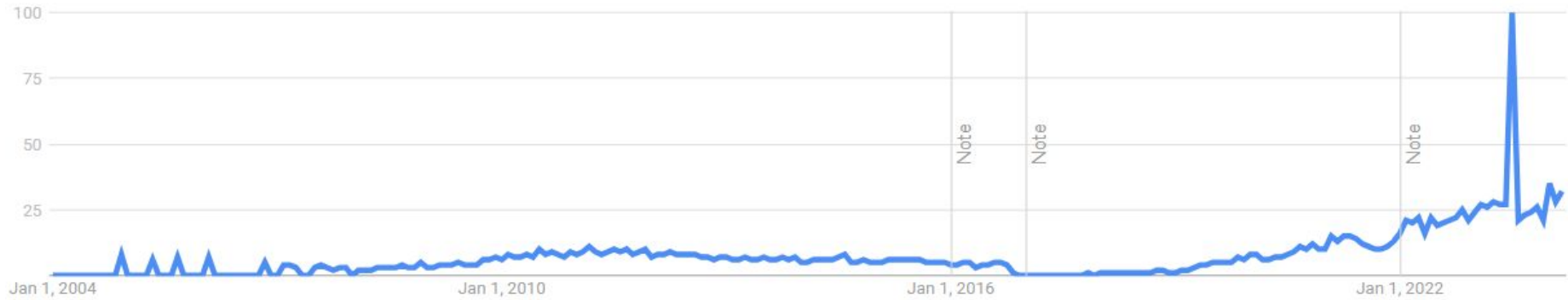


data

Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

André Mitri
@andremitri
12/03/2024

Interesse **percentual** ao longo dos anos no Google para o termo “Explainable Artificial Intelligence”



Definição de XAI

According to DARPA, XAI aims to “produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners



Objetivo de XAI

As stated by FAT, “is to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms”



Interpretável

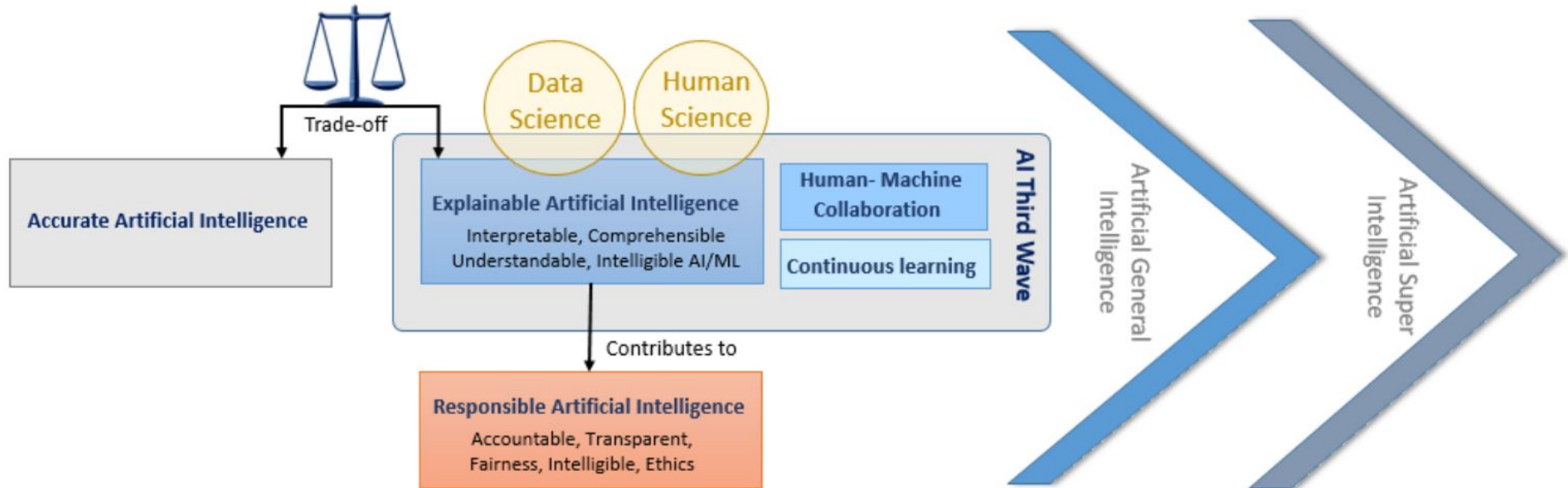
- **Entender como** decisões são tomadas
- Entender como o modelo processa os dados de **entrada e chega a uma saída** específica.
- Importante em campos onde a **transparência** é necessária

Explicável

- **Fornecer explicações** sobre as decisões ou previsões feitas por um modelo.
- Foco em comunicar **informações sobre o modelo** de uma maneira que seja compreensível para os usuários humanos



Alguns conceitos

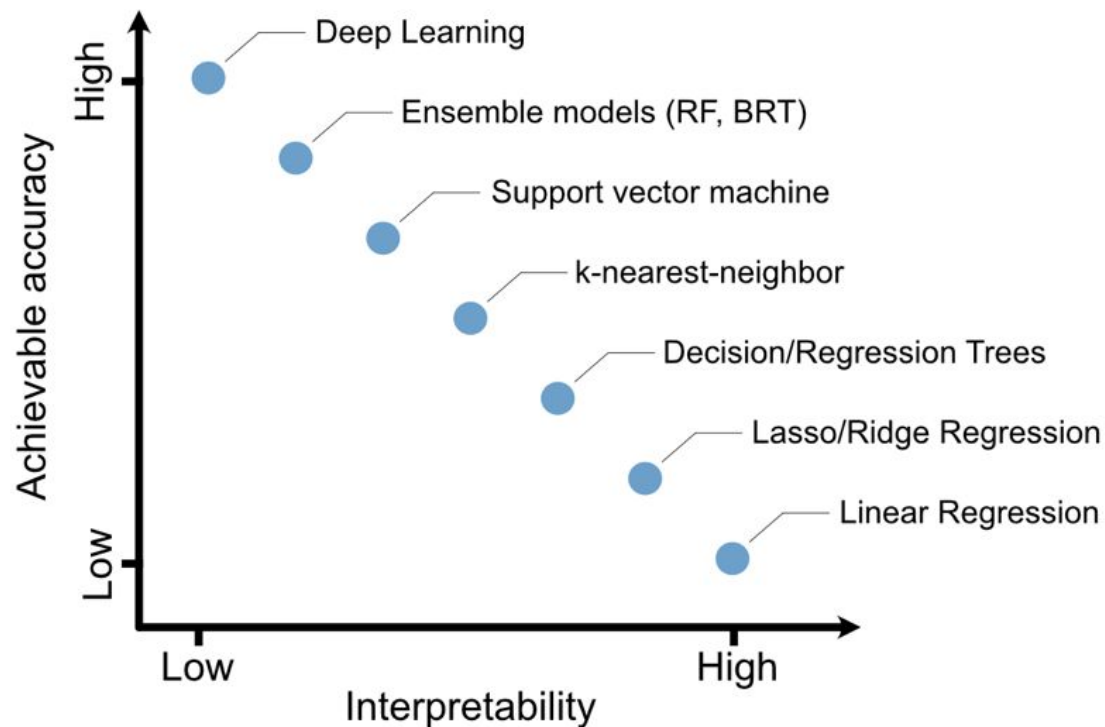


4

motivos XAI



Tradeoff:
Interpretabilidade
X
Acurácia





Estratégias de XAI



Two L-shaped lines, one blue and one magenta, framing the word 'Visualização'. The blue line is on the left, and the magenta line is on the right.

Visualização



Surrogate (Substituto)

Modelo mais simples
para explicar um
mais complexo
(Exemplo: LIME)

PARTIAL DEPENDENCE PLOT (PDP)

Representação gráfica
que ajuda a visualizar a
relação entre uma ou
mais variáveis de
entrada e as previsões
do modelo

INDIVIDUAL CONDITIONAL EXPECTATION

Extensão do PDP, releva
interações e diferenças
individuais
desagregando o output
do PDP





Métodos de Influência



Métodos de Influência

Análise Sensitiva:

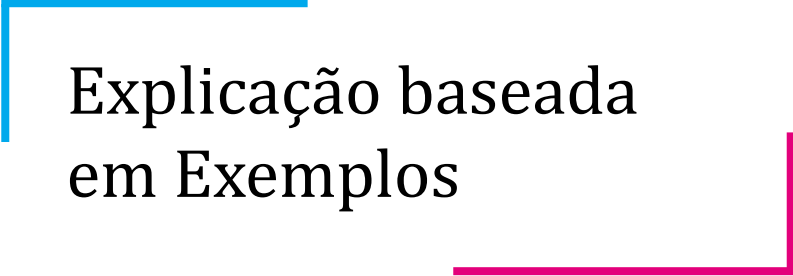
como uma rede neural é influenciada pela perturbação em seus pesos. Verifica se o comportamento e saída da rede se mantém estáveis.

LAYER-WISE RELEVANCE PROPAGATION

(LRP):

Redistribui a função de predição de trás para frente, começando na output layer e fazendo backpropagation para input layer. Explica predições relativamente ao estado de máxima incerteza.





Explicação baseada em Exemplos



Protótipos e Críticos

- Protótipos são seleções de instâncias dos dados
- Participação dos itens por similaridade aos protótipos
- Críticos são instâncias não bem representadas pelos protótipos
- Críticos são mostrados ao modelo

Explicações Contrafactuais

- Descrevem condições mínimas para tomada de uma decisão alternativa
- Explicação de uma única predição em contraste com exemplos adversários, com ênfase em reverter a predição e não em explicá-la





Avaliando Explicações



TABLE 2. Summary of explainability techniques.

Techniques	References	Intrinsic/Post-hoc	Global/Local	Model-specific/ Model-agnostic
<i>Decision trees</i>	[139], [140], [141], [142], [143]	I	G	SP
<i>Rule lists</i>	[66], [143], [144], [145], [146]	I	G	SP
<i>LIME</i>	[84], [85], [102], [147]	H	L	AG
<i>Shapely explanations</i>	[101]	H	L	AG
<i>Saliency map</i>	[87], [88], [89], [90], [91], [96], [97]	H	L	AG
<i>Activation maximization</i>	[82], [83]	H	G	AG
<i>Surrogate models</i>	[106], [107], [84]	H	G/L	AG
<i>Partial Dependence Plot (PDP)</i>	[108], [51], [110]	H	G/L	AG
<i>Individual Conditional Expectation (ACE)</i>	[112], [113]	H	L	AG
<i>Rule extraction</i>	[74], [114], [115], [116], [117], [118]	H	G/L	AG
<i>Decomposition</i>	[93], [94], [95]	H	L	AG
<i>Model distillation</i>	[49], [123], [124], [125], [126], [127]	H	G	AG
<i>Sensitive analysis</i>	[129], [130]	H	G/L	AG
<i>Layer-wise Relevance Propagation (LRP)</i>	[131]	H	G/L	AG
<i>Feature importance</i>	[113], [132], [86]	H	G/L	AG
<i>Prototype and criticism</i>	[133], [134], [135], [136]	H	G/L	AG
<i>Counterfactuals explanations</i>	[137]	H	L	AG

I: Intrinsic, H: Post-hoc, G: Global, L: Local, SP: Model-specific, AG: Model-agnostic

