Budding Data Scientists Hackathon

Hui Xiang Chua
National University of Singapore
datadoubleconfirm@gmail.com

Ee-Ling Chua
Hwa Chong Institution
chuael@hci.edu.sg

Kenneth Soo
Stanford University
kenneth@algobeans.com

ABSTRACT

The "Budding Data Scientists Hackathon" was a pilot program to bring data science into a high school's curriculum in Singapore. Unlike typical hackathons, this hackathon lasted for a few months. A total of seven teams comprising 22 students underwent one week of intensive training workshops and five months of mentoring to work on projects tackling social challenges using data science. The hackathon was made possible with the support of the KDD Impact Program [1].

1. INTRODUCTION

The "Budding Data Scientists Hackathon" set out to achieve the following objectives:

- Enhance data science community engagement;
- Expand outreach of data science;
- Increase diversity and participation in data science;
- Increase societal impact of data science;
- Influence public policy through data science.

The hackathon aimed to motivate upper secondary school students (i.e. grade 9/ 10 of the U.S. high school system) to develop an interest in data science and use data science to help a social cause. They worked in teams to help tackle social challenges of their interest using data science, with a possibility of improving the data maturity within Voluntary Welfare Organisations (i.e. non-profit organisations that provide welfare services and/or services that benefit the community at large). All teams presented their projects to a judging panel at the final showdown, and prize money were awarded to the top three teams. A special award was also given for best visualization.

As the current secondary school curriculum did not encompass data science, all participants underwent five training sessions (approx. 20 hours). The training covered different aspects of data science such as statistics, programming, visualization, data maturity framework, data pipelines etc.

This would be the first time students were able to gain real-world experience working with data science problems at earlier stages of their education in Singapore. The inaugural “Budding Data Scientists Hackathon” brought together five teams of students from Hwa Chong Institution, and two teams from the affiliated Nanyang Girls’ High School, with 3–4 students per team.

The final showdown was open to teachers and non-participating students to raise awareness of data science and its applications.

The various data science projects done during the hackathon can become use cases while the "Budding Data Scientists Hackathon" can be replicated across different high schools. Table 1 outlines the timeline of the hackathon.

Timeline	Deliverables
Jan - Mar 2018	Curriculum design
Mar 2018	Training week
Mar - Aug 2018	Mentoring
Aug 2018	Final showdown/ Project presentations

Table 1. Timeline of "Budding Data Scientists Hackathon".

2. OUR APPROACH

A. Training and Curriculum Design

The curriculum for the training sessions was developed by data science practitioners, namely Hui Xiang Chua and Kenneth Soo, and constituted data science concepts and hands-on exercises using tools such as R, Python and Tableau (see Table 2). The course materials were meticulously-designed and included the use of visuals, multimedia, and real-world examples. Homework was assigned to participants at the end of each session for them to apply what they had learnt, and was also used to evaluate the participants’ learning progress.

The curriculum was customized to fit the school's timetabling of five 4-hour sessions, as part of Sabbaticals Week where students learnt subjects of their interest outside of their core subjects. There was a focus on mathematical and computational analysis in the curriculum design as the sabbaticals were meant to further enhance students’ interests and develop their expertise in specific core subjects taught during lower secondary, namely mathematics and computing in this case.

B. Project Mentoring

During the project scoping stage, each team had to determine a topic of analysis that centered around public policy or a particular social cause. Participants could choose to make use of open data, reach out to voluntary welfare (non-profit) organisations of interest for data, and/ or collect their own data.

We partnered with Animal Concerns Research and Education Society (ACRES), a non-governmental organisation and a registered animal welfare charity with the Ministry of Culture, Community and Youth in Singapore, which focuses on tackling wildlife crime and humane education. Students could opt to work on projects that helped ACRES in their work.

To ensure that students received help and guidance whenever they needed, mentoring was done over a team collaboration platform, Slack. In addition, three face-to-face consultation sessions are held for progress updates and clarifications on project scoping, data collection, and data analysis.

Day	Curriculum
1	Lab 1: Software installation Theory 1: Introduction to various Data Science tasks Theory 2: Basic statistical concepts Lab 2: Basic statistical tests in R Homework #1: Share 3 things I learnt today and 1 question I still have.
2	Lab 3: Introduction to R (Basic R functions, Indexing, Sort) Lab 4: Data preparation in R (Merging, Recoding, Web Scraping) Lab 5: Plotting and Advance functions in R (IF and FOR) Homework #2: Using what you have learnt today, find interesting table(s) on Wikipedia, then use R to extract and plot something. Be sure to include plot title and axis labels.
3	Theory 3: Probability Theory 4: k-Nearest Neighbors Theory 5: Regression Lab 6: k-Nearest Neighbors Lab 7: Regression (Simple, Multiple) Lab 8: Decision trees Homework #3: Share 3 things I learnt today and 1 question I still have.
4	Lab 9: Data visualization and dashboarding in Tableau Homework #4: Build a dashboard containing three charts (at least two different chart types) and post to Tableau Public. Save your dashboard as image and create a Medium post inserting the image and your Tableau public dashboard URL.
5	Lab 10: Webscraping with Python Homework #5: Go to IMDb.com> Movies, TV & Showtimes> Most Popular Movies. Scrape Title, Rank, Rating, Advisory category, Run time, Genres, Date. Do a write-up on what kind of analysis can be done including a screenshot of your code.

Table 2. Curriculum of "Budding Data Scientists Hackathon".

C. Final Showdown and Blog

For the final showdown, teams had to do an eight-minute presentation, and a further three minutes were set aside for question-and-answer. Teams were judged based on five criteria, namely understandability, accuracy, creativeness, usefulness of findings, and teamwork.

A blog was set up to share the training materials/ curriculum and document growth stories, learning outcomes, and findings from the hackathon. [2] The blog content could be useful for individuals who are interested in learning about data science.

The projects done by the students included:

- An optimization of taxi services in Singapore
- Predicting traffic volume in the Central Business District of Singapore
- Understanding risk factors for diabetes
- Identifying illegal wildlife trading on an e-commerce platform (done by two teams independently)
- An analysis on wildlife trade
- Planning for Mass Rapid Transit delays in Singapore

3. OUTCOMES AND TAKEAWAYS

Students exhibited interest and enthusiasm throughout the hackathon, and were able to understand and apply most of the content that was taught.

A. Highlights and Effectiveness of Training Session

While most students were exposed to data science for the first time, it was heartening to see that students were interested and enthusiastic during class. Throughout the course, many participants asked questions that reflected their curiosity and desire to learn.

Homeworks #1 and #3 required students to summarise what they have learnt for the day, while Homeworks #2, #4, and #5 required students to find data from the Internet and apply what they had learnt. From the homework submissions, students demonstrated a high level of understanding of the course content, as well as creativity in applying the learnt skills to new data.

As expected, one challenging aspect of the course was teaching the coding component. Coding required students to think in a manner that was different from traditional subjects. The students had various coding backgrounds; some students learnt coding (in other languages) in school, whereas other students were completely new to coding. This resulted in different learning speeds. To solve this, peer-to-peer learning was encouraged. In addition, the faster students were given advanced materials to self-study while the slower students received additional help from the instructors. The use of annotated sample codes, interesting datasets, and examples were also critical to the students' learning.

During the training sessions, it was observed that the topic on statistical tests proved to be challenging for students to grasp. This could be due to a hastened introduction to such concepts without dwelling more on probability distributions and it might be more appropriate to cover this topic after the chapter on probability. In addition, more time should be spent on the general syntax of the Python programming language before the webscraping exercise.

B. Survey on Training Sessions

A survey was conducted to assess the effectiveness of the training sessions (see Figure 1). From the survey, most students agreed that the sessions were engaging, interesting, clear, and easy to follow. In addition, all students agreed that the course material aided their learning, and that the training sessions improved their knowledge of data science.

We also collected qualitative feedback and the following are some responses we obtained:

- The course is enriching and enjoyable and I have a better grasp of basic coding in R and Python after the 5 day sabbatical course. Hope this course could continue for juniors to attend.

- This course has not only helped us in learning the basics in data science, it has also given us a headstart in our project. Learning all of these useful skills will definitely be helpful in the future.
- I felt the course was meaningful and my knowledge of data science definitely improved.
- The course is really an enjoyable and meaningful learning experience for me. Although some content is very difficult and take me quite much time to figure out everything, I think these contents are very useful to both the project and future life.
- This course has not only helped us in learning the basics in data science, it has also given us a headstart in our project. Learning all of these useful skills will definitely be helpful in the future.

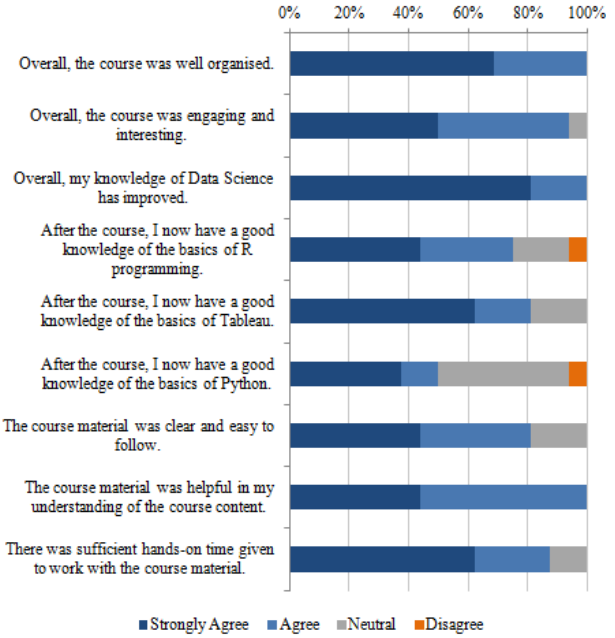

Fig 1. Survey results on training sessions of "Budding Data Scientists Hackathon" (n=16).

C. Highlights of Mentoring and Final Showdown

The five-month time frame for students to work on projects is extremely useful for students to explore and apply the concepts, skills and techniques learned. Students are motivated to put in the effort in these projects to fulfill the requirement of project work grading within the school curriculum, in addition to presenting at the final showdown. They had to balance other academic demands and commitment to the project. The school recognizes that the hackathon presented an opportunity for students to enhance their real-world problem solving skills and interact with working personnel to understand their business challenges. Students learn to recognize constraints in the real world and that there is no one perfect solution to the problem. This is also a platform for the school and students to contribute to society. As with any data science projects, data is critical and many non-profit organizations are not data-ready where data is either not collected in a proper manner, or not at all. Hence, most of the projects in this hackathon rely on public data and/ or APIs made available by the Singapore government. The pilot is successful and the school will be holding a second run in 2019.

4. ACKNOWLEDGMENTS

We would like to thank SIGKDD for establishing the KDD Impact Program and the funding of awards.

5. REFERENCES

[1] ACM *SIGKDD News: Announcing the SIGKDD Impact Program Recipients for 2018.* https://www.kdd.org/News/view/announcing-the-kdd-impact-program-recipients-for-2018

[2] *Budding Data Scientists Hackathon* homepage. https://medium.com/budding-data-scientists

About the authors:

Hui Xiang Chua graduated with a B.Sc.(Hons) in Statistics and M.Sc. in Business Analytics from National University of Singapore in 2012 and 2016 respectively. She is a Data Science for Social Good fellow and has over six years of experience solving problems using data in the public service as a research analyst. Her data science blog was recognised as 2018 Top 100 Data Science Resources on MastersInDataScience.com. (projectosyo.wixsite.com/datadoubleconfirm)

Ee Ling Chua completed her teacher training in National Institute of Education, Singapore in 2004 and subsequently Master's Degree in Education from the University of Western Australia in 2016. She is the Principal Consultant at Hwa Chong Institution for the Mathematics department and has served the school for eight years.

Kenneth Soo completed his MS degree in Statistics at Stanford University in 2017. He is the co-author of the best-selling book, Numsense! Data Science for the Layman: No Math Added, which was written as a gentle introduction to data science and its algorithms. He was the top student for all three years of his undergraduate class in Mathematics, Operational Research, Statistics and Economics (MORSE) at the University of Warwick.