What Makes It Testable?

Practical approach to quantifying the safety of Self-Driving cars

11/02/2017

Edward Schwalb, Ph.D
Lead Scientist
Machine Learning Group

Safety Statistics

How Good are Human Drivers?

Humans drive >100 million miles between catastrophic events

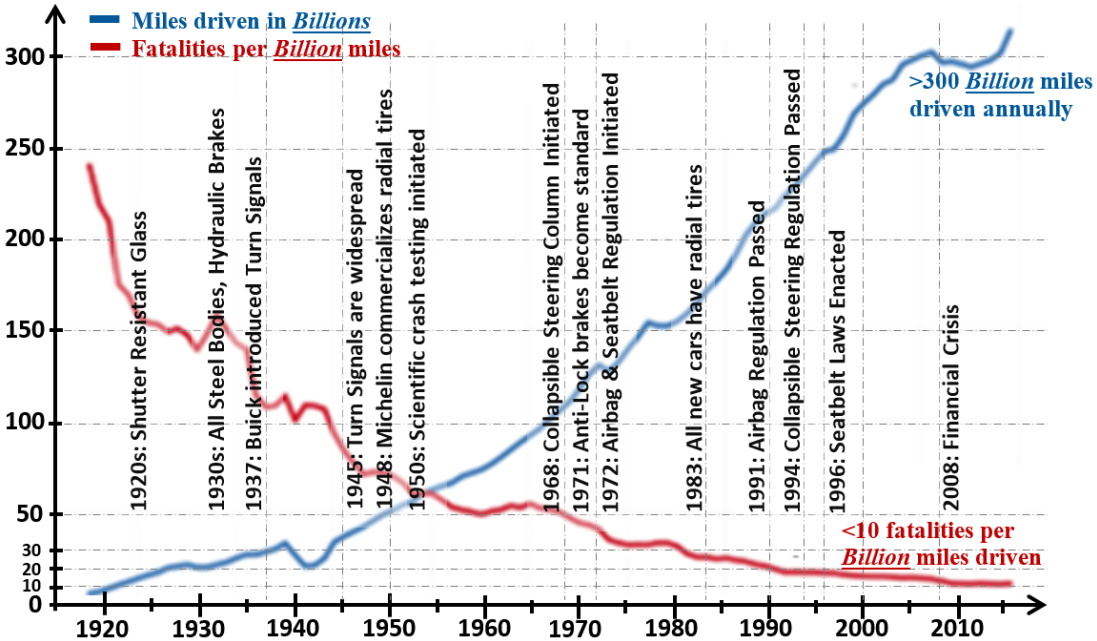
2015 Records show:

- 35,000 deaths
- 2.5 million injuries
- 6 million crashes

100 million miles:

- **1.1** fatalities
- **77** injuries
- **190** crashes

Despite Human competence, we see a large number of deaths and injuries because we drive lots of miles …



A.V. must be better !

Likelihood vs Risk	Negligible	Marginal	Critical	Catastrophic
Frequent $<10^{-3}$	X	X	X	X
Probable $<10^{-5}$		X	X	X
Remote $<10^{-7}$			X	X
Improbable $<10^{-9}$				X

Cannot use "traditional" metrics:

99.999% "accuracy" is *not* sufficient !!!

100 million miles = 10^{12} frames!

(10^4 f/m)

Multi-Pronged Testing Approach

Road + Track + Simulation Safety Quotient

Road Test

$10^?$ scenarios

Track Test

$10^?$ scenarios

Simulation Test

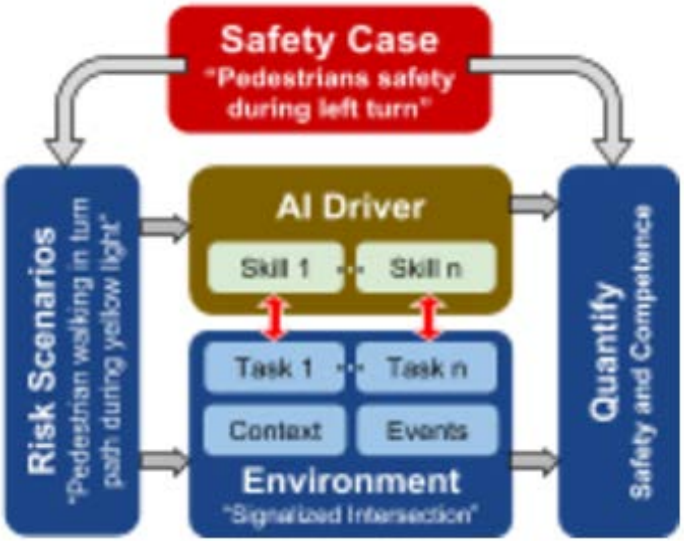
10^9 scenarios

Industry Consensus:

Cannot cover all scenarios in track nor road tests.

Simulation is a necessary for Autonomous Vehicle Development.

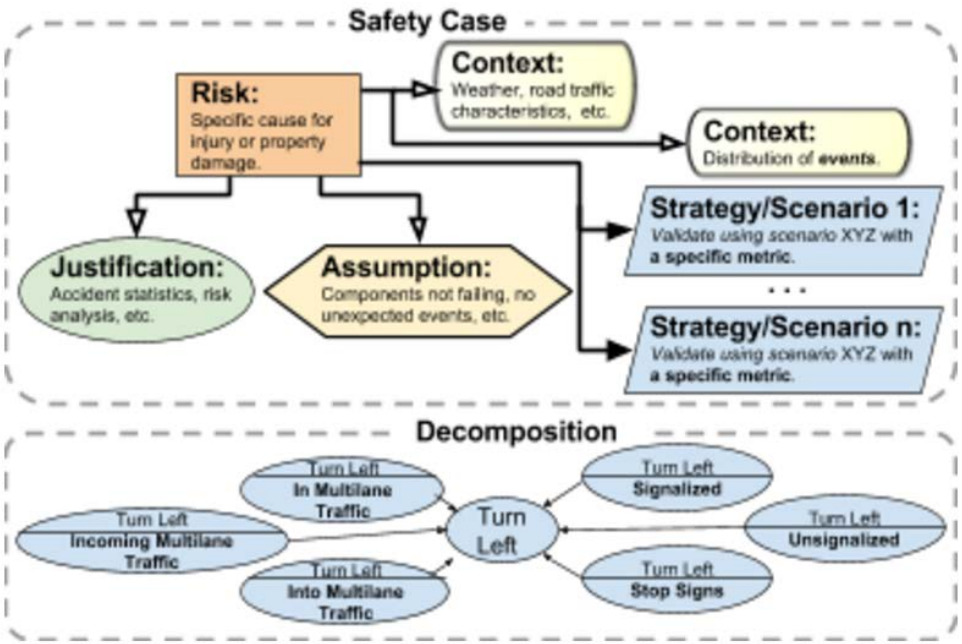
- Should be standardized
- Capabilities of MSC Toolchain
- OEM Intellectual Property



Risk Management Best Practices

Systematically Characterize and Prioritize Risk

Characterization of Risk + ***Prioritization of Risk***



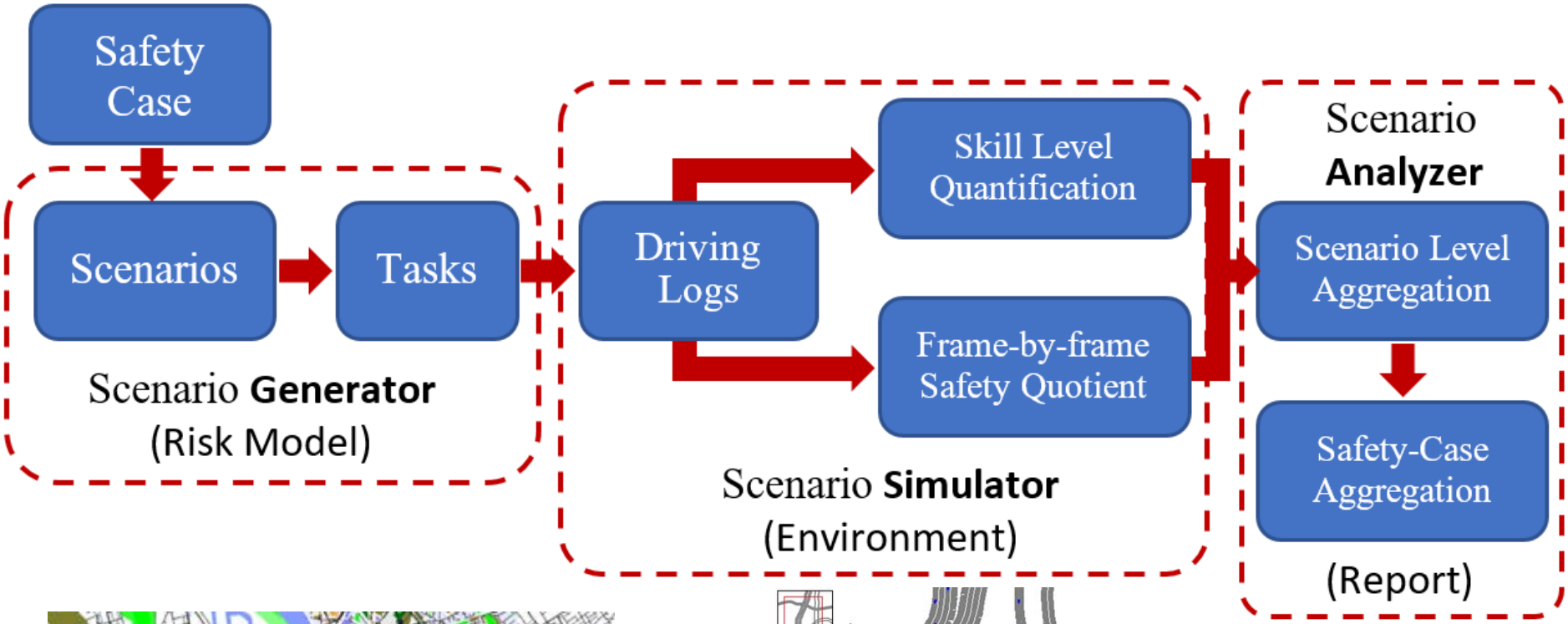
Controlability	Exposure	Severity			
		S0	S1	S2	S3
C1	E1	QM	QM	QM	QM
	E2	QM	QM	QM	QM
	E3	QM	QM	QM	A
	E4	QM	QM	A	B
C2	E1	QM	QM	QM	QM
	E2	QM	QM	QM	A
	E3	QM	QM	A	B
	E4	QM	A	B	C
C3	E1	QM	QM	QM	A
	E2	QM	QM	A	B
	E3	QM	A	B	C
	E4	A	B	C	D

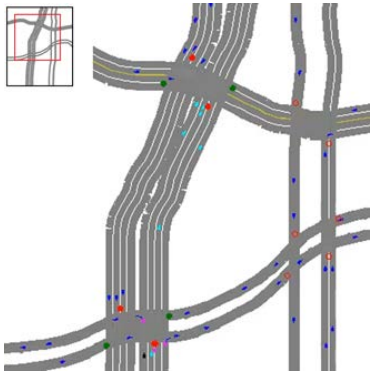
Severity $\times$ $P_{Exposure}$ $\times$ $P_{Controllability}$ $\times$ $P_{Integrity}$ ➡ Risk

Safety Critical Systems approach to AV development …

Quantifying Safety

Statistically Significant Testing using Simulation





Extensive FHWA Approach

Context, Events, Tasks, Metrics

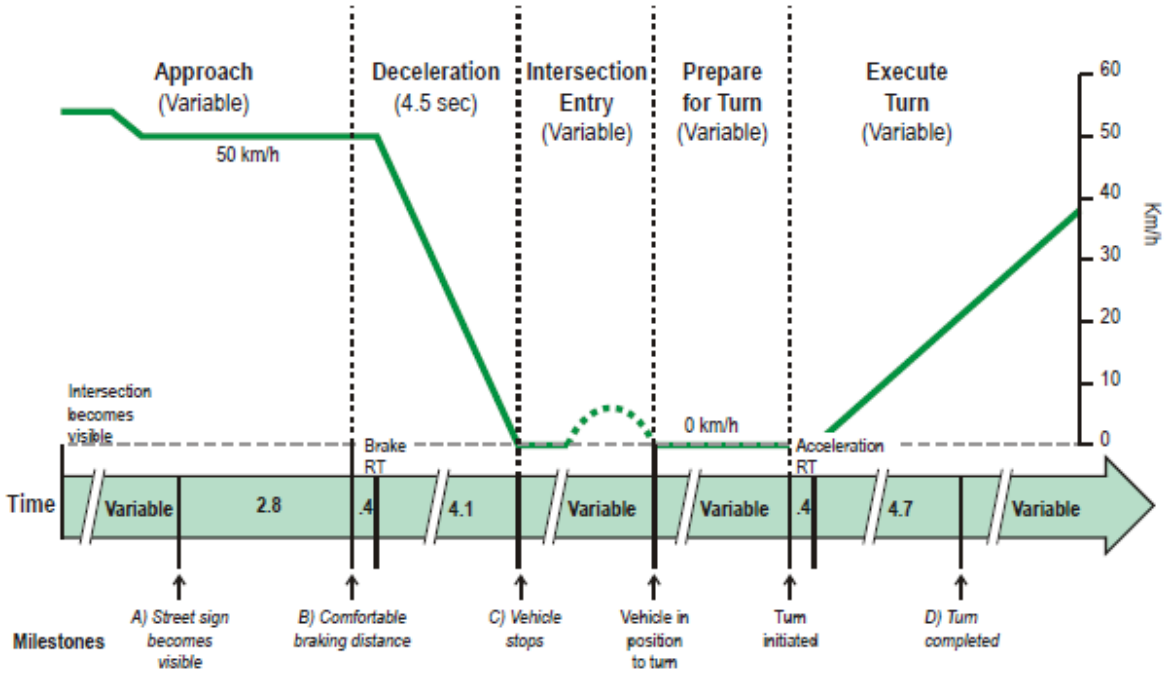
15-30 seconds of driving:

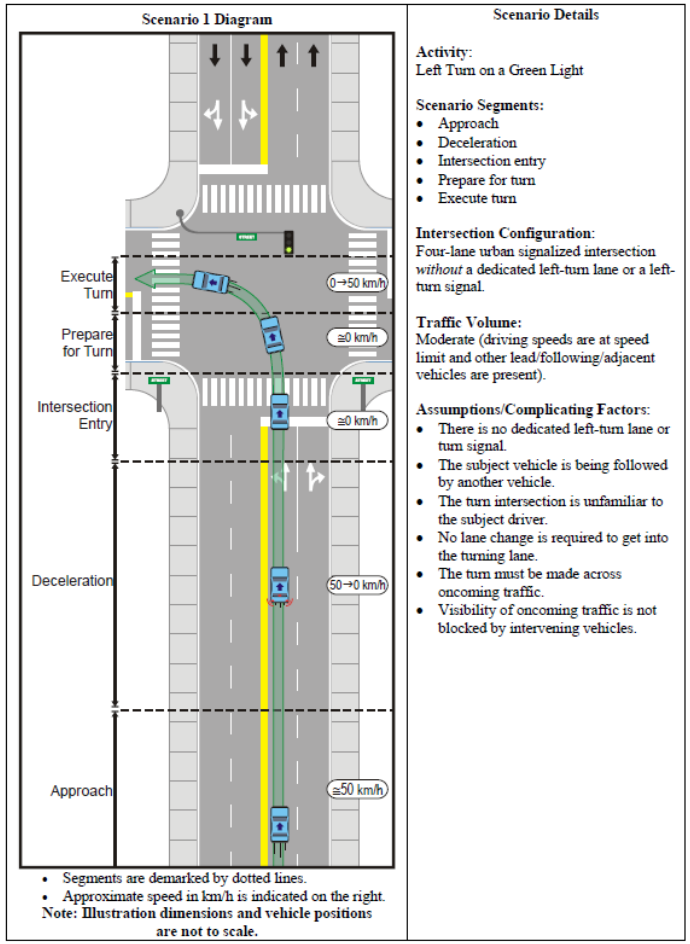
Context: Weather, Road Type, Traffic Conditions, Dynamic Vehicle Model, Sensors

Events: Route starts/ends, cut-ins, accidents, police intervention, **vehicle system failure**

Tasks: Turn left, onramp, exit, turn right, park

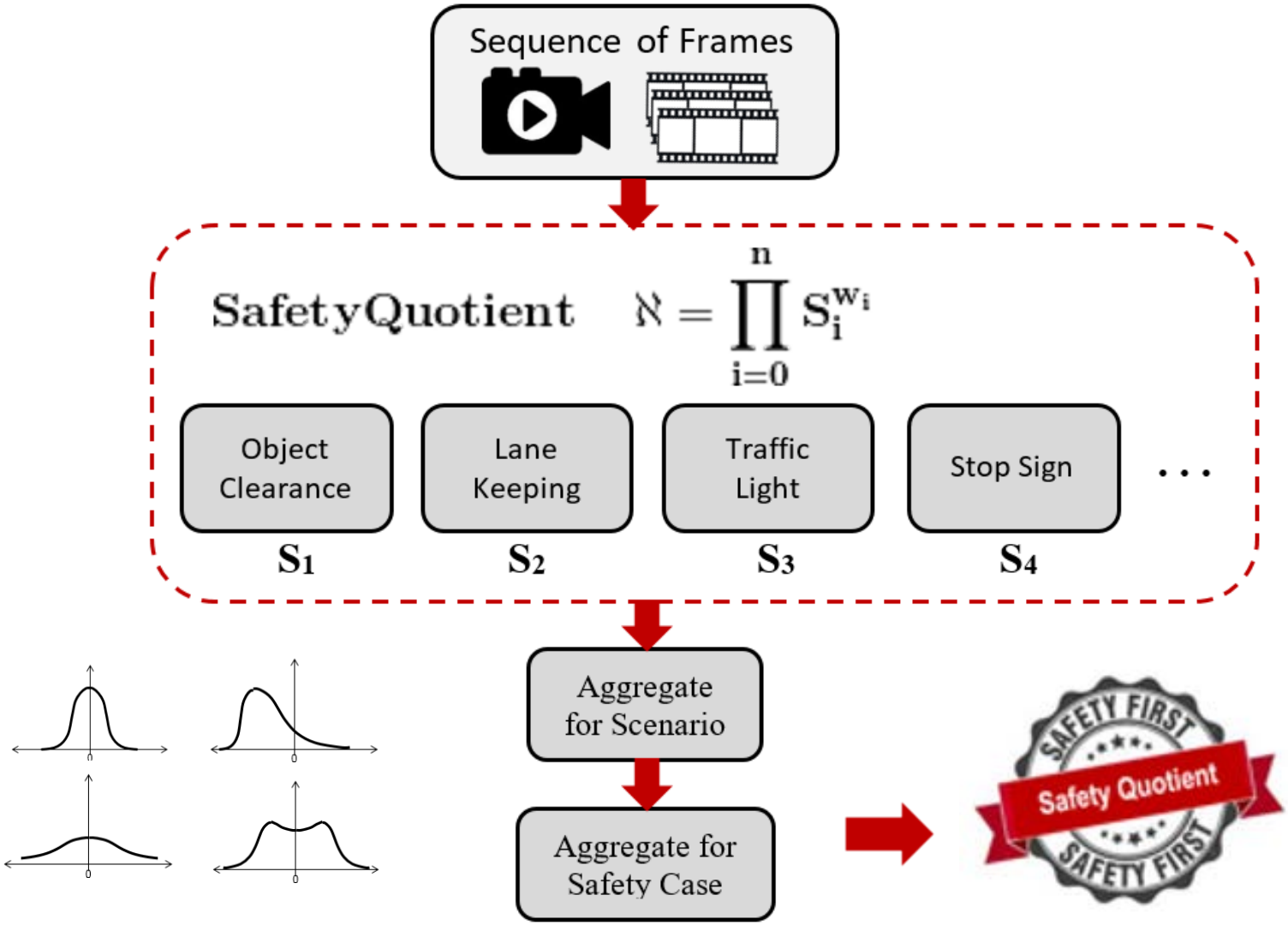
Metrics: Probability of Crashes, Rollovers, Near-misses
Degree of comfort (e.g., speed bumps)
Duration to completion (e.g., timid AI driver)





Compostable Safety Quotient

Quantifying Safety for A Sequence of Frames

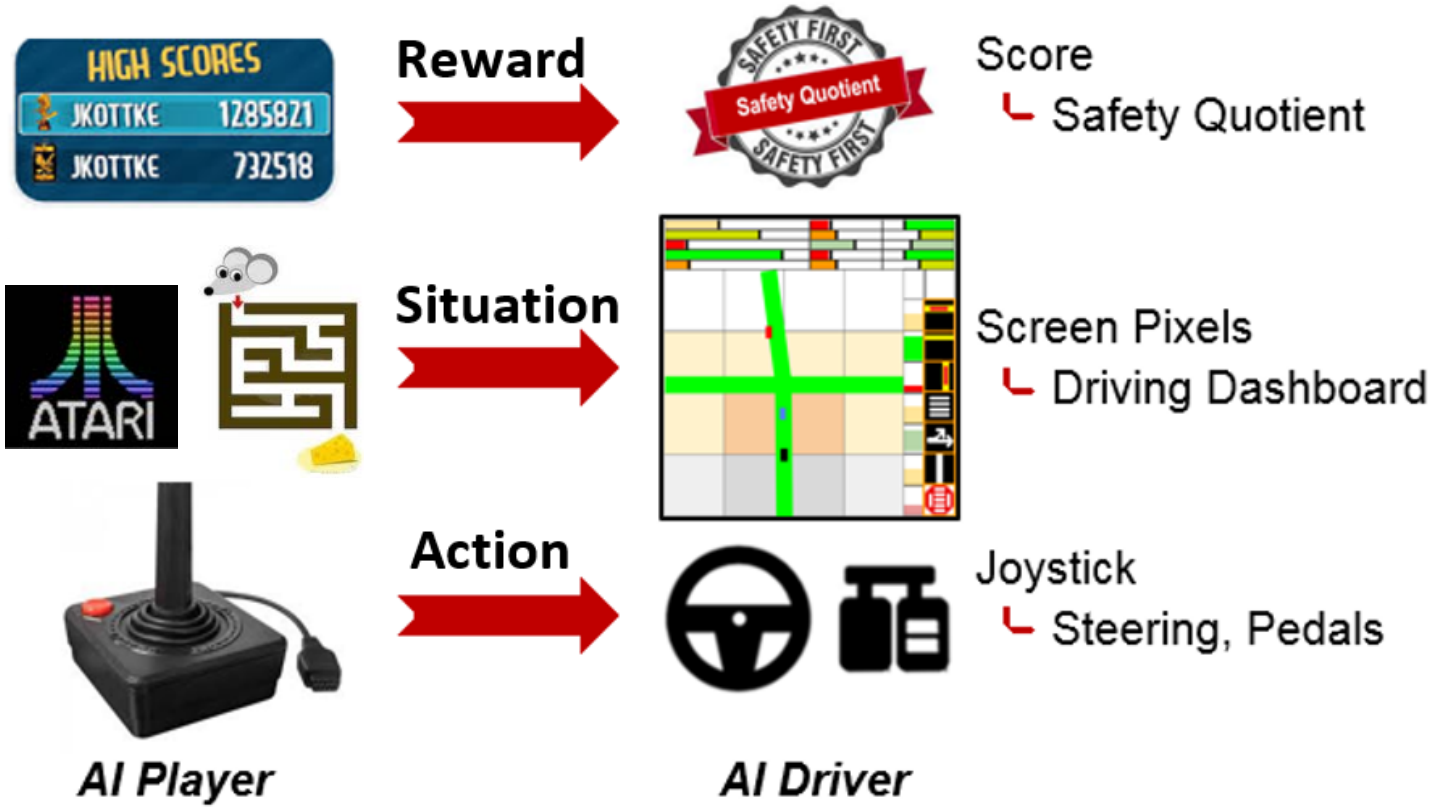


Beyond Copying Human Driving

From Superhuman Gamers to Superhuman Drivers

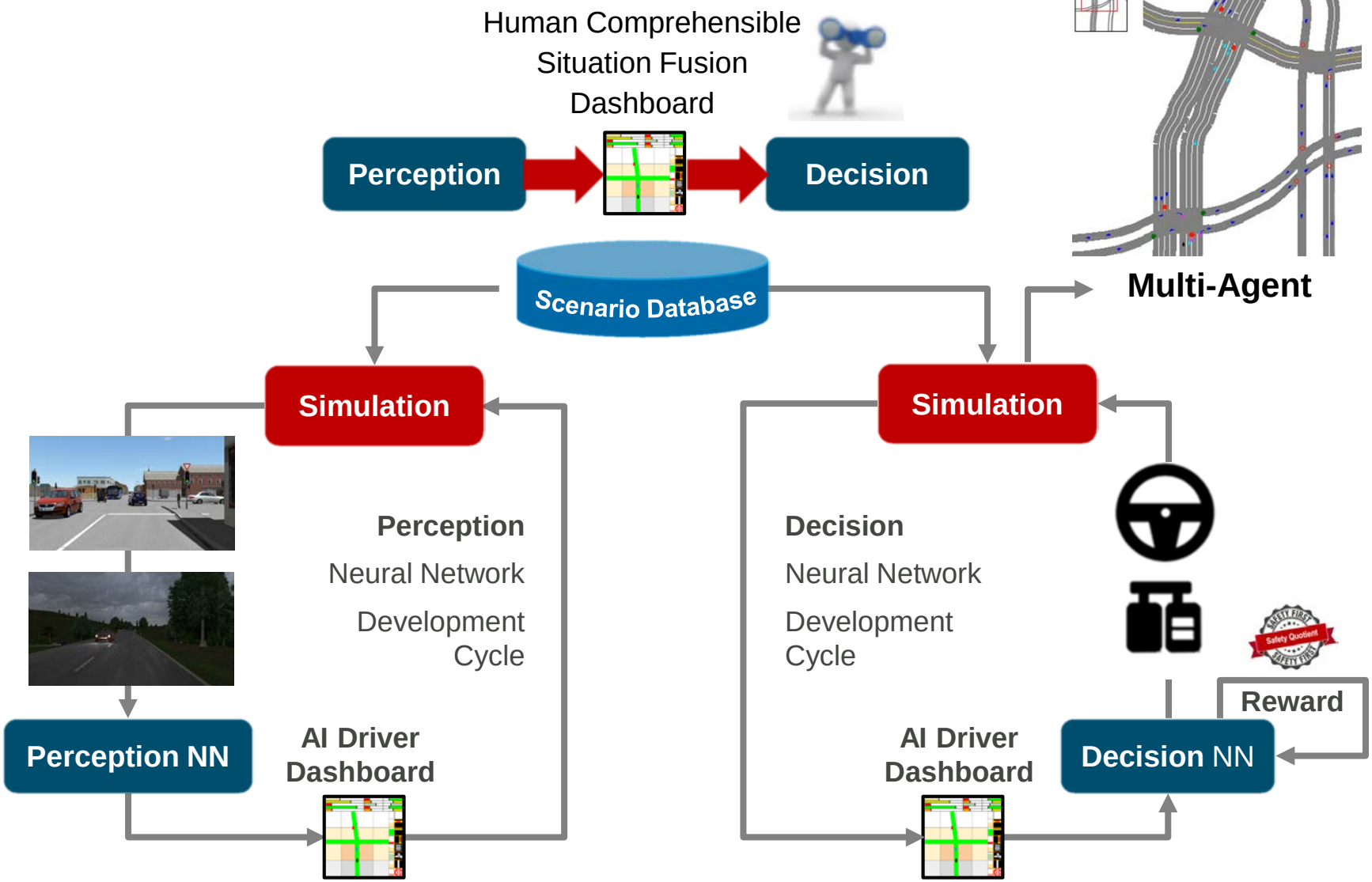
Strategy: Copying Humans does not improve on their capabilities.
Deep Reinforcement Learning *does*!

AlphaGo Zero: Tabula-Rasa RL achieved better than Human capabilities



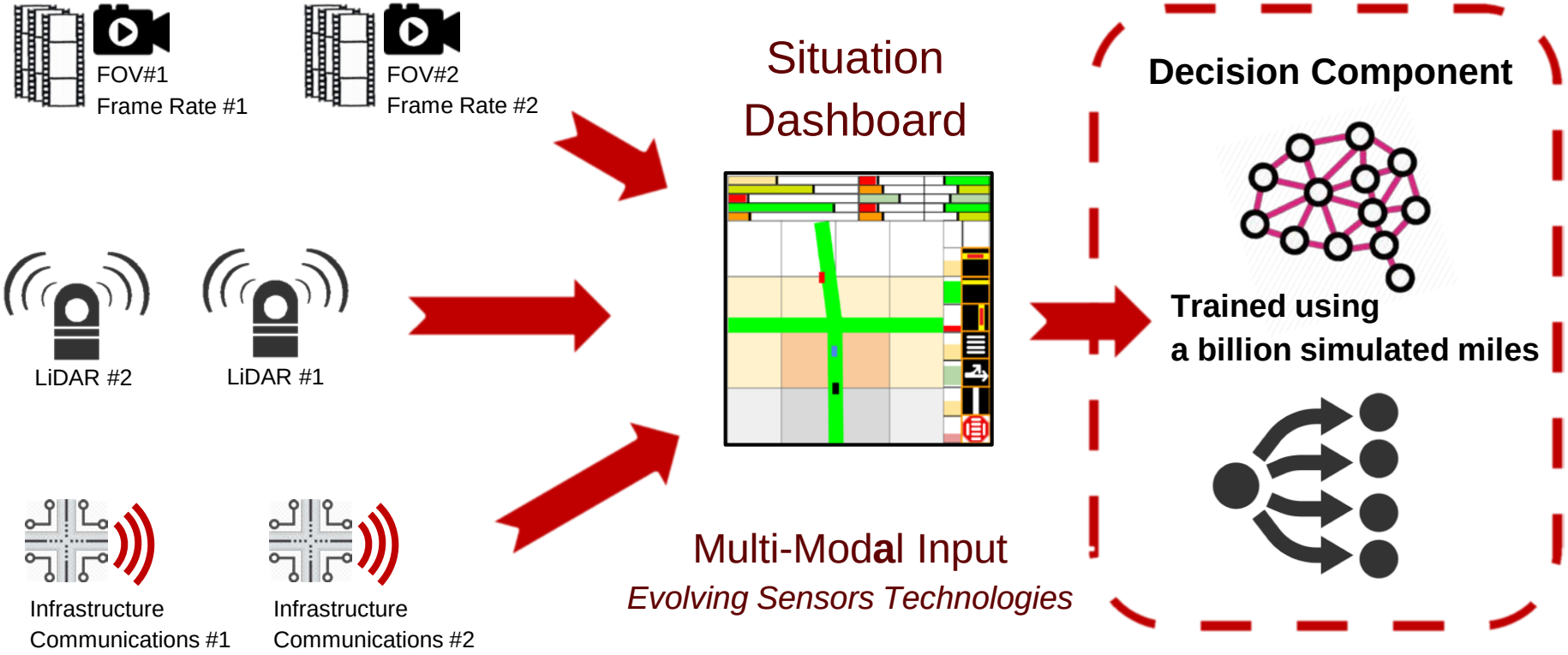
Anatomy of a testable AI Driver

Decouple Perception from Decision



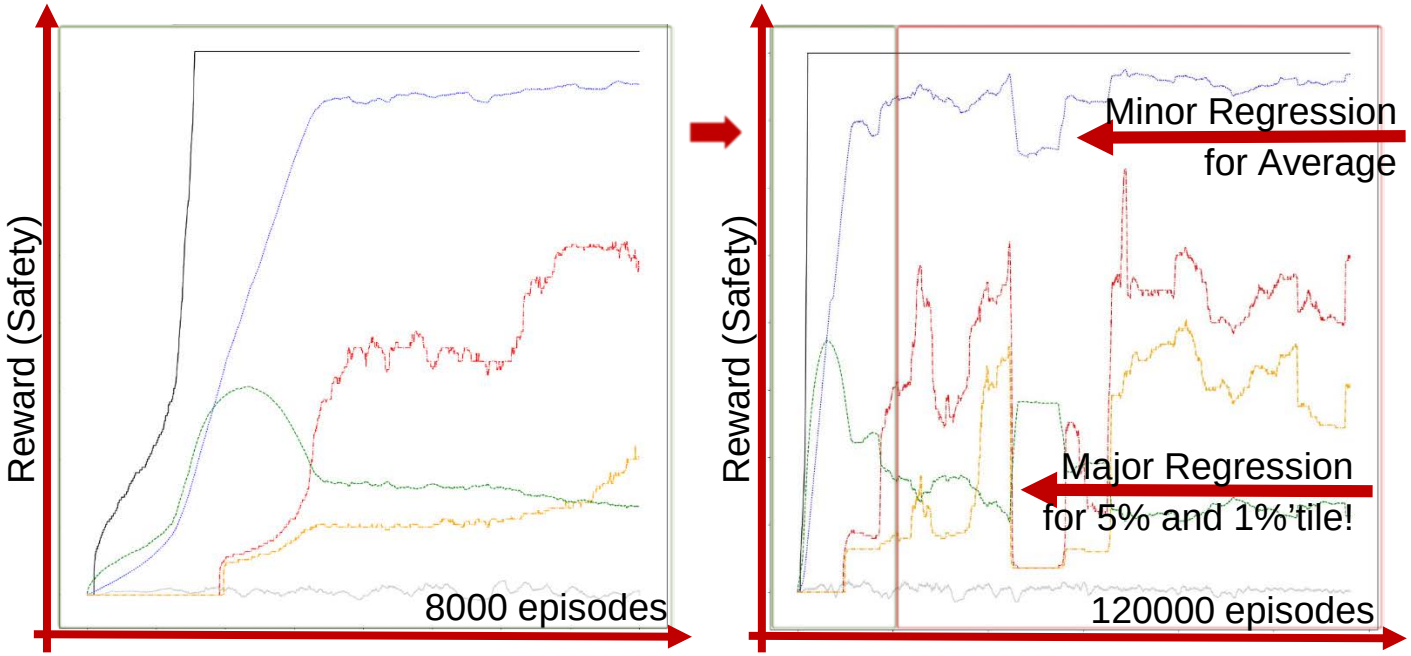
Preserve Skills Gained over Billions of Miles

As perception improves, must avoid retraining over billions of miles



Regression is Very Real

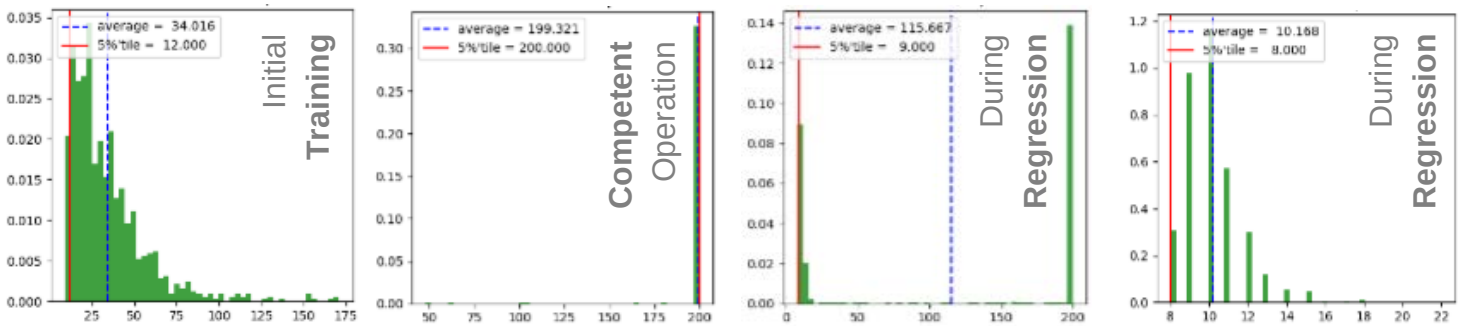
Confirmation bias can mask major risk



Regression for 1%'tile is much more significant than regression for averages.

Reward distributions during regression show clearly more than a single mode of operation.

- Max Reward
- DNN Loss
- Average Reward
- 5% of Reward
- 1% of Reward
- Variance



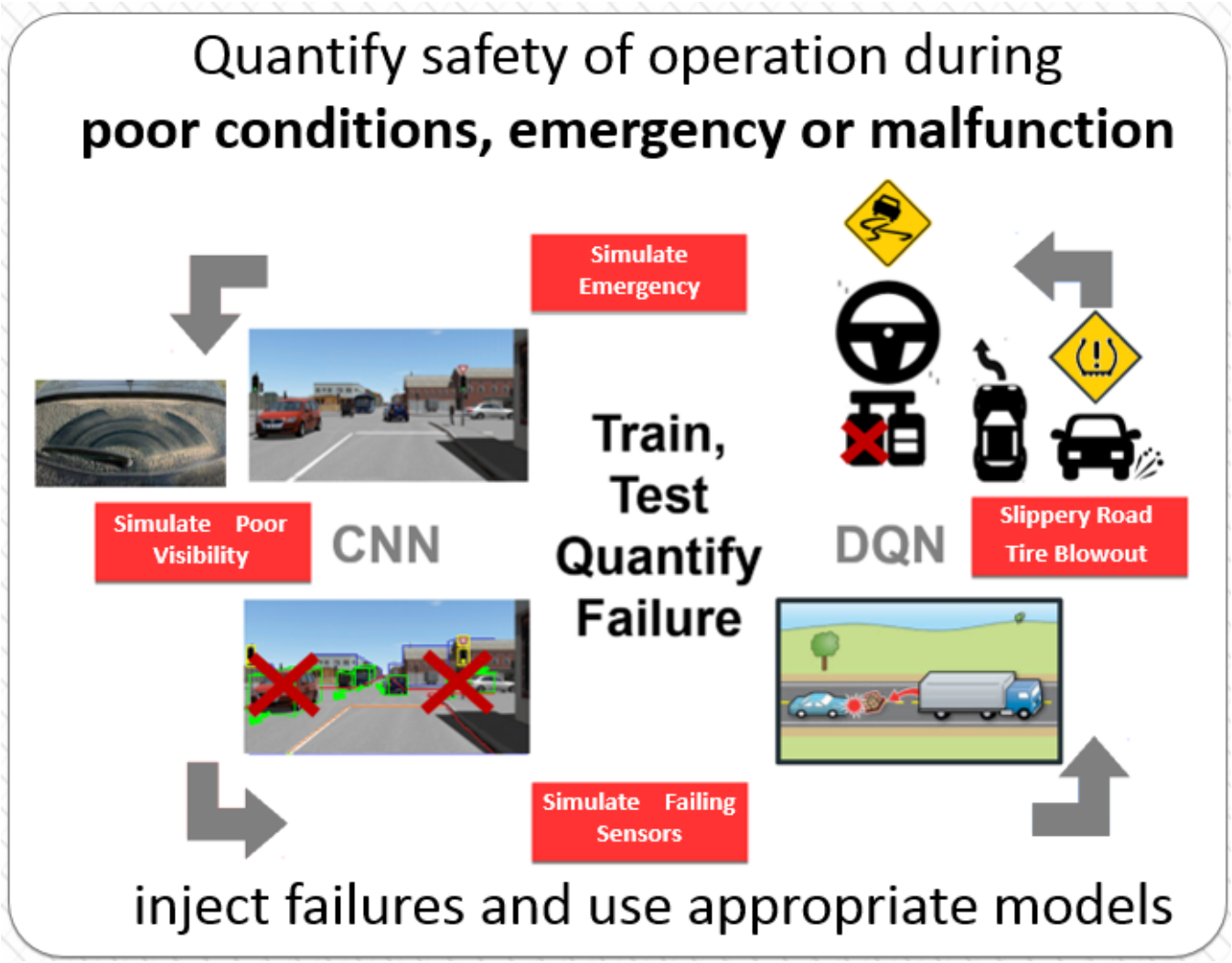
Episodes: 1-1000 43001-44000 44001-45000 45001-46000

How to Achieve and Quantify Operational Safety

Train and Test driving during Malfunctions and Emergencies

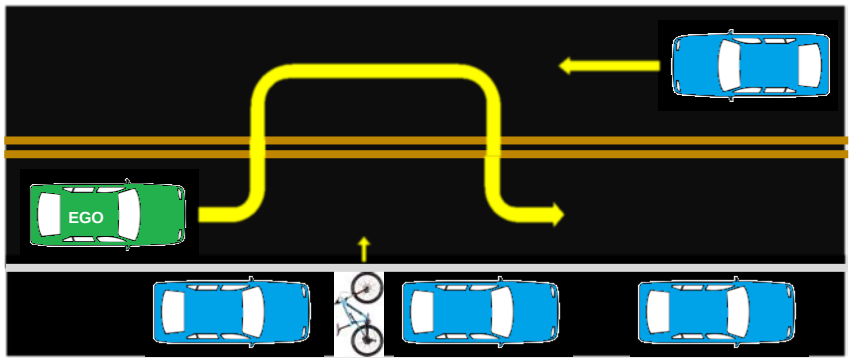
Failure and Emergency

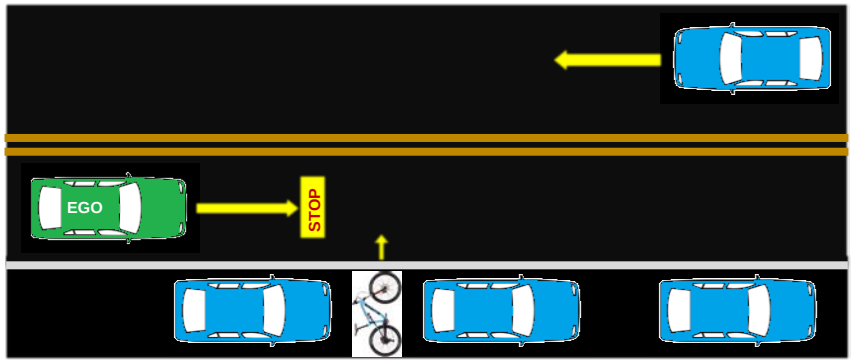
Safety Cases



Integration of Vehicle Dynamics

What is the best driving decision?





Steering
Angle, Velocity, and Torque

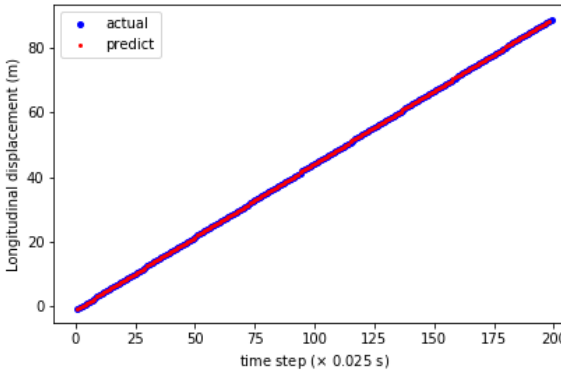
→

Acceleration
Lon., Lat., Ver., Roll, Pitch, Yaw

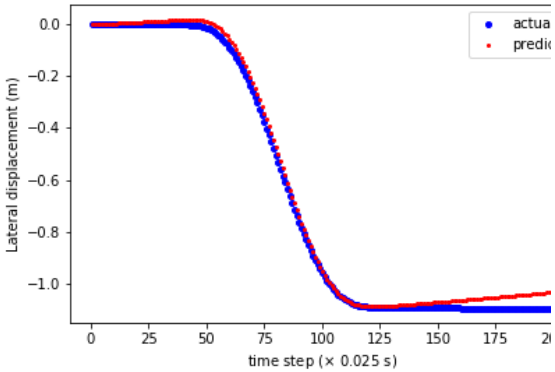
→

Displacement and Velocity
Lon., Lat., Ver., Roll, Pitch, Yaw

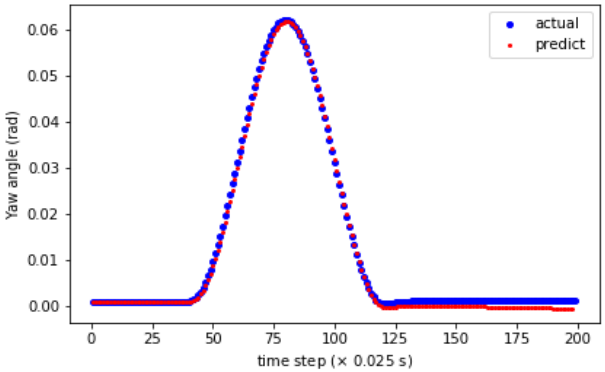
Longitudinal Displacement



Lateral Displacement



Yaw Angle



Testable Safe Approach

Quantify Distributions and Summary Statistics for Safety Quotient

Basic Approach	Safe Approach
Define Common Scenarios	Generate Scenarios
Record Sensor Data	Generate Environment
Create Labels	Train on Safety Quotient
Train	Summary Statistics
Road Test	Validate

Assemble libraries of *safety case scenarios*

and identify specific **tasks**

Assemble libraries of *environments*

and organize according to ***tasks***

Train models using Safety Quotient

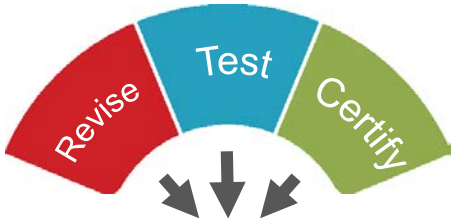
test on both naturalistic and synthetic data

Quantify confidence per scenario

evolve by improving on **tasks**

Quantify rate of failure and regression

distribution of Safety Quotient fir **tasks**



Tasks

Scenarios

Safety Cases

Compare **Human vs AI drivers** for ***Safety Cases Scenario and Task***

Backup Slides

Miles Needed for Testing
Achieving Statistically Significant Result

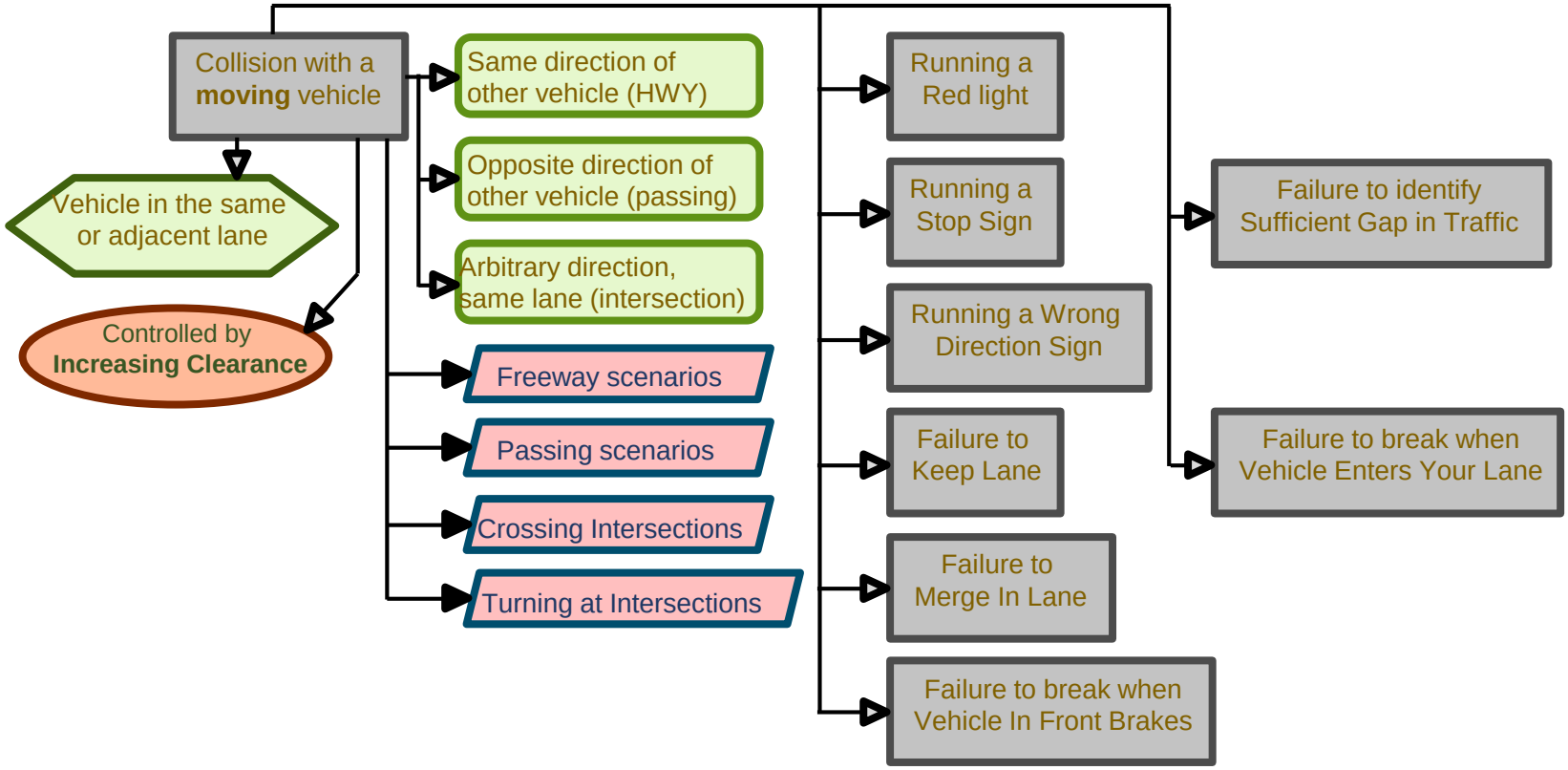
Statistical Question		Benchmark Failure Rate		
	How many miles (years[a]) would autonomous vehicles have to be driven...	**(A) 1.09 fatalities per 100 million miles?**	**(B) 77 reported injuries per 100 million miles?**	**(C) 190 reported crashes per 100 million miles?**
	(1) without failure to demonstrate with 95% confidence that their failure rate is at most...	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million miles (1 month)
	(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of...	8.8 billion miles (400 years)	125 million miles (5.7 years)	51 million miles (2.3 years)
	(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of...	11 billion miles (500 years)	161 million miles (7.3 years)	65 million miles (3 years)

[a] We assess the time it would take to compete the requisite miles with a fleet of 100 autonomous vehicles (larger than any known existing fleet) driving 24 hours a day, 365 days a year, at an average speed of 25 miles per hour.

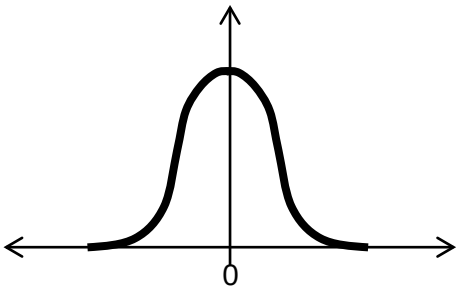
Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? N Kalra, and S Paddock, Santa Monica, Calif.: RAND Corporation, RR-1478-RC, Feb 2017.

Avoiding Collision in Intersection

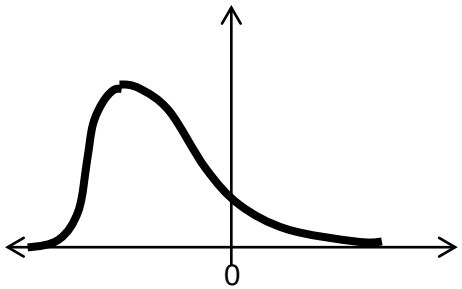
Safety Case Decomposition Example



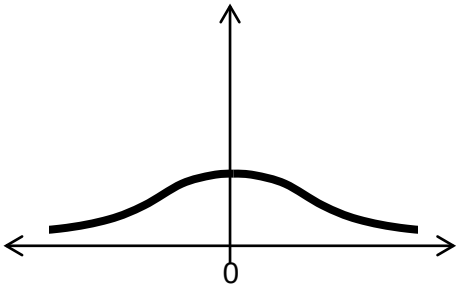
Safety Quotient Estimation
Expected Error Distribution



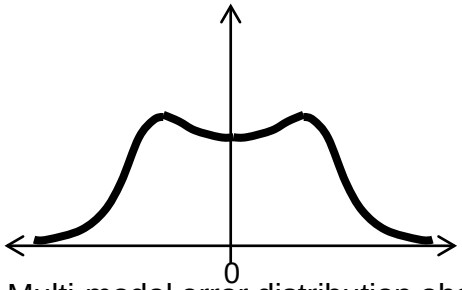
Target error distribution



Skewed distribution observed due to poor textures e.g., for poor visibility simulation

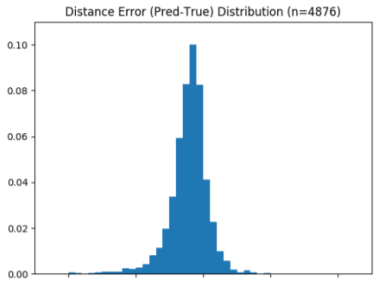


Flat wide distribution observed when testing on naturalistic models trained on synthetic

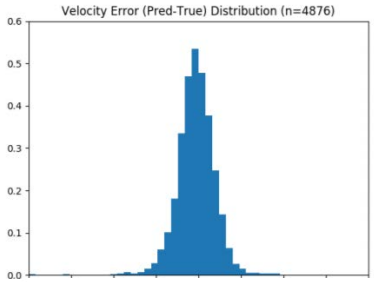


Multi-modal error distribution observed due to combination of overfitting of ensemble components and poor distribution of scenarios

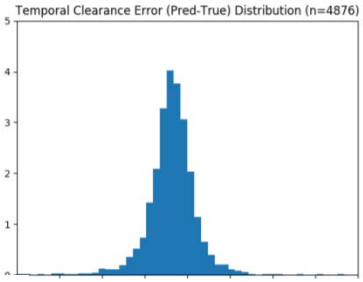
Safety Quotient Estimation: **Observed Error Distribution**



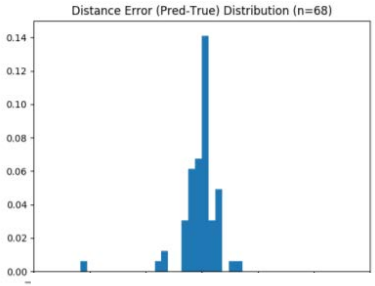
Distance Estimates Error Distribution for Synthetic



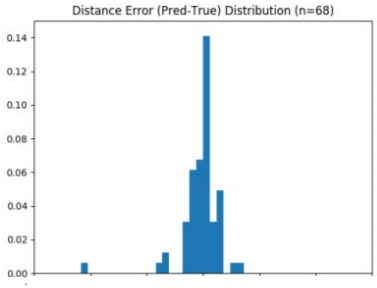
Velocity Estimates Error Distribution for Synthetic



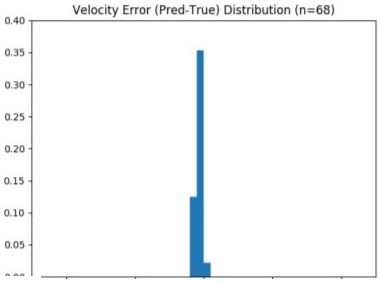
Temporal Clearance Estimates Error Distribution for Synthetic



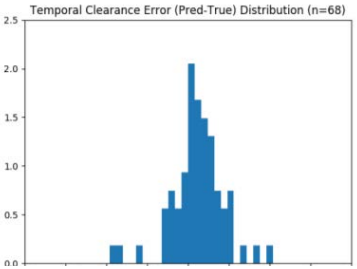
Distance Estimates Error Distribution for Naturalistic



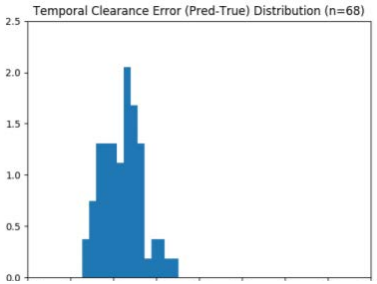
Velocity Estimates Error Distribution for Naturalistic



Temporal Clearance Estimates Error Distribution for Naturalistic

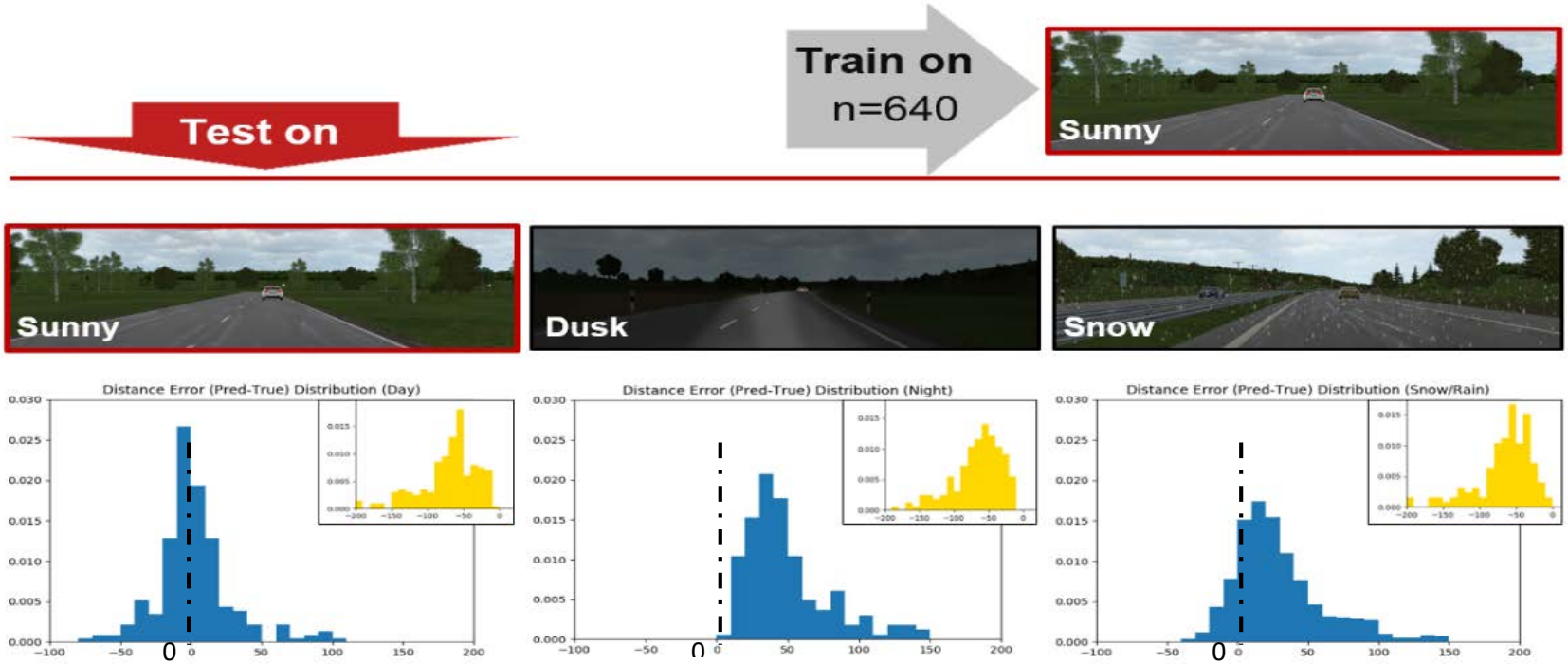


Temporal Clearance Trained Naturalistic, Tested on Reference Naturalistic



Temporal Clearance Trained Synthetic, Tested on Reference Naturalistic

Safety Quotient Estimation
Observed Error Distribution



Training and Testing on Sunny Weather

Training on Sunny Testing on Dusk

Training on Sunny Testing on Snow