Dia Dana
Professor Jordan Wirfs-Brock
CS/Math 215: Intro to Data Science

Data Manifesto

Data is immortal—ever-present, alive and everywhere. The dimension of a table, the archives of the library, the geometry of a tree, and the natural law of physics of this universe that is manifested in everything physical. Data has always been there long before humans existed - untouched and uncollected; from the moment we arrived, we began to interpret and collect it. Like the inevitability of a dropped stick striking the floor, certain patterns in nature unfold according to rules that transcend our prescriptions.

Albrecht Dürer's *Melencolia I* captures this tension perfectly. In the foreground, an angel wields advanced tools and intricate calculations, striving to decipher the unknown. Yet in the background, a divine presence hints at the limits of human understanding and the infinite mysteries beyond our reach. Data falls into place like those instruments of the angel, that of which it powers the humans' instruments and devices to figure out the unknown. Just as Dürer's angel grapples with tools that both illuminate and fall short, so too does modern data science, the most advanced instrument, push against the boundaries of the unknown. However advanced it may be, humans might never be able to grasp the unknown.

This manifesto is built on the backbone of the tension between the melancholy of the unknown infinity and data science at its peak pushing against this finite human boundary. I will translate this tension into four core principles:

1. Question your project's ethicality and consequences - everything has some impact
2. Get the basic skills under the belt
3. Know and learn deeply the technical required skills for every project
4. Put curiosity with humility at the heart of every project

By practicing these four principles, you will honor data's immortality: first by gathering raw observations and contextualizing them into meaningful information; next by analyzing that information to derive reliable facts and knowledge; and finally by applying those insights to real-world problems—always guided by curiosity and humility.

Starting off with the first principle: before every project, the data scientist must understand every consequence of their project- big or not. Data shapes a lot of decision making and a lot of times it could be biased and used for other unintended things. A great example of this is talked about by Cathy O'Neil in her third chapter "Arm Race" from Weapons of Math Destruction. Cathy mentioned how the *US-News* releases the rankings of colleges across the

country on different domains - graduation rate, living cost, SAT score, etc -, which U.S. News then with the help of an algorithm weights into an overall ranking. Despite it seeming harmless, this created a vicious cycle for colleges that ranked low. Ranking low as an institution meant lower admissions and students, so therefore even lower their ranking more. This example shows how a well-intentioned project can yield “lots of harmful unintended consequences” (Cathy O'Neil, “Arm Race”). Keeping this principle in mind is essential to ensure a successful project to the real world with no unintended consequences with the further principles.

The second principle—ensuring foundational skills are learned—is particularly important before the third principle. First of all, programming is a needed skill to start as a data scientist. Being an expert at Python/R will make the process much smoother. Next, and getting more specific to data science, data wrangling is the most time-consuming and essential skill to learn and get familiar with—finding a dataset, understanding it, uploading, cleaning, filtering, and preprocessing it for analysis. Visualization is then the skill to showcase the findings of your analysis. This phase is essential and generally uses libraries like Matplotlib and Seaborn. The learning process is also divided into multiple phases of trying out different graph types on the appropriate dataset and becoming comfortable with them. It is also important to understand other tools for building visualizations, such as Tableau and Microsoft Power BI, which allow interactive dashboards for broader audiences. A final skill to learn is how to communicate your findings clearly. It is useless to uncover insights if you cannot convey them effectively. This is an important phase at the end of the cycle to finalize the project. Despite these foundational skills, there are still more competencies needed for bigger and more complex analytical projects, which will be covered by the next principle.

Moving forward with the third principle, knowing and practicing the technical skills needed for a more complicated project is essential. Starting a project without the necessary technical foundation leads to confusion. You’ll spend extra time juggling learning and building, which slows progress. A personal example of this was during my project – Relational data using SQL & BigQuery. I had a very limited knowledge of SQL, so the project became tiresome and inefficient. If I knew the SQL basics before starting, I would have had a smoother process. On the other hand, learning the skills prior to starting makes the process much smoother and easier. Before giving a small guide to a more successful project, learning how to learn as a data scientist is essential despite the level of expertise as this kind of science is evolving everyday with new skills to learn. To learn the skill beforehand: get more familiar with the skill using resources like youtube videos and blogs; then start another smaller project to practice what you understood from the resources, just like the APIs and Web Scraping project that really translated what we learned from DataCamp to a small project for a deeper learning experience; and finally apply those skills learned to the actual analysis. Investing a few hours to solidify key skills saves days of trial-and-error and keeps your project on track. Sounds like a lot of work? With the third

principle, however, hard work will feel less like a grind and more like discovering the unknown infinite!

The final principle, by far the most important, is to put curiosity as the core motive during and towards every data science project with a balance of humility. Curiosity also falls down naturally during the first cycle of data science when formulating a question. It is best to wonder and formulate a question during that first phase to make sure no burnout occurs during the other phases. Building a project that sparks interest and touches the wonder mind will give a much smoother, consistent and joyful experience. On a separate note, data science typically answers empirical questions. However, this should not limit curiosity and still aim to bridge between the empirical and nonempirical truths with a humble approach that humans might never be able to grasp the divine unknown. Questions should not have a limit when formulated, but should remain humble.

In embracing these four principles—ethical foresight, foundational mastery, deep curiosity, and humble wonder—you honor data's timeless power while respecting its limits. By asking hard questions up front, building rock-solid skills, letting curiosity guide your investigations, and staying humble before the vast unknown, you can turn raw observations into meaningful insights that serve real people and real problems. In doing so, you not only unlock data's immortality, but you also ensure that your work remains responsible, impactful, and ever-attuned to the mysteries that lie just beyond our grasp.