Airbus Protect Artificial Intelligence Conference >>

MLEAP Stakeholders day #3

Paving the way for the future of Artificial Intelligence in Aviation

EASA European Union Aviation Safety Agency

LNE Laboratoire National de Métrologie et d'Essais

numalis

MLEAP project: [Machine Learning Application Approval]

January 25th, 2024

AIRBUS PROTECT

Agenda

Introduction of MLEAP project and the Partners

Presentation of MLEAP roadmap according to EASA objectives -

- **Q&A session,**

Review of work research results & presentation of experimental analyses -

- **Q&A session,**

Generic End to End AI development Pipeline proposal & joint conclusions -

- **Q&A session,**

Key takeaways & MLEAP Project next steps -

- **Q&A session,**

Afterwork with MLEAP Team & Stakeholders

Disclaimer

Machine Learning Application Approval (MLEAP) project is funded by the European Union, Horizon Europe Program.
EASA role in MLEAP project is limited to contract and technical management contract.
Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Union Aviation Safety Agency (EASA). Neither the European Union nor EASA can be held responsible for them.

Project page : https://www.easa.europa.eu/en/research-projects/machine-learning-application-approval

/ Introduction of MLEAP Partners

AIRBUS

Who we are > > > MLEAP TEAM

Consortium members :

AIRBUS

Founded in 1901 - Appointed by French government on testing, certification and metrology for Industry (all sectors)

AI evaluation Department

Development of evaluation standards
AI systems testing
Development of certification schemes
Development of testbeds
Professional training for industry

950+ systems evaluated in all major domains of AI and robotics since 2008

Development of softwares for AI evaluation and data preparation

Certification for AI processes (2021).

LEIA 1/2/3: testbeds for AI and robotics (simulation, physical, hybrid)

AIRBUS

Software:

AI Robustness
AI Explainability
Formal analysis
Trustworthy AI

Standardization:

ISO/IEC standard editor on AI robustness
Contributor to many other projects

Services:

Standardization ecosystem
Validation process
AI Audit

numalis

Numalis, the no-guess company

Formal methods for AI systems
Markets: Aeronautic, Defence, aerospace, railway, health
SaaS solution to
Measure robustness
Explain behavior
Prepare compliance of IA
23 persons, Montpellier

On-going projects:
HE MLEAP with EASA
2 EDIDP (Defence)
ESA…

AIRBUS

/ Airbus Protect
an {Airbus} company

bringing together outstanding expertise in
safety, cybersecurity and sustainability
we created a European leader in risk management

… delivering consulting, services & solutions

: What we do

Consulting

on Safety, Cybersecurity and Sustainability to optimise performance and support our customers on regulatory compliance and certification

Innovation

We are involved in research projects & member of institutional working groups

Training

We are a recognised training organisation

Software

Specialised software supporting end-to-end safe mobility activities

/ Presentation of MLEAP Objectives according to EASA Roadmap

AIRBUS

Introductory notes from EASA technical team

Guillaume Soudain
EASA AI Programme Manager
MLEAP Project Sponsor

Xavier Henriquel
EASA Safety Expert
MLEAP EASA Tech Lead Coordinator

François Triboulet
EASA ATM/ANS Expert

EASA AI Roadmap 2.0

Extended technical scope

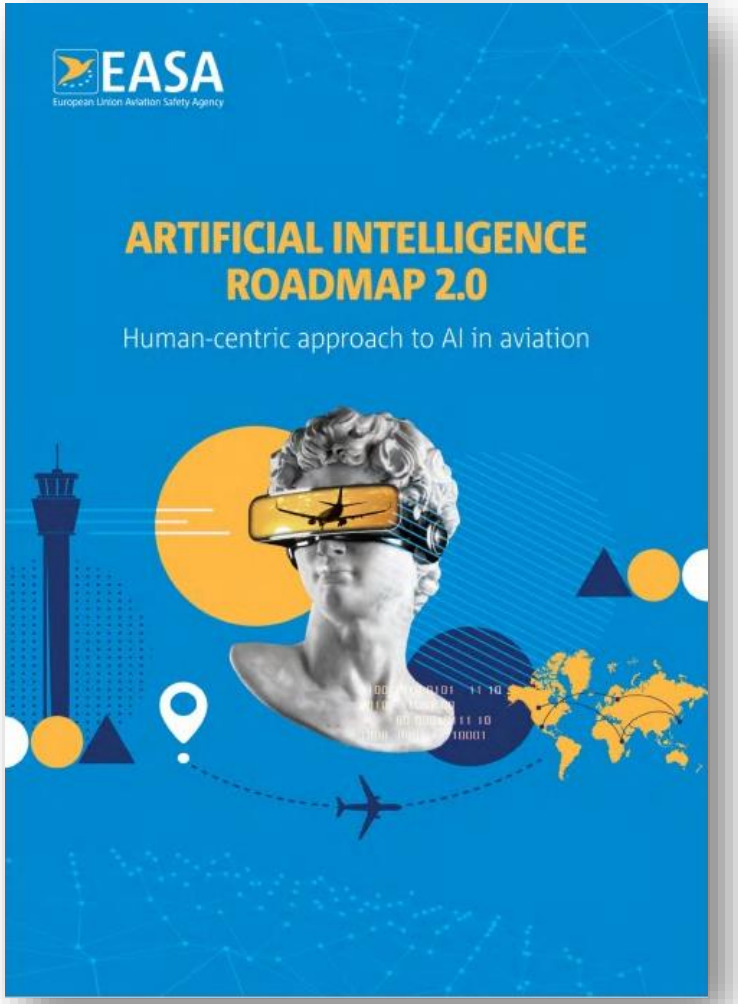
- Unsupervised and Reinforcement Learning
- Symbolic & Hybrid AI

Updated timeline

- Updated industry prognostics
- Augmented action plan

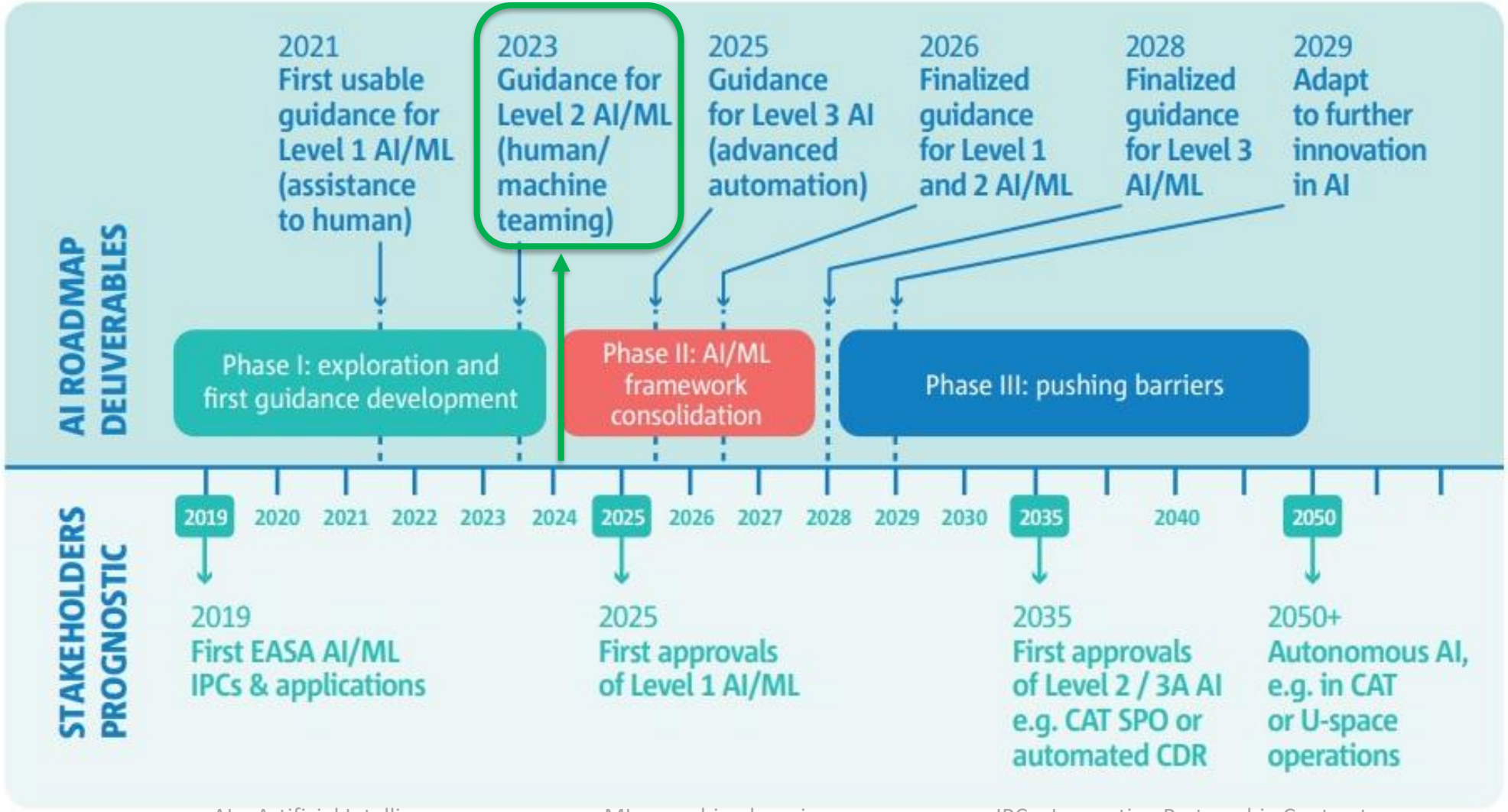
Anticipated consolidation phase II

- Overview of Rulemaking Concept
- Anticipated impact on all domains



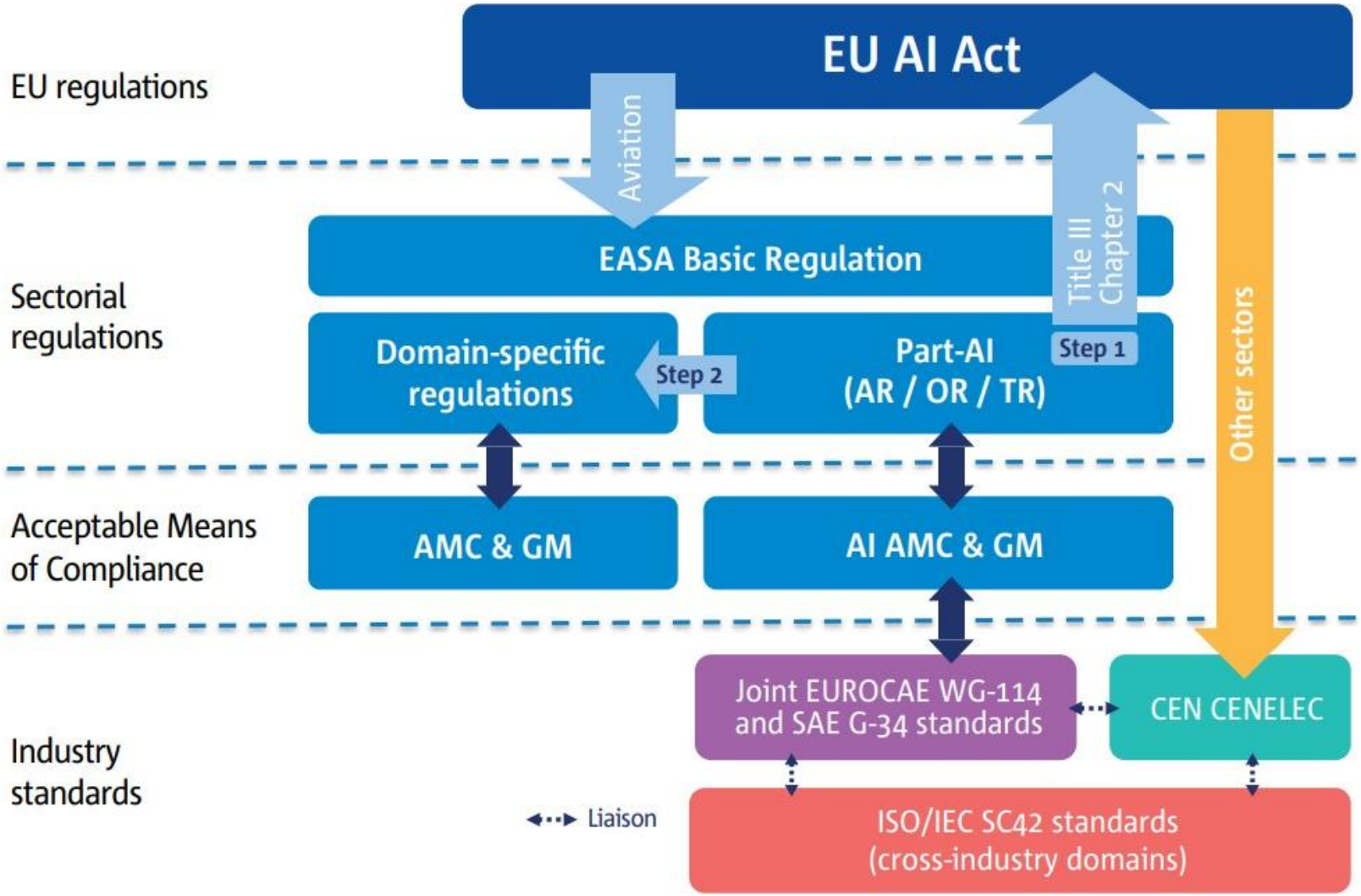
Updated timeline in Roadmap 2.0

Deliverable of Phase I = EASA AI Concept Paper for Level 1&2 AI



AI = Artificial Intelligence
ML = machine learning
IPC = Innovation Partnership Contract
CAT = Commercial air transport
SPO = Single Pilot operation
CDR = Conflict Detection & Resolution

AI Rulemaking Concept

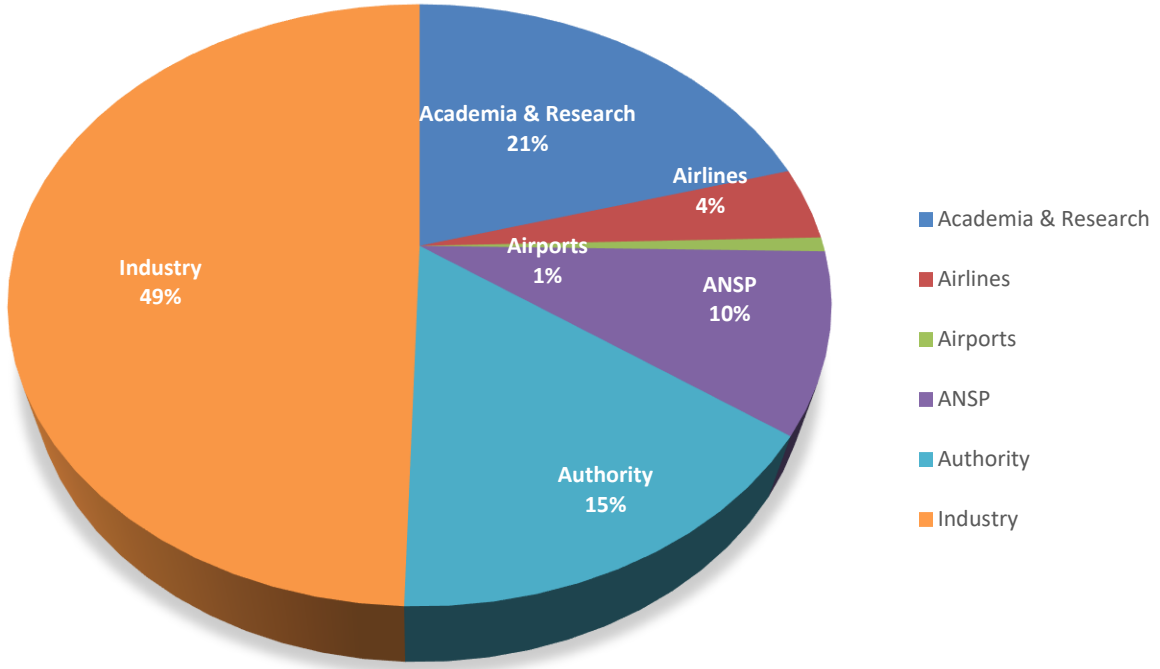


AMC = Acceptable Means of Compliance
OR = Organisation Requirements
AR = Authority Requirements
TR = Technical Requirements
GM = Guidance Material

EASA AI Concept Paper - Proposed Issue 02

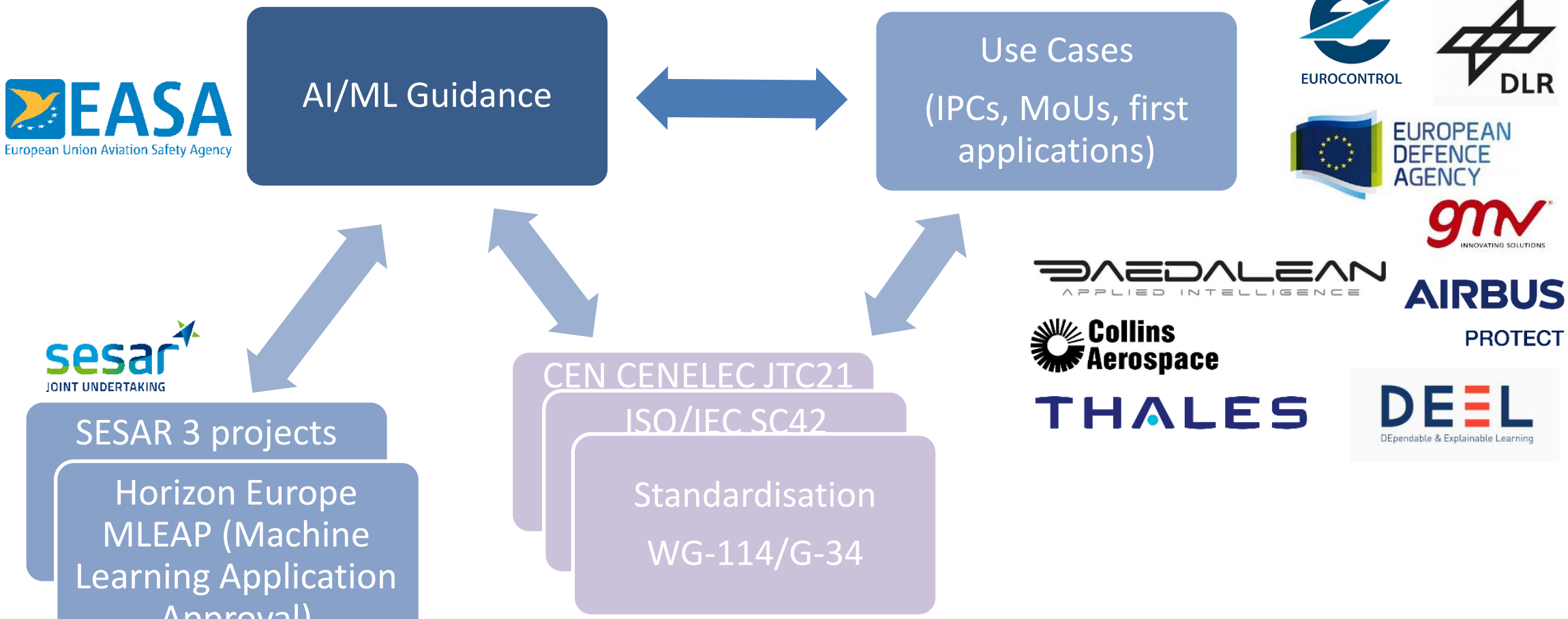
Published for consultation on 24th February 2023 (for 10 weeks)
https://easa.europa.eu/ai

917 comments from 34 stakeholders



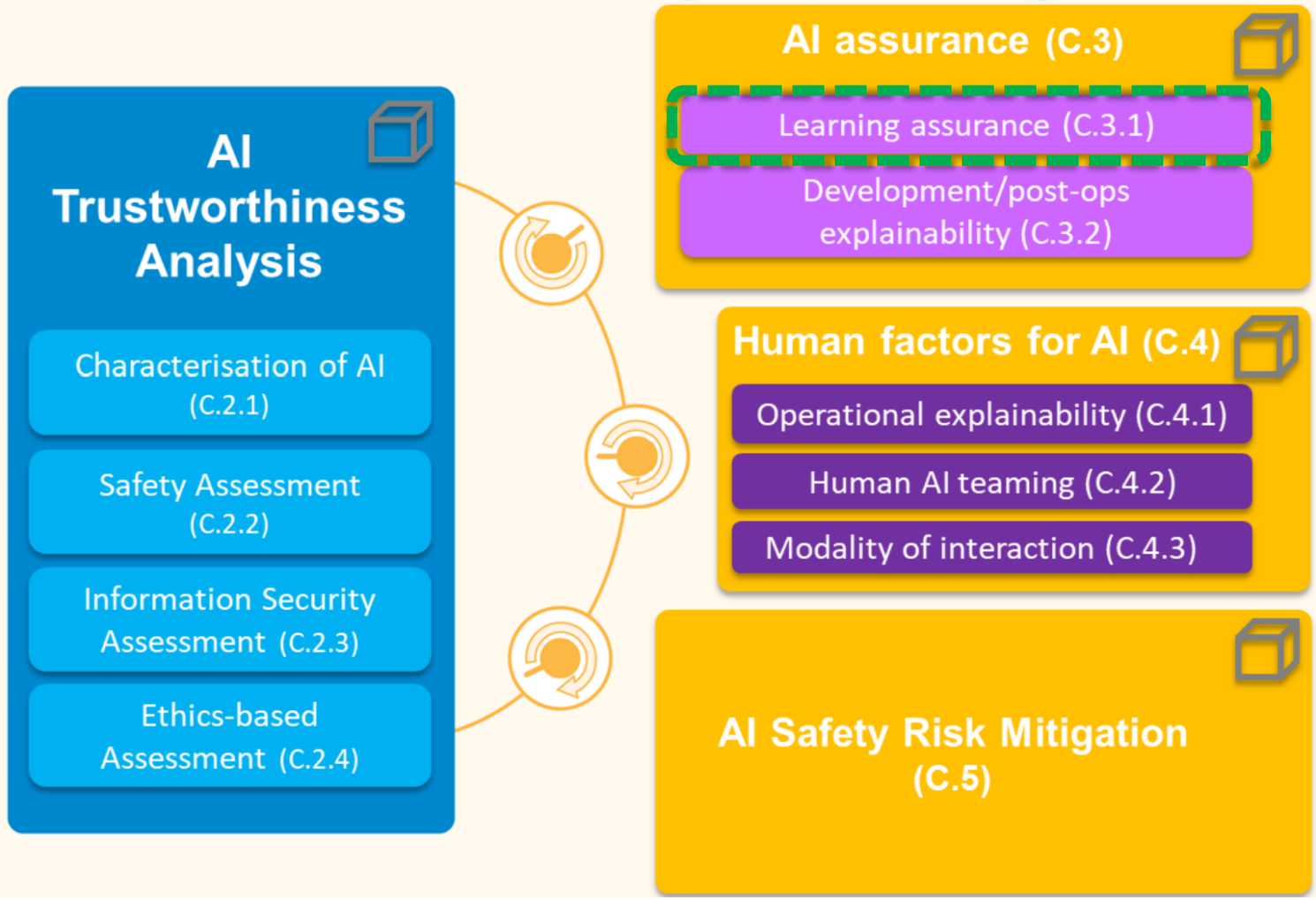
Release of final Issue 02 planned for end of February 2024

Exercising the AI Guidance with use cases



IPC = Innovation Partnership Contract
MoU = Memorandum of Understanding

EASA Concept paper - AI trustworthiness building-blocks



MLEAP Scope

Machine Learning Application Approval (MLEAP) project

Objectives

*"Streamline certification and approval processes by **identifying concrete means of compliance** with the **learning assurance objectives** of the EASA guidance for ML applications*

Research consortium

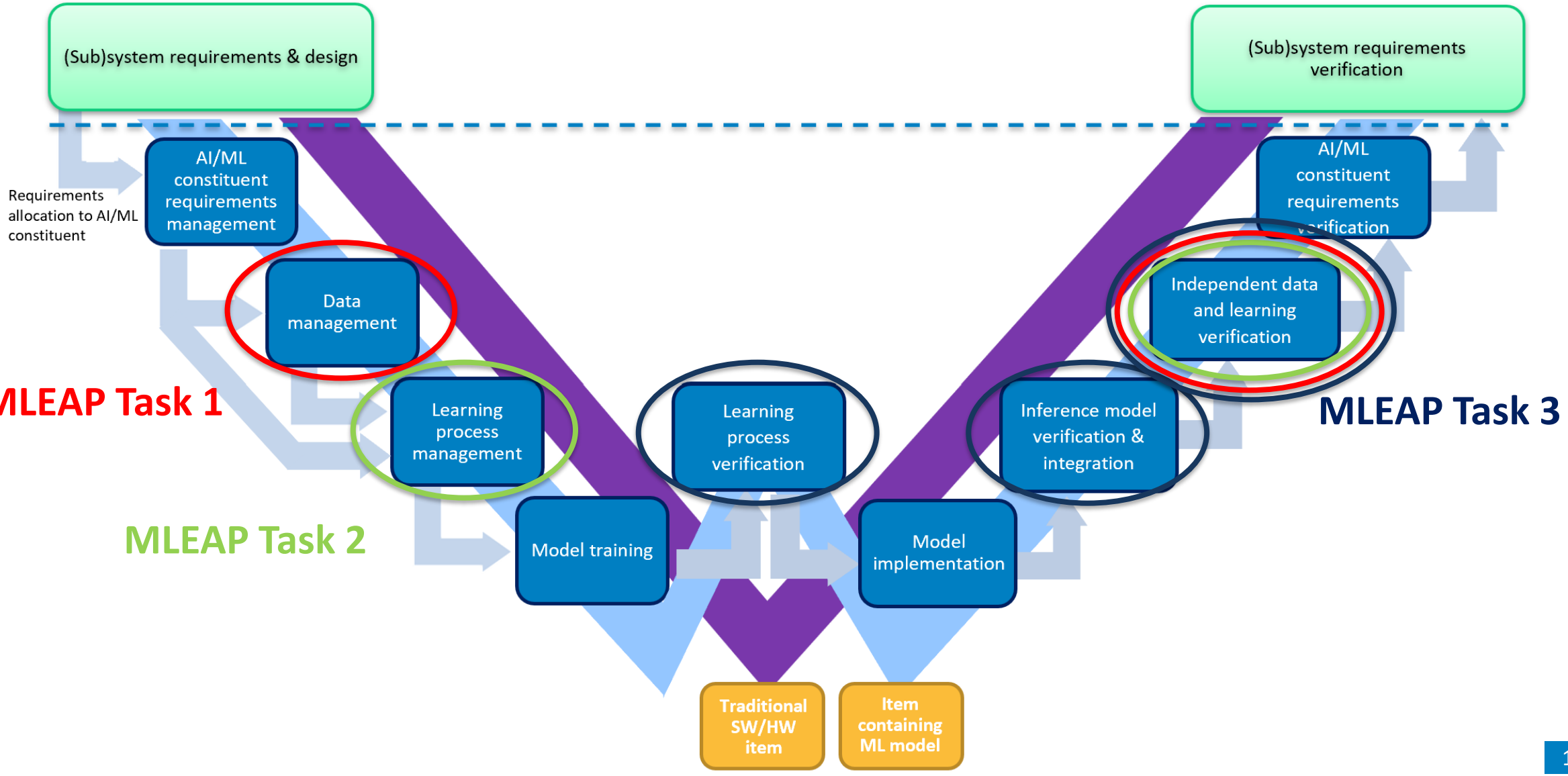
LNE - Airbus Protect - Numalis

Budget & timeline

1.475 m€ funded by EU Horizon Europe program

May 2022 - May 2024

W-shaped Learning Assurance concept



MLEAP project milestones

May 2023

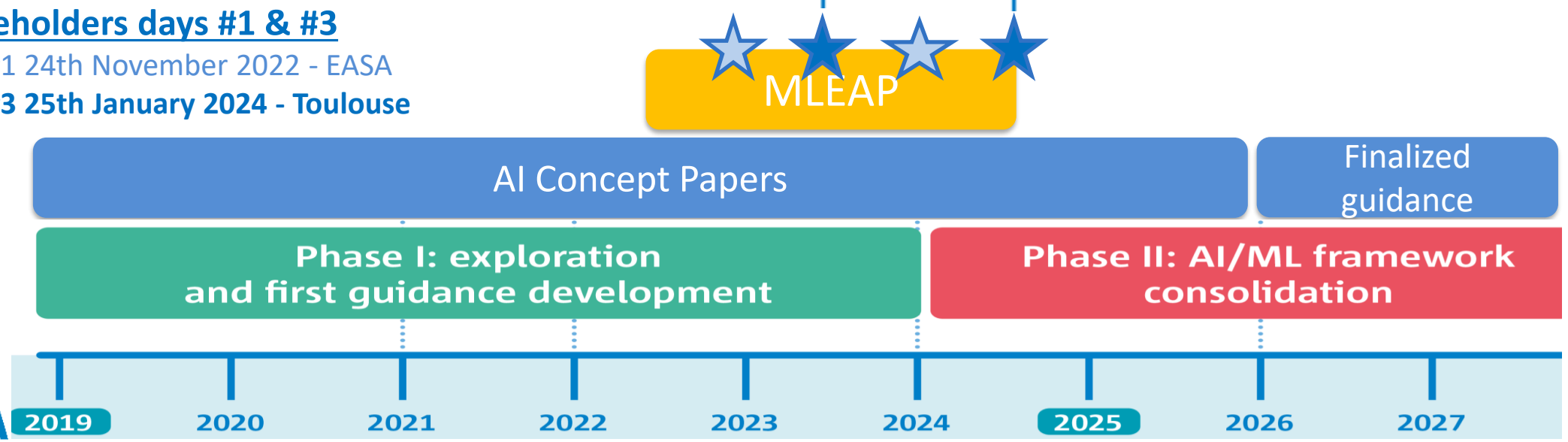
- **First public report & Exec Summary**
- Dissemination events and conferences:
  - "EASA AI days 2023" - 17th May 2023
  - #2 "Paris Air Show 2023" - 21st June 2023
  - "SG34&WG114 Köln Plenary" - 30th June 2023

May 2024

- **Final public report**
- **#4 Final event "EASA AI days 2024" 29th May 2024**

Stakeholders days #1 & #3

- #1 24th November 2022 - EASA
- **#3 25th January 2024 - Toulouse**



MLEAP project

Q&A

AIRBUS

/ Review of work research & presentation of experimental analyses -

AIRBUS

MLEAP workflow towards the objectives fulfillment

What?

Through the EASA's AI Roadmap, several issues have been raised, and MLEAP is launched to bring answers. A big picture of the targeted objectives for MLEAP and a workflow of research activities are provided ;

Why?

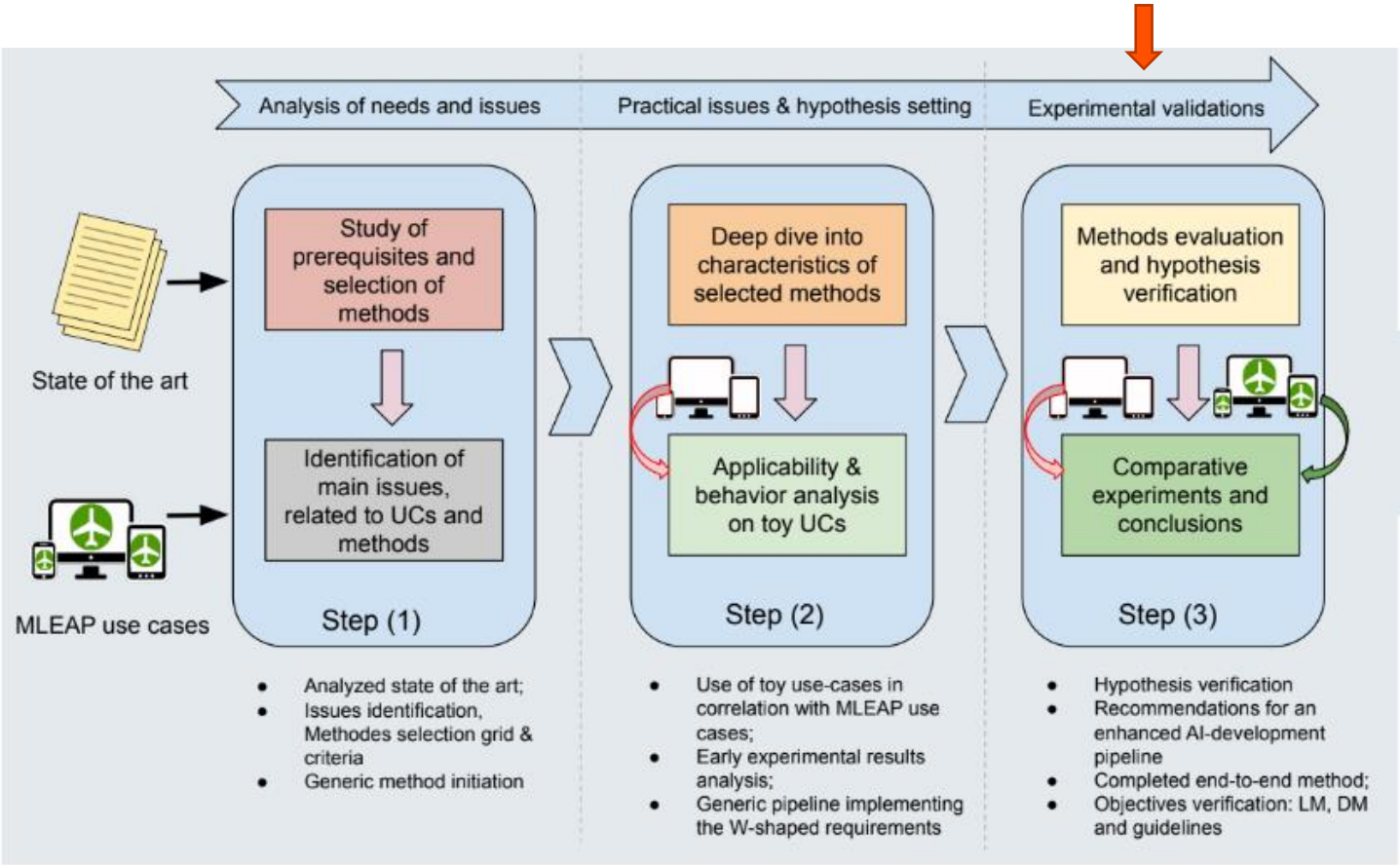
ML technologies to be used in safety related applications >> selection of relevant use-cases, representing real-life safety-related ML applications ;
Streamline the certification objectives of the EASA guidance >> focus on the objectives drawn from the EASA AI Roadmap

How?

For each of the identified objectives, set the MLEAP's strategy to meet the requirements and expectations, including an adapted experimental set up.

MLEAP workflow toward the objectives > > >

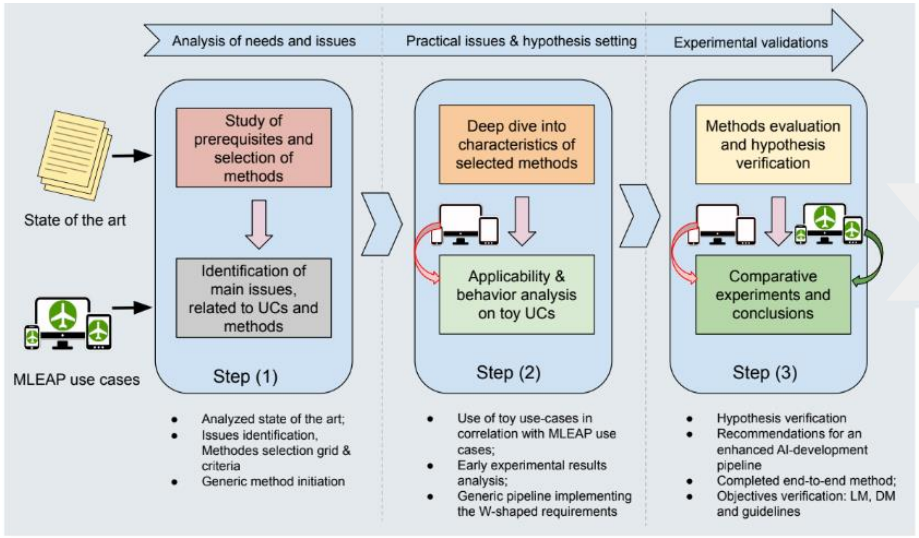
Project Steps

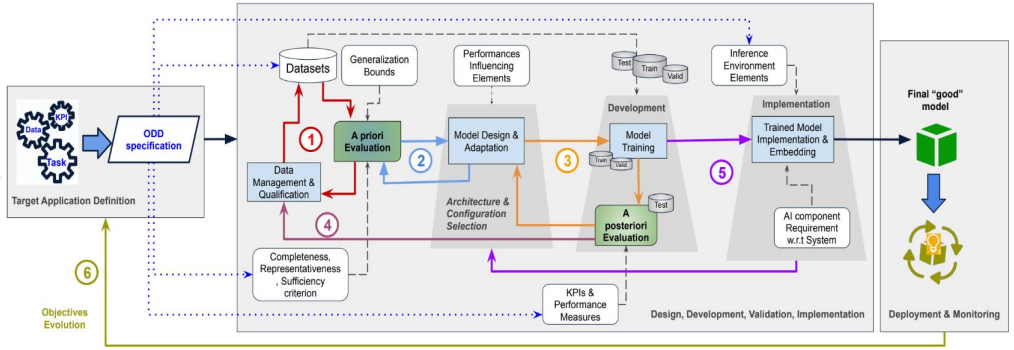


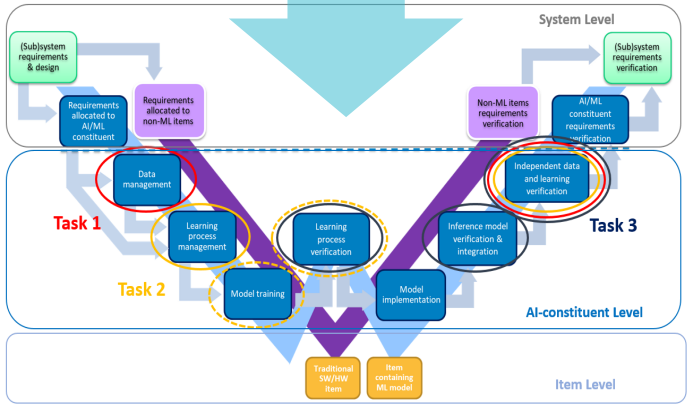
AIRBUS

MLEAP workflow toward the objectives > > >

Project Steps







AIRBUS

MLEAP workflow toward the objectives > > >

Use-Cases for experimental analysis

Toy use cases

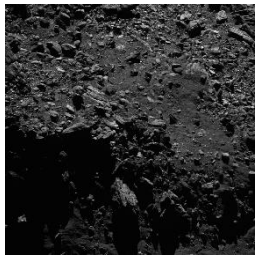
Simple, for preliminary analysis of selected methods

Real aviation use-cases

More complex, for approval of methods

Fashion-MNIST

MNIST



Rosetta

ATC-STT

AVI of aircrafts

Collision avoidance ACAS Xu

Several applications, with proprietary datasets & models and open source materials for public deliverables.

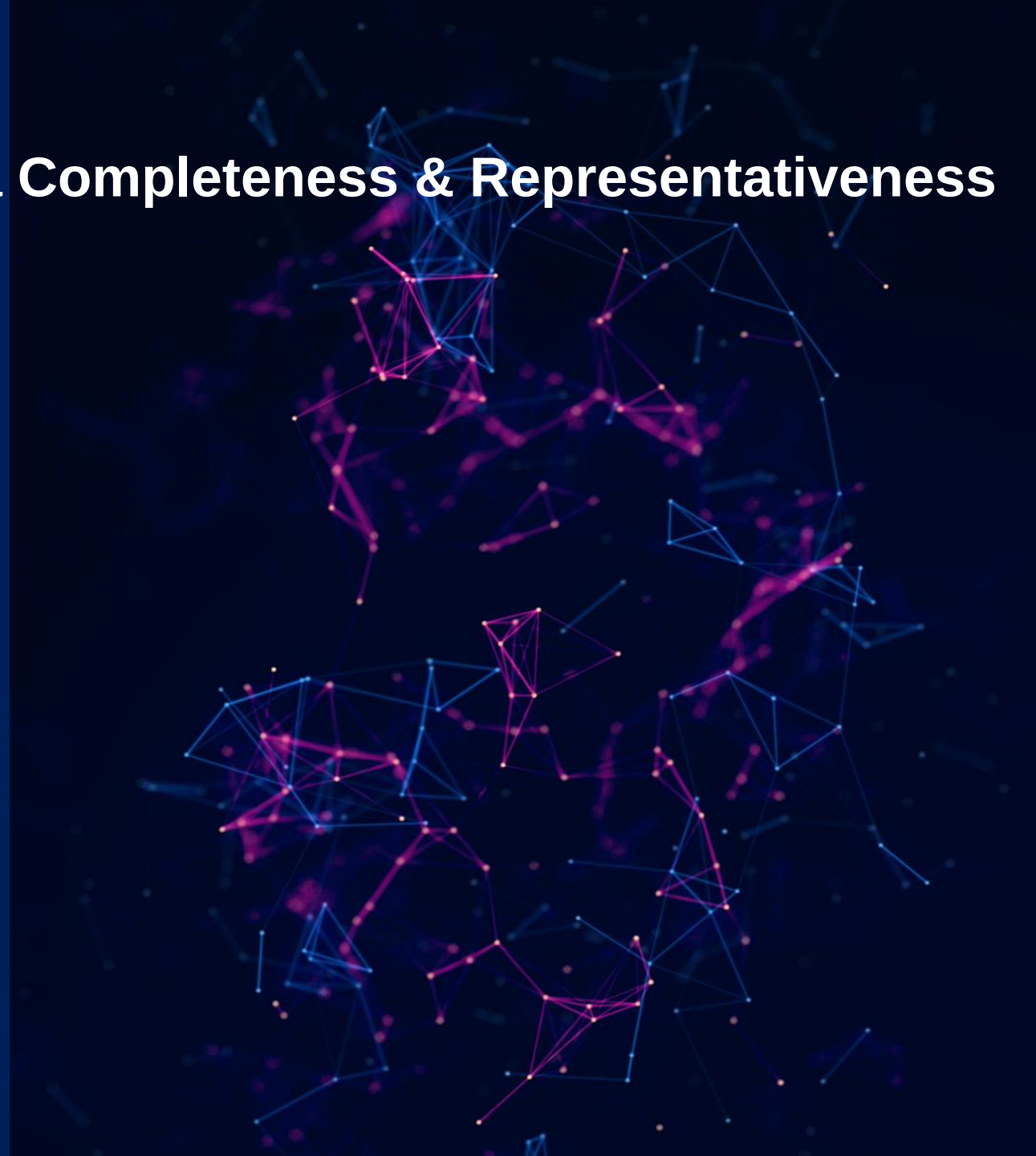
Source: https://github.com/zalandoresearch/fashion-mnist

AIRBUS

MLEAP – Task #1 milestones: Data Completeness & Representativeness

Completeness: *A data set is complete if it sufficiently covers the entire space of the operational design domain for the intended application*.

Representativeness: *A data set is representative when the distribution of its key characteristics is similar to the actual input space of the intended application*



Task 1 : Data completeness and Representativeness > > >

Task #1 objectives (so far)

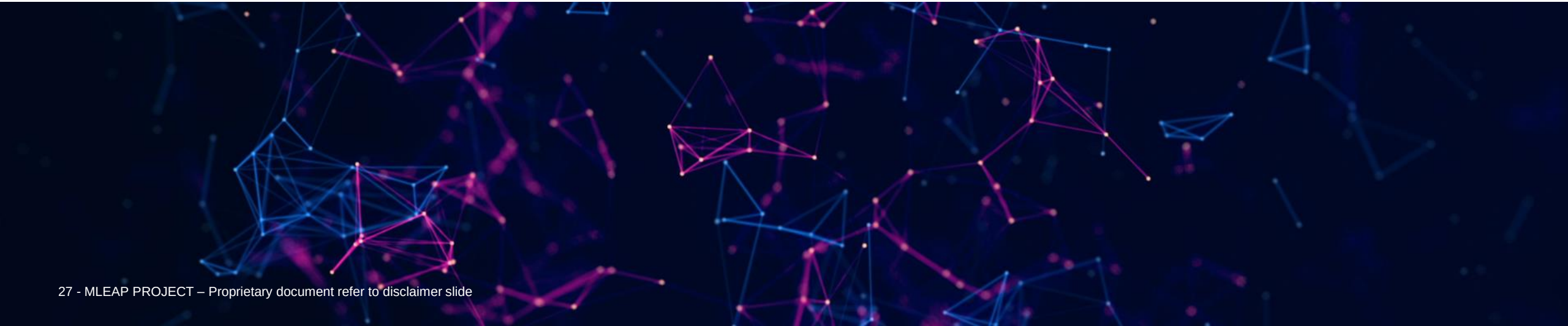
State-of-the-art (phase 1): Provide a list of factors influencing the choice of tools and approaches in order to assess the completeness and representativeness of databases, with corresponding justifications and bibliographical references.

80+ sources discussed

Synthesis (phase 1): Present a draft structure of the selection grid for the assessment tools and methods.

19 methods identified for testing

Testing (phases 2, 3 and 4): Identification or development of efficient and practicable methods and tools for the assessment of completeness and representativeness of data sets (training, validation and test) in the generic case of data-driven ML.



Task 1 : Data completeness and Representativeness > > >

Latest results

Off-the-shelf tools case study : Cleanlab

Already used in the industry (Google, Tesla)

Open source (mostly)

Preliminary experiments using MNIST dataset

Cleanlab detects :

- outliers
- near duplicates

} Model-dependent for non-tabular data

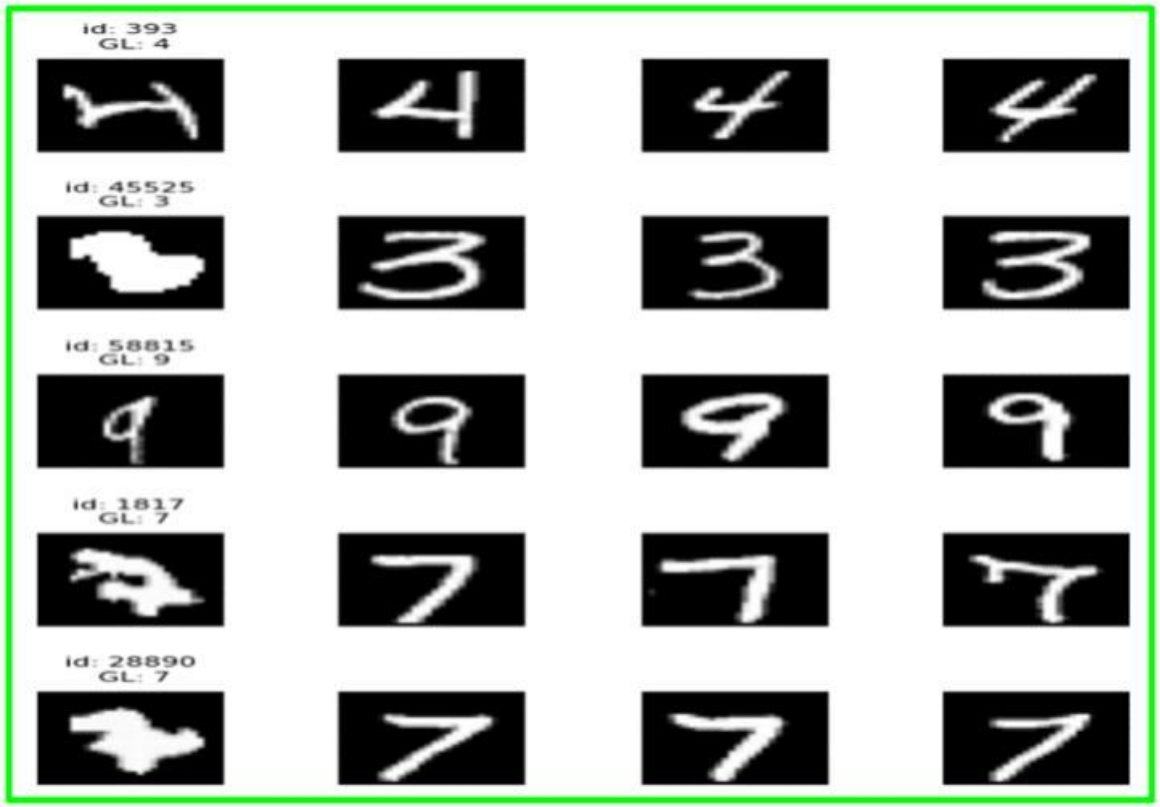
- labelling errors : Always model-dependent
- non i.i.d samples

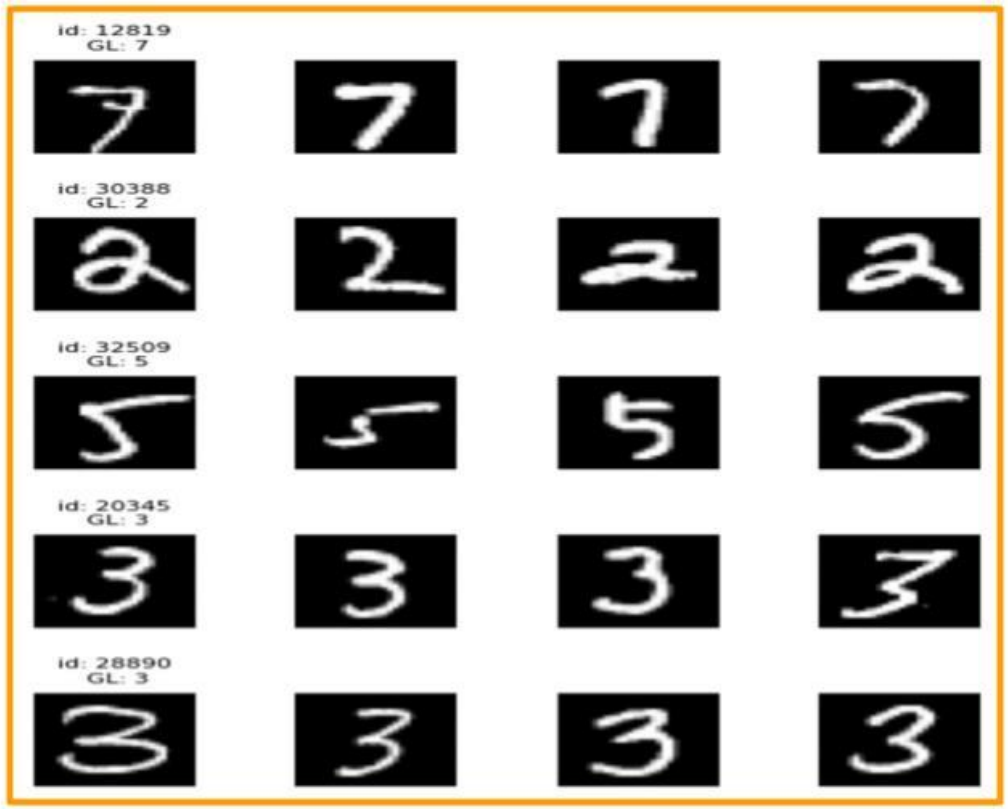
2 classifiers trained : 97% and 75% accuracy

Contrastive study

Task 1 : Data completeness and Representativeness > > >

Experiments: Outliers

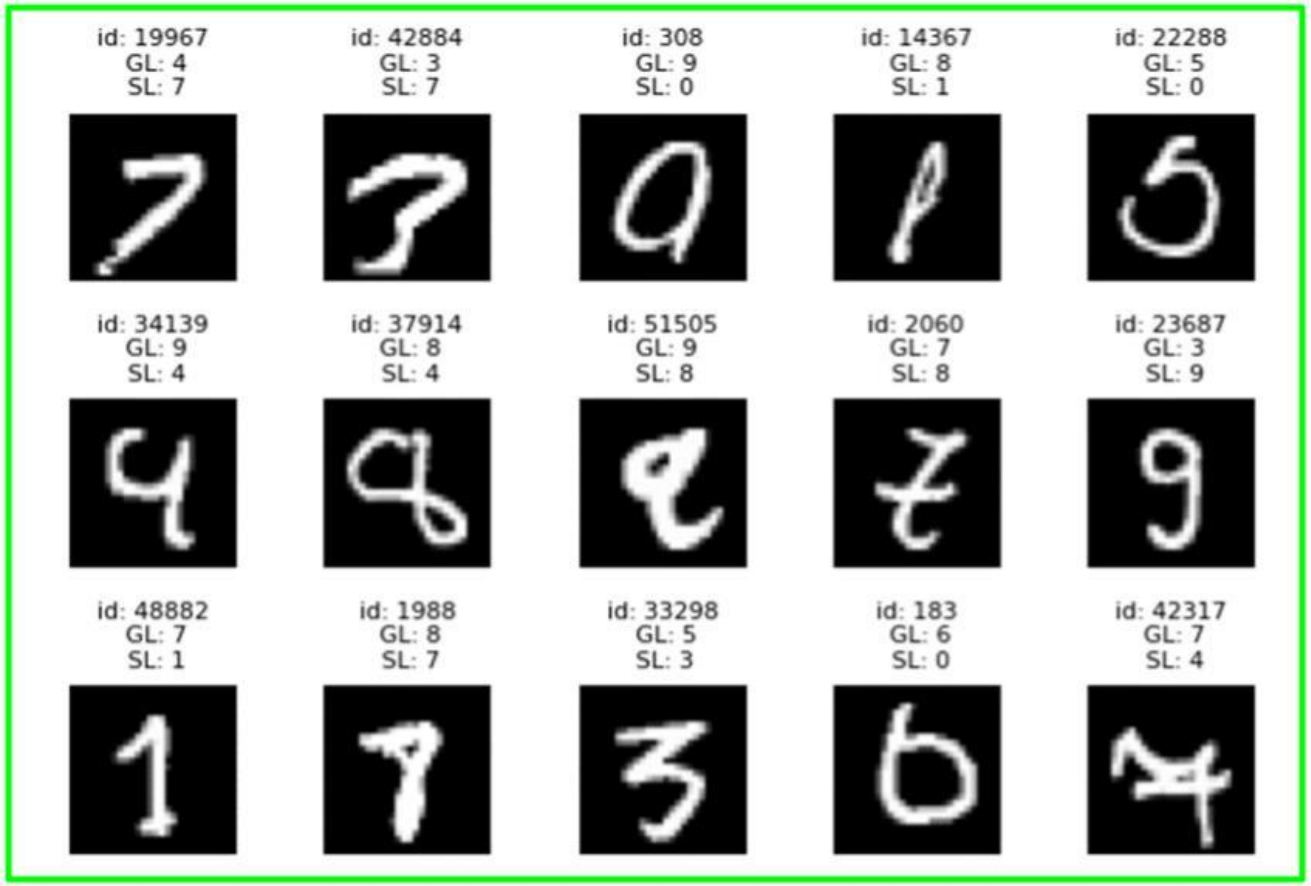


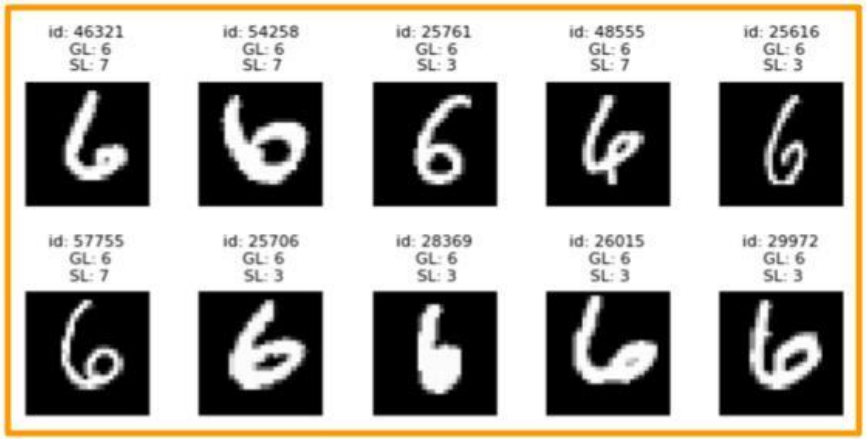


AIRBUS

Task 1 : Data completeness and Representativeness > > >

Experiments: Mislabelings

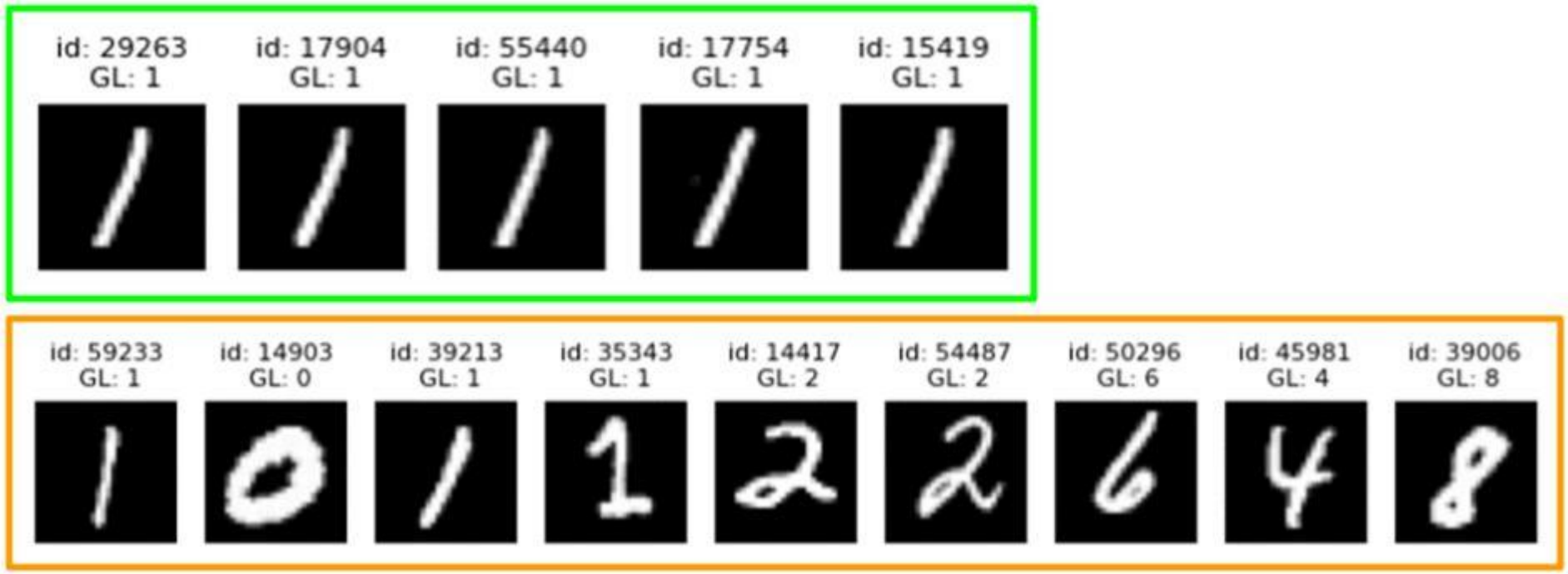




AIRBUS

Task 1 : Data completeness and Representativeness > > >

Experiments: Near-duplicates



AIRBUS

Task 1 : Data completeness and Representativeness > > >

Cleanlab: Takeaway

(Broad) conclusions on Cleanlab
Edge cases
Hard cases
Mislabelings

Model dependence helps tailor the data set to the model

Model dependence requires a mature model
Closer to production :

Human analysis is required
Large datasets will yield more issues requiring more manpower

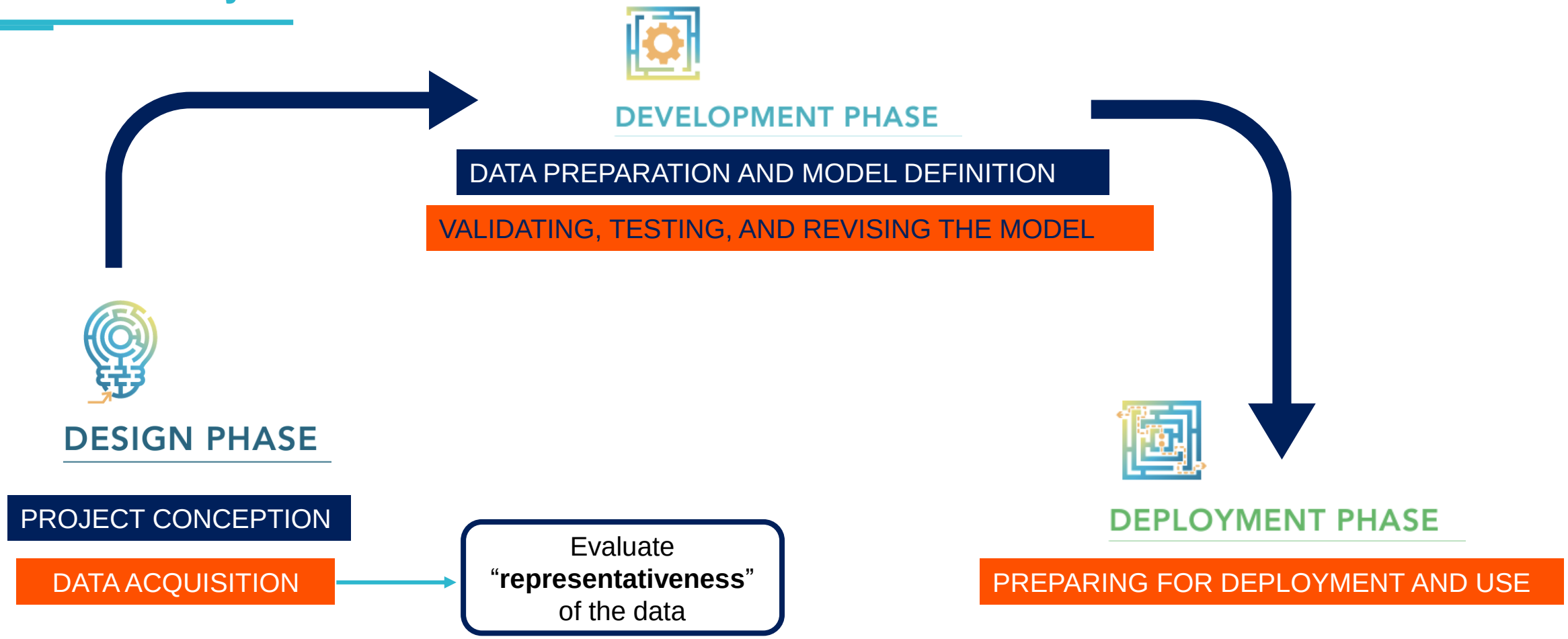
AIRBUS

Risk-based approach
Experimenting BSA‘s Framework

Designed for demographic data and bias risk

Task 1 : Data completeness and Representativeness > > >

BSA's AI Lifecycle



Source: BSA's AI bias framework

AIRBUS

Task 1 : Data completeness and Representativeness > > >

No homogeneity between the checkpoints

Compare demographic distribution of training data to the population where the system will be deployed	> > >	Requires a broad comparison
Assess whether there is sufficient representation of subpopulations that are likely to interact with your system	> > >	Requires an assessment + a threshold to be determined

Source: BSA's AI bias framework

Task 1 : Data completeness and Representativeness > > >

Non-explicit checkpoints

Compare demographic distribution of training data to the population where the system will be deployed

Assess whether there is sufficient representation of subpopulations that are likely to interact with your system

Source: BSA's AI bias framework

What is missing for a more rigorous analysis?

The wording of the checkpoints should hint about the required results

> > >

"The demographic distribution of training data is closely similar to the distribution of the population where the system will be deployed"

"The subpopulations that are likely to interact with the system are sufficiently represented."

Clearer and detailed instructions on how to proceed to the comparison

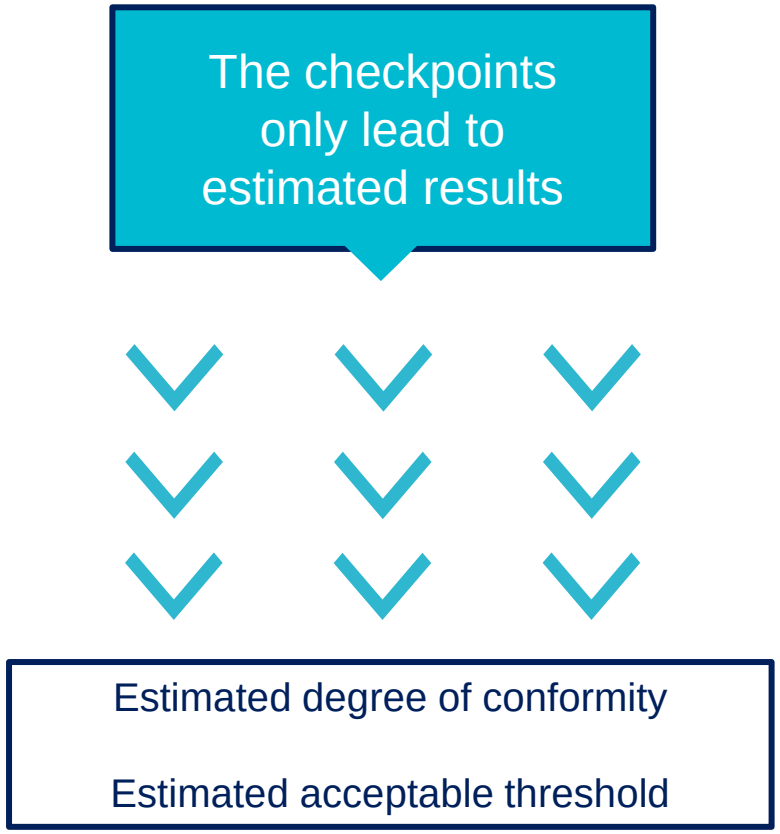
> > >

Relevant statistics ; expert analysis of the context of use

AIRBUS

Task 1 : Data completeness and Representativeness > > >

Only approximate results can be expected



AIRBUS

Task 1 : Data completeness and Representativeness > > >

Experimental method

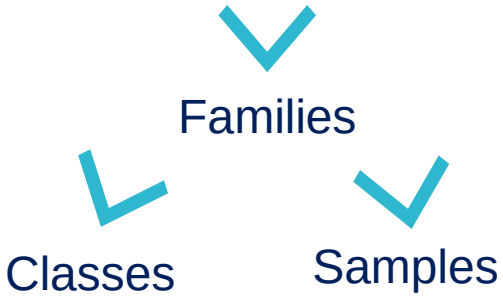
ROSE data set example

Weeds detection for agricultural machines in Europe

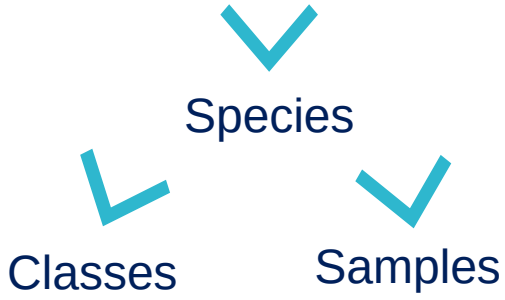
69714 images capturing the soil of a field

Crops and weeds plants annotated by polygonal bounding boxes

“The demographic distribution of training data is closely similar to the distribution of the population where the system will be deployed”



“The subpopulations that are likely to interact with the system are sufficiently represented.”



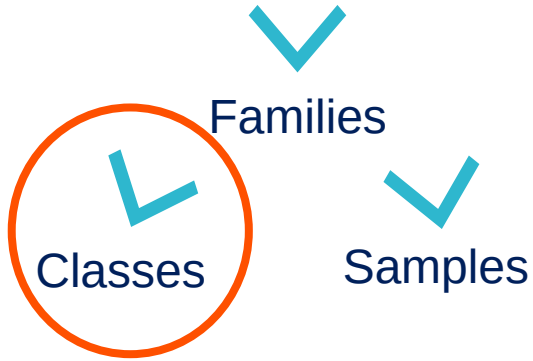
AIRBUS

Task 1 : Data completeness and Representativeness > > >

Analyzing by classes and samples

ROSE data set example

"The demographic distribution of training data is closely similar to the distribution of the population where the system will be deployed"



Representativeness ratio
= Number of families in the data set / Number of families in the target population (Europe)

= 4/48

= 0,08

Non-representative

AIRBUS

Task 1 : Data completeness and Representativeness > > >

Target population : the difficulty of finding and choosing relevant statistics

THEY MAY NOT EXIST

ROSE data set

Most relevant statistics :

Weeds species distribution in Europe

Existing statistics :

Only a survey of weeds that are increasingly spreading in Europe from 2005

THE CHOICE MAY NOT BE EVIDENT

Voxcrim data set

Most relevant statistics :

Demographics of France

OR

French prisons statistics

The “actual input state space” from EASA AI concept paper Iss2 is not considered

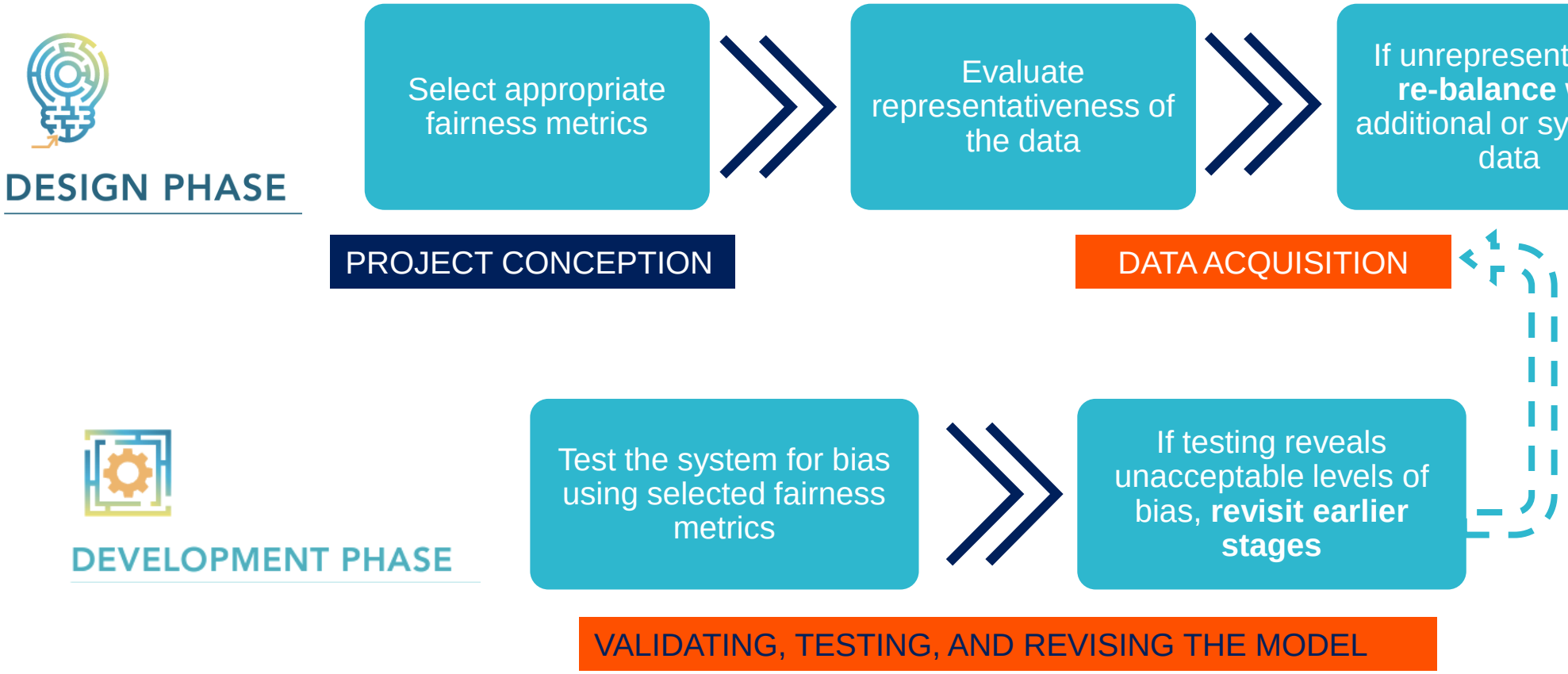
⚠ Risk of selection bias

⚖ Non-discrimination principle

AIRBUS

Task 1 : Data completeness and Representativeness > > >

Fairness constraints not taken into account during the evaluation of representativeness



AIRBUS

Task 1 : Data completeness and Representativeness > > >

A blind spot in BSA's method

A data set can be representative of the "actual input state space" (EASA, 2023), and at the same time reveal unfair biases.

Some cases need to prioritize fairness over obtaining a representative demographic distribution, and only seek sufficient representation of subpopulations.

Then, only the 2nd checkpoint may be relevant (required sufficient representation of subpopulations).

This scenario is not anticipated in BSA's method.

→ This can lead to counterproductive actions.

AIRBUS

Task 1 : Data completeness and Representativeness > > >

Suggested alternative mitigation practice

BSA only recommends to re-balance with additional or synthetic data.

Alternative

If re-balancing is inadequate or unrealistic :

The scope of the use case could narrow and become more adapted to the available data.

ROSE example :

From all European weeds to only 4 species

AIRBUS

Task 1 : Data completeness and Representativeness > > >

Takeaways & next steps

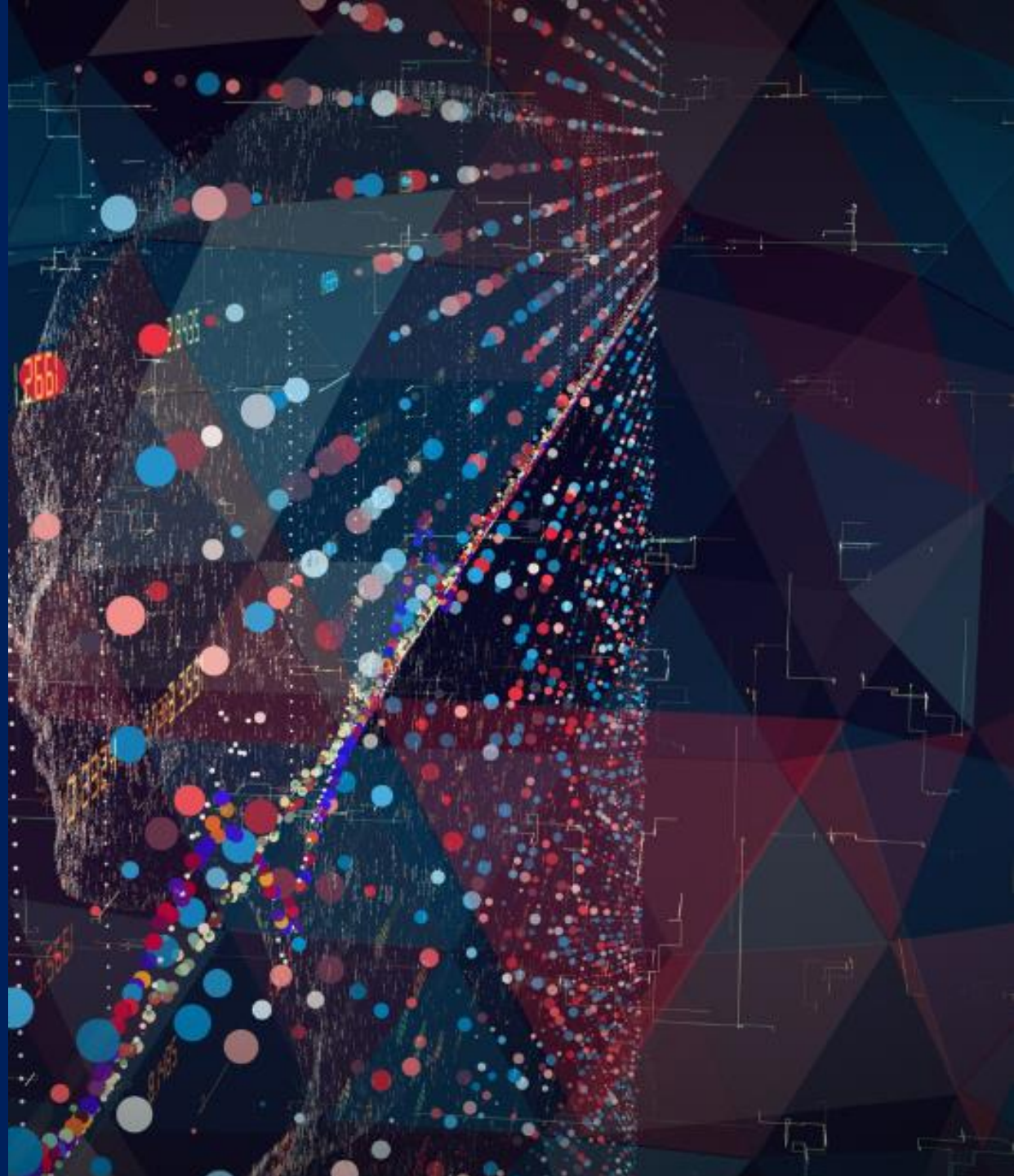
Takeways	Next steps
Need for a priori assessment based on the ODD… …but also a posteriori assessment based on the feedbacks of the model	Applying methods on the MLEAP use cases

Task #2: Generalization Properties

Objective:

Identification or development of efficient methods and tools for the quantification of generalization assurance level in the generic case of data-driven ML/DL development

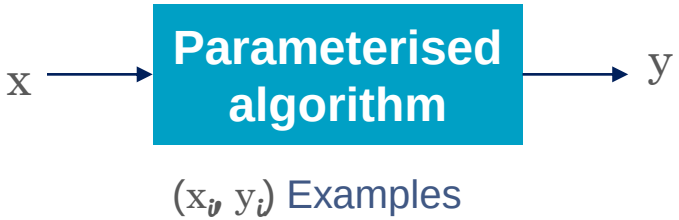
- Test available methods and tools to evaluate generalization bounds;
- Barriers in generalization guarantees for a given model: ML and DL;
- Identification/proposal of means to promote models generalization.



MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Supervised machine learning

Objective: Estimate the response *y* from the data *x*



Training: optimize algorithm parameters to minimize errors on the examples

Generalizability: We are expecting few errors on unseen data. It is based on the assumption that we have regularities behind the data which are discovered during the training phase.

Generalizability assessment:

- Performance measure on test and validation dataset
- Generalization bounds:
  - Upper bounding the Expected true risk
  - Generalization quality and “good” model identification
  - Practical guidance
- Development workflow steps influence

Algo.	Ref.	Bound
CNN	(Lin and Zhang, 2019)	$R_{\mathcal{D}}(F_C) \leq \hat{R}_{S,l_\eta}(F_C) + \tilde{O}\left(\frac{L^{\frac{L-1}{4}} \frac{3}{4} a^{\frac{1}{4}} c^{\frac{1}{2}} m^{\frac{1}{8}} r^{\frac{1}{2}}}{\sqrt{N}}\right) + \tilde{O}\left(\frac{L^{\frac{L-1}{4}} \frac{3}{4} a^{\frac{1}{4}} d}{\sqrt{N}}\right)$
NN for classification	(P. Jin et al., 2020)	$\mathbb{P}_{\mathcal{D}}\left[\forall h_\theta \in \mathcal{H}, \mathbb{E}_{x,y\sim\mathbb{P}(X,Y)}\left[\mathcal{R}(h_\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)\right] + \frac{\sqrt{d}.CD(\mathcal{D})}{\min(\delta_0, \kappa\delta_{\mathcal{D}})}\right] \geq 1-\delta$
NN	(Alquier, 2021)	Cantoni’s bound (PAC Bayes) $\mathbb{P}_{\mathcal{D}_0}\left[\forall\rho\in\mathcal{P}(\Theta), \quad \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}(h_\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}_0}_{emp}(h_\theta)\right] + \frac{\lambda C^2}{8N_0} + \frac{KL(\rho\|\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right] \geq 1-\varepsilon$
	(Alquier, 2021) (McAllester, 1998)	Mc Allester’s bound $\mathbb{P}_{\mathcal{D}_0}\left[\mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}_0}_{emp}(\theta)\right] + \sqrt{\frac{KL(\rho\|\pi) + \log\frac{1}{\varepsilon} + \frac{5}{2}\log(N_{\mathcal{D}_0}) + 8}{2N_{\mathcal{D}_0} - 1}}\right] \geq 1-\varepsilon$
	(Alquier, 2021) (Seeger, 2002)	Seeger’s bound $\mathbb{P}_{\mathcal{D}}\left[\forall\rho\in\mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}(h_\theta)] \leq kl^{-1}\left(\mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)] \middle| \frac{KL(\rho\|\pi) + \log\frac{2\sqrt{N_{\mathcal{D}}}}{\epsilon}}{N_{\mathcal{D}}}\right)\right] \geq 1-\epsilon$
	(Alquier, 2021) (Tolstikhin and Seldin, 2013)	Tolstikhin and Seldin’s bound $\mathbb{P}_{\mathcal{D}}\left[\begin{array}{c}\forall\rho\in\mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] \leq \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)] + \\ \sqrt{2\mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)]\frac{KL(\rho\|\pi) + \log\frac{2\sqrt{N_{\mathcal{D}}}}{\epsilon}}{2N_{\mathcal{D}}}} + 2\frac{KL(\rho\|\pi) + \log\frac{2\sqrt{nN_{\mathcal{D}}}}{\epsilon}}{2N_{\mathcal{D}}}\end{array}\right] \geq 1-\epsilon$

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

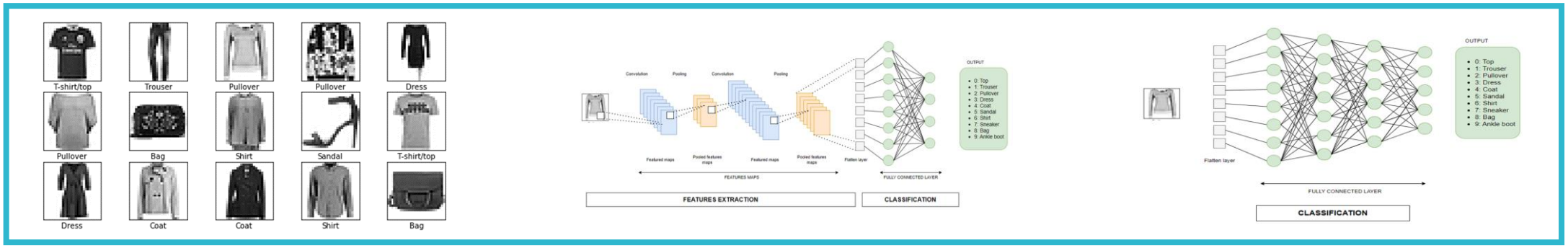
Generalization bounds - experimentation

Experimental objectives:

Test Generalization bounds as MoC to answer Objective SA-01, LM-04 and LM-09 (generalization guarantees by bounding empirical risk measure and true risk)
Check generalization bounds theories support in model architecture selection

Experimental protocol:

Targeted task: classification of fashion MNIST images (Assumption: correct completeness and representativeness)
Tests and analysis a priori results on 2 different architectures FCNN and CNN
Train and test several models
Analyse a posteriori generalization bounds regarding assurance level upper bounds



AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Generalization bounds - experimentation

A Priori evaluation:

Pessimistic as the theory remain valid in worst case and are vacuous in our case

Pac Bayes bounds, complexity bounds and margin bounds encourage minimum parameters (minimum complexity)

A Posteriori evaluation:

Tighter bounds but still too high to provide efficient assurance level regarding average loss

Naive application of the bounds do not provide accurate and self-sufficient means to guarantee the generalizability of the used models.

A priori generalization bounds / epsilon = 0.05 (95% confidence)							
		CNN	CNN	CNN	FCNN	FCNN	FCNN
	Assumptions for Apriori evaluation	1	2	3	1	2	3
Lin's Bound	spectral norm lower than 10 for FC layers Convolutional weights lower than 10	172	202	153	136	62218	11909
Jin's bound	Cover difference of the dataset						
Cantoni's bound	KL divergence upper bounded by a function of the number of parameters	55	45	306	21	829	134
McAllester's bound	KL divergence upper bounded by a function of the number of parameters	7	6	17	4	28	11
Seeger's bound							
Tolstikhin and Seldin's bound	KL divergence upper bounded by a function of the number of parameters	1664	1503	3918	1023	6438	2592
"Arora" bound	cushion is lower then 1/sqrt(#param)	9	21	4	3	13	13
Anthony's bound							
Neu and Lugosi's bound							
Feldman's bound	Stability w.r.t. Dtrain is 0.2	11	11	11	11	11	11
Hardt's bound	gradient of the loss function over iterations is lower than 1, Norm of parameters is lower than 1, and the number of iterations is 30	1.8	1.8	1.8	1.8	1.8	1.8
Lei's bound	delta (data Decision Boundary variability) is lower than 0.5 and delta is less than 1	10	10	10	10	10	10
Kawaguchi's bound							

Table 21. A priori evaluation of generalization bounds

A posteriori generalization bounds / epsilon = 0.05 (95% confidence)						
	CNN	CNN	CNN	FCNN	FCNN	FCNN
	1	2	3	1	2	3
Lin's Bound	11	19	101	1.77	147	2.17
Jin's bound	2.56	2.45	2.18	2.47	2.84	2.21
Cantoni's bound	14.4	14	27.8	9.8	66.8	20.3
McAllester's bound	1.8	1.8	2.9	1.2	4.9	2.4
Tolstikhin and Seldin's bound	6.7	6.5	17.4	3	48.6	11.4
"Arora" bound	9	21	4	3	13	13
Feldman's bound	11	11	11	11	11	11
Hardt's bound	1.62	1.54	1.59	1.74	1.67	1.6
Lei's bound	10	10	10	10	10	10

Table 22. A posteriori evaluation of generalization bounds

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Generalization bounds - takeaways

Generalization bounds:

- Difficult to obtain tight bounds with naive application
- Based on theorems of the bounds, some key elements are boosting generalization such as:
  - Regularization (dropout, batch size)
  - Early stopping and Optimization methods
  - Pooling for CNN

Alternative to generalisation bounds

- Alternatives to support overall safety case, architecture and hyperparameters selections could be:
  - Uncertainty Quantification
  - Conformal prediction
  - K-Fold Cross-Validation.

Next steps:

- Development workflow impact assessment of non naive application of generalization bounds
- Airbus use cases application
- Data preparation and analysis impact w.r.t. generalization

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: ATC-STT – Models evaluation

Objective: correctly translate spoken instructions ATCO to text for safer monitoring.

Target: 10% WER

Datasets:

AIRBUS dataset (real ATC exchange from French airports)

Open-source datasets (from European airports)

Models:

AIRBUS model, based on the Vosk API (no Deep Learning), trained on AIRBUS dataset

Open-source models, based on a transformers architecture, trained on the open-source datasets

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: ATC-STT – Models evaluation

Evaluation metric:

Word Error Rate (WER)

Results interpretation: Excellent performances of the AIRBUS mod
source datasets. Possible overfitting due to:

- Source of data (from a few French airports)
- Audio quality (noise, microphone used,…)
- Model technology (Vosk API)

Model \ Dataset		AIRBUS	ATCO2
Kaldi-based		11.43 %	91.05 %
transformer-based (1)	Original	43.70 %	45.54 %
	Fine-tuned	15.13 %	28.75 %
transformer-based (2)	Original	34.63 %	36.27 %
	Fine-tuned	14.76 %	29.85 %

Table 27 Comparison of the transformer-based models performances, in terms of WER measure, before and after fine-tuning on the AIRBUS training dataset. The evaluation is then performed on both the AIRBUS and ATCO2 datasets.

Pipeline analysis:

Model selection: real time constraints VS performance
Dataset representativity regarding the ODD

Next steps:

Training and optimization adaptation and models fine tuning w.r.t. the different objectives of the ML module

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: AVI – Models evaluation

Objective: help operators to perform the in-service damage detection, to reduce the aircraft maintenance duration, for scheduled and unscheduled events.
Target: 95% accuracy

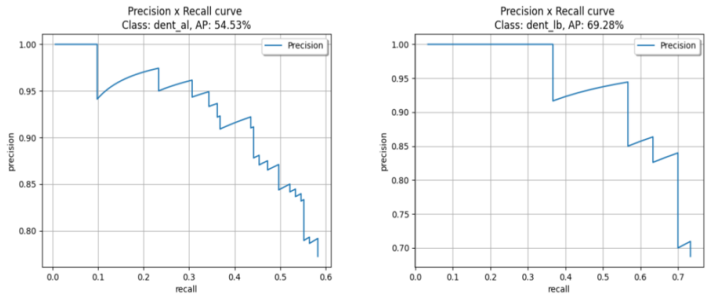
Datasets: AIRBUS dataset (pictures of surface damages detected and classified for lightning strikes and dents)

Models: YOLOv5 fine tuned model to minimize errors:

- damages location and dimension
- classification error
- no object detection error

Evaluation metric:

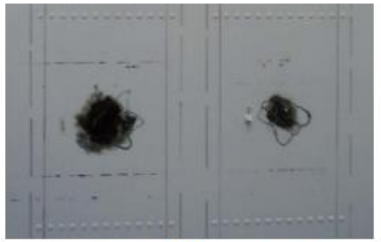
IoU (intersection over union)



(a) Classe dent_al (b) Classe dent_lb

Figure 65 Accuracy *versus* Recall *curves, with* $IoU_{th} = 0.5$*, corresponding to a trained YoloV5 model for detection of two types of dent instances.*

Dents Damages (1)



Lightning Strike impacts (2)

1) https://www.researchgate.net/figure/Wing-skin-metal-dent-examples_fig3_331961295
2) https://www.researchgate.net/figure/Structural-damage-in-the-outer-skin-in-the-Airbus-A400-M-airplane-after-the-lightning_fig8_305817924

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: AVI – Models evaluation

Results interpretation

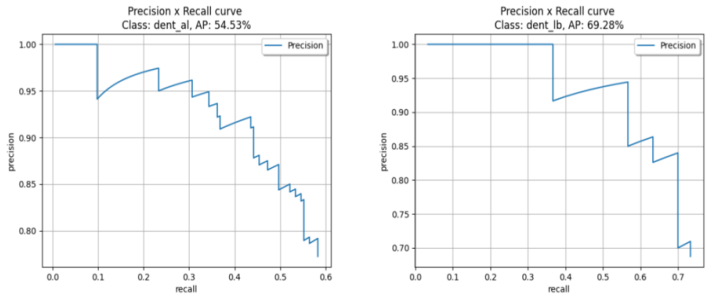
Due to limited data amounts, especially for lightning strikes, the obtained performances (**41%** on lightning strikes and **61.91%** on dents) do not meet the target objective of a 95% accuracy.

Pipeline analysis:

Limited amount of data -> low performance compared to the target

Model architecture could be adapted w.r.t. The targeted task

Loss function not normalized driving difficulties in model comparison

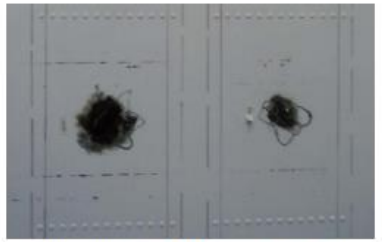


(a) Classe dent_al (b) Classe dent_lb

Figure 65 Accuracy *versus* Recall *curves, with* $IoU_{th} = 0.5$*, corresponding to a trained YoloV5 model for detection of two types of dent instances.*

Nest steps: Data augmentation with simulated data and Segmentation enabled model (YOLOv8)

Dents Damages (1)



Lightning Strike impacts (2)

1) https://www.researchgate.net/figure/Wing-skin-metal-dent-examples_fig3_331961295
2) https://www.researchgate.net/figure/Structural-damage-in-the-outer-skin-in-the-Airbus-A400-M-airplane-after-the-lightning_fig8_305817924

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: ACAS Xu Task – Models evaluation

Objective: reduce the storage space required to run ACAS Xu systems. Target: 100% accuracy

Datasets: Radio Technical Commission for Aeronautics (RTCA) Special Committee 147. The data consists of different entries of the LUTs from the RTCA SC-147 MOPS (600 Million of possible input)

Models: 45 neural networks - FNN with 6 hidden layers (is one NN for each pair time until loss of vertical separation and the last provided instruction)

Evaluation metric: Classification cross entropy

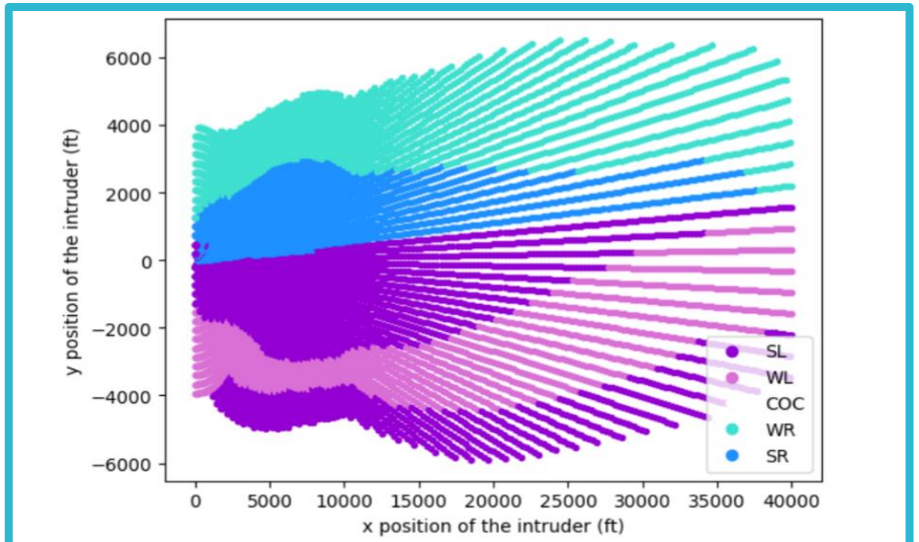


Figure 68 Instruction of the neural network $N_{0,COC}$ as w.r.t: the position of the intruder, the ownship being at (0,0) and moving along the x axis. With ($\psi = -\pi$) at and $v_{own} = v_{int} = 430$ ft/s.

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: ACAS Xu Task – Models evaluation

Results interpretation

Good models performance but not at 100% level regarding LUT approach
COC class overrepresented

Pipeline analysis:

Data unbalanced -> Introduction of weighted function to limit the impact
Model architecture adapted for classification task

Next steps:

Introduction of weighted function to limit the impact unbalanced class effect
Data augmentation: creation of intermediate artificial points for non COC classes

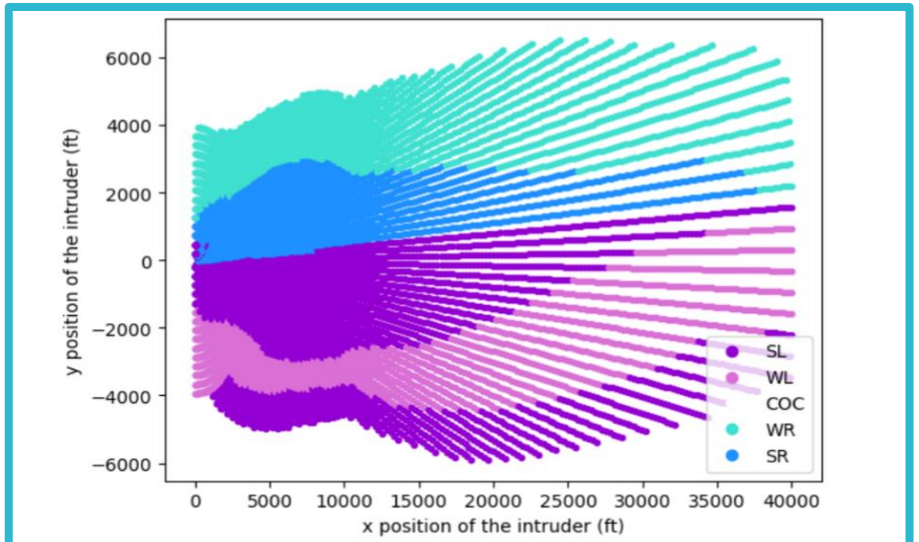


Figure 68 Instruction of the neural network $N_{0,COC}$ as w.r.t: the position of the intruder, the ownship being at (0,0) and moving along the x axis. With ($\psi = -\pi$) at and $v_{own} = v_{int} = 430$ ft/s.

AIRBUS

MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Experimentation: Conclusions

Generalization bounds:

Naive application of generalization bounds do not provide accurate and self-sufficient means to guaranty the generalizability of the used models

Confirmation that some methods are boosting generalization such as:

- Regularization (dropout, batch size)
- Early stopping and Optimization methods
- Pooling for CNN

Steps in development process - issues and limitations have been identified regarding the common practices:

Weak data processing when some hypothesis are violated (e.g independent and identically distributed hypothesis in test, train and validation datasets) and lack of data for optimal training

Gap between selected measures of performance and training objective (resulting of gap between the evaluation objectives and the industrial needs).

Model selection: architecture design w.r.t. Objectives and adaptation based on the detailed results

Task #3: Algorithme and model robustness

Task objective:

Review of methods and tools
Review of methods to identify corner cases and abnormal inputs
Identification of sources of instabilities during the design phase
Identification of sources of instabilities during the operational phase
Demonstration on a use-case for the intended application

MLEAP – Task #3 : Model evaluation – Robustness and Stability > > > – Summary

Multiple approaches available

Formal methods
- Solver
- Abstract interpretation
- Optimization
- Doable but with local results

Statistical methods
- Combining metrics
- Doable but through sampling

Empirical methods
- Field trial
- A posteriori
- Benchmarking
- Human intervention needed

Combining them is key

Property	Empirical	Statistical	Formal
Stability of the training algorithm			
Stability of the trained model			
Stability of the inference model			
Bias			
Variance			
Robustness (corner case exploration)			
Relevance			
Reachability			

} _Task 3: Model evaluation – Robustness and Stability

–#

Focus on the EASA concept

LM11: stability of the training algorithm

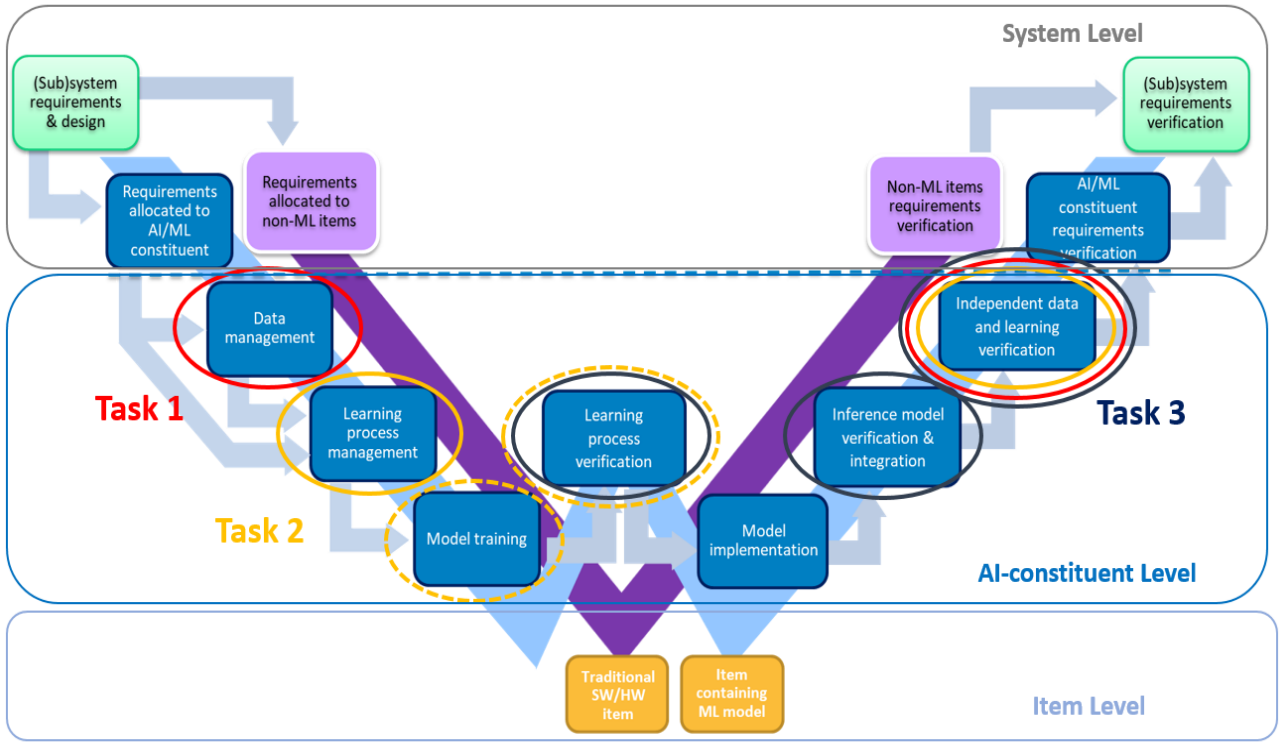
Very innovative requirement
Not much scientific results on the matter
Rather easy to setup
High risk of being difficult to fulfill

LM12: stability of the trained model

Already discussed in the standardization literature
Should be feasible with the right ODD
Low risk of being difficult to implement

LM13: robustness of the trained model

Already discussed in the standardization literature
Not necessarily easy to setup depending on the ODD
Medium risk of being difficult to implement



AIRBUS

} _Task 3: Model evaluation – Robustness and Stability

Toy examples

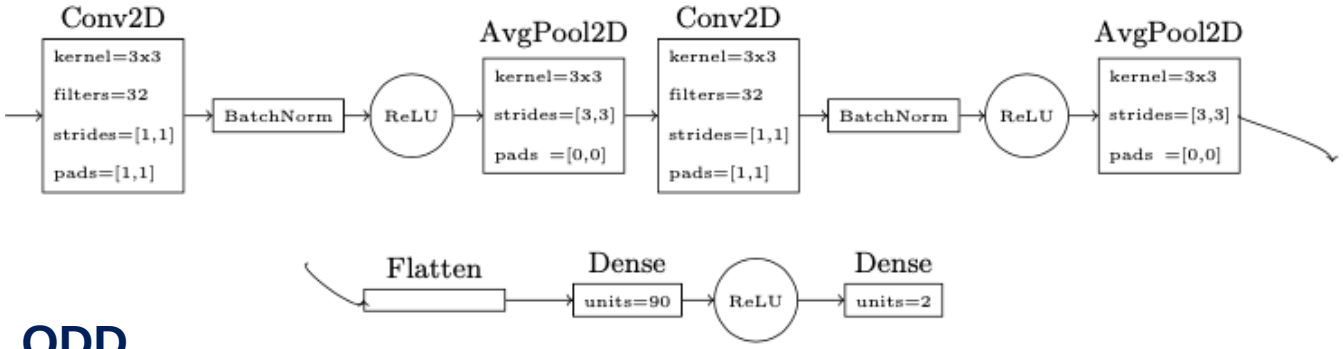
Model type	Origin	Data type	Dimensionality	LM	Actions to test
Classifier	Aerospace	Image	Small	LM11 LM12 LM13	Training algorithm stability General stability Stability against specific perturbations
Detector	Public domain	Image	High	LM12	Stability against generic perturbations
Classifier	Health care	Time series	Medium	LM11 LM12	General stability

} _Task 3: Model evaluation – Robustness and Stability

Imagine classifier

Statistical assessment of performance

2 classes
Confusion matrix >95% accuracy

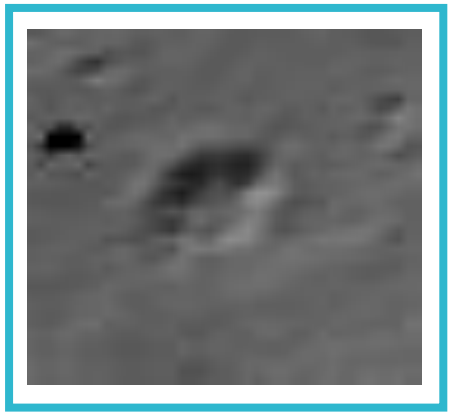


ODD

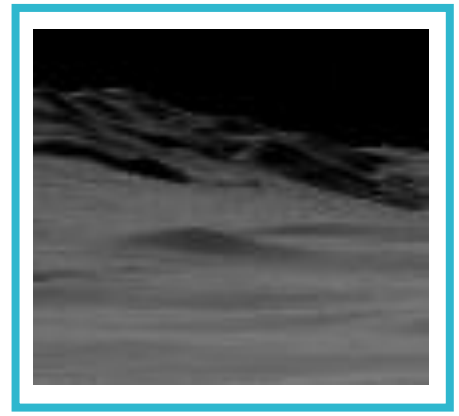
Can be defined by experts
But can still contained very unusual data points

Specific perturbations due to the space environment

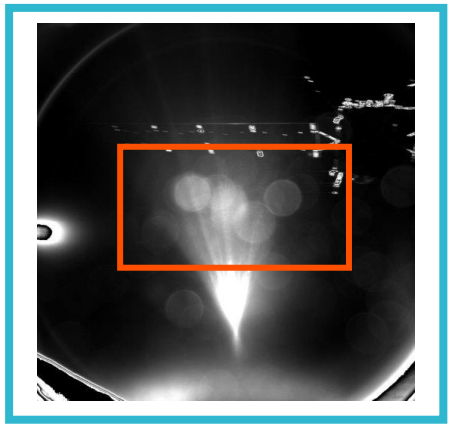
Flares
Radiation



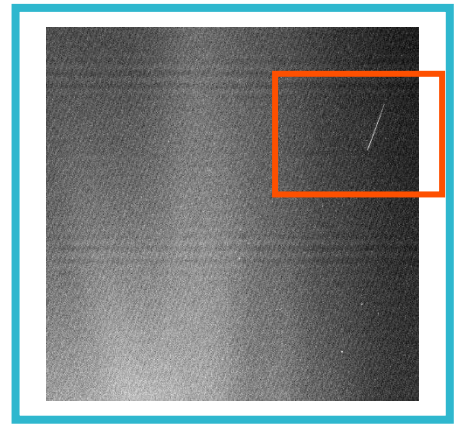
Crater



No crater



Flares



Radiation

•https://www.esa.int/Science_Exploration/Space_Science/Rosetta/Rosetta_image_archive_complete

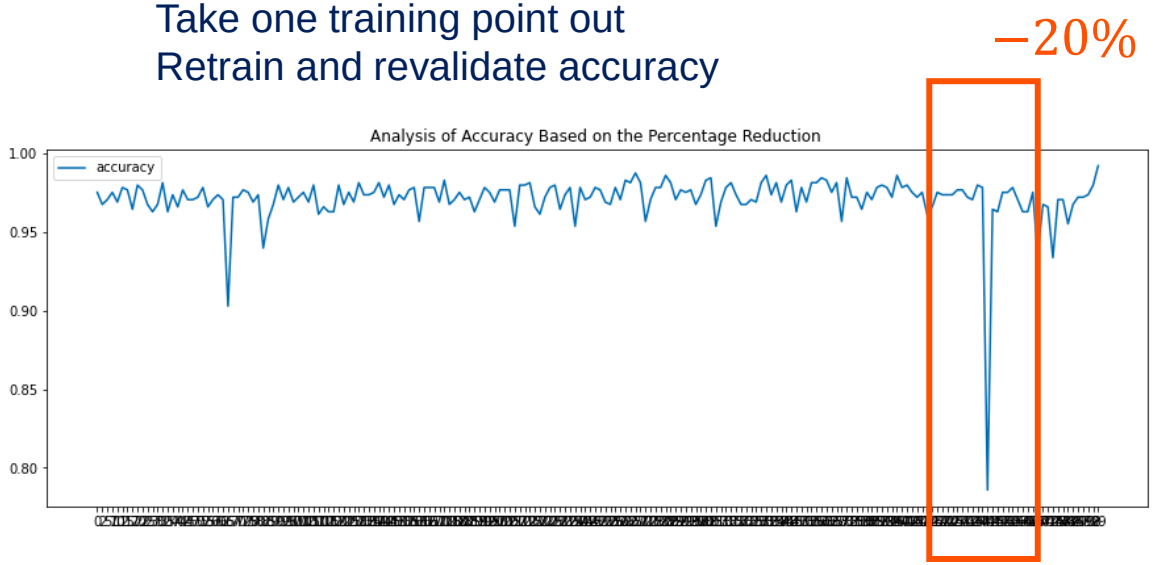
AIRBUS

} _Task 3: Model evaluation – Robustness and Stability

Imagine classifier

Training algorithm stability

Take one training point out
Retrain and revalidate accuracy

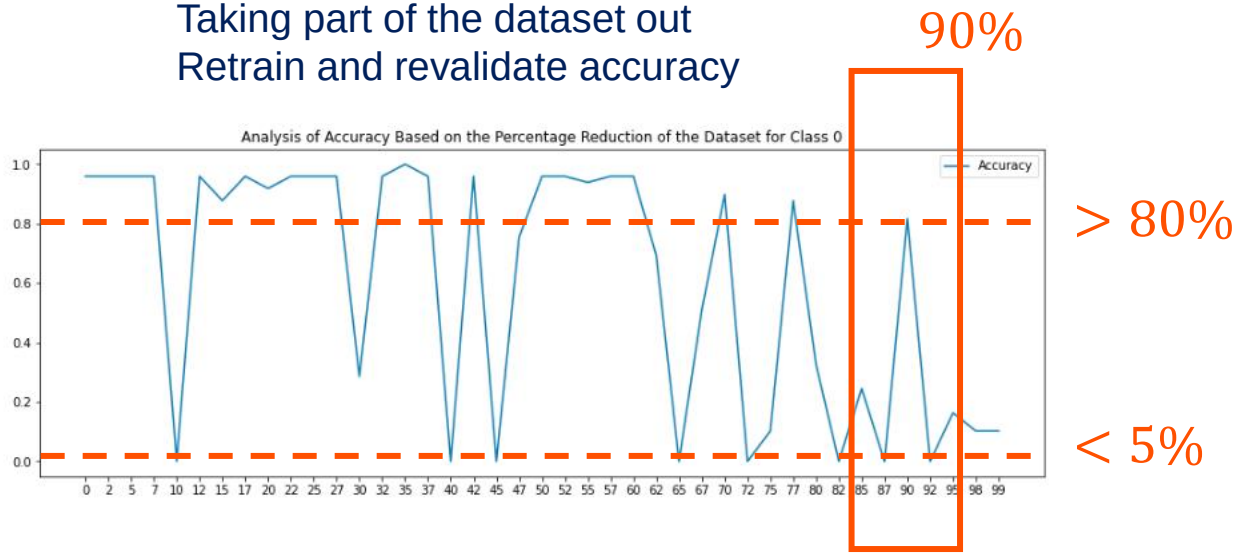


Could help measure training sensitivity

not really taken into account in the ecosystem

Training algorithm stability

Taking part of the dataset out
Retrain and revalidate accuracy



Could help measure the task inner difficulty

Link with Task 1 (dataset) and Task 2 (generalization)

LM11

AIRBUS

} _Task 3: Model evaluation – Robustness and Stability

Imagine classifier

General stability

Perturbation affecting all pixels
Formal methods to verify the stability of classification

	± 1 pixel variation	± 2 pixels variation
Zonotopes	1129 / 1312	72/1312
Polytopes	1212/1312	157/1312

Stability across the data set

Take Away

Model is easily unstable when considering variation on all pixels
Limitation of the formal approach or true vulnerability?

Future work

Check more local stability
Compare with adversarial attacks to found close counter-examples

LM12

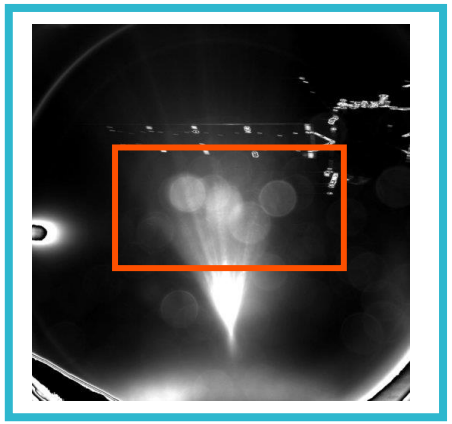
} _Task 3: Model evaluation – Robustness and Stability

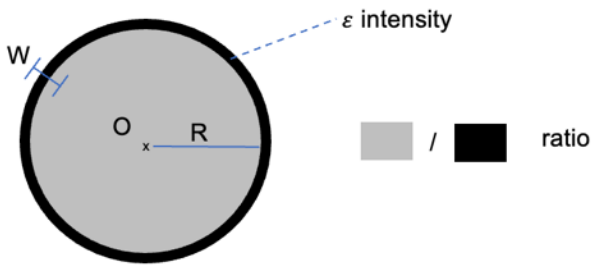
Imagine classifier

Stability against specific perturbations (related to the ODD)

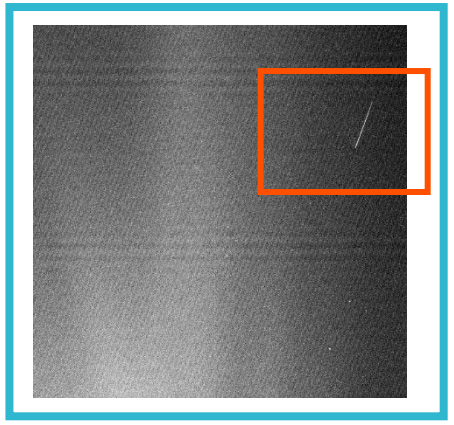
Requires a mathematical model of the perturbation for formal approaches
Validate on different levels of intensity of the perturbation

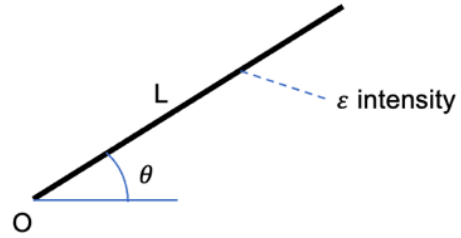
Flare





Radiation





LM13

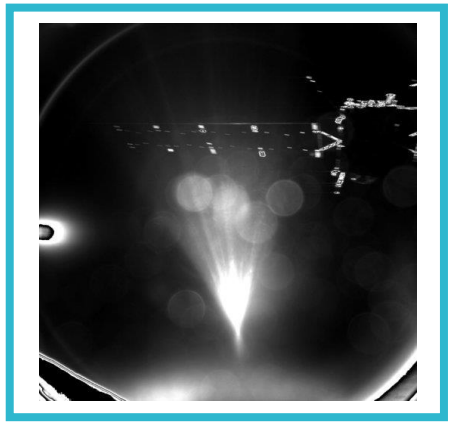
•https://www.esa.int/Science_Exploration/Space_Science/Rosetta/Rosetta_image_archive_complete

} _Task 3: Model evaluation – Robustness and Stability

Imagine classifier

Stability against specific perturbation (specific to the ODD)

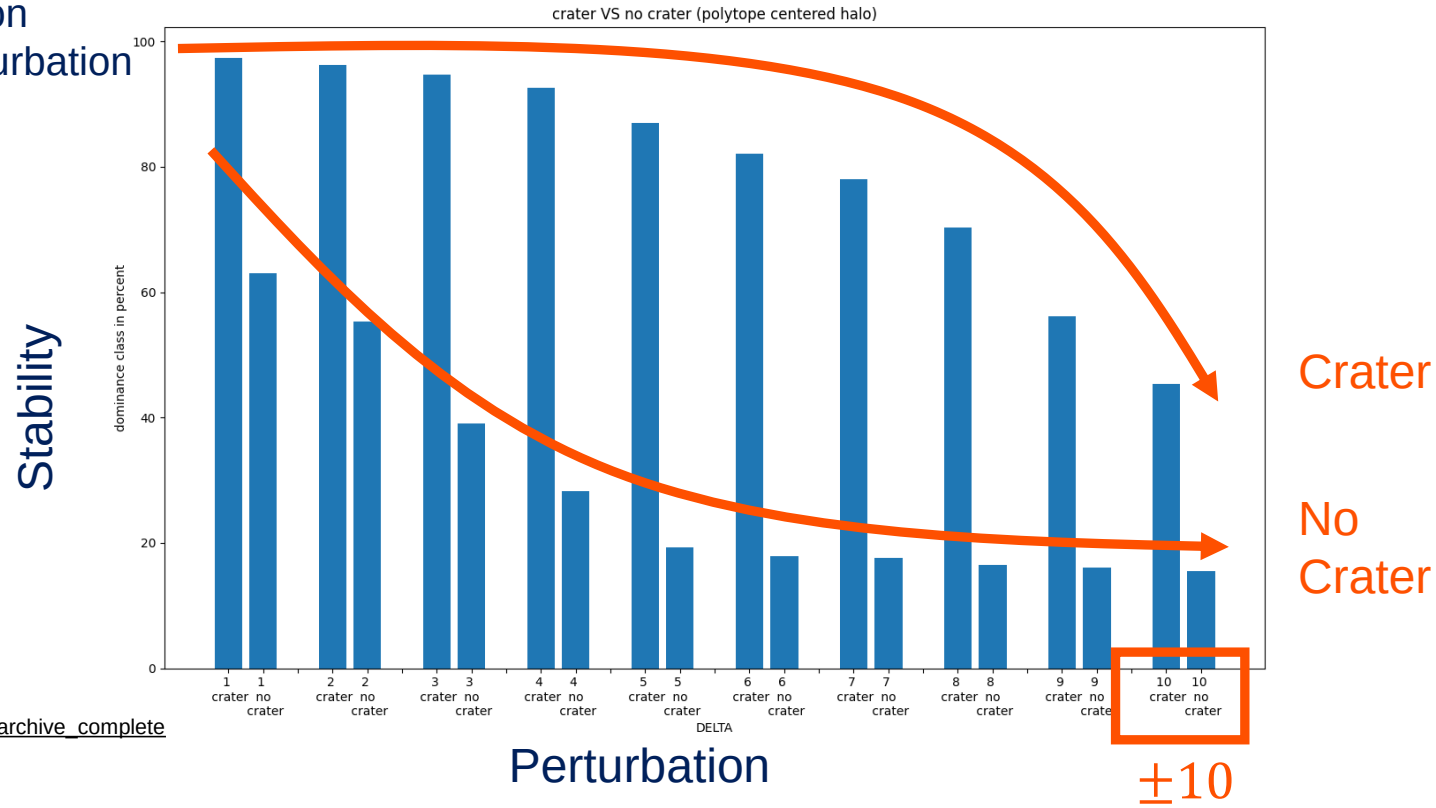
Requires a mathematical model of the perturbation
Validate on different levels of intensity of the perturbation



Flare

LM13

•https://www.esa.int/Science_Exploration/Space_Science/Rosetta/Rosetta_image_archive_complete



AIRBUS

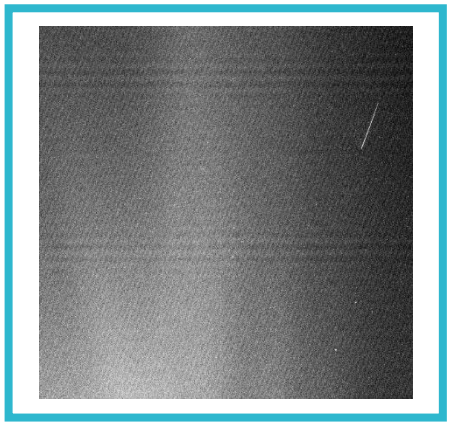
} _Task 3: Model evaluation – Robustness and Stability

Imagine classifier

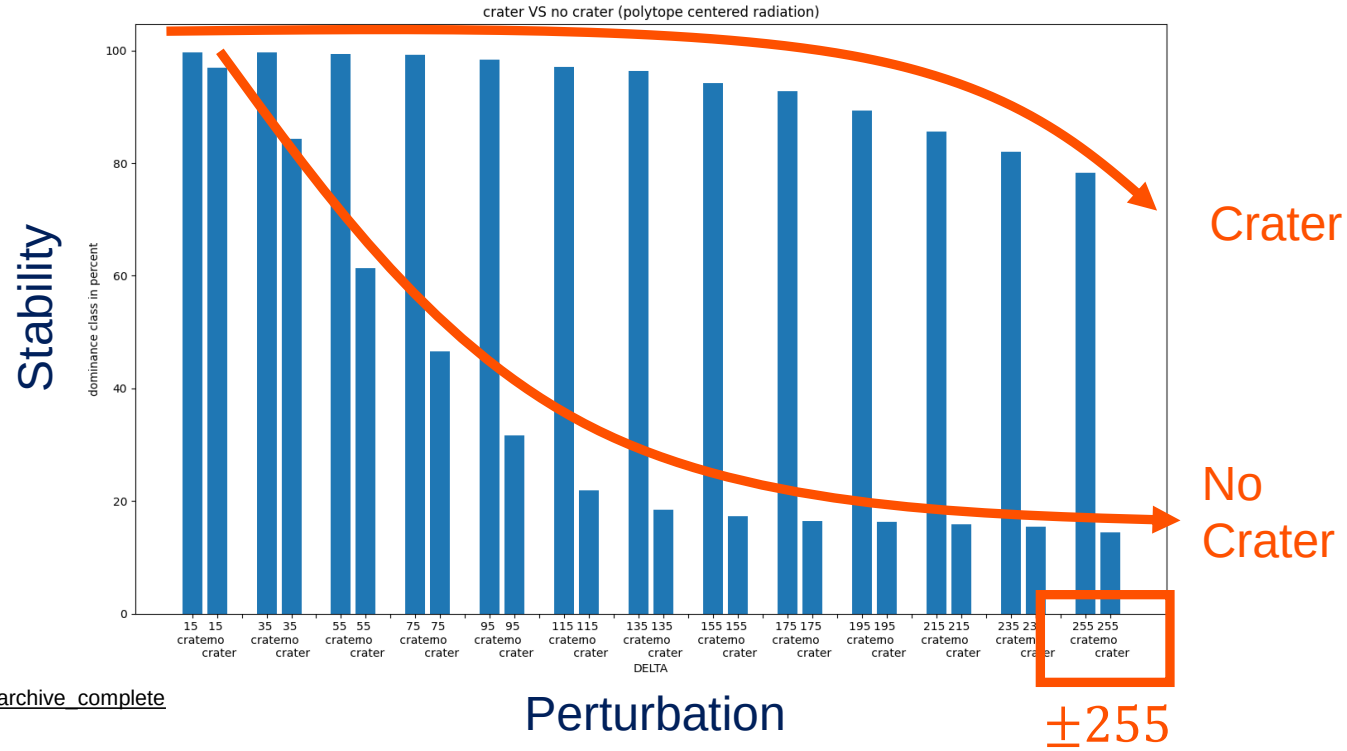
Stability against specific perturbation (specific to the ODD)

Requires a mathematical model of the perturbation

Validate on different levels of intensity of the perturbation



Radiation



LM13

•https://www.esa.int/Science_Exploration/Space_Science/Rosetta/Rosetta_image_archive_complete

} _Task 3: Model evaluation – Robustness and Stability

Imagine detector

Yolo (v3) architecture

LM12 tryout with formal methods
Feasibility is demonstrated
Computational time is still heavy (10+ minutes)

ODD

Extremely large
Can take up a very large amount of time to setup

LM12

•https://cocodataset.org/#overview

AIRBUS

} _Task 3: Model evaluation – Robustness and Stability

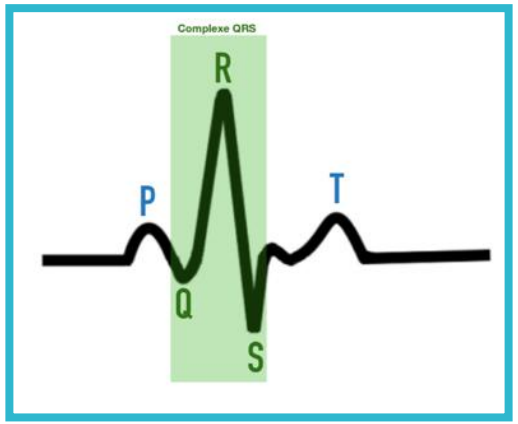
Time series classifier

Dataset

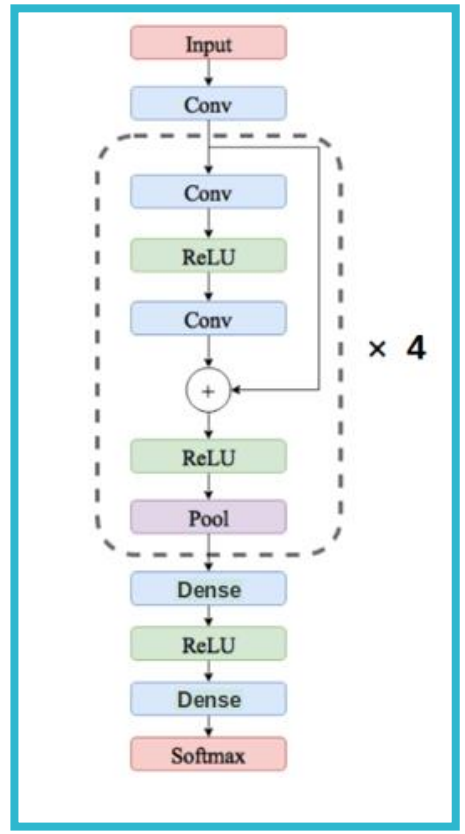
877K heart rythm
188 instants each
Class: 1 normal, 3 anormal, 1 unknown

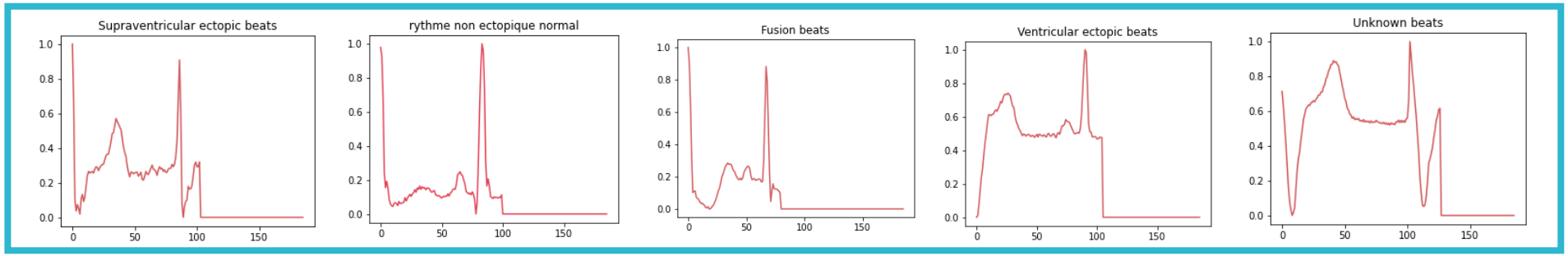
ODD

Can be defined by experts
But it is difficult to express abnormal cases



•https://arxiv.org/pdf/1805.00794.pdf





•https://physionet.org/content/mitdb/1.0.0/

AIRBUS

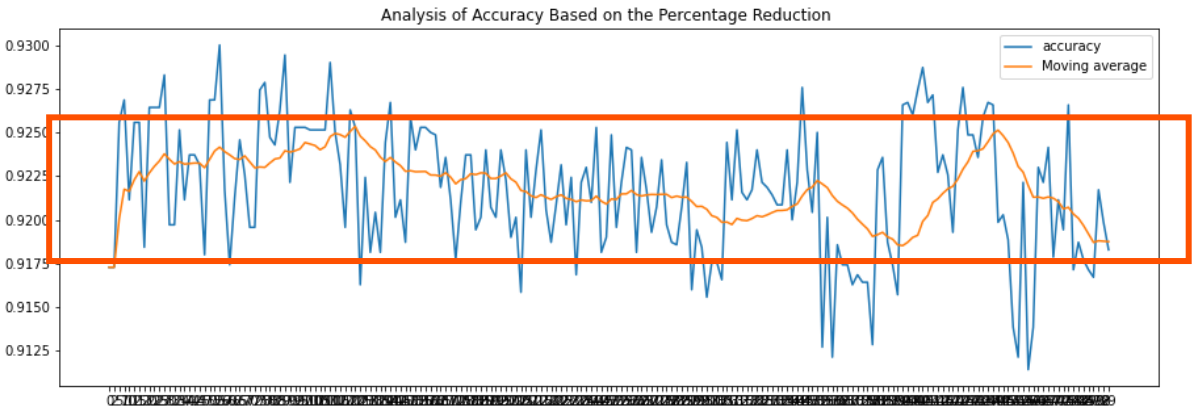
} _Task 3: Model evaluation – Robustness and Stability

Time series classifier

Training algorithm stability

Take one training point out
Retrain and revalidate accuracy

$\delta < \pm 1\%$



Better stability of the accuracy

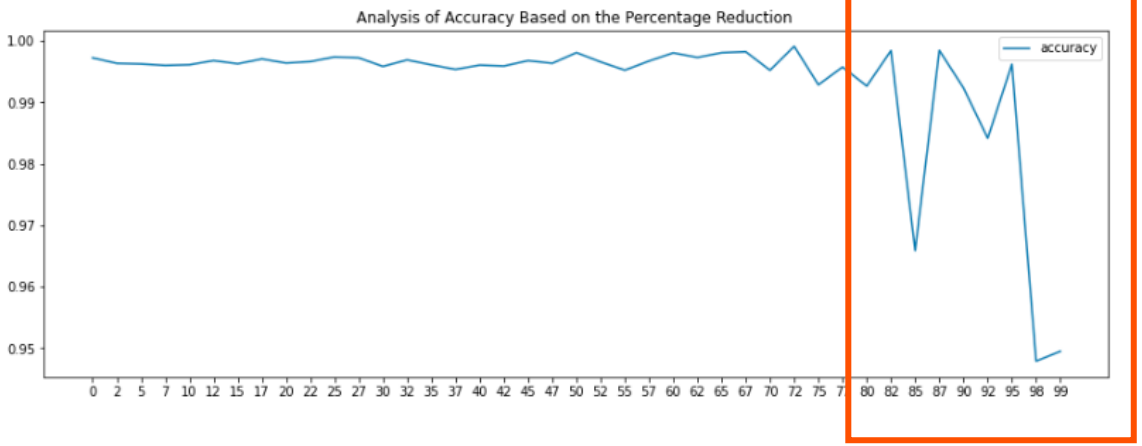
But is costly to do when have 500K+ data points

LM11

Training algorithm stability

Taking part of the dataset out
Retrain and revalidate accuracy

>85%



Most of the dataset does not contribute to the accuracy

Link with Task 1 (dataset) and Task 2 (generalization)

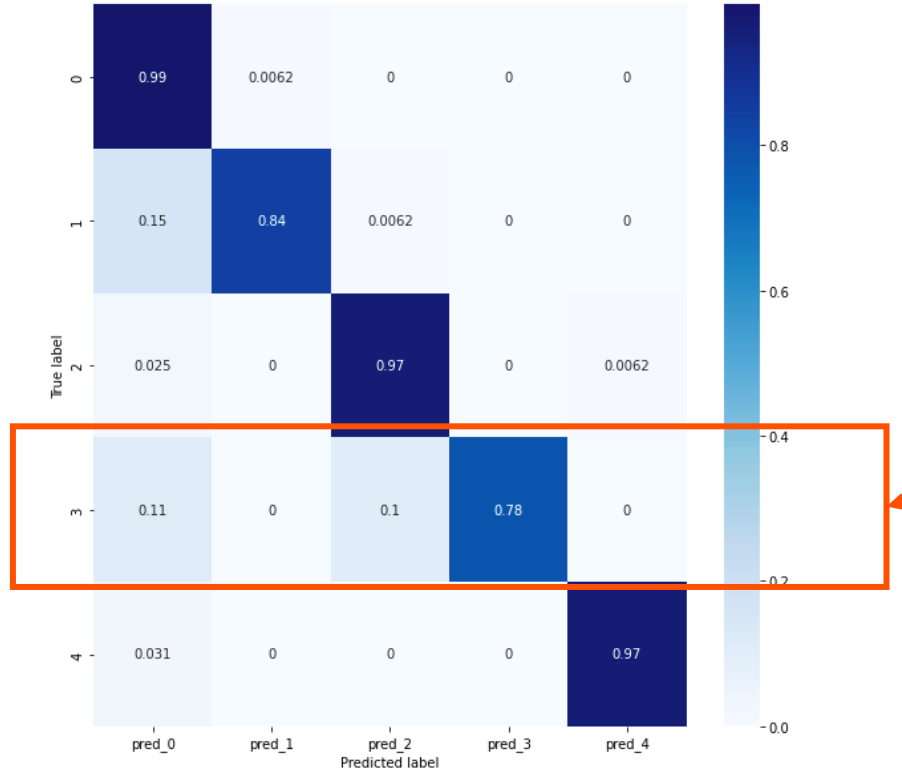
AIRBUS

} _Task 3: Model evaluation – Robustness and Stability – Time series classifier

Time series classifier

Statistical

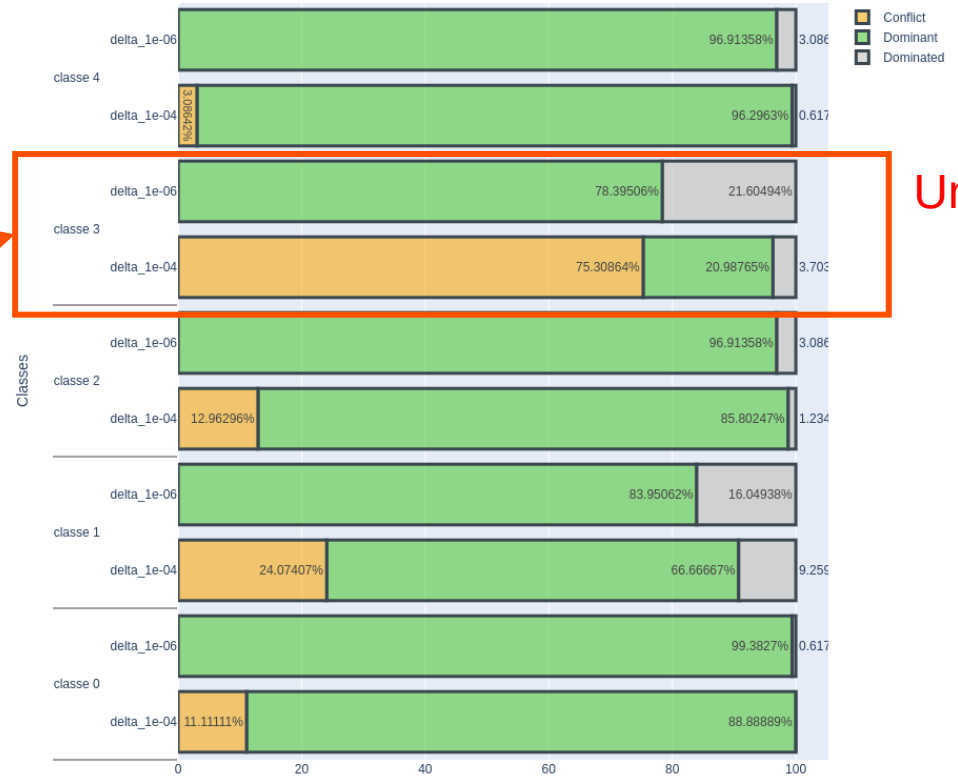
Confusion matrix
Accuracy 95+%



Slight decrease in accuracy

Formal

General stability



Unbalanced stability

LM12

AIRBUS

} _Task 3: Model evaluation – Robustness and Stability – Some good practices take away

Some good practices takeaways

Class separation -> Data -> Stability

- Detecting when and why classification change
- Ponder what can be done to better differentiate classes
- Adapt training dataset
- Measure again if stability has improved

ODD -> Perturbation -> Robustness

- Define clear specific perturbation using the ODD
- Measure how much the system can take
- Add more perturbated data (augmentation, simulation…)
- Measure again robustness has improved

Relevance (bias) -> Data -> Stability

- Detect incorrect relevance (manually or using segmentation)
- Identify pattern that can cause confusion (bias) (manually still)
- Adapt training dataset
- Measure again if stability has improved

Stability -> Wrong annotation -> Dataset

- Measure stability on each training data point
- Detect outlier in terms of maximum stability
- Control accuracy of the annotated data
- Correct if necessary

MLEAP project

Q&A

AIRBUS

/ Generic End to End AI Development Pipeline Proposal & Joint Conclusions -

AIRBUS

Application agnostic development pipeline:

Generic approach implementing the W-shaped process

What?

Pitfalls preventing AI projects from being released, along with their impact localization, and protocol to avoid them ;
A generic ML development approach implementing the W-shaped learning assurance ;

Why?

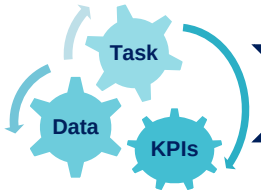
Bridge the gap between experimental objectives and industrial expectations (KPIs, business process, risk handling, …) ;
Help the development of AI-based systems respecting the means of compliance and certification guidelines of the EASA, through the W-shaped learning process.

How?

Identification, at each development stage, of the common issues and provide ways to overcome them ;
A complete pipeline, focused on the means of compliance objectives, to drive AI-based development.

Generic approach implementing the W-shaped process > > >

Common development mistakes: Identification

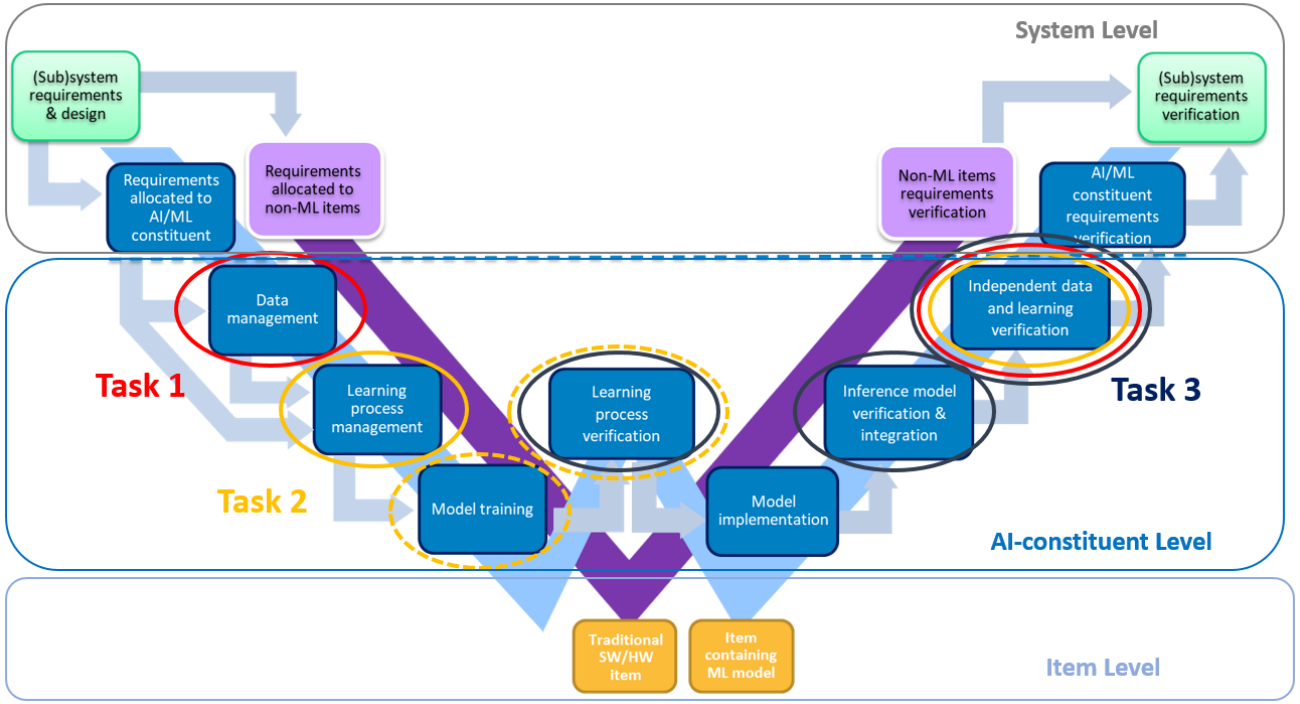


Modeling, Training, Evaluation/Testing, Adjusting, re-training, validating, preparing for release (embedding), implementation, behavior analysis, testing, validating.

Uncomplete ODD and system-level requirement definition

Inadequate data processing/representation w.r.t system-level requirements -> poor learning
Inappropriate datasets in training/testing -> lack of coverage, noisy, domain shift ...
Unhandled outliers, non-standardized data -> bad robustness, unstable model, …

Overfitting/Underfitting -> lack of performance
Mis-use/understanding of generalization bounds -> misleading for model design and evaluation
Inappropriate training objective, inappropriate evaluation measures -> gap between target objective and model performances
Inappropriate model capacity vs task complexity, non adapted optimization & regularization -> weak learner & bad performances

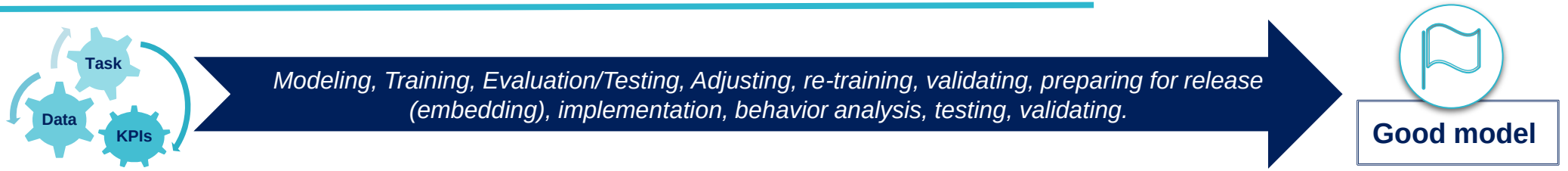


Rely only on testing -> not enough to state the model performances
Insufficient stability against specific noise -> non robust model
Learned bias -> weak robustness, stability & generalization
Incorrect annotations in training data -> bad performances, incorrect predictions

AIRBUS

Generic approach implementing the W-shaped process > > >

Common development mistakes: Recommendations & Checklist



(1) Drive the data management

The ODD as a centerpiece of data quality: completeness & representativeness

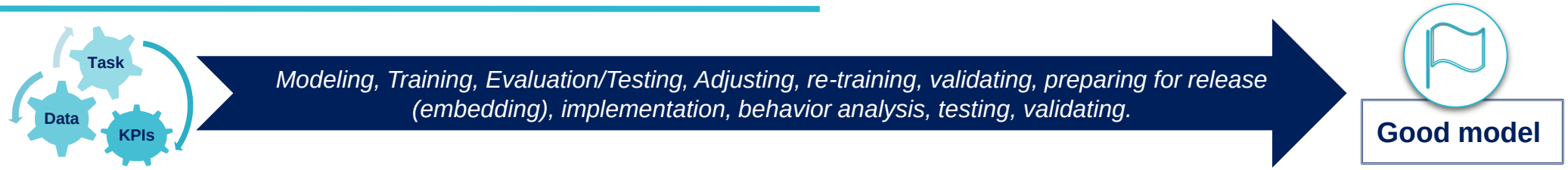
Sample of real world, but not the whole of it;
Include factors defining its limits, edge cases, and interactions;
Data requirements as meta-data & driver of the data collection & preparation;
Target performances specification for specific cases:
Estimation of volume needed and specific characteristics

The model as a necessary feedback source

Models behavior during training and evaluation results -> data patterns that are more/less complicated to be learned
Help finding a trade-off between completeness & representativeness

Generic approach implementing the W-shaped process > > >

Common development mistakes: Recommendations & Checklist



(2) Drive the model development and training

Rely on ODD analysis outcomes

- Data type (e.g img, txt…) and nature (e.g evolutive, static...) drive the ML design ;
- Task complexity, data volume and availability -> accurate model complexity ;
- Performances influencing elements (e.g appropriate data setting, weights, representation …) ;
- Target domain system-level requirements (model max volume, available resources, access, tolerable error-margins, uncertainty handling …)

Focus on target performance objectives from the industrial perspective

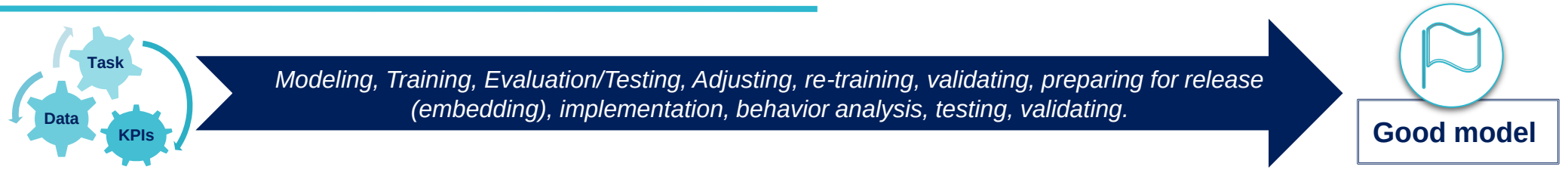
- Generalization assessment (bounds) & perf. evaluation (metrics) **vs** KPIs ;
- Critical systems requirements to be included -> no impact on safety ;
- Training objectives, eval. metrics selection/definition -> adaptations needed and acceptance criterion reviewed ;

Anticipate ways to enhance the performances

- Performance influencing elements handling & have a good error analysis to identify weaknesses of the model ;
- Good learner: regularization, optimization, learning objective adaptation … ;
- Architecture, settings, and parameters adaptation

Generic approach implementing the W-shaped process > > >

Common development mistakes: Recommendations & Checklist



(3) Reinforce the model robustness and stability

Using the class separation to improve stability

Maximum stability space per class (e.g formal methods to check the closest boundaries and distance of each data point -> in training, ensure sufficient distance inter-classes

Minimal level of perturbation required to change the classification decision -> monitor confusions

Using ODD perturbations to reinforce robustness

Edge-cases as borderline cases with perturbations;

Leverage existing ones and/or generate others using perturbation methods to reinforce stability;

Using relevance properties to avoid some bias

Identify learning bias of the model;

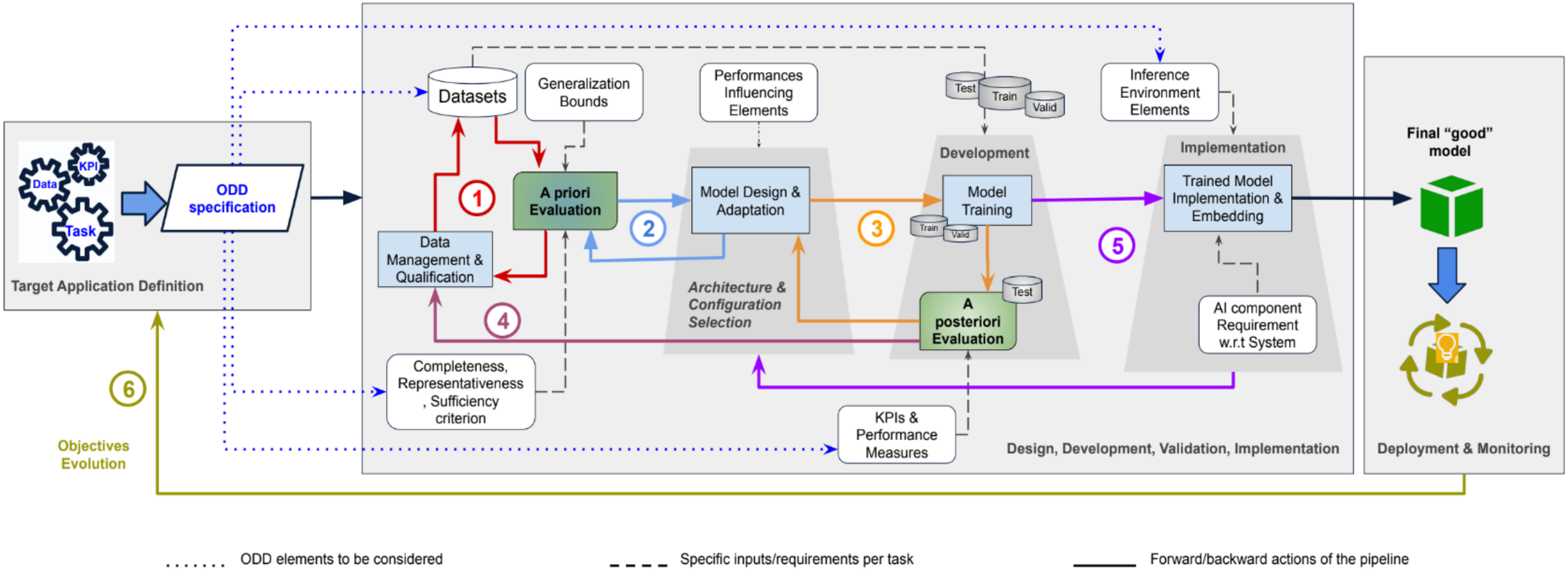
Semi-automatic relevance analysis -> help model training process (e.g fuzzy relevance means underfitting)

Using stability to crosscheck data sets

Lack of stability at some data point could be due to poor data annotation and/or representation -> max-stability space computation & identification of poor annotations

Generic approach implementing the W-shaped process > > >

Use-case agnostic development pipeline



AIRBUS

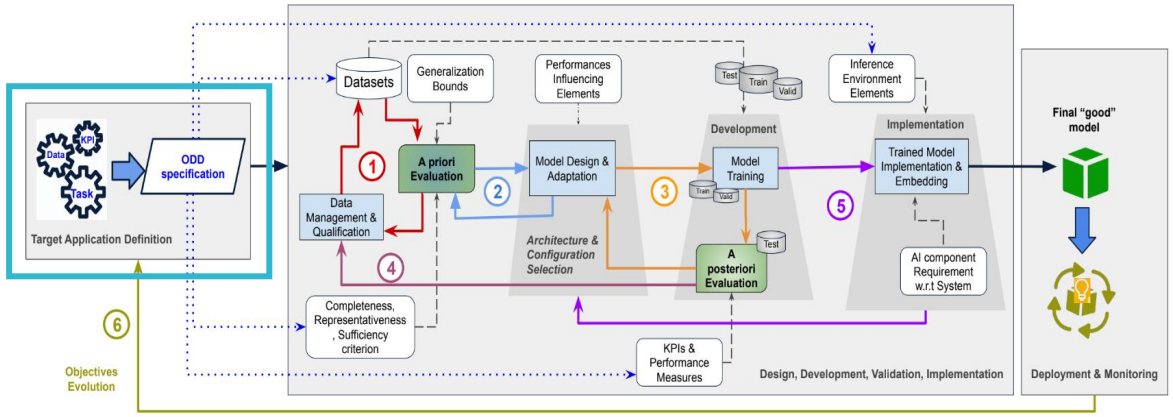
Generic approach implementing the W-shaped process > > >

Target application definition

Understanding the objectives & ODD specification

Datasets. input/output spaces and quality criterion (completeness, representativeness, and sufficiency)

Performances Influencing Elements. all known characteristics of the target environment that are more likely to influence the model. These are included directly on the model design.



KPIs & Performance Measures. expectations from the industrial perspective, including:

target performances, the ML-level requirements derived from the system-level requirements, including safety and certification requirements,

the acceptability criteria and conditions (e.g tolerable error margins and unacceptable errors)

Inference Environment Elements. Target application features and environment impacting results

elements related to the system-level requirements and operating conditions, having an impact on the ML-component

elements related to changing conditions that cannot be controlled at system or ML-component level (e.g. weather conditions and light intensity, which have an impact on video applications).

AIRBUS

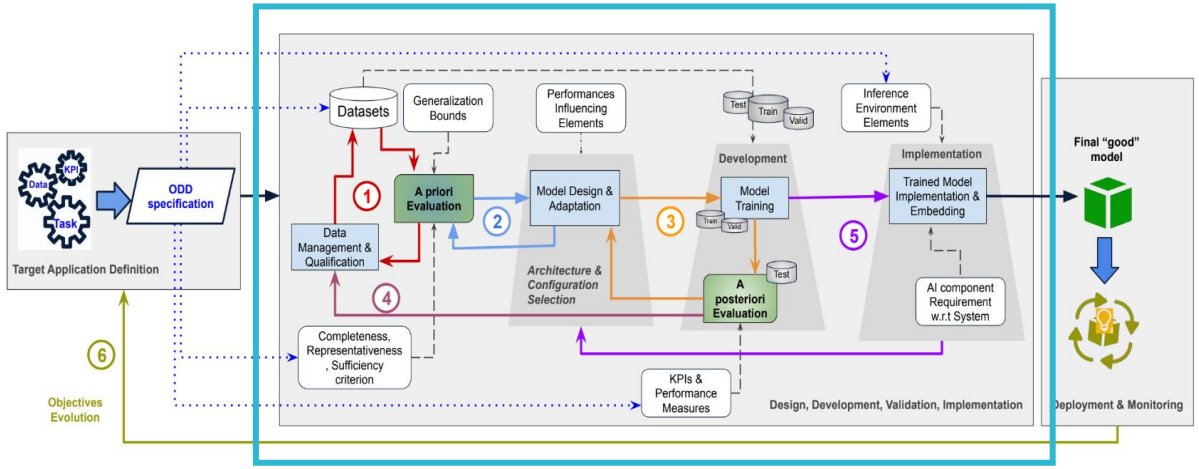
Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

Two-folds Evaluation

A priori evaluation. Before ML/DL design. Performance objectives assessment, in addition to data management:

- Development and design pitfalls and common issues investigation;
- Data quality and volume criteria requirements,
- Completeness and representativeness;
- Generalization bounds selection and computation;



A posteriori evaluation. After ML/DL training. Performances evaluation and verification:

- Focus on generalizability, robustness and performance stability.
- Integrates KPIs and selected performance measures
- Test dataset selected w.r.t several data management criteria (ODD conformity and training set representativeness)
- Evaluation metrics w.r.t. the target task and domain-specific (business) acceptance criteria
- Hypothesis on the performance requirements of the ML/DL model verification w.r.t system-level requirements

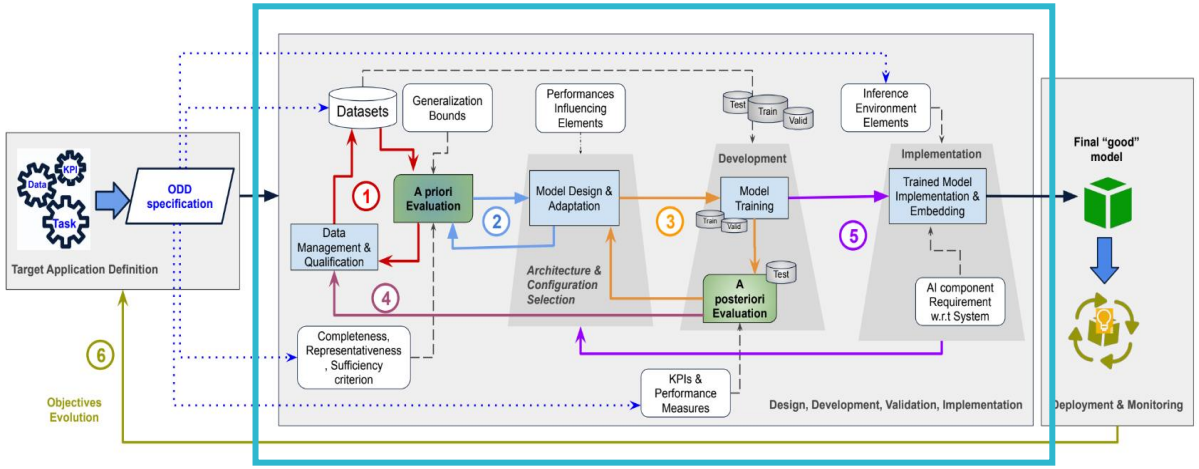
AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

(1) Data qualification and preparation

(a) Identify important criteria for the data quality (representativeness and Completeness), samples distribution analysis, corner/edge cases, outliers, impact on the training;
(b) ODD analysis: identify the requirements, in terms of data volume needed, specific cases handling on the data (specific measures for some outliers);
(c) If data is not collected yet, based on (a) and (b), data collection & preparation.



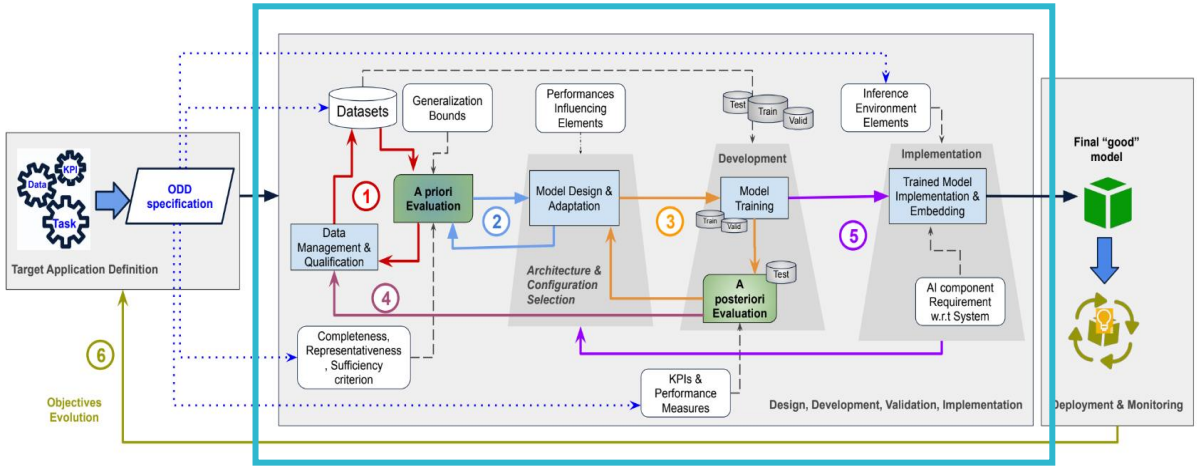
AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

(2) Model Design & Adaptation

(a) Architecture definition, approach that meets data and target application specificities;
(b) Models that is compliant with the constraints at the system-level and the target application (e.g real-time application, be embedded in a resources limited system …), data-related constraints (e.g. available data volume, inputs size and type);
(a) Use insights from the ODD analysis (performances influencing elements, system criteria …), data availability and features, estimated generalization (bounds)



AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

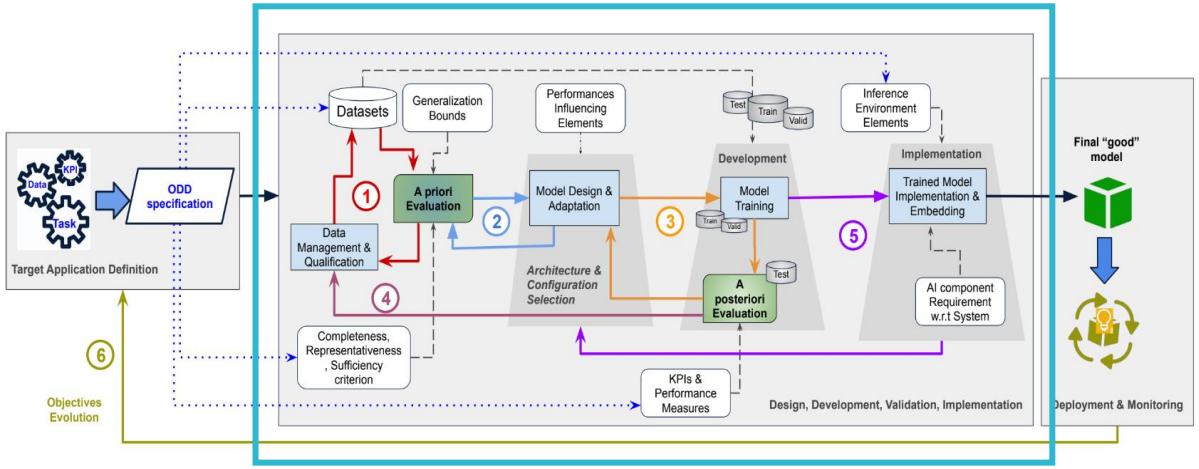
(3) Model development, training, and the a-posteriori evaluation

(a) using the qualified data sets in (1), and adapted training objective;

(b) benchmark including industrial KPIs, evaluation measures, and acceptability criteria,

(c) A posteriori evaluation of the trained model to ensure that it meets the industrial objectives (generalization, robustness, and stability)

(d) a backward action can be considered to re-work the model design and configuration if acceptance-criteria not verified



AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

(4) An iterative process for improvement and adaptation

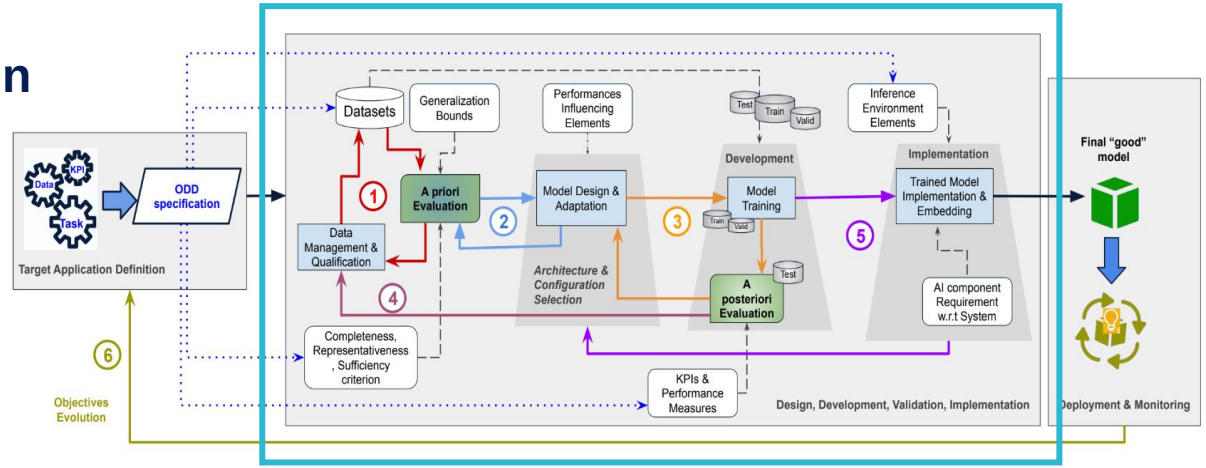
(a) both the training and test data as well as the construction of the model

(b) make each stage as secure as possible, with the necessary verifications to avoid backtracking;

(c) After training, if the model does not meet specified performance requirements, perform analysis and improvement actions:

-> identifying the main causes of the lack of performance,

-> poor training, non-adapted architecture, insufficient data or poor specifications.



Possible options:

- ❑ Combine assessment methods working directly on data (e.g. PCA) with methods using the model as feedback (e.g. Cleanlab);
- ❑ Observe the interaction between the data and the model;
- ❑ Ensure the reproducibility of the results of a trained model: handle the randomness of some ML/DL models (e.g NNs) and anticipate accurate configurations during the design (e.g fix the seeds parameter for random initialization).

AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

(5) The moment of truth is implementation

Is the expected objective met while interacting with target domain?

(a) Inference Environment Elements are consumed by the implemented model

(b) Verify performances in the target environment & AI component requirement w.r.t System requirements

(c) The model is either:

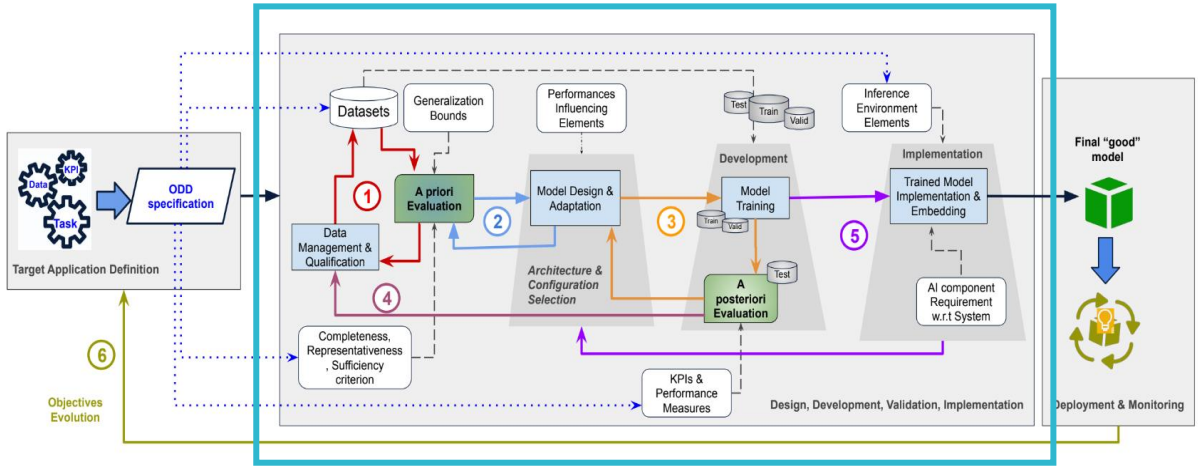
i. validated and go to the *Deployment & Monitoring phase*

ii. Rejected and a backward action is needed,

(d) if validation fails: -> *new model*

i. Adaptation of the model design-configuration, including influencing environment components

ii. Performances Influencing Elements are already included before training, rework their impact



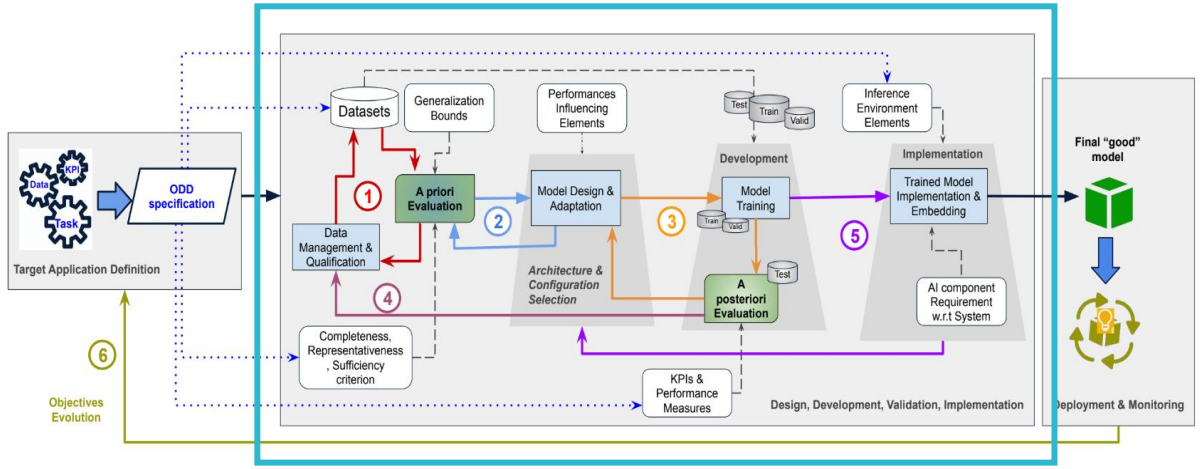
AIRBUS

Generic approach implementing the W-shaped process > > >

Design, development, validation, and implementation

(5) : Backtracking – Be Aware of:

- This impacts the previous validated choices (model configuration, generalization bounds, evaluation metrics) since target performances are not met;
- A new family of models will be selected with adapted set-up to take into account particularities of the implementation environment;
- Potential biases on data will be detected and feedback to the data management and preparation will be provided to enhance the quality of the datasets.



AIRBUS

Generic approach implementing the W-shaped process > > >

Deployment and monitoring

(6) System's objectives evolution after model deployment

System evolution, the monitoring could help integrating the new objectives of the system, without/without a new model
Changes of system-level objectives, means that the model may be inadequate to meet the new requirements

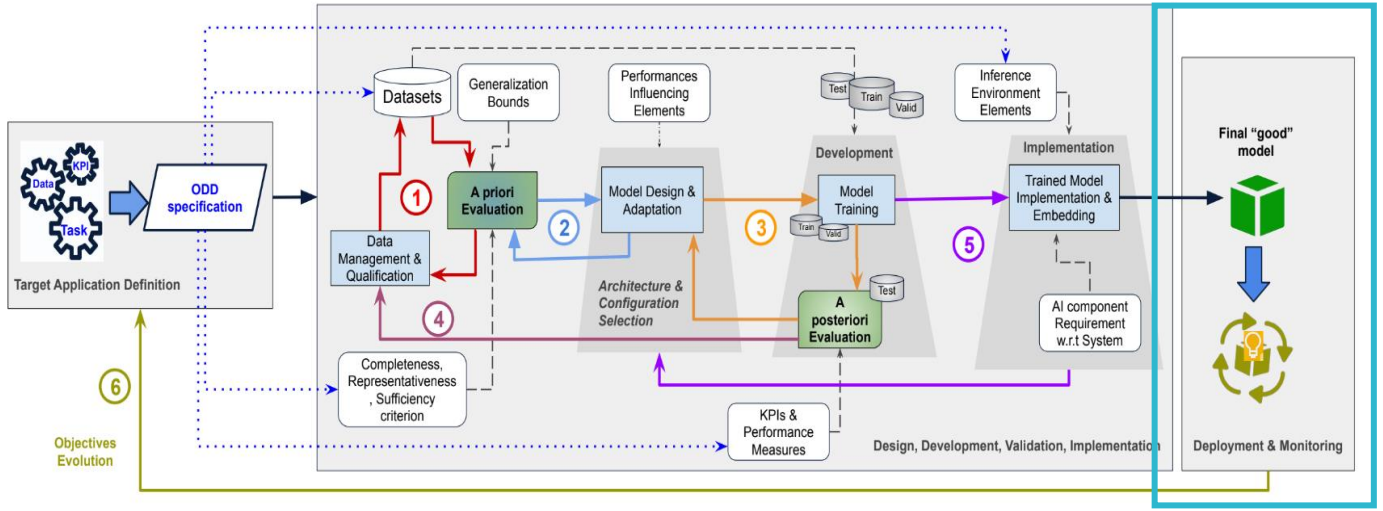
(a) definition of the ML component objectives to be reconsidered

(b) Major activities:

i. The definition of new objectives, and re-execution of the entire development pipeline;

ii. Re-using (retraining or fine-tuning) of the initially validated good model;

iii. Development of a new model using an architecture that is more adapted to the new objectives.



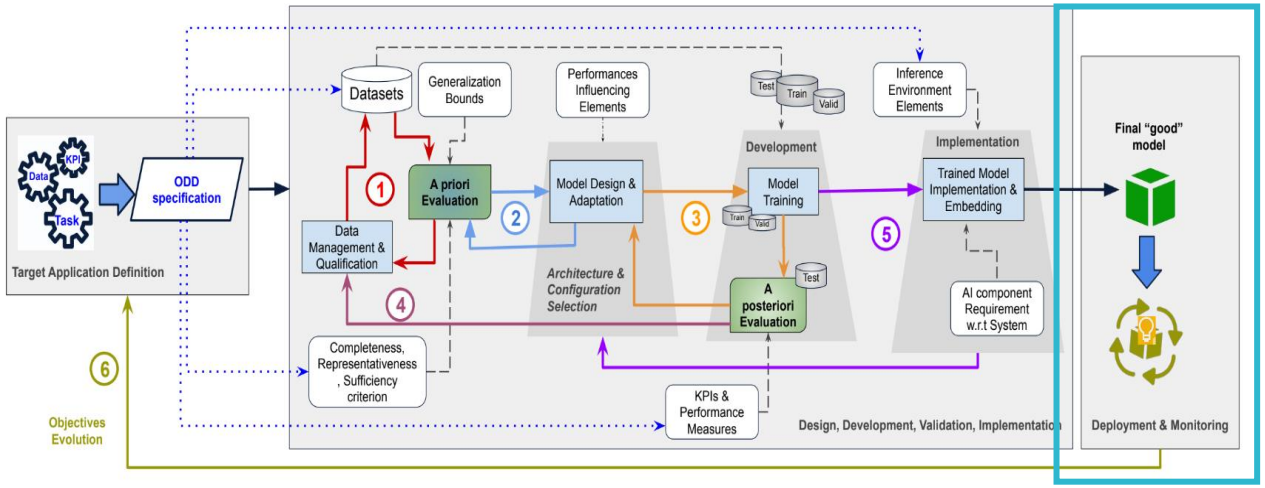
AIRBUS

Generic approach implementing the W-shaped process > > >

Deployment and monitoring

(6): Backtracking – Be Aware of:

It aims to include new objectives due to system-level evolution
In the case of model retraining, make sure to not reuse the same training data distributions
The already selected generalization bounds and evaluation measures will be revised
Take into account new requirements and adapt evaluation (KPIs, measures and acceptance criteria) accordingly
If same targeted performances for the new objectives (e.g ODD amplification) a new data qualification is required, including the verification of completeness and representativeness w.r.t the new task to be learned
The targeted performances may not be the same, different learning objectives, evaluation measures benchmarking to reconsider



AIRBUS

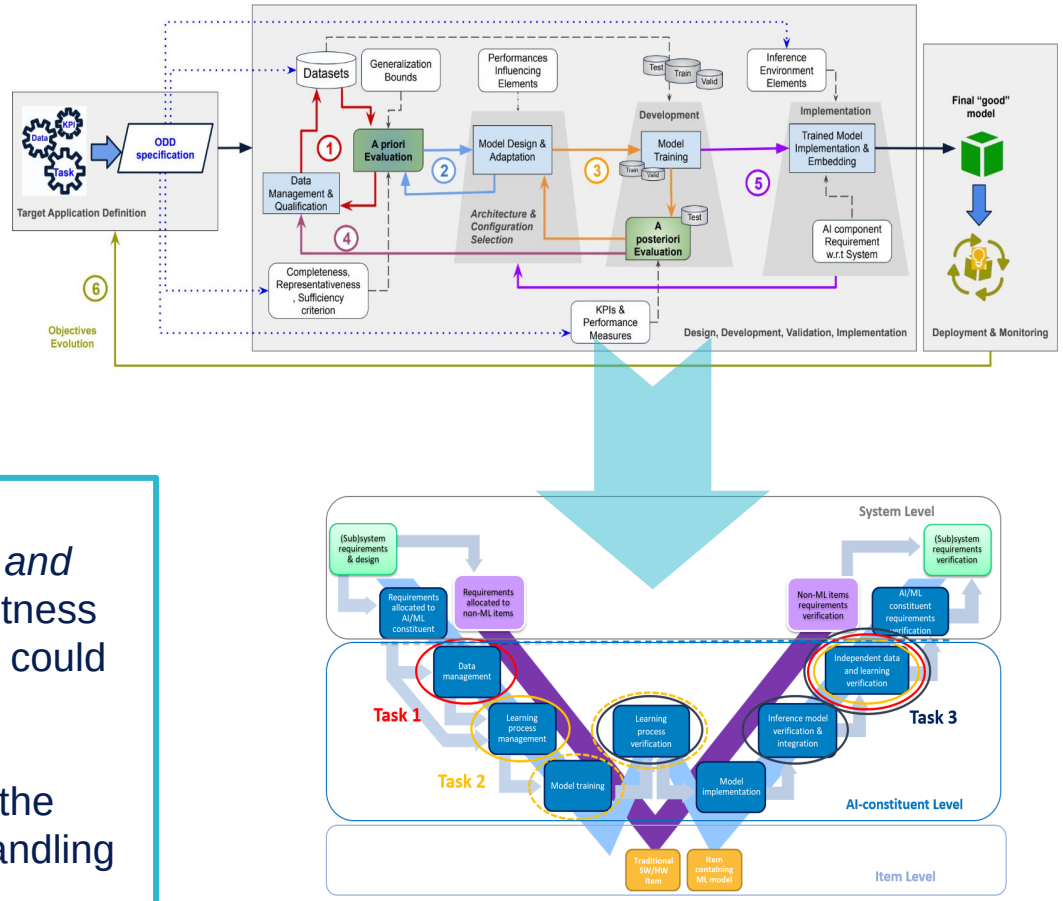
Generic approach implementing the W-shaped process > > >

Main Conclusions

In ML/DL development, the key point is to ensure a good data quality which highly impacts the generalization, robustness, and performances stability, after training and evaluation, which are decisive for model implementation ;

Depending on the use-case characteristics (*data, model, task complexity, and system-level requirement*), methods for generalization assessment, robustness evaluation, and data qualification could be used differently, interpretations could vary, and adaptations are needed ;

At different stages of the development, several common practices impact the model performances: data-related hypothesis, KPIs understanding and handling in train/dev, accurate evaluation setup, target environment analysis … ;



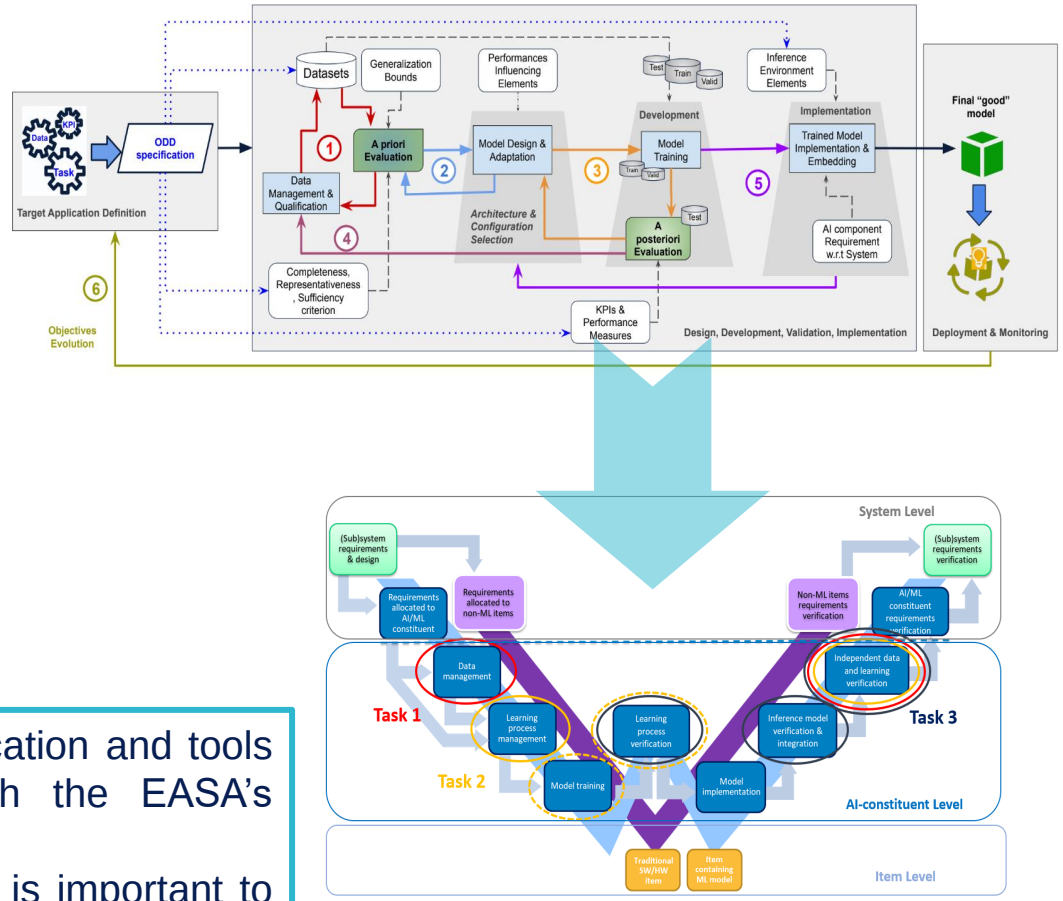
AIRBUS

Generic approach implementing the W-shaped process > > >

Main Conclusions

For a secure enough ML/DL development, stay close as possible to the target domain and system-level to have aligned objectives:

- Well understanding of the ODD and target application definition (data & characteristics) ;
- Consider system-level requirements (safety-related, results acceptance, monitoring definition) as well as item-level requirements (feasibility, available resources …) ;
- Before moving forward, make sure that there is no/less risk to make the model fail: test, evaluate, analyze (performances and error analysis and distribution on the test sets) ;



The generic pipeline implements the W-shaped process with steps specification and tools recommendation to ensure ML/DL performances and compliance with the EASA's guidelines and certification ;
Beyond the ML/DL development pipeline, after deployment, the monitoring is important to determine the model compliance with the system, and how system-level evolution impacts the AI-constituent-level evolution.

AIRBUS

MLEAP project

Q&A

AIRBUS

/ Key takeaways & MLEAP Project next steps -

AIRBUS

MLEAP – takeaways for each task

ODD is the centerpiece of the Learning Assurance concept
orienting the quality of the datasets and paving the way to model performance, stability, robustness and generalisability.

Task 1 (dataset completeness and representativeness)

- **Structuring the set of proposed methods into guidance for the applicants**
  - *Confirm the suitability of the methods for use cases depending on dimensionality*
  - *Segregate methods based on their goals (demonstration of lack or good completeness and/or representativeness)*
  - *Guide whether the method applies to a priori or a posteriori evaluation, and for which loop of the generic pipeline.*

Task 2 (ML models generalisation)

- **Ensuring generalisation remains a challenge**
  - *Set of methods experimented on "toy use cases" do not provides satisfactory generalisation bounds*
  - *Other methods should be further investigated*
  - *Generalisation is a key enabler for higher criticality levels AI-based systems.*
  - *Generalisation is a very active field of research to be monitored in the mid-term*

Task 3 (ML models stability and robustness)

- **Ensuring stability and robustness of the trained model**
  - *Statistical methods are the most straightforward way to analyse properties, however linked with preparation effort and limitations in high dimensionality.*
  - *Formal methods are confirmed to be usable for ML models stability, still subject to limitations in terms of scalability.*
  - *Empirical methods rely on expert judgment to make their evaluation, therefore remain case by case.*

MLEAP – Generic pipeline takeaways

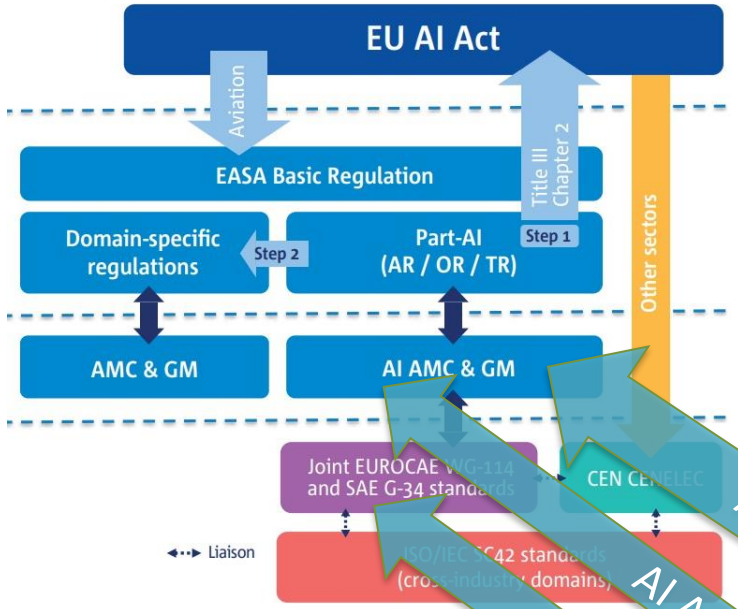
The generic pipeline provides a framework to organise the main verification activities for a machine learning model

- *It is introducing the notion of a-priori and a-posteriori verifications*
- *It covers a large portion of the necessary verification steps and properties from the Learning Assurance W-shaped process*

The generic pipeline is defined in the context of the three tasks of the MLEAP project

- *Its extension beyond the focus on generalisation property can be further refined in next steps of MLEAP*
- *Its extension of applicability to the full set of objectives of the learning assurance is to be confirmed for the overall scope of verification per the Learning Assurance W-shaped process.*
- *Its integration into industrial process frameworks is to be worked out (e.g. how to integrate the pipeline into an MLOps framework?)*

Key takeaways



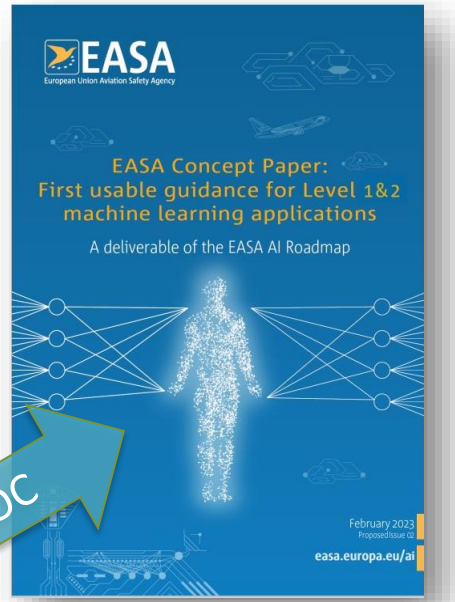
Rulemaking

AI GM

AI AMC

Need for ER/AIR?

Anticipated MOC



Final Issue 02

MLEAP project

Q&A

AIRBUS

WHAT's next for MLEAP?

PROJECT:

May 2024: End of the project

MLEAP Final report will be published

EVENTS:

May 2024: 29th: EASA AI DAYS – MLEAP Stakeholders day final conference!

STAY INFORMED AND FOLLOW US!

Websites

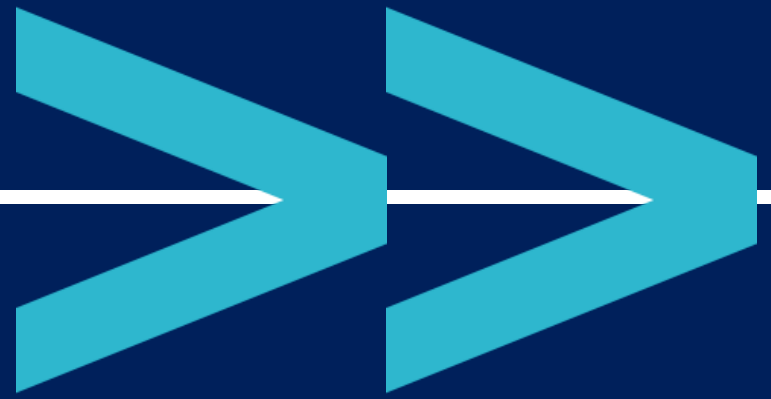
https://www.lne.fr/fr

https://www.protect.airbus.com/

https://numalis.com/

https://www.easa.europa.eu/en/research-projects/machine-learning-application-approval

{Thank you}



Afterwork:!

Let’s keep the party going!

AIRBUS