Foundations of Data Science

CMSE 830

Attendance Sign In

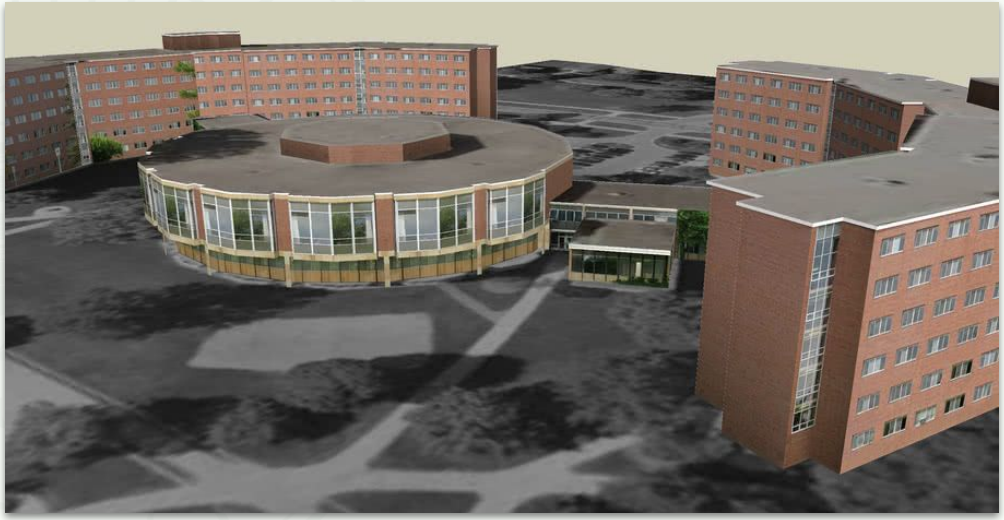
Prof. Murillo
Computational Mathematics, Science and Engineering
Michigan State University

Plan For Today

1. Discuss syllabus and course organization
    a. please read the syllabus in detail!
2. Lecture: What is Data Science?

Course Basics

- Tue-Thu
- 8:30 - 9:50am ET
- 27 Aug 2024 – 13 Dec 2024
- Wonders Hall C211
- 3 credits
- In Person
- Instructor: Prof. Murillo
- TA: Max Gregg



Organization of Course Content

There are two communication channels:

D2L™

- assignments
- course content
  - slides
  - PDF notes
- grades

- all other communication

PLEASE DO NOT USE EMAIL!

Schedule

Holidays/Breaks 2024-2025	
Holiday - University Closed	Monday, 9/2/24
Fall Break	Monday, 10/21/24 - Tuesday, 10/22/24
Holiday - University Closed	Thursday, 11/28/24 - Friday, 11/29/24
Holiday - University Closed	Tuesday, 12/24/24 - Wednesday, 1/1/25
Holiday - University Closed	Monday, 1/20/25
Classes Not Held	Thursday, 2/13/25
Spring Break	Sunday, 3/2/25 - Sunday, 3/9/25
Holiday - University Closed	Monday, 5/26/25
Holiday - University Open, Classes Not Held	Thursday, 6/19/25
Holiday - University Closed	Friday, 7/4/25

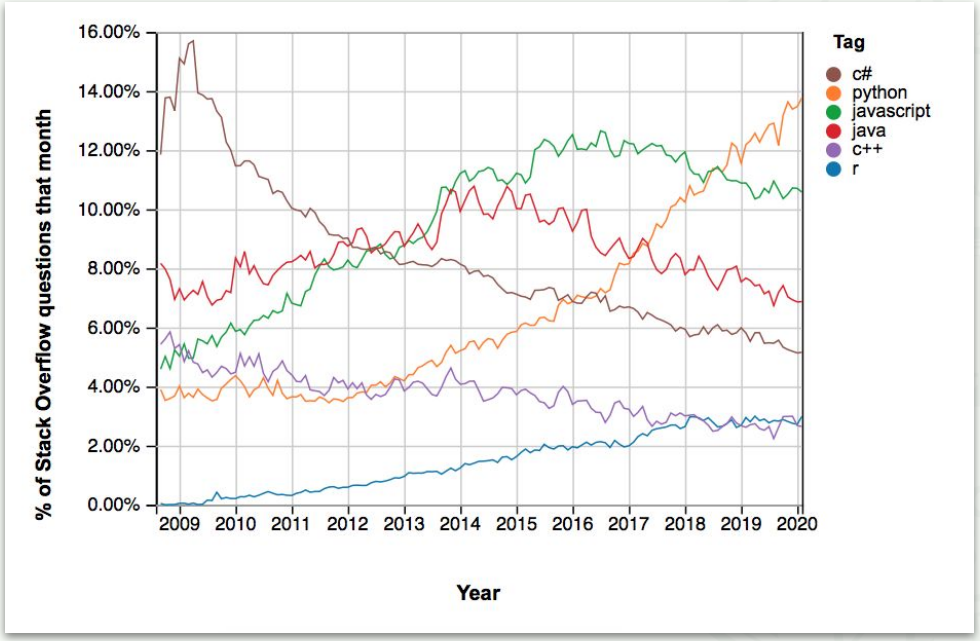
Week	Chapter	Content	Considerations
Early Weeks (1-5): Foundational Data Handling			
1-2	1	Data Types, Quality, and Cleaning	Covers all aspects of data quality and cleaning.
3	2	Visualization	Introduces visualization tools and techniques.
4	3	IDA and EDA	Deeper analysis techniques following visualization.
5	4	Missingness	Types of missing data, mechanisms causing missingness.
Mid Course (6-9): Complex Mathematical Methods and Imputation			
6	5	Imputation	Techniques for handling missing data, including NN, stochastic regression.
7-8	6	Linear Algebra (incl. SVD and PCA)	Foundational mathematical concepts applied in data science.
9	7	Structure of Data (incl. TDA and MDS)	Introduces topological data analysis and multidimensional scaling.
Advanced Topics (10-14): Specialized Data Types with Practical Applications			
10	8	Time Series Analysis	Specialized techniques for time-dependent data.
11	9	Geospatial Data	Handling and analysis of spatial data sets.
12	10	Graph Data	Focus on network and graph-based data analysis.
13	11	Image Data	Processing and analysis of image data.
14	12	Text Data (Natural Language Processing)	Introduction to NLP and its applications in data science.
Final Weeks (15-16): Review, Capstone Project, and Additional Topics			
15	13	Ethics, Security, and Privacy	Discusses ethical implications, data security, and privacy laws.
16	14	Review and Capstone Project	Comprehensive review and a real-world data science project.

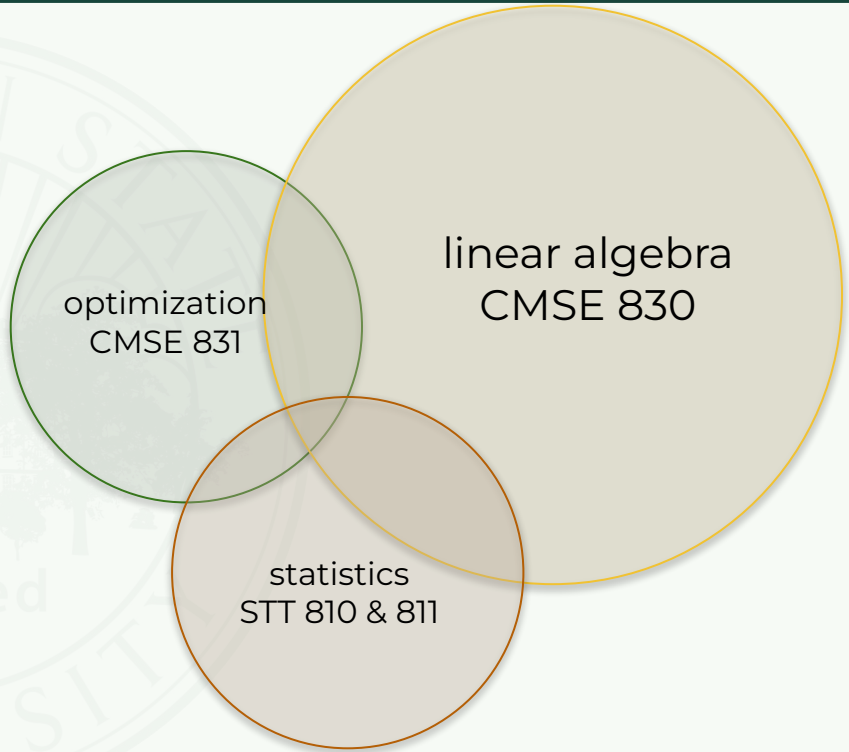
Office Hours

- **Prof. Murillo**
  - Tuesday and Thursday 10:00 – 11:00 in EGR 2501
- **Max**
  - Friday 1:00 – 3:00 in EGR 1503 (small CMSE CR)

Office hours are also available upon request.

Course Logic





Projects

- There are two projects.
- Projects are done individually, not in groups.
- Due in the middle of the semester (“midterm”) and end of the semester (“final”).
- Project 1: Data Science
  - write a deployable DS dashboard
- Project 2: Advanced Analysis (e.g., Machine Learning)
  - developed optimized models for prediction

This course has no exams! The projects count for 30% of your grade.

Projects and Homeworks

Projects *are done outside of class.*

However, I will guide you through the HWs.

HWs *will be partially:*

- *normal practice material*
- *new concepts*
- *pre-class material for in-class projects*
- *moving your projects along*

Note:

There is a document that covers the details of the projects. See D2L and Teams.

Laptop Setup

- Be sure to have your laptop with you on Thu.
- Ensure that it is charged.
- Bring all connectors/cables you might need.

Put Python on your laptop.

We will use Jupyter notebooks.

I recommend using the Anaconda Python installation.

MSU Honor Code

The Spartan Code of Honor Academic Pledge:

"As a Spartan, I will strive to uphold values of the highest ethical standard. I will practice honesty in my work, foster honesty in my peers, and take pride in knowing that honor in ownership is worth more than grades. I will carry these values beyond my time as a student at Michigan State University, continuing the endeavor to build personal integrity in all that I do."

Use of AI

There is are two documents – *see D2L and Teams* – with detailed instructions on how AI can be used.

In summary: *please use it!*

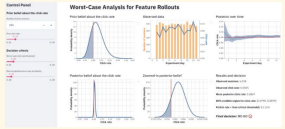
But, be extremely careful! And, read the AI document so that you follow the rules very carefully.

Documents in D2L and Teams

CMSE 830 Projects

This course is project based with two projects: you can think of them as your "midterm" and "final", since there are no exams.

The projects serve several goals. One of the goals is to ensure that you have something to show at the end of this course. To do this you will write dashboard web applications that are accessible from your personal GitHub account. We will all use Streamlit so that we can share our ideas and skills. In addition to learning to build web apps, share ideas and code through a repository, you will also learn excellent communication skills. You can link to your app from your CV and use your CMSE 830 projects in future job interviews (and your friends and family!).



Nearly all of the work on the projects is done *by you outside of class*. You know already that you will be tempted to procrastinate, so now is the time to think about how to trick yourself so that you don't do that. I will help you by encouraging you to make steady progress in your homeworks. I highly recommend using a calendar and putting milestones for this course into it with early reminders before due dates.

FAQ

Let's quickly cover some of the most obvious questions you have about the projects.

AI Policy for CMSE 830

Table of Contents

Introduction

As we navigate the rapidly evolving landscape of artificial intelligence (AI) in educatio
guidelines for its use in this course. AI tools, such as large language models and spec
become increasingly accessible and powerful. While these tools offer significant poter
experience, they also present challenges that need to be carefully considered.

Learn Smarter with AI: Your Personal Guide to Enhanced Education

Prof. Murillo
Computational Mathematics, Science and Engineering
Michigan State University

Foundations of Data Science: Homework 1

Welcome to your first homework assignment! This assignment will help you set up the necessary tools and familiarize yourself with course materials essential for your success in this Data Science course.

Objectives

By completing this homework, you will:

- Familiarize yourself with course expectations and policies
- Set up essential tools for the course
- Assess your current Python knowledge
- Begin engaging with course content

Foundations of Data Science — Fall 2024

CMSE 830-001

Foundations of Data Science

If you find any mistakes in this syllabus, please let me know.

Location and Time

- Wonders Hall C211 (note that this changed recently!)
- In Person (there is no hybrid or Zoom option)
- Tuesday and Thursday 8:30am – 9:50am ET
- 8/26/2024 – 12/13/2024
- Teams channel (primary method of communication)

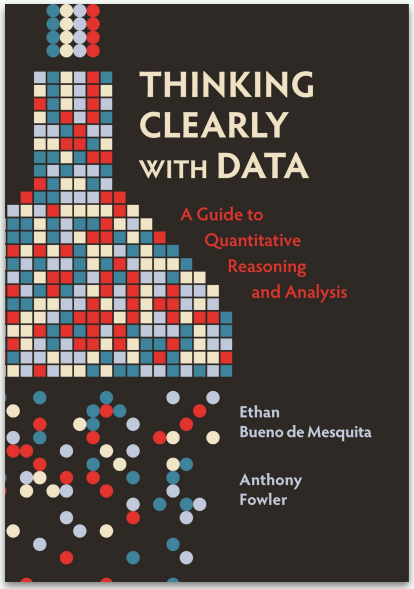
Textbooks

There are two textbooks for this class.

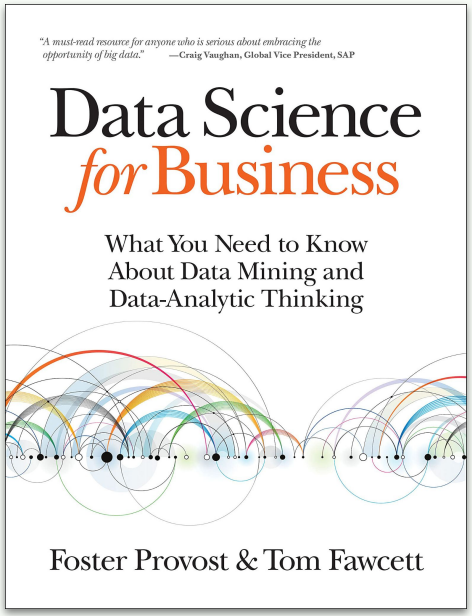
I will provide notes, in PDF form, that will amount to an informal text.

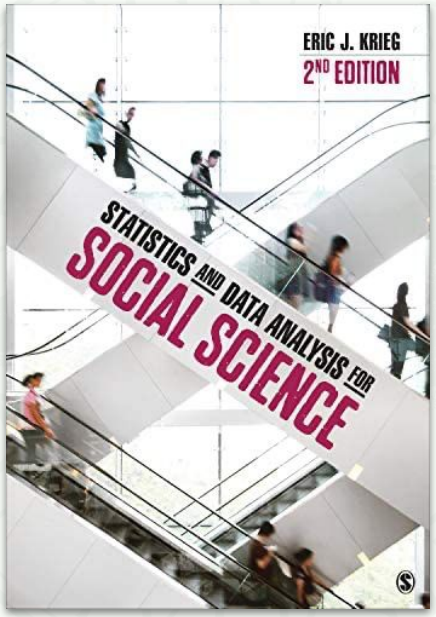
The PDF will be updated every week: you can always find it in D2L and Teams.

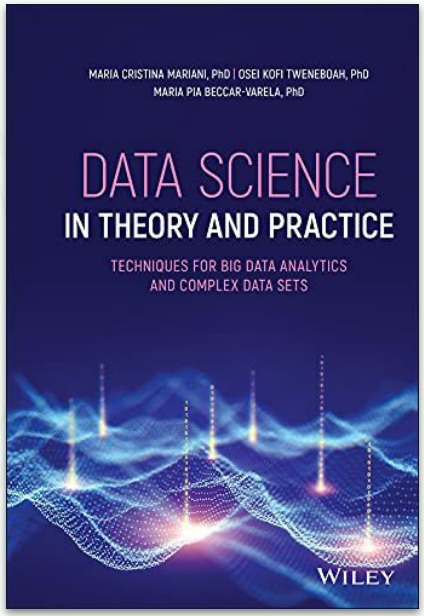
However, there are several books that I highly recommend (and some that I don't!).



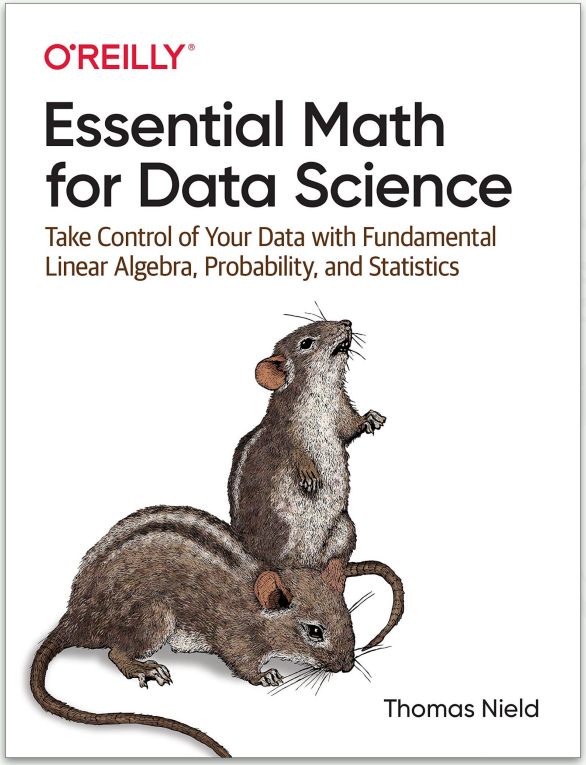
Textbooks I Don't Recommend

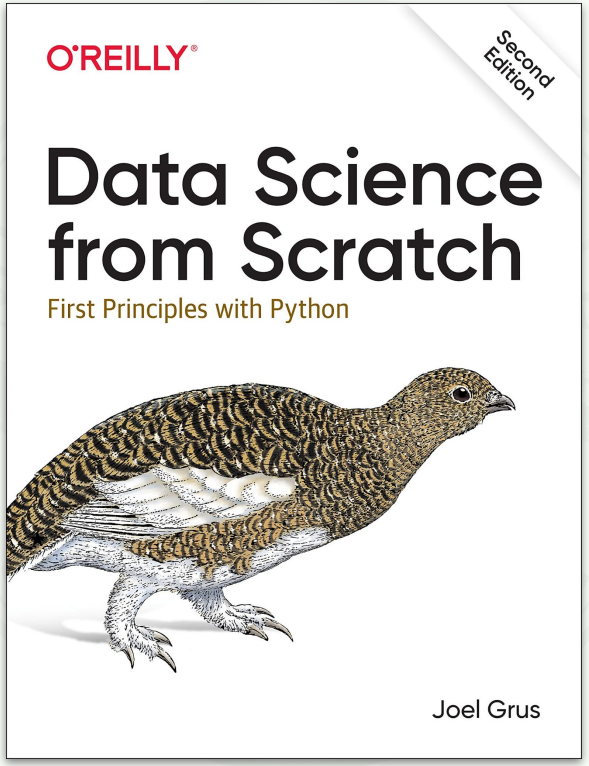


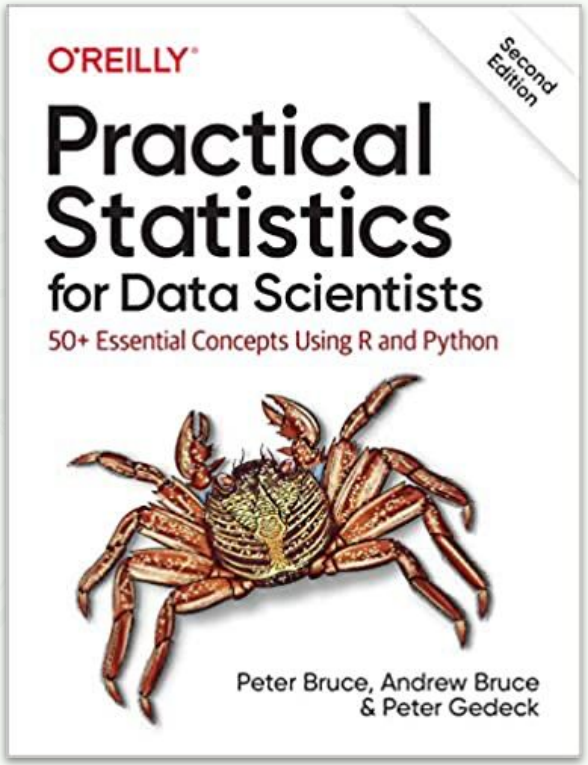




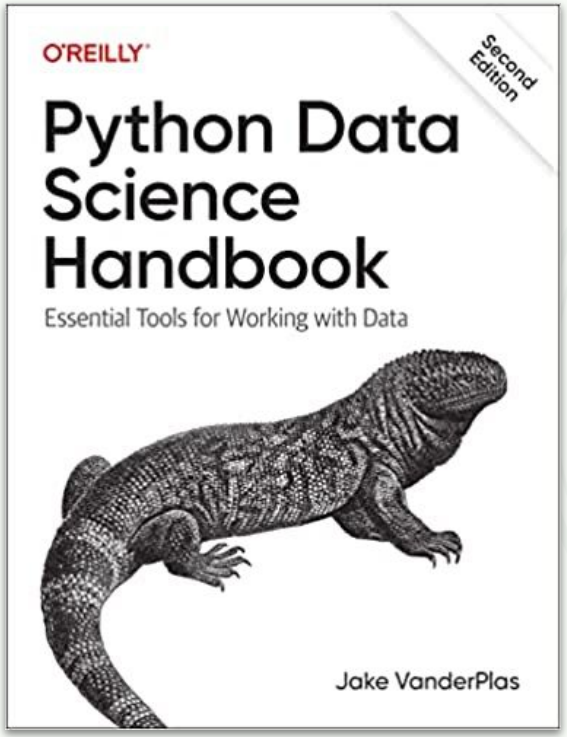
More Textbooks I Don't Recommend

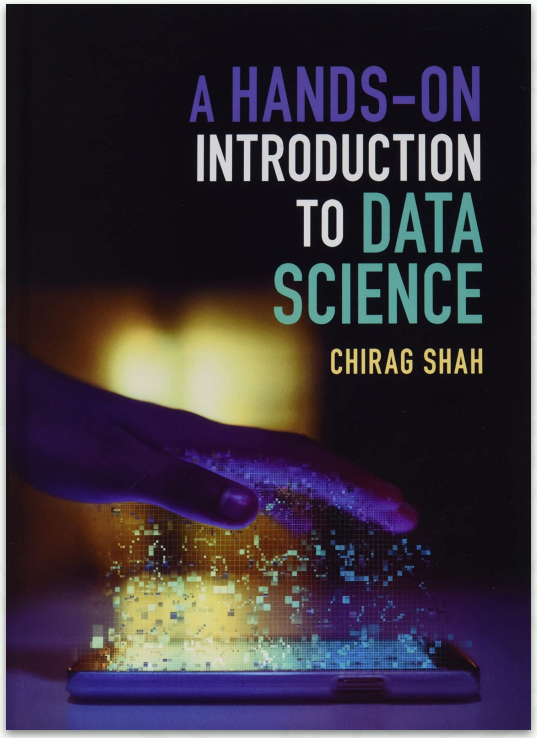


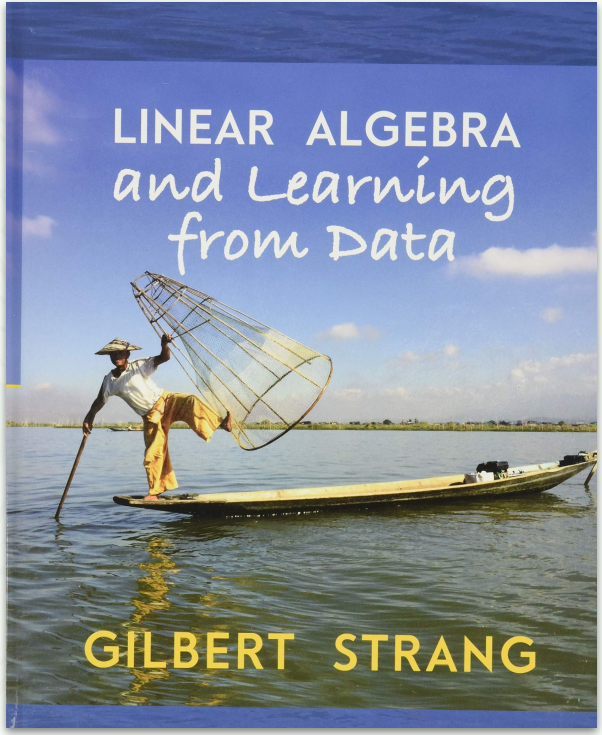




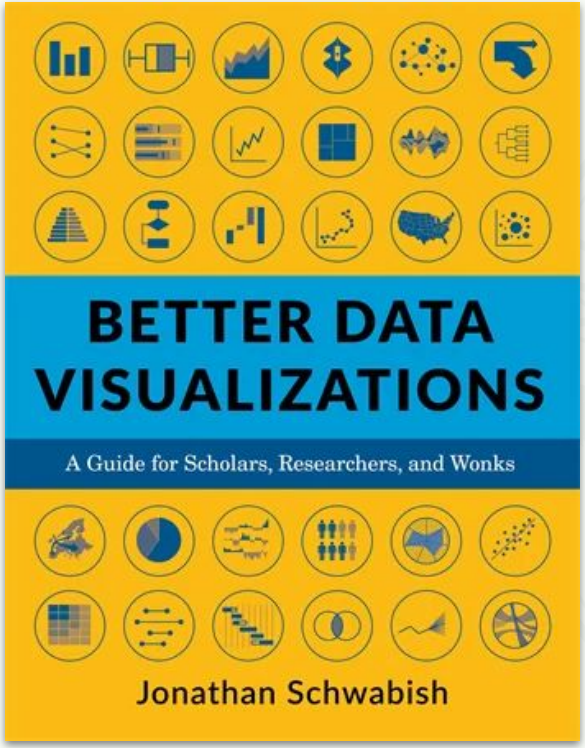
Textbooks That I Do Recommend

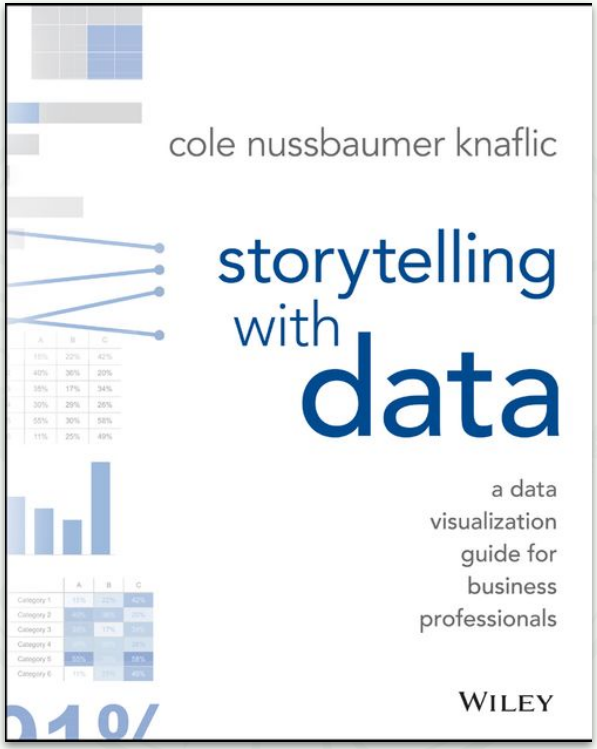


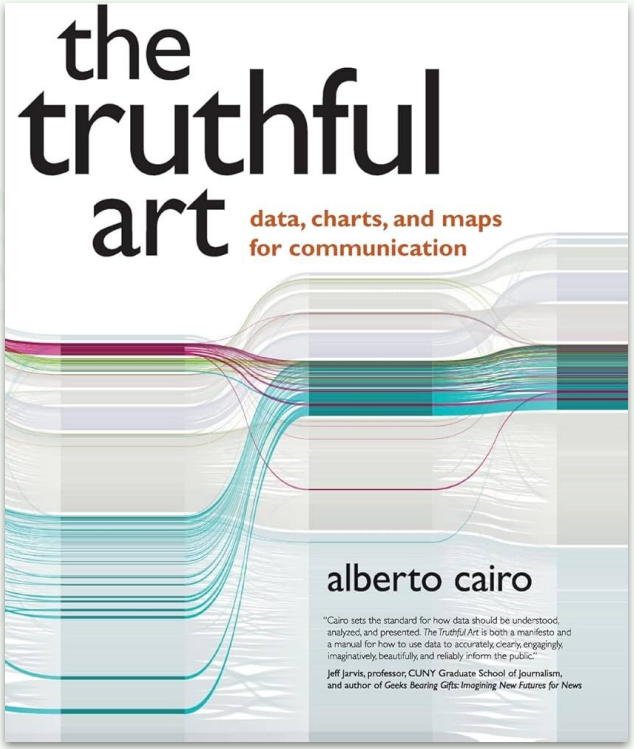




Textbooks That I Do Recommend: Visualization



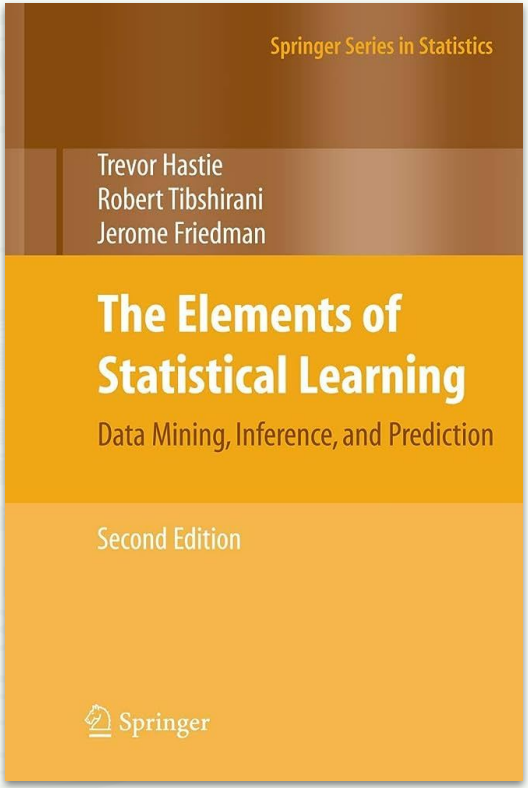




Textbooks That I Do Recommend

This is the best book overall for the mathematical details.

And, much of this applies to machine learning.



Course Organization

The general organization of the course is:

Much of the learning will be done through HW assignments, all of which are Jupyter notebooks.

There will be a HW every week due Sunday at midnight.

This pattern will be disrupted several times because of holidays and fall break.

Transition into Lecture....

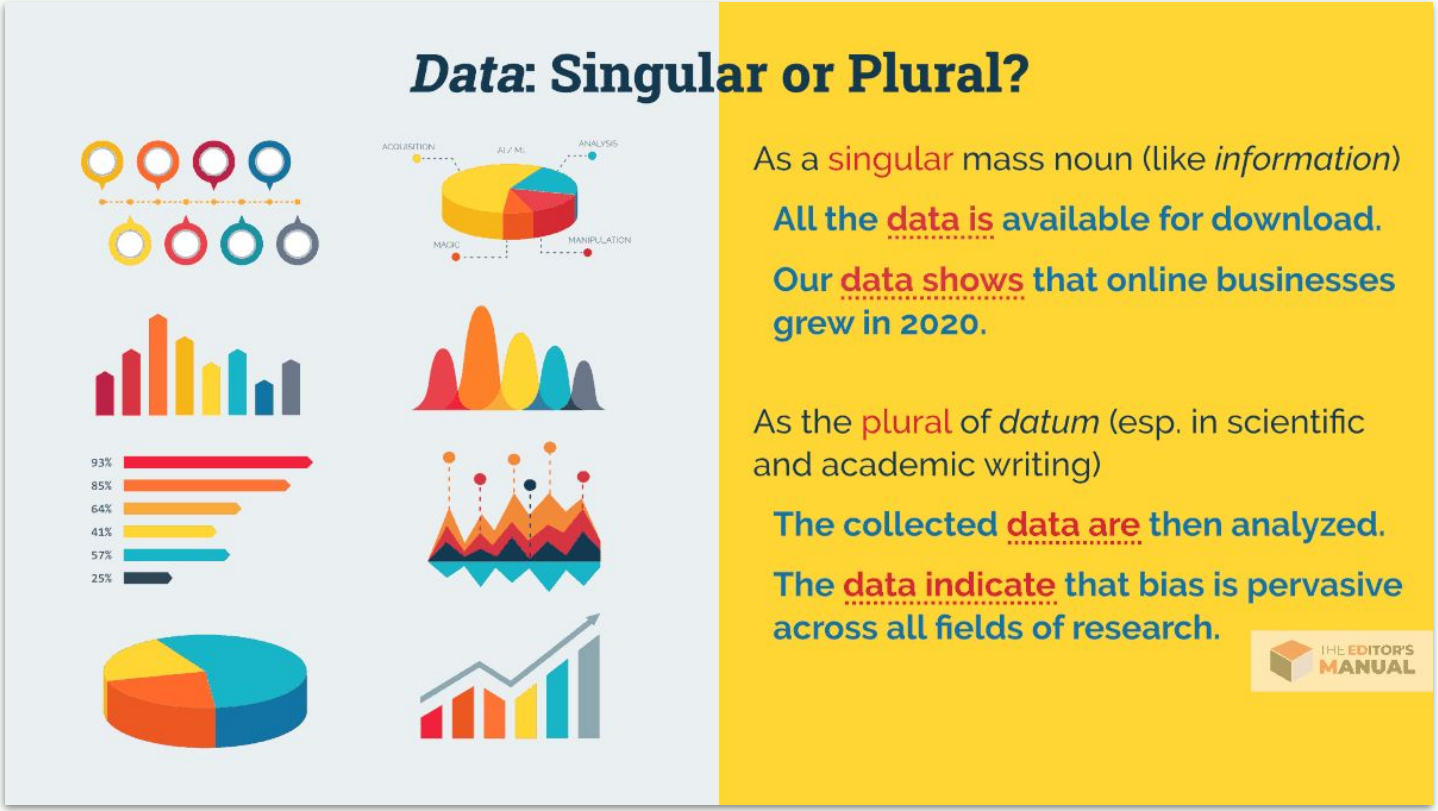
How is Data Defined?

`datum`: a piece of information

`data`: plural of datum

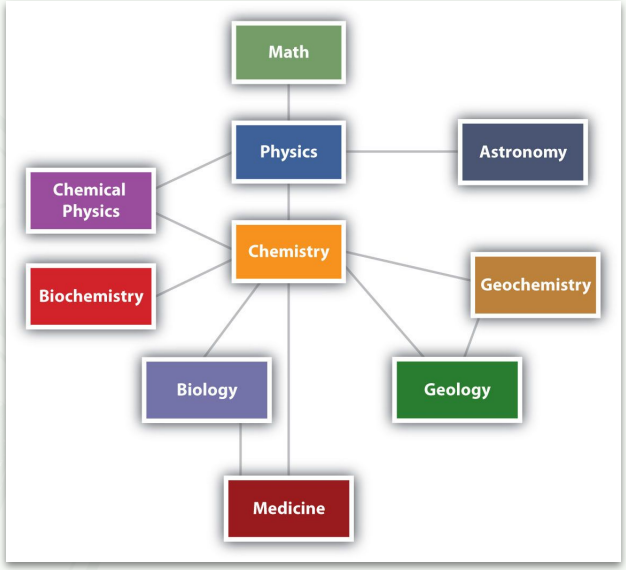
1. Today, almost no one uses the term “datum”. Rather, data is used for both the singular and the plural.
2. It gets messier: sometimes “data” is used as a *mass noun*, so you might hear both:
   a. “…data were…”
   b. “…data was…”
3. Does this mean that “data science” is the same as “information science”?

How is Data Defined?



How is Science Defined?

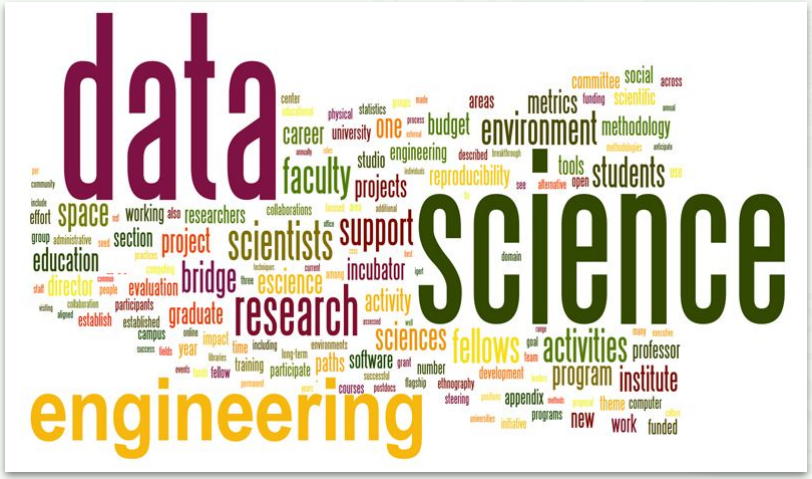
`science`: the systematic study of the structure and behavior of the physical and natural world through observation, experimentation, and the testing of theories against the evidence obtained



Probably data science is not science!

How is Science Alternatively Defined?

`science`: a systematically organized body of knowledge on a particular subject



Is Data Science “Just” Information Science?

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data,[1][2] and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.[3]

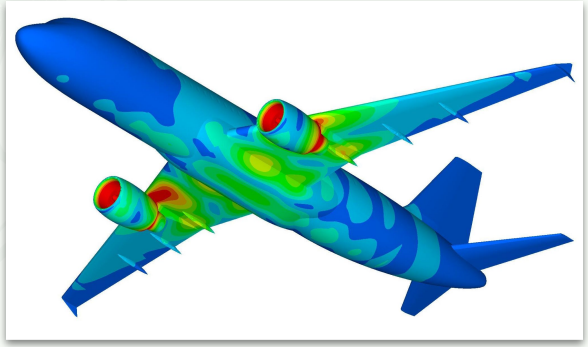
What is Data Science (DS)?

Solving problems using data.

1. How is DS different from solving problems other ways?
2. What is data?

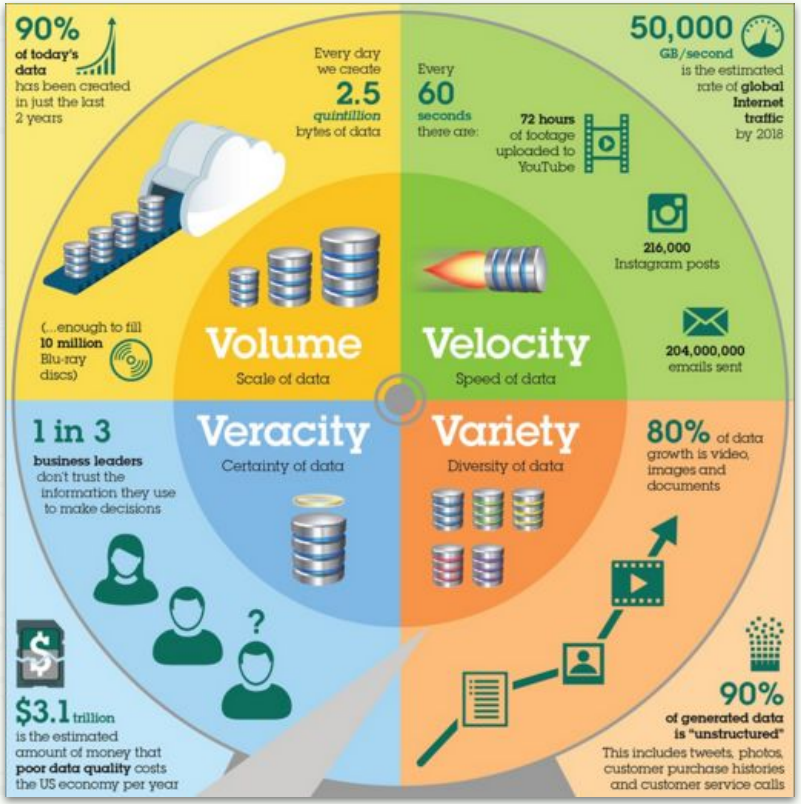
$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} + f(t, x)$$

data: the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.

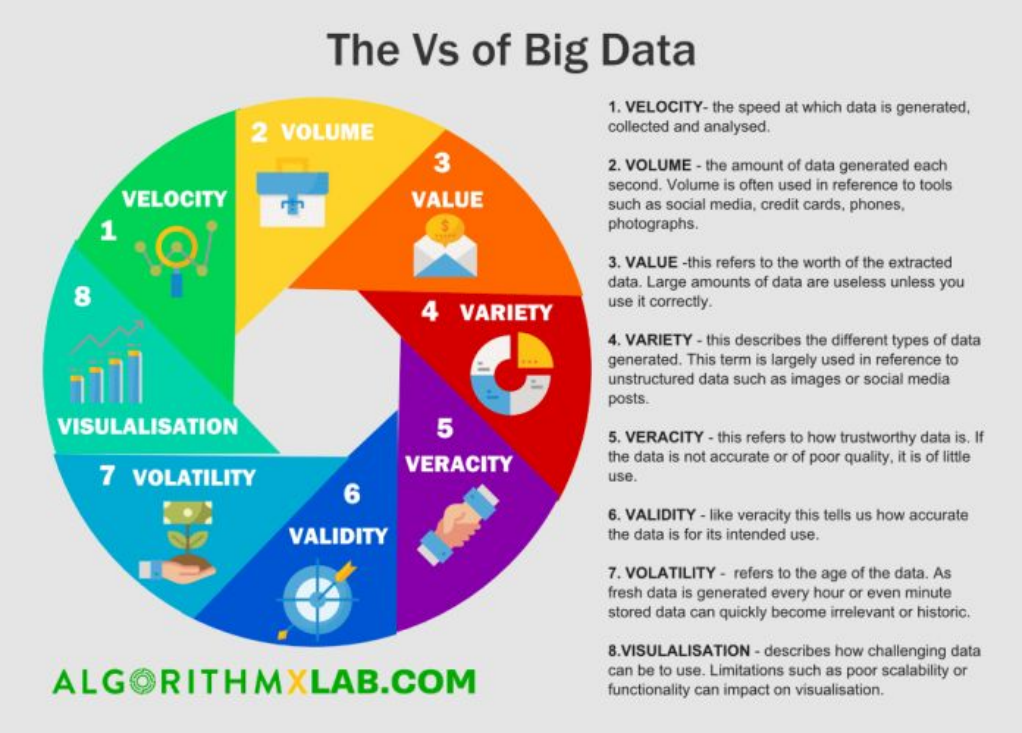


What is Big Data? The Four-V's of Big Data

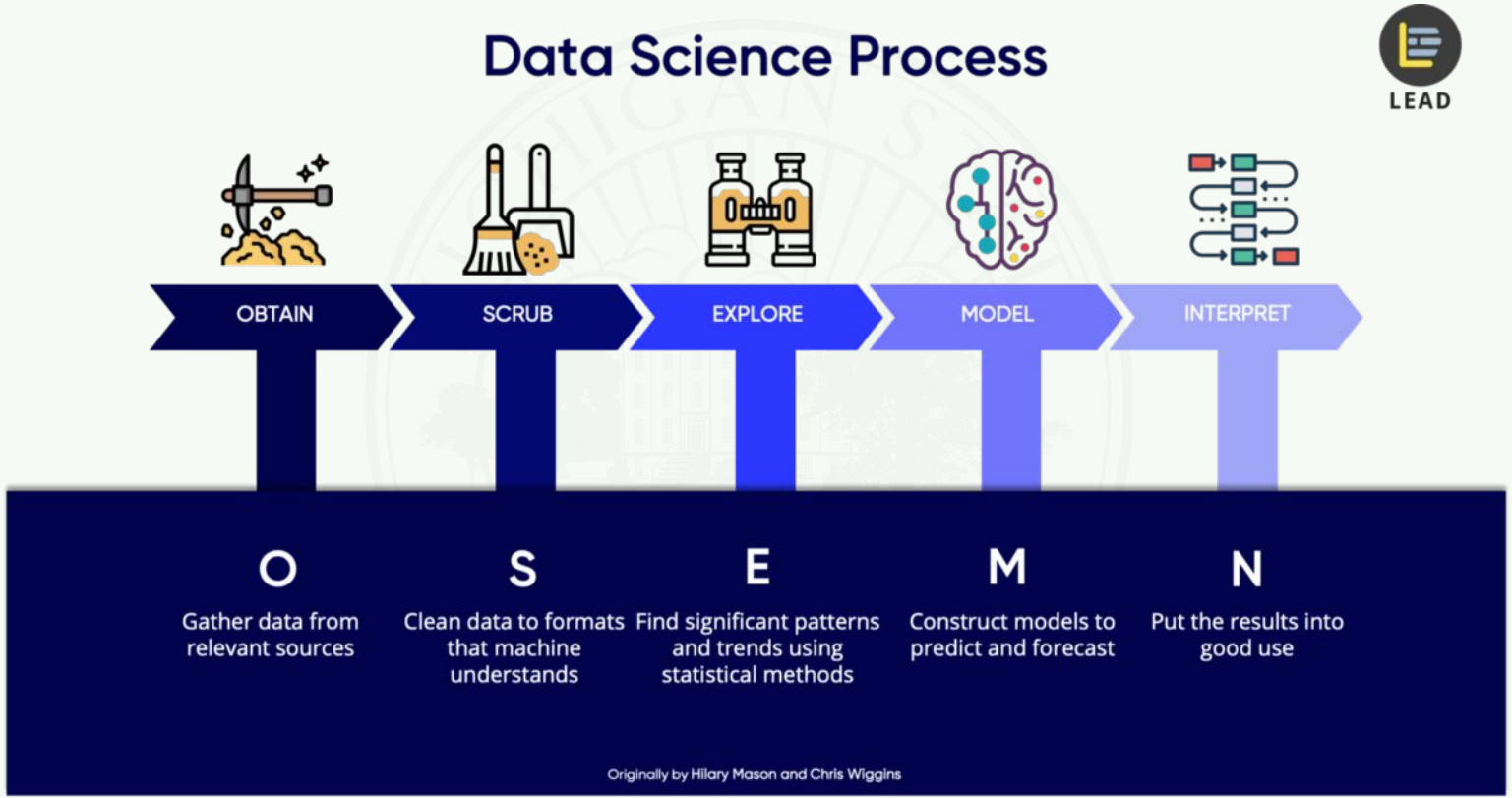
1. Volume
2. Velocity
3. Veracity
4. Variety



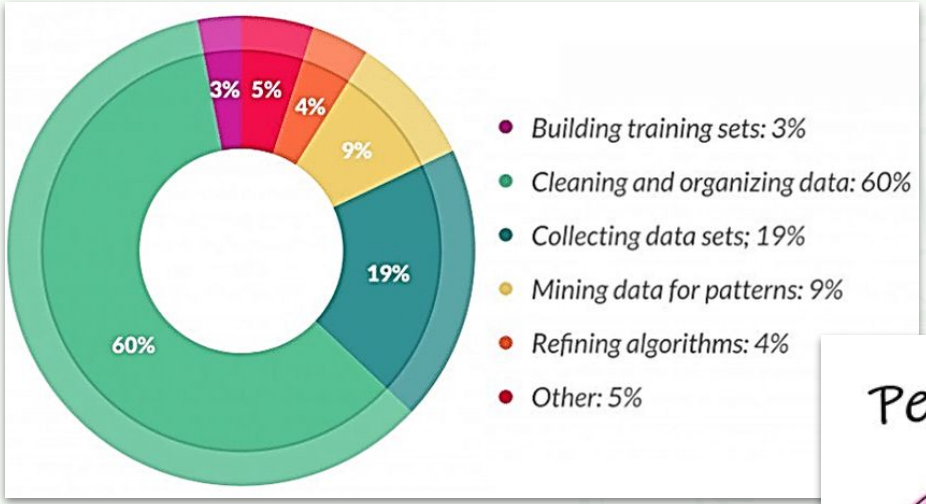
What is Big Data? The Eight-V's of Big Data

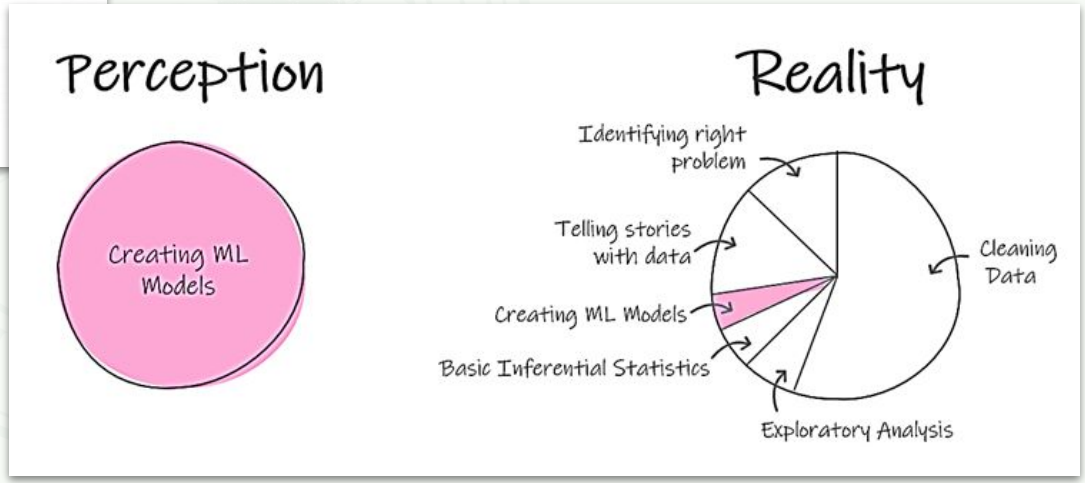


What is the Data Science Process/Workflow?

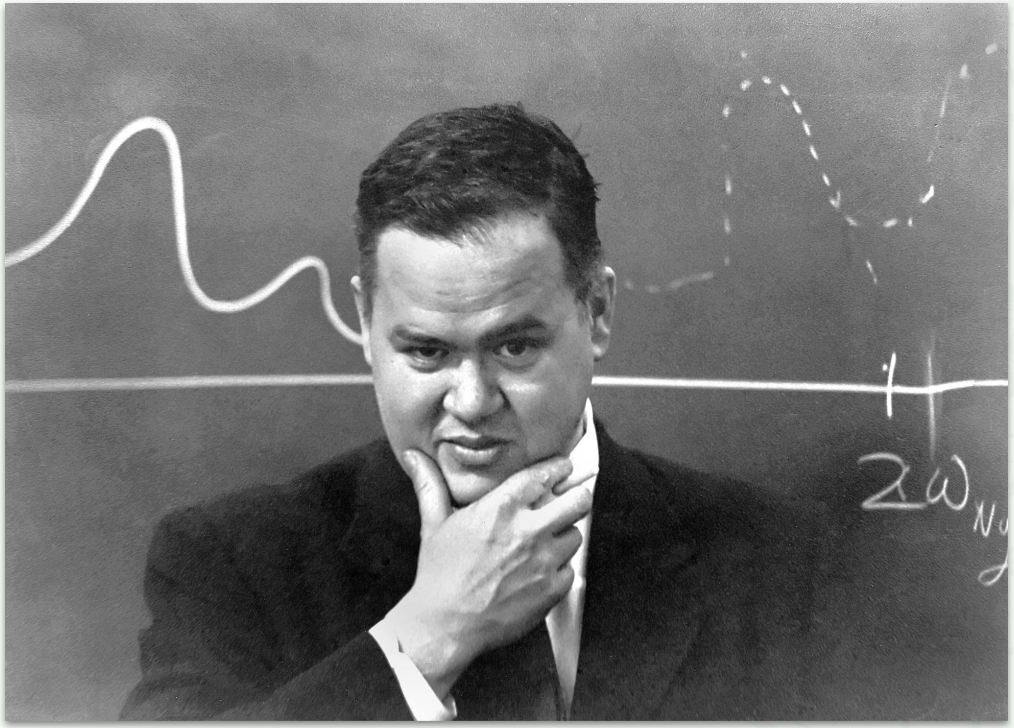


Where is the time spent?





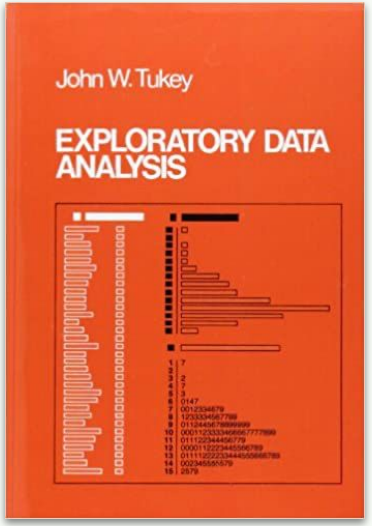
John W. Tukey



Tukey gave us important terms:

- software
- bit
- data analysis

Tukey is the originator of Exploratory Data Analysis (EDA).



Exploratory Data Analysis (EDA)

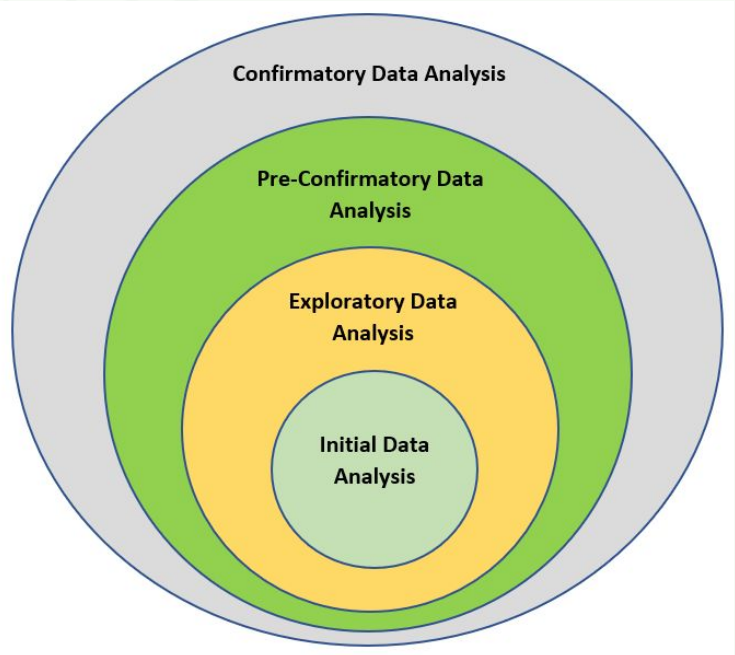
EDA is the analysis of datasets to summarize their main characteristics.

What message is in the data?!

Initial Data Analysis (IDA) and Confirmatory Data Analysis (CDA)

IDA is the **initial** analysis of a dataset that precedes EDA to determine the quality of the dataset.

CDA is the **final** analysis of a dataset in which your hypotheses are tested.



Let’s Look For a Message in a Dataset

Types of Integers

real: the datum is a “float”

cardinal: count (how many?)

ordinal: position

nominal: name

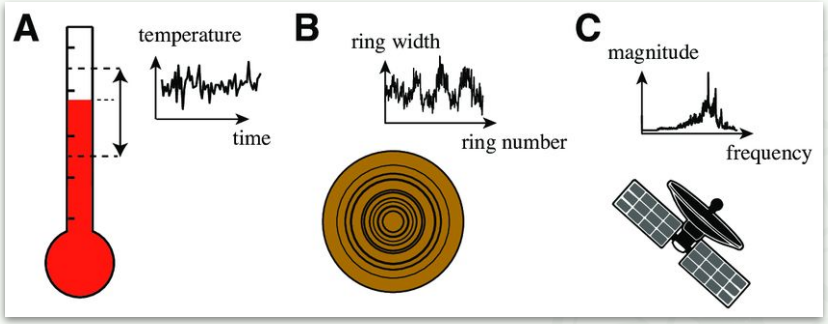
Data, Better Defined

Data is a collection of discrete objects:

- numbers
- words
- facts
- objects
- measurements
- observations
- descriptions
- and so on....

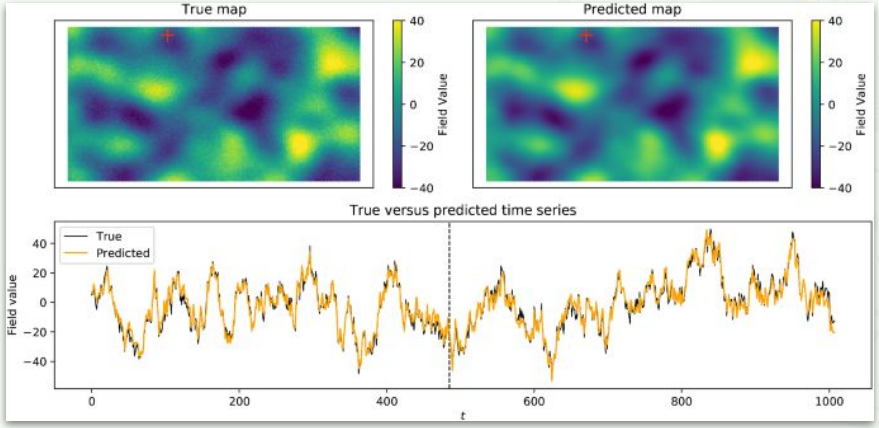
*"This is the **data** we were looking for."*

Data Contexts



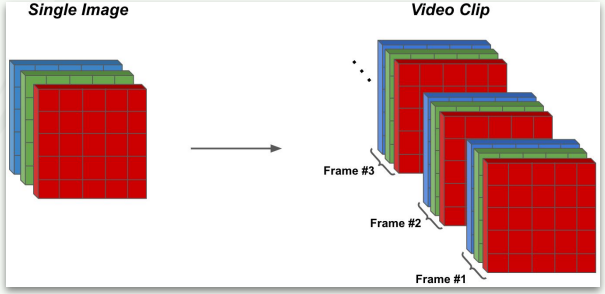
sequential

structured



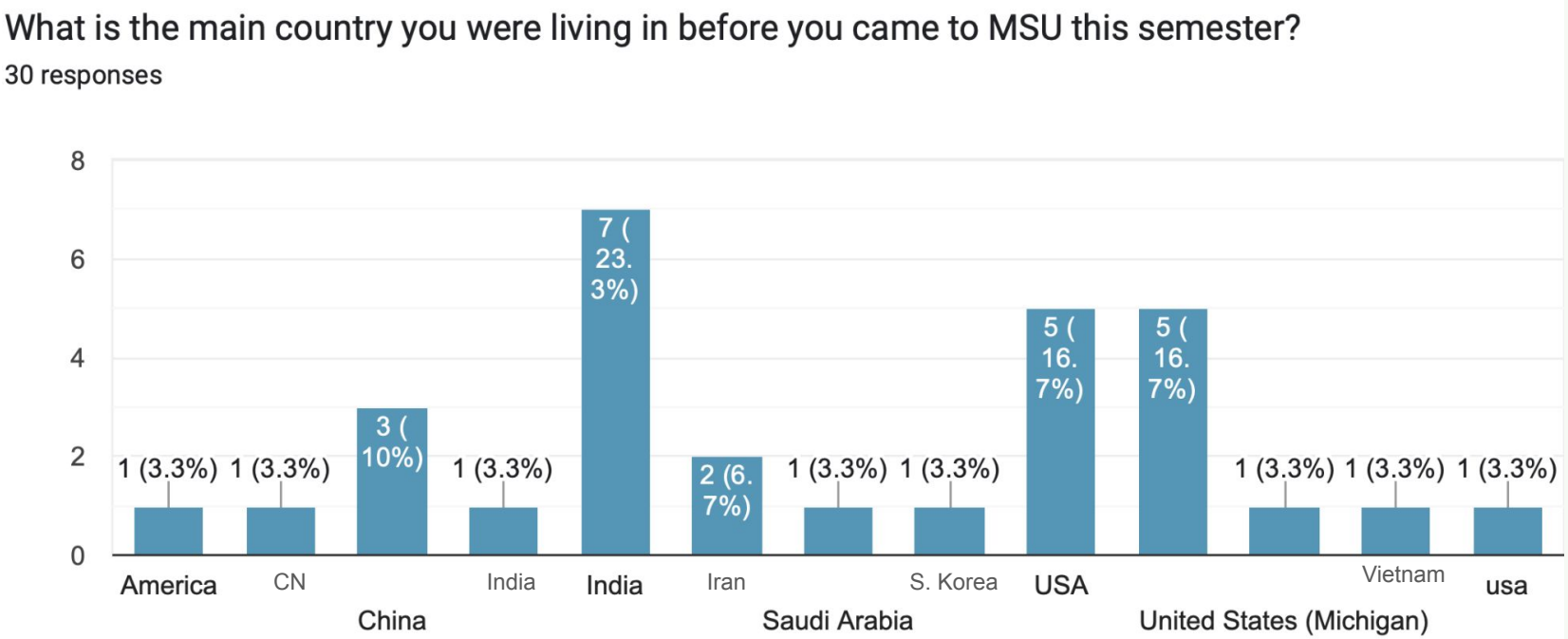
spatiotemporal

image/video



Another Data Set: You

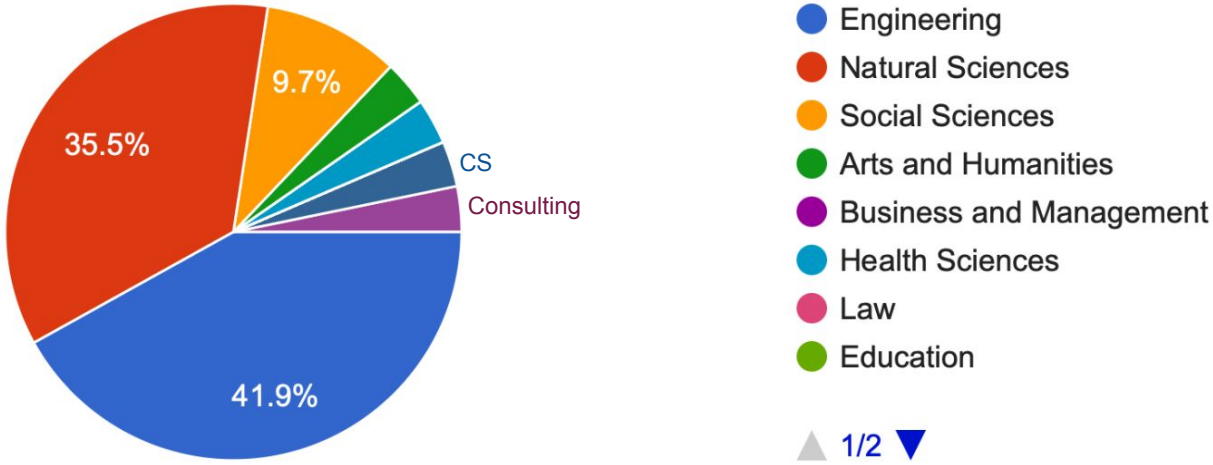
We Are a Diverse Group!



From Many Backgrounds!

What is the closest area of your background before taking this course?

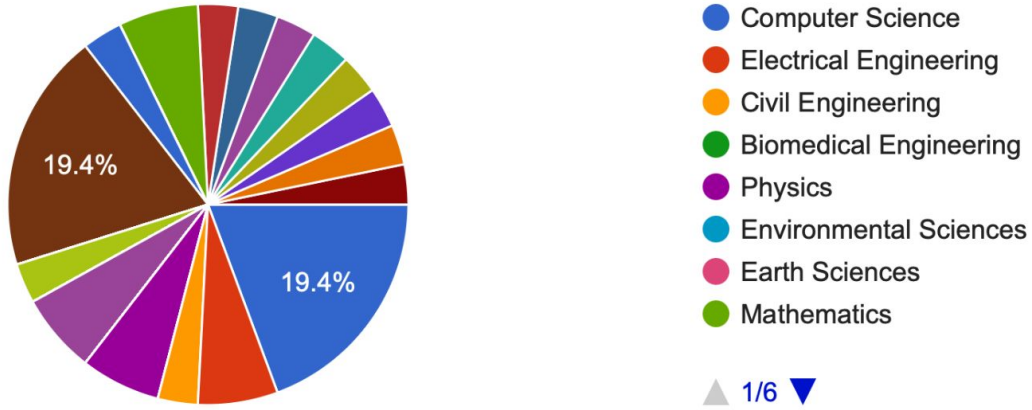
31 responses



What Specific Background Do You Have?

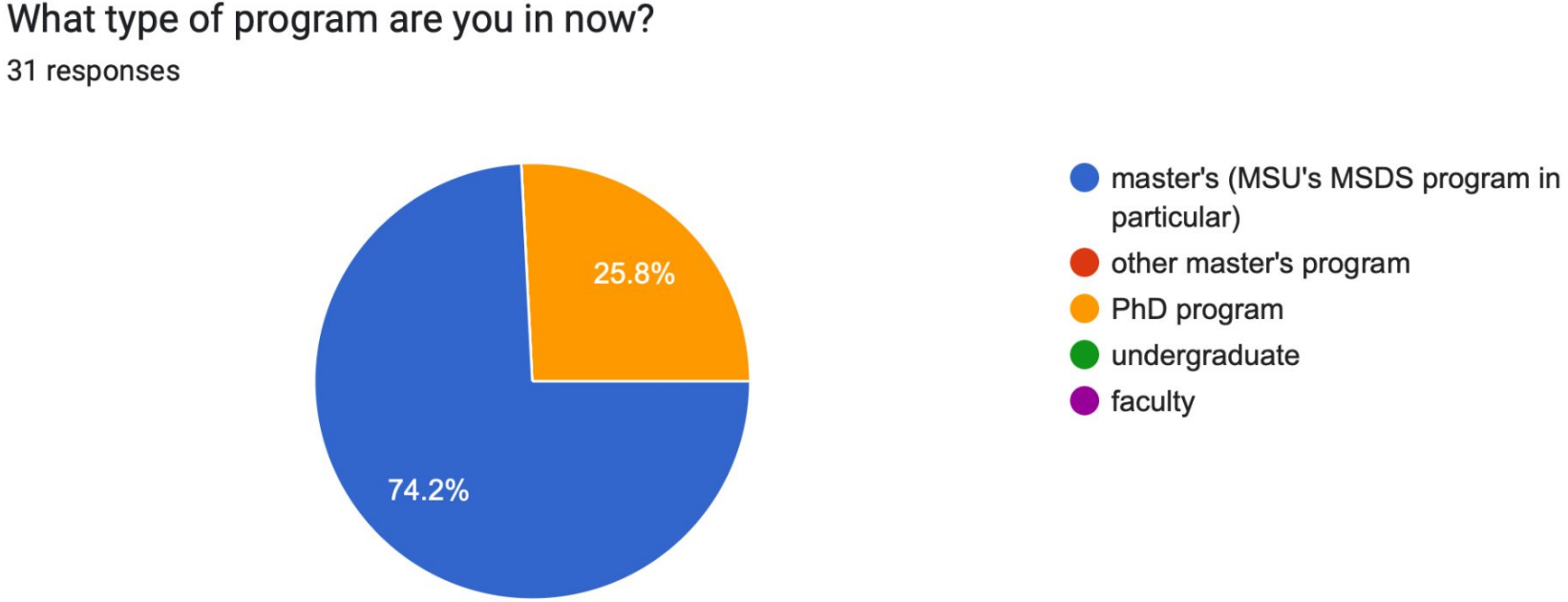
Let's resolve the previous question further. Again, which is the single closest area of your background?

31 responses

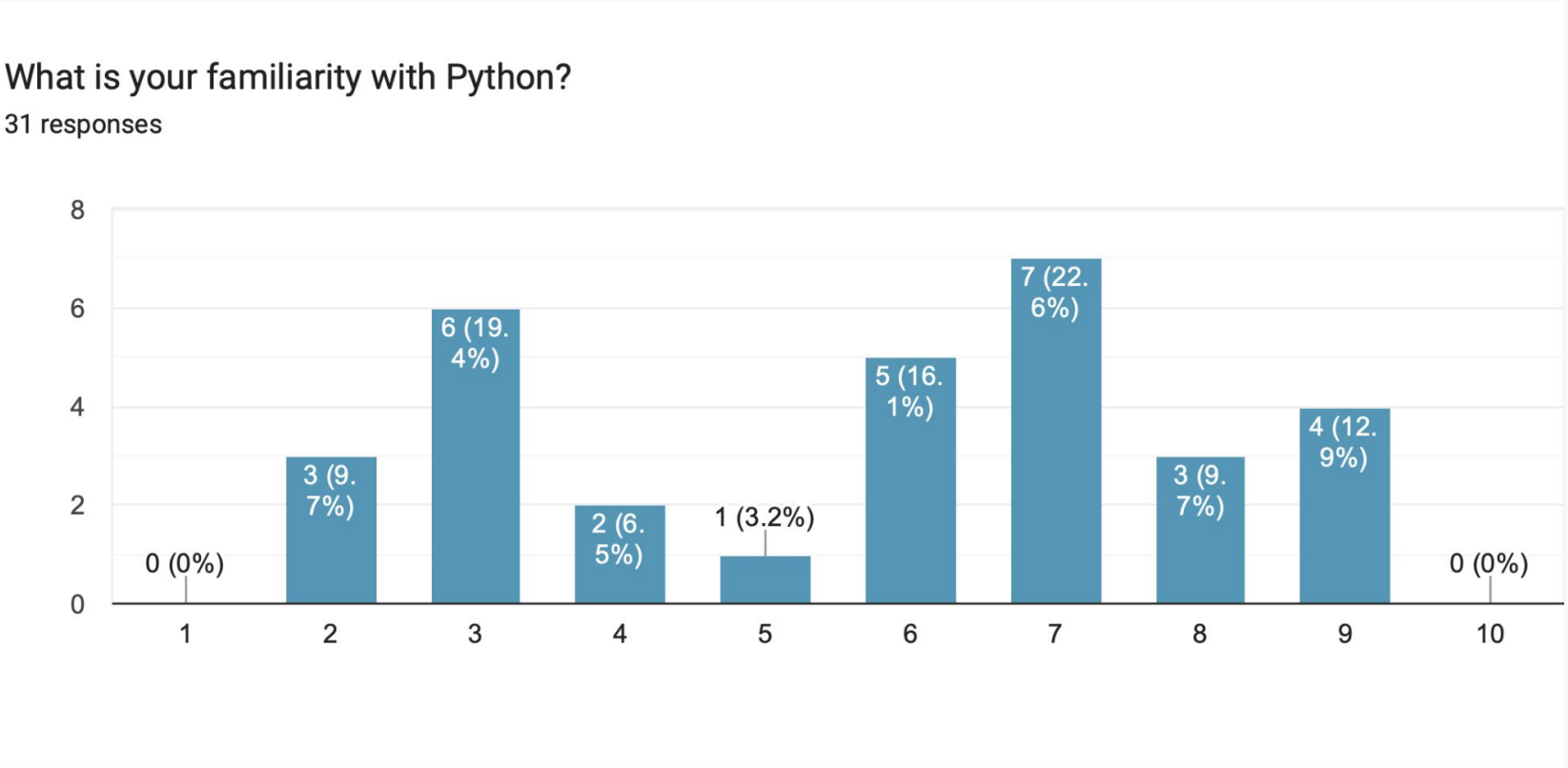


Data Science
Data Science
Psychology
Electrical Engineering
materials science
Psychology
Mechanical Engineering
Data Science
Mechanical Engineering
Information Technology
Civil Engineering
Computer Science
International Affairs
mechanical engineering
Cell and Molecular Biology
Agriculuture / Horticulture / Plant Breeding
Computer Science
Biology
Data Science
Data Science
Computer Science
Physics
Architecture
Computer Science
Electronics and Communication Engineering
Computer Science
Physics
Electrical Engineering
Computer Science
Medicine
Data Science

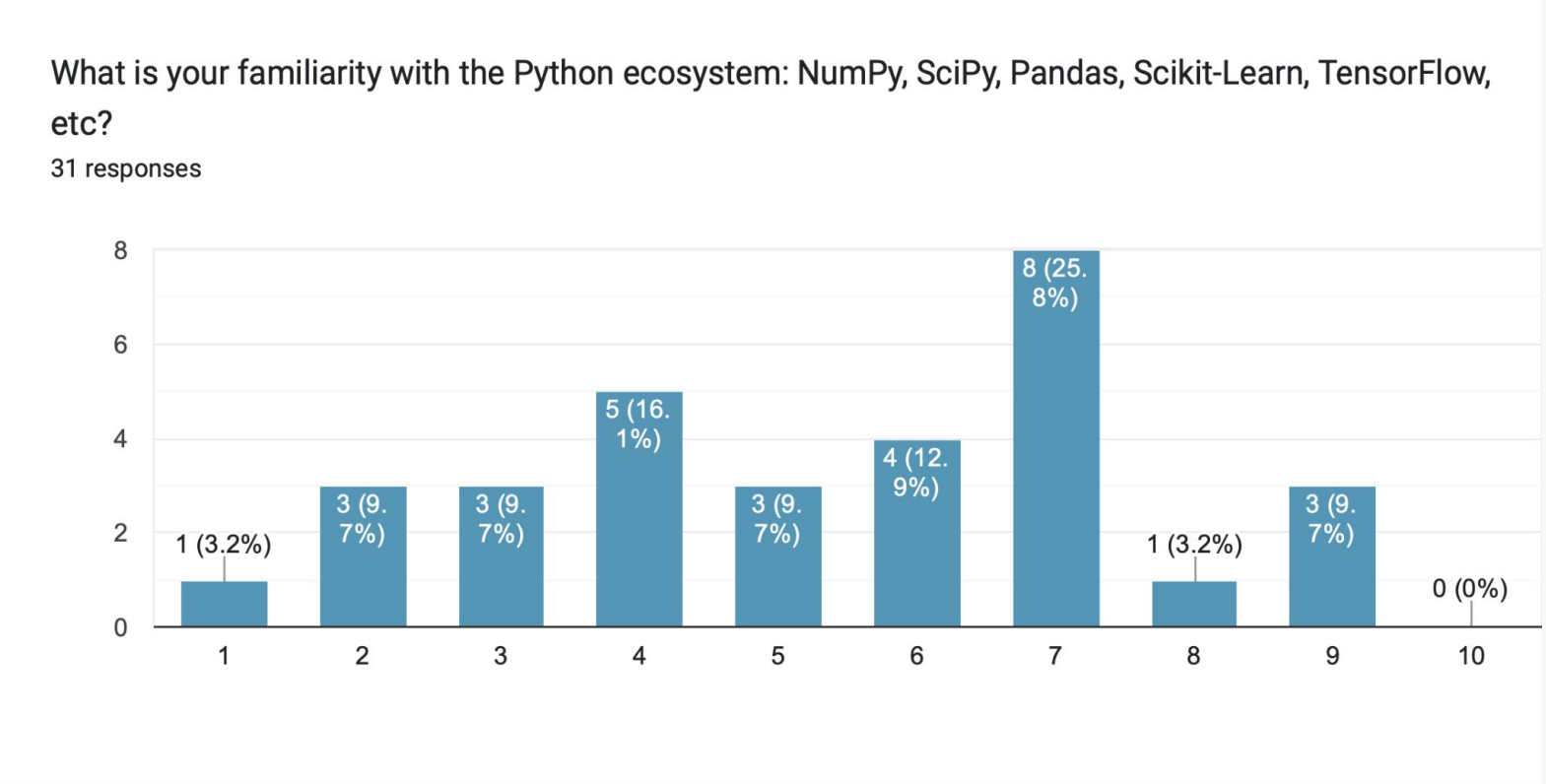
PhD or MSDS Program?



Python?


What is your familiarity with Python?
31 responses
8
6
4
2
0
0 (0%)
3 (9.7%)
6 (19.4%)
2 (6.5%)
1 (3.2%)
5 (16.1%)
7 (22.6%)
3 (9.7%)
4 (12.9%)
0 (0%)
1
2
3
4
5
6
7
8
9
10

Python Ecosystem?



What Are Our Learning Styles?

Which of the following best describes how you prefer to learn new information?

31 responses

For Thursday

1. Get the Jupyter notebook from D2L.
2. Install Anaconda on your laptop.
3. Charge your laptop and bring it Thursday.
4. Max will assign you to groups and post those on Teams.