Lecture 1: Course Introduction, Setup, Git

LING 1340/2340: Data Science for Linguists

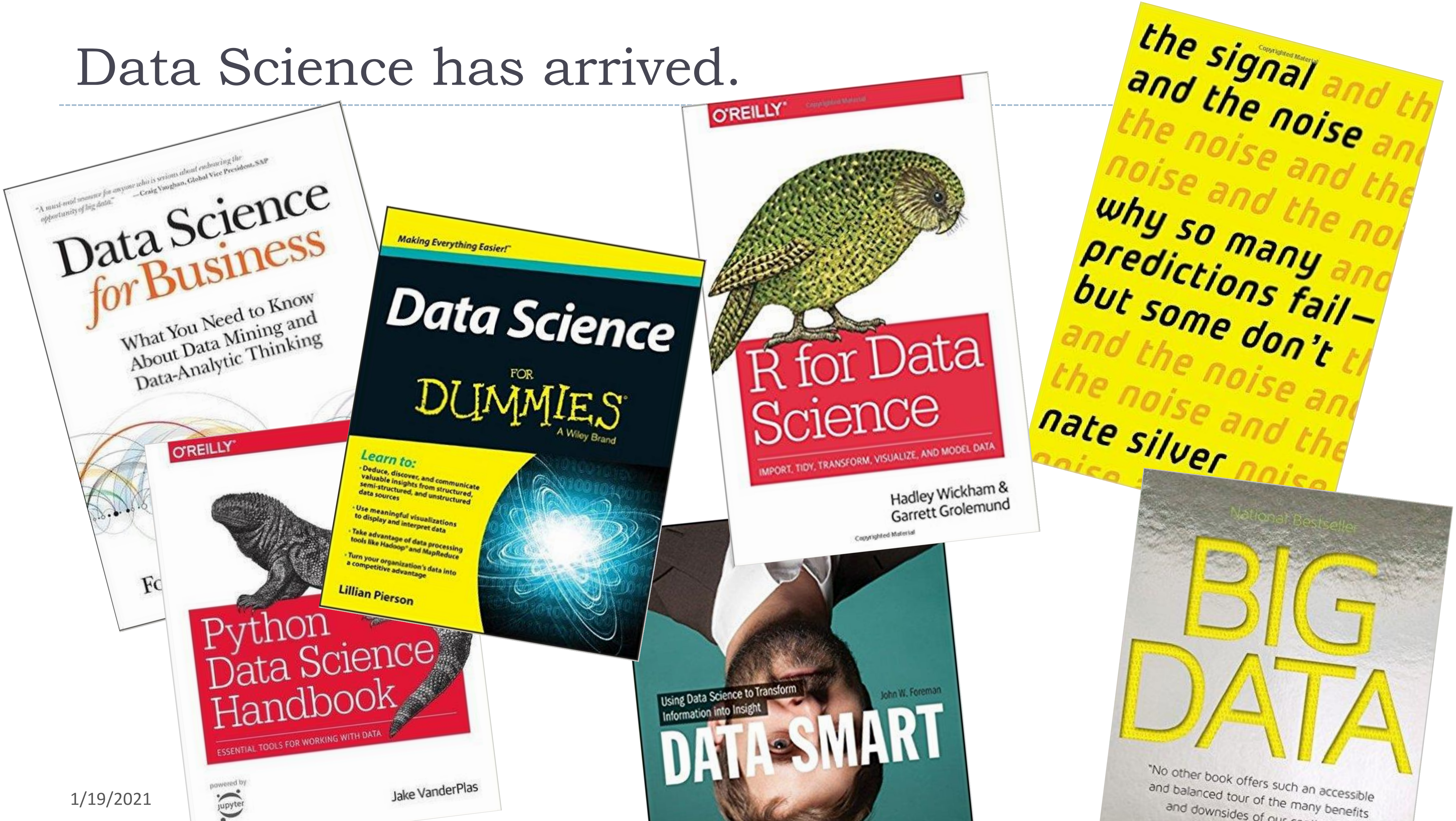
Na-Rae Han

Objectives

- Introduction: What is Data Science?
- Scope of this course
- Course logistics and policies

- Tools: Git

Data Science has arrived.


"A must-read resource for anyone who is serious about embracing the opportunity of big data." —Craig Vaughan, Global Vice President, SAP
Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas
Making Everything Easier!
Data Science FOR DUMMIES
A Wiley Brand
Learn to:
Deduce, discover, and communicate valuable insights from structured, semi-structured, and unstructured data sources
Use meaningful visualizations to display and interpret data
Take advantage of data processing tools like Hadoop and MapReduce
Turn your organization's data into a competitive advantage
Lillian Pierson
O'REILLY
R for Data Science
IMPORT, TIDY, TRANSFORM, VISUALIZE, AND MODEL DATA
Hadley Wickham & Garrett Grolemund
Using Data Science to Transform Information into Insight
John W. Foreman
DATA SMART
the signal and the noise
why so many predictions fail—but some don't
nate silver
National Bestseller
BIG DATA
"No other book offers such an accessible and balanced tour of the many benefits and downsides of our

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

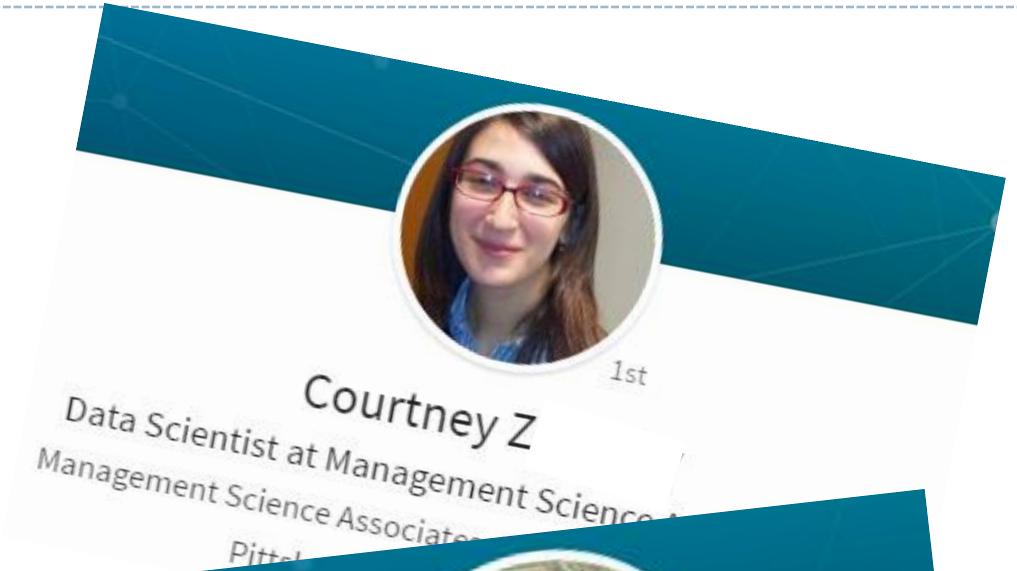
Data Science for Linguists (2)

- 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.



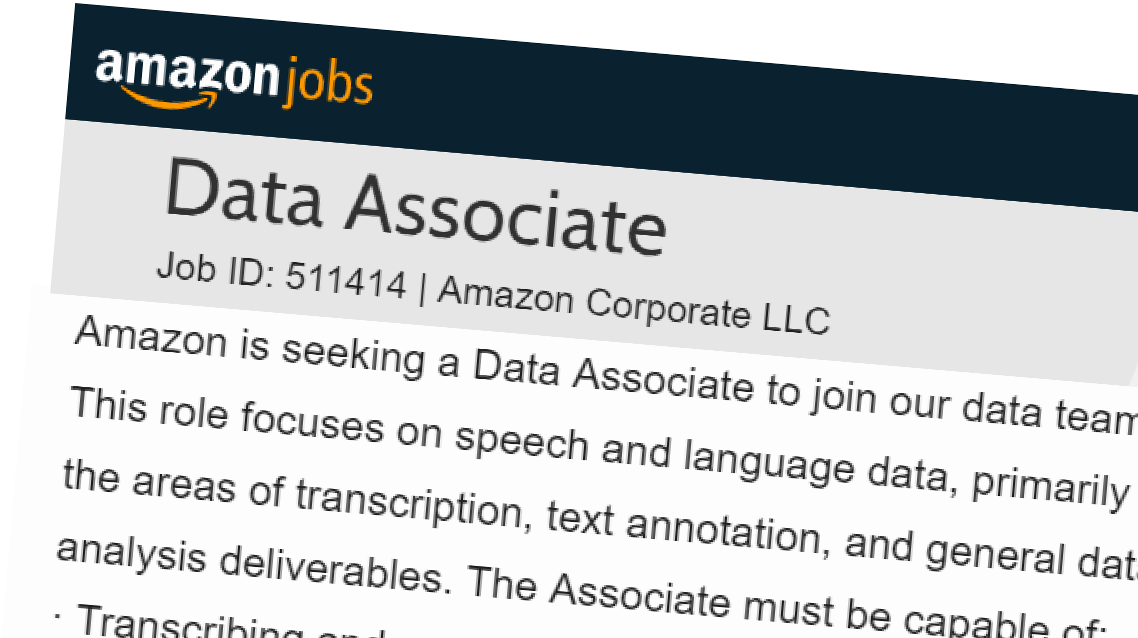
1st

Courtney Z

Data Scientist at Management Scienc

Management Science Associat

Pitt

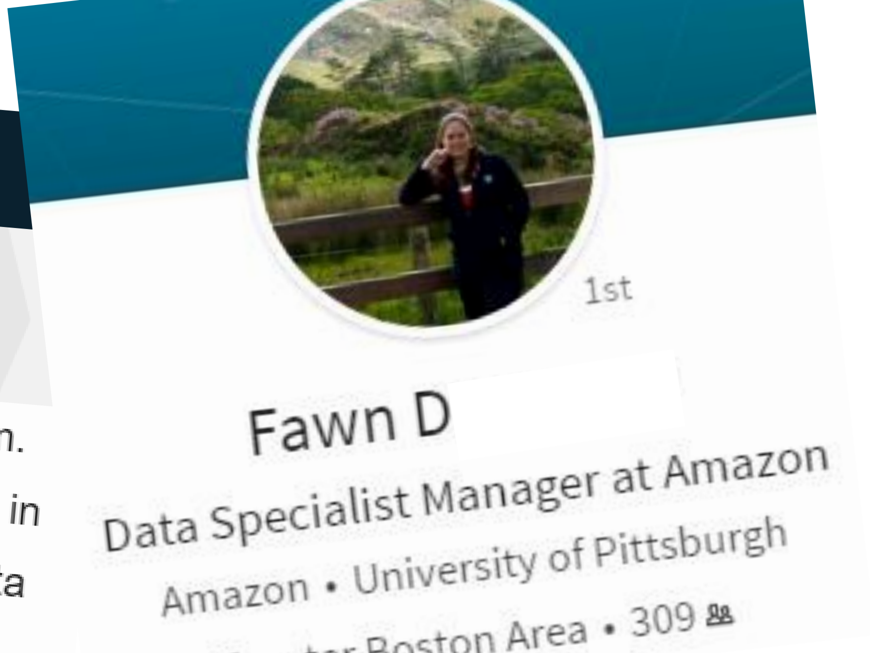


Data Associate

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



1st

Fawn D

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**.
  - (Note: our dept might also create an R-focused version.)

Review syllabus

- Course home page doubles up as syllabus:
  - https://naraehan.github.io/Data-Science-for-Linguists-2021/

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-2021/	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	GitHub org repository (public)	You (and instructors, commenters)	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
  - Atom, Notepad++, Sublime Text, ...
- Git
- Anaconda Python (Python version 3.8)
  - Jupyter Notebook

System-side house keeping

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

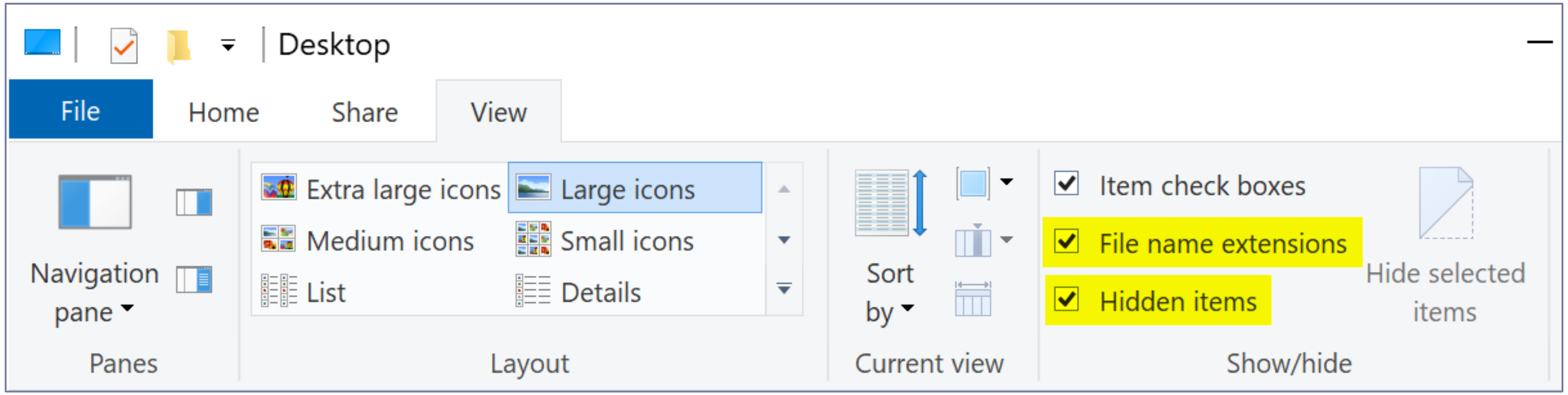
  `C:\Users\narae\Documents\Data_Science`

  ← Avoid space in folder name.

  ← AVOID SPACE in ALL file/folder names this semester.

- Have your system show: file name extension, hidden files, full path in window title bar.
  - Windows
  - Mac OS

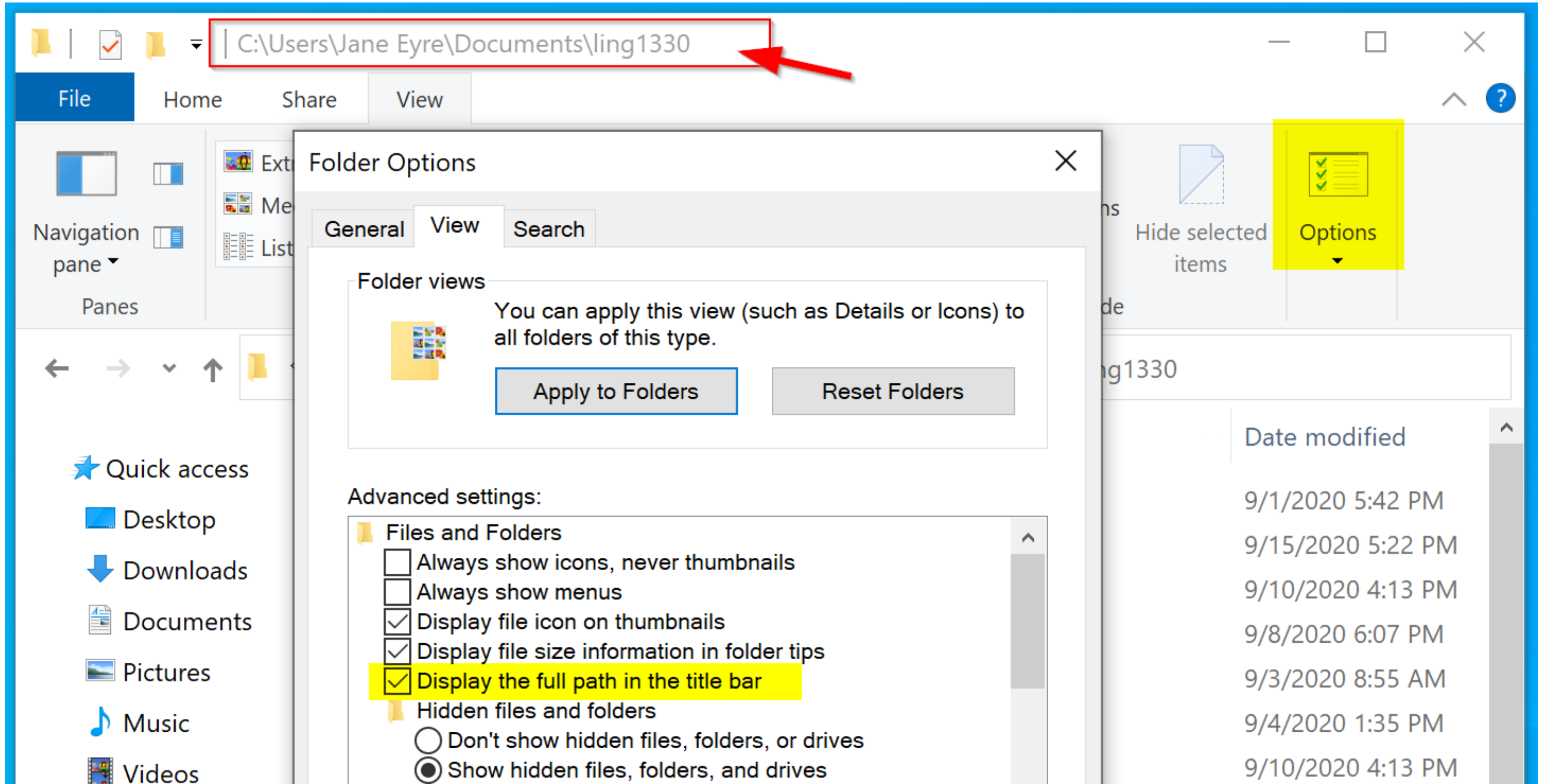
Windows Explorer "View" setup



☑ Item check boxes

☑ File name extensions

☑ Hidden items



C:\Users\Jane Eyre\Documents\ling1330

Folder Options

General | View | Search

Folder views

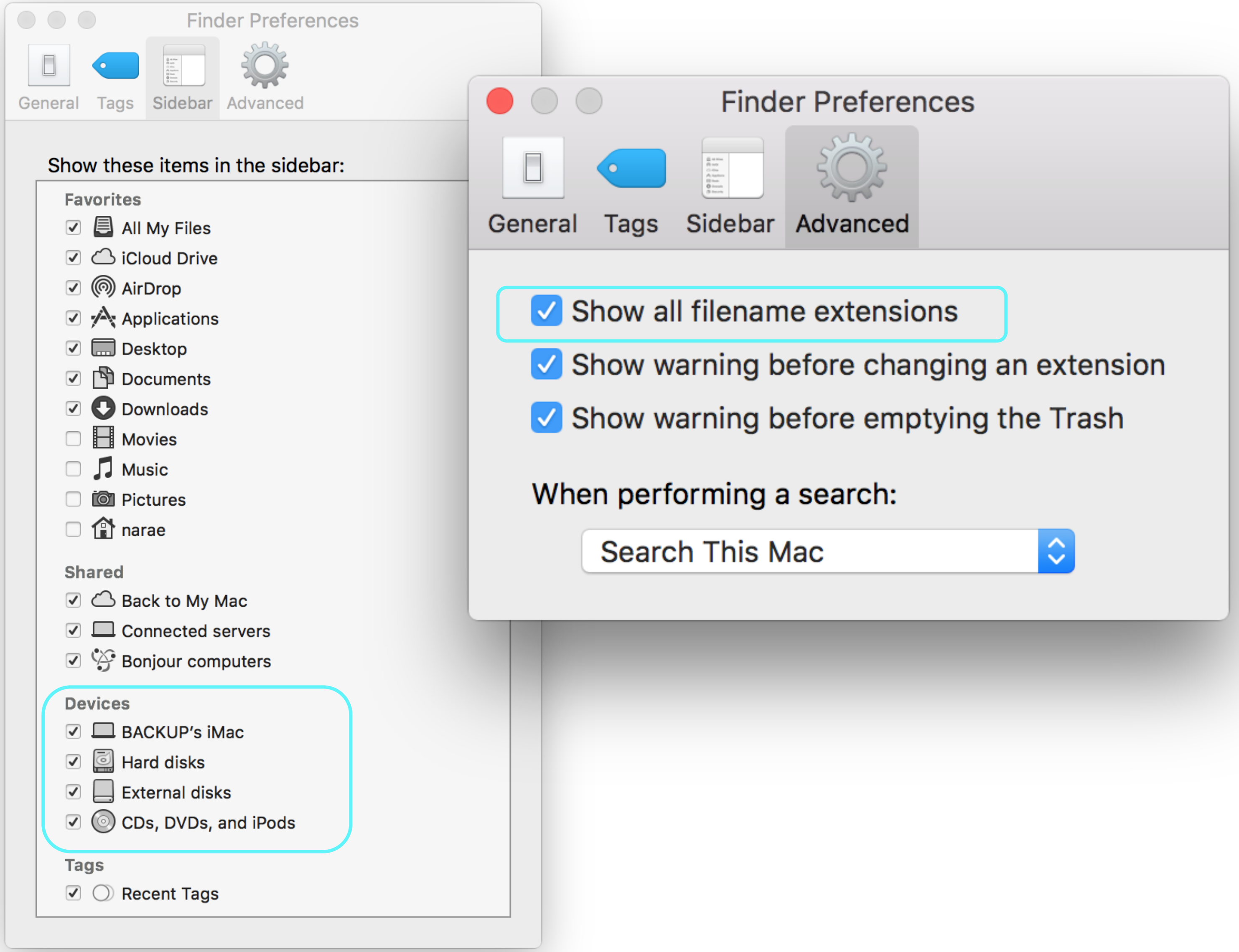
You can apply this view (such as Details or Icons) to all folders of this type.

Apply to Folders | Reset Folders

Advanced settings:

Files and Folders
- □ Always show icons, never thumbnails
- □ Always show menus
- ☑ Display file icon on thumbnails
- ☑ Display file size information in folder tips
- ☑ Display the full path in the title bar
- Hidden files and folders
  - ○ Don't show hidden files, folders, or drives
  - ● Show hidden files, folders, and drives

Mac Finder setup



Show these items in the sidebar:

Favorites
- ☑ All My Files
- ☑ iCloud Drive
- ☑ AirDrop
- ☑ Applications
- ☑ Desktop
- ☑ Documents
- ☑ Downloads
- □ Movies
- □ Music
- □ Pictures
- □ narae

Shared
- ☑ Back to My Mac
- ☑ Connected servers
- ☑ Bonjour computers

Devices
- ☑ BACKUP's iMac
- ☑ Hard disks
- ☑ External disks
- ☑ CDs, DVDs, and iPods

Tags
- ☑ Recent Tags

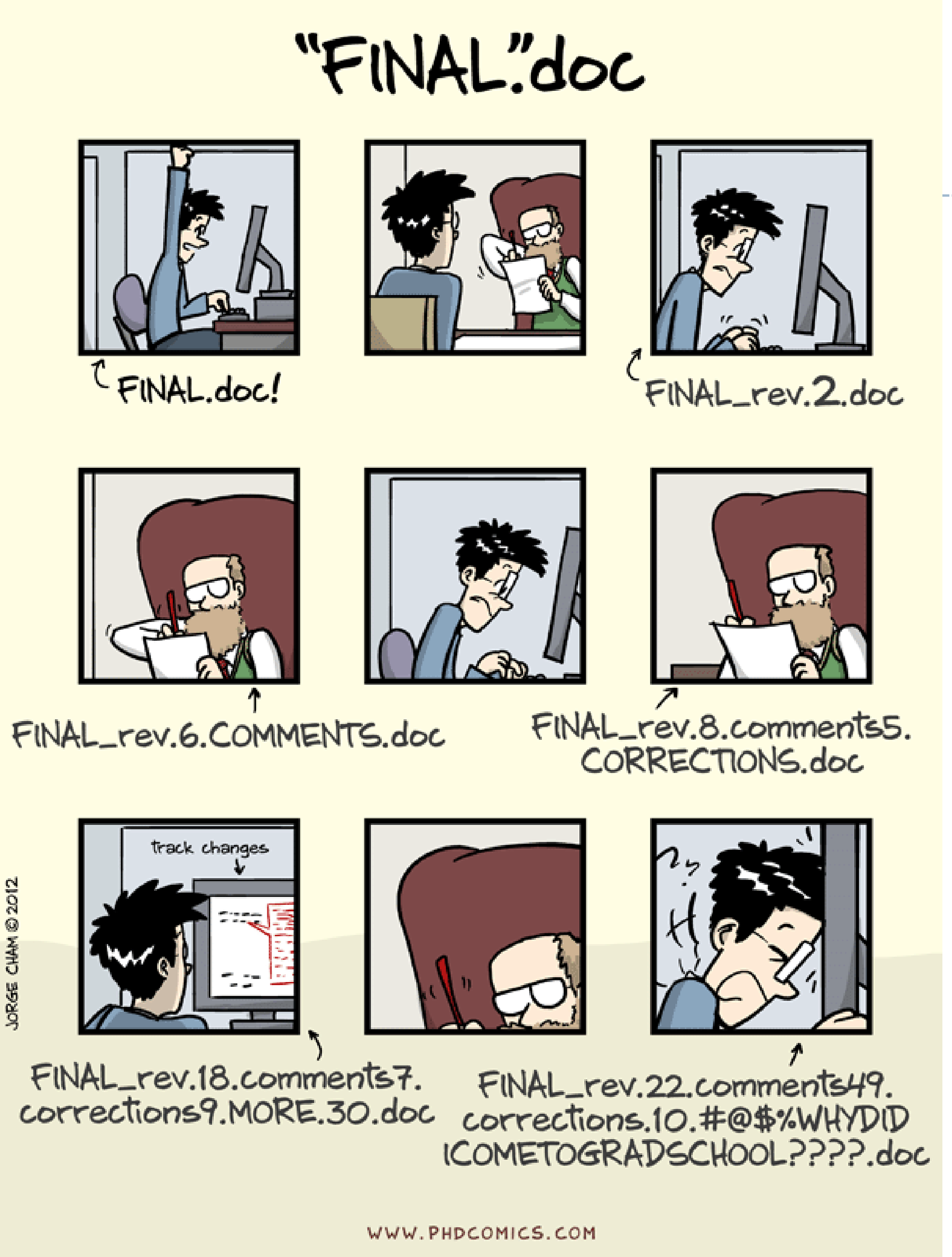
- ☑ Show all filename extensions
- ☑ Show warning before changing an extension
- ☑ Show warning before emptying the Trash

When performing a search:

Search This Mac

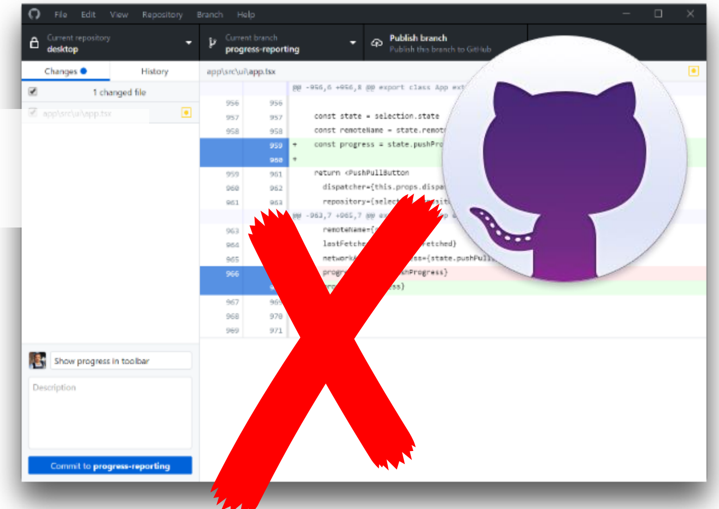
Version Control

- "Piled Higher and Deeper" by Jorge Cham
  http://www.phdcomics.com



Git

- What is Git?
  - One of the most popular *version control systems* in the coding community.
- LSA 2019 Workshop Tutorials:
  - Part 1 Intro to Git
  - Part 2 Linking Git with GitHub
- In this class, we will exclusively use the **command line** interface of git.
  - Ignore the GUI clients.
  - Likewise, do NOT install/use GitHub Desktop.



```
@@ -956,6 +956,8 @@ export class App ext

    const state = selection.state
    const remoteName = state.remot
+   const progress = state.pushPro
+
    return <PushPullButton
      dispatcher={this.props.dispa
      repository={select
@@ -963,7 +965,7 @@ ex
      remoteName={
      lastFetche                etched}
      network                ss={state.pushPull
      progr                  nProgress}
                             ss}
```

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 (anything other than vim!)

Your first local repository: getting started

Follow steps in Tutorial Part 1, Creating a Repository

1. Create a directory called `languages`
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
```

Your first local repository: tracking, history

Follow steps in Tutorial Part 1: Tracking Changes, A Commit Workflow, and Exploring History.

- To view entire version history:

```
git log
```

- To view changes:

```
git diff
git diff HEAD~1 file.txt
git diff --staged
```

- To view what changed in a particular version:

```
git show HEAD~1
```

- To scrap new changes since the last commit:

```
git checkout HEAD file.txt
```

- To restore an earlier version:

```
git checkout VERSION file.txt
```

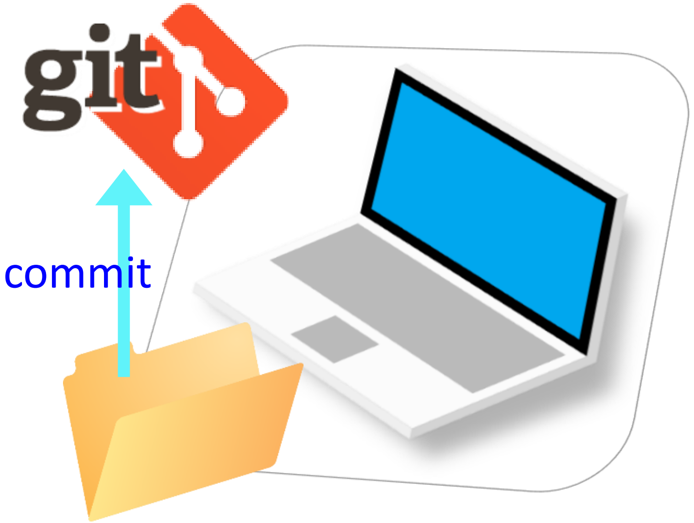
🡐 commit to make this the new HEAD

If thrown into pagination, use SPACE to page down, q to quit.

HEAD: the last committed version
HEAD~1: one before that

Your first local repository

- Your directory `languages/` was set up with a **Git repository**.



- `languages/` is now:
  - tracked by Git
  - all changes will be documented
  - able to revert back to earlier version, if needs be

- But is this all?
  - How about backup? collaboration? social?

GitHub

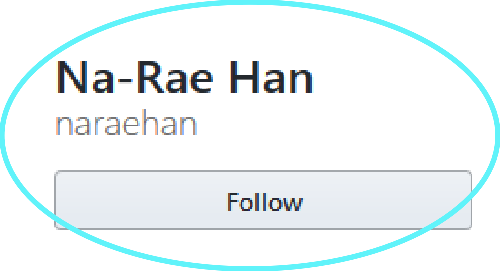
→ NEXT CLASS

Setting up GitHub

- Create a GitHub account at https://github.com/.
  - Use your **Pitt email address.**
  - If you already have an account with a different email, add your Pitt email to your account.
  - GitHub sends you a verification email. Confirm.

  ← The verification email might go to the SPAM folder. You MUST resolve and verify!

- Visit Na-Rae's GitHub profile, and follow →
  - https://github.com/naraehan



Wrapping up

- To-do #1 is out. Due next class.
- Homework #1 is also out: due next Tuesday.
  - Instructions relating to Git/GitHub & Jupyter Notebook will make little sense to you now. It will after Thursday.
- Create a GitHub account.
- I'll be sending DataCamp group invitation.
- Start learning:
  - Git, GitHub
  - Jupyter Notebook
  - numpy, pandas