INFO-UB 23: Introduction to Programming and Data Science

Katherine Hoffmann Pham

July 2, 2018

NYU Stern, Department of Information Systems

Welcome!

1. Please let me know if you are not on NYU Classes:
   `https://newclasses.nyu.edu`
2. Please log in to the Jupyterhub server:
   `jupyterhub.ipeirotis.org`
   If necessary, for today you can use Google Colab:
   `https://colab.research.google.com/`
3. Please ensure that you have access to our Slack account:
   `https://info-ub23-summer2018.slack.com/`
4. Please fill out the course survey:
   `https://goo.gl/forms/YDNQH7QCu0FkPW3o1`

Overview

Overview

INFO-UB23: Introduction to Programming and Data Science

aka "dealing with data"

1. Getting and cleaning data (Python)
2. Database design and management (SQL)
3. Visualization and analysis (Pandas, MatPlotLib)

Prerequisites and Focus

- Prerequisites:
  - Your computer and charger
- Focus: Data ...
  - Collection
  - Storage
  - Organization
  - Management
  - Analysis
- Not the focus:
  - Machine learning
  - Data mining
  - Big data

Why INFO-UB 23?

"Starting in 2012, my colleagues and I began taking a closer look at the hands-on experience of data scientists ... What we found was that the bulk of their time was spent manipulating data – a mix of data discovery, data structuring, and creating context. In other words, most of their time was spent turning data into a usable form rather than looking for insights."

- HBR, "The Sexiest Job of the 21st Century is Tedious, and That Needs to Change"

Course Philosophy

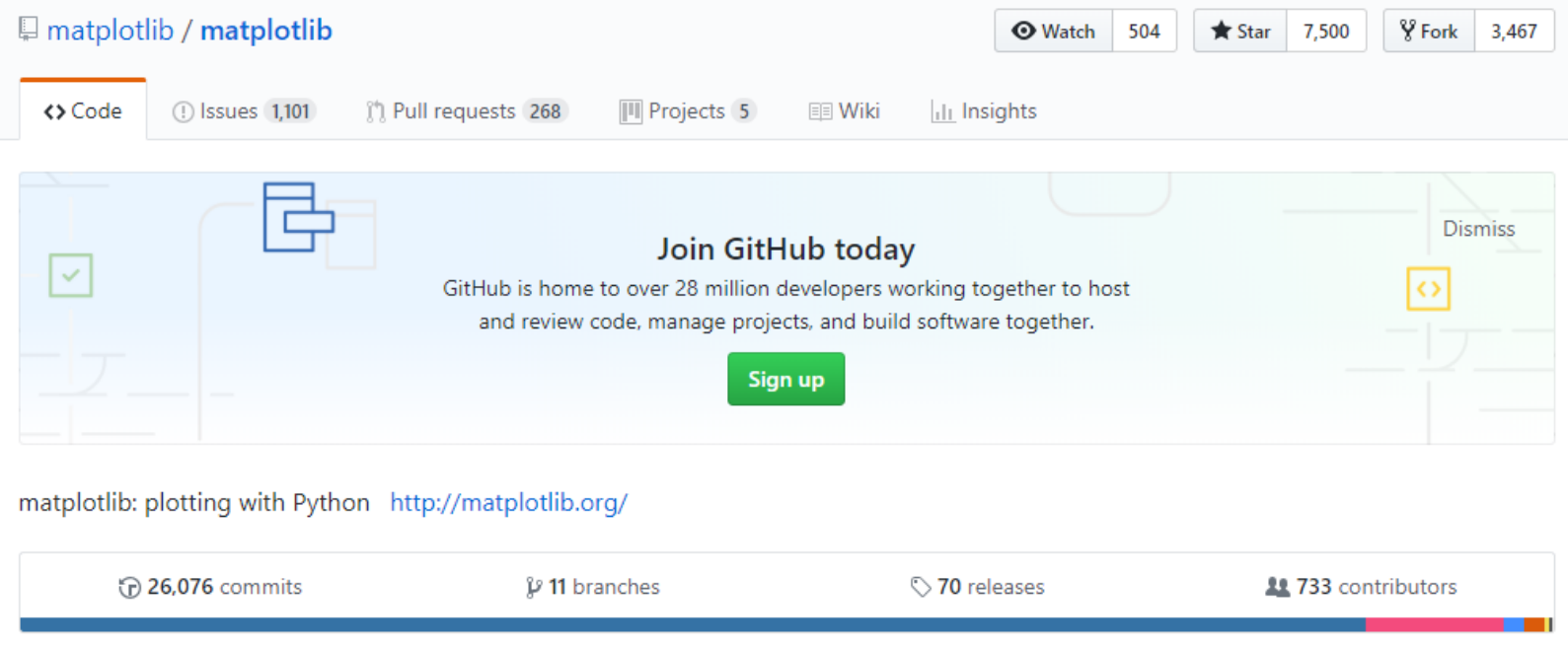
1. If it is repetitive, automate it.

Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.

Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.

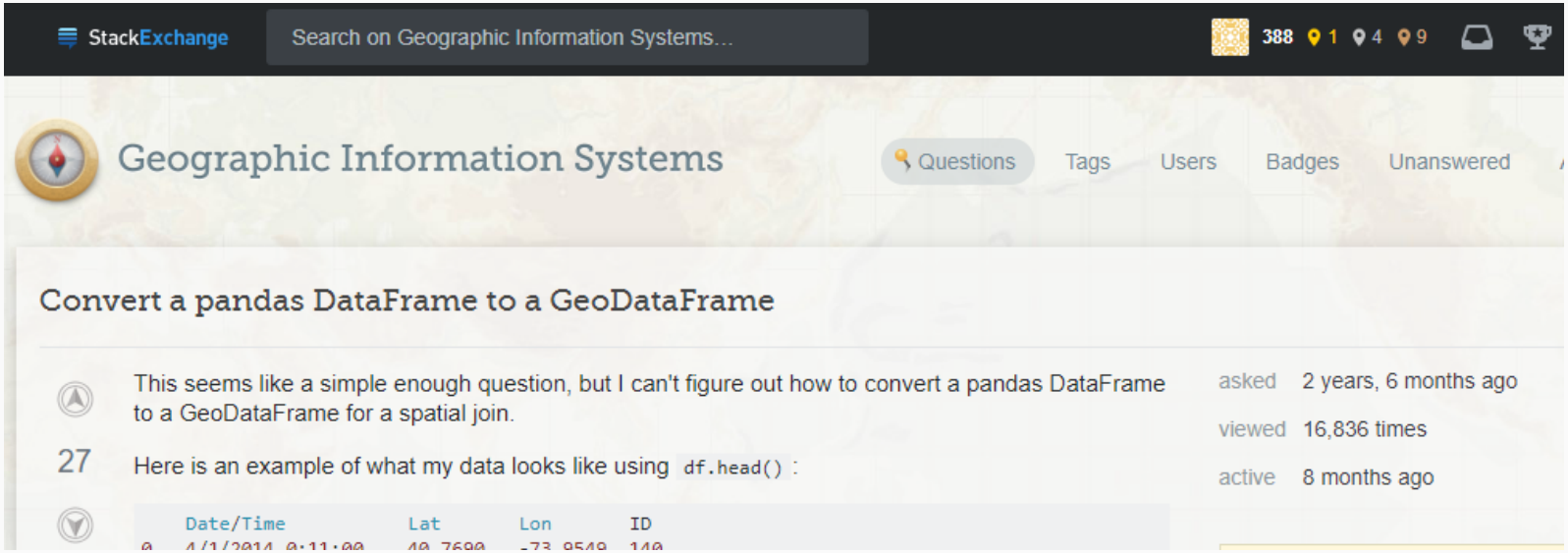


Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.
4. Expertise is relative.

Course Philosophy

1. If it is repetitive, automate it.
2. Build a reproducible workflow.
3. Data science is collaborative.
4. Expertise is relative.
5. Everyone has something to contribute.



Course Philosophy

Implications:

1. Stop me to ask questions at any time
2. Turn to each other for help

Where to next?

- INFO-UB 24 Projects in Programming and Data Science
- INFO-GB 3106 Data Visualization
- INFO-UB 57 Data Mining for Business Analytics
- NYU Center for Data Science
- Real life . . .

Introductions . . .

Let's give the short version of the survey:

- Name
- What you study
- Why this class?
- Experience working with data?
- Hobbies/interests

Course Logistics

Course page

Announcements, grades, materials via NYU classes:

`https://newclasses.nyu.edu/`

Otherwise, see the course webpage:

`http://people.stern.nyu.edu/khoffman/intro_programming_datasci/`

Course notebooks

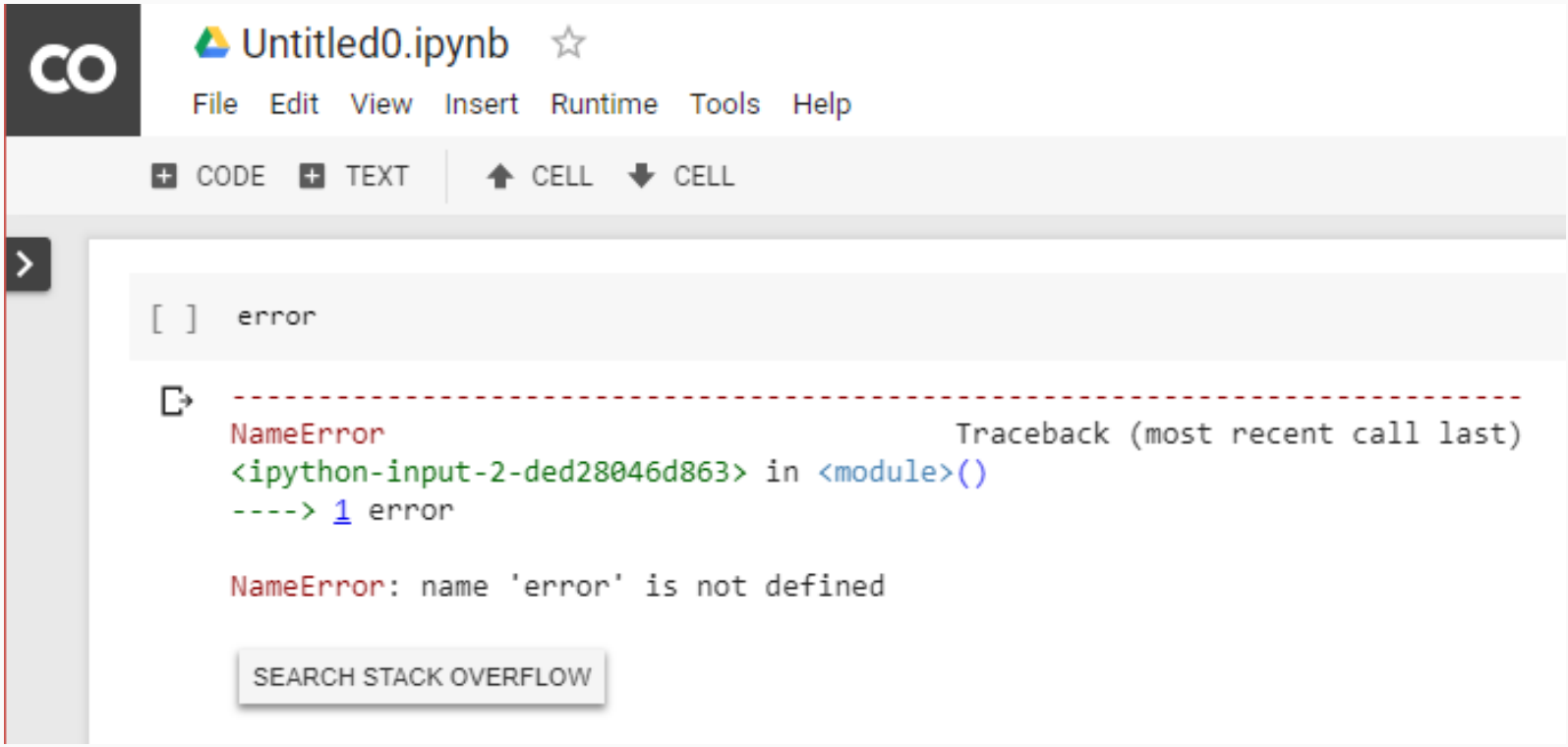
The course notes are available on Github.

- Our repository is:
  `https://github.com/khof312/Summer2018_ProfHoffmannPham`
- This is a fork of a master repository:
  `https://github.com/ipeirotis/dealing_with_data`

Getting Help

For troubleshooting and help, I recommend:

1. Google



Getting Help

For troubleshooting and help, I recommend:

1. Google
2. StackOverflow: `https://stackoverflow.com/`

Getting Help

For troubleshooting and help, I recommend:

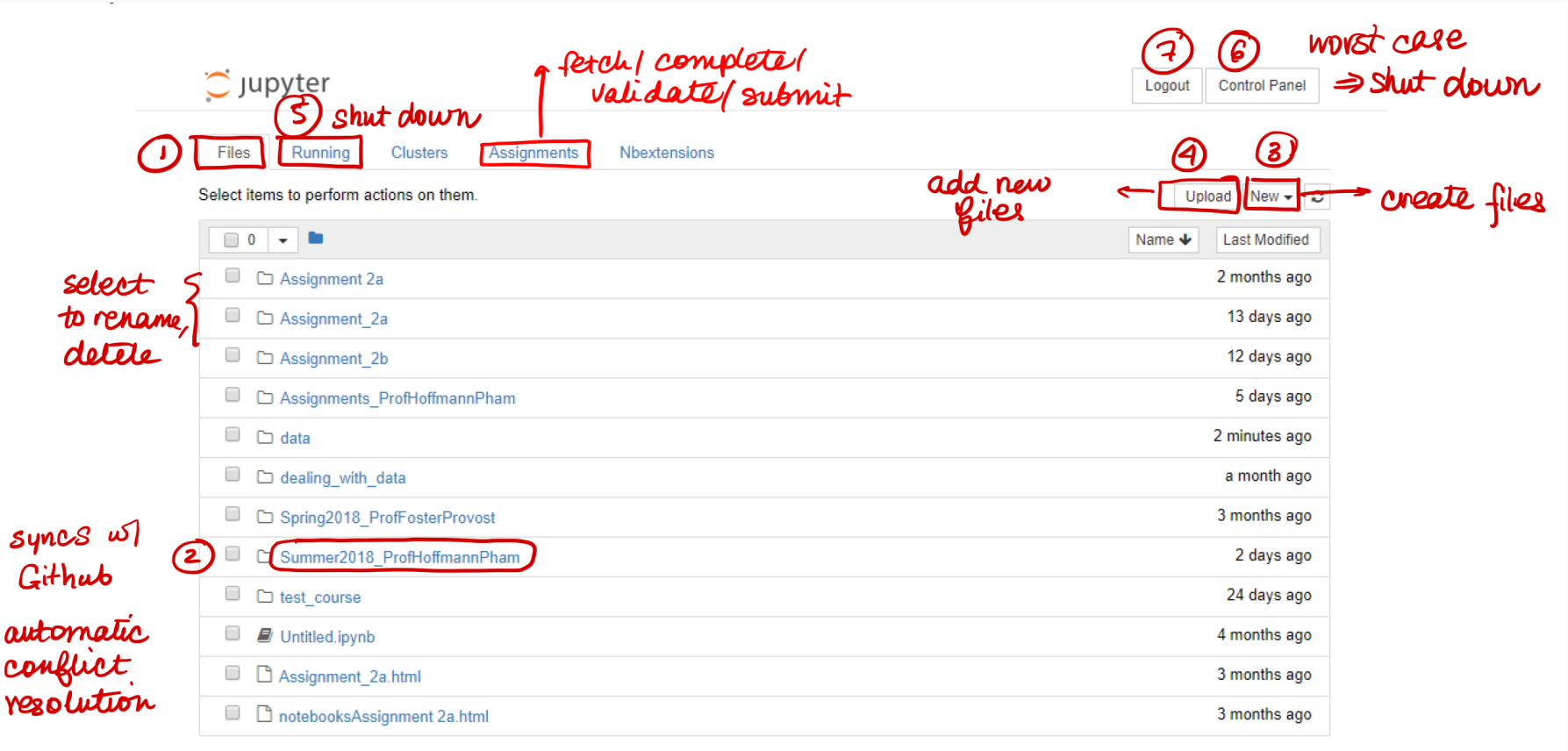
1. Google
2. StackOverflow: `https://stackoverflow.com/`
3. Slack: `https://info-ub23-summer2018.slack.com/`

Getting Started

Jupyterhub overview

- Jupyterhub is our online data science environment
- I encourage you to use it for several reasons:
  1. It provides a standardized environment
  2. I can automatically update the available files
  3. Homeworks can easily be tested and submitted

Jupyterhub overview



Jupyter notebook overview