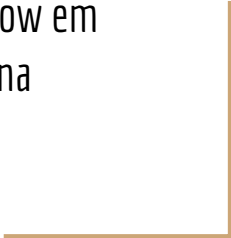




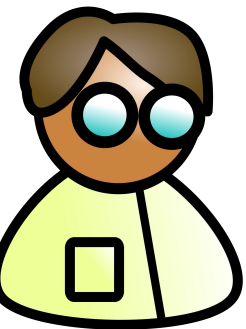
Projeto Orientado em Computação 2: Pitch Final

Relações entre Fairness, Privacidade e
Quantitative Information Flow em
Aprendizado de Máquina

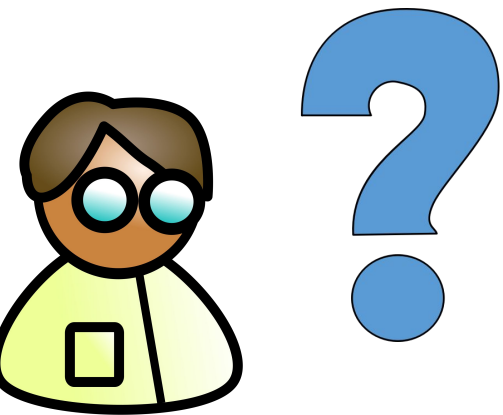


Nome: Artur Gaspar da Silva
Orientador: Mário Sérgio Alvim

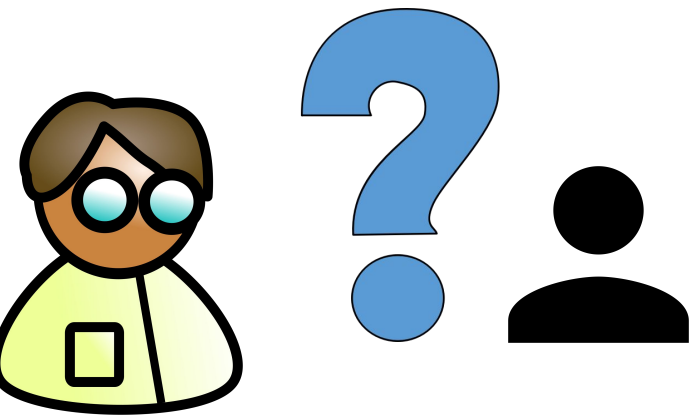
Local Differential Privacy



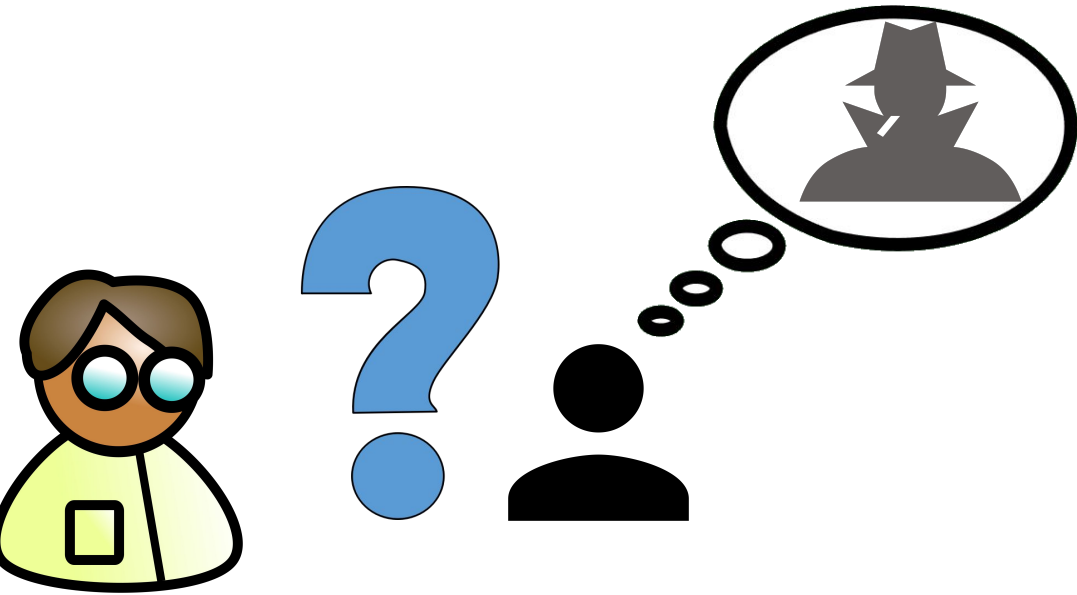
Local Differential Privacy



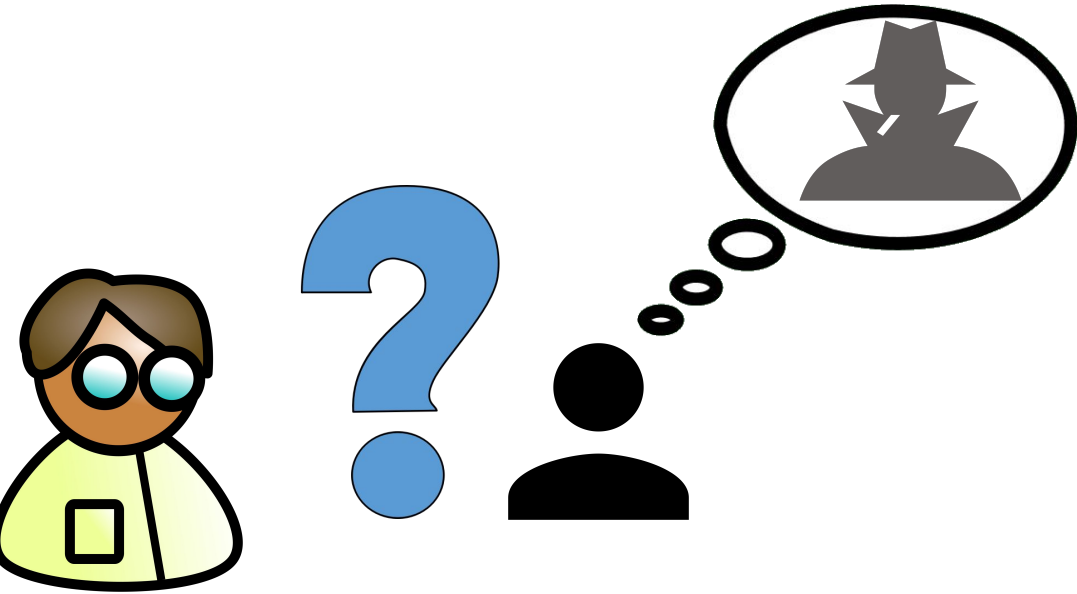
Local Differential Privacy



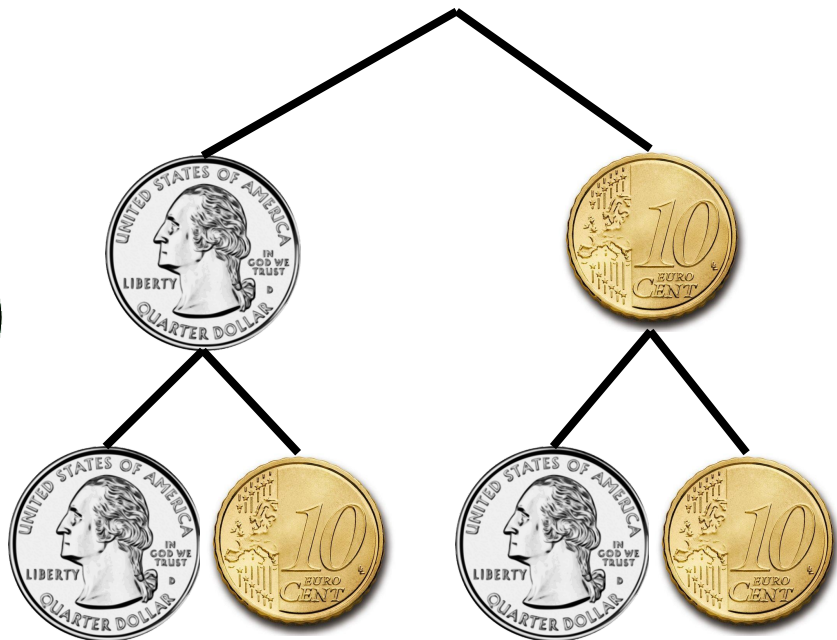
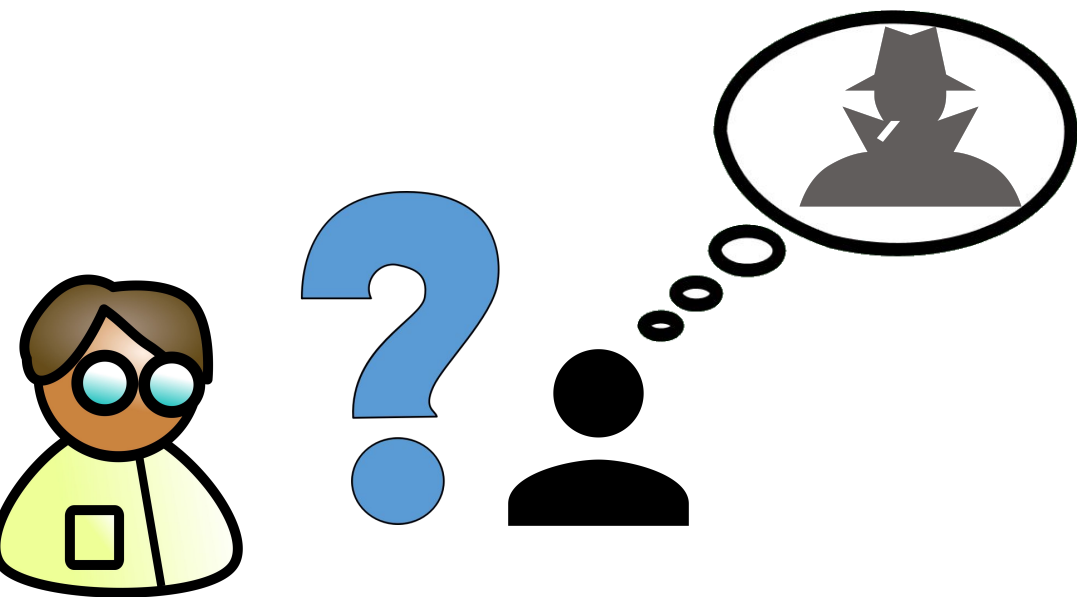
Local Differential Privacy



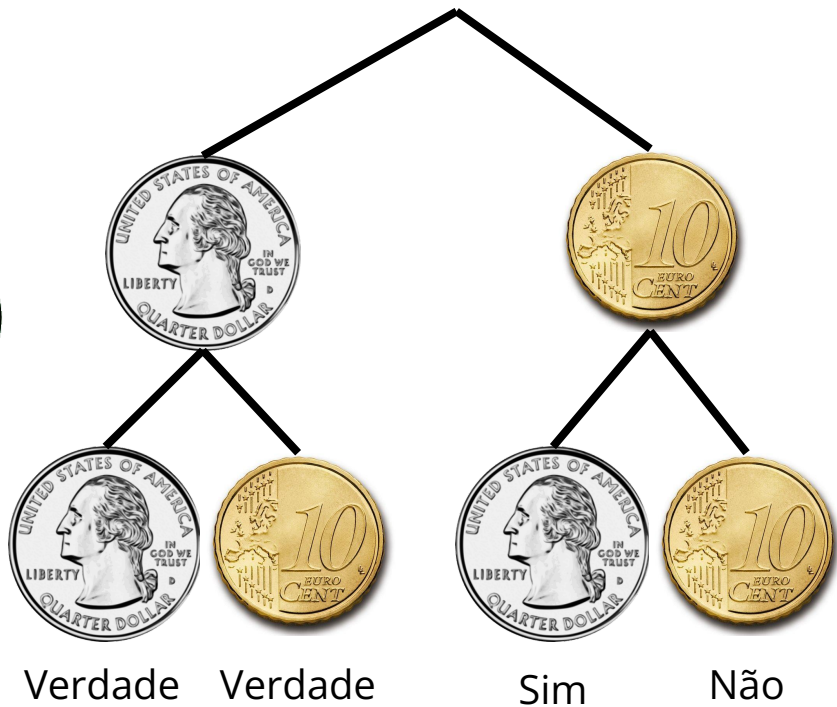
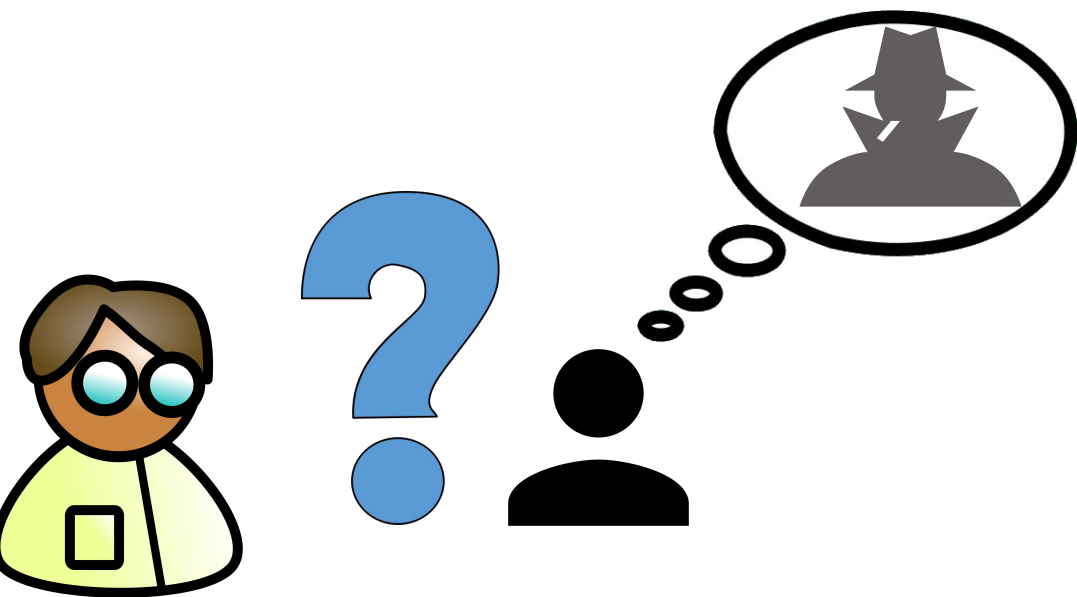
Local Differential Privacy



Local Differential Privacy



Local Differential Privacy



Fairness: (Conditional) Statistical Disparity



Fairness: Equal Opportunity Difference

$$P(\hat{Y}=1 | A=1, Y=1) - P(\hat{Y}=1 | A=0, Y=1)$$

Parâmetros epsilon e delta

- Epsilon representa o quanto de incerteza do adversário toleramos, no pior caso, entre o valor real e qualquer outro valor da informação secreta.
- Delta representa, de certa forma, qual a margem tolerável de erro para o valor de epsilon real estar errado.
- Um artigo recente de 2024 modelou epsilon usando Max-QIF, que considera cenários com probabilidade não nula.
- A ideia é considerar delta_Max-QIF , cenários com probabilidade pelo menos delta, pois cenários dentro da margem seriam automaticamente aceitos.

0 efeito em fairness sem reversão de ruído

- Artigos recentes consideram como obfuscar variáveis sensíveis pode automaticamente melhorar métricas de fairness.
- No entanto, esses artigos consideram que a etapa de reversão de ruído em LDP não é realizada, o que não condiz com a realidade.
- Problemas de privacidade surgem se LDP for aplicada apenas à variável sensível, que acaba não sendo protegida.
- Os resultados desses artigos **não condizem com a realidade** se a reversão de ruído é realizada.
- Vários estudos existentes consideram que não há reversão de ruído.

Distribuição do budget de privacidade

- Simplesmente aplicar ruído a variáveis sensíveis não é o bastante de um ponto de vista de privacidade, a menos que possamos garantir *a priori* que as outras variáveis não estão correlacionadas com as sensíveis.
- Aplicar ruído a todas as variáveis pode ser desnecessário.
- Possivelmente é mais eficiente aplicar ruído às variáveis sensíveis e a variáveis correlacionadas.
- Após verificar a viabilidade de fazer isso, concluímos que seria necessário uma modelagem do conhecimento a priori sobre a distribuição real dos dados.

Objetivos do POC1

- Estudar conceitos de Causalidade, Fairness, Privacidade, Acurácia, Fluxo de Informação e Explicabilidade, em Aprendizado de Máquina.
- Explorar conexões na literatura entre os tópicos mencionados.
- Explorar novas possíveis conexões entre os tópicos mencionados.

Objetivos do POC2

- Estudar mais a fundo conceitos de Privacidade Diferencial e Fairness
- Explorar impacto de métodos de obfuscagem em fairness
- Explorar como budget de privacidade (Local Differential Privacy, LDP) pode ser dividido entre variáveis de importância diferente
- Explorar conexões com Quantitative Information Flow

Etapas concluídas no POC1

- Exploração de conceitos chave de causalidade (como a *Ladder of Causality*) (capítulo 1)
- Estudo de diferentes ferramentas e modelos causais (cap. 1)
- Estudo inicial de inferência causal a partir de dados observados (cap. 2)
- Estudo de inferência de efeitos causais a partir de dados e observações qualitativas (cap. 3)
- Estudo da modelagem de Planos e Ações sob lentes causais (cap. 4)
- Estudo de Efeitos Diretos e Indiretos sob lentes causais (cap. 4)
- Análise de exemplos de ciências sociais e econômicas (cap. 5)
- Estudo do paradoxo de Simpson, Confounding e Colapsabilidade sob lentes causais (cap. 6)
- Estudo de contrafactuais (cap. 7)
- Estudo de como estimar limites e contrafactuais a partir de experimentos imperfeitos (cap. 8)
- Estudo de probabilidades de causa (causas necessárias e suficientes) (cap. 9)
- Estudo da noção de Actual Causes (cap. 10)
- Estudo de exemplos clássicos (cap. 11)
- Revisão da literatura em busca de conceitos chave de Fairness, Explicabilidade e Privacidade em aprendizado de máquina.
- Revisão da literatura em busca de relações entre os tópicos mencionados.

Etapas concluídas no POC2

- Exploração de conceitos chave de LDP e QIF.
- Várias tentativas de modelar o parâmetro delta de (delta, epsilon)-LDP com QIF. Conclusão de que novas abordagens serão necessárias no futuro.
- Discussão do impacto em fairness de métodos de ofuscação relacionados a privacidade. Conclusão de que muitos dos estudos existentes fazem a suposição forte de que o coletor de dados não executará o processo de reversão de ruído dos dados.
- Explorado como um budget de privacidade pode ser melhor distribuído entre variáveis com níveis de importância diferente. Conclusão de que uma modelagem exata da informação a priori sobre a distribuição real dos dados é necessária.

Fim!