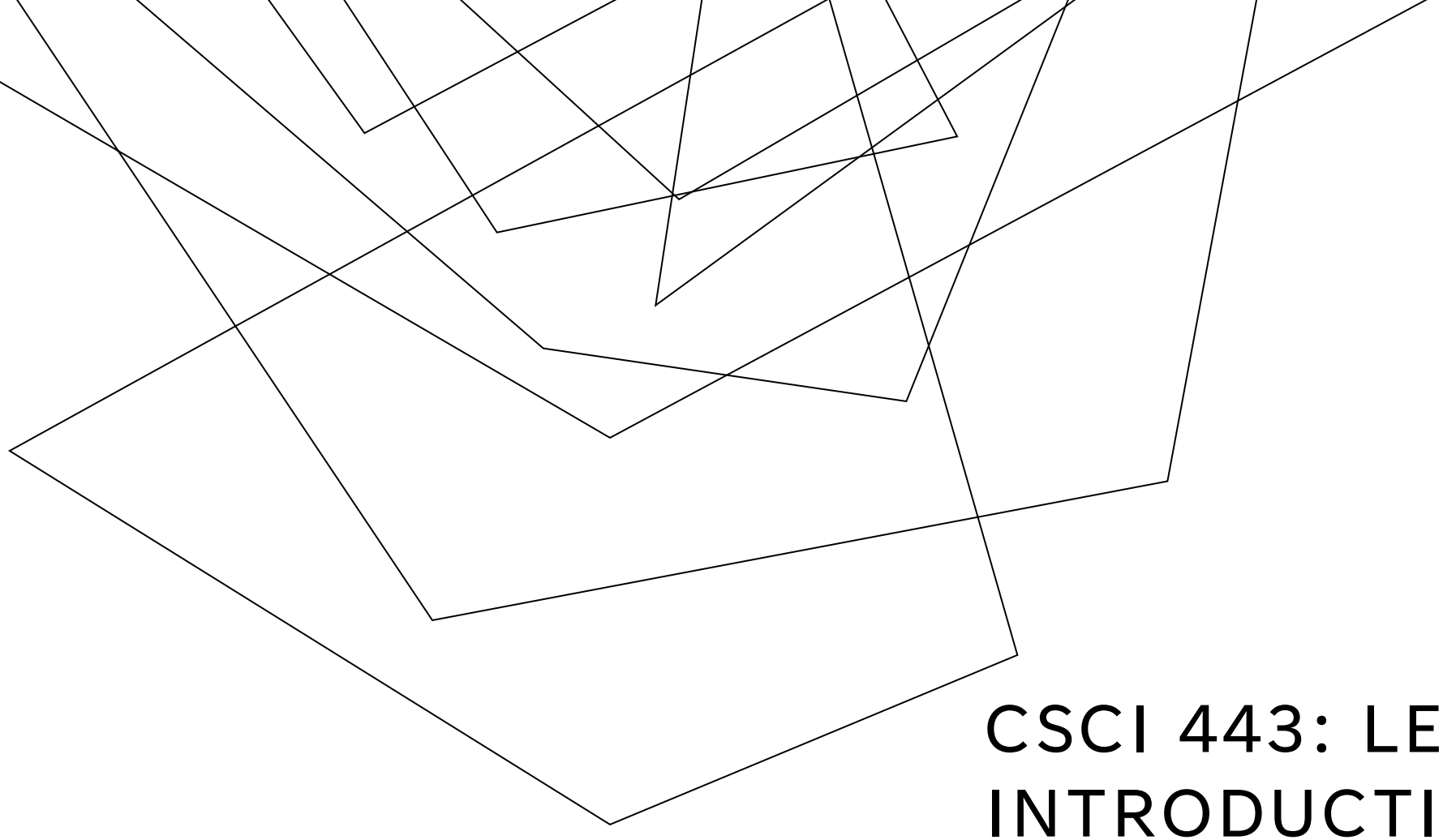
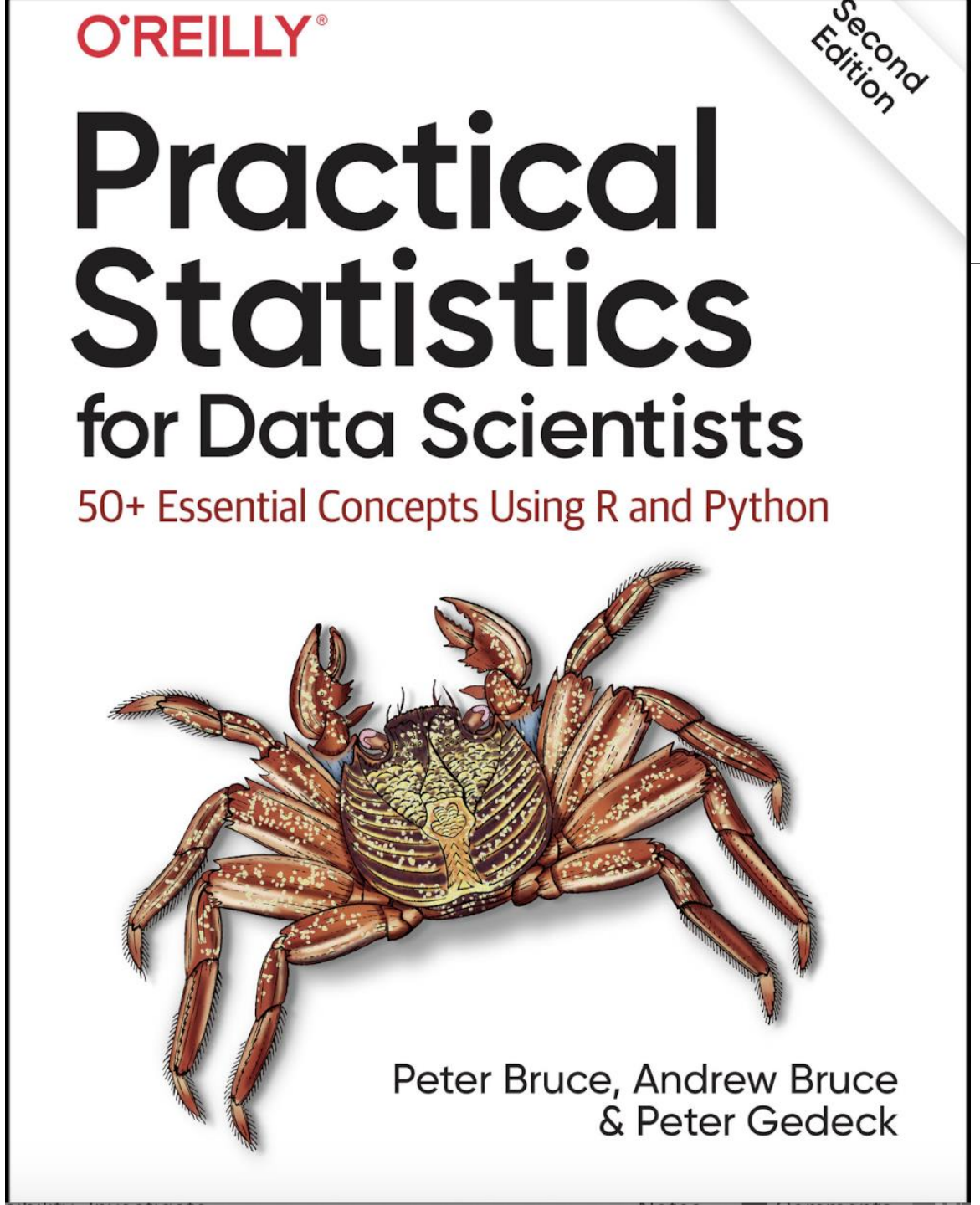


CSCI 443: LECTURE 1 INTRODUCTION, AND NUMPY

Professor David Harrison

TODAY: INTRODUCTION

- Who am I?
- Syllabus
- What is data science?
- What is data engineering?
- What does this course cover?
- What tools will we use?
- Some review of practical statistics
- Chapter 1 Practical Statistics
  - Up to page 18 in chapter 1.



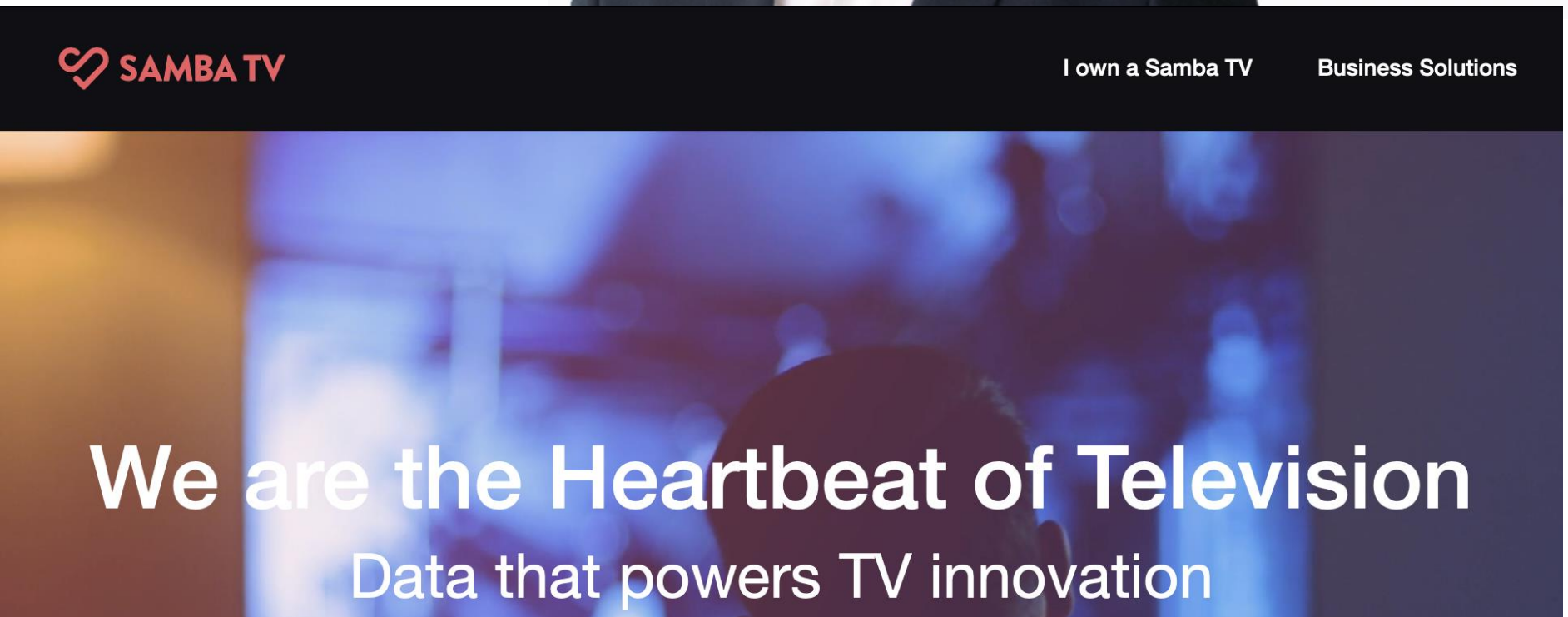
- Assistant professor in CS
- Previously

Dave Harrison

CTO & Co-Founder

David leads Samba engineering, ops and R&D.

Prior to Samba, David launched BitTorrent.org and invented BitTorrent's Streaming protocol. He previously held a post-doctoral position in the Video and Image Processing Lab at UC Berkeley.



SYLLABUS

CSCI 443 Advanced Data Science

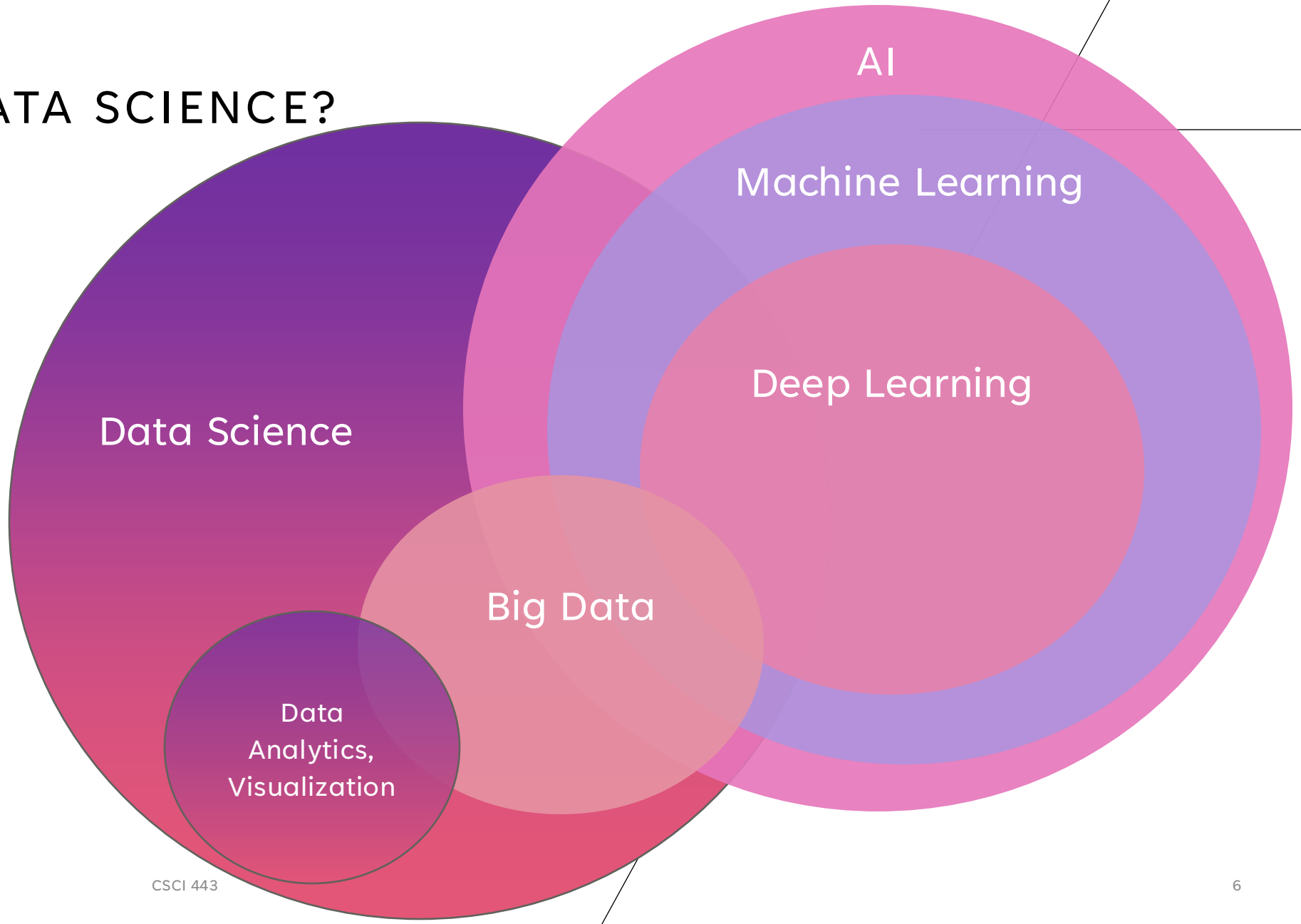
Course Overview

Building on CSCI 343, this course covers key statistical methods and engineering processes in data science, with a focus on large dataset analysis. Students will explore experiment design, data visualization, and data pipeline creation. Hands-on projects will highlight skills in both batch and real-time processing, preparing students for practical challenges in the field.

WHAT IS DATA SCIENCE?

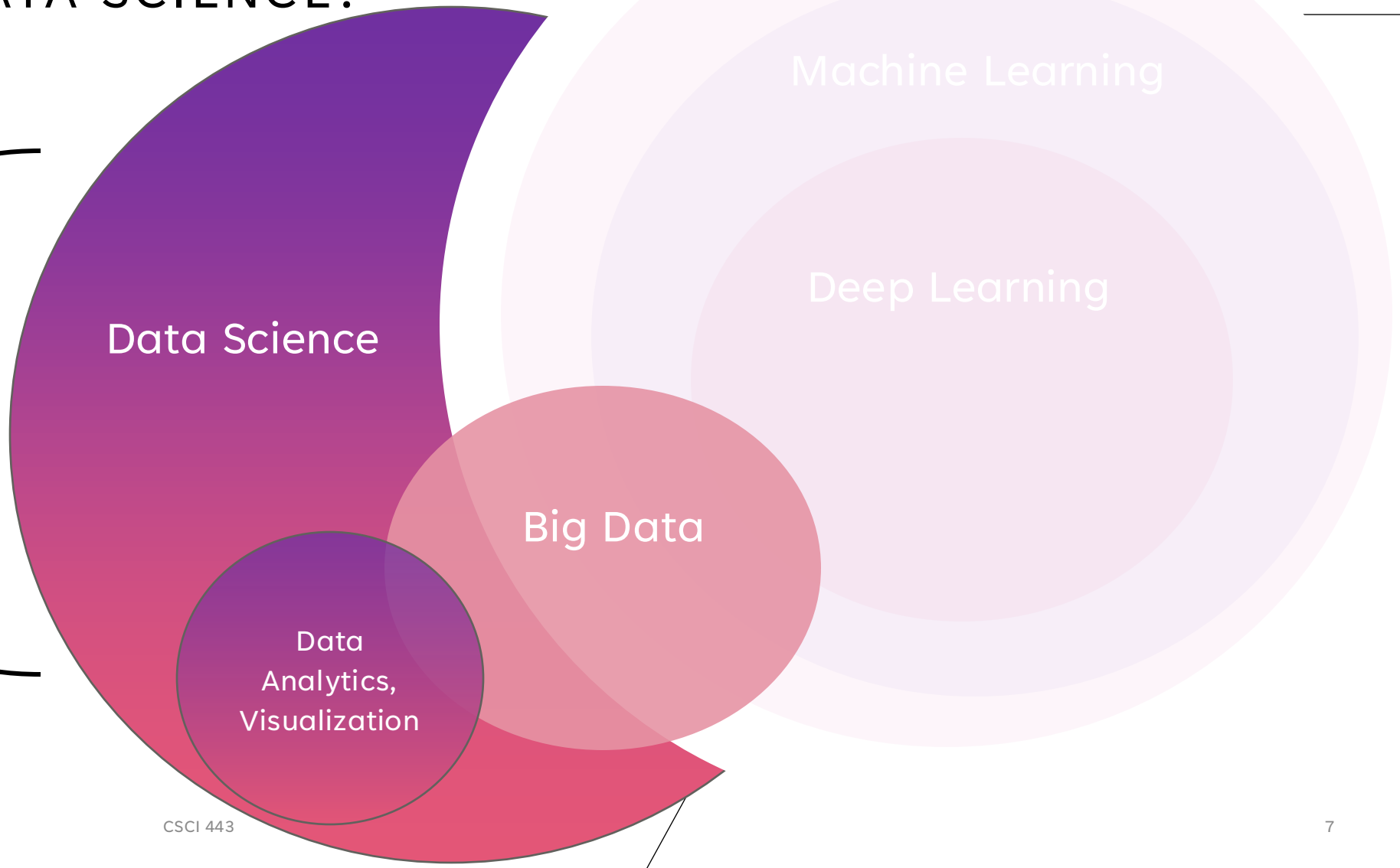
Data science encompasses the entire lifecycle of data processing and analysis, including data collection, cleaning, exploration, modeling, and interpretation. Its focus is on extracting insights and knowledge from data and involves developing methods of recording, storing, and analyzing data to effectively extract useful information.

WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?

This class focuses on data science.



DATA SCIENTISTS VS DATA ENGINEERS

- Typically 2-5 data engineers to each data scientist.
- Data scientist is often both internally and externally facing.
- Data scientist interfaces with key stakeholders to
  - Define a hypothesis, problem, question, ...
  - Design metrics
  - Design the experiments to answer the question.
  - Work with data engineers to understand, clean, and analyze data.
- Data engineers build it:
  - Data collection
  - Data Warehousing / Data Lakes
  - Cloud computing
  - Data pipelines
  - Dashboards and automated reporting
  - Data governance and security.

DATA SCIENTISTS VS DATA ENGINEERS

- Typically 2-5 data engineers to each data scientist.
- Data scientist is often both internally and externally facing.
- Data scientist interfaces with key stakeholders to
  - Define a hypothesis
  - Design metrics
  - Design the experiments to answer the question.
  - Work with data engineers to understand, clean, and analyze data.
- Data engineers build it:
  - Data collection
  - Data Warehousing / Data Lakes
  - Cloud computing
  - Data pipelines
  - Dashboards and automated reporting
  - Data governance and security.

This class

DATA SCIENTISTS VS DATA ANALYSTS

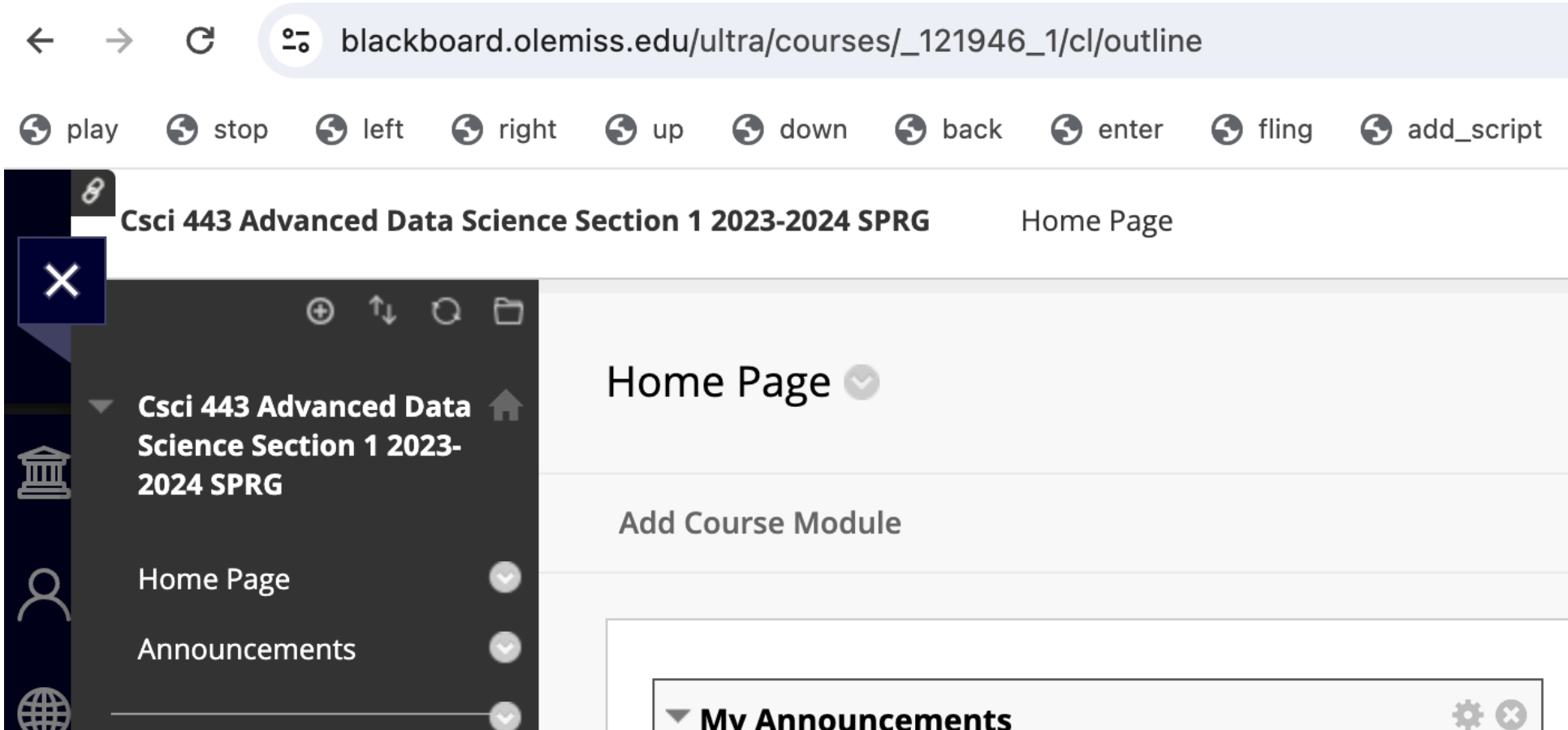
- Data scientists are both engineers and statisticians.
  - work on more complex and abstract tasks, such as developing new analytics methods, predictive models, and machine learning algorithms.
  - Often involved in research and development.

- Data analysts are skilled in data manipulation and visualization.
  - Often use processes put in place by a data scientist.
  - Often use tools implemented by data engineers.
  - Often support executives and sales
  - May be externally facing.
  - Often not well versed in engineering

TOOLS

BLACKBOARD

All lecture slides, homeworks, and solutions will appear on blackboard.



GITHUB

Example files I create during class will be put on github.

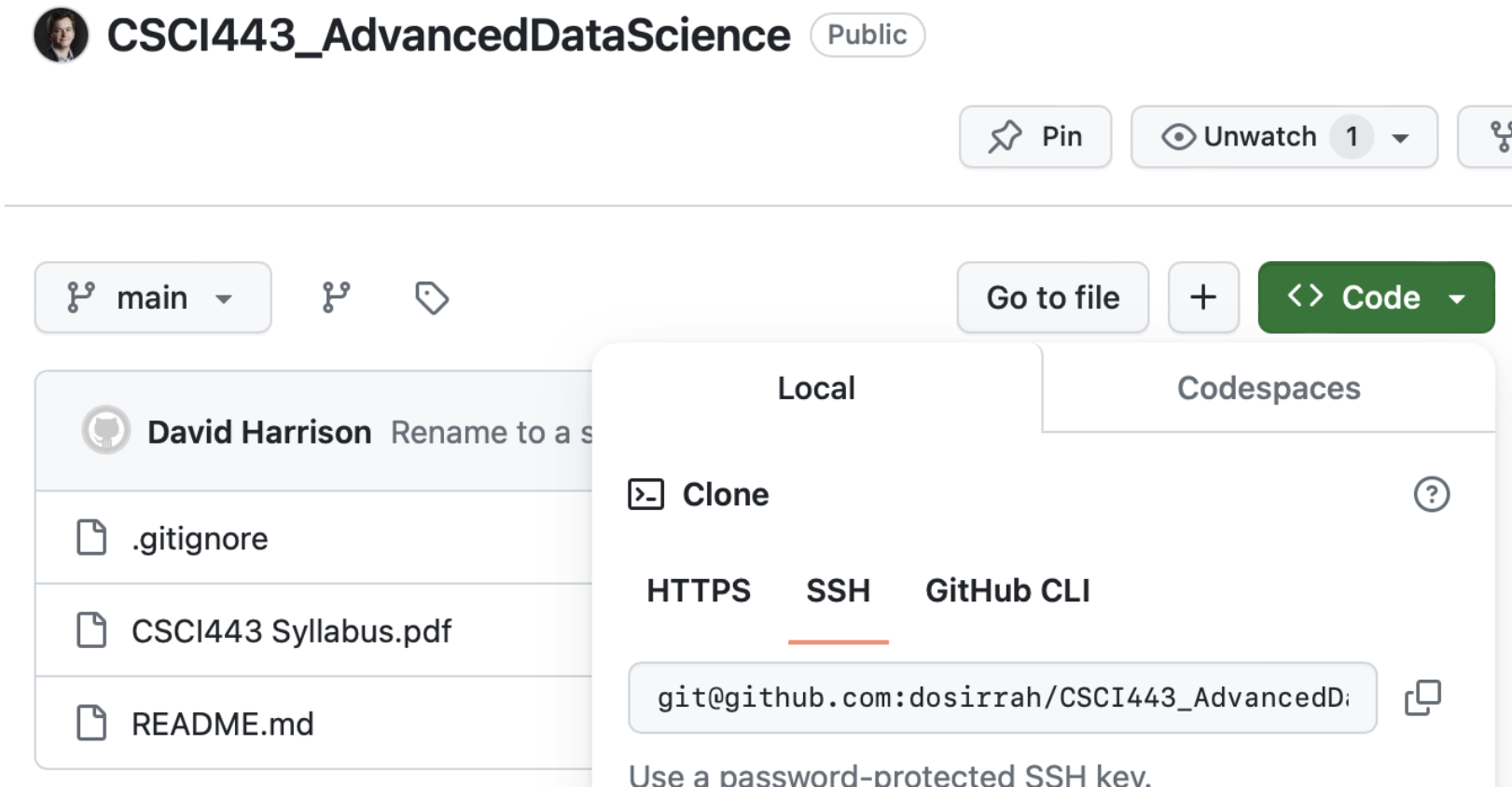
The project is at

https://github.com/dosirrah/CSCI443_25S_AdvancedDataScience

You will need to create a Github account independent of your olemiss accounts.

GitHub is free for our purposes.

I highly recommend committing any code you create to GitHub.

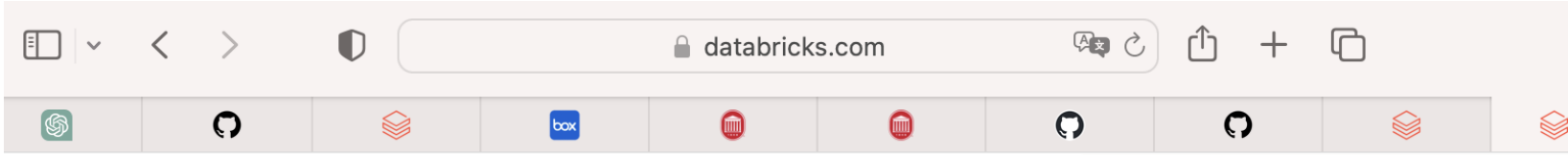


NO GITLAB...

Last year I used the department Gitlab. This semester I will only use github.

DATABRICKS

We will use the databricks community edition.
https://community.cloud.databricks.com/login.html



databricks

Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud. Sign-up with your work email to elevate your trial experience.

Create your Databricks account 1/2

Sign up with your work email to elevate your trial with expert assistance and more.

First name

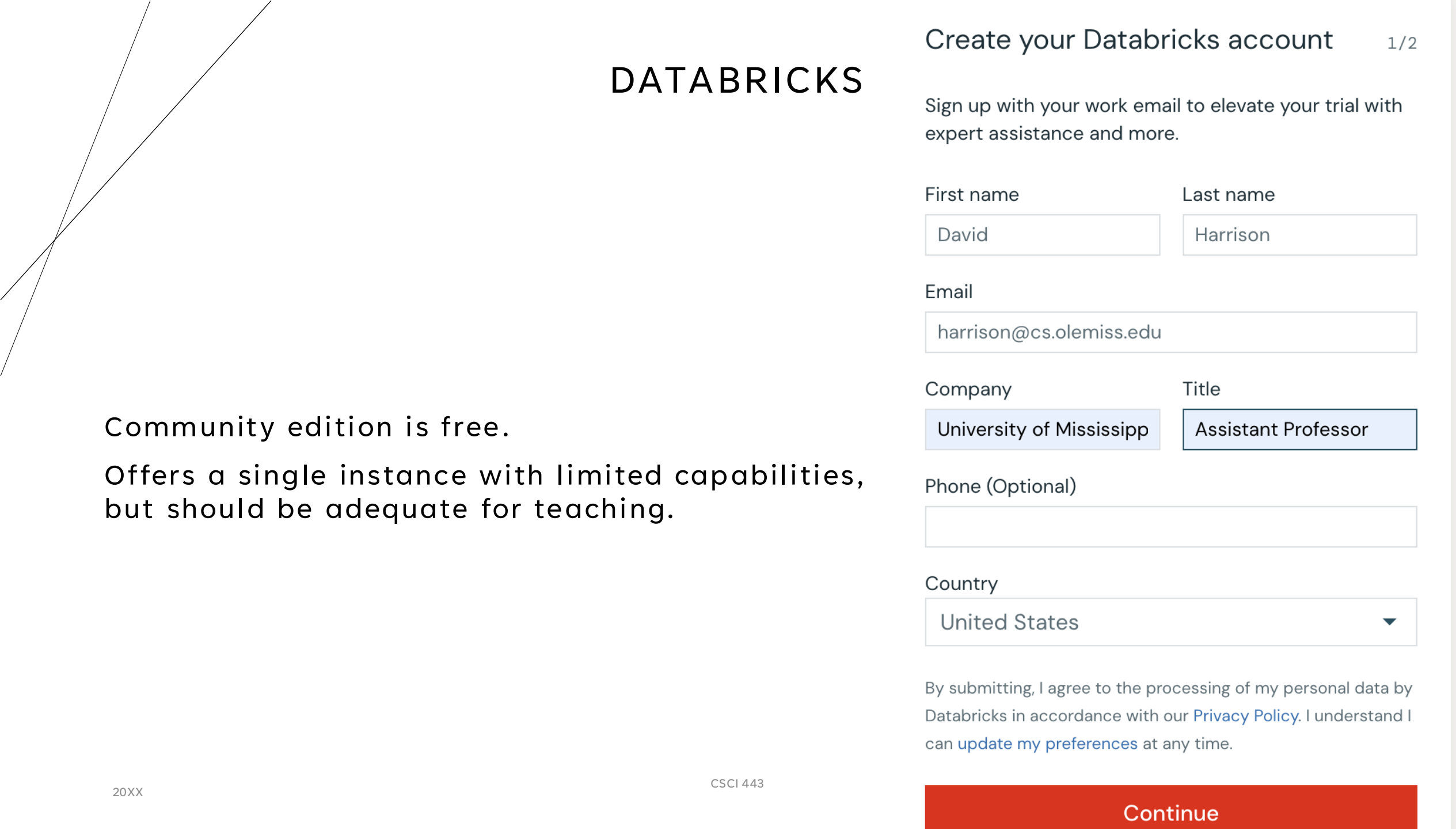
Last name

Email

DATABRICKS

Community edition is free.

Offers a single instance with limited capabilities, but should be adequate for teaching.

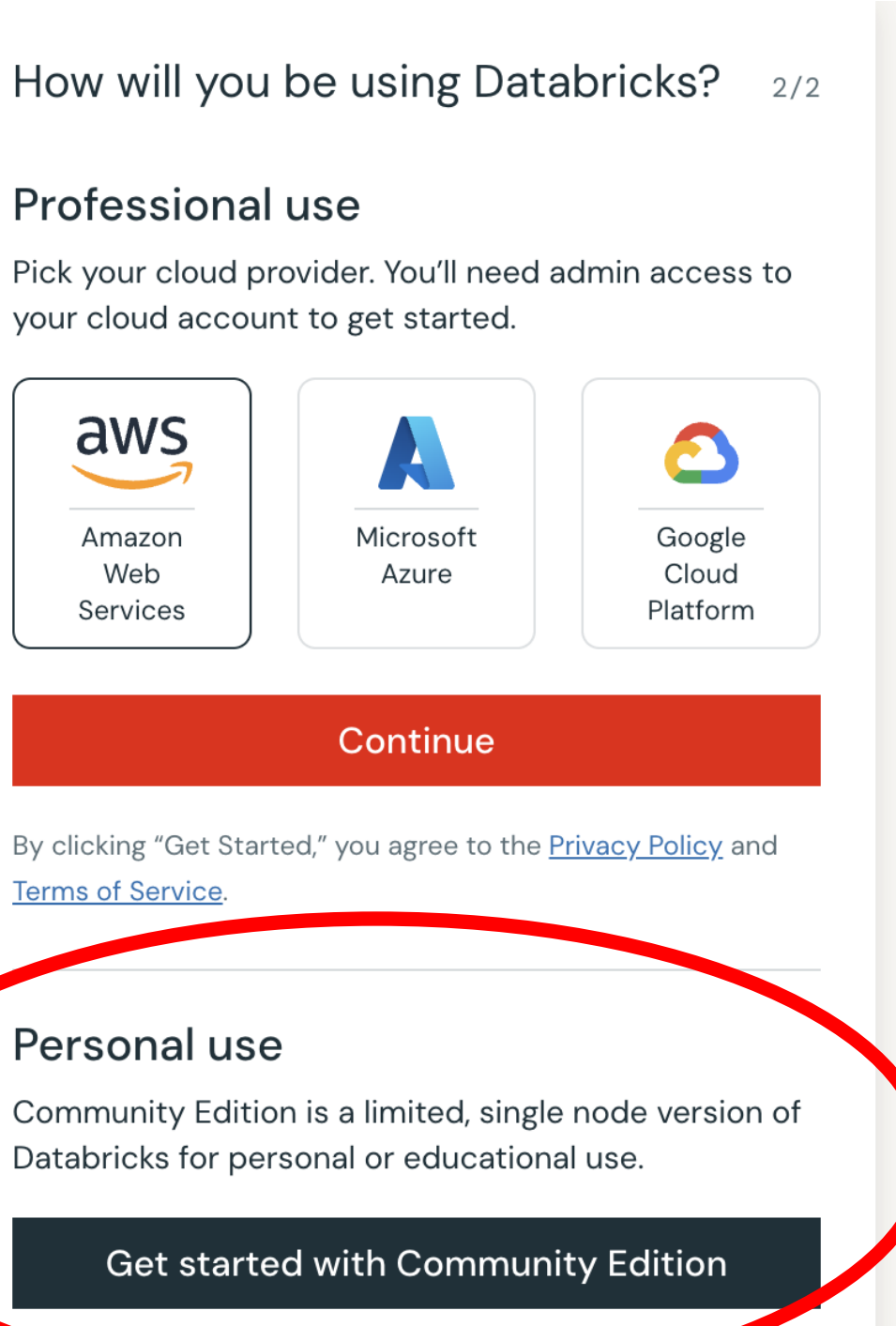


DATABRICKS

Community edition is free.

You do not need an AWS or Azure account.

You do not need to sign up for the 14-day trial.

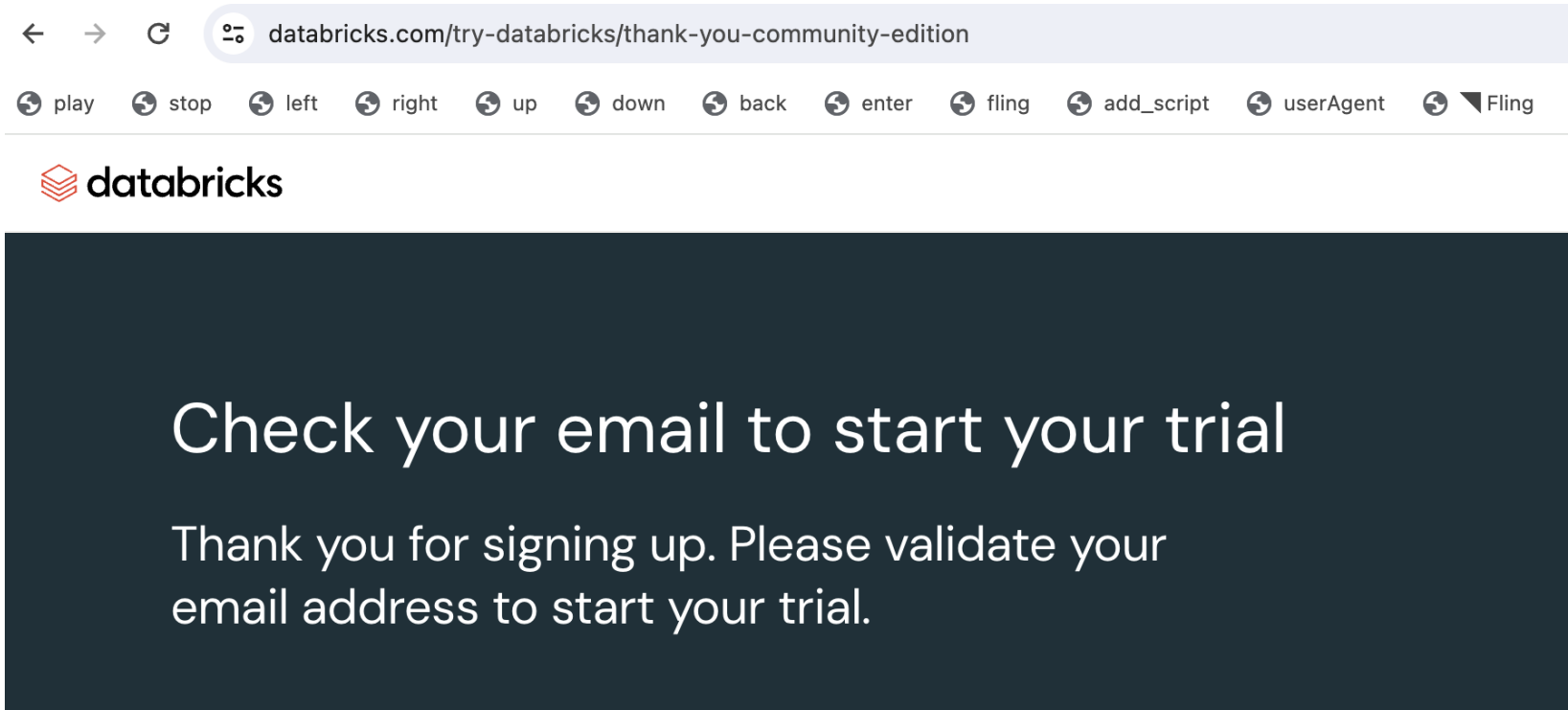


DATABRICKS

Community edition is free.

Don’t worry about “your trial.”

This is misleading.

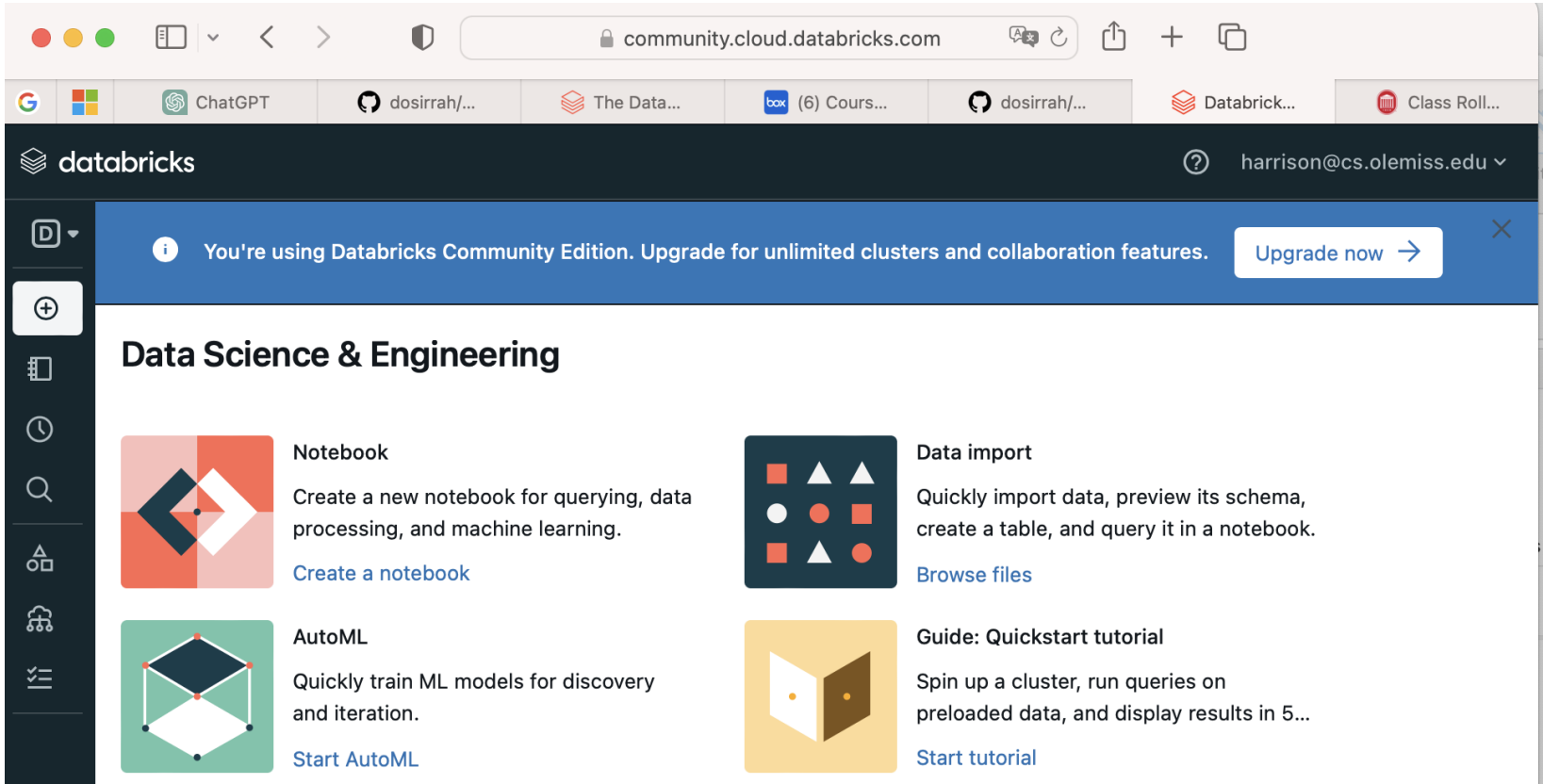


DATABRICKS

https://community.cloud.databricks.com

Once logged in, you should see options to start a notebook and to import data.

Ignore “Upgrade now”

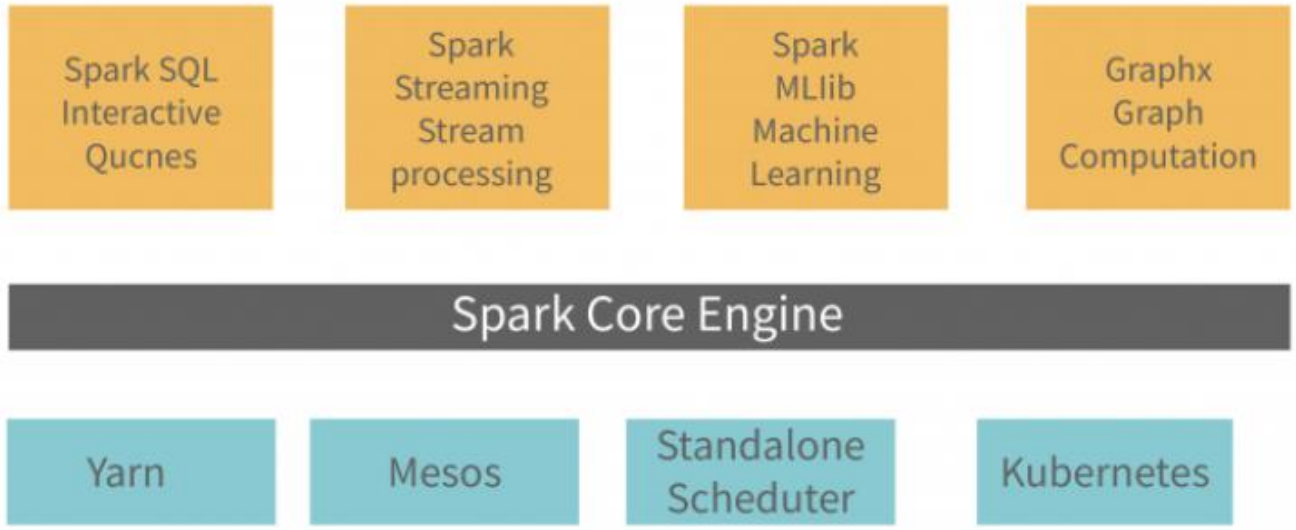


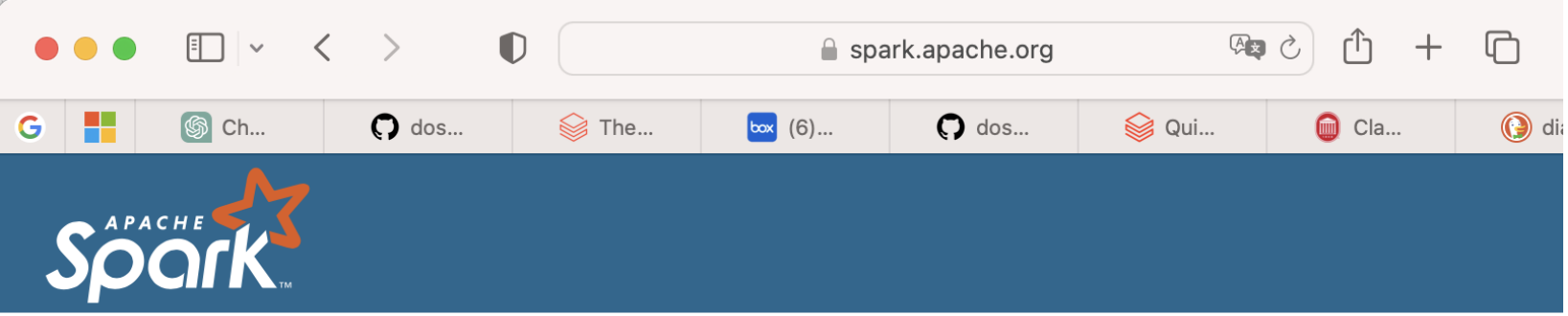
WHY DATABRICKS?

Databricks provides cluster management and a notebook (akin to Jupyter) interface to Apache Spark.

Spark unifies:

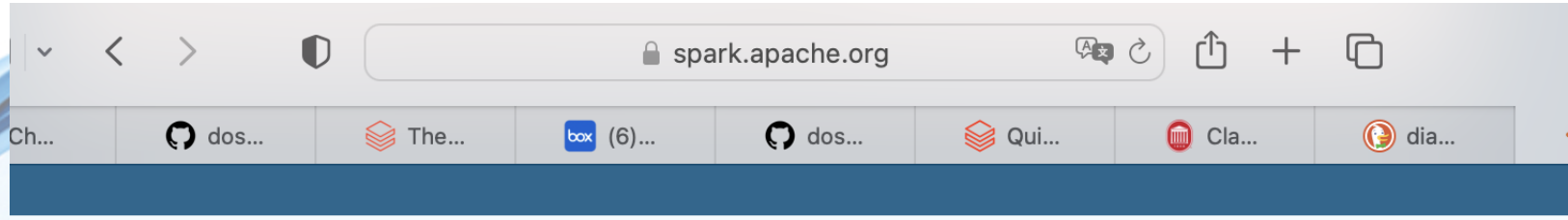
- Batch processing
- Real-time data processing
- Stream analytics (trending, dashboards, etc.)
- Machine learning
- Interactive SQL
- Successor and extension to what was traditionally done with Hadoop or other map-reduce systems.





Unified engine for large-scale data analytics

GET STARTED

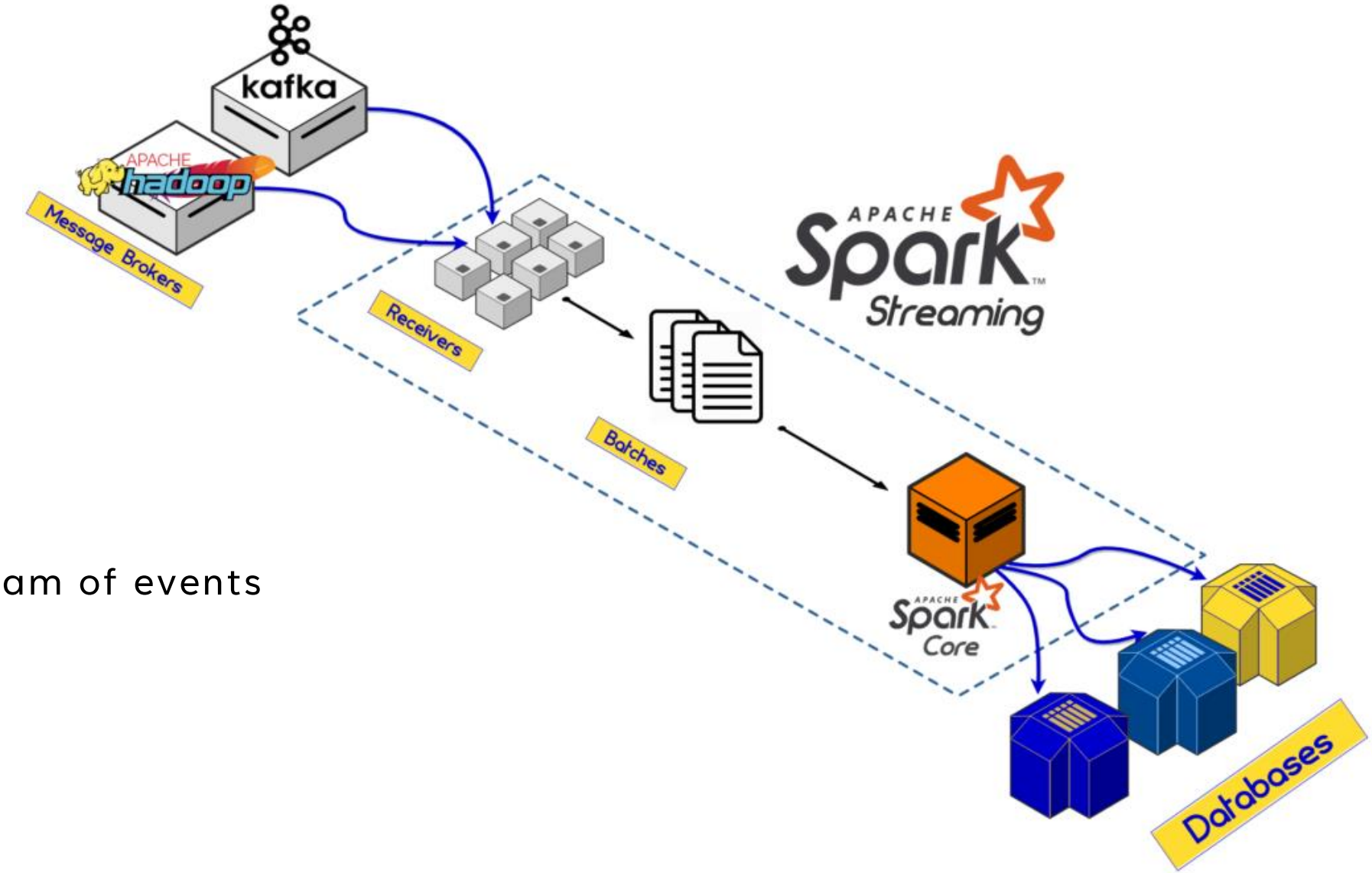


The most widely-used engine for scalable computing

Thousands of companies, including 80% of the Fortune 500, use Apache Spark™.

Over 2,000 contributors to the open source project from industry and academia.

APACHE SPARK



Real-time

- Processing a stream of events

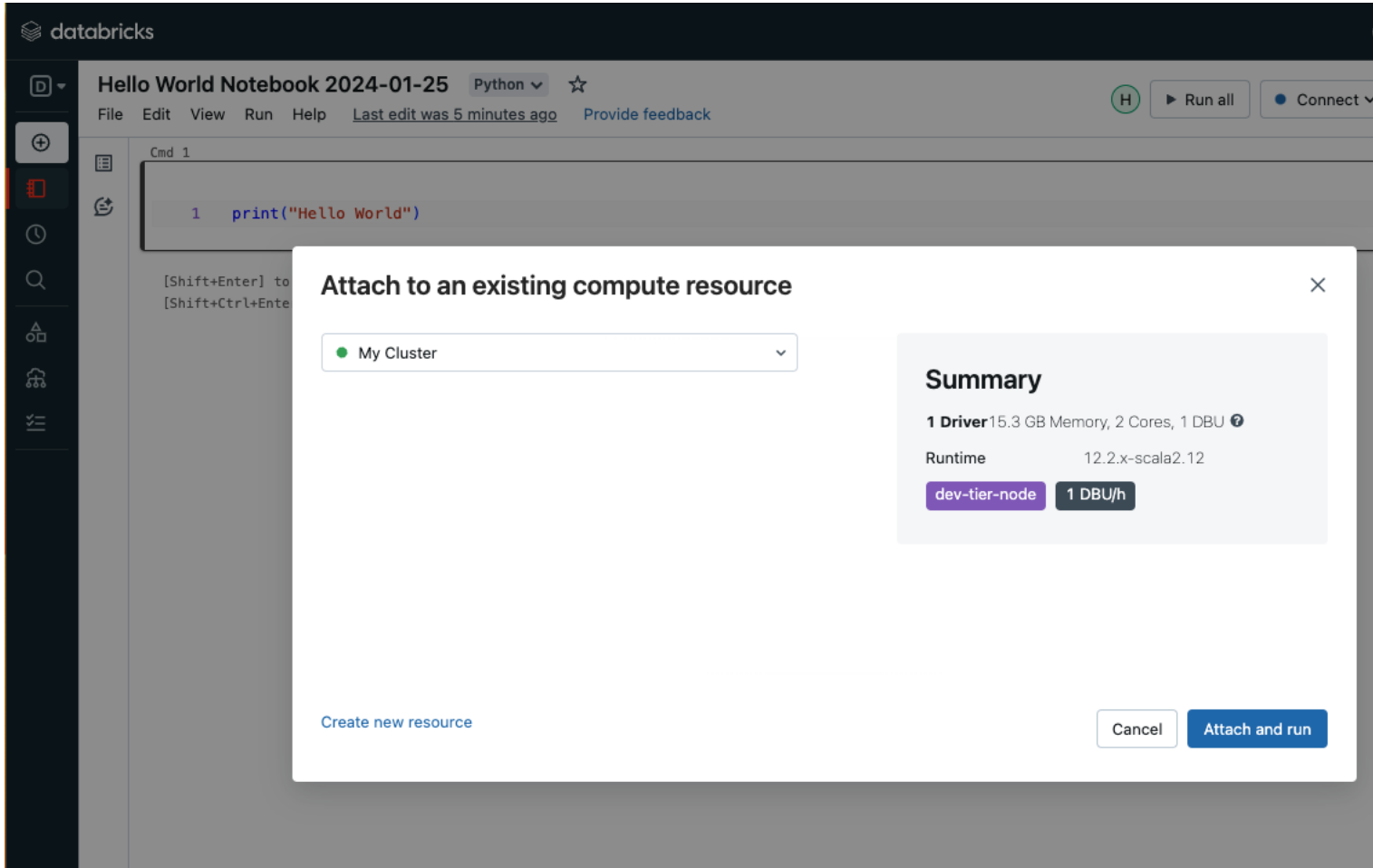
DATABRICKS NOTEBOOKS

Analogous to Google Colab.

Uses Apache Spark.

When you execute a notebook, the first command may take a while:

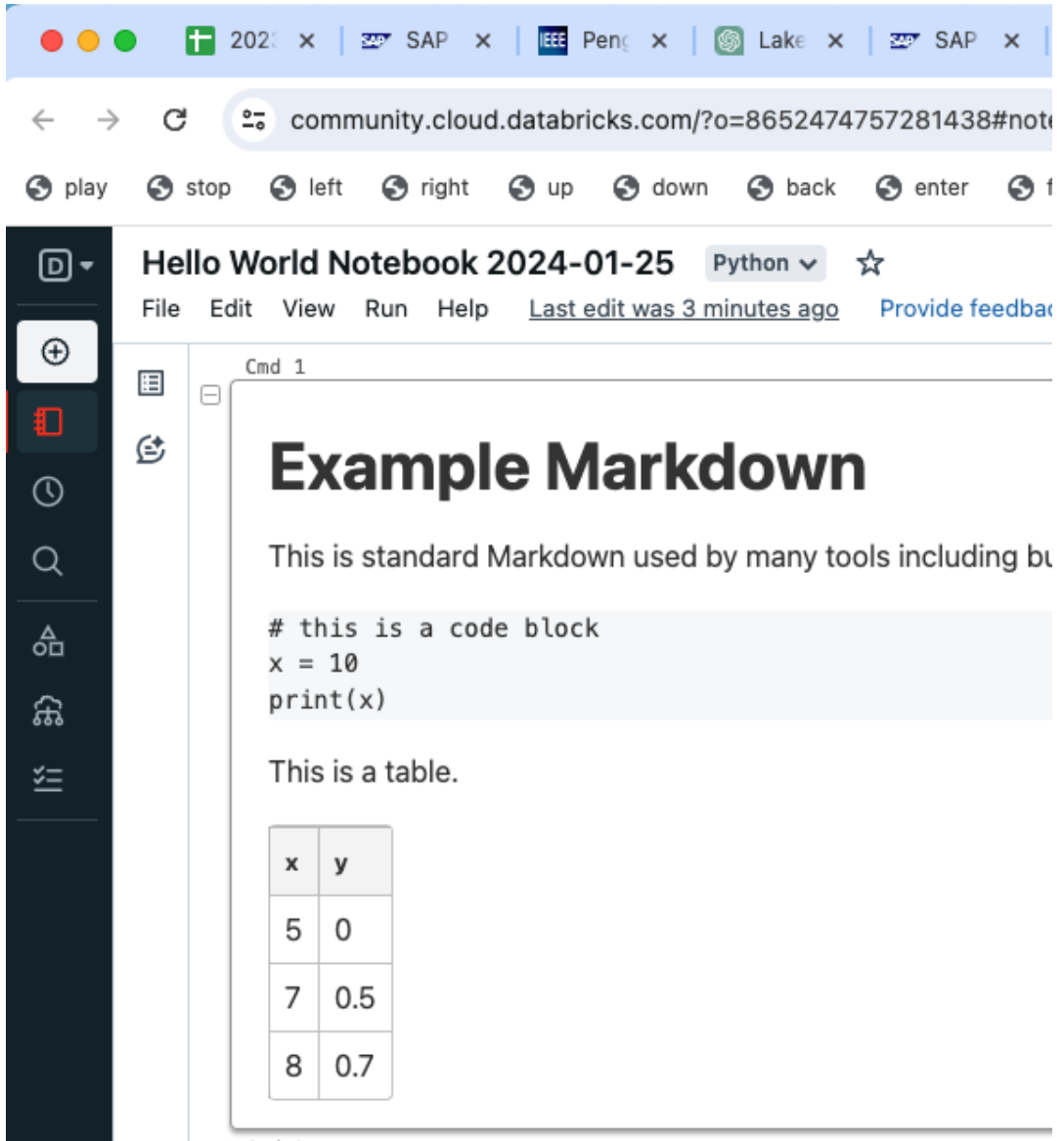
1. Starts a Spark cluster if one is not already running.
2. Attaches to the Spark cluster via the notebook interface.
3. Then you can start entering commands.
4. Much faster if your cluster is already running.



DATABRICKS NOTEBOOKS (2)

Within a notebook you can

- Run Python or R commands
  - **in this course I will focus on using Python**
- Include textual description using markdown
- Use Apache SQL
- Embed visualizations using matplotlib or seaborn.
- ~~Connect to github to commit your work.~~ (not possible unless someone figures out how)



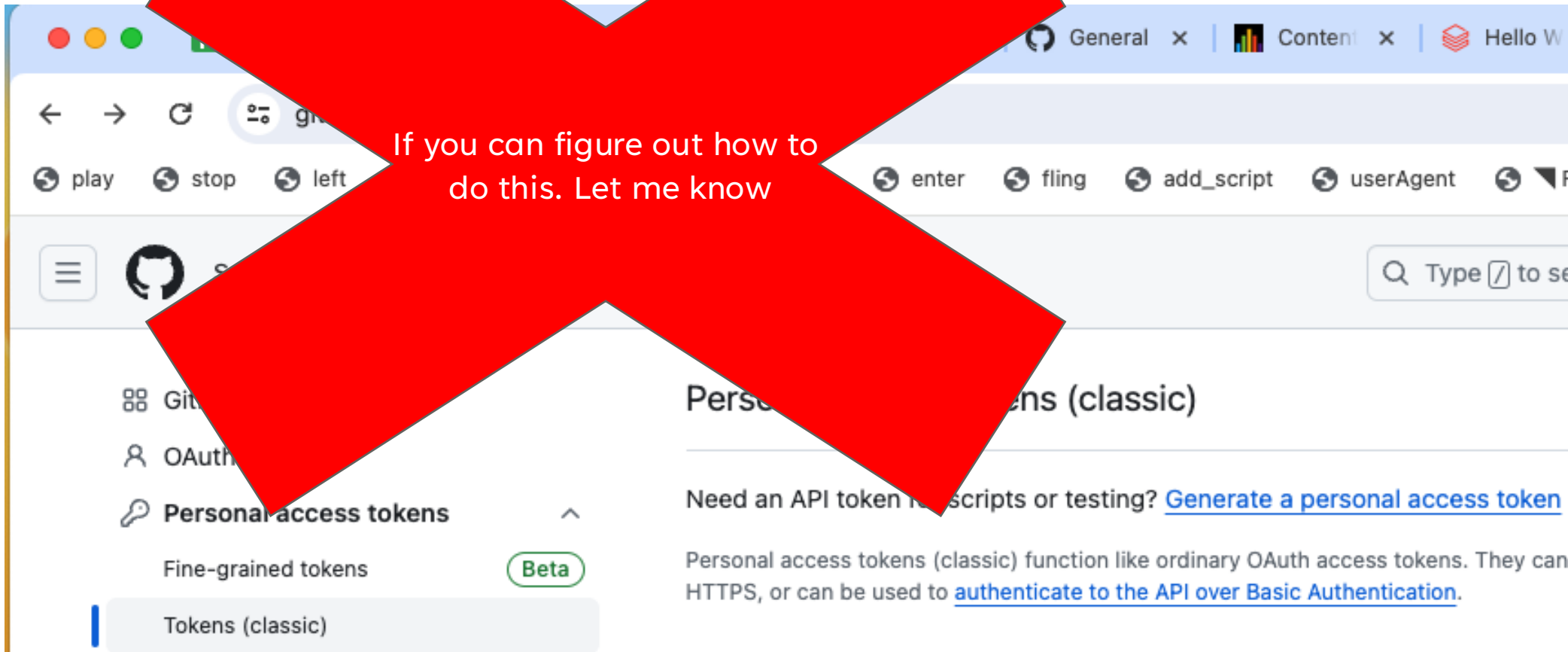
COMMIT NOTEBOOKS TO GITHUB

Please connect your notebooks to github.

Checkpoint your w regularly.

Ask chatgpt how

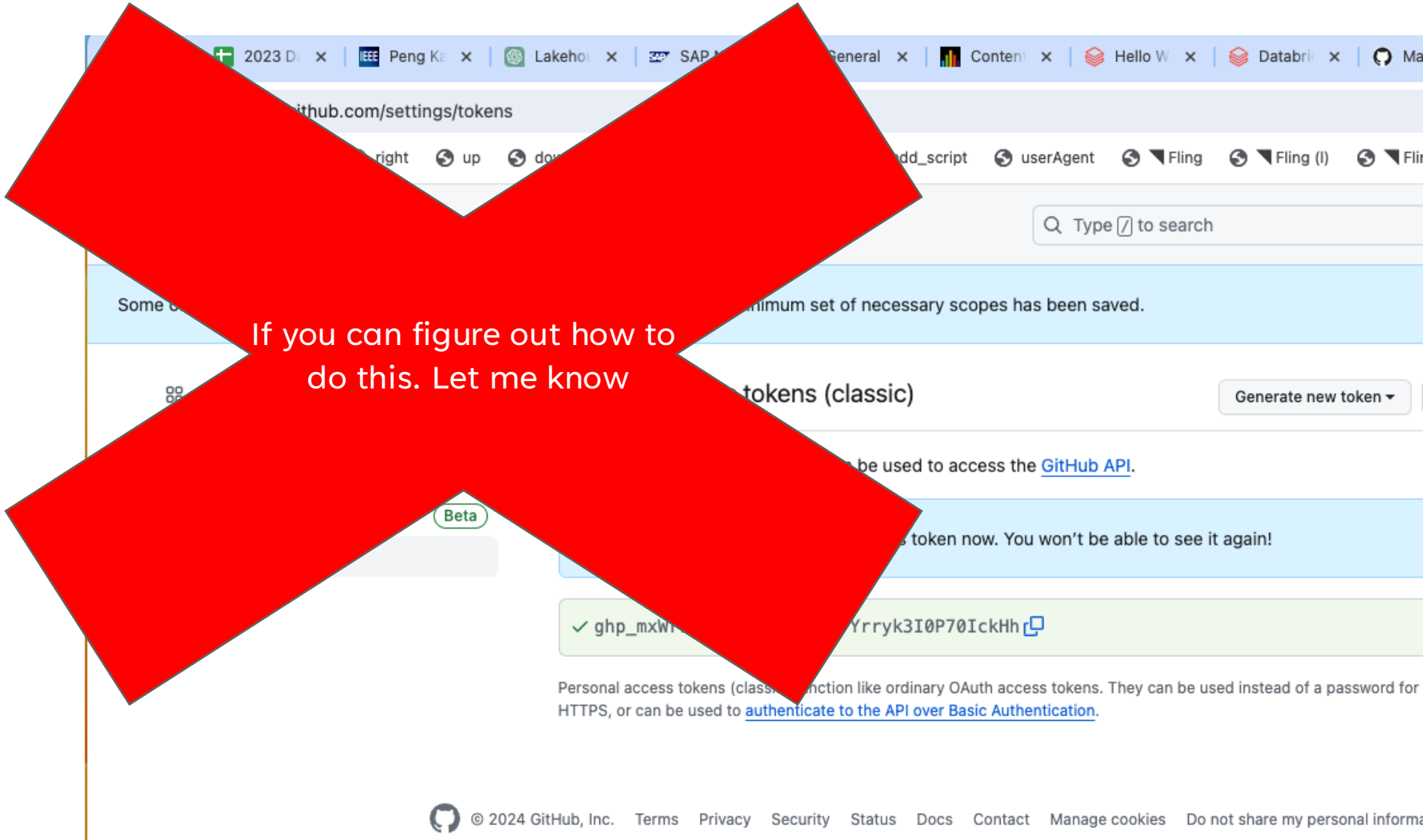
1. On github g Access T



If you can figure out how to do this. Let me know

COMMIT NOTEBOOKS TO GITHUB

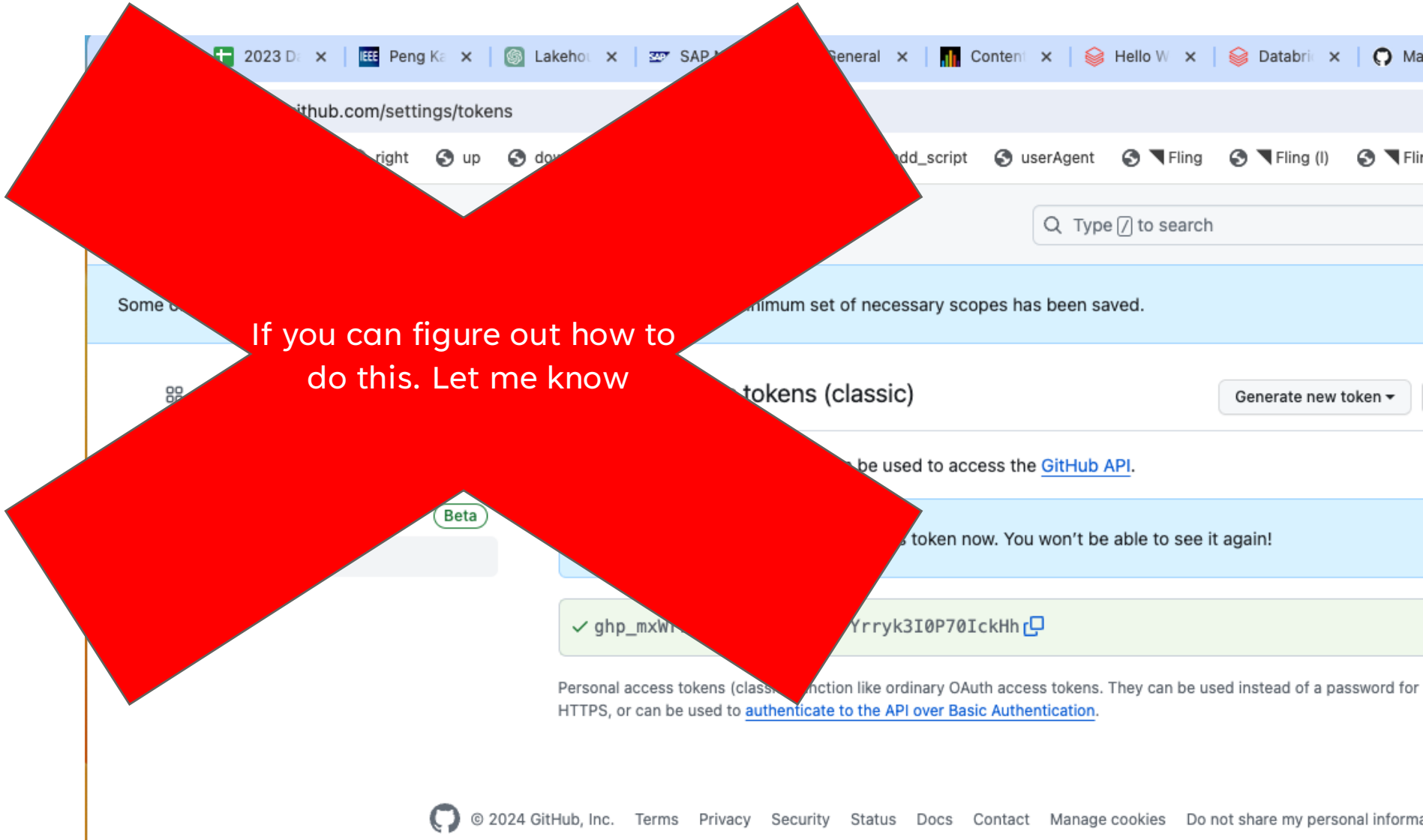
1. On github generate Personal Access Token.



If you can figure out how to do this. Let me know

COMMIT NOTEBOOKS TO GITHUB

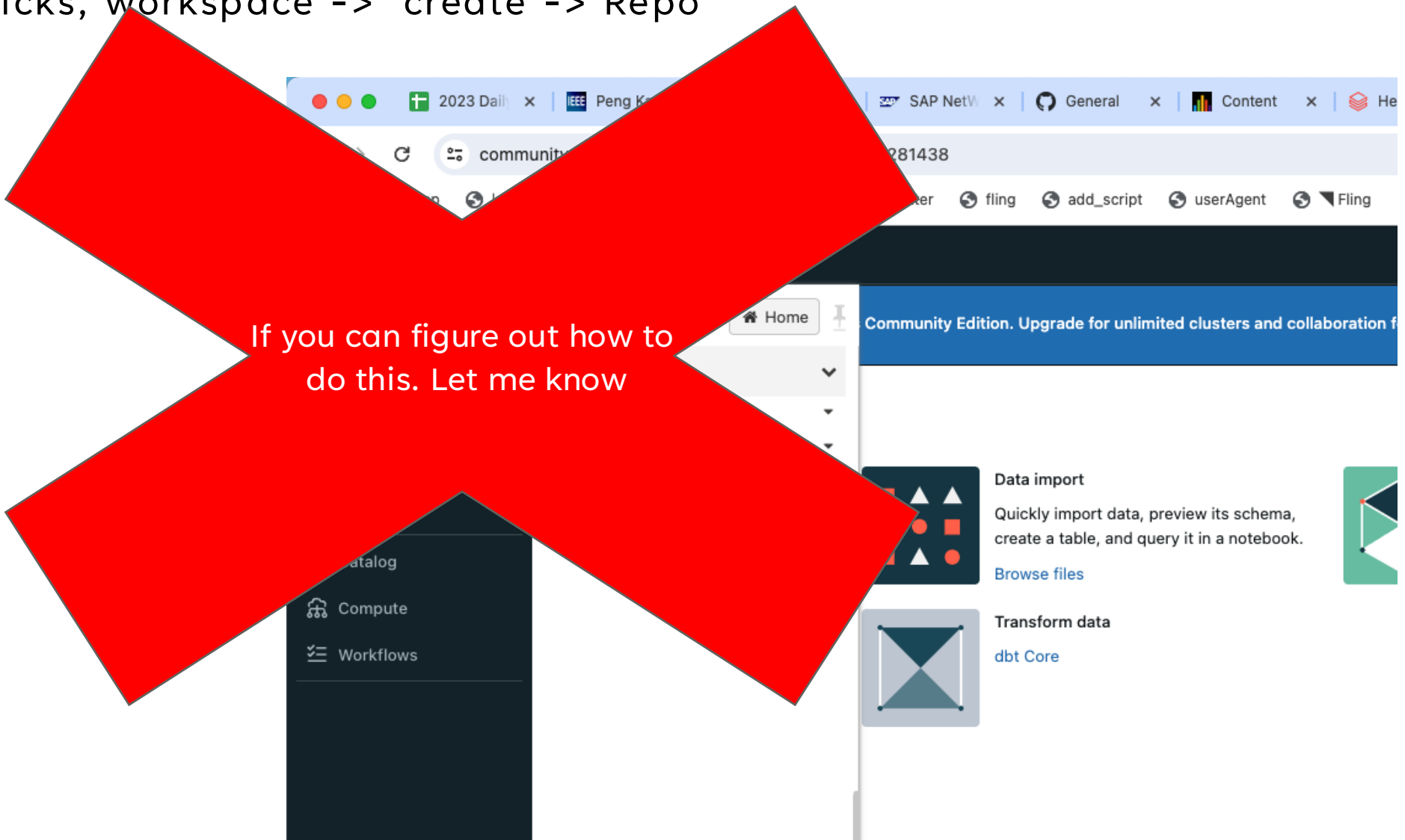
1. On github generate Personal Access Token.



If you can figure out how to do this. Let me know

COMMIT NOTEBOOKS TO GITHUB

2. On databricks, workspace -> create -> Repo



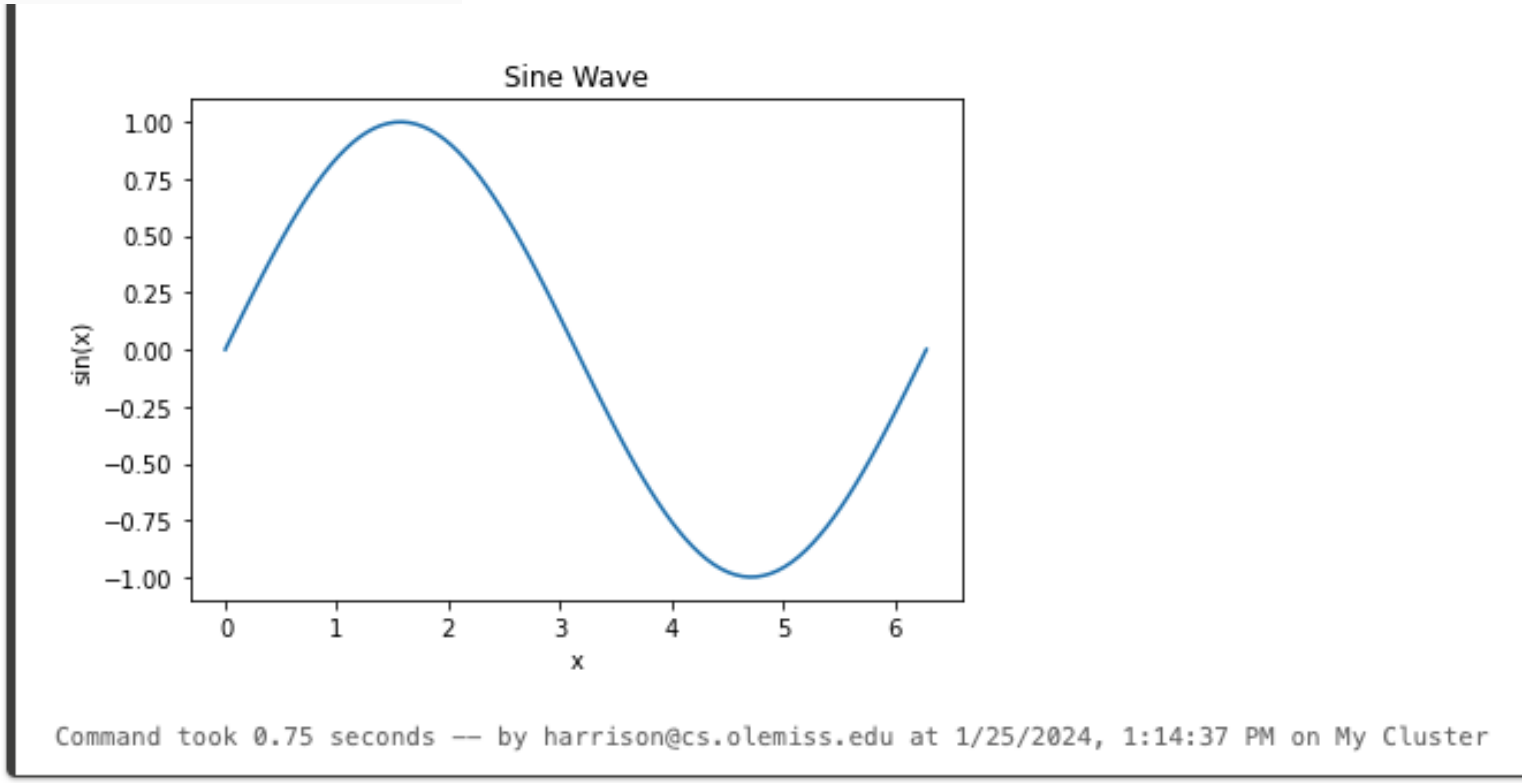
PYTHON

Python is the modern lingua franca of data science.

(show example Hello World using Python in a Databricks Notebook)

VISUALIZATION EXAMPLE USING MATPLOTLIB

```python
import matplotlib.pyplot as plt
import numpy as np

# Generate a range of x values
x = np.linspace(0, 2 * np.pi, 100)

# Calculate the sine of each x value
y = np.sin(x)

# Create the plot
plt.plot(x, y)

# Label the axes
plt.xlabel('x')
plt.ylabel('sin(x)')

# Add a title
plt.title('Sine Wave')

# Show the plot
plt.show()
```



PANDAS DATA FRAMES

A DataFrame is used to represent tabular data like in a spreadsheet

- But programmatic…

```
1   import pandas as pd
2   import matplotlib.pyplot as plt
3
4   # Create a sample DataFrame
5   data = {
6       'Year': [2015, 2016, 2017, 2018, 2019],
7       'Sales': [200, 300, 350, 280, 500]
8   }
9   df = pd.DataFrame(data)
```

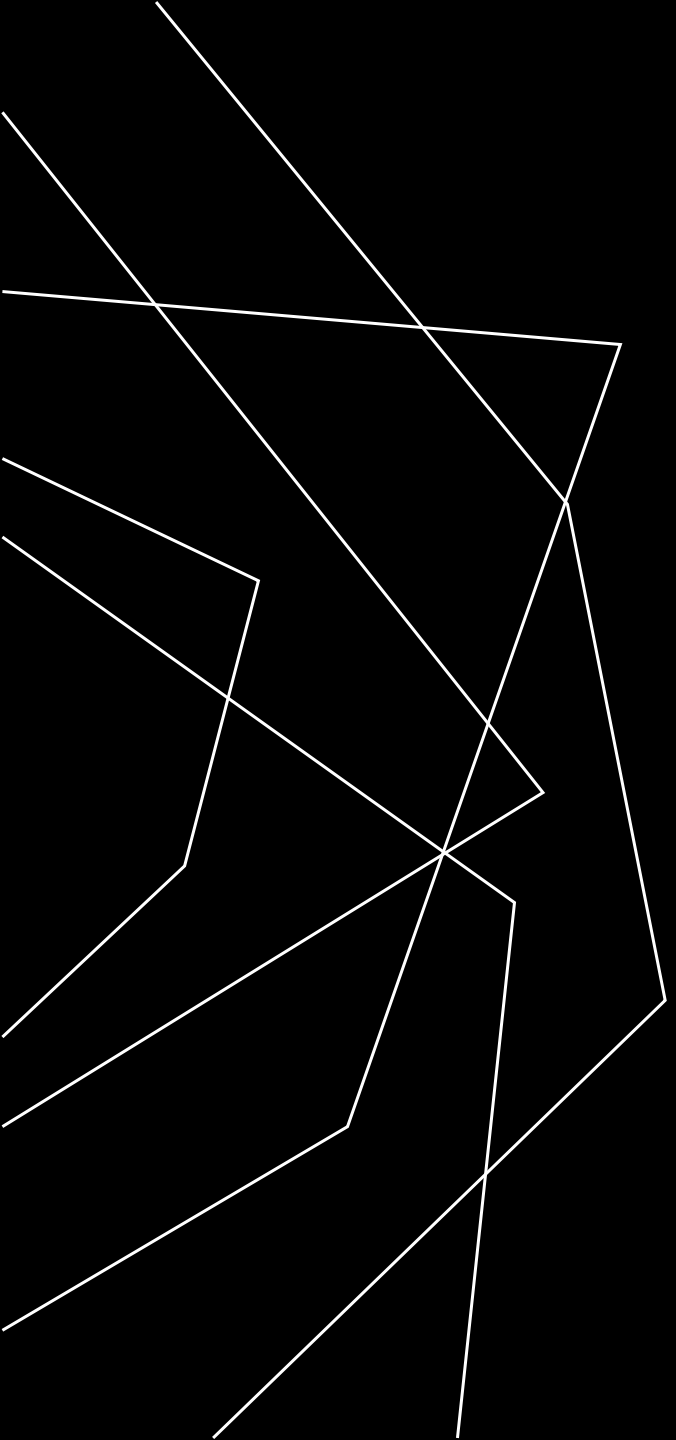
HOMEWORK 1

Get on blackboard.

The homework will be posted there tonight.

1. Setup an account with databricks.
2. Create a notebook
3. See homework on blackboard for problems.
    1. Some statistics review.
4. Familiarize yourself with DataFrames and visualization.

Due Thursday, Jan 30.



THANK YOU

David Harrison

Harrison@cs.olemiss.edu