What is Fair Data Processing ?

Benjamin Nguyen*
INSA Centre Val de Loire, Bourges, France – benjamin.nguyen@insa-cvl.fr

Abstract

Current data protection laws in France closely scrutinize personal data processing. Indeed, in the case of such a process many constraints apply: data collection must be limited, retention limits are imposed, and more generally, the processing must be *fair*. Conversely, such constraint do not exist if the data is anonymous (i.e. it is not possible or at least very difficult and costly to link a data item to a real individual) – again this can be viewed as *fairness*, since anonymous data is by definition harmless for the individuals concerned. However, data anonymization is still an open problem. Many state of the art anonymization techniques used in statistics (such as pseudonymization, or k-anonymization) cannot be mathematically proven to have any formal guarantees. Other techniques, such as differential privacy, although able to provide these guarantees, are on the contrary difficult to use in practice, and difficult to understand by the general public. Another field investigated is cryptographic techniques, which could fully enable private data processing, such as *fully homomorphic encryption*. For the moment (and for a foreseeable future), these techniques are not efficient enough to be used on *Big Data*.

Thus the question of fair data processing remains open: is anonymization a good road to follow? Shouldn't other aspects also be considered, such as the concepts promoted by the privacy field, such as openness, user control, auditability, etc? Finally, how should we design algorithms to run on Big Data be used in order to be fair?

Keywords: privacy, anonymization, fair data processing.