Explainable AI (XAI)

PTC3101 - Engenho e Arte do Controle Automático

Diego Lopes e Thomas Ferraz

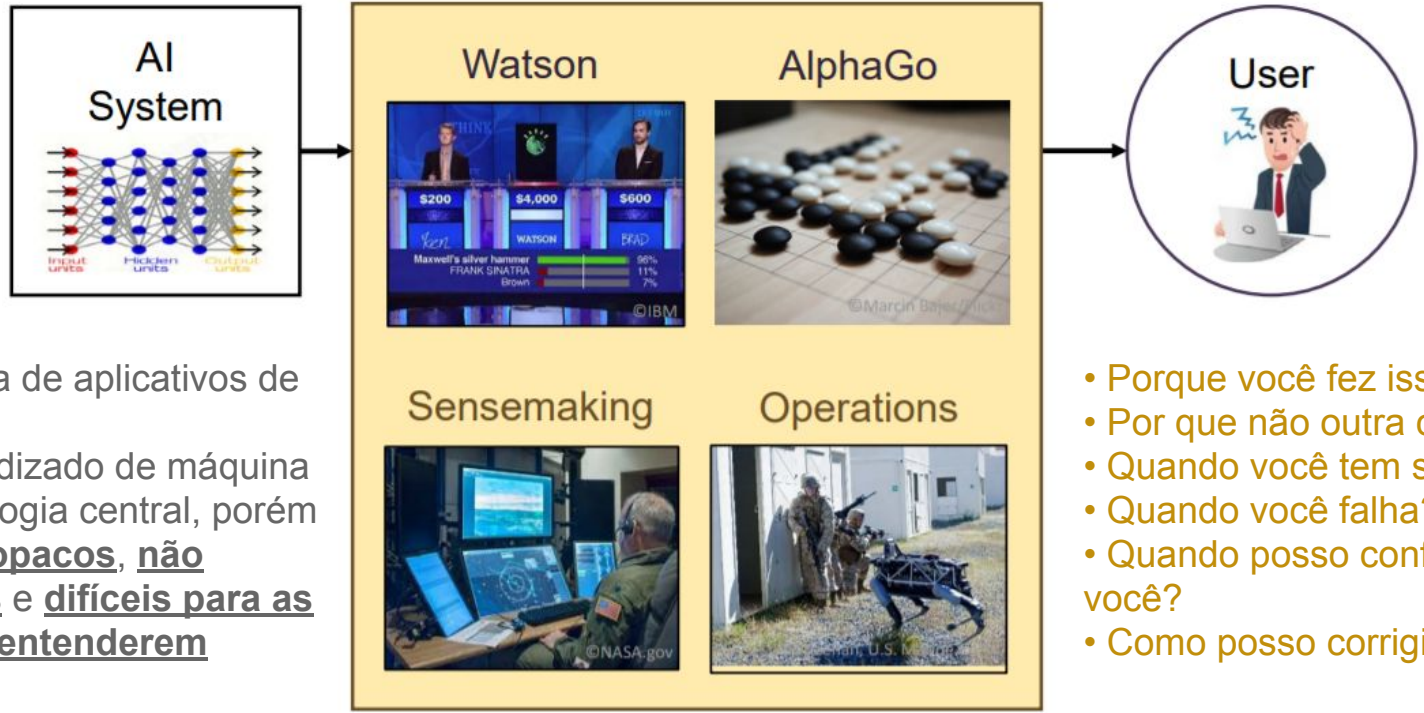
Introdução

- Panorama da Inteligência Artificial
  - Modelos baseados em aprendizado (em especial Deep Learning) estão em alta.
  - Capacidade de processamento em hardware tem possibilitado resultados impressionantes.
    - Para muitas aplicações, superando a capacidade humana de reconhecimento.
    - Ex.: GPT-3 (OpenAI, 2020) - A IA que produz IA
  - Outros tipos de IA ficaram um pouco de lado até pouco tempo.

Introdução

- Crise da humanidade (2020)
  - Vieses das redes neurais levam a perpetuação de racismo e sexismo (J ZOU e L SCHIEBINGER, Nature, 2018)
  - Onda de protestos nos EUA por direitos civis.
  - Microsoft, IBM e Amazon interrompem o investimento em desenvolvimento de pesquisa e de produtos em determinadas aplicações de Deep Learning.
- Novas áreas de pesquisa em IA.
  - Retomada pesquisa em lógica (Symbolic AI).
  - Explainable AI
  - Neuro-Symbolic AI / Neural Logic Machines (NLM)

Motivação

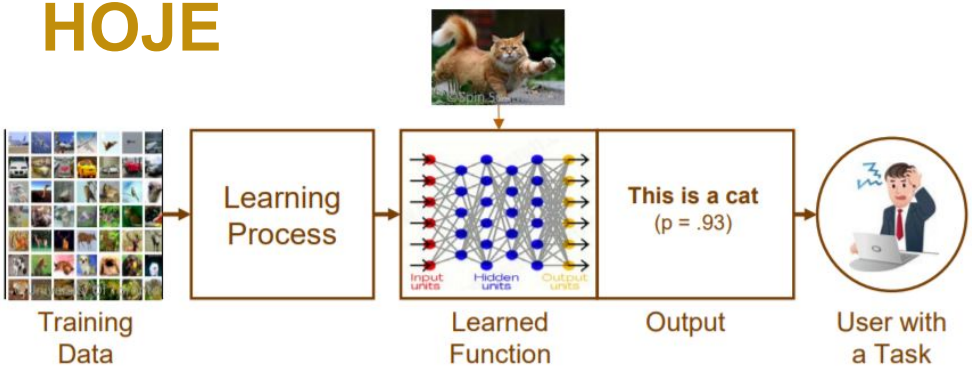


- Nova era de aplicativos de IA
- O aprendizado de máquina é a tecnologia central, porém eles são **opacos**, **não intuitivos** e **difíceis para as pessoas entenderem**

- Porque você fez isso?
- Por que não outra coisa?
- Quando você tem sucesso?
- Quando você falha?
- Quando posso confiar em você?
- Como posso corrigir um erro?

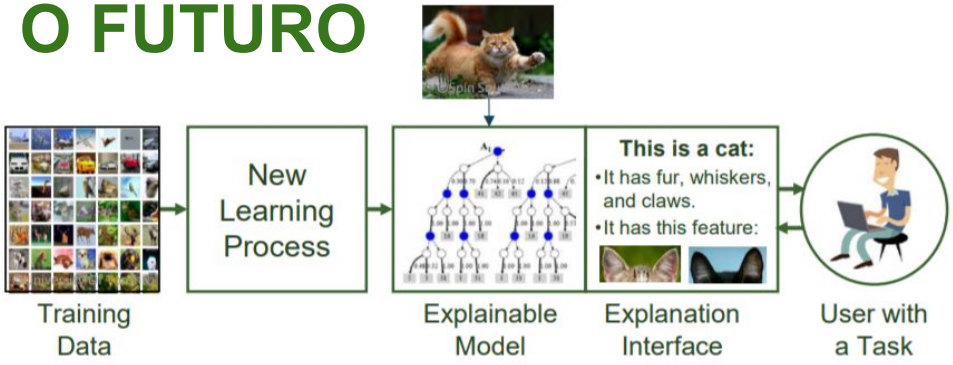
Objetivo da Explainable AI

HOJE



- Porque você fez isso?
- Por que não outra coisa?
- Quando você tem sucesso?
- Quando você falha?
- Quando posso confiar em você?
- Como posso corrigir um erro?

O FUTURO



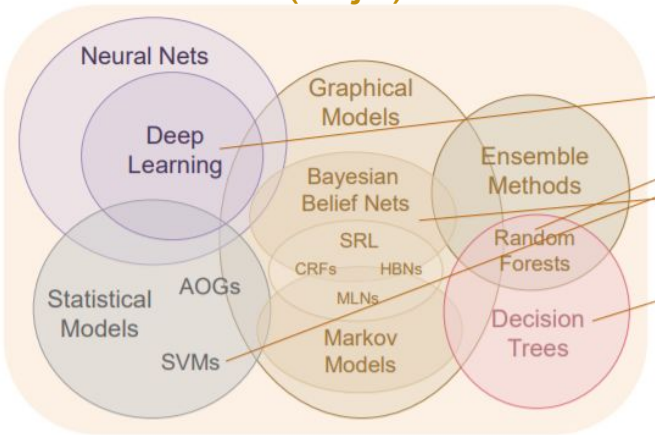
- Entendo o porquê
- Entendo porque não
- Sei quando você terá sucesso
- Sei quando você vai falhar
- Sei quando confiar em você
- Sei porque você errou

Desempenho x Explicabilidade

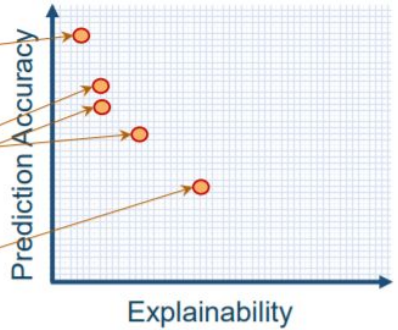
Desafio

Criar um conjunto de técnicas de aprendizado de máquina que produza modelos **mais explicáveis**, mas que mantenha um alto nível de desempenho de aprendizado.

Técnicas de aprendizado (hoje)



Explicabilidade (qualitativamente)

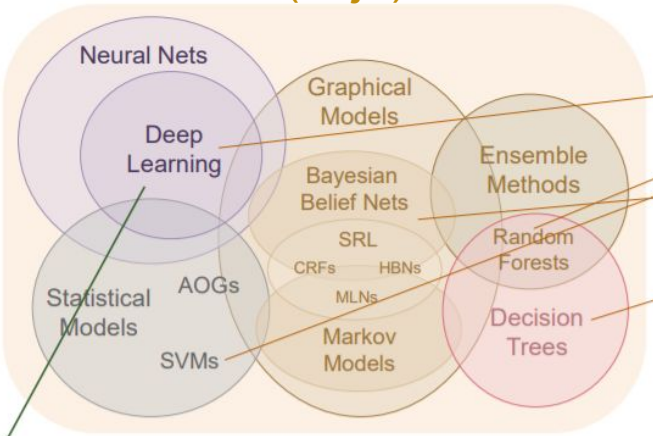


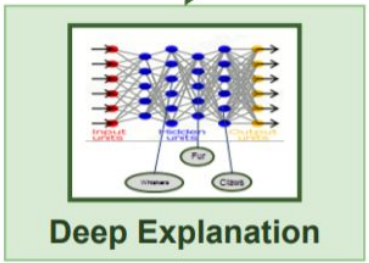
Desempenho x Explicabilidade

Desafio

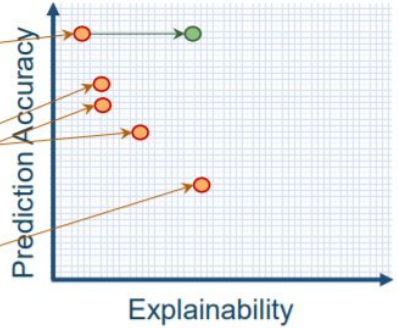
Criar um conjunto de técnicas de aprendizado de máquina que produza modelos **mais explicáveis**, mas que mantenha um alto nível de desempenho de aprendizado.

Técnicas de aprendizado (hoje)



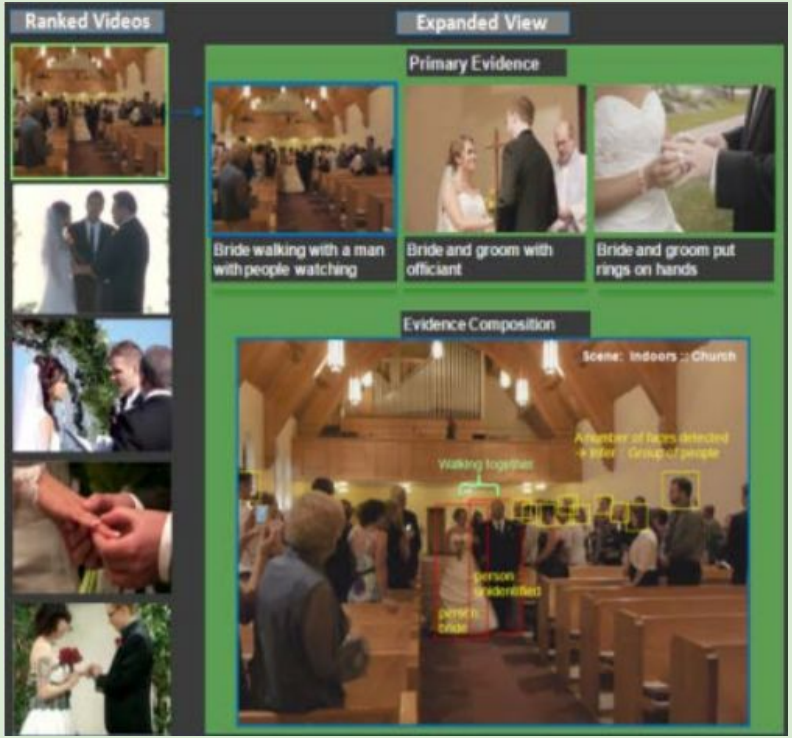


Explicabilidade (qualitativamente)



Deep Explanation Learning

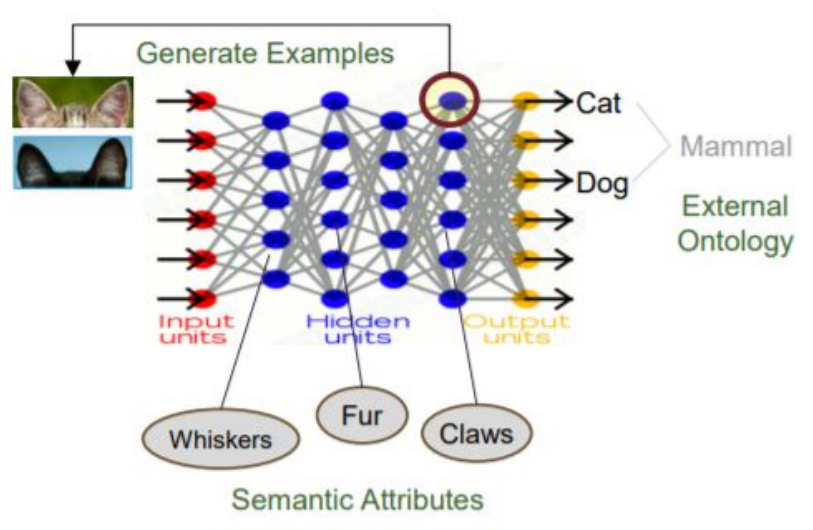
Classificação multimídia de eventos



- Exemplo de classificação de eventos.
- O sistema classificou este vídeo como um **casamento**.
- Os quadros acima apresentam suas evidências para a classificação como um casamento.

Deep Explanation Learning

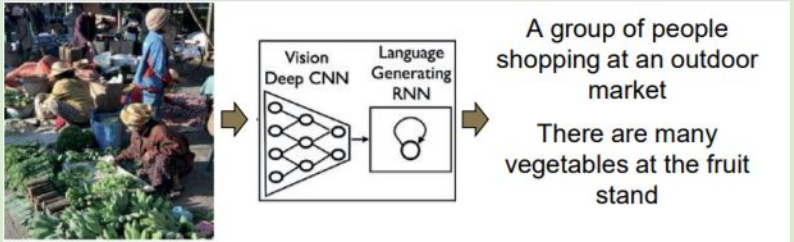
Aprendizado de Associações Semânticas



- Treinar rede neural para associar atributos semânticos a nós de camadas ocultas;
- Treinar rede neural para associar nós rotulados com ontologias conhecidas;
- Gerar exemplos de nós proeminentes, mas sem rótulos, para descobrir rótulos semânticos;
- Gerar grupos de exemplos de nós proeminentes;
- Identificar as melhores arquiteturas, parâmetros e sequências de treinamento para aprender os modelos mais interpretáveis;

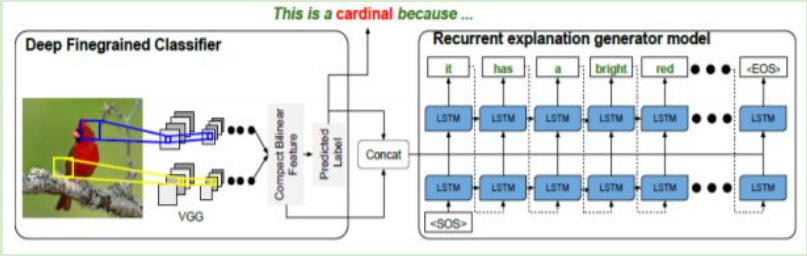
Deep Explanation Learning

Gerar legendas de imagens



- Um CNN é treinado para reconhecer objetos em imagens
- Um RNN gerador de linguagem é treinado para traduzir recursos da CNN em palavras e legendas.

Gerar Explicações Visuais



This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

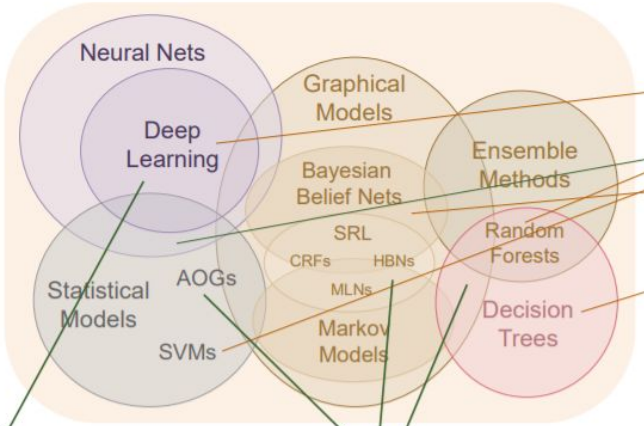
- Classificar as espécies de aves com 85% de precisão;
- Associar descrições de imagem (características discriminativas da imagem) com definições de classe (características discriminativas independentes de imagem da classe);
- Explicação limitada (na melhor das hipóteses indireta) da lógica interna e utilidade limitada para compreender os erros de classificação.

Desempenho x Explicabilidade

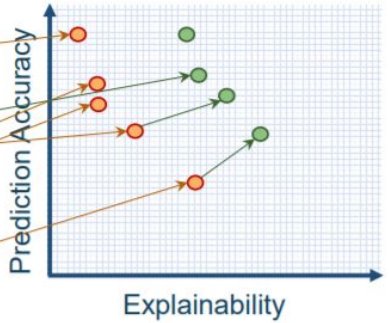
Desafio

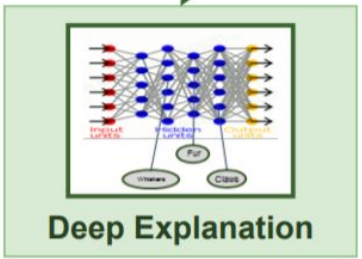
Criar um conjunto de técnicas de aprendizado de máquina que produza modelos **mais explicáveis**, mas que mantenham um alto nível de desempenho de aprendizado.

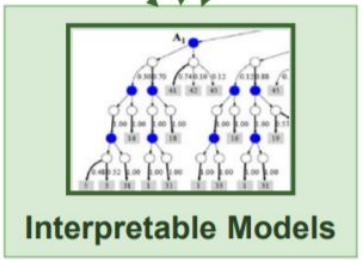
Técnicas de aprendizado (hoje)



Explicabilidade (qualitativamente)

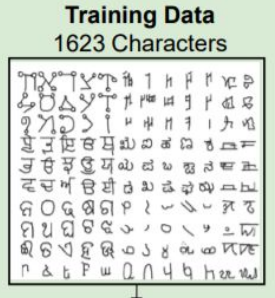






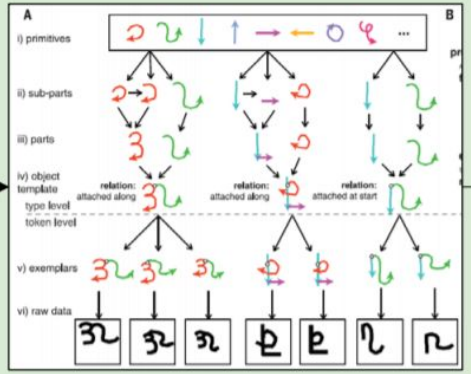
Aprender modelos mais interpretáveis

Aprendizado de conceito (Concept Learning) por meio de indução probabilística



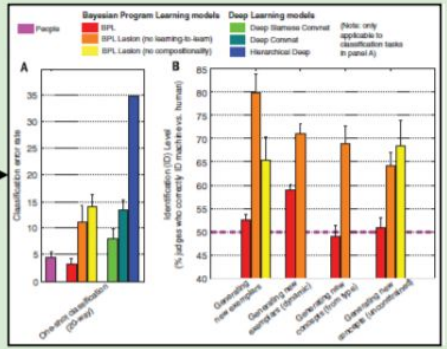
Seed Model

Seed: Um programa probabilístico simples que descreve os parâmetros de geração de caracteres.



Generative Model

Generative: Reconhece caracteres gerando uma explicação de como um novo caractere de teste pode ser criado (ou seja, a sequência mais provável de traços que criaria esse caractere)

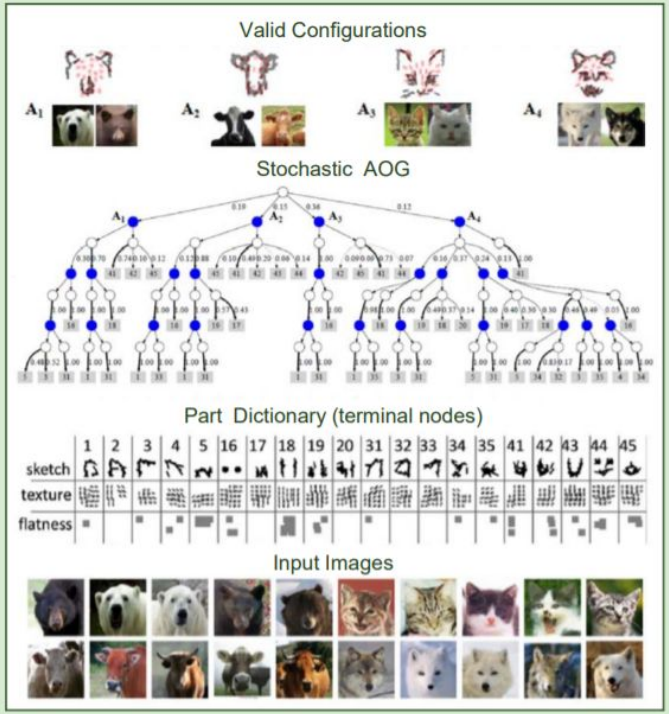


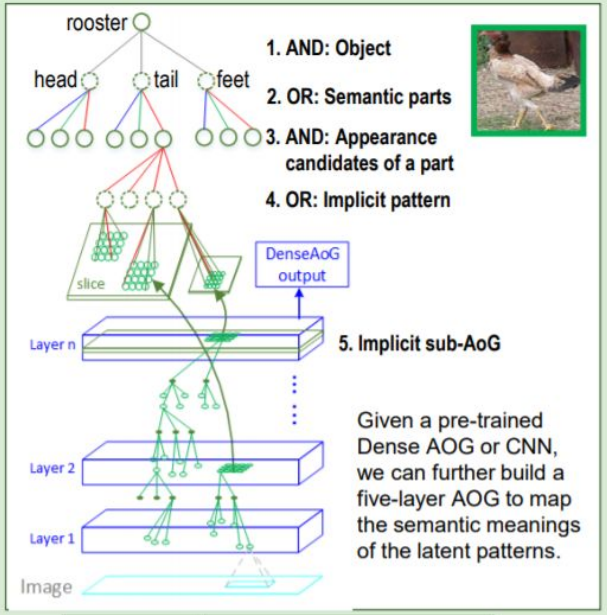
Performance

Resultado: Modelo alcança o desempenho humano e supera o por aprendizado profundo

Aprender modelos mais interpretáveis

AND-OR-Graphs (AOG)





$$L(\theta) = \frac{1}{M}\sum_{m=1}^{M} \underbrace{\log P(I_m, \theta)}_{\text{generative}} + \underbrace{L(pg_m^*, \hat{pg}_m)}_{\text{discriminative}}$$

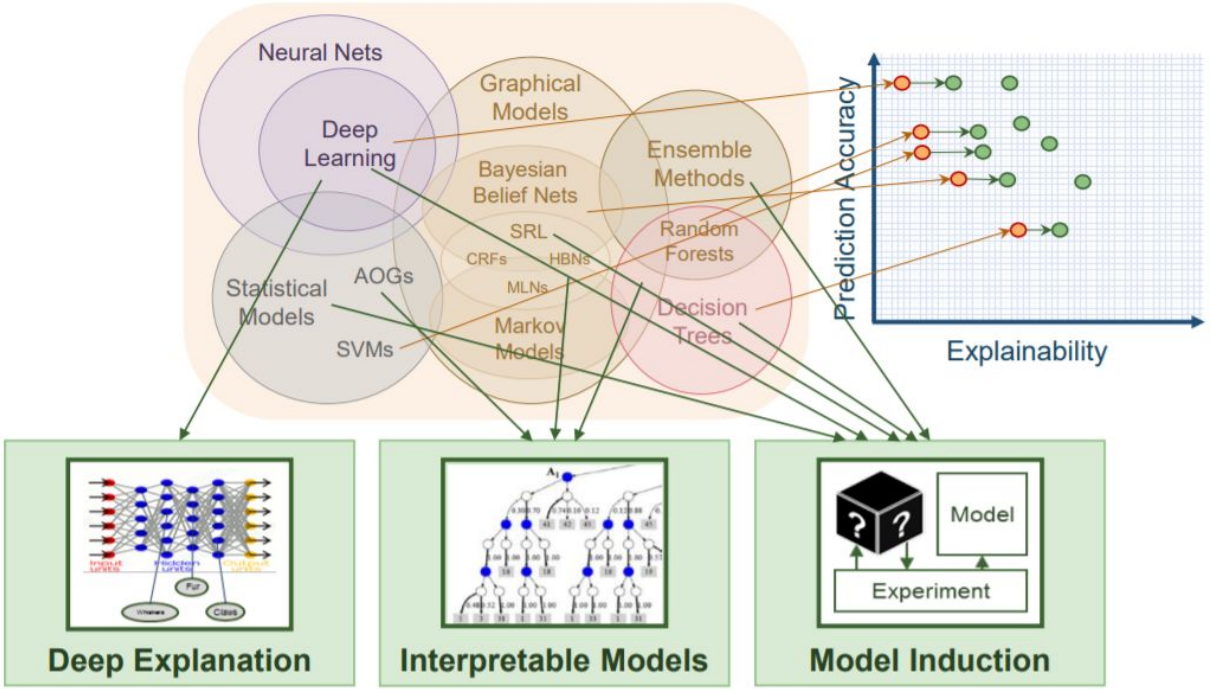
Desempenho x Explicabilidade

Desafio

Criar um conjunto de técnicas de aprendizado de máquina que produza modelos **mais explicáveis**, mas que mantenha um alto nível de desempenho de aprendizado.

Técnicas de aprendizado (hoje)

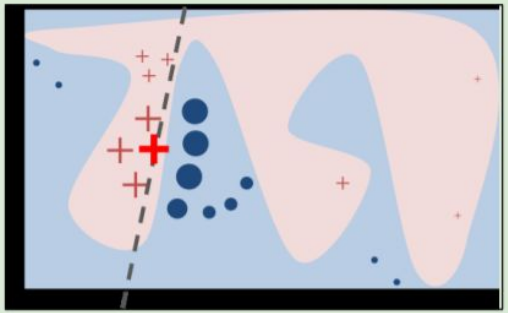
Explicabilidade (qualitativamente)



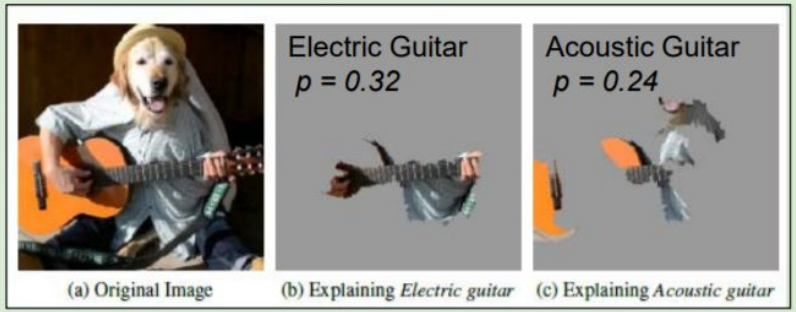
Indução de modelo

Algoritmo LIME (Local Interpretable Model-agnostic Explanations)

Black-box Induction



Example Explanation



- Fundo azul/rosa: A função F do modelo caixa-preta
- A cruz vermelha em negrito é o caso que está sendo explicado.
- O LIME coleta amostras de instâncias, obtém previsões usando F, e as pondera pela proximidade da instância que está sendo explicada.
- A linha tracejada é a explicação aprendida que é localmente fiel (mas não globalmente).

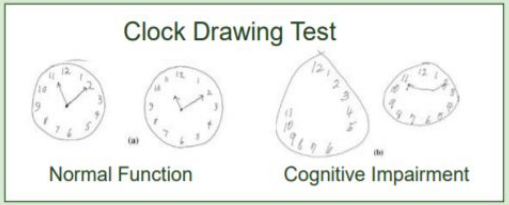
- **LIME** é um algoritmo que pode explicar as previsões de qualquer classificador de forma fiel, aproximando-o localmente com um modelo interpretável.

- **SP-LIME** é um método que seleciona um conjunto de instâncias representativas com explicações como forma de caracterizar todo o modelo.

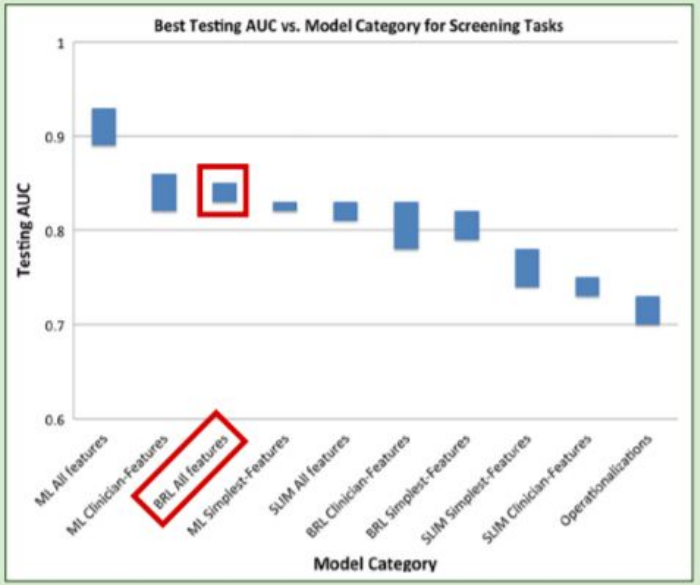
Indução de modelo

Bayesian Rule Lists (BRL)

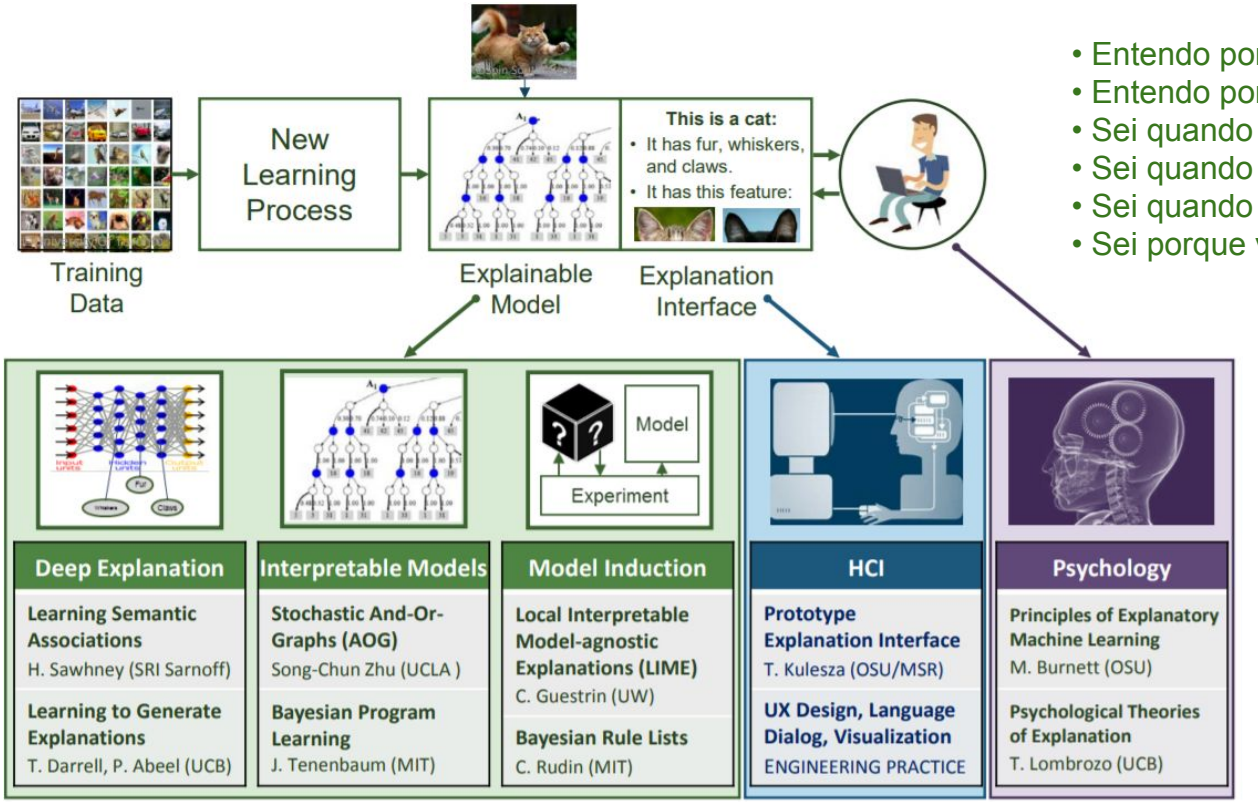
- **if** hemiplegia and age > 60
  - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
  - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
  - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
  - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
  - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
  - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)



• BRLs são listas de decisão - uma série de declarações IF-ELSE;
• BRLs discretizam um espaço de recursos multivariado de alta dimensão em uma série de declarações de decisão simples e prontamente interpretáveis.
• Experimentos mostram que BRLs têm acurácia em paridade com os algoritmos de ML atuais (aproximadamente 85-90% como efetivos), mas com modelos que são muito mais interpretáveis.

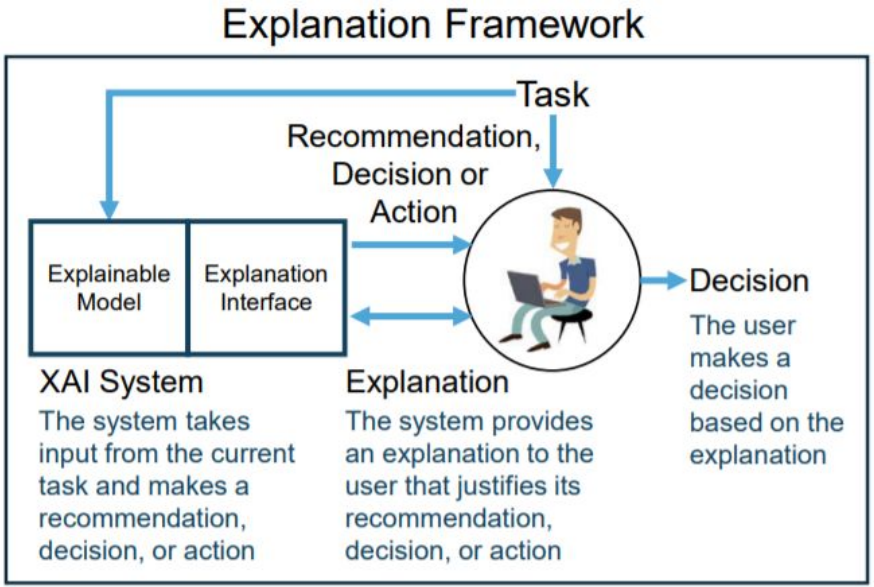


Panorama geral



- Entendo porque
- Entendo porque não
- Sei quando você terá sucesso
- Sei quando você vai falhar
- Sei quando confiar em você
- Sei porque você errou

Como avaliar se a explicabilidade é efetiva?



Satisfação do usuário

- Clareza da explicação (avaliação do usuário)
- Utilidade da explicação (avaliação do usuário)

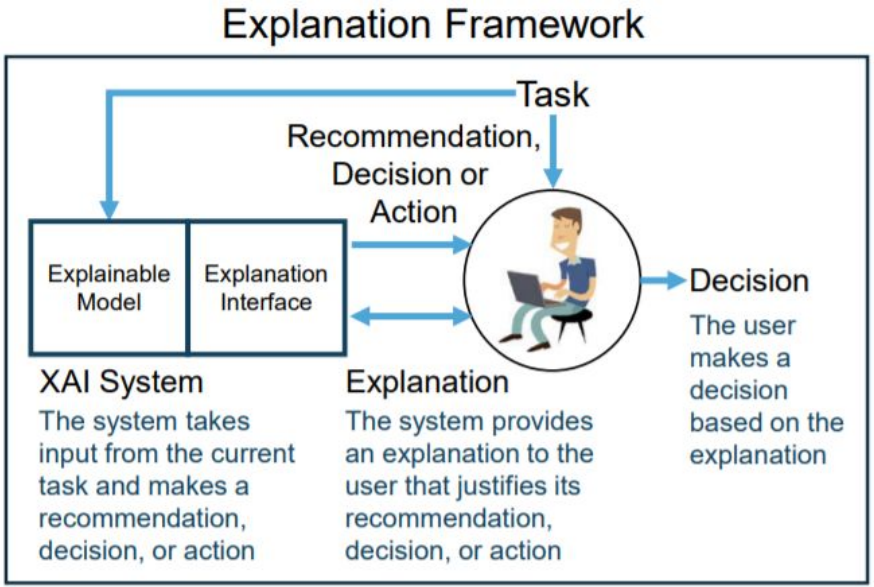
“Modelo mental”

- Compreender as decisões individuais
- Compreender o modelo geral
- Avaliar forças e fraquezas
- Predizer ‘o que vai acontecer’
- Predizer ‘Como posso intervir’

Desempenho da tarefa

- A explicação melhora a decisão do usuário, o desempenho da tarefa?
- Tarefas de decisão artificiais introduzidas para diagnosticar a compreensão do usuário

Como avaliar se a explicabilidade é efetiva?



Avaliação de confiança

- Uso futuro apropriado e confiável

Correção

- Identificação de erros
- Correção de erros
- Treinamento contínuo

Conclusão

- Explainable AI (XAI) não é uma técnica mas uma área guarda-chuva de diversas técnicas e estudos interdisciplinares
  - produzir modelos explicáveis
  - explicar modelos já existentes.
- Para algumas tarefas há abordagens que desempenham melhor que os modelos de Deep Learning
- Produzir modelos explicáveis demanda muito mais esforço que os modelos de aprendizado tradicional
- A área de XAI é muito nova, ainda existe muita pesquisa a ser feita.
  - No entanto, já está demonstrada a sua viabilidade.