Shia Aaron Lloyd Fisher

Paul J. Kelly, Tallal Ahmad, Yinka Ajiobola

Data and Algorithms: Ethics and Policy - Memo 4

14 July 2023

Individuals' privacy: differential privacy

Differential privacy is the concept of having output aggregate data remain anonymous. That is, readers of the data can glean information about a population, but it is not possible to use the information in any way the reveals an individual who participated in the input data that generated the output data. This approach addresses the wary of big data containing identifiable information. Identifiable information within output data does not need to be negative to be considered problematic from a deontological view. It only needs to be information which is shared without the explicit informed consent of the individual since this scenario would harm said individual by not respecting their autonomy.

Consider the example of the doctor from the 1970s British Doctors Study mentioned in Kearns et al. where Roger, the doctor, was re-identified as a smoker, as it was possible to reverse engineer the output data in the sense that knowing how the algorithm was designed one could backtrack to find the participant. To discuss the harm here let us suppose this doctor would have a more successful practice if their patients believed the doctor led, by example, a healthy lifestyle. It is a reasonable premise; however, the utilitarian believes the ends do justify the means, thus no differential privacy would be granted.

The implementation of differential privacy is intentional since it requires some mathematical principles to effectively map input to output data, treating the input stream with some Bayesian model in such a way that introduces some randomness in the output stream. The purpose of this is to remove the identifiable traits from the input completely. Another example was used to illustrate when this carefulness is not adhered to. Kearns et al also demonstrated that by having known a person's zip code, gender, and age one could, in theory, pretty closely reidentify an individual if this information is not dereferenced in the aggregated output.

Differential privacy is a necessary requirement in the deontological ethical framework for handling BigData. It has support from scientists who are simply interested in the findings of data rather than the individuals whose information was used to generate the data. It should be possible to remove a participant from the input stream without significantly changing the output.