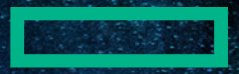


Hewlett Packard Enterprise

10/18/2019

# CAEML Research in Hardware Design and Optimization Using Machine Learning

Chris Cheng, Distinguished Technologist
Yongjin Choi, Master Technologist
Sumon Dey, Specialist Machine Learning Engineer

# Content

– Introduction to CAEML

– Unique types of machine learning models for IoT and hardware system management

– Example small learning : proactive hardware failure prediction

– Example medium learning : 56G PAM SerDes performance optimization

– Example deep learning : Dynamic resources demand forecast


# Vision

The center's goal is to enable fast, accurate design and verification of microelectronic circuits and systems by creating machine learning algorithms to derive models used for electronic design automation.

By speeding up the design and verification of microelectronic circuits and systems, CAEML will reduce development cost and time-to-market for manufacturers of microelectronic products, and will enable the development of optimized products, e.g. for low-power, high-reliability or security.

# CAEML Members 2019



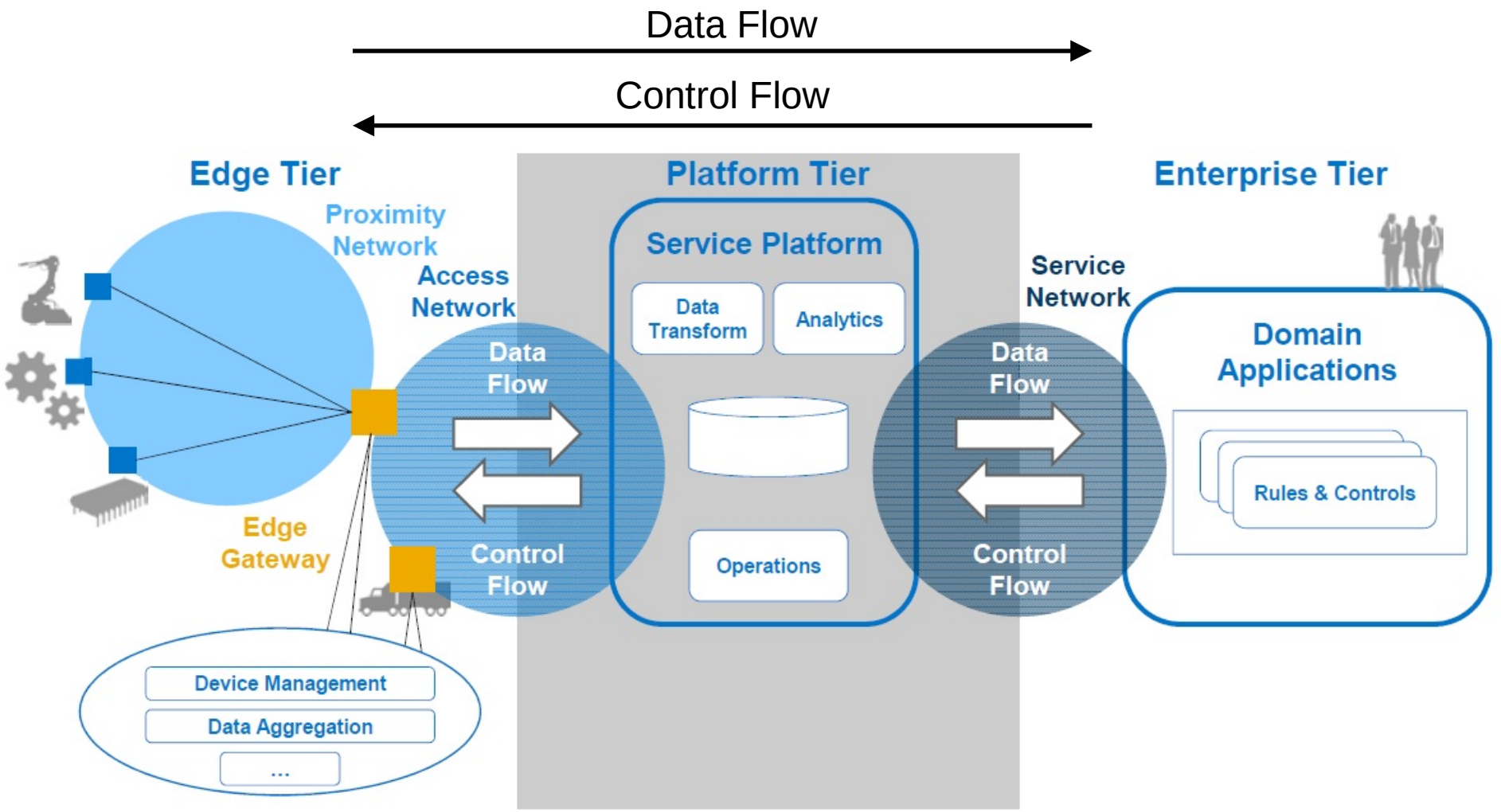






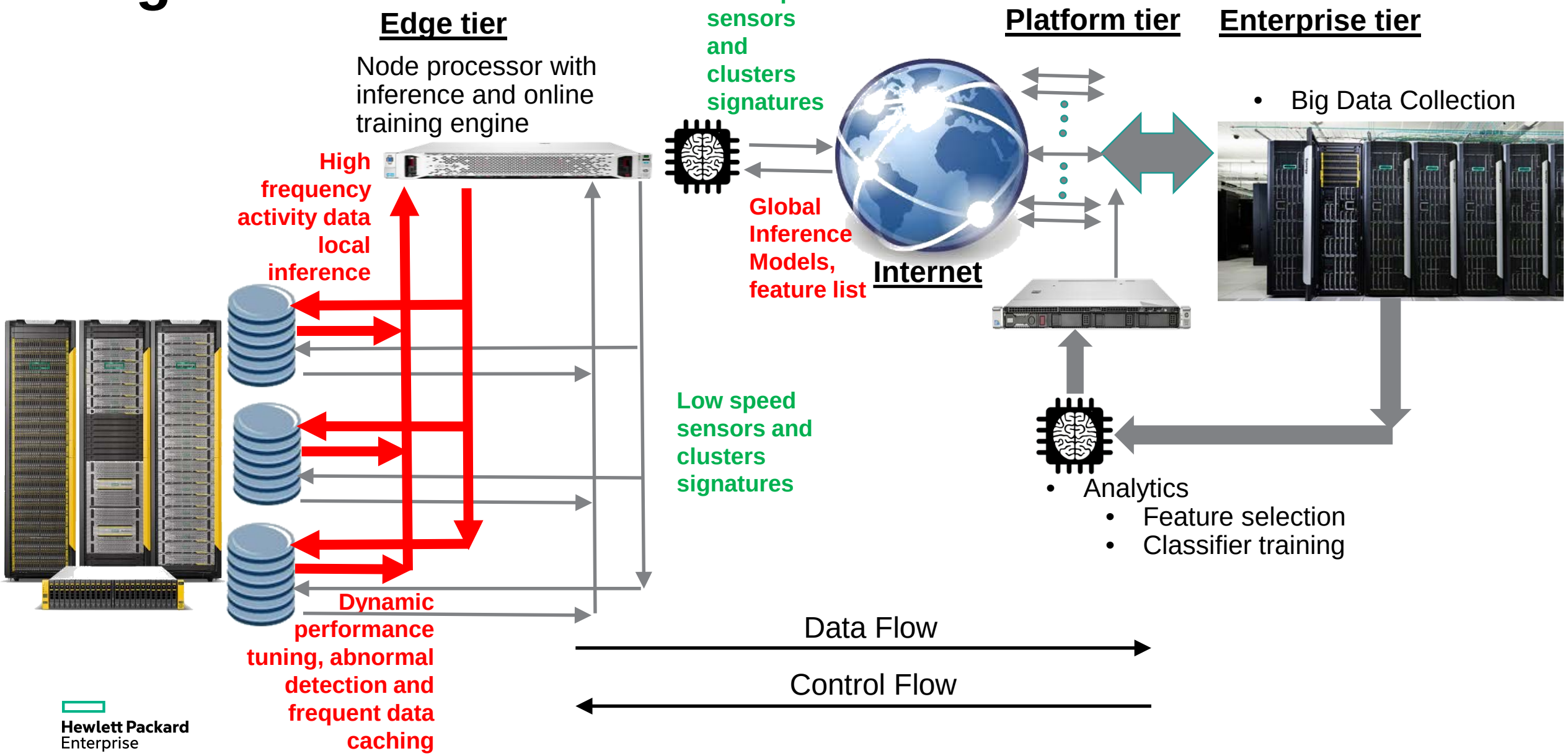



# IoT management model




IEC IoT white paper

# Edge inference



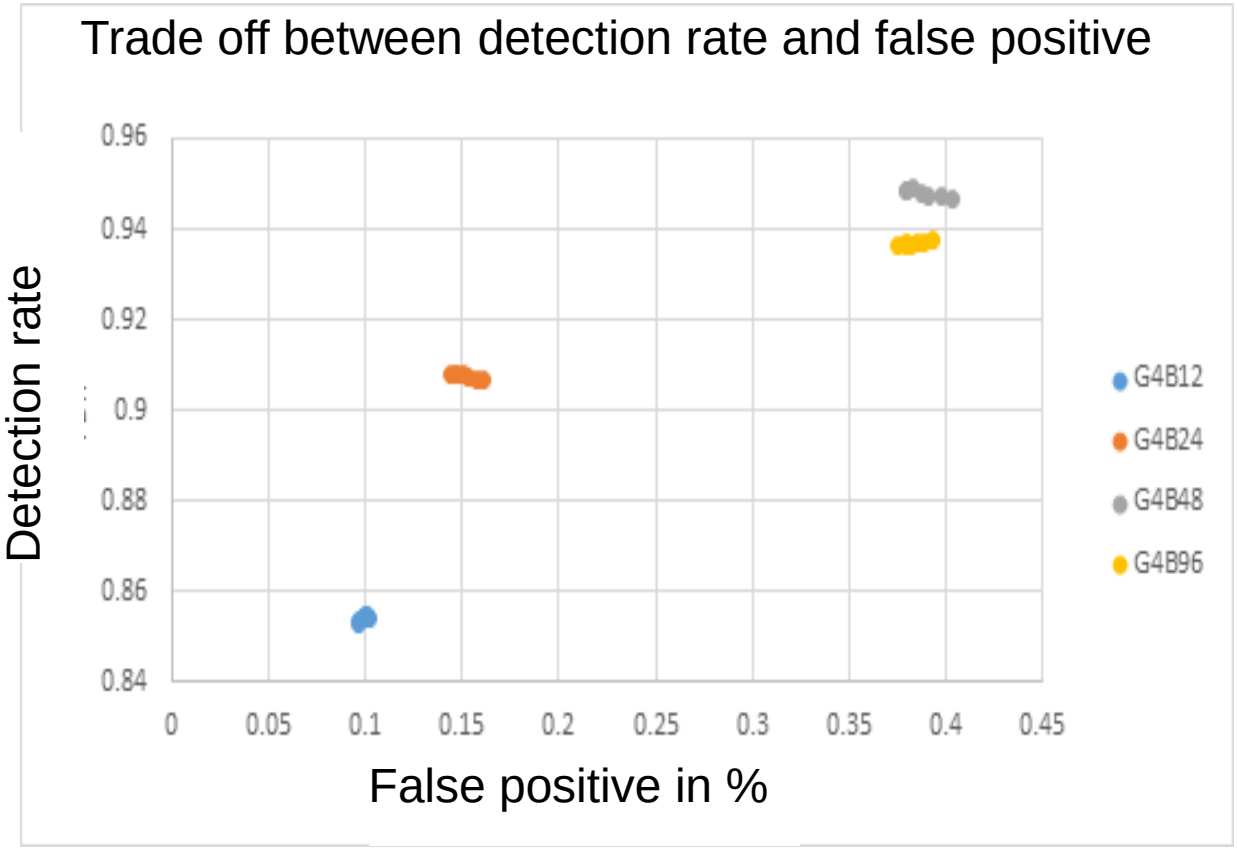

Hewlett Packard Enterprise

# 3 Types of IOT and hardware management models

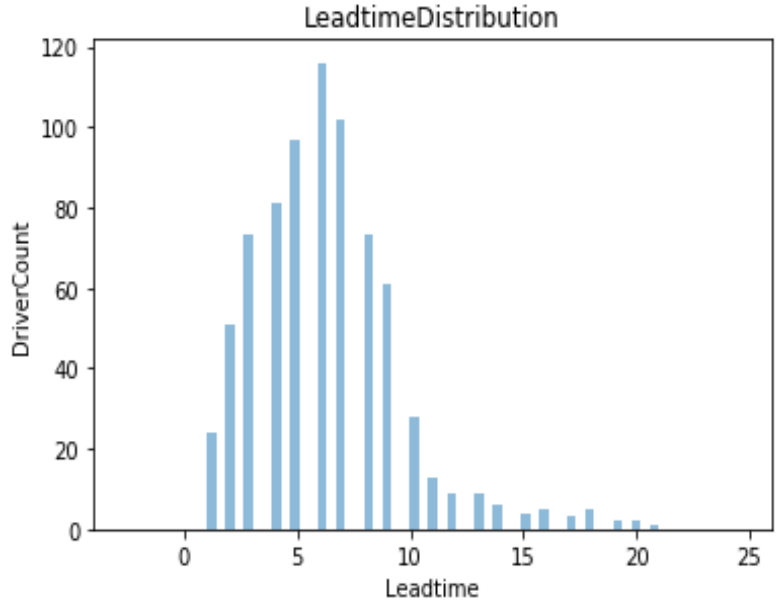
| | Feature size (number of variables or complexity) | Prediction throughput | Prediction interval | Machine learning engine | Examples |
|---|---|---|---|---|---|
|  Big data Small learning | Less than 100 | A few thousands to millions of predictions per sec | Variable from days to ms | Ensemble classifier | 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 | 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 |

# Small learning example : proactive hardware failure detection

# Proactive hardware failure prediction




- Window N =5 :




Average Lead Time =6.28 days

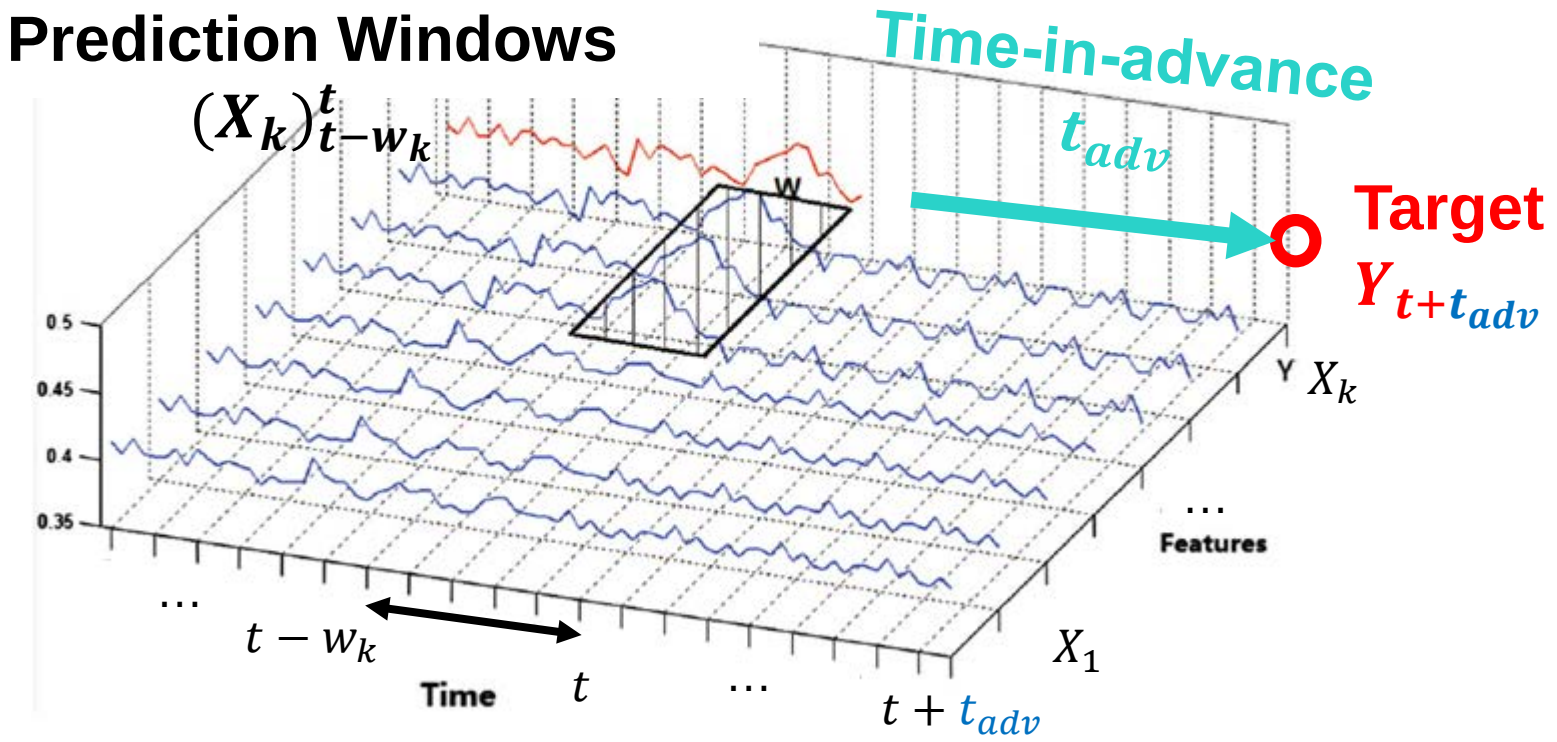
Recommended MATLAB package : Statistics and Machine Learning Tool Box

Hewlett Packard Enterprise

# Causal inference for feature selection

- Causal inference is used to pick the sensor signals and sample window
- 30% features reduction
- 15% accuracy improvement

Predict $Y_{T_n}$ from $(X_k)_{T_n-t_{adv}-w_k}^{T_n-t_{adv}}$ for each feature $X_k$




CAEML research

# Performance limitation of GPU on ensemble classifier

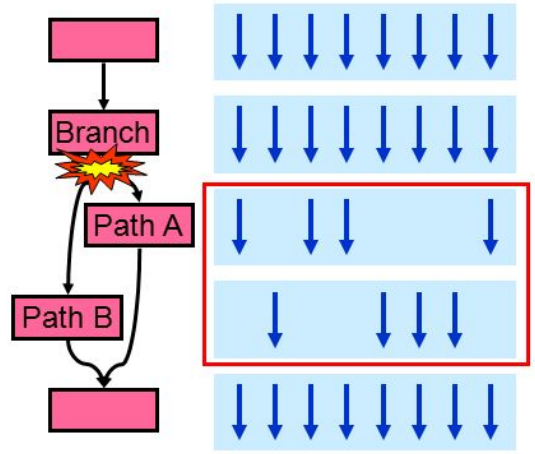
Relative inference time

| Number of samples | Intel CPU | CUDA-Tree |
|---|---|---|
| 10k | 1 | 2.2 |
| 20k | 1 | 2 |
| 50k | 1 | 1.8 |

Higher is slower execution time
GPU is slower than Intel CPU !

## Control Flow Problem in GPUs/SIMD

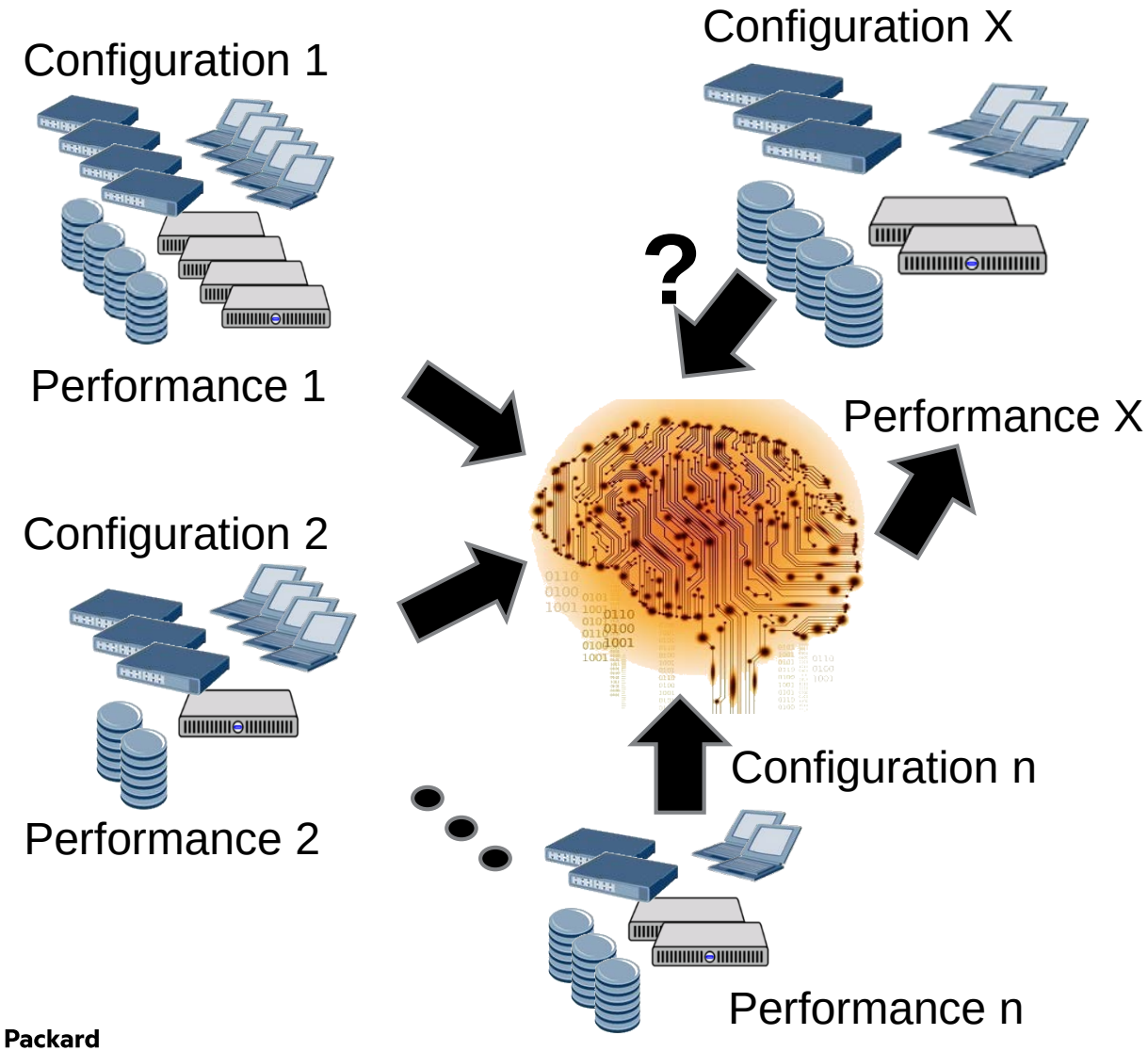
- GPU uses SIMD pipeline to save area on control logic.
  - Group scalar threads into warps
- Branch divergence occurs when threads inside warps branch to different execution paths.




Wilson Peng et al. UBC

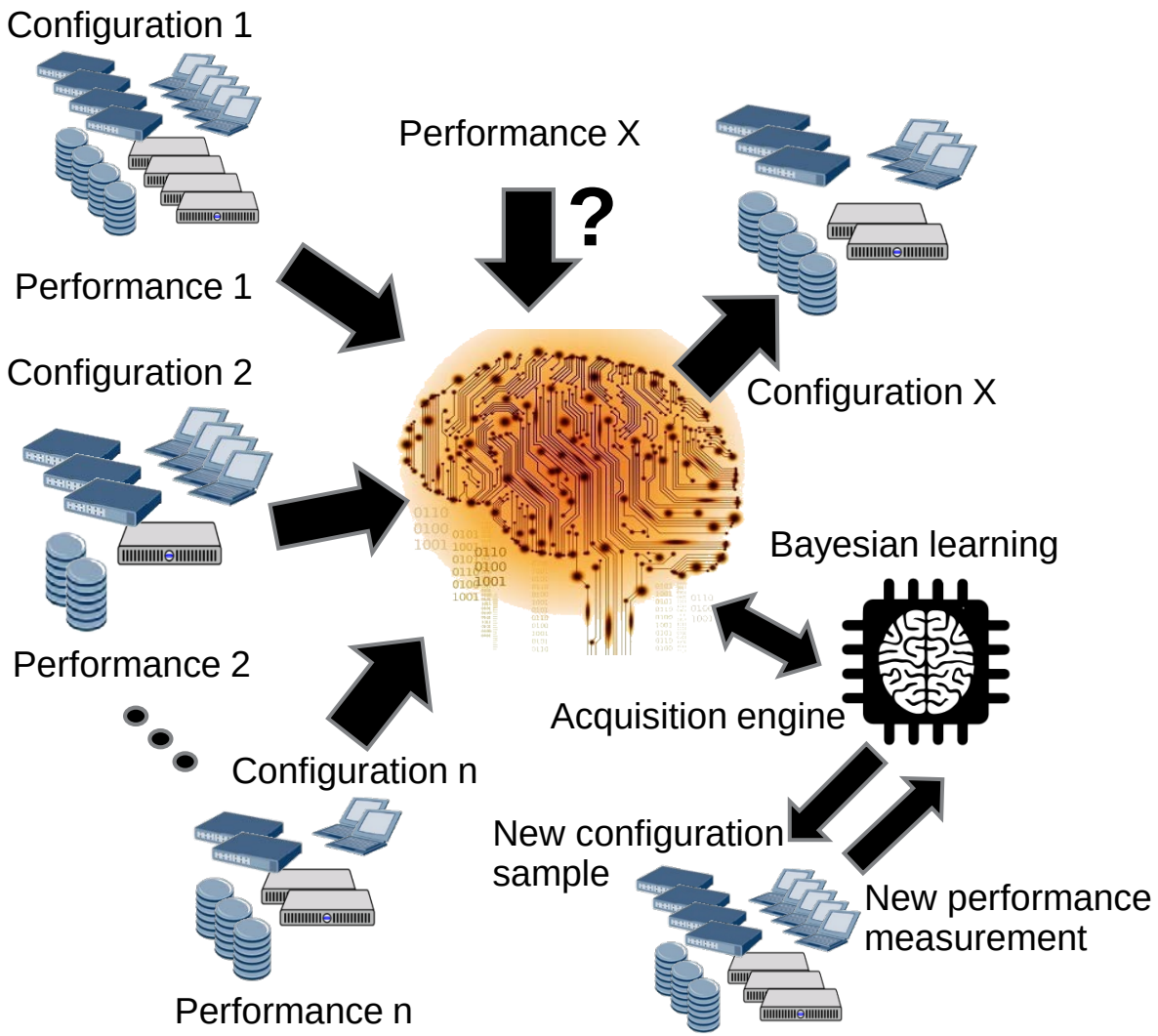
# Medium learning example : Surrogate models for system performance optimization

# Discriminative system model




- Collect n samples of system configurations and observed performance
- Build a model to predict the performance
  - Neural network
  - Linear/non-linear regression
- In the future, given a configuration X, model predict performance X

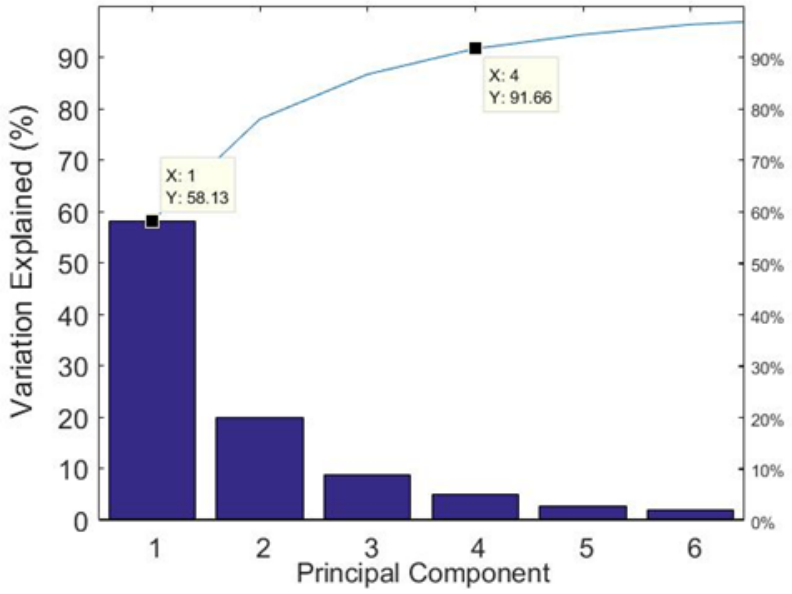
# Generative system model

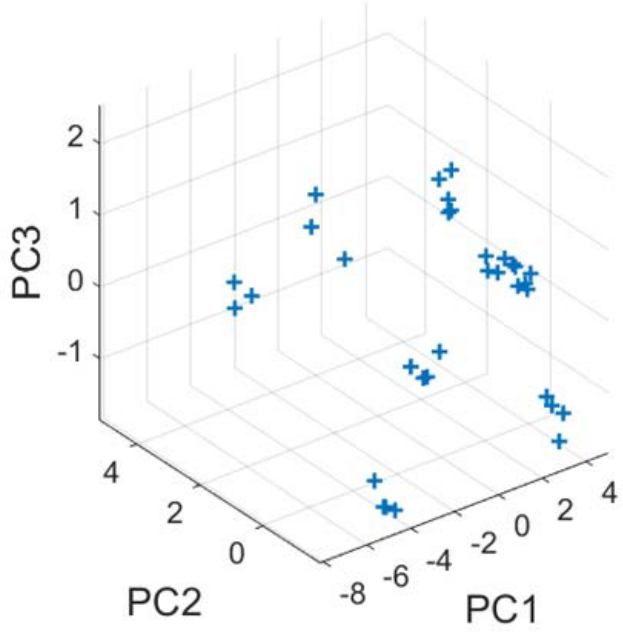



- Collect n samples of system configuration and performance
- Construct surrogate model and use Bayesian learning to construct additional configurations and performance measurement
- Keep iteration until target uncertainty is hit
- Given a new target performance, what is the optimal configuration to achieve, usually tie in with a cost function and optimize for the lowest cost

# Principal component analysis and surrogate models

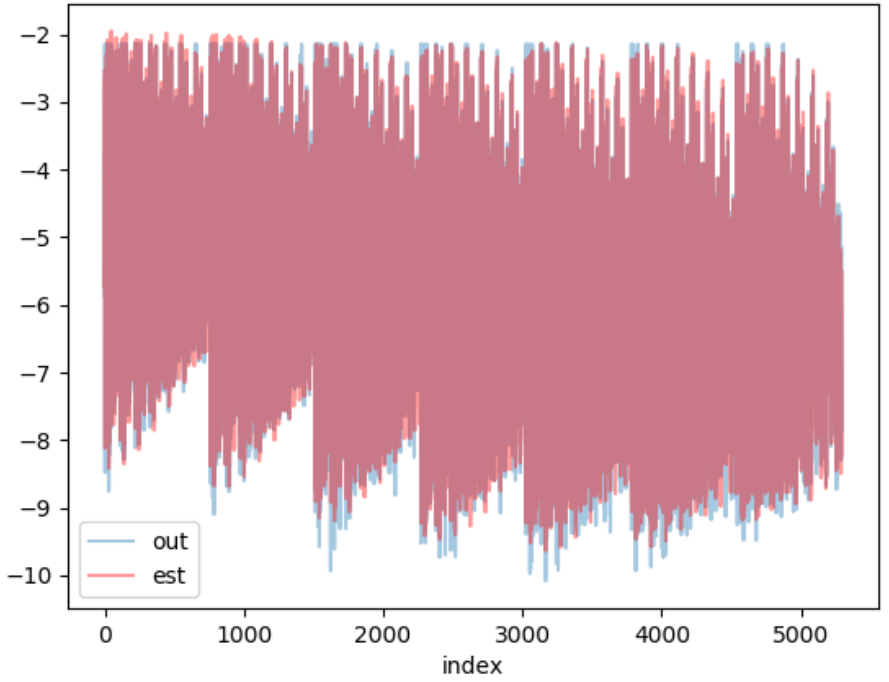
25 controlling taps in the 56G PAM SerDes is mapped into 4 principal component vectors that can cover 92% of solutions




Zhu et al, DesignCon 2019

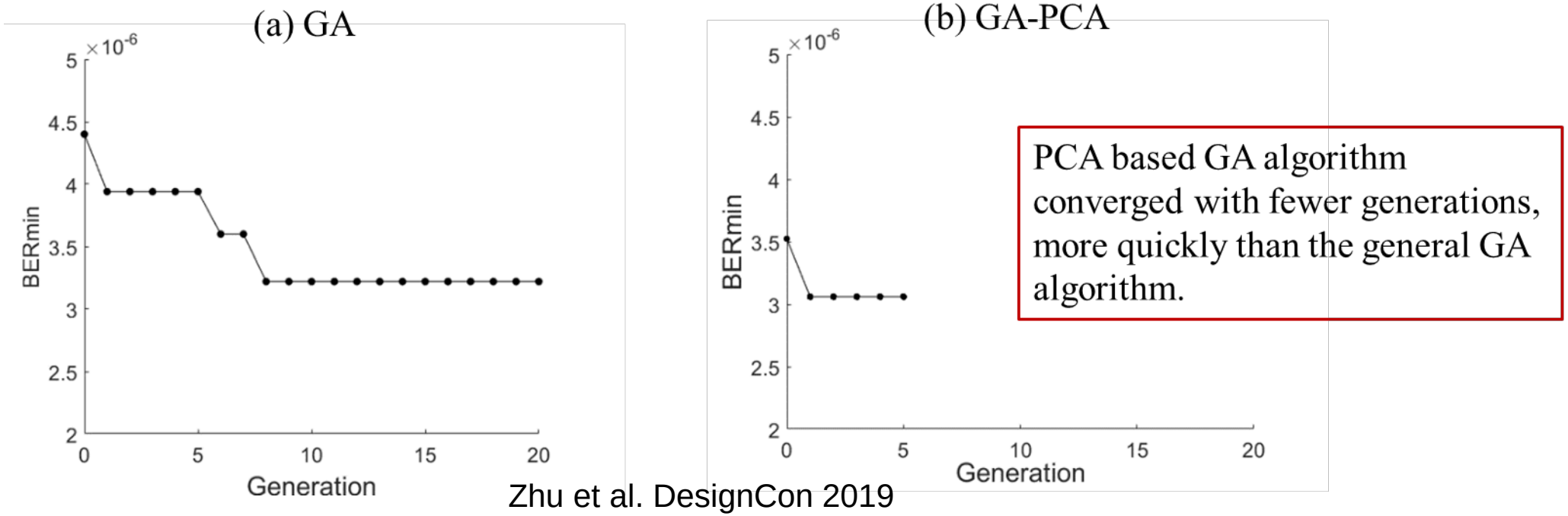
Surrogate models based on 5000+ measured samples




Recommended MATLAB package : Statistics and Machine Learning Toolbox

# Accelerated 56G PAM channel optimization using PCA

Best fitness function versus generation.

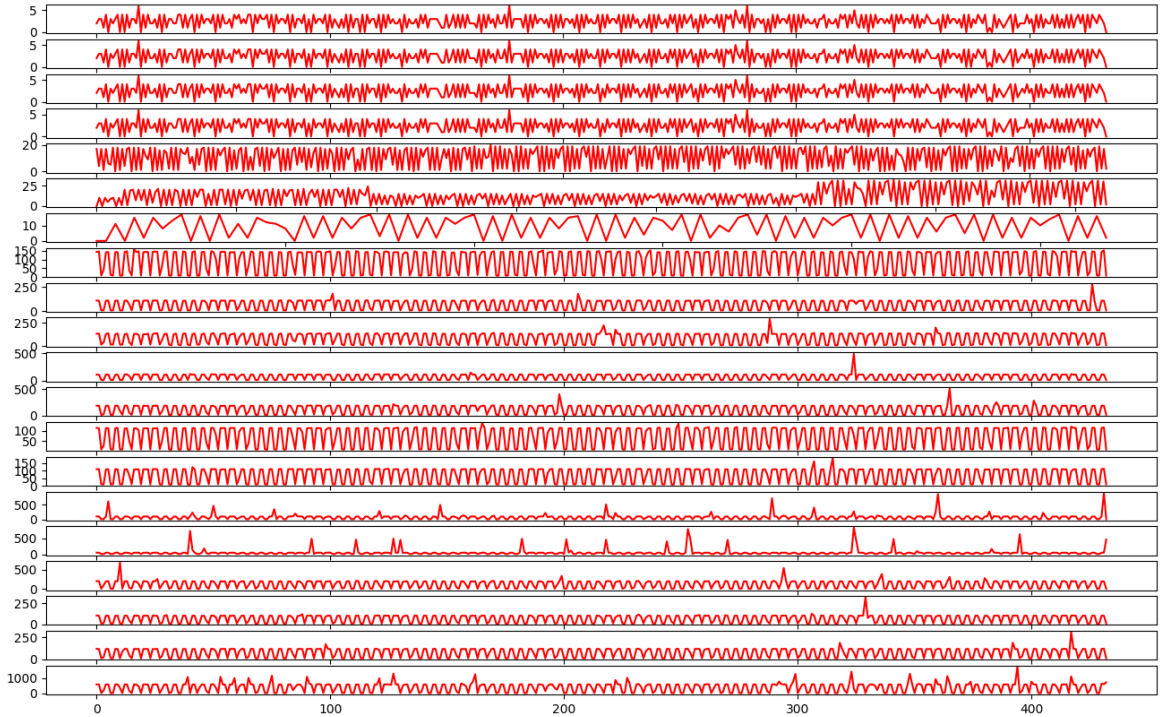



Zhu et al. DesignCon 2019

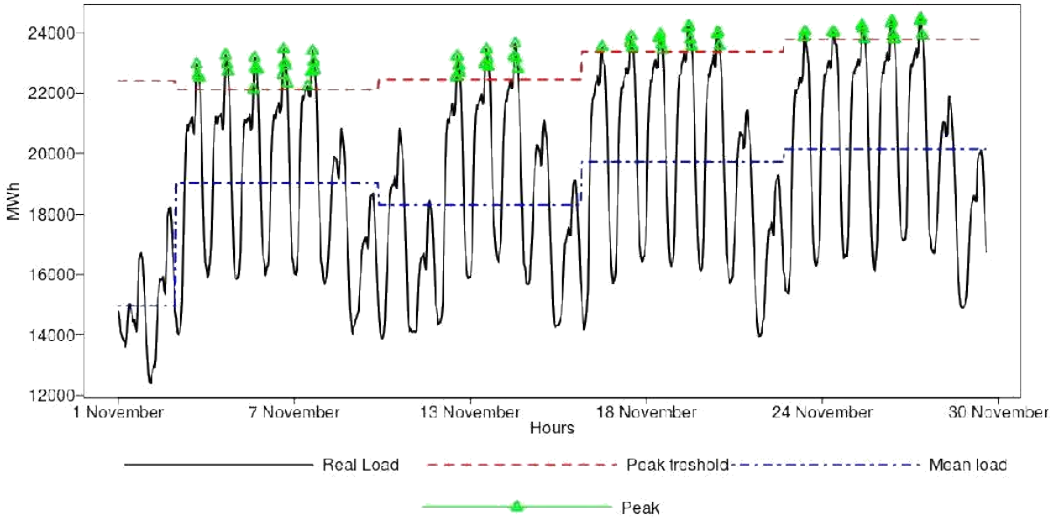
Recommended MATLAB Package : Global Optimization Toolbox

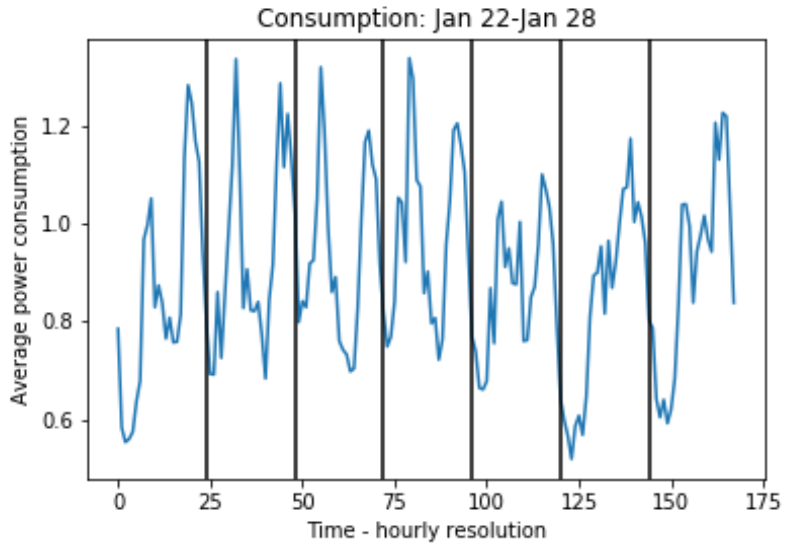
# Deep learning example : High dimension resource demand forecast

# High dimension resources demand forecast



Many system resources show clusters of users and daily pattern

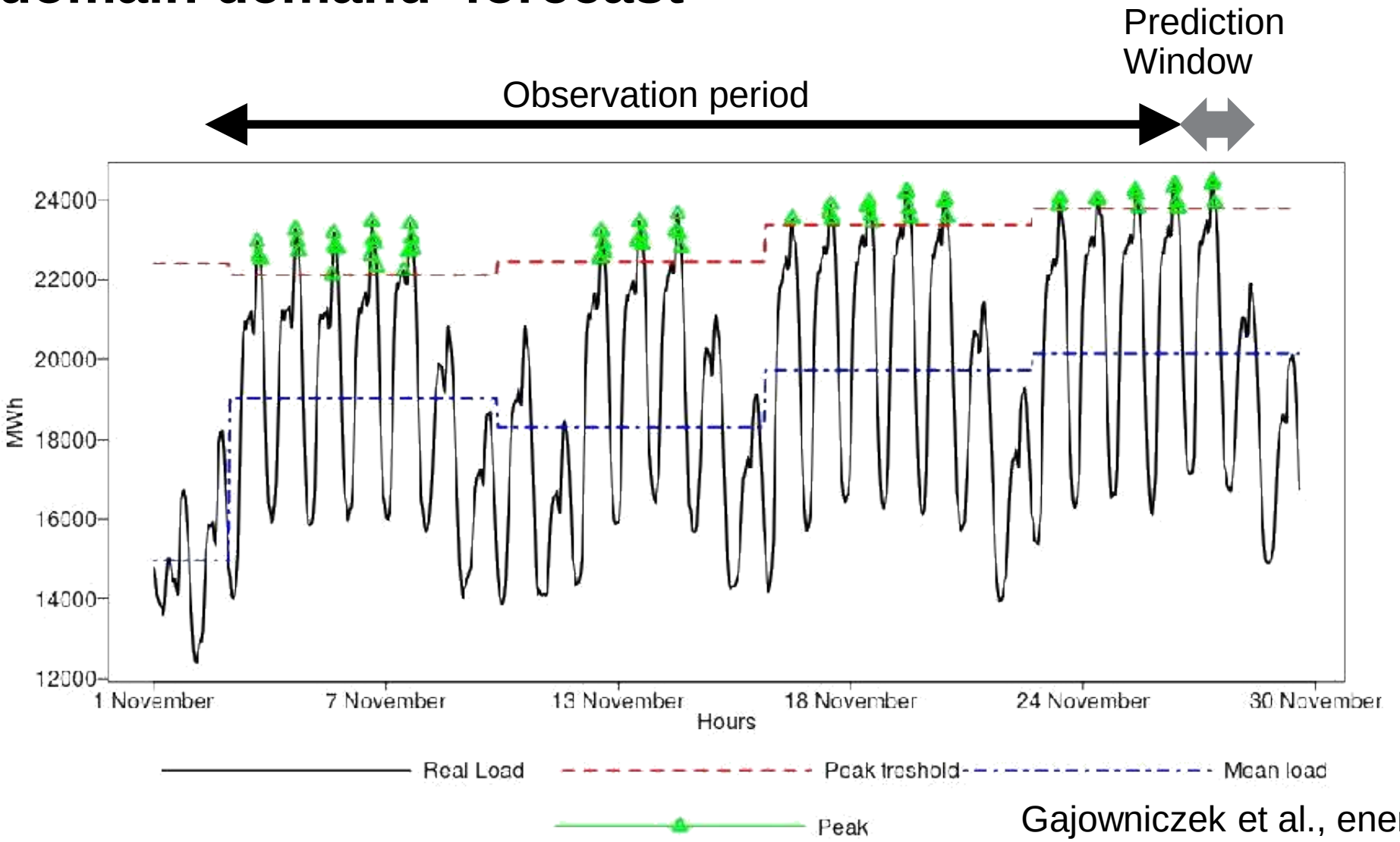



Gajowniczek et al., energies 2017

Hewlett Packard Enterprise

# Time domain demand  forecast




Gajowniczek et al., energies 2017

# Thank you