# 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? **Answer:** Kelsey and Ethan
2. Which article(s) did you read? **Answer:** Kelsey read “Rethinking Patient Data Privacy In The Era Of Digital Health.“ and Ethan read “Police Intelligence-Gathering and Surveillance: Better management needed to protect civil rights”
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? **Answer:** Jared and Kelsey
7. Which article(s) did you read? **Answer:** Jared read “How Trustworthy Is Big Data?” and Kelsey read  “Predictably inaccurate: The prevalence and perils of bad big data.”
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).

I think **data ownership** is the one I can relate to the most. The fact that our web search data is used is followed by companies like Facebook and sold to companies that can profit off that knowledge. A lot of the conversation around data ownership came to light during the Cambridge Analytica revelations. They bought data from Facebook and used it to shape the way people voted here in the United States. They were able to identify people, such as myself, that were Bernie Sanders supporters and that were wary about the Clinton campaign. People didn't even realize that they were being played by disinformation and targeting those people and adding more skepticism to them about Clinton despite the fact that Clinton likely was more politically aligned with them than the alternative. The election was very close, so any small influence matters at the end of the day, so this issue mattered. I'm not sure the default should be that everyone owns their own data. I do think a lot of value is created with these big companies using our data. I do think there should be more transparency though. We should know who has our data and get some idea of what they're using it for. If people knew that they were being politically targeted then perhaps they wouldn't have been influenced with ads and propaganda as easily. I want ads that appeal to my tastes and preferences, but I don't want to be manipulated by malicious actors. Has this issue been handled satisfactorily? Somewhat. GDPR and Apple have addressed these in real ways, but there is still an issue. The issue is tradeoffs too. As a data engineer or scientist, I would hope to respect peoples data rights and not use their data to affect them in a negative way at all. It's an interesting and new topic for our society and we need to talk about it more and figure it out so that we are creating wealth but also not approaching a more dystopian future.

**Data privacy** is somewhat similar to ownership. People do not realize what they're agreeing to when they click the privacy notices when they visit a website. Just the fact that we get these notifications is a sign that we're moving in the right direction. But people's privacy are still being invaded in my opinion. I would like to not invade peoples privacy, but I am also a believer in using data to generate value and increase insights, therefore I think there needs to be some kind of balance and certainly more regulation about collection practices and transparency.

Going back to elections, for **Data Validation**, I was thinking about the 2020 presidential election. I recall reading The Economist predictions for the winner. They claimed to be using advanced machine learning models to make the prediction that Biden was going to win by a very large percentage, somewhere in the 8 to 10 percent range. I recall it was higher than most media predictions. So either the machine learning model was off, or the data was not valid. Considering the pollsters were so off in general, I would say that there was a validation problem. Has this been handled satisfactorily? I'm not sure, we'll have to see what happens during the next couple of election cycles. It seems as though they were far off in 2016 and 2020 because they were undercounting Trump supporters, the theory being they are less likely to talk to pollsters. Or that there were hidden Trump supporters that lied about who they were supporting to pollsters. I would think that the loss of credibility will shift data collection practices in the future. As a professional, I would be looking at these discrepancies with the polls and with reality and come up with models that would weigh Trump-like voters more heavily when collecting polling data.

As far as Data **Fairness** goes, I would assume that, as a white man, I have benefited from data collection and machine learning models. A high percentage of workers data and machine learning workers are white and are men, and that these biases are baked into their models. This has been more discussed lately, since these biases have been discovered, and since the BLM protests in 2020, so I would hope that this has been addressed satisfactorily. As a professional, I would definitely be conscious about modeling and collecting data to make sure that everyone is being counted fairly and that all of these tools go towards benefiting all members of society, regardless of race, gender, or sexual/gender identification.