BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Ethics in Data Science

Biases & Fairness

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Topics

- Bias and Fairness
  - Types of Bias
  - Identifying Bias
  - Evaluating Bias
- Being a data skeptic – examples of misuse of Data
- Doing Good Data Science
- Five C’s
- Ethical guidelines for Data Scientist
- Ethics of data scraping and storage

Biases in Data Science

Bias and Fairness

Data Bias

- Data bias is that the available data is not representative of the population or phenomenon of study
- Bias also denotes:
  - Data does not include variables that properly capture the phenomenon we want to predict
  - Data includes content produced by humans which may contain bias against groups of people
- Except for data generated by carefully designed randomized experiments, most organically produced datasets are biased
- Data bias occurs due to structural characteristics of the systems that produce the data

Bias and Fairness

Data Biases

- Sample bias
- Exclusion bias
- Measurement bias
- Recall bias
- Observer bias
- Racial bias
- Association bias
- Automation bias
- Coverage bias
- Non-Responsive bias
- Group attribution bias
- Implicit bias
- Confirmation bias
- Experimenter's bias

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Bias and Fairness

Sample or Selection Bias

- Sample or Selection bias occurs if
  - Proper randomization is not used during data collection
  - a data set's examples are chosen in a way that is not reflective of their real-world distribution
  - a dataset does not reflect the realities of the environment in which a model will run
- Example 1:
  - Certain facial recognition systems trained primarily on images of white men
  - These models have considerably lower levels of accuracy with women and people of different ethnicities

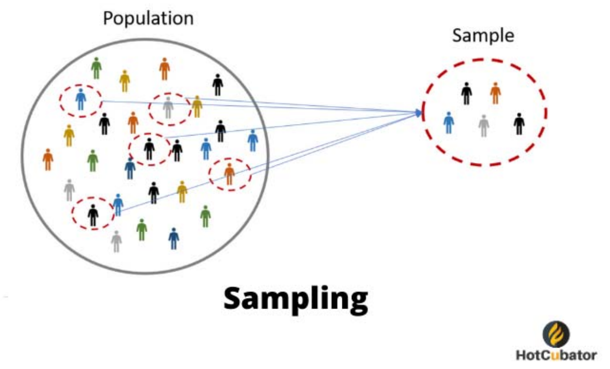


Image Source: https://hotcubator.com.au/research/a-complete-guide-to-sampling-techniques/

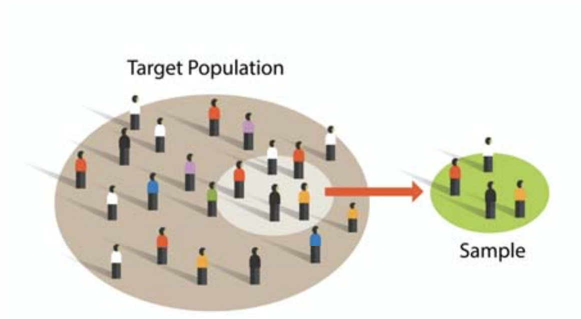


Image Source: https://medium.com/analytics-vidhya/what-is-the-sampling-bias-bbde6560fa

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Exclusion Bias

- Is most common at the data preprocessing stage
- Most often it’s a case of deleting valuable data thought to be unimportant
- However, it can also occur due to the systematic exclusion of certain information
- For example:
  - Imagine you have a dataset of customer sales in America and Canada
  - 98% of the customers are from America, so you choose to delete the location data thinking it is irrelevant
  - This means your model will not pick up on the fact that your Canadian customers spend two times more.

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Measurement Bias

- This type of bias occurs when faulty measurements result in data distortion
- Measurement bias results from poorly measuring the outcome you are measuring
- Example:
  - In image recognition datasets, training and test data comes from images taken from different types of cameras
  - The survey interviewers asking about deaths were poorly trained and included deaths which occurred before the time period of interest.
    - This would lead to an overestimate of the mortality rate because deaths which should not be included are included.
  - A machine to measure hemoglobin malfunctioned and was not checked, as should be done every day. It measured everyone's hemoglobin as 0.3 g/L too high.
  - This would lead to an underestimate of the prevalence of anemia because the readings would overestimate the hemoglobin for everyone measured by that team.

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Recall Bias

- When people remember past events, they don't usually have a complete or accurate picture of what happened.
- Recall bias is a systematic error that occurs when participants do not remember previous events or experiences accurately or omit details
- The accuracy and volume of memories may be influenced by subsequent events and experiences
- Our brains continuously rewrite memories, clouding them with current events, or even editing them completely
- Five years after an event, 50% of the memories for that event are completely lost
- Recall bias is a problem in studies that use self-reporting, such as case-control studies and retrospective cohort studies

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Recall Bias

- Bias in recall can be greater when the study participant has a poorer recall in general, and when events over a longer time interval are being asked about
- Other issues that influence recall include age, education, socioeconomic status and how important the condition is to the patient
- Furthermore, undesirable habits such as smoking or eating unhealthy foods tend to be underreported, and are therefore subject to recall bias
- Example
  - Parents of children diagnosed with cancer may be more likely to recall infections earlier in the child's life than parents of children without cancer
  - This may lead to observing an entirely or partially untrue association between childhood infection and cancer
  - Recall can be particularly problematic when the events of interest happened a long time ago.

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Observer Bias

- The tendency for observers to record data that may be biased as a result of personal expectations or motives, rather than recording what actually happens
- Observer bias is a type of detection bias that can affect assessment in observational and interventional studies
- "Systematic difference between a true value and the value actually observed due to observer variation"
- For example:
  - In the assessment of medical images, one observer might record an abnormality but another might not
  - Different observers might tend to round up or round down a measurement scale
  - Color change tests can be interpreted differently by different observers
  - Where subjective judgement is part of the observation, there is great potential for variability between observers, and some of these differences might be systematic and lead to bias
  - Observer bias may also occur if the researcher has a preconceived idea of what the blood pressure ought to be, leading to arbitrary adjustments of the readings.

towardsdatascience.com - Understanding Data Bias: Types and sources of data bias
Lionbridge.ai - 7 Types of Data Bias in Machine Learning

Bias and Fairness

Reporting Bias

- Reporting biases is an umbrella term that covers a range of different types of biases
- It is a distortion of presented information from research due to the selective disclosure or withholding of information by parties involved with regards to the topic selected for study and the design, conduct, analysis, or dissemination of study methods, findings or both
- Researchers have previously described seven types of reporting biases:
  - publication bias, time-lag bias, multiple (duplicate) publication bias, location bias, citation bias, language bias, and outcome reporting bias

Bias and Fairness

Reporting Bias

- Occurs when the frequency of events, properties, and/or outcomes captured in a data set does not accurately reflect their real-world frequency
- Can arise because people tend to focus on documenting circumstances that are unusual or especially memorable
- EXAMPLE:
  - A sentiment-analysis model is trained to predict whether book reviews are positive or negative based on a corpus of user submissions to a popular website
  - The majority of reviews in the training data set reflect extreme opinions (reviewers who either loved or hated a book), because people were less likely to submit a review of a book if they did not respond to it strongly
  - As a result, the model is less able to correctly predict sentiment of reviews that use more subtle language to describe a book

Bias and Fairness

Racial Bias

- An algorithm used to identify eligibility for care management programs reduced the number of black patients identified for extra care by more than half
  - Removing this disparity would result in a 28.8 percent rise in black patients receiving additional services.
- The bias was not intentional. However, in the algorithm, healthcare costs were used as a proxy measure for health needs
- Because black patients typically spend less money on healthcare, the algorithm underestimated the risk for these individuals
  - https://healthitanalytics.com/news/eliminating-racial-bias-in-algorithm-development
- Read the below article for racial biases in facial recognition algorithms
  - https://towardsdatascience.com/addressing-racial-bias-in-ai-a-guide-for-curious-minds-ebdf403696e3

Bias and Fairness

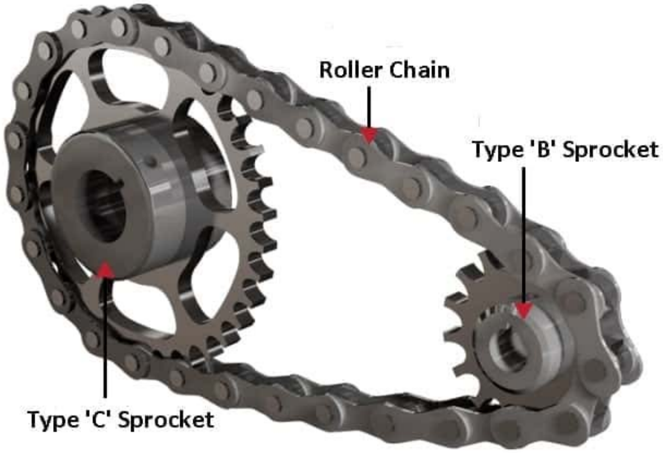
Association Bias

- Association bias is a tendency to be easily influenced by associations
- For example:
  - We usually associate high prices with quality goods. This association has proved to be true in our past experiences. Goods purchased at high prices did turn out to be high quality
  - So, in a world where we are overwhelmed with stimulus and information, it’s a lot easier to use the mental shortcut that high prices = high quality rather than to examine each product we buy with a magnifying glass to attempt to prove its quality
- Association bias could also be due to our association or identification with certain groups, political parties, sports clubs, etc.,.

Bias and Fairness

Automation Bias

- Tendency to favor results generated by automated systems over those generated by non-automated systems
  - irrespective of the error rates of each



CHAIN DRIVE

- EXAMPLE:
  - Software engineers working for a sprocket manufacturer were eager to deploy the new "groundbreaking" model they trained to identify tooth defects, until the factory supervisor pointed out that the model's precision and recall rates were both 15% lower than those of human inspectors.

Bias and Fairness

Coverage Bias

- Data is not selected in a representative fashion
- EXAMPLE:
  - A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product. Consumers who instead opted to buy a competing product were not surveyed, and as a result, this group of people was not represented in the training data.

Bias and Fairness

Non-Responsive Bias

- Also called as participation bias
- Non-response (or late-response) bias occurs when non-responders from a sample differ in a meaningful way to responders (or early responders)
- Data ends up being unrepresentative due to participation gaps in the data-collection process.
- EXAMPLE:
  - A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product and with a sample of consumers who bought a competing product. Consumers who bought the competing product were 80% more likely to refuse to complete the survey, and their data was underrepresented in the sample.

Group Attribution Bias

- Tendency to generalize what is true of individuals to an entire group to which they belong
- Two key manifestations are:
  - In-group bias
    - A preference for members of a group to which you also belong, or for characteristics that you also share.
    - EXAMPLE:
      - Two engineers training a résumé-screening model for software developers are predisposed to believe that applicants who attended the same computer-science academy as they both did are more qualified for the role.
  - Out-group homogeneity bias
    - A tendency to stereotype individual members of a group to which you do not belong, or to see their characteristics as more uniform.
    - EXAMPLE:
      - Two engineers training a résumé-screening model for software developers are predisposed to believe that all applicants who did not attend a computer-science academy do not have sufficient expertise for the role.

Bias and Fairness

Implicit Bias

- Occurs when assumptions are made based on one's own mental models and personal experiences that do not necessarily apply more generally
- EXAMPLE:
  - An engineer training a gesture-recognition model uses a head shake as a feature to indicate a person is communicating the word "no." However, in some regions of the world, a head shake actually signifies "yes."

Confirmation Bias

- Confirmation bias occurs when an individual looks for and uses the information to support their own ideas or beliefs.
- Model builders unconsciously process data in ways that affirm preexisting beliefs and hypotheses
- Example:
  - Imagine that a person holds a belief that left-handed people are more creative than right-handed people
  - Whenever this person encounters a person that is both left-handed and creative, they place greater importance on this "evidence" that supports what they already believe
  - This individual might even seek proof that further backs up this belief while discounting examples that don't support the idea
- Confirmation biases impact how we gather information, but they also influence how we interpret and recall information. For example
  - people who support or oppose a particular issue will not only seek information to support it, they will also interpret news stories in a way that upholds their existing ideas.

Source: https://www.verywellmind.com/what-is-a-confirmation-bias-2795024

Bias and Fairness

Experimenter's Bias

- Experimenter's bias is a phenomenon in which the outcome of an experiment tends to be biased towards a result expected by the experimenter
- A model builder may actually keep training a model until it produces a result that aligns with their original hypothesis
- EXAMPLE:
  - An engineer is building a model that predicts aggressiveness in dogs based on a variety of features (height, weight, breed, environment). The engineer had an unpleasant encounter with a hyperactive toy poodle as a child, and ever since has associated the breed with aggression. When the trained model predicted most toy poodles to be relatively docile, the engineer retrained the model several more times until it produced a result showing smaller poodles to be more violent.

Bias and Fairness

References

- https://catalogofbias.org/biases/
- https://www.youtube.com/watch?v=wEwGBIr_RIw&t=149s

Data Ethics

What is Data Ethics?

- With the use of data comes the misuse of data
- For example:
  - In the 2016 election, a company called Cambridge Analytica improperly accessed Facebook data and used that for political ad targeting
  - In 2018, an autonomous car being tested by Uber struck and killed a pedestrian (there was a "safety driver" in the car, but apparently she was not paying attention at the time)
  - Algorithms are used to predict the risk that criminals will reoffend and to sentence them accordingly
    - Is this more or less fair than allowing judges to determine the same?
  - Some airlines assign families separate seats, forcing them to pay extra to sit together

Data Ethics

What is Data Ethics?

- Should a data scientist have stepped in to prevent this?
- "Data ethics" intends to provide answers to these questions, or at least a framework for wrestling with them
- If you take the average of every definition we can find on ethics, we end up with something like:
  - Ethics is a framework for thinking about "right" and "wrong" behavior
- Data ethics, then, is a framework for thinking about right and wrong behavior involving data

Data Ethics

Should I Care About Data Ethics?

- You should care about ethics whatever your job is
- Decisions made by individuals working on technology problems (whether data-related or not) have potentially wide-reaching effects
- A tiny change to a news discovery algorithm could be the difference between millions of people reading an article and no one reading it
- A single flawed algorithm for granting parole that’s used all over the country systematically affects millions of people, whereas a flawed-in-its-own-way parole board affects only the people who come before it
- The broader the effects of our work, the more we need to worry about these things

Building Bad Data Products

- Some "data ethics" issues are the result of building bad products
- For example:
  - Microsoft released a chat bot named Tay that parroted back things tweeted to it, which the internet quickly discovered and enabled them to get Tay to tweet all sorts of offensive things
  - It seems unlikely that anyone at Microsoft debated the ethicality of releasing a "racist" bot
  - Most likely they simply built a bot and failed to think through how it could be abused
- Another example:
  - Google Photos at one point used an image recognition algorithm that would sometimes classify pictures of black people as "gorillas"
  - Again, it is extremely unlikely that anyone at Google explicitly decided to ship this feature (let alone grappled with the "ethics" of it)

Building Bad Data Products

- A likely problem here is some combination of bad training data, model inaccuracy, and the gross offensiveness of the mistake
  - If the model had occasionally categorized mailboxes as fire trucks, probably no one would have cared
- So, what's the solution?
  - How can we ensure that our trained model won't make predictions that are in some way offensive?
- Of course we should train (and test) our model on a diverse range of inputs, but can you ever be sure that there isn't some input somewhere out there that will make your model behave in a way that embarrasses you?
- This is a hard problem
  - Google seems to have "solved" it by simply refusing to ever predict "gorilla."
- We should think about how the things we build could be abused

Data Ethics

Trading Off Accuracy and Fairness

- Imagine we are building a model that predicts how likely people are to take some action
- You do a pretty good job

	Actual		
Prediction	No Action	Action	%
Unlikely	100	25	20
Likely	50	75	60

- Of the people you predict are unlikely to take the action, only 20% of them do
- Of the people you predict are likely to take the action, 60% of them do
- Seems not terrible

Data Science from Scratch, 2nd Edition by Joel Grus

Data Ethics

Trading Off Accuracy and Fairness

- Now imagine that the people can be split into two groups: A and B.
- Some of your colleagues are concerned that your model is unfair to one of the groups
- Although the model does not take group membership into account, it does consider various other factors that correlate in complicated ways with group membership
- When we break down the predictions by group, we discover surprising statistics

Group		Actual		
	Prediction	No Action	Action	%
A	Unlikely	80	20	20
A	Likely	10	15	60
B	Unlikely	20	5	20
B	Likely	40	60	60

Data Science from Scratch, 2nd Edition by Joel Grus

Data Ethics

Trading Off Accuracy and Fairness

- Is your model unfair? Data scientists make a variety of arguments:
- Argument 1:
  - The model classifies 80% of group A as "unlikely" but 80% of group B as "likely."
  - This data scientist complains that the model is treating the two groups unfairly in the sense that it is generating vastly different predictions across the two groups

Group		Actual		
	Prediction	No Action	Action	%
A	Unlikely	80	20	20
A	Likely	10	15	60
B	Unlikely	20	5	20
B	Likely	40	60	60

Trading Off Accuracy and Fairness

- Argument 2:
  - Regardless of group membership, if we predict "unlikely" we have a 20% chance of action, and if we predict "likely" you have a 60% chance of action
  - This data scientist insists that the model is "accurate" in the sense that its predictions seem to mean the same things no matter which group you belong to

Group		Actual		
	Prediction	No Action	Action	%
A	Unlikely	80	20	20
A	Likely	10	15	60
B	Unlikely	20	5	20
B	Likely	40	60	60

Data Ethics

Trading Off Accuracy and Fairness

- Argument 3:
  - 40/125 = 32% of group B were falsely labeled "likely," whereas only 10/125 = 8% of group A were falsely labeled "likely."
  - This data scientist (who considers a "likely" prediction to be a bad thing) insists that the model unfairly stigmatizes group B

Group		Actual		
	Prediction	No Action	Action	%
A	Unlikely	80	20	20
A	Likely	10	15	60
B	Unlikely	20	5	20
B	Likely	40	60	60

Data Science from Scratch, 2nd Edition by Joel Grus

Data Ethics

Trading Off Accuracy and Fairness

- Argument 4:
  - 20/125 = 16% of group A were falsely labeled "unlikely," whereas only 5/125 = 4% of group B were falsely labeled "unlikely."
  - This data scientist (who considers an "unlikely" prediction to be a bad thing) insists that the model unfairly stigmatizes group A

Group		Actual		
	Prediction	No Action	Action	%
A	Unlikely	80	20	20
A	Likely	10	15	60
B	Unlikely	20	5	20
B	Likely	40	60	60

Trading Off Accuracy and Fairness

- Which of these data scientists is correct? Are any of them correct? Perhaps it depends on the context.
- Possibly we may feel one way if the two groups are "men" and "women" and another way if the two groups are "R users" and "Python users."
- Or possibly not if it turns out that Python users skew male and R users skew female?
- Possibly we feel one way if the model is for predicting whether a Data Science user will apply for a job through the Data Science job board and another way if the model is predicting whether a user will pass such an interview
- Possibly your opinion depends on the model itself, what features it takes into account, and what data it was trained on
- In any event there can be a tradeoff between "accuracy" and "fairness"(depending on how you define them) and that these tradeoffs don’t always have obvious "right" solutions.

The Five C's

Guidelines for Building Data Products

The Five C's

- What does it take to build a good data product or service?
  - Not just a product or service that’s useful, or one that’s commercially viable, but one that uses data ethically and responsibly
- We often talk about a product’s technology or its user experience, but we rarely talk about how to build a data product in a responsible way that puts the user in the center of the conversation
- Five framing guidelines help us think about building data products
  - Consent
  - Clarity
  - Consistency and Trust
  - Control and Transparency
  - Consequences

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consent

- It is about the agreement between data collectors and data providers
- Establishing trust between the people who provide data and the people who use it requires some kind of agreement:
  - about what data is being collected and how that data will be used
- Agreement starts with obtaining consent to collect and use data
- Unfortunately, these agreements between a service's user (data provider) and the service itself (which uses data in many ways) is binary and lack clarity
  - meaning that you either accept or decline
- In businesses, contract negotiations happen between two parties through multiple iterations before the contract is signed
- But when a user is agreeing to a contract with a data service, they either accept the terms or they don’t get access
  - It’s non-negotiable

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consent - Examples

- Admissions into hospitals require us to sign a form that gives them the right to use our data
  - Generally, there's no way to say that our data can be used for some purposes but not others
- When we sign up for a loyalty card at your local pharmacy or grocery store, we're agreeing that they can use your data in unspecified ways
  - Those ways include targeted advertising (often phrased as "special offers")
  - They may also include selling your data (with or without anonymization) to other parties
- What happens to our data when one company buys another and uses data in ways that we didn't expect?

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consent – Examples

- Data is frequently collected, used, and sold without consent
- This includes organizations like Acxiom, Equifax, Experian, and Transunion, that collect data to assess financial risk
- Many common brands also collect data without consent
  - Google collected data from cameras mounted on cars to develop new mapping products
  - AT&T and Comcast both used cable set top boxes to collect data about their users
  - Samsung collected voice recordings from TVs that respond to voice commands
- At every step of building a data product, it is essential to ask whether appropriate and necessary consent has been provided

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Clarity

- Clarity is closely related to consent
- You can’t really consent to anything unless you’re told clearly what you’re consenting to
- Users must have clarity about what data they are providing, what is going to be done with the data, and any downstream consequences of how their data is used
- All too often, explanations of what data is collected or being sold are buried in lengthy legal documents that are rarely read carefully, if at all.
- Observant readers of Eventbrite’s user agreement recently discovered that listing an event gave the company the right to send a video team, and exclusive copyright to the recordings
- And the only way to opt out was by writing to the company
- The backlash was swift once people realized the potential impact, and Eventbrite removed the language.

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Clarity

- Facebook users who played Cambridge Analytica’s "This Is Your Digital Life" game may have understood that they were giving up their data
  - After all, they were answering questions, and those answers certainly went somewhere
    - Did they understand how that data might be used?
    - Or that they were giving access to their friends’ data behind the scenes?
- That’s buried deep in Facebook’s privacy settings.
- Even when it seems obvious that their data is in a public forum, users frequently don’t understand how that data could be used
- Most Twitter users know that their public tweets are, in fact, public; but many don’t understand that their tweets can be collected and used for research, or even that they are for sale

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Clarity

- Using data in public forums is not unethical
- However, the need isn’t just to get consent, but to inform users what they’re consenting to
  - That’s clarity
- We rarely get a simple explanation of what the service is doing with your data, and what consequences their actions might have
- Unfortunately, the process of consent is often used to obfuscate (obscure) the details and implications of what users may be agreeing to
- And once data has escaped, there is no recourse
  - You can’t take it back
- Even if an organization is willing to delete the data, it’s very difficult to prove that it has been deleted

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consistency & Trust

- Trust requires consistency over time
  - We can’t trust someone who is unpredictable
- People may have the best intentions, but if they don't honor those intentions when needed, we cannot trust them
- Once a trust is broken, restoring trust requires a prolonged period of consistent behavior
- Consistency, and therefore trust, can be broken either explicitly or implicitly
  - E.g., An organization that exposes user data can do so intentionally or unintentionally
- In the past years, we’ve seen many security incidents in which customer data was stolen:
  - E.g., Yahoo!, Target, Anthem, local hospitals, government data, data brokers like Experian, and the list grows longer each day
- Failing to safeguard customer data breaks trust—and safeguarding data means nothing if there is no consistency over time

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consistency & Trust

- We've also seen frustration, anger, and surprise when users don't realize what they've agreed to
- When Cambridge Analytica used Facebook's data to target vulnerable customers with highly specific advertisements, Facebook initially claimed that this was not a data breach
- And while Facebook was technically correct, in that data was not stolen by an intruder, the public's perception was clearly different
- This was a breach of trust, if not a breach of Facebook's perimeter
- Facebook didn't consistently enforce its agreement with its customers
- When the news broke, Facebook became unpredictable because most of its users had no idea what it would or wouldn't do
- They didn't understand their user agreements, they didn't understand their complex privacy settings, and they didn't understand how Facebook would interpret those settings
- https://www.youtube.com/watch?v=VDR8qGmyEQg

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Control and Transparency

- Once you have given your data to a service, you must be able to understand what is happening to your data
- Can you control how the service uses your data?
- For example:
  - Facebook asks for your political views, religious views, and gender preference
  - What happens if you change your mind about the data you've provided?
  - If you decide to rather keep your political affiliation quiet, do you know whether Facebook actually deletes that information?
  - Do you know whether Facebook continues to use that information in ad placement?
- All too often, users have no effective control over how their data is used
- Users are given all-or-nothing choices, or a convoluted set of options that make controlling access confusing
- It's often impossible to reduce the amount of data collected, or to have data deleted later
- However, there is a shift in data privacy rights to give users greater control of their data
- For example:
  - Europe's General Data Protection Regulation (GDPR) requires users' data to be provided to them at their request and removed from the system if they so desire.

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consequences

- Data products are designed to add value for a particular user or system
- As these products increase in sophistication, and have broader societal implications, it is essential to ask whether the data that is being collected could cause harm to an individual or a group
- Due to potential issues around the use of data, laws and policies have been put in place to protect specific groups. For example:
  - The Children's Online Privacy Protection Act (COPPA) protects children and their data
  - The Genetic Information Nondiscrimination Act (GINA) was established in 2008 in response to rising fears that genetic testing could be used against a person or their family
- Unfortunately, policy doesn't keep up with technology advances; neither of these laws have been updated
- Given how rapidly technology is being adopted by society, policies and laws are not updated at same pace

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

Consequences

- Even philanthropic approaches can have unintended and harmful consequences
- When, in 2006, AOL released anonymized search data to researchers, it proved possible to "de-anonymize" the data and identify specific users
- In 2018, Strava opened up their data to allow users to discover new places to run or bike
- Strava didn’t realize that members of the US military were using GPS-enabled wearables, and their activity exposed the locations of bases and patrol routes in Iraq and Afghanistan
- Exposure became apparent after the product was released to the public, and people exploring the data started talking about their concerns
- While Strava and AOL triggered a chain of unforeseen consequences by releasing their data, it’s important to understand that their data had the potential to be dangerous even if it wasn’t released publicly

The Five C's

Consequences

- Collecting data that may seem harmless and combining it with other data sets has real-world implications.
- Combining data sets frequently gives results that are much more powerful and dangerous than anything you might get from either data set on its own
- For example:
  - Data about running routes could be combined with data from smart locks, telling thieves when a house or apartment was unoccupied, and for how long
  - The data could be stolen by an attacker, and the company wouldn’t even recognize the damage.

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

The Five C's

Consequences

- In both Strava's and AOL's cases, well-intentioned data scientists were looking to help others
  - The problem is that they didn’t think through the consequences and the potential risks
- It is possible to provide data for research without unintended side-effects
- For example:
  - The US Internal Revenue Service (IRS), in collaboration with researchers, opened a similar data set in a tightly controlled manner to help understand economic inequality
    - There were no negative repercussions or major policy implications
  - Similarly the Department of Transportation releases data about traffic fatalities
  - The UK Biobank (one of the largest collections of genomic data) has a sophisticated approach to opening up different levels of data
  - Other companies have successfully opened up data for the public benefit, including LinkedIn’s Economic Graph project and Google Books’ ngram viewer
  - Many data sets that could provide tremendous benefits remain locked up on servers. Medical data that is fragmented across multiple institutions limits the pace of research
  - And the data held on traffic from ride-sharing and GPS/mapping companies could transform approaches for traffic safety and congestion
- But opening up that data to researchers requires careful planning.

Ethical Guidelines for a Data Scientist

Data Scientist's Checklist

- ❑ Have we listed how this technology can be attacked or abused?
- ❑ Have we tested our training data to ensure it is fair and representative?
- ❑ Have we studied and understood possible sources of bias in our data?
- ❑ Does our team reflect diversity of opinions, backgrounds, and kinds of thought?
- ❑ What kind of user consent do we need to collect to use the data?
- ❑ Do we have a mechanism for gathering consent from users?
- ❑ Have we explained clearly what users are consenting to?
- ❑ Do we have a mechanism for redress if people are harmed by the results?
- ❑ Can we shut down this software in production if it is behaving badly?
- ❑ Have we tested for fairness with respect to different user groups?
- ❑ Have we tested for disparate error rates among different user groups?
- ❑ Do we test and monitor for model drift to ensure our software remains fair over time?
- ❑ Do we have a plan to protect and secure user data?

Ethics and Data Science by Hilary Mason; Mike Loukides; DJ Patil

Thank You!