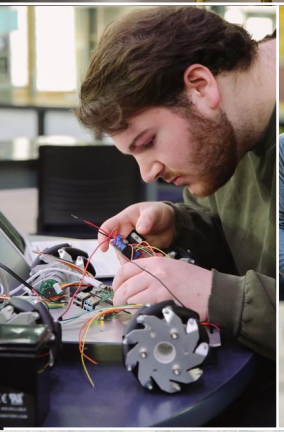





CIS-545-001 / CIS-545-002: Data Security & Privacy

Fall 2025

Project Presentation

Differentially Private Analysis of the COMPAS Dataset

Group 2:

Anthony Lewis

Lincoln Badejo

Margil Funtanilla

Loren Fang

Tom Tooma

Introduction

- Differential privacy (DP) is a leading framework for protecting sensitive data by adding carefully calibrated randomness to the mechanism that produces the outputs, such as statistics or model predictions, so individuals cannot be uniquely identified, while preserving overall data usefulness.
- In practice, DP is challenging to implement, privacy budgets (ε), mechanism choices, and privacy-utility trade-offs are often unclear.
- These challenges matter most in high-stakes domains like criminal justice, where datasets such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) contain highly sensitive personal information, and improper handling threatens privacy, fairness, and potential re-identification. Yet these datasets are routinely used for research and policy analysis.
- This project uses the COMPAS dataset as a case study to demonstrate differential privacy in action. We apply DP mechanisms to common analytical tasks, such as statistical queries, to examine how privacy-preserving noise affects both privacy (ε) and data utility.

Project Objectives

The main objective of the project is to demonstrate how differential privacy (DP) can be applied in real-world settings using the COMPAS dataset as a case study. Specifically,

- Explore a variety of queries to identify the most valuable insights from the dataset;
- Apply DP noise (Laplace and Gaussian mechanisms) to key statistical queries to evaluate their impact on privacy and utility;
- Implement visualizations that effectively represent the COMPAS dataset and facilitate comparison of DP mechanisms; and
- Show that strong privacy guarantees can be achieved without sacrificing interpretability.

Differential Privacy Framework

- **Intuition:** Differential privacy (DP) protects individuals by ensuring that the output of a mechanism *M* is nearly the same whether or not any single record is included.
- **Definition:** A randomized mechanism *M* (e.g., a set of queries) satisfies (ϵ,δ)-DP if outputs for neighboring datasets *d1*, *d2* are indistinguishable:

$$\Pr[M(d1) \in S] \leq e^{\varepsilon} * \Pr[M(d2) \in S] + \delta$$

- **Key Parameters:**
  - **ε (Privacy budget):** Smaller → stronger privacy, less utility
  - **δ (Failure probability):** Small chance DP guarantee fails; δ = 0 → pure DP
- **Mechanisms Implemented:**
  - **Laplace:** $M(d) = f(d) + Lap\left(\frac{s}{e}\right)$
  - **Gaussian:** $M(d) = f(d) + \frac{s * \sqrt{2 * log\frac{1.25}{\delta}}}{e}$
- **Sensitivity (Δf):** Maximum impact of one record on query output (Δf = 1 for counts)
- **Project Goal:** Add noise to queries to maximize privacy while preserving interpretability and utility

Design: Tools and Approach

- Programming Language used: Python
- Libraries utilized:
  - numpy
  - pandas
  - matplotlib.pyplot
- Utilized the Jupyter Notebook coding platform.
- Google Drive to store project files.
- Google Collab to work on Jupyter notebook as a group.
- Google Chat to communicate with members regarding the project.

Implementation

Data

- The COMPAS - Correctional Offender Management Profiling for Alternative Sanctions dataset contains variables used by the COMPAS algorithm in scoring criminal defendants, along with their outcomes within 2 years of the decision, in Broward County, Florida.
- It is an algorithm designed to assess a defendant's likelihood of recidivism.
- Recidivism is a term used to describe individuals who re-offend.
- COMPAS is used by criminal justice agencies to inform decisions regarding the placement, supervision and case management of offenders.
- Published data from Propublica, 'compas_scores_two_years.csv', was used in this analysis.

Implementation

- **Data Cleaning and Processing**
  - Filtered raw COMPAS two-year recidivism dataset using the ProPublica criteria:
    - +/- 30 days of screening
    - valid recidivism labels
    - removal of traffic cases
  - A final dataset was produced of 6,172 defendants with 53 attributes.
  - Audited fields for re-identify risk, classifying them as PII, SPII, and indirect identifiers (INDIR)
  - Two fields *start* and *end* were grouped as ‘Neither’ as none fit as identifiable information
  - 51/53 confirmed columns required privacy handling

PII, SPII, Indirect Identifiers

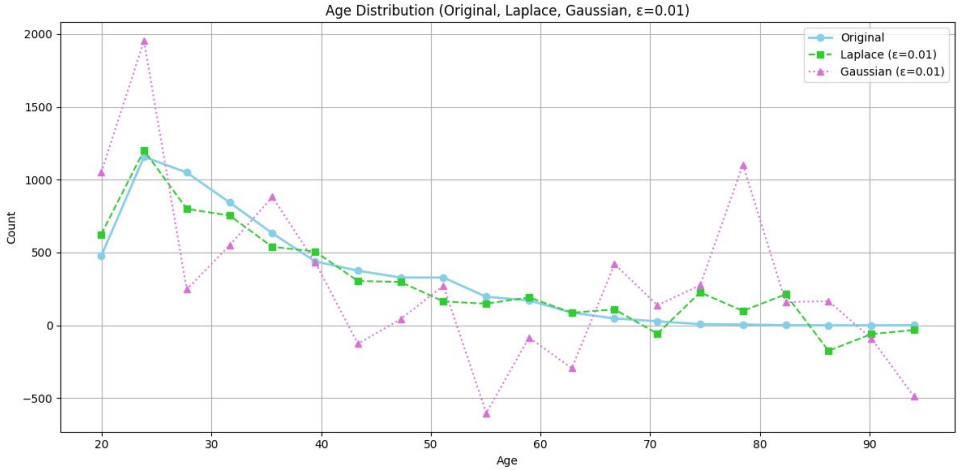
Category	Fields	Count
PII	full name, first, last, id	4
SPII	race, sex, dob, jail/arrest/case details, assessment	13
INDIR	age, days_b_screening_arrest, offense timing, prior counts	34
Neither	start, end	2

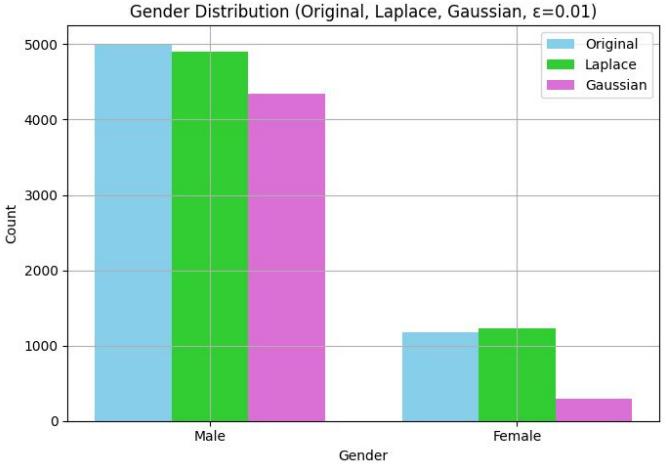
Analytical Workflow

- The analysis proceeded in two layers:
  1. A ground-truth (non-DP) layer
  2. A differentially private layer using both Laplace (ε-DP) and Gaussian ((ε, δ)-DP) mechanisms
- Various queries were evaluated to support our analysis of the COMPAS dataset.
  - These queries include the use of Laplace and Gaussian mechanisms on the distribution of:
    - Racial groups
    - Age
    - Gender
    - COMPAS decile score distribution
    - Two-year recidivism/re-offender rates or the proportion of defendants who committed a new crime over the two years
    - Differences in COMPAS score averages, decile score distribution by race.
  - Each query was executed under different privacy budget settings. For clarity, the following sections present results for a selected subset of these scenarios, while the full set of results is available in our implementation code.

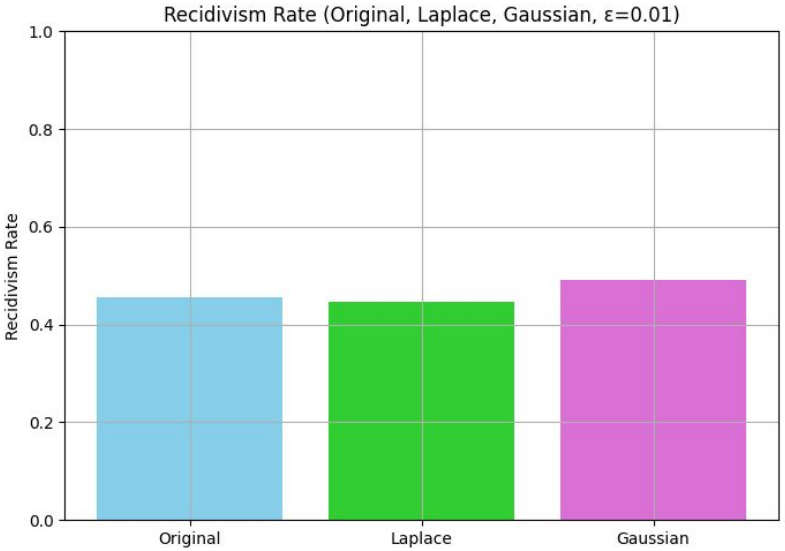
Experimental Evaluation

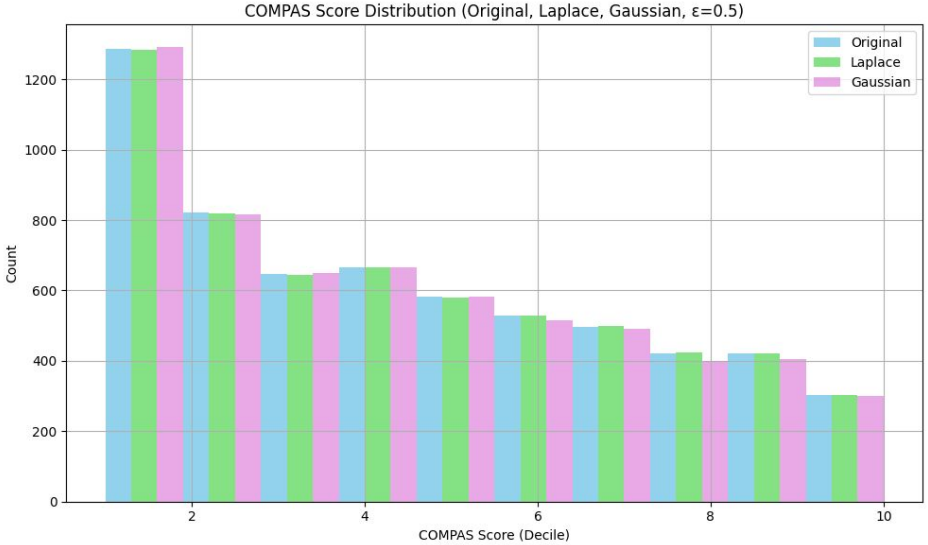
	Original		Laplace (ε=0.5)		Gaussian (ε=0.5)	
Race						
African-American	3175	51.4%	3173.4	51.4%	3175.6	51.5%
Caucasian	2103	34.1%	2104.8	34.1%	2103.9	34.1%
Hispanic	509	8.2%	509.4	8.3%	524.9	8.5%
Other	343	5.6%	342.8	5.6%	340.3	5.5%
Asian	31	0.5%	32.9	0.5%	27.2	0.4%
Native American	11	0.2%	6.6	0.1%	-5.5	-0.1%





Experimental Evaluation



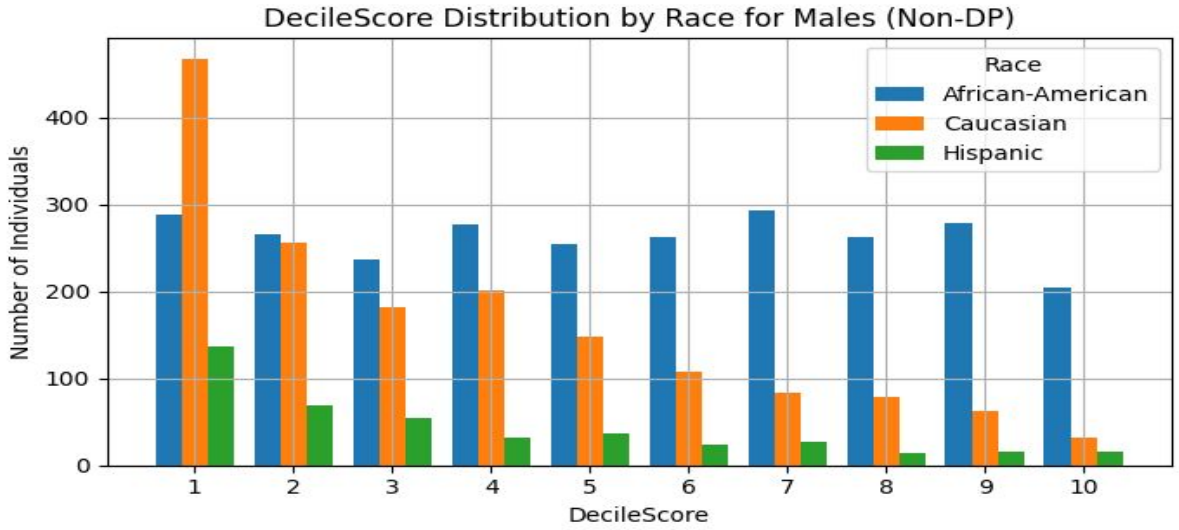



```
Epsilon = 0.01
Original recidivism rate: 45.51%
Laplace noisy recidivism rate: 44.53%
Gaussian noisy recidivism rate: 49.24%
```

```
Epsilon = 0.5
Original mean COMPAS score: 4.42
Laplace noisy mean: 4.42
Gaussian noisy mean: 4.42
```

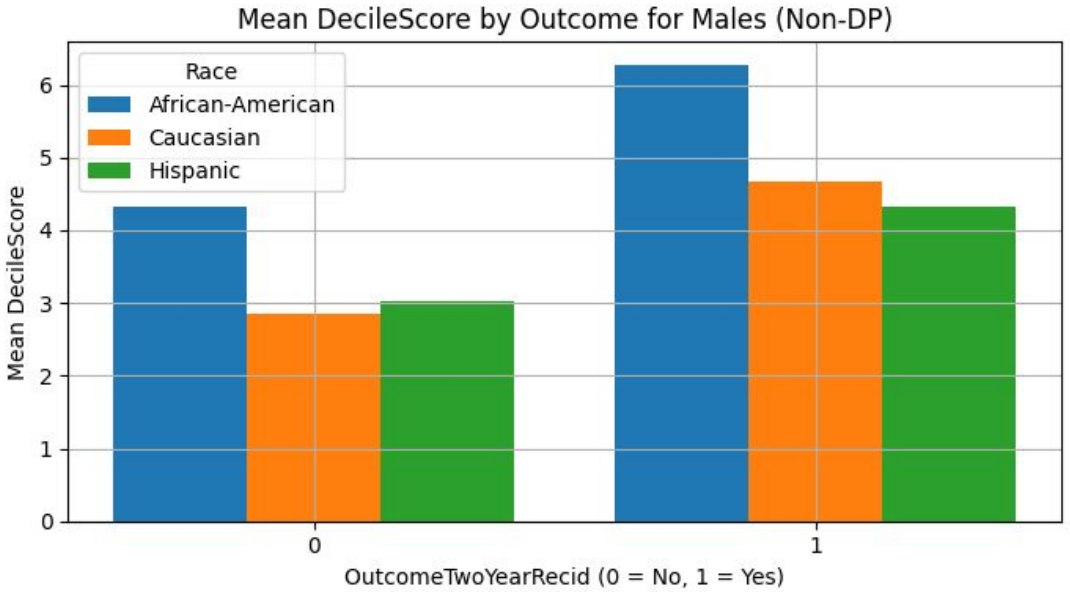
Experimental Evaluation Cont'd

- Query 1: Decile Score Distribution by Race
- Relevance of the query:
  - Baseline fairness / bias check
  - Are some races concentrated in higher COMPAS scores (8–10) more than others?
  - Do African-American defendants tend to receive higher risk scores than Caucasian defendants?



Experimental Evaluation Cont'd

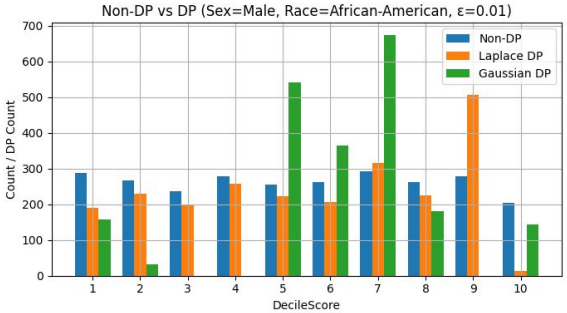
- Query 2: Mean Decile Score by Outcome
- Relevance of the query:
  - Is there a relationship between the Decile Score and Recidivism/re-offender rate?
  - Further Baseline fairness / bias check

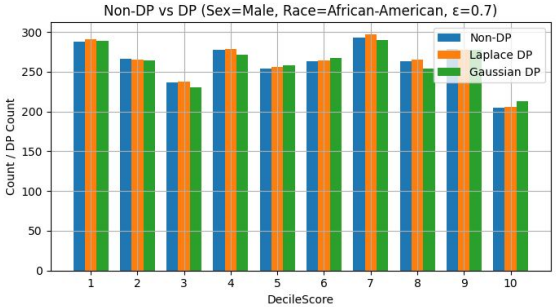


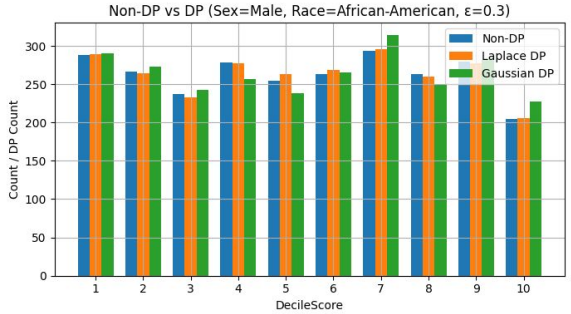
Experimental Evaluation Cont'd

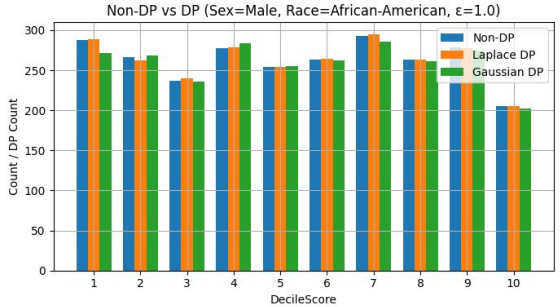
Differential Privacy Analysis on count query

- Compare Non-DP Query result with DP results
- Effect of tuning the Epsilon value on data utility.





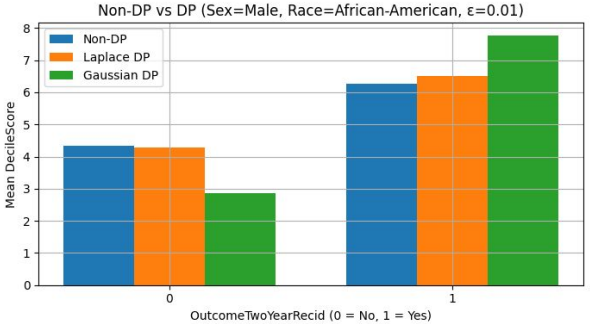


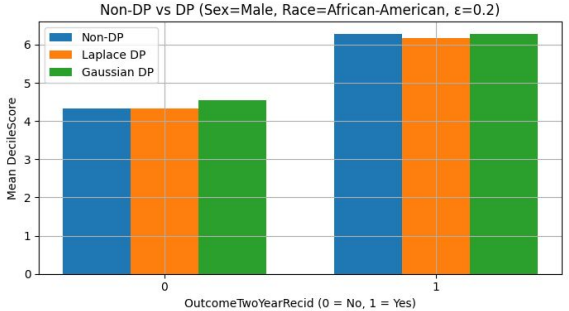


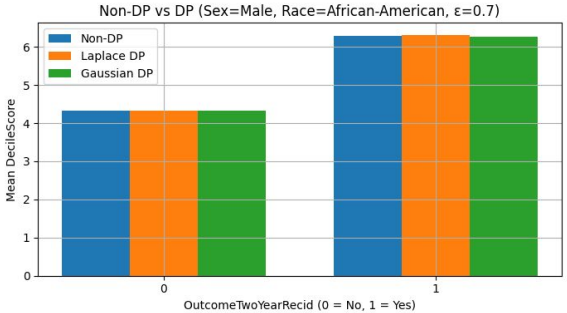
Experimental Evaluation Cont'd

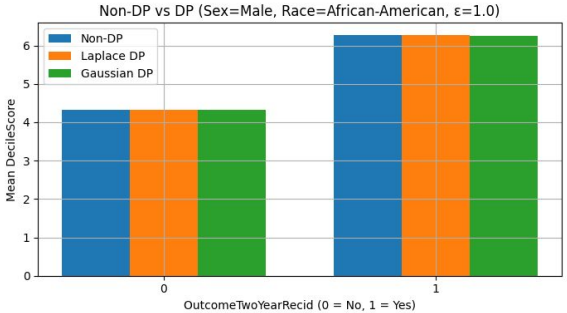
Differential Privacy Analysis on mean query

- Compare Non-DP Query result with DP results
- Effect of tuning the Epsilon value on data utility.





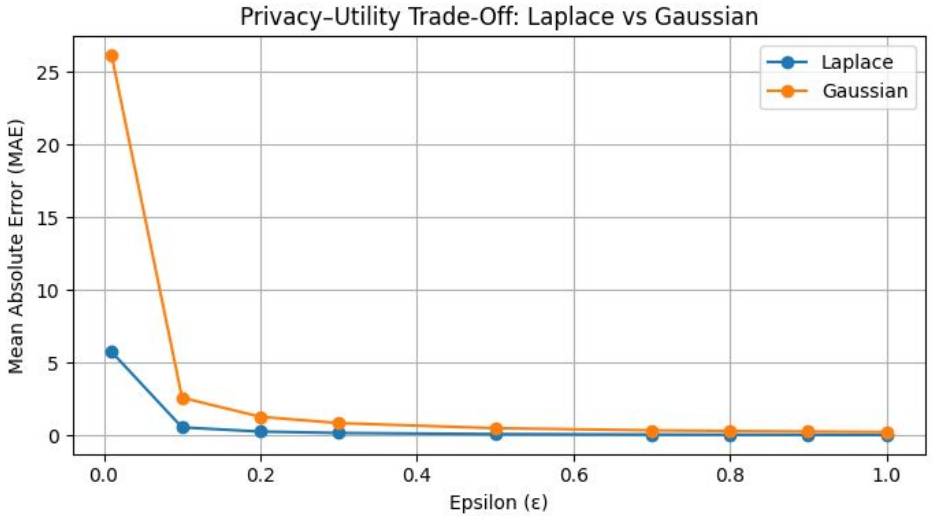




Privacy-Utility Trade Off

What's the take away?

- As **ε increases**, the **MAE drops sharply**
- Stronger privacy (small ε) causes large errors
- Weaker privacy (larger ε) gives much better accuracy.
- Laplace noise is generally more accurate than Gaussian
- There has to be a balance between private & utility.



Challenges / Limitations

Project challenges that occurred during the development of the project were:

- Deciding on the necessary queries to depict the dataset.
- Implementing the Laplace and Gaussian mechanisms to the queries.
- Plotting the data from the queries effectively.
- Maintaining a clean and concise coding structure.

Conclusion

- Our experiments using the COMPAS dataset demonstrate that strong privacy guarantees can be achieved with minimal loss of interpretability in aggregate statistics.
    - The Laplace mechanism generally produced lower data utility loss compared to Gaussian, especially at moderate and high privacy levels, making it well-suited for releasing summary statistics in sensitive domains.
- Despite introducing formal privacy noise, major patterns and disparities observed in the raw data, such as racial gaps in COMPAS scores and recidivism rates, still remained visible after differential privacy was applied. This suggests that differentially private methods can effectively protect individual-level privacy while still enabling meaningful analysis and accountability.

References

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings (Lecture Notes in Computer Science), Shai Halevi and Tal Rabin (Eds.), Vol. 3876. Springer, 265–284.

[2] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 9, 3-4 (2014), 211–407.

[3] Jeff Larson and Marjorie Roswell. 2017. COMPAS-analysis. Github. GitHub - propublica/compas-analysis: Data and analysis for 'Machine Bias'

[4] National Institute of Standards and Technology, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), SP 800-122. Gaithersburg, MD, USA: NIST, 2010. https://csrc.nist.gov/pubs/sp/800/122/final

[5] Protecting Americans From Harmful Data Broker Practices (Regulation V). Federal Register, vol. 89, no. 239, pp. 87621–87727, Dec. 13, 2024. https://www.federalregister.gov/documents/2024/12/13/2024-28690/protecting-americans-from-harmful-data-broker-practices-regulation-v

[6] L. Sweeney. 2000. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper 3, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2000.

[7] Northpointe. 2015. Practitioner's Guide to COMPAS Core. Practitioner-s-Guide-to-COMPAS-Core.pdf

Thank you!
Questions?