MODEL EXPLANATIONS WITH DIFFERENTIAL PRIVACY

Neel Patel, Reza Shokri, Yair Zick

ACM FAccT ‘22

Neel Patel, Reza Shokri, and Yair Zick. 2022. Model Explanations with Differential Privacy. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1895–1904.

https://doi.org/10.1145/3531146.3533235

Conduct

- From Ethos of MLC…
  - *https://mlcollective.org/wiki/code-of-conduct/*
- Highlights
  - *Expectation of Confidentiality*
  - *Reporting -> send me a direct message*

Summary

- Paper 30%
- Discussion 70%

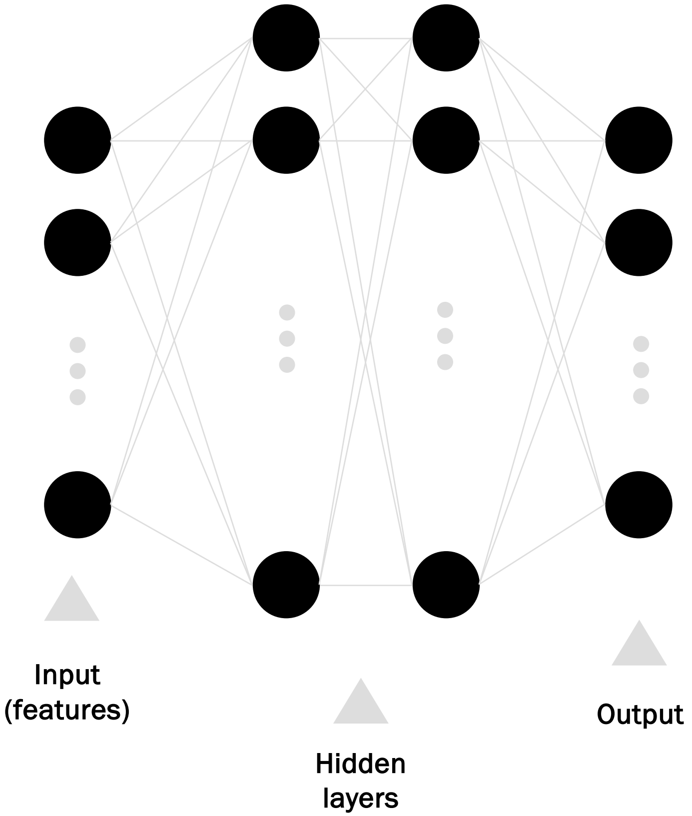
Problem

- Trustworthy Computing
  - *Fairness*
  - ***Explainability***
  - ***Privacy & Security***
  - *Robustness*
- Conflicts exist within trustworthy computing
  - *This paper focuses on conflicts between **Privacy** and **Explainability***
- **How does philosophy / ethics play a part in Trustworthy AI**

This is something I’m unsure / still thinking about

Neural Networks

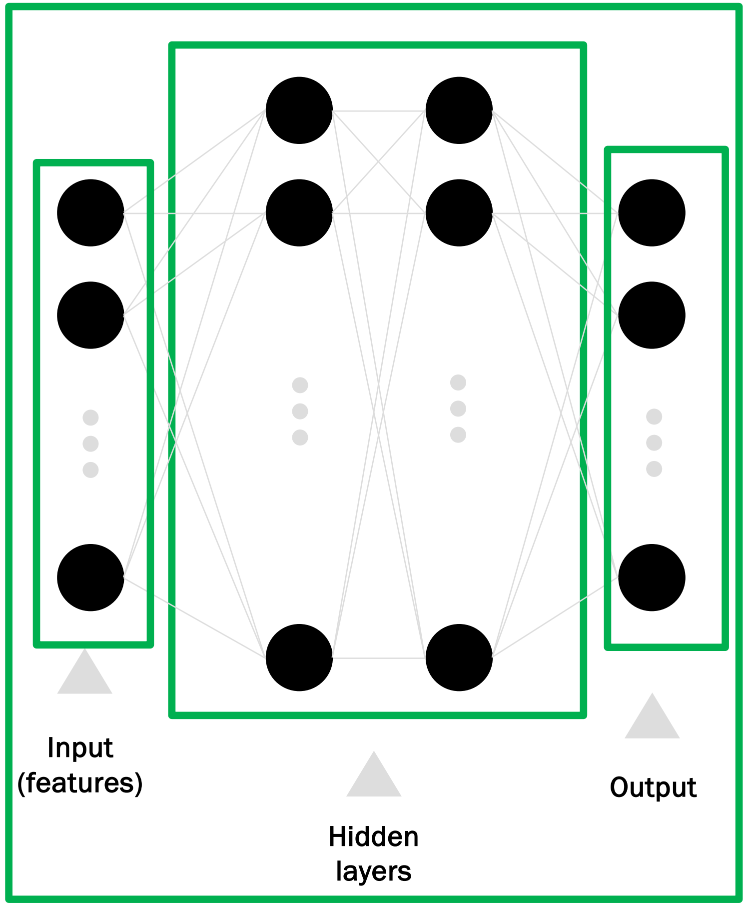
- Neural Networks automatically relate input (e.g., examples) to output
- We learn how inputs are related to a given output through **model training**
- Different relationships between input and output are compounded in the **hidden layers** of a model



Explainable AI (XAI)

- **Goal:** make model behavior more transparent or understandable
- Different XAI methods
  - *"One size fits all" (model-agnostic)*
  - *Model specific methods*
  - *Dimensionality Reduction*
- Explanations can explain the **Model** or **Output**
- **Interpretability v Explainability**

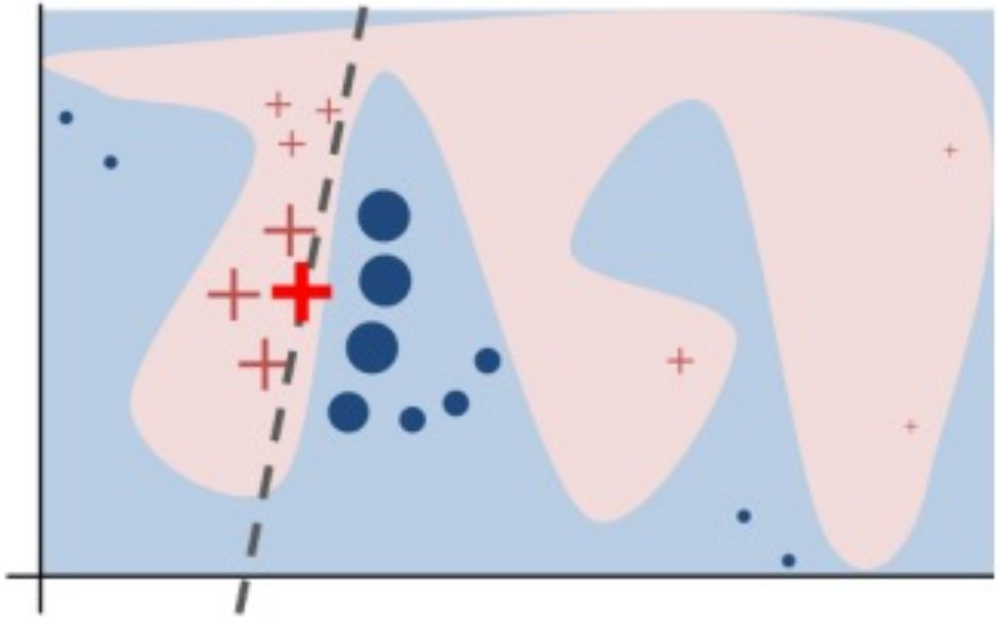
Explanation methods can be applied at different places in ML lifecycle



The Paper | Big Picture

- Produce Explanations that meet DP constraints
  - *Should maintain ε,δ privacy for a set of queries*
  - *Maintain high explanation quality*
- Maintain Privacy budget, ε
  - *Keep budget should be as low as possible*

LIME | a feature-based technique



Big Idea

- For an example, look at random examples (in red and blue) and identify most important features
- Identify a linear decision boundary
- Explanation should be easily understood

General Info

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

- Handles text and image classification
- Global explanation can be performed by gathering many local explanations

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

LIME

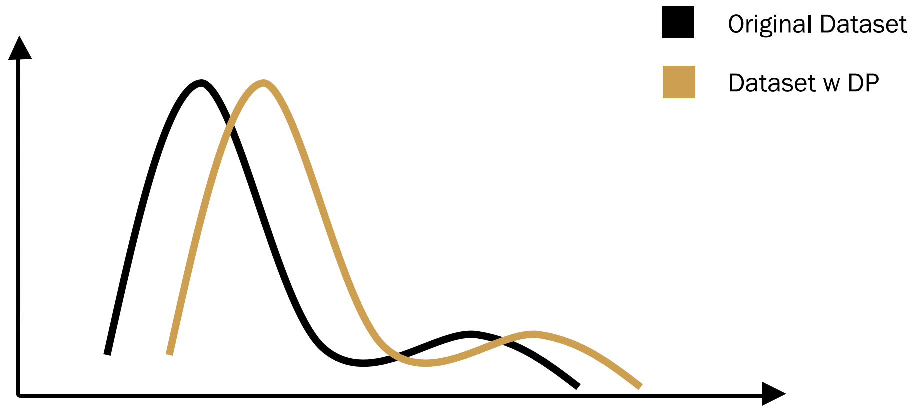
(a) Original Image

(d) Explaining *Labrador*

The top 3 classes predicted are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24) and "Labrador" (p = 0.21)

Differential Privacy

- Can be applied in 3 major areas
  - *Towards dataset*
  - *Towards gradient learning*
  - *Towards output*
- Downsides of Differential Privacy
  - *Privacy vs Utility*
  - *Longer training*
  - *High privacy spending*



Differential Privacy

- ε vs. (ε ,δ) differential privacy
- This paper applies DP to gradient calculations (training and explanations)

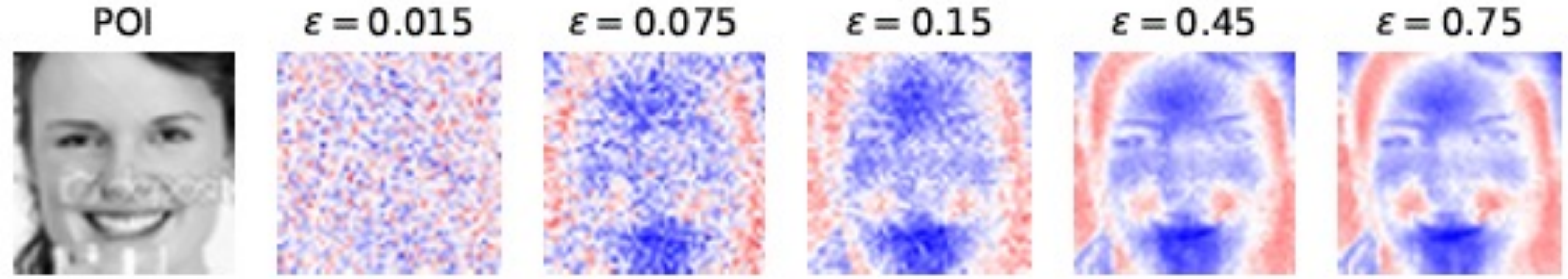
$$\frac{\Pr[\mathcal{A}(D_1) \in S]}{\Pr[\mathcal{A}(D_2) \in S]} \leq e^{\varepsilon},$$

Epsilon differential privacy

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta.$$

Epsilon, delta differential privacy

Differential Privacy



The effect of varying ϵ on explanation quality of Algorithm 1(In Appendix B). The color blue (red) indicates positive (negative) influence. Brighter colors indicate greater influence.

Previous Work

- QII
  - *Applies Differential Privacy to black-box explanations*
- DP Locally Linear Maps
  - *Represent classification using groups of matrices*
- Simplifying Neural Networks
  - *To a linear combination of logistic regression models*

Important Note

- *If a model is DP* then model explanations affect DP guarantees and information can be leaked
  - *Guarantees are dependent on how often information is accessed*
- *If a model is not DP*, then explanations can give an idea of the training data distribution

Methodology | Assumptions

- Any model can be used
- Any linear approximation explanation method can be used
- We can only explain a finite number of explanations
- We have access to the following:
  - *Explanations*
  - *Model Input / output*

Methodology

- **Paper Contribution**
  - *Adds privacy constraints to model explanations*
  - *Evaluate the effect of differential privacy on explanations*
- **Terminology**
  - *Model* $f_{\mathcal{D}} : \mathbb{R}^n \to \mathcal{R}$
  - *Dataset D*
  - *Explanation Dataset* $\mathcal{X} = \{(x_1, f_{\mathcal{D}}(x_1)), \ldots, (x_m, f_{\mathcal{D}}(x_m))\}.$
  - *Model Explanation* $\phi(\vec{z}, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$ *OR*
- **Prevent possible attack scenario**
  - *Explanations leak sensitive information* $\phi(\vec{z}, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X}))$

A big motivation
of paper

Methodology | Big Picture

- Keep Good Explanation Accuracy
- Maintain DP promises
- Reuse queried information when possible
  - *Very important to solution*
- Maximize given Privacy budget
  - *Use previous queries to set reduce training time*
- Just applying DP to Explanations isn’t enough. Look for for other potential privacy leaks

Methodology | Explanations

- Reuse explanations when possible
  - *A data point must come from an area that's already been explained*
  - *Reduces privacy spending*
- **Training and Explanations cost!**
  - *Choose initialization point wisely*
  - *Avoid overspending privacy budget*

$$||\nabla \mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_{h+1})||.$$

Initialization value should result in z_{h+1} giving a similar explanation

Methodology | Building a DP Explainability Metric

- We maintain DP constraints of explanations by controlling how our gradient function grows
  - *Apply a sensitivity bound to gradient calculation. Bounds weights*
- **Model explanations** and **Explanation Dataset** should offer no insight on private info

Procedure DPGD-Explain(ϕ, σ, T) :

$\phi^{\{t\}} \leftarrow \phi$

for $t = 1, \ldots T - 1$ **do** :

$$\xi_t \leftarrow \left(\phi^{\{t\}} - \eta(t)\left[\nabla \mathcal{L}(\phi, \vec{z}) + \mathcal{N}(0, \sigma^2 \mathrm{I})\right]\right)$$

$$\phi^{\{t+1\}} \leftarrow \underset{\phi \in C_{2,1}}{\arg\min} \|\phi - \xi_t\|$$

Return : ϕ^T

$$\mathcal{F}(c, \vec{z}) := \left\{ \alpha(\cdot) : \begin{array}{l} \alpha(\cdot) \textit{ is non-increasing and} \\ \forall \vec{x} \in \mathbb{R}^n, \alpha(\|\vec{x} - \vec{z}\|) \leq \frac{c}{2\|\vec{x}-\vec{z}\|(\|\vec{x}-\vec{z}\|+1)} \end{array} \right\}.$$

Sensitivity bound for gradient calculations

Methodology | Building a DP Explainability Metric

- Compare explanation quality
- Explanations should be ***“training data safe”***

$$\mathcal{E}(\phi, \vec{z}, f(\mathcal{X})) \triangleq \mathbb{E}\left[\mathcal{L}(\phi, \vec{z}, f(\mathcal{X}))\right] - \mathcal{L}(\phi^*(\vec{z}, f), f(\mathcal{X})).$$

$$\begin{aligned}&\Pr[\phi(\vec{z}_i, \mathcal{X}, f_{\mathcal{D}}(\mathcal{X})) : i = 1, \ldots k] \\ &\quad \leq \mathrm{e}^{\tilde{\epsilon}} \cdot \Pr[\phi(\vec{z}_i, \mathcal{X}, f_{\mathcal{D}'}(\mathcal{X})) : \forall i = 1, \ldots, k] + \tilde{\delta}, \qquad (5)\end{aligned}$$

Methodology

- 2 ways to apply Differential Privacy to model explanations:
    1. *Interactive DP*
    2. *Non-Interactive DP*

Methodology | Interactive Method

- Reuse previous explanations if they can be used to explain a different example

---

Algorithm 1: Adaptive DP for Model Explanation

Input: Queries $\{\vec{z}_1, \ldots\} \in \mathbb{R}^n$ arriving one by one, explanation dataset $\mathcal{X}$, privacy budget (ϵ, δ), the minimum per-query privacy loss $(\epsilon_{min}, \delta_{min})$, and the number of GD steps T;

$\mathcal{H} \leftarrow \emptyset$, ϵ_{spent}, $\delta_{spent} \leftarrow 0$;

$\epsilon_{ite} \leftarrow \frac{\epsilon_{min}}{\sqrt{8T \log \frac{2}{\delta_{min}}}}$; // Privacy budget to spend per iteration

$\sigma_{min} = \frac{\sqrt{2\log(2.5T/\delta_{min})}}{m \cdot \epsilon_{ite}}$; // Variance needed for Gaussian mechanism

$d \leftarrow \frac{\log T}{\sqrt{T}}$; // Distance bound required according to Thm. 4.1

for $h = 1, \ldots, \infty$ **do**

  if $\exists \vec{z}_j \in \mathcal{H}$ *with* $\|\vec{z}_h - \vec{z}_j\| \leq d$ & $\texttt{Flag}(\vec{z}_j) = \top$ **then**

    $\phi^{Priv}(\vec{z}_h) \leftarrow \phi^{Priv}(\vec{z}_j)$; // Use a nearby point (Thm. 4.1)

    report: $\phi^{Priv}(\vec{z}_h)$; // Report explanation of $\vec{z}_h$

    $\mathcal{H}.\text{append}(\vec{z}_h : \phi^{Priv}(\vec{z}_h), \texttt{Flag}(\vec{z}_h) = \perp)$

  else

    $\phi^{best}, \sigma, T' \leftarrow$ Parameters-DPGD$(\vec{z}_h, \mathcal{H}, \epsilon_{ite}, \sigma_{min}, \mathcal{X}, T)$;

    $\phi^{Priv}(\vec{z}_h) \leftarrow$ DPGD-Explain$(\phi^{best}, \sigma, T')$;

    Update ϵ_{spent}, δ_{spent}; // via the Strong Composition Theorem

    if $\epsilon_{spent} > \epsilon$ *or* $\delta_{spent} \geq \delta$ **then**

      break; // Privacy budget is exhausted

    end

    report: $\phi^{Priv}(\vec{z}_h)$;

    $\mathcal{H}.\text{append}(\vec{z}_h : \phi^{Priv}(\vec{z}_h), \texttt{Flag}(\vec{z}_h) = \top)$;

  end

end

Methodology

- Reuse explanations for previous points
  - *Possible only if in a subregion that is well explained*
- Apply DP to explanation

Algorithm 2: Adaptive DP for Parameter Selection

Input: $\vec{z}_h \in \mathbb{R}^n$, explanation dataset $\mathcal{X}$, History $\mathcal{H}$, privacy spending for parameter selection ϵ_{para}, and the number of GD steps T, minimum variance σ_{min};

Output: Input variables for `DPGD-Explain()` for the query $\vec{z}_h$;

Procedure `Parameters-DPGD`$(\vec{z}_h, \mathcal{H}, \epsilon_{para}, \sigma_{min}, \mathcal{X}, T)$

  if $\mathcal{H} == \emptyset$ **then**

    Arbitrary $\phi \in C_{2,1}$;

    return ϕ, σ_{min}, T;

  else

    $\phi^{best} \leftarrow \phi^{Priv}(\vec{z}_j)$ with

     $\Pr \propto exp\left(-m \cdot \epsilon_{para} \cdot \frac{\|\nabla \mathcal{L}(\phi^{Priv}(\vec{z}_j), \vec{z}_h)\|}{2}\right)$ for $\vec{z}_j \in \mathcal{H}$

    $\beta \leftarrow \|\nabla \mathcal{L}(\phi^{best}, \vec{z}_h)\|$;

    $\sigma \leftarrow \max\left(\frac{\beta}{\sqrt{n}}, \sigma_{min}\right)$;

    $a \leftarrow \frac{\log \frac{1}{\sqrt{n}\sigma}}{\log \log T}$; // Thm. 4.2

    if $a > \frac{1}{2}$ **then**

      $T' \leftarrow (\sqrt{n}\sigma)^{1-\frac{1}{2a}} T$;

    else

      $T' \leftarrow T$;

    end

  end

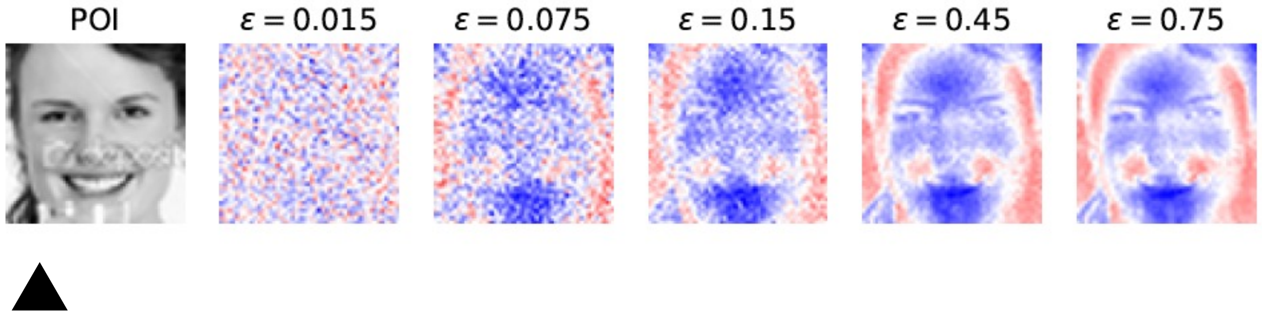
  return ϕ^{best}, σ, T';

Results

- 3 Datasets
  - *Face (Image)*
  - *Good v Bad Movie Recommendations (Text)*
  - *ACS13 (Tabular / Categorical)*

Table 1: Summary of mean ± Variance of loss in utility for each dataset for explanation generated by Algorithm 1 (Theorem B.1).

Dataset	$\epsilon = 0.01, \delta = 10^{-6}$	$\epsilon = 0.1, \delta = 10^{-6}$
Face	$1.4 \times 10^{-2} \pm 4.3 \times 10^{-3}$	$2.2 \times 10^{-3} \pm 2.1 \times 10^{-4}$
Text	$2.7 \times 10^{-3} \pm 7.8 \times 10^{-4}$	$4.3 \times 10^{-4} \pm 9.4 \times 10^{-6}$
ACS13	$5.7 \times 10^{-3} \pm 1.3 \times 10^{-3}$	$2.6 \times 10^{-4} \pm 6.3 \times 10^{-5}$



Stringent privacy constraints harm explanations

Results

1. Privacy Budget spending per explanation improves over time
2. Differential Privacy can make explaining some regions in a class hard
    1. *Randomized noise hurts explanation fidelity*

Future Work (Authors)

- Build privacy-preserving, interpretable models
  - *Understand how DP affects other XAI methods*
- Highlight tradeoffs between privacy and XAI
  - *Explanation accuracy is lower in DP models*
- Compare other privacy methods and XAI
- Assess Privacy and XAI tradeoffs for minority groups

Thoughts

- Paper is written from the viewpoint "Explanations are dangerous"
- What do domain specific explanation look like under DP?
- General: how do you go about making proofs for this?
  - *Do you observe behavior for a few instances?*
- I like that they give guidance on how to handle privacy / explanation utility tradeoff (e.g., how to choose privacy level)
- Do you trust the explanations given especially when noise has been added?
- Not a good choice for very big datasets (privacy will be less stringent)
- What are your thoughts?