Research Data and Computational Training Pilot Proposal

Summary

August 18, 2022

Purpose: Members of UBC Advanced Research Computing (ARC), UBC Okanagan Library, and Research Computing (RC) recommend that a pilot Research Data and Computational Training program be sponsored by their respective portfolios during the period of FY23-24. The objective of this pilot would be to evaluate whether a comprehensive course can be designed and implemented in a way that is sustainable and scalable and provides graduate students with foundational data, analysis, and coding skills, which are currently not offered.

Background

Data, analysis, and coding skills are progressively more important in academic work. However, most UBC graduate students do not have substantive understanding or skills in these areas, particularly in disciplines where computational research is still emerging.

Currently, UBC ARC and UBC Library offer training in the domains of data, coding, and analysis, however, these are one-off and generic in their approach. This can be challenging for new learners; they omit scaffolding and often neglect skills needed to move from the generic to practical research applications. This was noted by UBC researchers in the UBC DRI Needs Assessment:

> *"I attended some intro sessions last summer, but it was still too big of a jump from using a personal desktop to Sockeye or Compute Canada resources."*
>
> *"Sometimes the ARC training is very technical and unapproachable for students new to stats and HPC."*
>
> *"I think resources for… the direct application of the coding language within your environment and within your field is the gap that really prevents people from being able… to implement these skills at that higher level"*
>
> *"I took workshops, but the only ones that actually stuck were when I was actually doing work that was relevant [to my research]"*

Training Pilot Description

UBC ARC, UBC Library and RC share significant overlap in support during certain phases of the research life cycle. This pilot will develop and implement a comprehensive data, analysis, and coding curriculum, following the research data life cycle where this overlap exists, leveraging existing instructional resources as pick up and pass off points. This opportunity will provide graduate students with practical and adaptable research skills.

The proposed audience is a cohort of 10-12 students from UBCO. The model leverages known successful models from a handful of locations – UC Bolder and Duke University – described in more detail below. It also maps to a known successful model at UBCO - the Centre for Scholarly Communication's (CSC) Writer's Retreat - a multi-day intensive writing seminar for graduate students.

Rationale & Benefits

In order to ensure that UBCO attracts top-tier graduate students and provides a trajectory for these students to develop innovative research skills allowing them to be competitive in an academic job market, UBCO needs to provide robust supplementary training opportunities. UBC ARC, RC, and UBCO Library are in a unique position to provide foundational data, analysis, and coding training, bridging expertise and services. In line with the primary findings of the UBC DRI Needs Assessment, a collaborative service model will allow the institution to use its resources more efficiently and effectively than a siloed approach does.

Pilot Outline

To balance discipline specific needs against the lack of feasibility of running workshops for every discipline, we are proposing a three-stream approach, mirroring the Tri-Agency funding streams. The pilot would straddle aspects of these streams to maximize disciplinary scope for the purposes of addressing feedback and sustainability; for example, integrating in both qualitative and quantitative data making it relevant to a breadth of disciplines. Students would have a choice of preselected data sets to explore.

The format would be a one-week, in person, intensive course including didactic and collaborative learning. Each day would be split into two 2.5-hour blocks, the first block being used to introduce concepts and tools, the second block being used for collaborative problem solving with these concepts and tools. Over the course of 5 days, students would work through a miniature research question – establishing a question, identifying a data source, setting up a plan, acquiring, cleaning, modeling, and visualizing the data.

The pilot would be situated to allow learners to pursue more advanced opportunities, such as ARC specific training, and for the future potential of allowing faculty and labs to request tailored subsets of the curriculum for their students. In the assessment section we discuss the potential for future prospects of methods of delivery other than in person.

Key Objectives and their mapped Outcomes include:

Objective	**Outcome**
Acquire foundational data literacies	Describe common data types and structures used in computationally supported research and how choice of data structures impacts computational resources.

Demonstrate advanced data literacies	Discuss issues and concerns related to cross walking, data type conversions, and combining data sets of different origins with particular attention to concerns that arise in computational environments.
Engage in transparent, reproducible data centric research	Articulate the importance of data provenance and describe systems and meta data approaches to recording the provenance of information.
Gain awareness of data licensing and re-use	Be familiar with common licensing schemes used for data and software and articulate nuances related to data ownership and sovereignty.
Acquire core data science skills	Describe the data science workflow of ask, acquire, explore, model, and communicate and visualize.
Apply core data science computational skills	Demonstrate effective use of computationally reproducible tools to effectively acquire, explore, model, and communicate and visualize data.

Resource Analysis

January-March, 2023

- Key tasks:
    - Conduct environmental scan of comparable programs
    - Outline curriculum and begin building content
    - Participate in weekly data and computational consultations with the CSC
- Requested resources:
    - 2 graduate students
    - 12 hours/week @ $40/hour (wage of $33/hour + 20%)
    - Total: $11,520

April-August, 2023

- Key tasks:
    - Build out curriculum content to completion
    - Develop and execute communication strategy and guided recruitment
    - Deliver curriculum
    - Participate in weekly data and computational consultations with the CSC
- Requested resources:
    - 2 graduate students

    - 12 hours/week @ $40/hour (wage of $33/hour + 20%)
    - Total: $19,200

September-March, 2023

- Key tasks:
    - Assess program via learner feedback
    - Determine recommendations and next step based on assessment
    - Participate in weekly data and computational consultations with the CSC
- Resources requested:
    - 2 graduate students
    - 4 hours/week @ $40/hour (wage of $33/hour + 20%)
    - Total: $8,960

Total resources requested: $39,680*

*$11,520 fiscal 22/23; $28,160 fiscal 23/24

Implementation Considerations

In support of this work, the pilot will reference similar training models implemented at the University of Colorado Boulder (UCB), as well as Duke University.

UCB's Center for Research Data & Digital Scholarship (CRDDS) is a collaboration between Research Computing and the University Libraries, that offers a wide variety of data services and supports to UCB community. With numerous areas of expertise in their team, the CRDDS has developed an extensive catalogue of training offerings, and encourages classrooms and labs to book seminars of specific interest to their research and work.

Duke University's Innovation Co-Lab Training offers an extensive catalogue of digital skills, and organizes its courses into tracks in which students work through several workshops to get a deeper dive into a topic. Courses are also offered outside the track model.

This pilot seeks to leverage aspects of both models and early model development for the pilot would include meeting with representatives from UCB and Duke to explore lessons learned and guidance on an initial implementation.

Assessment

Pilot assessment will consider four elements: a) completion and delivery of the curriculum (was it manageable); b) qualitative feedback from the student cohort using a pre- and post- training assessment tool (quality of course and were skills gained); c) qualitative feedback from students hired to help with curriculum development (is it scalable); d) qualitative feedback and quantitative use of consultation services.

As a pilot format, an in-person model will facilitate assessment. Feedback from participants, designers, and relevant faculty will be used to explore alternative modalities of delivery,

whether that be virtual or asynchronous. This will also provide opportunity to consider how we cater learning outcomes and objectives to specific mediums of delivery, as well as how potential accreditation – if this was pursued - would be impacted by method of student engagement.

Sustainability & Collaboration

Of key importance to this work is its sustainability and its ability to scale to a larger or to a diversity of programs. Sustainability will be informed by the assessment process. If deemed successful -- manageable, skills building, scalable -- the pilot would be well situated to inform an ask for ongoing funding that would be shared between UBC ARC, UBCO Library, and RC, to broaden course development, delivery, and assessment over a longer period.

The collaborative model is inspired by UC Boulder's approach, which operates as a partnership with equal division of funding between their library and research computing teams. Using such a collaborative approach, we leverage expertise and complementary approaches to problem solving, outreach, and engagement based on respective exposure to researchers at various stages of their research careers and when in the research life cycle they might reach out to the Library, RC, or ARC. This collaborative approach also helps to limit redundant offerings, where multiple units might offer similar workshops, allowing increased streamlining of supports and services.