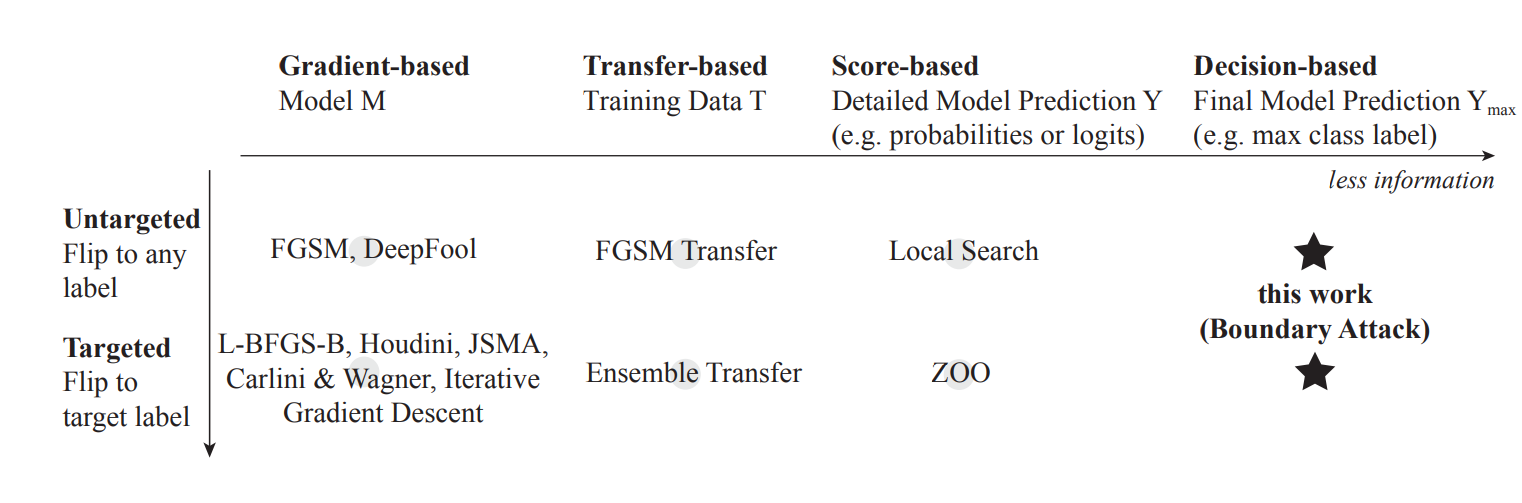
DECISION-BASED ADVERSARIAL ATTACKS: RELIABLE ATTACKS AGAINST BLACK-BOX MACHINE LEARNING MODELS

# 论文的Motivation

作者总结了目前所有的攻击方法，主要可以分为以下几大类：基于梯度的、基于迁移的、基于网络输出的logit或probability的。



这些方法或多或少在现实中有一点的缺陷。基于梯度的需要拿到模型的所有参数，基于Transfer-based需要拿到模型的训练集，基于score需要拿到模型的logit或probability。

然而，更现实的情况是，我们只能拿到模型输出的标签，或者被梯度掩盖后的hard label。