CYB-4203 / 6203 · UNIT 8 · PRESENTATION 17

# Differential Privacy & Federated Learning

Secure & Trustworthy AI · Spring 2026 · Dallas Elleman

Interactive version at
https://dallaselleman.github.io/cyb-4203-6203-spring-2026/course_materials/demos/differential-privacy-explainer.html

# The Illusion of Invisibility

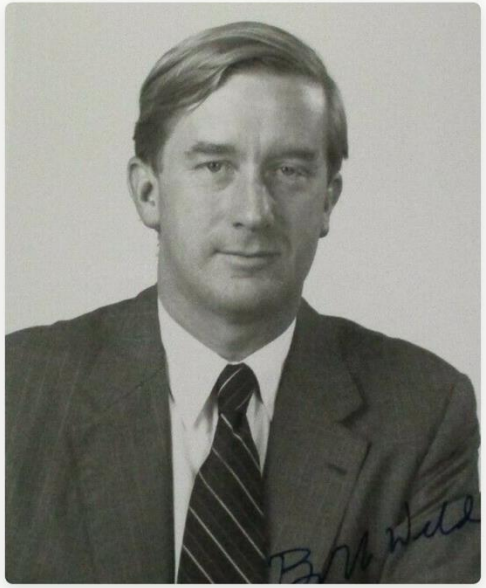
We want to collect and analyze data to **cure diseases**, to **train AI**, to **plan our cities**.

But the more detailed that data is, the higher the risk of *privacy harms* for the data sources — i.e., people.

The field that tries to resolve this paradox is called **privacy-enhancing technology**.
And to understand where it is now, we have to go back to **1996**.

# 1996





**Gov. William F. Weld**
Governor of Massachusetts

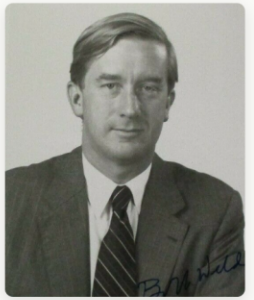
The internet is a toddler. Anonymization means taking a black marker to the obvious fields.

# Direct Identifiers

The **Group Insurance Commission** in Massachusetts decides (with good intentions) to release anonymized data on state employees' hospital visits — the idea was to let researchers crunch the numbers and find ways to lower healthcare costs.

**GIC HOSPITAL VISIT RECORDS (1996)**

| NAME | ADDRESS | SSN | ETHNICITY | SEX | BIRTH DATE | ZIP | VISIT DATE | DIAGNOSIS | PROCEDURE | MEDICATION | CHARGE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| William F. Weld | 15 Louisburg Sq, Boston | 042-31-8876 | White | M | 1945-07-31 | 02138 | 1996-05-14 | Hypertension | Blood panel | Lisinopril | $1,840 |
| Margaret O'Shea | 88 Beacon St, Boston | 019-44-2210 | White | F | 1962-03-11 | 02116 | 1996-02-22 | Asthma | Pulmonary test | Albuterol | $620 |
| David Chen | 412 Mass Ave, Cambridge | 127-88-5501 | Asian | M | 1978-09-04 | 02139 | 1996-08-03 | Migraine | MRI | Sumatriptan | $2,410 |
| Rosa Alvarez | 27 Blue Hill Ave, Roxbury | 553-12-6049 | Hispanic | F | 1955-12-19 | 02119 | 1996-06-30 | Diabetes, Type II | A1C panel | Metformin | $980 |
| James Whitaker | 9 Elm St, Worcester | 204-76-1183 | Black | M | 1949-11-02 | 01609 | 1996-04-17 | Arthritis | X-ray, knee | Naproxen | $1,275 |
| Linda Petrov | 56 Walnut Ave, Springfield | 396-55-7720 | White | F | 1970-06-24 | 01108 | 1996-07-09 | Depression | Psych eval | Fluoxetine | $540 |
| Marcus Johnson | 301 Seaver St, Dorchester | 611-09-3354 | Black | M | 1983-01-15 | 02121 | 1996-09-28 | Fractured radius | Cast, ortho | Oxycodone | $3,120 |
| Emily Tanaka | 74 Harvard St, Brookline | 445-22-9918 | Asian | F | 1990-04-07 | 02446 | 1996-11-05 | Pneumonia | Chest X-ray | Azithromycin | $1,680 |



Gov. Weld · a state employee, so his record is in the release.

Remove Direct Identifiers | Reset

To anonymize the data, they deleted **direct identifiers** — names, addresses, Social Security numbers — of patients and called it good.

Latanya Sweeney
MIT researcher, 1996

**Latanya Sweeney**, a researcher at MIT, saw a problem — and decided to prove a point.

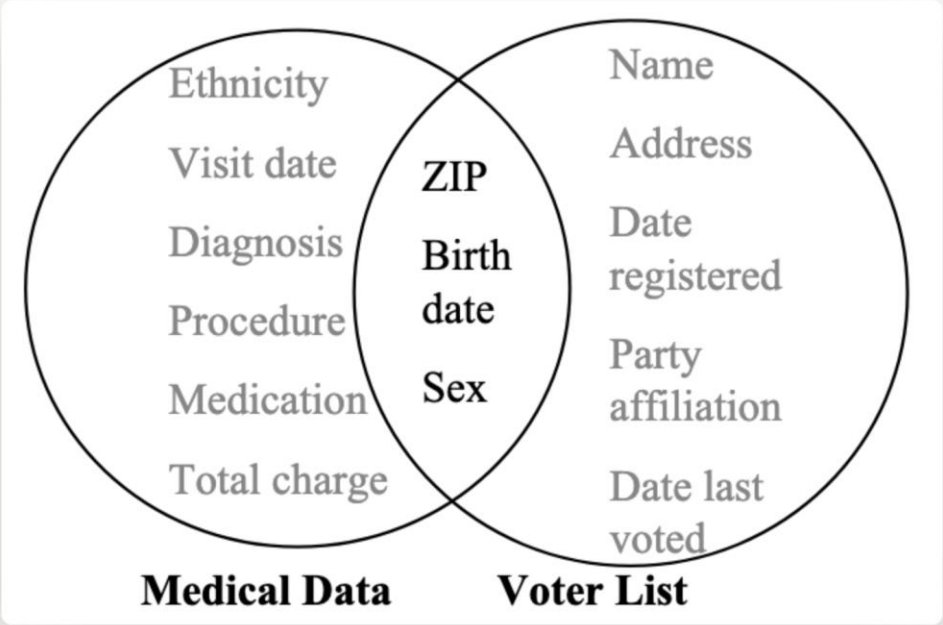
She bought the voter registration list (a public record) from Cambridge, MA for ***$20***.

She then cross-referenced this voter list with the released medical records, found Governor Weld's personal medical records in the "de-identified" data set, and **mailed them to him** to demonstrate the problem.

CAMBRIDGE, MA VOTER REGISTRATION LIST — PUBLIC RECORD, $20

| NAME | ADDRESS | ZIP | BIRTH DATE | SEX | DATE REGISTERED | PARTY | LAST VOTED |
|---|---|---|---|---|---|---|---|
| William F. Weld | 15 Louisburg Sq, Boston | 02138 | 1945-07-31 | M | 1968-09-12 | Republican | 1996-03-05 |
| Anne Driscoll | 210 Broadway, Cambridge | 02139 | 1952-02-18 | F | 1974-10-01 | Democrat | 1996-03-05 |
| Robert Kim | 47 Garden St, Cambridge | 02138 | 1965-11-22 | M | 1987-08-14 | Independent | 1994-11-08 |
| Sarah Goldberg | 12 Brattle St, Cambridge | 02138 | 1971-04-09 | F | 1989-06-03 | Democrat | 1996-03-05 |
| Michael O'Brien | 305 Mass Ave, Cambridge | 02139 | 1948-08-30 | M | 1970-02-19 | Democrat | 1995-11-07 |
| Priya Natarajan | 88 Prospect St, Cambridge | 02139 | 1980-06-17 | F | 1998-09-20 | Democrat | 1996-03-05 |
| Thomas Reilly | 59 Hancock St, Cambridge | 02139 | 1955-01-04 | M | 1977-11-08 | Republican | 1996-03-05 |
| Janet Williams | 22 Oxford St, Cambridge | 02138 | 1963-10-25 | F | 1984-04-15 | Independent | 1994-11-08 |

# The Linkage Attack




*Cross-reference two safe-looking datasets on their shared columns — and anonymity vanishes.*

INTERACTIVE DEMO

# Finding the Governor

Click each filter in sequence. Watch the two tables converge.

1. Birth date: 1945-07-31 | 2. Gender: Male | 3. ZIP: 02138 | Reset

**ANONYMIZED MEDICAL RECORDS**

| BIRTH DATE | ZIP | GENDER | DIAGNOSIS |
|---|---|---|---|
| 1945-07-31 | 02138 | M | Hypertension |
| 1945-07-31 | 02139 | M | Diabetes |
| 1945-07-31 | 02140 | M | Back Pain |
| 1945-07-31 | 02138 | F | Hypothyroid |
| 1945-07-31 | 02141 | F | Insomnia |
| 1945-07-31 | 02142 | F | Migraine |
| 1962-03-14 | 02138 | F | Depression |
| 1962-03-14 | 02138 | M | Heart Disease |
| 1978-11-22 | 02138 | M | Eczema |

**CAMBRIDGE VOTER LIST ($20)**

| NAME | BIRTH DATE | ZIP | GENDER |
|---|---|---|---|
| William F. Weld | 1945-07-31 | 02138 | M |
| James Chen | 1945-07-31 | 02139 | M |
| Michael Kowalski | 1945-07-31 | 02140 | M |
| Margaret Hayes | 1945-07-31 | 02138 | F |
| Patricia O'Brien | 1945-07-31 | 02141 | F |
| Elizabeth Wong | 1945-07-31 | 02142 | F |
| Robert Johnson | 1962-03-14 | 02138 | M |
| Jennifer Adams | 1962-03-14 | 02138 | F |
| David Park | 1978-11-22 | 02138 | M |

MEDICAL RECORDS REMAINING
20

VOTER RECORDS REMAINING
30

# 87%

of the US population can be uniquely identified using only **ZIP code**, **birth date**, and **gender**.

*— Sweeney, 2000*

Fields like these are called **quasi-identifiers** — data points that, while not unique alone, can be combined with other public or private information to re-identify individuals.

Data scientists realized they needed a new approach. They couldn't just delete names — they needed a **mathematical guarantee** of safety.

## Common Quasi-Identifiers by Category

### GEOGRAPHIC INFORMATION

- ZIP codes, census tracts, counties
- Business or facility names
- Movement patterns or locations
- School districts or neighborhoods

### DEMOGRAPHIC DETAILS

- Age ranges or birth years
- Gender or sex
- Education level or occupation
- Race or ethnicity
- Income brackets

### TEMPORAL DATA

- Service, admission, or discharge dates
- Purchase timelines or patterns
- Employment history or tenure
- Subscription duration
- Event attendance dates

### LIFESTYLE INDICATORS

- Vehicle ownership details
- Educational institutions attended
- Homeownership status
- Purchase categories or preferences
- Leisure activities or memberships

# k-Anonymity: The Spartacus Strategy

*“I am Spartacus.”*

The Romans demand the slaves identify Spartacus. One by one, every slave stands and claims, *“I am Spartacus.”* If everyone looks like the target, no one can be singled out.

This is **k-anonymity**: transform the data so that every record is indistinguishable from at least **k – 1** others on its quasi-identifiers. Instead of age *32*, store *30–40*. Instead of ZIP *02138*, store *021***. Suppress the outliers you can’t generalize.

If **k = 5**, every person in the release is hiding in a crowd of at least five identical-looking records. The Romans can’t pick the right one.

INTERACTIVE DEMO

# k-Anonymizing the Governor's Dataset

The same 20 medical records from *Finding the Governor*. Two tools: **generalization** (replace precise values with buckets) and **suppression** (drop rows that can't be hidden in a group of k).

k = 1 (raw) | k = 3 | k = 5 | Reset

**RAW MEDICAL RECORDS (K = 1)**

| BIRTH DATE | ZIP | GENDER | DIAGNOSIS |
|---|---|---|---|
| 1945-07-31 | 02138 | M | Hypertension |
| 1945-07-31 | 02139 | M | Diabetes |
| 1945-07-31 | 02140 | M | Back Pain |
| 1945-07-31 | 02138 | F | Hypothyroid |
| 1945-07-31 | 02141 | F | Insomnia |
| 1945-07-31 | 02142 | F | Migraine |
| 1962-03-14 | 02138 | F | Depression |
| 1962-03-14 | 02138 | M | Heart Disease |
| 1978-11-22 | 02138 | M | Eczema |

### GENERALIZATION

Replace precise values with buckets. Exact birth date → birth decade. Full ZIP → ZIP prefix ( `021**` ). Ages become ranges. The data loses detail but each row becomes indistinguishable from others in its bucket.

### SUPPRESSION

Some rows can't be hidden in a group of *k* — they're too rare. Drop them entirely. Outliers pay the price for everyone else's anonymity.

INTERACTIVE DEMO

# k-Anonymity & the Homogeneity Attack

k = 1 (raw) | k = 3 (generalized) | k = 5 (anonymized) | Reveal the Attack

RAW HOSPITAL DATASET (K = 1)

| AGE | ZIP | GENDER | DIAGNOSIS |
|---|---|---|---|
| 32 | 02138 | M | Heart Disease |
| 34 | 02139 | M | Heart Disease |
| 36 | 02141 | M | Heart Disease |
| 38 | 02142 | M | Heart Disease |
| 31 | 02140 | M | Heart Disease |
| 45 | 02138 | F | Arthritis |
| 48 | 02139 | F | Migraine |
| 51 | 02140 | F | Depression |
| 55 | 02141 | F | Diabetes |

# Differential Privacy

2006

**Dwork, McSherry, Nissim, Smith**

> *Stop protecting the data.*
> *Start protecting the* **process.**

Instead of trying to sanitize a dataset until it is "safe to release," define a mathematical property of the *release mechanism* itself: the output of any query should look effectively the same whether or not any single individual is in the database.

That property is called **ε-differential privacy**. The parameter ε controls how much any one person can influence the answer — and therefore how much they can be leaked.

PLAY →

**Differentially Private Tetris**

Feel the effect of ε on gameplay — low budget means noisy controls.

## Frosted Glass

*You can see the **shape**. You cannot see the pores.*

Privacy Budget (ε) ε = 1.00

Compute the true answer — then deliberately blur it with **mathematical noise**. Low ε = heavy blur = strong privacy. High ε = crisp detail = weak privacy.

INTERACTIVE DEMO

# The Epsilon Dial

Privacy Budget (ε) ε = 1.00 Re-roll Noise

TRUE MEAN $59,400

REPORTED MEAN $59,427

PRIVACY ERROR +$27

*The balance point — moderate privacy, moderate utility*

# Differential Privacy in the Wild

## Apple ↗

Your iPhone adds noise **before** data ever leaves the device. Apple learns trending emoji. It never sees your keystrokes.

## Google Maps ↗

"Popular Times" tracks device density with noise. They know the grocery store is busy. They don't know you're in it.

## 2020 US Census ↗

For the first time, the official population count has mathematical noise baked in — to defeat reconstruction attacks.

## Bank Collaboration (SMPC)

Two banks detect money laundering together **without ever sharing customer lists**. The secret files talk. The humans don't see.

# Minority Erasure: The Costs of Differential Privacy

If a group has **1,000 people**, adding ±5 is a rounding error.

If a group has **three people**...

*...they can mathematically cease to exist.*

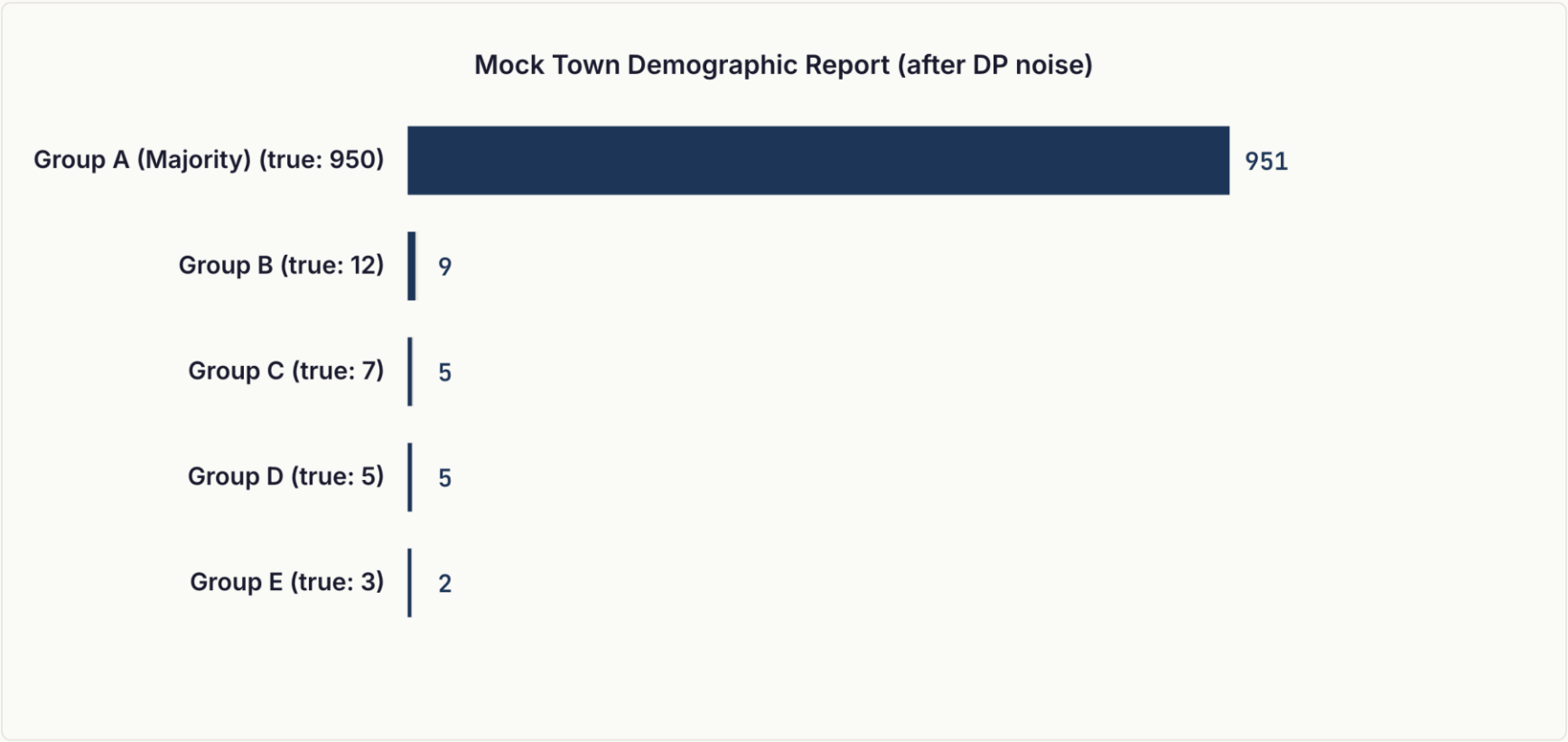
INTERACTIVE DEMO

# Minority Erasure

Privacy Budget (ε) ε = 1.00

Re-roll Noise

| GROUPS TRACKED | GROUPS ERASED | TRUE POPULATION |
|---|---|---|
| 5 | 0 | 977 |



# What's at Stake

FUNDING & REPRESENTATION

Census data decides where schools get built, where voting-district lines are drawn, and where federal aid lands. If the privacy noise washes out a small community, they **lose funding and political representation** — even though the people are still there.

VOTING RIGHTS ACT § 203

Federal law requires jurisdictions to provide ballots and voter materials in a non-English language whenever census data shows a language-minority population above a fixed threshold. If DP noise pushes those counts below the line, the trigger fails — and **voters lose translated ballots entirely**, disenfranchising communities the protection was designed for.

READ THE MEMO →

**Census Distortion Program: The Perils of Differentially Private Census Data**

Meredith Strohm Gunter, Weldon Cooper Center for Public Service, University of Virginia · Memorandum to Gov. Ralph Northam, January 23, 2020 · PDF

# Differential Privacy, AI/ML, and the "Matthew Effect"

*"For to everyone who has, more will be given... but from the one who has not, even what they have will be taken away."*

— MATTHEW 25:29

### MAJORITY WINS, MINORITY LOSES

The model learns the majority class **perfectly** — its signal is strong enough to survive the noise. But it **completely fails** to learn the minority class. Their signal is the same order of magnitude as the noise added to protect them.

### OUTLIERS CLIPPED

DP training mechanisms (e.g. DP-SGD) *clip* per-example gradients before adding noise. Rare and informative examples — the edge cases — get their influence capped. The model never hears from them.

### BIAS, AMPLIFIED

The result: an AI system that is **more biased than it would have been naturally**. The model doesn't just inherit bias from the data — the privacy mechanism itself makes it worse.

### ERASURE, NOT PROTECTION

If you're too unique, the algorithm **literally erases you**, distorting reality. The same mechanism that promises to protect you from identification also protects the model from learning you exist.

# Differential Privacy, Compliance, and Privacy Theater

EU

### GDPR 2018

The General Data Protection Regulation sets strict rules for collection, purpose limitation, and the right to erasure. “Anonymized” data — data that can no longer be linked back to a person — falls outside its scope entirely, creating a strong incentive to *really* anonymize, not just redact names.

QC

### Quebec Law 25 2022

Requires that personal data be **irreversibly anonymized** before release or reuse — the black-marker method doesn't qualify. Penalties reach **$25M CAD** or 4% of global revenue. This effectively forces organizations toward techniques like differential privacy.

THE LOOPHOLE

*“We use differential privacy.”*

**Great. What did you set ε to?**

The law says “use DP.” It usually doesn’t say *what ε to use*. A company can announce DP for the PR win, ship it with ε = 14 or 20, and be technically compliant while offering almost no real protection. Fancy math, safe door wide open. This is **privacy theater**.

> *Privacy is not a switch.*
> *It's a dial.*

*The hand turning it balances* **safety**, **accuracy**, *and* **fairness**.

FURTHER READING

**A Comprehensive Guide to Differential Privacy: From Theory to User Expectations**
Karmitsa et al., 2025 · arXiv:2509.03294

**Programming Differential Privacy**
Near & Abuah, 2025 · programming-dp.com

# Intro to Federated Learning

EXPLORABLE →

## Federated Learning

Google PAIR · pair.withgoogle.com