Is Private Learning Possible with Instance Encoding?
Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Abhradeep Thakurta, Florian Tramèr

What is private learning?
Goal: train a machine learning algorithm on a sensitive dataset without revealing particular details of the training data.

What is Instance Encoding?
“Encode” each image with a private encoding function, and train only on the encoded images.

Is InstaHide Private? (no)
InstaHide (ICML’20) is the leading candidate Instance Encoding scheme.

A simple attack: visual re-identification

Can Anything be Private? (no)
Informal theorem: If an encoding scheme can be used to learn two distinct functions, then it isn’t “strongly private”.

Defining learning with encoding
- Accuracy on encoded data
- Accuracy on original data

Defining ‘strong privacy’ for encoding
An indistinguishability security game

Our attack: (near) perfect reconstruction