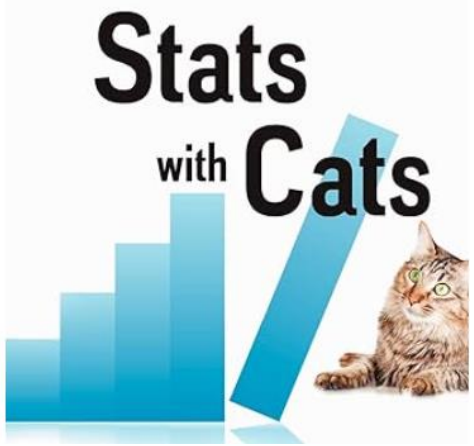
The Benefits of Implementing Git & GitHub in a Data Science Curriculum

Eric Burnheimer
Massachusetts Bay CC
eburnheimer@massbay.edu
ericburn.github.io

Hypothetical Student “Octo”*

- Octo had a keen interest in analyzing cat data.
  - Job Satisfaction in Data Science (DS) work is high (Baumer & Horton, 2023)



Credit: Charles Kufs

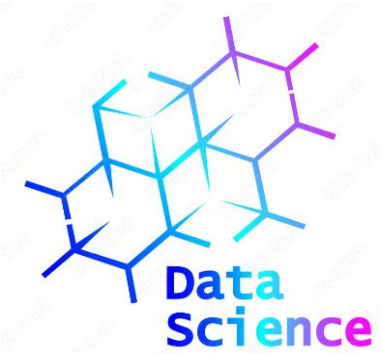
- Octo began to explore data science programs
  - Two-year colleges positioned to play large role in DS education
    (Kotz, 2020; Amstat, 2022; Glantz, et.al. 2023)

*Adapted from Baumer & Horton (2023)

Data Science Programs in Two-Year Colleges

MASSBAY COMMUNITY COLLEGE

- Octo's local two-year college had just started a DS certificate!
  - Important curricular decisions to be made for A.S. degree program



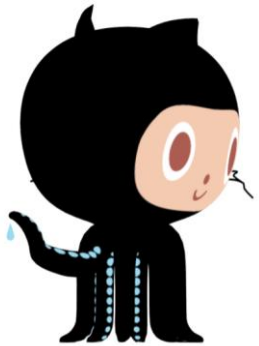
- One important curricular decision involves when and how to teach reproducibility and workflow (Baumer & Horton, 2023; Çetinkaya-Rundel & Rundel, 2018)
  - Version Control Systems (e.g., Git)
  - Collaboration and Project Management Venues (e.g., GitHub)

Two things happen to Octo during winter break…

Credit: Zoom

“It’s on my **GitHub**”

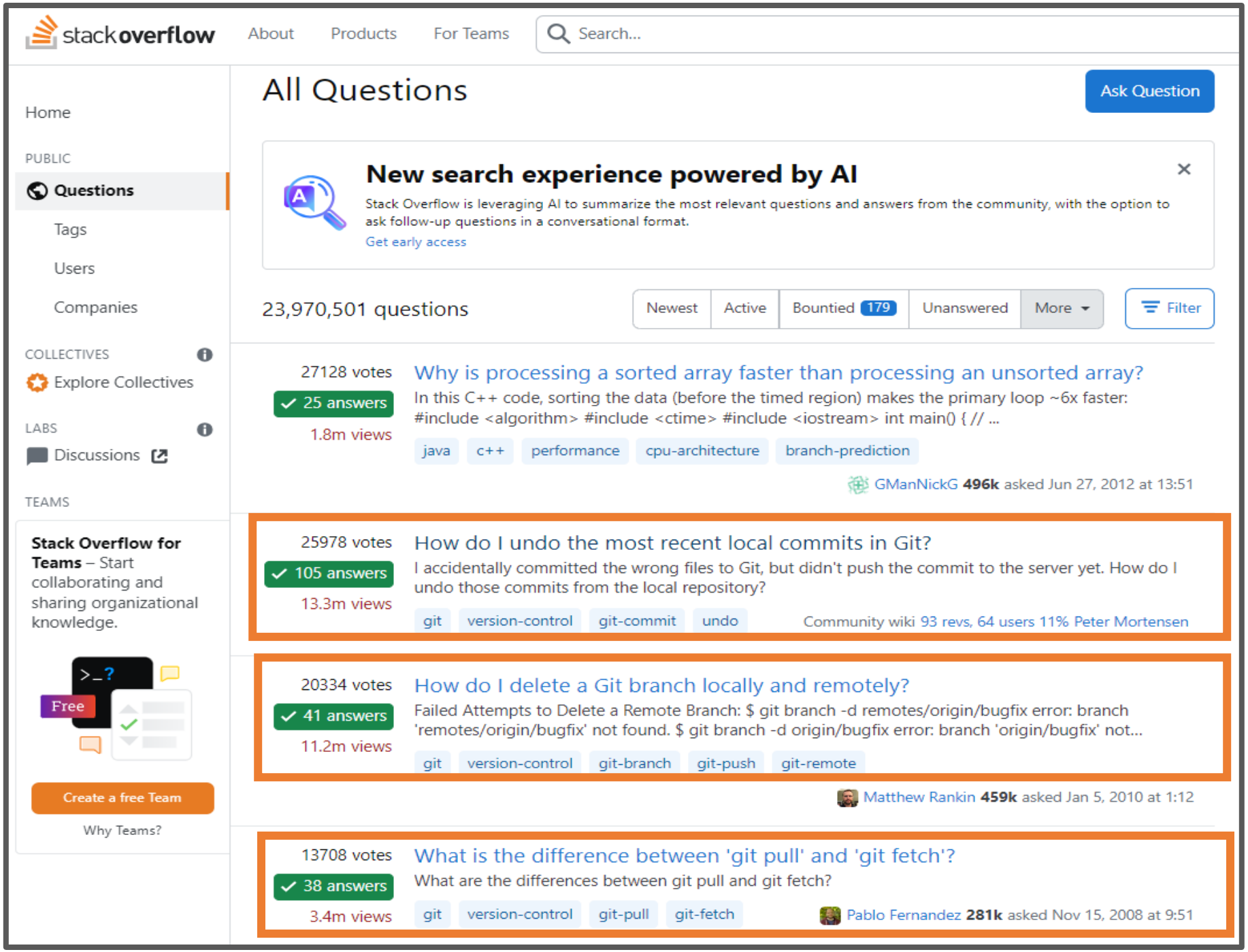
“Work collaboratively through **GitHub**”



“Most of the folks that we interview have their GitHub pages listed on their resume,” says CiBo Technologies talent acquisition manager Jamieson Vasquez. “I think that is important.”

Companies That Use Git & GitHub

Source: Modified from git-scm.com and Stack Share



MASSBAY COMMUNITY COLLEGE

3 of top 4 questions on Stack Overflow relate to **Git.**

MASSBAY COMMUNITY COLLEGE

"Whoa, I've just read this quick tutorial about git and oh my god it is cool. I feel now super comfortable using it, and I'm not afraid at all to break something."

— said no one ever (de Wulf)

MASSBAY COMMUNITY COLLEGE

What is Version Control and What Are Git & GitHub?

What is Version Control?

- “A system that records changes to a file or set of files over time…” (Chacon & Straub, 2014)

Credit: Jorge Cham

What is Git?

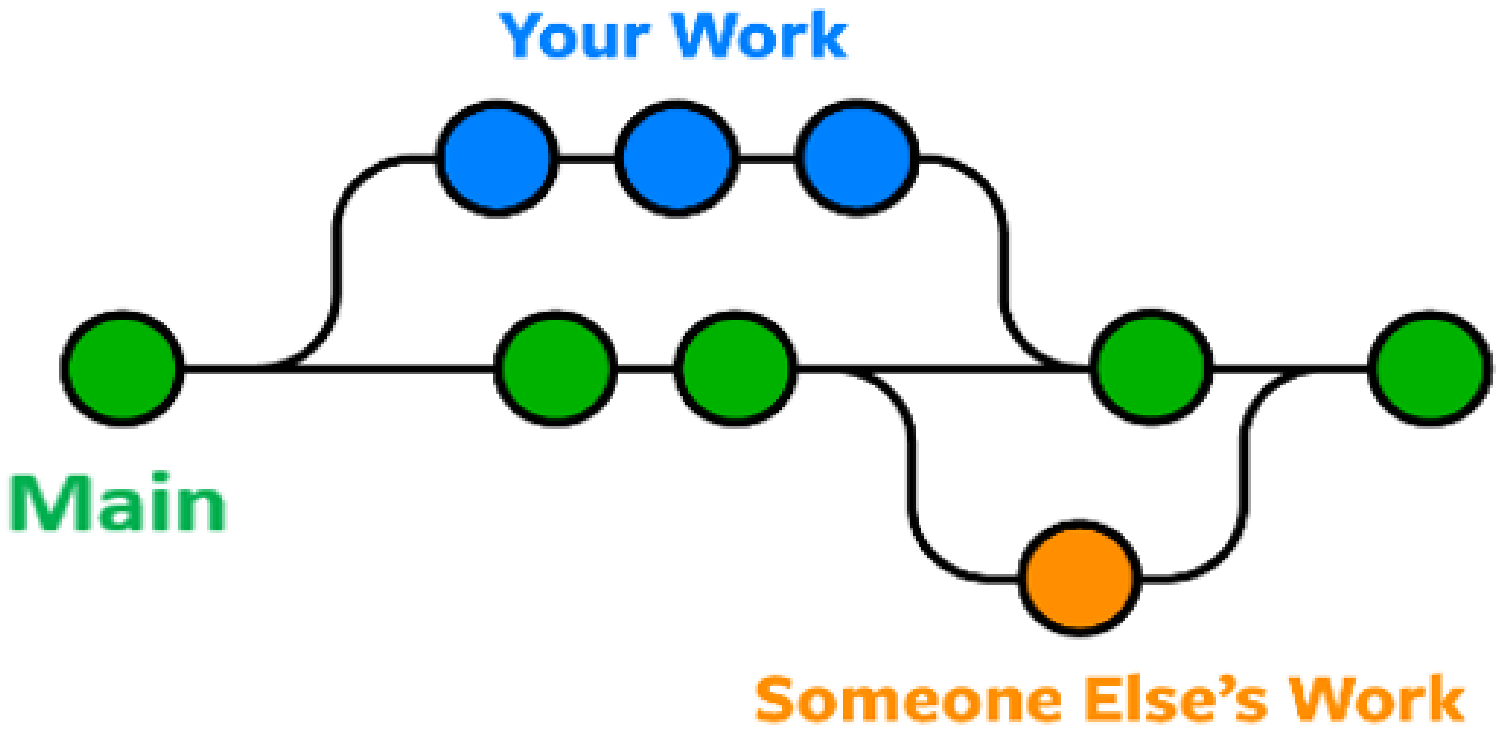
- **Git** is a free version control system
  - Various commands to track and manage changes to files over time
    - Revisit earlier versions of files
    - Compare changes between versions
    - Undo changes

- **Git** use is enabled through a repository (repo)
  - Place that tracks & manages files within a folder

Credit: Modified from Steele 2021

What is GitHub?

- **GitHub** hosts git repositories in the cloud
  - ➢ Allows teams (e.g., students) to work on files together (e.g., R scripts)



Credit: Noble Desktop

MASSBAY COMMUNITY COLLEGE

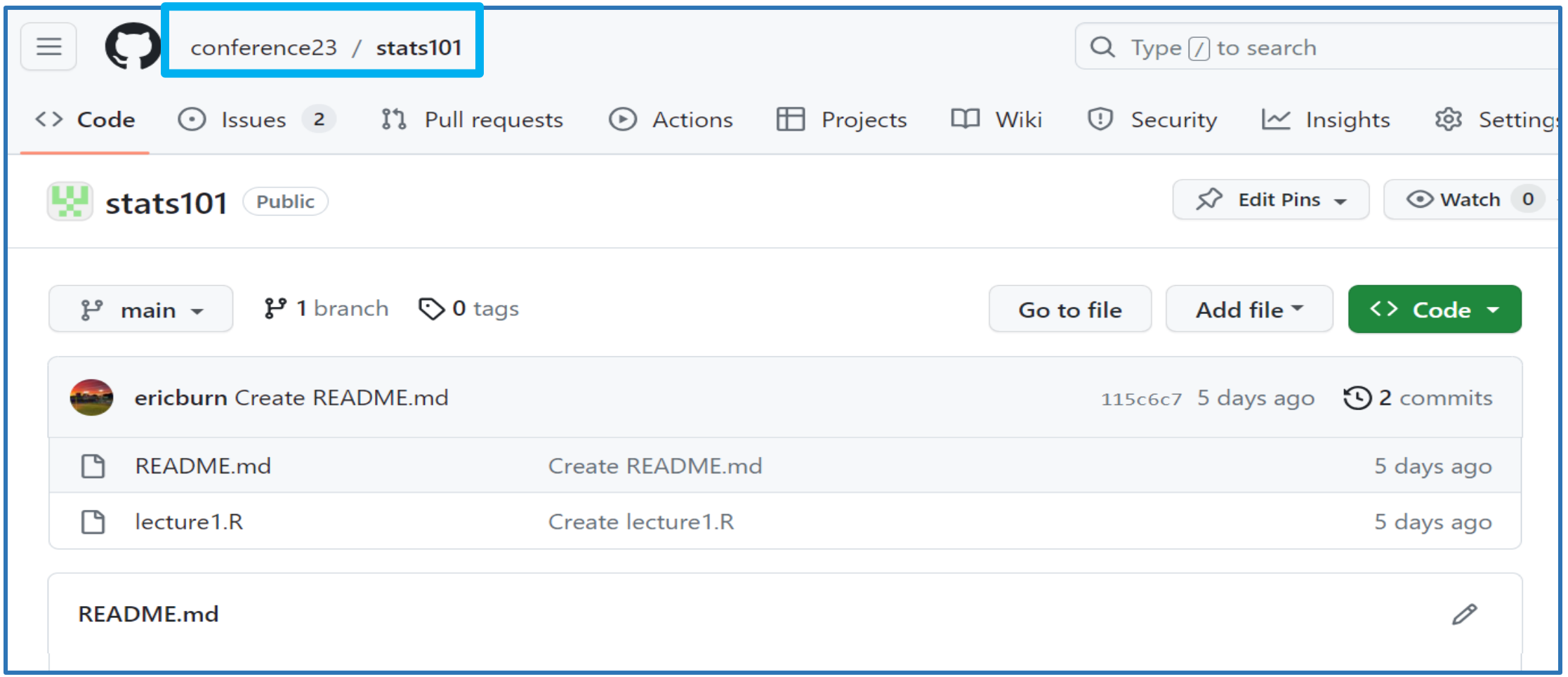
A Pathway for Implementing Git and GitHub in a DS Curriculum

MASSBAY COMMUNITY COLLEGE

Semester I: Use GitHub Issues to Collaborate in Introductory Statistics

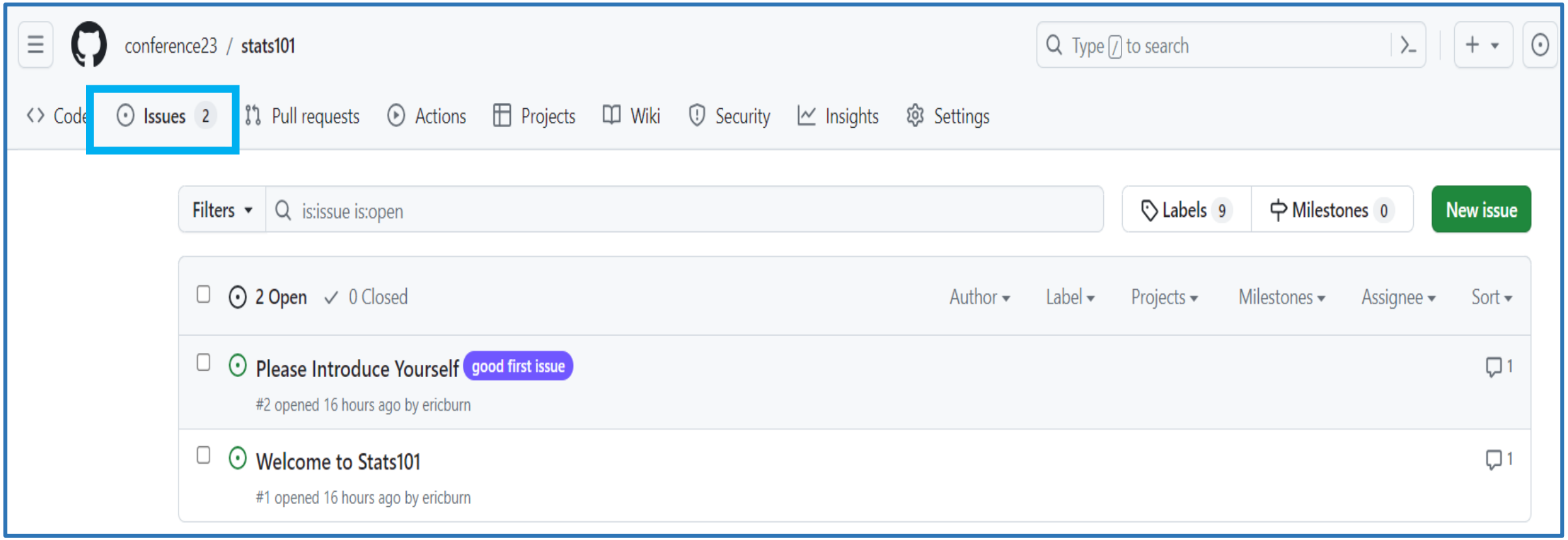
Part I: Create An Organization & Class Repo

- Invite students as collaborators to your repo (e.g., stats101)



Part II: Introduce GitHub (& Git) through GitHub Issues

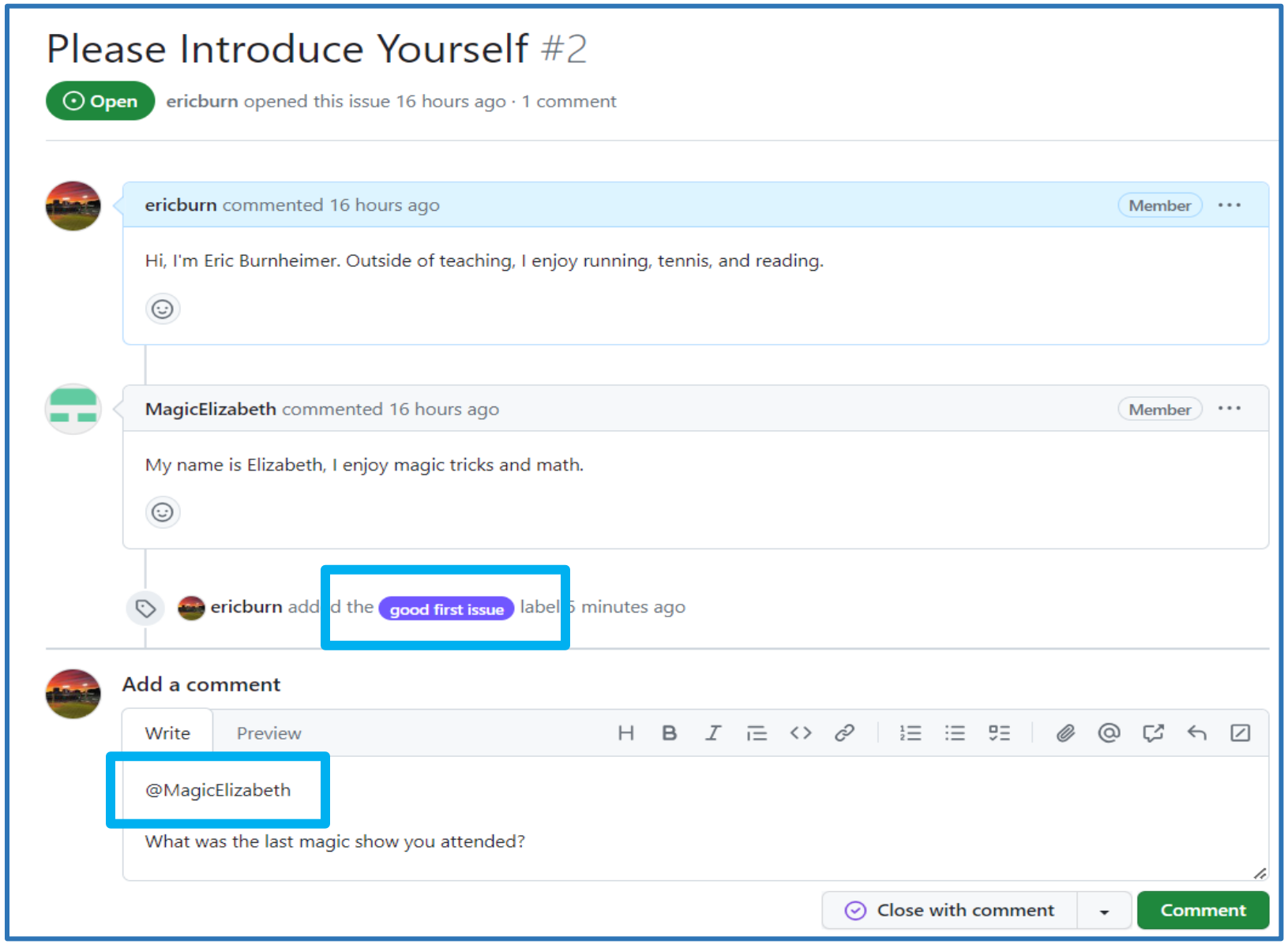
- **GitHub Issues** acts a message board



- Students can ask questions, give feedback, share ideas, etc.

Part II: Introduce GitHub (& Git) through GitHub Issues

- **Labels** can be added to posts; students can be mentioned
- Git (I mean get ☺) students used to the user interface right away



MASSBAY COMMUNITY COLLEGE

Mini-Course After Semester I:

Introduction to Git & GitHub

Intro to Git & GitHub Course / Workshop

MASSBAY COMMUNITY COLLEGE

- Create small groups of students to work together during the course
  - Create files for students to collaborate on in group repos within the organization

Week 1: Core Concepts

- Git Installation
  - Command Line
  - git init
- Basics
  - git add/commit/push
- Correcting Mistakes
  - git reset/revert
- Observing
  - git status/log/diff

Week 2: Branching

- Creating Branches
  - git branch/checkout
- Merging Branches
  - Fast forward/3-way
- Resolving Conflicts
  - git rebase
- Retrieving
  - git pull/fetch

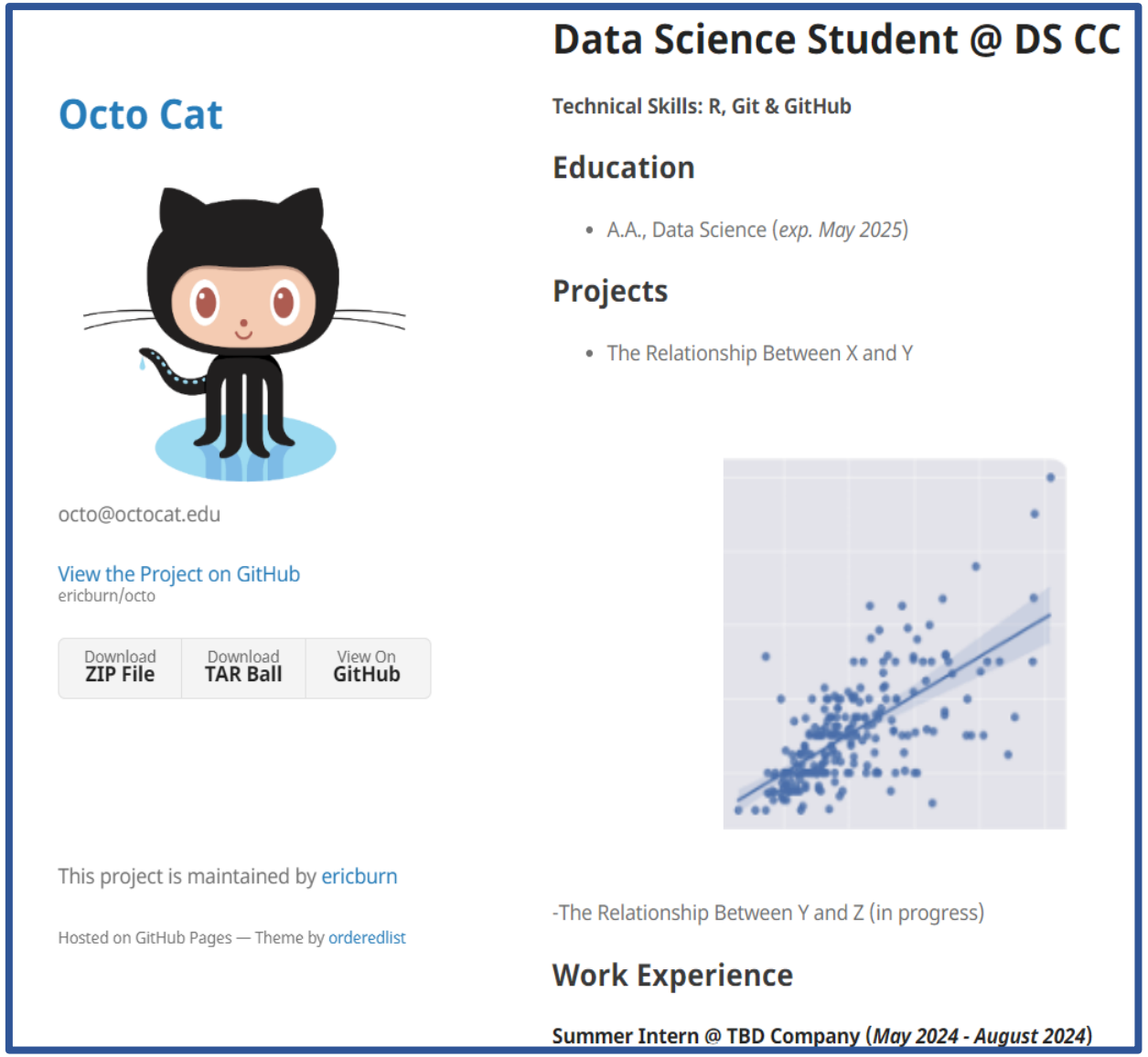
Week 3: Other

- Workflow
  - Pull Requests
- Create Personal Site
  - GitHub Pages
- GUIs
  - GitKraken

Other Topics to Explore

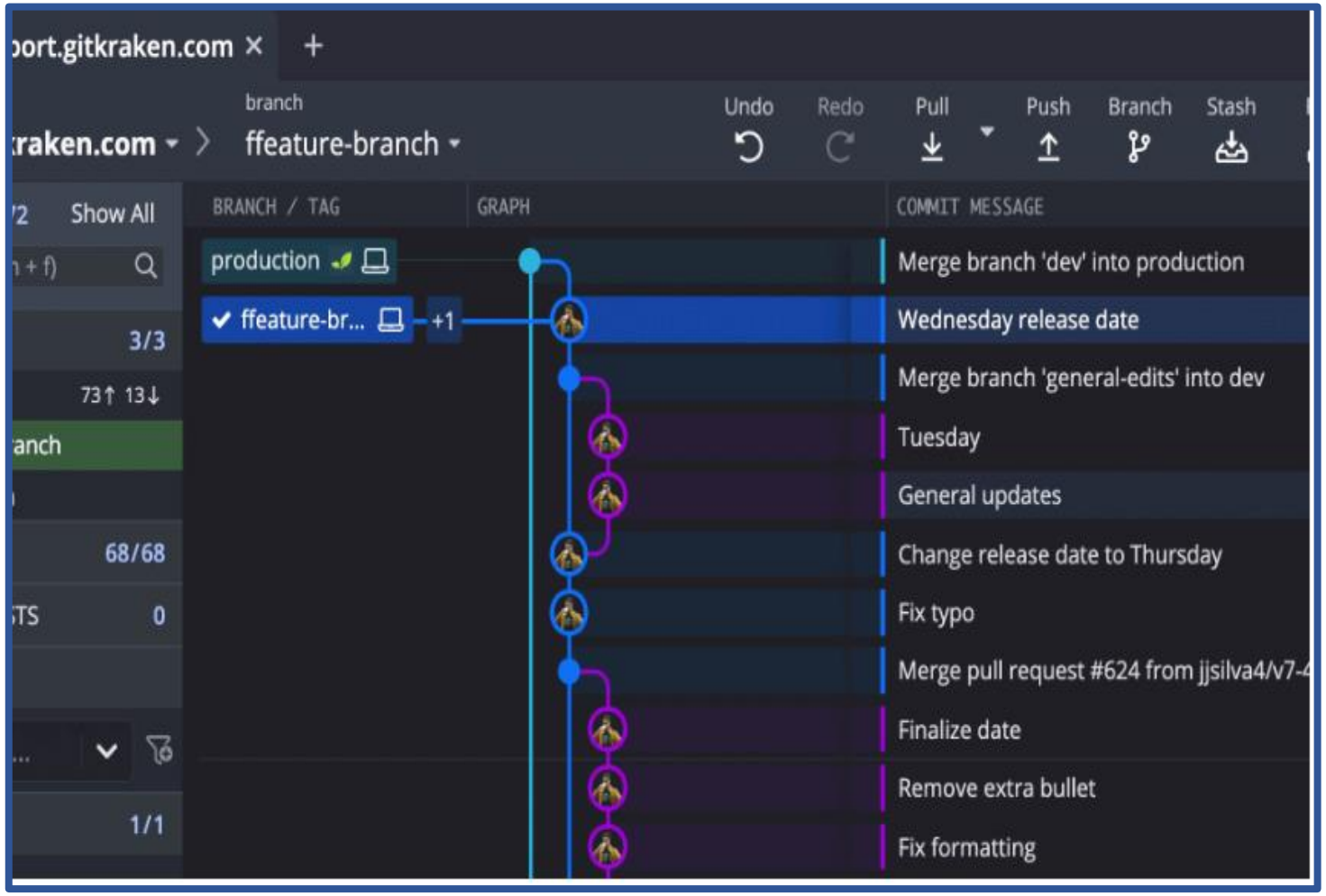
GitHub Pages Site

- **GitHub Pages** enables and eases the creation of a personal site



Graphical User Interface

- **GUIs** such as GitKraken
  - ➢ Nice visualization of branches



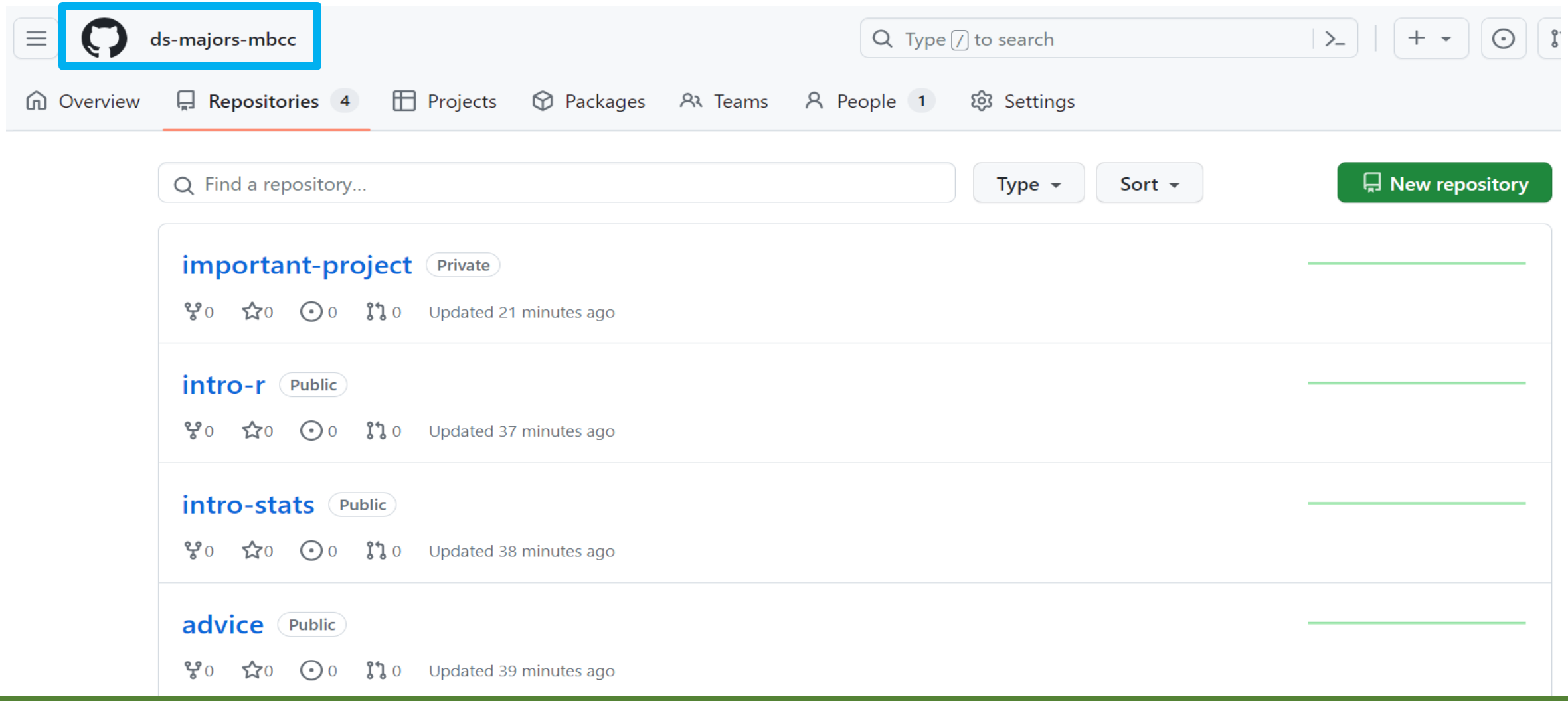
Credit: Medium

MASSBAY COMMUNITY COLLEGE

After Mini-Course: Access to Potential Data Science Majors Organization and Repos

Data Science Majors Organization

- Upon completing the course, students obtain access to the Data Science Majors Organization



Summary

MASSBAY COMMUNITY COLLEGE

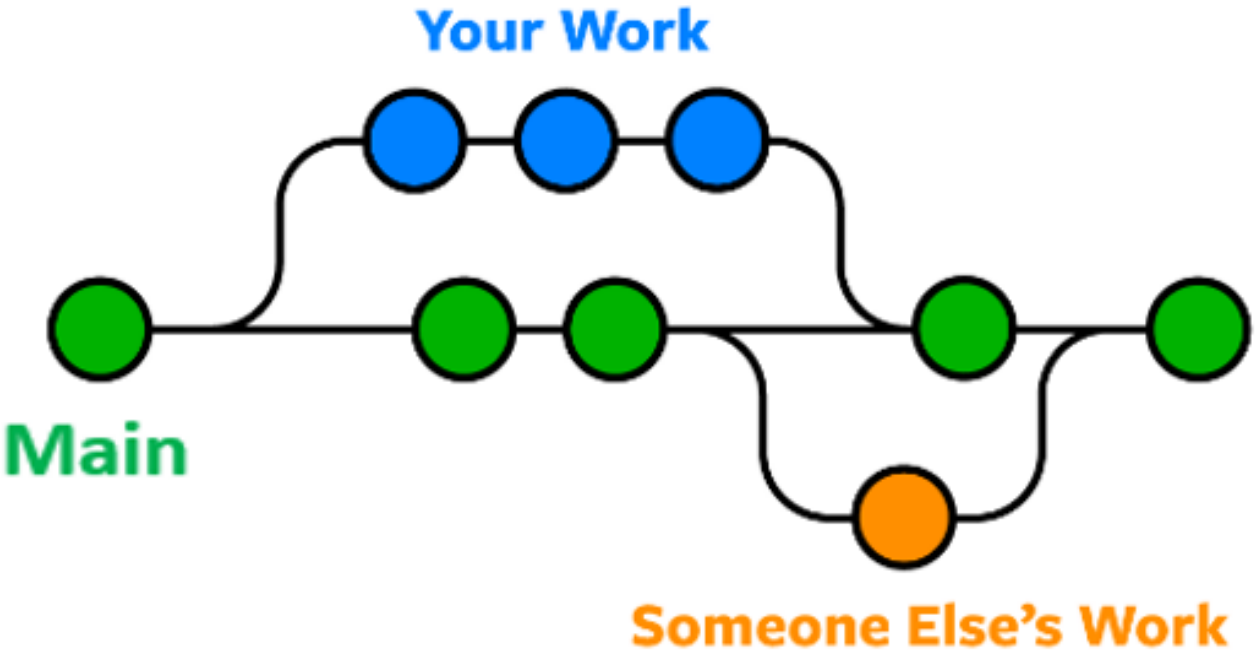
The Benefits of Implementing Git & GitHub in a DS Program

- Potential for Increased Collaboration
- Version Control and Project Management Skills
- Workforce / Transfer Readiness

A Pathway for Implementing Git & GitHub in a DS Program

- Semester I: Introduce GitHub (and preview Git) by communicating via Issues
- Mini-Course / Workshop after Semester I: Introduction to Git & GitHub
- Access to Data Science Majors Organization and Repos

Octo Will Be Ready to Collaborate!



Credit: Noble Desktop

Credit: Charles Kufs

Resources for Learning Git & GitHub

MASSBAY COMMUNITY COLLEGE

Free Resources

- git-scm.com
  - ➢ This is the official git site and includes full documentation on all commands
- *Pro Git* by Scott Chacon and Ben Straub
  - ➢ This book is affiliated with the official git site

Low-Cost Resources

- *The Git & GitHub Bootcamp* by Colt Steele on Udemy
  - ➢ This can be purchased for roughly $10-$15 during one of Udemy's many sales, and it provides great exercises for learning both Git & GitHub

References

Baumer, B. S., & Horton, N. J. (2023). Data Science Transfer Pathways From Associate's to Bachelor's Programs. *Harvard Data Science Review, 5*(1). https://doi.org/10.1162/99608f92.e2720e81

Çetinkaya-Rundel, M., & Rundel, C. (2018). Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician*, *72*(1), 58–65. https://doi.org/10.1080/00031305.2017.1397549

Chacon, S., & Straub, B. (2014). Pro Git (2nd edition). *Apress.*

Glantz, M., Johnson, J., Macy, M., Nunez, J. J., Saidi, R., & Velez, C. (2023). Students' Experience and Perspective of a Data Science Program in a Two-Year College. *Journal of Statistics and Data Science Education*, 1–10. https://doi.org/10.1080/26939169.2023.2208185

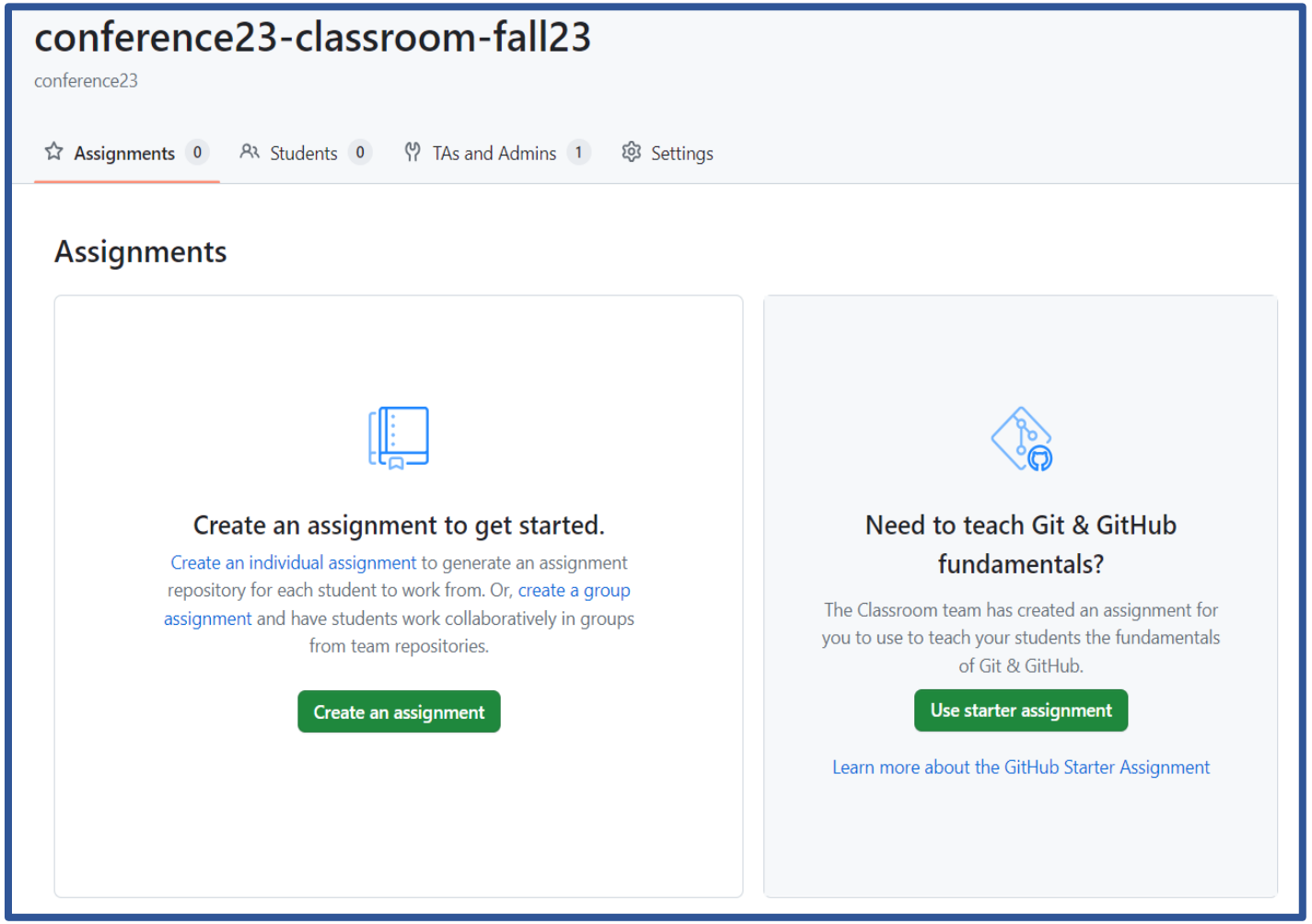
Kotz, B. (2020). The Opportunities of Two-Year College Data Science. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.75aed58b

New Two-Year College Data Science, Analytics Programs on the Rise. (2022, August 1). *Amstat News. https://magazine.amstat.org/blog/2022/08/01/new-two-year-programs/*

Steele, C. (2021). *The Git & GitHub Bootcamp.* Udemy.

Thank you!

Eric Burnheimer
Massachusetts Bay CC
eburnheimer@massbay.edu
ericburn.github.io

GitHub Classroom

- **GitHub Classroom**
  - ➢ Manage a virtual classroom
    - ❖ Make Assignments
    - ❖ Connect to some LMSs