## 当差分隐私遇到可解释性

1、PATEL N, SHOKRI R, ZICK Y. Model Explanations with Differential Privacy[J]//arXiv 20.

2、NAIDU R, PRIYANSHU A, KUMAR A, et al. When Differential Privacy Meets Interpretability: A Case Study[J]//arXiv 21.

## Model Explanations with Differential Privacy

## (新加坡国立大学)

# Importance of the application

* 机器学习模型目前应用于各种高风险领域，例如提供预测性医疗保健分析、评估保险政策和做出信贷决策。这些领域对高维数据的高预测精度，它们采用了越来越复杂的架构，使它们更难解释。使问题更具挑战性的是，这些模型通常用作黑盒算法：它们仅输出决策，而不提供有关其中间计算的详细信息。在高风险领域中对黑盒算法的不信任日益增加，导致对算法透明度的广泛呼吁。该术语泛指提供有关基础算法决策过程的附加信息的方法。这项工作专注于与模型无关的基于特征的解释。与模型无关的方法通常使用由黑盒模型标记的数据集来生成模型解释，称为解释数据集。该数据集通常是训练数据的一个子集，或者是从同一分布中采样的。
* 问题在于，提供额外的信息可能会导致重大的数据隐私奉献，因为会泄露有关基础解释和训练数据的敏感信息，这些信息可能会被推理攻击利用。一些基于特征的模型解释直接依赖于模型参数，这进一步增加了它们在白盒推理攻击方面的隐私漏洞。作为一个相关问题，模型解释也可以放大模型重建攻击的风险。到目前为止，在建模算法透明度和隐私之间的关系以及设计保护数据隐私的模型解释方面的工作很少。也没有研究随机隐私机制对模型解释保真度的破坏性影响。

# Problems cannot be solved with existing methods

* 模型解释容易受到成员推理攻击，这些攻击可以推断数据点是否是模型训练集的一部分。现有工作不提供任何因隐私而导致的效用损失的界限；重复查询导致不可接受的累积信息泄漏和松散的隐私保证，也不为训练数据集提供任何隐私分析。

# Formal definition of the problem

考虑一个在训练数据集上训练的黑盒决策机器学习模型。该模型返回预测的标签，并且不提供有关其决策的更多信息。目标是生成模型解释，其输入是兴趣点，解释数据集（用于生成解释），以及在上的输出。专注于基于特征的模型解释：是中的一个向量，其中衡量第个特征对预测标签的影响。

# What are the technical challenges with this approach?

* 如何有效地利用先前发布的差分隐私信息？

# Proposed Methods for these challenges

* 从一种简单的方法开始探索差分隐私模型解释：以差分隐私方式近似每个局部模型解释。图1中算法1提出一种梯度下降算法，通过在凸集的每次迭代中使用高斯机制来保证解释数据集的差分隐私。这是一种交互式差分隐私机制：对于每个查询，参考解释数据集来计算解释。

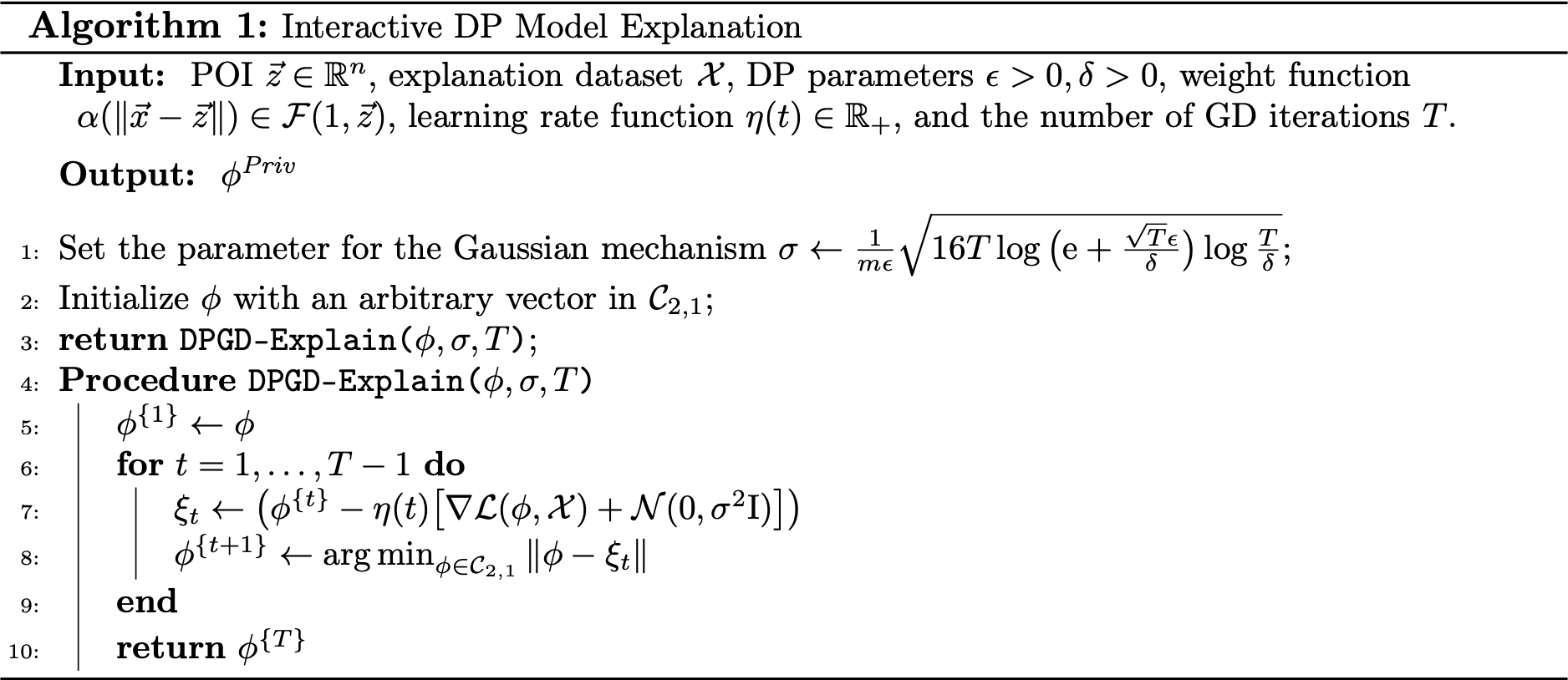
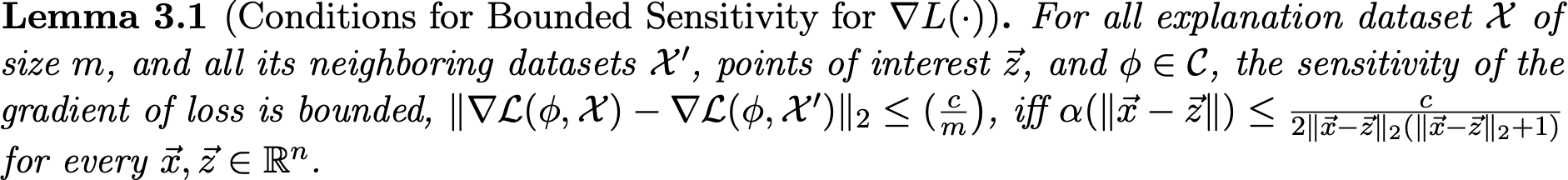


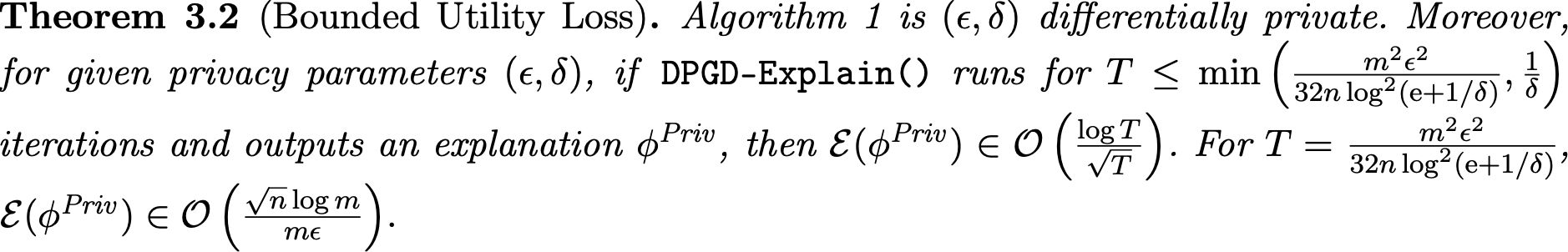
图1 交互式差分隐私模型解释

* + 损失函数的敏感度：



将理想的权重函数族定义为：

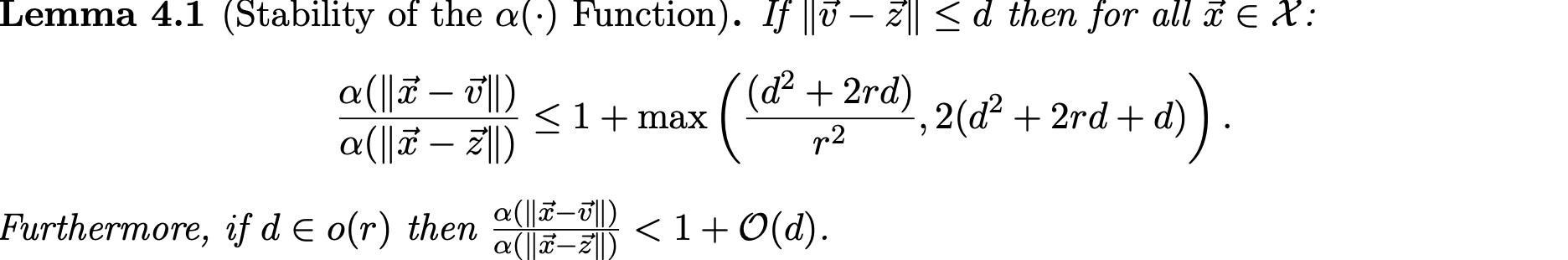
* + 隐私和效用保证：

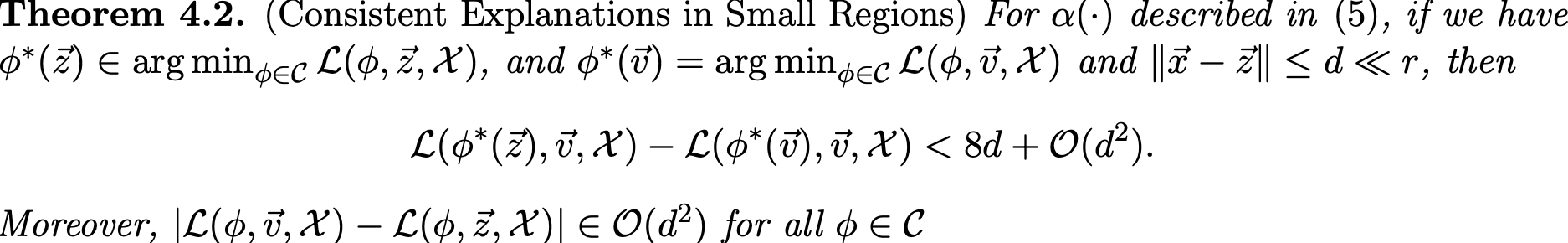


* 自适应差分隐私机制

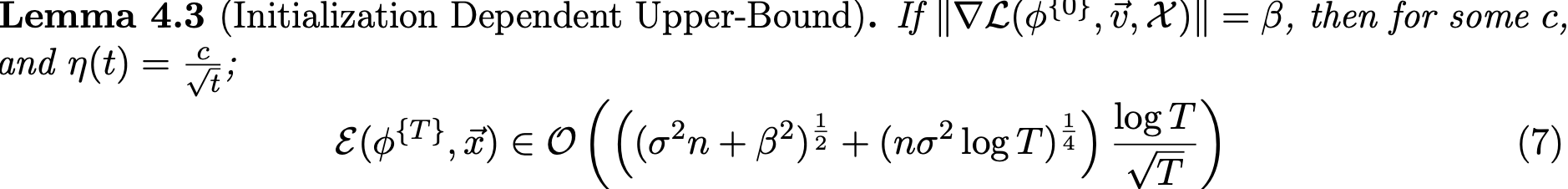
令是先前以不同的隐私方式解释的查询序列，以及它们对应的解释：。当收到一个新的查询时，目标是从中提取信息，这样新的查询需要更少的隐私预算，而不会影响其解释。

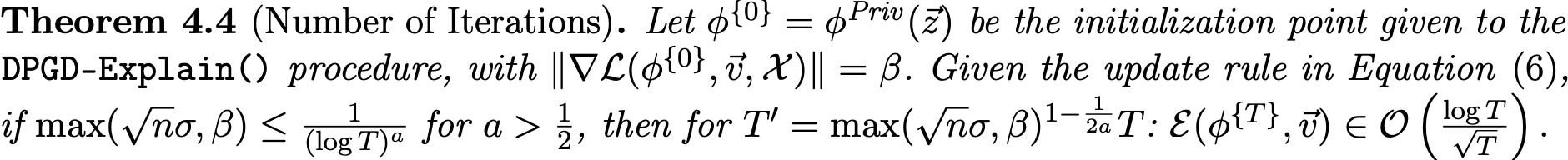
* + 以两种方式使用过去的信息：
    - 相似的数据点应该有相似的解释：如果和在数据空间中彼此非常接近，那么模型的解释和相似。



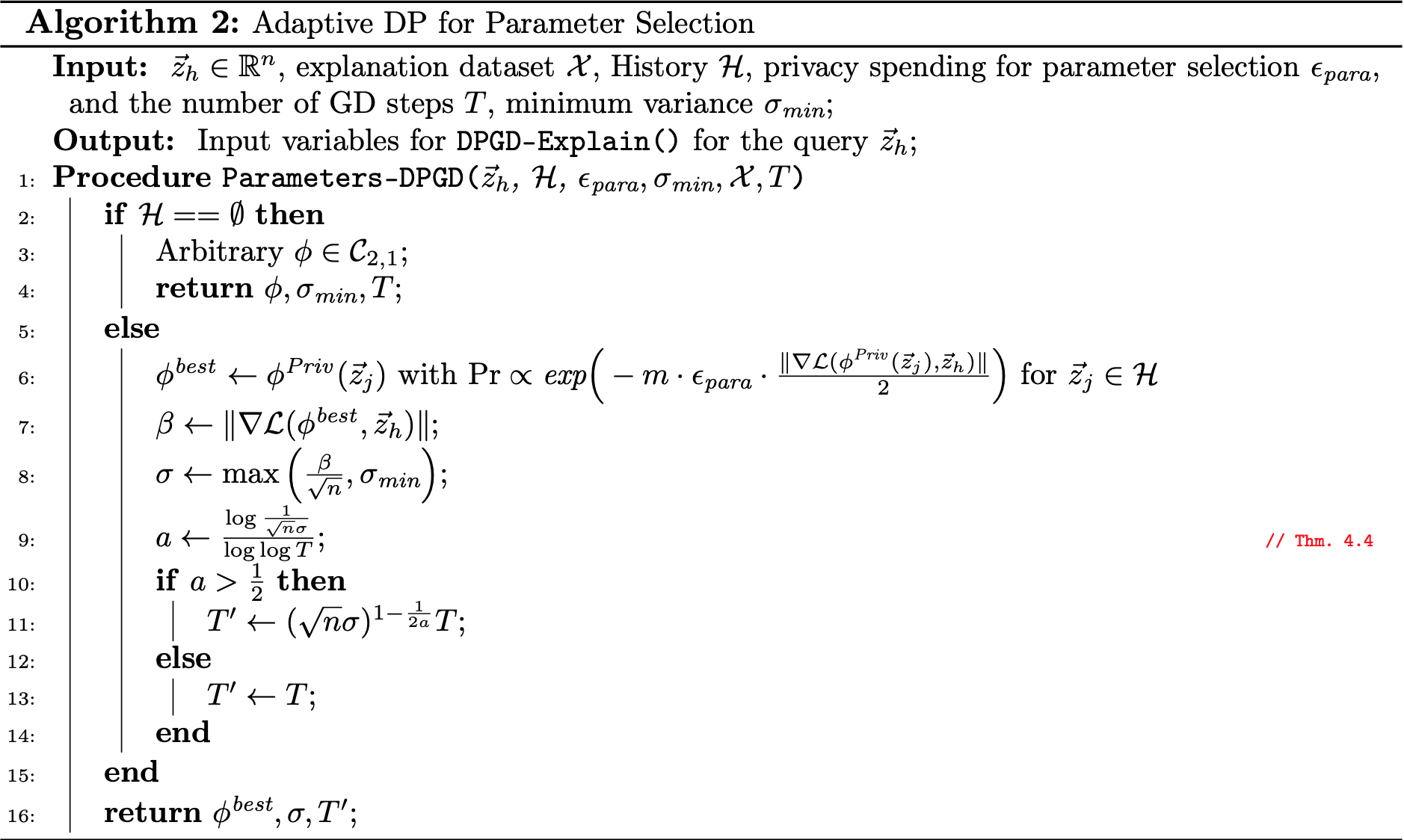


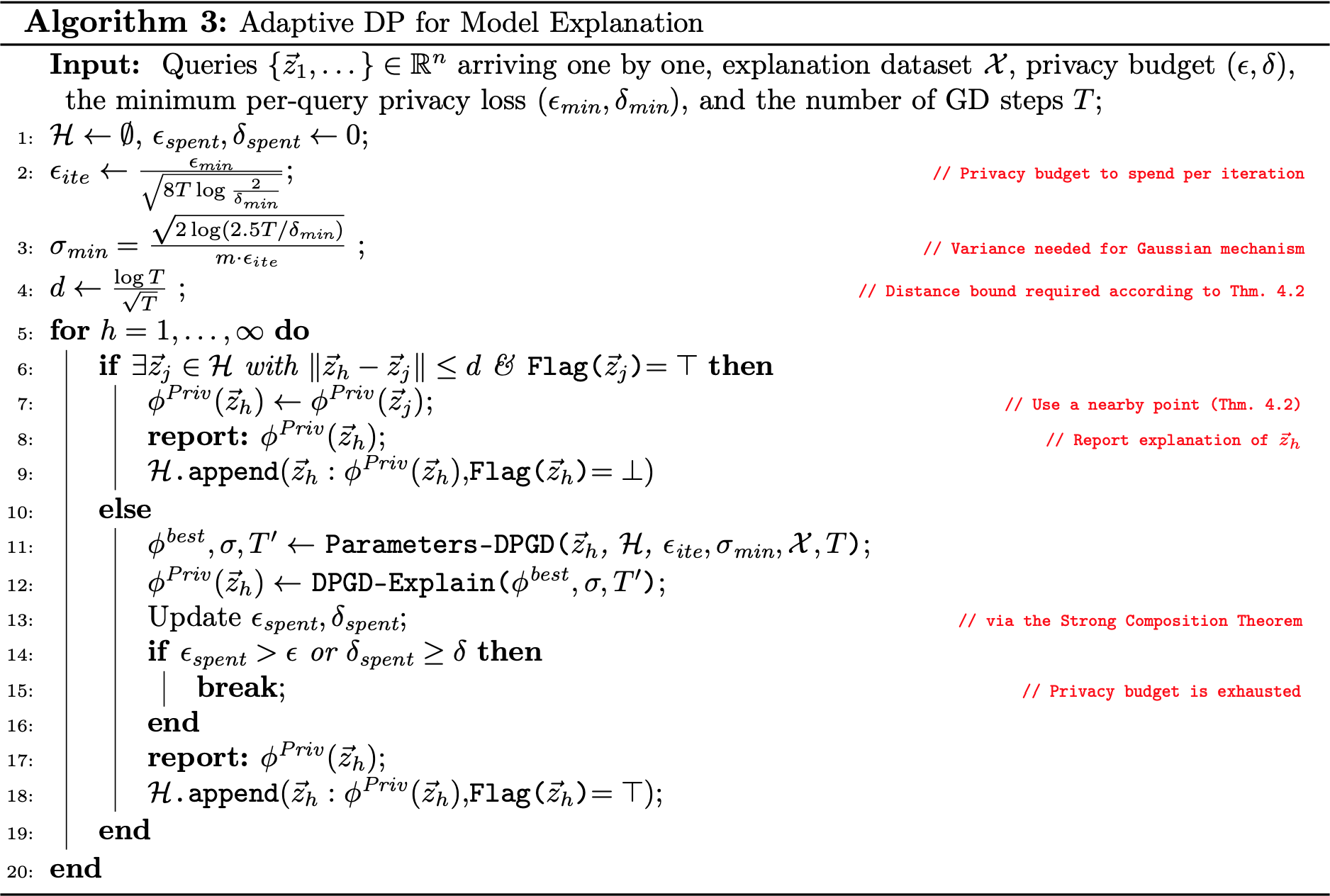
* + - 利用过去的信息来找到近似最优的开始，从而加快收敛速度并减少隐私支出。





* + 一旦算法发布了关于底层解释数据集的足够信息，它就会利用这些信息使用更少的隐私支出来解释新的查询。



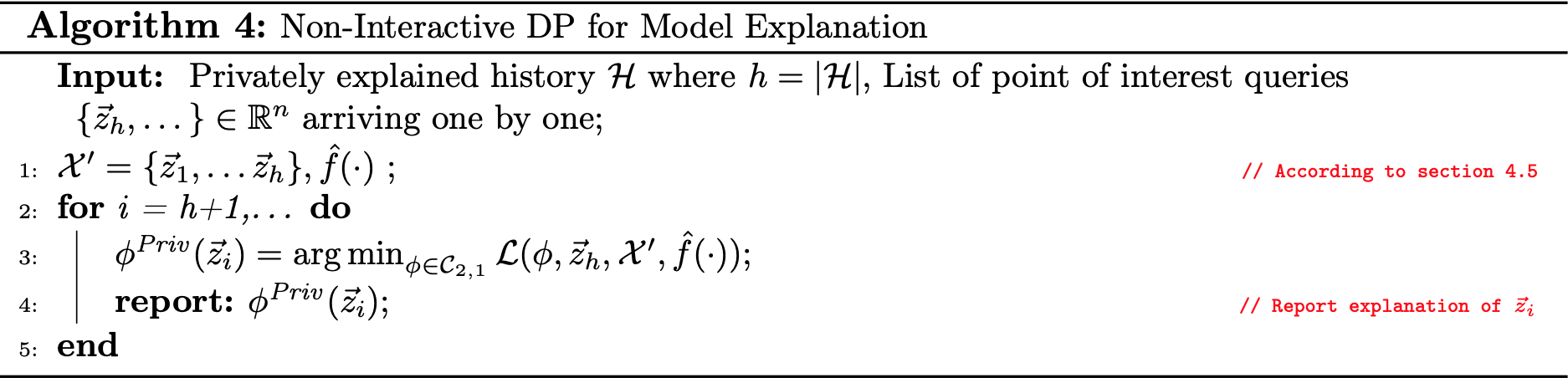


* + 提前终止和增强的自适应算法：

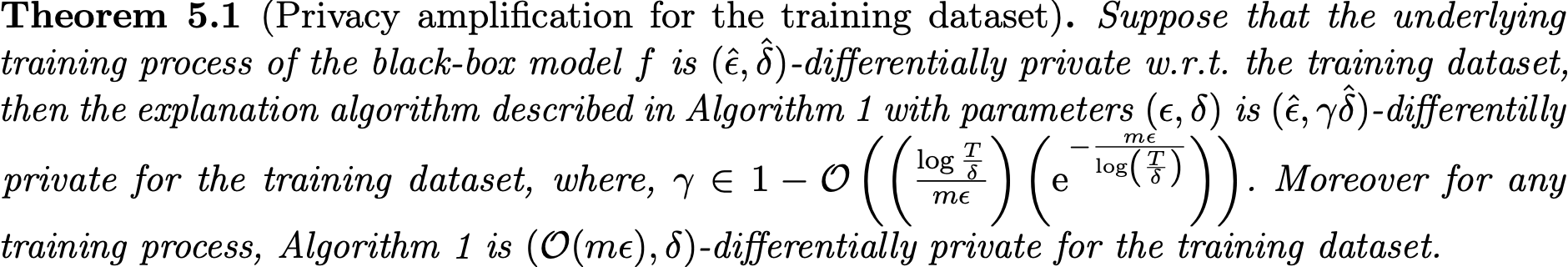
每当自适应算法发现的初始化时，增加迭代次数并不会导致更快的收敛，因为高斯噪声占主导地位：这些情况下的隐私支出提供损失改善不大。

* + 用于模型解释的非交互式差分隐私机制

如果包含有关底层黑盒模型的足够信息，则无需对解释数据集进行额外查询。



* 保护训练数据的隐私：

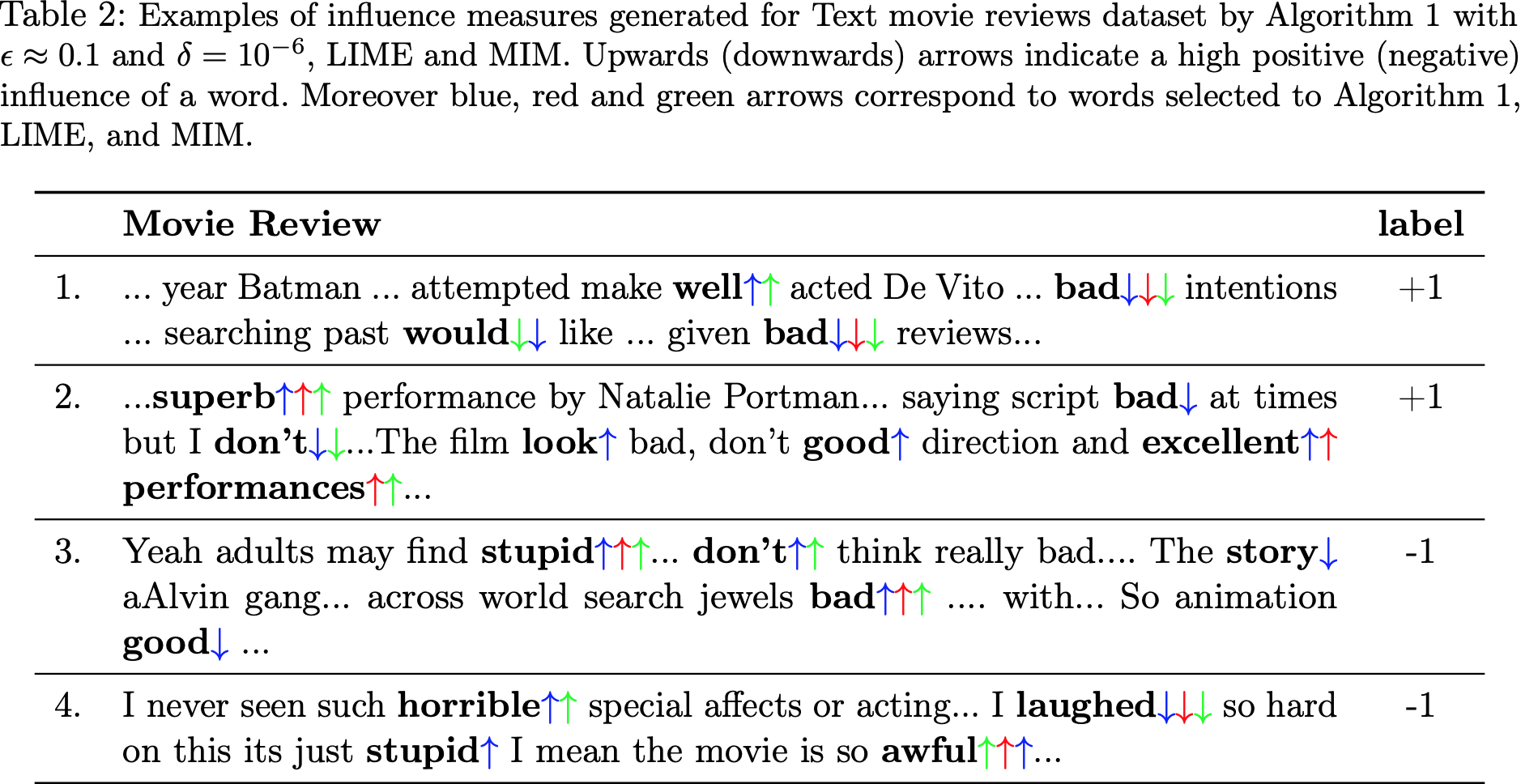


**Experiments and Datasets**

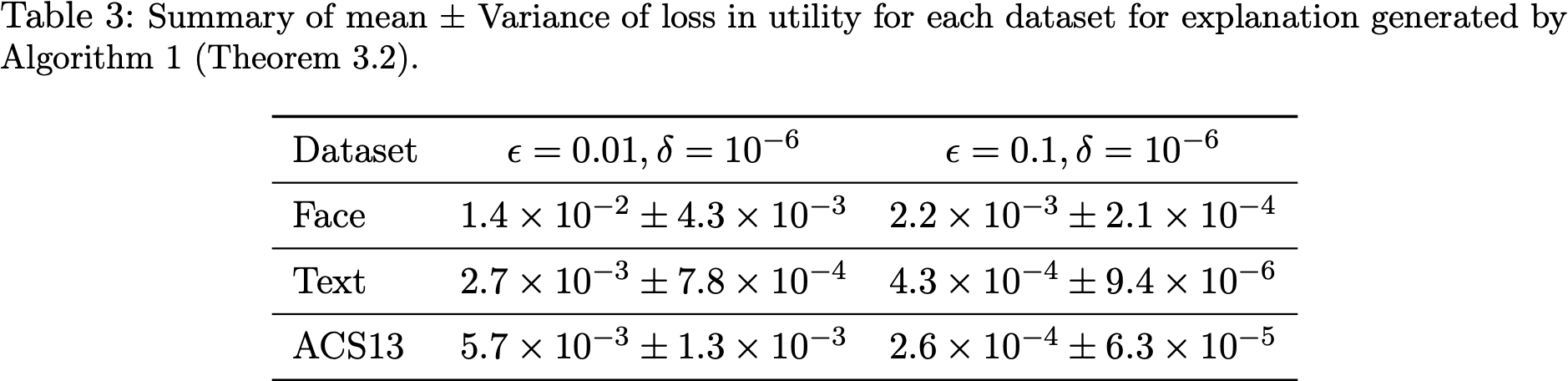
* 数据集

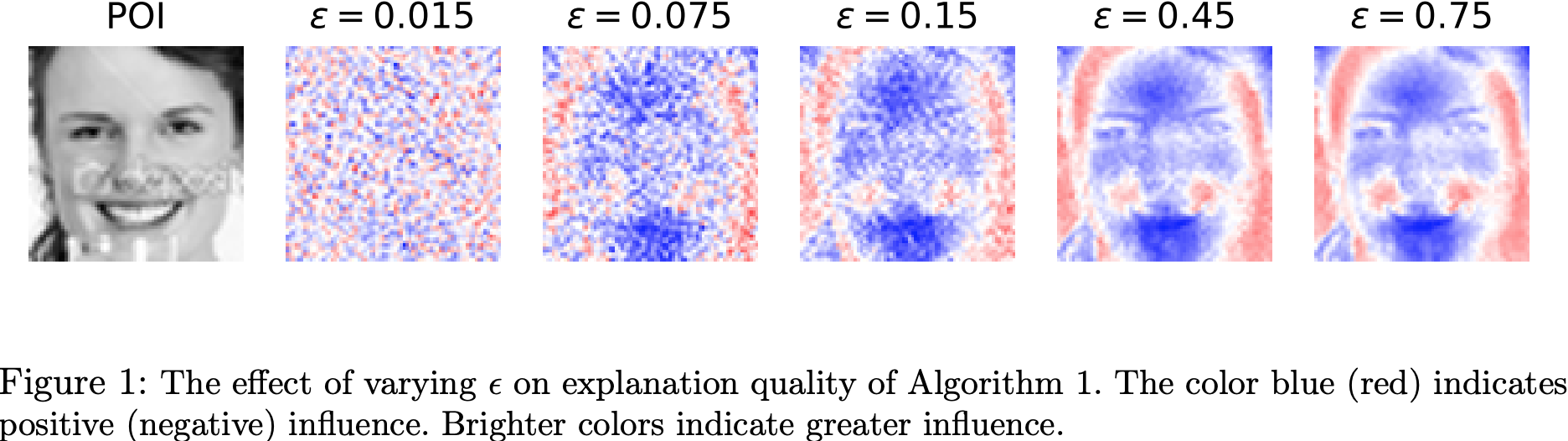
在标准机器学习数据集上评估模型解释。使用以下基准数据集（将整个数据集用作解释数据集）：

* + ACS13：该数据集包含1,494,974条记录并预测收⼊（> 50k$和 ≤ 50k$）。
  + IMDB/亚马逊电影评论（文本数据集）：该数据集包含来自亚马逊评论数据集的8,765,568条电影评论以及来自IMDB大型评论数据集的50,000条电影评论，使用500个高频词映射到二进制向量。每部电影评论都被标记为正面（+1）或负面（-1）评论。
  + 面部表情数据集：该数据集由12,156个48×48像素的面部灰度图像组成。
* 实验

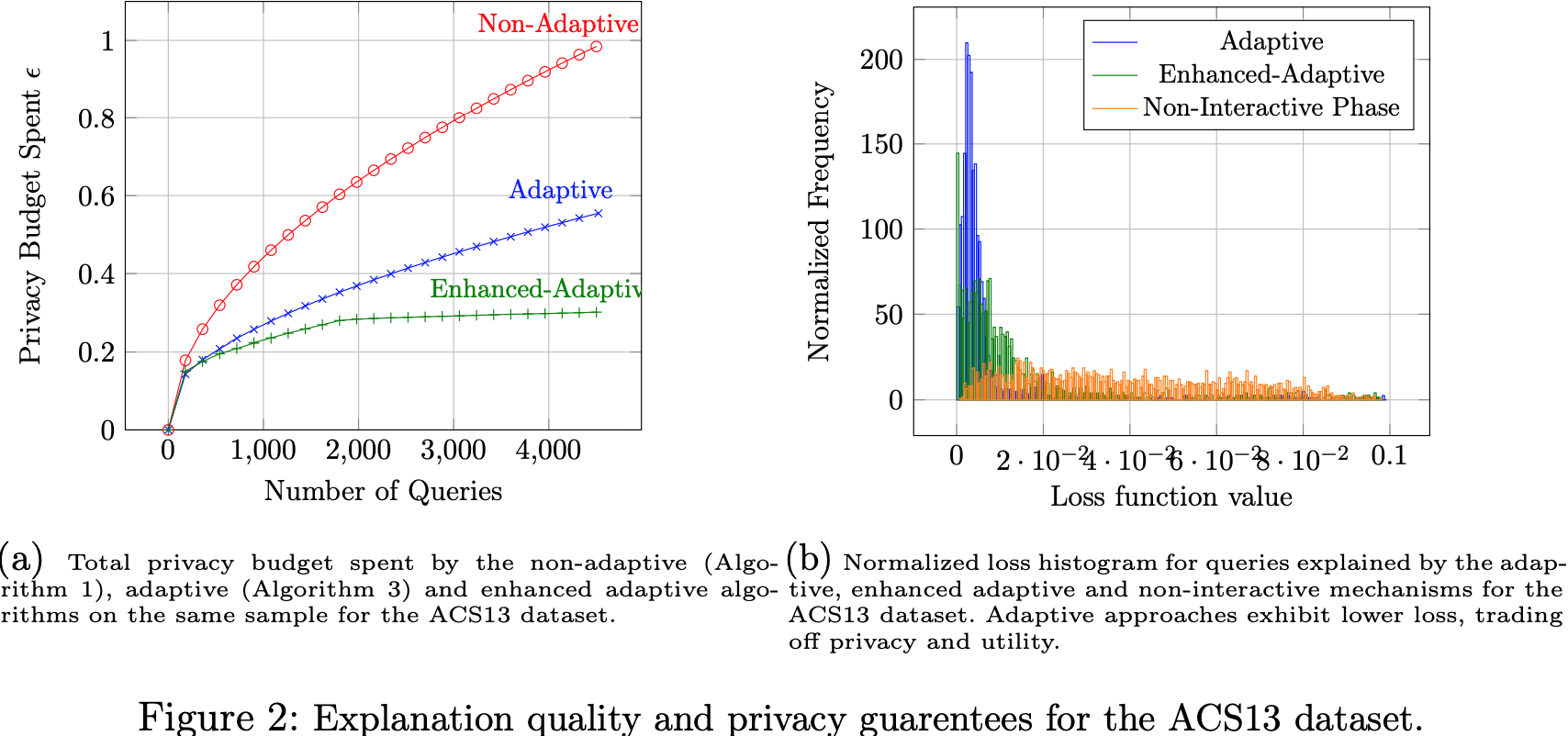


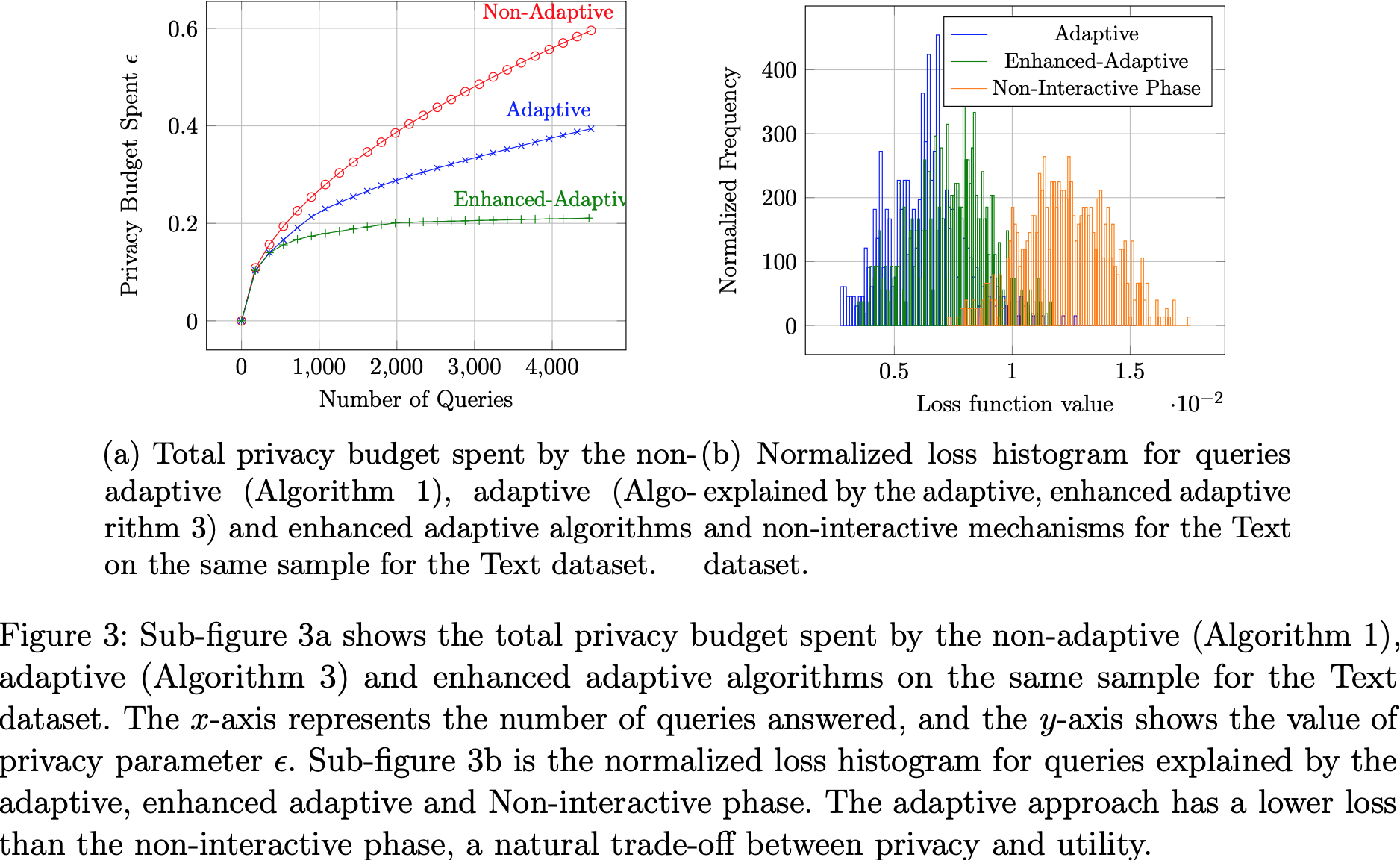
* + 随着对隐私要求严格，解释质量如何下降？

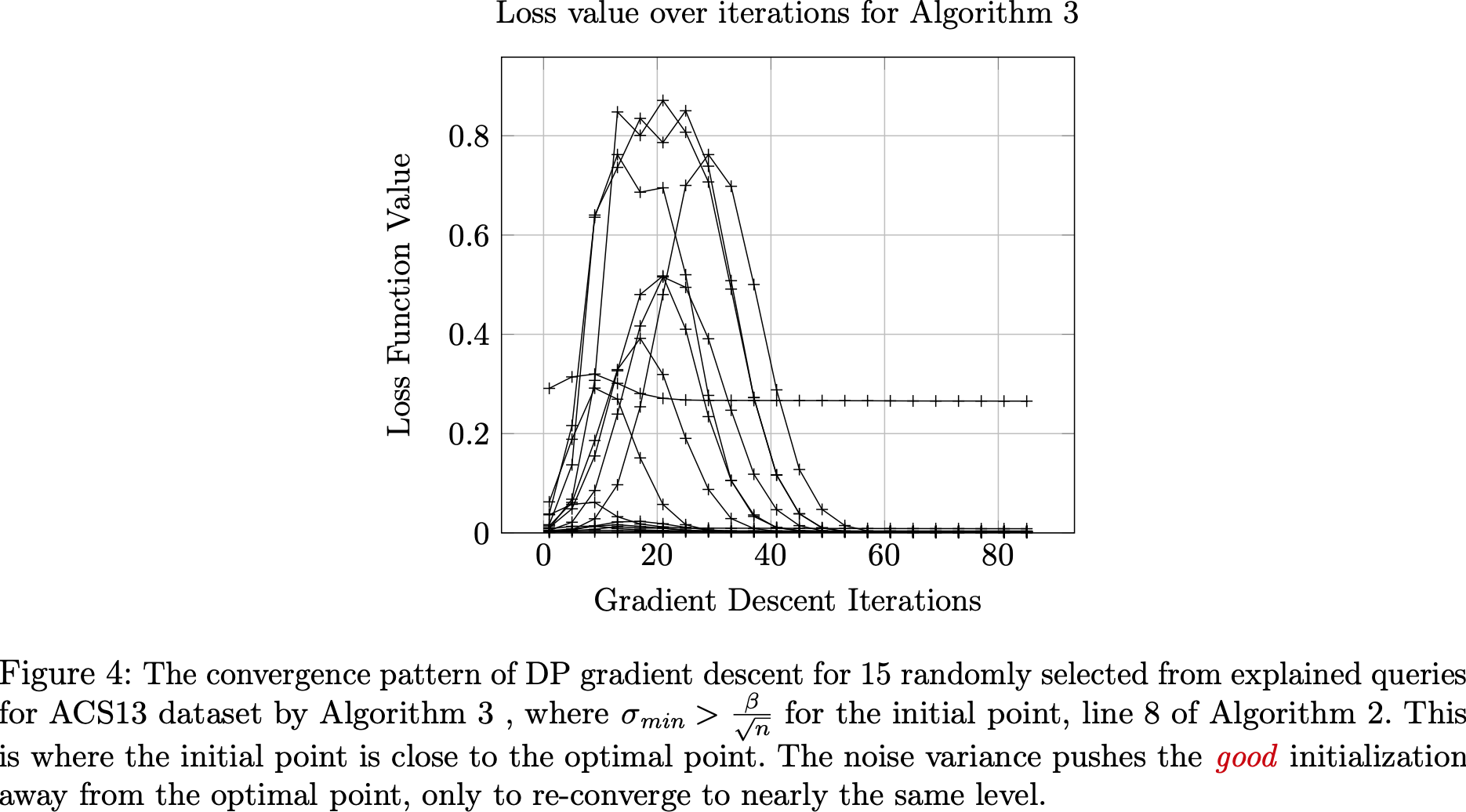


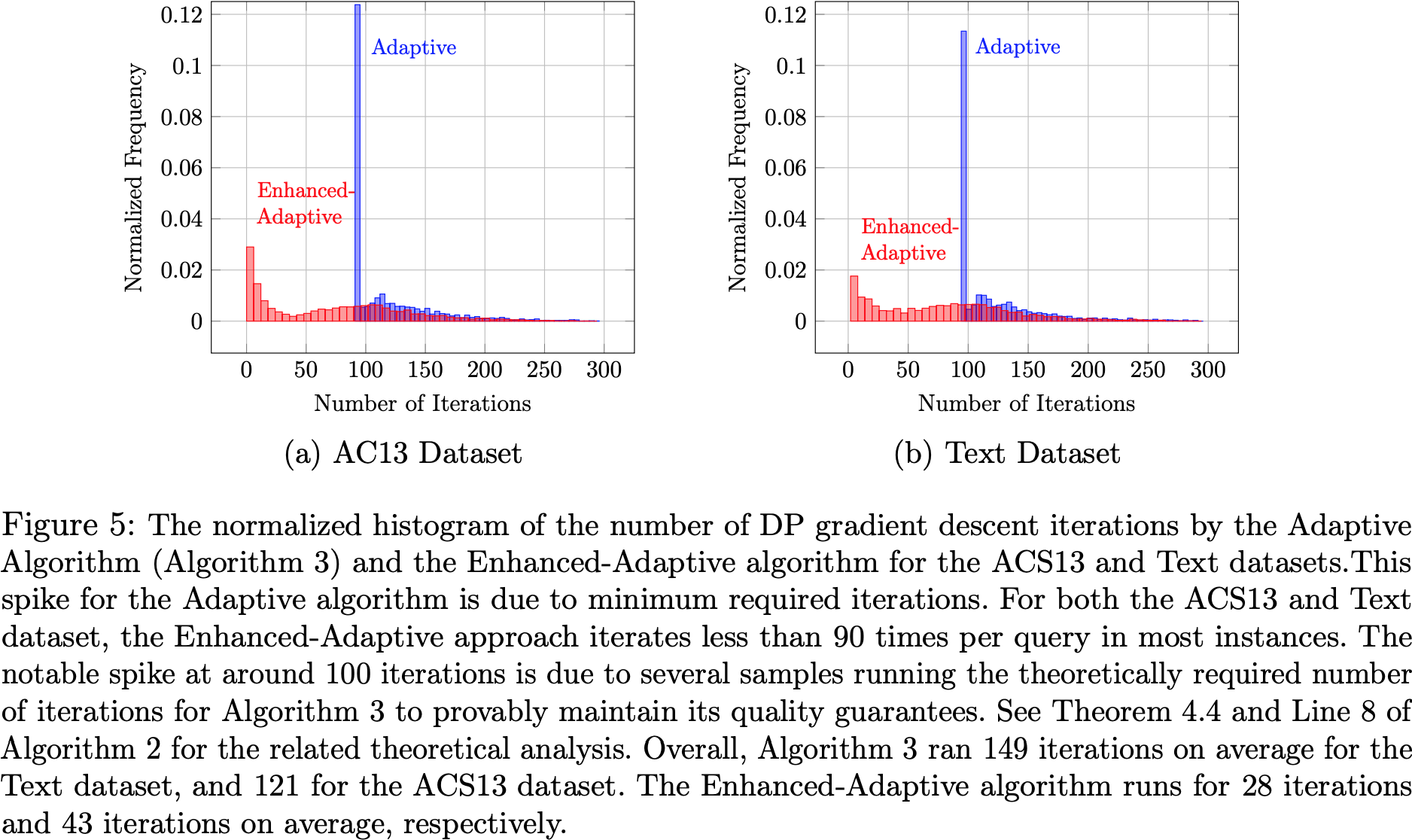


* + 通过自适应方法节省隐私预算

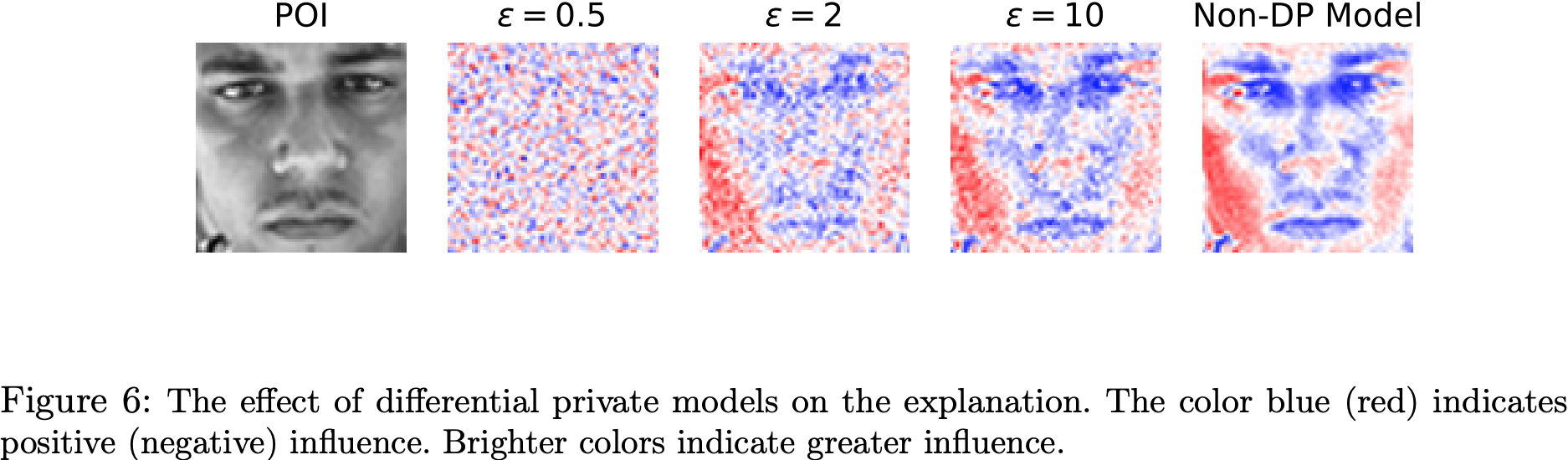


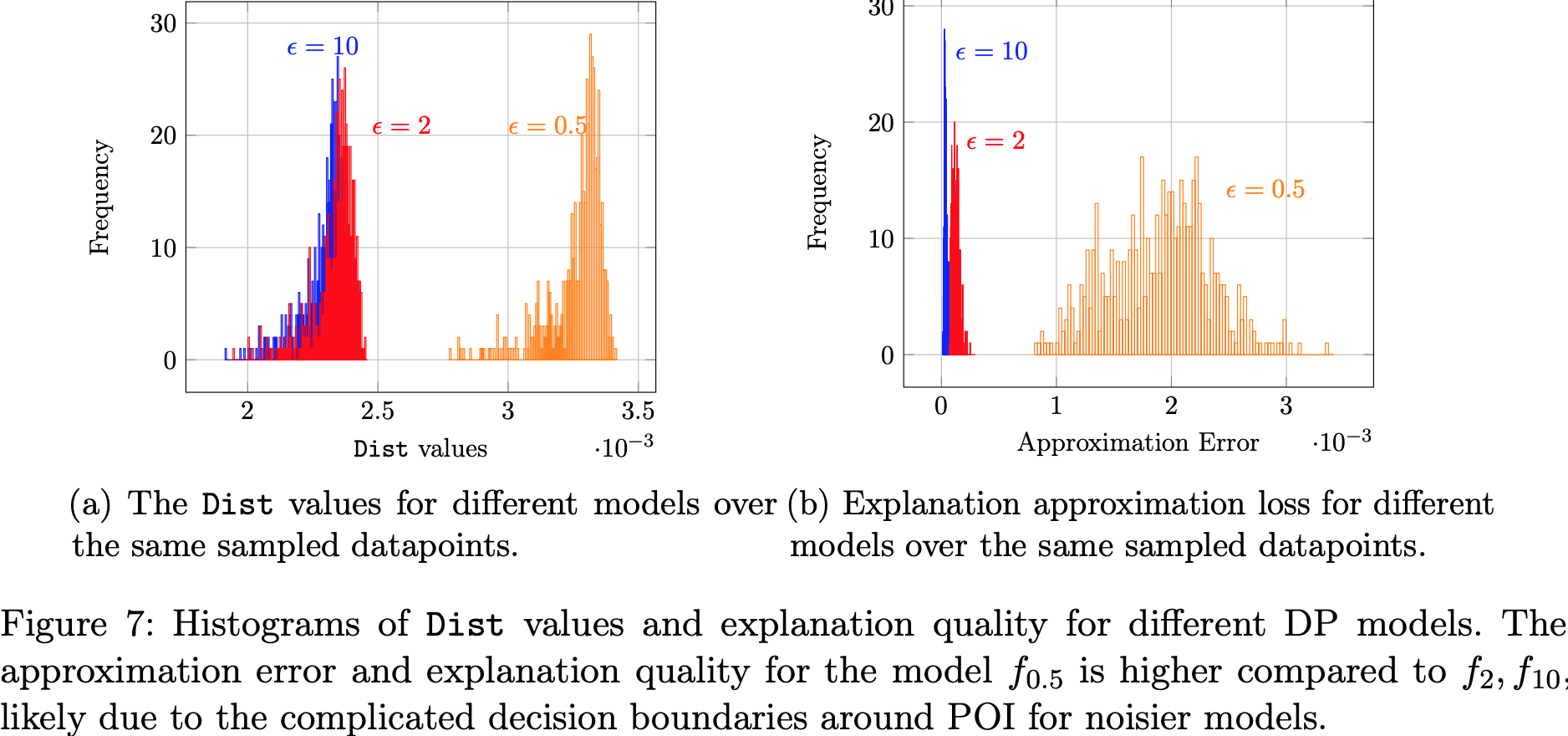


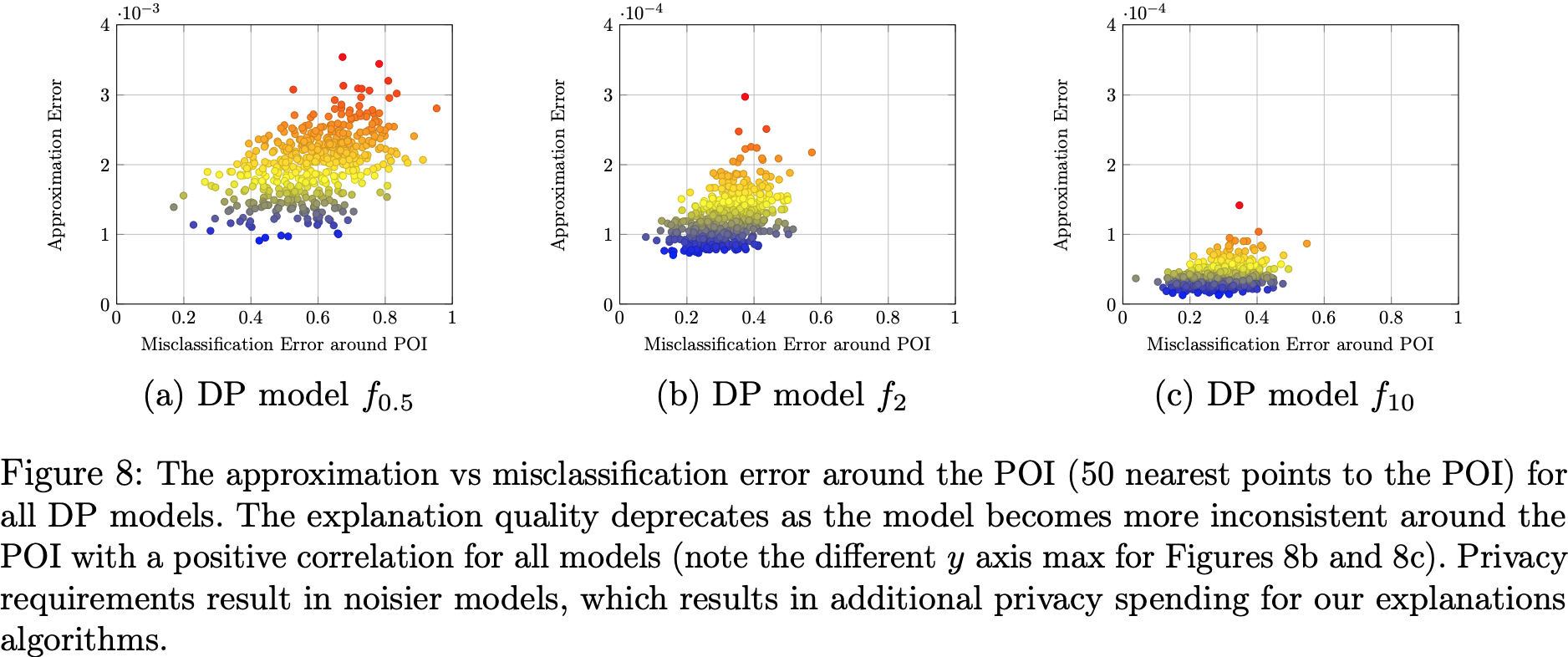




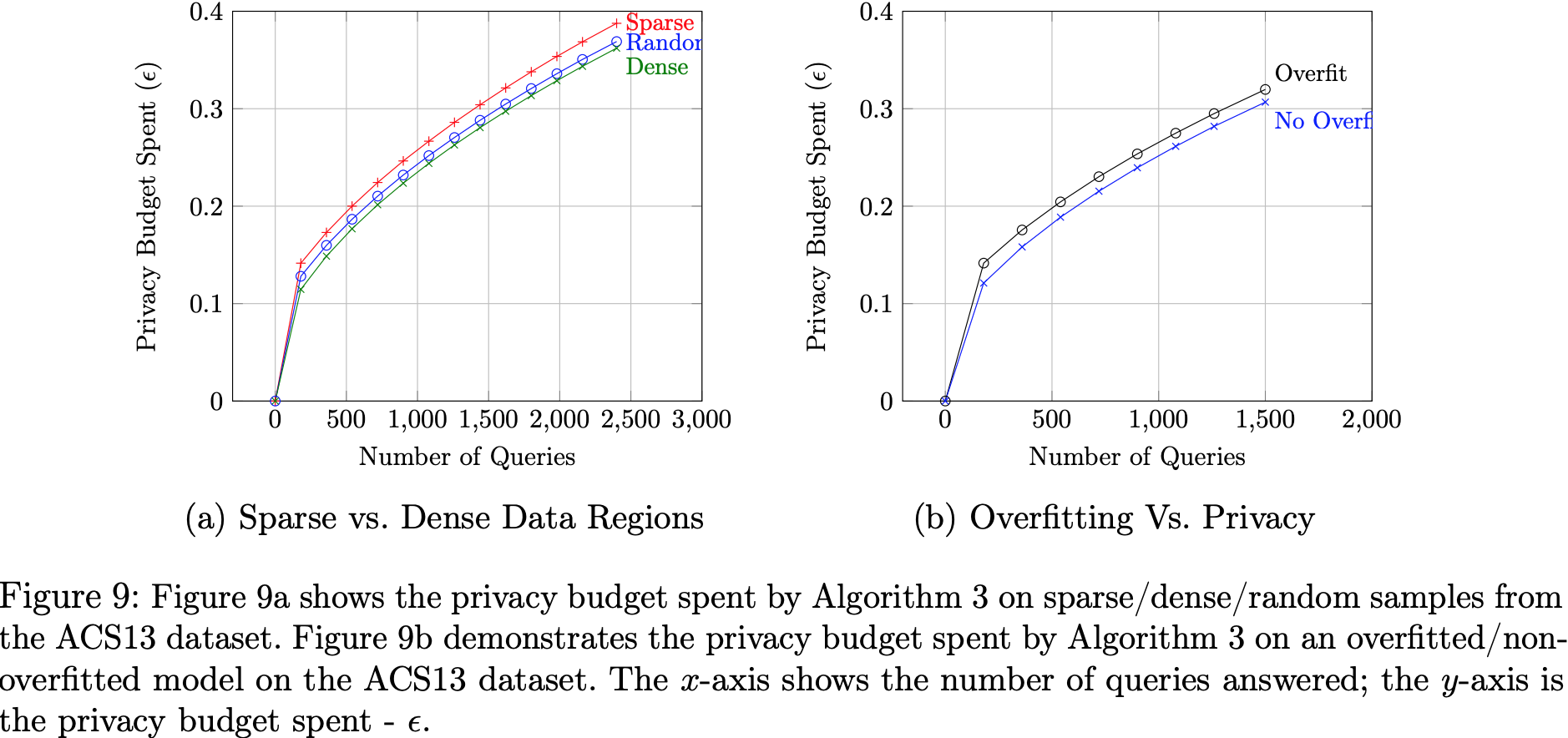
* + 差分隐私模型的差分隐私解释







* + 数据密度和过拟合的影响



# Differences of this Method compared to others

本文提出了用于计算模型解释的可证明合理的、与模型无关的和差分隐私的算法，以保护解释和训练数据集中的敏感信息。解释是合理的，因为它们与一些标准模型解释相似。方法可以适用于任何依赖于围绕试图解释的数据点生成准确的局部模型的解释方法。

首先设计了一个基线交互机制，输出基于特征的模型解释，同时保证每个查询的不同隐私。在此基线的基础上，设计了主要算法：用于模型解释的自适应差分隐私机制。在这种方法中解决的主要挑战是优化模型解释对所有查询的组合隐私损失，同时保持低解释错误。自适应算法有效地利用了先前发布的差分隐私的解释，显着减少了新模型解释查询的隐私支出。通过使用过去的查询为底层梯度下降算法选择一个更好的初始化点来实现这一点。改进了关于差分隐私梯度下降方法（在一个小假设下）的收敛性的界限，提供了更快的初始化相关收敛速度。进一步表明，当接近最优点时，差分隐私梯度下降算法会在最优点附近振荡，花费大量隐私预算，而仅获得可忽略不计的预期效用。这种观察导致了一种增强的自适应算法，它提供了更好的隐私保护，而准确性损失很小。最后，提出了一种非交互式差分隐私算法，该算法无需花费任何进一步的隐私预算即可生成解释。

算法限制了解释数据的差分隐私损失。训练数据的隐私保护通常在模型训练期间完成，或者通过差分隐私预测算法完成。然而，本文表明，机制使用底层模型作为黑盒，也可以适度放大训练数据的隐私保护。如前所述，保护敏感数据的隐私必须与保护解释质量同时进行：这至关重要，因为差分隐私机制中的随机性可能会降低解释的保真度。确保我们的机制保持解释质量需要仔细分析底层梯度下降机制：我们确定最小化解释查询所需的隐私预算支出所需的迭代次数，从而最大化整体解释质量。在使用差分隐私模型保护数据时，差分隐私模型决策边界的随机性降低了逼近算法的收敛速度，从而导致花费更多的隐私预算。这在我们的实证结果中得到了证明：对底层模型的更强隐私要求会降低解释质量。

**Limitation of this work or your discovery**

差异性影响：模型对数据过度拟合的区域往往更容易受到隐私/质量权衡的影响；换句话说，重要的是要确保少数群体不会以不可接受的方式受到不利影响。向少数群体提供低质量的模型解释是不可取的，但也不可冒险泄露他们的隐私信息。

## When Differential Privacy Meets Interpretability: A Case Study

## (曼尼帕尔理工学院、卡内基梅隆大学、OpenMined、宾夕法尼亚大学、焦特布尔印度理工学院、加州大学)

# Importance of the application

* 机器学习的应用与发展在医药卫生领域呈指数增长。随着最近在将人工智能应用于健康方面取得的进展，人们可以看到它拥有的巨大潜力。但是，医疗保健数据包含必须根据保护主体隐私的安全协议处理的敏感信息。同时，模型结果和预测必须是可解释的，允许相关的医学专家研究和验证评估。这清楚地确定了计算机视觉中的一个问题，对于某些用例，必须解决责任和隐私问题。
* 差分隐私定义为一个广泛的工具，对于给定数据分布的算法构成强大的隐私保证，通过提供数据集中的整体模式，同时隐瞒有关个人的信息。可解释的机器学习被定义为算法实现的集合，其中让用户了解模型如何获取具体决定。这两种技术，如果正确地利用可以进一步完善当前的实现和实用性，增加信任维度、纠正反馈和对黑盒模型的主张的可竞争性。

# What are the technical challenges with this approach?

* 应用差分隐私如何影响解释质量？

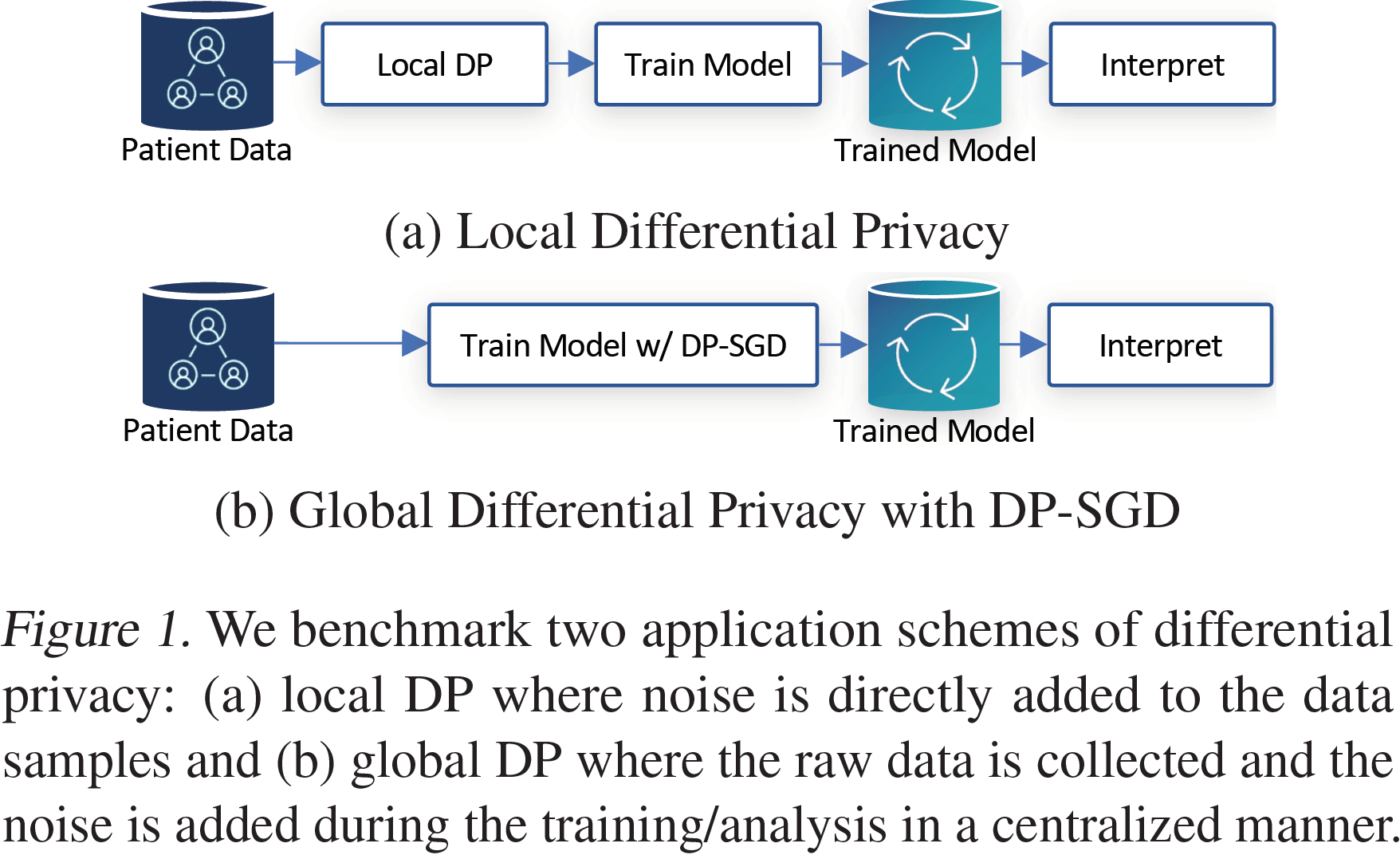
# Possible methods for this challenges

* 在健康案例中，更广泛地探索差分隐私模型可解释性。
* 提供了第一个探索可解释性的基准，特别是通过类激活映射，实用差分隐私训练深度神经网络。在大范围的隐私预算设置下，训练本地化和中心化的差分隐私的深度神经网络，研究它们对模型可解释性的影响。利用Grad-CAM作为可解释性方法，使用猫狗和APTOS数据集在通用和医疗设置中训练模型。

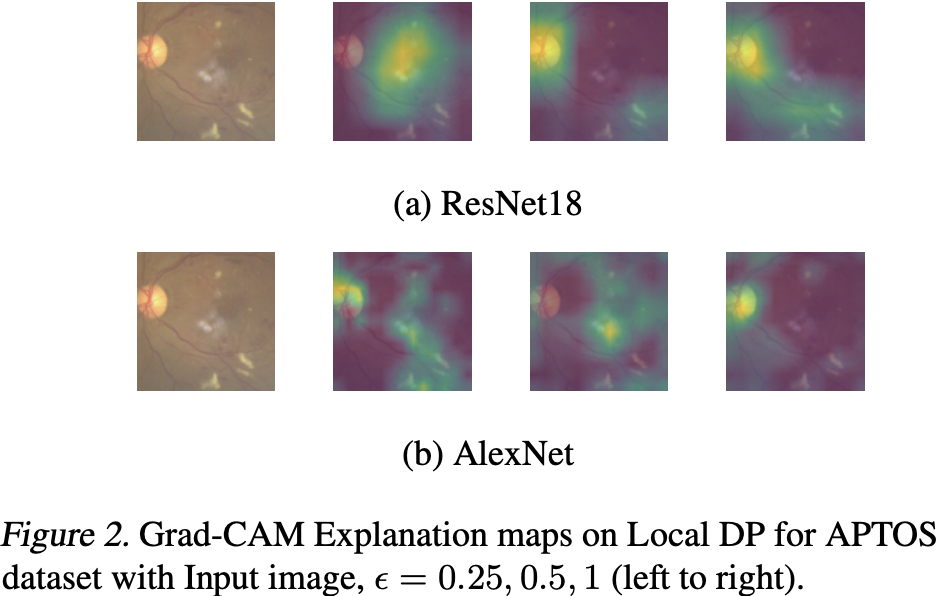
# Experiments and Datasets

* 数据集

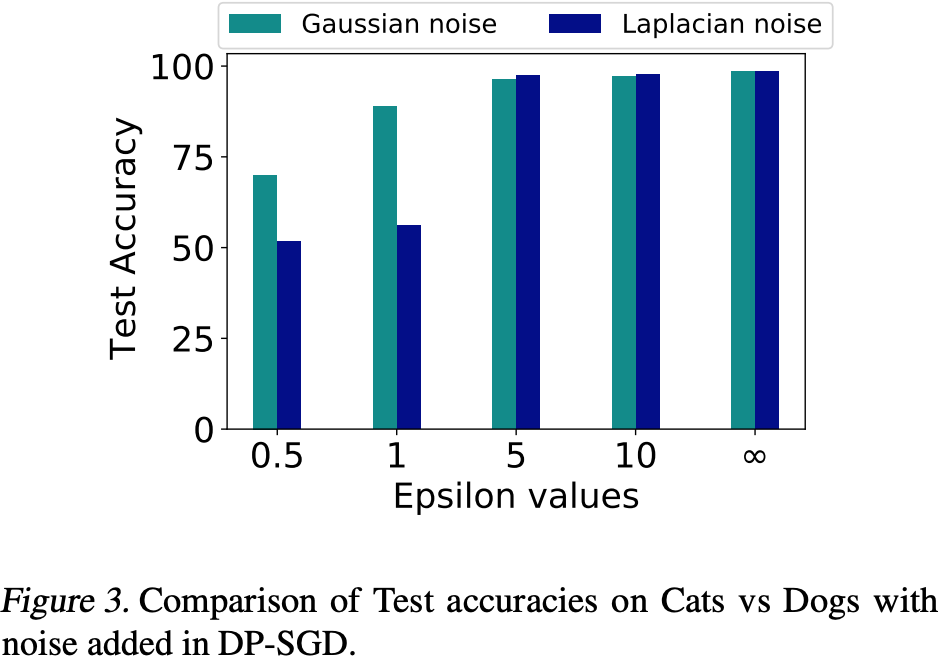
评估两个数据集：APTOS和猫狗数据集。



* 评价



随着增加（隐私程度降低），以更好的质量接近解释。



和之间的差距非常大（25%），通过噪声值采样的方差和散度来解释。

使用两个指标来评估差分隐私模型的可解释性：

* + ：是指使用输入图像的预测和使用显著图的预测所做出的预测中最大的正差。由给出。其中，是指输入图像在类上的预测值；是指对输入图像使用显著图在类上的预测值。
  + 置信度上升百分比：由给出。其中，是布尔函数，如果括号内的条件为真，返回1；否则返回0。

