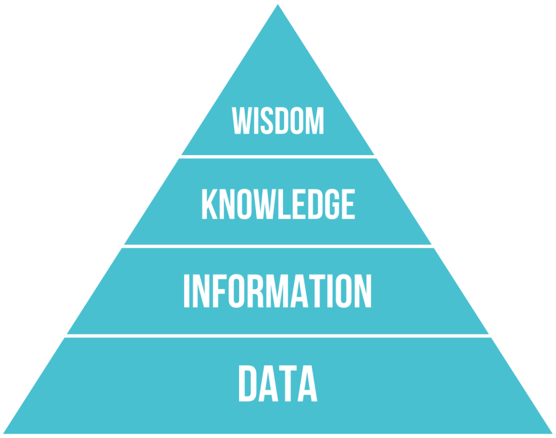
Data Manifesto:
Creating a Contemporary Data Scientist

Bernie Jones
CS 215
Professor Jordan Wifrs-Brock
Fall Semester 2023

WHY?

The Data, Information, Knowledge, Wisdom (DIKW) Pyramid provides a loose categorization of the underpinnings of contemporary data-science, representative of the interdisciplinary approach that the discipline is becoming increasingly challenged with. This holistic uprising emphasizes contextualizing the data, championing critical thinking, and questioning of not only collected data but also its source. Contemporary data science asks us questions of accessibility. and questions of inclusivity (and exclusivity). It acknowledges the human-centric aspect of a discipline currently polarized by growing concerns, and imparts the understanding that much like humanity, data is imperfect. In these shifts in thinking, our interactions with data reflect issues of social, economic, political ideology, and beyond, all at a time where recognition of these factors is becoming increasingly important as data continues to drive advancement. Whether you're a seasoned professional attempting to broaden your horizons, or a novice wanting to learn more, to a college student in an introductory data-science course, the principles presented in this manifesto will hopefully serve you well in your individual journey into the broad and ever expanding discipline that is data-science.



Acknowledge the ethical responsibilities that come with data-science.

In every project that you undertake, study, or come across, question the motivations that underlie it. What is missing? How would you do things differently? Ask yourself, what biases am I bringing to this project and my interactions with it? These questions are incredibly important to ask yourself as ‘data backed results’ drive our modern world. *Human Centered Data Science* posits, “the social impact of a data-science system can be broad and profound. We hope that you will take to heart the phrase "these are people" when you consider how your system and methods will affect others”.[1] By acknowledging the impact that a potential project has on society, this begs us to ask ourselves critical questions that ultimately contextualize the project. By understanding the factors that may be influencing our actions, we can begin to attempt to combat them. Much like sometimes our words can have unintended consequences so can the projects we create. By holistically approaching a project before undertaking it, we can address potential critiques before they occur. Further, this again reinforces the nature that ultimately our data-science projects are all grounded in a relationship with some aspect of humanity, even if this may seem far removed. *‘Raw Data’ Is an Oxymoron* theorizes this belief, “looking at the ways scientific knowledge is produced-rather than innocently “discovered”, for instance - resembles our project of looking into data or, better, looking under data to consider their root assumptions”.[2] Discovery does not arise from thin air, it comes from human centered curiosity on

---

[1] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. *Human-Centered Data Science*. MIT Press, 2022. (151).

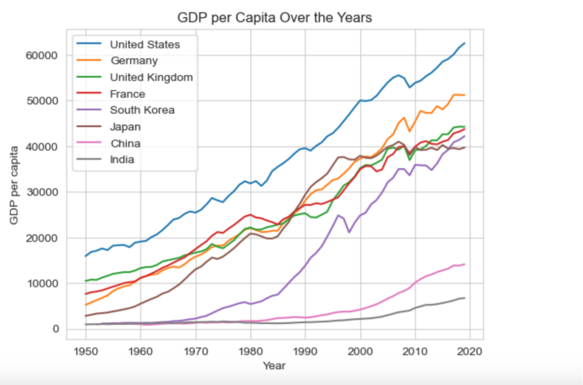
[2] "Raw Data" Is an Oxymoron, edited by Lisa Gitelman. MIT Press, 2013. (4).

some level. Humanity is flawed by its own existence and development, thus we must acknowledge our shortcomings and how this may be reflected within our work, even when attempting to be as careful as possible. Question the mysterious ‘black box’ nature of projects, even during the creation of your project, comment out your code, provide context, embrace the ethical responsibilities of transparency in your work, embrace criticism, and embrace the potential magnitude of your work.

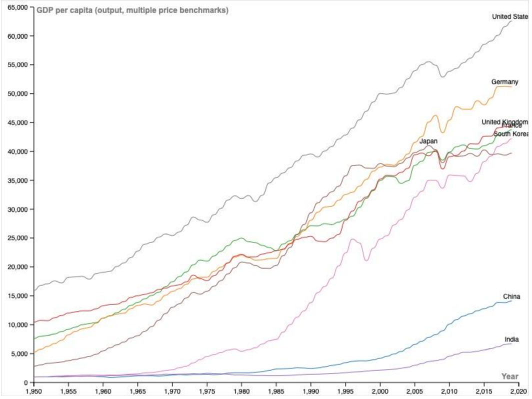
Strive for accessible communication in your project.

Both before and during our project process, we must always maintain an understanding of our intentions for what our project is attempting. This again returns to the importance of acknowledging our ethical responsibilities. We now shift our responsibility to the visual representations of a project. While infographics dominate contemporary media, oftentimes recognized for being ambiguously vague or hollow, by remaining grounded in our communication of our project we can deliver an accessible product.[3] *Human Centered Data Science* confirms “data needs to be imaged as data to exist and function as such, and the imagination of data entails an interpretive base”.[4] As interdisciplinary data-scientists, we must ask ourselves what story am I attempting to convey. By framing our project into the lens of crafting a story for our analysis, we improve not only the communication of our process but also remain rooted in this human-centered approach, “stories help us remember that communication is a process that involves others, not just ourselves and our intentions”.[5] Whether this communication is in the form of markdown cells explaining data, comments of code, or visualizations, each segment of our project requires the same ethical critiquing as our pre-project acknowledgements. Additionally, Label things, clarify them, add legends, develop an engaging project that remains transparent in its approach. There are a multitude of ways in which this can occur via visualization tools, packages, and more. We can even visualize the same analysis in different ways. In this day and age ,there’s little reason to not clearly display things. By the completion of your project it should be accessible enough to be reproducible, with the idea that it could be treated as open source! As *Data Humanism, The Revolution will be Visualized* asserts, “we should learn how to include and render the more qualitative and

Seaborn:



RawGraphs:



---

[3] Lupi, G. (2017, January 30). *Data humanism: The revolutionary future of*
https://www.printmag.com/article/data-humanism-future-of-data-visualizat

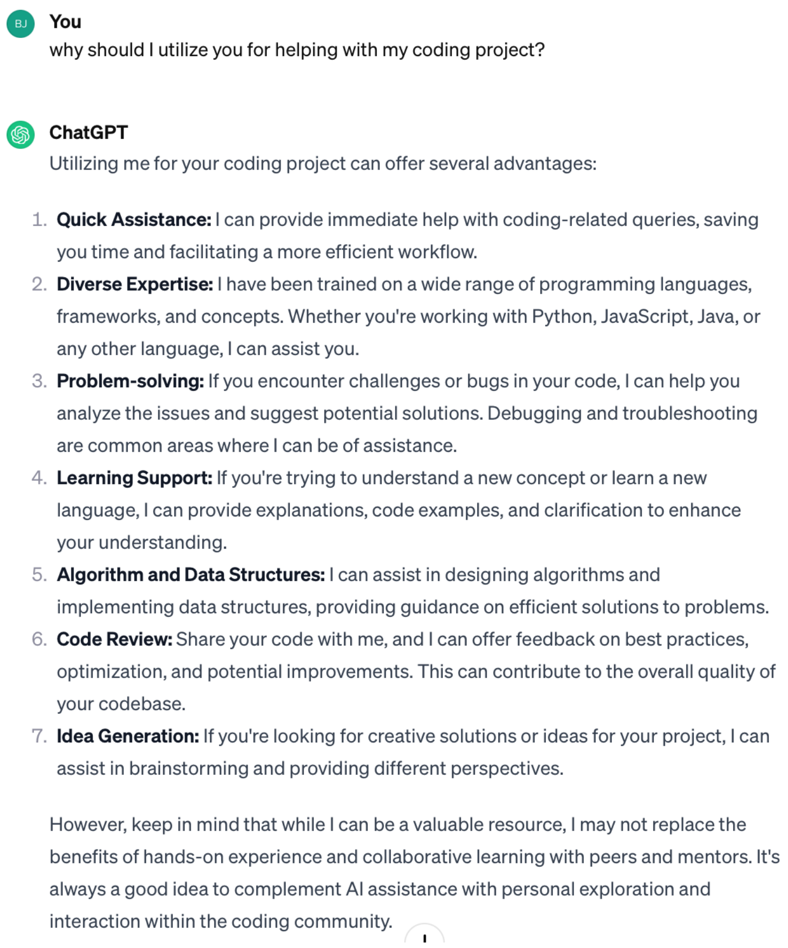
[4] Cecilia Aragon (150).

[5] Cecilia Aragon (150).

nuanced aspects of data. We should experiment with how to visualize uncertainty, possible errors and imperfections in our data".[6] Striving to communicate these details is not only a responsibility of today's data-scientists, but an engaging and thought provoking activity.

Embrace technology for the sake of efficiency, but do not forsake efficiency for understanding.

Today, many aspects of the data-science project process can be streamlined through innovative tools on the backs of machine learning or artificial intelligence, offering quick solutions to complex problems. Yes, you can leverage OpenAI's Chat-GPT to help you correct a line of code, yes you can even leverage it to write out an entire project.



You
why should I utilize you for helping with my coding project?

ChatGPT
Utilizing me for your coding project can offer several advantages:

1. **Quick Assistance:** I can provide immediate help with coding-related queries, saving you time and facilitating a more efficient workflow.
2. **Diverse Expertise:** I have been trained on a wide range of programming languages, frameworks, and concepts. Whether you're working with Python, JavaScript, Java, or any other language, I can assist you.
3. **Problem-solving:** If you encounter challenges or bugs in your code, I can help you analyze the issues and suggest potential solutions. Debugging and troubleshooting are common areas where I can be of assistance.
4. **Learning Support:** If you're trying to understand a new concept or learn a new language, I can provide explanations, code examples, and clarification to enhance your understanding.
5. **Algorithm and Data Structures:** I can assist in designing algorithms and implementing data structures, providing guidance on efficient solutions to problems.
6. **Code Review:** Share your code with me, and I can offer feedback on best practices, optimization, and potential improvements. This can contribute to the overall quality of your codebase.
7. **Idea Generation:** If you're looking for creative solutions or ideas for your project, I can assist in brainstorming and providing different perspectives.

However, keep in mind that while I can be a valuable resource, I may not replace the benefits of hands-on experience and collaborative learning with peers and mentors. It's always a good idea to complement AI assistance with personal exploration and interaction within the coding community.

Yet, in all of these quick solutions, understanding can be lost. Documentation platforms or open-threads such as *Stack Overflow* reflect years worth of dialogue across individuals, that now often with a quick well written prompt can answer even more complex questions in a matter of seconds via Chat-GPT. The instant gratification, clear coherent structure, and malleability of a custom prompt all are great benefits that point towards why this may be an effective solution. Embrace this! Just don't let it be your one tool kit in your arsenal. Do not overly rely on technology to fill gaps on each and every project that you do. For example, let's turn to an 'older' piece of technology: the calculator. The calculator represents a simplified analogy for the use of Chat-GPT in a myriad of ways. First, a calculator's computation is only as good as its inputs. This slightly differs with Chat-GPT as a poor prompt can be parsed and interpreted via its algorithm, yet still the result may not be the intended outcome (this also

---

[6] Lupi, G.

violates our other principles as it induces potential bias we must address due to the algorithm for Chat-GPT). Without a holistic understanding of a query's response, very quickly a project can reflect solutions that you do not entirely comprehend. Secondly, let's think of doing basic arithmetic using a calculator. Sure there is value in leveraging a calculator for these computations, but also your own knowledge can serve you just as well and even save you time. Really what we are looking for is embracing the fine line between effectively leveraging existing architecture and over-relying upon it. Ask yourself, is this something that I fully understand? What are the gaps or biases presented in this solution? The tools present today weren't here before and offer great insight into the relationship with us humans and technology. Embrace your resources, mastery of 'resource querying' and 'prompt creation' will continue to be essential skills moving forward. Embrace this, hone these skills, but do not lose the scope of your project, or yourself in this ease of use. Maintain the ethical responsibilities to your project. Embrace the struggle at times, and know that resources exist to help.

Remember, learning is continuous, enjoy it!

Each day is a new opportunity to learn and this extends beyond just data-science projects. Our world is constantly evolving, and within the data-science industry trends, requirements, and expectations are only increasing. Become a connoisseur of knowledge. This can come in various manners throughout all facets of the project cycle for example. This may be researching bias in algorithms, familiarizing yourself with new syntax, learning a new language or technique, the avenues for growth are continuous. Be comfortable with knowing that there will be unknown, and don't shy away from this. Refer to your resources, refer to your experiences, and remain enthusiastic for the challenge of critical thinking about your own work and others. Much like learning, the possibilities of data arise from our motivation, maintain this, allow it to build within you, let it guide you to interests and passions that in turn will build your knowledge base. By maintaining the connection of data-science as a human-driven endeavor, we can go beyond the concepts of data as just a "factual reality".[7] Shirk the idea of the 'black box', strive to understand both context and model building, and in turn leverage these in your own projects.

datacamp

Go on as a data-scientist.

While these principles will hopefully serve you well in becoming a critically-thinking, interdisciplinary data-scientist, we must again acknowledge the influence of this discipline upon mankind. *Algorithms of Oppression* soberingly reminds us, "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".[8] These principles all arise from issues such as the

---

[7] "Raw Data" Is an Oxymoron.

[8] Sofia Noble. *Algorithms of Oppression. (2018).*

aforementioned. A continued emergence of individuals incorporating a cross discipline approach to their work champions the accountability for the products' built by data-scientists. Afterall, data is human, it is imperfect, it is you, it is I, it is all around us.

Works Cited

Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. *Human-Centered Data Science*. MIT Press, 2022. (151).

Lupi, G. (2017, January 30). *Data humanism: The revolutionary future of data visualization*. PRINT Magazine. https://www.printmag.com/article/data-humanism-future-of-data-visualization/

"Raw Data" Is an Oxymoron, edited by Lisa Gitelman. MIT Press, 2013. (4).

Sofia Noble. *Algorithms of Oppression. (2018).*