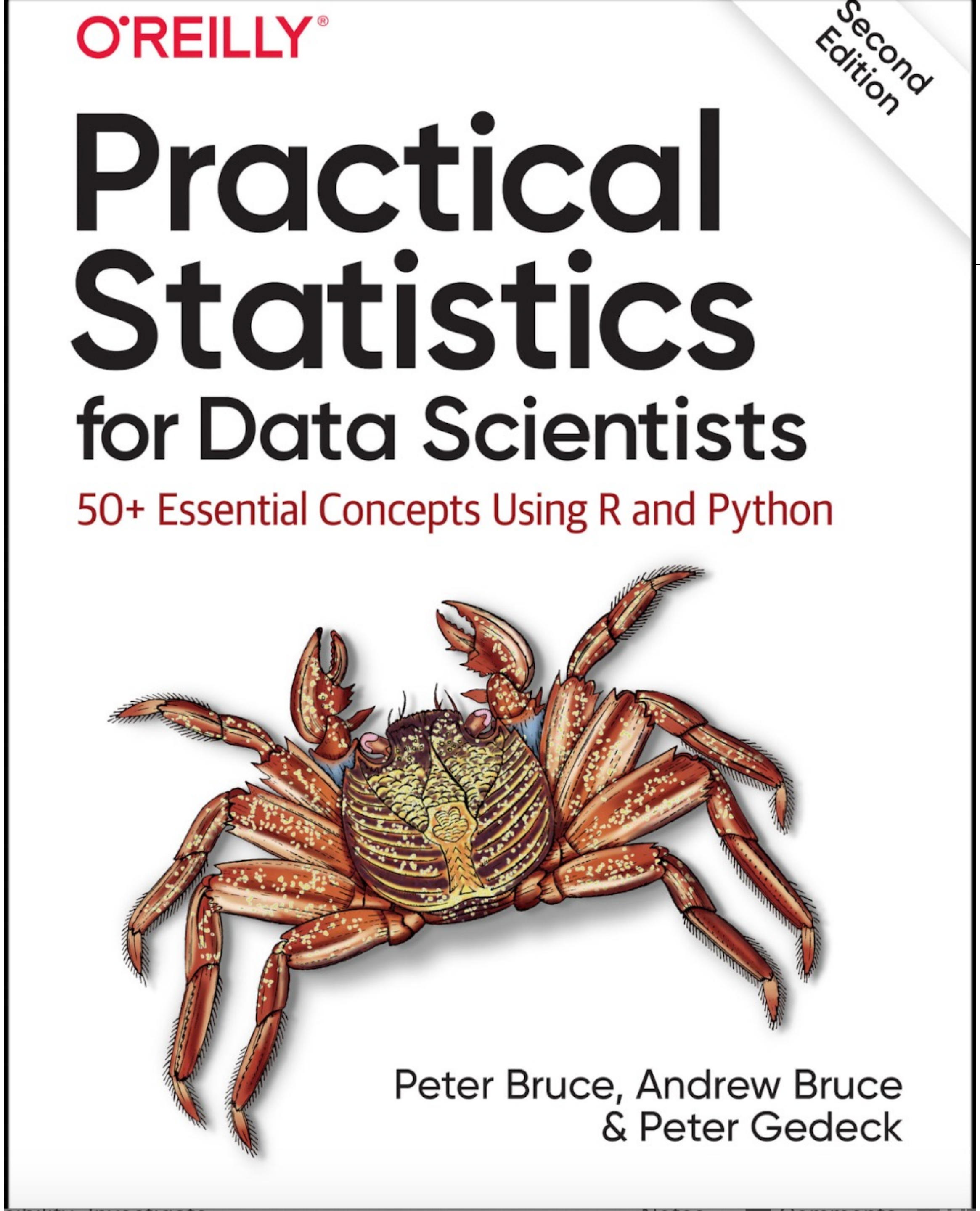
CSCI 443: LECTURE 1 INTRODUCTION

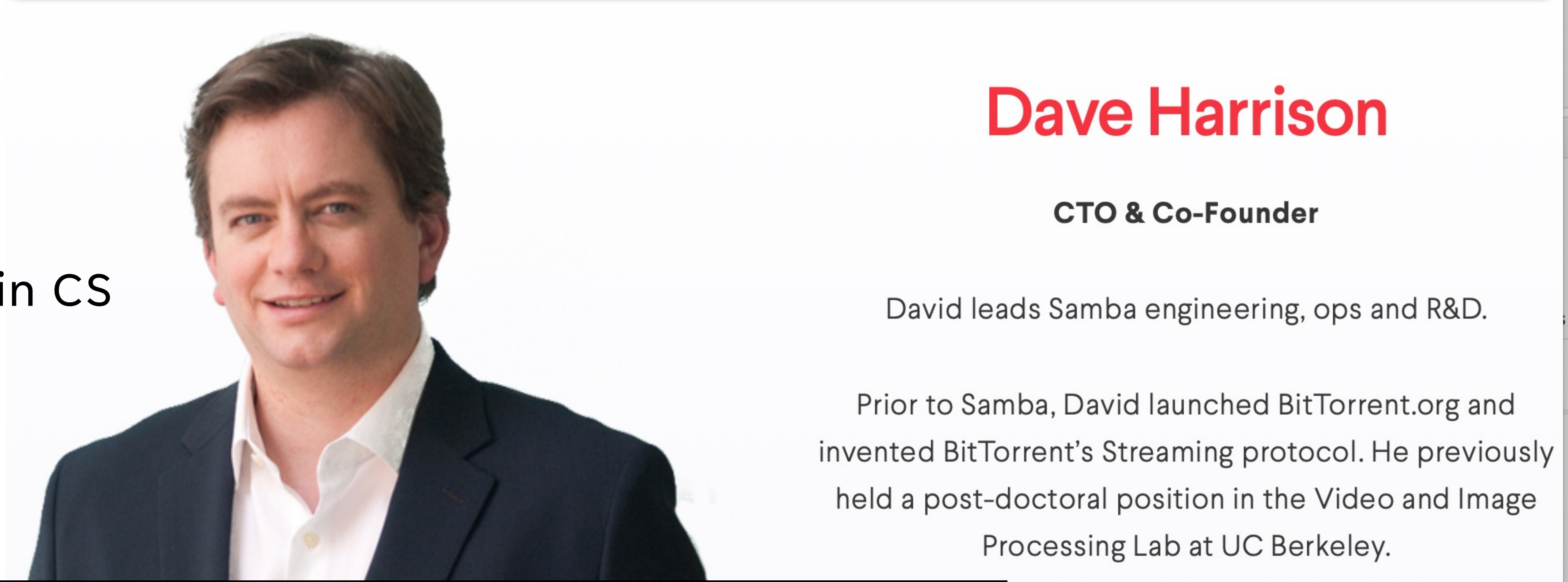
Professor David Harrison

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.


O'REILLY
Second Edition
Practical Statistics for Data Scientists
50+ Essential Concepts Using R and Python
Peter Bruce, Andrew Bruce & Peter Gedeck

- Assistant professor in CS
- Previously

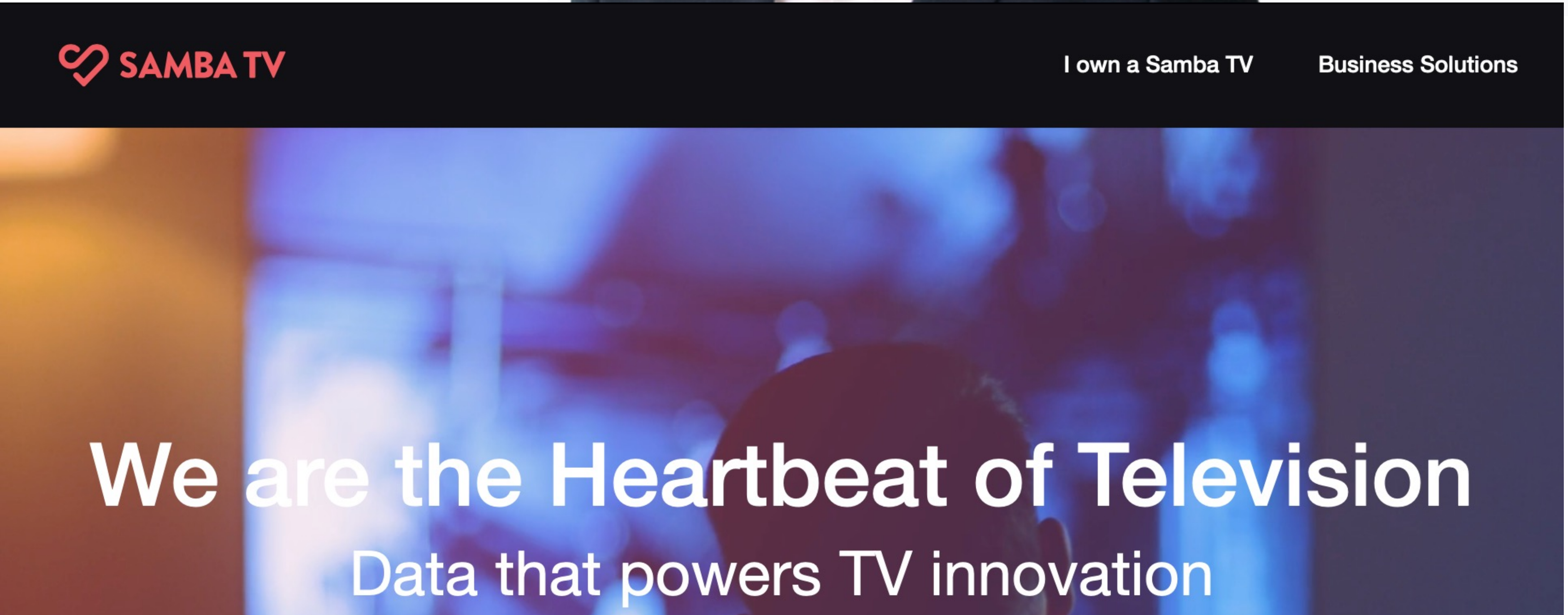


Dave Harrison

CTO & Co-Founder

David leads Samba engineering, ops and R&D.

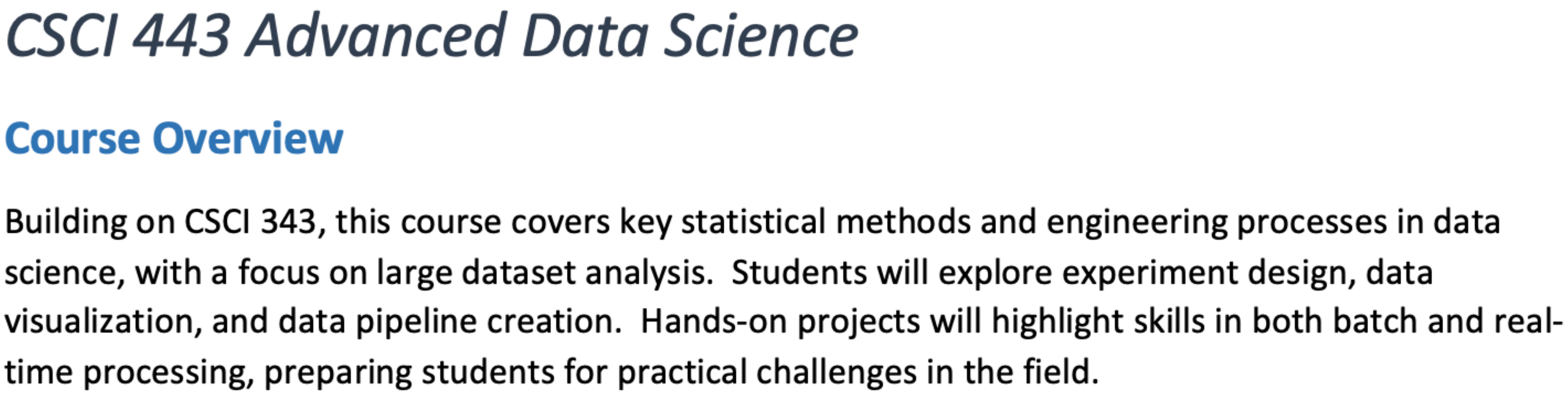
Prior to Samba, David launched BitTorrent.org and invented BitTorrent's Streaming protocol. He previously held a post-doctoral position in the Video and Image Processing Lab at UC Berkeley.



We are the Heartbeat of Television

Data that powers TV innovation

SYLLABUS



CSCI 443 Advanced Data Science

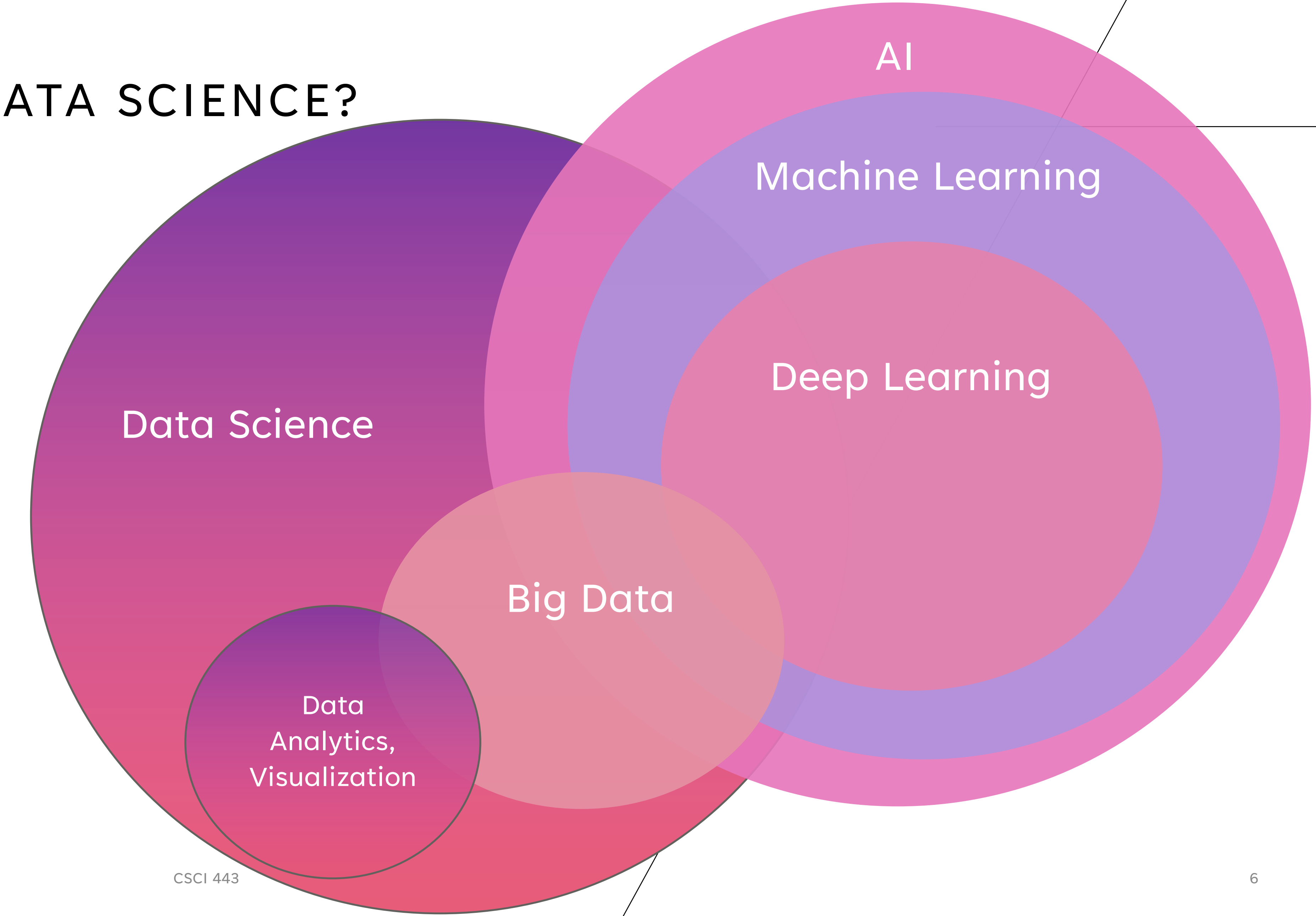
Course Overview

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?



AI

Machine Learning

Deep Learning

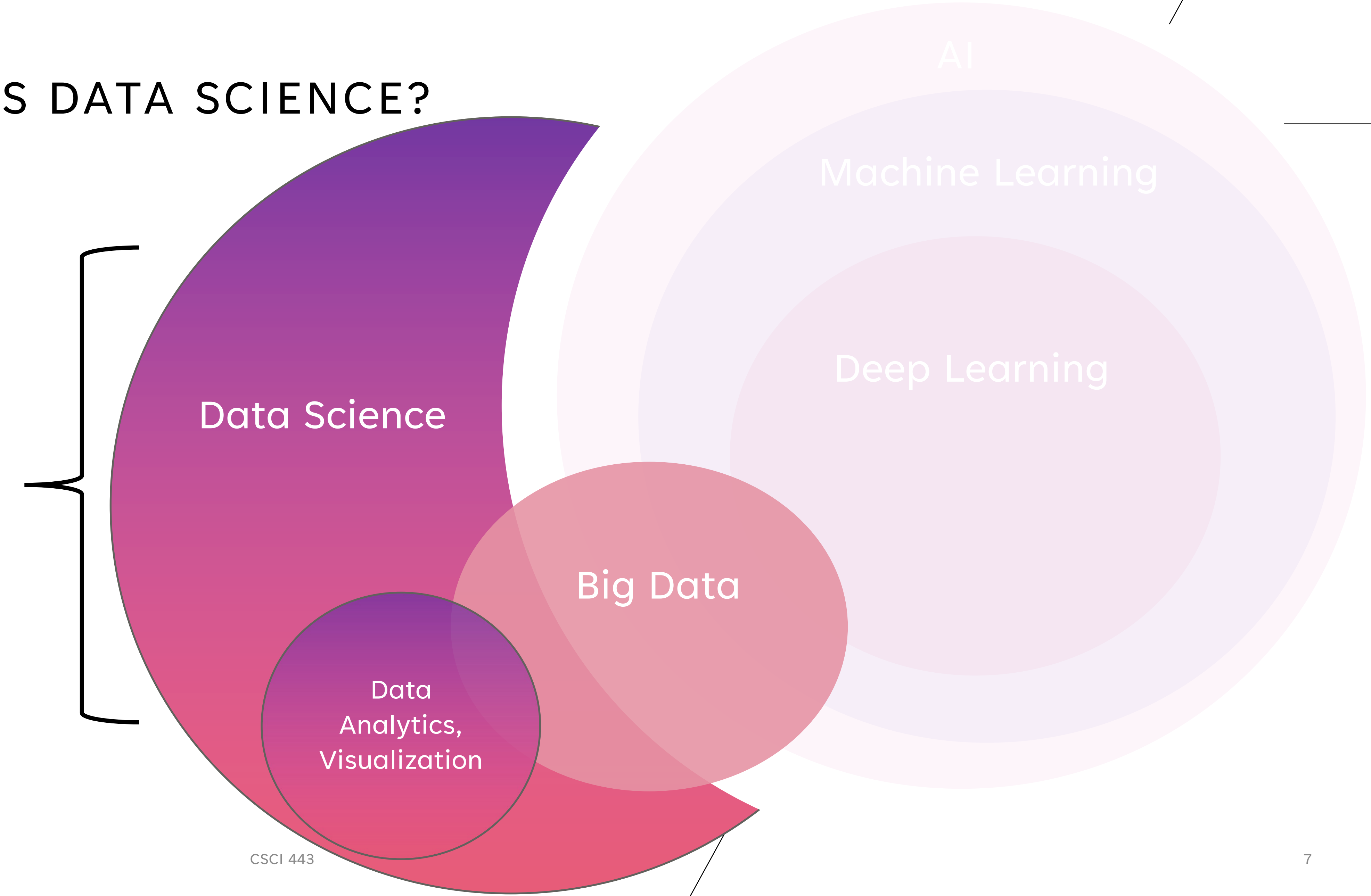
Data Science

Big Data

Data Analytics, Visualization

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.

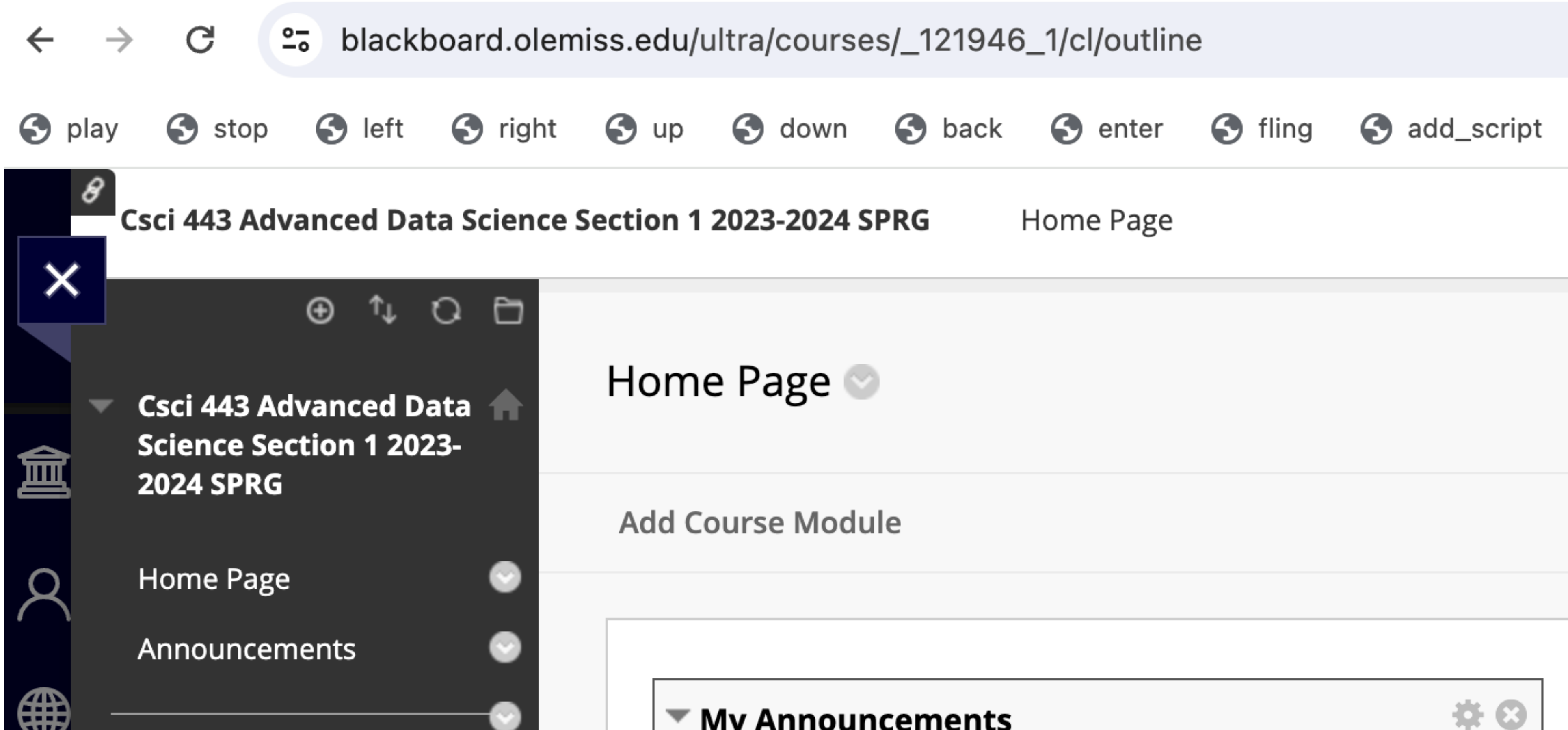
- Data analysts are skilled in data manipulation and visualization.
  - Often use processes put in place by a data scientist.
  - Often use tools implemented by data engineers.
  - Often support executives and sales
  - May be externally facing.
  - Often not well versed in engineering

TOOLS

-

BLACKBOARD

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



blackboard.olemiss.edu/ultra/courses/_121946_1/cl/outline

GITHUB

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

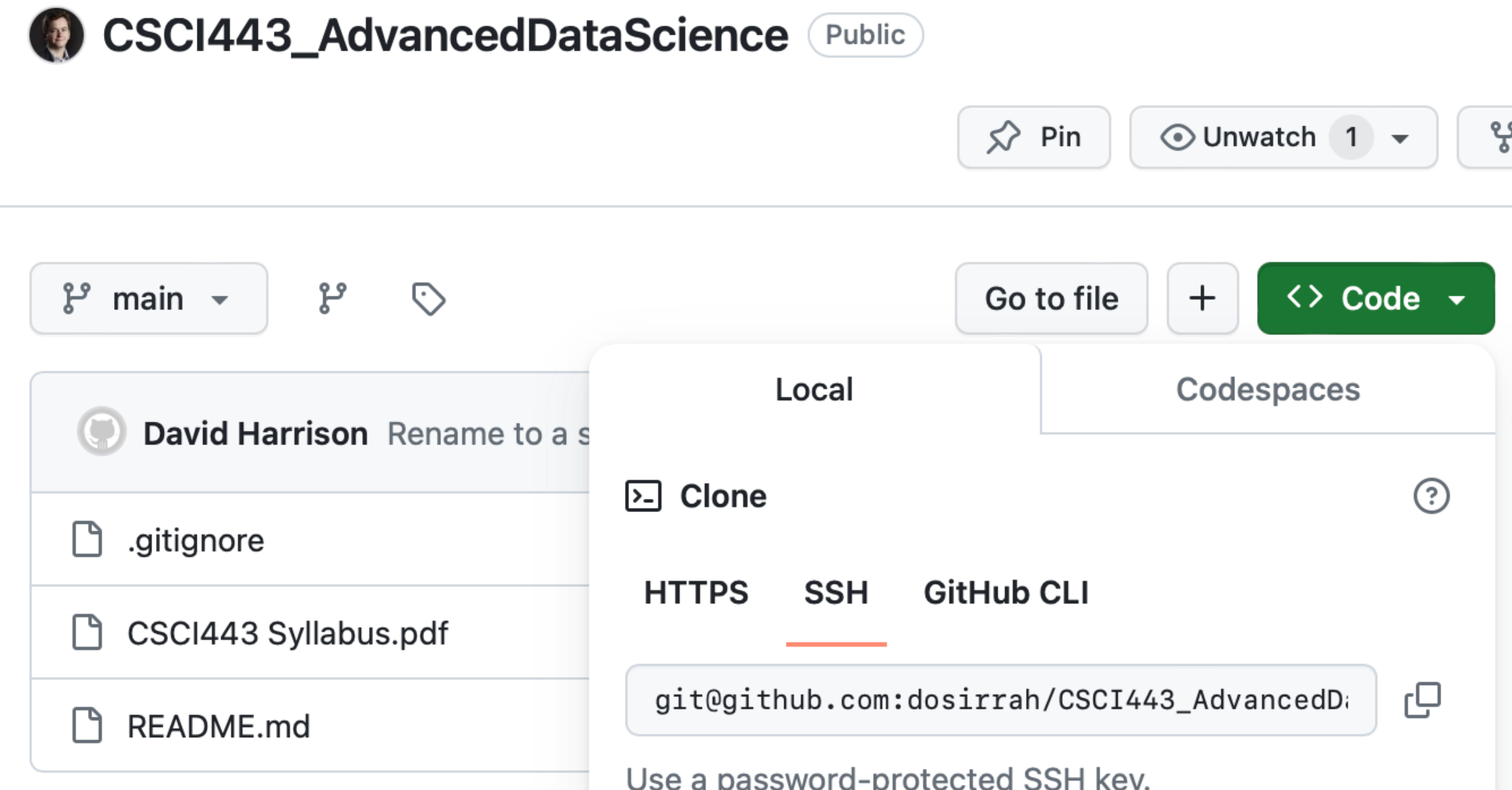
The project is at

https://github.com/dosirrah/CSCI443_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.



CSCI443_AdvancedDataScience Public

David Harrison Rename to a s

.gitignore

CSCI443 Syllabus.pdf

README.md

```
git@github.com:dosirrah/CSCI443_AdvancedD
```

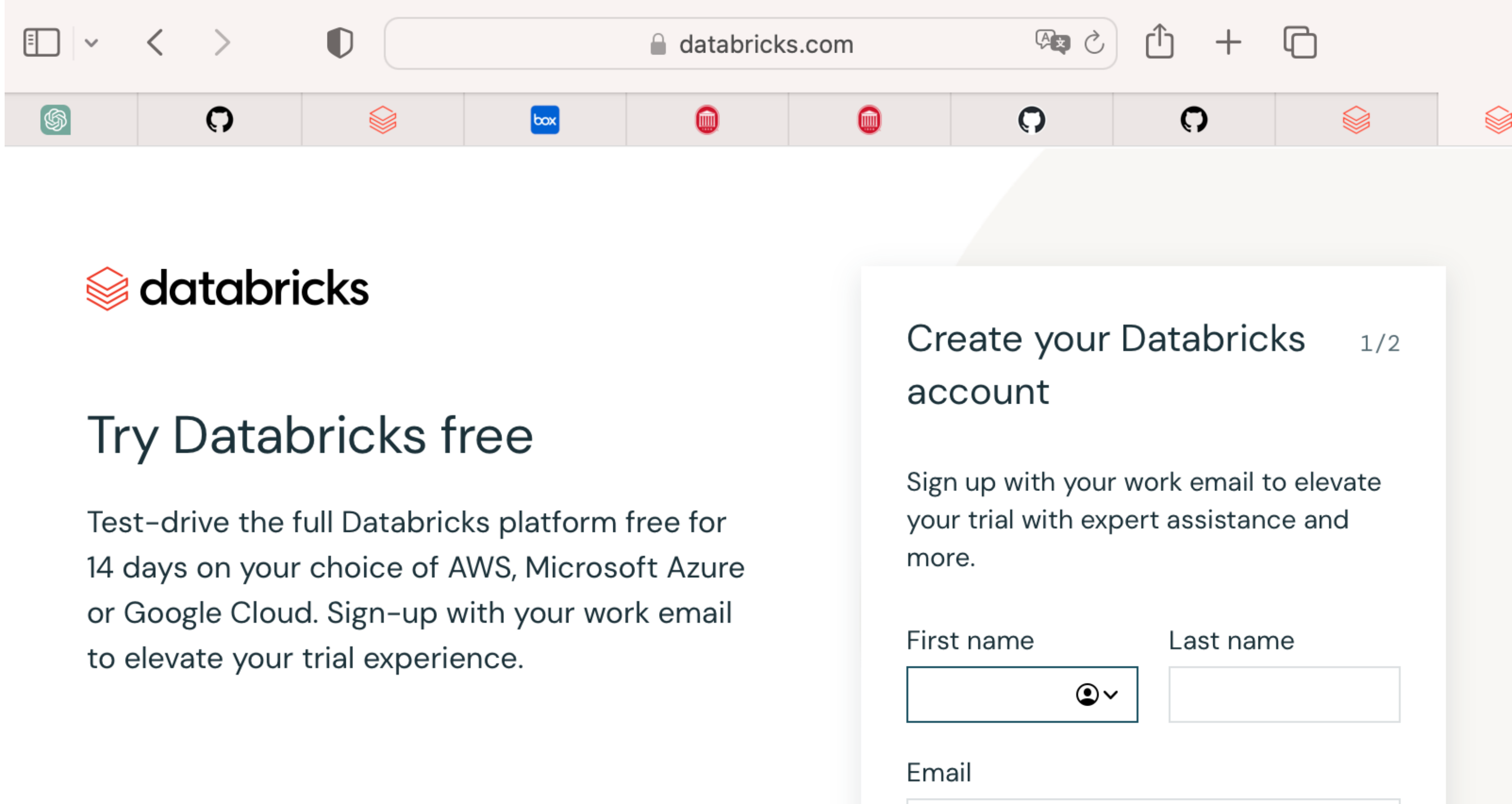
Use a password-protected SSH key.

NO GITLAB…

Last semester 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

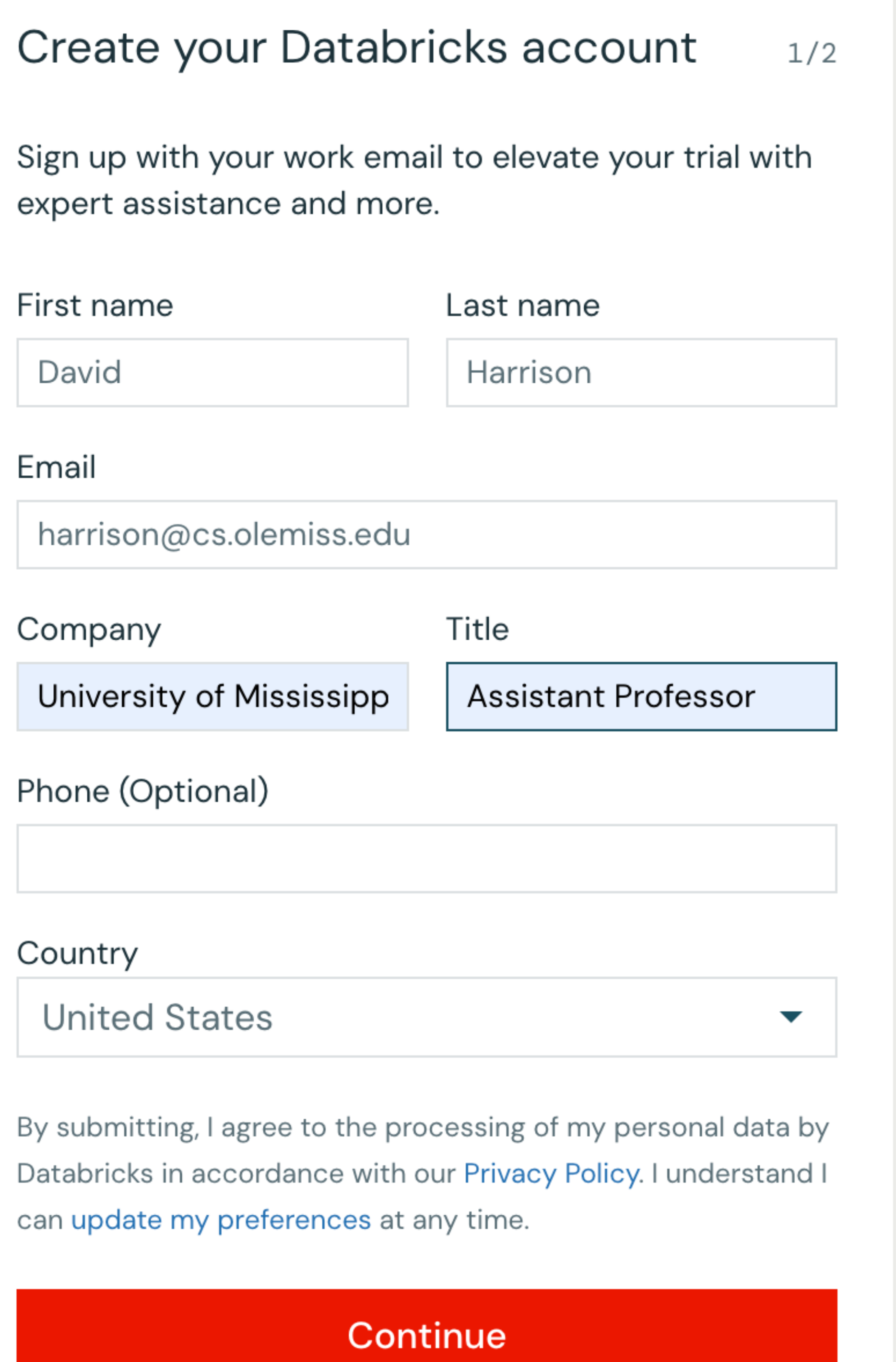
Last name

Email

DATABRICKS

Community edition is free.

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



Create your Databricks account 1/2

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

First name: David

Last name: Harrison

Email: harrison@cs.olemiss.edu

Company: University of Mississipp

Title: Assistant Professor

Phone (Optional):

Country: United States

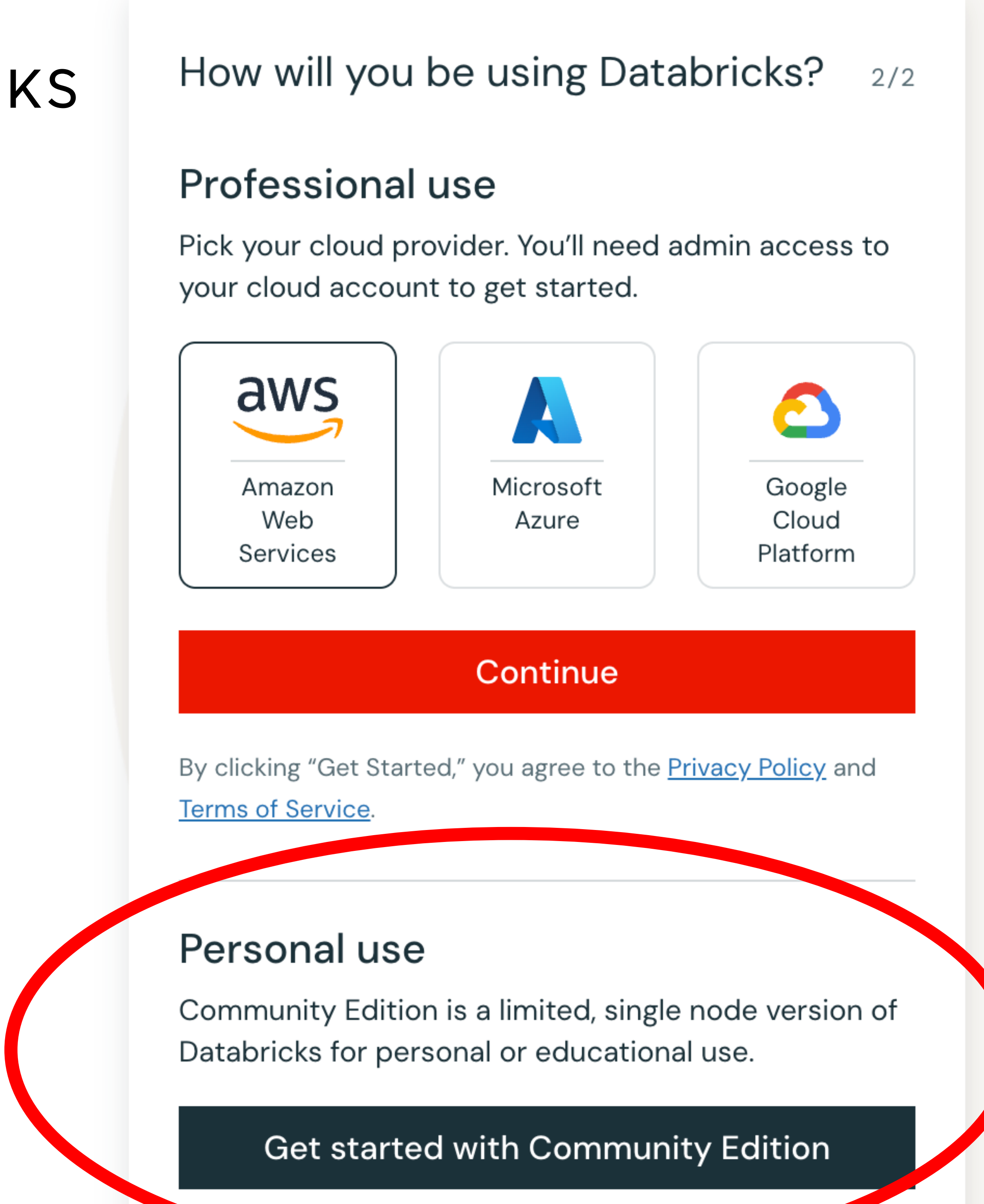
By submitting, I agree to the processing of my personal data by Databricks in accordance with our Privacy Policy. I understand I can update my preferences at any time.

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.



How will you be using Databricks? 2/2

Professional use

Pick your cloud provider. You'll need admin access to your cloud account to get started.

By clicking "Get Started," you agree to the Privacy Policy and Terms of Service.

Personal use

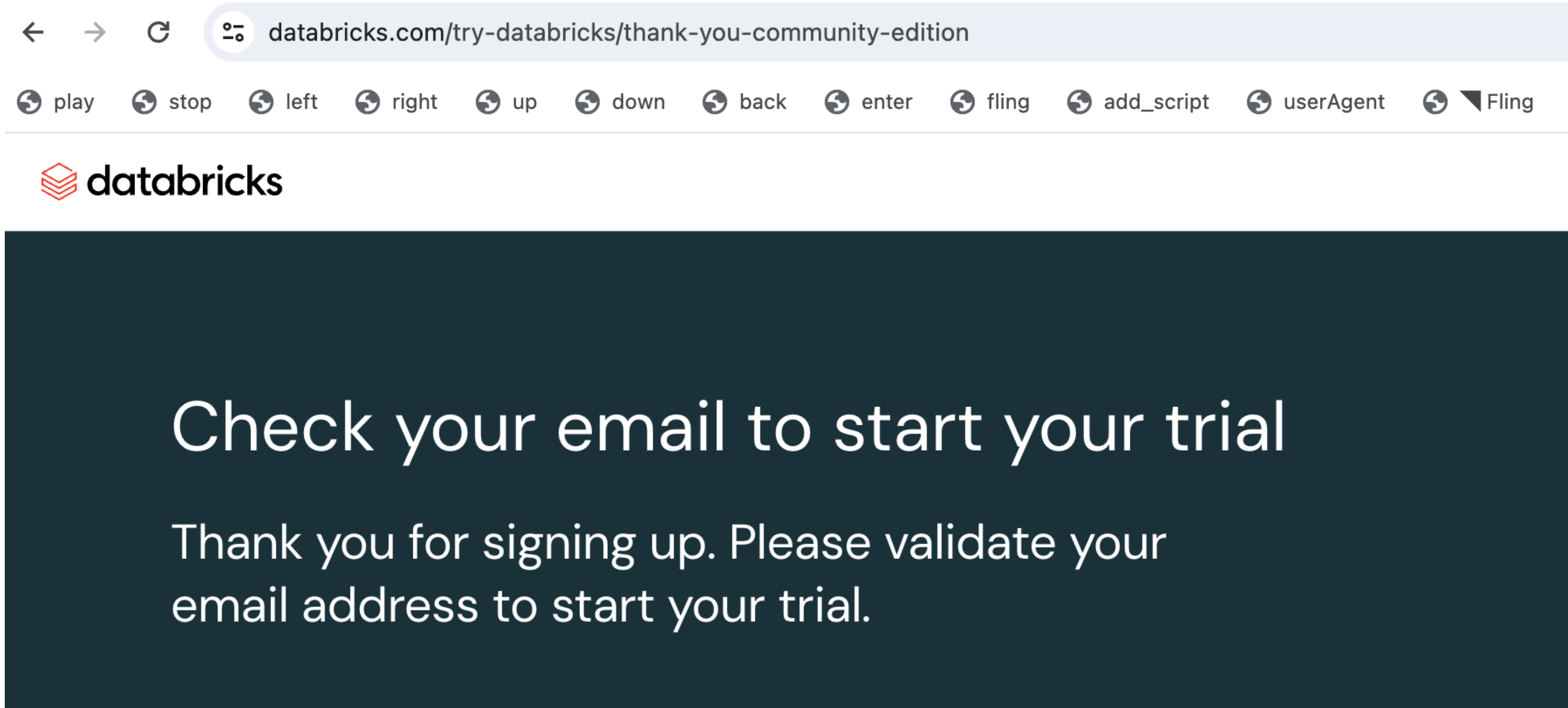
Community Edition is a limited, single node version of Databricks for personal or educational use.

DATABRICKS

Community edition is free.

Don’t worry about “your trial.”

This is misleading.



Check your email to start your trial

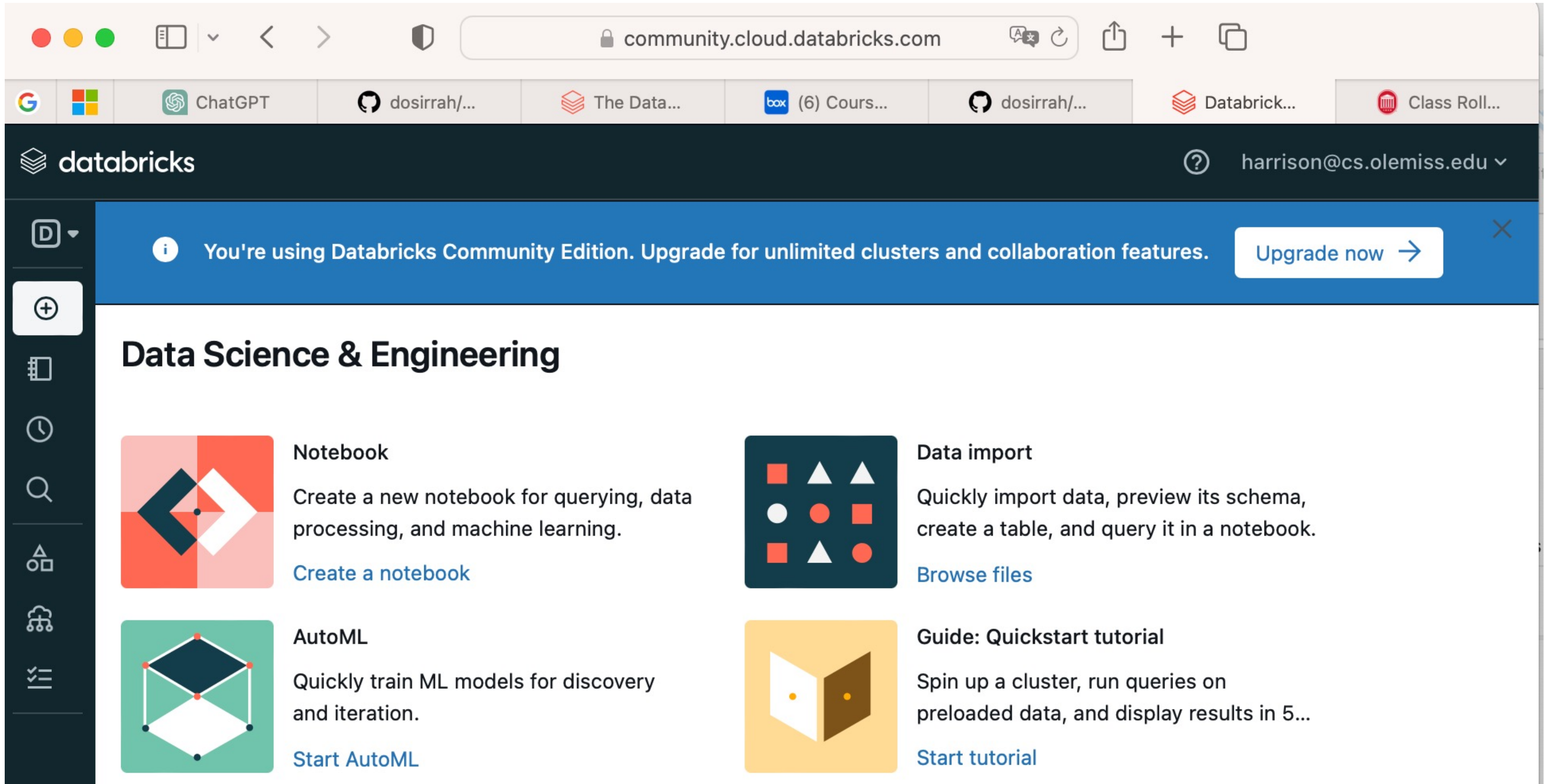
Thank you for signing up. Please validate your email address to start your trial.

DATABRICKS

https://community.cloud.databricks.com

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

Ignore "Upgrade now"



You're using Databricks Community Edition. Upgrade for unlimited clusters and collaboration features.

Data Science & Engineering

Notebook

Create a new notebook for querying, data processing, and machine learning.

Data import

Quickly import data, preview its schema, create a table, and query it in a notebook.

AutoML

Quickly train ML models for discovery and iteration.

Guide: Quickstart tutorial

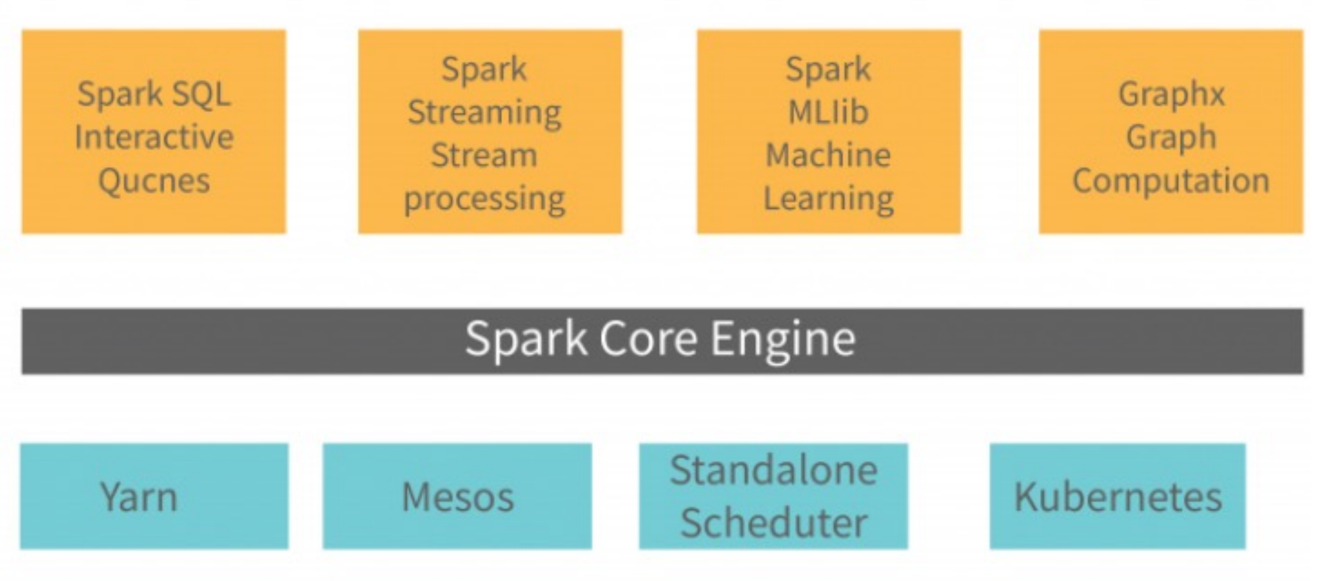
Spin up a cluster, run queries on preloaded data, and display results in 5...

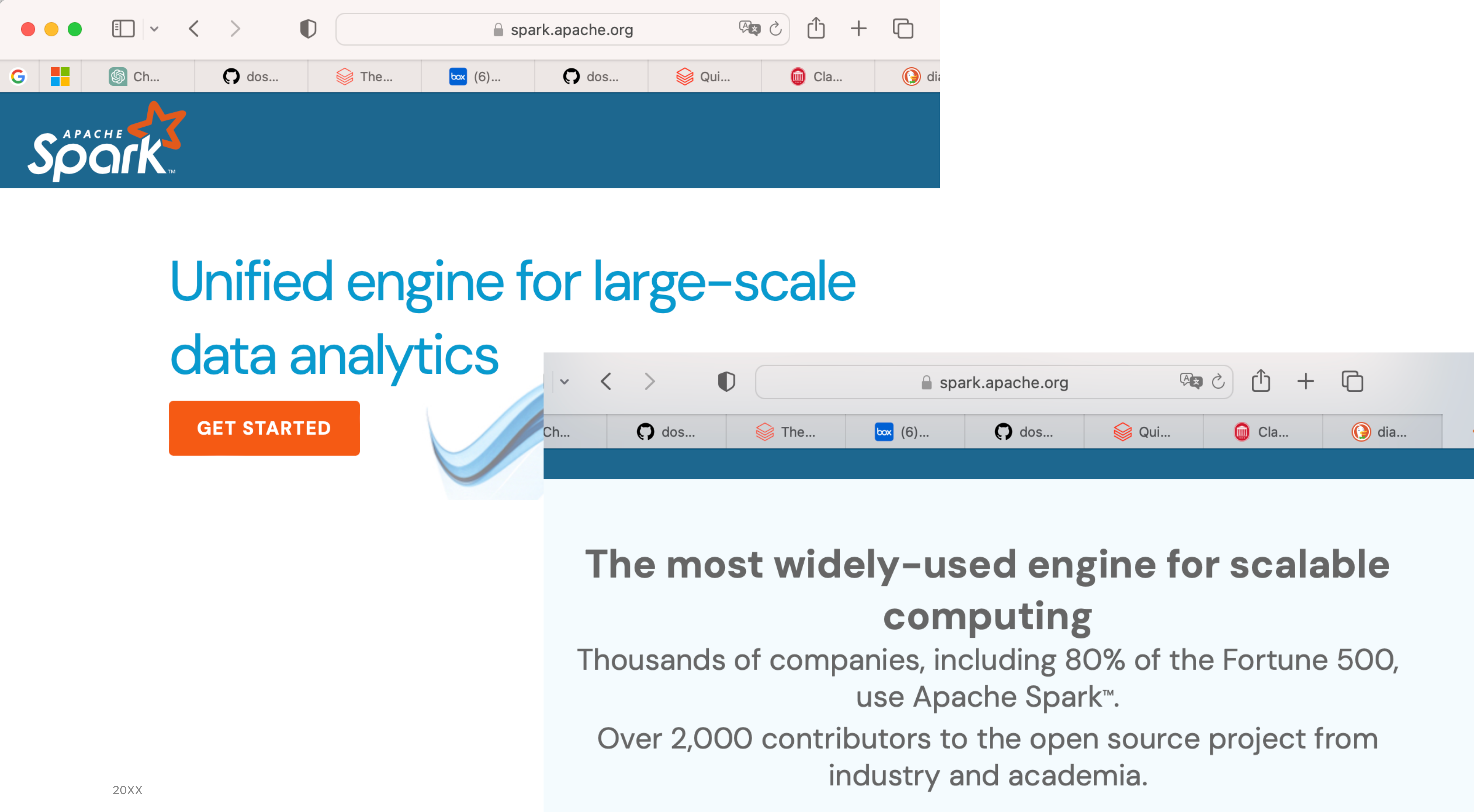
WHY DATABRICKS?

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.


Spark SQL Interactive Qucnes
Spark Streaming Stream processing
Spark MLlib Machine Learning
Graphx Graph Computation
Spark Core Engine
Yarn
Mesos
Standalone Scheduter
Kubernetes



Unified engine for large-scale data analytics

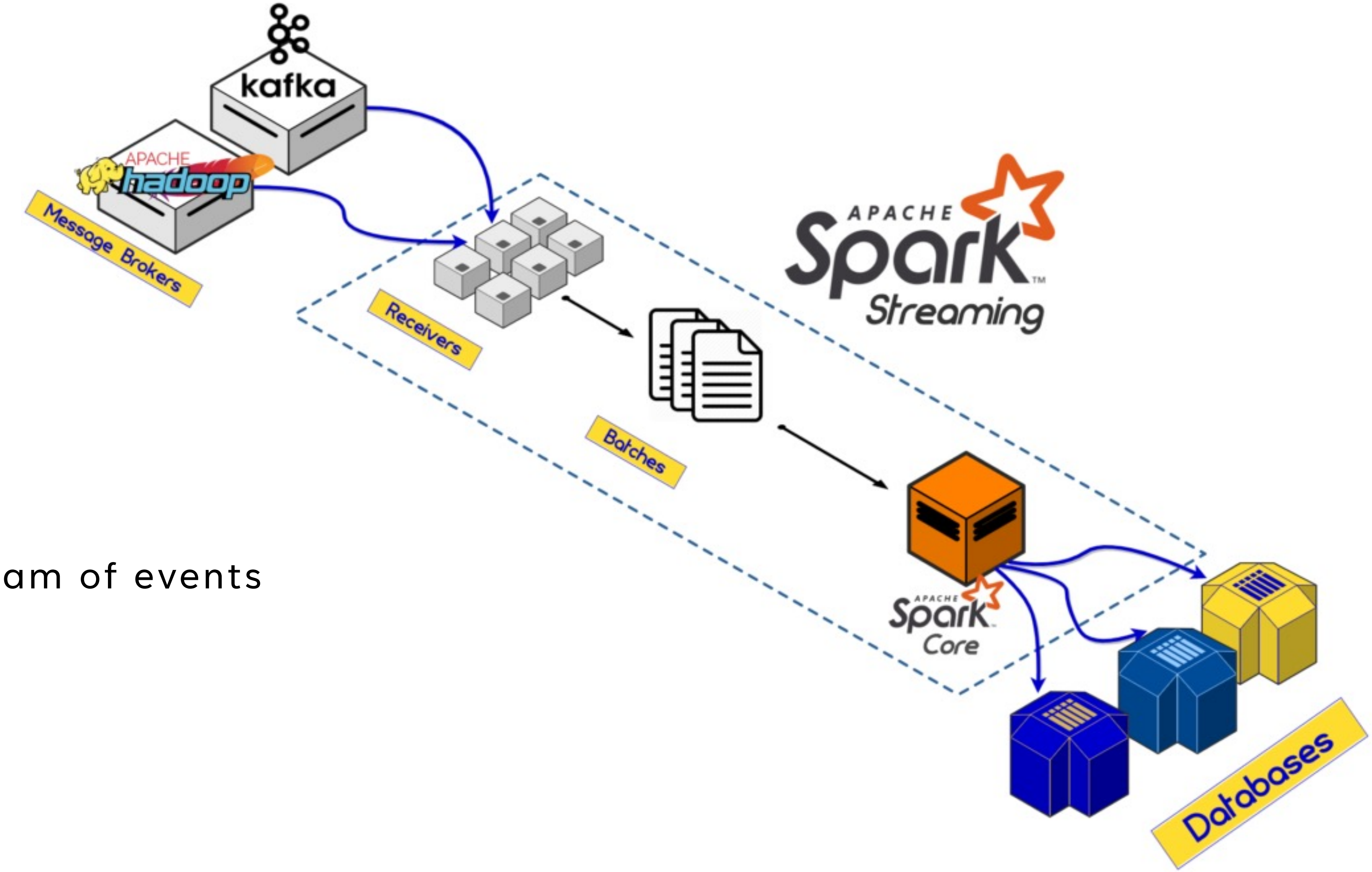
GET STARTED

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

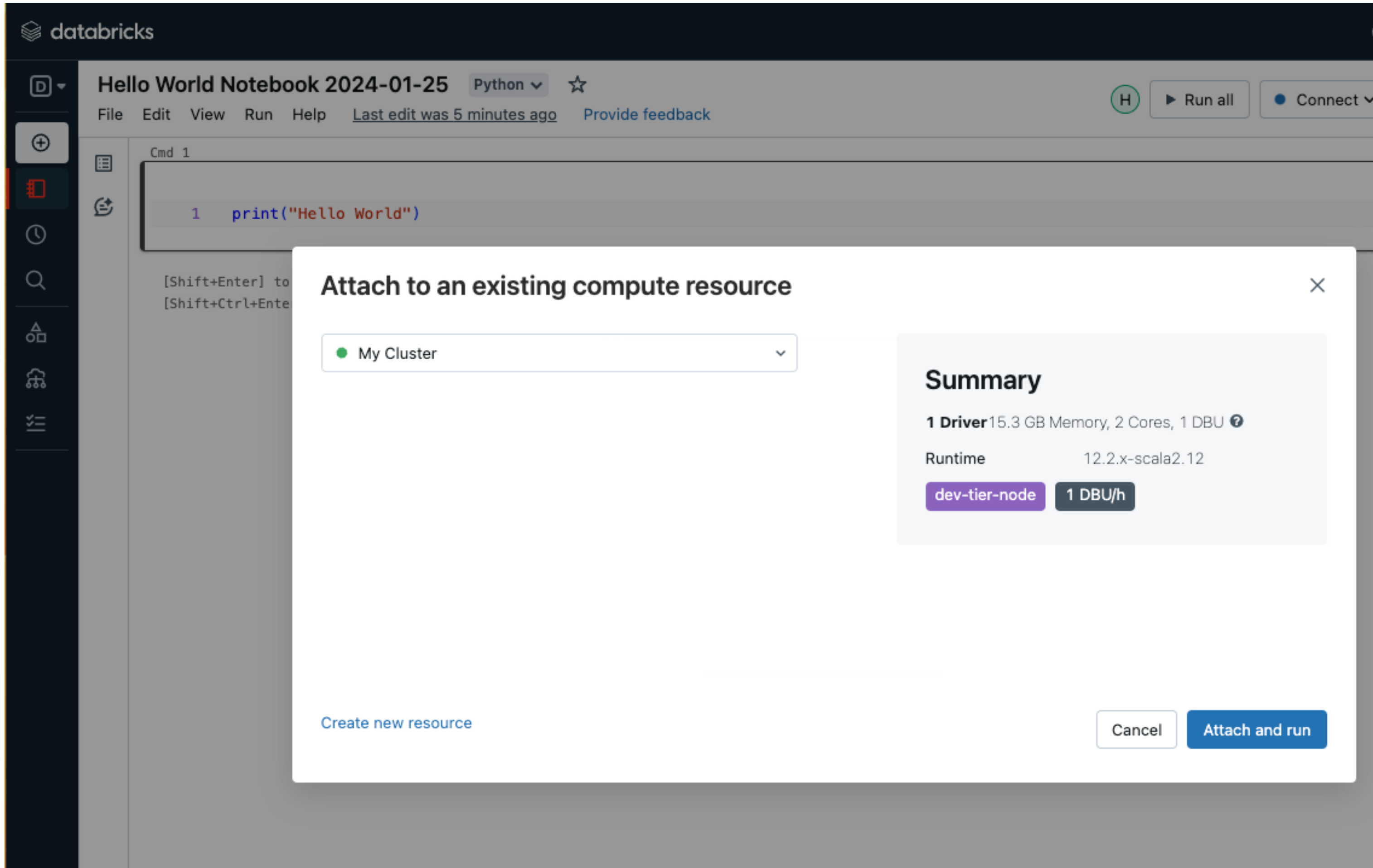
DATABRICKS NOTEBOOKS

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.



Hello World Notebook 2024-01-25

```
print("Hello World")
```

Attach to an existing compute resource

My Cluster

Summary

1 Driver 15.3 GB Memory, 2 Cores, 1 DBU

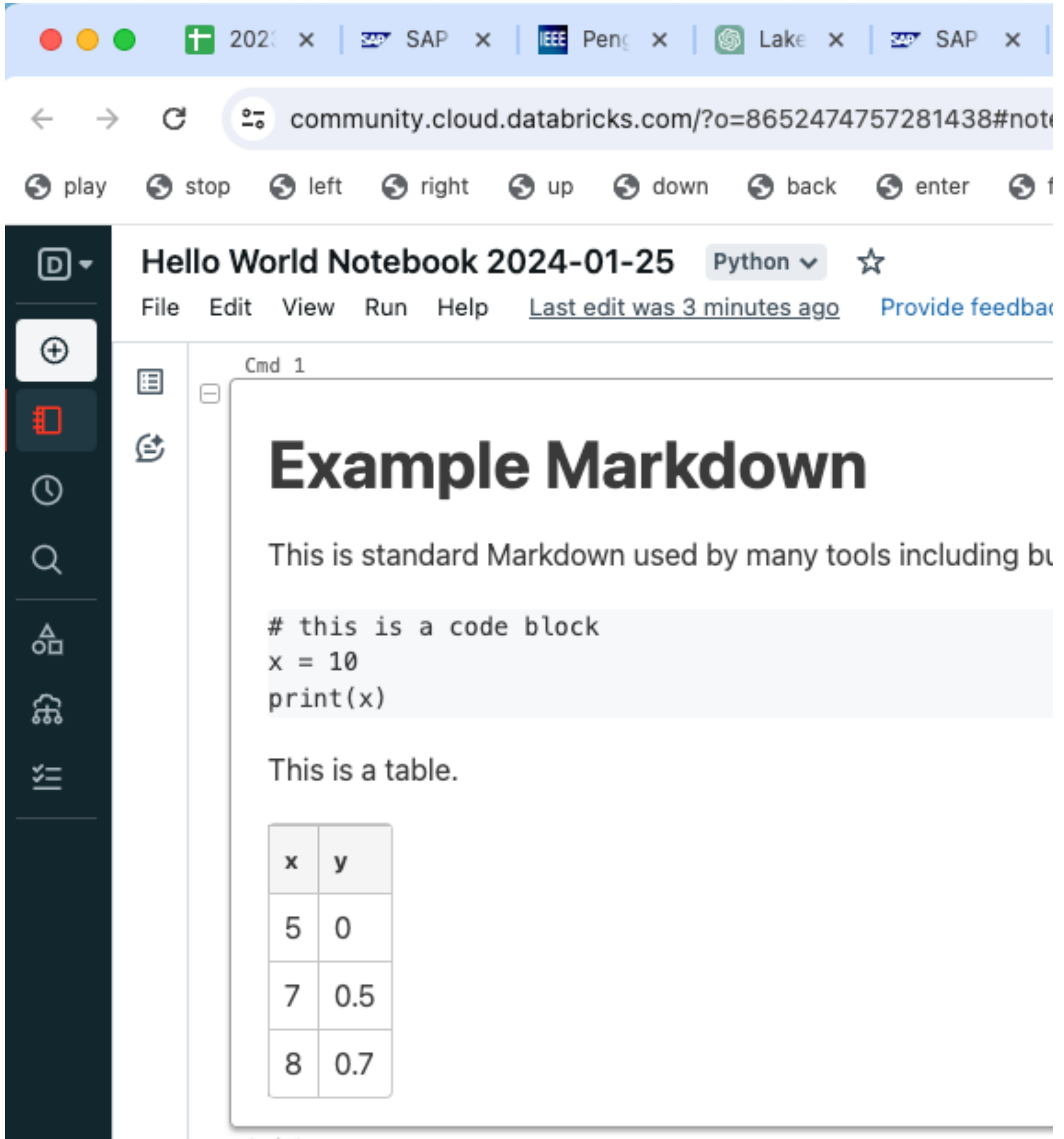
Runtime 12.2.x-scala2.12

dev-tier-node 1 DBU/h

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)



Hello World Notebook 2024-01-25

Example Markdown

This is standard Markdown used by many tools including bu

```
# this is a code block
x = 10
print(x)
```

This is a table.

x	y
5	0
7	0.5
8	0.7

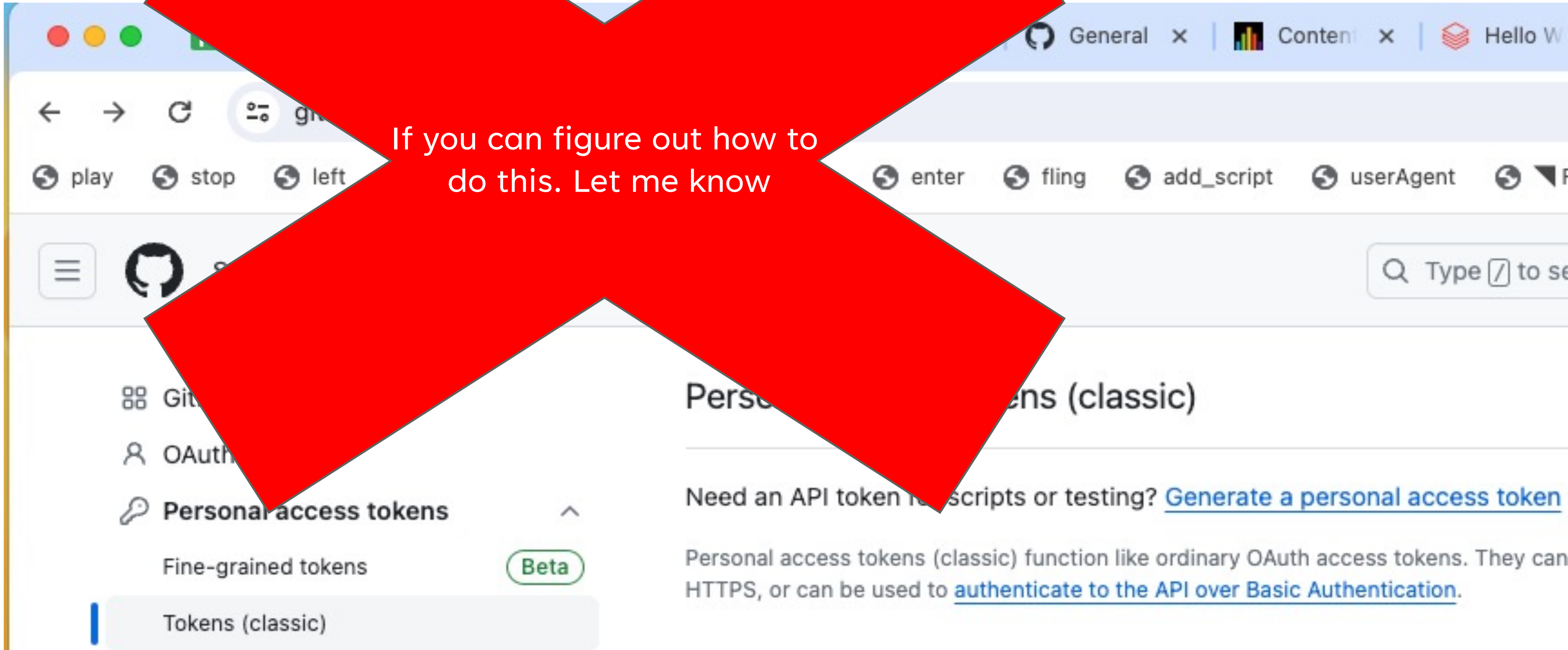
COMMIT NOTEBOOKS TO GITHUB

Please connect your notebooks to github.

Checkpoint your w regularly.

Ask chatgpt how

1. On github g Access T



Pers ens (classic)

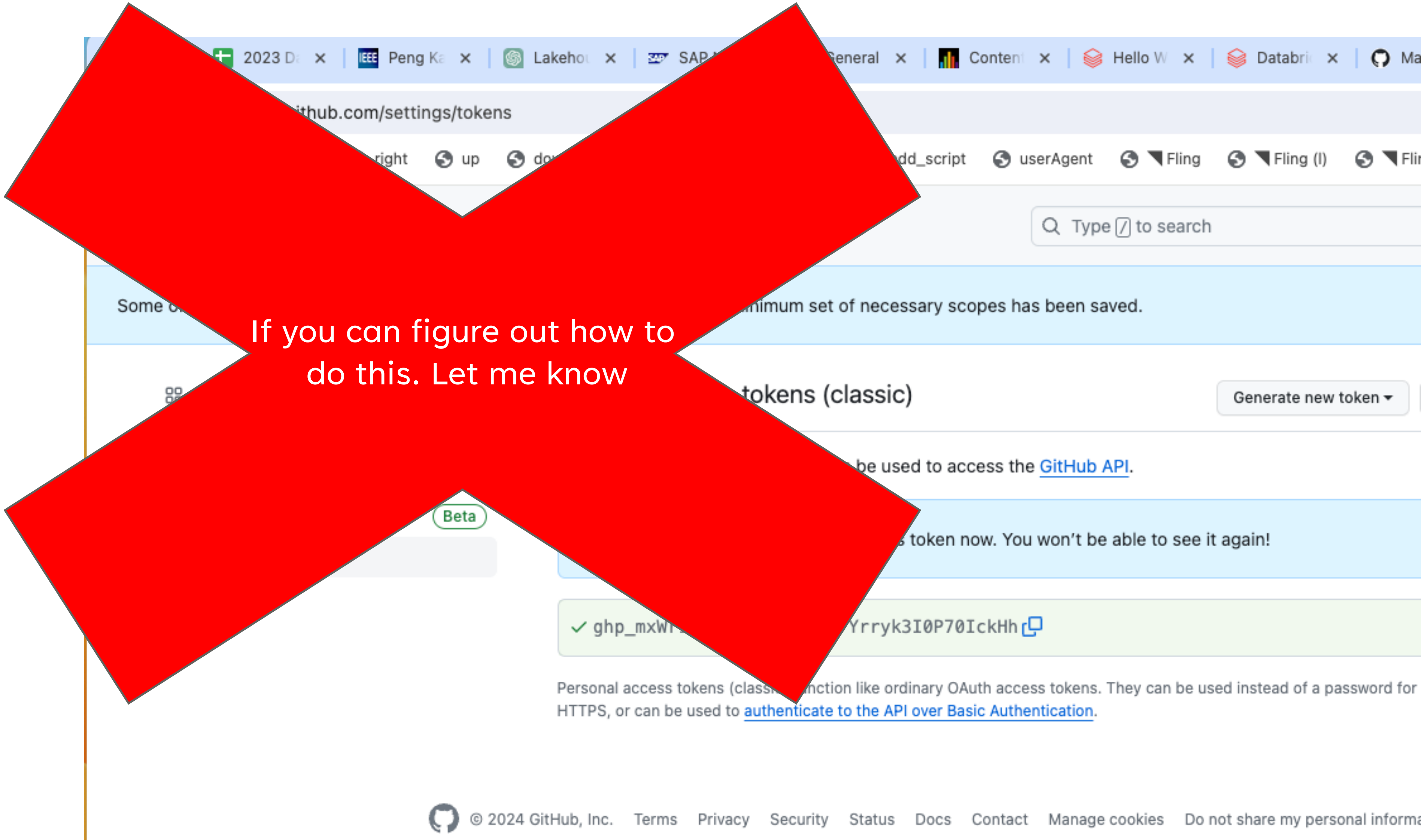
Need an API token f scripts or testing? Generate a personal access token

Personal access tokens (classic) function like ordinary OAuth access tokens. They can
HTTPS, or can be used to authenticate to the API over Basic Authentication.

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

Some o… imum set of necessary scopes has been saved.

… be used to access the GitHub API.

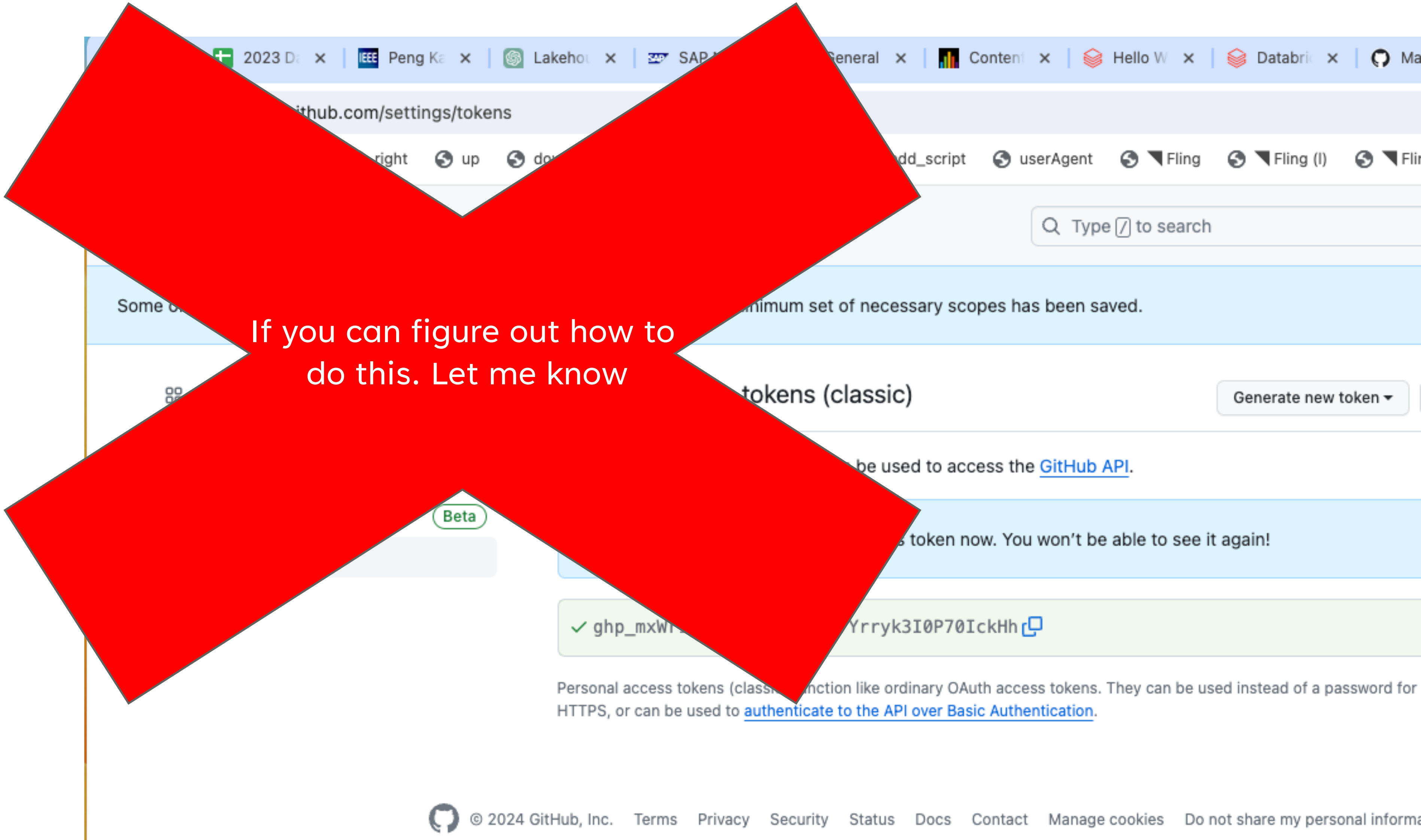
… token now. You won't be able to see it again!

✓ ghp_mxW… Yrryk3I0P70IckHh

Personal access tokens (class… nction like ordinary OAuth access tokens. They can be used instead of a password for HTTPS, or can be used to authenticate to the API over Basic Authentication.

COMMIT NOTEBOOKS TO GITHUB

1. On github generate Personal Access Token.



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

Some o

nimum set of necessary scopes has been saved.

tokens (classic)

be used to access the GitHub API.

token now. You won't be able to see it again!

✓ ghp_mxW

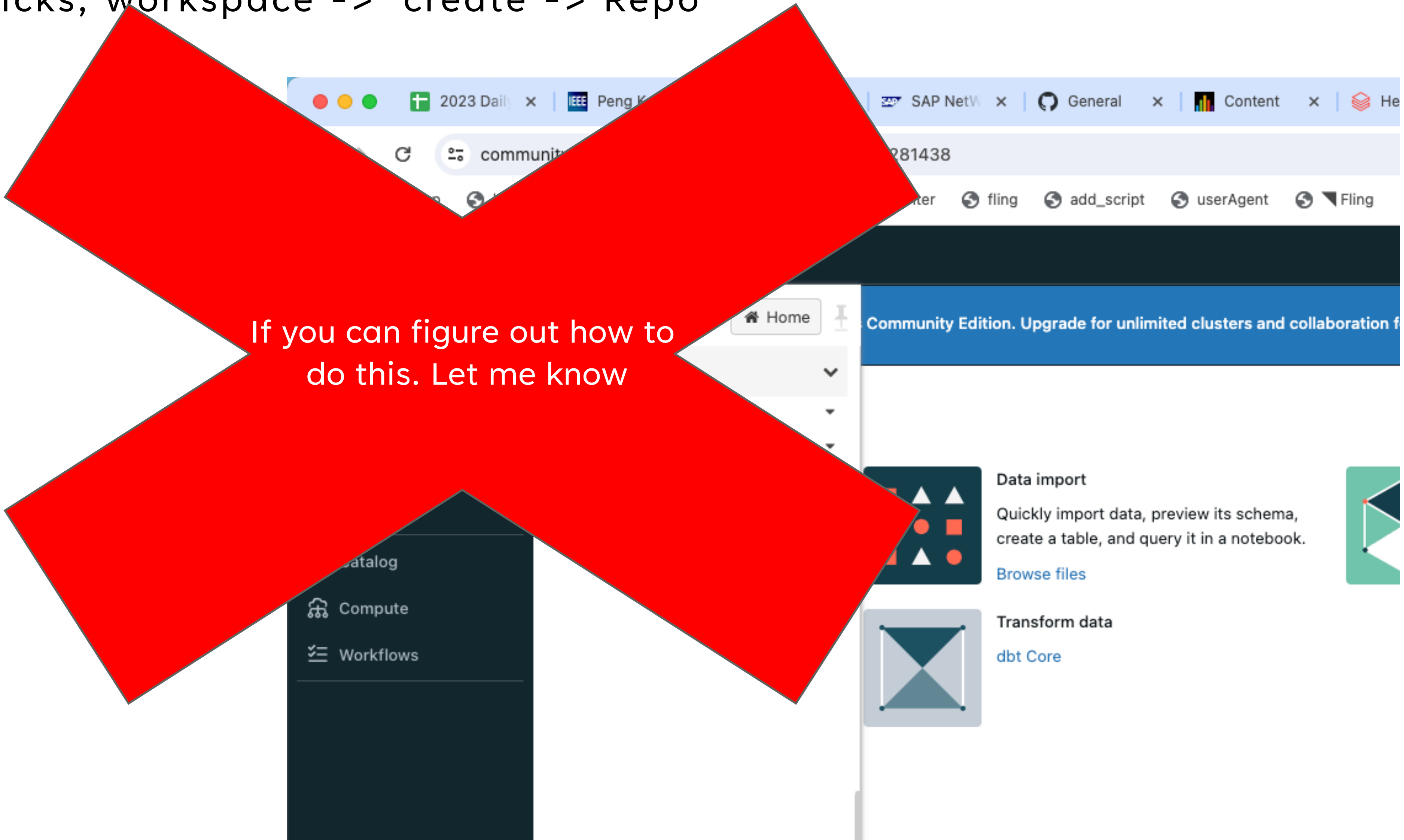
Yrryk3I0P70IckHh

Personal access tokens (class nction like ordinary OAuth access tokens. They can be used instead of a password for HTTPS, or can be used to authenticate to the API over Basic Authentication.

© 2024 GitHub, Inc. Terms Privacy Security Status Docs Contact Manage cookies Do not share my personal informa

COMMIT NOTEBOOKS TO GITHUB

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



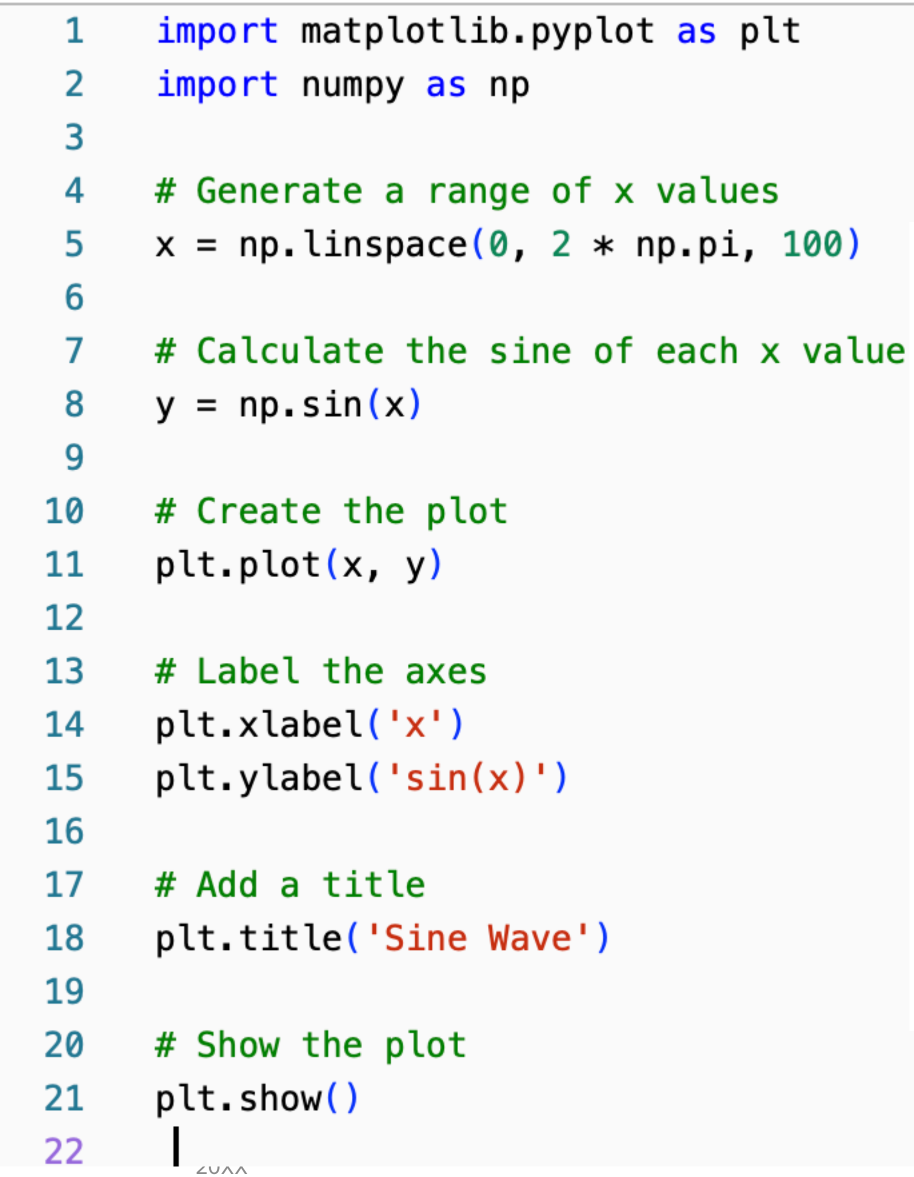
If you can figure out how to do this. Let me know

PYTHON

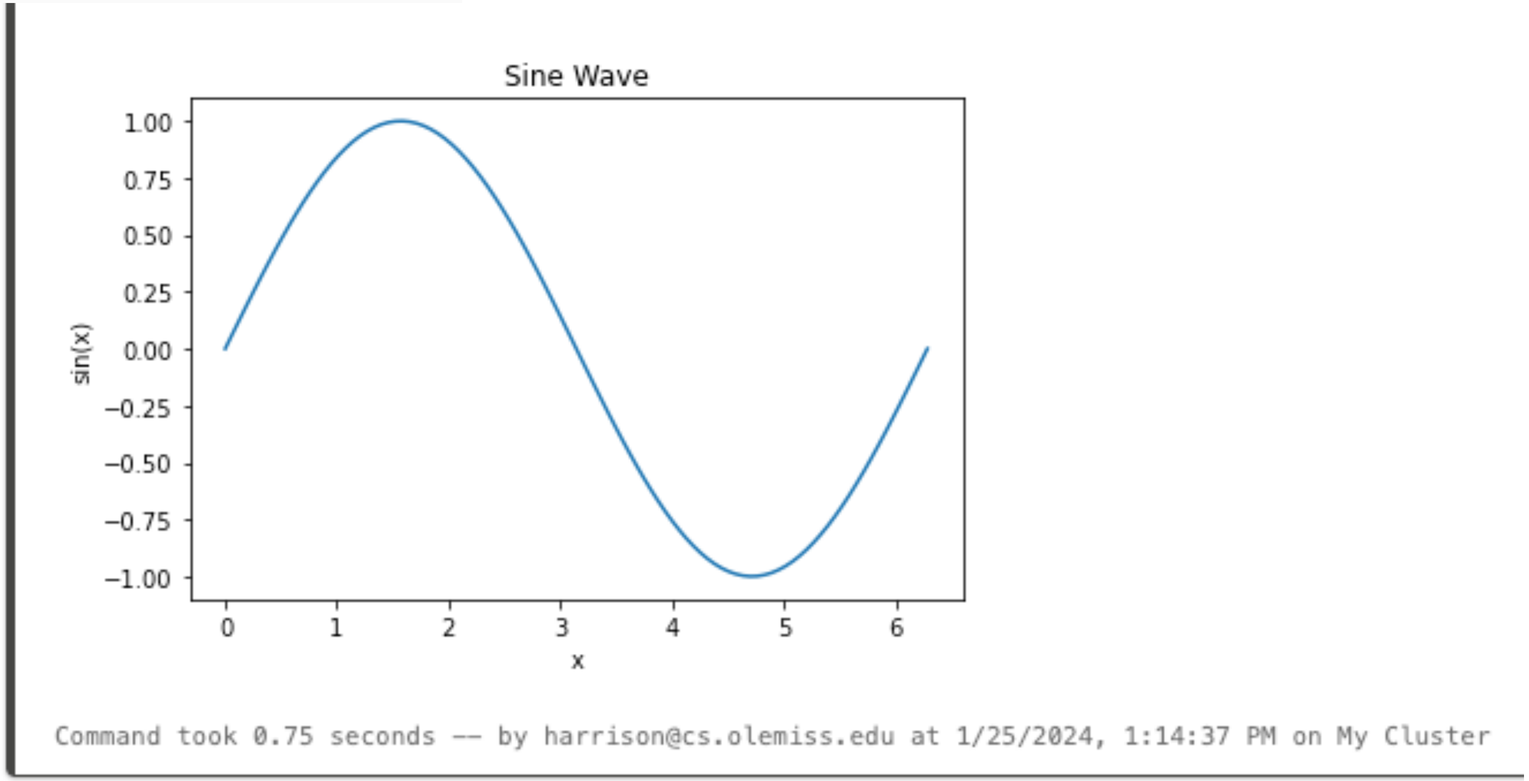
Python is the modern lingua franca of data science.

(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()

```

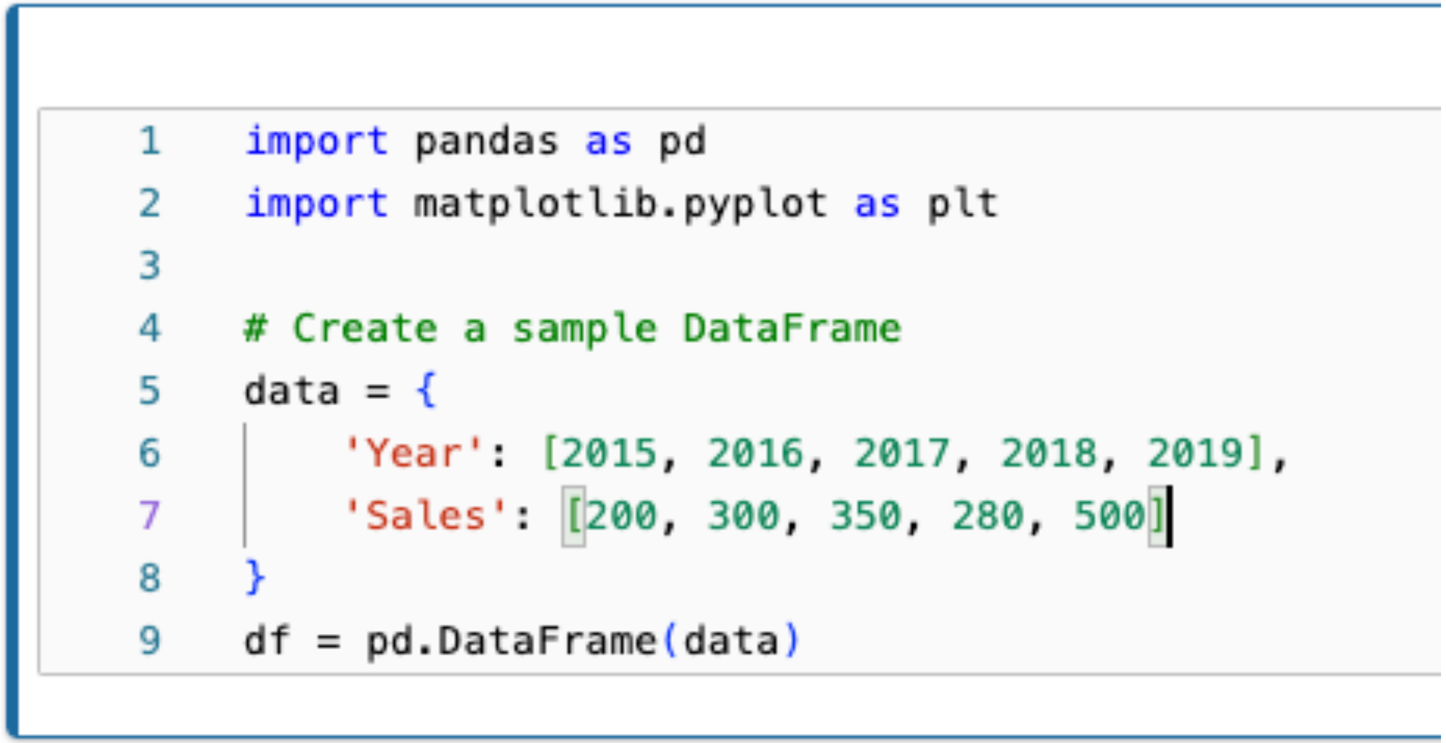


```
Command took 0.75 seconds -- by harrison@cs.olemiss.edu at 1/25/2024, 1:14:37 PM on My Cluster
```

PANDAS DATA FRAMES

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

- But programmatic…



```python
import pandas as pd
import matplotlib.pyplot as plt

# Create a sample DataFrame
data = {
    'Year': [2015, 2016, 2017, 2018, 2019],
    'Sales': [200, 300, 350, 280, 500]
}
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 next Thursday

THANK YOU

David Harrison

Harrison@cs.olemiss.edu