Data Manifesto

Meron Semere

Zooming In on What Matters:

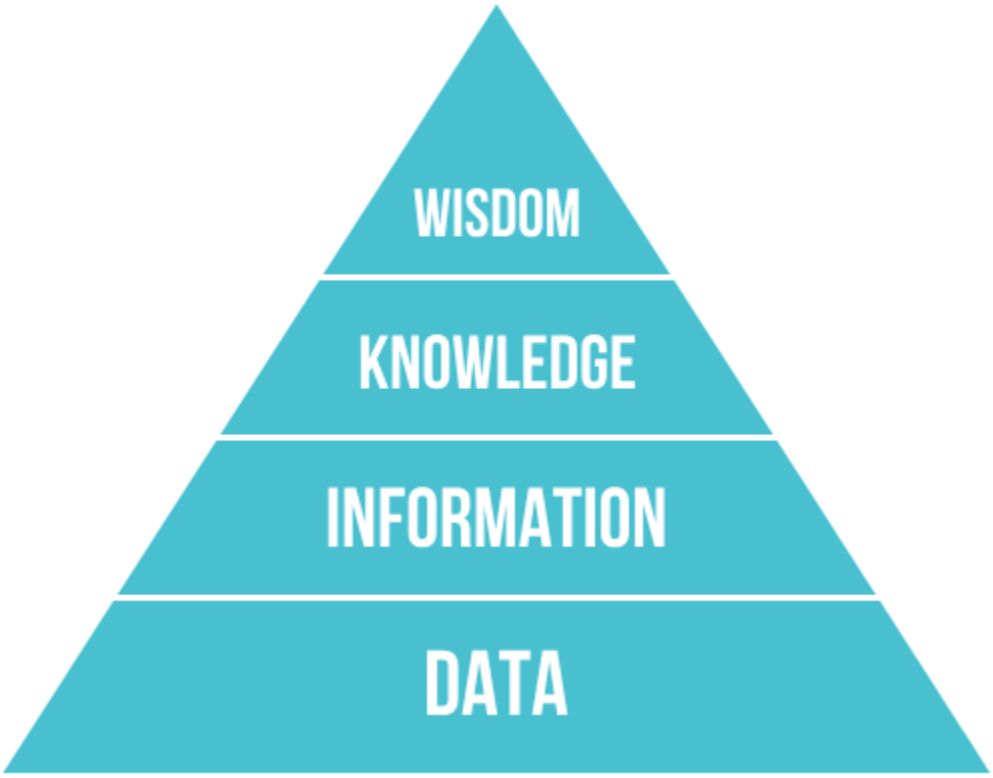
Principles for Thoughtful Data Work

I. Making Sense of Data

Data can feel overwhelming, and that is okay. It is everywhere and often messy, but that mess can become exciting when we break it down into manageable bite-sized pieces. By doing so, we make data more approachable and open up space for meaningful stories and insights to emerge from our work.

At its core, data is just raw input— numbers, symbols, or observations —that only becomes valuable when organized into meaningful information. When that information

is applied in context, it turns into knowledge, and then wisdom when used to make sound decisions. This is best described through the DIKW Pyramid (Data → Information → Knowledge → Wisdom), which reminds us that good data science is not just about writing code or crunching numbers but it's about interpretation, context, and making a meaningful impact on people and communities.



Another framework for thinking about how we move from data to meaning comes from Jill Lepore's filing cabinet metaphor. In this model, drawers represent broad categories of knowledge, folders within drawers represent specific topics, and the individual files inside stand for pieces of data. Lepore notes that the drawers are labeled "Mysteries,"

“Facts,” “Numbers,” and “Data”, with the most unknowable things like death or the reasons behind systems of power placed at the top.

> frame, just above the drawer pull. The drawers are labelled, from top to bottom “Mysteries,” “Facts,” “Numbers,” and “Data.” Mysteries are things only God knows, like what happens when you’re dead. That’s why they’re in the top

This metaphor shows that data is not neutral or complete. It reflects decisions about what gets collected, preserved, or left out. In a time when misinformation and selective truth-telling can shape public understanding, especially during politically charged moments like the Trump era, this framework reminds us that data must always be questioned, contextualized, and interpreted with an awareness of the power structures behind it.

II. The Role of a Data Scientist

Moving on from the DIKW Pyramid, we can start to reframe our role as data scientists—not just as coders but as translators and storytellers. If we cannot communicate our analysis to people in a way they understand, then the insight is useless. That’s where storytelling comes in. Good data science connects messy numbers to real context, helping people—whether clients, policymakers, or general public—make informed decisions. Storytelling bridges that gap and gives our work purpose.

III. Guiding Principles of Data Science Work

- **Flexibility:** Be open to changes in project direction. Sometimes, the data leads you somewhere unexpected and that's where insights lie.
- **Critical Thinking:** Don't take datasets at face value. Ask who collected the data, when, how, and why.
- **Bias Awareness:** All data contains assumptions. Check for sampling errors, exclusion, historical inequalities, and algorithmic bias.
- **Simplicity Meets Complexity:** While data science can deal with large, messy systems, clear structure and manageable sub-questions help make it digestible.
- **Storytelling Through Visualization:** Communicate insights clearly using thoughtful charts, maps, and graphics that engage a broad audience.

Flexibility

- **Let the data guide you:** Be willing to shift your research question or approach based on what the data reveals. Unexpected trends often uncover the most valuable insights.
- **Adapt to challenges:** Data work doesn't always go as planned. Missing values, or inconsistencies may require pivoting your methods or research focus.

- **Stay iterative:** View analysis as an evolving process. Revisiting earlier steps or rethinking your assumptions is not a setback but rather part of doing thoughtful work.

Critical Thinking

Understanding the context of a dataset is essential as it may contain hidden biases or structural issues. Without examining the data carefully before analysis, there is a risk that your findings could be misleading others or create unintentional harm.

Some key questions to consider when examining a dataset include:

1. For what purpose was this dataset created?
2. Who created the dataset?
3. Who funded its creation?
4. Are there any missing voices or perspectives?

As data practitioners, we have a responsibility to uphold ethical standards throughout the entire process from critically assessing the dataset to wrangling, transforming, and ultimately communicating the results.

To illustrate this approach, let's take the College Scorecard dataset from earlier in the semester as an example:

1) For what purpose was this dataset created?	The College Scorecard was developed to provide students and families with accessible and reliable information about college costs, graduation rates, student debt, and post-college earnings. Its primary goal is to promote transparency and help people make more informed decisions about higher education.
2) Who created the dataset?	It was created by the U.S. Department of Education.
3) Who funded its creation?	The dataset is publicly funded by the U.S. federal government through the Department of Education.
4) Are there any missing voices or perspectives?	Yes. The dataset lacks qualitative information, such as student satisfaction, campus culture, and lived experiences especially from marginalized groups. It also does not capture long-term outcomes such as career fulfillment. Also, it may be less relevant for non-traditional students.

Bias Awareness

Bias is an important consideration in any data project and recognizing how it influences results is essential to doing responsible work. One of the most common and often overlooked types is selection bias which occurs when the sample within the data does not accurately represent the broader population. This can lead to conclusions that are misleading or incomplete.

Selection bias can arise from the way data is collected, who is included or excluded, or even from mistaken assumptions about what the data reflects. If we don't pause to ask whether the dataset truly represents what we think it does, we risk drawing conclusions that don't hold up.

In *Calling Bullshit*, Carl Bergstrom and Jevin West explain how selection bias can subtly shape results, producing patterns that seem convincing at the surface level but fail to reflect the full reality. In the excerpt shown below, they illustrate this with a classroom example, even if only a few classes are large students may report disproportionately high class sizes simply because they are more likely to experience those larger settings. This example highlights how averages can distort lived experience and why understanding how data is structured is important to interpreting it correctly.

serve a disproportionately large number of students. Suppose that in one semester, the biology department offers 20 classes with 20 students in each, and 4 classes with 200 students in each. Look at it from an administrator's perspective. Only 1 class in 6 is a large class. The mean class size is [(20 × 20) + (4 × 200)] / 24 = 50. So far so good.

But now notice that 800 students are taking 200-student classes and only 400 are taking 20-student classes. Five classes in six are small, but only one student in three is taking one of those classes. So if you ask a group of random students how big their classes are, the average of their responses will be approximately [(800 × 200) + (400 × 20)] / 1,200 = 140. We will call this the *experienced mean class size,*[*4] because it reflects the class sizes that students actually experience.

Bias is not limited to sample selection. It can also arise from flawed measurement tools, outdated historical sources, or the framing of research questions. Even algorithms, which are often assumed to be neutral can replicate systemic inequities.

In *Algorithms of Oppression*, Safiya Noble examines how search engines and other technologies can reinforce discrimination. In the excerpt below, she discusses how algorithmic systems often lack social and human context which can disproportionately harm marginalized communities. She explains that search results can cause harm towards marginalized groups. These patterns are not random rather they are tied to historical inequalities that are deeply embedded in both media and technology.

> While organizing this book, I have wanted to emphasize one main point: 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. It is of particular concern for marginalized groups, those who are problematically represented in erroneous, stereotypical, or even pornographic ways in search engines and who have also struggled for nonstereotypical or nonracist and nonsexist depictions in the media and in libraries. There is a deep body of extant research on the harmful effects of stereotyping of women and people of color in the media, and I encourage readers of this book who do not understand why the perpetuation of racist and sexist images in society is problematic to consider a deeper dive into such scholarship.

Failing to address bias can have consequences that extend far beyond the dataset. Biased data can misinform decision-making and this is of particular concern in areas that impact large numbers of people such as public policy, healthcare, or urban planning. This not only deepens existing disparities but also risks eroding public trust in the institutions that rely on data to serve communities. That's why recognizing and challenging bias is not just a technical concern but it is a foundational part of ethical and responsible data work.

Simplicity Meets Complexity

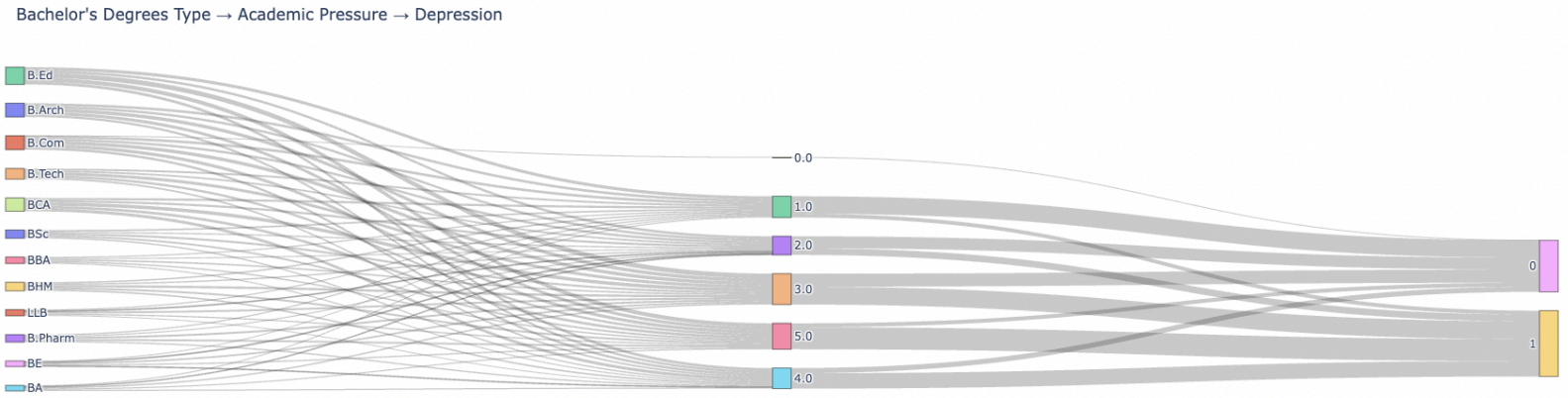
Working on a data project can feel overwhelming and it's not always easy to know where to begin. However, there are a few key practices that can help keep you focused and successfully carry out your work:

- Set clear goals
    - Begin by identifying what you want to communicate through your project. Having one main question or goal can help guide your focus and developing smaller sub-questions can support your exploration of that larger idea. This approach can lead to deeper insights, uncover meaningful patterns, and result in more engaging and compelling visualizations.
- Document your process
    - Comment not just big pieces of code but also the small steps like creating a new data frame, dropping rows with missing vales, or converting text into a query. Thorough documentation allows us to easily revisit and improve our work later on. It also allows others to follow our thinking and even build upon our analysis in new directions.
- Implement a realistic timeline
    - Create a timeline to help manage your project in simple manageable stages. Breaking the work into smaller steps reduces stress and helps you stay organized. By dedicating specific time blocks to each task, you can work more thoroughly and allow space for critical examination of potential issues like bias. This also gives you time to incorporate additional datasets or seek help if roadblocks arise.
- Communicate

    - It’s normal to hit roadblocks that make it feel like your project might not work out, but communication is essential for moving forward. Reaching out for support, whether from classmates, professors, or teammates can help you troubleshoot issues, discover new perspectives, and find alternative solutions that ultimately strengthen your work.

Storytelling Through Visualization

- Don’t be afraid to explore new techniques. Sticking to traditional visualization methods like histograms or bar charts may limit your ability to find more compelling ways to engage your audience. New techniques might be more effective than you expect just embrace the learning curve as part of the creative process. There are many resources available to help you learn such as Stack Overflow or documentation guides for Python and libraries like Seaborn or Plotly.
- Consider your audience. Do they have a background in data science or statistical modeling? Keeping this in mind can help you present your visualizations in a way that’s both clear and easy to understand. This can also shape the direction you take with visualizations.



One example that showcases this approach is the Sankey diagram from Project 9 which focuses on depression outcomes among college students. It shows how different bachelor's degree types connect to levels of academic pressure (rated from 0 for low pressure to 5 for high) and how these factors relate to whether students reported experiencing depression symptoms (with **0** meaning no symptoms and **1** meaning symptoms were reported). Using a Sankey diagram instead of a more traditional chart like a bar graph helps illustrate how variables flow and interact across categories. Exploring techniques like this can make the story behind the data more engaging especially when you're communicating with an audience that may not have the same level of familiarity with the data as you do.

IV. What Sets Great Data Scientists Apart

The best data scientists combine technical expertise with curiosity, a willingness to learn and adapt, and understanding of the people their work affects.

- **Technical Mastery**: Proficient in tools such as Python, SQL, and statistical modeling with the ability to clean, manipulate, and analyze data.
- **Problem Framing and Critical Thinking**: Skilled at asking the right questions, structuring problems clearly, and apply appropriate analytical methods to guide data exploration and interpretation.
- **Curiosity and Adaptability**: Eager to explore new techniques, learn from failure, and adapt to challenges and setbacks.
- **Ethical and Social Awareness**: Recognizes the broader impact of data work on individuals, communities, and institutions, and incorporates these considerations into responsible decision-making within data work.
- **Communication**: Able to transform complex analyses into compelling stories and visuals that inspire action.

Applying My Guiding Principles to a Hypothetical Project

One hypothetical project I would be interested in exploring focuses on how local tax policy impacts the distribution of public goods in historically underserved communities. Using datasets and reports from the Urban Institute's Tax Policy Center, I

would examine how differences in property tax rates, local revenues, and service investments correlate with demographic and geographic indicators such as income, race, and zip code.

To carry out this project responsibly and effectively, I would apply each of my guiding principles as follows:

- **Flexibility:**
    - Begin with a broad question about the relationship between tax policy and public services.
    - Allow the data to shape and potentially refine my focus. For example, shifting from general tax revenue analysis to a closer look at housing or education funding if patterns emerge.
- **Critical Thinking:**
    - Investigate who collected the local policy tax data, how communities were counted in the population (i.e. race), and whether certain groups are underrepresented. For example, undocumented immigrants may be undercounted which could affect how resources are allocated or how tax burdens are assessed.
    - Ask whether the dataset includes informal or overlooked economic activity that may affect local taxation and spending.
- **Bias Awareness:**

    - Examine the structural factors behind why some areas receive more funding than others, considering how redlining, disinvestment, or zoning laws might be reflected in the data.
    - Watch for misleading averages that hide inequities across different racial or income groups.
- **Simplicity Meets Complexity:**
    - Break the analysis into digestible steps such as one stage for comparing tax burdens and another for mapping investments.
    - Use clear sub-questions like: "How does per-capita spending on public education differ by income quartile?"
- **Storytelling Through Visualization:**
    - Use interesting visualizations such as Sankey diagrams, choropleth maps, or heat maps to communicate where resources go and who benefits.
    - Tailor the visuals to a broad audience possibly including local policymakers or nonprofit organizations to make findings accessible and actionable.

What I Bring to Data Science

As an economics major, I see data science as a powerful tool for shaping public policy in both positive and negative ways. My academic background has taught me to think critically about how data influences decisions around resource allocation, especially in

the context of tax policy. I'm particularly interested in how data can be used to ensure that public funding supports goods and services such as education, housing, and health care in a more equitable manner. At the same time, I understand that data can also reinforce existing inequalities if it is not approached carefully. This is why I bring a strong sense of ethical responsibility to my data work especially when it connects to policy. I believe data should be used to highlight disparities, amplify marginalized voices, and support the design of tax policies that more equitably distribute resources within local communities.

Advice

- Be patient, stay curious, and embrace the learning process. It can be challenging but it's rewarding.
- Data science isn't just about writing code. Focus on discovering patterns, telling stories, and thinking creatively.
- Stay flexible, projects often shift direction, and that's where some of the best insights emerge. Embrace the unexpected and leave room for exploration.
- Data cleaning may seem tedious, but it's the backbone of any meaningful and trustworthy analysis as it ensures your insights are built on solid ground.
- Learn tools like Python, SQL, or R/R Studio, but don't overlook the importance of communication and critical thinking.
- Embrace the mess, data science is a creative and evolving journey.