Data Ethics In-Class Activity (May 18)

Today we will discuss two of the Ethics sub-topics, Data Ownership and Fairness. 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. Just be ready to participate fully in the class discussion.

Data Ownership

Your job is to discuss the Data Ownership 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 Ownership?
   Chinmay Tawde.

2. Which article(s) did you read?
   Ownership : Who Owns Your Fingerprint? You Don't.

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?
      - Nigel Jaquiss wrote the article and he is a news reporter who mainly writes about politics, their role is to explain the story of an employee whose name is Chris Dresel who has raised an issue regarding biometric data collection and storage in the state of Oregon.

   b. What are the main points of the article?
      - The main points of the article are related to How there is no law regarding biometric data collection and storage in the state of Oregon. The article talks about a bill "Senate Bill 284" which was created to tackle such issues and bring in some regulations. The state senator "Monnes Anderson" played a major role in drafting and supporting this bill. In the end, the bill, although having a lot of meaning and value, didn't pass and must go through multiple iterations to actually let it come to fruition.

   c. What are the strong points of the article?
      i. Many employers and employees have no clue what happens to the biometric data that the companies utilize on a day to day basis. Who stores the data and what happens when the employee exits the company.
      ii. Biometric data such as fingerprints are more secure as they are difficult to replicate and add a whole new level of security as compared to passwords

and pins. But this strength is also their weakness when it comes to data breaches, etc. You cannot modify or alter the biometric details.

iii. Identity theft is a major issue, a bank account which is made secured with pins and passwords can be changed once compromised. But biometric data pertaining to one's physical information cannot be altered or modified once compromised. The person can then utilize your digital identity as he/she pleases without any means to stop them.

iv. There are very limited states that have laws regarding biometric data, those states are Texas, Illinois, and Washington. Those laws were highly opposed by big tech companies as such laws would then impact their market

4. What are the weak points (if any)?
    - None.

5. What did you learn from it or take away from it?
    - Biometric is one of the most secure ways of authentication assuming it can't be stolen. Else it is one of the most impactful security threats as once compromised cannot be altered or modified and can help the attacker to wander the digital world masquerading as you and there's not much we would be able to do about it.

6. Do you recommend that your teammates read this article?
    - No, As I have covered most of the points of the article and as it is mainly focused on the journey of the bill rather than providing any in-depth knowledge on regulation and usage of biometrics. So it wouldn't add any more value to spending time and reading articles.

7. All others in the group then should discuss and ask questions about the article.
8. Discuss the following questions:
    a. For each person on the team, have you ever encountered instances where your personal data was being used in a surprising way?
    Deepa Hegde: Random searches out of curiosity are used against us to promote and advertise.

    Chinmay Tawde: Agrees with the other two members. Similar experience.

    Varun Jaisundar Raju: Promotional emails, I tried searching for flight tickets to a trip back to my home country and now I am receiving so many unwanted emails which are promotional in nature.

    b. Who do you think should "own" data about you?
    Deepa Hegde: Whoever we give permission to

Chinmay Tawde: First myself, I should have total control of who I give permission to and also be able to revoke the permission as and when needed.

Varun Jaisundar Raju: Agrees to above both members

c. Can you propose an ethical yet profitable solution to the data ownership issue?
    - Absolute transparency of what is being collected and proving the control to the person who's data is being collected to add/modify/remove the information that he/she provides.

d. List all of the general categories of data about a person that might be collected automatically by computer systems. Rank them by how comfortable you are with somebody else owning that data (and using it for whatever legal purpose they choose)

We will rate from (1 to 6) as we have identified 6 different categories or types of data, 1 being the most comfortable to share with and 6 being the least.

Chinmay

- Location - 3
- Search history - 3
- Personal preferences - 3
- Photos & Videos - 5
- Voice & Text messages - 6
- Health Data or Medical Data - 2

Deepa

- Location - 2
- Search history - 1
- Personal preferences - 6
- Photos & Videos - 2
- Voice & Text messages - 2
- Health Data or Medical Data - 6

Varun

- Location - 2
- Search history - 1
- Personal preferences - 3
- Photos & Videos - 4
- Voice & Text messages - 6
- Health Data or Medical Data - 5

Fairness

Your job is to discuss the Fairness 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:

9. Who read any of the articles about Fairness?
   Varun Jaisundar Raju

10. Which article(s) did you read?
    Is Pokemon go racist?

11. 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?
       Allana Akhtar
    b. What are the main points of the article?
       How the racial discrimination is done through a Pokemon Go game
       Pokemon stops and gyms were mainly concentrated where white and rich people live
       There was racial discrimination as Bogado observed there were hardly few poke stops and gyms where most black and underprivileged people reside

    c. What are the strong points of the article?
       - Racial discrimination on digital platforms shows the world where poor and poverty regions are located
       - People who belong to African and poverty regions would have a huge disadvantage when compared to others even though they are contributing towards the game income
       - Bogado also says in her article this racial discrimination has even happened earlier with amazon prime delivery and google high speed internet link
       - Jeffrey Vagle, executive director of the Center for Technology, Innovation and Competition at the University of Pennsylvania in Philadelphia these racial inequities would go even further
       - Niantic CEO - John Hankle says even
       - Overall, Hispanics and African Americans are not well represented in tech. They make up 6% and 3% of the Silicon Valley tech workforce, respectively
       - At google they should also hire women employees which would “People who have critical minds and orientation to looking at social problems could be hugely valuable in technology development,”

d. What are the weak points (if any)?
   Pokemon go founders say they haven't done this one purpose they have taken the map from ingress which is an augmented reality gaming platform so they have no clue about this inequality
e. What did you learn from it or take away from it?
   Racial Discrimination in digital platform
f. Do you recommend that your teammates read this article?
   No, I would not recommend this. Because I have pretty much covered everything in my discussion and this article only explains how racial discrimination is done in various digital platforms

12. All others in the group then should discuss and ask questions about the article.
13. Discuss the following questions:
    a. Are you aware of any other issues/reports of unfairness related to data analysis? Bias? Other types of incorrect conclusions?
       Varun Jaisundar Raju: No
       Chinmay Tawde: No
       Deepa: No

    b. What can a Data Engineer do to prevent or lessen the impact of unfairness?
       Varun Jaisundar Raju: Make the game available globally to provide fair share distribution of data irrespective of geographical location.
       Chinmay Tawde: Research and include data from various sources which will not only make the dataset more accurate but also make it more valuable.
       Deepa: Seek help from someone who have neutral opinion to review the data

    c. How would you actually find/seek fairness issues earlier?
       Varun Jaisundar Raju: No
       Chinmay Tawde: It is difficult to detect fairness issues, but depending on the topic or dataset, we can make sure to perform some statistical analysis to figure out how diverse our dataset is.
       Deepa: We can review the results to check if there is any data that is concentrated based on certain opinions.

    d. When is appropriate/inappropriate to use “surrogates” in data analysis?
       Chinmay Tawde: No idea.
       Varun Jaisundar Raju: No idea
       Deepa: Don’t really know what surrogates mean in data analysis

    e. Was Pokemon Go racist?
       Varun Jaisundar Raju: Yes because Pokemon, Locations , Gym and Egg incubators to evolve Pokemonsters were majorly found near white and rich american regions as compared to african and poverty regions

Chinmay Tawde: I don't think so, major white regions are well covered through google maps whereas African and poverty regions are not. Simply putting game elements in random places would lead to accidents and even deaths. Such incidents were recorded in India (my home country) where these things were placed on train tracks, in the middle of roads, etc.
So thus, they probably reduced and decided to only feature those elements in a safe known environment.
Deepa: According to the paper reviewed by my teammate, pokemon go does seem to be racist as it focused mostly on white and rich american regions

f. How would you notice/find the type of bias found in Pokemon Go? How could they have noticed this before it turned into a news article?
Varun Jaisundar Raju: Color and Economic Bias
Chinmay Tawde: I think it was intentional and could have been very easily noticeable depending on the game algorithm and decisions.
Deepa: statistical analysis of the data would have revealed the data that were concentrated in specific regions.

g. Can a Data Engineer produce/provide bias-free data?
Chinmay Tawde: Often times it's not the data engineer's decision to produce biased data. It's the nature of the research or the factual nature of the dataset itself is biased.
Deepa: Yes it possible by reviewing the data over and over

h. Should a Data-related team include Social Scientists?
Chinmay Tawde: Yes, a SME (Subject Matter Expert) would definitely be able to provide more insightful knowledge and thus, refine the dataset.
Deepa: Yes. This would help data engineers to understand the data better and come up with efficient predictions/results

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 below. 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?
   Chinmay Tawde, Deepa Hegde
2. Which article(s) did you read?
   Rethinking Patient Data Privacy In The Era Of Digital Health
   How one country blocks the world on data privacy

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?

Lisa Bari is a healthcare information technology (IT) policy expert who recently served as the health IT and interoperability lead at the Centers for Medicare and Medicaid Services' Innovation Center, and as a senior technical adviser on artificial intelligence and emerging technology to the Department of Health and Human Services' Value-Based Transformation Initiative.

Daniel P. O'Neill has spent 10-plus years working in healthcare technology and recently served as a policy fellow on the professional staff of the US Senate HELP Committee, which has jurisdiction over the 21st Century Cures Act, the Public Health Service Act, and the Employee Retirement Income Security Act.

Their role is to just explain the current legislation and regulations around the health data that is being gathered, monitored and commercialized

    b. What are the main points of the article?

- It talks about the current rules and regulations around the health data that is being gathered, monitored and commercialized.
- It talks about Health Insurance Portability and Accountability Act (HIPAA), its scope,which entities are regular under it and individual rights surrounding it.
- It also talks about other such regulations such as GDPR European General Data Protection Regulation. its scope,which entities are regular under it and individual rights surrounding it.
- Combining that it then proposes the new and improved Adapted/Extended HIPAA regulations. Refer image below:

	HIPAA (current)	GDPR	Adapted / extended HIPAA (proposed)
Health data in scope	Individually identifiable health information (IIHI), relating to individual health (physical or mental), provision of care or payment for provision of care, *when that IIHI is held or transmitted by a covered entity or its business associate*	"Personal data" which includes direct or indirect identifiers and "expresses the physical, physiological, genetic, mental, commercial, cultural or social identity" individuals. Health data is a special category with heightened protection.	IIHI relating to an individual health (physical or mental), provision of care, or payment for provision of care, regardless of who collects, holds, processes or transmits that data
Regulated entities	*Covered entities (CEs)* – Health plans, health care providers and health care clearinghouses *Business associates* – person or organization performing functions on behalf of (or providing services to) a CE	*Controller* – Any person or organization which determines the purposes or means of processing personal data *Processor* – person or organization processing data on behalf of controller	*Custodian* – Any person or organization collecting or holding IIHI, or controlling the processing thereof *Processor* – person or organization processing data on behalf of custodian
Permitted uses of personal health data	Health care treatment, payment and operations	Treatment, public health, research, judicial proceedings, substantial public interest, by informed consent, or when processing is in the "vital interests" of person unable to consent	Health care treatment, payment and operations, regardless of the entities involved in those activities, when appropriately disclosed to the individual. All other uses require consent.
Security standards	✓	✓	✓
Breach notification requirements	✓	✓	✓
Individual right to: Access | Amend | Delete	✓ | ✓ | ✗	✓ | ✓ | ✓	✓ | ✓ | ✓

c. What are the strong points of the article?

- It talks about how companies try to gather personal information and commercialize it. They de-identify the information but it is very easy to connect the dots and re-identify the person whose information is shared.
- The article also discusses how we can create and append new health data but we are never allowed to delete it or have no control over the commercialization of it.
- Health data is personal and if used in a more aggressive way can lead to personal insights in a person's life which can lead to adverse mental effects.

d. What are the weak points (if any)?

- The article is too technical and not beginner friendly and tries to achieve a lot more in a few complex words.

e. What did you learn from it or take away from it?

- Currently the regulations related to collections and commercialization of health data are not stringent. Thus, a major change is needed to regulate these laws to protect people's privacy and personal information.

---

a. Who wrote the article, what is their role or point of view?

    - Nicholas Vinocur. He is a journalist and he is putting forward the thoughts of all the other EU members that are a part of GDPR

b. What are the main points of the article?
    - This is an article about the GDPR (General Data Protection Regulation) rule that was brought into action in the Europe Union during the year 2018
    - It is a privacy and security law enforcement by the EU in order to protect the privacy of their citizens and residents.
    - GDPR sets guidelines for collection and processing of personal information from the people who live in EU

c. What are the strong points of the article?
    - It explains the severity at which the GDPR was enforced across the EU
    - It explains how the EU is trying to protect private data of their citizens and residents

d. What are the weak points (if any)?
    - This weak point is not about the article but instead it is about the GDPR itself. The power should not reside with a single country. It should be equally distributed among different countries by having their respective representatives. This would solve the conflict of interest to an extent according to me

e. What did you learn from it or take away from it?
    - The differences and conflicts among the members when there is a law enforced.
    - When there is power residing with someone, he/she should be very vigilant and unbiased to take decisions in such situations

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?
        - General Data Protection Regulation - This is to protect people's private data in the European region

	GDPR
Health data in scope	"Personal data" which includes direct or indirect identifiers and "expresses the physical, physiological, genetic, mental, commercial, cultural or social identity" individuals. Health data is a special category with heightened protection.
Regulated entities	*Controller* – Any person or organization which determines the purposes or means of processing personal data *Processor* – person or organization processing data on behalf of controller
Permitted uses of personal health data	Treatment, public health, research, judicial proceedings, substantial public interest, by informed consent, or when processing is in the "vital interests" of person unable to consent
Security standards	✓
Breach notification requirements	✓
Individual right to: Access | Amend | Delete	✓ | ✓ | ✓

b. GDPR is a European effort, how does it relate to the USA?
   - Currently in the USA we have HIPAA law, which tries to protect personal health data related information. But it is not as strict as needed. So taking notes from the GDPR. The USA is trying to adopt a new improved HIPAA law that will protect and regulate the collection and commercialization of health related data.

c. How might a Data Engineer be involved in GDPR compliance?
   - One of the major activities of a Data Engineer is to validate and transform data. The data engineer can help protect personal data by abiding by GDPR compliance laws.

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.
        - Agreed, this happens all over the internet due to rampant commercialization and personalization using targeted advertisements. Even day-to-day internet users don't have any way to minimize such aggressive commercialization except using extensions like AdBlocks, etc.

    ii. Companies are scared, so they are spending bajillions protecting themselves. Bajillions that could be spent on things that actually benefit customers
        - I don't really agree, because they are earning more money trying to sell or benefit by collecting personal data. Thus, they are more than willing to spend Bajillions of money to protect themselves so they can profit in the long run.

    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.
        - Ireland plays a major role and is currently biased as being in the pocket of big tech companies but overall the world is being impacted and together we are moving towards a more

    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.
        - The data should be transparent and easily accessible but it also should be secure. For example, Bank applications. In those applications we can track our transaction history, etc. which reveals many private details but the app itself is very secure to not use this information for commercial purposes.

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?
   Deepa Hegde, Varun Jaisundar Raju

7. Which article(s) did you read?
   "6 problems that make data unreliable and how to fix them"
   How Trusty is 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?
      - Transform-Solutions. It is a consulting firm who are specialized in ecommerce services, data services and digital marketing.

   b. What are the main points of the article?
      - This article talks about the common problems which make the data unreliable
      - In order to avoid faulty predictions which can result in irreversible consequences, it is very important to ensure trustworthiness and reliability of content and data.
      - This article focuses on how to identify the common problems that make the data unreliable and how it can be fixed.

   c. What are the strong points of the article?
      - It covers all the important points that a data engineer needs to know which will help in identifying the problems
      - Well phrased, any layman can understand the article

   d. What are the weak points (if any)?
      - None

   e. What did you learn from it or take away from it?
      - Better understanding of the subject
      - Easy to relate to relate as we have identified some of the problems and fixed them as a part of our in class assignment

---

9. 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?
      Elliot Timo wrote this article
      The main idea of this article is the role of Big Data and its applications in multiple projects and organizations.
   b. What are the main points of the article?
      - Data needs to be validated no matter how accurate it is.

- Different Organizations use big data in different ways
- Key Practical Dimensions that organizations should consider when using big data

c. What are the strong points of the article?
   - Big data can help to predict people's emotions which is one of the important factors
   - Big data requires intensive processing, interpretations, and the use of data science algorithms
   - Any small mistake can lead to a disaster
   - Big data can also be dangerous because it can start collecting more of user personal data which might affect one's privacy

d. What are the weak points (if any)?
   - The big data collected from various devices such as sensors, and data lakes are just the approximate values it does not indicate the accurate values
   - Big data is captured in low-level detail
   - All parameters present in big data may not be used.

e. What did you learn from it or take away from it?
   - Role of Big Data in various industries.
   - Importance of Big Data, advantages and disadvantages of big data
   - Threats of big data

10. 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?
       - Data should be free from bias
       - Human errors and lack of precision - This can be avoided by automating all the data related operations and avoiding manual entries
       - Reliability and validity of the data sources - make sure to fetch data from a reliable source to get better data analysis results

    b. What specifically could/should a Data Engineer do to address the challenges listed in these articles?
       - Follow all the privacy related guidelines to secure the data he/she is working on

- Use of Data Engineering tools such as pandas, python in order to transform, validate and transform data
- Automate steps wherever possible to avoid human errors