2. Inferring Class Representatives

Prior work focused on properties of an entire class, e.g.:

Model Inversion [Fredrikson et al. CCS’15]

GAN attacks [Hitaji et al. CCS’17]

E.g.: given a gender classifier, infer what a male looks like

But...shouldn't useful machine learning models reveal something about population from which training data was sampled

Privacy leakage !=

Adv learns something about training data

5

2. Inferring Class Representatives

Prior work focused on properties of an entire class, e.g.:

- Model Inversion [Fredrikson et al. CCS'15]
- GAN attacks [Hitaji et al. CCS'17]

E.g.: given a gender classifier, infer what a male looks like

But…shouldn't useful machine learning models reveal something about population from which training data was sampled

Privacy leakage !=
Adv learns something about training data

Intuition