Welcome to the EASA AI Days High-Level Conference !

3rd July 2024

EASA

Welcome to EASA AI Day 2

Guillaume Soudain,

EASA Artificial Intelligence Programme Manager

Disclaimer

MLEAP project is funded by the European Union under the Horizon Europe Programme.

Views and opinions expressed are however 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.

This deliverable has been carried out for EASA by an external organisation and expresses the opinion of the organisation undertaking this deliverable. It is provided for information purposes. Consequently, it should not be relied upon as a statement, as any form of warranty, representation, undertaking, contractual, or other commitment binding in law upon the EASA.

_MLEAP STAKEHOLDERS DAY

CONFERENCE 4#

AIRBUS

- **Introduction of the MLEAP Project and of the Partners**
- **Presentation of the use cases**

Q&A session

- **Presentation of the outcome and recommendations of Task 1** *LNE*

Q&A session

- **Presentation of the outcome and recommendations of Task 2** *Airbus Protect*

Q&A session

- **Presentation of the outcome and recommendations of Task 3** *NUMALIS*

Q&A session

- **General conclusions and recommendations from MLEAP consortium**

Q&A session

- **EASA perspectives on MLEAP takeaways**

Q&A session

- **Conclusions of the EASA AI Days 2024**

} Who we are > > > MLEAP TEAM

Consortium members :

EASA
European Union Aviation Safety Agency

Willy Sigl,
Xavier Henriquel,
Guillaume Soudain,
François Triboulet

Michel Kaczmarek,
Thiziri Belkacem,
Jean-Baptiste Rouffet,
Jeremy Bascans,
Matthieu Rochambeau

AIRBUS
PROTECT

LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS
LNE

Olivier Galibert,
Swen Ribeiro,
Agnes Delaborde,
Sabrina Lecadre

Arnault Ioualalen,
Noémie Rodriguez

numalis

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)

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, 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 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

/ Introduction of the MLEAP Project

AIRBUS

MLEAP project
Introduction

Guillaume Soudain

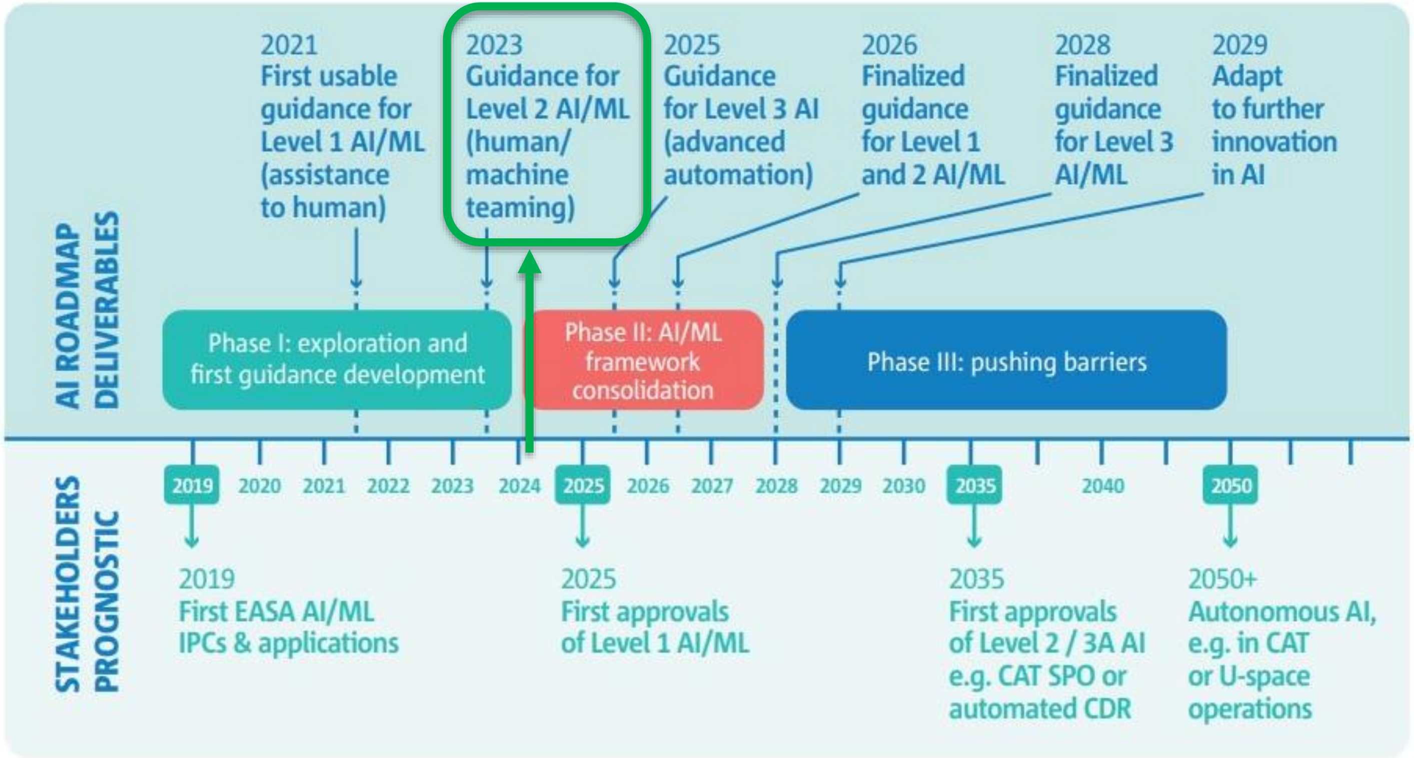
EASA AI Programme Manager

Xavier Henriquel

EASA MLEAP Tech lead

Timeline of EASA AI Roadmap 2.0

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



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

Use cases: a collaborative approach with Stakeholders

EASA
European Union Aviation Safety Agency

AI/ML Guidance

Use Cases

(IPCs, MoUs, first applications)

EUROCONTROL
DLR
EUROPEAN DEFENCE AGENCY
DAEDALEAN APPLIED INTELLIGENCE
gmv INNOVATING SOLUTIONS
Collins Aerospace
AIRBUS
THALES
DEEL DEpendable & Explainable Learning
BOEING
CAE

SESAR 3 projects

Horizon Europe MLEAP (Machine Learning Application Approval)

AIRBUS
PROTECT
LNE
numalis

CEN CENELEC JTC21

ISO/IEC SC42

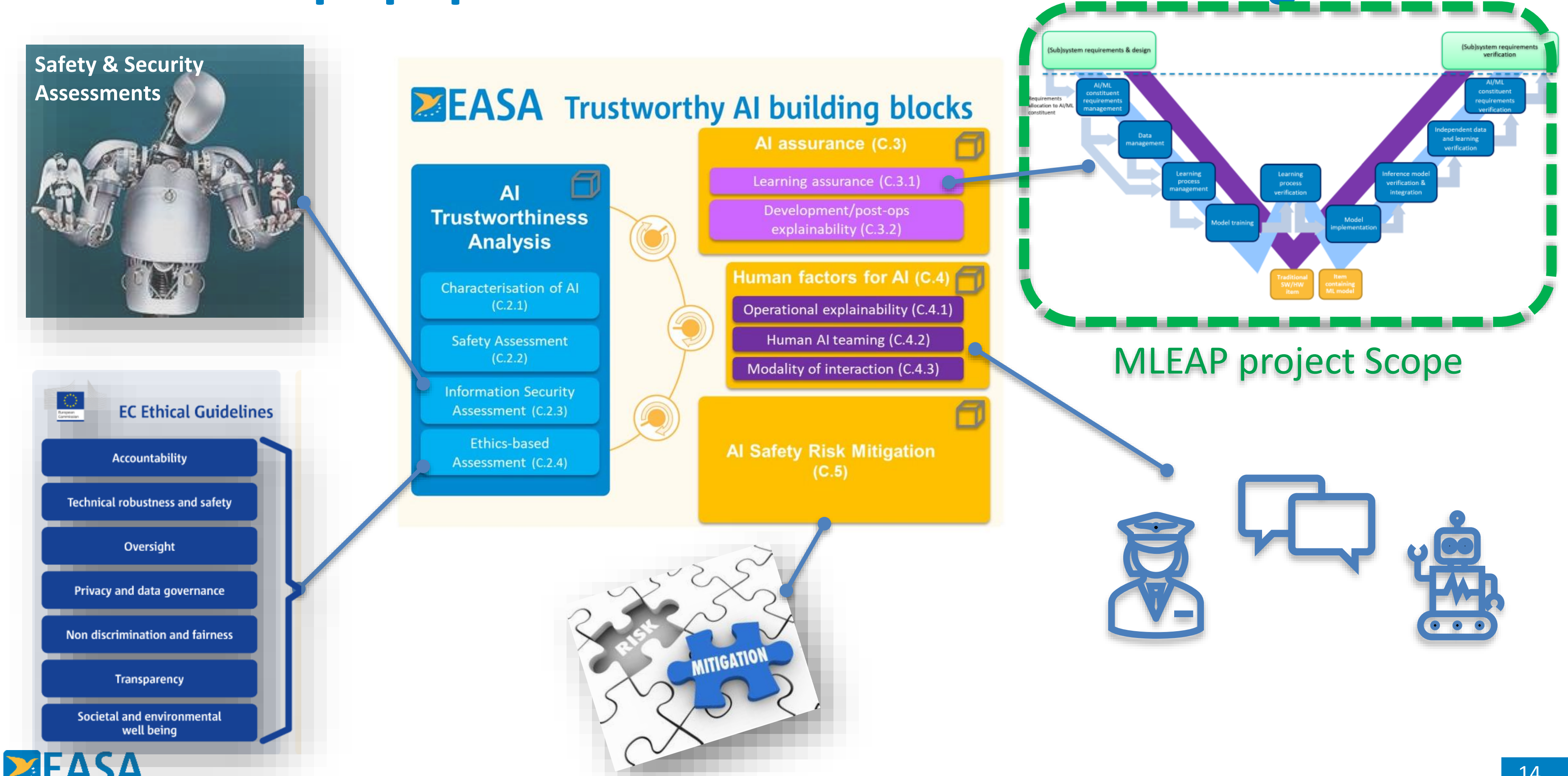
Working group

WG-114/G-34

EUROCAE
SAE INTERNATIONAL

IPC = Innovation Partnership Contract
MoU = Memorandum of Understanding

EASA Concept paper - AI trustworthiness building-blocks



Machine Learning Application Approval (MLEAP) project

Objectives

Streamline certification and approval processes by **identifying concrete means of compliance** with key objectives of **learning assurance objectives block of EASA Concept paper (CP).**

Research consortium

LNE - Airbus Protect - Numalis

Budget & timeline

1.475 m€ funded by EU
Horizon Europe program
May 2022 - May 2024

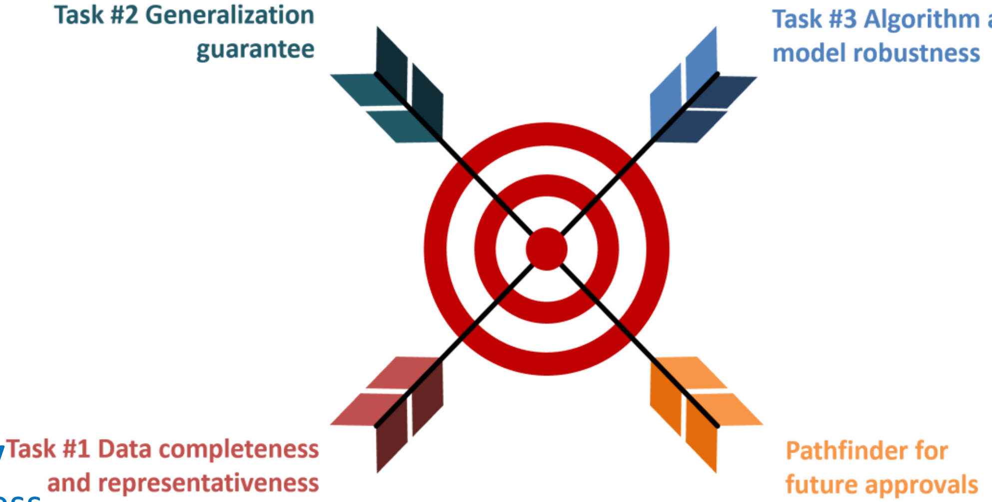
MLEAP Task 1 - Data completeness and representativeness

- **Overcoming Data Quality Obstacles**
  Ensuring data quality is complex and costly.

- **Addressing Completeness and Representativeness**
  The issues of data completeness and representativeness often go unaddressed. There is a notable lack of tools specifically designed for these tasks.

- **Balancing Representativeness and Diversity**
  Striking a balance between representativeness and diversity in data is a delicate task.

- **Main CP objectives:**
  DA-03, DA-04 and DM-07



MLEAP Task 2 - Generalization guarantee

- **Ability of AI/ML to scale up to unseen data** during training is one of main concern with safety critical applications

- Objective of Task 2 is to establish **protocols and strategies that improve the generalization capabilities** of deployed models. This involves:
  - taking into account data quality and volume.
  - obtaining quantifiable guarantees.

- **Main CP objectives:**
  LM-04, LM-07, LM-09, LM-10 and LM-14

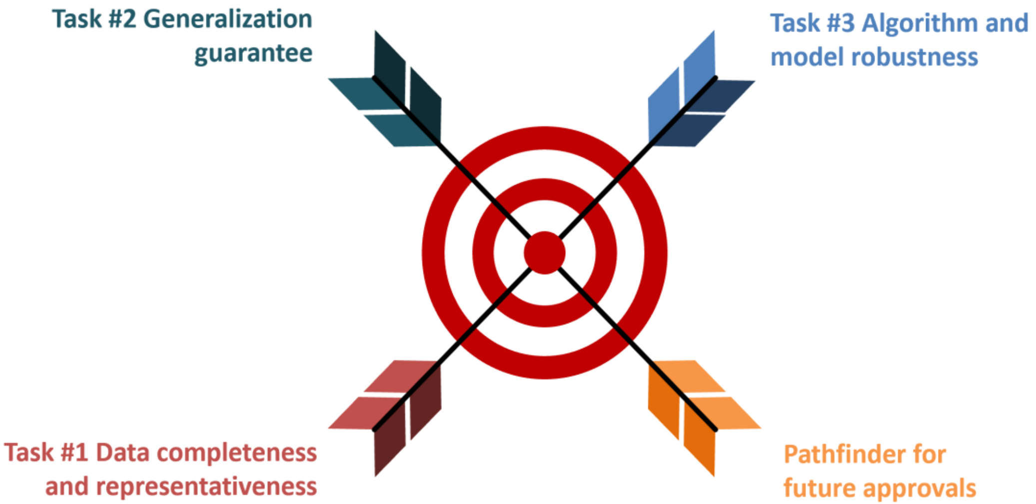
MLEAP Task 3 – Algorithm and model robustness

- **Aligning existing concept** in EASA Concept Paper, CoDANN I & II IPCs and ISO/IEC 24029
- **Variety of approaches available:** Empirical, statistical and formal methods
- **Continuation of the effort of evaluating formal methods benefits** (e.g. EASA-Collins Aerospace ForMuLA IPC)
- **Main CP objectives:** LM-02, LM-11, LM-12, LM-13

MLEAP - Pathfinder for future approvals

- **Practical aviation** AI/ML use cases
  - Provision for EASA access to detailed models & datasets
  - Utilization of public data or examples whenever feasible, enabling benchmarking with 3rd parties.

- **Knowledge sharing** and stakeholder guidance
  - Participation in public events
  - Project page with latest results
  - Public reports



/ MLEAP Objectives and work plan

MLEAP > > > Objectives & Roadmap

- Objectives Identification
  - Targeted objective

"The subject is the approval of machine learning (ML) ***technology for systems intended for use in safety-related applications in all domains covered by the EASA Basic Regulation (Regulation (EU) 2018/1139)****. The expected short-term effect of the research results will be to* ***streamline the certification and approval processes by identifying concrete means of compliance with the learning assurance objectives*** *of the EASA guidance for ML applications."*

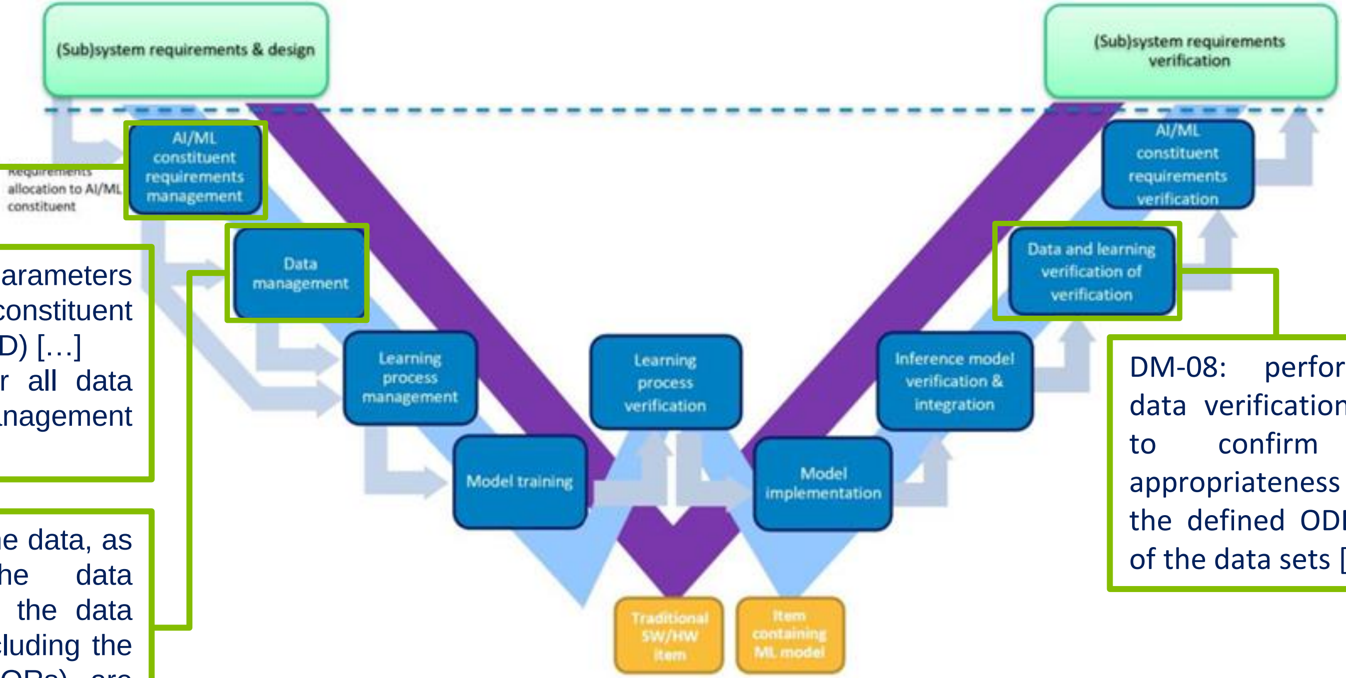
- Analysis of the objectives set by the EASA AI Roadmap
- Identify concrete means of compliance with the learning assurance objectives

- Selection of relevant use cases
- Real safety-related applications
- ML components are at the core of the systems' behaviour

- Set a development roadmap towards the objectives
- Put the conclusion all together for an end-to-end pipeline including the means of compliance

MLEAP > > > Objectives & Roadmap

- Objectives Identification
  - **Task 1**



DA-03: define the set of parameters pertaining to the AI/ML constituent operational design domain (ODD) […]
DA-04: capture the DQRs for all data pertaining to the data management process;

DM-07: ensure verification of the data, as appropriate, all along the data management process so that the data management requirements, including the data quality requirements (DQRs) are addressed.

DM-08: perform a data verification step to confirm the appropriateness of the defined ODD and of the data sets [..]

MLEAP > > > Objectives & Roadmap

- Objectives Identification
  - **Task 2**

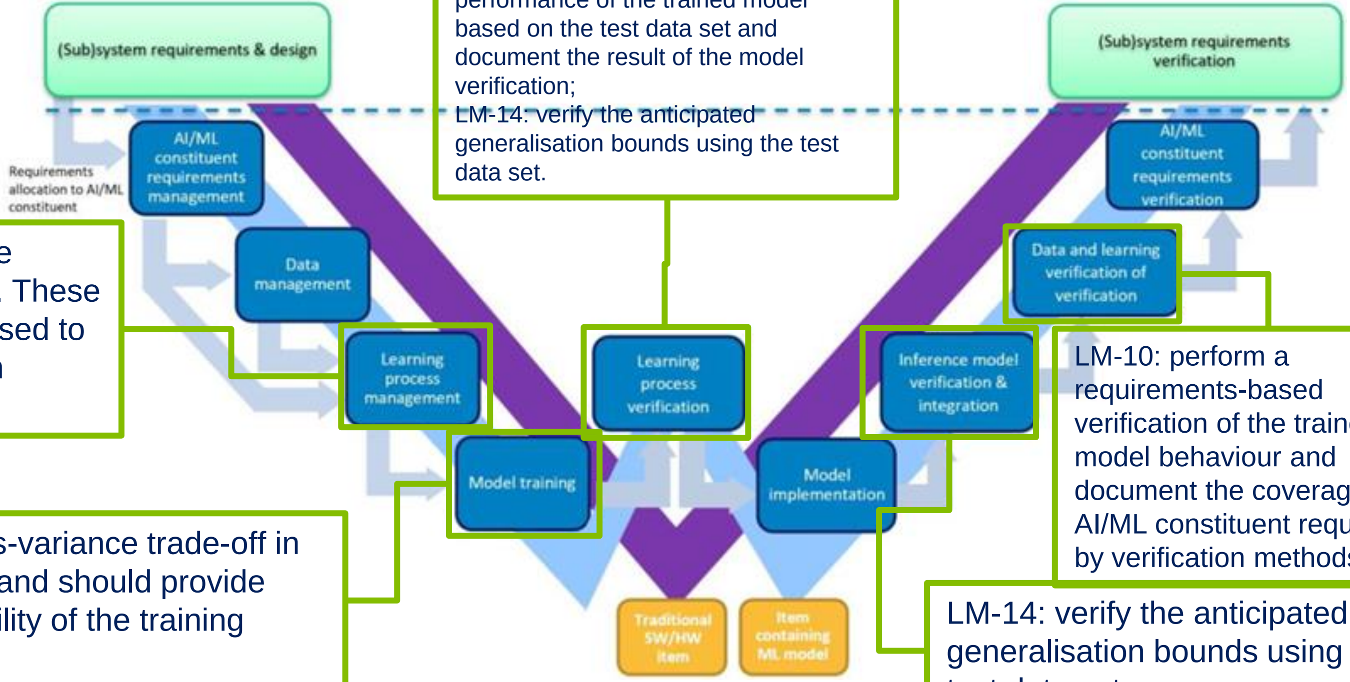
LM-09: perform an evaluation of the performance of the trained model based on the test data set and document the result of the model verification;
LM-14: verify the anticipated generalisation bounds using the test data set.

LM-04: provide quantifiable generalisation guarantees. These guarantees may then be used to support the Safety Case in Objective SA-01;

LM-07: account for the bias-variance trade-off in the model family selection and should provide evidence of the reproducibility of the training process;

LM-10: perform a requirements-based verification of the trained model behaviour and document the coverage of the AI/ML constituent requirements by verification methods;

LM-14: verify the anticipated generalisation bounds using the test data set.

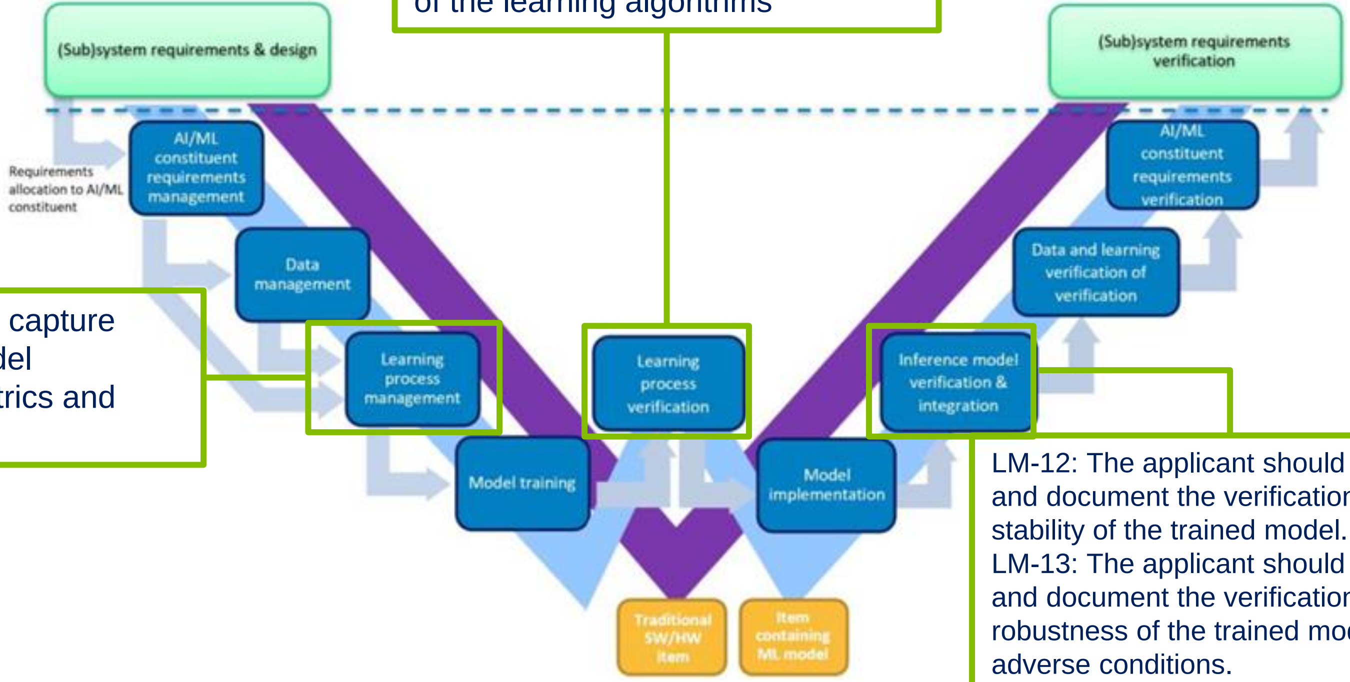


AIRBUS

MLEAP > > > Objectives & Roadmap

- Objectives Identification
  - **Task 3**

LM-11: The applicant should provide an analysis on the stability of the learning algorithms

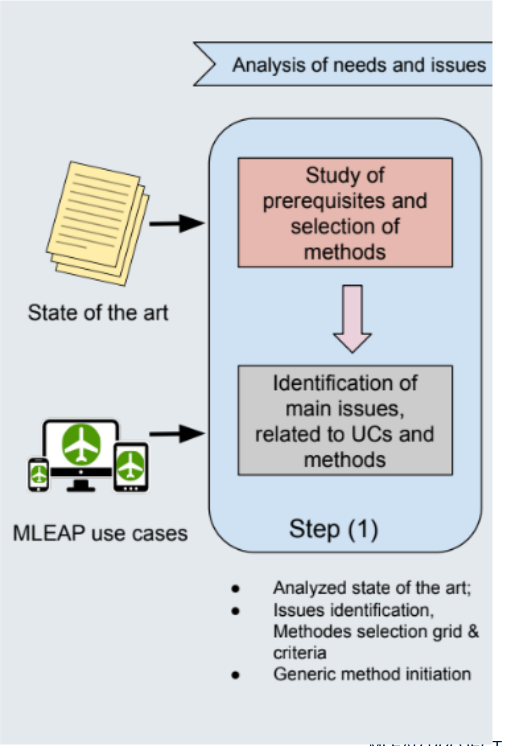


LM-02: the applicant should capture requirements related to model robustness and stability metrics and acceptable levels

LM-12: The applicant should perform and document the verification of the stability of the trained model.
LM-13: The applicant should perform and document the verification of the robustness of the trained model in adverse conditions.

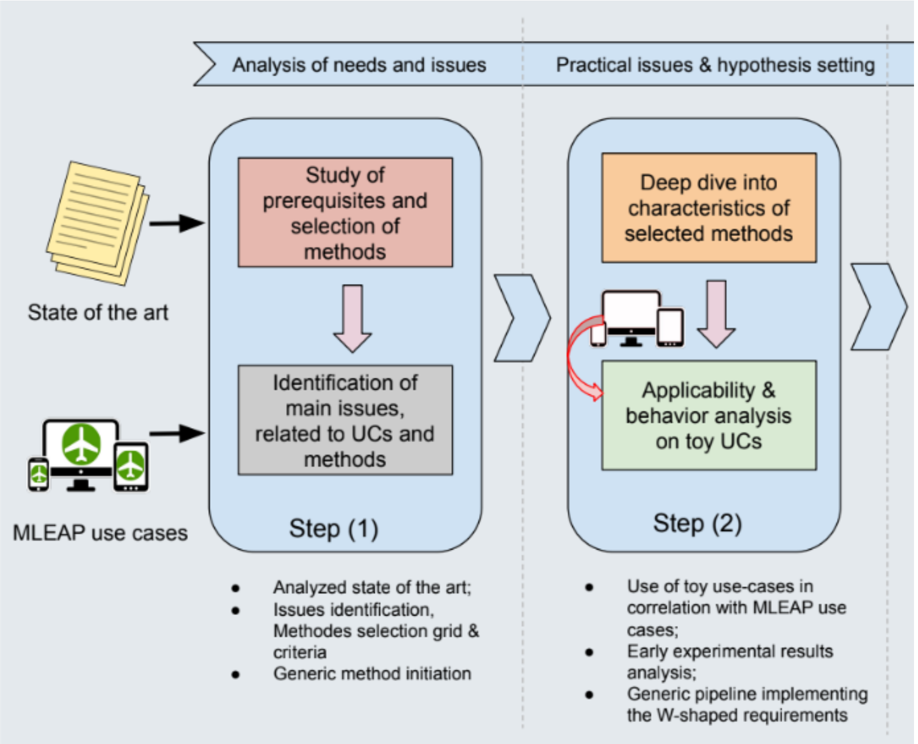
MLEAP > > > Objectives & Roadmap

- Roadmap



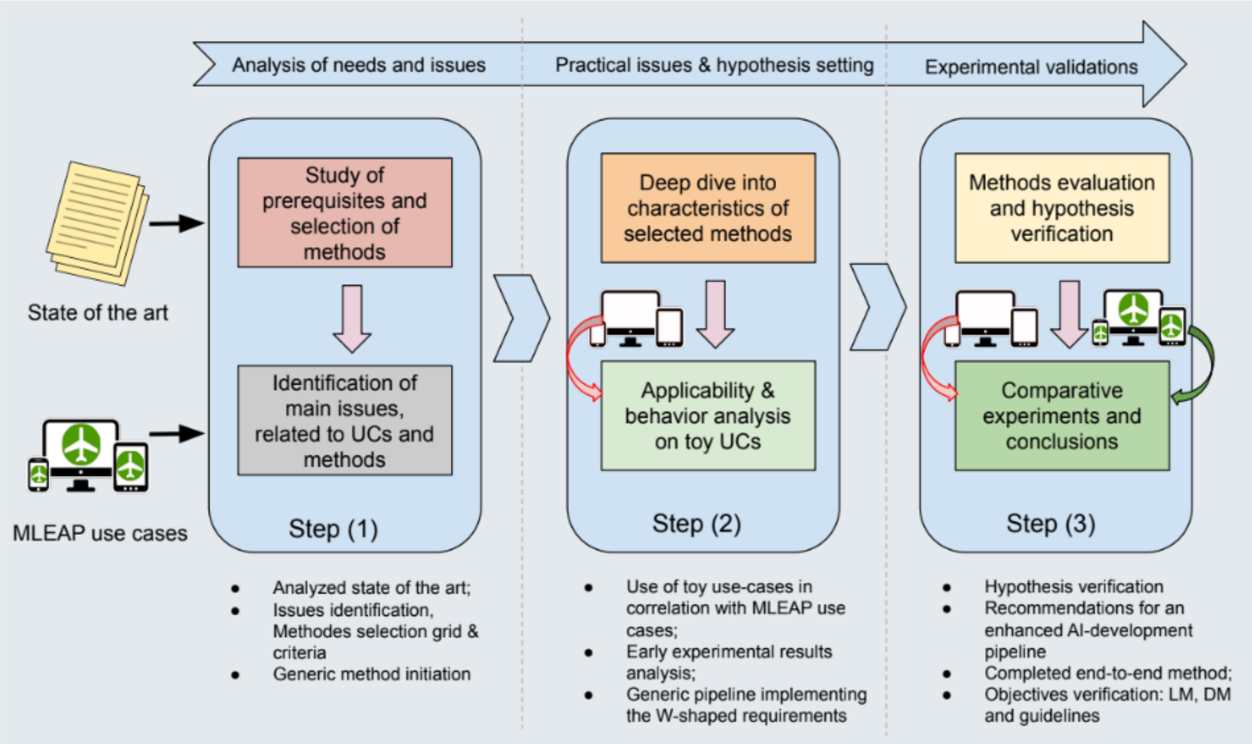
MLEAP > > > Objectives & Roadmap

- Roadmap



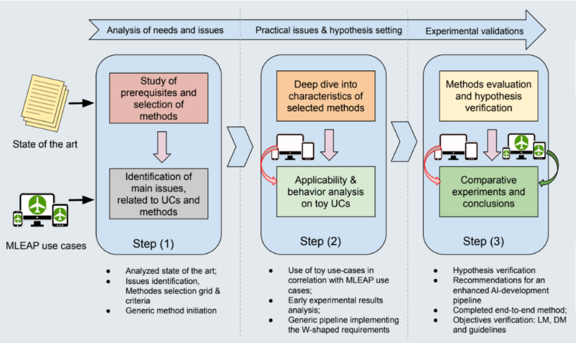
MLEAP > > > Objectives & Roadmap

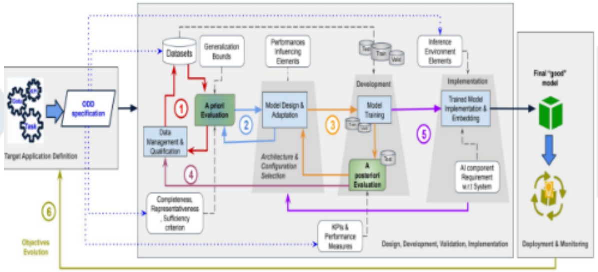
- Roadmap

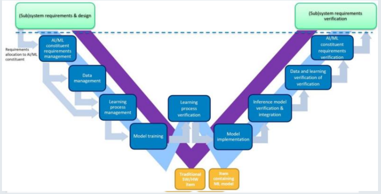


MLEAP > > > Roadmap & Objectives

- Roadmap







- Projection into the learning assurance: toward an enhanced w-shaped process
- Completed generic pipeline including the findings

Application-independent development process to meet the objectives of the target application and implement the certification requirements

/ Presentation of the use cases

Use Cases & Materials > > > Experimental Work

Toy use cases

- Less complex
  - Lower data dimensionality
  - Simpler tasks
- Open-source
  - Shareable results
  - Reproducibility of experimentations
- Applicability analysis
  - Equivalent applications to the target aviation use cases
  - Assess the method’s applicability and behaviour
  - Make a priori conclusions about the relevance of the selected methods towards the objectives

Aviation use cases

- More complex
  - Higher data dimensionality:
  - Complex tasks
- Real use cases relevant to the project’s objectives
- Validation of the a priori analysis of the selected methods
  - Applicability validation
  - Meeting objectives
- Make consistent conclusions supporting the roadmap of EASA
- Support the project conclusions with empirical results in known applications

Use Cases & Materials > > > Experimental Work

- Toy use cases

Application	Data set	Reference	Description
Images processing applications Classification & Objects Detection	FashionMNIST	https://github.com/zalandoresearch/fashion-mnist	Images classification (10 Zalando’s articles types); 60 000 training samples;
	MNIST	http://yann.lecun.com/exdb/mnist/	Images classification (10 digits); 60 000 training images;
	ROSE	https://www.challenge-rose.fr/	Plants detection & classification; 111 190 images;
	Rosetta	https://www.cosmos.esa.int/web/psa/rosetta	Object recognition (Craters detection in grey images); 1000 training samples;
Automatic Speech Recognition – Speech to Text	VoxCrim	https://lpp.cnrs.fr/la-recherche/projets-contrats/voxcrim/ https://voxcrim.univ-avignon.fr/#about	voice comparison systems used to identify criminals; 8338 audio samples of 400 speakers;
Time series	ECG Heartbeat	https://www.kaggle.com/datasets/shayanfazeli/heartbeat/data	Exploring heartbeat classification: normal and abnormal beats; 50 000 samples;

Use Cases & Materials > > > Experimental Work

- Aviation use cases

Rationales & Requirements	ATC-STT	ACAS Xu	AVI
High-level ODD	**Training Needs**: Acoustic and language models require complete data sets. **Data Completeness**: Includes noise types, airport checkpoint names, accents, and speech rates. **System Performance**: Full data ensures optimal system performance.	**Training Needs**: Data includes input points from RTCA SC-147 for ACAS-Xu's MOPS. **Data Completeness**: ODD is divided into sub-ODDs to fit 45 ML model elements. **System Performance**: Ensures ML model architecture aligns with operational standards.	**Training Needs**: Data is pictures of airframe structures under acceptable lighting and blur conditions. **Data Completeness**: Includes both indoor and outdoor pictures. **System Performance**: Outdoor weather conditions can influence lighting and blur state.
Performances and safety requirements derived from design & safety processes	**System requirements—Complex background noise.** The PESQ evaluation score represents operational conditions, 3.8 accepted, **System requirements – High speech rate** since ATC requires high timeliness **System requirements – Accents** The system must operate with French and Chinese accents	**System requirements – real-time 1s** The controller must execute with a period of 1s. **System requirements – anti-collision performance** Any implementation must behave similarly to the reference architecture **System requirements – ODD** The controller must operate on the ranges of the LUTs, i.e.	**ML-based requirements:** Focus on true positives ~ with 90% accuracy. **System requirements:** Solutions need to accommodate both indoor and outdoor environments. Detect both identified types of damage (**lightning strikes** and **dent impacts**).

Use Cases & Materials > > > Experimental Work

Aviation use cases

Speech-To-Text for Air Traffic Control (ATC-STT)

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

Correctly transcribe utterances into text

Support different accents of spoken English

Handle background noise

Model & Data: from Airbus internal project & open-source data/models

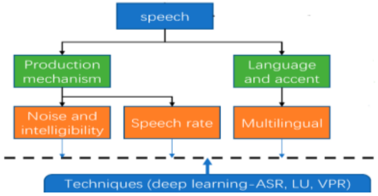
Models (classical and DL-based)

Airbus models: Kaldi STT models implemented with VOSK, accent/callsign models (DNN classifiers)

Open Source models: DL models, based on transformers facebook/wav2vec2-large-960h-iv60-self

MLEAP Challenges: robustness toward noise and different accents, accents detection, Callsign detection

The ASR research design concerned by the MLEAP project is part of a larger taxonomy provided in (Lin, 2021)



Data sets		Link	Whole Duration	Spoken Accent
Open Source	ATCO2 - ASR	https://www.atco2.org/data	1h 6 min	Yes: Czech, Slovak, German, French, Australian
	UWB	https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-CCA1-0	20h 35 min	Yes: Czech
	NIST LDC - Air Traffic Control Complete	https://catalog.ldc.upenn.edu/LDC94S14A	2h 02 min	No: US
	ATCOSIM	https://www.spsc.tugraz.at/databases-and-tools/atcosim-air-traffic-control-simulation-speech-corpus.html	10h 42 min	Yes: German, French
Proprietary	AIRBUS	-	150h	Yes: French, Chineese

Use Cases & Materials > > > Experimental Work

Aviation use cases

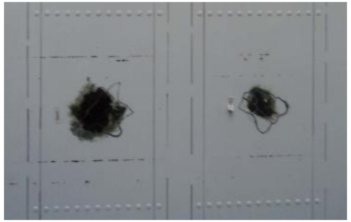
Automatic Visual Inspection (AVI)

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

Model & Data: from Airbus internal project & open-source

- **Data**: are made of two main parts**, lightning strikes** and **dent impacts**, with data augmentation (Changyu et al., 2014);
  - Acquisition of pictures is done from cameras and downloaded to the design/deployment environment;
  - Labeling is done using the VOTT tool, where every image can contain several damages of different classes;
  - Weighting samples to cope with imbalanced data sets
- **Model**: is made of a Siamese network constructed for a multitasking framework;
  - Aims to detect both the damage type (dent impact or lightning strike) and its characterization (severity level);
  - Using openCV library

Dents Damages (1)



Lightning Strikes (2)

(1) https://www.researchgate.net/figure/Wing-skin-metal-dent-examples_fig3_331961295

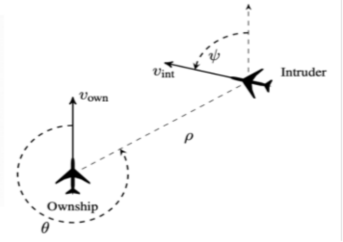
(2) https://www.researchgate.net/figure/Structural-damage-in-the-outer-skin-in-the-Airbus-A400-M-airplane-after-the-lightning_fig8_305817924

MLEAP Challenges:

- Automatic detection of external damages and their classification into two types: *lighting strike* impacts and *dents*;
- Targeted performance: >95% accuracy correctly detecting damages

Use Cases & Materials > > > Experimental Work

■ Aviation use cases



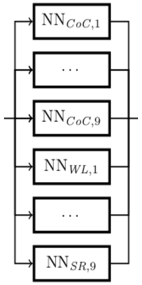
Next-Generation Airborne Collision Avoidance System for Unmanned Aircraft (ACAS Xu)

Objective:

solve ACAS problems (Bak and Tran, 2022) ACAS is a universal system-to-system collision avoidance

It issues horizontal turn advisories to avoid an intruder aircraft

Leverage NNs to



ML model elements of the ACAS Xu system

Model & Data:

The data consists of different entries of the LUTs from the RTCA SC-147 MOPS

The chosen action shall minimize the probability of collision

https://www.eurocontrol.int/publication/airborne-collision-avoidance-system-acas-guide

MLEAP Challenges:

In a context where the complete ODD is known, data quality is highly dependent on the LUTs

Models generalization & robustness are evaluated based on the ability of the model to compress LUTs correctly

Use Cases & Materials > > > Experimental Work

Dedicated Materials

MLEAP server hosted by Airbus Protect

CPU: Intel Xeon Gold 5220R 2.2GHz

RAM: 384 GB - 6x64GB

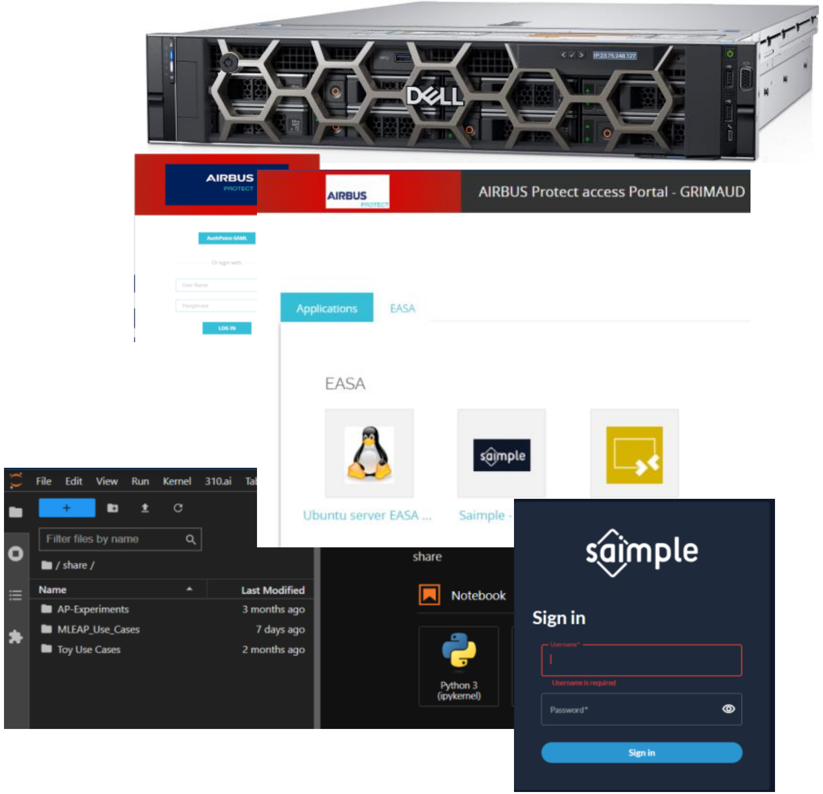
GPU: NVIDIA RTXTM A4000, 16Go, 4DP (Precision 7920T, 7820, 5820)

SSD: PCIe NVMe M.2 with 2TB extended to 4TB

Use cases and experiments accessible through a secured portal

Shared materials accessible in protected folders via **JupyterLab**

Numalis' proprietary tool (**Saimple**) installed locally



Q&A

www.sli.do
#AIDays
Passcode: hmkota

MLEAP project

AIRBUS

MLEAP > > > Coffee break / 10H20 – 11H00

www.sli.do

#AIDays

Passcode: hmkota

/ Presentation of the outcome and recommendations of Task 1

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*

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Context

- **Phase 1: Identifying assessment methods**
  - 80+ methods found and discussed
  - ~20 methods selected for further testing

- **Phase 2 & 3: Testing of methods on toy data sets**
  - Most methods are not « off-the-shelf »
  - Result analysis is not always a straightforward process
  - Some methods were filtered out

- **Phase 4 : Testing on MLEAP use cases**
  - Capitalizing on the experience of previous phases
  - Application to real-life data

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Key takeaways

- Methodology lacks structure
- Completeness harder ?
- Each AI task + dataset combo require a tailored assessment method
- 2 pillars for assessment : ODD vs model
- Trade off between completeness and representativeness

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Experimentations

- PCA
- Graph-based analysis
- Entropy analysis
- Sample-wise similarity
- Off-the-shelf tools
- Neuron Coverage
- Feature space characterization
- Completeness ratio
- Risk-based approach

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

PCA

- Dimension reduction technique for quantitative variables
- Applied on ACAS-Xu & AVI
- Intuition: **A complete and representative dataset yields a homogeneous scatter plot**
  - **ACAS-Xu** is a complete dataset, what happens if we visualize it ?
  - **AVI:** How data augmentation impacts completeness or representativeness?

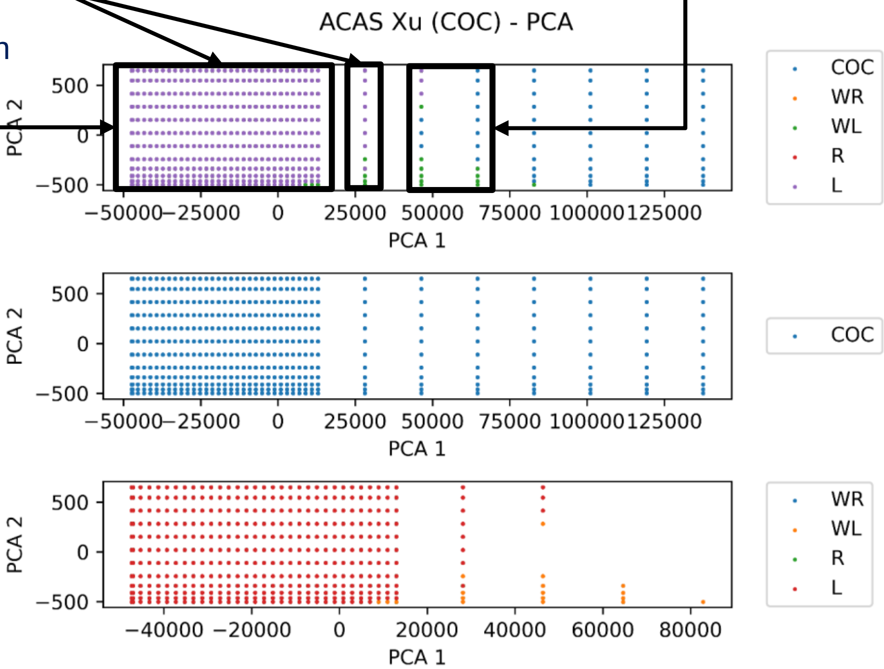
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

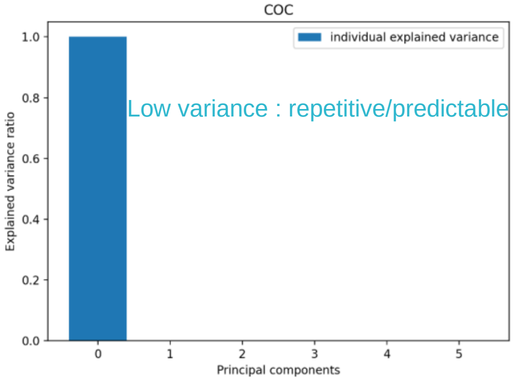
PCA : ACAS Xu

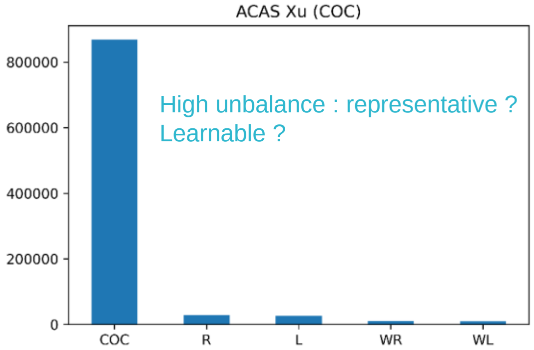
Regular patterns

Gradient of actions

More occurrences on sharper maneuvers

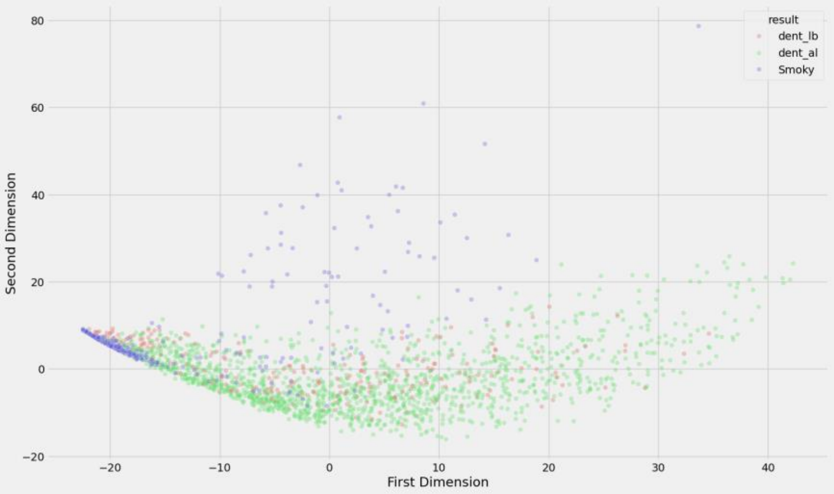




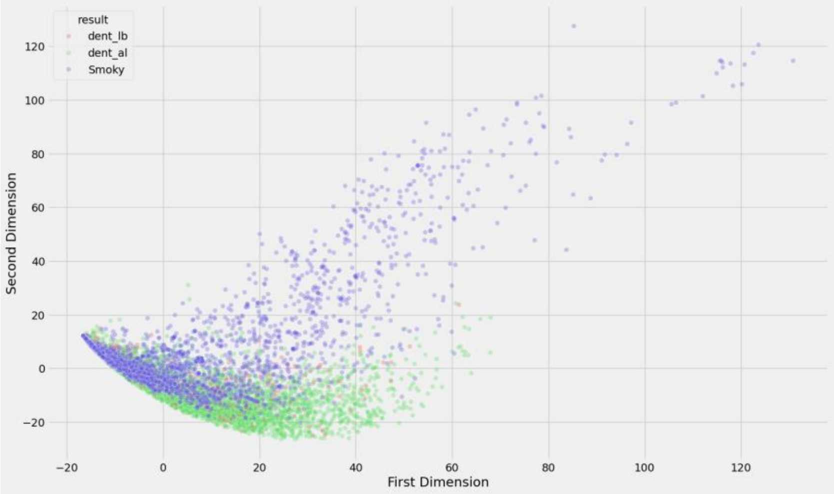


MLEAP – Task #1 Milestones Data completeness and Representativeness > > > PCA: AVI

AVI base



AVI augmented



- Higher density of data points : increased completeness
- Dent_lb not augmented
- Smaller spatial coverage : decreased representativeness

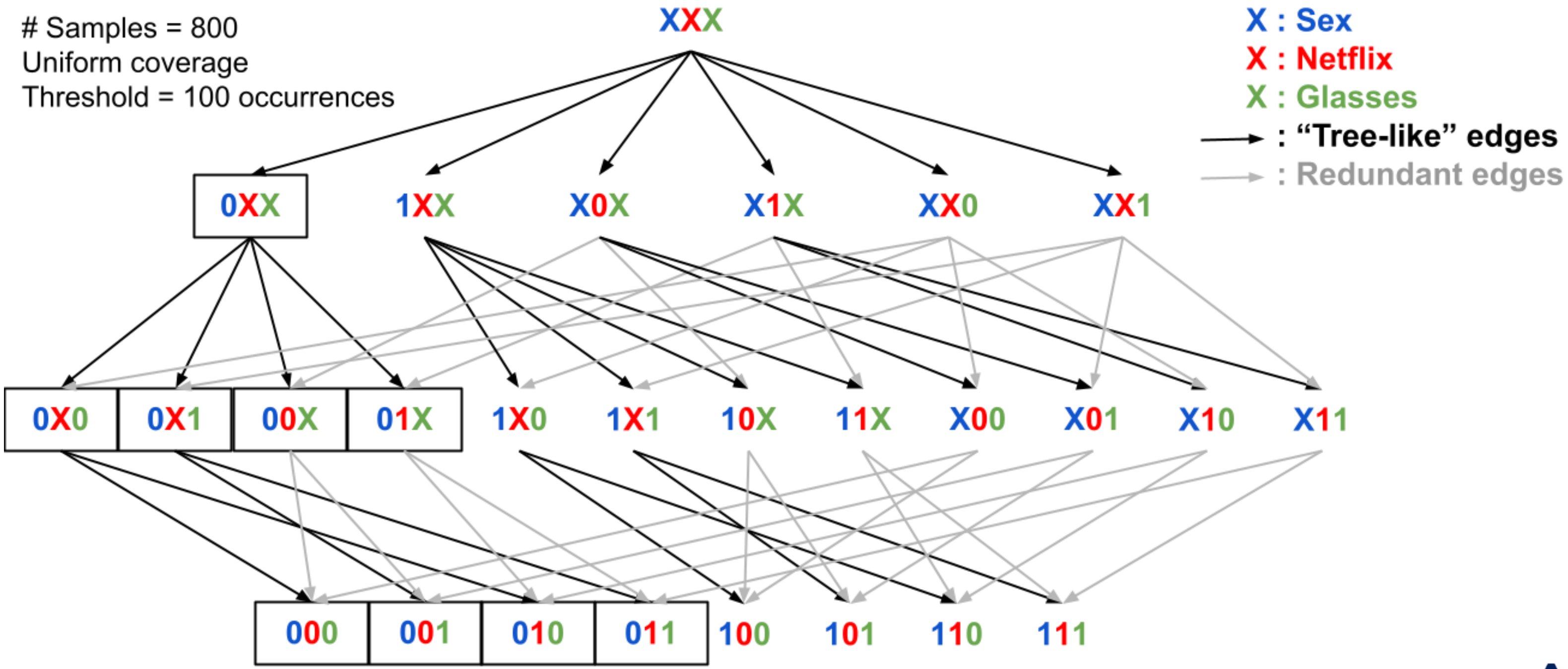
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Graph-based analysis

- Exhaustive coverage exploration
- Preferably for low-dimensional qualitative variables
- Mostly tested on toy datasets, implementation would benefit from more UX
- Identifies Maximum Uncovered Patterns

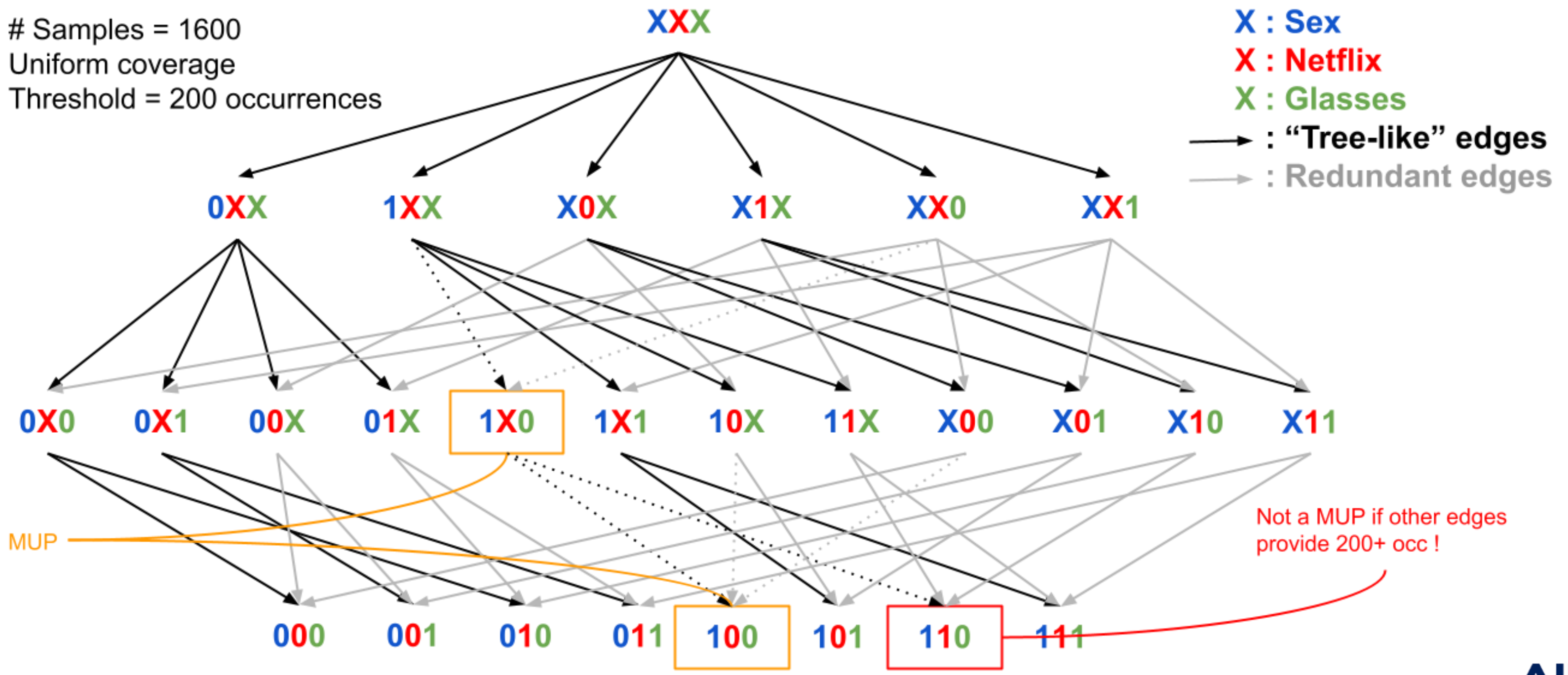
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Graph-based analysis



MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Graph-based analysis



MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Graph-based analysis

- Inherently useful for completeness
- Can be tweaked for representativeness
- Dependent of the chosen threshold

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Entropy analysis

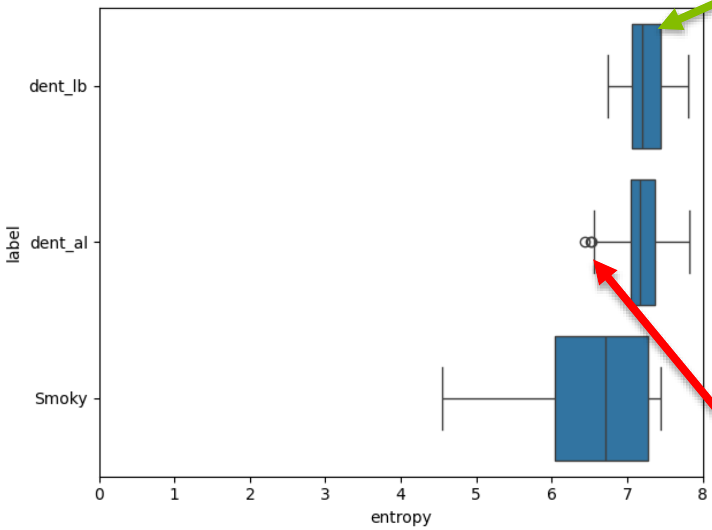
- Useful for high-dimensional data (image, audio)
- Tested on AVI
- Intuition: **heterogeneous entropy across classes might be indicative of representativeness discrepancy**

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

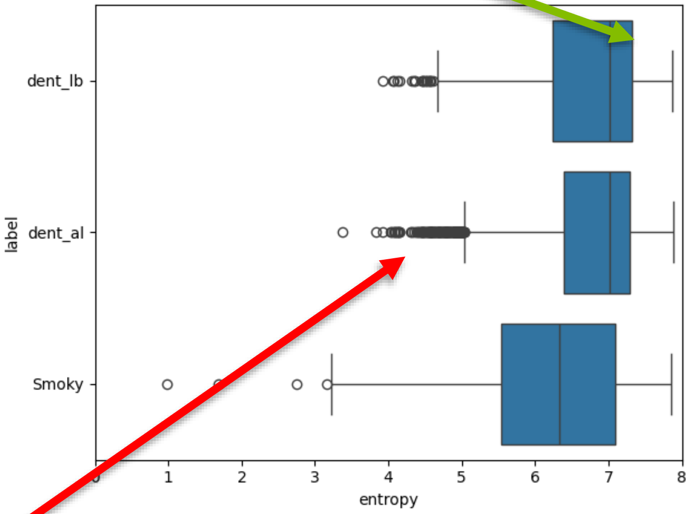
Broader box, similar mean : homogeneous extension

Entropy analysis

AVI base dataset (image-wise)



AVI augmented dataset (image-wise)

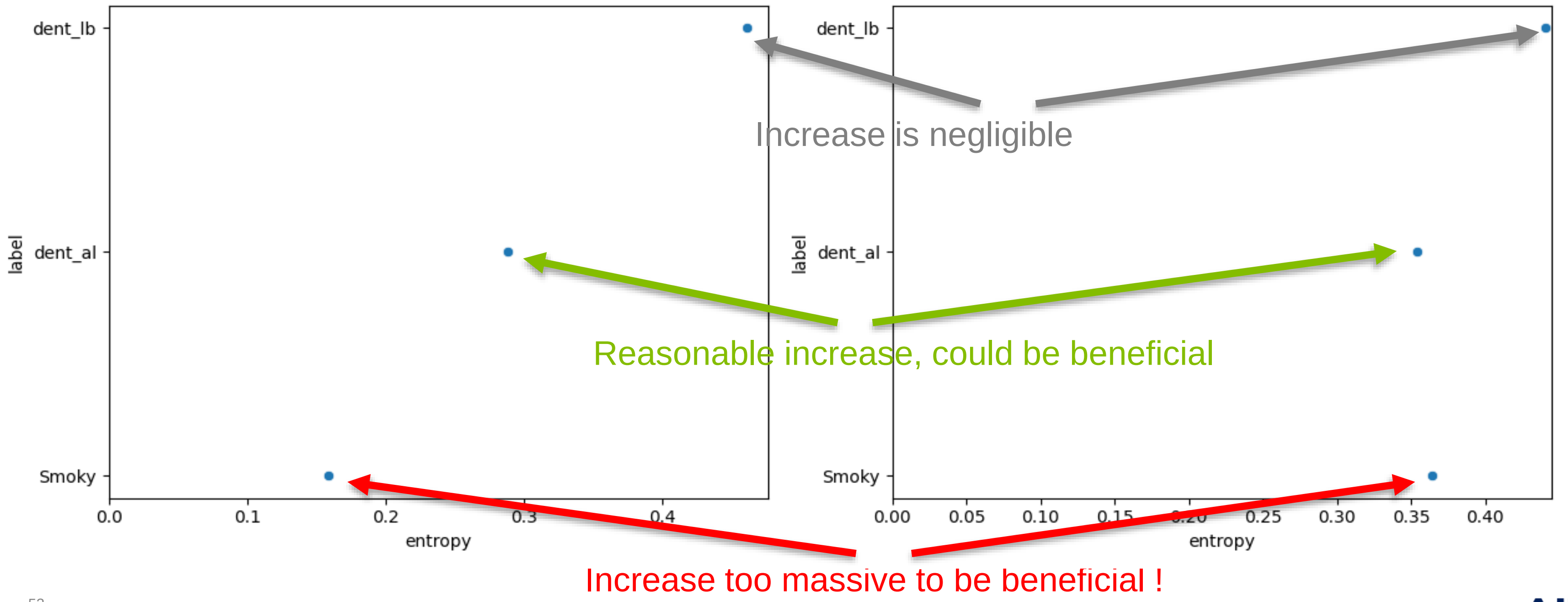


Larger whiskers and more outliers : heterogeneous addition

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Entropy analysis

AVI base dataset (label-wise) AVI augmented dataset (label-wise)



MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Entropy analysis

- A coarse grain tool but a good entry point
- Inter-class entropy might just be e.g. a « harder » class
  - Depends on the diversity of the classes

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

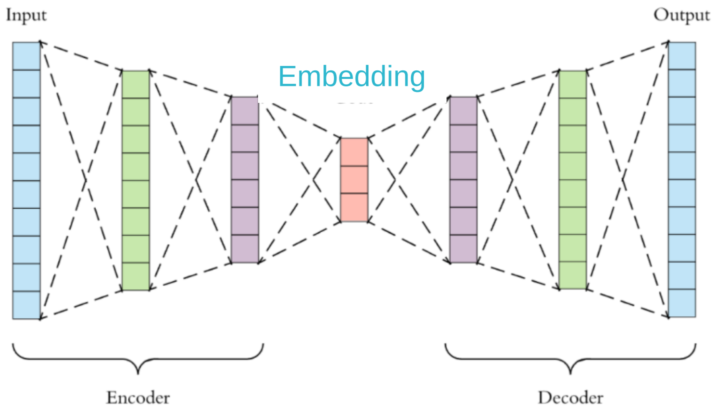
Sample-wise similarity

- Method for high-dimensional data
- Useful for hard-to-assess data such as audio
- Uses embeddings as proxies
- **Intuition: using the embedding space to assess latent properties**
- Not tested on aviation UC

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Sample-wise similarity

- What is an embedding ?
  - Input representation
  - Vectors space
  - « Low »-dimensional
- Objective: assessing the completeness of an audio data set (target: ATC-STT)
- Capacity needed: semantic similarity assessment
- 4 types of speech embeddings tested
- 0 have a semantic aspect



MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Sample-wise similarity

- Compatible with virtually any unstructured data set
- Brings structure !
- Depends on the properties encoded into the embeddings
- Requires a relevant similarity metric

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Off-the-shelf tools

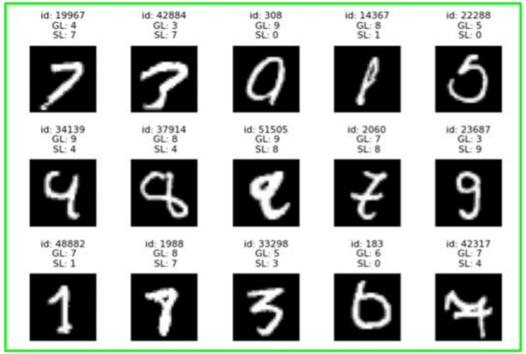
- Cleanlab tested
  - A prominent, open-source suite
  - Can process images, audio, text, tabular data
  - **Provides metrics about**
    - **Mislabellings**
    - **Outliers**
    - **Near-duplicates**
    - **Specific metrics e.g. odd-ratio for images**
- Tested on AVI

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

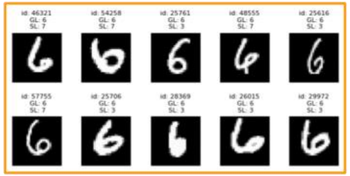
Off-the-shelf tools

MNIST: Image classification (outliers)

97% accuracy classifier



75% accuracy classifier



Total images : 60k
2602 outliers;
722 near duplicates**;**
120 labelling errors
0 blurry images;
0 dark images;
0 light images;
0 odd aspect ratio;
0 odd-size

AVI : Object detection

	Dents			Lightning strike		
	Train	Val	Test	Train	Val	Test
Total images	3659	1044	522	28	6	3
Blurry	284 (7.7%)	68 (6.5%)	35 (6.7%)	0	0	0
Low information	0	0	0	0	0	0
Dark	0	0	0	0	0	0
Light	0	0	0	0	0	0
Odd size	231 (6.3%)	73 (6.9%)	22 (4.2%)	0	1 (16.6%)	0
Odd aspect ratio	0	0	0	0	0	0
Grayscale	0	0	0	0	0	0
Near duplicate	143 (3.9%)	15 (1.4%)	5 (0.9%)	2 (7.1%)	0	0
Exact duplicate	0	0	0	0	0	0

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Off-the-shelf tools

- Cleanlab is not a silver bullet
- A useful suite for classification
  - Helps highlight edge/corner/hard cases
- Only on classification tasks
- Assessment heavily depend on the model
  - Need for a mature model
  - Is it worth it to backtrack on the data ?
- Cannot replace human examination
  - Reduces cost by highlighting points of interest

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

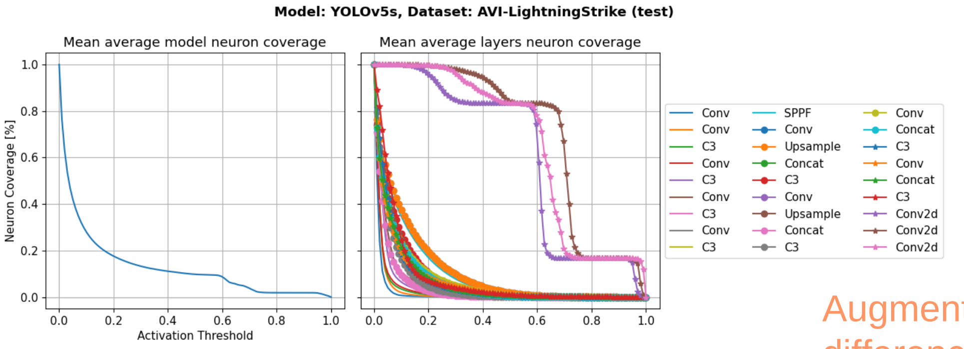
Neuron coverage

- Model-centric approach
  - Observing the activation states of a neural net
  - Data agnostic
- **Intuition: observe how the model reacts to data to infer possible lacks of completeness or representativeness**
- Tested on AVI

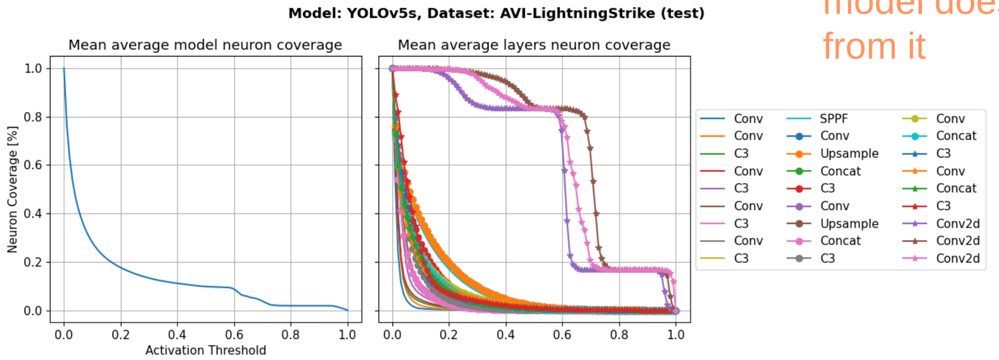
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Neuron coverage

AVI base (test set)



AVI augmented (test set)



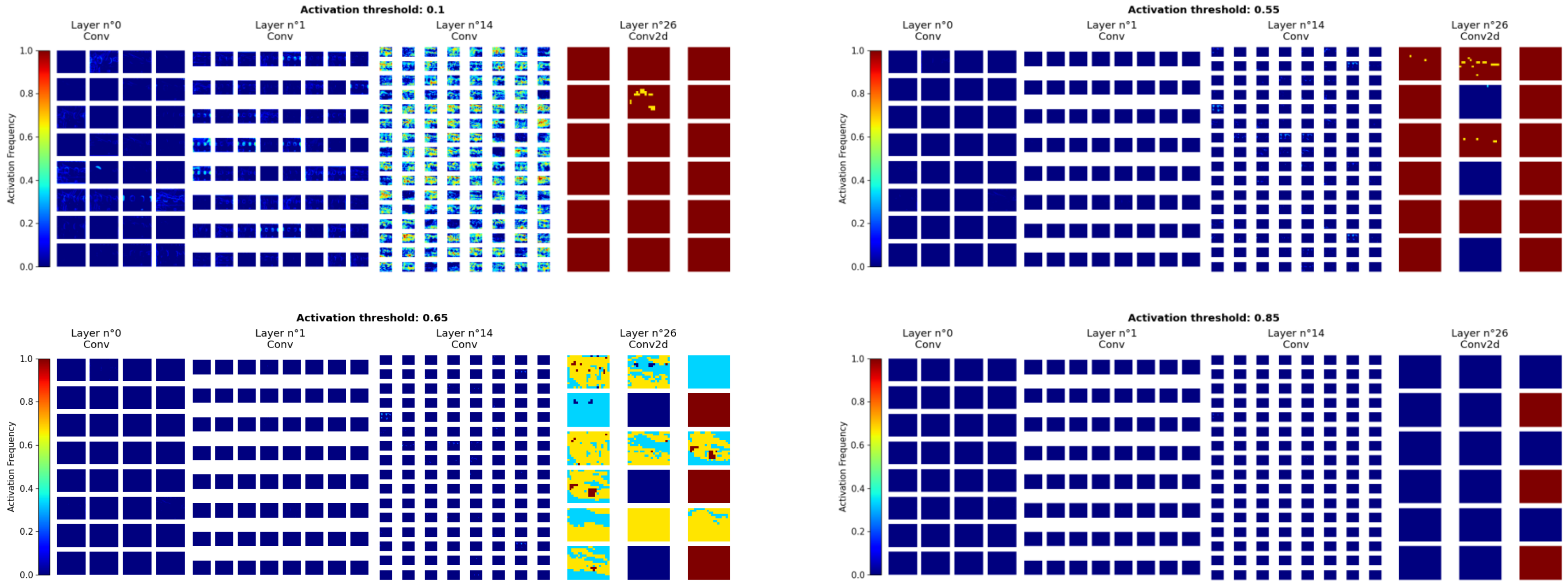
Augmentation shows no difference in trends: the model does not learn from it

AIRBUS

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Neuron coverage

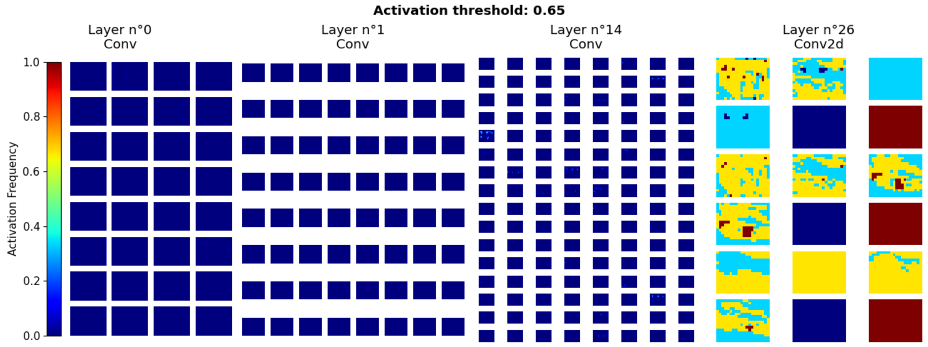
AVI base (test set)



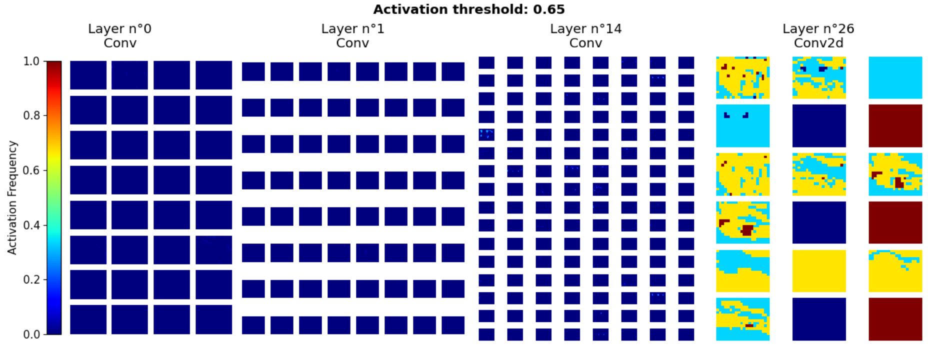
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Neuron coverage

AVI base (test set)



AVI augmented (test set)



Augmentation shows no difference in trends: the model does not learn from it

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Neuron coverage

- Very flexible in terms of possible visualisations
- Enables monitoring
- Requires white box access (better for in-house models)
- Takes some engineering

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Feature space characterization

- **Model-centric approach**
- 4 metrics:
  - Equivalence Partitioning
  - Centroid Positioning
  - Boundary Conditioning
  - Pairwise Boundary Conditioning
- Intuition: **a homogeneous feature space is indicative of a complete dataset (learning-wise)**
- Tested on AVI

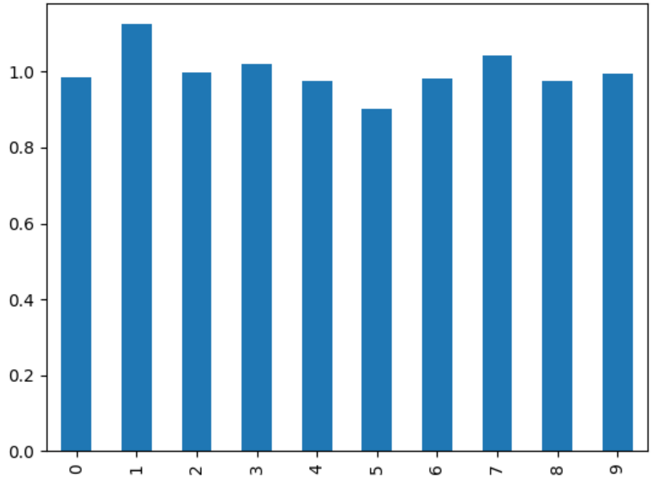
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Feature space characterization

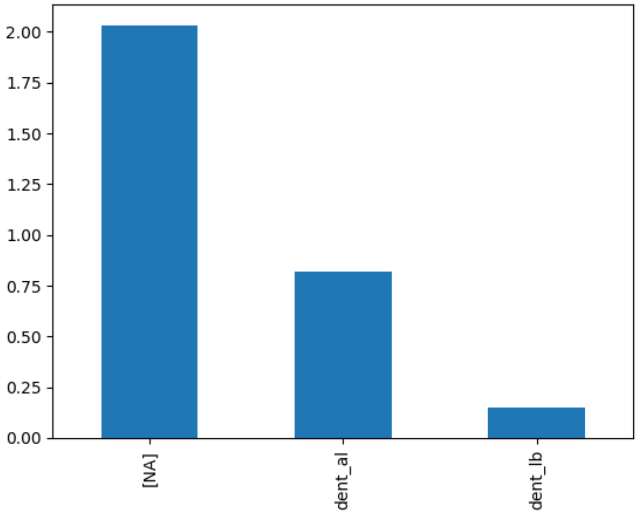
Equivalence partitioning

- Measures the class-wise balance of a dataset
- All classes should converge to 1

MNIST



AVI (base, dents only)



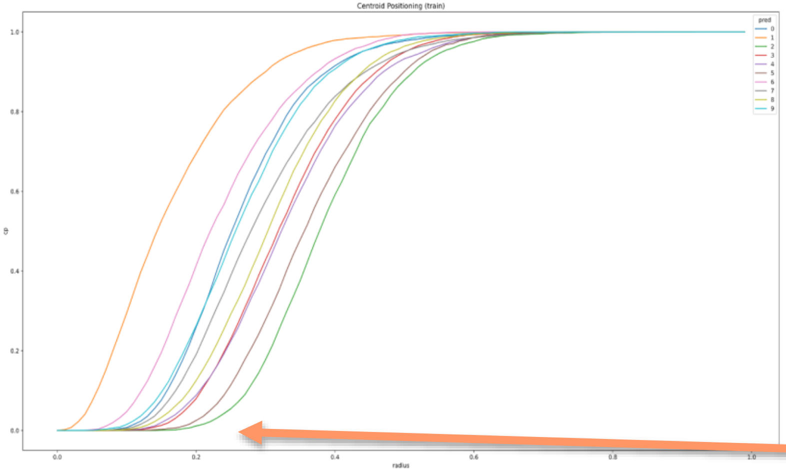
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Feature space characterization

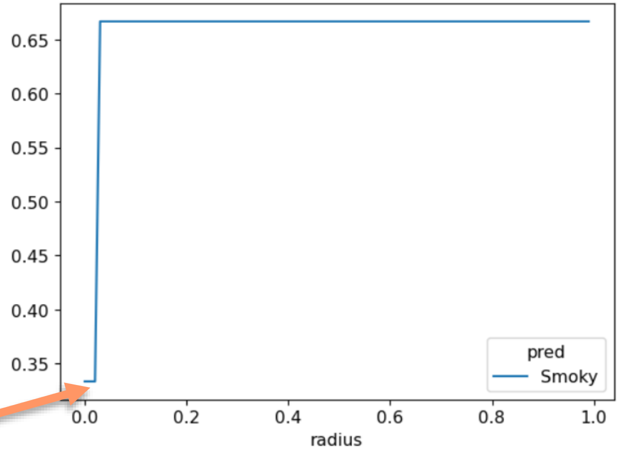
Centroid positioning

- Sample homogeneity score in a given radius
- The lower, the better

MNIST



AVI (base, lightning



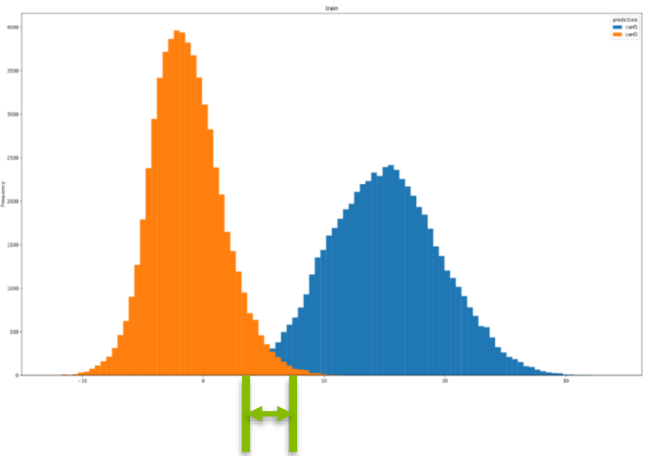
Both datasets diverge almost immediately

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

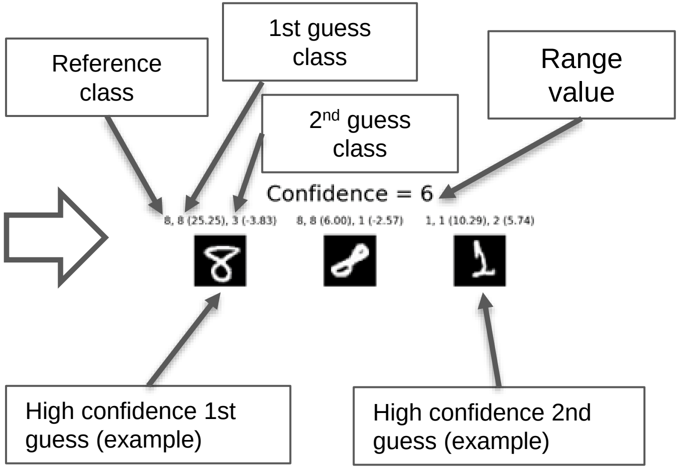
Feature space characterization

Boundary conditioning
- Compare confidence scores for best and second guesses
- Define a confidence range : the boundary

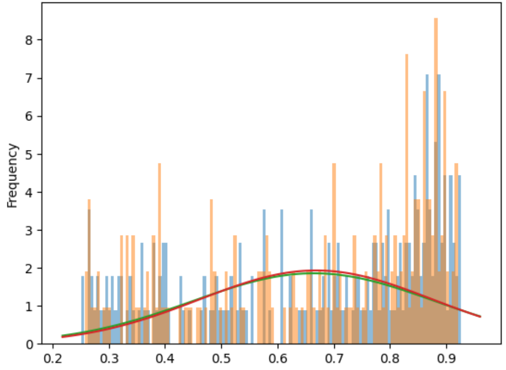
MNIST



Nice range (classification task ?)



AVI (base, dents only)



No identifiable range (detection task ?)

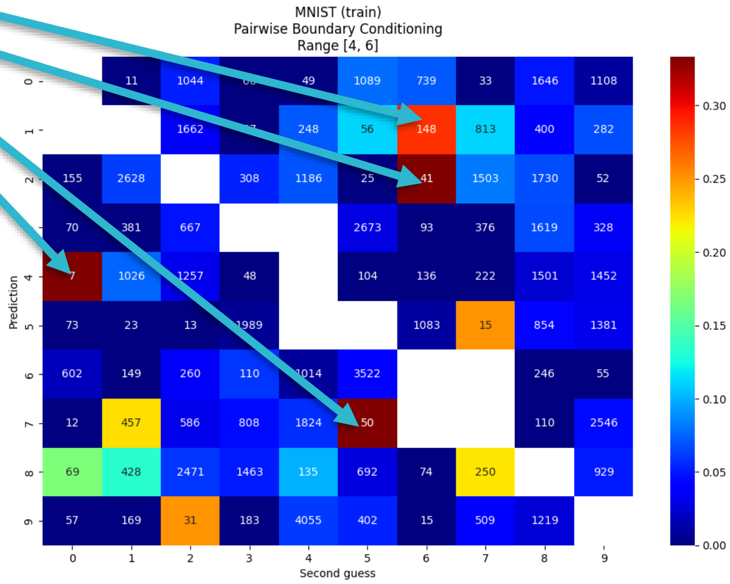
MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Feature space characterization

Pairwise Boundary conditioning
- Aggregate all boundaries for each class

Most « confusing » classes
- 1 & 6
- 2 & 6
- 7 & 5
- 4 & 0

MNIST



AVI: NA

AIRBUS

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Feature space characterization

- Data-agnostic…
- …but not task-agnostic

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Completeness ratios

- Metrics for tabular data (including metadata for more complex data sets)
- **Illustrate different notions of completeness**
- Not tested on aviation UC

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Completeness ratios

- 4 metrics from the literature
  - **Documentation**: ratio of complete samples (i.e. no missing features)
  - **Breadth**: distribution of feature completeness (as per **documentation**)
  - **Density**: # of samples with a given feature combination (cf graph-based)
  - **Predictive**: availability of sufficient information to predict an outcome

- 3 derived metrics
  - **G1**: column-wise feature completeness
  - **G2**: row-wise feature completeness
  - **G3**: absolute ratio of missing value

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Risk-based approach

- Methodology by the **B**usiness **S**oftware **A**lliance
- Aimed at adressing population bias in demographic data
- **Motivation : bias is a facet of representativeness**
- **Question: can this method be extended to any type of data set ?**

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

Risk-based approach

- **The method is indeed data-agnostic**
- Easy to apply : few tools required
- Rests heavily on expert knowledge
- Provides guidelines rather than a straightforward method
  - Without experts, the conclusions may remain too general

MLEAP – Task #1 Milestones Data completeness and Representativeness > > >

General conclusions

- **Not a prescriptive work**
- Data qualification is hard
  - MLEAP showcases some methods
  - Applicants can be a driving force in bringing methods to the table
  - Keeping in mind their accountability in the end

- Aeronautics is the tip of the spear for AI reliability
  - Pioneers of operational industrial-grade methods

Q&A

www.sli.do

#AIDays

Passcode: hmkota

MLEAP project

MLEAP > > > Lunch break / 12H00 – 13H00

www.sli.do

#AIDays

Passcode: hmkota

/ Presentation of the outcome and recommendations of Task 2

MLEAP – Task #2 milestones: 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 > > >

Context

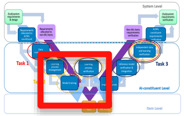
Quantification of generalization assurance level: main Concept paper objectives

- LM-04: provide quantifiable generalization guarantees.
- LM-09: performance evaluation of the trained model based on the test data set
- LM-14: verify the anticipated generalization bounds using the test data set.

Main focus

- Generalization bounds theory
- Drivers steps influencing generalization

Learning assurance process steps concerned



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

Work done

Phase 1: SOTA

- 13 generalization bounds selected
- Identification of methods to boost generalization and their limitations

Phase 2 & 3: First tests of methods identified

- Bounds evaluation coding and computation (Some have been filtered out)
- Trained models performance analysis w.r.t. generalization
- Issues identification and improvement proposal

Phase 4 : Tests on aviation use cases

- Capitalizing on the experience of previous phases
- Test improvements proposed
- Bounds evaluation on complex use cases

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

Generalization

WHAT is generalization?

Generalizability is the capacity of a model to generalize that is to say to keep same level of average performance on unseen data.

WHY are we interesting by generalization?

It is to demonstrate the ability of an AI trained model to handle real world variability and maintain performances across different operating conditions

How to assess generalizability ?

- Performance measurement on test and validation dataset
- Generalization bounds evaluation:
  - Upper bounding the Expected true risk
  - Generalization capacity and “good” model identification
  - Theoretical guidance
- Guidance during development workflow steps

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

Learning assurance process steps

Learning process mgt

- Preparatory step of the formal training phase.
- Selection and validation of key elements such as
  - → the training algorithm,
  - → the activation function,
  - → the loss function,
  - → the initialization strategy,
  - → the training hyperparameters
  - → The metrics that will be used for the various validation and verification steps

Model training

- Executing the training algorithm in the conditions defined in the previous step, using the training dataset from the data management process step.
- Model performance evaluation (bias and variance) using the validation dataset.

Learning process verification

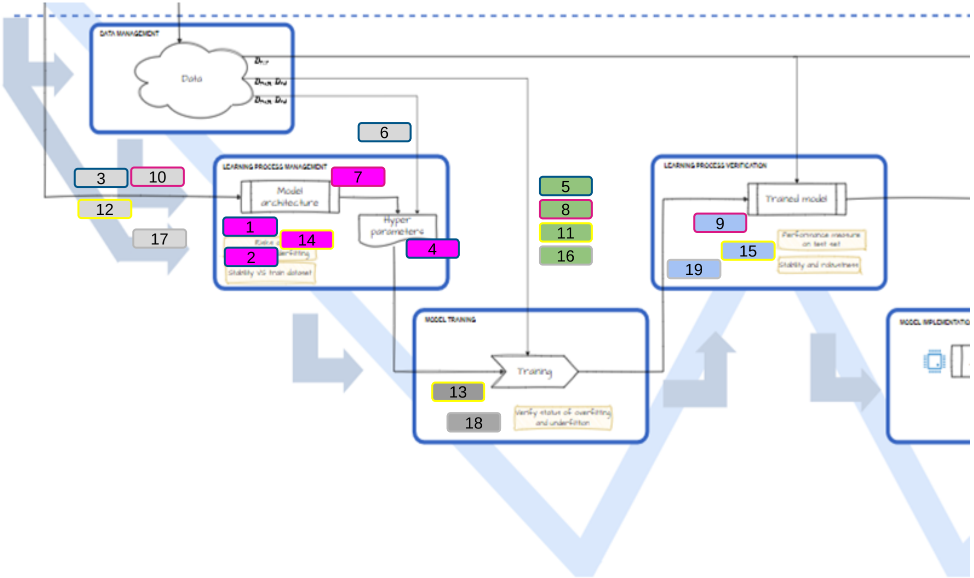
- Evaluation of the trained model on the test dataset
- Evaluation of the bias and variance of the trained model

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

Model type	Use case type	Task	Data type	Model type	Dimmensionality	Concept paper objectives	Experimentations
Toy	Fashion MNIST	Classifier	Images	DNN (FCNN & CNN)	High	LM-04, LM-07, LM-09, LM-14	Bounds theory wrt architecture selection - A priori Generalization bounds
							Architecture optimization to minimize generalization bounds
							Data augmentation influence on test performance
							Architecture selection based on hyper-parameters analysis
							A priori & A posteriori generalization bounds evaluation
							Training dataset size
Avionic	ATC-STT	Speech to text	Audio	Kaldi, transformers	High	LM-04, LM-07, LM-09, LM-14	architecture comparison
							A priori & A posteriori generalization bounds evaluation
							Performance evaluation on test dataset
							Training data representativeness wrt generalization
	AVI	Object detection	Images	Yolo	High	LM-04, LM-07, LM-09, LM-14	A priori & A posteriori generalization bounds evaluation
							Data augmentation & Training data representativeness wrt generalization
							Fine Tuning
							Architecture comparison yolov5 yolov8
							Performance evaluation on test dataset
	ACAS Xu	Regression	5 numerical values	FCNN	Low	LM-04, LM-07, LM-09, LM-14	A priori & A posteriori generalization bounds evaluation
							Data augmentation
							Weighted loss function

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

Model type	Use case type	Test	
Toy	Fashion MNIST	Bounds theory wrt architecture selection - A priori Generalization bounds	1
		Architecture optimization to minimize generalization bounds	2
		Data augmentation influence on test performance	3
		Architecture selection based on hyper-parameters analysis	4
		A priori & A posteriori generalization bounds evaluation	5
		Training dataset size	6
Avionic	ATC-STT	architecture comparison	7
		A priori & A posteriori generalization bounds evaluation	8
		Performance evaluation on test dataset	9
		Training data representativness wrt generalization	10
	AVI	A priori & A posteriori generalization bounds evaluation	11
		Data augmentation & Training data representativness wrt generalization	12
		Finetuning	13
		Architecture comparison yolov5 yolov8	14
		Performance evaluation on test dataset	15
	ACAS Xu	A priori & A posteriori generalization bounds evaluation	16
		Data augmentation	17
		Weighted loss function	18



Legend

- Learning process management
- Linked with data management
- Model training
- Learning process verification
- Learning process management and verification
- FashionMNIST
- AVI
- ACAS Xu
- ATC STT

AIRBUS

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

Generalization bounds

Generalization bounds aim to provide bound the gap between the true risk and the empirical one.

$$\forall \mathcal{D}\quad \mathbb{P}_{\mathcal{D}\sim\mathcal{S}}\left[\left|L_{\mathcal{D}}(W) - L_{\mathcal{S}}(W)\right| \le \underbrace{\boldsymbol{\varepsilon(\mathcal{H}, m, \delta, \mathcal{D}, \mathcal{S}, Optim, W)}}_{\text{Generalization bound}}\right] > 1 - \delta$$

Experimental objectives:

- Test Generalization bounds as MoC to answer Objective 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:

- Tests and analysis a priori results on 2 different architectures FCNN and CNN
- Train and test several models on different use cases
- Analyze a posteriori generalization bounds regarding assurance level upper bounds

Algo.	Ref.	Bound
CNN	(Lin and Zhang, 2019)	$R_{\mathcal{D}}(F_C) \le \hat{R}_{S,l_\eta}(F_C) + \tilde{O}\left(\frac{s^{\frac{L-1}{4}} L^{\frac{3}{4}} a^{\frac{1}{4}} c^{\frac{1}{2}} m^{\frac{1}{8}} r^{\frac{1}{2}}}{\sqrt{N}}\right) + \tilde{O}\left(\frac{s^{\frac{L-1}{4}} L^{\frac{3}{4}} a^{\frac{1}{4}} d}{\sqrt{N}}\right)$
NN for classification	(P. Jin et al., 2020)	$\mathbb{P}_{\mathcal{D}}\left[\forall h_\theta \in \mathcal{H}, \mathbb{E}_{x,y\sim\mathbb{P}(X,Y)}[\mathcal{R}(h_\theta)] \le \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)\right] + \frac{\sqrt{d}.CD(\mathcal{D})}{\min(\delta_0, \kappa\delta_{\mathcal{D}})}\right] \ge 1-\delta$
NN	(Alquier, 2021)	Cantoni's bound (PAC Bayes) $\mathbb{P}_{\mathcal{D}_0}\left[\forall\rho \in \mathcal{P}(\Theta), \quad \mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] \le \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}_0}_{emp}(h_\theta)\right] + \frac{\lambda C^2}{8N_0} + \frac{KL(\rho\Vert\pi) + \log\frac{1}{\varepsilon}}{\lambda}\right] \ge 1-\varepsilon$
	(Alquier, 2021) (McAllester, 1998)	Mc Allester's bound $\mathbb{P}_{\mathcal{D}_0}\left[\mathbb{E}_{\theta\sim\rho}[\mathcal{R}(\theta)] \le \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}_0}_{emp}(\theta)\right] + \sqrt{\frac{KL(\rho\Vert\pi) + \log\frac{1}{\varepsilon} + \frac{5}{2}\log(N_{\mathcal{D}_0}) + 8}{2N_{\mathcal{D}_0} - 1}}\right] \ge 1-\varepsilon$
	(Alquier, 2021) (Seeger, 2002)	Seeger's bound $\mathbb{P}_{\mathcal{D}}\left[\forall\rho \in \mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}(h_\theta)] \le kl^{-1}\left(\mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)\right] \middle\vert \frac{KL(\rho\Vert\pi) + \log\frac{2\sqrt{N_{\mathcal{D}}}}{\epsilon}}{N_{\mathcal{D}}}\right)\right] \ge 1-\epsilon$
	(Alquier, 2021) (Tolstikhin and Seldin, 2013)	Tolstikhin and Seldin's bound $\mathbb{P}_{\mathcal{D}}\left[\begin{array}{c}\forall\rho \in \mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] \le \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)] + \\ \sqrt{2\mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)]\frac{KL(\rho\Vert\pi) + \log\frac{2\sqrt{N_{\mathcal{D}}}}{\epsilon}}{2N_{\mathcal{D}}}} + 2\frac{KL(\rho\Vert\pi) + \log\frac{2\sqrt{nN_{\mathcal{D}}}}{\epsilon}}{2N_{\mathcal{D}}}\end{array}\right] \ge 1-\epsilon$
Fully connected NN & CNN	(Arora et al., 2018)	$\mathbb{P}_{\mathcal{D}_0}\left[\mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] \le \mathbb{E}_{\theta\sim\rho}\left[\mathcal{R}^{\mathcal{D}_0}_{emp}(\hat{h}_\theta)\right] + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x\in\mathcal{D}_0}\Vert h_\theta(x)\Vert_2^2 \sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\to}^2}}{\gamma^2 N_{\mathcal{D}_0}}}\right)\right] \ge 1-\delta$
Two class classifier	(Anthony, 2004)	$\mathbb{P}_{\mathcal{D}}\left[\mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] < \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)] + \sqrt{\frac{8}{N_{\mathcal{D}}}\left((n+k-1)\log\left(\frac{2eN_{\mathcal{D}}k}{n+k-1}\right) + \log\left(\frac{4}{\epsilon}\right)\right)}\right] \ge 1-\varepsilon$
Supervised learning	(Neu and Lugosi, 2022)	$\lvert\mathbb{E}[\mathrm{gen}(W_n, S_n)]\rvert \le \sqrt{\frac{4H(P_n)\mathbb{E}[\Vert\bar{l}(.,Z)\Vert_*^2]}{\alpha n}}$
γ-uniformly stable learning algorithm	(Feldman and Vondrak, 2018)	$\mathbb{P}_{\mathcal{D}}\left[\mathbb{E}_{\theta\sim\rho}[\mathcal{R}(h_\theta)] \le \mathbb{E}_{\theta\sim\rho}[\mathcal{R}^{\mathcal{D}}_{emp}(h_\theta)] + 8\sqrt{\left(2\gamma + \frac{1}{N_{\mathcal{D}}}\right).\log\frac{8}{\varepsilon}}\right] \le 1-\varepsilon$

17 bounds selected built from diferent theoretical framework:

- *Uniform convergence*
- *Uniform stability*
- *Algorithm robustness*
- *Measures related to optimization*

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

Generalization bounds – statistical guarantee

A Priori evaluation:

- Pessimistic as the theory remain valid in worst case and are vacuous for over-parametrized NN
- Pac Bayes bounds, complexity bounds and margin bounds encourage minimum parameters (minimum complexity)

A Posteriori evaluation:

- Tighter bounds but still too high for deep NN to provide efficient assurance level regarding average loss
- For small NN with large volume of data some bounds are providing tight results
- 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 – statistical guarantee

A Priori evaluation:

- Pessimistic as the theory remain valid in worst case and are vacuous for over-parametrized NN
- Pac Bayes bounds, complexity bounds and margin bounds encourage minimum parameters (minimum complexity)

A Posteriori evaluation:

- Tighter bounds but still too high for deep NN to provide efficient assurance level regarding average loss
- For small NN with large volume of data some bounds are providing tight results
- Naive application of the bounds do not provide accurate and self-sufficient means to guarantee the generalizability of the used models.

		BOUND			
		001	003	004	006
AVI dents	A priori evaluation	219999	2250	28372	1609964
	A posteriori evaluation	4642	52	15	468
ATC STT	A priori evaluation	810	4.10^7	1.106	3.10^9
	A posteriori evaluation	7	2255	90	16425
ACAS Xu	A priori evaluation	0.9	1.2	0.11	0.02
	A posteriori evaluation	0.1	0.014	0.06	0.008

		Fmnist ref		Fmnist Improved	
		A priori evaluation	A posteriori evaluation	A priori evaluation	A posteriori evaluation
BOUND	001	172	11	20,2	6,4
	002		2,56		1,6
	003	55	14,4	4,4	0,8
	004	7	1,8	3,9	1,3
	006	1664	6,7	31	3,6
	007	9	9	4,4	4,4
	010	11	11	11,4	8,8
	011	1,8	1,62	1,8	0,54
	012	10	10	3,6	3,6
Loss	Train		0,14		0,24
	Test		0,23		0,29
Acc %	Test		91		89

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

ATC-STT – Models evaluation

Targeted task: 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

Evaluation metric:

Word Error Rate (WER)

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

ATC-STT – Models evaluation

Results interpretation of the PoC:

Excellent performances of the AIRBUS model on the AIRBUS dataset and poor performances on open-source datasets.

Possible overfitting due to:

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

Pipeline analysis:

Model selection: real time constraints VS performance

Dataset representativity regarding the ODD

Optimization adaptation

Model finetuning

Model	Approach	Source	Training Dataset
AIRBUS	KALDI		AIRBUS dataset
DL 1	Transformers	HuggingFace	UWB and ATCOSIM
DL 2	Transformers	HuggingFace	UWB
DL 3	Transformers	HuggingFace	UWB and ATCOSIM
FT 3.1	Transformers	Finetuned DL 3 during 10 epochs	UWB, ATCOSIM and AIRBUS dataset
FT 3.2	Transformers	Finetuned DL 3 during 50 epochs	UWB, ATCOSIM and AIRBUS dataset
DL 4	Transformers	HuggingFace	UWB
FT 4	Transformers	Finetuned DL 4 during 50 epochs	UWB and AIRBUS dataset

Metric	Dataset / Model	ATCOSIM_de1	ATCOSIM_de2	ATCOSIM_fr1	ATCOSIM_fr2	ATCO2_cz1	ATCO2_cz2
MER	AIRBUS	93.77 %	93.11 %	91.25 %	91.32 %	88.97 %	100 %
	DL 1	10.14 %	5.14 %	5.96 %	6.57 %	31.97 %	31.18 %
	DL 2	20.91 %	27.34 %	27.17 %	26.59 %	24.97 %	29.18 %
	DL 3	6.48 %	5.27 %	6.73 %	6.78 %	26.63 %	26.99 %
	FT 3.1	22.97 %	27.53 %	26.21 %	25.58 %	42.13 %	41.71 %
	FT 3.2	12.82 %	15.23 %	16.96 %	16.92 %	29.71 %	29.65 %
	DL 4	15.54 %	21.72 %	24.63 %	24.83 %	24.47 %	28.55 %
	FT 4	17.31 %	20.79 %	20.51 %	21.33 %	28.06 %	27.61 %
WIL	AIRBUS	95.69 %	95.91 %	94.7 %	94.73 %	92.24 %	100 %
	DL 1	13.11 %	6.82 %	7.68 %	9.26 %	46.36 %	45.24 %
	DL 2	29.52 %	38.81 %	38.84 %	37.38 %	35.52 %	41.25 %
	DL 3	9.09 %	7.01 %	8.81 %	9.63 %	37.21 %	37.01 %
	FT 3.1	36.66 %	43.57 %	40.11 %	39.31 %	60.26 %	58.81 %
	FT 3.2	20.24 %	24.98 %	25.83 %	25.64 %	42.74 %	41.28 %
	DL 4	23.41 %	31.5 %	36.58 %	35 %	34.63 %	39.68 %
	FT 4	26.90 %	32.98 %	30.46 %	31.28 %	39.87 %	39.18 %
WIP	AIRBUS	4.31 %	4.09 %	5.3 %	5.27 %	7.76 %	0 %
	DL 1	86.89 %	93.18 %	92.12 %	90.74 %	53.64 %	54.76 %
	DL 2	70.48 %	61.19 %	61.16 %	62.62 %	64.48 %	58.75 %
	DL 3	90.91 %	92.99 %	91.26 %	90.37 %	62.79 %	62.99 %
	FT 3.1	63.34 %	56.25 %	59.89 %	60.69 %	39.74 %	41.19 %
	FT 3.2	79.76 %	75.02 %	74.17 %	74.96 %	57.26 %	58.72 %
	DL 4	76.59 %	68.5 %	63.42 %	65 %	63.37 %	60.32 %
	FT 4	73.01 %	67.02 %	69.54 %	68.72 %	60.13 %	60.82 %
WER	AIRBUS	97 %	93.61 %	94 %	95.59 %	89.18 %	100 %
	DL 1	10.54 %	6.41 %	6.32 %	6.84 %	32.61 %	31.67 %
	DL 2	21.4 %	33.37 %	28.09 %	27.02 %	25.43 %	29.5 %
	DL 3	6.86 %	6.31 %	7.04 %	7.05 %	27.16 %	27.26 %
	FT 3.1	23.44 %	28.35 %	26.35 %	25.73 %	42.99 %	42.44 %
	FT 3.2	12.95 %	16.04 %	17.29 %	17.14 %	30.43 %	30.16 %
	DL 4	15.96 %	23.2 %	25.56 %	25.17 %	24.93 %	28.9 %
	FT 4	17.37 %	22.18 %	20.73 %	21.55 %	28.68 %	27.99 %

Metric	Dataset / Model	ATCO2_ch1	ATCO2_ch2	ATCO2_ch3	ATCO2_nl	ATCO2_aus	UWB	AIRBUS_test
MER	AIRBUS	99.28 %	60.85 %	99.04 %	94.92 %	93.29 %	62.59 %	10.28 %
	DL 1	57.14 %	46.31 %	39.38 %	42.61 %	60.04 %	18.73 %	42.35 %
	DL 2	52.7 %	38.99 %	32.42 %	36.2 %	57.22 %	16.46 %	36.51 %
	DL 3	55.14 %	27.16 %	28.83 %	33.55 %	51.03 %	18.58 %	33.23 %
	FT 3.1	60.49 %	42.5 %	40.36 %	46.79 %	57.82 %	36.38 %	25.3 %
	FT 3.2	46.6 %	29.29 %	26.12 %	34.61 %	47.51 %	28.34 %	14.13 %
	DL 4	56.22 %	31.69 %	30.44 %	34.41 %	48.9 %	17.87 %	33.32 %
	FT 4	47.83 %	30.31 %	25.83 %	33.98 %	46.93 %	25.38 %	13.67 %
WIL	AIRBUS	99.64 %	71.35 %	99.37 %	96.39 %	95.71 %	73.57 %	13.73 %
	DL 1	73.76 %	63.99 %	56.64 %	59.94 %	77.99 %	26.47 %	58.26 %
	DL 2	67.12 %	54.87 %	47.01 %	51.11 %	74.41 %	22.67 %	49.88 %
	DL 3	68.18 %	38.77 %	42.18 %	47.71 %	66.61 %	26.27 %	45.49 %
	FT 3.1	74.81 %	60.56 %	58.81 %	66.11 %	74.2 %	53.41 %	38.76 %
	FT 3.2	61.4 %	42.13 %	39.24 %	49.71 %	63.31 %	41.77 %	20.32 %
	DL 4	68.92 %	44.94 %	43.27 %	48.64 %	64.02 %	35.19 %	45.47 %
	FT 4	61.81 %	44.19 %	38.54 %	47.58 %	62.66 %	36.82 %	19.48 %
WIP	AIRBUS	0.56 %	28.65 %	0.63 %	3.61 %	4.29 %	26.43 %	80.27 %
	DL 1	26.24 %	36.01 %	43.36 %	40.06 %	22.01 %	73.53 %	41.74 %
	DL 2	32.88 %	45.13 %	52.99 %	48.89 %	25.59 %	77.33 %	50.12 %
	DL 3	31.82 %	61.22 %	57.82 %	52.29 %	33.39 %	73.73 %	54.51 %
	FT 3.1	25.19 %	39.44 %	41.19 %	33.89 %	25.8 %	46.59 %	61.24 %
	FT 3.2	38.6 %	57.87 %	60.76 %	50.29 %	36.69 %	58.23 %	79.68 %
	DL 4	31.08 %	55.06 %	56.73 %	51.36 %	35.98 %	74.81 %	54.53 %
	FT 4	38.19 %	55.81 %	61.46 %	52.42 %	37.34 %	63.18 %	80.52 %
WER	AIRBUS	99.28 %	61.38 %	99.04 %	95.05 %	93.4 %	63.46 %	11.86 %
	DL 1	58.3 %	48.73 %	42.14 %	44.26 %	61.07 %	23.28 %	44.8 %
	DL 2	53.59 %	40.33 %	31.75 %	37.53 %	57.94 %	18.98 %	38.18 %
	DL 3	55.02 %	27.32 %	30.31 %	34.63 %	51.4 %	20.46 %	34.82 %
	FT 3.1	61.03 %	43.7 %	41.16 %	48.42 %	58.21 %	38.85 %	25.61 %
	FT 3.2	47.42 %	30.4 %	27.39 %	36.19 %	48.67 %	30.96 %	13.13 %
	DL 4	56.66 %	32.55 %	32.04 %	35.45 %	49.42 %	19.69 %	35.21 %
	FT 4	49.13 %	31.67 %	27.12 %	35.21 %	48.97 %	27.73 %	14.76 %

Dataset / Model		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.

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

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

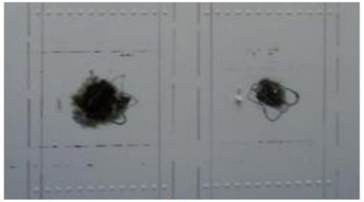
Datasets: AIRBUS dataset (pictures of surface damages detected and classified for lightning strikes and dents)

Models: YOLOv5 fine tuned model to minimize errors:

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

Evaluation metric: IoU (intersection over union)

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

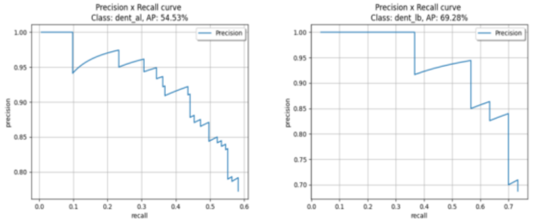
AIRBUS

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

AVI – Models evaluation

Results interpretation of the PoC

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 <u>95% accuracy</u>.



(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.

Pipeline analysis and experimentations:

Limited amount of data -> Data augmentation with simulated data

Model architecture influence YOLO V5 vs v8

Model finetuning

Metric	Model	Dents (1044 images, 316 labels)	Lightning strikes (6 images, 13 labels)
Precision %	Yolov5s	69.4	69.9
	Yolov8s	86.3	**98.9**
	Yolov8m	85.9	39.8
	Yolov8l	**88.5**	<u>90.1</u>
Recall %	Yolov5s	64.3	**50**
	Yolov8s	**84.9**	38.5
	Yolov8m	82.1	46.2
	Yolov8l	79.7	15.4
mAP@50 %	Yolov5s	64.4	**54.5**
	Yolov8s	**89.2**	44.8
	Yolov8m	88.6	26.8
	Yolov8l	86.6	28.3

Table 50: Performance's comparison of different Yolo architectures, trained in original and augmented datasets for AVI use case. The performances are % values of three main measures: precision, recall and mAP@50.

Metric	Model	Lightning strikes (6 images, 13 labels)
Precision %	Yolov5s	**69.9**
	Yolov5s finetuned on augmented data (100 epochs)	54
Recall %	Yolov5s	**50**
	Yolov5s finetuned on augmented data (100 epochs)	46.2
mAP@50 %	Yolov5s	**54.5**
	Yolov5s finetuned on augmented data (100 epochs)	39.9

Table 51. Comparison of the YoloV5 model trained in original data and the one trained in augmented data.

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

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 - FCNN 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

Neural nets ACAS Xu steps

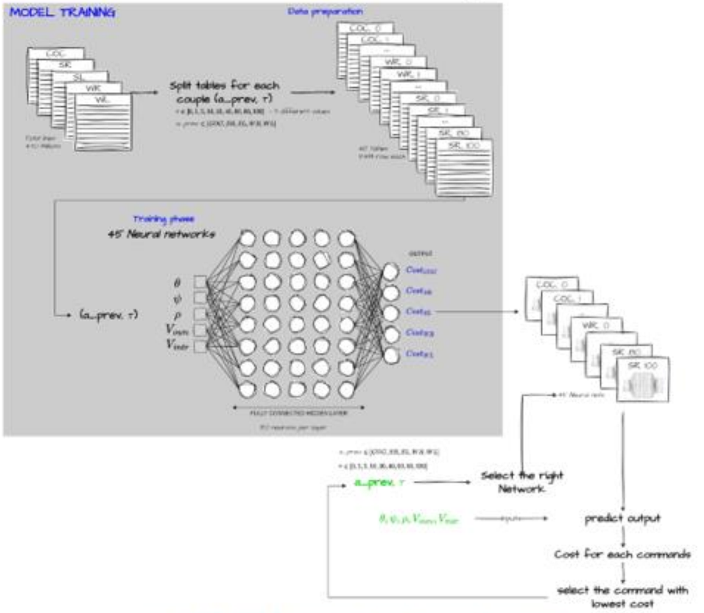


Figure 122: ACAS Xu neural network approach illustration

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

ACAS Xu Task – Models evaluation

Results interpretation

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

Pipeline analysis:

- Model architecture adapted for classification task
- Unbalanced dataset: data augmentation / Weighted loss function
  - *The positive effect could have been on the training error, which was already small. So, finally, it is difficult to conclude whether both approaches have a positive influence on generalisation. The benefits should be more focused on the stability and robustness of the models.*

		Reference		w/ data augmentation		w/ weighted loss function	
		A priori evaluation	A posteriori evaluation	A priori evaluation	A posteriori evaluation	A priori evaluation	A posteriori evaluation
BOUND	001	41,9	2,2	41,9	5,2	41,9	2,5
	002		1,6		1,6		1,6
	003	1,23	0,014	1,23	0,014	1,23	0,014
	004	0,17	0,06	0,17	0,06	0,17	0,06
	006	0,06	0,008	0,06	0,008	0,06	0,008
	007	2,5	2,5	2,5	2,5	2,5	2,5
	010	8	3,6	8	3,6	8	3,6
	011	0,6	0,05	0,6	0,05	0,6	0,05
	012	3,6	3,6	3,6	3,6	3,6	3,6

Table 52. Generalisation bounds comparison for ACAS Xu use case with data augmentation or weighted loss function

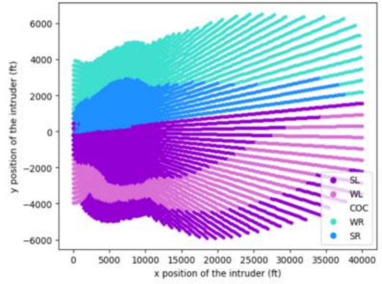
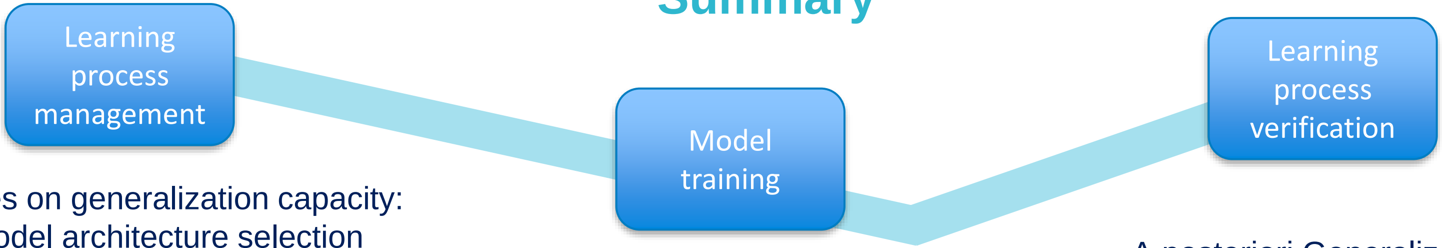


Figure 68 Instruction of the neural network $N_{0,COC}$ as w.r.t: the position of the intruder, the ownship being at (0,0) and moving along the x axis. With ($\psi = -\pi$) at and $v_{own} = v_{int} = 430$ ft/s.

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

Summary



Influences on generalization capacity:

- Model architecture selection
- Metrics selection
- Hyper parameters selection
- Volume of training data

A priori generalization bounds

Learning curves analysis
Bias and variance
Convergence stability

A posteriori Generalization bounds (trained model)
Performance on test dataset
Empirical gap measurement

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

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

Main takeways

Learning process management

Model training

Learning process verification

Generalization bounds (LM-04)

- For deep NN, difficult to use the theory to compare and select architecture
- For small network with large volume of data we have tight statistical guarantees

Methods to boost generalization and provide confidence

- Regularization
- Penalty methods
- Data expansion

Learning curves (LM-07)

- For deep NN, it is a key indicator to secure proper optimization
- Convergence

Training objective and Evaluation metrics

- Alignment between loss function selection and targeted application
- Representative of the targeted performance

Generalization bounds on trained model (LM-04)

- For deep NN, gap concerned by statistical guarantees are too big
- For small NN with large volume of data, small gap => learning assurance process

Performance on test data (LM-09)

- Test dataset volume and distribution
- Train dataset quality

Comparison (LM-14)

- Empirical gap measurement
- Issues detection

Q&A

www.sli.do

#AIDays

Passcode: hmkota

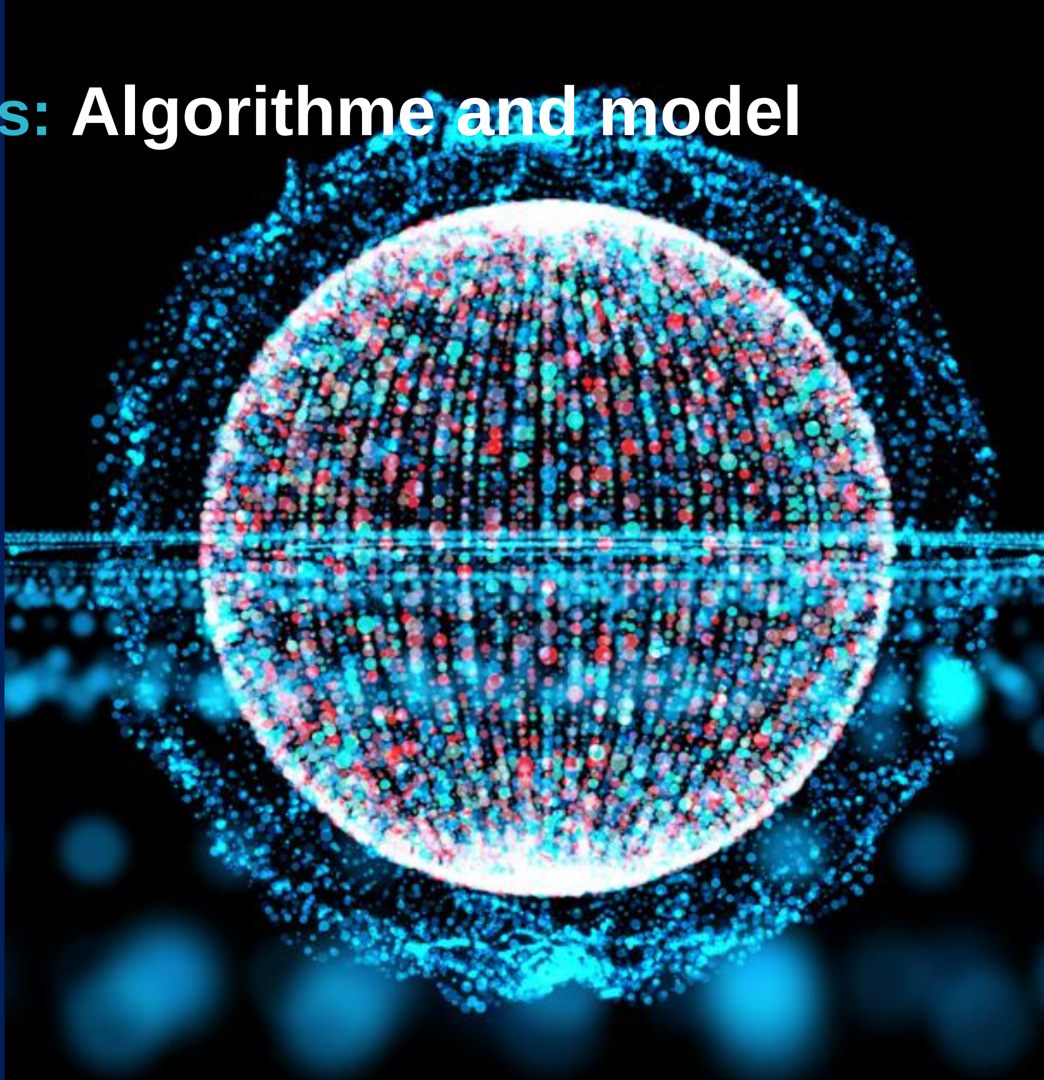
/ Presentation of the outcome and recommendations of Task 3

AIRBUS

MLEAP – Task #3 milestones: 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 Milestones: Algorithm and model robustness > > >

Why talking about robustness?

One of the key requirement from the HLEG

>

One of the key objective in the AI Act

>

Because it is one of the key issue with AI!

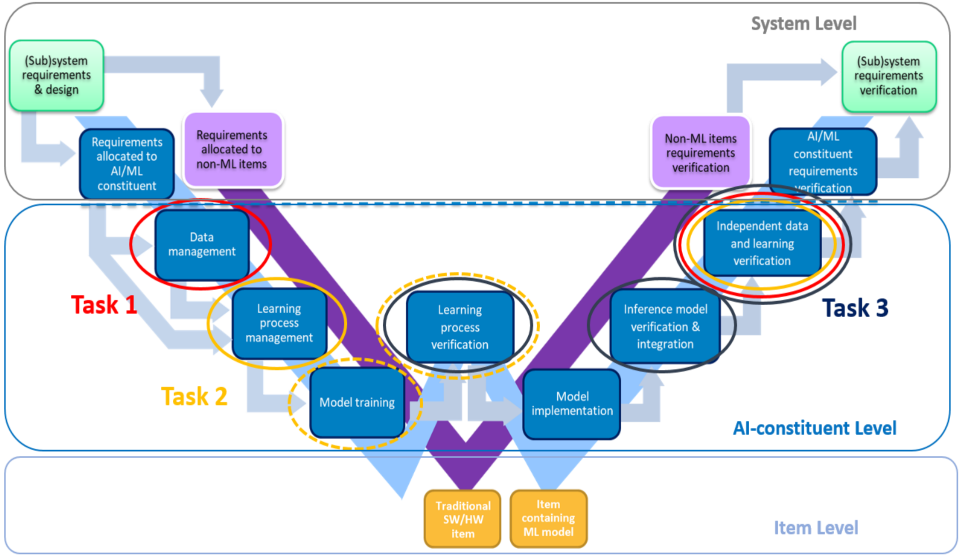
MLEAP – Task #3 Milestones: Algorithm and model robustness > > >

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



MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Why talking about robustness?

Robustness means keeping the performances on the domain of ODD

ODD in an open world can be challenging

Nominal case

Variation of nominal case

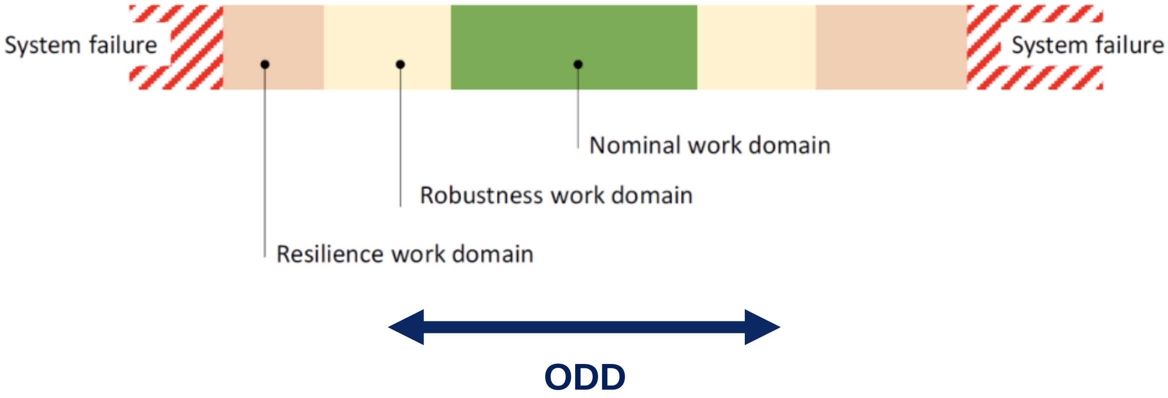
Adversarial case

A non-existent case

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Robustness assessment approaches

How to ensure that the system still works when it should?
Three types of approaches : statistical, formal, empirical



Picture from "DEEL White Paper on Machine learning in Certified System (DEEL Certification Workgroup, 2021")

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Different ways of defining the concept

Aligning several sources of the state of the art

- Different concepts robustness, stability, corner cases…
- Different requirements
- Different methods: statistical, formal, empirical

Studying the maturity of the ecosystem

- Scalability of the methods
- Applicability to the relevant use-cases

Preparing the application on the use case

Harmonized state of the art

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Common properties to assess

Stability (of the training algorithm, trained model and inference model)	$\|x' - x\| < \delta \Rightarrow \|\hat{f}(x') - \hat{f}(x)\| < \varepsilon$
Bias (~ underfitting)	$bias^2(\mathcal{F}, n) = \mathbb{E}_{x \sim \mathcal{X}}\left[(\overline{f_n}(x) - f(x))^2\right]$
Variance (~ overfitting)	$var(\mathcal{F}, n, x) = \mathbb{E}_{D \sim \mathcal{X}^n}\left[\left(\hat{f}^{(D)} - \overline{f_n}(x)\right)^2\right]$
Relevance (~ explainability)	Acceptability of contribution of each dimension of the input vector
Reachability	$\mathcal{E}^n\left(x, \hat{f}^n(x)\right) \notin Z$

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

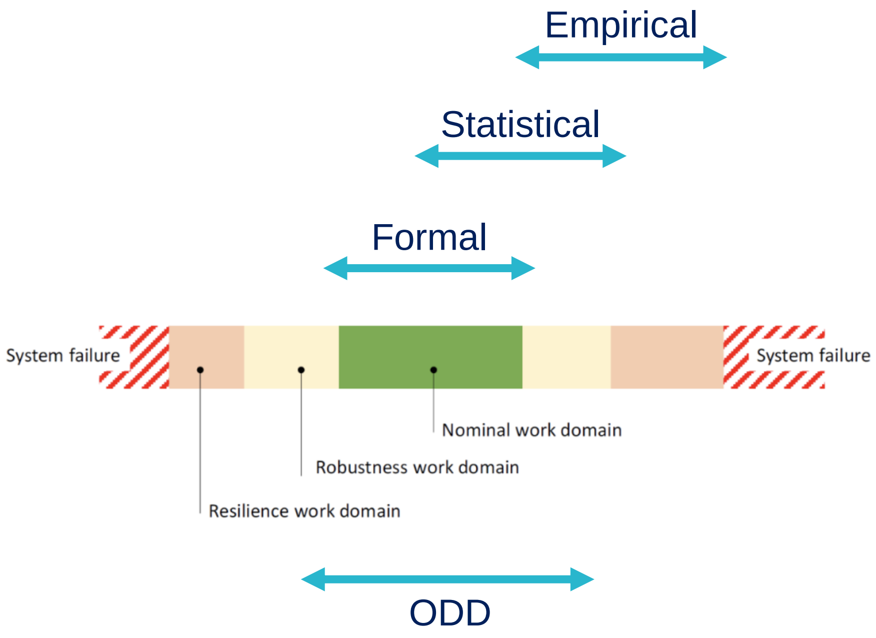
Complementarity of methods

Conceptual alignment is possible

- Stability around the nominal conditions
- Robustness to more difficult conditions
- Resilience to adverse conditions

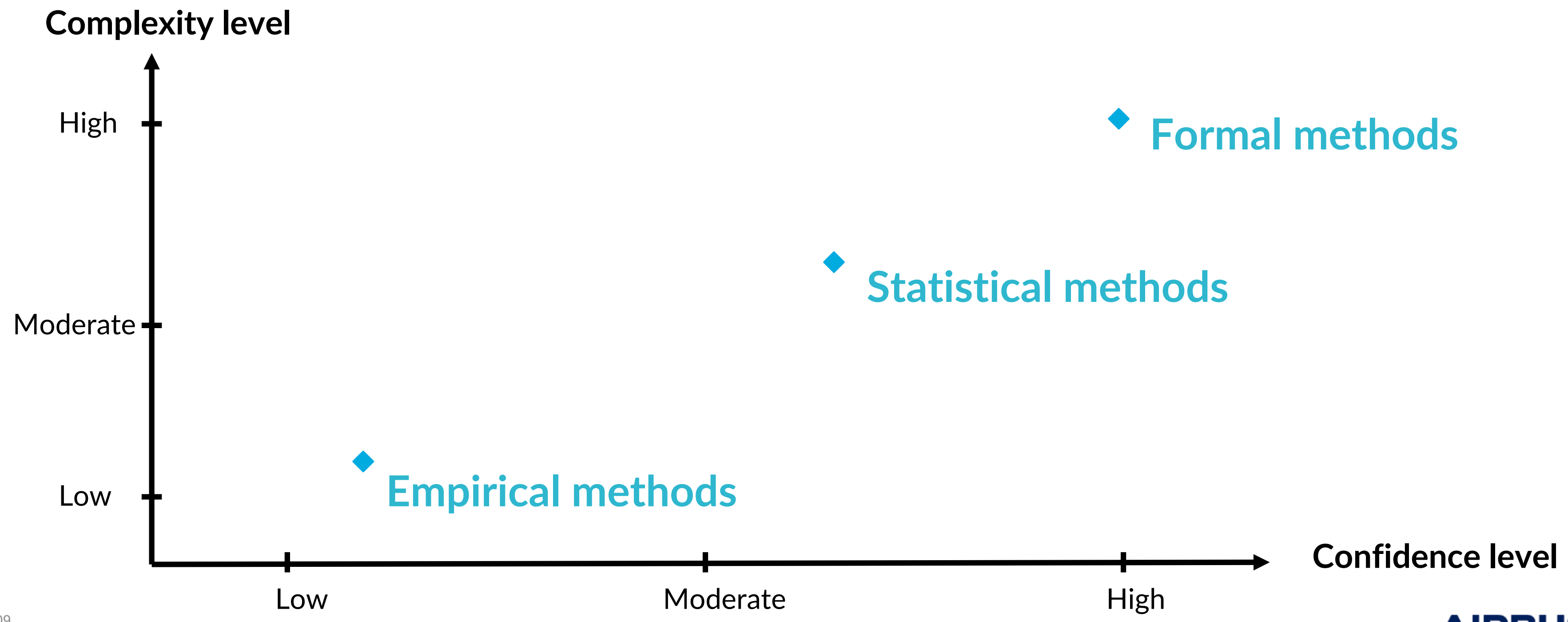
Methods are complementary

- Depends on the ODD description
- Combining approaches to match the requireme
- …but varying degree of scalability



MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Ease of use of methods



MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Corner case exploration

Different ways of exploring of the ODD

Different level to define corner case in the ODD (context: automotive)

- Scenario (several instants)
- Scene (one instant)
- Objects
- Domain (weather)
- Pixel (camera)

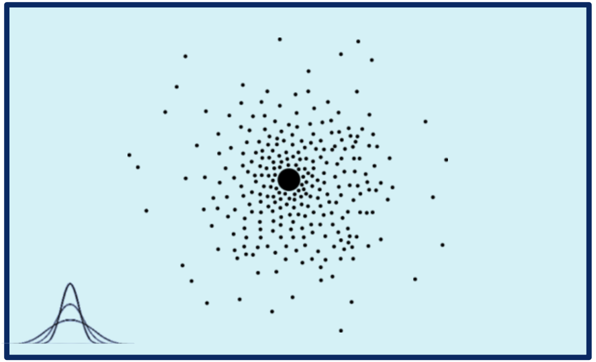
(From Heidecker et al., 2021)

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

3 approaches at a glance

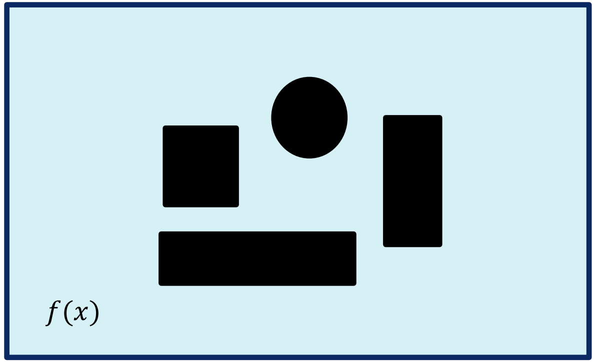
Each allow specific advantages and drawbacks

Statistical



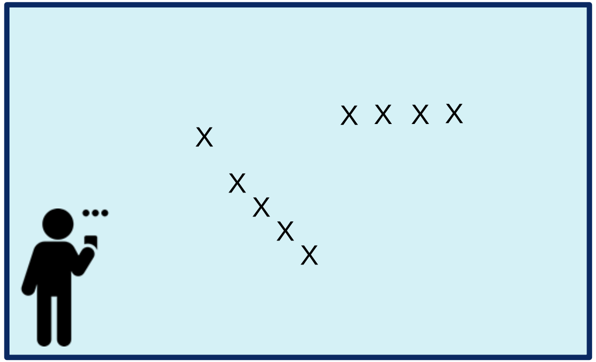
Easy to setup
Rely on data sets

Formal



Local guarantees
High dimensional sub-space

Empirical



Require human intervention
Experimental protocol

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Advantages and pitfalls

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

	Empirical methods	Statistical methods	Formal methods
Stability of the training algorithm	Not suitable	Suitable	Not suitable (training algorithm is still probably too large)
Stability of the trained model	Could be used but with limited confidence in the results	Suitable	Suitable
Stability of the inference model	Could be used but with limited trust in the results	Suitable	Suitable
Bias	Not really well suited	Suitable	Not really well suited
Variance	Not really well suited	Suitable	Not really well suited
Robustness (Corner case exploration)	Could be used for very specific catastrophic scenario	Suitable	Could be used in combination with statistical approach
Relevance	Expert judgment	Not suitable since it requires some form of symbolic analysis	Suitable in combination with empirical assessment
Reachability	Not suitable since it requires strong guarantees	Not suitable since it requires strong guarantees	Suitable

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Putting in practice

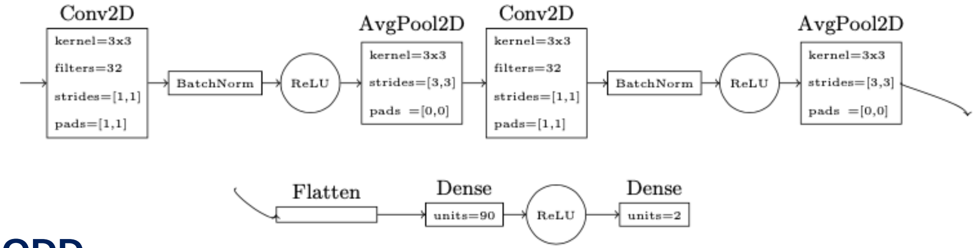
Example	Model type	Origin	Data type	Dimensionality	LM	Actions to test
Toy	Classifier	Aerospace	Images	Small	• LM11 • LM12 • LM13	Training stability General stability Stability against specific perturbations
	Detector	Public domain	Images	High	• LM12	General stability
	Classifier	Health care	Time series	Medium	• LM11 • LM12	Training stability General stability
Avionic	Detector	Avionic	Images	High	• LM11 • LM12 • LM13	Fine tuning stability General stability Stability against specific perturbations
	Speech to text	Avionic	Sounds	High	• LM12 • LM13	General stability Stability against specific perturbations
	Reachability	Avionic	Vector	Low	• LM12	General stability

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image classifier

Statistical assessment of performance

- 2 classes
- Confusion matrix >95% accuracy

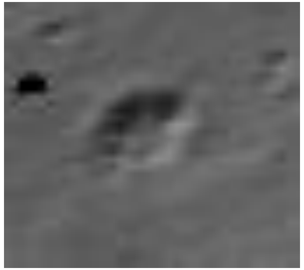


ODD

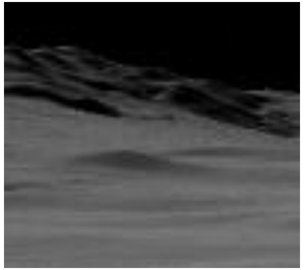
- Can be defined by experts
- But can still contained very unusual data points

Specific perturbations due to the space environment

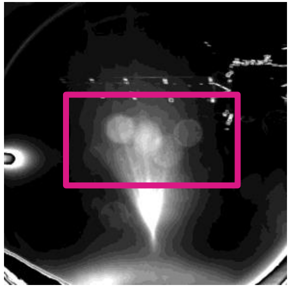
- Flares
- Radiation



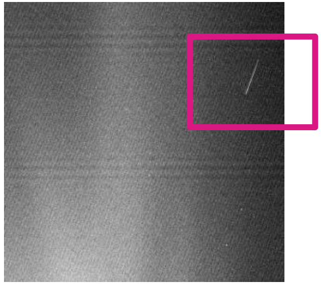
Crater



No crater



Flares

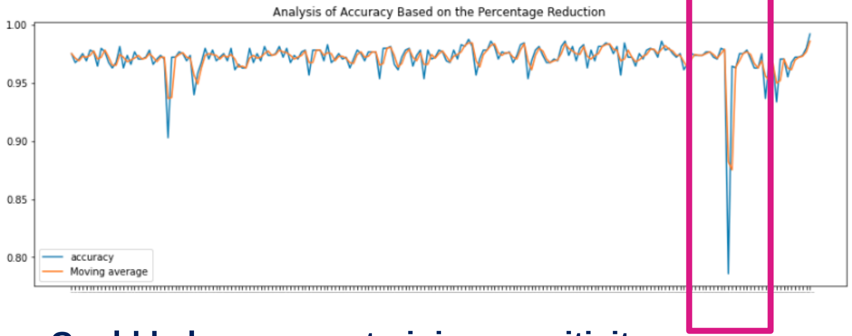


Radiation

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image classifier

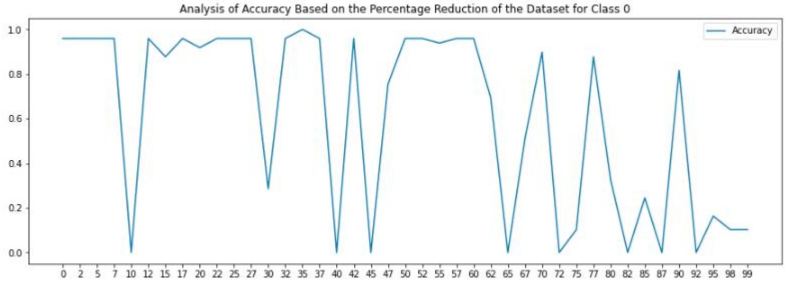
Training algorithm stability
- Take one training point out
- Retrain and revalidate accuracy



Could help measure training sensitivity
not really taken into account in the ecosystem

LM11

Training algorithm stability
- Taking part of the dataset out
- Retrain and revalidate accuracy



Could help measure the task inner difficulty
Link with Task 1 (datatset) tand Task 2 (generalization)

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image 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

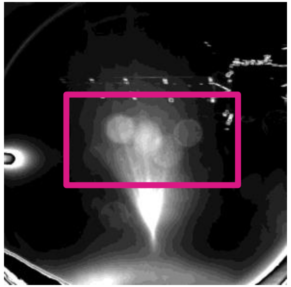
LM12

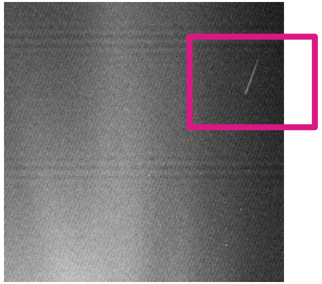
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

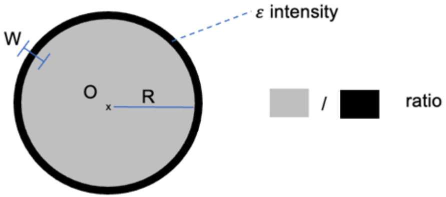
Image classifier

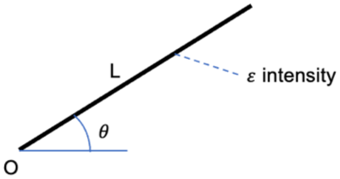
Stability against specific perturbations (related to the ODD)

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









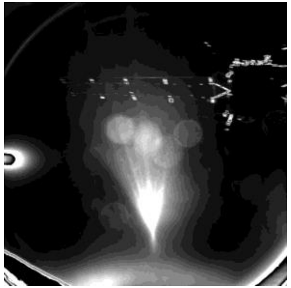
LM13

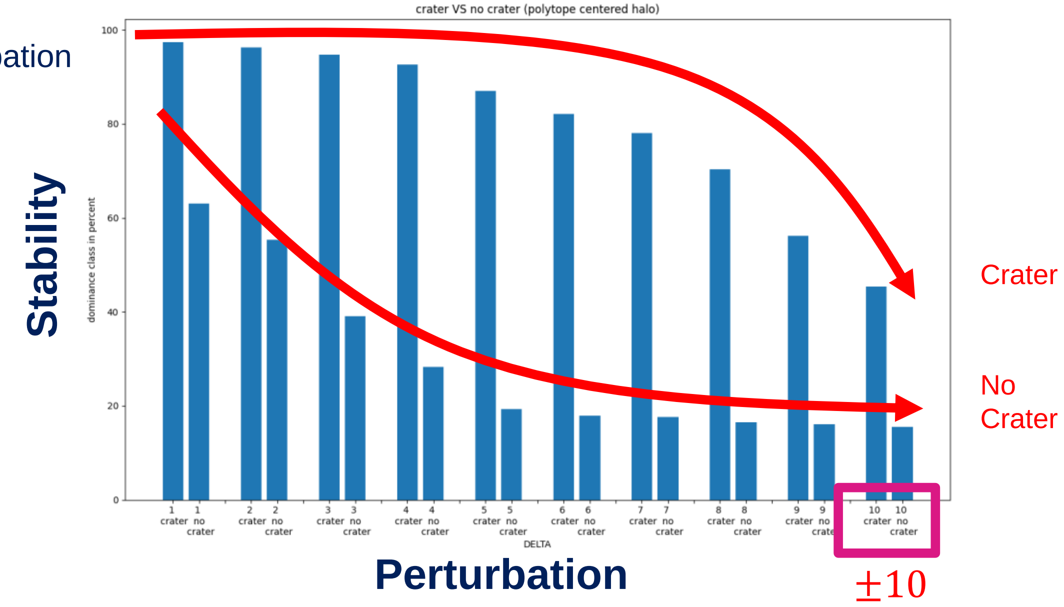
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image 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





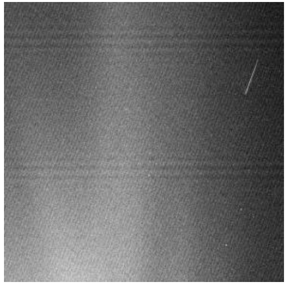
LM13

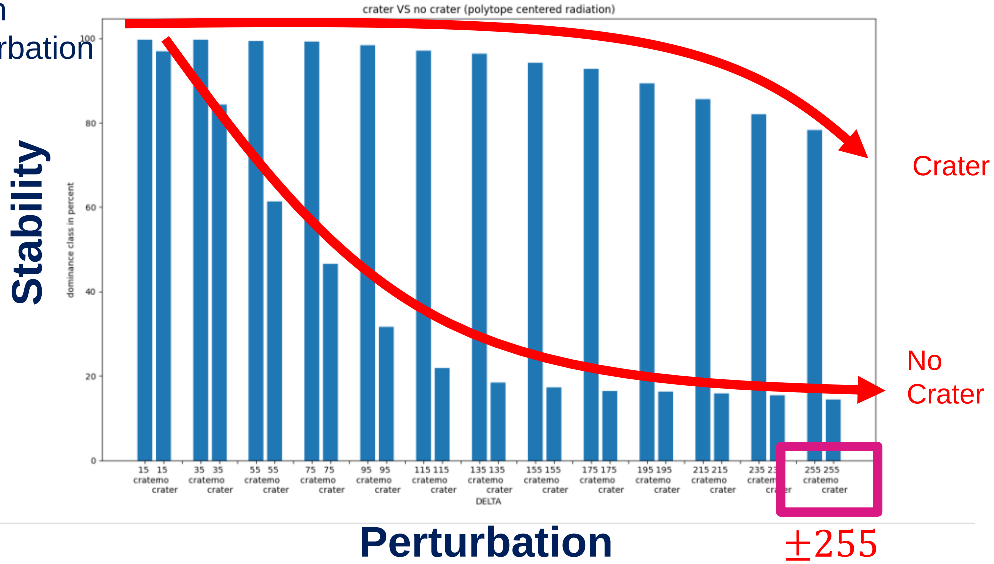
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image 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





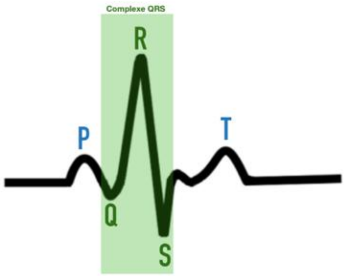
LM13

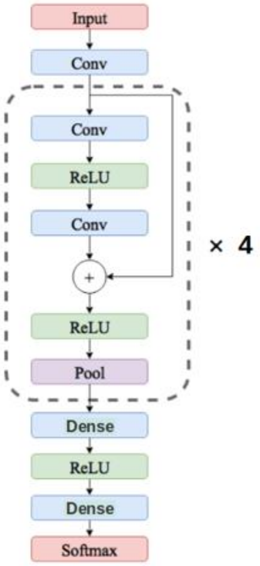
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Time series classifier

Dataset

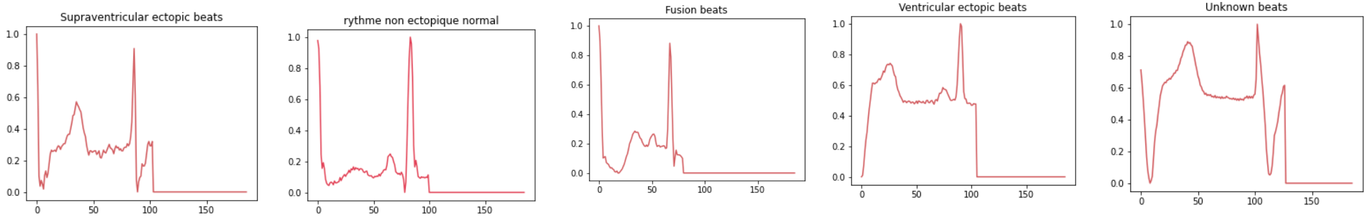
- 877K heart rhythm
- 188 instants each
- Class: 1 normal, 3 anormal, 1 unknown





ODD

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



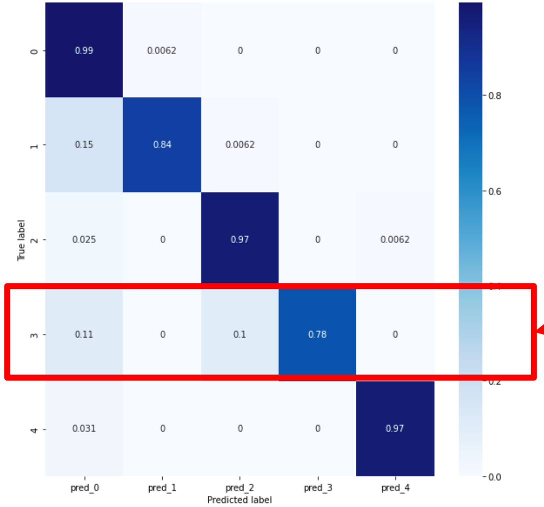
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Time series classifier

Statistical

- Confusion matrix
- Accuracy 95+%

Slight decrease in accuracy



LM12

Formal

- General stability

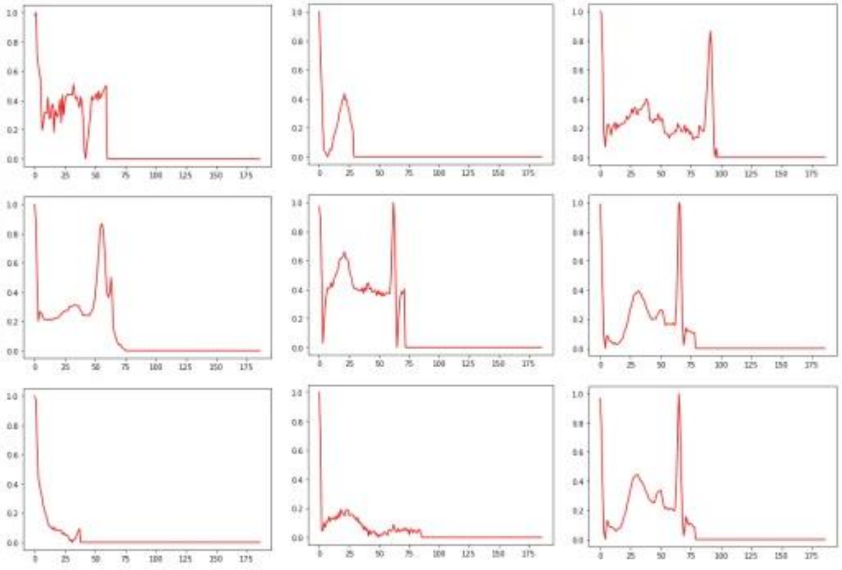
Unbalanced stability

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Time series classifier

Unbalanced stability

Wrong annotation



LM12

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

- Goal: improved maintenance
  - Finding dents
  - Finding lightning strikes

- Yolo v5 with SiLU or Leaky-ReLU activation

- Requirement tested
  - LM11
  - LM12
  - LM13

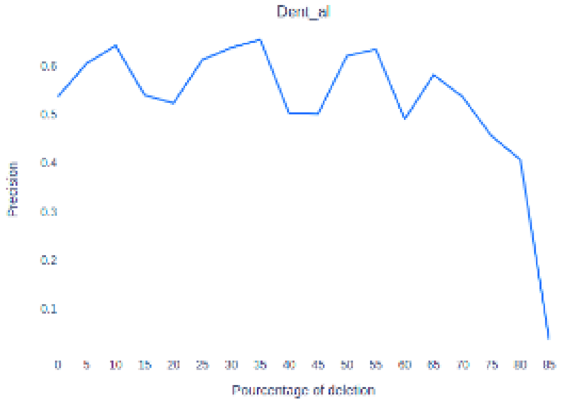
(credit PxHere)

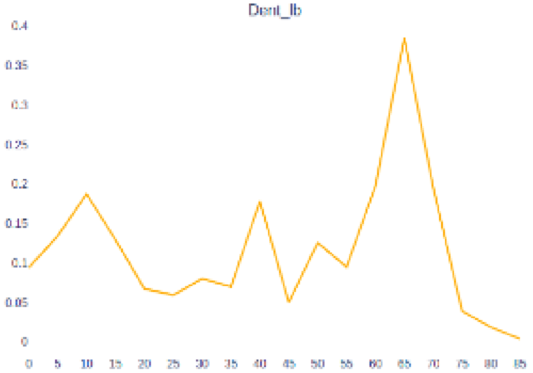
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

- **Reduce train data of finetuning**

- **For the "dent_al" class:**
  - **Accuracy remains stable until 75% of the training data is removed**
  - **Accuracy begins to decrease after 75%**
- **For the "dent_lb" class:**
  - **Accuracy remains constant on average (~0.1) until 55% of the training data is removed**
  - **Sudden increase after 55%, followed by a decrease similar to that of the "dent_al" class**





LM11

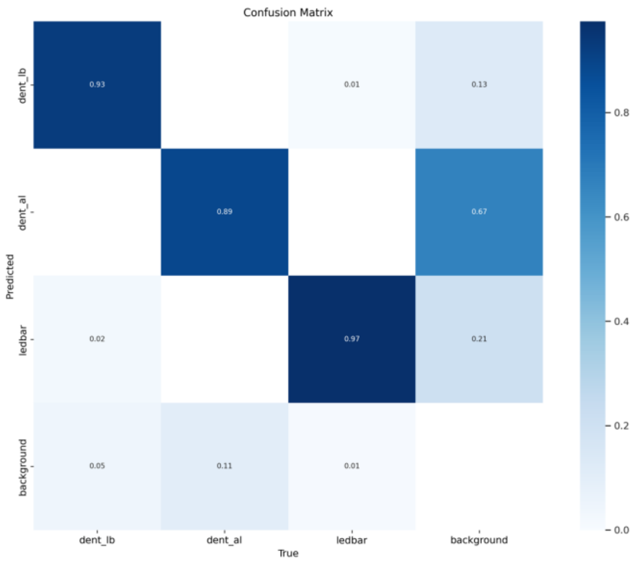
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

- AVI: LM12 Trained model stability
  - **SAIMPLE and statistical analysis**

Box number	class	Confidence	Objectness
1	Dent_al	[0.99727,0.99728]	[0.9296,0.9297]
2	Lebdar	[0.99739,0.99739]	[0.7836,0.7837]
3	Dent_al	[0.99462,0.99468]	[0.4477,0.4616]

LM12

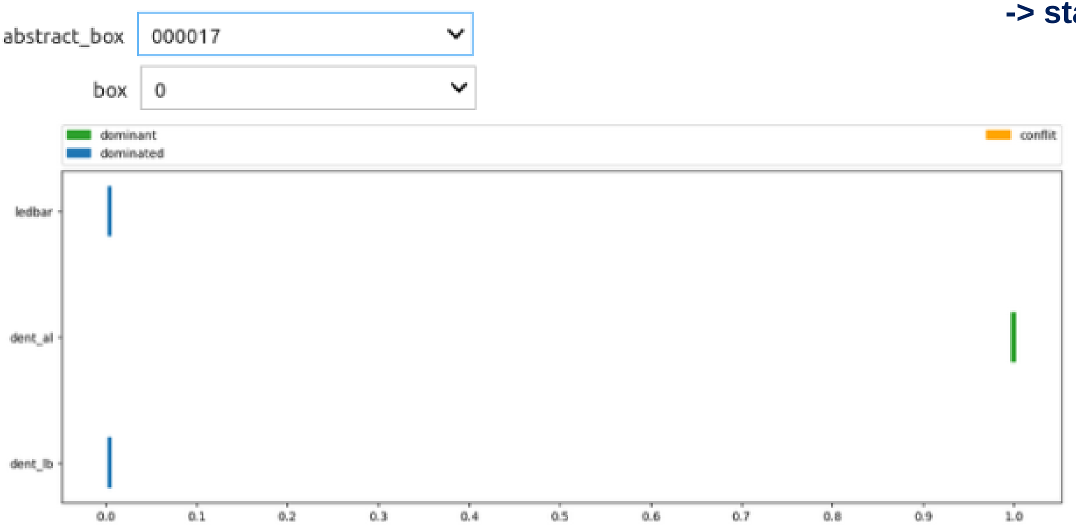


AIRBUS

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

- AVI: LM12 Trained model stability
  - SAIMPLE: Analysis of model stability



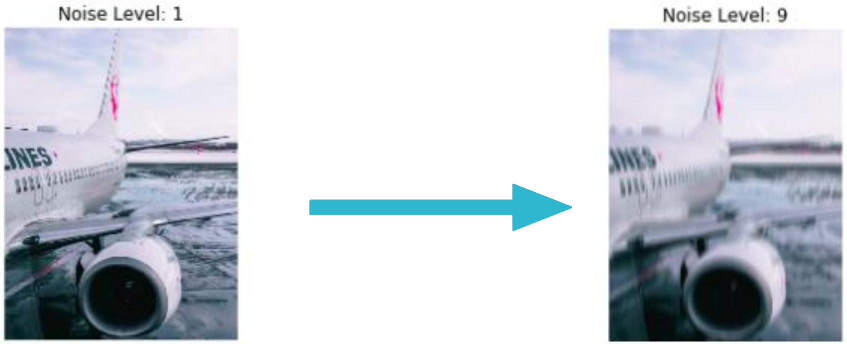
LM12

Box 0: good prediction, narrow interval length, and distant from other intervals -> stable prediction

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

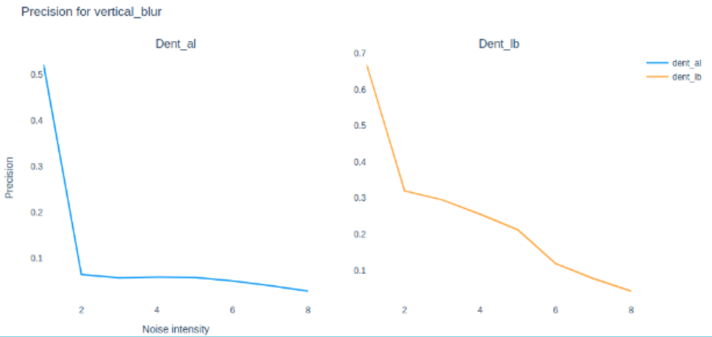
- AVI: LM13 Trained model robustness
  - **Analysis of the Yolov5-silu performance on different type of perturbation**
    - **Gaussian blur**
    - **Vertical blur**
    - **Horizontal blur**
    - **Brightness**
    - **…**

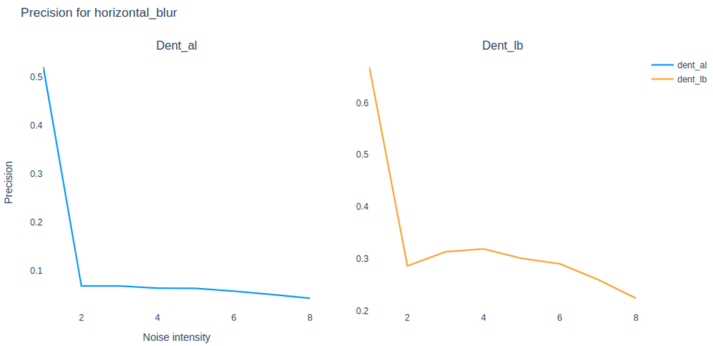


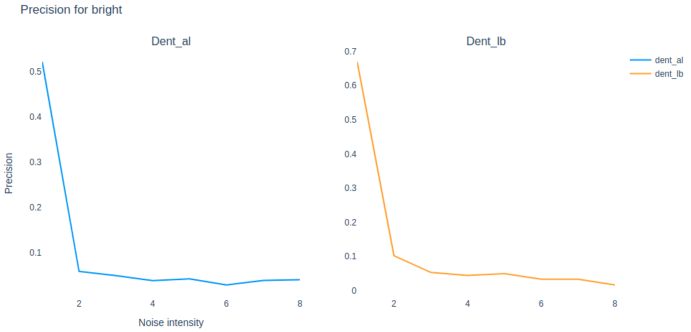
LM13

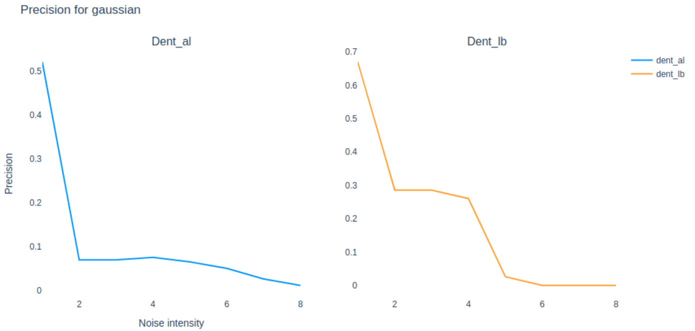
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors









LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Image detectors

- AVI: LM13 Trained model robustness
  - **Low robustness of the "dent_al" class to applied perturbations**
  - **Unlike the "dent_al" class, the "dent_lb" class also shows low robustness, although the performance drop is not as pronounced**
  - **A significant performance drop is observed for the "dent_al" class pointing to a high sensitivity to perturbations**
  - **Conversely, although the "dent_lb" class is not completely robust, it seems to withstand perturbations better than the "dent_al" class**

LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

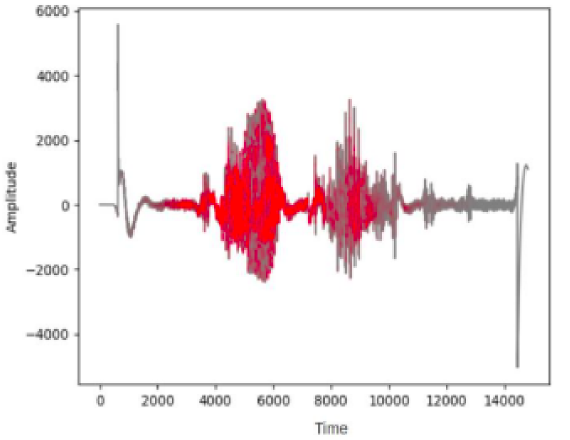
- Context: ATC communication
- Goal: improved communication processing
- Model:
  - Wav2Vec
  - Kaldi
- Requirement tested
  - LM12
  - LM13

(credit Kevin Blue)

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

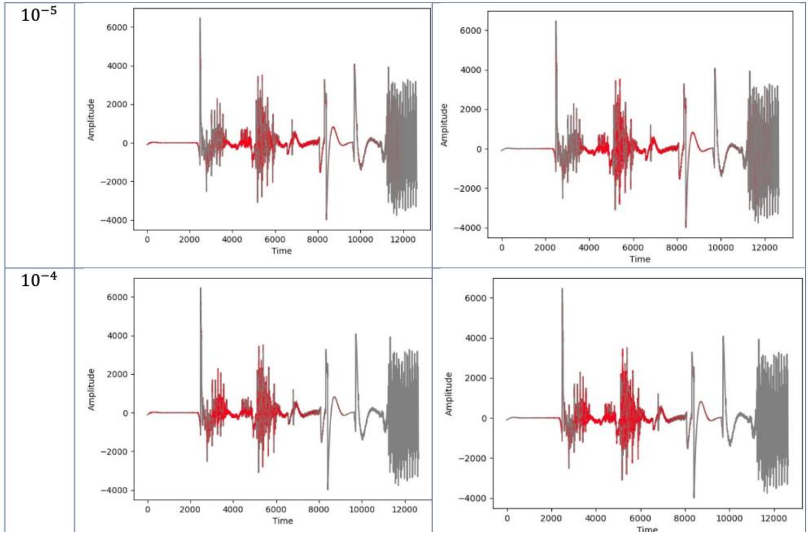
- STT: LM12 Trained model stability
  - Analysis Wave2vec

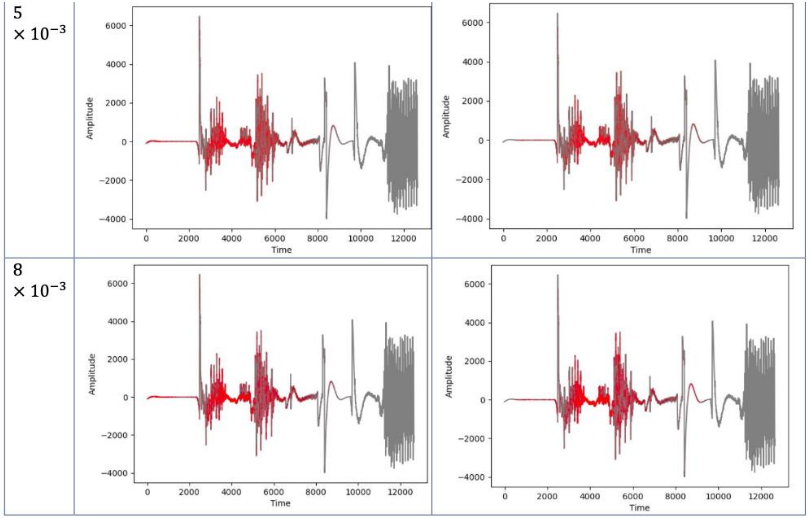
Re-transcription	Relevance
MINIMUM CLIMB SPEED	

LM12

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text



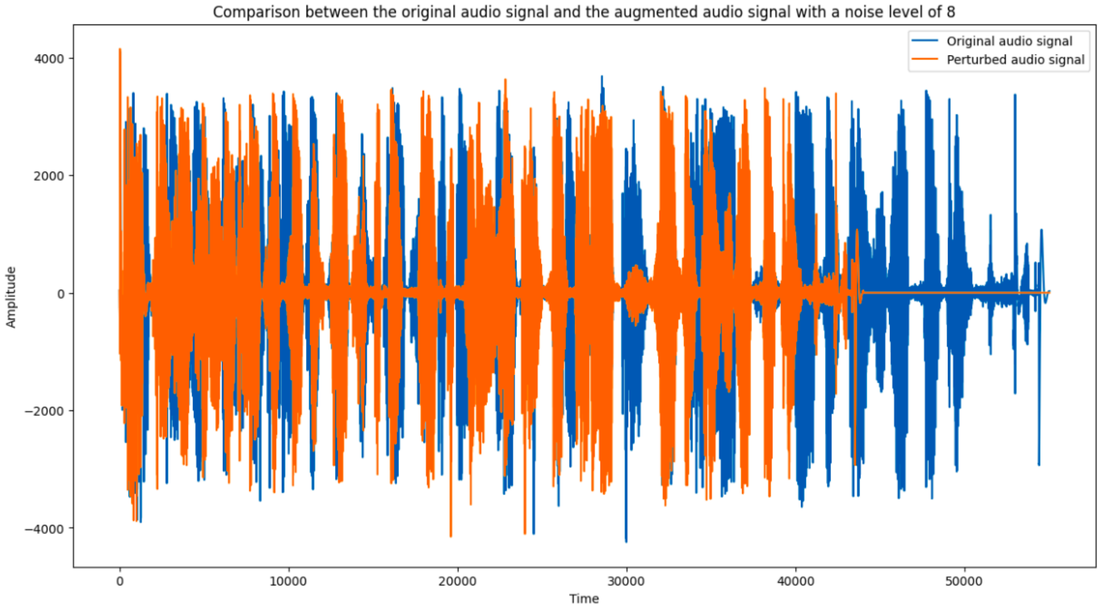


LM12

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

- STT: LM13 Trained model robustness
  - **Example of a perturbated recording under the speed perturbation (orange) from the original recording (blue).**

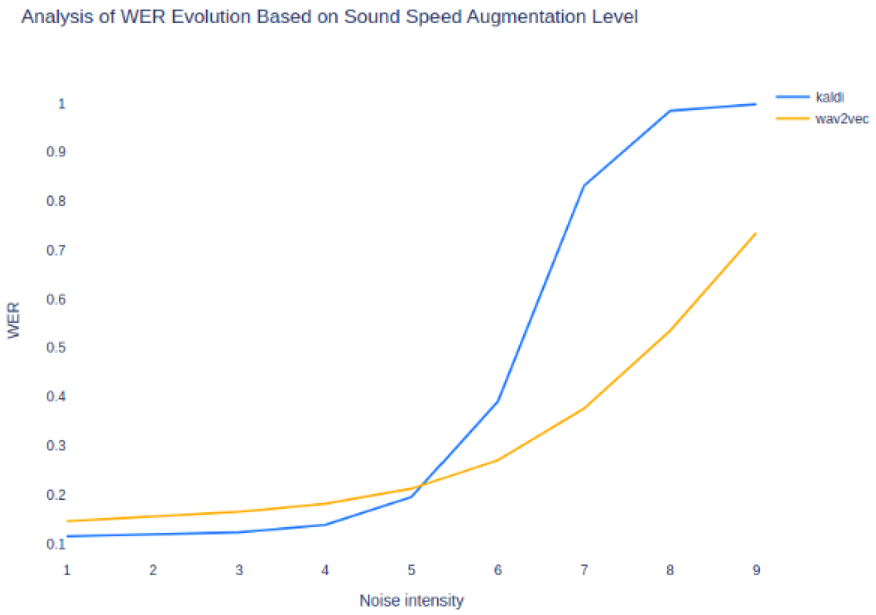


LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

- **Trained model robustness**
  - Evaluation against specific noise, such as speed rate, is insufficient to assess the model's robustness.
  - Given the use case nature, more particular perturbations should be considered to explore the ODD (Operational Design Domain) thoroughly.
  - More data points are required from external databases, which may also be biased.
  - A more empirical approach is needed to evaluate against such perturbations.
  - This type of validation is limited by subjectivity and may lack strong generalization properties over the ODD.

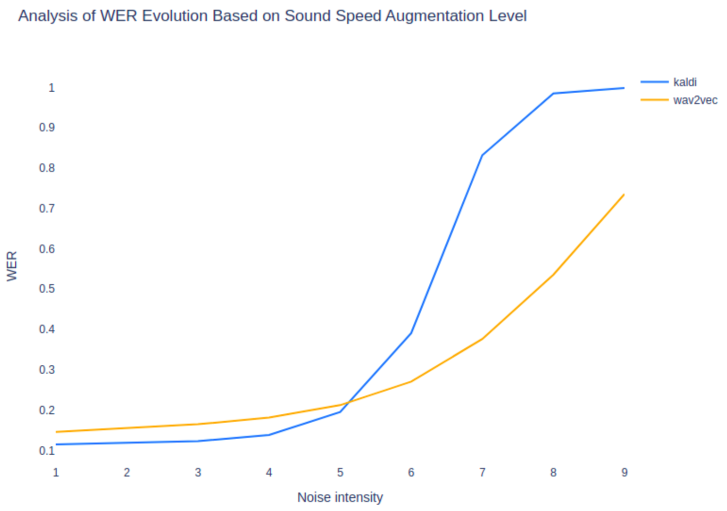


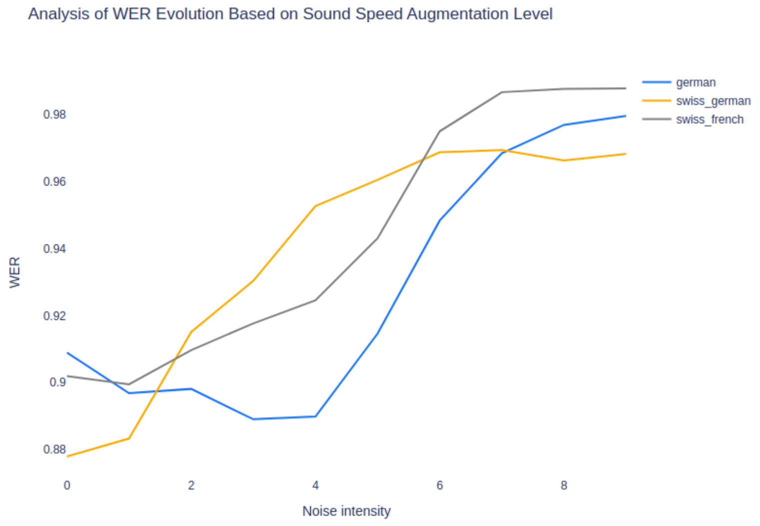
LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

- **STT: LM13 Trained model robustness**
  - Robustness to noise vs. Robustness to noise depending on the accent



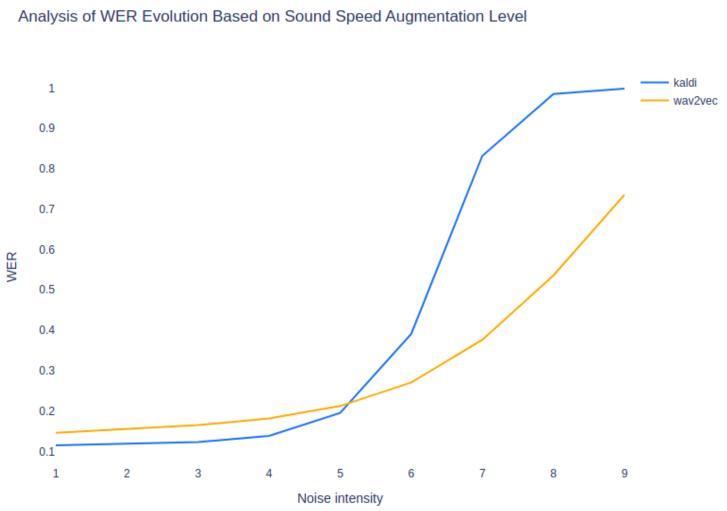


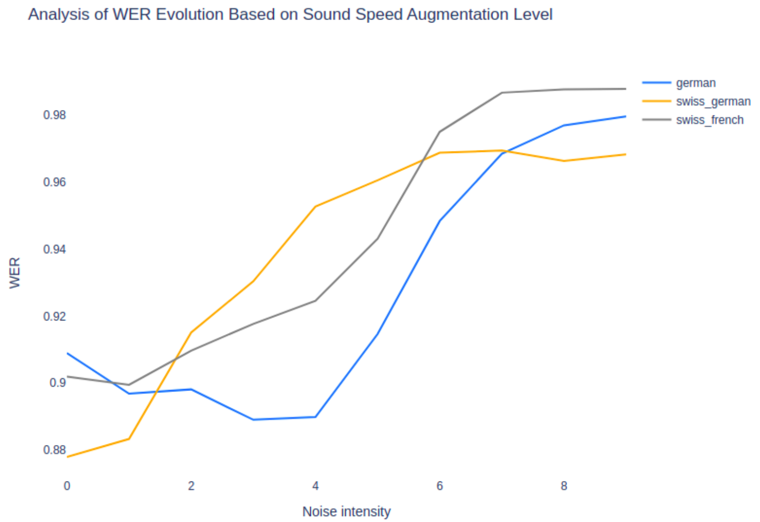
LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Speech to text

- STT: LM13 Trained model robustness
  - **Robustness to noise vs. Robustness to noise depending on the accent**





LM13

MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

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

www.sli.do

#AIDays

Passcode: hmkota

Q&A

MLEAP > > > Coffee break / 15H00 – 15H30

www.sli.do

#AIDays

Passcode: hmkota

/ General conclusions and recommendations from MLEAP consortium

Generic Pipeline > > > Way Forward

Purpose

Provide recommendations for each stage of AI development for critical aviation systems

Ensure **data** set **quality** (completeness, representativeness)
Assess, evaluate, and improve **generalisation**
Ensure **robustness** and **stability** of model performance

Methodology

- Mapping **MLEAP** project **tasks** to **W-shaped** development process stages
- Summarise **main issues** and discuss **strategies for ML/DL component development**
- Present **generic AI** development **pipeline** applicable to **various use cases**
- Provide way to **implement learning assurance** process with **requirements verification** for target applications

Generic Pipeline > > > Way Forward

Experimentation

Exploration of data-related and model-related **practices** to **enhance results**
Focus on **ways to minimise** the **gap** between **experimental** development and **industrial** objectives

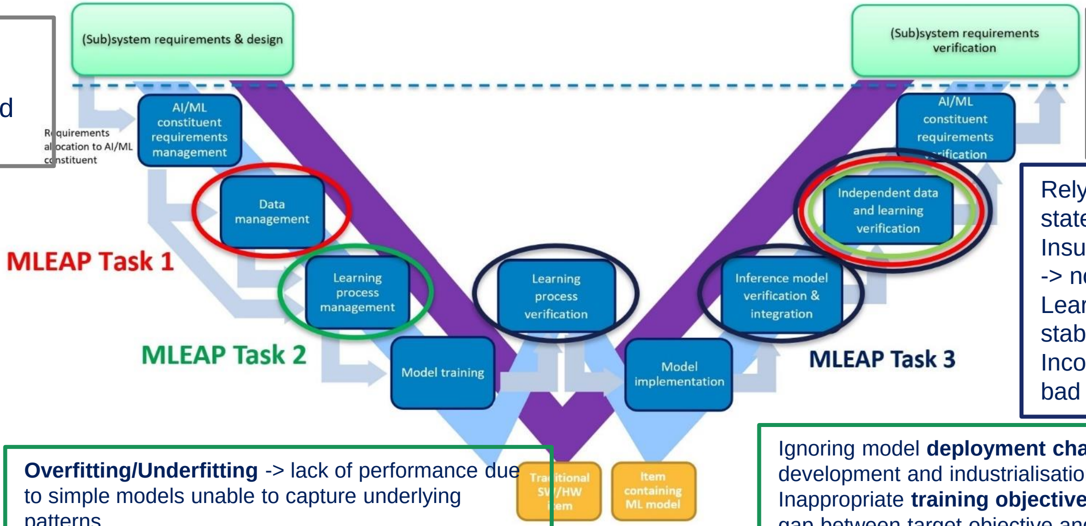
Conclusions

Focus on ways to **meet objectives of AI-based systems** development
Mapping of MLEAP **outcomes** regarding **data, models performances**, **to W-shaped** learning assurance
Methods and protocol **recommendation** to **meet the means of compliance**
Foreseen research **perspectives**

Generic Pipeline > > > weak Common Practices

Uncomplete/Unclear **ODD definition**, Inconsistency between **system-level** and **AI-level requirements**

Inadequate **data processing / representation** w.r.t AI-level requirements and intended use -> poor learning & non-relevant behaviour
Insufficient **data in training/testing** -> lack of coverage, underfitting, domain shift ...
Unhandled **outliers**, **non-standardised data** -> bad robustness, unstable model performance, …



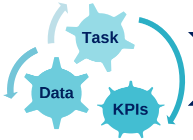
Overfitting/Underfitting -> lack of performance due to simple models unable to capture underlying patterns
Mis-use/understanding of **generalization bounds** -> misleading for model design and evaluation
Poor **hyperparameters tuning** -> poor generalization and handling data features

AI-level requirements can be met, but with inconsistency with system-level -> impact on safety

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

Ignoring model **deployment challenges** -> gap between experimental development and industrialisation purposes
Inappropriate **training objective**, inappropriate **evaluation measures** -> gap between target objective and model performances
Inappropriate **model capacity** vs **task complexity**, non adapted **optimisation** & **regularisation** -> weak learner & bad performances

Generic Pipeline > > > Practices Recommendation



Modeling, Training, Evaluation/Testing, Adjusting, re-Training, Validating, Preparing for release (deployment), Implementation, Behaviour Analysis, Testing, Validating.

Good model

(1) Drive the data management

Derived from system-level requirements, the ODD is 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:
Data volume needed and specific characteristics/monitoring

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

A priori assessment – Data Preparation

- PCA: dimensionality reduction, irrelevant features identification,
- MUP: relationships and correlations identification,
- Entropy: uncertainty and information richness

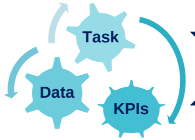
A posteriori assessment – Models Feedback

- CleanLab: model confidence, data mislabeling identification
- BSA: risk-based assessments, reliability and robustness
- Neuron Coverage: model behavior, coverage of

Data Enhancement – Adaptation & Augmentation

- Extend domain coverage and outliers handling
- Deployment domain features and impacting elements inclusion
- Adapt features engineering w.r.t operational conditions

Generic Pipeline > > > Practices Recommendation



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

Good model

(2) Drive the model development and training

Rely on ODD analysis outcomes

Data type and nature help to drive the ML design ;
Task complexity, data volume and availability analysis ;
Performances influencing elements of target environment ;
AI-level & system-level requirements (tolerance & monitoring) ;

Focus on target performance objectives – Industrial perspective

Generalisation assessment & perf. evaluation **vs** real KPIs ;
Critical system requirements to be included -> no impact on safety ;
Training objectives, eval. metrics selection/definition -> adaptations and acceptance criterion reviewed ;

Anticipate ways to enhance the performances

Performance influencing elements handling & exhaustive error analysis to identify weaknesses of the model ;
LM: regularisation, optimisation, and learning objective adaptation ;
Architecture, settings, and parameters adaptation

Model Design – ODD & Data outputs as driver

- Dimensionality: data characteristics (type & nature);
- VC-Dim: suitable model architecture and effective complexity
- Data available volume and outliers handling specific features

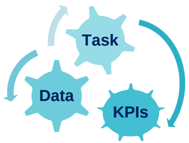
Model Development &Training – Target Performances

- Tuning: select accurate learning objectives, loss functions
- Translate KPIs to be included in training and evaluation
- Anticipate ways to enhance performances in iterative process

Model Validation – Behaviour Understanding and Monitoring

- Comprehensive performances evaluation: diverse metrics, tools
- Rigorous error analysis to understand and monitor errors distribution
- Include statistical tools to quantify generalisation, performances and uncertainty

Generic Pipeline > > > Practices Recommendation



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

Good model

(3) Reinforce the model robustness and stability

Using the class separation to improve stability

Maximum stability space identification per class, check the closest boundaries and distance of each data point;
Minimum perturbation changing the model's decision

Using ODD perturbations to reinforce robustness

Edge-cases as borderline cases with perturbations;
Leverage existing ones and generate others using perturbation methods to reinforce stability;

Using relevance properties to avoid bias

Identify learning bias of the model;
Model training analysis (e.g fuzzy relevance means underfitting);

Using stability to crosscheck data sets

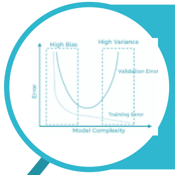
Lack of stability at some data point could be due to poor data; annotation and representation -> max-stability space computation & identification of poor annotations

Stability – Class Separation

- Formal methods: to study stability spaces
- Adjust training strategies to better separate classes
- Mitigation strategies and crosscheck data sets for stability

Robustness – ODD Perturbation

- Where the model is more likely to be confused (noisy data)
- Statistical methods: models behaviour under varying context
- Regularly evaluate robustness and incorporate findings in the model design



Bias – Relevance Properties

- Identify biased outputs, set requirements and justify model behaviour
- Automated relevance analysis and measures detecting mislabelling

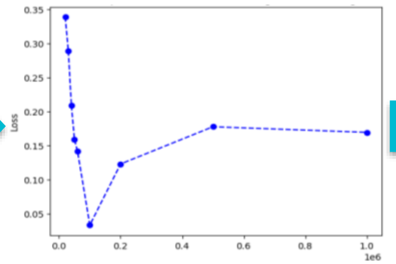
Generic Pipeline > > > Impact of Data Augmentation

Data Management and Model Performances

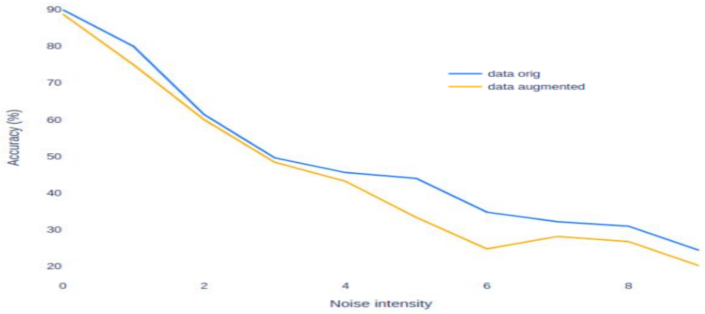
Experimentation – FashionMNIST

- Data volume enhancement and coverage increase
- More challenges (trends alteration in the original dataset)
- Requires revisiting experiments to understand its impacts

(1) Augmentation using ImageDataGenerator



(2) Training of the same classifier



(3) Robustness against gradual Gaussian noise

Data Analysis

• Random modifications using rotation with maximum angle of 10°
• Increased space coverage in augmented datasets => enhanced dataset completeness.
• Valuable information on both model learning and data augmentation effectiveness.

Learning Management

• Training with augmented data improved performances
• Increased stability and robustness against low rotation or small translations
• Deterioration of performance when augmented data are exceeds original ones
• Different augmentation methods may yield different results
• In real UCs the augmented data need to be confronted with ODD specification

Learning Verification

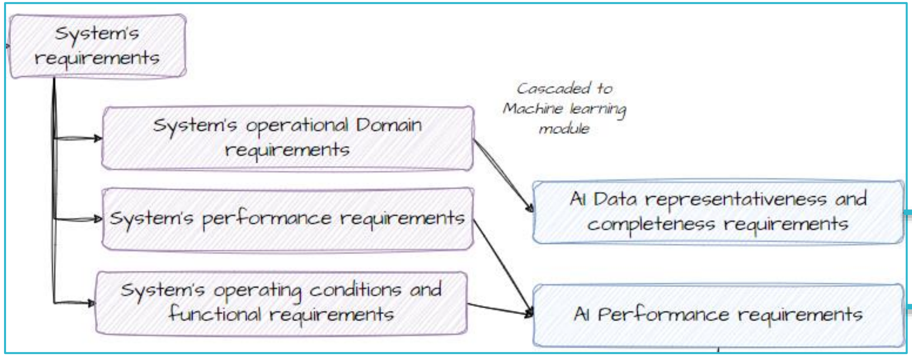
• The added noise had impacted robustness of both models, showing the small impact of the data augmentation on robustness
•The model trained with data augmentation demonstrates greater stability even ~ 90% of training data removed
• Data augmentation improved algorithm stability and accuracy retention

https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator

Generic Pipeline > > > System-level vs AI-level Requirements

Understanding Dependencies: System-Level – ML-Level

- Ensure AI-level performance aligns with system-level requirements
- Verify safety-related criteria and compliance at the AI-level



Key criteria of safety requirements at System-level cascaded to AI-level requirements

Data Management

- Data representative and complete w.r.t ODD and known outliers

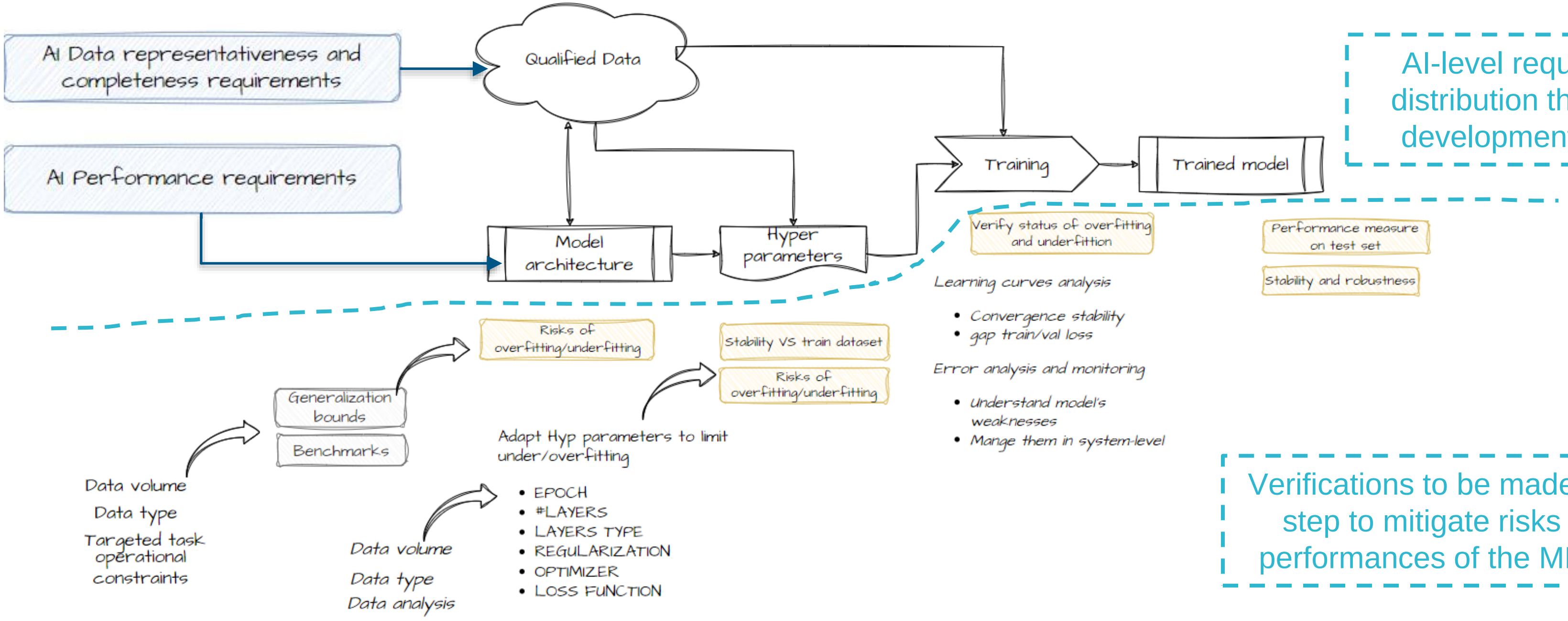
Learning Management and Verification

- No Overfitting/underfitting
- Performance on unseen test data OK
- Model stable W.r.t. inputs perturbation
- Model is robust to influencing elements
- Model is robust and stable w.r.t. train dataset
- Model's errors minimized and weaknesses known

Defining AI-level requirements and target application

Generic Pipeline > > > System-level vs AI-level Requirements

Understanding Dependencies: System-Level – ML-Level



AI-level requirements distribution through the development process

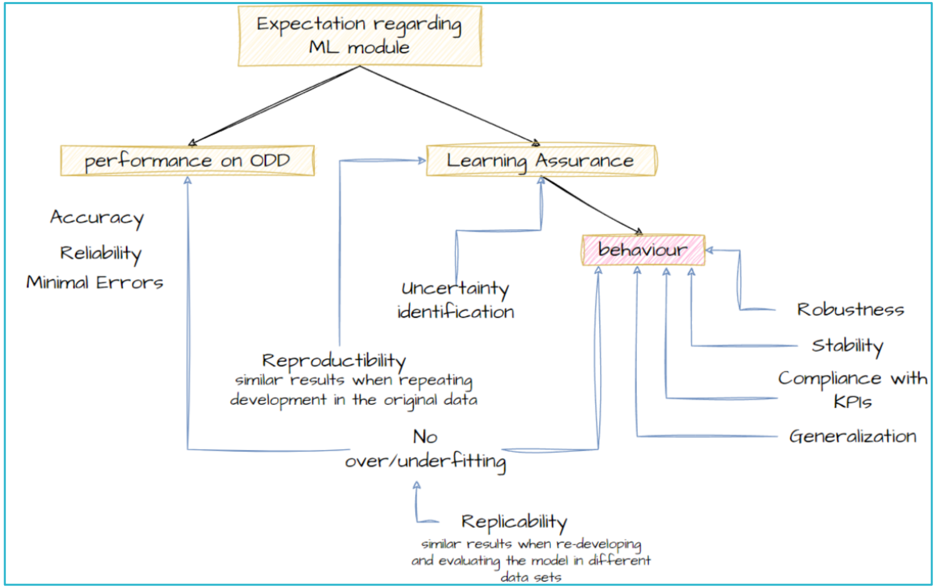
Verifications to be made at each step to mitigate risks of low performances of the ML model

AIRBUS

Generic Pipeline > > > System-level vs AI-level Requirements

AI-level Performances Requirements

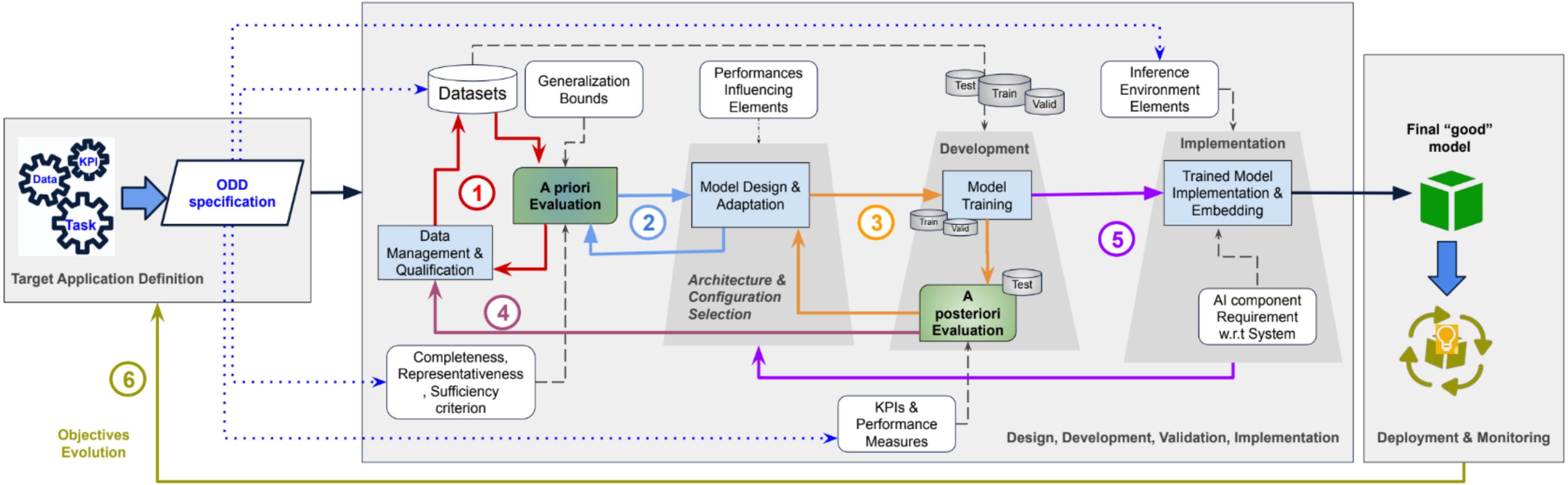
- **Criteria:**
  - Aligned with system-level objectives and efficiency.
  - Measurable and specified (e.g., accuracy, precision, maximum error rate).
  - Robust and stable model behaviour.
  - Verified performances in the Operational Design Domain (ODD).



- **Objectives**:
  - Promote ML models performances to be trustable and safe
  - Reduce impact of environmental impact on performance
  - Clear requirements specification with allowances/handling of uncertainty and variability
  - Establish mechanisms for monitoring and adapting to changing conditions

Generic Pipeline > > > Application Agnostic Pipeline

Framework Implementing the W-shaped Learning Assurance



.... ODD elements to be considered

– – – – Specific inputs/requirements per task

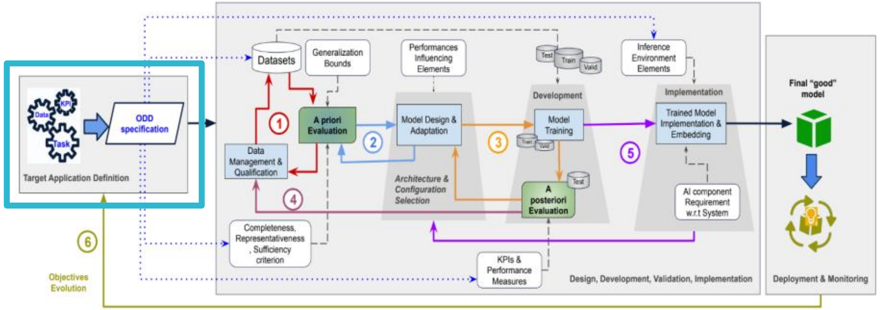
——— Forward/backward actions of the pipeline

Generic Pipeline > > > Application Agnostic Pipeline

Target Application Definition

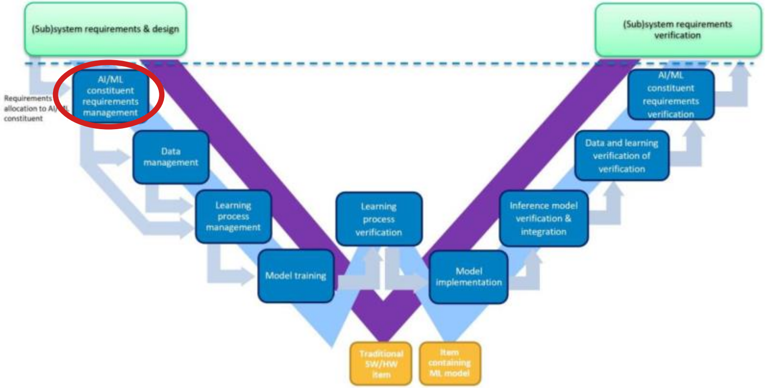
Understanding the objectives & ODD specifications

Datasets. input/output spaces, quality criterion (completeness, representativeness, and sufficiency), outliers & edge cases, OOD scoping;
Performances Influencing Elements. characteristics of the target environment that are more likely to influence the model, system-level specifications, AI-level working conditions.



KPIs & Performance Measures

Target performances. AI-level requirements derived from the system-level requirements, safety and certification related requirements,
Operating conditions & Monitoring. Acceptability criteria and conditions at AI-level



Inference Environment Elements.

Deployment environment features impacting results
System-level requirements and operating conditions having an impact on the ML-component
Possibilities/risks of changing conditions that cannot be controlled at AI-level (e.g. weather conditions and light intensity).

Generic Pipeline > > > Application Agnostic Pipeline

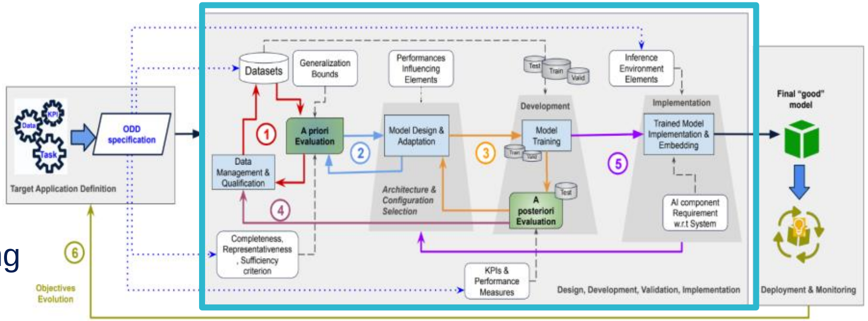
Design, development, validation, and implementation

Two-folds Evaluation

A priori evaluation. Before ML/DL design.

Performance objectives assessment, prerequisites understanding

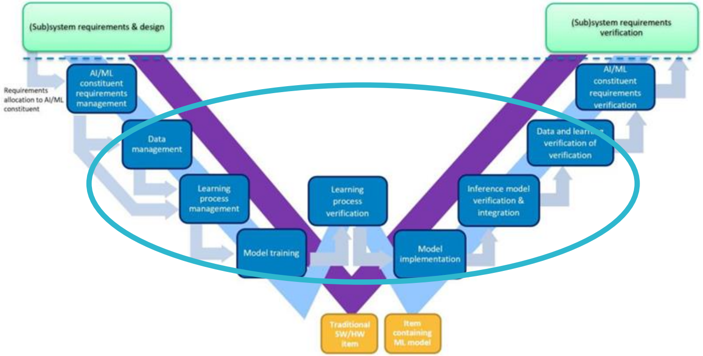
- Data quality and volume criteria requirements,
- Completeness and representativeness;
- Generalization bounds selection and computation;



A posteriori evaluation. After ML/DL training.

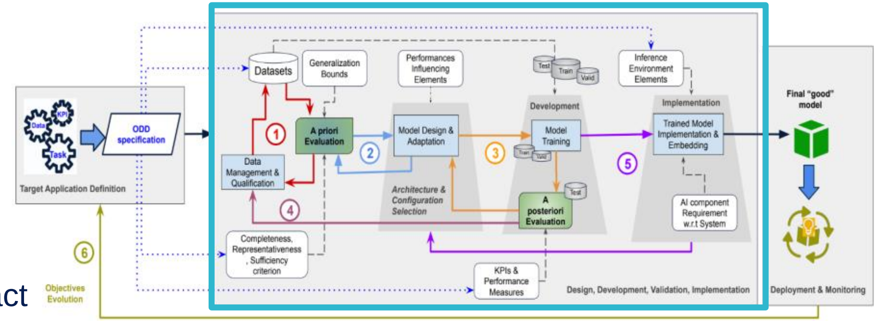
Performances evaluation and verification

- 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



Generic Pipeline > > > Application Agnostic Pipeline

Design, development, validation, and implementation

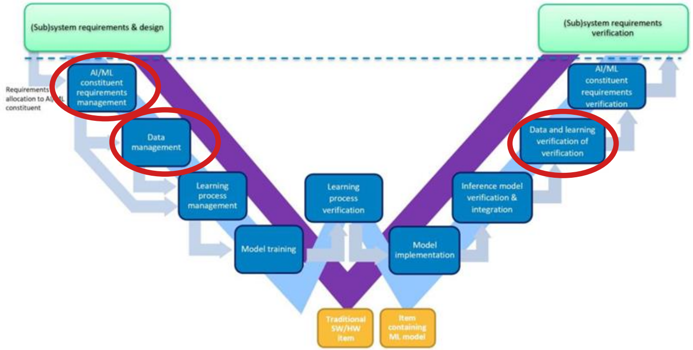


(1) Data qualification and preparation

a) Identify important criteria for the data quality (representativeness and Completeness), samples distribution analysis, corner/edge cases, outliers, impact on the training;

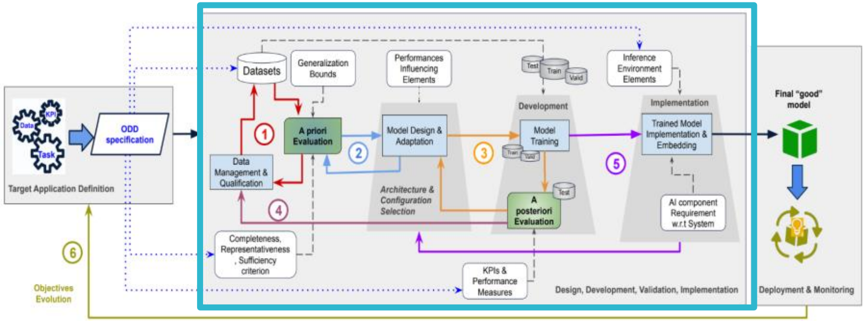
b) ODD analysis: identify the requirements, in terms of data volume needed, specific cases handling on the data (specific measures for some outliers);

c) If data is not collected yet, based on (a) and (b), data collection & preparation.



Generic Pipeline > > > Application Agnostic Pipeline

Design, development, validation, and implementation

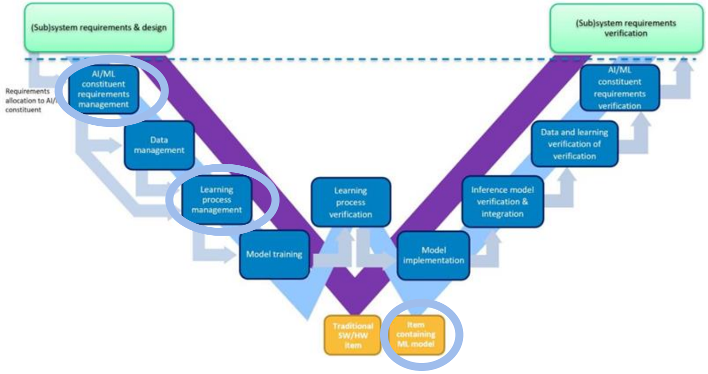


(2) Model Design & Adaptation

a) Architecture definition, approach that meets data and target application specificities;

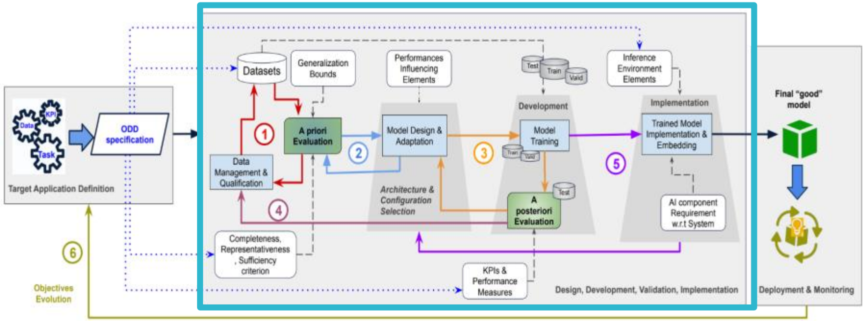
b) Model that is compliant with the constraints at the system-level and the target application (e.g real-time execution, be embedded in a resources limited system …), data-related constraints (e.g. available data volume, inputs size and type);

c) Use insights from the ODD analysis (performances influencing elements, system criteria …), data availability and features, estimated generalization (bounds)



Generic Pipeline > > > Application Agnostic Pipeline

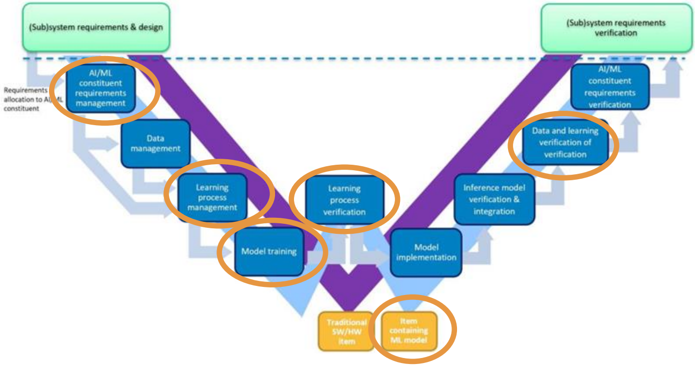
Design, development, validation, and implementation



(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)

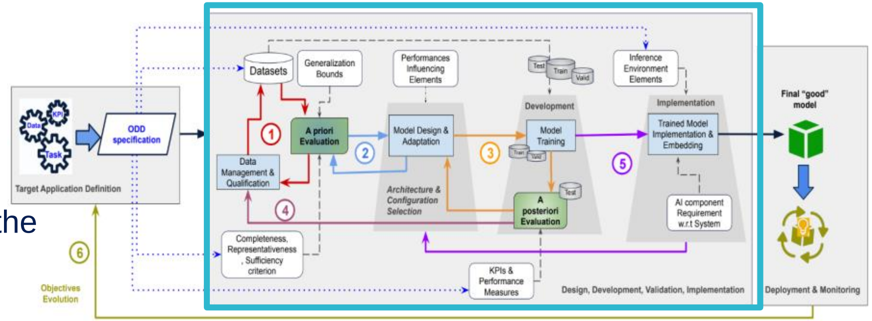
A backward action can be considered to re-work the model design and configuration if acceptance-criteria not verified



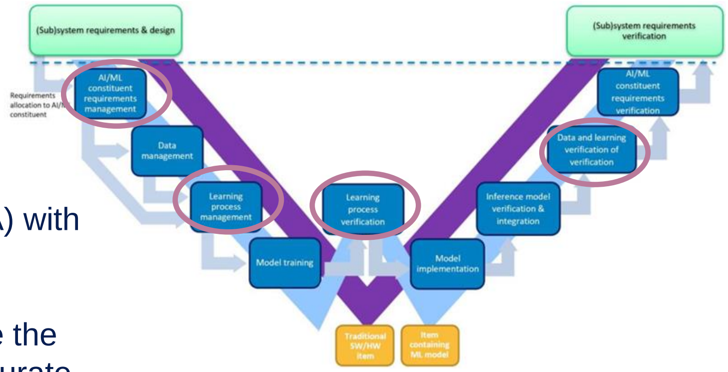
Generic Pipeline > > > Application Agnostic Pipeline

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…

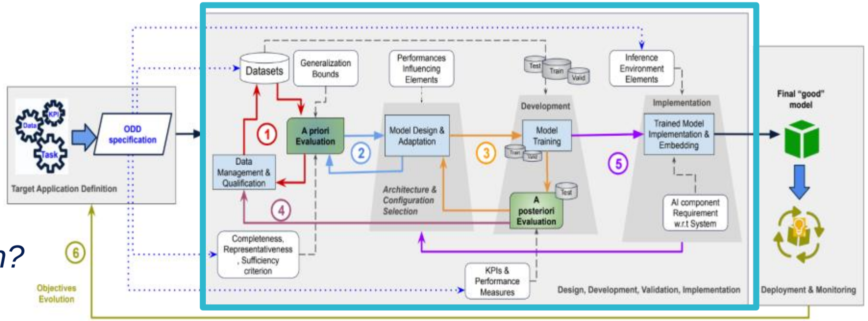


Possible options:

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).

Generic Pipeline > > > Application Agnostic Pipeline

Design, development, validation, and implementation

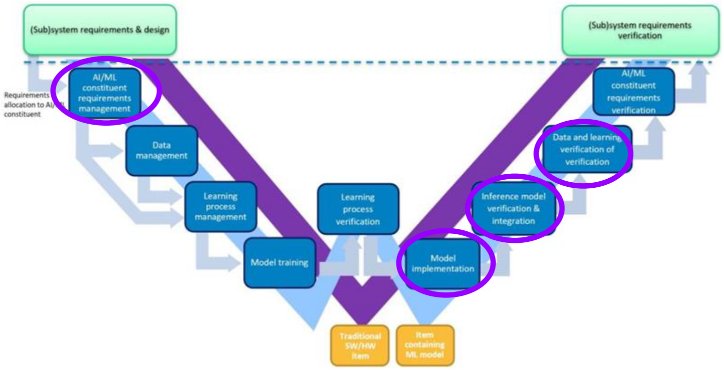


(5) Implementation & Verification

Is the expected objective met while interacting with target domain?

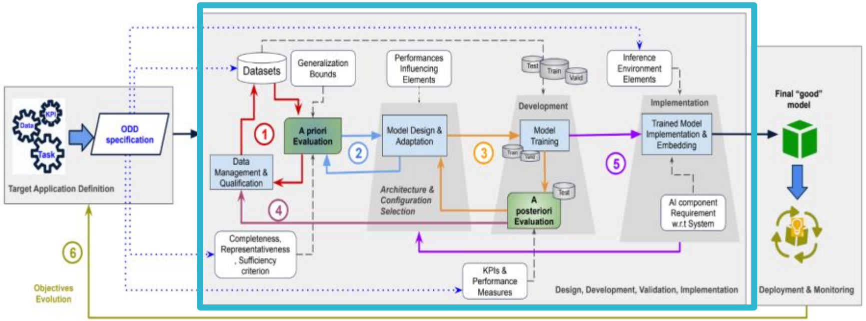
a) Inference Environment Elements are consumed by the implemented model
b) Verify performances in the target environment & AI component requirement w.r.t System requirements
c) The model is either:
   i. validated and go to the *Deployment & Monitoring phase*
   ii. Rejected and a backward action is needed,

- 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



Generic Pipeline > > > Application Agnostic Pipeline

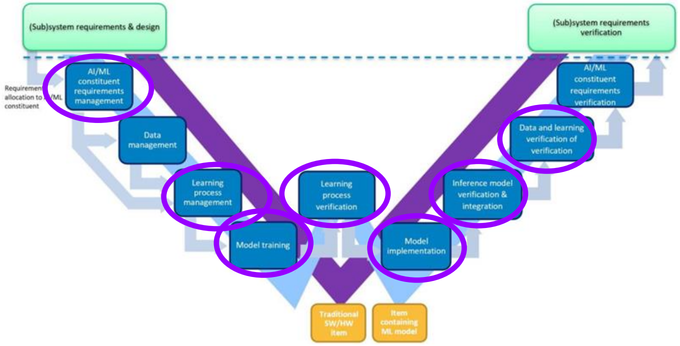
Design, development, validation, and implementation



(5) Implementation & Verification

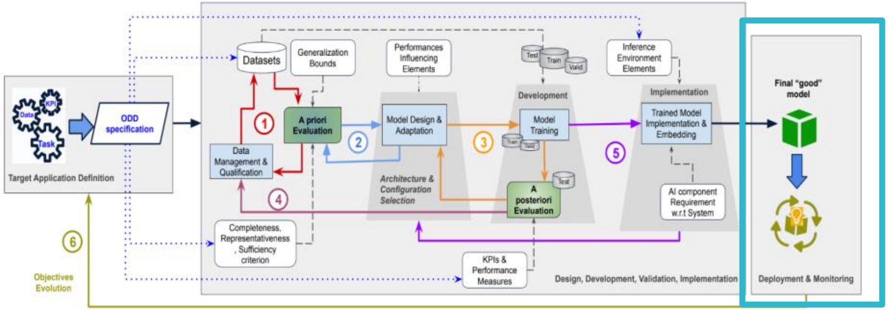
<u>Backtracking – Be Aware of:</u>

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.



Generic Pipeline > > > Application Agnostic Pipeline

Design, development, validation, and implementation



(6) System's objectives evolution after model deployment

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

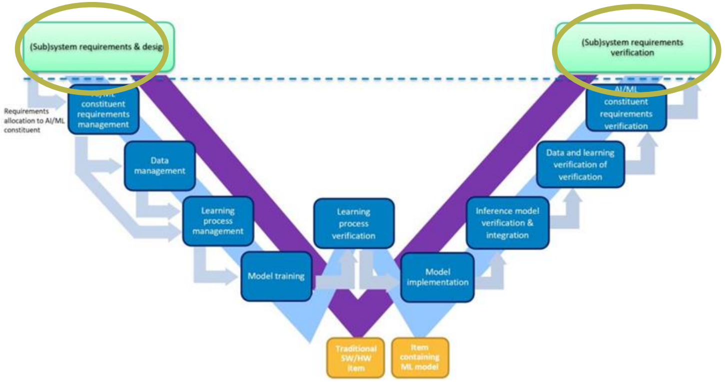
a) Definition of the ML component NEW objectives to be considered

b) Major activities:

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

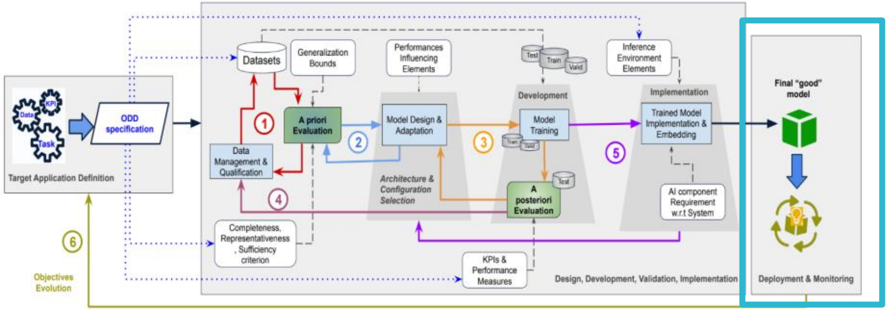
ii. Re-using (retraining or fine-tuning) of the initially validated good model;

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



Generic Pipeline > > > Application Agnostic Pipeline

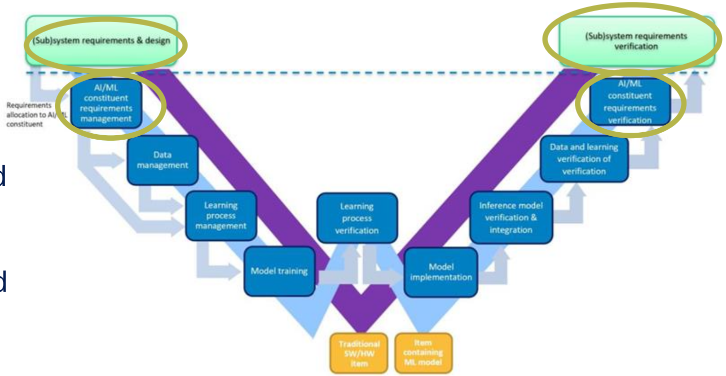
Design, development, validation, and implementation



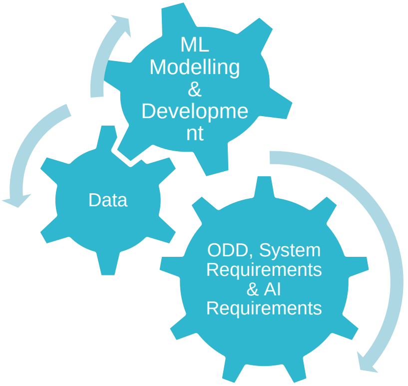
(6) System's objectives evolution after model deployment

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



Generic Pipeline > > > Conclusions



AI-level Main Development Components

Data Management & Qualification: Completeness and Representativeness

- A well defined ODD is needed, including operational conditions understanding;
- Diversify tools and metrics for a better assessment of the diversity & relevance of the data;
- Feedback from Model's learning behaviour and evaluation results ;

Model Design & Development: Generalization Assurance and Verification

- For a well generalizing model, avoid overfitting/underfitting (balanced);
- Generalization bounds: statistical tools, not sufficient guarantees to rely on;
- Deep investigations needed (error analysis, uncertainty & optimizations);
- Alignment of experimental & industrial objectives, focusing on operating conditions and ODD;

Learning Verification: Performances, Robustness & Stability

- Performances stability is highly related to the robustness of a model;
- Diverse approaches (formal, statistical, and empirical methods) can be used;
- Effective validation requires integrating various approaches to address the ODD and anticipated perturbations;

Q&A

www.sli.do
#AIDays
Passcode: hmkota

MLEAP project

/ EASA perspectives on MLEAP takeaways

MLEAP project

EASA perspective on MLEAP takeaways‘

European Union Aviation Safety Agency

Xavier Henriquel

EASA MLEAP Tech lead

François Triboulet

Project Manager ‘AI Assurance’

MLEAP – takeaways for each task

Datasets completeness and representativeness

Model generalisability

Model stability, robustness

ODD is the centerpiece of the Learning Assurance concept

MLEAP – takeaways for task#1

Structuring the set of proposed methods into guidance for the applicants

Guide whether the method applies to a priori or a posteriori evaluation, and for which loop of the generic pipeline.

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)

MLEAP – takeaways for task#2

Ensuring generalisation remains a challenge

Set of methods experimented on "toy use cases" <u>do not provides satisfactory generalisation bounds</u>

<u>Other methods</u> should be further investigated

Generalisation is a <u>key enabler for higher criticality levels</u> AI-based systems.

Generalisation is a very <u>active field of research</u> to be monitored in the mid-term

MLEAP – takeaways for task#3

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

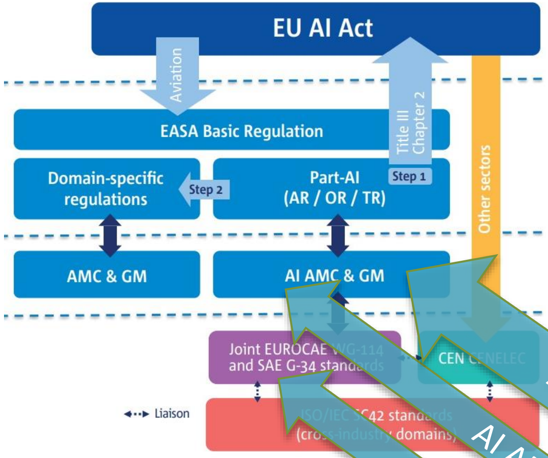
The generic pipeline provides a framework to organise the main verification activities for a machine learning model

- 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 now defined in the context of the three tasks of the MLEAP project

- 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?)

MLEAP outcome Implementation Plan



Rulemaking

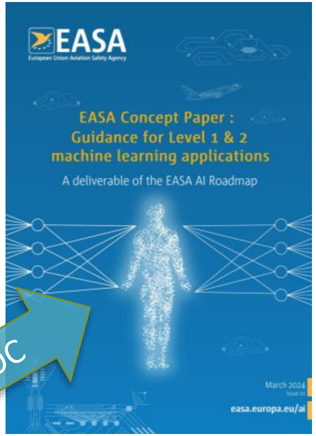
AI GM

AI AMC

Need for ER/AIR?

Anticipated MOC

Final report



Issue 02

Intermediate report

MLEAP final report actions
- WG114 ARP6983: Q1-2025
- EASA CP Issue 3: Q4-2025
- RMT.0742: Q2 2024-Q4 2027

MLEAP Intermediate report actions
Update of EASA CP Issue 2 - Q1 2024 (Done)

Wayforward - Use cases

Toy use cases and aviation use cases

- All MLEAP models, datasets, tools & methods and dedicated plateform remain available to EASA for the next 2 years

Possible Use of MLEAP artefacts

- Under assessment – large amount of data
- Identification of a limited number cases of interest in progress:
  It could be valuable to Aviation AI communities to have some shared use cases and examples for methods and tools.
- Inputs from audience / stakeholders welcome !

Way forward

Task 1 – Data quality

- Augment current MOCs with final report Chapter 4

Task 2 - Generalization

- Augment current MOCs with Chapter 5 and chapter 7 « recommendations and pipeline »

Task 3 - Robustness

- Improving the existing MOCs with MLEAP report Section 6
- Clarification of objective LM11 in EASA CP
- Explore benefit of « Relevance » properties

Research activities

- Lead by EASA, other authorities or external groups e.g. DEEL with Paper On the Feasibility of EASA Learning Assurance Objectives for Machine Learning Components
- Primarily on Task 1 and Task 2

Please use Slido
& raise your questions

www.sli.do

#AIDays

Passcode: hmkota

/ Conclusions of the MLEAP Stakeholders day #4

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 }

>>

Thank you for your participation to the EASA AI Days High-Level Conference !

Have a safe trip back!