

DS 300 Midterm 1 Study Guide

Josepheh Sepich Sept 27

1 Privacy Overview

Data is generated everywhere.

- Every click on-line
- user-generated
- Internet of things
- Health and Scientific computing

A 2013 study researchers were able to derive all the demographics of a person from a simple image on their social media account.

In the adversarial model sometimes the data controllers “dehumanize the enemy by calling them data”.

- To what extent can someone use data beyond the person’s intentions?
- Who owns data?
- How do we value different aspects of privacy?
- How do users get informed consent?
- What does it mean to be fair?

2 Data Science and Data

Predpol is a predictive policing service. The company adaptive an earthquake prediction algorithm to help predict crime hot spots that need police attention. In LA there’s been a 33% reduction in burglaries and 21% reduction in violent crimes in areas where the software is being used. Some details that are privacy concerning are:

- area specific data from crime incidents is used
- sensitive data could be derived from non-sensitive data

Macy and other companies have also developed price changing algorithms. These algorithms essentially change prices for products in real time adapting to demand and inventory.

Yet another application of data science is Google’s sponsored search. Google makes a revenue of \$50 billion a year and 97% of that revenue comes from marketing. The sponsored search service uses an auction. This creates pure competition where companies try to win access to customers. Using data given by Google companies creating auction algorithms to have the right model in order to reach the right audience. Google has 30 billion requests a month.

Other applications of Data Science:

- Transaction Databases: Recommendation system, Fraud detection
- Wireless Sensor Data: Internet of Things
- Text and Social Media Data
- Software Log Data: Automated Trouble Shooting

- Genotype Data: 23andMe, personalized medicine

Five V's of Big Data

1. Volume: Raw Data
2. Velocity: Change over time
3. Variety: Data types
4. Veracity: Data quality
5. Value: Information for decision making

It is hard to balance utility of data and privacy

Privacy: State of being let alone and able to keep certain matters to one's self.

Types of Privacy:

- Personal Privacy
- Information Privacy
- Legal Privacy

In a Ted Talk by a researcher on privacy, there was an experiment performed at CMU. In this experiment the researchers took a picture of someone and had them fill out a survey. While they filled out a survey they scraped the web with the picture. They were able to identify 1 in 3 people and could find SSN based on a picture and social media profiles.

Case Study: OkCupid Data

- Who owns the data?
- How do we value different aspects of privacy?
- How do we get informed consent?
- What does it mean to be fair?

Consider:

- Consent
- Transparency
- Terms of Service
- Adversarial Model
- Limits on inference: Where do we draw the line?

3 Overview of Data Privacy Issues

Privacy is a hard to define concept. Some examples of what it could be:

- Personhood
- Intimacy
- Secrecy
- Right to be let alone
- limited access to the self
- control over information

In this class we are focusing on privacy and not security, but security is a part of keeping data private. Some inherently identifying information cannot be anonymized:

- genome sequences
- ancestry.com

Desires about data involve questions of trust. People must be willing to share information such as medical records, but are more likely to share with researchers and not pharama companies. Privacy is **not a binary value**.

Limited Acces:

- Laws to prohibit or limit collection, disclosure, contact
- tech to facilitate anonymous transactions

Control:

- Laws to mandate choice
- facilitate informed consent
- enforce privacy preferences

There are four facets of privacy:

1. Solitude: Separate from group, free from observation
2. Intimacy: put in a small unit
3. Anonymity: public but finds freedom from id/surveillance

Information privacy deals with collection of data and how it is stored. Decisional privacy deals with how end users decide to release or keep data.

Privacy and stakeholders:

- Company
- Customer
- Government
- data analyst
- data snooper/adversary

Confidential data definitions:

- Easily connected to a person providing it
- Could be agree to be kept confidential
 - business income
 - health/medical details
- conditioned by a number of factors
 - ethical guidelines
 - legal requirements
 - research specific consent agreement

3.1 Data Analysis Workflow

1. Collection of Data
2. Filtering and Processing of Data
3. Analysis of Data
4. Publications and usage of data

3.2 Privacy Violations

1. Information leakage (cookies on the web)
2. Inference (predictive analysis)
3. Information/data used in unwanted way (recommendaions too personal)

Privacy Enhancing Technology

- VPN
- DuckDuckGo
- Wickr

Privacy Invasive Technology

- Amazon Echo
- Facial Recognition
- Location-based technology
- Venmo

4 Economics of Privacy