Privacy

- History Lesson Definitions
- Story
- Differential Privacy
- Ethics + Empathy

Example: Health Records

Hospital, Zip code, Cancer/No cancer

STORY

In 2000, Mass

- Released and records of all
  state employees
- Wiped IDs
  Repl: Zip code, birthday, gender

In Mass: buy voter data
- Names, birthday, zip code,
- Good student: Latanya Sweeney
• R-anonymity: public data → narrow down to at least \(k\) people

• \(l\)-diversity: (R-anonymity) + at least \(l\) distinct private traits

• \(t\)-closeness: \((l\)-diversity\) + distribution of private traits is t-close (in EMD)

Netflix Challenge 2006

Public: 
\[D_1 = \{ \text{userid, movie, date, grade}, \text{ (public data) } \} \]

Private: 
\[D_2 = \{ \text{userid, movie, date, grade}, \} \]

Evaluate: 
\[D_3 = \{ \text{iid, m, date, grade} \} \]

IMDB user-id, rating, time stamp, movie
Differential Privacy

Two data sets $D_1, D_2$:
- $D_1$, $D_2$ are similar statistical analysis
- change 1 data point $D_1 \rightarrow D_2$

$$D_1, D_2 \leq \{0, 1\}^n$$
- $\text{Ham}(D_1, D_2) = 1$ (change 1 bit)
- $\frac{\Pr [g(D_1) \in R]}{\Pr [g(D_2) \in R]} \leq \exp(\varepsilon) = 1 + \varepsilon$

$$D_1 = 10110011 \quad D_2 \oplus g \rightarrow 110000001 = 2$$

$$R = [0, 3]$$

1. Interactive Approach
   - $D_2 \leftarrow (D_1 + \text{noise})$
   - Is private
   - But asks questions.

2. Non-Interactive Approach
   - Release $D_2$

   $$D_1 \leftarrow \text{True height of Sly Stallone: 68}$$
   $$D_2 = D_1 + \text{Laplace Noise}$$

   $$\Pr [D_1 \geq 70] = \exp(-2\varepsilon)$$
   $$\Pr [D_2 \geq 70] = \exp(-3\varepsilon)$$

   $$\frac{\Pr [D_2 \geq 70]}{\Pr [D_1 \geq 70]} = \frac{\exp(-2\varepsilon)}{\exp(-3\varepsilon)} = \exp(\varepsilon) = 1 + \varepsilon$$