Behavioural data science is a unique field that combines principles of statistics and computer science to analyze and understand human behaviour using data. It involves collecting, processing, and analyzing large datasets related to human behaviour to identify patterns and insights that can inform decision-making and improve outcomes. The ultimate goal of behavioural data science is to use data-driven insights to better understand human behaviour and predict future actions. However, the data used in this field can often be private and require protection.

Keystroke dynamics and mouse movements are effective behavioural biometrics for data analysis as they provide useful and highly sensitive data, such as usernames, passwords, banking information, or text messages, which can be used for active authentication and predicting user intents. However, little attention has been paid to the privacy of collecting and transmitting keyboard and mouse data. Standard methods suffer from various vulnerabilities, which can allow attackers to steal personal and organizational information and even the victim's identity.

One of these methods is the anonymization technique, which involves passing the dataset through a mechanism that deletes some layers or fields that identify the personality of people participating in the dataset. The resulting de-identified dataset is then given to the analyst to run their models. However, the de-identified data is often not de-identified or not private, as demonstrated by the Netflix Prize example found on the internet. Therefore, it is crucial to develop more robust privacy protection methods to ensure the confidentiality and integrity of behavioural data in the field of data science. even though it seems pretty anonymous but it was not 2 computer scientists from the university of texas have published a paper that they have successfully revealed the users' identities by combining these data with data from IDMB, this type of attack is called a linkage attack.

Say we are training model on Mouse dataset of users moving their mouse on the screen and we need to predict if a certain user had actually move their mouse to that position on the screen and this model has been trained and comes and said that 55 precent confident that john has clicked on that position now say we took the same model and trained it on the same dataset but plus john's data and comes back and said actually I am 57 precent that johns have cancer that not necc means that johns has cancer , it may be that the machine learing that oh has more data and becomes more confident about the outcome however if we come again the datset with johns data and the machine learning model said that 88 precent john has clicked that is pretty obivious that john has actually clicked and there is a privacy leak about , the key thing to know is that we actually leaking information about a particulat data record everytime a particular otucome becomes likely or unlikely the original outcome when removing or adding a particual record and we can quantify this privact loss by comparing the ratios of the log of the two probalabilties either with the record included or removed and we call this ratio the privact budget and differintial privacy aims to limit this privacy loss ,

however, if we need to reach the peek of privacy we can set this epsilon of the privacy budget to 0 but reaching zero is not that easy we need a lot of obfuscation and noise added to afford that dataset performs the same with my record included or not.