Cs215 Ousmane Barry

Data manifesto.

Imagine a million years before today and a million years from today. These are two imaginary periods that are entirely different because they are characterizing different events. Since, a million years before now, would be in the past as history, this could be easy to understand. However, a million years from today could be tricky to see as the only ways for a possible outcome will either be through imagination or emphasizing which will be based on past and present. Making the past one of the future variables that can be used to predict what is coming. This can be tested in many different ways, as if today, I approach my parent and ask him to tell me about his childhood, most probably he might tell me about something that I did not experience during my childhood. So based on this, I can have a sense of what to expect from the coming generation of my kids: For example, 100 years ago telephone or internet was scarce making it impossible for an individual to possess one but nowadays, one person can possess more than 5 telephones, and even children can have access to one. From this example, we can give a perspective of what the world could be like within the next centuries. Maybe there will be more cars, more virtual work, or even the labor force will be replaced by robots. Now the questions we would ask are why the use of technology is progressing fast, and what are the factors that are contributing to this?

The use of technology has been increasing, thanks to the efficient use of data which causes innovation, which is seen as a great contributor to the positive increase in living standards. **Data** are all information that can be used, collected, transformed, analyzed, and

translated into a form that is efficient for movement or processing. **Data** can be used for a lot, for example, we can use it in research analysis, by collecting information and using them to solve for an outcome and interpret that outcome, as mentioned by *Rosling Rönnlund: " \Data has both tremendous powers— to predict the future, solve our most pressing concerns, and make us wealthy beyond our wildest dreams—and mysterious inaccessibility."* [1] So, this set of information gives us the necessary tools to understand the reason for the fast growth in terms of technology that we experienced in the past centuries. However, even though there is growth, I still think there are certain limitations to take into account when using data as it is not everything that can be represented as data. But, with the data definition above, we will be able to distinguish if data can be processed for information, fact, knowledge, and other terms or if it is just limited to exploring and running them. Data always comes with a goal and these goals can depend on the questions that it is aiming to solve. Therefore, the data will be processed, structured, organized, or even contextualized to provide an answer that can be considered as information to a person or a system. This can also be used to evaluate if a fact is correct or not as it is run through experiments, therefore, it's expected to provide evidence of facts that are driven from personal knowledge. Then, this makes data the primary set of information that all reasoning and knowledge prerequisites before making a claim.

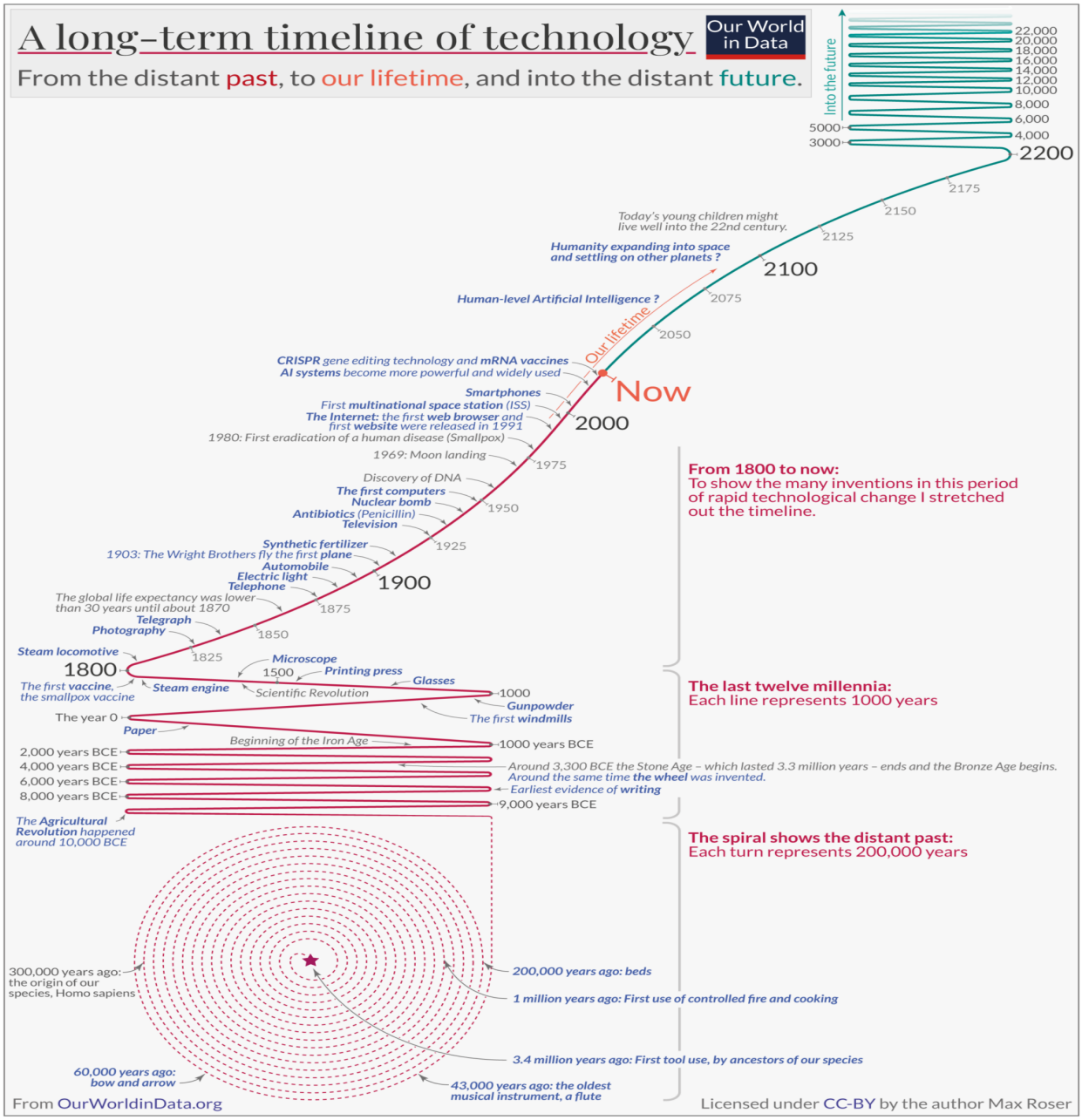
Nevertheless, this does not imply that all data we are presented with could be processed and used to solve problems and give an interpretation. Because the sources from which the data is collected are not accurate then there won't be a need to expect that, the based assumption could be valid, or if the data show biased. Meaning, their collection process favored one particular group's viewpoints points, or during the analysis are implemented by default as

[1] 1 *Book summary: Factfulness by Hans Rosling*. Sam Thomas Davies. (2023, March 29). https://www.samuelthomasdavies.com/book-summaries/business/factfulness/ page 2.

mentioned by *Carl Bergstrom & Jevin West: "Selection can have all sorts of interesting consequences, and when trying to understand patterns from data, it is worth thinking about whether there has been any selection bias or deliberate selection operating, and if so, how those factors affect the patterns you observe".*[2] To solve such issues and analyze, there are steps to follow, for example: Are the questions they chose to ask of the data in the first place associated with the data, Is the data processed with an expert(data scientist), is the processor used the right one for the type of data, is the programming language using the right one, is the line of code correct and is the code running, is the result valid, and can it be interpreted and produced a final result? Of course, if data are used with the necessary and quality resources it will produce outstanding results for the data scientist or researchers and an example of this see the image I attached below.

[2] Bergstrom, C. T., & West, J. Darwin. (2021). *Calling bullshit the art of skepticism in a data-driven world*. Amazon. https://www.amazon.com/Calling-Bullshit-Skepticism-Data-Driven-World/dp/0525509208 Chapter6

Ouamane Barry



[3]

Looking at the image we can see that there is a lot of different information provided, and this varies based on the event that happened in the human civilization. As we can see from the image, a reference to the starting point when humans used tools for the first time is noted with the red

[3] Roser, M. (2023, February 22). *Technology over the long run: Zoom out to see how dramatically the world can change within a lifetime*. Our World in Data. https://ourworldindata.org/technology-long-run

start, and at the end of the diagram, it shows the possible continuous innovation that would occur in the future.

Of course, when we see this data and the way it displays a timeline from which events that happened million years ago are revealed, we will be surprised. Because, if we think about it we might say that back in the day there were not any technological tools that could be used to collect or store data as there was no civilization. Therefore, those that come before writing existed are pretty much unaware of the importance of data but still, they were able to transmit data and this was done through oral tradition such as storytelling, chants, religious practices, and culture. Additionally, we can think about the past centuries, especially before and during the industrial revolution. During such a period most countries were facing development and were perplexed, and intimidated to have their data or information exposed to other countries. As it could cause them to equalize with other countries or even be dominated by other countries as in the case of colonization which is something that none of them was looking for. Therefore, data were more restricted from being accessed by every individual and it was mostly controlled by the government or legal institutions. Some of the ways they used it are measuring the increase in population, increase in production of goods and services, and increase in income as these were part of the addressing matter of development.

This back then, compared to the ways that are used, can be seen as one of the causes that limited innovation from fostering since then. Because with the world we are living in today, data are easily accessed by every individual, also because everything that we do is represented through a data form. So, due to this feasible and open data access, innovation has been growing the fastest of all time since civilization, see image above.

Ouamane Barry

Using data to study the past as shown in the images above is relevant because it helps us to understand not only our stand in terms of development but also the contribution we made, and also with the current information we can have an emphasized overview on what to expect in the future. However, what can make us certain that the result that we obtain from data that talked about the past or the future is accurate? See the first image above, which mentioned something that happened millions of years ago, for me whenever, I look back on this timeline, my brain developed a thousand questions. All these are based on how they found out this information and if their data was found, where did they find those data, and if they based their hypothesis on the present occurrence where are the similarities that they think could be used to emphasize that far of time? As a data scientist, these are questions that I can only answer based on assumptions from which I express conditions when interpreting the data. Because, for me, data are everything I can collect process, analyze and interpret, and this includes: books, articles, movies, institutional data measurement, and even personal experiences of my interaction on social media. I use different data processors such as programing language, and machine learning tools, and with this, I will generate responses to questions and prompts. Using these tools I can find ways to approach all types of data and implement designs to prioritize accuracy and reliability. I believe the learning process that I experienced from my data science(cs215), I am equipped with an awareness of all potential biased in data, and to take steps in mimicking them if needed. So, overall this would help analyze data from multiple sources and support evidence to ensure its accuracy.