

~~A Shift Towards Edge Machine-Learning Processing~~

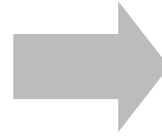
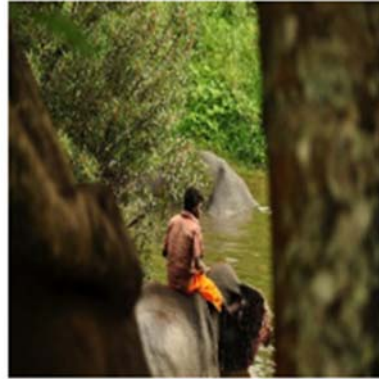
Olivier Temam
temam@google.com

Hardware Accelerators: Helping Deliver the AI Potential

Olivier Temam
temam@google.com

Style

The Starry Night,
Vincent van Gogh,
1889



Perceptual Losses for Real-Time Style Transfer and Super-Resolution (2016)



Celebrities (training set)

Generated Faces



Progressive Growing of GANs for Improved Quality, Stability, and Variation (2017)

Spectacular Progress In AI

It All (Re)Started With...

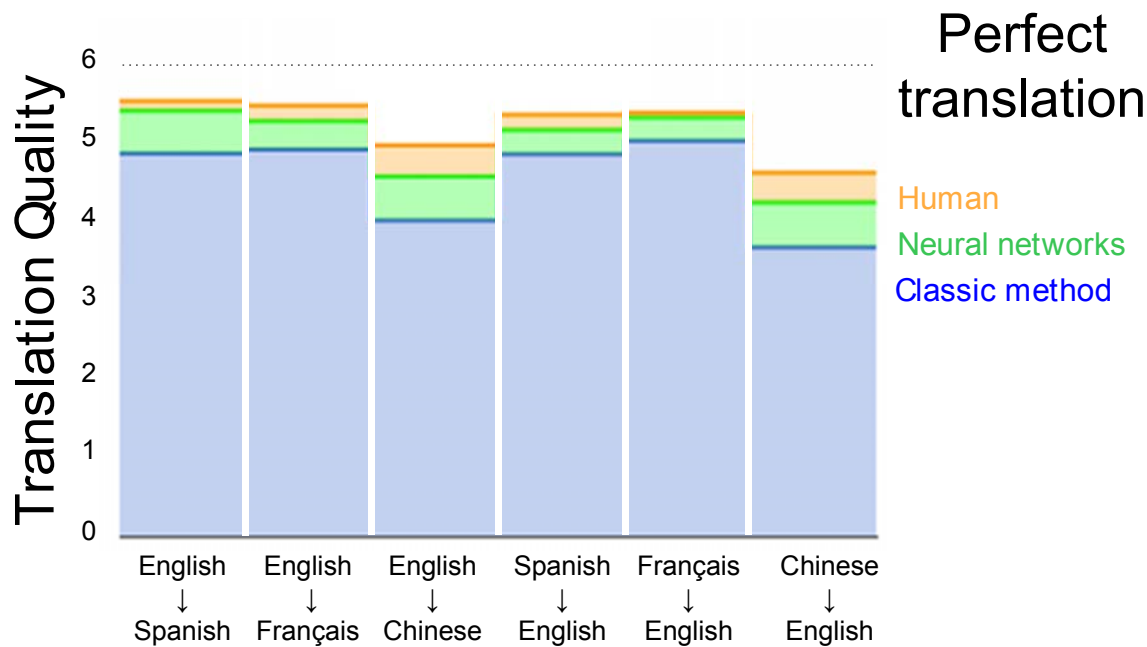


Large Scale Visual Recognition Challenge

Team	Year	Error
XRCE (Xerox) - pre-DNN	2011	25.8%
SuperVision (Univ. Toronto)	2012	16.4%
Clarifai	2013	11.2%
Inception (Google)	2014	6.7%
Andrej Karpathy (human-level)	2014	5.1%
MSRA (Microsoft)	2015	3.6%
WMW (Momenta/ Oxford Univ.)	2017	2.2%

Automatic Translation

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016)



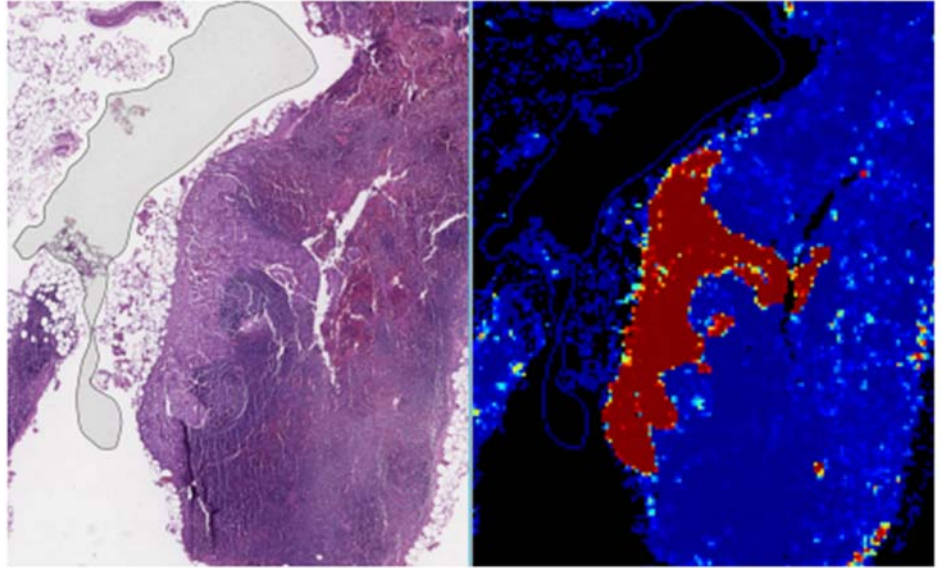
Detecting Cancer Metastases

Tumor localization score
(FROC):

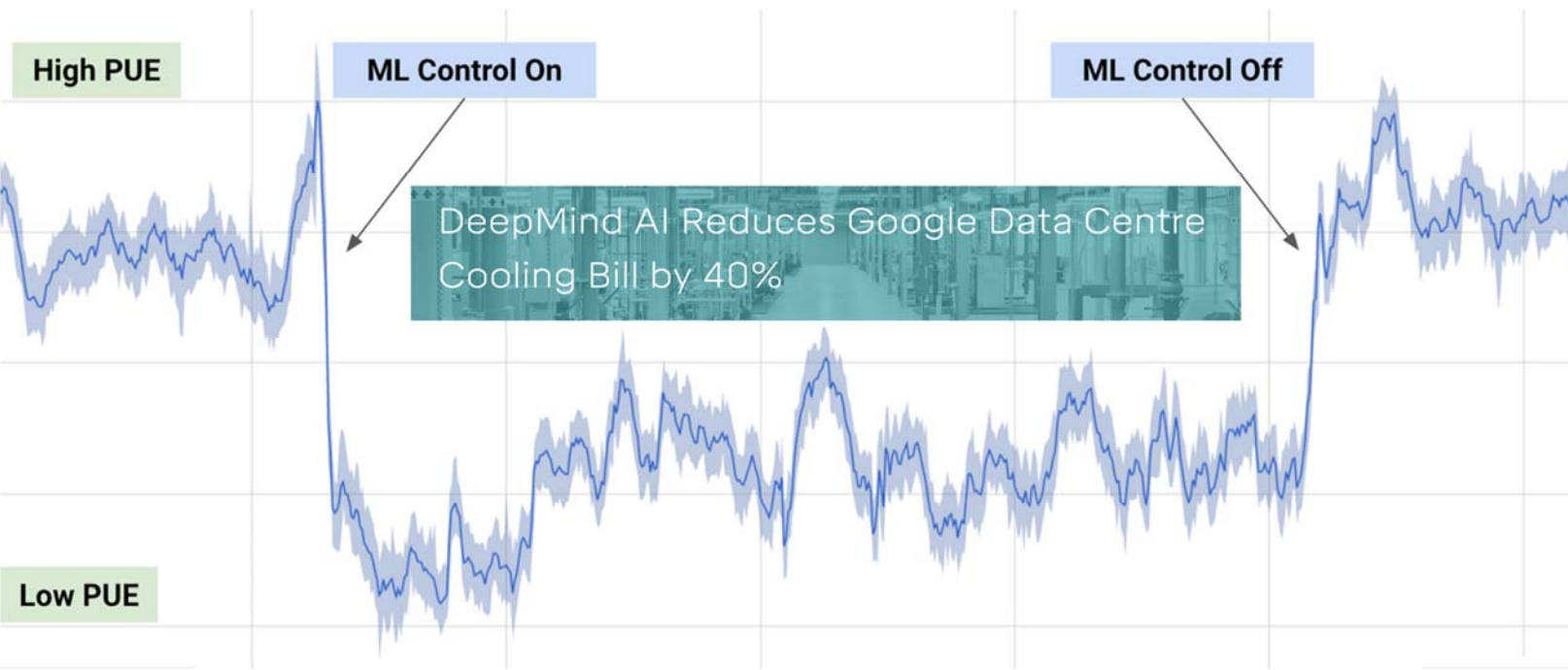
Pathologist: 0.73

AI model: **0.89**
(better)

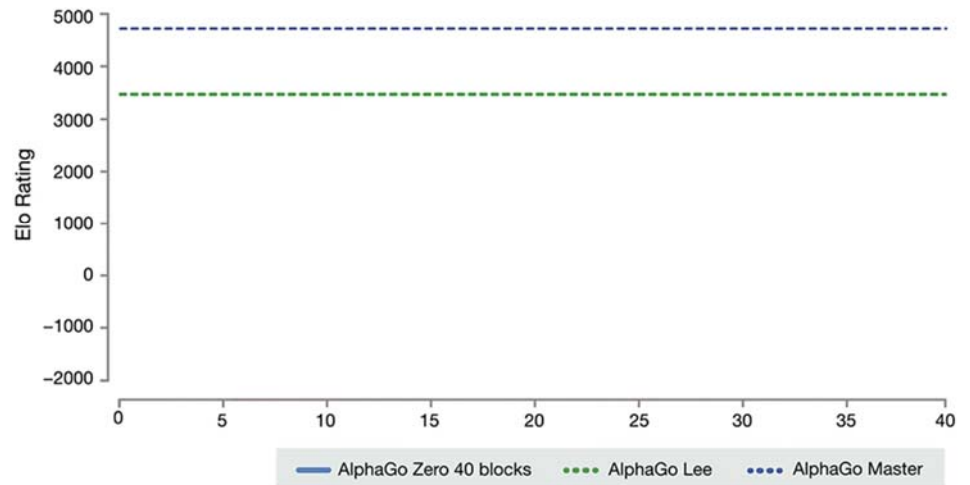
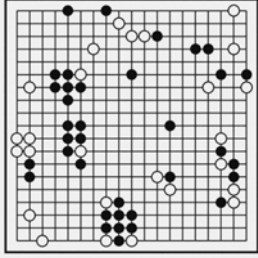
*Detecting Cancer
Metastases on Gigapixel
Pathology Images (2017)*



Google Data Center Energy



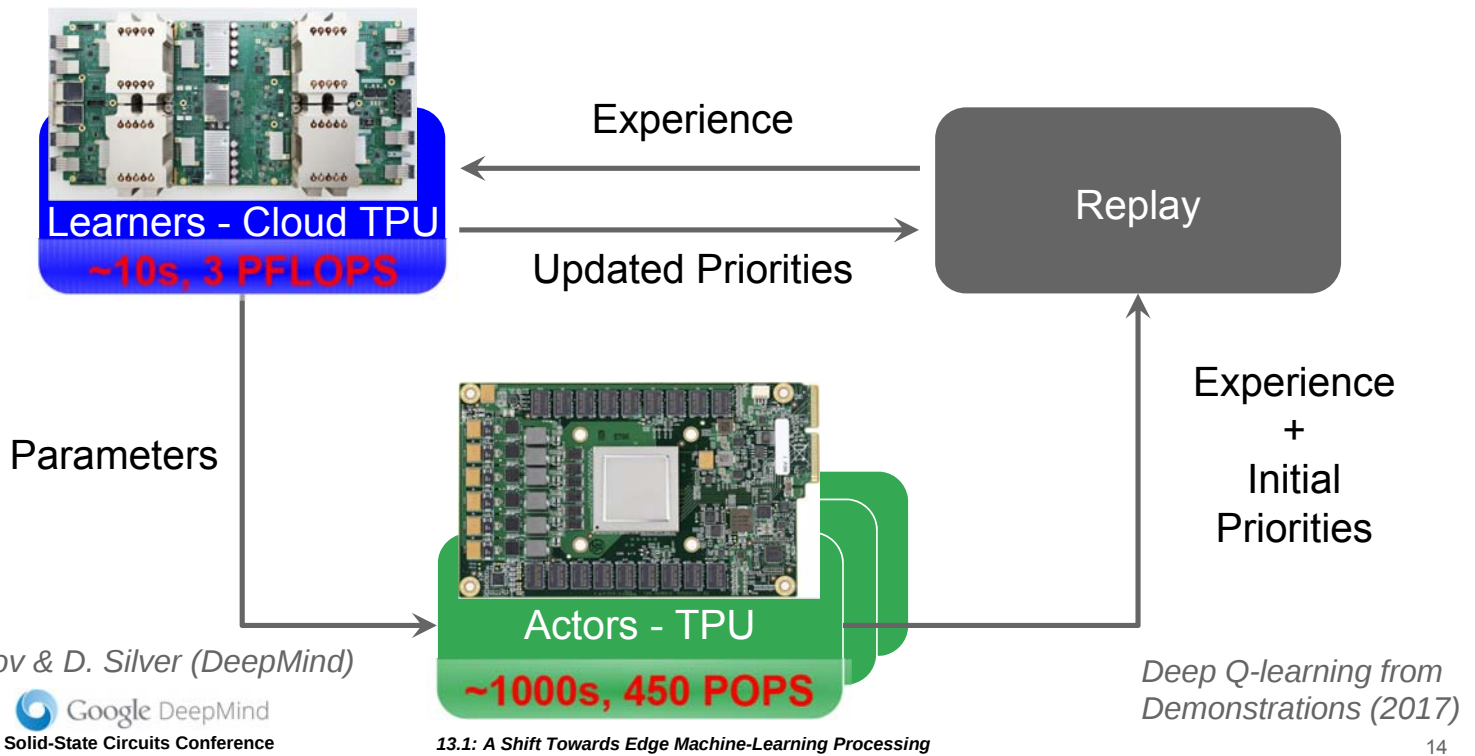
AlphaZero



AlphaZero only trained against itself.

Why ASICs for NNs ?

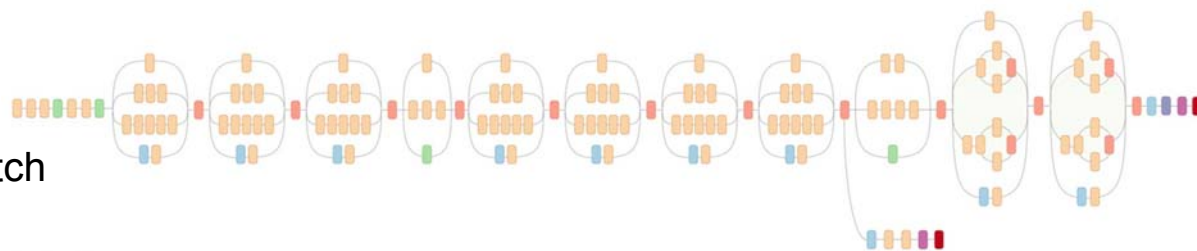
TPUs Used for AlphaZero



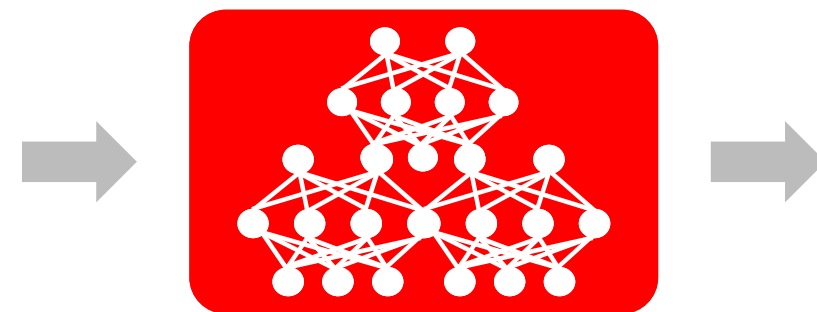
Economic Incentive

Inception v3:

- 299x299 input patch
- **11.4 Bops**



Photo

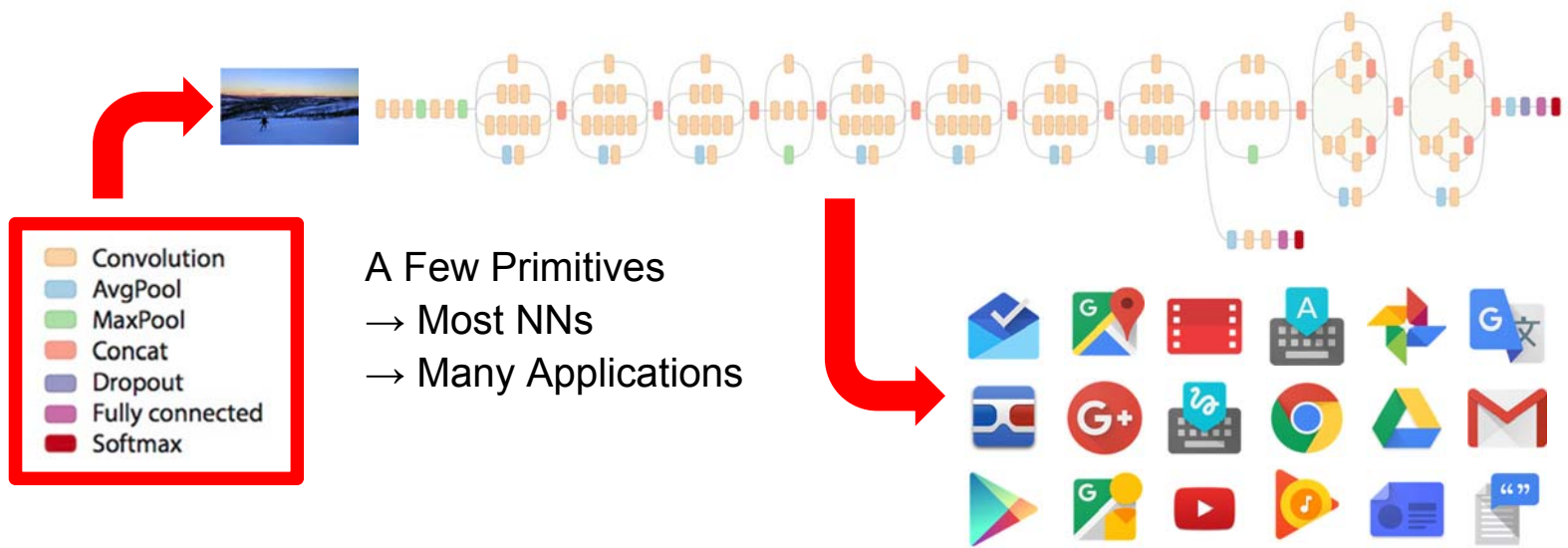


“ski”

Semantic

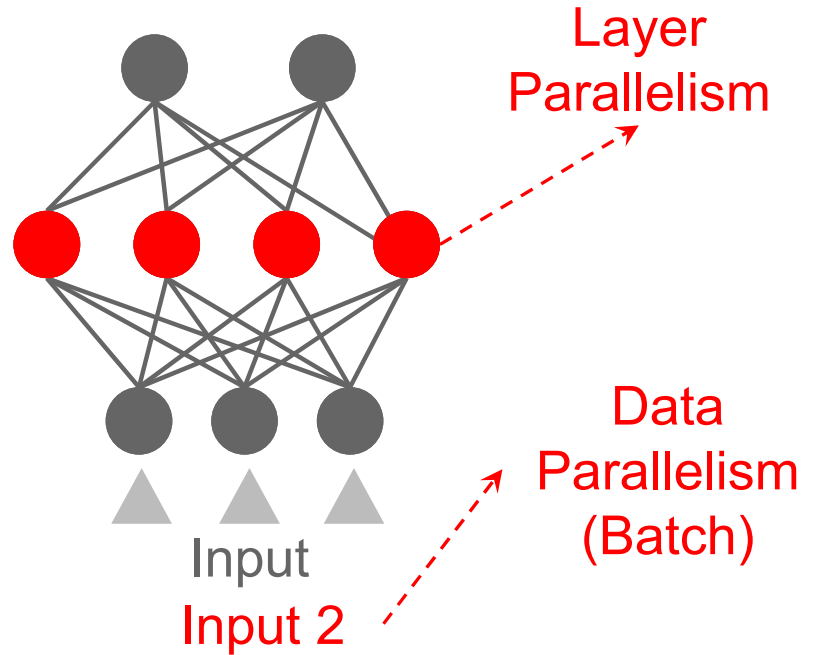
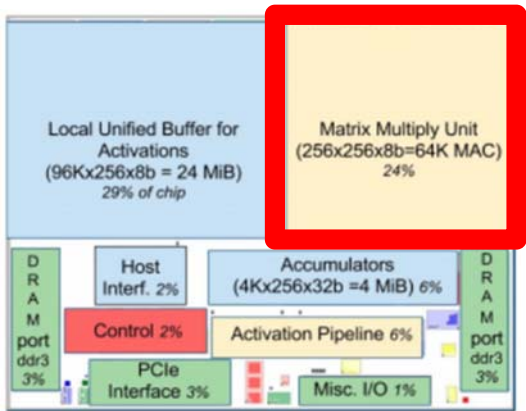
Google Photos Search

Technical Incentive



Reconcile **flexibility** and **efficiency**

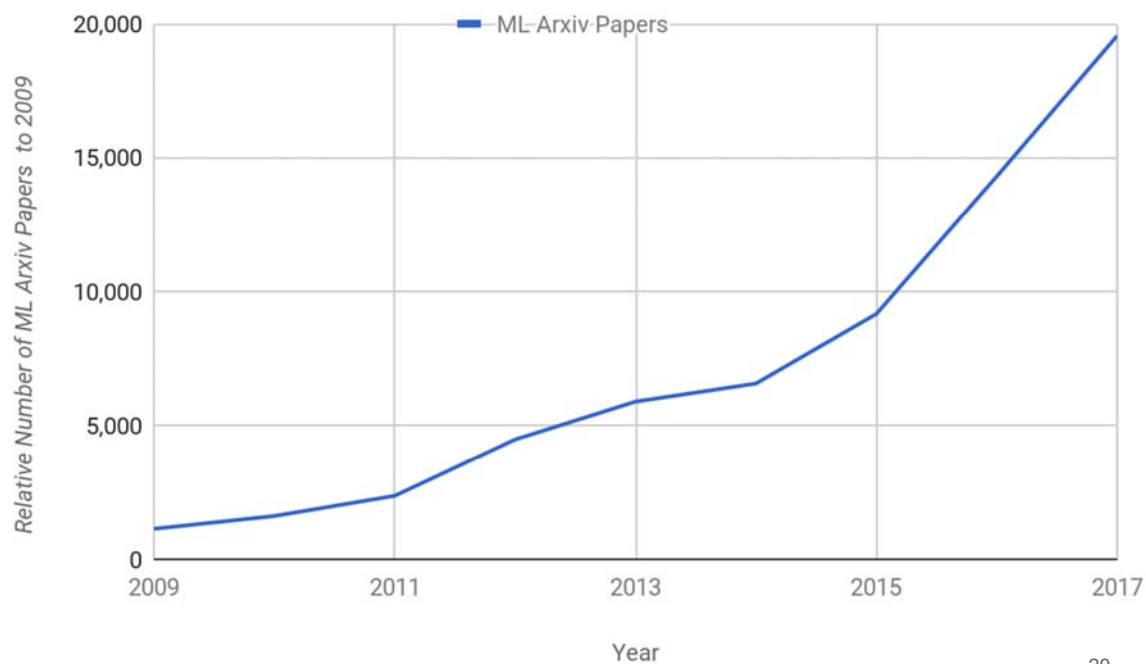
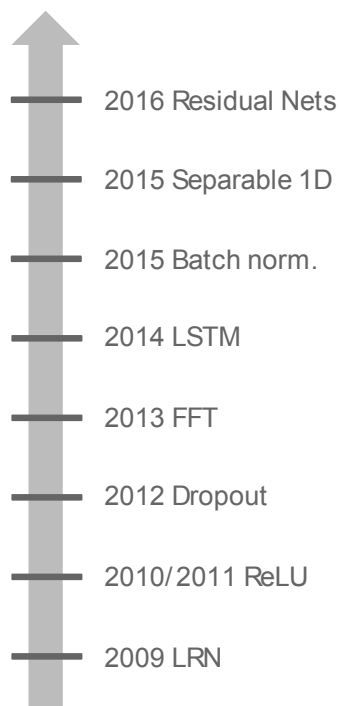
How ? Massive Computational Units



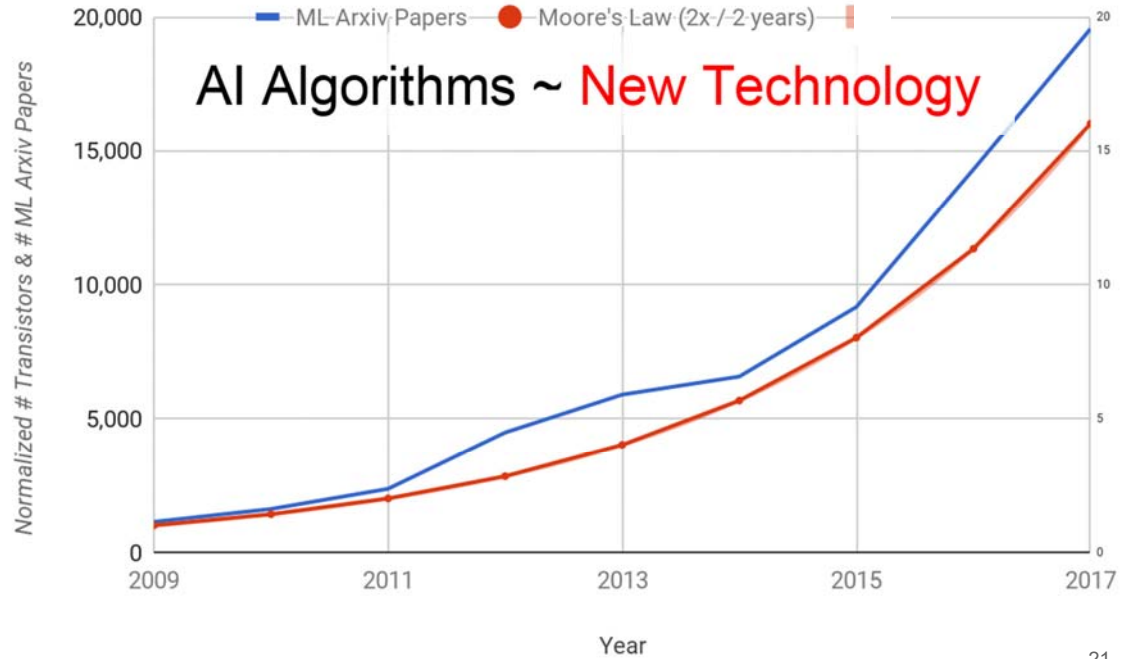
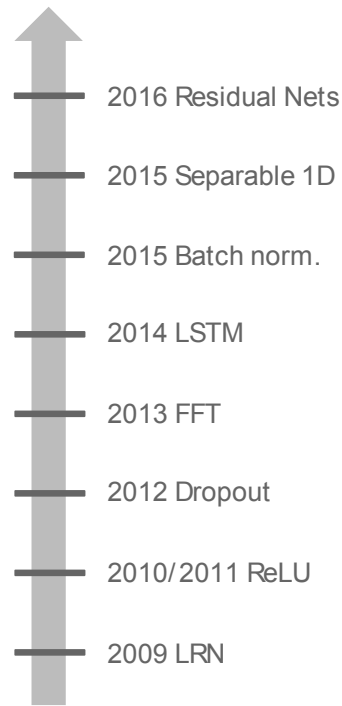
Most computations $\sim$ $M \times V$

Challenges & Opportunities

Software Challenge



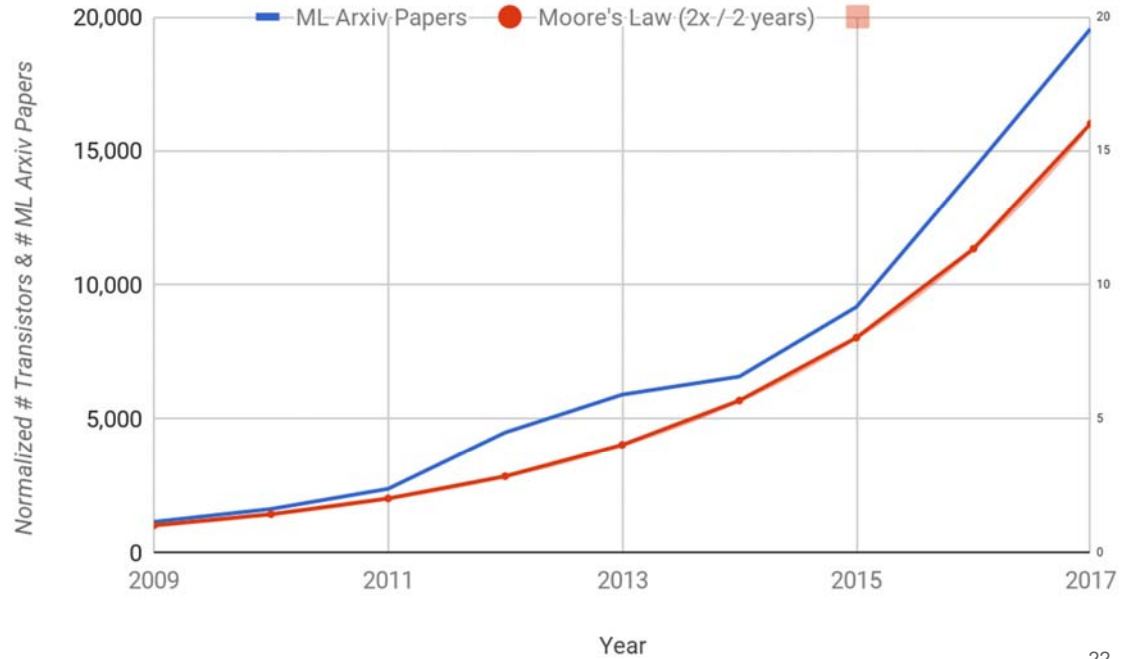
Rapid Evolution of AI Algorithms



Growing Diversity of AI Algorithms

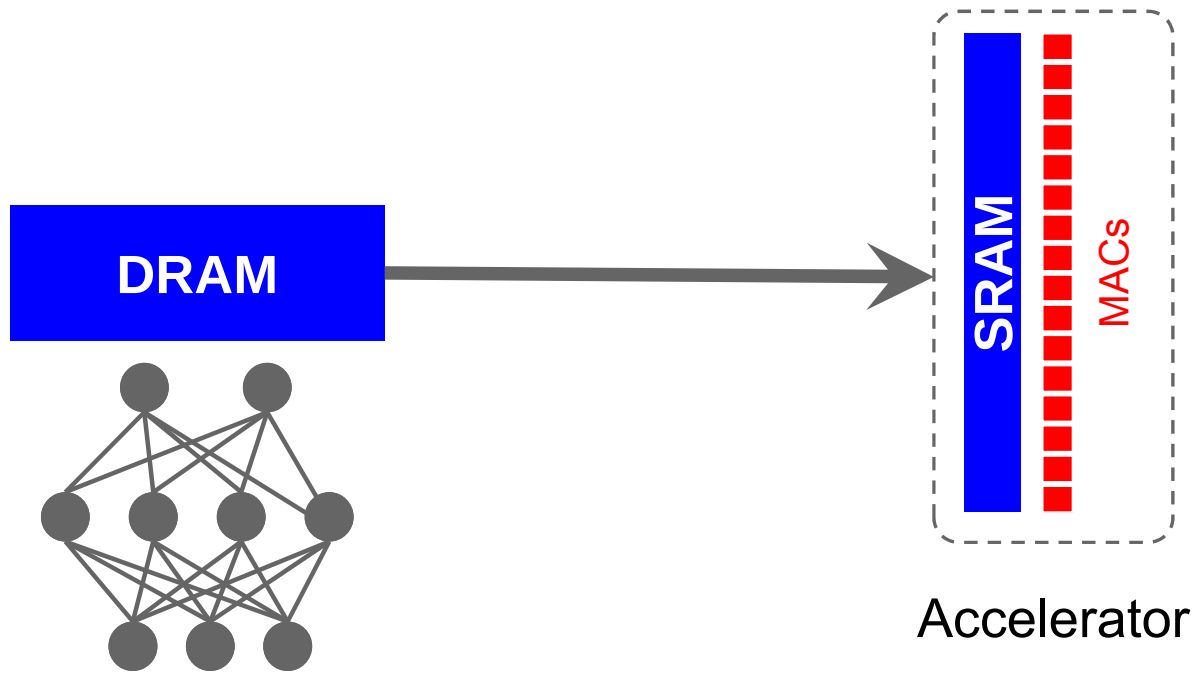
Batch norm.
Residual Nets
LRN
FFT
Separable 1D
ReLU
LSTM
Dropout

Tension between
HW customization
and ML (SW)
generality

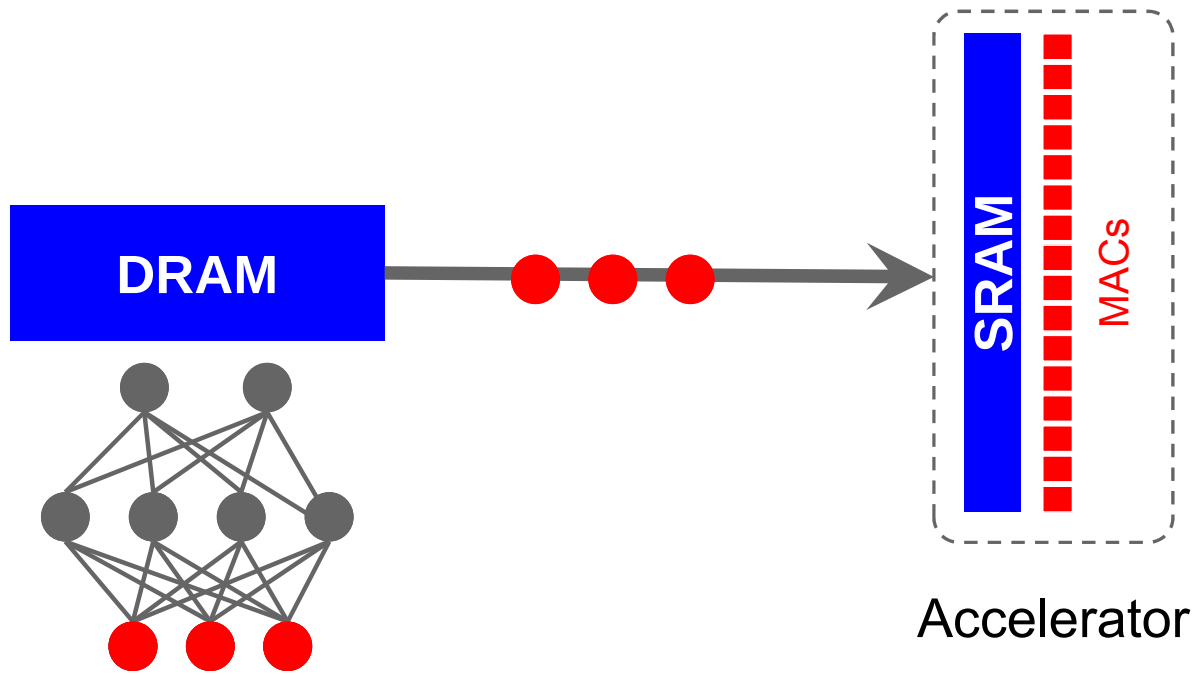


Memory Challenge

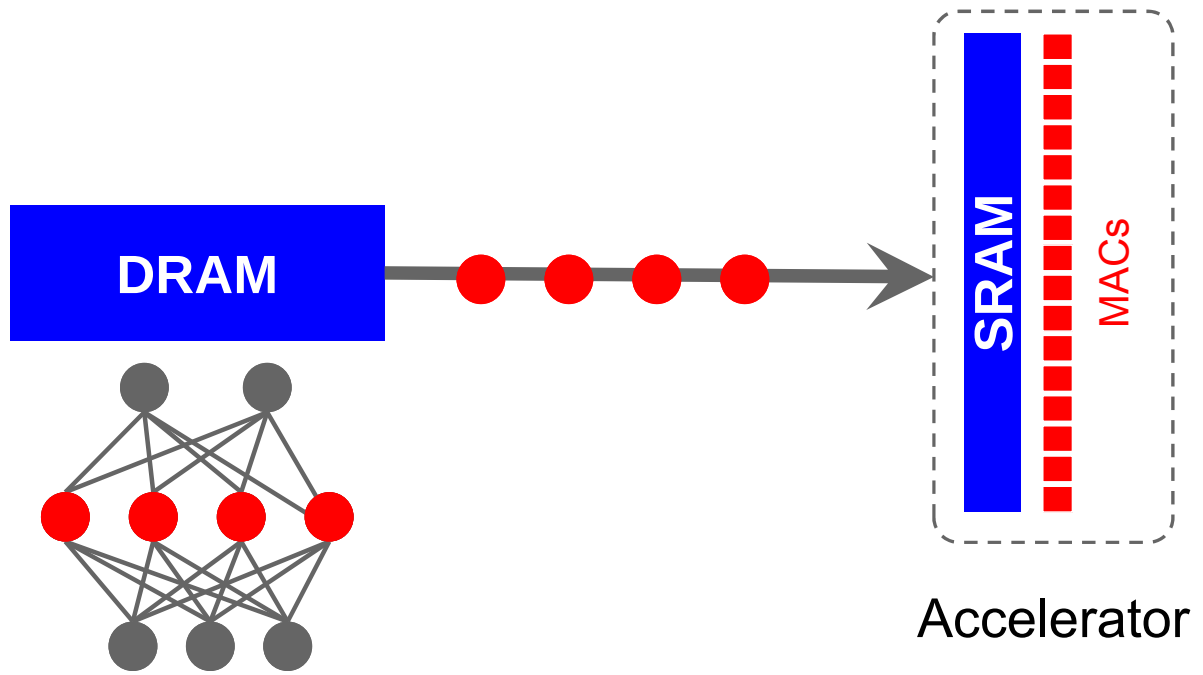
But... Still Von Neumann-Like Execution



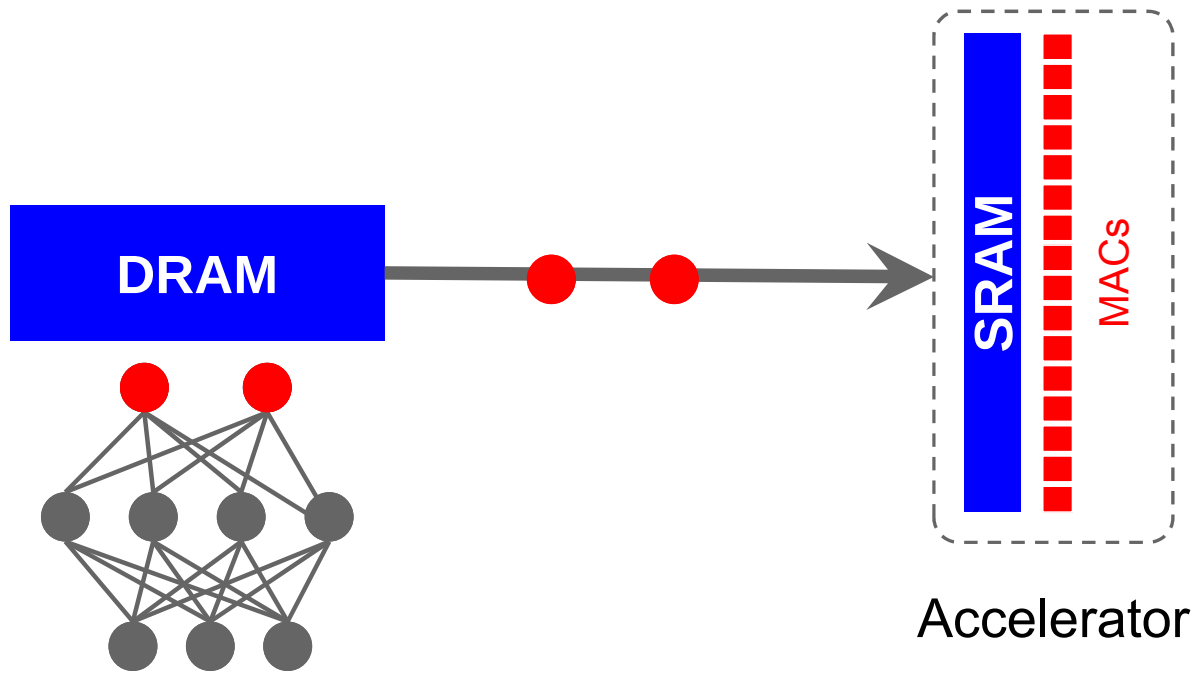
But... Still Von Neumann-Like Execution



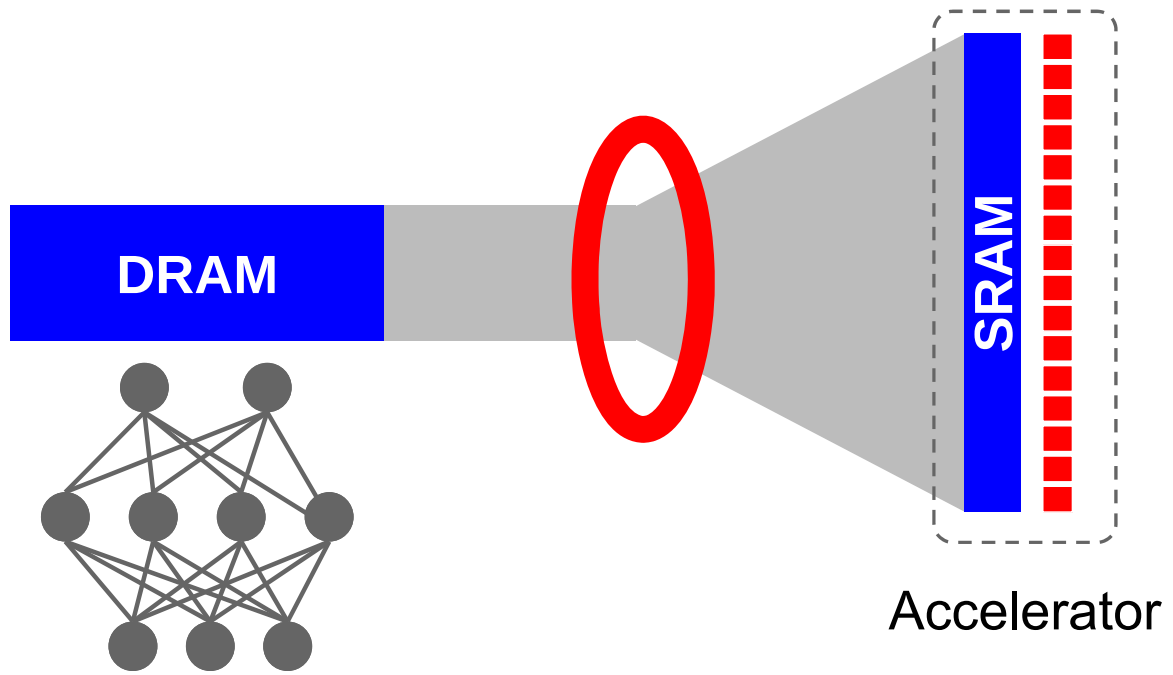
But... Still Von Neumann-Like Execution



But... Still Von Neumann-Like Execution



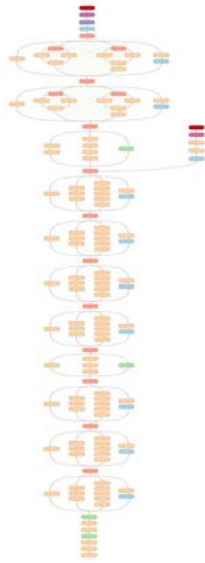
Biggest Limitation Is Memory Bandwidth



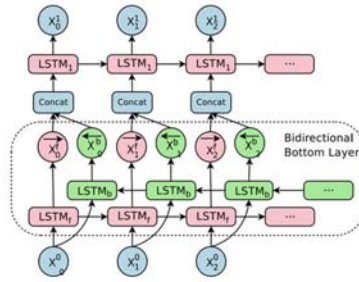
NNs Stored On-Chip Is Within Reach



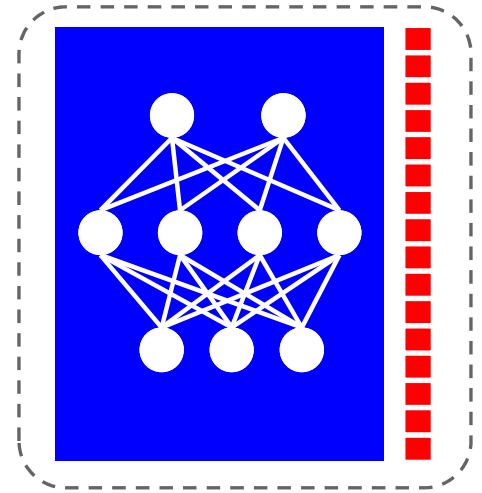
MobileSSD
O(1MB)



Inception
O(10MB)



NMT
O(100MB)



Accelerator

Need Near-Logic Memory Innovations



HBM

O(1TBps)



SRAM

O(10TBps)

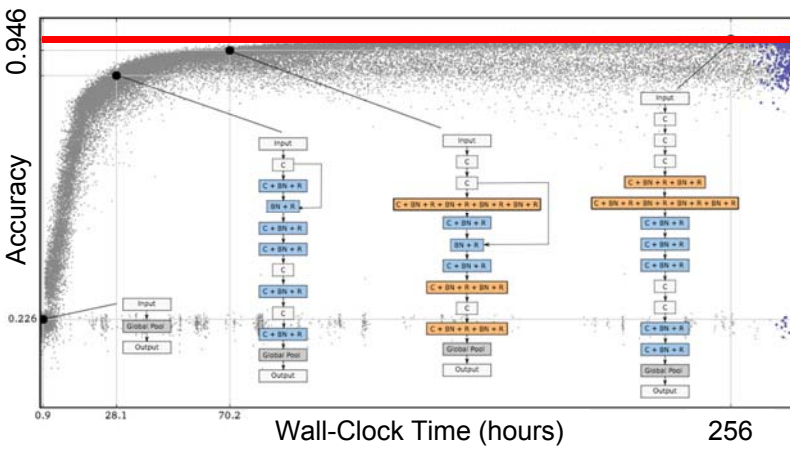


?

- O(100TBps)
- Bytes/mm² > DRAM
- Low leakage

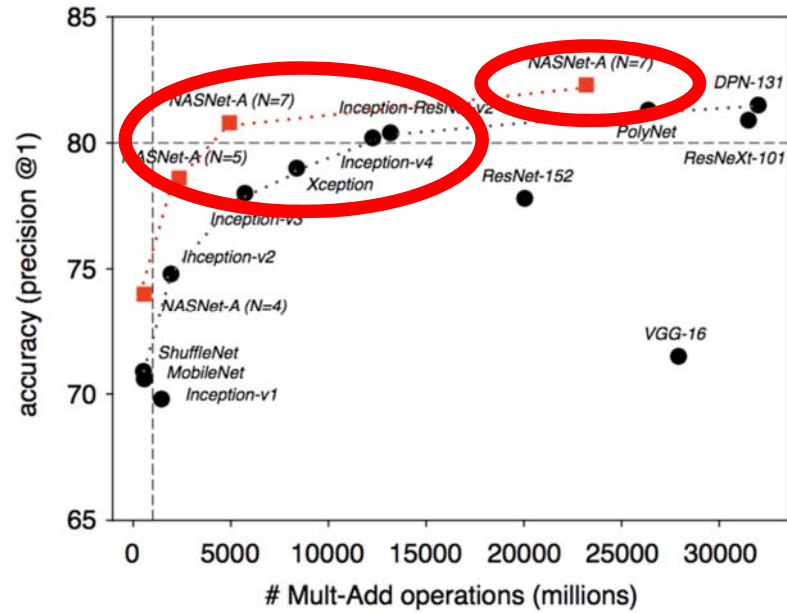
Efficiency Challenge & AutoML Opportunity

Towards Highly Efficient NNs

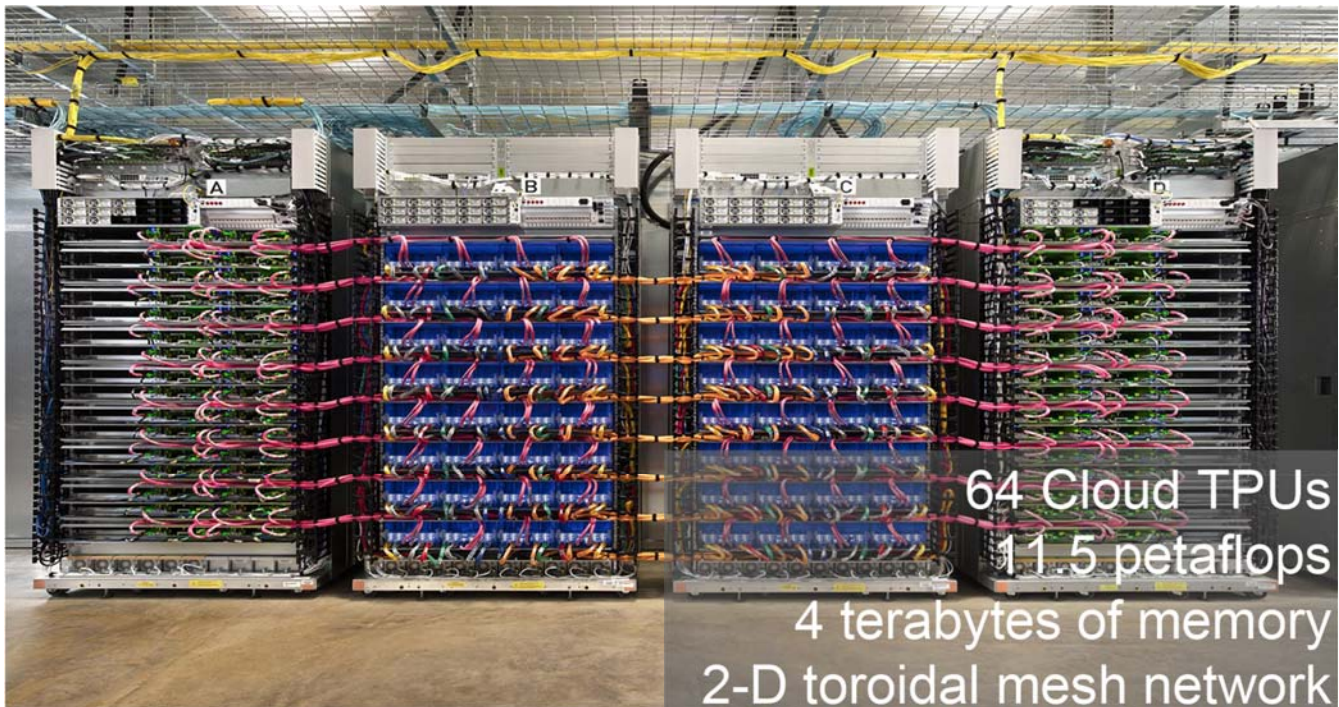


Large-Scale Evolution of Image Classifiers (2017)

Learning Transferable Architectures for Scalable Image Recognition (2017)



Towards Highly Efficient NNs



Cost Challenge

Make High-Quality AI Broadly Accessible



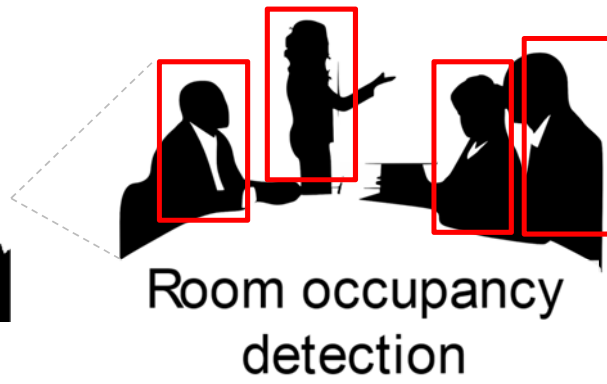
\$249 Google Clips



\$45 AIY Vision Kit

Edge Opportunity

Unlocking New Applications



Real-world
input



Raw data

Edge

Semantic
data



Cloud

The Many Benefits of Edge AI



Latency



Bandwidth



Privacy



Connection



Cost

Bandwidth



Upload bandwidth

5-8 Mbps

1492 Mbps
@30fps

5971 Mbps
@30fps

1080p
H.264

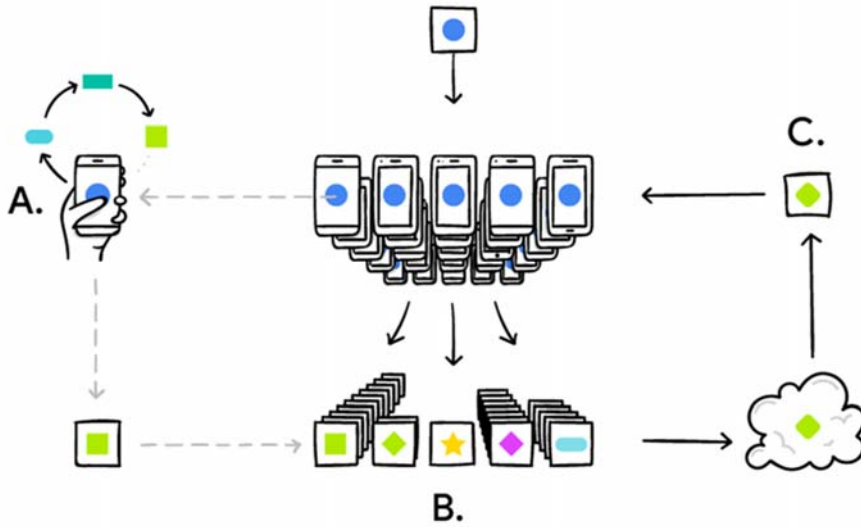
1080p
uncomp.

4K uncomp.

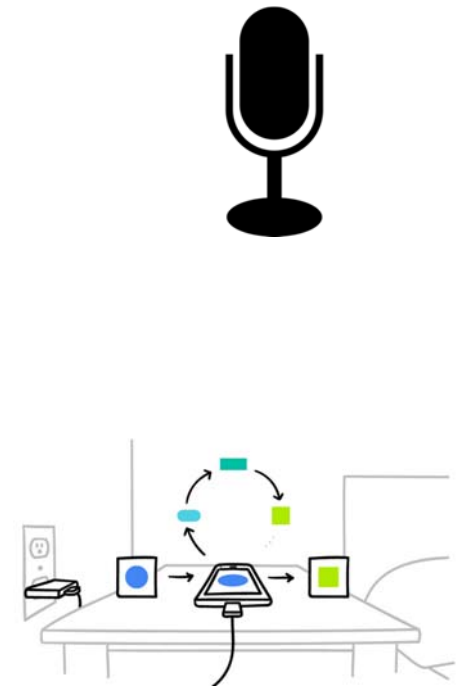


Face at 20ft

Data Privacy Compatible With Edge AI

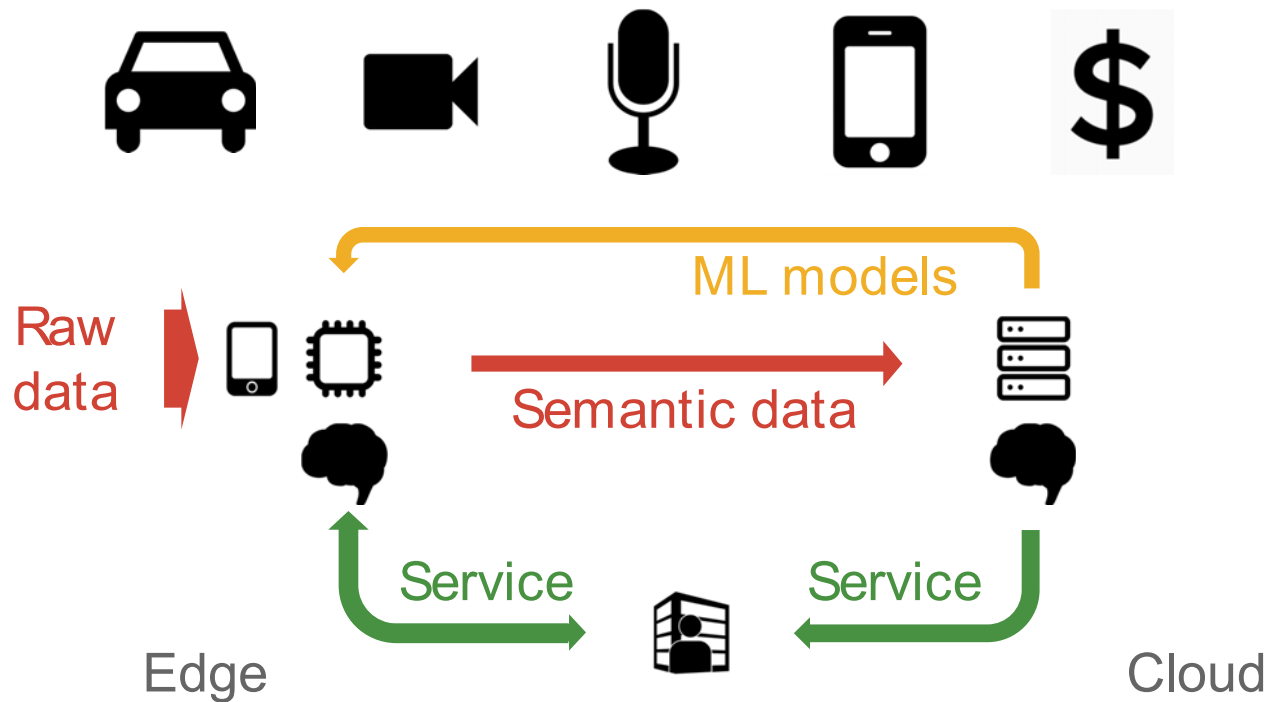


Federated Learning:
Training w/o sending data to Cloud

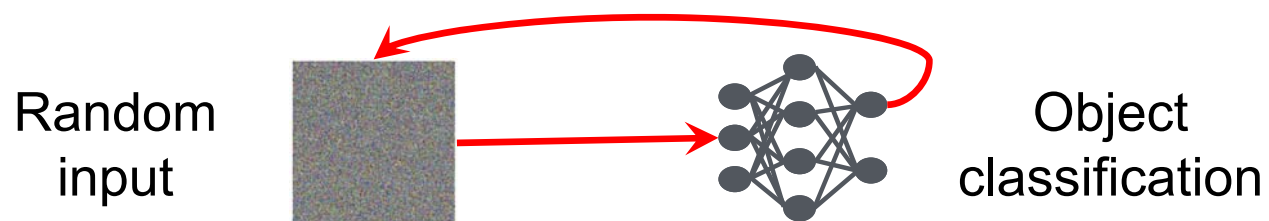


On-device training

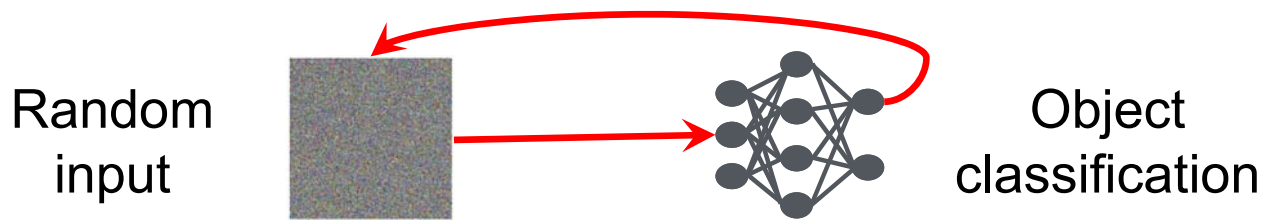
Towards A Cloud-to-Edge AI View



Thank You



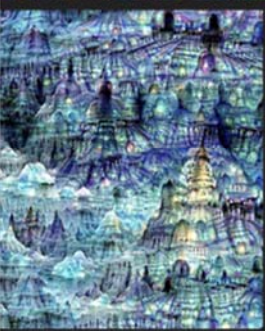
Deep NN "Dream"



Inceptionism: Going Deeper into Neural Networks (2015)



13.1: A Shift Towards Edge Machine-Learning Processing





"Deep NN dream"

Van Gogh's "The Starry Night"



"Deep NN dream"

Thank You