Explainable AI (XAI): Core Ideas, Techniques, and Solutions

RUDRESH DWIVEDI, Netaji Subhas University of Technology (formerly NSIT)
DEVAM DAVE, HET NAIK, and SMITI SINGHAL, Pandit Deendayal Petroleum University
RANA OMER, Cardiff University
PANKESH PATEL, University of South Carolina
BIN QIAN, ZHENYU WEN, TEJAL SHAH, GRAHAM MORGAN, and RAJIV RANJAN,
Newcastle University

Importância e Necessidade da IA Explicável (XAI)

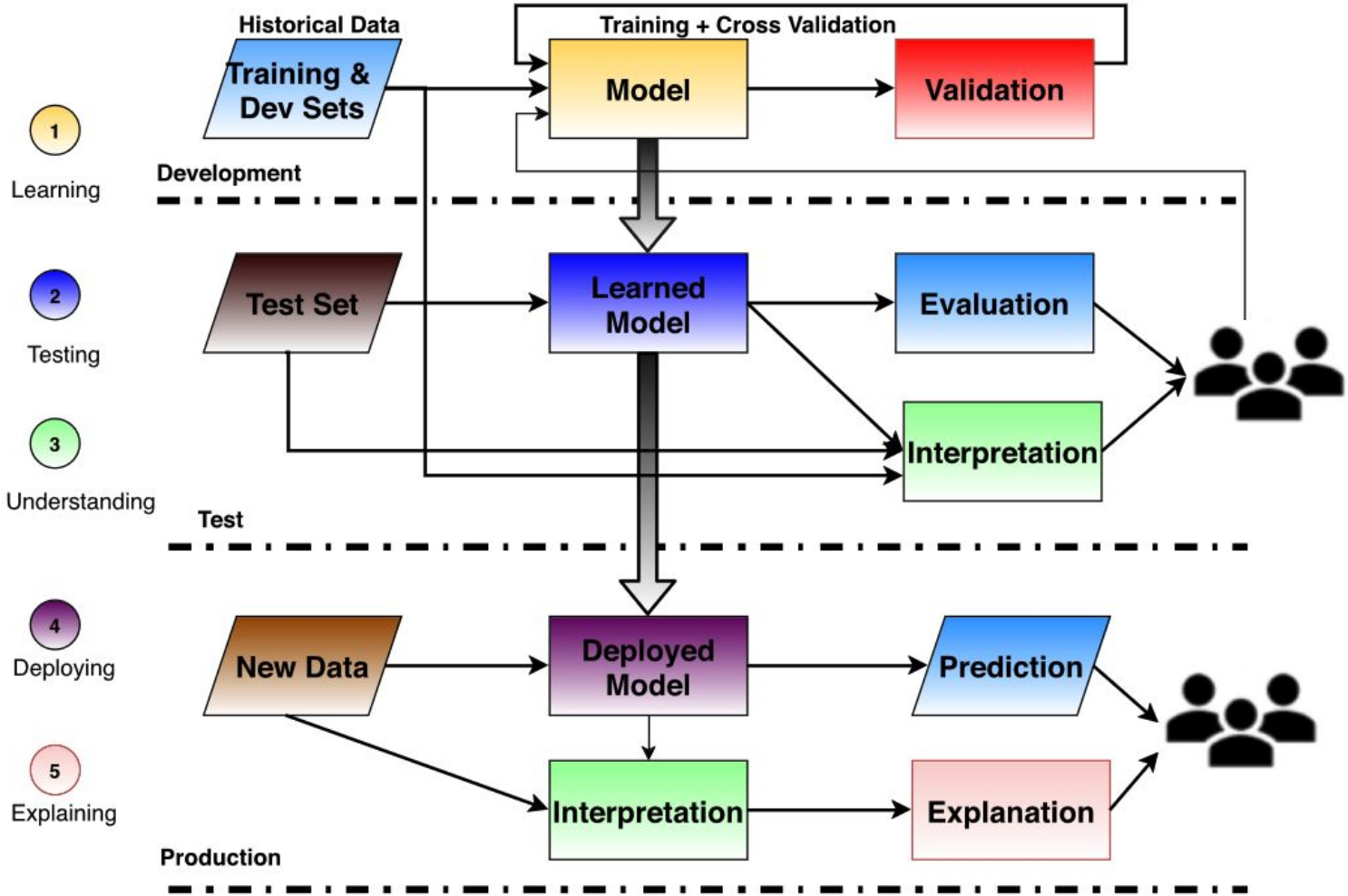
- Uso crescente de sistemas baseados em IA leva à necessidade de um maior **entendimento** sobre o funcionamento e resultados dos mesmos.
- Principalmente em setores críticos, como saúde, finanças e militares, é crucial entender com os sistemas tomam decisões para evitar consequências de possíveis erros.
- XAI visa tornar os processos e resultados de sistemas baseados em IA **transparentes** e **compreensíveis** para humanos.
- XAI se estende desde o início do processo, com a análise de dados, até os números finais oriundos de uma rede neural, por exemplo.
- Com isso em mente, esse *survey* traz um comparativo entre abordagens de XAI para o desenvolvimento de aplicações e ferramentas explicáveis.

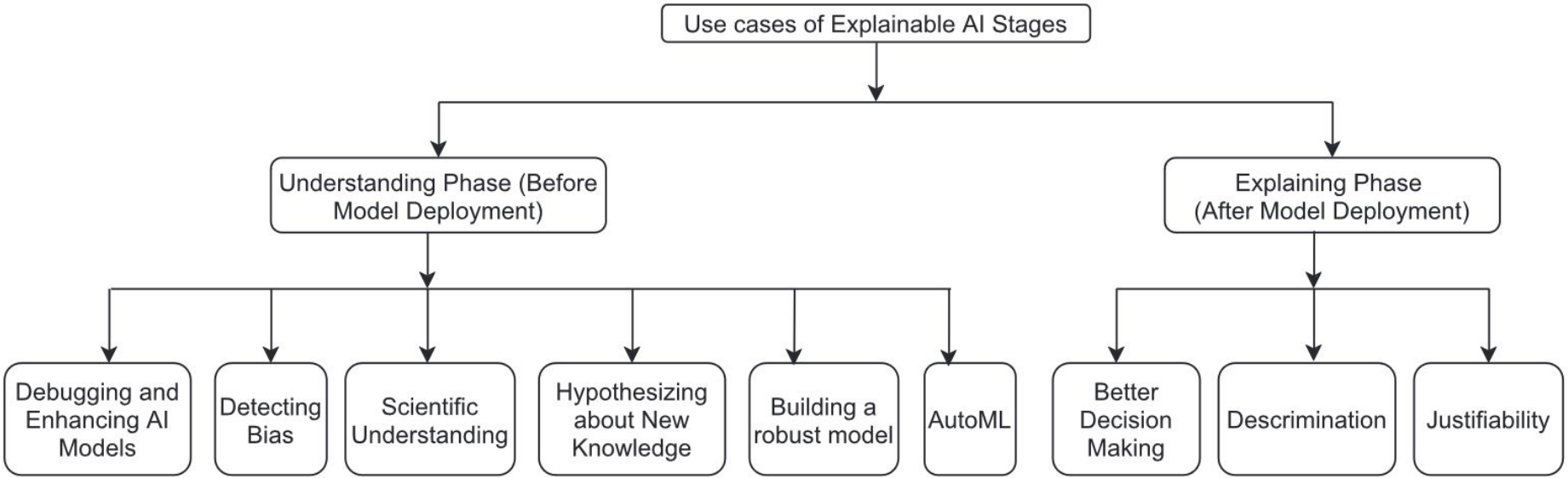
Desenvolvimento de Aplicações XAI

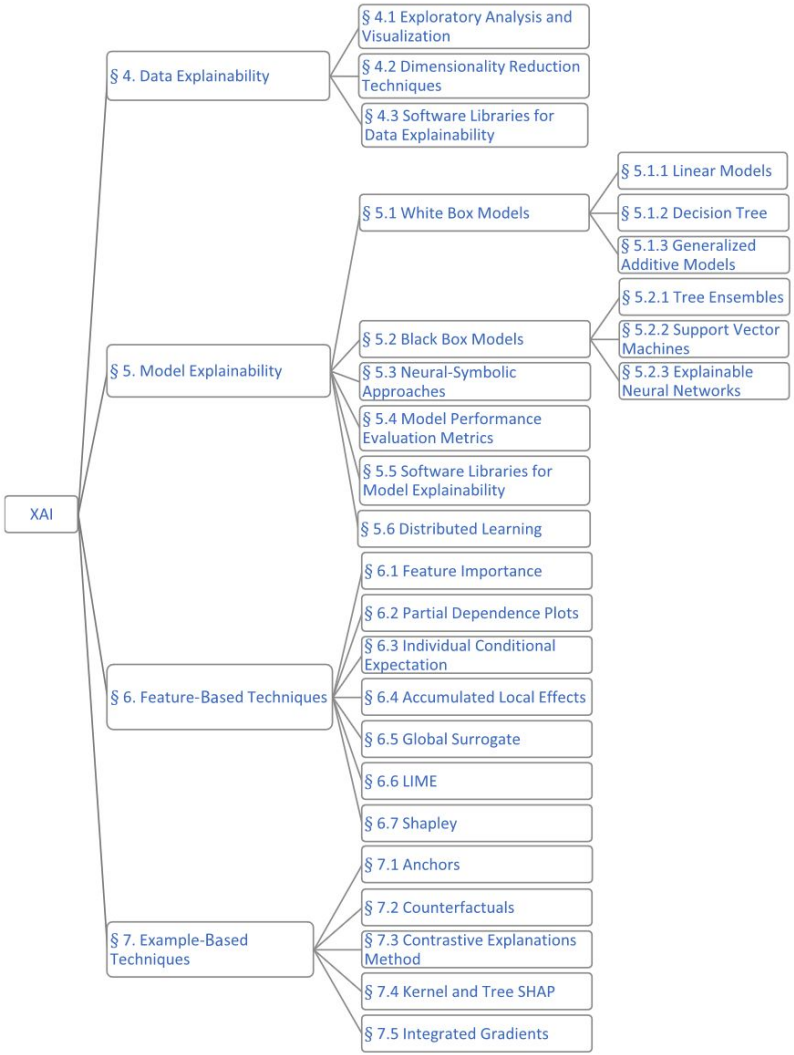
- Adição de duas novas etapas ao ciclo de vida de desenvolvimento de AI.
- **Fase de entendimento:** tem o objetivo de aprimorar o modelo durante a fase de treinamento.
  - Interpretação das variáveis e como elas interagem entre si.
  - Interpretação dos padrões que foram aprendidos pelos modelos.
  - Detecção e análise de vieses nos dados, evitando que sejam propagados ao modelo.
  - *Debug* e melhoria dos modelos de IA.
  - Entendimento científico.
  - Construção de um modelo robusto.
  - AutoML.
- Partes interessadas: desenvolvedores, acadêmicos, cientistas de dados.

Desenvolvimento de Aplicações XAI

- **Fase de explicação:** nessa fase, são interpretadas as predições realizadas após o modelo ser colocado no "mundo real" com dados do dia-a-dia. Com isso, os usuários podem receber explicações mais "legíveis" sobre os resultados dos modelos.
  - Melhora a tomada de decisões.
  - Explicação por trás de modelos com discriminação.
  - Justificação dos resultados dos modelos.
- Partes interessadas: usuários, consumidores, empresas, órgãos reguladores.


Historical Data
Training & Dev Sets
Training + Cross Validation
Model
Validation
1
Learning
Development
Test Set
Learned Model
Evaluation
2
Testing
3
Understanding
Interpretation
Test
4
Deploying
New Data
Deployed Model
Prediction
5
Explaining
Interpretation
Explanation
Production


Use cases of Explainable AI Stages
Understanding Phase (Before Model Deployment)
Explaining Phase (After Model Deployment)
Debugging and Enhancing AI Models
Detecting Bias
Scientific Understanding
Hypothesizing about New Knowledge
Building a robust model
AutoML
Better Decision Making
Descrimination
Justifiability


XAI
§ 4. Data Explainability
§ 4.1 Exploratory Analysis and Visualization
§ 4.2 Dimensionality Reduction Techniques
§ 4.3 Software Libraries for Data Explainability
§ 5. Model Explainability
§ 5.1 White Box Models
§ 5.1.1 Linear Models
§ 5.1.2 Decision Tree
§ 5.1.3 Generalized Additive Models
§ 5.2 Black Box Models
§ 5.2.1 Tree Ensembles
§ 5.2.2 Support Vector Machines
§ 5.2.3 Explainable Neural Networks
§ 5.3 Neural-Symbolic Approaches
§ 5.4 Model Performance Evaluation Metrics
§ 5.5 Software Libraries for Model Explainability
§ 5.6 Distributed Learning
§ 6. Feature-Based Techniques
§ 6.1 Feature Importance
§ 6.2 Partial Dependence Plots
§ 6.3 Individual Conditional Expectation
§ 6.4 Accumulated Local Effects
§ 6.5 Global Surrogate
§ 6.6 LIME
§ 6.7 Shapley
§ 7. Example-Based Techniques
§ 7.1 Anchors
§ 7.2 Counterfactuals
§ 7.3 Contrastive Explanations Method
§ 7.4 Kernel and Tree SHAP
§ 7.5 Integrated Gradients

Classification	XAI Techniques	Global	Local	Model Specific	Model Agnostic	White Box	Black Box
Data explainability	Commonly used data visualization plots	✓	✗	✗	✓	N.A.	N.A.
	Dimensionality reduction techniques	✓	✗	✗	✓	N.A.	N.A.
White box models	Linear model (Section 5.1)	✓	✗	✗	✓	✓	✗
	Decision tree (Section 5.1)	✓	✗	✗	✓	✓	✗
	Generalized additive models (GAMs) (Section 5.1)	✓	✗	✗	✓	✓	✗
	Tree ensembles (Section 5.1)	✓	✗	✗	✓	✓	✗
Artificial neural networks	Neural networks (Section 5.2)	✓	✗	✗	✓	✗	✓
	Neural-symbolic (Section 5.3)	✓	✓	✓	✓	✓	✗
Evaluation metrics	Model evaluation metrics (Section 5.4)	✓	✗	✗	✓	✓	✗
Feature-based XAI techniques	Feature importance (Section 6.1)	✓	✗	✗	✓	✗	✓
	Partial dependence plots (Section 6.2)	✓	✗	✗	✓	✗	✓
	Individual conditional expectation (Section 6.3)	✓	✗	✗	✓	✗	✓
	Accumulated local effects (ALE) (Section 6.4)	✓	✗	✗	✓	✗	✓
	Global surrogate (Section 6.5)	✓	✗	✗	✓	✗	✓
	Local interpretable model-agnostic explanations (LIME) (Section 6.6)	✗	✓	✗	✓	✗	✓
	Shapley value (Section 6.7)	✓	✓	✗	✓	✗	✓
Example-based XAI techniques	Counterfactuals (Section 7.2)	✗	✓	✗	✓	✗	✓
	Anchors (Section 7.1)	✗	✓	✗	✓	✗	✓
	Contrastive explanations method (Section 7.3)	✗	✓	✗	✓	✗	✓
	Prototype counterfactuals (Section 7.2)	✗	✓	✗	✓	✗	✓
	Integrated gradients (Section 7.5)	✗	✓	✗	✓	✓	✗
	Kernel SHAP (Section 7.4)	✓	✓	✗	✓	✗	✓
	Tree SHAP (Section 7.4)	✓	✓	✓	✗	✓	✗

Técnicas de XAI: Caixa Branca vs. Caixa Preta

Modelos de **caixa branca** são inerentemente mais transparentes devido à sua estrutura e lógica serem diretamente acessíveis e compreensíveis, como em **modelos lineares ou árvores de decisão**, onde as relações entre as entradas e saídas são explicitamente definidas e fáceis de seguir.

Modelos de **caixa preta**, como **redes neurais profundas**, são caracterizados por sua complexidade e falta de transparência, pois envolvem múltiplas camadas e conexões não-lineares que dificultam discernir como as entradas são transformadas em saídas. Essa natureza "oculta" dos modelos de caixa preta os torna poderosos em termos de capacidade de modelagem, mas desafiadores para análise e interpretação direta, exigindo técnicas de XAI para desvendar suas decisões internas.

Técnicas XAI Específicas de Modelo

- Projetadas para trabalhar com tipos **específicos** de modelos de aprendizado de máquina
- Aproveitam o conhecimento detalhado sobre sua arquitetura interna e mecanismos de funcionamento.
- Em redes neurais, técnicas como a **análise de camadas ocultas** e **visualização de filtros** são utilizadas para entender como as características de entrada são transformadas em previsões. Estas técnicas exploram diretamente a estrutura e os parâmetros do modelo, como os pesos em redes neurais ou as regras em árvores de decisão.
- Sua **aplicabilidade é limitada** ao tipo de modelo para o qual foram projetadas, o que significa que não podem ser transferidas para modelos de diferentes arquiteturas sem adaptações significativas.

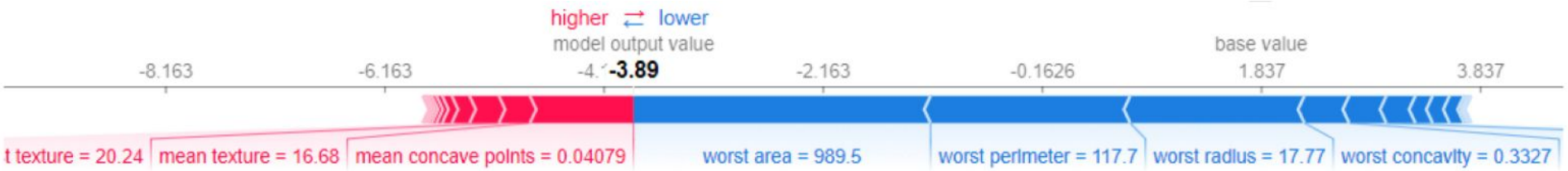
Técnicas XAI Agnósticas de Modelo

- Projetadas para serem aplicadas a **qualquer tipo de modelo** de aprendizado de máquina, independentemente de sua arquitetura interna.
- Tratam o modelo como uma "caixa preta" e se concentram em entender **como as características de entrada estão relacionadas às previsões de saída**.
- Exemplos incluem o LIME e o SHAP, que criam explicações aproximadas para previsões individuais, analisando as mudanças nas saídas do modelo em resposta a variações nas entradas.
- As técnicas agnósticas de modelo são particularmente valiosas devido à sua **flexibilidade** e **aplicabilidade** universal, tornando-as adequadas para um amplo espectro de modelos, desde simples regressões lineares até redes neurais complexas.

SHAP

- SHAP → Global e local, mais custoso computacionalmente
- LIME → Local, menos custoso computacionalmente

- SHAP (SHapley Additive exPlanations) é uma técnica de XIA baseada na teoria dos jogos, utilizada para interpretar modelos de machine learning.
- Atribui a cada *feature* de entrada um valor SHAP, **indicando sua contribuição para a previsão final do modelo**. Garante explicações justas e proporcionais, revelando não só a **importância de cada característica**, mas também como elas **interagem e influenciam o resultado**.
- O SHAP é especialmente útil em modelos complexos, como redes neurais, onde as relações entre características e previsões não são lineares, promovendo uma maior **transparência e confiança** nas decisões baseadas em IA.



Interpretação Global

- **Interpretação global** refere-se ao entendimento do modelo como um **todo**, abordando como o modelo geralmente toma decisões com **base em todas as características de entrada**.
- Este tipo de interpretação é crucial para compreender a lógica geral e o comportamento do modelo em uma ampla variedade de situações.
- Técnicas de interpretação global, como a importância de características agregadas e a análise de modelos simplificados (por exemplo, modelos lineares aproximativos), ajudam a revelar padrões gerais e tendências nas decisões do modelo.
- Isso é especialmente útil para identificar quais características são as mais influentes em geral e como elas interagem para influenciar as previsões do modelo.

Interpretação Local

- **Interpretação local**, em contraste, foca na explicação de **previsões individuais** feitas por um modelo de IA.
- Este tipo de interpretação serve para entender casos específicos, esclarecendo por que o modelo chegou a uma determinada conclusão para uma única instância de entrada.
- Técnicas como LIME e análise de contribuição de características individuais são comumente usadas para este fim.
- A interpretação local é particularmente valiosa em situações onde as decisões individuais têm consequências significativas, como no **diagnóstico médico** ou na **aprovação de crédito**.
- Ela permite aos usuários e especialistas avaliar e questionar a validade de uma decisão específica.

Explicabilidade dos Dados

- Análise Exploratória e Visualização de Dados
- Técnicas de Redução de Dimensionalidade
  - Principal Component Analysis (PCA)
  - Locally linear embedding (LLE)

XAI Techniques	numpy, pandas	matplotlib	seaborn	sklearn	wordcloud	networkX
Data visualization plots	✓	✓	✓	✓	✗	✗
Kernel density estimation (KDE)	✓	✓	✓	✓	✗	✗
Box and whisker plots	✓	✓	✓	✗	✗	
Correlation matrix	✓	✓	✓	✓	✗	✗
WordCloud	✗	✗	✗	✗	✓	✗
Network diagram	✗	✗	✗	✗	✗	✓
Principal component analysis (PCA)	✓	✓	✓	✓	✗	✗
HeatMaps	✓	✓	✓	✓	✗	✗
t-Distributed stochastic neighbor embedding (t-SNE)	✓	✓	✓	✓	✗	✗

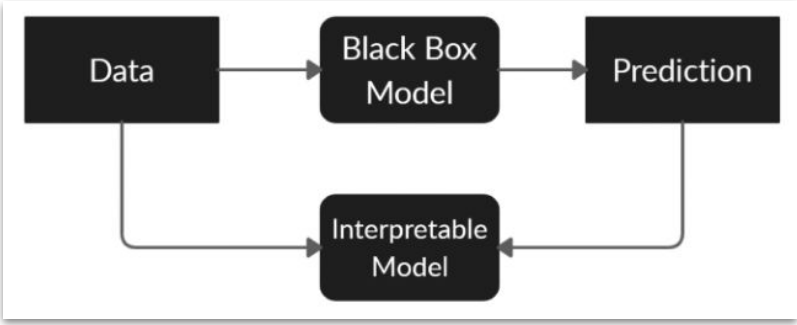
Explicabilidade de Modelos

- **Para modelos caixa branca:**
  - **Modelos Lineares:** visualização dos coeficientes
  - **Árvores de Decisão:** visualização das folhas e importância das features.
  - **Generalized Additive Models (GAMs):** modelos lineares + função de suavização.
- **Para modelos caixa preta:**
  - **Ensembles de Árvores:** converter uma árvore complexa em regras simples.
  - **Uso de Support Vector Machines (SVM)**
  - **Redes Neurais Explicáveis**
- **Abordagens Neuro-Simbólicas**
- **Métricas de Avaliação de Performance de Modelos**

Explicabilidade de Modelos: Bibliotecas

XAI Techniques	Basic Libraries*	TensorFlow	Keras	PyTorch	PyGAM
Linear model coefficients (Section 5.1)	✓	✓	✓	✓	✗
Decision tree (Section 5.1)	✓	✓	✗	✓	✗
Generalized additive models (GAMs) (Section 5.1)	✓	✗	✗	✗	✓
Neural networks (Section 5.2)	✓	✓	✓	✓	✗
Tree ensembles (Section 5.2)	✓	✓	✓	✓	✗
Model performance evaluation metrics (Section 5.4)	✓	✓	✓	✓	✗

Técnicas Baseadas em Features



- Importância das Features
- Gráficos de Dependência Parcial (PDP)
- Expectativa Condicional Individual (ICE)
- Substituto Global
- Interpretações Locais Agnósticas de Modelos: LIME e SHAP

Approach	Advantages	Disadvantages
PDP	• Intuitive • Easy to implement • Shows global effects	• Assumption of independence • Heterogeneous effects may be hidden
ICE	• Intuitive • Easy to implement • Can uncover heterogeneous relationships	• Can only display one feature meaningfully • Assumption of independence • Not easy to see the average
Feature importance	• Provides a highly compressed global insight • Comparable across problems • Automatically takes into account all interactions	• Not additive • Shuffling the feature added randomness • Need access to true data • Assumption of independence

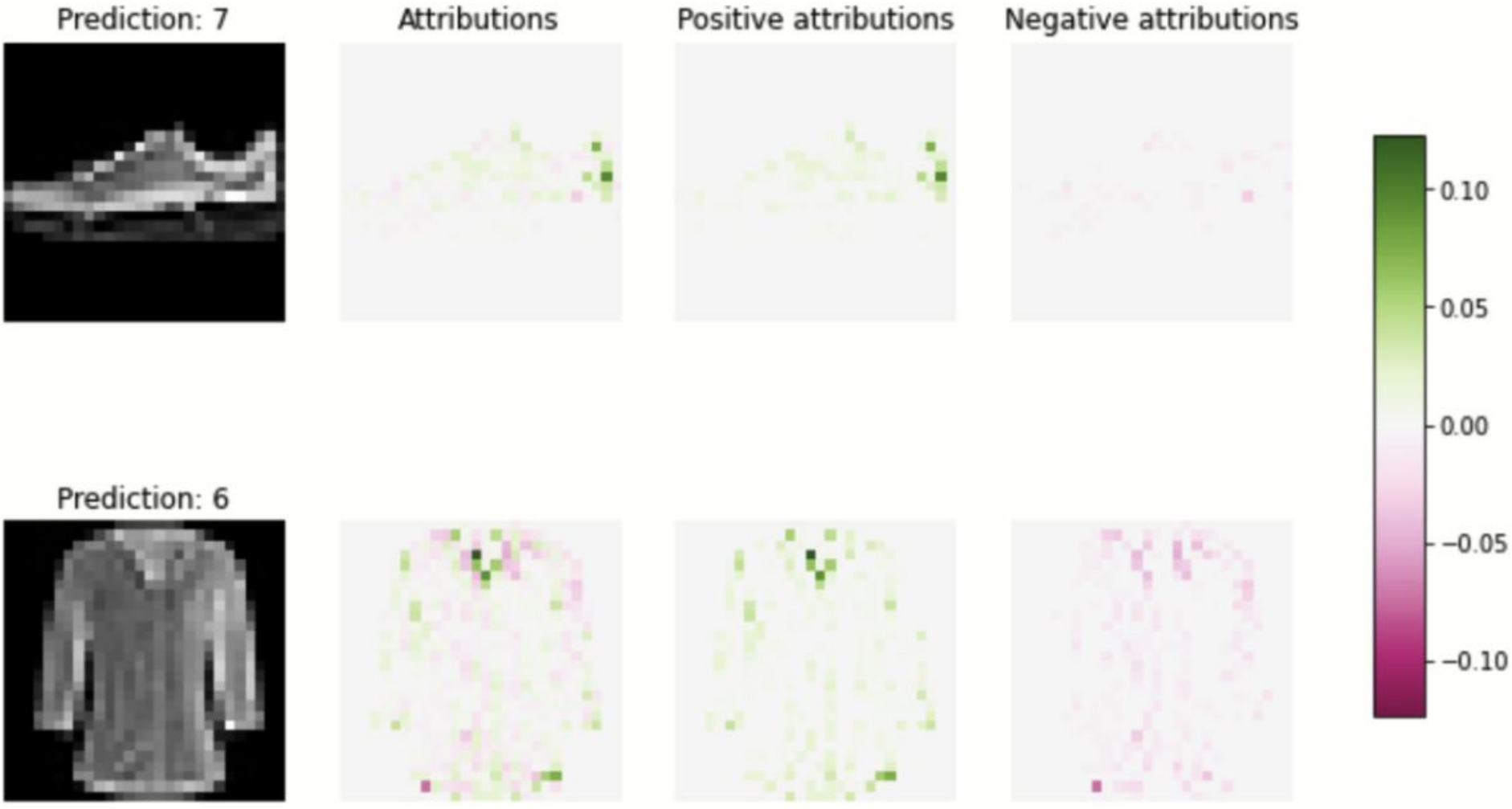
Técnicas Baseadas em Features: Bibliotecas

Feature-Based XAI Techniques	Basic Libraries*	Keras, TensorFlow, PyTorch	Lime	Shap	Skater	eli5	PDPbox	XAI
Feature importance (Section 6.1)	✓	✓	✗	✗	✗	✓	✗	✓
Partial dependence plots (Section 6.2)	✓	✓	✗	✗	✗	✗	✓	✗
Individual conditional expectation (Section 6.3)	✓	✓	✗	✗	✗	✗	✓	✗
Global surrogate (Section 6.5)	✓	✓	✗	✗	✗	✗	✗	✗
LIME (Section 6.6)	✓	✓	✓	✗	✓	✗	✗	✗
Shapely value (Section 6.7)	✓	✓	✗	✓	✓	✗	✗	✗
Accumulated local effects plot (Section 6.4)	✓	✓	✗	✗	✗	✗	✗	✗

Técnicas Baseadas em Exemplos

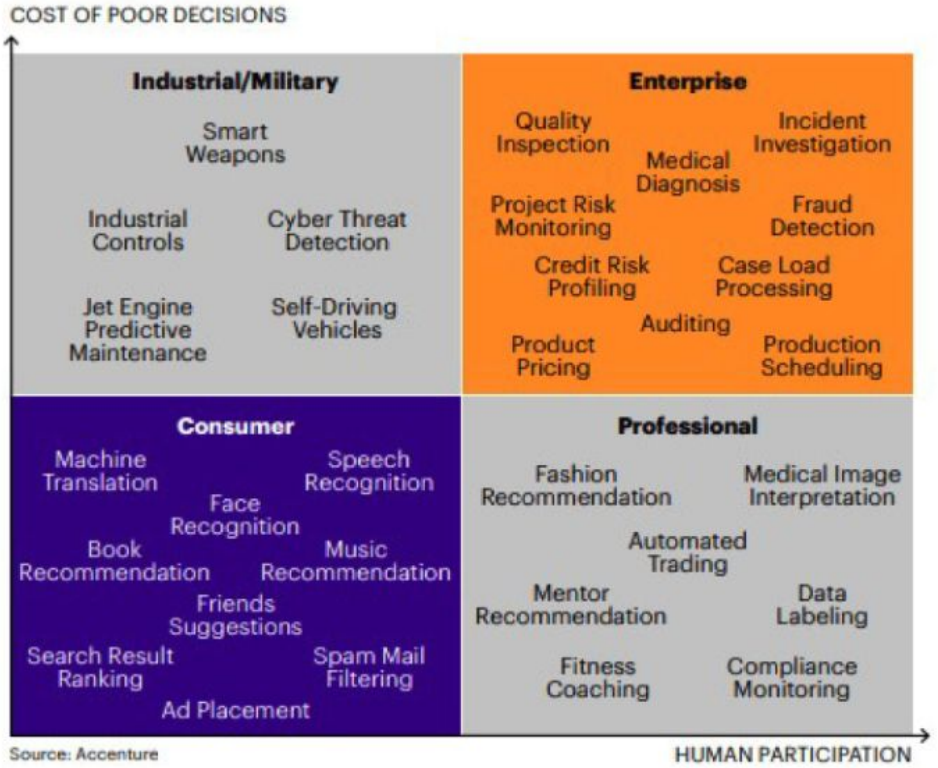
- Âncoras
- Explicações por Contraexemplos
- Métodos de Explicações Contrastantes
- Gradientes Integrados

Example-Based XAI Techniques	Basic Libraries*	Keras, TensorFlow, PyTorch	DiCE	Alibi	Tf-explain
Anchors (Section 7.1)	✓	✓	✗	✓	✗
Counterfactuals (Section 7.2)	✓	✓	✓	✓	✗
Prototype counterfactuals (Section 7.2)	✓	✓	✗	✓	✗
Contrastive explanations method (Section 7.3)	✓	✓	✗	✓	✗
Kernel SHAP (Section 7.4)	✓	✓	✗	✓	✗
Tree SHAP (Section 7.4)	✓	✓	✗	✓	✗
Integrated gradients (Section 7.5)	✓	✓	✗	✓	✓


Prediction: 7
Attributions
Positive attributions
Negative attributions
Prediction: 6
0.10
0.05
0.00
−0.05
−0.10

O Que Levar em Conta na Implementação de XAI

- Trade-off entre XAI e desempenho do modelo.
- Usuários diferentes possuem necessidades diferentes de explicabilidade.
- O *design* do sistema geralmente precisa balancear requisitos conflitantes.
- Qualidade dos dados faz parte da XAI
- Explicabilidade nem sempre é a prioridade no *design* de sistemas de IA.



Conclusão

Precisamos de mais explicabilidade em IA