

Abstract geometric lines in the top-left corner of the slide, consisting of several thin black lines forming various polygons and intersecting at different points.

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.

O'REILLY®

Second
Edition

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python



Peter Bruce, Andrew Bruce
& Peter Gedeck

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.



[I own a Samba TV](#)

[Business Solutions](#)

We are the Heartbeat of Television
Data that powers TV innovation

SYLLABUS

CSCI 443 Advanced Data Science

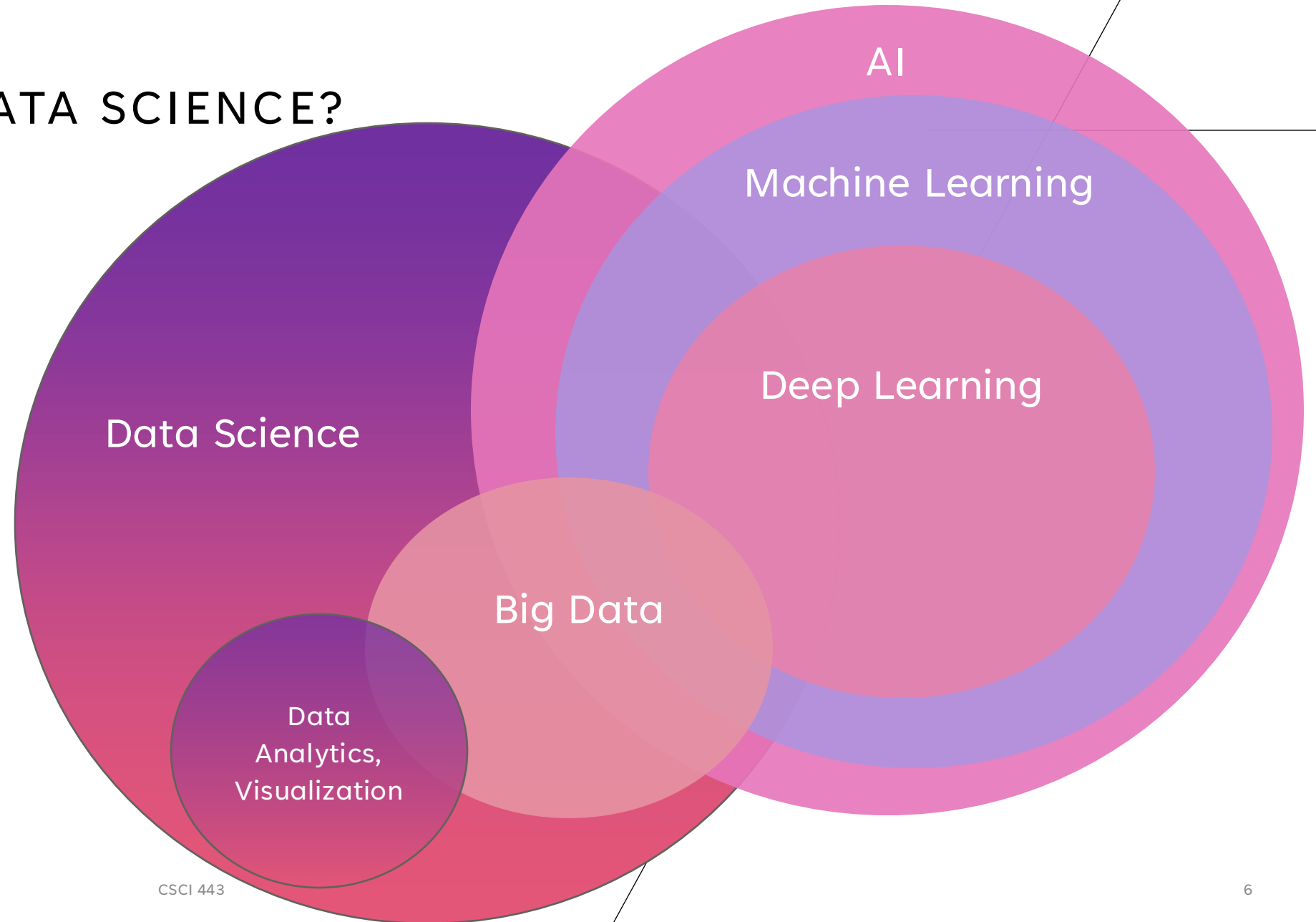
Course Overview

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?

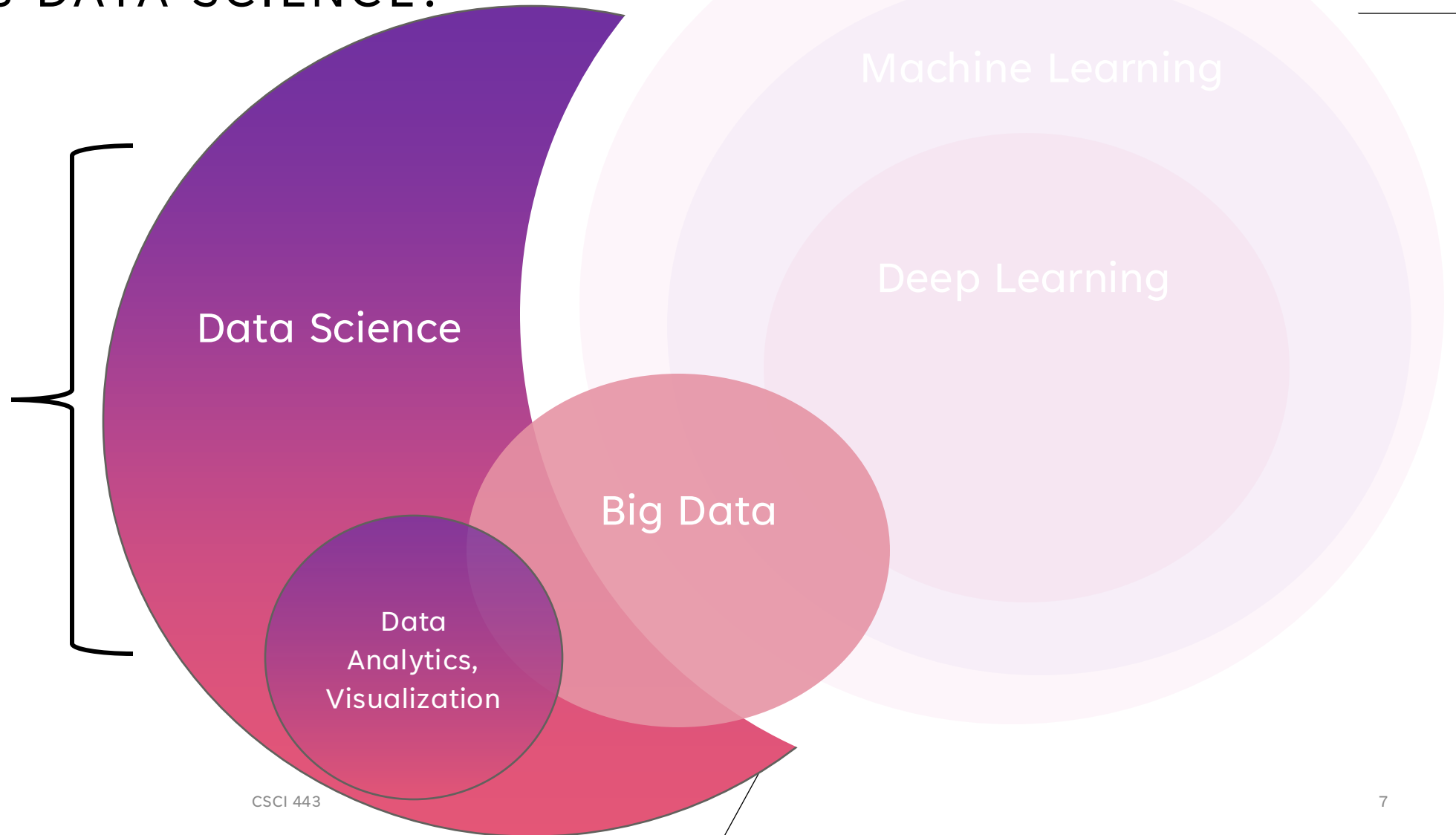
Data science encompasses the entire lifecycle of data processing and analysis, including data collection, cleaning, exploration, modeling, and interpretation. Its focus is on extracting insights and knowledge from data and involves developing methods of recording, storing, and analyzing data to effectively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

WHAT IS DATA SCIENCE?



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.

The screenshot shows a web browser window with the address bar displaying `blackboard.olemiss.edu/ultra/courses/_121946_1/cl/outline`. Below the address bar is a navigation bar with icons for play, stop, left, right, up, down, back, enter, fling, and add_script. The main content area has a dark blue sidebar on the left with a close button (X) and a list of course items: "Csci 443 Advanced Data Science Section 1 2023-2024 SPRG" (with a home icon), "Home Page" (with a dropdown arrow), and "Announcements" (with a dropdown arrow). The main content area on the right shows "Home Page" (with a dropdown arrow) and "Add Course Module". At the bottom right, there is a "My Announcements" section with a settings gear and a close button (X).

← → ↻ 🔍 blackboard.olemiss.edu/ultra/courses/_121946_1/cl/outline

🎮 play ⏹ stop ⏮ left ⏭ right ⬆ up ⬇ down ⏮ back ⏭ enter ⏮ fling ⏭ add_script

Csci 443 Advanced Data Science Section 1 2023-2024 SPRG Home Page

Home Page ▾

Add Course Module

My Announcements ⚙️ ✕

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.

The screenshot shows the GitHub interface for a repository named 'CSCI443_AdvancedDataScience' by user 'dosirrah'. The repository is public. At the top, there are buttons for 'Pin', 'Unwatch' (with a count of 1), and a share icon. Below this, there are buttons for 'main' (with a dropdown), a share icon, a tag icon, 'Go to file', a '+' button, and a 'Code' button (with a dropdown). A modal is open over the 'Code' button, showing options to 'Clone' the repository. The modal has tabs for 'Local' and 'Codespaces'. Under the 'Local' tab, there are three options: 'HTTPS', 'SSH' (which is selected and underlined), and 'GitHub CLI'. Below these options, a text box contains the SSH URL: 'git@github.com:dosirrah/CSCI443_AdvancedD:'. To the right of the text box is a copy icon. Below the text box, it says 'Use a password-protected SSH key.' In the background, the repository's file list is visible, showing files: '.gitignore', 'CSCI443 Syllabus.pdf', and 'README.md'. The repository owner's name 'David Harrison' is also visible with a 'Rename to a s' option.



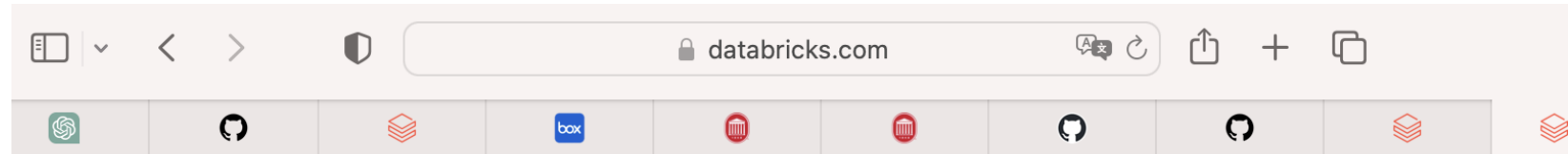
NO GITLAB...

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>



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 Mississippi

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.

Continue

DATABRICKS

Community edition is free.

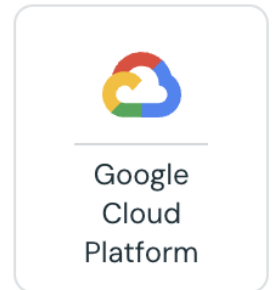
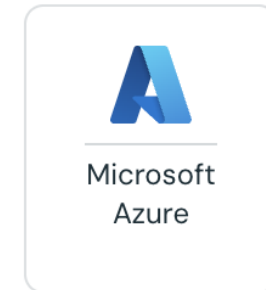
You do not need an AWS or Azure account.

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.



Continue

By clicking "Get Started," you agree to the [Privacy Policy](#) and [Terms of Service](#).

Personal use

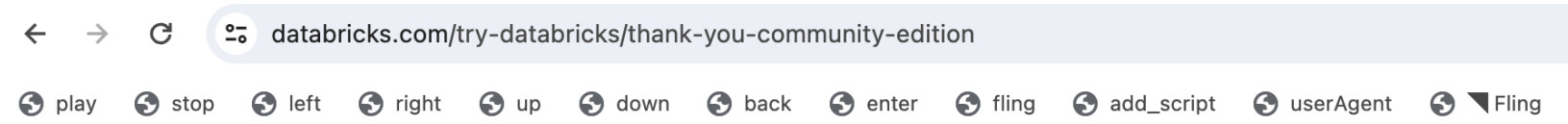
Community Edition is a limited, single node version of Databricks for personal or educational use.

Get started with Community Edition

DATABRICKS

Community edition is free.

Don't worry about "your trial."
This is misleading.



Check your email to start your trial

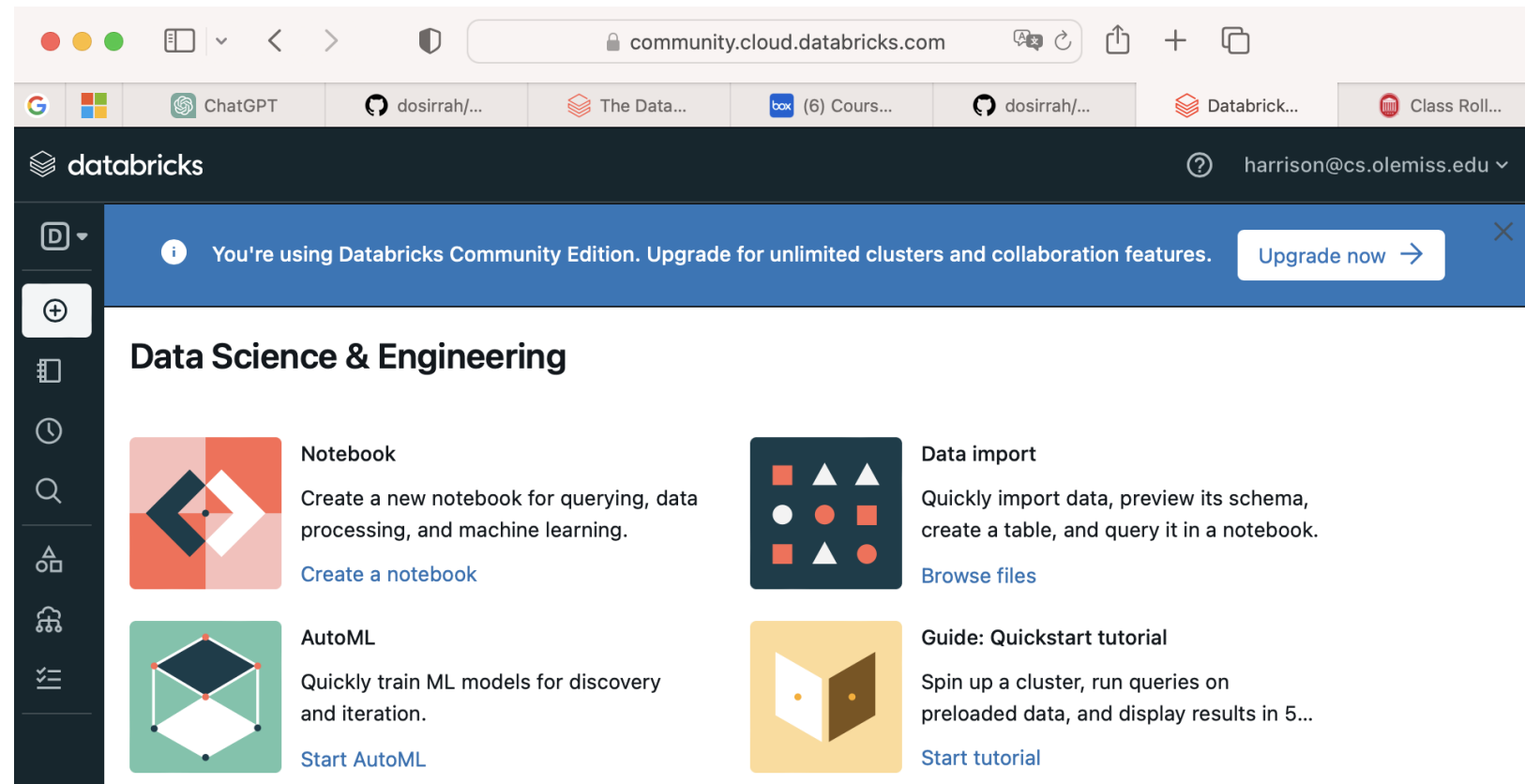
Thank you for signing up. Please validate your email address to start your trial.

DATABRICKS

<https://community.cloud.databricks.com>

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

Ignore “Upgrade now”

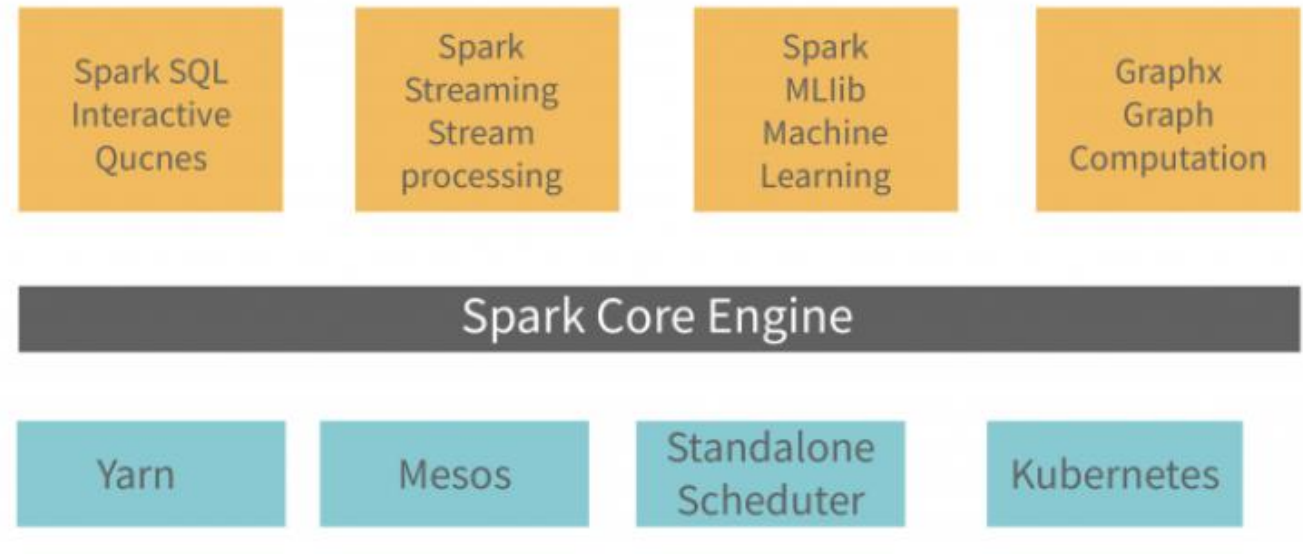


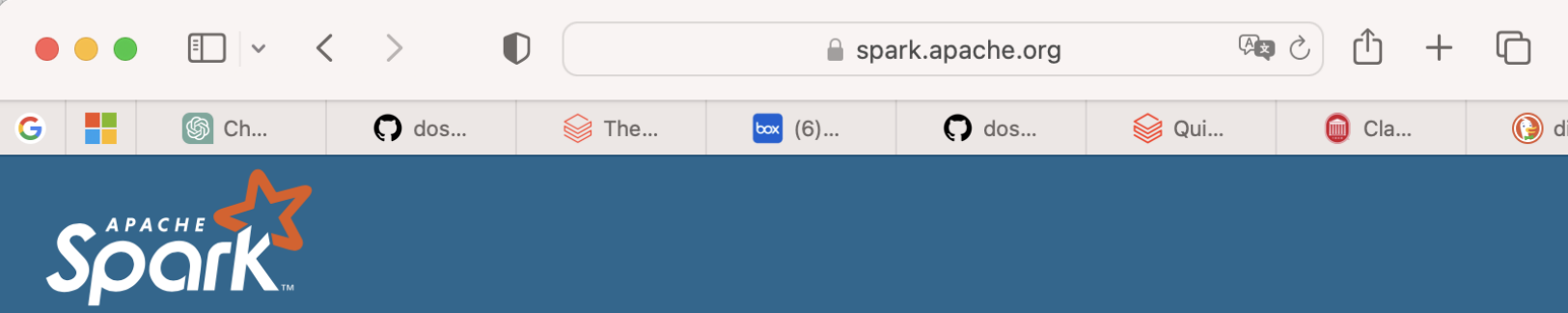
WHY DATABRICKS?

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

Spark unifies:

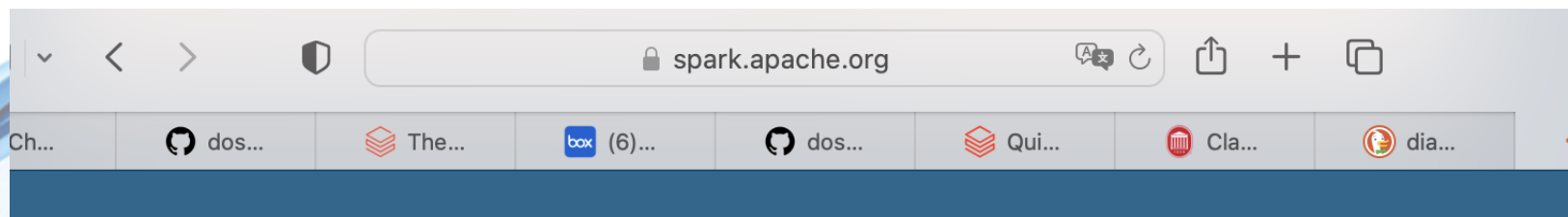
- Batch processing
- Real-time data processing
- Stream analytics (trending, dashboards, etc.)
- Machine learning
- Interactive SQL
- Successor and extension to what was traditionally done with Hadoop or other map-reduce systems.





Unified engine for large-scale data analytics

GET STARTED

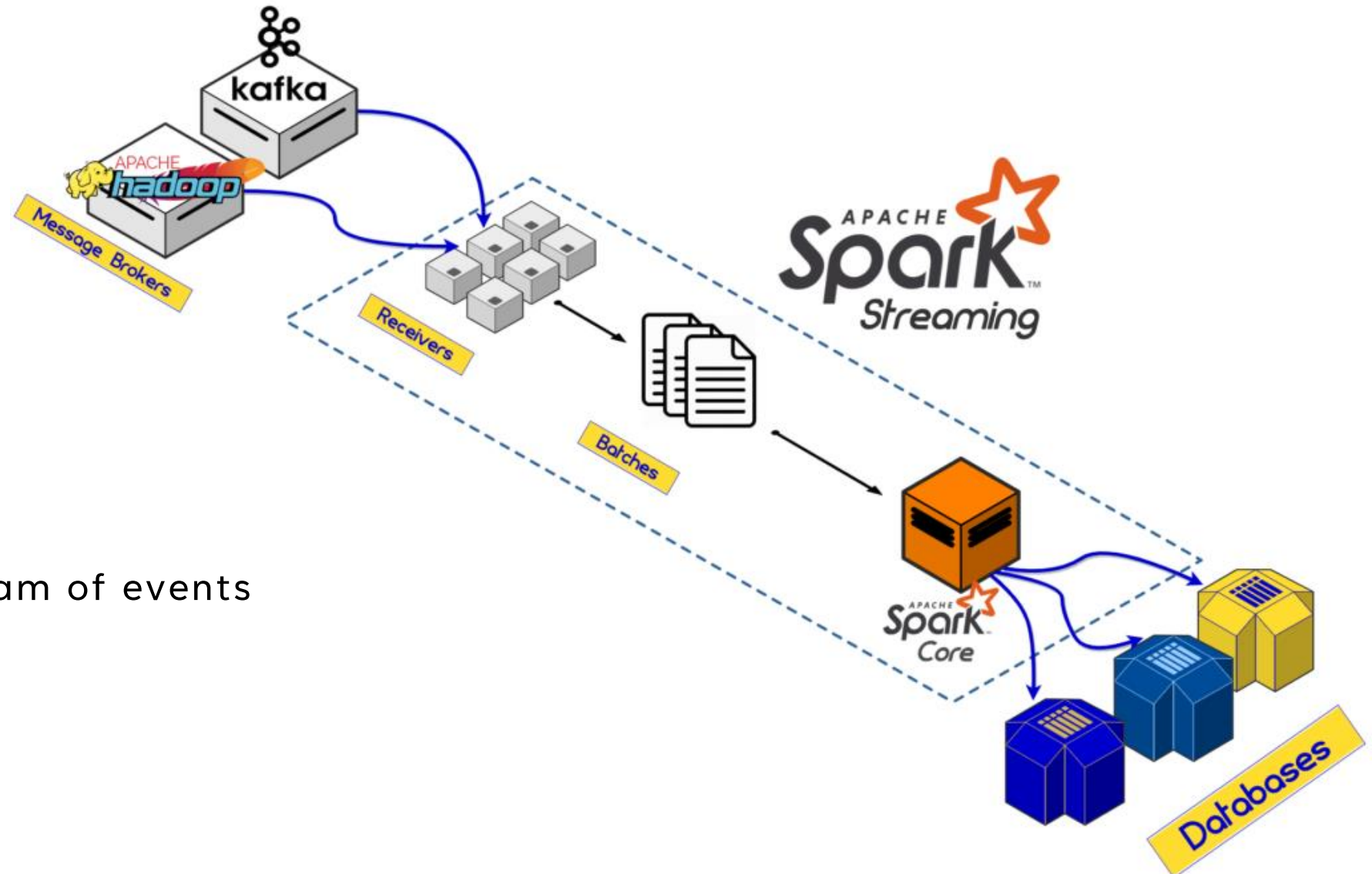


The most widely-used engine for scalable computing

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

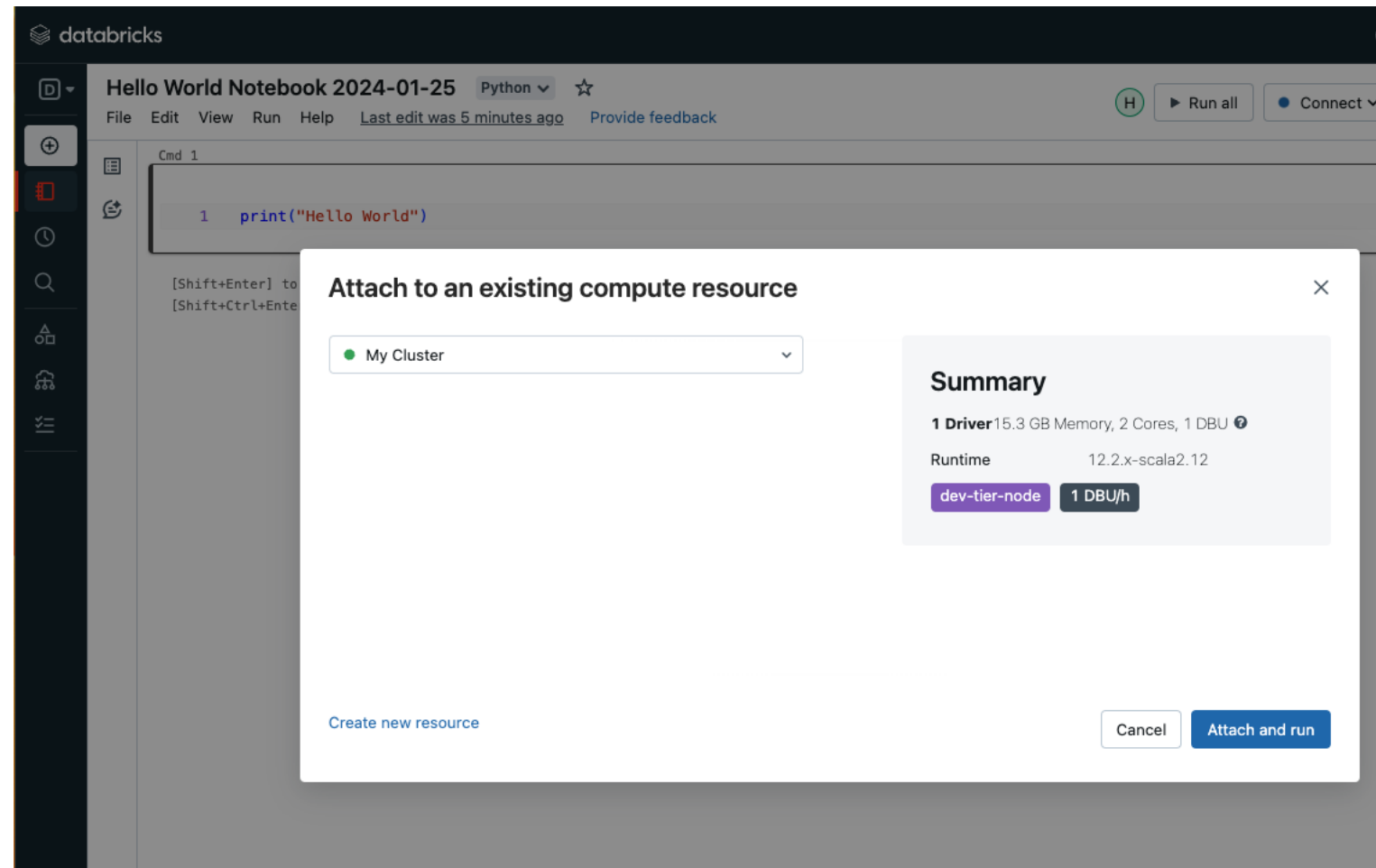
DATABRICKS NOTEBOOKS

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)~~

The screenshot shows a web browser window with multiple tabs. The active tab is a Databricks Notebook titled 'Hello World Notebook 2024-01-25' in Python. The notebook interface includes a left sidebar with icons for file management, a top navigation bar with 'File', 'Edit', 'View', 'Run', and 'Help' menus, and a main content area. The content area displays 'Cmd 1' with the following markdown and code:

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

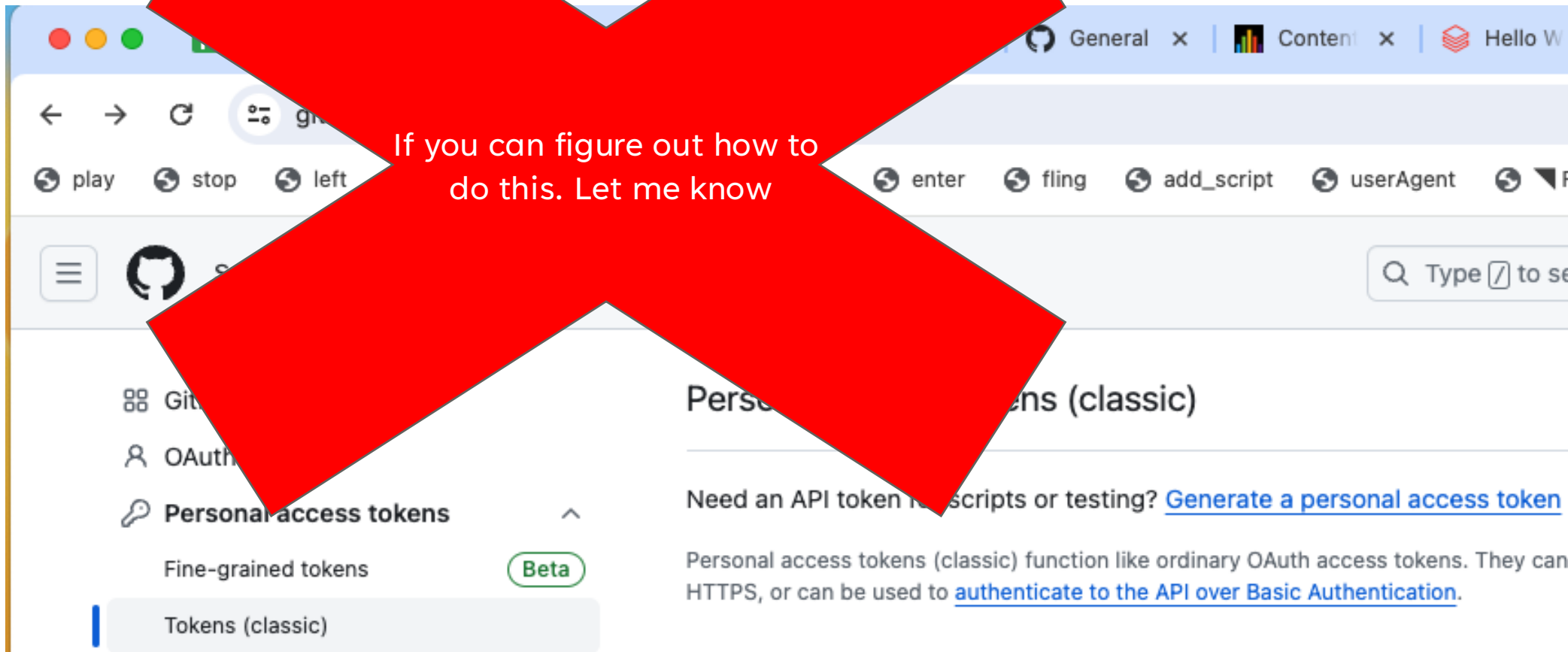
Please connect your notebooks to github.

Checkpoint your work regularly.

Ask chatgpt how

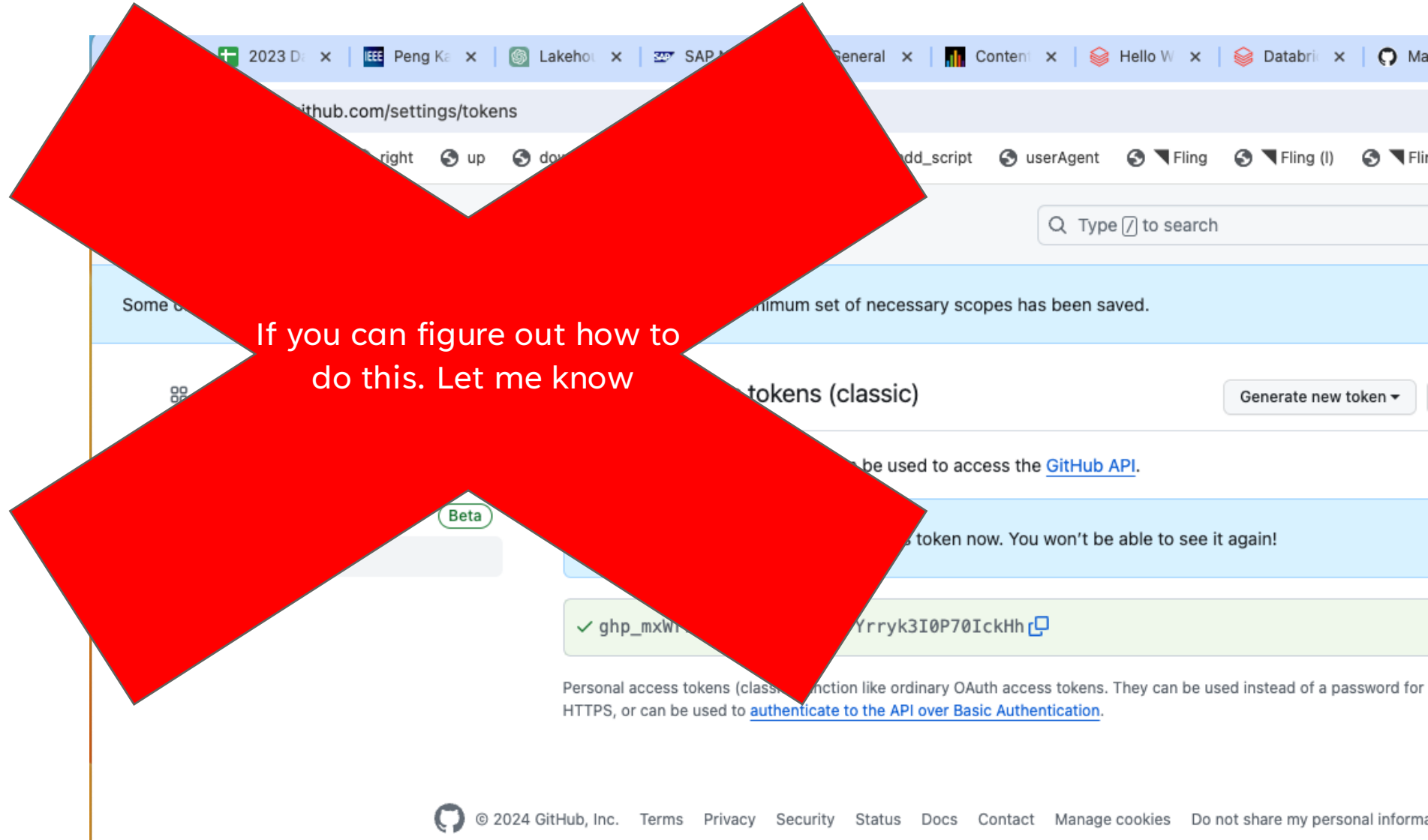
1. On github go to Access Tokens

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



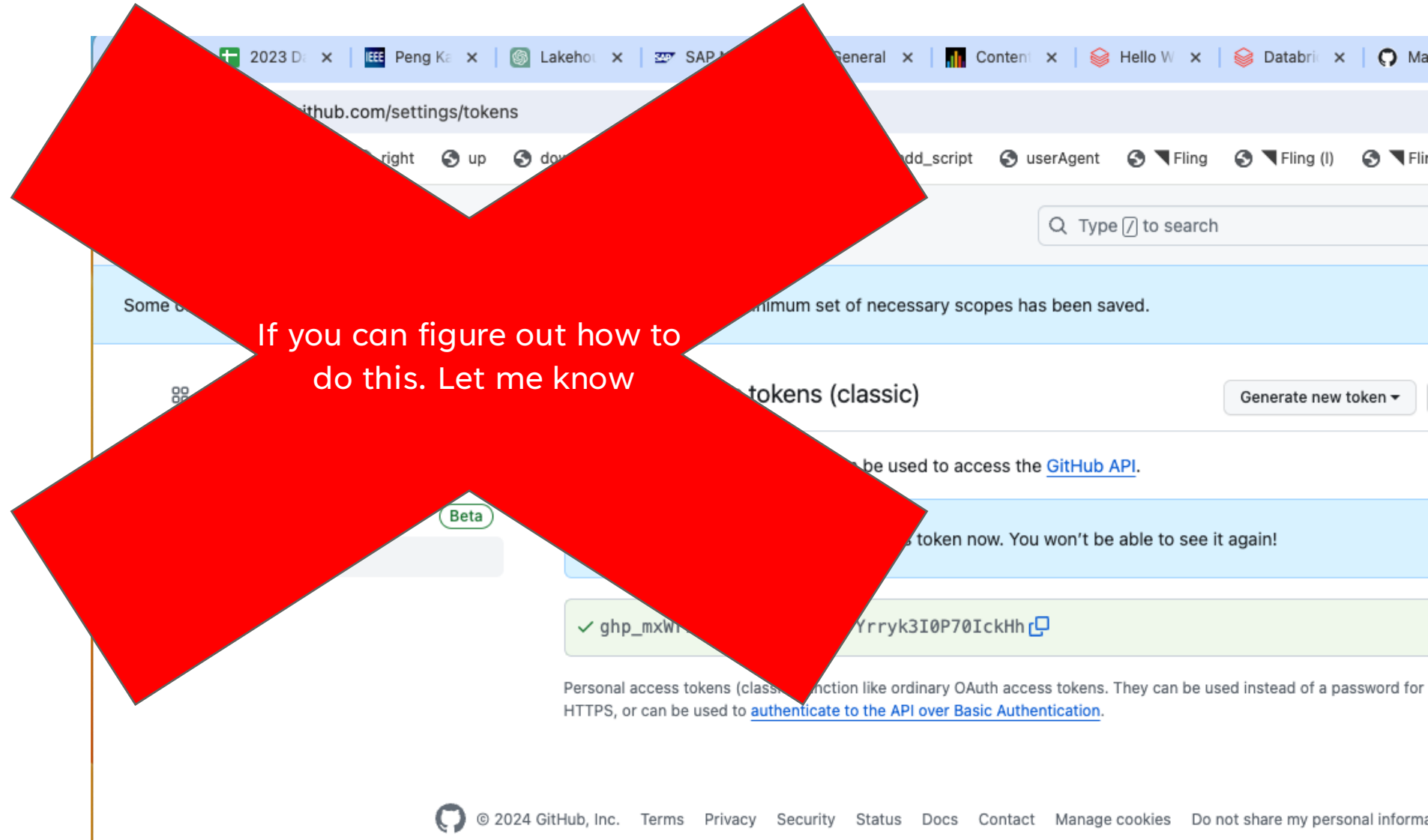
COMMIT NOTEBOOKS TO GITHUB

1. On github generate Personal Access Token.



COMMIT NOTEBOOKS TO GITHUB

1. On github generate Personal Access Token.



COMMIT NOTEBOOKS TO GITHUB

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

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



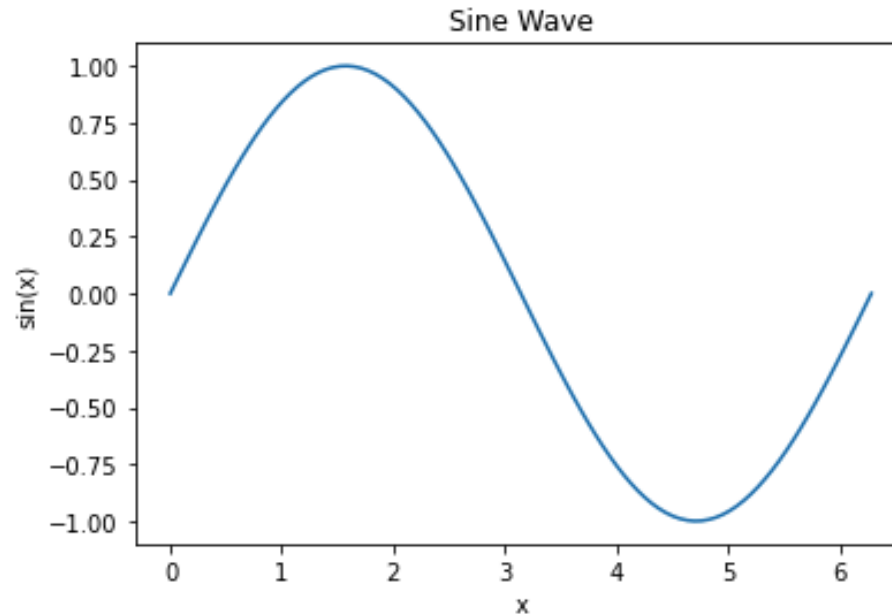
PYTHON

Python is the modern lingua franca of data science.

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

VISUALIZATION EXAMPLE USING MATPLOTLIB

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 # Generate a range of x values
5 x = np.linspace(0, 2 * np.pi, 100)
6
7 # Calculate the sine of each x value
8 y = np.sin(x)
9
10 # Create the plot
11 plt.plot(x, y)
12
13 # Label the axes
14 plt.xlabel('x')
15 plt.ylabel('sin(x)')
16
17 # Add a title
18 plt.title('Sine Wave')
19
20 # Show the plot
21 plt.show()
22 |
```



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...

```
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.

A series of white, overlapping geometric lines and polygons on a black background, located on the left side of the slide.

THANK YOU

David Harrison

Harrison@cs.olemiss.edu