

I ERAMIA-RS

MC07: Interpretação de Modelos de Aprendizado de Máquina

Bruno Iochins Grisci e Débora Cristina Santos de Sousa

Universidade Federal do Rio Grande do Sul
Instituto de Informática
Departamento de Informática Teórica

Estes slides utilizam em parte conteúdo adaptado da bibliografia “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable” de Christoph Molnar (Molnar, 2020).

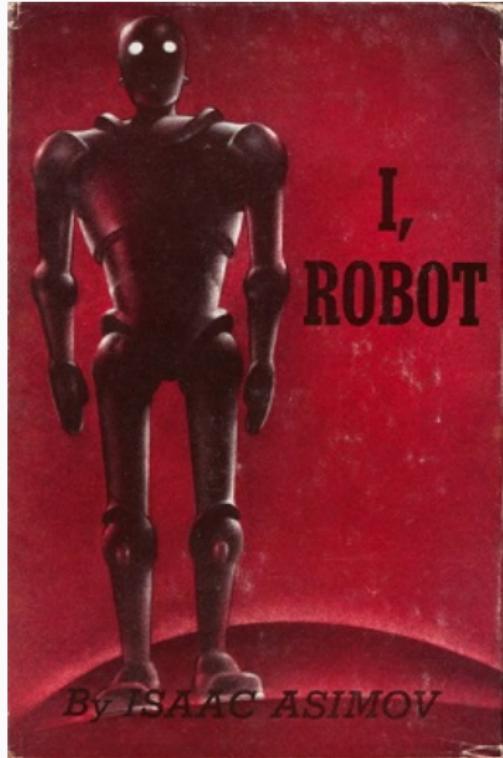
Prólogo

Eu, Robô (2004)



Eu, Robô (1950)

- Coletânea de contos de **Isaac Asimov**.
- Introduz as famosas **Três Leis da Robótica**.
- **Dra. Susan Calvin**: principal robopsicóloga da US Robots and Mechanical Men, Inc., considerada a maior fabricante de robôs do século XXI.



Introdução

Quem somos?

Bruno Lochins Grisci

Professor, Instituto de Informática — UFRGS

- Doutor, Mestre e Bacharel em Ciência da Computação (UFRGS)
- Experiência internacional: Reino Unido, Chile, Alemanha e Canadá
- Áreas de pesquisa:
 - Aprendizado de Máquina e Inteligência Artificial
 - Otimização Evolutiva e Interpretabilidade
 - Bioinformática e Visualização de Dados
- Coordenador do grupo de pesquisa em IA interpretável
- Melhor Tese AB3C, Menção Honrosa CAPES, Heidelberg Laureate Forum



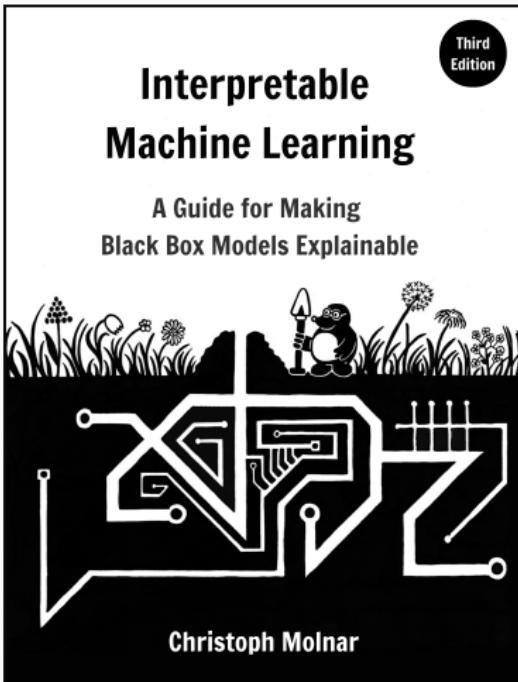
Quem somos?

Débora Cristina Santos de Sousa

- Bacharel em Biotecnologia com ênfase em Bioinformática (UFRGS)
- Mestranda em Ciências da Computação (UFRGS)
- Áreas de pesquisa:
 - Interpretabilidade em Aprendizado de Máquina
 - Biologia Computacional e Biologia Celular
 - Visão Computacional



Material do curso



Christoph Molnar (2020). **Interpretable machine learning.** Lulu. com
<https://christophm.github.io/interpretable-ml-book/>

Material do curso



<https://brunogrisci.github.io/interpretabilidaderamia>

O que é aprendizado de máquina?

Arthur Samuel (1959): "*Área de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados.*"



- Arthur Lee Samuel (December 5, 1901 – July 29, 1990).
- Um dos Pioneiros na área de *Machine Learning*.

O que é aprendizado de máquina?

Aurelien Geron (2017): "*É a ciência (e a arte) de programar computadores de modo que eles possam aprender a partir de dados*"



- Géron Aurélien (2019). "**Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**". In: *Concepts, tools, and techniques to build intelligent systems, 2nd ednn*

O que é aprendizado de máquina?

Aprendizado = Representação + Avaliação + Otimização

- **Representação:** quais *formas* o modelo pode assumir?
- **Avaliação:** como *medimos* a qualidade do modelo?
- **Otimização:** como *buscamos* a melhor função preditiva?

Exemplo: Regressão Linear = Pesos lineares + erro quadrático + descida do gradiente

(Domingos, 2012)

O que é aprendizado de máquina?

Pontos de dados

Guardam **protótipos** (observados ou computados).

Aplicação: *busca pelos vizinhos mais próximos.*

Regras de decisão

Guardam **estruturas se–então.**

Aplicação: *casar novos dados com as regras.*

Tensores de pesos

Guardam **pesos** (matrizes/tensores).

Aplicação: *multiplicar pesos por (meta)features intermediárias.*

Distribuições

Guardam **distribuições probabilísticas.**

Aplicação: *inferência via cálculos de probabilidade.*

(Molnar, 2025)

O que é interpretabilidade?

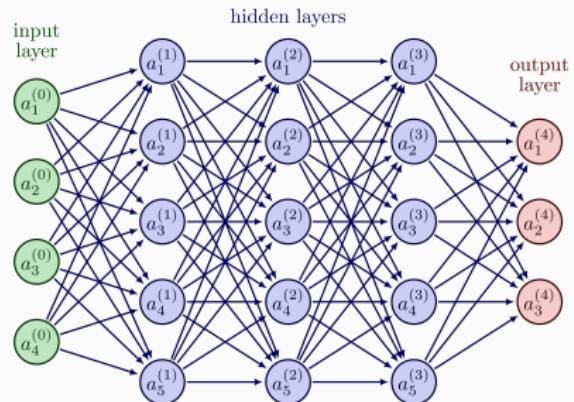
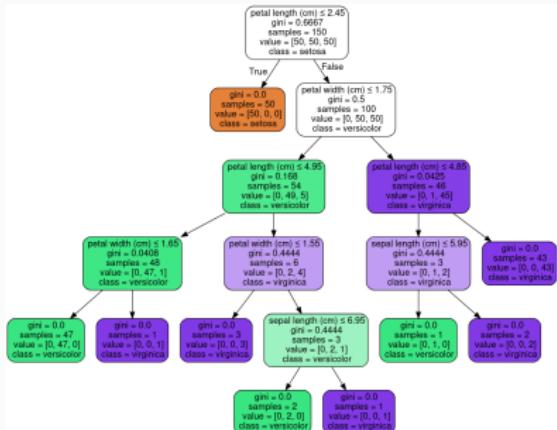
O que é interpretabilidade?

- “Não existe uma definição matemática de interpretabilidade.”
- “Interpretabilidade é o grau em que um ser humano consegue compreender a causa de uma decisão.”
- “Interpretabilidade é o grau em que um ser humano consegue prever de forma consistente o resultado de um modelo.”
- As previsões de modelos de caixa-preta são opacas para o usuário.

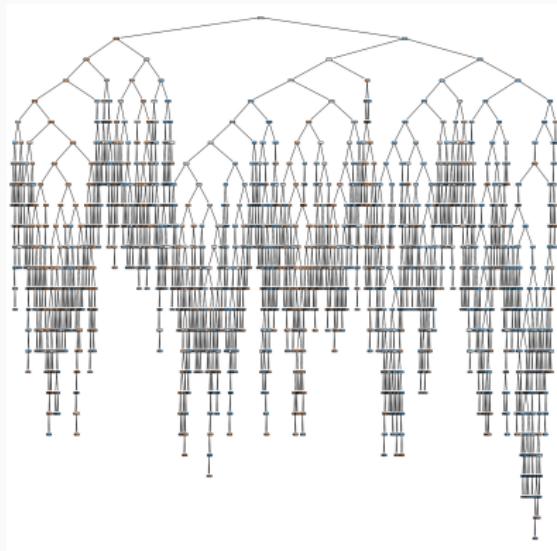
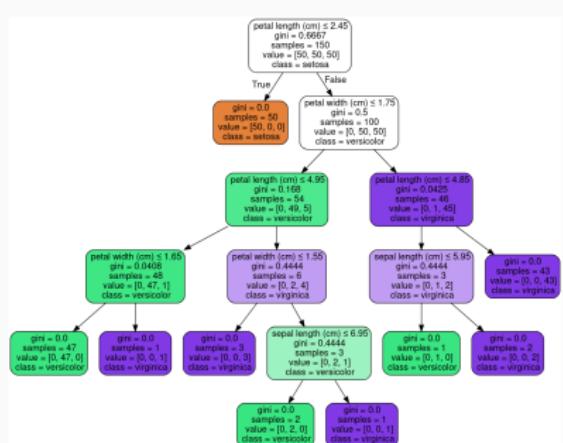
Alguns modelos são mais interpretáveis que outros

- Quanto mais **interpretável** um modelo, mais fácil entender **por que** ele fez uma previsão.
- Um modelo é mais interpretável que outro se suas decisões são mais fáceis de serem compreendidas por humanos.

Alguns modelos são mais interpretáveis que outros



Alguns modelos são mais interpretáveis que outros



Por que interpretabilidade?

- Clever Hans era um cavalo alemão que, no início do século XX, parecia resolver operações aritméticas e tarefas intelectuais.
- Em 1907, o pesquisador Oskar Pfungst demonstrou que o cavalo na verdade respondia a sinais involuntários do treinador humano ao perceber as expressões faciais e postura.
- A descoberta dessa comunicação não-intencional deu nome ao “efeito Clever Hans”, um alerta metodológico relevante para experimentos com humanos e animais.

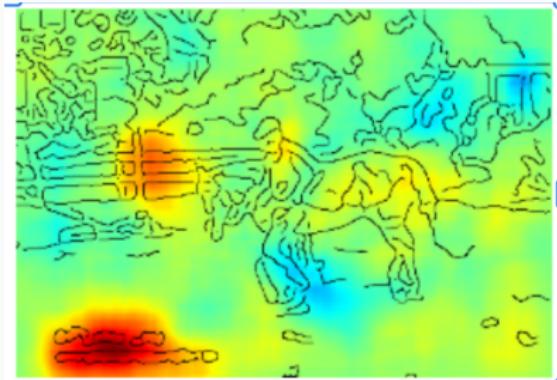


Por que interpretabilidade?

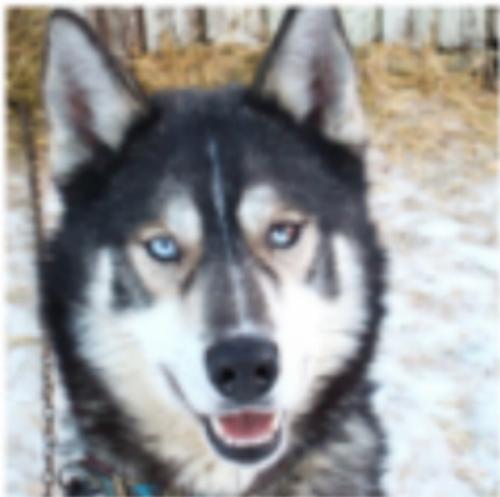


© Lothar Lenz
www.pferdeforarchive.de

Por que interpretabilidade?

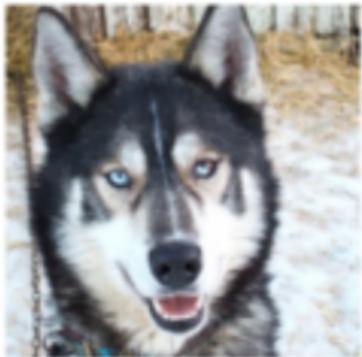


Por que interpretabilidade?



(a) Husky classified as wolf

Por que interpretabilidade?



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Por que interpretabilidade?

- Modelo para detectar baleias aprendeu a usar **artefatos nos arquivos de áudio**, e não o som das baleias em si. (DeLMA and Cukierski, 2013)
- Classificador de imagens passou a usar **textos presentes na imagem** ao invés das características visuais. (Lapuschkin et al., 2019)
- Classificador lobo vs. cachorro baseou-se na **presença de neve ao fundo**, e não no animal. (Ribeiro, Singh, and Guestrin, 2016a)

Em todos esses casos, as falhas não reduziram o desempenho preditivo no conjunto de teste!

Shortcut learning

Atalhos são regras de decisão que funcionam bem em benchmarks padrão, mas falham em cenários mais desafiadores (ex.: dados fora da distribuição) (Geirhos et al., 2020).

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillsides as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecongnisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context



Article: Super Bowl 50

Parvati N战士在比赛中成为第40届橄榄球赛上最年轻的四分卫。他也成为有史以来在超级碗上年龄最大的四分卫，当时39岁。过去的记录是由John Elway保持的，他在38岁时带领Broncos在超级碗XXXIII中获胜，并且目前是Denver的执行副总裁兼足球运营和总经理。四分卫Jeff Dean穿着37号球衣在超级碗XXXIV中上场。

Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Por que interpretabilidade?

Do ponto de vista jurídico, a interpretabilidade é fundamental para empresas que operam em diversos países.

Brasil

O Código de Defesa do Consumidor exige que bancos expliquem decisões tomadas com base em análises de crédito.

Europa

A Lei Geral de Proteção de Dados (GDPR) garante o direito à explicação de decisões tomadas por algoritmos. A falta de interpretabilidade pode inviabilizar o uso de redes neurais em diversas aplicações.

Por que interpretabilidade?

A falta de explicabilidade pode levar ao uso de modelos estatísticos ou lineares mais simples.

Questões:

- Menor poder preditivo
- Complexidade subjacente dos dados
- Não linearidade
- Ausência de aprendizado autônomo de features

(Montavon, Samek, and Müller, 2018)

Por que interpretabilidade?

A interpretabilidade surge como a chave para desvendar essa caixa preta. Ela permite que os usuários compreendam como o modelo chegou à sua conclusão, quais fatores influenciaram a decisão e qual a confiabilidade do resultado. Essa capacidade é essencial para:

Ciência

Validar descobertas, identificar novos conhecimentos e embasar conclusões em dados e análises transparentes.

Negócios

Tomar decisões mais assertivas, reduzir vieses algorítmicos e atender às exigências legais de explicabilidade.

Sociedade

Garantir justiça, equidade e transparência na aplicação de modelos de inteligência artificial.

Por que **não** interpretabilidade?

- Se o modelo não tem impacto significativo.
- Se o problema já foi bem estudado.
- Se há risco de manipulação do sistema.

Interpretabilidade vs. Explicabilidade

Termos (Roscher et al., 2020):

- **Interpretabilidade**: traduz conceitos internos do modelo para uma forma compreensível.
- **Explicabilidade**: exige **interpretabilidade + contexto adicional**.
- O termo **explicação** costuma se referir a métodos **locais** (explicar uma previsão específica).
- Na prática, os termos são vagos → tratamos como um **guarda-chuva** para:

Extrair conhecimento relevante sobre as relações nos dados ou aprendidas pelo modelo. (Murdoch et al., 2019)

O que é uma boa explicação?

Uma explicação é a resposta para uma pergunta do tipo “por quê” (Miller, 2019).

- Por que o tratamento não funcionou no paciente?
- Por que meu empréstimo foi negado?
- Por que o email caiu na caixa de *spam*?

O que é uma boa explicação?

- Explicações são **contrastivas**
- Explicações são **selecionadas**
- Explicações são **sociais**
- Explicações focam no **anormal**
- Explicações são **verdadeiras**
- Boas explicações são **consistentes** com as crenças prévias de quem recebe a explicação
- Boas explicações são **gerais e prováveis**

Áreas relacionadas

- **Explainable AI (XAI)**

- Foco: **entender e explicar** como e por que o modelo toma decisões.
- Métodos: importância de features, contrafactuals, visualizações.

- **AI Safety**

- Foco: **garantir que sistemas de IA sejam seguros**, previsíveis e **alinhados** com objetivos humanos.
- Inclui preocupações com riscos, comportamento emergente e sistemas muito poderosos (ex.: LLMs avançados).

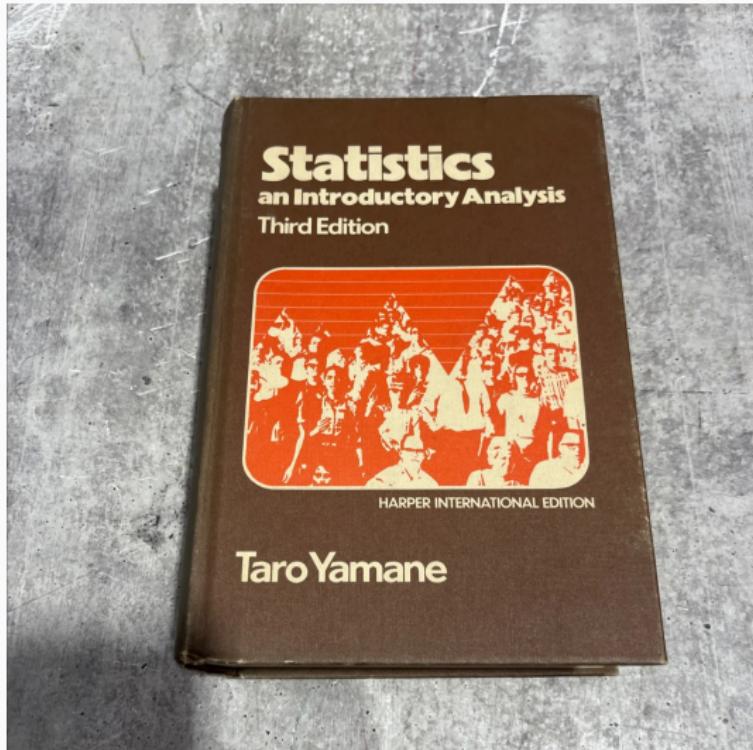
- **AI Fairness**

- Foco: **evitar vieses** e resultados injustos.
- Trabalha com equidade entre grupos, auditorias e métricas de imparcialidade.

Relações: XAI pode ajudar a identificar vieses (Fairness) e comportamentos arriscados (Safety), mas cada área tem **objetivos distintos**.

Começando: analisando os dados

Começando



O que é uma Feature?

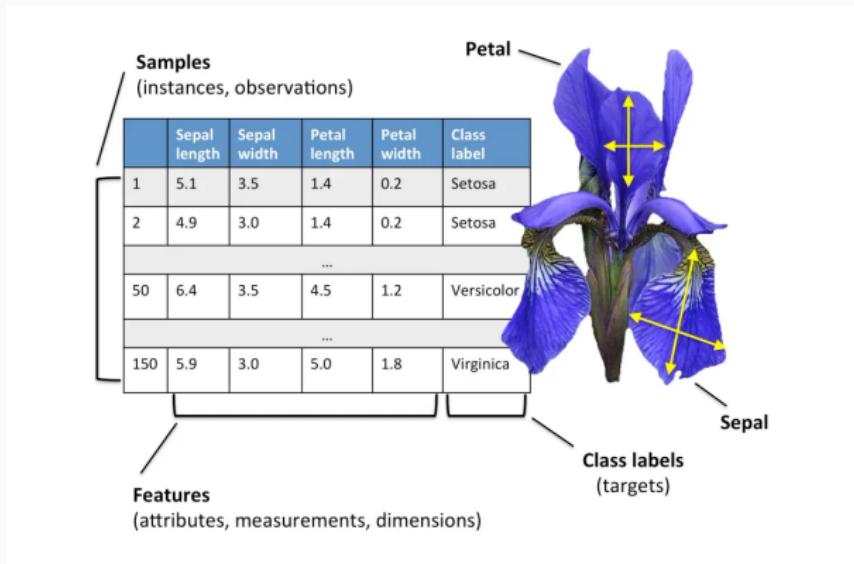
- Uma feature é uma propriedade ou atributo mensurável de uma amostra de dados.
- Em dados tabulares, as features são as colunas (dimensões) do conjunto de dados.
- Cada amostra (linha) é representada como um vetor de features.

Exemplos:

- Níveis de expressão gênica
- Intensidades de pixels em imagens
- Variáveis demográficas em questionários

(Barbieri, Grisci, and Dorn, 2024)

Features



<https://vinlab.medium.com/mastering-machine-learning-with-scikit-learn-an-experiment-with-the-iris-dataset-4c649dc65acf>

O que é Seleção de Features?

- Processo de identificar um subconjunto de features relevantes do conjunto original.
- Objetivo: Melhorar desempenho e interpretabilidade do modelo.
- Mantém o significado original das features (diferente de extração de features).

Ilustração de Seleção de Features

The diagram illustrates the process of feature selection. It starts with a large table containing five columns: Name, Employee ID, No. of year experience, Previous salary, and Salary. Three specific rows are highlighted: Rahul (Employee ID 1), Aman (Employee ID 34), and Ritika (Employee ID 31). Below this table, two arrows point downwards under the 'Employee ID' and 'No. of year experience' columns, both labeled 'Not useful'. This indicates that these two features are being removed. A large blue arrow points from the original table to a smaller table on the right, which contains only three columns: No. of year experience, Previous salary, and Salary. The rows in this smaller table correspond to the highlighted rows in the original table.

Name	Employee ID	No. of year experience	Previous salary	Salary
Rahul	1	2	20000	40000
Aman	34	3	30000	50000
Ritika	31	5	50000	70000

No. of year experience	Previous salary	Salary
2	20000	40000
3	30000	50000
5	50000	70000

<https://www.shiksha.com/online-courses/articles/feature-selection-beginners-tutorial/>

Redução de Dimensionalidade

Duas estratégias principais:

Seleção de Features

Seleciona um subconjunto das features originais.

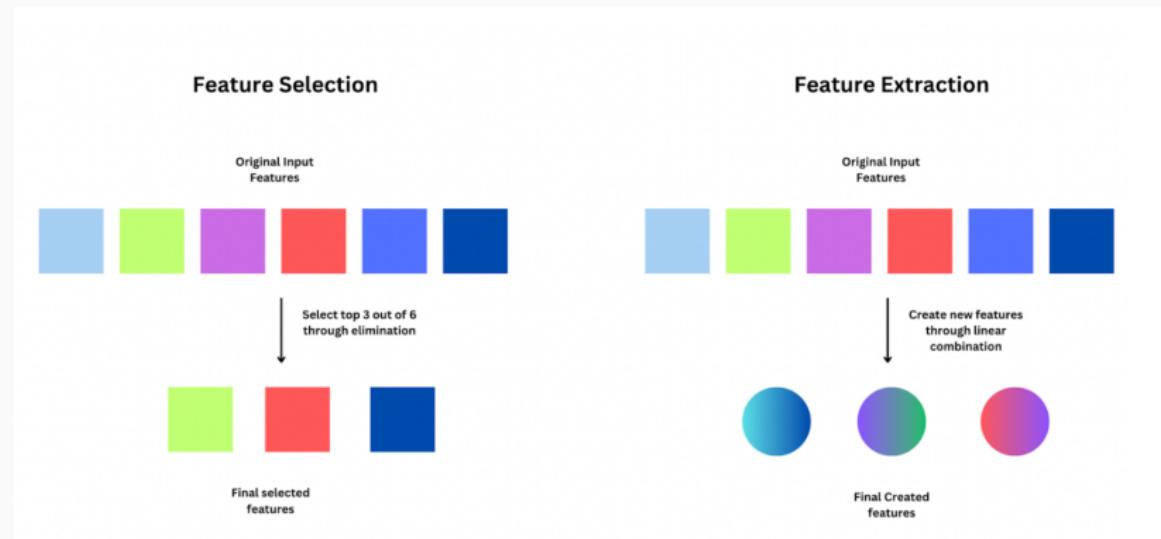
Extração de Features

Transforma as features para um novo espaço (ex.: PCA).

Comparação:

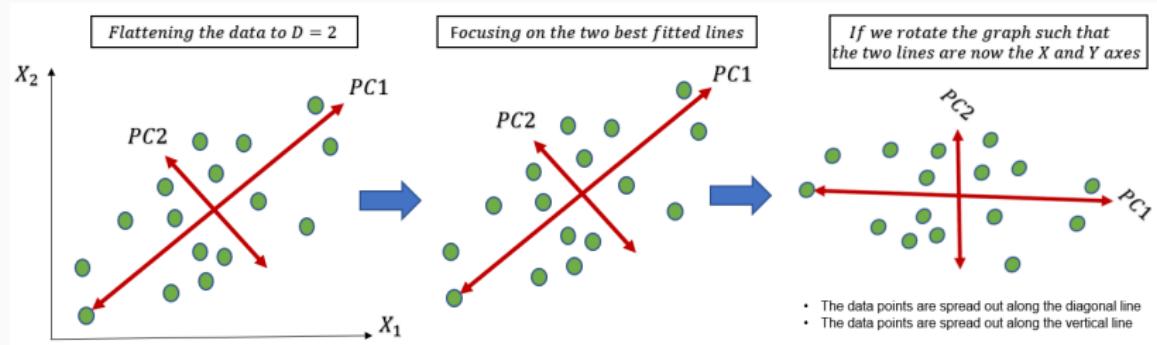
- **Seleção:** Preserva interpretabilidade
- **Extração:** Pode melhorar desempenho, mas reduz interpretabilidade

Seleção vs Extração



<https://viso.ai/deep-learning/feature-extraction-in-python/>

PCA



<https://www.linkedin.com/pulse/gentle-introduction-principal-components-analysis-michael-wynn/>

Por que Seleção de Features?

Motivações principais:

- **Acurácia:** Remove ruído e reduz risco de overfitting
- **Memória:** Reduz tamanho do modelo e do conjunto de dados
- **Interpretabilidade:** Ajuda a identificar variáveis relevantes

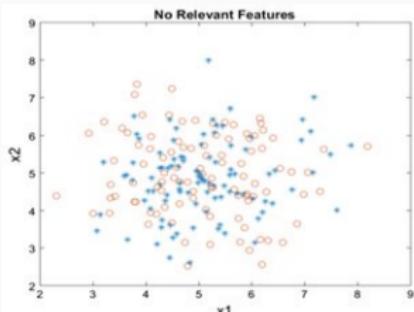
Útil em domínios com dados de alta dimensionalidade e poucas amostras
(ex.: expressão gênica).

Features Relevantes, Redundantes e Irrelevantes

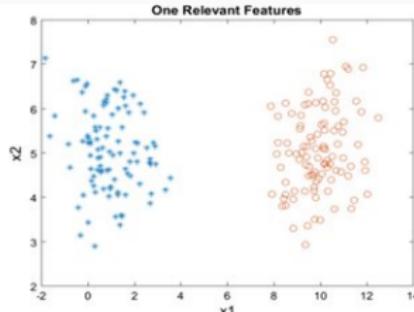
- **Relevantes:** Contêm informação útil para a predição
- **Redundantes:** Correlacionadas com outras features, não adicionam nova informação
- **Irrelevantes:** Não relacionadas à variável alvo

Objetivo da Seleção de Features: Retirar as relevantes, remover as redundantes/irrelevantes.

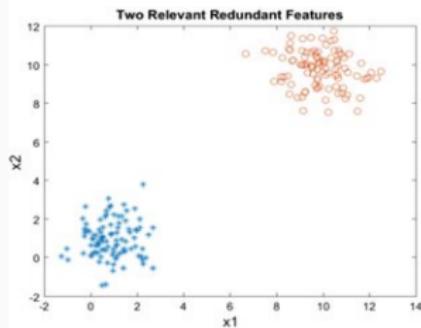
Features Relevantes, Redundantes e Irrelevantes



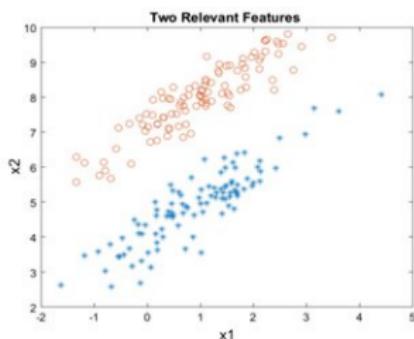
(a) Two-class example with no relevant features



(b) Two-class example with one relevant features

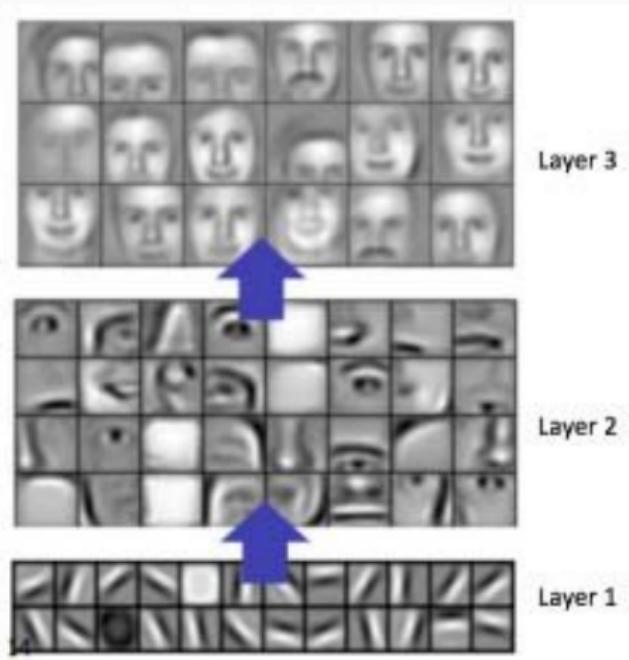


(c) Two-class example with two redundant features



(d) Two-class example with two relevant features

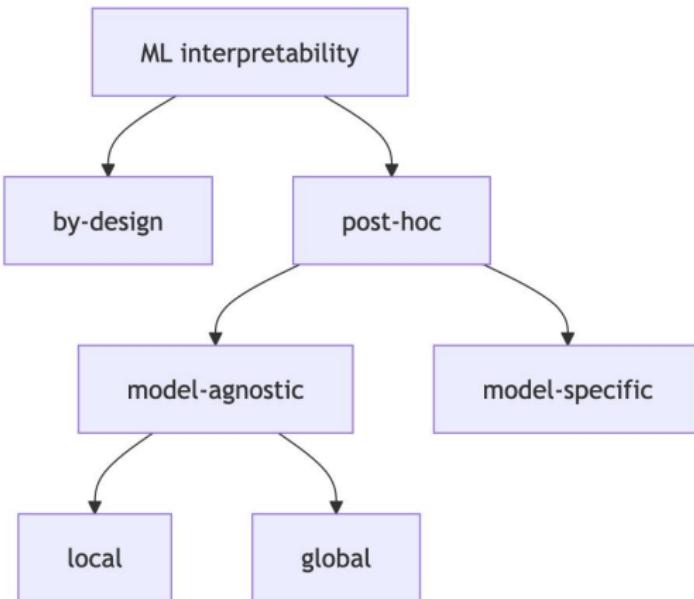
Features aprendidas



Features aprendidas por uma Rede Neural Convolucional (Al-Zawi, Mohammed, and Albawi, 2017)

Interpretabilidade

Taxonomia



Formas de Interpretabilidade

Interpretabilidade by design

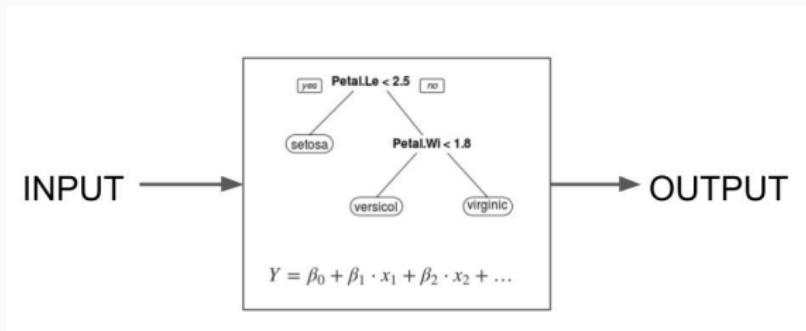
- O modelo já é inherentemente interpretável
- Ex.: Regressão logística ao invés de Random Forest

Interpretabilidade pós-treinamento (post-hoc)

- A interpretação é aplicada após o modelo estar treinado
- Pode ser:
 - **Agnóstica ao modelo**
 - Não depende do tipo de modelo
 - **Local:** explica previsões individuais
 - **Global:** explica padrões do conjunto de dados
 - Ex.: Importância de features por permutação
 - **Específica ao modelo**
 - Depende da estrutura interna do modelo
 - Ex.: análise das features aprendidas em redes neurais

Modelos intrinsecamente interpretáveis

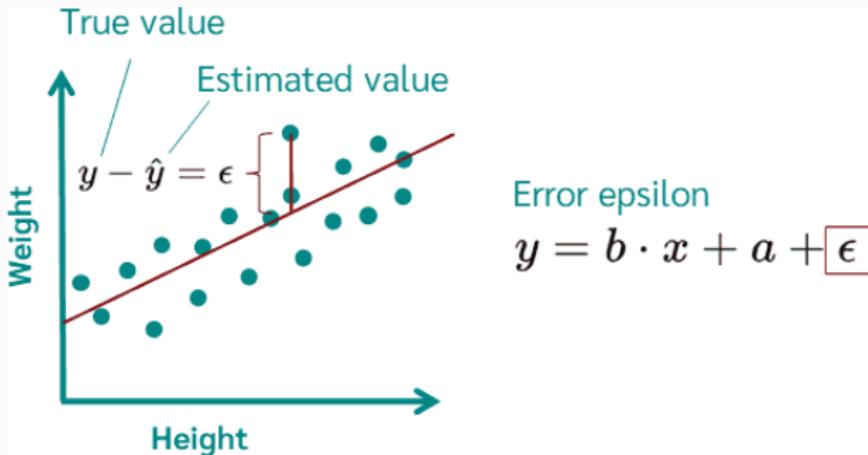
Modelos intrinsecamente interpretáveis



Exemplos:

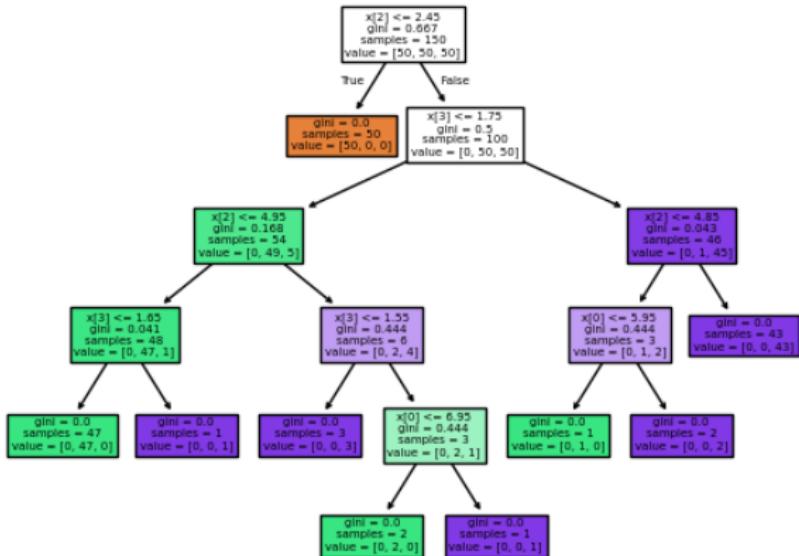
- Regressão linear: Ajusta um modelo linear minimizando a soma dos erros quadráticos.
- Regressão logística: Estende a regressão linear para classificação usando uma transformação não linear.
- Árvores de decisão: Dividem os dados recursivamente para criar modelos em forma de árvore.
- Regras de decisão: Extraem regras do tipo “se-então” a partir dos dados.
- RuleFit: Combina regras derivadas de árvores com regressão Lasso para aprender modelos esparsos baseados em regras.

Modelos intrinsecamente interpretáveis



Modelos intrinsecamente interpretáveis

Decision tree trained on all the iris features



Modelos intrinsecamente interpretáveis

O modelo é totalmente interpretável.

- Ex.: uma árvore de decisão pequena e visualizável; uma regressão linear com poucos coeficientes.
- Porém, “totalmente interpretável” é uma exigência **difícil** e um conceito um pouco **vago**.
- Em geral, só se aplica a modelos muito simples (regressão linear muito esparsa ou árvores muito curtas).

Modelos intrinsecamente interpretáveis

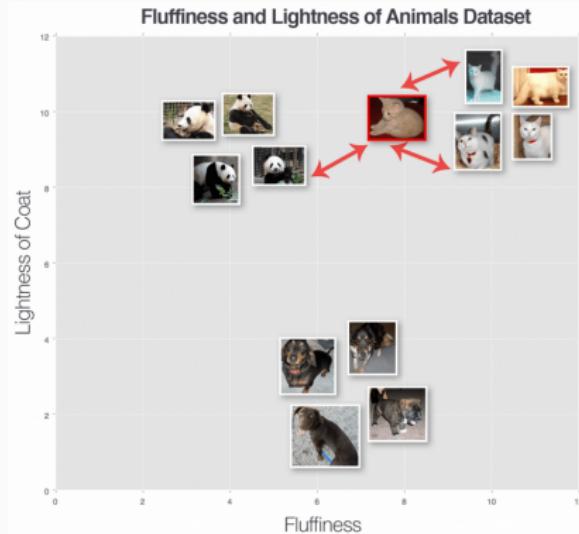
Partes do modelo são interpretáveis.

- Mesmo que o modelo completo seja complexo, algumas componentes podem ser interpretadas.
- Ex.: em regressão com muitas variáveis, ainda podemos interpretar cada coeficiente individualmente.
- Em listas ou árvores grandes, é possível inspecionar regras específicas.

Modelos intrinsecamente interpretáveis

As previsões do modelo são interpretáveis.

- Algumas abordagens permitem explicar **cada previsão individual**.
- Ex.: em um método tipo *k*-vizinhos para imagens, basta mostrar as imagens mais semelhantes usadas na decisão.
- Em árvores de decisão, a previsão pode ser explicada listando o caminho de regras seguido.



Modelos intrinsecamente interpretáveis

- **Fácil debugging e melhoria:** entendemos o funcionamento interno do modelo.
- **Boa justificativa das previsões:** facilita explicar decisões e validar com especialistas.
- **Uso comum em áreas científicas:** ex.: regressão logística em medicina.
- **Limite para descoberta de insights:**
 - Para usar o modelo como explicação do mundo, é preciso assumir que sua estrutura reflete a realidade.
 - Diferentes modelos podem ter desempenho similar, mas levar a interpretações distintas.

Efeito Rashomon

- Akira Kurosawa, 1950
- Num julgamento, os depoimentos são conflitantes, pondo em choque a verdade de cada personagem.



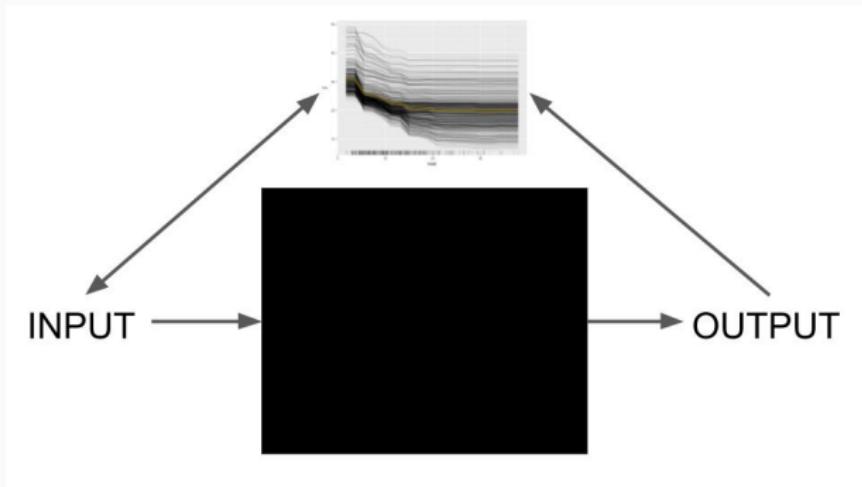
Existem diversos modelos igualmente bons, mas com explicações diferentes — difícil saber qual interpretar.

Interpretabilidade post-hoc

Interpretabilidade post-hoc

- Aplicada **depois** que o modelo já foi treinado.
- **Agnóstica ao modelo:**
 - Não considera a estrutura interna do modelo.
 - Analisa como a saída muda quando alteramos as entradas.
 - Ex.: permutar uma feature e medir o aumento do erro.
- **Específica ao modelo:**
 - Usa informações internas do modelo.
 - Ex.: neurônios que respondem a certos padrões em redes neurais; importância Gini em Random Forests.

Interpretabilidade post-hoc agnóstica



Interpretabilidade post-hoc agnóstica

Princípio SIPA (Scholbeck et al., 2019):

1. **Sample** — Amostrar dados
2. **Intervene** — Intervir nos dados
3. **Predict** — Obter previsões do modelo
4. **Aggregate** — Agregar resultados

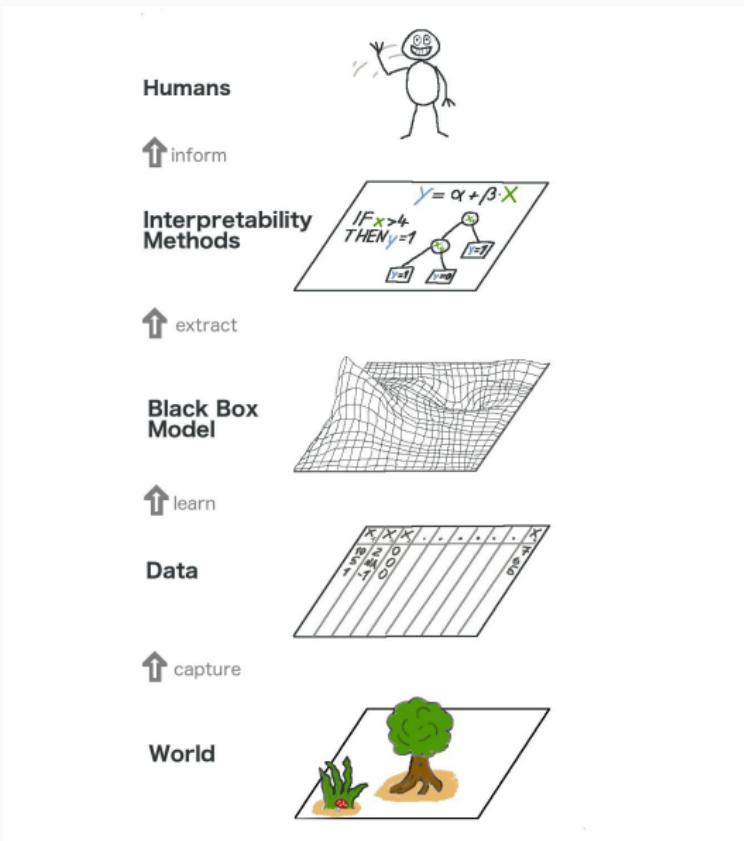
Exemplo: Importância por Permutação

- Pegamos uma **amostra** do conjunto de dados.
- **Intervimos** permutando os valores de uma feature.
- Obtemos as **previsões** novamente.
- **Comparamos** o erro antes e depois da permutação.

Por que é agnóstico?

Não precisamos acessar pesos, coeficientes ou estrutura interna do modelo.

Interpretabilidade post-hoc agnóstica



Interpretabilidade post-hoc agnóstica

Separar explicação do modelo dá flexibilidade (Ribeiro, Singh, and Guestrin, 2016b).

- **Flexibilidade na escolha do modelo:**

- Podemos trocar o modelo (ex.: XGBoost → outro) sem mudar a técnica de interpretação.

- **Flexibilidade na escolha do método de interpretação:**

- Ex.: trocar **PDP** por **ALE** sem retreinar ou alteração do modelo.

- **Flexibilidade na representação das features:**

- Ex.: explicações para imagens usando **regiões (patches)** em vez de pixels.

- Em contraste, modelos **intrinsecamente interpretáveis** mudam a forma de interpretação quando mudamos o modelo (ex.: regressão linear → classificador baseado em regras).

Interpretabilidade post-hoc agnóstica

- **Métodos Locais**

- Explicam **previsões individuais**.
- Foco: por que *este* exemplo recebeu *esta* saída.

- **Métodos Globais**

- Descrevem **como as features influenciam as previsões em média**.
- Foco: comportamento geral do modelo no conjunto de dados.

Métodos locais: quando usar?

- **Debugging do modelo:** fornecem uma visão “aproximada” de **previsões individuais**.
- Úteis para investigar **casos extremos** ou previsões com maior erro.
- Ajudam a identificar:
 - se o ponto é difícil de prever,
 - se o modelo é insuficiente,
 - se o dado pode estar rotulado incorretamente.
- Para **melhorias estruturais do modelo**, métodos **globais** costumam ser mais eficazes.

Exemplo de interpretabilidade local

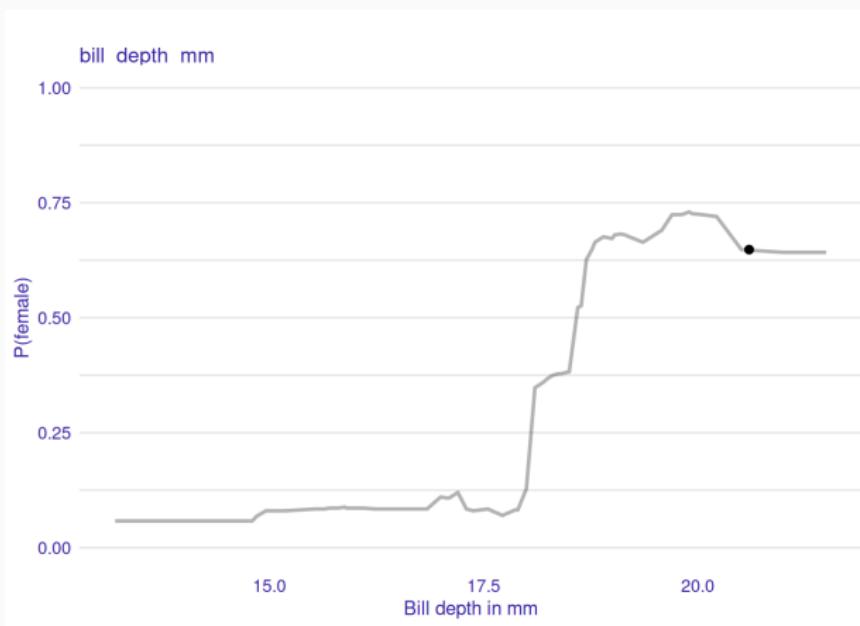


Gráfico Ceteris Paribus (“todo o mais é constante”) para a profundidade do bico e um pinguim específico. A linha mostra o valor previsto para esse pinguim ao variar a profundidade do bico. A profundidade real do bico do pinguim está marcada com um ponto.

Local para Justificação vs. Obtenção de Insights

- **Justificar previsões:** resultados mistos.
 - **Ceteris Paribus e Contrafactuals:** boas justificações (refletem o modelo diretamente).
 - **SHAP / LIME:** são modelos adicionais em cima do modelo → podem não ser ideais para decisões **críticas** (Rudin, 2019).
- **Insights sobre dados:**
 - SHAP e valores de Shapley permitem comparar a previsão atual com subconjuntos de referência.
 - **Ceteris Paribus e ICE** também ajudam a entender a resposta do modelo a variações de uma feature.
- A utilidade dos métodos locais **depende da qualidade do modelo**.

Métodos Globais Post-hoc Agnósticos ao Modelo

- Descrevem o **comportamento médio** do modelo no conjunto de dados.
- Não dependem da estrutura interna do modelo.
- Úteis para entender como o modelo funciona **de modo geral**.
- Frequentemente expressos como **valores esperados** baseados na distribuição dos dados.

Categorias de Métodos Globais

1) Efeitos de Features

Mostram a relação entre entrada e saída.

- Ex.: PDP, ALE, H-statistic, decomposições.

2) Importância de Features

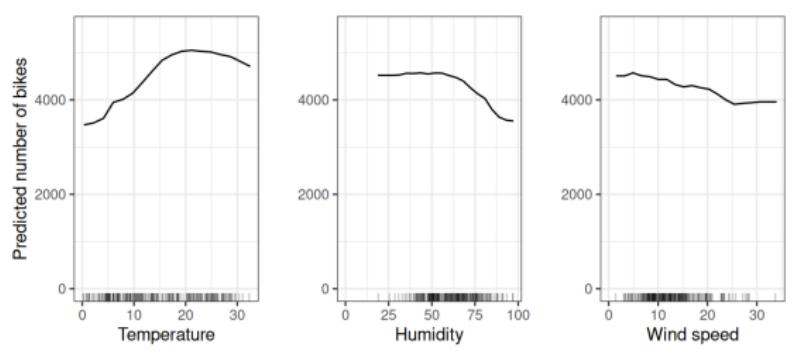
Ordenam as variáveis pela sua relevância.

- Ex.: PFI, LOFO, SHAP Importance.

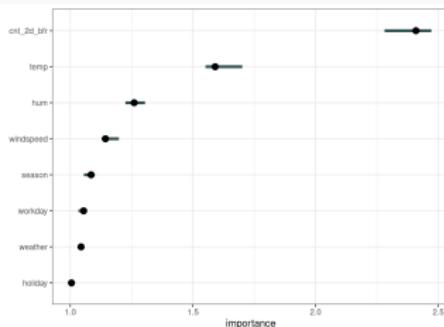
Resumo: efeitos explicam **como** uma feature influencia; importância diz **quanto** ela importa.

Efeitos de Features vs Importância de Features

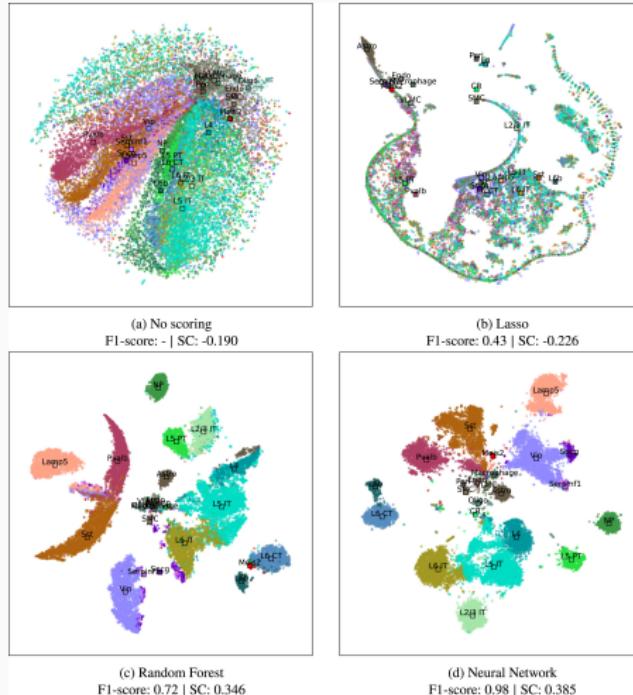
Efeitos de Features



Importância de Features

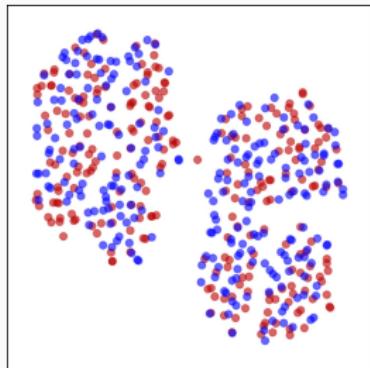


Weighted t-SNE

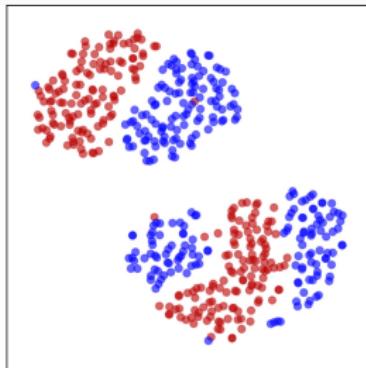


(Grisci, Inostroza-Ponta, and Dorn, 2025)

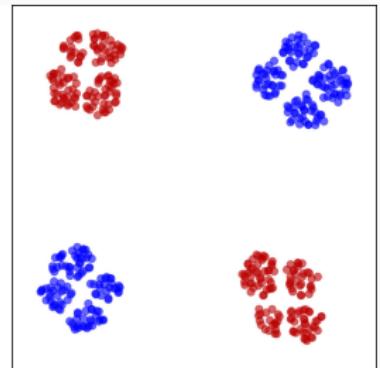
Weighted t-SNE



(a) Epoch: 45
Loss: 0.844
SC: -0.0008



(b) Epoch: 90
Loss: 0.726
SC: 0.0875



(c) Epoch: 135
Loss: 0.366
SC: 0.1992

(Grisci, Inostroza-Ponta, and Dorn, 2025)

Surrogate Models (Modelos Substitutos)

- **Ideia central:** Treinar um **modelo interpretável** para **aproximar as previsões** de um modelo caixa-preta.
- A interpretação é feita sobre o **modelo substituto**, não sobre a realidade.
- Método **agnóstico ao modelo**: não depende da arquitetura interna da caixa-preta, apenas de suas previsões.
- Qualquer modelo interpretável pode ser usado como substituto:
 - Regressão linear, árvores pequenas, listas de regras, etc.
- Usos em ML e também na engenharia (quando simulações são caras).

Vantagens e Limitações dos Surrogate Models

- **Vantagens**

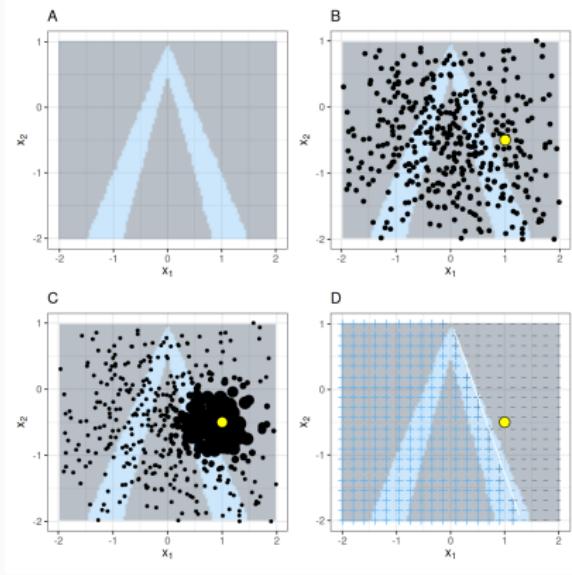
- Fácil de explicar para públicos diferentes.
- Permite trocar a caixa-preta sem mudar o método de interpretação.
- Dá flexibilidade para escolher o tipo de explicação (ex.: linear vs árvore).

- **Limitações**

- A interpretação reflete o **modelo**, não necessariamente o **mundo real**.
- Importa saber **quão próximo** o surrogate está da caixa-preta (ex.: R^2).
- Pode funcionar bem para alguns subgrupos e mal para outros.

- **Versão local**

- Ponderar os exemplos pela proximidade de uma instância → **surrogate local** para explicar uma previsão específica.
- Ex.: Local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, and Guestrin, 2016a)

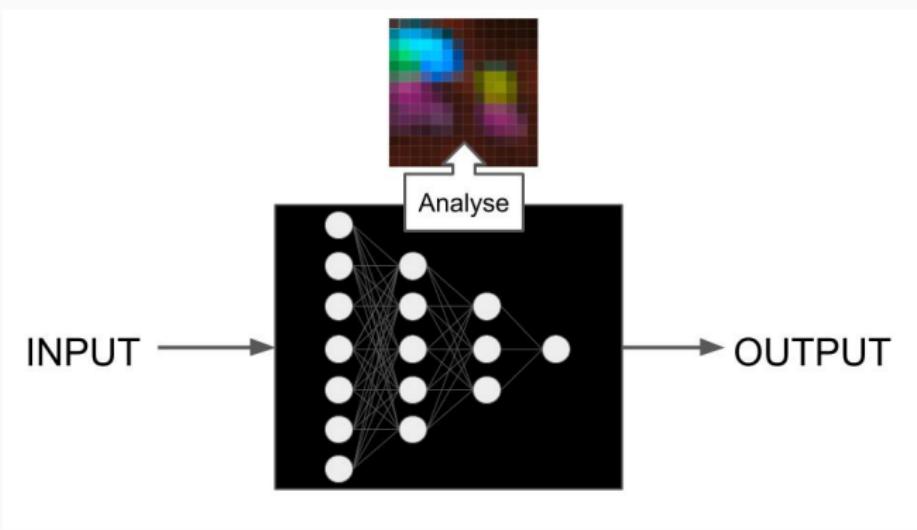


Algoritmo LIME para dados tabulares. A) Superfície de predição dada pelas features x_1 e x_2 . Classes previstas: 1 (escuro) ou 0 (claro). B) Instância de interesse (ponto grande) e dados amostrados (pontos pequenos). C) Atribuição de pesos com base na distância até a instância. D) Sinais (+/-) mostram as classificações do modelo local aprendido a partir das amostras ponderadas. A linha branca marca a fronteira de decisão ($P(c=1) = 0.5$).

Usos dos Métodos Globais

- **Debugging e melhoria do modelo**
 - LOFO se conecta diretamente à seleção de features.
- **Justificativa para stakeholders**
 - Permitem mostrar **quais variáveis foram mais relevantes**.
- **Combinam bem com modelos intrinsecamente interpretáveis**
 - Ex.: listas de regras podem explicar previsões individuais, enquanto métodos globais justificam **o modelo como um todo**.

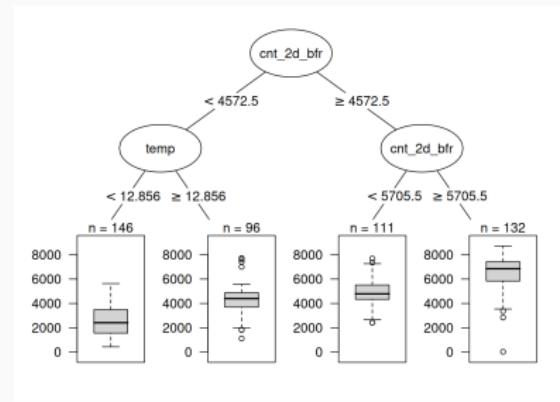
Métodos Post-hoc Específicos ao Modelo



Métodos Post-hoc Específicos ao Modelo

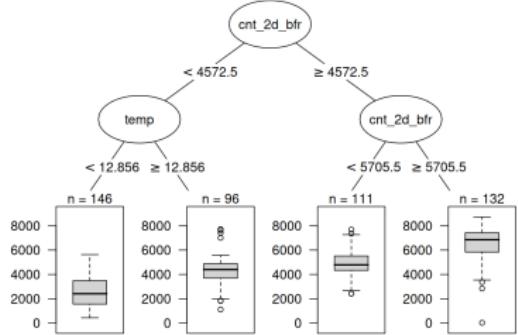
- São aplicados **após o treinamento**, mas funcionam apenas para **modelos específicos**.
- Exemplos:
 - Importância Gini em **Random Forests**
 - Razões de chances (odds ratios) em **Rregressão Logística**

Gini importance

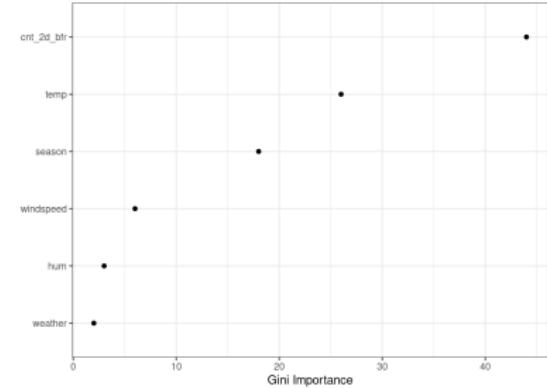


Árvore de regressão ajustada aos dados
de aluguel de bicicletas. Os boxplots
mostram a distribuição do número de
bicicletas alugadas no nó terminal.

Gini importance



Árvore de regressão ajustada aos dados de aluguel de bicicletas. Os boxplots mostram a distribuição do número de bicicletas alugadas no nó terminal.



A importância geral de uma variável em uma árvore de decisão pode ser calculada pela redução da variância ou índice de Gini nos nós onde ela é usada. A soma das importâncias é normalizada para 100.

Vantagens e Limites dos Métodos Específicos ao Modelo

- **Principal vantagem:** permitem aprender sobre o **modelo em si**.
 - Úteis para **melhorar** o modelo.
 - Facilitam **justificar** o modelo para outras pessoas.
- **Limitação para insights sobre os dados:**
 - Assim como modelos intrinsecamente interpretáveis, é necessário assumir que a estrutura do modelo **reflete a realidade**.
 - Ou seja: é preciso uma **justificativa teórica** para confiar que a interpretação reflete os dados.

Interpretando Redes Neurais

Interpretando Redes Neurais

- Previsões de redes neurais resultam de **muitas camadas** com multiplicações por **milhões de pesos** + transformações não lineares.
- Não é possível acompanhar manualmente o mapeamento entrada → saída.
- Para interpretar redes neurais, precisamos de **métodos específicos** porque:
 - Elas aprendem **características** (features) e **conceitos** em camadas ocultas.
 - Podemos explorar **gradientes** para métodos computacionalmente eficientes.

Interpretando Redes Neurais

Se tivermos uma rede neural que classifica imagens de cachorros, há cinco perguntas principais...

1. Como estão as ativações neurais?
2. Qual é um bom exemplo de cachorro?
3. Como é a aparência de uma imagem de cachorro?
4. O que faz com que esta imagem seja mais ou menos um cachorro?
5. O que faz com que esta imagem seja um cachorro?



Activation maximization



garbage truck



gasmask



milk can



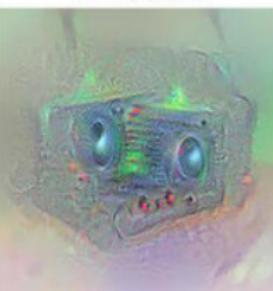
paddle



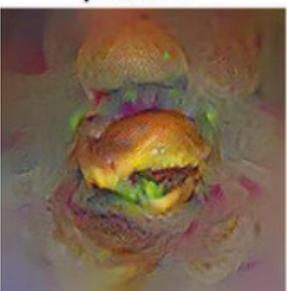
padlock



pirate



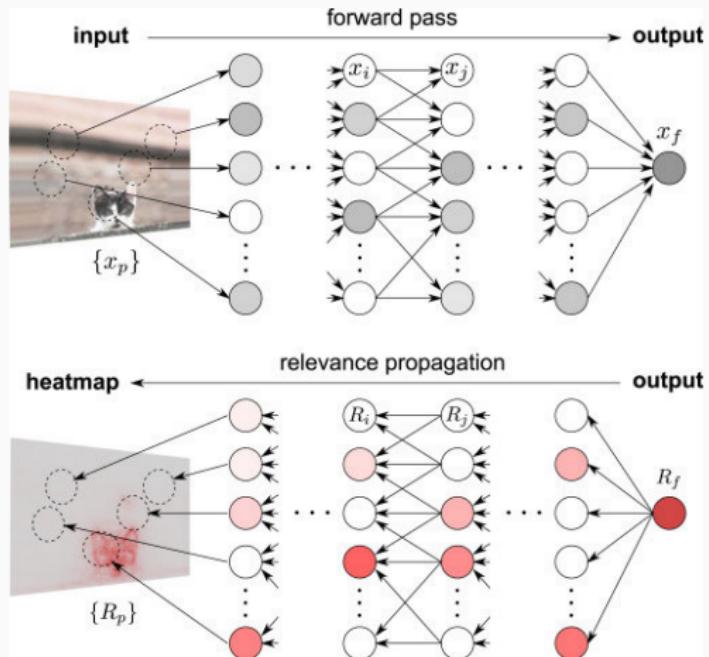
radio



cheeseburger

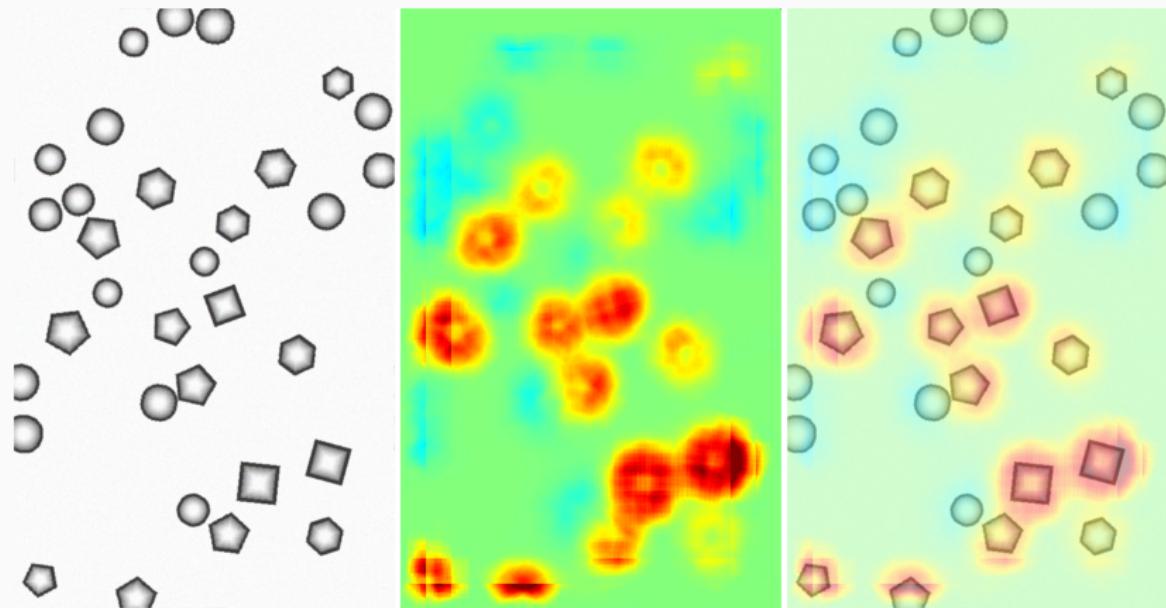
(Nguyen, Yosinski, and Clune, 2016)

Mapas de saliência



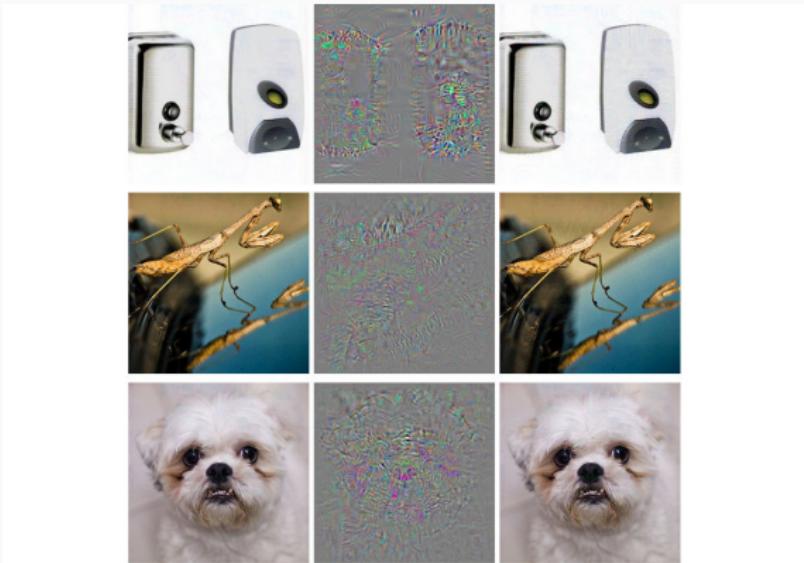
(Buhrmester, Münch, and Arens, 2021)

Layer-wise Relevance Propagation (LRP)



(Bach et al., 2015)

Ataques adversariais



Exemplos adversariais para o AlexNet por Szegedy et al., 2013. Todas as imagens na coluna da esquerda são classificadas corretamente. A coluna do meio mostra o erro (amplificado) adicionado às imagens para produzir as imagens na coluna da direita, todas classificadas (incorrectamente) como “aveSTRUZ” .

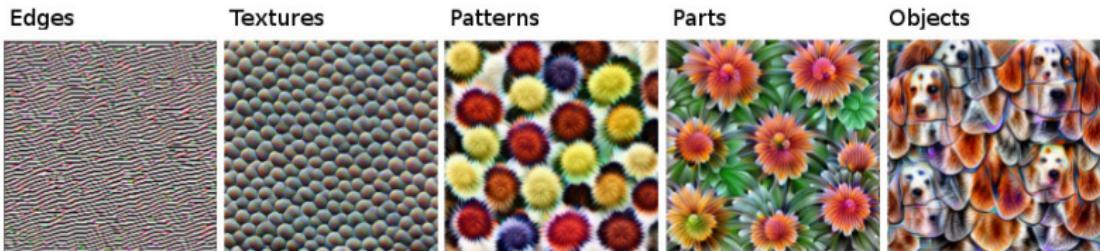
Mechanistic Interpretability

- Objetivo: **reengenheirar** redes neurais em termos de **mecanismos compreensíveis por humanos**.
- Foco comum: **Transformers e LLMs**, embora não seja limitado a eles.
- MI analisa o **interior do modelo**, portanto **não** utiliza métodos agnósticos como LIME ou SHAP.
- Na taxonomia geral, MI se enquadra em **interpretabilidade post-hoc específica ao modelo**.

Métodos Usados em Interpretabilidade Mecanicista

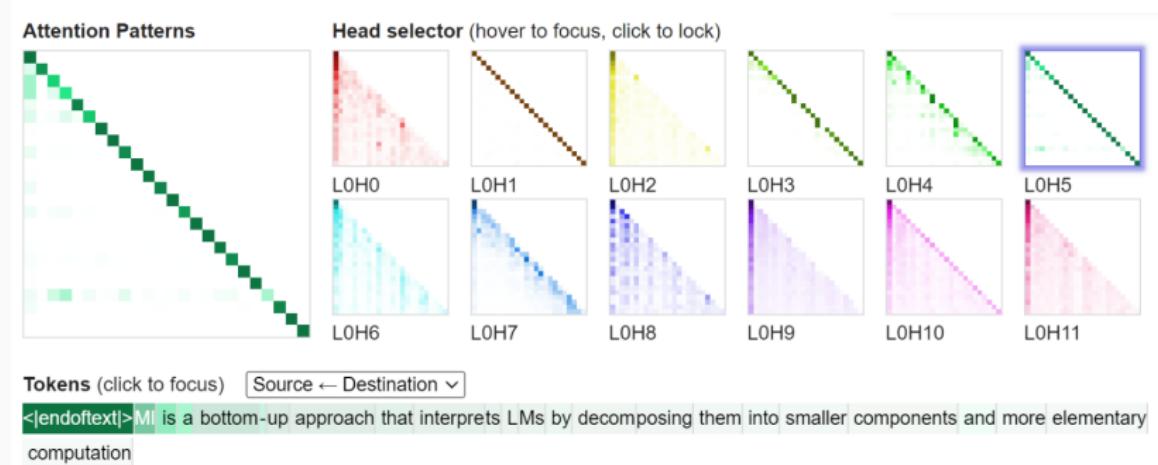
- **Visualização de features internas.**
- **Análise de circuitos:** como neurônios e cabeças de atenção se combinam para realizar funções.
- **Intervenções causais** dentro da rede.
- **Logit Lens:** aplica o logit final em camadas intermediárias (muito usado em transformers).
- Requer ferramentas para “abrir” e inspecionar representações internas.

Mechanistic Interpretability



Features aprendidas por uma rede neural convolucional (Inception V1) treinada no conjunto de dados ImageNet. As features variam de padrões simples nas camadas convolucionais inferiores (à esquerda) para características mais abstratas nas camadas superiores (à direita). Figura de Olah, Mordvintsev, and Schubert, 2017 (CC-BY 4.0) <https://distill.pub/2017/feature-visualization/appendix/>.

Mechanistic Interpretability



Visualização de atenção, criada utilizando a ferramenta de Cooney and Nanda, 2023.

Contexto Histórico da Interpretabilidade Mecanicista

- Termo cunhado em 2020 por **Chris Olah** (distill.pub, OpenAI → Anthropic).
- Inicialmente focada em **visão computacional**; com o hype dos **LLMs**, o foco mudou para **transformers**.
- A comunidade MI desenvolveu **linguagem própria** (ex.: “superposição” = conceitos entrelaçados ativando o mesmo neurônio).
- Há **interseção cultural** entre MI e **AI Safety** (inclui debates sobre AGI e riscos).
- Quando MI chegou ao NLP, houve **conflito** com a comunidade já existente: pesquisas anteriores foram ignoradas → “reinventando a roda”.
- Assim, MI é **também um marcador de comunidade**, não só um termo técnico — mas recentemente a comunidade tem se **ampliado**.

Objetos Fundamentais de Estudo em MI: Features

- Seguindo Olah, Cammarata, et al., 2020, a pesquisa em Interpretabilidade Mecanicista (MI) pode ser organizada em três áreas: **features**, **circuitos** e **universalidade**.
- **Features**: propriedades interpretáveis por humanos que são codificadas nas ativações do modelo.
- Exemplo:
 - Ao receber o token "dog", um modelo de linguagem pode ativar features como: *animal*, *pet*, *quatro patas*, etc.
- Meta da MI: **decodificar e interpretar** as representações aprendidas que aparecem nas ativações.

Features

Curves



3b:379

3b:406

3b:385

3b:343

3b:342



3b:388



3b:340



3b:330



3b:349



3b:324

Related Shapes (Circle, Spiral...)



3b:323



3b:325



3b:347



3a:101



4a:410



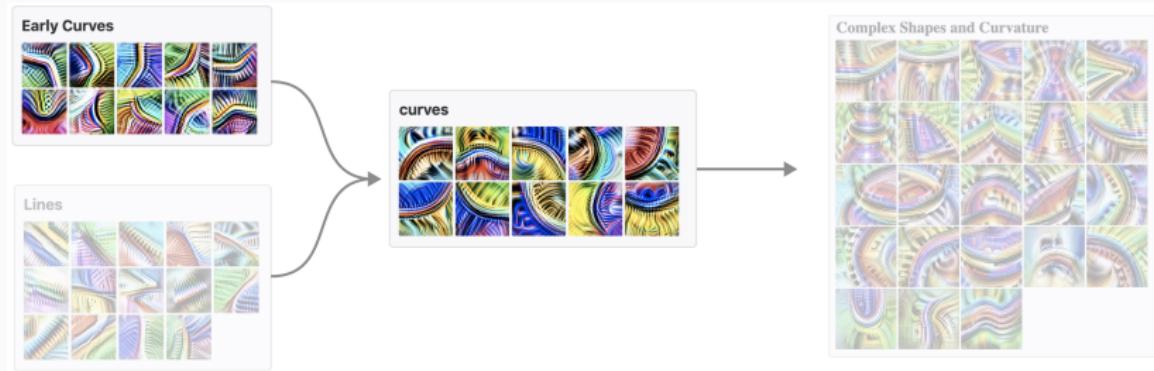
3a:176

(Olah, Cammarata, et al., 2020)

Objetos Fundamentais em MI: Circuitos

- O estudo de **features** mostra *o que* é representado nas ativações, mas não *como* essas informações são usadas.
- A Interpretabilidade Mecanicista investiga **circuitos**: caminhos computacionais que conectam features e possibilitam comportamentos específicos do modelo.
- **Definição (Olah, Cammarata, et al., 2020):**
 - Um circuito é um **subgrafo** do modelo que implementa um comportamento específico.
 - Nós = **features ou ativações de componentes do transformer**.
 - Arestas = conexões ponderadas entre essas ativações.
- Estudos posteriores ampliaram a definição para:
 - conexões entre cabeças de atenção, MLPs e camadas,
 - fluxos de informação mais gerais em LLMs.
- **Exemplo (Elhage et al., 2021):** Um **círculo de indução** em um pequeno LM usa duas cabeças de atenção para detectar e continuar sequências repetidas.

Circuitos

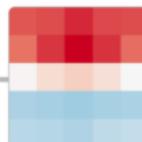


(Olah, Cammarata, et al., 2020)

Zoom In: An Introduction to Circuits

By studying the connections between neurons, we can find meaningful algorithms in the weights of neural networks.

Windows (4b:237) excite the car detector at the top and inhibit at the bottom.

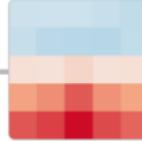


- positive (excitation)
- negative (inhibition)

Car Body (4b:491) excites the car detector, especially at the bottom.



Wheels (4b:373) excite the car detector at the bottom and inhibit at the top.



A car detector (4c:447) is assembled from earlier units.

(Olah, Cammarata, et al., 2020)

Circuitos

Group Related Nodes Into "Supernodes"

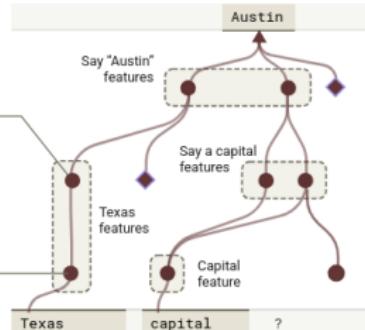
We group together features with related meanings that appear to play similar roles in the graph.

"Texas" feature #1

loved the "everything's **bigger in Texas**" joke implicit in the "te
became a state in 1845. Texas is **a big state with a big history**. T
nd a rodeo: Texas is known for its cowboys and cowgirls, and atte
also loved the "everything's **bigger in Texas**" joke implicit in the
Assistant: Here's a narrative about a trip to Texas: A Journey to

"Texas" feature #2

cy Hon. Pat M. Neff, governor of the state of Texas, that he ca
eo firsthand. 3. Explore the Big Bend National Park: The Big Ber
part of civil appeals for the Fourth supreme Judicial district to
SHANKLIN, Appellant, v. The STATE of Texas. No. PD. 0026-06.
heck. The Texas A&M University AgriLife Extension Service describe



Supernodes

Throughout the paper, we represent supernodes as stacked boxes

Say "Austin"

Hover over nodes for detailed feature visualizations. Select a feature to view in the top bar after hovering

Figure 3: Grouping related graph nodes into supernodes produces a simpler graph.

(Lindsey et al., 2025)

Objetos Fundamentais em MI: Universalidade

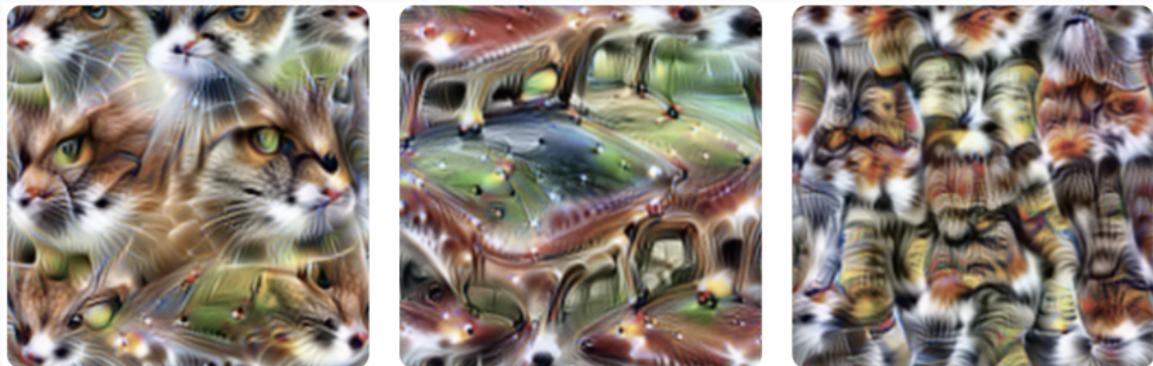
- Pergunta central: **features e circuitos descobertos em um modelo também aparecem em outros modelos ou tarefas?**
- **Universalidade** = grau em que estruturas semelhantes surgem em diferentes LMs e contextos (Olah, Cammarata, et al., 2020; Gurnee et al., 2024).
- **Se forem universais:**
 - Resultados obtidos em modelos pequenos podem **generalizar** para **LLMs maiores**.
 - Insights podem ser **reutilizados** e acelerar a interpretabilidade.
- **Se não forem universais:**
 - Cada modelo e tarefa exigirão **análise independente**.
 - A interpretabilidade pode se tornar **muito mais trabalhosa**.

Universalidade

	Curve detectors				High-Low Frequency detectors			
ALEXNET Krizhevsky et al. [34]								
INCEPTIONV1 Szegedy et al. [26]								
VGG19 Simonyan et al. [35]								
RESNETV2-50 He et al. [36]								

(Olah, Cammarata, et al., 2020)

Polissemia

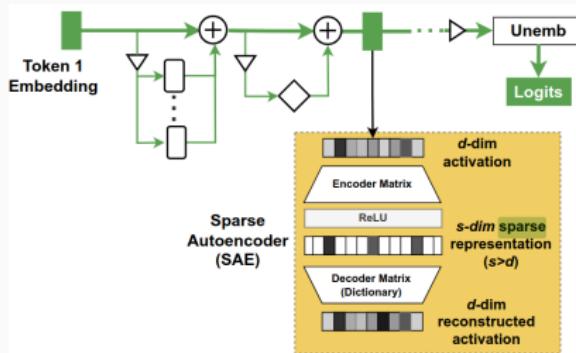


4e:55 é um neurônio polissêmico que responde a rostos de gatos, frentes de carros e pernas de gatos.

(Olah, Cammarata, et al., 2020)

Sparse Autoencoder (SAE)

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



Fractured Entangled Representations (FER)

GPT-3 on Counting Items

User:

I have 3 pencils, 2 pens, and 4 erasers. How many things do I have?

GPT-3:

You have 9 things. [correct in 3 out of 3 trials]

User:

I have 3 chickens, 2 ducks, and 4 geese. How many things do I have?

GPT-3:

You have 10 animals total. [incorrect in 3 out of 3 trials]

(Kumar et al., 2025)

Avaliando a interpretabilidade

Avaliação em Interpretabilidade

- Avaliar interpretabilidade é difícil porque **não existe ground truth** para explicações no mundo real.
- Benchmarks tradicionais (como em aprendizado supervisionado) **não se aplicam** diretamente.
- Quando precisamos de um “ground truth”, ele costuma ser **simulado**, não real.

Três níveis de avaliação:

- **Aplicação (tarefa real)** — testada por especialistas.
- **Humana (tarefa simples)** — testada com não especialistas.
- **Função (proxy)** — sem humanos; avalia propriedades internas ou estrutura.

Propriedades dos Métodos de Explicação

- **Expressividade:** forma da explicação (regras, árvores, pesos, linguagem natural).
- **Translucidez:** quanto depende de acessar o interior do modelo.
- **Portabilidade:** quão amplamente o método funciona entre diferentes modelos.
- **Complexidade Computacional:** custo para gerar a explicação.

Propriedades das Explicações Individuais

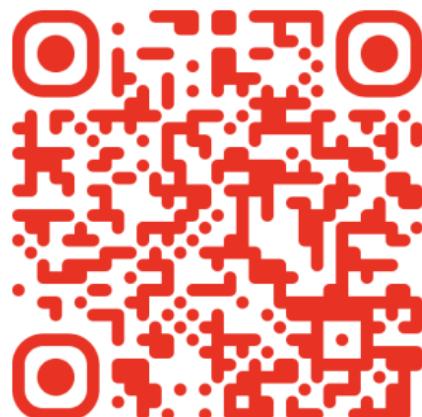
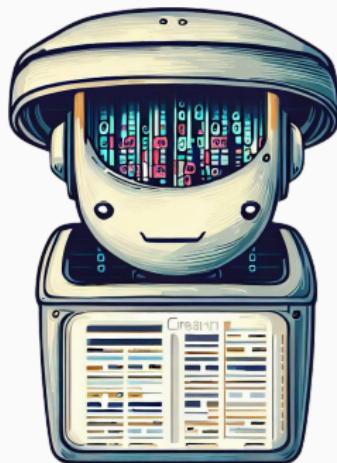
- **Acurácia:** capacidade de prever corretamente novos casos.
- **Fidelidade:** quão bem a explicação reflete a **caixa-preta**.
- **Consistência:** explicações semelhantes entre modelos com previsões semelhantes.
- **Estabilidade:** explicações semelhantes para instâncias parecidas.
- **Compreensibilidade:** quão bem humanos entendem a explicação.
- **Certeza:** se a explicação comunica confiança/risco.
- **Importância:** clareza sobre o que foi mais relevante na decisão.
- **Novidade:** se o ponto está fora da distribuição (baixa confiabilidade).
- **Representatividade:** quantas instâncias a explicação cobre (local x global).

Conclusão

Saiba mais

- Molnar, Christoph. Interpretable machine learning.
<https://christophm.github.io/interpretable-ml-book/> (Molnar, 2020)
- Prince, Simon JD. Understanding deep learning. MIT press, 2023.
<https://udlbook.github.io/udlbook/> (Prince, 2023)
- Barbieri, Matheus Cezimbra, Bruno Lochins Grisci, and Márcio Dorn. "Analysis and comparison of feature selection methods towards performance and stability." *Expert Systems with Applications* 249 (2024): 123667. <https://doi.org/10.1016/j.eswa.2024.123667> (Barbieri, Grisci, and Dorn, 2024)
- Rai, Daking, et al. "A practical review of mechanistic interpretability for transformer-based language models." *arXiv preprint arXiv:2407.02646* (2024). <https://arxiv.org/abs/2407.02646> (Rai et al., 2024)

ExpLAIn — Explainability Laboratory for Artificial Intelligence



- <https://brunogrisci.github.io/explain>
- <https://www.instagram.com/explain.ufrgs/>
- <https://www.linkedin.com/company/explainufrgs/>

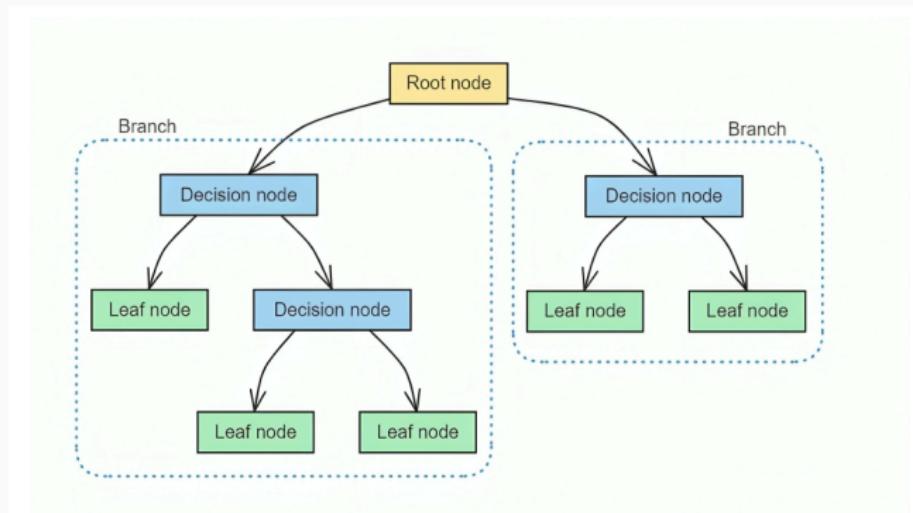
Atividade prática!

13/11 (10:30 - 12:00) — LAB 102/Prédio 43413

Atividade prática

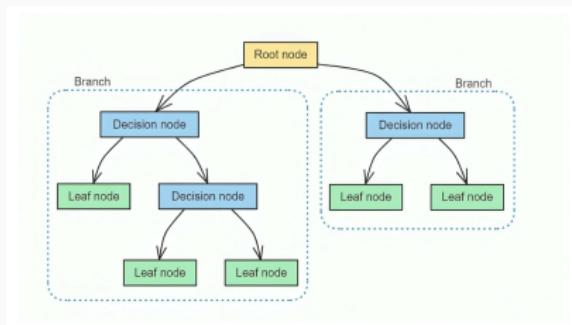
By design: Decision Tree (DT)

- Mapeamento hierárquico de uma sequência de escolhas para uma obtenção de uma classe/valor.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).



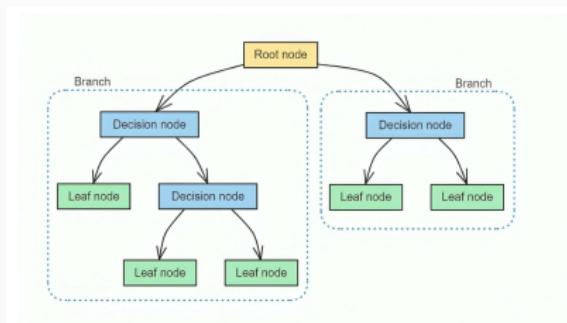
Model-specific: Feature Importance

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



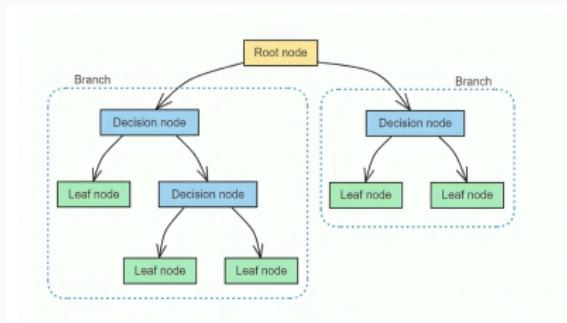
Local Model-agnostic: Ceteris Paribus (CP)

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



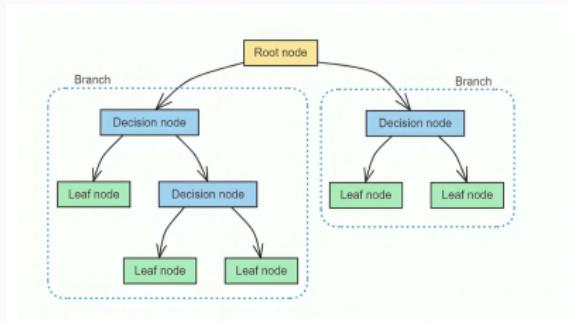
Local Model-agnostic: Individual Condition Expectation (ICE)

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



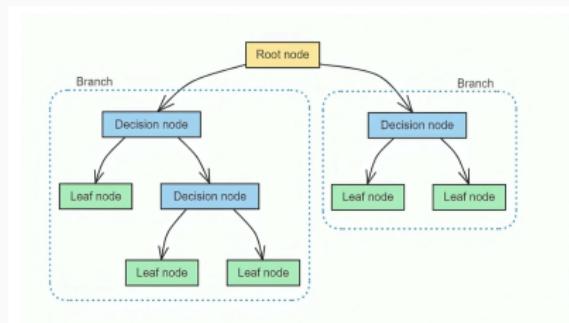
Global Model-agnostic: Partial Dependence Plot (PDP)

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



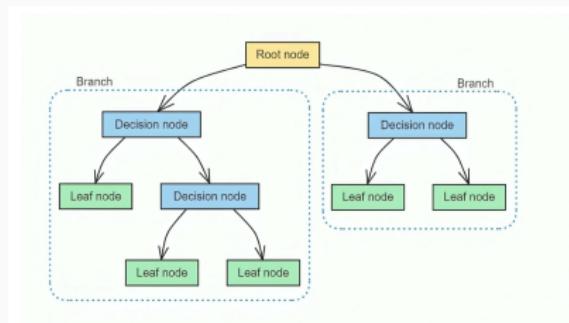
Global Model-agnostic: Permutation Feature Importance

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



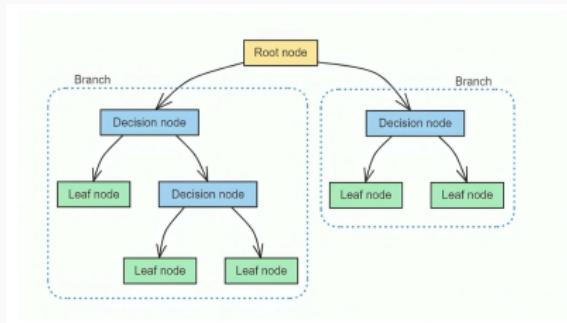
Shapley Values

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



Local Model-agnostic: SHapley Additive exPlanations (SHAP)

- Problema: **superposição** — ativações dos modelos codificam **mais features do que dimensões**.
- Resultado: **neurônios polissemânticos**, que respondem a vários conceitos não relacionados.
- SAEs projetam a ativação original (dimensão d) para uma **representação esparsa** de dimensão s ($s > d$).
- Essa nova representação tende a ter **neurônios monossêmicos**: cada um associado a **um único conceito**.
- Meta: tornar as ativações internas **mais interpretáveis**. (Rai et al., 2024)



References

-  Aurélien, Géron (2019). “**Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**”. In: *Concepts, tools, and techniques to build intelligent systems*, 2nd ednn.
-  Bach, Sebastian et al. (2015). “**On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation**”. In: *PloS one* 10.7, e0130140.
-  Barbieri, Matheus Cezimbra, Bruno lochins Grisci, and Márcio Dorn (2024). “**Analysis and comparison of feature selection methods towards performance and stability**”. In: *Expert Systems with Applications* 249, p. 123667.

Referências ii

-  Buhrmester, Vanessa, David Münch, and Michael Arens (2021). **“Analysis of explainers of black box deep neural networks for computer vision: A survey”**. In: *Machine Learning and Knowledge Extraction* 3.4, pp. 966–989.
-  Cooney, Alan and Neel Nanda (2023). **Circuitsvis**.
-  DeLMA and Will Cukierski (2013). **The ICML 2013 Whale Challenge - Right Whale Redux**.
<https://kaggle.com/competitions/the-icml-2013-whale-challenge-right-whale-redux>. Acessado em: 2025-03-09.
-  Domingos, Pedro (2012). **“A few useful things to know about machine learning”**. In: *Communications of the ACM* 55.10, pp. 78–87.
-  Elhage, Nelson et al. (2021). **“A mathematical framework for transformer circuits”**. In: *Transformer Circuits Thread* 1.1, p. 12.
-  Geirhos, Robert et al. (2020). **“Shortcut learning in deep neural networks”**. In: *Nature Machine Intelligence* 2.11, pp. 665–673.

-  Grisci, Bruno Iochins, Mario Inostroza-Ponta, and Márcio Dorn (2025). “**Assessing feature scorer results on high-dimensional datasets with t-SNE**”. In: *Neurocomputing*, p. 130561.
-  Gurnee, Wes et al. (2024). “**Universal neurons in gpt2 language models**”. In: *arXiv preprint arXiv:2401.12181*.
-  Kumar, Akarsh et al. (2025). “**Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis**”. In: *arXiv preprint arXiv:2505.11581*.
-  Lapuschkin, Sebastian et al. (2019). “**Unmasking Clever Hans predictors and assessing what machines really learn**”. In: *Nature communications* 10.1, p. 1096.
-  Lindsey, Jack et al. (Mar. 2025). **On the Biology of a Large Language Model**. Accessed: 2025-11-11. URL: <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.

-  Miller, Tim (2019). “**Explanation in artificial intelligence: Insights from the social sciences**”. In: *Artificial intelligence* 267, pp. 1–38.
-  Molnar, Christoph (2020). **Interpretable machine learning**. Lulu.com.
-  — (Nov. 2025). **Points, Rules, Weights, Distributions: The Elements of Machine Learning**. Accessed: 2025-11-09. URL: <https://mindfulmodeler.substack.com/p/points-rules-weights-distributions>.
-  Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). “**Methods for interpreting and understanding deep neural networks**”. In: *Digital signal processing* 73, pp. 1–15.
-  Murdoch, W James et al. (2019). “**Definitions, methods, and applications in interpretable machine learning**”. In: *Proceedings of the National Academy of Sciences* 116.44, pp. 22071–22080.

-  Nguyen, Anh, Jason Yosinski, and Jeff Clune (2016). “**Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks**”. In: *arXiv preprint arXiv:1602.03616*.
-  Olah, Chris, Nick Cammarata, et al. (2020). “**Zoom in: An introduction to circuits**”. In: *Distill* 5.3, e00024–001.
-  Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert (2017). “**Feature visualization**”. In: *Distill* 2.11, e7.
-  Prince, Simon JD (2023). **Understanding deep learning**. MIT press.
-  Rai, Daking et al. (2024). “**A practical review of mechanistic interpretability for transformer-based language models**”. In: *arXiv preprint arXiv:2407.02646*.

-  Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). **“” Why should i trust you?” Explaining the predictions of any classifier”**. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.
-  — (2016b). **“Model-agnostic interpretability of machine learning”**. In: *arXiv preprint arXiv:1606.05386*.
-  Roscher, Ribana et al. (2020). **“Explainable machine learning for scientific insights and discoveries”**. In: *Ieee Access* 8, pp. 42200–42216.
-  Rudin, Cynthia (2019). **“Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”**. In: *Nature machine intelligence* 1.5, pp. 206–215.

-  Scholbeck, Christian A et al. (2019). “**Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations**”. In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer, pp. 205–216.
-  Szegedy, Christian et al. (2013). “**Intriguing properties of neural networks**”. In: *arXiv preprint arXiv:1312.6199*.
-  Al-Zawi, SATAMS, T Mohammed, and S Albawi (2017). “**Understanding of a convolutional neural network**”. In: *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6.