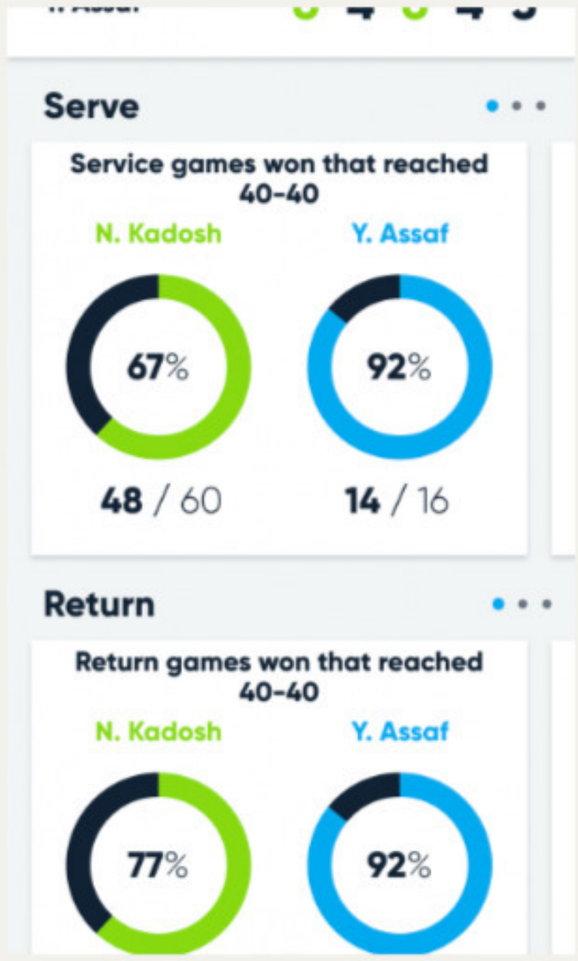
O'Farrill: Data Manifesto (CS-215-A)

I now see data as a foundational input of analysis, helping us to find trends or patterns, which can drive innovation.



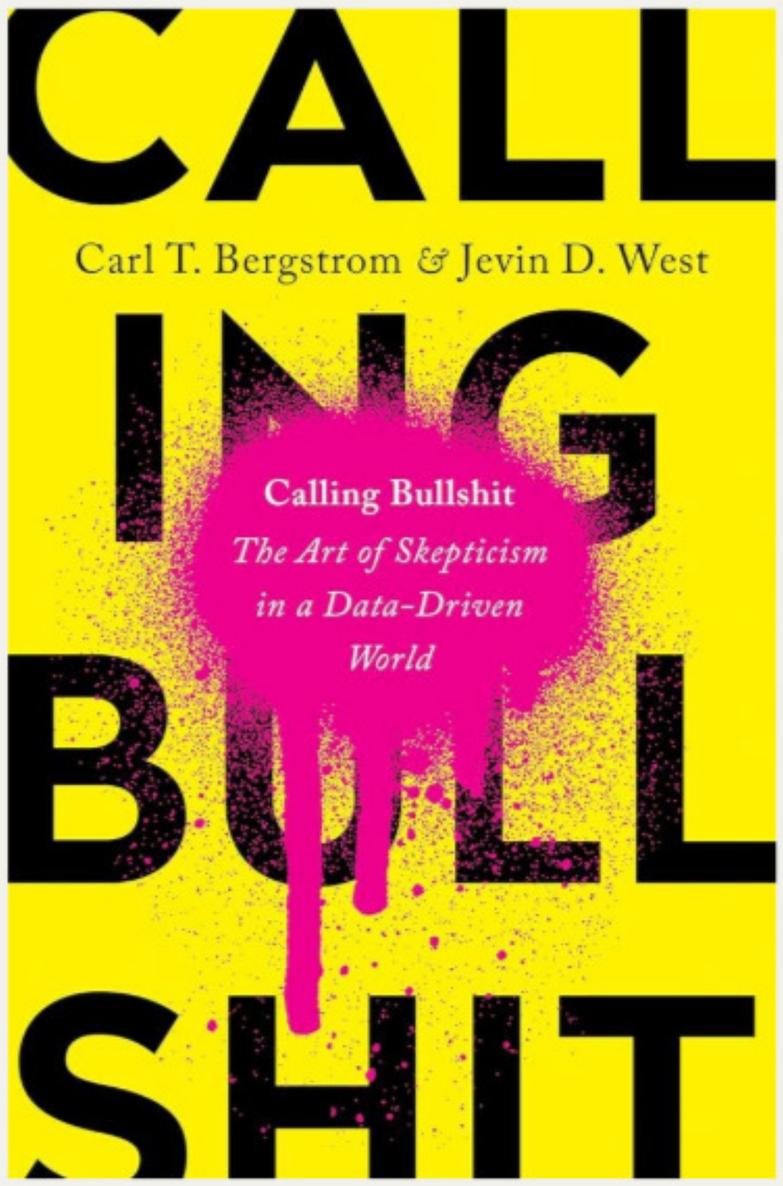
What is data, and how is it distinct from information, knowledge, facts, etc?

We don’t realize it at times, but our everyday actions and behaviors can become information if processed, structured, and contextualized (whether quantitatively or qualitatively). They serve as foundations for serving knowledge; however, they lack meaning until systematically analyzed. In other words, data is the raw material, while information is the interpreted and meaningful results derived from the processed data. Then, from the information, we can apply novel insights (knowledge).

For example, tennis can be seen as mundanely hitting tennis balls on a surface (meaningless information from the action itself). However, the mundane action can become useful when averaged to show trends in points (information). Through the information, knowledge is gained from the derived information. Now, it implies a deeper comprehension, which leads to us applying effective insights on the sport. For example, we can make the conclusion that points on serve are more dominantly won than on the receiver's end from average points by testing these conditions. Now, our choice of shot has purpose.

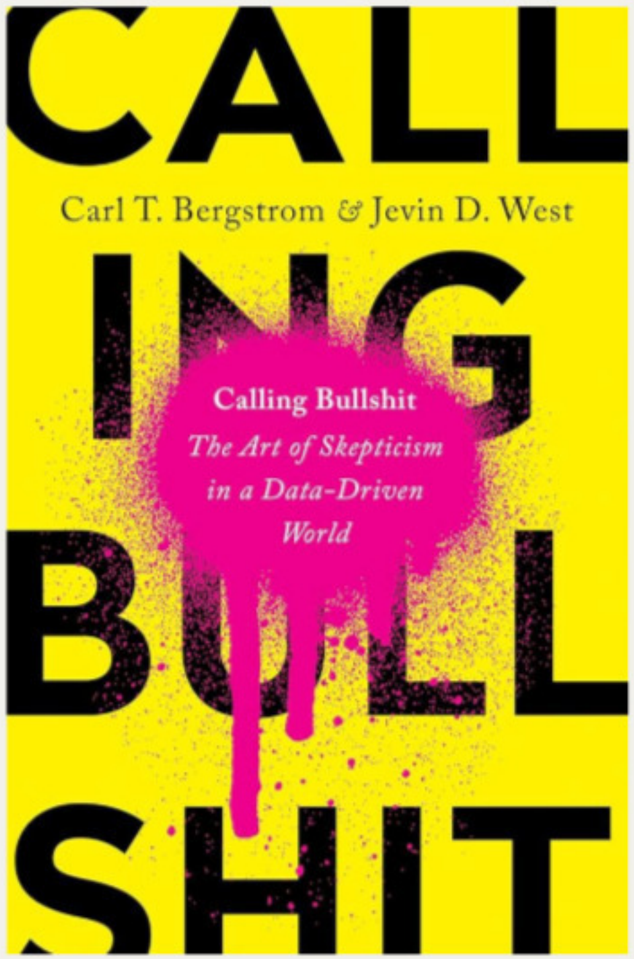
behaviors ⇾ information ⇾ knowledge

What does it mean to be a data scientist? Data represents real life. People facilitate the data to convey whatever they want through their methodological strategies. This can be a strength but also a limitation. We've learned that bias can adhere to the representation of data through misleading insights or poor decision-making. Bias can manifest through data collection or even its interpretation.



For example, "Selection Bias," from *Calling Bullshit* by Carl Bergstrom & Jevin West (2020), addresses how to spot misinformation, even when it comes wrapped in data.

They highlight the notion that the perspective we adopt informs the assumptions we make. This by itself can be considered bias since our own goals for what we want the data to look like can influence results, such as recruiting a single type of group for a study (leading to selection bias). Therefore, it's imperative that we diversify our groups when gathering samples.



For example, "Selection Bias," from *Calling Bullshit* by Carl Bergstrom & Jevin West (2020), addresses how to spot misinformation, even when it comes wrapped in data.

They highlight the notion that the perspective we adopt informs the assumptions we make. This by itself can be considered bias since our own goals for what we want the data to look like can influence results, such as recruiting a single type of group for a study (leading to selection bias). Therefore, it's imperative that we diversify our groups when gathering samples.

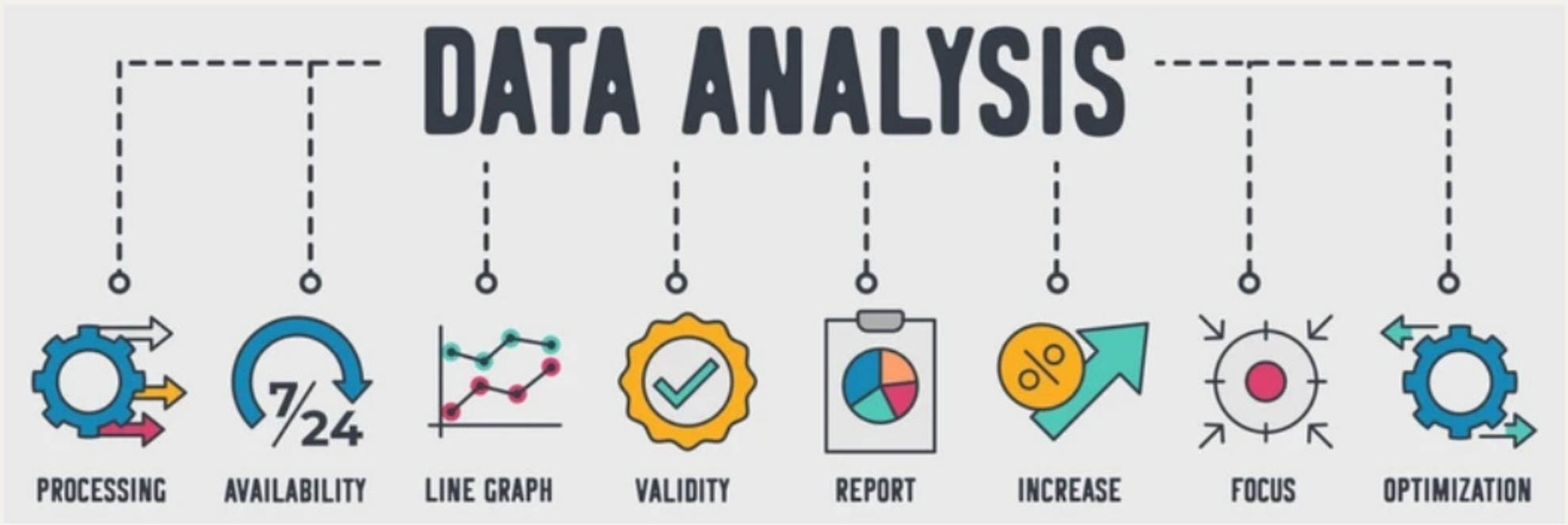
Therefore, being a data scientist means understanding that data is not inherently objective or neutral. It represents real-world scenarios but is also influenced by how it is collected and interpreted (as aforementioned). Moreover, a data scientist's role involves not just analyzing data but being critically aware of the methodologies used and the potential biases that may arise. This awareness allows them to have more accurate and transparent insights, ensuring that conclusions drawn from data are both meaningful and responsible.

What skills do you need to do data work?

I have compiled a list of skills that I've learned throughout this semester in the class. These are my thoughts:

1. **Communication Skills:** Have some knowledge to convey insights and perspectives on the data. Most importantly, elaborate on trends and findings. Even if they seem obvious. Perspective matters.
2. **Data Wrangling:** Have some skill in wrangling data for organizational purposes and to handle missing data. Not only for organization purposes, but also for better readability for the audience. Why include data that arent involved in analysis? Perhaps for context, but cluttering information can be overwhelming.
3. **Critical Thinking:** Some level of this skill means that you would go above and beyond in your relationship with the data. Given the nature of the data, what

3. **Critical Thinking:** Some level of this skill means that you would go above and beyond in your relationship with the data. Given the nature of the data, what information would you be able to convey? Data scientists will experience limitations in certain datasets. However, they persevere through its limited information and are able to display meaningful visualizations. This could tie into communication skills, but this skill is the root of the creativity.
4. **Statistical Knowledge:** Some statistical background is necessary to understand variability, uncertainty, and how bias can skew results. For example, because data can involve ubiquitous content, it's crucial to know the type of design that aligns best with the data.
5. **Interpretation:** This skill could tie into critical thinking but involves the context and the implication of findings. How does the data generalize to other populations? What other conclusions can be made?



Evidently, there are a variety of skills that make up data scientists, especially those that weren't listed. However, the above 5 skills listed are a priority for successful engagement because they acknowledge ethical concerns and a meticulous approach for data.

What advice might you give to someone who is hoping to become a data scientist?

"Stay curious, adaptable, and committed. Always! By continuously learning, you will set yourself up for success! Also, you'll be able to apply new knowledge in evolving contexts of data".

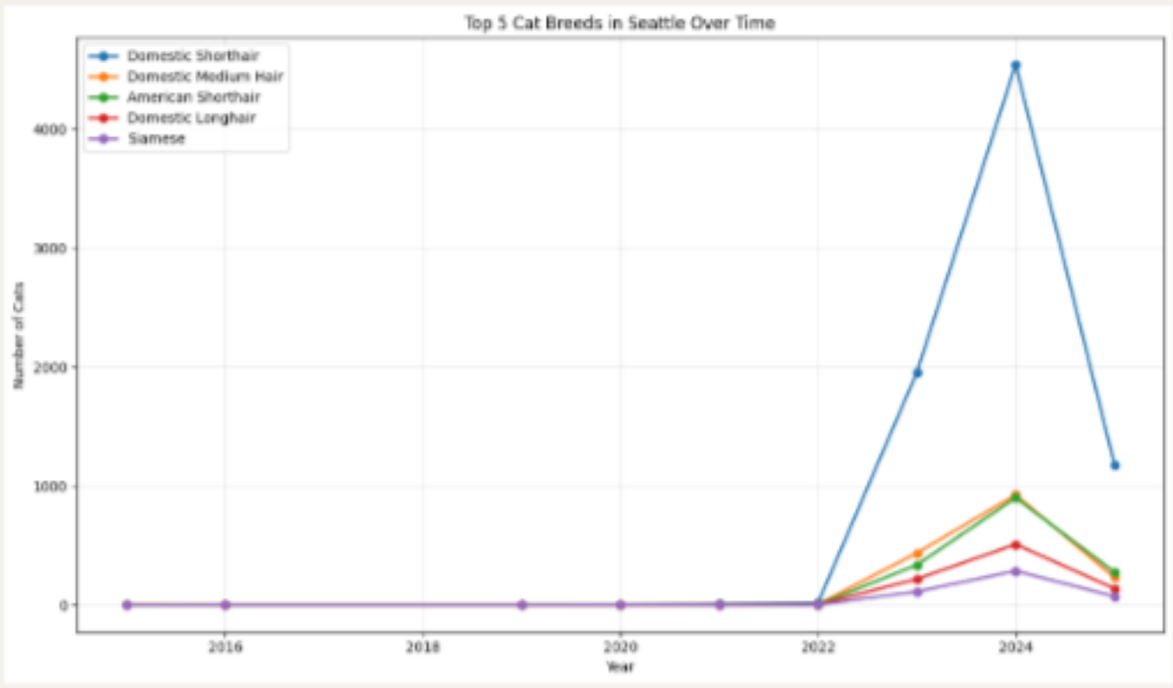
The above quote is verbatim what I would say.

To elaborate, staying curious would help the drive to investigate errors or burning questions of a dataset, and being adaptable builds adversity through the challenges we may experience when dealing with data.

Because looking at data can be daunting, and when we are challenged, we stray away from it. However, the challenge is a common occurrence when it comes to data. Therefore, I really emphasize versatility.

What kinds of problems can you solve / questions can you answer through data analysis and visualization?

When dealing with data, questions are imminent and ubiquitous. We wonder about data before we interact with it, during its interaction, and even after its finished product we may still have questions. Questions can also extend beyond the construction of the data. But we can also ask about its broader impact such as the following project.



The above visualization is derived from a cat license dataset from Seattle. We can see a spike in both licenses in general and also a specific breed, the Domestic Shorthair. It has had an overall increase throughout the trend; however, why did the spike in 2024 occur? From this example, we can see how our finished product with the datasets results in more questions.

3. **Gen Z is leading pet ownership growth**. In 2024, 18.8 million Gen Z households owned a pet, a 43.5% increase from 2023. This generation is also more likely to own multiple pets, with 70% of Gen Z pet owners reporting they have two or more animals.
4. More men, particularly Millennials and Gen Z, are acquiring pets. Among Gen Z dog owners, 58% are men, while 63% of Millennial dog owners are male, marking double-digit increases from the previous year. The biggest shift in cat ownership was also among younger men, with 38% of Gen Z cat owners and 46% of Millennial cat owners identifying as male, reflecting **substantial growth in male pet ownership.**

While the trend could exist for numerous reasons, the following link elucidates that the spike is a result of Gen Z and millennials, whose ownership of cats increased by around **45% in 2024**, proving that its not solely Seattle, but a reflection of the broader population

(https://www.petfoodindustry.com/pet-food-market/market-trends-and-reports/news/15741428/report-pet-ownership-expands-as-gen-z-shifts-trends#:~:text=Gen%20Z%20is%20leading%20pet,Gen%20Z%2C%20are%20acquiring%20pets.
)

Overall, we can see that data leads to information, and information leads to knowledge—all because of the questions we ask about the data. Asking questions sparks a motive for answers. In my case, I wanted to understand the reason behind the spike in 2024. After conducting research, I found the answer.

We can highlight 4 main principles that define the data science process from our overall discussion.

1. **Quality:** We want to figure out if data is reliable and accurate. Therefore, cleaning data by omitting missing values or columns for organization is ideal.
2. **Methodology:** Again, the methods we test are the result we get. Therefore, applying appropriate statistical methods is imperative for showcasing the results of data.
3. **Effective Communication:** Elaboration is expected when presenting data for overall perspective and understanding of patterns.
4. **Ethical Considerations:** We talked about how bias can intrude on the results of data. More broadly speaking, it can also threaten privacy for participants of the data by exposing identity, etc.