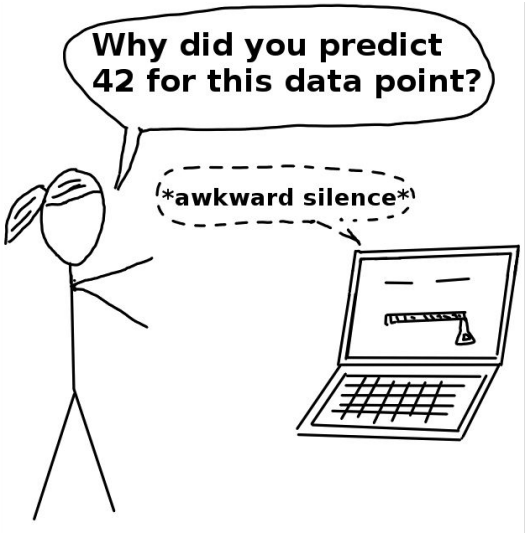
XAI resolvendo problemas da Medicina

REC.ai - Agosto de 2020

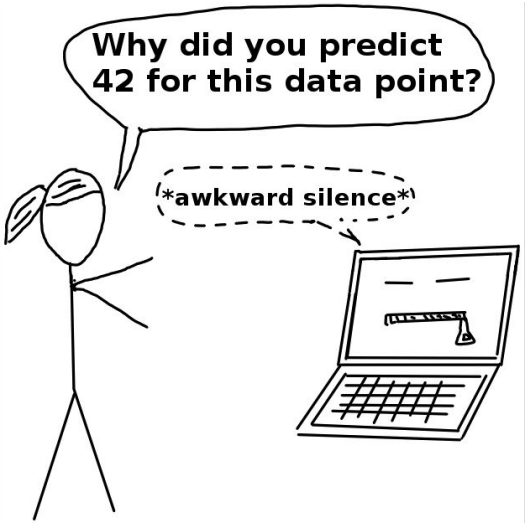
Explainable Artificial Intelligence (XAI)

Métodos que melhoram a **transparência**, **interpretabilidade** e **explicabilidade** de modelos complexos de ML


Why did you predict
42 for this data point?
awkward silence

Para algumas aplicações **99% acurácia não é suficiente**

O modelo é justo?



Pode ser auditado?

Os resultandos são confiáveis?

É possível saber se alguém “mexeu” nele?

O modelo é fácil de entender?

YouTube BR

thais viana meu modelo de aprendizado roubou

FILTRO

Meu modelo de aprendizado roubou pão na casa do João - Thaís Viana

Vinta Software • 143 visualizações • há 2 meses

A PythonXP 2020 é uma conferência online sobre Python e Django voltada para profissionais de tecnologia e estudantes de todo ...

YouTube BR

thais viana meu modelo de aprendizado roubou

FILTRO

Meu modelo de aprendizado roubou pão na casa do João - Thaís Viana

Vinta Software • 143 visualizações • há 2 meses

A PythonXP 2020 é uma conferência online sobre Python e Django voltada para profissionais de tecnologia e estudantes de todo...

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

YouTube BR

thais viana meu modelo de aprendizado roubou

FILTRO

There's

by Ju

The New York Times

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

230

YouTube BR

thais viana meu modelo de aprendizado roubou

NEWS

HURRICANE LAURA DECISION 2020 PLAN YOUR VOTE CORONAVIRUS U.S. NEWS

SOCIAL MEDIA

As algorithms take over, YouTube's recommendations highlight a human problem

A supercomputer playing chess against your mind to get you to keep watching.

By Rebecca Wexler

June 13, 2017

230

YouTube BR

thais viana meu modelo de aprendizado roubou

FILTRO

Meu modelo de aprendizado roubou pão na casa do João - Thaís Viana

Vinta Software • 143 visualizações • há 2 meses

A PythonXP 2020 é uma conferência online sobre Python e Django voltada para profissionais de tecnologia e estudantes de todo ...

As algorithms take over, YouTube's recommendations highlight a human problem

A supercomputer playing chess against your mind to get you to keep watching.

By Rebecca Wexler

June 13, 2017

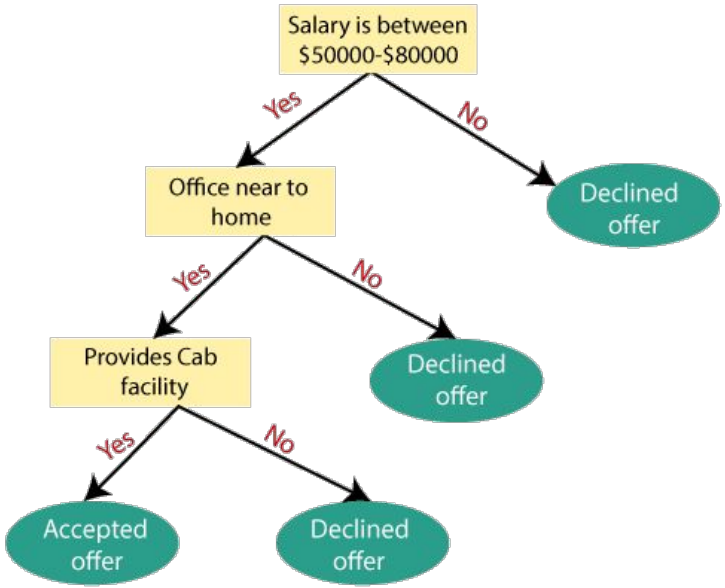
230

Interpretabilidade

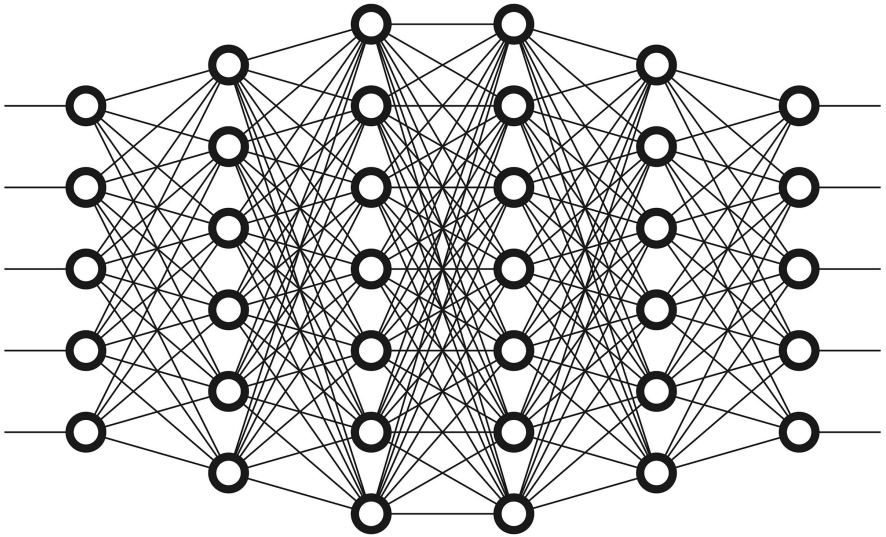
- Interpretabilidade é o grau em que um **ser humano** pode **compreender** a **causa** de uma **decisão**

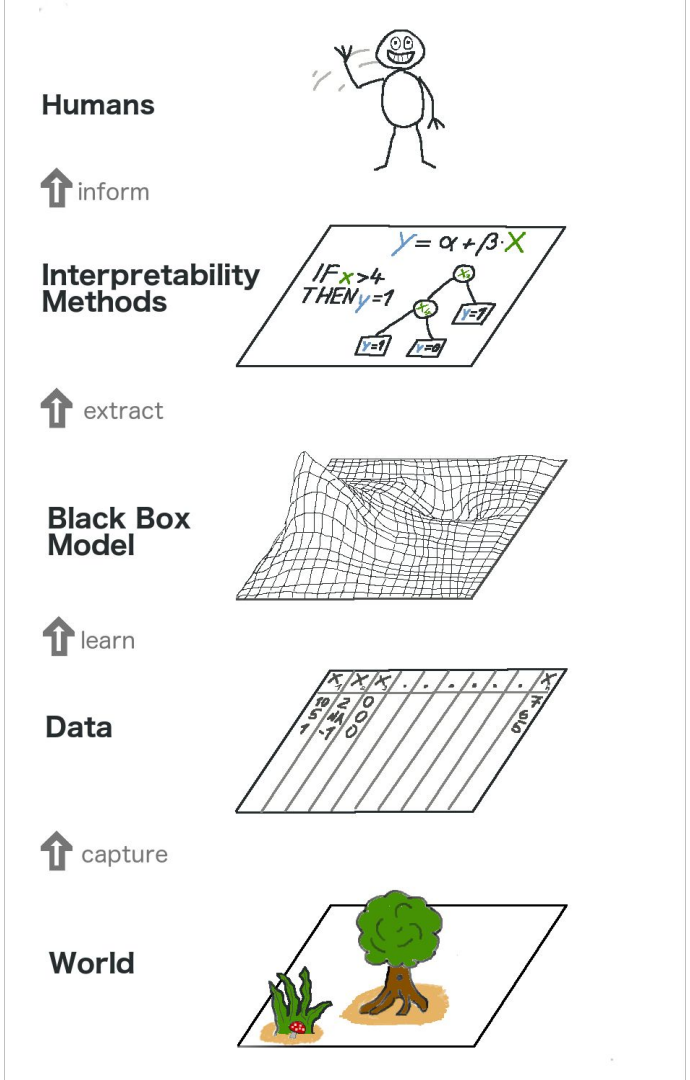
- Nem todo sistema precisa ser interpretável

Interpretável?



Interpretável?





Explicabilidade

- Relaciona os **valores** de uma **instância** à sua **previsão** de modelo
- Deve ser humanamente **compreensível**.

IA na Medicina

- Profissional precisa entender o “**como**” e o “**porque**” de cada decisão
  - Como o modelo é criado?
  - Como o modelo faz as previsões?
  - Como partes do modelo afetam as previsões?
  - Por que o modelo fez uma certa previsão para uma instância?
  - Por que o modelo fez previsões específicas para um grupo de instâncias?
- Razões éticas
- Confiança
- Transparência

XAI na Medicina

- Análise de marcha para classificar doenças neurodegenerativas

XAI na Medicina

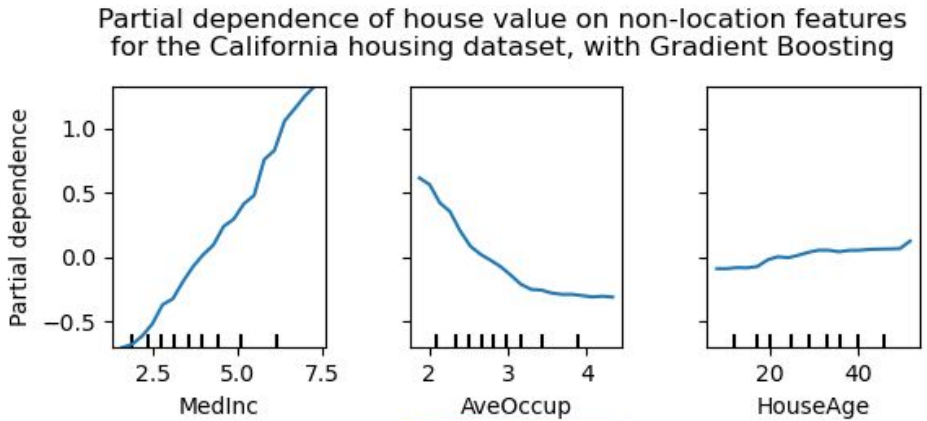
- Análise de marcha para classificar doenças neurodegenerativas

Partial Dependence Plot (PDP)

- Efeito que uma ou duas features têm sobre o resultado de um modelo
- Pode mostrar se a relação entre o resultado e uma feature é linear, monotônico ou complexo.
- Supõem que as features são independentes

Partial Dependence Plot

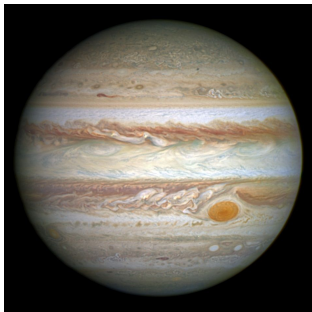
- California housing (sklearn)
- GradientBoostingRegressor



Partial Dependence Plot (PDP)

- Efeito que uma ou duas features têm sobre o resultado de um modelo
- Pode mostrar se a relação entre o resultado e uma feature é linear, monotônico ou complexo.
- Supõem que as features são independentes

- Alternativas: ALE (Accumulated Local Effects) and ICE (Individual Conditional Expectation)

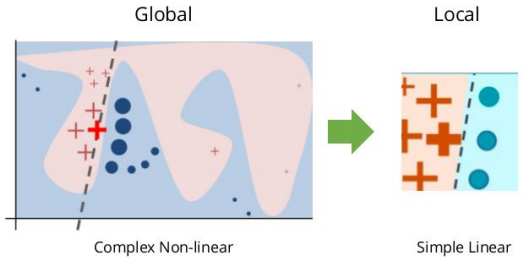


LIME 🇧🇷

- Local Interpretable Model-agnostic Explanations
- Usado para explicar predições individuais
- Um novo modelo é treinado para aproximar as previsões do modelo original

LIME 🇧🇷

- Manipula os dados de entrada (noise) para criar dados artificiais que são usados para explicar o modelo
- A saída é uma lista de explicações, refletindo a contribuição de cada feature para o resultado
- Aproximando o modelo complexo

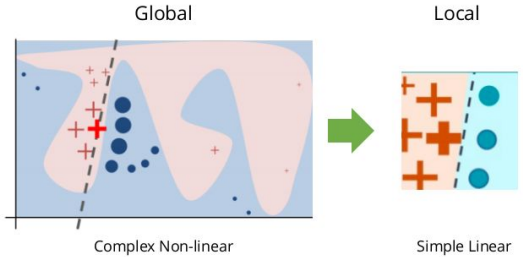


Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

LIME 🇧🇷

- Manipula os dados de entrada (noise) para criar dados artificiais que são usados para explicar o modelo
- A saída é uma lista de explicações, refletindo a contribuição de cada feature para o resultado
- Aproximando o modelo complexo

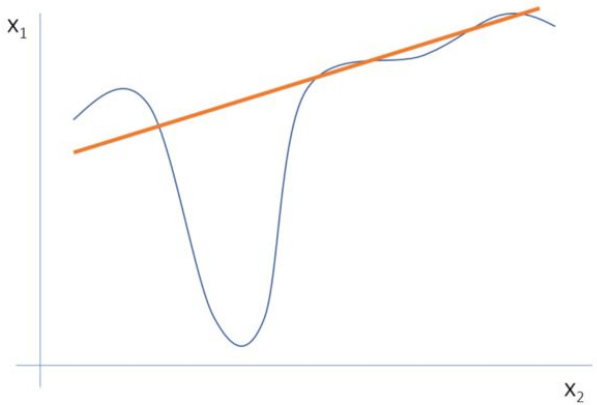


Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

LIME 🇧🇷

- Rápido
- Faz uma aproximação linear olhando para uma região pequena
- Ao expandir essa região o modelo linear pode não funcionar



SHAP

- SHapley Additive exPlanation
- Usados para explicar predições individuais
- Beaseado nos **Valores de Shapley** (Teoria dos Jogos)

SHAP

- **Valores de Shapley**

Um grupo de participantes com habilidades diferentes está cooperando por uma recompensa coletiva. Como dividir a recompensa de forma justa?

- O valor de Shapley é a contribuição marginal média de uma feature em **todas as combinações possíveis**.
- Mantendo alguns axiomas só existe **um único** jeito de dividir os ganhos
- **IA**: Quanto cada feature contribui na saída de um modelo?

SHAP

- **Valores de Shapley**

Um grupo de participantes com habilidades diferentes está cooperando por uma recompensa coletiva. Como dividir a recompensa de forma justa?

- O valor de Shapley é a contribuição marginal média de uma feature em **todas as combinações possíveis**.
- Mantendo alguns axiomas só existe **um único** jeito de dividir os ganhos
- **IA**: Quanto cada feature contribui na saída de um modelo?

SHAP

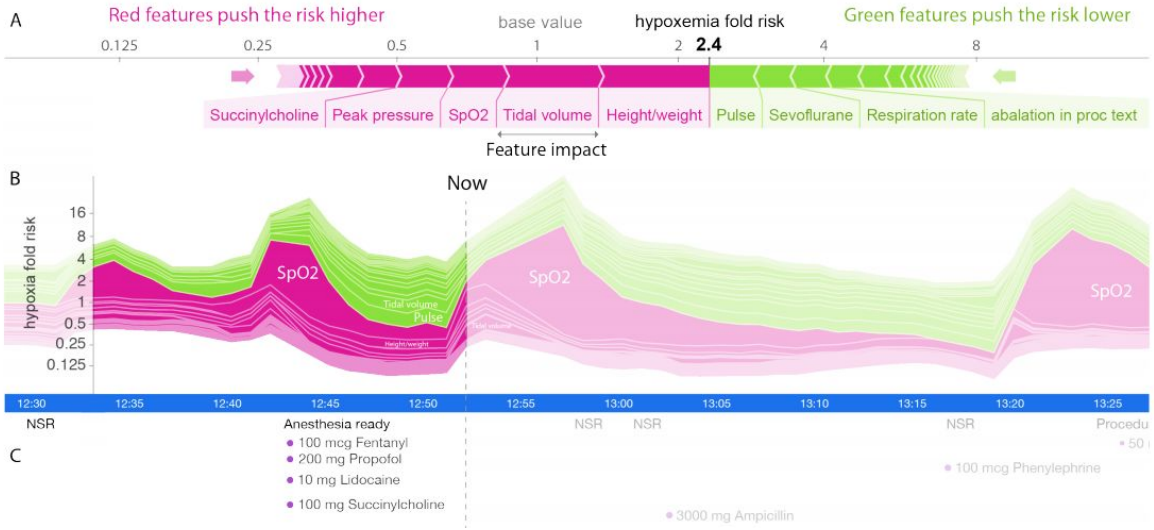
- **Valores de Shapley**

Um grupo de participantes com habilidades diferentes está cooperando por uma recompensa coletiva. Como dividir a recompensa de forma justa?

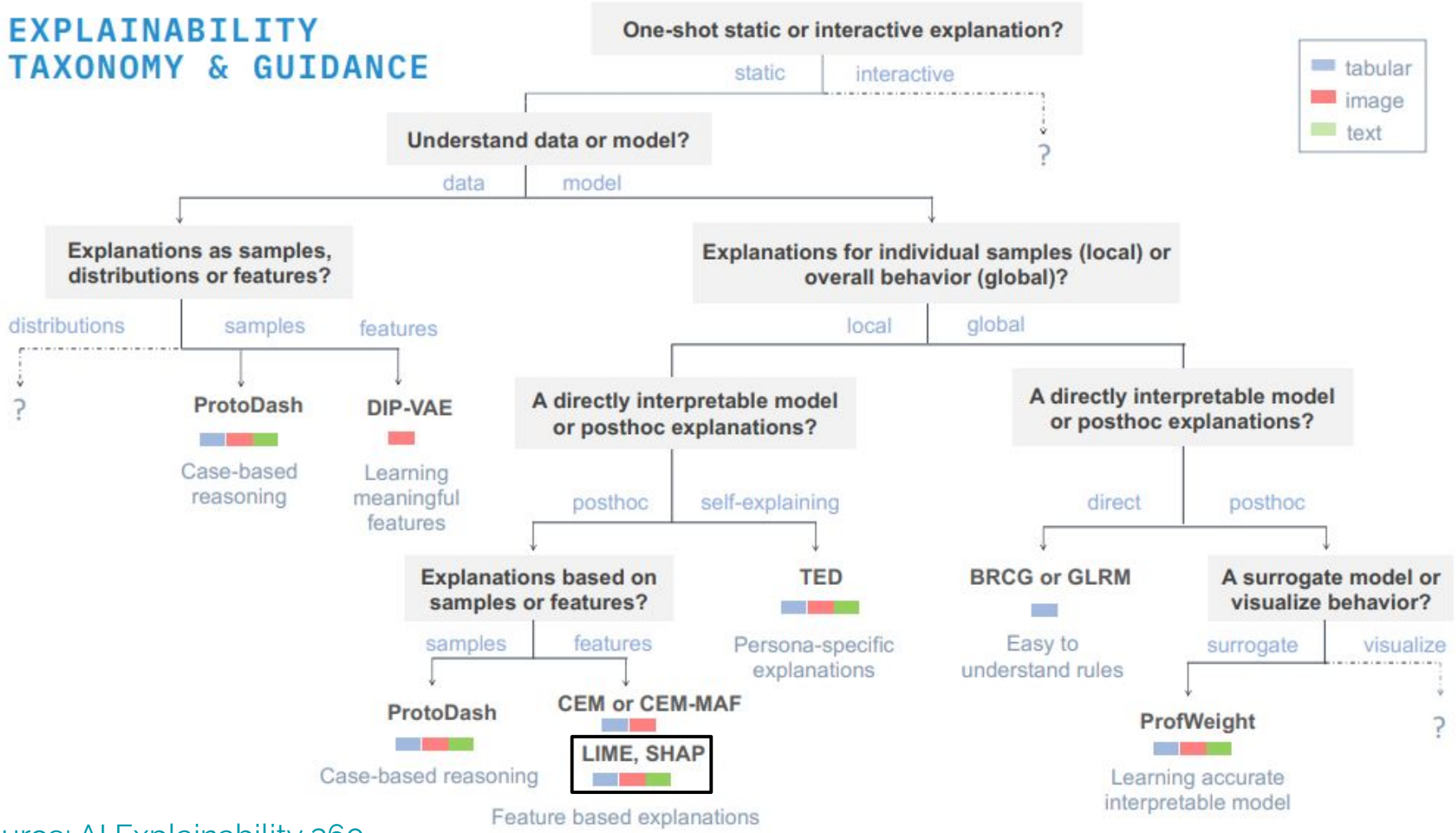
- O valor de Shapley é a contribuição marginal média de uma feature em **todas as combinações possíveis.**
- Mantendo alguns axiomas só existe **um único** jeito de dividir os ganhos
- **IA**: Quanto cada feature contribui na saída de um modelo?
  - https://github.com/slundberg/shap/tree/master/notebooks/general

SHAP

- Explainable machine-learning predictions for the prevention of hypoxaemia during surgery - Nature
- Prever risco de complicações em pacientes anestesiados



EXPLAINABILITY TAXONOMY & GUIDANCE



Source: AI Explainability 360

Conclusão

- Modelos explicáveis são indispensáveis
- Muita ferramentas disponíveis
- A interpretação desses métodos não é intuitivo
- A medida que a quantidade de dados aumenta a análise fica mais difícil

Obrigada!

Perguntas?

Código: https://github.com/rsarai/talks/

@_rebecasarai

Referências

- Palestra: Meu modelo de aprendizado roubou pão na casa do João - Thaís Viana (link)
- Machine Bias (link)
- When a Computer Program Keeps You in Jail (link)
- As algorithms take over, YouTube's recommendations highlight a human problem (link)
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016)
- Livro: Interpretable Machine Learning (link)