Private mouse and keyboard behavioral data

Final presentation – Bachelor Thesis

Hossam Elfar

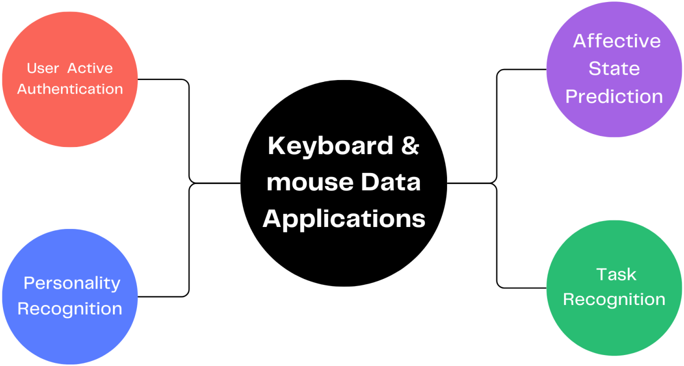
Supervisors: Guanhua Zhang and Mayar Elfares

August 23, 2023

Perceptual User Interfaces Group, University of Stuttgart

www.perceptualui.org

Motivation - Applications



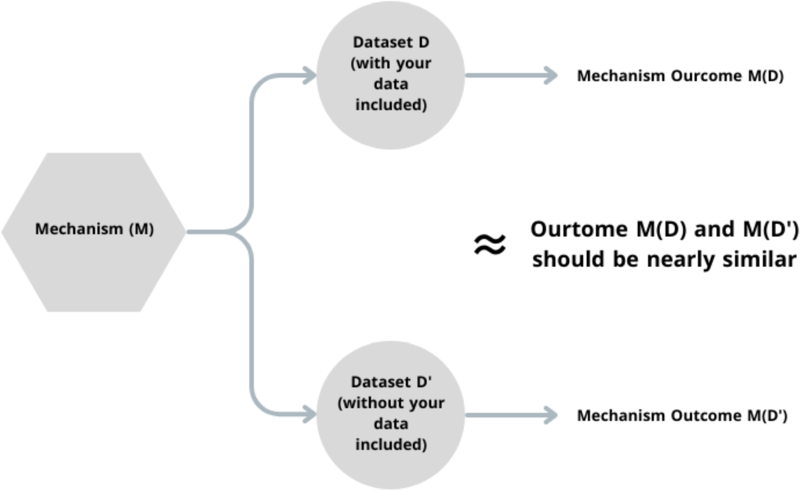
Challenge: Find a privacy-preserving mechanism that protects these sensitive datasets while maintaining their utility.

Motivation: Why is privacy important?

Mouse and keyboard data :

- can be utilized as:
    1. Biometric features, like fingerprints and eye prints [4].
    2. Affective state prediction [2].
    3. Personality recognition [3].
    4. Intent prediction [6].
- contains sensitive data such as mouse movements and keystrokes that can be used to identify tasks and anticipate user intentions.
- some datasets contain confidential information like personal messages, banking information, passwords and login credentials.

Background: Differential Privacy



Source: Differential Privacy

Background: Differential Privacy

- Differential privacy ensures statistical analysis doesn't compromise an individual's privacy [1].
- Perefect privacy is achieved when a mechanism produces indistinguishable outputs on any pair of datasets that only differ on one row.
- Howover Perfect privacy is often unattainable, but we can measure the privacy leak using the privacy parameter epsilon ϵ, where epsilon measures how much change could happen to the output.

Diffrential Privacy [1]

Definition: Algorithm $\mathcal{M}$ with domain $\mathcal{D}$ satisfies ε-differential privacy if for all pairs of adjacent datasets D and D' that differ in the data of a single individual.

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S]$$

- ε: privacy loss (small ε = stronger privacy protection)
- The inequality ensures that the probability of obtaining an output S from dataset D is approximately the same as the probability of obtaining the same output S from a neighboring dataset D', up to a multiplicative factor of e^{ε}.

[1] The algorithmic foundations of differential privacy - Dwork et al. - 2014

Contribution

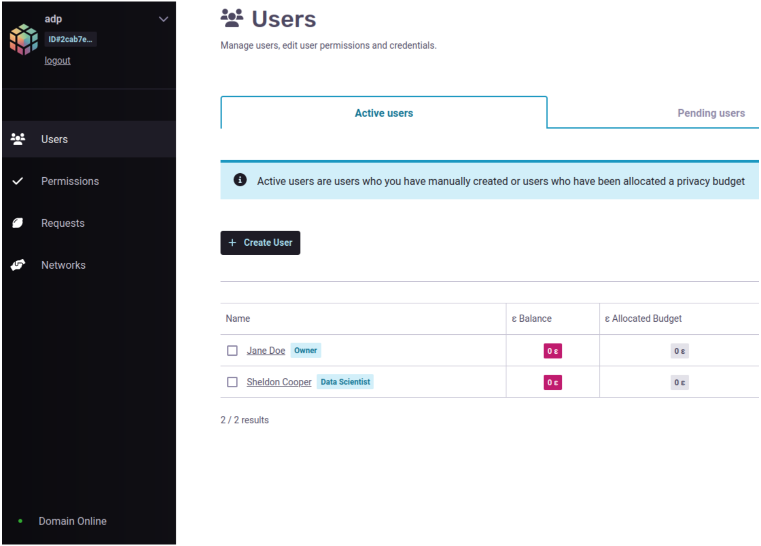
- Develop
    1. a **remote data science** technique, and
    2. a privacy-enhancing technique for **deep learning models**

  that enables data scientists to analyze behavioral mouse and keyboard data through differential privacy confidentially.
  which offers:
    - Usability
    - Scalability

Approach

Remote Data Science

Remote Data Science: Launch Domain Server



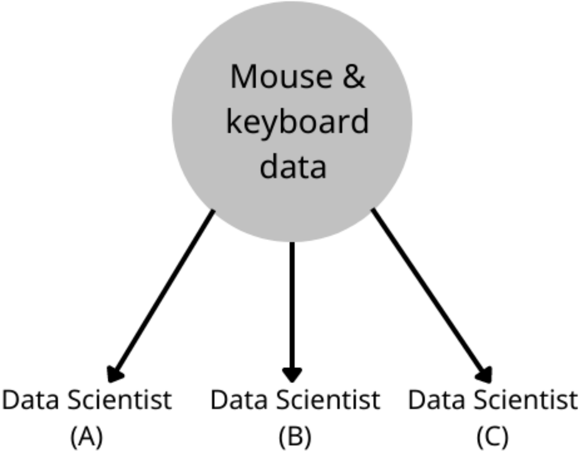
Source: Differential Privacy

Remote Data Science: Upload Dataset

- Everyday Mouse And Keyboard Interactions dataset [5]

Name	EMAKI
Users	39
Data	1.2M Mouse data, 210K Keyboard data
Tasks	Text Entry & Editing, Image Editing, Questionnaire Completion

Once the dataset is uploaded, we create multiple data scientists' accounts to query and process the dataset.



Source: Domain Server

Privacy in the domain server is maintained by two facts:

- Data scientists can only access the mock data and can not view the original data.
- Each data scientist receives a set amount of privacy budget to use when querying the data.

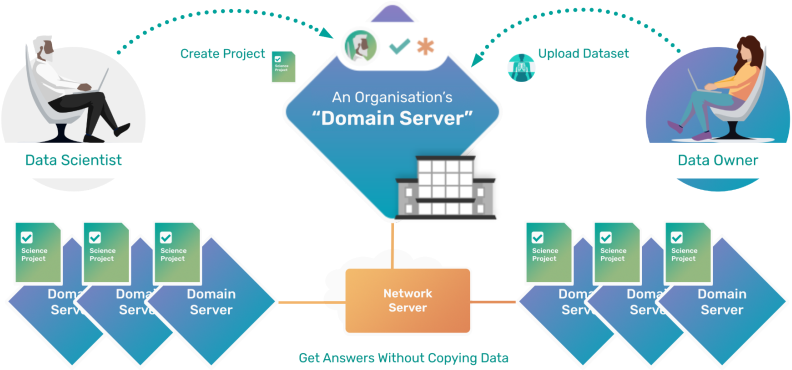
- After exploring the mock dataset, a data scientist can run his/her queries on it and then submit his/her code for review and approval before running it on the real dataset.
- After the data owner accepts the query, an amount of the privacy budget is deducted, and the results are sent back to the data owner.

How is a privacy budget deducted?

$$F(x) = f(x) + \mathrm{Lap}\left(\frac{s}{\varepsilon}\right)$$

- s: the sensitivity of the query.
- ε: the privacy loss.
- $\mathrm{Lap}(x)$: a sample from the Laplace distribution with scale parameter x.

Remote Data Science: Setup



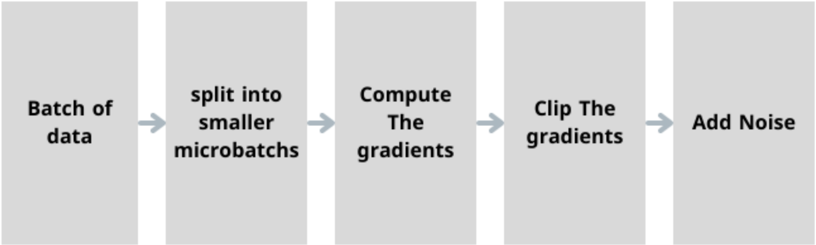
Source: Differential Privacy

Approach

Remote Learning

In deep learning, we achieve differential privacy with differential private stochastic gradient descent (DP-SGD), by adding noise to the gradients so that each data entry (individual's data) has plausible deniability.

Setup

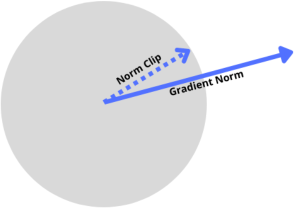


Source: Setup

Two modifications were added to the normal vanilla SGD optimizer:

1- Gradient Clipping: The sensitivity of each gradient needs to be bounded so that each data entry contributes to the model by a certain amount.

Through experiments, we've found numbers from 0.5 to 1.5 is working reasonably well and provide a good privacy level.

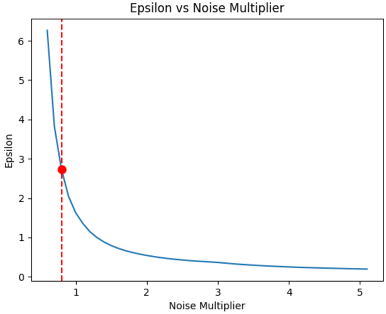


Source: Gradient Clipping

Noise multiplier

2- Noise multiplier: a sample of random noise is added to the clipped gradients to make it statically difficult to know whether or not a particular data point was included in the training set.

We have examined a wide range of noise and figured that a range of 0.8 to 4.5 provides a respectable level of privacy assurance.



Source: Epsilon vs Noise Multiplier

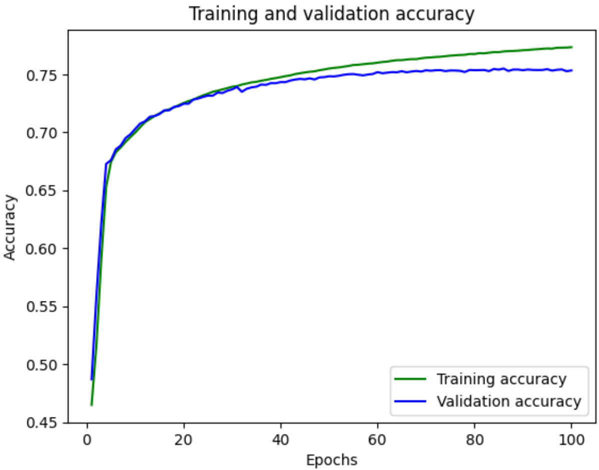
Prior to using differential privacy with DP-SGD, we first had to assess the model's accuracy using the vanilla SGD. A basic neural network was constructed using three layers :

- Bidirectional Gated Recurrent Unit (GRU).
- Dense Layer with RELU activation.
- Dense Layer with softmax activation.

Pre-processing

- Checking that no null values exist.
- Performing the train-test split.
- Performing the one-hot encoding to convert the categorical features into binary features.

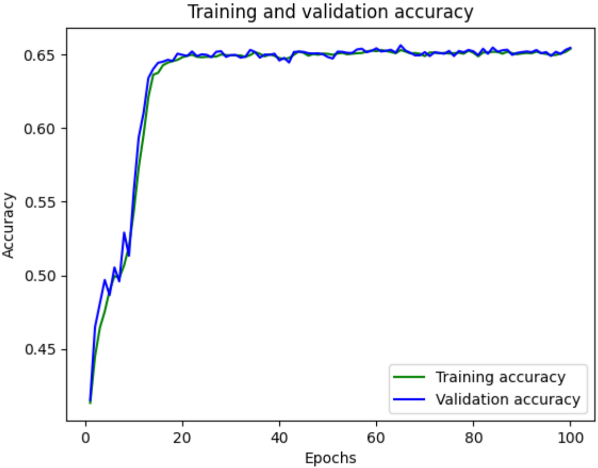
We can reach an accuracy of 75.4% in about 100 epochs without any privacy modification.



We have classified the results from the model with the DP-SGD optimizer into three levels:

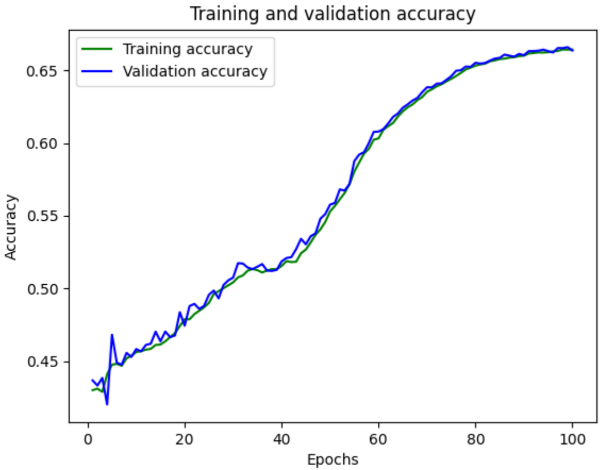
- High noise with epsilon $\epsilon = 0.57$.
- Medium noise with epsilon $\epsilon = 1.75$.
- Low noise with epsilon $\epsilon = 4.48$.

$\epsilon = 0.57$, Norm Clip $= 1.1$, Noise Multiplier $= 5.8$, Accuracy $=$ 65%



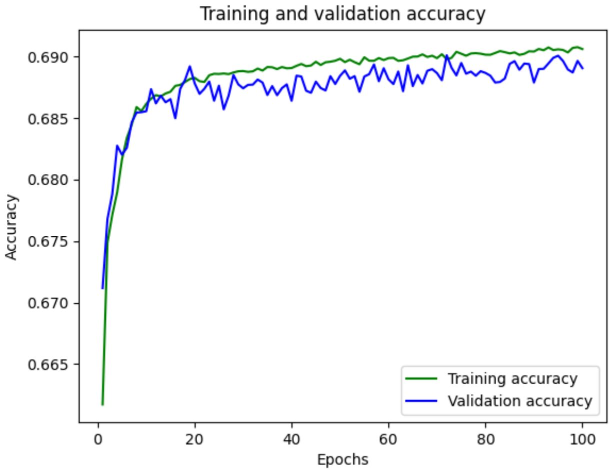
Medium noise

$\epsilon = 1.75$, Norm Clip $= 1.2$, Noise Multiplier $= 2.1$, Accuracy $=$ 65.8%



Low noise

$\epsilon = 4.48$, Norm Clip $= 1.5$, Noise Multiplier $= 1.1$, Accuracy $=$ 69.1%



Results Summary

Source: Comparison of Three Classes

	Features					
	Batch Size	Epochs	Norm Clip	Noise Multiplier	Epsilon	Accuracy
No Privacy	128	100	-	-	-	75.4%
High Noise	128	100	1.1	5.8	0.57	65%
Medium Noise	128	100	1.2	2.1	1.75	65.8%
Low Noise	128	100	1.5	1.1	4.48	69.1%

Limitation & Future Work

- One limitation pertains to computation cost. This encompasses the time needed for computations.
- Explore the potential of our technique on more interactive behaviour modalities like gaze and eye tracking data.

[1] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[2] Anis Elbahi, Mohamed Ali Mahjoub, and Mohamed Nazih Omri. Hidden markov model for inferring user task using mouse movement. In *Fourth International Conference on Information and Communication Technology and Accessibility (ICTA)*, pages 1–7. IEEE, 2013.

[3] Clayton Epp, Michael Lippold, and Regan L Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 715–724, 2011.

[4] Issa Traore, Isaac Woungang, Mohammad S Obaidat, Youssef Nakkabi, and Iris Lai. Combining mouse and keystroke dynamics biometrics for risk-based authentication in web environments. In *2012 fourth international conference on digital home*, pages 138–145. IEEE, 2012.

[5] Guanhua Zhang, Matteo Bortoletto, Zhiming Hu, Lei Shi, Mihai Bâce, and Andreas Bulling. Exploring natural language processing methods for interactive behaviour modelling. *arXiv preprint arXiv:2303.16039*, 2023.

[6] Yinghui Zhao, Danmin Miao, and Zhongmin Cai. Reading personality preferences from motion patterns in computer mouse operations. *IEEE Transactions on Affective Computing*, 13(3):1619–1636, 2020.

Thank you!