When Synthetic Data Met Regulation

Georgi Ganev[1,2]

University College London[1], Hazy[2]

Problem description

Main Question: Can we make synthetic data regulatory compliant?

Namely, we explore the legality of privacy-preserving synthetic data created by generative machine learning models trained on structured personal data.

Regulatory Definitions

Personal Data: ”Any information relating to an identified or identifiable living individual.”

Sufficient Anonymization: However, information that is effectively anonymized is not personal data and data protection law does not apply to it.

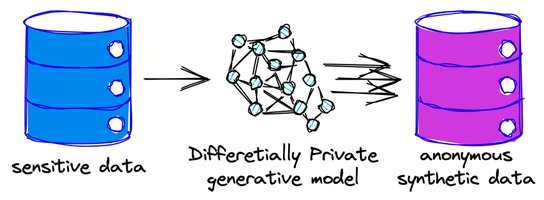
When assessing re-identification, the focus should be on objective factors such as cost/time, available technologies, and their developments over time.

Technical Tests: The following three key risks need to be reduced for sufficient anonymization (these risks should be looked through the *motivated intruder* test):

1. (singling out) any individual being isolated;
2. (linkability) any records/datasets (publicly available or not) being combined with synthetic data and thereby enabling the identification of an individual;
3. (inferences) an attribute being deduced with significant probability from the values of other attributes.

Synthetic Data as Anonymous Data

Combining **Generative Models** and **Differential Privacy (DP)** reduces all identifiability risks to sufficiently remote level and, therefore, the resulting data can be considered anonymous.



Generative Models: break the 1-to-1 mapping and to an extent reduce singling out and linkability but could be susceptible to various privacy attacks.

DP mechanisms formally protect against singling out, linkability, and other re-identifiability concerns even if faced with a resourceful and strategic adversary

Potential Limitations

DP often leads to utility reduction, particularly impacting outliers and underrepresented subgroups. Selecting both the right privacy budget and DP mechanism is non-trivial and highly context-specific.

So what?

Synthetic data produced by DP generative models can be sufficiently anonymized and, therefore, anonymous data and regulatory compliant. Our work aims to establish a foundation for broader Generative AI solutions.

Full paper:

georgi.ganev.16@ucl.ac.uk

1st ICML Workshop on Generative AI and Law (GenLaw ’23), part of ICML 2023