CS480/680: Introduction to Machine Learning

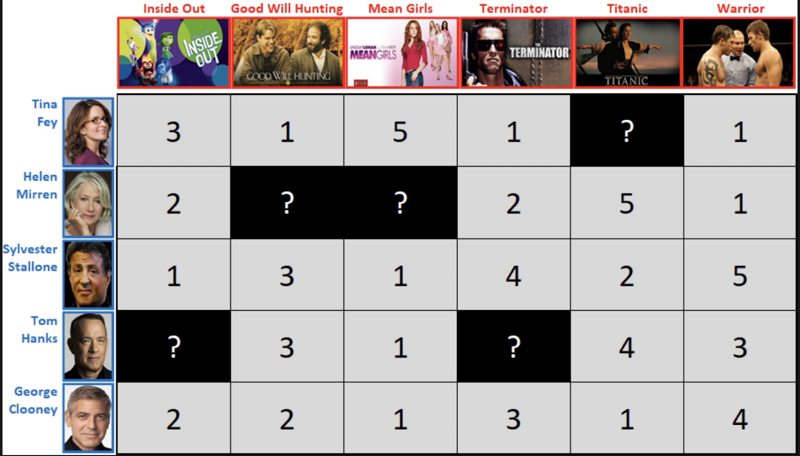
Lec 22: Differential Privacy

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE

July 24, 2024

The Netflix Challenge



	Inside Out	Good Will Hunting	Mean Girls	Terminator	Titanic	Warrior
Tina Fey	3	1	5	1	?	1
Helen Mirren	2	?	?	2	5	1
Sylvester Stallone	1	3	1	4	2	5
Tom Hanks	?	3	1	?	4	3
George Clooney	2	2	1	3	1	4

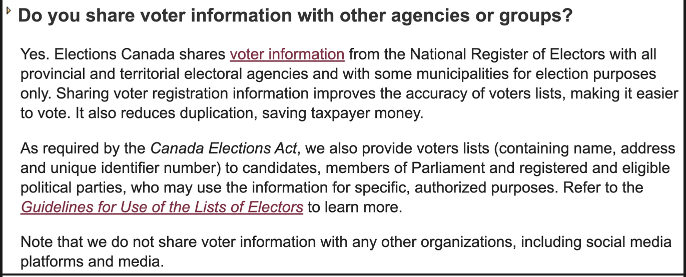
- <user, movie, date of rating, rating>
- ~1M ratings, .5M users, 20k movies

1M Prize

NETFLIX
2009
DATE 09 21 09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION 00/100
FOR The Netflix Prize
Reed Hastings

Lawsuit

Linkage Attack

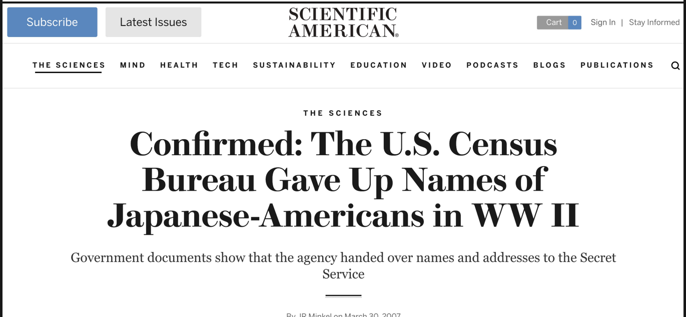


Do you share voter information with other agencies or groups?

Yes. Elections Canada shares voter information from the National Register of Electors with all provincial and territorial electoral agencies and with some municipalities for election purposes only. Sharing voter registration information improves the accuracy of voters lists, making it easier to vote. It also reduces duplication, saving taxpayer money.

As required by the *Canada Elections Act*, we also provide voters lists (containing name, address and unique identifier number) to candidates, members of Parliament and registered and eligible political parties, who may use the information for specific, authorized purposes. Refer to the *Guidelines for Use of the Lists of Electors* to learn more.

Note that we do not share voter information with any other organizations, including social media platforms and media.



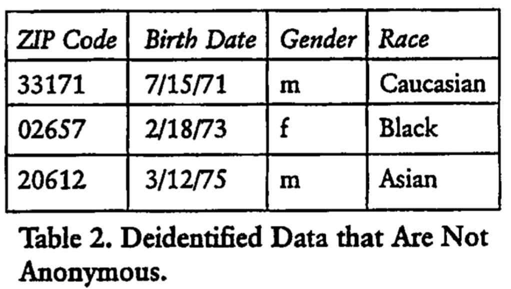
THE SCIENCES

Confirmed: The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II

Government documents show that the agency handed over names and addresses to the Secret Service

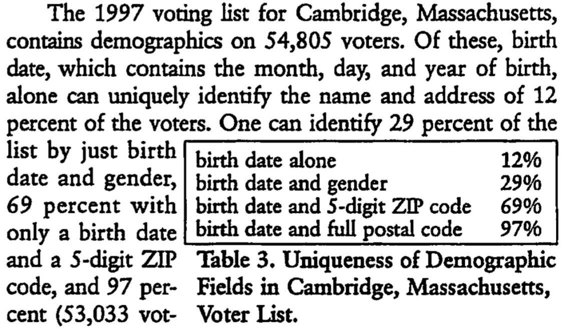
By JR Minkel on March 30, 2007

Anonymization is not Enough



ZIP Code	*Birth Date*	*Gender*	*Race*
33171	7/15/71	m	Caucasian
02657	2/18/73	f	Black
20612	3/12/75	m	Asian

Table 2. Deidentified Data that Are Not Anonymous.

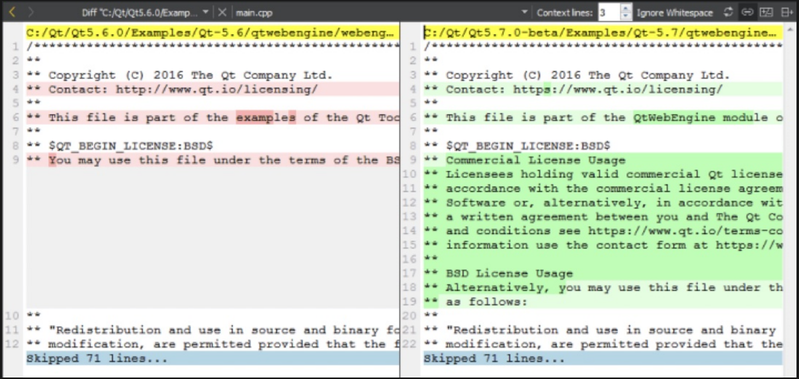


The 1997 voting list for Cambridge, Massachusetts, contains demographics on 54,805 voters. Of these, birth date, which contains the month, day, and year of birth, alone can uniquely identify the name and address of 12 percent of the voters. One can identify 29 percent of the list by just birth date and gender, 69 percent with only a birth date and a 5-digit ZIP code, and 97 percent (53,033 vot-

birth date alone	12%
birth date and gender	29%
birth date and 5-digit ZIP code	69%
birth date and full postal code	97%

Table 3. Uniqueness of Demographic Fields in Cambridge, Massachusetts, Voter List.

Differencing Attack

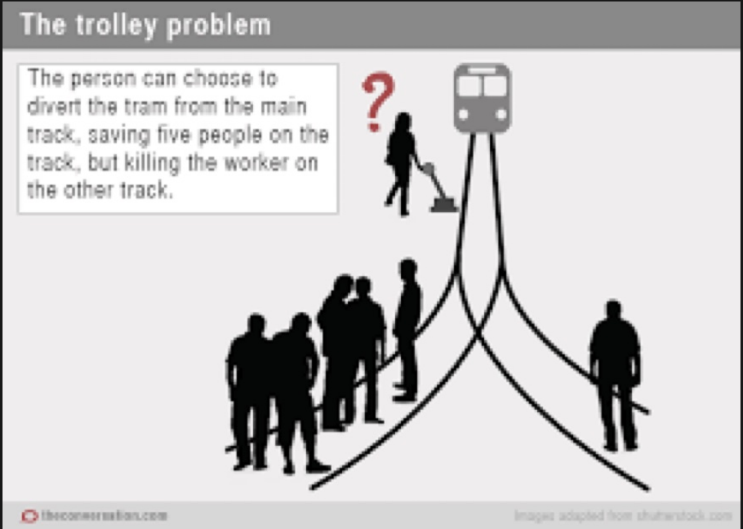


```
C:/Qt/Qt5.6.0/Examples/Qt-5.6/qtwebengine/webeng…
/*****************************************************
**
** Copyright (C) 2016 The Qt Company Ltd.
** Contact: http://www.qt.io/licensing/
**
** This file is part of the examples of the Qt Toc
**
** $QT_BEGIN_LICENSE:BSD$
** You may use this file under the terms of the BS
**
** "Redistribution and use in source and binary fc
** modification, are permitted provided that the f
Skipped 71 lines...
```

```
C:/Qt/Qt5.7.0-beta/Examples/Qt-5.7/qtwebengine…
/*****************************************************
**
** Copyright (C) 2016 The Qt Company Ltd.
** Contact: https://www.qt.io/licensing/
**
** This file is part of the QtWebEngine module o
**
** $QT_BEGIN_LICENSE:BSD$
** Commercial License Usage
** Licensees holding valid commercial Qt license
** accordance with the commercial license agreem
** Software or, alternatively, in accordance wit
** a written agreement between you and The Qt Co
** and conditions see https://www.qt.io/terms-co
** information use the contact form at https://w
**
** BSD License Usage
** Alternatively, you may use this file under th
** as follows:
**
** "Redistribution and use in source and binary
** modification, are permitted provided that the
Skipped 71 lines...
```

- “How many people have disease X?”
- “How many people, not named YYL, have disease X?”

Just Sacrifice A Few?



The trolley problem

The person can choose to divert the tram from the main track, saving five people on the track, but killing the worker on the other track.

theconversation.com

Images adapted from shutterstock.com

Restricted Access

Example

- Consider a medical study about smoking and cancer
- Should a smoker participate?
- If yes, may lead to higher insurance premium
- But may also benefit from learning health risks
- Has the smoker's privacy been compromised?

Participate or not, impact on the smoker is likely the same

Have you cheated in any exam?

Randomized Response

- Want to estimate the percentage of cheaters
- If ask bluntly, almost certainly will under-estimate
- Toss a coin: head, answer honestly; tail, answer randomly
  - cheaters: w.p. $\frac{3}{4}$ say yes
  - non-cheaters: w.p. $\frac{1}{4}$ say yes
  - $\frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{1}{2}p =$ percentage of yes
- Plausible deniability for everyone
- What happens if we ask this question repeatedly?

S. L. Warner. "Randomised response: a survey technique for eliminating evasive answer bias". *Journal of the American Statistical Association*, vol. 60, no. 309 (1965), pp. 63–69.

Differential Privacy

- Let M : $\mathcal{D} \to \mathcal{Z}$ be a randomized mechanism
- (ϵ, δ)-DP if for any $D, D' \in \mathcal{D}$ differing by one data point, for any event $E \subseteq \mathcal{Z}$,

$$\Pr[\mathsf{M}(D) \in E] \leq \exp(\epsilon) \cdot \Pr[\mathsf{M}(D') \in E] + \delta$$

  - dataset D, D' fixed; randomness from the mechanism
  - the smaller ϵ or δ is, the stricter the privacy requirement
- $(\epsilon, 0)$-DP if $\delta = 0$, a.k.a. ϵ-DP
- ϵ (roughly) bounds log odds ratio: $\boxed{\epsilon \leq 1}$ often considered "good"
- δ allows rare, possibly catastrophic event (to trade utility): often, $\boxed{\delta \ll 1/|\mathcal{D}|}$

C. Dwork and A. Roth. "The algorithmic foundations of differential privacy". *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4 (2014), pp. 211–407.

Randomized Response is $(\log 3, 0)$-DP

$$\log \frac{\Pr[\mathsf{M}(D) \in E]}{\Pr[\mathsf{M}(D') \in E]} = \log \frac{\int_E p(\mathbf{x})\,\mathrm{d}\mathbf{x}}{\int_E q(\mathbf{y})\,\mathrm{d}\mathbf{y}} = \log \int_E \frac{p(\mathbf{x})}{q(\mathbf{x})} \cdot \frac{q(\mathbf{x})}{\int_E q(\mathbf{y})\,\mathrm{d}\mathbf{y}}\,\mathrm{d}\mathbf{x}$$

$$(\text{Jensen's inequality}) \leq \int_E \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \cdot \frac{q(\mathbf{x})}{\int_E q(\mathbf{y})\,\mathrm{d}\mathbf{y}}\,\mathrm{d}\mathbf{x}$$

$$(\text{mean} \leq \text{max}) \leq \max_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \leq \epsilon$$

- Consider when D has a cheater and D' has a non-cheater:
  - $\log \frac{\Pr[\mathsf{M}(D)=\text{Yes}]}{\Pr[\mathsf{M}(D')=\text{Yes}]} = \log \frac{3/4}{1/4} = \log 3$
  - $\log \frac{\Pr[\mathsf{M}(D)=\text{No}]}{\Pr[\mathsf{M}(D')=\text{No}]} = \log \frac{1/4}{3/4} = -\log 3$

DP in Practice

- Apple: reportedly $\epsilon = 6$ in MacOS, $\epsilon = 14$ in iOS10 and $\epsilon = 2$ for health types
- Facebook: e.g., $\epsilon = 1.453$ and $\delta = 1e - 5$
- Google: e.g., ϵ up to 9
- LinkedIn: each query uses $\epsilon = 0.15$ and $\delta = 1e - 10$
- Microsoft: e.g., $\epsilon = 12$ and $\delta = 5.8e - 6$
- US Census Bureau: e.g., $\epsilon = 13.64$ and $\delta = 1e - 5$

https://desfontain.es/blog/real-world-differential-privacy.html

A Hypothesis Testing View

- Consider null hypothesis $H_0 : D$ and alternative hypothesis $H_1 : D'$
- Or simply two classes $\mathsf{Y} = 0$ vs. $\mathsf{Y} = 1$
- Treat $\hat{\mathsf{Y}} := [\![\mathsf{M}(\cdot) \in E]\!]$
  - $\Pr(\mathsf{M}(D) \in E) = \Pr(\hat{\mathsf{Y}} = 1 | \mathsf{Y} = 0)$: false positive rate; type-1 error
  - $\Pr(\mathsf{M}(D') \in E) = \Pr(\hat{\mathsf{Y}} = 1 | \mathsf{Y} = 1)$: true positive rate; power
- DP: $\mathtt{FPR} \leq \exp(\epsilon) \cdot \mathtt{TPR} + \delta$

J. Dong et al. "Gaussian Differential Privacy". *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 84, no. 1 (2022), pp. 3–37.

α Rényi-DP

$$\mathbb{D}_\alpha(\mathsf{M}(D)\|\mathsf{M}(D')) := \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{\mathsf{X}\sim q}\left(\frac{p(\mathsf{X})}{q(\mathsf{X})}\right)^\alpha \le \epsilon$$

$$\text{equivalently,}\quad \mathop{\mathbb{E}}_{\mathsf{X}\sim p} e^{(\alpha-1)r(\mathsf{X})} \le e^{(\alpha-1)\epsilon}$$

- p and q are the densities of $\mathsf{M}(D)$ and $\mathsf{M}(D')$, respectively
- Log odds ratio: $r = \log\frac{p}{q}$; a.k.a. privacy loss
- $\mathbb{D}_\alpha = \log\left[\mathbb{E}_{\mathsf{X}\sim p}(r(\mathsf{X}))^{\alpha-1}\right]^{\frac{1}{\alpha-1}}$ increasing w.r.t. $\alpha \ge 1$, in particular
  - $\alpha \downarrow 1 \implies \mathbb{D}_\alpha \to \mathsf{KL}$
  - $\alpha \to \infty \implies \mathbb{D}_\alpha \to \max_{\mathbf{x}} \log\frac{p(\mathbf{x})}{q(\mathbf{x})}$, used in $(\epsilon, 0)$-DP (see slide 13)

I. Mironov. "Rényi differential privacy". In: *IEEE 30th computer security foundations symposium*. 2017, pp. 263–275.

The Many Shades of DP

- ϵ-DP: log odds ratio r uniformly bounded by ϵ
- (ϵ, δ)-DP: roughly, with probability $1 - \delta$, we have $r \leq \epsilon$
  - anything can happen for the remaining δ probability
  - sacrificing some δ proportion for (much?) better utility
  - the smaller ϵ or δ is, the stronger the privacy guarantee
- α-DP: bounds the exponential moment of r
  - smoother transition than (ϵ, δ)-DP
  - implies (ϵ, δ)-DP by e.g. Markov's inequality
  - the bigger α or the smaller ϵ is, the stronger the privacy guarantee

Calculus for DP

- Post-processing: If M is DP, so is $\mathbf{T} \circ \mathsf{M}$ for any $\mathbf{T}$
- Parallel composition: $D = \cup_k D_k$, each M_k is DP, then $\mathsf{M}(D) := \big(\mathsf{M}_1(D_1), \ldots, \mathsf{M}_K(D_K)\big)$ is DP
- Sequential composition: $\big(\mathsf{M}(D), \mathsf{N}(D, \mathsf{M}(D))\big)$ is $(\alpha, \epsilon_N + \epsilon_M)$-RDP
  - cannot ask too many questions or run ML algorithms for too many epochs!
  - often been heavily abused in practice
- Differ by a group of k: $(k\epsilon, 0)$-DP
- Subsampling

J. Domingo-Ferrer et al. "The limits of differential privacy (and its misuse in data release and machine learning)". *Communications of the ACM*, vol. 64, no. 7 (2021), pp. 33–35.

Gaussian Mechanism

$$\mathsf{M}(D) := f(D) + \boldsymbol{\xi}, \quad \text{where} \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

- Sensitivity: $\Delta_2 f := \sup_{D \sim D'} \|f(D) - f(D')\|^2_{\Sigma^{-1}}$
- (α, ϵ)-RDP with $\epsilon = \frac{\alpha}{2}\Delta_2 f$
- (α, ϵ)-RDP $\implies$ $(\epsilon + \frac{1}{\alpha - 1} \log \frac{1}{\delta}, \frac{\delta}{\alpha})$-DP
  - note $\alpha \to \infty \implies \mathbb{D}_\alpha \to \max_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \implies (\epsilon, 0)$-DP
  - to achieve $\alpha \to \infty$ with Gaussian mechanism: $\epsilon = \frac{\alpha}{2}\Delta_2 f \to \infty$

DP-SGD

Algorithm 1: Differentially private stochastic gradient descent

```
Input: model $\mathbf{w}$; data $\mathbf{x}_1, \ldots, \mathbf{x}_n$; noise $\sigma$, gradient bound $C$, batch size $b$
for $t = 0, 1, \ldots$ do
    sample a random batch $B_t$ with size $b$
    for $i \in B_t$ do
        $\mathbf{g}_i \leftarrow \nabla_{\mathbf{w}} \ell(\mathbf{x}_i; \mathbf{w})$                                  // compute grad
        $\mathbf{g}_i \leftarrow \mathbf{g}_i / \max\{1, \|\mathbf{g}_i\|_2 / C\}$                     // grad clipping
    $\mathbf{g} \leftarrow \left[\frac{1}{b} \sum_{i \in B_t} \mathbf{g}_i\right] + \sigma C \boldsymbol{\xi}$                              // adding noise
    $\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{g}$                                                // grad descent
    $\mathbf{w} \leftarrow \mathrm{P}(\mathbf{w})$                                                       // projection
```

M. Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 308–318.

Application in Generative Models

- Modern generative models are powerful, e.g., ChatGPT, DALLE-2
  - We can release the generative model as a proxy of releasing data
  - We can conduct data analysis / ML downstream tasks using generated data
- How to protect privacy when sensitive data (medical records, face images) are used in training?
- One solution: Differentially Private Generative Models - equip generative models with DP guarantees

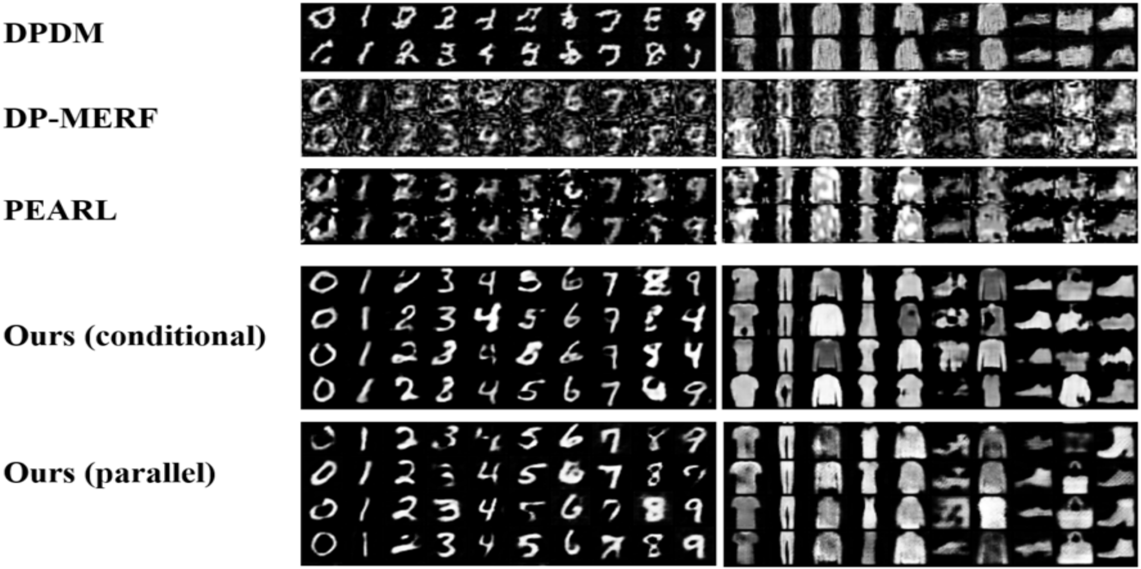


Figure 2: Qualitative comparison under $(0.2, 10^{-5})$-DP on MNIST and Fashion MNIST

D. Jiang et al. "Functional Rényi differential privacy for generative modeling" *Advances in Neural Information Processing Systems*, vol. 36 (2023).