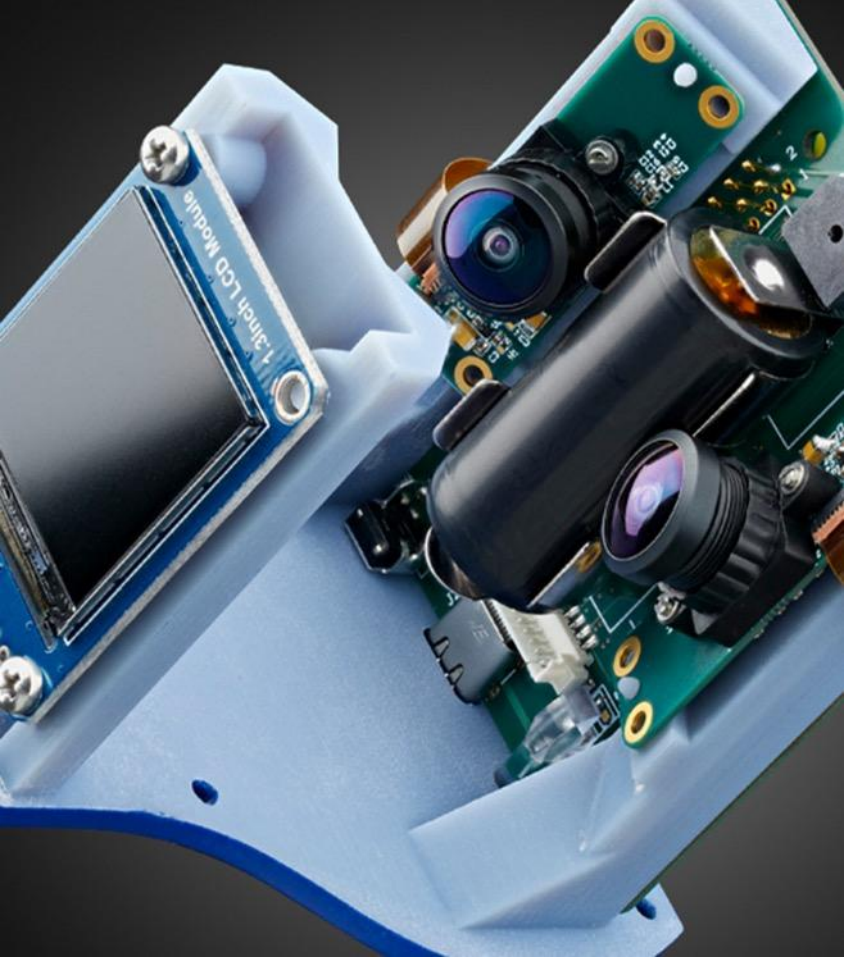
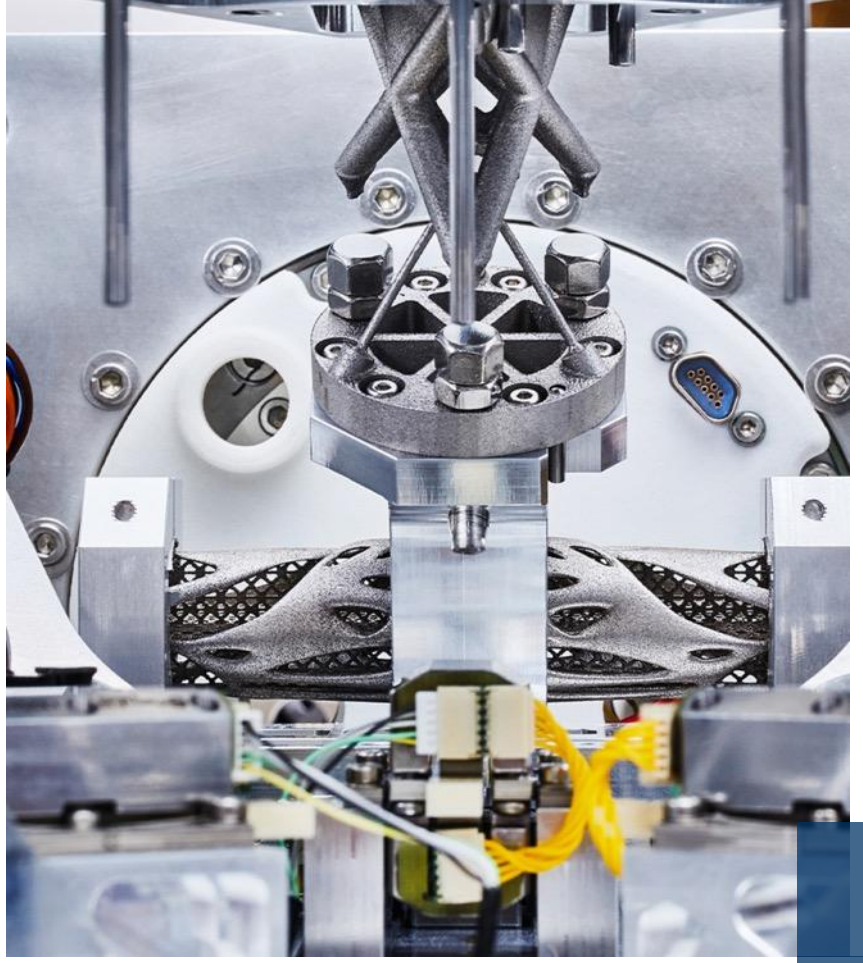
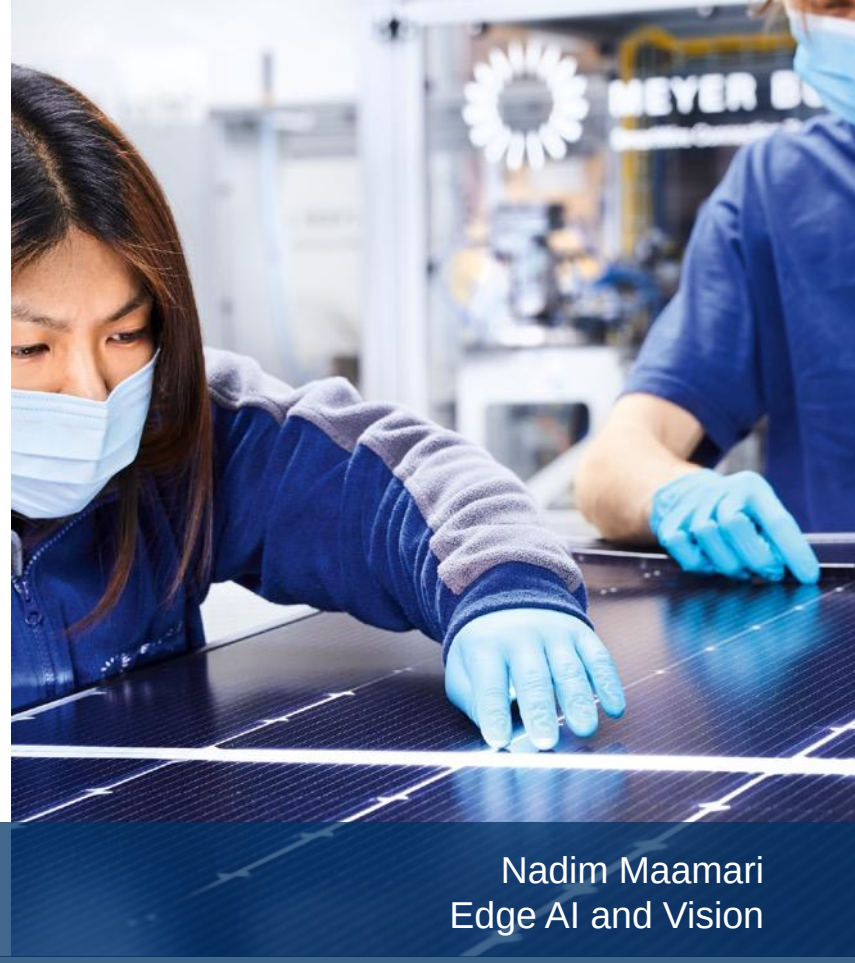




Nadim Maamari
Edge AI and Vision

08.07.2025

# DEFINING THE BOUNDARIES OF AI FROM FUNDAMENTAL LIMITATIONS TO RESOURCE CONSTRAINTS

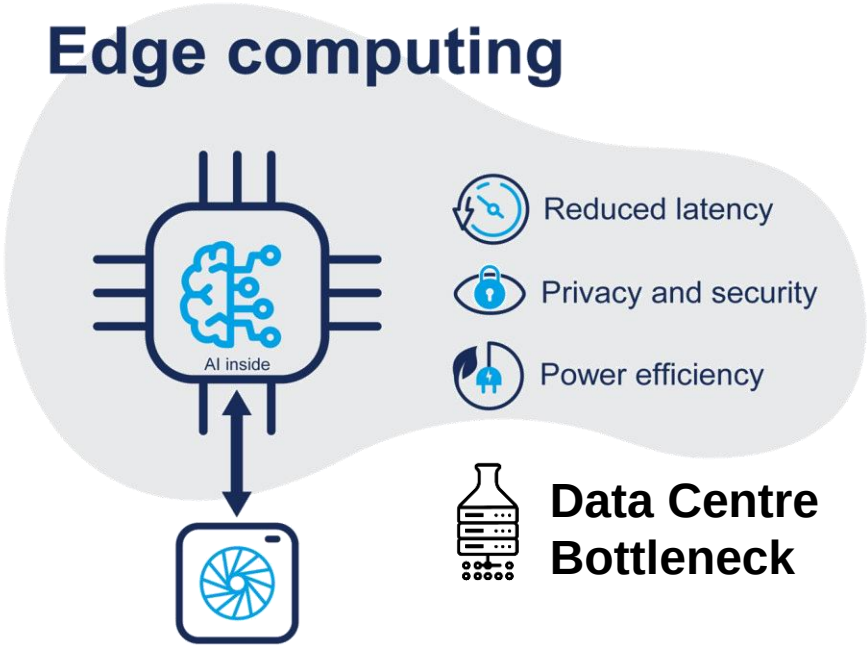
csem

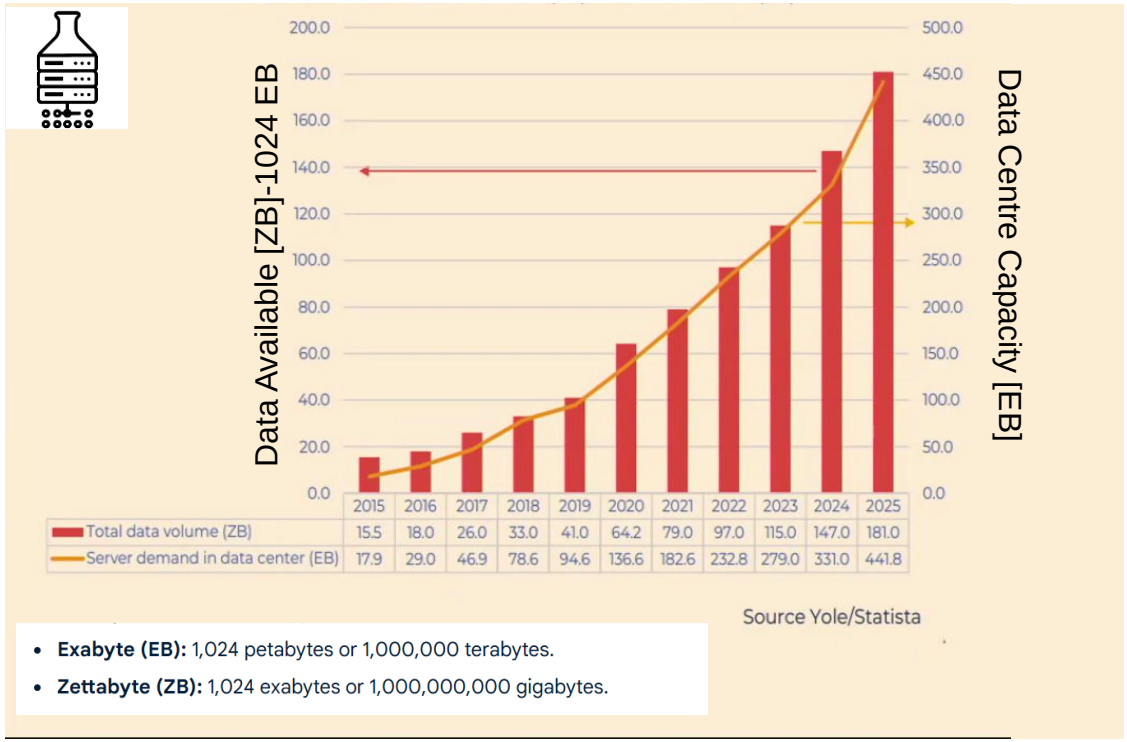
# EDGE AI

- **Enabler** – IoT Device And Data Center
- **Limitation –** Scaling law and Memory Wall
  - How to Address the limitation

# ENABLERS : DATA VS DATA CENTRES

- The Data available from IoT Device is 4 order of magnitude high than available data centre.
- Edge Computing is a Must to sustain growth.





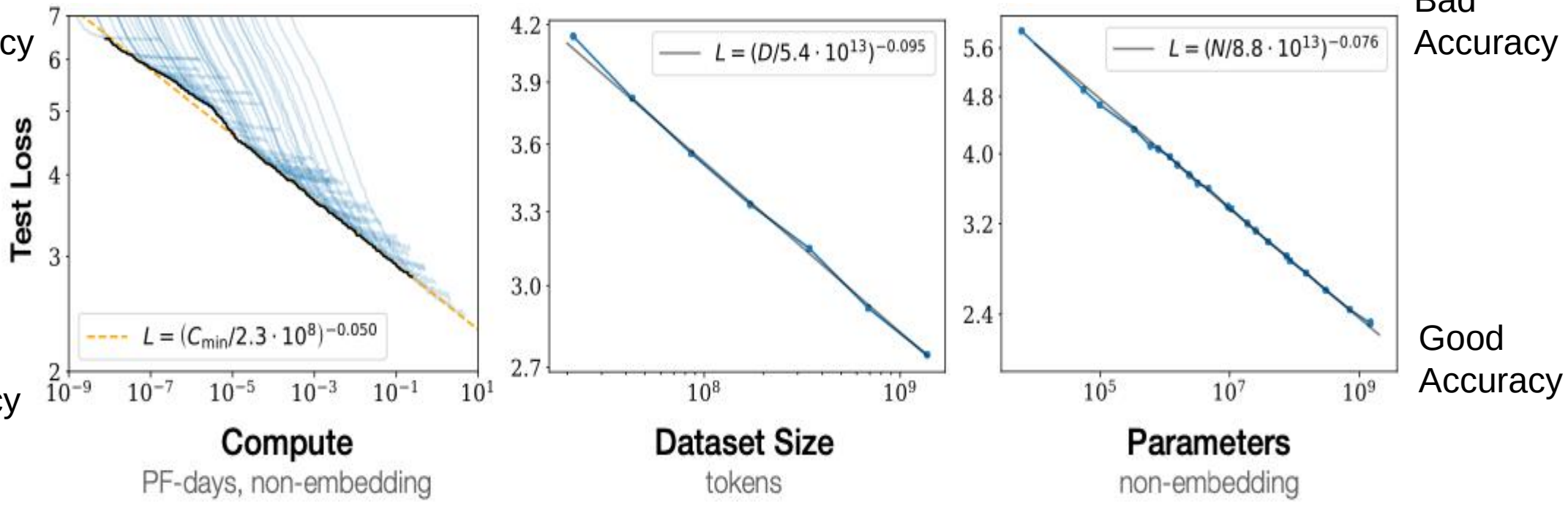

csem

# LIMITATION – BIGGER IS BETTER

- **N**eural **S**caling **L**aw

Bad Accuracy

Good Accuracy




Bad Accuracy

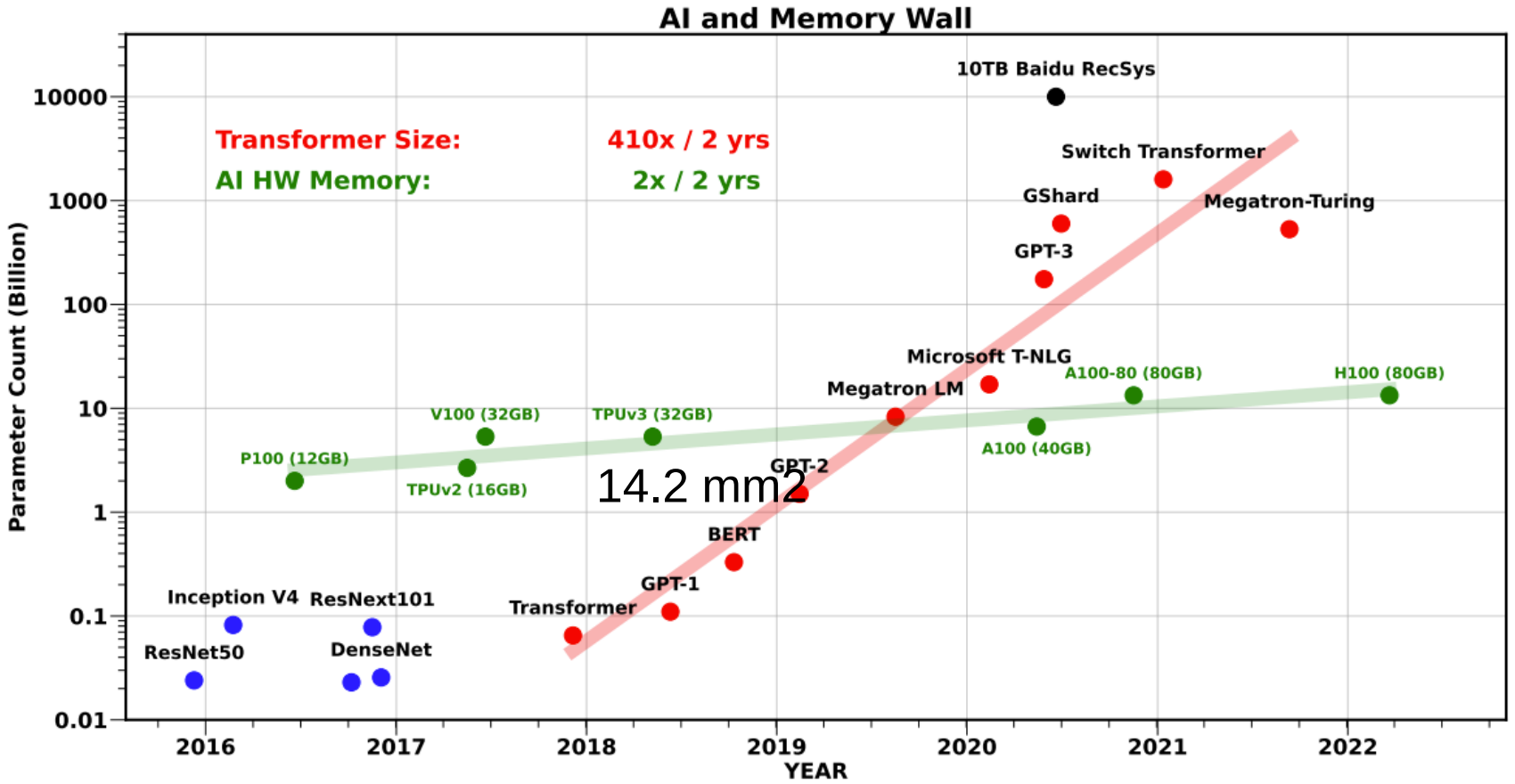
Good Accuracy

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *ArXiv Preprint* abs/2001.08361. https://arxiv.org/abs/2001.08361.

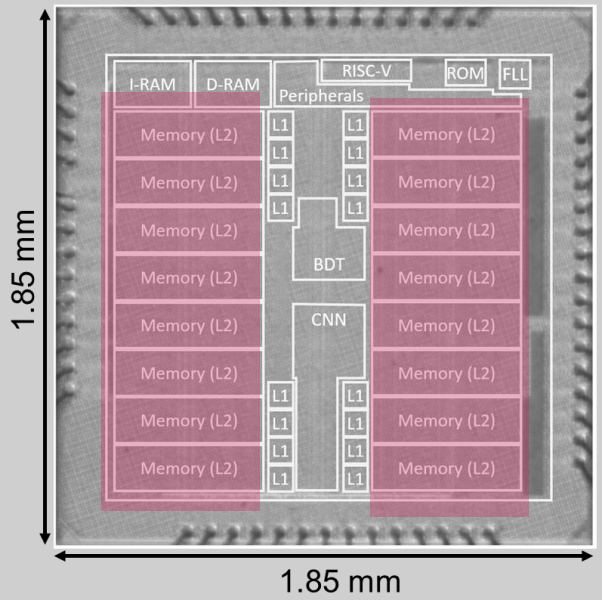


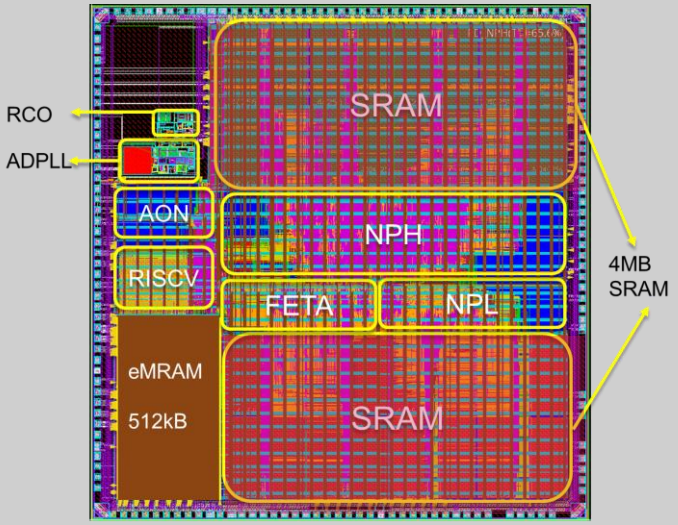
csem

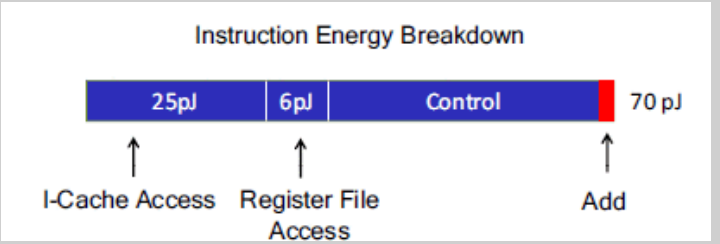
# MEMORY WALL




**AI and Memory Wall**
Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, Kurt Keutzer
https://arxiv.org/abs/2403.14123



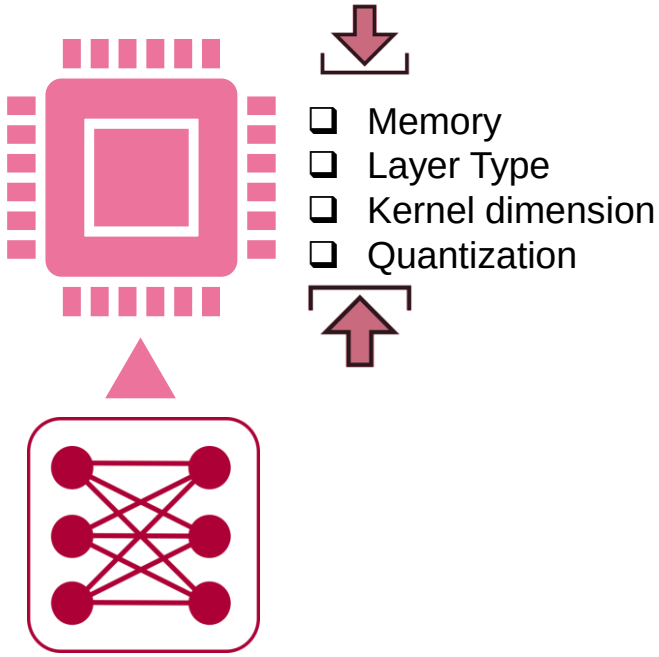



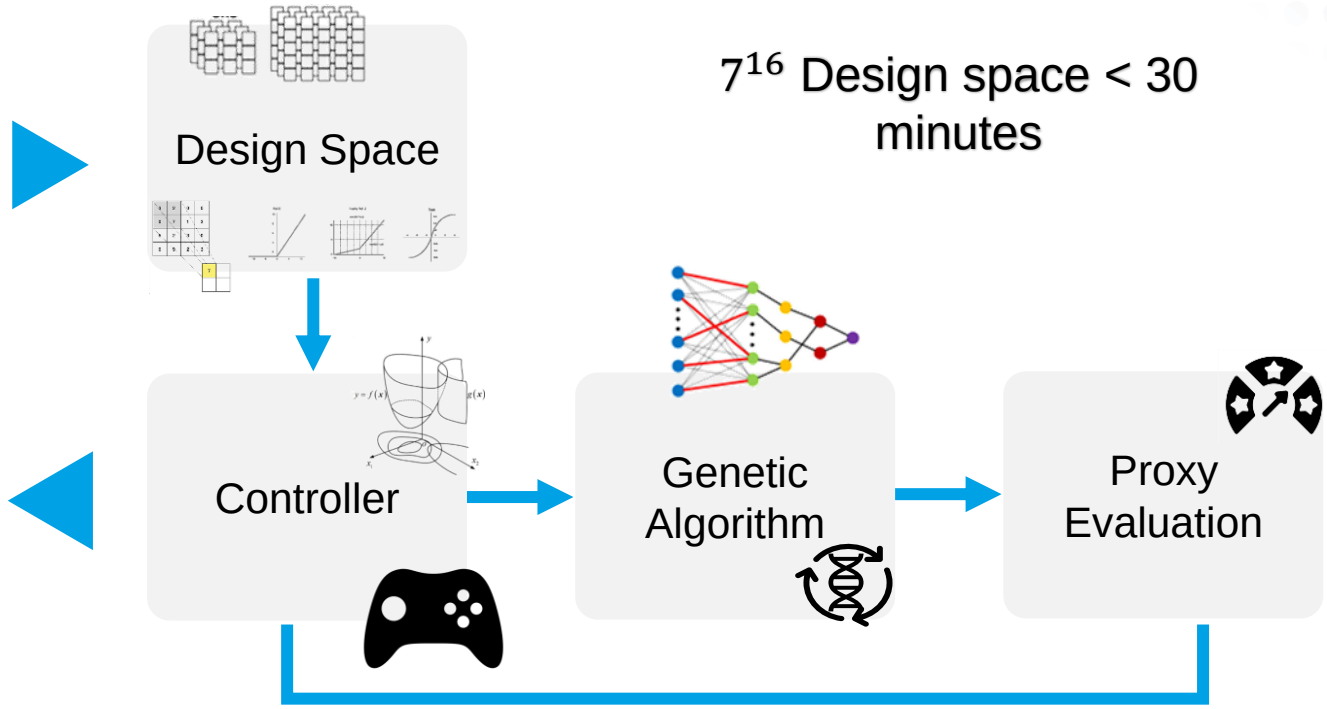
# HOW TO OVERCOME THIS LIMITATION

# HARDWARE AWARE AI

- Beyond Brut Force implementation



- ❑ Memory
- ❑ Layer Type
- ❑ Kernel dimension
- ❑ Quantization




$7^{16}$ Design space < 30 minutes

| Model | Latency (CPU) | Latency (TPU) | Accuracy (float32/int8) | Size (float32/int8) |
|---|---|---|---|---|
| MobileNetV2 | 16.3 ms | 2.86 ms | 69.8 / 69.8 % | 9.42 / 2.82 MB |
| ZiCo (2M limit) | 88.8 ms | 4.39 ms | 69.3 / 69.0 % | 7.14 / 2.13 MB |
| ZiCo-BC (2M limit) | 47.5 ms | 4.04 ms | 69.1 / 68.8 % | 7.39 / 2.21 MB |
| ZiCo (1.5M limit) | 49.8 ms | 4.92 ms | 68.5 / 68.3 % | 5.70 / 1.70 MB |
| ZiCo-BC (1.5M limit) | 35.1 ms | 4.29 ms | 69.5 / 69.5 % | 5.71 / 1.71 MB |

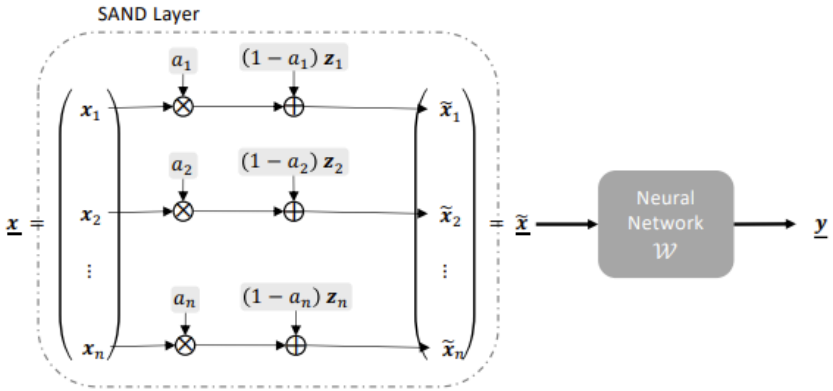
Constrained Zero-Shot Neural Architecture Search on Small Classification Dataset
Remy Vuagniaux et al .
https://ieeexplore.ieee.org/iel8/10675726/10675735/10675864.pdf

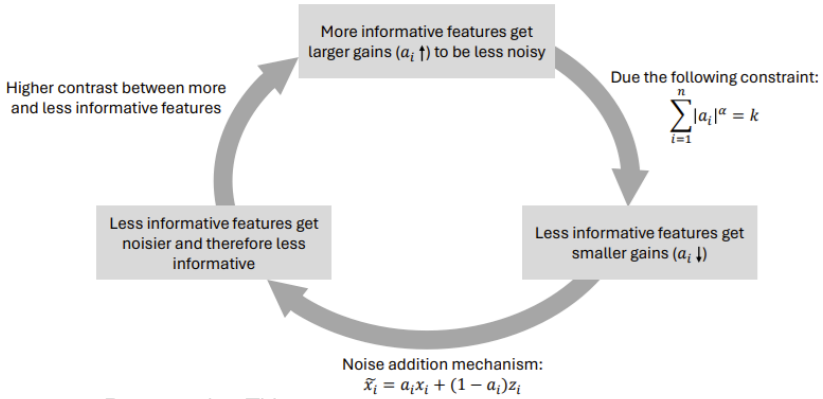


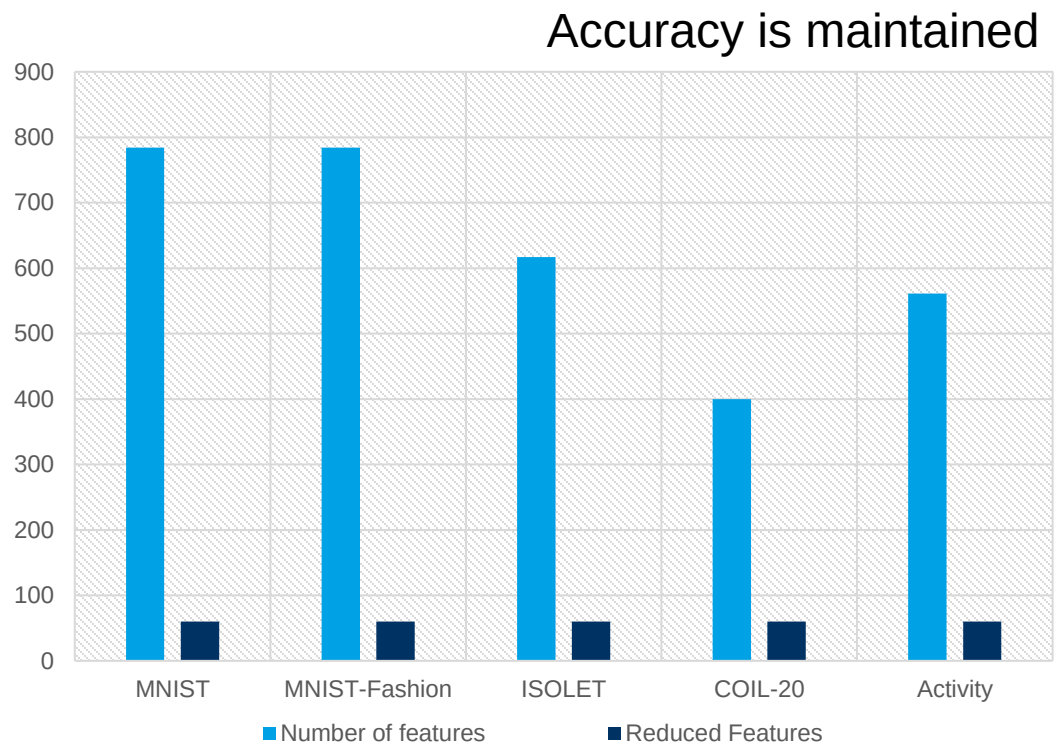
csem

# FEATURE SELECTION $K$

- **One-Shot** Feature **S**election with **A**dditive **N**oise **D**istortion
- No need for model retraining.



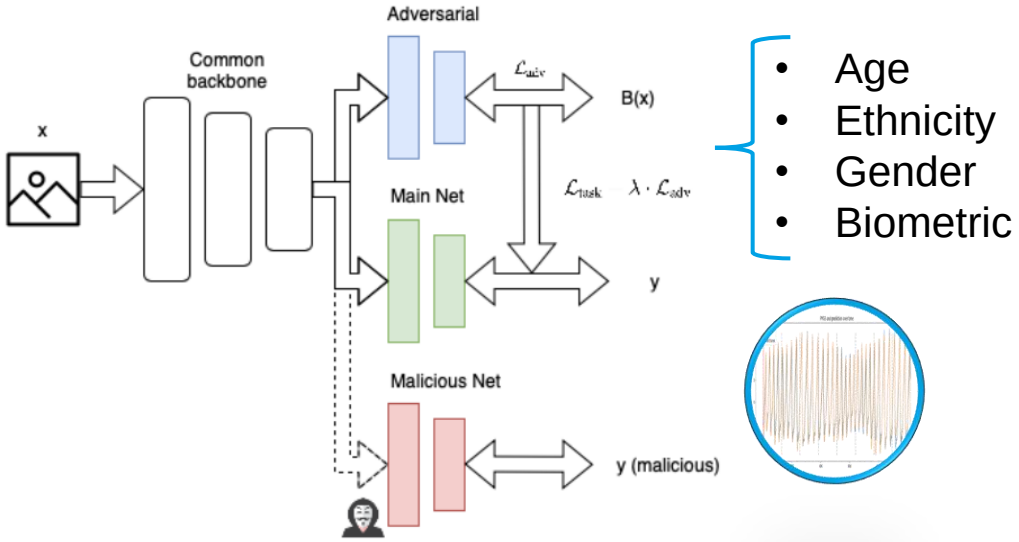

**SAND: One-Shot Feature Selection with Additive Noise Distortion** (ICML)
Download PDF

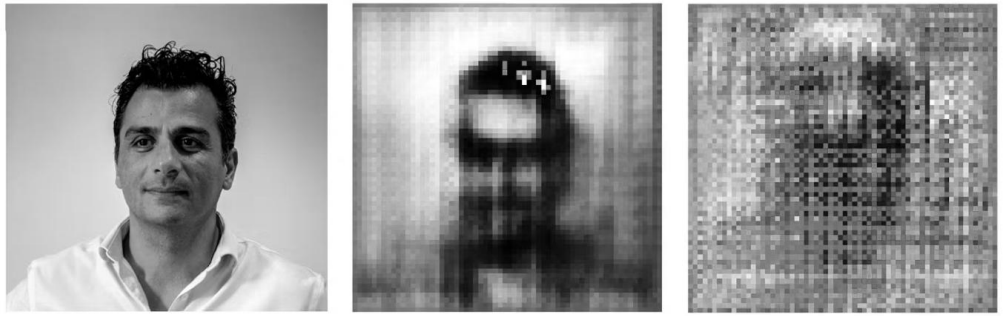


csem

# ADVERSARIAL TRAINING : VITAL SIGNS

- AI-driven solution for contactless vital sign monitoring (rPPG)



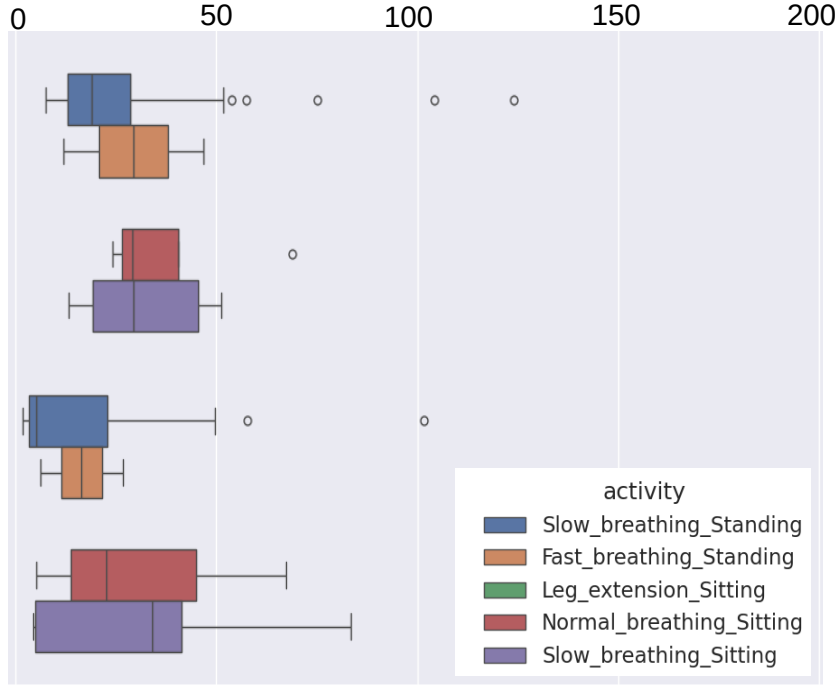



Vuagniaux, R.; Dia, M.; Saeedi, S.; Narduzzi, S.; Maamari, N. Unified Adversarial Training for Bias Mitigation and Privacy Preservation.

Mean absolute error [$ms$]

Remote Measurement

Contact base measurement






csem

# ENABLING AI SOLUTION ON EDGE

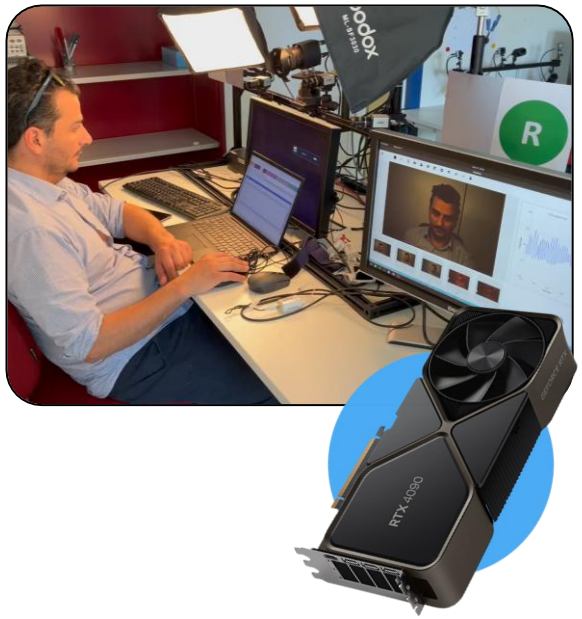
| Pure software | Software + dedicated Sensors | Software on COTS | Dedicated ASIC |
|---|---|---|---|

CLOUD / DATA CENTER

MW

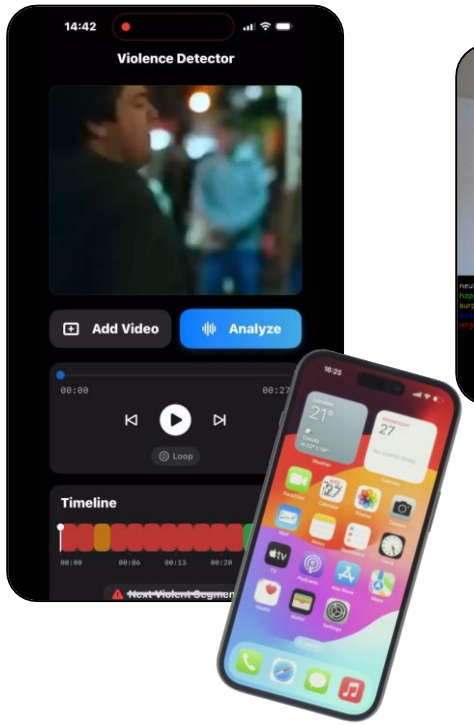
LOCAL COMPUTER

kW



EDGE & MOBILE DEVICES

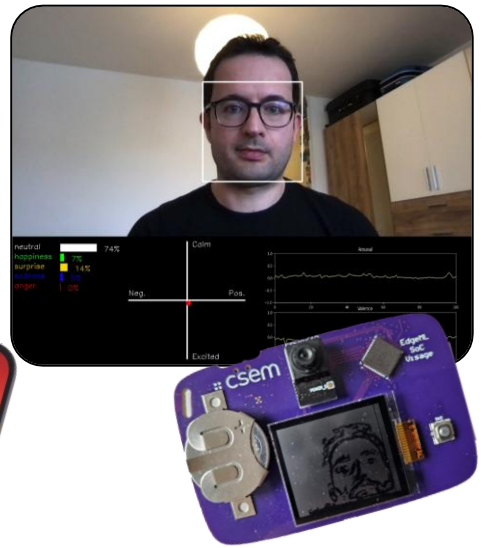
W





SMART CONNECTED SENSORS

mW






csem

# CONCLUSION

- Enablers -> Data Vs Data Centre Availability
- Limitation -> Neural Scaling law + Memory Wall.

The Way Around It :

1. Hardware-aware AI designs networks with balanced arithmetic intensity as constrain.
2. Feature selection method reducing the model by an order of magnitude.
3. Adversarial training enables privacy, removes bias and improves generalisation.