Course Proposal:
Analytical Workflows
IB 5XX

Mark Novak & Ben Dalziel

Course Motivation

Graduate students are amassing ever larger datasets to ask ever more challenging questions with ever more sophisticated statistical techniques. Many others propose dissertation chapters entailing computationally demanding simulations or the analysis of parameter-rich mathematical models. Keeping pace in their field often requires students to straddle empirical, statistical and theoretical techniques. These multifaceted and ever more collaborative approaches lie at the heart of our department's *integrative* approach to biology and have become standard practice in many of the career paths our students pursue. And yet, nowhere in our curriculum are students provided the opportunity to learn the skills, basic tools and best practices with which to manage the complexity of their projects.[1] This poses a non-trivial challenge to our students that is fundamentally distinct from the learning of specific content, methods or platforms.

Our course provides students the skills and tools they need to organize and manage their projects from inception to publication. Working with their own data or models, students learn and practice: (re)organizing projects into efficient, reproducible and easily-modified "*analytical workflows*"; writing modular, transparent computer code in the programming language of their choice; and tracking and communicating progress and hurdles to others using stand-alone and cloud-based version control and collaboration software. Although these have proven to be game-changing for graduate students at all levels, the target audience for this course is students who have substantial data in hand or have an implementable framework for their analyses established (i.e. 2^{nd} to 4^{th} year Ph.D. and 2^{nd} to 5^{th} term M.S. students).

Student-targeted Advertisement

Have you proposed a modeling chapter for your dissertation but need support getting things up and running? Will you soon be sitting on a complete data set ready for your planned analyses but don't know how or where to begin? Maybe you're far along in a series of analyses and feel "lost in the trees."

This course will help you with these challenges by practicing the development and implementation of efficient, reproducible *workflows* for your projects. Every project should (and can) be modular and fully automated, hence reproducible, portable and easily modified. Rerunning an analysis with a different set of parameters should (and can) be as simple as a few keystrokes. Regenerating all figures and tables for your manuscript after finding a typo in your code or dataset should (and can) be painless.

Efficient workflows start at project conception and end only if the project idea is itself a dead end. Thus, in this course, we'll work to practice (1) refining and articulating project goals and benchmarks, (2) creating modular and automated analyses, and (3) using best practices in coding and project management. We'll learn how to use Git, GitHub, LaTeX, Markdown, and High Performance Clusters. The instructors will mostly use R within RStudio, but users of other programming languages and text editors are welcome and encouraged. You will need either (1) a dataset and a visualization or analysis goal, or (2) a model or simulation (or sufficiently well-developed ideas for one). The use of other people's data or published models is also encouraged, as needed.

Course Catalog Description

Theory and implementation of efficient, reproducible workflows – including best practices in scientific programming, project management, and collaboration – for computational, analytical, and data-driven biological research.

[1] See *Related Courses at OSU* below.

Student Testimonials

We have taught two prior IB 599 versions of this course: Spring 2019 (2 *cr.*) & Fall 2020 (4 *cr.*). Students ranged from 1^{st} year M.S. to 5^{th} year Ph.D. from across three colleges. Student eSET scores were 6.0 and 6.0 in 2019 and 5.8 and 5.8 in 2020 for "the course as a whole" and "the instructor's contribution" respectively[2]. The following are excerpts of the feedback that was received via email and the eSET evaluations.

> *I am of the opinion that this Analytical Workflow class is indispensable. Engaging in research comes with a massive deluge of papers, files, data, output, etc, and yet prior to this class I had never before been exposed to organizational best practices recommendations.*
>
> *I started this course with nothing but raw data and anxiety, and I am leaving with code that works, a reproducible workflow, and the confidence and resources to press on.*
>
> *I'm so happy this class exists, and it was instrumental to much of the progress I've made this term.*
>
> *Ben and Mark helped turn a daunting task that I'd been putting off [...] into a well-organized reality! Five stars.*
>
> *This class has been extremely useful for the conceptual organization and technical execution of my research. Thank you for organizing and leading this class!*
>
> *Who should take this course? Anyone working on a project that uses coding or modeling.*
>
> *I really appreciate that you shared your experience and learning process with us (metacognitive reflection! best teaching practices!).*
>
> *After taking the class, [...] I feel empowered to solve issues with my model on my own.*
>
> *It was reassuring and encouraging to be able to share feelings and experiences with my colleagues and have them validated, listen to their counsel, and know that I am not the only one facing these challenges.*

Details

Credits:	4
Frequency:	Annually
Quarter:	Winter
Course times:	Tuesday & Thursday 10:00-11:50am
Instructors:	Mark Novak & Ben Dalziel (of record alternating annually)
Prerequisites:	Graduate standing, or by instructor permission
Enrollment cap:	18

Learning Outcomes

After successful completion of this course, students will be able to:

1. Translate a research plan into an explicit analytical workflow;
2. Apply best practices in scientific programming to construct reproducible research;
3. Manage and collaborate on complex research projects using a version control system;
4. Apply analytical workflows to advance their dissertation research.

Course Philosophy

Our primary goal in this course is for students to develop more efficient research skills. An important secondary goal is to have students make significant progress on their thesis work. Our philosophy is that students can achieve both because our primary goal is best achieved by having students practice new tools while working on their own research.

[2]By contrast, the department's median scores were 4.9 and 5.1 in Spring 2019, and 5.1 and 5.4 in Fall 2020, out of 6.0 points possible.

Course Materials

For complete access to all teaching materials and learning resources, see https://github.com/analyticalworkflows/TeachingMaterials.

Schedule

Week 1: Course overview & Philosophy
Structuring projects & Version control with Git

Week 2: Project proposals
Workflow diagrams

Week 3: Coding best practices
Hack-a-thon

Week 4: Git w/ GitHub (*Project management & collaboration*)
Hack-a-thon

Week 5: Typesetting with Markdown
Data visualization

Week 6: Project progress presentations
Hack-a-thon

Week 7: Faster computing (*Vectorization & parallel computing*)
Hack-a-thon

Week 8: Faster computing (*High performance computing*)
Hack-a-thon

Week 9: Typesetting with LaTeX
Hack-a-thon

Week 10: Project presentations
Project presentations & Wrap-up

Related courses at OSU

OSU's relevant statistics courses focus exclusively on data visualization. Existing omics and informatics courses introduce students to specific analysis pipelines in particular programming environments and do not address issues relating to project management as a whole. Existing courses are also not relevant to the many IB students pursuing non-*omic* research.

BB 485/585 - Applied Bioinformatics
Fundamental concepts needed to understand the software and methods used in bioinformatics. Includes contemporary techniques such as databases, gene and genome annotations, functional annotations, sequence alignment, motif finding, secondary structure prediction, phylogenetic tree construction, high-throughput sequence data, ChIP-Seq peak identification, transcriptome profiling by RNA-Seq, microRNA discovery and target prediction. *Prerequisites*: BB 314 or BB 314H.

ST 537 - Data Visualization *E-campus only*
Perceptual principles for displaying data; critique and improvement of data visualizations; use of color in visualization; principles of tidy data; strategies for data exploration; select special topics. *Prerequisites*: ST 512 or ST 517 or ST 552.