The Data Science Incubation Program

Submit a proposal for Spring Quarter (https://catalyst.uw.edu/webq/survey/dhalperi/225163)

The goal of the Data Science Incubator is to enable new science by bringing together data scientists and domain scientists to work on focused, intensive, collaborative projects to enable new science through the development of new techniques and technologies or the application of existing techniques and technologies in new ways. Projects frequently, but not exclusively, involve a non-trivial software engineering component. Our team of data scientists can provide expertise in state-of-the-art technology and methods in large-scale data manipulation and analytics (e.g., Hadoop, GraphLab, Myria, SciDB), cloud and cluster computing, statistics and machine learning, and visualization to help researchers extract knowledge from large, complex, and noisy datasets.

Overview

To apply to the program, any faculty, research staff, or student (typically, but not exclusively, at UW) can submit a short project proposal (details below) describing the science goals, the relevant datasets, and the expected technical challenges. Each project will also identify one or more researchers willing to physically co-locate with our team for at least 2-3 days a week for the duration of the project (typically three weeks to three months, and up to twelve months). We find that collaboration in a shared space is important for deeper technical engagement and provides opportunities for "cross-pollination" among multiple concurrent projects. The pilot program will operate out of the eScience space in Sieg 326, moving to a new Data Science Studio planned to open Fall 2014. Incubator projects are not "for-hire" software jobs -- each project will be led by the representatives of the applicant's team working in collaboration with the data scientists and the broader eScience community.

Areas of Focus

Each project will be different, but we emphasize projects in the following categories:

Scalable Analytics:
As data sizes continue to explode, parallel methods have become critical at every step. Scripts in Python and R are not natively parallel and are difficult to apply to datasets larger than main memory. Our team can help triage your problem and adapt it for use with parallel data manipulation and machine learning platforms such as Hadoop/MapReduce, parallel SQL databases, GraphLab, SciDB, and advanced research systems such as UW's own Myria. We also design and implement new parallel algorithms for large datasets independent of existing platforms.

Data Management and Automation:
Our collaborators report spending 90% of their time "handling" data as opposed to analyzing data: data discovery, acquisition, file format conversions, cleaning, restructuring, loading, sharing, etc. Leveraging technology from cloud providers and SQLShare, we aim to simplify or eliminate these data manipulation tasks and let researchers focus on the science.

Visualization:
We have experience building data-driven visualizations to help scientists make sense of data. We focus on web-enabled, interactive visualizations using platforms like D3.

Reproducibility and Open Science:
We can help you share your code, data, and results with collaborators and with the general public. We favor projects that emphasize open data and open source, allowing other researchers to recreate your results with minimal effort. We advocate alternative metrics and can help you maximize recognition and credit for ensuring reproducibility and open access. Suitable incubator projects may include organizing and uploading data into suitable public repositories, reviewing and publishing your code on GitHub, identifying venues for publishing papers describing your data or code (they exist!), or migrating your application to a commercial cloud to improve access.

We structure our work according to agile methodologies (http://agilemanifesto.org/), typically breaking large projects into multiple short-term sprints of a few weeks each.

Success Stories

Our team has a strong track record of building systems that get real use. Below are listed some of our previous collaborations.

- *Cruise telemetry dashboard for real-time analytics* (Armbrust Lab (http://armbrustlab.ocean.washington.edu/)). View the website (http://uwescience.github.io/seaflow-viz/km1314/).
- *Experiment to re-express bioinformatics workflows in pure SQL* (Roberts Lab (http://faculty.washington.edu/sr320/)).
- Migration of R scripts to database queries (Armbrust Lab (http://armbrustlab.ocean.washington.edu/)). Read the CiSE journal article (http://homes.cs.washington.edu/~dhalperi/pubs/CS_CiSESI-2012-10-0093.R1_Howe.pdf).
- Interdisciplinary environmental data integration (Armbrust Lab (http://armbrustlab.ocean.washington.edu/)). Read the SSDBM conference paper. (http://homes.cs.washington.edu/~dhalperi/pubs/halperin_2013_ssdbm_geomics_case_study.pdf)
- Parallelization of the InfoMap algorithm (http://mapequation.org/) for graph-based community detection (with Martin Rosvall and Jevin West).
- Migration of computationally intensive animation rendering tasks to the cloud (with Barbara Mones and Stephen Spencer in CSE).
- Studying protein/protein interactions (https://github.com/uwescience/sqlshare/wiki/Interolog-queries-in-sql) with SQL (Kimmen Sjolander and Cyrus Afrasiabi at Berkeley).

You can learn more by reviewing the slides (incubator_info_session_2_20_14.pptx) from the information session from February 2014.

How to Get Started

Do you have interesting data challenges? You can submit a project proposal (https://catalyst.uw.edu/webq/survey/dhalperi/225163) for Spring quarter. Proposals should include:

- Contact information for the project lead -- the one who will join us in the studio and be responsible for carrying out the project.
- Project summary / objective (1 page).
- A description of your data. At least the size, formats, where the data currently resides, and any privacy and access restrictions. We strongly favor projects that have already collected the relevant data rather than "preparatory" projects that involve building software in the anticipation of future data collection activities.
- A list of the key science questions the data will help answer, and a discussion of the publications that you anticipate resulting.
- A list of key technical challenges you face in answering these questions: Do you need new methods or algorithms? Do you need to scale up existing methods? Do you need to integrate data so it can be analyzed? Do you need to publish data and/or code to improve collaborative opportunities and reproducibility?
- The timeframe for your work.
- The names of those researchers who will be physically joining the team to lead the project.

These proposals are prioritized based on the following criteria:

- Good clustering between proposals; ideally, we seek a cohort of proposals with a common theme.
- Alignment with sponsor and program goals
- Participant availability and engagement
- Ability to answer fundamentally new research questions
- Clarity and shovel-readiness
- Capacity for measurable outcomes
- Capabilities of the incubator staff

We expect that some good proposals will not meet every criteria.

Important Dates for Spring 2014 Session

- 2/20: Incubator information meeting
- 3/10: Proposals are due
- 3/14: Follow-up requests sent
- 3/21: Pilot participants notified
- 3/31: Spring launch
- 4/21: First project milestone
- 5/12: Second project milestone
- 6/2: Third project milestone
- 6/6: Poster / networking event


GORDON AND BETTY
MOORE
FOUNDATION
ALFRED P. SLOAN
FOUNDATION
NSF