AIMLAI @ ECML PKDD 2026
September 7th 2026

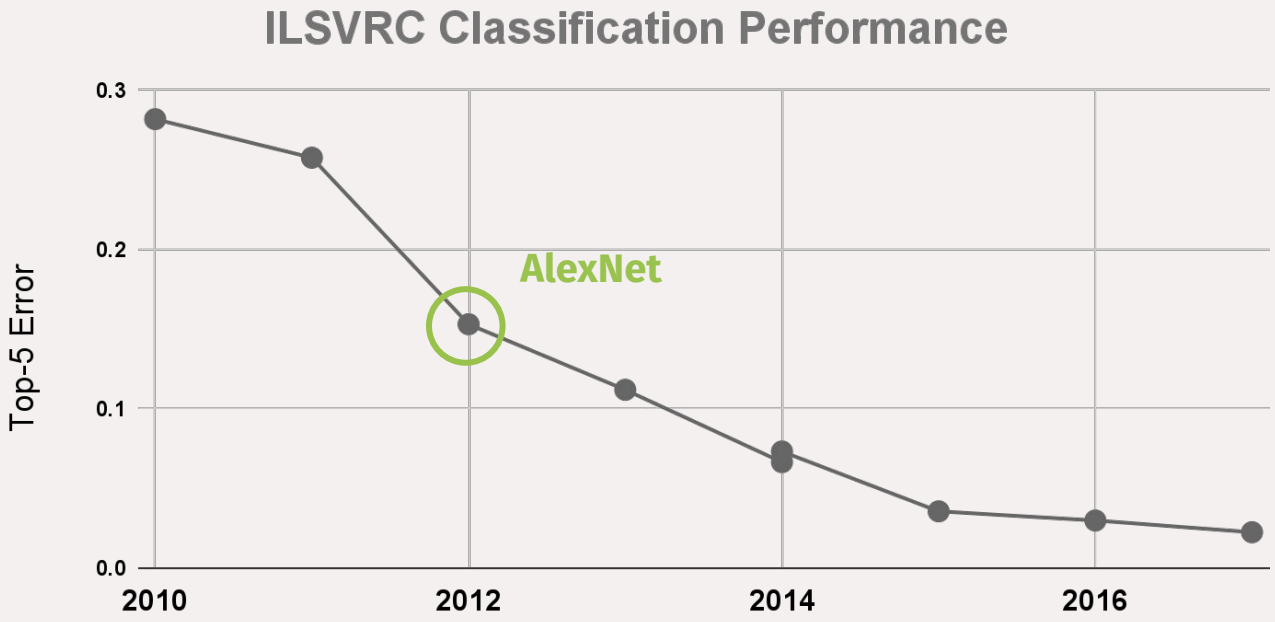
On the Evolution of Interpretability Methods

Ward Gauderis · Thomas Dooms · Geraint Wiggins · José Oramas

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP
Universiteit Antwerpen

fwo Flanders AI Research Program

13 years since AlexNet, 8 since Attention and 3 since ChatGPT... Deep models are here to stay



Adoption has not only increased in volume but in sophistication.

Microsoft's speech recognition engine listens as well as a human

"This is an historic achievement" - Xuedong Huang

Andrew Tarantola, @terrortola
10.18.16 in Personal Computing

Intelligent Machines

Deep-Learning Machine Listens to Bach, Then Writes Its Own Music in the Same Style

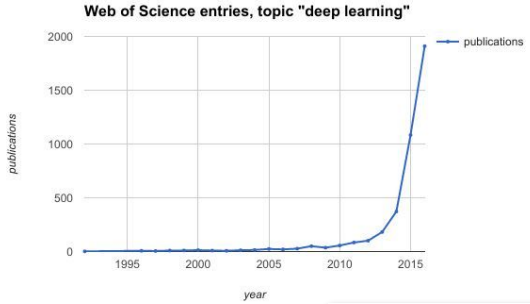
Can you tell the difference between music composed by Bach and by a neural network?

by Emerging Technology from the arXiv December 14, 2016

The Big Read Driverless vehicles + Add to myFT

Driverless cars inspire a new gold rush in California

MAY 23, 2017 by: Leslie Hook and Tim Bradshaw



WIRED BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

TOM SIMONITE BUSINESS 10.25.2019 03:00 AM

Google Search Now Reads at a Higher Level

The company is incorporating new software that better understands subtleties of language, with the biggest changes for queries outside the US.

Article | Open Access | Published: 29 August 2019

Deep Learning to Improve Breast Cancer Detection on Screening Mammography

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

Scientific Reports **9**, Article number: 12495 (2019) | Cite this article

9229 Accesses | **2** Citations | **27** Altmetric | Metrics

Yes! Interpreting these models is necessary for trusting their predictions in high-stakes scenarios.

https://www.itmastersmag.com/transformacion-digital/que-el-gdpr-no-sea-una-amenaza/

Clever Hans Effect

Article | Open access | Published: 11 March 2019

Unmasking Clever Hans predictors and assessing what machines really learn

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek ✉ & Klaus-Robert Müller ✉

Nature Communications **10**, Article number: 1096 (2019) | Cite this article

Yes! Interpreting these models is necessary for trusting their predictions in high-stakes scenarios

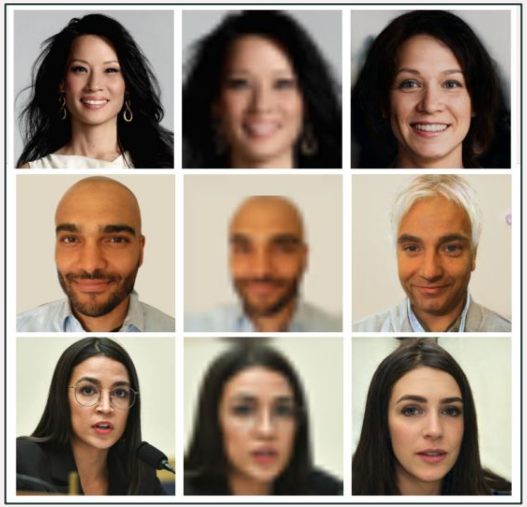
Article | Published: 31 May 2021

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave, Joseph D. Janizek & Su-In Lee

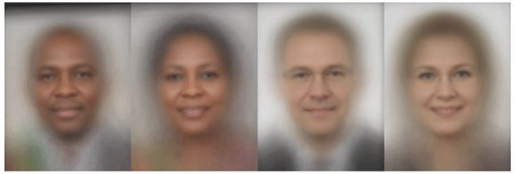
Nature Machine Intelligence **3**, 610–619 (2021) | Cite this article

It can exploit shortcuts ...



... and hallucinate

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

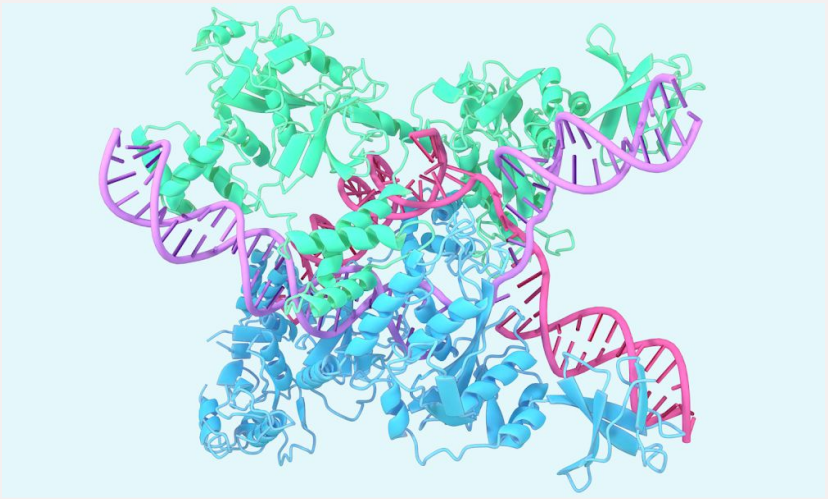


... be biased & unfair...

Current models outperform human experts. What do these models know that we don't?

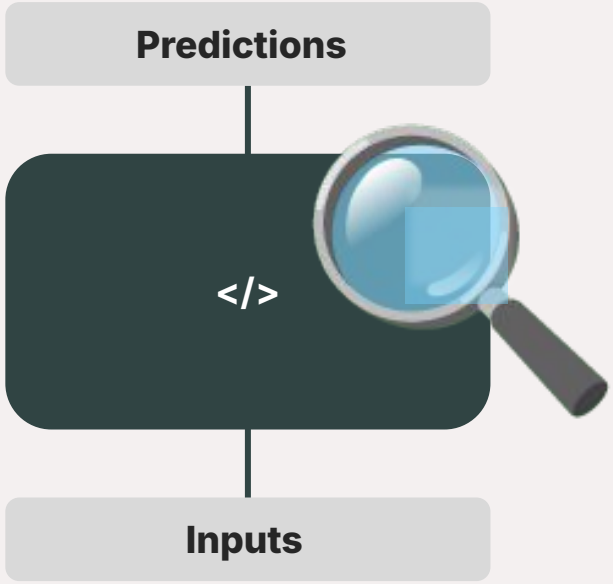
Games (2016)

Sciences (2018)



Transparency - AI Interpretability

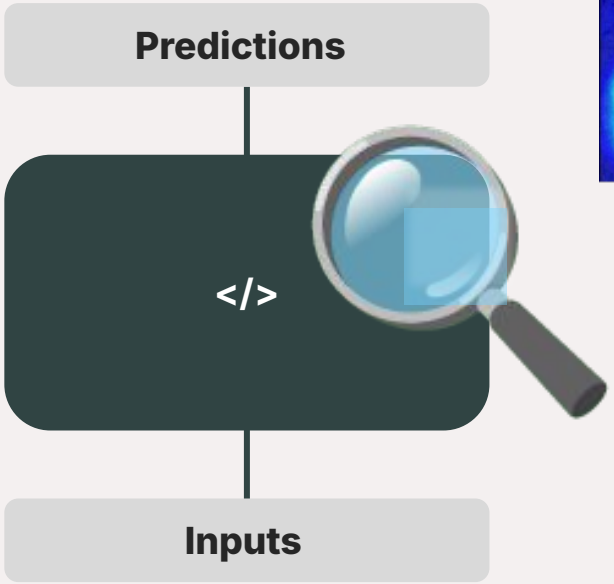
Desideratum: Understanding the inner-workings of the decision-making process

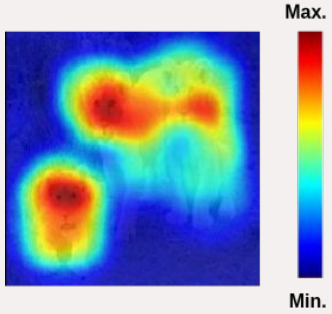


Discovery

Transparency - AI Interpretability

In practice: Relating inputs to outputs



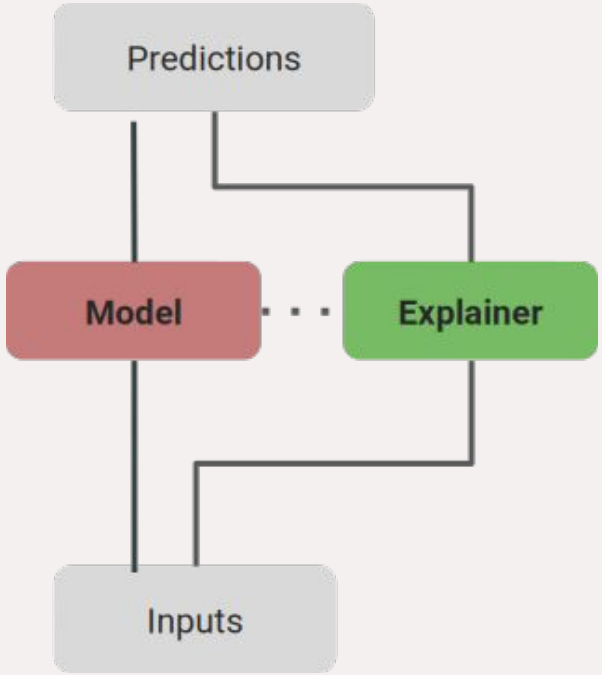


Attribution

Transparency - AI Interpretability

In practice: Relating inputs with outputs

Post-Hoc Attribution (XAI)

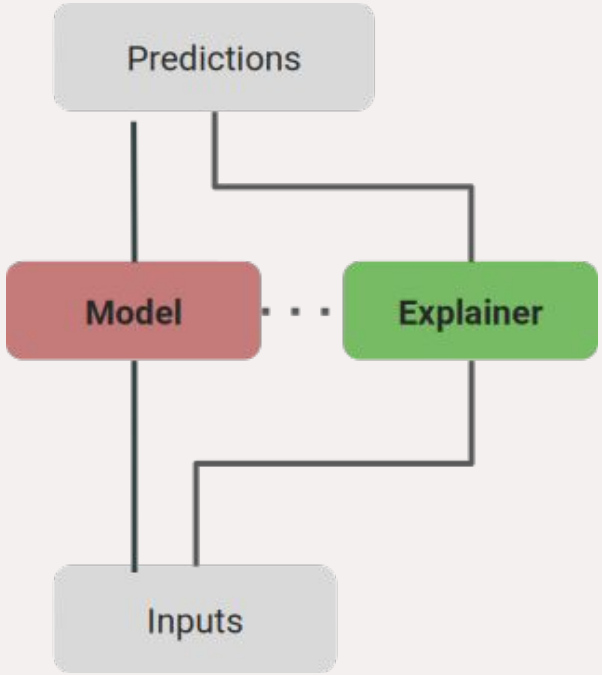


- Gradient and inversion-based methods
- Occlusion methods (e.g. RISE)
- Surrogate methods (e.g. LIME)

Transparency - AI Interpretability

Issue: Produced explanations are not always faithful

Post-Hoc Attribution (XAI)



- Gradient and inversion-based methods
- Occlusion methods (e.g. RISE)
- Surrogate methods (e.g. LIME)

ARTICLE | FREE ACCESS

Sanity checks for saliency maps

Authors: Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim | Authors Info & Claims

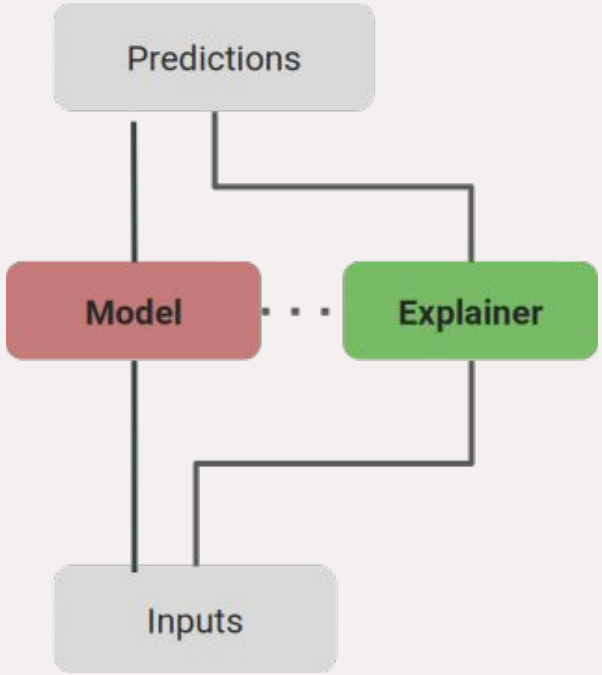
NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems
Pages 9525 - 9536

Published: 03 December 2018 Publication History

Transparency - AI Interpretability

Design models that are interpretable in first place

Post-Hoc Attribution (XAI)



Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

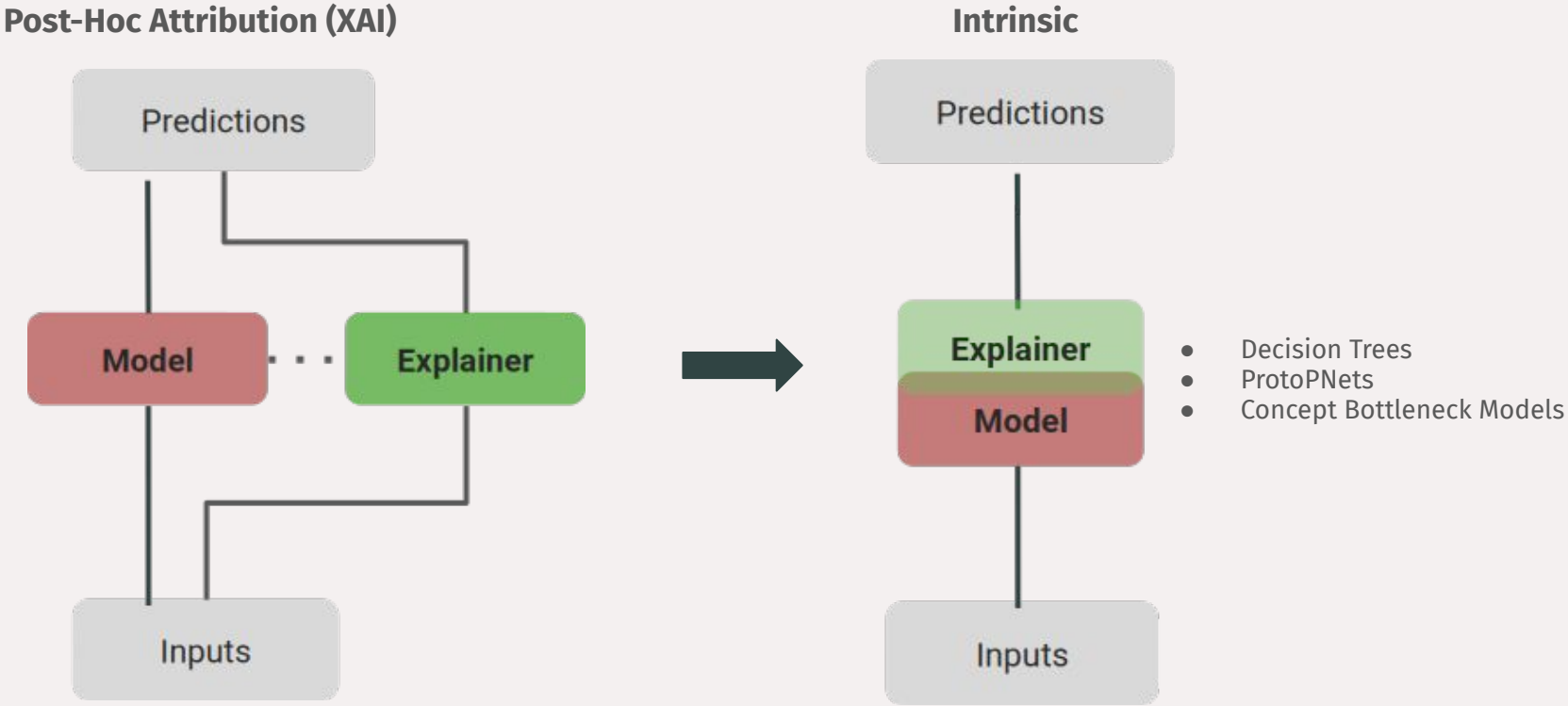
Cynthia Rudin

Nature Machine Intelligence 1, 206–215 (2019) | Cite this article

- XAI provides a false sense of security
- Not every problem requires a complex model

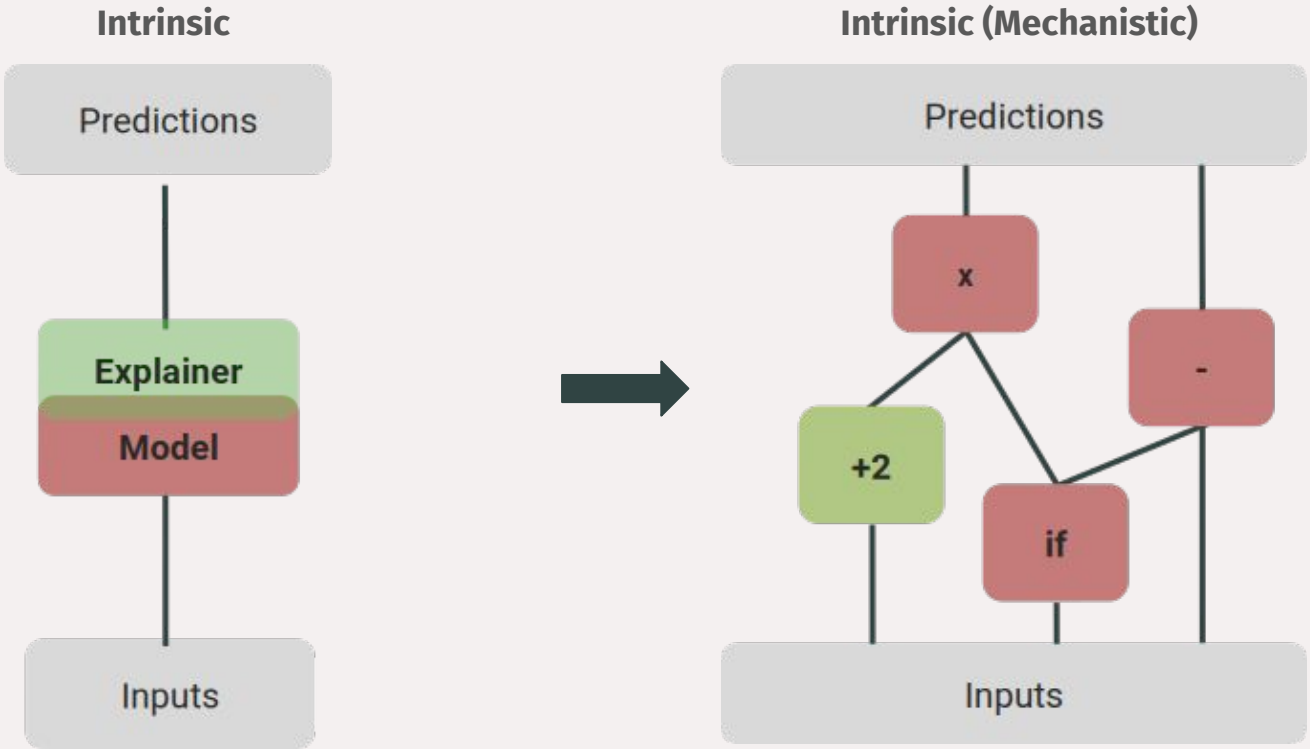
Transparency - AI Interpretability

Integrate explanation capabilities at design time



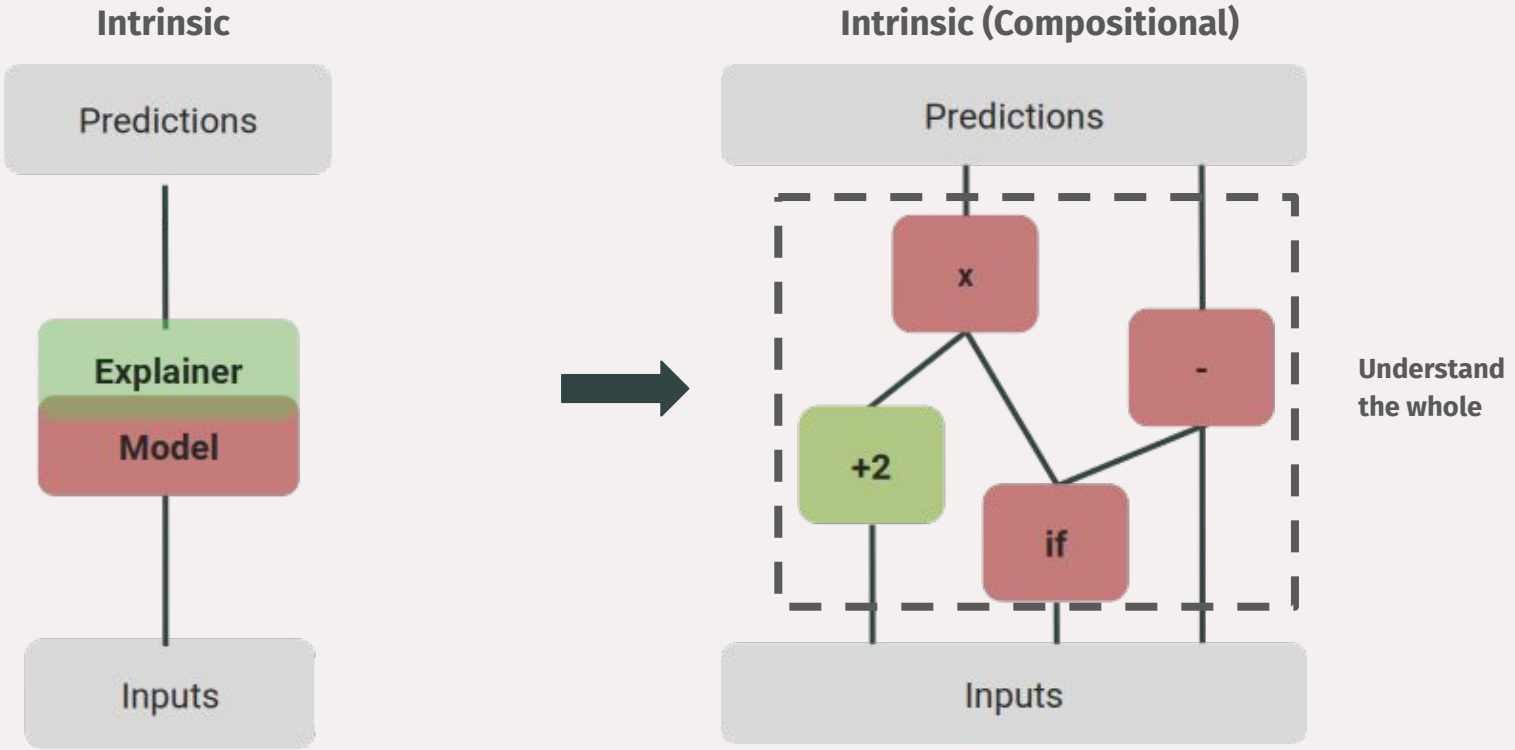
Transparency - AI Interpretability

Going from feature occurrence to understanding internal processes



Transparency - AI Interpretability

From internal processes to understanding the whole



Coming up...

Early Interpretability Methods

José Oramas

From Classical to Mechanistic Interpretability

Ward Gauderis, José Oramas & Thomas Dooms

From Mechanistic to Compositional Interpretability

Ward Gauderis & Thomas Dooms