Columbia Biostatistics Computing Club Seminar: How to Set Up your Computer as a Statistician - 9/11/2017

Meeting Notes & Summary

At this seminar, Professor Jeff Goldsmith gave us a tour on how he organizes his computer and files, then there was a panel with MS and PhD biostatistics students (Yutao Liu, Julia Wrobel, Brady Rippon, and Karissa Whiting) who shared their unique perspectives, tips, and tricks on organizing code and files, what applications they find essential, methods to share large data files and collaborate on projects, and more.

Jeff's Presentation

Setting up laptop

What kind to get: get whatever you're comfortable and happy with!

What software:

- Download most recent version of R and RStudio (Jeff does all statistical computing in R)
- LaTeX – download full LaTeX distribution or use online platform like Overleaf
- Word processor
- Dropbox
- Git client (e.g. SourceTree)

Organization

Collaborative Projects Folders (Stored on DropBox) – Separated by Year (e.g. YYYY --> *YYYYMM_CollaboratorProject*)
Everything for that project inside of that folder, with each project folder organized the same way (subfolders for Drafts, Application (Code- RProject inside), Results, Response, and Reviewer comments. He uses same naming structure to organize emails for the project.

He has a completely separate data folder on local computer that contains datasets he's working with (no data on Dropbox especially if you're working with patient data).

Sharing Data with collaborators

- USB drive for collaborators at Columbia
- P drive – shared and secure across Columbia. Gives restricted collection of file access. Can go through process to get secure access (Jeff prefers to have a local copy)

Other Advice

-If you use Mac, use the stuff! Mail, Calendar
-Use spotlight to quickly navigate (Windows equivalent Cortana)

Question and Answer with the Panel (Yutao Liu, Julia Wrobel, Brady Rippon, Karissa Whiting, Jeff Goldsmith)

Q: "How to keep track of different versions of analysis in different drafts?"

Jeff's answer:

- He used to add dates to files… that got messy
- Now uses Git- internal repository on your own computer (keeps everything local and private, but with version control). Hosting on GitHub makes your code public (can get free academic account for unlimited private repositories though)
    - Git is method for version control (check out resources page of our website for slides from last year's Computing Club seminar on *Using GitHub for collaboration and version control*)
    - If data changes: best case scenario is you can rerun your code with the new data, worst case is you may need to start a new script and start analysis over (e.g. if new variables were added or data significantly changed).

Q for Yutao: "What do you need to use the computing cluster?"

Yutao's answer:

Cluster has a huge number of computers that will run a lot of stuff at the same time. If each simulation takes 20 minutes, and you have 100, if you can run them at the same time it will improve computing time.

- Command line interface "Here's code, here's how many times I want you to run it" and it spits something else
    - Can create one job and multiple tasks per job.
    - Specify which program, which version, how much memory you need, how much time, specify paths to R libraries
- Your laptop may even have multiple cores! (Karissa's Macbook has 4 physical cores)- you can test doing parallel computing using your own laptop with multiple cores.
    - R package DoParallel, can use foreach
- Overall, practice UNIX/Shell/Terminal script, this will be essential if you need to use the computing cluster for any projects.

Q for Brady "Can you tell us about using LaTeX?"

Brady's answer:

- LaTeX is like Word, but for math. It's the preferred way to write a document that is math intensive
- MacTeX is mac version, MicTeX for Windows, but Brady's all about Overleaf which is the online interface
    - LaTeX has a harsh learning curve but Overleaf is very user friendly. In LaTeX, you do your code and convert it to PDF… but won't convert if there are any errors. Overleaf updates in real time, with an Editor pane and real-time updating PDF pane. It tells you exactly where errors are.
    - Overleaf is also like a DropBox- all your documents in LaTeX are saved in your personal folder. You can also collaborate on documents using Overleaf.
    - Overleaf has templates that you can use to start you off

- Yutao: Overleaf has also online repository for packages. If you're working locally on LaTeX (MacTex or MicTex), then you need to download and keep packages up to date.

Q: "RMarkdown vs LaTeX?"
If you're going to use RMarkdown for typing up math/equations, it also uses LaTeX in the background.

- RMarkdown does not look AS nice as something in Overleaf or a TeX program on your computer. But it's easy to integrate analysis and visuals, so great for homeworks or things you don't care as much about how it looks.
  - The Panel recommended LaTeX for things like research papers, articles etc.

Q: "How do you organize your references?"
Jeff & Julia:

- bibteX file for references. Can put that into LaTeX or RMarkdown.
  - From Julia: natbib is the latex library that allows you to integrate bibtex files (which are stored as .bib) with .tex documents. This link provides more information to those curious.
- Handy way to output references and knit them with written document.
- Jeff has one bibtex file with all different sources
- Exports citations from Google Scholar or wherever

Q: "How do you keep track of all the papers you've read?"

- Julia: Readcube (highest recommendation)
- Jeff kind of recommended: Mendeley (GUI where you put your stuff in)
  - Can also have it autoupdate your bibtex file
- Zotero (iffy suggestion)

Q: "How do you collaborate with people remotely especially when using different programming languages?"
Karissa's answer:

- When collaborating with people who use different languages, GIT IS THE MOST IMPORTANT
  - you actually do the collaboration through GitHub, which stores the code online, and under the hood GitHub uses git for version control.
- Also uses Slack- messaging software integrates with git and github
  - Pushes repository update to shared slack channel to update your collaborators

Karissa only uses open source data so private data isn't an issue for her
But you can use Gitignore – keeps things locally, if you don't want it on GitHub

- Jupityr notebook- interactive coding environment

Q: "For data visualization: ggplot vs Tableau?"

- Check out presentation on ggplot2 from last year's Computing Club resources for visualization in R
- Tableau is user friendly for quick graph, but not as strong analysis as SAS or R