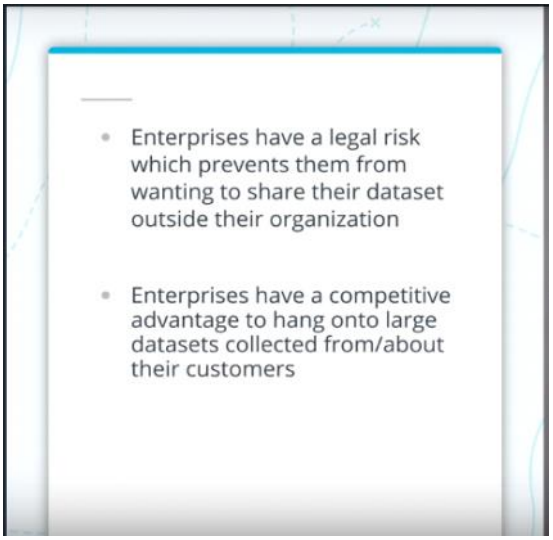
Introducing Differential Privacy

03 July 2019 22:42

Data privacy is a very important concern for government and companies these days. On one hand we are moving forward with having high- quality data available and on the other hand it is also necessary to keep data safe from both intentional and accidental leakage.
In real world most datasets are isolated within large enterprises for two reasons:

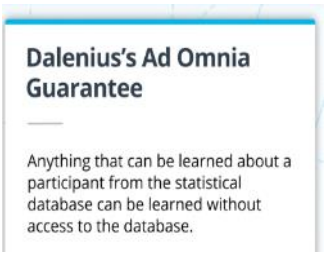
The more personal the data is and the more potential the use of data in society is,the more restricted it is for use. So in this whole course we are going to over about how to build secure models using state of the art techniques like differential privacy and federated learning.

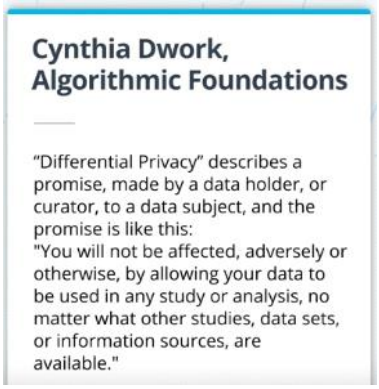


What is Differential Privacy(DP)?
According to Wikipedia Differential privacy is a constraint on the algorithms used to publish aggregate information about a statistical database which limits the disclosure of private information of records whose information is in the database.
"Differential privacy" describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available."

The General Goal of DP is to ensure that different kinds of statistical analysis don't compromise privacy. So statistical analysis in most general form means that we have some training dataset or dB or some individual data and we want to make sure that the statistical analysis that we did does not compromise on the privacy of any particular individual contained within that dataset.

What is Privacy?





Can we Anonymize Data?

Even if you do great job of anonymizing the data and if someone else releases a related anonymized private dataset often it is possible to divulge the private aspects of information you were trying to hide by studying these two datasets.
One Example is that of Netflix best Recommendation engine competition in which even though the data used was anonymized with some integer values and also names of the movies were not disclosed but despite that a paper was published in which they were able to de-anonymize both the names and movie names they did it by first scrapping the IMDB review site and then use statistical analysis to find individuals who were rating at both IMDB and on Netflix this allowed to de-anonymize large percentage of users on Netflix as well as the title of the movies they were watching.[1]
Another Example was looking at multiple separate anonymized datasets as well as online voter registration records(public) which lead to re-identification of the medical records of the governor of Massachussetts.[2]

Key idea of Differential Privacy is the ability to ask this question: "When querying a DB if I removed someone from DB, would the output of query be any different?". In order to do this we construct parallel databases which are simply databases with one entry removed. Basically we are going to construct a query which doesn't changes no matter who we remove from the database.
Now for the first coding challenge we learn how to generate parallel databases. For this refer to my Differential Privacy repository on Github[3]