Project 2: Group 4 + 40

Master set data - Technical

A. How do you store this master data set? Explain your answer.

In order to store the master data set, it was necessary to examine and define the data. We categorized data from the metadata file as either Wi-Fi clients or access points and locations, and data in the time series files, is specific readings of either of those types of measurements. There exist multiple other types of metadata but we have chosen to ignore either by scope or because the data is incomplete - more on that in the next section. We use the serialization framework Apache Avro[1] in order to define schemas and (de)serialize these objects. We defined an Avro-schema both for the provided data files and for data model, such that cleaning could be done efficiently and dynamically.

Parsing the data is conducted using the Google GSON[2] library with the defined Avro schemas and serializing the parsed, cleaned and transformed data into files of serialized Avro objects. By creating these schemes it makes it easier to perform analysis as it is possible to have access to related data easily.

We wanted to store data in Hive[3] tables using Kite SDK[4] on top of Avro, but because of difficulties with other parts of the exercise we chose to postpone this implementation, and eventually, we did not reach a point with time enough left to implement this. Some of our choices with regards to splits of metadata still looks like it is meant to be placed in tables though.

The resulting metadata contains roughly the same information as in the provided data files. We did parse the path-property into multiple properties though, and remove some of the data that are unused for our batch views.

Furthermore the schema contains a definition of the readings that contains the UUID and a collection of Reading-objects that each contain a timestamp and a value, rather than being a list of numbers with length 2.

The schema of the master data set has been drawn in Figure 1 below. Note that AccessPointMeasure and WiFiClientMeasure are enumerations where each value is in the comma separated list.

[1] http://avro.apache.org/
[2] https://github.com/google/gson
[3] https://hive.apache.org/
[4] http://kitesdk.org/

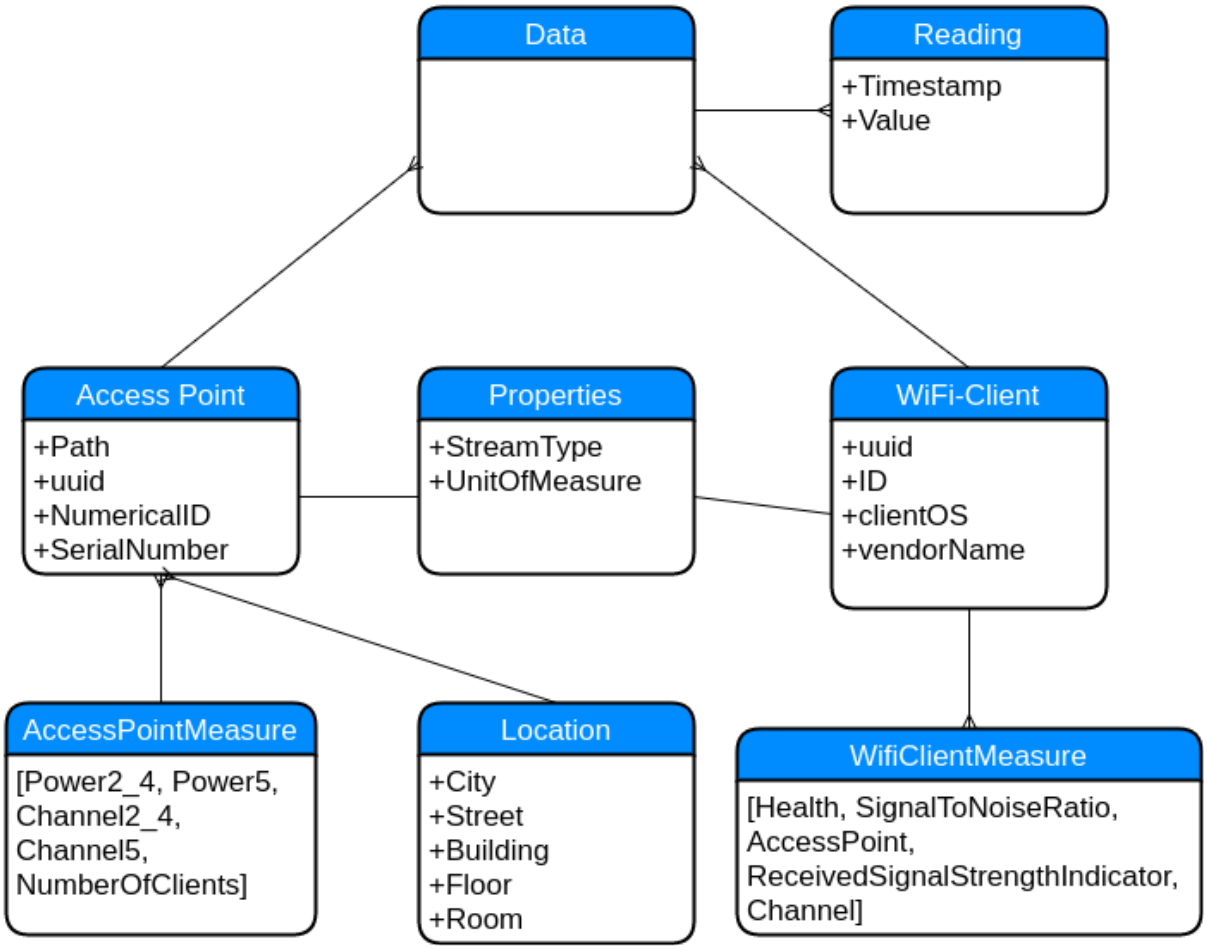


Figure 1 - The master dataset model

Finally Avro supports serializing to binary files, so the data is stored in 4 different binary files, with each type of data. It would have been useful to have a partitioning framework to be able to partition the data for each day, but this was not done in time for the project handin.

B. What is the sampling interval of the data? Are there missing data in the dataset?

The sampling interval of access point information is approximately one sample every four minutes. This calculation is only made on a subset of data, by applying the following formula:

$$sampling\ interval = \frac{time\ of\ latest\ reading - time\ of\ earliest\ reading}{\frac{number\ of\ readings\ for\ access\ points}{number\ of\ access\ points}}$$

As mentioned in the first section we ignore certain types of data. In the data files information concerning student course base, calendar bookings, thermostat readings, is available. In the case of the thermostat readings, the data is incomplete, since it is only present for one room. In the imported data, we also have missing information, for example is the information about the vendor name and the OS name for Wi-Fi clients often unknown. Furthermore, we have chosen not to include the student course base and calendar bookings, not because data is missing but because we see it as out of scope for the analysis we want to perform.

C. **Define a procedure to clean the data set and handle the missing data. Give arguments for your approach.**

There are multiple approaches on how to handle missing data. We have chosen to both remove, by not including data, and ignore certain types of missing data. When we read the raw data and transform it to our master data format, we do not include any entities not related to Wi-Fi clients or access points. This is done for both the metadata file and the readings files. This effectively removes information about the student course base, thermostat readings and calendar bookings.

For the other type of missing information, we simply ignore it when performing the analysis. By keeping data concerning these entities but ignoring them it is possible to perform analysis on the rest of the entity even if this specific information is missing. This approach has the obvious downside by requiring more work and thought to complete analysis.

D. **Generate a clean data set.**

Do you use Hadoop for this batch process? Explain your answer.

We run the cleaning of data as a simple java program, which does not use any Hadoop[5] features. As of now running the program, and moving the data is a manual task but it would be possible to change the program to use Hadoop and MapReduce, which would make it possible to automate the process at some point. This is not done, due to time pressure and failed tries to do so.

That said, at the moment our solution is not prepared for performant data ingestions and data partitioning. In the future we are planning to use Kite together with Apache Crunch[6] to import, clean, partition and ingest data into our master data set. This solution would utilize all the features offered by Hadoop.

How many instances of missing data did you find?

We do not have a specific number of how much data was missing but as mentioned before, thermostat readings were only available for one room, therefore implying that data was missing from all the other rooms at ITU. Furthermore, there is a lot of missing data regarding vendor and OS name used by Wi-Fi clients.

Personal data - Critical

A. **What are these data about? What is known/not known about Wi-Fi use in this dataset? What is made obvious/visible? What is overlooked?**

The data are about Wi-Fi use at the IT University of Copenhagen (ITU). The nature of the data is continuous and not real-time. It comes in batches every day making us able to say it

[5] http://hadoop.apache.org/
[6] https://crunch.apache.org/

is about the Wi-Fi use over the course of 7 days, i.e. the daily Wi-Fi use at the university. We employ, the term ‘use’ is in a broad sense in this report so far. From a technical viewpoint, the data tells us that activity is occurring at the hardware level, a kind of Wi-Fi infrastructure at the university (based on metadata on the Wi-Fi Clients as identified in the technical investigation). Devices are connecting to the Wi-Fi access points throughout the building, but we do not know for what purpose they are doing so. We can for instance pinpoint at the room level where activity on the access points is occurring and many other characteristics of this so-called use. The context, however, is unknown--or made invisible. The visible is that these data can say something about the behaviour of the people at ITU as a collective. What is more hidden is the fact that with the right analyses, we can begin to track and identify individuals.

B. What kinds of stories can these data tell about people at the ITU (what can these data reveal about individuals if anything)?

Lack of context brings us to the question of the stories the data can tell us as well as what privacy concerns may be tied into these. Stories have main characters, or protagonists, whose actions and behavior tie seamlessly into the progression of the story. From data about Wi-Fi use at ITU we are arguably not able to know much about the individuals, whereby the stories we can tell become quite flat. We also cannot know what the Wi-Fi is enabling the users to do, that is, what they are using it for, for example the websites they are browsing or the files they are down or uploading. The ‘who’ remains at the aggregate level, and the ‘what’ entirely unknown. So asking what stories we can tell is relevant, but the answer would most likely be: not very compelling ones. This can can be argued if you just take the data at face value. However, if we apply a critical view, we may be able to tell the compelling stories: stories that highlight how this data can be used for activities that puts assumptions about privacy, identifiability and ethics at stake. We will do so later in this project paper.

C. What can these data reveal about all occupants at ITU in general? Can you say anything about things other than the devices connecting to Wi-Fi access points and the locations of these access points in the building?

The data can reveal frequency of use. We may see spikes in activity on the access points and Wi-Fi Clients during weekdays where the ITU would be host to more people than in the weekends. The activity may also be spread out throughout the building to account for the heterogenous sites of activity (students as well as faculty). On weekdays more rooms will be in use and thus more access points will be connected to from various devices. On afternoons in the weekdays, the usage may be isolated to a few rooms and locations throughout the building with students undertaking study activities such as group work in single rooms. Outliers in activity can occur from the ITU hosting conferences and other such events, and in the same vein, Friday nights (until 2 AM) may also see a spike in the area around Scroll Bar. The argument can be made, that this information puts privacy at stake to some extent, as the concept of privacy exists and is negotiated in many different ways in theory. A definition of privacy can be the individual’s right “to control, edit, manage, and delete information about them[selves] and decide when, how, and to what extent information is communicated to others” (Westin 1967). As such, privacy functions as a timeout from the

practices taking place around an individual at ITU, which this individual feels is compromising this right. To better illustrate the potential of this dataset as it is subjected to more and various analyses and tracking, we will define three views or scenarios in which tracking can have implications of different sorts.

Batch layer - Technical/Critical

A. **Define three views that can be used to get insights about this data set**.

View 1: How can Wi-Fi data be used for purposes of evaluation of teaching staff?

With this view we are implicitly concerned with how use of the Wi-Fi at the room level can say anything about the turnout of students for a particular teaching activity. Our hypothesis is that the number of registered participants should match the number of Wi-Fi clients connected to the access points in the room of the lecture. If the access point in the room has a number of Wi-Fi clients connected which does not match the expected number, this might indicate that students do not show up and in turn indicate if course is popular or not. More specifically, connecting to the access point could indicate how many users connected their laptops, presumably for note-taking and participation of other activities in the course. Tracking the data over time, could be a way to see if fewer students are attending throughout a semester.

Attendance may in some evaluations be seen as indicators of the success of a course. This assumes that a successful course draws a lot of students to the lectures consistently, and therefore this indicator ultimately reflects on the quality of the teaching staff affiliated with the course.

There are a lot of assumptions which must be made to conclude this. First of all we assume that a Wi-Fi client is one attendance. Furthermore we assume that everyone attending is using the Wi-Fi, which might not be the case for all courses. It also assumes that every room the course has is used only by students of that course. Even though some of these assumptions seems like a stretch one could take some of these problems into account and use it in collaboration with the qualitative analysis of the course evaluation.

View 2: How can Wi-Fi data be used for tracking an individual at ITU?

We have a hypothesis that information about what access points a client has been connected to makes it possible to track a single person at ITU. Since each Wi-Fi client has an unique ID in the data set, tracking that ID around the ITU through various access points in certain rooms, one could connect this information to teaching activities. One could essentially build a schedule corresponding to a person, the holder of the unique ID, and by cross referencing the public course base, identify any student or teacher.

It is necessary to assume that every person is connected to Wi-Fi whenever they are at ITU and even more important that they are connected to the access points in the rooms that they have courses in.

Interestingly it should be noted that if a person stalked another person around ITU, writing down where and when a person was there with his/her device, it would be possible to identify the ID of that person with the current data alone. As soon as a person has identified the ID to a person, it is possible to track that person for the rest of the lifetime of that device. The fact that this is possible is not visible directly by looking at the data, but it raises interesting ethical questions about the data. One could imagine ITU tracking a specific student to make sure that they participate in the courses they are registered in.

View 3: How can Wi-Fi data be leveraged as a commodity by ITU to sell to third parties with commercial interests?

The data can reveal details about the devices connected to the access points. In a situation where ITU would wish to monetize the data in question, it could be sold to third parties with commercial interests so that these may target advertising towards people using particular brands of software. Through analysis where the device is correlated with the unique Wi-Fi Client ID and the courses the holder of this ID attends, a commercial third party may be able to pinpoint and target the creative Design and Communication student that uses Apple's Macbook Pro for completing design projects. If this individual has strong brand loyalty, advertisers can target ads for other Apple products towards them, but the advertiser could also choose to base their marketing on other factors of the segmentation of software users of a given course (by referencing the course base). The critical question one should raise in this regards is "If I am exposed to this content, what content will/am I then then not exposed to?" (Turow 2011). The mediation of communication we as individuals are exposed to, in the example ads, determines how our perception of the world is shaped. However the notion of the *world being shaped* is critical since it encourage certain behaviours, leaving the freedom of will to being no more than an illusion, played out by internalized practice by marketers.

If the hypothesis is true, then the buyers of the information would be the ones who had to make assumptions. Of course there would still be assumptions about Wi-Fi clients being people, and people taking specific courses and so on, to be able to sell the data. The point here being: The moment the data shift hands, it also shifts domains, thus the domain one works in also determines what questions one should and can ask. In this instance, the data goes from an ITU domain to a marketing/advertising domain,however the assumptions and thereby also the disciplines that are build up around each domain will be different from each other even though the data remains the same.

B. Implement the corresponding batch processes that take the clean data as input.

Implementing the analysis is done using MapReduce. In each example the readings are taken as keys. For each view a mapper is created which filters and transforms the readings while also adding relevant information from the corresponding meta data files. The reducer is then often used to aggregate or sum up the information. The batch view is simple text files, but it would be possible to serve this in another way in an extension of the project. The result of the batch views is still huge amount of data, but it is transformed in such a way that a query could easily retrieve interesting information about specific entities. For example the tracking view shows where every Wi-Fi client has been to at every point in time, and

therefore a query on a specific Wi-Fi client will be necessary to track a single person. As of now running the batch views is a manual process but it would be easy to automate to run once a day for example.

Three examples of the outputs of the batch views can be seen below:

Course popularity: ***number of "people" in room Aud3 at 8, 9 AM on a tuesday***

```
...
AUD32-3A56: 2016-10-25_08 74
AUD32-3A56: 2016-10-25_09 86
...
```

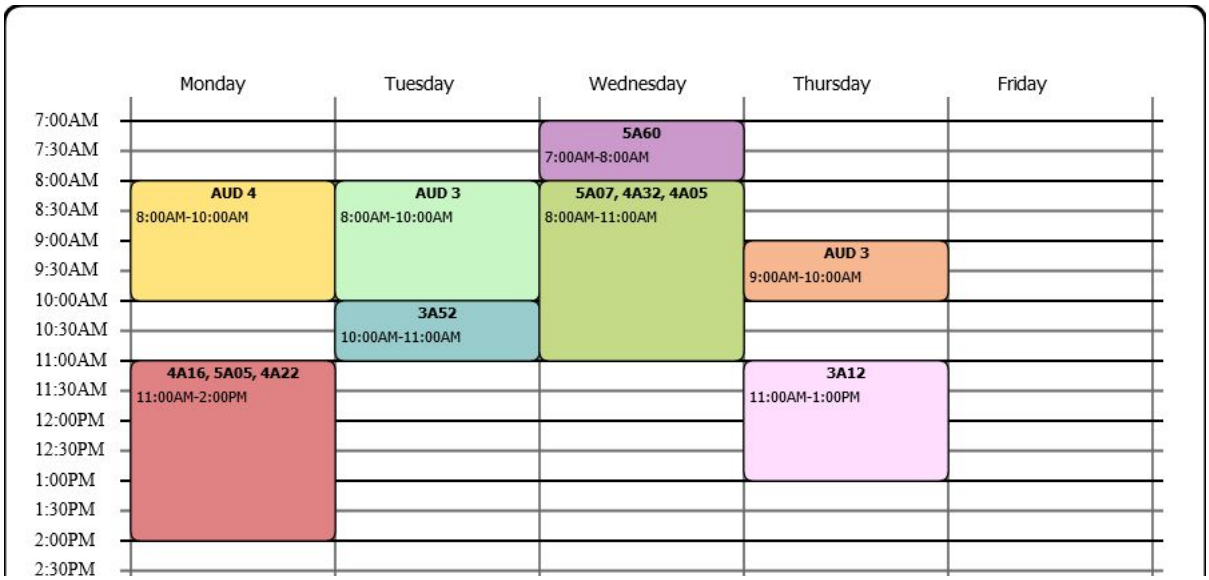
This result implies, that the course, Big Data Management (Technical), have 74-86 people showing up to the course the 25th of October '16. However, the course base states that only 65 people are taking this course[7]. This could be explained by people having multiple devices, or people sitting outside the auditorium being connected to the same access point. Therefore as expected, our assumptions does not hold in all cases.

Tracking info: ***snippet of tracking info of a person who's been around room Aud3 at 8, 9 AM on a tuesday.***

```
...
fd958189-5ad3-5586-a7ad-d3fe4e6f4695  4A32-2016-10-12:11,
AUD44A60-2016-10-24:08,               AUD44A60-2016-10-24:09,
AUD32-3A56-2016-10-04:09,          AUD32-3A56-2016-10-04:08,
5A60-2016-10-12:07,                      4A58-2016-10-24:08,
AUD44A60-2016-10-10:09,            AUD32-3A56-2016-10-04:10,
3A12-2016-10-06:11,  3A12-2016-10-06:12,  5A07-2016-10-12:11,
AUD32-3A56-2016-10-13:09,                5A05-2016-10-31:11,
5A07-2016-10-12:10,  3A52-2016-10-25:11,  3A52-2016-10-25:10,
AUD44A60-2016-10-24:10,                  4A58-2016-10-24:09,
AUD44A60-2016-10-31:10,                  4A16-2016-10-24:14,
5A07-2016-10-05:08,  4A16-2016-10-24:11,  4A16-2016-10-24:13,
4A16-2016-10-24:12,  5A07-2016-10-12:08,  5A07-2016-10-12:07,
AUD32-3A56-2016-10-25:10,            AUD44A60-2016-10-10:10,
4A58-2016-10-24:10,  5A07-2016-10-12:09,  4A16-2016-10-10:12,
4A16-2016-10-10:11,  4A16-2016-10-10:14,  4A16-2016-10-10:13,
4A22-2016-10-31:13,                      4A05-2016-10-12:11,
AUD32-3A56-2016-10-06:09,          AUD32-3A56-2016-10-13:10,
5A07-2016-10-05:09, AUD32-3A56-2016-10-25:09,
...
```

[7] https://mit.itu.dk/ucs/cb/course.sml?course_id=1835814&mode=search&semester_id=1820419

With these results, we can put together a schema for the person and match it with course information and TimeEdit, to find the rooms the person have been around. The most plausible result is, that it is a 1st year GBI student, since the schemas match.



Schema composed by Wi-Fi readings

1st year GBI schema from TimeEdit [8]

[8] http://bit.ly/2ebulGt

Vendor info: ***room Aud3 at 8, 9 AM on a tuesday***

```
...
AUD32-3A56: 2016-10-25_08-Android 9
AUD32-3A56: 2016-10-25_08-Apple iOS 2
AUD32-3A56: 2016-10-25_08-Mac OS X 3
AUD32-3A56: 2016-10-25_08-Windows 7/Vista 2
AUD32-3A56: 2016-10-25_08-Windows 8 3
AUD32-3A56: 2016-10-25_08-unknown 55
AUD32-3A56: 2016-10-25_09-Android 7
AUD32-3A56: 2016-10-25_09-Mac OS X 6
AUD32-3A56: 2016-10-25_09-Windows 7/Vista 3
AUD32-3A56: 2016-10-25_09-Windows 8 2
AUD32-3A56: 2016-10-25_09-unknown 68
...
```

This batch view is actually very much alike the course popularity, and only required small changes in the code. Adding the ClientOS to the output key of the mapper was the only thing needed. This showcases the flexibility of the MapReduce and the batch setup. Unfortunately it is visible that a large amount of data is not clean since the OS is unknown. In this batch view the unclean data is not ignored but shown, since Windows 10 for example obviously should be present but is not.

C. Do you use Hadoop to answer Q3B? Explain your answer.

The implementation uses Hadoop since it is implemented with MapReduce. This is done to make it possible to scale the results. Even though the amount of data right now is not that big, a lot of data gets added every day and at some point it will be difficult to handle with regular relational databases. Therefore, using Hadoop will ensure that the system can stay in place and if needed, it is possible to extend the system horizontally. Furthermore, Hadoop and MapReduce, makes it easier to extend the solution with different frameworks should the scope of the project extend. In the implemented batch views, it was possible for us to do all the work in one mapper and one reducer. Hence, we did not need more features, like pipelining.

Log - Technical/Critical

List the problems/challenges you faced during this project and explain how you tackled them.

Technical:

- We had high expectations at the start. We had planned to try to use Kite to create Avro schemas for us and then use kite to create a Hive dataset and import the data into Hive, and then either query directly on Hive or use Crunch[9] on top to query. But as looked more into the frameworks and tried Kite, we saw that it was harder to do than expected.
- The Kite tutorials seemed to be easy at first, but the further in we got, the more complex it became. The tutorials either seemed too easy and broad or too specific and complex to be used in our case.
- Errors and problems regarding Kite CLI in Hadoop with Hive were too specific to be googled, making it harder to progress.
- We had several problems regarding Kite CLI
    - It took a long time to run the program, because the tutorial didn't explicitly tell how with Hadoop. Thereafter, it also took a lot of time to figure out the right commands.
    - We had problems with Hadoop, which seemed like we were missing some dependencies either in the Kite .jar-file or in Hadoop. E.g. ClassDefNotFound, ClassNotFound
    - Furthermore, the newest Kite .jar-file from their own site doesn't seem to work. So we ended up downloading an older version from another site.
        - The older version could not format or make schemas with json-files (only csv), but unfortunately the data was in json.
        - The older version could only be used to complete the tutorial
    - Hive wasn't installed, when we tried to make the dataset in Hive with Kite. Instead we stored it in HDFS; but that way it couldn't be queried.
- The Hadoop installation seems to be missing classpaths and dependencies, and so in MapReduce we had to explicitly include every dependency in the compiled jar of the MapReduce job. The weird part is that our Parser job uses the same libraries, where this is not an issue. The difference between the jobs is that the MapReduce job runs on the cluster, where the Parser job runs on the Hadoop client machine that we have access to.
- The project description and assignment is so vague, that it makes it hard to decide which tools to use and why
    - We spent a lot of time on deciding what to do, and in the end we couldn't get Kite, Hive etc. to work within the time available, so we opted for raw MapReduce instead.
- It was hard to integrate the technologies / frameworks together
- When we got the project it was hard to know, where to start.

[9] https://crunch.apache.org/

- One of the difficulties was to communicate the data model with critical students, because they are lacking technical knowledge within data modeling and manipulation. Hence, they were not able to generate questions for which our big data infrastructure needs to provide answers. In order to make this gap smaller we needed to create an appendix explaining the data model and the provided data we are interpreting.
- We haven't learned anything about cleaning yet, and so it was hard to know how to.
- Hard to do all parts of the project together, which means that our knowledge is scattered.
    - Some did MapReducing
    - Some looked more into frameworks
    - Some did a little bit of both
- Initially a lot of time was spent understanding what the data was about and how different data was related.

Critical:

- First C&T Group Work: Prepared data was delayed
    - The plan for Project 2 was to have the Technical (T) team prepare the dataset prior to our first meeting. Due to miscommunication at the course administration level, this did not happen. But it gave us a chance to gain valuable insights into aspects of the data preparation process. The first meeting on Project 2 gave us a chance to look over the shoulders of T while they gained an overview of the data's properties as well as took initial steps to model the data on a whiteboard for both T and C to see. All the while they were answering any questions we had about data preparation in general and the concrete data we were given to work with. It was also an occasion where we could directly question any choices to exclude or include data that T would be taking--and be given, what we must and do assume, valid reasoning for either or.

- Views and close collaboration (C&T) is more productive for C
    - It is difficult to do the work on this project separate from T. At the very least, it seems we are more productive when the work is done together. Though we are C, we enjoy staying close to the data so we avoid being too abstract and trying to answer the project questions too broadly (as done in the earlier project), forgetting the scope we are given. Defining scenarios and views helps, though. And it reflects the approach we are taking to the data: we are looking at the data and seeing what we can do with it, as opposed to stating a question and finding the data that can answer this question.

- A quite creepy discovery
    - The iterative process of working with the data meant that interesting discoveries could be made throughout the project. From the data, T managed to compile a schedule for an individual whose data was present in the dataset. This brings about aspects of ethics to the fore, which are arguably

just as interesting to dive into as the view we chose to focus on in the 5th part (evaluation of teachers and their courses), but time constraints make us unable to treat them with our critical expertise. Nevertheless, the discovery makes us feel that we might be dealing with very creepy data that certainly warrants some critical thoughts.

Ethics/consent - Technical/critical

A. **If you were charged with a problem of coming up with ways to make these data useful to ITU in new ways what might be some options for doing so? Select and describe one potential way you might implement a way to use these data - what kind of system would this be?**

The batch views defined above all provides some interesting ways for ITU to gain information. Especially the course popularity view could be used to support the qualitative analysis currently done through the course evaluation. We consider this a plausible extension of the current system at ITU, more so than the two other views. The view should of course try to mitigate the insecure assumptions mentioned in the section, by doing extra analysis, by for example making sure that they only count unique Wi-Fi clients, with specified operating systems and so on. Mitigating the assumptions in this way is taking a critical approach in order to avoid drawing conclusions based solely on correlations, because "[…] Big Data correlations suggest causations where there might be none. We become more vulnerable to having to believe what we see without knowing the underlying whys." (Zwitter, 3 2014).

The system should be automated to take batches of data each day from the IT-department's wifi data. These data should then automatically be cleaned and stored with the corresponding batch view. Storage of the batch view could be done in a relational database, which could be queried from mit.itu.dk or a similar intranet. This way both students, teachers and administration could check the course popularity based on both the course evaluation and the wifi use.

B. **What would you implement as your consent procedure? Do you even need consent here?**

It is interesting to notice that at this point it is not transparent that the collection of data is happening and is available to ITU staff members, and furthermore that it is possible to use the data for analysis on this topic. By actually performing the analysis, it is clarified that privacy in some sense is violated, since the users of the Wi-Fi does not know that the data is used for analysis. Therefore implementing a consent procedure if ITU chose to perform any of these batch views, would be a good idea. Though it is difficult to see why ITU would track the individual student, it is easy to imagine why ITU would try to track the course popularity, which would still have some individual effect.

Consent when you set up Wi-Fi on your device. The data is largely anonymized. And that fact is what is visible. But we have now seen that actually a lot is possible if you implement analysis.

C. What are some of the ethical issues you would need to think about? (Critical students, think about the "unraveling effect" - week 7 lecture).

Since tracking the popularity of a course is indirectly tracking an individual's performance, raises ethical questions. The derived information can have personal implications if the teacher is fired from doing so.

There are aspects of the unraveling effect (Peppet 2011) that become crucial in the ethical concerns of using the Wi-Fi data for evaluating teaching staff. We may imagine that decision-makers at ITU are able to argue for using the data as an indicator of popular courses that by their popularity reflect on the capabilities of teachers. In a show of concern for ethics regarding data use, these same decision-makers could ask teachers to give consent to having the data be used. Decision-makers would contend that the data will be used to improve evaluation, leaving teachers to take a stand to this claim when they consider whether to give consent to the collection and use or not. Teachers that give consent will be in line with the interests of the administration, whereas teachers who do not appear to stand in opposition and may see consequences because of it. In this scenario such consequences could ultimately be the loss of the job. Not because they are poorly evaluated, but because they do not consent to the practices that decision-makers incentivize and claim are a part of the evaluation process itself.

Appendix

Literature

1. Peppet, S. R. (2011). Unraveling privacy: The personal prospectus and the threat of a full-disclosure future. Nw. UL Rev., 105, 1153.
2. Turow (2011) The Daily You. Yale University Press - Introduction & Ch 1
3. Zwitter, A. (2014). Big Data ethics. Big Data & Society, 1(2).