COGS 9 – Discussion Section A01

Kunal Rustagi (TA): Wed 9AM (Zoom)
Jiesen Zhang (IA): Tues 4PM (Zoom Pass: 804882)
Bobby Zhu (IA): Tues 12:30PM (CSB 114 and Zoom Pass: 110985)
Boyong Liu (IA): Wed 2PM (Zoom)

Discussion Sections Schedule

Week 3: Introductions + Course Logistics + Making teams + Reading 1 + Python Basics (time permitted)

Week 4: Reading 2 + Python basics + Getting data and wrangling it using Pandas

Week 5: Reading 3 + Assignment 1 + Basics of programming for data science

Week 6: Final project - I discussions

Week 7: Reading 4 + Assignment 2 + Data Visualizations and Data exploration

Week 8: Assignment 3 + Machine Learning demo

Week 9: Reading 5 + Closing thoughts

Week 10: Final project - II discussions

Deadlines

- Quiz 2, due Friday, 3rd February

Data’s day of reckoning

By Mike Loukides, Hilary Mason, and DJ Patil

Data: Benefit or Harm?

- Recommender systems
- Autonomous vehicles
- Maps and Navigation

However, when it is not properly used…

- Facebooks' violation of individual privacy
- The case of Henrieta Lacks

Ethics and security training

- Integrate ethics ideas into school assignment
- Add ethics course to core curriculum
- integrated into every course at colleges, universities, online courses, and programming boot camps.

Developing guiding principles

- Tight schedule
- Checklists are the solution

❑ 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 and use the data?

❑ Do we have a mechanism for gathering consent from users?

Building ethics into a data-driven culture

- An individual needs to be empowered to stop the process before damage is done.
- Anyone should be able to escalate issues for remediation without fear of retaliation
- An ethical challenge should be part of the hiring process
- Product reviews must ask questions about the product’s impact
- Teams must reflect diversity of thought, experiences, race, and background
- Corporations must make their own principles clear

Regulation

- The Common Rule
- U.S. Federal Trade Commission (FTC)
- Nuclear Regulatory Commission (NRC)
- Consumer Finance Protection Bureau (CFPB)


Ethics, Privacy, Security

What is ‘PII’

“any information relating to an [...] natural person [...] who can be identified, directly or indirectly, in particular by reference [...] to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity.”

-Data Protection Directive

PII and Privacy Protection Technologies

Examples: *k*-anonymity, *l*-diversity

“these methods aim to make joins with external datasets harder by anonymizing the identifying attributes.”

The methods modifies quasi-identifiers to satisfy various syntactic properties to prevent

Problem: Simply not enough to do the job😅

Re-identification without PII

It turns out there's a wide spectrum of human characteristics that enable re-identification as long as they satisfy the following key properties:

1. They are reasonably stable across time and contexts
2. The corresponding data attributes are sufficiently numerous and fine-grained that no two people are similar, except with a small probability.

Re-identification algorithms take advantage of such properties to re-identify individuals using other attributes.

Differential Privacy

Differential privacy (DP) is a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset.

Python basics
+
Data wrangling with Pandas