It has been shown that machine learning techniques memorize patient data and can be reverse-engineered to identify patients. This makes medical institutions wary of distributing data for researchers to train machine learning models.

Data scientists require a way to build medical machine learning models while maintaining patient privacy.

The flowchart below highlights our solution to this problem. It allows hospitals to keep their data entirely on-site and prevents machine learning algorithms from memorizing information.

We created a model to predict breast cancer under varying privacy requirements. We then compared the accuracies in the figure shown below. Note that smaller $\varepsilon$ values correspond to higher levels of privacy.

1. <akhann13@jhu.edu> Department of Biomedical Engineering, Department of Applied Mathematics and Statistics, Johns Hopkins University. 2. Department of Computer Science, Yale University. 3. Department of Biomedical Informatics, Columbia University; Core Faculty, New York Genome Center. 4. Department of Molecular Biophysics and Biochemistry, Department of Computer Science, Department of Statistics and Data Science, Yale University.