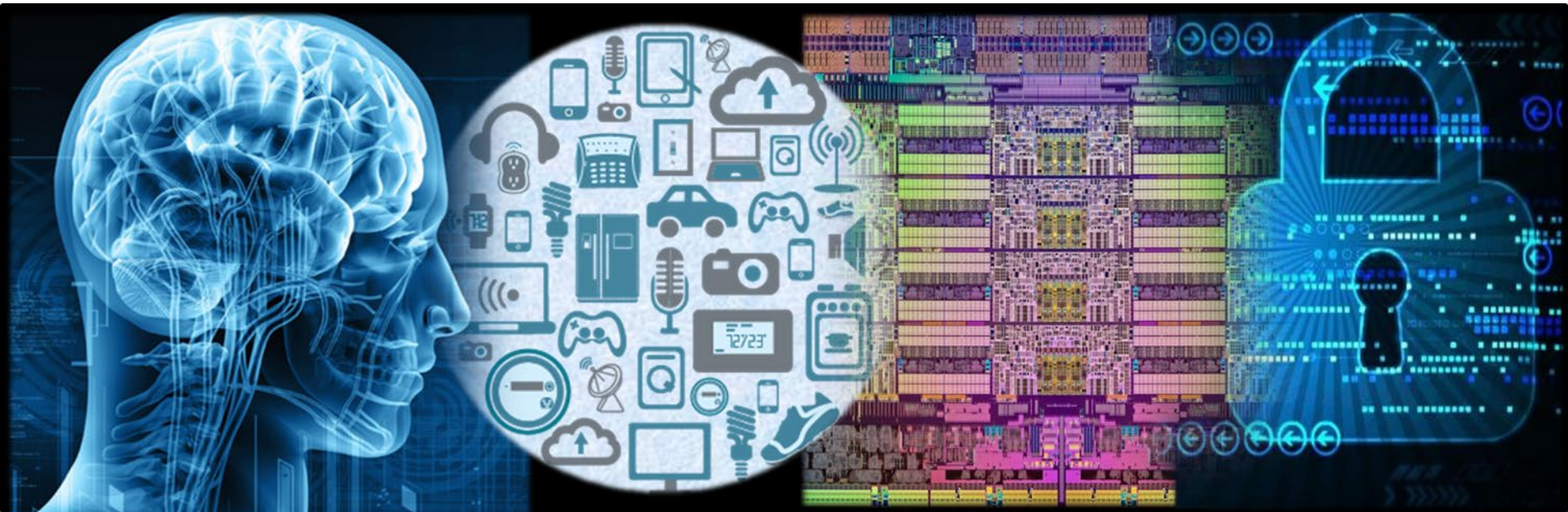
TECHNISCHE UNIVERSITÄT WIEN

CARE Tech
Computer Architecture and Robust Energy-Efficient Technologies Group

Security for Machine Learning Systems

Attacks, Defenses, & Reconfigurability

M. Shafique



Who Ruled the World!

Age of Power

Man-Power (#), Skills, Strength, Courage, etc.

Age of Resources and Industry

Fuel, Industrial Tech., Economic Politics, etc.

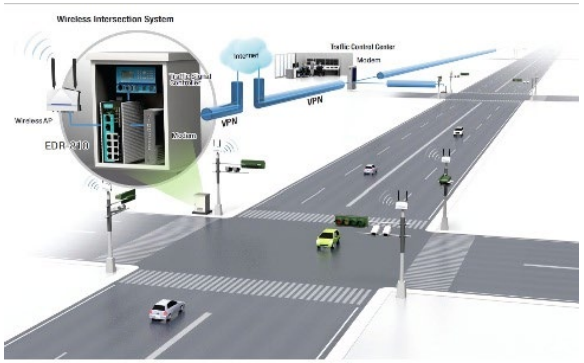
Age of Data and AI

Data is the New Fuel

Innovation in Technology is the New Politics

Nation-wide Race for Dominance in AI

Smart Cyber Physical Systems & Internet-of-Things



Smart Traffic Control

https://www.emaze.com/@ACIOWOWR/IMSA-Slide-Show

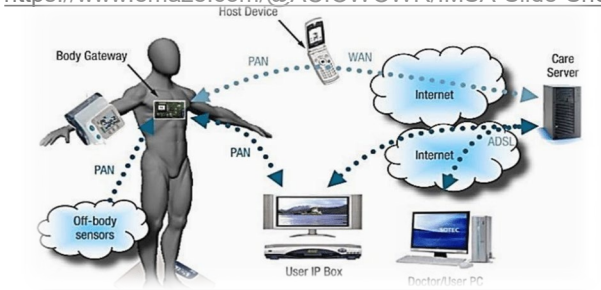
Smart Automobiles

http://www.it5g.com/latest-software-enhancements-in-the-auto-industry/

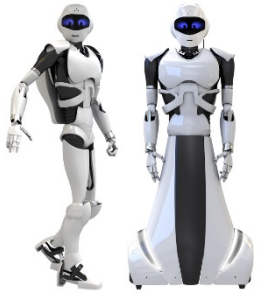
Smart Transport Systems

https://www.automotiveworld.com/analysis/automotive-cyber-physical-systems-next-computing-revolution/

Applications



Smart Health Care



Smart Robots

http://alpha-smart.com/alphaboten

Industry 4.0:
Smart Industrial Automation

https://vimeo.com/145877805

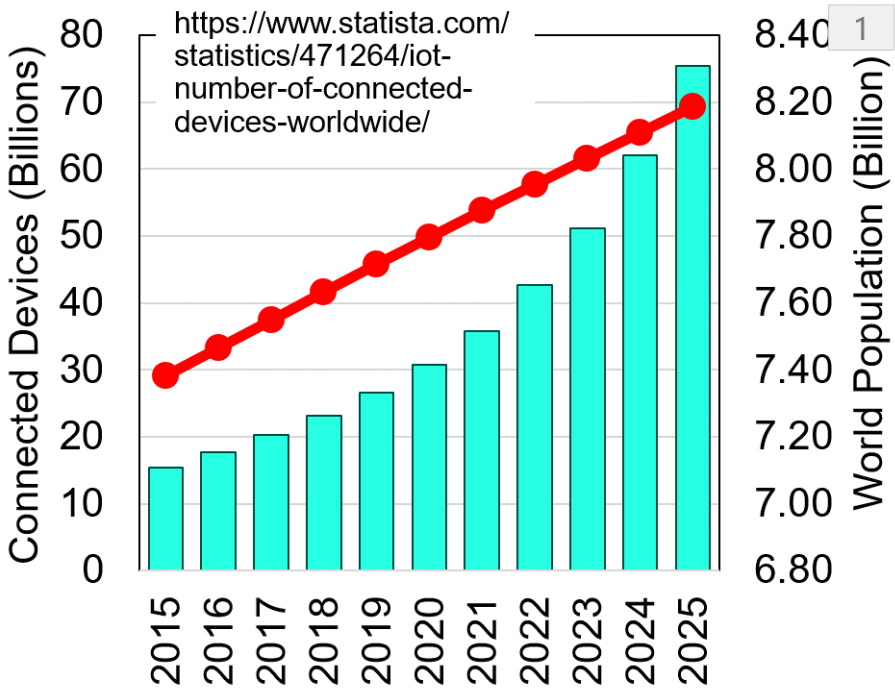
Smart Houses

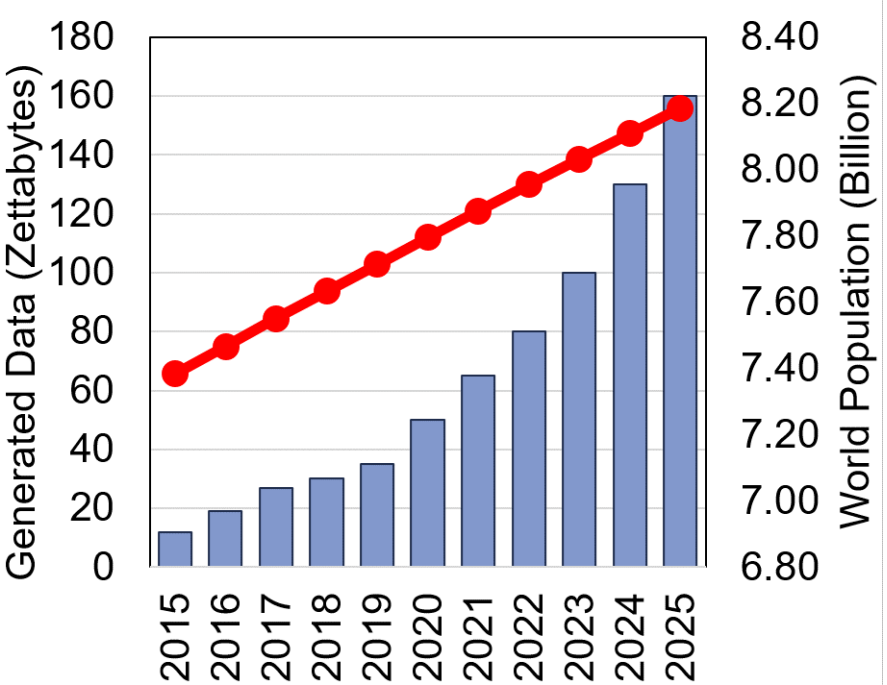
https://www.linkedin.com/pulse/smart-homes-private-secure-future-intelligent-home-tripti-jha

Smart Grids

http://solutions.3m.com/wps/portal/3M/en_EU/SmartGrid/EU-Smart-Grid/

Smart CPS & IoT => The Big Data Processing Challenge!

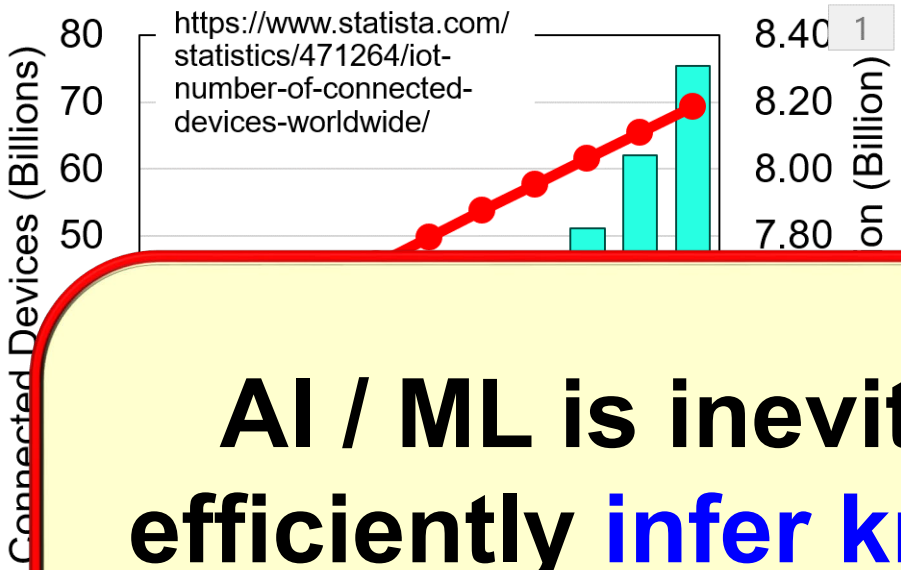




- By 2025: **75 Billion** connected devices
- By 2025: **4,785 interactions** per connected person in one day.

- By 2025: **160 Zettabyte** measured data
- By 2020: for every person on earth, **1.7 MB data/sec** will be generated.

Smart CPS & IoT => The Big Data Processing Challenge!



AI / ML is inevitable, we have to efficiently infer knowledge from the big data, and *derive predictions*

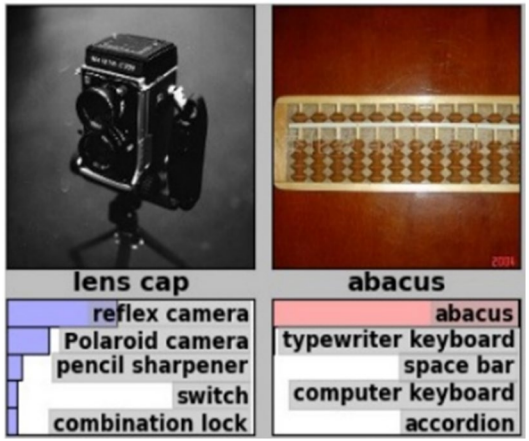
- By 2025: **75 Billion** connected devices
- By 2025: **4,785 interactions** per connected person in one day.

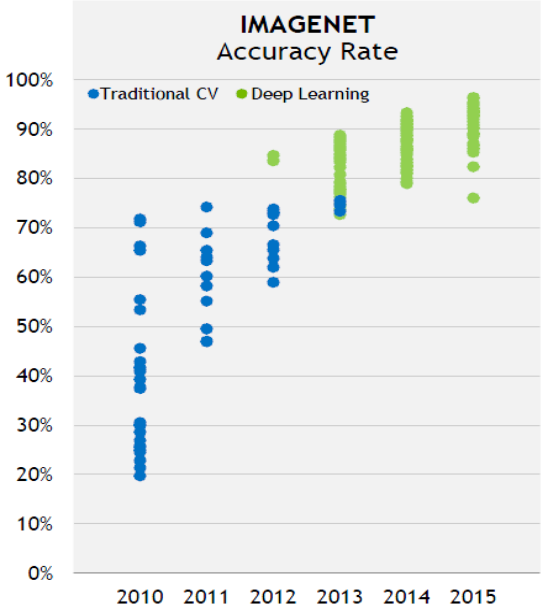
- By 2025: **160 Zettabyte** measured data
- By 2020: for every person on earth, **1.7 MB data/sec** will be generated.

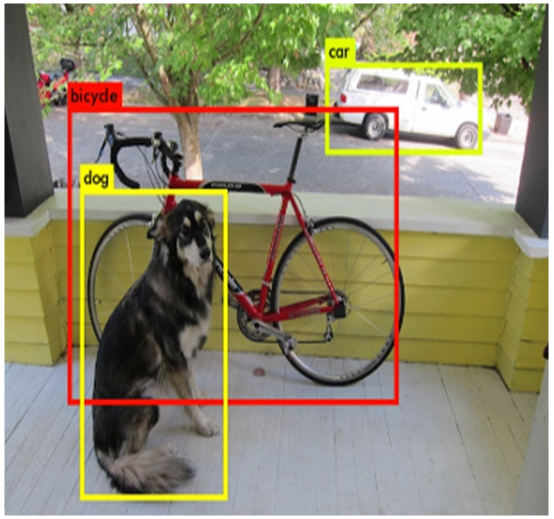
ML Applications => requiring High Efficiency Gains

Autonomous Driving

Image Classification



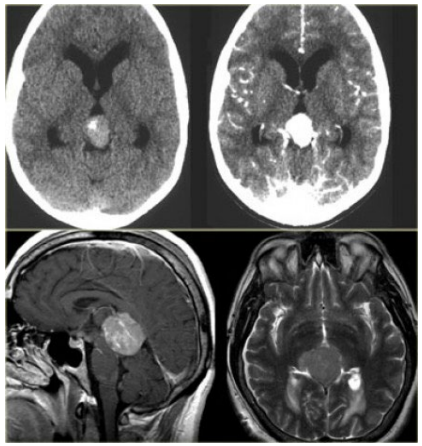




Object Detection & Localization

Machine Translation

Strategy Games



Cancer Detection

Natural Language Processing

Forex/Stocks Trading

Machine Learning and the Computing Efficiency Gap!

Deep Blue vs. Garry Kasparov

IBM RS/6000 SP high-performance computer

900W + Power for 480 dedicated ASICs

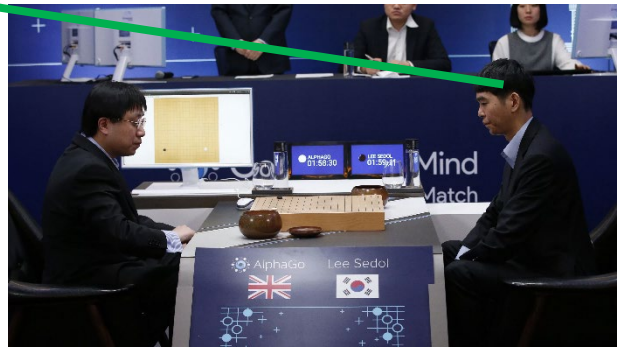
30 P2SC nodes + 480 dedicated ASICs to play chess

20W

1MW

1,920 CPUs and 280 GPUs

AlphaGO vs. Lee Sedol



Google's cloud computing

1996 Chess

2011 Jeopardy

2016 GO

IBM Watson vs. Brad Rutter and Ken Jennings

200KW

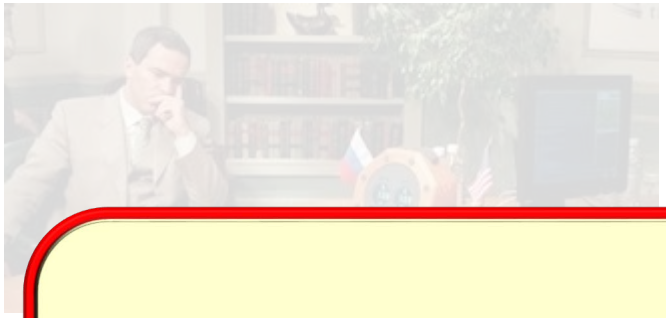
2,880 POWER7 processor threads and 16 terabytes of RAM

IBM Blue Gene supercomputer

Data Source: https://jacquesmattheij.com/another-way-of-looking-at-lee-sedol-vs-alphago

Machine Learning and the Computing Efficiency Gap!

Deep Blue vs. Garry Kasparov



900W + Power for 480 dedicated ASICs

30 P2SC nodes + 480 dedicated

20W

1MW

1,920 CPUs

AlphaGO vs. Lee Sedol

We need 10,000x Gains to Bridge this Gap!!!

200KW

2,880 POWER7 processor threads and 16 terabytes of RAM

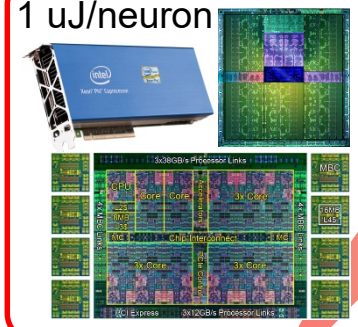
IBM Blue Gene supercomputer

Data Source: https://jacquesmattheij.com/another-way-of-looking-at-lee-sedol-vs-alphago

Brain-Inspired Computing: Research@CARE-Tech.

High Performance, Energy Efficiency, Configurability, Reliability & Security

1 uJ/neuron



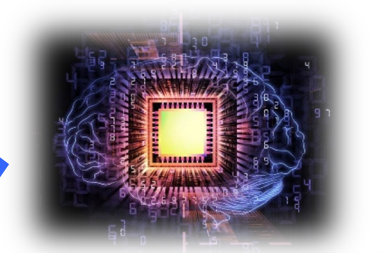
Software (Multi-Cores, GPUs)

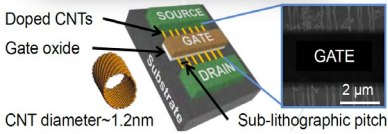
Approximate Computing

Deep Learning Architectures

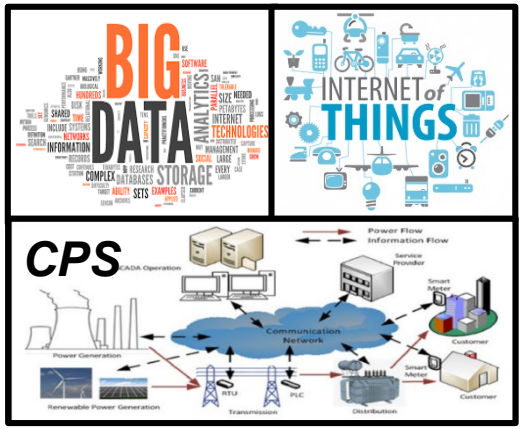
Neuromorphic Architectures and *Beyond*

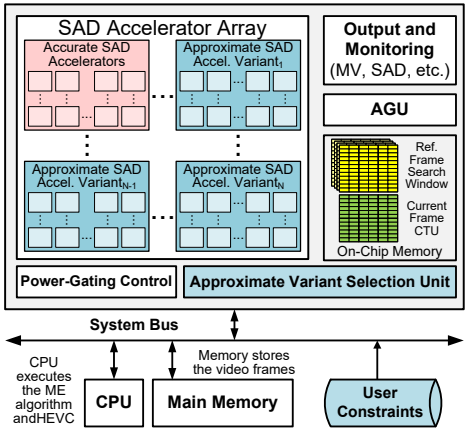
Post-CMOS Technologies





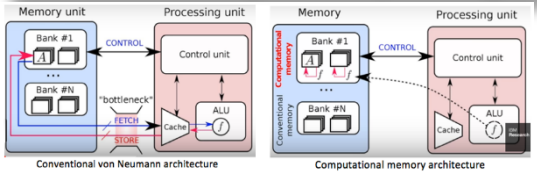
Source: *S. Mitra, P. Wong (Stanford), C. Mackin (MIT), J. Zhang (Google)*



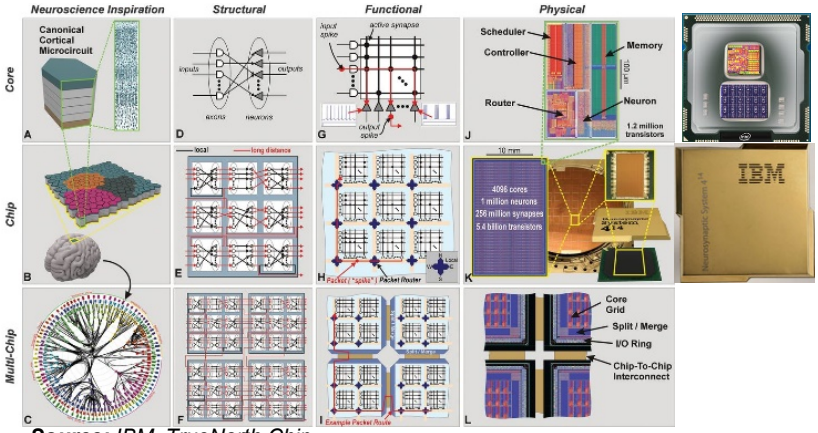


In-Memory Computing

Source: *Huawei Kirin Chip*



Source: IBM Research



Source: *IBM, TrueNorth Chip*

Robustness for Machine Learning: News Feed

Beware: Galaxy S10's Facial Recognition Easily Fooled with a Photo

Jesus Diaz • Freelance Writer
Updated Mar 11, 2019

Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital

Hackers trick a Tesla into veering into the wrong lane

https://www.youtube.com/watch?v=a7L51u23YoM

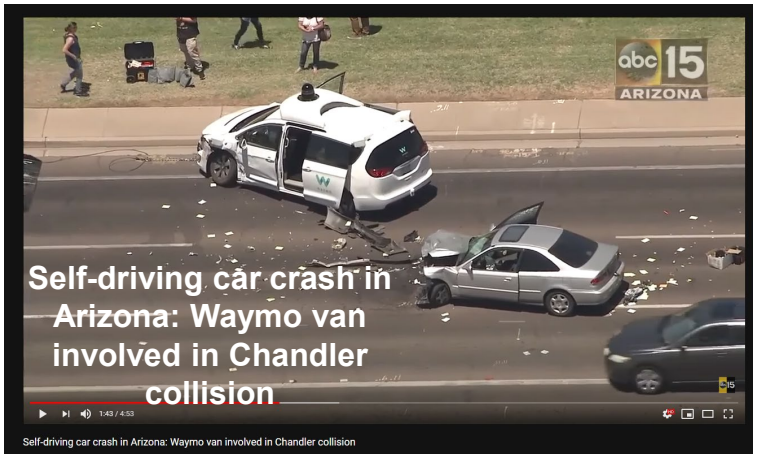
Tesla Model 3: Autopilot engaged during fatal crash

BBC

17 May 2019

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to distinguish a white tractor-trailer crossing the highway against a bright sky



https://www.technologyreview.com/f/613254/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic/

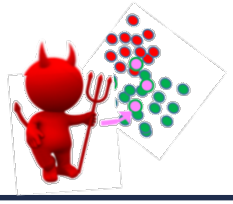
Risks that Arise from Adversarial Examples

FACE RECOGNITION

"Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition" [Sharif et. al. CCS 2016]

TEXT CLASSIFICATION (Spam)

"Crafting Adversarial Input Sequences for Recurrent Neural Networks" [Papernot et. al. 2016]

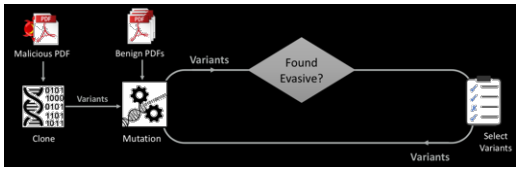


SPEECH RECOGNITION

"Hidden Voice Commands", [Carlini et. al. USENIX 2016]

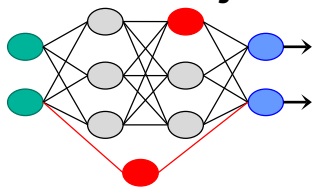
MALWARE DETECTION

"Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers" [Xu et. al. NDSS2016]

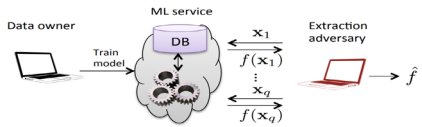


Reliability and Security for Machine Learning Systems

Neural Trojans



Model Stealing Attacks

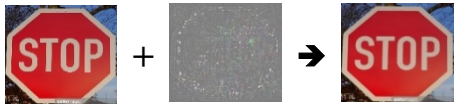


Src: Tramer@*UsenixSec16*

Backdoor Attacks

Src: Reiter@*CCS2016*

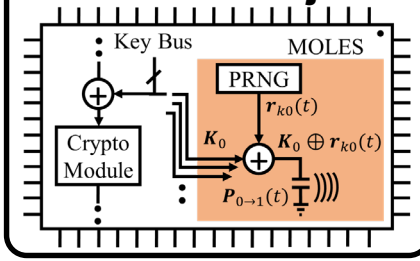
Adversarial Attacks



Stop Sign

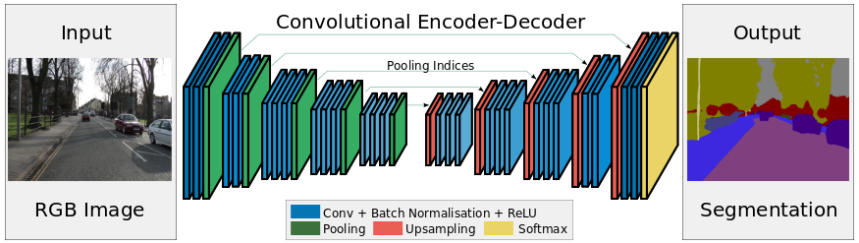
Speed Limit (60km/h)

Hardware Trojans

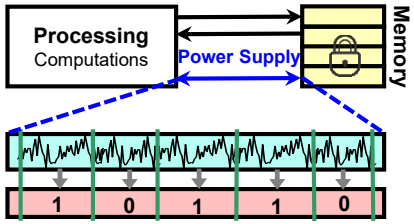


Machine Learning-based System

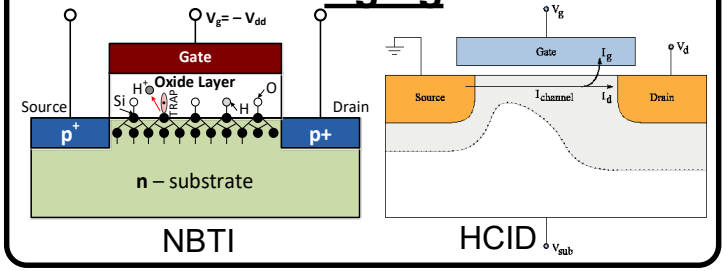
Src: google Image



Side Channel Attacks



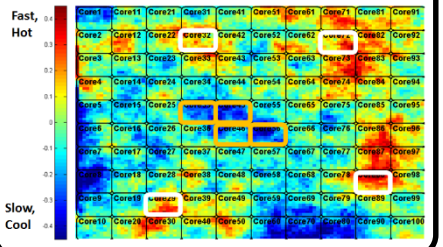
Aging



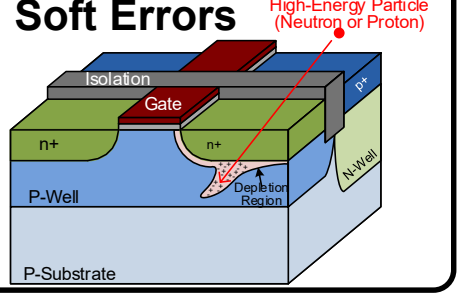
NBTI

HCID

Process Variations



Soft Errors



- M. A. Hanif, F. Khalid, R. V. W. Putra, S. Rehman, M. Shafique, "Robust Machine Learning Systems: Reliability and Security for Deep Neural Networks", in IOLTS-2018, Platja d'Aro, Spain, pp. 257 - 260.
- F. Kriebel, S. Rehman, M. A. Hanif, F. Khalid, M. Shafique, "Robustness for Smart Cyber-Physical Systems and Internet-of-Things: From Adaptive Robustness Methods to Reliability and Security for Machine Learning", ISVLSI-2018, Hong Kong, China, pp. 581-586.

Details: Refer to Our Works

- **Tutorial** on *"Adversarial ML and Vehicular Networks: Strategies for Attack and Defense"* at IEEE 89th Vehicular Technology Conference (VTC).

- **Selectd Research Papers**
  - **DAC'19:** "Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities".
  - **ICML'19** Workshop on Robust ML: "CapsAttacks: Robust and Imperceptible Adversarial Attacks on Capsule Networks"
  - **DATE'19:** "FAdeML: Understanding the Impact of Pre-Processing Noise Filtering on Adversarial Machine Learning"
  - **IOLTS'19:** "TrISec: Training Data-Unaware Imperceptible Security Attacks on Deep Neural Networks".
  - **IOLTS'19:** "QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks".

Details: Refer to Our Works

- **Tutorial** on *"Adversarial ML and Vehicular Networks: Strategies for Attack and Defense"* at IEEE 89th Vehicular Technology Conference (VTC).

- **Selectd Research Papers**

RTML Workshop at ITC 2019

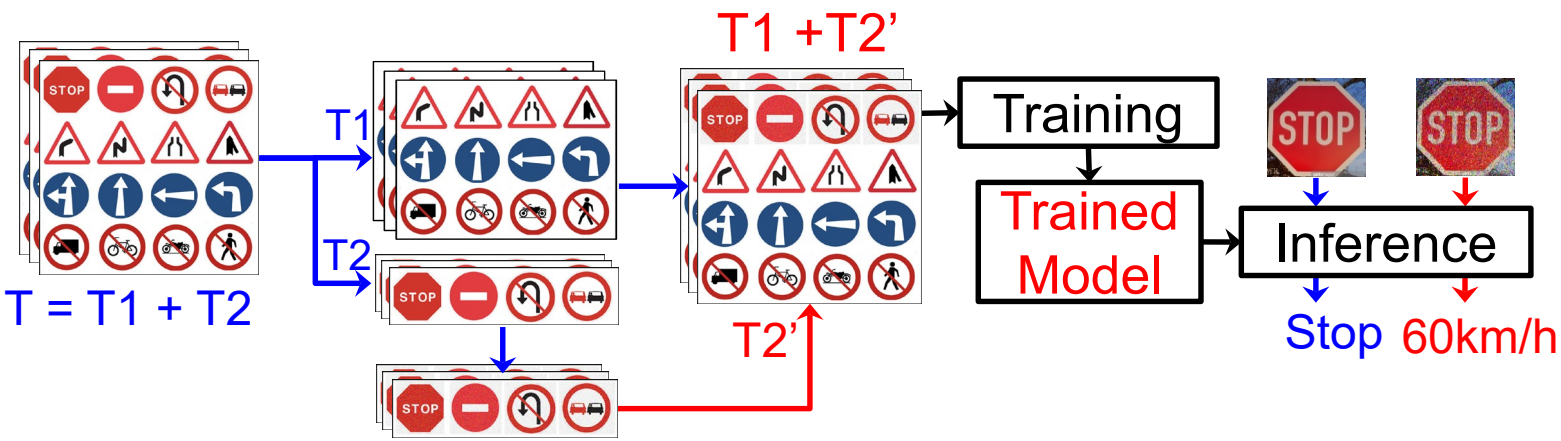
"Open for Submissions"

Processing Noise Filtering on Adversarial Machine Learning"

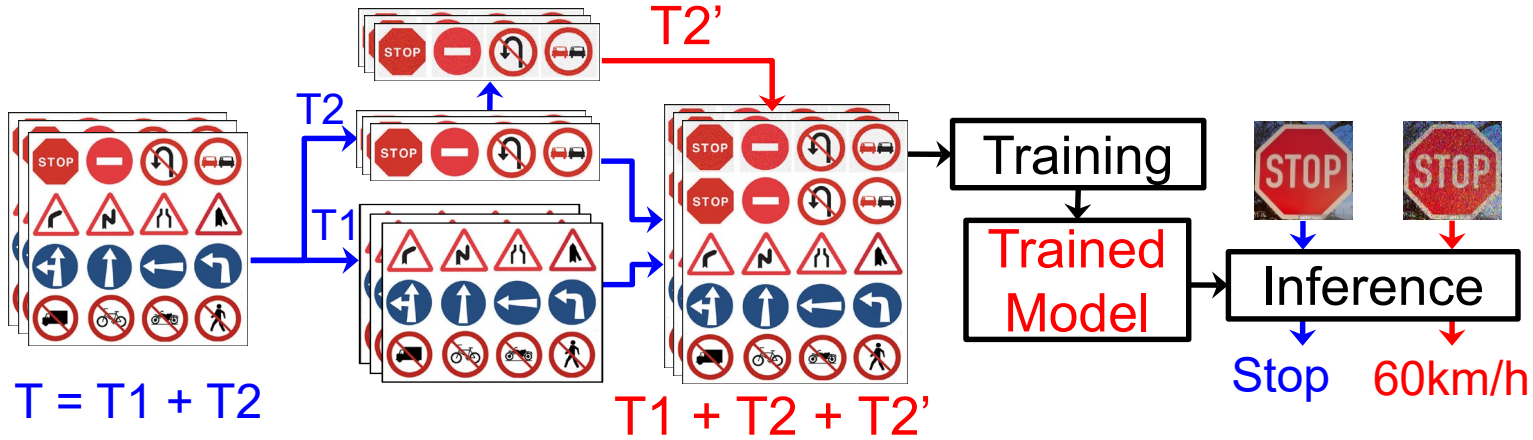
- **IOLTS'19:** "TrISec: Training Data-Unaware Imperceptible Security Attacks on Deep Neural Networks".
- **IOLTS'19:** "QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks".

Data Poisoning Attacks during Training

Attack 1: In this attack certain portion of the original dataset is intruded which is then used for training.



Attack 2: Instead of perturbing the original dataset, in this attack we extend the dataset with intruded samples which is then used for training.



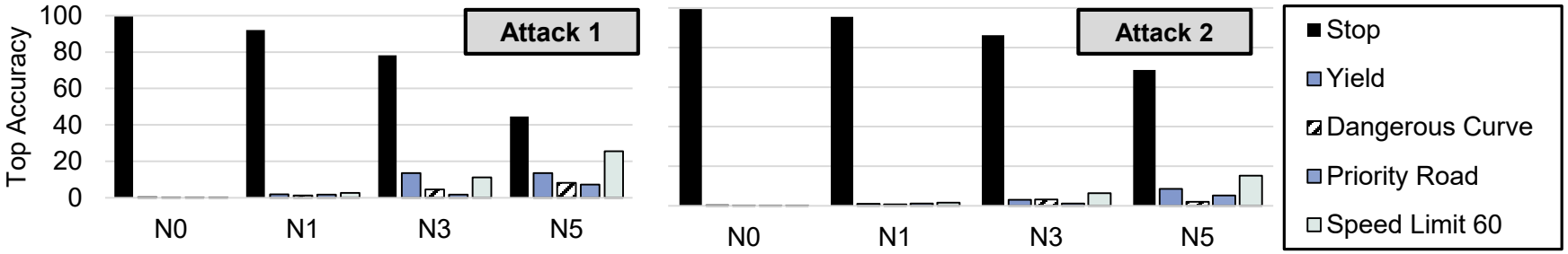
Data Poisoning Attacks: Training

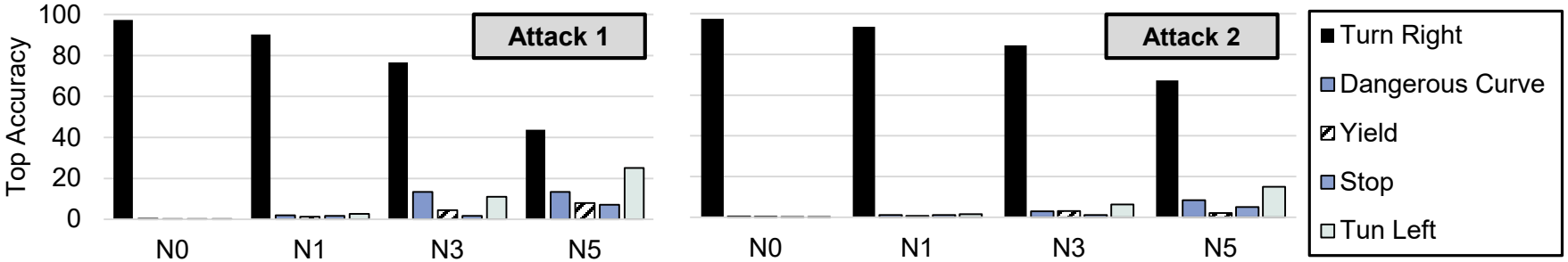
N0: No Noise

N1: 1% Noise

N2: .% Noise

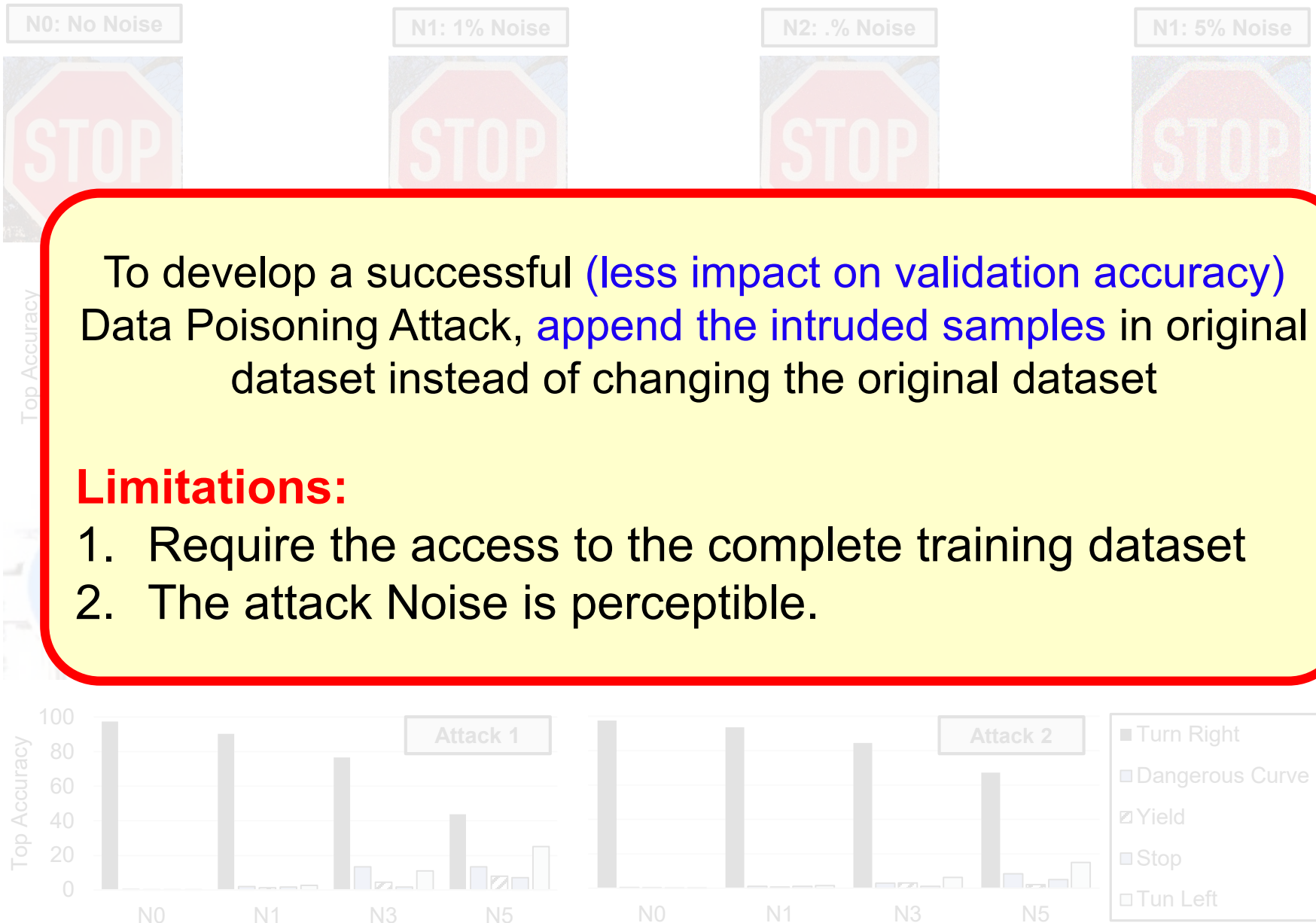
N1: 5% Noise





Faiq, Shafique et al. @ FIT 2018, DSD'18, IOLTS 2019

Data Poisoning Attacks: Training



To develop a successful (less impact on validation accuracy) Data Poisoning Attack, append the intruded samples in original dataset instead of changing the original dataset

Limitations:

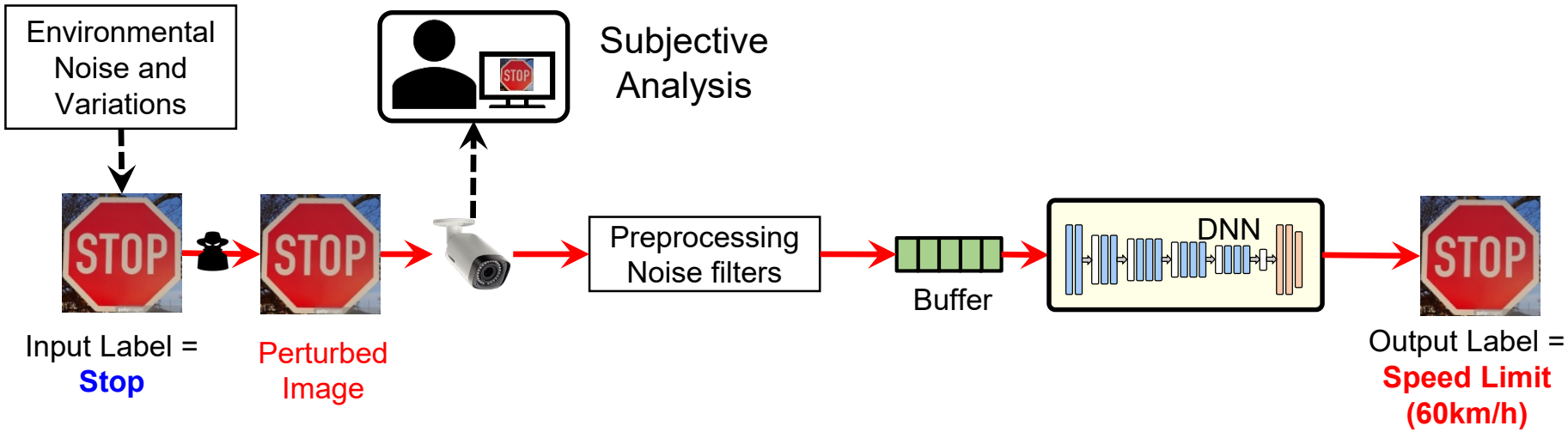
1. Require the access to the complete training dataset
2. The attack Noise is perceptible.

Faiq, Shafique et al. @ FIT 2018, DSD'18, IOLTS 2019

Adversarial Attacks on ML-Systems

Specially crafted **imperceptible noise** to **misclassify** the ML System

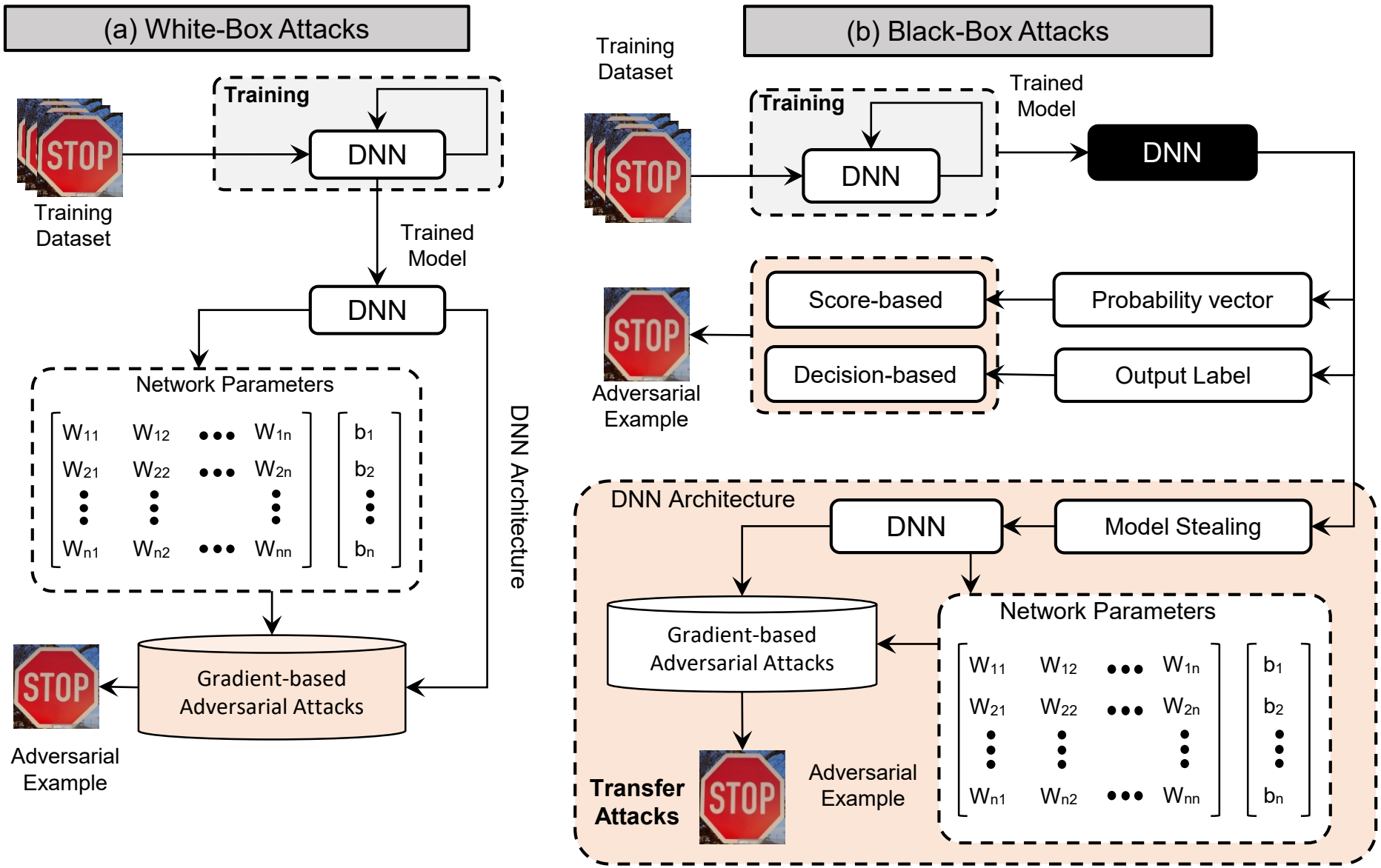
- ❑ **Imperceptibility**
- ❑ **Robustness**



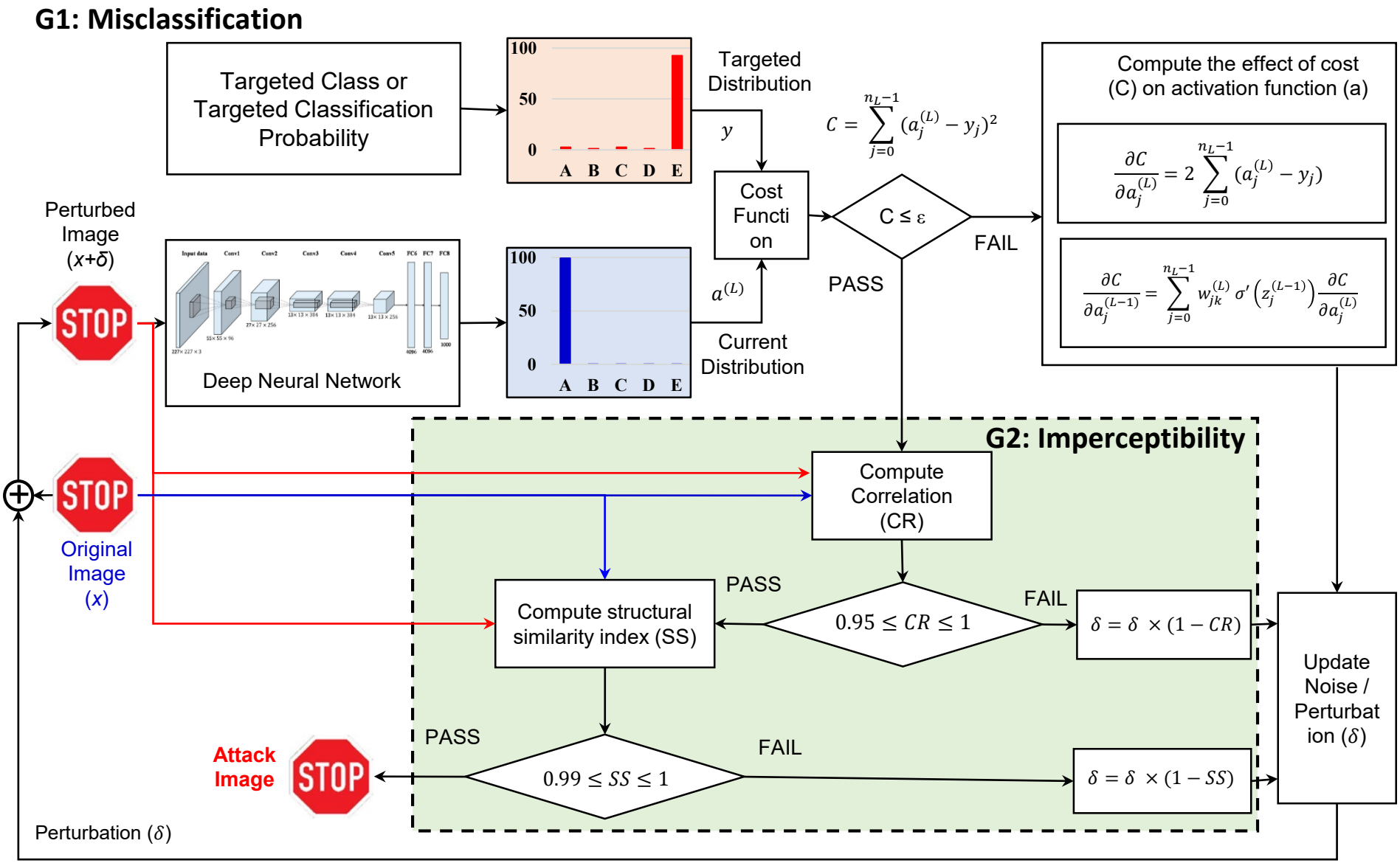
Challenges

- Ensure the imperceptibility
- Handling the impact of environmental Noise

Adversarial Attacks



TrISec: Training Data Unaware Adversarial Attack



F. Khalid, M. A. Hanif, S. Rehman, M. Shafique, "TrISec: Training Data-Unaware Imperceptible Security Attacks on Machine Learning Modules of Autonomous Vehicles", IOLTS2019. [Link: https://arxiv.org/abs/1811.01031]

Imperceptible Data Poisoning Attacks during Inference

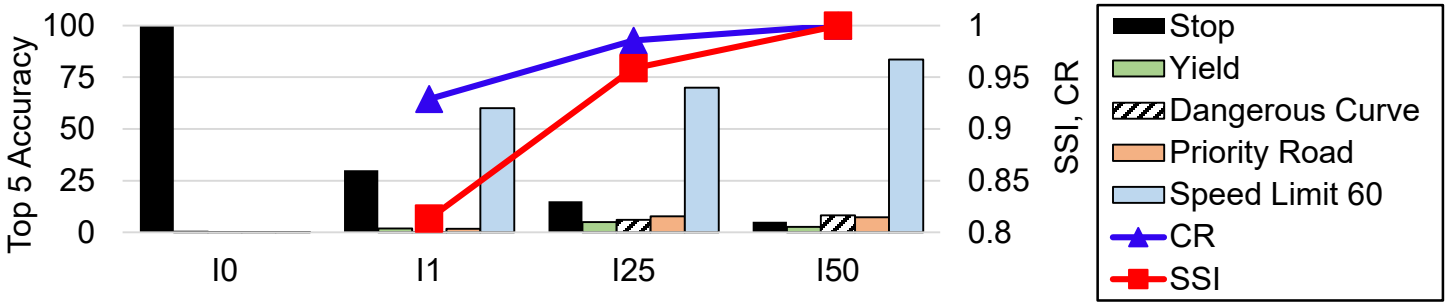
I0: No Noise

I1: 1st Iteration

I25: 25Kth Iteration

I50: 50Kth Iteration

Top 5 Accuracy, Correlation Coefficient (CR) and Structural Similarity Index (SSI)

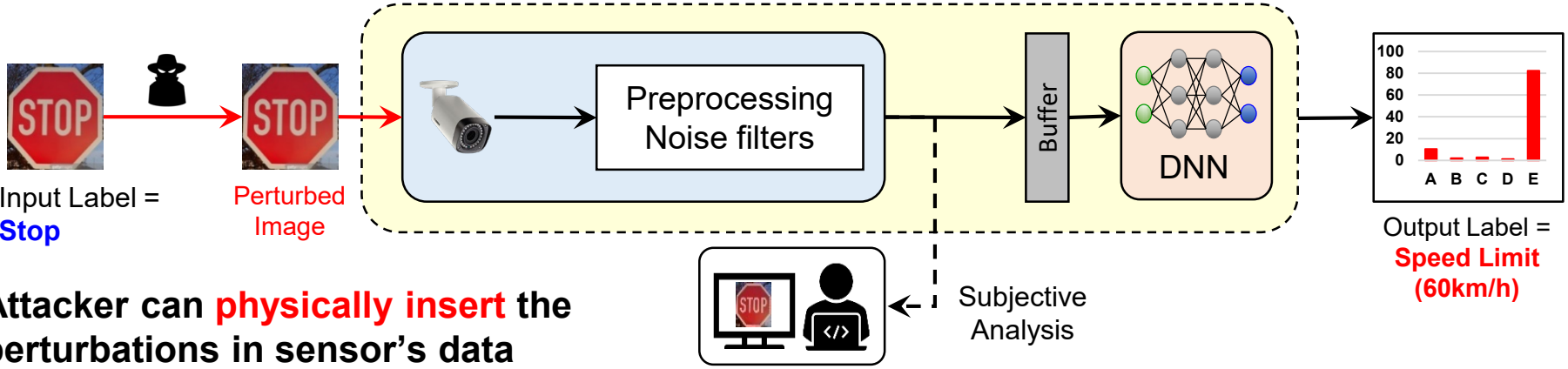


Key Insights:

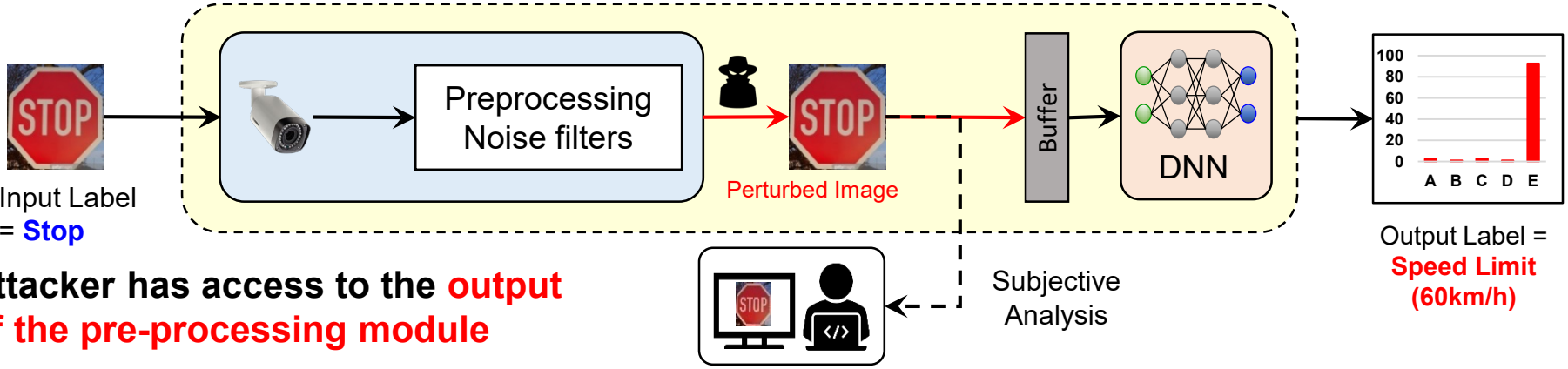
- Instead of training data samples, the probability vector at the output can be used to optimize the attack noise.
- To ensure the imperceptibility, additional similarity and correlation checks are required

F. Khalid, M. A. Hanif, S. Rehman, M. Shafique, "**TrISec**: Training Data-Unaware Imperceptible Security Attacks on Machine Learning Modules of Autonomous Vehicles", IOLTS 2019. [Link: https://arxiv.org/abs/1811.01031]

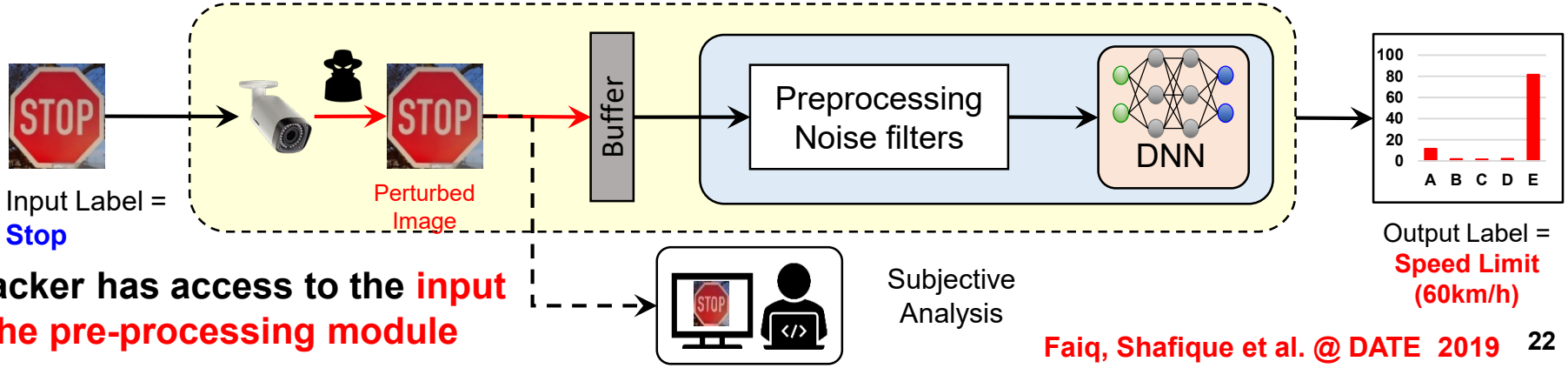
Practical Implications of Adversarial Perturbations



Attacker can physically insert the perturbations in sensor's data



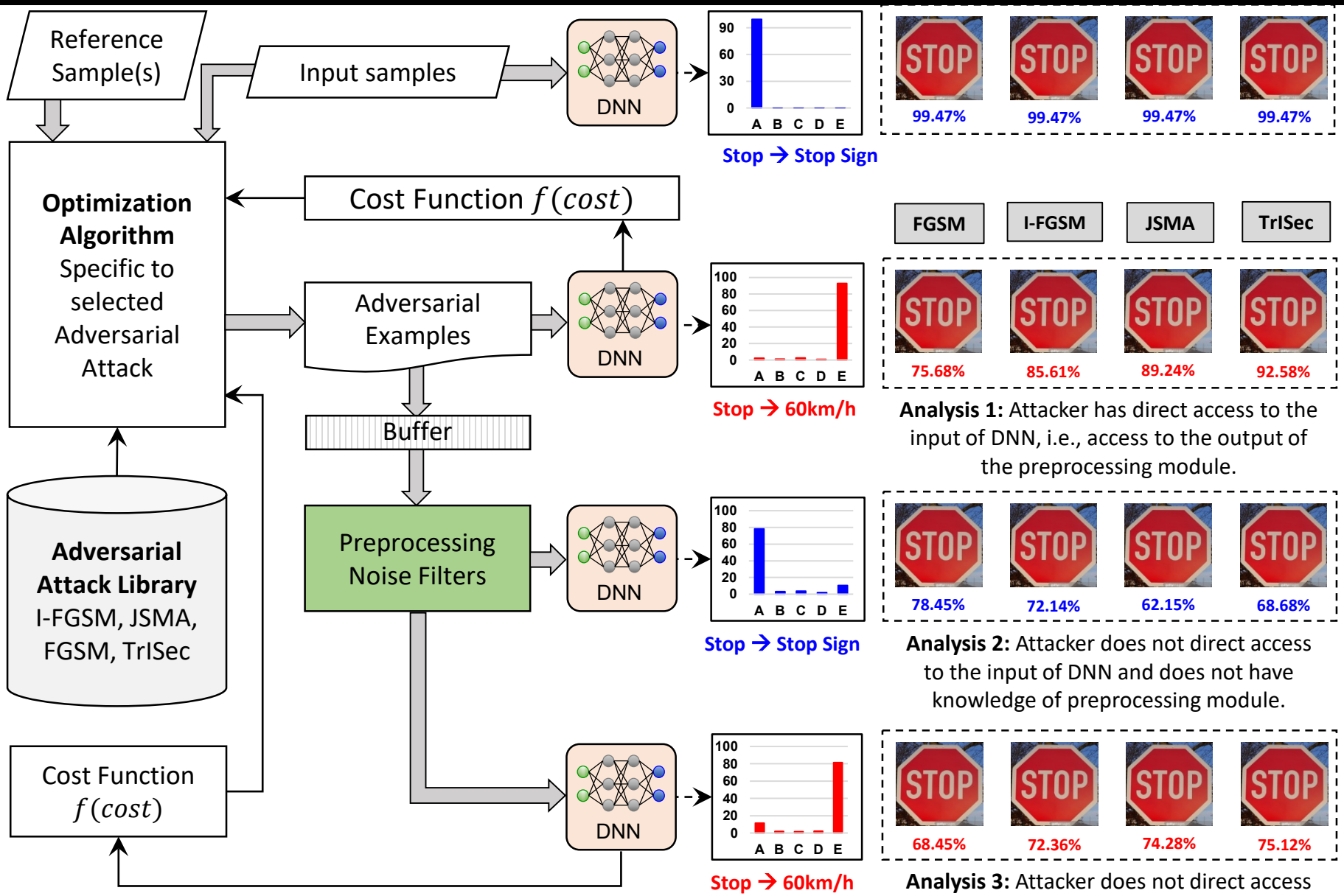
Attacker has access to the output of the pre-processing module



Attacker has access to the input of the pre-processing module

Faiq, Shafique et al. @ DATE 2019

Potential Defenses: FAdeML



Faiq, Shafique et al. @ DATE 2019

Defenses Against Adversarial Examples

- **Goal**
  - Increase the **perceptibility** of the attack noise

Original | Before Defense | After Defense

Automobile | Deer | Deer

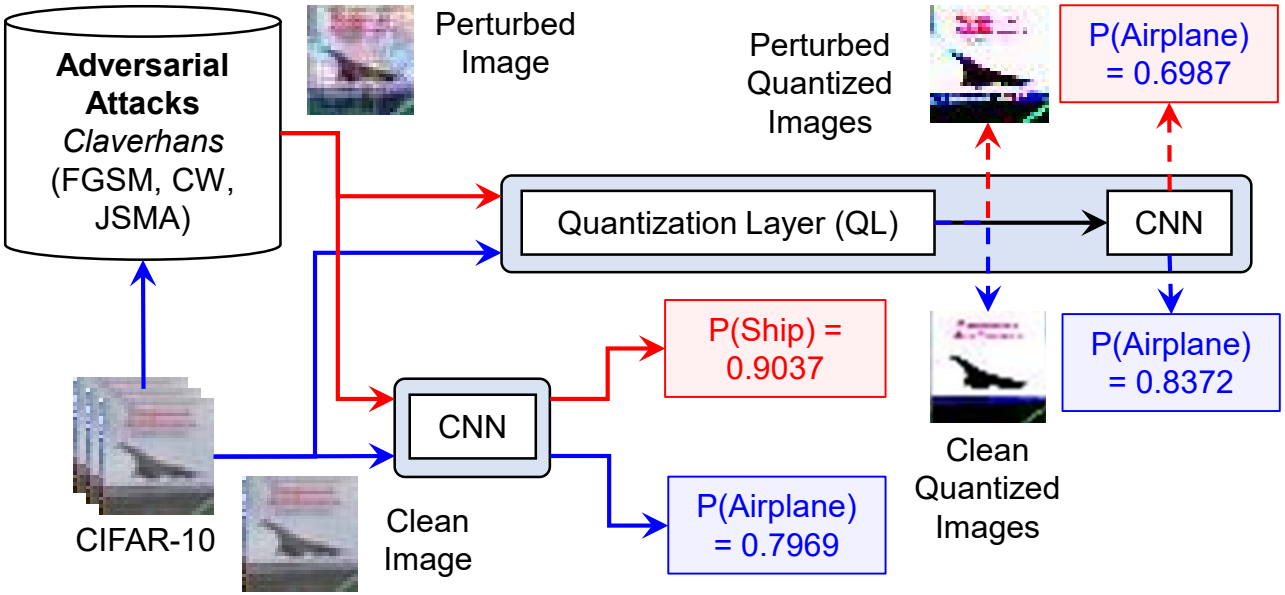
  - Decrease the **robustness** of the attack, i.e., nullify the effects of imperceptible attack Noise

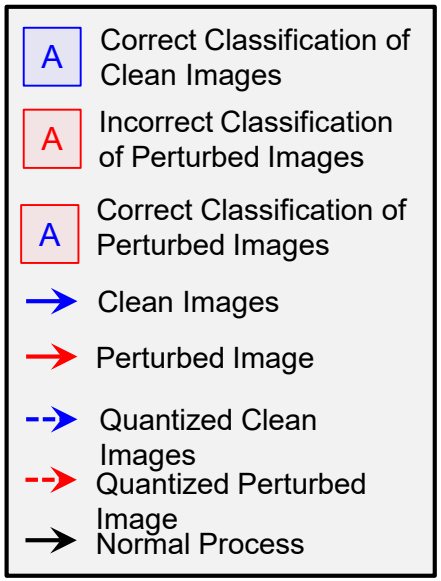
Original | Before Defense | After Defense

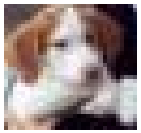
Automobile | Airplane | Automobile

Faiq, Shafique et al. @ IOLTS 2019

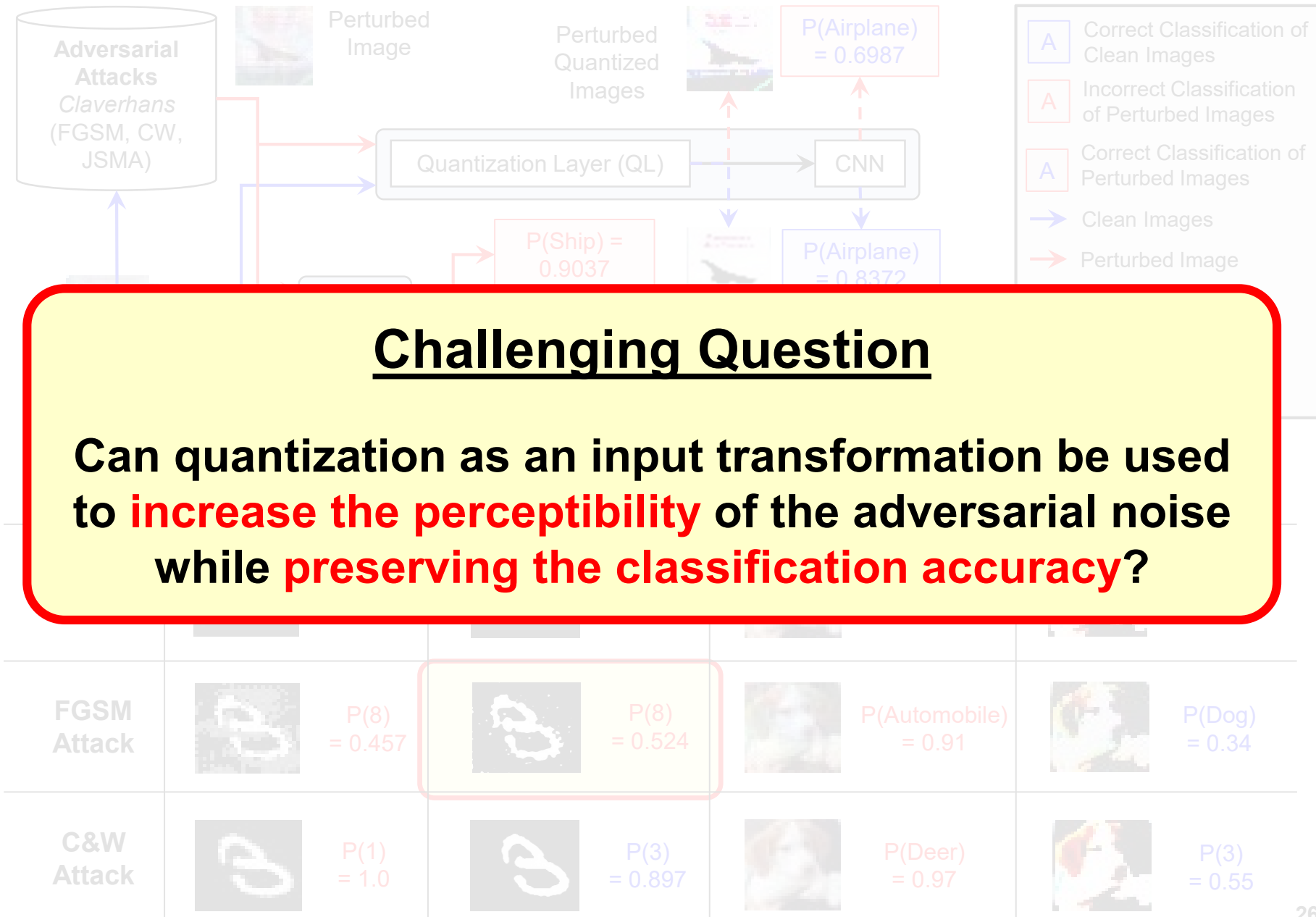
IOLTS'19: Effects of Quantization on Adversarial ML





	Without Quantization Layer	**With** Quantization Layer	**Without** Quantization Layer	**With** Quantization Layer
No Attack	P(3) = 0.862	P(3) = 0.911	P(Dog) = 0.75	P(Dog) = 0.69
FGSM Attack	P(8) = 0.457	P(8) = 0.524	P(Automobile) = 0.91	P(Dog) = 0.34
C&W Attack	P(1) = 1.0	P(3) = 0.897	 P(Deer) = 0.97	P(3) = 0.55

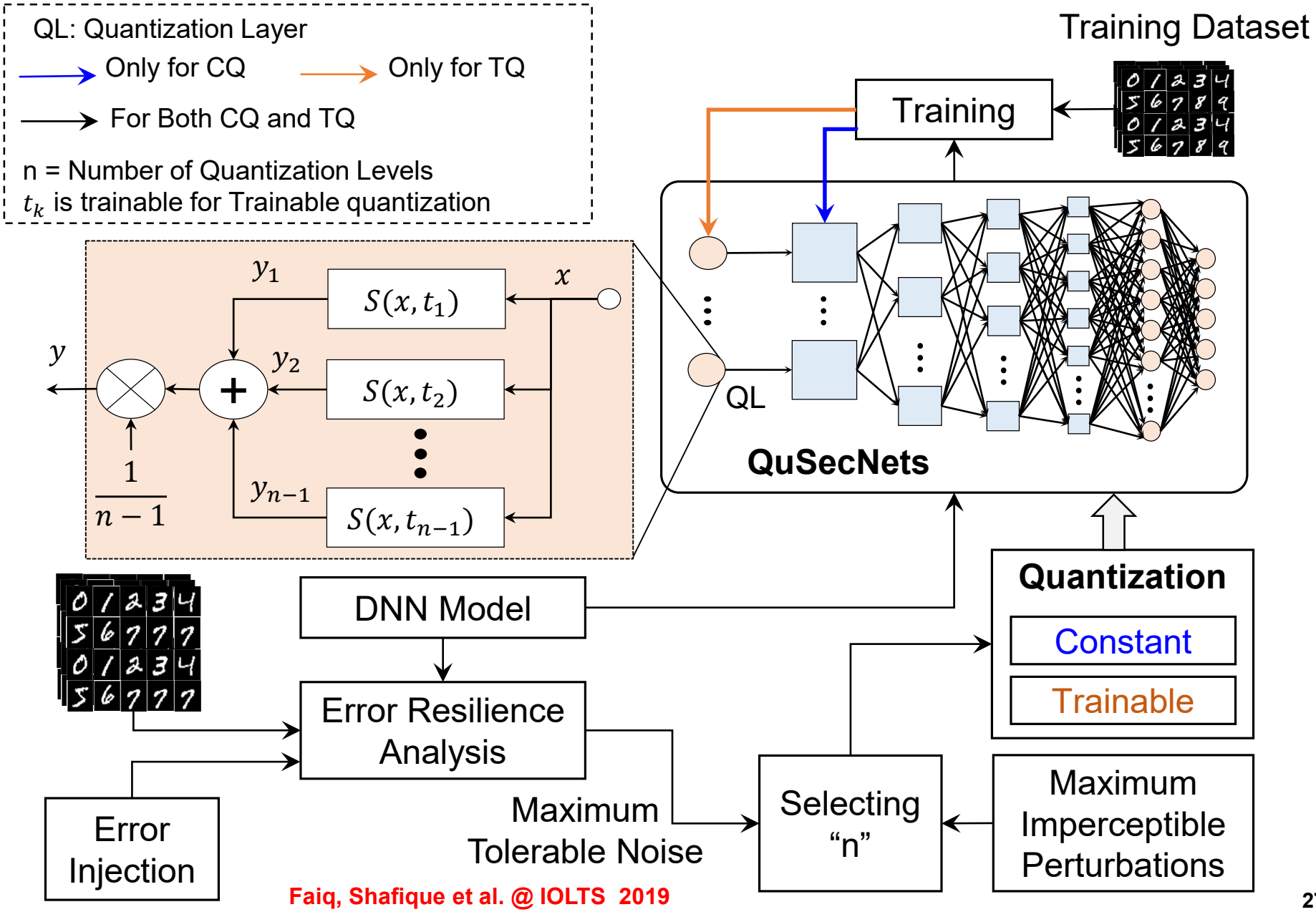
iOLTS'19: Effects of Quantization on Adversarial ML



Challenging Question

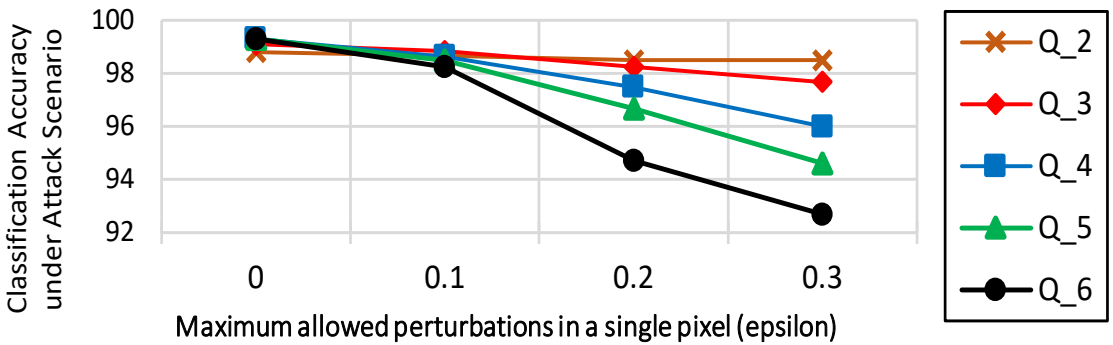
Can quantization as an input transformation be used to increase the perceptibility of the adversarial noise while preserving the classification accuracy?

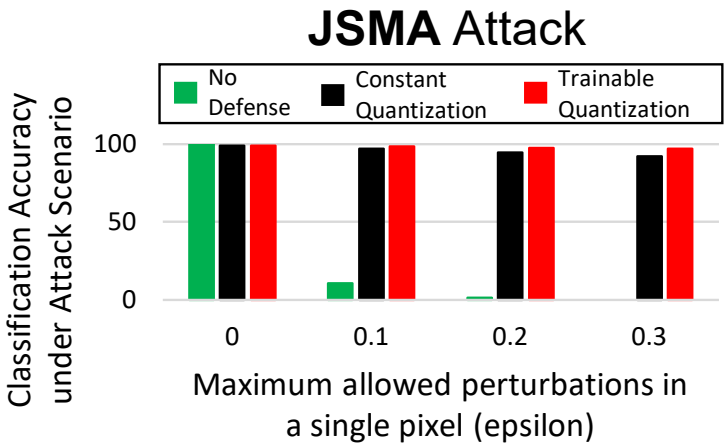
IOLTS'19: QuSecNets – Quantization-based Defense



Faiq, Shafique et al. @ IOLTS 2019

Experimental Results





FSGM Attack

Automobile

No Defense

Airplane

Airplane

Constant Quantization

Bird

Ship

Trainable Quantization

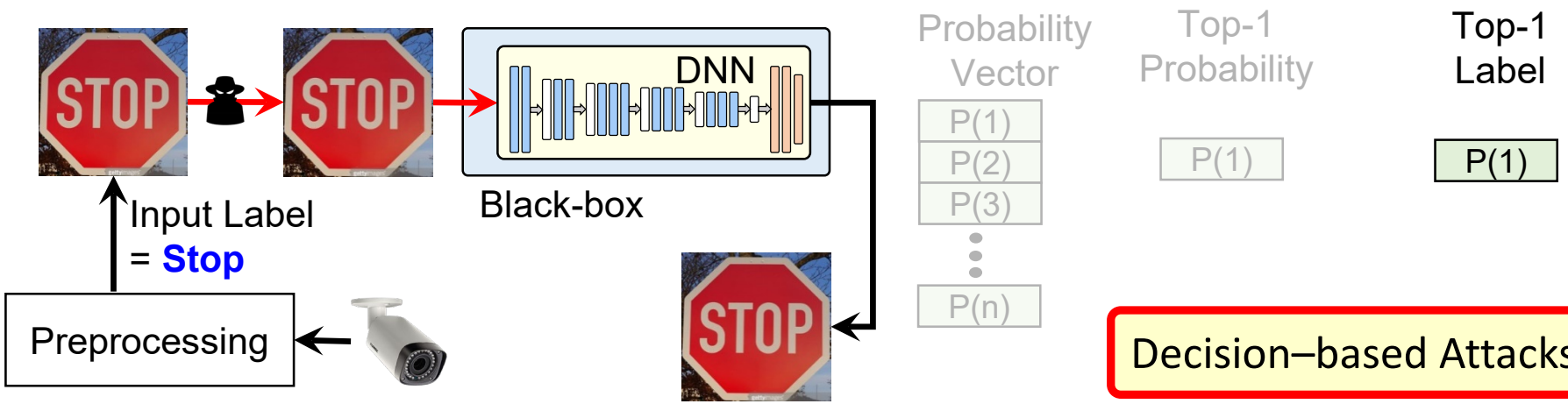
Deer

Frog

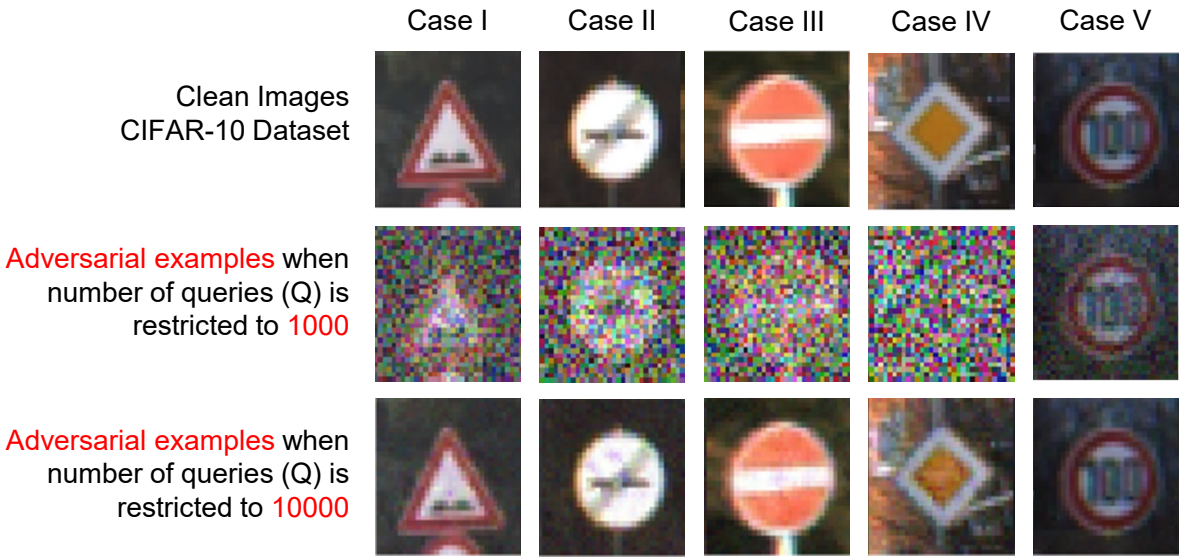
Trainable Quantization performs better than Constant Quantization

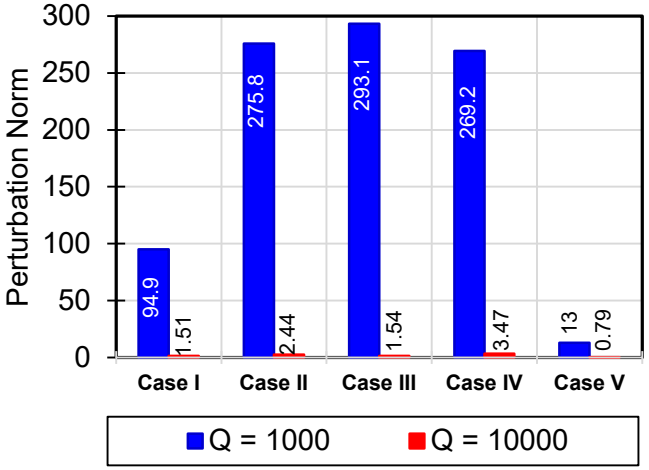
- ❑ Increase the perceptibility of the adversarial noise
- ❑ Increase the classification accuracy under attack

Decision-based Adversarial Attacks



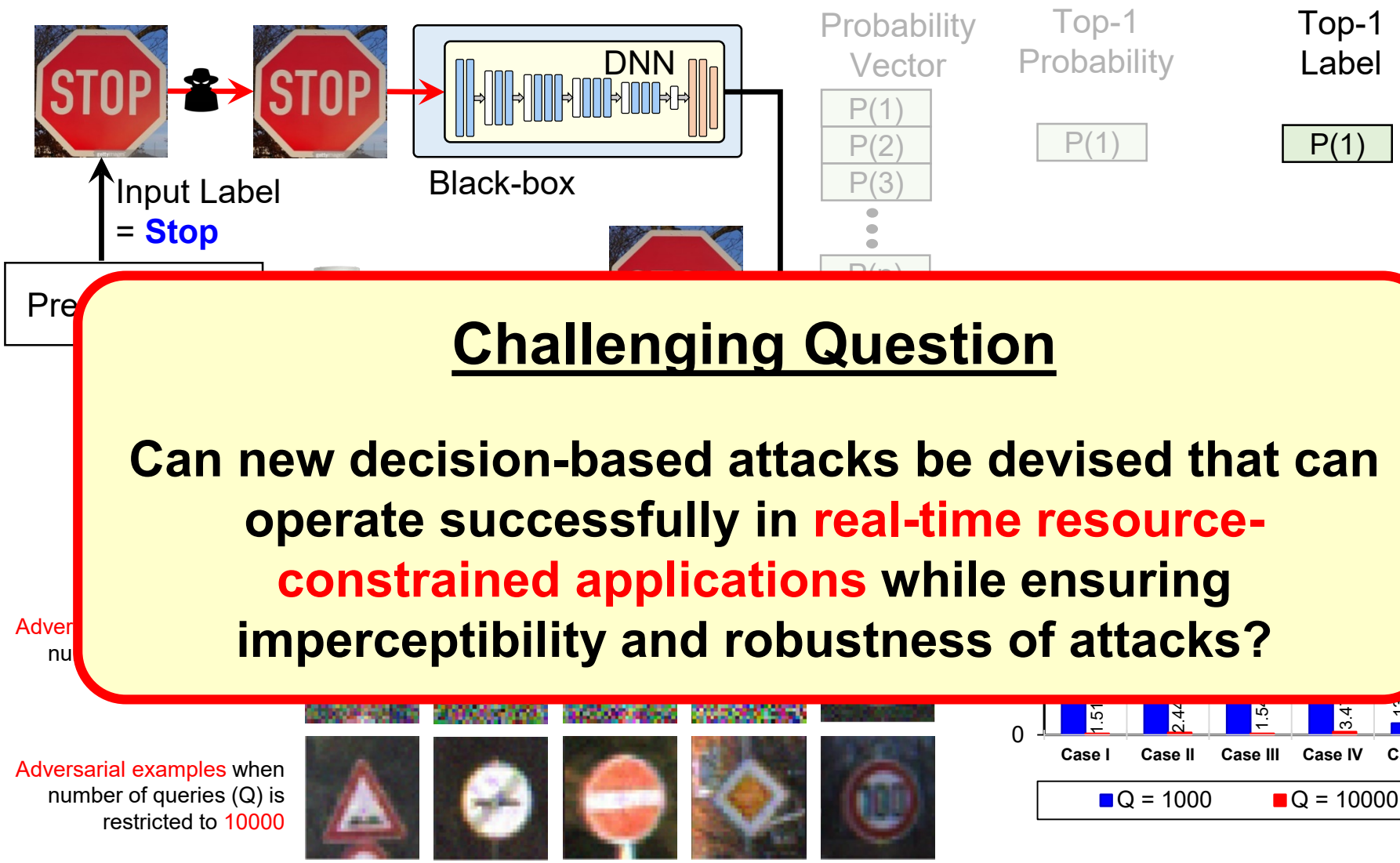
Decision–based Attacks





The imperceptibility of the adversarial noise increases with the increase in number of the allowed queries.

Decision-based Adversarial Attacks



Challenging Question

Can new decision-based attacks be devised that can operate successfully in real-time resource-constrained applications while ensuring imperceptibility and robustness of attacks?

The imperceptibility of the adversarial noise increases with the increase in number of the allowed queries.

Decision-based Adversarial Attacks

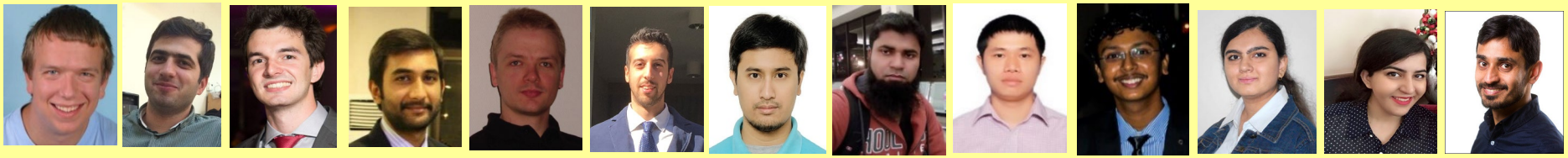
F. Khalid, M. Shafique et al., "**RED-Attack**: Resource Efficient Decision based Attack for Machine Learning",
[Link: https://arxiv.org/abs/1901.10258]

Adversarial examples when number of queries (Q) is restricted to 10000

The imperceptibility of the adversarial noise increases with the increase in number of the allowed queries.

My Research Team and Collaborators

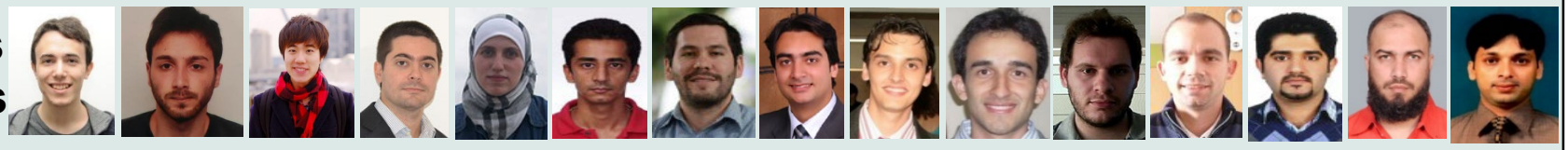
Post-Docs and PhDs



MS/BS Students

Collaborators

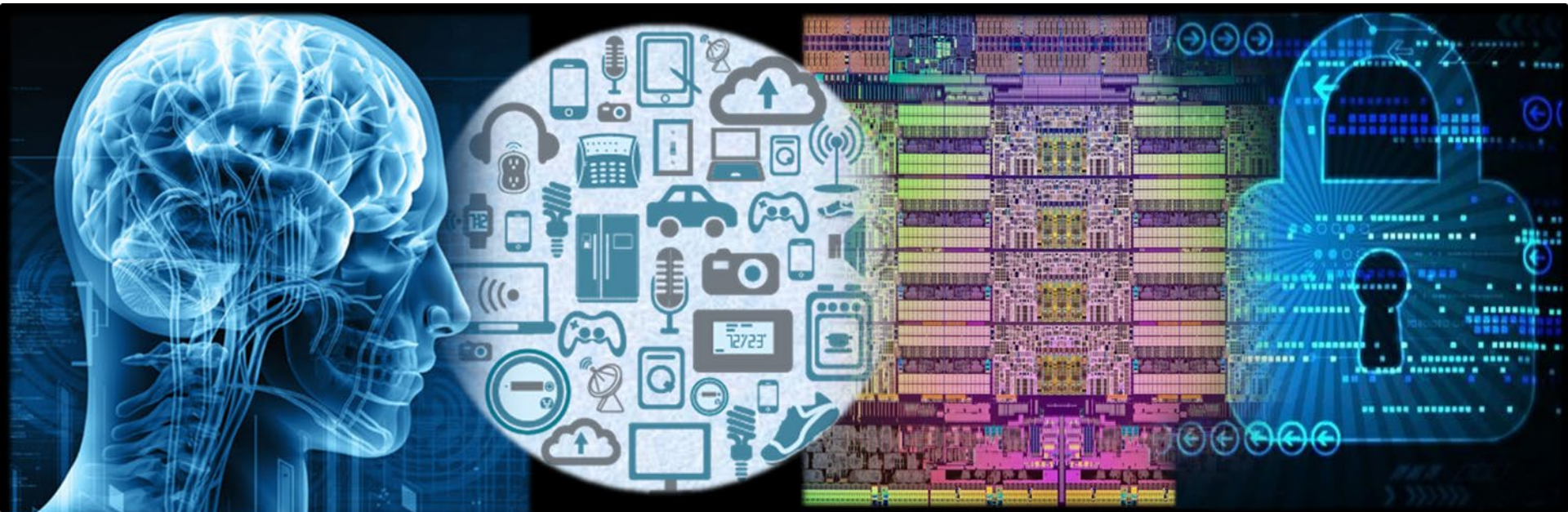
Previous Students



TECHNISCHE UNIVERSITÄT WIEN

CARE Tech Computer Architecture and Robust Energy-Efficient Technologies Group

Thank you!
Questions?

M. Shafique

swahlah1979@gmail.com