# MACHINE LEARNING AND AI IN HIGH SPEED SYSTEM DESIGN

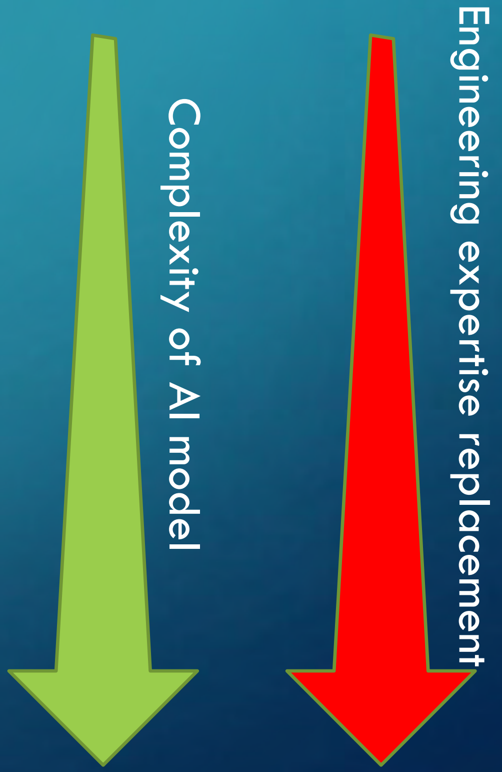

CHRIS CHENG, DISTINGUISHED TECHNOLOGIST, STORAGE DIVISION, HPE

YONGJIN CHOI, MASTER TECHNOLOGIST, STORAGE DIVISION, HPE

SUMON DEY, SENIOR MACHINE LEARNING ENGINEER, STORAGE DIVISION, HPE

# WORK IN PROGRESS, THE JOURNEY CONTINUES….

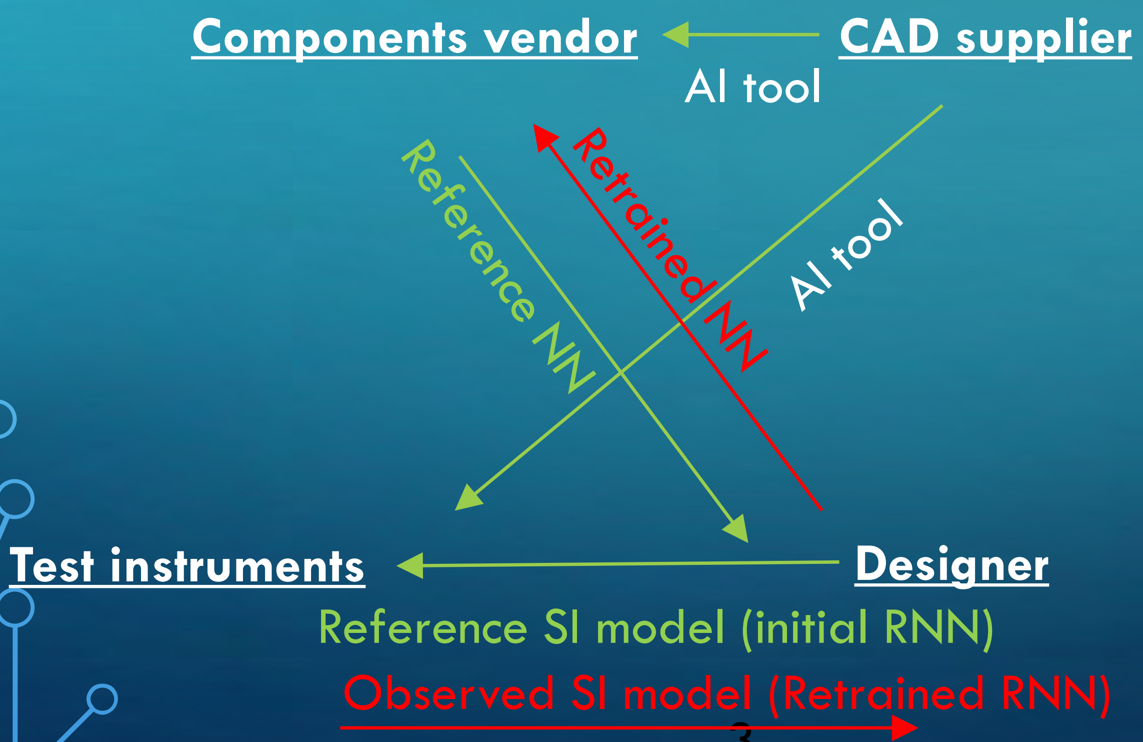
- Year one : Improving an existing process
  - Self correcting simulation models with neural network
- Year two : Speeding up and optimizing a design
  - Accelerating 56G PAM4 SerDes tuning with PCA vectors
- Year three : Digital Twins
  - Automatic channel condition detection and SerDes tuning using digital twins
- Year four : Deep learning of a design process
  - GAN modeling of SerDes

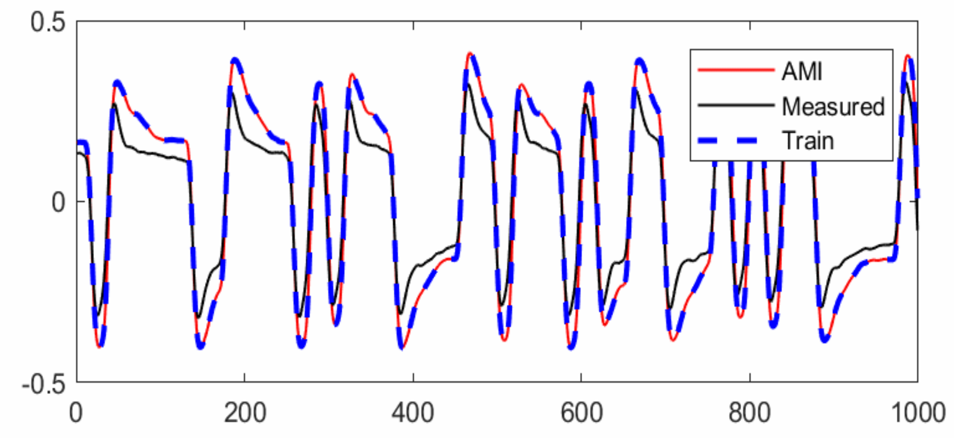


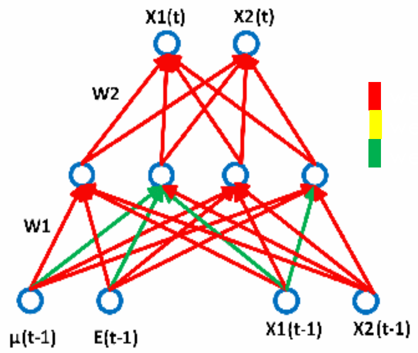
# Year 1 : Self correcting models

- Objective :

Machine learning as an alternative for existing solution



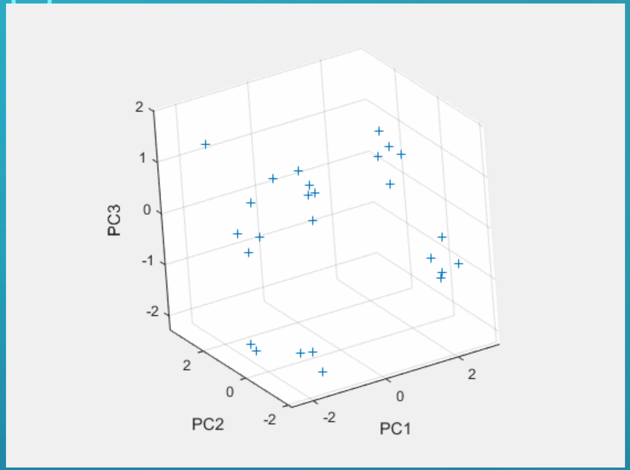



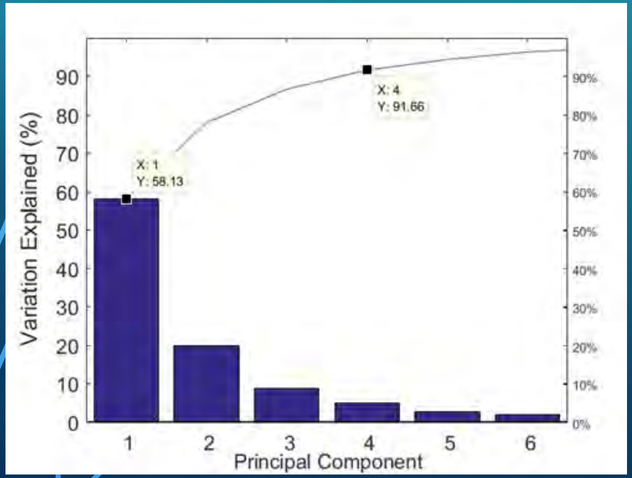



# YEAR 2 : ACCELERATING 56G PAM4 SERDES TUNING

- Objective :

Using Machine learning to accelerate optimizing a design

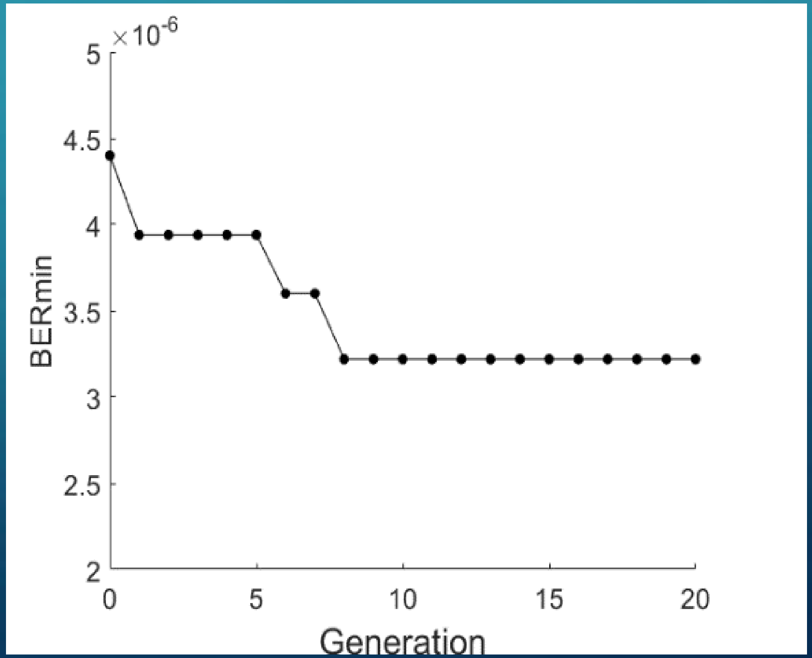


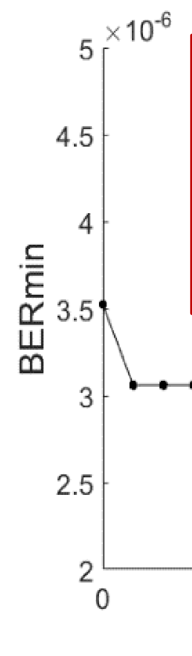

Zhu et al, DesignCon 2019

Best fitness function versus generation.

(a) GA

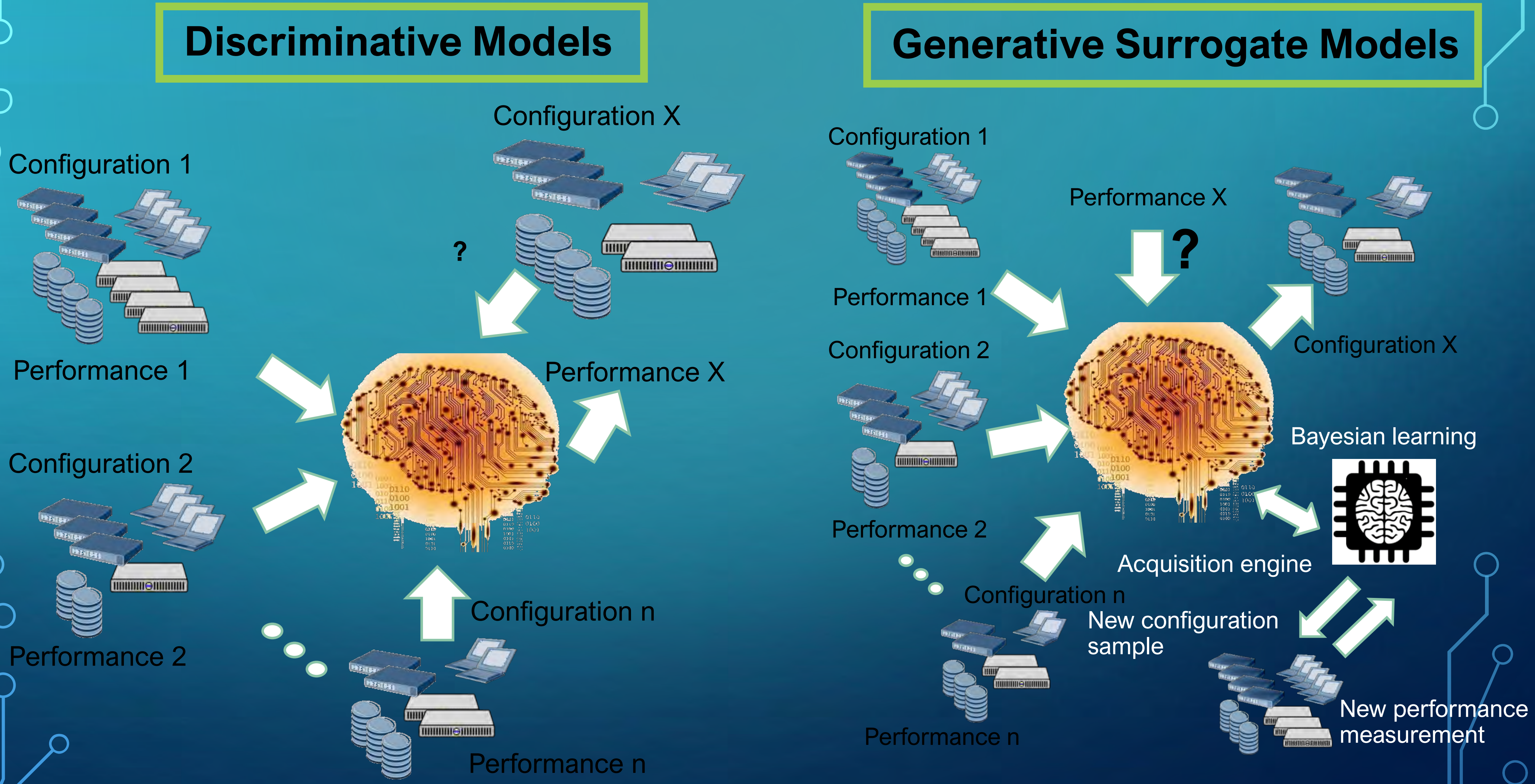



(b) GA-PCA

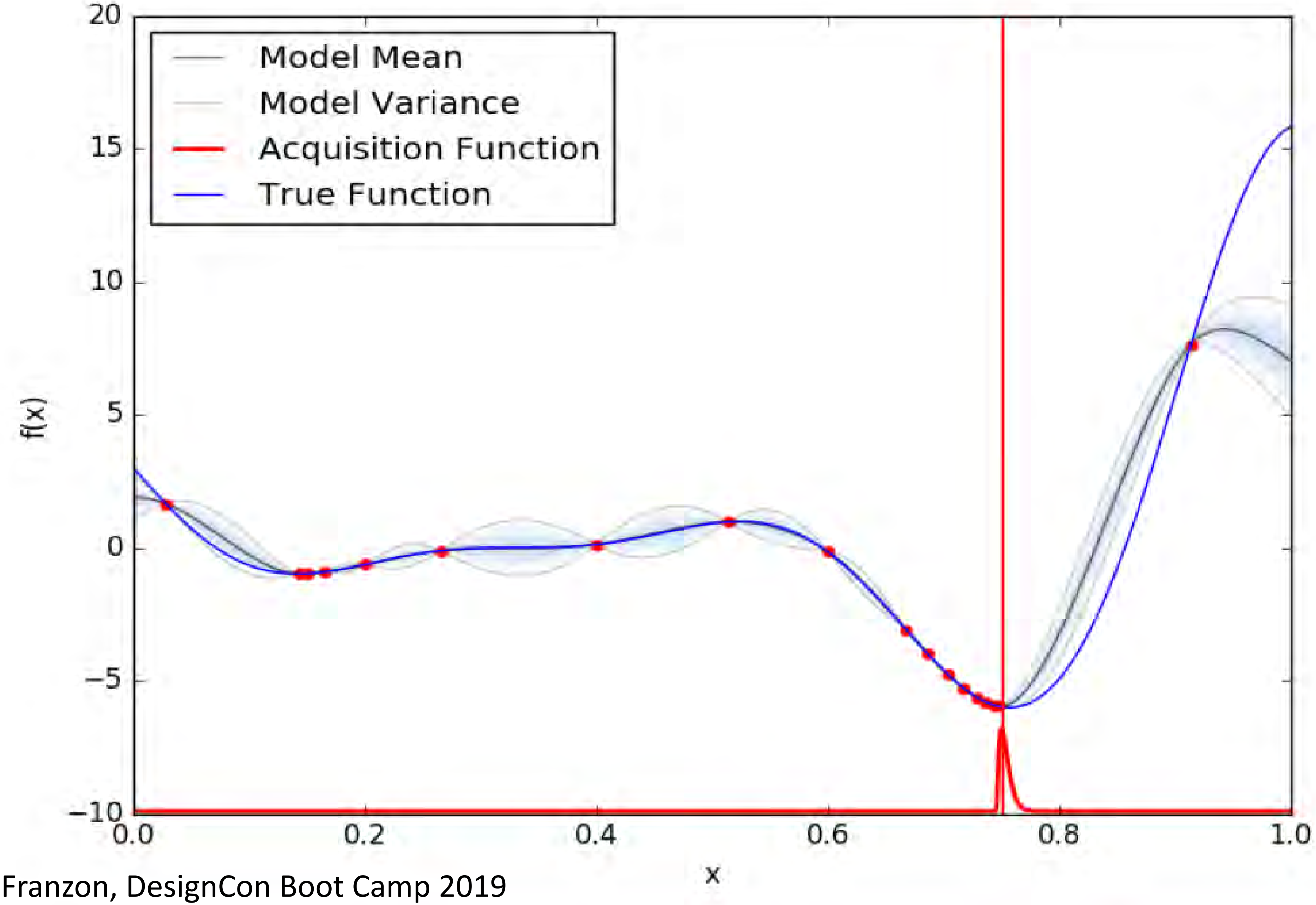



PCA based GA algorithm converged with fewer generations, more quickly than the general GA algorithm.

# YEAR 3 : GENERATIVE MODELS AS DIGITAL TWIN



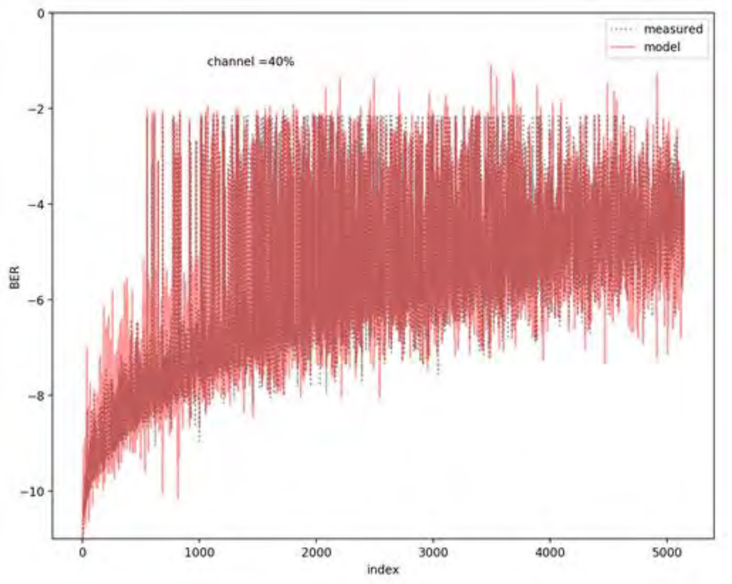
# BAYESIAN LEARNING 1D DEMO



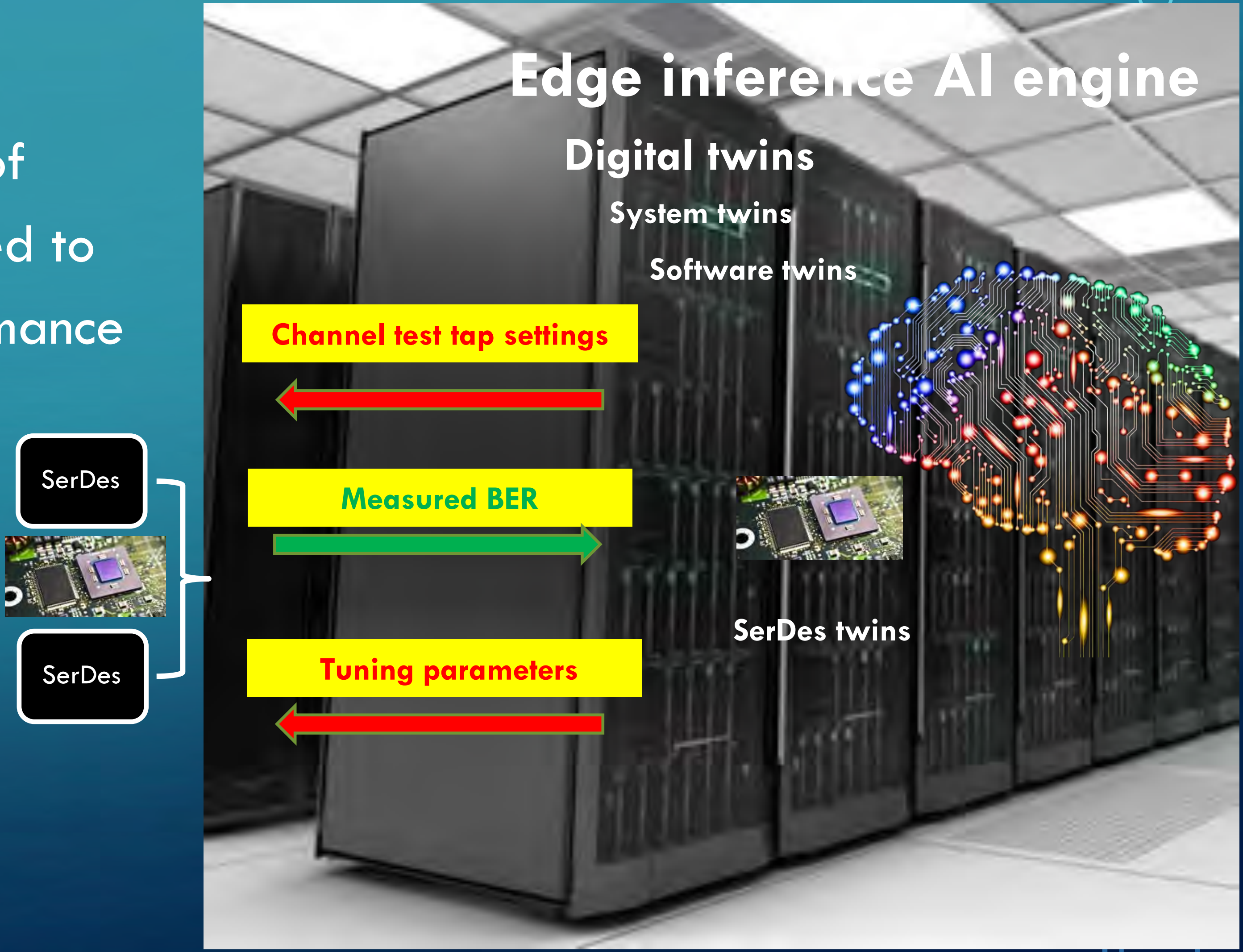

Franzon, DesignCon Boot Camp 2019

# DIGITAL TWINS FOR INTELLIGENT EDGE COMPUTING

- Next generation platforms will have build-in AI engine to perform inference at the edge
- Digital twins are surrogate models of system performance and can be used to dynamically tune the system performance



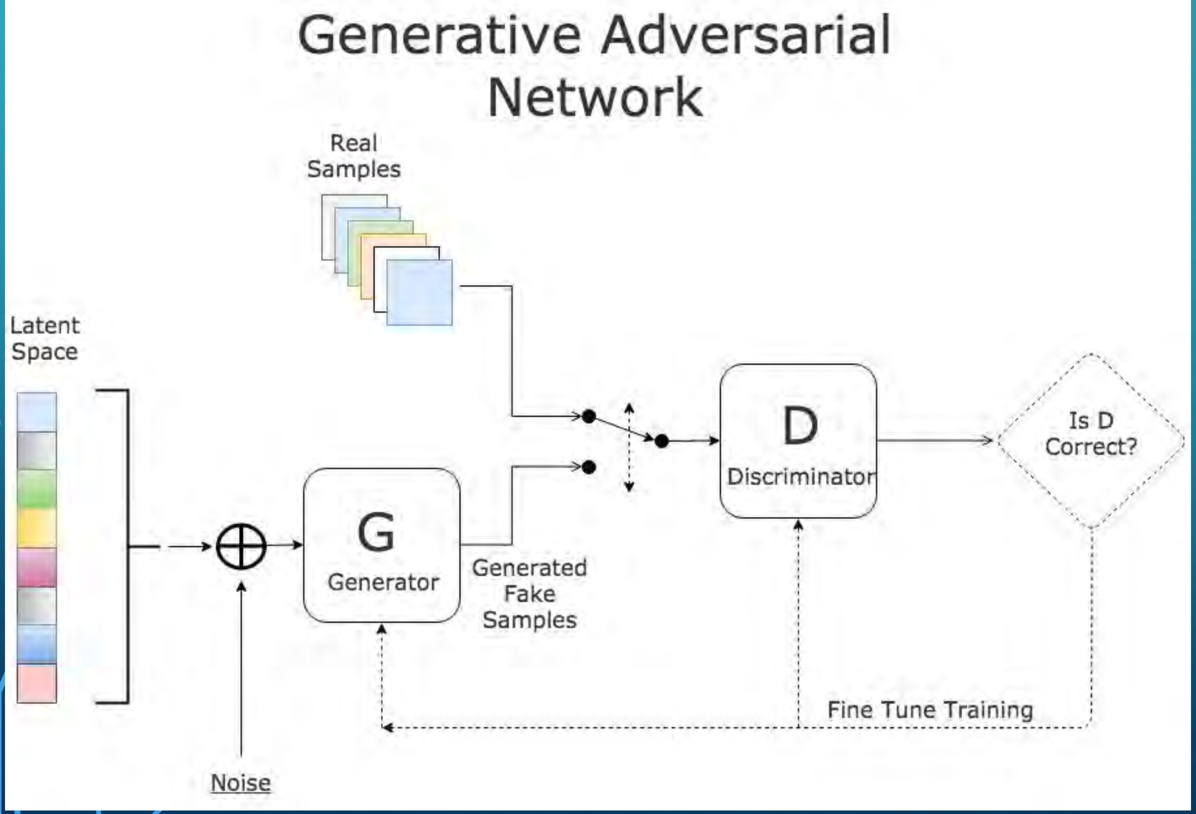

Choi et al, DesignCon 2020

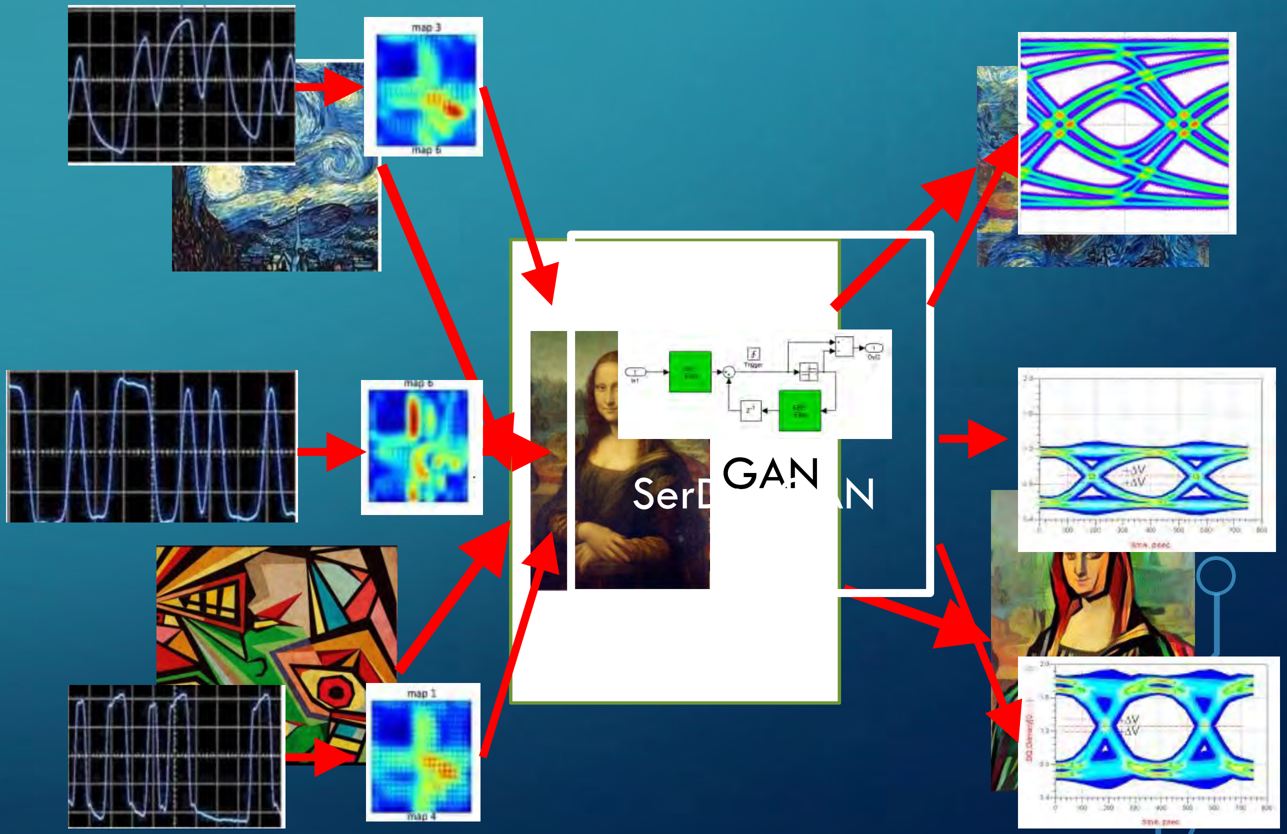


# YEAR 4 : DEEP LEARNING OF DESIGN PROCESS

- Objective

Deep learning of SerDes modeling using Generative Adversarial Networks

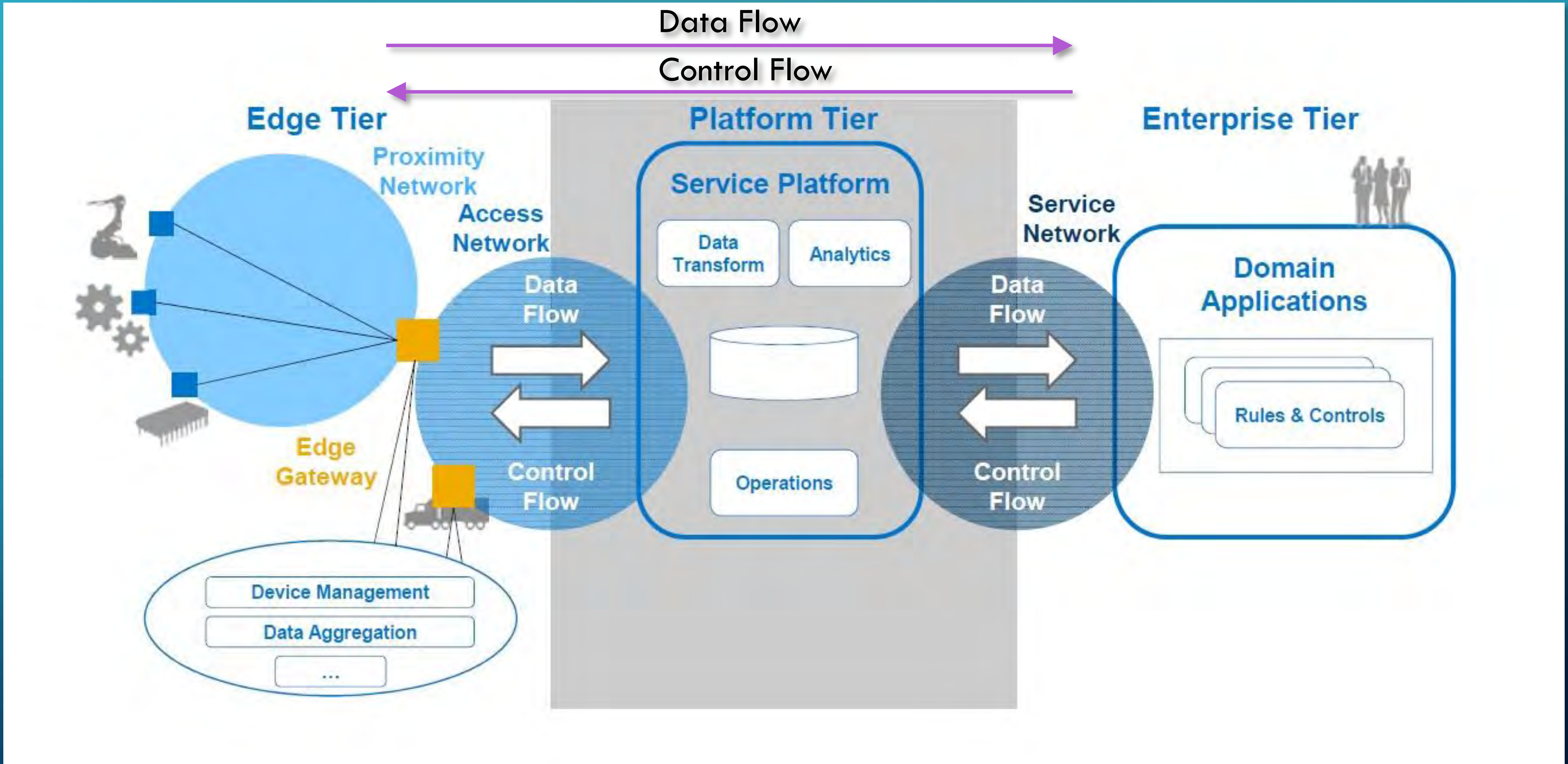


# INTELLIGENT EDGE FOR THE 5G/IOT GENERATION

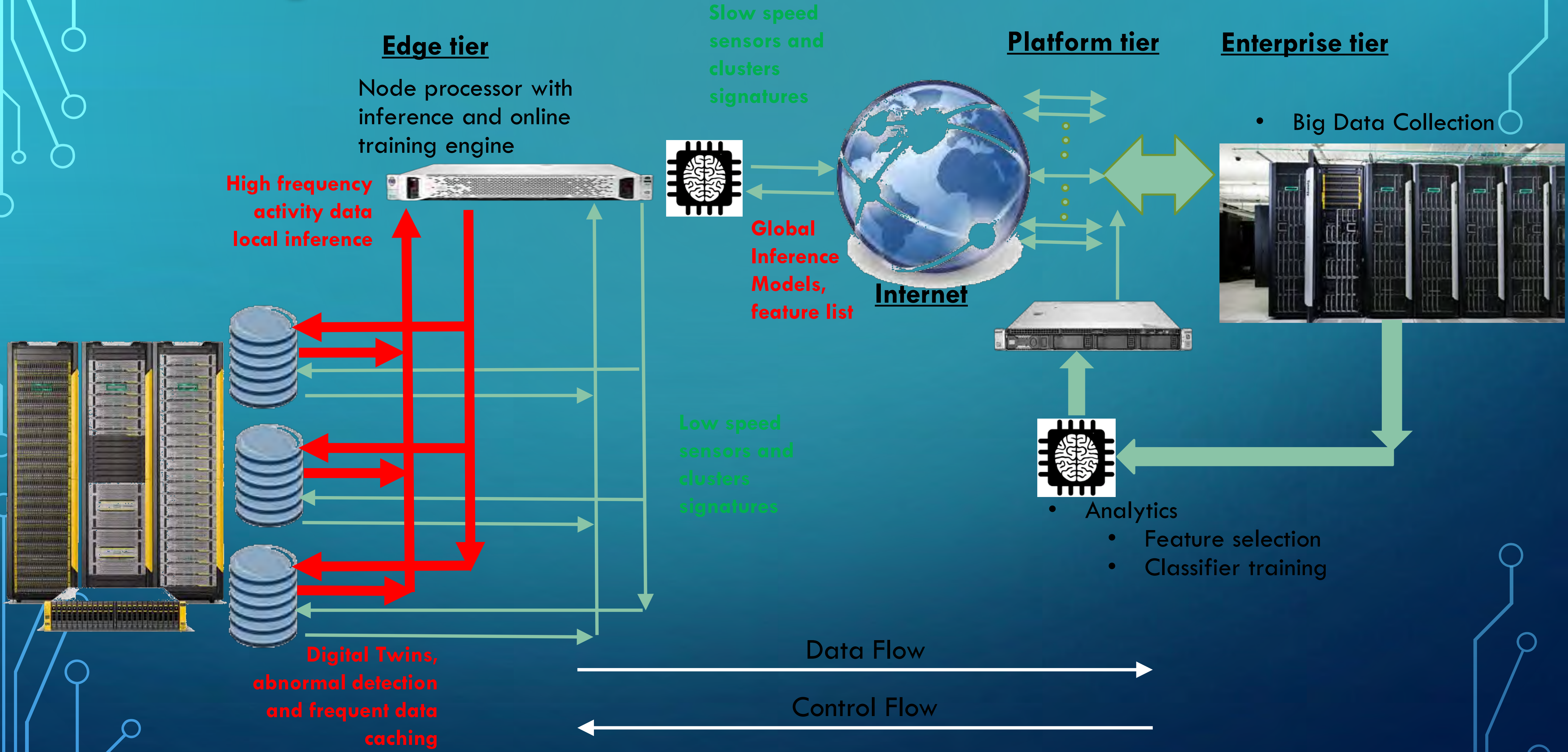
AI FOR INTELLIGENT EDGE MANAGEMENT AND SECURITY

# IoT management model




IEC IoT white paper

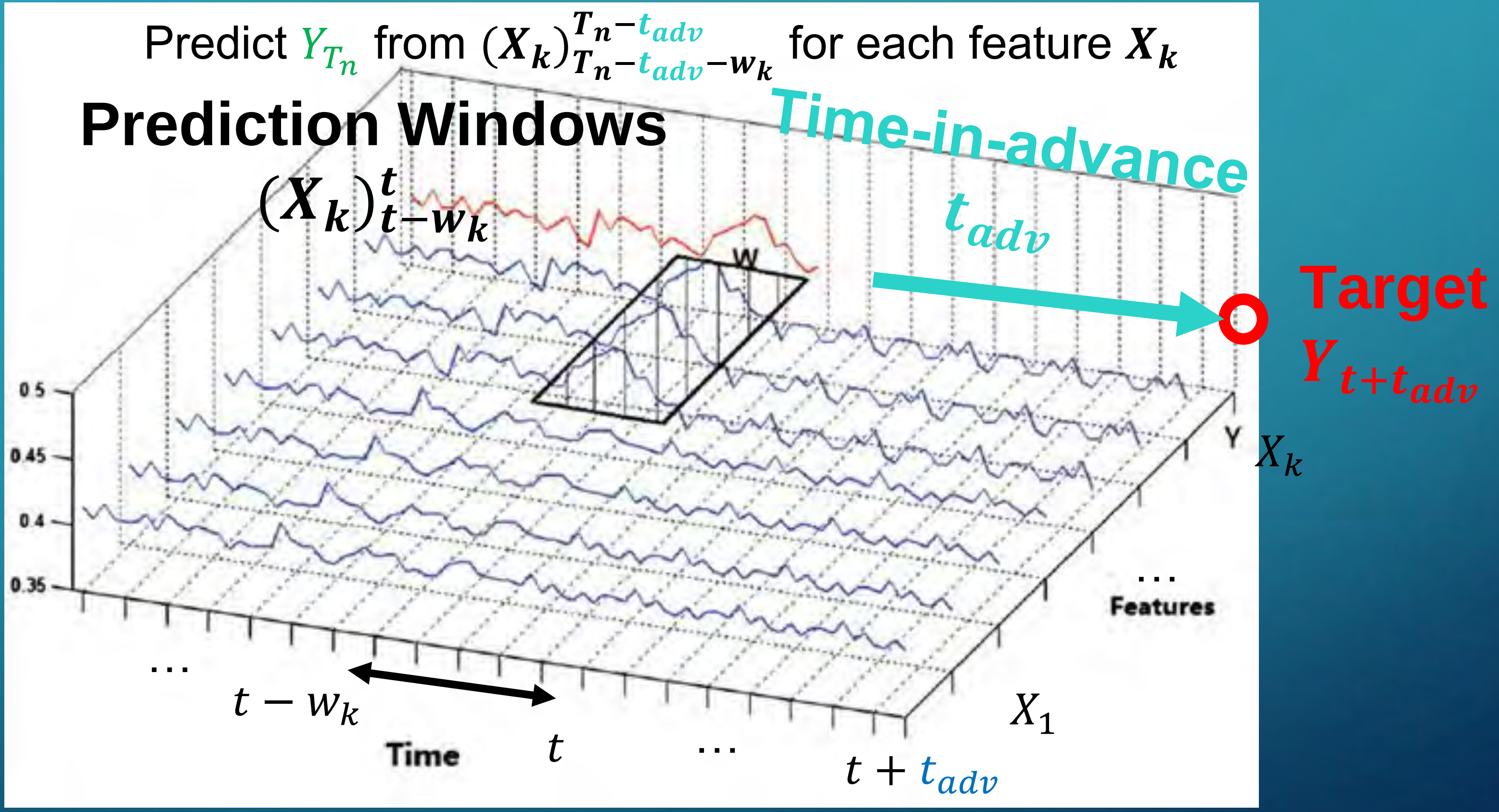
# Edge inference



# 3 Types of IoT and hardware management models

| | Feature size (number of variables or complexity) | Prediction throughput | Prediction interval | Machine learning engine | Examples |
|---|---|---|---|---|---|
| Big data Small learning | Less than 100 | A few thousands to millions of predictions per sec | Variable from days to ms | Ensemble classifier | Hardware failure prediction<br>Software failure prediction<br>Automatic application detection<br>Storage security applications (ransomware detection)<br>System abnormally detection |
| Big data Medium learning | Between 100 to 500 | A few hundred predictions per secs | A few secs | Generative performance surrogate model with Bayesian learning | Digital Twins for dynamic system performance optimization |
| Big data Deep learning | 100s to 1000s | 1000's of predictions in an hours | 5-15 mins | Deep Markovian Models<br>Deep learning neural networks | High dimensional time series for resource demand prediction |

# Casual Inference for feature selection

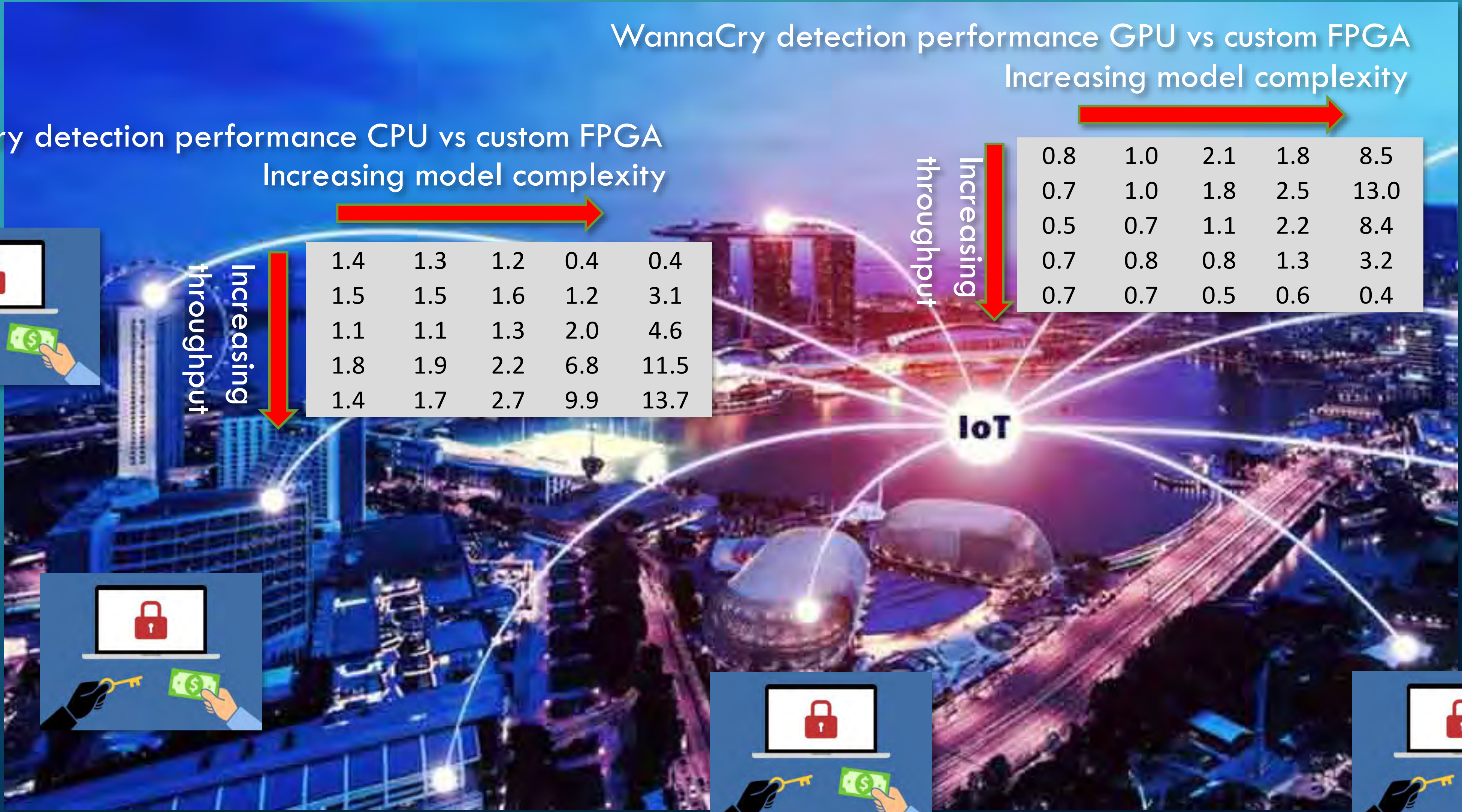



- 30% features reduction
- 15% accuracy improvement

CAEML research

# SECURING THE EDGE : RANSOMWARE DETECTION

- Machine learning has been shown to be effective in detecting day zero attacks
- Latency and throughput are both important




WannaCry detection performance CPU vs custom FPGA

Increasing model complexity →

Increasing throughput ↓

| | | | | |
|---|---|---|---|---|
| 1.4 | 1.3 | 1.2 | 0.4 | 0.4 |
| 1.5 | 1.5 | 1.6 | 1.2 | 3.1 |
| 1.1 | 1.1 | 1.3 | 2.0 | 4.6 |
| 1.8 | 1.9 | 2.2 | 6.8 | 11.5 |
| 1.4 | 1.7 | 2.7 | 9.9 | 13.7 |

WannaCry detection performance GPU vs custom FPGA

Increasing model complexity →

Increasing throughput ↓

| | | | | |
|---|---|---|---|---|
| 0.8 | 1.0 | 2.1 | 1.8 | 8.5 |
| 0.7 | 1.0 | 1.8 | 2.5 | 13.0 |
| 0.5 | 0.7 | 1.1 | 2.2 | 8.4 |
| 0.7 | 0.8 | 0.8 | 1.3 | 3.2 |
| 0.7 | 0.7 | 0.5 | 0.6 | 0.4 |

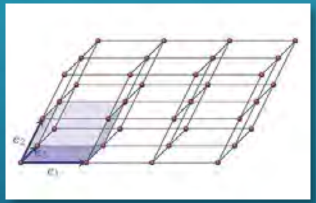
# SECURING THE EDGE : POST-QUANTUM SIDE CHANNEL ATTACH

Quantum computers is expected to be able to crack factor-based encryptions such as RSA

Homomorphic encryption

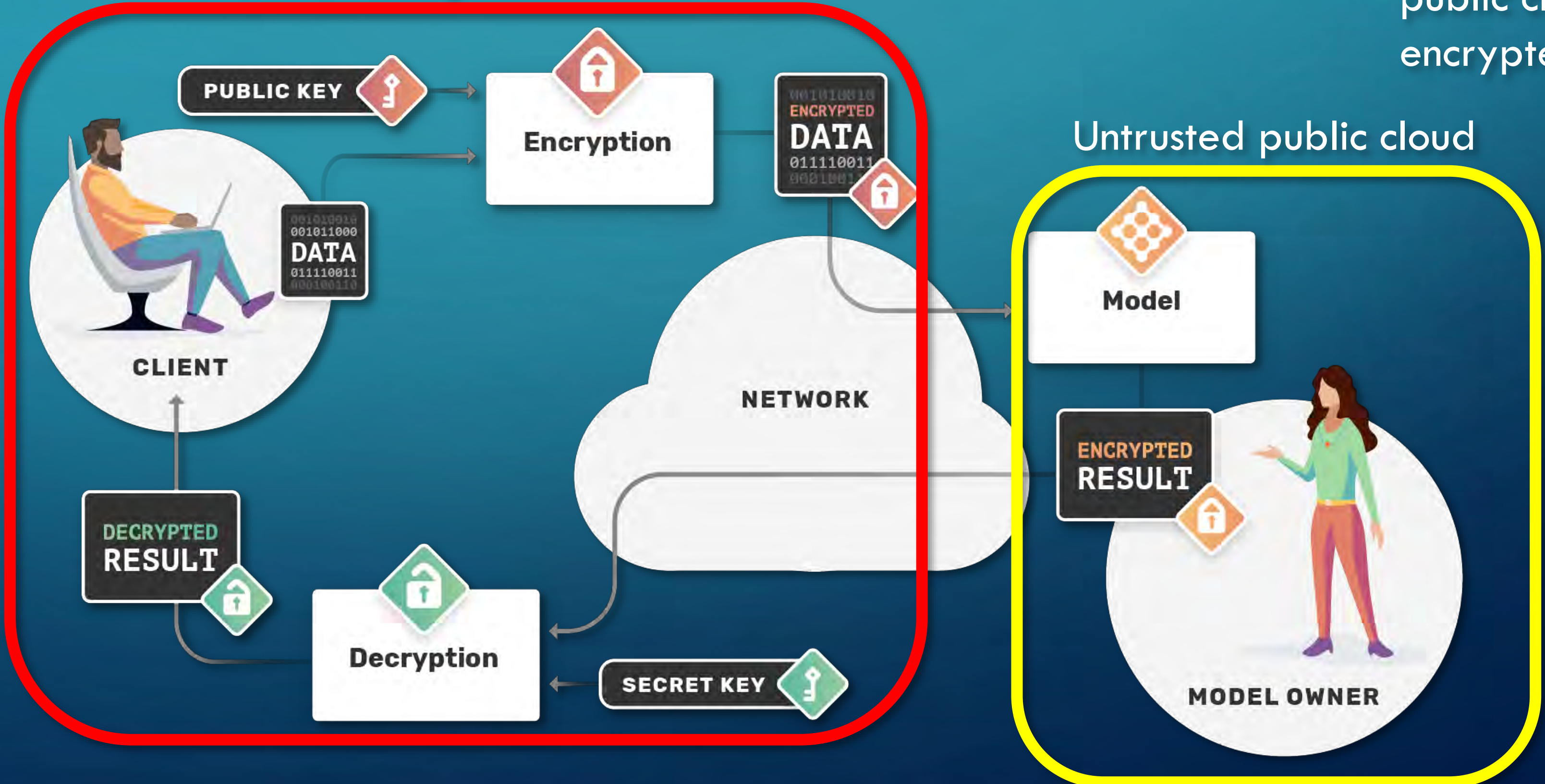
- Compute and operations done in public cloud with encrypted data

Lattice based
Post quantum
Homomorphic encryption

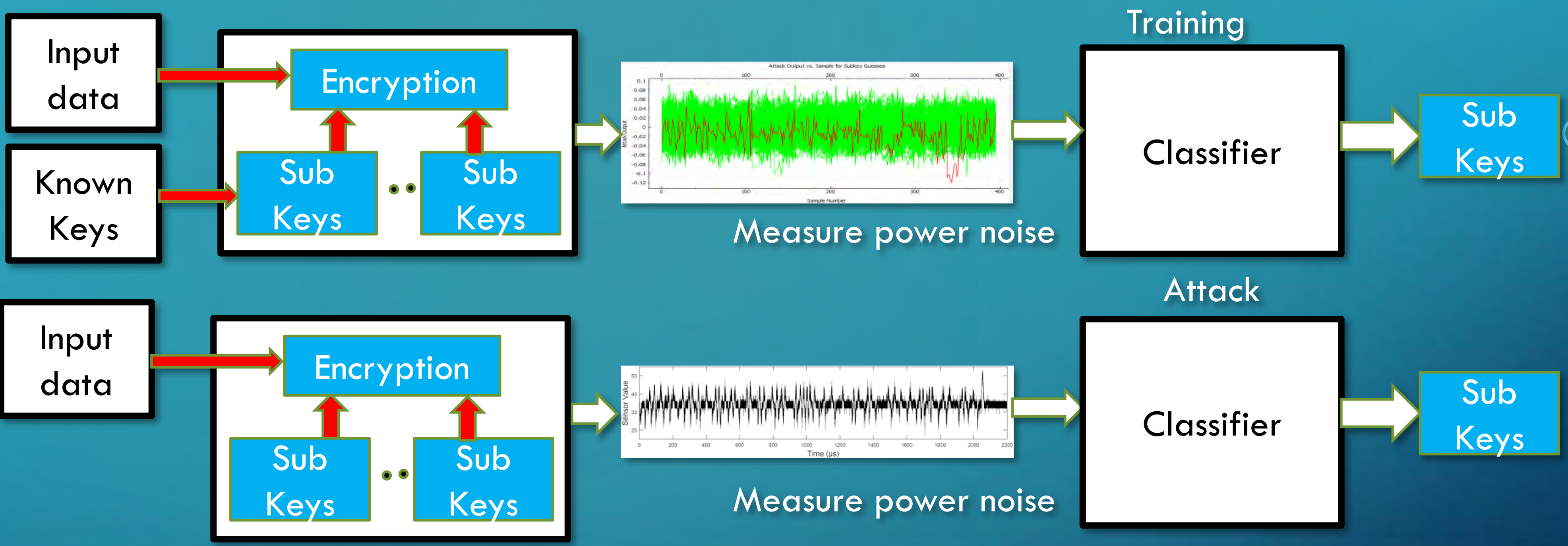


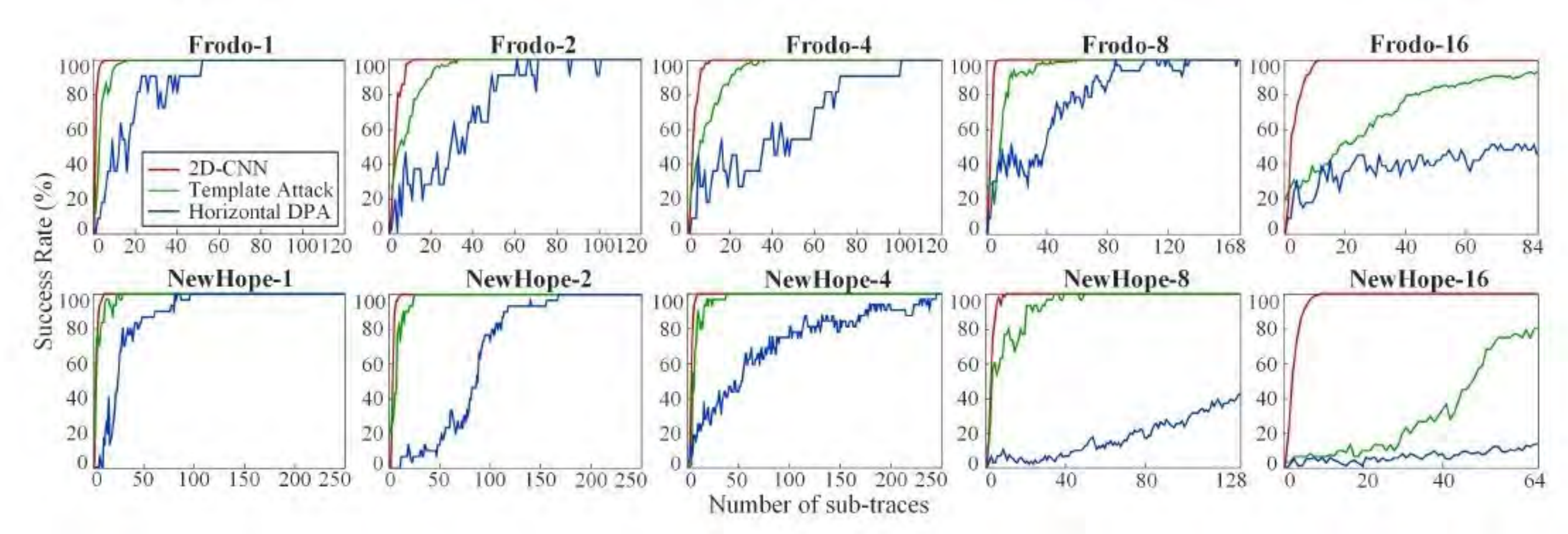
Secured edge device

Untrusted public cloud



# SIDE CHANNEL ATTACK POWER ANALYSIS




Aydin et al, SAMOS 2020

# FINAL THOUGHTS

- The learning continues but..
- Real world applications of ML/AI for SI/PI are here and many more to come
- Set up your own goals of using ML/AI and search for readily available solutions first
- Be prepare to spend 80-90% of your development time in data preparation, extract translate load (ETL)
- Enjoy the journey !