UNT

Department

of Computer Science and Engineering

CSCE 5300: Introduction to Big Data and Data Science

Chapter 7 (7.2): Privacy and Ethics

3 November 2024

7.2. Privacy and Ethics

- **Ethics**
- ***What is Ethics?***
- *The term "ethics" comes from the Greek word Ethos, which means "habit" or "custom."*
- *Ethics instructs us on what is good and wrong.*
- Philosophers have pondered this crucial topic for a long time and have a lot to say about it.
- Most people associate ethics with morality: a natural sense of what is "good."
- We as humans live in a society, and society has rules and regulations.

7.2. Privacy and Ethics

- We must be able to decide what is right and what is wrong.
- Ethics deals with feelings, laws, and social norms which determine right from wrong.
- Our ways of life must be reasonable and live up to the standards of society.
- ***Why Ethics in Data Science is important?***
- Today, data science has a significant impact on how businesses are conducted in disciplines as diverse as medical sciences, smart cities, and transportation.
- Whether it's the protection of personally identifiable data, implicit bias in automated decision-making, the illusion of free choice in psychographics, the social impacts of automation, or the apparent divorce of truth and trust in virtual communication, the dangers of data science without ethical considerations are as clear as ever.

7.2. Privacy and Ethics

- The need for a focus on data science ethics extends beyond a balance sheet of these potential problems because data science practices challenge our understanding of what it means to be human.
- ***Algorithms, when implemented correctly, offer enormous potential for good in the world.***
- *When we employ them to perform jobs that previously required a person, the benefits may be enormous: cost savings, scalability, speed, accuracy, and consistency, to name a few.*
- *And because the system is more precise and reliable than a human, the outcomes are more balanced and less prone to social prejudice.*

7.2. Privacy and Ethics

- ***Misconfigured databases, spyware, theft, or publishing on a public forum can all lead to data leaks.***
- *Individuals and organizations must use safe computing practices, conduct frequent system audits, and adopt policies to address computer and data security.*
- ***Companies must take appropriate cybersecurity steps to prevent the leakage of data and information.***
- *This is more important for banks and financial institutions which deal with customers' money.*
- ***Protections must be maintained even when equipment is transferred or disposed of, according to policies.***

7.2. Privacy and Ethics

- ***What is big data ethics?***
  Much research has gone into the field of big data ethics in the past decade as academics and business leaders alike attempt to grapple with public push-back on the use of big data.

- *The field of big data ethics itself is defined as outlining, defending and recommending concepts of right and wrong practice when it comes to the use of data, with particular emphasis on personal data.*
- ***Big data ethics aims to create an ethical and moral code of conduct for data use.***

7.2. Privacy and Ethics

- ***There are five main areas of concern in big data ethics that outline the potential for immoral use of data:***

- ***Informed Consent***
- *To consent means that you give uncoerced permission for something to happen to you.*
- ***Informed Consent*** *is the most careful, respectful and ethical form of consent.*
- *It requires the data collector to make a significant effort to give participants a reasonable and accurate understanding of how their data will be used.*

7.2. Privacy and Ethics

- In the past, informed consent for data collection was typically taken for participation in a single study.

- Big data makes this form of consent impossible as the entire purpose of big data studies, mining and analytics is to reveal patterns and trends between data points that were previously inconceivable.

- In this way, consent cannot possibly be ‘informed’ as neither the data collector nor the study participant can reasonably know or understand what will be garnered from the data or how it will be used.

7.2. Privacy and Ethics

- Revisions to the standard of informed consent have been introduced.
- *The first is known as ‘broad consent’, which pre-authorises secondary uses of data.*
- *The second is ‘tiered consent’, which gives consent to specific secondary uses of data, for example, for cancer research but not for genomic research. Some argue that these newer forms of consent are a watering down of the concept and leave users open to unethical practices.*

7.2. Privacy and Ethics

- Further issues arise when potentially ‘unwilling’ or uninformed data subjects have their information scraped from social media platforms.

- Social media Terms of Service contracts commonly include the right to collection, aggregation and analysis of such data.
- However, Ofcom found that 65% of internet users usually accept terms and conditions without reading them.
- So, it’s not unreasonable to assume that many end-users may not understand the full extent of the data usage, which increasingly extends beyond digital advertising and into social science research.

7.2. Privacy and Ethics

- ***Privacy***
  The ethics of privacy involve many different concepts such as liberty, autonomy, security, and in a more modern sense, data protection and data explosure.
- We can understand the concept of big data privacy by breaking it down into three categories:
  - ***The condition of privacy***
  - ***The right to privacy***
  - ***The loss of privacy and invasion***
- The scale and velocity of big data pose a serious concern as many traditional privacy processes cannot protect sensitive data, which has led to an exponential increase in cybercrime and data leaks.

7.2. Privacy and Ethics

- One example of a significant data leak that caused a loss of privacy to over 200 million internet users happened in January 2021.
- A rising Chinese social media site called Sociallarks suffered a breach due to a series of data protection errors that included an unsecured ElasticSearch database.
- **A hacker was able to access and scrape the database which stored:**
- ***Names***
- ***Phone numbers***
- ***Email addresses***
- ***Profile descriptions***
- ***Connected social media account login names***

7.2. Privacy and Ethics

- ***Follower and engagement data***
- ***Locations***
- ***LinkedIn profile links***
- A further concern is the growing analytical power of big data, i.e. how this can impact privacy when personal information from various digital platforms can be mined to create a full picture of a person without their explicit consent.
- For example, if someone applies for a job, information can be gained about them via their digital data footprint to identify political leanings, sexual orientation, social life, etc.
- All of this data could be used as a reason to reject an employment application even though the information was not offered up for judgement by the applicant.

7.2. Privacy and Ethics

- ***Data ethics is a new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).”***

- ***Simply speaking, in data ethics, we learn about all the ethical problems that appear during our use of data.***

- In this era of rapid technological development, we are living in a “Data-fied World.”

7.2. Privacy and Ethics

- Data collection is a vital part of nearly every aspect of our lives, from the phones in our pockets to the cars we drive.
- Almost every human behavior and every operation we do with a tool like a computer can be collected as data.
- Over the years, as technology progressed and we aimed for a better life, we began to use data generated from day-to-day actions to conduct complex analysis with the help of strengthened computing powers and new analytical tools.
- Advanced technologies related to data science, like Machine Learning and AI, have brought a lot of benefits to our life.
- However, as humans begin to step away from hands-on analysis and let automated machines do most of the work for us, different issues such as fairness, privacy, and

7.2. Privacy and Ethics

- representation emerge.
- We will cover a couple of cases about those issues in detail below, so keep reading!

- Ever since Data Science became a buzzword in the technological industry, colleges and universities have been scrambling to open a Data Science Program to satisfy the world's growing demand for data scientists, engineers, and analysts.
- As Data Scientists, we often deal with big sets of data that are driven by people, so it is our duty to keep private data secured and use it responsibly.
- To better incorporate human values like justice and equity in data-driven technologies, we need to also understand the underlying human and social structures.

7.2. Privacy and Ethics

- **How should we incorporate Data Ethics in our work as students?**
- When we are doing a data science project, we need to make sure that we understand the potential ethical consequences of our work.
- Some tips for you to be an ethical data scientist are:
- ***first, be aware of privacy issues such as data breaches and find ways to adequately secure the data.***
- If you are not familiar with the danger of a data breach, check out this news article about the *Facebook Security Breach*.
- ***Second, be transparent with your data usage.***
- ***Get user consent before you use their data in any way.***

7.2. Privacy and Ethics

- ***Third, despite the difficulty of being completely objective, you should try your best to make sure there is no bias involved in your model.***
- In fact, to make employees follow data ethics principles, many companies and organizations have incorporated certain codes of ethics and conduct.
- It addresses common ethical dilemmas that data scientists from the industry, academia, and the public sector may face.
- Feel free to take a look at it.

7.2. Privacy and Ethics

- **Here's the Data Ethics Checklist:**
- Big data analytics raises several ethical issues, especially as companies begin monetizing their data externally for purposes different from those for which the data was initially collected.
- The scale and ease with which analytics can be conducted today completely change the ethical framework.
- We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do.

- While there is still no black or white, experts agree on a few principles:

7.2. Privacy and Ethics

- ***Private customer data and identity should remain private:*** *Privacy does not mean secrecy, as personal data might need to be audited based on legal requirements, but that private data obtained from a person with their consent should not be exposed for use by other businesses or individuals with any traces to their identity*.

- ***Shared private information should be treated confidentially:*** *Third-party companies share sensitive data — medical, financial or locational — and need restrictions on whether and how that information can be shared further.*

7.2. Privacy and Ethics

- ***Customers should have a transparent view*** *of how our data is being used or sold and the ability to manage the flow of their private information across massive, third-party analytical systems.*

- ***Big Data should not interfere with human will:*** *Big data analytics can moderate and even determine who we are before we make up our minds.*
- *Companies need to consider the kind of predictions and inferences that should be allowed and those that should not*.

7.2. Privacy and Ethics

- ***Big data should not institutionalize unfair biases*** *like racism or sexism.*
- *Machine learning algorithms can absorb unconscious biases in a population and amplify them via training samples.*

- There are certainly more principles we need to develop as more powerful technology becomes available.
- Data scientists, data engineers, database administrators, and anyone involved in handling big data should have a voice in the ethical discussion about how data is used.
- Companies should openly discuss these dilemmas in formal and informal forums. When people do not see ethics playing in their organization, they go away in the long run.

7.2. Privacy and Ethics

- To protect individuals' privacy, ensure you're storing data in a secure database so it doesn't end up in the wrong hands.
- Data security methods that help protect privacy include dual-authentication password protection and file encryption.
- For professionals who regularly handle and analyze sensitive data, mistakes can still be made.
- One way to prevent slip-ups is by de-identifying a dataset.
- A dataset is de-identified when all pieces of private information are removed, leaving only anonymous data.
- his enables analysts to find relationships between variables of interest without attaching specific data points to individual identities.

7.2. Privacy and Ethics

- **Ethical use of algorithms**
- If your role includes writing, training, or handling machine-learning algorithms, consider how they could potentially violate any of the five key data ethics principles.

- Because algorithms are written by humans, bias may be intentionally or unintentionally present.

- Biased algorithms can cause serious harm to people.
- In Data Science Principles, the following ways bias can creep into your algorithms:

7.2. Privacy and Ethics

- ***Training:*** *Because machine-learning algorithms learn based on the data they're trained with, an unrepresentative dataset can cause your algorithm to favor some outcomes over others.*
- ***Code:*** *Although any bias present in your algorithm is hopefully unintentional, don't rule out the possibility that it was written specifically to produce biased results.*
- ***Feedback:*** *Algorithms also learn from users' feedback.*
- As such, they can be influenced by biased feedback.
- For instance, a job search platform may use an algorithm to recommend roles to candidates.
- If hiring managers consistently select white male candidates for specific roles, the algorithm will learn and adjust and only provide job listings to white male candidates in the future.

7.2. Privacy and Ethics

- The algorithm learns that when it provides the listing to people with certain attributes, it's "correct" more often, which leads to an increase in that behavior.

- ***"No algorithm or team is perfect, but it's important to strive for the best," Tingley says in Data Science Principles.***

- "Using human evaluators at every step of the data science process, making sure training data is truly representative of the populations who will be affected by the algorithm, and engaging stakeholders and other data scientists with diverse backgrounds can help make better algorithms for a brighter future."

7.2. Privacy and Ethics

- **Using data for good**
- ***While the ethical use of data is an everyday effort, knowing that your data subjects' safety and rights are intact is worth the work.***

- When handled ethically, data can enable you to make decisions and drive meaningful change at your organization and in the world.

7.2. Privacy and Ethics

- **Ethics in Data Science and Proper Privacy and Usage of Data**
- Data may be utilized to make decisions and have a large influence on businesses.
- However, this valuable resource is not without its drawbacks.
- How can businesses acquire, keep, and use data in an ethical manner?
- What are the rights that must be protected? Some ethical practices must be followed by data-handling business personnel.
- ***Data is someone's personal information and there must be a proper way to use the data and maintain privacy.***

7.2. Privacy and Ethics

- **Some Ethical Practices**

- ***Making Decisions:***
- ***Data scientists should never make judgments without contacting a client, even if the decision is for the interest of the project.***
- The aims and objectives of projects must be understood by both data scientists and clients.
- ***Let's say a data scientist wishes to take action on behalf of a customer on a certain ongoing project.***
- ***Even if the action is advantageous to the client and the project, it must be explained to the client, and no choice should be made on their behalf.***

7.2. Privacy and Ethics

- Data scientists should only make decisions when it is expressly stated in the contract or when their authority allows them to.

- ***Privacy and Confidentiality of Data:***
- Data scientists are continually involved in producing, developing, and receiving information.
- Data concerning client affiliates, customers, workers, or other parties with whom the clients have a confidentiality agreement is often included in this category.
- ***Then, regardless of the sort of sensitive information, it is the data scientist's responsibility to protect it.***

7.2. Privacy and Ethics

- Only when the customer provides permission for data scientists to share or talk about this type of information should it be disclosed or spoken about.
- ***Complete privacy of clients' or customers' data must be maintained.***
- Even if a consumer consents to your organization collecting, storing, and analyzing their personally identifiable information, that doesn't mean they want it made public.

- To preserve people's privacy, make sure you're keeping the information in a secure database so it doesn't get into the wrong hands.
- Dual-authentication password protection and file encryption are two data security solutions that assist safeguard privacy.

7.2. Privacy and Ethics

- ***Data Ownership:***
- One of the important concepts of ethics in Data Science is that the individual has data ownership.
- ***Collecting someone's personal data without their agreement is illegal and immoral.***
- As a result, consent is required to acquire someone's data.
- Signed written agreements, digital privacy policies that require users to accept a company's terms and conditions, and pop-ups with checkboxes that allow websites to track users' online behavior using cookies are all typical approaches to get consent.
- ***To prevent ethical and legal issues, never assume a consumer agrees to you gathering their data; always ask for permission.***

7.2. Privacy and Ethics

- ***Good intentions with Data:***
- ***Intentions of data collection and analyzing data must be good.***

- Data professionals must be clear about how and why they use the data.

- If a team is collecting data regarding users' spending habits, to make an app to manage expenses, then the intention is good.

7.2. Privacy and Ethics

- ***Transparency:***
- Data subjects have a right to know how you plan to acquire, keep, and utilize their personal information, in addition to owning it.
- ***Transparency should be used when acquiring data.***
- ***You should create a policy that explains how cookies are used to track user's activity and how the information gathered is kept in a secure database, as well as train an algorithm that gives a tailored online experience.***

- It is a user's right to have access to this information so that they may choose whether or not to accept your site's cookies.

7.2. Privacy and Ethics

- ***We must consider whether we should to process big data.***

- ***If the effects of big data are more likely to be negative than positive, then it would be the moral responsibility of workers in the field to redirect their research.***

- *Many new technologies have had unintended negative side effects: nuclear fission brought Chernobyl and the threat of global destruction; the internal combustion engine brought air pollution, global warming, and the paving-over of paradise.*