Safety of Autonomous Systems

Stuart Reid PhD, FBCS

(stureid.test@gmail.com / www.stureid.info)

Scope of the Talk

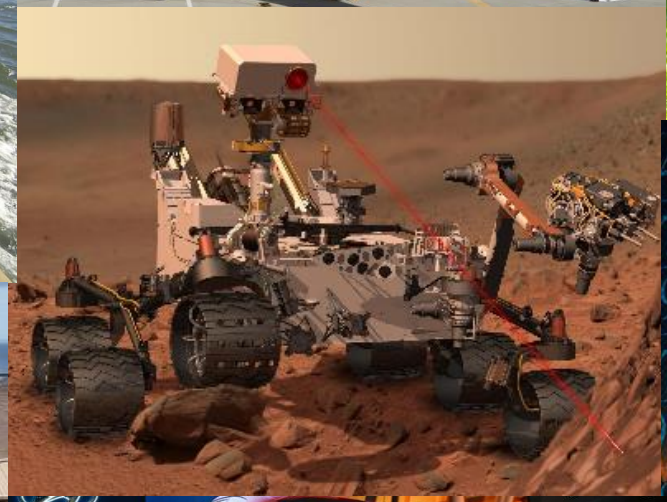
- Introduction to Autonomous Systems
- Specifying Objectives (Safely)
- Online vs Off-Line Machine Learning
- Machine Learning Challenges
- Black Box Testing
- White Box Testing
- The Necessity of Virtual Test Environments
- Conclusions

Introduction to Autonomous Systems

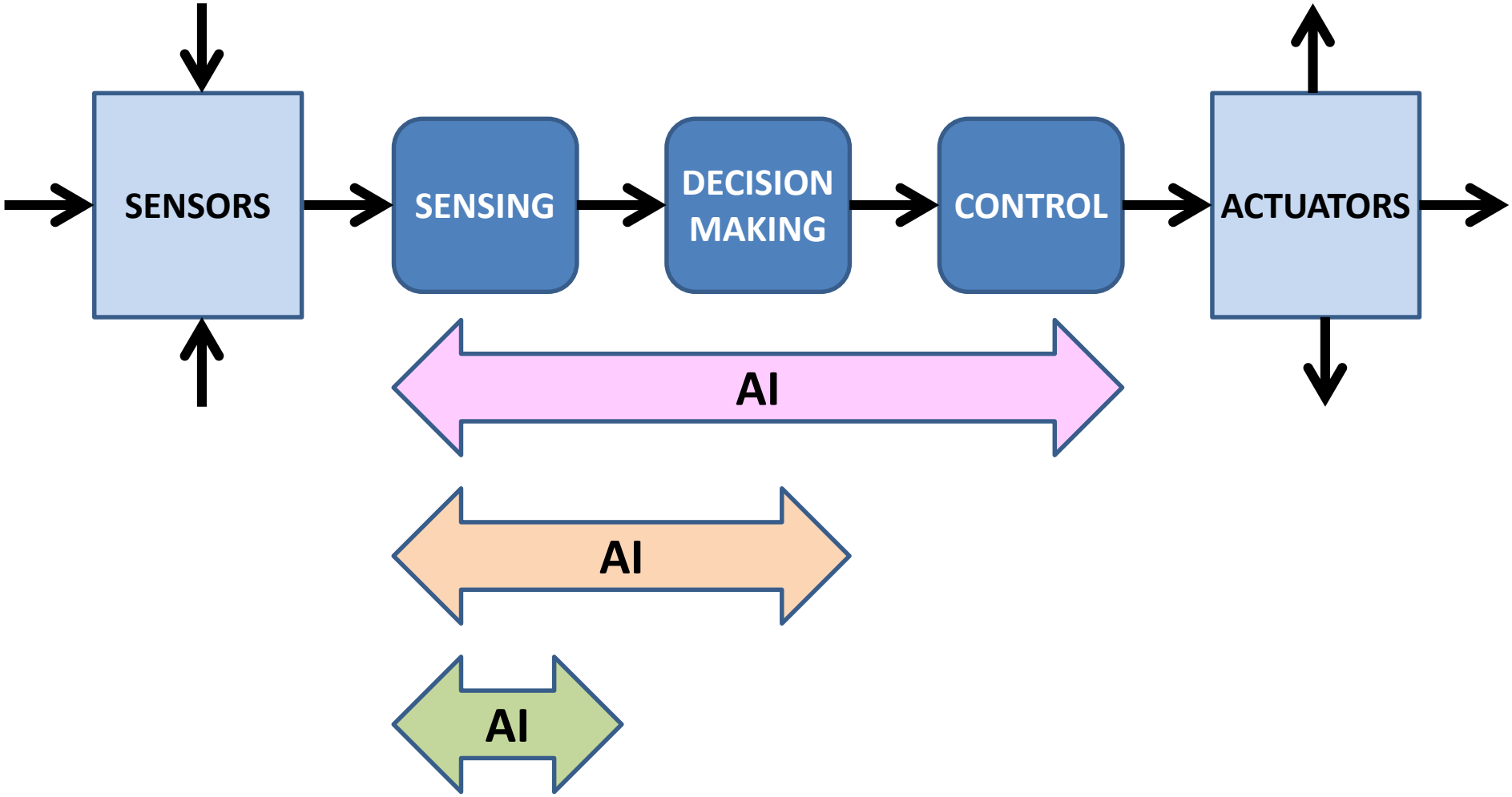
Definition used for Autonomous System

- **Autonomy**
  - the capacity to make an informed, un-coerced decision. Autonomous organizations or institutions are independent or self-governing
  - the ability to act independently of direct human control and in unrehearsed conditions
- **Autonomous System**
  - system that changes its behaviour based on its experiences and the current situation to achieve given objectives without human control

Examples: Autonomous Systems

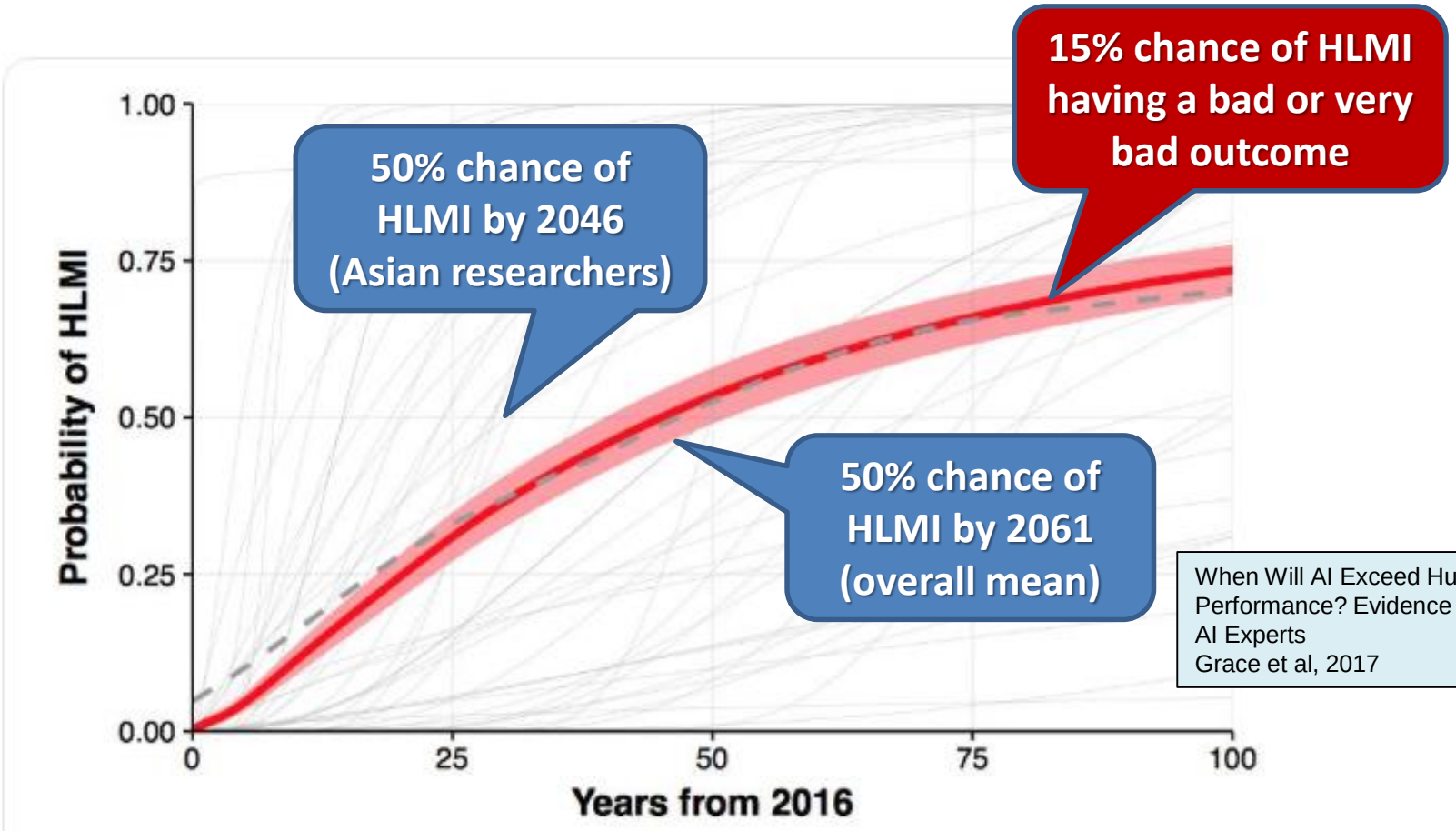


Basic Autonomous System Framework



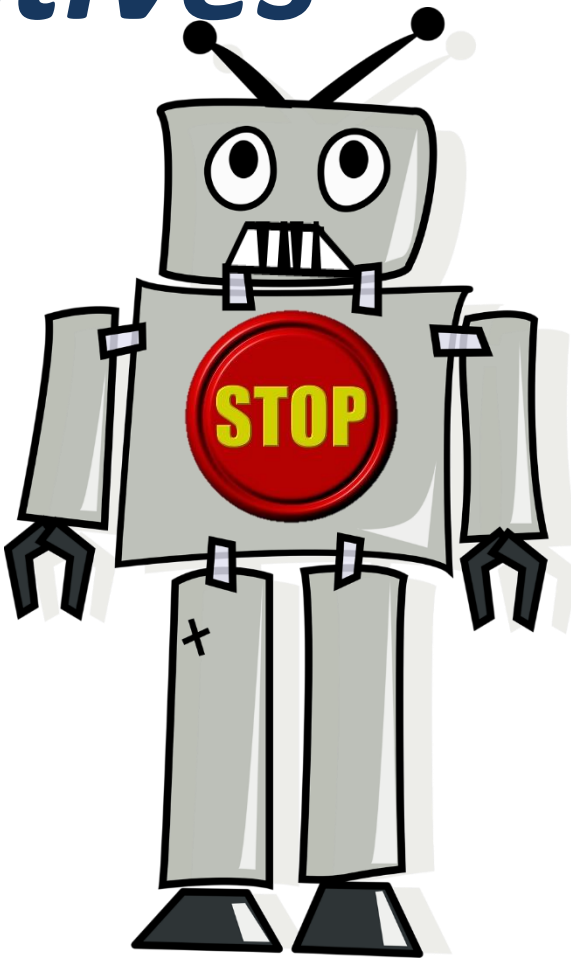
High-Level Machine Intelligence
- as predicted by published AI researchers

"High-level machine intelligence" (HLMI) is achieved when unaided machines can accomplish every task better and more cheaply than human workers.



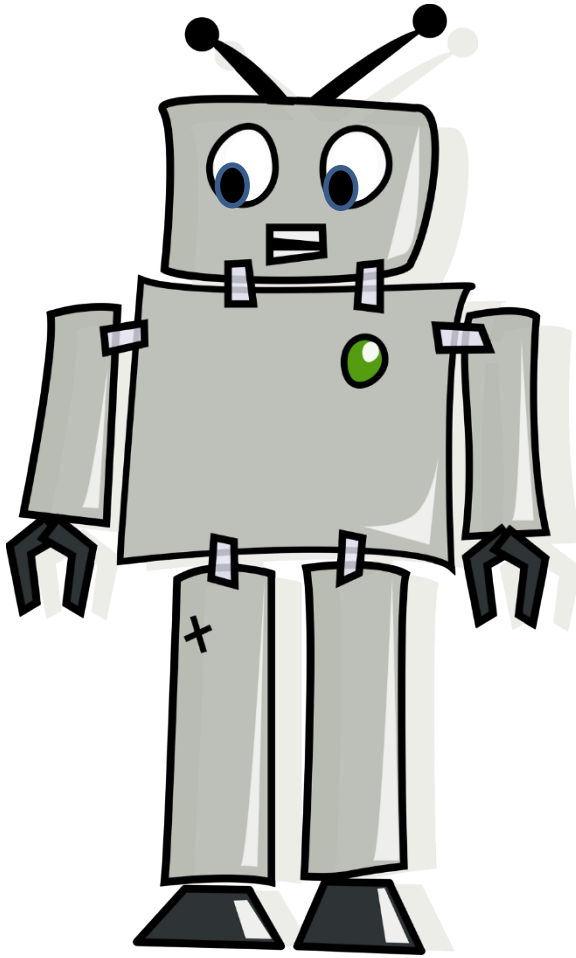
When Will AI Exceed Human Performance? Evidence from AI Experts
Grace et al, 2017

Specifying Objectives (Safely)

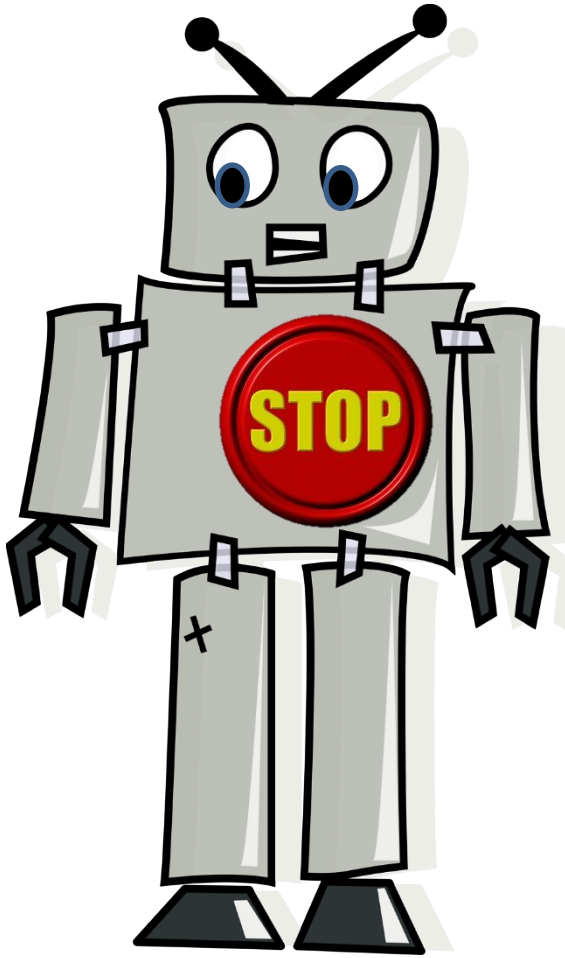


The Midas Problem “마이더스의 손”

“I’m hungry! Make me dinner”



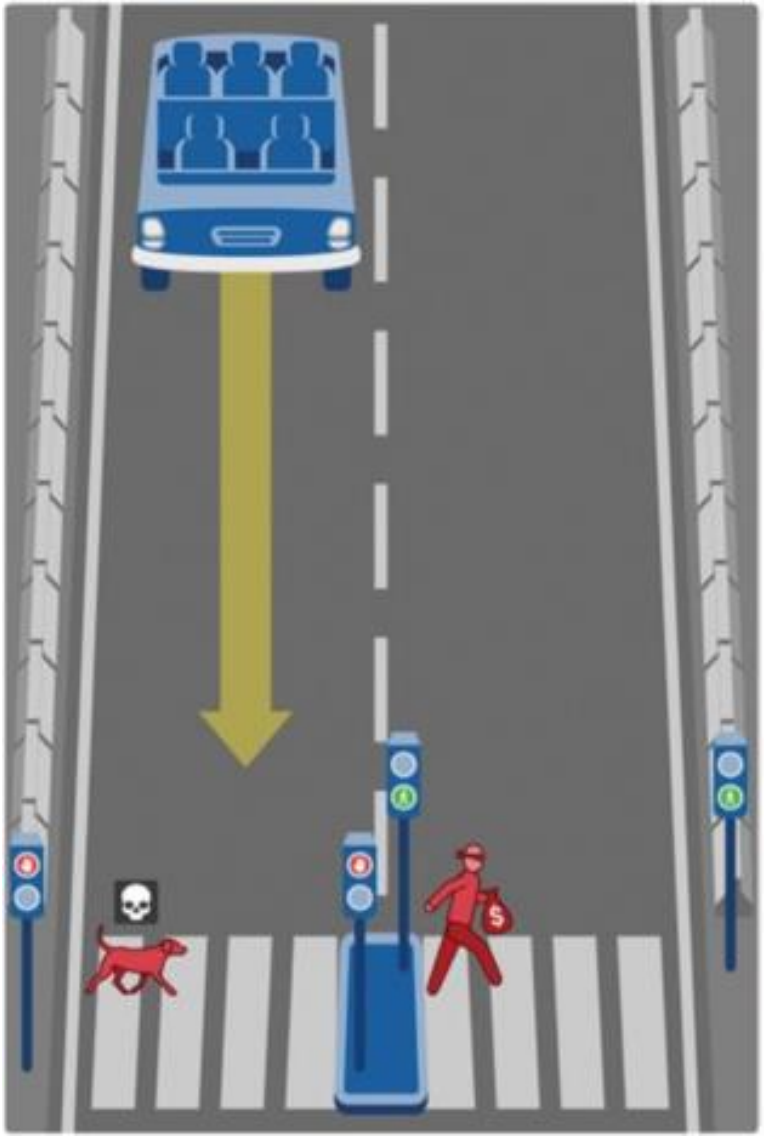
“Keep the kitchen clean”

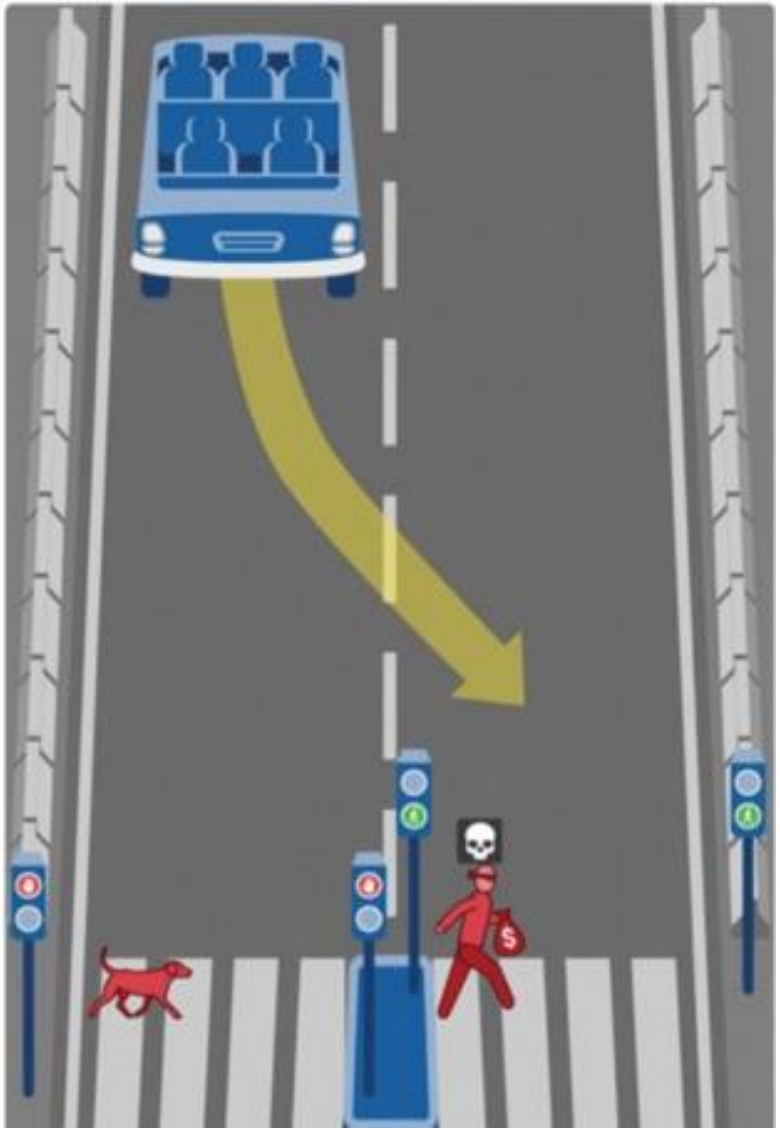


Side-Effects, Reward Hacking and Role Models

- **Reinforcement learning involves the system being rewarded for achieving objectives**
  - must be aware of side-effects
  - however problems can arise with ‘reward hacking’ when the system ‘hacks’ the objectives
- **Instead, we can get systems to learn from human demonstrations**
  - and get feedback from humans
- **BUT**
  - make sure the humans are representative
  - recognize that human values change over time
  - humans aren’t always the best role models…

MIT's Moral Machine (moralmachine.mit.edu)

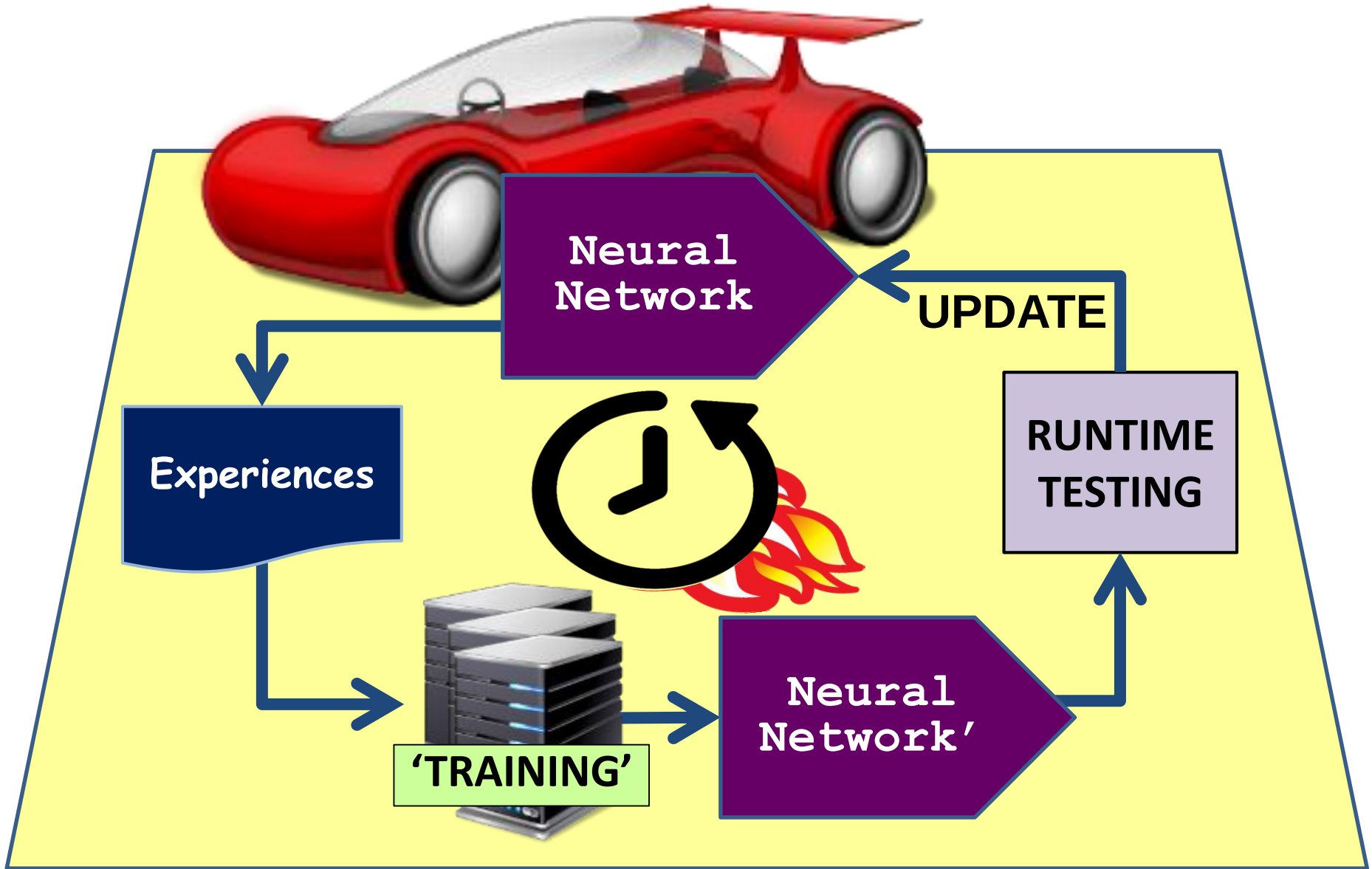




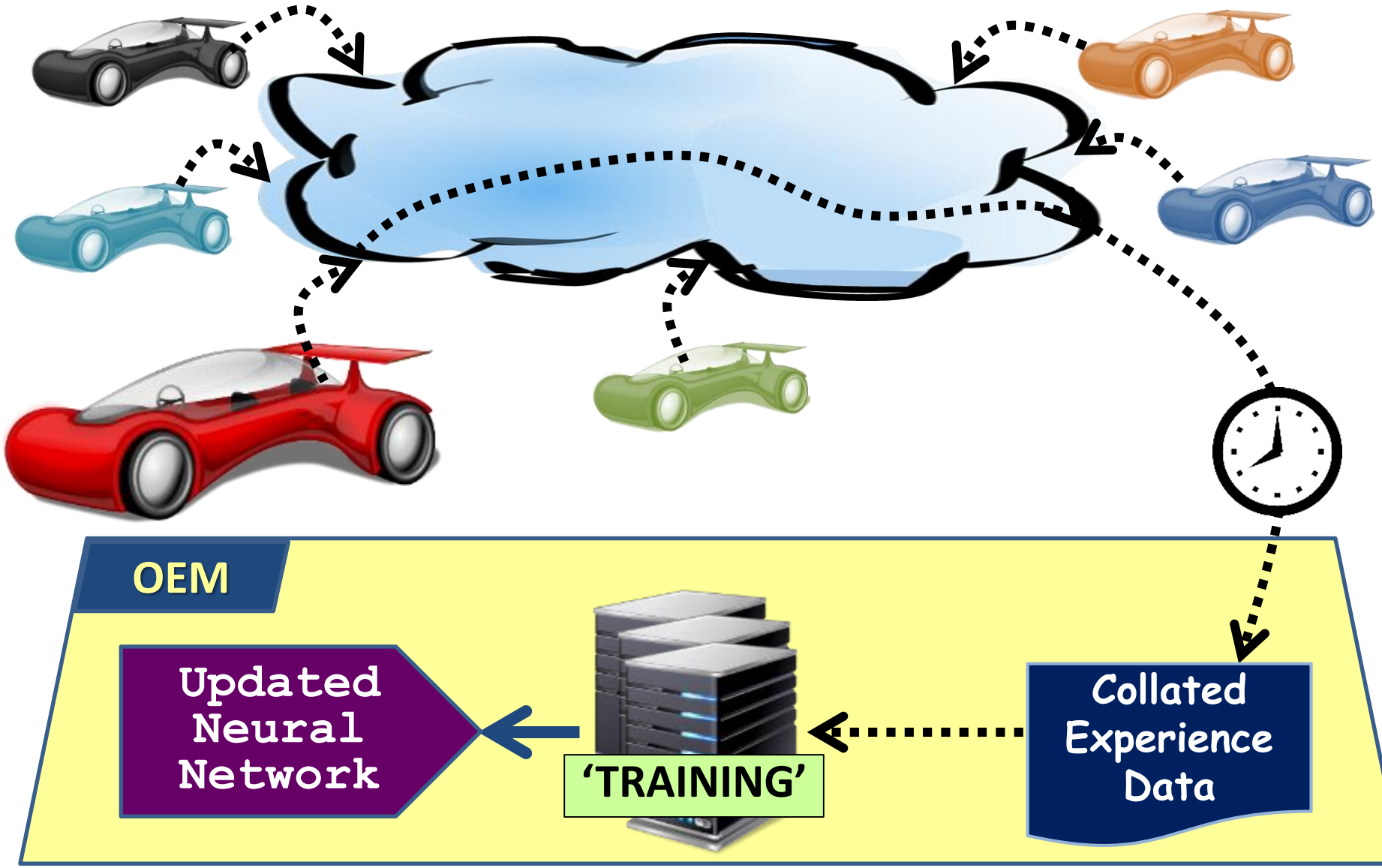
Better than Humans?

Online vs Off-Line Machine Learning

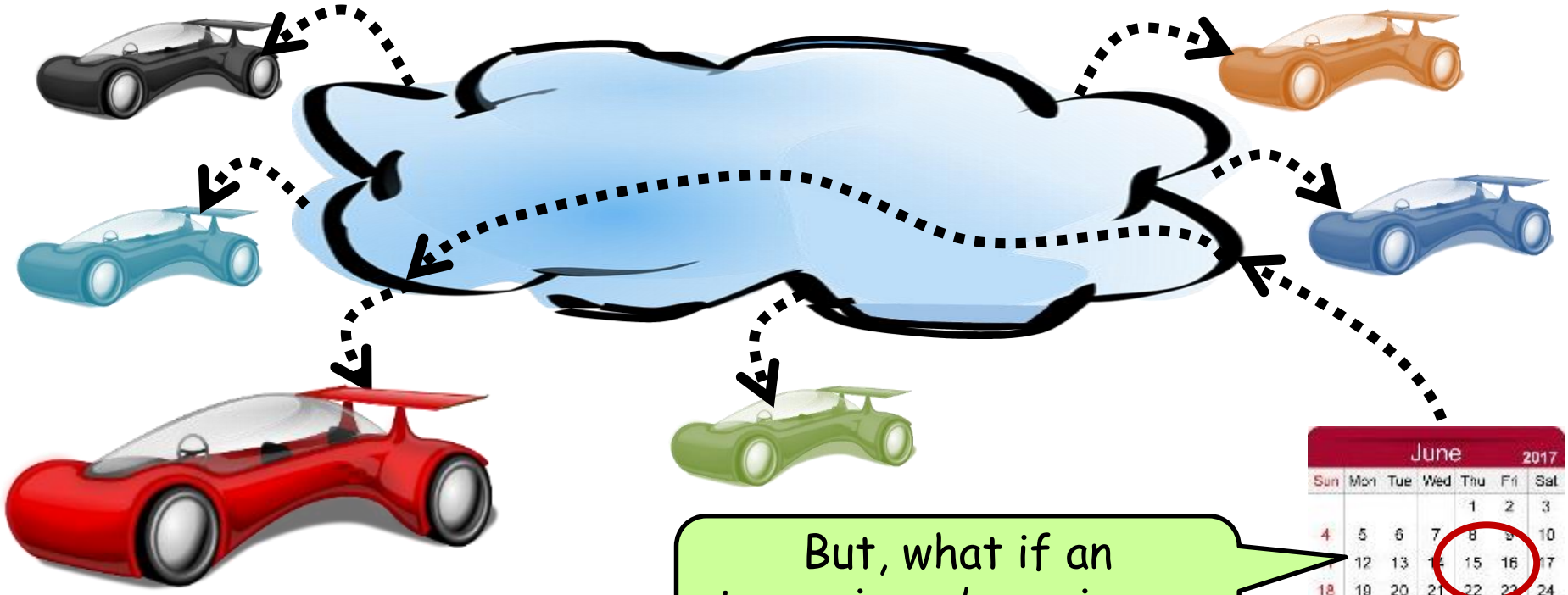
Continuous Online Learning



Off-Line Learning – from Day-to-Day Use



Performance Updates - Over-The-Air



But, what if an 'experience' requires a quick fix?

OEM

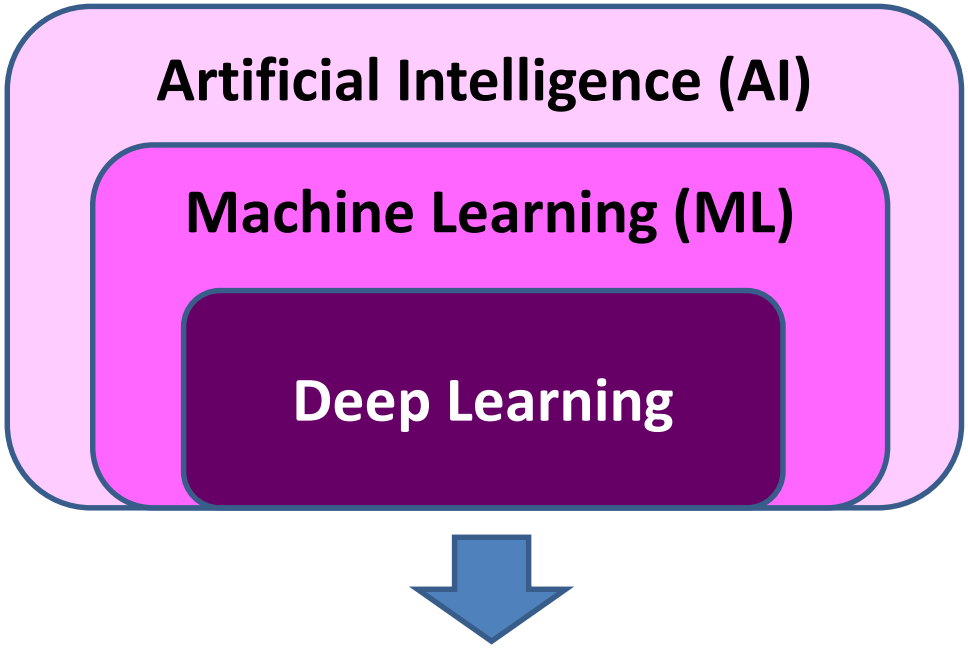
Updated Neural Network

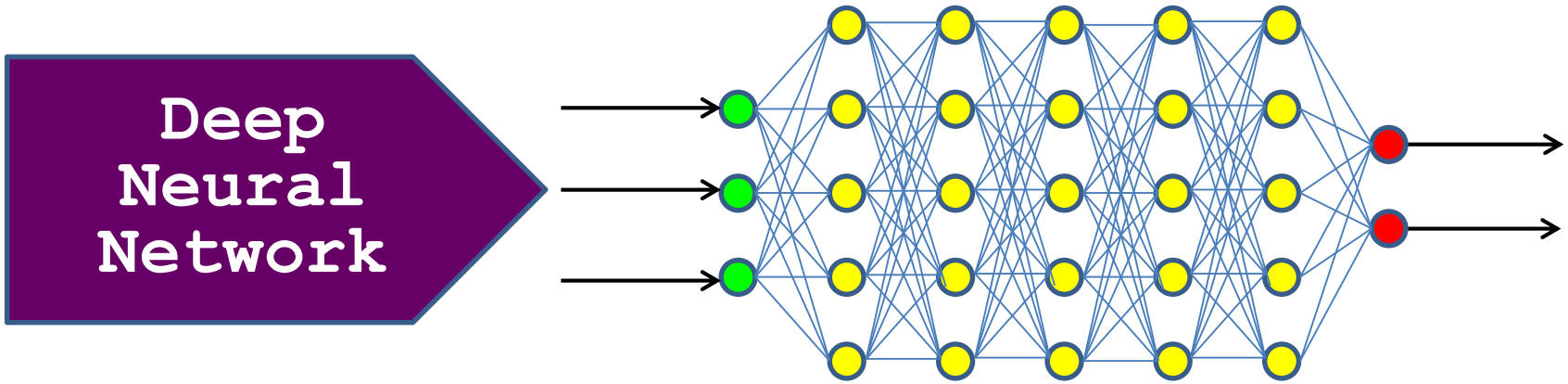
OFFLINE TESTING

UPDATE

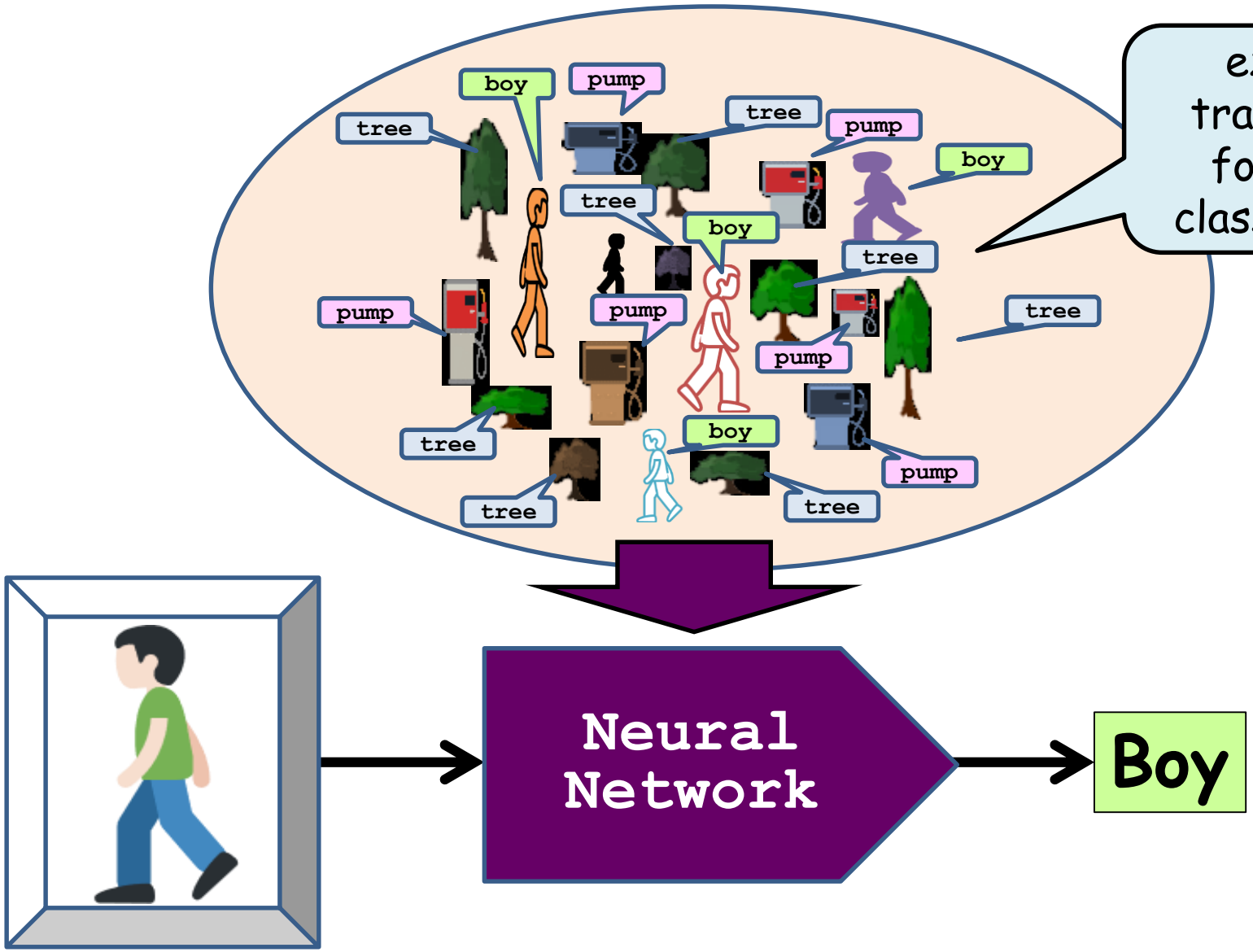
Machine Learning Challenges

Deep Learning Systems

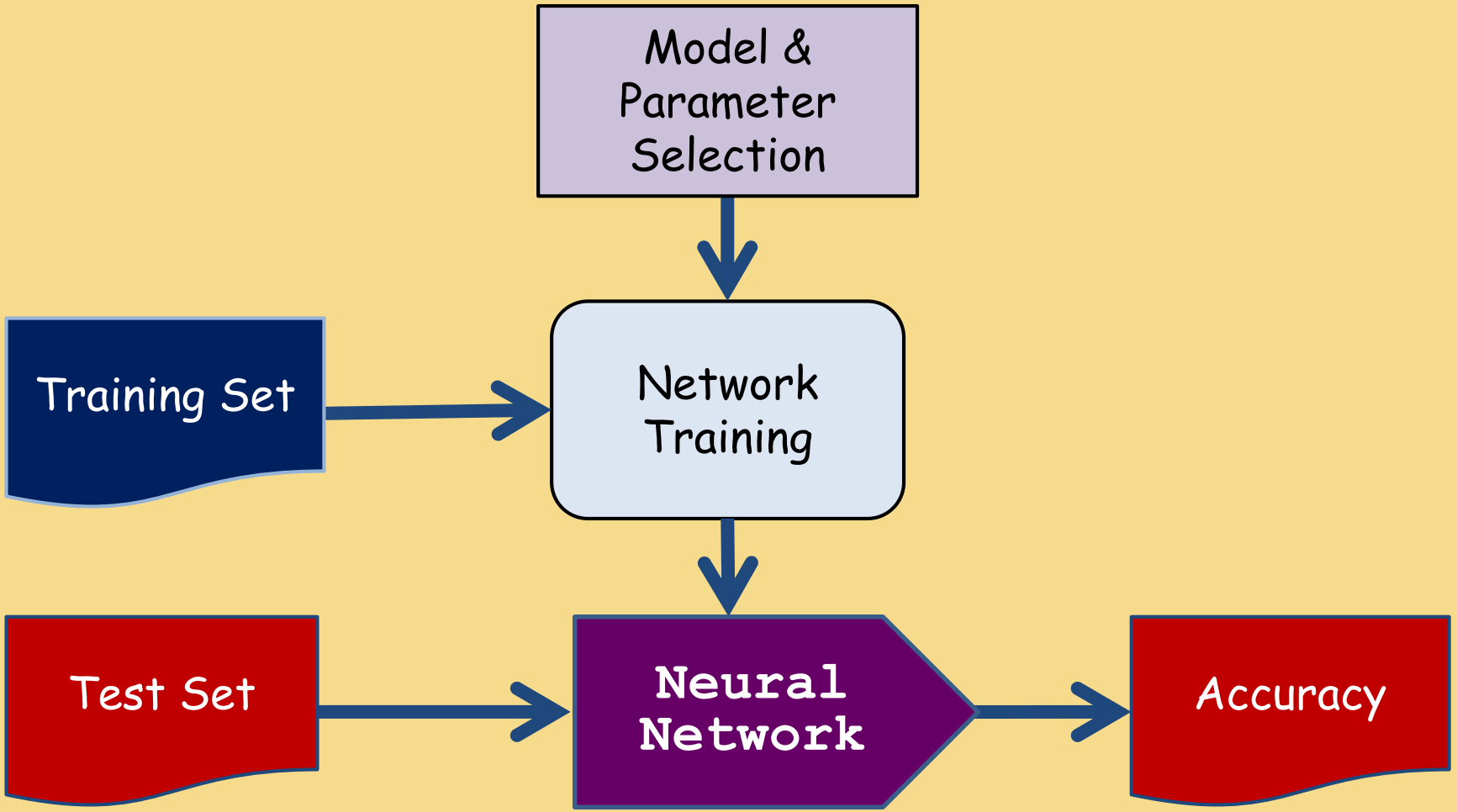




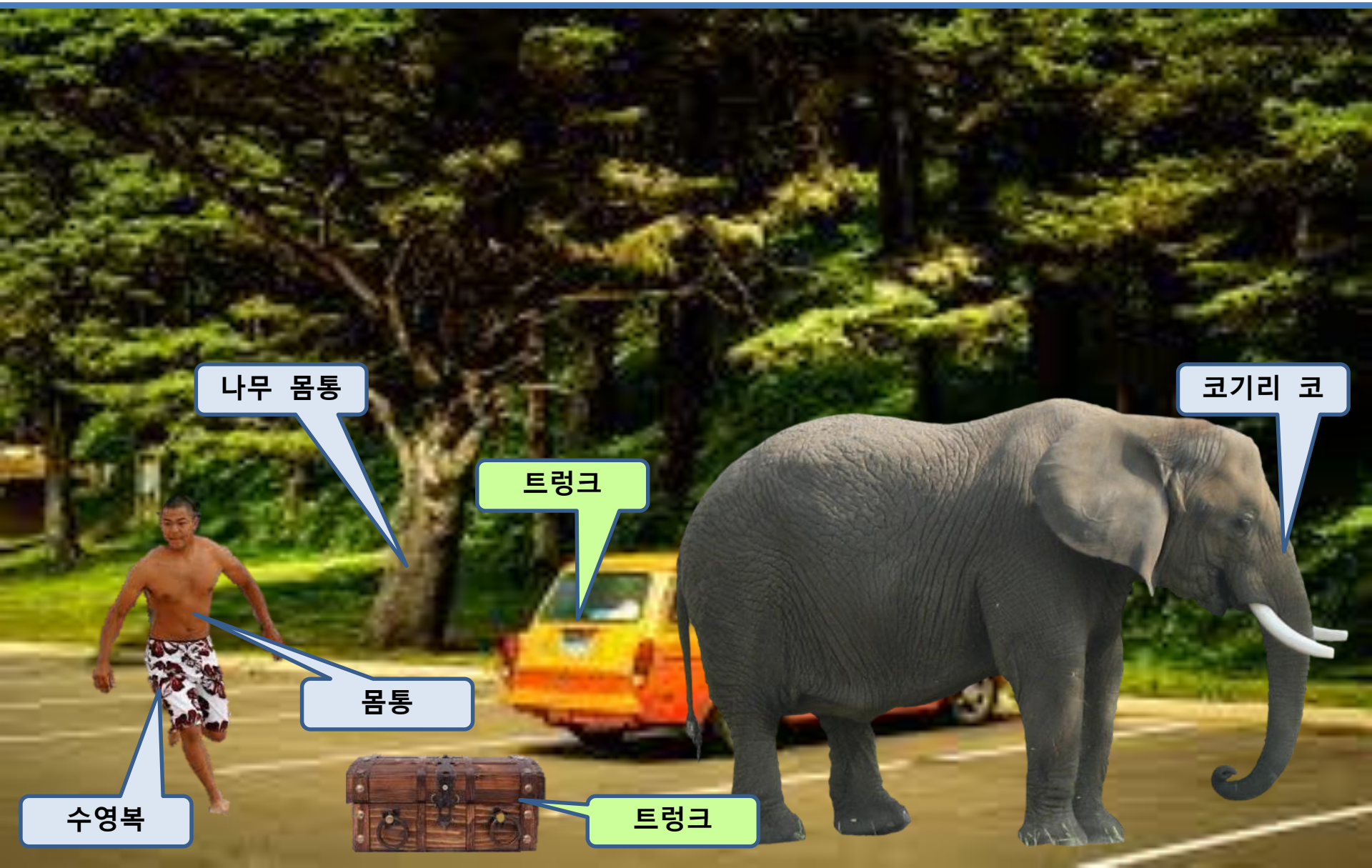
Example of Machine Learning



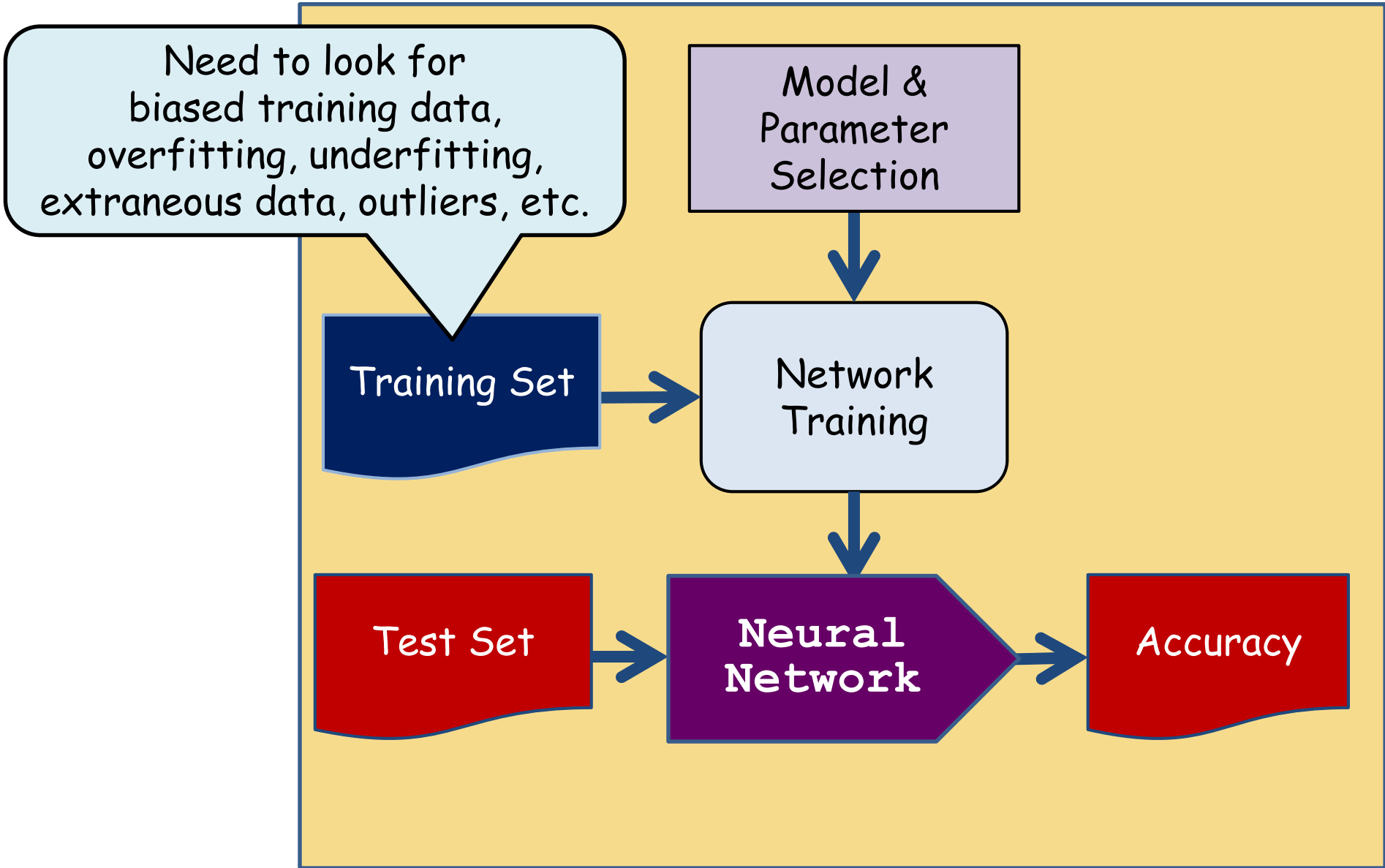
Supervised Machine Learning



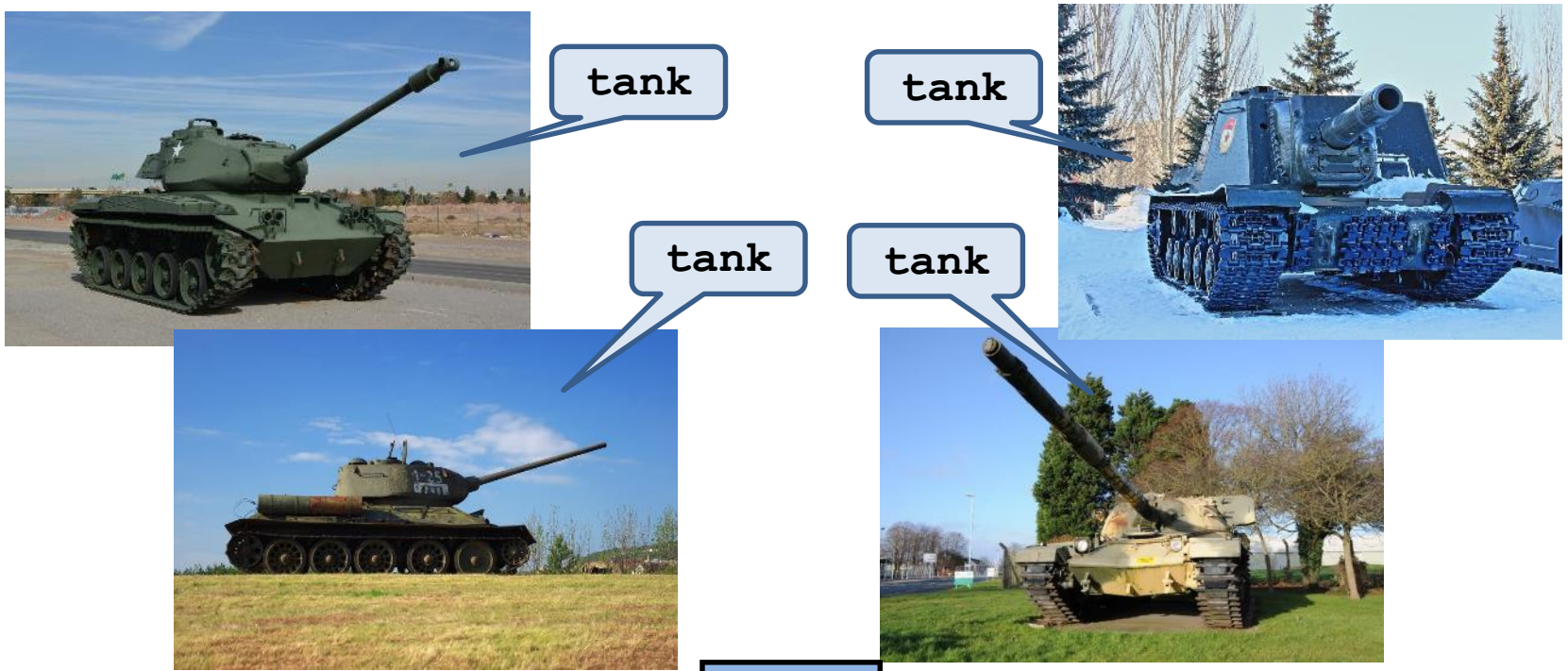
잘못된 분류 – 한글?



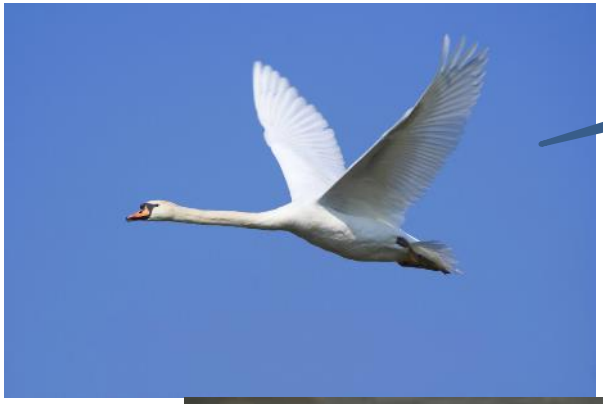
Checking the Training Set



Misunderstanding – Data Bias



Incomplete Training Set

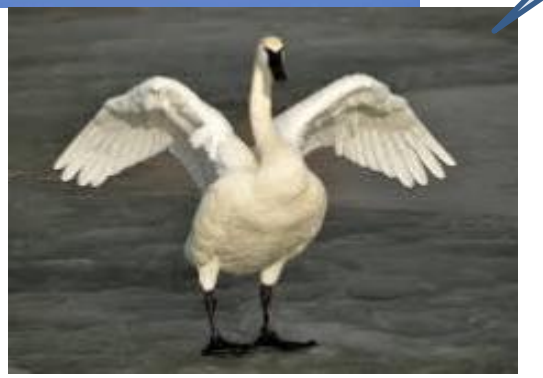


swan

swan

swan

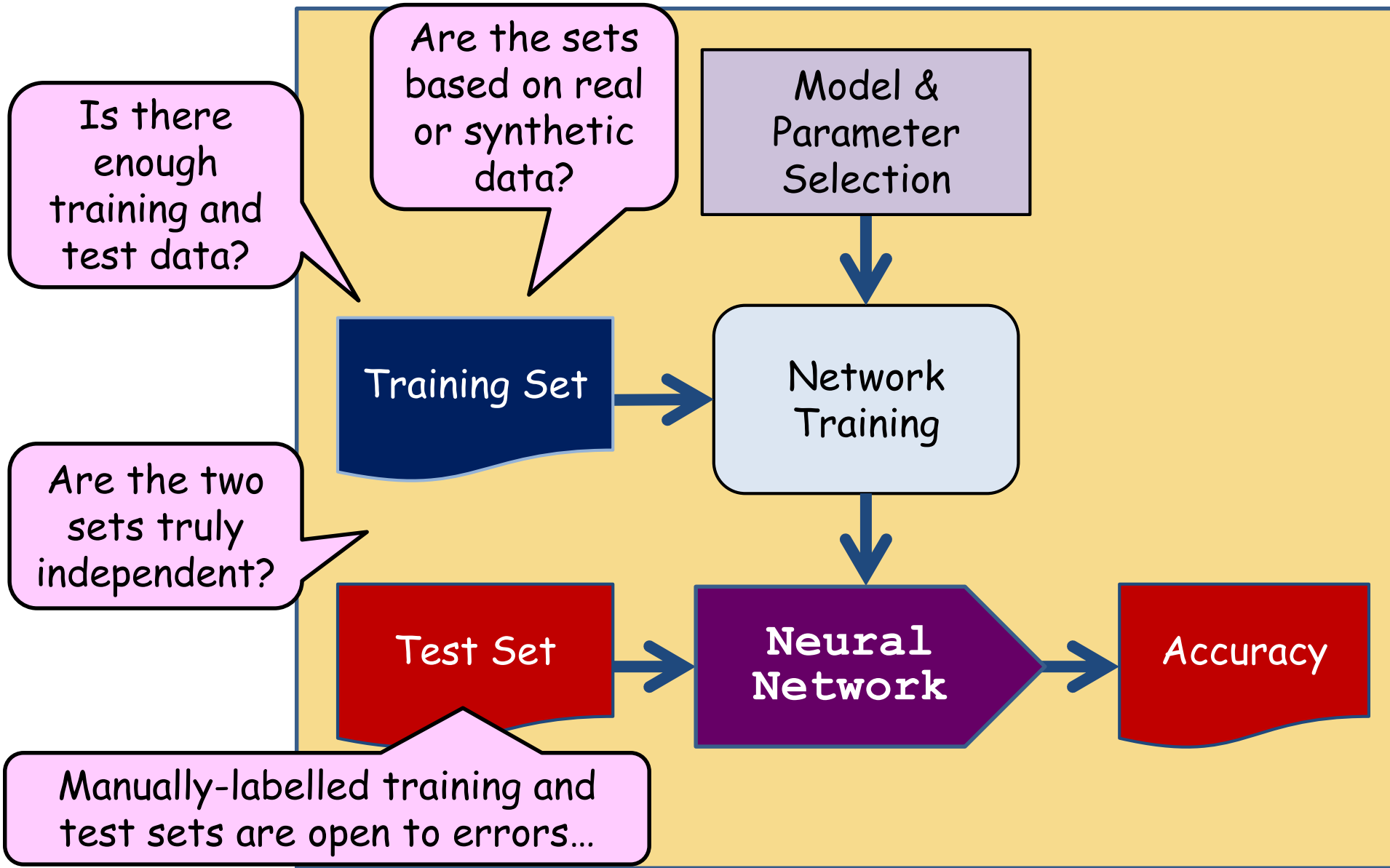
swan



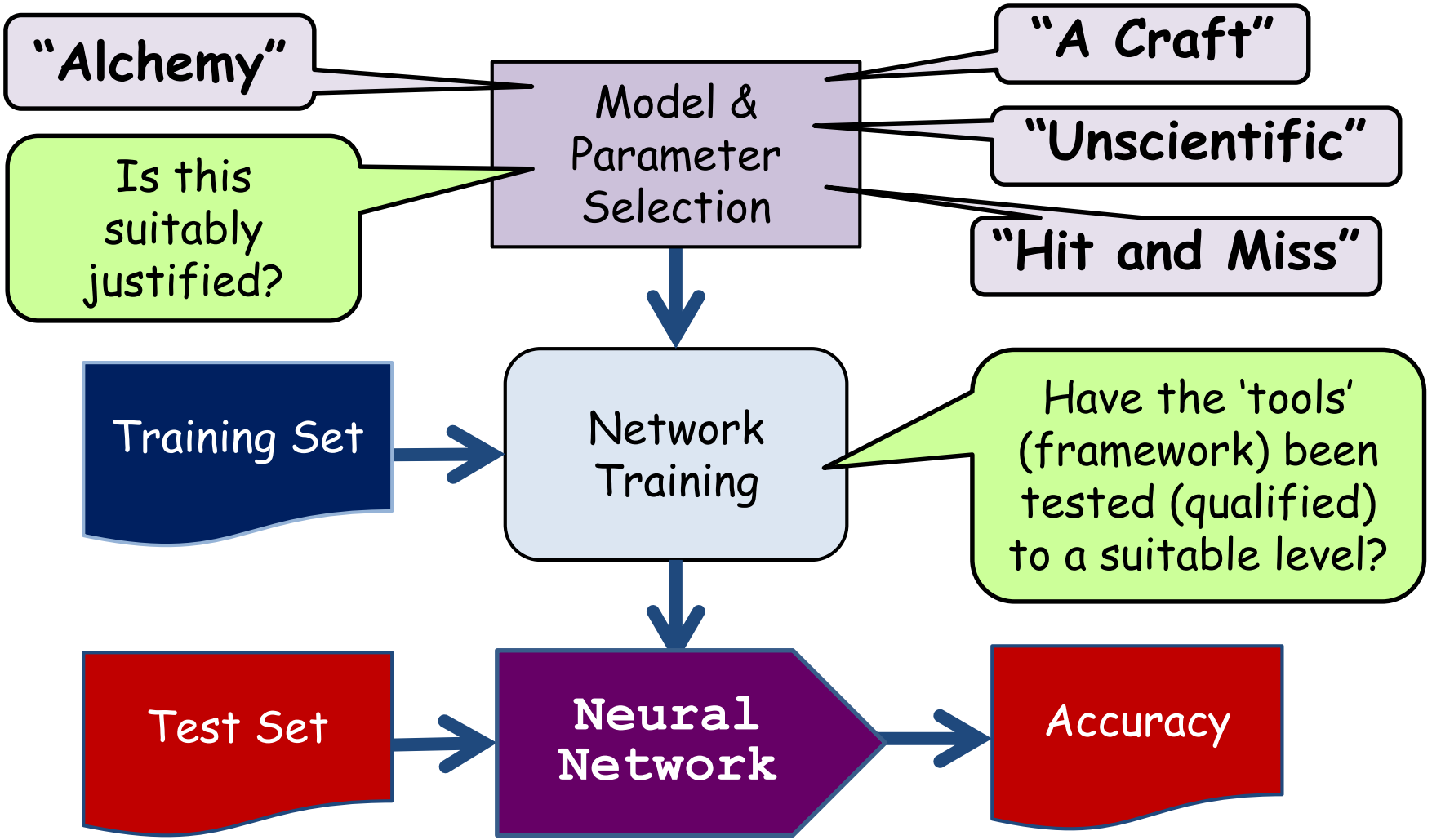
Neural Network

Not a swan

Checking the Training & Test Sets



Checking the Training

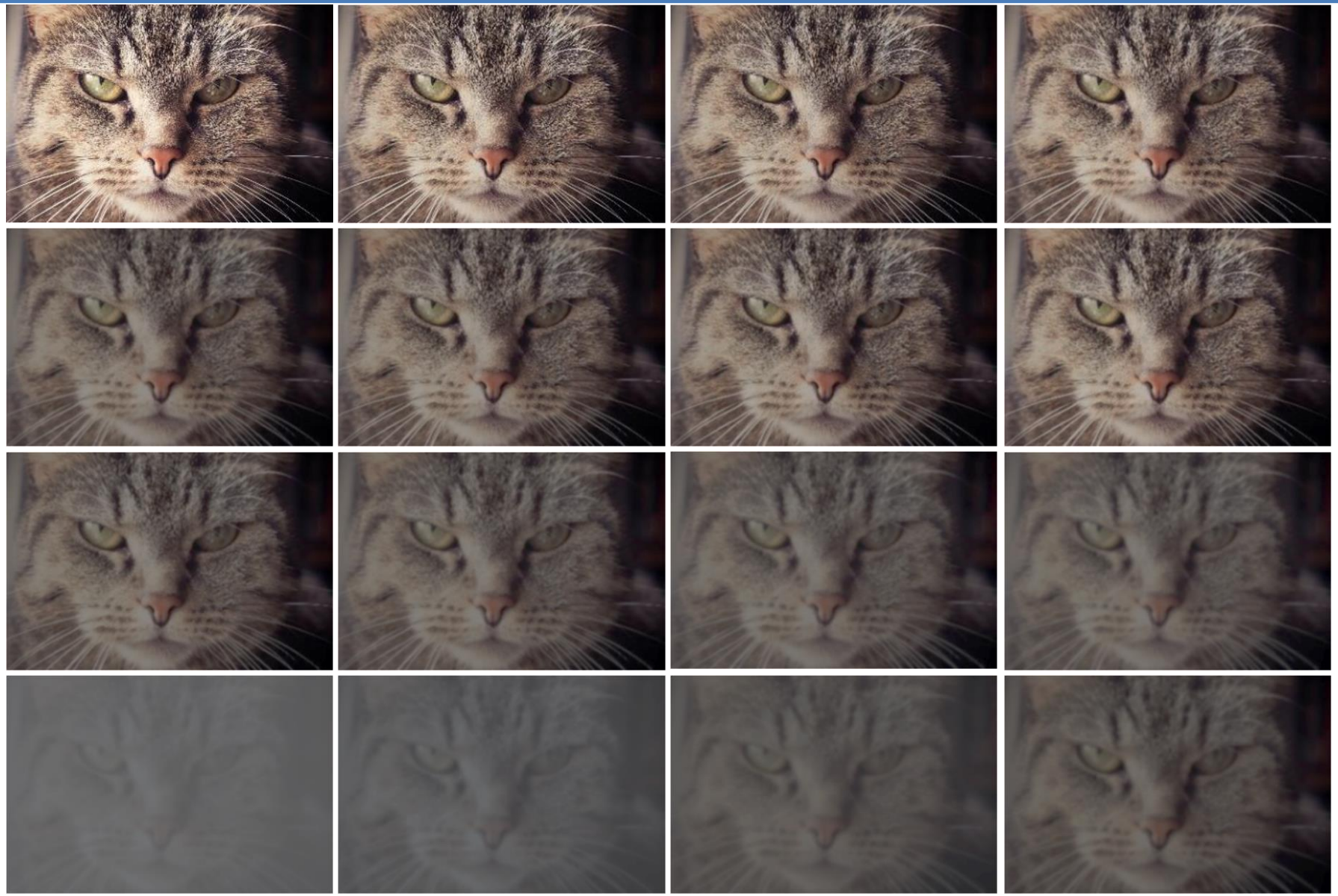


Black Box Testing of Autonomous Systems

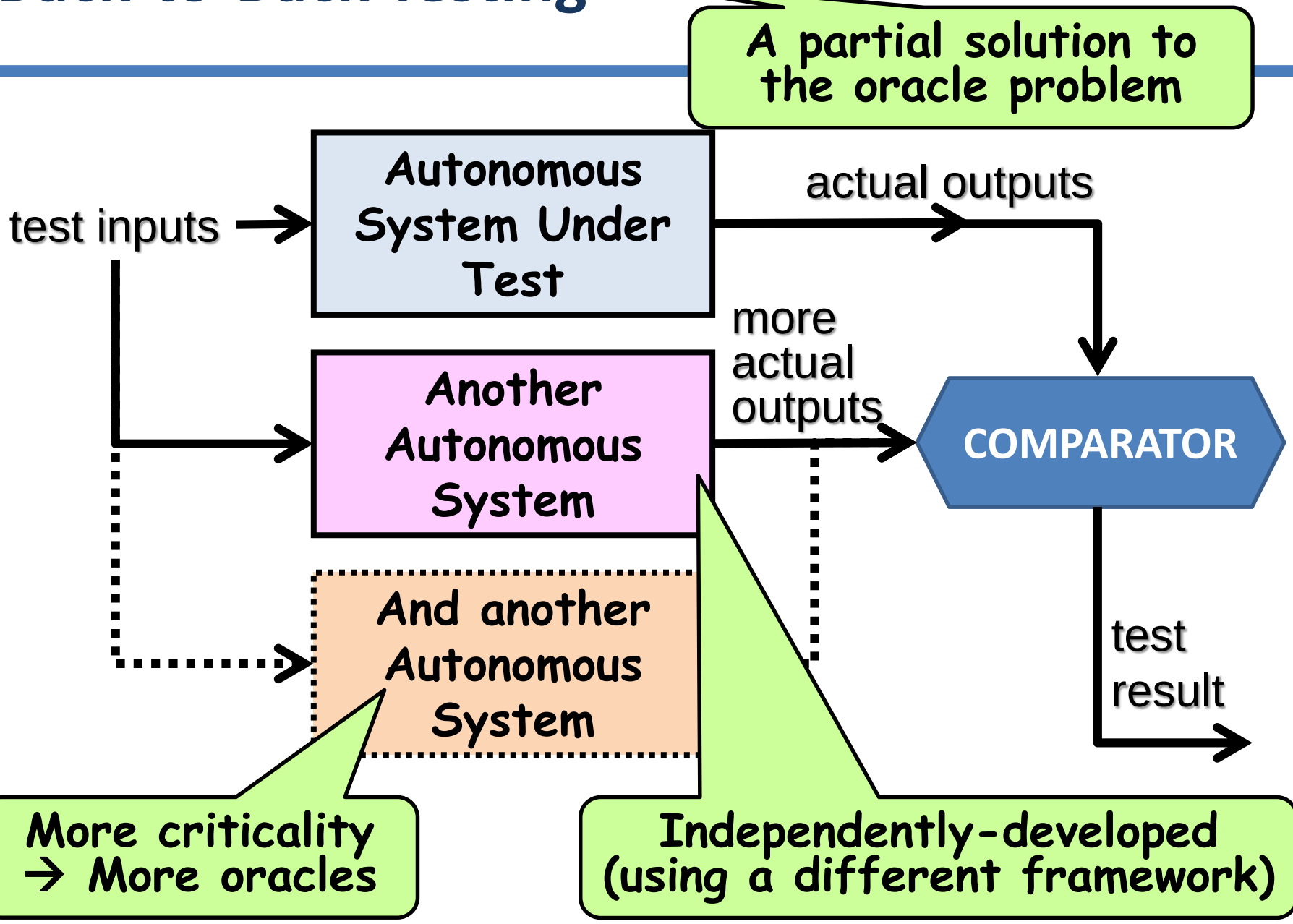
Test Challenges of Autonomous Systems

- **Expected Results (Test Oracle)**
  - if we struggle to set the objectives, then determining expected results will be equally difficult
- **Probabilistic Systems and Non-Determinism**
  - the probabilistic nature means that predicting expected results is difficult
    - we need many more tests to be statistically confident
  - non-determinism causes real problems for regression testing
- **Complexity**
  - autonomous systems are difficult to understand - and to test
  - interacting autonomous systems may cause ‘special’ failures
  - many sensors can create many tests…

Example - Sensor Degradation Testing

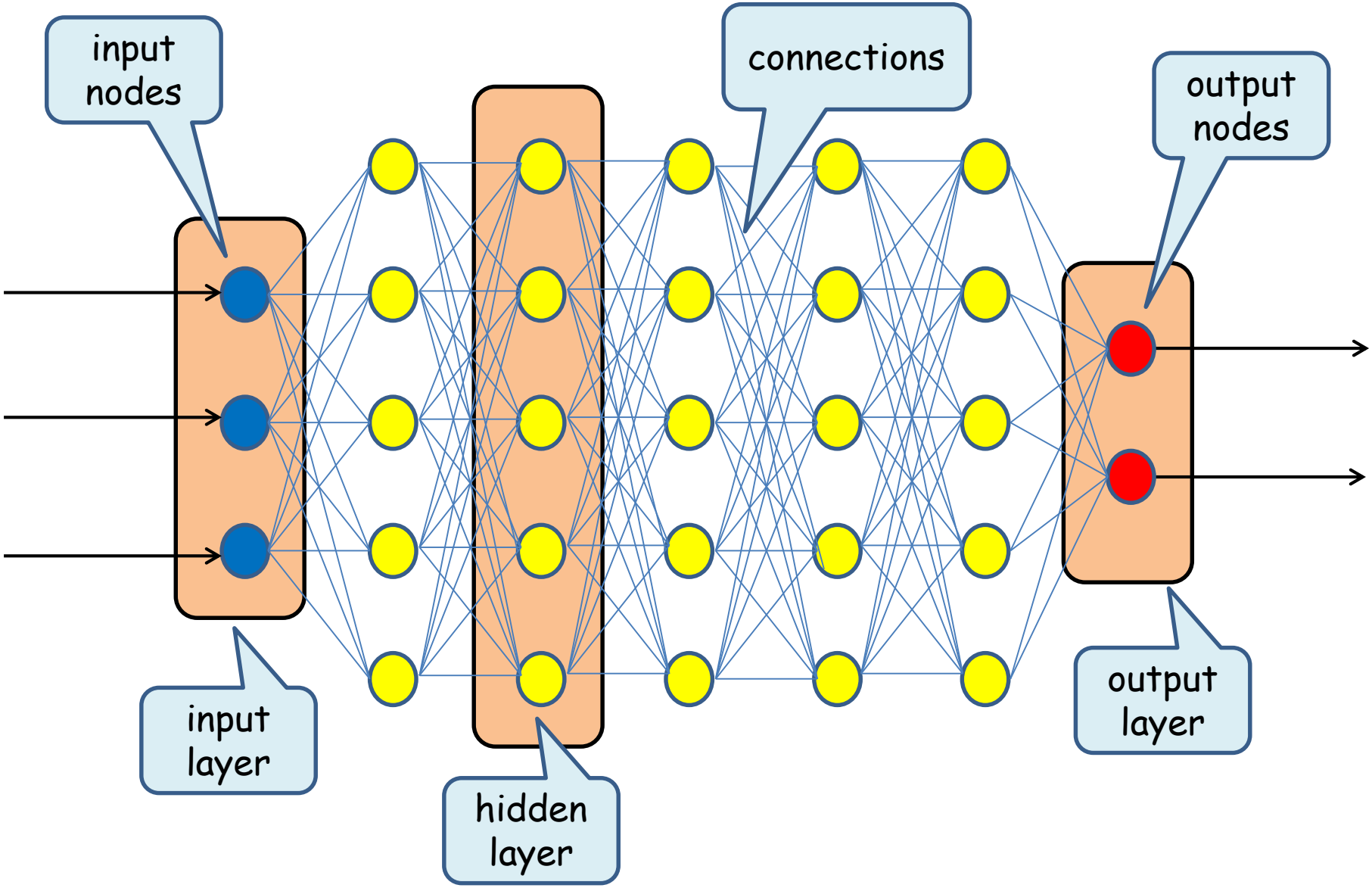


Back-to-Back Testing

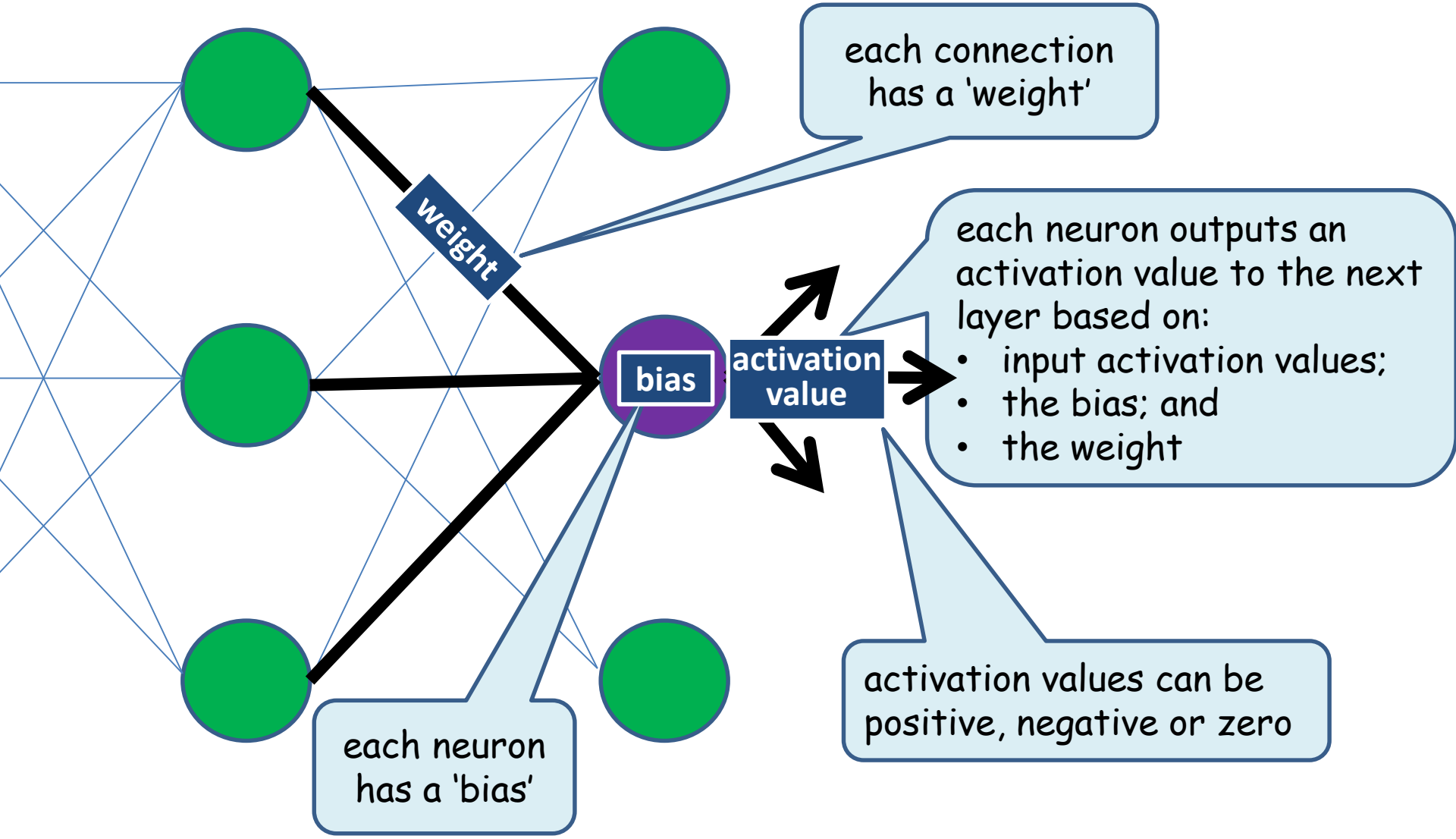


White Box Testing of Autonomous Systems

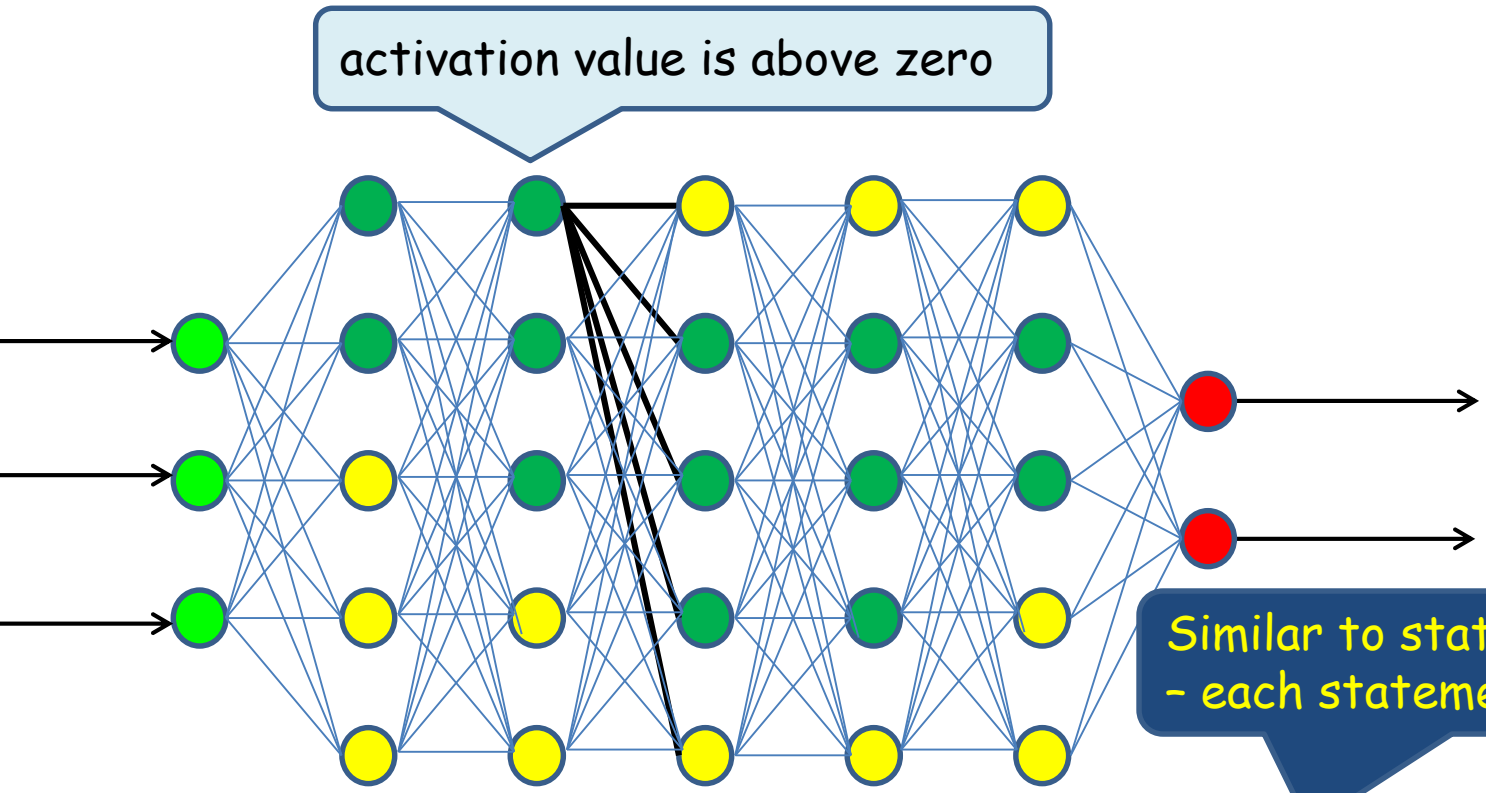
Deep Neural Net



Activation Values



‘Neuron’ Coverage

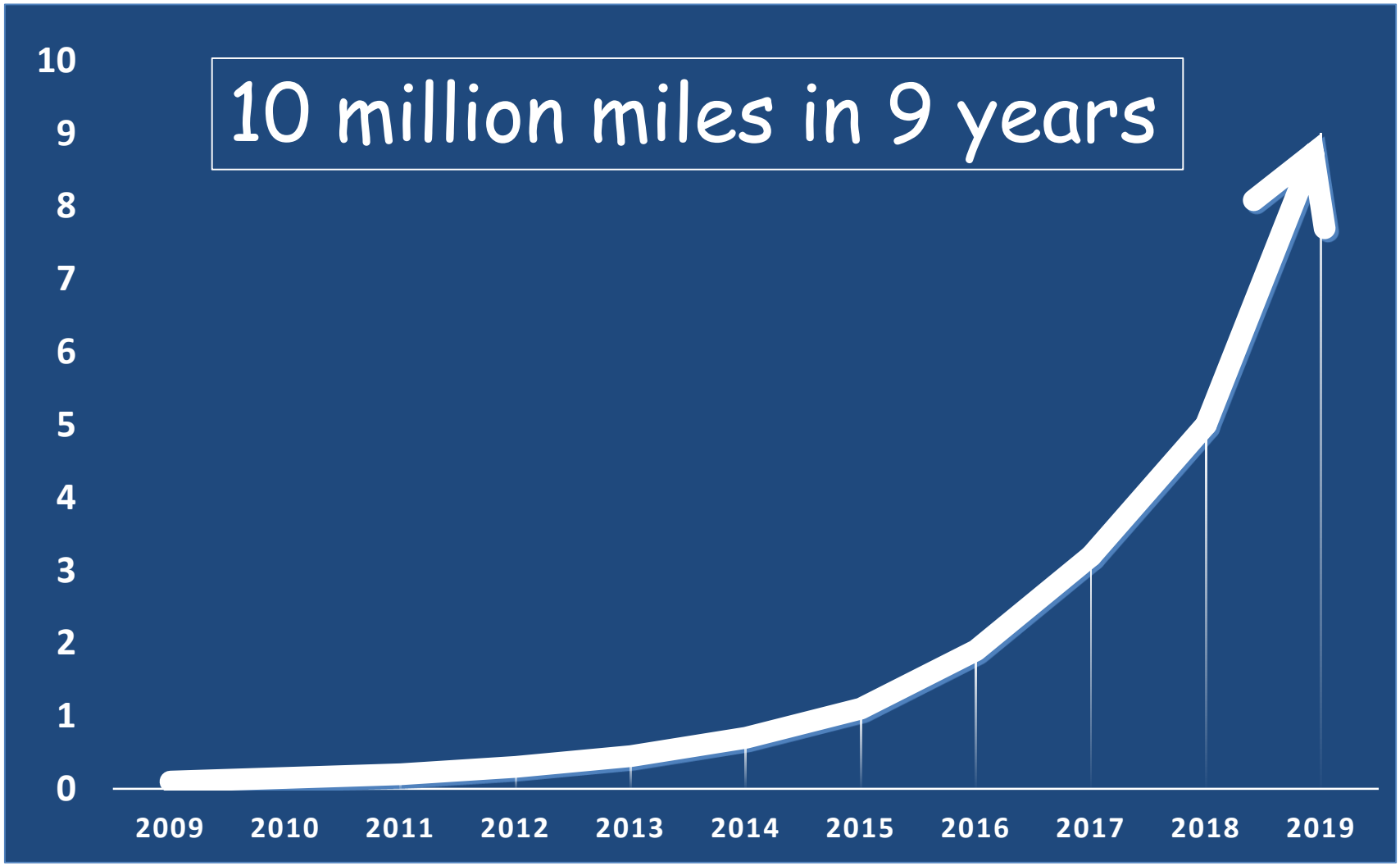


Similar to statement coverage – each statement is exercised

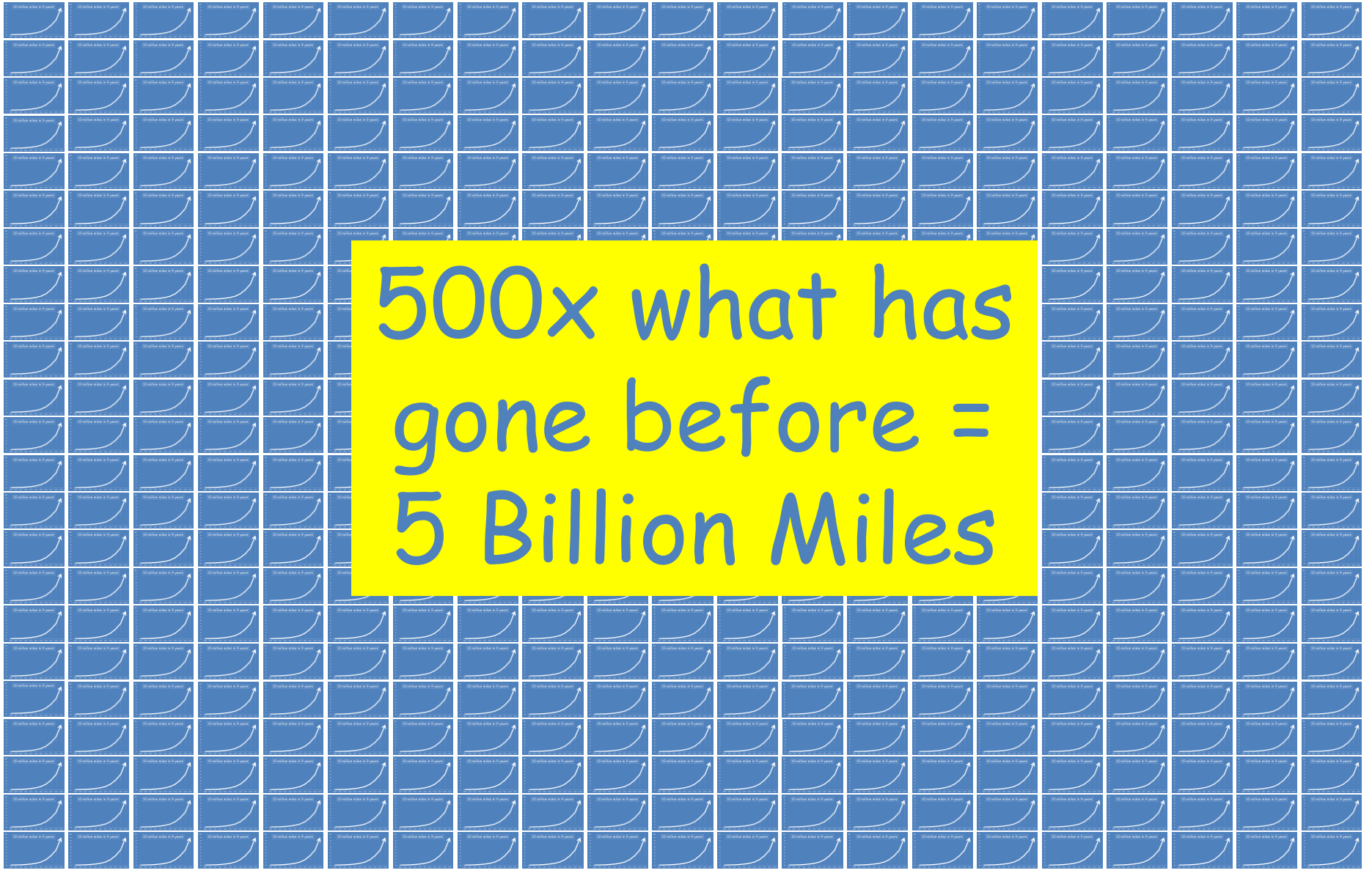
Full ‘neuron’ coverage shows that every neuron is ‘activated’ (value above zero) at least once (but - basic coverage - typically finds no adversarial examples)

The Necessity of Virtual Test Environments

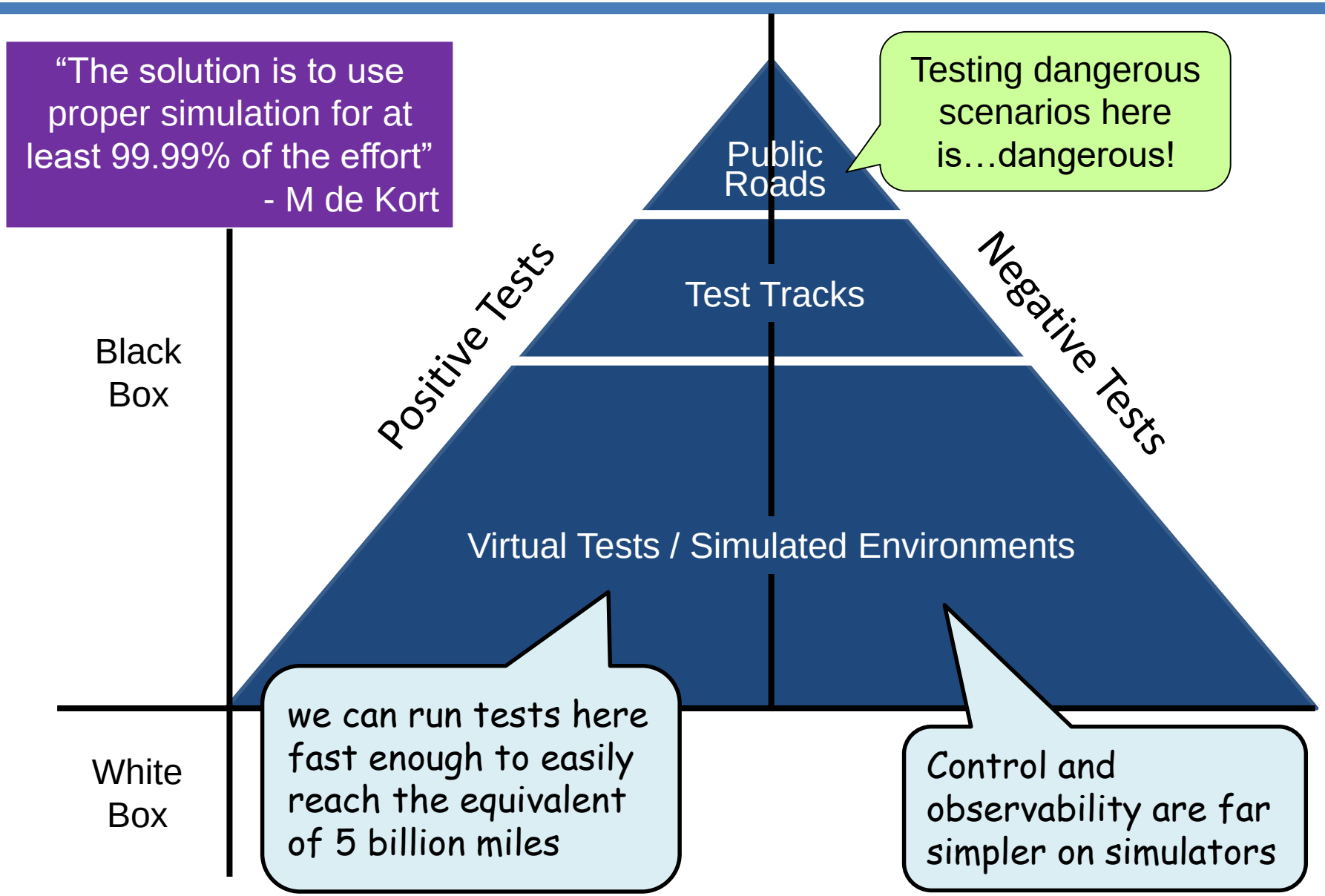
Waymo On-Road Test Miles (millions)



20% Better (than human drivers)

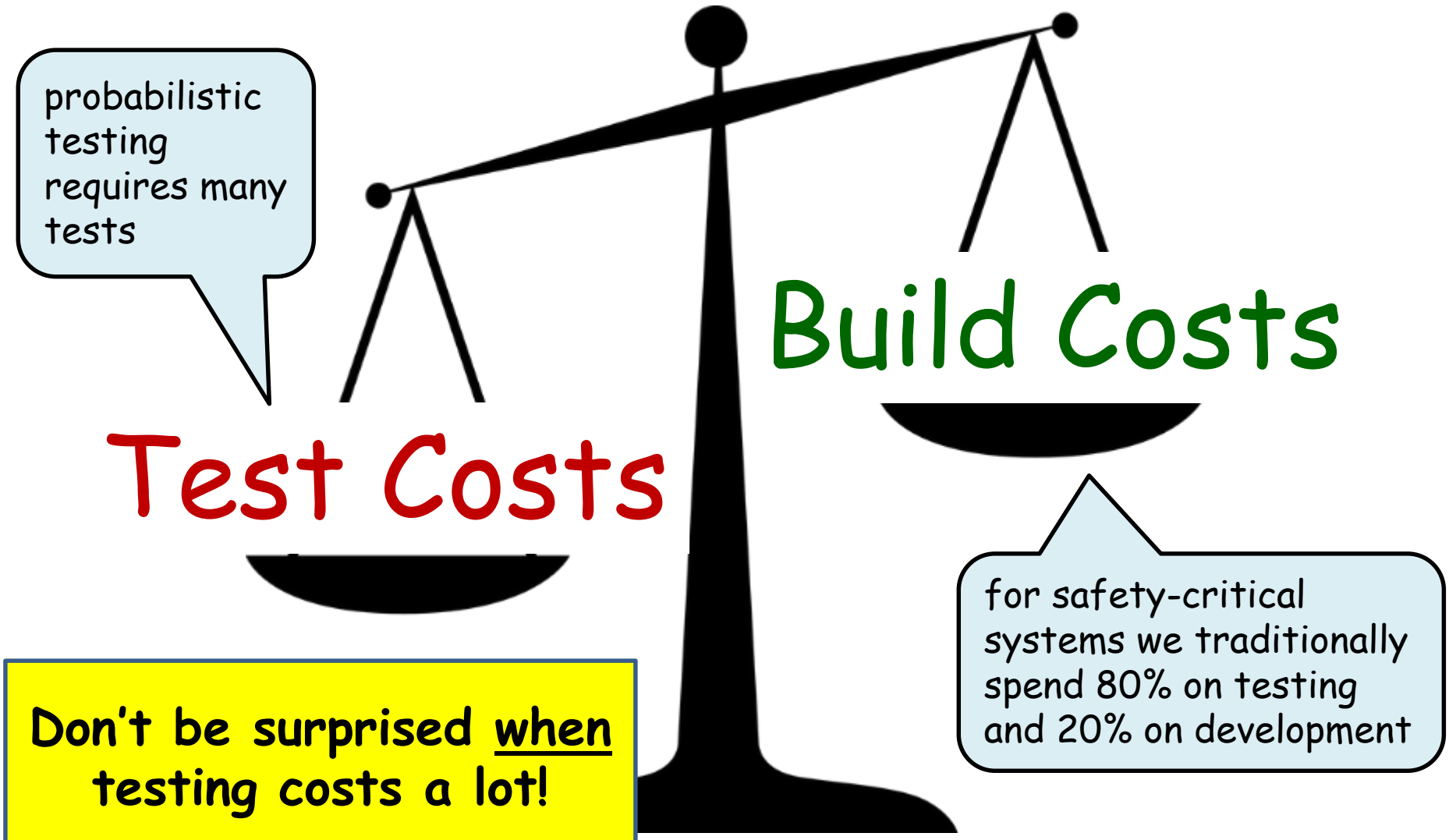


Autonomous Cars – Test Environments



Conclusions

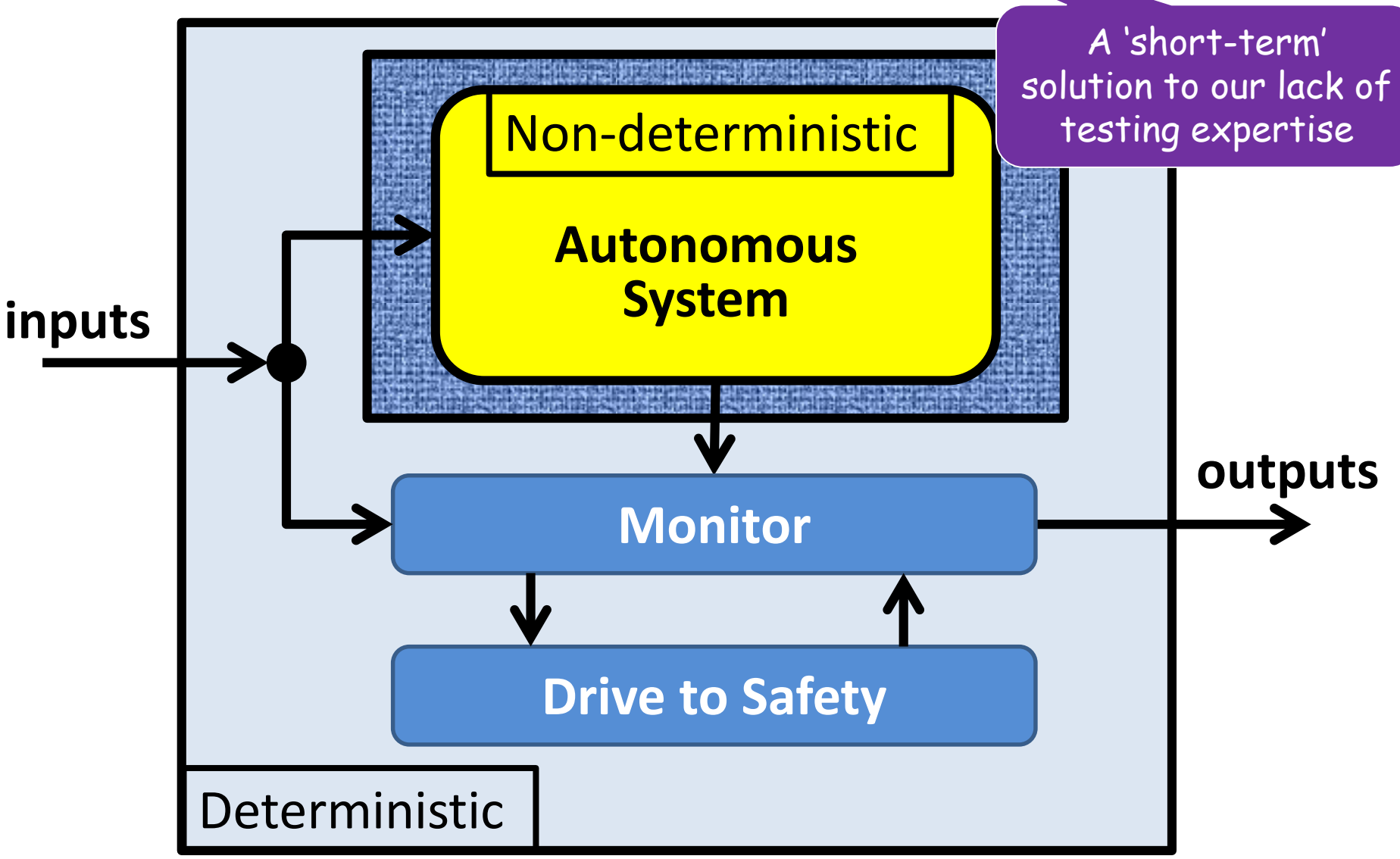
Autonomous System Costs



Conclusions – Safety of Autonomous Systems

- **For the ‘simple’ case of off-line systems we need:**
  - both black and white box testing
  - new test approaches and measures (with evidence)
  - more tests to assure these probabilistic systems
  - the support of sophisticated virtual test environments
- **For the learning on-line systems we need:**
  - to understand the new dangers these systems bring
- **Until we reach maturity, we should use a safety net…**

Safety Shell Architecture



Thank you for listening