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

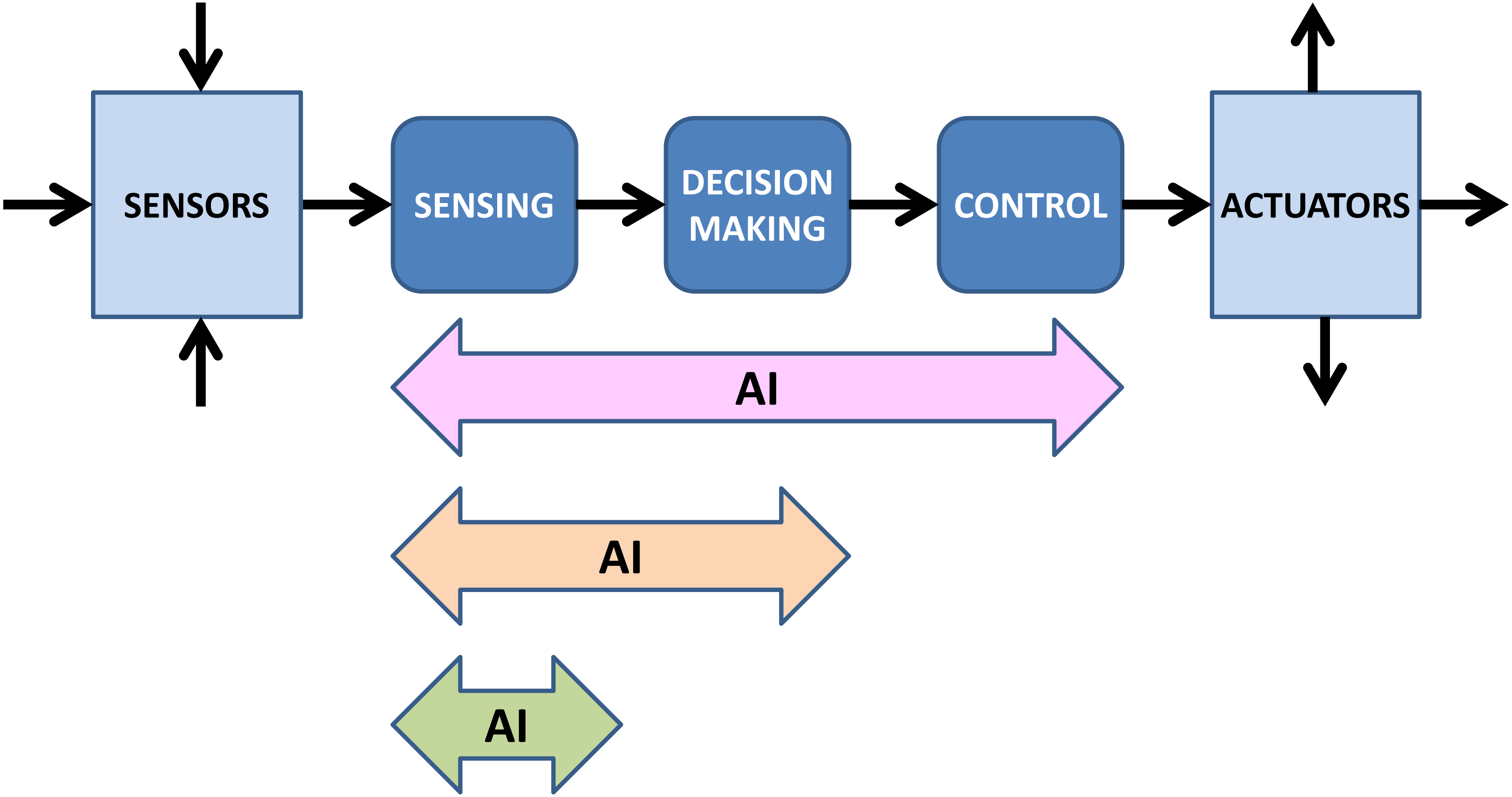
Introduction to Autonomous Systems

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



SENSORS

SENSING

DECISION MAKING

CONTROL

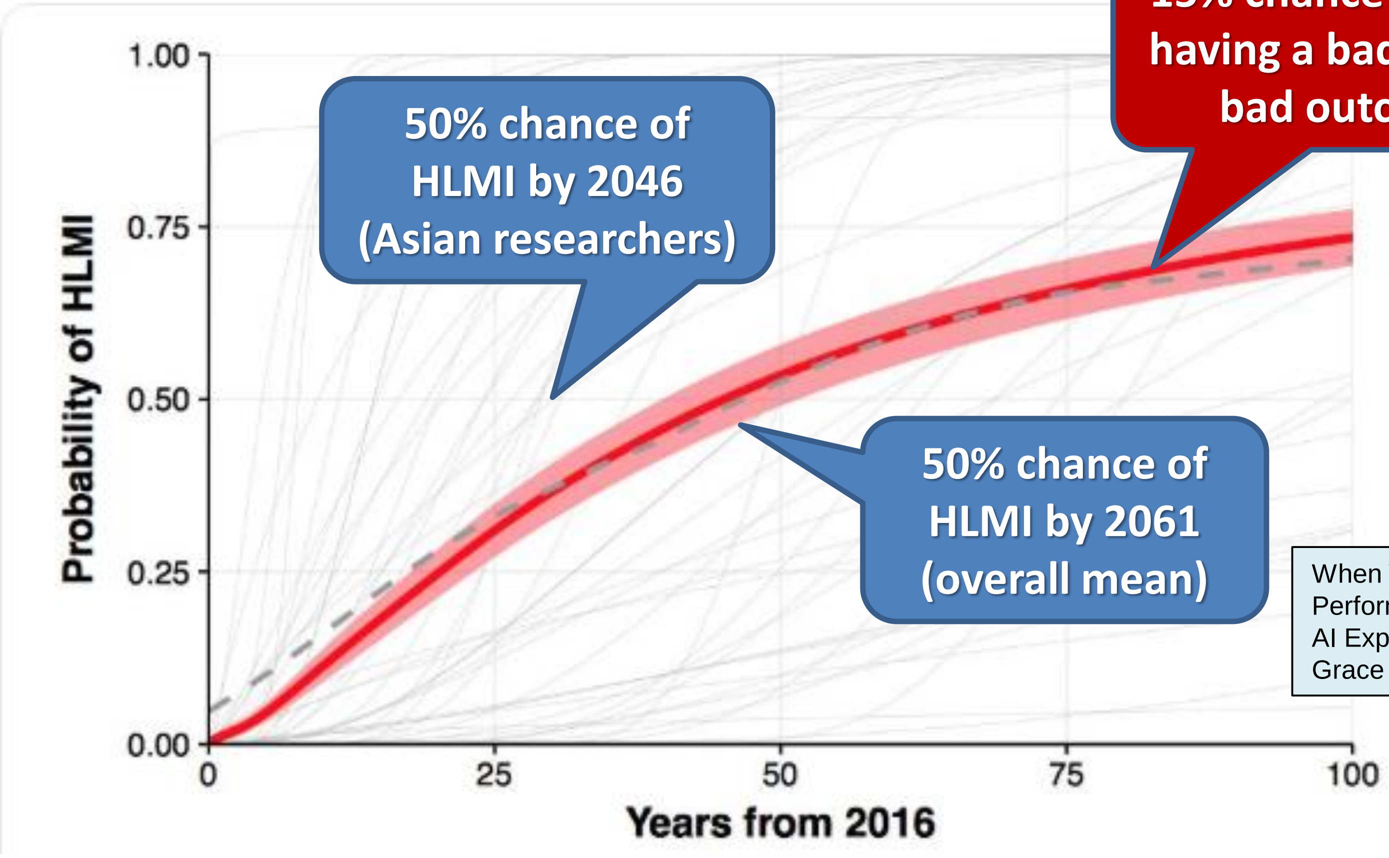
ACTUATORS

AI

AI

AI

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



50% chance of HLMI by 2046 (Asian researchers)

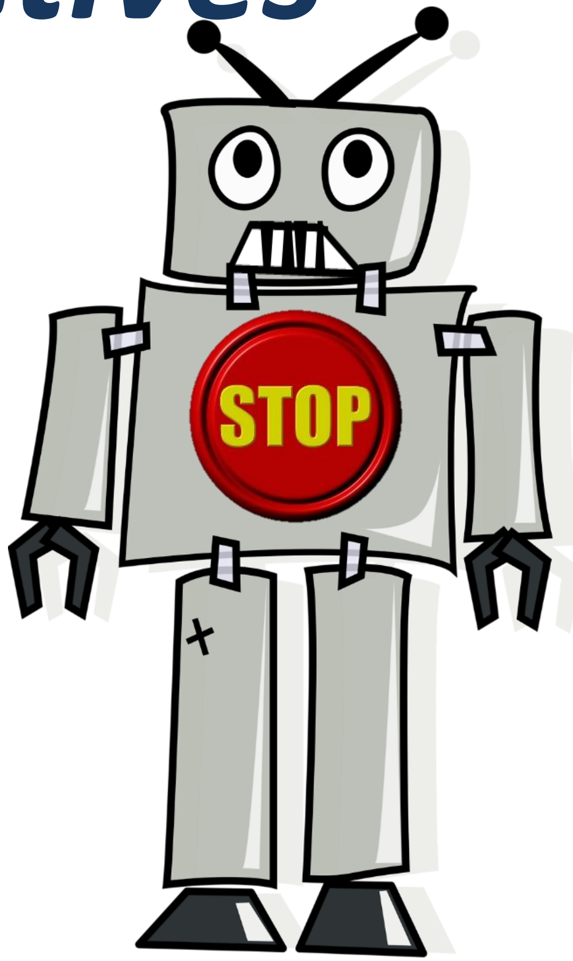
15% chance of HLMI having a bad or very bad outcome

50% chance of HLMI by 2061 (overall mean)

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

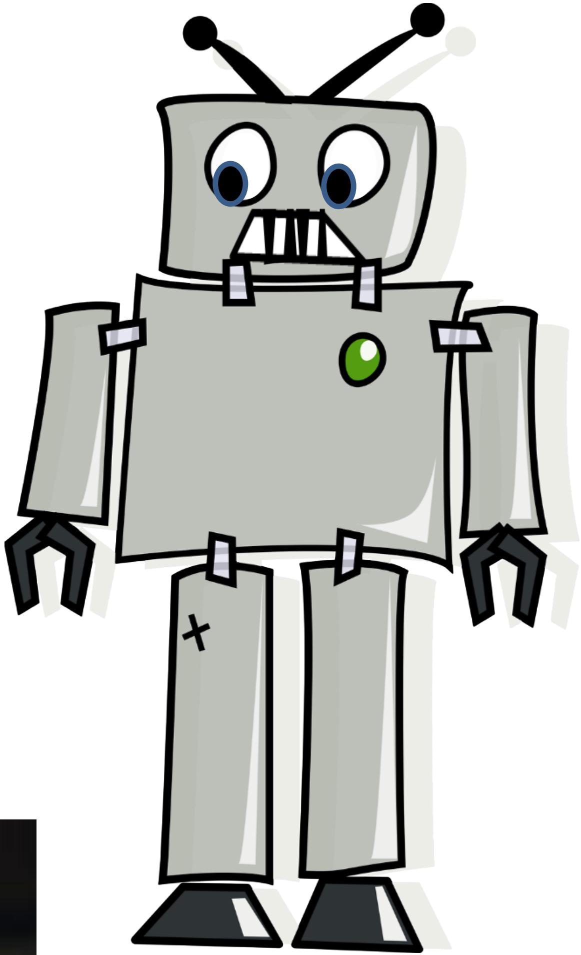
"High-level machine intelligence" (HLMI) is achieved when unaided machines can accomplish every task better and more cheaply than human workers.

Specifying Objectives (Safely)

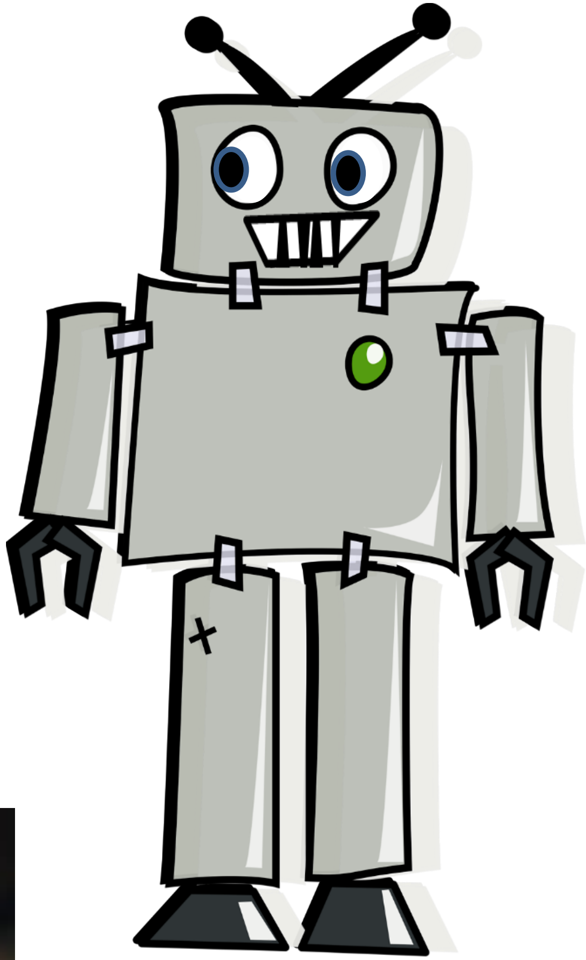


The Midas Problem “마이더스의 손”

“I’m hungry!   Make me dinner”



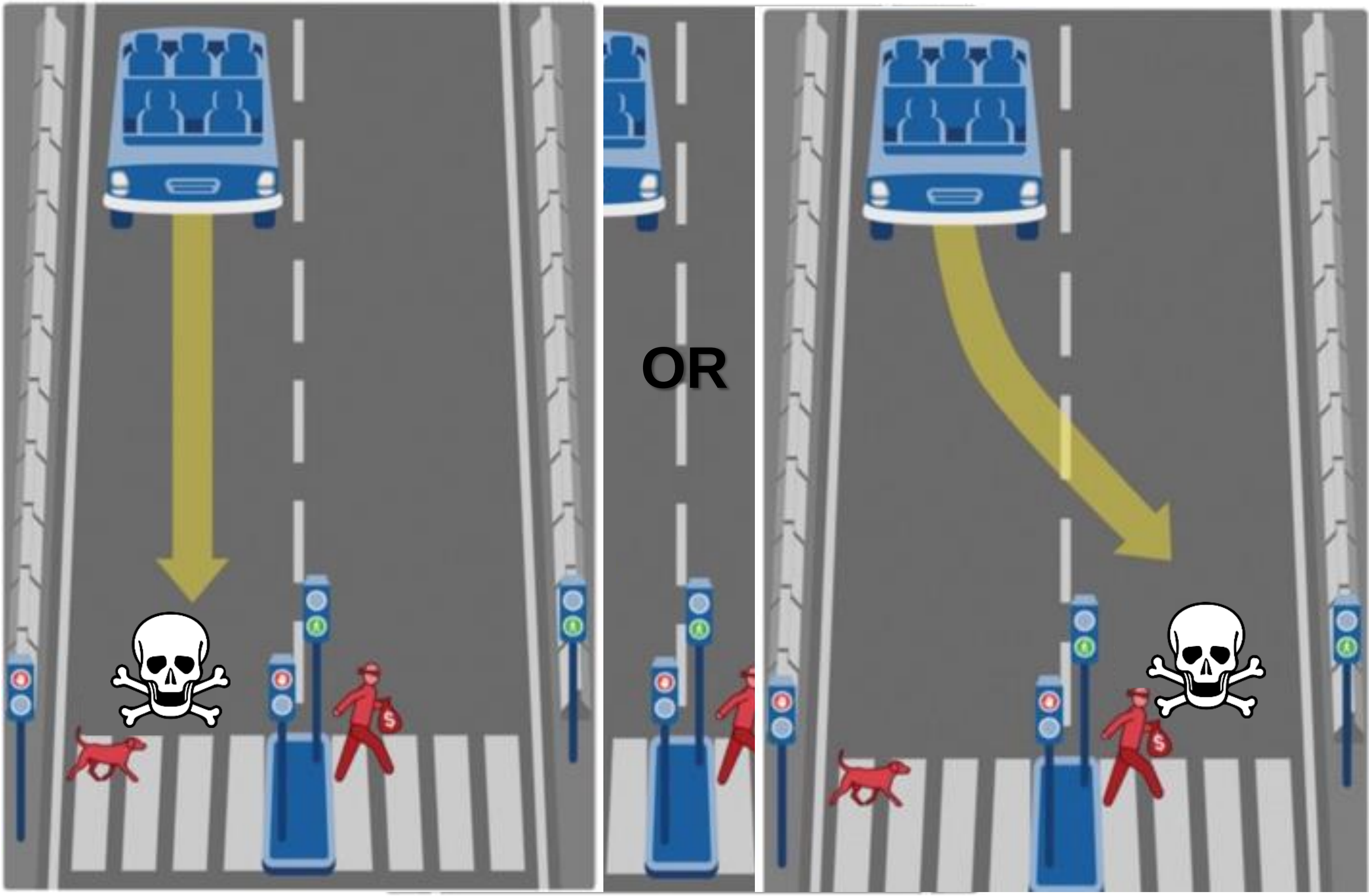
“Keep the kitchen clean”



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)



Go to **www.menti.com** and use the code **41 56 35**

Who should die?

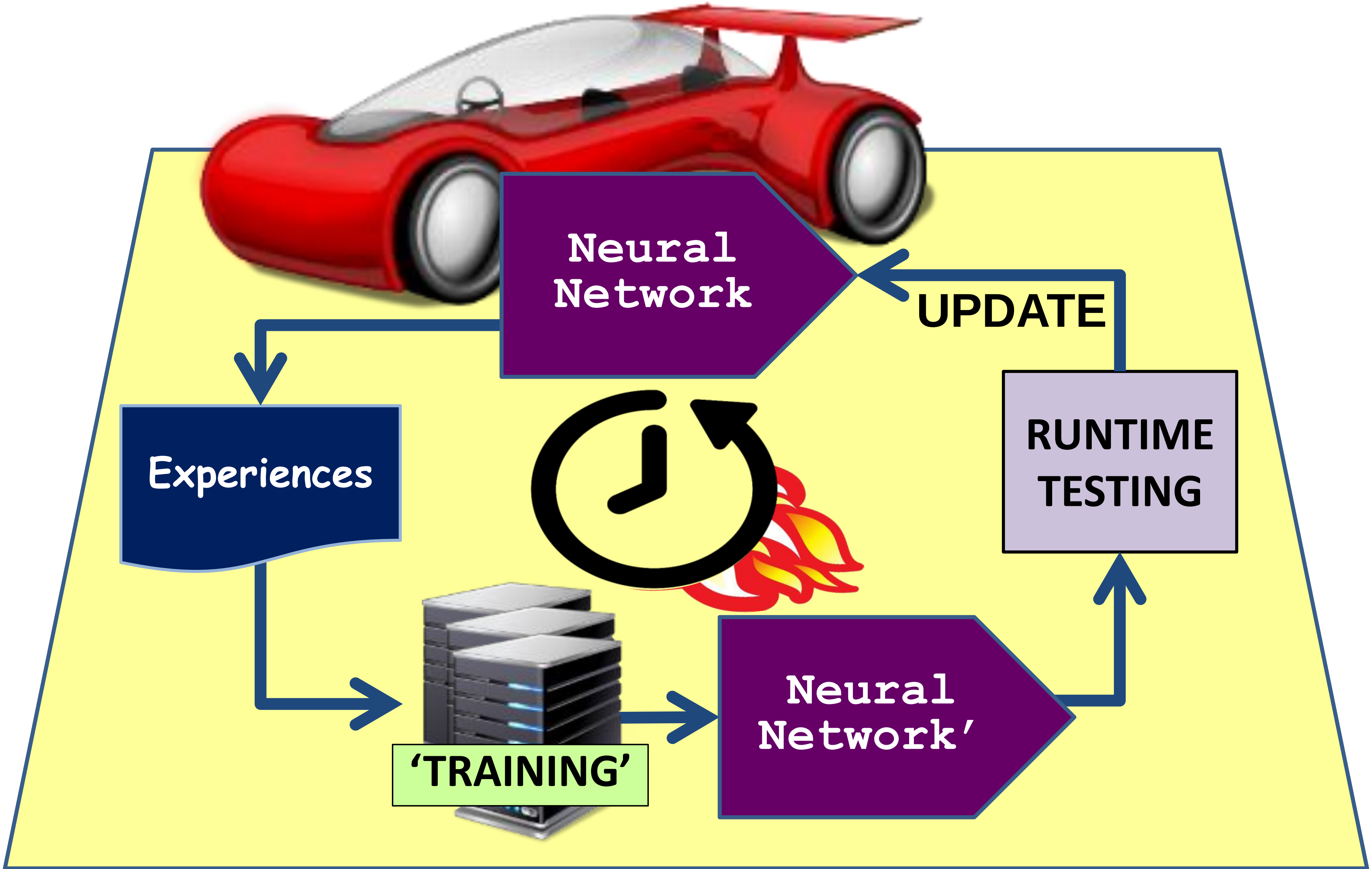
Slide is not active [Activate]

0

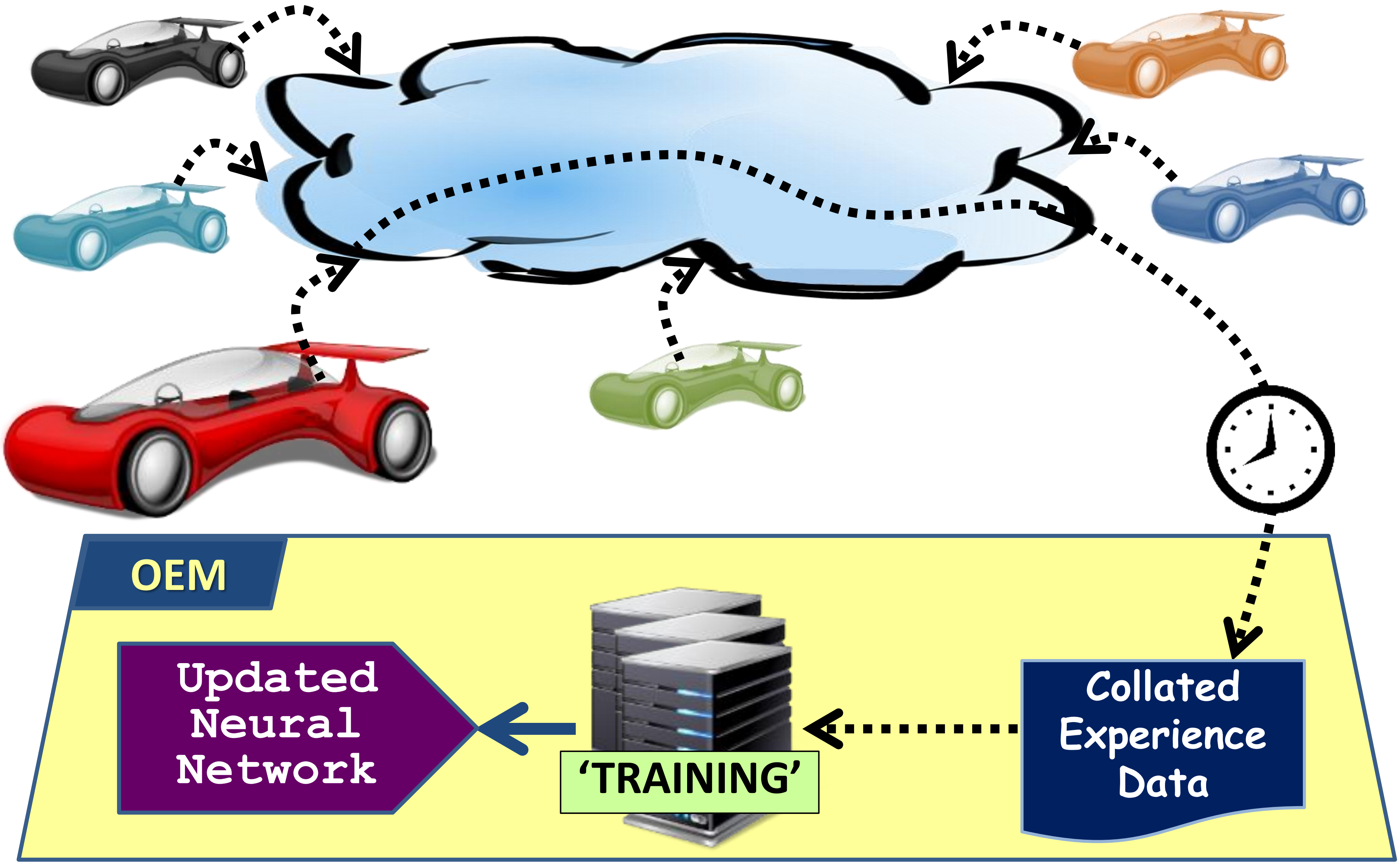
Better than Humans?

Online vs Off-Line Machine Learning

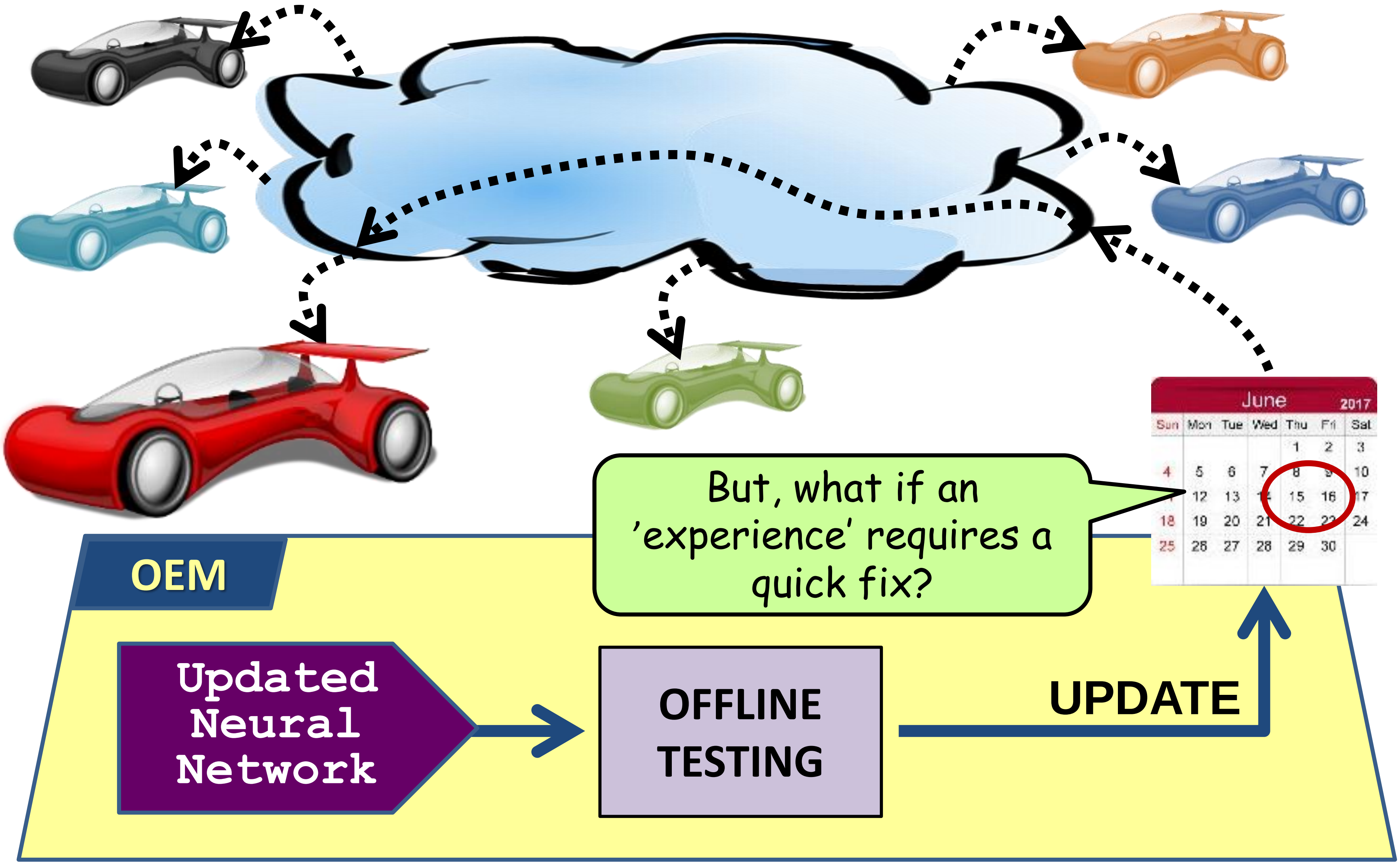
Continuous Online Learning



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

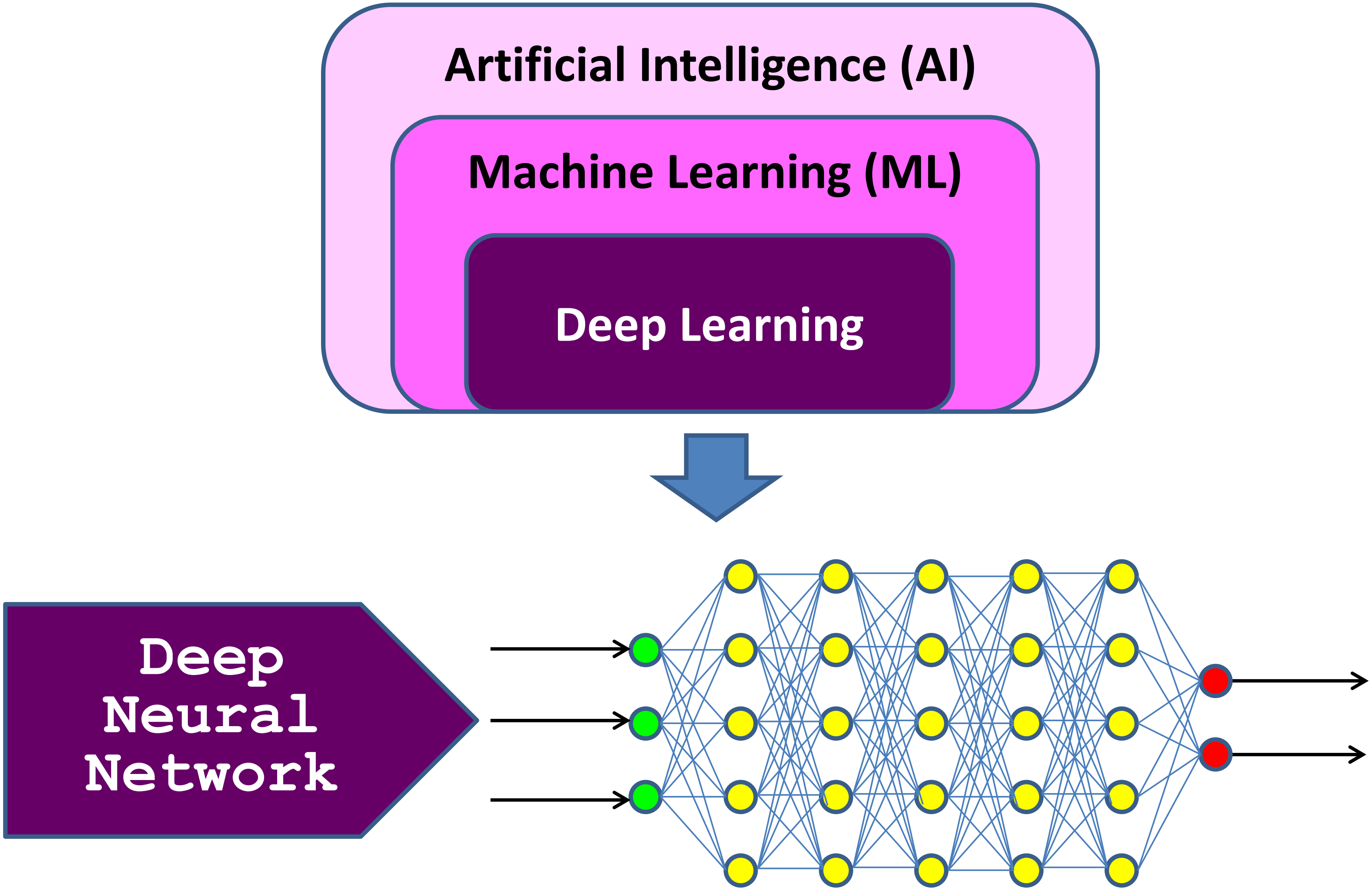
UPDATE

June 2017

Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Machine Learning Challenges

Deep Learning Systems



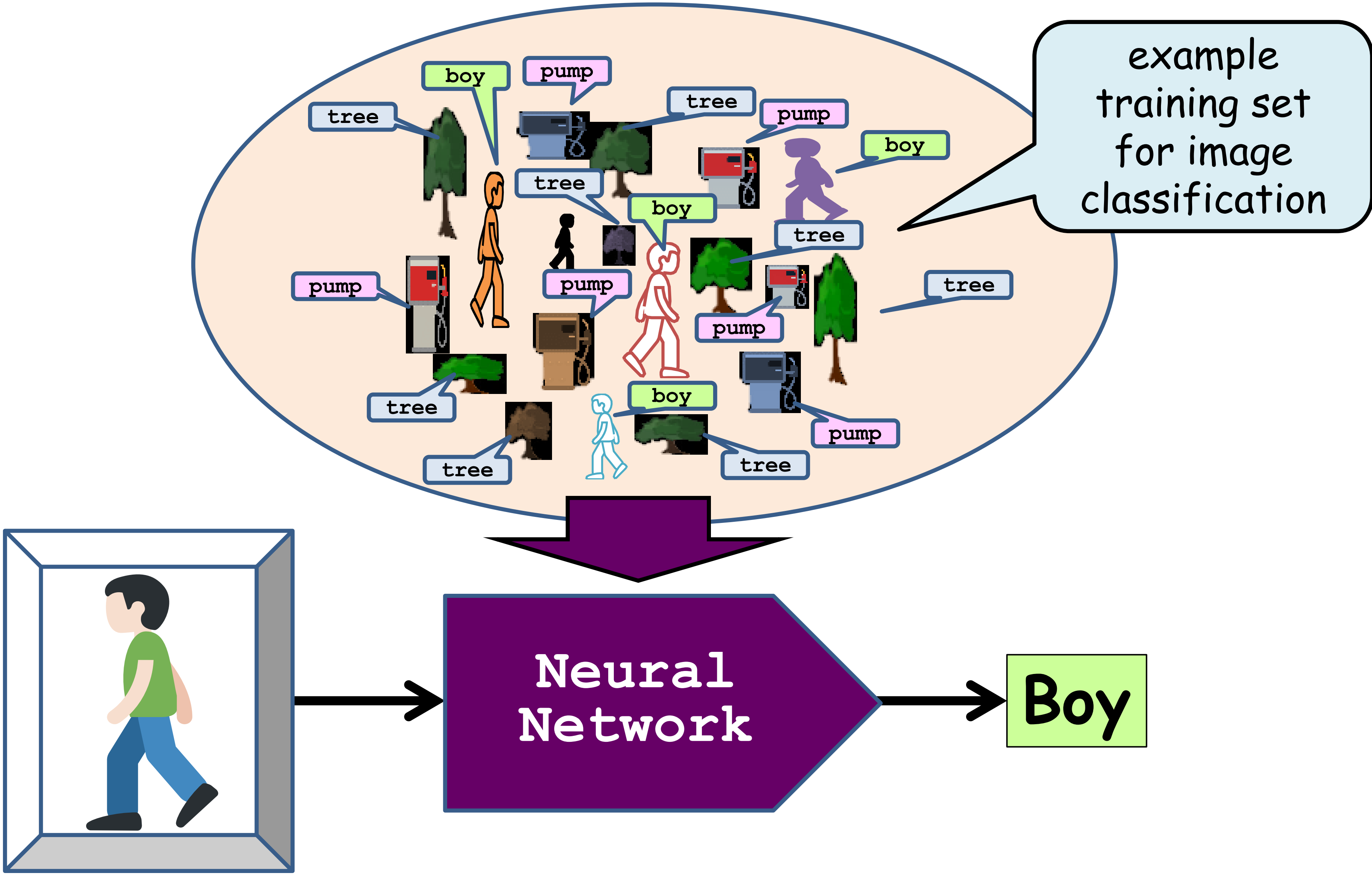
Artificial Intelligence (AI)

Machine Learning (ML)

Deep Learning

Deep
Neural
Network

Example of Machine Learning

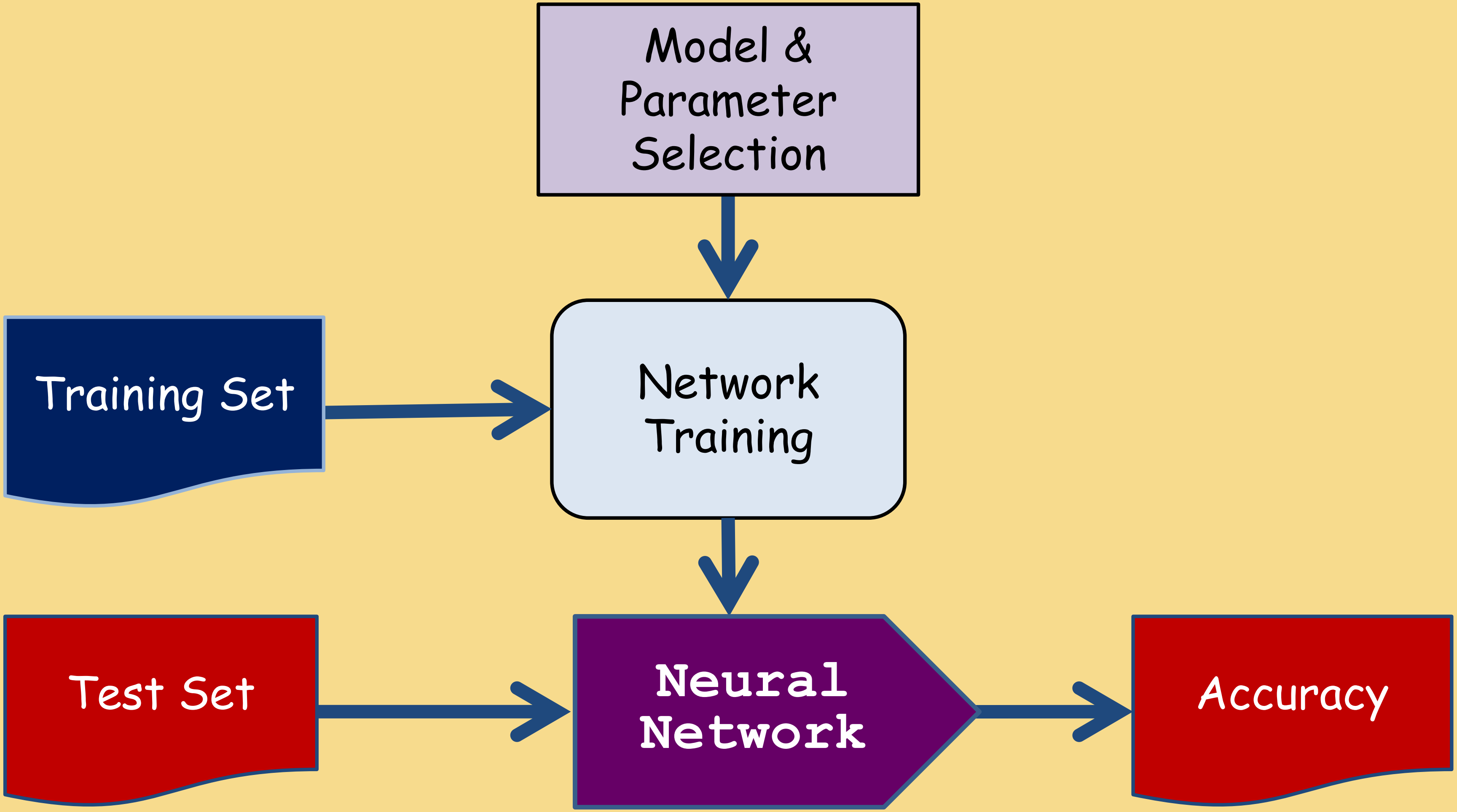


example training set for image classification

Neural Network

Boy

Supervised Machine Learning


Model & Parameter Selection

Training Set

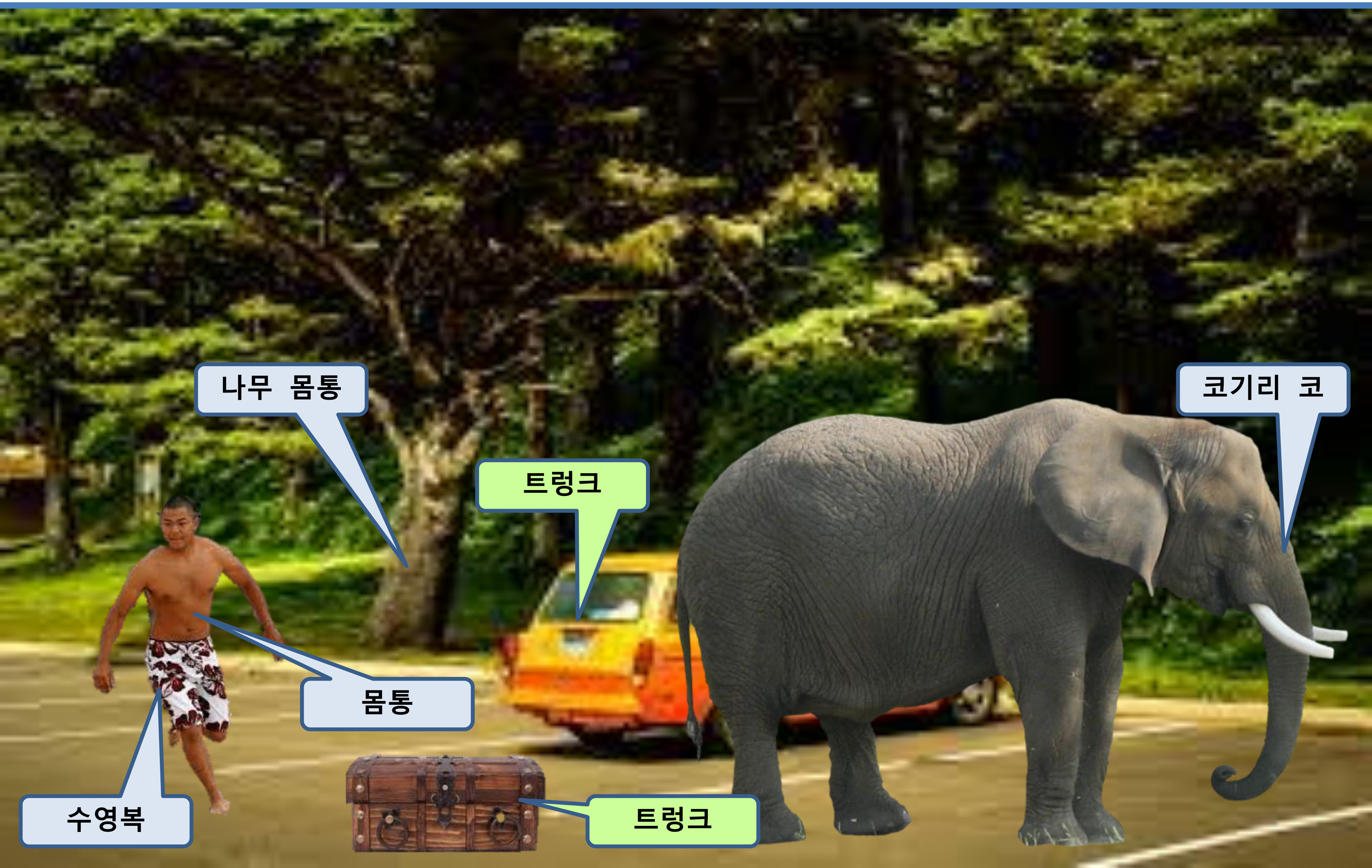
Network Training

Test Set

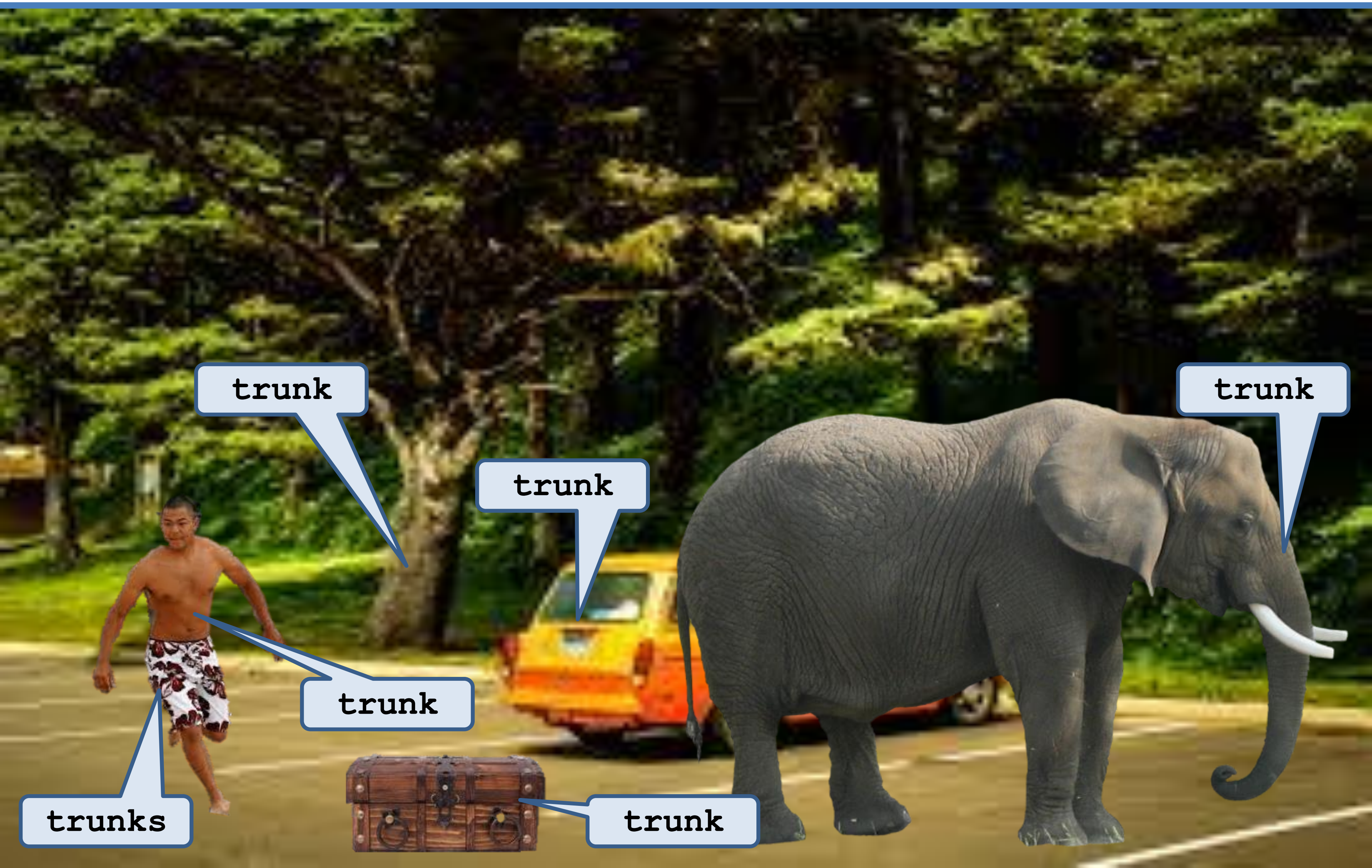
Neural Network

Accuracy

잘못된 분류 – 한글?

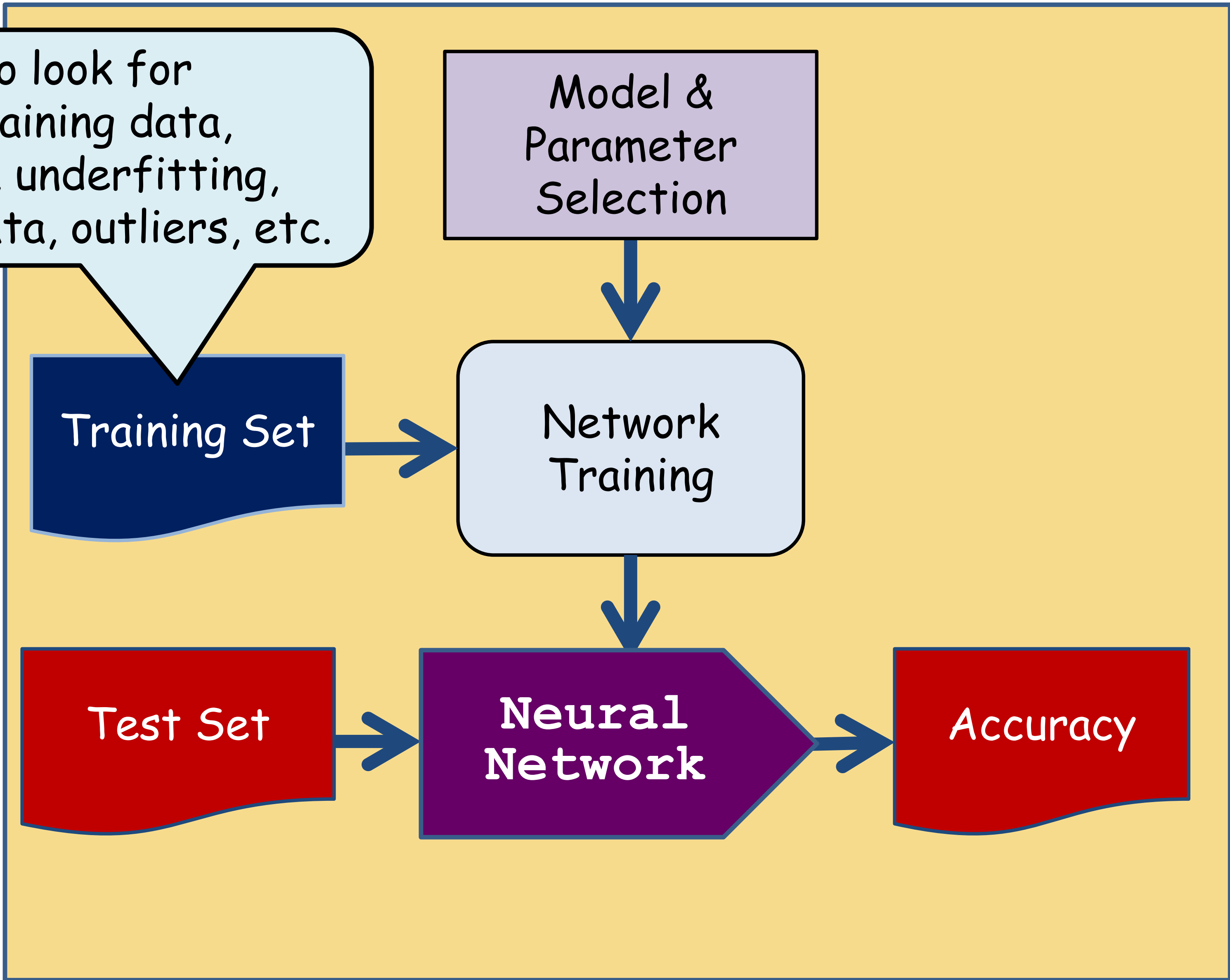


Mis-Classification

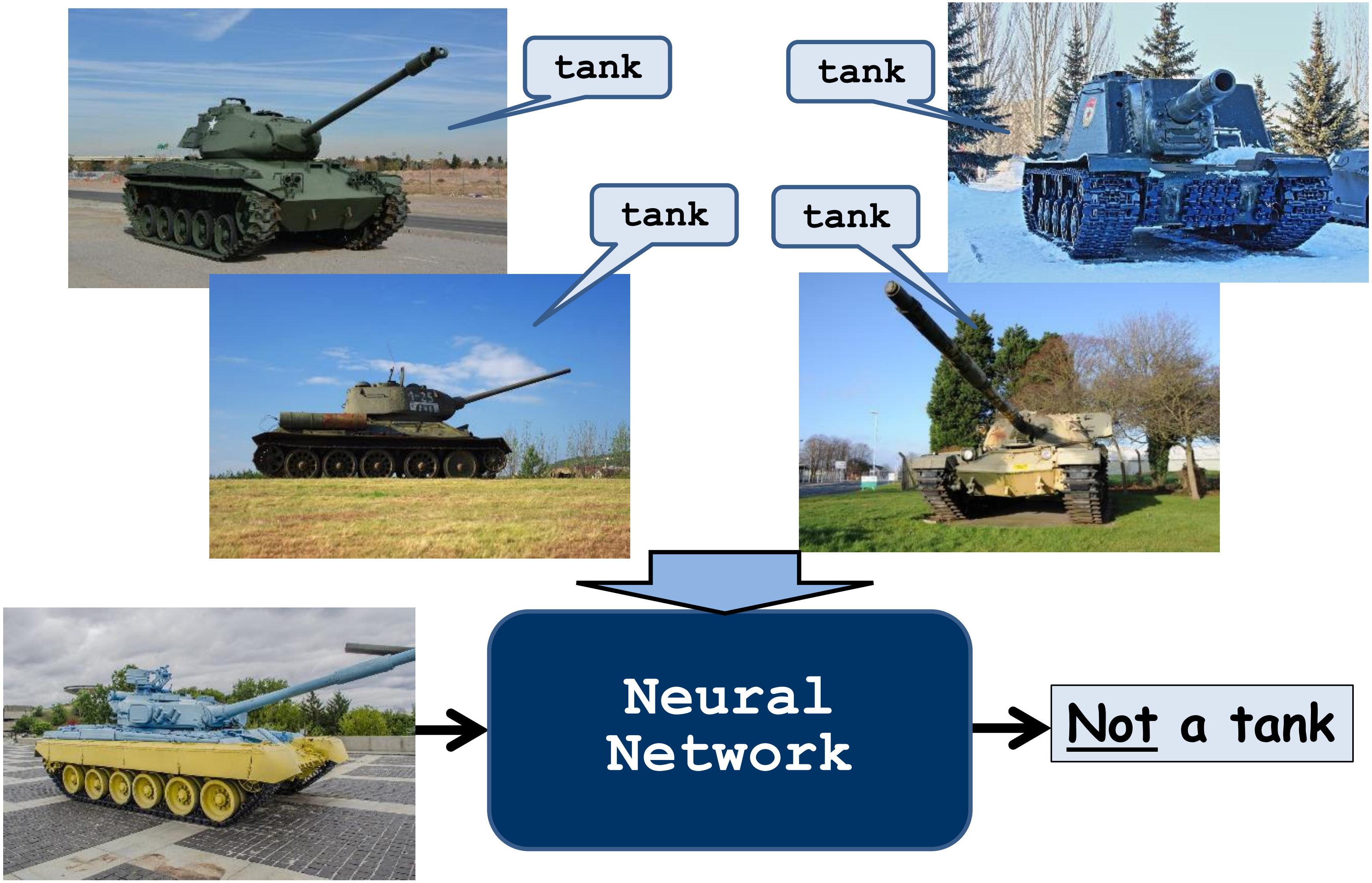


Checking the Training Set

Need to look for biased training data, overfitting, underfitting, extraneous data, outliers, etc.



Misunderstanding – Data Bias



tank

tank

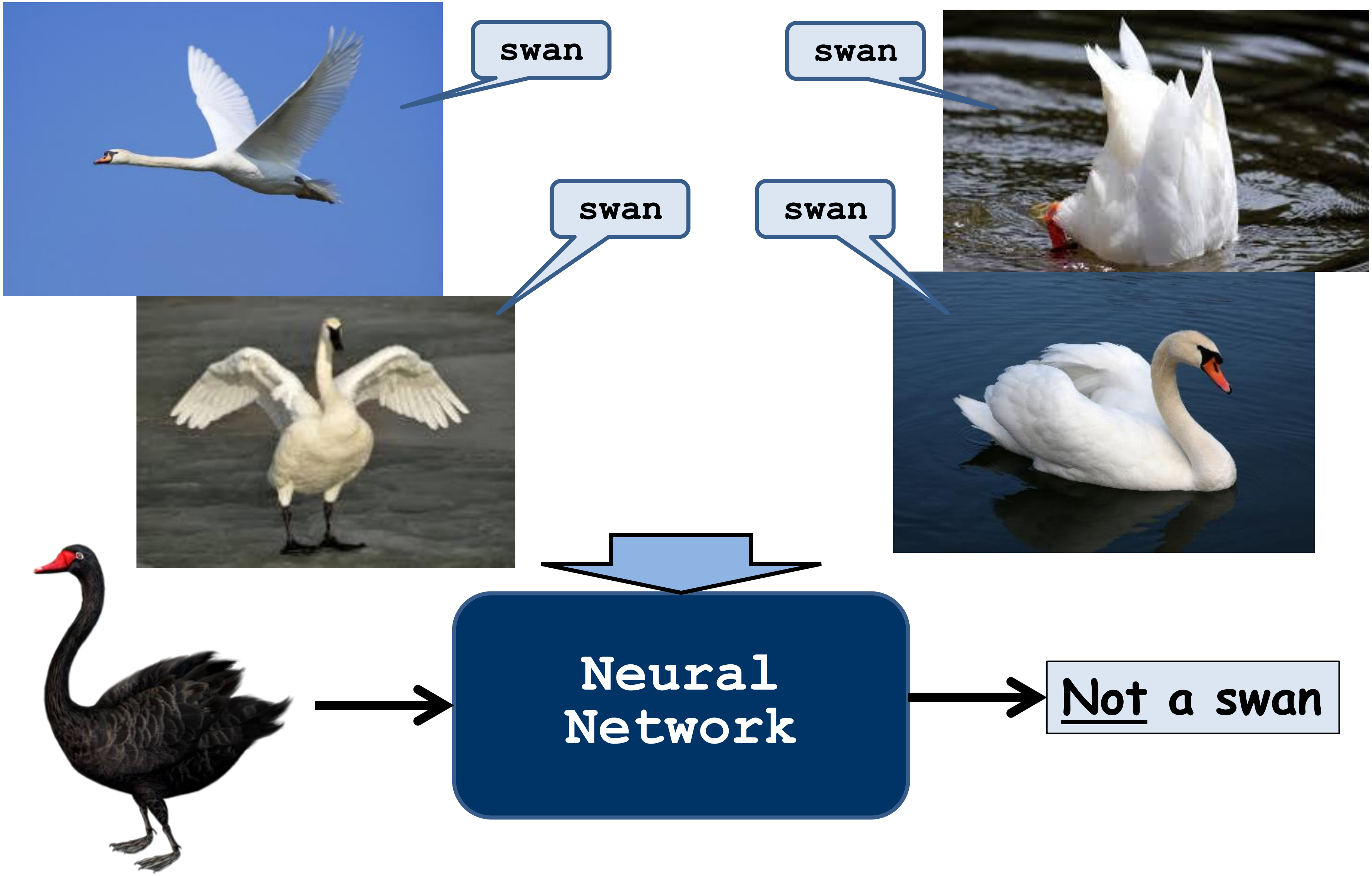
tank

tank

Neural Network

Not a tank

Incomplete Training Set



swan

swan

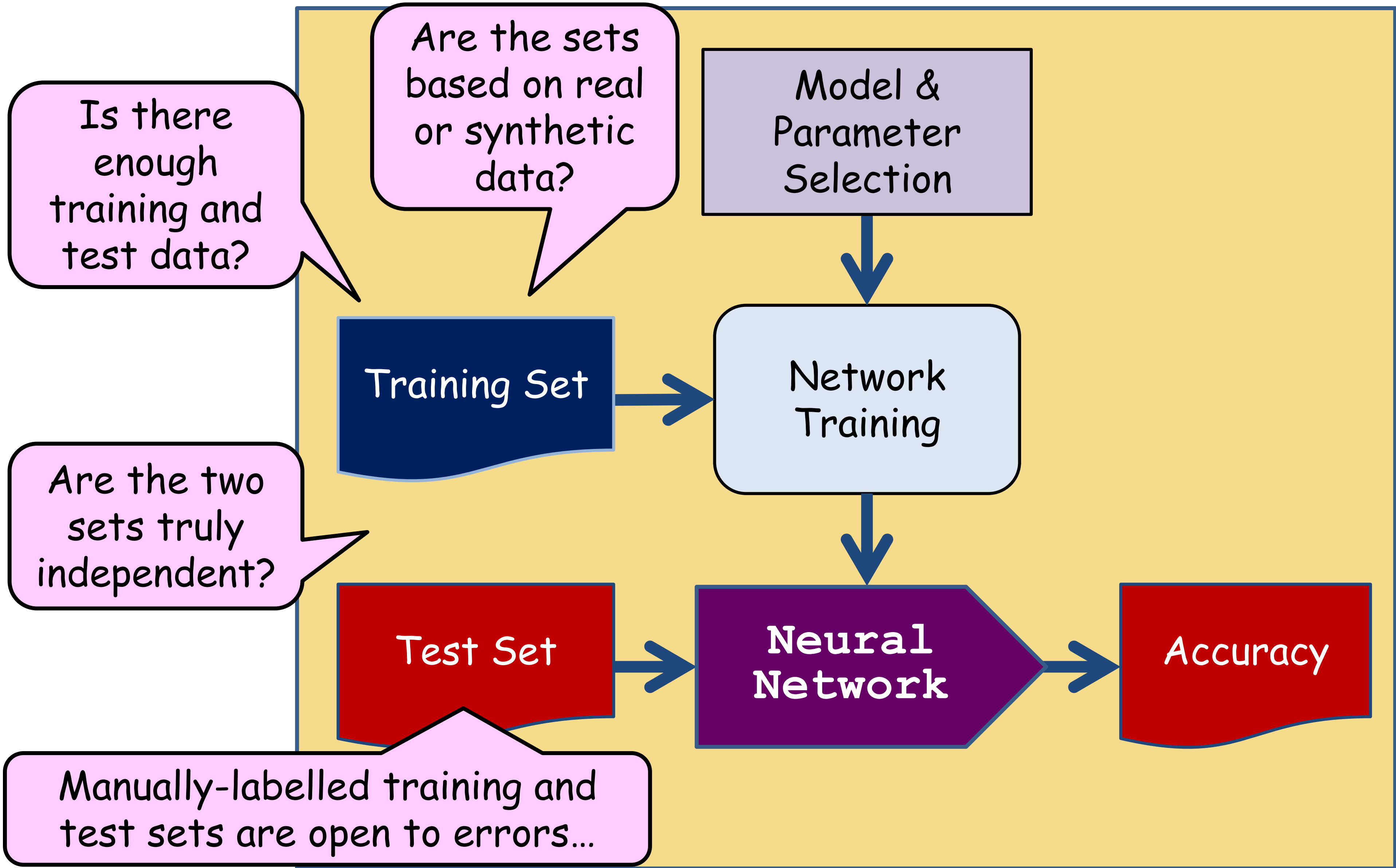
swan

swan

Neural Network

Not a swan

Checking the Training & Test Sets



Is there enough training and test data?

Are the sets based on real or synthetic data?

Model & Parameter Selection

Training Set

Network Training

Are the two sets truly independent?

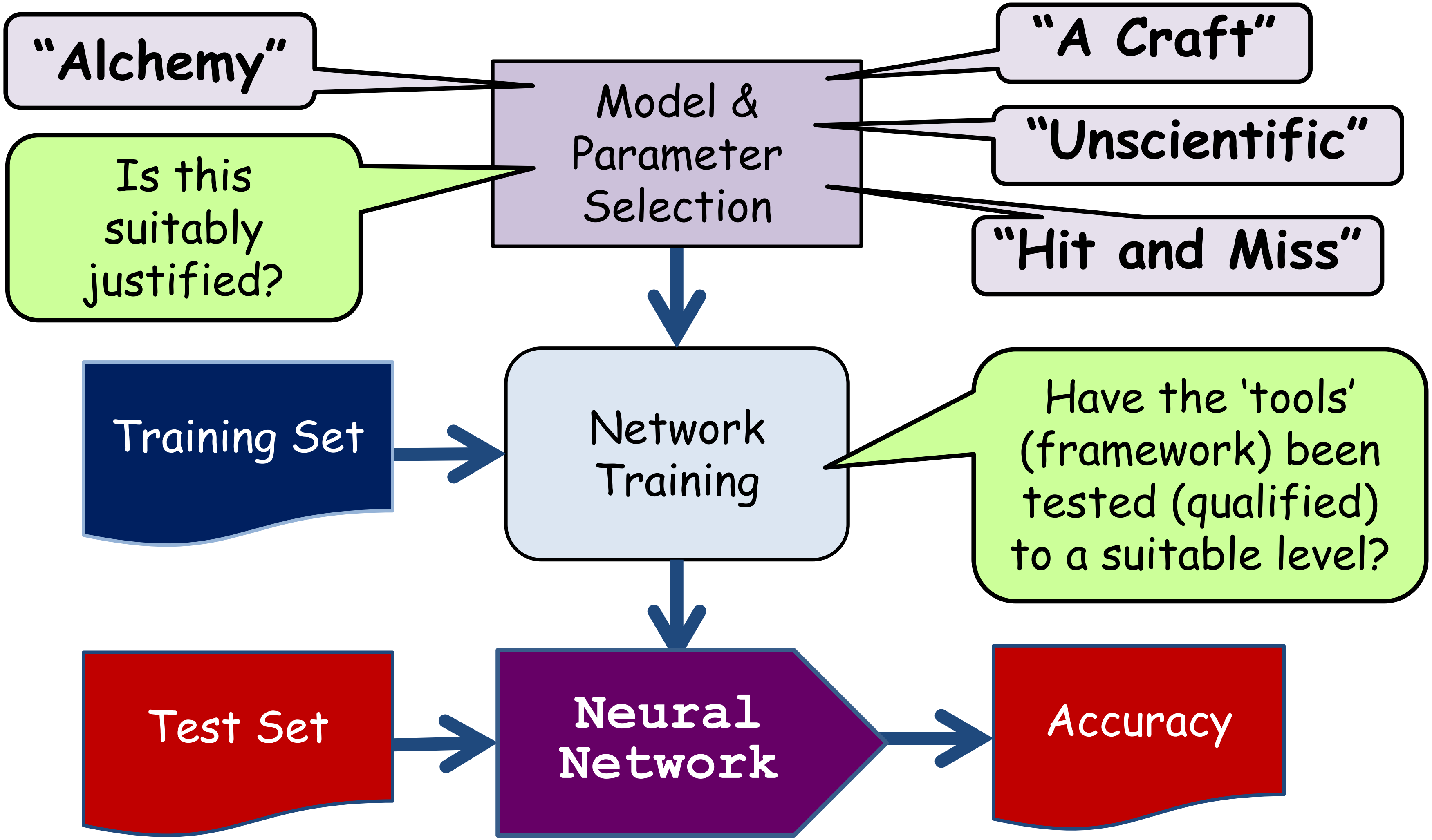
Test Set

Neural Network

Accuracy

Manually-labelled training and test sets are open to errors…

Checking the Training

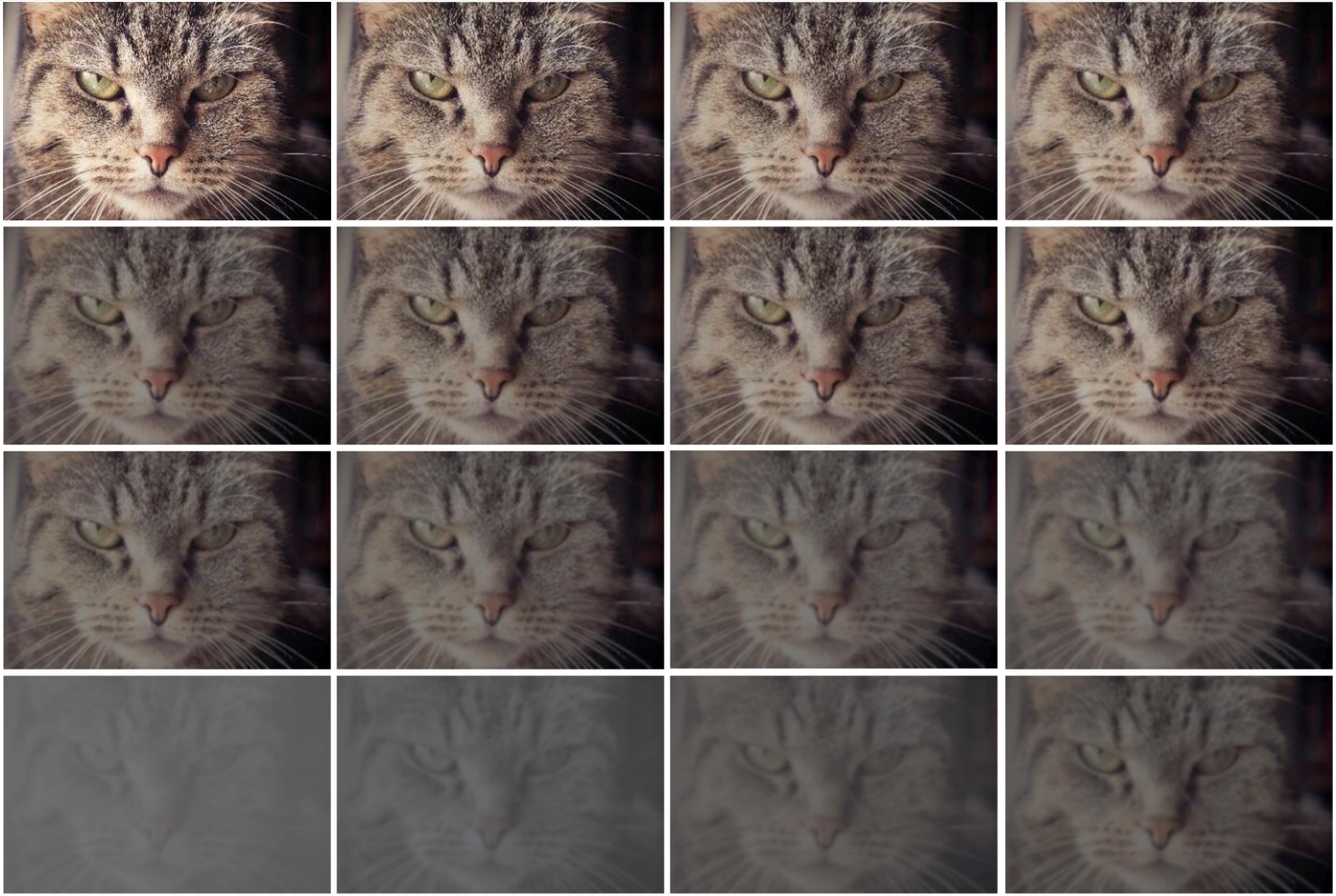

"Alchemy"
Model & Parameter Selection
"A Craft"
"Unscientific"
"Hit and Miss"
Is this suitably justified?
Training Set
Network Training
Have the 'tools' (framework) been tested (qualified) to a suitable level?
Test Set
Neural Network
Accuracy

Black Box Testing of Autonomous Systems

Test Challenges of Autonomous Systems

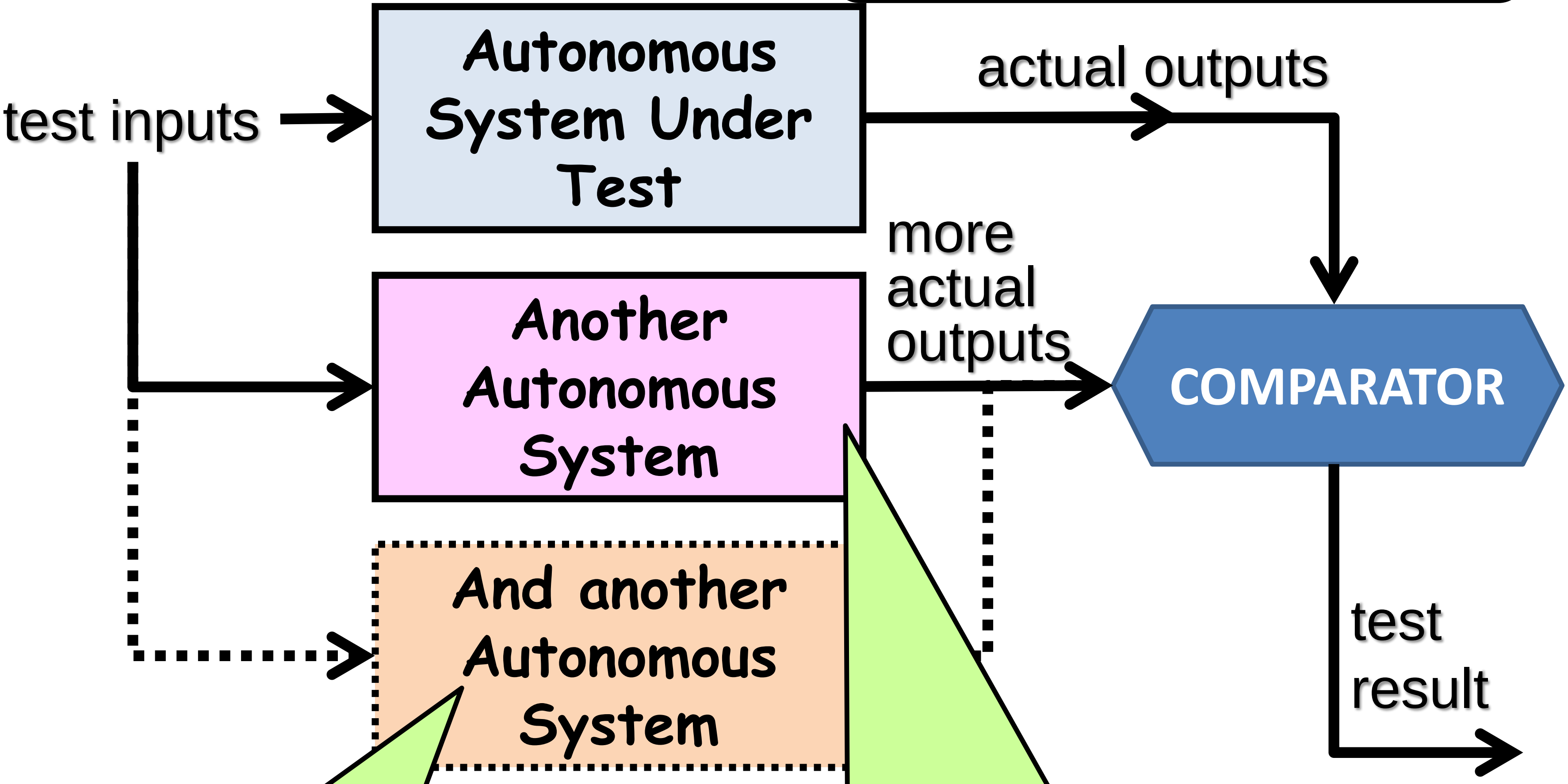
- **Expected Results (Test Oracle)**
  - if we struggle to set the objectives, then determining expected results will be equally difficult
- **Probabilistic Systems and Non-Determinism**
  - the probabilistic nature means that predicting expected results is difficult
    - we need many more tests to 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

A partial solution to the oracle problem

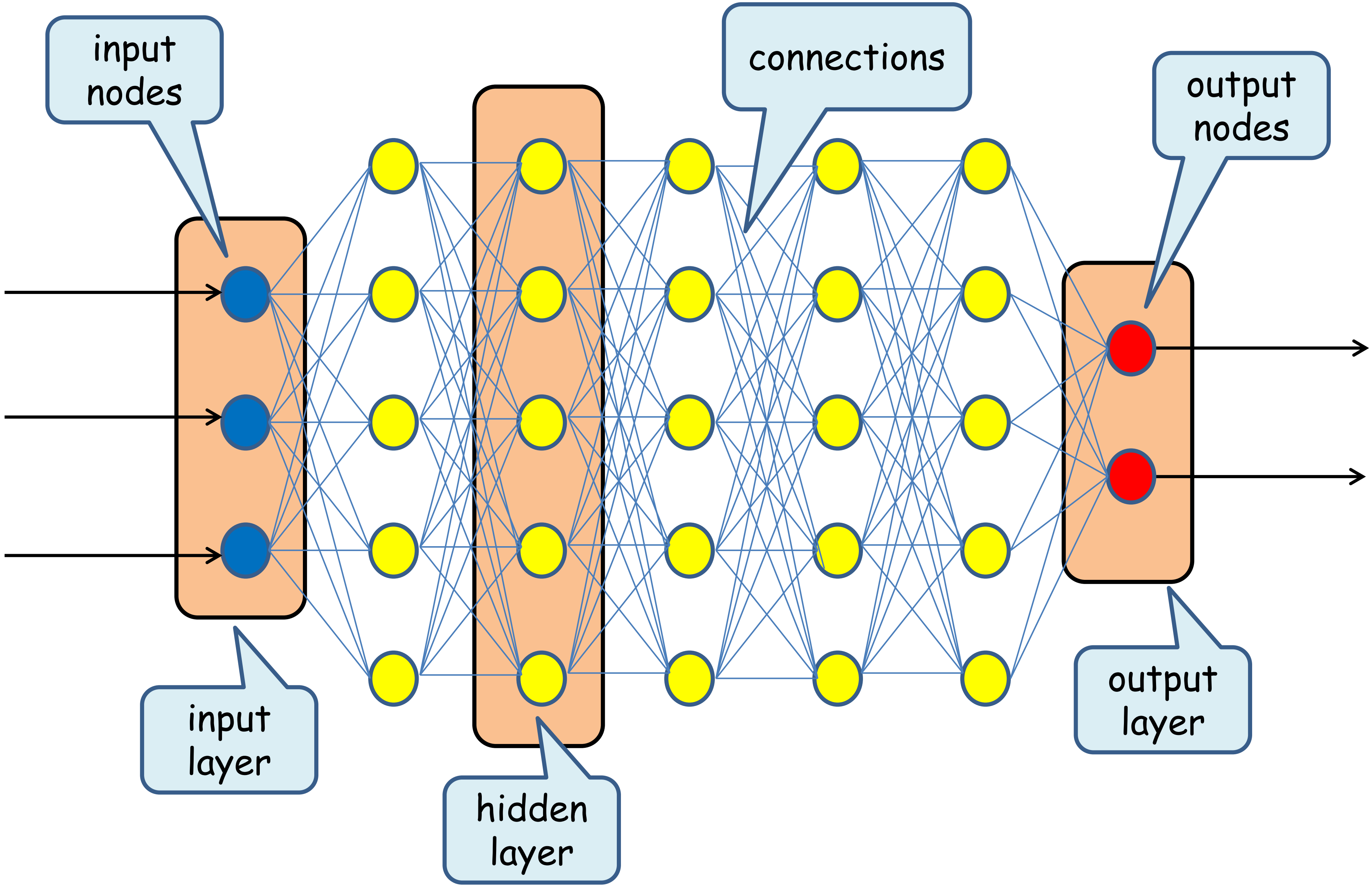

test inputs
Autonomous System Under Test
actual outputs
Another Autonomous System
more actual outputs
COMPARATOR
And another Autonomous System
test result

More criticality → More oracles

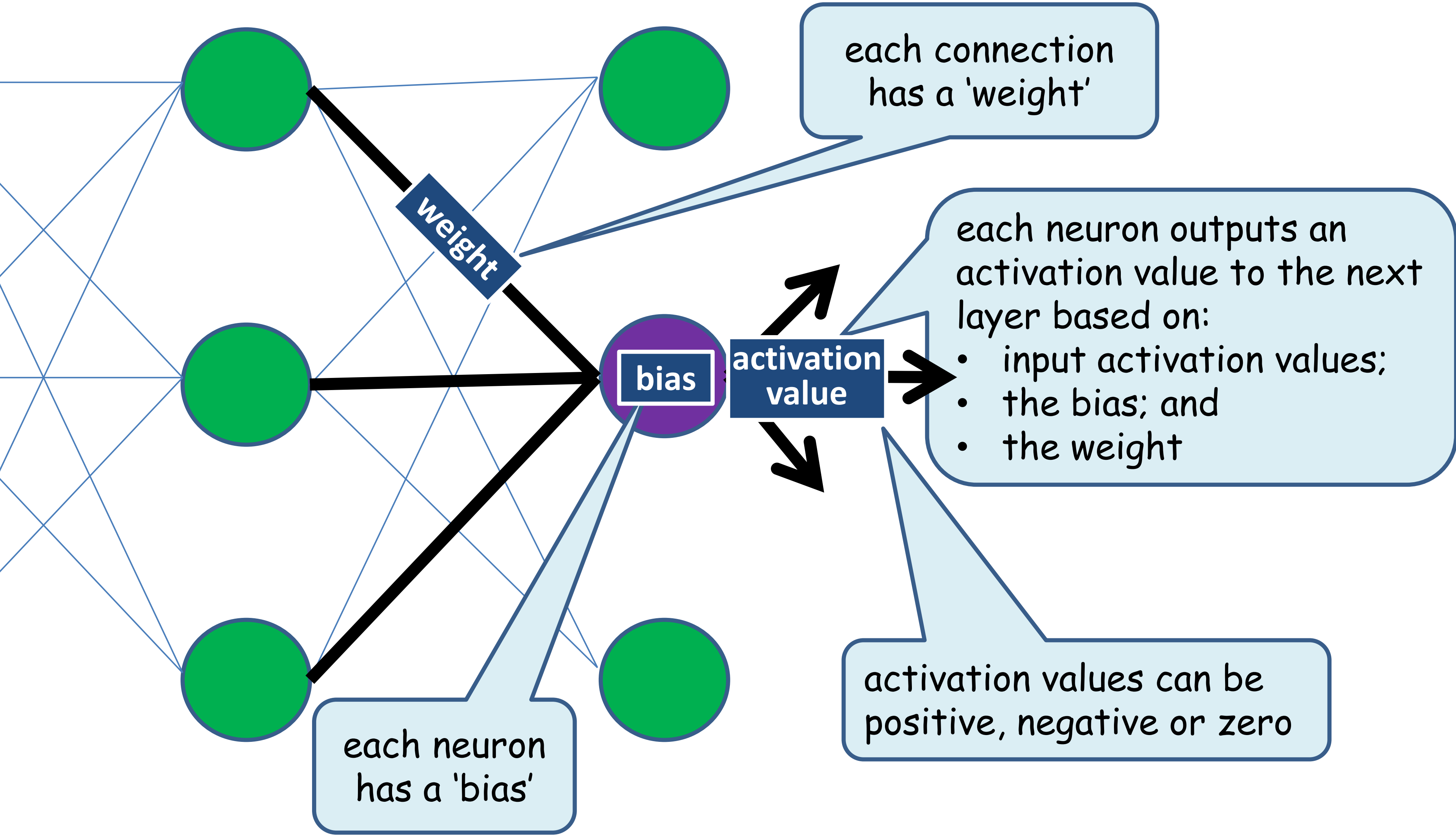
Independently-developed (using a different framework)

White Box Testing of Autonomous Systems

Deep Neural Net



Activation Values



each connection has a 'weight'

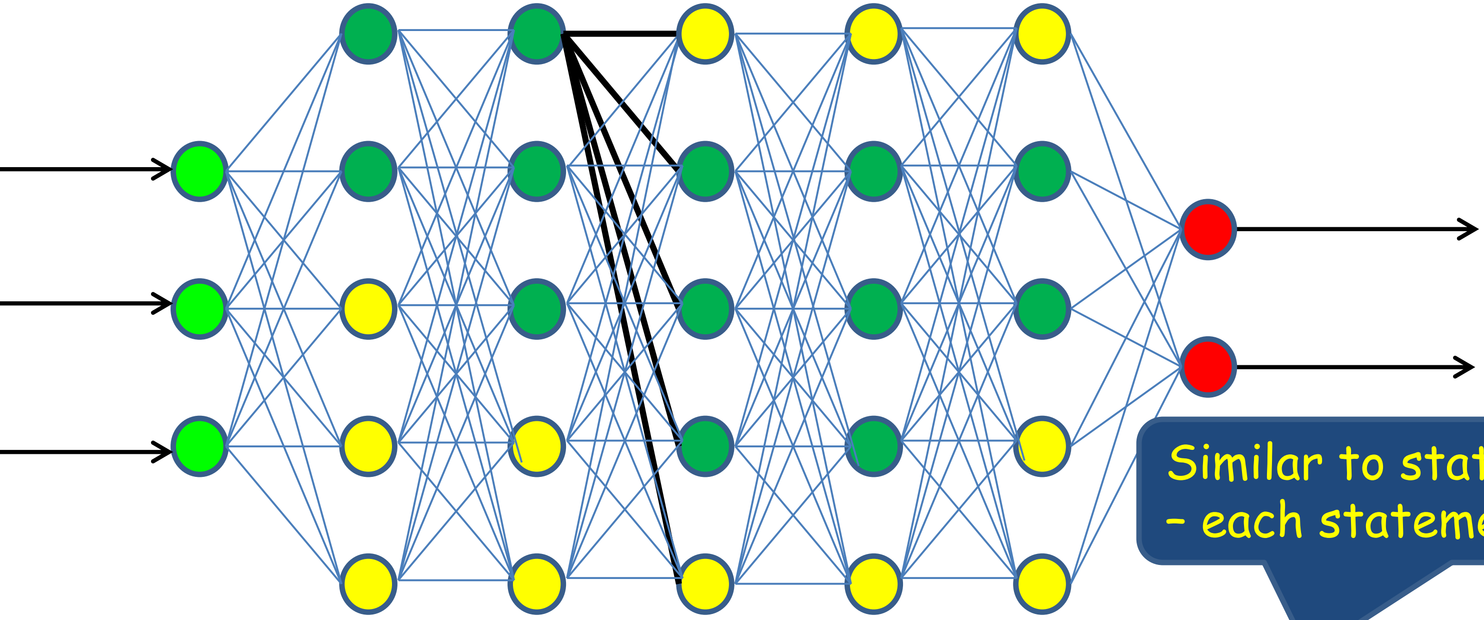
each neuron outputs an activation value to the next layer based on:

- input activation values;
- the bias; and
- the weight

each neuron has a 'bias'

activation values can be positive, negative or zero

‘Neuron’ Coverage



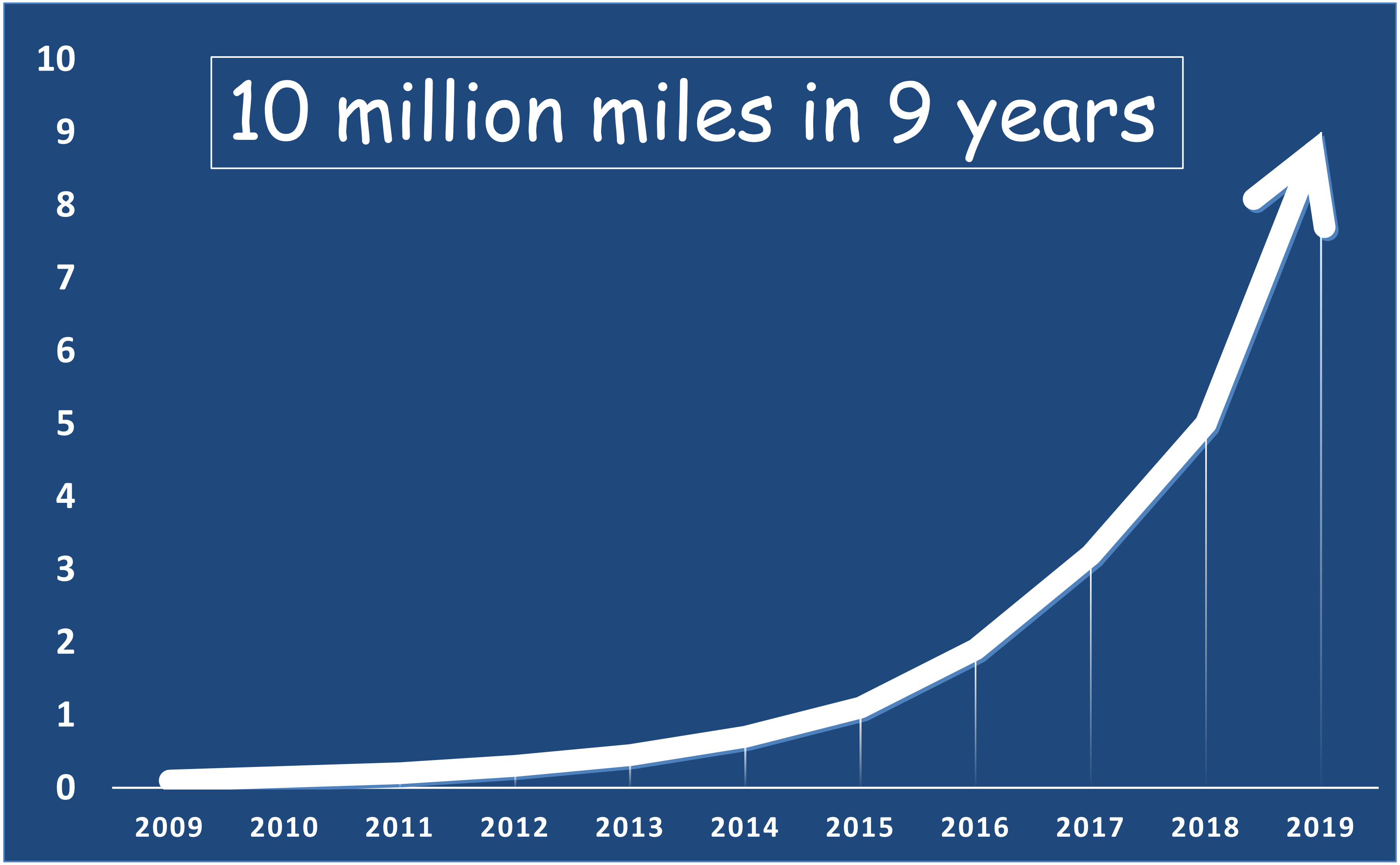
activation value is above zero

Similar to statement coverage – each statement is exercised

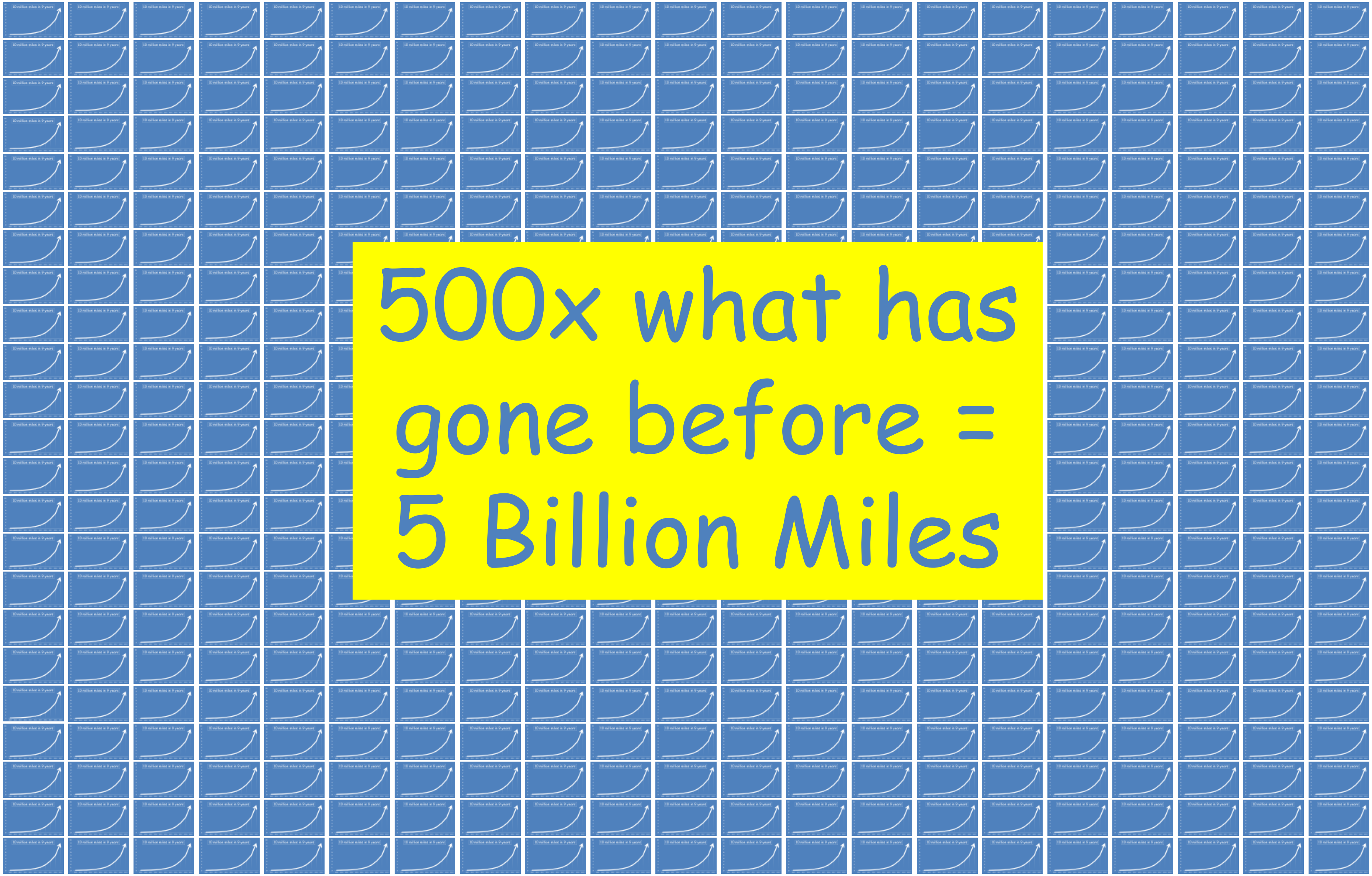
Full ‘neuron’ coverage shows that every neuron is ‘activated’ (value above zero) at least once (but very basic coverage – easy to achieve with a few tests)

The Necessity of Virtual Test Environments

Waymo On-Road Test Miles (millions)



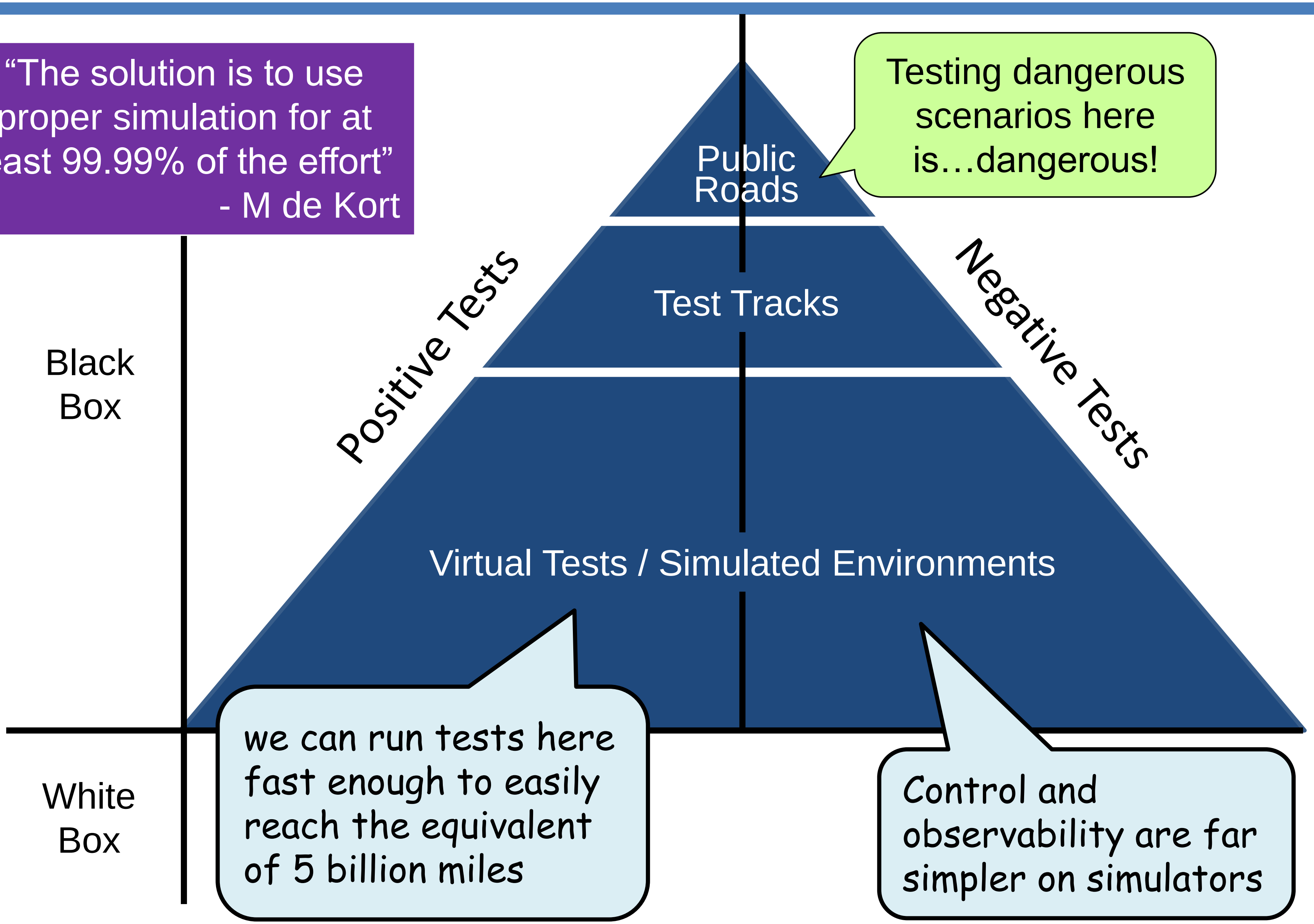
20% Better (than human drivers)



500x what has gone before = 5 Billion Miles

Autonomous Cars – Test Environments

"The solution is to use proper simulation for at least 99.99% of the effort"
- M de Kort



Public Roads

Test Tracks

Virtual Tests / Simulated Environments

Positive Tests

Negative Tests

Black Box

White Box

Testing dangerous scenarios here is…dangerous!

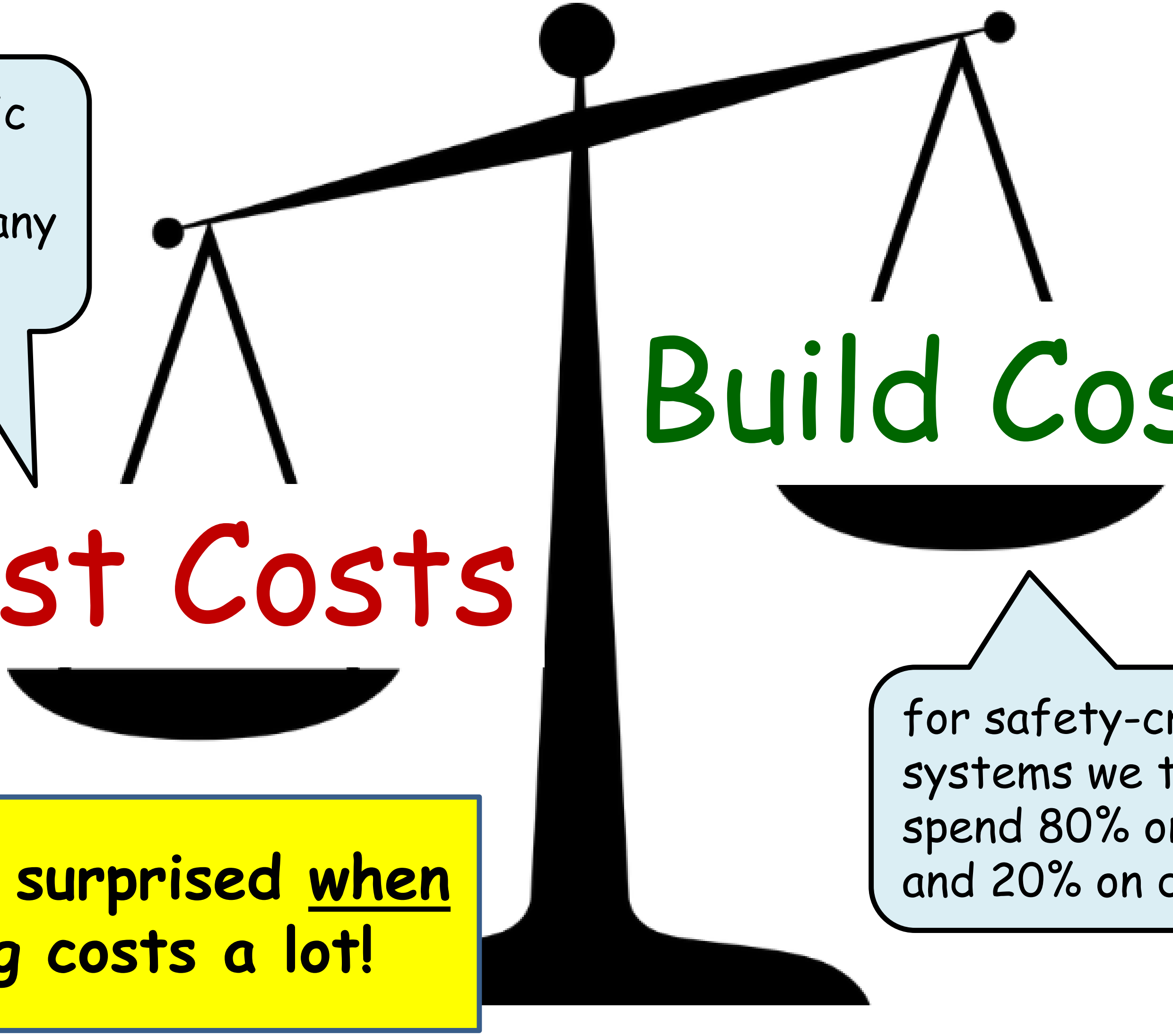
we can run tests here fast enough to easily reach the equivalent of 5 billion miles

Control and observability are far simpler on simulators

Conclusions

Autonomous System Costs

probabilistic testing requires many tests



Build Costs

Test Costs

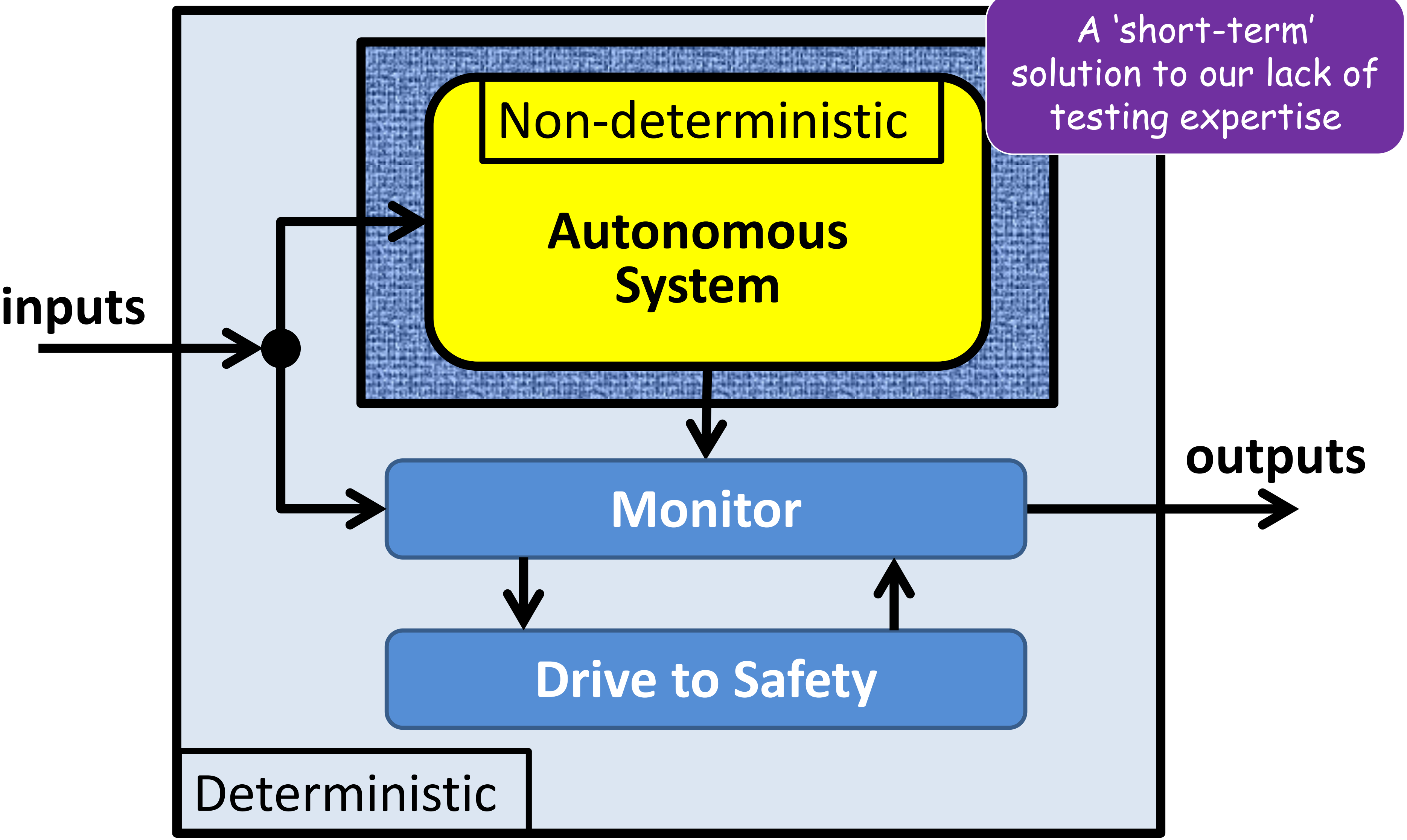
for safety-critical systems we traditionally spend 80% on testing and 20% on development

Don't be surprised when testing costs a lot!

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



A 'short-term' solution to our lack of testing expertise

Thank you for listening

Any Questions?