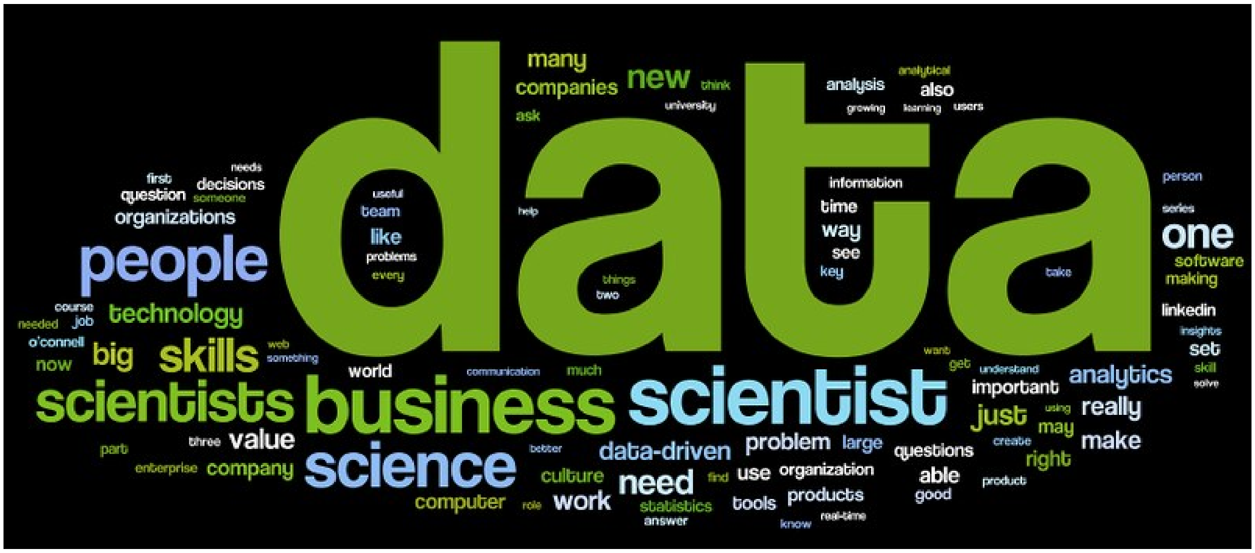
Data Science, an Overview on How to Approach Data in the Age of Tech

We entered the tech era years ago, now it is time to apply these technologies to the field of data science. Let’s dive deeper into how we can think about data.



What is data, and what does it mean to be a data scientist?

Data

Data is raw and unprocessed facts or figures. **Data is the building blocks of information**: numbers, measurements, observations, symbols, and characters all fall under the umbrella of ‘data.’ **Information is data that has been processed or given context;** it is the result of cleaning, organizing, analyzing, and interpreting data in order to make it meaningful. Information can be subjective or objective depending on the interpretation and context. Usually, information leads to an increased understanding and reduces uncertainty in the applicable area (a.k.a., knowledge). Knowledge is information that has been integrated into a larger body of understanding–so the ways in which information is applied to solve problems, make decisions, and navigate new situations. **Knowledge tends to be gathered from both data, information, and personal**

experience; it can be taught to others as well (i.e. someone who has the knowledge of how to read a CSV file in pandas can pass this knowledge to someone else).

Data science

Being a data scientist is a multifaceted role that involves extracting meaningful insights from data to answer questions, solve problems, and drive informed decision-making. **Data science is more than just lines of code, it is a canvas for unveiling stories and problem sets that are hidden in mountains of information.** Data scientists sculpt insights from raw numbers, and bring these insights to life through visualizations and presentations. **Being a data scientist means to embrace this intricate process of organizing big data, uncovering patterns and making a difference in today's world.**

CREATED 3Y VECTORPORTAL.COM

Facts are individual pieces of information that are true and verifiable. This is crucial for the state of the fact to be true, many people confuse fact and opinion. Facts can be simple or complex, and can lead different people down different paths of interpretation.

It is crucial to note the difference between these terms as you proceed forward in the data science process. The bigger data becomes, the more these lines are blurred. **We have to maintain consistency and validity in the way we organize data and make assumptions.**

Principles of Data Science

Understand the Problem & How to Approach It.

Know the purpose behind every project.

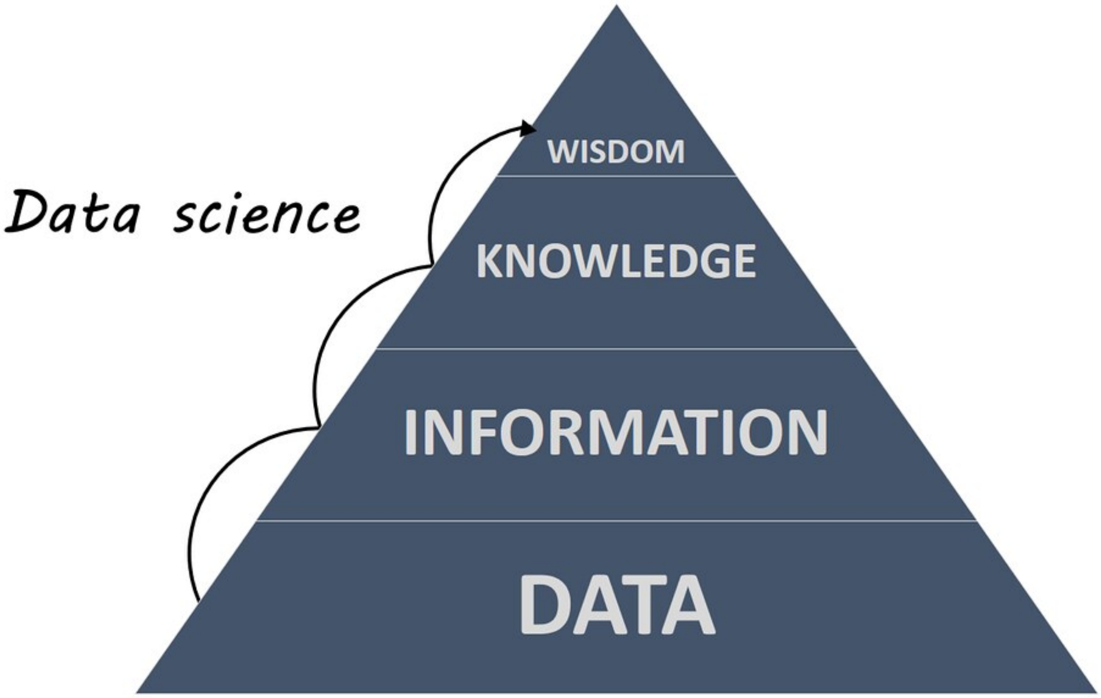
{PIC}

Before diving deeper into the numbers, **I prioritize comprehending the context, biases, objectives, and potential challenges of the project.** I want my analysis to be as aligned with the goal at hand as possible and avoid going down irrelevant paths.

Often the problems at hand require data wrangling and organization. Two methods that I have adapted to think about organizing data are the **DIKW pyramid** and **Jill Lepore's file cabinet metaphor**. These recognized methods are a great way of tackling the initial question:

How do I approach data?

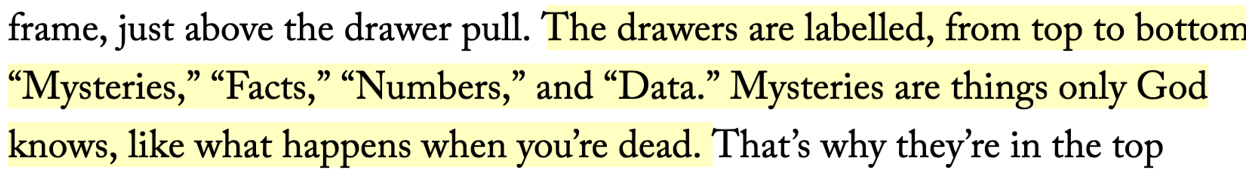
DIKW pyramid is a visual representation of the hierarchical relationship between data, information, knowledge and wisdom. The DIKW pyramid base begins with data, and continues up the pyramid with information, knowledge, and finally wisdom.



I find this to be a very nice way of visually distinguishing the relationship between data and information. **The pyramid emphasizes the need to move beyond simply collecting data to gain true understanding**. It encourages data scientists to analyze, interpret, and apply information to acquire knowledge and ultimately reach wisdom. It also highlights all the different stages of data collection and analysis. If we take care to approach data knowing where it leads us, we will be able to stay on track when we actually dive into the process.

Another way of thinking of data that has impacted my journey as a data scientist is Jill Lepore's filing cabinet metaphor: drawers represent different categories or domains of knowledge; folders represent specific topics or concepts within each drawer; files represent individual pieces of information. The filing cabinet metaphor highlights the importance of efficiently organizing and easily accessing data. This system allows data scientists to make connections between different pieces of data, synthesize information, and form new knowledge.

This snippet is directly from Jill Lepore's "The Data Delusion".

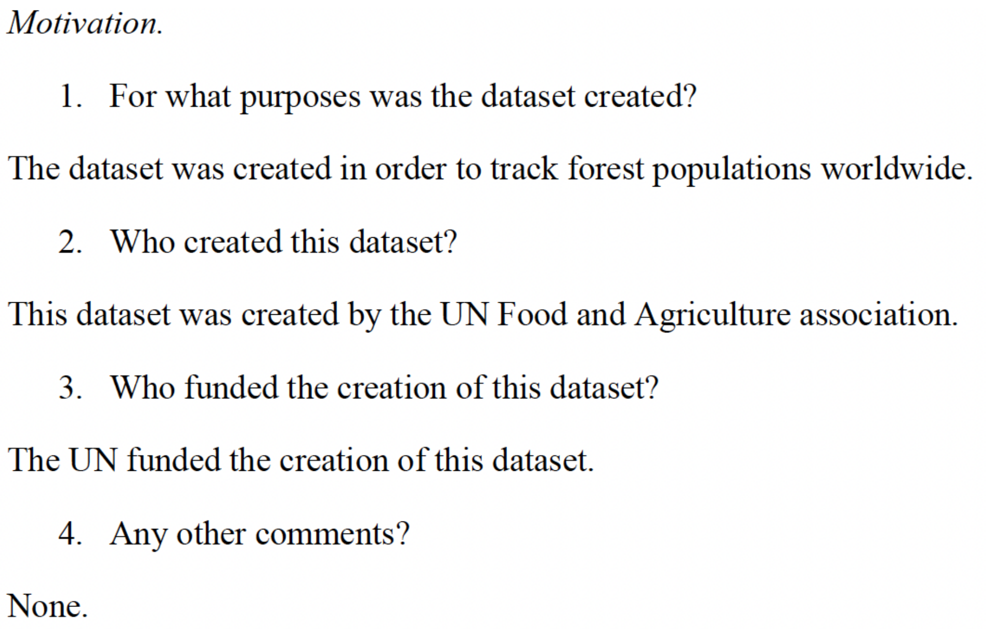


frame, just above the drawer pull. The drawers are labelled, from top to bottom "Mysteries," "Facts," "Numbers," and "Data." Mysteries are things only God knows, like what happens when you're dead. That's why they're in the top

These models show how we can gain knowledge from data. This is what allows us to begin.

Critically examine your dataset

The context of the data is different for every dataset. **It is crucial that you critically examine the data set, looking out for biases or other potential problems in the dataset.** You will want to **know the motivation behind collecting the data and who the data affects**.



Motivation.

1. For what purposes was the dataset created?

The dataset was created in order to track forest populations worldwide.

2. Who created this dataset?

This dataset was created by the UN Food and Agriculture association.

3. Who funded the creation of this dataset?

The UN funded the creation of this dataset.

4. Any other comments?

None.

For example, if I wanted to conduct a data science project that answers my questions about deforestation worldwide (like I did in project 4), I would have to think about when the data was taken, the method by which the data was collected (how does anyone count the number of trees worldwide?), and the people it might affect. In this specific example, the amount of trees per area did not have negative effects on the population, so I was able to proceed.

Project 4 provides a good example of how to critically examine a dataset. The more we evaluate datasets, the closer we come to understanding where they came from, how they were created, how they might be used, and any limitations.

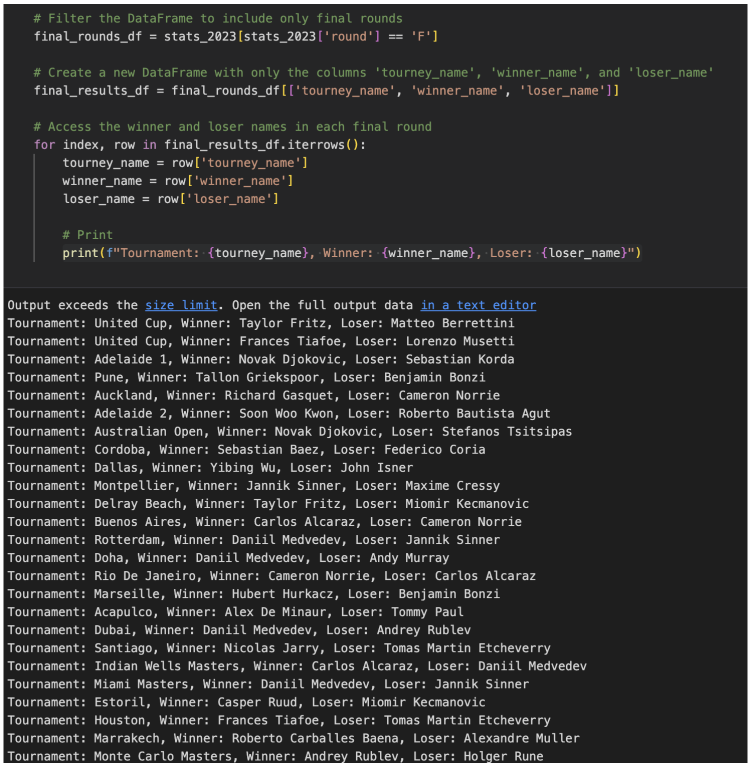
If you don't have this base understanding about your data, you could end up using inaccurate, biased, or harmful data in unintentional ways. It is essential to remember your responsibility to track and maintain ethical strategies for data collection and transformation.

Stay on track

The first step is to **define clear goals and objectives**. **Document your process!** Write down the things you want to know! Think about what you want to achieve with your project and what you will need to accomplish in order to reach your goal. It can be helpful to turn bigger problems into smaller, more manageable problems.

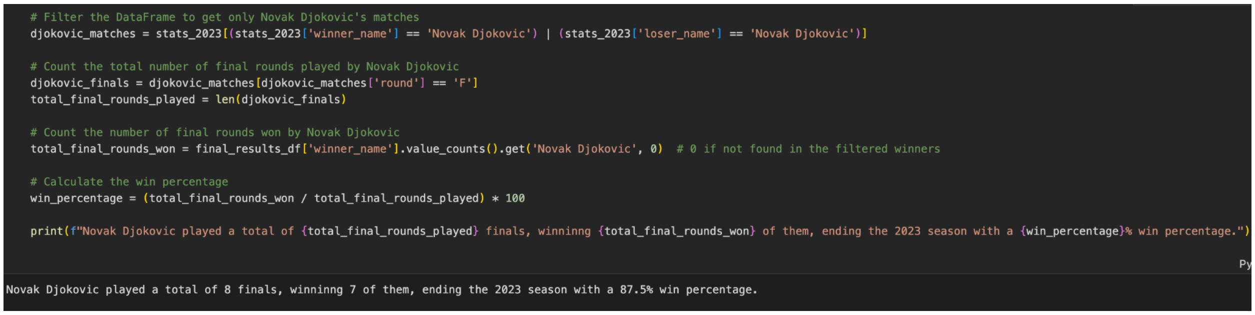
In project 9 when we were creating a data science project of our own, I had to spend some time generating a list of questions and thinking carefully about what I wanted to do with the data.

Identify the most critical tasks to your data science project and **do those first.** Often information is revealed early in the process, so getting the most important information first will help guide the process and ignite curiosity.



```
# Filter the DataFrame to include only final rounds
final_rounds_df = stats_2023[stats_2023['round'] == 'F']

# Create a new DataFrame with only the columns 'tourney_name', 'winner_name', and 'loser_name'
final_results_df = final_rounds_df[['tourney_name', 'winner_name', 'loser_name']]

# Access the winner and loser names in each final round
for index, row in final_results_df.iterrows():
    tourney_name = row['tourney_name']
    winner_name = row['winner_name']
    loser_name = row['loser_name']

    # Print
    print(f"Tournament: {tourney_name}, Winner: {winner_name}, Loser: {loser_name}")
```

```
Output exceeds the size limit. Open the full output data in a text editor
Tournament: United Cup, Winner: Taylor Fritz, Loser: Matteo Berrettini
Tournament: United Cup, Winner: Frances Tiafoe, Loser: Lorenzo Musetti
Tournament: Adelaide 1, Winner: Novak Djokovic, Loser: Sebastian Korda
Tournament: Pune, Winner: Tallon Griekspoor, Loser: Benjamin Bonzi
Tournament: Auckland, Winner: Richard Gasquet, Loser: Cameron Norrie
Tournament: Adelaide 2, Winner: Soon Woo Kwon, Loser: Roberto Bautista Agut
Tournament: Australian Open, Winner: Novak Djokovic, Loser: Stefanos Tsitsipas
Tournament: Cordoba, Winner: Sebastian Baez, Loser: Federico Coria
Tournament: Dallas, Winner: Yibing Wu, Loser: John Isner
Tournament: Montpellier, Winner: Jannik Sinner, Loser: Maxime Cressy
Tournament: Delray Beach, Winner: Taylor Fritz, Loser: Miomir Kecmanovic
Tournament: Buenos Aires, Winner: Carlos Alcaraz, Loser: Cameron Norrie
Tournament: Rotterdam, Winner: Daniil Medvedev, Loser: Jannik Sinner
Tournament: Doha, Winner: Daniil Medvedev, Loser: Andy Murray
Tournament: Rio De Janeiro, Winner: Cameron Norrie, Loser: Carlos Alcaraz
Tournament: Marseille, Winner: Hubert Hurkacz, Loser: Benjamin Bonzi
Tournament: Acapulco, Winner: Alex De Minaur, Loser: Tommy Paul
Tournament: Dubai, Winner: Daniil Medvedev, Loser: Andrey Rublev
Tournament: Santiago, Winner: Nicolas Jarry, Loser: Tomas Martin Etcheverry
Tournament: Indian Wells Masters, Winner: Carlos Alcaraz, Loser: Daniil Medvedev
Tournament: Miami Masters, Winner: Daniil Medvedev, Loser: Jannik Sinner
Tournament: Estoril, Winner: Casper Ruud, Loser: Miomir Kecmanovic
Tournament: Houston, Winner: Frances Tiafoe, Loser: Tomas Martin Etcheverry
Tournament: Marrakech, Winner: Roberto Carballes Baena, Loser: Alexandre Muller
Tournament: Monte Carlo Masters, Winner: Andrey Rublev, Loser: Holger Rune
```



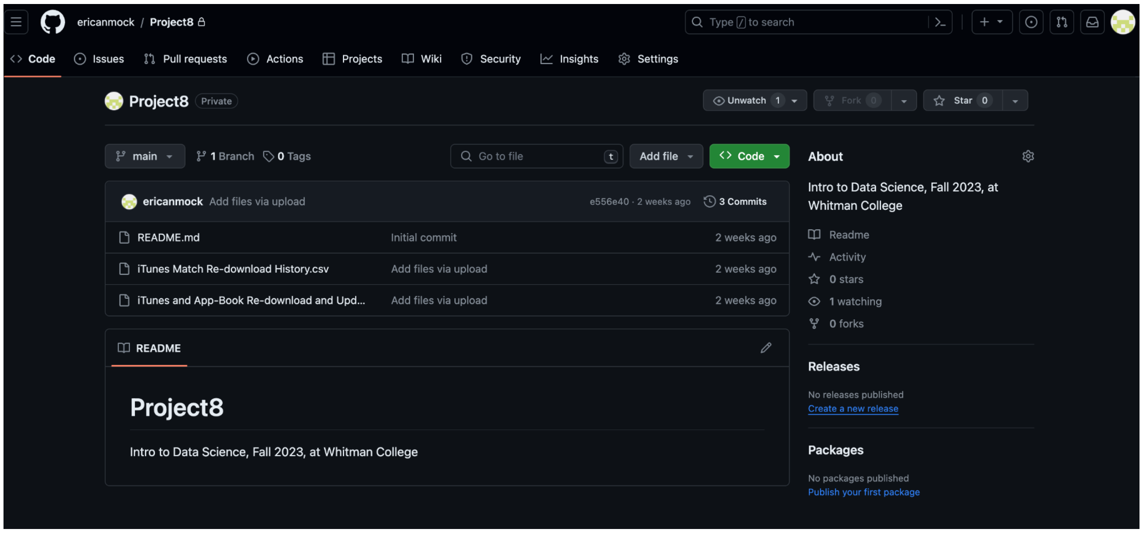
```
# Filter the DataFrame to get only Novak Djokovic's matches
djokovic_matches = stats_2023[(stats_2023['winner_name'] == 'Novak Djokovic') | (stats_2023['loser_name'] == 'Novak Djokovic')]

# Count the total number of final rounds played by Novak Djokovic
djokovic_finals = djokovic_matches[djokovic_matches['round'] == 'F']
total_final_rounds_played = len(djokovic_finals)

# Count the number of final rounds won by Novak Djokovic
total_final_rounds_won = final_results_df['winner_name'].value_counts().get('Novak Djokovic', 0)  # 0 if not found in the filtered winners

# Calculate the win percentage
win_percentage = (total_final_rounds_won / total_final_rounds_played) * 100

print(f"Novak Djokovic played a total of {total_final_rounds_played} finals, winninng {total_final_rounds_won} of them, ending the 2023 season with a {win_percentage}% win percentage.")
```

```
Novak Djokovic played a total of 8 finals, winninng 7 of them, ending the 2023 season with a 87.5% win percentage.
```

In Project 9, I started with the most important thing to me: the players that won tournaments

For some people, **implementing a timeline is helpful**. Setting sufficient time for each stage of the project and achieving these milestones can be encouraging and show that your hard work is paying off.

Spend time managing your data and **use collaborative software products** like GitHub to document your process. Automate repetitive tasks through code and add comments to your code so that anybody could follow along.



ericanmock Add files via upload e556e40 · 2 weeks ago 3 Commits

README.md	Initial commit	2 weeks ago
iTunes Match Re-download History.csv	Add files via upload	2 weeks ago
iTunes and App-Book Re-download and Upd...	Add files via upload	2 weeks ago

Project8

Intro to Data Science, Fall 2023, at Whitman College

Intro to Data Science, Fall 2023, at Whitman College

No releases published
Create a new release

No packages published
Publish your first package

In project 8 we used GitHub to collaborate and share files with each other. This made the process much smoother as files were always accessible.

Learn from others and communicate your roadblocks and challenges. Often you are faced with a roadblock that you think is preventing you from continuing, but the roadblock might actually take you down the wrong path. Be cautious and communicate.

In project 4 when we had to give peer feedback to another group's project, we were practicing our collaboration skills. Peer insights can lead to a better result and can help when you get stuck.

Embrace Data Diversity, Quality, and Equality.

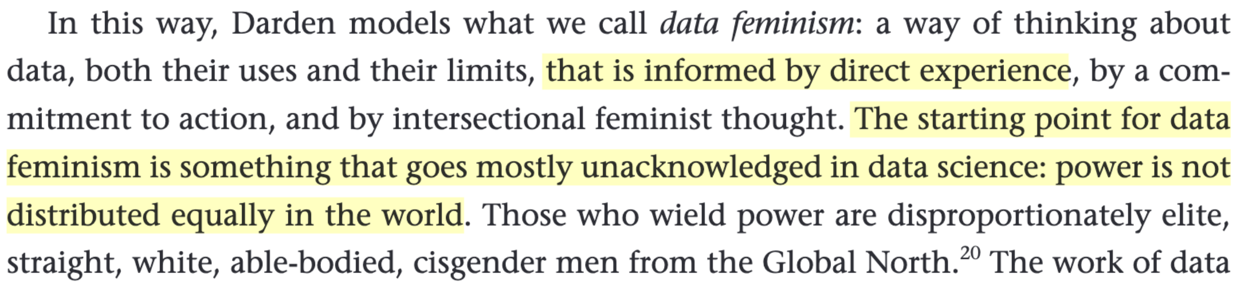
I recognize that **data comes in various forms, structured and unstructured, quantitative and qualitative**. I actively seek out diverse data sources to build a more holistic understanding, while employing quality checks to ensure data accuracy and completeness. Likewise, I also seek out information about where the data originates from, who was responsible for the data collection, and are there potential biases in the dataset. It can be difficult to notice and remove biases, though it is essential to maintain an ethical and equal data science practice.

Take note of biases and problems in dataset

Data, on its own, is a collection of meaningless points. It's context that provides meaning, the backstory, and the surrounding circumstances, the "why" behind the numbers. Without this crucial framework, we're left with blind interpretations, missed patterns, and ultimately, poor decision-making. **Context is the bridge between data**

and understanding, allowing us to see the bigger picture, build accurate models, and communicate insights effectively.

Also, **despite data’s supposed objectivity, it can be riddled with biases**. From skewed samples and faulty measurements to flawed algorithms and historical inequalities, these biases can distort reality and lead to misleading conclusions. It's vital to be aware of these pitfalls, choose methods carefully, and interpret results with a critical eye.



In this way, Darden models what we call *data feminism*: a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought. The starting point for data feminism is something that goes mostly unacknowledged in data science: power is not distributed equally in the world. Those who wield power are disproportionately elite, straight, white, able-bodied, cisgender men from the Global North.[20] The work of data

Catherine D'Ignazio and Lauren Klein. “Data feminism”

“Data feminism” is a paper written by Catherine D'Ignazio and Lauren Klein. It explains biased data and the problems with data science in how **data can build on prejudice and be harmful to minority groups**.

After critically examining your dataset, take note of potential biases and harm that the data itself, or you, have the potential to cause.

Include a diverse dataset

Including diverse data is crucial for accurate and unbiased analysis, but it requires a proactive approach. When you are in the data acquisition process, **seek diverse perspectives and sources. Don't just look where it is convenient!** Actively look for datasets from trustworthy sources and make sure to document every step of the way.

As mentioned above, address biases in your dataset. After critically examining your dataset and noting the power structures hidden within the data, you should be ready to address any bias and have a plan of how to deal with it. Further, you will want to **check for sampling discrepancies, missing values, or any other sort of odd imbalance.**

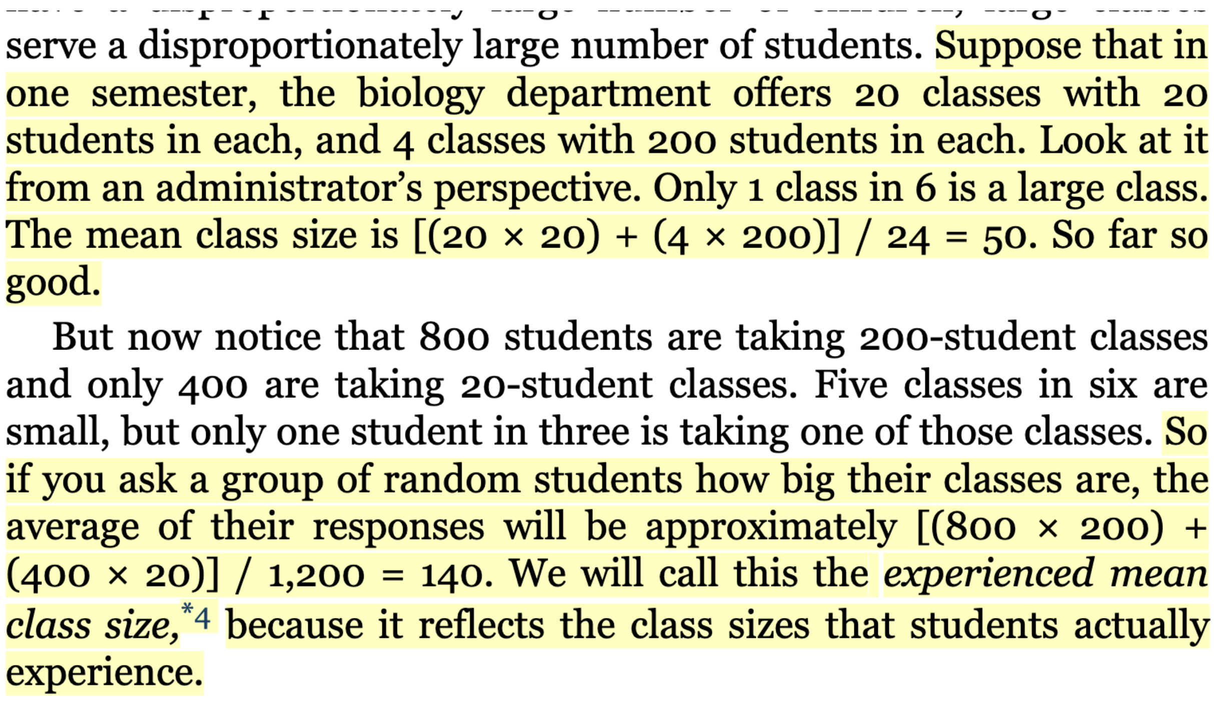
In project 4, halfway through the process, I realized my dataset had a ton of missing values. When it came to plotting my results, the missing values totally threw everything off, and I wasn’t able to account for them, or distinguish that they were missing values. I should have noted that from the beginning and devised a plan of how to deal with it.

How do I conduct research in an unbiased fashion?

One key way bias can come up in the research process is through **selection bias**. Selection bias is a critical concept in the data science realm. This is the idea that conclusions drawn from a sample are inaccurate because the sample does not adequately represent the entire population it's meant to represent.

This could be due to the sampling method, measurement errors, or exclusion, but could also just be a misinterpretation about what is happening in the data.

In **"Selection Bias," from *Calling Bullshit*** by Carl Bergstrom & Jevin West, they nicely explain this concept:



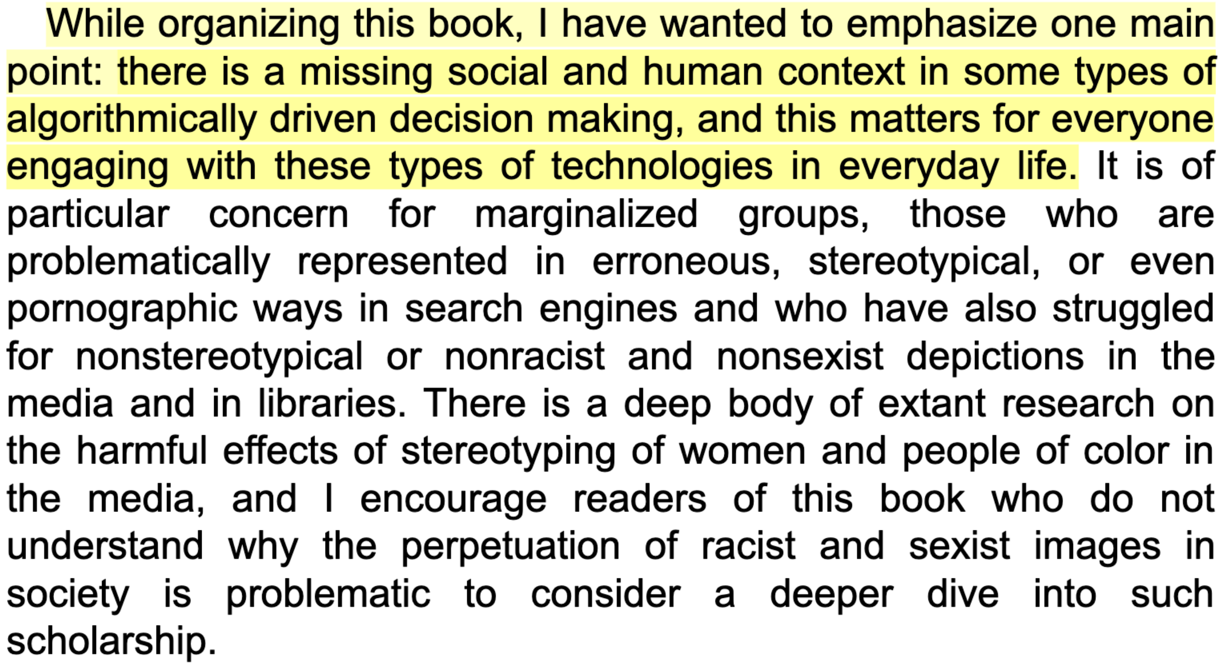
serve a disproportionately large number of students. Suppose that in one semester, the biology department offers 20 classes with 20 students in each, and 4 classes with 200 students in each. Look at it from an administrator's perspective. Only 1 class in 6 is a large class. The mean class size is [(20 × 20) + (4 × 200)] / 24 = 50. So far so good.

But now notice that 800 students are taking 200-student classes and only 400 are taking 20-student classes. Five classes in six are small, but only one student in three is taking one of those classes. So if you ask a group of random students how big their classes are, the average of their responses will be approximately [(800 × 200) + (400 × 20)] / 1,200 = 140. We will call this the *experienced mean class size,*[*4] because it reflects the class sizes that students actually experience.

This is a great example of how selection bias can infiltrate datasets and lead to erroneous results.

Of course, selection bias is not the only type of bias out there. There are also forms of bias in measurements, historical contexts, and confirmation. There are also algorithmic biases in overfitting and underfitting models. Furthermore, when it comes to interpretation, there are framing biases at play.

***Algorithms of Oppression* by Sofia Noble** is a great read that highlights the bias in algorithms and how these power structures ultimately cause harm. Here is a snippet:



> While organizing this book, I have wanted to emphasize one main point: there is a missing social and human context in some types of algorithmically driven decision making, and this matters for everyone engaging with these types of technologies in everyday life. It is of particular concern for marginalized groups, those who are problematically represented in erroneous, stereotypical, or even pornographic ways in search engines and who have also struggled for nonstereotypical or nonracist and nonsexist depictions in the media and in libraries. There is a deep body of extant research on the harmful effects of stereotyping of women and people of color in the media, and I encourage readers of this book who do not understand why the perpetuation of racist and sexist images in society is problematic to consider a deeper dive into such scholarship.

Ignoring bias in research can lead to very dark paths, findings can be wildly off course, leading to flawed and harmful conclusions or decisions. **Watch out for those biases!**

Prioritize Interpretability and Complexity

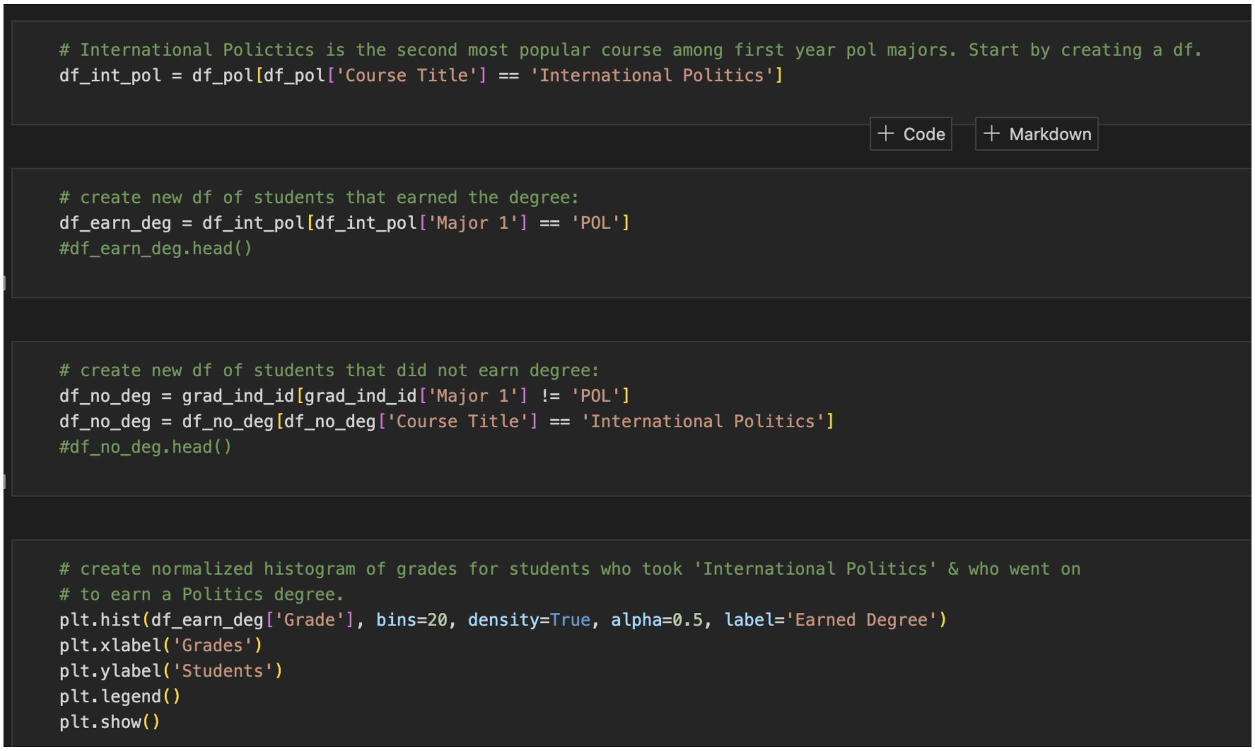
My aim is to **make data insights accessible and actionable, even in the face of complex datasets that produce complex models**. I favor approaches that provide clear explanations for their findings, empowering an audience to learn and make informed decisions.

Skills

There are many skills needed to be a successful data scientist: **technical skills, analytical thinking, communication skills, curiosity, problem-solving, creativity, programming, math, and statistics.** Wow, that is a long list! Let’s briefly go over them.

Technical skills refer to building tools using code and software; this goes hand-in-hand with programming, it will be hard to make it far in this field without some base understanding of coding. Further, base understanding in fields like math and statistics is necessary to understanding the numbers game. At times, data scientists may use formulas and algorithms to extract meaning from, make sense of uncertainty within, and interpret patterns and probabilities hidden in data. At this point, analytical thinking is necessary in order to unpacking problems, sitting through data, and drawing logical conclusions.

In every project, I had to sit with the data and write code that explained it to me.



```
# International Polictics is the second most popular course among first year pol majors. Start by creating a df.
df_int_pol = df_pol[df_pol['Course Title'] == 'International Politics']
```

```
# create new df of students that earned the degree:
df_earn_deg = df_int_pol[df_int_pol['Major 1'] == 'POL']
#df_earn_deg.head()
```

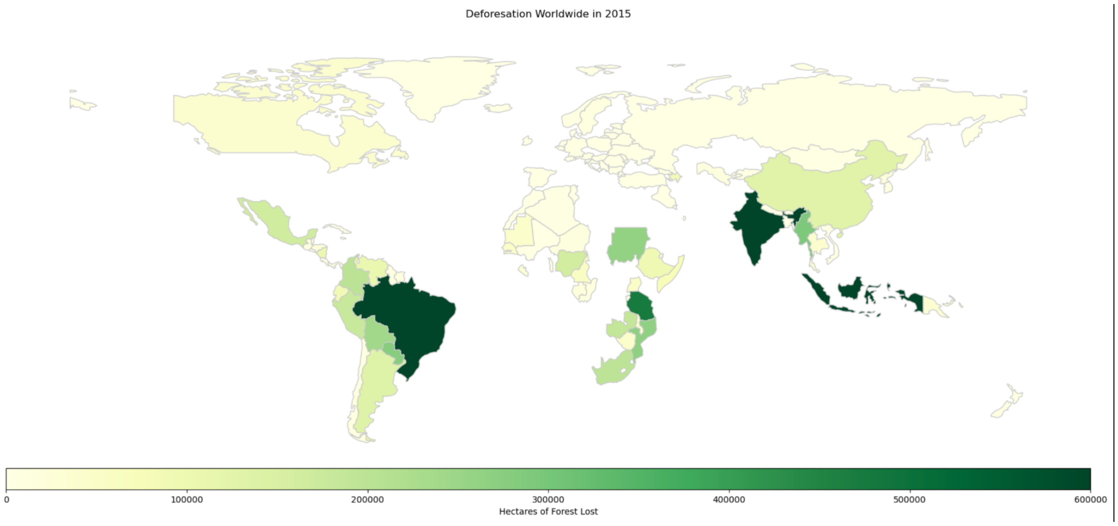
```
# create new df of students that did not earn degree:
df_no_deg = grad_ind_id[grad_ind_id['Major 1'] != 'POL']
df_no_deg = df_no_deg[df_no_deg['Course Title'] == 'International Politics']
#df_no_deg.head()
```

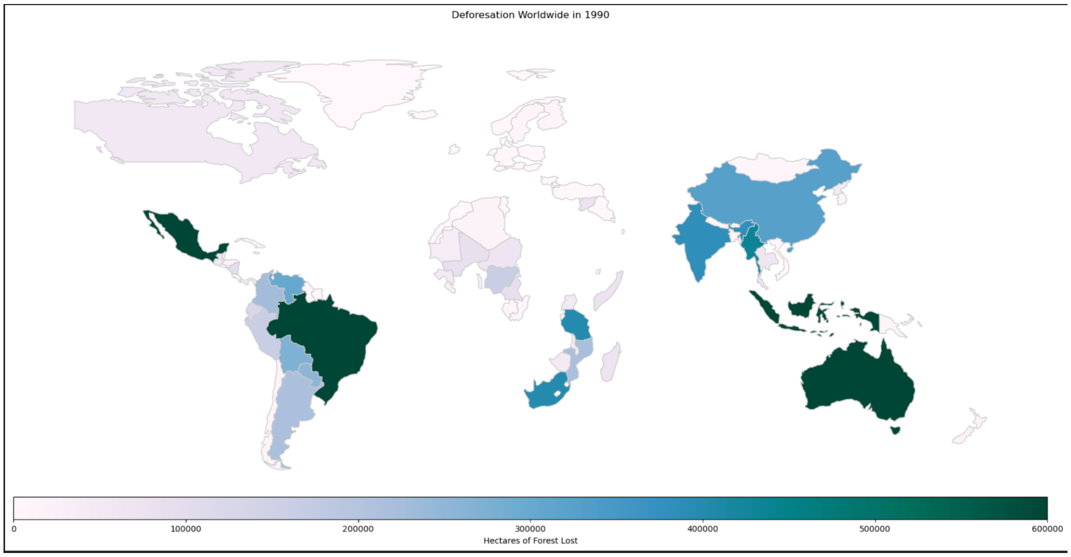
```
# create normalized histogram of grades for students who took 'International Politics' & who went on
# to earn a Politics degree.
plt.hist(df_earn_deg['Grade'], bins=20, density=True, alpha=0.5, label='Earned Degree')
plt.xlabel('Grades')
plt.ylabel('Students')
plt.legend()
plt.show()
```

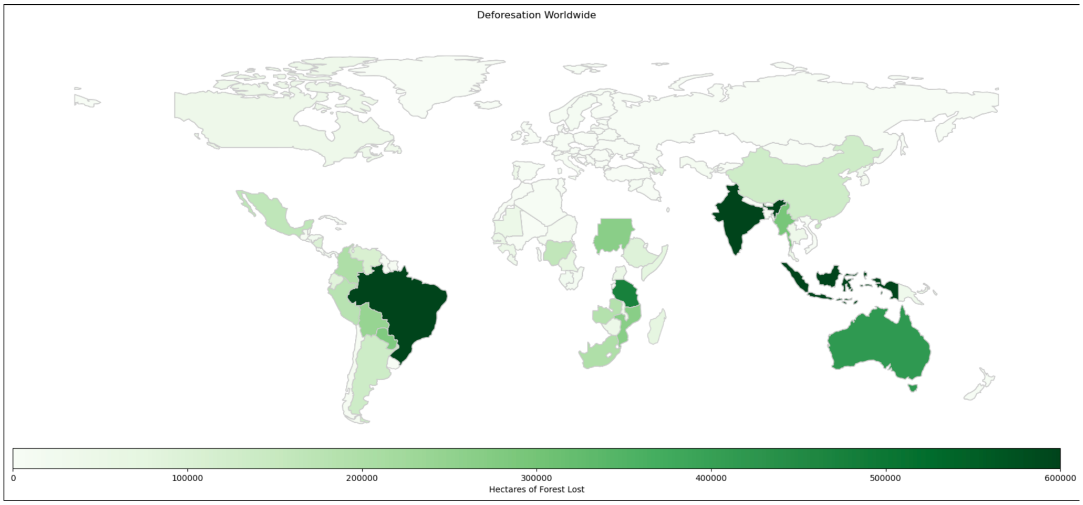
This is a code snippet from Project 3, it shows some of the technical skills you will need to know in order to proceed with a data analysis project.

Further, those who have **creative minds will be able to find innovative ways of analyzing and visualizing data**. The data science world embraces unique perspectives and a creative touch on any project. So, it is import to keep curiosity ignited as a data scientist.

In Project 4, I wanted to map my data on an actual world map, so I figured out how to install geopandas onto my computer and make these cool graphs(!):


Deforesation Worldwide in 2015
0
100000
200000
300000
400000
500000
600000
Hectares of Forest Lost


Deforesation Worldwide in 1990
0
100000
200000
300000
400000
500000
600000
Hectares of Forest Lost

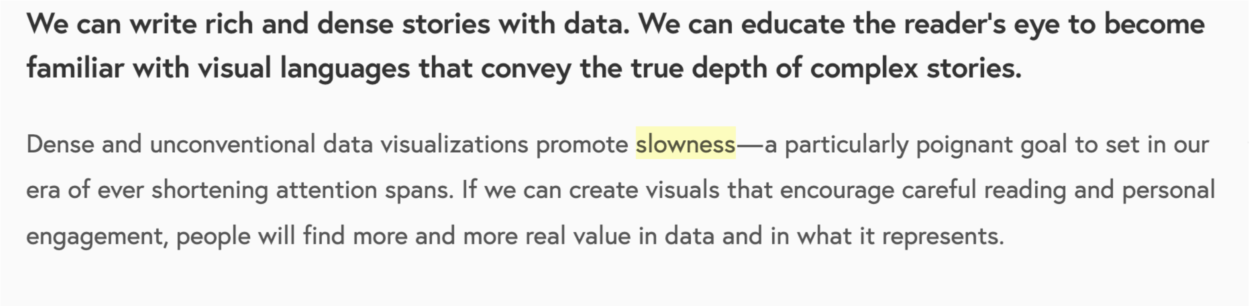

Deforesation Worldwide
0
100000
200000
300000
400000
500000
600000
Hectares of Forest Lost

When things seem too complicated. . . Embrace it!

Embracing complexity in datasets can be daunting, but it's also the key to unlocking deeper insights and richer understanding. Simple, sanitized data may be easier to handle, but it lacks detail. The messiness, the contradictions, the unexpected connection–these are the surprises that complex data holds. By diving into complexity, head first, we can discover patterns and develop solutions, and predict real-world outcomes.

Don’t fear complexity, embrace it!

In Giorgia Lupi’s “Data Humanism” she emphasizes this need for complexity. She refers to it as ‘slow’ data.



We can write rich and dense stories with data. We can educate the reader's eye to become familiar with visual languages that convey the true depth of complex stories.

Dense and unconventional data visualizations promote slowness—a particularly poignant goal to set in our era of ever shortening attention spans. If we can create visuals that encourage careful reading and personal engagement, people will find more and more real value in data and in what it represents.

Giorgia Lupi’s call for “data humanism” through packed visualizations resonates with me. In our era of information overboard, Lupi urges us to embrace slowness–give the time and energy to the dataset you are working with. Unconventional visuals invite us to linger, ponder, and discover the deeper meanings and stories hidden within the data.

Interpreting and communicating results.

Interpreting and communicating results in a data science project is a crucial step. The goal here is to make your results actionable. Clearly explain how your findings can be used to inform decisions, solve problems, or improve processes.

Start by evaluating your project’s performance as a whole, paying special attention to accuracy in the results. Identify any key insights or problems found in the results. Look for patterns, trends, and correlations between variables. Check for biases (as bias can be inserted at any part of the process, it is important to continually check yourself) overfitting, or any statistical significance to ensure that your conclusions are reliable.

Make your findings human-readable! It is important to include the backstory, or just frame the findings as a story overall.

There are many tools and techniques for how you deliver your conclusions to an audience, for example, in Project 9 we gave short presentations with slides that included photos of our visualizations and code. I am biased in thinking this, but I think visualizations are the best way of communicating results.

Learn and Adapt Throughout the Process

Data science is an iterative journey. I constantly evaluate results, refine my approach, and document my process. It is a journey of continuous learning and progress.

Solvable Problems

You might ask **what kind of problems can I solve through data science?** Well, you can answer descriptive questions referring to trend, relationships between variables, and look for most common values. You can also ask predictive questions and make assumptions based on past data; also, you can find patterns and factors that influence specific outcomes and draw conclusions.

However, there are still questions that can’t really be solved with data science–data can be gathered about open-ended and ethical questions, but usually there needs to be human judgement behind these sorts of questions. Further, if there is not much data available about a certain topic, it can be hard or impossible to make future predictions.

Advice

I would tell someone that is hoping to become a data scientist to **be patient, be curious, enjoy the process, and sharpen technical skills**. Focus on programming languages like Python, SQL, or R. Spend time wrangling, cleaning and processing data–this is part of the process. Practice critical thinking and be patient with this learning process, sometimes it can be daunting to dive into these large data problems. Finally, try to enjoy the process and your findings throughout the process. Data science can be an entertaining field.