

Lecture 1: Course Introduction, Set up, Git

LING 1340/2340: Data Science for Linguists

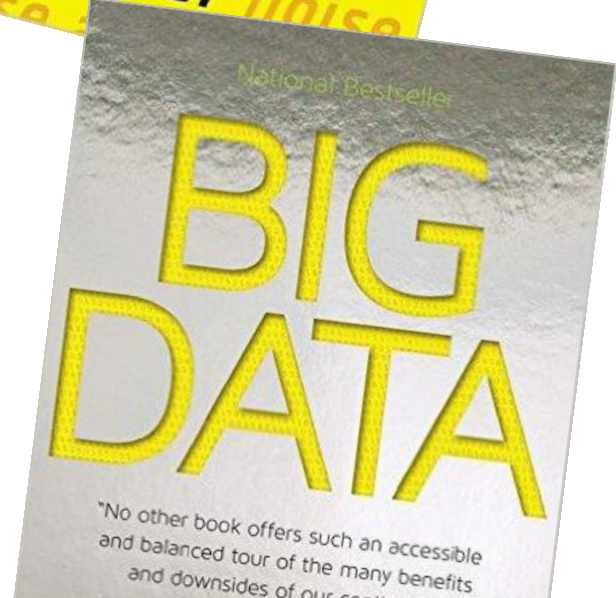
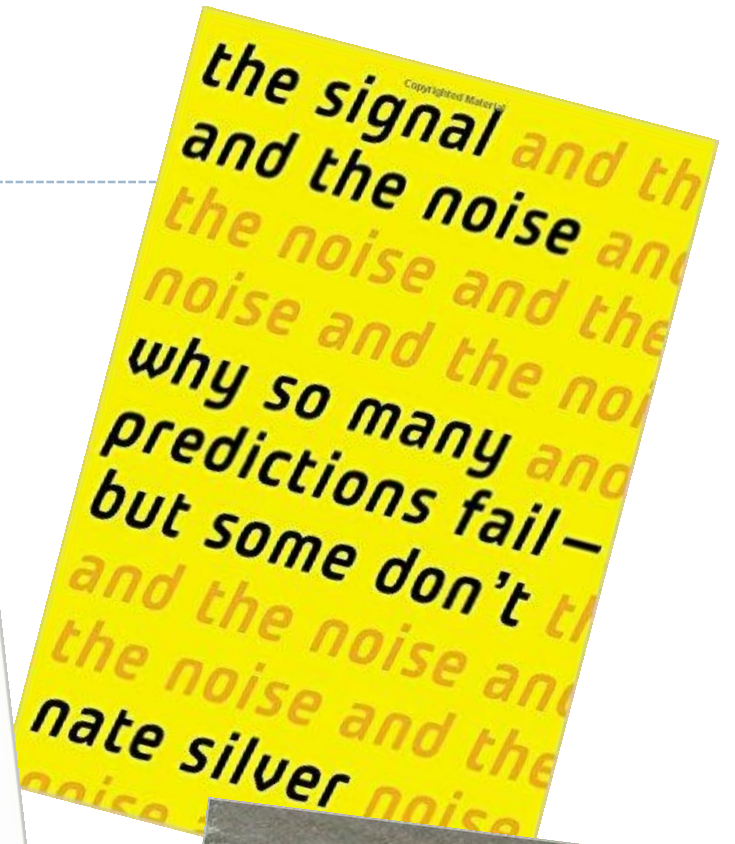
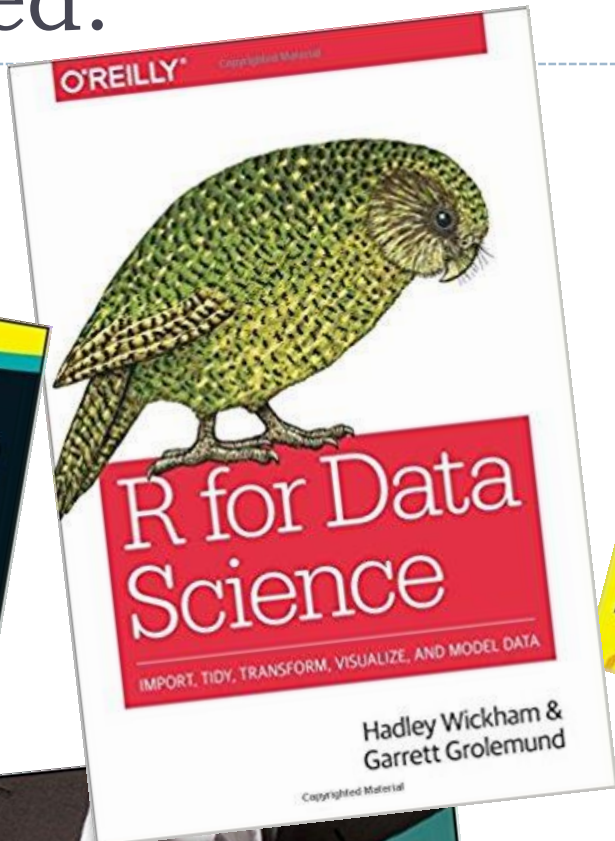
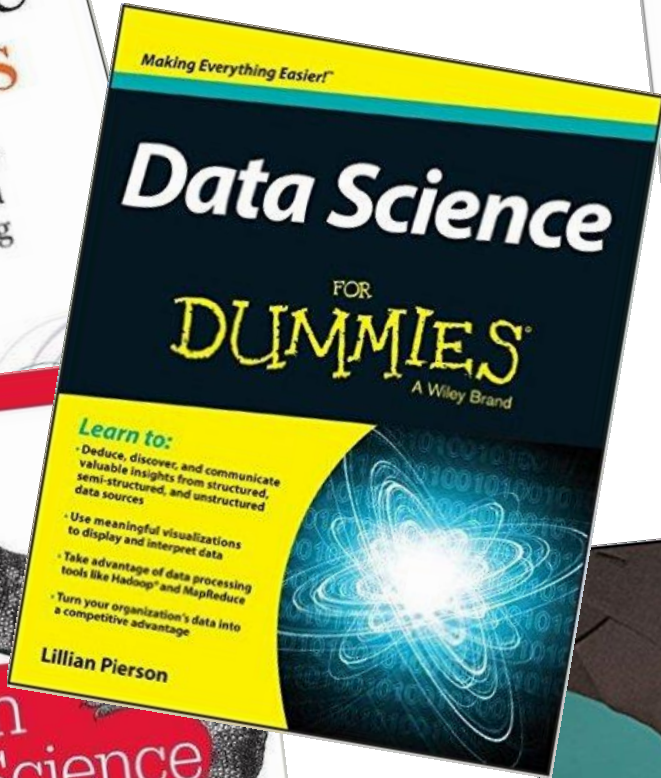
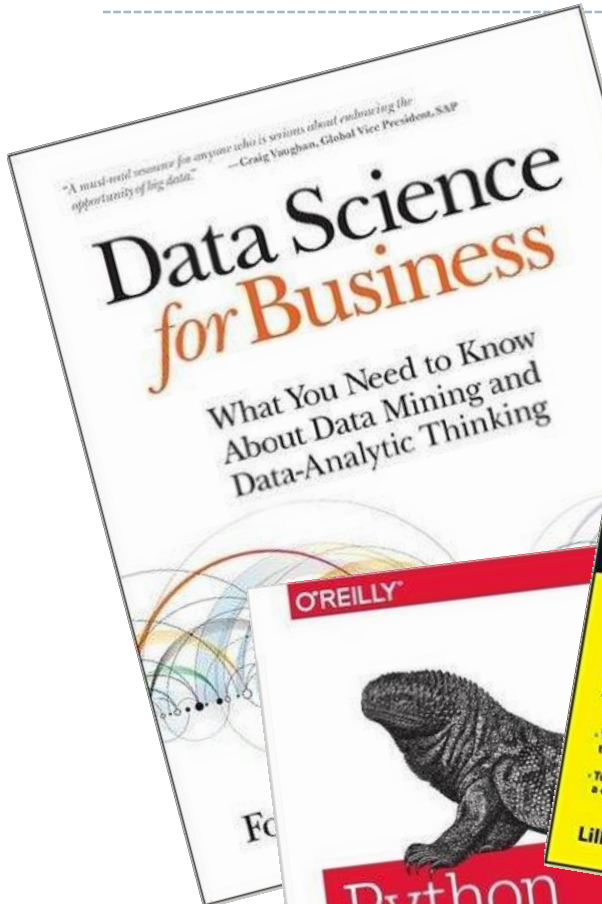
Na-Rae Han

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 is "Data Science"?

▶ Wikis and articles:

- ◆ [https://en.wikibooks.org/wiki/Data Science: An Introduction/A Mash-up of Disciplines](https://en.wikibooks.org/wiki/Data_Science:_An_Introduction/A_Mash-up_of_Disciplines)
- ◆ <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>
- ◆ <https://www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/>

▶ Data Science degree programs:

- ◆ <https://datascience.berkeley.edu/about/what-is-data-science/>
- ◆ <http://datascience.nyu.edu/what-is-data-science/>
- ◆ Pitt is also starting one! (Math + Stats + CS)

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>

Before "Data Science" was a thing

- ▶ Business Analytics

- ◆ But is it just about number crunching, as in stock prices?

- ▶ Data Analytics

- ▶ Data Mining

- ◆ "Data miner" as a job description? Hmm...

- ▶ Data Science

- ◆ Sounds like a proper discipline, and "data scientist" has a nice ring as a job title.
- ◆ Stresses the "unstructured" and "organic" nature of today's data.

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.



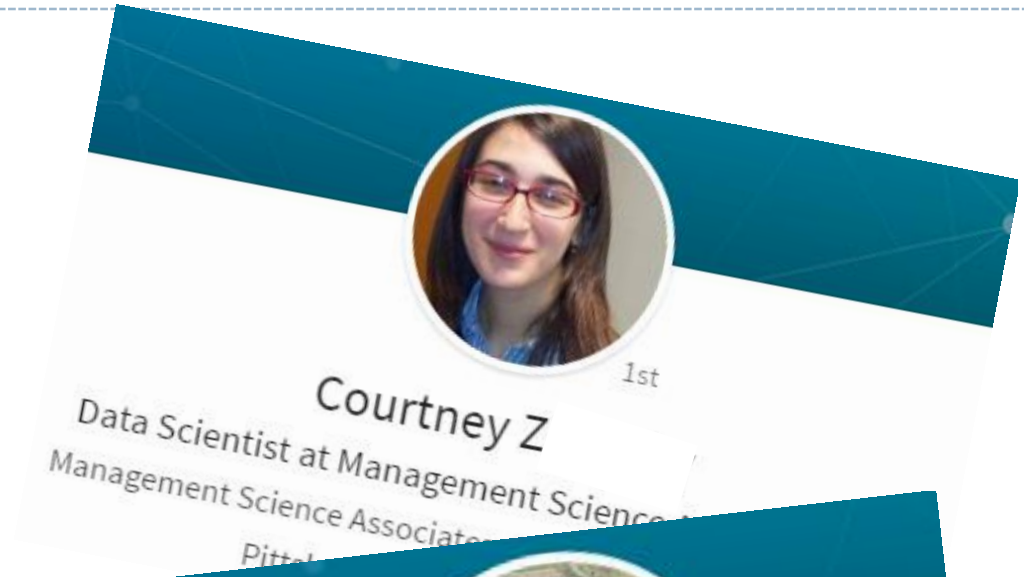
amazon jobs


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

Courtney Z

Data Scientist at Management Science Associates
Management Science Associates
Pittsburgh, PA



 1st

Fawn D

Data Specialist Manager at Amazon
Amazon • University of Pittsburgh
Greater 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**. (We may cover a bit of R later.)

Review syllabus

- ▶ Course home page doubles up as syllabus:
 - ◆ <https://naraehan.github.io/Data-Science-for-Linguists-2019/>

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, "correctly working code" 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-2019/	Instructors	Entire world can view
Your grades, announcements	CourseWeb	Instructors	Instructors and you
Data distribution, assignment submission	GitHub repositories (private)	Everyone in this class	Everyone in this class
Your term project	GitHub organization repository (public)	You (and instructors, commenters)	Entire world can view

Additionally: PSC account, online video tutorials, textbooks

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:

- ▶ Anaconda Python (Python version 3.7)
 - ◆ Jupyter Notebook
- ▶ Git
- ▶ A text editor
 - ◆ Atom, Notepad++, Sublime Text, ...

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.

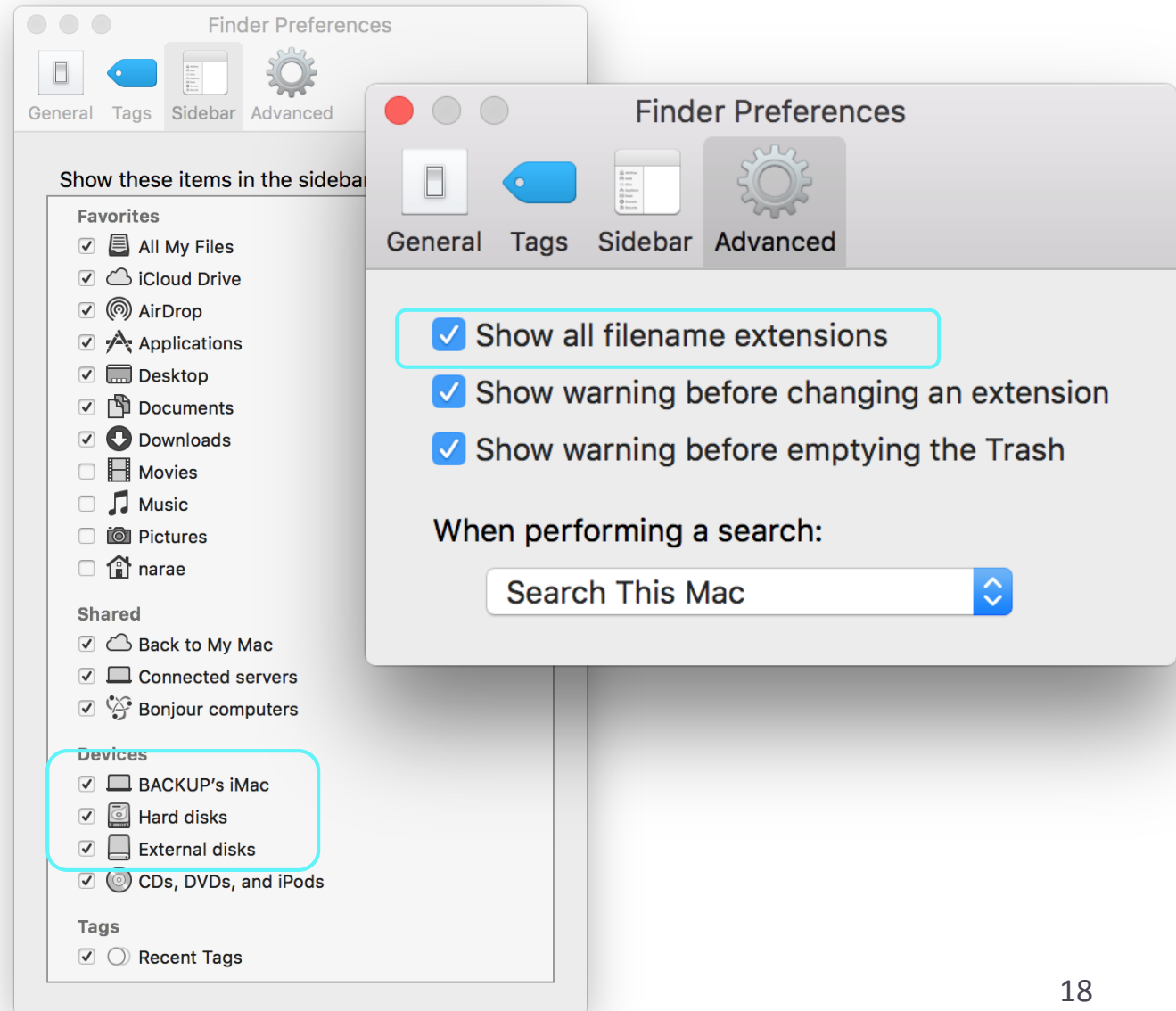
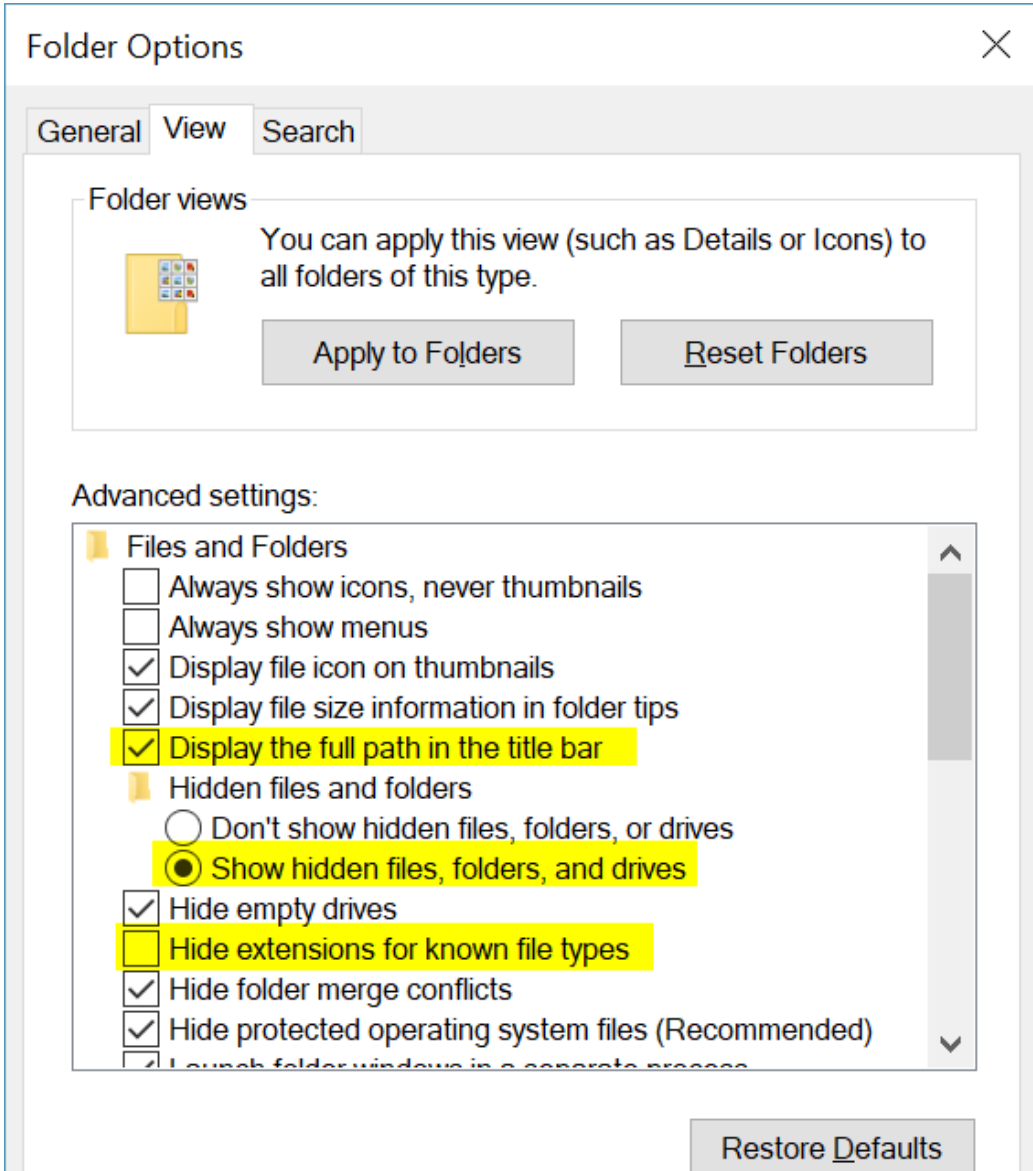
← **AVOID SPACE** in ALL file/folder names this semester.



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

- ◆ Windows
- ◆ Mac OS

Windows, Mac Finder set up



Version Control

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



Git



▶ What is Git?

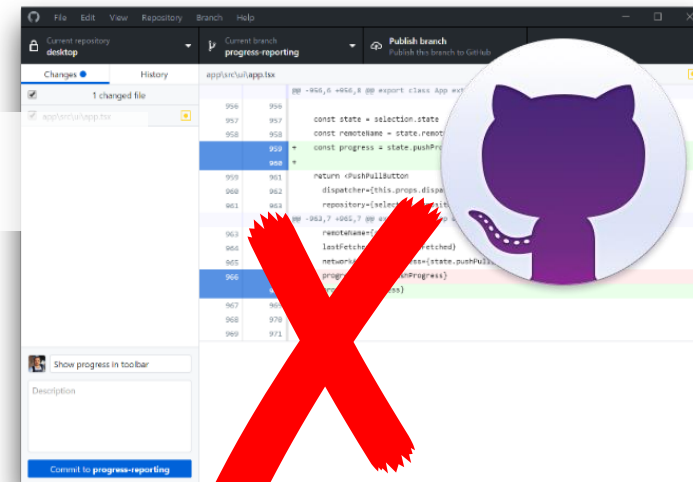
- ◆ One of the most popular *version control systems* in the coding community.

▶ 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](#).



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 new changes:

```
git diff
```

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

```
git diff --staged
```

To view what's 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.

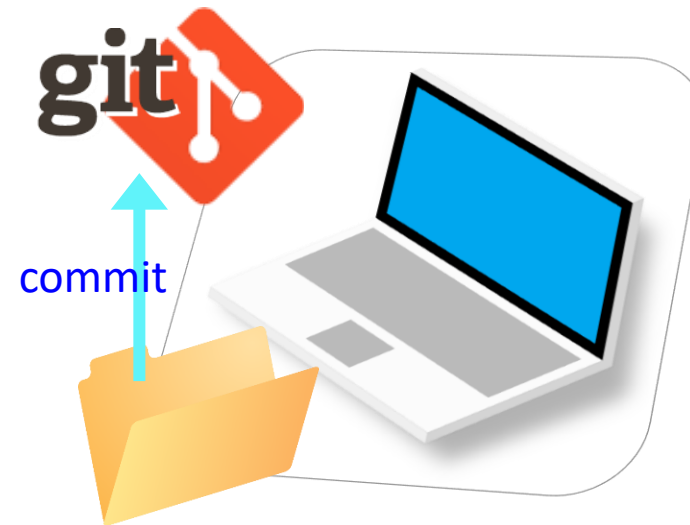
HEAD: the last committed version
HEAD~1: one before that

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?



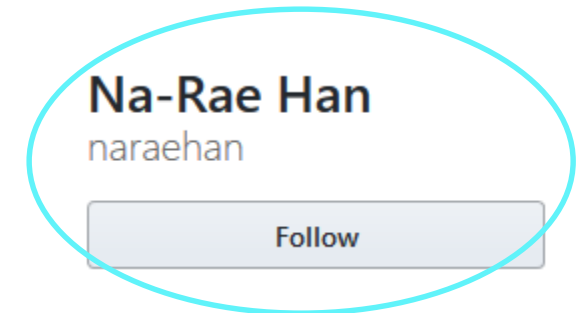
GitHub

➔ NEXT CLASS

Setting up GitHub



- ▶ Create a GitHub account at <https://github.com/>.
 - ◆ Use your **Pitt email address** if you can.
 - ◆ If you already have an account with a different email, **add your Pitt email** to your account.
 - ◆ GitHub sends you **a verification email. Confirm.**
- ▶ Visit Na-Rae's GitHub profile, and **follow** →
 - ◆ <https://github.com/naraehan>
- ▶ **OPTIONAL:** Go to GitHub Education page: <https://education.github.com/> and click on "**Request a discount**".
 - ← Optional because, as of TODAY, free accounts also get private repositories



Wrapping up

- ▶ To-do #1 is out. Due next class.
 - ◆ Corpus vs. non-corpus data type
- ▶ 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.
- ▶ I'll be sending DataCamp group invitation.
- ▶ Start learning:
 - ◆ Git, GitHub
 - ◆ Jupyter Notebook
 - ◆ numpy, pandas