DATA7001

Human centred problem formulation in

data science

数据科学中以人为中心的问题公式化

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

截止日期:8月28日17:00，通过黑板提交10分，个人的

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.

在这个作业中，你将应用你从设计思维的讲座中所学到的知识来为一个数据问题进行以人为中心的表述。赋值由一个简短的(约。2页)报告，展示你的调查结果，从半结构化的采访利益相关者和（研究的数据密集型领域，即学习分析。

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.

领域概要:昆士兰大学(UQ)圣卢西亚在一个典型的学期工作日内接待超过5万名游客(Charles-Edwards and Corcoran 2016)。这群人包括学生、工作人员和参观设施的其他访客和校园服务。这个种群是高度动态的，在一天的过程中随着其人多少、组成和分布而变化。对校园人口的持续监测有可能为有关校园空间利用的一系列决策提供依据。

Stakeholders:

利益相关者:

Vern Bawden (Senior Manager, Data and Identity Services)

Vern Bawden(数据和身份识别的高级经理)

Sasenka Abeysooriya (Senior Strategic Adviser, Data Strategy and Governance)

Sasenka Abeysooriya(数据战略和治理高级战略顾问)

Lachlan Kuhn (Manager, Analytics Data Services)

Lachlan Kuhn(分析数据服务经理)

Berg Lloyd-Haig (Senior Systems Programmer, Analytics Data Services)

Berg Lloyd-Haig(高级系统程序员，分析数据服务)

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

与利益相关者的半结构化访谈如下。如有需要，学生亦可在课程广场的网站上查询

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

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

以“设计思维的论文作为访谈的问题”为主题

#设计思维：共情，定义，构思，原型和测试。共情（Empathize），也称移情，指理解用户的需求。这个阶段的核心价值是以人为中心，一切需求的出发点是“人”，通过观察、倾听、访谈等方法和用户产生共情，进而分析出用户的核心诉求。定义（Define），以人为中心重新组织和定义问题。在定义阶段通常可以用一句话来描述问题：谁？（用户User）有什么需要？（需求Need）我发现了什么？（洞察Insight），简称POV法。定义阶段的核心价值是收敛，排定优先顺序，在此我们分辨出对用户来说什么是真正重要的，什么是我们应该花更多时间去投入的。构思（Ideate），在创意阶段发散思维产生许多点子或想法。在这一阶段，我们可以用各种方法来加强创造性，头脑风暴和草图是最为常用的，目标是产出尽可能多不同的概念，然后将它们可视化。原型（Prototype），设计产品原型或问题的解决方案。构思阶段结束后，产生了许多点子和想法，我们从中选取一些想法形成基本的概念模型，设计出相对详细的解决方案。原型的核心价值是MVP精神，也即生成最小可行性产品。测试（Test），验证设计原型，并改进方案。测试阶段，可以找同事进行，当然最佳方式是对用户进行测试，更具启发性，然后将测试的结果反馈到产品的下一个迭代版本中，所以，测试阶段的核心价值是迭代精神。

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:

你的报告将包括以下几个部分。以简单的介绍开始报告，说明文件的目的及其内容;以简短的结论结束报告。使用以下各节编写报告正文:

Data profile:

数据概要:

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.

本节将从人类或社会的角度描述数据:什么是数据?什么类别?参数是什么?有多少数据?多久更新一次?它能追溯到什么时候?这些是初始的，不是详尽的示例问题。

Stakeholders:

利益相关者：

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.

确定并描述利益相关者。谁是拥有数据的人？还有谁对此感兴趣？谁从中受益？怎么获利？您可以使用图表来表示利益相关者组，并将小组成员作为利益相关者与您确定的其他利益相关者一起放入此图中。

Scenarios of use:

使用场景：

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.

确定数据的三个用例。第一个问题应该是你所理解的典型用例谁在使用数据，在什么情况下，他们如何使用数据，他们在探索什么样的关系，他们可以找到什么样的问题的答案，以及为谁。第二个问题应该稍微远一点，例如，不同类型的利益相关者，不同类别的问题。第三个应该是一个“边缘”场景：一个可能但不典型的场景，与经常被忽视的数据使用有关，以及一组环境、动机和/或技能集，以寻求与此使用相关的数据问题。

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.

对于每个用例，说明为什么要这么做，将您从小组讨论中获得的信息与您从场景陈述中推断出的结论/需求联系起来，然后通过数据中的答案确定哪些人受益，哪些人可能被边缘化#（知道结果，但不知道过程）。

Limits:

限制：

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).

确定两个问题（只是）超出了数据的范围，但如果有更多的数据，或其他类型的数据，或外部数据源进行比较等可以得到答案。同样，对于每个人，确定谁可能受益，谁可能被边缘化。确保你提供了一个基本原理，将你从小组讨论中获得的信息与你的问题联系起来(提示:有时这些信息可能来自没有说出来的内容)。

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.

对于你所确定的最后两个“限制”问题，请概述你认为你将如何处理这些数据;为了得到这个问题的答案，你需要进行什么样的数据收集或分析?在可能的情况下，清楚地说明哪一种方式足以算作一个结论性的答案，而不是什么是不确定的，以及为什么。

Assessment:

评价:

Your report will be assessed on the following criteria:

我们会根据以下准则评估你的报告:

• Sensibleness: How reasonable and grounded in human experience, research or evidence is the report? How much sense does it make? (2 marks)

•敏感性:报告的合理性和基于人类经验、研究或证据的程度如何？这有多合理？（2分）

• 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)

•范围:在调查数据的属性、可能感兴趣的人及其目的方面，该报告是否完整?(2分)

• Creativity: How well does the report show imagination for how people may use the data and for what purposes? (2 marks)

•创造力:报告对人们如何使用数据以及用于什么目的的想象力表现得如何?(2分)

• 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)

•理解:报告如何证明对与人类议程和问题相关的数据和数据科学可能性的理解?(2分)

• 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)

•清晰性和风格：报告在专业性和清晰性方面的沟通情况如何？演示文稿是否简洁、中肯且在页数限制内？（2分）

Interview Transcript

面试笔录

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.

在面试记录下面找到分析，其中Q代表面试官，A代表面试官的回答。每一行都有编号，以便于分析和参考。

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

Q：什么是空间利用分析，你为什么要用它？

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.

A：空间利用是大学正在进行的研究领域。为了更好地了解我们校园内的教学和活动空间，我们的研究和服务是如何更好地利用这些空间的。这项研究对于预测未来的空间需求至关重要，这反过来又为有关建筑建造、维修和翻新以及教学活动安排的优化提供决策依据。

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

Q：你在收集什么数据？你在找什么？

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.

A：传统上，每学期在教学第三周左右对选定的教学和学习空间进行一次实体调查（一个以人为中心的过程，包括人工计算人数），并由昆士兰大学的物业和设施部门进行协调。这些调查形成了一个基本的事实，以比较其他计数技术，如Cohera系统（它处理图像并对进出房间的人进行计数），并且在一些教学入口进行了测试。Cohera系统有一个内部模型，它产生一个简单的整数计数时间序列向量，表示当前在空间中的人数。

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.

信息技术服务部门的分析数据服务团队，致力研究利用WiFi会议的可行性，以提高现有空间利用数据集的时间丰富性。该项目旨在创建一个简单的膨胀系数模型，可应用于WiFi会话聚合，以达到给定空间的估计出现人数。在对所有数据集进行数据质量检查后，将这些估计值与物理调查和Cohera相机计数器结果进行比较，以衡量膨胀模型的准确性。

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

Q：我们如何收集数据？谁拥有这些数据？多久收集一次？

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.

A：物理测量的位置应考虑物理属性，以合理表示圣卢西亚校园内的空间多样性。在一周的时间里，每10分钟对楼层和房间进行人口统计；然后由测量员将这些计数数字化到Excel电子表格中。Cohera统计数据由Properties&Facilities导出为CSV格式，每隔10分钟提供一次。Archibus（存放与建筑物、楼层、房间相关的数据的系统）、Cohera和测量信息均归财产与设施部门所有。

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.

三个WiFi数据集被研究。Cisco Prime Sessions提供不同用户会话的时间序列，以及与特定无线接入点相关联的持续时间(以秒为单位)。另外一个Cisco系统被称为CMX提供了另外两个数据集:楼层数和客户端快照。Cisco CMX楼层计数生成检测到（但未连接）、已连接和总计（已检测和已连接的总和）的5分钟间隔快照。Cisco CMX客户端快照数据集生成系统是由每个无线客户端的5分钟间隔快照，以及它们最后已知的状态，以及最后看到的时间。客户端快照包括一个位置三角测量系统，该系统可用于估计设备在给定建筑物地板上的位置，并为每个客户端生成一个地理点，以及一个置信因子值，该值等于95%置信边界框长度的一半（英尺），生成的地理点位于其中心。建筑信息，包括来自Archibus的楼层和地质地形，丰富了WiFi数据集，提供了将数据与勘测数据集和Cohera数据集进行比较所需的环境。

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.

无线基础设施由ITS网络和数据中心团队操作，他们是无线数据集的保管人。主要的会议是在前一天午夜后从无线接入点控制器收集的，对于前一天，CMX数据集每五分钟使用AWS Lambda函数从Cisco CMX REST API收集一次。这三个Cisco数据集的单个事件通过ITS数据中心提供的Archibus站点、构建和楼层表进行丰富，然后作为aws3对象存储在压缩的JSON行文件中。这些事件还被发送到一个Elasticsearch集群，该集群在ITS中集中使用，使用Kibana前端应用程序进行近乎实时的操作观测。在创建之后，S3对象将其s3uri发布到awsns主题中，该主题又用于添加到信息技术与分析数据服务帐户中的AWS SQS队列，然后使用Lambda脚本复制压缩的S3对象。一个气流调度的DAG(管道)每天执行一个Apache Spark (pyspark)脚本，该脚本将压缩后的JSON对象加载到一个Spark数据帧中，执行一些额外的ETL，最后将数据及其模式以列式数据格式Apache Parquet存储回S3。

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.

最后一步是使用Apache Spark和R Studio的组合来比较数据集以生成相关性和可行性报告。将提出一个模型，然后实施该模型以产生额外数据集，这些数据集将以Apache Parquet格式存储模型输出，并通过JDBC/ODBC接口为商业智能单元提供访问，同时提供定制的仪表板和信息图形工具。

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

Q：为什么收集这些数据很重要？谁将使用这些数据？

A The developmen tof 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.

A：开发一个足够精确的纵向空间利用模型，将提供校园内所有无线位置一年内的季节趋势（而不是学期繁忙期间物理调查的目标最大值）。物业和设施可以利用从空间利用模型中获得的见解，对拟建的建筑项目做出明智的决策，同时考虑到受审查的物理空间的特殊属性。

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

Q：它在哪里使用？它还能用在哪里呢？

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.

A：仪表盘（包括信息图形和摘要信息）已从结果数据集中生成，使物业和设施能够获得校园空间利用率的最新视图。还开展了更多的工作，将空间利用数据集与课程注册和活动时间表数据集相结合，以便分析时间表效率和空间普及率。例如，可以标记和调查没有任何注册学生的预定教学活动，以便在时间表中腾出空间。可以研究未充分利用的教学空间的属性，并计划改善它们以增加利用率。可以比较房间容量与注册人数，以确保有效利用空间，确保学生舒适，同时避免过度供应。在COVID-19流感大流行期间，无线会话数据集还可用于服务昆士兰卫生部的接触追踪请求。

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

Q：有任何利益相关者会从中受益吗?此外，思考利益相关者所面临的不利影响。

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.

A：尽管空间利用项目提出了促进UQ学术项目改进的见解，但当涉及到可能侵犯学生和员工隐私时，关于个人身份和位置信息的道德和隐私考虑是最重要的。虽然源数据集确实包含有关用户及其设备位置的具体信息，但模型输出仅生成不包含任何个人识别信息的聚合信息，并严格遵守昆士兰大学的隐私管理政策（1.60.02）和2009年昆士兰信息隐私法。遵循道德准则的模型可以降低任何相关风险。