Mathematical Methods for Privacy-Preserving Machine Learning

Thomas Strohmer
Center for Data Science and AI Research
Department of Mathematics
University of California, Davis

Texas A&M
October 11, 2021

CeDAR: cedar.ucdavis.edu

Acknowledgements

Research in collaboration with:

Roman Vershynin
UC Irvine

March Boedihardjo
UC Irvine

This work is sponsored by NSF-DMS and
the National Geospatial-Intelligence Agency.

Once we searched Google. Now it searches us

- Google's Street View mapping project scooped up passwords, e-mail and other personal information from people. Google says this was done unintentionally . . .
- Google used secret code to bypass Safari's antitracking security setting
- Google tracks your Android phone even if you turn off location services, stopped using apps and removed your SIM card from your device. Android users cannot opt out.
- Google sells this info to third parties. Eg., if you are near a specific store, that store can send you targeted advertising.
- Android was created to extract our personal information.

Facebook is a serial violator of privacy

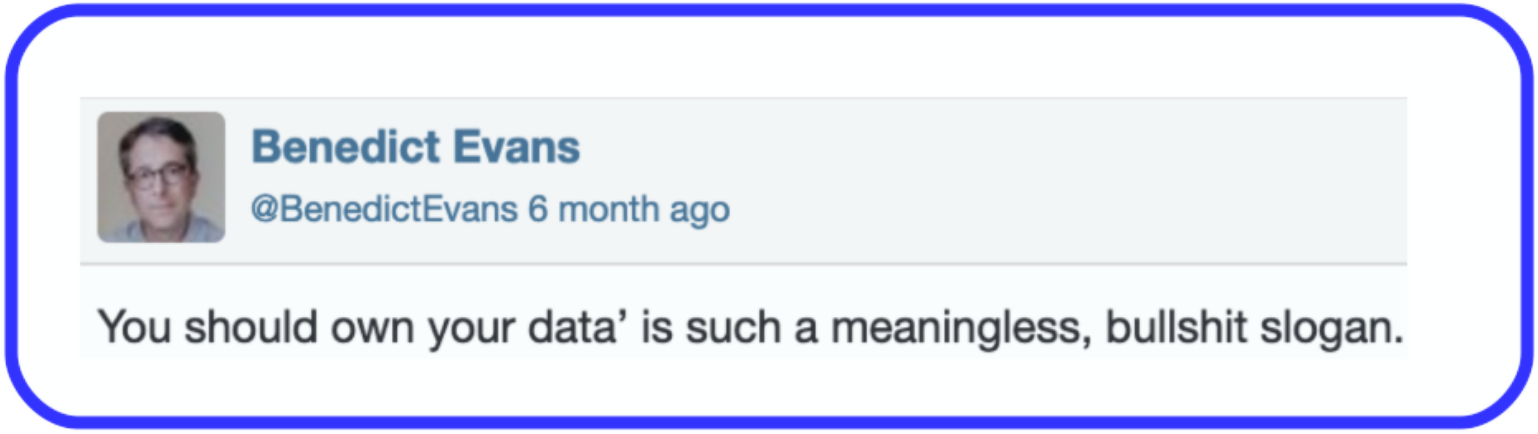
- Without users' consent, Facebook has been sharing their data with more than 150 businesses, including Amazon, Microsoft, Netflix, Spotify, ...
- Facebook pretended to apply Europe's new privacy laws to all its users outside the US and then secretly switched all non-European users to the spineless US privacy law.
- Apps are sharing sensitive data with Facebook without informing users; and even when users are not logged in through Facebook, or do not have a Facebook account.
- Facebook wants banks to hand over their customers' sensitive financial data to offer better service to its users – *give us your data, we give you our users*.

The combined data from all the different apps paint a detailed and intimate picture of people's activities, interests, behaviors.

Companies claim they care about privacy ...

The privacy of our customers' personal information is very important to us ... but not as important as our profits!

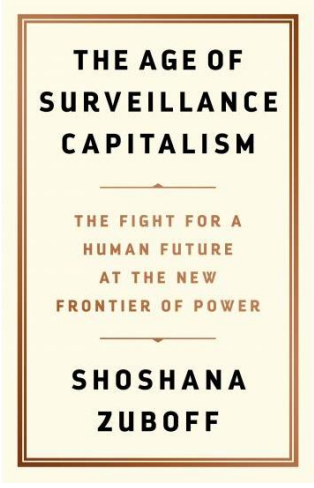
Zuckerberg voted down a shareholder proposal for more accountability and transparency regarding privacy.
Google's Page and Brin have voted down a similar proposal.



Happily retweeted by Facebook's Chief AI Scientist ...

Surveillance Capitalism in a nutshell

Surveillance capitalism: is a new economic system, which pursues the exploitation and control of human nature, thereby threatening our social fabric.



S. Zuboff: By providing free services that billions of people cheerfully use, it enables companies like Google, Facebook, Amazon, ... not only to monitor the behavior of those users in astonishing detail - often without their explicit consent.

Our means of social participation have been conflated with the means through which surveillance capitalists collect their data and seek to modify our behavior.

... with disastrous consequences for democracy and freedom!

Privacy in times of a pandemic

Trusting Google and Facebook with privacy is like trusting Exxon and BP with environmental regulations.

Google

facebook

ExxonMobil

bp

Privacy-preserving machine learning

Privacy-preserving machine learning aims to protect data security, privacy and confidentiality, while still permitting useful conclusions from the data or its use for model development.

Privacy-preserving machine learning plays a key role in democratizing data science and AI

Privacy-preserving techniques:

- Homomorphic encryption
- Federated learning/On-device learning
- Anonymization
- Differential privacy
- Synthetic data

Privacy-preserving machine learning

Privacy-preserving machine learning aims to protect data security, privacy and confidentiality, while still permitting useful conclusions from the data or its use for model development.

Privacy-preserving machine learning plays a key role in democratizing data science and AI

Privacy-preserving techniques:

- Homomorphic encryption
- Federated learning/On-device learning
- **Anonymization**
- **Differential privacy**
- **Synthetic data**

Privacy-preserving techniques

The path to privacy is paved with NP-hard problems!

Anonymization: k-anonymity

Intuition: The information for each person contained in the dataset cannot be distinguished from at least $k-1$ individuals whose information also appear in the dataset. [Sweeney, 2002].

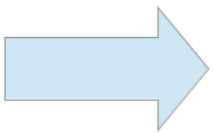
Definition: Let $\mathcal{X}$ denote the input data. An algorithm $\mathcal{A}(\mathcal{X})$ is k-anonymous if the preimage $\mathcal{A}^{-1}(\mathcal{Y})$ of any point $\mathcal{Y}$ under $\mathcal{A}$ has cardinality at least k.

Anonymization: *k*-anonymity

Intuition: The information for each person contained in the dataset cannot be distinguished from at least $k-1$ individuals whose information also appear in the dataset. [Sweeney, 2002].

Definition: Let $\mathcal{X}$ denote the input data. An algorithm $\mathcal{A}(\mathcal{X})$ is *k*-anonymous if the preimage $\mathcal{A}^{-1}(\mathcal{Y})$ of any point $\mathcal{Y}$ under $\mathcal{A}$ has cardinality at least *k*.

first	last	age	race
Harry	Stone	34	Afr-Am
John	Reyser	36	Cauc
Beatrice	Stone	47	Afr-Am
John	Ramos	22	Hisp



$k = 2$

first	last	age	race
*	Stone	30-50	Afr-Am
John	R*	20-40	*
*	Stone	30-50	Afr-Am
John	R*	20-40	*

The privacy guarantees offered by *k*-anonymity are limited, but its simplicity has made it quite popular and a standard method in the arsenal of privacy enhancing technologies.

Differential privacy [Dwork et al. 2006]

Intuition: An algorithm satisfies differential privacy (DP) if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not.

Definition: [Dwork et al. 2006] A randomized function $\mathcal{M}$ gives ε-differential privacy if for all databases D_1 and D_2 differing on at most one element, and all measurable $S \subseteq \text{range}(\mathcal{M})$,

$$\mathbb{P}[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(D_2) \in S],$$

where the probability is with respect to the randomness of $\mathcal{M}$.

The lower the value ε, the more indistinguishable the results, and therefore the more each individual's data is protected.

Differential privacy and noise

Almost all existing mechanisms to implement DP are based on adding noise to the data or the data queries.

Laplacian mechanism: $\text{Lap}(\sigma) \sim \exp(-|x|/\sigma)$
adding Laplacian noise $\text{Lap}(\sigma)$ with $\sigma \sim 1/\varepsilon$ to data queries

The lower the value ε, the more noisy the data are.

Challenge: how to choose "privacy budget" ε in practice.
Typical value for ε in practice: $\varepsilon \leq 4$.

DP used in connection with releasing the Census 2020 data

Synthetic Data

Synthetic data are generated (typically via some randomized algorithm) from existing data such that they maintain the statistical properties of the original data set, but do so without risk of exposing sensitive information.

Intrinsic conflict between privacy and utility: Synthetic data should be different from the original data, but still very similar.

Extreme cases:
Synthetic data = original data: Perfect utility, zero privacy
Synthetic data = random data: Perfect privacy, zero utility.

Combining synthetic data with DP has great promise to mitigate key weaknesses of DP [Bellovin 2019, Kearns 2020]

Classical way to privatize data

Anonymization

Diff.Privacy

$\Longrightarrow$ Loss of useful information!

Privacy via synthetic data

Synthetic data generation

None of the faces in the right panel are real. But hopefully they are a faithful representation of the original face dataset

How do we know that the synthetic dataset captures the nuances of the original dataset?
How do we know that privacy is preserved?

Privacy and health care data

In 2013 the UK National Health Services planned to sell patient data to drug and insurance firms. Patients could not opt out.

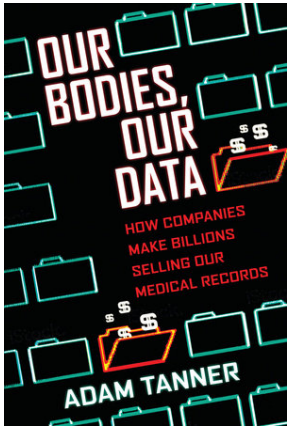
After major complains, the program was modified so that patients could opt out. Eventually the whole program was cancelled (for now ...).

Privacy and health care data

In 2013 the UK National Health Services planned to sell patient data to drug and insurance firms. Patients could not opt out.

After major complains, the program was modified so that patients could opt out. Eventually the whole program was cancelled (for now ...).

Major misconception:
Patients should not have to opt out in the first place, since medical data are among the most personal data of an individual.



The legal right of businesses to harvest and sell the information of individual patients without their permission has been upheld by the US Supreme Court. [564 U.S. 552 (2011)]

Privacy and health care data

The intensive care unit is becoming one of the most data-driven clinical environments.

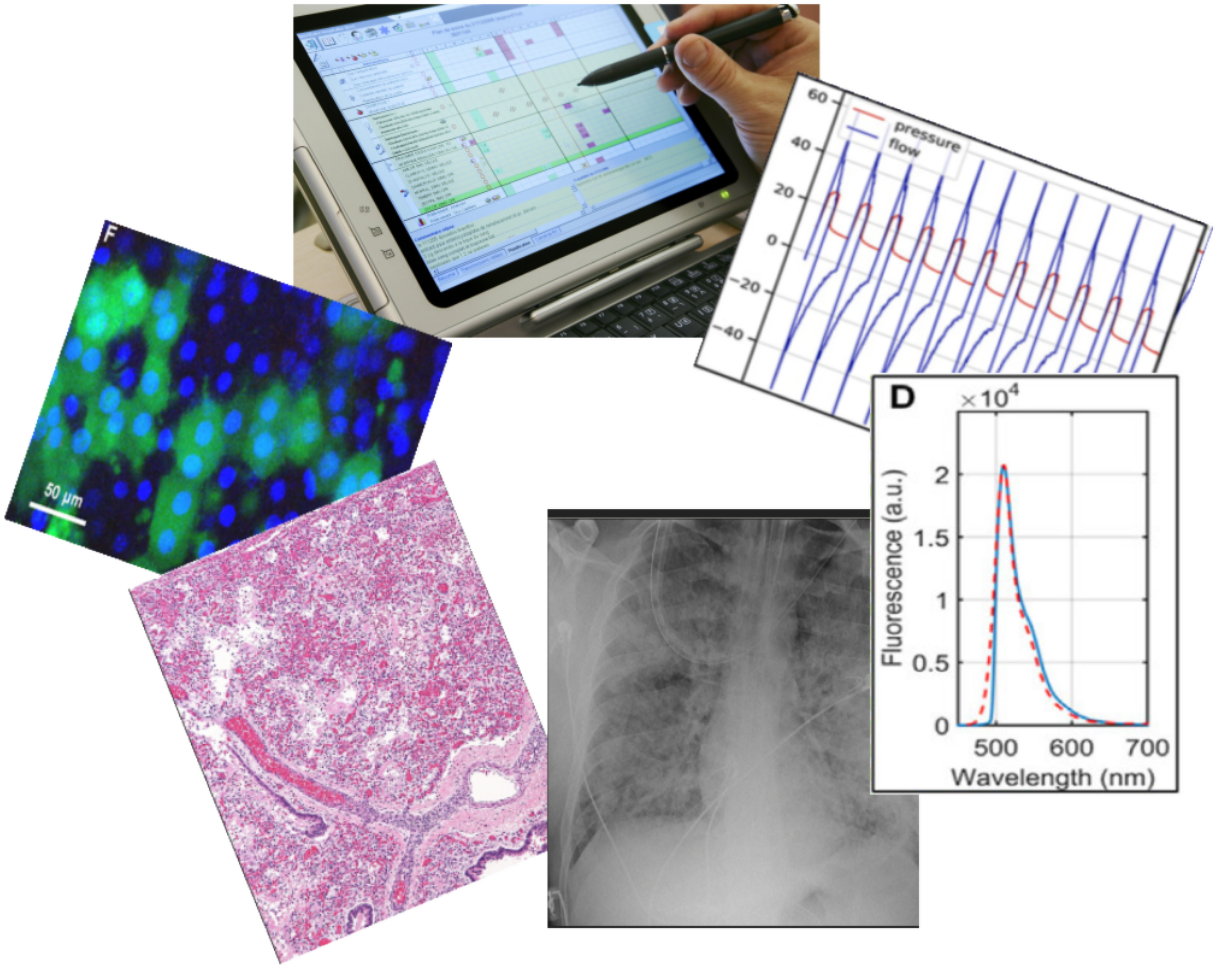
Data analysis approaches that are tailored to the specific needs and limitations of the ICU environments are needed.

But lack of availability and access to sufficient data is a main road block for medical experts and AI scientists towards the development of advanced medical decision support systems

Thus ICU is a prototypical setting where (high-quality) synthetic data would be tremendously helpful to break through this data bottleneck, while respecting health data privacy laws.

Privacy and health care data

Challenge: Data are heterogeneous, dynamic, multimodal, ...

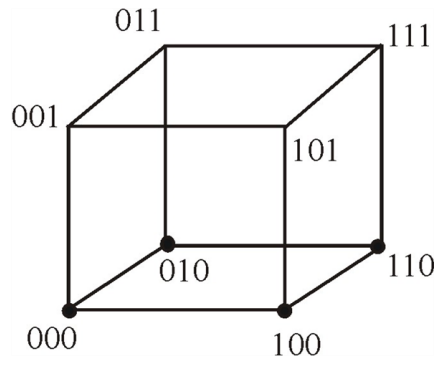


Benchmark model: Boolean cube

Let us start with a seemingly simple data model ...

Boolean cube $\{0,1\}^p$ as benchmark model for the dataset $\mathcal{X}$.

$\mathcal{X}$ has n rows and p columns.
$\mathcal{X}$ might represent an electronic health record (EHR) with n patients, each patient is represented by a row $x \in \{0,1\}^p$.



We can also represent categorical data (gender, occupation, etc.) or numerical data (by splitting them into intervals) on the Boolean cube via binary or one-hot encoding.

Utility

Accuracy: We measure accuracy by comparing the marginals of true and synthetic data.
A d-dimensional marginal of the true data has the form

$$\frac{1}{n}\sum_{i=1}^{n} x_i(j_1)\cdots x_i(j_d)$$

for some given indices $j_1, \ldots, j_d \in [p]$.

EHR: a d-dimensional marginal is the fraction of the patients whose d given parameters all equal 1.
The one-dimensional marginals encode the means of the parameters, and the two-dimensional marginals encode the covariances, e.g., nr. of patients who smoke and have diabetes.

Accurate and private synthetic data

Given data $x_1, \ldots, x_n \in \{0,1\}^p$, our goal is to design a randomized algorithm that satisfies:

(i) **synthetic data:** the algorithm outputs a list of vectors $y_1, \ldots, y_m \in \{0,1\}^p$;
(ii) **efficiency:** runtime is polynomial in n and p;
(iii) **privacy:** the algorithm is differentially private;
(iv) **accuracy:** the low-dimensional marginals of $y_1, \ldots, y_m$ are close to those of $x_1, \ldots, x_n$.

Accurate and private synthetic data

Given data $x_1, \ldots, x_n \in \{0,1\}^p$, our goal is to design a randomized algorithm that satisfies:

(i) **synthetic data:** the algorithm outputs a list of vectors $y_1, \ldots, y_m \in \{0,1\}^p$;
(ii) **efficiency:** runtime is polynomial in n and p;
(iii) **privacy:** the algorithm is differentially private;
(iv) **accuracy:** the low-dimensional marginals of $y_1, \ldots, y_m$ are close to those of $x_1, \ldots, x_n$.

[Ullman-Vadhan 2011]: Achieving (i),(iii),(iv) is NP-hard.
Thus we cannot satisfy (i)-(iv) simultaneously.

Accurate and private synthetic data

Given data $x_1, \ldots, x_n \in \{0,1\}^p$, our goal is to design a randomized algorithm that satisfies:

(i) **synthetic data:** the algorithm outputs a list of vectors $y_1, \ldots, y_m \in \{0,1\}^p$;
(ii) **efficiency:** runtime is polynomial in n and p;
(iii) **privacy:** the algorithm is differentially private;
(iv) **accuracy:** the low-dimensional marginals of $y_1, \ldots, y_m$ are close to those of $x_1, \ldots, x_n$.

[Ullman-Vadhan 2011]: Achieving (i),(iii),(iv) is NP-hard.
Thus we cannot satisfy (i)-(iv) simultaneously.

Why not create synthetic data by adding noise to true data and use randomized rounding to map noisy data to Boolean cube?
Problem: We need to add a lot of noise to ensure DP.
Hence, resulting data would not be accurate.

Fighting surveillance capitalism is NP-hard ...

We will circumvent the NP-hardness of the problem in two different ways:

1. Relax accuracy of low-dimensional marginals from “all marginals” to “most marginals”. We will show that in this case we can achieve (i)-(iv).
2. Statistical framework: Ullman-Vadhan is a worst-case no-go result. We will show that for typical data we can achieve (i)-(iv).

k-anonymity

k-anonymity: The information for each person contained in the dataset cannot be distinguished from at least $k - 1$ individuals whose information also appear in the dataset.

k-anonymity: a simple idea, how hard can it be?

k-anonymity

k-anonymity: The information for each person contained in the dataset cannot be distinguished from at least $k - 1$ individuals whose information also appear in the dataset.

k-anonymity: a simple idea, how hard can it be?

NP-hard!

- Finding the optimal partition into k-anonymous groups is NP-hard [Meyerson&Williams, 2004]
- Optimal multivariate microaggregation is NP-hard for $k \geq 3$ [Oganian&Domingo-Ferrer 2001,Thaeter&Reischuk 2020]

k-anonymity and microaggregation

Given a dataset $\mathcal{X}$ consisting of n elements (e.g., patients) each described by p real-valued attributes, i-th element is represented by a vector $x_i \in \mathbb{R}^p$.

Microaggregation:

1. Cluster the vectors x_i into clusters C_j, each of size $\geq k$.
2. All elements x_i of a cluster C_j are replaced by a cluster representative y_j, thus giving k-anonymity.

In order to keep the quality of the data in $\mathcal{X}$ one would like to generate as little distortion as possible when replacing individual attribute vectors x_i by cluster representatives y_j.

Related to k-means, but more challenging, since in microaggregation each cluster must have (at least) k elements.

Numerous papers on microaggregation for privacy:
[Domingo-Ferrer&Sanchez, Laszlo&Mukherjee, Monedero, ...]
But no guarantees regarding utility.

Challenges of k-anonymity

- Real-world datasets are high-dimensional and very sparse
- Need to find $k-1$ neighbors for each data point: How to do this computationally efficiently?
- How to control utility? Existing methods provide no utility guarantees.
- Projection to low dimensions: may lose all information, and how to control utility in projected space?
- How to protect against linkage attacks?

Challenges of k-anonymity

- Real-world datasets are high-dimensional and very sparse
- Need to find $k-1$ neighbors for each data point: How to do this computationally efficiently?
- How to control utility? Existing methods provide no utility guarantees.
- Projection to low dimensions: may lose all information, and how to control utility in projected space?
- How to protect against linkage attacks?

We will address these issues by first solving a completely different problem

Covariance Loss

A fundamental question from probability:

How much information is lost when we take conditional expectation?

Given a random variable X and a sigma-algebra $\mathcal{F}$,
consider the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$
Law of total expectation: Y gives unbiased estimate of mean:

$$\mathbb{E}\, X = \mathbb{E}\, Y$$

Law of total variance:

$$\mathrm{Var}(X) - \mathrm{Var}(Y) = \mathbb{E}\, X^2 - \mathbb{E}\, Y^2 = \mathbb{E}(X - Y)^2$$

$\Longrightarrow$: taking conditional expectation underestimates variance.

How much variance is lost?

Covariance Loss

If X is bounded, say $|X| \leq 1$, we can decompose the interval $[-1, 1]$ into k subintervals of length $2/k$ each, take F_i to be the preimage of each interval under X, and let $\mathcal{F} = \sigma(F_1, \ldots, F_k)$ be the sigma-algebra generated by these events.

Since X and Y takes values in the same subinterval a.s., we have $|X - Y| \leq 2/k$ a.s. Thus, the law of total variance gives

$$\mathrm{Var}(X) - \mathrm{Var}(Y) \leq \frac{4}{k^2}.$$

Covariance Loss

Generalize this to high dimensions: $X \in \mathbb{R}^p$ is a random vector and $Y = \mathbb{E}[X|\mathcal{F}]$, the law of total expectation holds unchanged. Let $\Sigma_X = \mathbb{E}(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^\mathsf{T}$ be the covariance matrix of X. The law of total variance becomes the law of total covariance:

$$\Sigma_X - \Sigma_Y = \mathbb{E}\,XX^\mathsf{T} - \mathbb{E}\,YY^\mathsf{T} = \mathbb{E}(X - Y)(X - Y)^\mathsf{T}$$

In particular: $\Sigma_X \succeq \Sigma_Y$.

Heuristically, the simpler the sigma-algebra $\mathcal{F}$ is, the more variance gets lost.

- What is the best sigma-algebra $\mathcal{F}$ with a given complexity?
- How small can the covariance loss be?

We face the curse of dimensionality:
The unit Euclidean ball in $\mathbb{R}^p$ cannot be partitioned into k subsets of small diameter, unless k is exponentially large in p.

Covariance Loss - Beating the curse of dimensionality

Theorem (Covariance loss, [BSV, 2021])

Let X be a random vector in $\mathbb{R}^p$ such that $\|X\|_2 \leq 1$ a.s. Then, for every $k \geq 3$, there exists a partition of the sample space into at most k sets such that for the sigma-algebra $\mathcal{F}$ generated by this partition, the conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ satisfies

$$\|\Sigma_X - \Sigma_Y\|_2 \leq C\sqrt{\frac{\log\log k}{\log k}}.$$

The rate is optimal up to a $\sqrt{\log\log k}$ factor.
The partition can be made with exactly k sets, all of which have the same probability $1/k$.
Moreover, the result extends (magically?) to higher moments:

$$\|\mathbb{E}\, X^{\otimes d} - \mathbb{E}\, Y^{\otimes d}\|_2 \leq 4^d \|\mathbb{E}\, X^{\otimes 2} - \mathbb{E}\, Y^{\otimes 2}\|_2 = 4^d \|\Sigma_X - \Sigma_Y\|_2.$$

From covariance loss to microaggregation

True data are $x_1, \ldots, x_n \in \mathbb{R}^p$. Let $X(i) = x_i$ be the random variable on the sample space $[n]$ equipped with uniform probability distribution. Obtain a partition $[n] = I_1 \cup \cdots \cup I_m$ from the Covariance Loss Theorem and assume for simplicity that all sets I_j have the same cardinality $|I_j| = n/k$.
The conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ on the sigma-algebra $\mathcal{F} = \sigma(I_1, \ldots, I_m)$ generated by this partition takes values

$$y_j = \frac{k}{n} \sum_{i \in I_j} x_i, \quad j = 1, \ldots, k.$$

with probability $1/k$ each.
Thus, the synthetic data $y_1, \ldots, y_k$ are obtained by taking local averages, or by *microaggregation* of the input data $x_1, \ldots, x_n$.
Crucial: synthetic data is obviously (n/k)-*anonymous*.

From covariance loss to microaggregation

True data are $x_1, \ldots, x_n \in \mathbb{R}^p$. Let $X(i) = x_i$ be the random variable on the sample space $[n]$ equipped with uniform probability distribution. Obtain a partition $[n] = I_1 \cup \cdots \cup I_m$ from the Covariance Loss Theorem and assume for simplicity that all sets I_j have the same cardinality $|I_j| = n/k$.
The conditional expectation $Y = \mathbb{E}[X|\mathcal{F}]$ on the sigma-algebra $\mathcal{F} = \sigma(I_1, \ldots, I_m)$ generated by this partition takes values

$$y_j = \frac{k}{n} \sum_{i \in I_j} x_i, \quad j = 1, \ldots, k.$$

with probability $1/k$ each.
Thus, the synthetic data $y_1, \ldots, y_k$ are obtained by taking local averages, or by *microaggregation* of the input data $x_1, \ldots, x_n$.
Crucial: synthetic data is obviously (n/k)-*anonymous*.

What about accuracy of the synthetic data?

Law of total expectation $\mathbb{E}\, X = \mathbb{E}\, Y \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{k}\sum_{j=1}^{k} y_j$.
Thus one-dim. marginals are preserved exactly.
Marginals of higher dimension: assume $\|x_i\|_2 \leq 1$ for all i.
Then Covariance Loss Theorem yields

$$\|\frac{1}{n}\sum_{i=1}^{n} x_i^{\otimes d} - \frac{1}{k}\sum_{j=1}^{k} y_j^{\otimes d}\|_2 \lesssim 4^d \sqrt{\frac{\log\log k}{\log k}}.$$

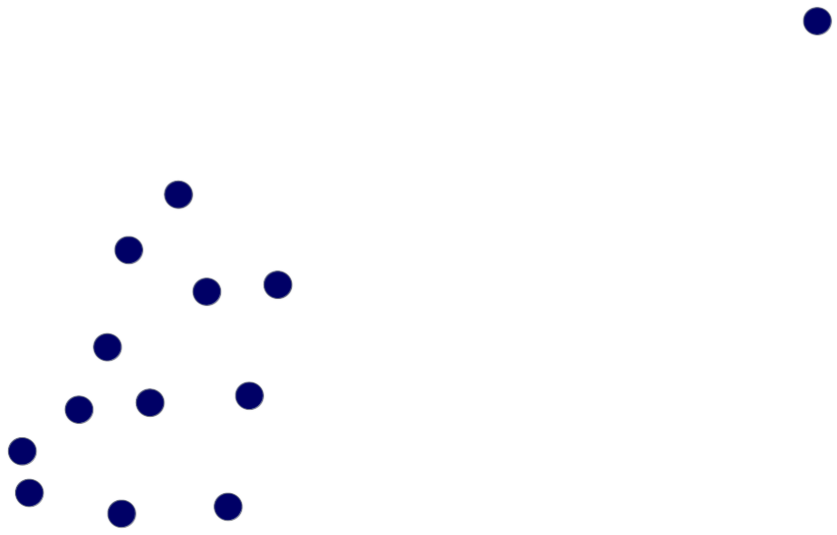
Thus, if $k \gg 1$ and $d = O(1)$, the synthetic data is accurate in the sense of the L^2-average of marginals.

Privacy

How can we upgrade anonymity to differential privacy?

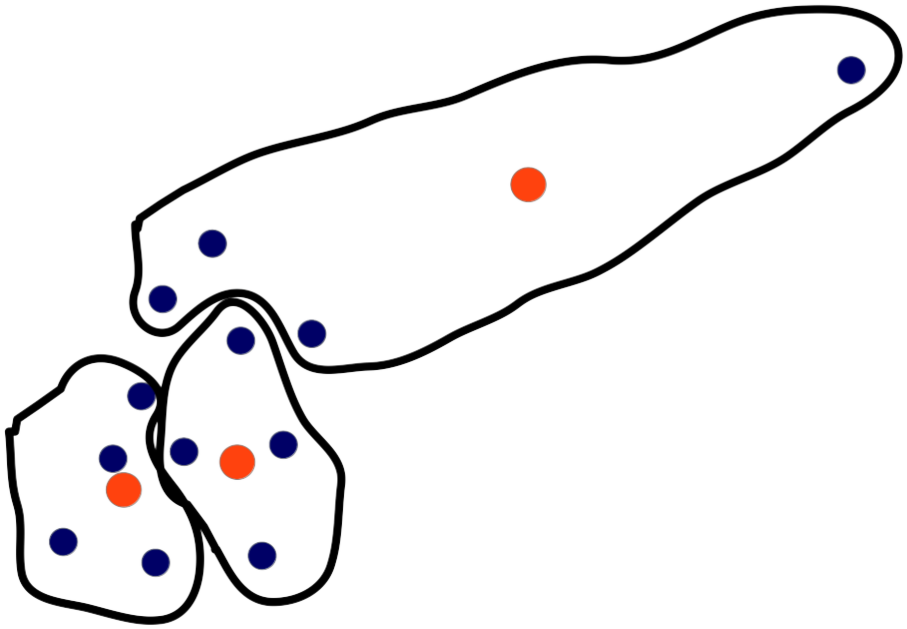
Microaggregation reduces sensitivity of the synthetic data, but by itself it is not differentially private.

Microaggregation and privacy



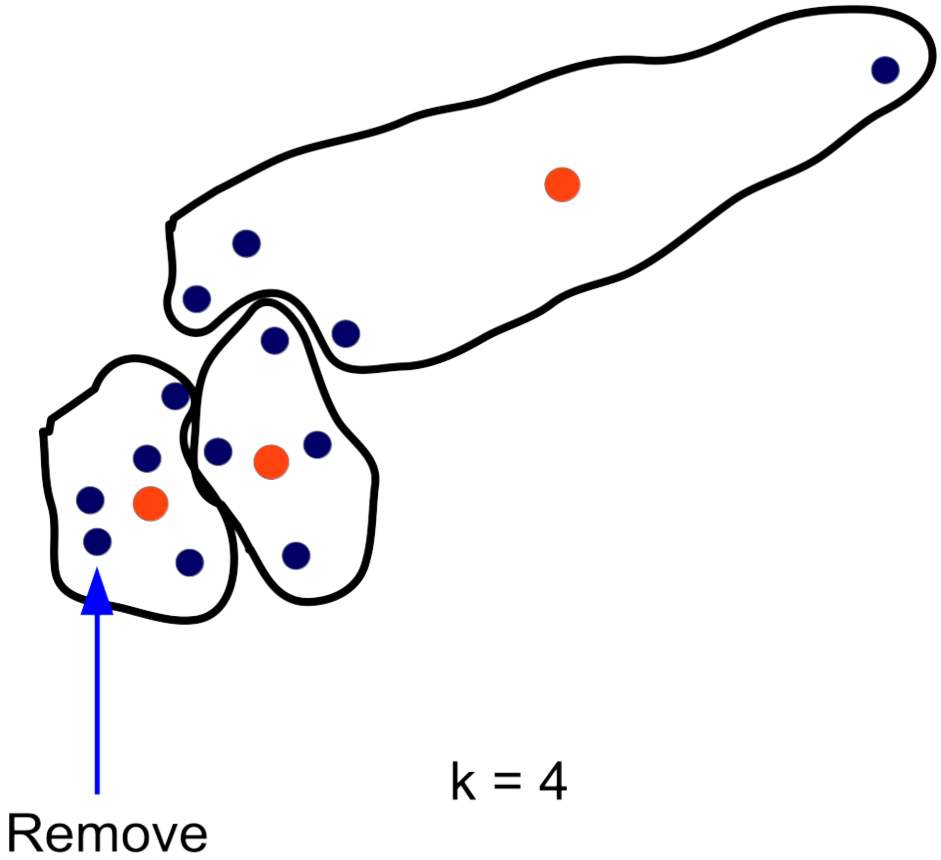
k = 4

Microaggregation and privacy

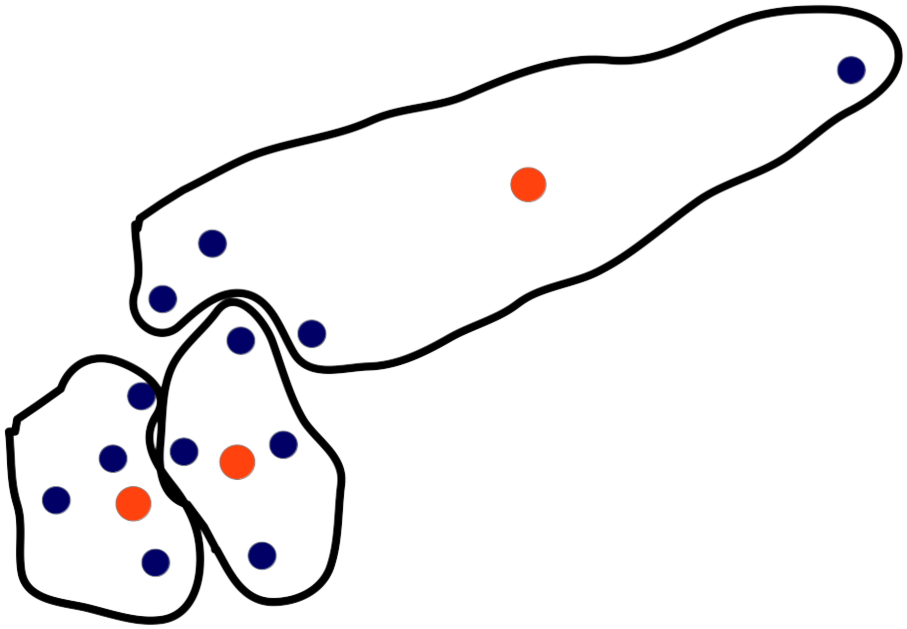


k = 4

Microaggregation and privacy

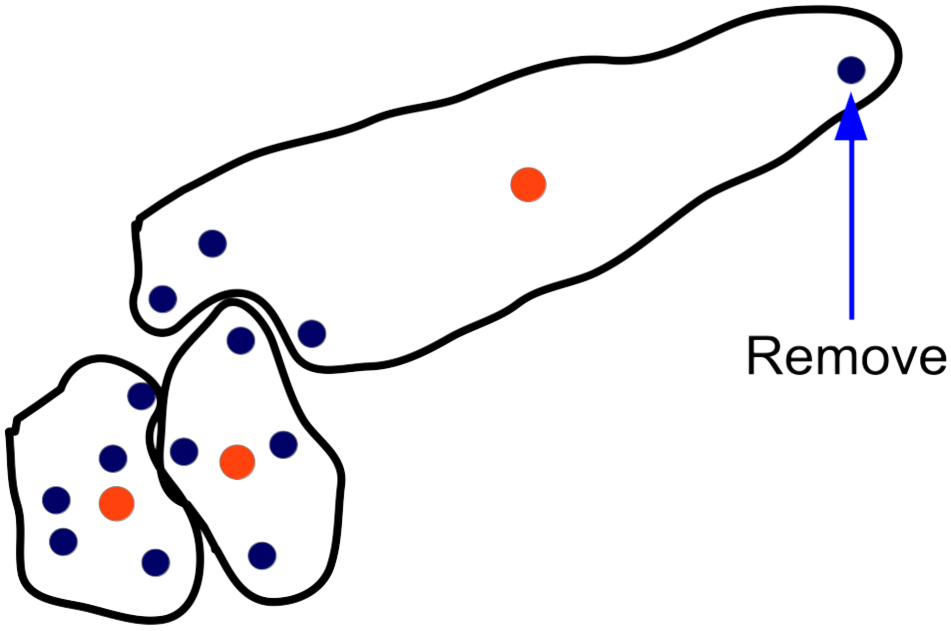


Microaggregation and privacy



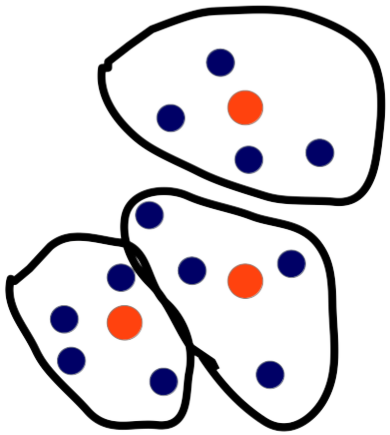
k = 4

Microaggregation and privacy



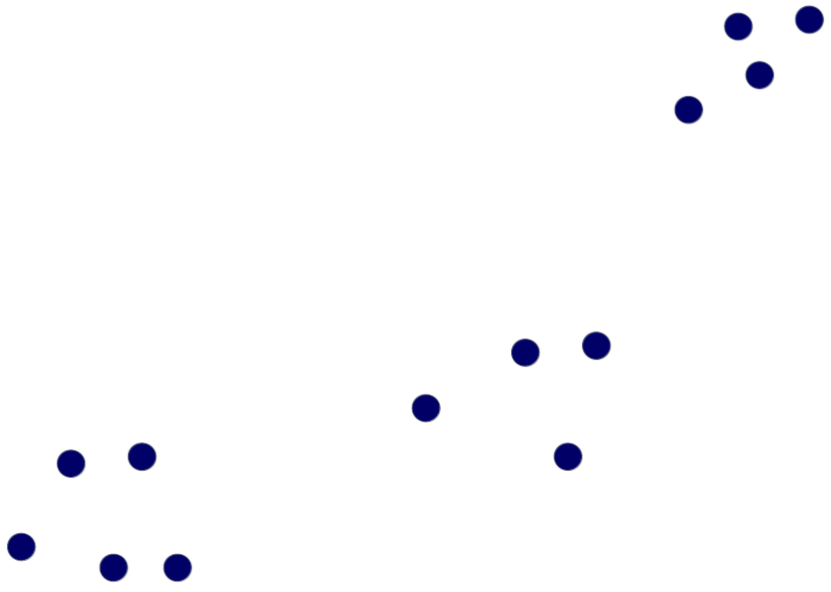
k = 4

Microaggregation and privacy



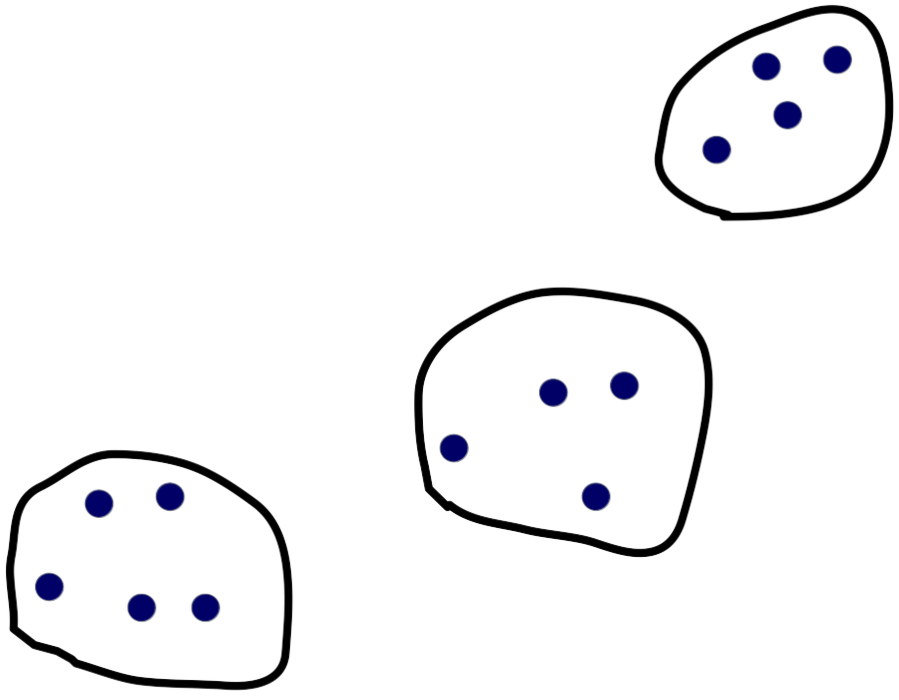
k = 4

Microaggregation and privacy



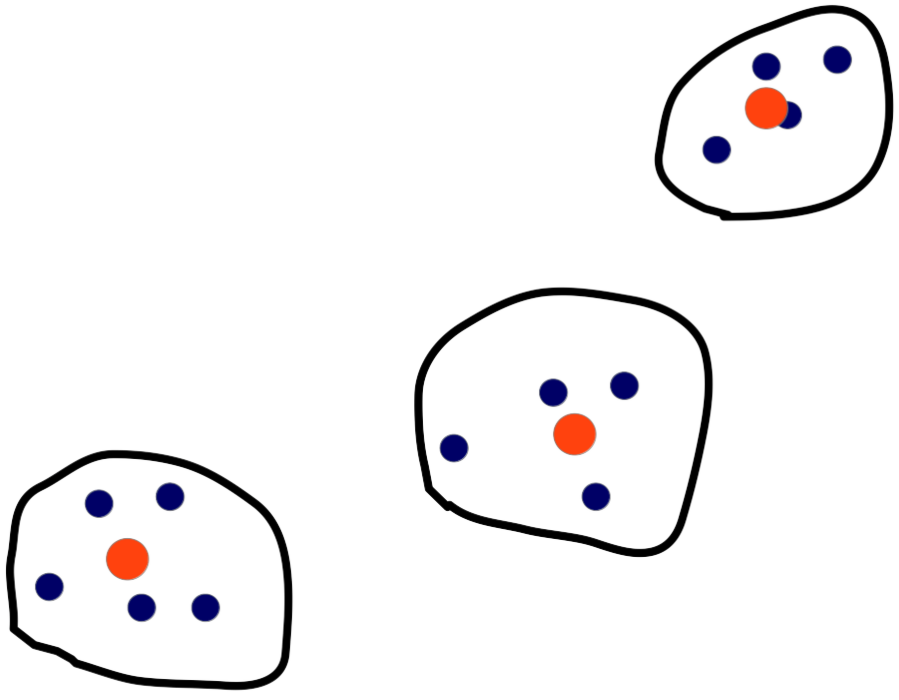
k = 4

Microaggregation and privacy



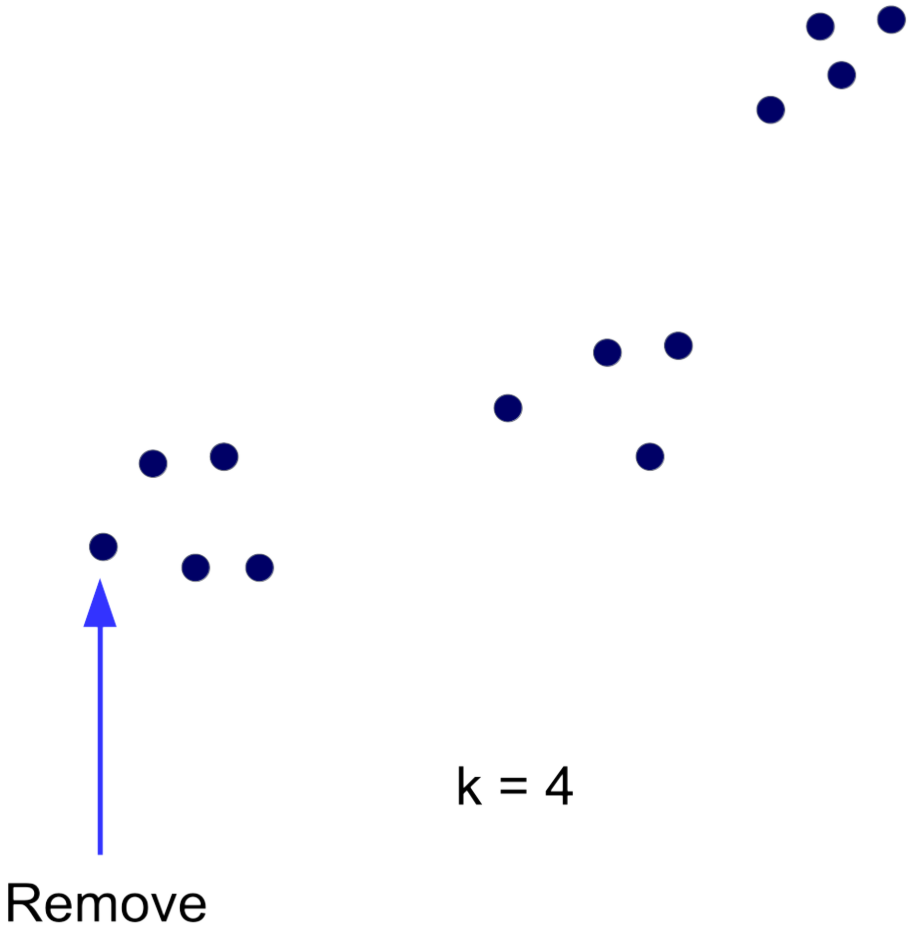
k = 4

Microaggregation and privacy

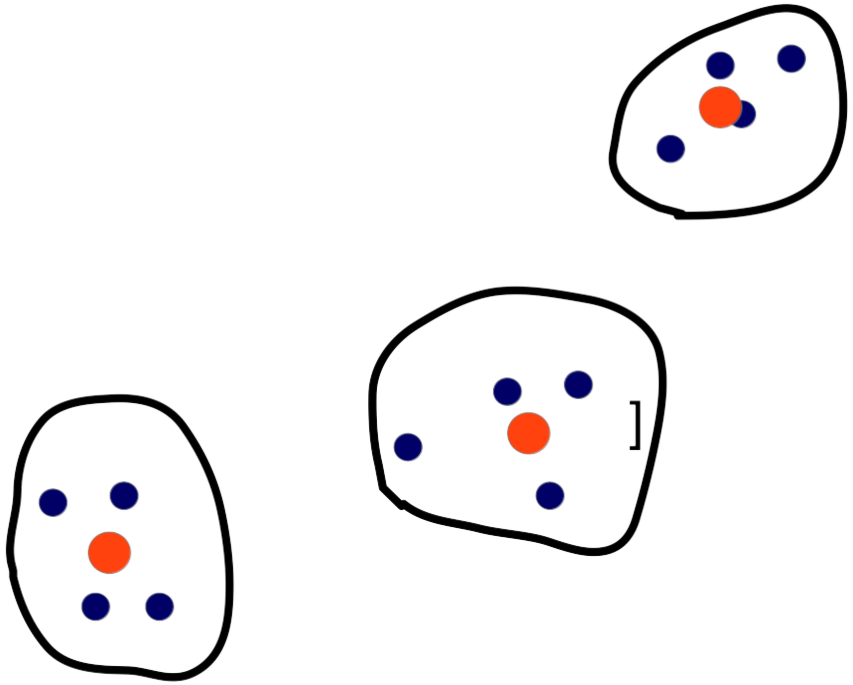


k = 4

Microaggregation and privacy

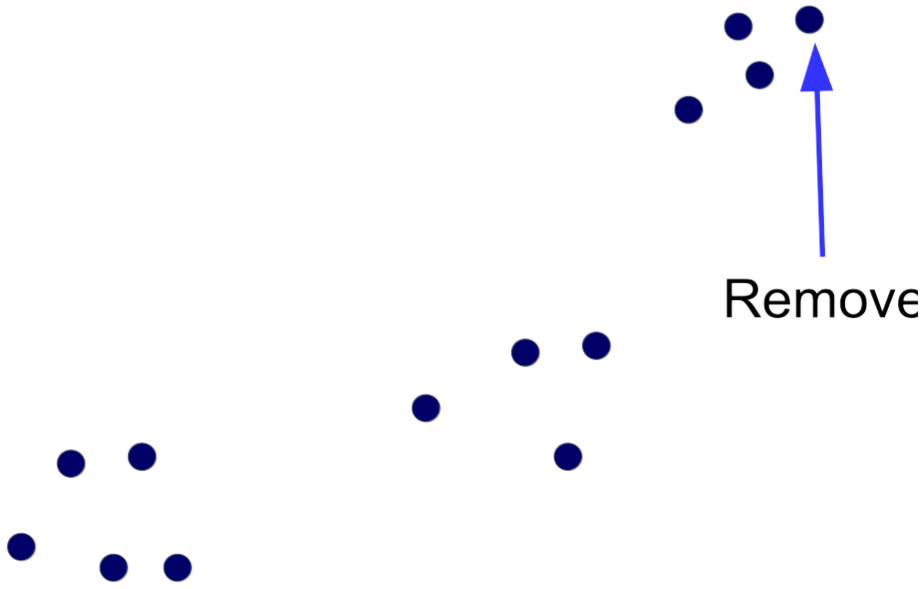


Microaggregation and privacy



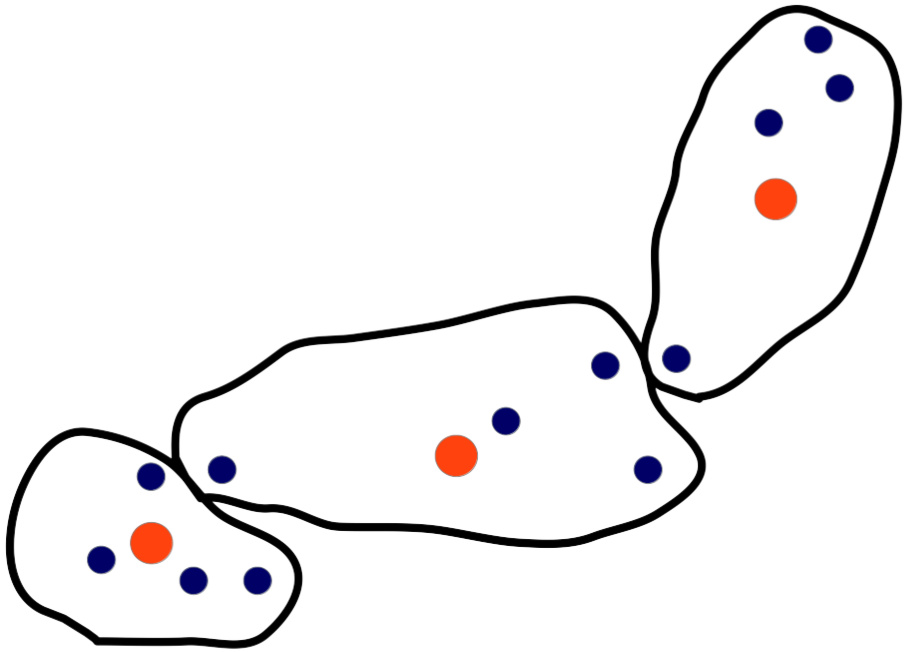
k = 4

Microaggregation and privacy



k = 4

Microaggregation and privacy



k = 4

Accurate, private synthetic data via microaggregation

Given a sequence of points $x_1, \ldots, x_n$ on the cube $\{0,1\}^p$ (true data), our algorithm comprises the following steps:

1. Spectral projection of data onto leading eigenvectors of the (noisy) second-moment matrix $S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$.
   (key to computational efficiency)
2. Nearest neighbor partition in projected space
3. Damped microaggregation
   (key to addressing instability of microaggregation)
4. Add Laplacian noise (for privacy)
5. Metric projection
   (adding noise for privacy may move data outside cube)
6. Bootstrapping and randomized rounding
   (To create unlimited number of synthetic data)

Accurate, private synthetic data via microaggregation

Theorem

Let $\varepsilon \in (0,1)$. There exists an ε-differentially private randomized algorithm that transforms input data $x_1, \ldots, x_n \in \{0,1\}^p$ into the output data $z_1, \ldots, z_m \in \{0,1\}^p$ in such a way that the synthetic data are $o_\varepsilon(1)$-accurate for d-dimensional marginals on average. The algorithm runs in time polynomial in p, n and linear in m, and is independent of d.

Challenges in 30-page proof:

- Keep track of information loss in projected space
- Right balance of noise (privacy) and accuracy
- Handle sensitivity of microaggregation to DP
- Extension to higher-dim. marginals, when input is only about two-dim. marginals

Accurate, private synthetic data via microaggregation

Strengths of the algorithm:
No model assumed about data
Not limited to specific queries
Applicable to data living in a convex set, beyond Boolean cube

Drawback: Guarantee only for most marginals

A statistical framework for synthetic data

Recall: Making DP synthetic data that preserve all two-dim. marginals with accuracy $o(1)$ is NP-hard [Ullman-Vadhan 2011]

This is a worst-case result, for the worst kind of data.
Yet the *worst* kind of data, for which the problem is NP-hard, are rarely seen in practice.

Perhaps things are better in the "typical case"?
This suggests: consider a statistical framework

A statistical framework for synthetic data

Assume the true data $\mathcal{X}$ is a random sample drawn from some probability space (Ω, Σ, ν). The probability distribution ν specifies the population model of the true data.

We assume that we neither know ν, nor that we can sample according to ν, thereby generating more true data.

Suppose, however, that we can sample from Ω according to some other, known, probability measure μ.

Example: We may not know the population distribution ν of the patients in the Boolean cube $\Omega = \{0, 1\}^p$, but we can still sample from the cube according to the uniform measure μ.

Similarly, while we may not know the population distribution ν of written notes in patient health records, there do exist generative models μ that generate texts.

A statistical framework for synthetic data

Linear statistics of the data $\mathcal{X} = (x_1, \ldots, x_n)$ are sums of the form $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ for $f \in \mathcal{F}$, where $\mathcal{F}$ is a set of functions from Ω to $[-1, 1]$. For instance, linear statistics include marginals.

We would like the synthetic data $\mathcal{Y}$ to approximately preserve all these sums, up to a given additive error δ:

$$\max_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} f(y_i) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right| \leq \delta.$$

In this case we say that the synthetic dataset is δ-*accurate*.

A statistical framework for synthetic data

Often $|\Omega|$ is too large for computations while $|\mathcal{F}|$ is reasonable. E.g., if $\mathcal{F}$ encodes all d-dimensional marginals of $\{0,1\}^p$, $|\Omega|$ is exponential in p, while $|\mathcal{F}|$ is polynomial in p.

To circumvent the computational hardness of the problem we *subsample* Ω: replace Ω by a much smaller random subset Ω^* that is sampled according to the distribution μ.
Then we generate synthetic data in Ω^* by fitting the desired linear statistics $\mathcal{F}$of the true data as close as possible.

A statistical framework for synthetic data

Often $|\Omega|$ is too large for computations while $|\mathcal{F}|$ is reasonable. E.g., if $\mathcal{F}$ encodes all d-dimensional marginals of $\{0,1\}^p$, $|\Omega|$ is exponential in p, while $|\mathcal{F}|$ is polynomial in p.

To circumvent the computational hardness of the problem we *subsample* Ω: replace Ω by a much smaller random subset Ω^* that is sampled according to the distribution μ.
Then we generate synthetic data in Ω^* by fitting the desired linear statistics $\mathcal{F}$of the true data as close as possible.

But is this even possible?

This subsampling idea can only work if the sampling distribution μ has some "correlation" with the population distribution ν.

Rènyi condition number

We quantify this correlation using the Rènyi condition number:

$$\kappa(\nu\|\mu) := \int \left(\frac{d\nu}{d\mu}\right)^2 d\mu = \int \frac{d\nu}{d\mu}\, d\nu,$$

which equals the exponential the *Rènyi divergence* of order 2.

Conceptually, $\kappa(\nu\|\mu)$ is similar to the notion of the condition number in numerical linear algebra: the smaller, the better. The best value of $\kappa(\nu\|\mu)$ is 1, achieved when $\nu = \mu$.

Algorithm

Input: (a) the true data $x_1, \ldots, x_n \in \Omega$;
(b) a family $\mathcal{F}$ of test functions from Ω to $[-1, 1]$;
(c) the reduced space $\Omega^* = \{z_1, \ldots, z_m\}$.

1. **Add noise:** For each test function $f \in \mathcal{F}$, generate an independent Laplacian random variable $\lambda(f) \sim \mathrm{Lap}(\sigma)$.
2. **Reweight:** Compute a density h^* on Ω^* whose linear statistics are uniformly as close as possible to the linear statistics of the true data perturbed by Laplacian noise:

$$h^* = \operatorname{argmin} \left\{ \max_{f \in \mathcal{F}} \left| \sum_{i=1}^{m} f(z_i) h(z_i) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \lambda(f) \right| \right\},$$

   where h^* is a density.
3. **Bootstrap:** Create a sequence $y_1, \ldots, y_k$ of k elements drawn from Ω^* independently with density h^*.

Output: synthetic data $y_1, \ldots, y_k$.

Private, accurate, efficient synthetic data

Theorem (Privacy)

Let $\delta > 0, \gamma > 0$ *and set* $\sigma = \delta / \log(|\mathcal{F}|/\gamma)$. *If*

$$n \geq 2(\varepsilon\delta)^{-1}|\mathcal{F}| \log(|\mathcal{F}|/\gamma),$$

then the algorithm is ε*-differentially private.*

Theorem (Accuracy)

Let $\min(n, k) \geq \delta^{-2} \log(|\mathcal{F}|/\gamma)$ *and* $m \geq \delta^{-2} K|\mathcal{F}|/\gamma$. *Set* $\sigma = \delta / \log(|\mathcal{F}|/\gamma)$. *Assume that the Rènyi condition number satisfies* $\kappa(\nu\|\mu) \leq K$. *Then with probability at least* $1 - 4\gamma$ *the synthetic dataset generated by the algorithm is* (8δ)*-accurate.*

Computational efficiency: Computing h^* amounts to solving a linear program with $|\Omega^*| \leq m$ variables, thus complexity of the algorithm is polynomial in $|\Omega^*|$.

Statistical framework - challenges

- While the proposed method provides a simple and efficient roadmap to construct private synthetic data that preserve with high accuracy linear statistics of the original data, we may require our synthetic data to accurately model other features of the data that are not (fully) captured by linear statistics.
  How well do linear statistics inform other kinds of data analysis, e.g., clustering, classification, regression, ...?
- We do not know the population distribution ν, thus we may not know how to choose a good sampling distribution μ. Using various generative models seem a natural choice for certain types of data, such as text and images. Using those, we may hope to build the sampling distribution μ that has enough "overlap" with the population distribution ν (as measured by the Rènyi condition number).

Differential privacy without the noise

- DP is always implemented by adding some form of noise.
- But noise will negatively affect utility and can inject systematic errors, hence bias, into the data!
- Can we achieve DP in a noisefree way?

Differential privacy without the noise

- DP is always implemented by adding some form of noise.
- But noise will negatively affect utility and can inject systematic errors, hence bias, into the data!
- Can we achieve DP in a noisefree way?

Yes, we can!

Private sampling:
Can modify the algorithm presented in the statistical framework so that DP can be achieved without adding noise but via a carefully calibrated random sampling strategy.
Details and proofs: rather technical.
Mathematical tools include: Boolean Fourier analysis, hypercontractivity, duality, and empirical processes.

Differential privacy without the noise

Lemma (Private sampling)

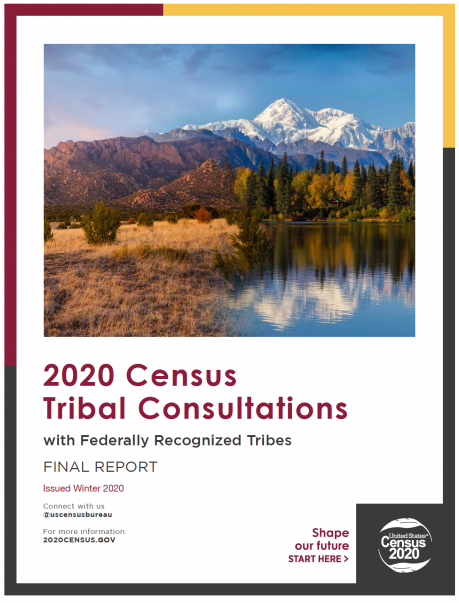
Let Ω be a finite set. Let f be a mapping that takes a dataset $\mathcal{X}$ as input and returns a probability mass function $f(\mathcal{X})$ on Ω. Suppose $\varepsilon > 0$ and $k \in \mathbb{N}$ are chosen so that

$$\|f(\mathcal{X}_1)/f(\mathcal{X}_2)\|_\infty \leq \exp(\varepsilon/k)$$

for all datasets $\mathcal{X}_1$ and $\mathcal{X}_2$ that differ on a single element. Then the algorithm that takes $\mathcal{X}$ as input and returns a sample of k points drawn from Ω independently and according to the distribution $f(\mathcal{X})$ is ε-differentially private.

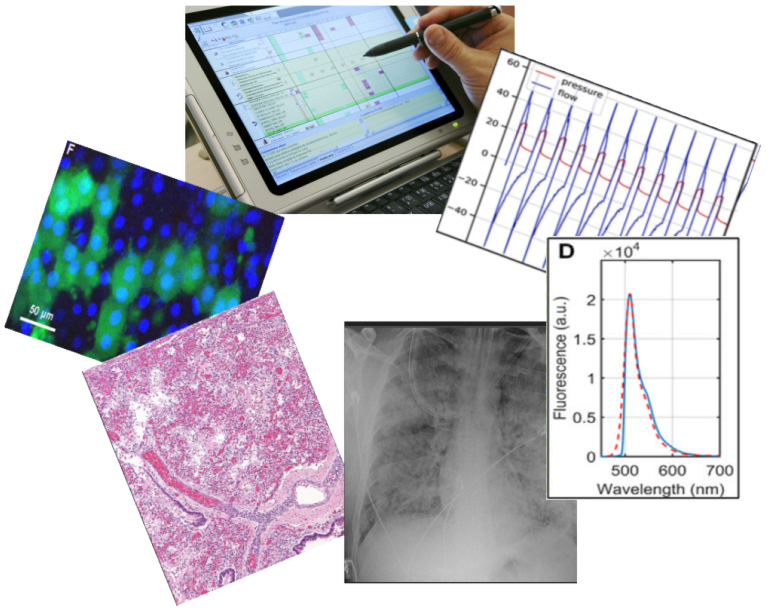
Ongoing work: Private sampling and Census 2020

The U.S. Census Bureau plans to use differential privacy in the release of the Census 2020 data.
However, initial simulations have shown that the DP Census data has a strong negative effect on small communities and minorities.



Can private sampling help to mitigate the negative effects of noise-induced DP?

Ongoing work: Synthetic data for the ICU



Further Readings

These papers can be found on the arvix or my homepage:

1. M. Boedihardjo, T. Strohmer, and R. Vershynin. Private sampling: a noiseless approach for generating differentially private synthetic data. Preprint, 2021.
2. M. Boedihardjo, T. Strohmer, and R. Vershynin Privacy of Synthetic Data: A Statistical Framework. Preprint, 2021.
3. M. Boedihardjo, T. Strohmer, and R. Vershynin. Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data. Preprint, 2021.

Conclusion and Outlook

- We have developed several mathematical frameworks for computationally efficiently creating private and accurate synthetic data
- Many open challenges: How to extend this beyond linear statistics? How to handle multimodal synthetic data? ...
- Privacy-preserving synthetic data ecosystems “democratize" data science research
- Synthetic data are a piece of the puzzle toward fighting surveillance capitalism
- See my talk “Pandemics, Privacy, and Paradoxes - Why We Need a New Paradigm for Data Science and AI”, https://www.youtube.com/watch?v=T5AWRe1aqJs