Prof. Philip Koopman

Edge Cases and Autonomous Vehicle Safety

IFIP WG 10.4
25 January 2019

Carnegie Mellon University

@PhilKoopman

Edge Case Research

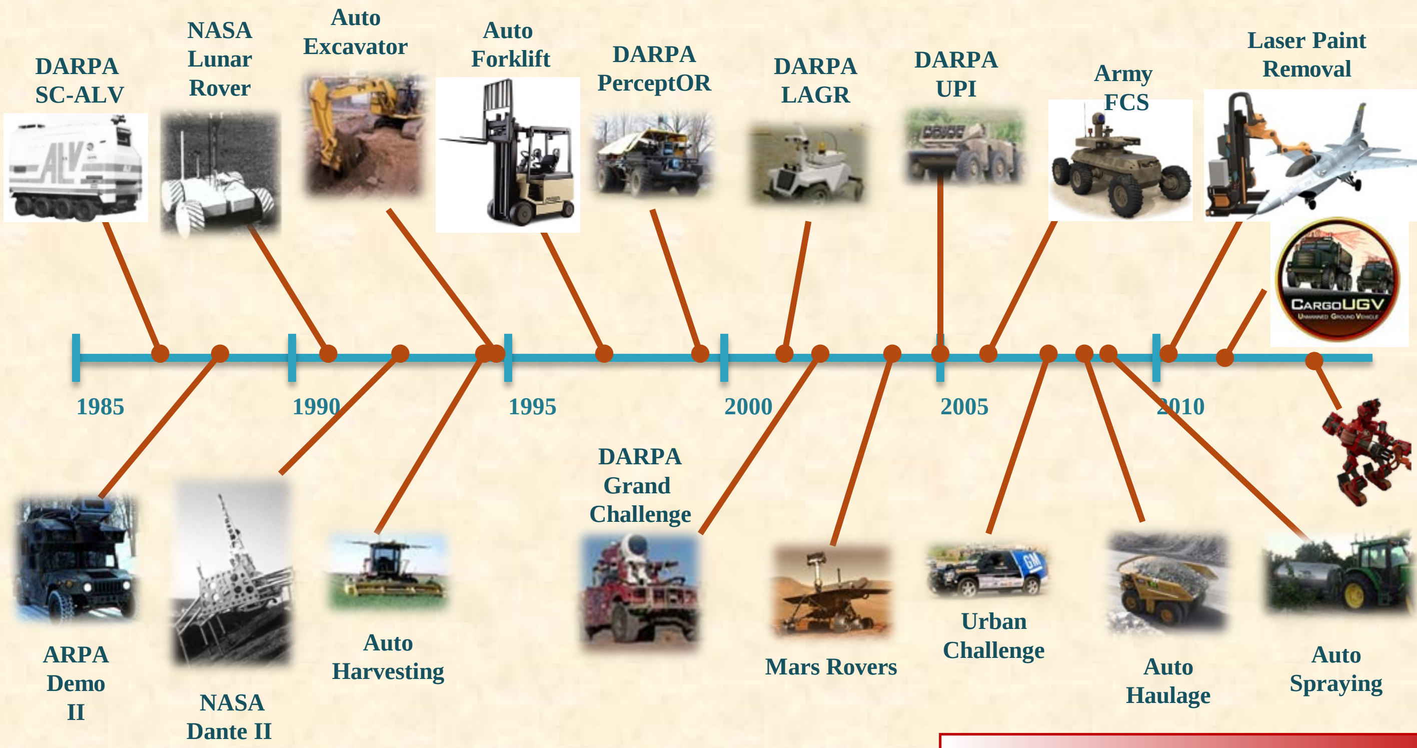
Overview

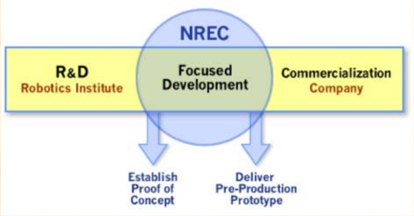
- **Edge cases matter**
  - Robust perception matters
- **The heavy tail distribution**
  - Fixing stuff you see in testing isn't enough
- **Perception stress testing**
  - Finding the weaknesses in perception

[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

98% Solved For 20+ Years

NO HANDS ACROSS AMERICA
Carnegie Mellon University • Delco Electronics • AssistWare Technology

July 1995

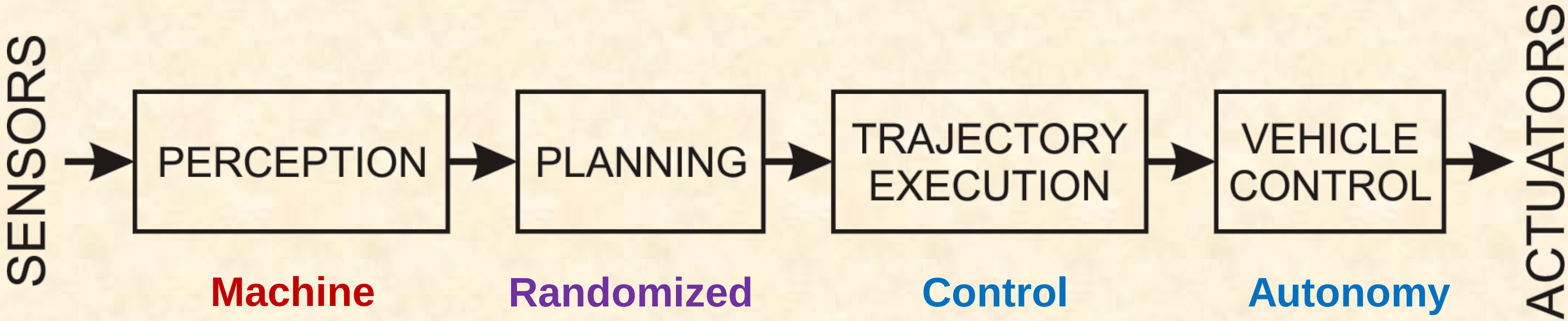
TRIP COMPLETE !!!
2797/2849 miles (98.2%)

- **Washington DC to San Diego**
  - CMU Navlab 5
  - Dean Pomerleau
  - Todd Jochem

  https://www.cs.cmu.edu/~tjochem/nhaa/nhaa_home_page.html

- **AHS San Diego demo Aug 1997**

Validating an Autonomous Vehicle Pipeline



SENSORS → PERCEPTION → PLANNING → TRAJECTORY EXECUTION → VEHICLE CONTROL → ACTUATORS

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

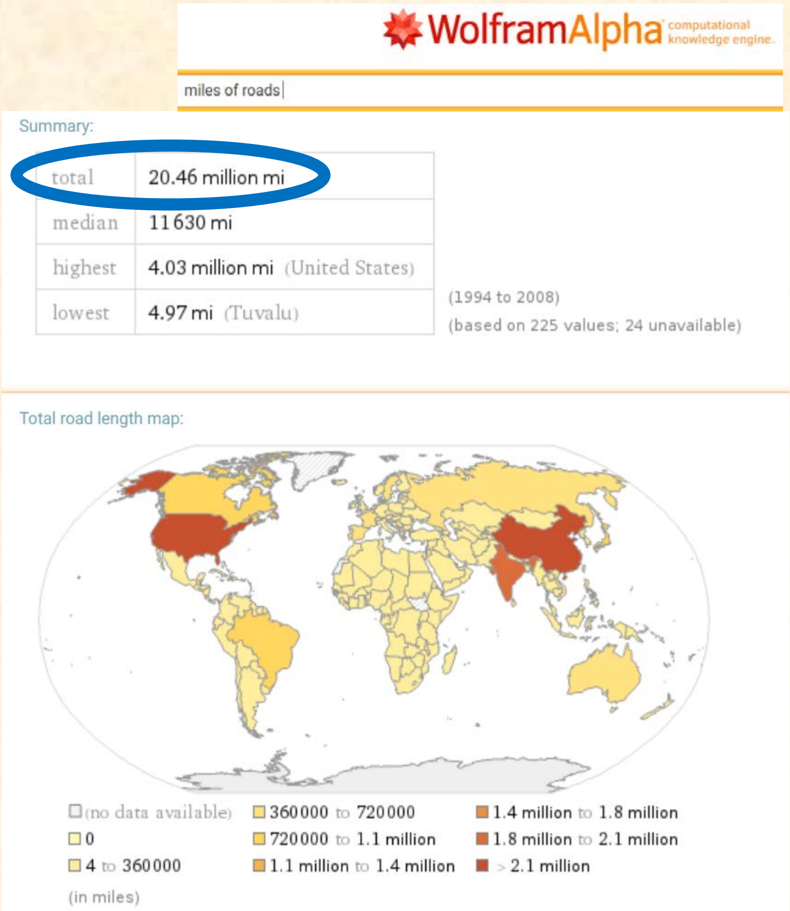
Perception presents a uniquely difficult assurance challenge

Validation Via 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

…



WolframAlpha computational knowledge engine

miles of roads

Summary:

total	20.46 million mi	
median	11 630 mi	
highest	4.03 million mi (United States)	(1994 to 2008)
lowest	4.97 mi (Tuvalu)	(based on 225 values; 24 unavailable)

Total road length map:

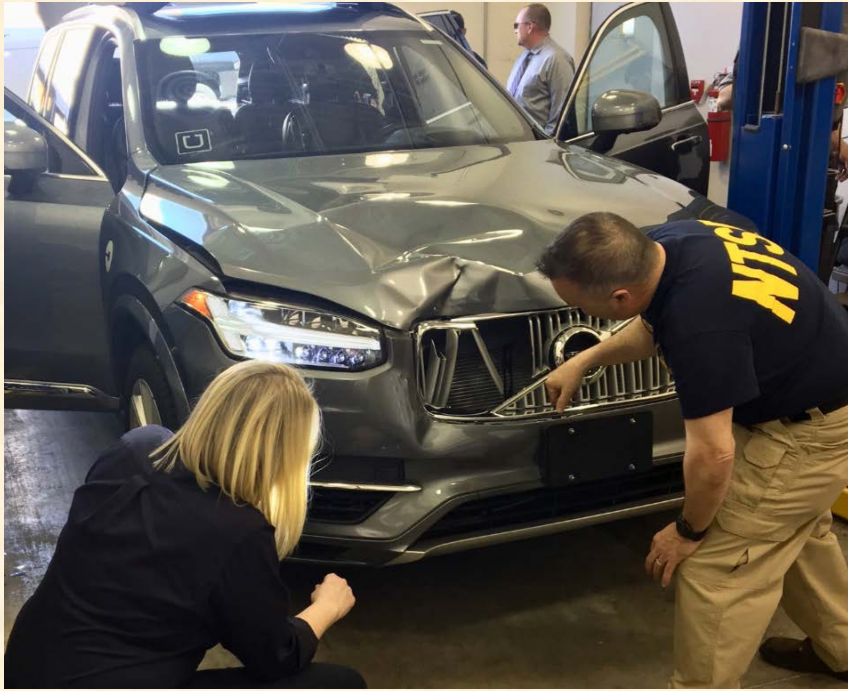
□ (no data available)	□ 360000 to 720000	■ 1.4 million to 1.8 million
□ 0	□ 720000 to 1.1 million	■ 1.8 million to 2.1 million
□ 4 to 360000	■ 1.1 million to 1.4 million	■ > 2.1 million

(in miles)

Brute Force AV Validation: Public Road Testing

- **Good for identifying “easy” cases**
  - Expensive and potentially ***dangerous***

http://bit.ly/2toadfa



Closed Course Testing

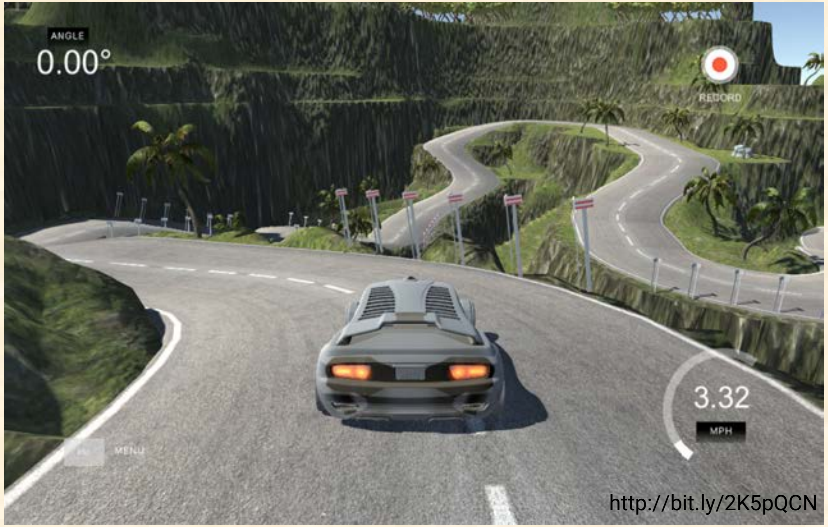
- **Safer, but expensive**
  - Not scalable
  - Only tests things you have thought of!

Volvo / Motor Trend

Simulation

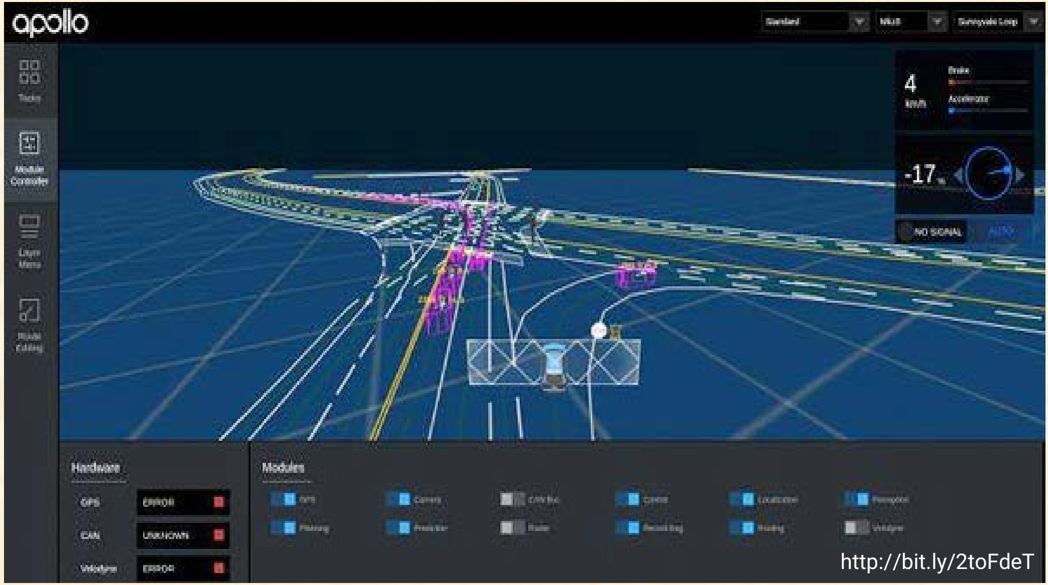
Carnegie Mellon University

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



http://bit.ly/2K5pQCN

Udacity



Hardware

GPS ERROR

CAN UNKNOWN

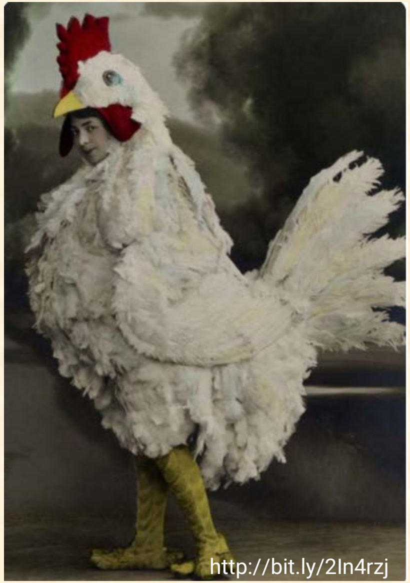
Velodyne ERROR

http://bit.ly/2toFdeT

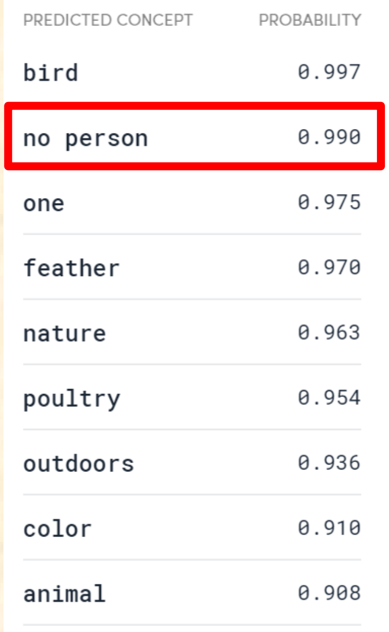
Apollo

What About Edge Cases?

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



http://bit.ly/2In4rzj



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

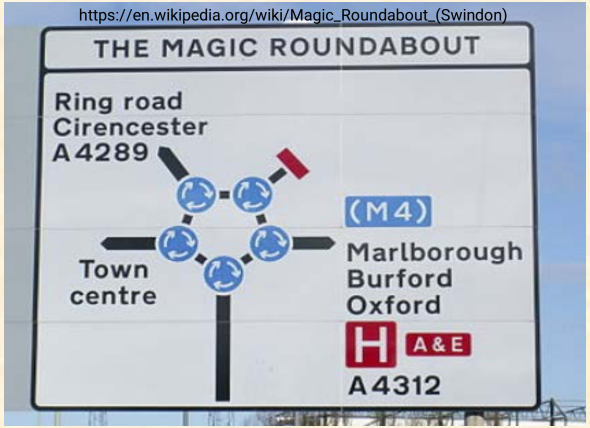
- Unusual road obstacles & obstacles
- Extreme weather
- Strange behaviors

http://bit.ly/2top1KD

https://goo.gl/J3SSyu

http://bit.ly/2tvCCPK

https://dailym.ai/2K7kNS8


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

Why Edge Cases Matter

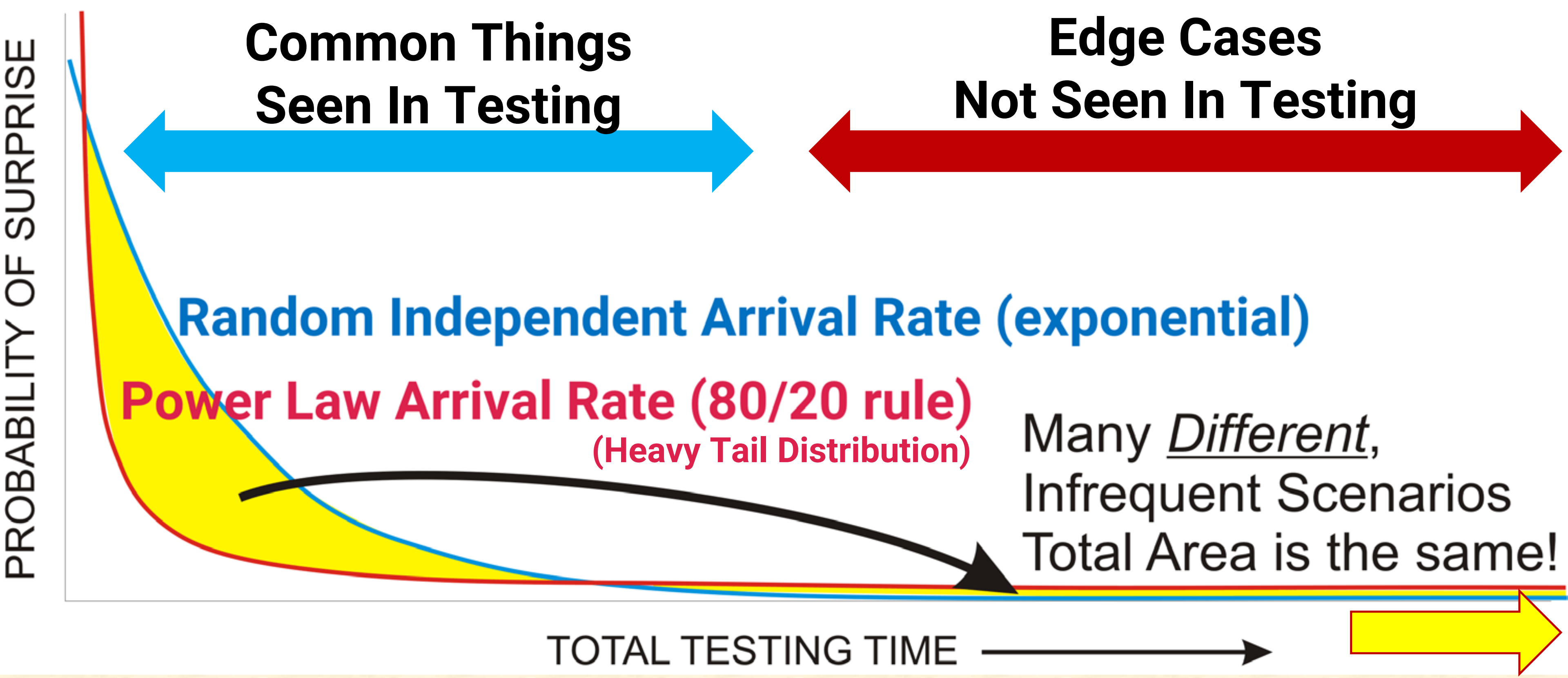
- **Where will you be after 1 Billion miles of validation testing?**

- **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)

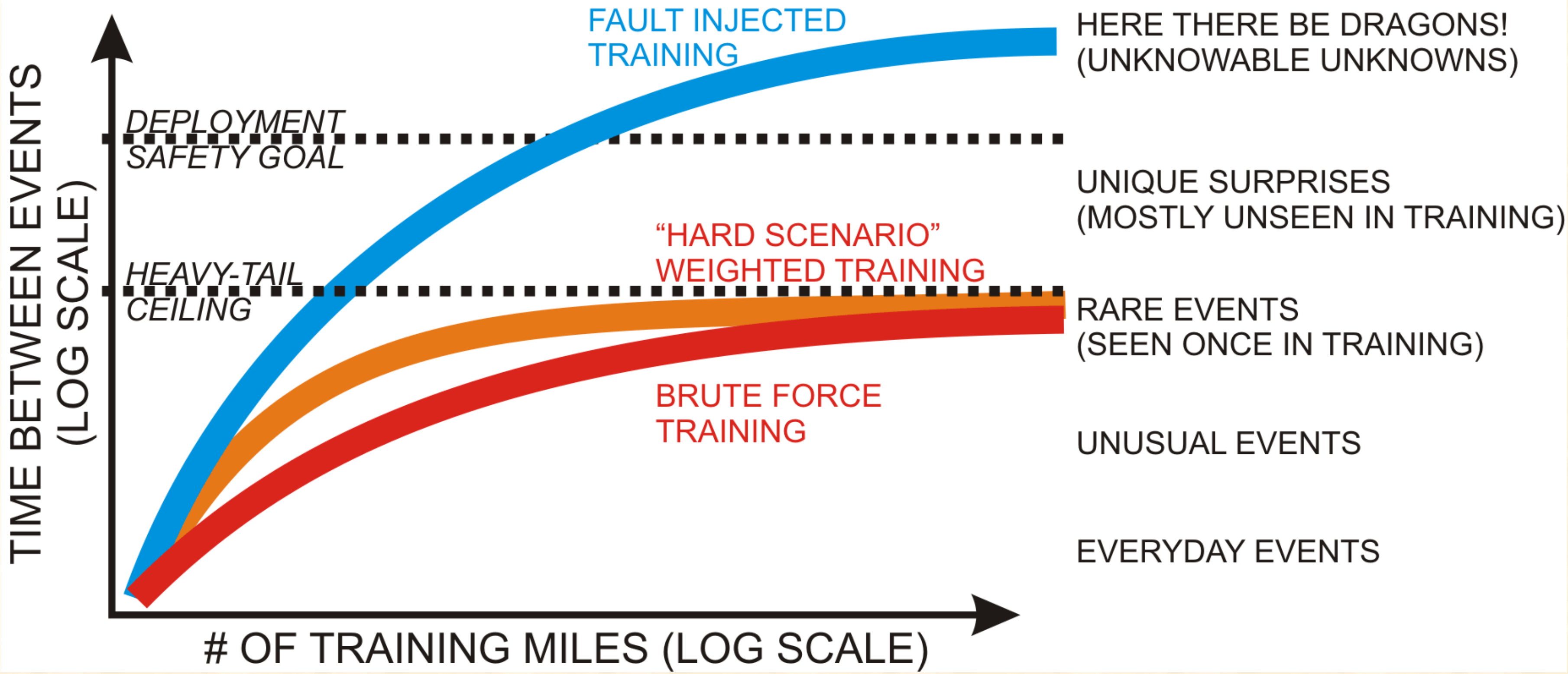
https://goo.gl/3dzguf

The Real World: Heavy Tail Distribution(?)

Carnegie Mellon University



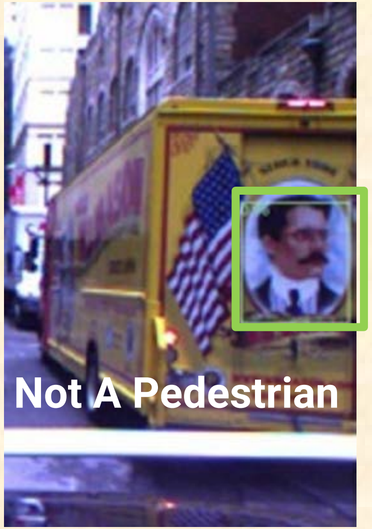
The Heavy Tail Testing Ceiling



Edge Cases Part 1: Triggering Event Zoo

- **Need to collect surprises**
  - Novel objects
  - Novel operational conditions
- **Corner Cases vs. Edge Cases**
  - Corner cases: infrequent combinations
    - Not all corner cases are edge cases
  - Edge cases: combinations that behave unexpectedly

https://goo.gl/Ni9HhU

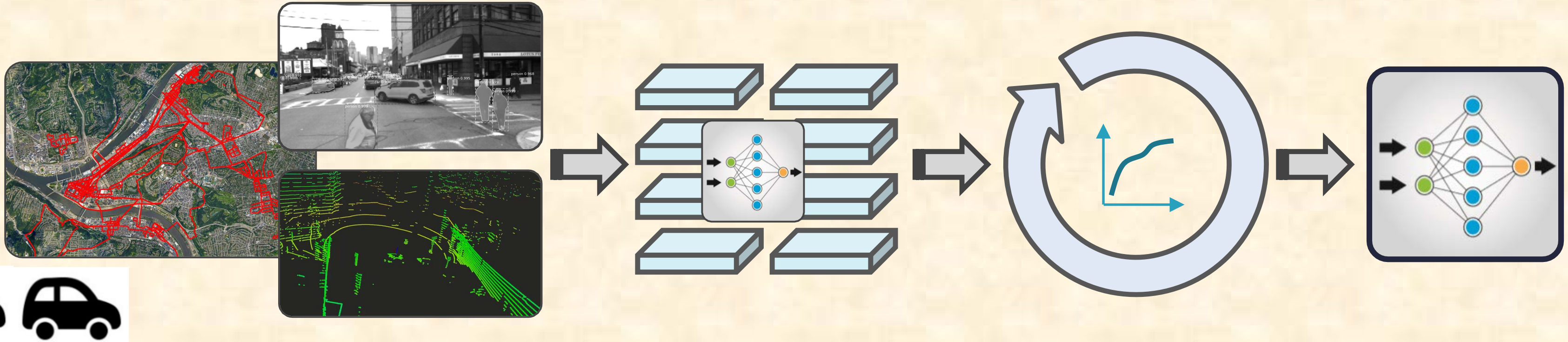


- **Issue: novel for person ≠ novel for Machine Learning**
  - ML can have “edges” in unexpected places
  - ML might train on features that seem irrelevant to people

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 identifies weaknesses & helps retrain your CNN

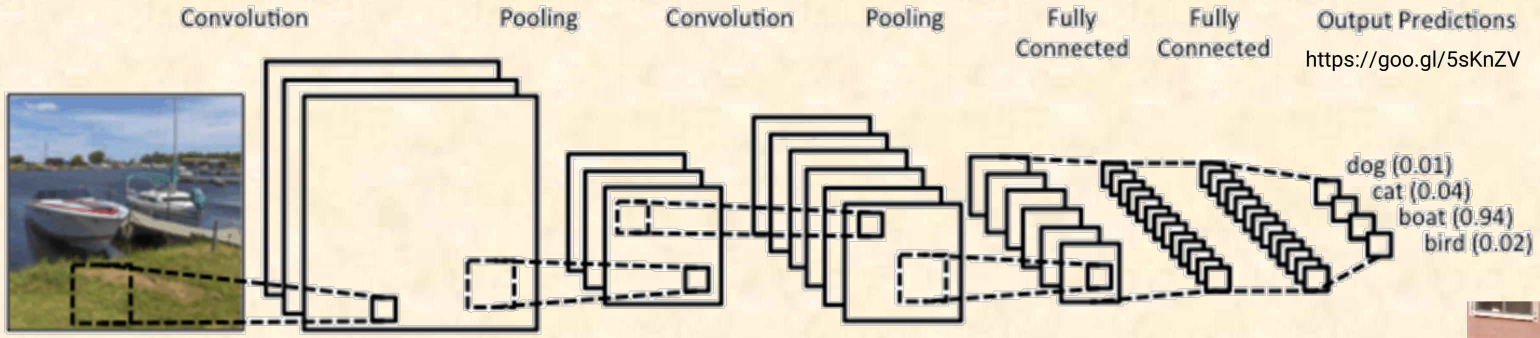
Your CNN becomes more robust

hologram

SAFER PERCEPTION FOR AUTONOMY

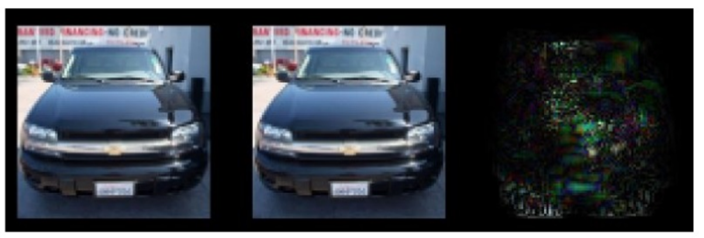
Edge Cases Part 2: Brittleness

Malicious Image Attacks Reveal Brittleness:



https://goo.gl/5sKnZV

QuocNet:

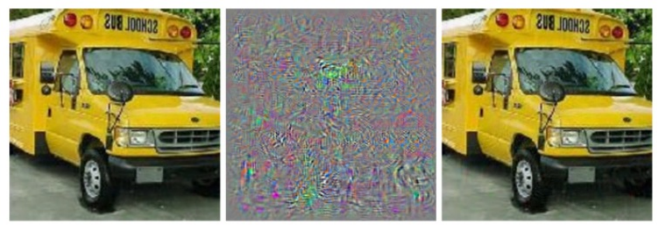


Car

Not a Car

Magnified Difference

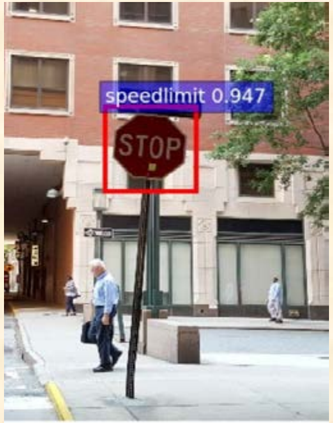
AlexNet:



Bus

Magnified Difference

Not a Bus

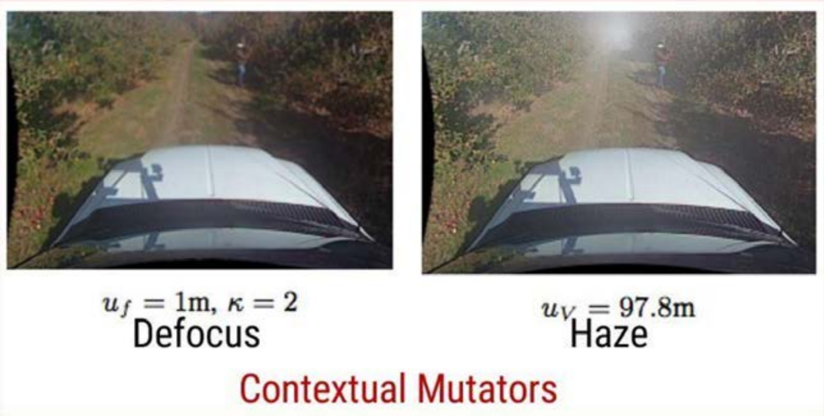


https://goo.gl/ZB5s4Q
(NYU Back Door Training)

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

ML Is Brittle To Environment Changes

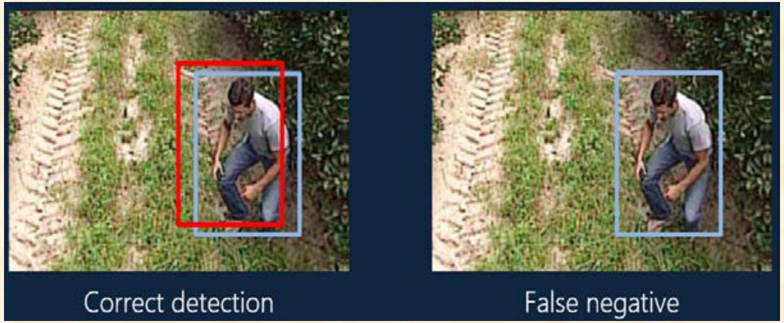
- **Sensor data corruption experiments**



Defocus & haze are a significant issue

Synthetic Equipment Faults

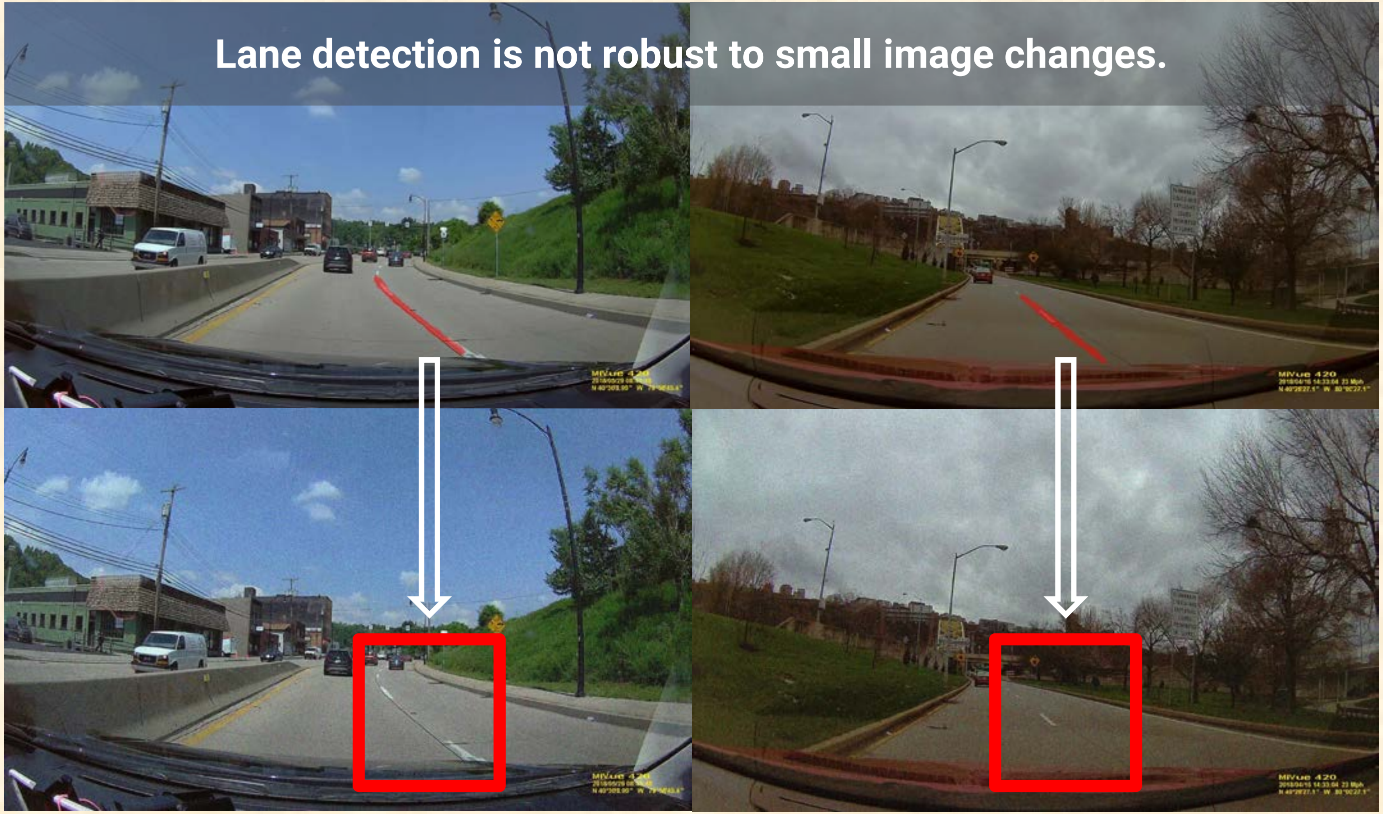
Gaussian blur



Gaussian Blur & Gaussian Noise cause similar failures

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

Noise Susceptibility



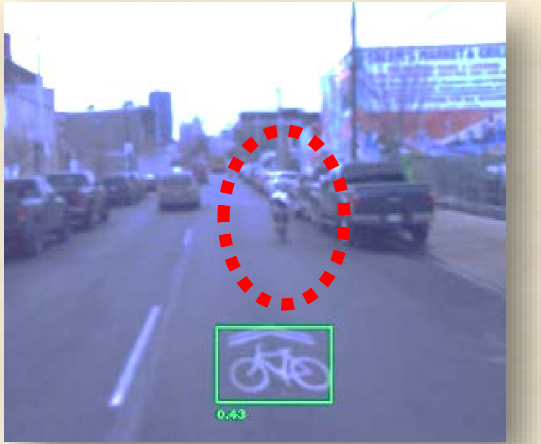
Lane detection is not robust to small image changes.

LaneNet Original

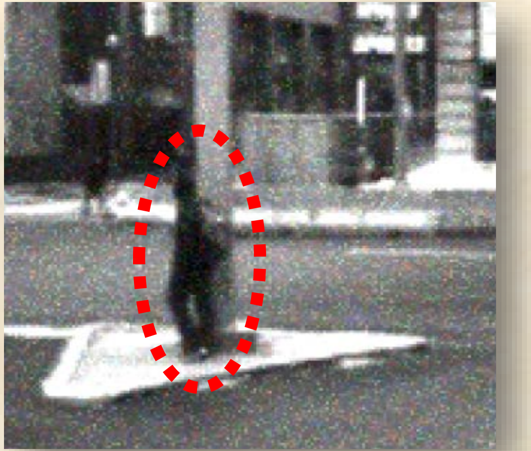
LaneNet With Gaussian Noise

Context-Dependent Perception Failures

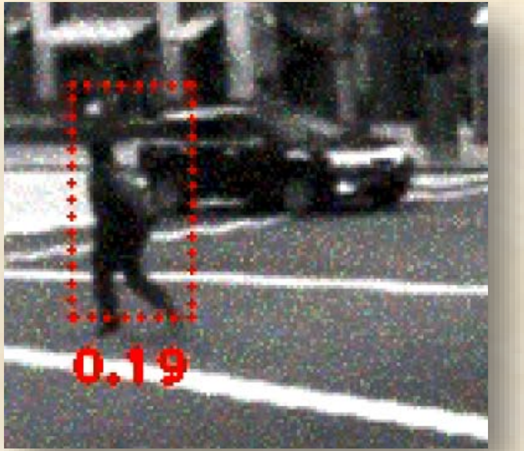
- **Perception failures are often context-dependent**
  - False positives and false negatives are both a problem



False positive on lane marking
False negative real bicyclist



False negative when
person next to light pole

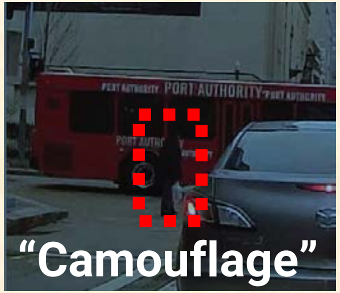


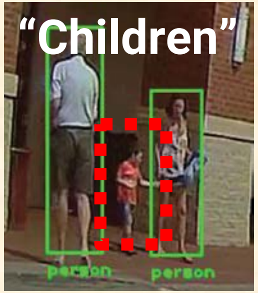
False negative when
in front of dark vehicle

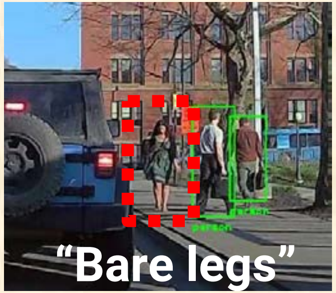
Will this pass a “vision test” for bicyclists?

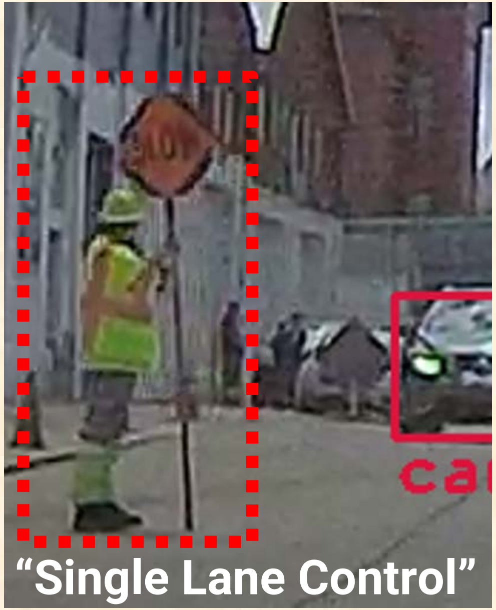
Example Triggering Events via Hologram

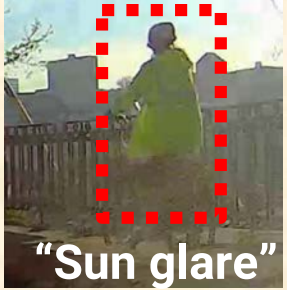
- Mask-R CNN: examples of clusters we found

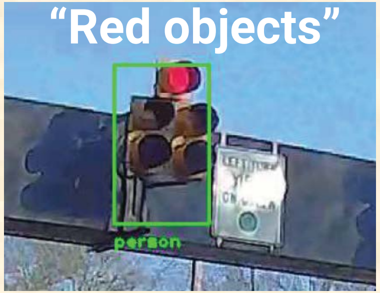


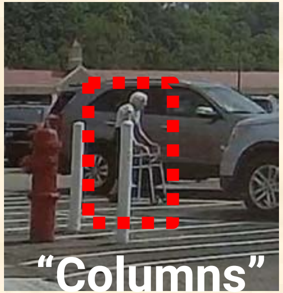












Notes: These are baseline, un-augmented images.
(Your mileage may vary on your own trained neural network.)

Ways To Improve AV Safety

- **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 surprises and brittleness**
  - Data collection and feedback on field failures

Thanks!

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

hologram

SAFER PERCEPTION FOR AUTONOMY

EDGE CASE RESEARCH

info@ecr.guru