

May-Summer 2024 Data Science Boot Camp

Introduction

Welcome!

- Welcome to the May-Summer 2024 Data Science Boot Camp!
- In this boot camp we will:
 - Learn some python
 - Learn some data science
 - Complete a data science project

Top two resources

- Boot Camp Website,
<https://www.erdosinstitute.org/programs/may-summer-2024/data-science-boot-camp>
- Erdős Institute Slack
 - [may-summer-2024-cohort](#) is a private channel.
 - You should already be a member!
 - [may-summer-2024-data-science](#) is a public channel you should join.

Lecturer

- Steven Gubkin, PhD
 - Head of Training and Assessment at Erdős since 01/01/24
- Graduated from OSU Math in 2016
- Taught math at Cleveland State from 2016 - 2023



Your contact for access

- Amalya Lehmann, PhD
 - PhD in Music History, Literature, and Theory from UC Berkeley
- Your top contact for:
 - Slack channel access
 - GitHub repository access



Group Project Coordinator

- Alec Clott, PhD
 - Head of Data Science Projects
 - Sr. Principal, Quantitative Analytics and Data Science at Gartner
- Graduated from OSU Political Science in 2021



The Erdős Institute Projects

May 2024

Goals

- An opportunity to work with real-world data and produce findings in a short time-span
- Focus on substantive areas (environment, health, finance, etc.) using techniques from the bootcamp.
 - The focus should be on using what we learn.
 - Okay to use more advanced methods. Just make sure to compare their performance to the best model you could make using methods covered in the bootcamp.
- Building your portfolio is crucial in the data science market, provides a framework for job interviews

Projects

- Portfolio-worthy data science project/product
- Includes:
 - 5-minute overview video and slide show presentation
 - Annotated GitHub
 - Executive Summary
- Reviewed by project judges
- Top 5 projects will present to all participants in our closing ceremony for the Spring 2024 Bootcamp

Team Formation

Background of boot camp attendees

- Hundreds of students from all over the world
- Some of you may know other attendees, others of you won't
- Many different backgrounds (subject areas, experience with coding)*
- Various types of data science career goals
- Various goals for the bootcamp
- Various goals for the projects

*And that is totally fine and expected!

Read these documents

<https://www.erdosinstitute.org/programs/may-summer-2024/data-science-boot-camp>

(Project Information at Bottom)

Project Expectations

Overall Structure

- **Team size:** 3-5 people
- **Goals:** “portfolio” project
 - Can be used in job interviews (when the time comes)
 - Results have business value
 - Communicate to lay-people and team of data scientists
- **Structure**
 - Group meetings -- each group decides how much time they want to spend
 - Check-in with project mentor on a regular basis (15-30 min)

Project Requirements

- **Instructions at the bottom of the May 2024 Data Science Bootcamp Page**
- **In order to get an Erdős certificate, you must complete a data science project start to finish**
 - Project must be coded in Python
 - Have an annotated GitHub repository
 - Executive summary of your project results and implications
 - ***For presentation day:***
 - **5-min** pre-recorded PowerPoint presentation detailing project process from start to finish
 - Judges will vote on winners!
 - More info will be given closer to project day

Your To-Do List

First Important Dates:

Schedule

Click on any date for more details

Problem Solving Session 1

May 6, 2024 at 11:00:00 AM

EVENT

Office Hours

May 6, 2024 at 12:00:00 PM

EVENT

Lecture 1: Introduction

May 6, 2024 at 3:00:00 PM

EVENT

Problem Solving Session 2

May 7, 2024 at 11:00:00 AM

EVENT

Office Hours

May 7, 2024 at 12:00:00 PM

EVENT

Lecture 2: Data Collection

May 7, 2024 at 3:00:00 PM

EVENT

Problem Solving Session 3

May 8, 2024 at 11:00:00 AM

EVENT

Office Hours

May 8, 2024 at 12:00:00 PM

EVENT

Lecture 3: Regression I

May 8, 2024 at 3:00:00 PM

EVENT

May 10, 2024 Submit Team Proposal to Project Formation Page

11:59 PM

If you want to propose a project, or have an idea for a project, submit it by this date.

May 12, 2024 Finalized Teams with Preliminary Project Idea

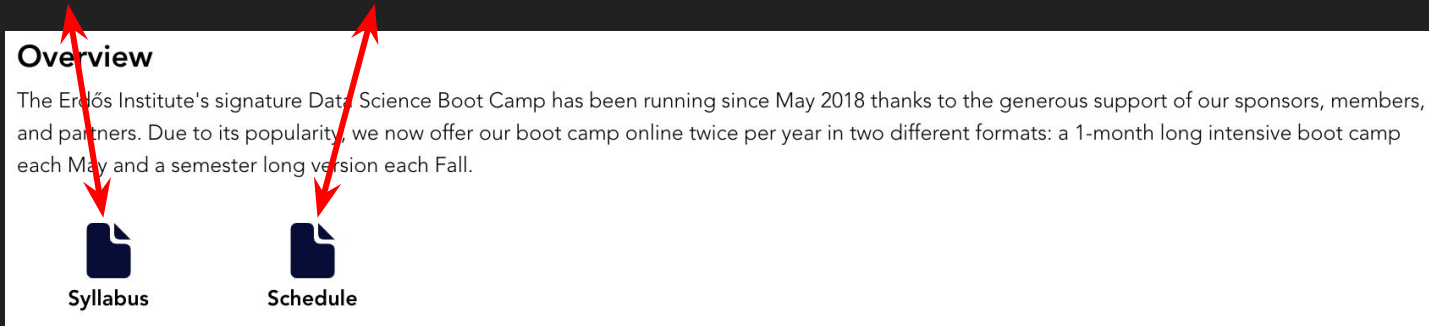
11:59 PM

Teams need to be finalized by this point. If you proposed or created a project, you must have others in your group. If you did not propose or create a project, you must join an open group.

Note: You can find these dates at the bottom of the course website



Boot Camp Format: Non-Project Portion

- 12 Live Lectures
- 12 Problem Solving Sessions
- All Zoom links can be found in your Erdős profile or on the course website
- Syllabus and Schedule can be found on the course website



Overview

The Erdős Institute's signature Data Science Boot Camp has been running since May 2018 thanks to the generous support of our sponsors, members, and partners. Due to its popularity, we now offer our boot camp online twice per year in two different formats: a 1-month long intensive boot camp each May and a semester long version each Fall.

 **Syllabus**  **Schedule**

Two red double-headed arrows are overlaid on the image. One arrow connects the 'Overview' heading to the 'Syllabus' link. The other arrow connects the 'Overview' heading to the 'Schedule' link.

Lectures

- Live lectures 3:00 - 4:30 PM ET every M/T/W/R until May 23rd
 - Will be recorded and uploaded to the website
- Every lecture jupyter notebook already has a pre-recorded lectures on the website.

Problem Sessions

- One hour to work on problem sets in small groups
- Every M/T/W/R 11:00 AM - 12:00 PM ET until May 23rd
 - Will not be recorded
- TAs will rotate between groups to assist and observe
- Many problem sessions also have a “prep notebook” with prerequisite practice.

Office Hour

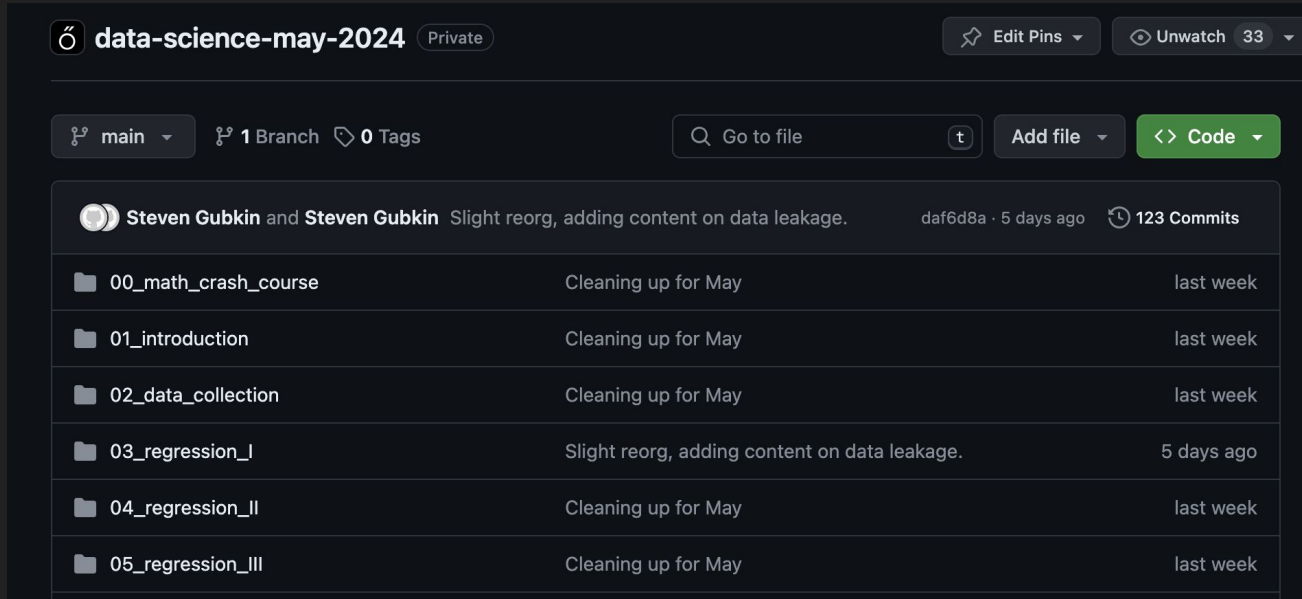
- Office Hour is every M/T/W/R/F 12:00PM to 1:00PM ET and by appointment.
 - Ask anything about course content, projects, debugging, etc.

Getting Set Up

- Clone the repository
- Be able to open a jupyter notebook

The GitHub Repository

- Link can be found on the course website
- Contains all of the educational content for the boot camp



The screenshot shows the GitHub repository page for 'data-science-may-2024'. The repository is marked as 'Private'. At the top, there are buttons for 'Edit Pins', 'Unwatch', and a count of '33'. Below this, the 'main' branch is selected, with '1 Branch' and '0 Tags' indicated. A search bar 'Go to file' and buttons for 'Add file' and 'Code' are present. The commit history shows a recent commit by 'Steven Gubkin' with the message 'Slight reorg, adding content on data leakage.' and '123 Commits' in total. Below the commit history, a table lists the repository's structure:

File/Folder	Commit Message	Time
00_math_crash_course	Cleaning up for May	last week
01_introduction	Cleaning up for May	last week
02_data_collection	Cleaning up for May	last week
03_regression_I	Slight reorg, adding content on data leakage.	5 days ago
04_regression_II	Cleaning up for May	last week
05_regression_III	Cleaning up for May	last week

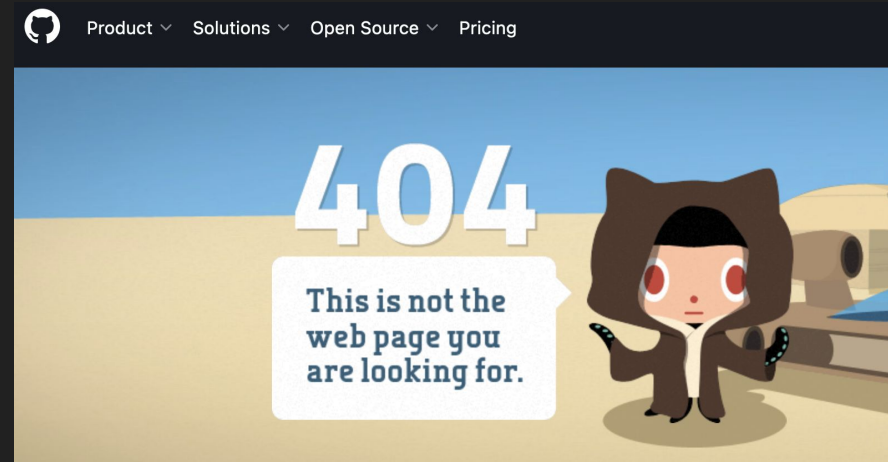
The GitHub Repository - Steps

- Sign into your GitHub account
- Clone the repository onto your computer
 - Can find instructions in the “First Steps” section of the website
- Everyday of the boot camp you will need to “pull” the updates to the repository
 - Look for “Getting Started with GitHub” in the “First Steps” section of the website
- Either make a folder where you copy over files you want to work on (leaving the git repo folder “clean”) or make a local branch where you do your work.

The GitHub Repository - 404 Issue

If you receive the 404 Error when clicking repo link:

- Check you are signed in
- Check that you have added your GitHub link to your Erdős profile
- Message Amalya about being added to the repository



Jupyter Notebooks

- All educational content contained in jupyter notebooks
- Allows combination of markdown and python code
- Let's look at an example

Jupyter Notebooks - Getting Set Up

- Follow Step 3 Under “First Steps” on website
- Lots of options:
 - Visual Studio Code: I use this one
 - Jupyter Notebook
 - Anaconda Navigator
 - Many other options

Conda Environment

- If you want the most streamlined experience possible this semester, you should set up an `erdos_may_2024` conda environment and run all of the notebooks with this environment.
 - Instructions in the repo README document
- Make sure you can run the following notebooks with this environment to confirm everything is working correctly:
 - `02_data_collection/problem_session_2/1_Problem_Session_2_Complete.ipynb`
 - `07_time_series_II/problem_session_7/1_Problem_Session_7_Complete.ipynb`
 - `11_ensemble_II/problem_session_11/Problem_Session_Complete.ipynb`
 - `12_neural_networks/problem_session_12/1_Problem_Session_12_Complete.ipynb`

Questions & Concerns?