Kyle Gustafson

Data Manifesto

Data science is NEVER entirely objective.

Introduction

The existence of massive amounts of data about so many aspects of our society has given us unprecedented access to information that we can use to form knowledge. Tech companies have collections of user data that represent a shockingly coherent digital profile of each user, and most of us use enough apps and services to extend this data to most aspects of our lives. Scientific data has exploded with the creation of autonomous sensing, satellites, health databases, etc. Nations have extensive intelligence about their enemies, allies, and citizens. Any of this data can be communicated to any of us through the internet and social media. In short, the last 75 years - and especially the last 15 or so - has seen an explosion of data measurement, collection, and distribution on a scale that is hard to comprehend.

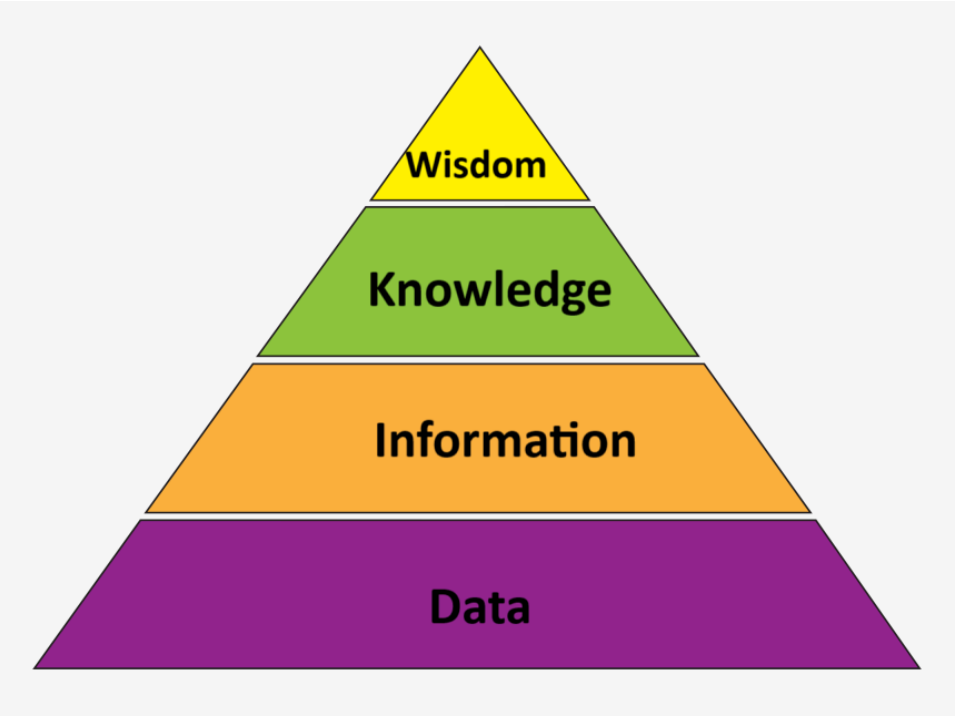
The speed of this increase left little time for the public to consider the state of data collection and think about its implications, purpose, and intentions. We haven't had enough time to understand data before it has become a fundamental part of modern thought. Most of the time, we see data as a type of information that is fundamentally different from other types: opinions, findings, and claims are for debating and criticizing, but data, at its basic form, is just the world converted to bits. Why should we question it? My experience with data science has led me to believe that we should question data and its derivatives as much as any other form of information.

My philosophy for working with data science emphasizes skepticism and intentionality. In this document, I'll outline my principles for thinking about and working with data, my understanding of data and its limitations, and the ideas that led me to these conclusions.

Principles

When thinking about data science, it is helpful to create a structure to visualize the process by which data science creates new understandings. One of the most common representations is the DIKW pyramid, which depicts the process as:

Data → Information → Knowledge → Wisdom



In this model, data is the foundation for all human thought created by its analysis. Data is the result of compiling measurements into numeric form. Information is the result of processing and compiling data into organized 'facts'. Knowledge is the result of the analysis of information to understand what it means. Wisdom is what we learn from the process, and how that informs our decisions.

While this model is useful to understand the different structural products of data science, I think that it implies a more linear, objective process than actually exists. This leads me to the first principle I use to think about data science:

---

1. **The entire process of data science - from collection to interpretation to communication - is influenced by human thoughts and decisions.**

One of our early readings from the book *Human Centered Data Science* strongly influenced my first principle:

> "One of the difficulties in dealing with the "data deluge" is a facile assumption that the data can tell us everything-that it is unbiased, neutral, and somehow possessing wisdom far beyond the human. "If it's big enough, it contains everything." Our approach puts human

responsibility at the center of data science. People are involved at every stage of the cycle of collecting, cleaning, analyzing, and communicating data science results."

When I think critically about data analysis I see online, I mostly think about the claims an author is making and the way they present their data, but not as much about where they get their data or how they select from it. Decision making at every step of the data analysis process influences how we interpret and understand the conclusions we make about that data. From the beginning, responsibly creating and sourcing data is just as important as any other step in the data science process. If no consideration is given to the impact that data selection and collection has on the possible outcomes of analysis, we shouldn't be sure of the knowledge we produce from that data. This leads me to my second principle:

---

2. Ensuring data is accurate, impartial, and informative is necessary to create further forms of information.

To emphasize the importance of considering how humans produce data, I'll use weather as a personal example to my interests. Weather data is collected in many different ways, including weather stations, buoys, satellites, balloons, and airplanes. This data may seem inherently unbiased; why should we have any reason to be critical of this data? While weather data is much more objective than many other types of data, the human decisions that go into how and when measurements are performed can have a profound impact on the data that is created. Weather stations have to be built and placed somewhere; Seattle's official station is located nearly 30 miles south of the city at the SeaTac airport. Not only does this station regularly experience different small-scale weather than other parts of Seattle, it could be affected by its proximity to the vast concrete surfaces of the airport. When working with raw data, it is easy to forget that we first have to create that data, which adds human biases to the process of data science right at the beginning.

After ensuring sound data, we can start to analyze that data to uncover patterns and make inferences that can lead to new knowledge or understanding. After responsibly gathering and wrangling data for my projects, I move on to the process of exploratory data analysis. The textbook *Veridical Data Science* defines exploratory data analysis as "the task of visually and numerically summarizing the patterns, trends, and relationships that a dataset contains in the context of the domain problem that we are solving" (VDS Ch. 5). At this stage, I am not trying to

make any inferences about my data. The most important part of data exploration to me is keeping an open mind to see what my data tells me without a desired answer in mind. As such, my third principle is:

---

3. Think critically and openly about the results of your data analysis and the conclusions you can reach from it.

I want to use data science to better understand reality. If I analyze weather trends, and I hypothesize that climate change has an impact, I should make sure to be open to any results of my data, even if I do not see said impact. Unexpected results should be seen as an opportunity to further investigate the topic you are working with (and to double-check your data).

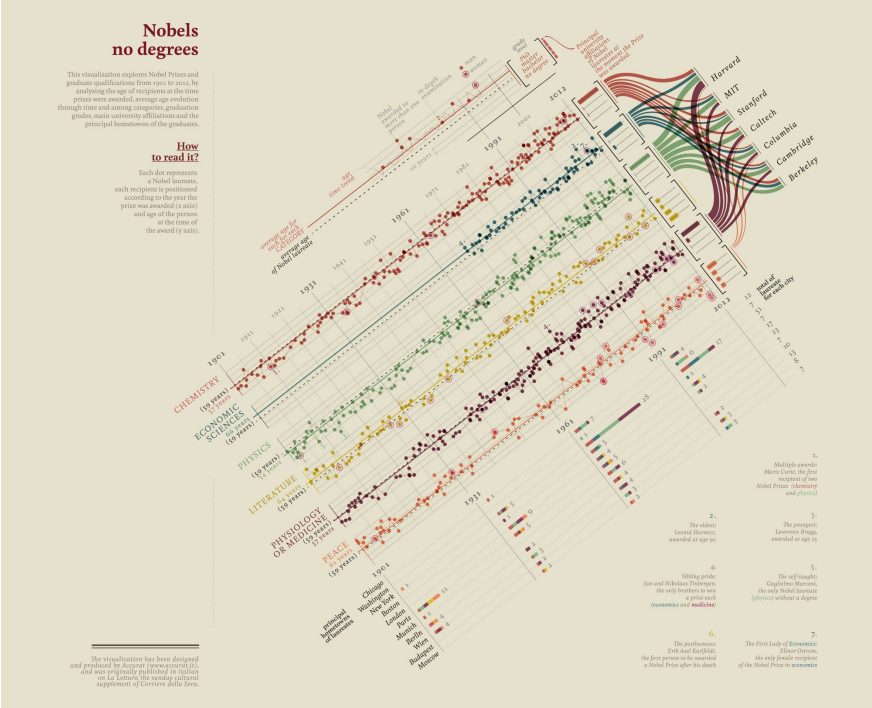
Once I have thoroughly explored my data and understand the results, I move on to the final stage of my data science process. In this stage, I am focused on presenting and communicating what I have learned from my analysis. In this class, I worked with a number of analysis techniques (such as cluster analysis) and visualization types (such as heatmaps). While I have a lot more to learn before I can achieve the depth I want to be able to answer more complex questions, I learned how best to create visualizations that highlight what the data means. My fourth and final principles are:

---

4. Data presentations should focus on balancing accessibility and depth.

5. Effective data presentations will let viewers reach their own conclusions.

In reading *Data Humanism* by Giorgia Lupi, we were exposed to the idea that we should "embrace complexity" in our data presentations. She gave examples of visualizations that have many different interacting visualizations within them to create a rich narrative:



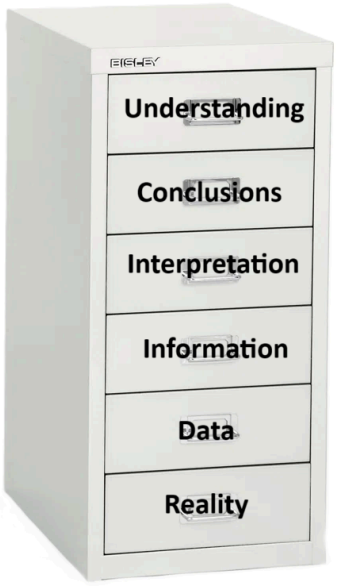
I think that it is extremely important to ensure that my data presentations have enough context to present a complete narrative about the data. My goal is to fully explain and interpret my data so that anyone viewing my analysis receives the information they need to reach their own conclusions. A good data analysis will include a depth of visualizations and explanations to clearly demonstrate the patterns and conclusions it presents.

At the same time, it is important to me that my analyses are widely accessible. The Nobel visualization above is beautiful and incredibly interesting to explore, but is also daunting to approach, especially for those not involved in data science. An effective data analysis should be approachable and easy to understand as well as retaining depth.

Conclusion

This course has developed my understanding of data science methods and processes and influenced how I think about the nature of human knowledge. Data science plays a larger role in society than it ever has before. With the rise of machine learning and artificial intelligence, humans are stepping back from the process of creating knowledge. Ensuring accurate and unbiased data is more important than ever when we train AI systems that synthesize knowledge off of this data.

As a representation of how my understanding of the evolution of knowledge has changed, I updated the DIKW model to reflect my data principles. I used Jill Lepore's filing cabinet metaphor to represent how human thought is transferred between these categories. Because data itself is a product of human creation, I added a category below it for reality. Before any step of data science, there exists a basic reality that we are trying to understand or analyze. This category would include the storms that produce precipitation data. Data is the result of our attempt to represent reality numerically. Information represents the categorization and organization of data to display patterns. I added a category for interpretation, which is how we describe and explain our information. From our interpretations, we reach conclusions about what our data shows and what that means for our understanding of the topic. These conclusions can then go on to influence collective understandings of the topic. Moving files up into higher cabinets introduces an inherent level of subjectivity based on the decisions we make to form higher levels of thought.



Data Cookbook

With my data manifesto, I have laid out my philosophy and approaches to using data science. My data cookbook is a collection of the most important programming tools I learned this semester to perform data analysis.
https://colab.research.google.com/drive/1MCp9xuELun5lsVzu8fJ9M2lvgmAulcGT?usp=sharing