SAIDA

Security of AI for Defense Applications

~~Teddy Furon, LinkMedia~~

~~Centre Inria Rennes Bretagne Atlantique~~

Présenté par Laurent Amsaleg, LinkMedia

IRISA, Rennes

Proposal

- Call
  - Chair of research and teaching in artificial intelligence

- Agence de l’Innovation de Défense
  - 4 projects / 40 selected projects

  - Topics of interests
    - Data processing from various sensors (radar, sonar, SAR and IR imagery, hyperspectral ...)
    - Reliability, robustness, vulnerabilities and countermeasures of A.I.
    - Distributed processing and applications for network communications
    - AI for cyber-security, risks of misinformation and fake news

- Chaire SAIDA supported by
  - DGA, Thales, Airbus Defense & Space, Naval Group, ZAMA

Motivations

- Robustness gives a false sense of Security
  - Robustness: To operate as expected even under perturbations (Innocuous)
  - Security: To operate as expected even in hostile environnments (Malicious)

- Little bits of history repeating
  - I've seen it before: Digital Watermarking
  - I've seen again: Content Based Image Retrieval
  - The next big thing is here: Machine Learning

- Motto: « Security of M.L. before M.L. for security »
  - Better study the intrasinc security of a tool before using it in security applications

Goal

- Establish the principles for designing reliable and secure AI systems
  - a reliable AI maintains good performance even under uncertainties
  - a secure AI resists attacks in hostile environments

  - at training and testing time

- Combining theory with applied and heuristic studies
  - to guarantee the applicability
  - to cope with real world settings

Scope

1. Theoretical investigations

1.1 Local Intrinsic Dimensionality–LID	collab. NII, Japan
1.2 Reliability and Rare Event analysis	Ph.d Thales
1.3 Immune training	_

2. Lessons learned from Information Forensics and Security

2.1 Inputs from Watermarking	Ph.d. DGA
2.2 Inputs from Steganalysis and Image Forensics	
2.3 Black box Attacks	Ph.d Inria

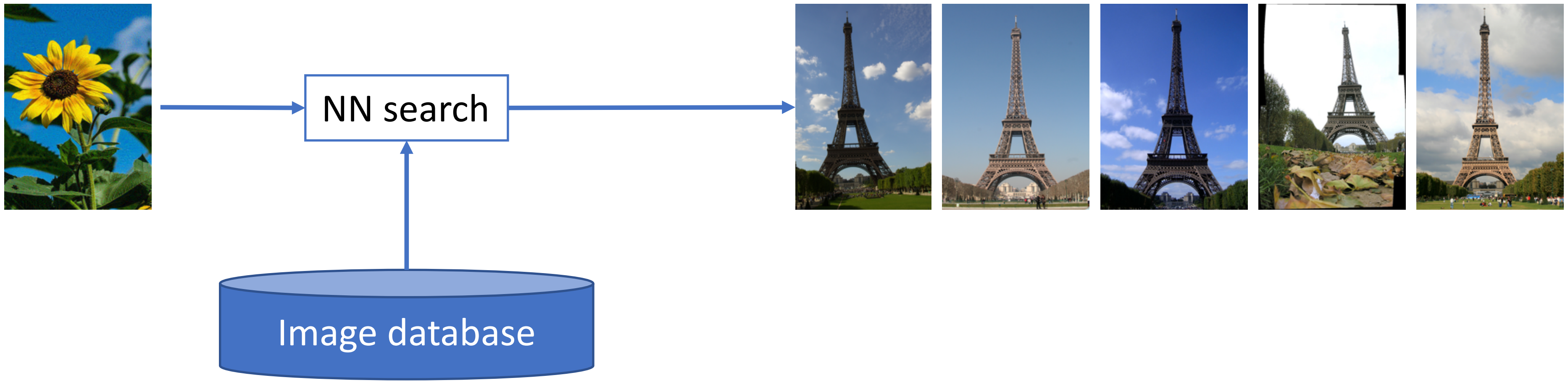
3. Protection of the data/network

3.1 Leakage about training data	_
3.2 Poisoning of training data	_
3.3 Secret-keyed network	Ph.d ZAMA.ai

Focus #1: High LID facilitates adversarial attack

Deluding Nearest Neighbors Search in large collection

- k-NN is ubiquous in data mining



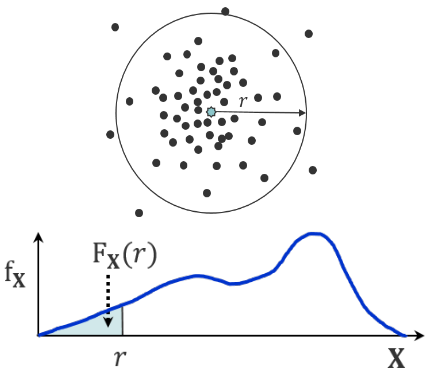
Query with a Flower to Retrieve the Tower, Tolias et al., CVPR19
Deluding image recognition by attacking keypoints, Do et al., ICASSP12

Focus #1: High LID facilitates adversarial attack

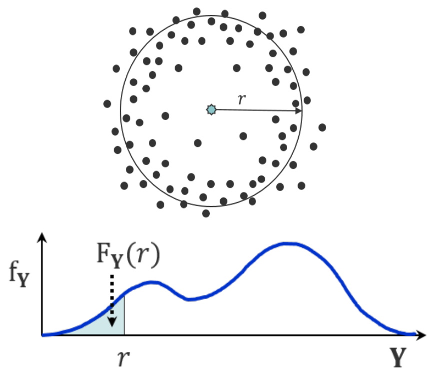
Our work: Theoretical evidence

Local Intrinsic Dimensionality caracterizes the neighbourhood of a point

Low *LID*



Large *LID*



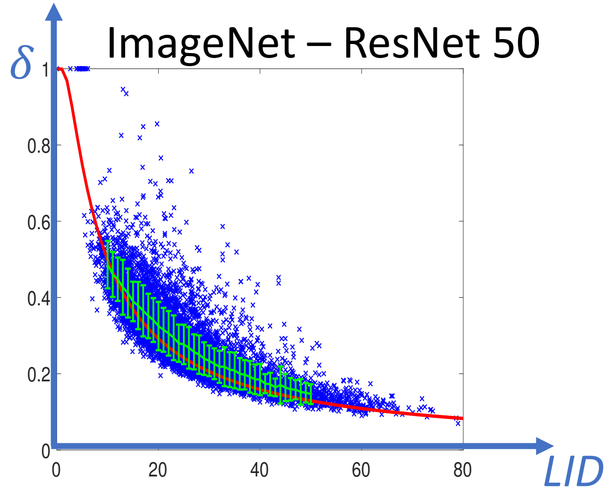
Scenario: perturbate query s.t. *k*-th NN becomes 1st

$$\delta \approx 1 - k^{-\frac{1}{LID(x)}}$$

amount of perturbation of the query x

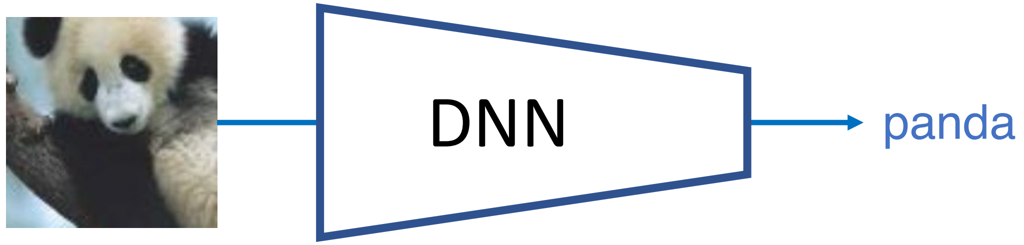
*k*th NN becomes 1st NN

LID around query x



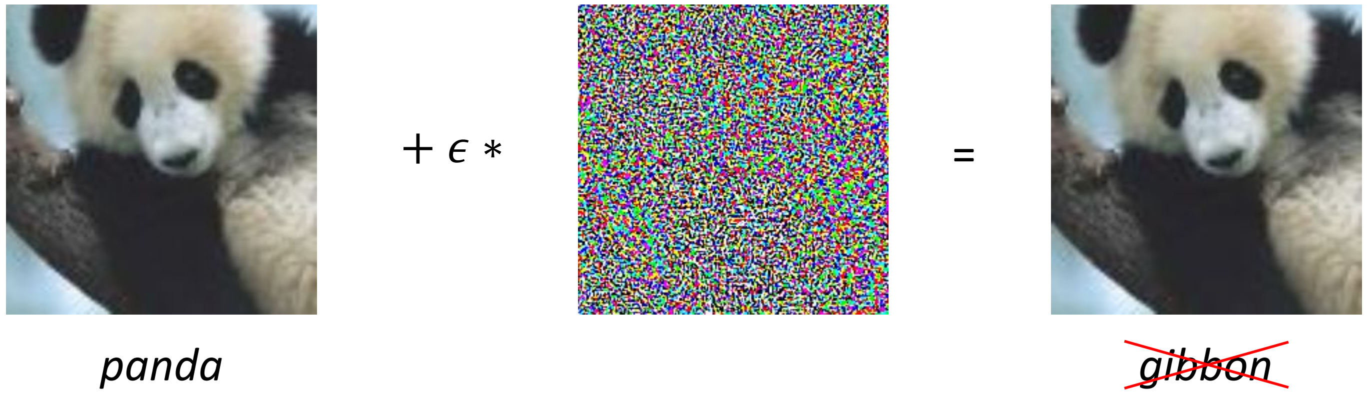
Focus #2: Adversarial examples

Perturbate input image to delude a classifier



In literature, most attacks forge adversarial images … which are not images!

- Machine learners work with floating point $x \in [0,1]^{3*L*C}$
- Naïve rounding ruins the attack



Explaining and harnessing adversarial examples, Goodfellow et al., 2015

Focus #2: Adversarial examples

Our work: design a quantization maintaining adversariality

- Apply your favorite attack
- We turn it into real images (PNG or JPEG)

original
shopping_cart

JPEG75
shopping_cart

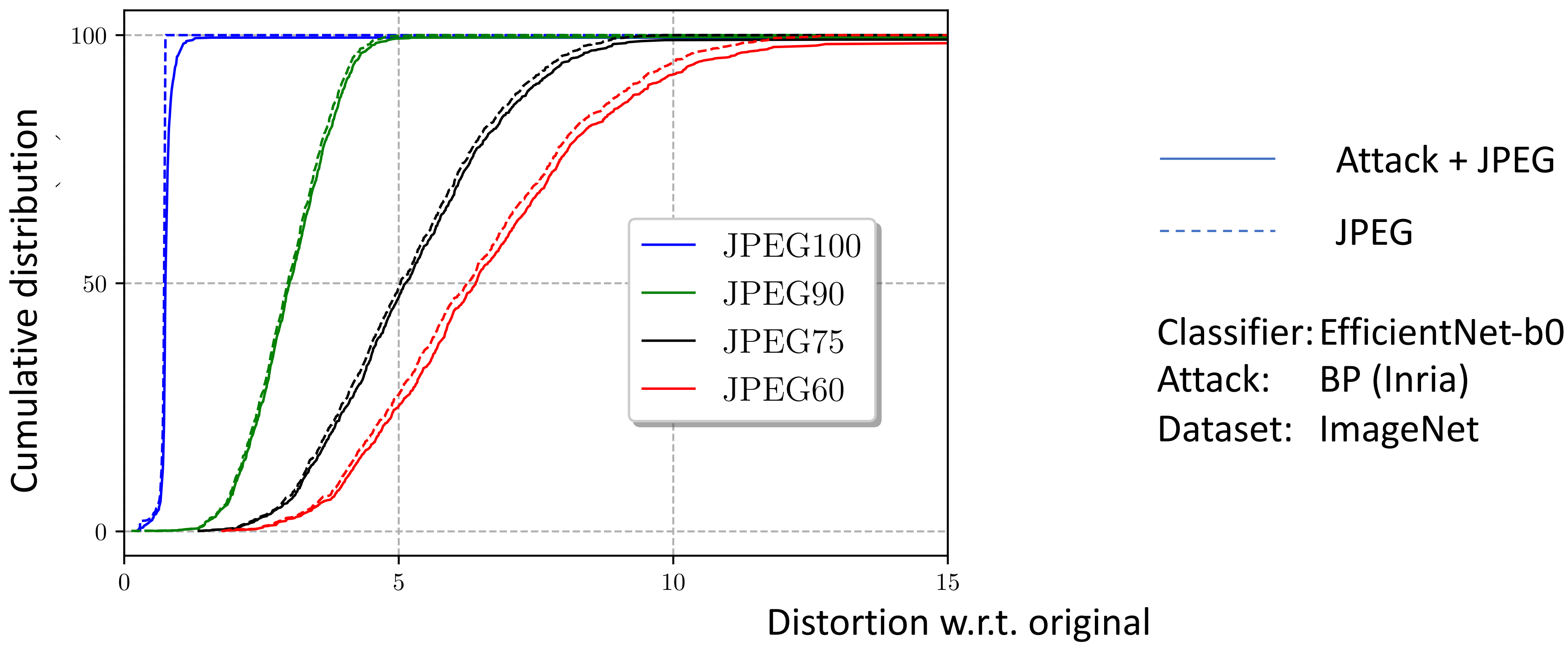
Attack+PNG
basset_hound

Attack+JPEG75
basset_hound

Focus #2: Adversarial examples

Surprizingly:

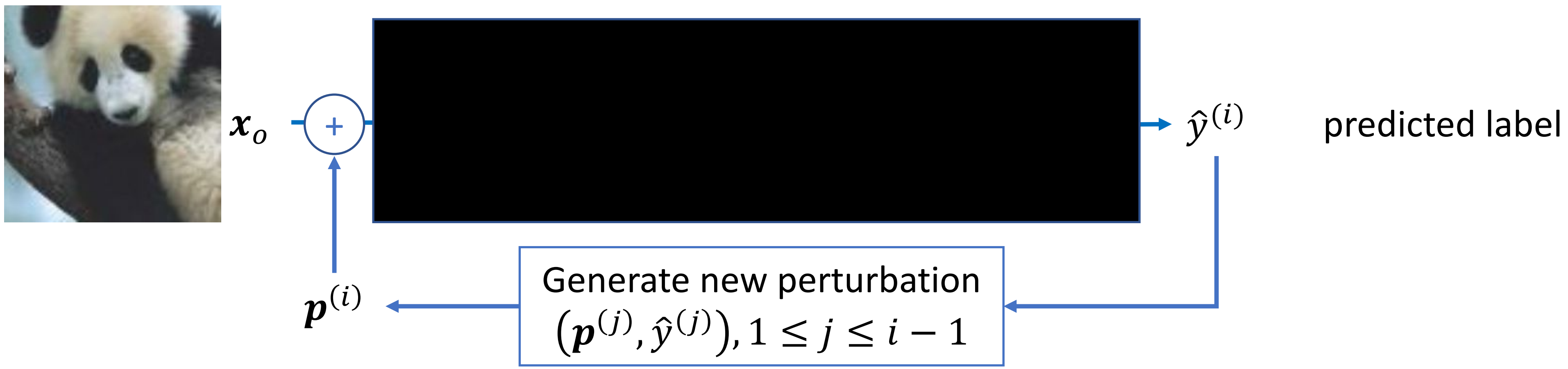
- Quantization is not a strong constraint (if treated carefully)
- The attack is for free w.r.t. distortion



Classifier: EfficientNet-b0 (google)
Attack: BP (Inria)
Dataset: ImageNet

Focus #3: Black box attack

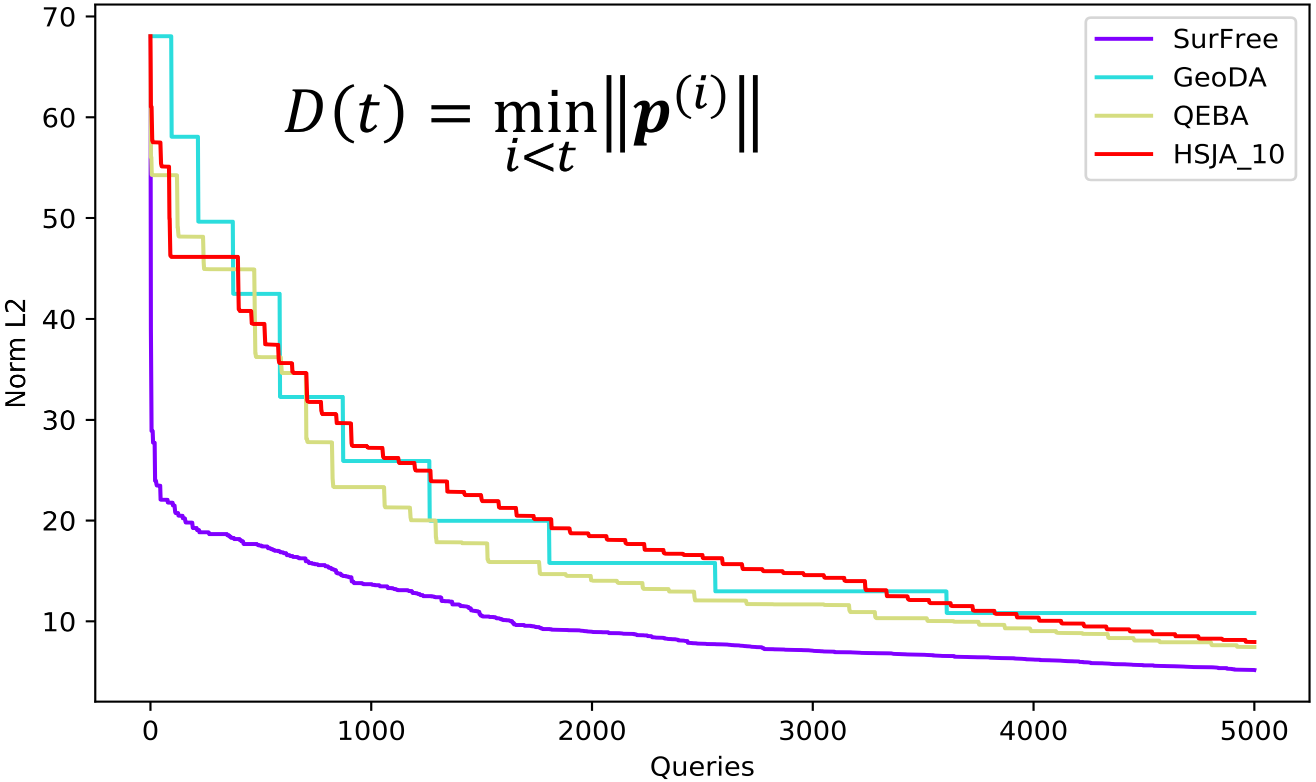
- Difficult scenario
  - No knowledge of the classifier
  - Access as an oracle
    - Choose input, observe ouput (hard predicted label)

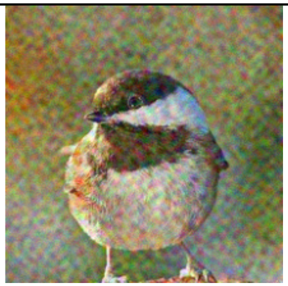
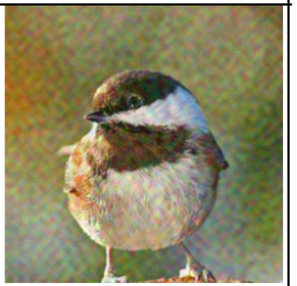


- SotA attacks are very long (~5,000 calls per image)

Focus #3: Black box attack

- Our work: SurFree
  - Designed for speed (few calls to the oracle)
  - Still competitive in the long run

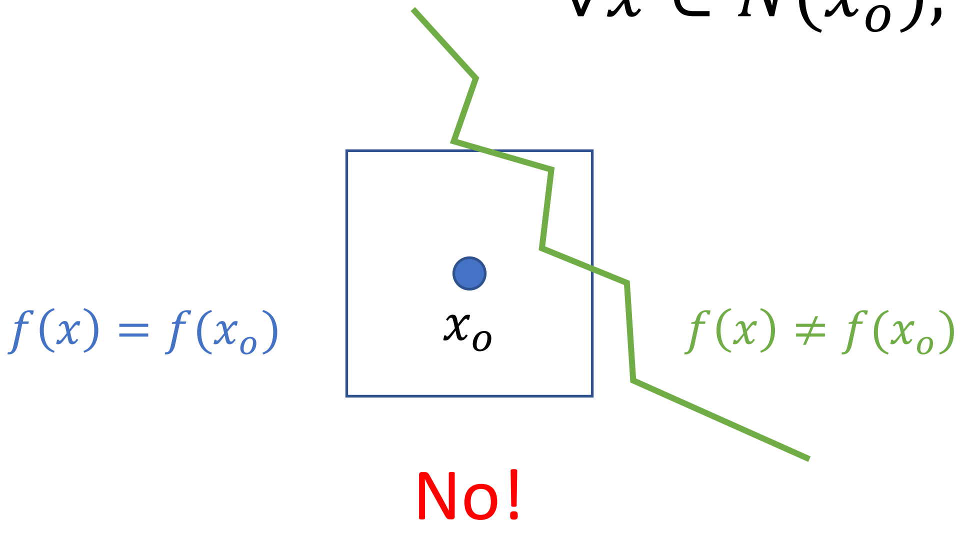


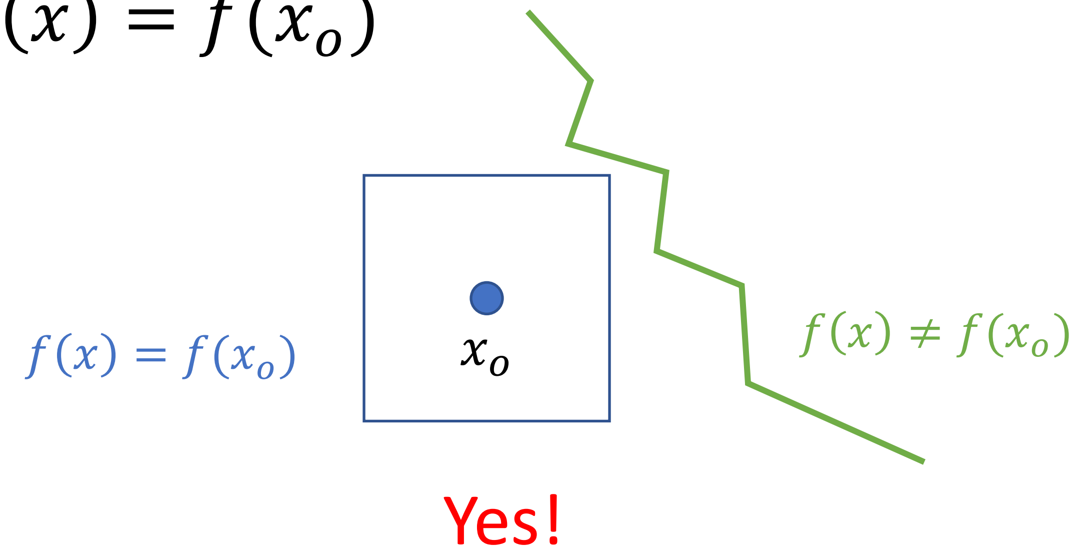
attack	$K = 100$	$K = 500$	$K = 1000$
SurFree	 amer. dipper- 2.6	 amer. dipper- 1.3	 amer. dipper- 0.9
QEBA [13]	 stingray- 60.6	 stingray- 33.7	 stingray- 20.8
GeoDA [22]	 brambling- 18.9	 brambling- 9.7	 brambling- 5.8

Focus #4: Certification of neural networks

- Is this property true?

$$\forall x \in N(x_o), \qquad f(x) = f(x_o)$$

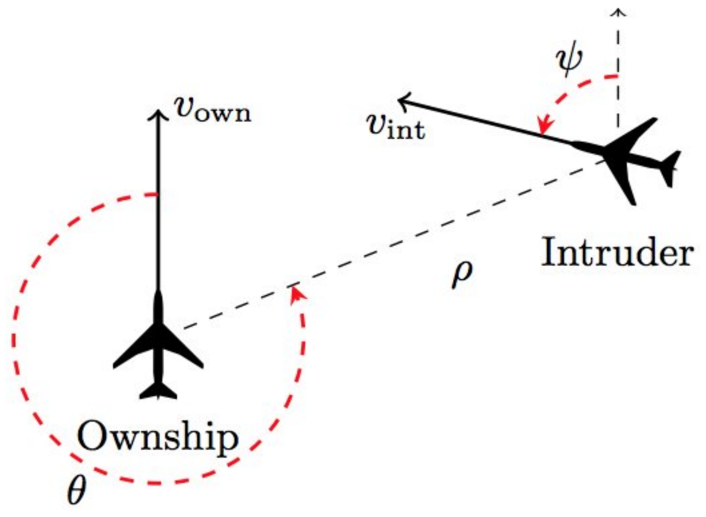


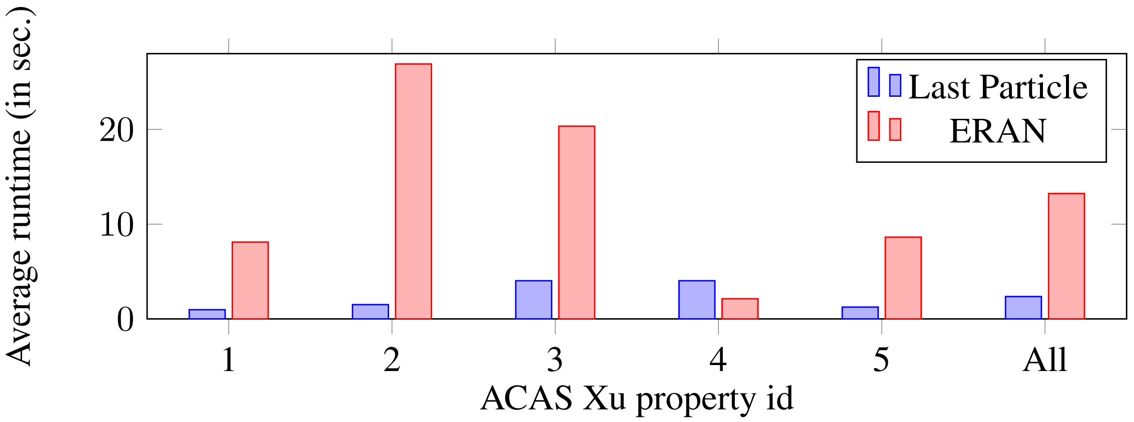


- Formal proof
  - NP-hard for Deep Neural Networks
  - Some librairies (ReLuPLEX, ERAN, PROVEN)
    - simple networks, simple neighborhoods
    - May time out, may give up

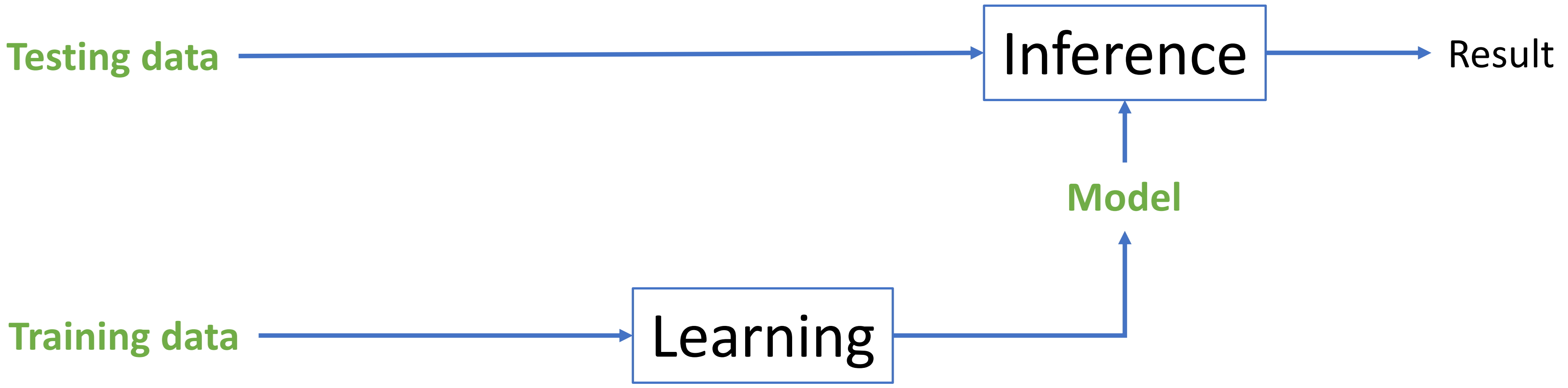
Focus #4: Certification

- Our work: statistical approach
    1. Consider random input $X \sim \mathcal{U}[N(x_o)]$
    2. Estimate $p = \text{Prob}\big(f(X) \neq f(x_o)\big)$ with Rare Event Simulation
    3. Certify if $p < p_c$ with p_c extremely small $\sim 10^{-30}$
- Fast but not sound
    - Incorrect if $0 < p \ll p_c$





The global picture: Security of M.L.



Extension to different data types and learning frameworks (X - learning)

These three contents need protection

- Values to be protected
  - Integrity
  - Confidentiality
  - Ownership