100. Synthetic Data

Abstract

Generative artificial intelligence is a general-purpose tool with many applications, though image generators and deepfakes often take the limelight. Of particular interest to privacy scholars, however, generation can also be pointed at tabular data found in a dataset. In turn, this allows for the creation of artificially produced data—synthetic data—that could potentially be both useful and privacy-preserving (*i.e.*, a silver bullet for data anonymity). This entry will engage with synthetic data and show how it is no silver bullet, though it does have notable advantages over typical, subtractive forms of deidentification such as redaction.

Definition: At its most general, synthetic data is data that has been generated using a computer. In the machine learning context, the output of synthetic data is like "replacing the pieces of a jigsaw puzzle to create a different picture; even though all the puzzle pieces fit together in the same way (*i.e.*, each piece has similar, yet synthetic, attributes), the overall image has changed." In turn, synthetic data replaces original data and thereby permits useful data analysis without, potentially, encumbering the sensitivity of the original, raw data.

Introduction

Artificial intelligence's current summer phase predominantly rests on scale. Massive amounts of data, large models, and cutting-edge hardware have given rise to artificial intelligence performance that matches or outpaces human expertise on a variety of tasks. A crucial piece in this equation, however, may not be as straightforward as it seems. Data, in particular to machine learning, is not created equal. Some data, like a social security number, has privacy or sensitivity implications; other data, like creative works of expression, have legal implications such as copyright; and still, other data is simply low quality, whether that be because, in practice, there is not enough of it to be useful or, in reality, the data is too unique and is difficult to draw patterns from. Moreover, there exists a reasonable question of whether scale itself will plateau: Is there enough data in the world to maintain the appetite of the massive machine learning models we are creating?

What all of this pushes toward is synthetic data, a potential solution for a variety of data problems—scale, privacy, sensitivity, and legality. This entry will unpack the idea of synthetic data by providing a simple definition of the term, an overview of the history behind synthetic data, and a discussion of recent privacy scholarship related to synthetic data. The entry will end with a few comments on where synthetic data might be headed in the future.

Terminology

At the outset, it is important to note that synthetic data is a term that has at least two meanings found in the literature. The difference between these two meanings is one of process: How was the synthetic data created?

The original concept of synthetic data refers to the creation of a "dummy" dataset by cloning a "real" dataset in terms of the original dataset's statistical properties. Statistical properties, a term of art, refer to those

properties identified as important by a programmer or domain expert. For example, if you wanted to create a synthetic dataset of an array of numbers (*e.g.*, [1,2,3,4,5]) then you may use an algorithm (*i.e.*, a repeatable recipe taking input and producing output) that focuses on the replica data being dense (*i.e.*, all integers in the original appear in the replica) and random (*i.e.*, the order of the integers in the synthetic dataset should not follow commonly known patterns like "least to greatest"). Applying those two criteria to the original dataset could produce a replica, synthetic dataset (*e.g.*, [4,1,2,3,5,1]). The replica has similar statistical properties to the original and may be used in place of the original.

The newer term "synthetic data" likewise refers to the creation of a dummy dataset based on a real dataset, but the process used to create the data is through generative artificial intelligence as opposed to careful assessment. Machine learning can figure out on its own what statistical properties to focus on. All the developer needs to do is train a machine-learning model on real data with the goal of producing similar-looking, replica data. Specialized types of generative models will be most helpful for this task (*e.g.*, a generative adversarial network).

Privacy

The idea that synthetic data can be used in machine learning is appealing. What could be the downside of removing the barriers of low-quality data by replacing that data with new, high-quality data? Moreover, when most machine learning systems replicate data to create something like a synthetic dataset, the result appears, at face value, quite different from the original. This motivation and appearance-based perspective, however, elides the potential privacy concerns that synthetic data encumbers.

Similar to the failures of anonymization experienced by the AOL search-query debacle and the Netflix Prize affair, synthetic data is vulnerable to unverified claims of anonymity. In these two headliner privacy failures, AOL and Netflix made efforts to anonymize their datasets before releasing data into the wild. Each company used techniques commonly available at the time. For example, the companies would engage in something like generalization (*e.g.*, turning a column of zip codes, like 20009, into the country where the zip code is from, like "United States") or suppression (*e.g.*, replacing the last four digits of a social security number with asterisks, like 111-42-****). Indeed, these techniques make datasets appear anonymous.

Unfortunately, given the large amount of data that exists beyond siloed datasets, researchers and journalists in both cases were able to patch up the holes in the sanitized data and reidentify individuals, leading to large-scale privacy concerns. What is more, these basic hole-patching techniques have been around for a long time, as Professor Sweeney persuasively demonstrated in 1997 by reidentifying hospital records of a Massachusetts governor after the records had been "anonymized."

Similarly, synthetic data presents largely unverified claims of anonymity, though it does have notable advantages like being generative rather than subtractive (*i.e.*, synthetic data uses noise to replace data rather than remove the data entirely). In fact, computer science researchers looking into the ability of synthetic data to create privacy have identified these very concerns: Synthetic data creates "highly variable privacy gain" and "unpredictable utility loss." On top of this, there is a legal concern that "vanilla" synthetic data (*i.e.*, generative data that did not undergo any additional privacy-preserving processing like differential privacy) will be accepted as sufficient sanitization by certain data-protective statutes without a thorough analysis of the risks. In the United States, the Video Privacy Protection Act (VPPA), for example, permits data sharing even if a skilled investigator can reidentify individuals—*i.e.*, reidentification is still possible, but only with special skills. It is likely that a synthetic dataset with "variable privacy gain" leaks private information, though only someone with special skills may be able to identify a leak. Therefore, caution needs to be taken on blanket approval of synthetic data as truly anonymous; true enough, the VPPA standard may be permissible for a statute aimed at video rental history, but vanilla synthetic data likely should not be acceptable when dealing with health data or data from children. In summary, synthetic data, though additive instead of subtractive, is far from a "silver bullet" for protecting anonymity and may create a more ambiguous scenario given the difficulty of measuring privacy loss and matching that loss with a data-protective statute.

Conclusion

It is undeniable that some amount of synthetic data is useful to artificial intelligence. It is a nearly unanimous practice to use a simple type of synthetic data via single-image data transformations (*e.g.*, flipping an image horizontally or zooming in on an image) when training something like an image generator. Moreover, most production models used by large companies are trained on more advanced synthetic data given a need to balance training datasets. However, when improperly used, synthetic data makes promises it cannot keep, privacy being the predominant concern illustrated in this entry. With that in mind, researchers are starting to look into more formal guarantees of privacy using tools like differential privacy. Differential privacy, "allows one to learn the statistics of a group without also learning the statistics of the individuals making up the group." Adding differential privacy to a synthetic data pipeline seeks to offer a best-of-both-worlds approach, privacy plus utility, but inherits issues with a necessary amount of data loss (*i.e.*, utility loss). Whenever non-real data (*e.g.*, noise) is added to a dataset, there is a tradeoff that occurs between privacy and utility. A perfectly private dataset has no data in it; a perfectly useful dataset does not protect privacy in any way. The aim of future work is to better balance the privacy–utility tradeoff using synthetic data in a way that allows the data to be useful, but private.

References

[1] Meng Wang, Weijie Fu, Xiangnan He, Shijie Hao, and Xindong Wu. A survey on large-scale machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2574–2594, 2020. doi: 10.1109/TKDE.2020.3015777.

[2] Steven M. Bellovin, Preetam K. Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stanford Technology Law Review*, 22:1–52, 2019.

[3] Nathan Reitinger and Amol Deshpande. Epsilon-differential privacy, and a two-step test for quantifying reidentification risk. *Jurimetrics*, 63:263–317, 2023.

[4] Jim Gray, Prakash Sundaresan, Susanne Englert, Kenneth Baclawski, and Peter J. Weinberger. Quickly generating billion-record synthetic databases. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, pages 243–252. ACM, 1994. doi: 10.1145/191839.191886.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, 2014.

[6] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701–1777, 2010.

[7] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009. doi: 10.1109/SP.2009.22.

[8] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data—Anonymisation groundhog day. In *Proceedings of the 31st USENIX Security Symposium*, pages 1451–1468. USENIX Association, 2022.

[9] Sergey I. Nikolenko. *Synthetic Data for Deep Learning*, volume 174 of *Springer Optimization and Its Applications*. Springer, Cham, 2021. ISBN 978-3-030-75177-7. doi: 10.1007/978-3-030-75178-4.

[10] Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. Differential privacy and machine learning: A survey and review, 2014. URL http://arxiv.org/abs/1412.7584. Accessed 12 June 2024.

[11] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–487, 2014. doi: 10.1561/0400000042.