Data Manifesto

Konon Matsumoto

Data Science:

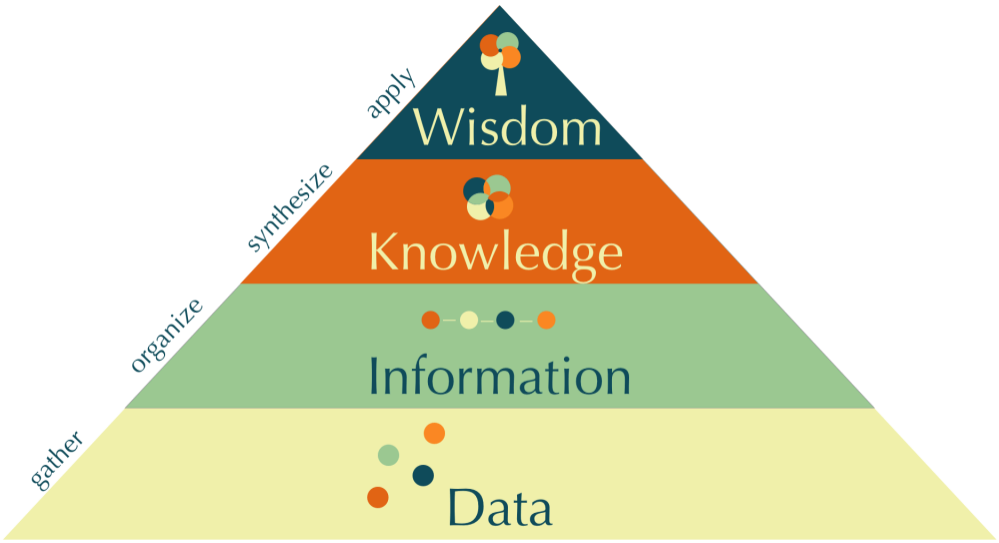
Decoding Digital Signals in a Complex World

I. Making Sense of Data in the Tech Era

We've entered the age of information overload—information envelops us in mounting amounts, looking chaotic and overwhelming. But amidst this apparent chaos lies tremendous potential if we deal with it in a systematic and intentional manner. As we

take apart massive data sets into their component elements, we transform raw numbers into compelling narratives that break down real tales about our world.

Essentially, data is raw input—binary signals, measurements, or observations—that is only valuable when organized into valuable information. In context, that information becomes knowledge, and eventually wisdom when applied to make sound decisions. This is best described by the DIKW Pyramid (Data → Information → Knowledge → Wisdom), reminding us that data science is more than coding and computation, and encompasses interpretation, contextualization, and significant application.



Another model for thinking about data is Jill Lepore's filing cabinet analogy. In this system, drawers are broad categories of knowledge, folders in drawers are particular subjects, and individual files hold discrete points of data. Interestingly, Lepore titles the

drawers "Mysteries," "Facts," "Numbers," and "Data," and the most mysterious ideas such as death or power structures are 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 wonderfully illustrates that data is neither neutral nor whole. What becomes collected, archived, or barred out symbolizes deliberate decisions subject to current power dynamics. In our current environment, where biased data representation has the potential to drastically alter public opinion, this paradigm leads us to question data by asking ourselves: What's omitted? Who made this? What's the story here?

We've entered the tech age—but what does that mean for data science? If data is everywhere, how do we make sense of it responsibly? This manifesto is my personal reflection on how I now approach data work, informed by my experiences in class and the insights from the readings we explored together.

II. The Data Scientist as Translator and Storyteller

Outside the DIKW Pyramid, we must redefine the data scientist's role away from technical skill. We are translators and narrators operating between raw data and human understanding. If our analysis cannot be effectively told to stakeholders—executives, policymakers, or the broader public—then even the most sophisticated models deliver little value. Storytelling plugs this vital gap, adding meaning and impact to our work.

In our Data Science course, this principle was realized in Project 4's visualization recreation assignment. When asked to re-create the same visualization using three tools—Python, spreadsheet software, and specialized visualization tools like Flourish or Tableau—I learned how much the medium affects the message. Producing the same chart in various forms showed that although the same data was being used, every tool's advantages and disadvantages radically influenced the perception of the information. The Python version provided accuracy but took immense effort, the spreadsheet version was easy to access but not so customizable, and the visualization software offered a trade-off between looks and usability. That exercise demonstrated that successful data communication has nothing to do with the digits themselves, but with choosing the right medium for stimulating well-rounded decision-making within specific audiences and applications.

We can write rich and dense stories with data. We can educate the reader's eye to become familiar with visual languages that convey the true depth of complex stories.

Dense and unconventional data visualizations promote slowness—a particularly poignant goal to set in our era of ever shortening attention spans. If we can create visuals that encourage careful reading and personal engagement, people will find more and more real value in data and in what it represents.

This emphasis on slowness and slow reflective engagement with data visualizations acts as a corrective to the immediacy of information consumption that marks so much of our experience in the online realm. By establishing visualizations worthy of attention and reflection, we invite a more immersive understanding of what stories data has to tell.

III. Guiding Principles of Data Science Practice

Principle 1: Begin with the Why

- Clarify the intent behind the data: Before jumping into visuals or metrics, I asked what this dataset was trying to communicate. For example for project 5 I asked where the data was just documenting park locations—or highlighting public access to green space?
- Recognize the broader issue: I realized that park availability is a proxy for deeper urban equity questions: Who has easy access to green space? Are amenities clustered in privileged ZIP codes?
- Let equity reshape your goals: Initially, I considered counting parks per ZIP code. But that didn't account for the lived experience of proximity. Once I reframed the "why," I focused instead on how many residents truly lived within walkable distance of a park.
- Challenge assumptions about completeness: This dataset may look comprehensive, but it required cross-checking with population density and geographic service zones to reveal real accessibility gaps.

Principle 2: Critically Examine Your Dataset

Understanding a dataset's provenance is essential before beginning analysis. Without critically examining its source and structure, our findings may propagate hidden biases or lead to harmful conclusions.

When evaluating any dataset, consider these key questions:

- What motivated this dataset's creation?
- Who compiled the data and with what resources?
- Who funded or commissioned the collection?
- Which perspectives or populations might be underrepresented?

As data practitioners, we have the obligation to uphold ethical principles throughout our workflow—from analysis to communication.

For example, when I worked on Project 5 (Parks Clustering Analysis):

- Purpose: The data set was utilized to examine public parks and recreation facilities in the 100 largest U.S. cities with the goal of exploring patterns in access and resource distribution using clustering techniques.
- Creator: The original data is from the Trust for Public Land; our instructor cleaned and prepared it for analysis.
- Funding: The Trust for Public Land is a non-profit organization, and they released the dataset for public use for research and educational purposes.
- Missing perspectives: The data set focuses on quantitative metrics (e.g., acres per resident, number of amenities) and may leave out qualitative aspects like safety, accessibility for people with disabilities, or cultural importance of parks. It may also not include inequities within cities, e.g., differences in park access by race or income.

Principle 3: Take Bias Awareness and Ethical Practice

Bias appears in a variety of guises throughout the data life cycle, and an awareness of its effects is essential to ethical analysis. Selection bias—when sample data isn't representative of the wider population—is perhaps the most frequent but overlooked form of bias in data science.

This bias can be due to methodology of collection, inclusion/exclusion parameters, or false assumptions regarding representation. By not thoroughly considering who or what is in our dataset, we risk making conclusions that are sound but not based on reality.

> In this way, Darden models what we call *data feminism*: a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought. The starting point for data feminism is something that goes mostly unacknowledged in data science: power is not distributed equally in the world. Those who wield power are disproportionately elite, straight, white, able-bodied, cisgender men from the Global North.[20] The work of data
>
> Catherine D'Ignazio and Lauren Klein. "Data feminism"

This critical lens of data science acknowledges that data science is not an isolated phenomenon but also reflects the same power structures that govern the rest of society. In embracing data feminism, we're making a commitment to examining how power structures inform what data is being collected, how it's being analyzed, and who benefits from the conclusions drawn.

In "Calling Bullshit," Bergstrom and West illustrate this concept using examples from the classroom: students consistently tell us they have higher mean class sizes than institutional records indicate because they're disproportionately enrolled in bigger classes. This discrepancy shows that lived experience and statistical means can diverge significantly, and therefore understanding data structure is essential to accurate interpretation.

> 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 extends to measurement tools, historical context, and even question design. Algorithms, which are typically considered to be unbiased, simply reflect systemic disparities.

In Algorithms of Oppression, Safiya Noble examineshow search engines and other technologies like them reproduce discrimination. As Noble describes, these patterns are

not accidents but rather the outcome of deep-seated inequalities that have become fixed both in data sets and technical systems.

> 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.

Neglecting to confront bias has repercussions that stretch well beyond statistical accuracy. Biased information can mislead decision-making, especially in situations that affect public policy, access to healthcare, or allocation of resources. This not only widens current inequalities but also threatens to destroy public trust in the institutions that depend on data to support communities.

Principle 4: Data Is a Conversation, Not a Conclusion

Data readily gets out of control if there isn't good organization. These routines help remain grounded in the presence of complexity:

- Set solid objectives: Begin with a principal study question, then develop supportive sub-questions to guide inquiry. In our project on network analysis, our starting question on social graph community development became manageable when broken down into specific metrics like centrality values and clustering coefficients.
- Document your workflow: Label your code carefully, including what may appear to be insignificant steps such as creating dataframes, missing value handling, or applying transformations. This documentation was incredibly useful during debugging of our sentiment analysis project, enabling us to easily determine where preprocessing influenced classification accuracy.

```
# Select only numeric columns
numeric_cols = df.select_dtypes(include=['float64']).columns

# Normalize the numeric columns by dividing each by its standard deviation
df_normalized = df.copy()
df_normalized[numeric_cols] = df[numeric_cols] / df[numeric_cols].std()

# Display the first few rows of the normalized dataset
print("The normalized dataset:")
df_normalized
```

```
The normalized dataset:
```

	City	Population	Acres per 1,000 people	Parks per 10,000 residents	Parks as % City Area	Fields/ Diamonds	Tennis_dedicdated	Pickleball_dedicated	Pickleball_combined	Hoops	Community_garden_sites	Dog_parks	Playgrounds	Rec_senior_centers	Restroo
0	Albuquerque, NM	0.567914	0.129078	2.618802	1.932526	2.675051	1.693940	1.470240	2.638993	1.333337	0.000000	2.950869	1.787147	2.013198	1.0452
1	Anaheim, CA	0.354636	0.044274	0.865378	1.462808	1.957079	0.750315	2.102183	1.320652	0.396539	0.357604	0.859187	0.938341	0.519990	1.3949
2	Anchorage, AK	0.296059	10.026918	3.572272	8.182003	2.370345	1.175312	1.007244	0.632780	0.520668	0.535447	2.058361	1.685994	0.373724	1.2531
3	Arlington, TX	0.407615	0.036063	1.158310	0.724488	1.778387	0.602579	0.804740	1.286883	2.063360	0.077781	0.560636	2.285868	1.176253	1.5574
4	Arlington, VA	0.252164	0.024098	2.771109	1.144397	3.333435	3.652685	0.000000	1.485860	4.504337	1.257309	3.020835	2.947228	2.193895	1.2908
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	Tulsa, OK	0.427568	0.064704	1.501789	0.664422	4.328673	2.824424	1.115907	0.788676	0.771649	0.222455	0.534474	2.334853	0.690069	1.0991
96	Virginia Beach, VA	0.473901	0.185675	0.009963	1.684870	1.822546	1.338926	3.586740	2.253294	1.871762	0.000000	0.803696	2.855577	0.544776	0.9568
97	Washington, DC	0.717279	0.044109	1.573203	2.431623	1.644947	2.496909	0.623614	0.940254	1.836139	6.851233	2.017788	1.090242	3.599295	2.2874
98	Wichita, KS	0.392056	0.035603	1.396964	0.411021	2.124342	1.722856	2.357908	1.481306	0.848442	0.161736	0.971476	1.230732	0.846647	1.6823
99	Winston-Salem, NC	0.250861	0.051537	1.675062	0.455600	2.059630	3.671649	1.307590	0.821466	0.905545	0.000000	0.910955	1.083313	2.499323	1.8076

100 rows × 20 columns

- Adhere to realistic timelines: Break down the work into phases with specified blocks of time for each aspect. With Project 9, collaborative work ensured that we had to coordinate our timelines—precise time was spent on data preparation,

development of visualizations, and most importantly, rigorous feedback sessions. This group process not only enriched our student mental health analysis but also provided us sufficient time to include peer feedback and refine our visualizations prior to submission.

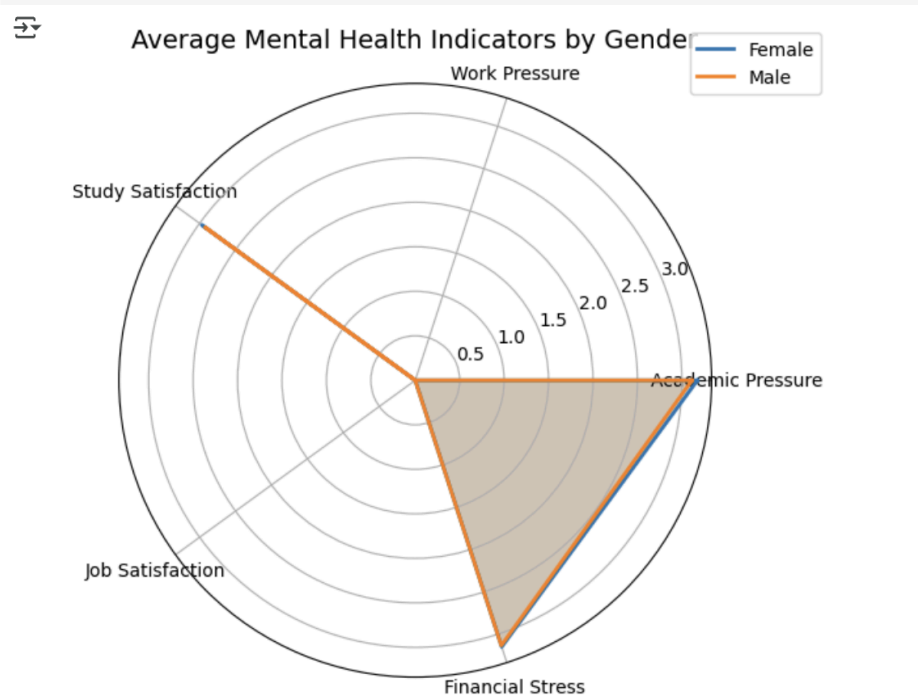
- Communicate: If faced with obstacles—whether conceptual or technical—ask colleagues, instructors, or internet forums. In our cluster analysis, asking peers for advice revealed other distance measures that significantly improved our results.

Principle 5: Storytelling Through Visualization

- Don't be afraid to acquire new techniques: Familiar plots like scatter plots and histograms serve their purposes, but incorporating some visualization tools in your toolkit amplifies the potential of communication.
- Keep your audience in mind: Adapt visualizations to technical background and domain expertise of viewers. Our presentation to computer science professors had graphs of algorithm complexity in detail, whereas the same project presented to a general audience used simple visual metaphors with a lot of documentation of the code and what the visualization is communicating.

One particularly effective example from our coursework was the radar chart we created for Project 9, which focused on student mental health patterns. Rather than using standard bar charts to display average outcomes, we implemented a radar chart to show how multiple mental health indicators—such as academic pressure, financial

stress, job satisfaction, and study satisfaction—varied between male and female students.



This visualization immediately revealed patterns that traditional charts couldn't capture. For instance, while both genders reported similar academic pressure and financial stress levels, the radar plot showed sharper divergence in job and study satisfaction. By exploring this multidimensional approach instead of defaulting to familiar methods, we uncovered nuanced differences in student experiences that weren’t visible in the raw numbers alone.

IV. What Distinguishes Exceptional Data Scientists

The greatest data scientists have technical proficiency paired with intellectual curiosity, flexibility, ethical awareness, and good communication skills. Their technical skills—whether in programming languages like Python, database system interactions such as SQL, or using statistical models—allow them to manipulate and analyze large data sets effectively. Just as important is their ability to frame questions: they can take imprecise or broad questions and convert them into testable hypotheses and select analytical techniques best suited to the question.

What actually sets them apart, however, is their intellectual curiosity. They are constantly seeking to learn, to pick up new methods and failures as an opportunity to improve. Behind this push is a deep ethical decision—a sense that data work has broader social stakes and should be approached with respect and accountability. Finally, outstanding data scientists are effective communicators. They abstract analyses and bring them back down to earth as simple, engaging stories and pictures, making data not only accessible but actionable to a diverse set of people.

Applying My Guiding Principles to a Hypothetical Project: International Education Costs

To demonstrate how I would apply my guiding principles to a new project, let's assume I'm working with a dataset of international education costs—tuition, living costs, and country-level data for international students studying overseas.

1. Begin with the Why

Why was this dataset created, and by whom? Is the aim to assist policy makers? Guide student decisions? I'd clarify its purpose, then ask who benefits from the way it's presented. I'd consider how my own framing might unintentionally reinforce global power imbalances—especially if costs are used to "rank" countries.

2. Critically Examine the Dataset

I'd assess which countries are included—and which are excluded. Are developing world regions underrepresented? Are fees based on real student reports or institutional estimates? Is currency conversion equitably applied (e.g., adjusted for PPP)? If the dataset includes only Western institutions, conclusions would need major caveats.

3. Be Aware of Bias and Ethics

Bias can also seep in through omission (countries left out), generalization (averages concealing inequality), or framing (high price as high quality). I would carefully consider the ethics of cross-country comparison. For instance, labeling low-cost countries as "cheap" rather than "affordable" can reinforce stereotypes. Following Algorithms of Oppression, I'd avoid visual rhetoric that privileges powerful dominant powers.

4. Treat Data as an Ongoing Conversation

Instead of attempting to discover the one solution—e.g., "Where is the cheapest place to study?"—I would tackle this as an iterative project. I might start with cost maps, then

introduce layers of visa availability, post-grad work policies, and student satisfaction. These incremental layers would compel me to revisit my assumptions. I'd solicit peer and user feedback throughout.

5. Storytelling Through Visualization

I would prioritize visualizations that communicate not just cost but also context. For instance, a choropleth map of the average tuition per country could be deceptive unless layered with income information or visa challengingness. A radar chart may be more successful at communicating the nuance: cost, quality, safety, post-grad opportunity, etc.

Inspired by Giorgia Lupi's slow data philosophy, I would design these visuals to promote reflective thinking—not impulsive responses. I'd avoid busy comparisons or misleading ranking graphs. I'd aim to create visuals that enable students and policymakers to understand trade-offs in international study, and feel the nuance behind each data point.

What I Bring to Data Science

As a computer science student with experience in algorithmic design and systems thinking, I'm drawn to data science for the ways in which technical systems influence human results and social formations. My technical training has given me strong programming skills and algorithm analysis capabilities, but I've since come to see these as tools, not ends.

I am especially drawn to investigating how technology systems mediate access to opportunity and resources between communities. This is because I think that data science has a lot to lose if implemented in action in automated decision-making systems such as hiring systems, financial approval systems, and resource allocation mechanisms—areas where algorithms directly influence people's life chances.

At the same time, I understand that technical practice is only half of ethical data science. The education I received in ethics and social consequences of computing has shown me how unambiguously framed technical decisions can have mixed impacts on diverse groups. This sensitivity encourages me to be as interested in the individuals beyond the data as I am in the data, bringing attention to each assignment with regard for whose interests are benefited or possibly harmed by choices made during analysis and subsequent uses.

Final Thoughts: Advice for Future Data Scientists

It takes time to become a responsible and effective data scientist, so be patient, inquisitive, and patient. It is frustrating at times, but it is also profoundly rewarding. Practice slowing down—complexity isn't your enemy, it's your ally. Data science requires working across many diverse disciplines, and mastery comes gradually with repeated effort. In the meantime, remember that even the best algorithms can't compensate for poorly phrased questions or lack of knowledge about the data itself.

Heavily document your work; your future self will thank you when you return to a project after months. While data cleaning may be tedious, it is the basis of any meaningful and credible analysis. Start simple and add complexity incrementally. From a solid base generally leads to greater insights than bypassing and moving directly to high-level techniques. And while technical proficiency is needed, so too is communication. Insights are futile if they are not communicated properly to decision makers.