Machine Learning for Civil Aviation

Focus on Robustness against Evasion Attacks

Machine Learning in Civil Aviation

Why Machine Learning?

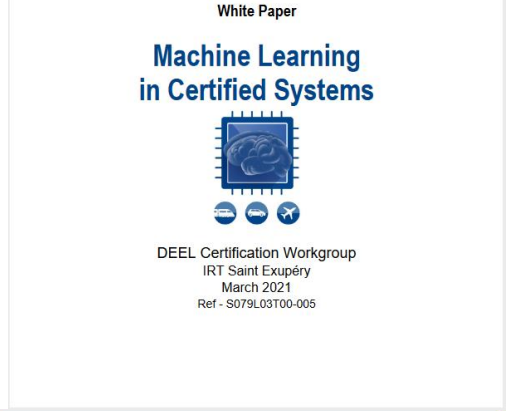
Safety-critical communities **do not** like new software methods

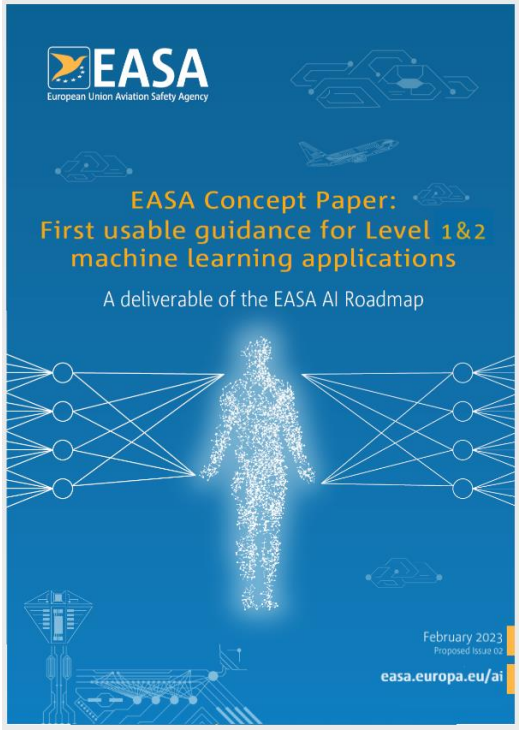
Example: Object-Oriented Programming (e.g., polymorphism) for avionics

→ only in 2011 (DO-332, supplement to DO-178C)

ML solves previously **intractable** problems (natural language processing, object recognition) by **extracting complex correlations in huge datasets and exploiting them in real time**.

Some examples in civil aviation:



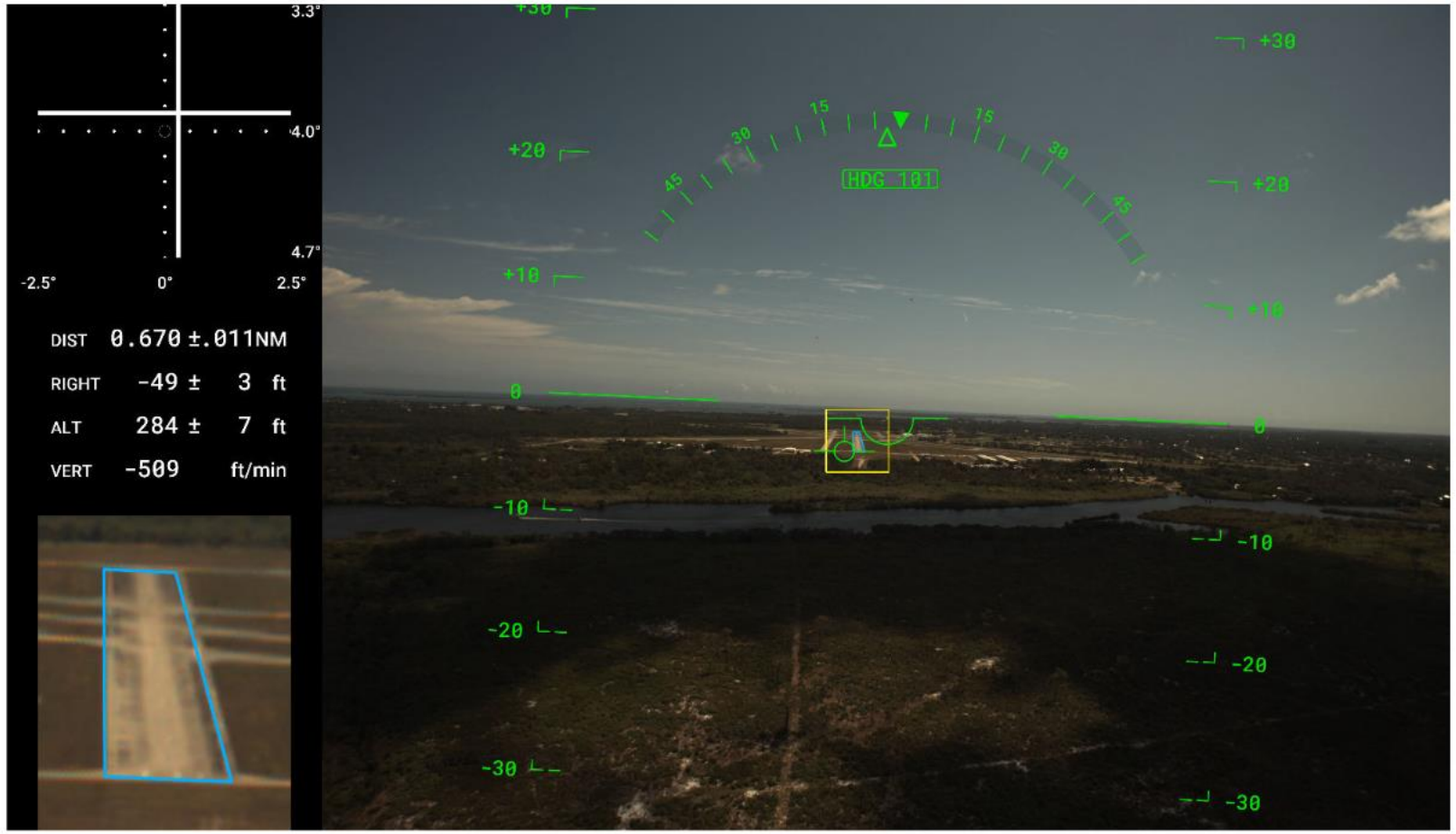


Machine Learning in Civil Aviation

Aircraft design and operations

- *Visual Landing System*

FAA WJH Technical Center, Aviation Research Division. Neural Network Based Runway Landing Guidance for General Aviation Autoland. DOT/FAA/TC-21/48. 2021



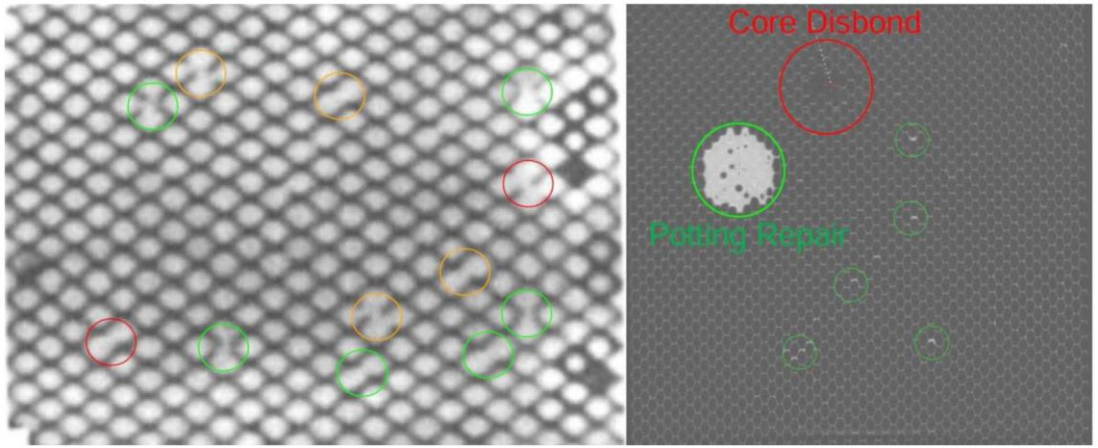
Machine Learning in Civil Aviation

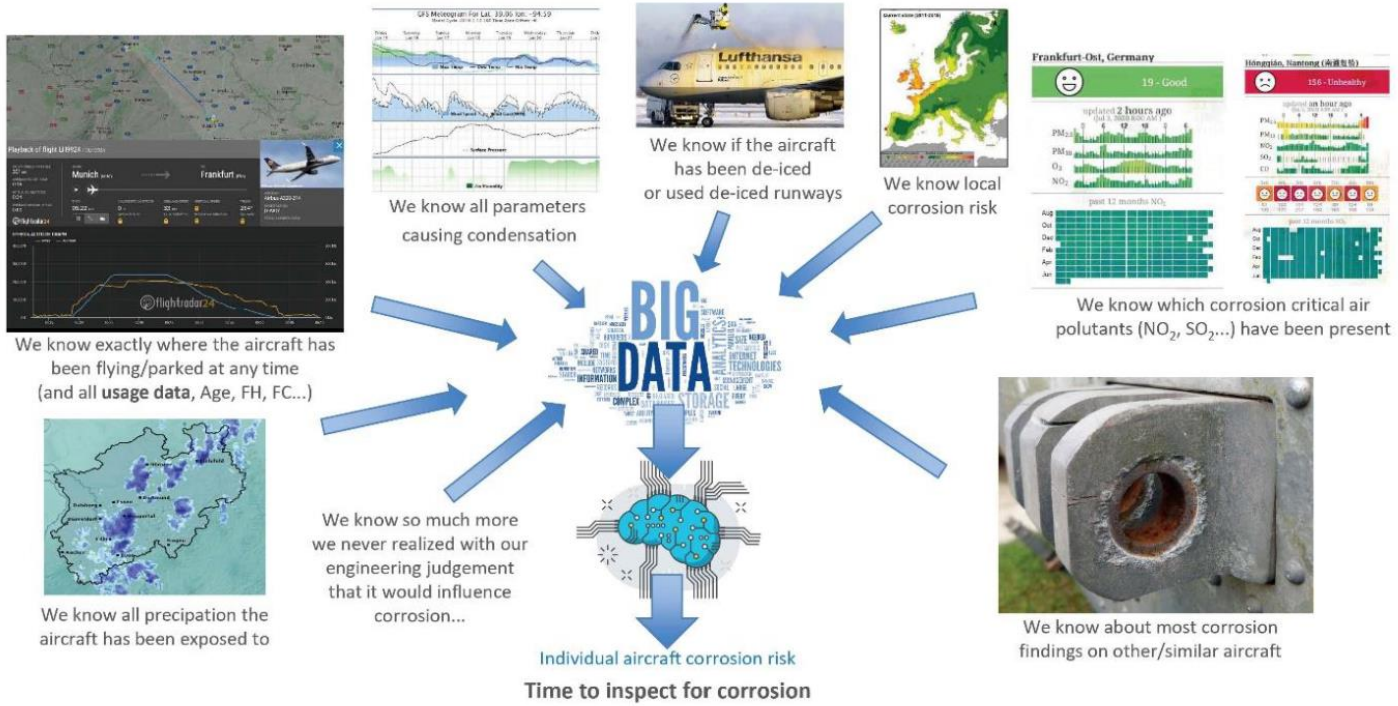
Aircraft design and operations

- *Voice recognition and suggestion of radio frequencies*
- *Single-pilot operations with a virtual co-pilot (Pilot AI teaming)*

Aircraft production and maintenance

- *Controlling corrosion by usage-driven inspections*
- *Damage detection in images*





Machine Learning in Civil Aviation

Air Traffic Management / Air Navigation Services

- *AI-based augmented 4D trajectory prediction (climb & descent rates)*
- *Congestion prediction for Extended ATC Planner with LSTM*

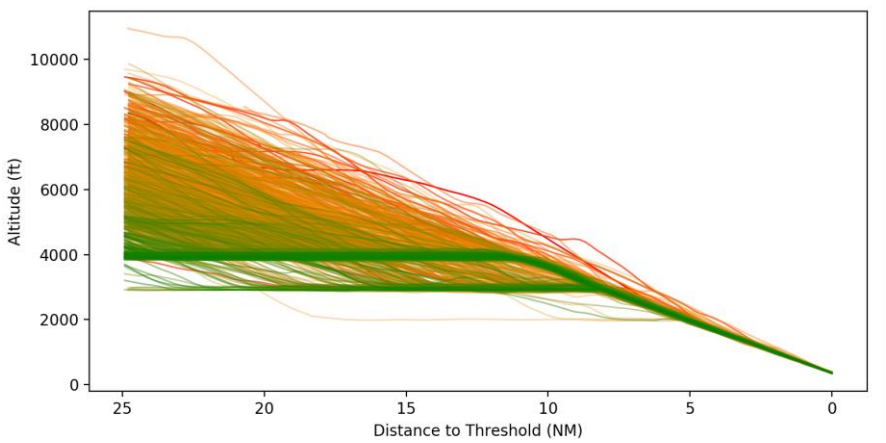
L. Shi-Garrier, D. Delahaye, N. C. Bouaynaya. Predicting Air Traffic Congested Areas with Long Short-Term Memory Networks. 14th USA/Europe ATM R&D Seminar, 2021

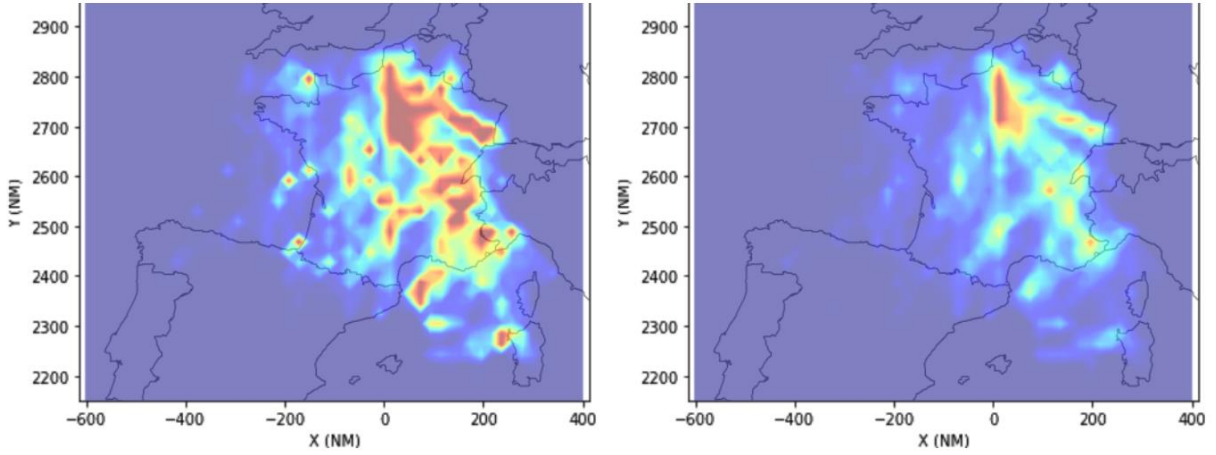
- *Detection of abnormal approaches with GAN*

G. Jarry, N. Couellan, D. Delahaye. On the use of generative adversarial networks for aircraft trajectory generation and atypical approach detection, EIWAC 2019

- *Air traffic structuration with reinforcement learning*

P. Juntama, S. Chaimatanan, D. Delahaye. Air Traffic Structuration based on Linear Dynamical Systems. 10th SESAR Innovation Days, 2020





Machine Learning in Civil Aviation

Aerodromes

- *AI-based screening for airport security systems*
- *Detection of foreign object debris on the runway*

Urban Air Mobility

- *ACAS Xu*

G. Katz, C. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, Computer Aided Verification, 2017

Overview

1. Concepts and challenges for **Trustworthy AI/ML** in aviation

2. Focus on evasion attacks and defenses

3. Robustness against evasion attacks: verification of ML models

A brief overview of the certification process

Airworthiness Regulation Requirements
Federal Airworthiness Requirements / Certification Specification (EASA)

FAR/CS 25.1309 "Equipment, Systems and Installations"

AC/AMC 25.1309 "System design and analysis"
Advisory Circular – Acceptable Means of Compliance

Airworthiness Standards for Safety

Law

Aircraft & System Level

Recommended Practices for system safety assessment

ED-79A (ARP4754A)
"Certification considerations for highly-integrated or complex aircraft systems"

ED-135 (ARP4761)
"Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment"

Component Level

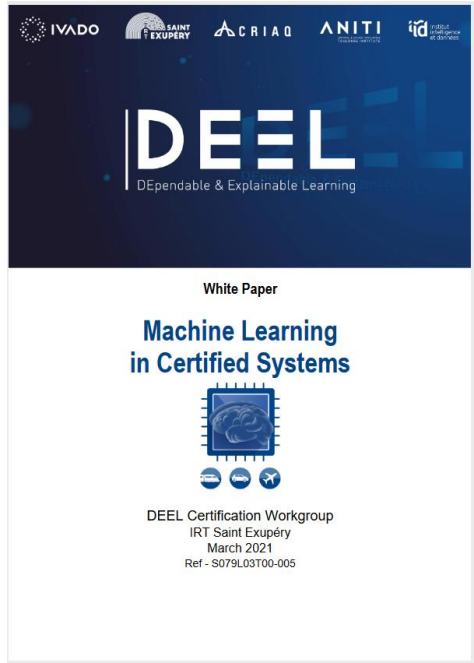
ED-80 (DO-254)
"Design assurance guidance for airborne electronic hardware"

ED-12C (DO-178C)
"Software considerations in airborne systems and equipment"

ED-14G (DO-160G)
"Environmental conditions and test procedures for airborne equipment"

Industrial Rules to build compliance

Guides



Limitations of current standards w.r.t. ML

Traditional standards are **requirement-based** through the V-cycle

- High-level functions are refined into requirements allocated to subsystems and items
- Each line of code must be traced back to the requirements

Avionic systems should safely perform their intended function under all foreseeable operating and environmental conditions.

ML models are **data-driven**

- including **supervised**, **unsupervised**, and **reinforcement** learning

ML models are **black-box**

- State-of-the-art models are **"neural" networks** trained with **gradient descent** (differentiable composition of linear maps and non-linear "activations")

ML models are **non-deterministic**

Challenges of ML certification

- **Specifiability**: correct and complete capture of item requirements
  - Does the data correctly represent the Operational Design Domain (ODD)?
  - Robustness
- **Traceability**: relationship between item requirements and code (learned parameters)
  - Requires **explainability/transparency** to be confident that the model implements the intended function correctly and safely
- **Innocuity**: no unintended behavior

Vidot et al. Certification of embedded systems based on Machine Learning: A survey, arXiv, 2021.

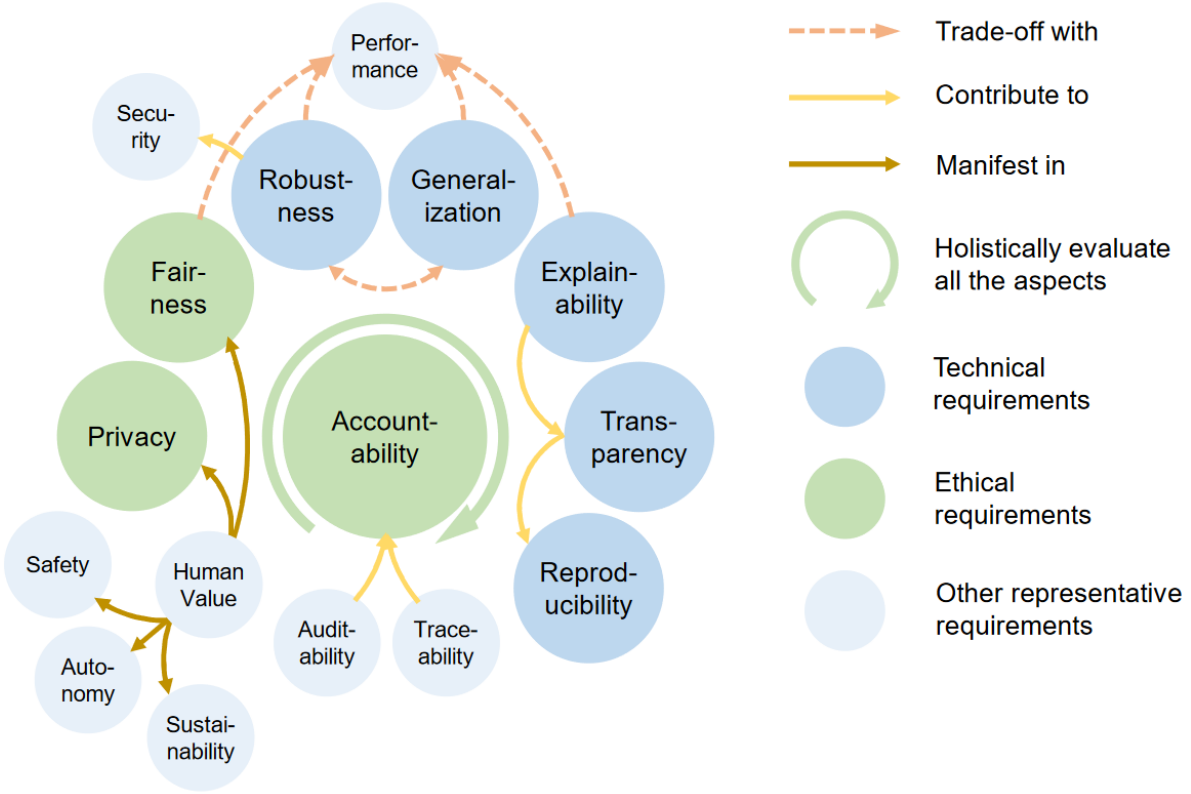
The core concept: **Trustworthy AI**

What is AI trustworthiness?

Trustworthiness = “**certification**” + “**explanation**”

... and **generalization**, **data management**, **ethics**, **accountability** etc.

→ Trade-off

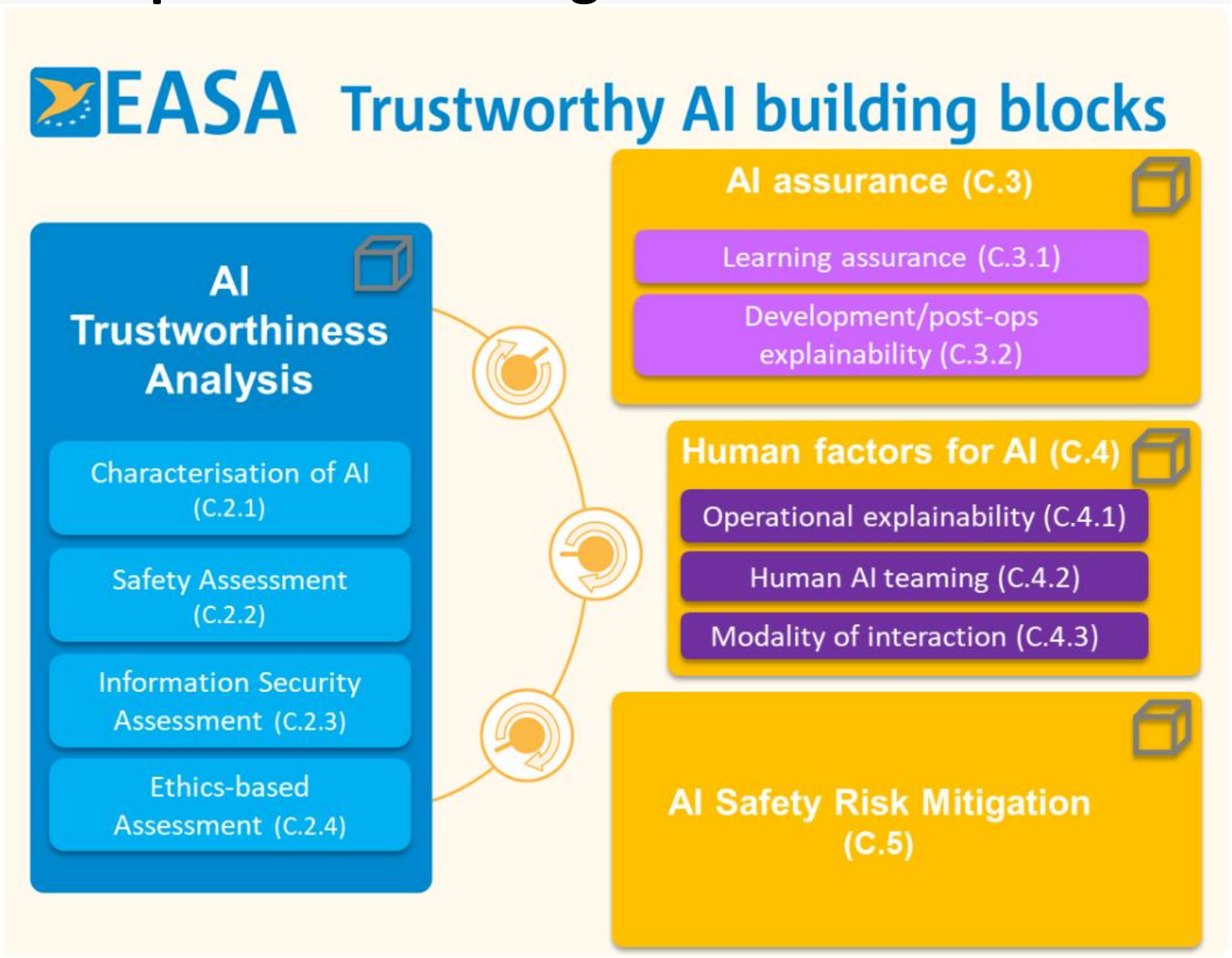


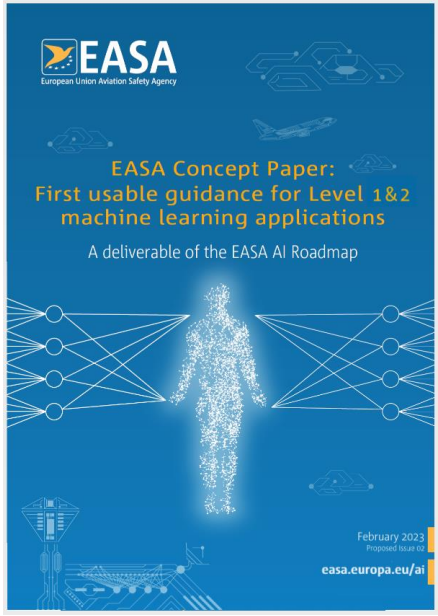
Huang et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Computer Science Review, 2020.

Li et al. Trustworthy AI: From Principles to Practices, ACM Computing Surveys, 2023.

Trustworthy AI: EASA AI Roadmap

Focus on **offline supervised learning**





EASA classification of AI applications

Level 1 AI : assistance to human	Level 2 AI : human/machine teaming	Level 3 AI : more autonomous machine
• Level 1A: Human augmentation • Level 1B: Human cognitive assistance in decision and action selection	• Level 2A: Human and AI-based system cooperation • Level 2B: Human and AI-based system collaboration	• Level 3A: The AI-based system performs decisions and actions, overridable by the human. • Level 3B: The AI-based system performs non-overridable decisions and actions.

Robustness

Ability of a system to maintain its level of performance under all foreseeable conditions.

Robustness in adverse conditions (outside the ODD)

- Edge cases
- Out-of-distribution samples
- Distribution shift

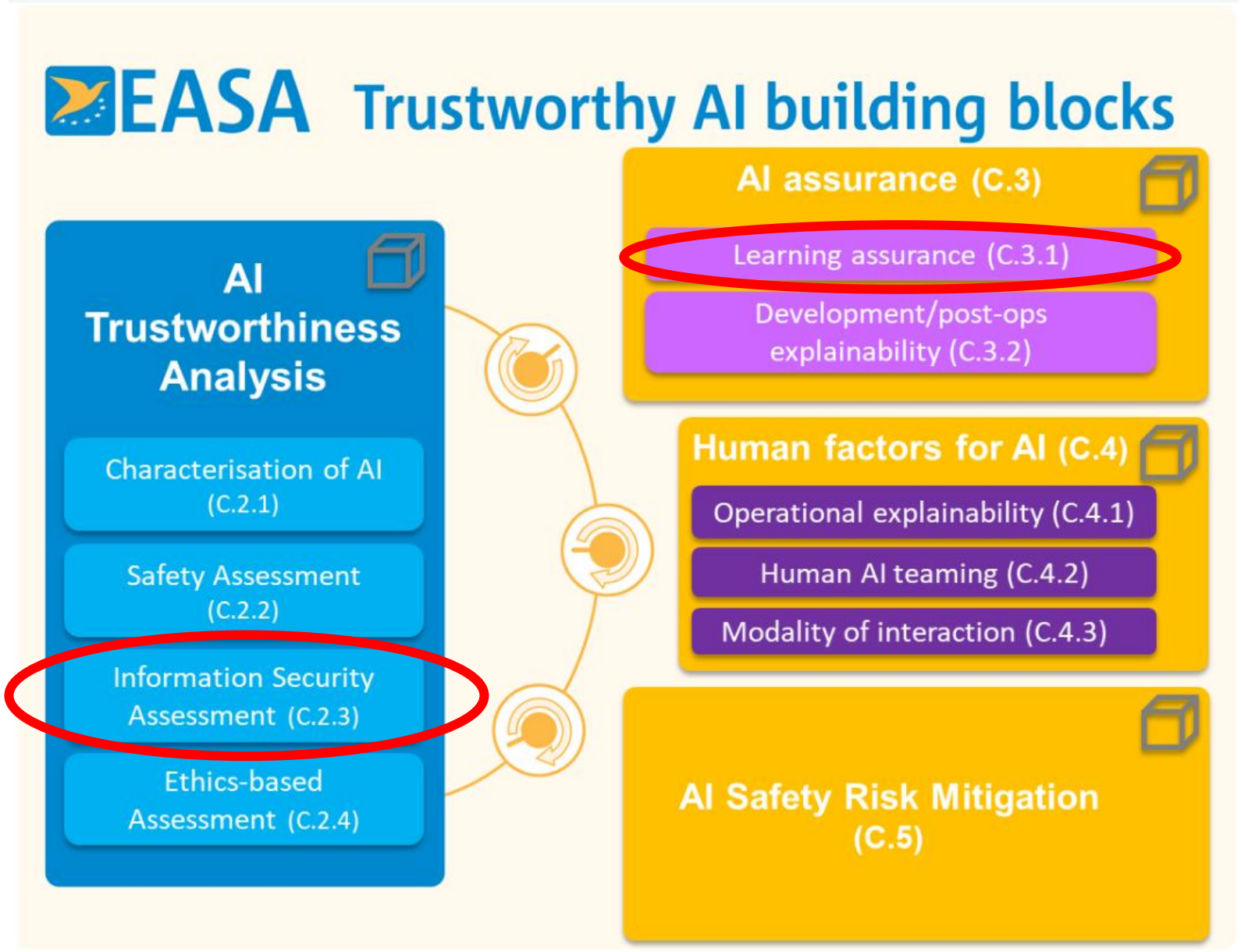
Model stability (inside the ODD)

"Adversarial attacks" = instability to small perturbations

May be caused by: noise, sensor fault, human error, deliberate attack …

Stability w.r.t. parameters and hyperparameters (Embeddability)

Robustness

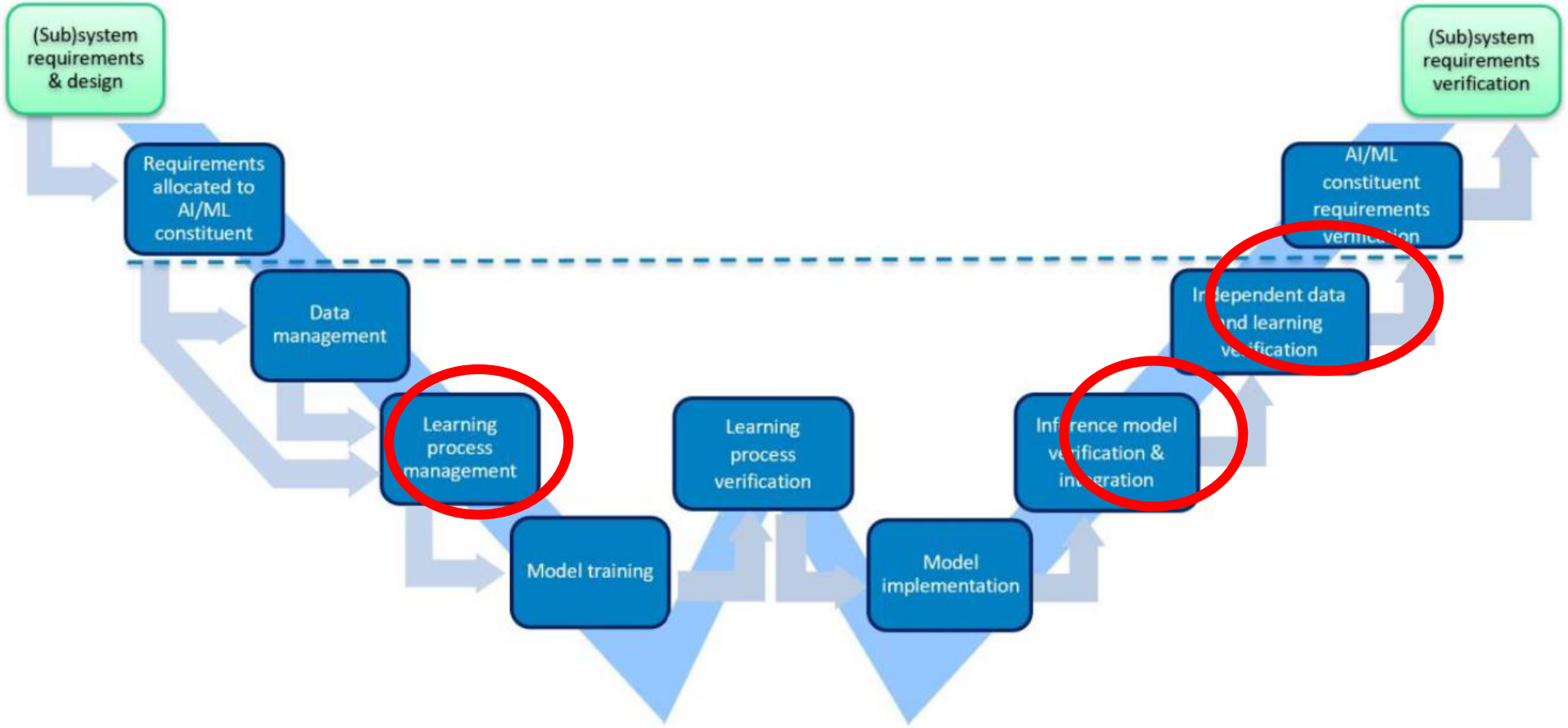


cf next slide

Security-oriented robustness

Robustness

Learning assurance: W-shaped cycle



Objective LM-02: "model robustness metrics and acceptable levels"

Objective LM-13/IMP-08: "document the verification of the robustness of the trained/inference model"

Objective DM-14: "the expected guarantees (generalization, robustness) on the model have been reached"

Adversarial attacks

Depend on the goal, knowledge, and access capabilities of the attacker

- Poisoning attacks (e.g., backdoor attacks, Byzantine attacks)
- **Evasion attacks**
- Model/data extraction (privacy)
- …

Evasion attacks

- White-box, gray-box, physical attacks
- Black-box: zeroth-order attacks, transferability
- Targeted, untargeted, confidence reduction
- Constrained, unconstrained
- What dissimilarity measure?
  - L_p norms, semantic/spatial attacks, invariance attacks, optimal transport
- Universal attacks
- Data type & tasks
  - mostly images, but also text, time series, graphs …
  - mostly classification, but also segmentation, object detection, speech recognition …

Attacks and defenses

Why should we be interested in adversarial attacks and defenses?

- It may provide insights to understand the causes of adversarial vulnerability
- It may help to design strategies to quantify adversarial vulnerability, mitigate it, assess its safety/security impact, and provide guarantees

→ Robustness, Verifiability, Provability

Some classical attacks

- Fast Gradient Sign Method (**FGSM**): L_∞
  - $x_{adv} = x + \epsilon \text{sign}(\nabla_x L(f(x), y))$ where L is the loss function and f is the model
- Projected Gradient Descent (**PGD**): L_∞ , L_2
  - Iterations of FGSM with projections to stay in the L_p ball
- Jacobian Saliency Map Attack (**JSMA**) : L_0
- DeepFool: L_p
  - Iterations of optimal attack for affine models
- Carlini & Wagner attack (**C&W**): L_p for some target class t
  - Minimize $||\epsilon||_p + c\,(\max_{i \neq t} l_i(x+\epsilon) - l_t(x+\epsilon))$ where l(x) are the logits

→ All these methods are **gradient-based** white-box attacks

Some classical defenses

- Adversarial training
- Gradient masking: input transformation, randomization
  - ➢ Confound the adversary but do not eliminate adversarial examples
- Adversarial examples detection
  - ➢ Using another neural network, using statistics
  - ➢ Using a Bayesian framework to quantify the uncertainty

G. Carannante, D. Dera, G. Rasool, N. C. Bouaynaya, L. Mihaylova. Robust Learning via Ensemble Density Propagation in Deep Neural Networks, MLSP 2020

- Model robustification: regularization, robust architectures (Parseval network, 1-Lipschitz networks)
- Robust learning: label smoothing, logit squeezing, distillation, dimensionality reduction

Fisher information and adversarial attacks

Why Fisher information?

The output $f(x) = (p_1, \dots, p_c)$ can be interpreted as a probability distribution

The dissimilarity between two distributions can be measured with the Kullback-Leibler divergence $KL(f(x_1), f(x_2))$

We have $KL\big(f(x), f(x+\epsilon)\big) \approx \frac{1}{2}\epsilon^T J_x^T G_{f(x)} J_x \epsilon$ where $G_{f(x)}$ is the Fisher information matrix.

Fisher information and adversarial attacks

Fisher information has even stronger properties

- It is the **"infinitesimal" metrics** of a large family of "divergences" (including KL divergence)
- It is positive definite and covariant under reparametrization of the parameter spaces (x or p) thus it is a **Riemannian metric** over the manifold of probability distributions (i.e., statistical manifold)
- Chentsov's theorem states that it is the **unique** Riemannian metric on a statistical manifold that is invariant under sufficient statistics (which includes **reparametrization of the sample space**)

Fisher information and adversarial attacks

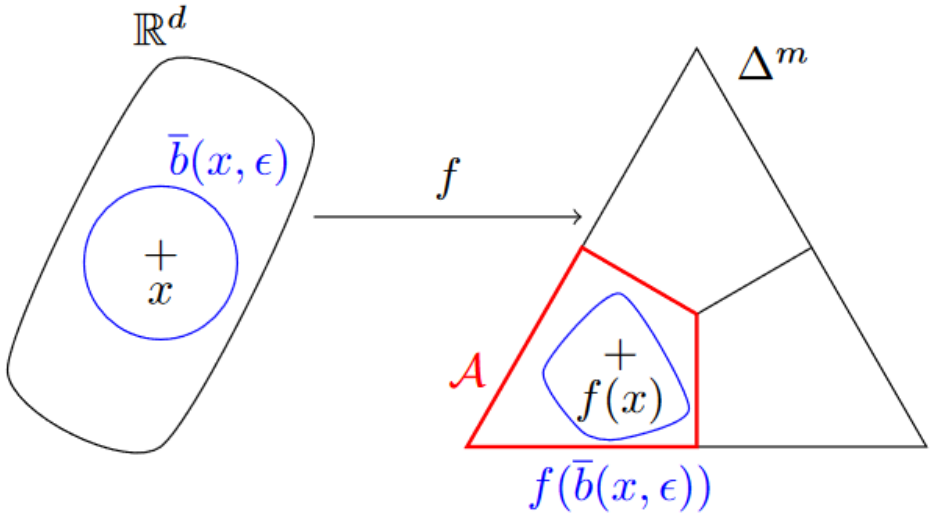
- One-step spectral attack & adversarial example detection

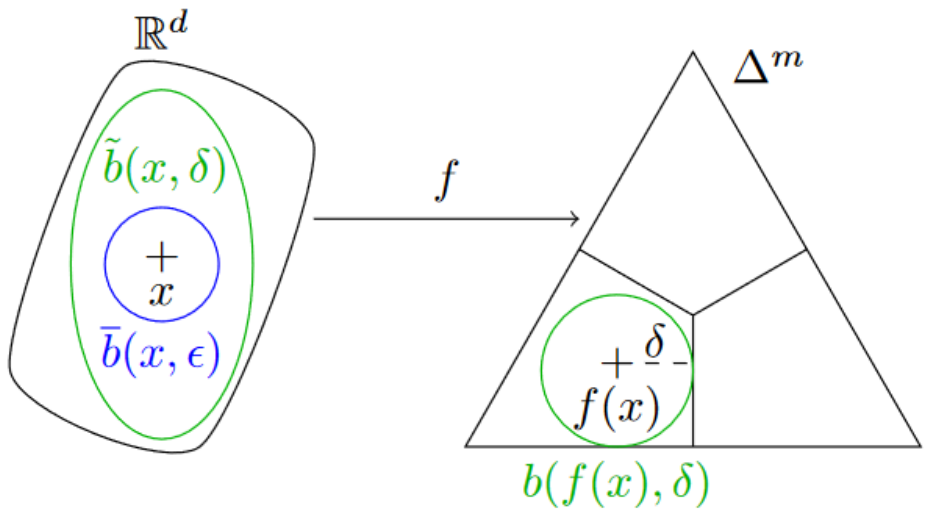
Zhao et al. The Adversarial Attack and Detection under the Fisher Information Metric, AAAI 2019.

- Fisher information regularization
  - ➢ Equivalent to Label smoothing

Shen et al. Defending Against Adversarial Attacks by Suppressing the Largest Eigenvalue of Fisher Information Matrix, arXiv, 2019.

Fisher information and adversarial attacks





Local robustness if $J_x^T G_{f(x)} J_x \leq \delta^2/\epsilon^2$

We have $G_{f(x)} = P_x^T P_x$ where P_x can be efficiently computed as the Jacobian matrix of a well-chosen reparametrization.

Then, we have local robustness if $||P_x J_x||_2 \leq \delta/\epsilon$

What adversarial attacks mean

In terms of cybersecurity: limited relevance

- Real adversaries have little knowledge about the model and no direct access to it
- Real adversaries rely on domain knowledge and social engineering
- Real adversaries have economic constraints

Apruzzese et al. "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice, arXiv, 2022.

What adversarial attacks mean

The expression "*adversarial attack*" is misleading (and redundant)

→ No need of an adversary!

Let h be a human oracle, x such that $f(x) = h(x)$, $\eta(x)$ a "neighborhood" of x.

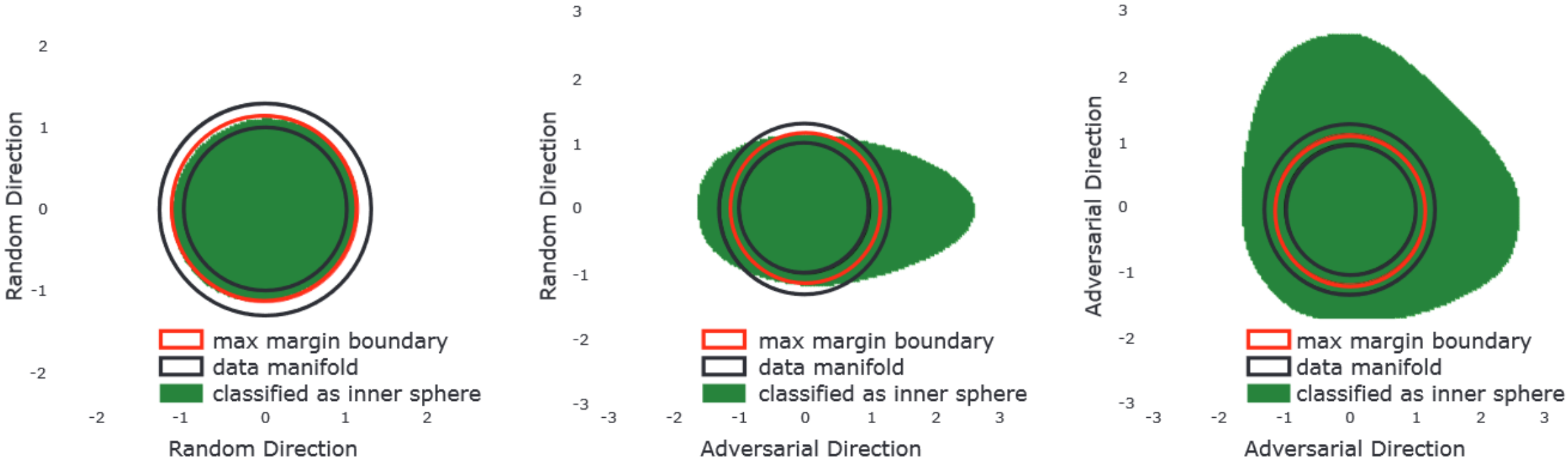
x' is an adversarial example if $\left(x' \in \eta(x)\right) \wedge \left(f(x') \neq f(x)\right) \wedge \left(h(x') = h(x)\right)$

"Adversarial attacks" highlight an inconsistency of the decision boundaries between ML models and human oracles.

→ Machine Learning models don't know what they are doing!

Illustration

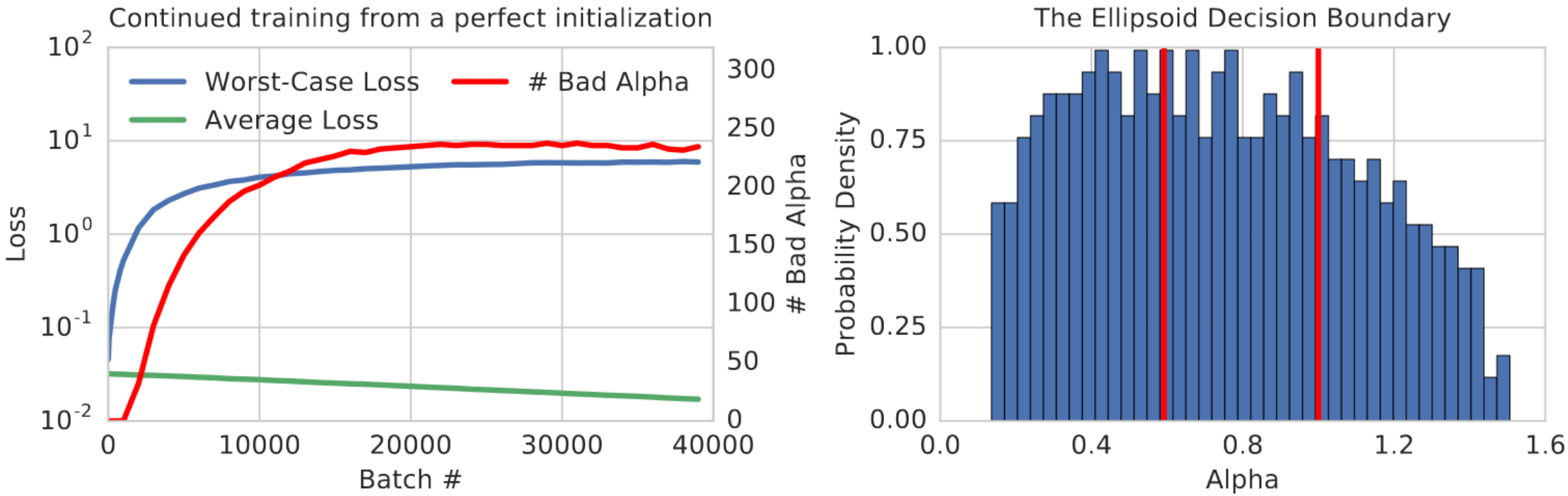
Synthetic dataset: two concentric spheres of dimension $d = 500$



Gilmer et al. Adversarial Spheres, arXiv, 2018.

Illustration

Using a “quadratic network”: $N(x) = \sum \alpha_i z_i^2 - 1$



“The statistical model sums “incorrect” numbers together and obtains the correct answer.”

Gilmer et al. Adversarial Spheres, arXiv, 2018.

Choice of the dissimilarity measure

"Invariance attack": small L_p norm, but different class

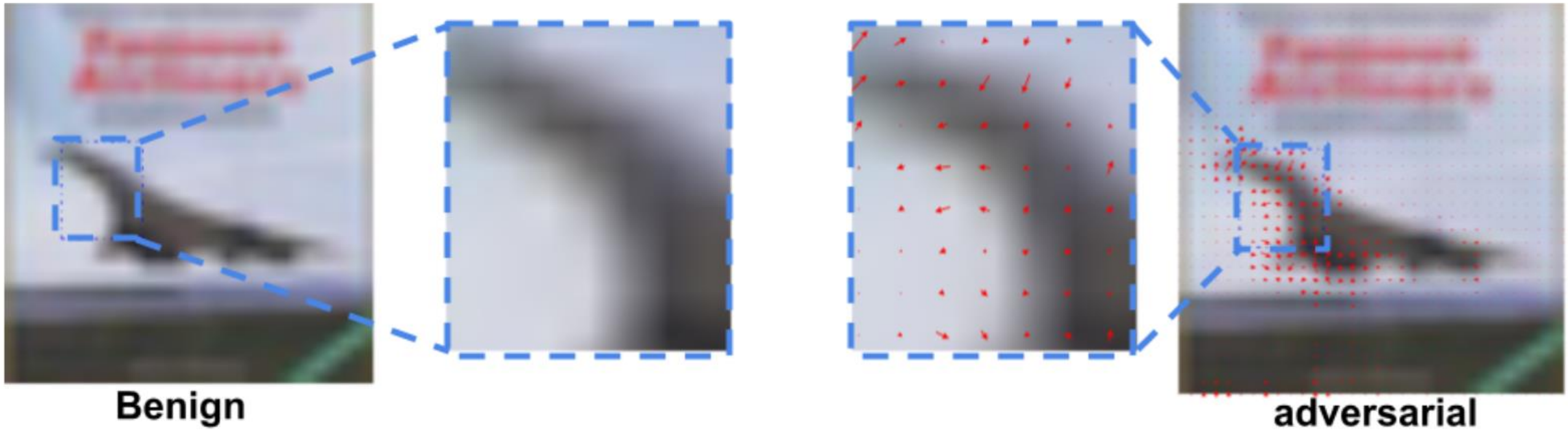
Physical attack

Label: No entry
Predict: No left turn
Confidence: 52.5%

Large L_p norm, visible perturbation but seems natural

Choice of the dissimilarity measure

Spatial attack: Large L_p norm, but still indistinguishable



Tramer et al. Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations, ICML 2020.
Liu et al. Adversarial Attack with Raindrops, arXiv, 2023.
Xio et al. Spatially Transformed Adversarial Examples, arXiv, 2018.

Explaining adversarial vulnerability

No general agreement, because no rigorous foundations

- Too little or too much “linearity”
- The “manifold hypothesis”
- Overfitting, no enough data
- High dimensionality
- Adversarial features
- Computational complexity / Accuracy-robustness trade-off
- …

What adversarial attacks mean

What are the implications for safety?

- Can extensive testing really build trust against adversarial attacks?
- Are rigorous AI assurance process and data management enough to build trust?
- **Robustness** and **explainability** are not independent from each other.
- Robustness is a consequence of the training process.

Certification of model stability

How to certify the model stability of ML models?

Certification of traditional software can be extended to ML models:

- **Reviews** (of both development process and items properties)
- **Testing**
  - ➢ coverage criteria, test case generation, mutation testing, simulation, scenario …
  - ➢ for relevant evaluation of adversarial robustness, use specialized benchmark relying on adaptive attacks (e.g., AutoAttack)

Croce & Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, PMLR 2020.

- **Verification** (formal methods): by-design, a posteriori

Verification of model stability

What properties?

- Local adversarial robustness
- Interval property
- Reachability
- Lipschitz constants

Huang et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Computer Science Review, 2020.

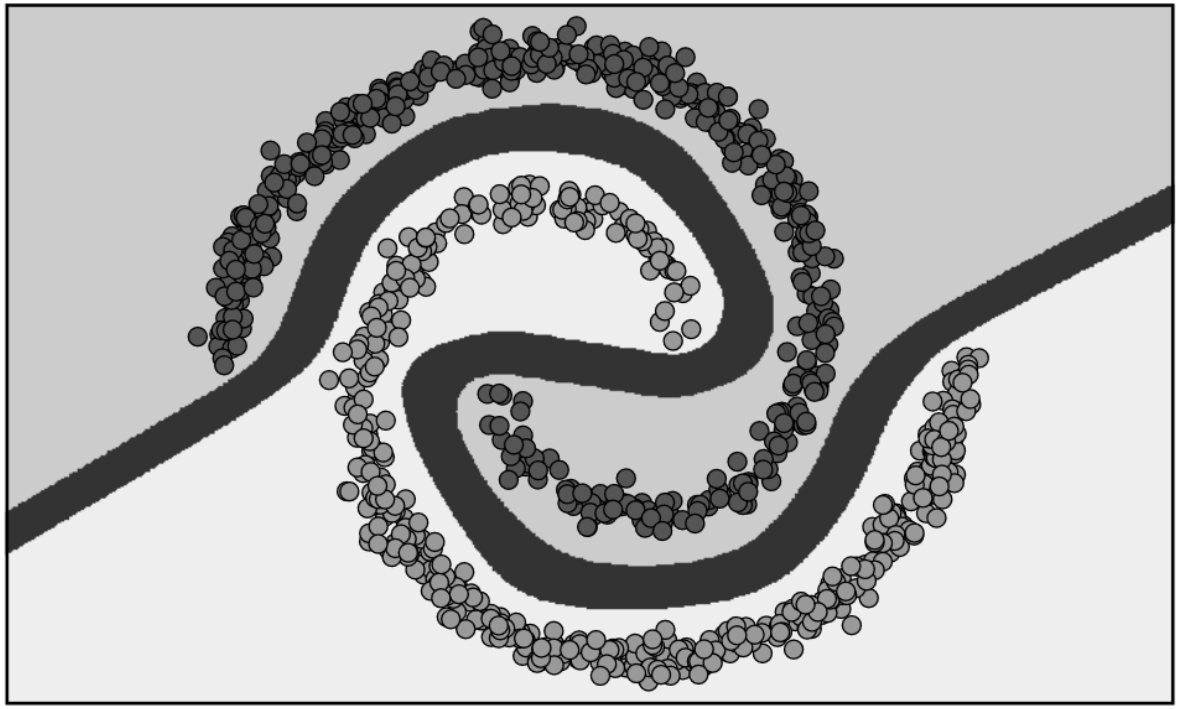
Verification of model stability

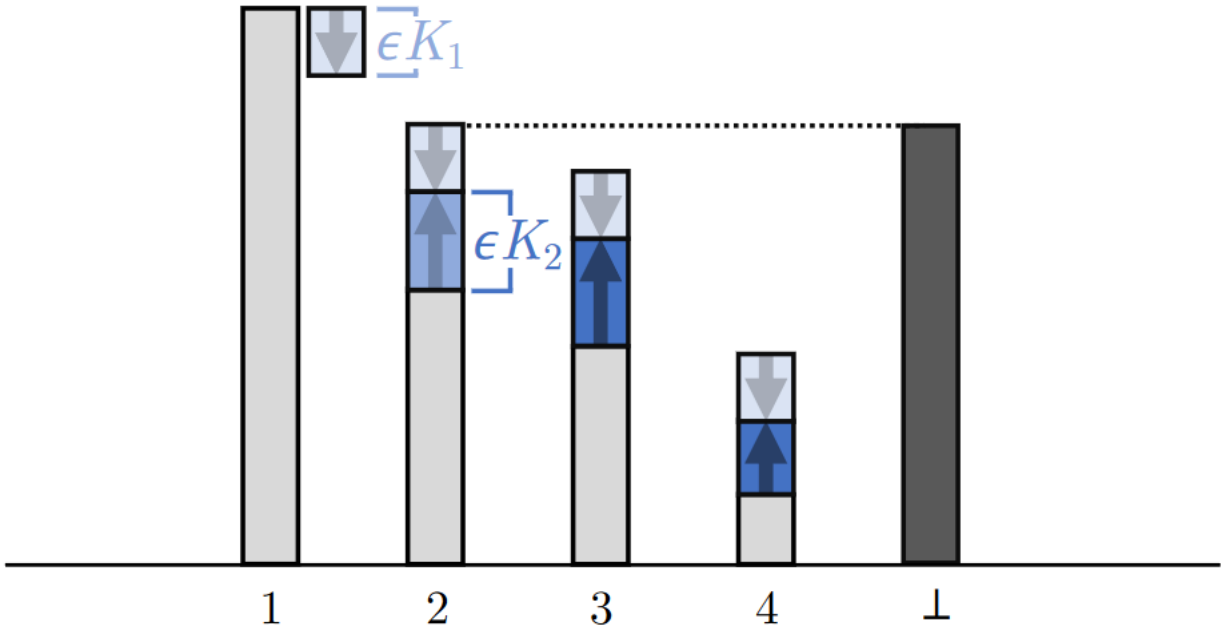
What guarantees?

- Deterministic (sound and complete) ~100 neurons
  - SAT/SMT (Reluplex), MILP
  - NP-hard
- Approximate (sound but not complete) ~10,000 neurons
  - Abstract interpretation
  - Convex optimization based methods
  - Interval arithmetic
- Converging (sound but not complete) $\geq 10^6$ neurons
- Statistical

Global robustness guarantees

- Based on an upper-bound of the global Lipschitz constant
- Additional class for points that cannot be verified as robust





Leino et al. Globally-Robust Neural Networks, arXiv, 2021.

Randomized smoothing

- Statistical verification method
- Scalable, model-agnostic

Base classifier f, input point x, parameter $\sigma > 0$.

Smoothed classifier:

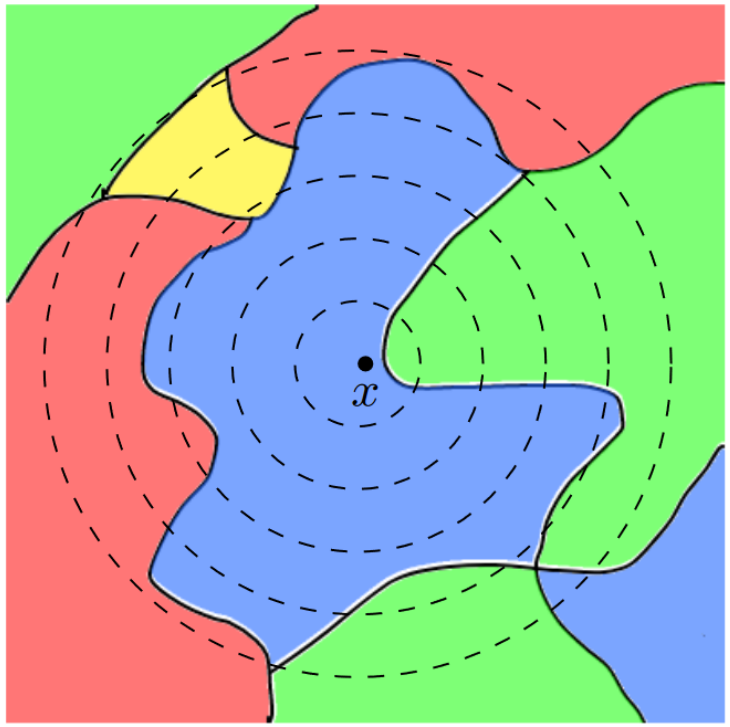
$$g(x) = \text{argmax}_c \, P(f(x + \epsilon) = c)$$

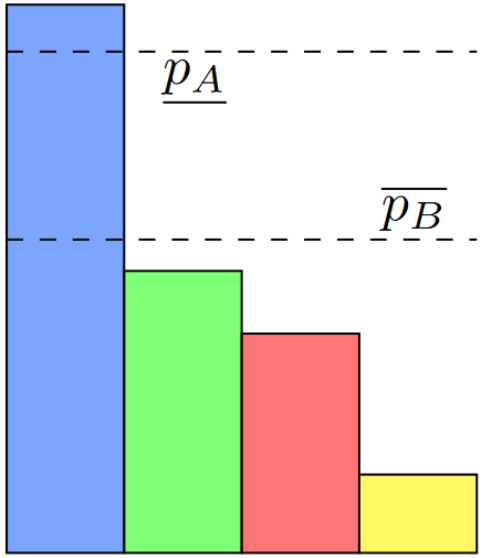
with $\epsilon \sim N(0, \sigma^2 I)$.

If $P(f(x + \epsilon) = c_A) \geq p_A \geq p_B \geq \max_{c \neq c_A} P(f(x + \epsilon) = c)$,

Then $g(x + \delta) = c_A$ for all:

$$||\delta||_2 < \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$$





Cohen et al. Certified Adversarial Robustness via Randomized Smoothing, PMLR, 2019.

Takeaways

- Industries (and society at large) will become more and more **dependent** on ML.
- ML models performance should not hide the fact that their foundations are **not understood**.
- How can we trust a technology that becomes more and more **opaque**, where small areas of **catastrophic instability** are hiding?
- Since we cannot “open the black-box”, how can we produce **quantified safety assessment**?
- **Verification methods** are a solution, relying on the field of formal methods, or developing new approaches specific to ML.
- Another solution is the **resilience** of the systems containing ML items.