Iris Çano

Intro to Data Science

Prof. Jordan Wirfs-Brock

Spring 2023

Data Manifesto

Mind the gap! is not just a mundane cautionary phrase uttered to passengers before embarking upon the London Underground's subway system. It is a crucial reminder of the dangers that lie beneath this beautiful city, and a warning to be careful while traversing the subterranean labyrinth. Data and public transportation might seem like two unrelated concepts, however, they are both intertwined and both play a crucial role in our lives. The transportation industry has heavily relied on data in order to become accessible to people by optimizing its routes and schedules and making optimal decisions and use of their resources. Different companies have used data to detect potential hazards such as travel hazards and faulty equipment and enhance

their safety. However, unlike the London underground, we are often not told about the existence of the data gap. Data has had invisible power across history, however, it is now acknowledged to have extensive influence over our daily lives. While reading articles I have always noticed the bold statistics statements that are used to give the argument credibility. Oftentimes people consider data to be objective, however that might not be the case. The history of how data has been used or in some cases how it is still indicated otherwise. After a lot of consideration, I have realized that reading data information or working with data requires a lot of critical thinking and questioning.

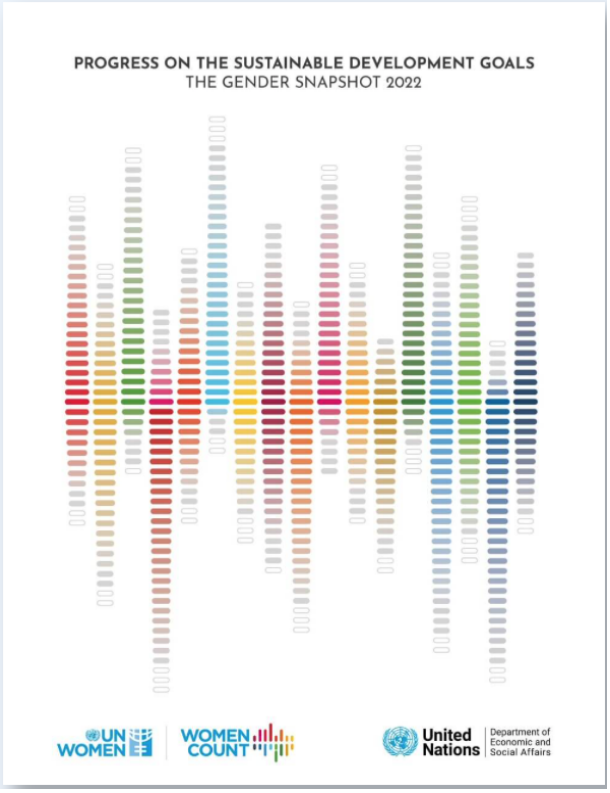
Data is not merely an objective reality, it is the building blocks of everything that surrounds us. Similarly to making a building, data pieces are used to generate new insights and new predictions and create new forms of knowledge. Data sometimes requires cleaning, preprocessing and transformation before it can be analyzed further similar to how raw materials used in construction. Different materials have various properties and require different maintenance similar to how data requires different processing methods and are suitable for different analytical approaches. Just as manufacturers need to manage their raw materials inventory to ensure they have enough materials to meet demand, organizations need to manage their data effectively to ensure they have the information they need to make informed decisions and achieve their goals. Construction requires a lot of time, planning and organization in order for it to be useful, data also requires planning on how to collect and analyze it, the lack of which can cause further challenges. Another relevant aspect of data is its validity. Analyzing incorrect or biased data can lead to misinformation that can be potentially harmful to people utilizing it later. The correctness of the data collection process is like the foundations of a new building. The

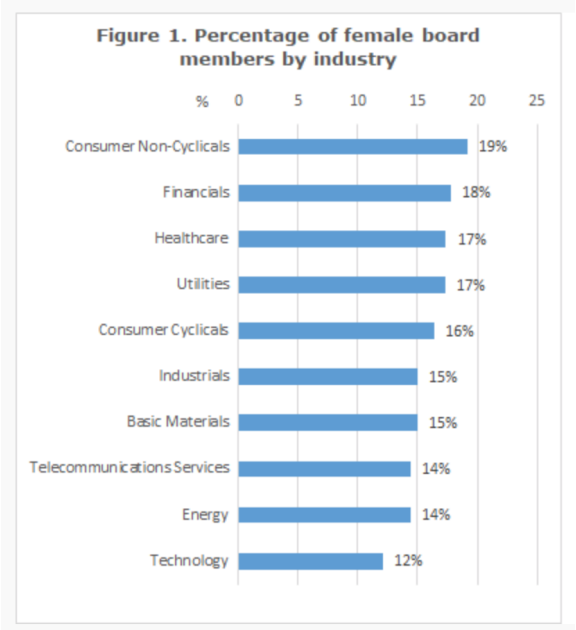
foundations of a building provide support for the entire structure, similarly to how the quality and completeness of data provides the foundation for which decisions based on data can be made. A poorly built foundation will not be able to withstand natural disasters and internal factors. Similarly, good quality and well organized data can provide great support for people who use it to make decisions and rely on it. Being able to work with data demands self awareness of the history of how data has been used to inform policy and strategies.

There are a few examples of how data has been used to cater to men in various ways. One of them would be at what is considered to be optimal room temperature. What we consider to be room temperature was determined decades ago based on metabolic rate data on a 40 year old man who weighs 150 pounds. Now that women represent a greater part of the workforce, and on average have slower metabolic rates at rest, different considerations need to be taken into account to accommodate their thermal needs. This is a very simple example of how data can create more comfort or discomfort in people's lives. There are various examples of how the way data is analyzed can amplify bias towards women. There is a significant lack of data concerning women, whether that concerns female biology and hormones or other factors.

Data is often seen as objective and not biased, however, it is a clear reflection of who we are as a civilization, as much as language, culture and traditions are. The systemic ignoral towards a significant portion of the population is not by coincidence, it is a relevant indicator of our society and portrays how women have been ignored and undermined throughout history. Women have been ignored in several steps of the data analysis process including the data collection methods, gender bias in data analysis, lack of disaggregated data as well as limited

representation in the decision making process. Historically women have been excluded from the data collection process or have ignored asking questions that are mainly relevant to women's experiences.





Ignoring women's experiences during the process of collecting data has led to a lack of data concerning women's lives, experiences and needs. Furthermore, another relevant aspect of analyzing data is about who makes decisions concerning it. As women have not been part of the workforce for a very long time, they have not been allowed to make decisions regarding how data is analyzed and later interpreted. There are many studies on the gender wage gap that acknowledge that overall men tend to get paid more than women. Some of these studies are often misinterpreted and might attribute the gender wage gap to women having less technical skills and ignore systemic differences that contribute to the labor market differences between men and women. This perspective is a very male centric one that fails to acknowledge the systemic

barriers that mostly women have to endure. For example, women's career prospects are generally affected by having children, which reduces their ability to secure employment and hence, their productivity in the future. In addition, women experience more discrimination in the workplace and the barriers of entry women face to enter the labor force. Reported data also fails to recognize the intersectionality of women's experiences on the basis of nationality, ethnicity and race. Data sometimes fails to understand the differences between the struggles that women of different backgrounds face. As Dr. Fei-Fei Li, a well-known data scientist has stated: *"Data is a reflection of our world, and if we don't have women and people of different backgrounds represented in that data, then we're not going to have algorithms that work for everyone."* (). It is essential for there to be data to be representative of people of different backgrounds as that would create inaccurate models and algorithms that reinforce negative biases.

These are some of the ways data can be misinterpreted and be utilized to perpetuate negative stereotypes about women in different fields. Therefore data science work requires skills that go beyond technical as well as knowledge and self awareness on the influence of our results. It is important to know that data inherently carries the bias of those who collect it and that it takes a lot of critical work to address it. It is very important to acknowledge the ways in which data intersects with forms of oppression and how it can have a great impact on people's lives.

Bibliography: