

Data for People: A Manifesto

Emily Lockman

May 9th, 2023

CS215: Data Science

Professor Wirfs-Brock

Data has become a critical driver of progress and innovation in many fields, from business to healthcare to scientific research to solving issues like the climate change crisis. With the rise of advanced technologies and digital platforms, we are generating and collecting vast amounts of data every day, enabling us to gain insights and solve complex problems that were once impossible. However, this abundance of data also brings challenges and complexities for investing in the right tools and techniques to make sense of the massive volumes of information we have access to. Despite this, the potential benefits of data science are too great to ignore, and it is clear that data will continue to play a vital role in shaping the future of our world.

As a student of Data Science, I have narrowed in on four fundamental principles that encapsulate the significance, the perils and the complexities of working with data:

- Decision Making
- Interdisciplinarity and Collaboration
- Ethical Considerations
- Continuous Growth

By adhering to these four pillars of data science, professionals in the field can leverage their skills and knowledge to tackle complex data sets and extract meaningful insights that can be used to make sense of the vast amounts of data available today, paving the way for a better future powered by data-driven insights.

DECISION MAKING

Data refers to raw, unprocessed information that is collected and stored in a structured format. As someone who is taking Data Structures this semester, my understanding of data includes its representation as a sequence of binary digits, consisting of 1's and 0's, which hold meaning and are processed by *computers* at a fundamental level. However, in data science, the term "data" is used more broadly to refer to any raw, unprocessed information that can be collected and stored in a structured format, including text, images, videos, and even other forms of data to be understood and interpreted by *humans*.

Once data has been collected, it is processed and organized in a way that makes it useful and meaningful. This processed and organized data is known as information. **Information is data that has been compiled in a way that is understood and interpreted by humans.** In its raw form, the data may be difficult to understand or interpret, much like a disorganized file cabinet can make it hard to find the desired information. However, just as a well-organized file cabinet can make finding information easy, data that is processed and organized in a meaningful way becomes information that can be easily understood and interpreted by humans. (Jill, 2023). The process of transforming data into information (or its organization into categories) is a crucial step in unlocking the insights and, in turn, better decision making strategies.

From this information, knowledge can be gained. Knowledge is the understanding and insight that is gained from information. It is the result of processing and analyzing information, and involves

interpreting and making sense of the information to gain a deeper understanding of the subject matter. Knowledge is typically more abstract than information and is often based on experience and intuition, something that is inherited from the human and less computer brain.

The DIKW pyramid is a model that illustrates the relationship between data, information, knowledge, and wisdom. It suggests that data is the foundation upon which information, knowledge, and wisdom are built. Each layer builds upon the layer below it, with data forming the foundation

upon which information is built, information forming the foundation upon which knowledge is built, and knowledge forming the foundation upon which wisdom is built. (Raddaoui, 2023).

[Figure 1: Raddaoui, 2023]

Data scientists use this framework to transform raw data into meaningful insights that can be used to make informed decisions. By

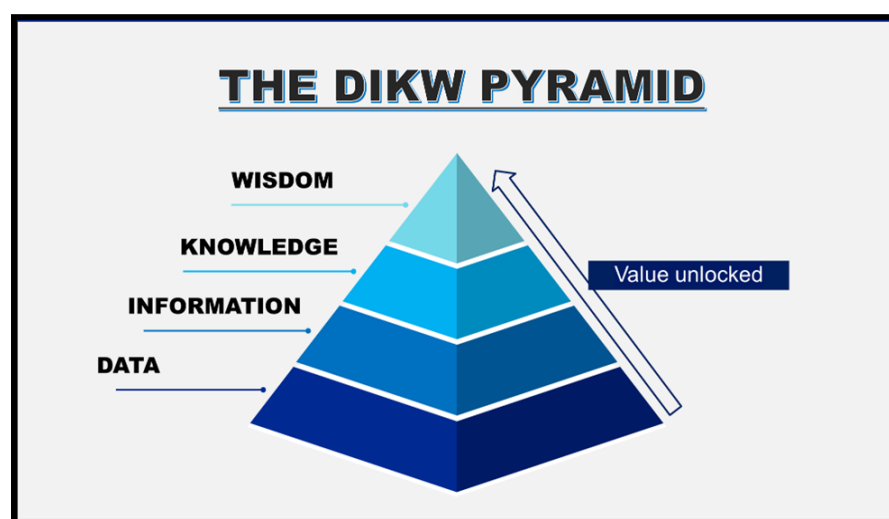
analyzing and interpreting data, data scientists can create information that is relevant and useful to a particular problem or decision. They can then use this information to gain knowledge about the problem or decision, and ultimately use this knowledge to make wiser decisions.

ETHICAL CONSIDERATIONS

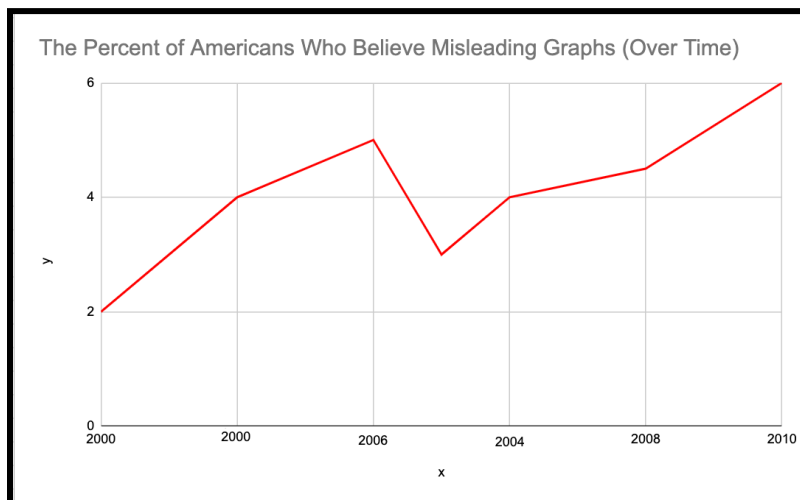
Ethics are a crucial consideration in the field of data science, encompassing the principles and values that guide the collection, analysis, and use of data. With the exponential growth of data being generated every day, it is important to ensure that data privacy is protected, and bias is avoided in data analysis. Transparency and accountability are critical to building trust in data-driven decision making and ensuring that the benefits of data are realized by everyone.

Data analysis and visualization offer the potential to solve a wide range of problems and answer a variety of questions. However, certain questions, such as ethical or moral ones, require a different kind of reasoning and decision-making that is **objective to humans rather than machines**. In addition, data analysis is limited by the quality and quantity of available data, as well as the accuracy of the models used to analyze it.

Data scientists are entrusted with the collection and analysis of large amounts of data that often contain sensitive and personal information about individuals. As a result, they have a crucial responsibility to conduct their work in an ethical manner. Furthermore, it is important for data



scientists to recognize these ethical implications of their work and ensure that the data they collect, analyze, and utilize is accurate and unbiased. This is essential because data has the potential to have



far-reaching ethical implications, underscoring the need for data scientists to be acutely aware of the impact of their actions.

[Figure 2: Lockman, 2023]

Data bias is a significant ethical consideration in data science. As data scientists work with datasets, they need to be aware that the data they use might include biases that can

influence their analyses and models. Biases can result from the data collection process or the presence of incomplete data. As a result, the models that are developed based on such data can be inaccurate or biased towards specific outcomes. To avoid these issues, data scientists need to ensure that they use unbiased and representative data and develop models that do not perpetuate or exacerbate existing biases. (Kangralkar, 2021)

Another ethical concern is the privacy of individuals. Data scientists must be mindful of the personal information they are handling and take steps to ensure that this information is not misused or mishandled. This includes obtaining informed consent from individuals before collecting their data and ensuring that the data is stored securely and only used for its intended purpose.

In the first chapter of "Everyday Adventures with Unruly Data," Feinberg delves into the ethical considerations that arise when working with data. She posits that data is not an impartial entity, but rather is molded by human decisions and biases, and thus, transparency regarding the sources and constraints of data is crucial.

[Figure 3: Tribune Content Agency, 2014]

She emphasizes the significance of considering how data may be utilized to reinforce or challenge power dynamics and inequalities, as well as the need for informed consent and the



potential risks that may arise from data collection and analysis. This is a noteworthy perspective, as Feinberg acknowledges the inevitability of data bias due to its human connection and potential for human error. (Feinberg, 2022).

INTERDISCIPLINARY AND COLLABORATIVE

Interdisciplinary and collaborative approaches are crucial to data science as it involves the integration of knowledge and techniques from multiple fields. Data science often requires collaboration among experts from different disciplines such as computer science, statistics, mathematics, domain-specific fields, and more. Collaboration between these experts is necessary

for the successful analysis of complex data and the development of actionable insights.

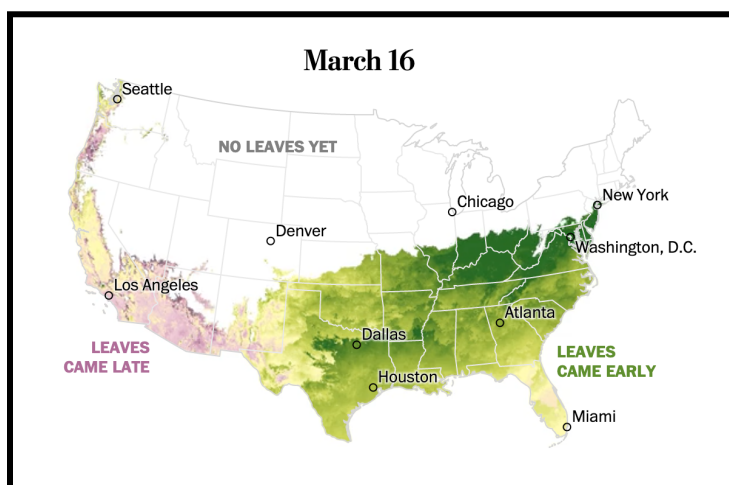
[Figure 4: Stevens, 2023]

Harry Steven's article in the The Washington Post, "When Will Spring Come?" exemplifies climate change and its impacts is a prime example of the interdisciplinary nature of data. As we know, climate change is a complex issue that requires the integration of data from various fields, including meteorology,

ecology, geography, and sociology – just to name a few. The article presents data in various forms, including maps, graphs, and interactive visualizations, to help readers understand the interconnectedness of different factors affecting climate change, such as temperature, precipitation, sea level rise, and human activity. The data used in this article has been collected and analyzed by experts from multiple fields, demonstrating how collaboration and sharing of data across disciplines is essential to addressing complex problems like climate change.



[Figure 5: Stevens, 2023]

Furthermore, the "check my work" section of the article also relates to the collaborative nature of data and data science. This section allows readers to verify the author's



Check my work

- ◆ The data in this article comes from the USA National Phenology Network, the National Oceanographic and Atmospheric Administration, and Mark D. Schwartz of the University of Wisconsin at Milwaukee, who emailed me the data file. I got the list of U.S. city coordinates from [this GitHub repository](#).

The code I wrote to produce the map of leaf arrival timing is in [this computational notebook](#) . The code and data to produce the chart comparing leaf index and temperature from January to May is in [this computational notebook](#) .

You can use the code and data to produce your own analyses and charts — and to make sure mine are accurate. If you do, email me at harry.stevens@washpost.com, and I might share your work in my next column.

calculations and ensure the accuracy of the data presented. This highlights the importance of transparency and openness in data analysis and emphasizes the collaborative efforts of the data science community. The use of platforms such as GitHub and Observable further emphasizes the collaborative approach, as they enable users to share and build upon each other's work. This ultimately encourages cross-disciplinary collaboration and fosters a more expansive and inclusive data analysis. (Stevens, 2023).

CONTINUOUS GROWTH

Continuous learning in data science involves staying up-to-date with new technologies, tools, techniques, and best practices. Machine learning frameworks, data visualization tools, and big data platforms are constantly evolving, and data scientists must stay current with these technologies to use them effectively. The field of data science is also developing new techniques and algorithms for analyzing data, and data scientists must stay current with these developments to use them effectively and remain competitive.

Moreover, data sources are constantly evolving, with new types of data becoming available and existing data sources changing. Staying current with these changes to ensure that they are using the most relevant and accurate data in their analyses. Additionally, industry trends and best practices shape the field of data science, and data scientists must stay current with these trends and best practices to ensure that they are following ethical and effective data practices.

An immediate example of continuous learning in data science is the use of AI. AI is constantly evolving, and data scientists must stay current with these developments to use AI effectively in their

work. For example, the field of natural language processing (NLP) has seen rapid advancements in recent years, with new techniques and algorithms for processing and analyzing language data. Data scientists who work with NLP must stay current with these developments to ensure that they are using the most effective and ethical techniques. (Hirschberg, 2015).

[Figure 6: Stevens]

In her book, "You Look Like A Thing and I Love You", Janelle Shane highlights how machine learning algorithms learn from data and use it to improve their performance over time. This process encapsulates continuous learning, as the

BASIC CLAM FROSTING

main dish, soups
1 lb chicken
1 lb pork, cubed
½ clove garlic, crushed
1 cup celery, sliced
1 head (about ½ cup)
6 tablespoon electric mixer
1 teaspoon black pepper
1 onion—chopped
3 cup beef broth the owinls for a fruit
1 freshly crushed half and half; worth water

With pureed lemon juice and lemon slices in a 3-quart pan.
Add vegetables, add chicken to sauce, mixing well in onion.
Add bay leaf, red pepper, and slowly cover and simmer covered for 3 hours. Add potatoes and carrots to simmering. Heat until sauce boils. Serve with pies.

If the liced pieces cooked up desserts, and cook over wok.
Refrigerate up to ½ hour decorated.
Yield: 6 servings

algorithms must continuously analyze new data and adjust their parameters to enhance their accuracy and effectiveness. The development of an AI system that learns to generate new recipes or a knock knock joke demonstrates how continuous learning is crucial in data science. (Shane, 2019). All data scientists can start with a baseline level of accuracy, but with time and dedication, they can develop their skills and add new tools to their toolbox, allowing them to produce more and more accurate results at a faster pace.

I personally experienced this during the course of a semester, where I dedicated myself to continuous learning and was able to add new skills to my repertoire. By the end of the semester, my skills had significantly improved, surpassing my past self's capabilities. I believe that part of being a data scientist is to be adaptable and interactive with the skills and practices you learn.

In conclusion, the principles of decision making, interdisciplinarity and collaboration, ethical considerations, and continuous growth are essential to the practice of data science. By adhering to these principles, data scientists can leverage their skills and knowledge to tackle complex data sets and extract meaningful insights that can be used to make sense of the vast amounts of data available today, paving the way for a better future powered by data-driven insights. It is important to recognize that the field of data science is still evolving, and as such, there will continue to be new challenges and opportunities that arise in the coming years. By staying current with the latest developments and continuously learning, data scientists can remain at the forefront of this exciting field and help shape the future of our world.

On a final, personal note, I find it an honor to be involved in such a complex field of study. As stated earlier, my class on Data Structures integrated how computers understand data and while in contrast, this class on Data Science tackled how humans interpret and interact with data. Both subjects have shown me the immense power of data and its potential to drive innovation and progress in various ways. I encourage anyone interested in this field to continue exploring and learning, as the possibilities for data science are endless. With the right skills, tools, and mindset, we can use data to solve complex problems, improve decision-making, and ultimately make a positive impact on the world.

References

- Elliott, Timo. "Analytics Cartoons." Timo Elliott's Blog, <https://timoelliott.com/blog/cartoons/analytics-cartoons>. Accessed 10 May 2023.
- Feinberg, Melanie. "Everyday Adventures with Unruly Data." MIT Press, 2022.
- Hirschberg, Julia, and Christopher D. Manning. "Advances in Natural Language Processing." *Science*, vol. 349, no. 6245, 2015, pp. 261–266, <https://doi.org/10.1126/science.aaa8685>.
- Jill Lepore, "The Data Delusion." *The New Yorker*, 2023.
- Kangralkar, Swapnil. "Types of Biases in Data." Medium, 26 Aug. 2021, towardsdatascience.com/types-of-biases-in-data-cafc4f2634fb.
- Lupi, G. (2019). Data Humanism: The Revolution will be Visualized. Retrieved from <http://giorgialupi.com/data-humanism-my-manifesto-for-a-new-data-world>.
- Shane, Janelle. "You Look Like a Thing and I Love You." Voracious/Little Brown, 2019.
- Stevens, Harry. "When Will Spring Come? Or Has It Already? Look up Where You Live." *The Washington Post*, 16 Mar. 2023, www.washingtonpost.com/climate-environment/interactive/2023/spring-early-late-climate-change/.
- Raddaoui, Omar. "Supercharge Your UX Pursuits with the DIKW Framework." *Medium*, 6 Feb. 2023, uxplanet.org/supercharge-your-ux-pursuits-with-the-dikw-framework-9fa87b2586be.