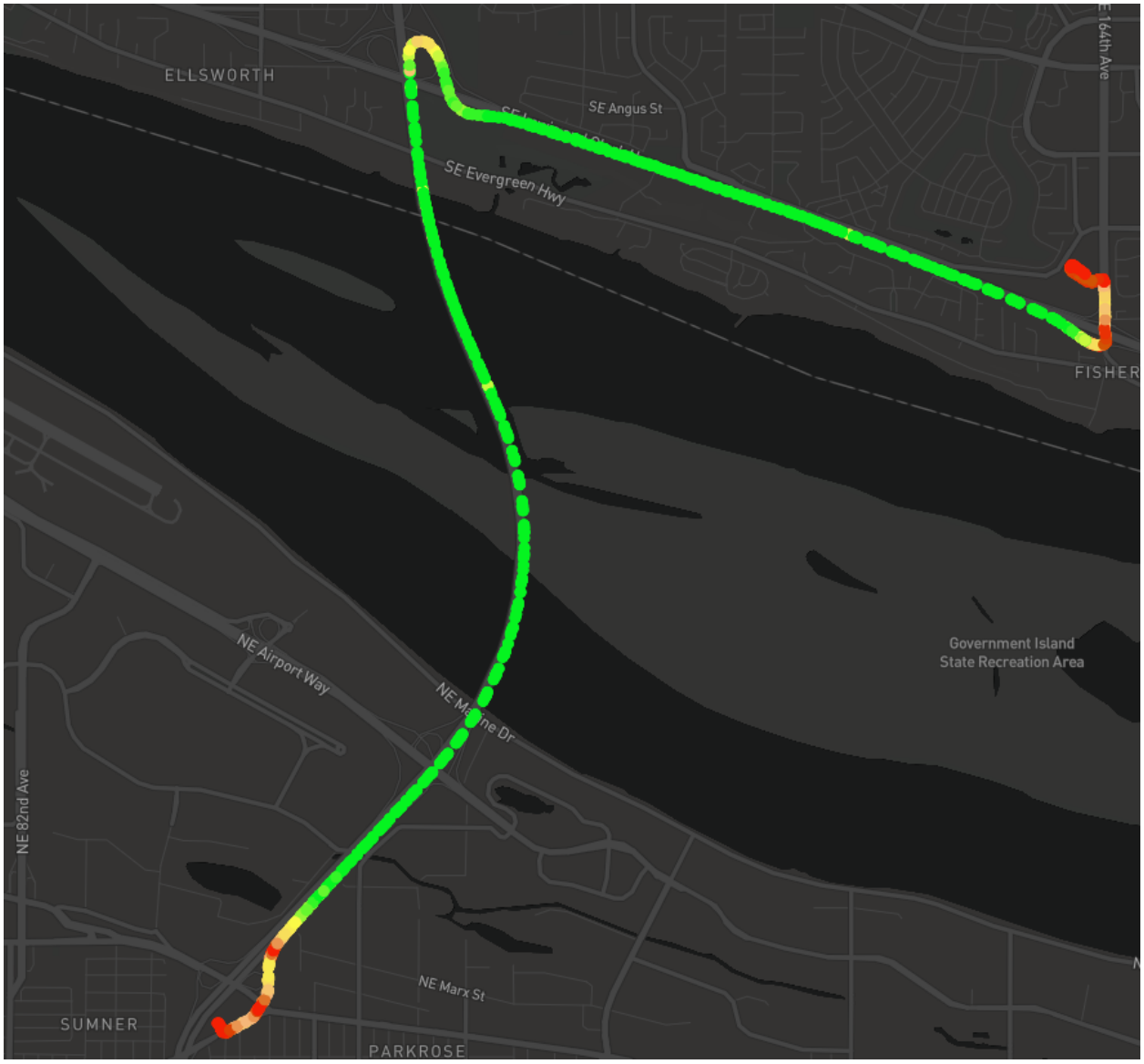
Visualization 2. All outbound trips that occurred on route 65 on any Friday (you choose which Friday) between the hours of **4pm and 6pm**.

Database Query:

```
SELECT latitude, longitude, speed
FROM Trip
INNER JOIN BreadCrumb
On Trip.trip_id=BreadCrumb.trip_id
WHERE
   tstamp::date='2020-10-12' AND
   Trip.direction='Out' AND
   tstamp::time>='16:00:00' and tstamp::time<='18:00:00' AND
   route_id=65;
```

Note: at the time of doing this query, we didn't have any trips with route_ids for a Friday, so we went with a Monday instead.

Answer:



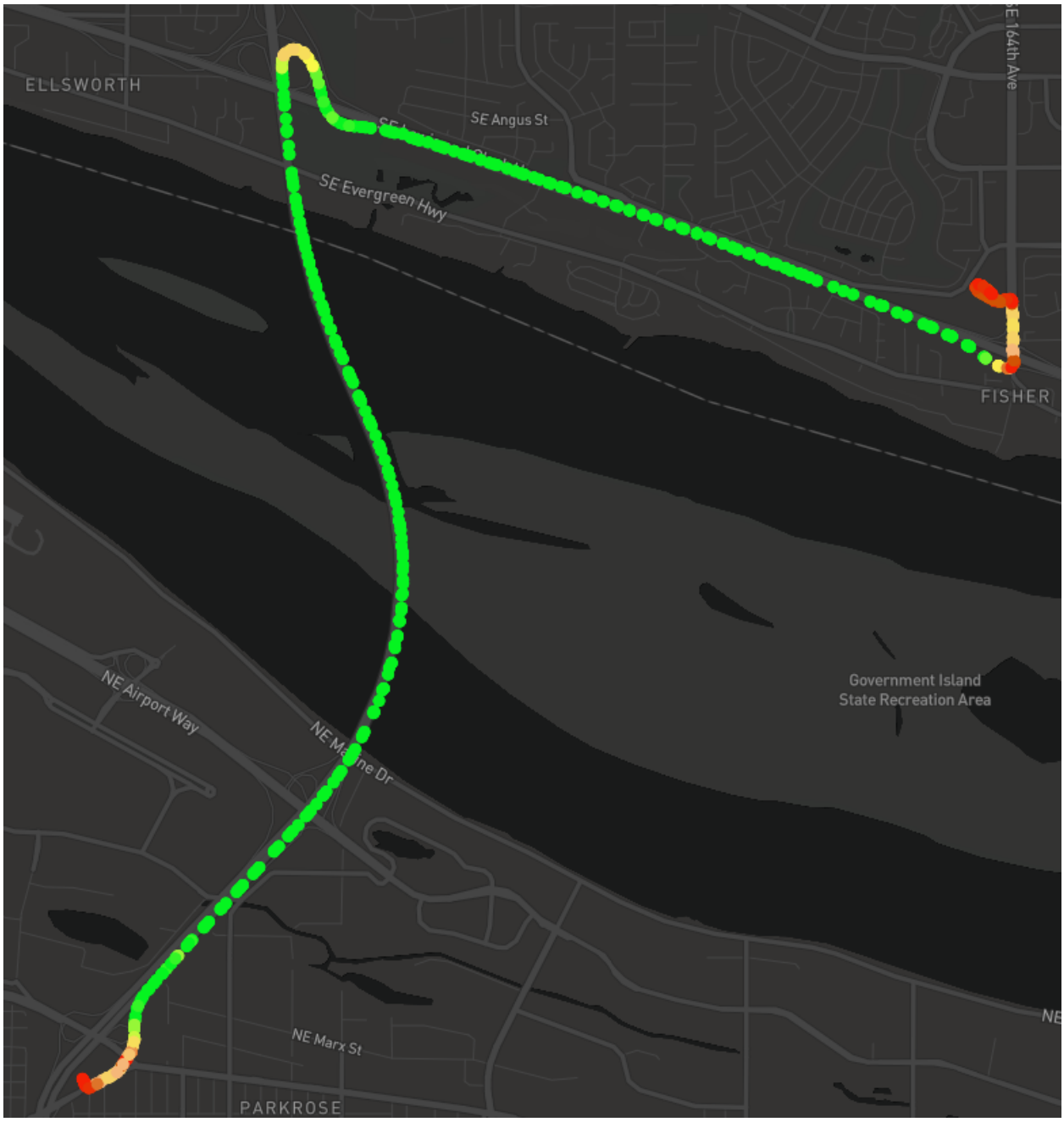
Visualization 3. All outbound trips for route 65 on any Sunday morning (you choose which Sunday) between 9am and 11am.

Database Query:

```
select latitude, longitude, speed
from Trip
INNER JOIN BreadCrumb
On Trip.trip_id=BreadCrumb.trip_id
Where
    tstamp::date='2020-10-11' and
    Trip.direction='Out' and
```

tstamp::time>='09:00:00' and tstamp::time<='11:00:00' and
route_id=65;

Answer:



Visualization 4. The longest (as measured by time) trip in your entire data set. Indicate the date, route #, and trip ID of the trip along with a visualization showing the entire trip.

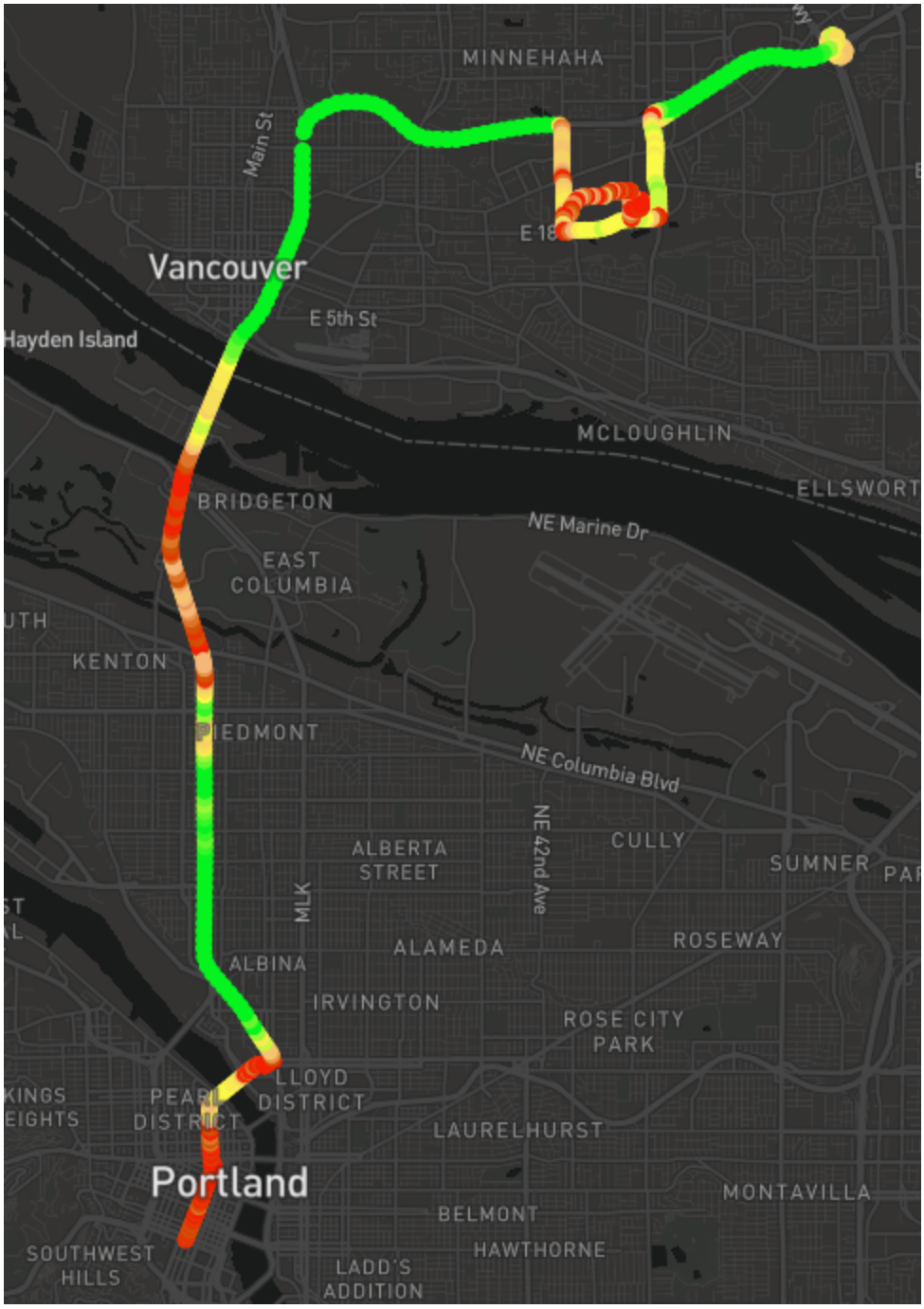
Query:

```
select latitude, longitude, speed
from BreadCrumb
Where
    Trip.trip_id=169302880;
```

Answer:

Date: 2020-10-01
Route_id: 0
Trip ID: 169302880

Note: we don't have the route_id data for this trip, and we have confirmed that this trip (169302880) is the longest trip (in terms of time)



Visualization 5a, 5b, 5c, …. Three or more additional visualizations of your choice. Indicate why you chose each particular visualization.

5a: What do the speeds look like between 6AM to 9AM (rush hour) on 2020-10-12 (a Monday)?

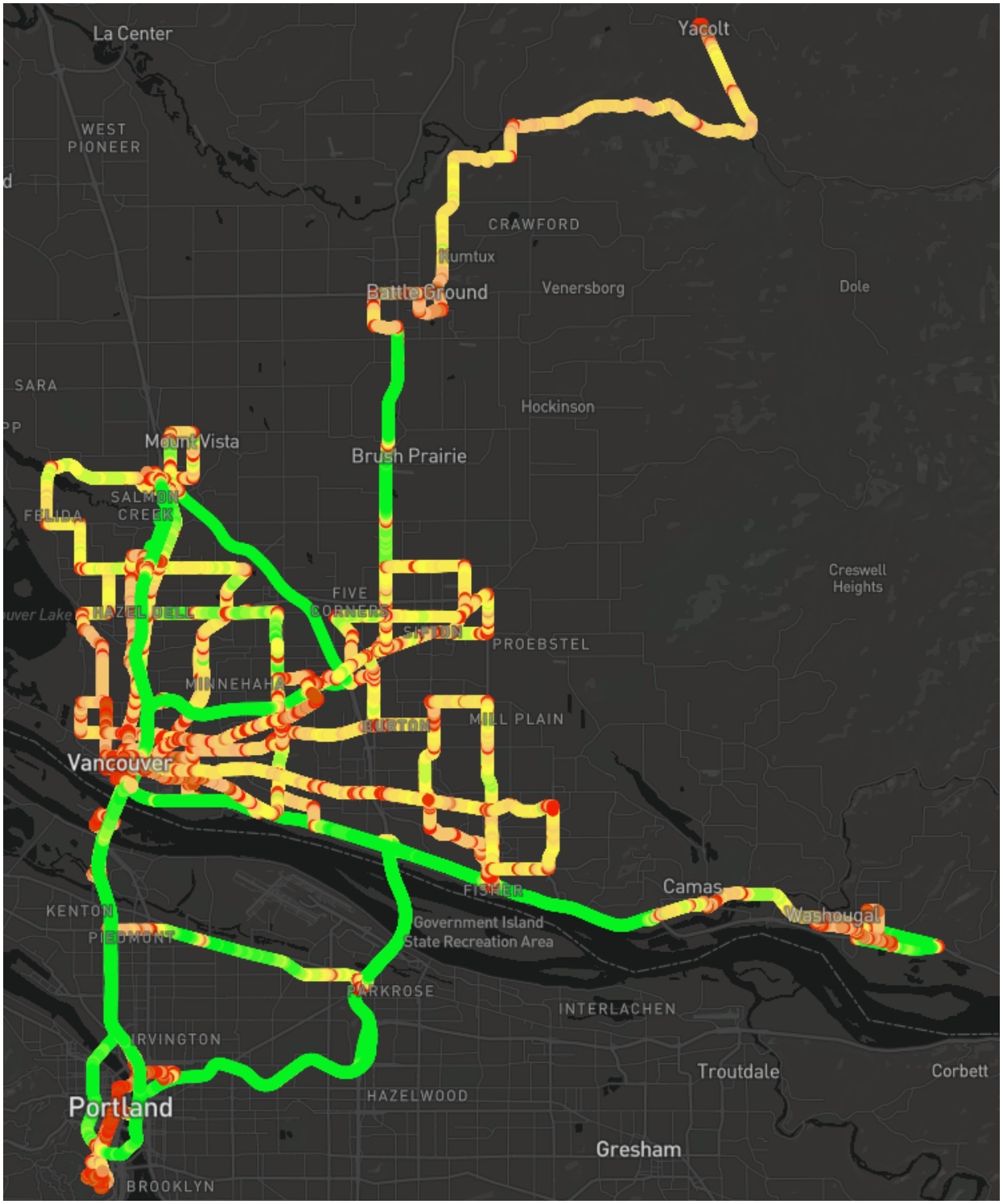
Reason:

We want to see the visualization during the A.M rush hour.

Query:

```
select latitude, longitude, speed
from BreadCrumb
Where
   tstamp::date='2020-10-12' and
   tstamp::time>='06:00:00' and tstamp::time<='09:00:00';
```

Answer:

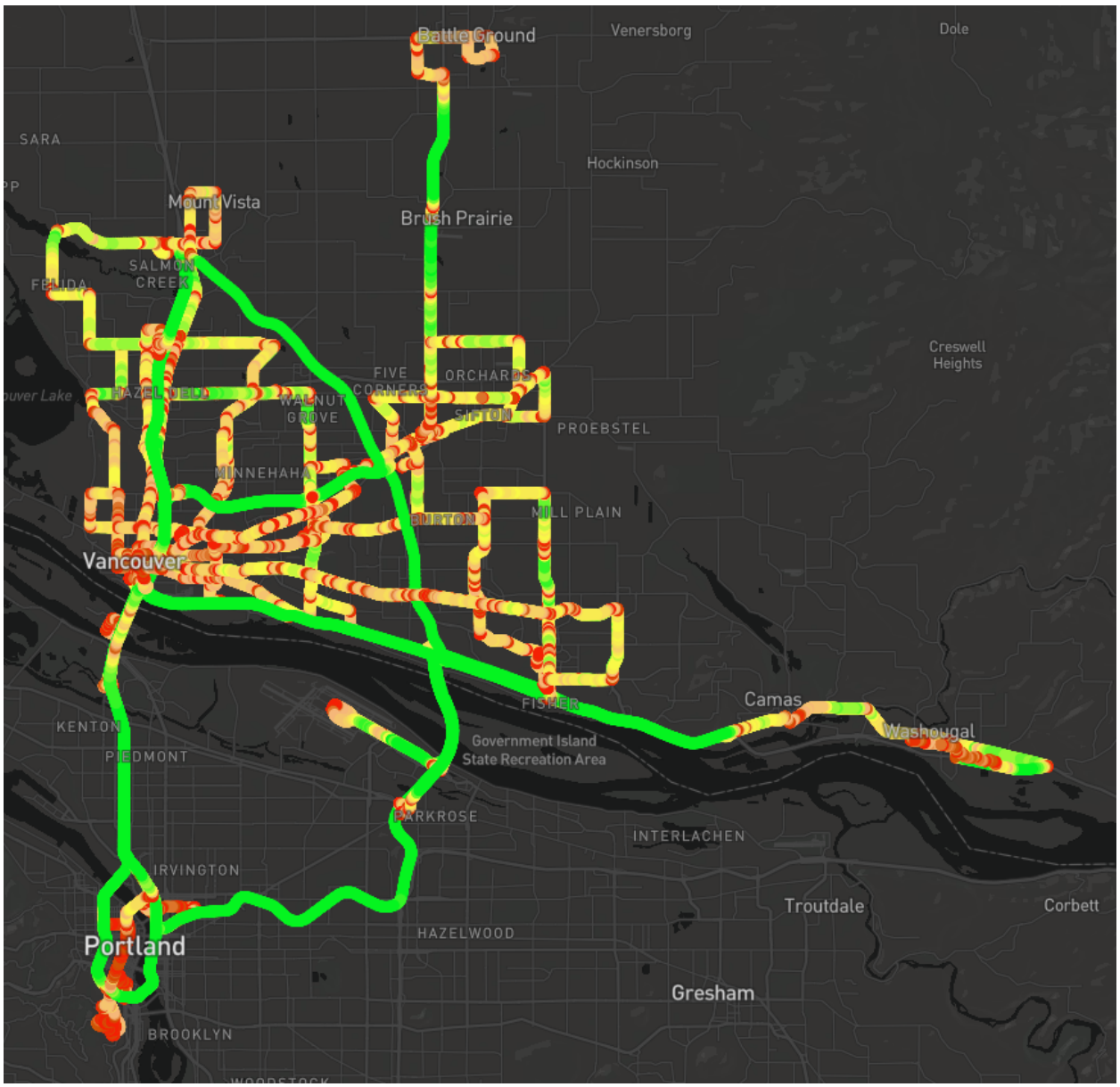

La Center
Yacolt
WEST PIONEER
CRAWFORD
Kumtux
Battle Ground
Venersborg
Dole
SARA
Hockinson
Mount Vista
Brush Prairie
SALMON CREEK
FELIDA
Creswell Heights
FIVE CORNERS
HAZEL DELL
PROEBSTEL
MINNEHAHA
BURTON
MILL PLAIN
Vancouver
FISHER
Camas
Washougal
KENTON
PIEDMONT
Government Island State Recreation Area
PARKROSE
INTERLACHEN
IRVINGTON
Troutdale
Corbett
HAZELWOOD
Portland
Gresham
BROOKLYN

5b: What do the speeds look like between 3PM to 6PM (rush hour) on 2020-10-12 (a Monday)?

Reason:
We want to see the visualization during the P.M rush hour.

Query:
```
select latitude, longitude, speed
from BreadCrumb
Where
   tstamp::date='2020-10-12' and
   tstamp::time>='15:00:00' and tstamp::time<='18:00:00';
```

Answer:

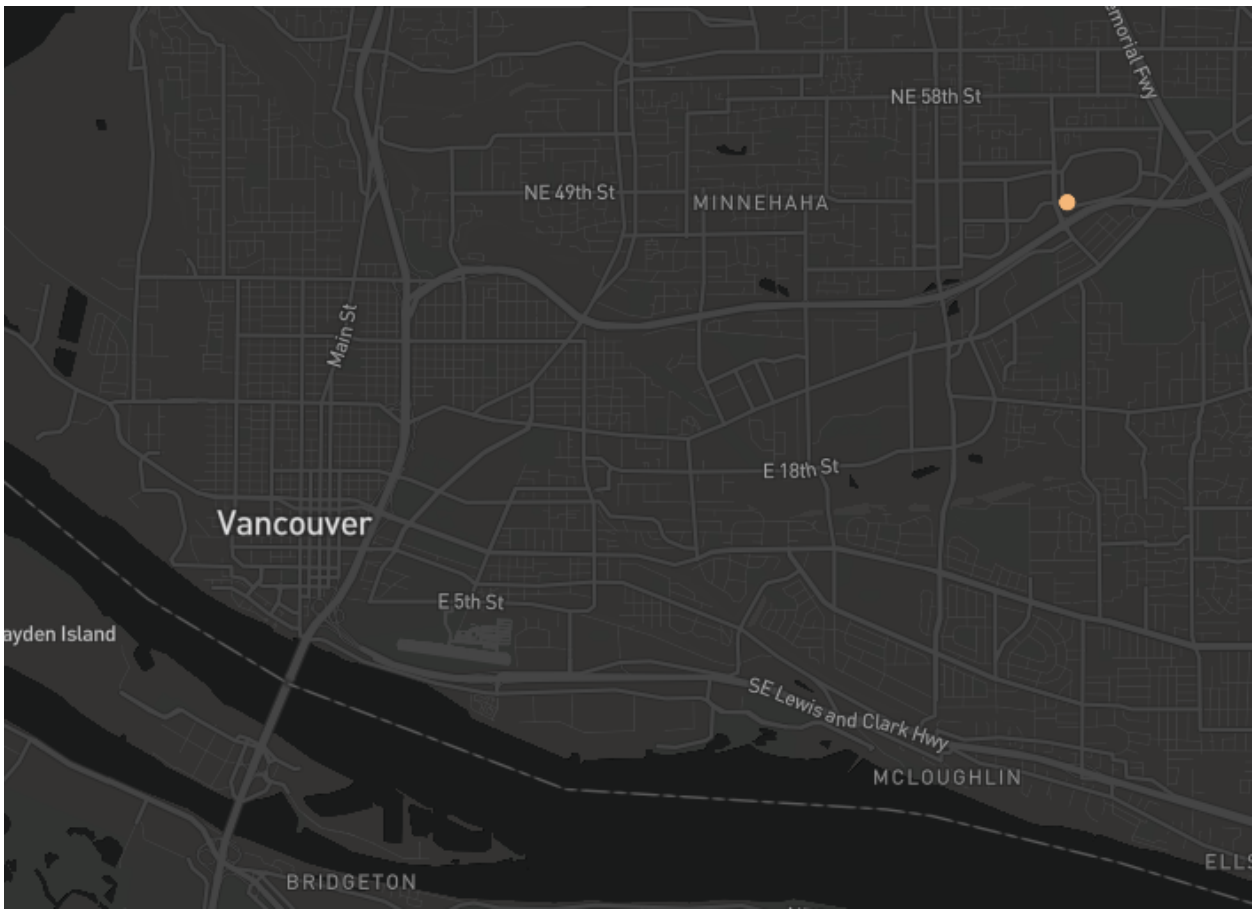


5c: What is the shortest bus route, with regards to time?

Reason:
We want to see what the shortest bus route looks like.

Query:
```
select latitude, longitude, speed
from BreadCrumb
Where
   trip_id=169763001;
```

Answer:



Note: It seems that there was only 1 stop on this trip (we had several of those actually). Maybe most of the data points got validated out?

Your Code

Provide a reference to the repository where you store your code. If you are keeping it private then share it with Bruce (bruce.irvin@gmail.com), David and Aman (github references TBD).

GitHub Reference: https://github.com/alex-bailey1/Data_Eng_project