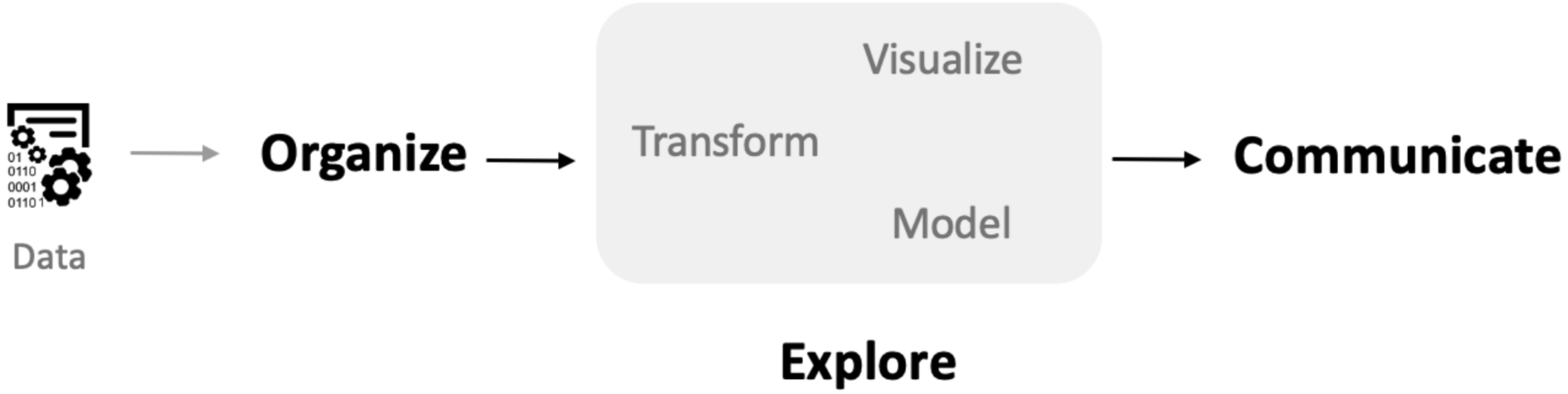
DSST289: Introduction to Data Science

Welcome!

1. Brief Overview

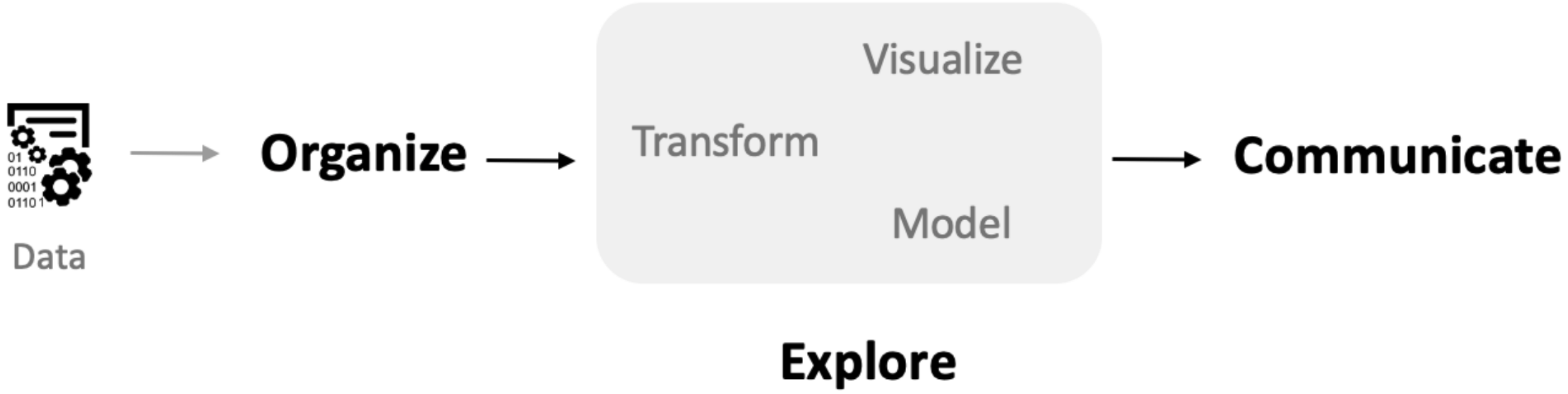
Data Science Pipeline

A standard, highly abstract diagram showing the flow of information when doing data science work.



Data Science Pipeline

A standard, highly abstract diagram showing the flow of information when doing data science work.



We are going to learn how to do all of these elements using a variety of tools, including **spreadsheets** (data collection) and **Python** (most of the other steps) as well as some custom hardware/software for certain kinds of data.

2. My Background

My area of study focuses on large-scale **text** and **image** analysis. I've had a variety of positions (full-time and consulting) in industry and work with data from a variety of domains.

Industry Roles

IBM

AT&T Bell Laboratories

DARPA

Pfizer

TRAVELERS

Academic Positions

Yale

CONNECTICUT COLLEGE 1911

UNIVERSITY OF RICHMOND

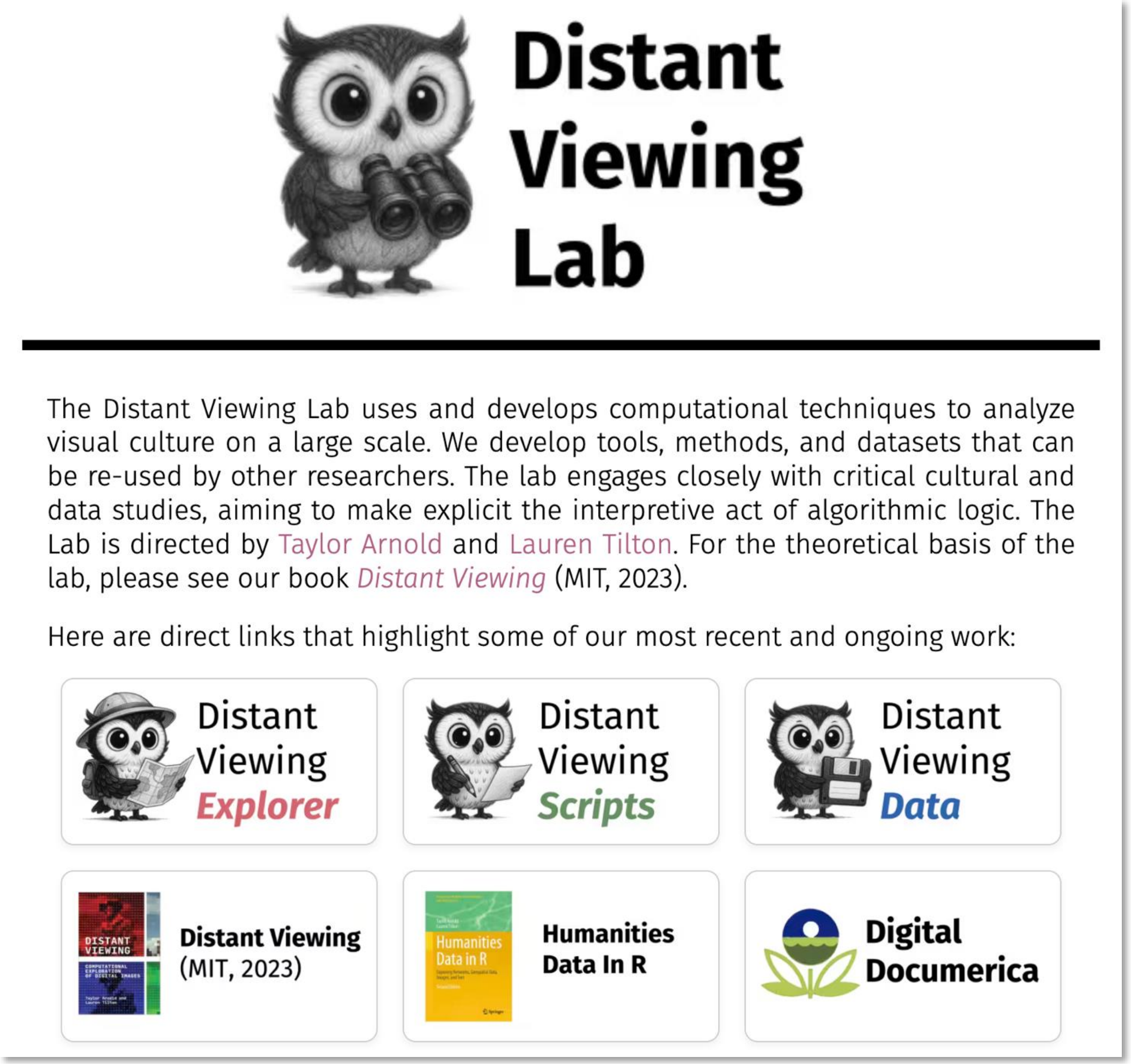
COLLEGIUM DE LYON Université de Lyon

pictoria CONSORTIUM HUMA-NUM

Université Paris Cité

Recently, focusing on the research laboratory I direct along with Lauren Tilton (RHCS) at UR.

I’ll be integrating real examples of the data we study (i.e., television and photography) into our class this semester.



The Distant Viewing Lab uses and develops computational techniques to analyze visual culture on a large scale. We develop tools, methods, and datasets that can be re-used by other researchers. The lab engages closely with critical cultural and data studies, aiming to make explicit the interpretive act of algorithmic logic. The Lab is directed by Taylor Arnold and Lauren Tilton. For the theoretical basis of the lab, please see our book *Distant Viewing* (MIT, 2023).

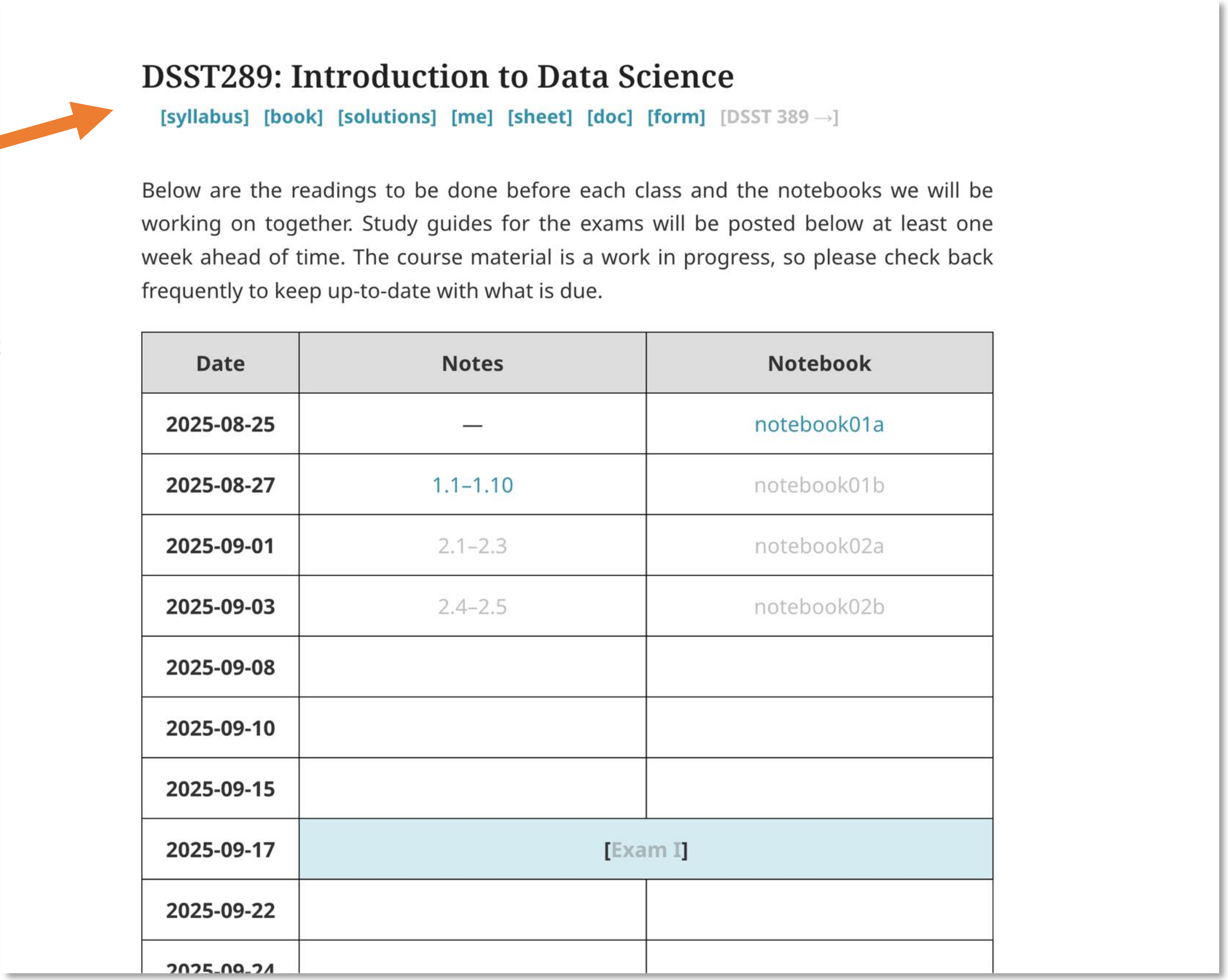
Here are direct links that highlight some of our most recent and ongoing work:

3. Syllabus

Course Structure

All of the notes and materials that you need this semester can be found on the public **course website.**

Let’s go through a few of these elements.



DSST289: Introduction to Data Science

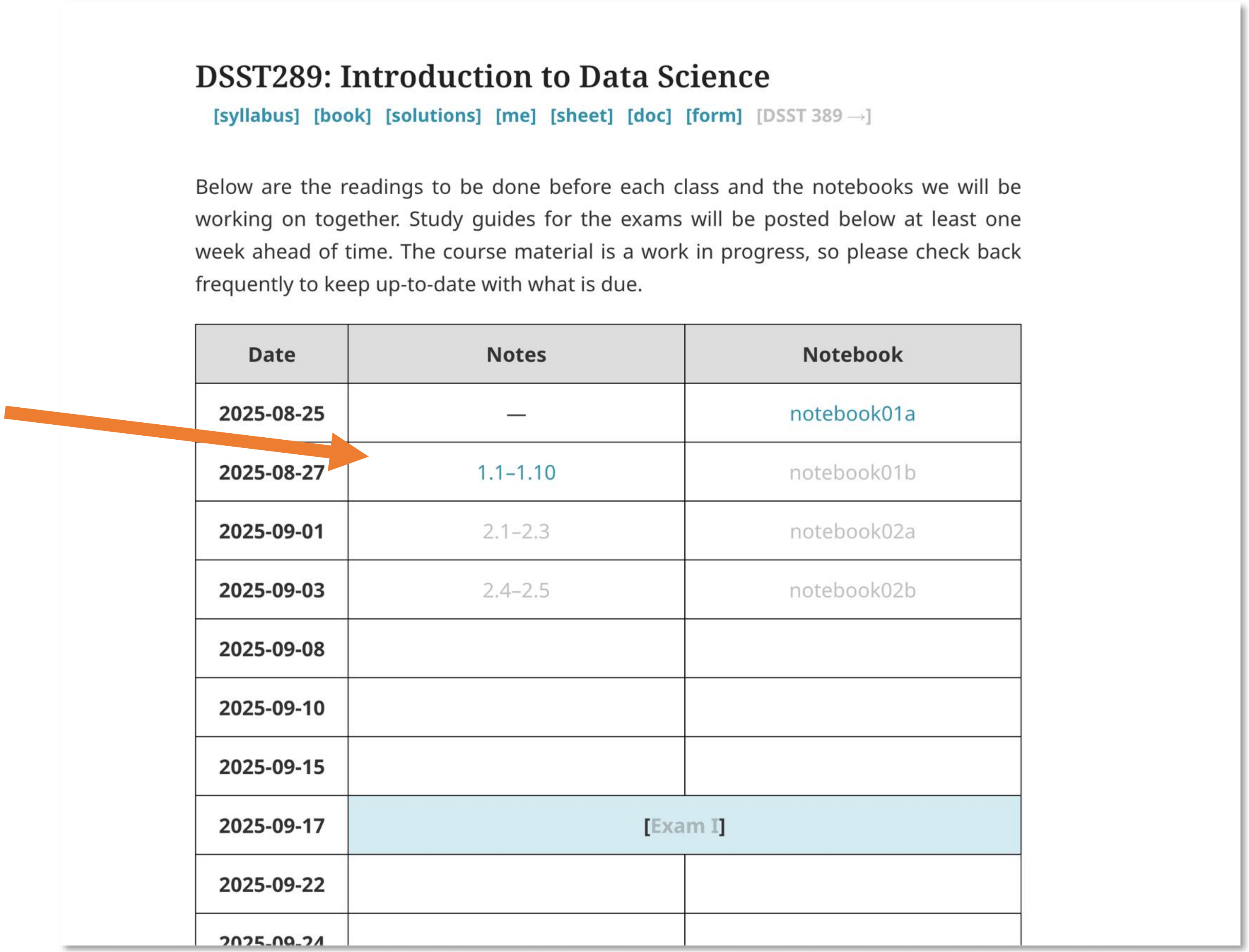
[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Course Structure

Prior to each course meeting, you should do the reading assigned in the **Notes** section on the website.



DSST289: Introduction to Data Science

[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

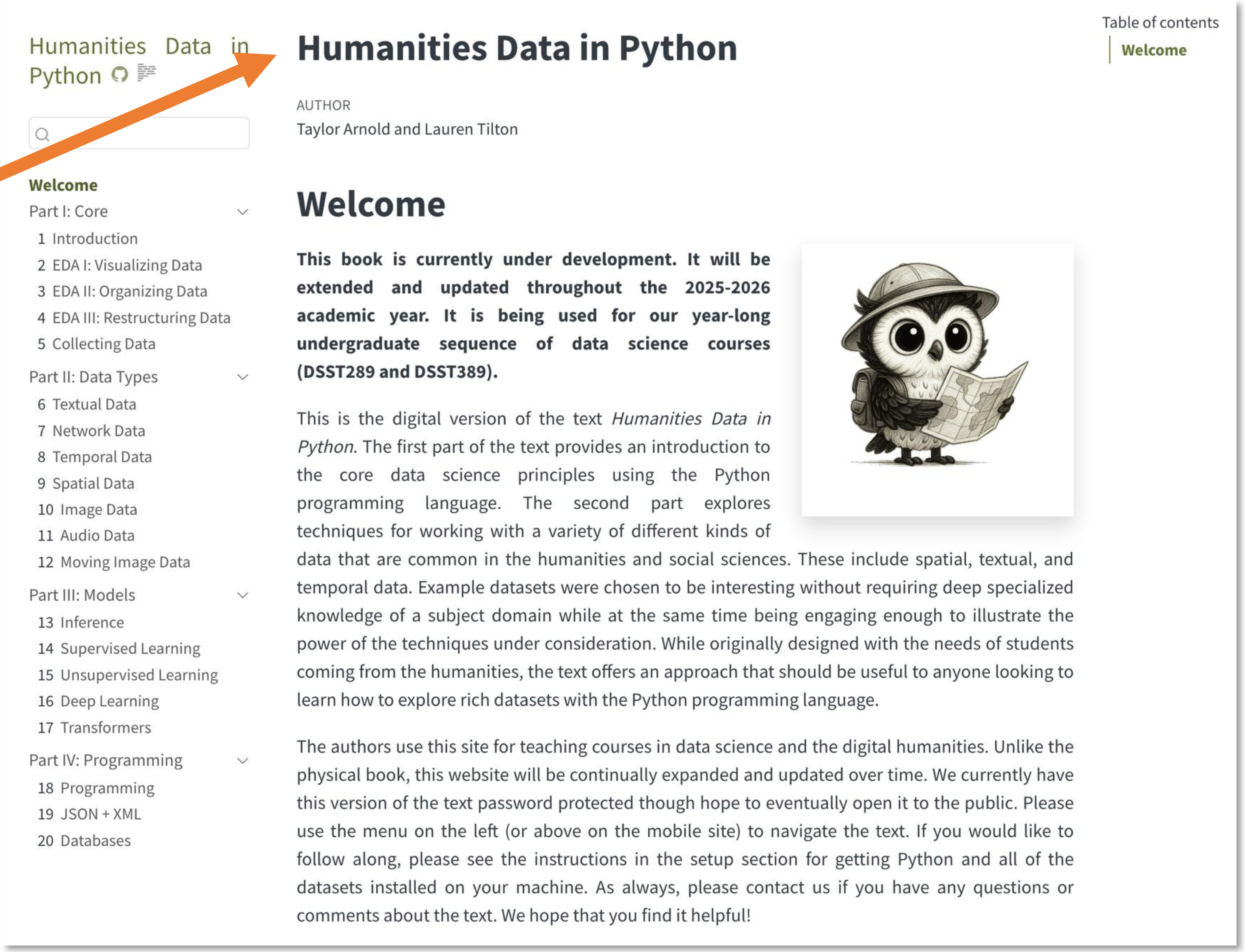
Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Course Structure

Most readings will come from the book *Humanities Data in Python*. I am in the middle of writing this and will be making changes throughout the semester.

And don't worry, it's not just for the humanities!



Humanities Data in Python

Table of contents

Welcome

Humanities Data in Python

AUTHOR

Taylor Arnold and Lauren Tilton

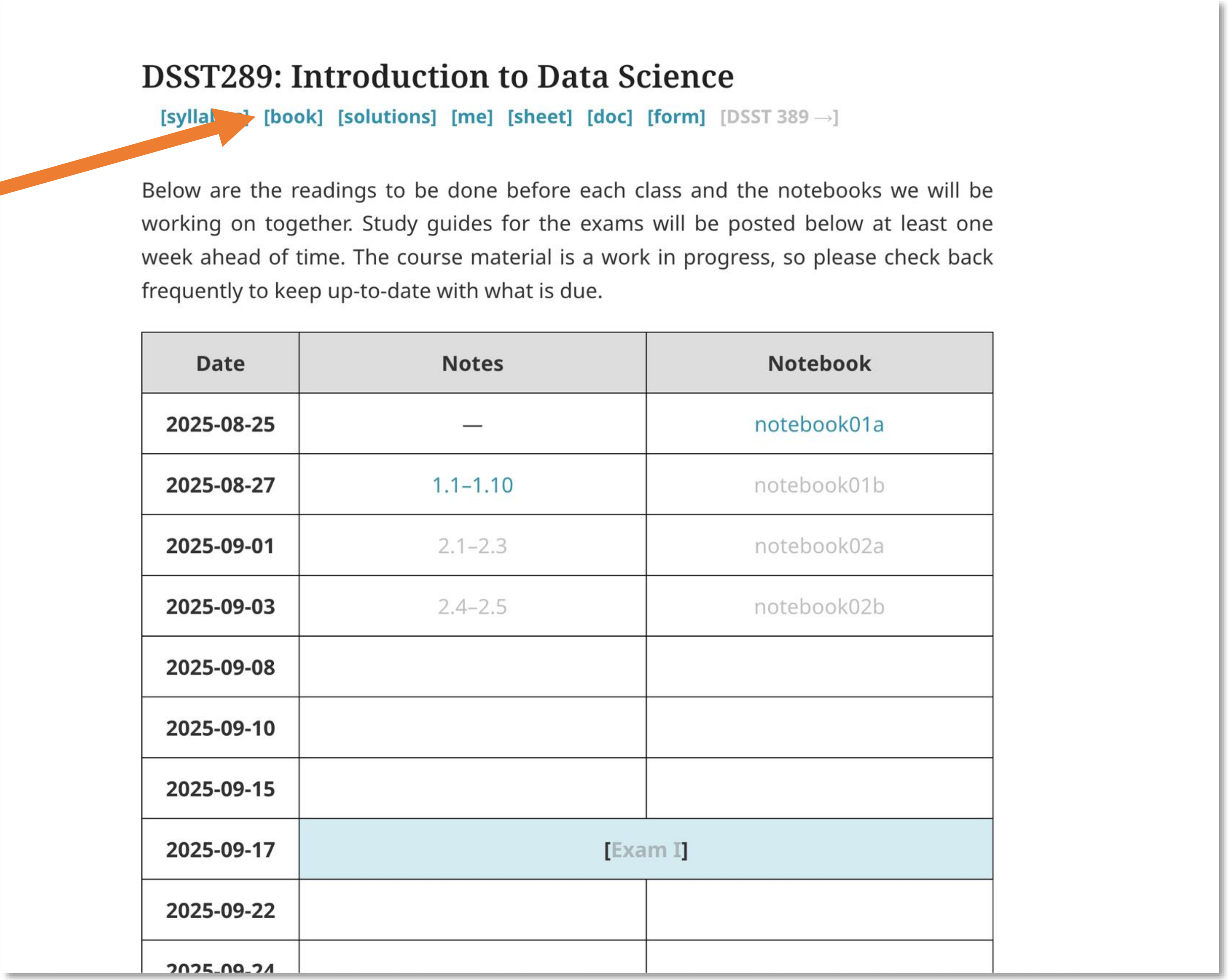
Welcome

This book is currently under development. It will be extended and updated throughout the 2025-2026 academic year. It is being used for our year-long undergraduate sequence of data science courses (DSST289 and DSST389).

This is the digital version of the text *Humanities Data in Python*. The first part of the text provides an introduction to the core data science principles using the Python programming language. The second part explores techniques for working with a variety of different kinds of data that are common in the humanities and social sciences. These include spatial, textual, and temporal data. Example datasets were chosen to be interesting without requiring deep specialized knowledge of a subject domain while at the same time being engaging enough to illustrate the power of the techniques under consideration. While originally designed with the needs of students coming from the humanities, the text offers an approach that should be useful to anyone looking to learn how to explore rich datasets with the Python programming language.

The authors use this site for teaching courses in data science and the digital humanities. Unlike the physical book, this website will be continually expanded and updated over time. We currently have this version of the text password protected though hope to eventually open it to the public. Please use the menu on the left (or above on the mobile site) to navigate the text. If you would like to follow along, please see the instructions in the setup section for getting Python and all of the datasets installed on your machine. As always, please contact us if you have any questions or comments about the text. We hope that you find it helpful!

Course Structure



DSST289: Introduction to Data Science

[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

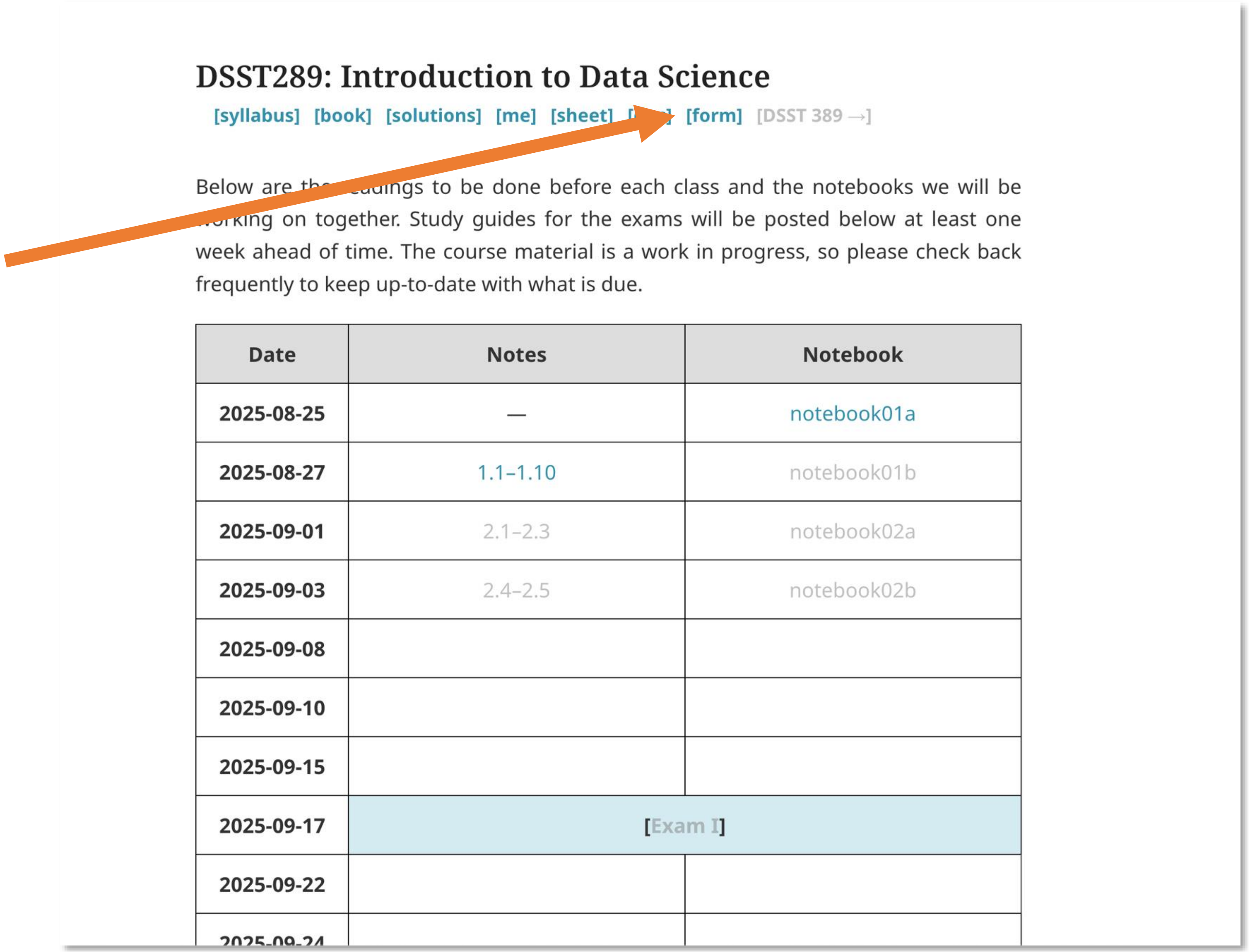
Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

There is also a direct link at the top of the page to the book.

Course Structure

At the start of class, we will fill out the course form together. This is a form of checking attendance and sometimes doing a bit of data collection.



DSST289: Introduction to Data Science

[syllabus] [book] [solutions] [me] [sheet] [illegible] [form] [DSST 389 →]

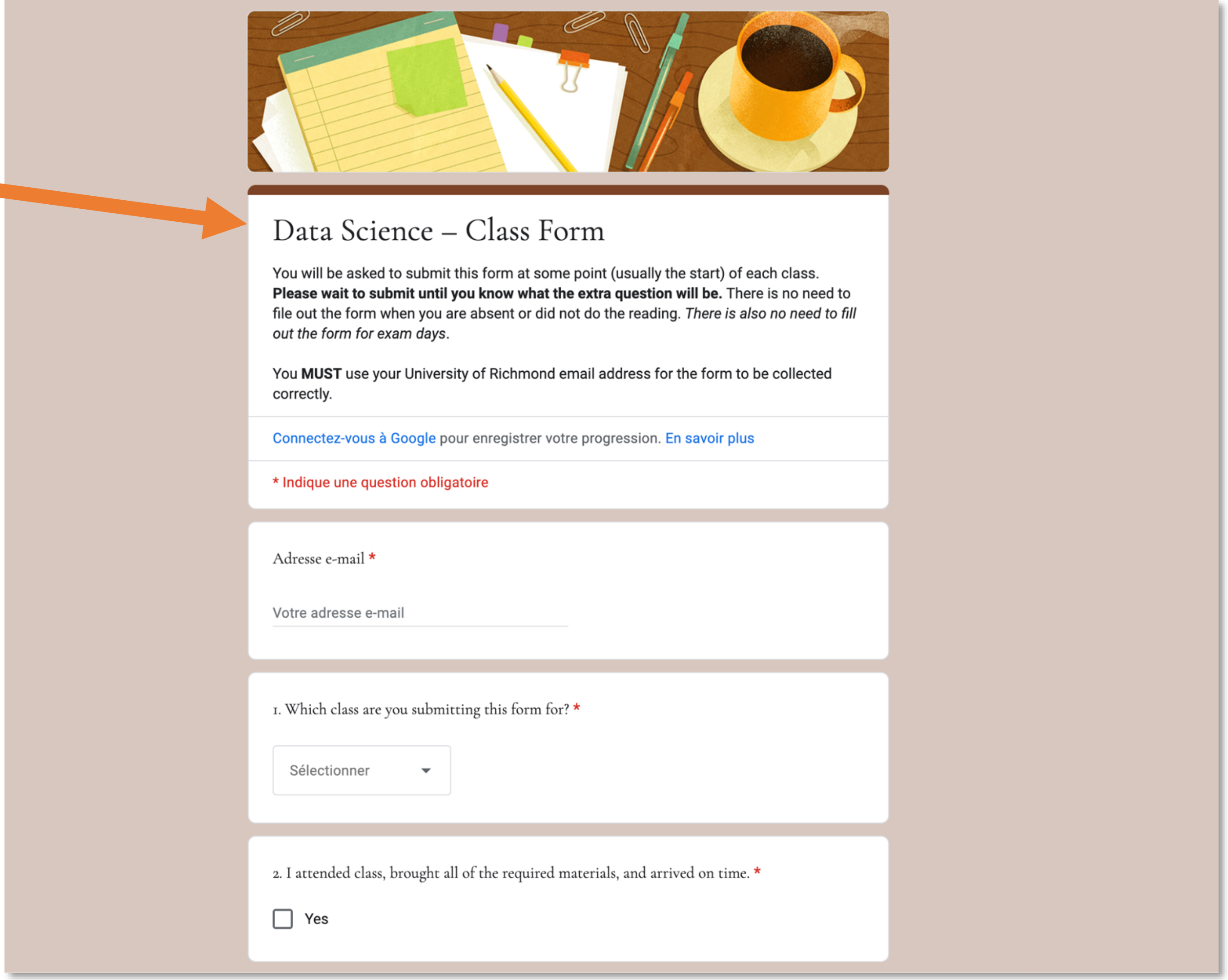
Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Course Structure

At the start of class, we will fill out the course form together. This is a form of checking attendance and sometimes doing a bit of data collection.

It will ask you whether you are in class, whether you've done the reading, and sometimes asks for some extra information.



Data Science – Class Form

You will be asked to submit this form at some point (usually the start) of each class. **Please wait to submit until you know what the extra question will be.** There is no need to file out the form when you are absent or did not do the reading. *There is also no need to fill out the form for exam days*.

You **MUST** use your University of Richmond email address for the form to be collected correctly.

Connectez-vous à Google pour enregistrer votre progression. En savoir plus

* Indique une question obligatoire

Adresse e-mail *

Votre adresse e-mail

1. Which class are you submitting this form for? *

Sélectionner ▾

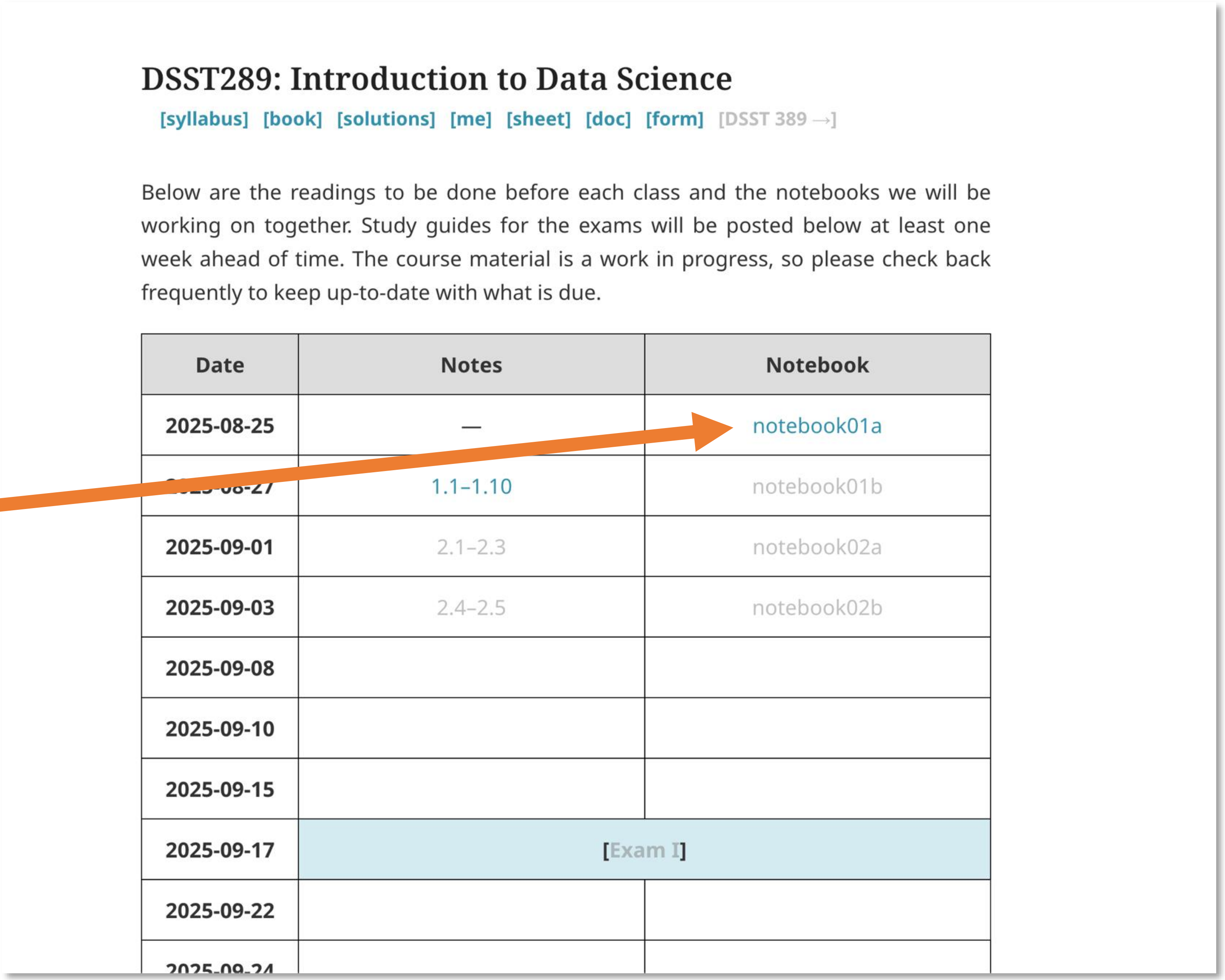
2. I attended class, brought all of the required materials, and arrived on time. *

□ Yes

Course Structure

During class, we will spend most of our time going through a **Notebook** (coding) together and/or doing some data collection.

These are neither submitted nor graded.



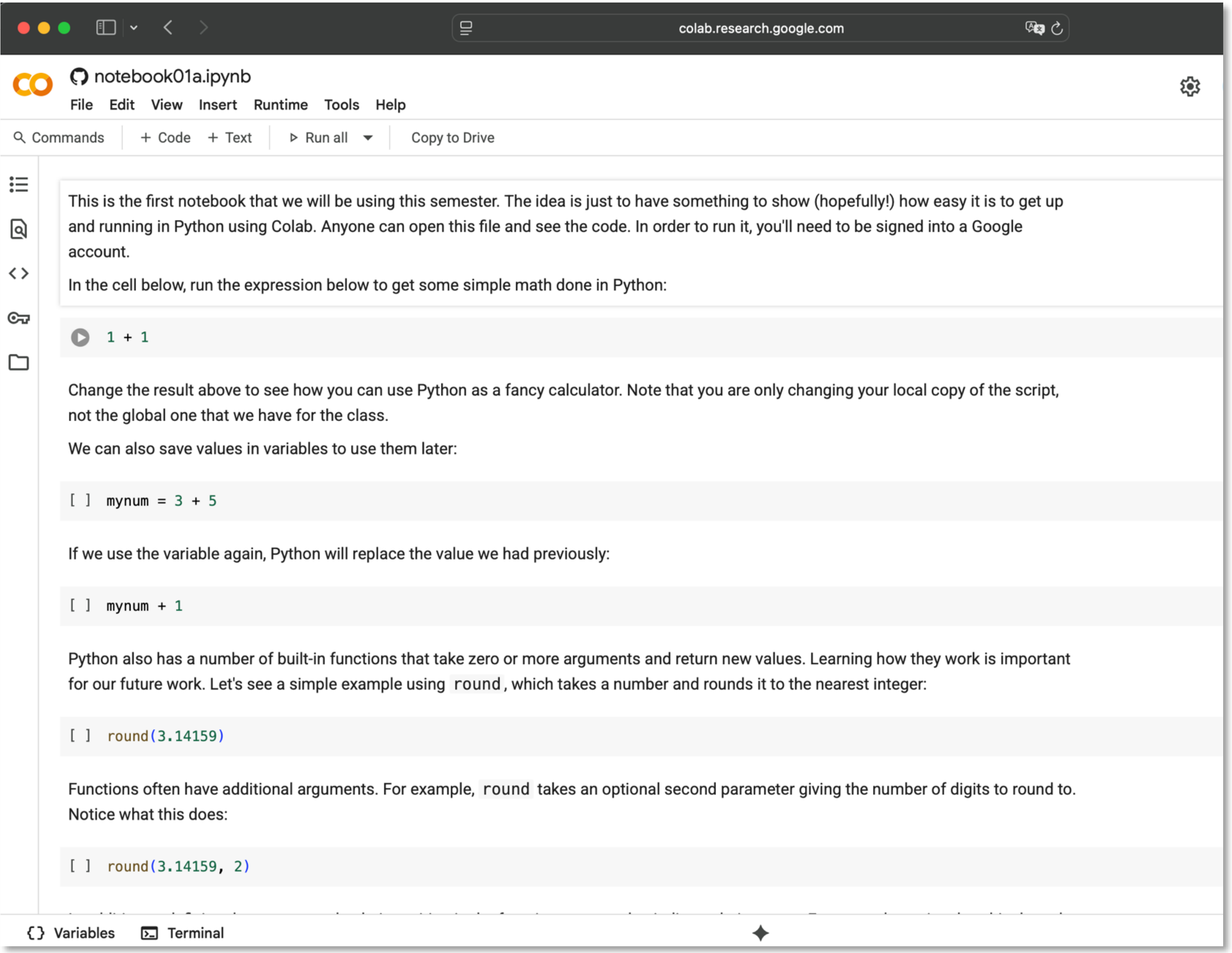
DSST289: Introduction to Data Science

[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Course Structure


colab.research.google.com
CO
notebook01a.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text ▷ Run all Copy to Drive
{} Variables
Terminal

This is the first notebook that we will be using this semester. The idea is just to have something to show (hopefully!) how easy it is to get up and running in Python using Colab. Anyone can open this file and see the code. In order to run it, you'll need to be signed into a Google account.

In the cell below, run the expression below to get some simple math done in Python:

```
1 + 1
```

Change the result above to see how you can use Python as a fancy calculator. Note that you are only changing your local copy of the script, not the global one that we have for the class.

We can also save values in variables to use them later:

```
[ ] mynum = 3 + 5
```

If we use the variable again, Python will replace the value we had previously:

```
[ ] mynum + 1
```

Python also has a number of built-in functions that take zero or more arguments and return new values. Learning how they work is important for our future work. Let's see a simple example using `round`, which takes a number and rounds it to the nearest integer:

```
[ ] round(3.14159)
```

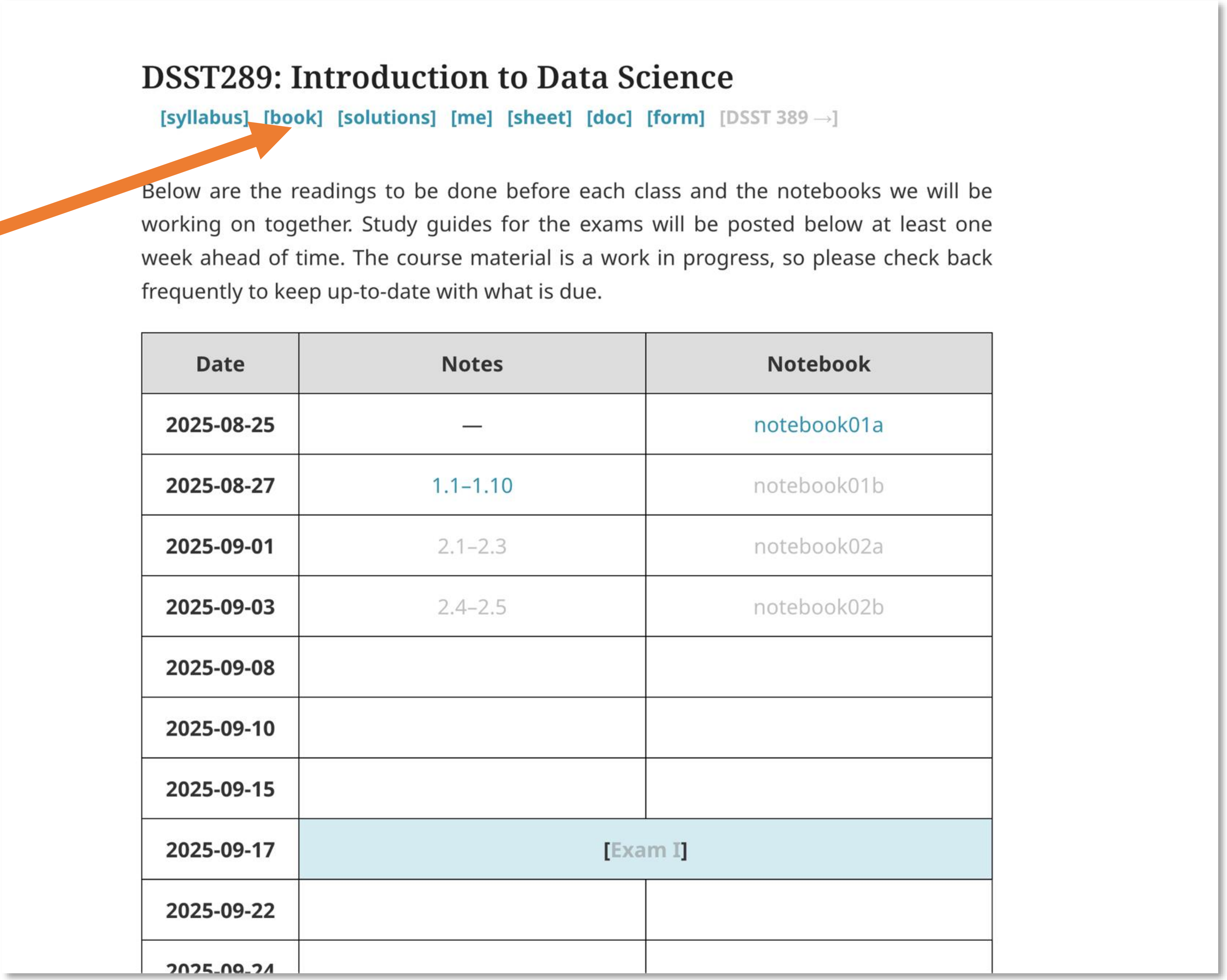
Functions often have additional arguments. For example, `round` takes an optional second parameter giving the number of digits to round to. Notice what this does:

```
[ ] round(3.14159, 2)
```

During class, we will spend most of our time going through a **Notebook** (coding) together and/or doing some data collection.

These are neither submitted nor graded.

Course Structure



DSST289: Introduction to Data Science

[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

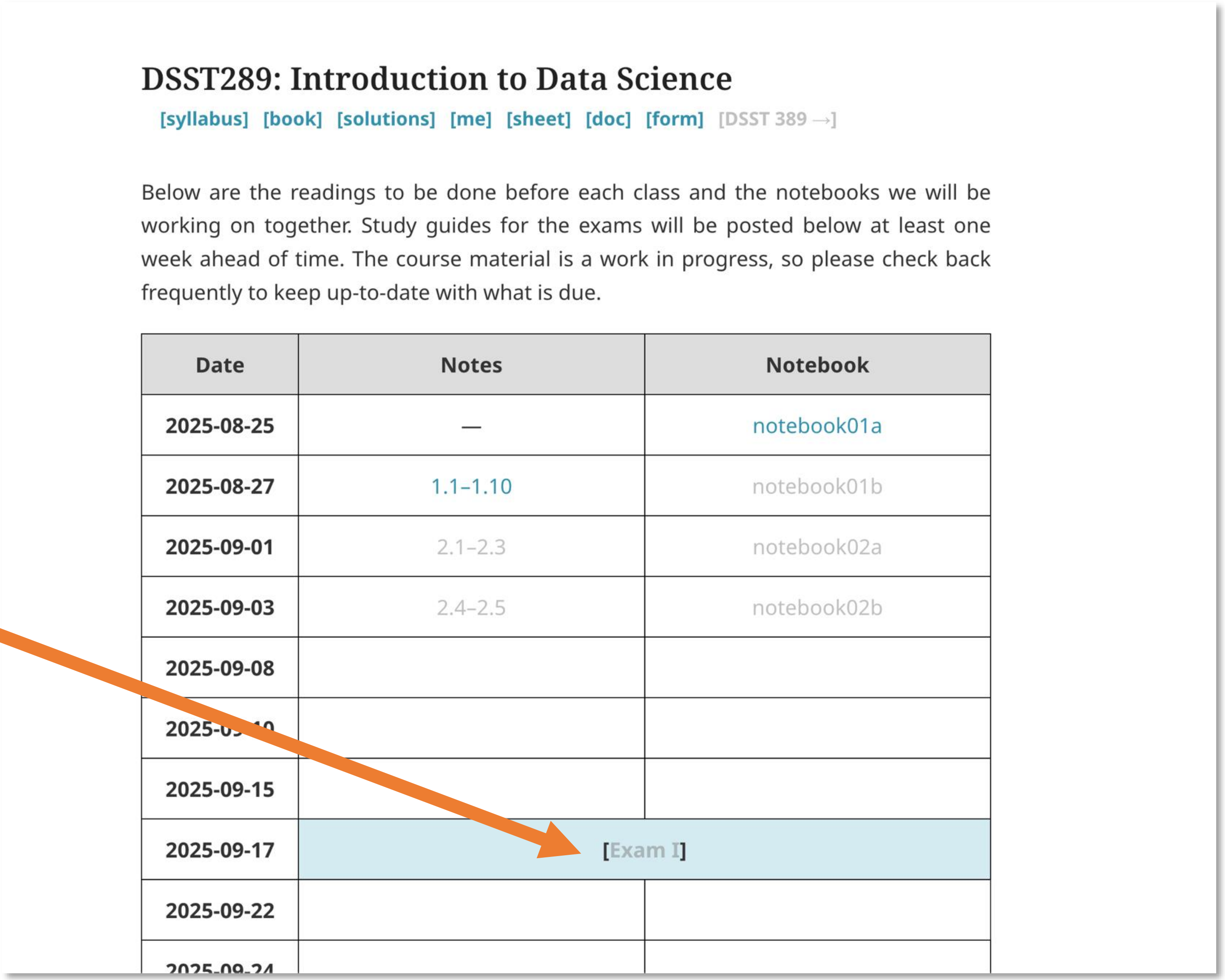
Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Solutions will be posted at the **solutions** page, linked to at the top of the page.

Course Structure

We will have four exams on the days listed on the course website (spread out more-or-less evenly through weeks 2-12).



DSST289: Introduction to Data Science

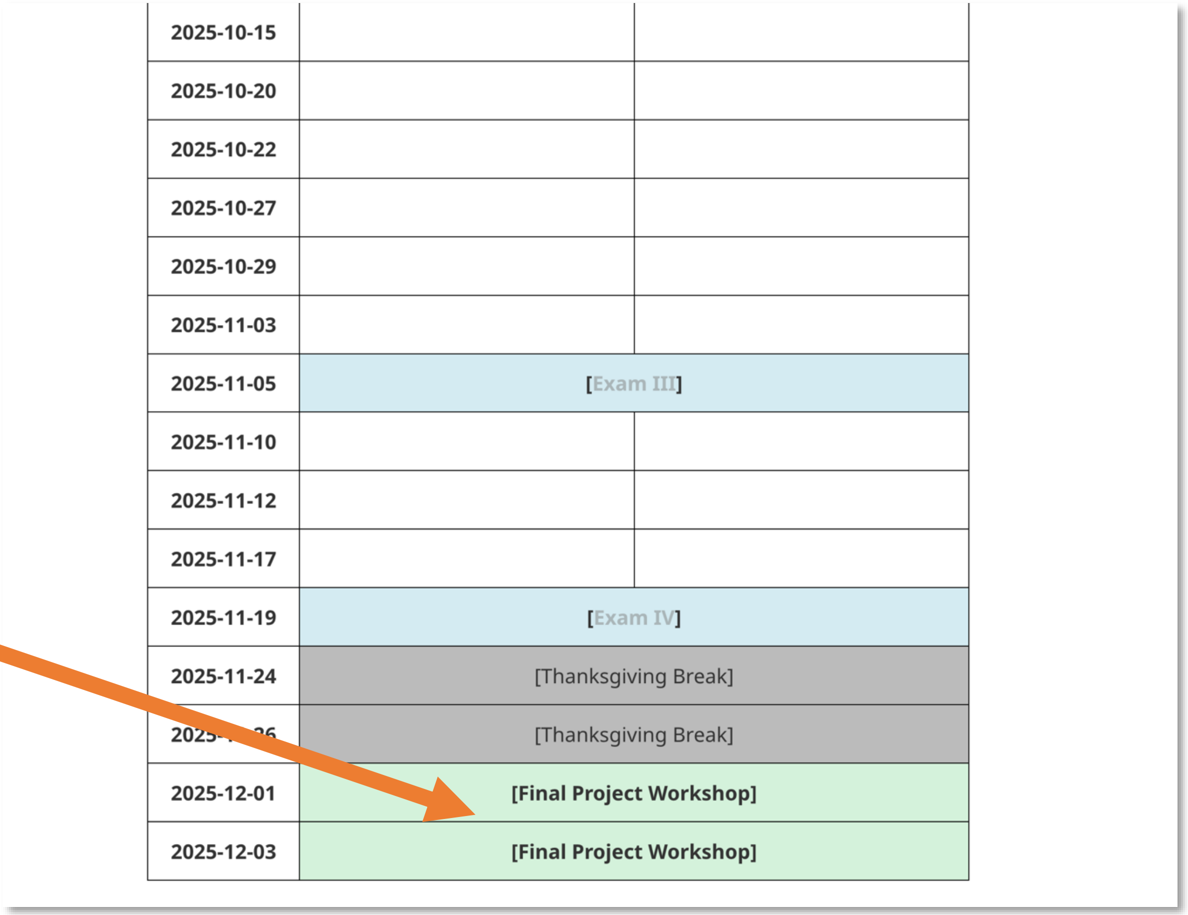
[syllabus] [book] [solutions] [me] [sheet] [doc] [form] [DSST 389 →]

Below are the readings to be done before each class and the notebooks we will be working on together. Study guides for the exams will be posted below at least one week ahead of time. The course material is a work in progress, so please check back frequently to keep up-to-date with what is due.

Date	Notes	Notebook
2025-08-25	—	notebook01a
2025-08-27	1.1–1.10	notebook01b
2025-09-01	2.1–2.3	notebook02a
2025-09-03	2.4–2.5	notebook02b
2025-09-08		
2025-09-10		
2025-09-15		
2025-09-17	[Exam I]	
2025-09-22		
2025-09-24		

Course Structure

In the final week of the semester, we will do a final project. More details on that soon.



2025-10-15		
2025-10-20		
2025-10-22		
2025-10-27		
2025-10-29		
2025-11-03		
2025-11-05	[Exam III]	
2025-11-10		
2025-11-12		
2025-11-17		
2025-11-19	[Exam IV]	
2025-11-24	[Thanksgiving Break]	
2025-11-26	[Thanksgiving Break]	
2025-12-01	**[Final Project Workshop]**	
2025-12-03	**[Final Project Workshop]**	

Grading

I try to keep grading simple. You’ll get a grade for each exam, your class forms (100% - 5% for each missing form beyond 2), and your final project. These six grades are averaged, and a letter grade is assigned as follows:

A (93–100), **A-** (90–92),
B+ (87–89), **B** (83–86), **B-** (80–82),
C+ (77–79), **C** (73–76), **C-** (70–72)
F (0–69)

Website Link

taylor-arnold.github.io/courses/dsst289-f25/

Generative AI

I assume you are getting a lot of different rules/advice about using generative AI in all of your classes. We are actually going to be studying it here, so prepare to hear a lot more!

For the actual work this semester, you are welcome to Gen AI in any way you would like outside of class. During class, refrain from the temptation to have the technology answer your questions for you. I shouldn't see any ChatGPT, Claude, Gemini, Grok … or similar running on your machine while we are working together in class unless there has been a specific instruction to do so.

4. Class Form

Class Form

For the extra question today:

What is your previous background with R/Python/programming/spreadsheets?

None is totally fine! And if you have a lot, no need for an exhaustive list. Just trying to get a sense for everyone's background

5. Our First Notebook

Python

A big change this semester is that I am transition this class from R to Python, for a number of inter-related reasons. The core content is the same, but Python let’s us do some extra things (machine-learning related, largely) that were tricky in R. It will also allow us to run code through Google Colab instead of requiring a setup on your machine. While we are using the Google product, everything you learn can be applied to any version of Python that you want to make use of.

Google colab

6. A Bit of Data

Recipe

Let’s do a little data work today to start thinking about the core questions of data science.

On a piece of paper, write down a recipe for a favorite dish of yours. Something moderately complicated (around 5-10 ingredients) is perfect.

Include the ingredients and instructions for how to make it.

Recipe

As you read the section for next class, consider what it would take to create a tabular version of your recipe as a structured dataset. Consider how we could create a large dataset of hundreds or thousands of these.

What if we needed a way to do things such as detect recipes that are vegetarian, vegan, kosher, gluten free? What if we wanted to calculate how to scale up/down you version? Convert from/to metric? Figure out what things you could make with a certain ingredient?

Think about all of these things as a concrete example as you read for Wednesday's class!