Big data - Project 1 (Group 17)

Authors: Søren Harrison, Jakob Reinbach Fauerskov and Jan Vium Enghoff

Stage 1

Copenhagen:

1. Data available include traffic measurements, health and health care, job and integration. The data are available as Excel-sheets, PDFs, and CSV-files as well as geo-data for a few files. The data are mainly represented as statistics – how many do A and how many do B and the analyses that may be conducted on these are mainly correlations and further statistics.
2. Data missing are among others education, safety, business. Private sector type information is not available at all. Nothing about culture.
3. We have for instance average traffic on each street distributed between bikes, cars, and trucks. We could use the number of trucks and bikes and correlate to the number of (right-turn) accidents (if available?) on the busiest roads, for instance Knippelsbro which has 35000 cyclists every day on average and about 1500 trucks (24700, 6.2 % trucks). Perhaps even correlate this with the number of traffic signs on Knippelsbro or speed. Problem is to know what the abbreviations mean…
4. They may be used to see if there is a problem at a given location or not and that could lead to a specialized action toward that place in particular. Doing this on a larger scale might provide insights to troublesome places in Copenhagen. The problem with the date from Copenhagen is that it is relatively new, so there is not enough to show a tendency.

London:

1. Traffic measurements, health and health care, education, sports, culture, crime, planning, environment, demographics. Has almost everything. The data are available as Excel, CSV, PDF, Maps, XML, HTML and several other formats. The data are mainly represented as statistics – how many do A and how many do B and the analyses that may be conducted on these are mainly correlations and further statistics.
   Data are in two forms, namely reports and as raw data. Many of the reports are in a human readable form, but are hard to process as data for data mining. Some of the raw data is available through the sites API, but mostly the API serves metadata for looking up the available resources. Analysis on the data in the London data store would mostly be on the reports, more than on the data that it derived from. The raw data that are available are historical data on groups, with limited availability to real-time data.

2. Mainly public sector information is there so private business information is missing. The data is mostly grouped by time and London city districts, which means that finer grained analysis might not be possible to perform. This includes behavior of single individuals.
3. Could run the same analysis on data here as with Copenhagen.
4. Could run the same analysis on data here as with Copenhagen.

New York City:

1. Business, city government, education, environment, health, housing & development, public safety, recreation, social services, transportation, NYC Big/Apps. The data available mostly isn’t human readable, but it seems to be very computer friendly, offering a wide variety of export formats. The data is very varied, jumping from being quite detailed about air pollution levels at a specific point in the city, to water complaints in a general area, or what seems to be flooding plans. The data is also available through their API making it easy to data crawl and build up a huge dataset for later use.
2. Data about citizen health is not quite as clear as it was in the Copenhagen dataset (e.g. explicitly counting runners at certain areas). The scope of the data is very different, and usually not available outside of that scope. This means that the data might be scoped in an unfortunate way for the particular type of analysis one might wish to perform.
3. It is possible to answer many of the same questions as with the other cities. We could, for example, use the data about air quality to perform analysis about pollution from vehicles at different times of the day and year, which might tell us a - quite coarse grained - tale about how how the traffic develops in various time frames, and which times of year the traffic is the most dense.
4. They might be useful, depending on whether the data needed is available at the right scope for what one needs in the particular situation.

Stage 2

In general, a data-driven city is capable of making significantly more informed decisions, by asking various questions of their available datasets.
Questions and decisions based on the data:

- How much pollution is present at certain roads? How many runners are on these roads?
  - Should something be done about the amount of pollution, can it be more evenly distributed so it's not a health hazard to run here.
- How is traffic distributed at certain times of day?
  - Would it make sense to redirect traffic on roads with heavy traffic to other roads at certain times of day?
- How about parking at certain times of day?
  - Should some road be taken as parking zones or are there too many parking zones?
- How many people are using the parks?
  - Should they be moved, replaced, expanded?

The kind of data we have here is not real-time which means that analysis can only be based on historical data.
Possible questions during decision making:

- How should the infrastructure of the major cities be structured to optimize traffic throughput.
  - Does road work and temporary closing of roads affect the neighboring area in a positive way? Pollution due to longer routes? Discomfort for the locals? More visitors to the shops around the area due to more exposure from the roads? Answers to these questions and more could help making a decision in improving the city's infrastructure.
- Citizens want to plan their running route starting from their home. The data could answer the questions:
  - How do the citizen run in the most green areas?
  - How do the citizen avoid the heavily trafficked streets?
- The municipality want to build a new car park. The data could answer the questions:
  - Where are the existing car parks most at their capacity?
  - How will placing a car park affect the surrounding buildings.
  - How are the ground around the city? What areas are best suited for placing the new car park?

Of course, issues will arise in situations where a need for datasets that haven't been created yet, lacking the necessary data to perform the analysis. This issue is of course compounded by the fact that having sufficient statistical data to perform a decent analysis will take a while, meaning that from the date that you figure out that you need some arbitrary dataset that you're lacking, and waiting for that dataset to be sufficiently mature to perform data mining upon it, might take a while. No matter how many, and how varied, the datasets are for a given city, they will probably always figure out that there’s some obscure dataset that they’d like to look at to figure out the solution to some problem, only to realise that it isn’t available. A lot of questions requiring more fine grained data cannot be answered. An example could be:

- The traffic light are not updated in real time which makes it hard for businesses to provide a service for calculating the fastest route through the city. This makes cars break a lot and is damaging for the environment. A dataset with real-time data for traffic lights would make it possible to provide such a service to the drivers. The reason these aren’t available are most likely due to the technical limitations as well as the demand of such a dataset.