How about if we inferred properties of a subset of the training inputs…

…but not of the whole class?

In a nutshell: given a gender classifier, infer race of people in Bob’s photos

7

Property Inference Attack

How about if we inferred properties of a subset of the training inputs…

…but not of the whole class?

In a nutshell: given a gender classifier, infer race of people in Bob’s photos

Property Inference Attack

3. Reconstruction/Model Extraction Attacks

More recent attacks… sorry, no time!