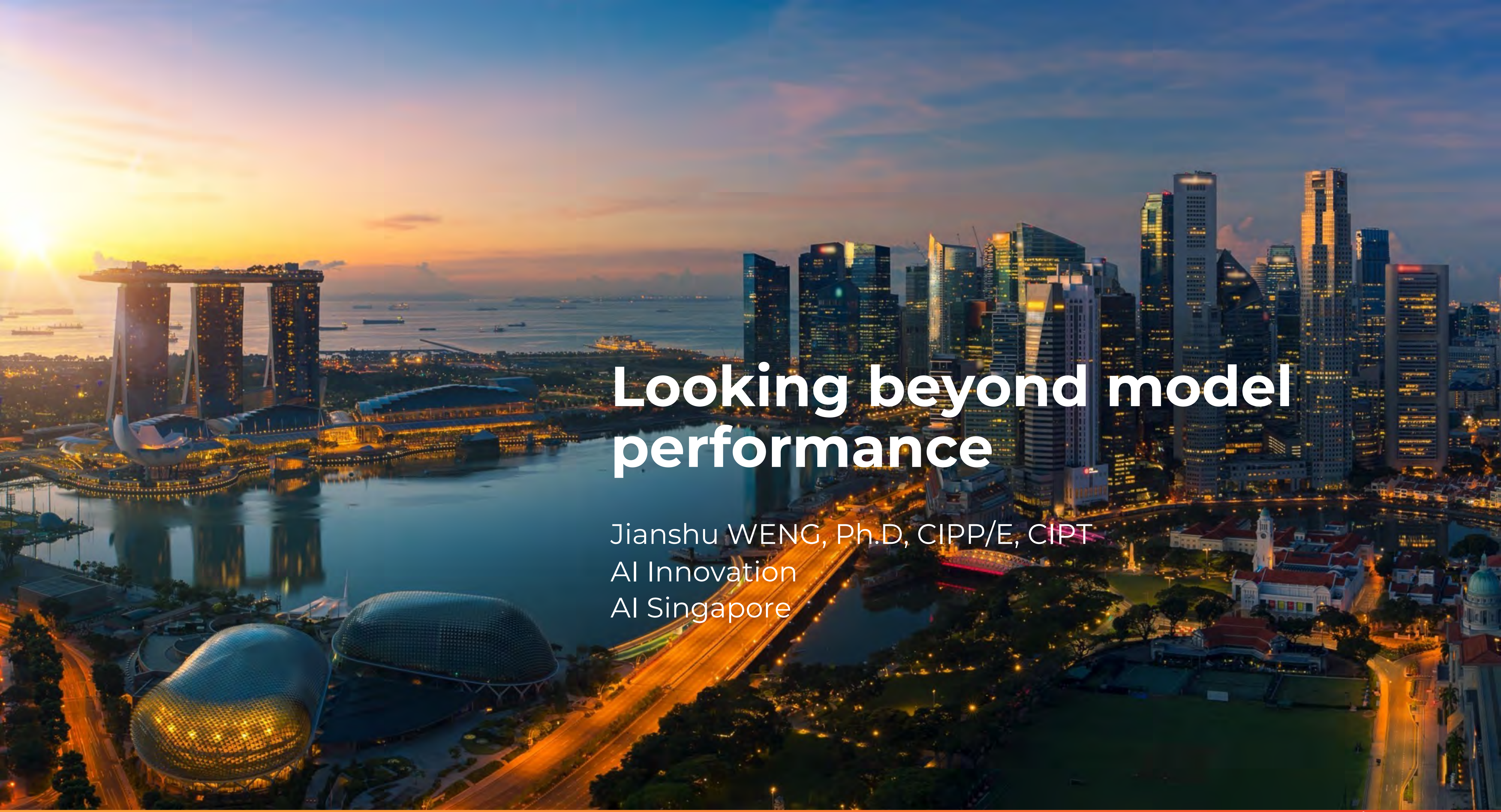
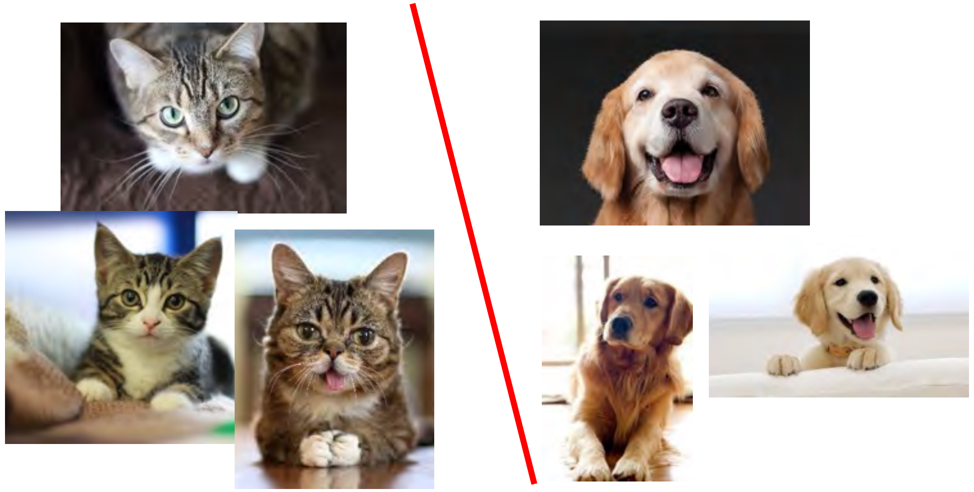


Looking beyond model performance

Jianshu WENG, Ph.D, CIPP/E, CIPT
AI Innovation
AI Singapore

AI/ML Crash Course: Supervised vs Unsupervised Learning

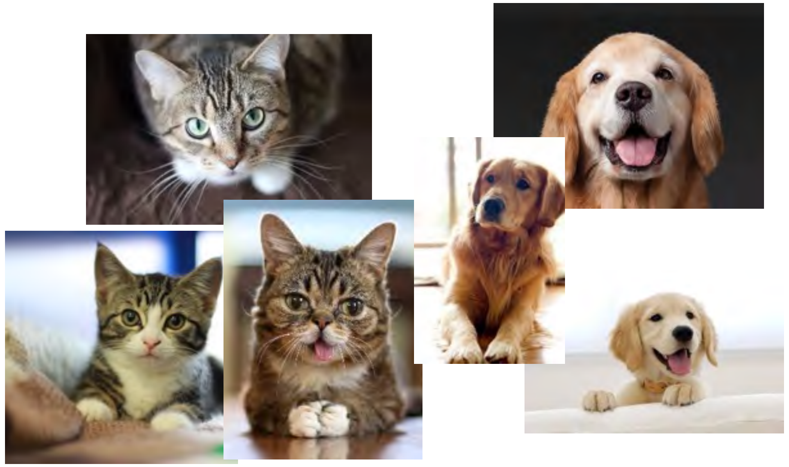
Supervised learning



This looks like the dogs I saw before, it should be a dog too

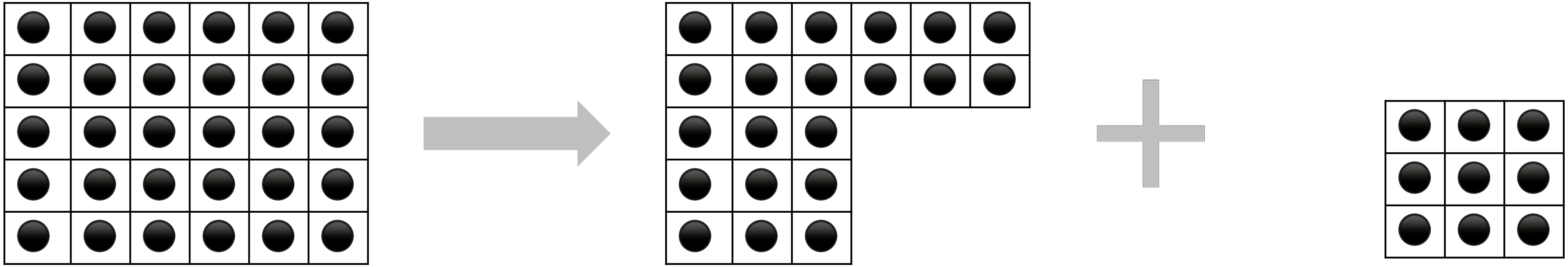
It has different names in different contexts:
- predictive modelling
- pattern recognition
- …

Unsupervised learning



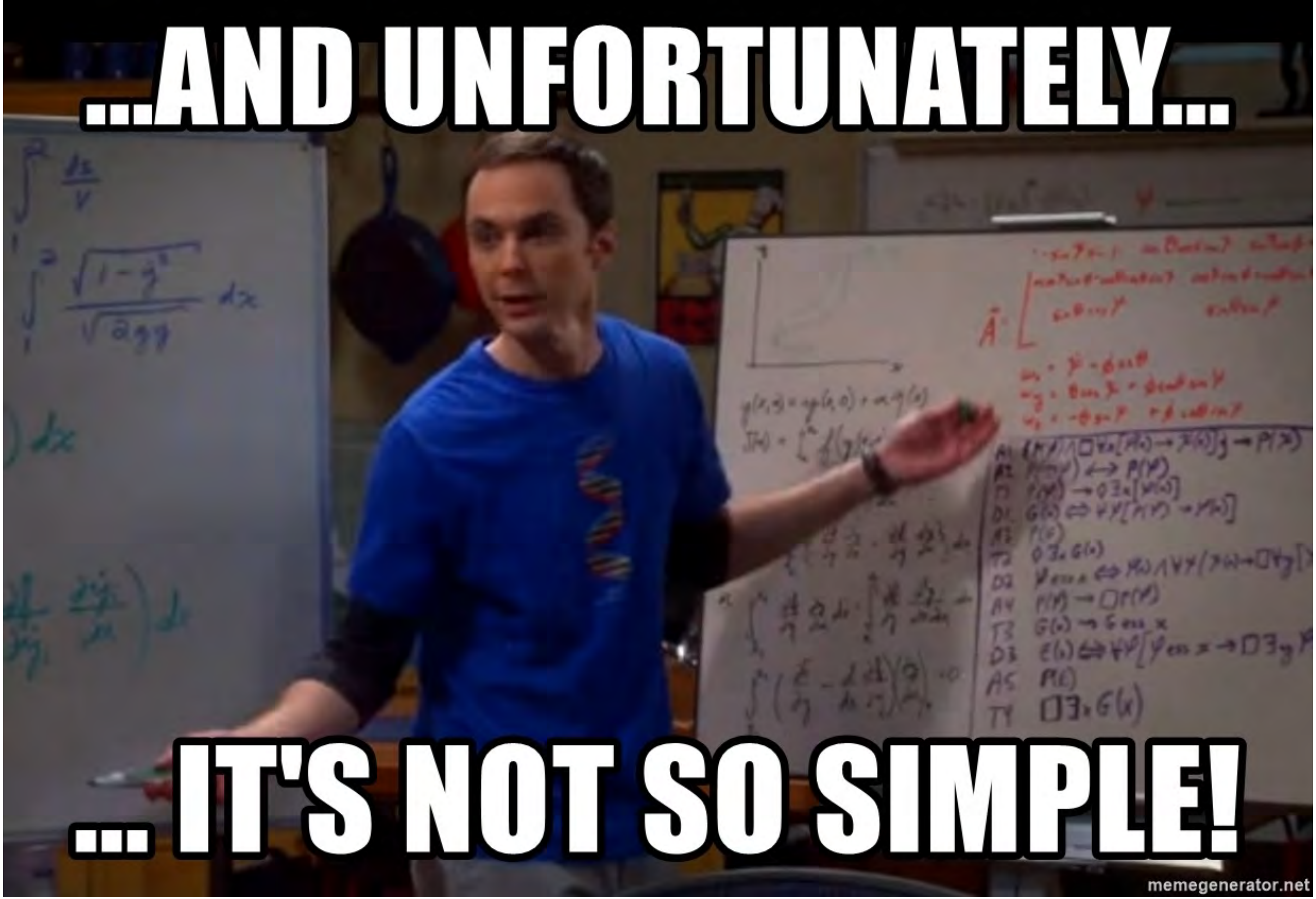
I can see there are two types of animals. But I don't know what they are. Can anybody tell me what they are?

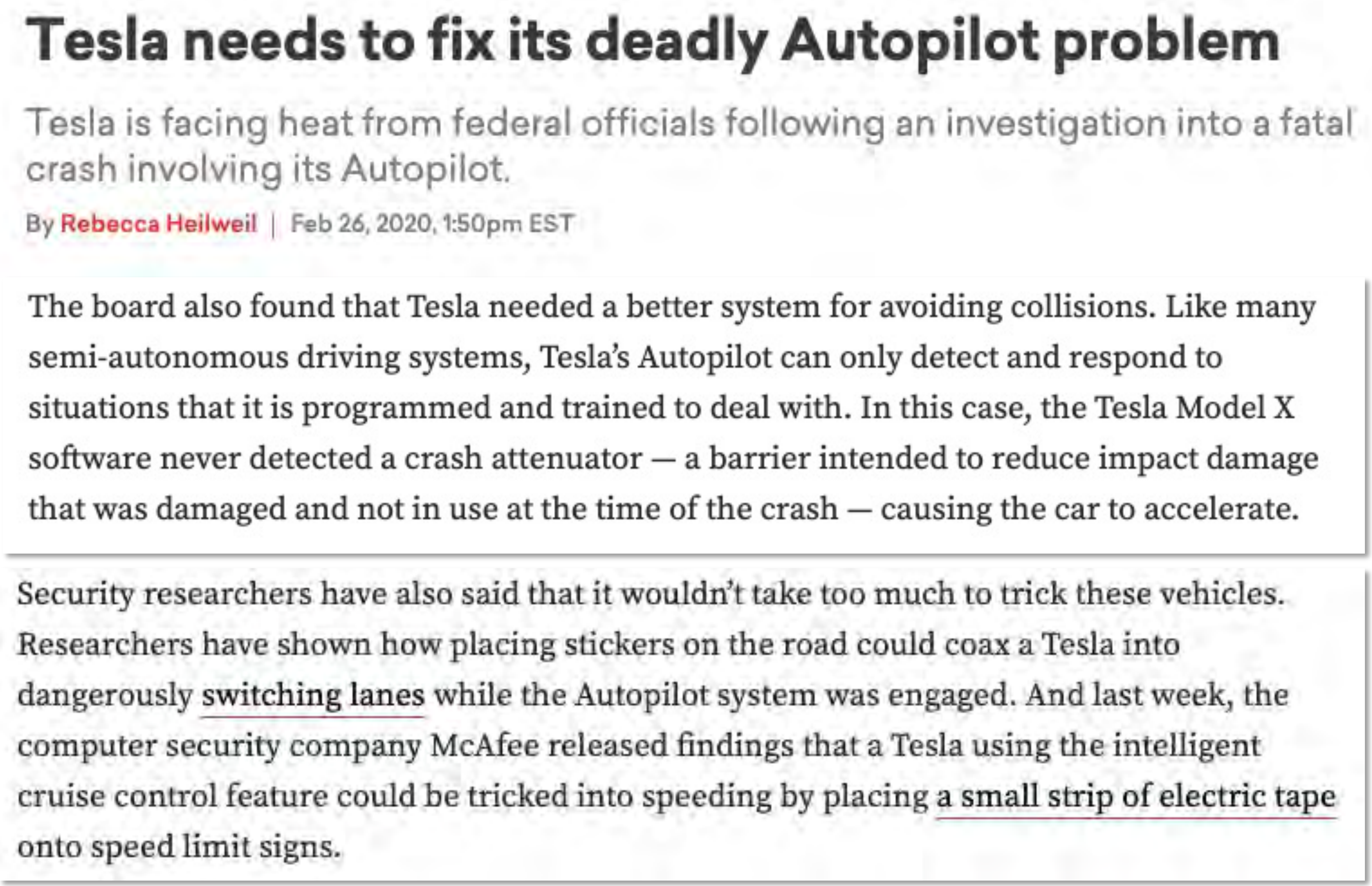
How do we evaluate model performance nowadays



training

testing/hold-out


...AND UNFORTUNATELY...
... IT'S NOT SO SIMPLE!
memegenerator.net



Tesla needs to fix its deadly Autopilot problem

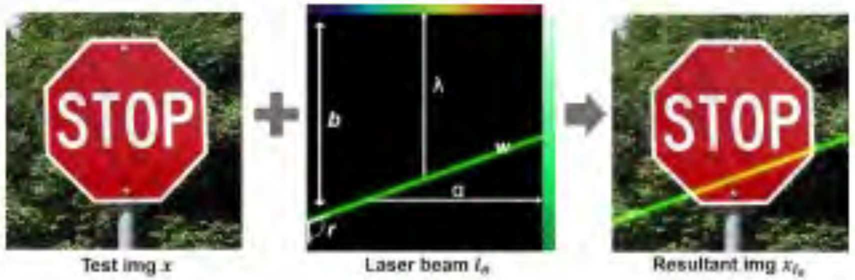
Tesla is facing heat from federal officials following an investigation into a fatal crash involving its Autopilot.

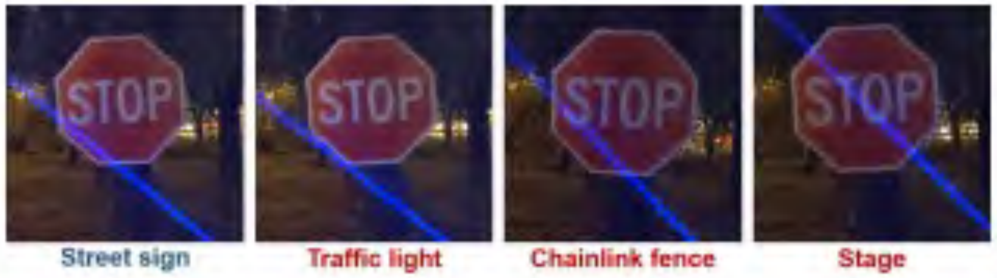
By Rebecca Heilweil | Feb 26, 2020, 1:50pm EST

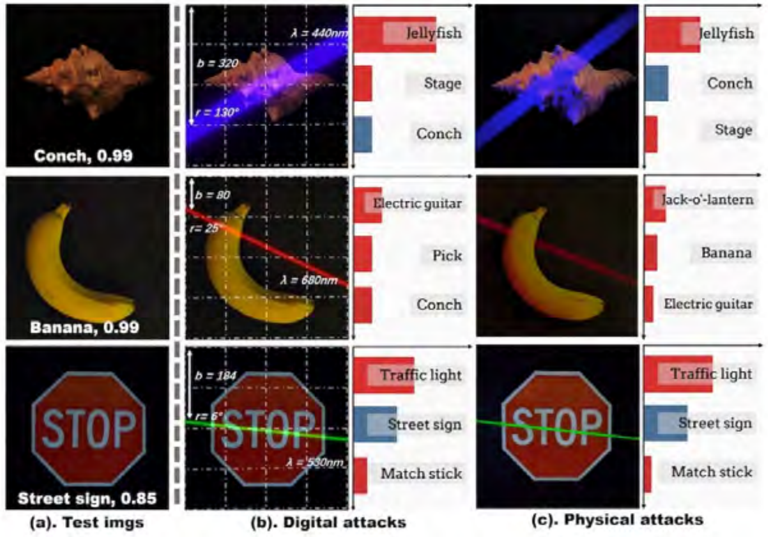
The board also found that Tesla needed a better system for avoiding collisions. Like many semi-autonomous driving systems, Tesla's Autopilot can only detect and respond to situations that it is programmed and trained to deal with. In this case, the Tesla Model X software never detected a crash attenuator — a barrier intended to reduce impact damage that was damaged and not in use at the time of the crash — causing the car to accelerate.

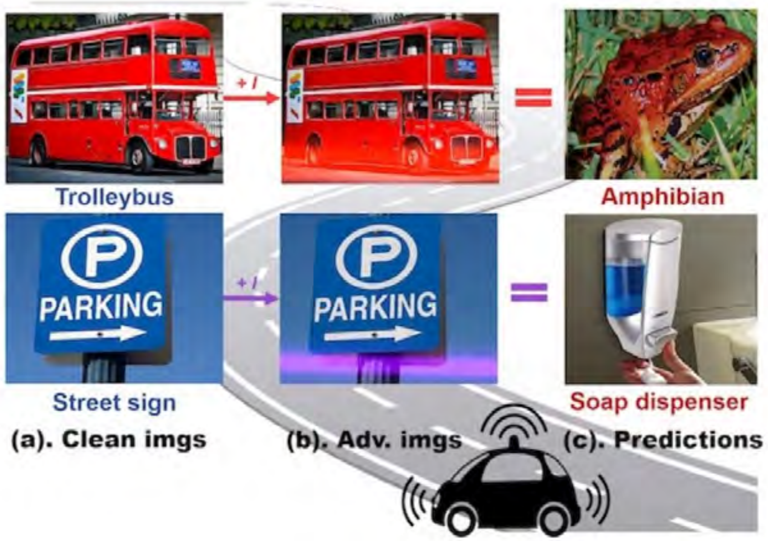
Security researchers have also said that it wouldn't take too much to trick these vehicles. Researchers have shown how placing stickers on the road could coax a Tesla into dangerously switching lanes while the Autopilot system was engaged. And last week, the computer security company McAfee released findings that a Tesla using the intelligent cruise control feature could be tricked into speeding by placing a small strip of electric tape onto speed limit signs.

A laser beam could do the trick too

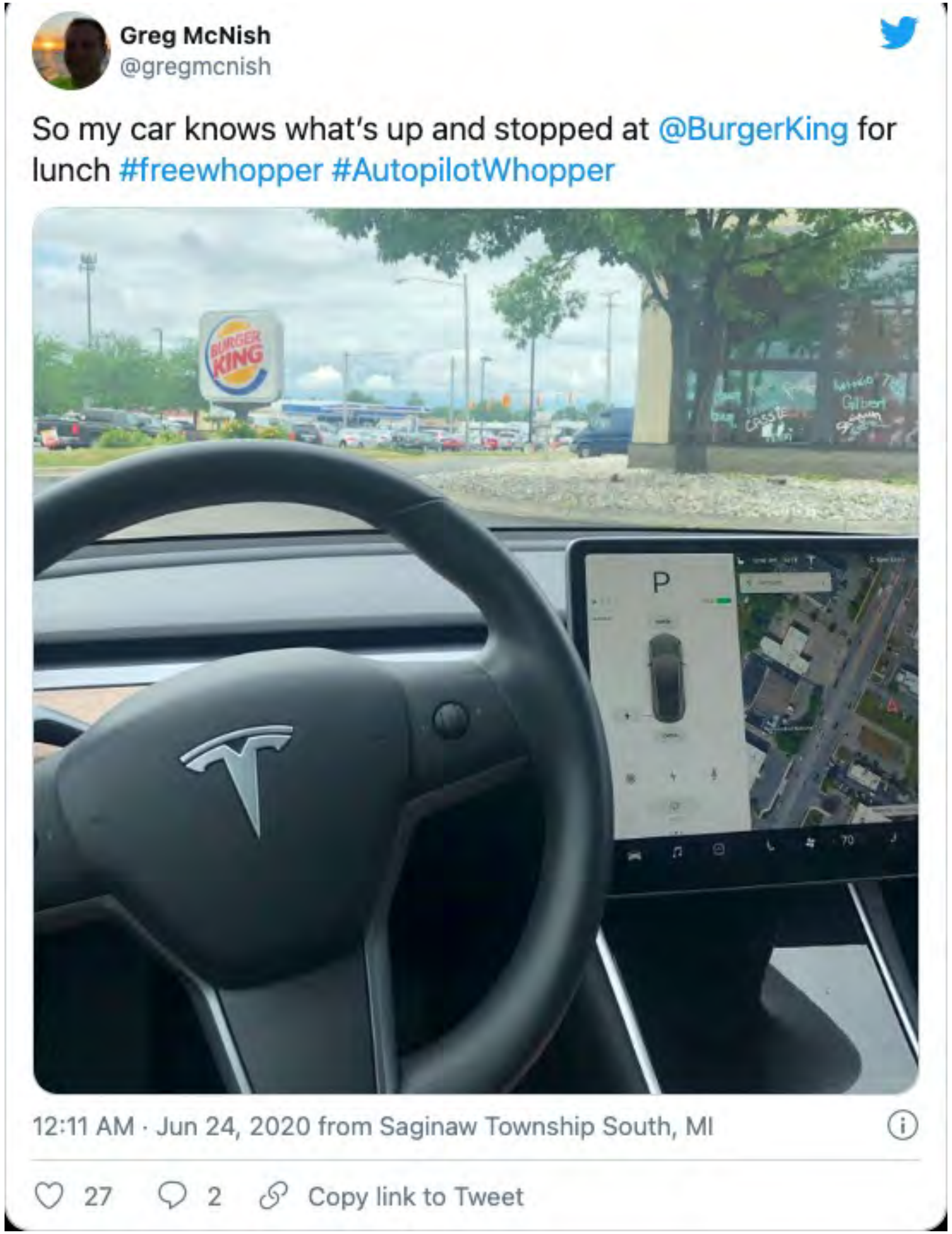








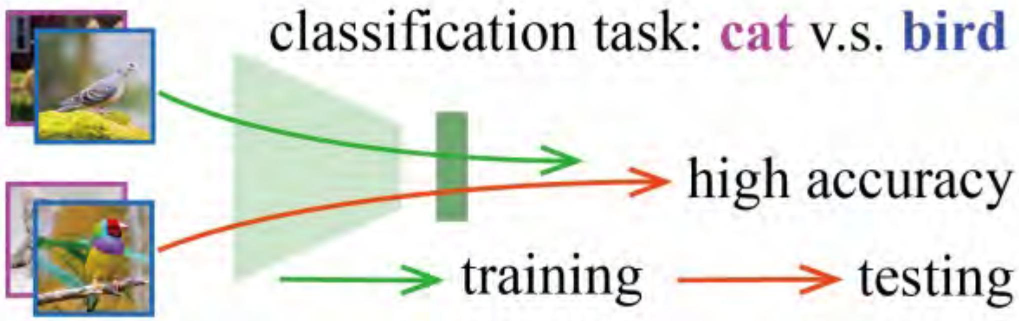
Four key parameters to define a laser beam l_θ, including wavelength (λ), layout (r, b), width (w), and intensity α. A linear image fusion method is adopted in the attack, i.e. $x_{l_\theta} = x + l_\theta$



Greg McNish
@gregmcnish

So my car knows what's up and stopped at @BurgerKing for lunch #freewhopper #AutopilotWhopper

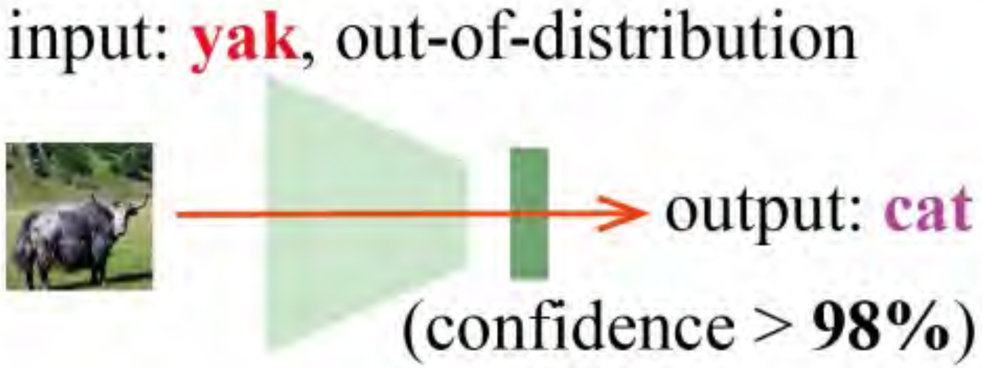
12:11 AM · Jun 24, 2020 from Saginaw Township South, MI



classification task: cat v.s. bird

high accuracy

training → testing

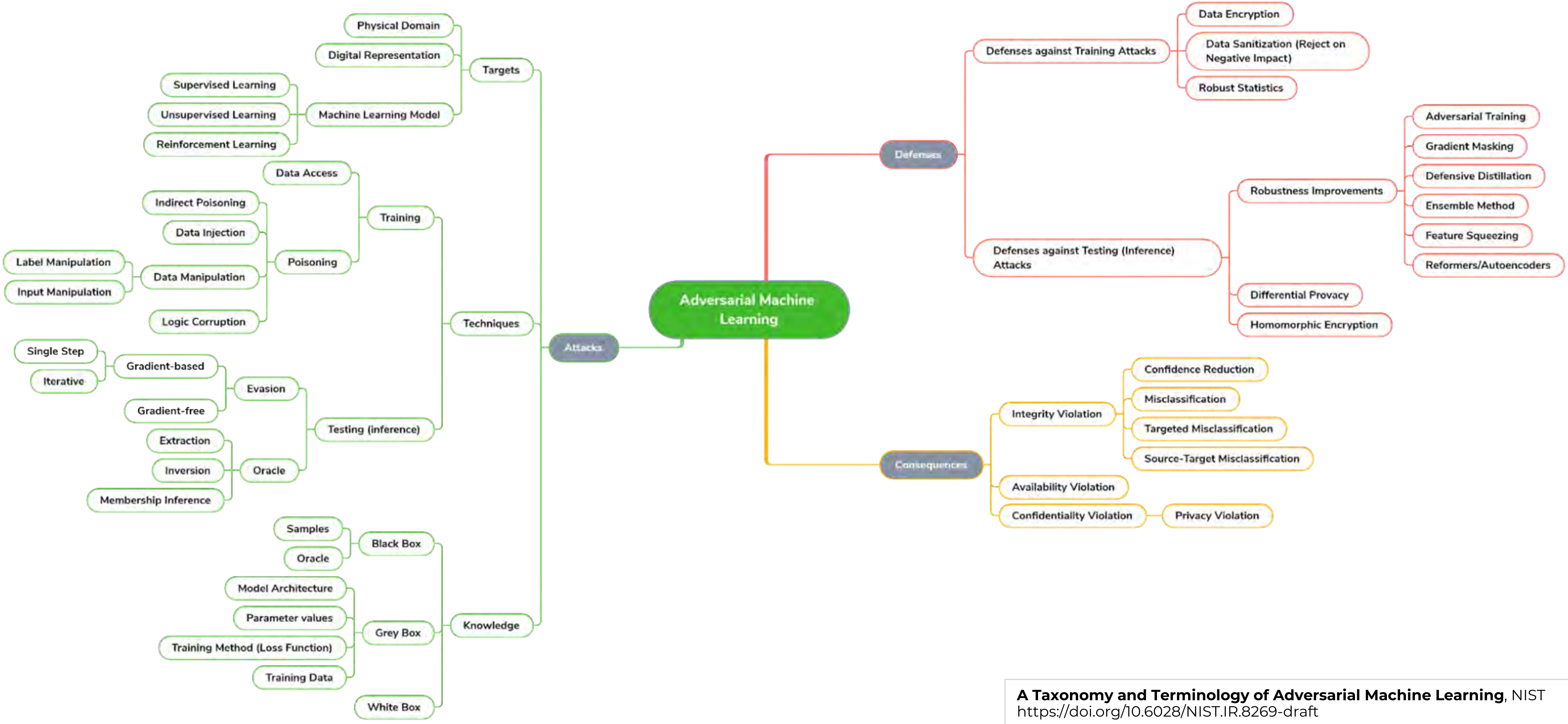


input: yak, out-of-distribution

output: cat

(confidence > 98%)

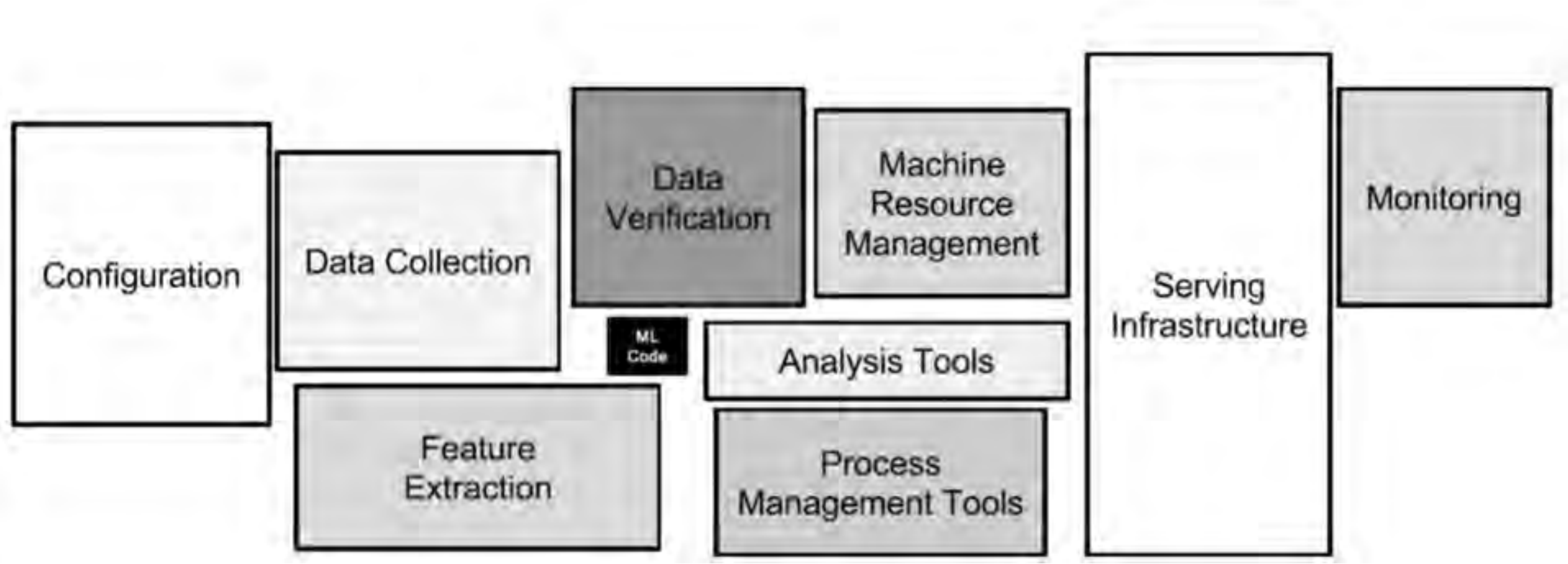
A taxonomy of Adversarial ML



- Adversarial Machine Learning
  - Attacks
    - Targets
      - Physical Domain
      - Digital Representation
      - Machine Learning Model
        - Supervised Learning
        - Unsupervised Learning
        - Reinforcement Learning
    - Techniques
      - Training
        - Data Access
        - Poisoning
          - Indirect Poisoning
          - Data Injection
          - Data Manipulation
            - Label Manipulation
            - Input Manipulation
          - Logic Corruption
      - Testing (inference)
        - Evasion
          - Gradient-based
            - Single Step
            - Iterative
          - Gradient-free
        - Oracle
          - Extraction
          - Inversion
          - Membership Inference
    - Knowledge
      - Black Box
        - Samples
        - Oracle
      - Grey Box
        - Model Architecture
        - Parameter values
        - Training Method (Loss Function)
        - Training Data
      - White Box
  - Defenses
    - Defenses against Training Attacks
      - Data Encryption
      - Data Sanitization (Reject on Negative Impact)
      - Robust Statistics
    - Defenses against Testing (Inference) Attacks
      - Robustness Improvements
        - Adversarial Training
        - Gradient Masking
        - Defensive Distillation
        - Ensemble Method
        - Feature Squeezing
        - Reformers/Autoencoders
      - Differential Provacy
      - Homomorphic Encryption
  - Consequences
    - Integrity Violation
      - Confidence Reduction
      - Misclassification
      - Targeted Misclassification
      - Source-Target Misclassification
    - Availability Violation
    - Confidentiality Violation
      - Privacy Violation

A Taxonomy and Terminology of Adversarial Machine Learning, NIST
https://doi.org/10.6028/NIST.IR.8269-draft

Only a small fraction of real-world AI/ML systems is composed of the ML code.



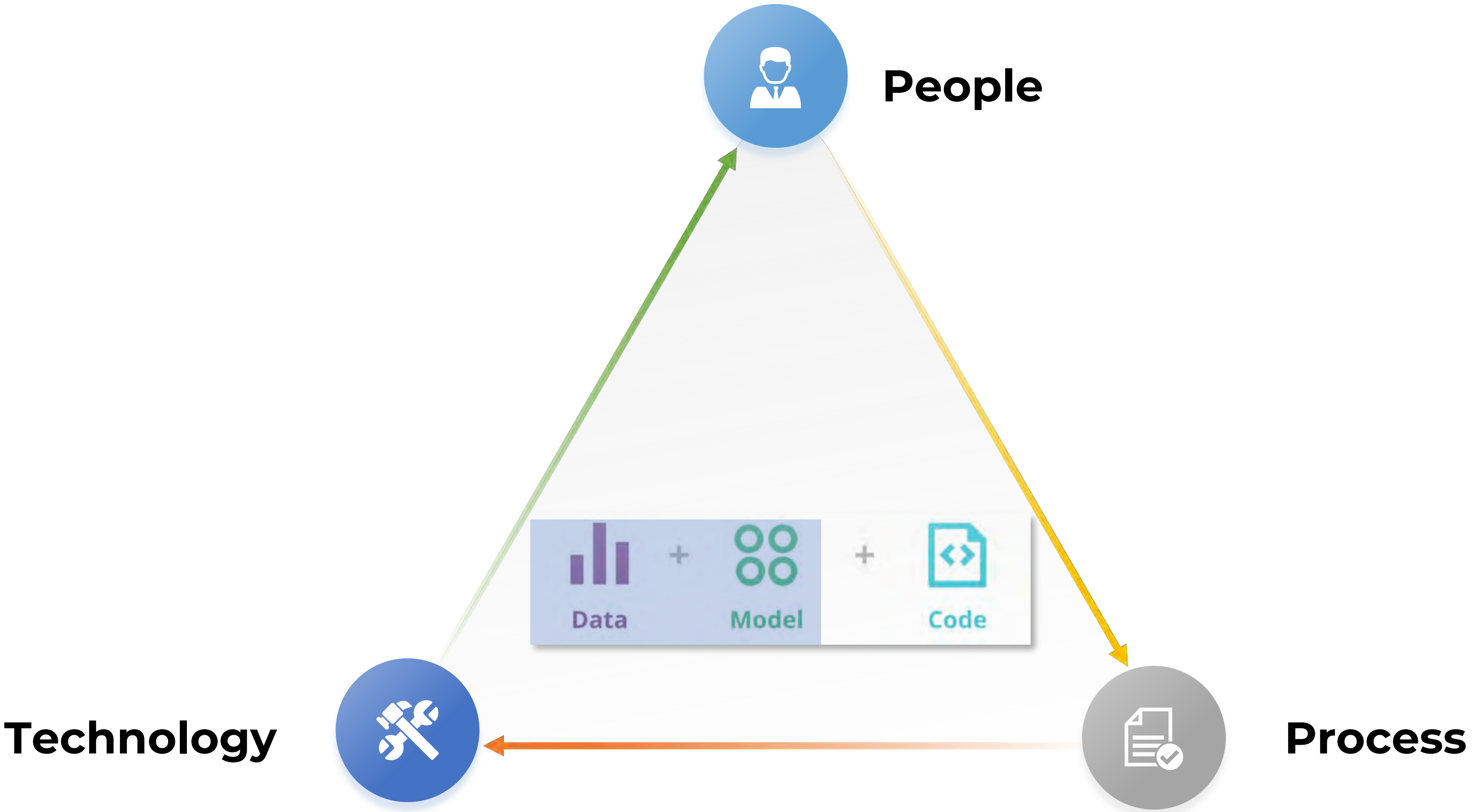
Hidden Technical Debt in Machine Learning Systems, NIPS 2015
https://dl.acm.org/doi/10.5555/2969442.2969519

Top 10 risks of ML Systems

	Data	Model	Infra	Human interaction
Adversarial examples	🟨	🟨		
Data poisoning	🟨			
Online system manipulation	🟨	🟨		🟨
Transfer learning attack		🟨		
Data confidentiality	🟨		🟨	
Data trustworthiness	🟨			
Reproducibility			🟨	
Overfitting		🟨		
Encoding integrity	🟨			
Output integrity			🟨	🟨

An Architectural Risk Analysis of Machine Learning Systems: Toward More Secure Machine Learning, BIML

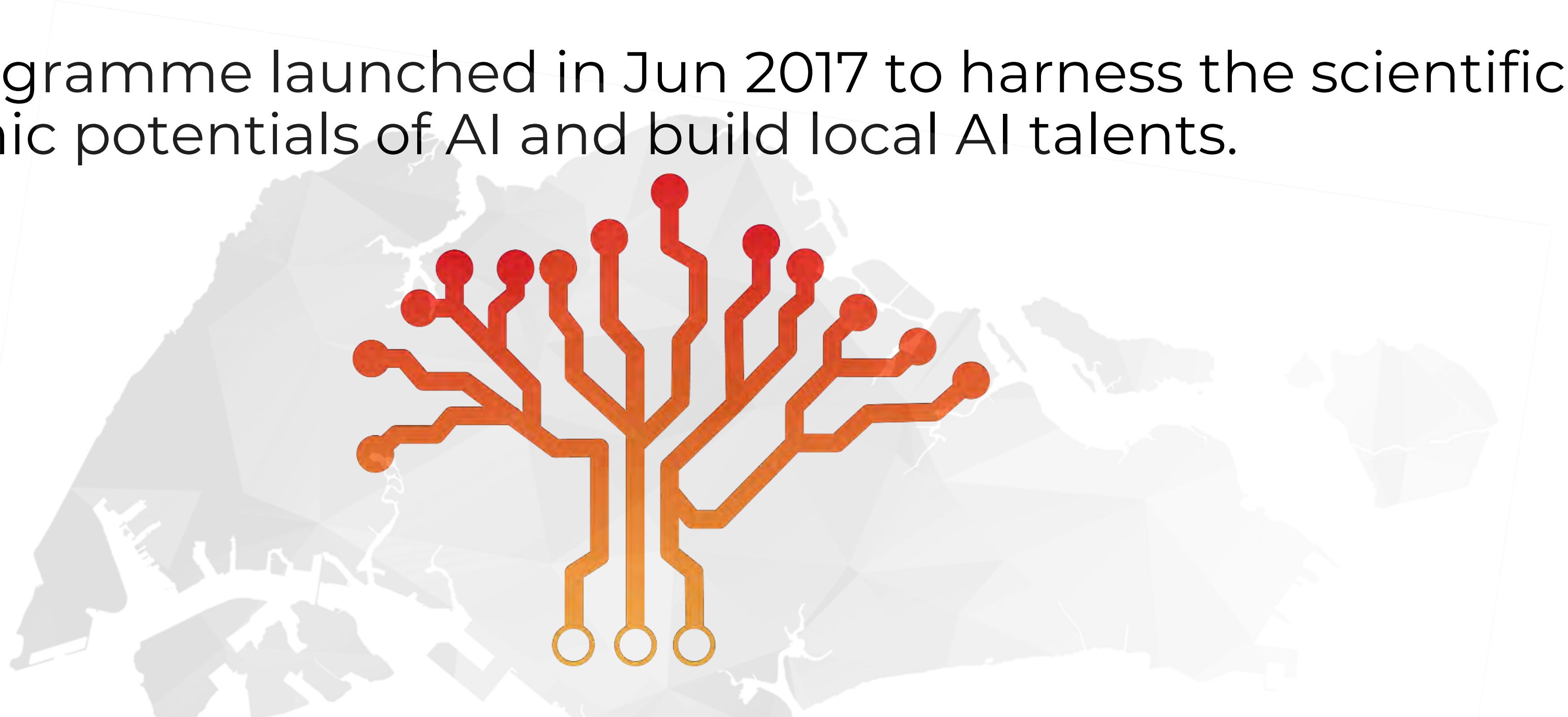
From DevSecOps to MLSecOps



What is AI Singapore doing

AI Singapore® (AISG)

National programme launched in Jun 2017 to harness the scientific and economic potentials of AI and build local AI talents.



AI SINGAPORE

NATIONAL RESEARCH FOUNDATION
PRIME MINISTER'S OFFICE
SINGAPORE

Smart Nation SINGAPORE

GOVTECH SINGAPORE

Ministry of Education SINGAPORE

MINISTRY OF HEALTH SINGAPORE

DIGITAL INDUSTRY SG

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

EDB singapore

Enterprise Singapore

Key pillars

AI RESEARCH

Invest in next gen AI strategic to SG

- Research grant calls
- PhD Fellowship Programme

AI TECHNOLOGY

Use AI to solve national challenges

- AI Grand Challenges
- Prize-based Challenges
- Technology Challenges

AI INNOVATION

Build AI capabilities and capacities for industry

- 100 Experiments
- AI Apprenticeship Programme®
- AI Data Apprenticeship Programme
- AI for Everyone®, AI for Students®, AI for Industry®, AI for Kids®
- AI Discovery Clinics

SCIENTIFIC IMPACT ⟷ ECONOMIC IMPACT

Research Publications

AI Technologies Deployed

Industry R&D Spending

Value-add to Economy

Trained and Certified Manpower

AI Jobs

Talent Programmes so far...

AI Apprenticeship Programme (AIAP)®

- 9-months full-time apprenticeship
- Monthly stipend of SGD 3,500 to 5,500
- Work alongside AI engineers and mentors to build real world deployable AI solutions
- IMDA TeSA supported
- **Only for Singaporeans**

AI Data Apprenticeship Programme

- 6-months full-time on-the-job training
- Monthly stipend of SGD 1,800
- Learn data curation, data engineering techniques and hands-on on real world AI projects
- **Only for Singaporeans**

AI for Industry (AI4I)®

- 143h fully online course
- Learn data science, machine learning, artificial intelligence and visualization in Python.
- Course fee is $224.70 for SC/SPR, $845.30 for Others

AI for Everyone (AI4E)®

- 3h fully online course
- Learn AI basics and build a simple AI model with online tools.
- FREE

AI for Students (AI4S)®

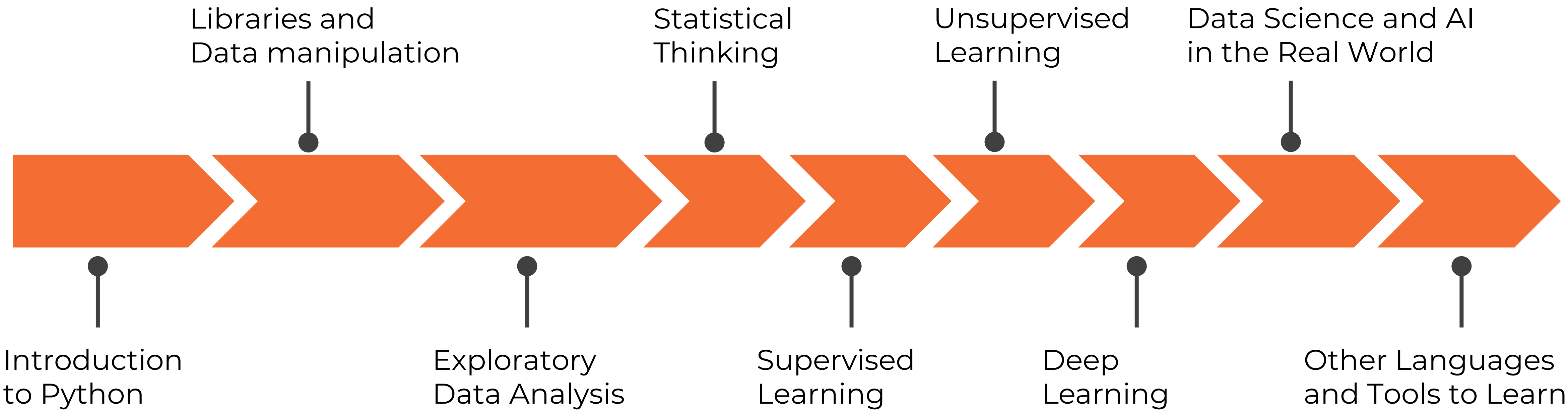
- Partnership with Datacamp to offer FREE access to all premium modules for Singapore educators from secondary school onwards
- FREE

AI for Kids (AI4K)®

- 9h blended learning including online self-learning and interactive bootcamp
- Suitable for children aged between 10 to 12 years old.
- Certified Instructor Course to teach schoolteachers and parent volunteers to deliver AI bootcamps

AI for Industry (AI4I)®

A fully online programme to help learners PLUS-skill themselves and learn data science, machine learning, artificial intelligence and visualization in Python.



Introduction to Python

Libraries and Data manipulation

Exploratory Data Analysis

Statistical Thinking

Supervised Learning

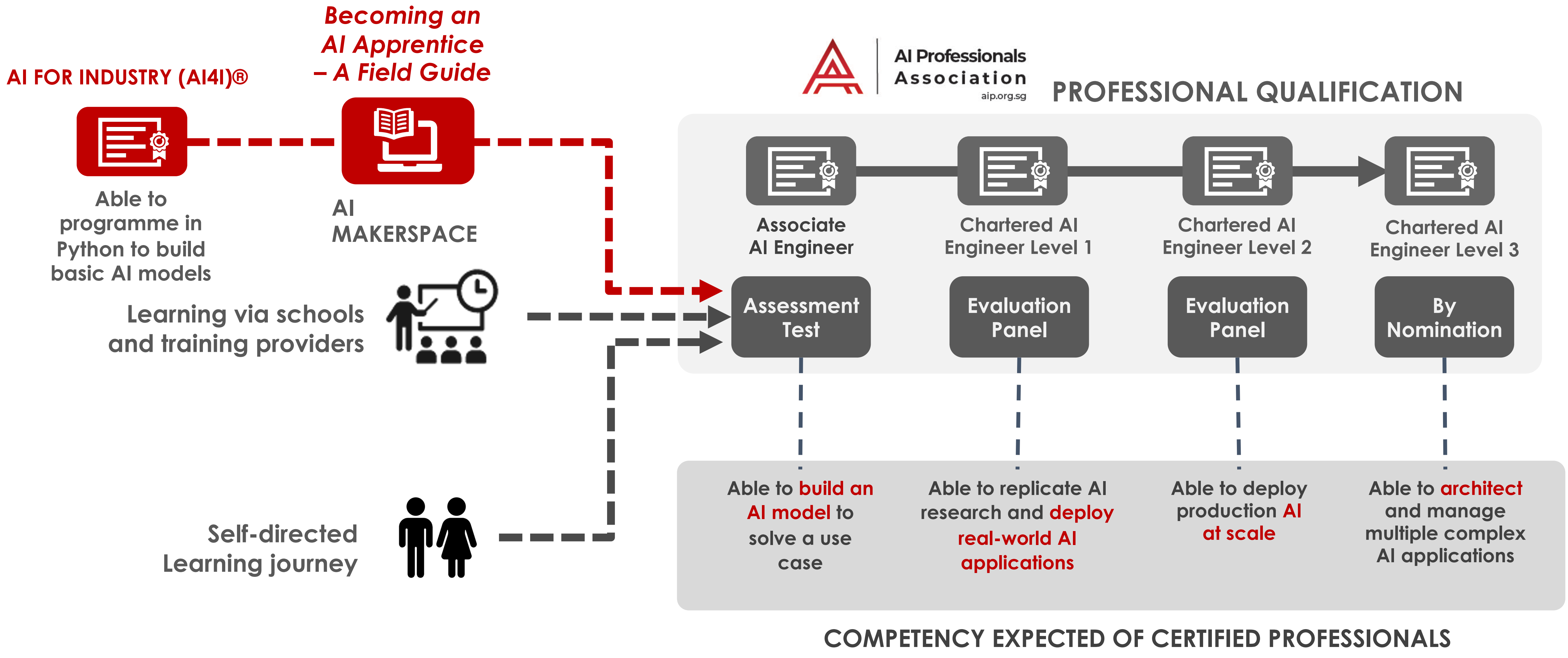
Unsupervised Learning

Deep Learning

Data Science and AI in the Real World

Other Languages and Tools to Learn

Becoming a qualified AI engineer



AI FOR INDUSTRY (AI4I)®

Able to programme in Python to build basic AI models

Becoming an AI Apprentice – A Field Guide

AI MAKERSPACE

Learning via schools and training providers

Self-directed Learning journey

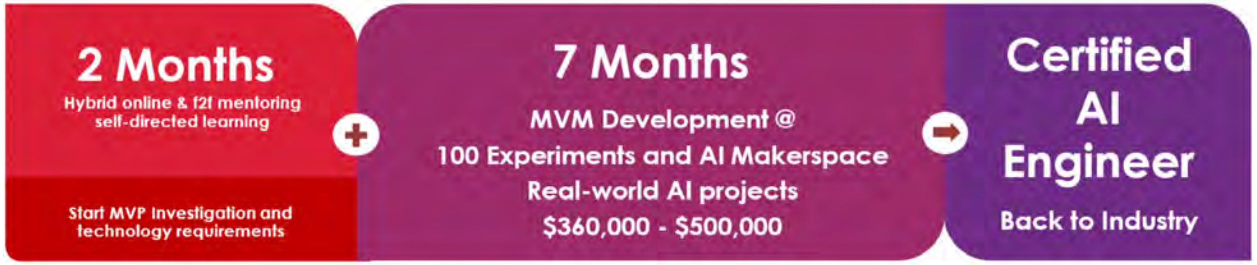
AI Professionals Association aip.org.sg

PROFESSIONAL QUALIFICATION

Associate AI Engineer	Chartered AI Engineer Level 1	Chartered AI Engineer Level 2	Chartered AI Engineer Level 3
Assessment Test	Evaluation Panel	Evaluation Panel	By Nomination
Able to **build an AI model** to solve a use case	Able to replicate AI research and **deploy real-world AI applications**	Able to deploy production **AI at scale**	Able to **architect** and manage multiple complex AI applications

COMPETENCY EXPECTED OF CERTIFIED PROFESSIONALS

AI Apprentices



2 Months

Hybrid online & f2f mentoring self-directed learning

Start MVP Investigation and technology requirements

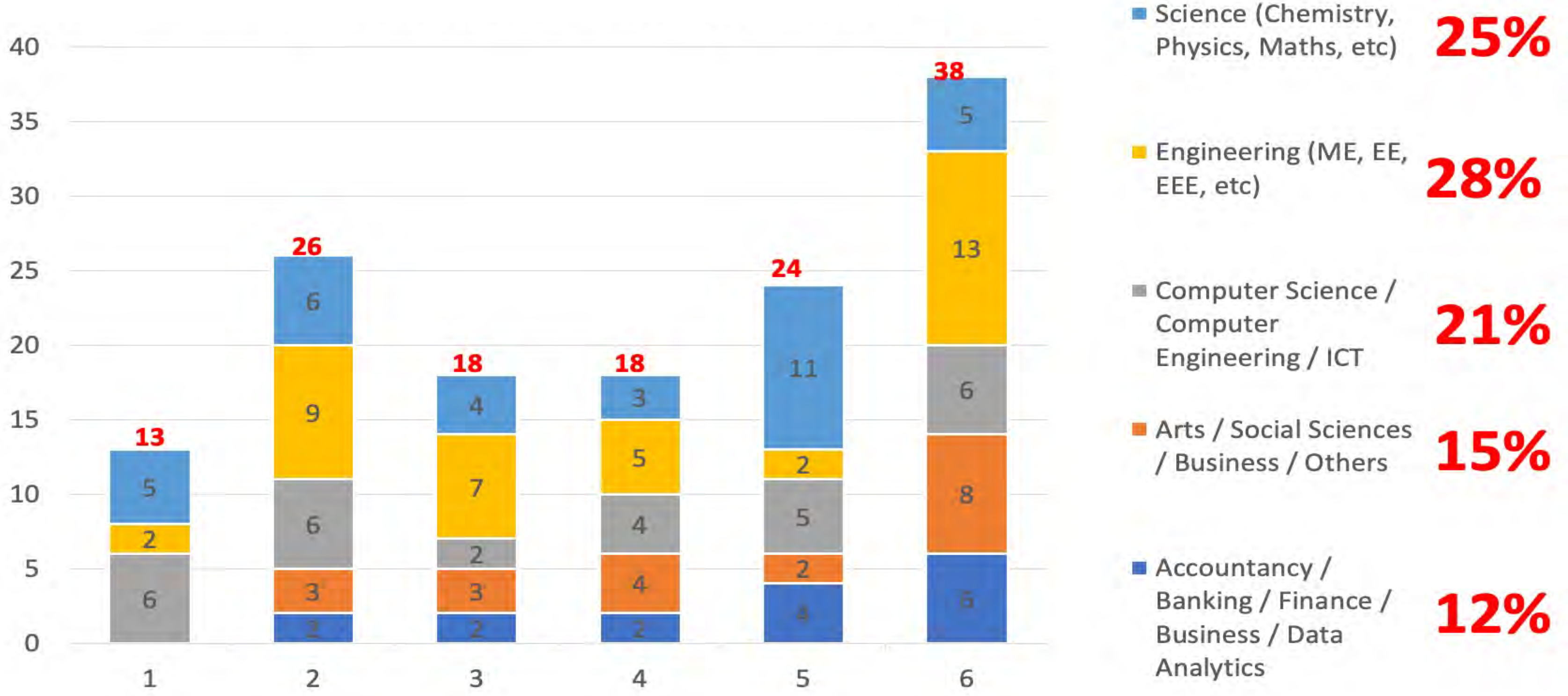
7 Months

MVM Development @ 100 Experiments and AI Makerspace Real-world AI projects $360,000 - $500,000

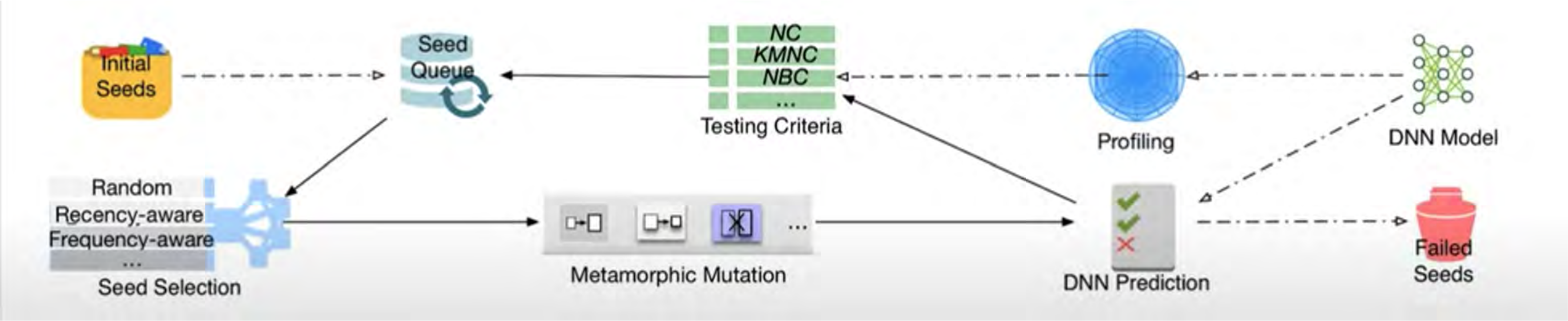
Certified AI Engineer

Back to Industry

AI Apprenticeship Programme® intakes (academic background)



Technology AISG is working on with our collaborators



AI SINGAPORE

Thank you

AISingapore

AISingapore

ai_singapore

aisingapore