DATA7001

Human centred problem formulation in data science

Due 28 August 17:00, submit via Blackboard
10 Marks, Individual

In this assignment, you will apply your learning from the design thinking lecture to undertake human centred formulation for a data problem. The assignment consists of a short (approx. 2-page) report that presents the results of your investigation from a semi-structured interview with stakeholders and study of a data intensive domain, namely *Learning Analytics*.

Summary of the domain: The University of Queensland (UQ) St Lucia is host to over 50,000 visitors on a typical semester weekday (Charles-Edwards and Corcoran 2016). This population is comprised of students, staff and other visitors accessing the facilities and services on Campus. This population is highly dynamic and varies with respect to its size, composition and distribution over the course of the day. Continuous monitoring of the population on Campus has the potential to inform a range of decisions with respect to space utilisation on Campus.

Stakeholders:
Vern Bawden (Senior Manager, Data and Identity Services)
Sasenka Abeysooriya (Senior Strategic Adviser, Data Strategy and Governance)
Lachlan Kuhn (Manager, Analytics Data Services)
Berg Lloyd-Haig (Senior Systems Programmer, Analytics Data Services)

The semi-structured interview with the stakeholders is available below. If needed students can also ask further questions on the course Piazza site

piazza.com/uq.edu.au/semester22020/data7001/home

using the subject of “Interview Question for Design Thinking Assignment”

Your report will have several sections as detailed below. Begin the report with a simple introduction that explains the purpose of the document and its contents; end the report with a short conclusion. Prepare the body of the report using the sections below:

Data profile:
This section will describe the data in human or social terms: What is the data? What categories? Parameters? How much data is there? How often is it updated? How far does it go back in time? These are initial and not exhaustive example questions.

Stakeholders:
Identify and describe the stakeholders. Who are the people who own the data? Who else has interests in it? Who benefits from it? How? You may use a diagram to represent stakeholder groups and place the panellists as stakeholders into this diagram along with the other stakeholders you identify.

Scenarios of use:

Identify three use cases of the data. The first should be what you understand to be the typical use case—who is using the data, in what circumstances, how they are working with it, what kind of relationships they are probing, what kind of questions they can find answers to, and for whom. The second should be a little farther afield, e.g. a different type of stakeholder, a different class of question. The third should be a "fringe" scenario: a possible but atypical scenario, relating to an often overlooked use of the data, and the set of circumstances, motives and/or skill set required to pursue the data questions related to this use.

For each use case, provide a rationale that links the information you obtained from the panel discussion to values / needs that you had deduced to your statement of the scenario, then identify who benefits and who may be marginalised by its answer from the data.

Limits:

Identify two questions that are (just) beyond the scope of the data, but that with a little more data, or other kinds of data, or an external source of data to compare etc. could be answered. Again, for each, identify who is likely to benefit and who may be marginalised. Ensure that you provide a rationale that links the information you obtained from the panel discussion to your questions (tip: sometimes this information may come from what is not said).

For these last two "limit" questions you have identified, outline how you think you would have to work with the data; what kind of data collection or analysis you would need to conduct in order to get an answer to the question. Where possible, articulate what would be sufficient to count as a conclusive answer one way or the other, versus what would be indeterminate, and why.

Assessment:

Your report will be assessed on the following criteria:

- Sensibleness: How reasonable and grounded in human experience, research or evidence is the report? How much sense does it make? (2 marks)
- Scope: How complete is the report in terms of probing the properties of the data, the people and their purposes that may have interests in it? (2 marks)
- Creativity: How well does the report show imagination for how people may use the data and for what purposes? (2 marks)
- Understanding: How well does the report demonstrate an understanding of data and data science possibilities as they relate to human agendas and issues? (2 marks)
- Clarity and style: How well does the report communicate professionally and clearly? Is the presentation succinct, to the point and within page limit? (2 marks)

Interview Transcript

Find below the interview transcript for analysis, where Q refers to the interviewer and A refers to the interviewee's responses. Each line is numbered for ease in your analysis and referencing for your own purpose.

Q **What is space utilisation analytics and why do you use it?**

A Space Utilisation is an ongoing area of research for the university. Its aim is to help better understand the usage of physical spaces across our campuses, and how these spaces are used for teaching, learning, research activity, and professional services. This research is critical for forecasting future space demands, which in turn informs decision making concerning building construction, renovation & refurbishment, as well as the optimisation of teaching activity scheduling.

Q **What data are you collecting? What are you looking for?**

A Physical surveys (a human-centered process involving the manual counting of persons) of selected teaching and learning spaces have traditionally been conducted once per semester around the third week of teaching, and are coordinated by UQ's Properties & Facilities division. These surveys form a ground truth to compare other counting technologies, such as the Cohera system (which processes images and counts people entering and leaving rooms), and has also been used in testing in several teaching and learning entryways. The Cohera system has an internal model that produces a simple integer count time-series vector representing the number of people currently in the space.

The Analytics Data Services team within the Information Technology Services department were engaged to study the feasibility of using WiFi session information to improve the temporal richness of the existing Space Utilisation datasets. The project aimed to create a simple inflation factor model that can be applied to WiFi session aggregations to arrive at an estimated attendance value for a given space. After a data quality inspection of all datasets, these estimates are compared against the physical survey and the Cohera camera counter results to measure the inflation model's accuracy.

Q **How we collect data? Who owns the data? How frequently is it being collected?**

A The locations for the physical survey were chosen with a spread of physical attributes to provide a reasonable representation of space diversity across the St Lucia campus. People counts of floors and rooms were conducted at 10 minute intervals over the course of a week; these counts are then digitised into Excel spreadsheets by the surveyors. The Cohera statistics were exported by Properties & Facilities into CSV format, and provided at 10 minute intervals. Archibus (the system which houses data relating to buildings, floors, rooms), Cohera, and the survey information are all owned by the Properties & Facilities division.

Three WiFi datasets were investigated. Cisco Prime Sessions provides a time series of distinct user sessions and their duration in seconds associated with a particular wireless access point. An additional Cisco system called CMX provides two more datasets: Floor Counts and Client Snapshots. The Cisco CMX Floor Counts produced 5 minute interval snapshots of devices detected (but not connected), connected, and total (the sum of detected and connected). The Cisco CMX Client Snapshots dataset produces 5 minute interval snapshots of every wireless client across the system and their last known state, along with a last seen timestamp. The Client Snapshots includes a location triangulation system that can be used to approximate a device's location on a given building's floor, and generates a geopoint for each client, along with a confidence factor value that is equal to half the length in feet of a 95% confidence bounding box, where the generated geopoint is at its centre. Building information, including floor and geopoints from Archibus enrich the WiFi datasets to provide the additional context required to compare the data with the survey and Cohera datasets.

The wireless infrastructure is operated by the ITS Networks & Data Centres team who are the custodians of the wireless datasets. The Prime Sessions are collected from the wireless access point controllers shortly after midnight for the previous day, the CMX datasets are collected from the Cisco CMX REST API every five minutes using an AWS Lambda function. The individual events of the three Cisco datasets are enriched with Archibus site, building, and floor tables sourced via the ITS Data Hub and then stored in compressed JSON line files as AWS S3 objects. The events are also sent to an Elasticsearch cluster that is used centrally in ITS for near real-time operational observations using the Kibana front-end application. Upon creation, the S3 objects publish their S3 URI into an AWS SNS topic which in turn is used to populate an AWS SQS queue in the Information Technology Services Analytics Data Services account, where the compressed S3 objects are then copied using a Lambda script. An Airflow scheduled DAG (pipeline) executes an Apache Spark (pyspark) script on a daily basis that loads the compressed JSON objects into a Spark data frame, performs some additional ETL, and finally stores the data and its schema back to S3 in the columnar data format Apache Parquet.

The final step is to compare the datasets to produce correlation and feasibility reports using a combination of Apache Spark and R Studio. A model will be proposed and then implemented to produce additional curated datasets that will store the model outputs in Apache Parquet format and make it accessible via JDBC/ODBC interfaces for the Business Intelligence unit, alongside bespoke dashboard and info-graphic tools.

Q **Why is collecting this data important? Who will use the data?**

A The development of a sufficiently accurate longitudinal space utilisation model will provide visibility of seasonal trends over the course of a year (as opposed to targeted max values of a physical survey during the busy time of semester), across all wireless locations of the campuses. Properties & Facilities can use the insights gained from the space utilisation model to make informed decisions on proposed construction projects, taking into account the particular properties of the physical spaces under scrutiny.

Q **Where is it being used? Where else can it also be used?**

A Dashboards (including infographics and summary information) have been produced from the resultant datasets that allows Properties & Facilities to get up-to-date views on campus space utilisation. Additional bodies of work have been undertaken to blend the space utilisation datasets with course enrolment and activity timetabling datasets to enable the analysis of timetable efficiency and space popularity. For instance, scheduled teaching activities that do not have any enrolled students turning up can be flagged and investigated to free up space in the schedule. Properties for teaching spaces that are underutilised can be studied and plans to improve them for increased utilisation can be realised. Room capacity vs. enrolment numbers can be compared to ensure efficient use of spaces as well as ensuring student comfort, while avoiding over-provisioning. The wireless sessions datasets have also been useful in servicing requests from Queensland Health for contact tracing during the COVID-19 pandemic.

Q **Do any stakeholders stand to benefit from this? Further, reflect upon any adverse effect(s) faced by stakeholders.**

A Despite the insights that are produced to drive enhancements to the UQ academic venture from the Space Utilisation project, ethical and privacy considerations regarding personally identifiable and location information are at the forefront when it comes to potential invasions of student and staff privacy. While the source datasets do include specific information about users and their device locations, the model output only produces aggregate information that does not included any personally identifiable information, and strictly adheres to the the University of Queensland's Privacy Management Policy (1.60.02) and the Queensland Information Privacy Act 2009. The model following ethical guidelines mitigates any associated risk.