PTAS Project Report (for REGULAR PROJECT GRANTS)

Project Title:
Data Fairs, Matchmaking and Collaboration Patterns for Data Science Teaching

Project type: **B - Innovation Project** (introduction and evaluation of an educational innovation, usually taking a practical approach)

Principal Investigators: Dave Murray-Rust, Benjamin Bach
Schools/department: Design/ECA, School of Informatics

Team members (including Schools and Departments):
Anouk Lang / Languages, Literatures and Cultures
Catherine Lai / School of Informatics

For further details, please contact:

Project teams must submit a report within 4 months of the conclusion of their project. Copies of dissemination material (eg journals/newsletter articles, conference papers, posters should be listed and attached (separate to the word count). The brief report will be published on the IAD web pages.

Report (maximum 1500 words)

What did you do?

Goals: In this project, we developed the concept of a "Data Fair", which aims to bring together people and agencies with data in need of analysis with students who need to learn about data analysis. The goal was to create an engaging and compelling learning and teaching environment for both sides: students would be engaged by real-world problems, learn about specific domains, and develop skills particular to this solution. For data holders, the experience would include working with a fresh set of minds and learn about particular practices in current design university education. In general, data fairs are meant to be a means to

- Avoid the use of 'toy' datasets with simplified findings
- learn about the steps needed to work with messy realtime data

- make the learning feel worthwhile, as the findings matter to the data holders
- develop skills in carrying out collaborations around data
- develop skills in communicating about data as well as analysing it

Context: As part of the course, *Data Science for Design,* students learned about the basic concepts of data science and exploratory data analysis, data visualization, data ethics and privacy, as well as general design practice and methodology. Approximately half of the students came from a design background (graphic design, product design, architecture..) while the other half came from an engineering background (e.g. compute science). The course offered a first introduction into programming with python to get both student groups familiar with data analysis programming. The 2nd part of the course, eight weeks in total, students engaged in the data fair and the work with their project partners.

Data Fairs: People with interesting data submit *data challenges* describing their data, context and questions; students meet them at a 4h interactive event where they pitch their questions and start discussions.

Challenges: The challenges that we wanted to address in this PTAS project included:

- Providing the right set of instructions to both students and data holders
- Providing students with an informative description of the respective data projects
- manage both students' and data holders expectations towards the project, including communication and outcomes.
- Supporting group selection, allocation, as w ell as communication with data holders and students during the course.
- Understanding working patterns and key practices that made the collaborations successful - Generating guidelines for running future data fairs.

Project spending: The project paid for

- assistants to help with running data fairs and gather data about students feedback and engagement, as well as
- the development of a data fair website which we used to publish data challenges and allocate groups.

Process: Over the course of this PTAS project, we ran two courses (2018/19 and 2019/20), based on a first 'pilot' version i n 2017/18 which brought us the idea of data fairs. A list of partners in these 2 data fairs is provided below.

- **Data challenges:** based on student feedback from our 2017/18 data fair, we created a template for a data challenge including information about the origin and type of

data, a description of the current analysis progress, skills required, potential stake-holders, as well as the description of a challenge for this project.

- **Website:** To populate and access these information to students, we developed a website which was send to data holders in advance. Students can sign up to a class and register their interest to specific projects. This helped us course organizers assessing popular and less popular topics and to mitigate during the data fair event.
- **Instruction material:** We informed two external data fairs run by colleagues as part of a Business Analytics Programme in Paris, France. We supported their data fair through guidelines and guidance material, the data challenge template, the website, and questionnaires for gathering student and data holders feedback .
- We created a website with information about our data fairs, instructions, and guidance, and example projects: https://sites.google.com/view/datafairs .

Project partners include:

- **2018:** NHS, School of Literatures, Languages and Cultures, School of Social and Political

Sciences, School of Informatics, Scottish Government Statistics division, Wiki Media

- **2019:** National Library of Scotland, School of Social and Political Sciences, Space and Satellite programme Scotland, Scottish Government Statistics division, Wikimedia foundation, University of Edinburgh Student office, School of Law, School of Informatics, School of Literatures, Languages and Cultures

Impact: Over the two data fairs we run, student projects managed to make the following impact:

- The **National Libraries Scotland (**NLS **)** was involved in 4 projects all well received by management and advertised on the NLS's data foundry website [1] and demoed to the Prime Minister of Scotland Nicola Sturgeon as part of her visit to NLS (she liked the creativity students bring into such projects). As a result, NLS was running a fellowship (worth £7000) over this summer to encourage further creative projects.
- A project about Scottish witchcraft, proposed by **Wiki Media** foundation was featured largely on Twitter and in talks, with students presenting a lasercut map, experientially showing trial data from medieval Scotland.
- A project using network science to show similarity and evolution of worldwide trade agreements (with Dr Anouk Lang) was presented at the Digital Humanities 2019 conference (DH2019).

[1] https://data.nls.uk

- **MSc projects:** Two students decided to continue their projects during their consequent masters project working on the visualization of historical migration data and geographic networks.
- One collaboration with an external partner (Dr. Campagnolo) led to a **successful grant application** on creating online data visualization course f or professionals (£80k).
- **PhD student application:** One project lead to the formulation of a larger research grant worth £250k, which made it to the 2nd stage for the funding application.
- One student built on the course to work on a commercialisation activity with the Scotch Malt Whisky Society combining natural language processing and visualisation

What did you find out?

Through interviews with students and data holders, as well as personal reflections on the program and the data fairs, we conclude with the following key findings.

- **Expectations are crucial** : students need specific support in figuring out what the data holders role is, and h ow best to make use of them. This included making more use of the data holder's time to understand what is p resent in the data, but also knowing when to stop bombarding them with repetitive requests for information.
- **Simple data** can have powerful results through creative communication and distillation so simple data that allows students to come to results quickly is safer than larger, complex data sets.
- **Data fairs require organizational infrastructure** to contact potential data holders, collect and coordinate applications, respond to questions from data holders and students, manage student allocation and mitigate group work. Learning analytics approaches could be a possible solution here, as following group progress is particularly important.
- **Guidance and project discussions** are necessary during the course. We provided for r**egular check-in sessions** with data holders and course instructors. Even short (15min weekly sessions), helped to address design issues, create a common language and define common goals. It was extremely helpful to both students, teachers, and students, to hear each other's opinions, ideas, and language.
- **Differences among data holders:** Some of the data holders engaged quite strongly with the students while others were more distant, leading to different student experiences and learning curves. Setting up specific expectations and check-ins helps with this.
- **Differences among data challenges:** There was a large range of domains and complexity to the data provided, with more or less open challenges. Developing a summary of the challenge that students can understand is crucial, especially when

students have limited skills, and must select data that fits with their skill levels as well as relates to their interests.
- **Working with data holders** early and closely improves the quality of data challenges they provide, ensuring they fit with student capabilities and course constraints, and that expectations are reasonable.
- **Clear guidelines and timelines are necessary** for both students and data holders in order to keep pace with the very short timeline of a course project.

How did you disseminate your findings?

We have published one paper covering the high level outline of the method used: *Joseph Corneli, Dave Murray-Rust and Benjamin Bach: "Towards Open-World Scenarios: Teaching the Social Side of Data Science"* at the AISB50 Symposium on Cybernetic Serendipity.

We have a further paper in development with empirical data: *"Data Fairs for Education and Collaboration in Visual Data Science: An Experience Report"* aimed at publication later this year at a venue suitable for data science education such as IEEE VIS, or IEEE Transactions in Visualization and Computer Graphics (TVCG). That paper will
- describe our detailed course methodology
- report on student and data holder feedback
- report on case studies project examples
- report on our findings, i.e., an extension of the above mentioned findings. -embed our findings into related work

Our website for public information and with detailed information on time schedule etc.: https://sites.google.com/view/datafairs/home.

Data holders helped with the dissemination by advertising student projects on their respective websites and communities.

Word of mouth is particularly important for this kind of activity, especially through the data holders as we get requests from people who have heard about our data fair from previous data fair participants.

What have been the benefits to student learning?
At the end of each data fair, we conducted interviews with the student cohort. Key learning was around:

- **Communication skills with external partners** : this includes pitching ideas and communicating an idea through sketches, written text, and face-to-face meetings. For many students this was their first project working closely with external partners.
- **Understanding real-world problems** and better understanding how the students' skills and interests can contribute to providing a potential solution to such a problem.
- **Iteration, balancing creativity and constraint:** Students learn that not every idea they have may be an appropriate solution to the problem, and that it may take many iterations to reach a solution that is practicable and acceptable in a real-world context. At the same time, many challenges were not narrowly defined allowing a creative process and demonstrating novel solutions to data holders.
- **Specific problems** that students are interested in but which are not covered in the courses. This includes NLP and dealing with textual data, geographic data, temporal data, etc. The course leaves students with room to explore areas of interest in an applied and practical environment.
- **Dealing with messy data** : most of our data holders' data came in their natural (messy) form; the ability to understand and process raw data is a crucial skill future data scientists need and one of the learning outcomes of our Data Science for Design course.
- **Create strong portfolio and contacts** into industry for internships and job search.

Wider benefits: Besides benefits for students, we found additional benefits to other groups involves in this project.

- Data holders: gained insight on their data and public visibility as noted above.
- **Course programme:** The course program (*Master in Design Informatics*) benefitted from both a wider exposure to external partners a s they could see and experience what this program is teaching students, as well as exposure to real-world problems and potential external partners. Data fairs acted as an outreach exercise showing externals what we do while demonstrating student creativity and skills.
- **University researchers and course instructors** were able to explore projects with course instructors specializing in data science and visualization. This opened new and strengthened existing collaborations in these areas.
- **Course instructors** were freed from their duty to provide for clean data examples and manage very similar class projects. In turn, their duties shifted towards managing a broad set of projects as well as expectations of external partners.

How could these benefits be extended to other parts of the university?

Our Data fairs can be extended to other parts of the data science curriculum with potentially minimalistic changes: working with more advanced students or those with particular skills, and broadened to include more academic partners as well as outside

organisations. This could be themed, e.g., a set of Data Fairs on data for public policy, or social good, or similar themes, to bring in a larger number of external partners and a committed cohort of students. Data fairs could also be apply to taught programmes other than design informatics that have a strong design background or, even to programmes that do exactly lack a strong design component. To achieve these goals, our website will help streamlining the sign-up process and our paper will help with the protocol and reflections

As future work, we are planning:

- to run data fairs for different courses (e.g., pure data visualization course, run by PI Bach) - apply the concept to MSc and other student (group) projects.
- include training / education / best-practices for cross-disciplinary work into programme curriculum, e.g., how to pitch an idea, how to understand a data challenge, how to communicate effectively with external partners, how to deal with private and confident data.
- Apply for more funding to extend the scheme and address some the open questions and challenges mentioned in this report.

Financial statement

This project has utilised the funding awarded to it by the PTAS adjudication committee and the Principal Investigator or School Administrator appropriate can provide financial statements showing the funding usage as and when required by the UoE Development Trusts who may require it for auditing purposes.

Please send an electronic PDF copy of this report to:

Email: iad.teach@ed.ac.uk