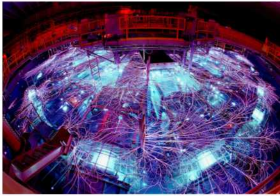
Exceptional service in the national interest

Sandia National Laboratories

SAND2020-7484PE



What is Differential Privacy and Why Should I Care?

Evercita C. Eugenio

July 29, 2020

U.S. DEPARTMENT OF

ratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia,
neywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract

Motivation

Motivation

Sandia National Laboratories

Motivation


e-health
Search
Medical Record
A-F
G-L
M-R
T-Z
HN
Home
Schedule
Health plan
Appointment


Strongly agree
Agree
Disagree
disagree

Motivation


e-health
Search
Medical Record
A-F
G-L
M-R
T-Z
Home
Schedule
Health plan
Appointment


Strongly agree
Agree
Disagree
y disagree

Motivation


e-health
Search
Medical Record
A-F
G-L
M-R
T-Z
HN
Home
Schedule
Health plan
Appointment


Strongly agree
Agree
Disagree
disagree

Motivation


e-health
Search
Medical Record
A-F
G-L
M-R
T-Z
Home
Schedule
Health plan
Appointment


Strongly agree
Agree
Disagree


VOTER REGISTRATION

Motivation

Motivation

Horror stories about data privacy breaches include:

Motivation

Horror stories about data privacy breaches include:

Motivation

Horror stories about data privacy breaches include:

Motivation

Horror stories about data privacy breaches include:

Motivation

Horror stories about data privacy breaches include:

Motivation

Sandia National Laboratories

- While privacy breaches continue to be a huge concern, it has not stopped the requests for data sharing or data release.

Motivation

- While privacy breaches continue to be a huge concern, it has not stopped the requests for data sharing or data release.
- Government agencies, business, survey and research organizations, and medical institutions are constantly being asked to release and share more and more of their data for transparency and accountability.

Motivation

Sandia National Laboratories

- While privacy breaches continue to be a huge concern, it has not stopped the requests for data sharing or data release.
- Government agencies, business, survey and research organizations, and medical institutions are constantly being asked to release and share more and more of their data for transparency and accountability.
- So handling all this data in a way that protects the confidentiality of the data subjects' identities and sensitive attributes while maintaining the statistical usability/accuracy of the data set has developed into a critical area of study.

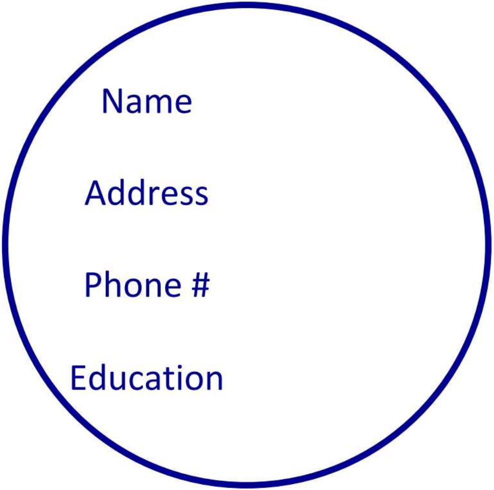
Motivation

Sandia National Laboratories

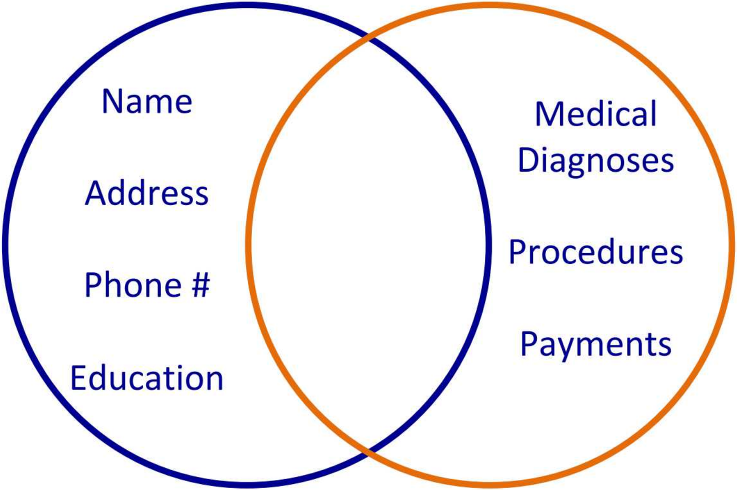
- While privacy breaches continue to be a huge concern, it has not stopped the requests for data sharing or data release.
- Government agencies, business, survey and research organizations, and medical institutions are constantly being asked to release and share more and more of their data for transparency and accountability.
- So handling all this data in a way that protects the confidentiality of the data subjects' identities and sensitive attributes while maintaining the statistical usability/accuracy of the data set has developed into a critical area of study.
- One of the most common ways to "protect" data is to simply anonymize the data (i.e. remove identifying information or sensitive characteristics).

Motivation: Record Linkage

Motivation: Record Linkage

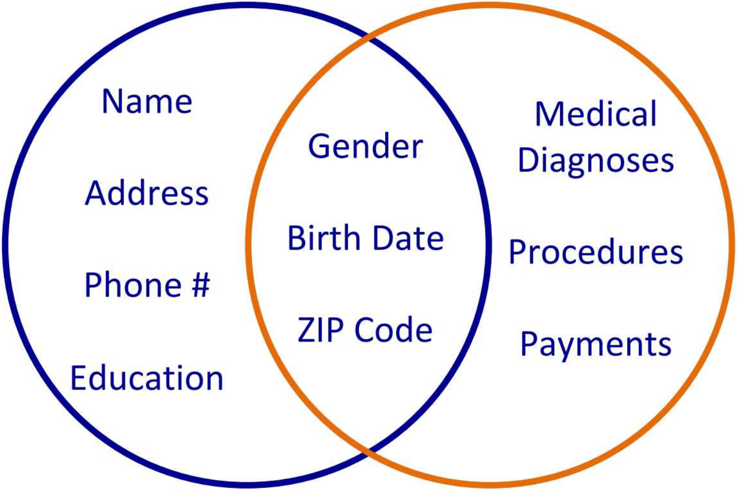


Motivation: Record Linkage



Motivation: Record Linkage

Sandia National Laboratories



Motivation

- The old ways of anonymizing data do not always work!

Motivation

- The old ways of anonymizing data do not always work!
- It has been shown that anonymized data can be linked with other publicly available datasets. This record linkage can possibly lead to re-identification.

Motivation

- The old ways of anonymizing data do not always work!
- It has been shown that anonymized data can be linked with other publicly available datasets. This record linkage can possibly lead to re-identification.
- So instead of using the old techniques of anonymization, new privacy techniques have been formulated in the field of differential privacy.

What is differential privacy?

Differential Privacy

Differential privacy ensures that the addition or removal of a single database item does not substantially affect the outcome of any analysis.

Queries in Differential Privacy

- Queries are defined to be the questions of interest with regards to a data set or a database.

Queries in Differential Privacy

- Queries are defined to be the questions of interest with regards to a data set or a database.
- A data curator manages queries sent to a data set.

Queries in Differential Privacy

- Queries are defined to be the questions of interest with regards to a data set or a database.
- A data curator manages queries sent to a data set.
- The curator can provide a response either using differential privacy or not.

Differential Privacy (Dwork 2006)

Let $\mathcal{K}$ be a mechanism (to be defined later), and let D_1 and D_2 be two databases that differ in at most one element. A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$e^{-\epsilon} \leq \frac{Pr[\mathcal{K}(D_1) \in S]}{Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

Differential Privacy (Dwork 2006)

Let $\mathcal{K}$ be a mechanism (to be defined later), and let D_1 and D_2 be two databases that differ in at most one element. A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(\mathcal{K})$,

$$e^{-\epsilon} \leq \frac{Pr[\mathcal{K}(D_1) \in S]}{Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

The total privacy budget for a set of queries is ϵ, based on the definition of ϵ-differential privacy.

Differential Privacy

$$e^{-\epsilon} \leq \frac{Pr[\mathcal{K}(D_1) \in S]}{Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

$\mathcal{E} =$ Amount of Privacy Used
or
Information Leaked

Differential Privacy

$$e^{-\epsilon} \leq \frac{Pr[\mathcal{K}(D_1) \in S]}{Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

$\varepsilon \uparrow$

Leak MORE information

Differential Privacy

$$e^{-\epsilon} \leq \frac{Pr[\mathcal{K}(D_1) \in S]}{Pr[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

ε ↓

Leak LESS information

Differential Privacy Example: Notre Dame students

Differential Privacy Example: Notre Dame students

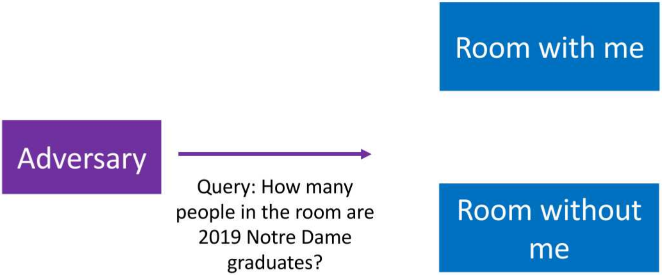
Adversary

Differential Privacy Example: Notre Dame students

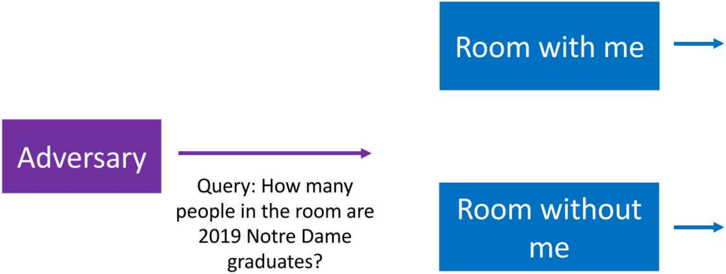
Adversary

Query: How many people in the room are 2019 Notre Dame graduates?

Differential Privacy Example: Notre Dame students



Differential Privacy Example: Notre Dame students



Differential Privacy Example: Notre Dame students

Adversary → Query: How many people in the room are 2019 Notre Dame graduates?

Room with me → 1

Room without me → 0

Can deduce that I am the only one who is a 2019 Notre Dame graduate.

Differential Privacy Example: Notre Dame students

Room with me → ~~1~~

Adversary →

Query: How many people in the room are 2019 Notre Dame graduates?

Room without me → ~~0~~

Under Differential Privacy should not be able to distinguish between results.

Differential Privacy Example: Notre Dame students

Room with me

Adversary

Query: How many people in the room are 2019 Notre Dame graduates?

Add noise under ϵ-differential privacy

Room without me

Under Differential Privacy should not be able to distinguish between results.

Differential Privacy Example: Notre Dame students

Room with me

Adversary

Query: How many people in the room are 2019 Notre Dame graduates?

1

Room without me

Under Differential Privacy should not be able to distinguish between results.

Differential Privacy Example: total income

Differential Privacy Example: total income

Database for South Bend, IN
with ME

ID	Income
1	300K
2	50K
3	21K
4	32K
5	150K
...	
100,000	78K

Total income = *f(X)*

Differential Privacy Example: total income

Database for South Bend, IN with ME

ID	Income
1	300K
2	50K
3	21K
4	32K
5	150K
...	
100,000	78K

Total income = *f(X)*

Differential Privacy Example: total income

Database for South Bend, IN
with ME

ID	Income
1	300K
2	50K
3	21K
4	32K
5	150K
…	
100,000	78K

Total income = $f(X)$

Database for South Bend, IN
without ME

ID	Income
1	300K
2	50K
4	32K
5	150K
…	
100,000	78K

Total income = $f(X_{-3})$

Differential Privacy

- If the curator answered **WITHOUT** any differential privacy, and if you knew that I was going to move to another area, then simply querying this database before and after my move would allow you to deduce my income.

Differential Privacy

- If the curator answered **WITHOUT** any differential privacy, and if you knew that I was going to move to another area, then simply querying this database before and after my move would allow you to deduce my income.
- If the curator answered **WITH** differential privacy, then the risk to my privacy would not be substantially (as bounded by ϵ) increased as a result of participating in the statistical database.

Differential Privacy

- If the curator answered **WITHOUT** any differential privacy, and if you knew that I was going to move to another area, then simply querying this database before and after my move would allow you to deduce my income.
- If the curator answered **WITH** differential privacy, then the risk to my privacy would not be substantially (as bounded by ϵ) increased as a result of participating in the statistical database.
- To apply differential privacy, we can add some noise using the mechanism $\mathcal{K}$ to the result of a query on our dataset to ensure the formula for ϵ-differential privacy holds.

Noise in Differential Privacy

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.
- Also, many noise adding mechanisms exist for specific statistical analyses. They include

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.
- Also, many noise adding mechanisms exist for specific statistical analyses. They include
  - contingency tables,

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.
- Also, many noise adding mechanisms exist for specific statistical analyses. They include
  - contingency tables,
  - principal component analysis,

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.
- Also, many noise adding mechanisms exist for specific statistical analyses. They include
  - contingency tables,
  - principal component analysis,
  - location privacy, and

Noise in Differential Privacy

- Differential privacy requires some noise to be added to queries of interest in order to protect privacy.
- Many noise adding mechanisms exist for general query release. They include
  - Laplace mechanism,
  - Exponential mechanism,
  - Gaussian mechanism, and
  - median mechanism.
- Also, many noise adding mechanisms exist for specific statistical analyses. They include
  - contingency tables,
  - principal component analysis,
  - location privacy, and
  - graphs and social networks.

Laplace Mechanism

Laplace Mechanism

Sensitivity (Dwork 2006)

For $f : D \to R^k$, the sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

for all D_1, D_2 differing in at most one element.

Laplace Mechanism

Sensitivity (Dwork 2006)

For $f : D \rightarrow R^k$, the sensitivity of f is

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

for all D_1, D_2 differing in at most one element.

Captures how large a difference between the value of f on two databases (differing in a single element) must be hidden by the additive noise generated by the curator.

Laplace Mechanism (Dwork 2006)

When the query is numeric, adding Laplace random noise independently to each of the components of $f(X)$ guarantees ϵ-differential privacy.

$$f(X) + Lap(\Delta f/\epsilon)$$

Laplace Mechanism (Dwork 2006)

When the query is numeric, adding Laplace random noise independently to each of the components of $f(X)$ guarantees ϵ-differential privacy.

$$f(X) + Lap(\Delta f/\epsilon)$$

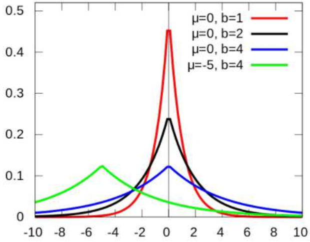


Image credit: `https://en.wikipedia.org/w/index.php?title=Laplace_distribution&oldid=697827002`

Privacy Budget

Privacy Budget

The total privacy budget for a set of queries is ϵ, based on the definition of ϵ-differential privacy.

Privacy Budget

The total privacy budget for a set of queries is ϵ, based on the definition of ϵ-differential privacy.

Types of Composition (McSherry 2009)

Privacy Budget

The total privacy budget for a set of queries is ϵ, based on the definition of ϵ-differential privacy.

Types of Composition (McSherry 2009)

- Sequential Composition: Each query uses δ_i privacy, so for each query the portion of the total privacy budget used is $\frac{\delta_i}{\epsilon}$. This places a restriction on the number of queries that can be asked or the amount of privacy that the query actually has.

Privacy Budget

The total privacy budget for a set of queries is ϵ, based on the definition of ϵ-differential privacy.

Types of Composition (McSherry 2009)

- Sequential Composition: Each query uses δ_i privacy, so for each query the portion of the total privacy budget used is $\frac{\delta_i}{\epsilon}$. This places a restriction on the number of queries that can be asked or the amount of privacy that the query actually has.
- Parallel Composition: If D_i are disjoint subsets of the original database and M_i provides differential privacy for each D_i, then the sequence of M_i provides differential privacy. The ultimate privacy guarantee only depends on the worst of the guarantees of each analysis, not the sum.

Beyond Queries

Beyond Queries

- Queries allow one to answer very specific questions about the data set.

Beyond Queries

- Queries allow one to answer very specific questions about the data set.
- What happens when we want to do more with the data?

Beyond Queries

- Queries allow one to answer very specific questions about the data set.
- What happens when we want to do more with the data?
- Can perform differentially private data synthesis, which yields differentially private synthetic data sets.

Why should I care about differential privacy?

Who uses differential privacy?

Who uses differential privacy?

Google

Who uses differential privacy?

Google

Microsoft

Who uses differential privacy?

Google

Microsoft

Who else uses differential privacy?

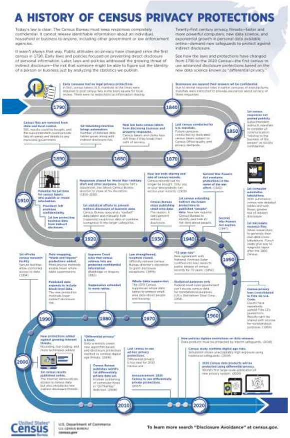
Who else uses differential privacy?

Uber

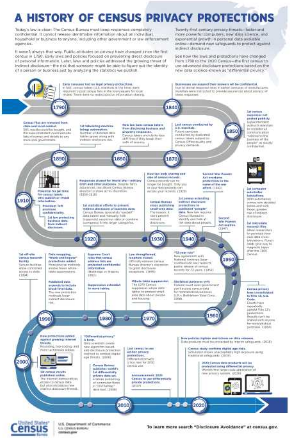
Who else uses differential privacy?

What about outside the tech world?

What about outside the tech world?



What about outside the tech world?



What about outside the tech world?

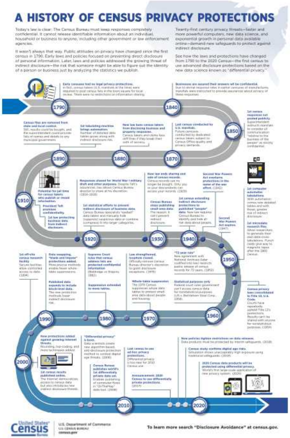


Image credit:
`https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html`

What else can differential privacy be used for?

What else can differential privacy be used for?

- Contact tracing which can include mobility reports or movement patterns.

What else can differential privacy be used for?

- Contact tracing which can include mobility reports or movement patterns.
- Mobile data sharing.

What else can differential privacy be used for?

- Contact tracing which can include mobility reports or movement patterns.
- Mobile data sharing.
- Models and model parameters being shared across boundaries (federated learning).

Ongoing Research in Differential Privacy

Ongoing Research in Differential Privacy

- Ensuring that differential privacy is understood by the general public.

Ongoing Research in Differential Privacy

- Ensuring that differential privacy is understood by the general public.
- Open-sourcing tools to allow researchers to apply differential privacy themselves (e.g. Google DP Library, IBM Research DP Library or Microsoft/Havard's OpenDP).

Ongoing Research in Differential Privacy

- Ensuring that differential privacy is understood by the general public.
- Open-sourcing tools to allow researchers to apply differential privacy themselves (e.g. Google DP Library, IBM Research DP Library or Microsoft/Havard's OpenDP).
- Differentially private mechanism design for other statistical analyses or model types.

Ongoing Research in Differential Privacy

- Ensuring that differential privacy is understood by the general public.
- Open-sourcing tools to allow researchers to apply differential privacy themselves (e.g. Google DP Library, IBM Research DP Library or Microsoft/Havard's OpenDP).
- Differentially private mechanism design for other statistical analyses or model types.
- Applying differential privacy to new applications areas.

Questions?

References I

Bowen, C. and Liu, F. (2016), Differential Private Data Synthesis Methods (arXiv:1602.01063v1), Journal of the Royal Statistical Society: Series A (invited revision).

A History of Census Privacy. From `https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html`

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265-284, Berlin, Heidelberg. Springer-Verlag.

Dwork, C. (2008). Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC'08, pages 1-19, Berlin, Heidelberg. Springer-Verlag.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). Generalization in adaptive data analysis and holdout reuse. *CoRR*, abs/1506.02629.

Eugenio, E. C., and Liu, F. (2018). Cipher: Construction of differentially private microdata from low-dimensional histograms via solving linear equations with tikhonov regularization. arXiv preprint arXiv:1812.05671.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94-103, Washington, DC, USA. IEEE Computer Society.

McSherry, F. D. (2009). Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 19-30, New York, NY, USA. ACM.