U.S. NAVAL RESEARCH LABORATORY

Breaking and Fixing Autonomous Cyber-Physical Tactical Systems

Dr. Ramesh Bharadwaj

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

Ramesh.Bharadwaj@nrl.navy.mil

Speaker Introduction

Ramesh Bharadwaj

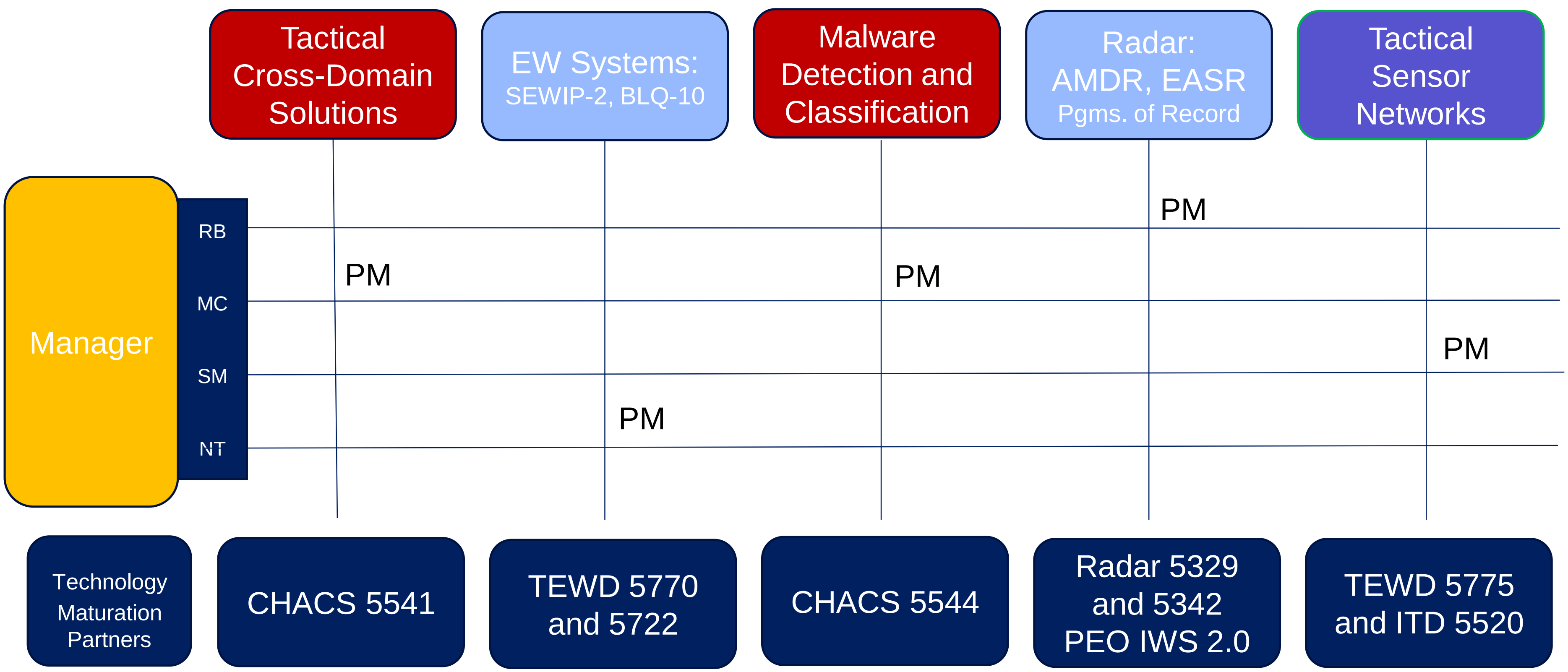
- PhD, Computer Engineering, Communications Research Laboratory, McMaster University, Hamilton, ON Canada
- MEE, Electronics Engineering, Philips/Eindhoven International Institute, Eindhoven, The Netherlands
- BE, Electronics and Communications Engineering, National Institute of Engineering, Mysore, India

- Current Position: Researcher, Assured Autonomous Systems

- Previous Positions:
  - Research: Philips Research Laboratories (Eindhoven), Tata Institute of Fundamental Research (Mumbai), Stanford University (Palo Alto), AT&T Bell Laboratories (Murray Hill), Fraunhofer FOKUS (Berlin)
  - Teaching: National Centre for Software Technology (Mumbai), KTH Royal Institute of Technology (Kista), George Washington University and Catholic University of America (Washington DC)
- Background:
  - Ten years' experience in Modeling & Simulation and Electronic Warfare (EW) systems
  - Five years' experience in Virtual Integration of Electronic Warfare Systems (ViEWS)
  - Subject Matter Expert on multifunction radars and EW systems including AN/SPS49A(V)1 and AN/SLQ-32(V)6

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



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.”

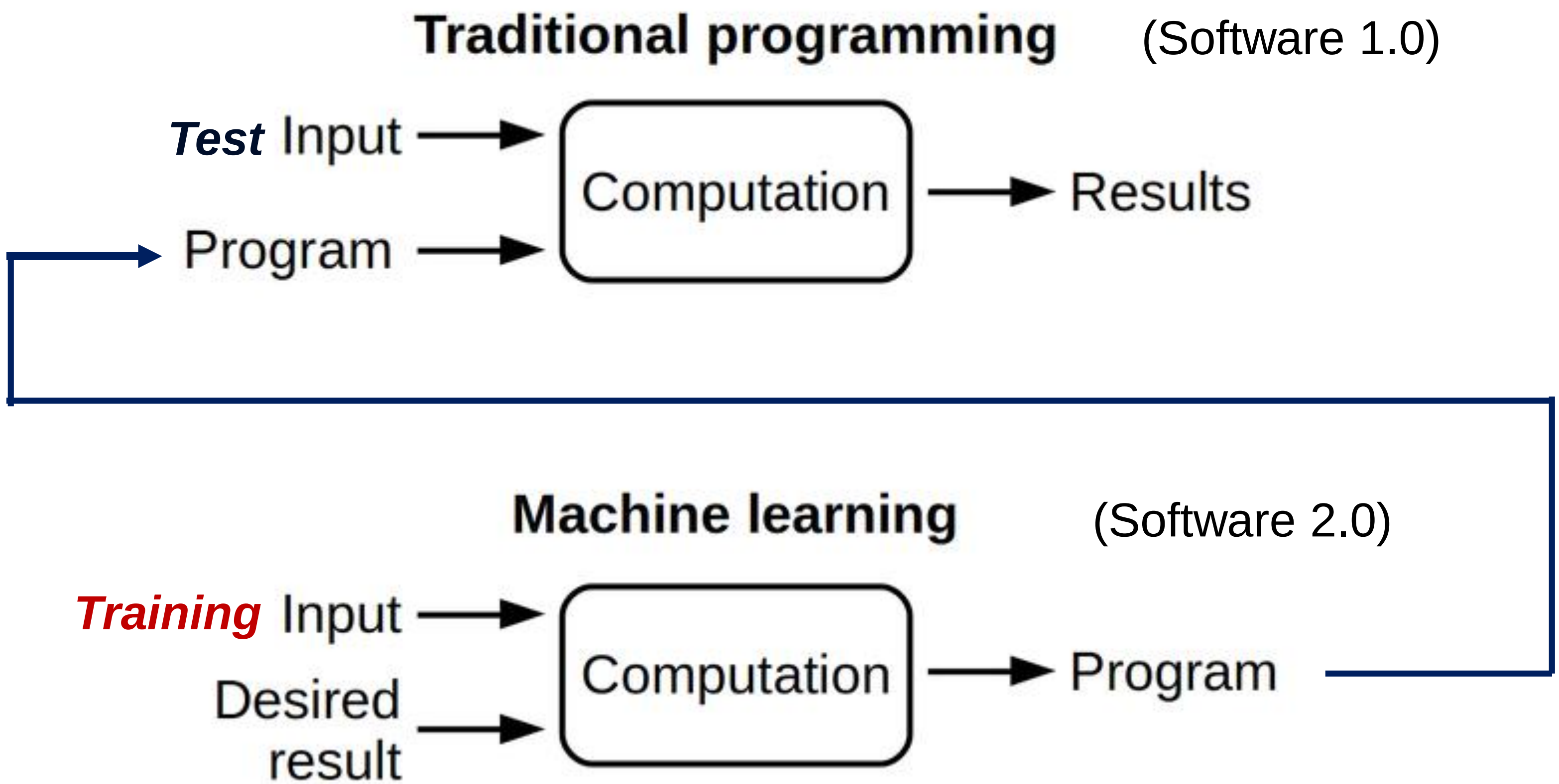
Goodwood Aerodrome, West Sussex, 4 July 2019

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

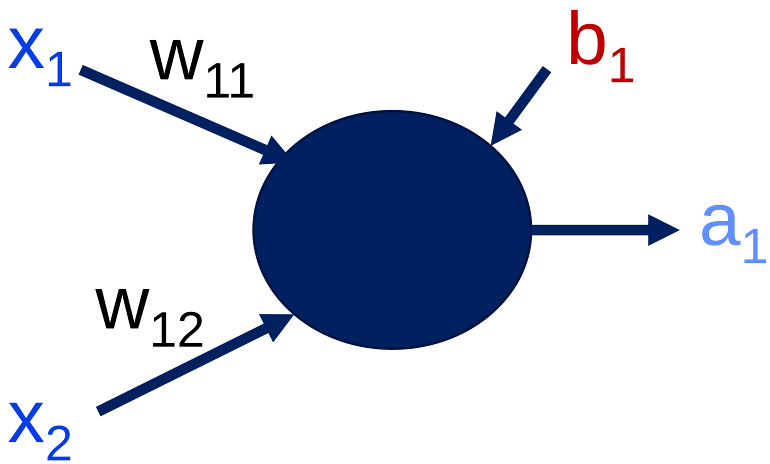
What is Machine Learning (ML)

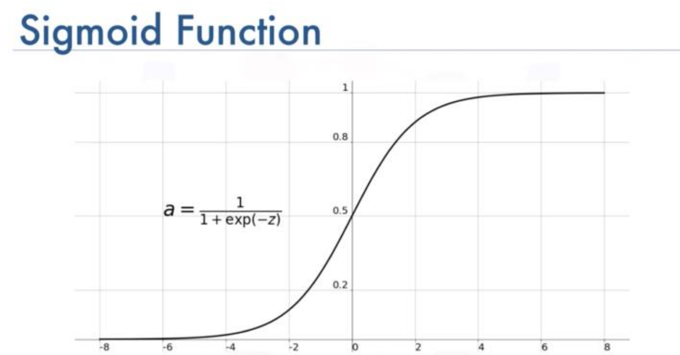
A Branch of Artificial Intelligence (AI)



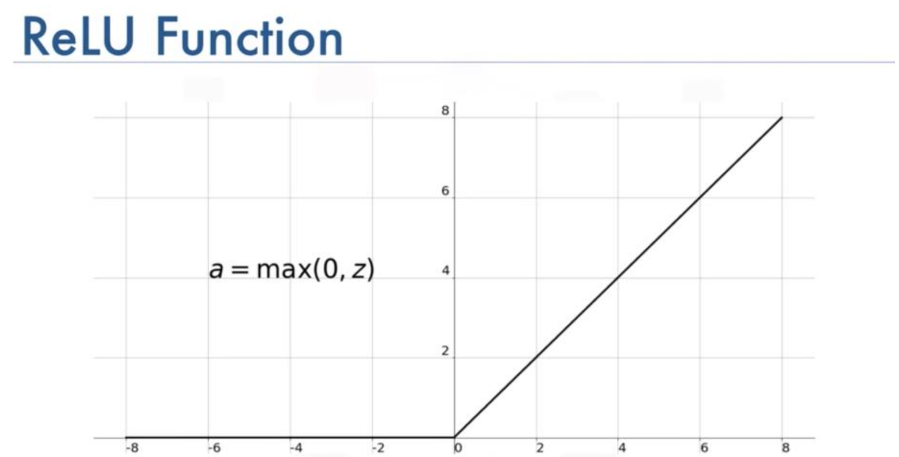
What is a Neuron?

Perceptron (a.k.a., “Neuron”)





$$a_1 = f(x_1 * w_{11} + x_2 * w_{12} + b_1)$$



OUT = Red if $a_1 < 0.5$

Blue if $a_1 \geq 0.5$

$$a_i = \max(0, \sum w_{ij} x_j + b_i)$$

Who Invented the Perceptron?

Frank Rosenblatt at Cornell (1957)

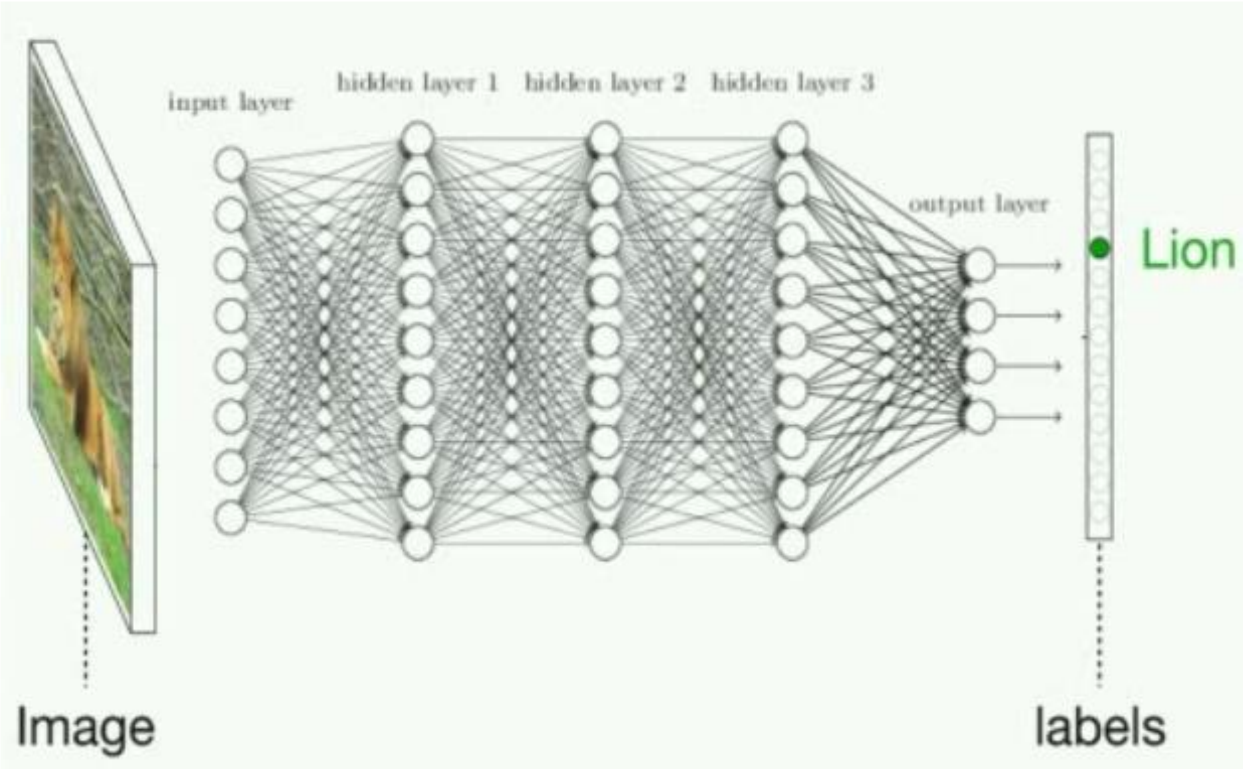
Funded by the Office of Naval Research!!!

U.S. NAVAL RESEARCH LABORATORY

Perceptron Demo

https://www.cs.utexas.edu/~teammco/misc/perceptron/

Multi-Layer Perceptron
a.k.a. Deep Neural Network (DNN)



D. Anderson and G. McNeill, Artificial Neural Networks Technology, ELIN: A011, Rome Laboratory, NY, August 1992.

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"

The Four Fallacies of AI

Fallacy 1: Narrow intelligence is on a continuum with general intelligence

- Deep Blue was "was hailed as the first step of an AI revolution"
- Watson system [is] "a first step into cognitive systems…….."
- OpenAI's GPT-3 [is] a "step toward general intelligence"

Fallacy 2: Easy things are easy and hard things are hard

- John McCarthy (who coined the term "Artificial Intelligence") lamented that "AI was harder than we thought"
- Marvin Minsky explained that this is because "easy things are hard"

Fallacy 3: The lure of wishful mnemonics

- "Neural Networks" have nothing to do with neurons or the brain
- "Machine Learning" and "Deep Learning" do not resemble human learning
- "Watson can read all of the health-care texts in the world in seconds"
- "AlphaGo's goal is to beat the best human players not just mimic them"
- "We can always ask AlphaGo how well it thinks it's doing during the game. ...It was only at the end of the game that AlphaGo thought it would win"

Fallacy 4: Intelligence is all in the conscious mind

- "A physical symbol system has the necessary and sufficient means for general intelligent action"
- Herb Simon (a Nobel winning economist) said: "[To] understand cognition, we don't have to worry about unconscious perceptual processes."

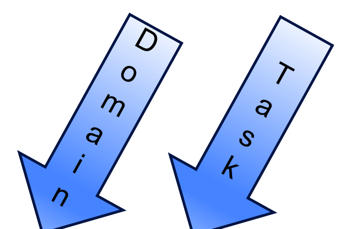
Assurance Objective

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

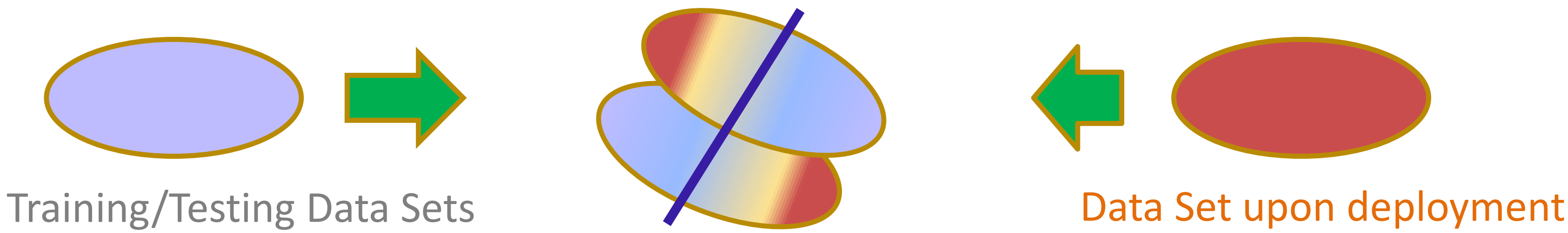


(1D, 2D and 3D variants; generic 2D variant for $\boxed{\text{image}}$ $\boxed{\text{classification}}$)

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. Discussed at TADM 2021!!

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

Department of Defense Directive 3000.09: Autonomy in Weapon Systems, November 21, 2012
"Establish guidelines [to] … minimize consequences of failure that may lead to unintended engagements."

TADM 2021

Organizers: Ramesh Bharadwaj (NRL) and Ilya Parker (3D Rationality LLC)

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

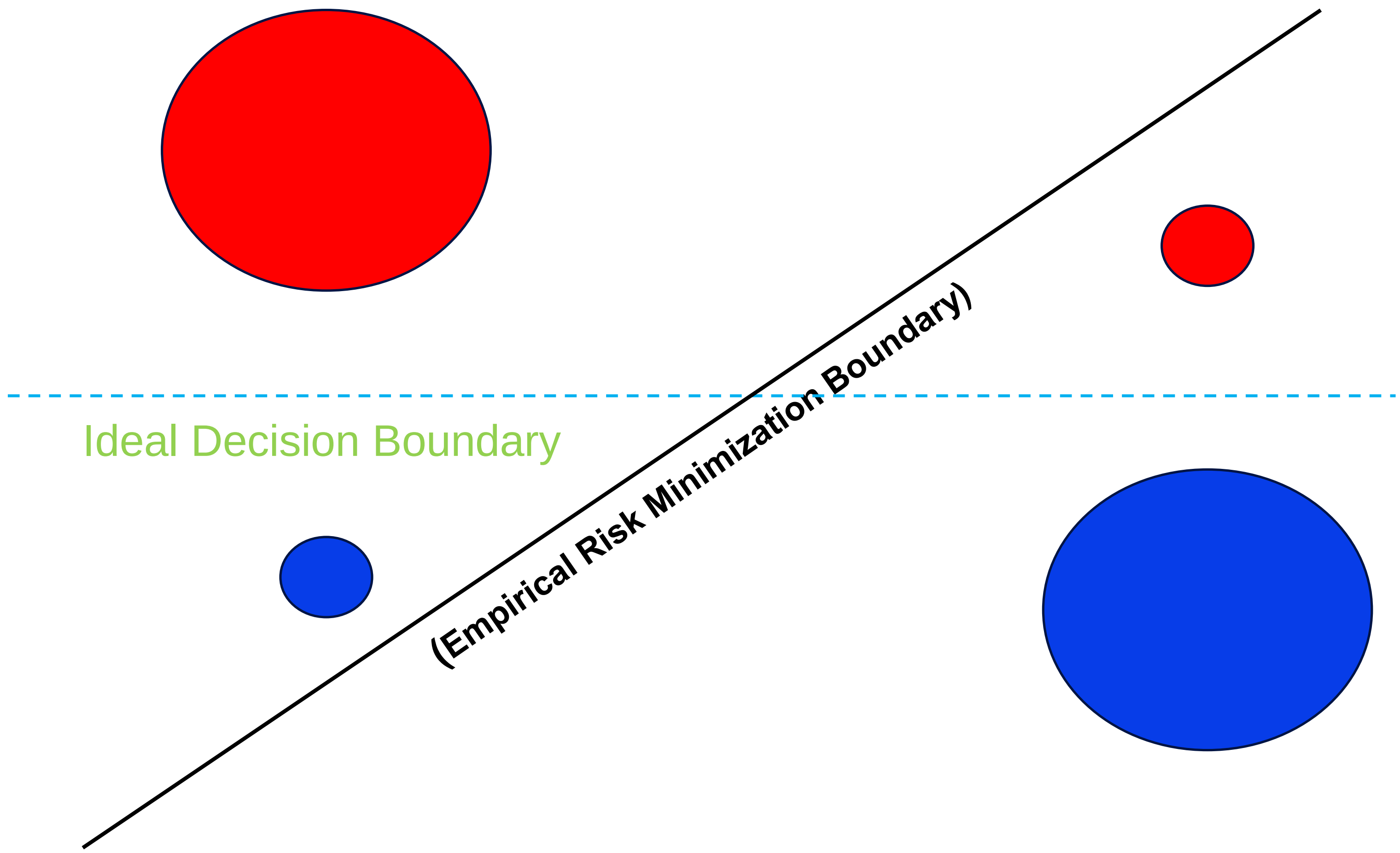
Prof. Michael I Jordan, Berkeley
Prof. Cynthia Rudin, Duke
Prof. Wendell Wallach, Yale

TADM workshop website:

https://3drationality.com/TADM2021

DNN Assurance Challenge: “Good Enough” Decisions are not Accurate

Subclass labels for the data are often unavailable



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: Taking commercial (or other) technologies developed for Virtual-only Systems and attempting to implement them in the Cyber-Physical Domain

U.S. NAVAL RESEARCH LABORATORY

CNN Assurance Challenge: Adversarial Perturbations

“Cat”

Add Noise

“Panda”

Adversarial Examples: Attack at a distance!

“We demonstrate that 13 defenses recently published at ICLR, ICML and NeurIPS---and which illustrate a diverse set of defense strategies---can be circumvented despite attempting to perform evaluations using adaptive attacks.”

Florian Tramer, Nicholas Carlini, Wieland Brendel, Aleksander Madry,
“On Adaptive Attacks to Adversarial Example Defenses,” in Proceedings Advances in Neural Information Processing Systems 33 (NeurIPS 2020)

U.S. NAVAL RESEARCH LABORATORY

Local Robustness: Mathematical Formulation

“Cat”

Add Noise

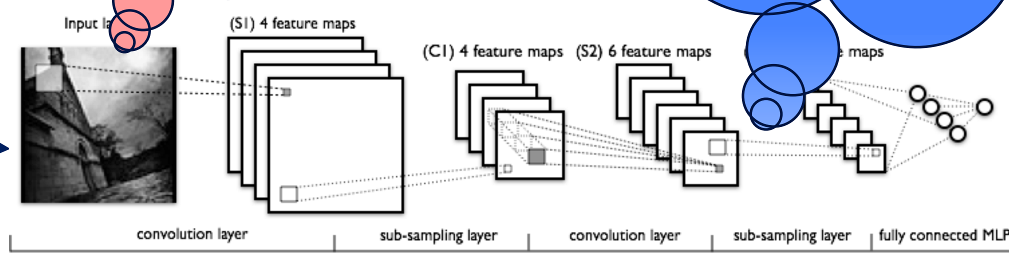
“Panda”

- Human Cognition: $f = \mathbb{R}^n \to C$
- Multi-layer Feed-Forward Network computes an approximation of f: $\hat{f} = \mathbb{R}^n \to C$
- M training examples: $\{ (x^i, c^i) \}_{i = 1, n}$
- Adversarial perturbations:
  - $x^i \to \hat{f} \to c^i$ 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

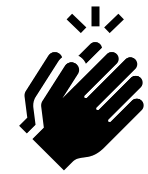
System Hardening Process for DNNs

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

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