CertifAI: certification of ML-based systems

16/11/2023

ANITI

Université de Toulouse

Members of the chair

ONERA

Claire
Pagetti
(COVNI)

Kevin
Delmas
(RIME)

Charles
Lesire
(SEAS)

LAAS

Jérémie
Guiochet

IRIT

Thomas
Carle

Airbus

Mélanie
Ducoffe
(plan de relance)

Adrien
Gauffriau
(MAD)

CS

Mohammed
Belcaid (MAD)

Thésards ANITI

Iryna
De Albuquerque

Iban
Guinebert

Noémie
Cohen

Anthony
Faure-Gignoux

Post doc ANITI

Joris Guérin
(2020 – 2022) now CR @IRD

Stagiaires
M1/M2
9

Lightening talks

ANITI
Université de Toulouse

Scope – certification

Certification

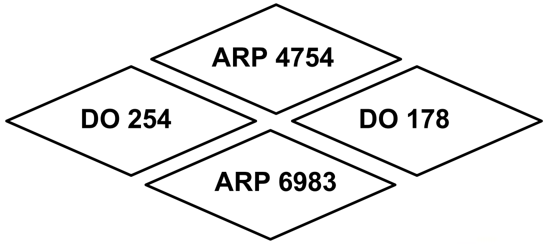
- evaluation of an **argumentation**, to convince that a system (i.e., its architecture, its settings, including mitigation means. . .) is compliant with the regulatory requirements
- accepted mean of compliance with the requirements is to rely on **mature standards**

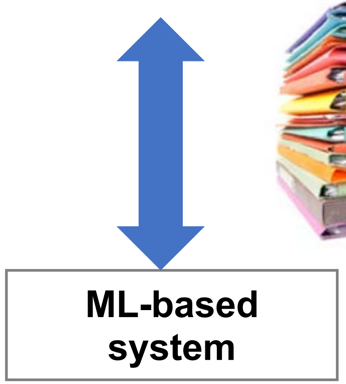
Applicative scope:

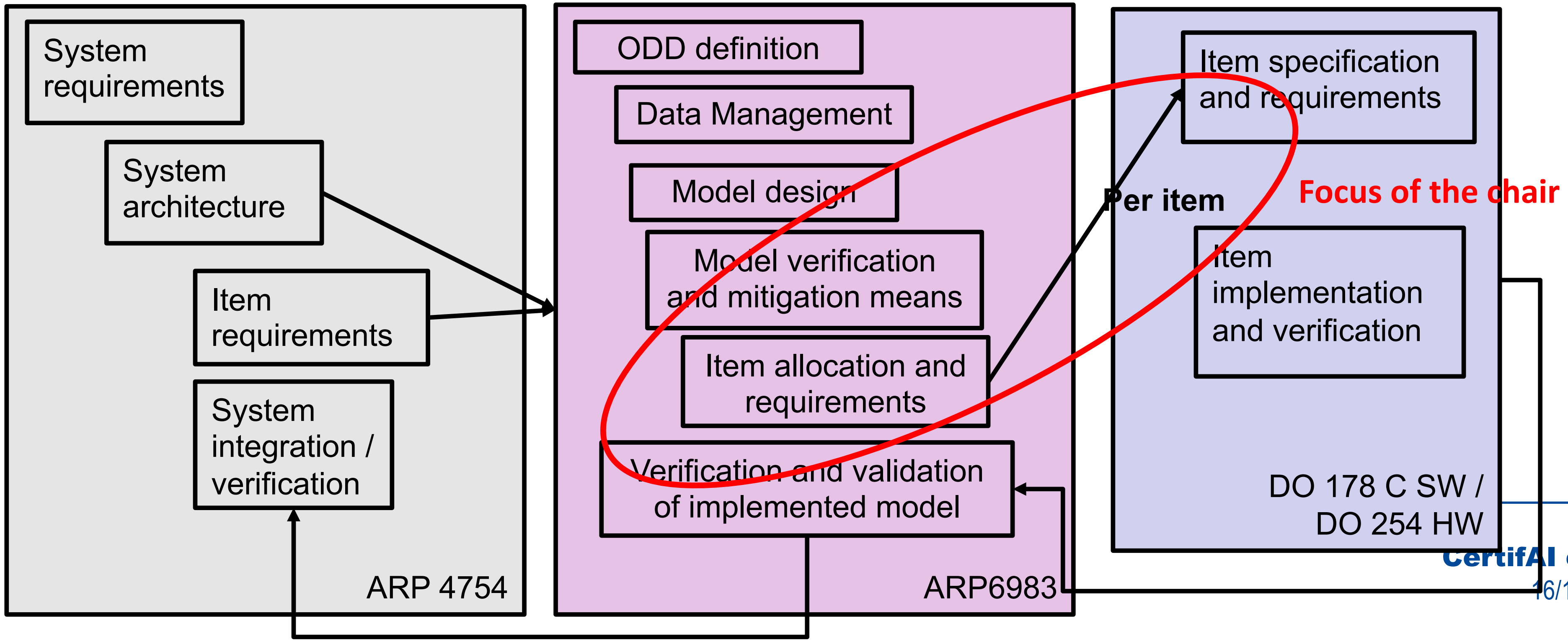
- ML (Machine Learning) based systems

Active contribution to

- DEEL mission certif
- EUROCAE/FAA ED 324 / ARP 6983 (more particularly on the implementation section)

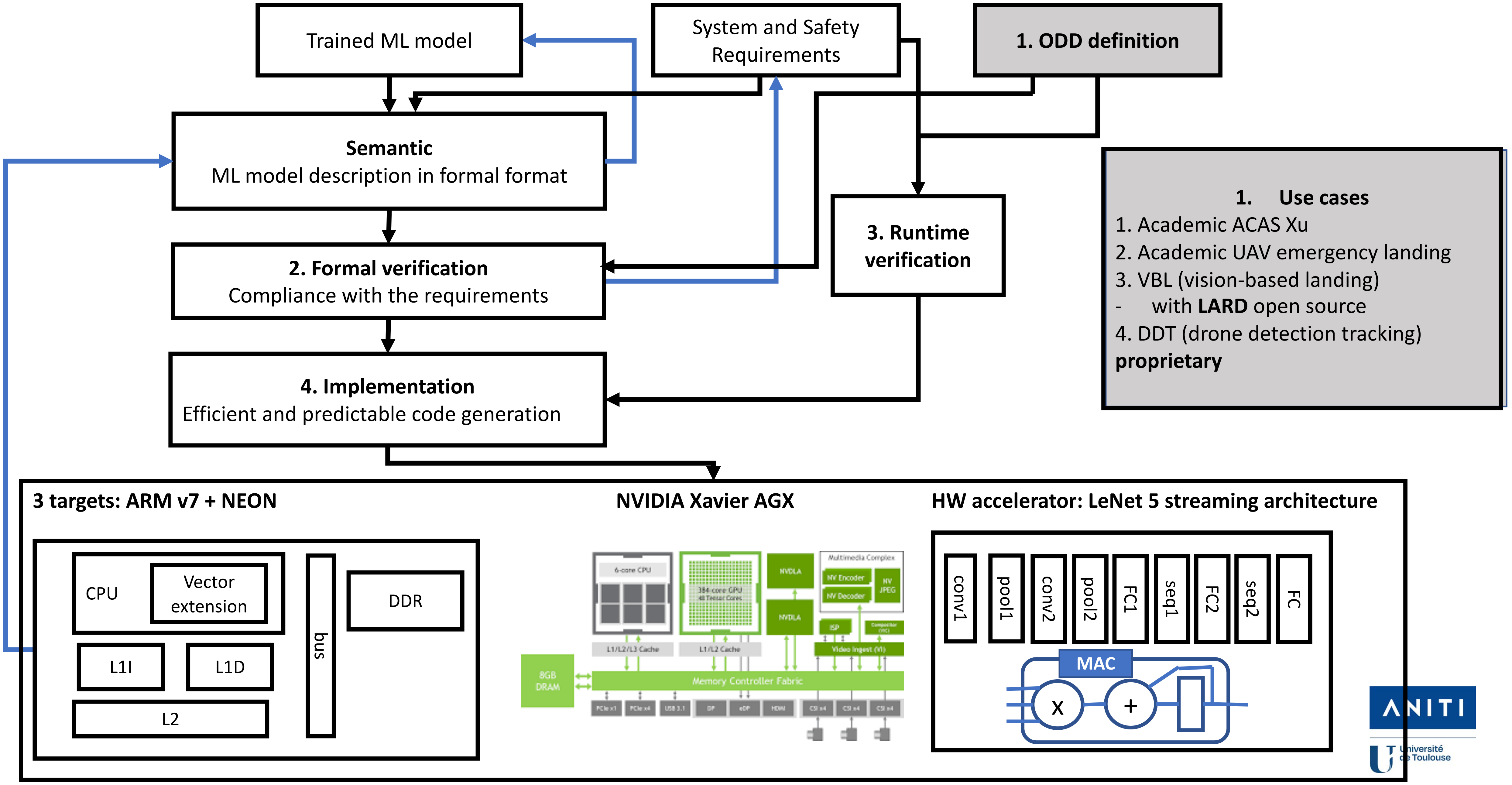






General objective: End-to-end development process to achieve the expected level of performances and provide some of the evidences required by certification.

Outline & contributions



ACAS Xu

Avoidance Collision System for vertical and horizontal cooperative and non-cooperative avoidance (Multi-Intruders)

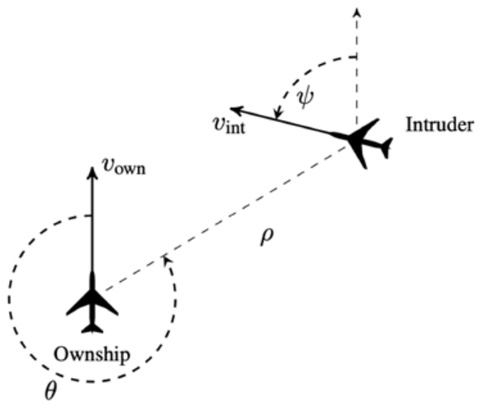
Intended function "the intruder should not enter in the ownship envelope”

ODD: pre-defined ranges of inputs

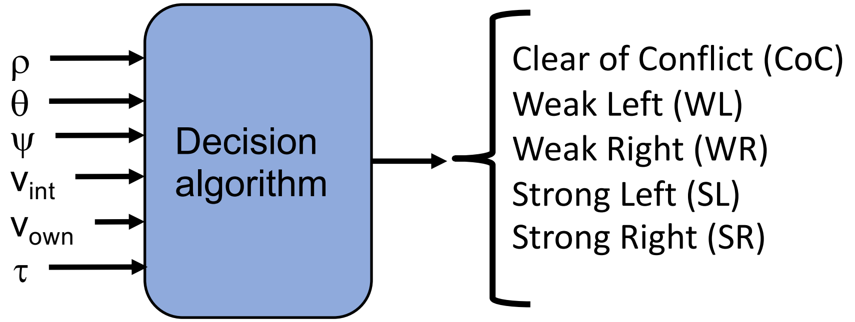
- Initially based on look-up tables (LUT)
- Replacing the LUT by neural networks proposed by Standford
  - Interest: Gain in memory footprint (from 4Go to 3Mo)

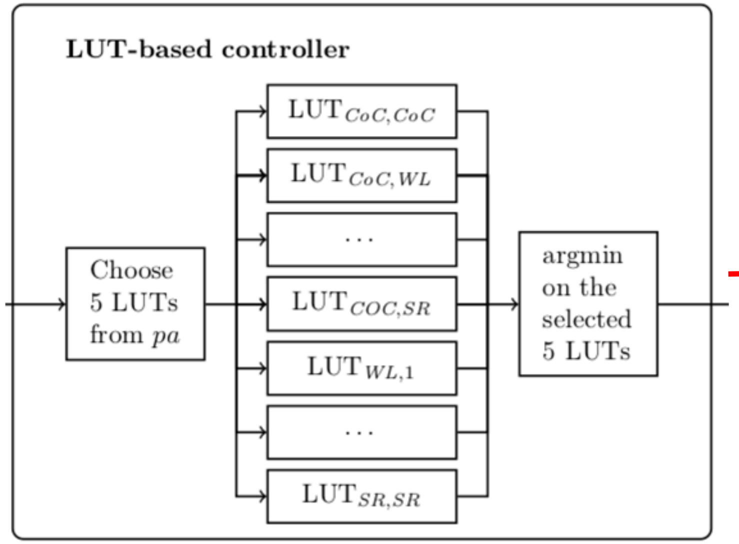
ANITI focus:

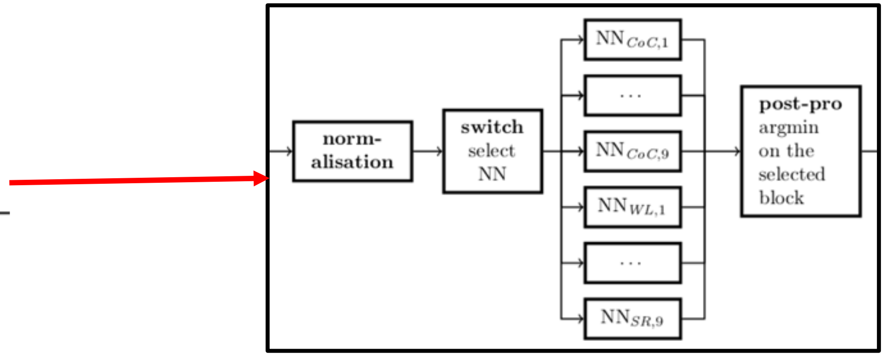
- Formal verification
- Hybrid architecture definition (safety net as a backup when the NNs do not replicate well the LUT)
- Implementation



State of the intruder relative to ownship

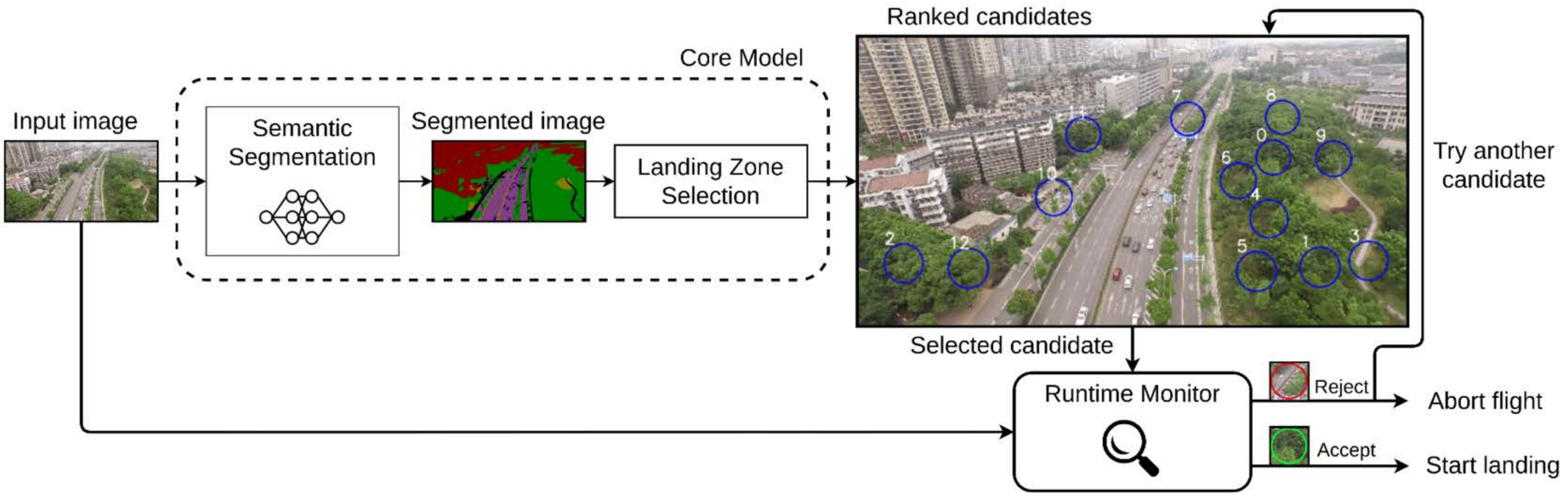






45 NN (fully connected NN with 6 layers of 50 neurons each)

UAV emergency landing

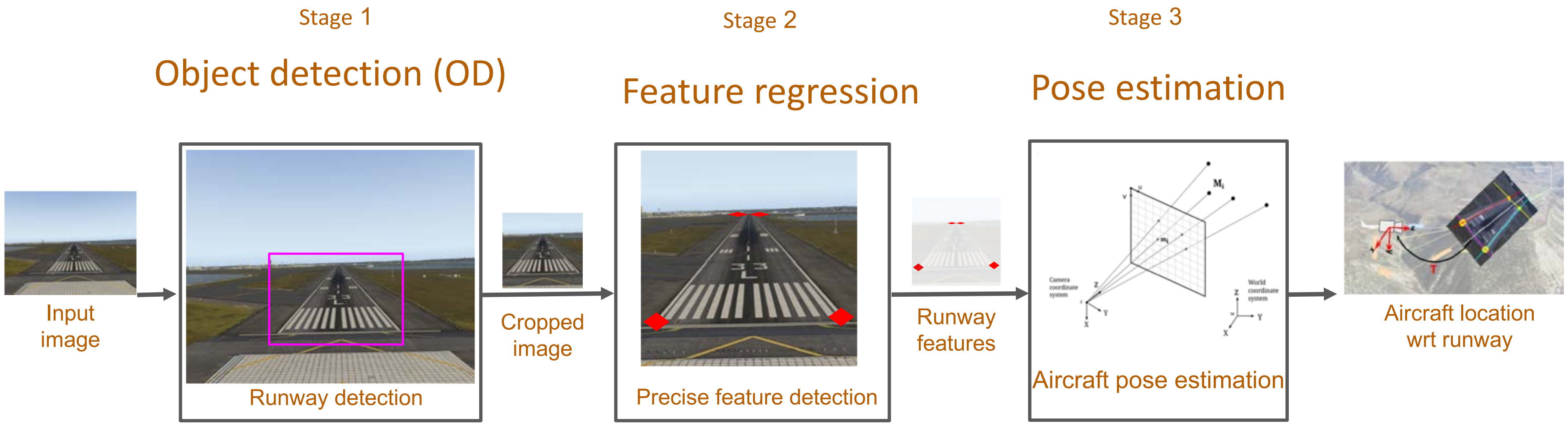


Intended function: identify a landing zone in urban environment ensuring a safe emergency landing. If no suitable landing is found, the flight is aborted.

ANITI focus:

- Runtime verification

VBL (vision-based landing)



Intended function: computing the position of the aircraft from the position of the runway within an image taken during the approach and landing phases of an aircraft.

ANITI focus:

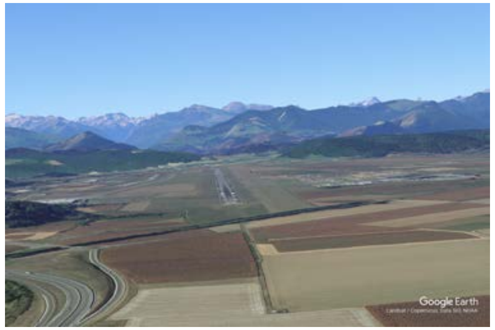
- Dataset design
- ODD definition
- Formal verification
- Implementation

LARD – Landing Approach Runway Detection – Dataset

- Training dataset
  - Google Earth Studio and Microsoft Flight Simulator synthetic images of 33 runways
- Test dataset
  - synthetic images of 79 runways
  - real footage of 38 runways

Collaboration with **DEEL** mission certif

LANDING APPROACH RUNWAY DETECTION



LARD -- Landing Approach Runway Detection--Dataset for Vision Based Landing . Ducoffe et al. 2023. ArXiv
https://github.com/deel-ai/LARD

Tarbes: Comparison of a real landing footage (left) with synthetic replicas (Google Earth Studio center, Microsoft Flight Simulator right)

DDT – drone detection tracking

Inteded function: camera-based detection of intruder drone:

- Focus and tracking of the drone by the camera

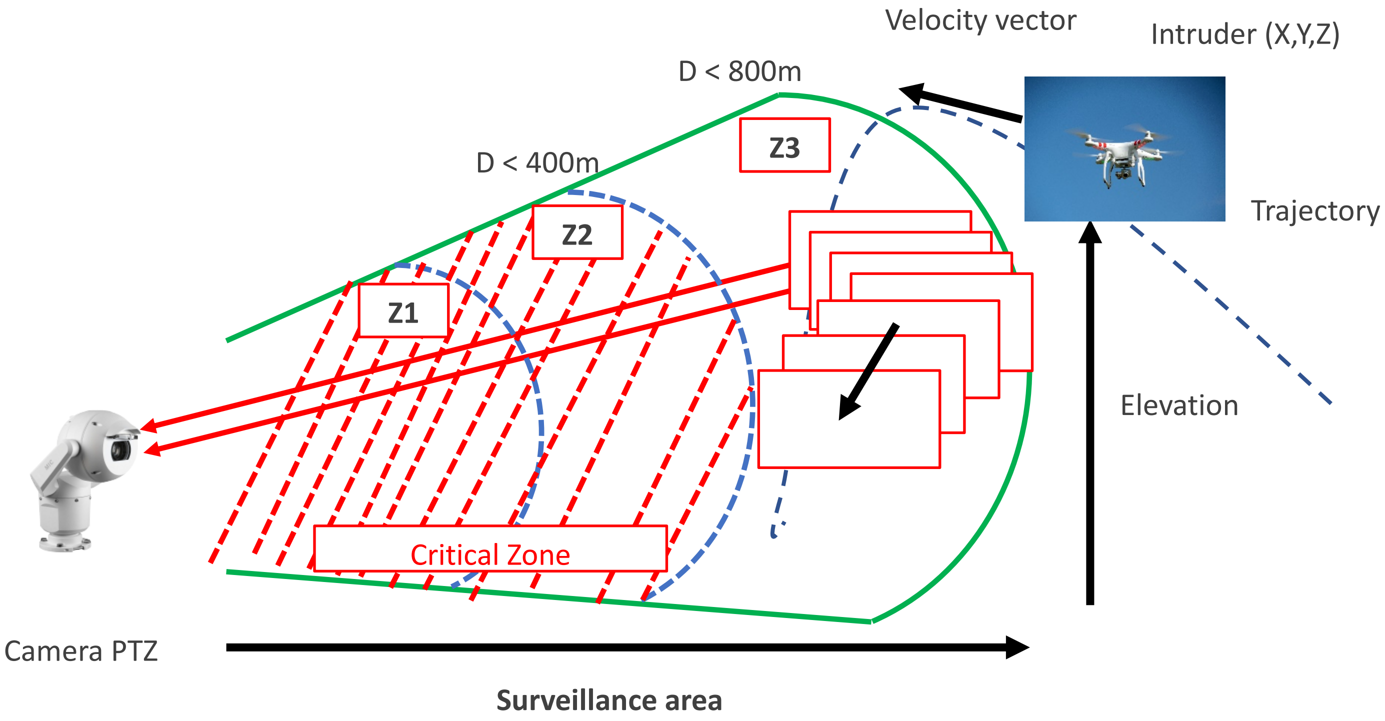
ANITI focus:

  - Extension of existing academic and CS Group proprietary dataset
  - ODD definition
  - Model design
  - Implementation on the NVIDA Xavier

- **Main difficulties**
  - Object size ≤ 25 pixels^2
  - Birds and drones

  - Position within the image

- **Collaboration with IRT Saint Exupéry – CS Group within ARCHEOCS project**



a Sopra Steria company

ANITI
Université de Toulouse

DDT – dataset extension

ODD definition via several operational scenarios

- Category of objects (bird, drone), size of the objects, range of velocities
- Possible trajectories
- Diversity of backgrounds (empty grass background, buildings …)
- Possible out of ODD (helicopter, plane)

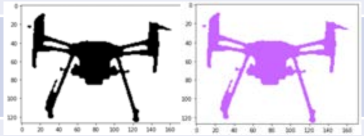
Existing dataset

- Example of academic datasets: Distant Bird Detection Dataset for Safe Drone Flight https://github.com/kakitamedia/drone_dataset
- Internal company collection of data

➔ limitations: not all the ODD is covered and many biases in the dataset (position of the drone in the image, type of background…)

Example of extension

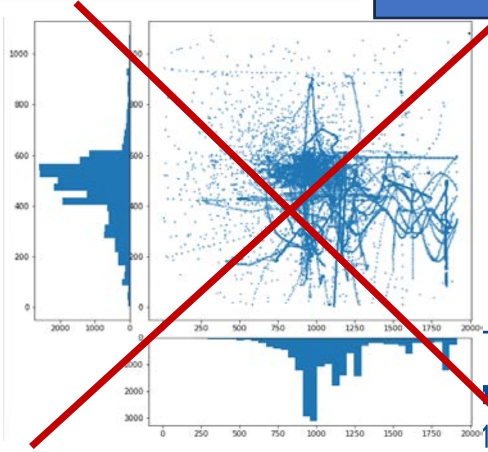
- unbiased position of the object

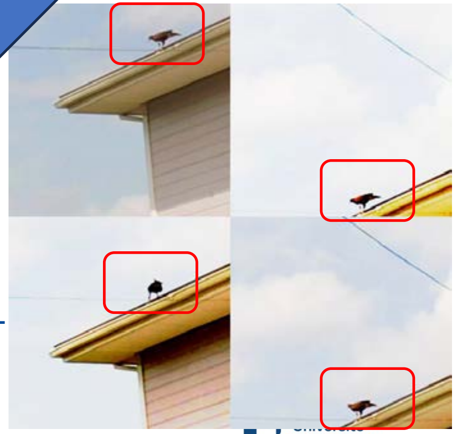
Parameters Test Set	Combinatorial - Range
Classes (Objets)	
Object Size	Min 20 pixels – Max 400 pixels
Background	Sky, buildings, Landscape
Position	Uniform Spatial distribution
Texture	
Brightness	Monotonic Function
Geometric Transformation	Rotation – Symmetric

Distribution of object position in the dataset

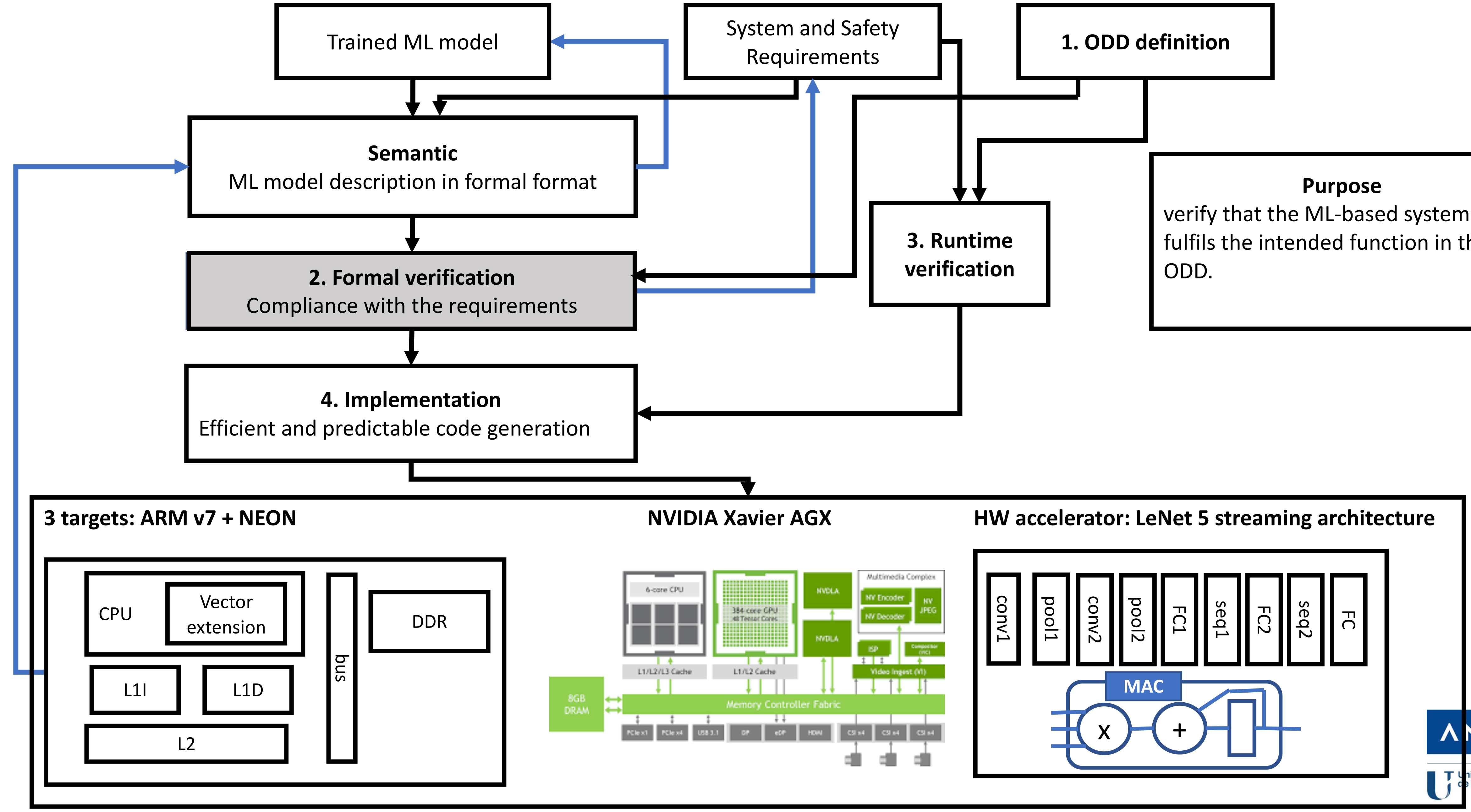
Transformation to unbiase the Position

Modification with uniform spatial distribution





Outline & contributions



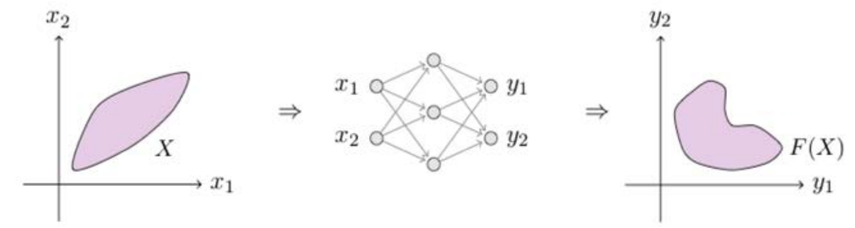
Reminder on formal verification

Reachability problems

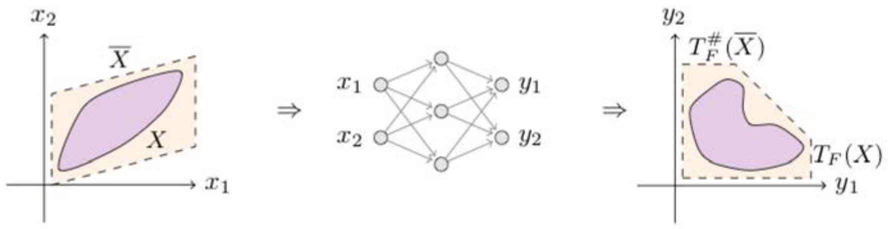
- **Property:** Given an input set X and a NN model realising the function F, what is the reachable set F(X)?
- **Practical property:**
  - X is approximated with an abstract domain
  - solver computes (an over-approximation of) F(X)
- **Existing solvers:**
  - Exact solvers: Reluplex/Marabou, Planet, …
  - Approximating solvers: Auto-Lirpa (IBP, Crown, …), DecoMon, …

What you know:

- Last year ANITI Days presentation with Mélanie
- Lightning talk of Noémie



property



solved problem

DecoMon – verification tool developed by Mélanie

Build a toolbox to ease the use of **formal methods** among Airbus Data Scientists.
https://github.com/airbus/decomon

model

```
decomon_model =
convert(model)
```

Automatic Keras Conversion
(Formal Training)

```
get_upper_box
get_lower_box
```

Automatic Common properties built with **divisions**

Business

Open-source Airbus library
Used by BUs
Airbus ML-flow

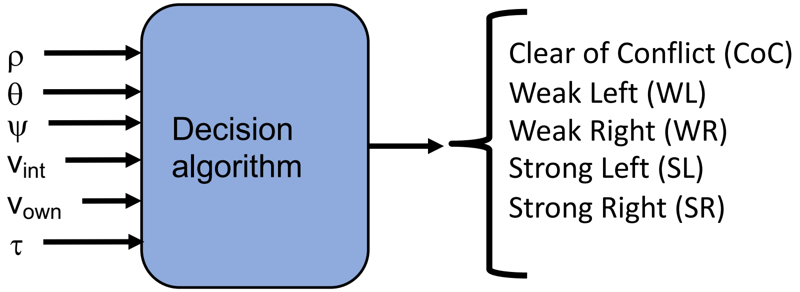
Compatibility with ANITI's libraries

Research

CVPR 2023: FM for XAI
ICAF 2023: FM for predictive maintenance
Ongoing submission: NASA FM …

decomon Public

Application on ACAS Xu



Definition [p-box]: $p \in N$, a p-dimensional box $[b]^p$ is a set of R^p defined as the cartesian product of p intervals:

$$[b]^p = \times_{i<p+1} [l_i,u_i]$$

Definition (Similar behaviour).

- A = {*CoC,WL,SL,WR,SR*}.

We consider that a $NN_{pa,range}$ *behaves similarly to the* LUT_{pa} *on* l *if*

$$\textbf{decisions } NN_{pa,range}(l) \subseteq \textbf{decisions } LUT_{pa}(l)$$

Collaboration with DEEL mission certif, Collins

Case 1 3D (fixing v_{own}, v_{int}): 304 000 p-boxes

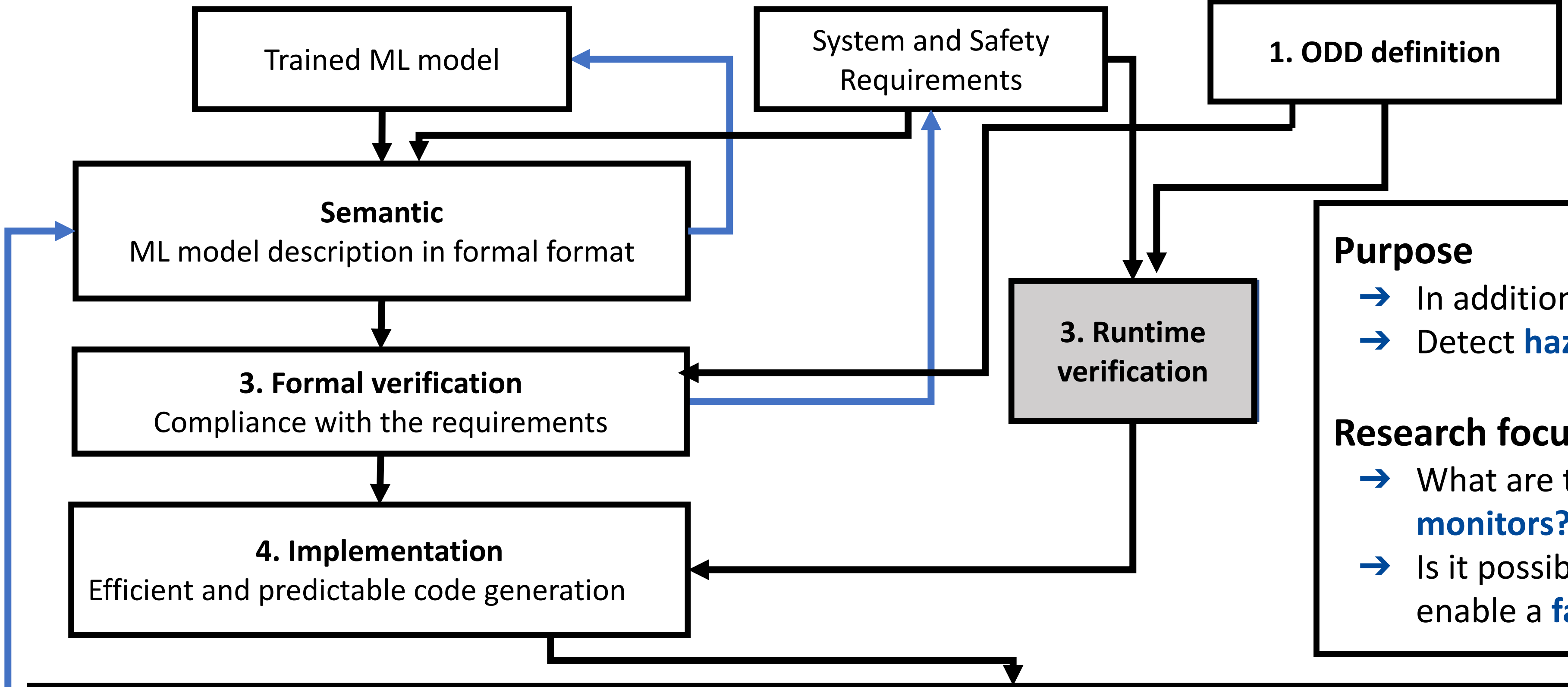
	Verif time	Number of true	Success rate
Reluplex NN	5 days	254 670	84%
adversarial	17 hours	286 023	94%
Corner	32 hours	272 212	90%
deellip	26 min	280 000	92%

Case 2 5D: 36.10^6 p-boxes

	Verif time	Number of true	Success rate
Reluplex NN	> year		
adversarial	29 days	34 352 549	93.4%
Corner	> 1 month		
deellip	3 days	34 173 698	92.9%

Towards certification of a reduced footprint acas-xu system: A hybrid ml-based solution. Damour et al. Safecomp. 2021
Toward the certification of safety-related systems using ML techniques: the ACAS-Xu experience. Gabreau et al. ERTS. 2022

Outline & contributions

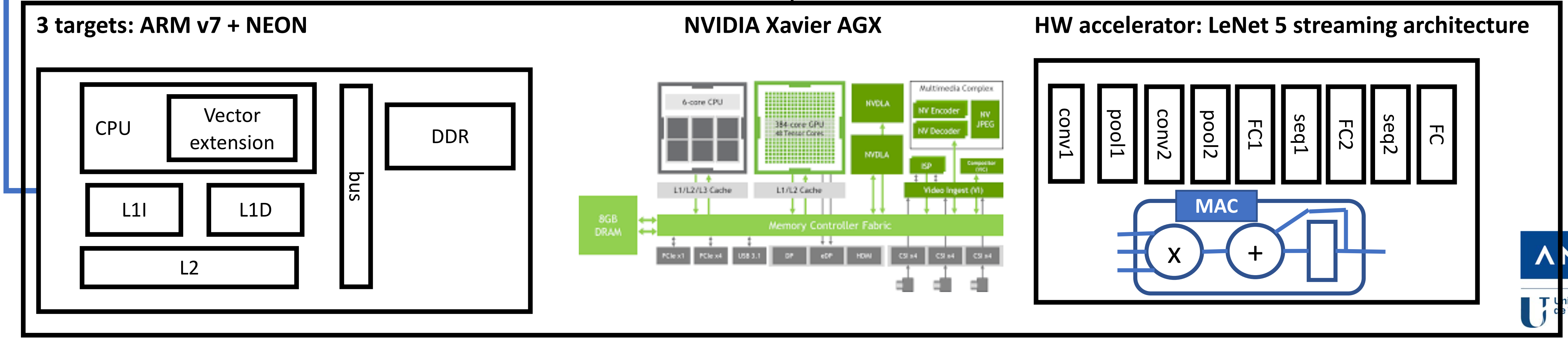


Purpose

- ➔ In addition to formal verification
- ➔ Detect **hazardous behaviors** at runtime

Research focus: Evaluation of safety monitors

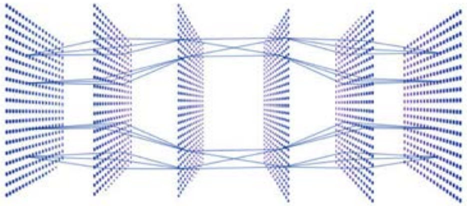
- ➔ What are the relevant metrics to assess **safety monitors?**
- ➔ Is it possible to unify the evaluation method to enable a **fair comparison** of safety monitors



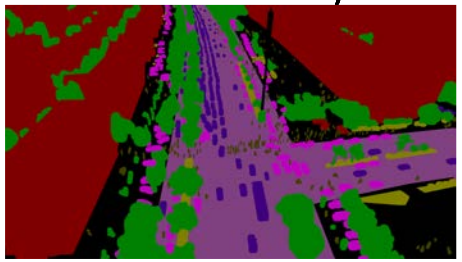
What is Safety Monitoring of Deep Neural Networks ?

Sensory input (e.g., Image)

Deep Neural Network



Prediction used by critical



Safety Monitor (may observe many parts of the DNN)

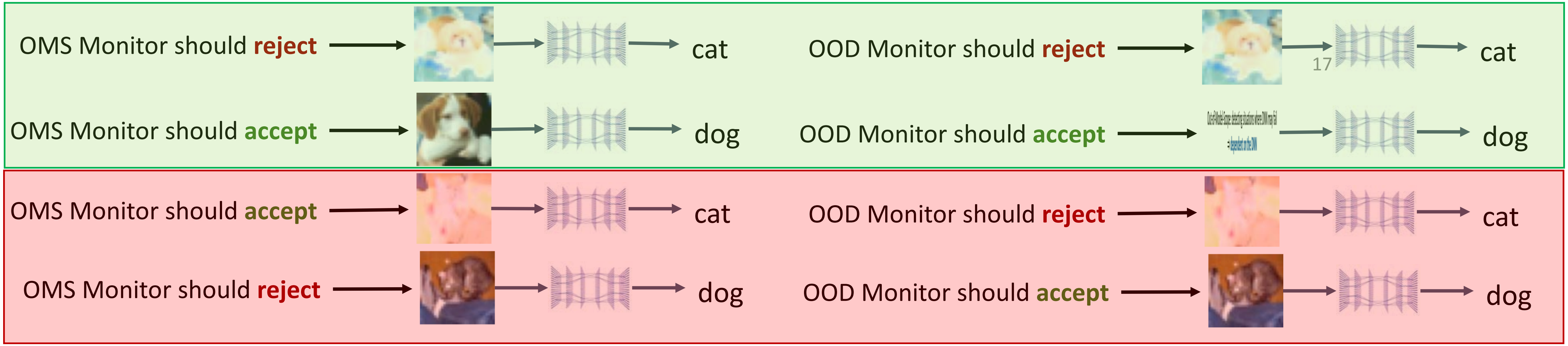
Warning
unsafe to use the DNN predictions

Images: *Cordts et al. "The cityscapes dataset for semantic urban scene understanding."*

CertifAI Chair Contribution : Unambiguous evaluation measures for safety monitors

Out-of-Model-Scope: detecting situations where DNN may fail
⇒ **dependent on the DNN**

Out-of-Distribution: detecting situations that are not sampled according to the training distribution
⇒ **independent from the DNN (dependent on the dataset)**



- ➔ The definition of OMS is **objective**
- ➔ OMS is not assuming which situations are difficult for the DNN

- ➔ The definition of OOD is **ambiguous** e.g., how much perturbation is required to define OOD?
- ➔ A perfect OOD detector can **discard valid predictions** the DNN
- ➔ The best monitor for OOD is not always the best to **detect errors**

Out-of-Distribution Detection Is Not All You Need, *Joris Guérin, Kévin Delmas, Raul Ferreira & Jérémie Guiochet,* AAAI 2023

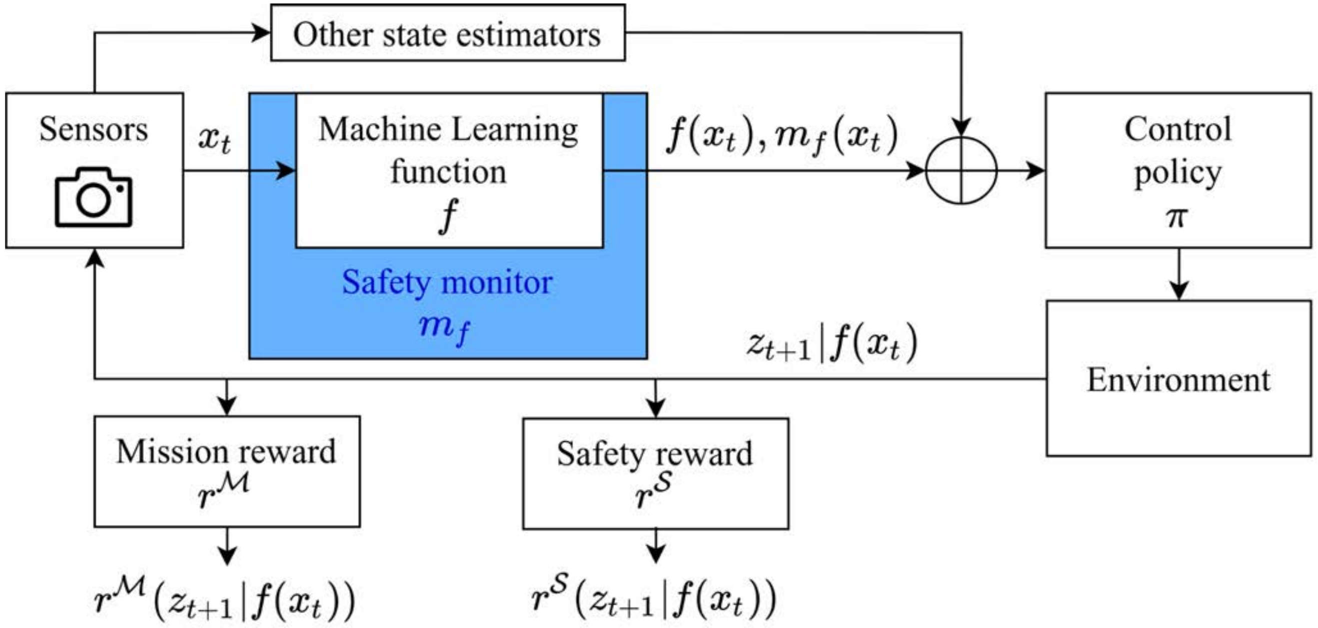
CertifAI Chair Contribution: OMS-based monitor evaluation framework

Evaluation framework inspired from RL training process

- ➔ **Assess the safety and capacity** to complete the mission (i.e., mission reward)
- ➔ Applicable to **any monitor** for perceptive functions
- ➔ Request **a clear formalisation** of the assumptions that are made regarding the impact of a ML error on the system's safety

Applied on automotive and UAV use cases and demonstrates that:

- ➔ Monitors developed for OOD **may not be the best** to ensure the safety



Unifying Evaluation of Machine Learning Safety Monitors, *Joris Guérin, Raul Ferreira, Kévin Delmas & Jérémie Guiochet,* ICRA 2022 & ISSRE 2022

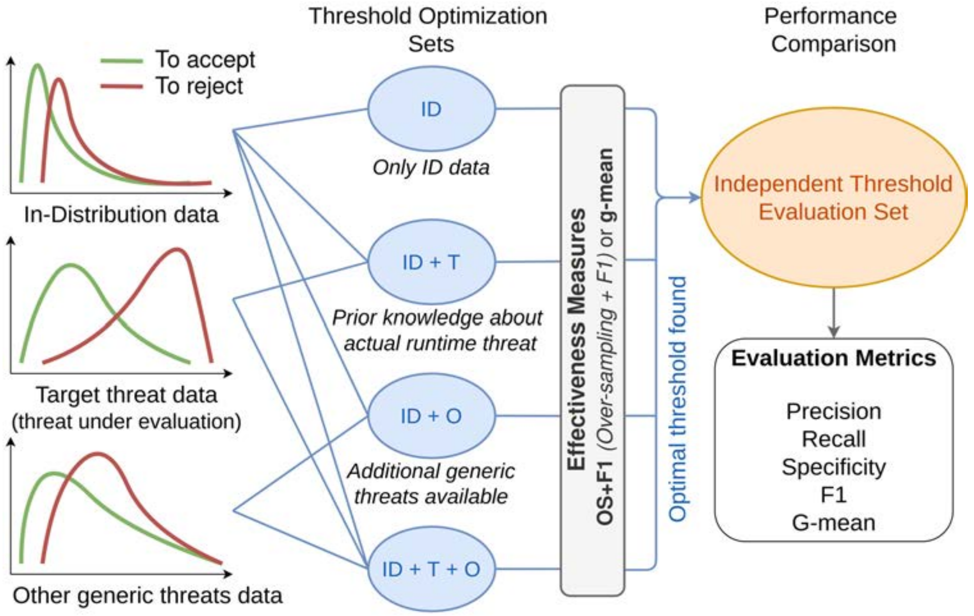
CertifAI Chair Contribution: Designing and tuning process for efficient safety monitors

Selecting a rejection threshold is pivotal during the design of a safety monitor but:

- ➔ Many works use **threshold-agnostic** evaluation metrics (e.g., AUROC)
- ➔ Very-few works are adressing the problem of selecting an **optimal threshold**
- ➔ The actual evaluation the safety monitor should be done for the **selected rejection threshold**

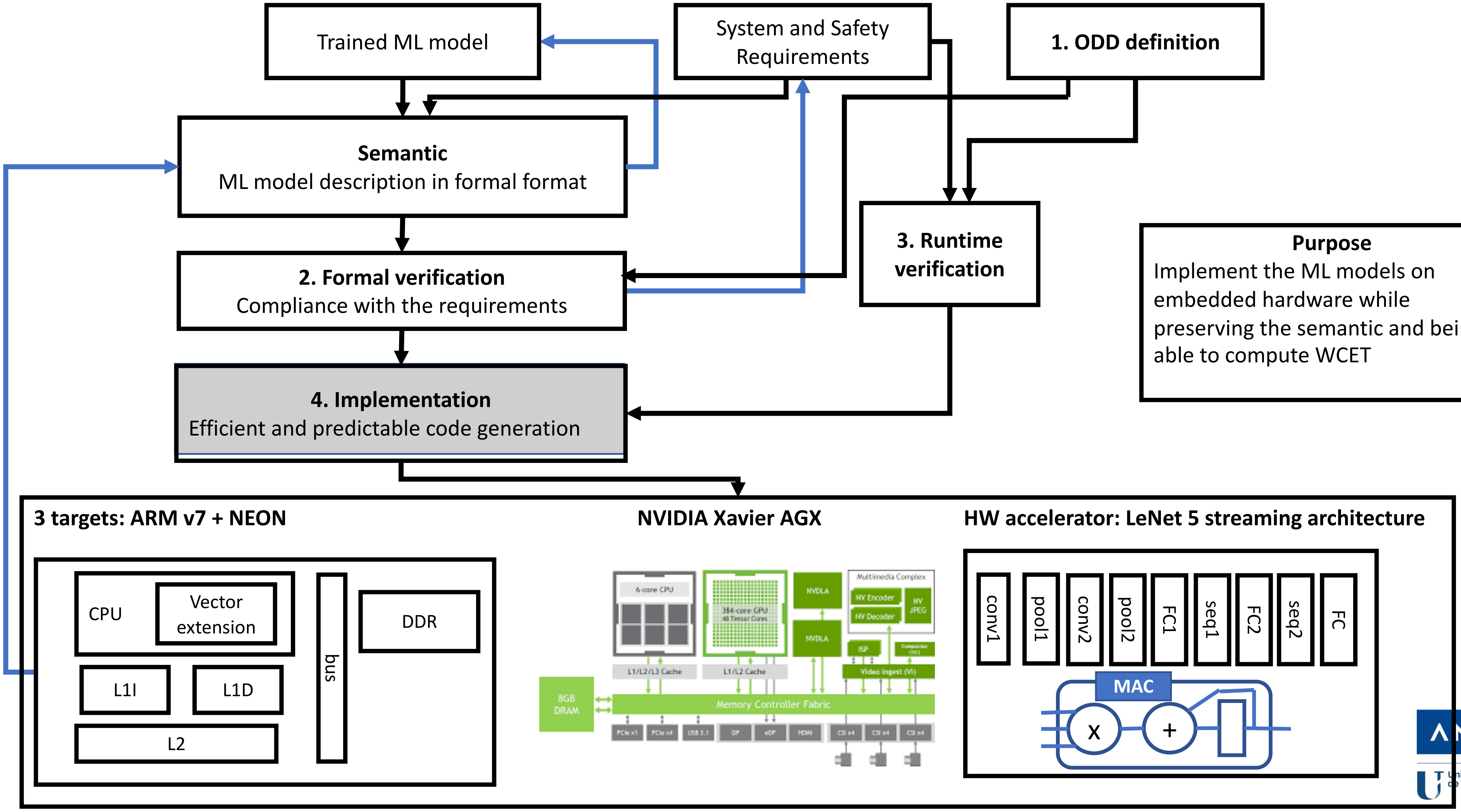
Develop a threshold optimisation method for safety monitors:

- ➔ Assess the impact of **prior knowledge** on problematic situations (i.e., runtime threats)
- ➔ Select **threshold-aware metrics** to evaluate a monitor
- ➔ Provide an **automated optimization** of the rejection threshold



Red Pill or Blue Pill? Thresholding Strategies for Neural Network Monitoring, *Tran Khoi, Joris Guérin, Kévin Delmas & Jérémie Guiochet,* ICLR 2024 (under review)

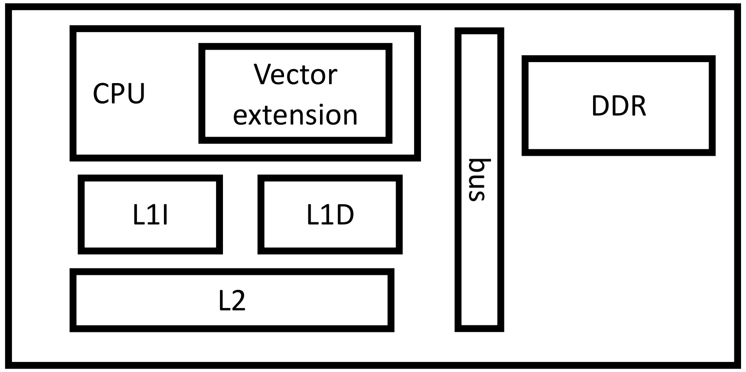
Outline & contributions



ANITI
Université de Toulouse

Certification context (DO 178-C)

- Traceability between the requirements and the (source) code
- Capacity to compute tight WCET
- Intense testing



ACETONE Automatic sequential C code generation from inference model

https://github.com/idealbuq/NNCodeGenerator

- Convolution, 3 implementation: direct conv, naïve gemm, gemm (block matrices)

Criteria:

- Semantic preservation: similar results in the order of 10^{-6}
- WCET
- MET

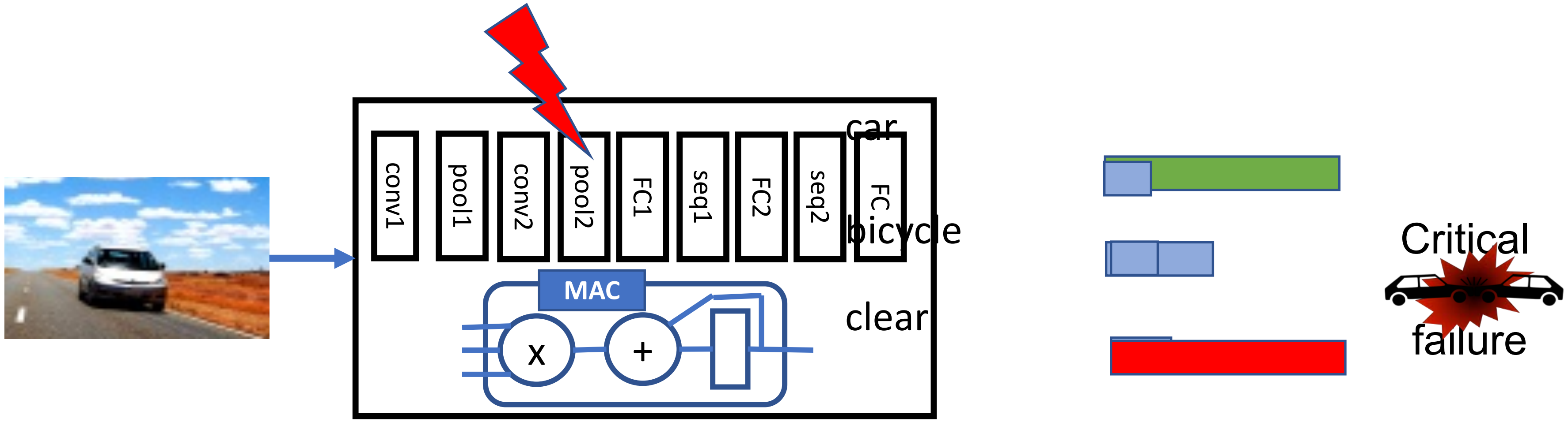
Architecture / Framework	Execution time [cycles] ACAS-Xu *decr128*	LeNet-5	CifarNet	WCET [cycles] ACAS-Xu *decr128*	LeNet-5	CifarNet
ACETONE	533 767	12 186 378	233 450 428	6 128 253	165 718 749	3 018 534 290
Keras2C	1 104 134	25 786 401	642 390 830	36 838 054	1 160 385 934	97 959 064 345
uTVM static	681 708	10 201 249	193 599 362	6 765 413	113 449 651	3 215 754 680

Extending a predictable machine learning framework with efficient gemm-based convolution routines. De Albuquerque et al. RTS. 2023
ACETONE: Predictable Programming Framework for ML Applications in Safety-Critical Systems. De Albuquerque et al. ECRTS. 2022

Neural network hardware accelerator

Purpose: Impact of hardware faults on the execution of DNN

- **Type of accelerator:** streaming architecture



- HW fault injection
- Formal methods to assess the quality of fault injection strategy
- Reduction of fault injection points while preserving the coverage

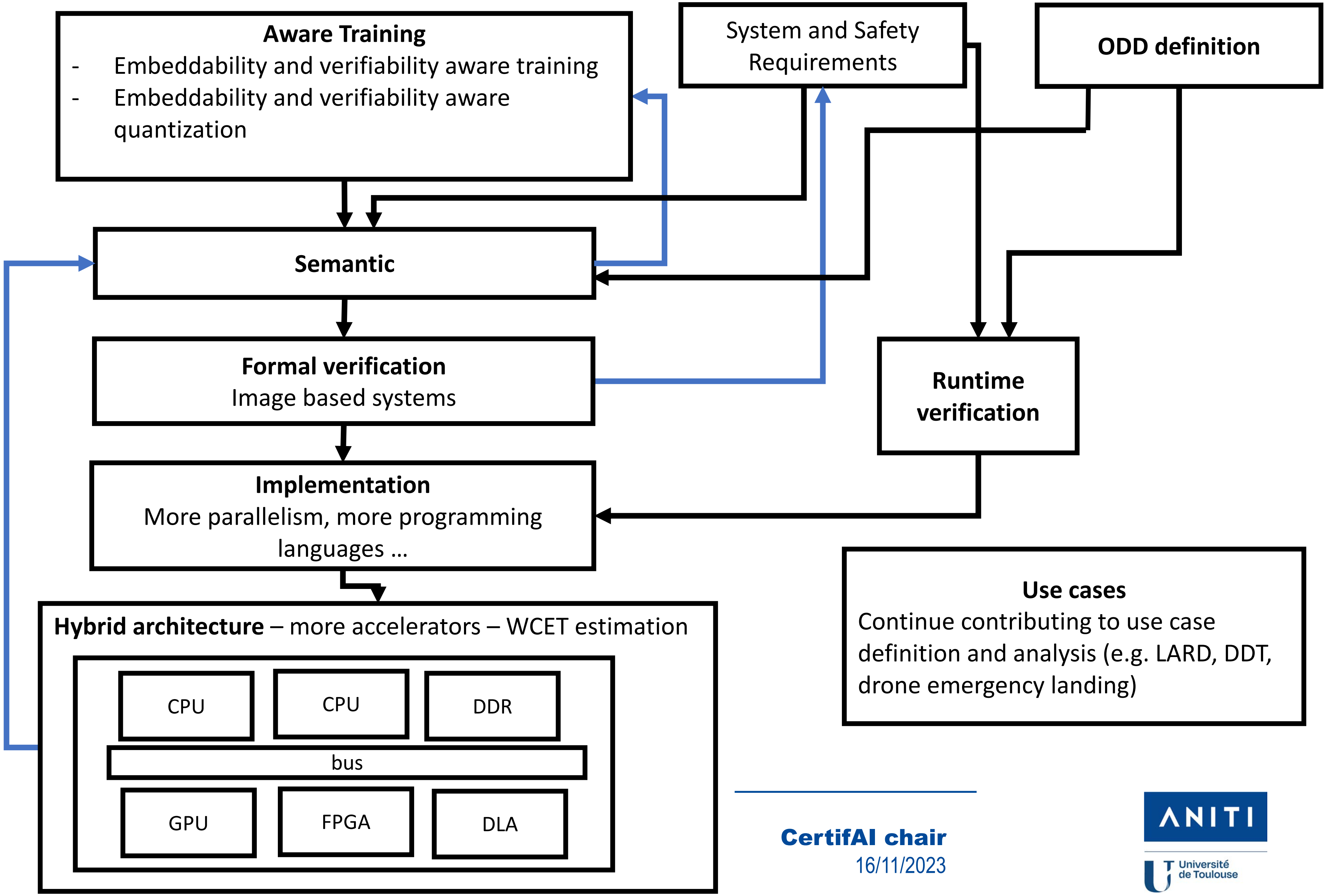
Collaboration with NXP

Quality of Fault Injection Strategies on Hardware Accelerator. Guinebert et al. Safecomp 2022.
Fault injection strategies: identifying HW failures with functional impact. Guinebert et al. ETS industrial paper. 2023

Conclusion & future work

End-to-end development process to achieve the expected level of performances and provide some of the evidences required by certification.

ANITI2: industrial chair on "Embeddability and safety assurance of ML-based systems under certification (CertifEmbAI)"



Use cases
Continue contributing to use case definition and analysis (e.g. LARD, DDT, drone emergency landing)