Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality

Imane EL Missaoui
Nene Sidibe Bakary
Elnaz Sherafat
Ezechiel Djohi

Pr. Benjamin GUINHOUYA

24/04/2024

M2 Data science en santé

Plan

1. Introduction
2. Confidentialité différentielle (Differential privacy)
3. Cas d’application : Projet Crisper
4. Apports
5. Limitations et défis
6. Conclusion et perspectives

“Human intuition about what is private is not especially good. Computers are getting more and more sophisticated at pulling individual data out of things that a naive person might think are harmless.”

— Frank McSherry, Co-inventor of differential privacy,
Co-founder and Chief Scientist, Materialize, Inc.

“Data is the pollution problem of the information age, and protecting privacy is the environmental challenge.”

– Bruce Schneier

Contexte : Partage d'informations

IMPORTANCE DU PARTAGE DES DONNÉES

Soins

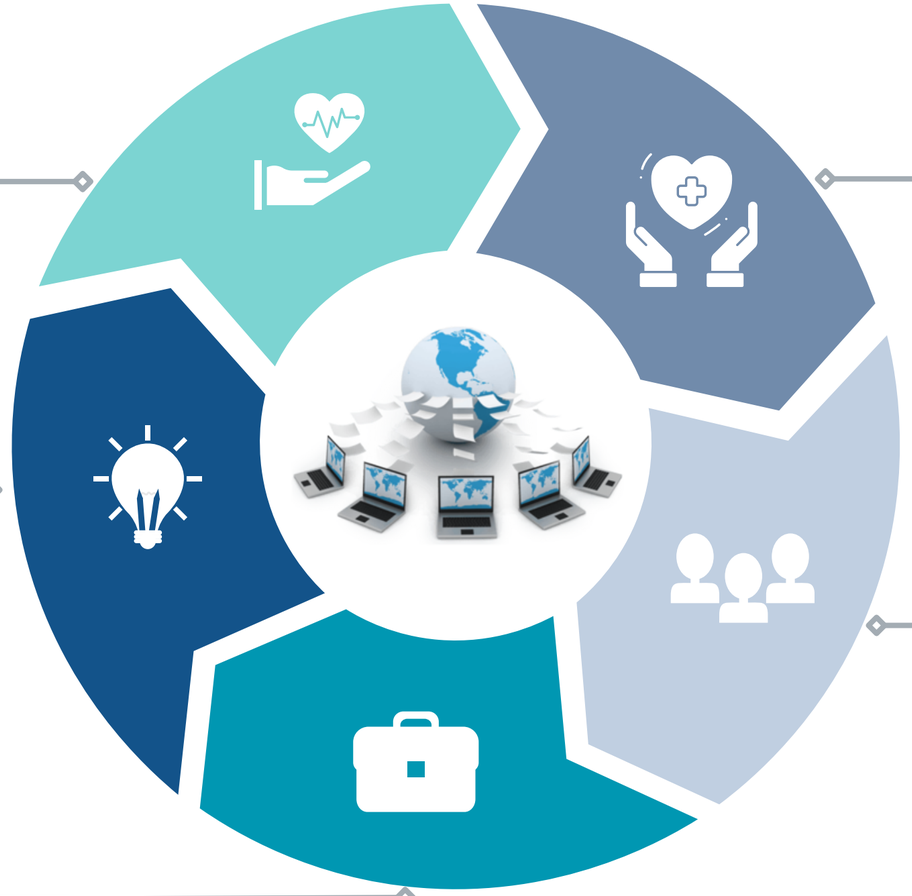
réduire les erreurs médicales, améliorer la gestion des dossiers médicaux, la coordination des soins, et la communication entre les professionnels de santé

R&D

mener des études, effectuer des analyses et développer de nouvelles solutions ou innovations dans différents domaines

Formation

former les étudiants et les professionnels, leur permettant d'acquérir de nouvelles compétences et connaissances.



Santé publique

surveiller les épidémies, les maladies infectieuses, et d'autres problèmes de santé publique

Innovation et Collaboration

encourager l'innovation en permettant aux différentes parties prenantes de collaborer, partager des idées et développer de nouvelles solutions ensemble.

Contexte : Protection des données

Protection de la vie privée

Importance

1 **Confiance du Public**

Favorise une participation volontaire et une adhésion aux programmes de santé publique

2 **Respect des Droits des Individus**

Respect des droits de confidentialité des informations personnelles

3 **Prévention de la Stigmatisation et de la Discrimination**

Utilisation inappropriée des données sensibles (état de santé, antécédents médicaux ou caractéristiques génétiques)

4 **Conformité aux Réglementations et Normes**

Eviter les sanctions, amendes et conséquences juridiques associées à la violation de la vie privée des données de santé

Contexte : Protection des données

Protection de la vie privée

Mesures actuelles

1. **De-identification**
   Suppression des identifiants personnels et remplacement par des identifiants

2. **Agrégation**
   Agrégation des données individuelles en résumés statistiques

3. **Authentification**
   Vérification de l'identité numérique
   Choix de mots de passe robustes et les changer fréquemment

4. **Chiffrement**
   Utilisation d’algorithmes pour crypter et décrypter les données par des personnes autorisées

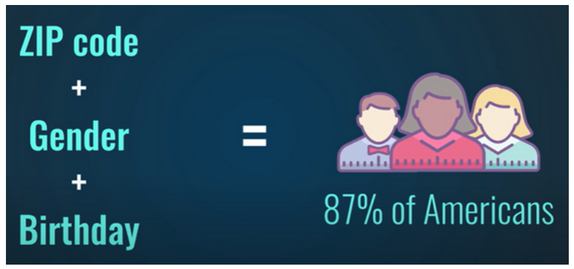
Contexte : Linkage attack

Linkage attack

Massachusetts

Sa commission d'assurance collective a choisi de rendre publiques les visites hospitalières des fonctionnaires de l'État

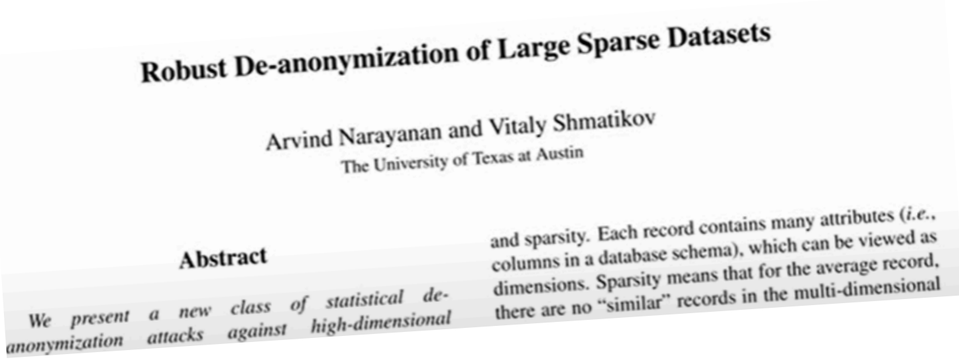
Données médicales publiques  **Données de votes**



Compétition : NETFLIX Prize (2006)

480K users
100M notes
17K films

Anonymisation des données



Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional

and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no "similar" records in the multi-dimensional

Contexte : Cyberattaque

Cyber attack

11 hôpitaux attaqués en 2022

Le plus médiatisé

Centre hospitalier de Corbeil-Essonnes

rançon de 1 million d'euros

Cartographie des cyberattaques en france

2022

La liste des hôpitaux touchés par une cyberattaque en 2022

Le plus récent

Cyberattaque mardi 16 avril à l'hôpital de Cannes

La fuite concernerait une base de données de plus de "900 000 patients avec leurs adresses, leurs numéros de téléphone, leurs antécédents médicaux…"

11 février 2024

Contexte : Confidentialité Différentielle

CONFIDENTIALITÉ DIFFÉRENTIELLE

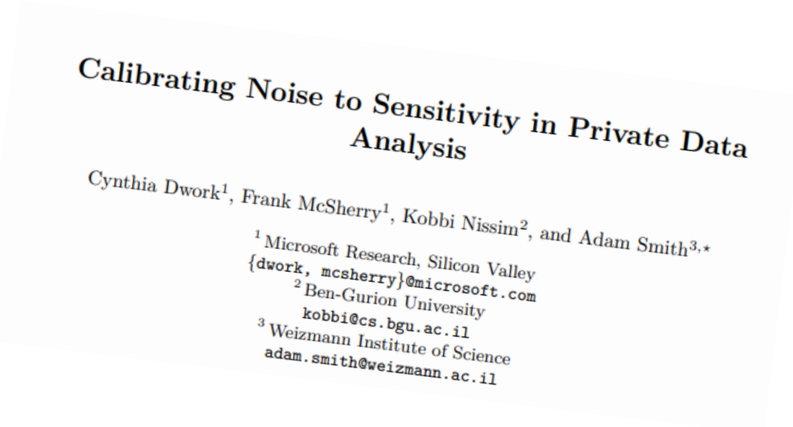
Approche mathématique permettant de collecter des informations sans compromettre la confidentialité des données individuelles

abordé en 2006 par Cynthia Dwork et Frank McSherry, et al. dans deux articles intitulés "Calibrating Noise to Sensitivity in Private Data Analysis" et "Differential Privacy".

Largement utilisé par :

Uber

Microsoft



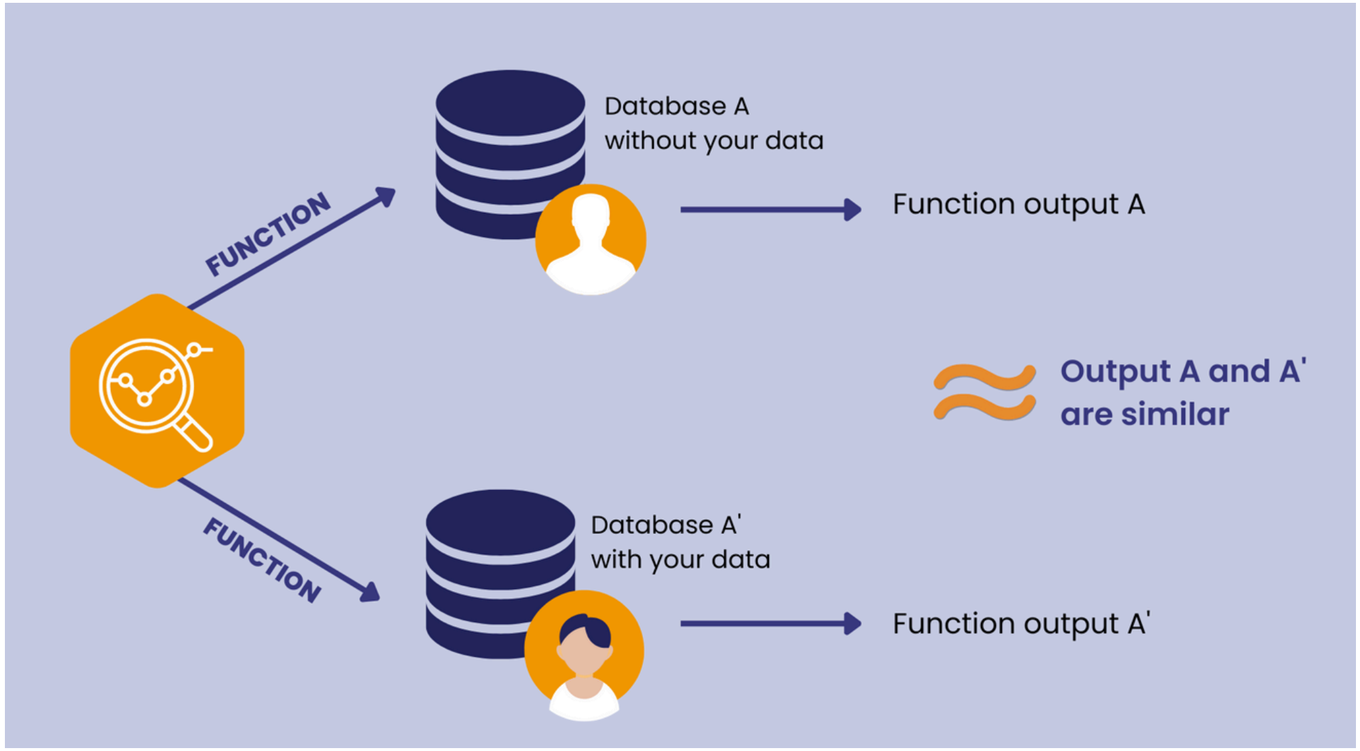
Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3,⋆]

[1] Microsoft Research, Silicon Valley
{dwork, mcsherry}@microsoft.com
[2] Ben-Gurion University
kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science
adam.smith@weizmann.ac.il

Trés utilisé pour les grands jeu de données

Fondements de la confidentialité différentielle “Differential privacy”



Notion de vie privée = risque cumulatif

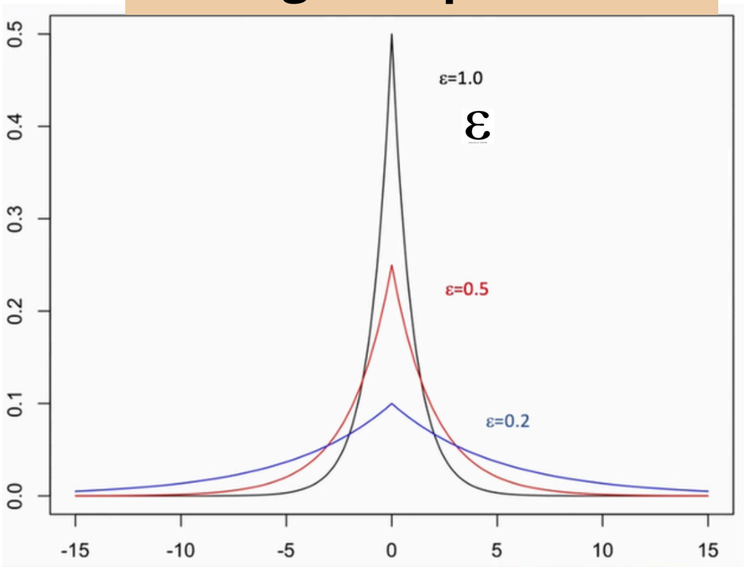
"Mécanisme" = tout calcul pouvant être effectué sur les données.

Ajout du bruit aux données de manière contrôlée de manière à protéger la vie privée des individus tout en permettant une analyse utile des données.

Mécanisme de Laplace

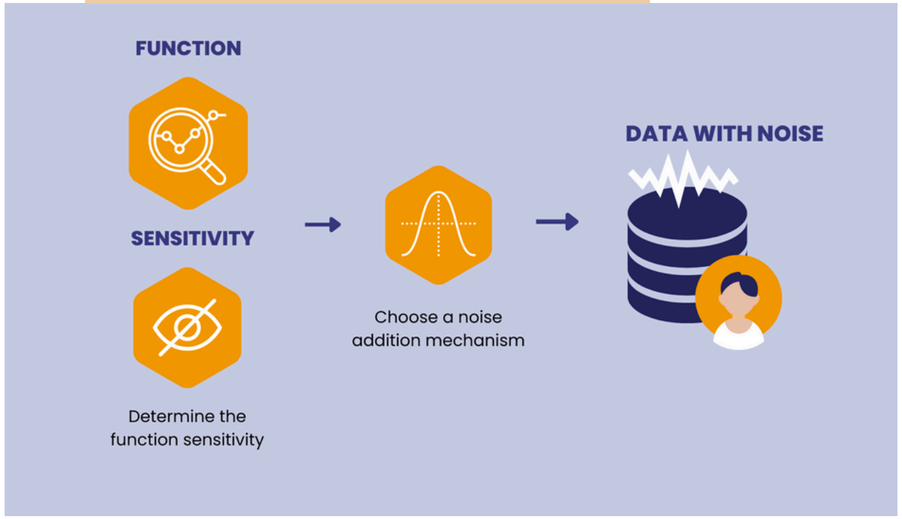
Du bruit est ajouté à la sortie d'une fonction. La quantité de bruit dépend de la sensibilité de la fonction et est tirée d'une distribution de Laplace.

Budget de protection



Budget de protection = degré de confidentialité vs degré de précision

Sensibilité

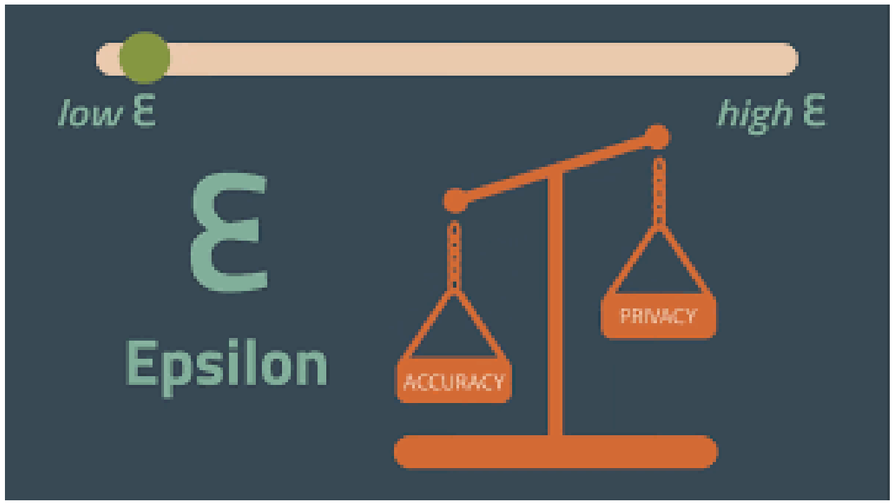


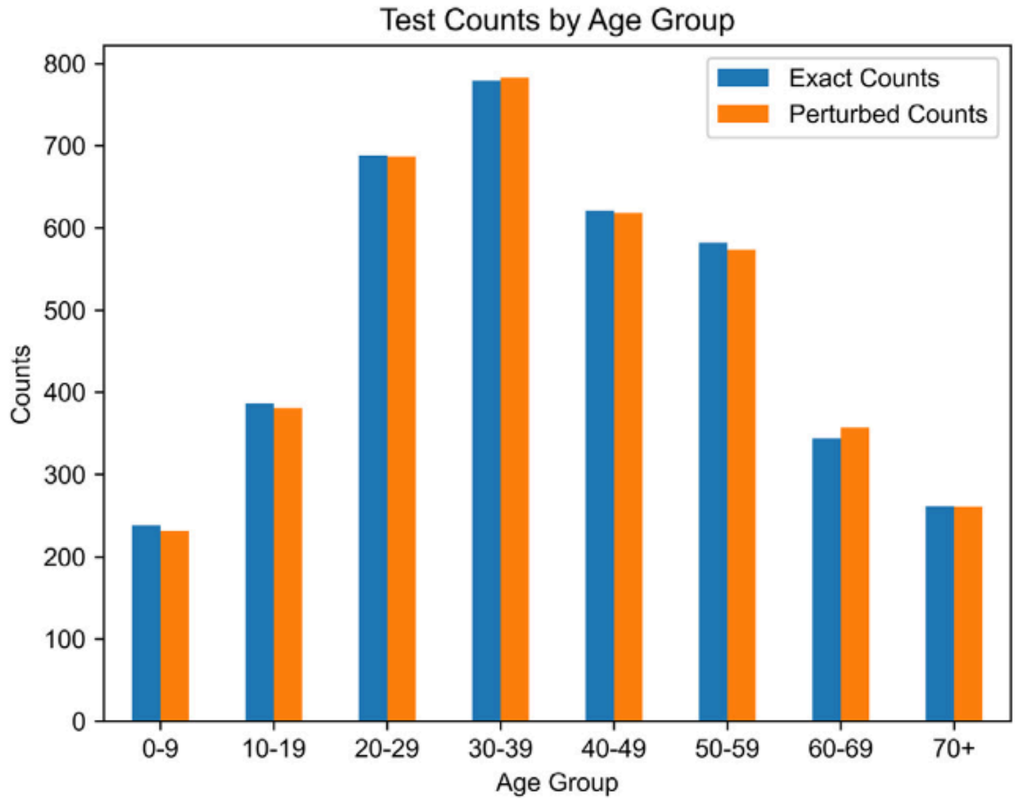
Sensibilité = impact des résultats modifiés

Bruit = Sensibilité / Budget de protection

Plus **la sensibilité** est **grande**, plus il faut **ajouter** du **bruit**

Requêtes par histogramme :





Exemple : base de données COVID-19

- Comporte le groupe d'âge de chaque personne ayant subi un test COVID-19 à une date donnée.
- Une **requête par histogramme** compte le nombre de personnes dans chaque groupe d'âge.
- L'histogramme peut être rendu différentiellement privé en ajoutant un bruit de Laplace indépendant au décompte de chaque valeur possible.

Cas d'utilisation: Projet CRISPER

CRISPER
COVID-19 Real-time Information System for Preparedness and Epidemic Response

Mission

- Un outil de cartographie interactive développé par l'équipe des auteurs en Australie pour visualiser et interagir avec les données COVID-19
- Utilise un algorithme de confidentialité différentiel par le biais d'un moteur de données afin de protéger les données qui ne sont pas accessibles au public.
- Utilise actuellement des racleurs de données et des API pour analyser des données provenant d'un certain nombre de sources publiques telles que les sites web des services de santé.

Autres études

Analysis of Application Examples of Differential Privacy in Deep Learning

Zhidong Shen and Ting Zhong

School of Cyber Science and Engineering, Wuhan University, Wuhan, Hubei, China

Correspondence should be addressed to Zhidong Shen; shenzd@whu.edu.cn

Received 6 May 2021; Accepted 8 September 2021; Published 26 October 2021

Making Differential Privacy Work for Census Data Users

Cory McCartan
Center for Data Science
New York University

Tyler Simko
Department of Government
Harvard University

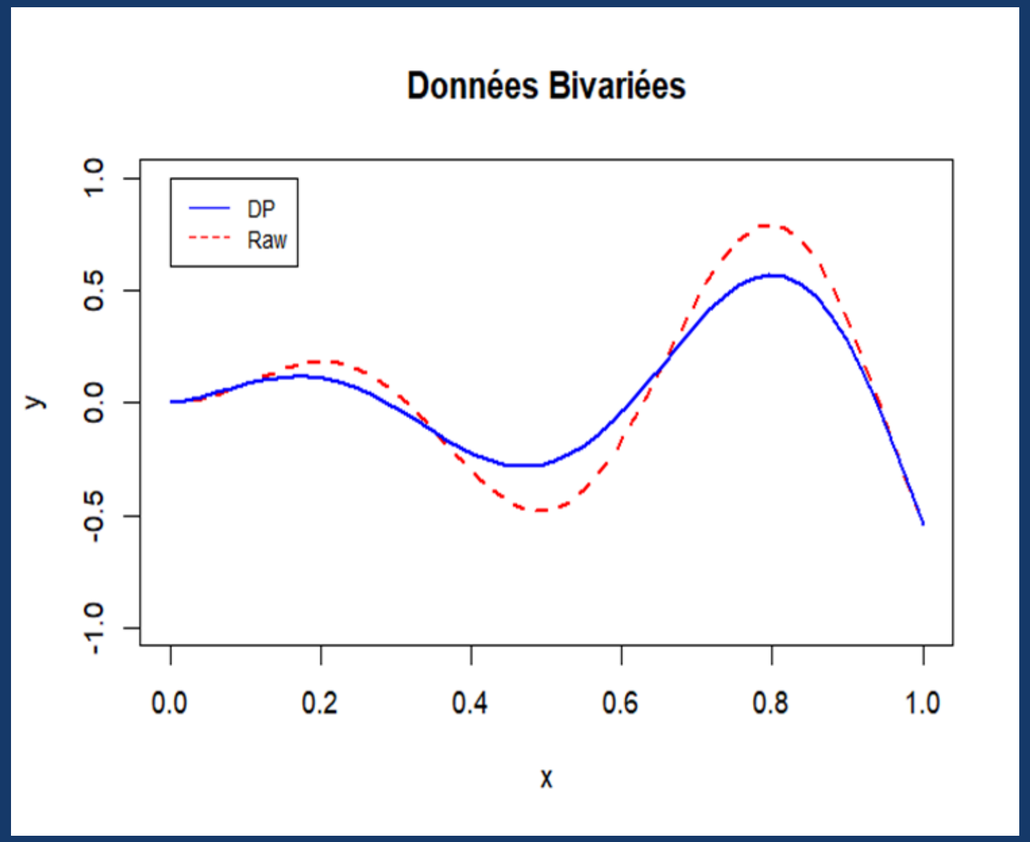
Kosuke Imai*
Department of Government
Department of Statistics
Harvard University

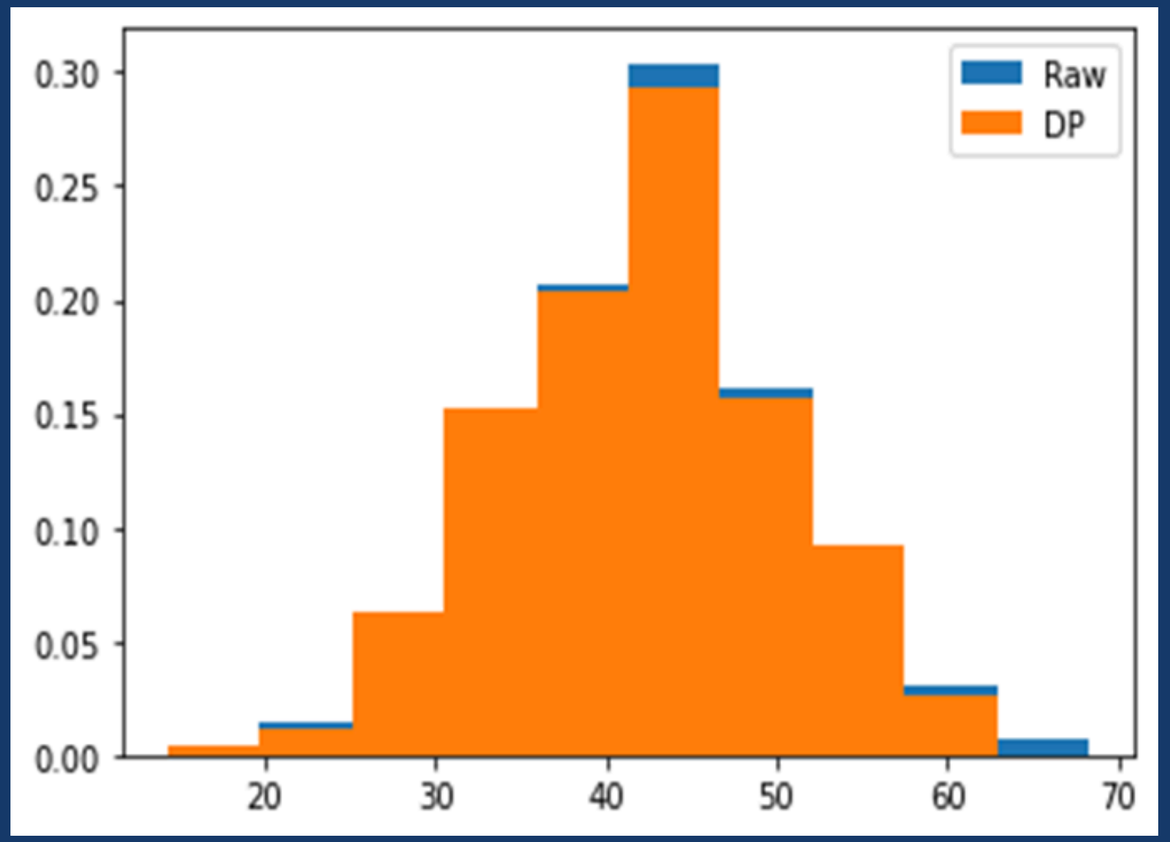
October 7, 2023

Differentially Private Model Publishing for Deep Learning

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, Stacey Truex
School of Computer Science, College of Computing, Georgia Institute of Technology
Email: leiyu@gatech.edu, {ling.liu,calton.pu}@cc.gatech.edu, {memregursoy,staceytruex}@gatech.edu

Application





Xicor = 0.7611

Private mean: 45.577

Raw mean: 45.337

python™ **diffpriv**

diffprivlib

Avantages et Limites

Avantages	Limites
• Partage de données pour la recherche et l'analyse • Facilitation de la recherche sur des groupes sensibles • Garantie une confidentialité • Équilibre entre la vie privée et l'utilité des données	• Bruit et perte d'utilité des données • Mise en place difficile en cas de plusieurs variables et de petit jeu de données • Choix difficile du Privacy budget • Pas adapté à la prise de décision sensible • Limites inhérentes à la protection de la vie privée

Perspectives

- Une méthode prometteuse
- Compétences techniques et une compréhension approfondie
- Adoption de cette méthode dans le domaine de la santé publique

CONCLUSION

- Un outil précieux pour le partage de données
- Permet de protéger la vie privée des individus
- Optimiser son utilisation pratique