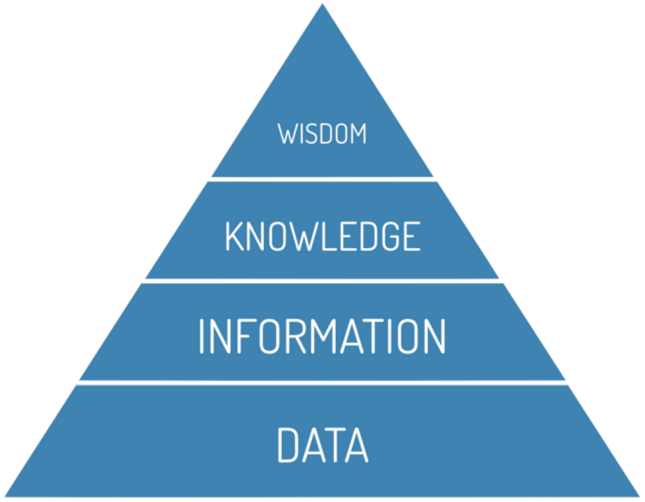
Personal Data Manifesto

Friday, May 5, 2023 8:43 PM

The concept of a utopia is now considered both essential and attainable, thanks to the rapid advancement of Artificial Intelligence. Topics surrounding Artificial Intelligence have been around for decades; yet, it has grown exponentially only within the last few years due to sudden increases in computational power. We all have seen the capabilities of such technologies and how they are complemented in all different fields and domains all over the world. It promotes efficiency and allows people to continue exploring the unknown creatively. However, as AI continues to advance and its impact on society grows, ethical considerations and potential risks are major concerns and should be understood for the betterment of humanity.

What role does data play in this?



Data comes in many forms such as numbers, words, quantitative data, audio, and even videos. While left alone, data is meaningless; yet, it plays a vital role in creating such AI. Data scientists work primarily with the data level of the DIKW pyramid, as they collect, process, and analyze it in order to establish information and knowledge. The knowledge level of the pyramid refers to the domain-specific knowledge that data scientists possess, such as finance, biology, chemistry, economics, and so on. This knowledge allows Data scientists to understand the context of the data and apply appropriate techniques to analyze and interpret it.

Companies all over the world are utilizing data to gain valuable insights - which comes with a high demand for Data Scientists - as it holds monetary value. Data science can be applied to many domains! Below are my 4 main principles that define my data science process.

PUSH THE BOUNDARIES (in your domain)

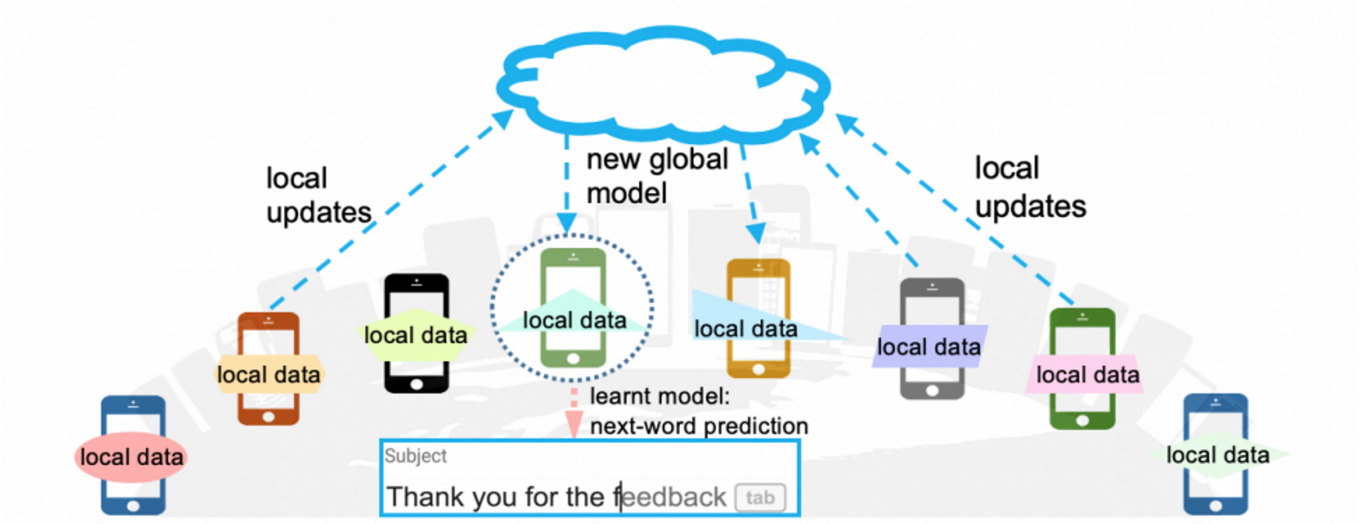
AI is impacting a wide range of fields and domains, as its applications are diverse and constantly expanding. ChatGPT, for instance, is being taken advantage and relied upon from many companies. I asked ChatGPT to provide a few different examples of domains AI and Data Science are impacting and here is what I was reported back…

1. Healthcare: AI is being used in medical research, diagnosis, and treatment. AI algorithms are helping to identify patterns and insights in medical data, which can be used to improve patient outcomes and develop more effective treatments for diseases.
2. Finance: AI is being used in banking, trading, and financial analysis. AI algorithms are used to detect fraud, identify investment opportunities, and predict market trends.
3. Manufacturing: AI is being used to optimize production processes and improve quality control. AI algorithms are used to analyze sensor data from manufacturing equipment, identify patterns and anomalies, and make real-time adjustments to production processes.
4. Retail: AI is being used to optimize supply chain management, inventory management, and customer service. AI algorithms are used to predict customer behavior, analyze shopping patterns, and recommend products to customers.
5. Education: AI is being used to personalize learning experiences and improve educational outcomes. AI algorithms are used to analyze student data, identify learning gaps, and provide customized feedback to students.
6. Transportation: AI is being used in autonomous vehicles, traffic management, and logistics. AI algorithms are used to optimize routes, reduce congestion, and improve safety on roads and highways.

With the rise of big data and the need to derive meaningful insights from vast amounts of information, data scientists have become critical players in many industries, including healthcare, finance, manufacturing, retail, education, and transportation (according to ChatGPT).

As data-driven decision-making becomes increasingly important in the digital age, the demand for data scientists is only expected to grow. Having a strong background and being able to physically deliver the basic responsibilities as a data Scientist is extremely important; however, I think it is extremely important to encourage creativity related to people's passions, as they are more likely to take a deep interest in the data they are analyzing and the insights they are uncovering. Take the following example…

Federated Learning



Training machines and models requires a substantial amount of quality data. While data is gathered everyday in large datasets, there are masses of inaccessible information due to privacy concerns. Federated learning is Machine Learning on decentralized data that enables edge devices to do Machine Learning without centralizing data and secures privacy. Federated learning involves training a shared model, which allows clients to keep their information confidential.

I think emphasizing privacy concerns and further improvements in Federated learning for healthcare is critical. Due to the massive number of devices and communications through a network, for instance, computation can be slower and expensive. Each edge device varies in terms of network capability, hardware variability, and power. Thus, it is necessary to develop communication efficient methods that convey model updates as part of the training process, as opposed to sending the entire dataset over the network. Federated learning is a newer machine learning technique that will continue to develop; however, it is important to understand the need for these systems in many domains, not just healthcare.

The impact of AI spans across numerous domains and industries, making it increasingly important for data scientists to invest deeply in their chosen field. I truly believe that it is not sufficient enough to be a strong programmer; moreover, a true data scientist must possess a deep understanding of their domain and a desire to effect positive change. Whether it's in healthcare,

finance, manufacturing, or any other domain, data scientists have the opportunity to utilize their skills and expertise to make a lasting impact and contribute to the betterment of society.

UNDERSTAND THE DATA

Data consists of raw values or measurements and if not analyzed or contextualized, it is meaningless. A new interpretation of the DIKW pyramid - specifically the information stage - can be demonstrated as below…

$$Information = \sum(Data)$$

Here, we see the potential and possibilities of data. While all data might not be relevant, its raw nature provides means for Data scientists to potentially further explore the unknown. The value of data increases with the number of data scientists who are able to analyze and interpret it effectively.

Data scientists work to transform raw data into information by identifying patterns, relationships, and trends within the data. They use techniques such as data visualization and statistical analysis to present this information in a meaningful way to help organizations make informed decisions.

It is important to ensure the quality of the data we are either accessing or using. Similar to how we were told growing up to NOT use Wikipedia for essays, ensuring the quality of data is presented as indispensable for a variety of reasons. This means validating the source of data, and performing necessary data cleaning and transformation to ensure consistency and completeness. We must also be aware of potential errors in data, and continue taking steps to identify them.

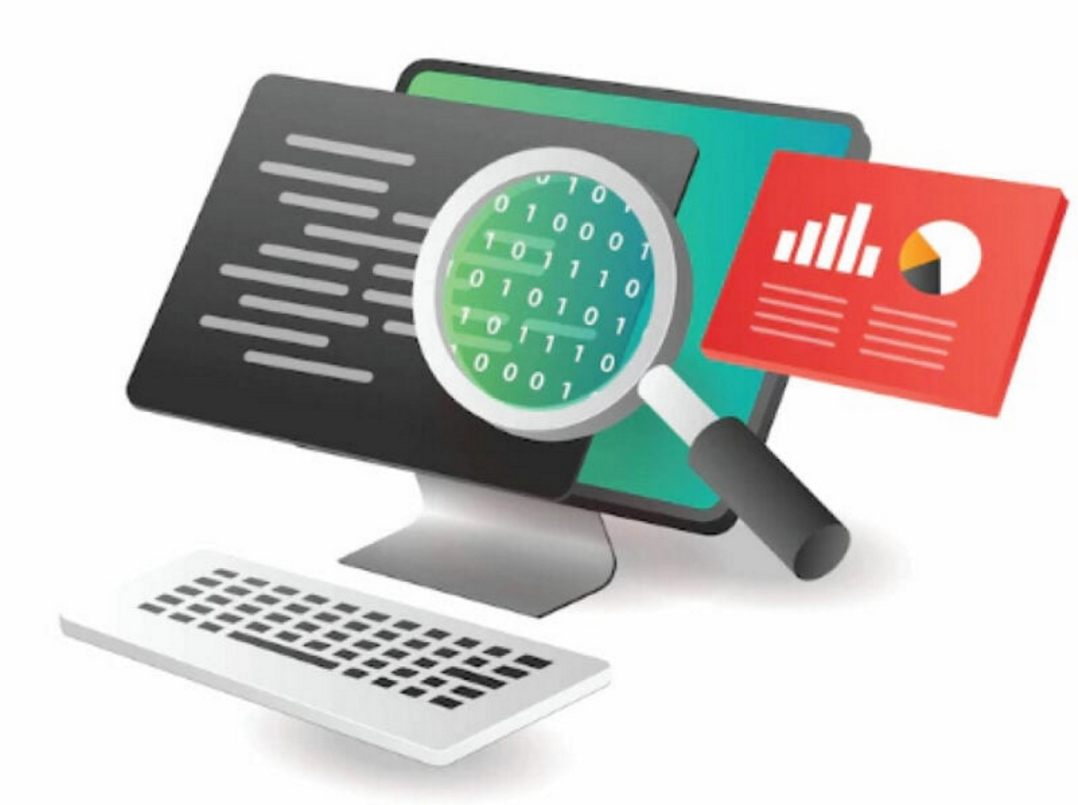
By prioritizing data quality, we can ensure that our analyses and models are based on accurate and trustworthy data, leading to more accurate and actionable insights. Take the following conversation generated by ChatGPT as an example of gendered bias…

The doctor gave the nurse money because she owed it to him. Who owed money to whom?

In the given sentence, "she" refers to the nurse. Therefore, the sentence means that the nurse owed money to the doctor, and as a result, the doctor gave her money.

I pulled this from a Whitman CS Slack post from Professor Janet Davis to highlight that bias exists in AI models, such as ChatGPT. This is a critical issue as it relates not only to gender but to race, ethnicity, and other socio-cultural factors. So many companies are relying on AI models such as ChatGPT for performance enhancements. Students are interacting more and more with newer technologies. AI is being viewed as potent and perfect, which is extremely dangerous.

In addition to analyzing the quality of the source, being able to work with all kinds of data is a basic essential tool for any data scientist.



Data wrangling is the process of cleaning and preparing raw data for analysis. I believe that before we can talk about the other three principles for the data science process, the importance of being able to actually work with data should be emphasized. Below is a table of basic analysis data scientists regularly do…

Reshaping Data	Handling missing values	Grouping Data

`df.sort_values()`
`df.rename(columns={"A" : "new_A"})`
`df.sort_index()`
`df.drop(columns={"A"})`
	`df.dropna()`
	`df.fillna(-1)`
		`df.groupby()`
		`df.pivot_table()`

Data wrangling is an essential step in the data science process as it involves transforming and cleaning raw data into a format that

can be analyzed, allowing for more accurate and meaningful insights to be generated. Without proper data wrangling, data scientists may encounter issues such as incomplete or inconsistent data, which can negatively impact the accuracy and reliability of their analyses.

I wanted to close this section with a quote from "Data Feminism" that I feel is essential when viewing AI and to highlight the importance of quality data…

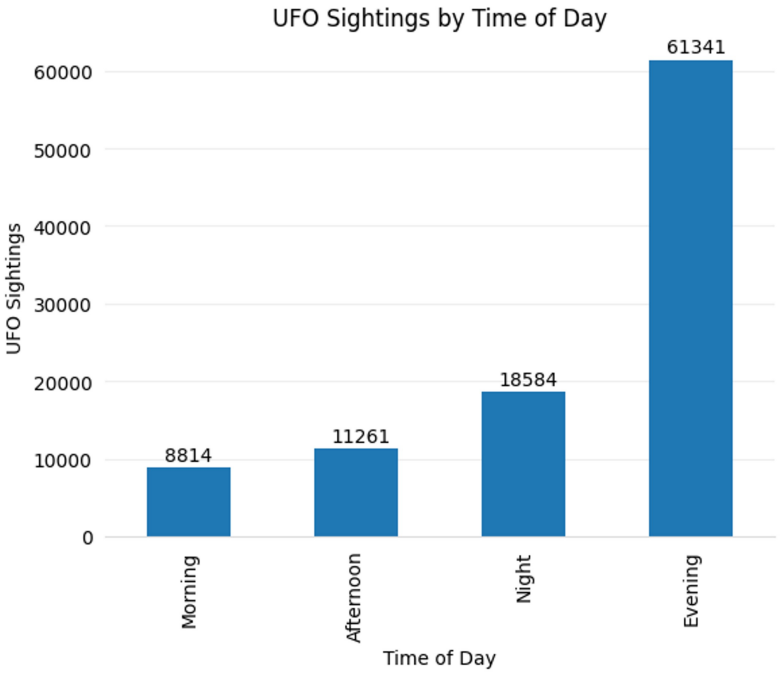
> *"Underlying data feminism is a belief in and commitment to co-liberation: the idea that oppressive systems of power harm all of us, that they undermine the quality and validity of our work, and that they hinder us from creating true and lasting social impact with data science," (9).*

COMMUNICATING DATA FINDINGS

Proper communication is widely regarded as an essential skill that individuals should possess. While part of being a data scientist is performing high level analysis, being able to communicate about findings and newer insights is equally as important.

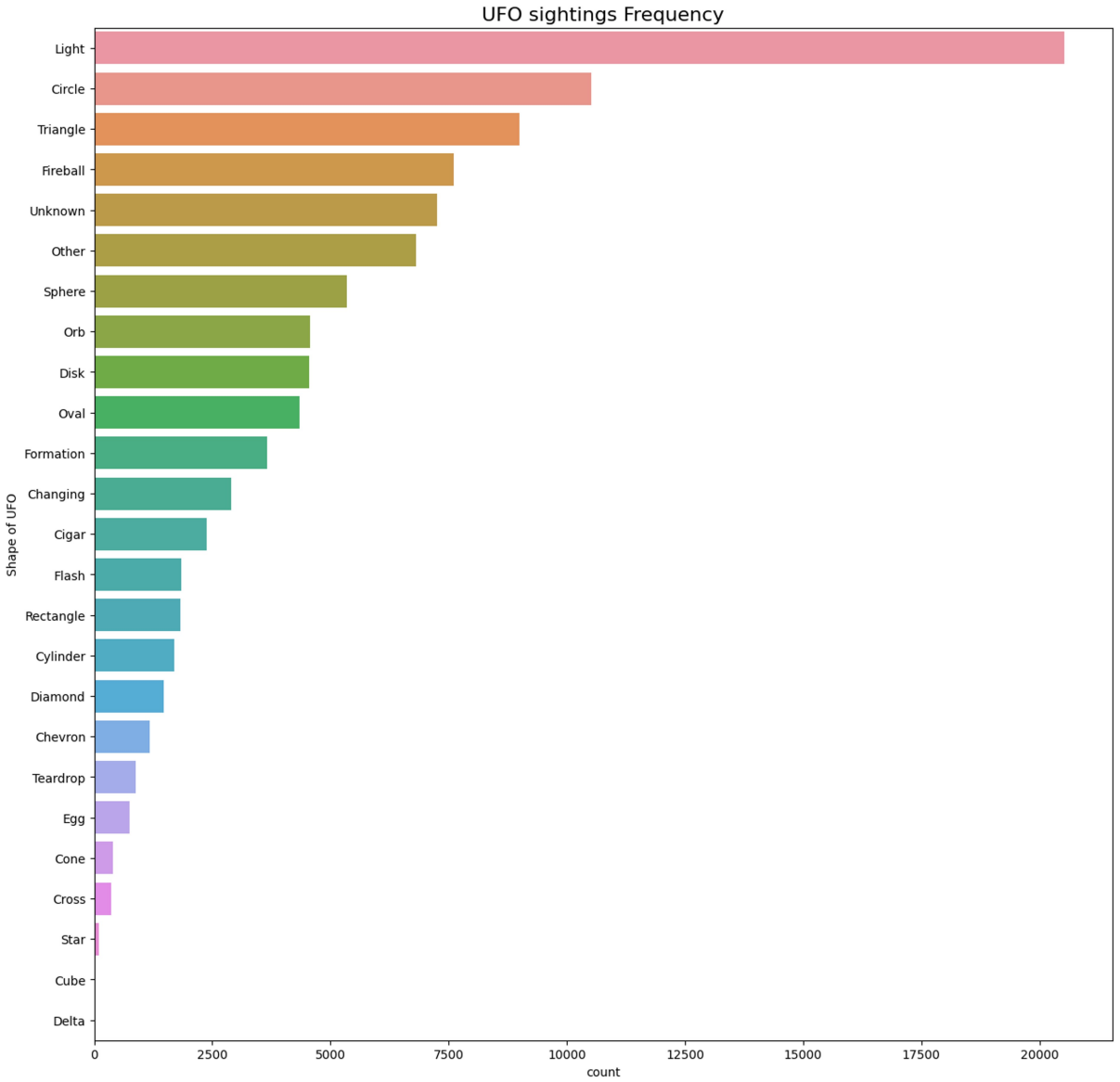
Python contains many libraries that allows for data scientists to visualize their data. Looking at one of my past projects on UFOs that I worked on with my partner, Colby, here are a few examples of different visualizations that we made…

Matplotlib



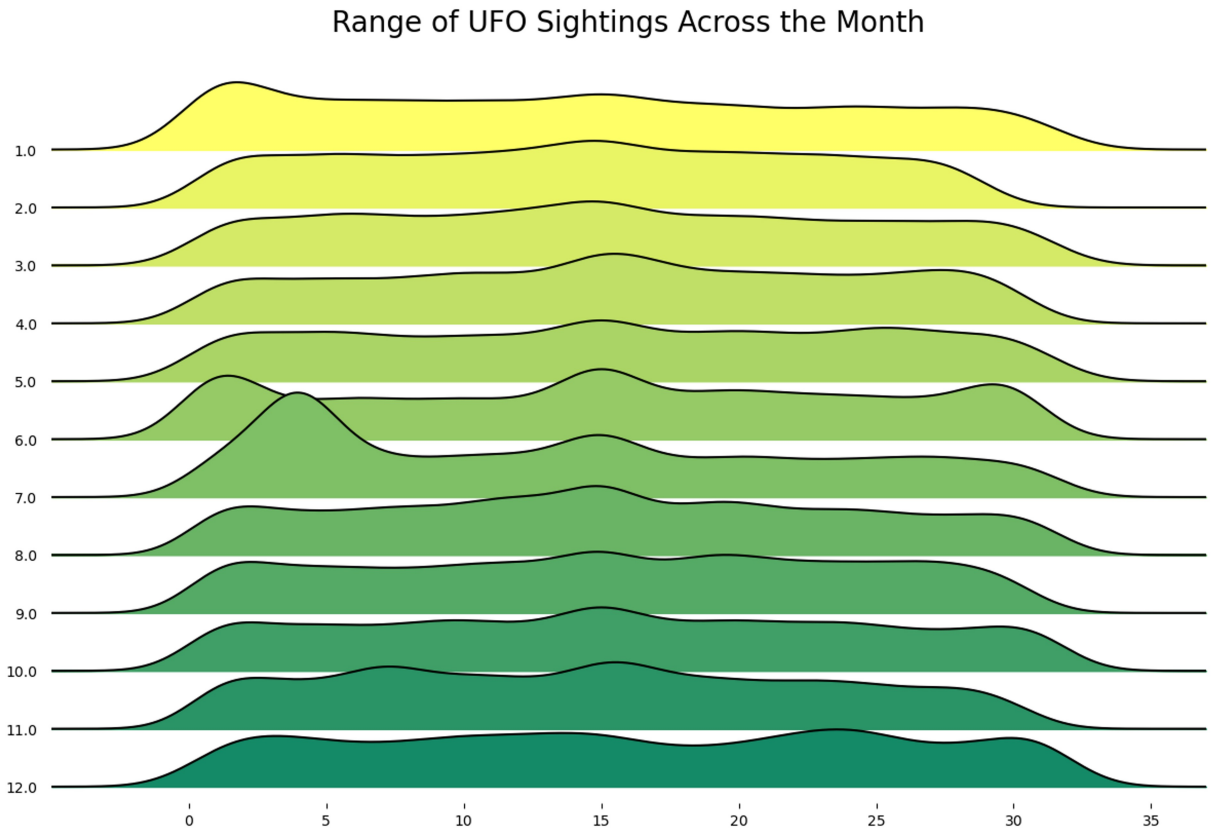
This plot showcases UFO sightings count based on the time of day. It makes sense that evening sightings are more common because it is dark outside but still early where most people are outside looking at the stars.

Seaborn



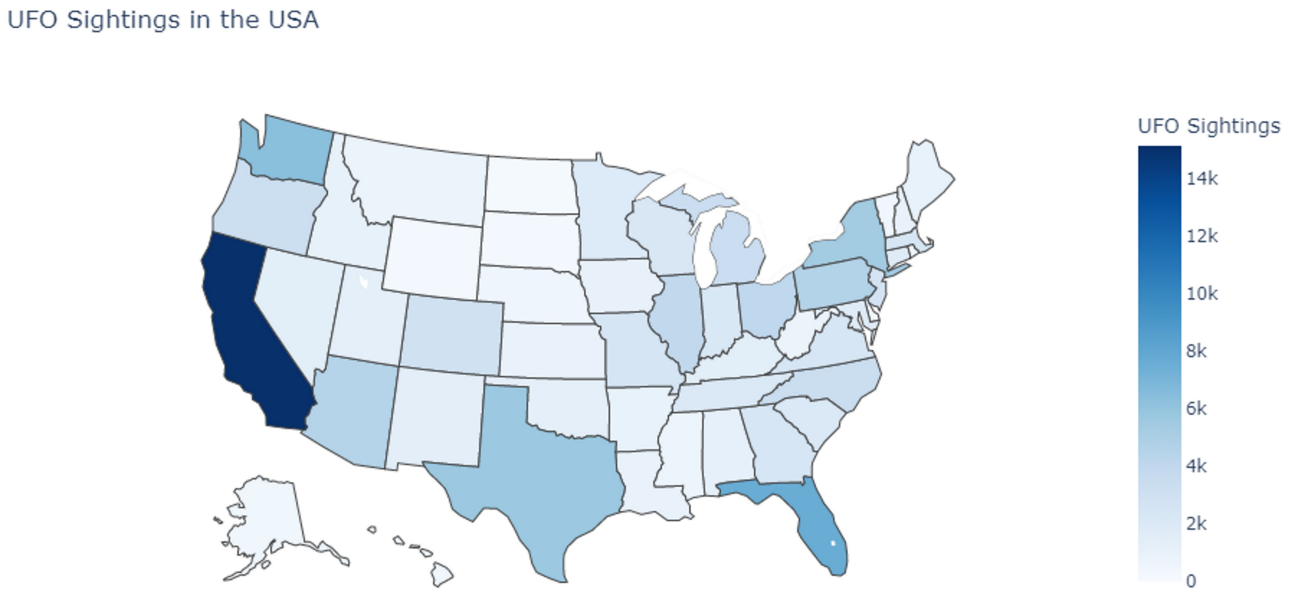
This plot demonstrates different UFO shapes and their corresponding frequency count. It seems that a light is the most common UFO sighting according to this plot?

Joyplot



This plot shows the sightings in a month day relationship. It is interesting to note that before we found that light UFOs are more common and the more common days for sightings are on days where there are a lot of fireworks (July 4th, New Years).

Plotly



This plot shows the distribution of sightings in the whole United States. Looks like California has the most sightings for UFOs.

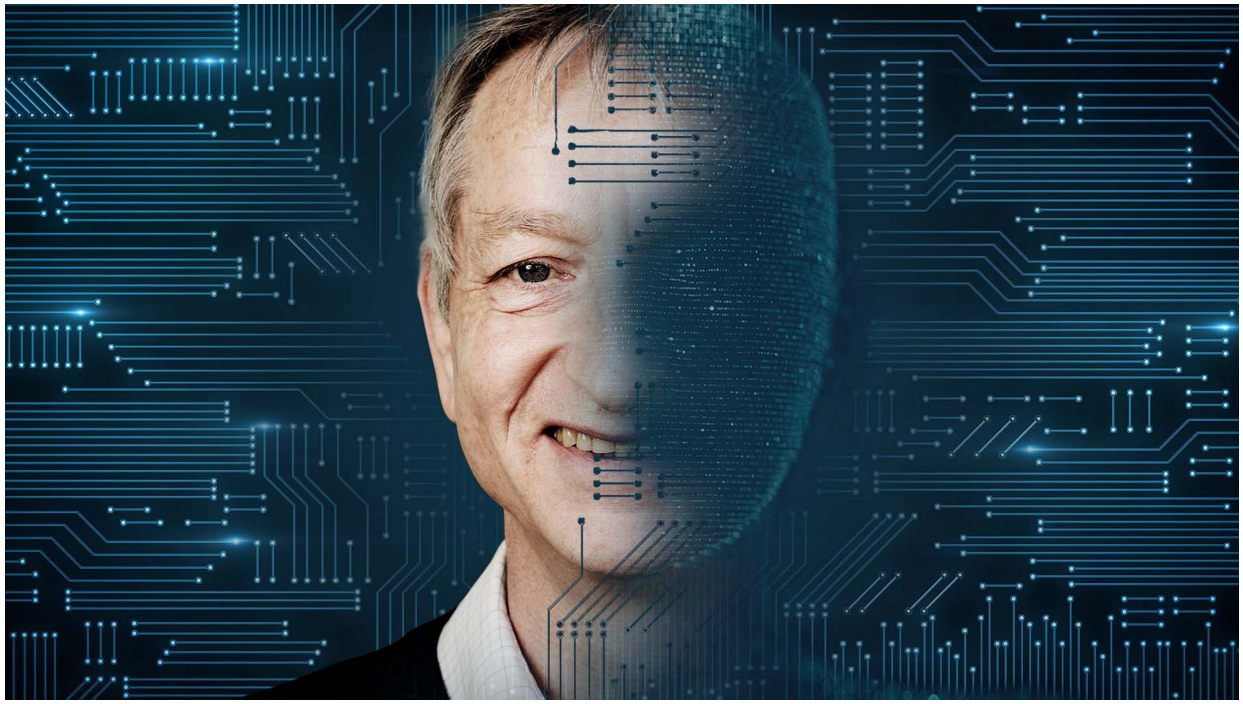
Data science involves presenting insights and findings in a clear and understandable manner. Being able to create clear visualizations and find correlations, is a powerful tool. It requires a combination of technical skills and communication skills to present data in a clear, understandable, and meaningful way. The choice of visualization will depend on the specific goals of the analysis and the type of data being analyzed.

I was a bit surprised myself when adding these graphs to this Data Manifesto that it took two different plots (one from Colby and another from myself) to see that correlation that most sightings happen on days where there are fireworks and the fact that light UFOs are reported most frequently!

BE HUNGRY FOR CONTINUOUS LEARNING

We have seen an exponential growth in AI in just the last decade. Deep learning, for instance, has been around since the 1960s; yet, proper implementation began only in the early 2000s. Whether that is a change in techniques or trends, staying up-to-date is a critical component as the level of computational power increases.

Who is Geoffrey Hinton



Geoffrey Hinton, widely known as the "godfather of AI," recently departed from Google citing concerns about the potential impact of his work. Hinton fears and believes that as AI algorithms scales exponentially, they might outstrip their human creators within a few years, “I used to think it would be 30 to 50 years from now. Now I think it's more likely to be five to 20.”

The risks of AI is so extreme. Imagine chatbots - similar to ChatGPT - spreading misinformation and manipulates more and more people. The consequences of AI can be not just harmful but deadly.

In the realm of data science, it is crucial to stay on top of current events and emerging techniques. While data itself may appear insignificant, it can serve as a powerful ingredient that fuels potentially harmful AI applications. The field of data science is constantly evolving, demanding continuous learning and adaptation.

Embrace a mindset of growth and development. Dedicate yourself to the pursuit of knowledge and find a domain that genuinely piques your interest. By immersing yourself in that domain, you can pour your heart into your work and make a meaningful impact. Strive to exercise responsibility and ethical considerations in your work, understanding the potential unintended consequences that AI can cause.