U.S. NAVAL RESEARCH LABORATORY

# Breaking and Fixing
## Autonomous Cyber-Physical Tactical Systems

**Dr. Ramesh Bharadwaj**
Ramesh.Bharadwaj@nrl.navy.mil

High Assurance Tactical Systems Engineering Group
Center for High Assurance Computer Systems
US Naval Research laboratory
4555 Overlook Avenue SW
Washington DC 20375 USA

U.S. NAVAL RESEARCH LABORATORY

# Speaker Introduction

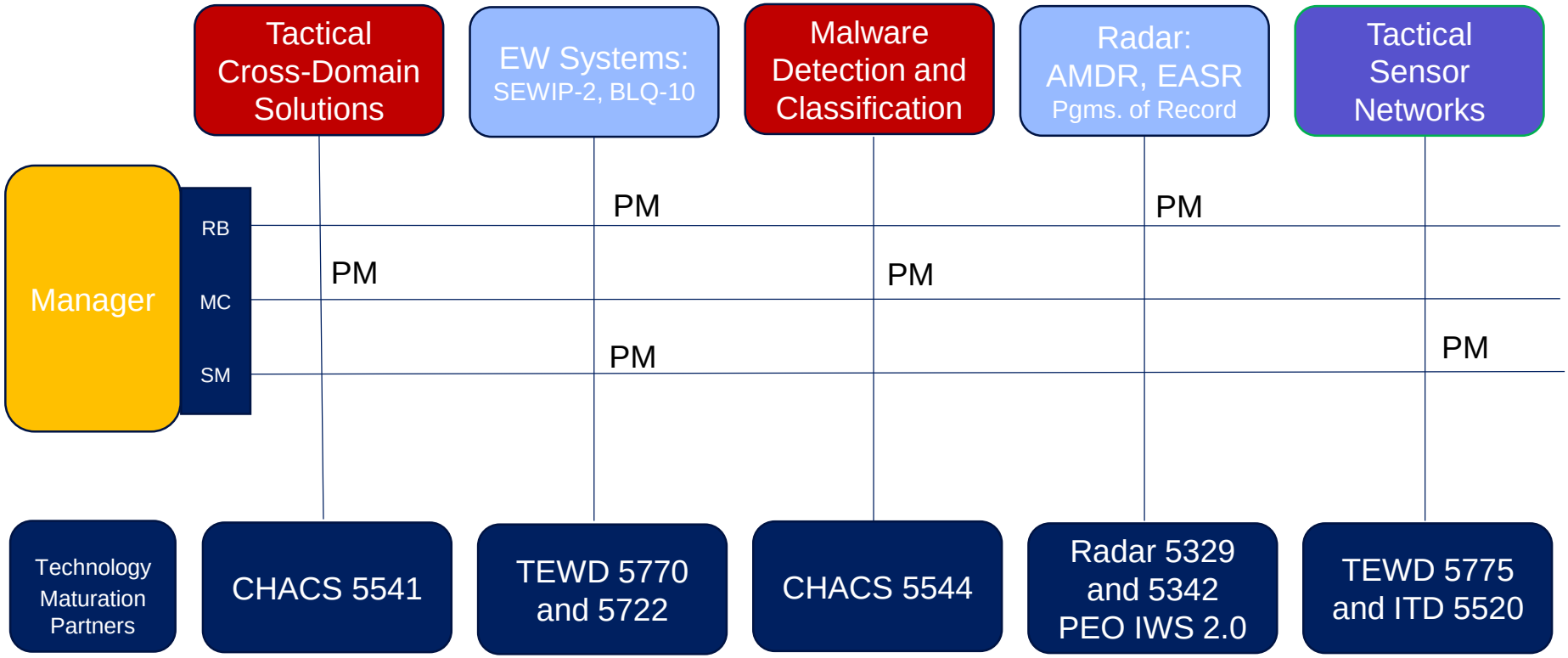
## Ramesh Bharadwaj

- PhD, Computer Engineering, Communications Research Laboratory, McMaster University, Hamilton, ON Canada
- MEE, Electronics Engineering, Philips/Eindhoven International Institute, Eindhoven, The Netherlands
- BE, Electronics and Communications Engineering, National Institute of Engineering, Mysore, India
- Current Positions:
  - Project Manager: Assured Autonomy (NRL Base Program), Cognitive Coherent Networked Apertures (ONR Code 312)
  - Member, Software Working Group, AMDR and EASR Programs of Record (AN/SPY6(V)1/2/3)
  - Member, US Delegation, Indo-US Joint Technical Group on C4ISR
  - Member, SAE G-34/ EUROCAE WG-114, “Artificial Intelligence in Aeronautical Systems”
  - Member, Program Committee: “Implementing AI Ethics,” AAAI Spring Symposium Series (virtually) Stanford, CA
  - Organizer, “1st International Workshop on Trusted Automated Decision-Making,” at ETAPS 2021, (virtually) Luxembourg
- Previous Positions:
  - Philips Research Laboratories (Eindhoven), Tata Institute of Fundamental Research (Bombay), Stanford University (Palo Alto), AT&T Bell Laboratories (Murray Hill), Fraunhofer FOKUS (Berlin)
  - KTH Royal Institute of Technology (Kista, Sweden), George Washington University and Catholic University of America (Washington DC)
- Background:
  - Ten years’ experience in Modeling & Simulation and Electronic Warfare systems
  - Five years’ experience in Virtual Integration of Electronic Warfare Systems (ViEWS);
  - Subject Matter Expert on multifunction radars and electronic warfare systems including AN/SPS49A(V)1 and AN/SLQ-32(V)6

U.S. NAVAL RESEARCH LABORATORY

# High Assurance Tactical Systems Engineering Research

| Disruptive Innovation in Tactical Systems Engineering |
| --- |
| Objective: Machine Learning for High Assurance |
| Approach: High Levels of Automation for Low Code |
| Tools, Theories, and Processes for High Assurance |
| Underlying Theories: Mathematical Logic; Statistical Learning |
| Products: Research Prototypes, Technology Demonstrators |



U.S. NAVAL RESEARCH LABORATORY

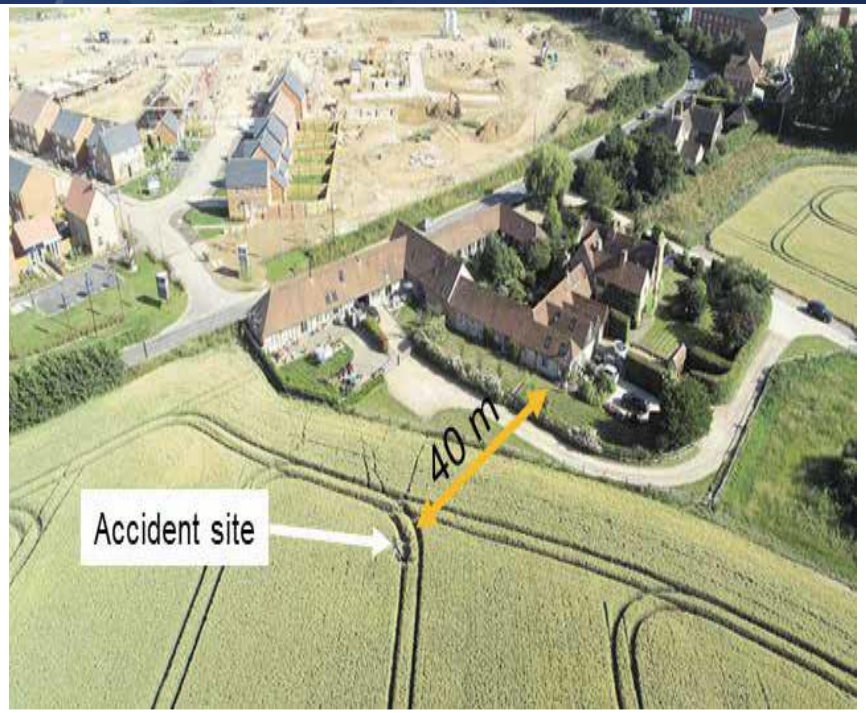
# News Headline: Unmanned Aircraft Crashes

News Report: “Control of a prototype unmanned aircraft, an Alauda Airspeeder Mk II, was lost resulting in a fly-away and eventual crash.”

Goodwood Aerodrome, West Sussex, 4 July 2019

U.S. NAVAL RESEARCH LABORATORY

# Post-Mortem of Events Leading to the Crash




Sequence of events:

- Remote pilot lost control of the 95 kg unmanned craft
- Safety “kill switch” was activated, but had no effect
- The craft climbed to 8000 ft, into controlled airspace
- Crashed in a field of crops approximately 40m from occupied houses and 700m outside of its designated operating area

U.S. NAVAL RESEARCH LABORATORY

# FAA/ EASA Response to the Drone Problem

## SAE G34/ EuroCAE WG-114 Working Group on "Artificial Intelligence in Aviation"

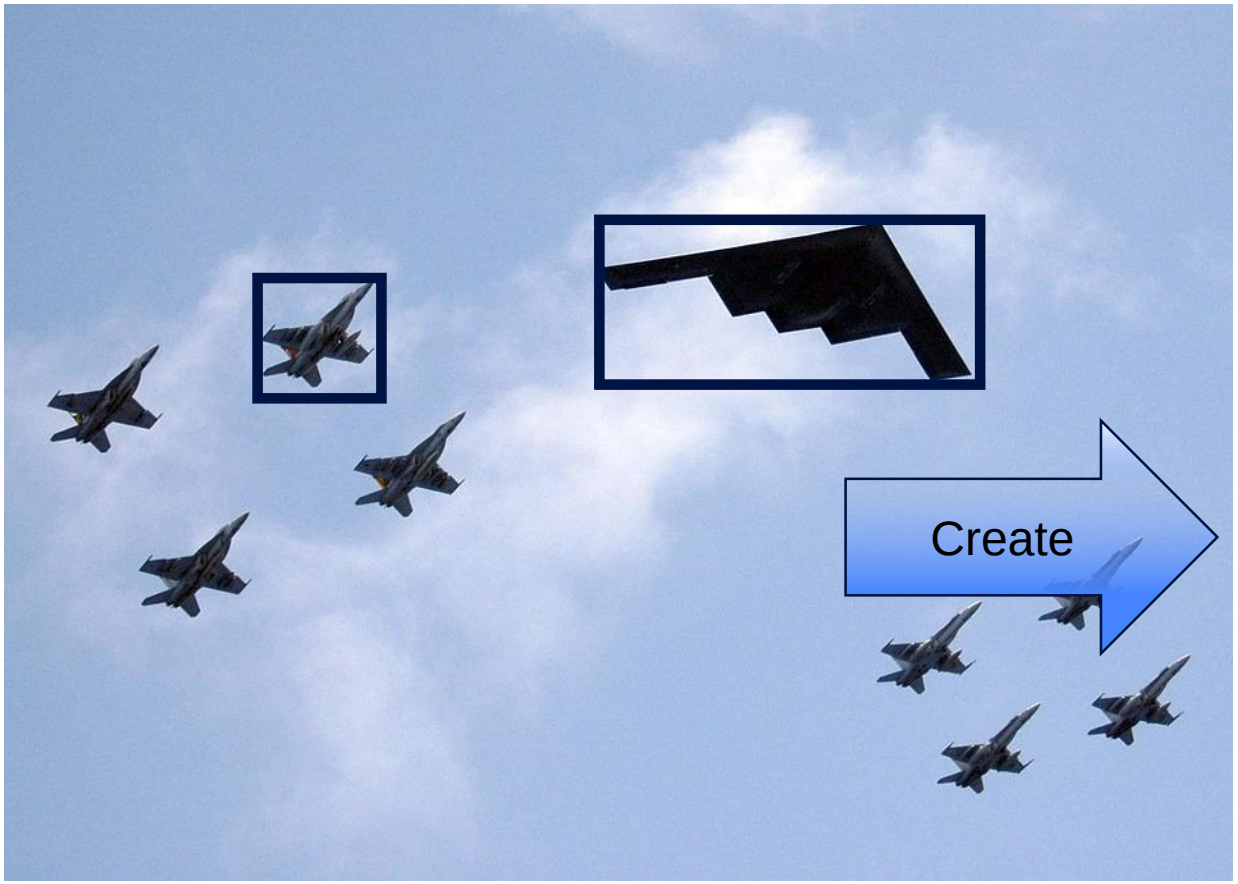
Circling back to the Airspeeder Mk II crash: Our group's charter is to "prepare technical standards required to support development and certification of aeronautical systems implementing AI-technologies."

AIR6988 "Artificial Intelligence in Aeronautical Systems: Statement of Concerns"

AIR6983 "Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI"

U.S. NAVAL RESEARCH LABORATORY

# Background: *Deep Learning*





Labeled DoD Dataset

Deep Learning Process:

1. Aquire Labeled DoD Dataset
2. Train Network with Labeled Training Da et
3. Test Network on Labeled Test Dataset

Generalization

Deep learning is predicated upon the availability of large corpora of labeled data

U.S. NAVAL RESEARCH LABORATORY

# Assurance Objective

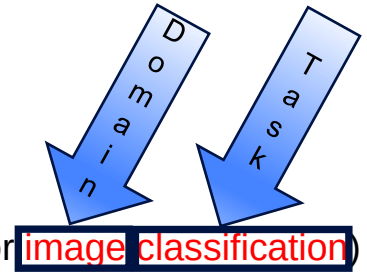
## Dependability of Naval Autonomous Systems based on Machine Learning (ML), in particular Deep Learning (DL)

1. Systems based on ML **will be** deployed on a wide range of DoD systems – surveillance and recommendation systems, radar and EW, cruise missiles, and systems for long-duration unmanned missions such as UUVs, USVs, and UASs
2. ML-based systems trained by deep learning are prone to misclassification errors
3. Assurance of DoD autonomous systems that rely on ML algorithms is paramount

Vocabulary

- ML: Machine Learning
- DL: Deep Learning
- DNN: Deep Neural Network
- CNN: Convolutional Neural Network

(1D, 2D and 3D variants; generic 2D variant for image classification)



U.S. NAVAL RESEARCH LABORATORY

# Why is this a DoD problem?

World's first self-driving shuttle crashes on first day (11/8/17)
Tesla's Autopilot steers Model X into highway median (2/25/20)


Self-driving Uber Volvo in high-speed crash (3/24/17)
Self-driving Uber Volvo kills bicyclist (3/19/18)

- **"Design for commercial systems rarely considers the possibility of high-regret outcomes in complex, unpredictable, and contested environments …. Fielded commercial autonomous systems do not yet face a motivated adversary attempting to defeat normal operations*"**
- Commercial autonomous systems can rely upon law enforcement and social and legal penalties to discourage adversaries

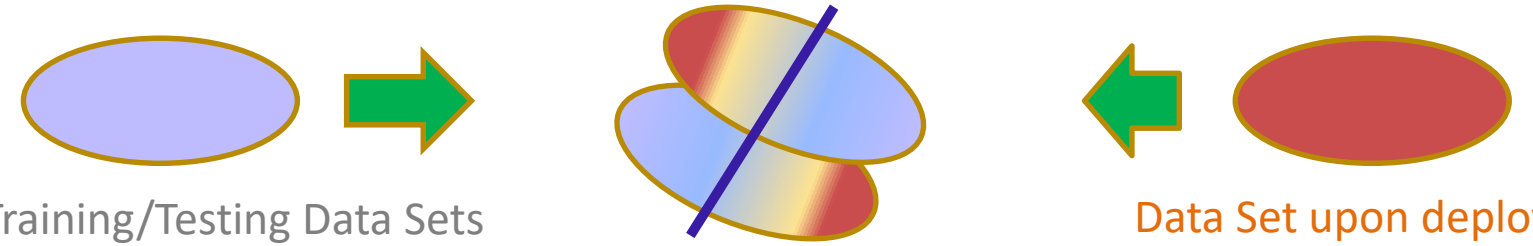
*** Report of the Defense Science Board Summer Study on Autonomy (June 2016)**

U.S. NAVAL RESEARCH LABORATORY

# Dependability of Systems Based on ML

We define "Dependability" as follows:

1. Safety[1]: No "unintended engagements" with other agents in the system's environment (under any circumstances)
2. Reliability[2]: Robust operation under all fielded conditions
   - Natural or Adversarial Distribution Shifts




3. Trust[1]: System actions are interpretable, secure and fair

Still a research question. Do participate in TADM 2021!!

[1] Proved by logical arguments [2] Established by statistical metrics

**DoD Directive 3000.09: "… minimize consequences of failure that may lead to unintended engagements."**

U.S. NAVAL RESEARCH LABORATORY

# Levels of Criticality in ML-based Systems

## Level 0: Non-critical

Netflix recommendation system; Face-tagging photos/videos on Instagram; Bird species identification

## Level 1: Pecuniary

Credit card fraud alerts; Automated trading; Creditworthiness assessment; COVID “Health Passports”

## Level 2: Lifestyle

Recidivism assessment; Biometric id for apprehending criminals/traffickers; Automated Radiologist

## Level 3: Safety-critical

Autonomous vehicles (ground/drones); Firefighting; Explosive/radioactive ordnance detection/disposal

## Level 4: Mission-critical

Nuclear reactor and power grid control; Automated warfighting systems; Nuclear-tipped ballistic missiles

Virtual (no physical interaction with environment)

Cyber-Physical (human safety is at risk)

**Most insidious issue: Use of commercial (or other) technologies developed for Virtual-only Systems and attempting to implement them in the Cyber-Physical Domain**

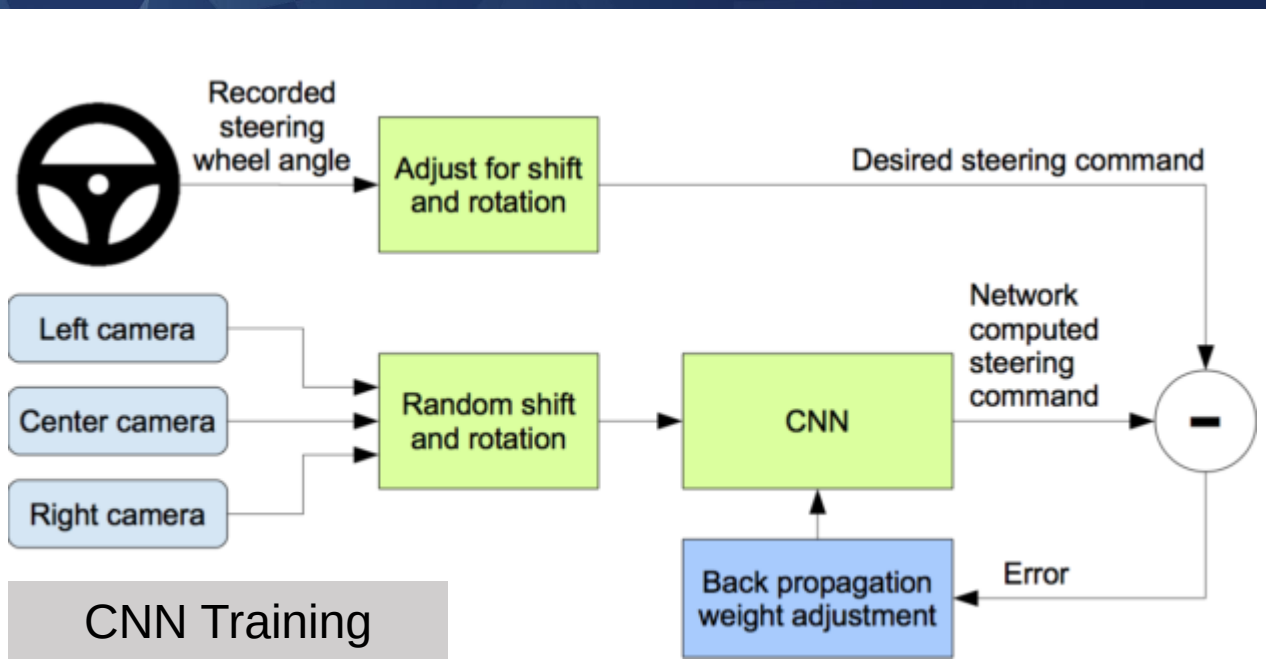
U.S. NAVAL RESEARCH LABORATORY

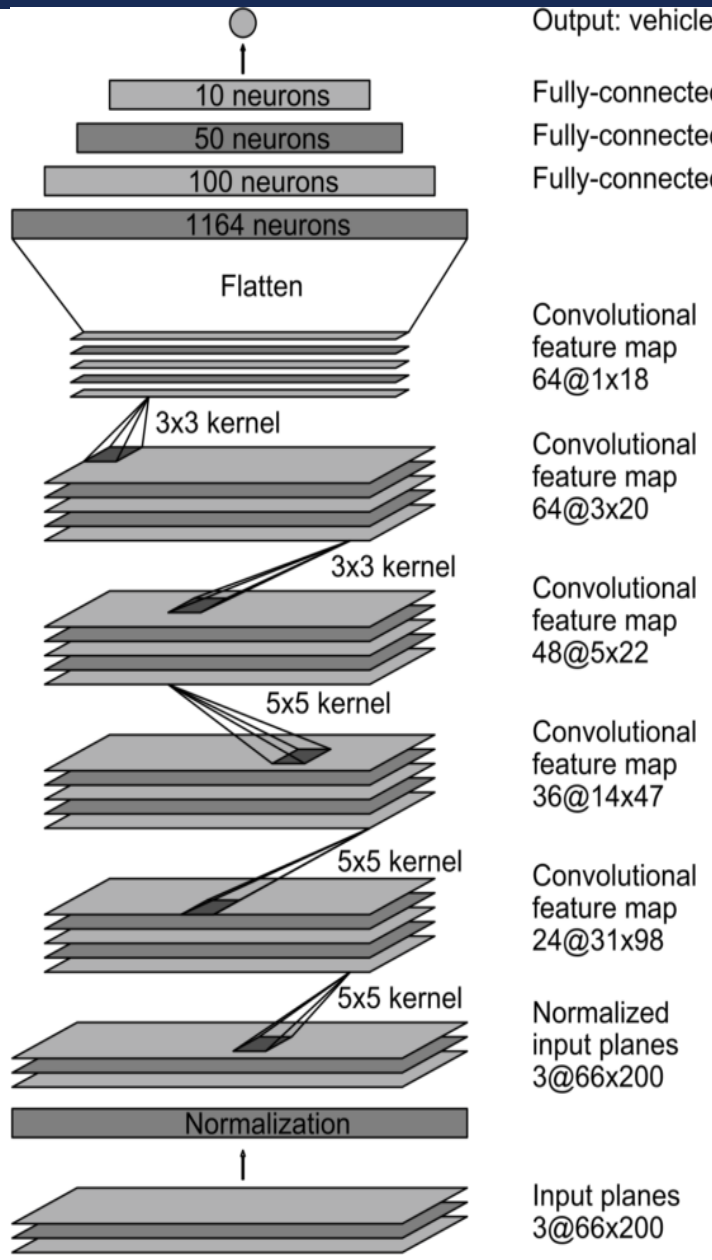
# Assurance of Autonomous Systems Controlled by CNNs


NVIDIA 'BB8' AI uses End-to-End Deep Learning to steer self-driving car

ML-based computer vision and image processing architectures for autonomous system operations rely on complex machine-learning algorithms that are poorly understood

# Example: NVIDIA's PilotNet

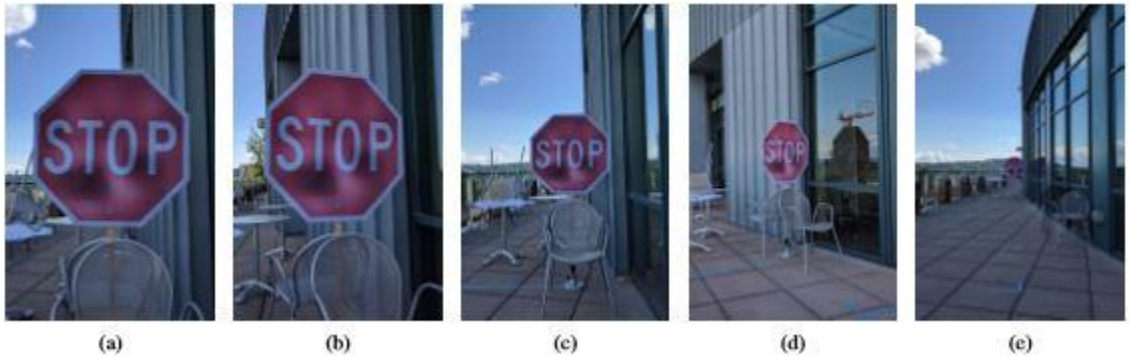


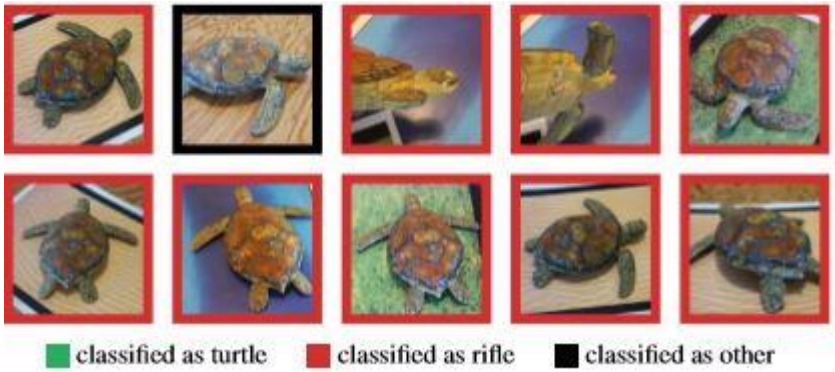
U.S. NAVAL RESEARCH LABORATORY

# CNN Assurance Challenges

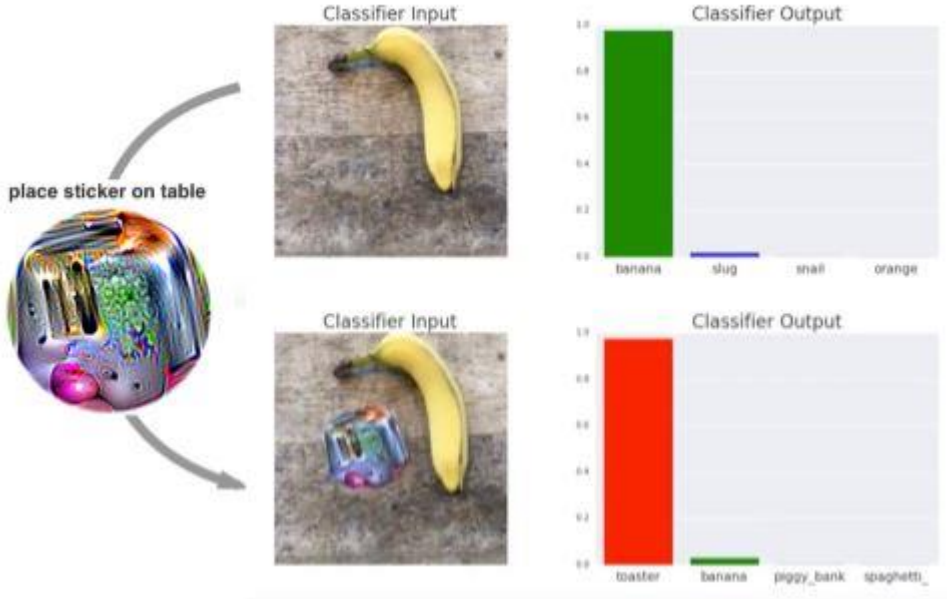

Neural Networks can be fooled




Targeted Physical Perturbations STOP sign misclassified as Speed Limit 45 sign




Poses of a single 3D-printed turtle adversarially perturbed to be misclassified as a rifle



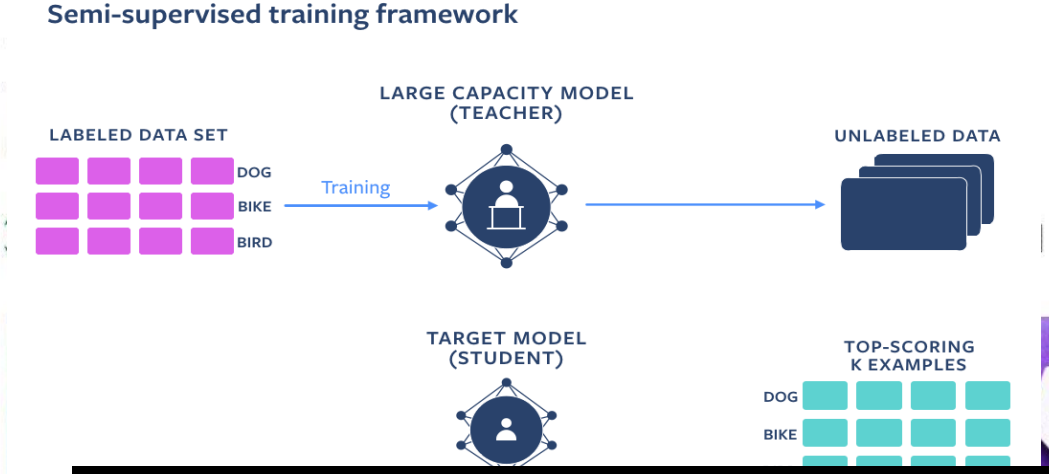

Adversarial Examples: Attack at a distance!

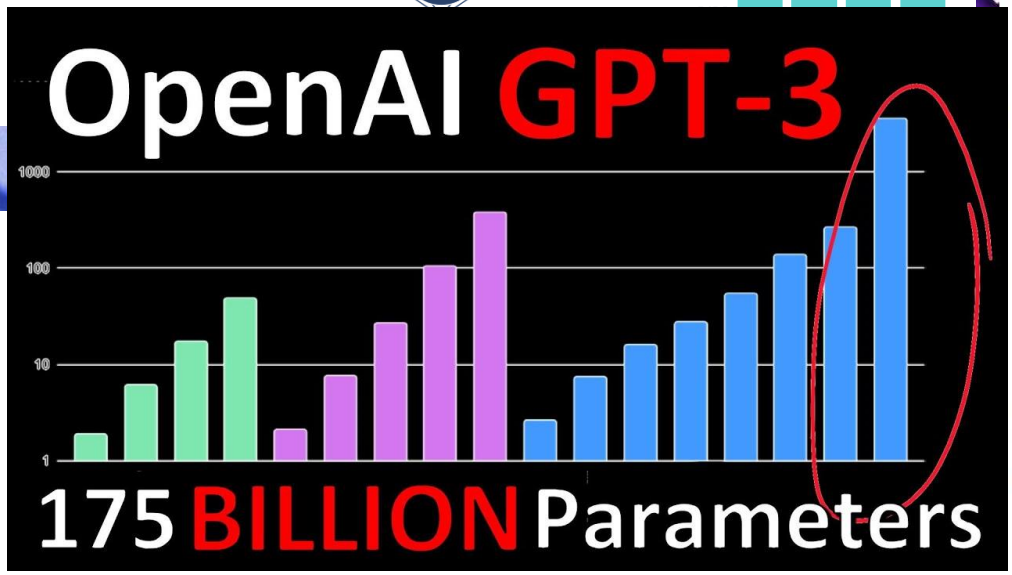
U.S. NAVAL RESEARCH LABORATORY

# Spectacular Advances in Image and NLP

GENERAL INTELLIGENCE

## Facebook Scraped 1 Billion Pictures From Instagram to Train Its A.I. —

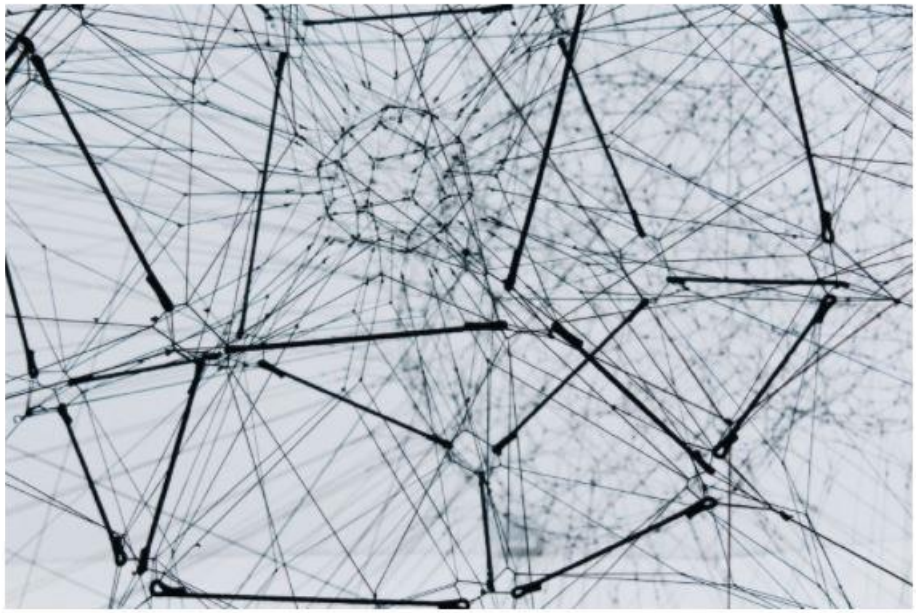



## How to Try CLIP: OpenAI's Zero-Shot Image Classifier

Earlier this week, OpenAI dropped a bomb on the computer vision world — you can now make image classifications with no training required.

Jacob Solawetz Jan 8 · 6 min read



(cite)

Earlier this week, OpenAI dropped a bomb on the computer vision world: two new groundbreaking models that hint at what's to come as massive GPT3-esque Transformer models encroach on the vision domain. While

U.S. NAVAL RESEARCH LABORATORY

# Spectacular Failures in Image and NLP

| | |
|---|---|
| Granny Smith | 85.6% |
| iPod | 0.4% |
| library | 0.0% |
| pizza | 0.0% |
| toaster | 0.0% |
| dough | 0.1% |

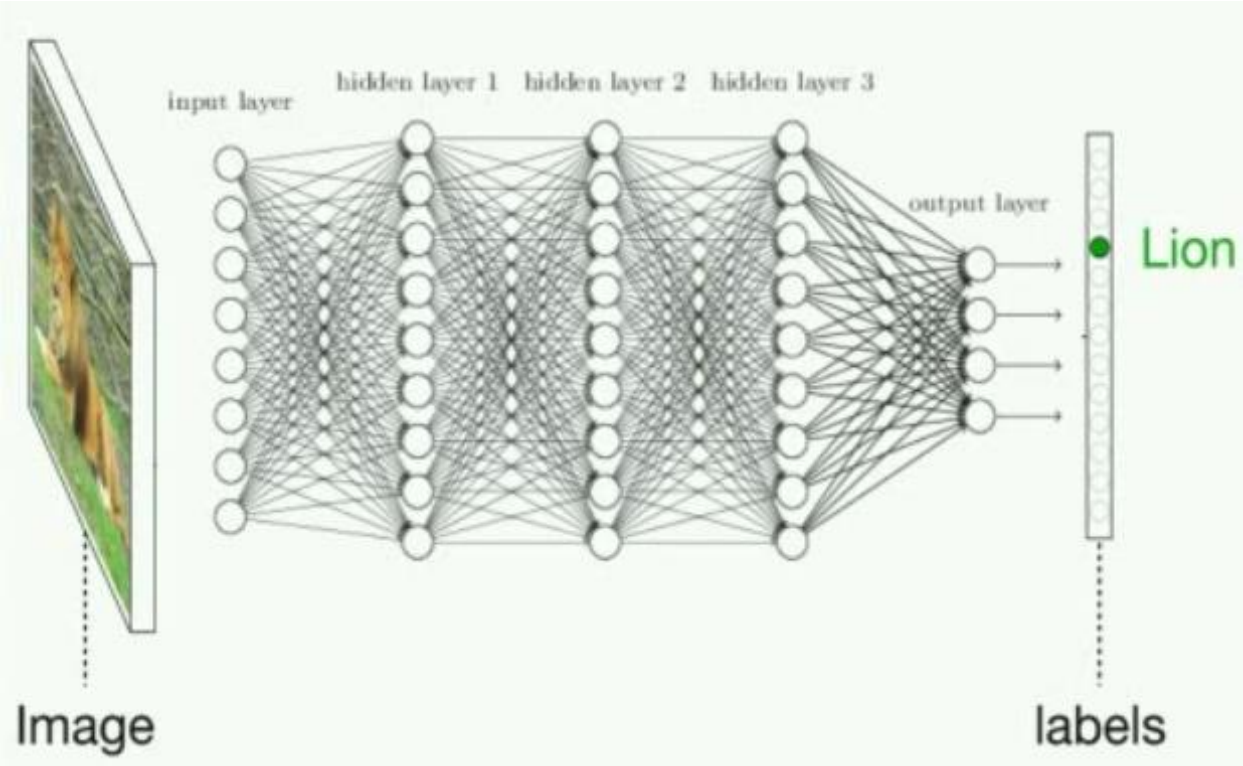

| | |
|---|---|
| Granny Smith | 0.1% |
| iPod | 99.7% |
| library | 0.0% |
| pizza | 0.0% |
| toaster | 0.0% |
| dough | 0.0% |

| | |
|---|---|
| chainsaw | 91.1% |
| lawn mower | 7.0% |
| power drill | 1.0% |
| vacuum cleaner | 0.4% |
| wheelbarrow | 0.1% |
| tractor | 0.1% |


| | |
|---|---|
| piggy bank | 70.1% |
| chainsaw | 1.5% |
| slot machine | 1.1% |
| wheelbarrow | 0.9% |
| hammer | 0.8% |
| mousetrap | 0.6% |

U.S. NAVAL RESEARCH LABORATORY

# A Crash Course on DNNs




A Deep Neural Network N with n input nodes, m output nodes and h hidden nodes computes a function

$$(y_1, \ldots, y_m) = f_N(x_1, \ldots, x_n), \text{ with } f_N : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

A Rectified Linear Unit computes a non-linear function given by

$$z_i = \max(0, \sum a_{ij}x_j + \sum b_{ij}z_j + c_i)$$

Slide Courtesy Michael Colon

U.S. NAVAL RESEARCH LABORATORY

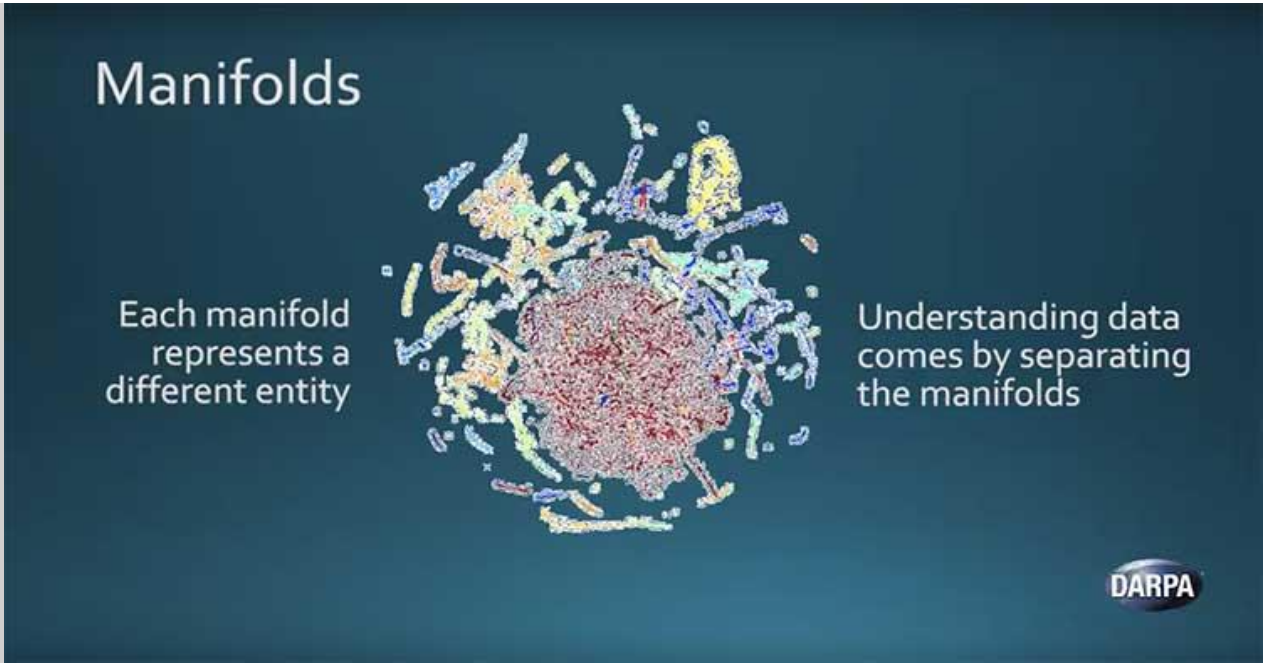
# Generalization: Fundamental Goal of ML

- ***Generalize*** beyond the examples in the training set
- No matter how much training data we provide, unlikely we'll see the same data during test
- Common mistake: test using training data, providing an illusion of success
- "No free lunch" theorem of Wolpert and Macready –
  No learner can beat random guessing over all possible functions to be learned
  - How so?
    - Consider learning a Boolean function of 100 variables from a million examples
    - There are $2^{100}$ possible classes to be learned from $10^6$ examples
    - There are $2^{100} - 10^6 = 1.3 \times 10^{30}$ examples whose classes are unknown
    - Clearly, there is no way to do this that beats flipping a coin
  - Or is there? *Why* does deep learning work?

The "manifold hypothesis* " (MH) provides a plausible explanation.

MH: High-dimensional natural data tend to clump and be shaped differently when visualized in lower dimensions.

* Manifold Assumption and Defenses Against Adversarial Perturbations [ICLR 2018]



U.S. NAVAL RESEARCH LABORATORY

# Is CNN Assurance Attainable?

1. Multilayer feedforward networks are a class of universal approximators.
   - They are capable of approximating any measurable finite-dimensional function*
     * Kurt Hornik “Multilayer Feedforward Networks are Universal Approximators,” Neural Networks, Vol. 2, pp. 359-366, 1989
   - This does not imply that there’s an effective training procedure to learn this function
2. Academic papers on adversarial examples are sensational and fatalistic
   - Many of them work on simpler classifiers, e.g., PCA, SVMs, linear regression
   - Many of the adversarial learning algorithms additionally perturb background imagery


3. Conjecture: Generalization works only if underlying random processes are Ergodic
   - Ergodic random processes exhibit both stationarity and ergodicity
   - Stationarity property: Guarantees that statistical properties do not change
   - Ergodicity: Every sequence or sizable sample is equally representative of the whole
   - Underlying measurable spaces are mapped by Random Variables or Measurable Functions
   - For non-Ergodic processes, it is conjectured that a CNN trained on examples will generalize poorly

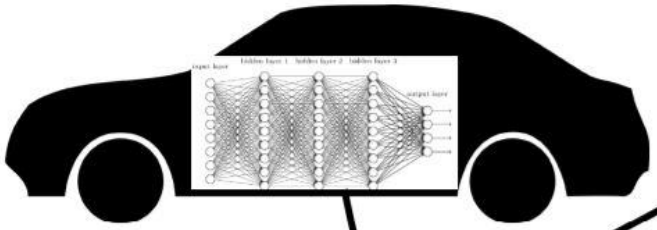
Hypothesis: CNNs can be trained to be robust against adversarial attacks

U.S. NAVAL RESEARCH LABORATORY

# Can CNNs be formally verified?

## Challenges for Verified AI

S. A. Seshia, D. Sadigh, S. S. Sastry.
*Towards Verified Artificial Intelligence*. July 2016. https://arxiv.org/abs/1606.08514.



System **S**

Environment **E**

Specification φ

Does S || E satisfy φ?

YES [+ proof]

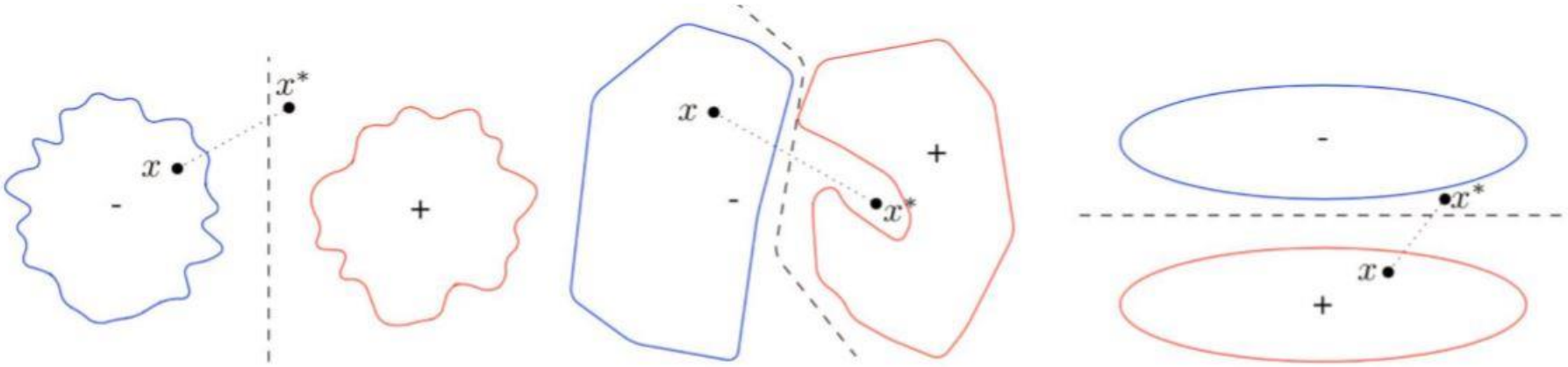
NO
[+ counterexample]

**Need to Search Very High-Dimensional Input and State Spaces**

**Design Correct-by-Construction?**

U.S. NAVAL RESEARCH LABORATORY

# Understanding CNN Vulnerability



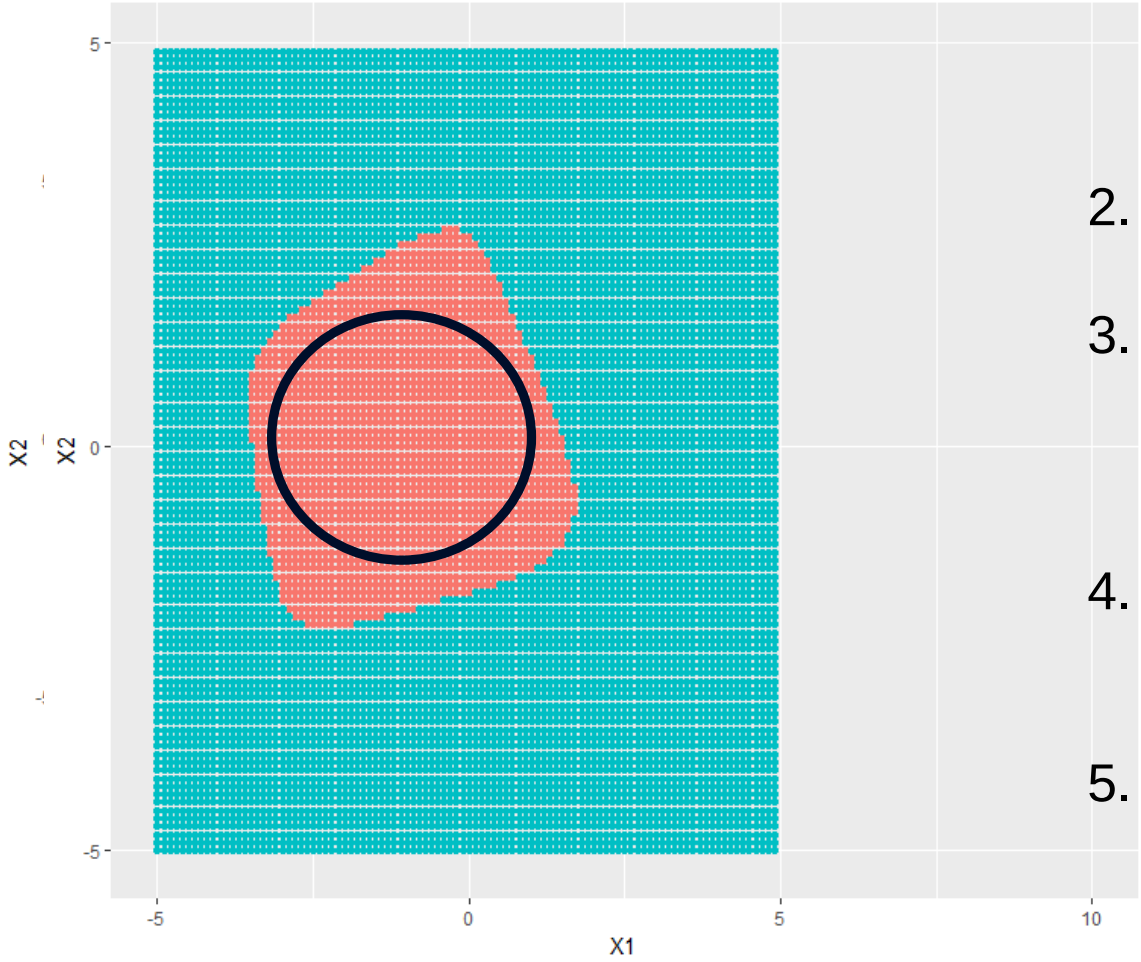

(a) Two simple 2D submanifolds. (b) One submanifold has a 'pocket'. (c) Nearby 2D submanifolds.

1. Based on the hypothesis that adversarial examples lie off data manifolds
2. We show three interesting configurations of adversarial examples:
   a. The adversarial example lies far from each sub-manifold
   b. Far from the classification boundary but near an adjacent manifold, but not on it
   c. Near the classification boundary and near the adjacent manifold, but not on it
3. Representative defenses include (a) Defensive Dropouts and (b) Distillation as a Defense
   - However, these defenses are vulnerable to the attacks of Carlini and Wagner
   - We plan to explore a new defense approach known as Stochastic Activation Pruning

* Carlini, N and Wagner, D. "Towards Evaluating the Robustness of Neural Networks," IEEE Symposium on Security and Privacy, 2017

U.S. NAVAL RESEARCH LABORATORY

# Evaluating CNN Robustness
## *Underlying Cause of Misclassification*



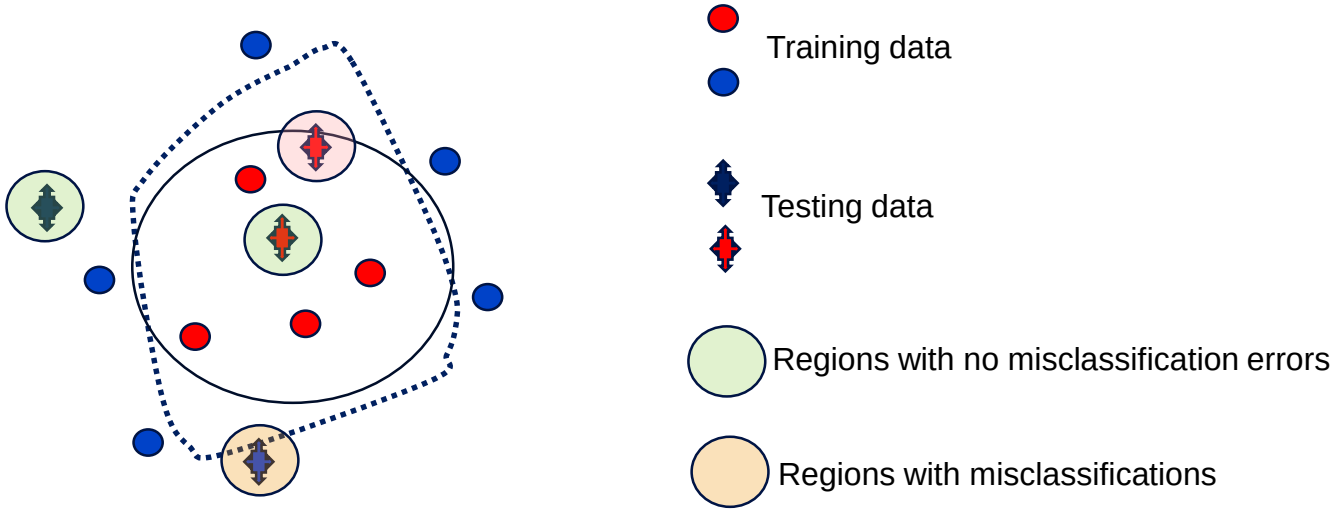

Notes:

1. We trained a three-layer feed-forward network with inputs drawn from one of two 2D Gaussian processes A and B
2. The decision problem is to assign each input to hypothesis $H_A$ or $H_B$
3. The learned classification boundary is shown overlaid over the theoretical (Bayesian) optimum boundary
4. We determined underlying cause for classification errors during testing
5. Computing this mapping is feasible for inputs of low dimensions; infeasible for higher input dimensions

U.S. NAVAL RESEARCH LABORATORY

# Evaluating Robustness for CNNs of Low Dimensionality

**Assess misclassification by safe perturbations of input data samples**




Technical approach addresses challenges of high dimensionality, nonlinearity, huge scale

U.S. NAVAL RESEARCH LABORATORY

# Local Robustness: Mathematical Formulation

- Human Cognition: $f = \mathbb{R}^n \rightarrow C$
- Multi-layer Feed-Forward Network computes an approximation of $f$: $\hat{f} = \mathbb{R}^n \rightarrow C$
- M training examples: $\{ (x^i, c^i) \}_{i = 1, n}$
- Adversarial perturbations:
  - $x^i \rightarrow \hat{f} \rightarrow c^i$ i.e., $\hat{f}(x^i) = c^i$
  - $\hat{f}(x^i + \mathbf{\Delta} x^i) \neq c^i$ while $f(x^i + \mathbf{\Delta} x^i) = c^i$
  - where $\mathbf{\Delta} x^i$ is an adversarial perturbation
  - $x^i + \mathbf{\Delta} x^i$ is an adversarial example
- Resizing, cropping, changing lighting, maliciousness are sources of adversarial perturbations
- Problem formulation: Probability of misclassification of adversarial example should be low
- Statistical robustness: Average minimum distance ($\mathbf{\Delta} x^i$) for misclassification should be high

Original image classified as a panda with 60% confidence.

Tiny adversarial perturbation.

Imperceptibly modified image, classified as a gibbon with 99% confidence.

U.S. NAVAL RESEARCH LABORATORY

# Verification of DNN Local Robustness

A DNN N with n input nodes, m output nodes and h hidden nodes can be represented as logical formula $F_N$, which is a propositional combination of linear constraints:

$$(w_1 \leq 0 \wedge z_1 = 0 \vee w_1 > 0 \wedge z_1 = w_1) \wedge w_1 = \sum a_{1j}x_j + \sum b_{1j}z_j + c_1 \wedge$$

...

$$(w_h \leq 0 \wedge z_h = 0 \vee w_h > 0 \wedge z_h = w_h) \wedge w_h = \sum a_{hj}x_j + \sum b_{hj}z_j + c_h \wedge$$

$$y_1 = \sum d_j z_j + e_1 \wedge$$

...

$$y_n = \sum d_j z_j + e_n$$

A DNN N satisfies the property (P, Q) iff $Q(f_N(x_1, \ldots, x_n))$ whenever $P(x_1, \ldots, x_n)$

- Example: robustness with respect to adversarial inputs:
  $P(x_1, \ldots, x_n)$: input image is within δ units of "Panda" image
  $Q(y_1, \ldots, y_m)$: output classification is "Panda"

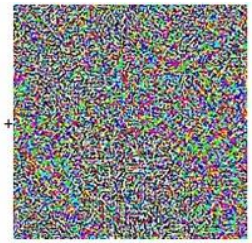
If P and Q are expressible as propositional combinations of linear constraints, then N can be verified using a satisfiability checker for linear constraints

Slide Courtesy Michael Colon

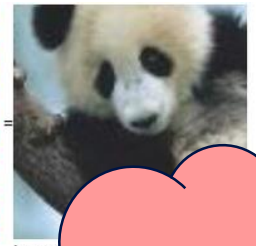
U.S. NAVAL RESEARCH LABORATORY

# System Hardening for Learning-Enabled Systems

Original image classified as a panda with 60% confidence.



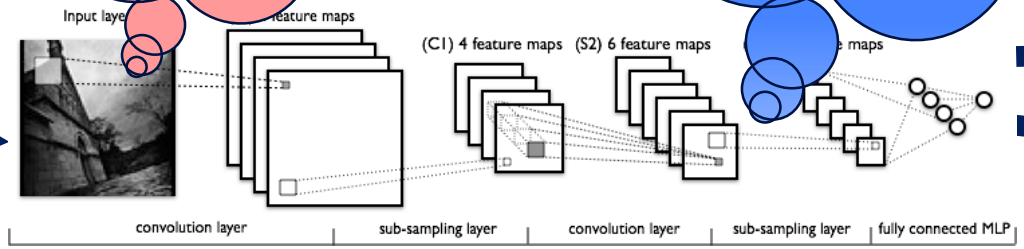
Tiny adversarial perturbation.



Adversarial Perturbations using $L_1, L_2, and\ L_\infty$ Norms *

Semantic Perturbations

Defenses:
Adversarial Learning
Domain Adaptation
Filters/Noise Injection
Stochastic Activation
Pruning
Decision Safety




No

Perturbation results in misclassification?

Yes

Augmented Training Examples

Key Property: NN is invariant to perturbations indistinguishable by a human

Assumption of "minimality of manipulations"

- Local adversarial robustness at a given point
- Exhaustive search for adversarial misclassifications (with a given norm)

Verification: Guarantee a misclassification is found if exists

Falsification: Re-work the network for mitigation

U.S. NAVAL RESEARCH LABORATORY

# Naval Relevance

Naval Unmanned and Autonomous Systems are deployed for missions that are "dirty, dull, or dangerous."

- ASV Unmanned Systems

- Marine Corps MAGTF

- ONR autonomous boats

- DARPA unmanned vessel

- ONR underwater vehicles

Machine learning is an increasingly important component of a broad range of defense systems, including autonomous systems [...] the DoD laboratories should establish research and experimentation programs around the practical use of machine learning in defense systems with efficient testing, independent verification and validation (IVV), and resiliency and hardening as the primary focus points. […] They should create and promulgate a methodology and best practices for the construction, validation, and deployment of machine learning systems, including architectures and test harnesses.
DSB Report on **Design and Acquisition of Software for Defense Systems (February 2018)**

**Guarantee safety of autonomous system operations and performance**

U.S. NAVAL RESEARCH LABORATORY

# TADM 2021

Call for Participation

TADM 2021: Trusted Automated Decision-Making
Co-located with ETAPS 2021 Virtually in Luxembourg, Luxembourg, March 27-28, 2021

We invite you to participate in TADM 2021. The format of the workshop will be informal, to solicit preliminary work and to foster future collaboration among disparate disciplines. We're delighted to have the following three keynote speakers:

Prof. Michael I Jordan, Berkeley
Prof. Cynthia Rudin, Duke
Prof. Wendell Wallach, Yale

TADM workshop website and ETAPS registration links follow:

https://3drationality.com/TADM2021

https://etaps.org/2021/registration