Data Ethics In-Class Activity (May 23)

Today we will discuss two of the Ethics subtopics, Data Privacy and Validity. Refer to the corresponding articles listed in the previous activity document. There is no need to submit any written report about the topics or about the discussion. Be sure to stay in your breakout room even if you finish your group discussion early, as there is a class discussion at the end of the session.

Data Privacy

Your job is to discuss the Data Privacy topic first within your breakout group (for approximately 30 minutes) and then join the main room for a full class discussion of the topic.

Within your group:

1. Who read any of the articles about Data Privacy?
2. Which article(s) did you read?
3. Each person who read any of the articles should take a few minutes to give an overview of the article, specifically:
    a. Who wrote the article, what is their role or point of view?
    b. What are the main points of the article?
    c. What are the strong points of the article?
    d. What are the weak points (if any)?
    e. What did you learn from it or take away from it?

4. All others in the group then should discuss and ask questions about the article.
5. Be ready to discuss the following questions with the full class
    a. What is the GDPR?
    b. GDPR is a European effort, how does it relate to the USA?
    c. How might a Data Engineer be involved in GDPR compliance?
    d. Discuss the following questions:
        i. Popups everywhere. It’s annoying and the average internet user has no idea how to control/configure data privacy consent, so they just agree to everything.
        ii. Companies are scared, so they are spending bajillions protecting themselves. Bajillions that could be spent on things that actually benefit customers.

iii. The whole thing is toothless. Only Ireland can bring an actual judgment, and they are in the pocket of big tech. So there have not been many significant cases or judgements so far.

iv. It requires private data to be transparent and easily accessible by the users, and that makes it easier for hackers to obtain private data by impersonating users.

Validity

Discuss the Validity topic first within your breakout group (for approximately 30 minutes) and then join the main room for a full class discussion.

Within your group:

6. Who read articles about Validity?
7. Which article(s) did you read?
8. For each article read by at least one person in the group:
    a. Who wrote the article, what is their role or point of view?
    b. What are the main points of the article?
    c. What are the strong points of the article?
    d. What are the weak points (if any)?
    e. What did you learn from it or take away from it?
9. Discuss the following questions:
    a. The articles list many problems with data validity. Which of these problems could be helped by a Data Engineering approach?
    b. What specifically could/should a Data Engineer do to address the challenges listed in these articles?

Submit

Create a copy of this document (or create a new document if you prefer), and use it to answer the following question.

For each of the four major areas of Data Ethics, mention a situation that you have experienced that involved the corresponding area of Data Ethics. Say whether or not (in your opinion) the issue was handled satisfactorily. Finally, state how you might improve the handling of Data Ethics in similar situations in the future.

Use the in-class assignment submission form to submit your response(s).

Fairness:

In the Pokemon Go article, we also discussed how certains areas or communities had less access to certain places. Like poorer areas didn't have much compared to other areas that are wealthier. Like you have to drive much further to get to your dental appointments or even go to your local gym. Like the Pokemon Go article these wealthier areas had more infrastructure which in turn, meant there are more pokestops compared to areas with less infrastructure. We don't think this is handled fairly and it makes the situation worse since many areas and neighborhoods depends on what is nearby to them. One idea for improvement was to reduce/remove human bias and look at the general layout of the land rather than looking at the population and the wealth on certain areas. This can potentially give more people more overall access.

Validity:

For validity, we experienced it with the Trimet project. We experience validation and transformation and discuss what we would do with inaccurate data/human error. We also along with Genevieve about how the data is from October 2020. In the midst of Covid where it definitely has impacted the amount of Trimet riders during these times. We discuss that in the future we should compared and even use other data from different years since they made be more accurate and valid and that the year 2020 is a anomaly. I believe that this issue would be handled satisfactorily because we took a look at the data as a whole and study where the data came from and discuss steps on what to do with them. In the future, we discuss that one improvement would be to having someone with a different perspective to look at the data. This would help fight against human bias that may occurred in the data.

Ownership:

For data ownership, we talked about our experiences with how Google or Amazon knows so much about you and that the data that is being collected is being used against you. Such as ads, where you may be looking to buy a new phone and all of the sudden there is many links regarding phones on your browsers. At times, we kind of accept this especially if we are shopping certains items at certain sites, but we don't think this is ethical in all cases. Because there is times we do things and not know that the data is being collected. Like watching a YouTube video about a product, and now Google thinks you are interested in buying that particular product. One idea to improve data ethics in ownership is having more transparency between the company and the user. So the user knows exactly what and how the data is being used and if they are okay with that or not.

Privacy:

Similar to ownership, we talked about big companies like Google, Amazon, and Facebook about data privacy. These companies collect data about us constantly with many users not knowing or many aren't aware of the power of data collection. Companies can gather data about users and learn all about their tendencies and

preferences and the users wouldn't know any better. Lots of services do have a terms of service and is essentially a consent form for the data being gathered, however, the issue with these terms of services are that they are too easily ignored and easily bypassed. Often times they are extremely length, full of jargon, and simply require the user to click a box to confirm they have read it. There isn't an easy solution to allow users to have more control over their data privacy, but spreading awareness can be a first step.