# Membership Inference Attacks Against Machine Learning Models

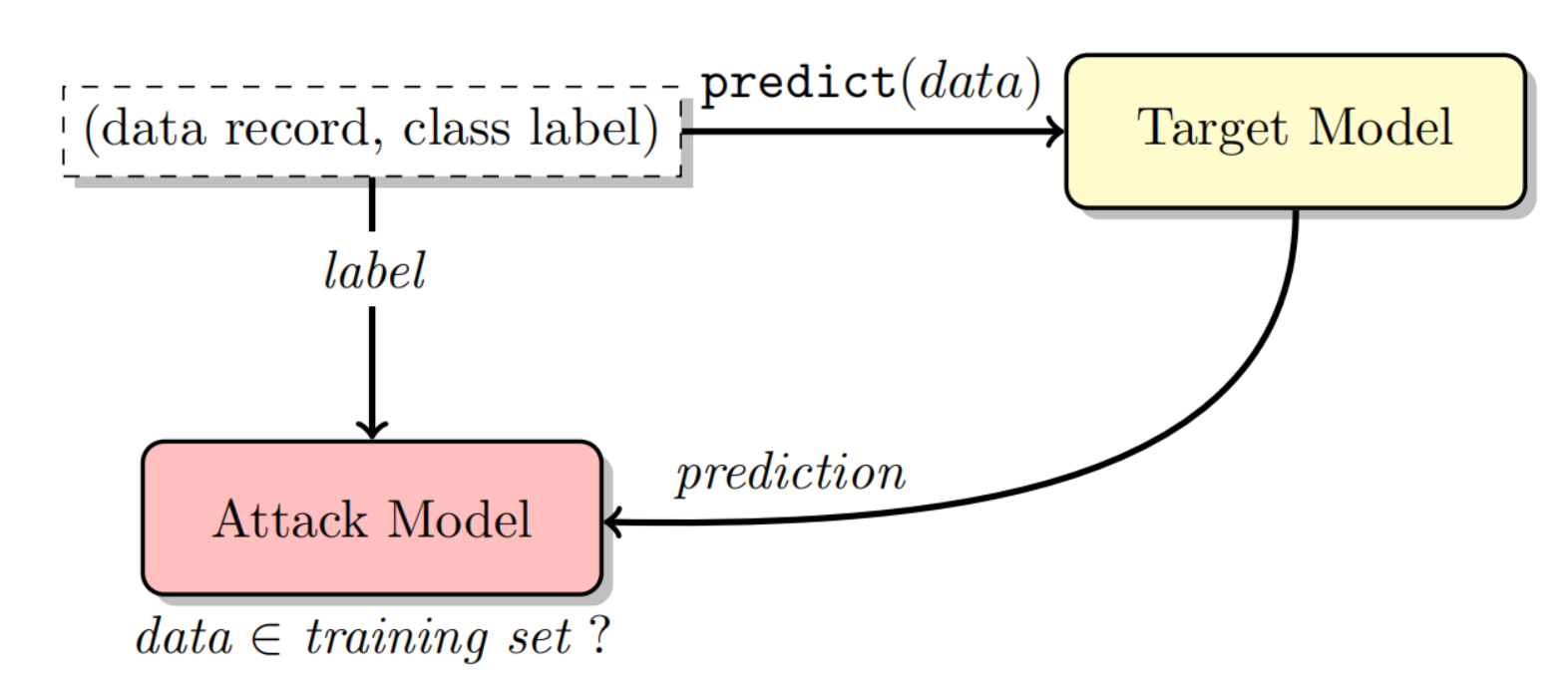
**研究背景：**机器学习算法可以帮助我们更好地理解和分析复杂的数据。机器学习的目标是学习数据(data)和标签(label)之间的关系，并构造一个可以推广到训练集(train set)之外的数据记录(records)的模型。在监督学习模式下，算法学习标记好的数据，生成一个给定输入能够做出好的预测的模型。然而在训练过程中，算法会最小化模型在训练集上的误差，但是模型有可能会过于适合训练集，即**过拟合**，在其他同分布的数据集上表现较差。

**成员关系推理攻击：**给定一个数据记录(data record)的情况下，判断这个数据记录是否用于目标模型(target model)的训练，即该数据记录是否在目标模型的训练集(train set)中。成员关系推理攻击是通过对目标模型在训练集与非训练集数据的输出的量化，预测该数据记录是否在训练集中。

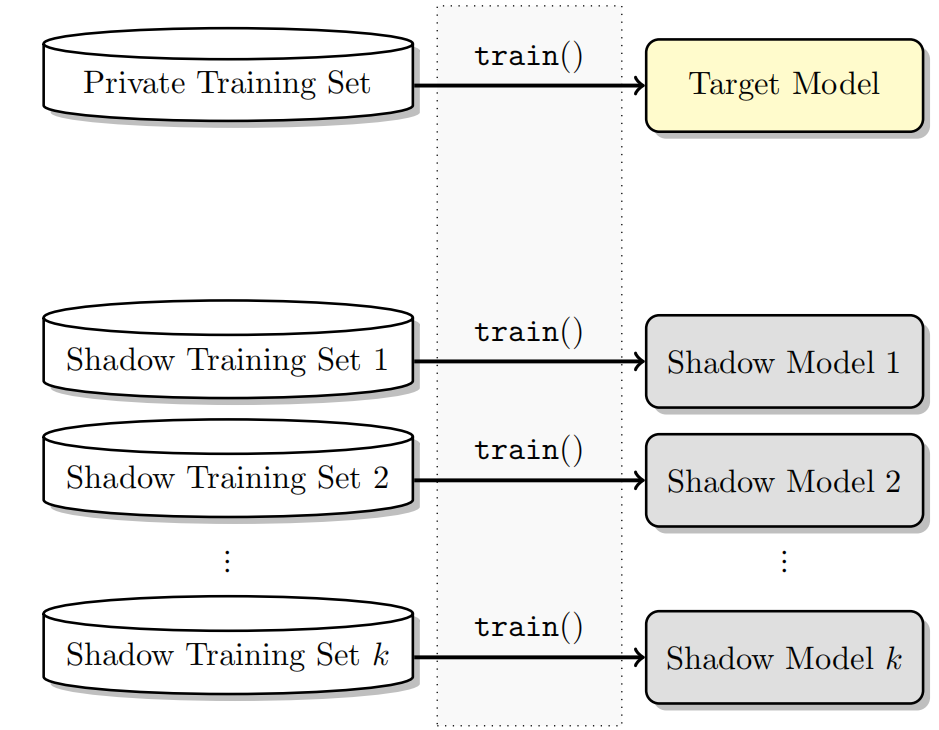
影响成员关系推理攻击的主要因素有：

1. 目标模型的泛化能力，或者是目标模型过拟合的程度
2. 目标模型训练集的选择

模型的攻击组织架构：



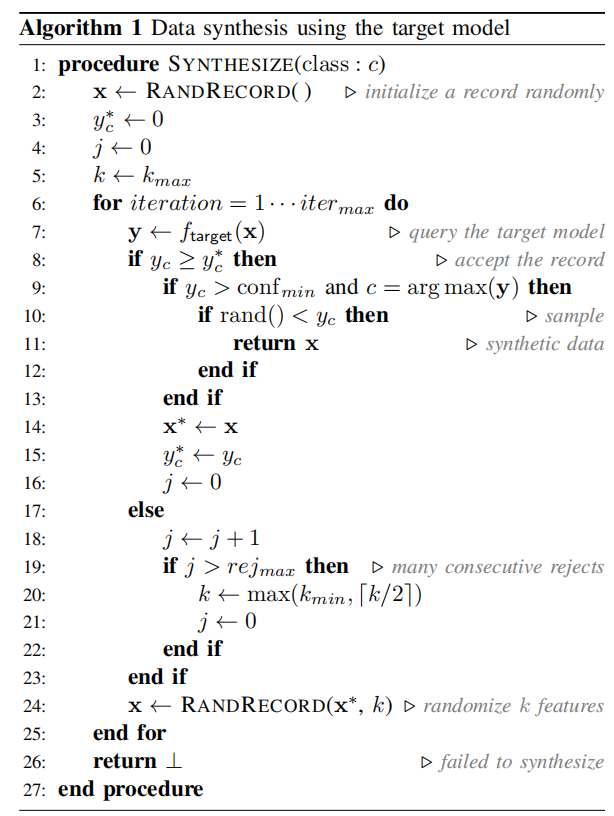
将数据的标签和目标模型的预测值作为一组数据提供给攻击模型进行预测，记为((𝑥,𝑦),𝑓(𝑥))，但仅仅是通过黑盒访问一个目标模型来推测该数据记录是否属于训练集是比较困难的，于是提出**影子模型**降低攻击模型训练数据的获取难度。



**影子模型（shadow models）**是多个与目标模型结构相同的模型，每个影子模型都在同分布的数据集上进行与目标模型相同算法的训练。投入的影子模型越多，则最后训练出的攻击模型会更加准确，因为攻击模型训练判别影子模型上的训练集和非训练集，更多的影子模型意味着更大的训练集。

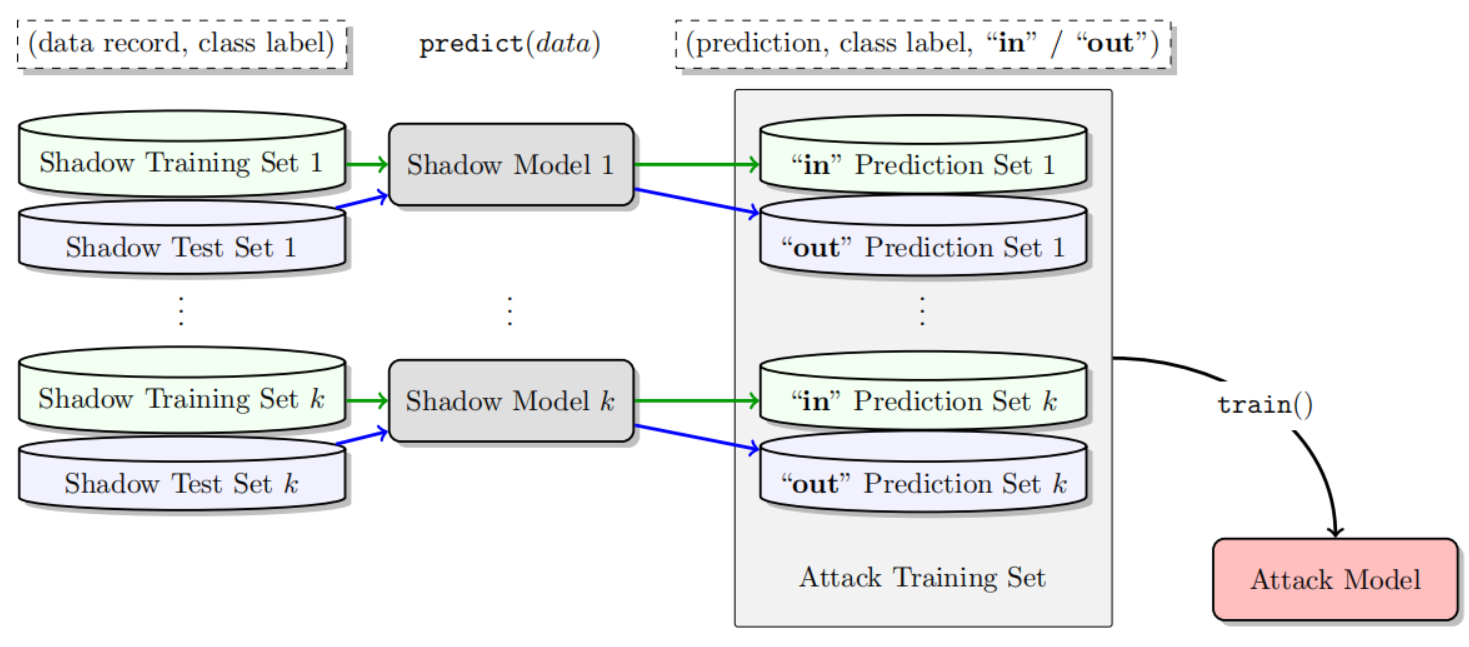
**训练数据的获取：**由于攻击只能访问模型以及其类型，对于数据方面的获取可能存在部分缺陷，因此又提出了训练数据的多种获取方式：

1. 基于模型合成训练数据



1. 使用爬山算法搜索可能的数据空间，而且数据记录能被目标模型高置信度的分类
2. 从这些数据中提取并合成数据样本
3. 基于统计分布合成，从同一分布的数据中提取训练集
4. 对真实数据添加噪声

**训练攻击模型：**

****

对于，计算其预测值，则表示为

对于，计算其预测值，则表示为

由上面两个组成攻击训练集，并且按每个类分割训练集，对于每个类训练一个独立的攻击模型

**总结：**使用该攻击模式即使攻击者对目标模型训练数据分布的假设不是很准确，但是攻击依旧稳定，在只对目标模型有一定了解的情况下，无需对目标模型训练数据分布有任何的先验知识。

基于攻击的经验总结，对于成员关系推理攻击的防御手段如下：

1. 限制预测向量，如只返回最有可能的标签
2. 降低预测的精度
3. 增加预测向量的熵
4. 使用正则化