

# LOGIC-BASED EXPLAINABLE ARTIFICIAL INTELLIGENCE

---

Joao Marques-Silva

ICREA & Univ. Lleida, Catalunya, Spain

ESSLLI, Bochum, Germany, July 2025

# Brief bio sketch...

- Research:
  - Boolean satisfiability (SAT) – clause learning used in CDCL
  - Optimization & inconsistency
  - Quantification, model checking, uses of CEGAR
  - Logic encodings, description logics
  - Model-based diagnosis
  - ...
  - Rigorous explainability – new topic since 2019
- Career:
  - Research Professor – ICREA & Univ. Lleida, Catalunya, [Spain](#)
  - Directeur de Recherche – CNRS & IRIT, Toulouse, [France](#)
  - SFI Stokes Prof. of CS – Univ. College Dublin, [Ireland](#)
  - Full Professor of CS – Univ. Southampton, [United Kingdom](#)
  - Full Professor of CS – Univ. Lisbon – FCUL, [Portugal](#)
  - Full Professor of CS – Univ. Lisbon – IST, [Portugal](#)
  - ...
  - PhD at Univ. Michigan, Ann Arbor, MI, [USA](#) (1995)

# Lecture 01

# Remarkable ML successes – e.g. LLM ChatGPT for knowledge

Explain Sudoku puzzles. Give me an example.  
Thanks!



## What is Sudoku?

Sudoku is a logic-based number placement puzzle. It consists of a 9x9 grid divided into nine 3x3 subgrids, also called "regions." The goal is to fill the grid with numbers from 1 to 9 such that:

1. Each row contains the numbers 1 through 9, without repetition.
2. Each column contains the numbers 1 through 9, without repetition.
3. Each 3x3 subgrid contains the numbers 1 through 9, without repetition.

The puzzle starts with some numbers pre-filled, and your task is to fill in the blanks following the rules above. The difficulty of the puzzle depends on how many numbers are pre-filled and their arrangement.

## Example Puzzle

Here's an example of a Sudoku puzzle:

5	3		7					
6			1	9	5			
	9	8				6		
8			6			3		
4			8	3		1		
7			2			6		
6			2	8				
			4	1	9	5		
			8			7	9	

Copy

## How to Solve It

1. **Start with easy numbers:** Look for rows, columns, or subgrids where most numbers are already filled in. Use the process of elimination to determine the missing numbers.
2. **Use pencil marks:** If you're unsure, pencil in all possible candidates for a blank cell and narrow them down as you progress.
3. **Keep the rules in mind:** Continuously check the rows, columns, and subgrids to ensure you don't duplicate any numbers.
4. **Look for patterns:** Advanced strategies like "naked pairs" and "hidden singles" may help solve more challenging puzzles.

Would you like help solving this example or generating a new one?



# Remarkable ML successes – e.g. LLM ChatGPT for programming

code for selecting columns of pandas  
dataframe matching some value

To select columns in a pandas DataFrame that match a specific value, you can use a combination of boolean indexing and `.any()` or `.all()` methods. Here's a simple example:

## Example Code

Let's assume you want to select all columns containing a specific value (say, 42).

```
python Copy  
  
import pandas as pd  
  
# Example DataFrame  
data = {  
    'A': [10, 42, 10],  
    'B': [5, 5, 42],  
    'C': [42, 1, 42],  
    'D': [2, 2, 2]  
}  
df = pd.DataFrame(data)  
  
# Specify the value to match  
value_to_match = 42  
  
# Selecting columns that contain the specified value  
matching_columns = df.columns[(df == value_to_match).  
  
# Filter the DataFrame to include only matching column  
df_matching = df[matching_columns]  
  
print(df_matching)
```

## Explanation

1. `(df == value_to_match)` creates a DataFrame of booleans where `True` indicates cells that match the specified value.
2. `.any()` is used on columns (default `axis=0`) to determine if any cell in each column matches the value.
3. `df.columns[...]` is used to filter the columns to include only those with matching values.
4. `df[matching_columns]` creates a new DataFrame with only the matching columns.

## Output

For the example given above, if `value_to_match` is 42, the output will be:

Copy

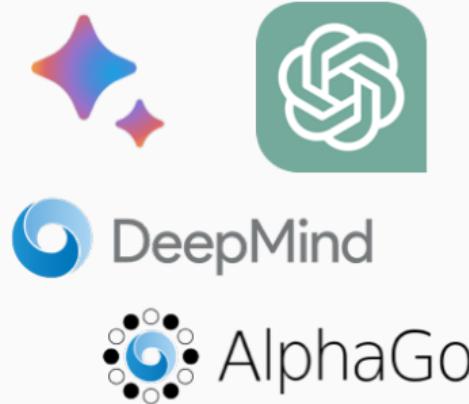
	A	B	C
0	10	5	42
1	42	5	1
2	10	42	42

Only columns A, B, and C are returned, as they contain the value 42.

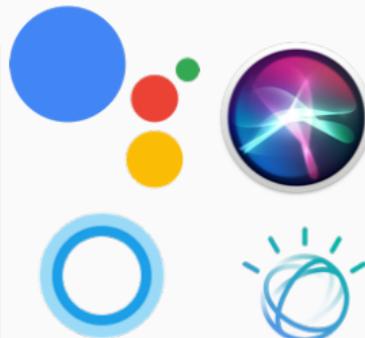
# Recent & ongoing ML successes



<https://en.wikipedia.org/wiki/Waymo>

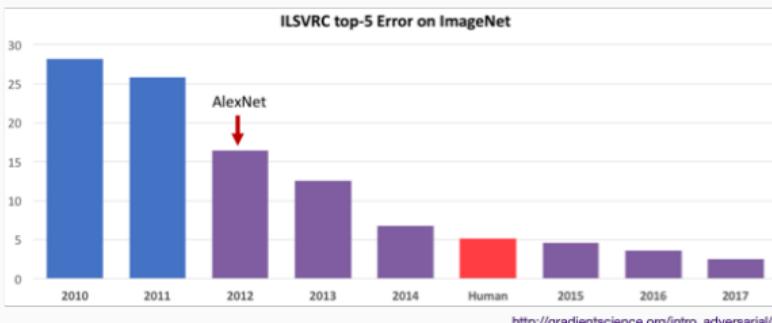


AlphaGo Zero & Alpha Zero



[https://fr.wikipedia.org/wiki/Pepper\\_\(robot\)](https://fr.wikipedia.org/wiki/Pepper_(robot))

## Image & Speech Recognition



## Troubling ML issues – e.g. LLM ChatGPT for puzzles (Jan'25)

Count the number of occurrences of the letter 'L' in the word 'LOLLAPALOOZA'

To count the occurrences of the letter 'L' in the word **LOLLAPALOOZA**, we can visually inspect it:

The word is: **LOLLAPALOOZA**.

Now, let's count the 'L's:

- There are **3 L's** in the word.

**Answer:**

The number of occurrences of the letter 'L' is **3**.

<https://arxiv.org/html/2405.19616v2>

# Troubling ML issues – e.g. LLM ChatGPT for puzzles (Jan'25)

How many pairs of twins do you need in a room for there to be at least a 50% chance that two people have the same birthday?

This question is a variation of the famous **Birthday Problem**, where the goal is to determine how many people are needed in a room for at least a 50% chance of two sharing the same birthday.

Here, you're dealing with **pairs of twins**. A pair of twins share the same birthday, so the effective number of unique birthdays to consider is reduced compared to regular individuals.

**Step-by-Step Solution:**

- 1. Revisiting the Regular Birthday Problem:** For a 50% chance of two people sharing a birthday in a room of unrelated individuals, you need about **23 people**. This assumes 365 equally likely days for a birthday.
- 2. Impact of Twins:**
  - Each pair of twins reduces the diversity of birthdays because both members of the pair share a single birthday.
  - Therefore, each pair of twins counts as **1 effective individual** in terms of the Birthday Problem.
- 3. Calculation for Pairs of Twins:**
  - For a 50% chance of a shared birthday, you need **23 effective individuals**.
  - Since each pair of twins is equivalent to **1 effective individual**, you need **23 pairs of twins** to achieve the same probability.

**Final Answer:**

You need **23 pairs of twins** in the room for at least a 50% chance that two people have the same birthday.

↓

<https://arxiv.org/html/2405.19616v2>

# Can we trust ML models?

- Accuracy in training/test data
- Complex ML models are **brittle**
  - Extensive work on finding adversarial examples
  - Extensive work on learning robust ML models
- More recently, complex ML models **hallucinate**
  - LLMs also fail at basic reasoning tasks
- Need to validate operation of ML model, with rigor
  - Explanations; robustness; verification

## ML models are brittle – adversarial examples



## ML models are brittle – adversarial examples



## ML models are brittle – adversarial examples



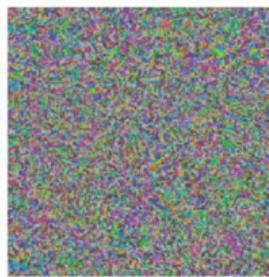
# Adversarial examples can be very problematic

Original image



+ 0.04 ×

Adversarial noise



Adversarial example



Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.



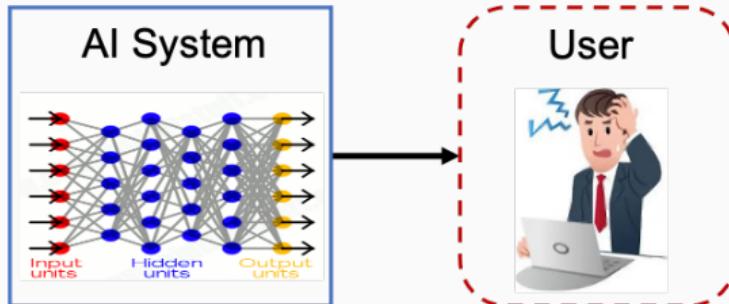
Perturbation computed by a common adversarial attack technique.

Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.



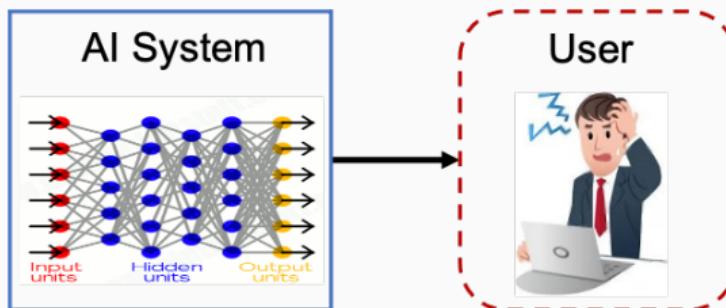
Finlayson et al., Nature 2019

# The road to trust – eXplainable AI (XAI)



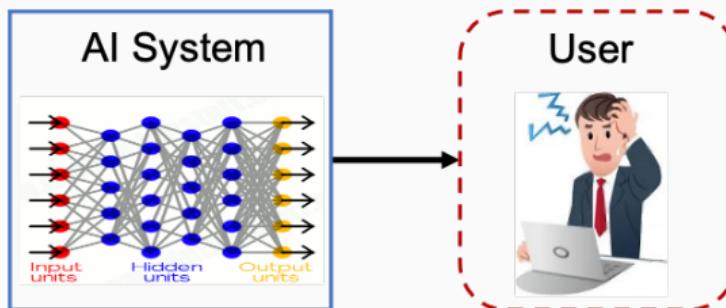
- Complex ML models are **opaque**
- Goal of XAI: **to help humans understand ML models**

# The road to trust – eXplainable AI (XAI)



- Complex ML models are **opaque**
- Goal of XAI: **to help humans understand ML models**
- How to XAI?
  - Answer to **Why (the prediction)?** questions
  - Answer to **Why Not (some other prediction)?** questions

# The road to trust – eXplainable AI (XAI)



- Complex ML models are **opaque**
- Goal of XAI: **to help humans understand ML models**
- How to XAI?
  - Answer to **Why (the prediction)?** questions
  - Answer to **Why Not (some other prediction)?** questions
- What about **rigor**?
  - Sub-symbolic (i.e. non-symbolic) methods?
  - Symbolic (i.e. logic-based) methods?

# Importance of XAI

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making  
and a “right to explanation”

Bryce Goodman,<sup>1\*</sup> Seth Flaxman,<sup>2</sup>

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE  
(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION  
LEGISLATIVE ACTS

## Explainable Artificial Intelligence (XAI)



David Gunning  
DARPA/I2O  
Program Update November 2017



© DARPA

European Commission > Strategy > Digital Single Market > Reports and studies >

Digital Single Market

REPORT / STUDY | 8 April 2019

Ethics guidelines for trustworthy AI

# Importance of XAI

European Union regulation  
and a “right to explanation”

Bryce Goodman

■ We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.

REGULATION (EU) 2016/679

In order to trust deployed AI systems, on the part of the user and the provider, we must not only improve their robustness,<sup>5</sup> but also develop ways to make their reasoning intelligible. Intelligibility will help us spot AI that makes mistakes due to distributional drift or incomplete representations of goals and features. Intelligibility will also facilitate control by humans in increasingly common collaborative human/AI teams. Furthermore, intelligibility will help humans learn from AI. Finally, there are legal reasons to want intelligible AI, including the European GDPR and a growing need to assign liability when AI errs.

Weld & Bansal, CACM, Jun'19  
Due November 2017

THE COUNCIL

(Data protection and on the free movement of data in the internal market and on the free movement of electronic communications (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation))

Proposal for a

EUROPEAN PARLIAMENT AND OF THE COUNCIL

RULES ON ARTIFICIAL INTELLIGENCE (AMENDING DIRECTIVE 2014/52/EU AND AMENDING CERTAIN UNION LEGISLATIVE ACTS)

European Commission > Strategy > Digital Single Market > Reports and studies >

Digital Single Market

REPORT / STUDY | 8 April 2019

Ethics guidelines for trustworthy AI



© DARPA



Search

European Commission > Strategy > Digital Single Market > Reports and studies >

Digital Single Market

REPORT / STUDY | 8 April 2019

## Ethics guidelines for trustworthy AI

Following the publication of the draft ethics guidelines in December 2018 to which more than 500 comments were received, the independent expert group presents today their ethics guidelines for trustworthy artificial intelligence.

About Artificial intelligence

Blog posts

News

# XAI & the principle of explicability



European Commission > Strategy > Digital Single Market > Reports and

## Digital Single Market

### REPORT / LIST

#### The principle of explicability

- **Explicability** is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.<sup>33</sup>

About Artificial  
intelligence

Blog posts

News

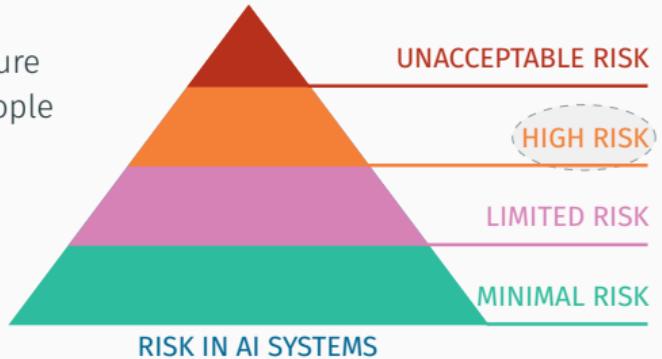
& thousands of recent papers!

# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people
- ...

[EU21b, EU21a]



# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and processing
- ...

otherwise incorrect or unjust manner. Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, explainable and documented.

[21b, EU21a]

HIGH RISK

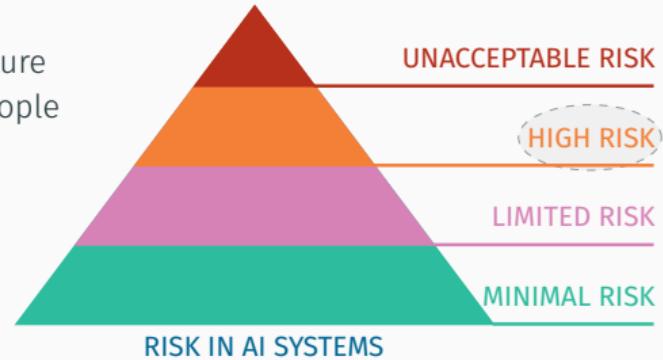


# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people
- ...

[EU21b, EU21a]



- And **safety-critical**:

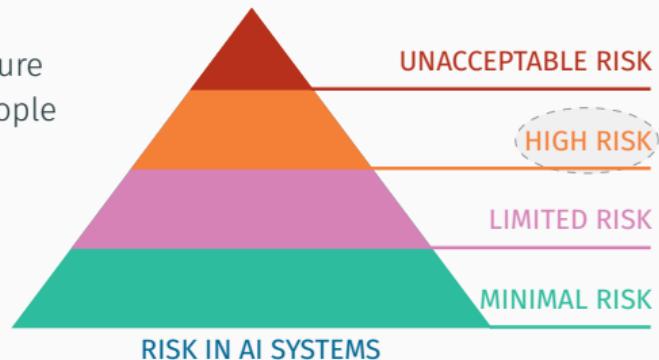
- Self-driving cars
- Autonomous vehicles
- Autonomous aerial devices
- ...

# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people
- ...

[EU21b, EU21a]



- And **safety-critical**:

- Self-driving cars
- Autonomous vehicles
- Autonomous aerial devices
- ...

PERSPECTIVE  
<https://doi.org/10.1038/s42256-019-0048-x>

nature  
machine intelligence

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

Cynthia Rudin

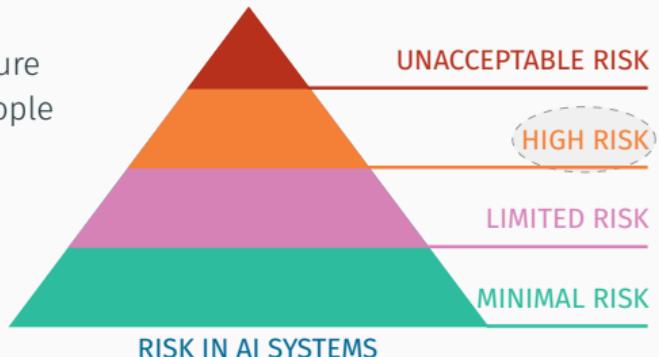
May 2019

# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people
- ...

[EU21b, EU21a]



- And **safety-critical**:

- Self-driving cars
- Autonomous vehicles
- Autonomous aerial devices
- ...

- **Correctness of explanations is paramount!**

- To build trust
- To help debug AI systems
- To prevent (catastrophic) accidents
- ...

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0048-x>

nature  
machine intelligence

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

Cynthia Rudin

May 2019

# XAI for high-risk & safety-critical applications

- **High-risk** (EU regulations):

- Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people
- ...

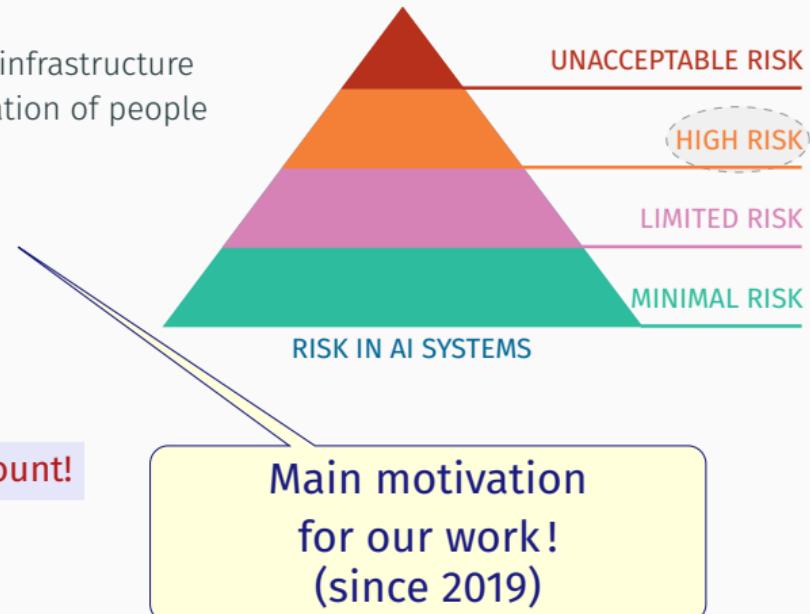
[EU21b, EU21a]

- And **safety-critical**:

- Self-driving cars
- Autonomous vehicles
- Autonomous aerial devices
- ...

- **Correctness of explanations is paramount!**

- To build trust
- To help debug AI systems
- To prevent (catastrophic) accidents
- ...



# Can we trust (non-symbolic) XAI? – some questions

- Many proposed **solutions** for XAI
  - Most, and the better-known, are heuristic
  - I.e. no guarantees of rigor
- Many proposed **uses** of XAI
- Regular complaints about issues with existing (heuristic) methods of XAI

# Can we trust (non-symbolic) XAI? – some questions

- Many proposed **solutions** for XAI
  - Most, and the better-known, are heuristic
  - I.e. no guarantees of rigor
- Many proposed **uses** of XAI
- Regular complaints about issues with existing (heuristic) methods of XAI
- **Q:** Can heuristic XAI be trusted in high-risk and/or safety-critical domains?
- **Q:** Can we validate results of heuristic XAI?

# What have we been up to? 1. Created the field of symbolic (formal) XAI – I

[MI22, Mar22, MS23, Mar24]

- Rigorous, logic-based, definitions of explanations
  - Relationship with abduction – abductive explanations (AXps)
  - Contrastive explanations (CXps) [Mil19]
- Duality between AXps & CXps
  - AXps are MHSes of CXps and vice-versa

# What have we been up to? 1. Created the field of symbolic (formal) XAI – I

[MI22, Mar22, MS23, Mar24]

- Rigorous, logic-based, definitions of explanations
  - Relationship with abduction – abductive explanations (AXps)
  - Contrastive explanations (CXps) [Mil19]
- Duality between AXps & CXps
  - AXps are MHSes of CXps and vice-versa
- Tractability results
  - Devised efficient poly-time algorithms

# What have we been up to? 1. Created the field of symbolic (formal) XAI – I

[MI22, Mar22, MS23, Mar24]

- Rigorous, logic-based, definitions of explanations
  - Relationship with abduction – abductive explanations (AXps)
  - Contrastive explanations (CXps) [Mil19]
- Duality between AXps & CXps
  - AXps are MHSes of CXps and vice-versa
- Tractability results
  - Devised efficient poly-time algorithms
- Intractability results
  - Devised efficient methods
  - Links with automated reasoners

# What have we been up to? 1. Created the field of symbolic (formal) XAI – I

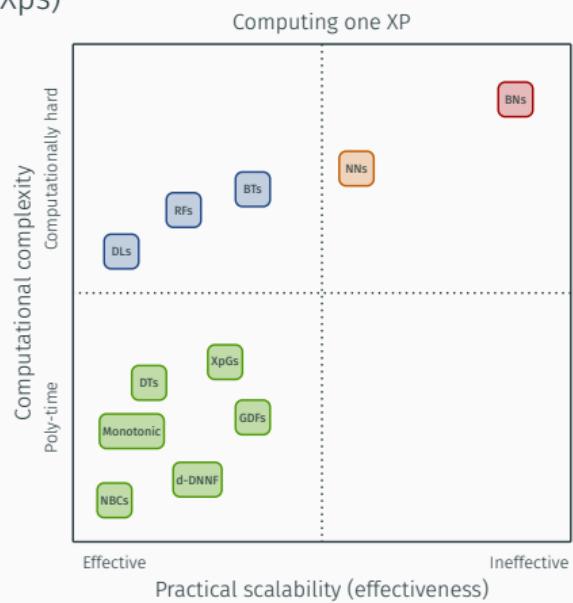
[MI22, Mar22, MS23, Mar24]

- Rigorous, logic-based, definitions of explanations
  - Relationship with abduction – abductive explanations (AXps)
  - Contrastive explanations (CXps) [Mil19]
- Duality between AXps & CXps
  - AXps are MHSes of CXps and vice-versa
- Tractability results
  - Devised efficient poly-time algorithms
- Intractability results
  - Devised efficient methods
  - Links with automated reasoners
- Wealth of computational problems related with AXps/CXps

# What have we been up to? 1. Created the field of symbolic (formal) XAI – I

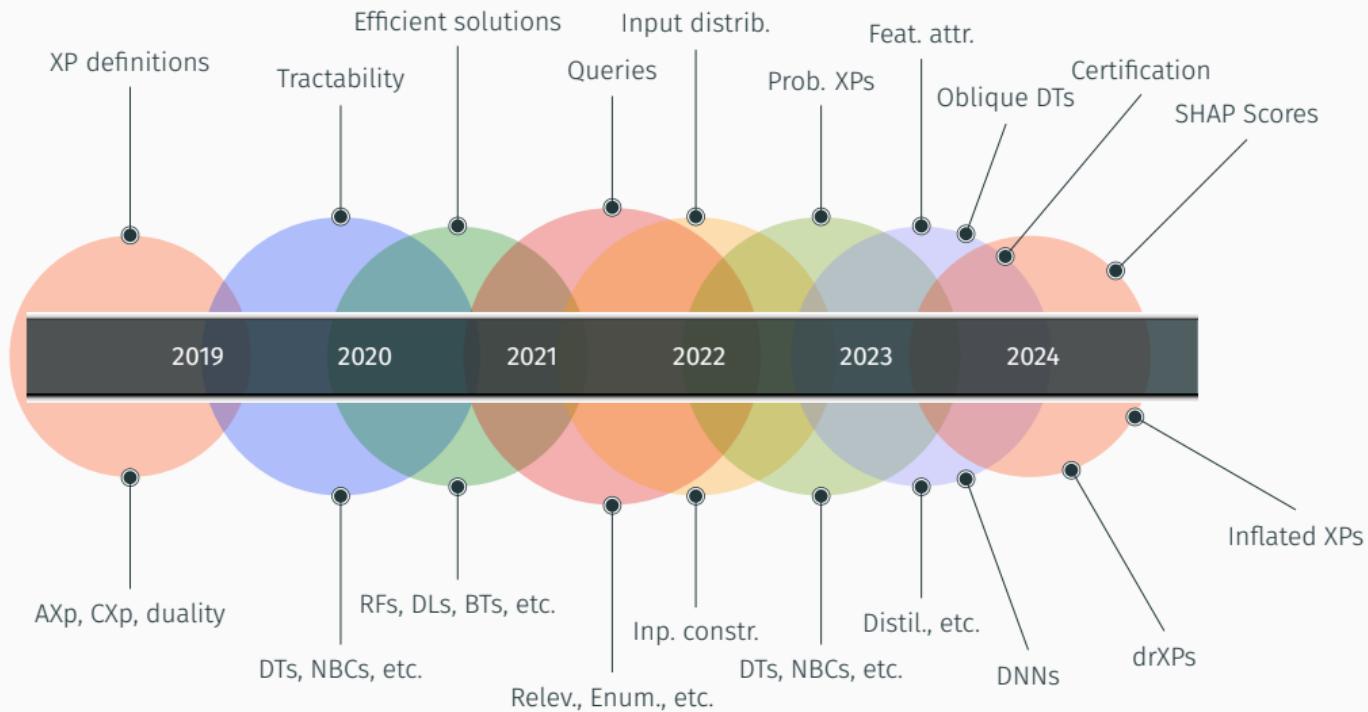
[MI22, Mar22, MS23, Mar24]

- Rigorous, logic-based, definitions of explanations
  - Relationship with abduction – abductive explanations (AXps)
  - Contrastive explanations (CXps) [Mil19]
- Duality between AXps & CXps
  - AXps are MHSes of CXps and vice-versa
- Tractability results
  - Devised efficient poly-time algorithms
- Intractability results
  - Devised efficient methods
  - Links with automated reasoners
- Wealth of computational problems related with AXps/CXps



# What have we been up to? 1. Created the field of symbolic (formal) XAI – II

[MI22, Mar22, MS23, Mar24]



# What have we been up to? 2. Uncovered key myths of non-symbolic XAI – I

[RSG16, LL17, RSG18, Rud19]

## LIME “Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro  
University of Washington  
Seattle, WA 98105, USA  
marcotcr@cs.uw.edu

Sameer Singh  
University of Washington  
Seattle, WA 98105, USA  
sameer@cs.uw.edu

Carlos Guestrin  
University of Washington  
Seattle, WA 98105, USA  
guestrin@cs.uw.edu

## PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0048-x>

nature  
machine intelligence

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

Intrinsic Interpretability

Cynthia Rudin



## A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg  
Paul G. Allen School of Computer Science  
University of Washington  
Seattle, WA 98105  
slundi@cs.washington.edu

Su-In Lee  
Paul G. Allen School of Computer Science  
Department of Genome Sciences  
University of Washington  
Seattle, WA 98105  
suinlee@cs.washington.edu



## Anchors: High-Precision Model-Agnostic Explanations

Marco Tulio Ribeiro  
University of Washington  
marcotcr@cs.washington.edu

Sameer Singh  
University of California, Irvine  
sameer@uci.edu

Carlos Guestrin  
University of Washington  
guestrin@cs.washington.edu



Anchor

[MSH24, HMS24, HM23]

# research and advances



DOI:10.1145/3635301

**When the decisions of ML models impact people, one should expect explanations to offer the strongest guarantees of rigor. However, the most popular XAI approaches offer none.**

BY JOAO MARQUES-SILVA AND XUANXIANG HUANG

# Explainability Is Not a Game

## » key insights

- Shapley values find extensive uses in explaining machine learning models and serve to assign importance to the features of the model.
- Shapley values for explainability also find ever-increasing uses in high-risk and safety-critical domains, for example, medical diagnosis.
- This article proves that the existing definition of Shapley values for explainability can produce misleading information regarding feature importance, and so can induce human decision makers in error.

# Plan for this course

- Lecture 01 – unit(s):
  - #01: Foundations
- Lecture 02 – unit(s):
  - #02: Principles of symbolic XAI – **feature selection**
  - #03: Tractability in symbolic XAI (& myth of interpretability)
- Lecture 03 – unit(s):
  - #04: Intractability in symbolic XAI (& myth of model-agnostic XAI)
  - #05: Explainability queries
- Lecture 04 – unit(s):
  - #06: Recent, emerging & advanced topics
- Lecture 05 – unit(s):
  - #07: Principles of symbolic XAI – **feature attribution** (& myth of Shapley values in XAI)
  - #08: Corrected feature attribution – nuSHAP
  - #09: Conclusions & research directions

Unit #01

Foundations

## Classification problems

- Set of features  $\mathcal{F} = \{1, 2, \dots, m\}$ , each feature  $i$  taking values from domain  $D_i$ 
  - Features can be categorical, discrete or real-valued
  - Feature space:  $\mathbb{F} = \Pi_{i=1}^m D_i$
- Set of classes  $\mathcal{K} = \{c_1, \dots, c_K\}$

# Classification problems

- Set of features  $\mathcal{F} = \{1, 2, \dots, m\}$ , each feature  $i$  taking values from domain  $D_i$ 
  - Features can be categorical, discrete or real-valued
  - Feature space:  $\mathbb{F} = \Pi_{i=1}^m D_i$
- Set of classes  $\mathcal{K} = \{c_1, \dots, c_K\}$
- ML model  $\mathcal{M}_C$  computes a (non-constant) classification function  $\kappa : \mathbb{F} \rightarrow \mathcal{K}$ 
  - $\mathcal{M}_C$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathcal{K}, \kappa)$

# Classification problems

- Set of features  $\mathcal{F} = \{1, 2, \dots, m\}$ , each feature  $i$  taking values from domain  $D_i$ 
  - Features can be categorical, discrete or real-valued
  - Feature space:  $\mathbb{F} = \prod_{i=1}^m D_i$
- Set of classes  $\mathcal{K} = \{c_1, \dots, c_K\}$
- ML model  $\mathcal{M}_C$  computes a (non-constant) classification function  $\kappa : \mathbb{F} \rightarrow \mathcal{K}$ 
  - $\mathcal{M}_C$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathcal{K}, \kappa)$
- Instance  $(\mathbf{v}, c)$  for point  $\mathbf{v} = (v_1, \dots, v_m) \in \mathbb{F}$ , with prediction  $c = \kappa(\mathbf{v})$ ,  $c \in \mathcal{K}$ 
  - Goal: to compute explanations for  $(\mathbf{v}, c)$

# Regression problems

- For regression problems:
  - Codomain:  $\mathbb{V}$
  - Regression function:  $\rho : \mathbb{F} \rightarrow \mathbb{V}$  (non-constant)
  - ML model:  $\mathcal{M}_R$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathbb{V}, \rho)$

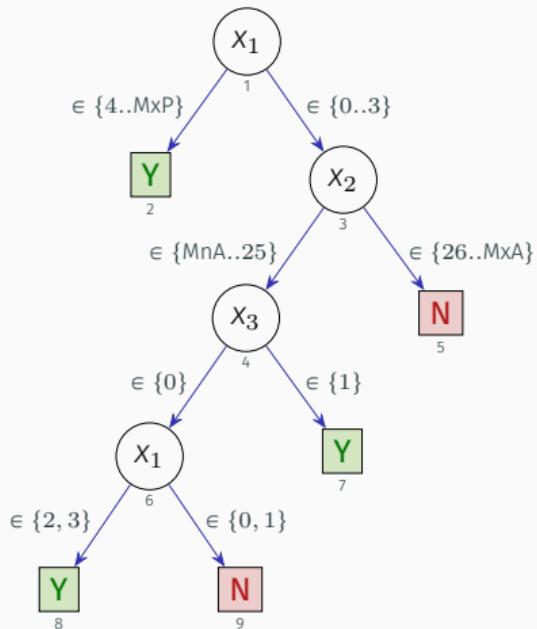
# Regression problems

- For regression problems:
  - Codomain:  $\mathbb{V}$
  - Regression function:  $\rho : \mathbb{F} \rightarrow \mathbb{V}$  (non-constant)
  - ML model:  $\mathcal{M}_R$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathbb{V}, \rho)$
- General ML model:
  - $\mathbb{T}$ : range of possible predictions
  - Non-constant function  $\tau : \mathbb{F} \rightarrow \mathbb{T}$
  - ML model:  $\mathcal{M}$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathbb{T}, \tau)$

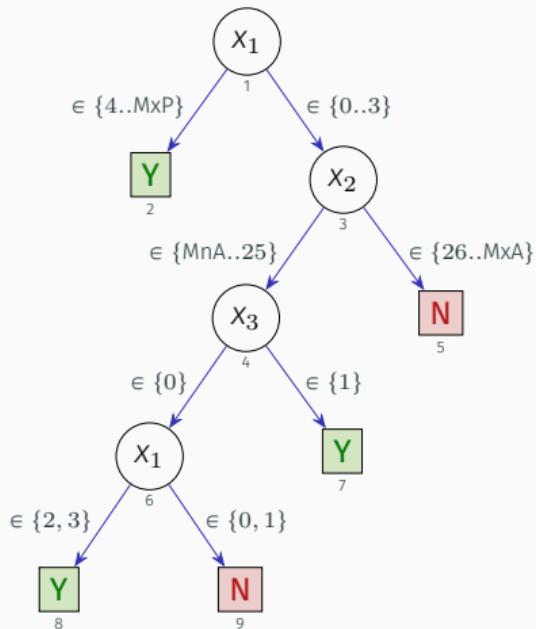
# Regression problems

- For regression problems:
  - Codomain:  $\mathbb{V}$
  - Regression function:  $\rho : \mathbb{F} \rightarrow \mathbb{V}$  (non-constant)
  - ML model:  $\mathcal{M}_R$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathbb{V}, \rho)$
- General ML model:
  - $\mathbb{T}$ : range of possible predictions
  - Non-constant function  $\tau : \mathbb{F} \rightarrow \mathbb{T}$
  - ML model:  $\mathcal{M}$  is a tuple  $(\mathcal{F}, \mathbb{F}, \mathbb{T}, \tau)$
- Instance:  $(\mathbf{v}, q), q \in \mathbb{T}$

## Example ML models – classification – decision trees (DTs)

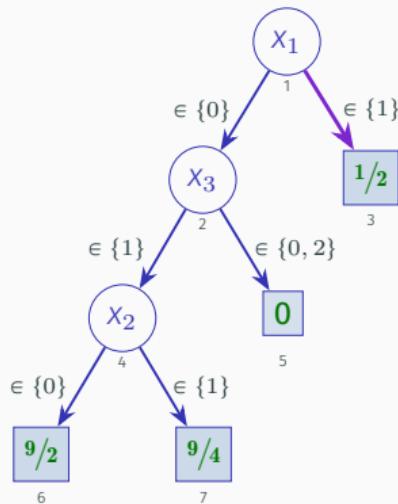


## Example ML models – classification – decision trees (DTs)



- Literals in DTs can use  $=$  or  $\in$

## Example ML models – regression – regression trees (RTs)



- Literals in RTs can use  $=$  or  $\in$

## Example ML models – classification – rules

- Ordered rules – decision lists (DLs):

```
IF       $x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_2 \vee x_3$  THEN predict N  
ELSE                THEN predict Y  
 $\mathcal{F} = \{1, 2, 3\}; \mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}_3 = \{0, 1\}; \mathcal{K} = \{Y, N\}$ 
```

## Example ML models – classification – rules

- Ordered rules – decision lists (DLs):

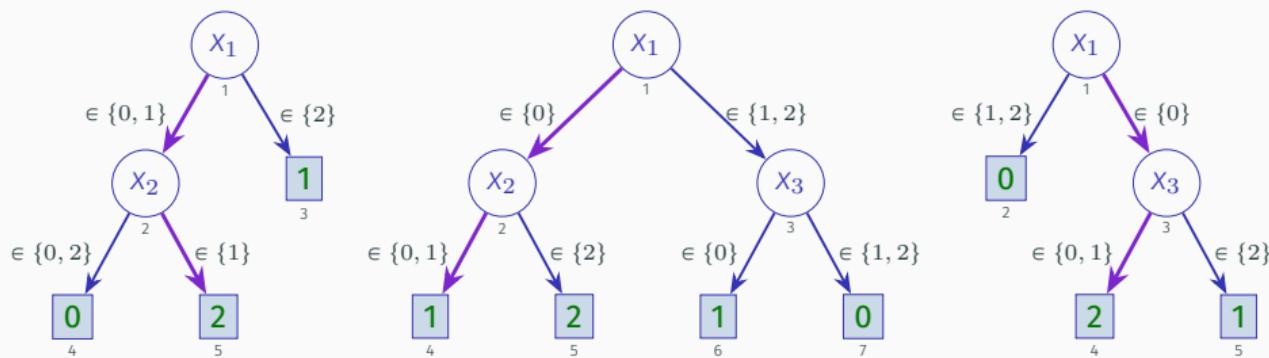
$$\begin{array}{lll} \text{IF } & x_1 \wedge x_2 & \text{THEN predict Y} \\ \text{ELSE IF } & \neg x_2 \vee x_3 & \text{THEN predict N} \\ \text{ELSE } & & \text{THEN predict Y} \\ \mathcal{F} = \{1, 2, 3\}; \mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}_3 = \{0, 1\}; \mathcal{K} = \{\text{Y}, \text{N}\} \end{array}$$

- Unordered rules – decision sets (DSs):

$$\begin{array}{lll} \text{IF } & x_1 + x_2 \geq 0 & \text{THEN predict } \blacksquare \\ \text{IF } & x_1 + x_2 < 0 & \text{THEN predict } \blacksquare \\ \mathcal{F} = \{1, 2\}; \mathcal{D}_1 = \mathcal{D}_2 = \mathbb{R}; \mathcal{K} = \{\blacksquare, \blacksquare\} \end{array}$$

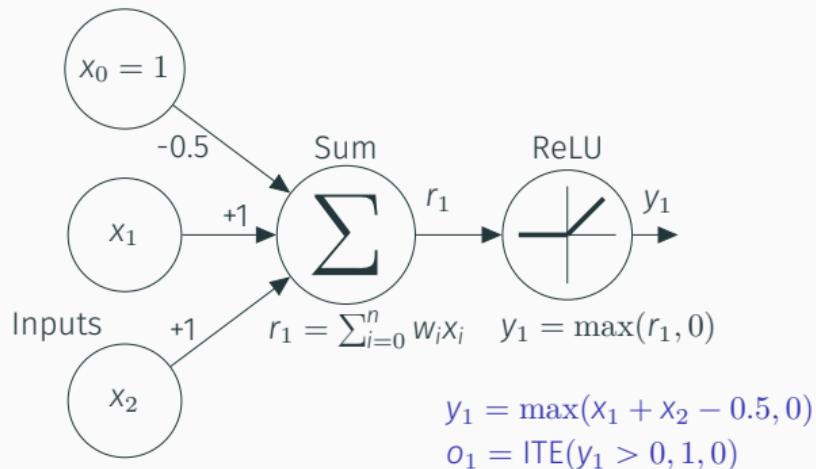
- Issues of DSs: overlap; incomplete coverage

## Example ML models – classification – random forests (RFs)



- For each input, each DT picks a class
- Result uses majority or weighted voting of the DTs

## Example ML models – classification – neural networks (NNs)



# Outline – Unit #01

ML Models: Classification & Regression Problems

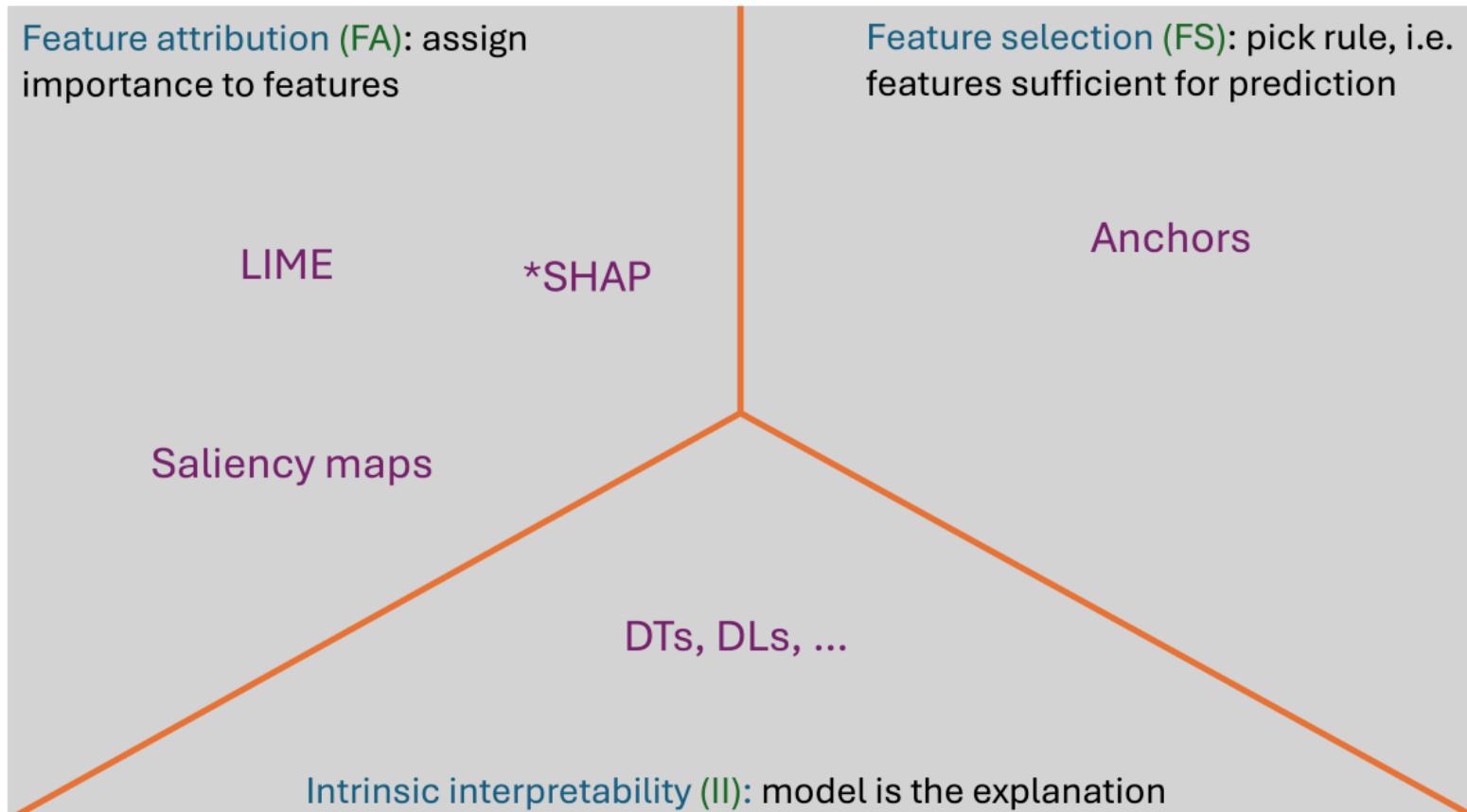
Basics of (non-symbolic) XAI

Motivation for Explanations

Brief Glimpse of Logic

Reasoning About ML Models

Understanding Intrinsic Interpretability

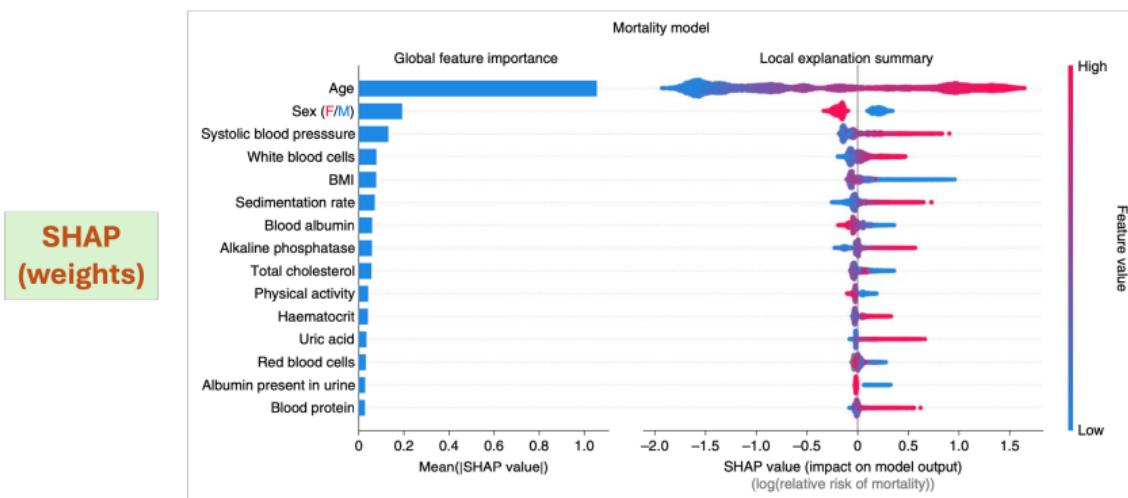


# Some non-symbolic XAI methods

[LL17, RSG18]

**IF Country = United-States AND Capital Loss = Low  
AND Race = White AND Relationship = Husband  
AND Married AND  $28 < \text{Age} \leq 37$   
AND Sex = Male AND High School grad  
AND Occupation = Blue-Collar  
THEN PREDICT Salary > \$50K**

**Anchors  
(rules)**



# Some non-symbolic XAI methods

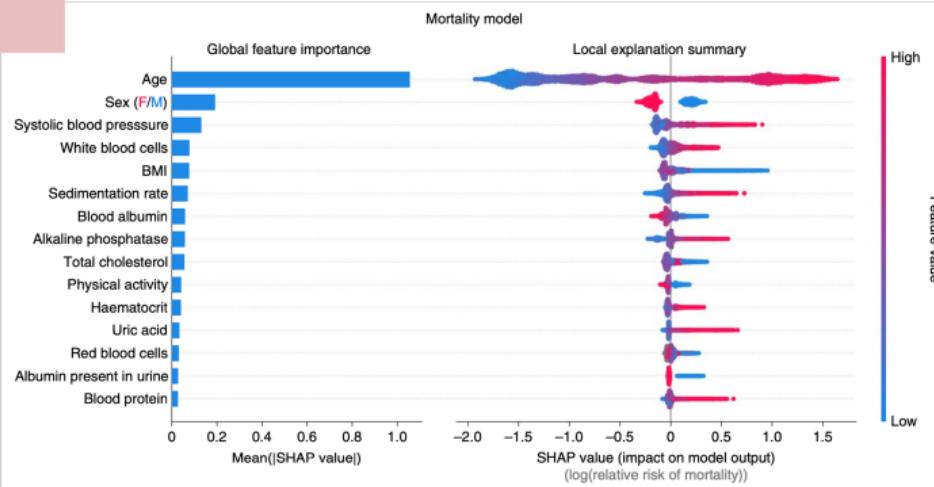
[LL17, RSG18]

**IF Country = United-States AND Capital Loss = Low  
AND Race = White AND Relationship = Husband  
AND Married AND  $28 < \text{Age} \leq 37$   
AND Sex = Male AND High School grad  
AND Occupation = Blue-Collar  
THEN PREDICT Salary > \$50K**

**Anchors  
(rules)**

model-agnostic, sample-based,  
& no guarantees of rigor...

**SHAP  
(weights)**



# Outline – Unit #01

ML Models: Classification & Regression Problems

Basics of (non-symbolic) XAI

Motivation for Explanations

Brief Glimpse of Logic

Reasoning About ML Models

Understanding Intrinsic Interpretability

## What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”

## What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF <COND> THEN  $\kappa(x) = c$

# What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF <COND> THEN  $\kappa(x) = c$

- **Explanation:** set of **literals** (or just **features**) in <COND>; **irreducibility matters!**
  - <COND> is **sufficient** for the prediction

# What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF   <COND>   THEN    $\kappa(x) = c$

- **Explanation:** set of **literals** (or just **features**) in **<COND>**; **irreducibility matters!**
  - **<COND>** is **sufficient** for the prediction
- **Obs:** rules are used in tools like Anchors
  - An **anchor** is a “high-precision rule”

[RSG16]

[RSG16]

# What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF <COND> THEN  $\kappa(x) = c$

- **Explanation:** set of **literals** (or just **features**) in <COND>; **irreducibility matters!**
  - <COND> is **sufficient** for the prediction
- **Obs:** rules are used in tools like Anchors
  - An **anchor** is a “high-precision rule”
- We seek a **rigorous** definition of rules for answering **Why?** questions such that,

[RSG16]

[RSG16]

# What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF <COND> THEN  $\kappa(x) = c$

- **Explanation:** set of **literals** (or just **features**) in <COND>; **irreducibility matters!**
  - <COND> is **sufficient** for the prediction
- **Obs:** rules are used in tools like Anchors
  - An **anchor** is a “high-precision rule”
- We seek a **rigorous** definition of rules for answering **Why?** questions such that,
  - <COND> is **sufficient** for the prediction
  - <COND> is **irreducible**

[RSG16]

[RSG16]

# What explanations do we seek? I.e. how to answer **Why?** questions?

- How to answer a **Why?** question? I.e. “ Why (the prediction)? ”
  - Our answer to a **Why?** question is a **rule**:

IF <COND> THEN  $\kappa(x) = c$

- **Explanation:** set of **literals** (or just **features**) in <COND>; **irreducibility matters!**
  - <COND> is **sufficient** for the prediction
- **Obs:** rules are used in tools like Anchors
  - An **anchor** is a “high-precision rule”
- We seek a **rigorous** definition of rules for answering **Why?** questions such that,
  - <COND> is **sufficient** for the prediction
  - <COND> is **irreducible**
- We also seek the algorithms for the rigorous computation of such rules

[RSG16]

[RSG16]

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

- Explanation for why  $\kappa(1, 1, 1, 1, 1) = \text{N}$ ?

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

- Explanation for **why**  $\kappa(1, 1, 1, 1, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_1 = 1) \wedge (x_4 = 1) \wedge (x_5 = 1)$  **THEN**  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_1 = 1, x_4 = 1, x_5 = 1\}$  suffice for DL to predict **N**

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

- Explanation for **why**  $\kappa(1, 1, 1, 1, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_1 = 1) \wedge (x_4 = 1) \wedge (x_5 = 1)$  **THEN**  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_1 = 1, x_4 = 1, x_5 = 1\}$  suffice for DL to predict **N**
- Explanation for **why**  $\kappa(1, 0, 0, 0, 0) = \text{Y}$ ?

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

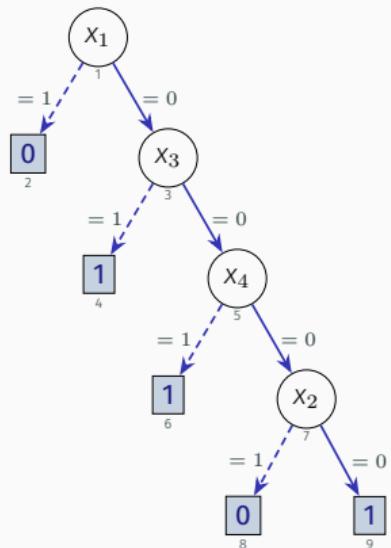
- Explanation for **why**  $\kappa(1, 1, 1, 1, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_1 = 1) \wedge (x_4 = 1) \wedge (x_5 = 1)$  **THEN**  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_1 = 1, x_4 = 1, x_5 = 1\}$  suffice for DL to predict **N**
- Explanation for **why**  $\kappa(1, 0, 0, 0, 0) = \text{Y}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_4 = 0)$  **THEN**  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_4 = 0\}$  suffices for DL to predict **Y**

## A decision list example

```
IF       $\neg x_1 \wedge x_2$  THEN predict Y  
ELSE IF  $\neg x_1 \wedge x_3$  THEN predict Y  
ELSE IF  $x_4 \wedge x_5$  THEN predict N  
ELSE                THEN predict Y
```

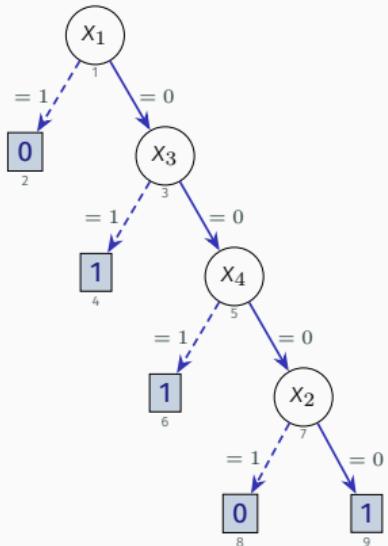
- Explanation for **why**  $\kappa(1, 1, 1, 1, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_1 = 1) \wedge (x_4 = 1) \wedge (x_5 = 1)$  **THEN**  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_1 = 1, x_4 = 1, x_5 = 1\}$  suffice for DL to predict **N**
- Explanation for **why**  $\kappa(1, 0, 0, 0, 0) = \text{Y}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_4 = 0)$  **THEN**  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_4 = 0\}$  suffices for DL to predict **Y**
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ ,  
**IF**  $(x_5 = 0)$  **THEN**  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_5 = 0\}$  also suffices for DL to predict **Y**

# A decision tree example



$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

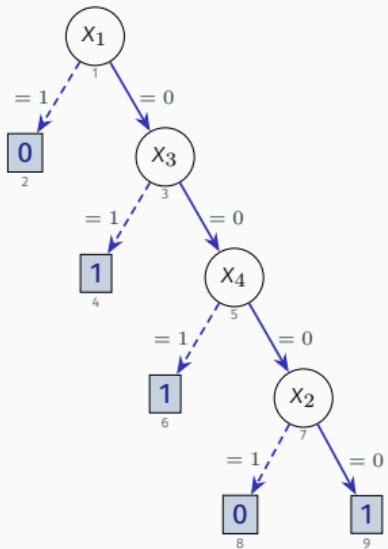
# A decision tree example



- Explanation for why  $\kappa(0, 0, 0, 0) = 1$ ?

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

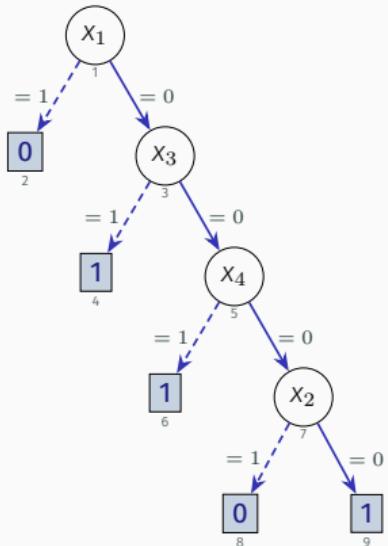
# A decision tree example



- Explanation for **why**  $\kappa(0, 0, 0, 0) = 1$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,
  - IF**  $(x_1 = 0) \wedge (x_2 = 0)$  **THEN**  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_1 = 0, x_2 = 0\}$  suffice for DT to predict 1

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

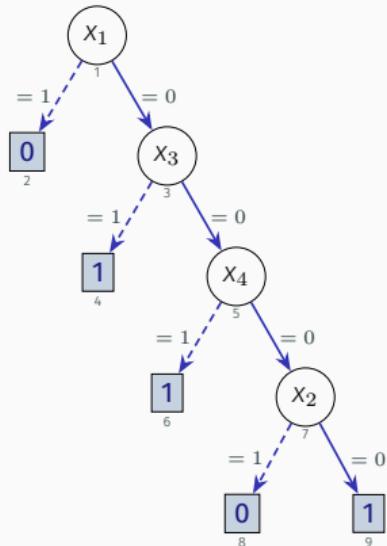
# A decision tree example



- Explanation for why  $\kappa(0, 0, 0, 0) = 1$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,
  - IF  $(x_1 = 0) \wedge (x_2 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_1 = 0, x_2 = 0\}$  suffice for DT to predict 1
- Explanation for why  $\kappa(1, 1, 1, 1) = 0$ ?

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	0
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	0

# A decision tree example

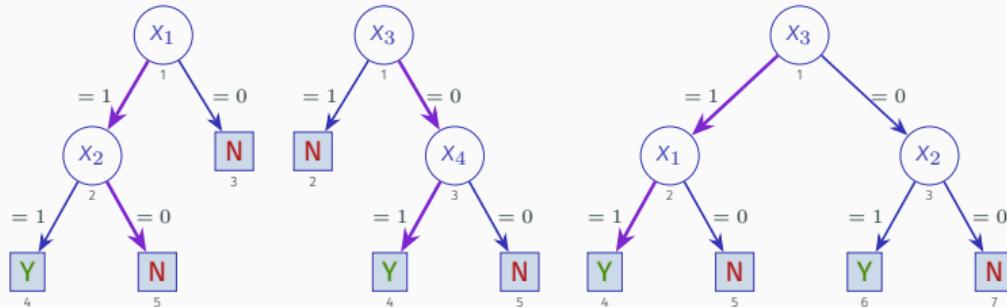


- Explanation for **why**  $\kappa(0, 0, 0, 0) = 1$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,  
**IF**  $(x_1 = 0) \wedge (x_2 = 0)$  **THEN**  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_1 = 0, x_2 = 0\}$  suffice for DT to predict **1**
- Explanation for **why**  $\kappa(1, 1, 1, 1) = 0$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,  
**IF**  $(x_1 = 1)$  **THEN**  $\kappa(\mathbf{x}) = 0$
  - i.e.  $\{x_1 = 1\}$  suffices for DT to predict **0**

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	<b>1</b>
0	0	0	1	<b>1</b>
0	0	1	0	<b>1</b>
0	0	1	1	<b>1</b>
0	1	0	0	<b>0</b>
0	1	0	1	<b>1</b>
0	1	1	0	<b>1</b>
0	1	1	1	<b>1</b>
1	0	0	0	<b>0</b>
1	0	0	1	<b>0</b>
1	0	1	0	<b>0</b>
1	0	1	1	<b>0</b>
1	1	0	0	<b>0</b>
1	1	0	1	<b>0</b>
1	1	1	0	<b>0</b>
1	1	1	1	<b>0</b>

# A random forest example

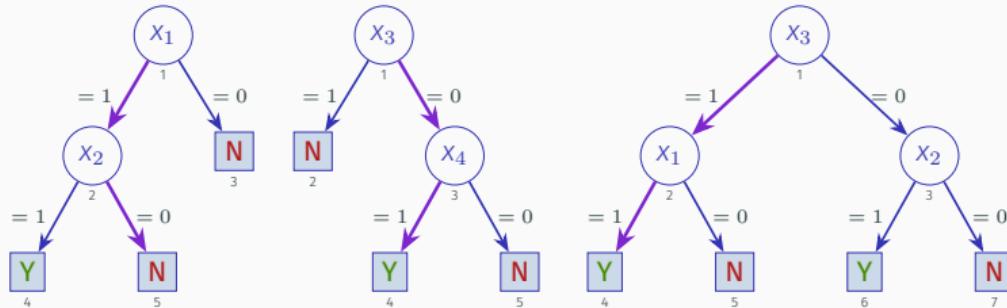
[IMS21]



X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	$\kappa(\mathbf{x})$
0	0	0	0	N	N	N	N
0	0	0	1	N	Y	N	N
0	0	1	0	N	N	N	N
0	0	1	1	N	N	N	N
0	1	0	0	N	N	Y	N
0	1	0	1	N	Y	Y	Y
0	1	1	0	N	N	N	N
0	1	1	1	N	N	N	N
1	0	0	0	N	N	N	N
1	0	0	1	N	Y	N	N
1	0	1	0	N	N	Y	N
1	0	1	1	N	N	Y	N
1	1	0	0	Y	N	Y	Y
1	1	0	1	Y	Y	Y	Y
1	1	1	0	Y	N	Y	Y
1	1	1	1	Y	N	Y	Y

# A random forest example

[IMS21]

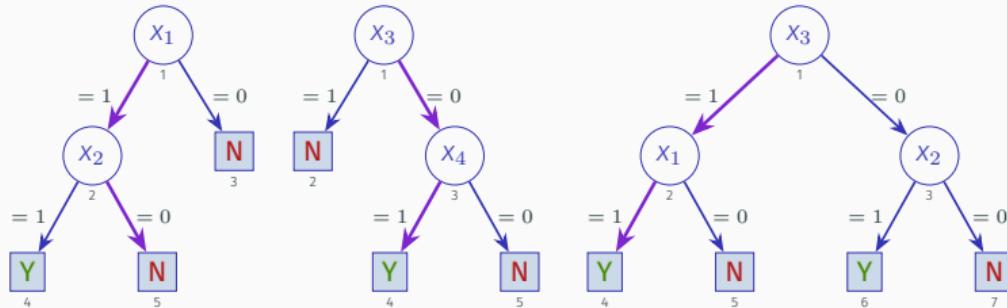


- Explanation for why  $\kappa(1, 0, 0, 1) = \text{N}$ ?

X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	$\kappa(\mathbf{x})$
0	0	0	0	N	N	N	N
0	0	0	1	N	Y	N	N
0	0	1	0	N	N	N	N
0	0	1	1	N	N	N	N
0	1	0	0	N	N	Y	N
0	1	0	1	N	Y	Y	Y
0	1	1	0	N	N	N	N
0	1	1	1	N	N	N	N
1	0	0	0	N	N	N	N
1	0	0	1	N	Y	N	N
1	0	1	0	N	N	Y	N
1	0	1	1	N	N	Y	N
1	1	0	0	Y	N	Y	Y
1	1	0	1	Y	Y	Y	Y
1	1	1	0	Y	N	Y	Y
1	1	1	1	Y	N	Y	Y

# A random forest example

[IMS21]



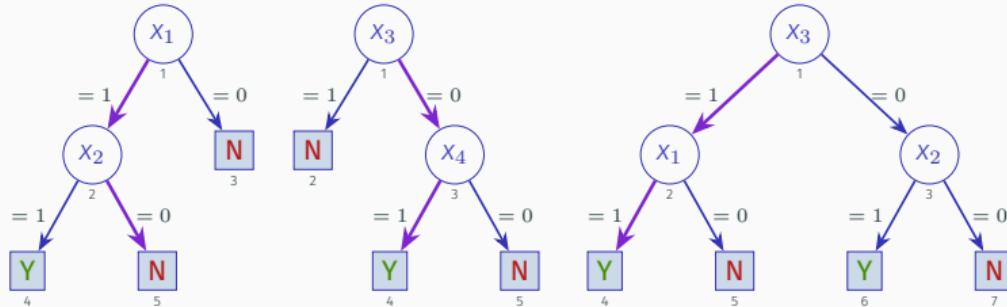
- Explanation for why  $\kappa(1, 0, 0, 1) = \mathbf{N}$ ?

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF ( $x_2 = 0$ ) THEN  $\kappa(\mathbf{x}) = \mathbf{N}$
- I.e.  $\{x_2 = 0\}$  suffices for RF to predict **N**

$x_1$	$x_2$	$x_3$	$x_4$	$T_1$	$T_2$	$T_3$	$\kappa(\mathbf{x})$
0	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
0	0	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	0	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
0	1	0	1	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
0	1	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
1	0	1	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	0	1	1	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	1	0	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	0	1	<b>Y</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
1	1	1	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	1	1	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>

# A random forest example

[IMS21]

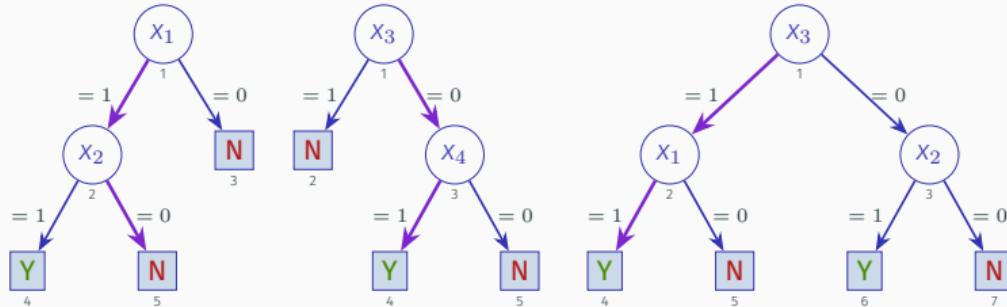


- Explanation for why  $\kappa(1, 0, 0, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF ( $x_2 = 0$ ) THEN  $\kappa(\mathbf{x}) = \text{N}$
  - i.e.  $\{x_2 = 0\}$  suffices for RF to predict **N**
- Explanation for why  $\kappa(1, 1, 1, 1) = \text{Y}$ ?

X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	T <sub>1</sub>	T <sub>2</sub>	T <sub>3</sub>	$\kappa(\mathbf{x})$
0	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
0	0	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	0	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
0	1	0	1	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
0	1	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
1	0	1	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	0	1	1	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	1	0	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	0	1	<b>Y</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
1	1	1	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	1	1	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>

# A random forest example

[IMS21]

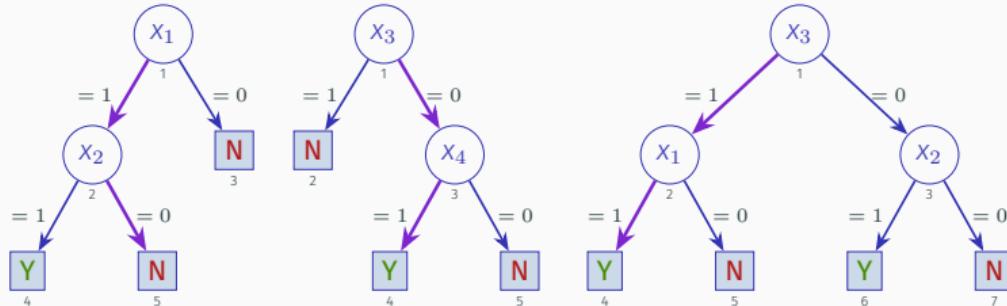


- Explanation for why  $\kappa(1, 0, 0, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_2 = 0)$  THEN  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_2 = 0\}$  suffices for RF to predict **N**
- Explanation for why  $\kappa(1, 1, 1, 1) = \text{Y}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_1 = 1) \wedge (x_2 = 1)$  THEN  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_1 = 1, x_2 = 1\}$  suffice for RF to predict **Y**

$x_1$	$x_2$	$x_3$	$x_4$	$T_1$	$T_2$	$T_3$	$\kappa(\mathbf{x})$
0	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
0	0	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	0	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	0	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
0	1	0	1	<b>N</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
0	1	1	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
0	1	1	1	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	0	<b>N</b>	<b>N</b>	<b>N</b>	<b>N</b>
1	0	0	1	<b>N</b>	<b>Y</b>	<b>N</b>	<b>N</b>
1	0	1	0	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	0	1	1	<b>N</b>	<b>N</b>	<b>Y</b>	<b>N</b>
1	1	0	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	0	1	<b>Y</b>	<b>Y</b>	<b>Y</b>	<b>Y</b>
1	1	1	0	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>
1	1	1	1	<b>Y</b>	<b>N</b>	<b>Y</b>	<b>Y</b>

# A random forest example

[IMS21]

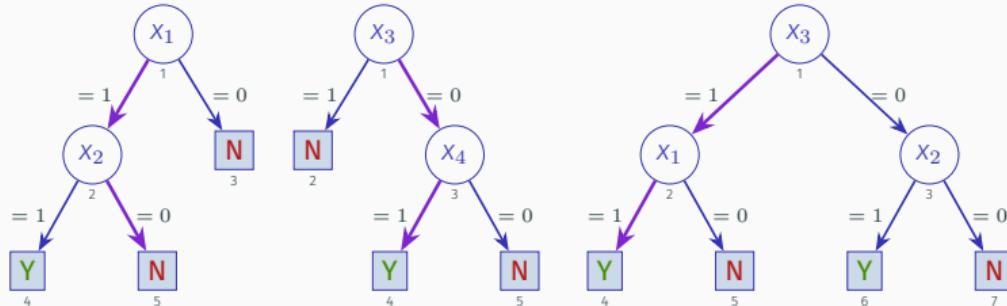


- Explanation for why  $\kappa(1, 0, 0, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_2 = 0)$  THEN  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_2 = 0\}$  suffices for RF to predict N
- Explanation for why  $\kappa(1, 1, 1, 1) = \text{Y}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_1 = 1) \wedge (x_2 = 1)$  THEN  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_1 = 1, x_2 = 1\}$  suffice for RF to predict Y
- Explanation for why  $\kappa(0, 1, 1, 1) = \text{N}$ ?

$x_1$	$x_2$	$x_3$	$x_4$	$T_1$	$T_2$	$T_3$	$\kappa(\mathbf{x})$
0	0	0	0	N	N	N	N
0	0	0	1	N	Y	N	N
0	0	1	0	N	N	N	N
0	0	1	1	N	N	N	N
0	1	0	0	N	N	Y	N
0	1	0	1	N	Y	Y	Y
0	1	1	0	N	N	N	N
0	1	1	1	N	N	N	N
1	0	0	0	N	N	N	N
1	0	0	1	N	Y	N	N
1	0	1	0	N	N	Y	N
1	0	1	1	N	N	Y	N
1	1	0	0	Y	N	Y	Y
1	1	0	1	Y	Y	Y	Y
1	1	1	0	Y	N	Y	Y
1	1	1	1	Y	N	Y	Y

# A random forest example

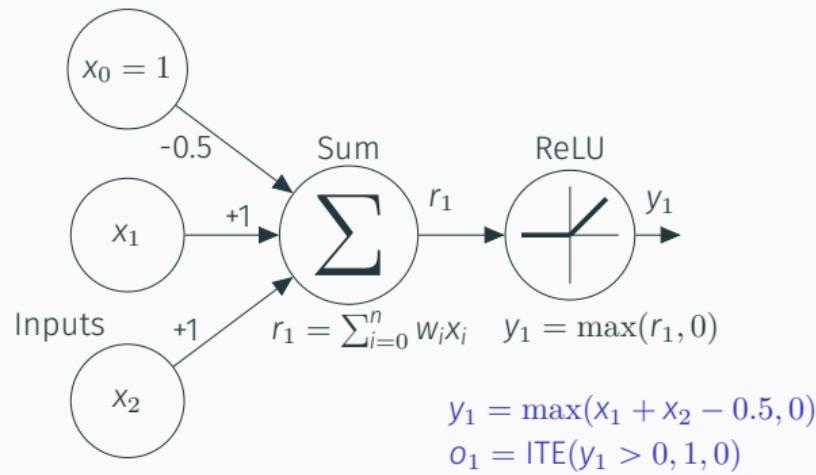
[IMS21]



- Explanation for why  $\kappa(1, 0, 0, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_2 = 0)$  THEN  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_2 = 0\}$  suffices for RF to predict N
- Explanation for why  $\kappa(1, 1, 1, 1) = \text{Y}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_1 = 1) \wedge (x_2 = 1)$  THEN  $\kappa(\mathbf{x}) = \text{Y}$
  - I.e.  $\{x_1 = 1, x_2 = 1\}$  suffice for RF to predict Y
- Explanation for why  $\kappa(0, 1, 1, 1) = \text{N}$ ?
  - Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ , IF  $(x_1 = 0) \wedge (x_2 = 1) \wedge (x_3 = 1)$  THEN  $\kappa(\mathbf{x}) = \text{N}$
  - I.e.  $\{x_1 = 0, x_2 = 1, x_3 = 1\}$  suffices for RF to predict N

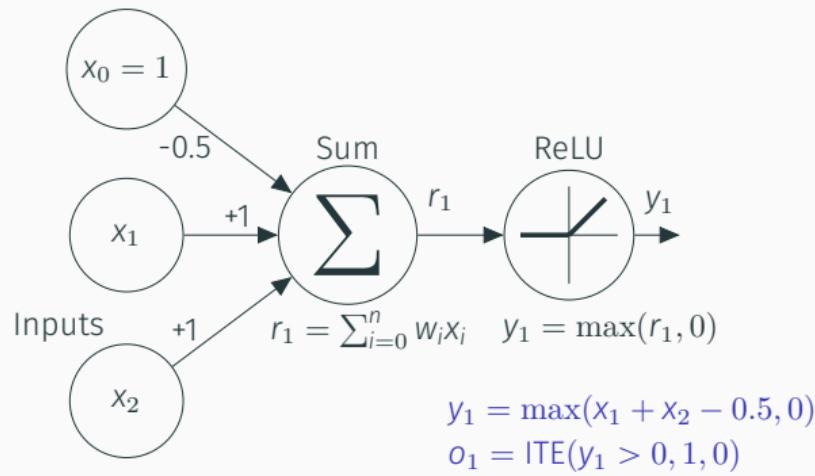
$x_1$	$x_2$	$x_3$	$x_4$	$T_1$	$T_2$	$T_3$	$\kappa(\mathbf{x})$
0	0	0	0	N	N	N	N
0	0	0	1	N	Y	N	N
0	0	1	0	N	N	N	N
0	0	1	1	N	N	N	N
0	1	0	0	N	N	Y	N
0	1	0	1	N	Y	Y	Y
0	1	1	0	N	N	N	N
0	1	1	1	N	N	N	N
1	0	0	0	N	N	N	N
1	0	0	1	N	Y	N	N
1	0	1	0	N	N	Y	N
1	0	1	1	N	N	Y	N
1	1	0	0	Y	N	Y	Y
1	1	0	1	Y	Y	Y	Y
1	1	1	0	Y	N	Y	Y
1	1	1	1	Y	N	Y	Y

# A neural network example



$x_1$	$x_2$	$r_1$	$y_1$	$\kappa(\mathbf{x})$
0	0	-0.5	0	0
0	1	0.5	0.5	1
1	0	0.5	0.5	1
1	1	1.5	1.5	1

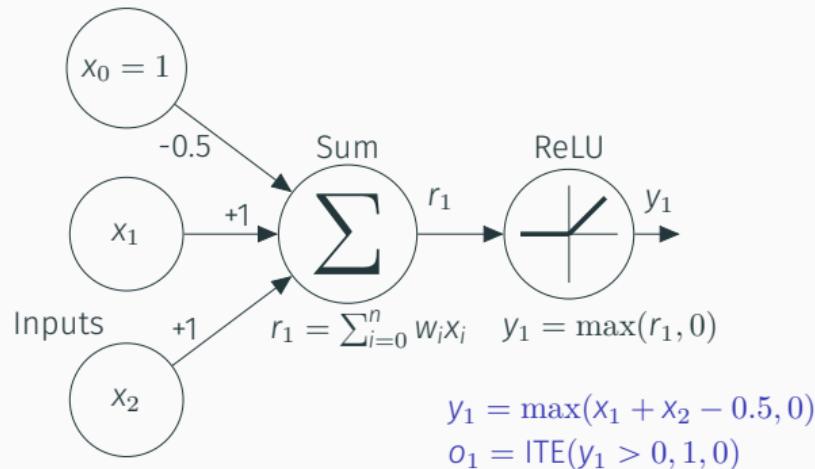
# A neural network example



$x_1$	$x_2$	$r_1$	$y_1$	$\kappa(\mathbf{x})$
0	0	-0.5	0	0
0	1	0.5	0.5	1
1	0	0.5	0.5	1
1	1	1.5	1.5	1

- Explanation for why  $\kappa(1, 1) = 1$ ?

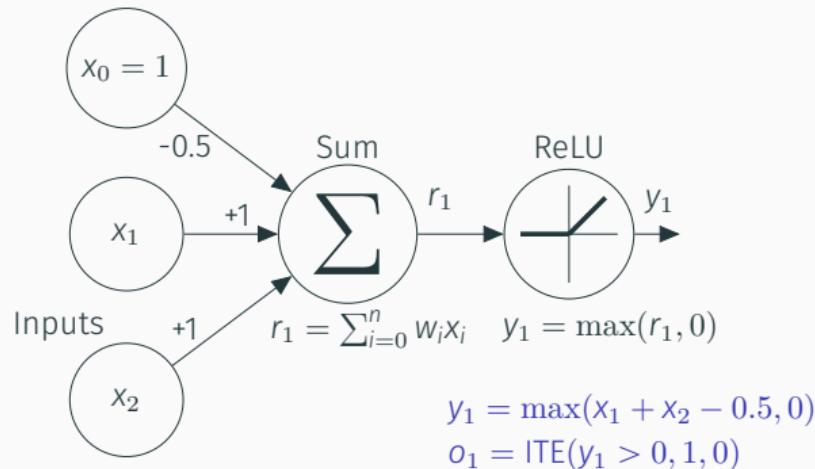
# A neural network example



$x_1$	$x_2$	$r_1$	$y_1$	$\kappa(\mathbf{x})$
0	0	-0.5	0	0
0	1	0.5	0.5	1
1	0	0.5	0.5	1
1	1	1.5	1.5	1

- Explanation for why  $\kappa(1, 1) = 1$ ?
  - Given  $\mathbf{x} = (x_1, x_2)$ , IF  $(x_1 = 1)$  THEN  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_1 = 1\}$  suffices for NN to predict 1

# A neural network example



$x_1$	$x_2$	$r_1$	$y_1$	$\kappa(\mathbf{x})$
0	0	-0.5	0	0
0	1	0.5	0.5	1
1	0	0.5	0.5	1
1	1	1.5	1.5	1

- Explanation for why  $\kappa(1, 1) = 1$ ?
  - Given  $\mathbf{x} = (x_1, x_2)$ , IF  $(x_1 = 1)$  THEN  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_1 = 1\}$  suffices for NN to predict 1
  - Given  $\mathbf{x} = (x_1, x_2)$ , IF  $(x_2 = 1)$  THEN  $\kappa(\mathbf{x}) = 1$
  - i.e.  $\{x_2 = 1\}$  suffices for NN to predict Y

# An arbitrary classifier

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

# An arbitrary classifier

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 0, 0, 0), 1)$

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

# An arbitrary classifier

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 0, 0, 0), 1)$

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,

IF  $(x_1 = 0) \wedge (x_3 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$

- I.e.  $\{x_1 = 0, x_3 = 0\}$  suffices for RF to predict 1

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

# Outline – Unit #01

ML Models: Classification & Regression Problems

Basics of (non-symbolic) XAI

Motivation for Explanations

Brief Glimpse of Logic

Reasoning About ML Models

Understanding Intrinsic Interpretability

# Standard tools of the trade

- **SAT**: decision problem for propositional logic
  - Formulas most often represented in CNF
  - There are optimization variants: MaxSAT, PBO, MinSAT, etc.
  - There are quantified variants: QBF, QMaxSAT, etc.
- **SMT**: decision problem for (decidable) fragments of first-order logic (**FOL**)
  - There are optimization variants: MaxSMT, etc.
  - There are quantified variants
- **MILP**: decision/optimization problems defined on conjunctions of linear inequalities over integer & real-valued variables
- **CP**: constraint programming
  - There are optimization/quantified variants

# Standard tools of the trade

- **SAT**: decision problem for propositional logic
  - Formulas most often represented in CNF
  - There are optimization variants: MaxSAT, PBO, MinSAT, etc.
  - There are quantified variants: QBF, QMaxSAT, etc.
- **SMT**: decision problem for (decidable) fragments of first-order logic (**FOL**)
  - There are optimization variants: MaxSMT, etc.
  - There are quantified variants
- **MILP**: decision/optimization problems defined on conjunctions of linear inequalities over integer & real-valued variables
- **CP**: constraint programming
  - There are optimization/quantified variants
- Background on SAT/SMT:
  - <https://alexeyignatiev.github.io/ssa-school-2019/>
  - <https://alexeyignatiev.github.io/ijcai19tut/>

Basic knowledge on  
SAT & SMT assumed.  
See links below.

[BHvMW09]

## SAT/SMT/MILP/CP solvers used as oracles – more detail later

- Deciding satisfiability, entailment
- Computing prime implicants/implicates
- Computing MUSes, MCSes
  - Algorithms: Deletion, QuickXplain, Progression, Dichotomic, etc. [MM20]
- Enumeration of MUSes, MCSes
  - Algorithms: Marco, Camus, etc. [LS08, LPMM16]
- Solving MaxSAT, MaxSMT
  - Algorithms: Core-guided, Minimum hitting sets, branch&bound, etc. [MHL<sup>+</sup>13]
- Solving quantification problems, e.g. QBF
  - Algorithms: Abstraction refinement [JKMC16]

# Basic definitions in propositional logic

- Atoms ( $\{x, x_1, \dots\}$ ) & literals ( $x_1, \neg x_1$ )
- Well-formed formulas using  $\neg, \wedge, \vee, \dots$
- Clause: disjunction of literals
- Term: conjunction of literals
- Conjunctive normal form (CNF): conjunction of clauses
- Disjunctive normal form (DNF): disjunction of terms
- Simple to generalize to more expressive domains

# Basic definitions in propositional logic

- Atoms ( $\{x, x_1, \dots\}$ ) & literals ( $x_1, \neg x_1$ )
- Well-formed formulas using  $\neg, \wedge, \vee, \dots$
- Clause: disjunction of literals
- Term: conjunction of literals
- Conjunctive normal form (CNF): conjunction of clauses
- Disjunctive normal form (DNF): disjunction of terms
- Simple to generalize to more expressive domains
- $\text{CO}(\psi(\mathbf{x}))$  decides whether  $\psi(\mathbf{x})$  is **satisfiable** (i.e. whether it is **consistent**), using an oracle for SAT/SMT/MILP/CP/etc.

## Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$

# Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$ 
  - We say that  $\tau$  **entails**  $\varphi$ , written as  $\tau \vDash \varphi$ , if:

$$\forall(\mathbf{x} \in \mathbb{F}).[\tau(\mathbf{x}) \rightarrow \varphi(\mathbf{x})]$$

# Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$ 
  - We say that  $\tau$  **entails**  $\varphi$ , written as  $\tau \vDash \varphi$ , if:

$$\forall(\mathbf{x} \in \mathbb{F}).[\tau(\mathbf{x}) \rightarrow \varphi(\mathbf{x})]$$

- We say that  $\tau(\mathbf{x})$  is **sufficient** for  $\varphi(\mathbf{x})$

# Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$ 
  - We say that  $\tau$  **entails**  $\varphi$ , written as  $\tau \vDash \varphi$ , if:

$$\forall(\mathbf{x} \in \mathbb{F}).[\tau(\mathbf{x}) \rightarrow \varphi(\mathbf{x})]$$

- We say that  $\tau(\mathbf{x})$  is **sufficient** for  $\varphi(\mathbf{x})$
- To decide entailment:
  - $\tau \vDash \varphi$  if  $\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x})$  is **not** consistent, i.e.  $\text{CO}(\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x}))$  does not hold

# Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$ 
  - We say that  $\tau$  entails  $\varphi$ , written as  $\tau \vDash \varphi$ , if:

$$\forall(\mathbf{x} \in \mathbb{F}).[\tau(\mathbf{x}) \rightarrow \varphi(\mathbf{x})]$$

- We say that  $\tau(\mathbf{x})$  is sufficient for  $\varphi(\mathbf{x})$
- To decide entailment:
  - $\tau \vDash \varphi$  if  $\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x})$  is not consistent, i.e.  $\text{CO}(\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x}))$  does not hold
- An example:
  - $\mathbb{F} = \{0, 1\}^2$
  - $\varphi(x_1, x_2) = x_1 \vee \neg x_2$
  - Clearly,  $x_1 \vDash \varphi$  and  $\neg x_2 \vDash \varphi$
  - Also,  $\text{CO}(x_1 \wedge (\neg x_1 \wedge x_2))$  does not hold

# Entailment

- Let  $\varphi$  represent some formula, defined on feature space  $\mathbb{F}$ , and representing a function  $\varphi : \mathbb{F} \rightarrow \{0, 1\}$
- Let  $\tau$  represent some other formula, also defined on  $\mathbb{F}$ , and with  $\tau : \mathbb{F} \rightarrow \{0, 1\}$ 
  - We say that  $\tau$  entails  $\varphi$ , written as  $\tau \vDash \varphi$ , if:

$$\forall(\mathbf{x} \in \mathbb{F}).[\tau(\mathbf{x}) \rightarrow \varphi(\mathbf{x})]$$

- We say that  $\tau(\mathbf{x})$  is sufficient for  $\varphi(\mathbf{x})$
- To decide entailment:
  - $\tau \vDash \varphi$  if  $\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x})$  is not consistent, i.e.  $\text{CO}(\tau(\mathbf{x}) \wedge \neg\varphi(\mathbf{x}))$  does not hold

- An example:
  - $\mathbb{F} = \{0, 1\}^2$
  - $\varphi(x_1, x_2) = x_1 \vee \neg x_2$
  - Clearly,  $x_1 \vDash \varphi$  and  $\neg x_2 \vDash \varphi$
  - Also,  $\text{CO}(x_1 \wedge (\neg x_1 \wedge x_2))$  does not hold

- Another example:
  - $\mathbb{F} = \{0, 1\}^3$
  - $\varphi(x_1, x_2, x_3) = x_1 \wedge x_2 \vee x_1 \wedge x_3$
  - Clearly,  $x_1 \wedge x_2 \vDash \varphi$  and  $x_1 \wedge x_3 \vDash \varphi$
  - Also,  $\text{CO}(x_1 \wedge x_2 \wedge ((\neg x_1 \vee \neg x_2) \wedge (\neg x_1 \vee \neg x_3)))$  does not hold

## Entailment & explanations – how do we construct explanations?

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 1, 0, 0), 1)$

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

## Entailment & explanations – how do we construct explanations?

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 1, 0, 0), 1)$

- Localized explanation:** any irreducible conjunction of literals, consistent with  $\nu$ , and that entails the prediction

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

## Entailment & explanations – how do we construct explanations?

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 1, 0, 0), 1)$

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

- Localized explanation:** any irreducible conjunction of literals, consistent with  $\mathbf{v}$ , and that entails the prediction

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,

IF  $(x_1 = 0) \wedge (x_3 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$

## Entailment & explanations – how do we construct explanations?

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 1, 0, 0), 1)$

- Localized explanation:** any irreducible conjunction of literals, consistent with  $\mathbf{v}$ , and that entails the prediction

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,

IF  $(x_1 = 0) \wedge (x_3 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$

- Global explanation:** any irreducible conjunction of literals, that is consistent, and that entails the prediction

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

# Entailment & explanations – how do we construct explanations?

- Classification function:

$$\kappa(x_1, x_2, x_3, x_4) = \neg x_1 \wedge \neg x_2 \vee x_1 \wedge x_2 \wedge x_4 \vee \neg x_1 \wedge x_2 \wedge \neg x_3 \vee \neg x_2 \wedge x_3 \wedge x_4$$

- Instance:  $((0, 1, 0, 0), 1)$

- Localized explanation:** any irreducible conjunction of literals, consistent with  $\mathbf{v}$ , and that entails the prediction

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,

IF  $(x_1 = 0) \wedge (x_3 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$

- Global explanation:** any irreducible conjunction of literals, that is consistent, and that entails the prediction

- Given  $\mathbf{x} = (x_1, x_2, x_3, x_4)$ ,

IF  $(x_1 = 0) \wedge (x_2 = 0)$  THEN  $\kappa(\mathbf{x}) = 1$

$x_1$	$x_2$	$x_3$	$x_4$	$\kappa(\mathbf{x})$
0	0	0	0	1
0	0	0	1	1
0	0	1	0	1
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	1

# Outline – Unit #01

ML Models: Classification & Regression Problems

Basics of (non-symbolic) XAI

Motivation for Explanations

Brief Glimpse of Logic

Reasoning About ML Models

Understanding Intrinsic Interpretability

# Decision sets with boolean features

- Example ML model:

Features:  $x_1, x_2, x_3, x_4 \in \{0, 1\}$  (boolean)

Rules:

IF	$x_1 \wedge \neg x_2 \wedge x_3$	THEN	predict $\boxplus$
IF	$x_1 \wedge \neg x_3 \wedge x_4$	THEN	predict $\boxminus$
IF	$x_3 \wedge x_4$	THEN	predict $\boxminus$

# Decision sets with boolean features

- Example ML model:

Features:  $x_1, x_2, x_3, x_4 \in \{0, 1\}$  (boolean)

Rules:

IF	$x_1 \wedge \neg x_2 \wedge x_3$	THEN	predict $\blacksquare$
IF	$x_1 \wedge \neg x_3 \wedge x_4$	THEN	predict $\square$
IF	$x_3 \wedge x_4$	THEN	predict $\square$

- Q:** Can the model predict both  $\blacksquare$  and  $\square$  for some instance, i.e. is there overlap?

# Decision sets with boolean features

- Example ML model:

Features:  $x_1, x_2, x_3, x_4 \in \{0, 1\}$  (boolean)

Rules:

IF	$x_1 \wedge \neg x_2 \wedge x_3$	THEN	predict $\blacksquare$
IF	$x_1 \wedge \neg x_3 \wedge x_4$	THEN	predict $\blacksquare$
IF	$x_3 \wedge x_4$	THEN	predict $\blacksquare$

- Q:** Can the model predict both  $\blacksquare$  and  $\blacksquare$  for some instance, i.e. is there overlap?

- Yes, certainly: pick  $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$

# Decision sets with boolean features

- Example ML model:

Features:  $x_1, x_2, x_3, x_4 \in \{0, 1\}$  (boolean)

Rules:

IF	$x_1 \wedge \neg x_2 \wedge x_3$	THEN	predict $\blacksquare$
IF	$x_1 \wedge \neg x_3 \wedge x_4$	THEN	predict $\square$
IF	$x_3 \wedge x_4$	THEN	predict $\square$

- Q: Can the model predict both  $\blacksquare$  and  $\square$  for some instance, i.e. is there overlap?

- Yes, certainly: pick  $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$
- A formalization:

$$\begin{aligned}y_{p,1} &\leftrightarrow (x_1 \wedge \neg x_2 \wedge x_3) \wedge \\y_{n,1} &\leftrightarrow (x_1 \wedge \neg x_3 \wedge x_4) \wedge \\y_{n,2} &\leftrightarrow (x_3 \wedge x_4) \wedge (y_p \leftrightarrow y_{p,1}) \wedge \\(y_n &\leftrightarrow (y_{n,1} \vee y_{n,2})) \wedge (y_p) \wedge (y_n)\end{aligned}$$

... and solve with SAT solver (after clausification)

Or use PySAT

[Tse68, PG86]

[IMM18]

$\therefore$  There exists a model iff there exists a point in feature space yielding both predictions

## Decision sets with ordinal features

- Example ML model:

Features:  $x_1, x_2 \in \{0, 1, 2\}$  (integer)

Rules:

IF  $2x_1 + x_2 > 0$  THEN predict  $\text{田}$

IF  $2x_1 - x_2 \leq 0$  THEN predict  $\text{□}$

## Decision sets with ordinal features

- Example ML model:

Features:  $x_1, x_2 \in \{0, 1, 2\}$  (integer)

Rules:

IF  $2x_1 + x_2 > 0$  THEN predict  $\text{■}$

IF  $2x_1 - x_2 \leq 0$  THEN predict  $\text{□}$

- Q:** Can the model predict both  $\text{■}$  and  $\text{□}$  for some instance, i.e. is there overlap?

## Decision sets with ordinal features

- Example ML model:

Features:  $x_1, x_2 \in \{0, 1, 2\}$  (integer)

Rules:

IF  $2x_1 + x_2 > 0$  THEN predict  $\text{■}$

IF  $2x_1 - x_2 \leq 0$  THEN predict  $\text{□}$

- Q:** Can the model predict both  $\text{■}$  and  $\text{□}$  for some instance, i.e. is there overlap?

- Yes, of course: pick  $x_1 = 0$  and  $x_2 = 1$

# Decision sets with ordinal features

- Example ML model:

Features:  $x_1, x_2 \in \{0, 1, 2\}$  (integer)

Rules:

IF  $2x_1 + x_2 > 0$  THEN predict  $\blacksquare$

IF  $2x_1 - x_2 \leq 0$  THEN predict  $\blacksquare$

- Q: Can the model predict both  $\blacksquare$  and  $\blacksquare$  for some instance, i.e. is there overlap?

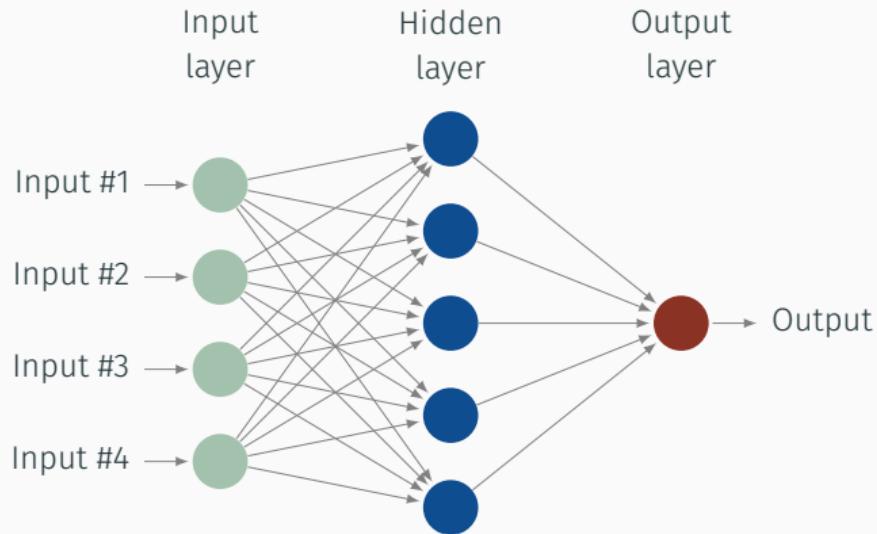
- Yes, of course: pick  $x_1 = 0$  and  $x_2 = 1$
- A formalization:

$$y_p \leftrightarrow (2x_1 + x_2 > 0) \wedge y_n \leftrightarrow (2x_1 - x_2 \leq 0) \wedge (y_p) \wedge (y_n)$$

... and solve with SMT solver (many alternatives)

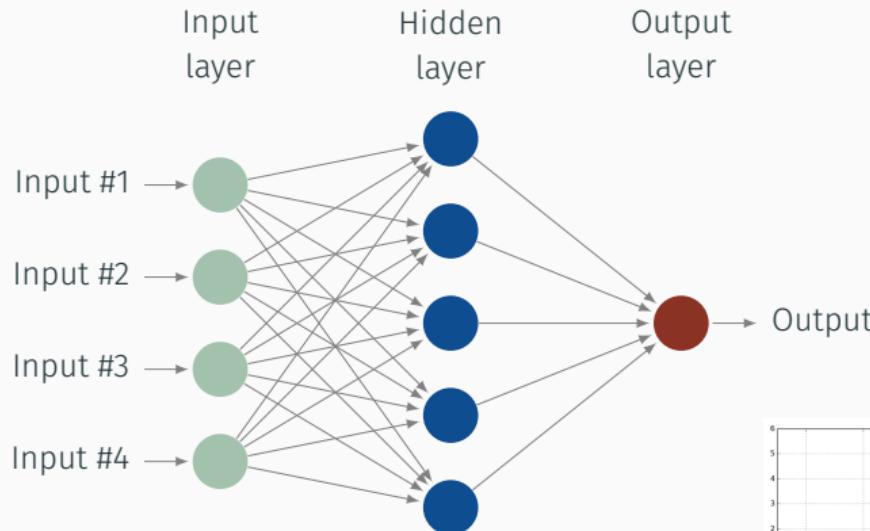
. $\therefore$  There exists a model iff there exists a point in feature space yielding both predictions

# Neural networks

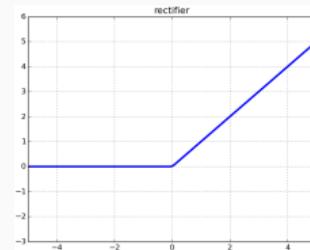


- Each layer (except first) viewed as a **block**, and
  - Compute  $\mathbf{x}'$  given input  $\mathbf{x}$ , weights matrix  $\mathbf{A}$ , and bias vector  $\mathbf{b}$
  - Compute output  $\mathbf{y}$  given  $\mathbf{x}'$  and activation function

# Neural networks



- Each layer (except first) viewed as a **block**, and
  - Compute  $\mathbf{x}'$  given input  $\mathbf{x}$ , weights matrix  $\mathbf{A}$ , and bias vector  $\mathbf{b}$
  - Compute output  $\mathbf{y}$  given  $\mathbf{x}'$  and activation function
- Each unit uses a **ReLU** activation function



[NH10]

## Encoding NNs using MILP

Computation for a NN ReLU **block**, in two steps:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$

$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

# Encoding NNs using MILP

Computation for a NN ReLU **block**, in two steps:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$

$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

Encoding each **block**:

[FJ18]

$$\sum_{j=1}^n a_{i,j}x_j + b_i = y_i - s_i$$

$$z_i = 1 \rightarrow y_i \leq 0$$

$$z_i = 0 \rightarrow s_i \leq 0$$

$$y_i \geq 0, s_i \geq 0, z_i \in \{0, 1\}$$

Simpler encodings exist, but **not** as effective

[KBD<sup>+</sup>17]

# Encoding NNs using MILP

Computation for a NN ReLU **block**, in two steps:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$

$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

Modeling ML models  
with logic is not only  
possible but also **simple !**

Encoding each **block**:

[FJ18]

$$\sum_{j=1}^n a_{i,j}x_j + b_i = y_i - s_i$$

$$z_i = 1 \rightarrow y_i \leq 0$$

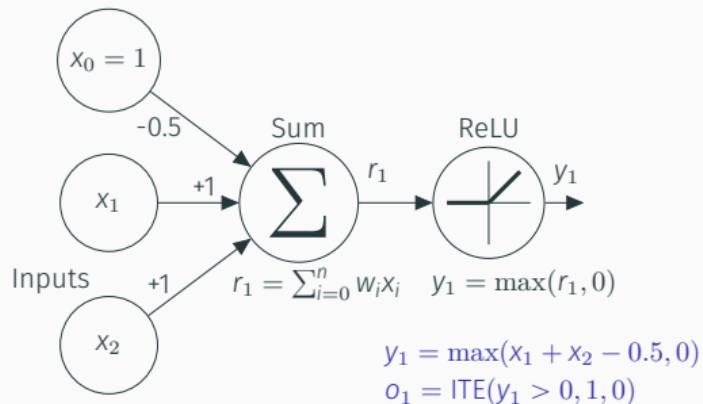
$$z_i = 0 \rightarrow s_i \leq 0$$

$$y_i \geq 0, s_i \geq 0, z_i \in \{0, 1\}$$

Simpler encodings exist, but **not** as effective

[KBD<sup>+</sup>17]

## Example – encoding a simple NN in MILP



$x_1$	$x_2$	$r_1$	$y_1$	$o_1$
0	0	-0.5	0	0
1	0	0.5	0.5	1
0	1	0.5	0.5	1
1	1	1.5	1.5	1

MILP encoding:

$$x_1 + x_2 - 0.5 = y_1 - s_1$$

$$z_1 = 1 \rightarrow y_1 \leq 0$$

$$z_1 = 0 \rightarrow s_1 \leq 0$$

$$o_1 = (y_1 > 0)$$

$$x_1, x_2, z_1, o_1 \in \{0, 1\}$$

$$y_1, s_1 \geq 0$$

Instance:  $(\mathbf{x}, c) = ((1, 0), 1)$

$$1 + 0 - 0.5 = 0.5 - 0$$

$$1 \vee 0.5 \leq 0$$

$$0 \vee 0 \leq 0$$

$$1 = (0.5 > 0)$$

$$x_1 = 1, x_2 = 0, z_1 = 0, o_1 = 1$$

$$y_1 = 0.5, s_1 = 0$$

Checking:  $\mathbf{x} = (0, 0)$

$$0 + 0 - 0.5 = 0 - 0.5$$

$$0 \vee 0 \leq 0$$

$$1 \vee 0.5 \leq 0$$

$$0 = (0 > 0)$$

$$x_1 = 0, x_2 = 0, z_1 = 1, o_1 = 0$$

$$y_1 = 0, s_1 = 0.5$$

# Outline – Unit #01

ML Models: Classification & Regression Problems

Basics of (non-symbolic) XAI

Motivation for Explanations

Brief Glimpse of Logic

Reasoning About ML Models

Understanding Intrinsic Interpretability

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

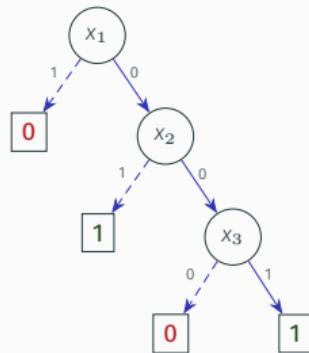
[Lip18]

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]

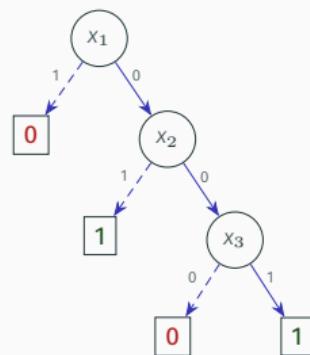


# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



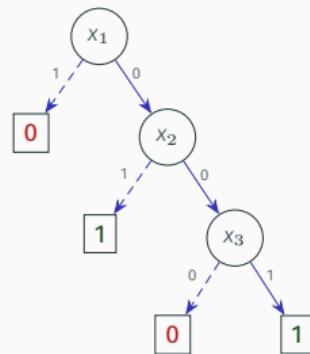
- What is an explanation for  $((0, 0, 1), 1)$ ?

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



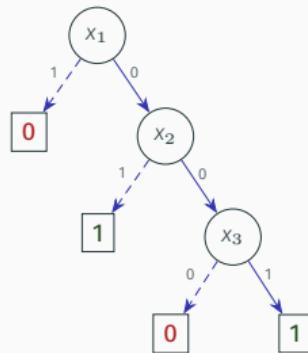
- What is an explanation for  $((0, 0, 1), 1)$ ?
- Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



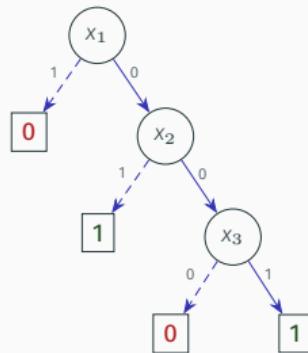
- What is an explanation for  $((0, 0, 1), 1)$ ?
- Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$ 
  - $\{\neg x_1, \neg x_2, x_3\}$  or  $\{1, 2, 3\}$  is an explanation

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective...*

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



- What is an explanation for  $((0, 0, 1), 1)$ ?
- Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$ 
  - $\{\neg x_1, \neg x_2, x_3\}$  or  $\{1, 2, 3\}$  is an explanation

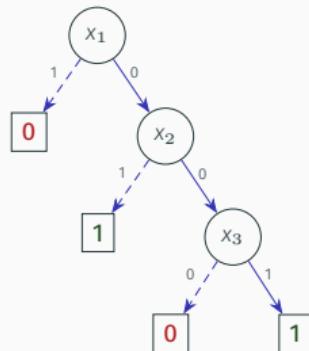
Really?

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



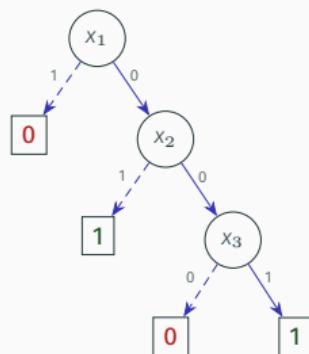
- What is an explanation for  $((0, 0, 1), 1)$ ?
- Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$ 
  - $\{\neg x_1, \neg x_2, x_3\}$  or  $\{1, 2, 3\}$  is a **weak** explanation!
- It is the case that: IF  $\neg x_1 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$   
 $\therefore \{1, 3\}$  is also **sufficient** for the prediction!

# What is intrinsic interpretability?

- Goal is to deploy *interpretable* ML models
  - E.g. Decision trees, decision lists, decision sets, etc.
- The explanation is the model itself, because it is *interpretable*
- **But:** definition of *interpretability* is rather *subjective*...

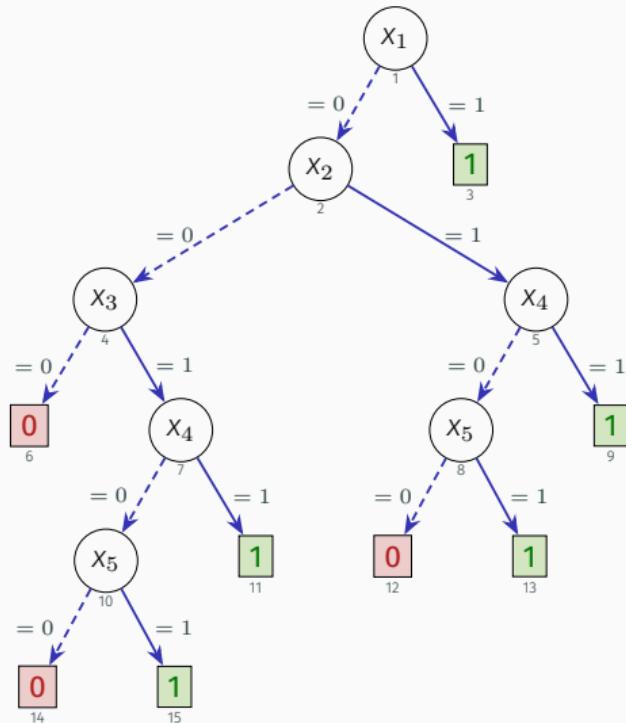
[Rud19, Mol20, RCC<sup>+</sup>22, Rud22]

[Lip18]



- What is an explanation for  $((0, 0, 1), 1)$ ?
- Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$ 
  - $\{\neg x_1, \neg x_2, x_3\}$  or  $\{1, 2, 3\}$  is a **weak** explanation!
- It is the case that: IF  $\neg x_1 \wedge x_3$  THEN  $\kappa(\mathbf{x}) = 1$ 
  - $\{1, 3\}$  is also **sufficient** for the prediction!
  - $\{1, 3\}$  is easier to grasp; also, it is **irreducible**

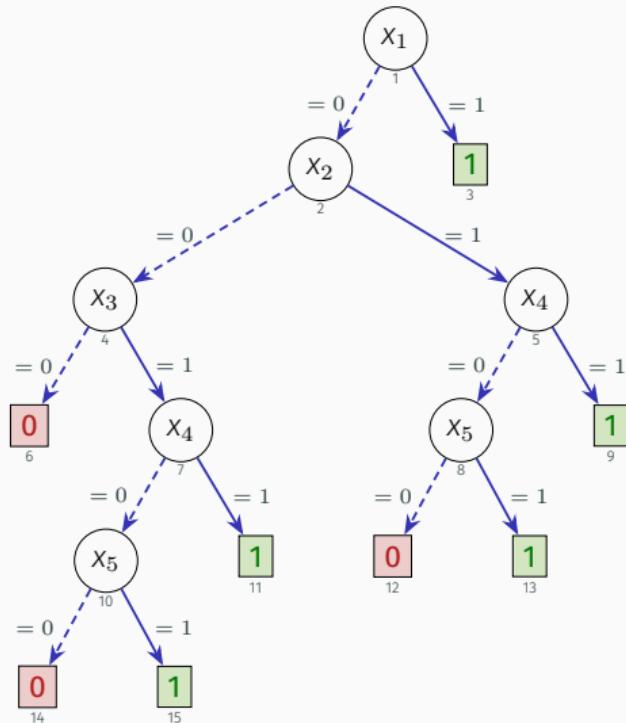
# Are *interpretable* models really interpretable? – DTs



- Case of **optimal** decision tree (DT)
- Explanation for  $(0, 0, 1, 0, 1)$ , with prediction 1?

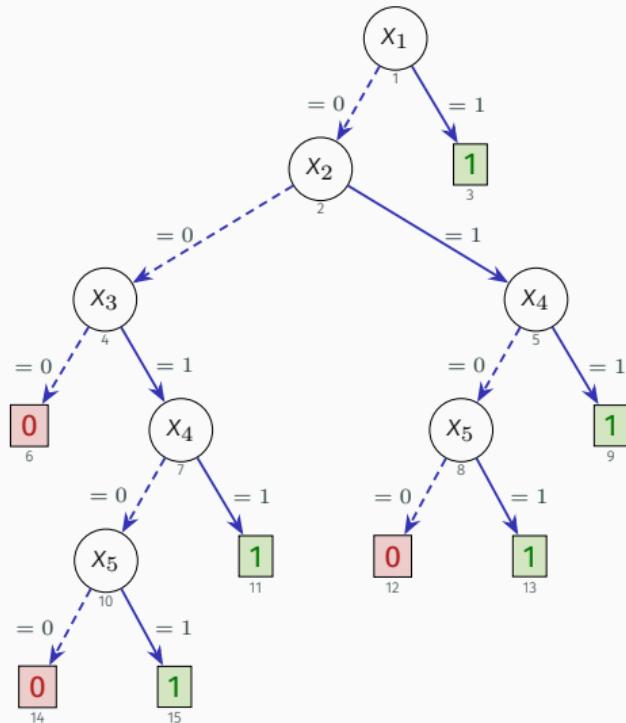
[HRS19]

# Are *interpretable* models really interpretable? – DTs



- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for  $(0, 0, 1, 0, 1)$ , with prediction 1?
  - Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$  THEN  $\kappa(\mathbf{x}) = 1$

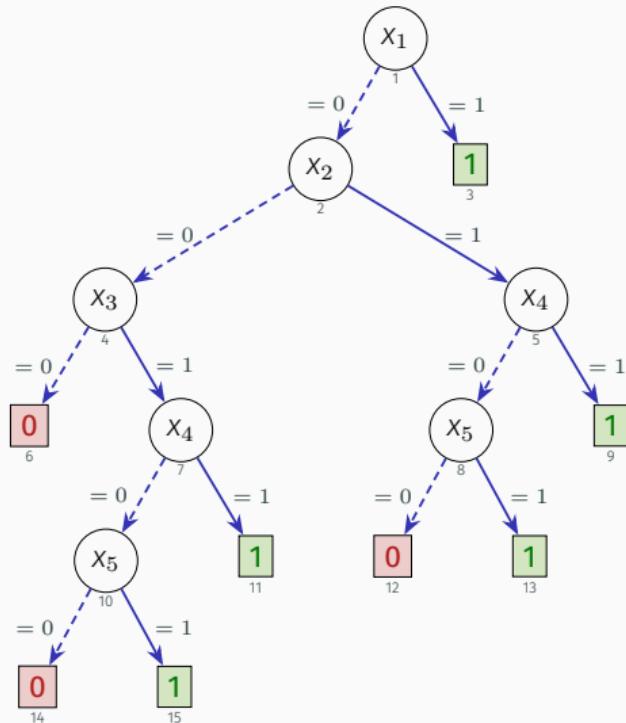
# Are *interpretable* models really interpretable? – DTs



- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for  $(0, 0, 1, 0, 1)$ , with prediction 1?
  - Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$  THEN  $\kappa(\mathbf{x}) = 1$
  - But,  $x_1, x_2, x_4$  are **irrelevant** for the prediction:

$x_3$	$x_5$	$x_1$	$x_2$	$x_4$	$\kappa(\mathbf{x})$
1	1	0	0	0	1
1	1	0	0	1	1
1	1	0	1	0	1
1	1	0	1	1	1
1	1	1	0	0	1
1	1	1	0	1	1
1	1	1	1	0	1
1	1	1	1	1	1

# Are *interpretable* models really interpretable? – DTs



- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for  $(0, 0, 1, 0, 1)$ , with prediction 1?
  - Clearly, IF  $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$  THEN  $\kappa(\mathbf{x}) = 1$
  - But,  $x_1, x_2, x_4$  are **irrelevant** for the prediction:

$x_3$	$x_5$	$x_1$	$x_2$	$x_4$	$\kappa(\mathbf{x})$
1	1	0	0	0	1
1	1	0	0	1	1
1	1	0	1	0	1
1	1	0	1	1	1
1	1	1	0	0	1
1	1	1	0	1	1
1	1	1	1	0	1
1	1	1	1	1	1

$\therefore$  fixing  $\{3, 5\}$  suffices for the prediction  
Compare with  $\{1, 2, 3, 4, 5\}...$

$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires

$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires
- What is an explanation for the prediction?

$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires
- What is an explanation for the prediction?
- Fixing  $\{3, 4, 6\}$  suffices for the prediction

$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires
- What is an explanation for the prediction?
- Fixing  $\{3, 4, 6\}$  suffices for the prediction
  - Why?
    - We need 3 (or 1) so that  $R_1$  cannot fire
    - With 3, we do not need 2, since with 4 and 6 fixed, then  $R_4$  is guaranteed to fire

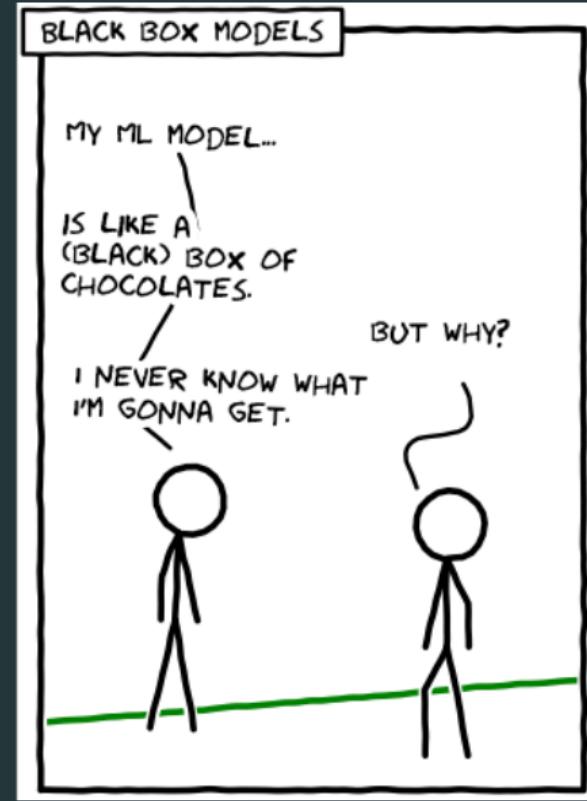
$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires
- What is an explanation for the prediction?
- Fixing  $\{3, 4, 6\}$  suffices for the prediction
  - Why?
    - We need 3 (or 1) so that  $R_1$  cannot fire
    - With 3, we do not need 2, since with 4 and 6 fixed, then  $R_4$  is guaranteed to fire
  - Some questions:
    - Would average human decision maker be able to understand the irreducible set  $\{3, 4, 6\}$ ?

$R_1 :$	IF	$(x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_2 :$	ELSE IF	$(x_2 \wedge x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_3 :$	ELSE IF	$(\neg x_1 \wedge x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_4 :$	ELSE IF	$(x_4 \wedge x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_5 :$	ELSE IF	$(\neg x_1 \wedge \neg x_3)$	THEN	$\kappa(\mathbf{x}) = 1$
$R_6 :$	ELSE IF	$(x_6)$	THEN	$\kappa(\mathbf{x}) = 0$
$R_{DEF} :$	ELSE			$\kappa(\mathbf{x}) = 1$

- Instance:  $((0, 1, 0, 1, 0, 1), 0)$ , i.e. rule  $R_2$  fires
- What is an explanation for the prediction?
- Fixing  $\{3, 4, 6\}$  suffices for the prediction
  - Why?
    - We need 3 (or 1) so that  $R_1$  cannot fire
    - With 3, we do not need 2, since with 4 and 6 fixed, then  $R_4$  is guaranteed to fire
  - Some questions:
    - Would average human decision maker be able to understand the irreducible set  $\{3, 4, 6\}$ ?
    - Would he/she be able to compute the set  $\{3, 4, 6\}$ , by manual inspection?

# Questions?



## References i

- [BBM<sup>+</sup>15] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek.  
**On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.**  
*PloS one*, 10(7):e0130140, 2015.
- [BHvMW09] Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors.  
**Handbook of Satisfiability**, volume 185 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2009.
- [EU21a] EU.  
**European Artificial Intelligence Act.**  
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>, 2021.
- [EU21b] EU.  
**European Artificial Intelligence Act – Proposal.**  
<https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>, 2021.
- [FJ18] Matteo Fischetti and Jason Jo.  
**Deep neural networks and mixed integer linear optimization.**  
*Constraints*, 23(3):296–309, 2018.

## References ii

- [HM23] Xuanxiang Huang and João Marques-Silva.  
**The inadequacy of Shapley values for explainability.**  
*CoRR*, abs/2302.08160, 2023.
- [HMS24] Xuanxiang Huang and Joao Marques-Silva.  
**On the failings of Shapley values for explainability.**  
*International Journal of Approximate Reasoning*, page 109112, 2024.
- [HRS19] Xiyang Hu, Cynthia Rudin, and Margo Seltzer.  
**Optimal sparse decision trees.**  
In *NeurIPS*, pages 7265–7273, 2019.
- [IMM18] Alexey Ignatiev, António Morgado, and João Marques-Silva.  
**PySAT: A python toolkit for prototyping with SAT oracles.**  
In *SAT*, pages 428–437, 2018.
- [IMS21] Yacine Izza and Joao Marques-Silva.  
**On explaining random forests with SAT.**  
In *IJCAI*, pages 2584–2591, July 2021.
- [JKMC16] Mikolás Janota, William Klieber, Joao Marques-Silva, and Edmund M. Clarke.  
**Solving QBF with counterexample guided refinement.**  
*Artif. Intell.*, 234:1–25, 2016.

## References iii

- [KBD<sup>+</sup>17] Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer.  
**Reluplex: An efficient SMT solver for verifying deep neural networks.**  
In *CAV*, pages 97–117, 2017.
- [Lip18] Zachary C. Lipton.  
**The mythos of model interpretability.**  
*Commun. ACM*, 61(10):36–43, 2018.
- [LL17] Scott M. Lundberg and Su-In Lee.  
**A unified approach to interpreting model predictions.**  
In *NIPS*, pages 4765–4774, 2017.
- [LPMM16] Mark H. Liffiton, Alessandro Previti, Ammar Malik, and Joao Marques-Silva.  
**Fast, flexible MUS enumeration.**  
*Constraints*, 21(2):223–250, 2016.
- [LS08] Mark H. Liffiton and Karem A. Sakallah.  
**Algorithms for computing minimal unsatisfiable subsets of constraints.**  
*J. Autom. Reasoning*, 40(1):1–33, 2008.
- [Mar22] João Marques-Silva.  
**Logic-based explainability in machine learning.**  
In *Reasoning Web*, pages 24–104, 2022.

## References iv

- [Mar24] Joao Marques-Silva.  
**Logic-based explainability: Past, present & future.**  
*CoRR*, abs/2406.11873, 2024.
- [MHL<sup>+</sup>13] António Morgado, Federico Heras, Mark H. Liffiton, Jordi Planes, and Joao Marques-Silva.  
**Iterative and core-guided MaxSA solving: A survey and assessment.**  
*Constraints*, 18(4):478–534, 2013.
- [MI22] João Marques-Silva and Alexey Ignatiev.  
**Delivering trustworthy AI through formal XAI.**  
In *AAAI*, pages 12342–12350, 2022.
- [Mil19] Tim Miller.  
**Explanation in artificial intelligence: Insights from the social sciences.**  
*Artif. Intell.*, 267:1–38, 2019.
- [MM20] João Marques-Silva and Carlos Mencía.  
**Reasoning about inconsistent formulas.**  
In *IJCAI*, pages 4899–4906, 2020.

## References v

- [Mol20] Christoph Molnar.  
*Interpretable machine learning.*  
Lulu.com, 2020.  
<https://christophm.github.io/interpretable-ml-book/>.
- [MS23] Joao Marques-Silva.  
**Disproving XAI myths with formal methods – initial results.**  
In *ICECCS*, 2023.
- [MSH24] Joao Marques-Silva and Xuanxiang Huang.  
**Explainability is Not a game.**  
*Commun. ACM*, 67(7):66–75, jul 2024.
- [MSI23] Joao Marques-Silva and Alexey Ignatiev.  
**No silver bullet: interpretable ml models must be explained.**  
*Frontiers in Artificial Intelligence*, 6, 2023.
- [NH10] Vinod Nair and Geoffrey E. Hinton.  
**Rectified linear units improve restricted boltzmann machines.**  
In *ICML*, pages 807–814, 2010.

## References vi

- [PG86] David A. Plaisted and Steven Greenbaum.  
**A structure-preserving clause form translation.**  
*J. Symb. Comput.*, 2(3):293–304, 1986.
- [RCC<sup>+</sup>22] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong.  
**Interpretable machine learning: Fundamental principles and 10 grand challenges.**  
*Statistics Surveys*, 16:1–85, 2022.
- [RSG16] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.  
"why should I trust you?": Explaining the predictions of any classifier.  
In *KDD*, pages 1135–1144, 2016.
- [RSG18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.  
**Anchors: High-precision model-agnostic explanations.**  
In *AAAI*, pages 1527–1535. AAAI Press, 2018.
- [Rud19] Cynthia Rudin.  
**Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.**  
*Nature Machine Intelligence*, 1(5):206–215, 2019.

## References vii

- [Rud22] Cynthia Rudin.  
**Why black box machine learning should be avoided for high-stakes decisions, in brief.**  
*Nature Reviews Methods Primers*, 2(1):1–2, 2022.
- [Tse68] G.S. Tseitin.  
**On the complexity of derivations in the propositional calculus.**  
In H.A.O. Slesenko, editor, *Structures in Constructive Mathematics and Mathematical Logic, Part II*, pages 115–125, 1968.