Datathons: An Experience Report of Big Data Hackathons

Craig Anslow, John Brosz, Frank Maurer
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
{canslow,jdlbrosz,fmaurer}@ucalgary.ca

Mike Boyes
Department of Psychology
University of Calgary
Calgary, Alberta, Canada
boyes@ucalgary.ca

ABSTRACT

Hosting successful hackathons is difficult. With the accessibility of open data there is now an opportunity to host hackathons focused on big data, however these are new and not clearly defined about how to do this process. In this paper we present our experience at hosting four big data hackathons called *datathons* that involved students and members from the community coming together to solve challenging problems with publicly open data and data from not for profits. The resources developed for our datathons about our experience will help inform others who also wish to host big data hackathons.

Categories and Subject Descriptors

K.3.2 [**Computer and Information Science Education**]: Computer science education

Keywords

Big Data, Data Science, Hackathon, Open Data

1. INTRODUCTION

Big data is a term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. While open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

We need better tools and techniques for exploring big data and open data. We also need people to develop the necessary skills in order to gain better insight about large data sets that exist within organizations and available through open data repositories. Universities and schools are slowly adopting their practices by introducing courses on data science but these courses are still in their infancy. To address this challenge many community based organizations are running events (workshops and tutorials) to help people acquire the necessary data science skills.

In this paper we present our experience at hosting four big data hackathons called *datathons* that aimed at helping students and members from the community to come together to solve challenging problems with publicly open data and data from not for profits. The resources developed for our datathons about our experience will help inform others who also wish to host big data hackathons.

2. RELATED WORK

Games and hackathons
Stitch fest paper[1]
explain what big data is
explain we are exploring big data hackathons

3. DATATHONS

A datathon is an event similar to a hackathon where people come together over a certain time period, commonly 24 hours to work on problems with a specific dataset. There are a number of steps involved in a datathon including preparing, planning, recruiting participants, arranging an appropriate venue, preparing the data for exploration, logistics, and hosting the event. We describe the benefits of datathons and outline our experience at running four datathons.

3.1 Benefits

3.2 Data For Good Case Study

Data for Good (DFG)[1] is a community organization inspired by DataKind.org and is working for positive social action through "data in the service of humanity". Data for Good in Calgary brings together data scientists with social organizations through a collaborative approach that leads to shared insights, greater understanding, and positive action of data. Data for Good leads a community of data scientists from the community and university to inspire a new way of using the skills and tools of corporations & governments, to meet the needs of the NFP/NGO and social innovation sector.

DDG organizes regular meetups for it's members and hosts weekend datathon events. A DataThon is a weekend event that matches up selected social organizations (that have well-defined data problems) with a team of volunteer data

[1]http://www.meetup.com/Data-for-Good-Calgary/

Submitted to SIGCSE 2016

scientists to tackle their data-related challenges over a 24-48 hours period. The participants are fed throughout the weekend event & the results are presented to the social organizations at the end of the event. Some may refer to these types of events as "Hack-a-Thons", etc. These events are completely free for the participants and serve to energize, educate, and provide direct benefit to the NFP/NGO organizations, as well as to enlighten social sector groups about the power of being data-driven.

Suatainable Alberta Association[2]

Distree Centre[3]

3.3 Canadian Open Data Experience (CODE) Case Study

The Canadian Open Data Experience (CODE)[4] is an intense 48-hour coding sprint where innovators from coast to coast compete to build the best app utilizing federal government data from Canada's Open Government portal. We organized two hubs as part of the CODE competition in 2014 and 2015.

In 2014 four teams participated with four members per team. One team made the grand final of the top 15 teams and presented their app in front of judges in Toronto.

4. DISCUSSION

There are a number of aspects that we would like to draw upon based on our experience at running big data hackathons over a two year period.

Participants. recruitment, group dynamics, community engagement

Data. structuring and cleaning, open data, storing

Tools. Participants used a variety of tools to solve the problems as part of the datathons. Some used statistics focused tools such as R, SPSS, and Matlab. While others used some visualization tools such as Tableau and Excel. While more participants who were developers used toolkits like D3 to develop visualizations.

Logistics. room setup, break out rooms, food

Things to avoid. Going all night, Grouping all technical people into one team

5. CONCLUSIONS

Acknowledgments

Thanks to the Government of Canada, City of Calgary for providing open data sets. Thanks to the Data For Good Calgary Meetup group for assisting with the datathon event, in particular Geoff Zakaib and Kathryn Winkler. Some of this research was funded by Mitacs.

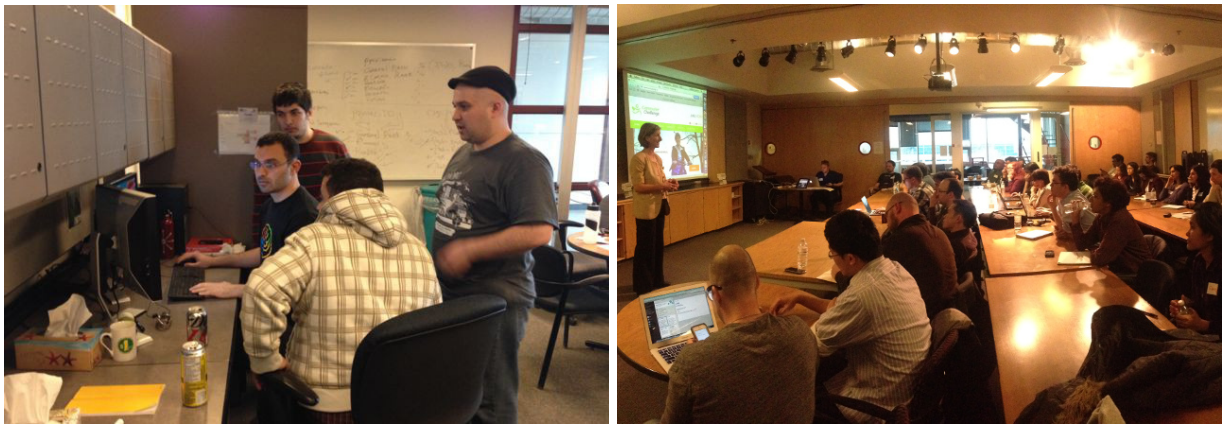
6. REFERENCES

[1] G. T. Richard, Y. B. Kafai, B. Adleberg, and O. Telhan. Stitchfest: Diversifying a college hackathon to broaden participation and perceptions in computing. In *Proc. of the Symposium on Computer Science Education (SIGCSE)*, pages 114–119, New York, NY, USA, 2015. ACM.

[2]http://www.meetup.com/Data-for-Good-Calgary/events/175671942/

[3]http://www.meetup.com/Data-for-Good-Calgary/events/221869098/

[4]https://www.canadianopendataexperience.ca/



(a) CODE 2014 - Brainstorming Session. (b) Data For Good 2014 - Introductory Session.

Figure 1: Datathons Case Study - illustrating different aspects of the datathons.