Ruru Rajbhandari

Prof Jordan Wirfs-Brock

CS-215/ Math-215

12 Dec 2023

Data Manifesto

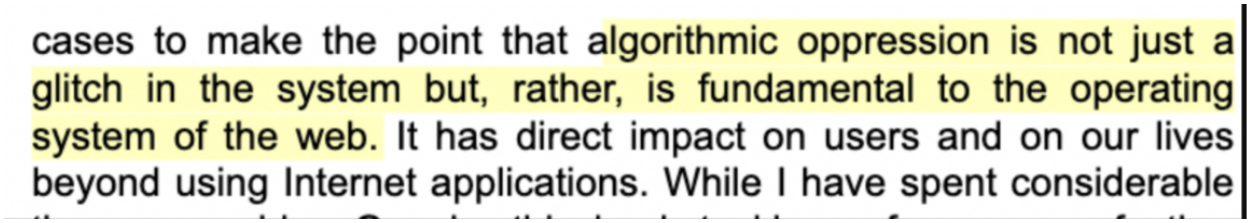
In my perspective data is undefined raw material that holds the possibility of having great insight. Data Science is the technique of refining this untouched and unrefined raw material with the ability of creating meaningful information and extracting knowledge. Data Science is more or less the same as it sounds. It is about taking a scientific approach of analyzing unrefined data and creating meaningful insight from it. It is about analyzing patterns in data and presenting them in a neat coherent way.

Throughout this data science course I have been introduced to a wide variety of data science topics and have changed my perspective on data and the internet. These include projects with Pandas, Matplotlib, Github, requesting personal data, learning APIs and WebScraping, SQL Geospatial Data and many different types of annotation.

As A CS major I genuinely believe that this class has taught me how to code more than any other CS class I have taken at Whitman. This class introduced me to such a wide variety of topics that the skills learned in this class will definitely help me transcend the scope of the class. All the bugs that I ran into while coding and the hours upon hours I have spent in this class has shown me that even though something may seem difficult at first, it is not impossible and that everything has a solution. This is the best CS class I have taken here at Whitman. A previous student of yours named Carl highly recommended this class and I am very glad I ended up taking it.

Alongside coding activities Professor Wirfs-Brock also implemented weekly reading annotations in this class. These were sometimes not always related to coding but were always extremely interesting widening my scope of knowledge beyond just coding.

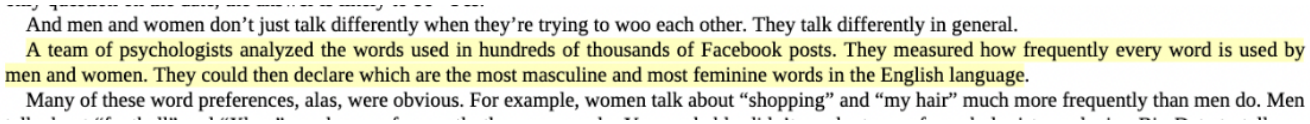
One of the best readings I completed in this class was titled "Intro to Algorithms of Oppression" By Safia Noble. This reading widened my perspective on the creation of algorithms. Up until this reading, although I knew that algorithms were written by humans I did not completely grasp that idea. In the back of mind I always thought that algorithms were written by some machine or some extra-terrestrial object. But after this reading that perspective quickly changed. I realized that the judgements and biases present in an algorithm are directly due to the creators of an algorithm. Algorithms are also designed exactly as they are made.


cases to make the point that algorithmic oppression is not just a glitch in the system but, rather, is fundamental to the operating system of the web. It has direct impact on users and on our lives beyond using Internet applications. While I have spent considerable

The highlighted quote above is from the reading itself. This part of the reading made me truly realize that algorithms are written by humans. The quote is basically saying that algorithms are designed to work exactly how a human designs it to be. So, any present biases are because of either the data or the ways the data was analyzed. The biases present in this algorithm also have an impact on real world populations so, it is always important to take data analysis with a grain of salt.

Another interesting reading that changed my perspective on data was the reading "Data Imagined" by Seth Stephens-Davidowitz. The article basically talks about

how data is used to come up with conclusive evidence using a dataset that may not necessarily be true. This is shown perfectly in an example in this article. This article talks about a data scientist trying to find a masculine and feminine words using a dataset.


And men and women don't just talk differently when they're trying to woo each other. They talk differently in general.

A team of psychologists analyzed the words used in hundreds of thousands of Facebook posts. They measured how frequently every word is used by men and women. They could then declare which are the most masculine and most feminine words in the English language.

Many of these word preferences, alas, were obvious. For example, women talk about "shopping" and "my hair" much more frequently than men do. Men

The article talks about using Facebook posts as a metric of determining "masculine" and "feminine" terms. Although it seems pretty straight forward, the conclusion resulting from this dataset could be quite misleading. The dataset is very limited, using Facebook as their main source. Alongside this, the dataset misleads on other important factors that need to be accounted for such as: age, education level, ethnicity and much more.

When it comes to skills required in data science, these can mainly be divided into: technical skills and critical thinking skills. For technical skills it's important to be proficient in a programming language such as Python as it is the primary language used for data science.

```
#filtered out data for the specific countries used in Our World in Data
entities = ['United States', 'United Kingdom', 'China', 'India', 'World'
filtered_co2 = co2_emissions[co2_emissions['Entity'].isin(entities)]
```

Knowing the basics of python such as lists, strings, and indexing is a crucial skill in data science. Other good technical skills to learn would be for loops, functions, and learning how to read and edit different types of file

```
# Merge the datasets on the rounded timestamp
df_combined = pd.merge(df2, df, on='rounded_timestamp', suffixes=('_yours', '_prof'), how='inner')
```

Learning how to merge files and create columns is also a good tool to use. Alongside basic python knowledge it is important to be able to look stuff up on the internet when you get stuck. As a programmer/data scientist you are bound to get stuck at some point of coding. During these times it's important to use good resources. I always recommend resources such as: Stack Overflow, Github, geeksforgeeks, and Chatgpt.

A strong grasp on data structures, algorithms and machine learning is also recommended. But most importantly a data scientist must be good at their craft. Data scientists with genuine interest and a lot of projects under their belt are likely to succeed in the field. In terms of critical thinking skills: the ability to problem solve and be persistent is key in the field of data science.

Data science is an entrance to a big thrilling world with endless possibilities. The best advice I would go to an aspiring data scientist is being proficient at coding languages such as Python or R. I would look up online tutorials such as freecodecamp in order to first understand the basics of the languages. After learning the basics I would dive straight into making projects. I would make projects about whatever interests me. It's important to remain curious in the journey and to realize that data science is a field that is constantly changing. I would start with easier projects but move on to ones that are considered to be more difficult. Alongside just technical skills I would also try to network. Meeting people in the field that you are aspiring to work in is a great way to boost your career. I would also try to boost problem-solving skills by learning data structures and algorithms and practicing leetcode style problems.

The scope of data science is endless. As a new and incoming field data science has the ability to predict trends, give valuable information and make decisions. The ability to grab large chunks of data and make conclusive knowledge is one that we take granted of.

Although the scope is endless, the field does have some limitations. Data science is directly dependent on the quantity and quality of the data involved leading to biased or poor quality data. This was shown in college admission data that we talked about in class. Data science also scratches the thin line of devling through privacy concerns and potential misuse of personal data.

Four main principles: The first principle is that Curiosity drives Data: At the end of the day a dataset is just a set of numbers or letters until it tells a story. Therefore, it is important to approach a dataset with a sense of curiosity. This means asking the right questions, and exploring new perspectives. Using curiosity we can drive innovative problem solving approaches coming up with different methods and tools to uncover hidden patterns in a dataset. I used this principle specifically in Project 5. In this assignment the task was to recreate a visualization a website was displaying. The first part of this assignment was required to be completed in Python.

Visualization recreation in Python using Pandas

In [8]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

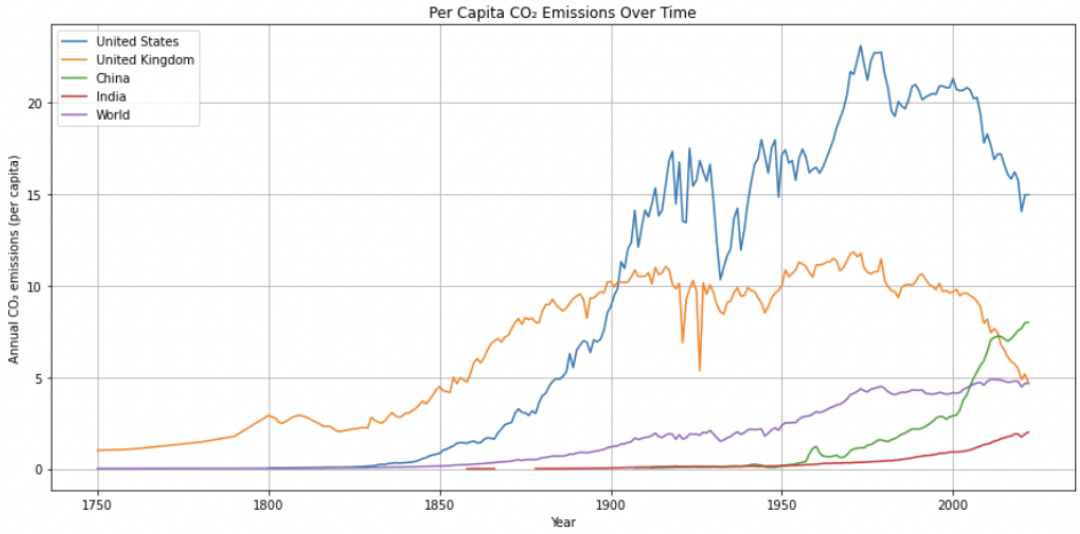
In [9]:
```
#reading the csv file
co2_emissions = pd.read_csv('co-emissions-per-capita.csv')
```

In [10]:
```
#filtered out data for the specific countries used in Our World in Data
entities = ['United States', 'United Kingdom', 'China', 'India', 'World']
filtered_co2 = co2_emissions[co2_emissions['Entity'].isin(entities)]
```

In [11]:
```
#indexes the rows by year, columns represent each entity and cells contain each annual co2 emission
pivoted_co2 = filtered_co2.pivot(index='Year', columns='Entity', values='Annual CO₂ emissions (per capita)')
```

In [13]:
```
#was running into errors here so ended up using chatgpt
#used plt.plot
plt.figure(figsize=(15, 7))
for entity in entities:
    plt.plot(pivoted_co2.index, pivoted_co2[entity], label=entity)

plt.title('Per Capita CO₂ Emissions Over Time')
plt.xlabel('Year')
plt.ylabel('Annual CO₂ emissions (per capita)')
plt.legend()
plt.grid(True)
plt.show()
```



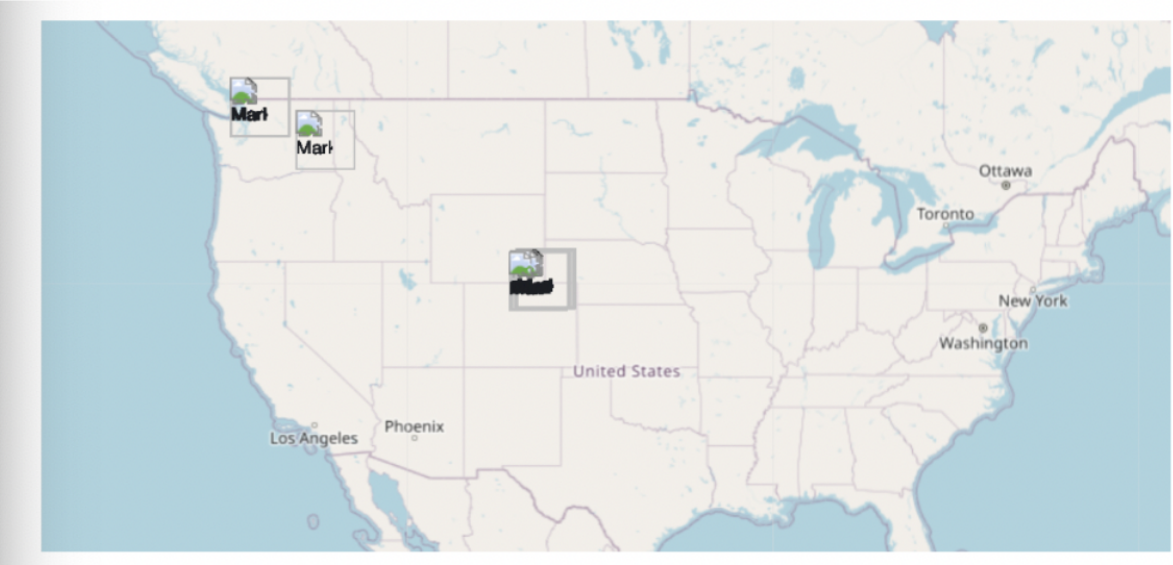
The principle of Curiosity driving Data really pertains to this assignment. I had to think long about how to approach this assignment and how to make the visualization exactly how it looked on the website. Without curiosity this assignment would have never been possible as most would lose interest.

The second principle is Effective Communication: As A data scientist communication is extremely important. Rarely would a data scientist be working on a project alone. Most of the time data scientists would work in teams and groups which

emphasizes the idea of communicating even more important. This principle was particularly prevalent in a group project I completed. This was Project 9: A Data Analysis of your own. I worked on this project with Anthony Maniko. Throughout this project, I realized the importance of communication and team work throughout the project. I realized that its important to dedicate certain tasks for each group member. This is exactly what we did for this project.

> Work Distribution amongst Group: Anthony Maniko and Ruru Rajbhandari both worked on this project equally.
>
> Anthony Maniko:
>
> found the dataset and performed a lot of the data wrangling process whereas
>
> Ruru Rajbhandari:
>
> was mainly responsible for part 3 of the presentation and some of the plots shown below. Was also responsible for the above and beyond component.

Here we evenly divided up the tasks between me and Anthony Maniko,my group mate. This made the task easier as we were able to openly communicate.
The third principle is Persistent problem solving: In data science it is important to continuously keep working towards a solution regardless of the time understanding that solutions only emerge after a certain amount of effort. In coding assignments in general, persistence is a big key in managing frustration and staying motivated even when the progress is slow or non-existent. I believe this is the most important of the four principles. There was a project where persistence was key. This was the assignment where students had to map the location data of Professor Wirfs-Brock using observables. This task was extremely difficult and took me hours on end because I was unfamiliar with the tools Java Script and Observable. After using multiple sources including Geeks for Geeks, Stack OverFlow and ChatGpt. I was finally able to figure something out and came up with this map.



Using professor Wirfs-Brock's location data I was able to plot markers on the United States map and was able to see where she traveled in her time through her visit to Whitman College.

The fourth principle is Respecting privacy: Although data science is based on using and manipulating data, respecting one's privacy is an extremely important task. Obtaining informed consent from individuals whose data is being collected and used is essential. Alongside this, working with sensitive data to remove or mask personal identifiers is crucial. Organizations should also develop and adhere to ethical guidelines and standards.

Reflecting on my experience in the class "Introduction to Data Science" I realize that it has been a transformative journey. Not only has my technical proficiency increased but also my understanding of the role of data in today's society. The course has honed my coding skills in Python, increasing my problem-solving skills and enabled me to approach data from an unbiased perspective. I have learned that data in its raw

form is full of infinite possibilities. Between bar charts, scatter plots, spreadsheets e.t.c there are millions of ways you can represent a simple dataset.

The course has significantly changed my perspective on data and the implications it has on me and society. The reading in this course, especially "Intro to Algorithms of Oppression" by Safia Noble, has been eye-opening, going to show the human biases that can be in these "unbiased" algorithms. The insight from articles underscore the importance of ethical considerations in data science.

Moreover, the course has emphasized the balance between technical skills and behavioral critical skills. As I move forward in my Computer Science degree and further my career, I strive to continuously keep learning like I did in this class. Networking, project based learning, reading, and annotation are all essential parts of learning and it is something that I want to continue. The importance of having good communication skills and tenacity are all shown in this class.