**Open-ended questions**

**Namrata Acharya**

Consider implications of data collection, storage, and data biases you would consider relevant here considering Data Ethics, Business Outcomes, and Technical Implications

1. Ethical considerations: There are many ethical considerations one must keep in mind when analyzing data. For data collection, it is important to have prior and clear consent from the participants whose data is being used, regardless of how anonymous it is. In terms of storage, it is important to discard any data that could compromise the identity of the participant. For instance, it is not a good idea to store the participant's name, as combined with their major, university and year, it could give away their identity. It is also imperative that there are robust security measures in place to guard the data. If it is stored without much security, it could be at risk of being accessed by unwanted or malicious parties. Data biases could also arise, such as only surveying students from a particular major/year/university or being more active during data collection at certain times of the day than at others.

2. Business outcome implications: If data is not correctly collected or stored, it could create a negative impact for a business. If the data is not properly stored and gets compromised, customers lose faith in the business and could face risks due to privacy invasion. If the data collected has a lot of bias, the models being trained will be trained on that bias, and lead to misguided and biased insights. This could mean that the business could lose customers, invest in unprofitable scenarios or be unable to capitalize on important attributes.

3. Technical implications: It is important to safeguard the data and ensure the correct security protocols are in place. This requires considerable research into existing tools and consulting with good data engineers. Depending on the tool used to collect data, bias needs to be handled. For instance, letting customers enter their own information instead of relying on the server to do so, or expecting mandatory participation to avoid skewed data. If a lot of data bias is present in the data, the trained model's results are likely to be biased as well. This would make the data scientist's or analyst's job considerable harder, as they would try to get better results from a flawed dataset. It could also become more time consuming to reduce this bias from the data.

Given the work required to bring a solution like this to maturity and its performance, what considerations would you make to determine if this is a suitable course of action?

Some considerations I may make are:

1. The skill set of the current team. It is important that the team is well versed with the pipeline of data analytics, as there are many statistical intricacies that exist under the hood. The team should be comfortable with this process and the functionality of different predictive models.

2. The introduction of bias. If I choose to explore only a certain subset of data, I might miss important inferences and results, and draw incorrect conclusions. Analysts should be trained well to minimize this bias, otherwise the process may end up being futile.

3. Size of the dataset/quality of data: Since data is the backbone of this process, it is important that we have high quality data and a larger dataset to aid training. If the dataset is extremely skewed, unbalanced or has quality issues, it may take considerable time to process it and may not yield expected results.