Project 3: Group 4 + 40

Question 1 (C): Defining your goals

We have chosen to work with scenario 1.

A city may be thought of in namely two ways. In one way it is a place. It is given a name and it describes a what and where in representations of many other whats and wheres, for example on a map. But we may also think of a city in terms of a space that engenders relationships between objects and humans. Some of these relationships come to be managed within the context of city planning and traffic management. A city has infrastructure and facilities that support the lives of residents in various ways. As such, a scenario that the city planners and traffic managers have to deal with could be improving mobility infrastructure for these residents with attention to the forms of transport that exist, which in the case of this paper is pedestrians and vehicles. These are the means through which residents travel from A to B for whatever purpose. Mobility tracking and transport visibility is in itself certain view of a city and a certain understanding of the relationship between its objects.

There are, however, limitations to the movement that pedestrians and vehicles can make within the city. For vehicles, movement has to take place via designated roads. Pedestrians may travel more freely and deviate from the roads. In this way there are things we know about the mobility of residents in our city space because we have designed the space physically. But there is arguably value in knowing how the space is being used because this can help to improve the infrastructure for residents (or other stakeholders) if the insights make their way to those in power to make decision that money should be spent towards this goal of improvement.

An important step towards mobility improvement, then, is tracking mobility and making the movement of our relevant actors, visible, so we can make sense of it as a reality that exists in our city. What we really want is to see movement, even just partially. A way to do this could be having cameras scattered across the city. Another way, which we are able to imagine in our city, is to have our residents and their vehicles equipped with data-producing technology that can be retrieved by the city government for use.

In our case tracking is enabled by the fact that a fair percentage of vehicles have networking capabilities. They are able to produce location data in the form of longitude and latitude matched to a vehicle. These will be visible in the data. Vehicles without such capabilities will not be visible, which poses a problem of representation for the residents driving these vehicles, but also for the accuracy of whatever further knowledge is made based on the data. The behavior of these un-networked vehicles has an effect on the city mobility in the physical environment, but as the behavior cannot be seen in the data, we also cannot know how this behavior impacts the behavior that IS seen. This would have to be mitigated somehow.

Pedestrians are also producing data and are visible in the dataset in question. The same point can be made about pedestrians as above: for whatever reasons, some pedestrians may not be equipped with the technology to produce and send data to the city government. Only the data-producing and those contributing to the view of data are represented. In the case that and they are contributing data it is detrimental to their representation as citizens, and for those other pedestrians the accuracy of improvements for those who ARE represented.These are the human actors represented in the data. The other aspects of infrastructure visible is location in the form of longitude and latitude. We cannot know where every resident is at any given time. The data we have is timestep data that records the location and other data points of a certain pedestrian of vehicle at a point in time. Using the data effectively means we have to look at throughout recorded or captured time or in real time.

Areas where data can be used for improving person mobility and transport management are for example congested areas. Roads that are bottleneck at certain times of day for a period of time. We may also adopt an environmentally friendly strategy and aim to improve how the city is planned to have green areas further away from congested areas so that pollution does not take away from the positive environmental effects of parks and green spaces. A city could look into having areas be car-free at certain times, maybe even permanently. Signs could be implemented to provide better updated information on congested roads.

We want to identify patterns of stopping of vehicles. This could indicate that they break a lot and that the road is congested. This can will lead to higher fuel consumption and time consumption for the people driving. We can propose improvements so the city can become more green and more efficient.

Question 2 (C/T): Batch layer

1. **Car density**

How can we use the data to gain insight to the density of cars, and why does such insight matter? It can matter if you infer that a high density of cars is cause for elevated pollution levels in areas affected by this. Knowing where car density is high can allow urban planners to take this factor into account when implementing green spaces and outdoor recreation facilities. Two approaches present themselves when you have the insight: you can either have more green spaces like parks in areas with high car density to counter the emissions’ effect on air quality and aural environment in those same areas. The parks and recreational areas would then essentially act as safe havens for residents of noisy and polluted neighborhoods. The other way to take action upon the insight would be to use car density measurements for mapping areas where green spaces and recreational areas should not be implemented, and instead moving clusters of such facilities to areas further from the noise and pollution, so the benefits of green and recreational spaces are not hamstrung. In this view we would have to take the lanes as best indicator of geographical space, which is arguably a poor indicator of such. Since lanes is a part of an edge, which then may only be a stretch of an entire road, based on our analysis of that value.

Data: lane, AVG(lat, lng), COUNT(cars), grouped by lanes, ordered by count desc
Extract: Top 1000
After: Put top 1000 in map (tableau) and look if there's any recreational areas around

2. **Person density**

Car free days and car free areas are popular concepts in city planning. In order to utilize the city space we want to implement car free areas in Copenhagen. In the process of choosing a car free area, we want to create a view that enables us to identify areas with high density of pedestrians. This could also be combined with the density of cars, so the disturbance of car traffic is minimised.

Data: edge, average (lat, lng), count of persons, grouped by edges, ordered by count desc
Extract: Top 1000
After: Put top 1000 in map (tableau)

3. **Count number of times a car stops and starts on specific edges**

We want to optimise the flow of traffic through the city. This is desirable both because of increased efficiency for the commuting people and good for the environment, since the cars are spending less time on the road queuing. In order to establish where in the city there are bottlenecks in the traffic flow, we are establishing a view where we can calculate the starts and stops on a given road. Our hypothesis here is that fewer stops and starts equals a smoother flow of the traffic.

So what is a possible way to gauge the number of stops for several vehicles?

With a real time overview of the traffic flow, we could be able to send direction information to the smart cars in the grid and guiding them to take the optimal route through the city. With machine learning the algorithm can be optimised with historical data of driving patterns in order to optimise the route for the specific car and the specific driver.

Data: Count the number the cars stop and start again on lanes, ordered by average speed and number of cars
Extract: Top 1000

Question 3 (T): Master data set

The data is made available in XML.

A. **Describe the pros and cons of two different systems to store and manipulate this data (to answer this question you will need to make assumptions about the kind of processing required that are consistent with your answers for Question 2)**

1. Solution with plain mapreduce

Pros	Cons
● Very low-level in the sense that we have absolute control of the mapping and reducing logic and direct access to the HDFS. ● Automatically allow parallelization. ● Does not have to load in all data for processing, if the master data set is partitioned.	● Very low-level, because we have to take care of details that could be hidden from the use of frameworks (for instance setting up the files that are used in the job). ● Specifics of the storage of data has to be taken care of directly. ● If multiple MapReduce iterations must be used, the output from one reducer, must be written to HDFS, before it can be used from the next mapper.

2. Solution with Hive

Pros	Cons
● Tez provides pipelining ● Easy batch processing - Batch views can be written as HiveQL queries, which looks like SQL. ● Benefits from MapReduce ● Easy abstraction of data as structured data (high level abstraction)	● harder to integrate frameworks with each other ● harder to inject data, because there's more restrictions defined by the frameworks ● hard to understand the underlying processes, since its high level

B. Ingest the data into the system of your choice in a way that supports the definition of views defined in Question 2.

We decided on implementing the solution with Hive. To approach the difficult problem of ingesting 80GB of XML data, two things must be taken into account. First of all 80GB of data cannot be put into memory and probably neither into swap space, therefore we need to make sure that when we ingest the data, it is done in a manner which splits the input into chunks and handles them individually. This leads us to the second problem. The XML format is a tree format structure, and it is difficult to split into parts which can be handled by different processes, therefore a flatter format will do better. Because of this we decided to convert the XML data file to CSV which is a flat file format. Then we used Hive SerDe to import the CSV file into a table format in Hive. By using Hive we are able to express the batch view via an SQL interface, but still using map-reduce underneath such that we harness the power of that framework. We used the following command to create the tables in Hive and import the data:

```
CREATE TABLE IF NOT EXISTS sumo_data (timestep_time DOUBLE,
vehicle_id INT, vehicle_x DOUBLE, vehicle_y DOUBLE, vehicle_z
DOUBLE, vehicle_angle DOUBLE, vehicle_type VARCHAR(25),
vehicle_speed DOUBLE, vehicle_pos DOUBLE, vehicle_lane VARCHAR(25),
vehicle_slope DOUBLE, person_id INT, person_x DOUBLE, person_y
DOUBLE, person_z DOUBLE, person_angle DOUBLE, person_speed DOUBLE,
person_pos DOUBLE, person_edge VARCHAR(25), person_slope DOUBLE)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' WITH
SERDEPROPERTIES ( "separatorChar" = "\;", "quoteChar" = "'",
"escapeChar" = "\\" ) STORED AS TEXTFILE;
```

Creating the table with SerDe import options

```
LOAD DATA LOCAL INPATH 'project3/FCDOutput.csv' INTO TABLE
sumo_data;
```

Loading data into the created table

This command turned out to import all columns as strings when using SerDe to import the CSV, although we did create a typed table beforehand. We could probably achieve a noticeable speedup by making the column types correct, and splitting the input into two tables, one for persons and one for vehicles, or by creating indexes on the most used columns.

It should be noted that we were not successful in ingesting the data in such a way that view 3 is easily computable (though it is possible), but as we will see in the next section an approximation of it is still possible.

C. Implement the derivation processes that produce the views defined in Question 2

For computing the batch views we used Hive SQL scripts, which are stated below.

View 1: **Vehicle density**

```
SELECT COUNT(DISTINCT vehicle_id) AS vehicle_count,
       AVG(CAST(vehicle_x AS DOUBLE)),
       AVG(CAST(vehicle_y AS DOUBLE)),
       vehicle_lane
FROM   sumo_data
WHERE  vehicle_id IS NOT NULL
GROUP BY vehicle_lane;
```

This query calculates the vehicle counts on different lanes (which can be mapped to actual roads if wanted) over the entire data set (2 hours). Furthermore we include an average of the latitude and longitude (vehicle_x and vehicle_y) of the cars on that lane. This information is only included to have a rough idea of where the lane is located, without merging the data with map information. The reason for the cast to double is because the data is stored as strings as written in the previous section. A snippet of the results can be seen below.

vehicle_count	vehicle_x	vehicle_y	vehicle_lane
648	12.48229964	55.72693068	122492607#3_0
557	12.49144177	55.72215528	119507350_0
555	12.54158527	55.66268498	27409296#1_0
531	12.5412276	55.66199099	26482048#0_1
528	12.54150479	55.66246743	:10437896_1_0
523	12.51613423	55.6636757	25912891#2_0

This table for example shows us that lane *10437896_1_0* has *528* distinct cars in total over the two hours. We can also see that the lane with the most traffic is the first with about 20% more cars than the other lanes. It is clear that given this view a number of interesting queries can be made.

View 2: **Person density**

```
SELECT COUNT(DISTINCT person_id AS person_count,
       AVG(CAST(person_x AS DOUBLE)),
       AVG(CAST(person_y AS DOUBLE)),
       person_edge
FROM   sumo_data
WHERE  person_id IS NOT NULL
GROUP BY person_lane;
```

This query is almost identical to the one above. It just calculates number of people on edges, rather than vehicles on lanes.

person_count	person_x	person_y	person_edge
3208	12.58723354	55.67475359	-1694109
2970	12.58780263	55.67437389	-322164666
2903	12.58865936	55.67388881	-1694110#1
2876	12.57654324	55.67056482	225964063#0
2863	12.51608017	55.66418092	115678974#0
2861	12.51618019	55.66432078	-78412354#2

The table shows us that on edge *-1694109* there have been a total of *3208* people, and that edge is the edge with the most people in the table above. Just like the table above it is clear that given this view a number of interesting queries can be made.

View 3: **Count number of times a car stops and starts on specific edges**

```
SELECT AVG(vehicle_speed) AS avg_speed,
       COUNT(DISTINCT vehicle_id) AS vehicle_count,
       AVG(CAST(vehicle_x AS DOUBLE)),
       AVG(CAST(vehicle_y AS DOUBLE)),
       vehicle_lane
FROM   sumo_data
WHERE  vehicle_id IS NOT NULL
GROUP BY vehicle_lane
ORDER BY vehicle_count DESC, avg_speed ASC;
```

This query calculates the average vehicle speed per lane. As the other queries it includes an estimate of the center of the lane. This view should have counted the number of times a car stop on a given lane, but this is hard to express in HiveQL if we do not want to include the same car multiple times, if it is stopped for more than one second. This table might not be interesting in itself, but if we match this information with data about speed limits for the given lane (or road), then we can see roads where the majority of cars is driving at a much slower speed than the limit, which might indicate a problem on the road.

avg_speed	vehicle_count	vehicle_x	vehicle_y	vehicle_lane
1.735935597	648	12.48229964	55.72693068	122492607#3_0
0.7336198794	557	12.49144177	55.72215528	119507350_0
2.339899297	555	12.54158527	55.66268498	27409296#1_0
13.08320411	531	12.5412276	55.66199099	26482048#0_1
7.8700208	528	12.54150479	55.66246743	:10437896_1_0
1.25324843	523	12.51613423	55.6636757	25912891#2_0

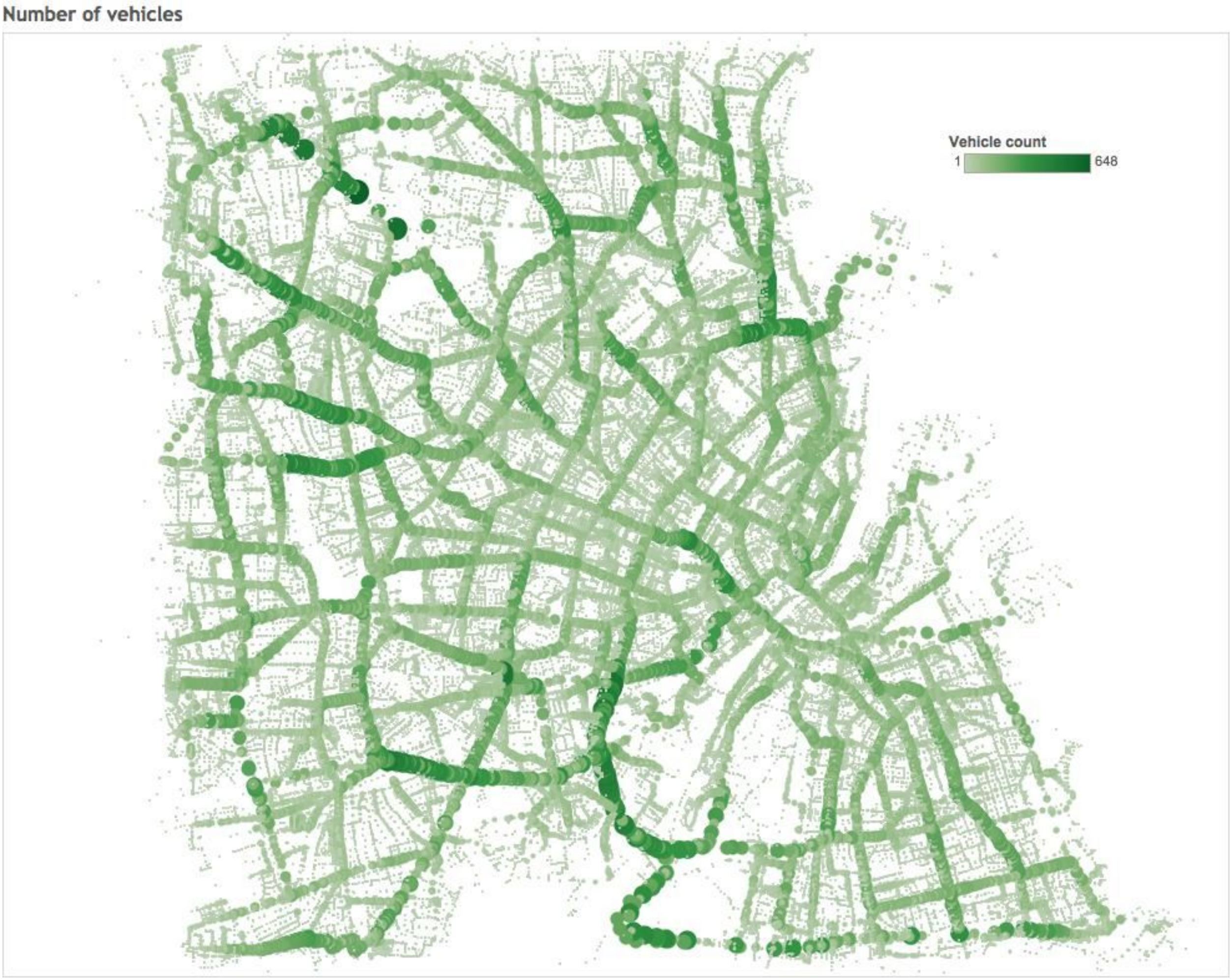
The table above shows that the average speed of vehicle lane 119507350_0 is 0.734, which is by far the lowest value of the lanes shown above. This might indicate that the vehicles are stopping often, for example due to traffic or intersections.

Interestingly enough, the originally planned batch view would have been more easily expressed in pure Map-Reduce with java since it requires some fine manipulation to find consecutive speeds of zero for the same cars and mapping them to one entity.

Question 4 (C/T): Data processing

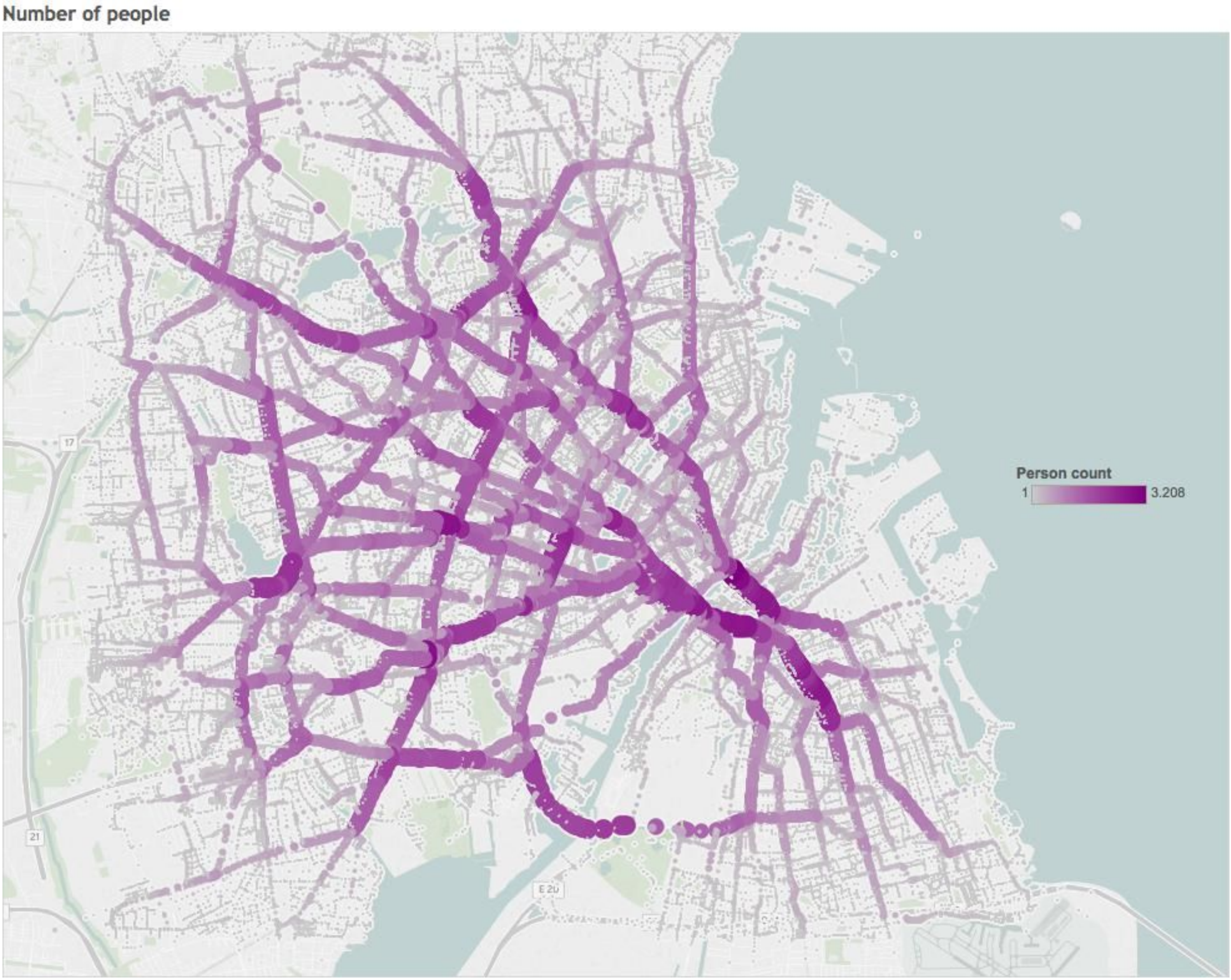
Car density

With the number of cars on each lane, we can visualize where in the city the car density is highest and lowest. The visualization of number of vehicles below is made without a background map, but you are still able to see the outline of the city roads. As expected the main gateways to the city carry more cars, but you can also use the map to investigate which of the smaller roads have unexpected high usage.



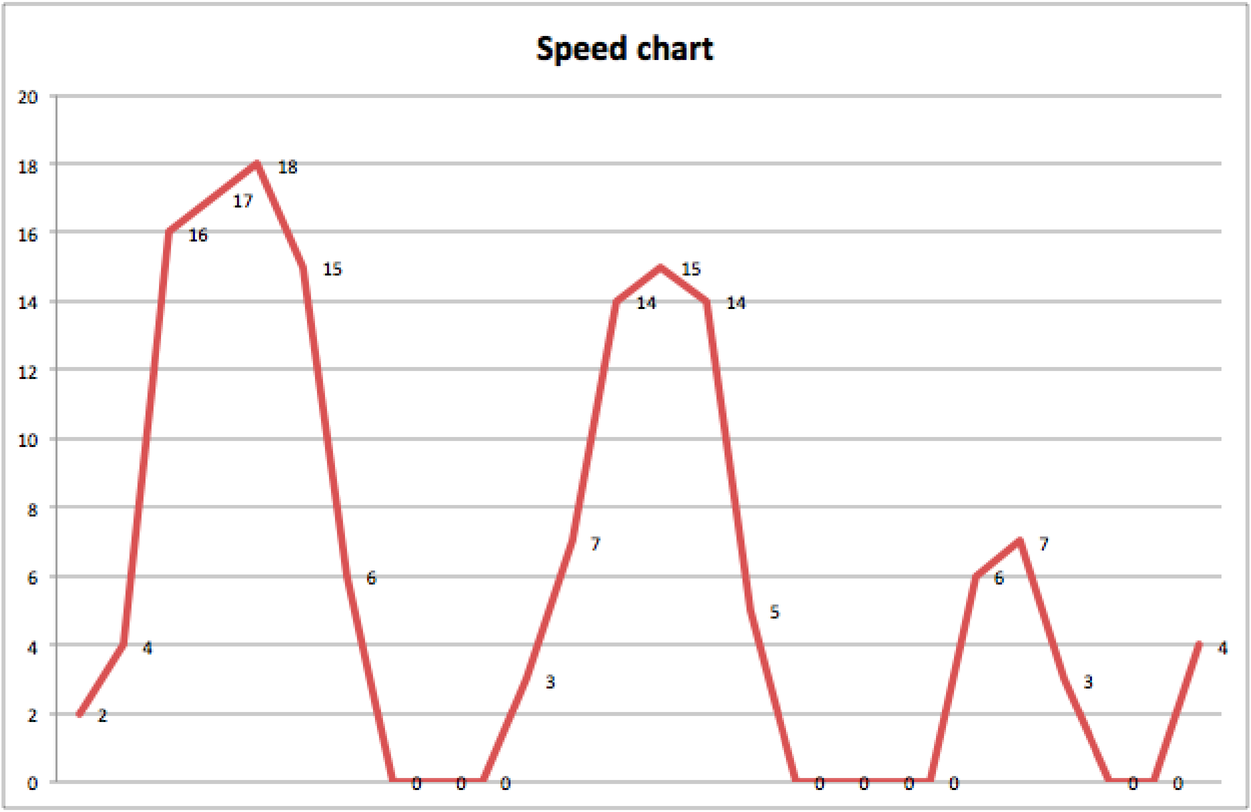
Person density

Combined with the previous vehicle count map, we can use the person count map to see where we could potentially implement car free zones. If we identify an area with many cars, the implications of making this area car free will be too immense and disruptive for the traffic flow. On the other hand, if there are no people in the area, not many would be reaping the benefits of the car free zone, which makes the intervention redundant.



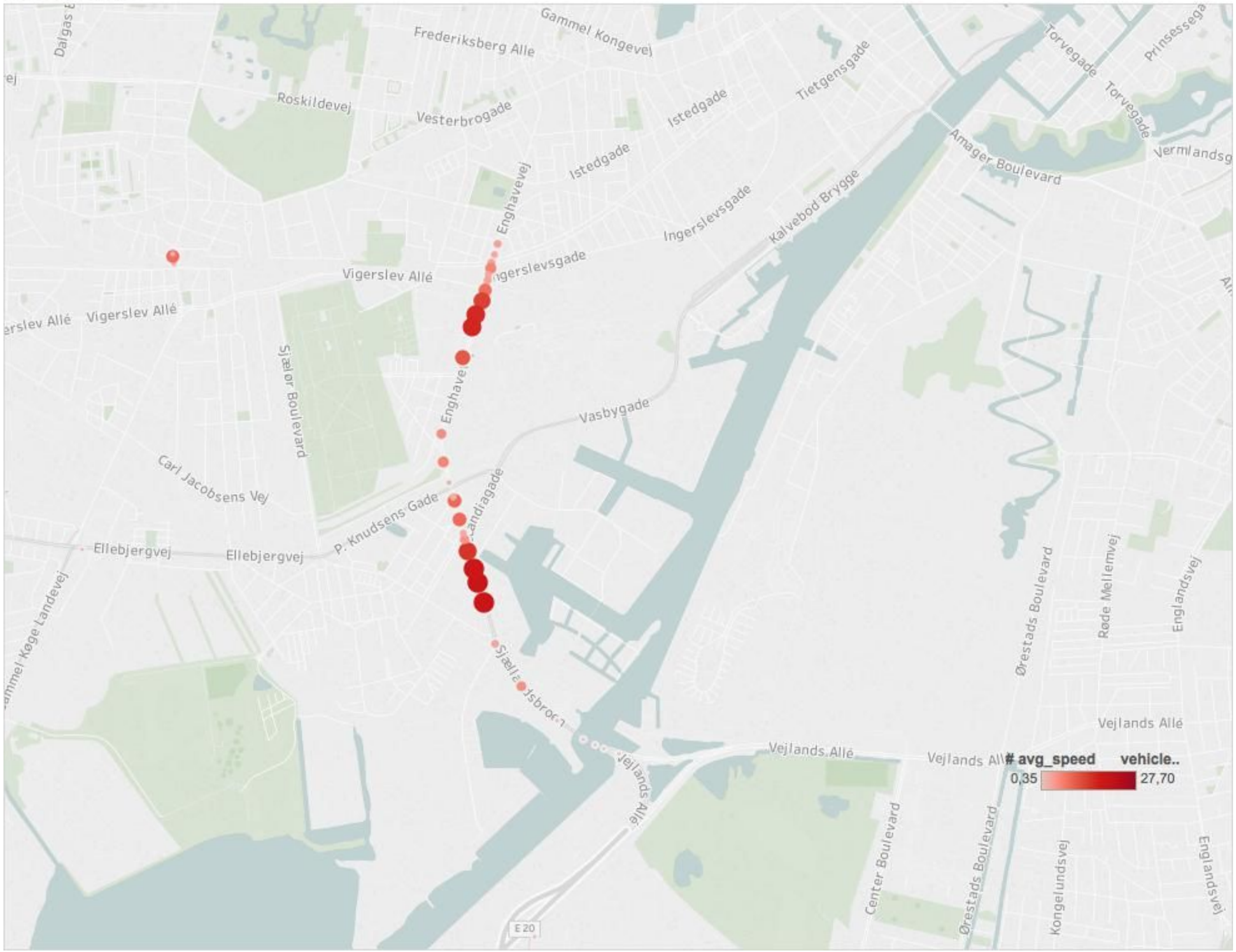
Count number of times a car stops and starts on specific edges

In order to count the number of stops and starts, we need to establish a way to define a stop. On the 'Speed chart' figure we have exemplified data from a vehicle. On the y-axis we have speed and then we have time on the x-axis. When we see the data like this, we can see that there are three events where the vehicle is stopping, but there are nine events of the speed being zero. We can therefore not just count the number of zeroes, but we have to add the condition that if the previous speed were zero, then the event is not a stop, and if the previous event is not zero, then the event will be defined as a stop.



In our final iteration of the data processing we did not manage to produce the stop/start count. We did manage to extract the average speed on the top 100 lanes in the dataset. This gives us an idea of how we can use the data. In the average speed visualization, we can see how the speed distribution is on a given road. If this view could be converted to a real time visualization, we would be able to analyse events that reduces speed.

Average speed



Question 5 (C/T): Log

Critical:

- 08-11-16: We have chosen a scenario (1), and can start defining mobility tracking etc. and the focus of our committee in developing this plan (i.e. if we want to be greener, smarter, more efficient). **But to choose points of action** we need to look at the data which is still being downloaded.

- 08-11-16: How to go from source data to **data we can use in visualization programs** (we discuss Tableau). This is for the purpose of storytelling in the pitch/proposal (Question 6).

- 08-11-16: We start on C questions while T discuss **how to store and handle the data**.

- 17-11-16: Defining views within the “green vision”--maybe expanding to a smart city discourse. Batch views that are the same for T, we see as different stories still.

- 17-11-16: Discussing with T what is possible in terms of batch views correlated with what might also be interesting, given that we have to frame, storytell and pitch. Given the attributes in the data, how can we combine them, and potentially outside sources and data, to do something interesting.

- 17-11-16: T are experiencing a bottleneck with some of the tools and platforms they have to use, which means we also stall at some points, or may come to. But we have defined our views into the data and we, C, will now work on articulating them further.

- 17-11-16: There will be some sort of “data handover” on Thursday next week, so we can start to clean data for visualization purposes.

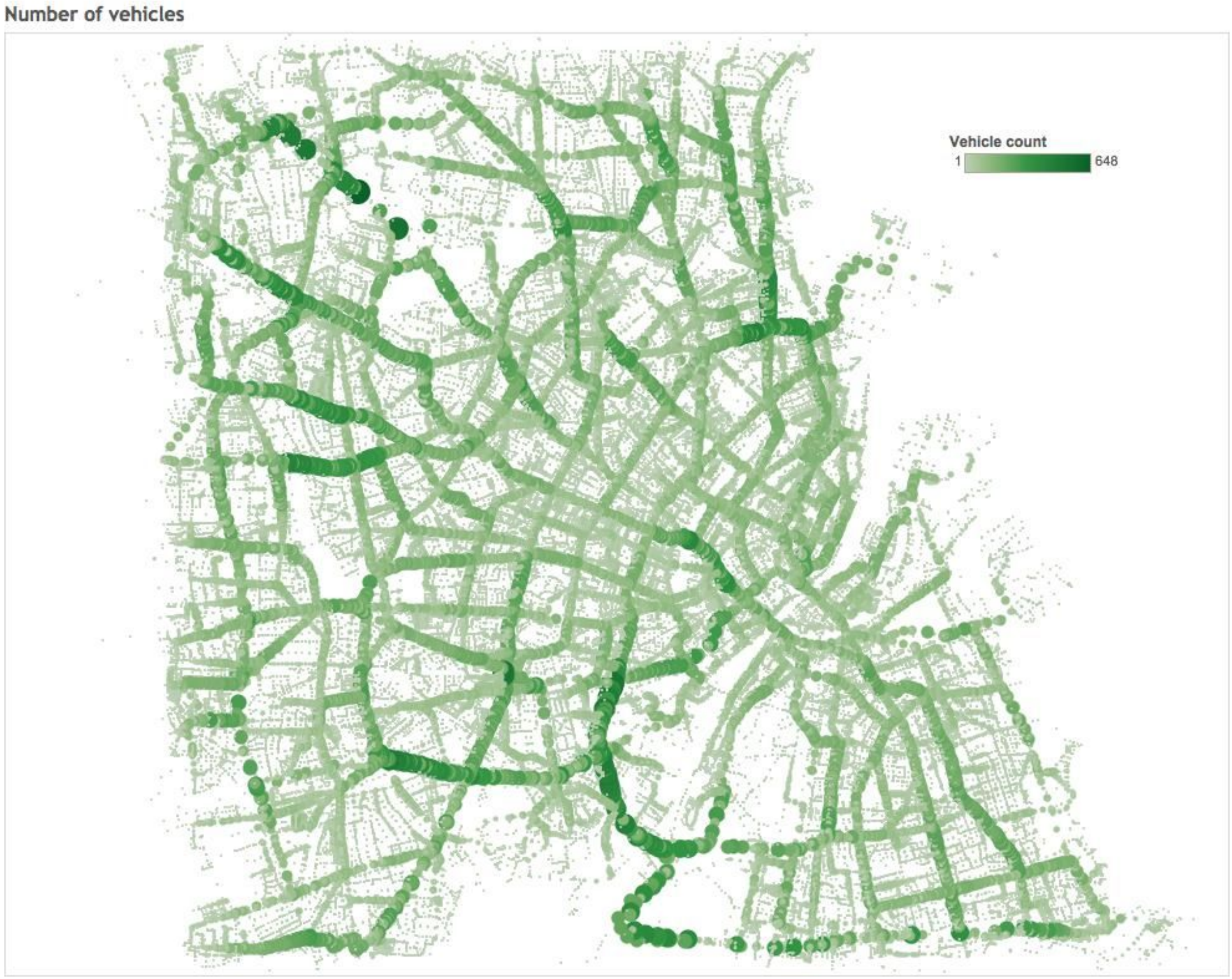
- 22-11-16: We work separate from T today, because we have managed to arrange other times for meeting this week. They work on the batch views, while we work on framing the views. We agree to be “less critical” in rhetoric when storytelling and proposing/pitching, but make sure to note the critical reflections for the designated paragraph. We hope to have data available for visualization later this week.

- 29-11-16: “Getting data” for visualization can be something entirely different than “getting it” and using it for analysis in other ways (as we have done in previous projects). Data for visualization in a program like Tableau means a lot of cleaning data, is our assumption. This creates a kind of missing link in the workflow, which is then not exclusive to either T or C.

Technical:

- The entire cluster had a lot of down time (including the host of the virtual machines).
- The different frameworks on the cluster, that we were using were down on the server.
- After importing the csv to Hive with SerDe, we realised that SerDe overwrites every data type to String.
- We first tried to compute the XML to CSV in memory, because of wrong instructions on how the xml2csv-tool worked, but obviously it didn't work.
- Then we tried using the CSV Philippe had made, but we were missing some columns for our views.
- We had to adapt the XML schema to get the columns we needed. For instance, in the schema-file that was referenced from the data-xml-file, there was no representation of persons at all.
- When preparing for this course another time, please be sure to have space enough for all groups. We know that some of it is for illustrative purposes, but it does not make sense to block out groups from doing projects, because other groups are filling up the entire drives.
- It is quite problematic that we only have a single entry point to the HortonWorks cluster. If one group are running a job on the server (not the cluster), then it is impractical for other groups to do anything on that server.
- It is practically impossible to do any work on the cluster during the exercise lessons, so instead we have opted for working in the weekends sometimes, or use our own laptops for computation, which was not the intent for this course. It is also hard to coordinate, because we have to work, when no one else is using the cluster, and it is hard to know when that's the case before trying.
- We tried to import the CSV to Hive in one of the exercise lessons, but we had to wait to import it until the following Sunday, because the cluster and/or Hadoop, YARN, Tez, HDFS, Hive or similar was down untill then.

Question 6 (C/T): Selling it



The visualization above makes city mobility visible in one way. It focuses on a crucial part of how our citizens move through our city, with the use of cars. Representing the city in this way gives us the ability to make sense of and intervene in traffic management.

Cars making multiple stops on the roads is bad for both the flow of vehicular traffic and air quality in our city. In that way it is bad for all residents of our city. Wanting to improve mobility with a green view in mind needs special attention focused on minimizing the stopping. We need to stop stopping and start driving. The cars that move through our city are increasingly equipped with networking capabilities, which provides a good foundation of existing infrastructure that just needs to be integrated with the correct data practices and analyses.

From data to algorithms

To give you an idea of how data harnessed from pervasive data-producing technologies can be used towards the goal of stop stopping let us turn our attention towards the simulation dataset. Scaling such a dataset provides the means for analysis that can determine the amount of stops cars have on certain roads. But what is the use of such a metric? We argue that knowing where stopping is most prevalent should be integrated in traffic optimization algorithms. We would use these algorithms to direct the movement of vehicles throughout the city, recommending alternate routes of making it from A to B. This would happen in real-time given we have the means of implementing a speed layer into our existing data capturing infrastructure.

From algorithms to drivers

The envisioned scenario when the algorithms and the infrastructure are in place is that we have a automated system that simultaneously analyse traffic flow while directing vehicles by the optimal route through the city. The system will be able to learn from previous data in order to optimize its recommendations. This combination of Big Data and machine learning is bringing our city infrastructure into the future of transportation.

Benefits for pedestrians

We have not forgotten about citizens who choose to walk our streets by foot. There are certainly more aspects to managing city space than the flow of vehicular traffic. Our plans to optimize the flow of cars on the roads has positive effects for foot pedestrians as well, because minimizing stopping is at the same time a way to minimize emissions so that air quality becomes better. We are not only concerned with bettering the roads we travel, but also the air we breathe.

Some challenges

Some of the additions to the current big data processing infrastructure are arguably very costly. The best use of data for traffic flow management in the form that we propose requires a speed layer that gives access to data real-time. There is no doubt that we need to have a system of data governance in place. Ethical concerns includes not getting consent to capture data for learning about the amounts of, but also for the aspect of the system that aims to communicate with drivers of cars. If we want to push information about alternate routes to drivers. In setting up practice for getting consent, we can anticipate an unravelling effect which needs to be addressed. We need support in the form of policies that continue to incentivize buying cars that have the necessary technological capabilities to provide us data as well as receive our notifications of redirection. We do need to be aware of how changed mobility can affect other stakeholders than just drivers. Certain areas, such as residential ones, may be planned and commodified in ways that rely on the roads being less quiet, polluted and trafficked. This is an environmental impact that warrants concern.

If we are able to critically reflect on the above we have no doubt that the system can be designed to best alleviate the concerns while having a positive impact on the flow of traffic in our city.

Critical reflection

Our proposal is arguably very focused on increasing mobility of traffic for vehicles. Although we argue for some benefits for pedestrians, we can question whether there is too heavy a focus on benefits for those driving cars. There is also the concern that we use this data to design a system with which we aim to make better use of less trafficked roads to relieve the most trafficked ones. But in doing so we should take into account that the areas surrounding the former will be impacted by the new flow of traffic. Some citizens may have taken up residency in areas with less trafficked roads precisely because they are quieter and cleaner. The environmental impact, although briefly addressed in the above, is no small matter. It suggests that traffic management should happen closely with urban planning.

In our proposal we argued for pushing transportation policy towards increasing the number of cars that can take part in the improved mobility infrastructure. There might be social drawbacks to this, because it requires that citizens are able to spend the money needed to either equip their cars with the necessary technology or buy cars that are already equipped. Regarding consent, we draw attention to this in our proposal. But the fact remains, that there is a real chance of marginalizing those who do not give consent to have their data collected, stored and used. It takes the form of an unraveling effect (Peppet 2011) that makes those that agree to give consent viewed as in line with the interests of policy-makers, while does who do not wish to participate in the data handover find themselves underrepresented.

Designing and implementing a system where an algorithm is tasked with directing the flows of a city means we are introducing a new knowledge logic (Gillespie 2014) into how mobility is practiced within the city. This logic relies on classifying certain instances of data on traffic and making decisions upon those classifications means that other data and instances of traffic are omitted (Bowker 2005). This is effectively a way of simplifying a very complex reality of mobility in our city.

In terms of how we reach out to potential investors and adopters of our proposed system, we use the aid of visually representing data. *“Visualization can allow humans to interface with and make sense of a large amount of ever-changing big data”* (Kim et al 2013, p. 5.) We wanted the audience or intended readers of our proposal to think of mobility in a way that blends well with how mobility data could be visualised, by the use of maps.

References

Bowker, G. Introduction - from Memory Practices in the Sciences (2005) MIT Press

Kim, J., Lund, A., & Dombrowski, C. (2013). Telling the story in big data. interactions, 20(3), 48-51.

Gillespie, T. (2014). The Relevance of Algorithms. in Gillespie, T., Boczkowski, P. and Foot, K. (Eds) Media technologies: Essays on communication, materiality, and society, Cambridge, MA: MIT Press

Peppet, S. R. (2011). Unraveling privacy: The personal prospectus and the threat of a full-disclosure future. Nw. UL Rev., 105, 1153.