UNISS
UNIVERSITÀ
DEGLI STUDI
DI SASSARI

# Trustworthy AI:

# Verification of Neural Networks

*Dario Guidotti*

CPS Summer School – Alghero, 25/09/2025

*This activity is funded by the EU funded projects SECURED*
*"Scaling Up Secure Processing, Anonymization And Generation Of Health Data For EU Cross Border Collaborative Research And Innovation"*
*with Grant Agreement no. 101095717.*

SECURED
EU Project

Dario Guidotti
University of Sassari

# AIMet Lab

**Luca Pulina**
*Full Professor @UNISS*

**Laura Pandolfo**
*Assistant Professor @UNISS*

**Dario Guidotti**
*Assistant Professor @UNISS*

- Research in **AI**, **Knowledge Representation & Reasoning**, **Formal Methods**
- Application domains: *safety-critical systems*, *industry 4.0*, *digital libraries*, *environmental monitoring*, *tourism & cultural heritage*

Dario Guidotti
University of Sassari

# Human Intelligence vs Artificial Intelligence

## Human Intelligence

*The ability to perceive or deduce information, retain it as knowledge, and apply it to adaptive behaviour.*

**Key aspects**:

*Reasoning*: analyse information and draw logical conclusions
*Learning*: acquire knowledge and skills through experience
*Adaptation*: adjust behaviour to environmental changes

**Types**: Logical, Emotional, Social

## Artificial Intelligence

*A branch of computer science creating systems that perform tasks requiring human-like intelligence.*

**Key objectives:**

*Automation*: perform tasks without direct human intervention

*Simulation of human abilities*: understand language, speak, recognise images, solve problems

Dario Guidotti
University of Sassari

# Weak AI vs Strong AI

## Weak AI

- Designed for specific tasks
- Appears intelligent but lacks real understanding or awareness
- Relies on data and pattern recognition
- *Current real-world systems* (e.g. speech recognition, recommendation engines)

## Strong AI

- *Theoretical* concept of human-level intelligence and consciousness
- Would understand its actions and make autonomous decisions
- *Not yet realised -- considered a long-term research goal*
- Hypothetical example: a robot able to perform any intellectual task as a human

Dario Guidotti
University of Sassari

# Inductive AI vs Deductive AI

## Inductive AI

- *Learns* from data and examples
- *Identifies* patterns and *generalises*
- **Strengths**: powerful when large datasets are available
- **Limits**: needs lots of data; may fail if patterns change

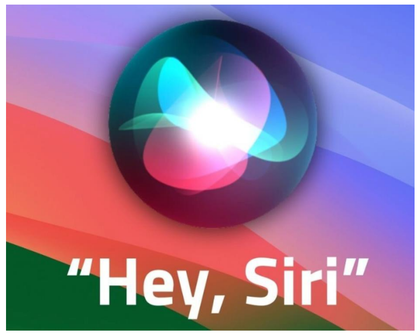
## Deductive AI

- Starts from *rules* and *principles*
- Applies logic to draw conclusions
- **Strengths**: works without big datasets; effective with clear rules
- **Limits**: restricted to well-defined problems; less effective in complex or ambiguous contexts

Dario Guidotti
University of Sassari

# AIs in everyday life

## Voice Assistants

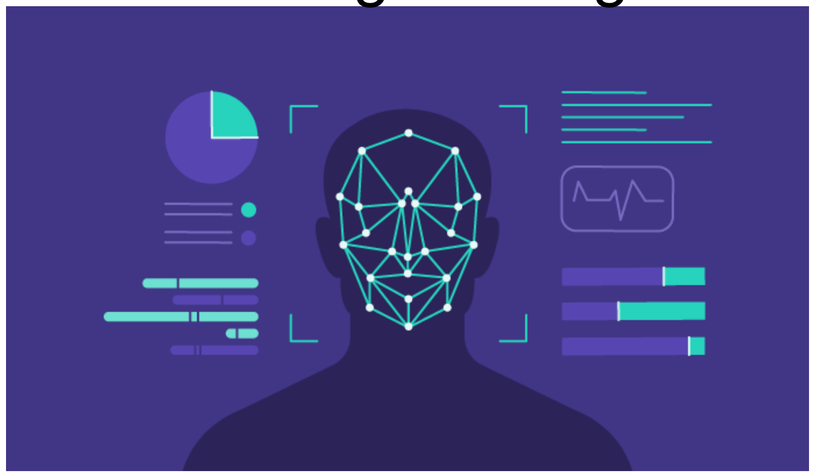




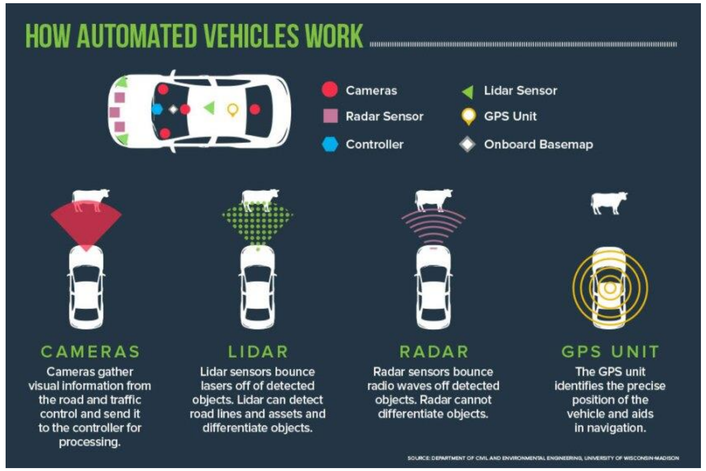
## Recommender Systems



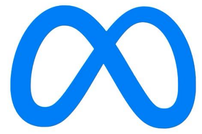
## Face/image Recognition



## Self-driving vehicles




## Chat-bots



MISTRAL AI_

Dario Guidotti
University of Sassari

# Why focus on Weak, Inductive AI

- **Extremely Popular**: many real-world AI systems belong to this category (data-driven).
- **Ubiquitous**: from everyday services to safety- and security-critical domains (*e.g.*, recommender systems but also automotive, medical imaging, and other domains)
- **High stakes**: failures can have severe consequences for people, organisations, and society.
- **Hard to certify**: their decisions are often opaque, and reliability is difficult to demonstrate beyond empirical performance on data.

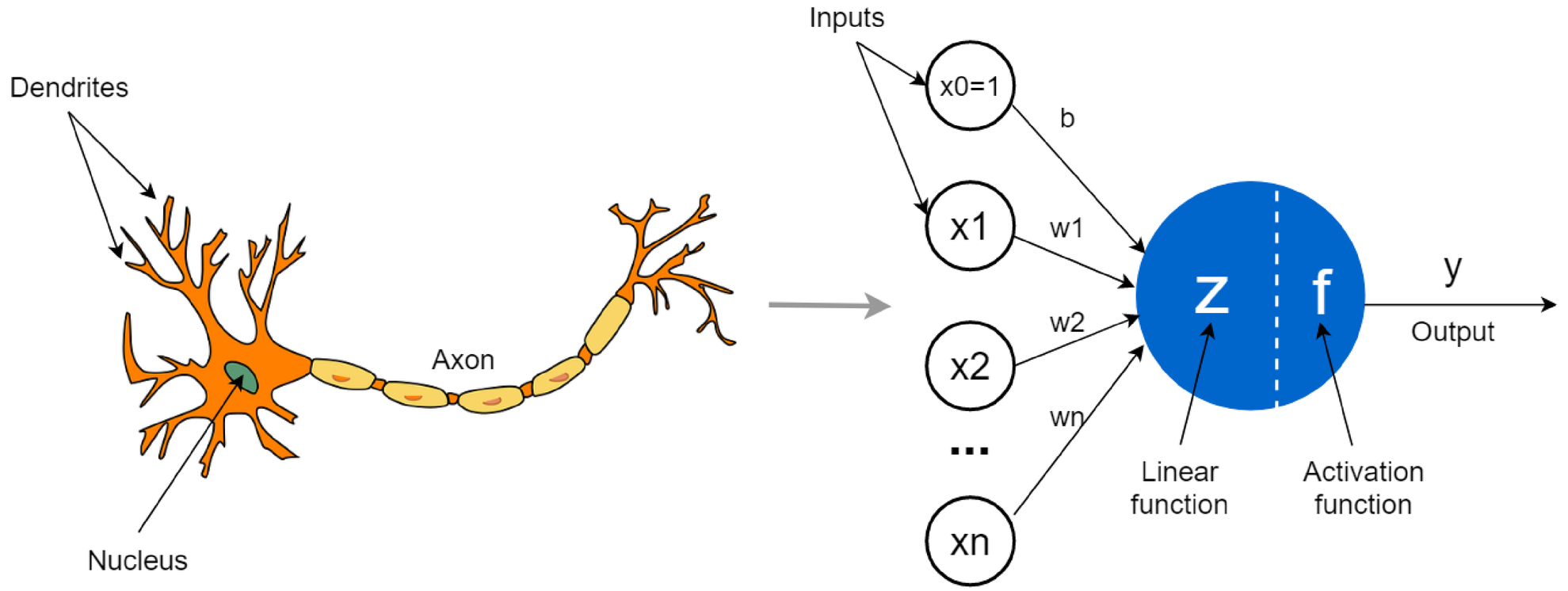
Dario Guidotti
University of Sassari

# What Are Neural Networks and Why Are They Popular?

- **Inspired by the brain**: networks of interconnected “neurons” that process signals and transmit activations.
- **Data-driven learning**: they adjust their internal connections (weights) to recognise patterns and make predictions.
- **Universal approximators**: capable of modelling complex, non-linear relationships between inputs and outputs.
- **Scalability**: with enough data and computational power, they outperform many traditional AI methods.
- **Versatility**: applied successfully in vision, speech, language, recommendation, robotics, and more.
- **Popularity boost**: rapid progress in recent years thanks to big data, GPUs, and improved training techniques.

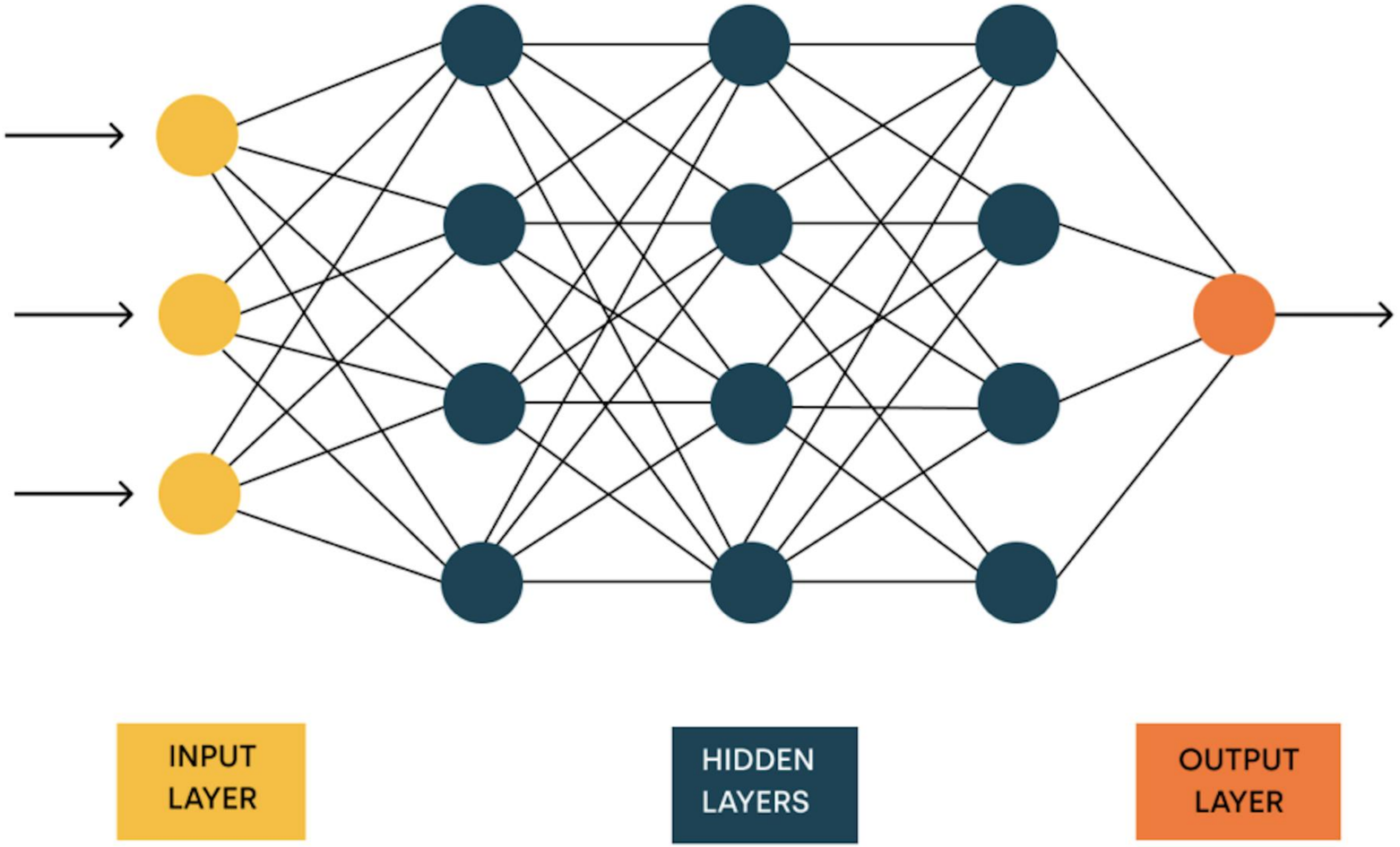
Dario Guidotti
University of Sassari

# Real Neuron vs Artificial Neuron



Dario Guidotti
University of Sassari

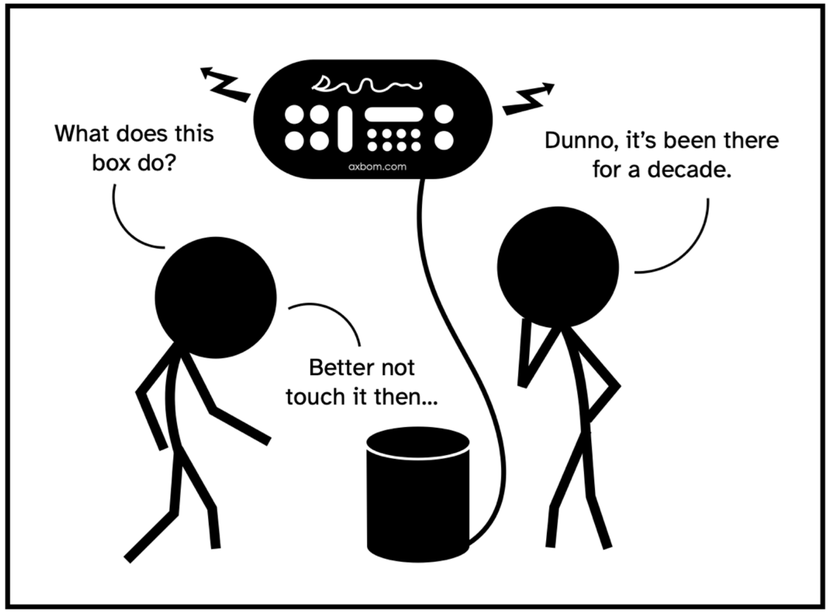
# Generic Architecture

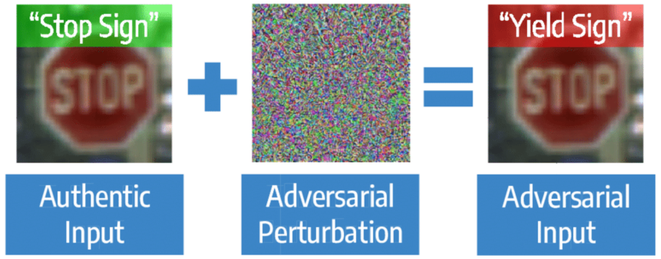


Dario Guidotti
University of Sassari

# Why Reliability & Trustworthiness Matter for Neural Networks

- **Low explainability**: decisions arise from millions of parameters, making causes of errors hard to trace and certify.
- **Adversarial perturbations**: tiny, often imperceptible input changes can flip predictions, with security and safety implications.
- **Overconfidence & poor calibration**: models may be confidently wrong, undermining risk-aware decisions in the real world.
- **System integration**: model outputs feed larger pipelines and human workflows, where small model errors can cascade.



Dario Guidotti
University of Sassari

# Building Trustworthy AI: Many Approaches

- **Explainability**: making model behaviour interpretable to humans.
- **Safe Training**: designing and training models which are robust from the start.
- **Model Repair**: correcting faulty behaviours after training.
- **Synthetic Data Evaluation**: changing the focus to validating the data produced instead of the models.
- **Neuro-Symbolic AI**: combining learning with reasoning for stronger guarantees.
- **Verification**: formally proving properties hold in all possible cases.

Dario Guidotti
University of Sassari

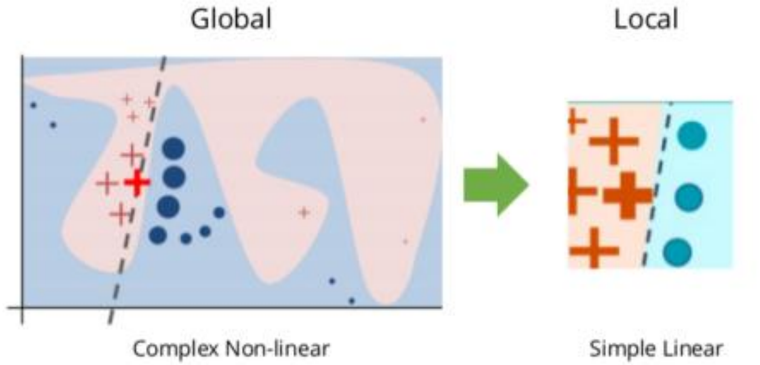
# Explainability

*Explainability is the ability to understand and interpret the decision-making process of a neural network or another "black box" model. It aims to provide insights into why a model produces a given output.*

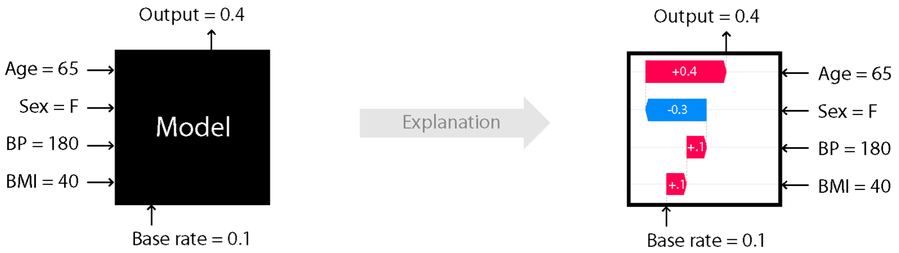
- Builds trust between humans and AI systems.
- Essential for accountability in high-stakes domains (medicine, security, finance).
- Helps in debugging and improving model performance.
- *Methodologies*:
  - ❖ **LIME**: Creates local surrogate (explainable) models to approximate decisions.
  - ❖ **SHAP**: Quantifies contribution of each input feature to the output.
- *Challenges*: current explanations are **often approximate**, may **vary across methods**, and **can be misleading** if interpreted as full transparency.

Dario Guidotti
University of Sassari

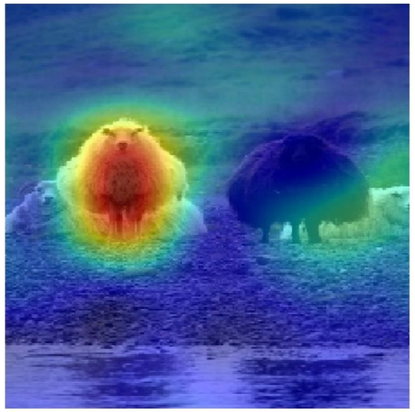
# Explainability: some visual examples



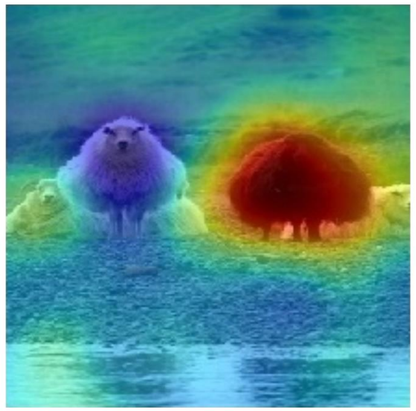

SHAP




(a) Sheep - 26%, Cow - 17%



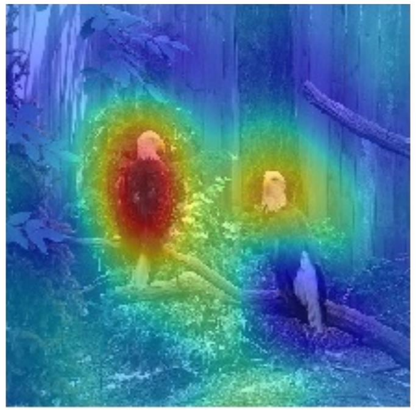
(b) Importance map of '*sheep*'



(c) Importance map of '*cow*'

(d) Bird - 100%, Person - 39%

(e) Importance map of '*bird*'



(f) Importance map of '*person*'

Dario Guidotti
University of Sassari

# Safe Training

*Safe training is the design of learning procedures that aim to make neural networks more robust, fair, and reliable from the start.*

- Reduces vulnerability to adversarial perturbations.
- Mitigates bias and unfair outcomes in sensitive applications.
- Enhances generalisation and stability in real-world environments.
- *Methodologies*:
  - ❖ **Adversarial training:** incorporate adversarial examples during training to enhance model's robustness.
  - ❖ **Fairness-aware training:** include fairness objectives or constraints in optimisation.
- *Challenges*: Even advanced safe training techniques **cannot give guarantee** regarding the correct behaviour of the models of interest.

Dario Guidotti
University of Sassari

# Model Repair

*Model repair refers to techniques that modify a trained neural network to correct undesired or faulty behaviours without retraining from scratch.*

- Fixes critical errors discovered after deployment.
- Saves time and resources compared to full retraining.
- Allows adaptation to new requirements or constraints.
- *Methodologies*:
  - **Weight modification:** directly adjust specific parameters to enforce desired outputs.
  - **Local patching:** apply targeted updates that affect only problematic regions of the input space.
- *Challenges*: repairs **may not generalise**, risk **introducing new errors**, and can be **difficult to scale** to large models.

Dario Guidotti
University of Sassari

# Synthetic Data Evaluation

*The process of assessing the quality, fidelity, and utility of data generated by synthetic data models.*

- Ensures synthetic data preserves the key statistical properties of the real data.
- Critical in sensitive domains where data quality directly affects downstream analysis.
- Supports safe adoption of synthetic datasets for training, evaluation, and sharing when real data are limited or private.
- *Methodologies*:
  - ❖ **Statistical similarity measures**: compare distributions of real vs. synthetic data.
  - ❖ **Feature-based metrics**: use embeddings from pretrained models to evaluate realism and diversity.
- *Challenges*: **High similarity** to real data may **risk privacy leaks**, while **too much divergence reduces usefulness** for real tasks.

Dario Guidotti
University of Sassari

# Synthetic Data Evaluation in Practice: The SynthVal Framework

**The SECURED Project**

- EU-funded initiative creating the SECURED Innohub, a trusted collaboration hub for health data.
- Provides tools for secure multi-party computation, data anonymization, and synthetic data generation.
- Aims to enable cross-border health data collaboration, supporting research, SMEs, and innovators.
- Promotes synthetic data to boost health data availability, reduce bias, and support medical education.


**The Role of SynthVal**

- *SynthVal* was developed within SECURED to evaluate the quality of synthetic health data, focusing on medical images.
- Combines feature extraction (radiology-informed + foundation models) with evaluation metrics.
- Applied to mammography datasets, showing how evaluation ensures trustworthy, safe AI adoption in healthcare.

Dario Guidotti
University of Sassari

# Neuro-symbolic AI

*Neuro-Symbolic AI is the integration of neural networks with symbolic reasoning systems. It aims to combine the perception power of deep learning with the interpretability and logical structure of symbolic AI.*

- Enables reasoning over complex data while maintaining explainability.
- Bridges the gap between pattern recognition and knowledge representation.
- Supports robust decision-making in high-stakes domains (medicine, autonomous driving, cultural heritage).
- *General approach*:
  - **Neural modules** for perception (vision, language, speech).
  - **Symbolic modules** for rules, ontologies, and logical inference.
  - **Hybrid training strategies** that align subsymbolic embeddings with symbolic constraints.
- *Challenges*: **integration** remains complex; symbolic reasoning is **hard to scale**; neural components may still introduce **opacity**; lack of standard frameworks for **seamless fusion**.

Dario Guidotti
University of Sassari

# Neuro-symbolic AI in Practice: CCN+ Framework

**Problem**: Neural networks find hidden patterns but often fail to comply with explicit requirements (rules, domain knowledge).

**Solution**: CCN+ integrates propositional logic requirements directly into the neural network's output layer. Ensures predictions are compliant by design. Adapts the loss function to enforce logical constraints.

**Benefits**: Combines statistical learning with symbolic reasoning. Produces outputs that respect real-world constraints (safety, regulations). Improves performance by leveraging domain knowledge.

**Tested on**: Autonomous driving dataset with logical safety rules (e.g., "if stop sign detected → no acceleration"). CCN+ outperforms standard deep learning and state-of-the-art baselines.

CCN+: A neuro-symbolic framework for deep learning with requirements – Giunchiglia et al. – International Journal of Approximate Reasoning, 2024

Dario Guidotti
University of Sassari

# Verification

*Verification is the process of formally ensuring that an AI system (often a neural network) behaves according to specified requirements or constraints.*

- Ensures compliance with domain-specific requirements.
- Critical for safety in high-stakes applications (autonomous driving, healthcare, cybersecurity).
- Provides guarantees beyond empirical testing.
- **Challenges**: current verification techniques can be computationally expensive, may not scale to very large networks, and often require simplifying assumptions.

Dario Guidotti
University of Sassari

# NN Verification: a more formal definition

*Given a Neural Network* $f\colon \mathbb{R}^n \rightarrow \mathbb{R}^m$*, an Input Sub-Domain* $\mathcal{D} \subset \mathbb{R}^n$ *and an Output Sub-Domain* $\mathcal{O} \subset \mathbb{R}^m$*, we want to prove that:*

$$\forall x \in \mathcal{D}, \qquad y = f(x) \rightarrow y \in \mathcal{O}$$

*To avoid the use of universal quantifier the problem is usually reframed as its negation:*

$$\nexists x \in \mathcal{D}, \qquad y = f(x) \rightarrow y \notin \mathcal{O}$$

In the case of **Robustness to Adversarial Perturbation**:

$$\nexists \mathrm{x} \quad \mathrm{s.t.} \quad \|\mathrm{x} - \hat{\mathrm{x}}\|_\infty < \varepsilon, \qquad \mathrm{y} = \mathrm{f}(\mathrm{x}) \rightarrow \mathrm{y} \neq \hat{\mathrm{y}}$$

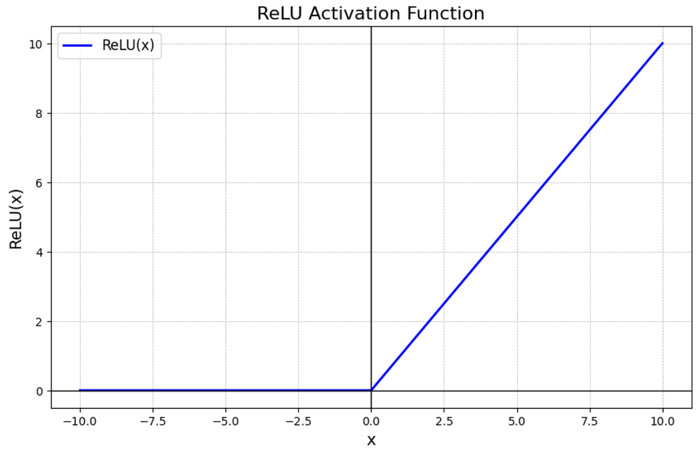
Where $\hat{\mathrm{x}}$ and $\hat{\mathrm{y}}$ are, respectively, the original input image and its assigned class.

Dario Guidotti
University of Sassari

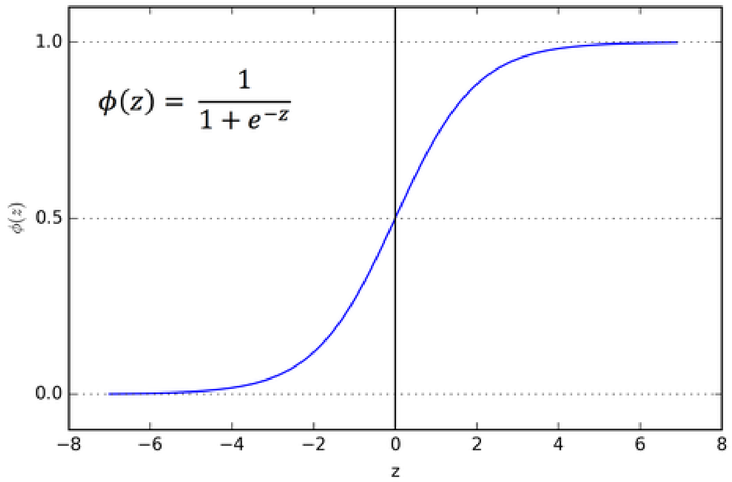
# NN Verification: why is it so challenging?

***In one word: non-linear activation functions!***

- If f(x) was a linear, convex function, the problem could be easily and efficiently solved using LP!
- Unfortunately, for NNs to be Universal Approximators we need at the very least piece-wise linear activation functions (*e.g.*, ReLU, Sigmoid)!




y = max(0, x)



Dario Guidotti
University of Sassari

# How can we manage non-linearity?

## Complete Approach

- We divide the problem in linear sub-problem
- Feasible only for piece-wise linear activation functions
- Exponential complexity, very limited scalability
- Can certifiably prove if a ***Neural Network is safe <u>or not</u>***

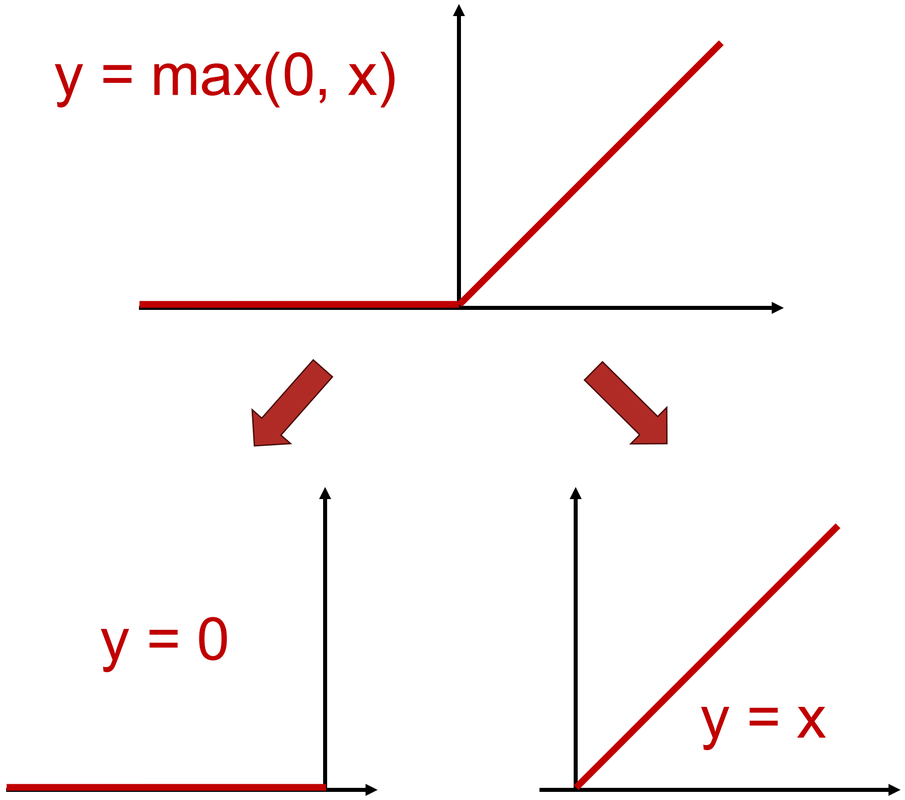
## Incomplete Approach

- We use a linear over-approximation of the activation
- Can be used for piece-wise linear, non-linear, transcendent
- Scalability and complexity can be tuned
- Can certifiably prove only if a ***Neural Network is safe***

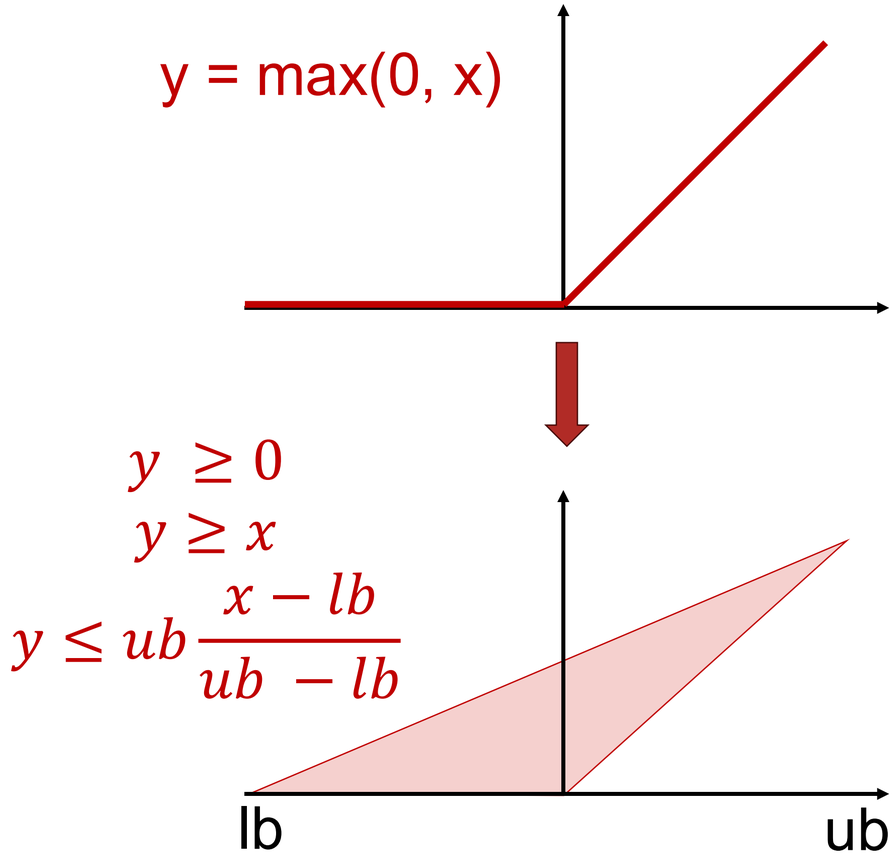
Dario Guidotti
University of Sassari

# A Graphical Example

## Complete

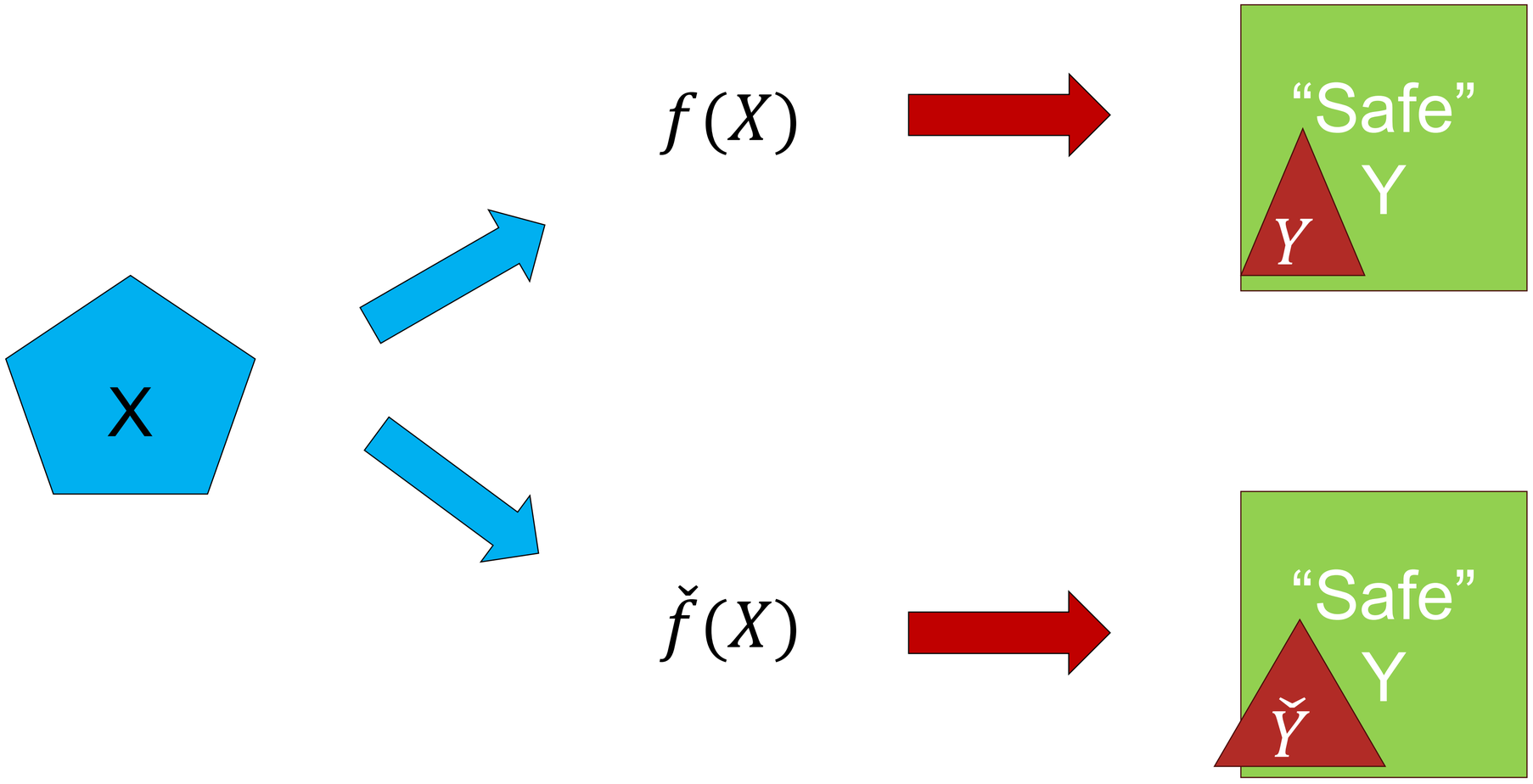



## Incomplete



Dario Guidotti
University of Sassari

# Why Incomplete cannot prove «unsafety»?



Dario Guidotti
University of Sassari

# Useful resources: ONNX, VNNLIB, and VNNCOMP

- **ONNX** is the standard format for exchanging neural networks
  - Provides a flexible graph structure that can represent most architectures
- **VNNLIB** is the standard for sharing neural network verification benchmarks
  - Uses .vnnlib files (SMT-LIB style) to define input/output specifications
  - Networks themselves are provided in ONNX format
- **VNNCOMP** is the international competition on neural network verification
  - First held in 2020, now a yearly event with growing participation
  - Typically co-located with the CAV conference

Dario Guidotti
University of Sassari

# State of the art: VNNCOMP 2024

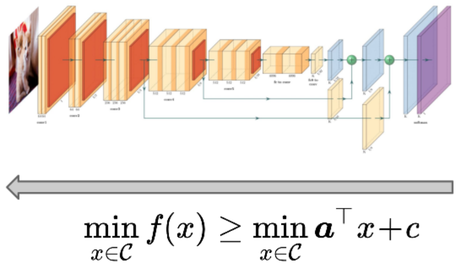
| Name | Institution | Open Source |
|---|---|---|
| alpha-beta-CROWN | UIUC, CMU, UCLA, Drexel, Columbia, RWTH Aachen University | yes |
| CORA | TU München | yes |
| Marabou | Stanford | yes |
| NeVer2 | University of Genoa, University of Sassari | yes |
| NeuralSAT | GMU | yes |
| nnenum | Stony Brook University | yes |
| NNV | Vanderbilt, University of Nebraska-Lincoln | yes |
| PyRAT | Université Paris-Saclay, CEA, List | no |

| # | Tool | Score |
|---|---|---|
| 1 | $\alpha$-$\beta$-CROWN | 1200.0 |
| 2 | PyRAT | 1000.8 |
| 3 | Marabou | 751.0 |
| 4 | nnenum | 572.5 |
| 5 | NNV | 530.0 |
| 6 | NeVer2 | 262.3 |
| 7 | CORA | 251.7 |
| 8 | NeuralSAT | 0 |

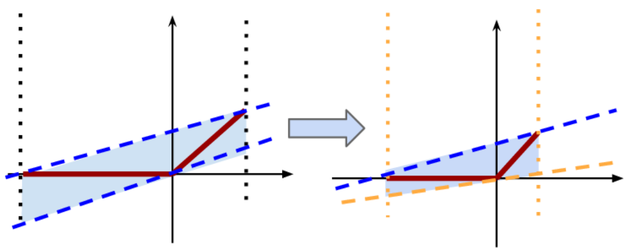
| Network Type | # Parms | Input Dim |
|---|---|---|
| FC + ReLU | 4k | 30 |
| Conv, FC, Residual + ReLU, Sigmoid | 33k - 37M | 1-308 |
| FC + ReLU | 0.5M - 1.1M | 784-3072 |
| FC, Residual + ReLU | 203k | 4 |
| FC + ReLU, Sigmoid | 342k-855k | 792 |
| FC, Conv, Residual + ReLU, BatchNorm | 2.5M - 3.8M | 3072 |
| FC, Conv, Residual + ReLU, BatchNorm | 3.6M | 9408 |
| FC + ReLU | 13k | 5 |
| FC, Conv, ConvTranspose, Residual + ReLU, BatchNorm, AvgPool | 500k-68M | 5 |
| Conv + ReLU, Dropout | 60k - 262k | 400-800 |
| Conv, FC + ReLU | 466k-7.4M | 5376 |
| FC + ReLU | 17k - 67M | 2 |

Dario Guidotti
University of Sassari

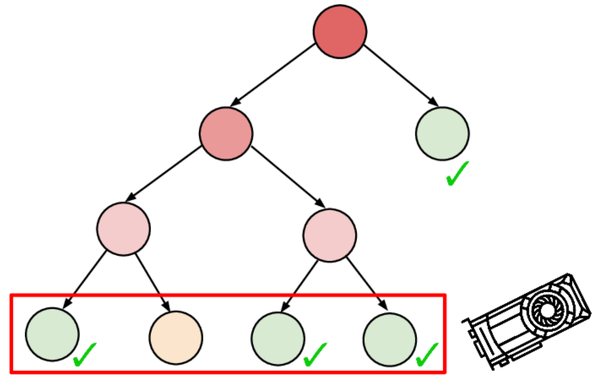
# State of the art: α-β-Crown




Efficient bound propagation (**CROWN**)



GPU optimized relaxation (***α*-CROWN**)

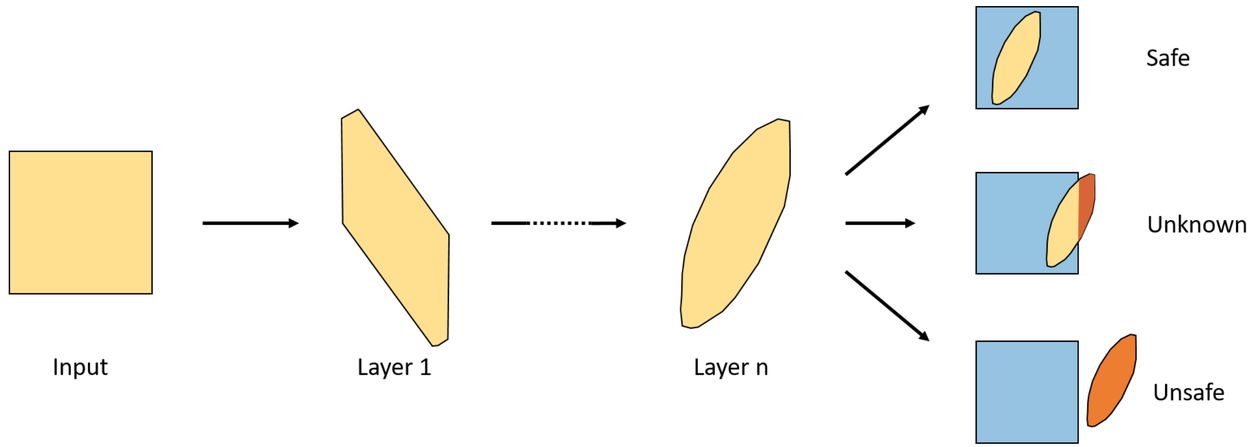


Parallel branch and bound (***β*-CROWN**)

- Uses linear bound propagation with branch-and-bound, accelerated on GPUs for scalability to large models.
- Handles CNNs, ResNets, and diverse activation functions via the auto_LiRPA library.
- Provides certified robustness against adversarial attacks and verifies properties like stability in control.

Dario Guidotti
University of Sassari

# State of the art: PyRat

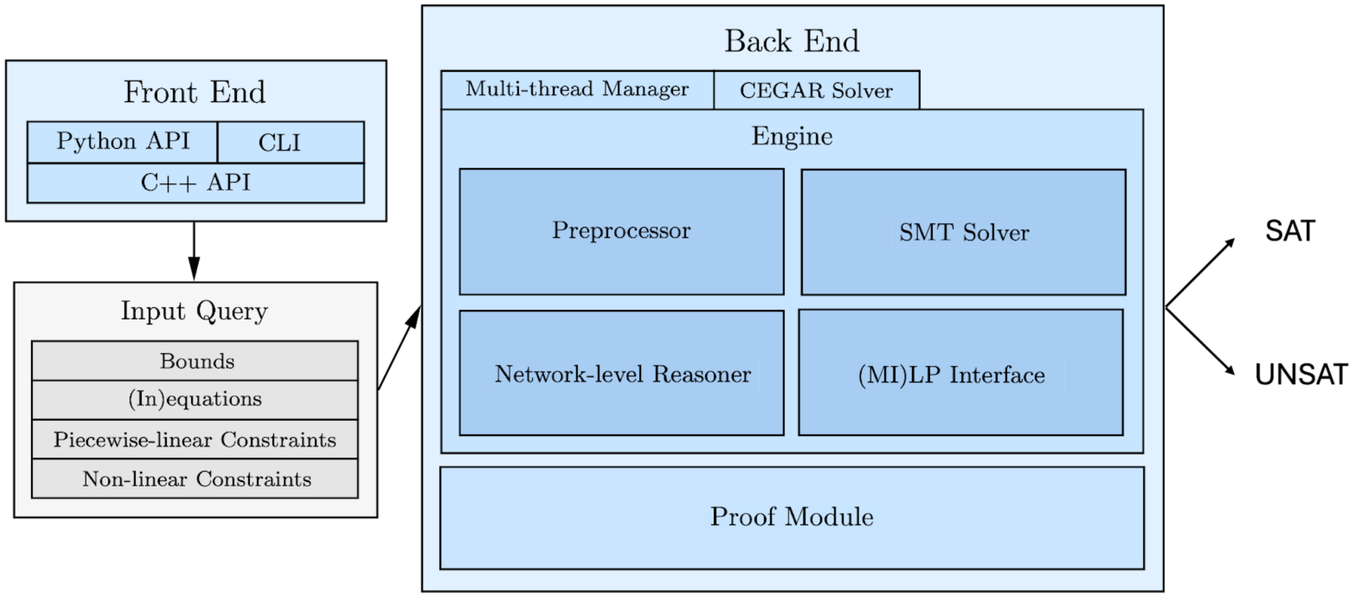



PyRAT

- An abstract interpretation based verifier that checks whether certain states can be reached in a neural network.
- Propagates abstract domains (Box, Zonotopes, Polyhedra, etc.) through layers, supporting DNNs, CNNs, and residual architectures with multiple activation functions.
- Provides conservative guarantees: if the abstract output fits the desired property, the network is verified; otherwise, the result is inconclusive.

Dario Guidotti
University of Sassari

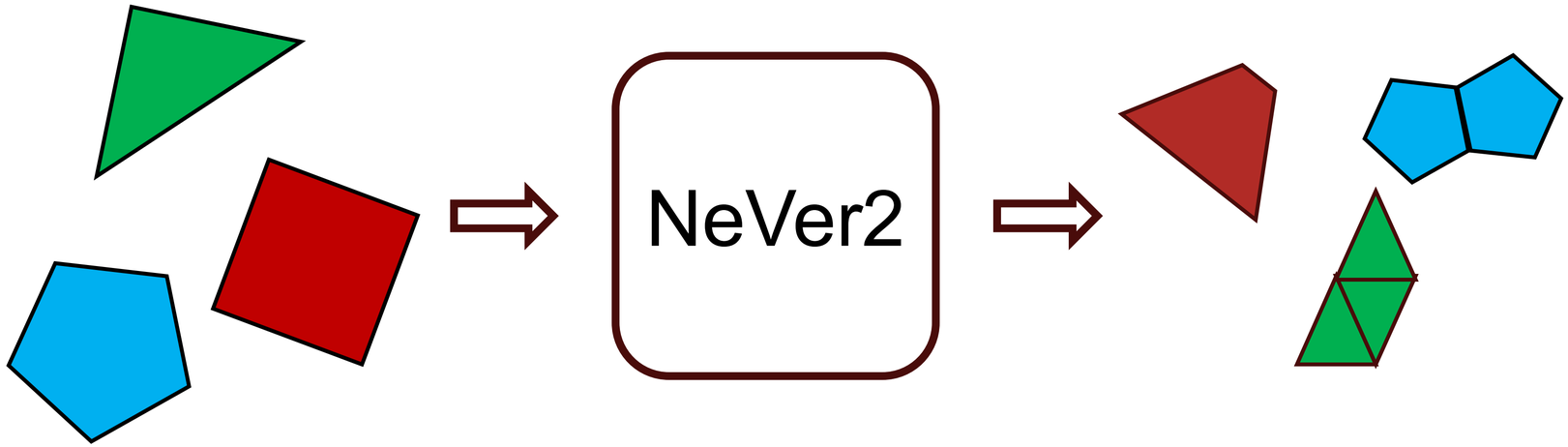
# State of the art: Marabou




- An SMT-based verifier that checks whether a neural network satisfies a given property, or produces a counterexample if violated.
- Supports multiple formats (ONNX, .nnet, TensorFlow) and feed-forward networks with various activation functions.
- Properties can be defined via a Python interface or in the VNNLIB format.

Dario Guidotti
University of Sassari

# State of the art: NeVer2




- Abstract-interpretation–based verifier using Star Sets and a search algorithm.
- Supports ONNX, PyTorch, and VNNLIB property specifications.
- Provides a Python API for design, learning, and verification of feed-forward NNs with ReLU activations.
- Freely available at https://github.com/NeVerTools

Dario Guidotti
University of Sassari

# NeVer2: Starset and Affine Transformation

- Leverages the Star-set abstract domain: as expressive as Convex Polytope but less computationally expensive for affine transformation (FC layers) in neural networks.
- Able to represent any kind of convex domain as:

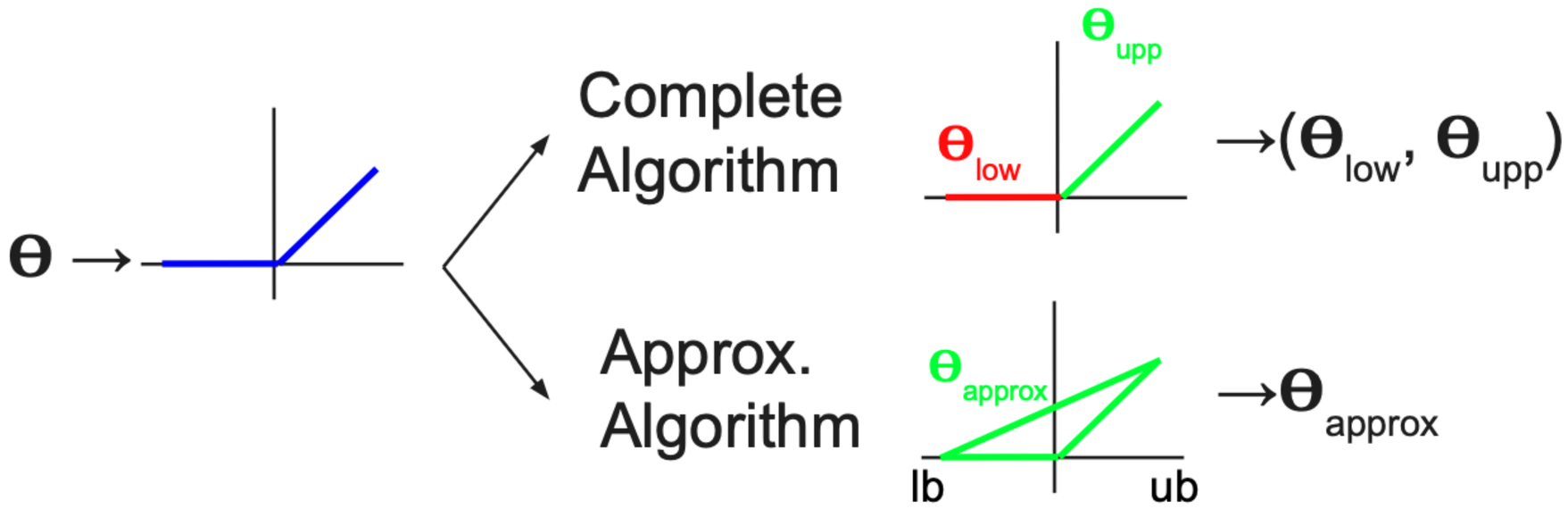
$$\boldsymbol{\theta} = (c, V, B, d) = \{z \in \mathbb{R}^n \mid z = Vx + c \quad s.t. \quad Bx \leq d\}$$

- Easy affine transformation $f(\boldsymbol{\theta}) = A\boldsymbol{\theta} + b$ as:

$$\widehat{\boldsymbol{\theta}} = \left(\hat{c}, \hat{V}, B, d\right) \text{ with } \hat{c} = Ac + b \text{ and } \hat{V} = AV$$

Dario Guidotti
University of Sassari

# NeVer2: ReLU




- If we want complete we split the stars in two adding constraints for x < 0 and x > 0 to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}_{low}$ and $\boldsymbol{\theta}_{upp}$
- Otherwise we add 3 constraints to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}_{approx}$
- P.s. to do the same two operations for polytope we would need to switch between V-representation and H-representation! Very computationally expensive!

Dario Guidotti
University of Sassari

# NeVer2: Applications

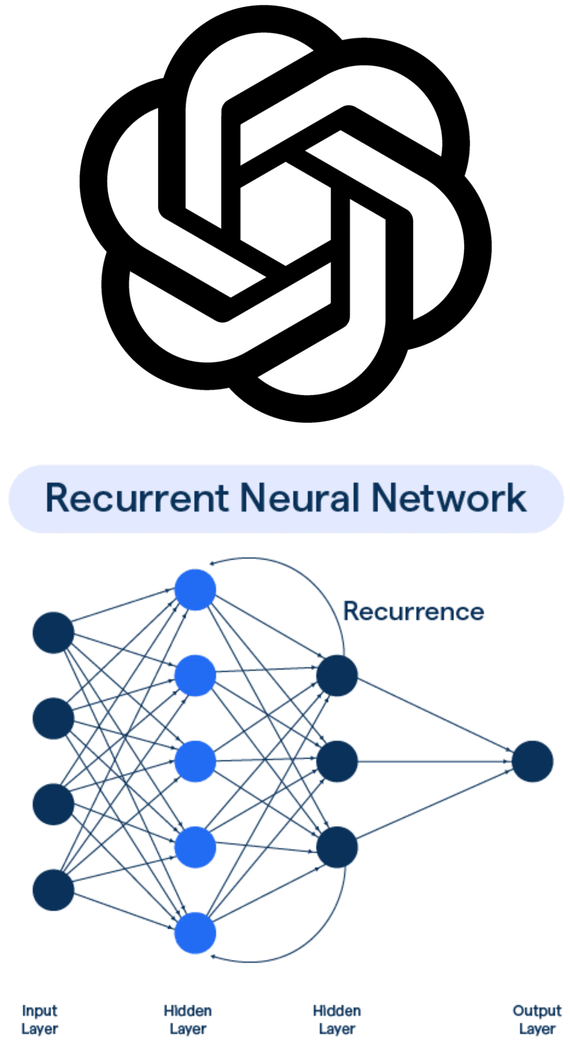
- **AIDOaRt**: H2020-ECSEL project focused on developing an AI-driven framework to enhances productivity, quality, and predictability in the continuous engineering of Cyber-Physical and large industrial systems.
  - **NeVer2** was used to design, train and verify neural networks for image classification and object detection in the automotive domain in collaboration with Abinsula

- **IMOCO4.E**: H2020-ECSEL project focused on enhancing energy-efficient, secure, and resilient Industry 4.0 manufacturing systems across sectors such as semiconductors, robotics, and healthcare.
  - **NeVer2** was again used as pipeline to produce high-performance and safe neural networks for anomaly detection and recognition in elevator controls system (with WEG and Siemens), and in industrial packaging machines (with CRIT Research).

Dario Guidotti
University of Sassari

# Open Challenges

- **Scalability**: what about multi-billion parameters models?
- **Architectures**: what about RNN, LSTM, Transformers, etc?
- **Properties**: what about more complex properties? Global Robustness, Stability, etc.



UNISS
UNIVERSITÀ
DEGLI STUDI
DI SASSARI

# Thank you for your attention!

# Questions?

Contact: dguidotti@uniss.it

*This activity is funded by the EU funded projects SECURED*
*"Scaling Up Secure Processing, Anonymization And Generation Of Health Data For EU Cross Border Collaborative Research And Innovation"*
*with Grant Agreement no. 101095717.*

SECURED
EU Project