Data Ethics

An Expanding the Data Analytics Technician Pipeline Module

Data Ethics

While data-driven decision making is powerful and impactful for businesses, it also exposes ethical and privacy risks that need to be addressed. Strong individual and business ethics are necessary for collecting, storing, and using data. Multiple levels of protections must be in place to prevent misuse of

- personally identifiable information (PII)
- personal sensitive information (PSI)
- personal health information (PHI)

To have data ethics is to examine the moral implications and obligations to data gathered, and to determine if collecting and examining this data is the right thing to do. Also, consider if the data collected and examined is necessary, or if there is possibly a better way to collect, or protect, the data.

Vocabulary

- personal identity information (PII) - any information that permits the identity of an individual of an individual to be directly, or indirectly inferred. (dhs.gov) It can be sensitive or non-sensitive.
- personal health information (PHI) - health care, health status, or payments for health care that is created or collected, and can be linked to a specific individual
- personal sensitive information (PSI) - Personally identifiable information (PII) not available through public records.
- non-sensitive personal information - personally identifiable information that is available through public record, such as directory information
- ownership - who owns the values in the data
- transparency - showing what actions are performed on the data. Not hiding any data or transformations on the data. Explaining why actions were performed on the data
- privacy - being free from observation, disturbance or public attention
- intention - the goal, plan, or reason why data was collected, transformed, or analyzed
- outcomes - the way things turn out
- Algorithmic bias - intentionally or unintendedly changing the outcome of computer calculations due to racial, ethnic, religious, political, or sexual orientation bias. One of the most famous is the bias in the most widely used facial recognition software
- Disparate impact - inadvertent harm to a person or group of persons. Unlawful under the Civil Rights Act

Personally Identifiable Information (PII)

There are two forms of PII - sensitive information and non-sensitive information. Non-sensitive information is information from openly and freely available public sources. One example of non-sensitive would be directory information (address, telephone number). Another example would be public access information for court cases. Because court cases are public, information regarding court cases is also publically available. This is why you can view arrest records of individuals, birth and death certificates, and access clerk of courts data.

Sensitive data is data that leads to identity theft. It can include protected health information, education records, confidential personnel information, genetic or biometric data, race, ethnicity, religion, political persuasion, sexual orientation, or union membership status. When collecting, storing, transmitting, or analyzing sensitive information, care must be taken to ensure that it is not revealed verbally, or visually, to someone who is not authorized to have access to it.

It is important to note that non-sensitive data can lead to sensitive information. Like a trail of breadcrumbs, combining multiple bits of non-sensitive information can provide clues or reveal intention, or potential operations that should be considered sensitive, or classified. Learn more at the Homeland Security's web page.

PHI

Personal Health Information is protected under the Health information Patient Privacy Act (HIPPA). This includes any medical, dental, behavioral, emotional, or financial information. This applies to the data, reports and audiovisuals that contain the data, as well as transmission of data. Revealing HIPPA information can result in both fines and jail time.

Principles of Data Ethics

Obtained from Harvard Business School Blog. (https://online.hbs.edu/blog/post/data-ethics)

1. Ownership: An individual owns their own personal information. It is unethical and unlawful to collect personal data without an individual's consent.
2. Transparency: individuals have a right to know how their personal information is being collected, stored, and used. Deception, or withholding this information is unlawful.
3. An individual's privacy is to be protected at all times.
4. Intention: be able to explain why you're collecting this data, what will be gained from it, and what changes you will make to it, and how it will be conveyed. It is not ethical to obtain and use data for illegal, deceptive, or malicious means.
5. Data analysis outcomes should not be revealed if disparate impact occurs, which is why intention is very important. Disparate impact is inadvertent harm to a person or group of persons, and is unlawful under the Civil rights Act

Activity 1

In this exercise, we are going to be researching the public records of any J. Smith in Montgomery County, Ohio.

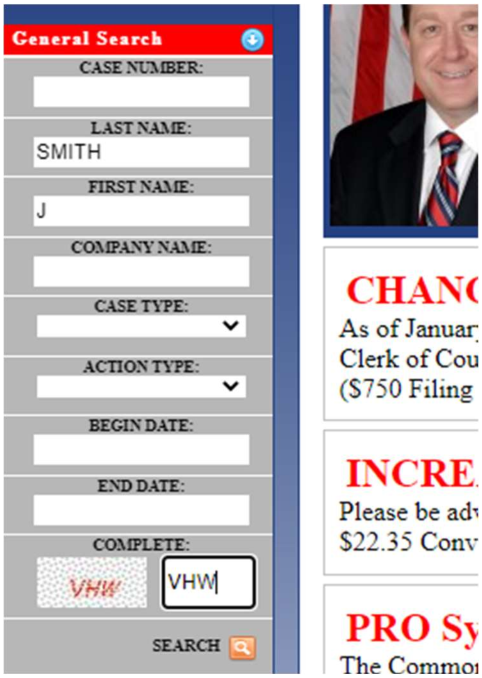
Resources

Montgomery County Clerk of Courts Records (https://pro.mcohio.org/pro)

Montgomery County Auditor (https://www.mcrealestate.org/search/commonsearch.aspx?mode=owner)

Steps

1. Visit the Montgomery County Clerk of Courts Records site and complete the form at this site as shown below. Do not forget the very small captcha (in red) at the bottom. In this example, the captcha is VHW. Browse what you see. What can you learn from a Clerk of Courts records?



General Search

CASE NUMBER:

LAST NAME: SMITH

FIRST NAME: J

COMPANY NAME:

CASE TYPE:

ACTION TYPE:

BEGIN DATE:

END DATE:

COMPLETE: VHW VHW

SEARCH

2. Visit the Montgomery County Auditor site, and type in the last name "Smith". Select any property. What do you learn from this site? Be sure to explore Deed information to discover additional document search of public records.
3. Be prepared to discuss with the class:
    a. How might an individual be affected by disparate impact should data from these sites be revealed?
    b. What are some ways the data collected from this site could be misused?

c. How might data from the site be used to perform scams, frauds, or blackmail?

Activity 2

Resources

ACM Magazine: (https://cacm.acm.org/magazines/2021/3/250698-can-the-biases-in-facial-recognition-be-fixed-also-should-they/fulltext)

Steps:

1. With a partner, summarize the most important points in the article. Discuss and come to a consensus on whether you agree or disagree. Use the further readings, and any supporting videos to support your reasoning. Submit your points, and your conclusions, along with citations, to your instructor.

Activity 3:

Resources:

Digital Footprints, Breadcrumbs

Military movements revealed by smart phone data

Steps:

1. After reading the articles, brainstorm with the class (or in small groups) additional situations in which breadcrumb data could lead to revealing PII, as well as preventative measures to protect the data. Example: How might vulnerable groups such as teens, LGBTQ+, or women be targeted using breadcrumbs?

Activity 4:

Steps

1. With a partner, research 4 internet sources less than 3 years old that identify real life examples of algorithmic bias, bias in digital marketing and marketing and political ads, and in facial recognition software, which is used globally.
2. Create a presentation to inform the “average” person how they may be unconsciously influenced by computer algorithms and deep fakes. Use MLA Citation.

Activity Extension:

Create a humorous, obviously false deep fake video to share with the class using https://deepfakesweb.com/

This material is based upon work supported by the National Science Foundation under Grant DUE 2055411.Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.