Privacy Protecting Data Science

MODELS WITH INTEGRATED PRIVACY PROTECTION & GDPR COMPLIANT LEARNING

Christoffer Långström

What is Privacy in the Age of Information?

Factors

- "Data as an asset" & data driven business models
- Data collection technology is abundant
- What is the risk to individuals in all this?

This Work

- Supported by Combient Mix
- Supervised by Raazesh Sainudiin, UU & Combient Mix

Aims:

- To provide an overview of available privacy preserving measures in data science today.
- Provide scalable implementations of pseudonomyzation & privacy preserving analysis methods
- Characterize the loss of accuracy as a result of these privacy preserving methods

Structure:

1. Introduction
2. GDPR Guidelines
3. Privacy in Data Storage
4. Privacy in Data Analysis

Challenges: The Problem of Identifiers

Direct identifiers & Quasi identifiers

Name	Gender	ZIP	Adress	DoB
John Doe	Male	12345	123 Homestreet	1980-02-01

Challenges: The Problem of Identifiers

Direct identifiers & Quasi identifiers

Name	Gender	ZIP	Adress	DoB
~~John Doe~~	Male	12345	~~123 Homestreet~~	1980-02-01

Challenges: The Problem of Identifiers

Direct identifiers & Quasi identifiers

Name	Gender	ZIP	Adress	DoB
John Doe	Male	12345	123 Homestreet	1980-02-01

- It has been found* that 87% of Americans can be identified using only {5 digit ZIP, Gender, DoB} *
- In general, few (quasi) identifiers are needed to identify an individual

* L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population

Challenges: The Problem of Revealing Statistics

- Results of data analysis & computation might itself expose individuals, or may be reverse engineered to reveal unintended information

Challenges: The Problem of Revealing Statistics

- Results of data analysis & computation might itself expose individuals, or may be reverse engineered to reveal unintended information

- **Example**: HR keeps track of how many employees have children in the office. The current figure is 36. A new employee is hired. The figure is now 37.

How can data be collected and utilized without risking personal privacy?

Can the loss of utility due to privacy measures be quantified?

GDPR & Privacy Legislation

- The General Data Protection Regulation; 2018
- Aims: *"…to protect natural persons with regard to the processing of personal data and rules related to the free movement of personal data"* (Art.1)

Principles:

- Data Protection By Design
- Right To Know
- Right Of Erasure*
- Right of Explanation

Effects:

- Data protection must be built in to the system
- Must be able to give out information on data subjects
- Must be prepared to handle data removal

Guidelines for GDPR Compliant Data science:

- ✓ Data Protection (Local encryption)
- ✓ Pseudonymization (Non-Destructive)
- ✓ Be prepared to explain: Technical Report, Understand Deployment, Educate The Subject

Part I: Privacy in Data Protection

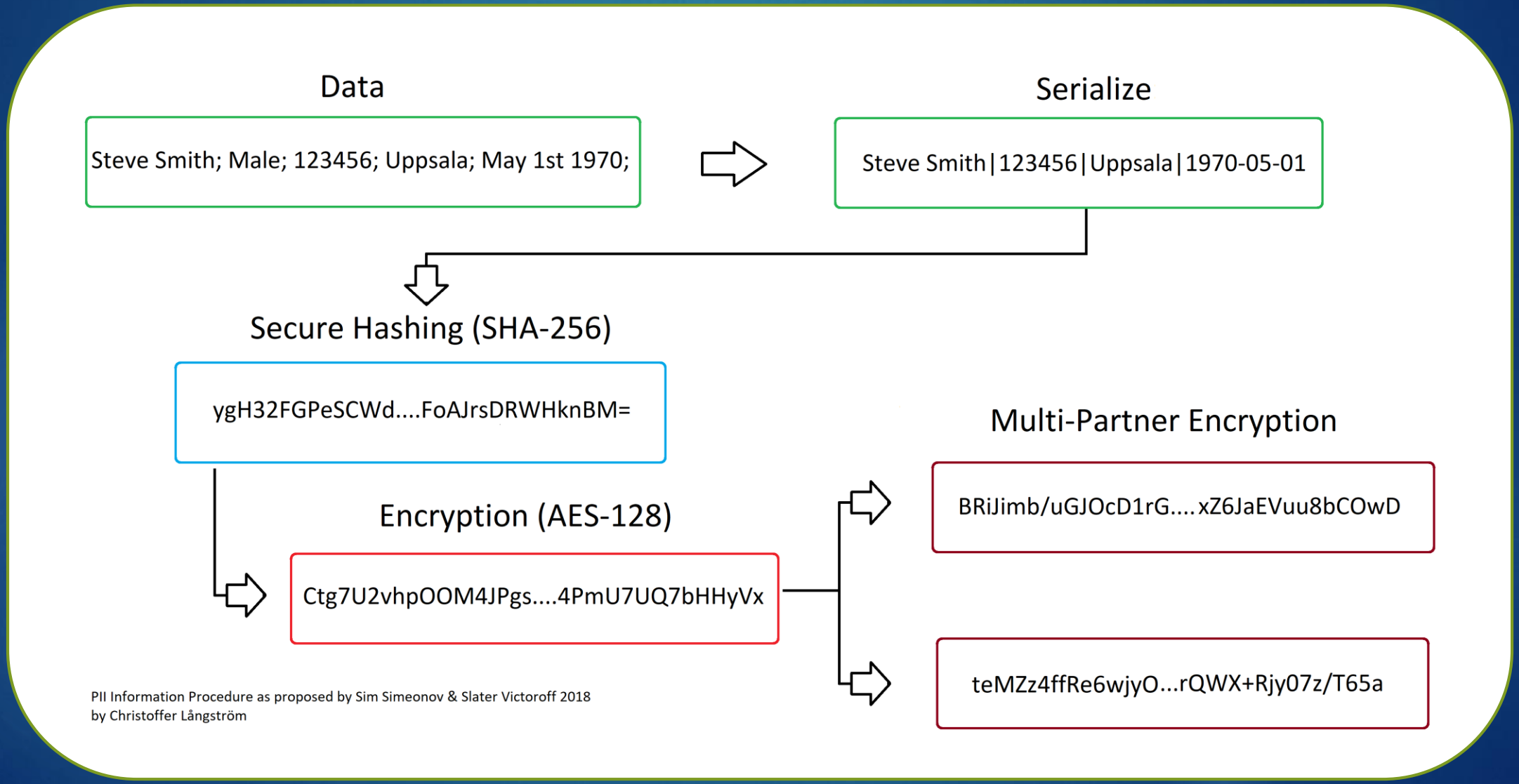
Checklist:

- ✓ Data gathered must be protected!
- ✓ Protection against intruders
- ✓ Protection against joining attacks
- ✓ Must be queryable

Pseudonomyzation & K-anonymity

Framework for Pseudonomyzation:

As suggested by Simenov & Victoroff at the 2018 Spark AI Summit



K – Anonymity

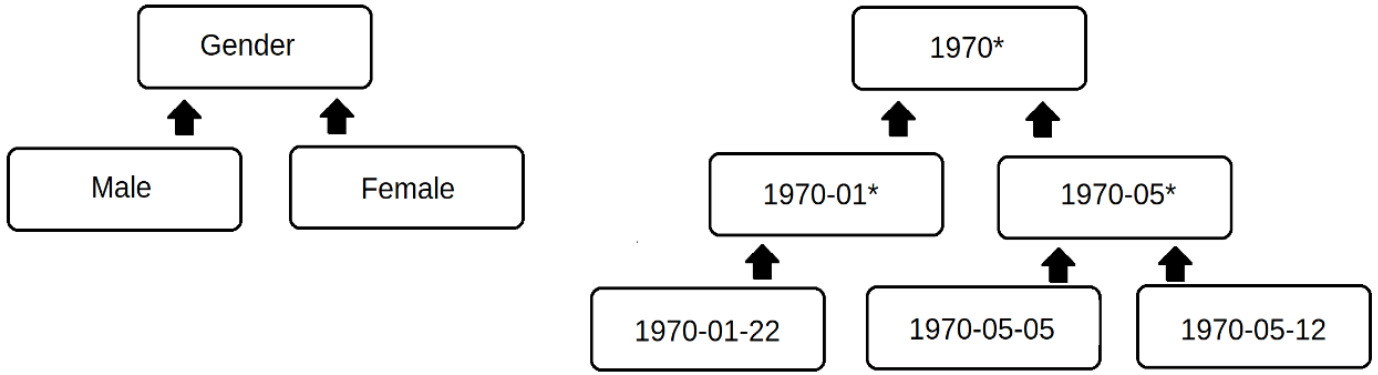
- Concept of Data Privacy (Sweeney et. al 1998)
- Enforces privacy by decreasing data resolution

> *"A data set is K-anonymous if, no matter what attributes you select by, you can never narrow it down to less than K rows"*

- Optimal K-anonymity is NP-Complete (Williams, Meyerson, 2004)
- Domain-Specific, Greedy Algorithms (Incognito, Mondrian)
- L-diversity: Privacy sensitive information is dependent on the frequency distribution of sensitive attributes

K – Anonymity: Incognito Algorithm

I) Define domain hierarchies

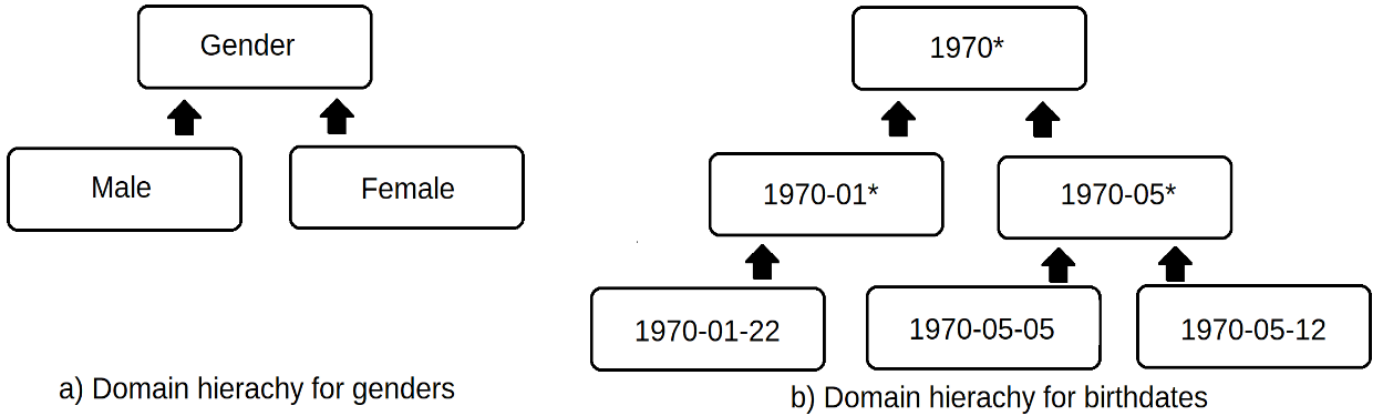


a) Domain hierachy for genders

b) Domain hierachy for birthdates

K – Anonymity: Incognito Algorithm

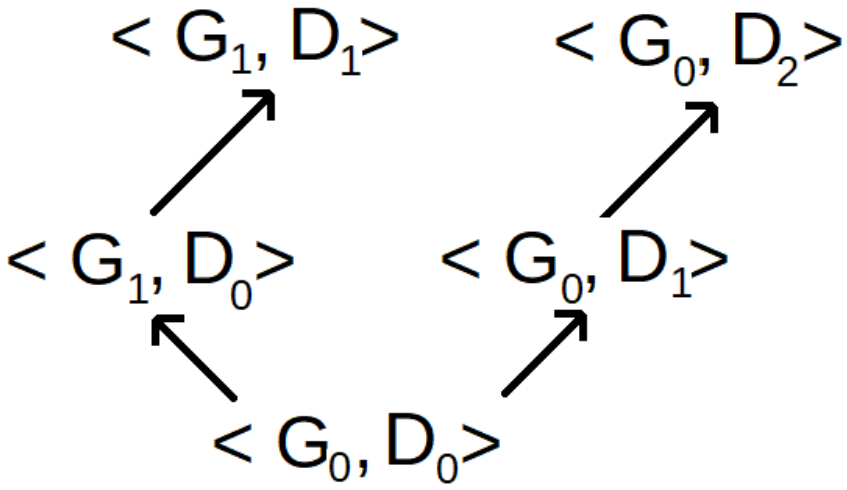
I) Define domain hierarchies



a) Domain hierachy for genders

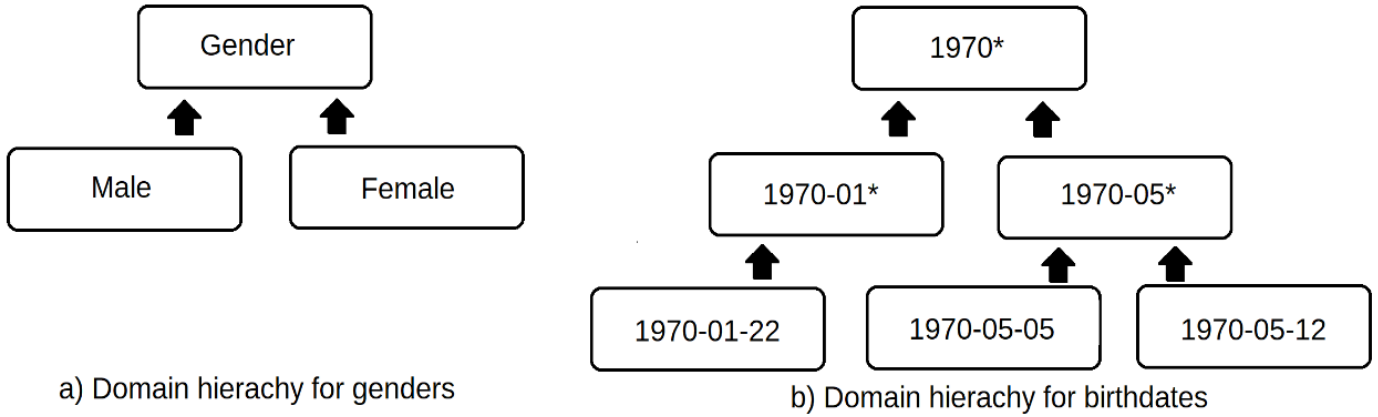
b) Domain hierachy for birthdates

II) Construct directed graphs for all generalizations

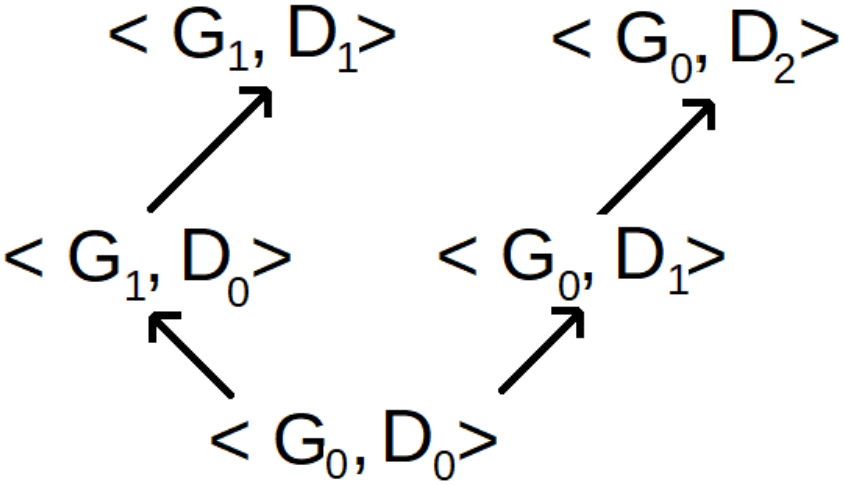


K – Anonymity: Incognito Algorithm

I) Define domain hierarchies



II) Construct directed graphs for all generalizations



III) Iterate through a queue, ordered by height in tree, check for k-anonymity

$< G_0, D_0>$ $< G_1, D_0>$ $< G_0, D_1>$ $< G_1, D_1>$ $< G_0, D_2>$

Part II: Privacy in Data Analysis

Why do we need privacy in analysis?

Differential Privacy:

"The data set with the best privacy (for you) is the one you aren't in. "

Current Work in Diff. Privacy:

Google:

- RAPPOR (2014)
- Deep Learning w/ DP (2016)
- Deep Learning with Private Data (2017)
- TensorFlow Privacy (Python ML with DP, git repository)

Apple:

- Learning with Privacy at Scale (2018)

Locally Private Analysis:

- Locally Private Estimators (Jordan et. al)

Differential Privacy

Definition 1: Let D_1 & D_2 be two datasets that differ on at most one element, and let $\mathcal{M}$ be a randomized algorithm and S be a subset of the outcome space of $\mathcal{M}$. Then for a real number $\epsilon \geq 0$, $\mathcal{M}$ is said to be ϵ-differentially private if

$$P\Big(\mathcal{M}(D_1) \in S\Big) \leq e^{\epsilon} P\Big(\mathcal{M}(D_2) \in S\Big)$$

- Stochastic Method, acts on the *algorithm*
- Goal: Make it so that the outcome of the algorithm is "relatively unaffected" by the inclusion of any specific data point (provide "plausible" deniability).
- Extension: (Eps, delta)
- Common Method: Add noise to the query or statistical computation (Laplace Mechanism)

Differential Privacy

Definition 1: Let D_1 & D_2 be two datasets that differ on at most one element, and let $\mathcal{M}$ be a randomized algorithm and S be a subset of the outcome space of $\mathcal{M}$. Then for a real number $\epsilon \geq 0$, $\mathcal{M}$ is said to be ϵ-differentially private if

$$P\Big(\mathcal{M}(D_1) \in S\Big) \leq e^{\epsilon} P\Big(\mathcal{M}(D_2) \in S\Big) + \delta$$

- Stochastic Method, acts on the *algorithm*
- Goal: Make it so that the outcome of the algorithm is "relatively unaffected" by the inclusion of any specific data point (provide "plausible" deniability).
- Extension: (Eps, delta)
- Common Method: Add noise to the query or statistical computation (Laplace Mechanism)

Basic Differential Privacy: Sensitivity & The Laplace Mechanism

Definition 2: Let f be a function operating on a dataset D. The l_2 sensitivity, Δf of f w.r.t D is the quantity

$$\Delta f = \sup_{D'} \|f(D) - f(D')\|_2$$

where D' is a dataset that differs from D in a single row.

Theorem 1: Let f be a function operating on a dataset D as above, and let $Z \sim Laplace\left(\frac{\epsilon}{\Delta f}\right)$. Then $f' = f + Z$ is differentially private in the sense of (1)

Privatized SGD for Binary Classification Model

Model:

$$E(Y;\theta|\mathbf{X}) = g^{-1}(\theta^T\mathbf{X})$$

$$\log\left(\frac{p}{1-p}\right) = \theta^T X$$

$$p = \frac{1}{1+e^{\theta^T\mathbf{X}}} =: \sigma(\theta^T\mathbf{X})$$

Parameter Estimation (MLE)

$$l(\theta, y) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x^i)) + (1-y_i)\log(1-\sigma(\theta^T x^i))$$

Gradient Descent:

$$x_{i+1} = x_i - \eta \nabla f(x_i)$$

Privatized SGD Optimizer:

Algorithm 4: Differentially Private SGD (DPSGD)

Input: Data set of observations $X \in \mathbb{R}^{n \times k}$
Parameters: Number of iterations T, privacy parameters ϵ, δ, norm bound C, step size η, lot size L

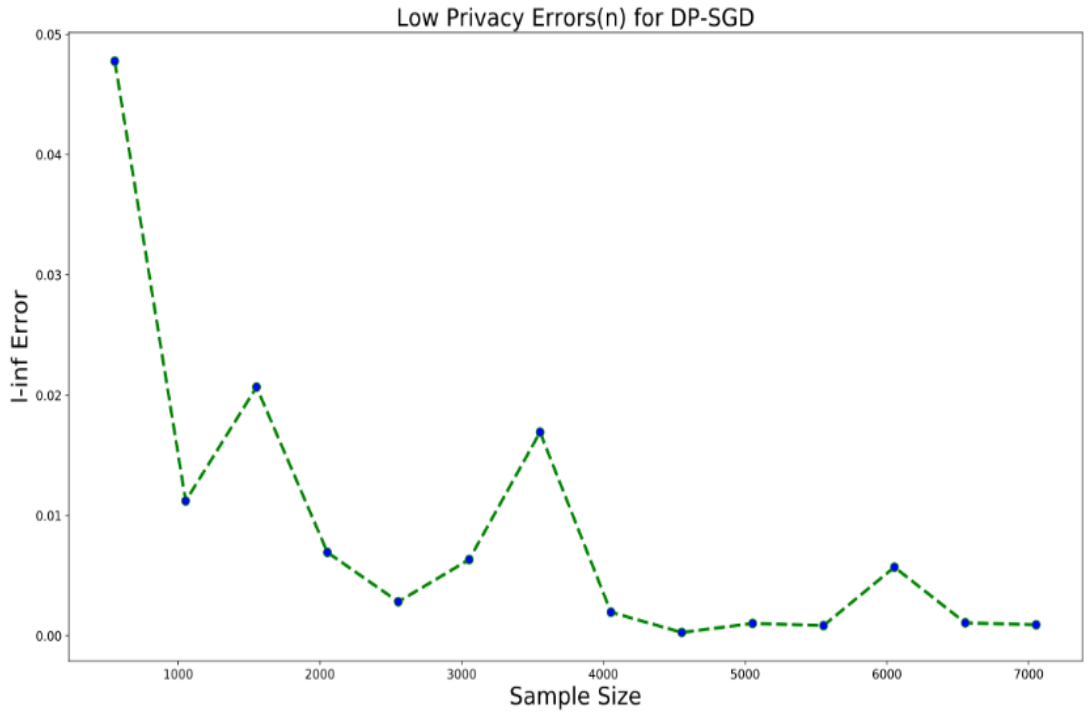
1. Initialize θ_0 randomly.
2. Compute the noise factor $\sigma = \sqrt{2\log(1.25/\delta)}/\epsilon$
3. For each $0 \leq t \leq T$
   (a) Sample L rows from X uniformly into one batch $x_t = \{x_1, x_2, ..., x_L\}$
   (b) Compute the gradient $g(x_t, \theta_t) = \nabla l(x_t, \theta_t)$
   (c) Clip the gradient: $G(x_t) = g(x_t)/\min(1, \frac{\|g(x_t)\|_2}{C})$
   (d) Add noise: $\hat{G}(x_t, \theta_t) := G(x_t, \theta_t) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}_k)$
   (e) Perform the update:
   $\theta_{t+1} = \theta_t - \eta \hat{G}(x_t, \theta_t)$
4. Return θ_T as the minimum

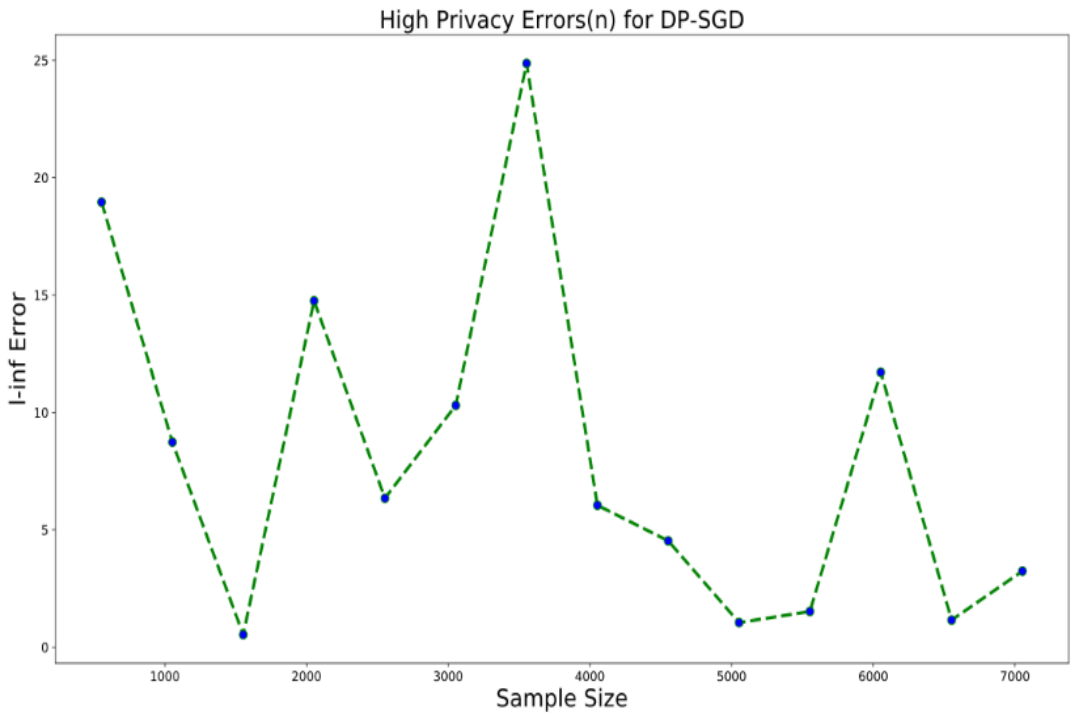
- Privacy parameters: eps (or sigma), delta
- Noise may be generated ahead of runtime -> Save computational effort
- Gradient clipping bound taken as C = median(g(x_t,theta_t))

Google Research, *Deep Learning With Differential Privacy*, 2016

Case Study I:

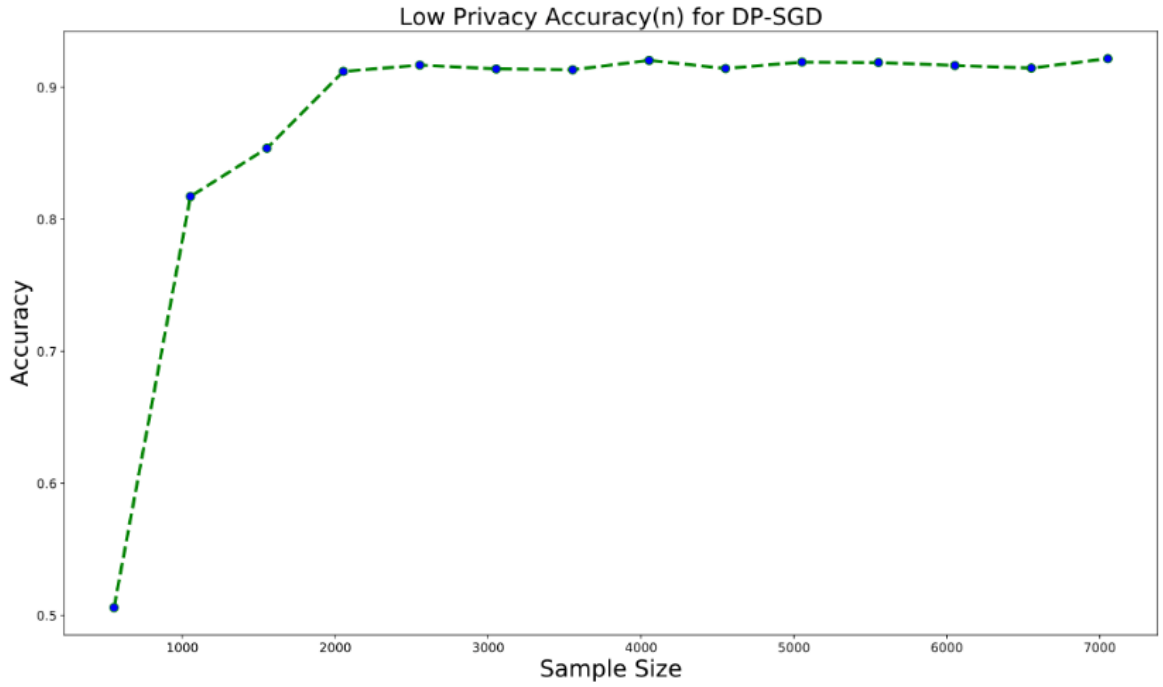
- Y ~ (X1, X2), graphs show convergence of l_inf error in parameter estimates, averaged to account for stochastic noise
- Total data set n = 10,000, each subset 70/30 split for training/testing
- High privacy requires rougly 10 times larger sample sizes for comparable error size

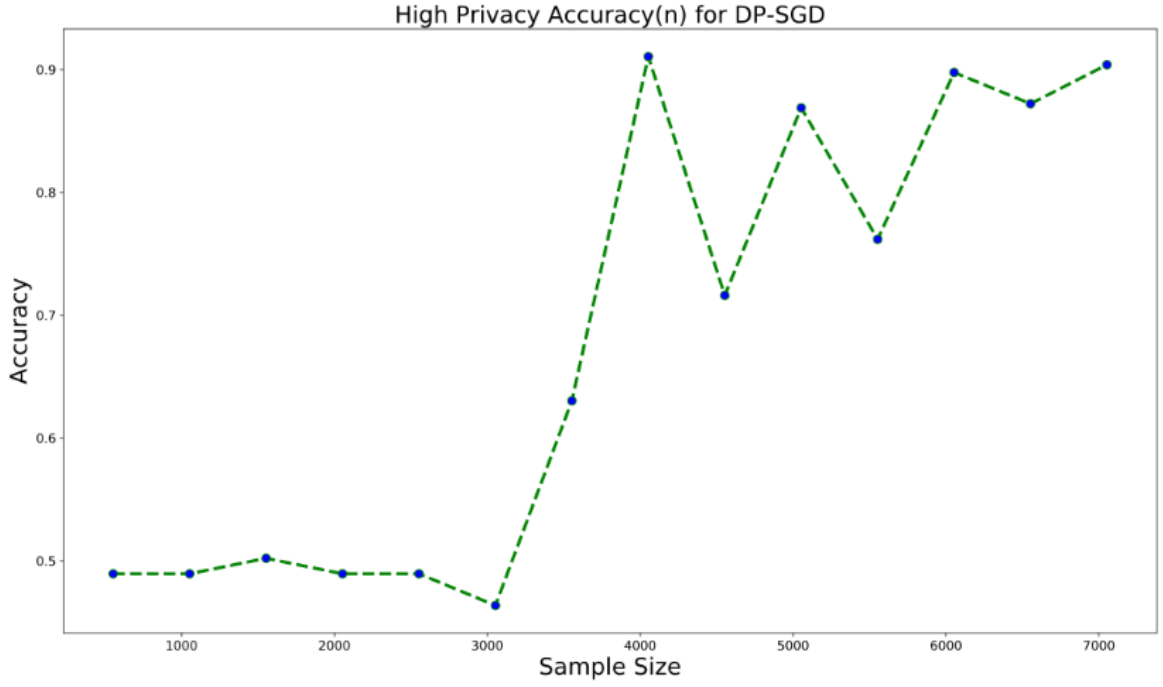




Case Study I: Accuracy(n)

- Predictive accurary of the model as a function of sample size
- Low privacy reaches >90% predictive accuracy for ¼ of the sample size of high privacy at best





Locally Private Analysis

- Min/Max Optimal Procedures for Locally Private Estimation
  Duchi, Jordan et. Al 2017
- Stochastic Method, acts on the *data*

Definition 3: Let $X_i \in \mathcal{X}, Z_i \in \mathcal{Z}$ be random variables and $\sigma(\mathcal{Z})$ be an suitable sigma-field on $\mathcal{Z}$. For a given privacy parameter α, Z_i is said to be an α-locally differentially private view of X_i if

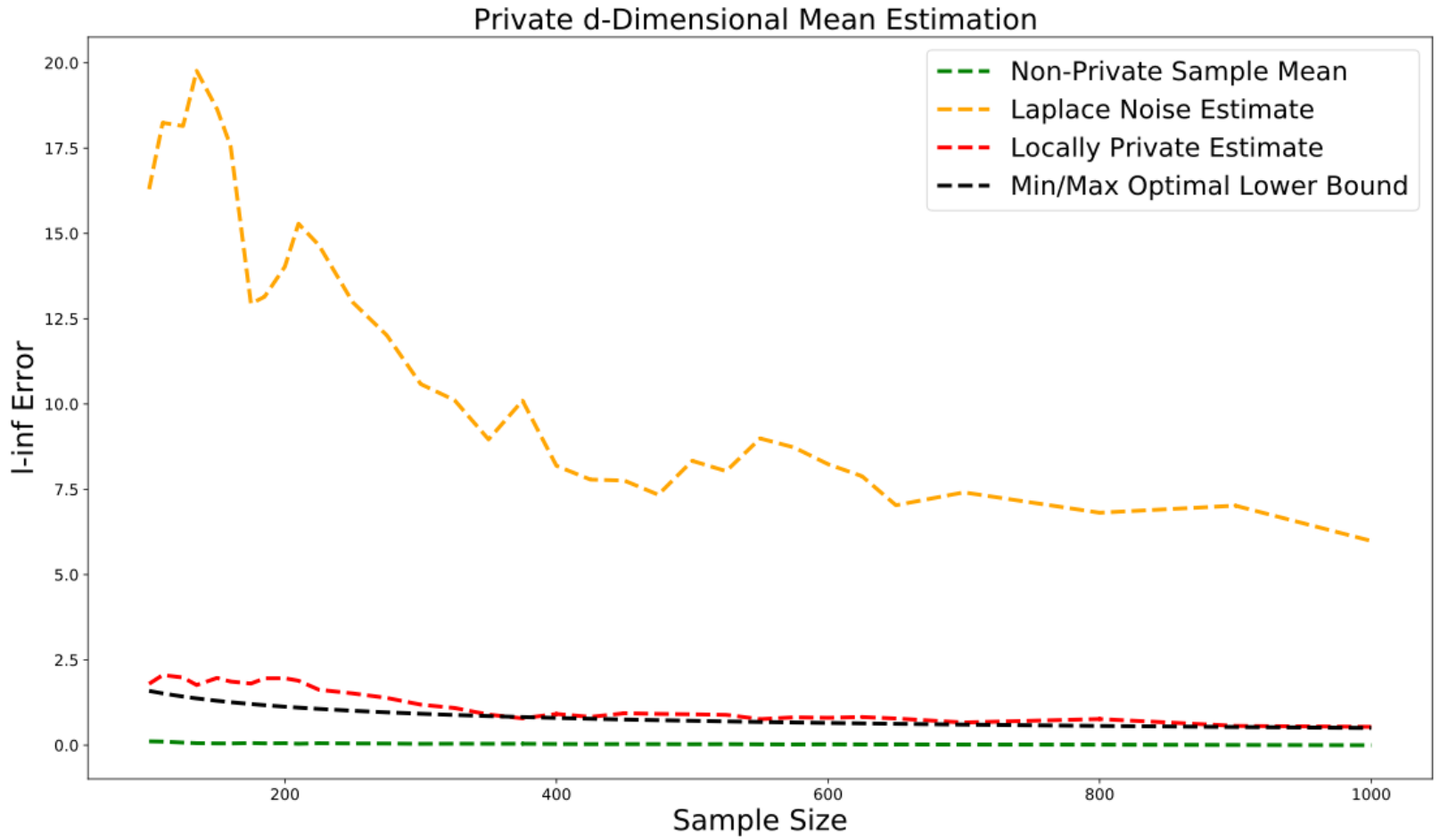
$$\sup_{S \in \sigma(\mathcal{Z})} \frac{Q_i(Z_i \in S | X_i = x, Z_1 = z_1, ..., Z_{i-1} = z_{i-1})}{Q_i(Z_i \in S | X_i = x', Z_1 = z_1, ..., Z_{i-1} = z_{i-1})} \leq e^{\alpha}$$

for all $z_1, ...z_{i-1}$ and $x, x' \in \mathcal{X}$.

- Privatization procedures must be customized for each statistical method
- Methods developed for mean estimation, medians, glm, density estimation
- In general, involves quite complicated sampling procedures

Example: Locally Private Estimation

- Multidimensional Mean Estimation (d=11)
- Graph shows l_inf error in estimation as a function of sample size



- Locally Private Min/Max Optimal Estimators are significantly more accurate than the laplace mechanism

Locally Private SGD for Binary Classification Model

Algorithm 5: α-Locally Differentially Private SGD (LPSGD)

Input: Data set of observations $X \in \mathbb{R}^{n \times k}$
Parameters: Number of iterations T, privacy parameters α, step size η, lot size L

1. Initialize θ_0 randomly.
2. For each $0 \leq t \leq T$
   (a) Sample L rows from X uniformly into one batch $x_t = \{x_1, x_2, ..., x_L\}$
   (b) Compute the gradient estimate $g_t(x, \theta_t) = \nabla l(x_t, \theta_t)$
   (c) Apply the ℓ_∞ sampling procedure Q_α to the gradient estimate

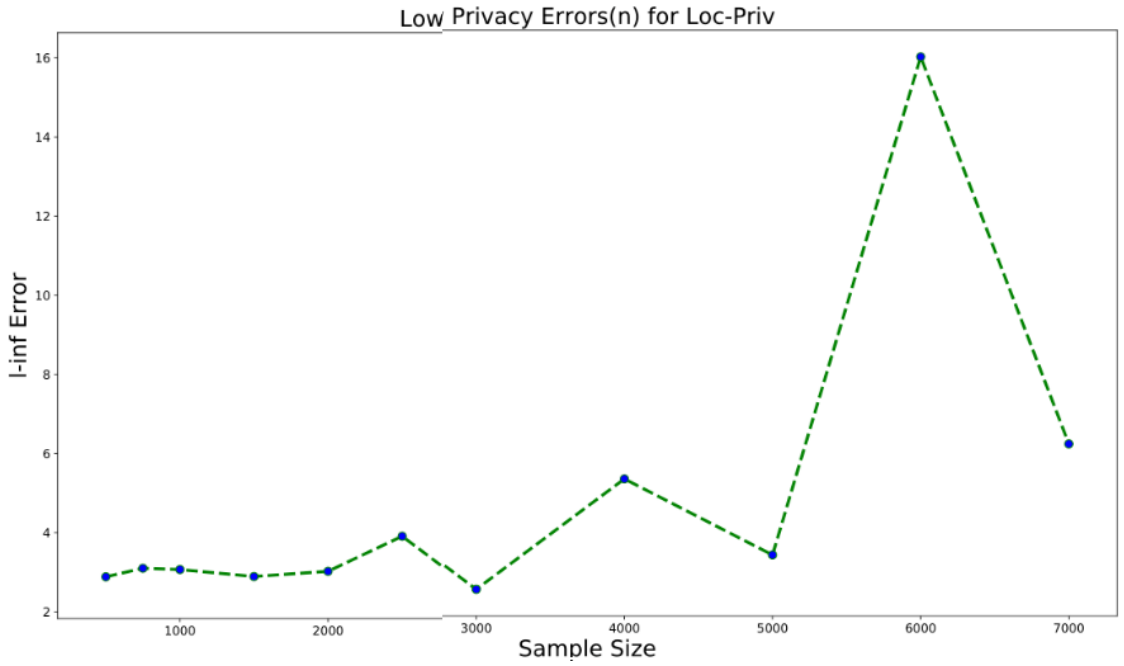
$$G(x_t, \theta_t) := Q_\alpha\Big(g_t(x_t, \theta_t)\Big)$$

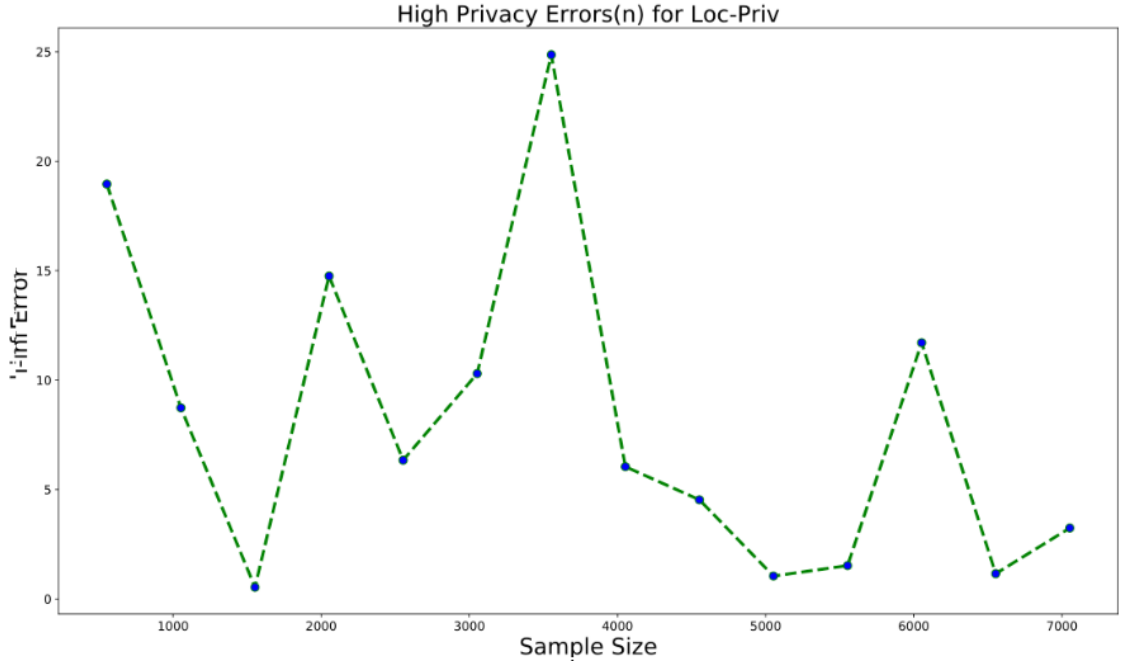
   (d) Perform the update:
   $\theta_{t+1} = \theta_t - \eta G(x_t, \theta_t)$
3. Return θ_T as the minimum

- Single privacy parameter alpha
- Resampling the gradient is done at each iteration –> computational cost

Case Study II: Error(n)

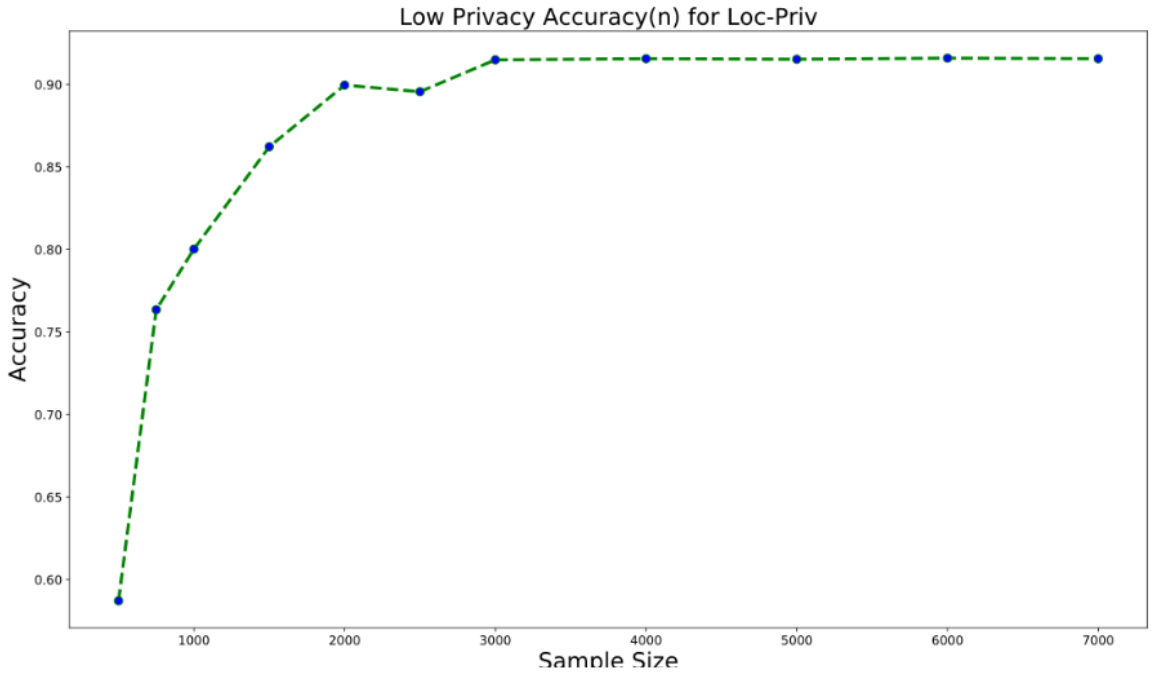
- Even for low level of privacy, errrors do not converge
- Greater effects of the stochastic resampling

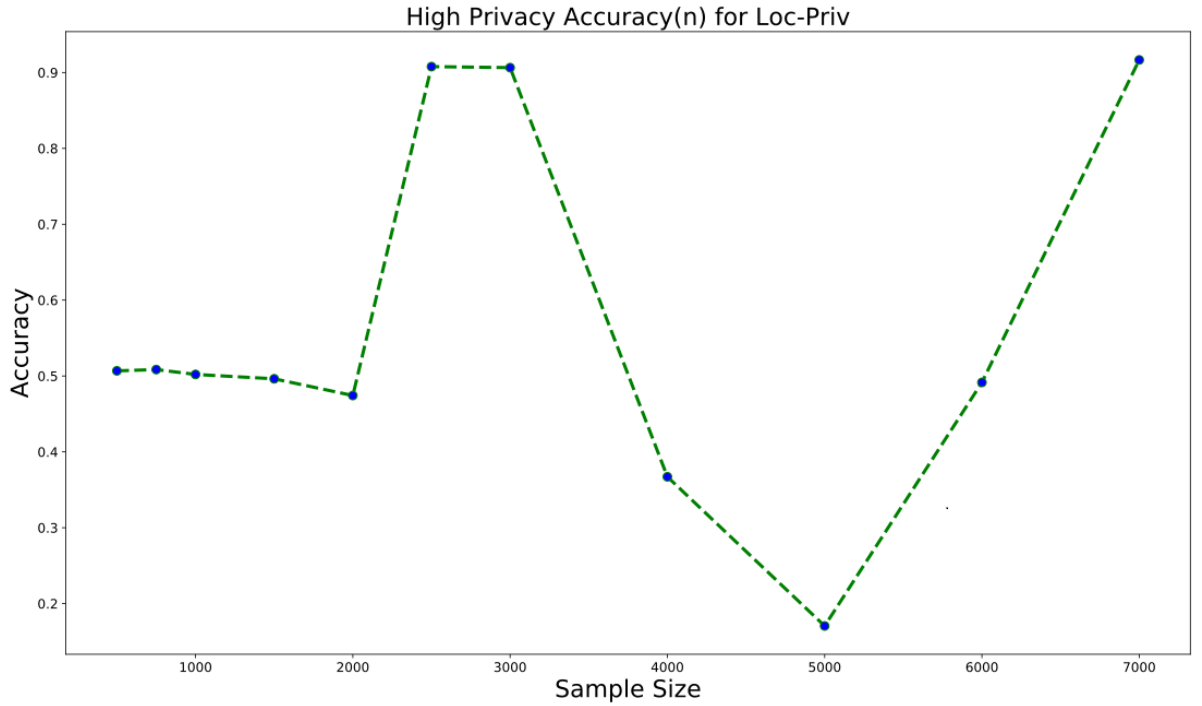




Case Study II: Accuracy(n)

- High privacy protection causes significant degredation in the predictive accuracy





Conclusion:

- GDPR compliant protection is simple with no loss of data utility; use destructive hashing with encryption
- K-anonymity is strong, but can lead to severe resolution loss, best suited for large data sets, with significant cost of implementation.
- Differential Privacy: Existing approches of varying complexity are avaliable, but for higher levels of privacy performance is greatly reduced. Sophisticated methods require significant investment in development and computation cost, but simple applications are straightforward.

Databricks notebooks are available, with guides and implementations

christoffer.langstrom.88@gmail.com

Thank You!