Lecture 1:
Course Introduction, Setup, Git

LING 1340/2340: Data Science for Linguists
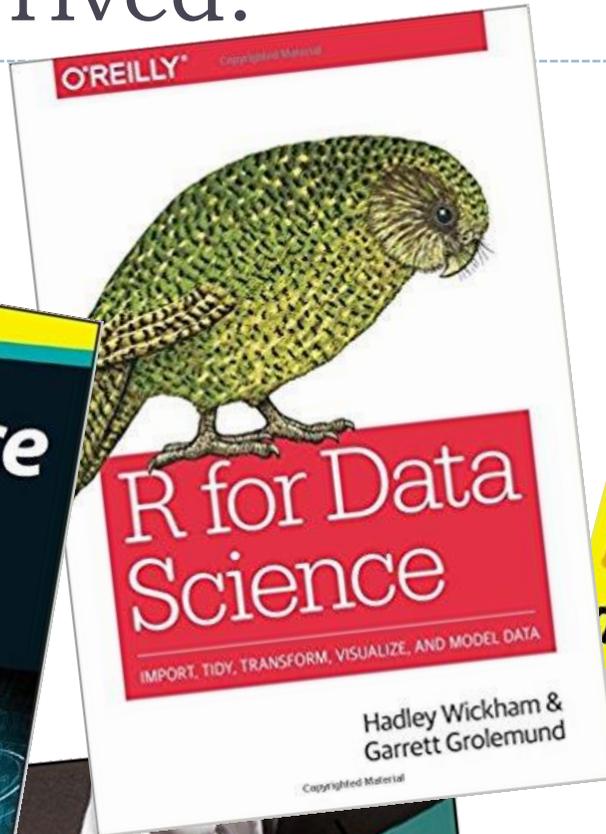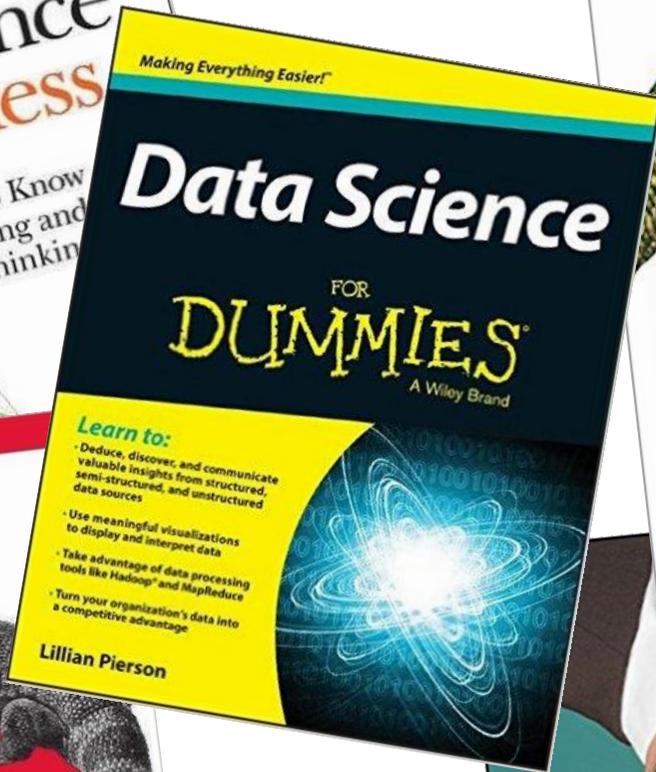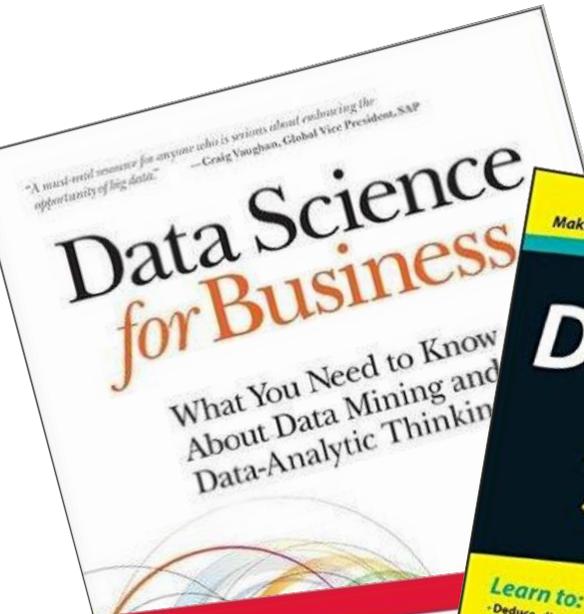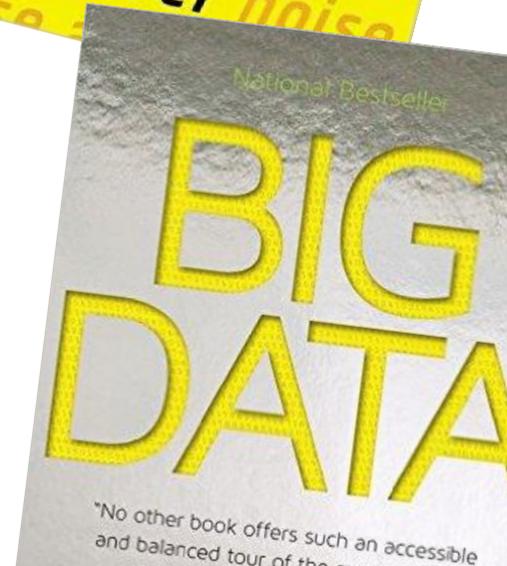
Na-Rae Han

Objectives
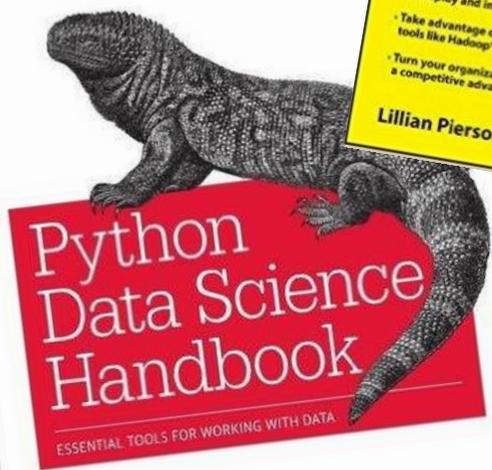
▸ Introduction: What is Data Science?

▸ Scope of this course

▸ Course logistics and policies

▸ Tools: Git

Data Science has arrived.

What is "Data Science"?

▸ Is it about doing *science* with *data*?

▸ Well, there's more than that. In its contemporary incarnation, Data Science implies:

   ◆ *Big* data.

   ◆ Data in the wild exists in organic, unstructured form. It's the job of "data scientists" to source, sift through, clean, organize, and finally make some sense out of it.

   ◆ Predictive modelling, using statistics and machine learning methods.

   ◆ Story-telling and visualization.

What data scientists do

"More than anything, what data scientists do is make discoveries while swimming in data. It's their preferred method of navigating the world around them. At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. They identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set. In a competitive landscape where challenges keep changing and data never stop flowing, data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data."

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Data Science for Linguists (1)

▶ We linguists have always been doing "science" with "language data". Our methods are analytical.

▶ We are therefore uniquely positioned to:

  ◆ add linguistic knowledge to raw language data through annotation
  ◆ plan, develop, and manage language data in a scientific way

  ◆ bring our data practices up-to-date, to be in line with current trend & standards in data-intensive research and industry

← Our role: "curator" of language data

Data Science for Linguists (2)

▶ We also need to keep pace in the era of "big data".

- Language data too is going *organic* and *big*.

- We can't just leave interpretation to computer scientists and statisticians with no linguistic training.

- Basic competency in **statistics**, **machine learning**, and **computer science** is a must.

← Our role: "user" and "interpreter" of language data, from a domain expertise vantage point

Linguists make fine Data Scientists

▸ Recent graduates have been hired for "Data ..." positions.

▸ More and more tech companies are hiring for "Data..." positions.

Courtney Z
Data Scientist at Management Science
Management Science Associate
Pitt

1st
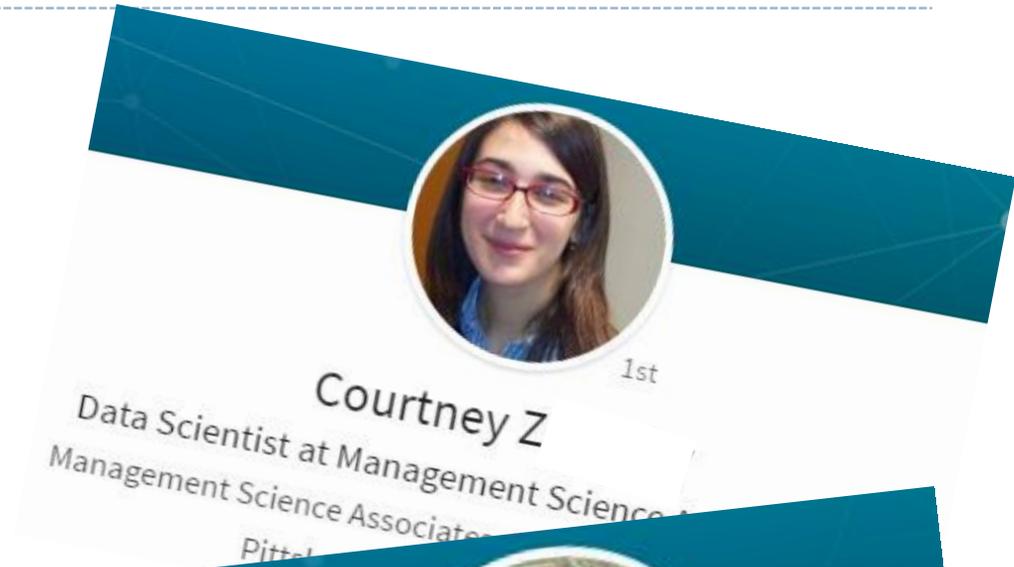
amazonjobs

Data Associate

Job ID: 511414 | Amazon Corporate LLC

Amazon is seeking a Data Associate to join our data team.
This role focuses on speech and language data, primarily in
the areas of transcription, text annotation, and general data
analysis deliverables. The Associate must be capable of:

· Transcribing and

Fawn D

1st

Data Specialist Manager at Amazon

Amazon • University of Pittsburgh

ter Boston Area • 309

Doing data science

▶ [Python](#) and [R](#) have emerged as the two primary tools of the trade.

▶ Statisticians are more likely to use R, computer scientists Python.

▶ In this class, we will work with **Python**.

  ◆ Our department has an R version for grad students, taught by Dan Villarreal: LING 2020

Review syllabus

▶ Course home page doubles up as syllabus:

♦ https://naraehan.github.io/Data-Science-for-Linguists-2025/

Learning expectations

▶ Unlike in LING 1330/2330, much of learning in this class will be open-ended.

- There will be less hand-holding from me.
- You should be in charge of your own learning.
- I will be pointing you towards many external learning resources, whose scope might exceed your immediate needs.

▶ In Ling1330/2330, "code producing correct output" was a good target. Not in this class. Your code will be evaluated upon:

1. Working correctly
2. Computational efficiency and elegance
3. Presentational: documentation and readability

Where is *everything?*

Content and activities	Where	Authors	Access
Course home page, syllabus, lecture slides, homework assignments	https://naraehan.github.io/Data-Science-for-Linguists-2025/	Instructors	Entire world can view
Your grades, announcements	Canvas	Instructors	Instructors and you
Forum, chat, office hours	MS Teams	Everyone in this class	Everyone in this class
Data distribution, assignment submission	GitHub org repositories (private)	Everyone in this class	Everyone in this class
Your term project	Your own GitHub org repository (public)	You	Entire world can view

Additionally: online video tutorials (DataCamp), textbooks, CRC

Your work, open access, and privacy

▸ Coding has become open & social.

▸ The spirit of open access, sharing, and collaboration is very much part of data science's central tenets.

▸ Your homework submissions will be shared with class.

▸ Your term project must remain public throughout the semester.

System setup

You have already installed:

▶ A text editor

  ◆ Notepad++, Sublime Text, VS Code… (what else?)

▶ Git

▶ Anaconda Python (Python version 3.12 or 3.11)

  ◆ Jupyter Notebook

System-side house keeping

▸ Decide on a folder where you will keep your course work.

  ◆ **Mac** recommendation:

    `/Users/narae/Documents/Data_Science`

  ◆ **Windows**

    Recommended: `C:\Users\narae\Documents\Data_Science`

    Also good: `C:\Users\narae\Data_Science`

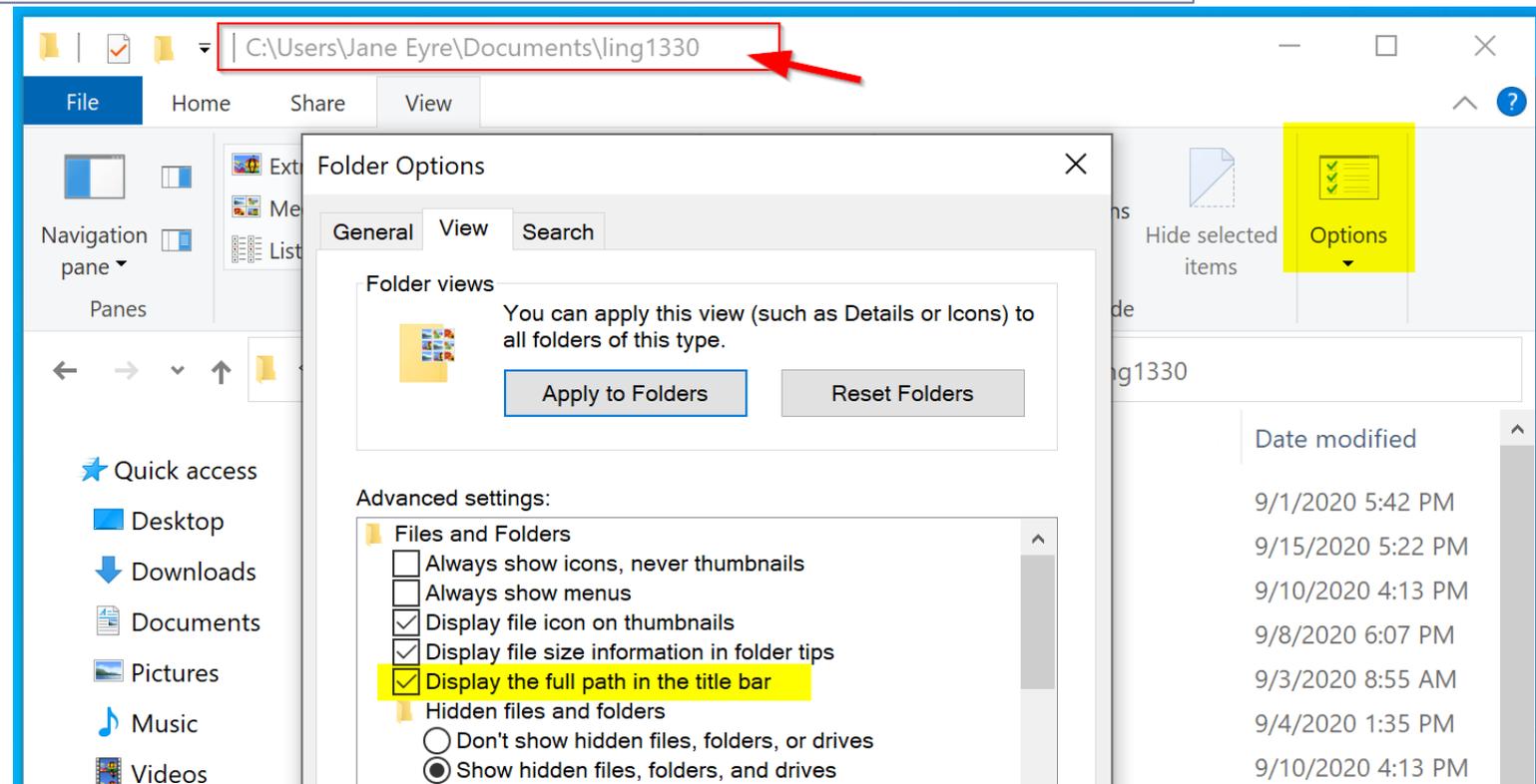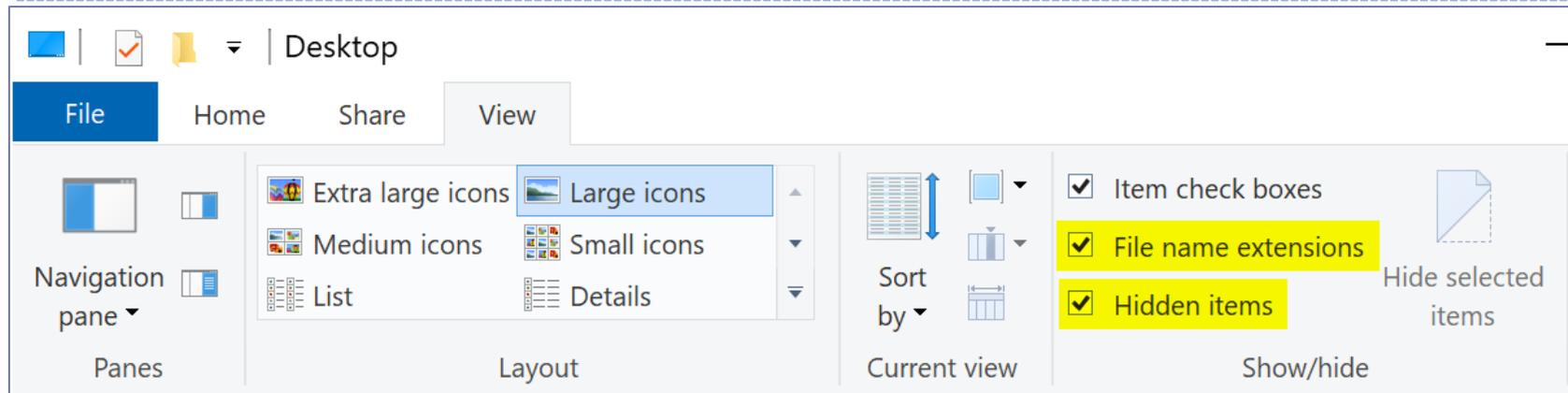   BAD ➔ `C:\Users\narae\`**`OneDrive`**`\Documents\Data_Science`

    ← It's a really bad idea to have your Data_Science folder live inside a **cloud-synced folder** like **OneDrive**.
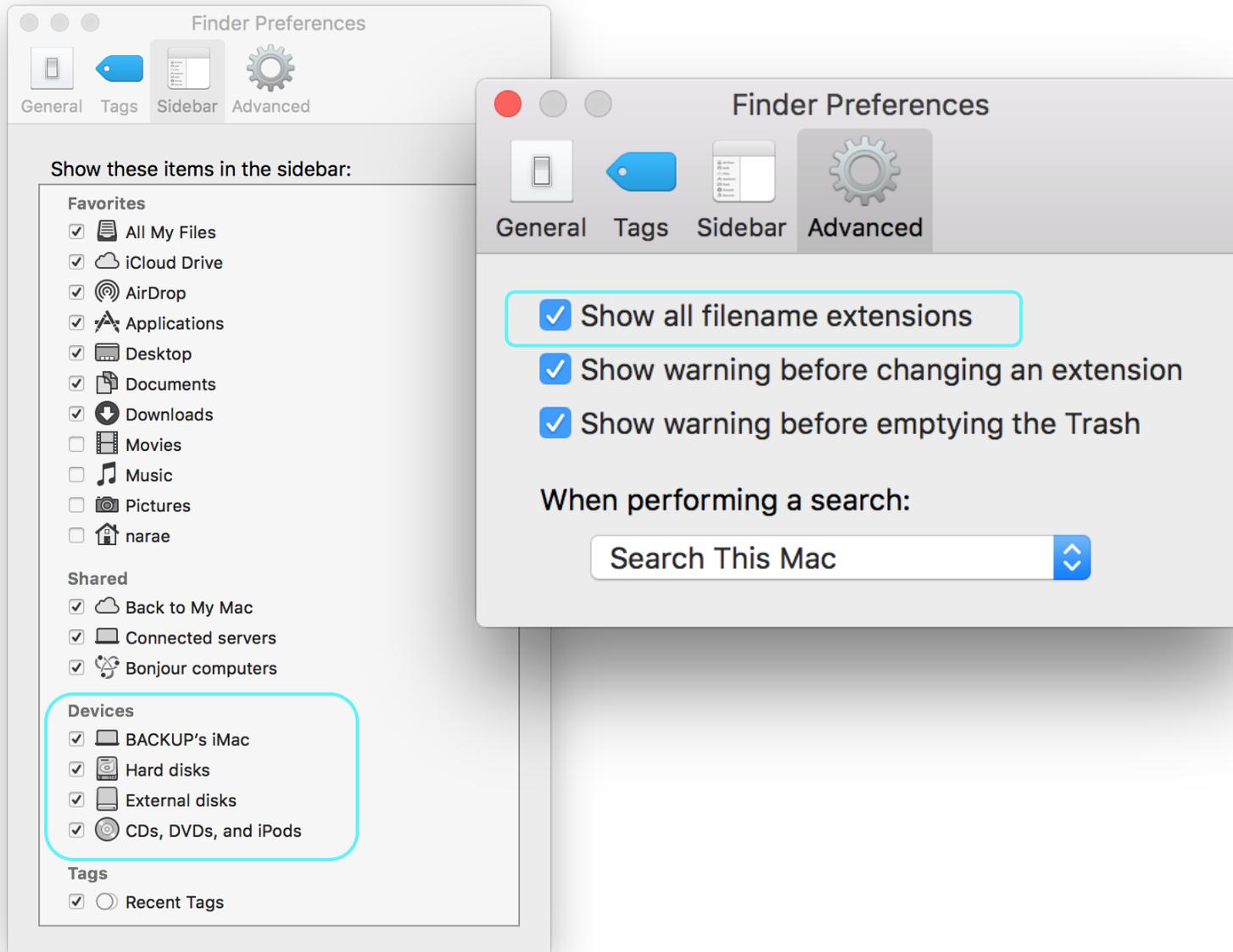
> AVOID SPACE in ALL file/folder names!

▸ Have your system show: file name extension, hidden files, full path in window title bar.

  ◆ Windows & Mac instructions ➔ next slides
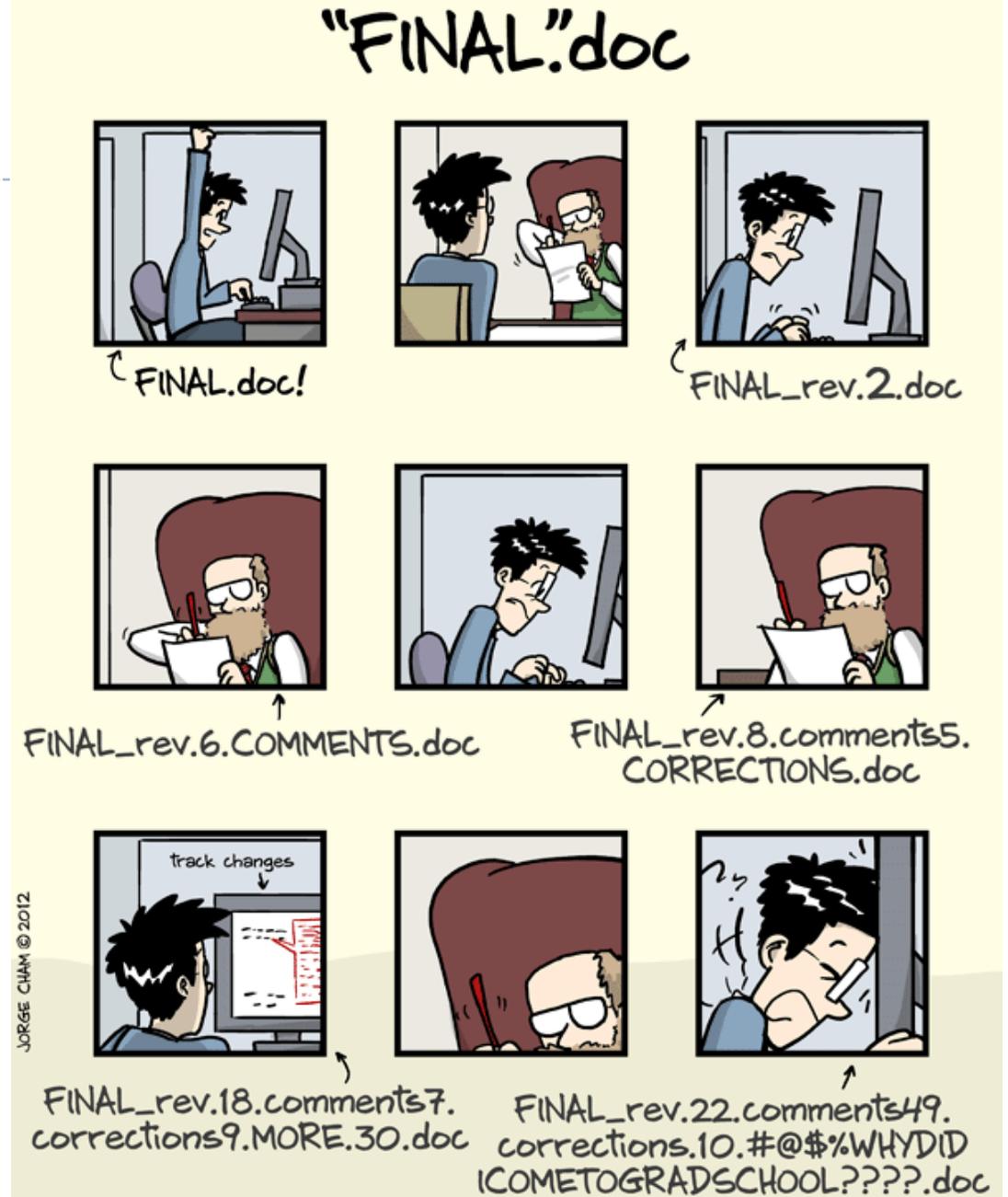
Windows Explorer "View" setup

1/8/2025

Mac Finder setup

Version Control

- "Piled Higher and Deeper" by Jorge Cham

- https://phdcomics.com/comics/archive.php?comicid=1531

What's the problem with this approach?
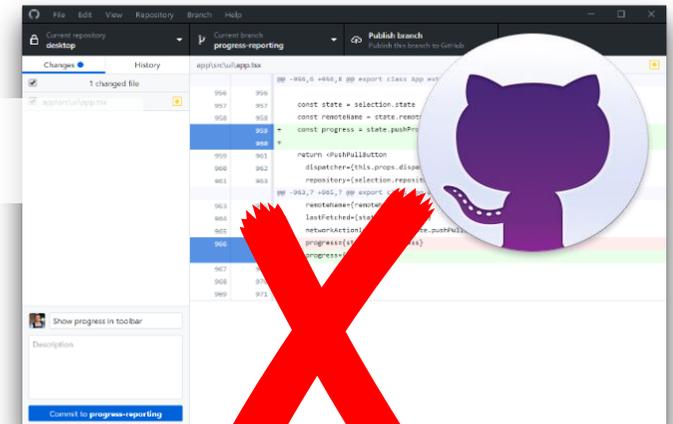

1/8/2025

Git

- **What is Git?**
  - One of the most popular *version control systems* in the coding community.

- **Na-Rae's LSA 2019 Workshop Tutorials:**
  - Part 1 Intro to Git
  - Part 2 Linking Git with GitHub → will cover next class

- **In this class, we will exclusively use the command line interface of git.**
  - Ignore the GUI clients.
  - Likewise, do NOT install/use GitHub Desktop.

Configuring your Git

Tutorial Part 1, [Setting Up Git](#)

▸ Mac users: open up a Terminal.

▸ Windows users: open up a Git Bash terminal.

▸ Display current global configuration:

  ◆ `git config --list --global`

▸ We will configure:

  ◆ Your name

  ◆ Your email (use Pitt email)

  ◆ Your editor (Don't leave it as vim! Your own text editor, or set it to nano if you're unsure.)

Your first local repository: getting started (To-do #1)

Follow steps in Tutorial Part 1, [Creating a Repository](#)

1. Create a directory called "languages". Move into it.

2. Initiate it as a Git repository:

   `git init`

3. Create a new text file 'zulu.txt', add lines to it

4. Add files to staging area:

   `git add zulu.txt`

5. Commit the change:

   `git commit -m "started zulu"`

6. Edit the text file again

7. Add files to be committed:

   `git add zulu.txt`

8. Commit the change:

   `git commit -m "details on…"`

Check status between steps:
`git status`

Check changes, history:
`git diff`
`git log`

Wrapping up

▶ To-do #1 is out. Due next class.

◆ Learn Git!