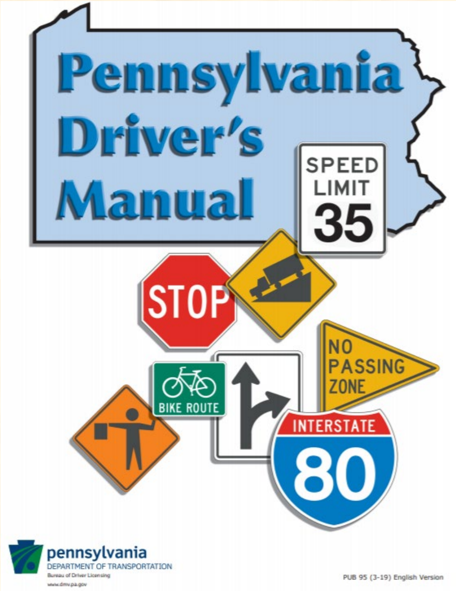
Validating Machine Learning-Based Systems

Why Not Use a Traditional Driver Test?

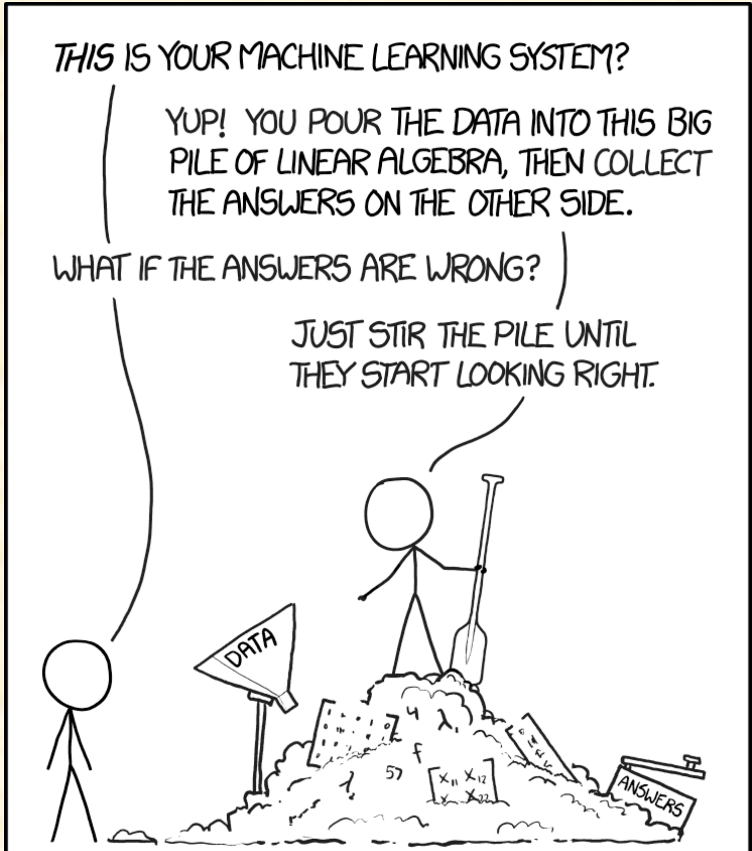
Carnegie Mellon University



- Written test
  - Does ADS know traffic laws?
  - Does ADS know behavioral expectations?
- Road test
  - Can ADS execute traffic laws?
  - Can ADS negotiate effectively with human drivers?
  - Does ADS exhibit good driver hygiene?
  - Can ADS resolve potentially ambiguous driving situations?
- Being a 16 year old human
  - How do we measure ADS judgment maturity?
  - Does the ADS know when it doesn't know what to do?

Machine Learning Challenges

- Inductive learning
  - Collect lots of training data
  - Adjust learned model; iterate
  - Declare success when tests pass
- Fundamental challenges:
  - Assurance on novel inputs
    - What if it over-fitted to data?
    - Gaps in training data
  - Did it learn what you hoped?
    - Prone to "gaming" the learning
  - What was actually learned?



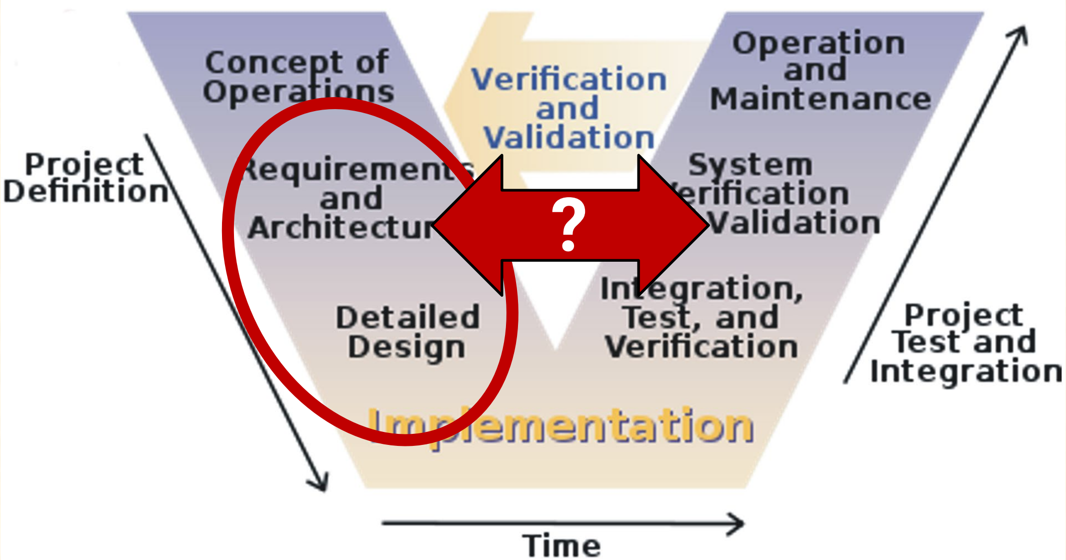
https://xkcd.com/1838/

Traditional Validation Vs. Machine Learning

- 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



Early Testing: Public Road Testing

Carnegie Mellon University

- Good for identifying "easy" cases
  - Expensive and potentially <u>dangerous for closed loop testing</u>

Validation Via Brute Force Road Testing?

- If 200M miles/critical mishap…
  - Test 3x–10x longer than mishap rate
    ➔ Need 2 Billion miles of testing

- That's ~50 round trips on every road in the world
  - With fewer than 10 critical mishaps

- And what if the answer is: "not safe enough; try again?"

WolframAlpha computational knowledge engine.

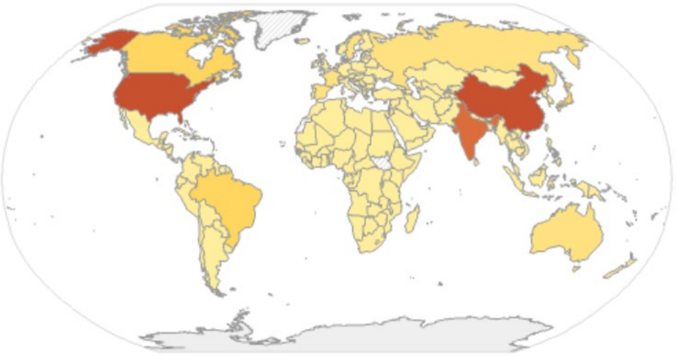
miles of roads

Summary:

total	20.46 million mi
median	11 630 mi
highest	4.03 million mi (United States)
lowest	4.97 mi (Tuvalu)

(1994 to 2008)
(based on 225 values; 24 unavailable)

Total road length map:



Closed Course Testing

Carnegie Mellon University

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

Volvo / Motor Trend

Simulation

- Highly scalable; fidelity vs. cost tradeoff
  - Need to build highly detailed models
  - Challenge of matching real world data into simulation models
  - Only tests things you have thought of!

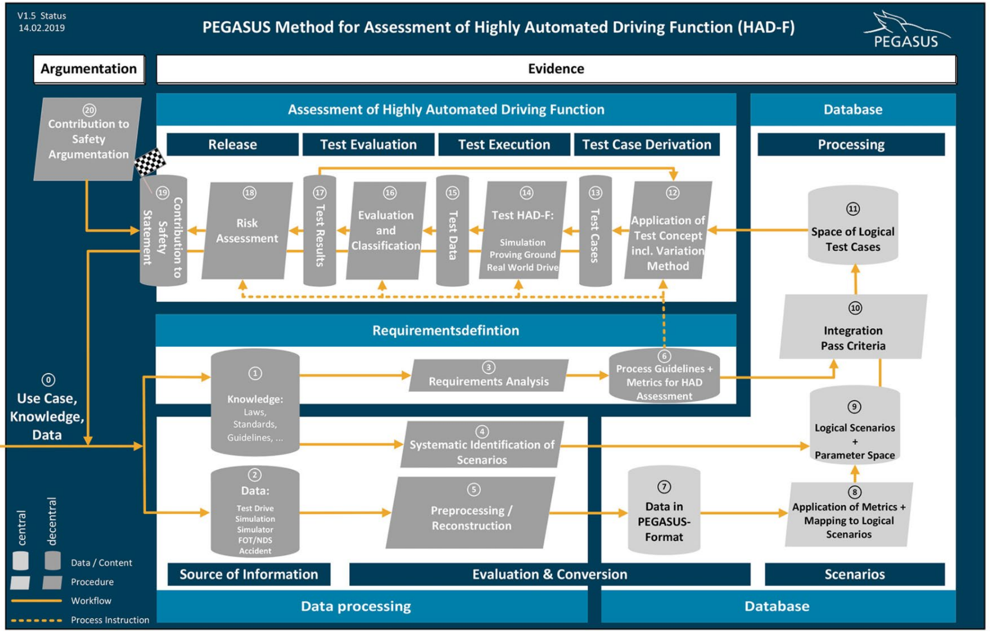
ANSYS

Scenario-Based Simulation

- Scenarios must cover Operational Design Domain (ODD)

THE PEGASUS METHOD

https://www.pegasusprojekt.de/en/pegasus-method



NHTSA-inspired pre-crash scenarios

We have selected 10 traffic scenarios from the NHTSA pre-crash typology to inject challenging driving situations into traffic patterns encountered by autonomous driving agents during the challenge.

Traffic Scenario 01: Control loss without previous action

- **Definition**: Ego-vehicle loses control due to bad conditions on the road and it must recover, coming back to its original lane.

Traffic Scenario 02: Longitudinal control after leading vehicle's brake

- **Definition**: Leading vehicle decelerates suddenly due to an obstacle and ego-vehicle must react, performing an emergency brake or an avoidance maneuver.

Traffic Scenario 03: Obstacle avoidance without prior action

- **Definition**: The ego-vehicle encounters an obstacle / unexpected entity on the road and must perform an emergency brake or an avoidance maneuver.

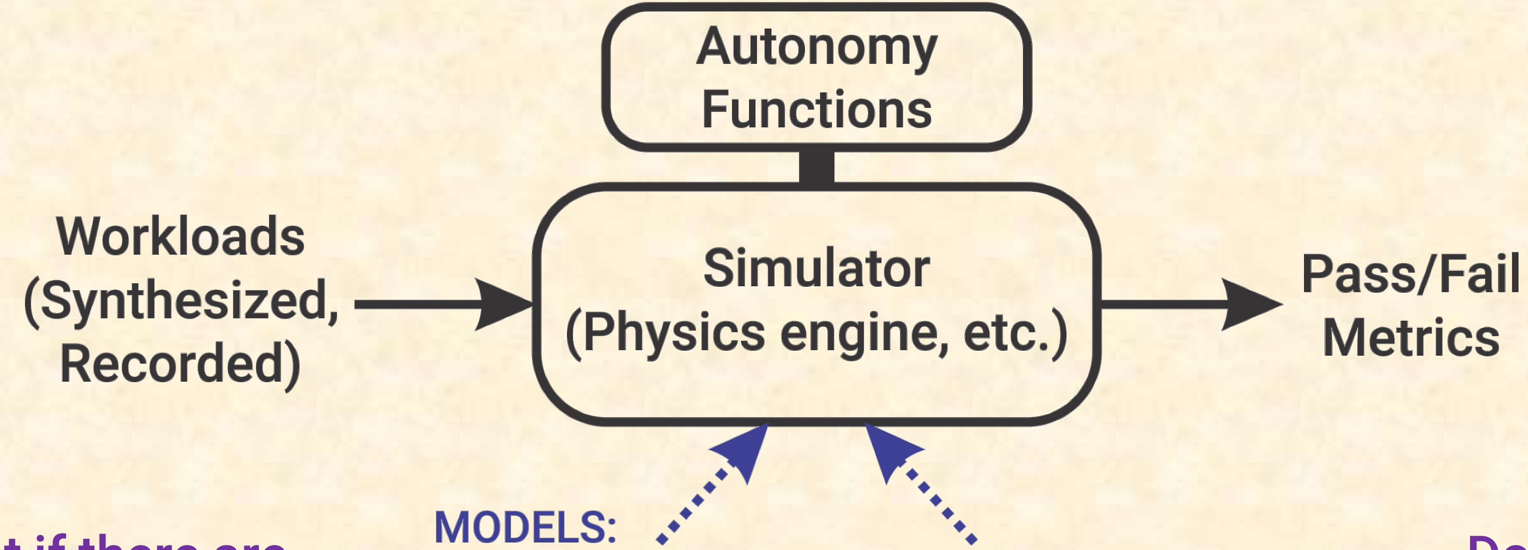
Traffic Scenario 04: Obstacle avoidance with prior action

- **Definition**: While performing a maneuver, the ego-vehicle finds an obstacle / unexpected entity on the road and must perform an emergency brake or an avoidance maneuver.

https://carlachallenge.org/challenge/nhtsa/

Simulation Components

Carnegie Mellon University



MODELS:
- Roads, weather, ...
- Sensors
- Vehicle dynamics
- Other road users
- Faults/failures
- What "safe" means ...

Design of Experiments

What if there are simulation software/model defects?
Will somebody die?

Does your simulation approach include perception?

Simulation Validity

- Fidelity & qualification
  - Environment; road users
  - Perception as well as vehicle motion
  - Appropriate safety metrics
  - Tool & model qualification

CARLA https://youtu.be/2c-KIQ8SFcc

- Experimental design
  - Coverage of ODD & high-risk edge cases
  - Matching simulated scenario to real-world scenario
  - Experimental design for validation of simulation itself

"All models are wrong, but some are useful." – George Box

What Does It Mean for a Test To Pass?

Carnegie Mellon University

- Traditional test paradigm:
  - You think design is right
  - Test validates engineering done properly
    - Test traces to requirements/design
- Inductive training test paradigm:
  - You think system was trained properly
  - Test determines whether training worked
    - Weak traceability to test set, if any
    - Hope to detect training data gaps, overfitting
  - BUT: nondeterministic, opaque "design"
    - Robust test plan is essential

https://goo.gl/cFCknY

https://goo.gl/QdTYVV

Changing Validation Approaches

- Machine Learning (ML) breaks the "V"
- Simulation validity (including models & test plan)
- Are you simulating perception (the hardest part)?

Carnegie Mellon University

