Good morning, and welcome to my bachelor midterm presentation. I am "hossamelfar," and today I will be discussing the topic of private mouse and keyboard data. To provide an overview, let's begin with the table of contents. Firstly, we will have a recap of the project, including its motivation and approach. Then, I will delve into the progress made and discuss what has been accomplished so far. Afterwards, I will outline the next steps we have planned. Finally, I will conclude by presenting the project's schedule.

Let's begin by discussing our motivation for this project. It is important to note that behavioural datasets, such as mouse and keyboard data, can contain sensitive information about individuals. Some examples of this sensitive information include:

1. Passwords and login credentials: Mouse and keyboard data can potentially capture users' input while they enter their passwords and login credentials, which are highly sensitive and must be protected.
2. Personal messages and communication: User interactions with their keyboards and mice can involve typing and clicking during personal messaging or communication activities. These interactions may contain private conversations that should be kept confidential.
3. Banking information: Users often utilize their keyboards and mice to perform online banking activities, such as entering credit card numbers, account details, or conducting financial transactions. This information needs to be safeguarded to prevent unauthorized access and financial fraud.

Absolutely, keyboard and mouse data can be vulnerable to attacks that potentially expose information about individuals in the dataset. Key frequency analysis is one such technique that can be used to extract insights from the data. By counting the occurrence of each key press or key combination, it becomes possible to identify frequently used keys or shortcuts. This analysis can be performed by creating a frequency distribution or calculating the percentage of occurrence for each , and help determine commonly used words, phrases, or even passwords

Approach :

Moving to our approach, we have chosen to employ differential privacy as our privacy-preserving mechanism. Differential privacy is a concept that aims to protect individual records from being identified while still allowing for analysis and learning from the overall population behaviour.

Differential privacy provides a rigorous mathematical framework that quantifies the privacy guarantees of a given algorithm or mechanism. It allows for a fine balance between data utility and privacy protection, ensuring that meaningful insights can still be derived from the dataset while minimizing the risk of exposing sensitive information about individuals.

the formal definition of differential privacy states that for any two neighboring datasets, which differ in only one data point, the probability of the outcome of a given mechanism (M), which can be any computation or query performed on the dataset, should be bounded. This bound is determined by the privacy loss parameter (epsilon), where a smaller epsilon guarantees higher privacy.

To elaborate further, if we denote the outcome of mechanism M on dataset D as M(D) and the outcome on a neighboring dataset D' as M(D'), differential privacy ensures that the probability of M(D) and M(D') deviating significantly is limited. The level of deviation is controlled by the privacy loss parameter (epsilon). When epsilon is set to 0, the outcome of M(D') should be exactly the same as the outcome on D, providing perfect privacy. However, it is important to note that setting epsilon to 0 may lead to reduced utility or information loss, as the noise added to preserve privacy can affect the accuracy or fidelity of the results.

To summarize, differential privacy is a quantitative mechanism that ensures the presence or absence of an individual record in a dataset does not significantly impact the outcome of a given mechanism or computation. In order to achieve differential privacy, algorithms need to introduce some form of noise or randomness to the outcome of the mechanism, thus preserving privacy while allowing for meaningful analysis.

One commonly used method for adding noise is the Laplace mechanism, which is also employed in our project's framework. The Laplace mechanism adds random noise following the Laplace probability distribution to the outcome of the mechanism. The amount of noise added is determined by the privacy loss parameter (epsilon) and the sensitivity of the computation being performed. The sensitivity refers to how much the output of the mechanism can change when a single record is added or removed from the dataset.

In terms of progress, we have used the pysyft libray whih provides secure and private remote Data Science in Python . the concept of remote data science is based on the idea that our ability to answer important queries or teaching deep learning models is limited because we can not access exisiting data of several organization because these data are very private and theses oragnizations have many concerns to share these data with you so that I don't misuse this data . Remote data science is composed of three main components , domain server which imitated a gpu server that holds our dataset but this time in a secure way which allow data scientist to perform their computation remotly in our server without seeing the actual data or acquiring a copy of this data , A data owner which represent the holder of mouse and keyboard data and the data scientits or the end users who wants to perform computations on our private datasets . but it is clear that doing our computation remotly in the organizations domain server will not compromise our data but what a bout the results leaving the organization, here comes dp where each data owner when they join the domain server we assign to him an inital privacy budget the he can uses to perform queries on the dataset and every time and based of the amount of the noise added to his query and how much budget he has a code reveiwer from the domain server should accept or reject the execution of the query and if accepted an amount of his privacy budget is detucted .

but how can a data scientist do good data science without acutually seiing the data here we introduced a mock data which is imitiating the real data