Welcome to the EASA AI Days High-Level Conference !

Day 2: MLEAP Stakeholders' Day

17th May 2023

Welcome to the EASA AI Days High-Level Conference !

Day 2: MLEAP Stakeholders’ Day

Guillaume Soudain - EASA AI Programme Manager

Welcome to the EASA AI Days High-Level Conference !

_MLEAP STAKEHOLDERS DAY

#2

Paving the way for the future of Artificial Intelligence in Aviation

EASA
European Union Aviation Safety Agency

LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS LNE

numalis

MLEAP project: [Machine Learning Application Approval]

May 17th 2023

AIRBUS
PROTECT

Agenda

- **Introduction of the MLEAP project and of the Partners**
- **Presentation of the use cases**
- **Presentation of the single public deliverable**
- **Q&A session**

COFFEE BREAK

- **Presentation of the objectives and progress of Task 1 (data management)** *Swen RIBEIRO, LNE*
- **Presentation of the objectives and progress of Task 2 (generalisation guarantees)***Thiziri BELKACEM – Jean-Baptiste ROUFFET, Airbus Protect*
- **Presentation of the objectives and progress of Task 3 (robustness guarantees)** *Arnault IOUALALEN, NUMALIS*
- **Conclusions & Next Steps**
- **Networking Lunch**

Who we are > > >

Consortium members :

EASA

Willy Sigl, Xavier Henriquel, Guillaume Soudain, François Triboulet

Airbus Protect

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

MLEAP Team

LNE

Olivier Galibert, Swen Ribeiro, Agnes Delaborde, Sabrina Lecadre

Numalis

Arnault Ioualalen, Noémie Rodriguez

LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS

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

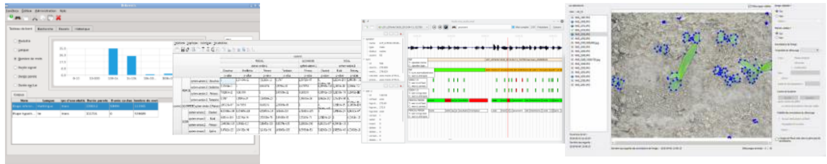
AI evaluation Department

- 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



www.lne.fr/logiciels/lne-matics

Certification for AI processes (2021)

https://www.lne.fr/en/service/certification/certification-processes-ai

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

LEIA 1 LEIA 2 LEIA 3

AIRBUS

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*
- *20 persons, Montpellier*

On-going projects:

HE MLEAP with EASA
2 EDIDP (Defence)
ESA…

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

/ Airbus Protect

: What we do

Training

We are a recognised training organisation

Consulting

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

Software

Specialised software supporting end-to-end safe mobility activities

Innovation

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

an {Airbus} company

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

… delivering consulting, services & solutions

R&T & software development projects in AI:

DEEL project for IRT Saint Exupéry and ANITI
Confiance AI project
EPI project for IRT SYSTEMX (Consortium with STELLANTIS, NAVAL Group, EXPLEO, LIP6)
PRISSMA project for French Ministry of Transportation

Day 2: MLEAP Stakeholders’ Day
Introductory notes from EASA technical team

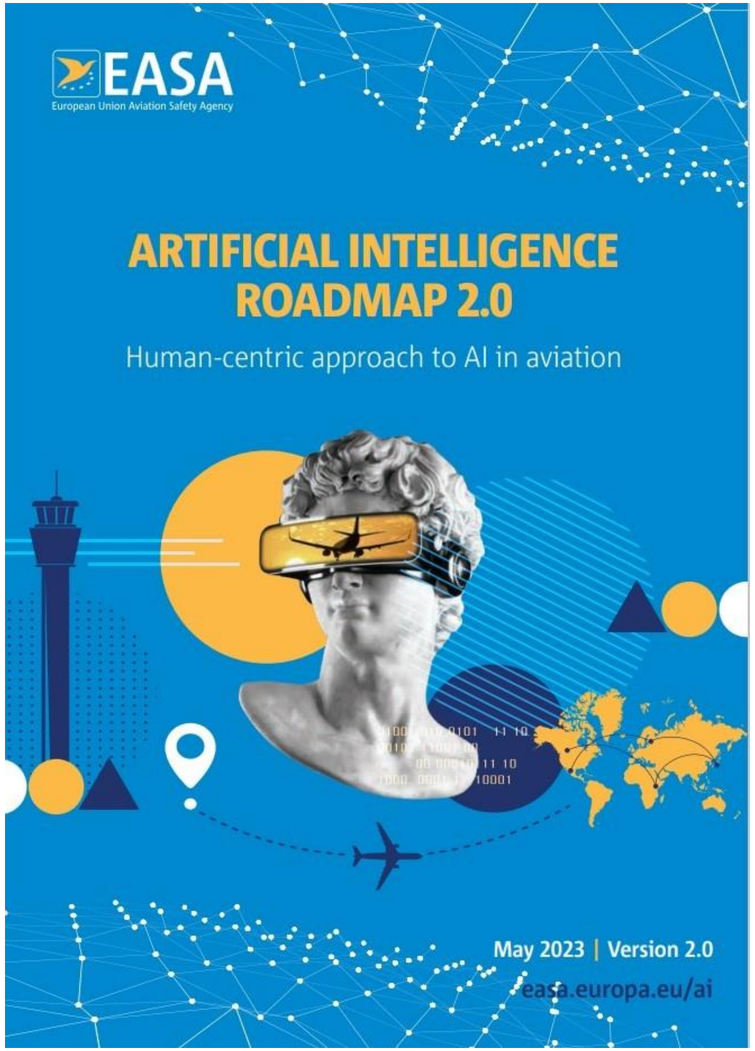
Guillaume Soudain
EASA AI Programme Manager
MLEAP Project Sponsor

Xavier Henriquel
EASA Safety Expert
MLEAP Tech Lead

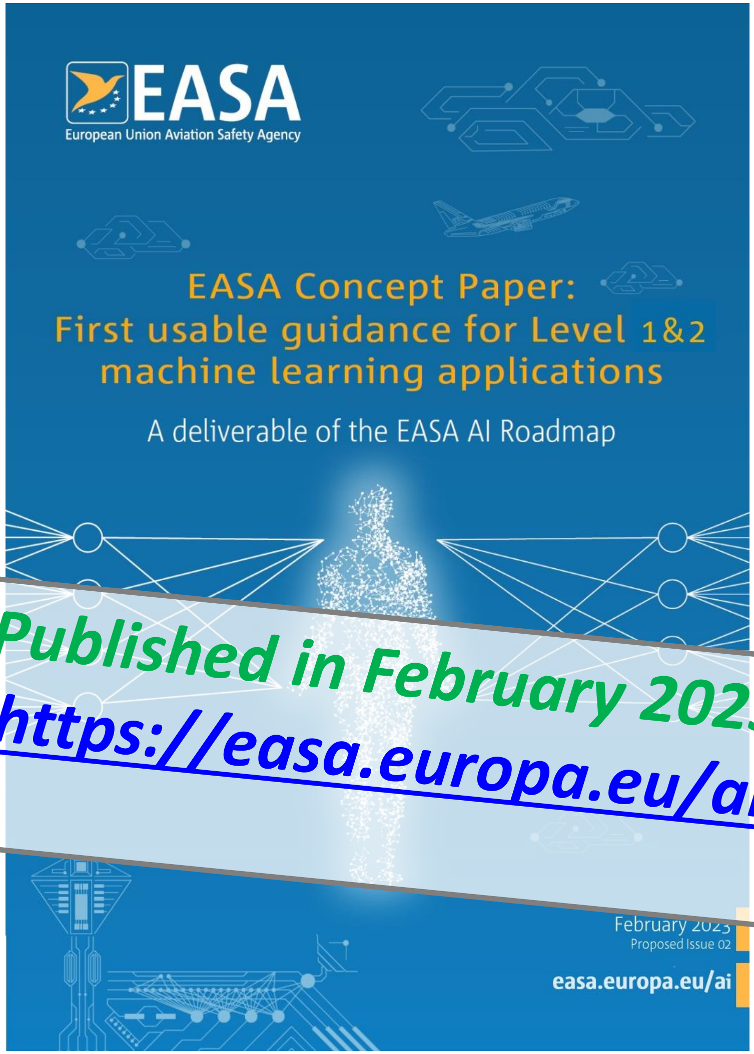
François Triboulet
EASA ATM/ANS Expert
Coordinator

EASA AI Roadmap – Towards AI trustworthiness

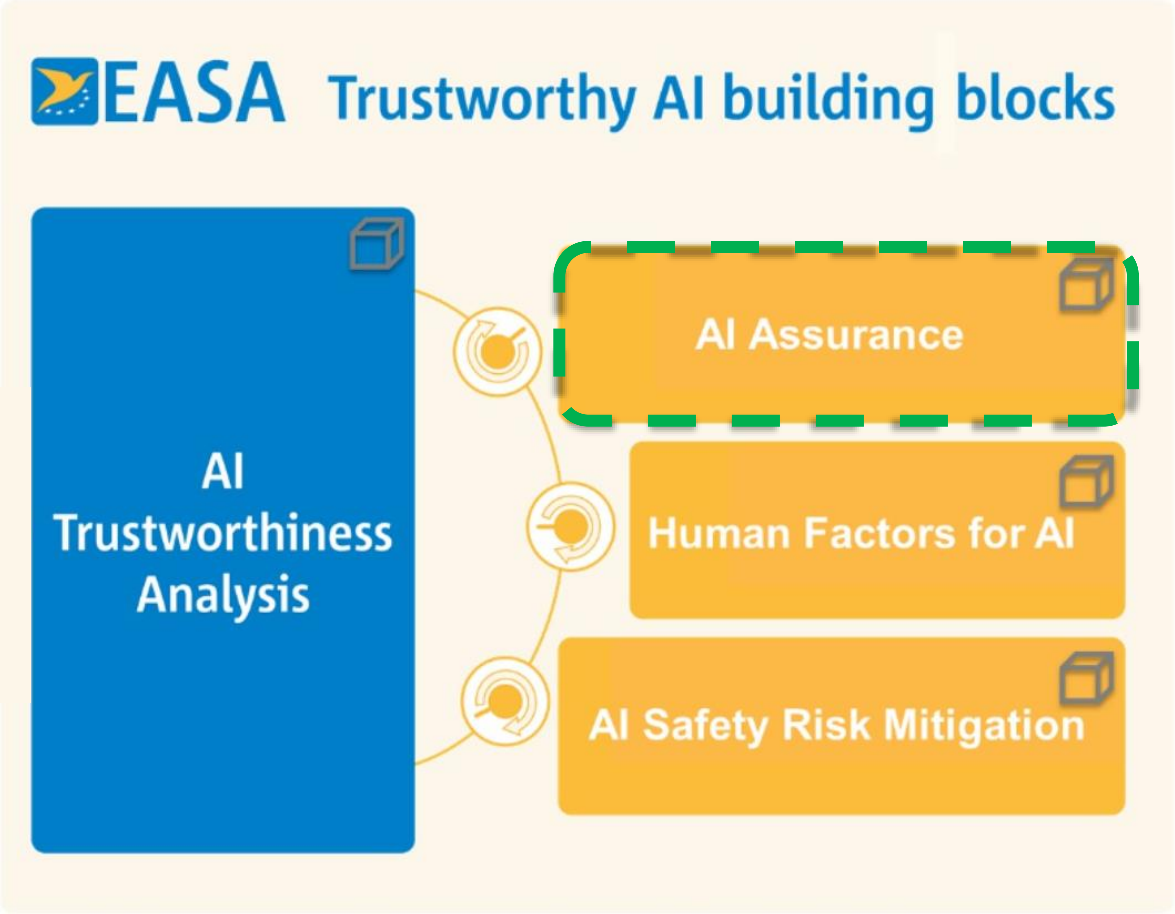
- → Impact on all aviation domains
- → Common issues for safety-related applications
- → « AI trustworthiness » concept is the key!



EASA guidance for Level 1 & 2 ML* applications



Published in February 2023
https://easa.europa.eu/ai



* ML = Machine Learning

TOP3 challenges for Level 1&2 ML guidance

1. **Anticipate means of compliance for Learning Assurance objectives on ML Model guarantees (generalization and robustness)**

   →Exploit the Horizon Europe Research project MLEAP on 'Machine LEarning applications APproval'

Partnering on research projects is a key driver for the guidance!

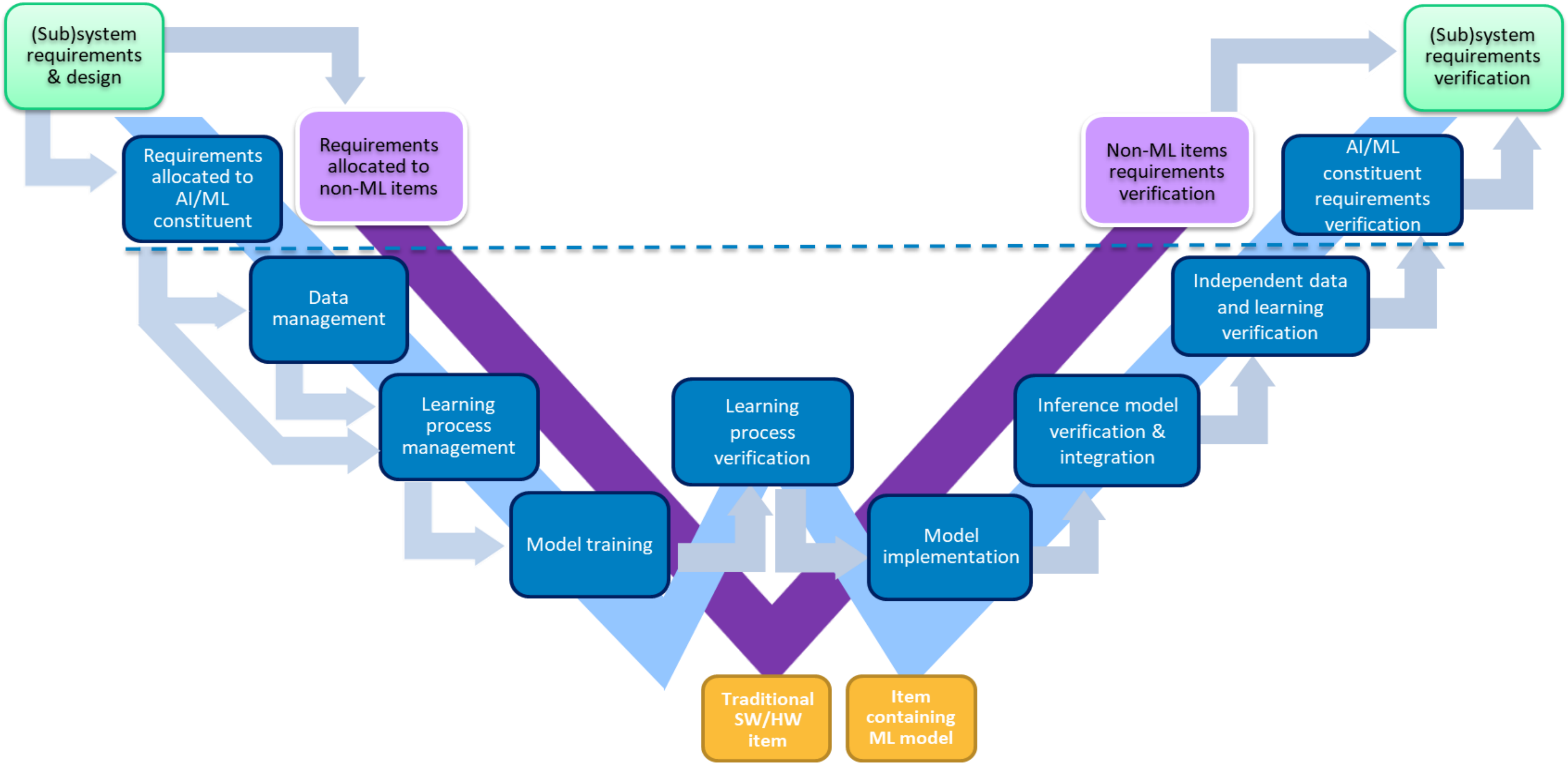
2. **Operational explainability & human centric aspects of AI**

   →Foster trust in the human-AI teaming by developing specific Human Factors guidance.

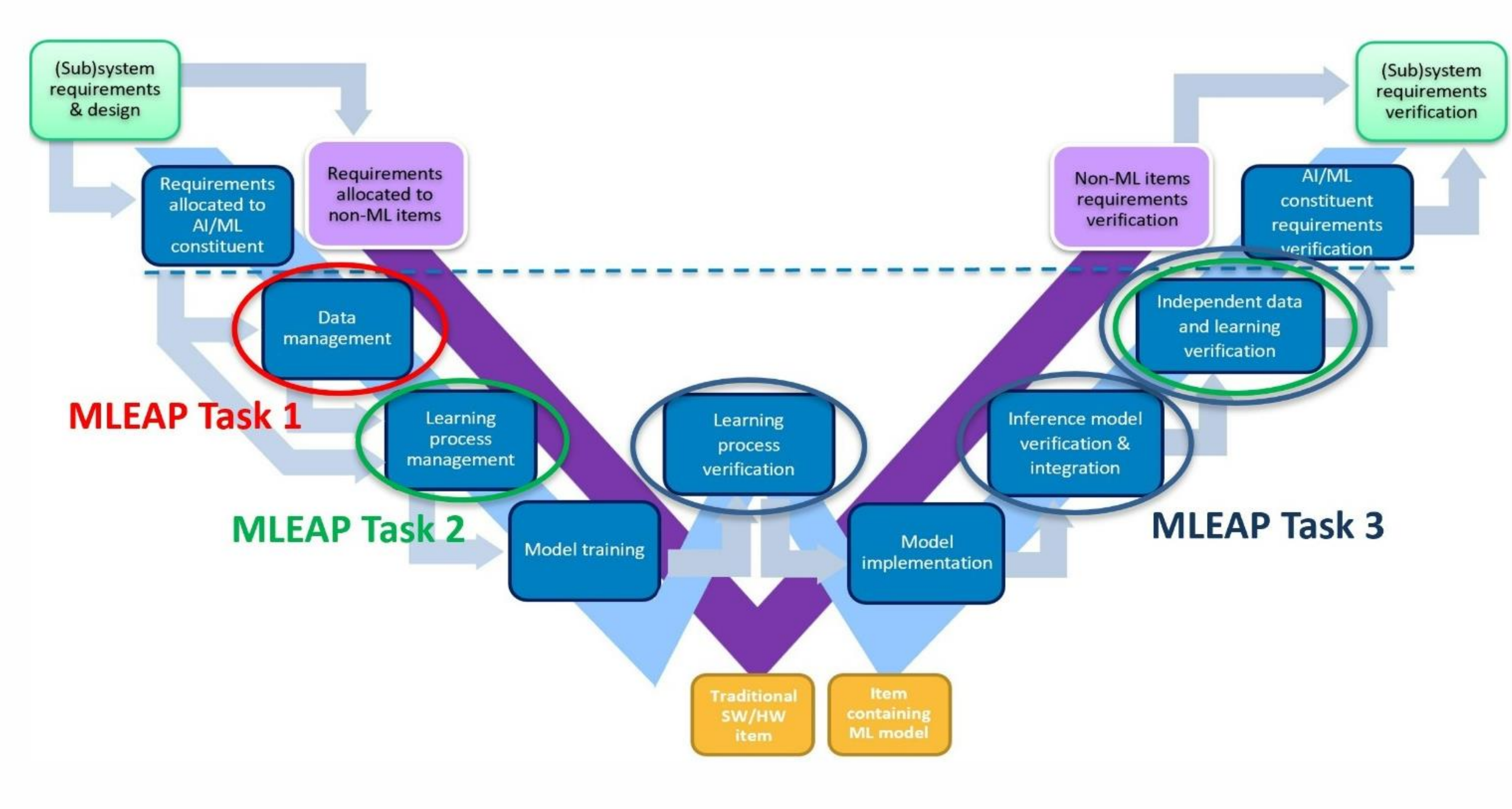
3. **Ethics-based assessment – social & societal aspects**

   →Evaluate and refine guidance based on use cases

W-shaped assurance process



W-shaped assurance process



Machine Learning Application Approval (MLEAP) project

Objectives

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

Budget

1.475 Million Euros funded by EU Horizon Europe

Timeline

May 2022 - May 2024

Research consortium

Airbus Protect - LNE - Numalis

What is MLEAP project ?

Task #2 Generalization guarantee

Task #3 Algorithm and model robustness

Task #1 Data completeness and representativeness

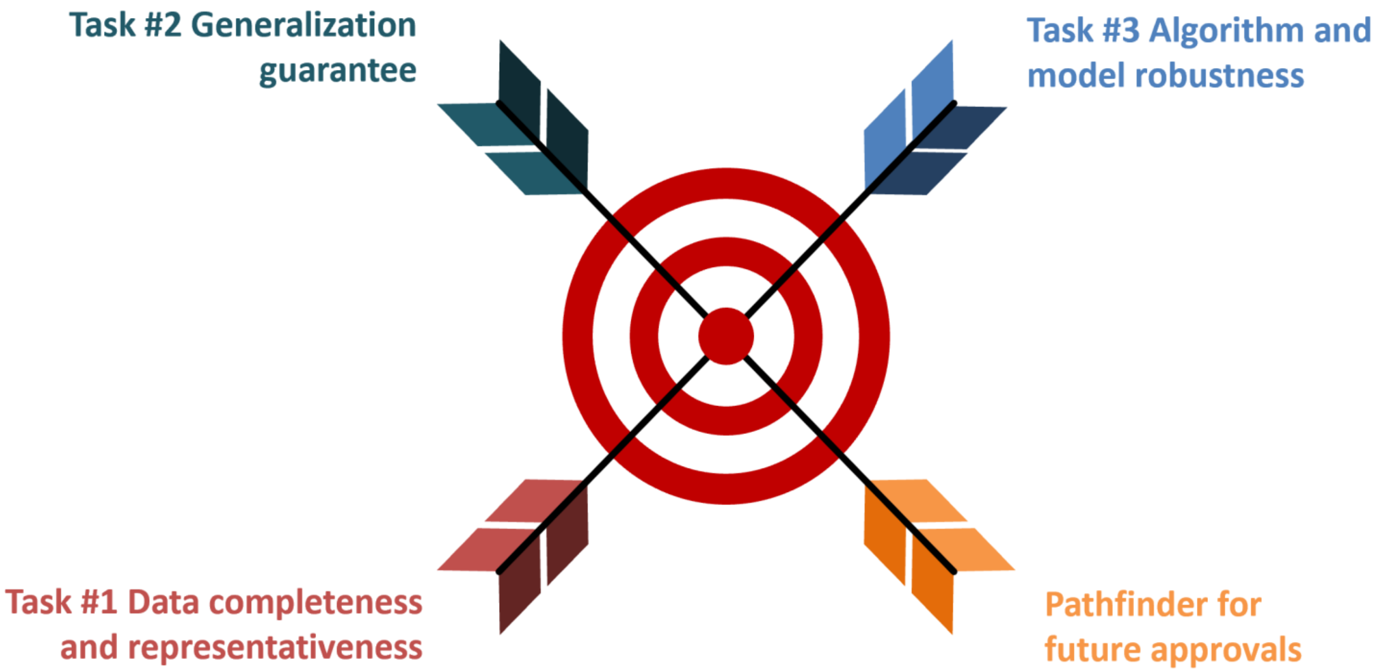
Pathfinder for future approvals

MLEAP Task 1 - Data completeness and representativeness

- Data quality is a challenge due to inherent costs

- Data completeness and representativeness are usually not addressed per se:
  - Almost no dedicated tools
  - Tradeoff between representativity and diversity

  …But crucial to AI/ML performances & guarantees



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

- This task aims at defining protocols and strategies to enhance the ability of released models to generalize well

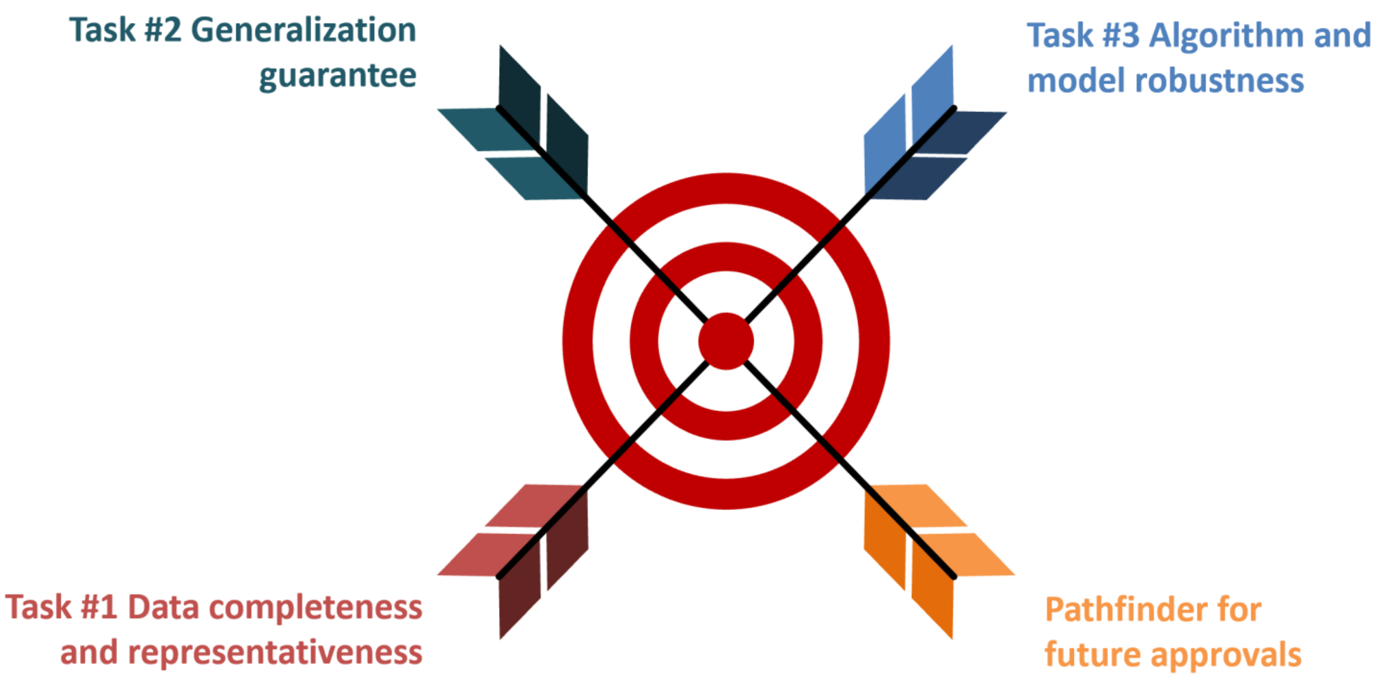
  … accounting for data quality and volume and obtaining quantifiable guarantees.

MLEAP Task 3 – Algorithm and model robustness

- Aligning existing concepts and definitions between EASA Concept Paper, CoDANN I & II IPCs and ISO/IEC 24029

- Variety of approaches available:
  Empirical, statistical and formal methods

- Part of the ongoing effort of evaluating formal methods benefits (e.g. EASA-Collins Aeropsace ForMuLA IPC)



MLEAP - Pathfinder for future approvals

- Practical aviation AI/ML use cases
  - EASA access to detailed models & datasets
  - Public data/examples used when possible to allow comparison with 3rd parties

- Knowledge sharing
  - Events organized every 6 months
  - Project [page](#) with latest results
  - Public reports

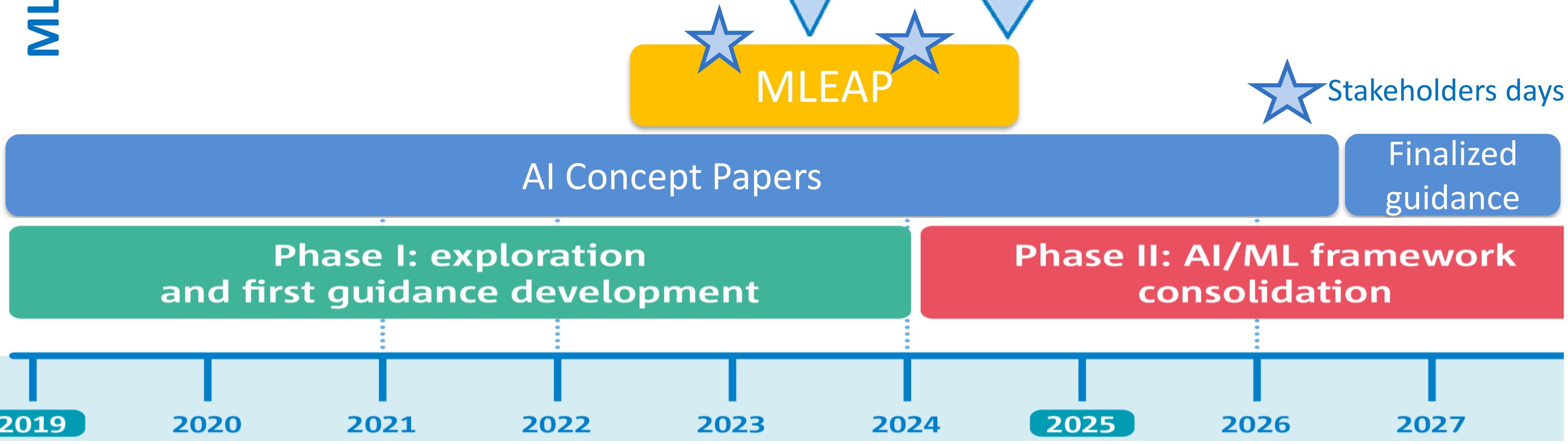
- EASA AI Concept paper regularly updated with MLEAP outputs

MLEAP project milestones

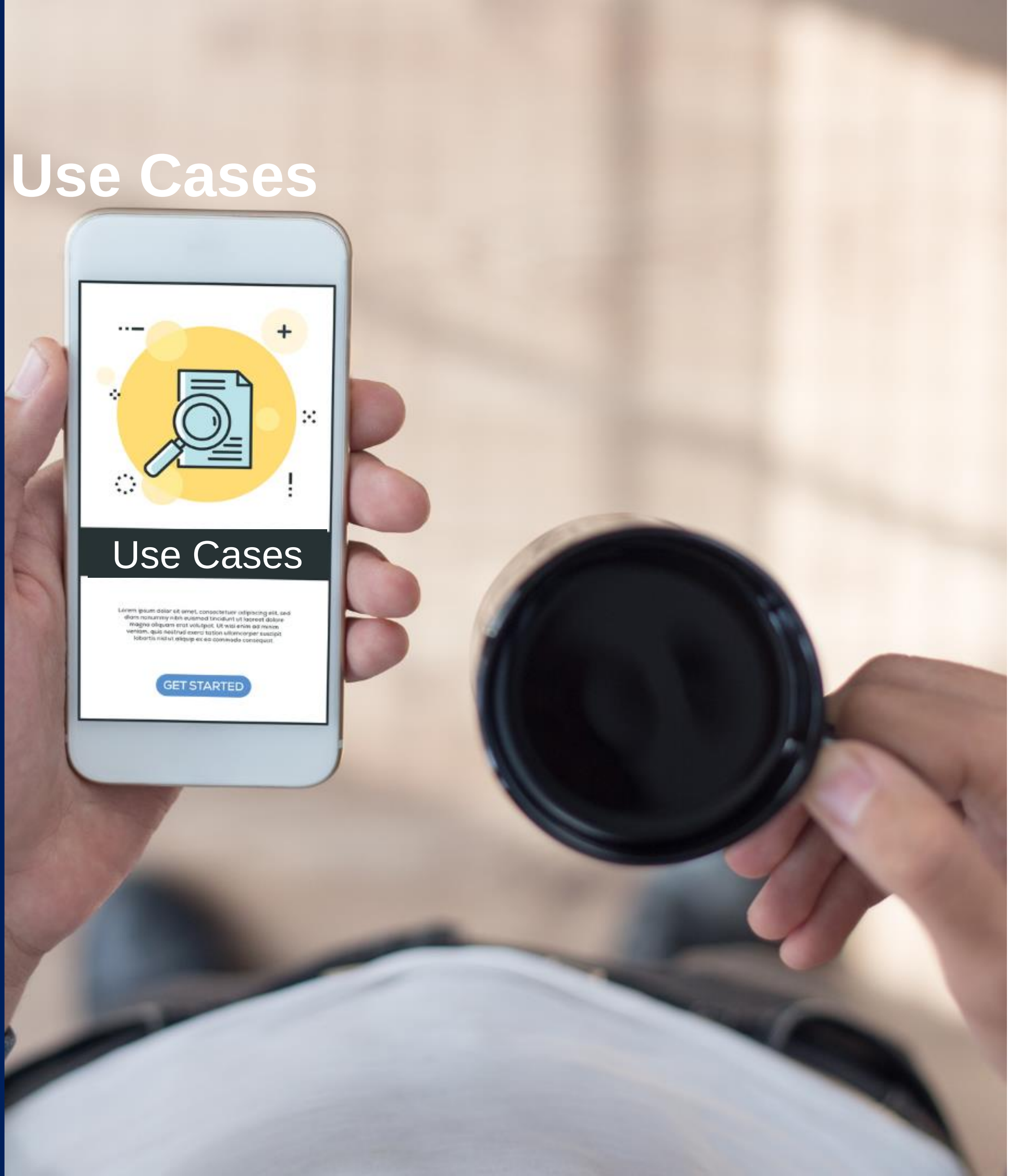
MLEAP deliverables and events

May-June 2023
- **First public report – 11th May 2023**
- Stakeholders day & Dissemination events
  - “EASA AI days” – **17th May 2023**
  - “Paris Airshow 2023” – **21st June 2023**

May-June 2024
- **Final public report**
- Stakeholder day & events



MLEAP – Presentation of the Use Cases



Objective:

Lead and support the methods/tools selection process: Data qualification, Models evaluation, and Performance verification

Perform a comparative evaluation, of selected methods and tools, to assess their efficiency

Help making recommendations for possible means of compliance

Speech to text
STT - ATC

Automated visual inspection
AVI

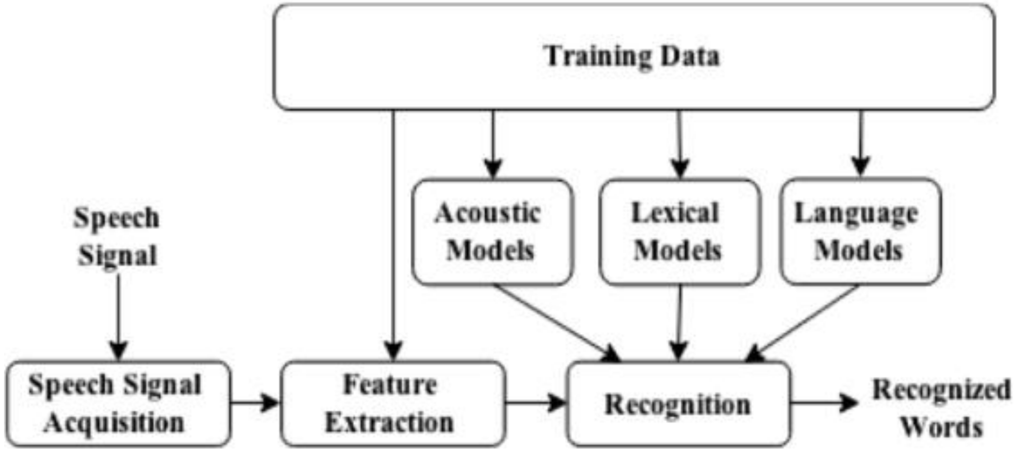
Collision avoidance
ACAS - Xu

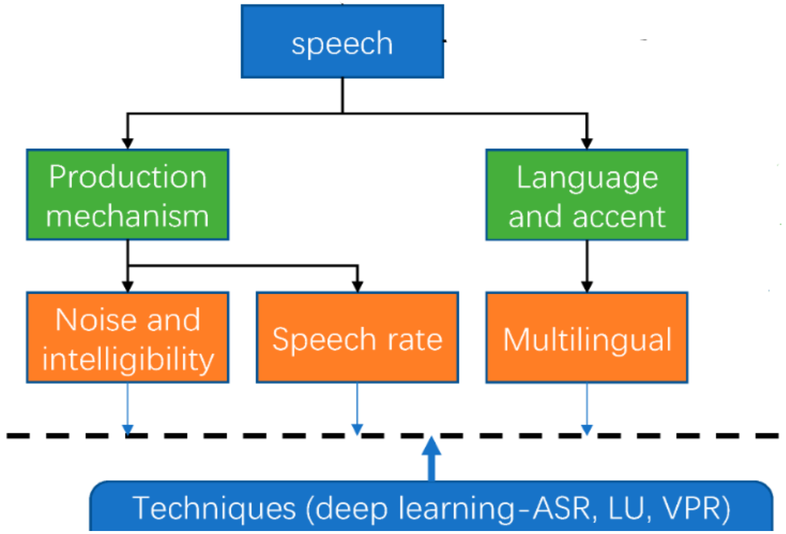
MLEAP – Use Cases Description> > >

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

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

- *Language Understanding (LU):* (Raju et al., 2021) systems provide both text and semantics associated with every input utterance.
- *Spoken Instruction Understanding (SIU):* (Lin, 2021) correctly interpret the ATCO instructions communicated between the control tower and the pilots
- *VoicePrint Recognition (VPR):* (Saquib et al., 2011) or Speaker Recognition Systems (SRS), aim to validate a user's claimed identity using characteristics extracted from their voices





AIRBUS

MLEAP – Use Cases Description> > >

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

➢Model & Data:

From Airbus internal project & open-source data/models

Data (utterances + transcriptions):

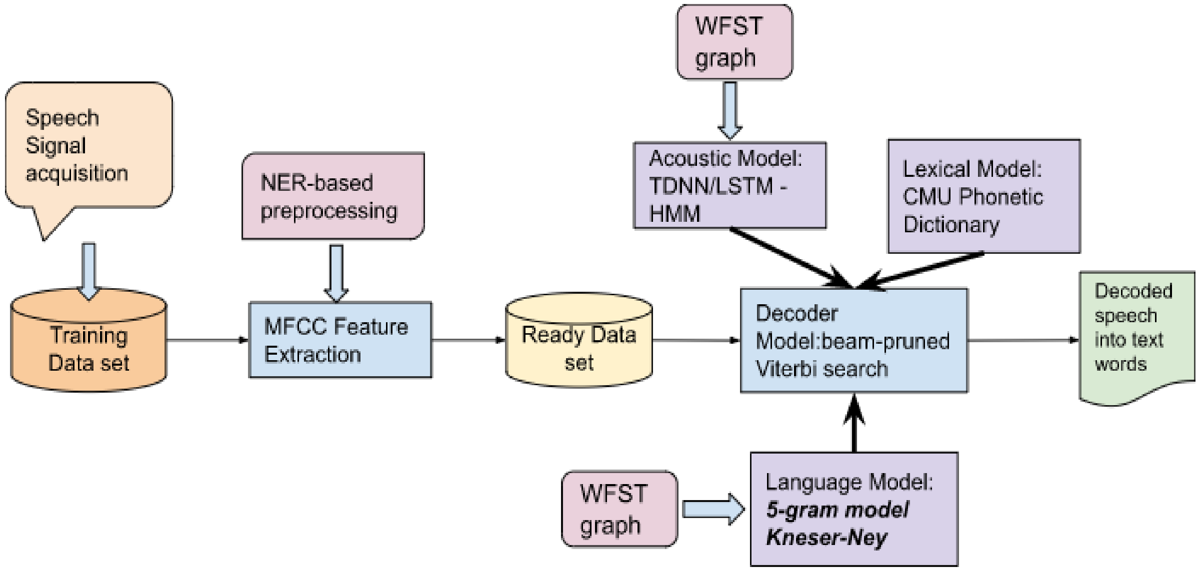
Airbus data: ATC interactions, in English, 100h French accent, 50h Chinese accent

Open Source: real (ATCO2, UWB, NIST LDC-ATC), simulated (ATCO Sim), several accents (Chinese, French, German, Slovak, Australian), US, ~44h30min

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

MLEAP – Use Cases Description> > >

Automatic Visual Inspection (AVI)

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

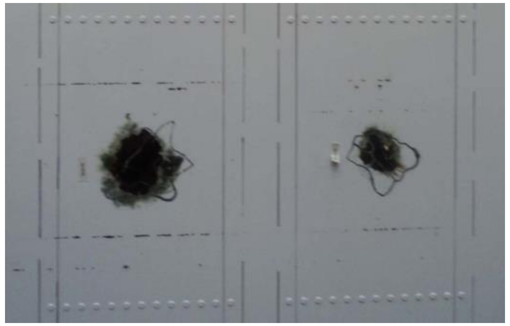
Model & Data: from Airbus internal project & open-source (TBC)

- 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;
  - Labelling 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 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

MLEAP Challenges:

- Automatic detection of external damages and their classification into two types: lighting strike impacts and dents;
- It is an on-going project, materials (metrics, models and data) are still under development
  - Find acceptable metrics to bring computer vision models to human abilities on surface damage detection
- First targeted performance: >95% accuracy correctly detecting damages

Dents Damages (1)

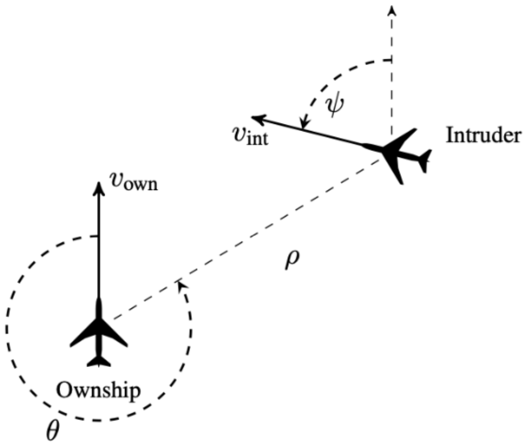


Lightning Strike impacts (2)

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

MLEAP – Use Cases Description > > >

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



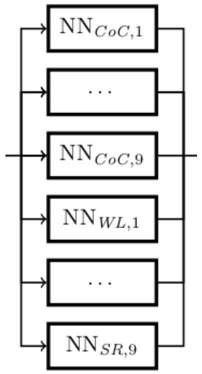
Objective:

ACAS is a universal system-to-system collision avoidance
It issues horizontal turn advisories to avoid an intruder aircraft
Leverage NNs to solve ACAS problems (Bak and Tran, 2022)

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:

- ρ (ft): Distance from ownship to intruder
- θ (rad): Angle to intruder relative to ownship heading
- ψ (rad) : Heading angle of intruder relative to ownship heading direction
- vown (ft/s) : Speed of ownship
- vint (ft/s) : Speed of intruder
- τ (s) : Time until loss of vertical separation



ML model elements of the ACAS Xu system

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 correctly compress LUTs

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

MLEAP Report

First version of the MLEAP deliverable, next and last version in a year.

A nice 260 pages document.

MLEAP – Report – The Topics

Data

Representativeness and Completeness
Corner cases and outliers

Models

Generalization properties

Evaluation

Robustness & stability

MLEAP – Report – The common steps

Definitions

What are the meanings of the terms
What do the various documents (standards, CP, …) define
What meaning do we choose for the report

State of the Art

Review of scientific littérature
Review of existing methods and tools
Construction of selection grids to associate use-cases and methods/tools

360

Experimentation

How the tools and methods actually behave with various data or models
Experiment around scaling
Try with aviation use-cases

Projection into the W-shaped process

Generalize the methodologies as much as reasonably possible
Structure inputs for use by the EASA in their works

MLEAP – Report – The document

Introduction

Use cases

Data: representativeness and completeness

Model development: generalization properties

Model evaluation: robustness and stability

Conclusions

MLEAP – Report – The next steps

Complete the selection grids
Complete the experimentations

Apply the methods to the use cases

Consolidate the projection to the W-shaped process

Listen to the feedback from all of you

ai@easa.europa.eu

AIRBUS

MLEAP – Task #1 milestones: Data Completeness & Representativeness

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

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

Task #1 : Data Completeness and Representativeness

Task #1 objectives (so far)

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

Task #1 : Data Completeness and Representativeness

Task #1 objectives (so far)

- State-of-the-art: Provide a list of factors influencing the choice of tools and approaches in order to assess the completeness and representativeness of databases, with corresponding justifications and bibliographical references.
- Synthesis: Present a draft structure of the selection grid for the assessment tools and methods.

Task #1 : Data Completeness and Representativeness

Task #1 objectives (so far)

- State-of-the-art: Provide a list of factors influencing the choice of tools and approaches in order to assess the completeness and representativeness of databases, with corresponding justifications and bibliographical references.
- Synthesis: Present a draft structure of the selection grid for the assessment tools and methods.
- Testing: Identification or development of efficient and practicable methods and tools for the assessment of completeness and representativeness of data sets (training, validation and test) in the generic case of data-driven ML.

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

State-of-the-art: influence factors identified

Technical requirements

- Intended behavior
- Model architecture
- Data dimensionality
- Intended level of autonomy
- Intended level of performance
- Intended level of robustness and resilience
- Intended level of stability

Processes

- Data Management requirements (specs)
- Data Quality improvement (augmentation…)
- Data synthesis
- Data sampling
- Labelling
- Pre-processing

Other DQRs

- Balance
- Relevance
- Diversity (discriminative power)
- Diversity (absence of non representative sampling bias)
- Currentness

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the state-of-the-art

Assessment of data quality in general lacks maturity in the field of AI:

< 10 works are explicitly considering influence factors in their relationship to Completeness/Representativeness
Influence factors and target properties are not studied in a structured way

Exhaustive data quality of the data set may be actually hard and challenging to attain

Operations required to enhance data quality attributes may be mutually exclusive (e.g. ensuring relevance can be detrimental to representativeness)
Importance of expert contextual trade-off

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the state-of-the-art

In literature, the burden of sorting the wheat from the chaff often still rests on the model.

No "off-the-shelf" method to quantify the relationship between a factor of influence and Completeness/Representativeness.

High-dimensionality challenges rarely addressed. Adaptability of the methods to high-dimensional data needs to be explored.

AIRBUS

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Synthesis: Building the selection grid

80+ sources explored, among which 60+ assessment methods analysed

20 methods selected for testing

Sufficient maturity

In line with the project objectives

Technical requirements

- Intended function
- Model architecture
- Data dimensionality
- Intended level of autonomy
- Intended level of performance
- Intended level of robustness and resilience
- Intended level of stability

6 methods selected (from 11 identified)

Processes

- Data Management requirements (2 methods)
- Data Quality improvement (3 methods)
- Data synthesis (1 method)
- Data sampling (1 method)
- Labelling (2 methods)
- Pre-processing

11 methods selected (from 33 identified)

Other DQRs

- Balance (1 method)
- Relevance
- Diversity (discriminative power)
- Diversity (absence of bias) (1 method)
- Currentness (1 method)

3 methods selected (from 18 identified)

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

- Test task: Classification
- Associated UC: ACAS-Xu
- Test data sets
  - ACAS-Xu
  - Gas sensor array (external)

AIRBUS

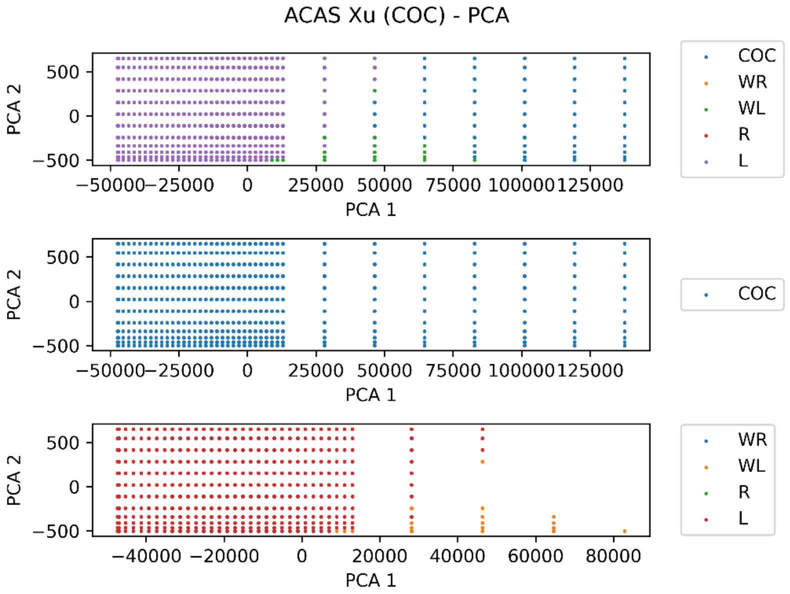
MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA
- Test task: Classification
- Associated UC: ACAS-Xu
- Test data sets
  - ACAS-Xu
  - Gas sensor array (external)

PCA
- PCA highlighted particularities of the ACAS-Xu dataset
- Triggered further investigations
- Note: ACAS-Xu is probably at the edge of relevance for this method



MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

Entropy (Shannon)

- Test task: Image Segmentation
- Associated UC: AVI
- Test data sets
  - CIFAR-100 (external)
  - ROSE (LNE)

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

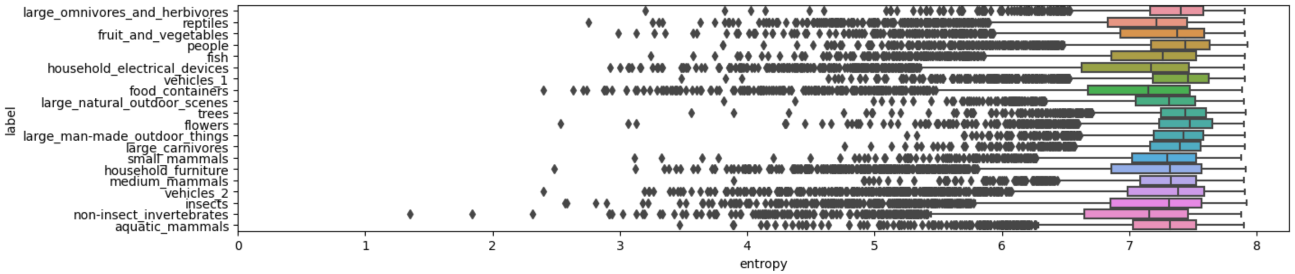
PCA

Entropy (Shannon)

- Test task: Image Segmentation
- Associated UC: AVI
- Test data sets
  - CIFAR-100 (external)
  - ROSE (LNE)

Entropy

- Can be used at different level (label-wise, pixel-wise…)
- Provides coarse-grain information
- Should preferably be combined with other metrics (yet to be determined)



AIRBUS

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

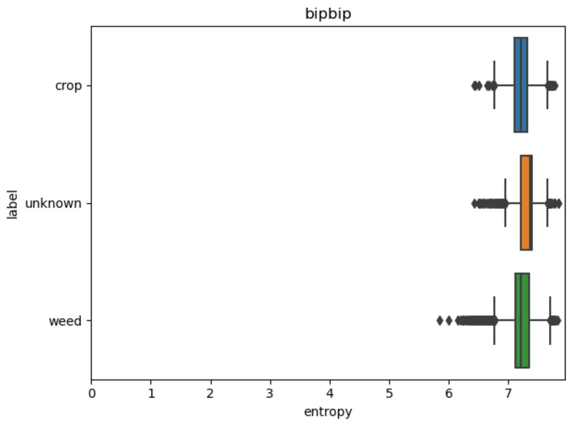
PCA

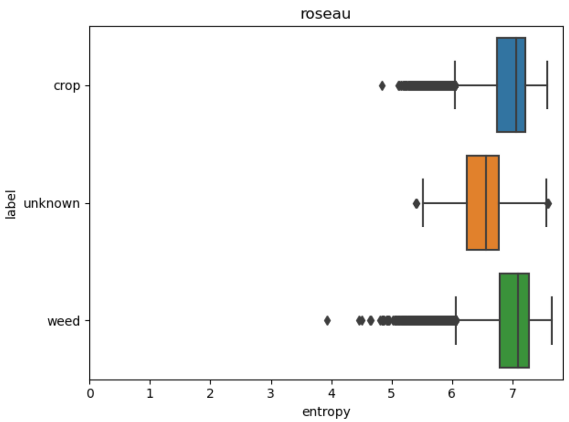
Entropy (Shannon)

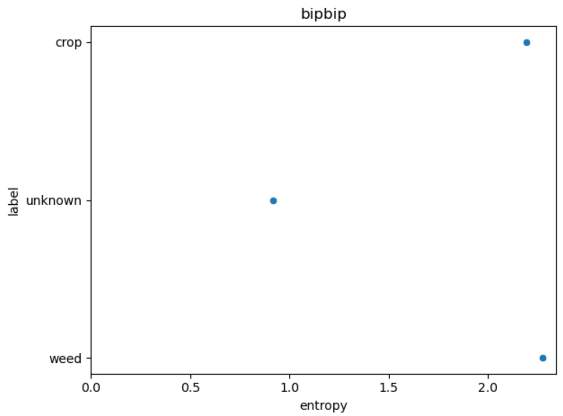
- Test task: Image Segmentation
- Associated UC: AVI
- Test data sets
  - CIFAR-100 (external)
  - ROSE (LNE)

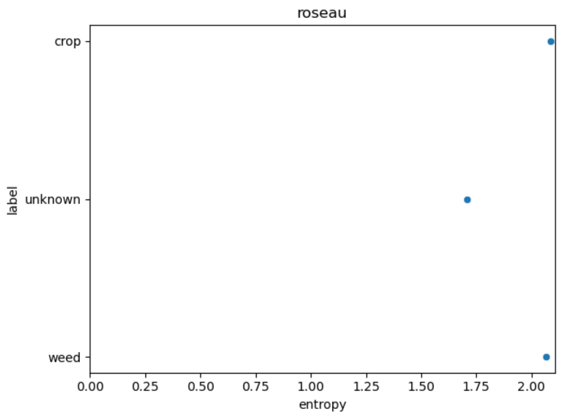
Entropy

- Can be used at different level (label-wise, pixel-wise…)
- Provides coarse-grain information
- Should preferably be combined with other metrics (yet to be determined)









MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

Entropy

Graph (feature combination distribution)

- Test task: Classification
- Associated UC : ACAS-Xu
- Test data set
  - Titanic (external)

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

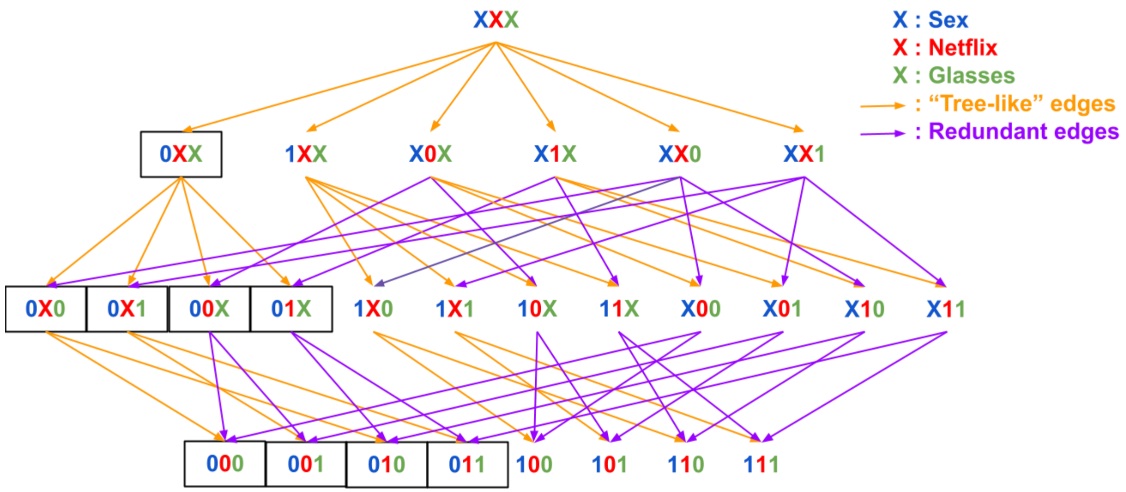
Entropy

Graph (feature combination distribution)

- Test task: Classification
- Associated UC : ACAS-Xu
- Test data set
  - Titanic (external)

Graph

- Results are easy to interpret
- Can be used as an efficient visual tool (like PCA)
- Must be tested at scale



MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

Entropy

Graph (feature combination distribution)

Sample similarity (Degree of Correspondence)

- Test task: Speech recognition
- Associated UC: ATC-STT
- Test data set
  - Fluent Speech Commands (external)

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Testing Phase

PCA

Entropy

Graph (feature combination distribution)

Sample similarity (Degree of Correspondence)

- Test task: Speech recognition
- Associated UC: ATC-STT
- Test data set
  - Fluent Speech Commands (external)

Sample similarity

- DoC is inconsistent and intractable on speech embeddings
- This specific method will be put aside
- Similarity-based analysis remains interesting

MLEAP – Task #1 Technical Feedback > > >

Task #1 : Data Completeness and Representativeness

Main take aways of the testing phase

No method is **self-sufficient**
They need to be combined to provide meaningful insight

No method is **universal**
The method and their combination must be tailored to each type of task/data

Completeness and representativeness can only be estimated w.r.t ODD specifications
No "absolute measure"

Trade-off between completeness and representativeness for e.g. corner cases

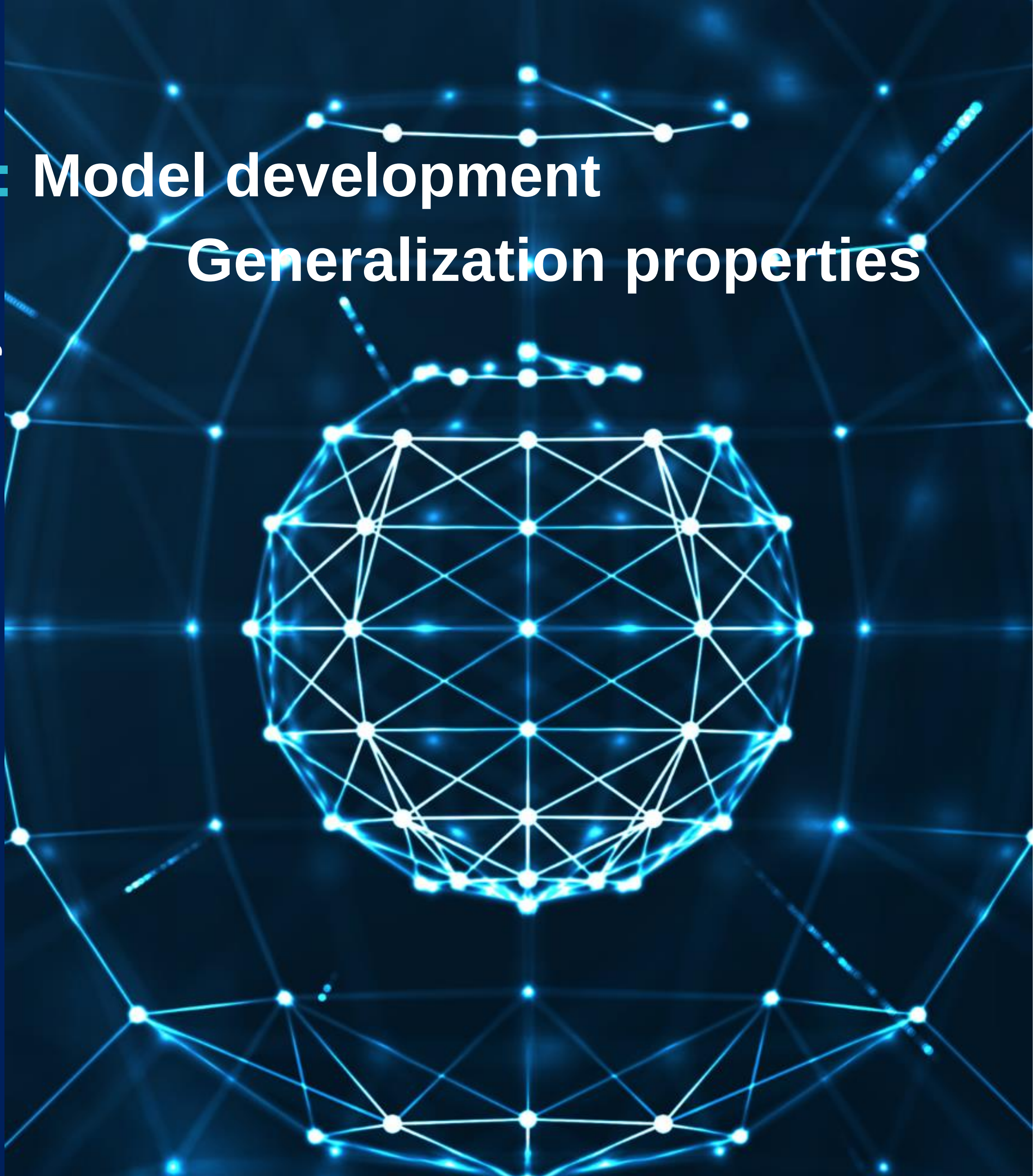
Task #1 : Data Completeness and Representativeness

Next step for Task 1

Adapting identified methods to work on high scale datasets

Continue to test methods of the selection grid

MLEAP – Task #2 Milestones: Model development Generalization properties



State-of-the-art analysis:

Available methods and tools to evaluate generalization bounds;
Barriers in generalization guarantees: ML and DL;
Limitation of available methods and common practices;

Identification/selection of suitable methods:

Methods selection;
Projection into the W-shaped approach: ML development pipeline;

Experimentation & Evaluation

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

Task #2 : Model generalization

Supervised machine learning

Objective: Estimate the response y from the data x

x  What is this animal? y = cat

$x \longrightarrow$ **Parameterised algorithm** $\longrightarrow y$

(x_i, y_i) Examples

Training: optimize algorithm parameters to minimize errors on the examples

Machine learning:

- Approximation
- Optimization
- Estimation

Generalization: We are expecting few errors on unseen data. It is based on the assumption that we have regularities behind the data .

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

Task #2 : Model generalization

Generalizability

Definition

Model's ability to generalize the learned knowledge to a new context or environment

Success estimator

Statistical tools that estimate how well the model generalizes to unseen data

For $\delta \in (0,1)$ the generalizability of model $\hat{f} \in F$ on w.r.t. data set D is:

$$G(\hat{f}, D) \leq \sqrt{\frac{func(model\ class\ F\ complexity) + \log(1/\delta)}{\|D_{train}\|}}$$

Success indicator

- Evaluation-based: Good performances (w.r.t. some criteria) for Dtest ≠ Dtrain
- Testing-based: correctness of results during adversarial attacks and spot failure modes

Generalization Guarantees		Algorithm Dependent: Yes	Algorithm Dependent: No
Data Dependent	Yes	• PAC-Bayesian • PAC-Bayesian bounds for NNs (+) more precise, better distributional properties of the learning algorithm	• Rademacher Complexity (RC) • RC and regularized Empirical Risk Minimization (ERM) (+) better estimation
Data Dependent	No	• Model Compression • Based on Model Distillation (-) do not take into account data features (+) focuses on the model enhancement	• VC-dimension • VC-dimension for NNs (-) Not practical for particular use-cases (Dar et al., 2021) (+) widely applicable

Data Dependent		• Statistical guarantees
		  ○ Data statistics
		  ○ Error gradient during training
		• Geometry analysis bounds (combining input, output spaces and the mapping)	

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

Task #2 : Model generalization

Generalization Bounds

Objective: bounding the deviation of the true risk of the learned hypothesis from its empirical measurement

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

Several bounds are defined in the littérature based on different theoretical framework, such as:

- Uniform convergence based (Sharpness-based measures)
- Uniform stability based
- Algorithm robustness based
- Mutual information
- Measures related to the optimization procedures

For example: bound based on VC dimension

$$\boldsymbol{\varepsilon \rightarrow \varepsilon(\mathcal{H}, m, \delta) \;\sim\; \mathcal{O}\left(\sqrt{\frac{VCdim(\mathcal{H}) + ln(\frac{2}{\delta})}{m}}\right)}$$

Algo.	Ref.	Bound
CNN	(Lin and Zhang, 2019)	$R_{\mathcal{D}}(F_C) \leq \hat{R}_{S,l_\eta}(F_C) + \mathcal{O}\left(\left(\frac{\lVert X\rVert_F \mathcal{R}_C}{\eta}\right)^{\frac{1}{4}} n^{-\frac{5}{8}} + \sqrt{\frac{\ln(1/\delta)}{n}}\right)$
RNN	(Chen et al., 2019)	$R(f_t) \leq \hat{R}(f_t) + \tilde{O}\left(\frac{L \times Complexity}{\sqrt{m}} + B\sqrt{\frac{\log(1/\delta)}{m}}\right)$
NN for classification	(P. Jin et al., 2020)	$\mathcal{E}(f) \leq \frac{\sqrt{d}.(1-\rho_\tau)}{\min(\delta_0, \kappa\delta_\tau)} = \alpha(\tau).CC(\tau)$
NN	(Alquier, 2021)	Catoni's bound (PAC Bayes) $\mathbb{P}_S\left(\forall\rho \in \mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{KL(\rho\Vert\pi) + log\frac{1}{\epsilon}}{\lambda}\right) \geq 1-\epsilon$
	(Alquier, 2021) (McAllester, 1998)	Mc Allester's bound $\mathbb{P}_S\left(\forall\rho \in \mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \sqrt{\frac{KL(\rho\Vert\pi) + \log\frac{1}{\epsilon} + \frac{5}{2}\log(n) + 8}{2n-1}}\right) \geq 1-\epsilon$

17 bounds selected based on:

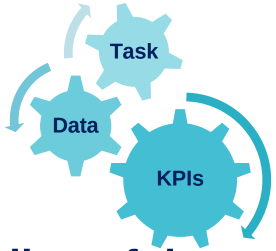
- genericity of the bound
- Use cases applicability

AIRBUS

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

Task #2 : Model generalization

AI issues: from problem analysis to model release



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

Good model

Misunderstanding of the generalization bounds

- Some norm-based measures negatively correlate with generalization
- Conventional bounds based on uniform convergence or uniform stability are inadequate for over-parameterized models

Common mistakes and pitfalls in practice

- Inappropriate training objective
- Inappropriate data representation, volume, split (train, test, valid), quality (noisy, high sparsity)
- Inappropriate model complexity to perform the task, and evaluation metrics

Gap between expectations from evaluation vs the real-world application

- How far away the empirical assessment reflects the reality about the model efficiency?
- Appropriate performance indicators to the application domain cannot ALWAYS be translated by existing evaluation metrics
- How to define a good model ? what constitutes a good AI/ML model?
- What about the uncertainty tolerance: how a 85% accuracy is good? how the 15% uncertainty is tolerable ?
- How the final model will behave in the target system/environment?

Unhandled ML/DL testing limitation and challenges

- How to define exhaustively the testing scenarios? How to deal with "black boxes in DL"?

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

Task #2 : Model generalization

Towards application independent ML development pipeline to promote generalizability

Objectives

Deal with models overfitting/underfitting in industry

- Regularization techniques, training adaptation (warm-up and fine-tuning)
- Model/Network architecture and complexity adequacy with the target task

Bridge the gap between experimentation and industrial expectation

- Adopt a multicriteria/additional validation phases;
- Include KPIs (industrial target performance) in the learning objectives and the evaluation metrics as well
- Leverage ML testing properties to promote the quality assurance and help to identify defects and flaws

Better handle the OOD samples and reduce the impact on the safety of the AI system

- Deal with rare cases with high impact on the confidence of the model, in order to minimize the risks.

Build an enhanced data and model development pipelines reducing the impact of common practices and pitfalls that result in a weak generalization ability of an ML/DL model, after release/implementation

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

Task #2 : Model generalization

Target application definition

$$D \xrightarrow{T_f} (X, Y), \quad with$$
$$T_f(d) = (x, y), \quad and \;\; x \in X \;\; and \;\; y \in Y$$

$$T = \begin{cases} f \;\in F\,, & (1) \\ X:\; x_i = \left[x_i^0, \dots, x_i^n\right], & (2) \\ Y:\; y_j = \left[y_j^0, \dots, y_j^m\right], & (3) \\ f\colon X \xrightarrow{f(x)} Y, & (4) \\ M = \{m_1, \dots, m_k\}, & (5) \\ B = \{b_1, \dots, b_l\}, & (6) \\ b_t\big(m_t \circ f(x_t)\big) \in \{0, 1\} & (7) \\ E = \{e_1, \dots, e_z; (e_i \odot x_i) = x'_i\} & (8) \end{cases}$$

1) The selected model
2) The input space
3) The output space
4) The mapping function
5) Set of SMART objectives & metrics to evaluate their achievements
6) Verification scheme and target performances validity/acceptance indicators
7) Benchmarking of the model w.r.t (6)
8) Elements and/or conditions that directly impact the inputs, and hence the outputs after implementation.

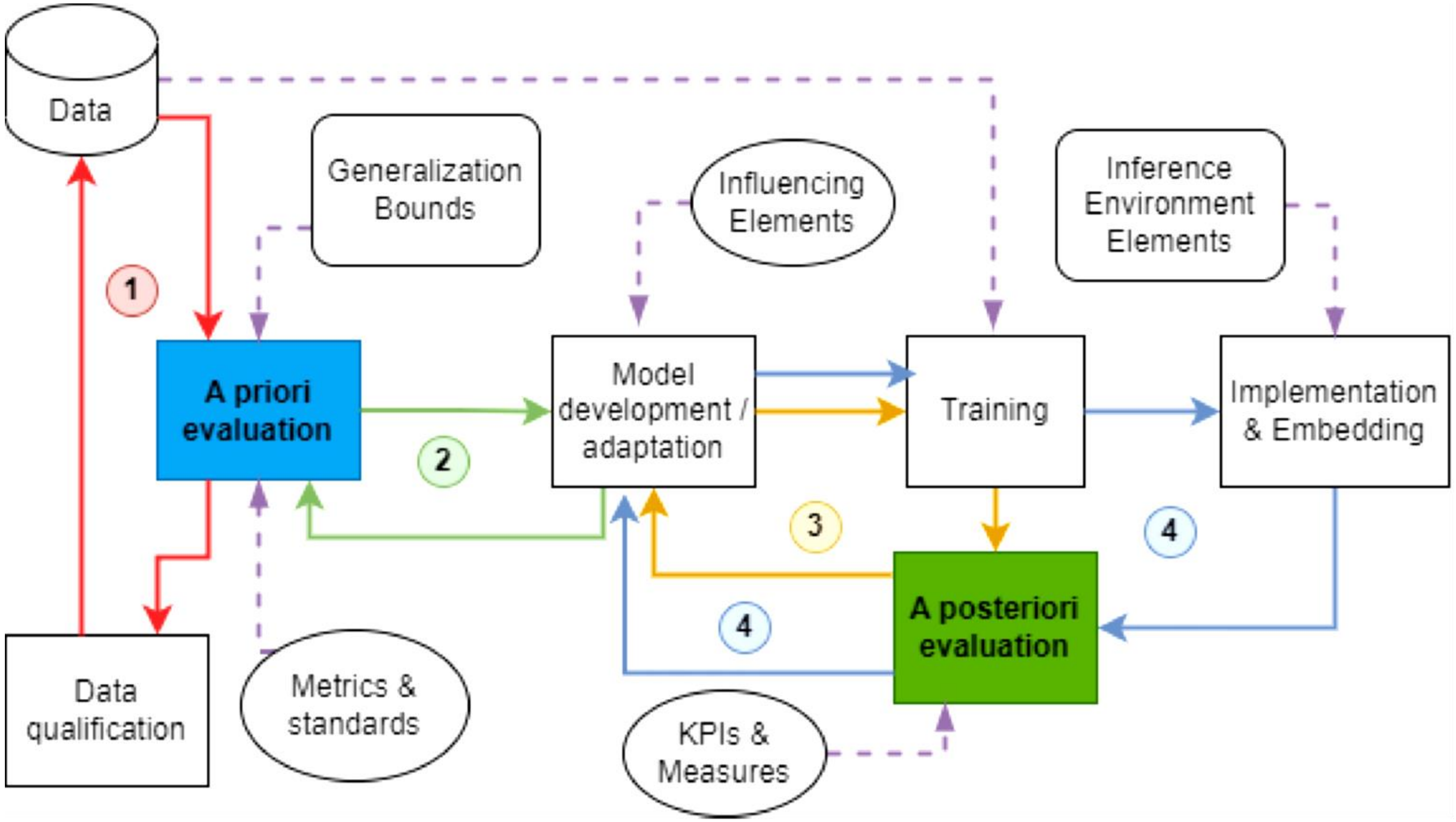
MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Task #2 : Model generalization

ML Pipeline / W-shaped process projection

(1) Data evaluation and qualification (<=> Task#1)

a. Minimal size of data set needed
b. Data quality evaluation (completeness, representativeness)
c. Enhancement operations: data augmentation, processing, cleansing, balancing, and splitting;



AIRBUS

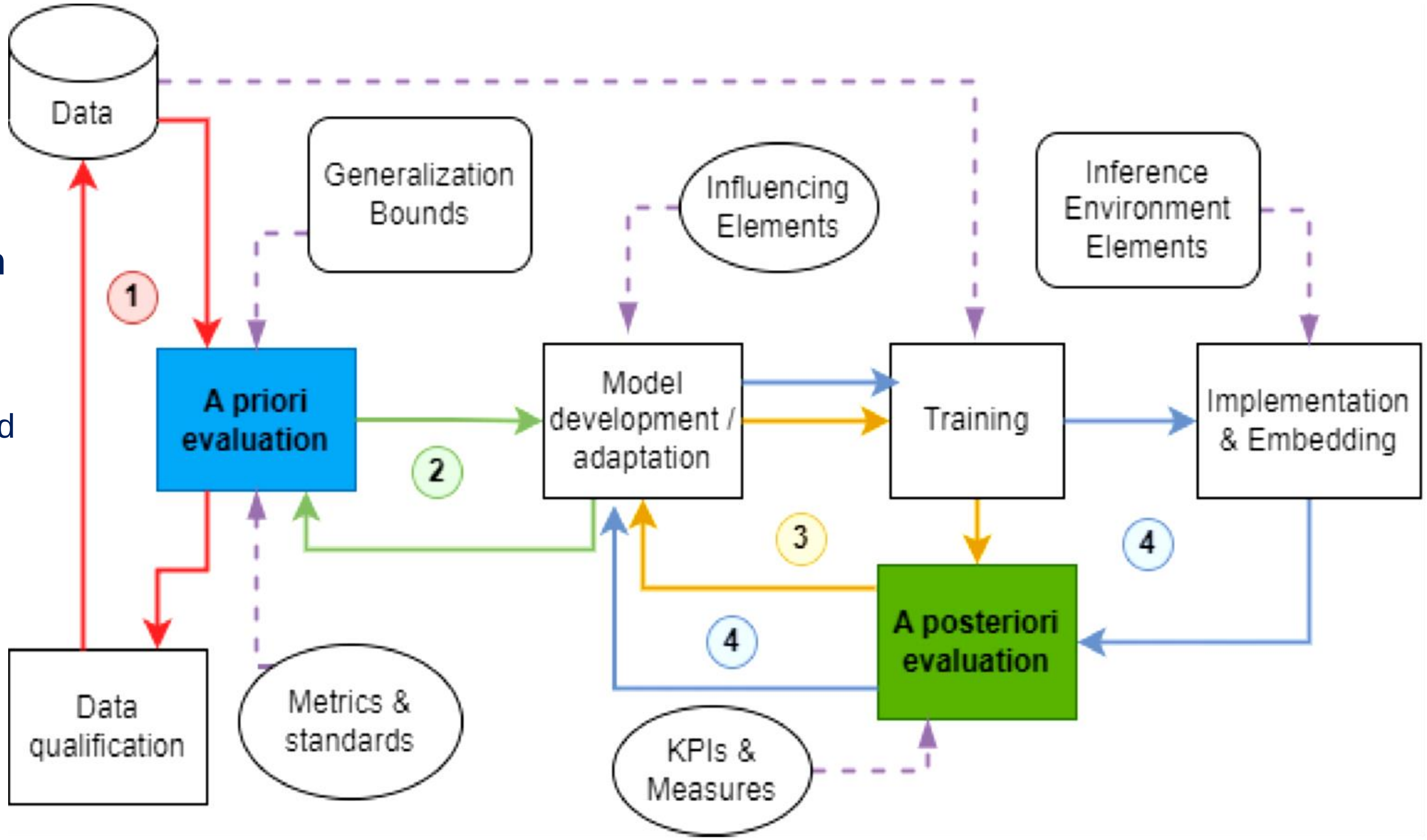
MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Task #2 : Model generalization

ML Pipeline:

(2) Model development and adaptation

a. Data Constraints: data size and type, alignment, balance …
b. The mappings between the inputs and outputs
c. Generalization bounds estimation ;



AIRBUS

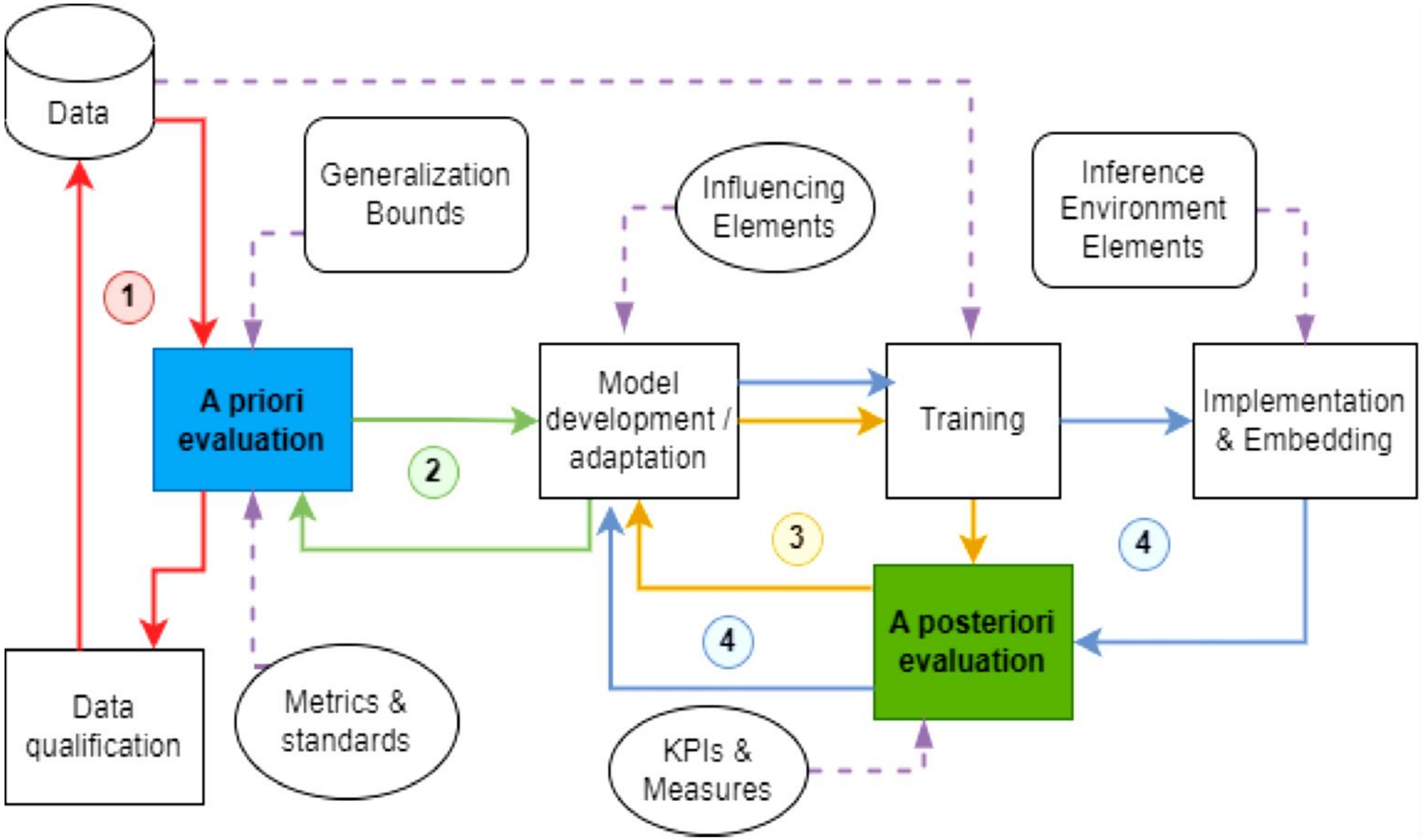
MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Task #2 : Model generalization

ML Pipeline:

(3) Model training on the optimized data set (<=> Task#3)

a. Benchmark including a set of industrial KPIs
b. Adapted evaluation measures/metrics/thresholds
c. A posteriori evaluation of the trained model: generalization & robustness
d. Measures and loss functions should be adapted to meet the target application objectives



AIRBUS

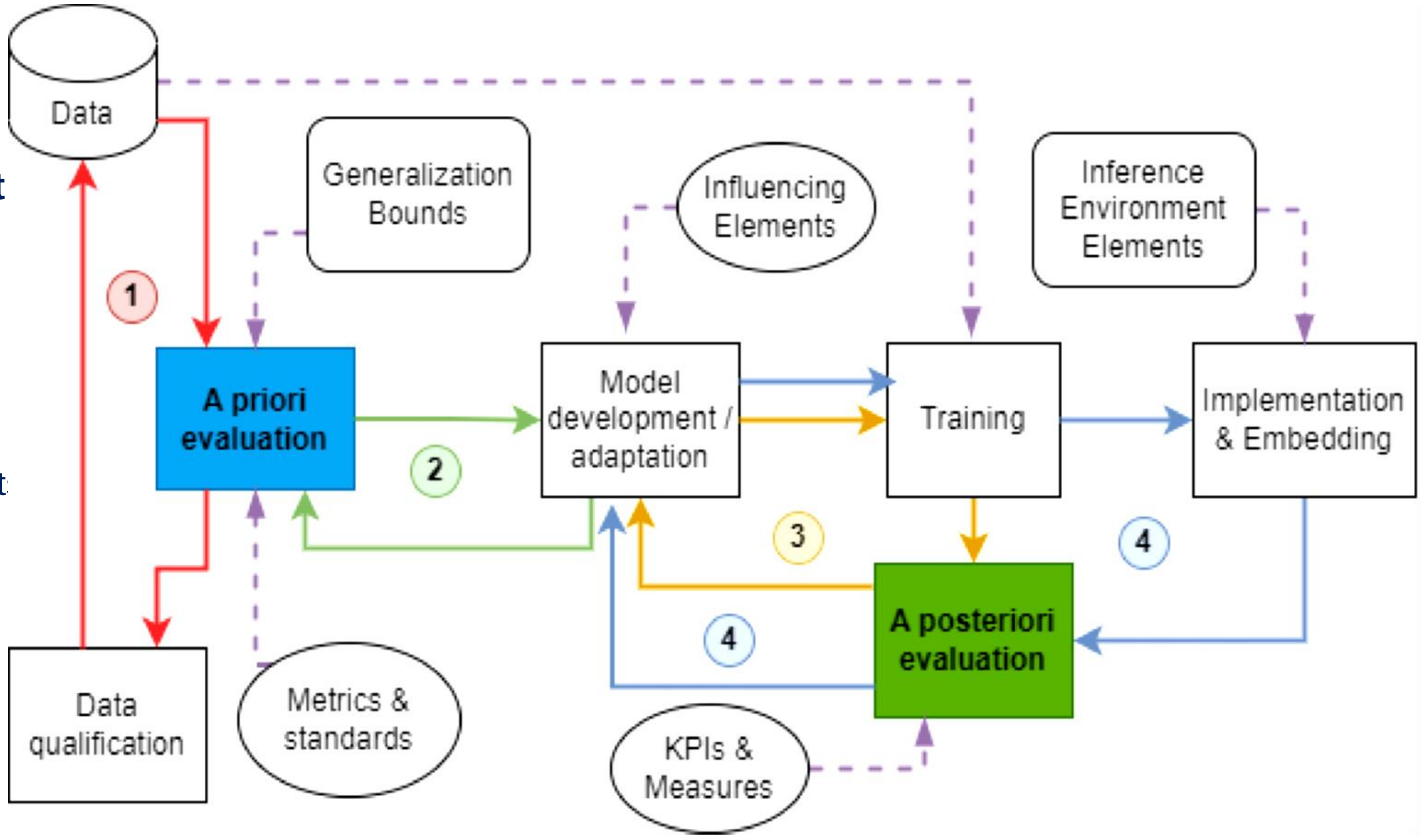
MLEAP – Task #2 Milestones : Model development – Generalization properties > > >

Task #2 : Model generalization

ML Pipeline:

(4) Performance verification in the target environment

a. Verify the performances after implementation
b. Different environment and system elements impacting performances
c. System/target performance requirements are involved
d. Possible step-back if important drop in performances



AIRBUS

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

Task #2 : Model generalization

Evaluation Objectives

1. Analysis of the development pipelines to identify the limitations;
2. Analysis of the data quality & volume w.r.t. target performance;
3. Analysis of the evaluation schemes: metrics, KPIs, training objectives …;
4. Compare the estimated generalizability VS the real performances;
5. Make suggestions: metrics, data OPs, methods to improve existing results and pipelines;
6. Validate the suggestions on real-life use-cases.

Experimentation: ATC-STT Task – Models evaluation

Datasets:

– AIRBUS dataset (real ATC exchange from French airports)

– Open-source datasets (from European airports)

Models:

– AIRBUS model, based on the Vosk API & Kaldi (no Deep Learning), trained on AIRBUS dataset (FR accent included)

– Open-source DL models, based on a transformers architecture, trained on the open-source datasets, fine-tuned in AIRBUS data

Evaluation metric:

– Word Error Rate (WER) = $\frac{S+D+I}{N}$

- S is the number of substitutions
- D is the number of deletions
- I is the number of insertions
- N is the total number of words in the reference

– Accuracy = 1-WER

AIRBUS

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

Task #2 : Model generalization

Experimentation: ATC-STT Task – Performance Comparison

AIRBUS Vosk Model vs DL Open-source

Excellent performances of the AIRBUS model on the AIRBUS data set and **poor** performances on open-source data sets

Possible bias:

→Source of data (from a few French airports)
→Audio quality (noise, microphone used,…)
→Model architecture and implementation (Vosk API & Kaldi)

Open-source models are trained on larger data sets, and their complexity is more important, but performance on AIRBUS data are **average.** **High-performance** on open-source data, regardless of the recording context (accent, noise, etc.) and therefore more robust

Transfer Learning

→Zero-shot evaluation
→Fine-tuning on the AIRBUS data

Models \ Data		AIRBUS data set	ATCO2	ATCOSIM	UWB
AIRBUS Model Vosk		11.50%	91.05%	95.05%	63.46%
Zero-shot	Jzuluaga/wav2vec2-large-960h-lv60-self-1	34.63%	36.27%	6.82% ★	20.46% ★
	Jzuluaga/wav2vec2-large-960h-lv60-self-2	34.89%	37.14%	22.98%	19.69% ★
Fine-tuned	Jzuluaga/wav2vec2-large-960h-lv60-self-1	15.13%	35.81%	15.85%	30.96%
	Jzuluaga/wav2vec2-large-960h-lv60-self-2	In progress			

Zero-shot evaluation results showing averaged WER of the models

★ *Means that the model is trained in the Dtrain part of the data set*

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

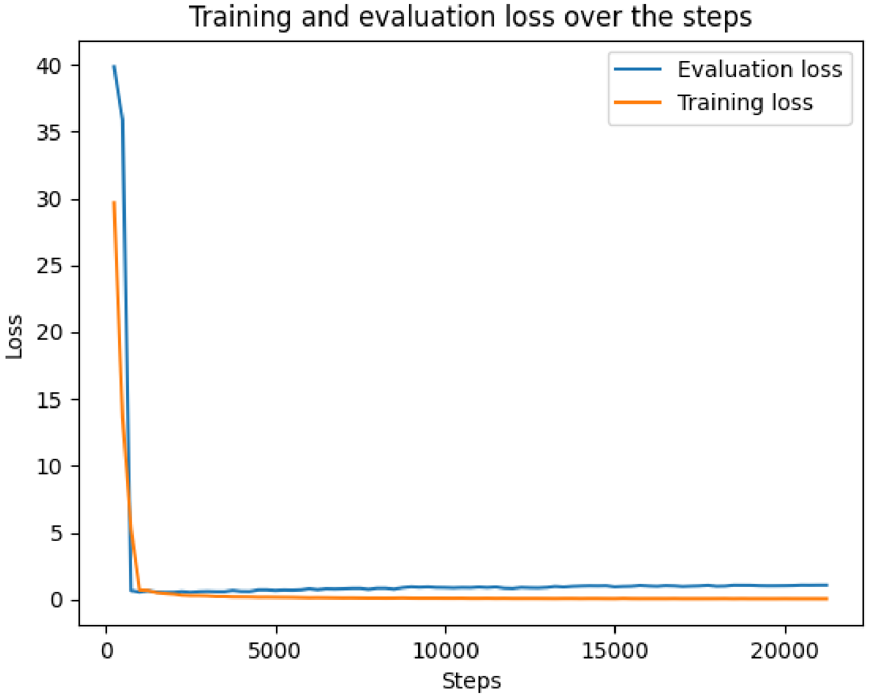
Task #2 : Model generalization

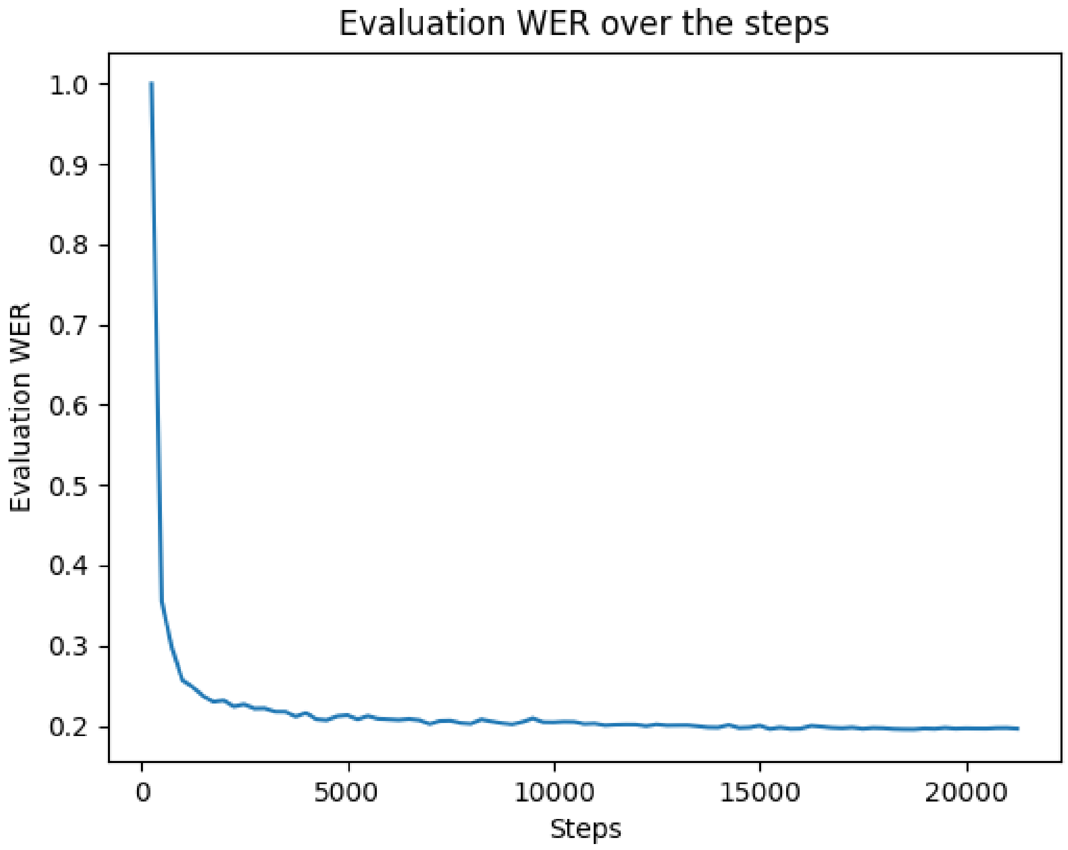
Experimentation: ATC-STT Task – Performance Comparison

Fine-tuning configuration:

AIRBUS data set : 6 826 utterances (~5h11)

50 training epochs, batch size=16





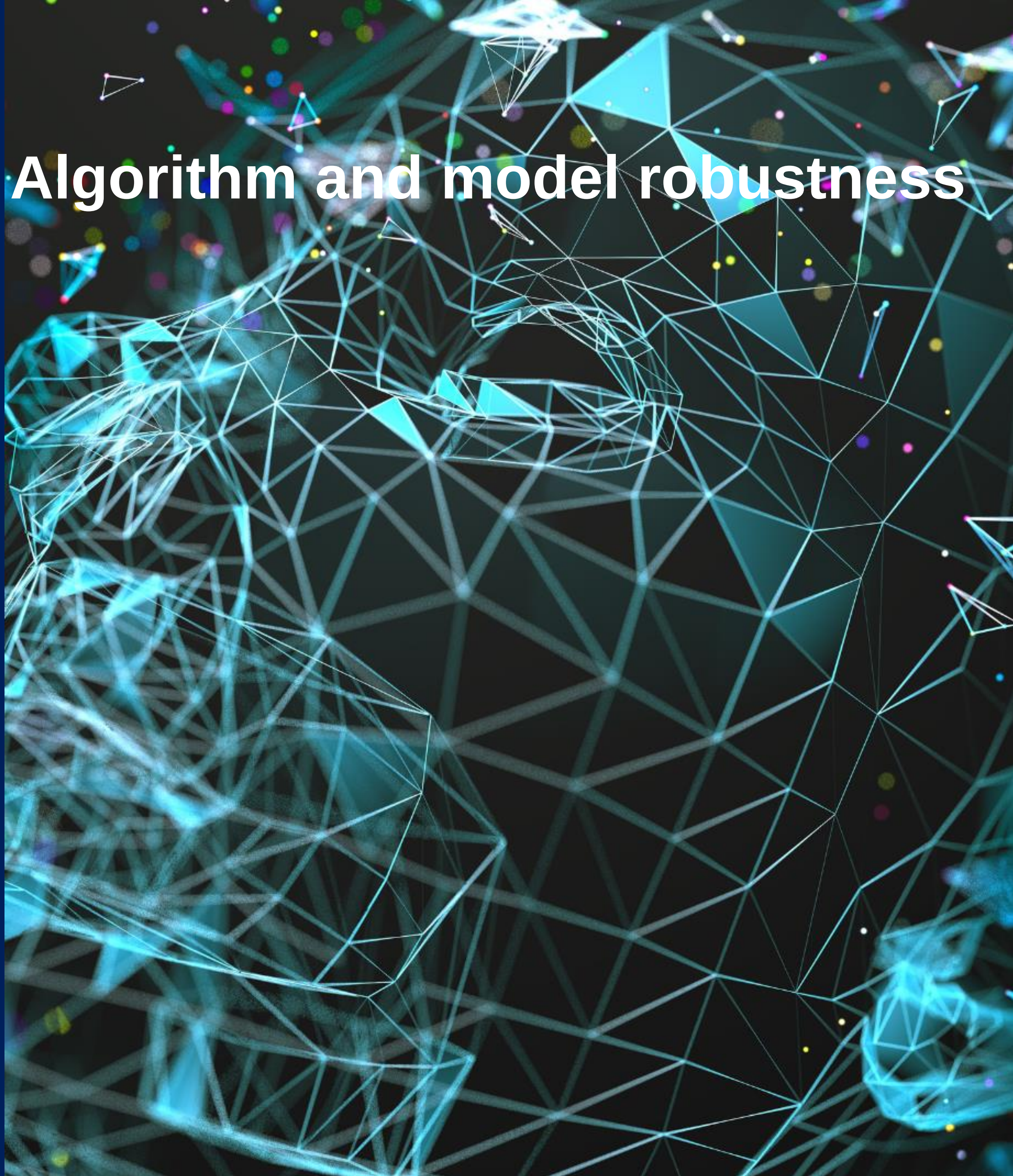
Task #2 : Model generalization

Next step for Task 2

Experimentation and Evaluation:

General framework development and tests of identified bounds and methods

MLEAP – Task #3 Milestones: Algorithm and model robustness



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> > >

Task #3 : 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> > >

Task #3 : 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

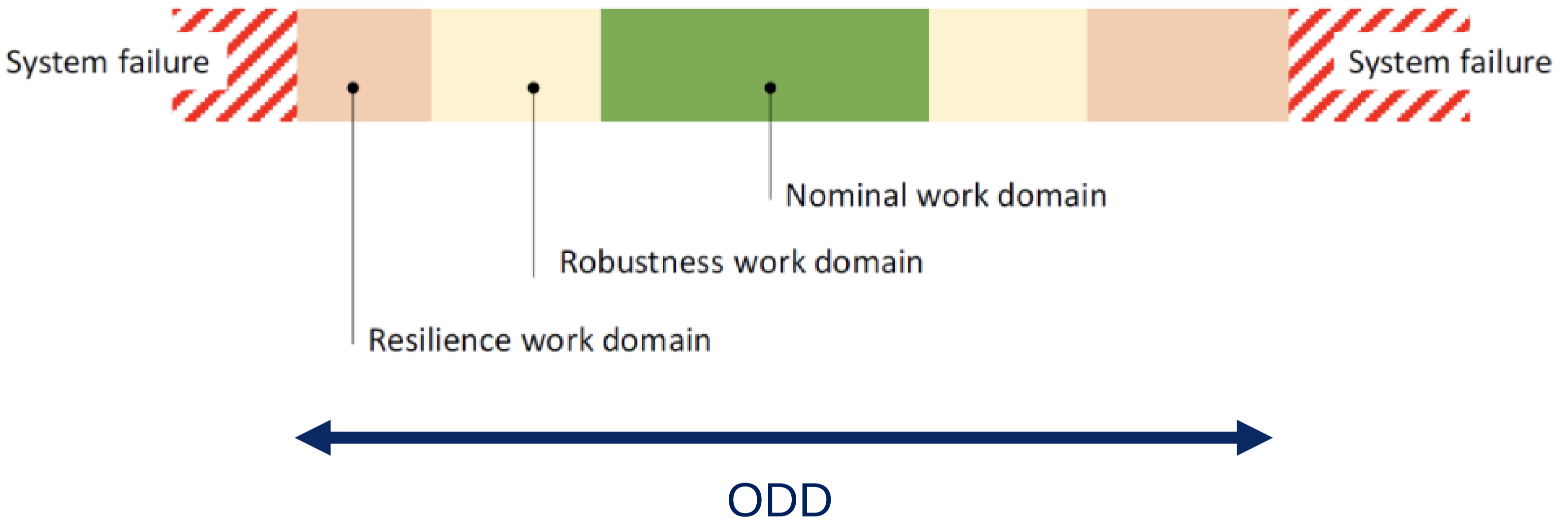
A non-existent case

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

Task #3 : 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> > >

Task #3 : Algorithm and model robustness

Different ways of defining the concept

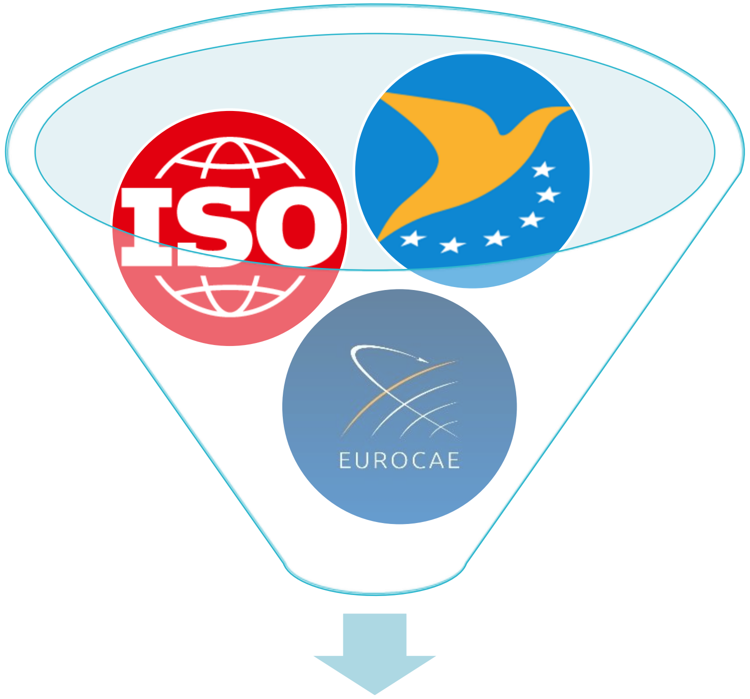
Aligning several sources of the state of the art

- 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> > >

Task #3 : 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)	$bias^2(\mathcal{F}, n) = \mathbb{E}_{x \sim \mathcal{X}}[(\overline{f_n}(x) - f(x))^2]$
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	$\mathcal{E}^n\left(x, \hat{f}^n(x)\right) \notin Z$

AIRBUS

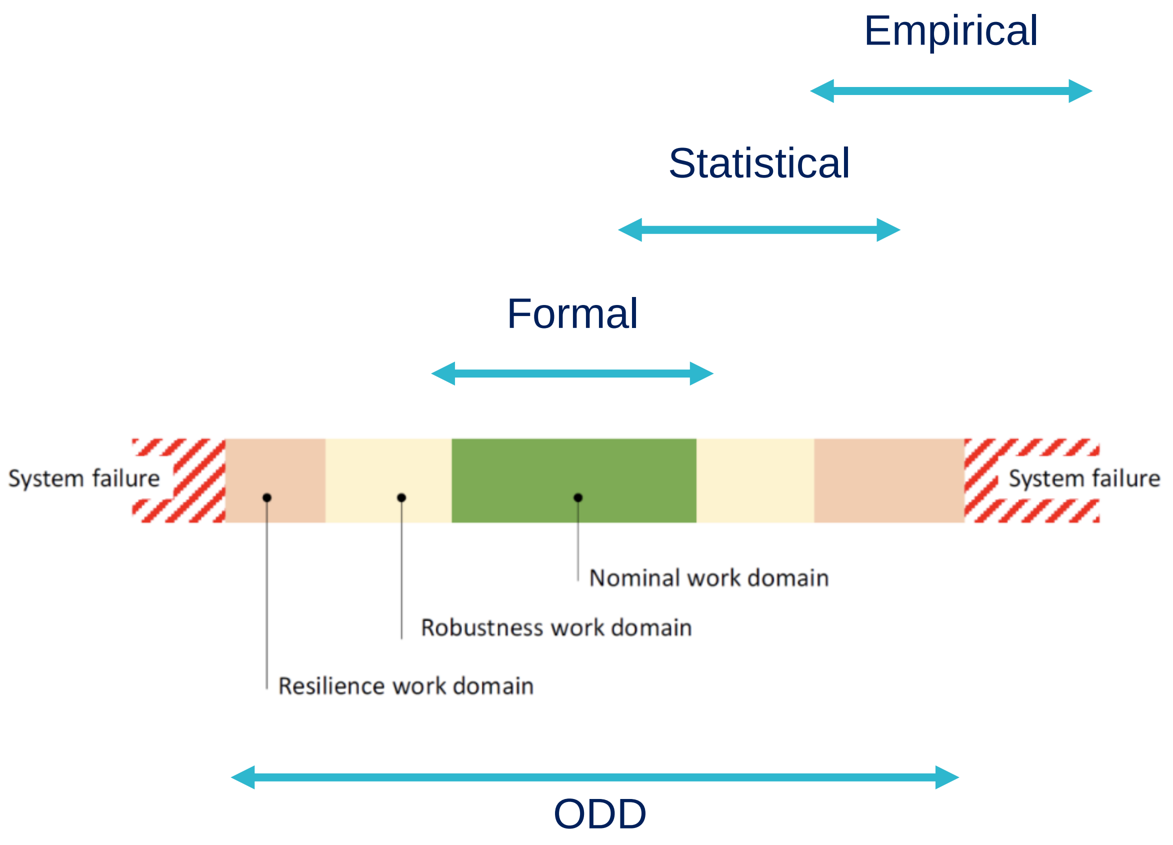
MLEAP – Task #3 Milestones: Algorithm and model robustness> > >

Task #3 : 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 requirements
- …but varying degree of scalability



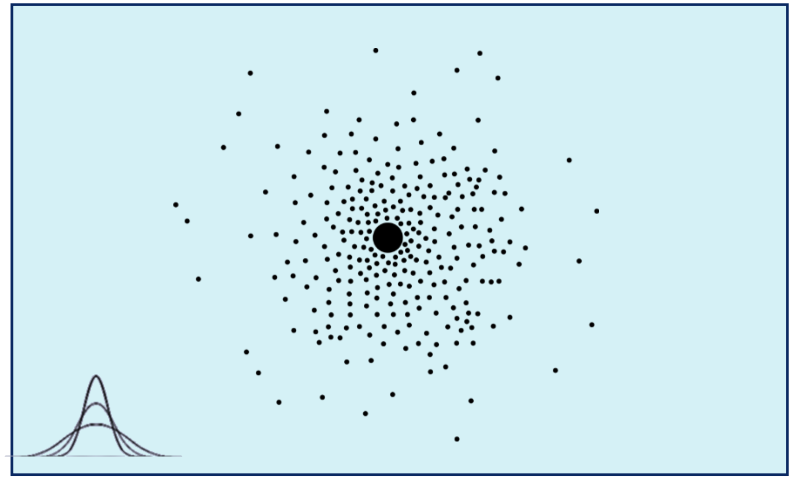
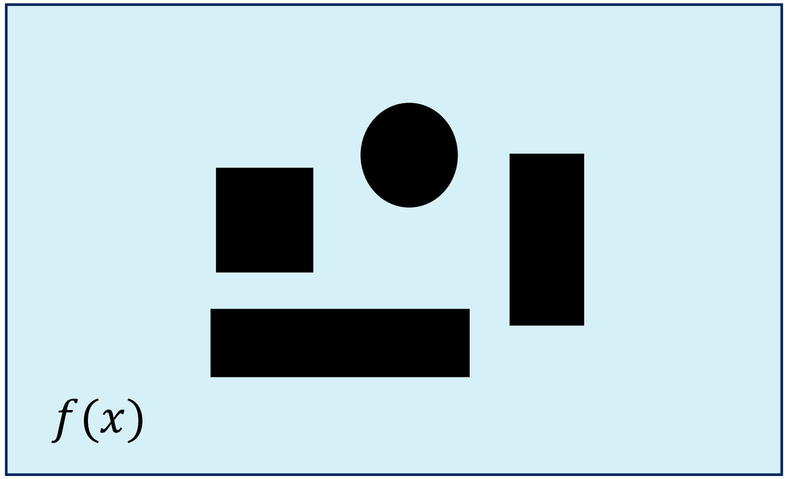
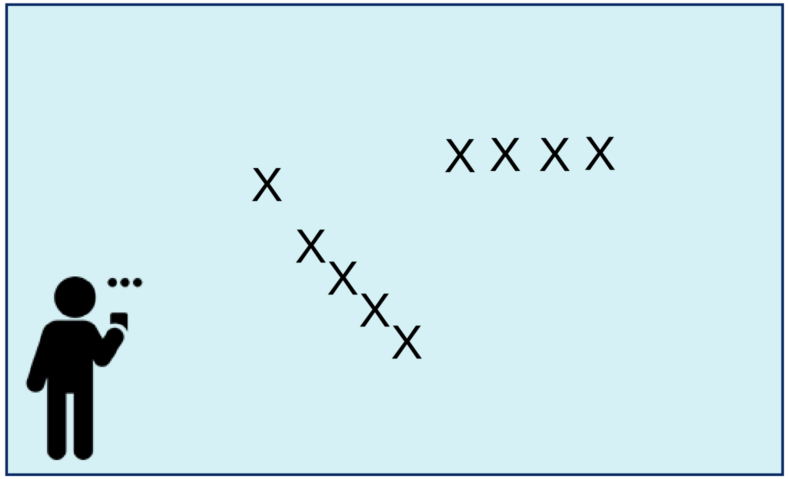
Picture from "DEEL White Paper on Machine learning in Certified System (DEEL Certification Workgroup, 2021"

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

Task #3 : Algorithm and model robustness

3 approaches at a glance

Each allow specific advantages and drawbacks

Statistical	Formal	Empirical
		
Easy to setup Rely on data sets	Local guarantees High dimensional sub-space	Require human intervention Experimental protocol

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

Task #3 : 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> > >

Task #3 : Algorithm and model robustness

A priori assessment of suitability

	Empirical methods	Statistical methods	Formal methods
Stability of the training algorithm			
Stability of the trained model			
Stability of the inference model			
Bias			
Variance			
Relevance			
Reachability			
Corner case exploration			

Scalability	Human intervention needed	Doable but through sampling	Doable but locally
Methods	• Field trial • A posteriori • Benchmarking	• Combining metrics	• Solver • Abstract interpretation • Optimization

AIRBUS

Task #3 : Algorithm and model robustness

Next step for Task 3

Applying a panel of suitable approaches on the different use cases to exemplify the guidance

What’s next for MLEAP Project?

Airbus Protect / Artificial Intelligence Days
At Paris Air Show
June 21st >>

June 21st:

Awareness session conference:

from 10 to 11am at Airbus Pavillon - On Invitation only

Knowledge Sharing Conference & Networking:

from 3 to 5pm - at VIPARIS Conference center - Conference Room N°2 -

Our partners:

numalis LABORATOIRE NATIONAL DE MÉTROLOGIE ET D'ESSAIS LNE

EASA European Union Aviation Safety Agency

AIRBUS

WHAT's next for MLEAP?

PROJECT:

MLEAP Final report in 1 year from today

EVENTS:

January 2024: MLEAP Stakeholders day #3

Awareness session conference #2

April 2024: Knowledge sharing conference #2

May 2024: 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

https://events.airbus.com/airbus-protect-easa-paris-air-show/

{Thank you}

NETWORKING LUNCH!

Let’s keep the party going!

Thank you for your participation!

Let's continue the discussion until 14:00 around a lunch sponsored by AIRBUS PROTECT

Any question?

Please contact us: ai@easa.europa.eu