A Differential Testing Framework to Identify Critical AV Failures Leveraging Arbitrary Inputs

Trey Woodlief[1], Carl Hildebrandt[1], Sebastian Elbaum


University of Virginia

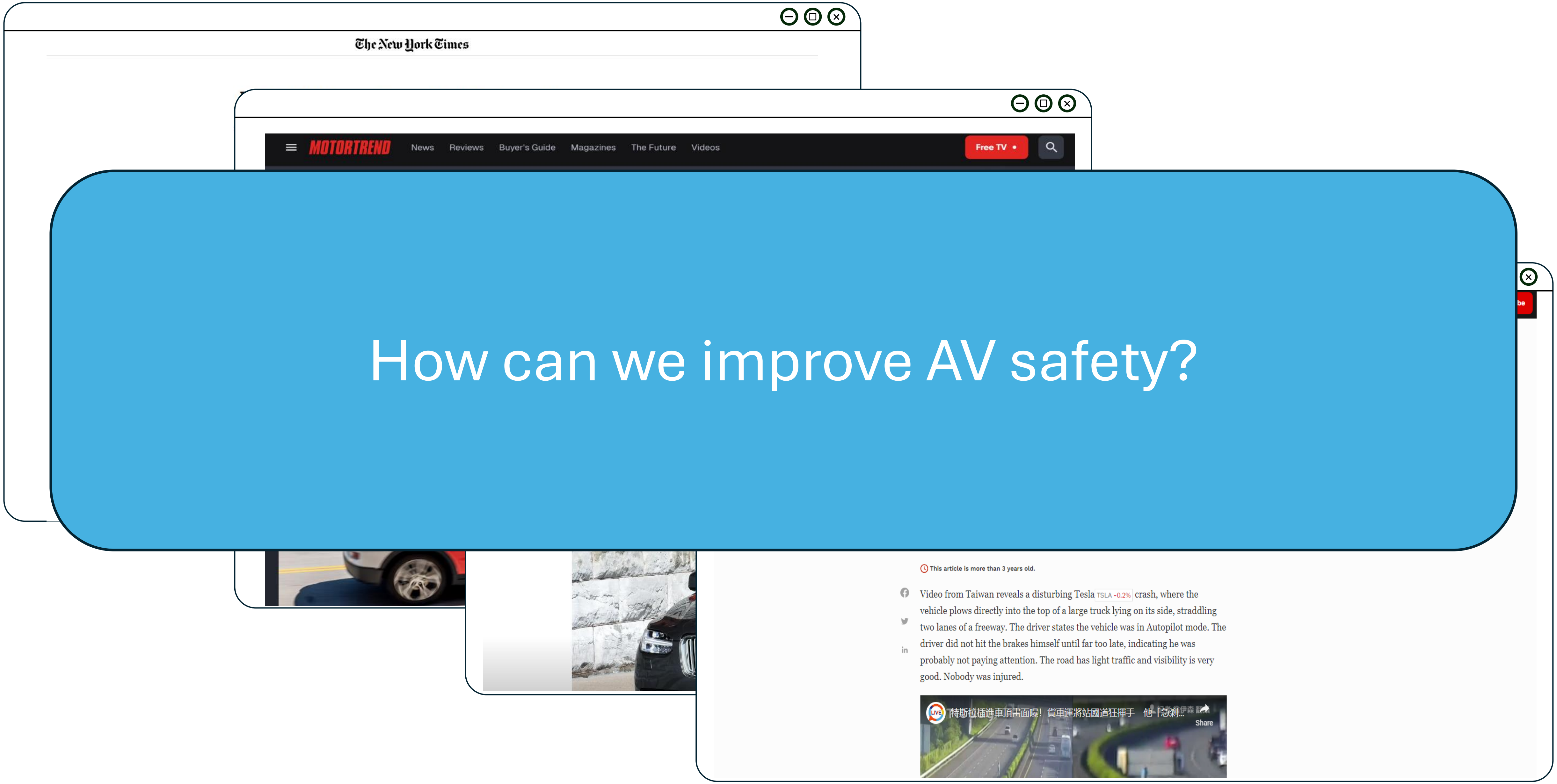
Trey Woodlief

WILLIAM & MARY

CHARTERED 1693

I am joining William & Mary Fall’25 as Assistant Prof

AV Failures



How can we improve AV safety?

How do AVs work?

Input

AV

Output

Goal: produce safe outputs for every input

How do AVs work?

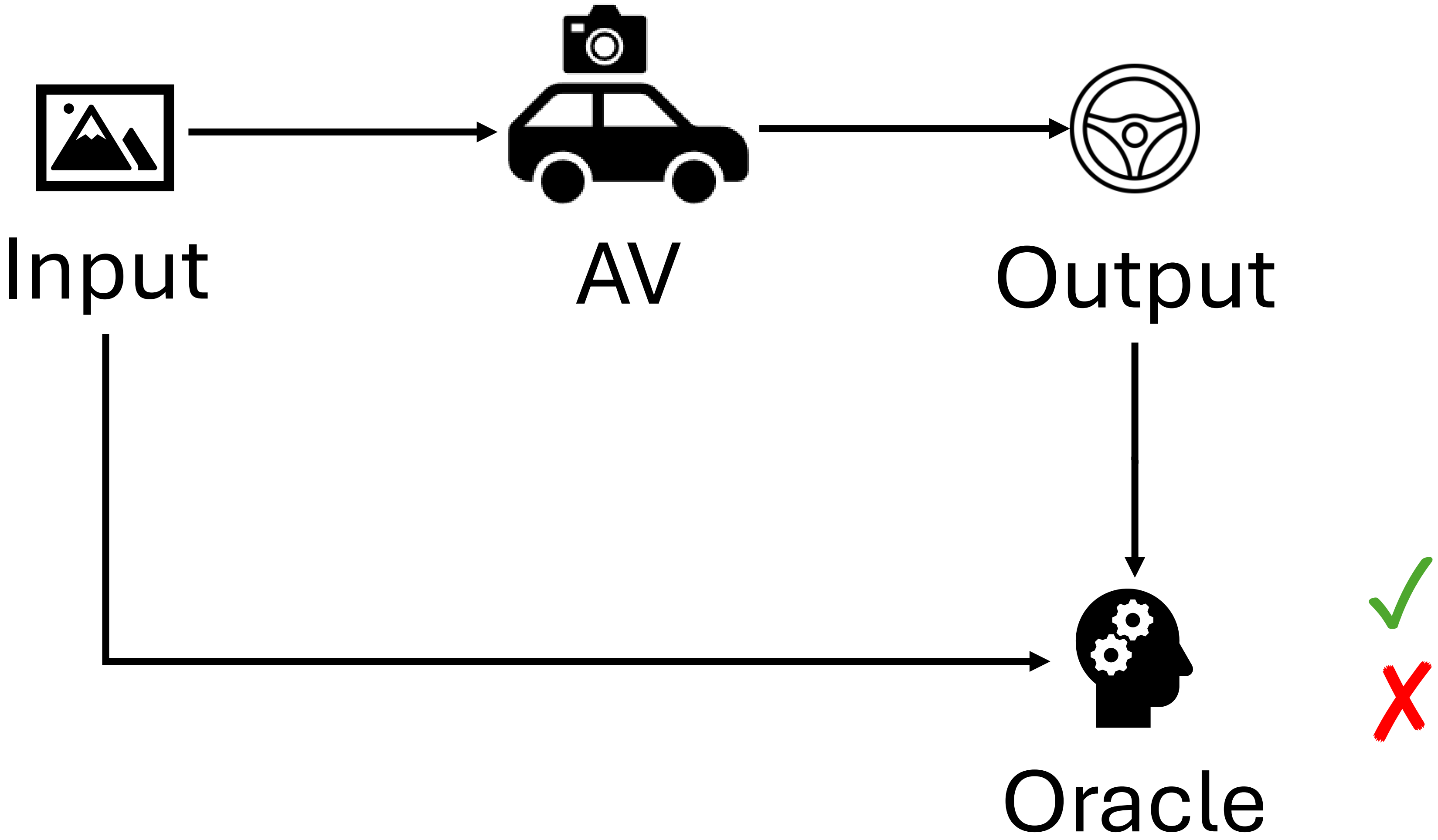
Input

AV

Output

Validate: produce safe outputs for every input

Validation



Test Inputs

IEEE Spectrum | Tesla's Autopilot Depends on a Deluge of Data

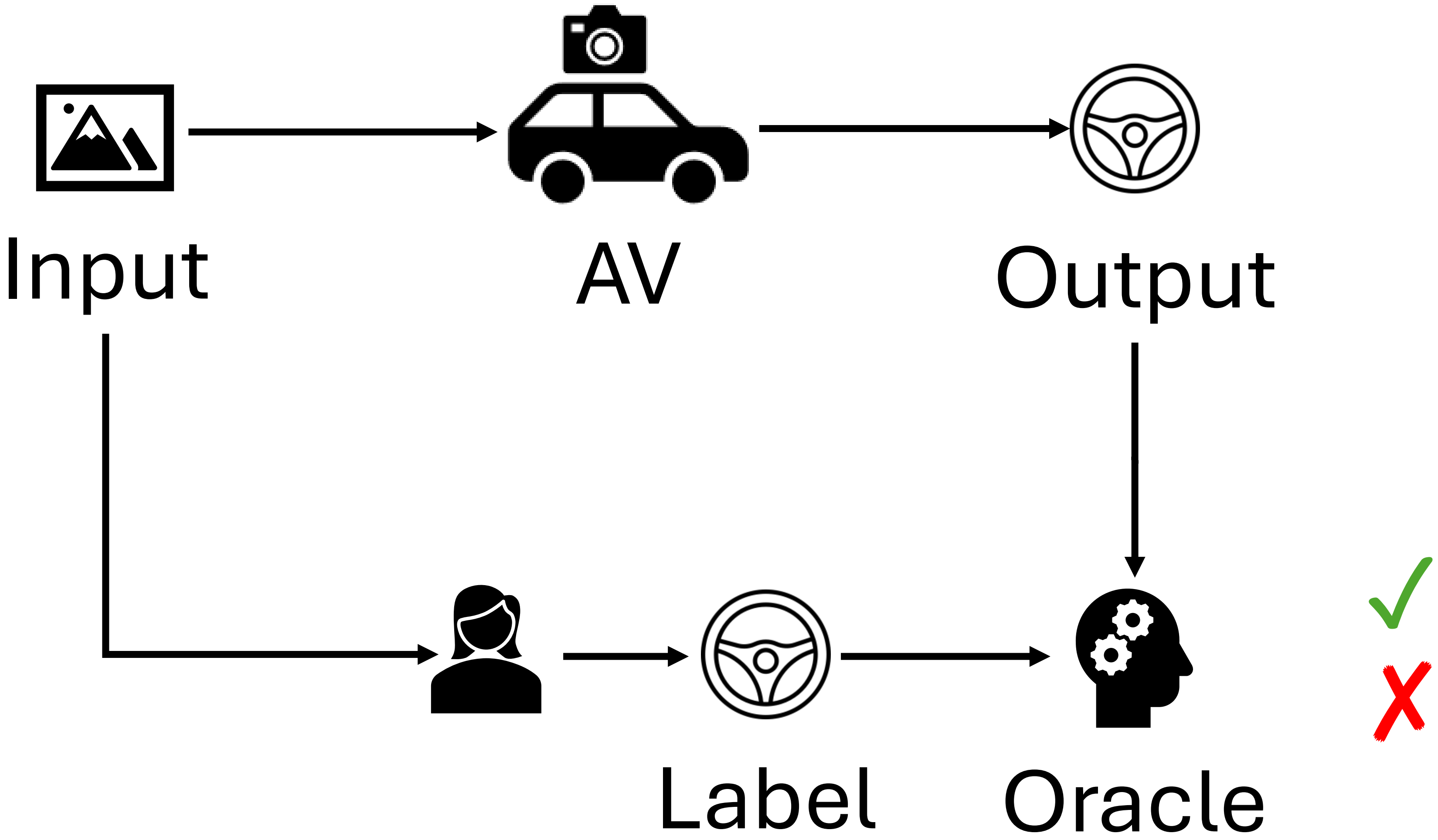
Type to search

PHILIPP MANDLER/UNSPLASH

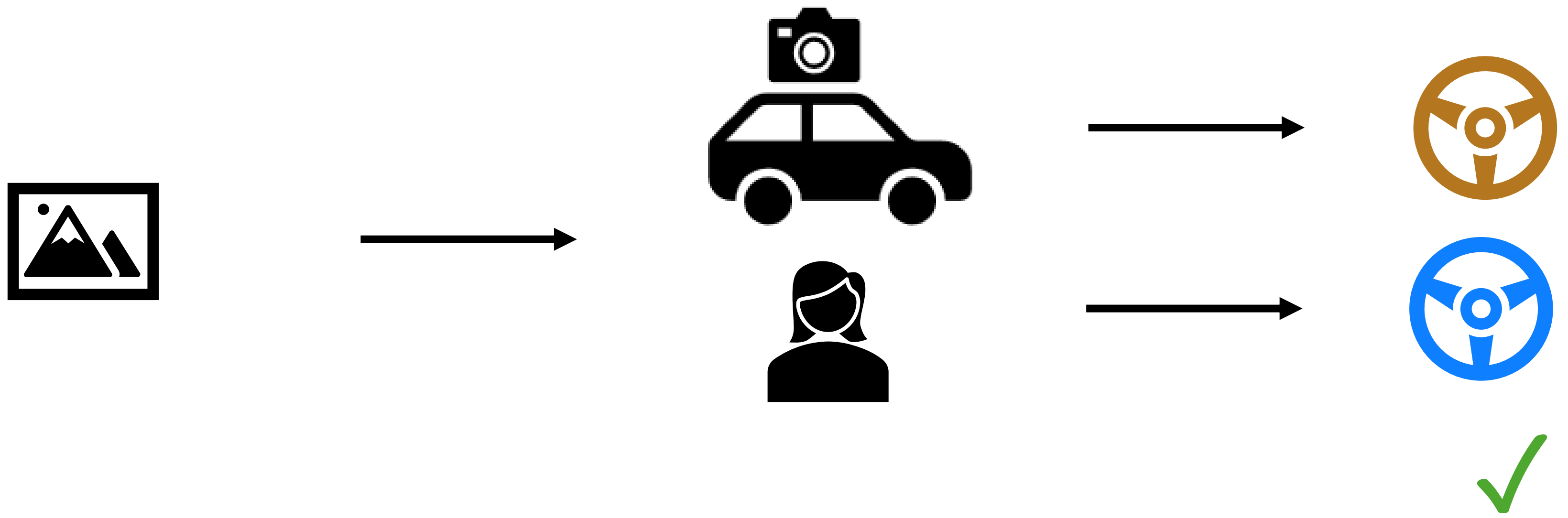
In Shadow Mode, operating on Tesla vehicles since 2016, if the car's Autopilot computer is not controlling the car, it is simulating the driving process in parallel with the human driver. When its own predictions do not match the driver's behavior, this might trigger the recording of a short "snapshot" of the car's cameras, speed, acceleration, and other parameters for later uploading to Tesla. Snapshots are also triggered when a Tesla crashes.

In 2015, Tesla obtained sensor data from 1 million miles every 10 hours

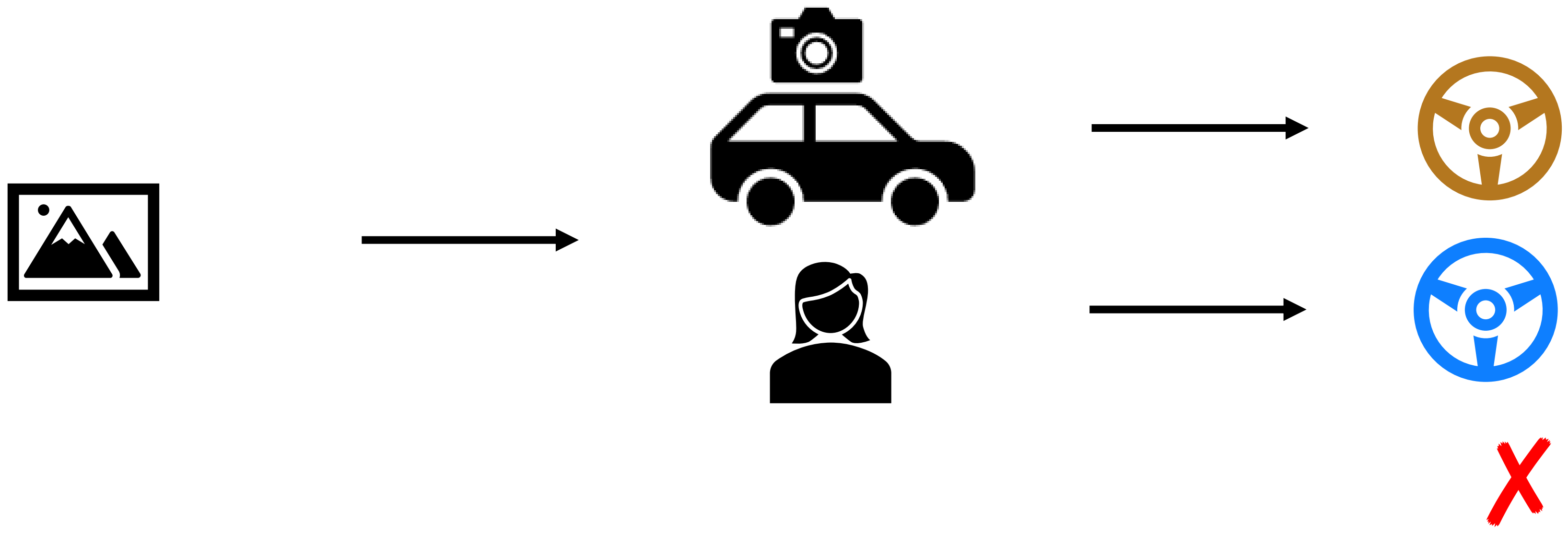
Test Oracles



Existing Test Suites



Existing Test Suites



Existing Test Suites

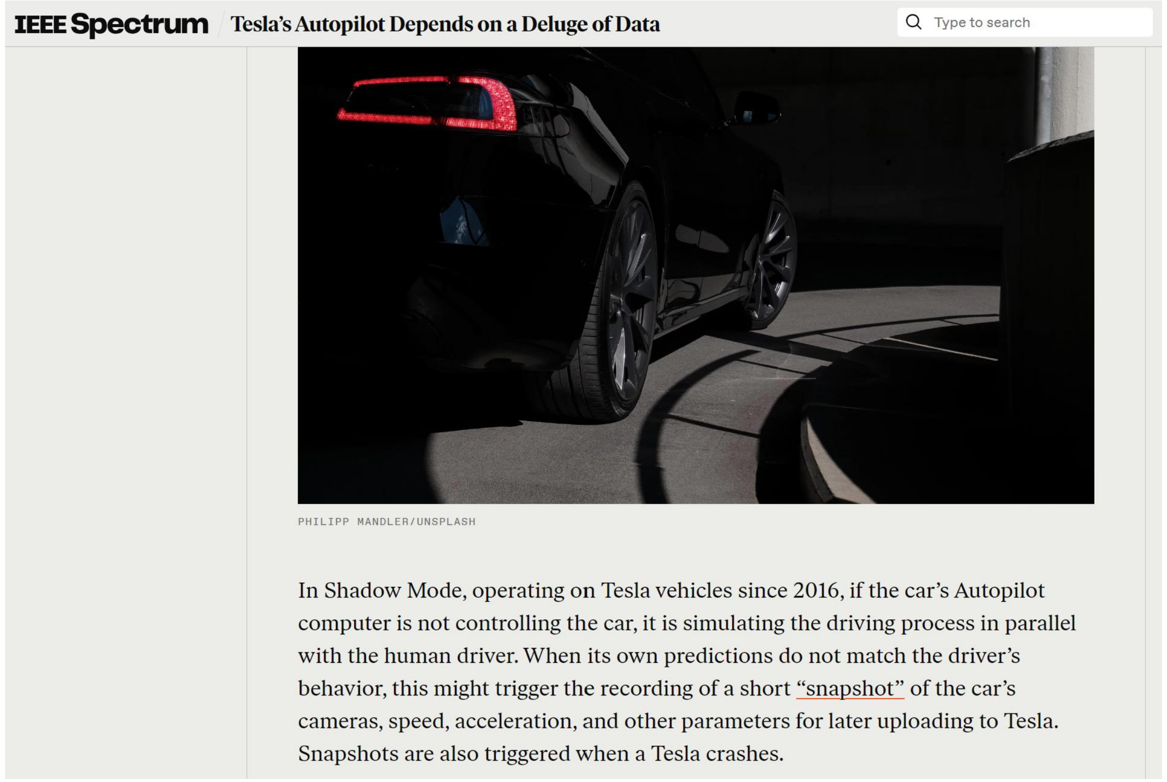
How do we get enough labeled data?
What if the original label is wrong?

Cautious Aggressive Drunk Distracted

Differential Testing

- Use unlabeled data
- Leverage multiple systems
  to find the correct answer

WAYMO



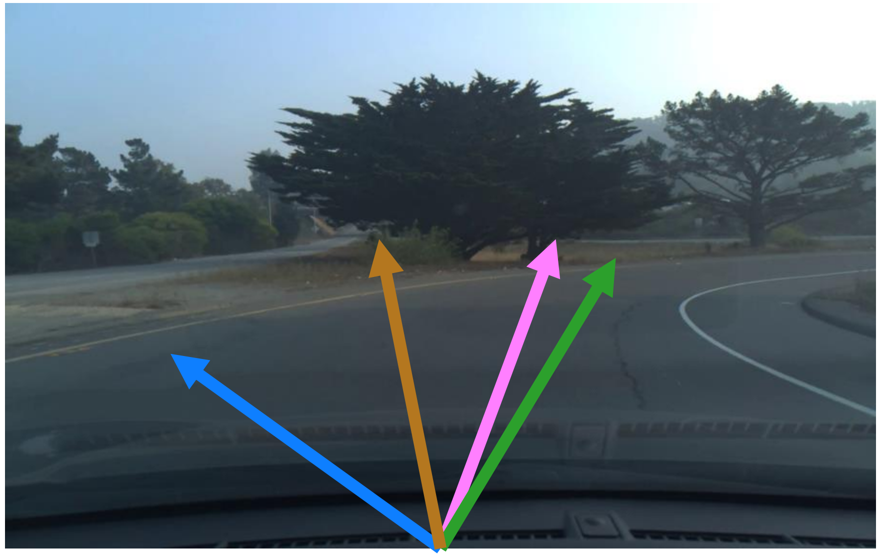
IEEE Spectrum Tesla's Autopilot Depends on a Deluge of Data

Type to search

PHILIPP MANDLER/UNSPLASH

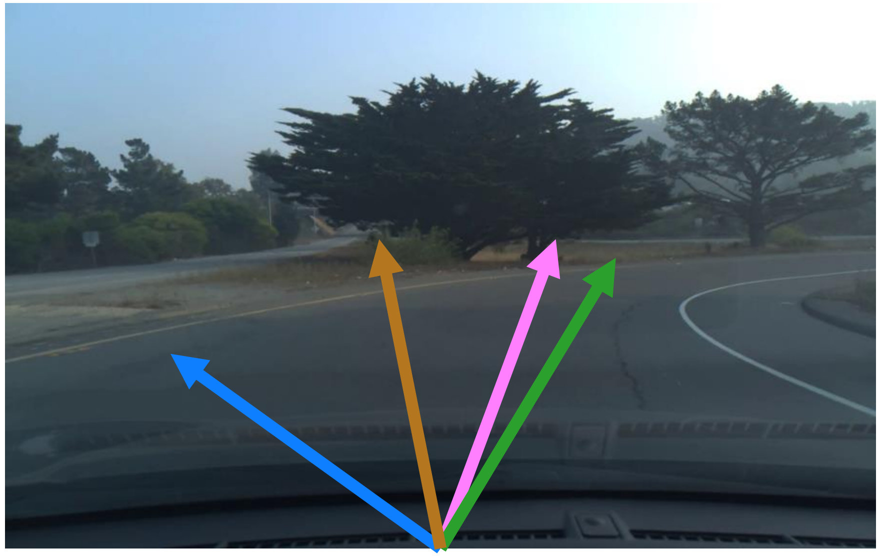
In Shadow Mode, operating on Tesla vehicles since 2016, if the car's Autopilot computer is not controlling the car, it is simulating the driving process in parallel with the human driver. When its own predictions do not match the driver's behavior, this might trigger the recording of a short "snapshot" of the car's cameras, speed, acceleration, and other parameters for later uploading to Tesla. Snapshots are also triggered when a Tesla crashes.

Differential Testing



Cautious Aggressive Drunk Distracted

Differential Testing

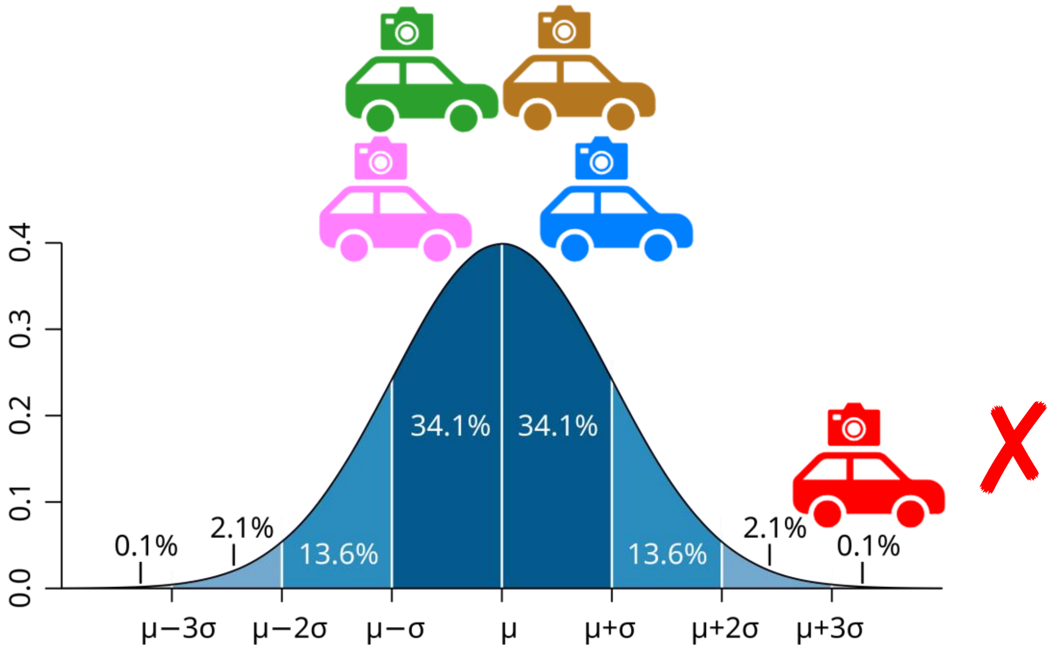


AV1 AV2 AV3 AV4

Differential Testing

- Use unlabeled data
- Leverage multiple systems
  to find the correct answer

Statistical Outlier Detection

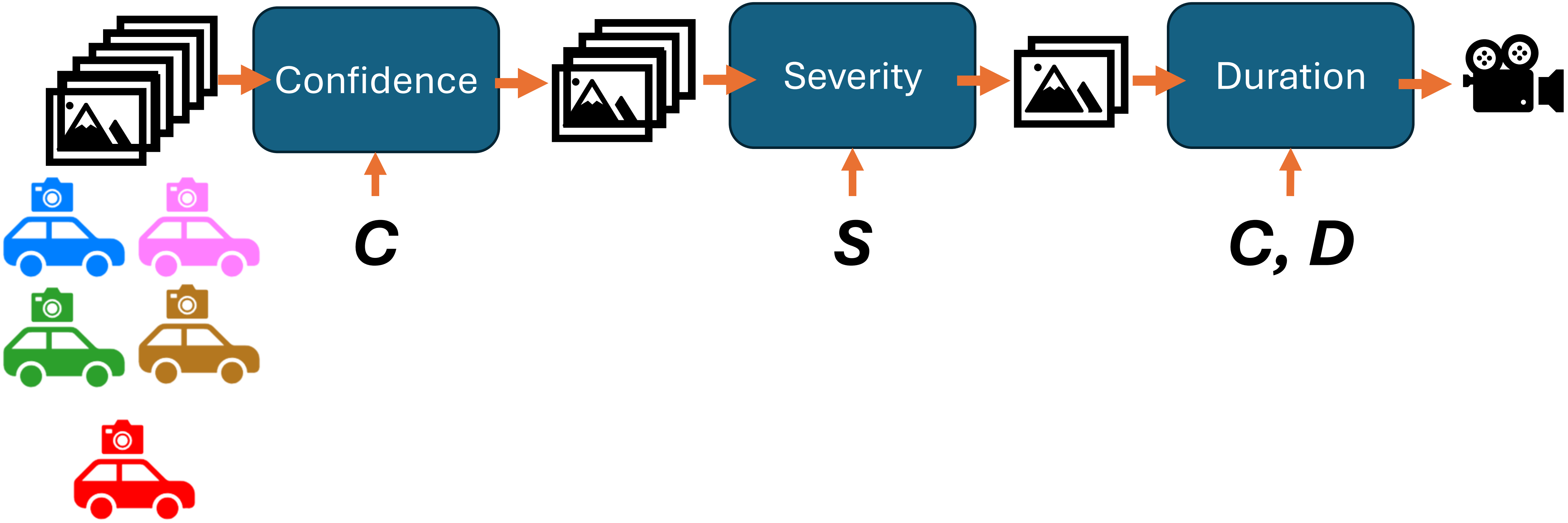


This requires knowing the distribution!

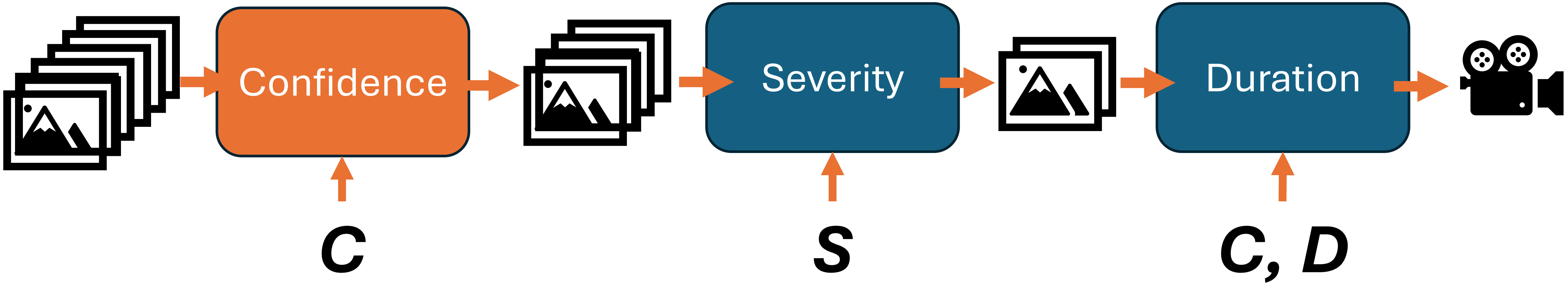
Building an Oracle

- Confidence
- Severity
- Duration

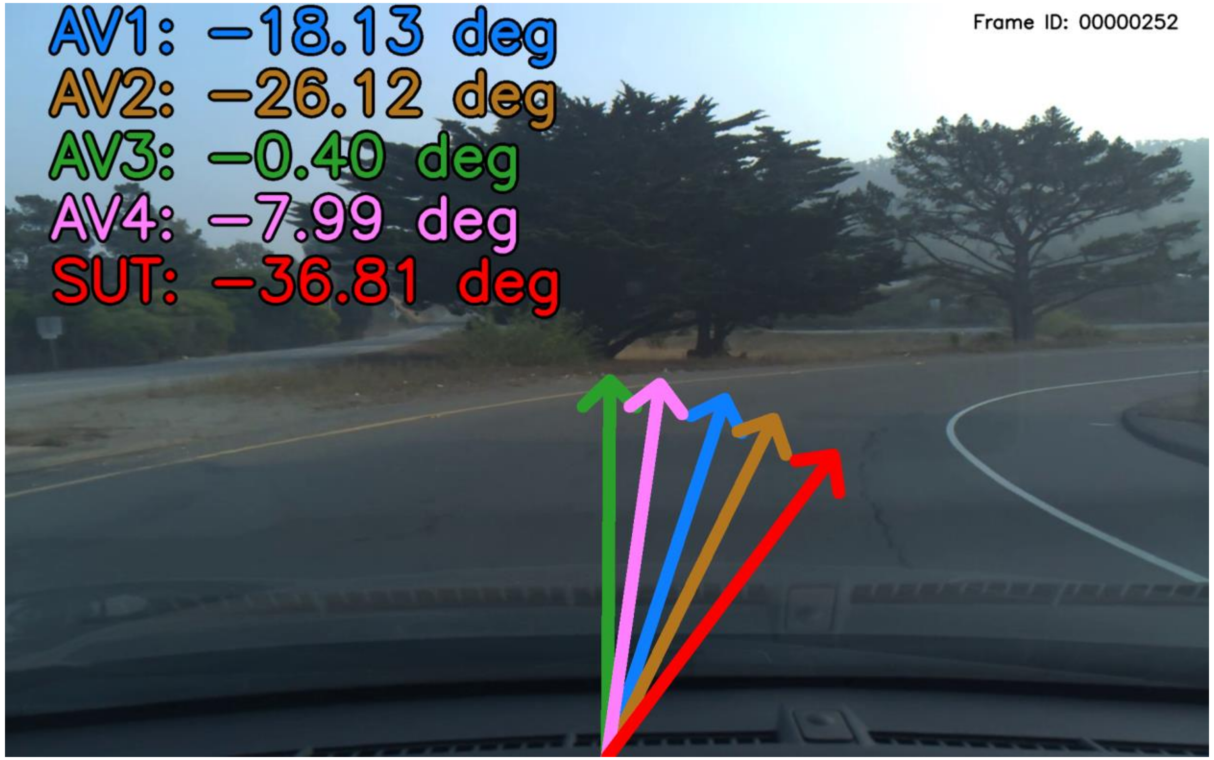
DiffTest4AV Approach



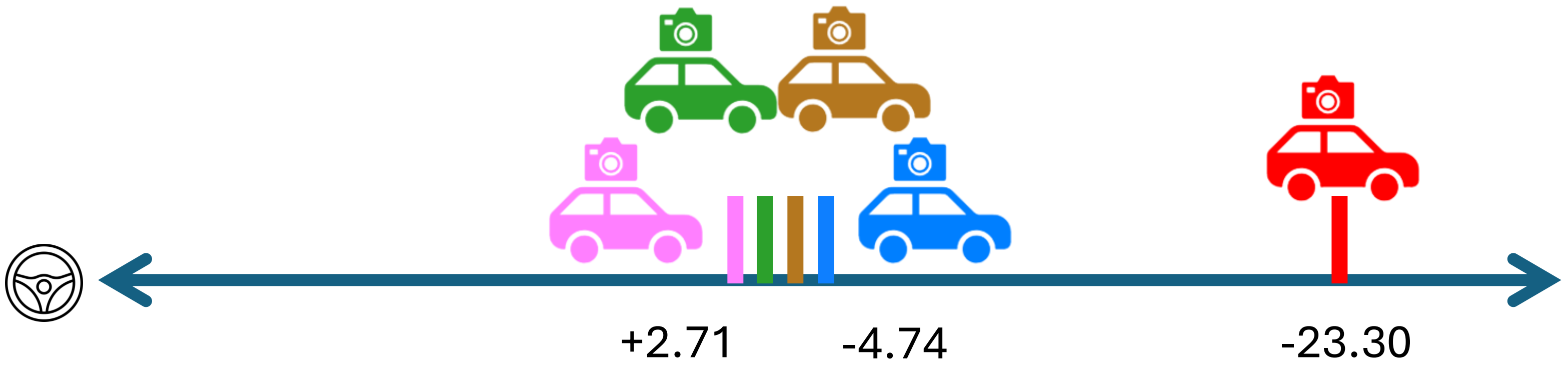
DiffTest4AV Approach



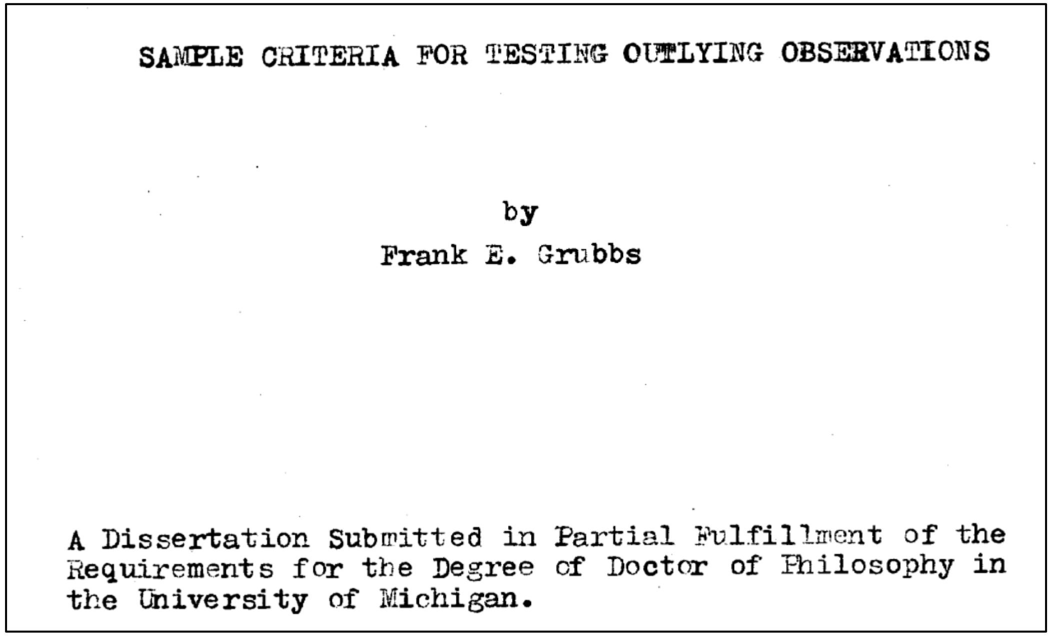
Why do we need confidence?



Confidence Through Outliers



$$Q = \frac{\max_{i} \left| Y_i - \bar{Y} \right|}{Y_{max} - Y_{min}}$$



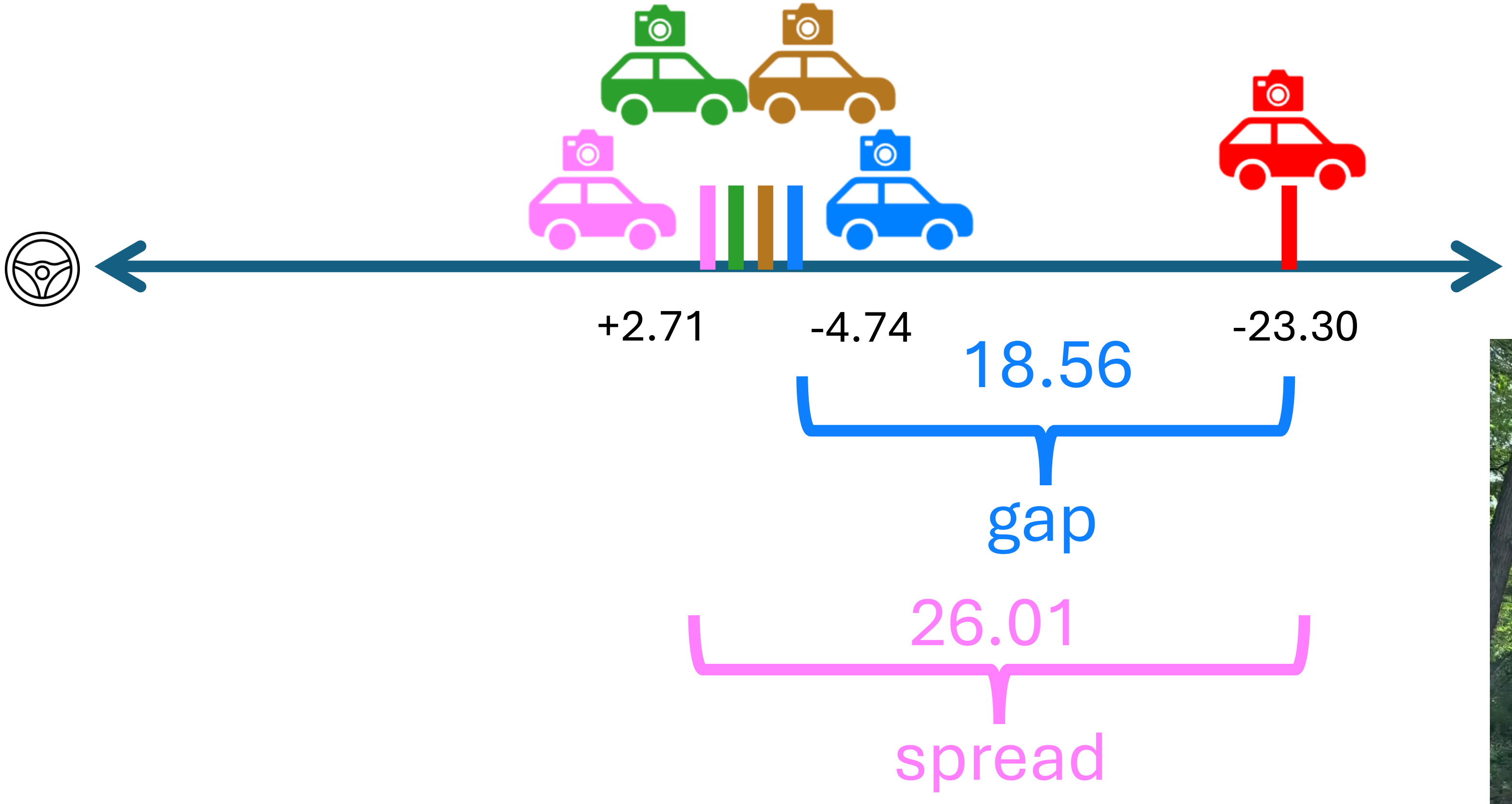
SAMPLE CRITERIA FOR TESTING OUTLYING OBSERVATIONS

by

Frank E. Grubbs

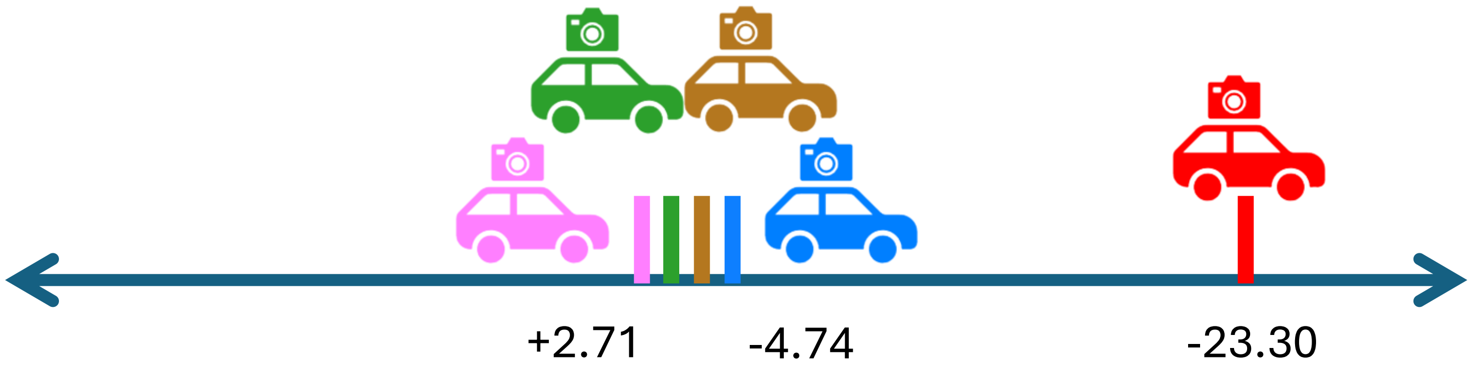
A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the University of Michigan.

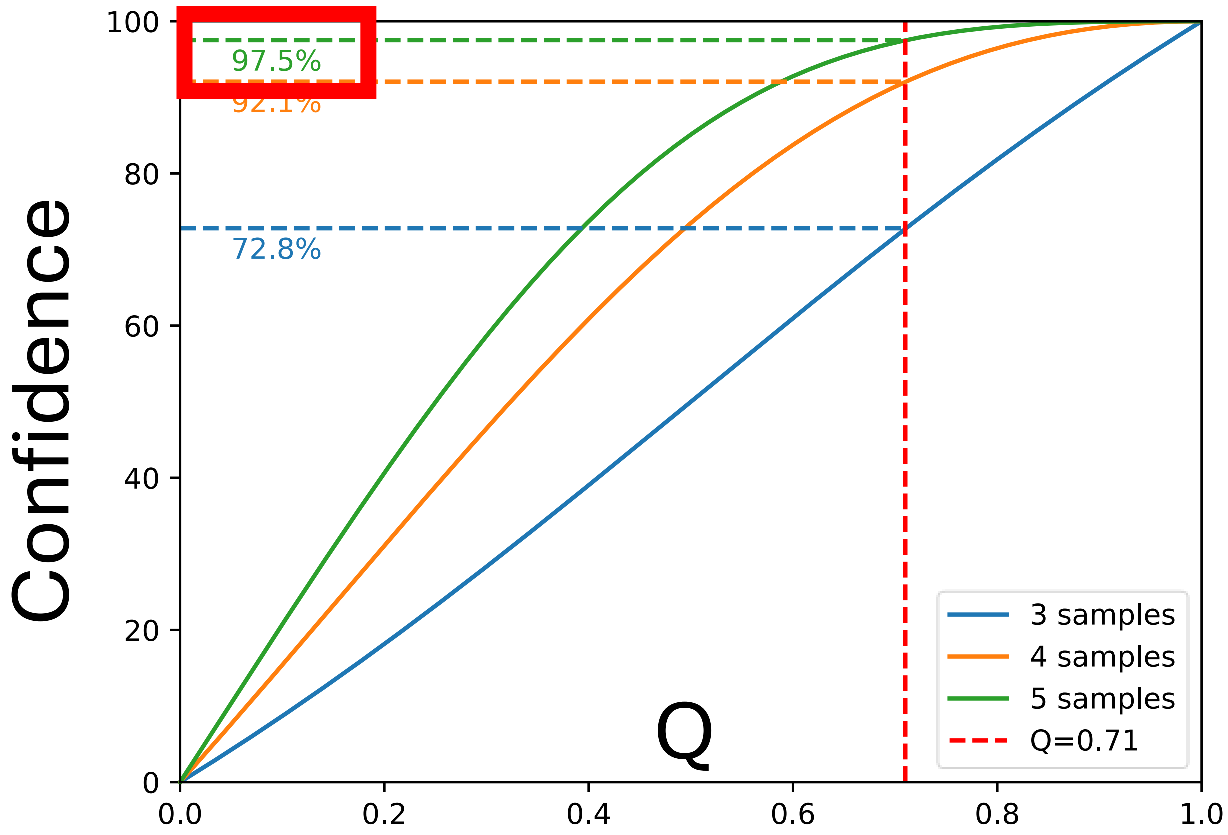
Confidence: Grubbs's Q Test



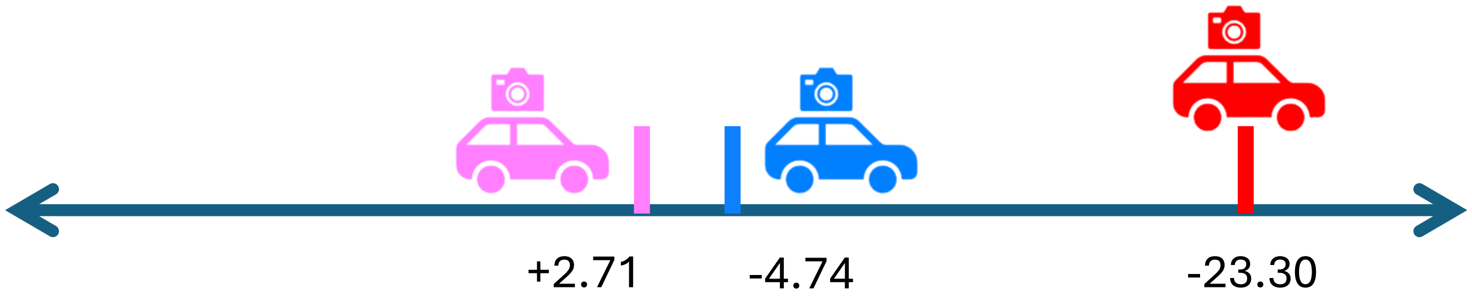
$$Q = \frac{gap}{spread} = \frac{18.56}{26.01} = .71$$

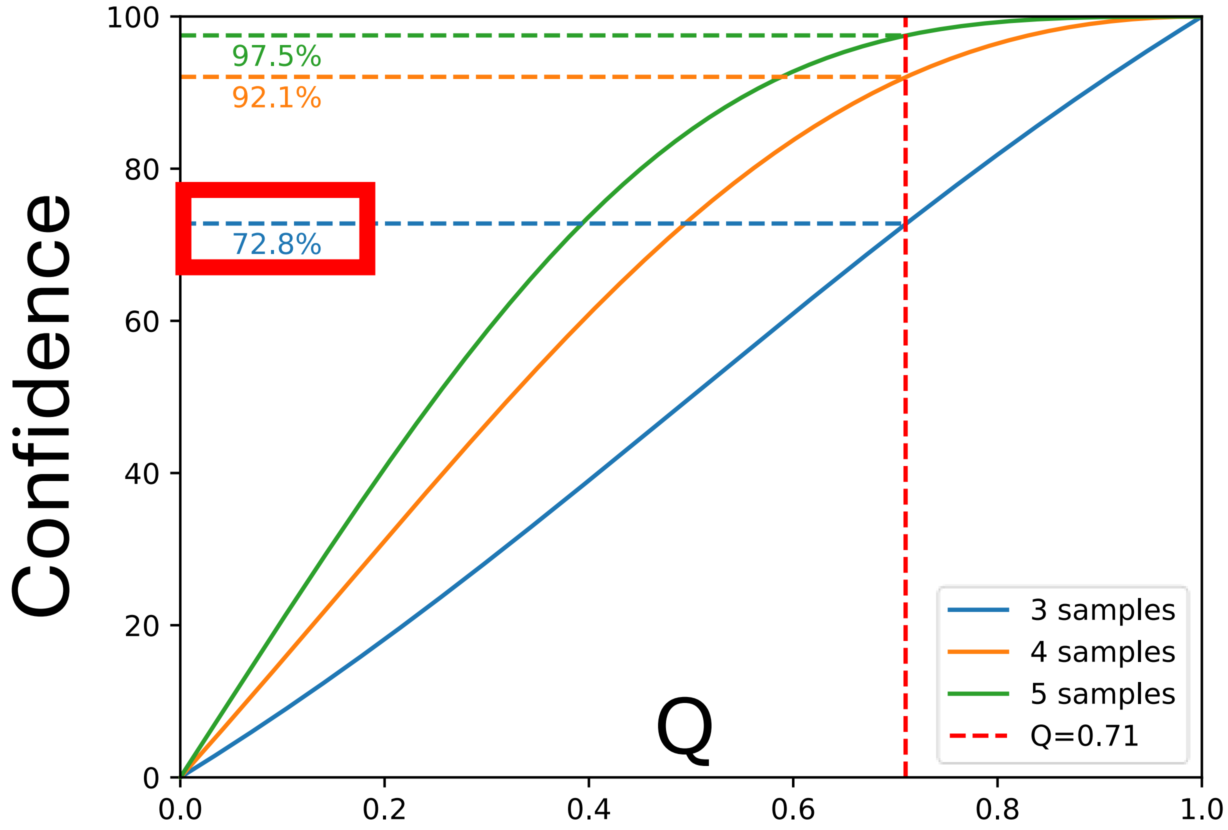
Confidence: Grubbs's Q Test



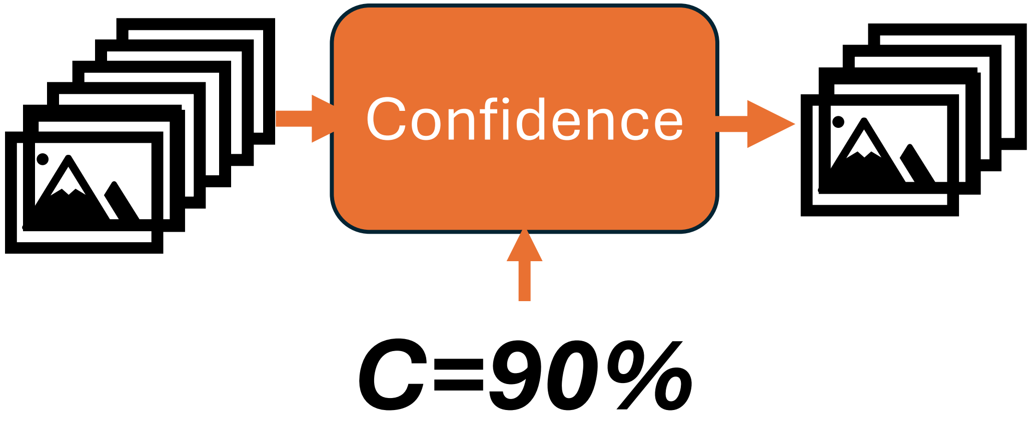


Confidence: Grubbs's Q Test





Threshold by Confidence

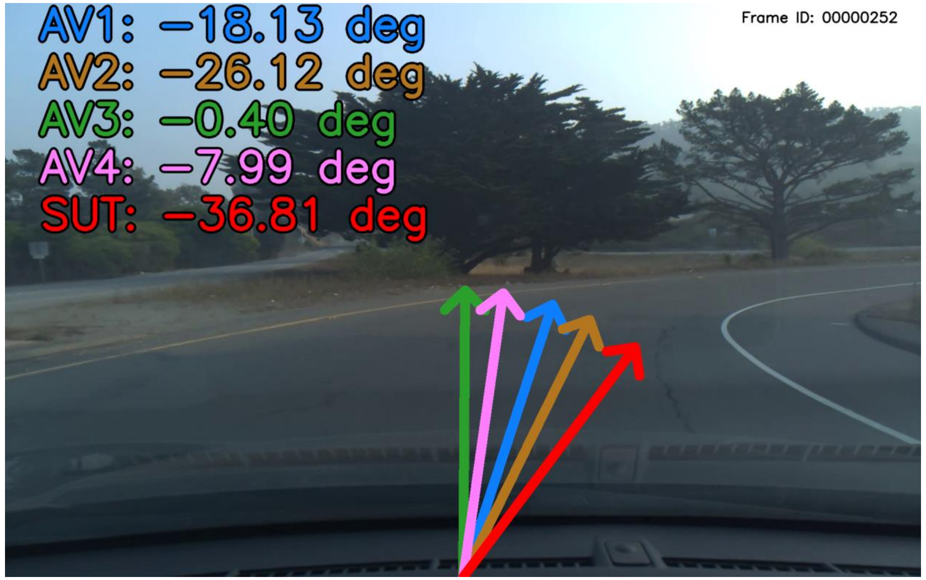


Confidence: 97.3%

✓

The Need for Confidence

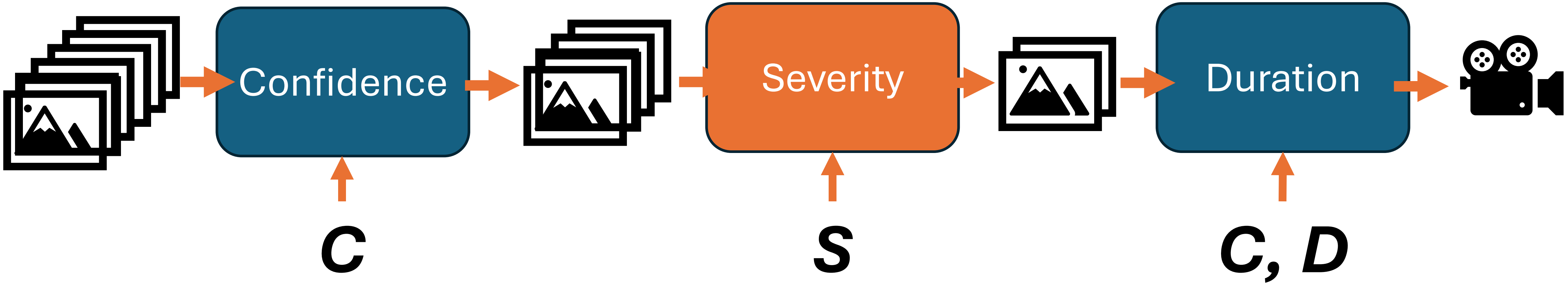
C=90%



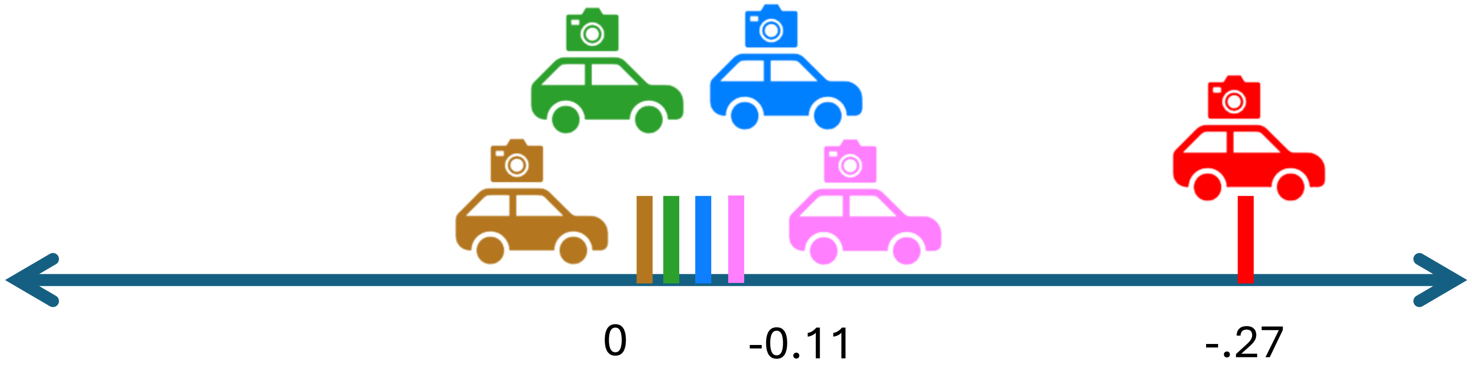
$$Q = \frac{gap}{spread} = \frac{10.7}{36.4} = .26$$

Confidence: 57.6% ✗

DiffTest4AV Approach



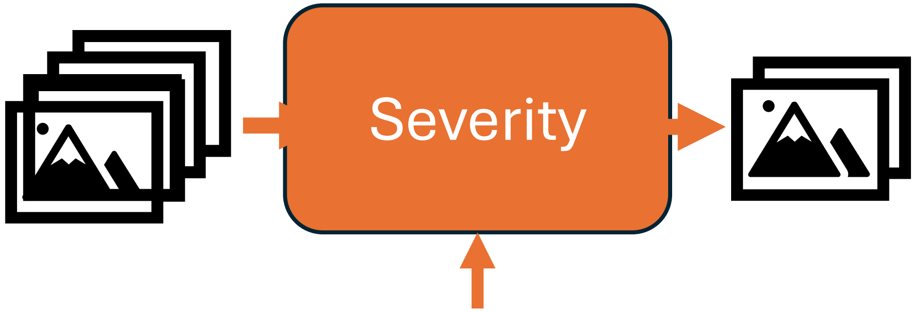
The Need for Severity



$$Q = \frac{gap}{spread} = \frac{0.16}{0.27} = .59$$

$$confidence = 92.3\%$$

Finding High Impact Failures



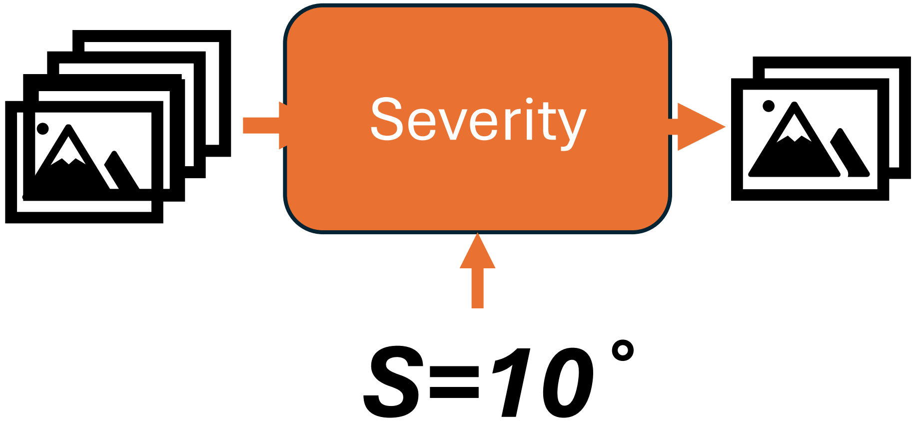
$S=10^\circ$

Confidence: 92.3%

Severity: 0.16°

✗

Finding High Impact Failures

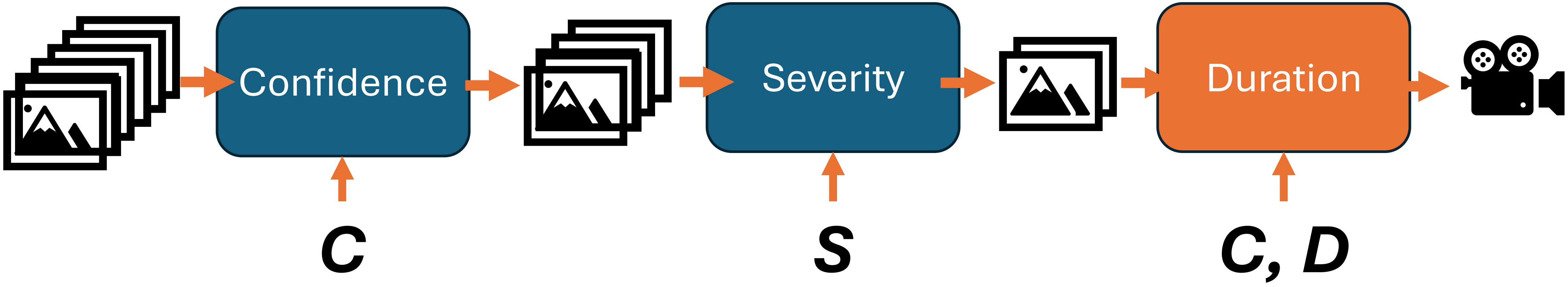


Confidence: 97.3%

Severity: 18.56°

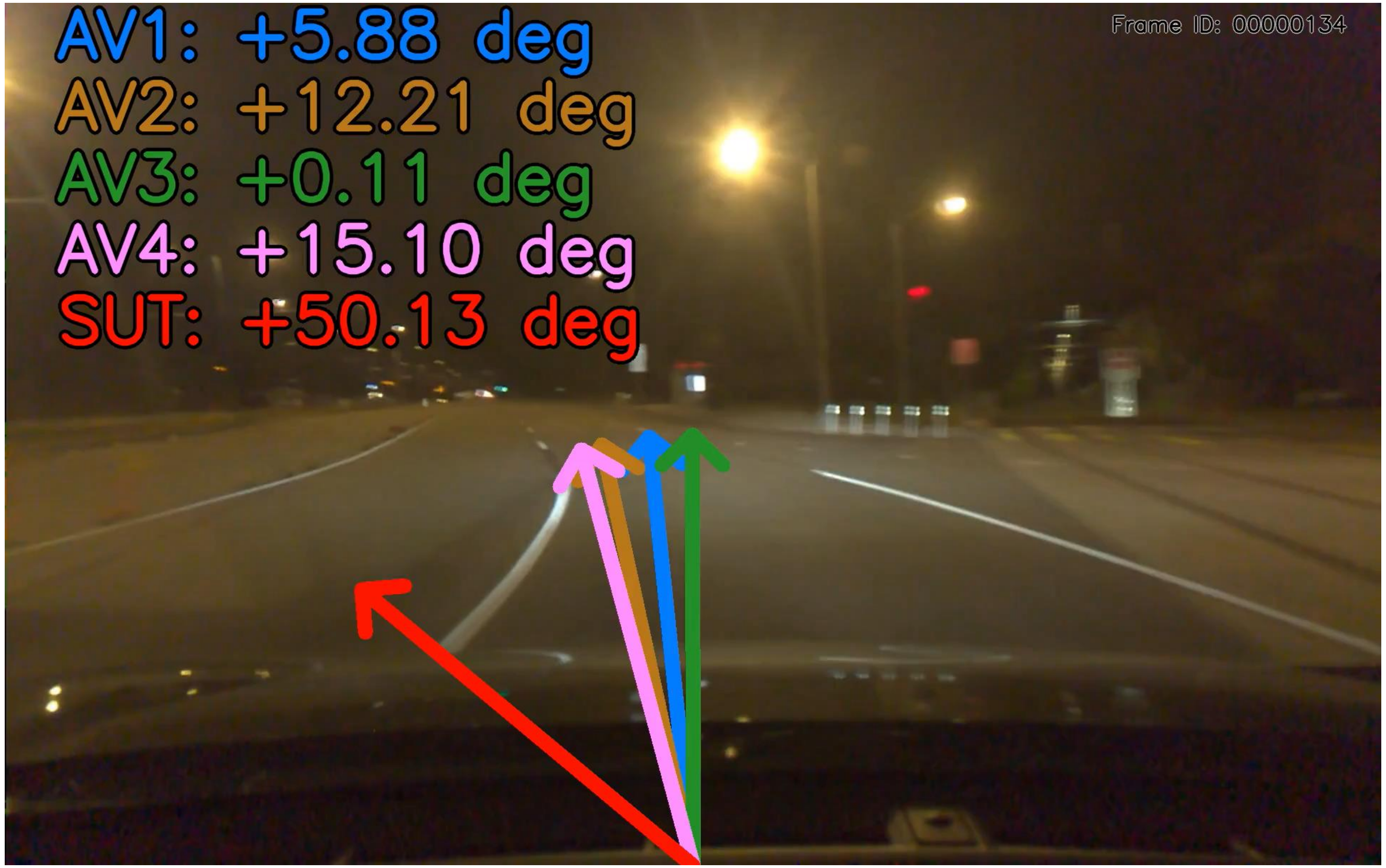
✓

DiffTest4AV Approach



Duration

- Continuous failures escalate to system failures



Duration

- Continuous failures escalate to system failures

Frame	142	143	144	145	146	147	148	149	150	151	152
Conf: (%)	98.78	98.93	98.65	99.16	99.48	99.51	99.51	99.7	99.7	99.88	99.87

Duration

- Continuous failures escalate to system failures

Frame	142	143	144	145	146	147	148	149	150	151	152
Conf: (%)	98.78	98.93	98.65	99.16	99.48	99.51	99.51	99.7	99.7	99.88	99.87

Option 1: Filter by minimum confidence **98.65%**

Duration

- Continuous failures escalate to system failures

Frame	142	143	144	145	146	147	148	149	150	151	152
Conf: (%)	98.78	98.93	98.65	99.16	99.48	99.51	99.51	99.7	99.7	99.88	99.87

Option 1: Filter by minimum confidence **98.65%**

Option 2: Cumulative Confidence **93.37%**

$$\prod_{i=j}^{j+m} confidence(t_i)$$

Duration

- Continuous failures escalate to system failures

$$\prod_{i=j}^{j+m} confidence(t_i)$$

Frame	142	143	144	145	146	147	148	149	150	151	152
Conf: (%)	98.78	98.93	98.65	99.16	99.48	99.51	99.51	99.7	99.7	99.88	99.87

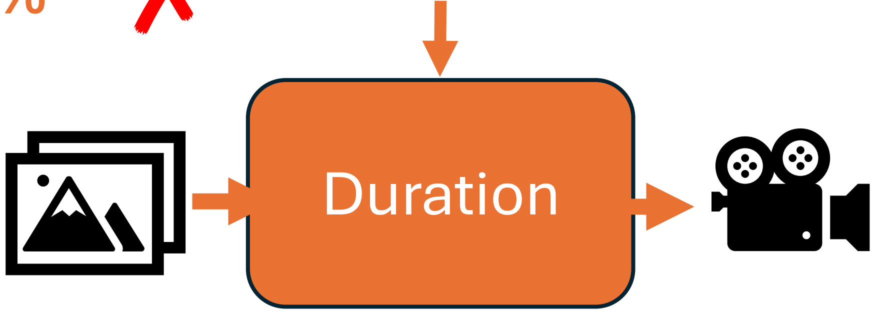
Option 1: Filter by minimum confidence **98.65%** ✓

Option 2: Cumulative Confidence **93.37%** ✗

$$\prod_{i=j}^{j+m} confidence(t_i)$$

D=10

C=95%

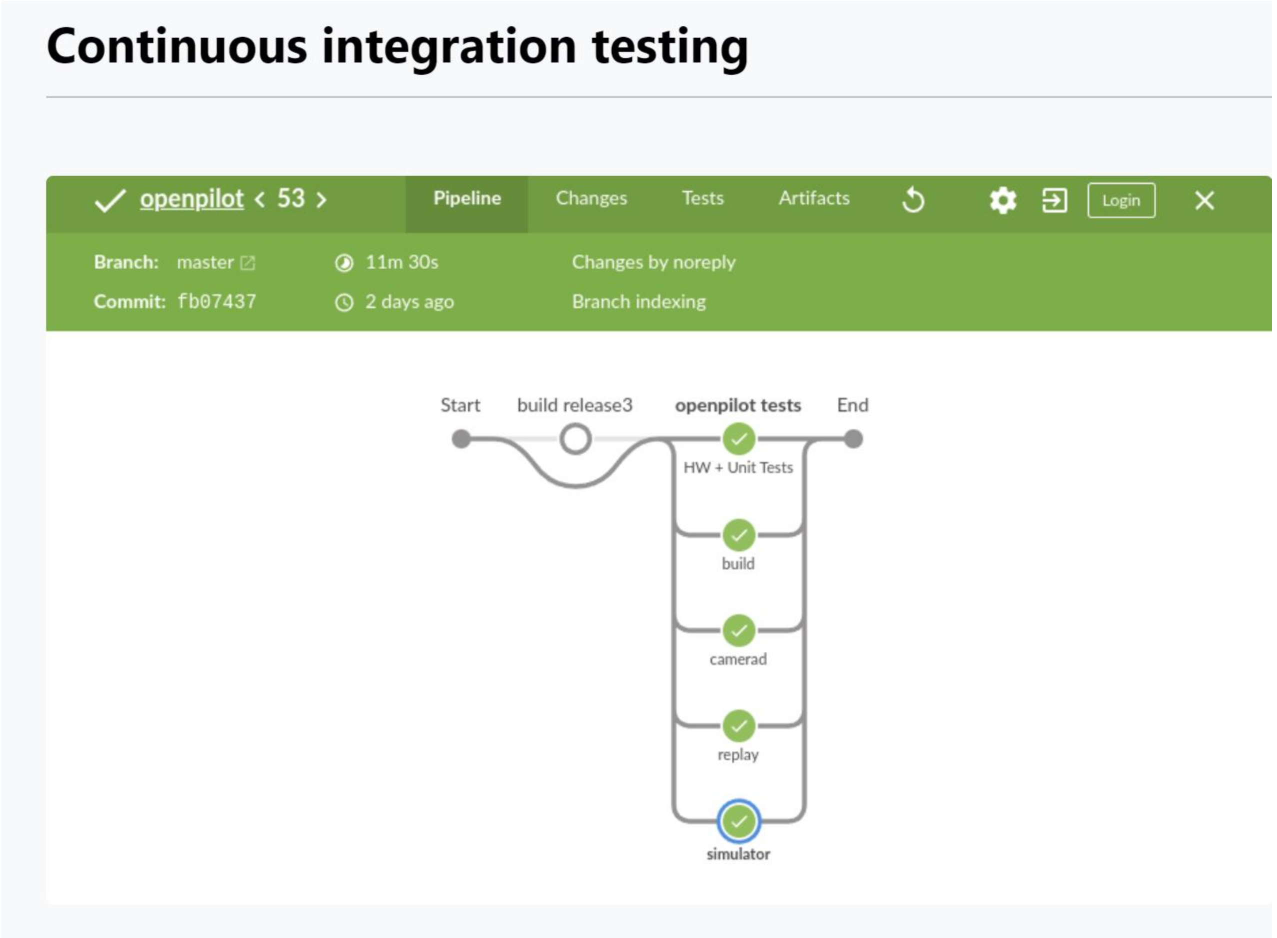


Study

- RQ1: High-confidence failures
- RQ2: High-confidence & High-severity failures
- RQ3: High-impact & long-running failures

Experiment Setup: AV

- comma.ai OpenPilot



CI tests that run on every openpilot commit on comma 3 hardware

Experiment Setup: AV

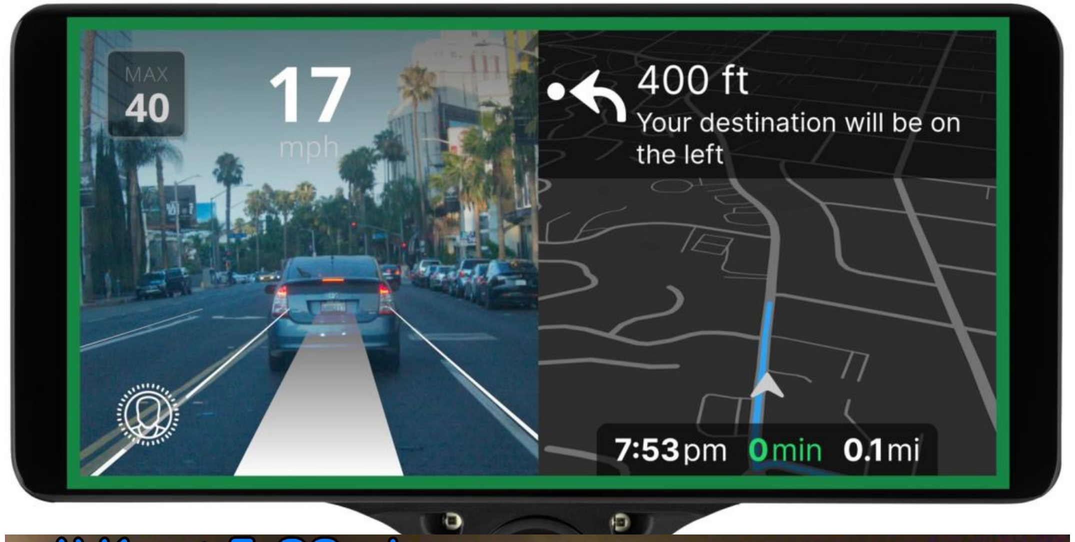
- comma.ai OpenPilot
  - Apr 2022 (AV1)
  - Jul 2022 (AV2)
  - Nov 2022 (AV3)
  - Mar 2023 (AV4)
  - Jun 2023 (SUT)

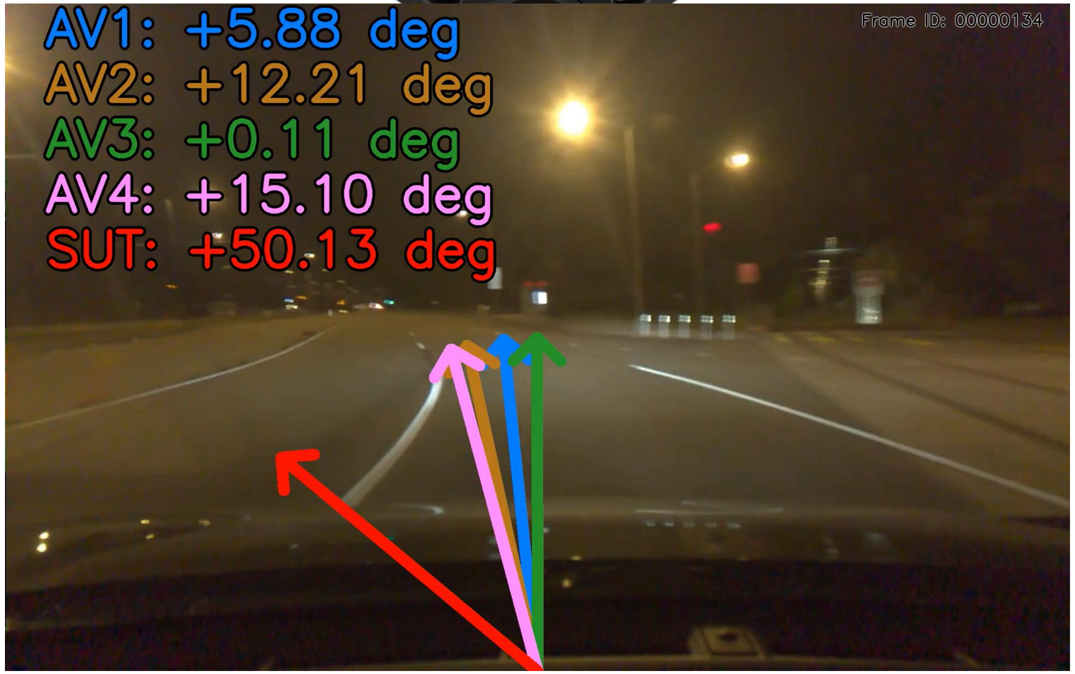
Experiment Setup: Data

- comma.ai 2016 
  - 11 videos; 391,843 images
- comma.ai 2k19 
  - 2035 videos; 1,825,111 images
- External JUtah
  - 50 videos; 2,362,708 images

Some Interesting Data

- At 90% Confidence:
  - External JUtah has the most failures, but ~2% >10 °
  - Comma.ai 2k19 has *all* the >50° failures and the longest failure





Takeaways

From 4,579,662 inputs
identify 81 (0.002%)
high-impact failures at
90% confidence and 50° severity
Including failures up to 72 frames (4.8s)

Takeaways

In 2015, Tesla obtained
1 million miles every 10 hours

Even if Tesla failed 1 million times less often, DiffTest4AV would find 31 failures per year!

IEEE Spectrum | Tesla's Autopilot Depends on a Deluge of Data

Type to search

PHILIPP MANDLER/UNSPLASH

In Shadow Mode, operating on Tesla vehicles since 2016, if the car's Autopilot computer is not controlling the car, it is simulating the driving process in parallel with the human driver. When its own predictions do not match the driver's behavior, this might trigger the recording of a short "snapshot" of the car's cameras, speed, acceleration, and other parameters for later uploading to Tesla. Snapshots are also triggered when a Tesla crashes.

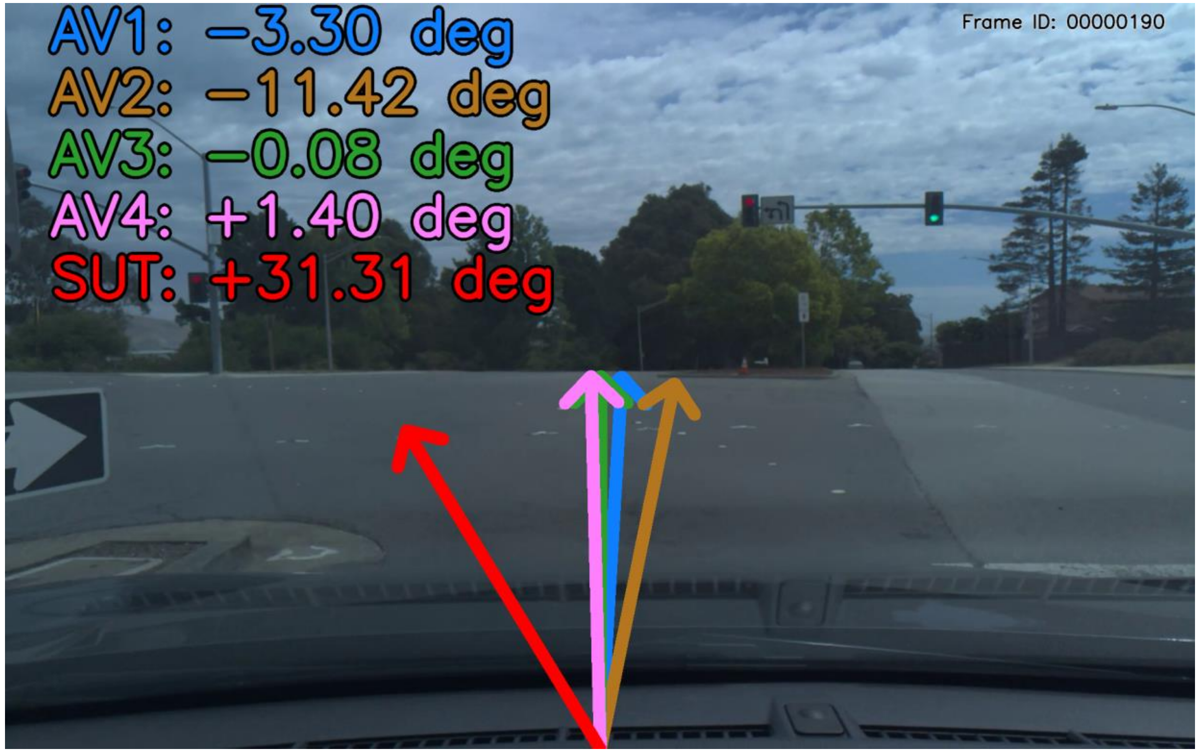
Future Work

- Removing False Positives:

  Check Requirement Preconditions: ODD

Limitations of openpilot ALC and LDW

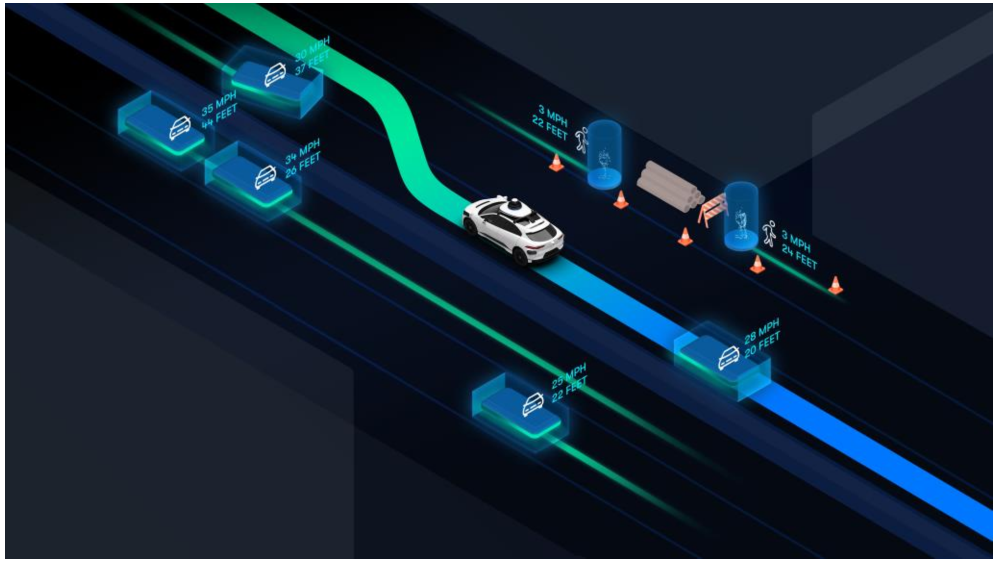
- When in sharp curves, like on-off ramps, intersections etc.; openpilot is designed to be limited in the amount of steering torque it can produce.



Future Work

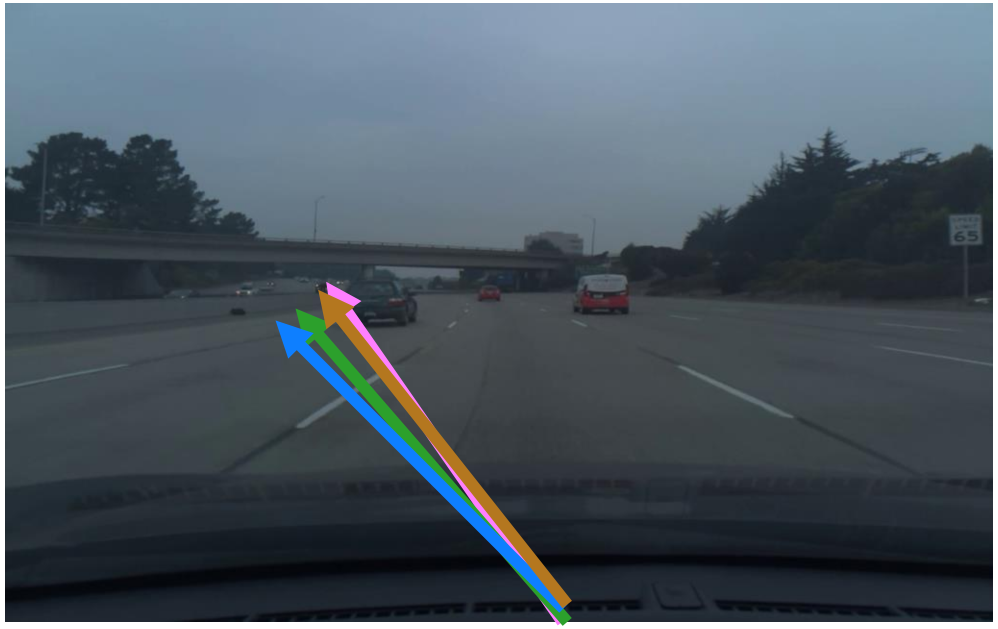
Future Work

- Multidimensional Behavior



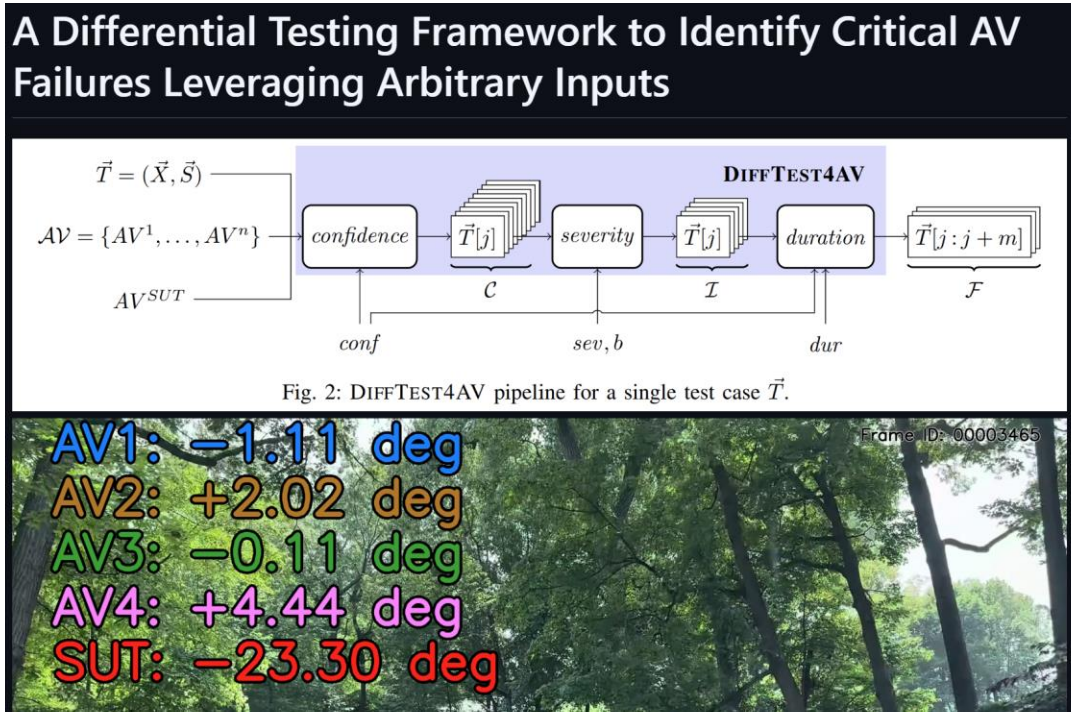
- Statistical Assumptions & Oracle Strength

✓

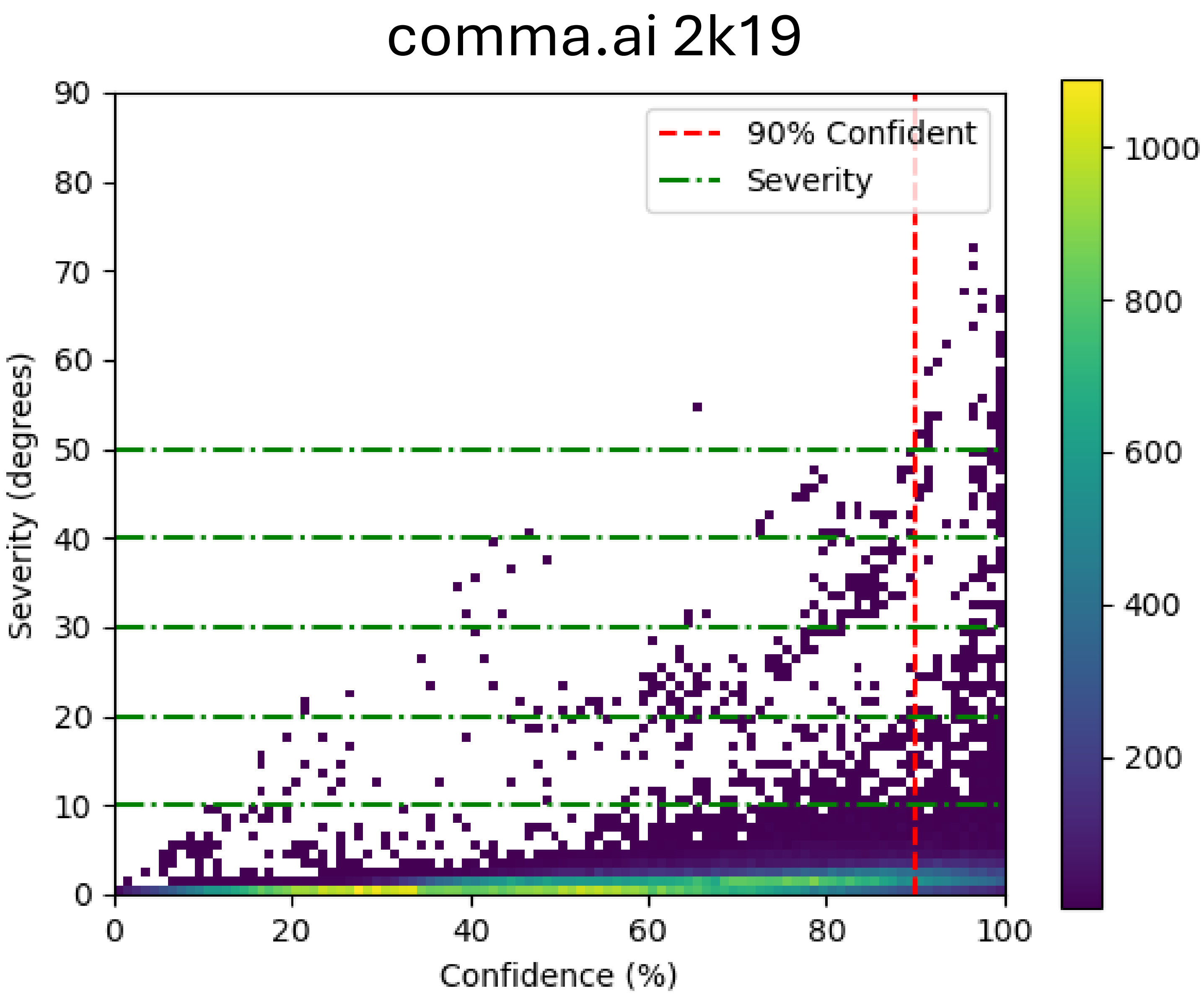


Questions?

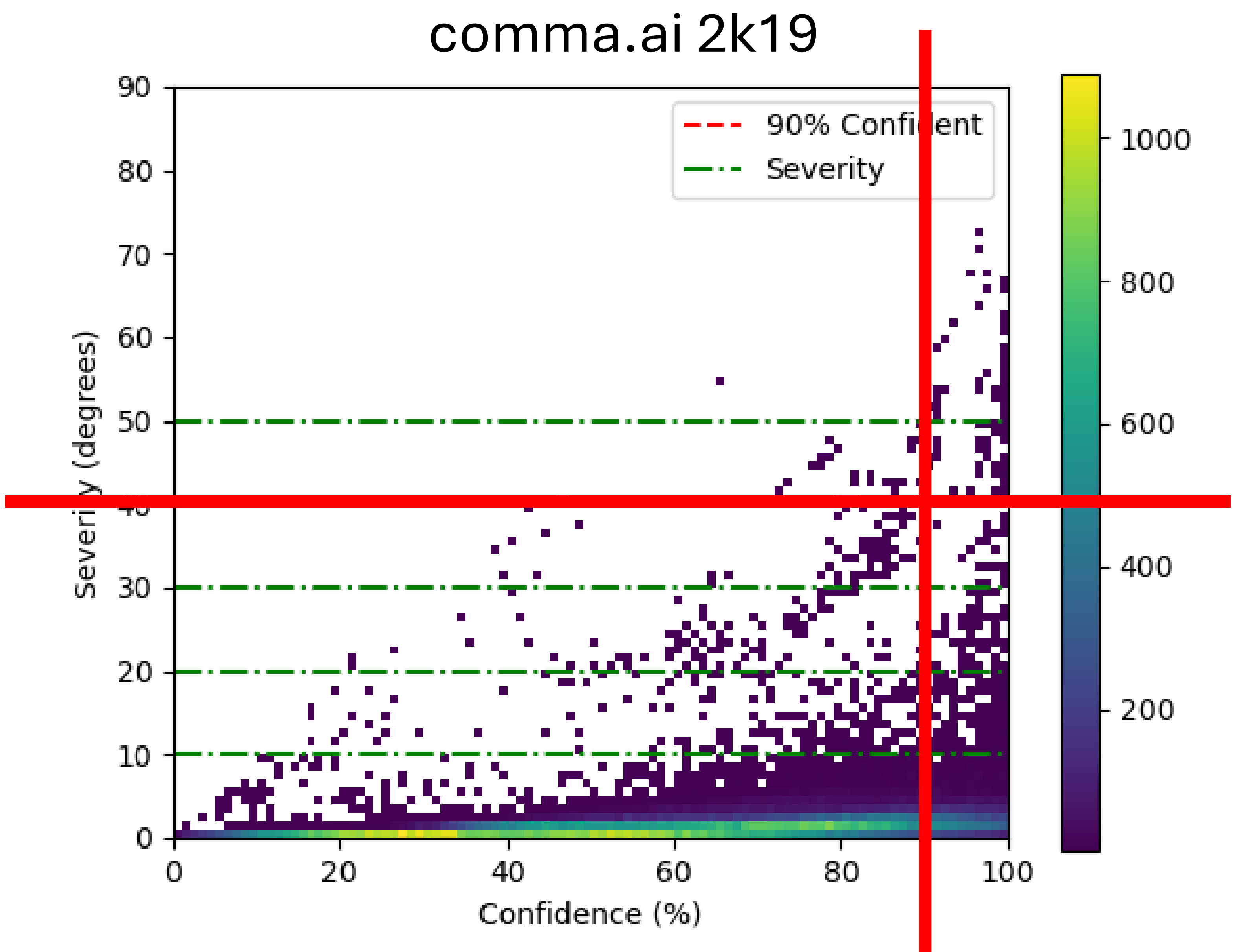
https://github.com/less-lab-uva/DiffTest4AV



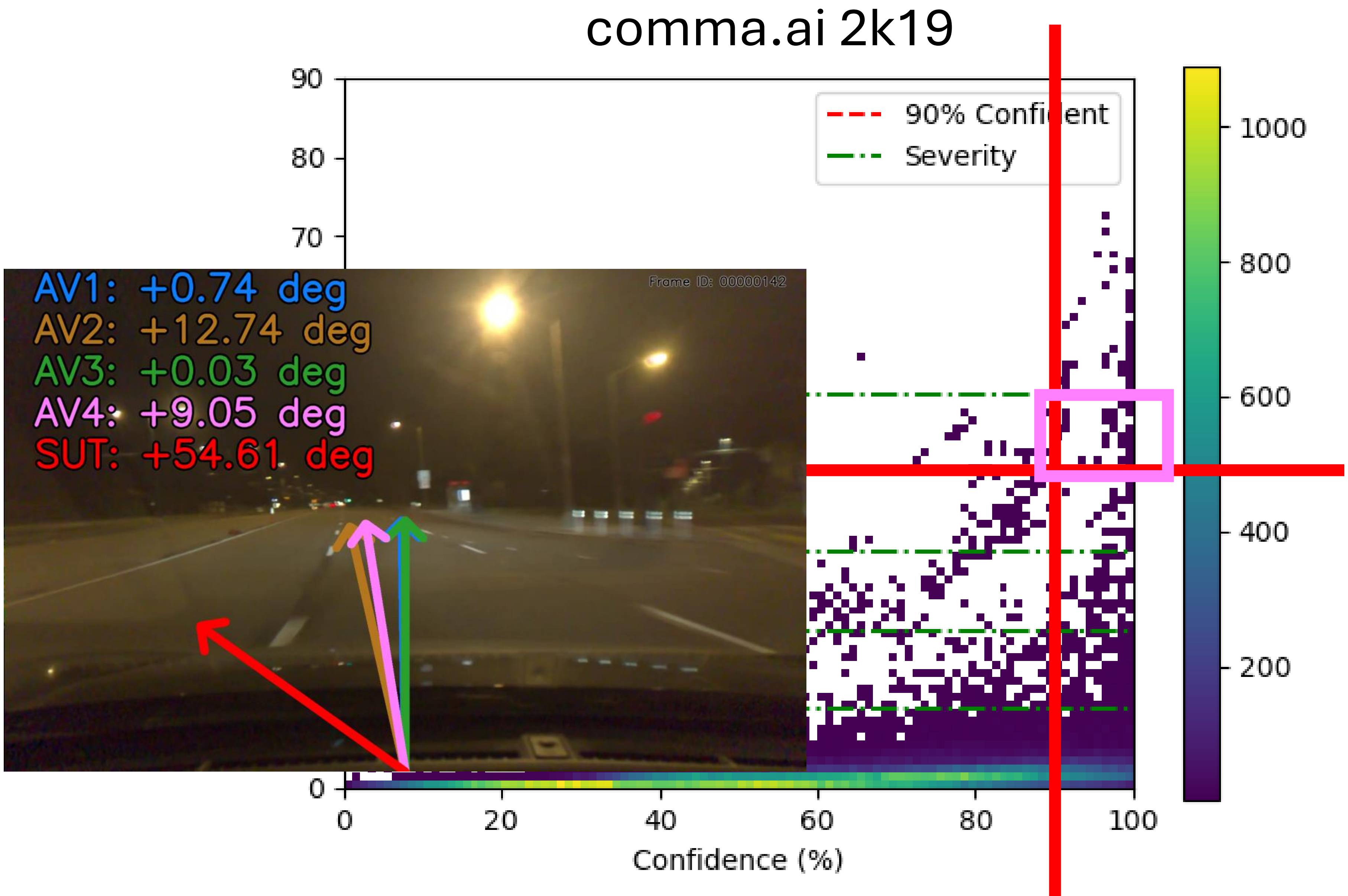
Experimental Results: Confidence vs Severity



Experimental Results: Confidence vs Severity



Experimental Results: Confidence vs Severity



Study

- RQ1: High-confidence failures
- RQ2: High-confidence & High-severity failures
- RQ3: High-impact & long-running failures

Experimental Results: Duration

 # Frames

	50%	75%	90%	95%	99%
10˚	72	58	34	22	14
20˚	64	56	34	22	14
30˚	52	48	34	20	14
40˚	42	42	27	19	11
50˚	33	33	27	19	5