Prof. Philip Koopman

Software Risk and Autonomy

NASA Autonomy Workshop
Oct 11, 2018

Carnegie Mellon University

Electrical & Computer ENGINEERING

Edge Case Research

Overview

- **Control & Planning safety**
  - Breaking robots for fun and profit

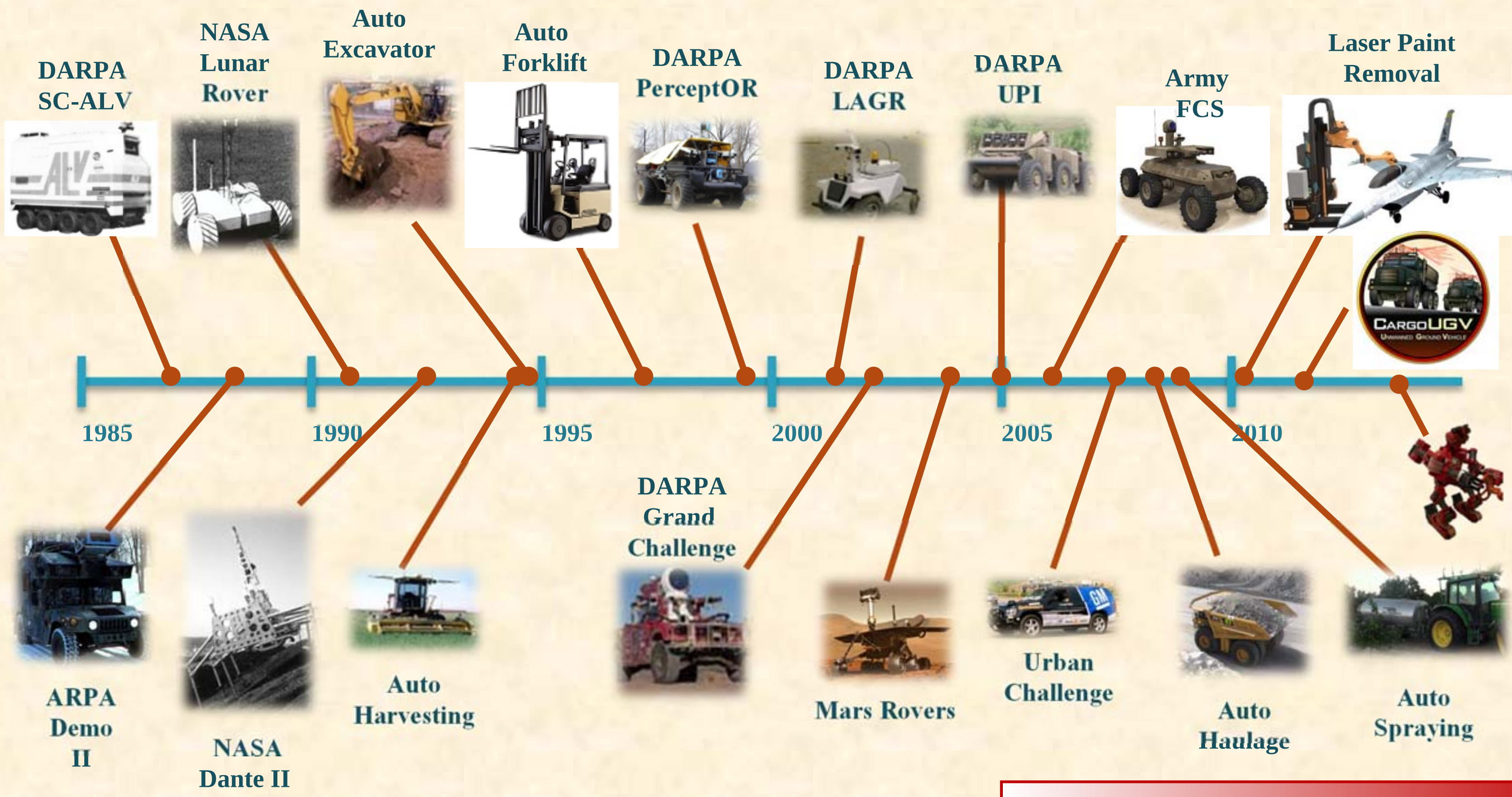
- **Perception safety**
  - It's a bird. It's a plane.
    It's … what the heck is that?

- **Edge cases**
  - Back to breaking robots for fun and profit

[General Motors]

NREC: 30+ Years Of Cool Robots

NATIONAL ROBOTICS ENGINEERING CENTER



Software Safety

Carnegie Mellon University Faculty, staff, students
Off-campus Robotics Institute facility

Before Autonomy Software Safety

- The **Big Red Button** era

APD (Autonomous Platform Demonstrator)

Safety critical speed limit enforcement

NATIONAL ROBOTICS ENGINEERING CENTER

TARGET GVW: 8,500 kg

TARGET SPEED: 80 km/hr

Approved for Public Release. TACOM Case #20247 Date: 07 OCT 2009

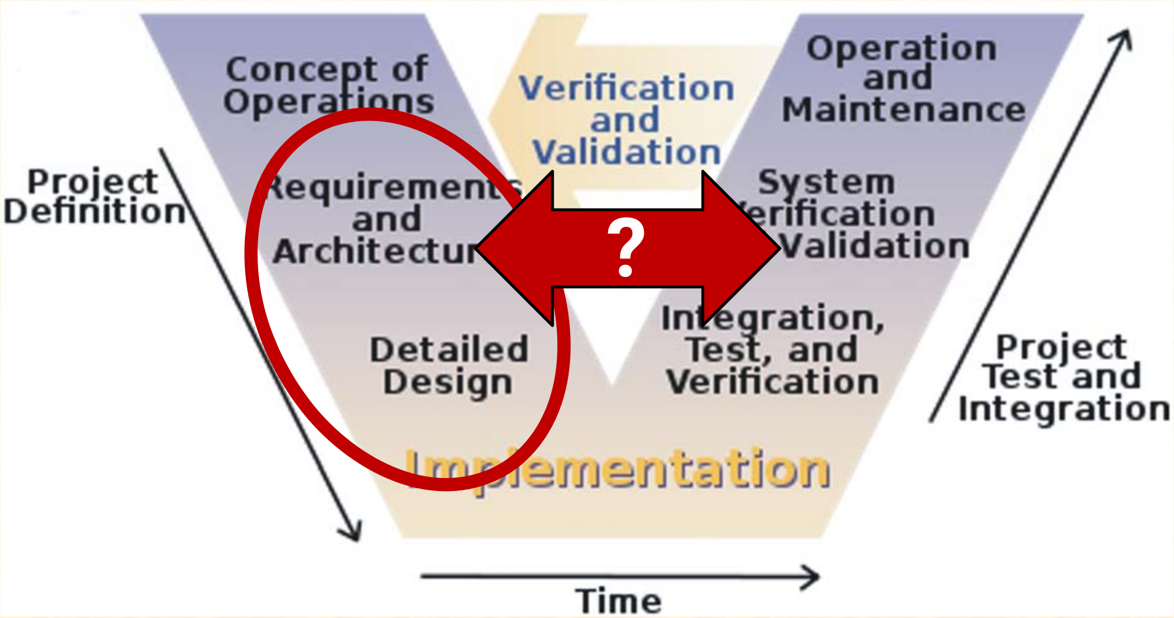
Traditional Validation Meets Machine Learning

Carnegie Mellon University

- Use traditional software safety where you can

..BUT..

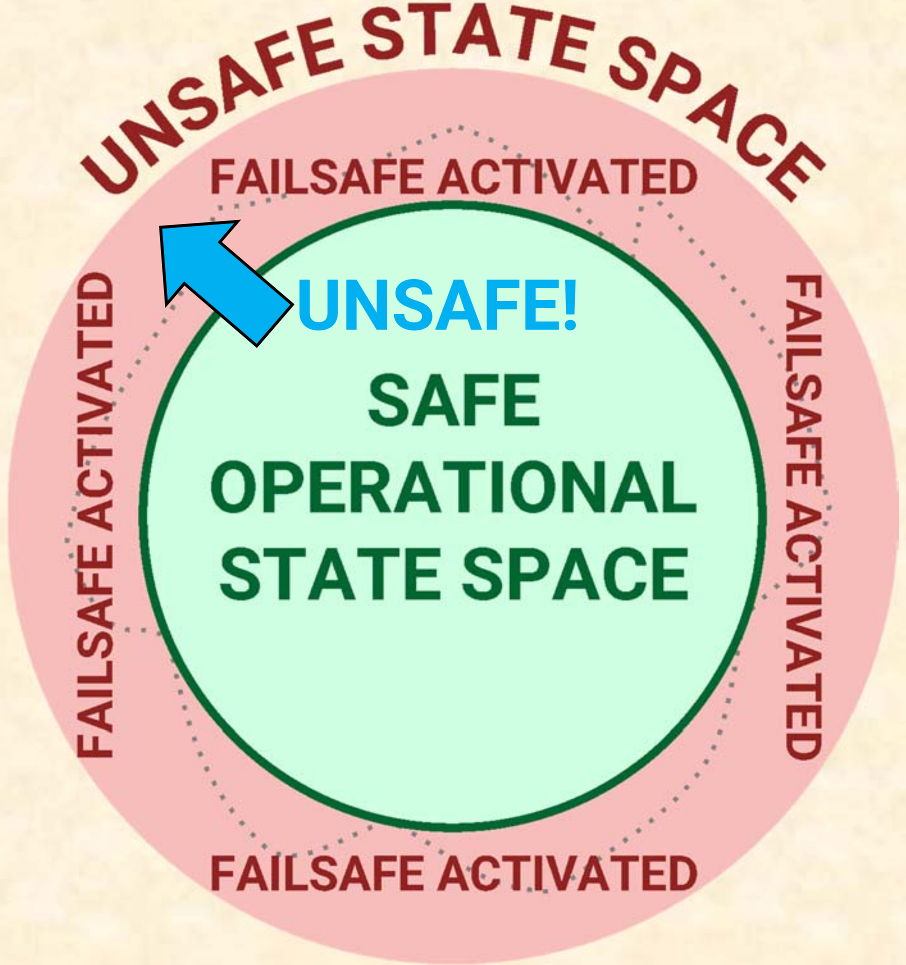
- Machine Learning (inductive training)
  - No requirements
    - Training data is difficult to validate
  - No design insight
    - Generally inscrutable; prone to gaming and brittleness



Safety Envelope Approach to ML Deployment

Carnegie Mellon University

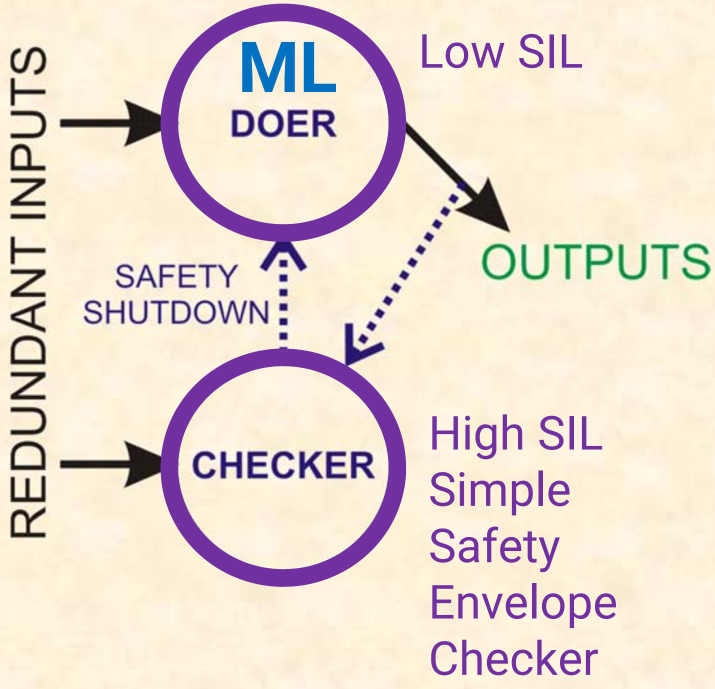
- **Specify unsafe regions**
- **Specify safe regions**
  - Under-approximate to simplify
- **Trigger system safety response upon transition to unsafe region**



Architecting A Safety Envelope System

- **"Doer" subsystem**
  - Implements normal, untrusted functionality
- **"Checker" subsystem – Traditional SW**
  - Implements failsafes (safety functions)
- **Checker entirely responsible for safety**
  - Doer can be at low Safety Integrity Level
  - Checker must be at higher SIL

Doer/Checker Pair

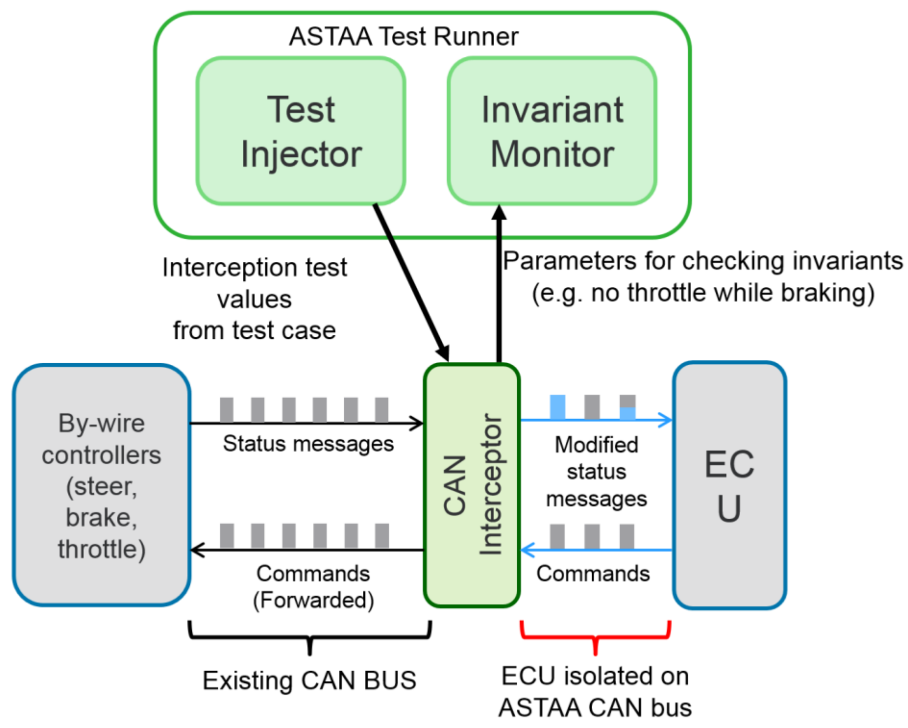


(Also known as a "safety bag" approach)

Robustness Testing

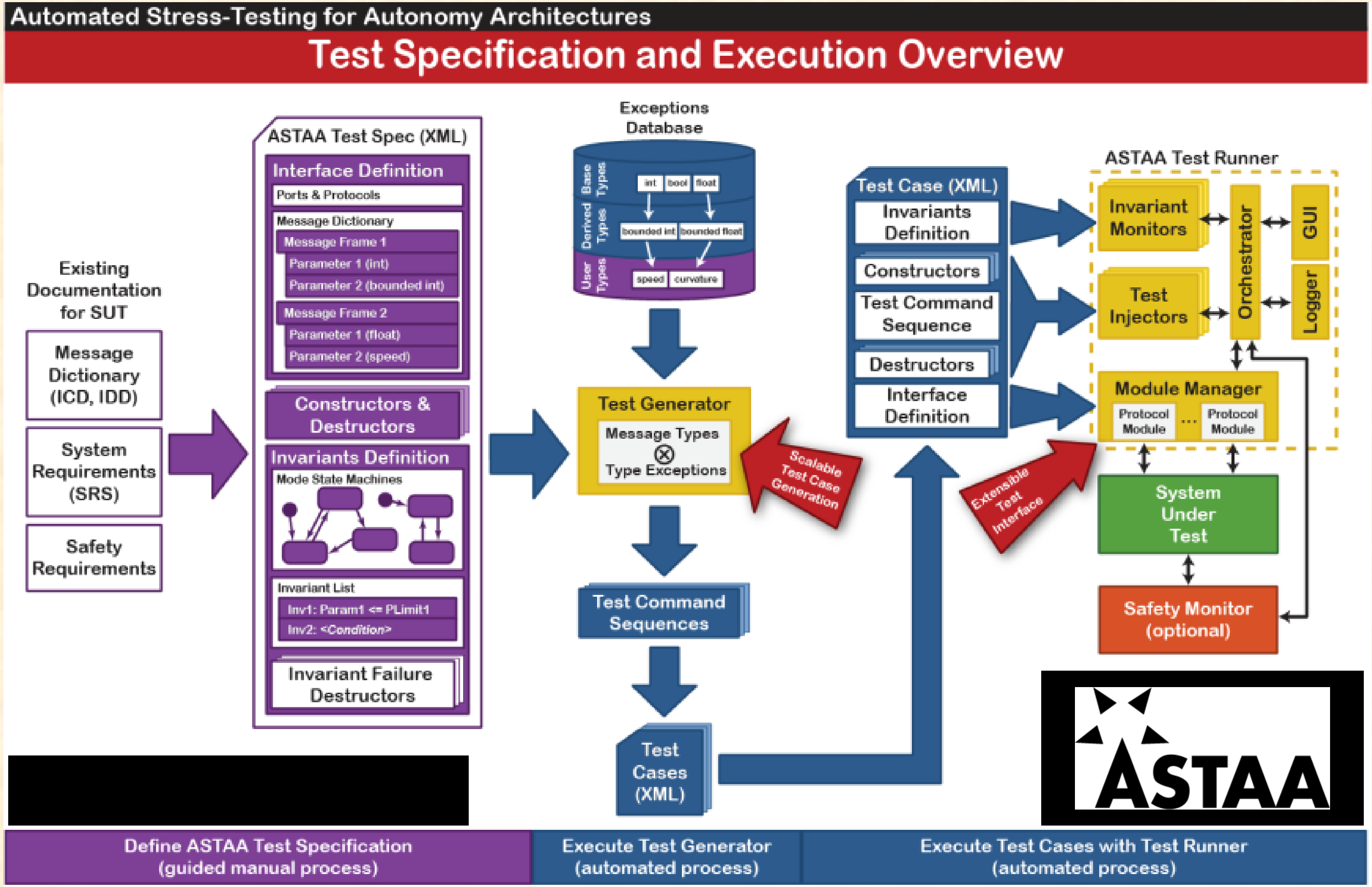
Carnegie Mellon University

- **ASTAA: Automated Stress Testing of Autonomy Architectures**
  - Key idea: combination of exceptional & normal inputs to an interface
- **Example: Ground Vehicle network**
  - Test Injector
    - Selectively modifies CAN messages on the fly
    - Modification based on data type information
  - Invariant monitor
    - Reads messages for invariant evaluation
    - "Checker" invariant monitor detects failures
- **Commercial tool build-out:**
  - Edge Case Research Switchboard (software & hardware interface testing)



DISTRIBUTION A – NREC case number STAA-2013-10-02

Robustness Test + Monitor → ASTAA



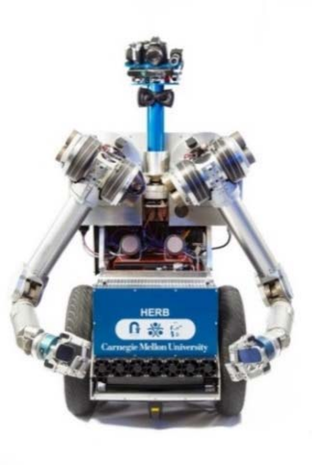
DISTRIBUTION A –
NREC case numbers
STAA-2012-10-23,
STAA-2013-10-02

Researchers evaluated 150 bugs from 11 distinct projects over 4 years [ICSE 2018]

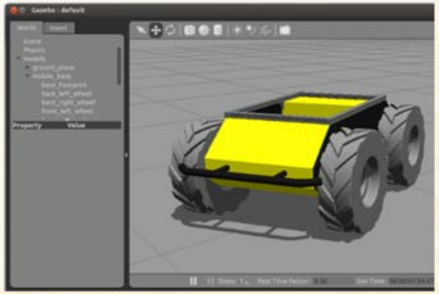
From "RIOT Expanded Technical Brief, NAVAIR Public Release- 2016-842 'Approved for Public Release; distribution is unlimited'.

Robustness Testing Finds Problems

- **Improper handling of floating-point numbers:**
  - Inf, NaN, limited precision
- **Array indexing and allocation:**
  - Images, point clouds, etc…
  - Segmentation faults due to arrays that are too small
  - Many forms of buffer overflow with complex data types
  - Large arrays and memory exhaustion
- **Time:**
  - Time flowing backwards, jumps
  - Not rejecting stale data
- **Problems handling dynamic state:**
  - For example, lists of perceived objects or command trajectories
  - Race conditions permit improper insertion or removal of items
  - Garbage collection causes crashes or hangs

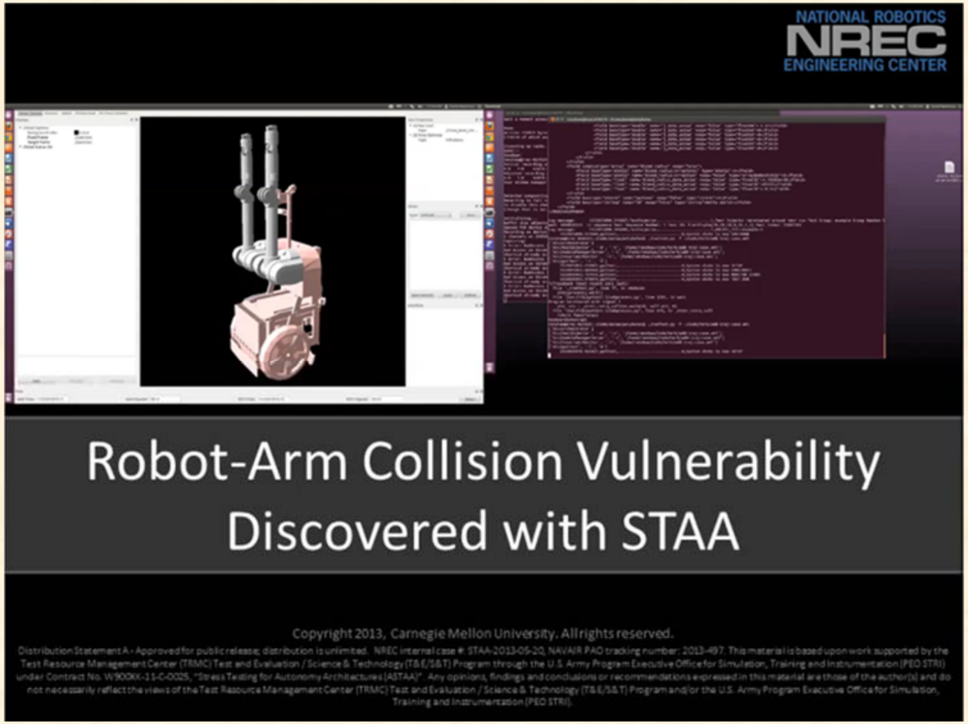


ROS



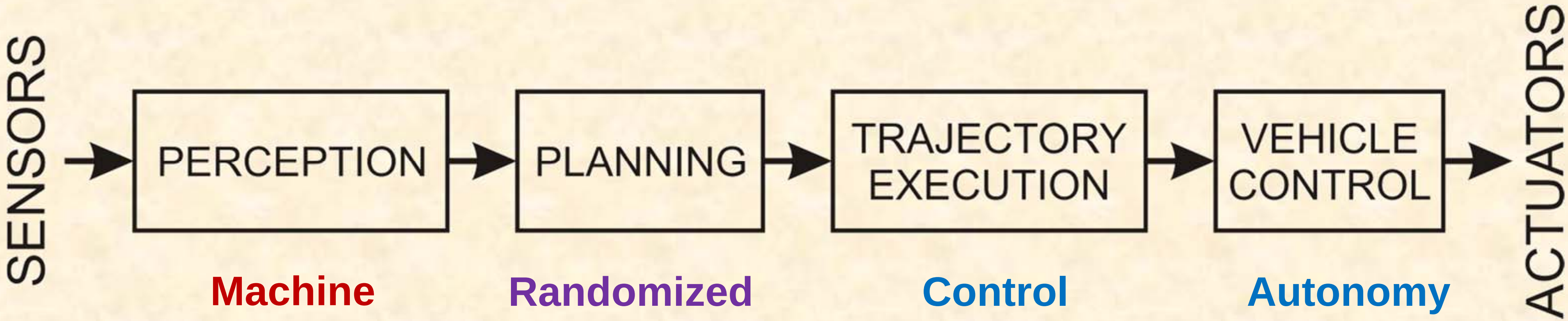
Non-Machine Learning Robustness Lessons

- **Protect your robots from data assumptions**
  - Don't trust that your configuration is valid
  - Time is not always monotonic
  - Semantically redundant field mismatches
- **Floats and NaNs useful but dangerous**
  - Do not use floats as iterators
  - NaNs propagate
- **Plan for the system to fail**
  - Nodes should not fail silent
  - Good logging is invaluable
- **Common sense?**
  - (Not so common it turns out)



Send of "infinity" floating point joint angle causes unsafe wind-milling

Validating an Autonomous Vehicle Pipeline



Machine Learning Based Approaches

➔ ???

Randomized & Heuristic Algorithms

➔ Run-Time Safety Envelopes

➔ Doer/Checker Architecture

Control Systems

➔ Control Software Validation

➔ Doer/Checker Architecture

Autonomy Interface To Vehicle

➔ Traditional Software Validation

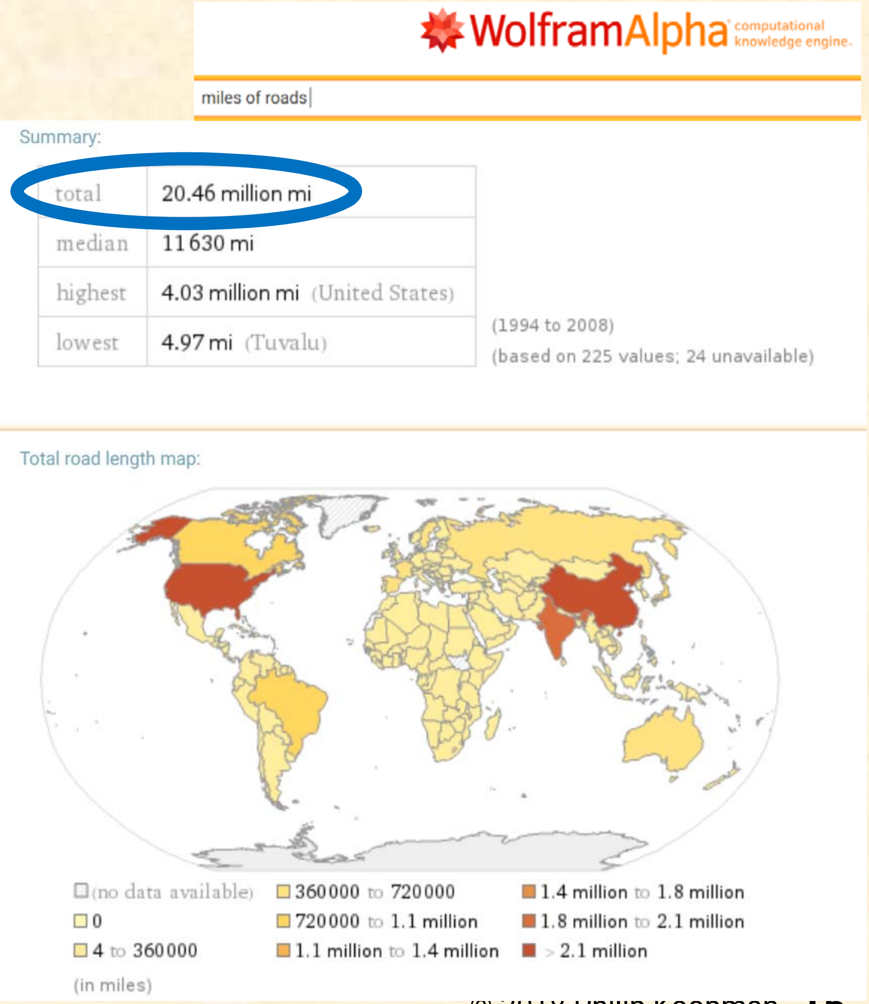
Perception presents a uniquely difficult assurance challenge

Brute Force Road Testing

- **If 100M miles/critical mishap…**
  - Test 3x–10x longer than mishap rate
    ➔ Need 1 Billion miles of testing

- **That's ~25 round trips on every road in the world**
  - With fewer than 10 critical mishaps

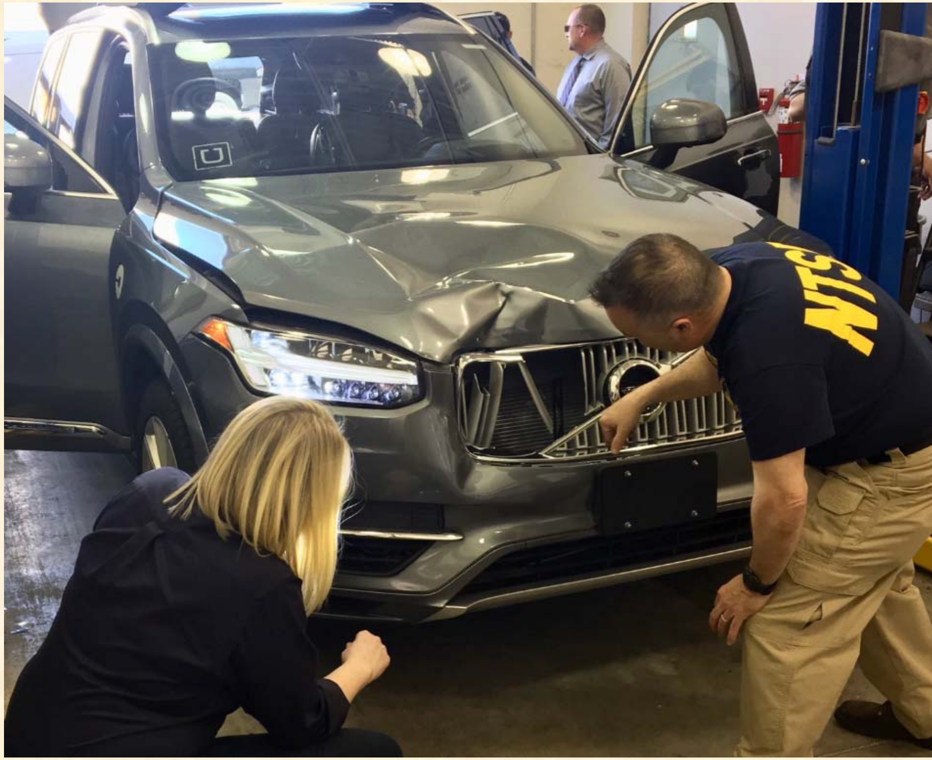
…



Brute Force AV Validation: Public Road Testing

Carnegie Mellon University

- **Good for identifying "easy" cases**
  - Expensive and potentially ***dangerous***



Did We Learn The Right Lesson from Tempe?

Carnegie Mellon University

- **NOT: Blame the victim**
  - Pedestrian in road is **expected**
- **NOT: Blame the technology**
  - Immature technology under test
    – **Failures are expected!**
- **NOT: Blame the driver**
  - A solo driver drop-out is **expected**

- **The real AV testing lesson:**
  ➔ **Ensure safety driver is engaged** ➔
  - Safety argument: Driver alert; time to respond; disengagement works

Can Safety Driver React In Time?

Carnegie Mellon University

- **Safety Driver Tasks:**
  - Mental model of "normal" AV
  - Detect abnormal AV behavior
  - React & recover if needed

Jan 20, 2016; Handan, China

- **Example: obstructed lane**
  - Does driver know when to take over?
  - Can driver brake in time?
    - – Or is sudden lane change necessary?

- **Example: two-way traffic**
  - What if AV commands sudden left turn into traffic?

Closed Course Testing

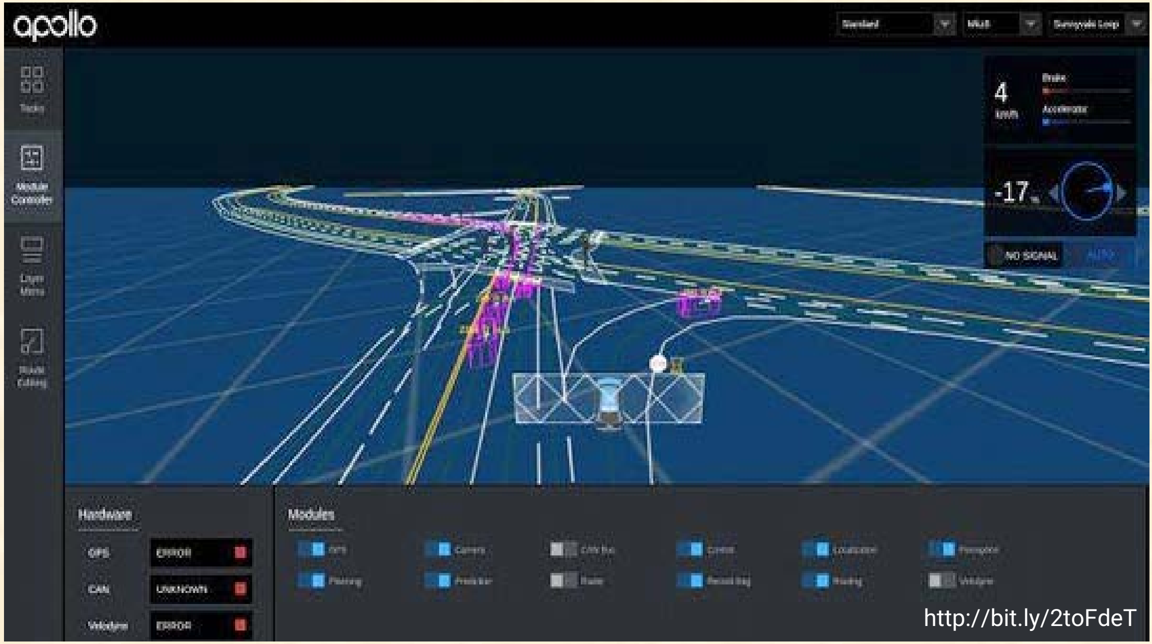
- **Safer, but expensive**
  - Not scalable
  - Only tests things you have thought of!

Volvo / Motor Trend

Simulation

- **Highly scalable; less expensive**
  - Scalable; need to manage fidelity vs. cost
  - Only tests things you have thought of!

Udacity



Apollo

What About Edge Cases?

- **You should expect the extreme, weird, unusual**
  - Unusual road obstacles
  - Extreme weather
  - Strange behaviors

PREDICTED CONCEPT	PROBABILITY
bird	0.997
no person	0.990
one	0.975
feather	0.970
nature	0.963
poultry	0.954
outdoors	0.936
color	0.910
animal	0.908

https://www.clarifai.com/demo

- **Edge Case are surprises**
  - You won't see these in testing

➔ Edge cases are the stuff you didn't think of!

Just A Few Edge Cases

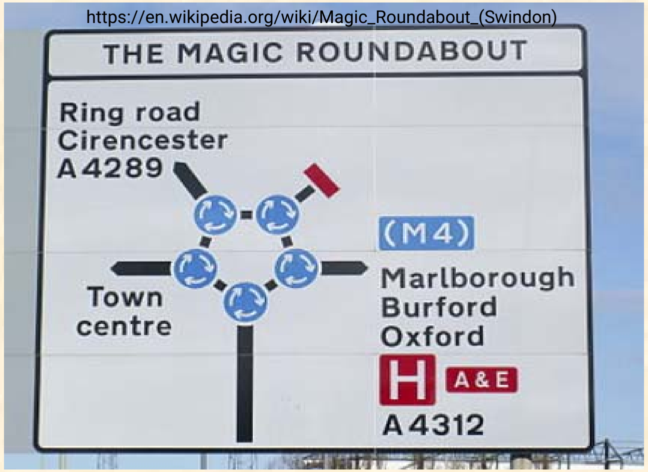
- Unusual road obstacles & obstacles
- Extreme weather
- Strange behaviors

https://goo.gl/J3SSyu

http://bit.ly/2top1KD

https://dailym.ai/2K7kNS8

http://bit.ly/2tvCCPK



https://en.wikipedia.org/wiki/Magic_Roundabout_(Swindon)

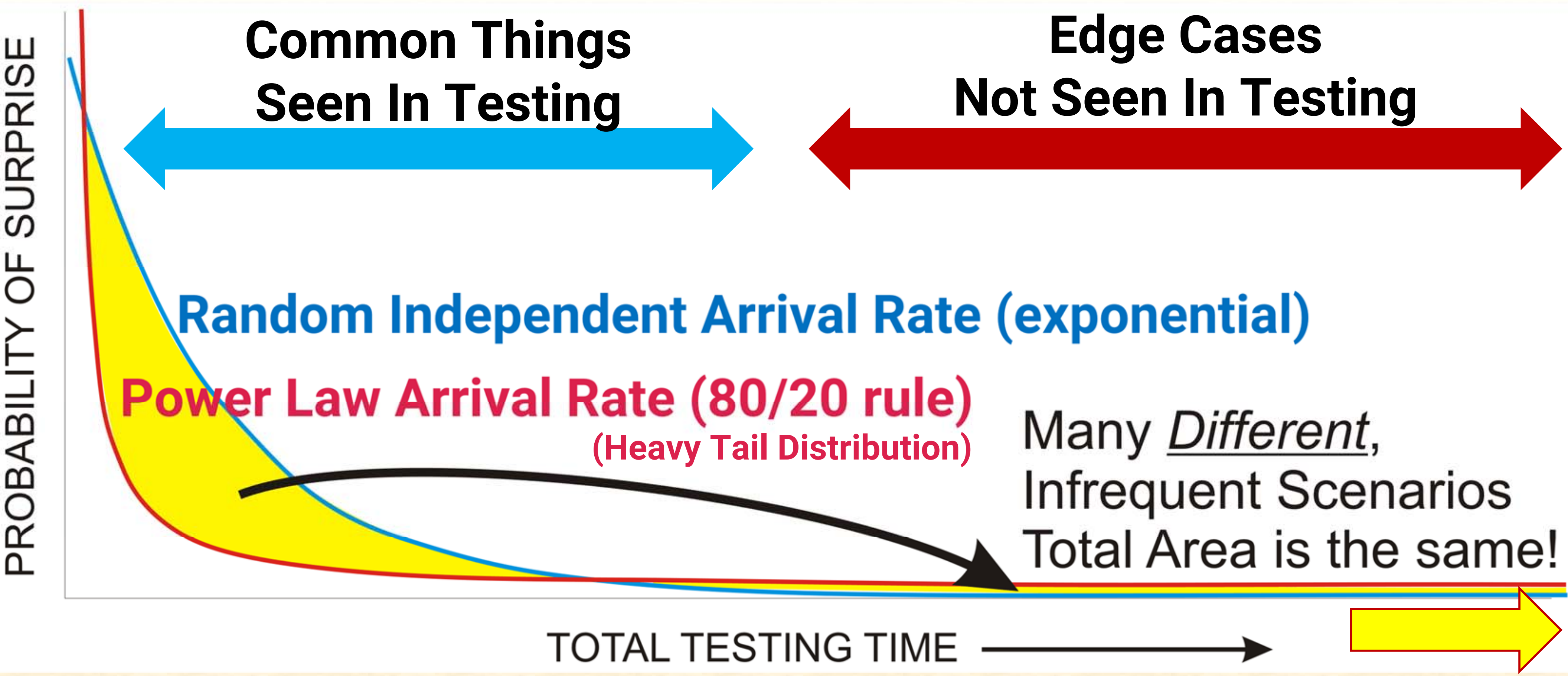
Why Edge Cases Matter

- **Where will you be after 1 Billion miles of validation testing?**

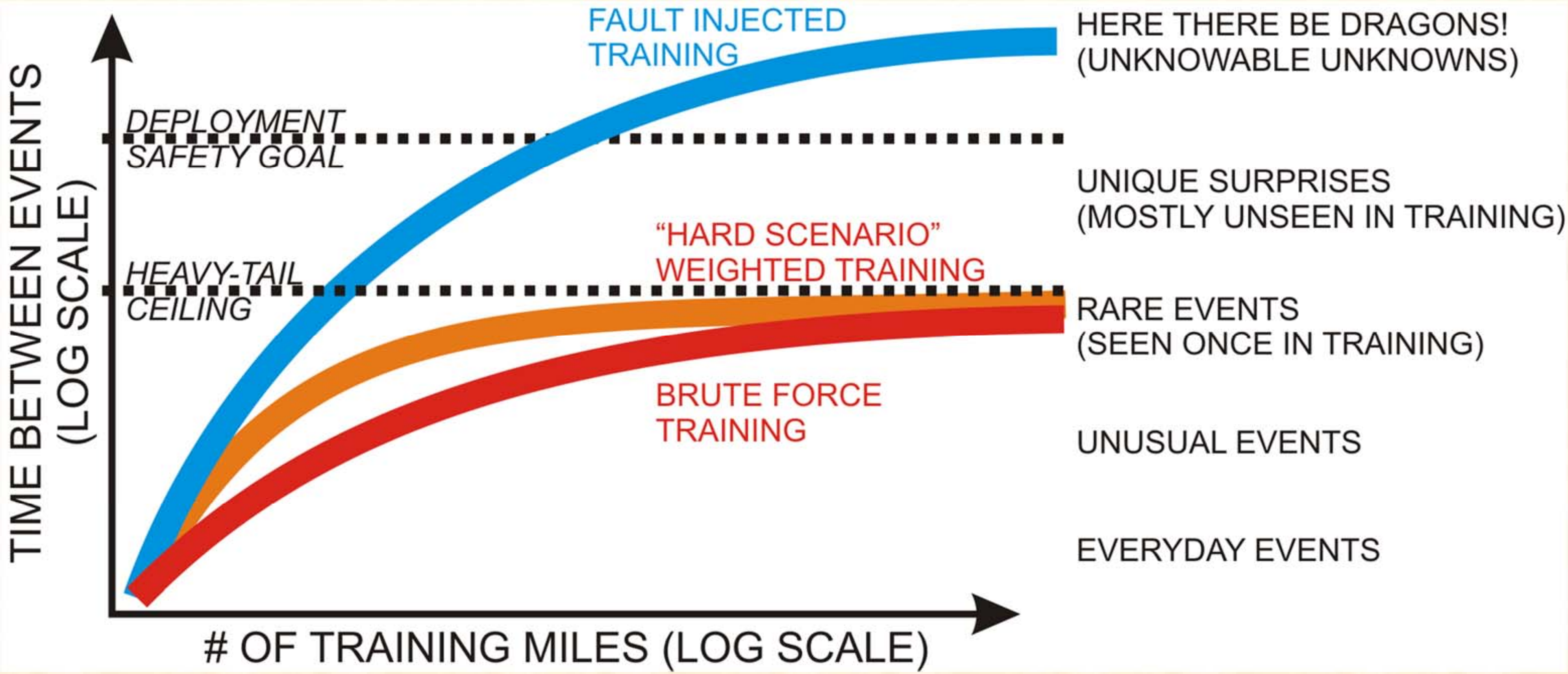
- **Assume 1 Million miles between unsafe "surprises"**
  - Example #1:
    100 "surprises" @ 100M miles / surprise
    - All surprises seen about 10 times during testing
    - With luck, all bugs are fixed
  - Example #2:
    100,000 "surprises" @ 100B miles / surprise
    - Only 1% of surprises seen during 1B mile testing
    - Bug fixes give no real improvement (1.01M miles / surprise)

https://goo.gl/3dzguf

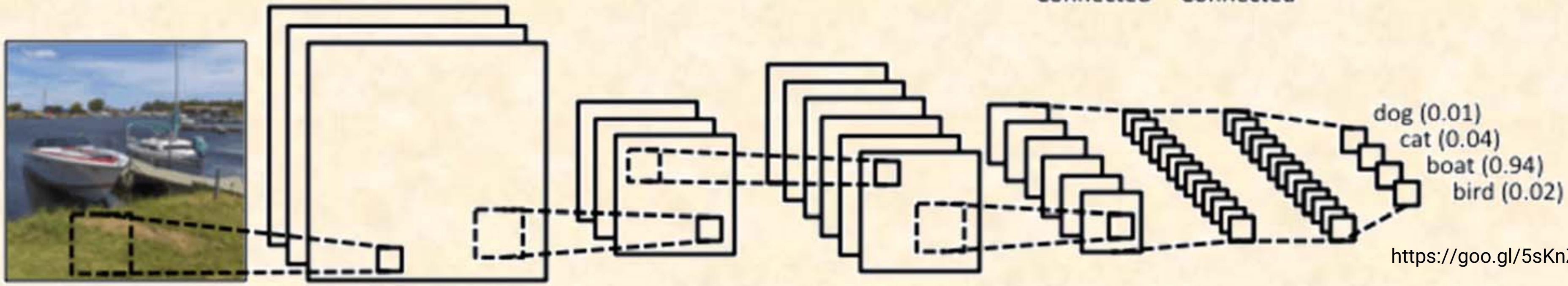
The Real World: Heavy Tail Distribution(?)



The Heavy Tail Testing Ceiling



Malicious Image Attacks Reveal Brittleness

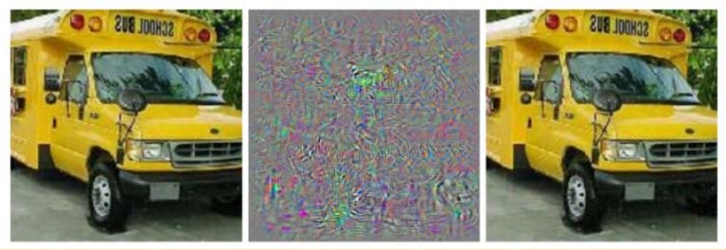


https://goo.gl/5sKnZV

QuocNet:

Car | **Not a Car** | ***Magnified Difference***

AlexNet:

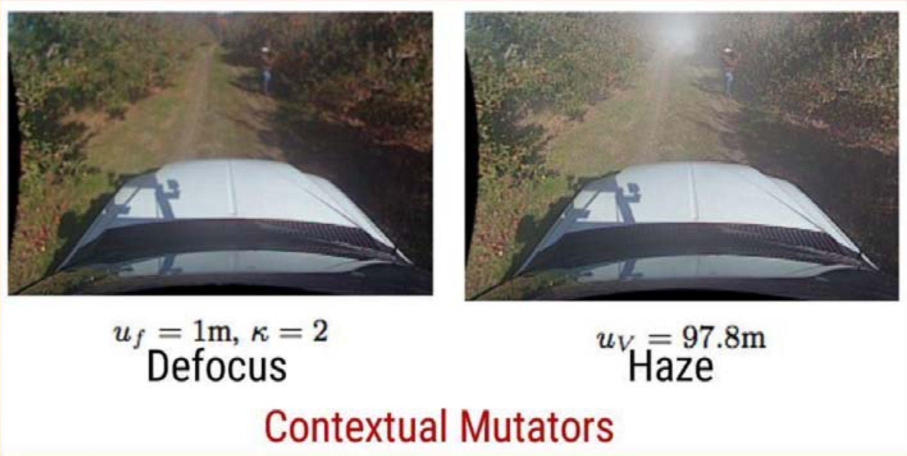


Bus | ***Magnified Difference*** | **Not a Bus**

Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

ML Is Brittle To Environment Changes

- **Sensor data corruption experiments**



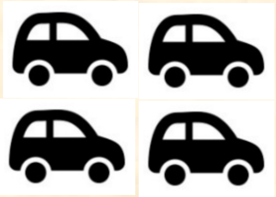
Defocus & haze are similarly a significant issue

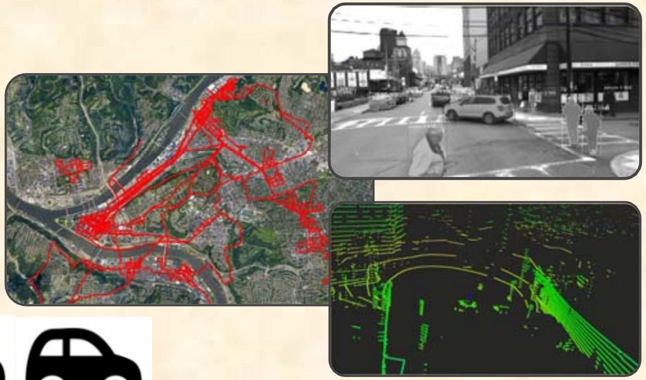
Exploring the response of a DNN to environmental perturbations from "Robustness Testing for Perception Systems," RIOT Project, NREC, DIST-A.

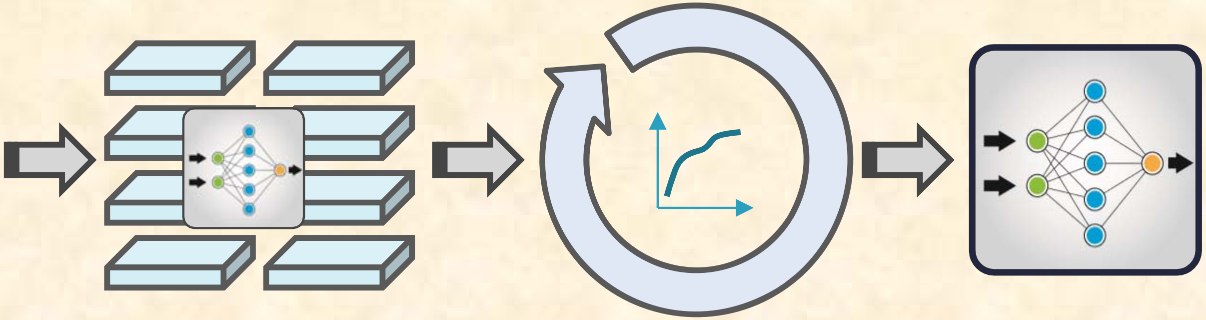
What We're Learning With Hologram

Edge Case Research

- A scalable way to test & train on Edge Cases







Your fleet and your data lake

Hologram cluster tests your CNN

Hologram cluster trains your CNN

Your CNN becomes more robust

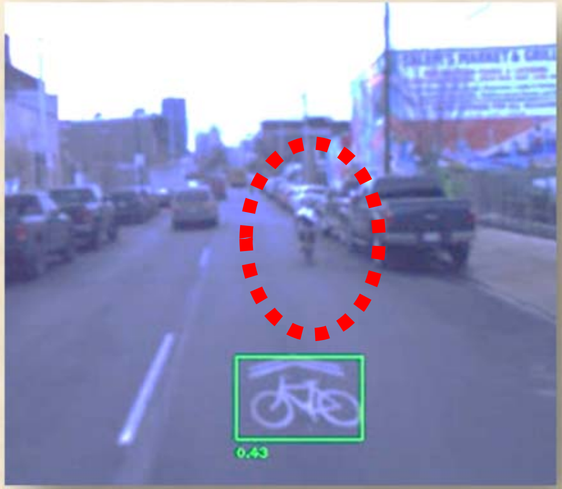
hologram

SAFER PERCEPTION FOR AUTONOMY

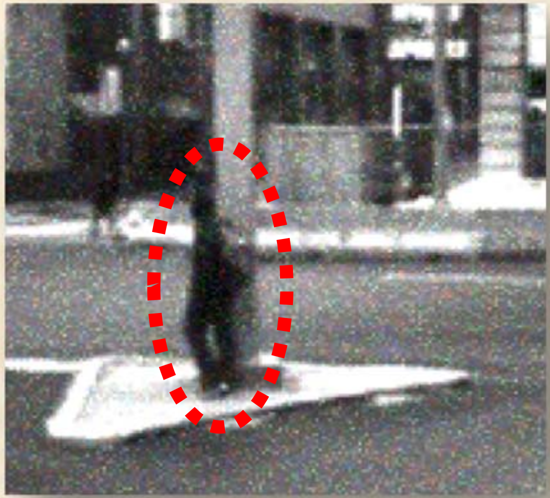
Context-Dependent Perception Failures

Edge Case Research

- **Perception failures are often context-dependent**
  - False positives and false negatives are both a problem
  - ***This is an active research area … technology still in development***



False positive on lane marking
False negative real bicyclist



False negative when
person next to light pole

False negative when
in front of dark vehicle

Will this pass a "vision test" for bicyclists?

Ways To Improve AV Safety

Carnegie Mellon University

- **More safety transparency**
  - Independent safety assessments
  - Industry collaboration on safety
- **Minimum performance standards**
  - Share data on scenarios and obstacles
  - Safety for on-road testing (driver & vehicle)
- **Autonomy software safety standards**
  - Traditional software safety … ***PLUS*** …
  - Dealing with uncertainty and brittleness
  - Data collection and feedback on field failures

http://bit.ly/2MTbT8F (sign modified)