Algorithmic Privacy

Dominic Bordelon, Research Data Librarian
University of Pittsburgh Library System

November 8, 2022

Agenda

1. The story so far; why k-anonymity isn't enough
2. Differential privacy
3. Mathematical formalism of differential privacy
4. Hands-on: examples from *Programming Differential Privacy* (Near and Abuah 2021)
5. Closing thoughts

The story so far

Classic approaches to data privacy

- Cryptography influences
- Suppression: not publishing
  - Suppressing an entire dataset
  - Suppressing particular fields; “de-identification” by dropping “PII” or “PHI”[1] columns
  - Suppressing particular subsets (e.g., very small groups)

1. “Personally identifiable information” and “personal/protected health information”

Classic approaches to data privacy

- Cryptography influences
- Suppression: not publishing
- Sampling: publishing "only" a subset of records
- Jittering by hand; swapping
- Generalization
  - Binning (age `54` $\longrightarrow$ `50–60`), rounding
  - Publication of summary statistics and contingency tables

What are some shortcomings of these techniques?

MA Gov. Weld re-identification (1997)

- Latanya Sweeney, then-PhD student at MIT, discovered state employees' health records would be released (de-identified but with gender, DOB, zip code remaining)
- "A mental calculation surprised me: There are 365 days in a year, two *[sic]* genders, and people live about 78 years. Multiplying these numbers gives **56,940 unique combinations**. However, the average five-digit ZIP code in the United States has only about **25,000 people**." (Sweeney 2015)
- Governor William Weld's DOB, gender, and zip code were unique in voter rolls (available for $20) and in the de-identified release, allowing him to be re-identified by Sweeney—example of a *re-identification attack*
- HIPAA (Health Insurance Portability and Accountability Act) subsequently adopted a Privacy Rule which continues to be updated
  - Example: where a zip code appears, only first three digits may be reported publicly
- Subsequent experiments by Sweeney were court-sealed, and she had difficulty publishing and obtaining funding

The concept of k-anonymity

- k answers the question, for a given row in a dataset, how many other rows are identical?
- Or: when grouping by every variable, what is the smallest group size?
- $k = 1$ when there exists one unique record in the dataset (i.e., at least one record is uniquely identifiable, so there is no guarantee of privacy)
- If you are in the dataset, and $k = 10$, you are indistinguishable from 9 other participants; therefore, we want k to be a high number
- Proposed by Samarati and Sweeney (1998)
- Since shown to *not* satisfy privacy needs on its own, but still a useful concept

AOL search log release (2006)

- For three days in 2006, AOL made available a log file of ~20m queries of ~650k users for a 3-month period
- Only “PII” column was a numeric user ID
  - Some (even now) might naively consider these data to be anonymized!
- User Thelma Arnold (among others) was re-identified by The New York Times; publication of this breach, with Arnold’s permission, led to resignations and firings.
- Class-action lawsuit (settled 2013)

The Netflix Prize (2006–2010)

- The Netflix Prize was a public contest by Netflix from 2006-2009 which had competitors use released, de-identified data about viewings in order to optimize Netflix's recommendation algorithm.
- Narayanan and Shmatikov (2008) used the Netflix Prize training data, in combination with IMDB reviews and statistical analysis ("entropic de-anonymization"), in order to successfully link Netflix users to IMDB accounts.
- Netflix cancelled further iterations of the Prize and settled a lawsuit.
- By 2009-10 (time of the controversy), differential privacy had been formulated by Dwork et al. (2006), but was not in adoption yet.

What can we learn from public records and a news story?

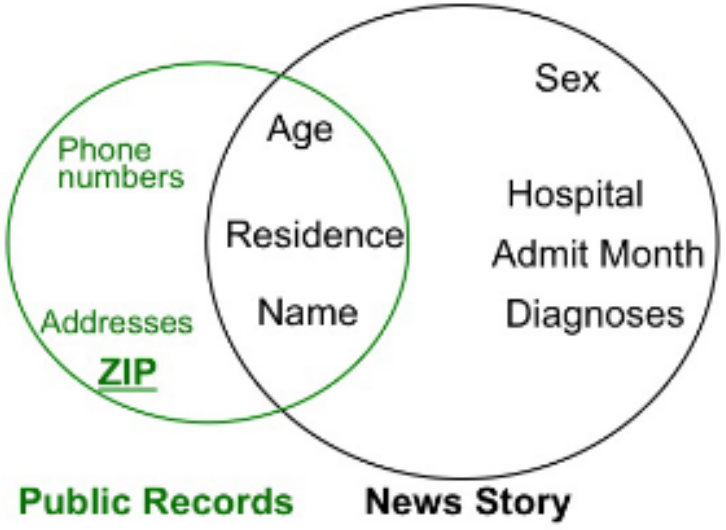


Figure 5. Acquire five-digit ZIP codes from public records using {name, residence information, age} from the news story. Age in years is from news and date of birth from public records.

Source: Sweeney (2015)

What can we also learn from obtainable hospital data?

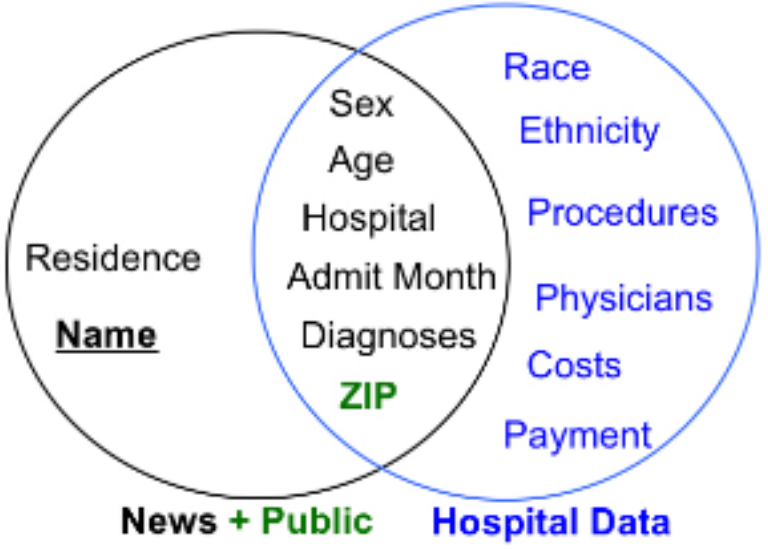


Figure 6. The automated approach matches news information and ZIP (from public records) to hospital data on a combination of {gender, age, hospital, admit month, diagnoses related to incidence, ZIP}, thereby putting a name to a medical record. Age is in years and months and the month of birth comes from public records.

Source: Sweeney (2015)

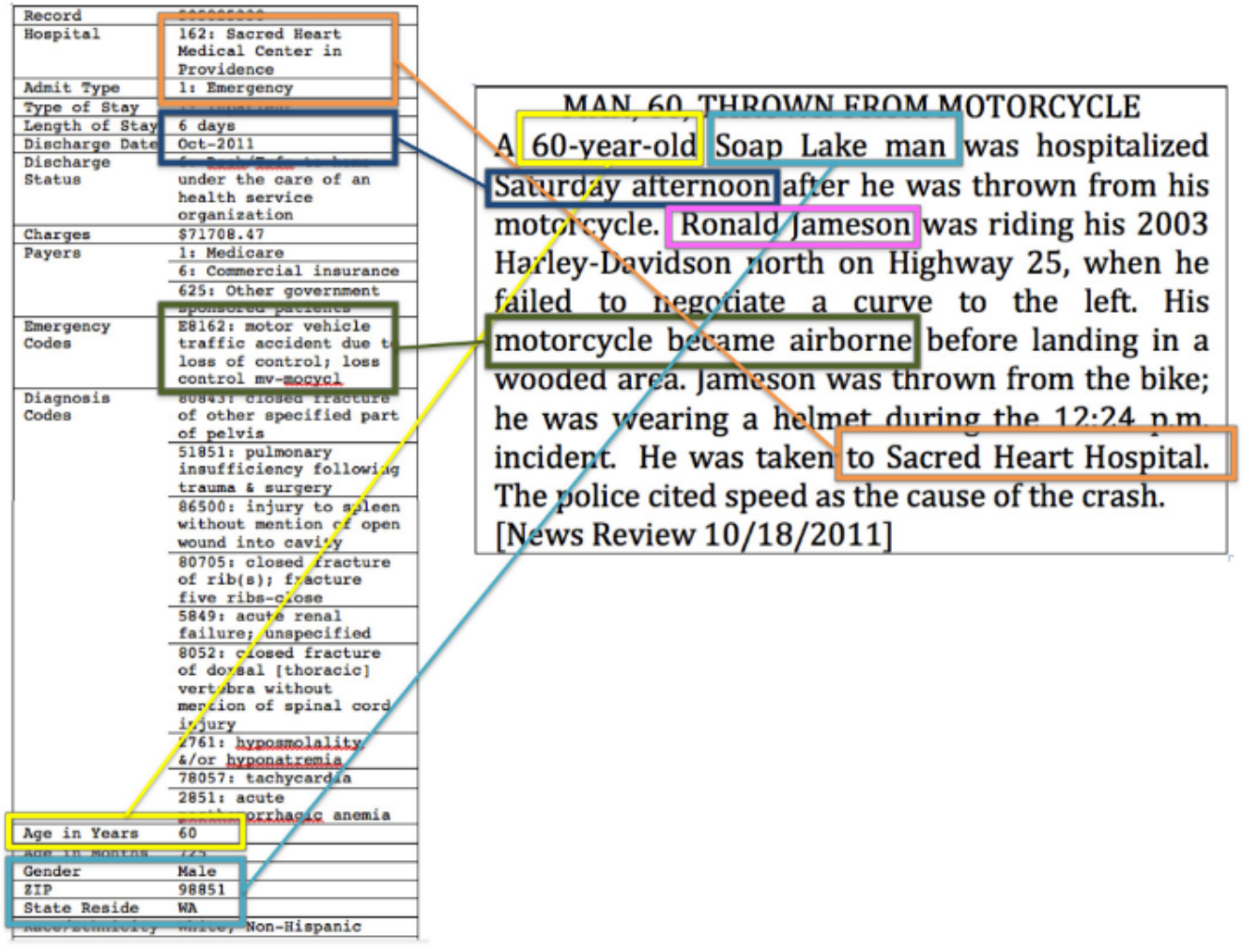


Figure 7. Example of information from news stories uniquely and exactly matching a medical record in publicly available Washington State health data.

Source: Sweeney (2015)

Genomic data privacy

Besides being based in relatively new technologies, genomics poses new privacy challenges…

- Necessarily sensitive and persistent information
- Raw data are not tabular (different algorithms, or transformation, may be needed)
- Use cases span ancestry tracing, forensics, medical diagnostics and screening—each with their own complex ethical issues
- A right to solitude: privacy "could include the right not to be (re)contacted about ancillary findings generated from genomic testing or discovery-driven investigations into existing genomic data sets or by previously unknown relatives" (Wan et al. 2022).

Genomic data

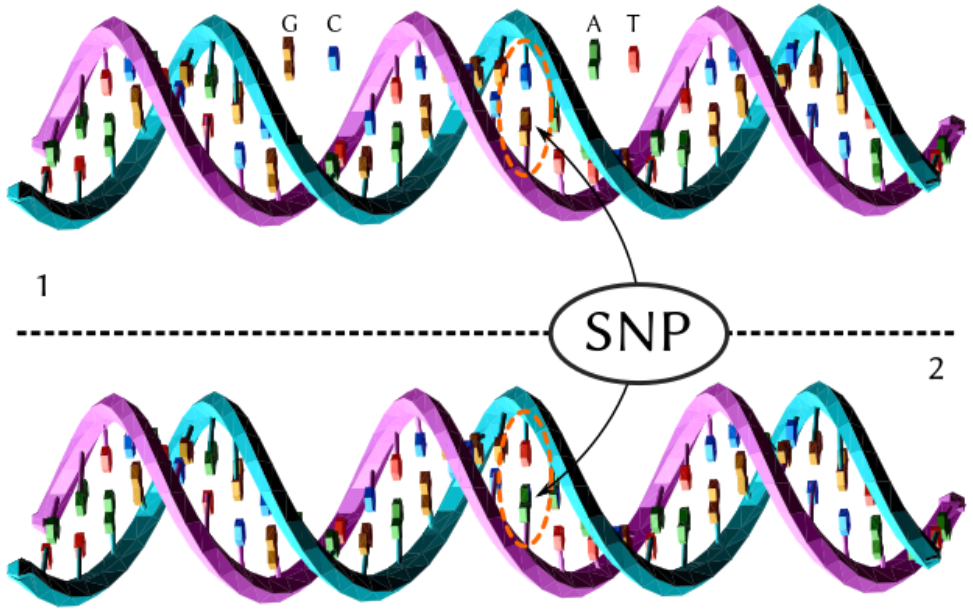
A typical datum in this area is a single-nucleotide polymorphism (SNP), which is a difference in a single nucleotide of genetic code.

SNPs may or may not result in phenotypic differences—and many of them are in non-coding regions of DNA, anyway—but taken altogether, they form a unique “fingerprint.”

SNPs are the raw data; attacks often combine these data and accompanying metadata (e.g., patient’s demographic info).

Single-nucleotide polymorphisms

What a SNP looks like in genetic data.
Source: genome.gov



What SNPs look like in DNA.
Source: SNP model by David Eccles (gringer), CC BY 4.0

Types of genomic privacy intrusion

- Malin (2005) began applying the wider data privacy conversation to genomics, deploying four re-identification techniques:
  - Family structure (i.e., genomic + genealogical data) and combinatorics
  - Genotype-phenotype inference
  - “Trails” of geoinformation (e.g., location of study or hospital)
  - “Dictionary” attacks which decrypt simplistic privacy measures
- Erlich and Narayanan (2014) comprehensively outline vectors of attack in genetic data:
  - Identity tracing attacks: surname inference, DNA phenotyping, demographic identifiers, pedigree structure, side-channel leaks
  - Attribute disclosure attacks using DNA: n=1, genotype frequencies, linkage disequilibrium, effect sizes, trait inference, gene expression
  - Completion attacks: imputation of a masked marker, genealogical imputation of single or multiple relatives
- Wan et al. (2022) provide a very current and comprehensive review

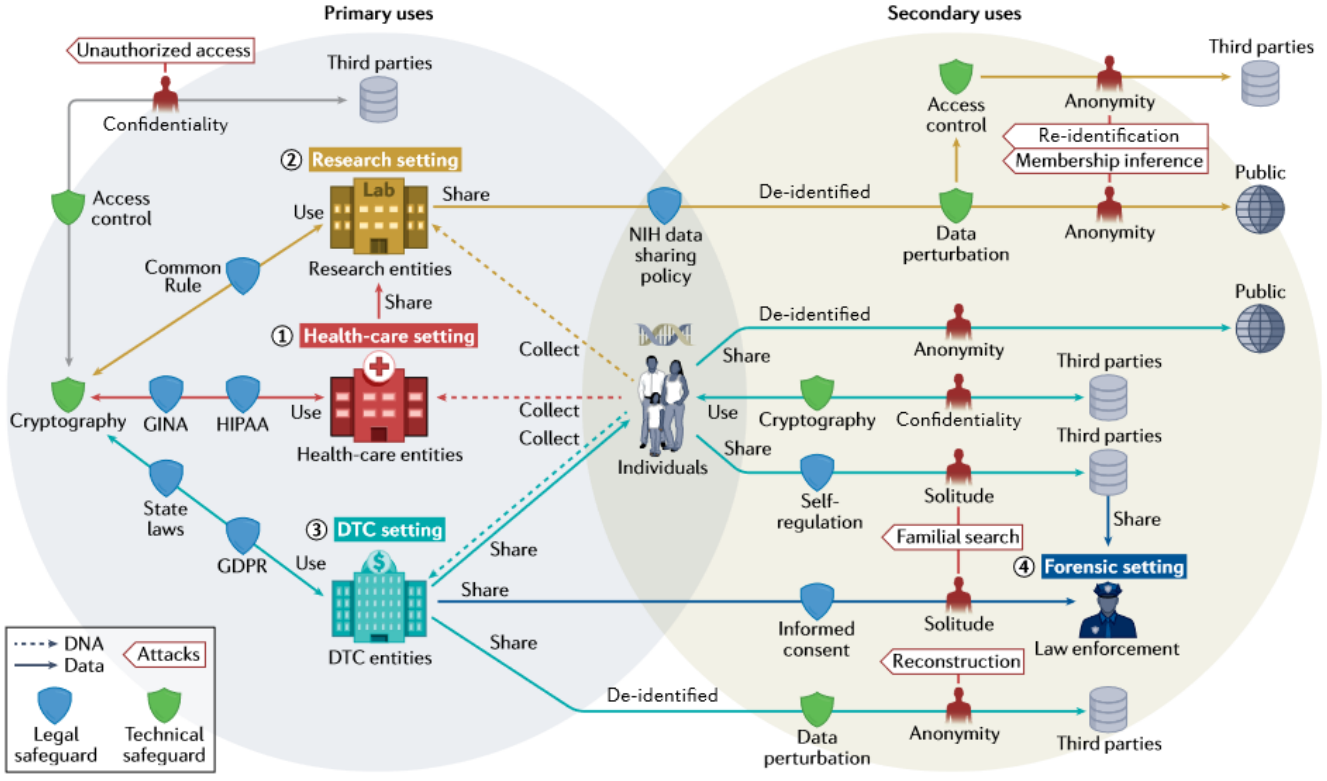


Fig. 1 | **An overview of privacy intrusions and safeguards in genomic data flows.** The four routes of genomic data flow (as indicated by the arrow colours) represent four settings in which data are used or shared: health care (red), research (gold), direct-to-consumer (DTC; green) and forensic (dark blue). The grey line represents a combination of the first three settings. In the health-care setting, data collected by a health-care entity (for example, Vanderbilt University Medical Center) are protected by the Genetic Information Nondiscrimination Act of 2008 (GINA)[128] and the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[116,117] for primary uses. In the research setting, data collected by a research entity (for example, 1000 Genomes Project, Electronic Medical Records and Genomics (eMERGE) network or All of Us Research Program) are primarily protected by the Common Rule[14,124] for primary uses and protected by the US National Institutes of Health (NIH) data sharing policy[37,38] for secondary uses. In the DTC setting, data collected by a DTC entity are protected by the European Union's General Data Protection Regulation (GDPR)[12] and/or the US state privacy laws (for example, California Consumer Privacy Act[130], California Privacy Rights Act[131] or Virginia Consumer Data Protection Act[132]) for primary uses and protected by self-regulation (for example, data use agreements[36], privacy policies[173] or terms of service[174]) for secondary uses. In the forensic setting, data shared with law enforcement are protected by informed consent[192]. A first party refers to the individual to whom the data correspond, whereas a second party refers to the organization (or individual) who collects and/or uses the data for a purpose that the first party is made aware of. By contrast, third parties refer to users (or recipients) of data who have the ability to communicate with the second party only and might include malicious attackers. Examples of third parties include researchers who access data from an existing research study or a pharmaceutical company that partners with a DTC genetic testing company. The data flow from a DTC entity to a research entity is represented by the arrow at the bottom. Confidentiality is mostly concerned when data are being used, whereas anonymity and solitude are mostly concerned when data are being shared. Specifically, cryptographic tools[31] protect confidentiality against unauthorized access attacks, whereas access control[27] and data perturbation approaches[83] protect anonymity against privacy intrusions such as re-identification and membership inference attacks. We simplify the figure by omitting the impacts of GDPR and data use agreements in the research setting.

An overview of privacy intrusions and safeguards in genomic data flows in Wan et al. (2022)

Privacy is a requirement for a democratic society

- Freedom of speech also entails freedom of thought (“intellectual freedom”)
- Knowing we’re being watched changes our behavior: self-censorship or “chilling effect”
- 2001 PATRIOT Act empowered law enforcement to (for example) more easily compel borrower records from libraries
  - Librarians’ response: sorry, we don’t retain those records!
- Given the big data we all generate, and the precise tools available to many industries such as health insurance and advertising, ought we not be concerned about privacy in all forms?

Why k-anonymity isn't enough

- Data may be published as k-anonymous, but can still be combined with other data to trace or re-identify an individual.
- Sweeney (2015) re-identified 43% of records studied (n = 81) in the preceding newspaper examples and using a state health database (exact, not probabilistic matches); confirmed with newspaper reporter
- k is a *combinatorial* property; it can shrink from 100 to 1 quickly (as separate datasets are joined)
- So… which information is "sensitive"? Potentially all of it! Then how do we protect it?

Differential privacy

Differential privacy

- A guarantee: there are two adjacent databases, D and D'. (Adjacent = same except for removal of one record, yours.) Given the ability to query both databases, an adversary won't be able to determine (1) anything about you, nor (2) definitively determine whether you are a member of D and/or D' .
- Proposed in Dwork et al. (2006)
- Calculates for us the *right* amount of statistical perturbation (noise) to both:
  1. Protect against re-identification and tracing attacks
  2. Keep the same statistical qualities as the original data
- We can identify real-world correlations without jeopardizing any individual's information or identity
- Believed to be future-proof (but Dwork still contends: "anonymized data, isn't").

Differential privacy

> "Perhaps the most surprising property of differential privacy is that, despite its protective strength, it is compatible with meaningful data analysis."

this is because

> "data snoopers are not interested in the population, rather, they are interested in this specific realization of data from the population, namely the database itself, D. . . . In a sense, D is a population, and a data snooper is trying to learn about its fixed unknown parameters"

(Dwork et al. 2017; Matthews and Harel 2012)

What *doesn't* differential privacy protect against?

- Information the adversary knows from *outside* of the dataset
  - Example: Cancer study data won't allow the adversary to see Bob's information (e.g., whether he is a smoker), nor whether he's even *in* the study. But they might find out that Bob is a smoker through other means.
  - However, because of DP perturbation, the adversary will not be able to join/cross-reference the study data with whatever they know about Bob.
- Application of inferences that an analyst generates from the dataset
  - If the data in the previous example show a correlation between smoking and incidence of lung cancer, the privacy regime does *not* prevent the adversary inferring that Bob is at a higher risk of lung cancer.

Approaches to differential privacy

We can think about the topic of differential privacy in some different ways:

- What kind of risks are faced by the user? How does the proposed mechanism help?
- Is there a trusted curator of the data?
  - *central model* of DP: yes, there is a trusted curator with access to true data, who receives queries and serves results to an untrusted analyst
  - *local model* of DP: data nodes don't trust anyone; DP is applied by the data creator before uploading to untrusted curator

Approaches to differential privacy

We can think about the topic of differential privacy in some different ways:

- What kind of risks are faced by the user? How does the proposed mechanism help?
- Is there a trusted curator of the data?
- Which data are perturbed?
  - *input* techniques: modify underlying data $\rightarrow$ query/compute on faked data (local model)
  - *output* techniques: query/compute on real data $\rightarrow$ modify results before serving to analyst (central model)

Approaches to differential privacy

We can think about the topic of differential privacy in some different ways:

- What kind of risks are faced by the user? How does the proposed mechanism help?
- Is there a trusted curator of the data?
- Which data are perturbed?
- Which perturbation mechanism is deployed: Gaussian, Laplace, exponential…
- What is the structure of the data (tabular, social network, imagery…), and how to comply with specific sociocultural requirements (e.g., protection against facial recognition adversaries)
- For a non-summarized data release, how to apply DP principles in order to synthesize fake data (which are nevertheless statistically valid samples of the population)

Properties of differential privacy

- Sequential composition: “the ϵs add up”
  - $\epsilon =$ our privacy budget, determined by the curator (the symbol is a lowercase epsilon)
  - Sequential composition “bounds the total privacy cost of releasing multiple results of differentially private mechanisms on the same input data.” (Near and Abuah 2021)
  - More queries on a dataset $\rightarrow$ less privacy, as represented by the budget adding up
  - Given two mechanisms with budgets of 1 and 2 respectively, applying them sequentially results in a dataset with a total budget of 3: ϵ-1 + ϵ-2 = ϵ-3
  - Each mechanism enforces its ϵ mathematically
  - Compare with k-anonymity, which changes multiplicatively—differential privacy is a stronger guarantee

Properties of differential privacy

- Sequential composition: “the ϵs add up”
- Parallel composition: chunking the dataset, each with its own DP mechanism
- Post-processing: if a dataset is differentially private, the mechanism cannot be reversed through any kind of post-processing.
  - Applying a noise reduction algorithm on a differentially private dataset will *not* be able to resolve the data to its true (i.e., vulnerable) form.

Some implementations of differential privacy

- As of the 2020 Census, the U.S. Census is the first anywhere to publish differentially private data.
- Apple: Siri in iOS
- Microsoft: Windows telemetry

Mathematical formalism of differential privacy

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{A}(D_2) \in S]$$

Hands-on: examples from *Programming Differential Privacy*

A nice figure: https://www.nature.com/articles/s41576-022-00455-y/figures/2

Closing thoughts

What is sensitive or vulnerable is not always obvious.

Who is the adversary?

- Scenario: Suppose you've used a social media site for some time, and it changes policies and/or ownership. Now, suppose you no longer trust the company that owns the site. Wouldn't you hope that the site's developers, data architects, and data engineers had designed their platform to differentially privatize your data?

When working with data, we should cultivate and maintain a critical awareness of the potential power embedded in our sociotechnical systems—we owe it to our fellow members of society.

- The adversary may not be a hacker, but an unscrupulous coworker, your employer, or your government.

References

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. “Calibrating Noise to Sensitivity in Private Data Analysis.” In, edited by Shai Halevi and Tal Rabin, 265–84. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11681878_14.

Dwork, Cynthia, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. “Exposed! A Survey of Attacks on Private Data.” *Annual Review of Statistics and Its Application* 4 (1): 61–84. https://doi.org/10.1146/annurev-statistics-060116-054123.

Erlich, Yaniv, and Arvind Narayanan. 2014. “Routes for Breaching and Protecting Genetic Privacy.” *Nature Reviews Genetics* 15 (6): 409–21. https://doi.org/10.1038/nrg3723.

Malin, Bradley A. 2005. “An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future.” *Journal of the American Medical Informatics Association* 12 (1): 28–34. https://doi.org/10.1197/jamia.M1603.

Matthews, Gregory J., and Ofer Harel. 2012. “Assessing the Privacy of Randomized Vector-Valued Queries to a Database Using the Area Under the Receiver Operating Characteristic Curve.” *Health Services and Outcomes Research Methodology* 12 (2): 141–55. https://doi.org/10.1007/s10742-012-0093-y.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. “2008 IEEE Symposium on Security and Privacy (Sp 2008).” In, 111–25. https://doi.org/10.1109/SP.2008.33.

Near, Joseph P., and Chiké Abuah. 2021. *Programming Differential Privacy*. https://programming-dp.com/.

Samarati, Pierangela, and Latanya Sweeney. 1998. “Protecting Privacy When Disclosing Information: K-Anonymity and Its Enforcement Through Generalization and Suppression,” 19. https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf.

Sweeney, Latanya. 2015. “Only You, Your Doctor, and Many Others May Know.” *Technology Science*. https://techscience.org/a/2015092903/.

Wan, Zhiyu, James W. Hazel, Ellen Wright Clayton, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley A. Malin. 2022. “Sociotechnical Safeguards for Genomic Data Privacy.” *Nature Reviews Genetics* 23 (7): 429–45. https://doi.org/10.1038/s41576-022-00455-y.