DESIGNING ARTIFICIAL INTELLIGENCE FOR EMBEDDED SYSTEMS

Introduction to Embedded AI

IEEE Real Time Conference

UDS Université de Sherbrooke

WELCOME

INTRODUCTION

Name

Home institution

Research

First time at IEEE RTC?

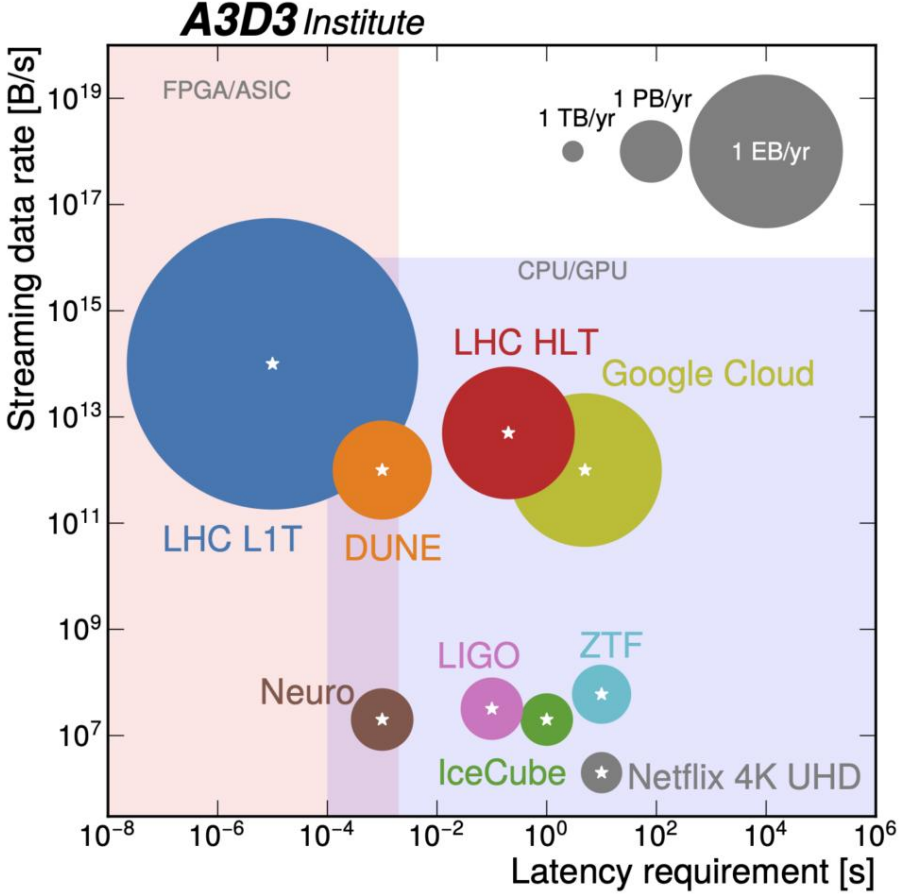
SCHEDULE

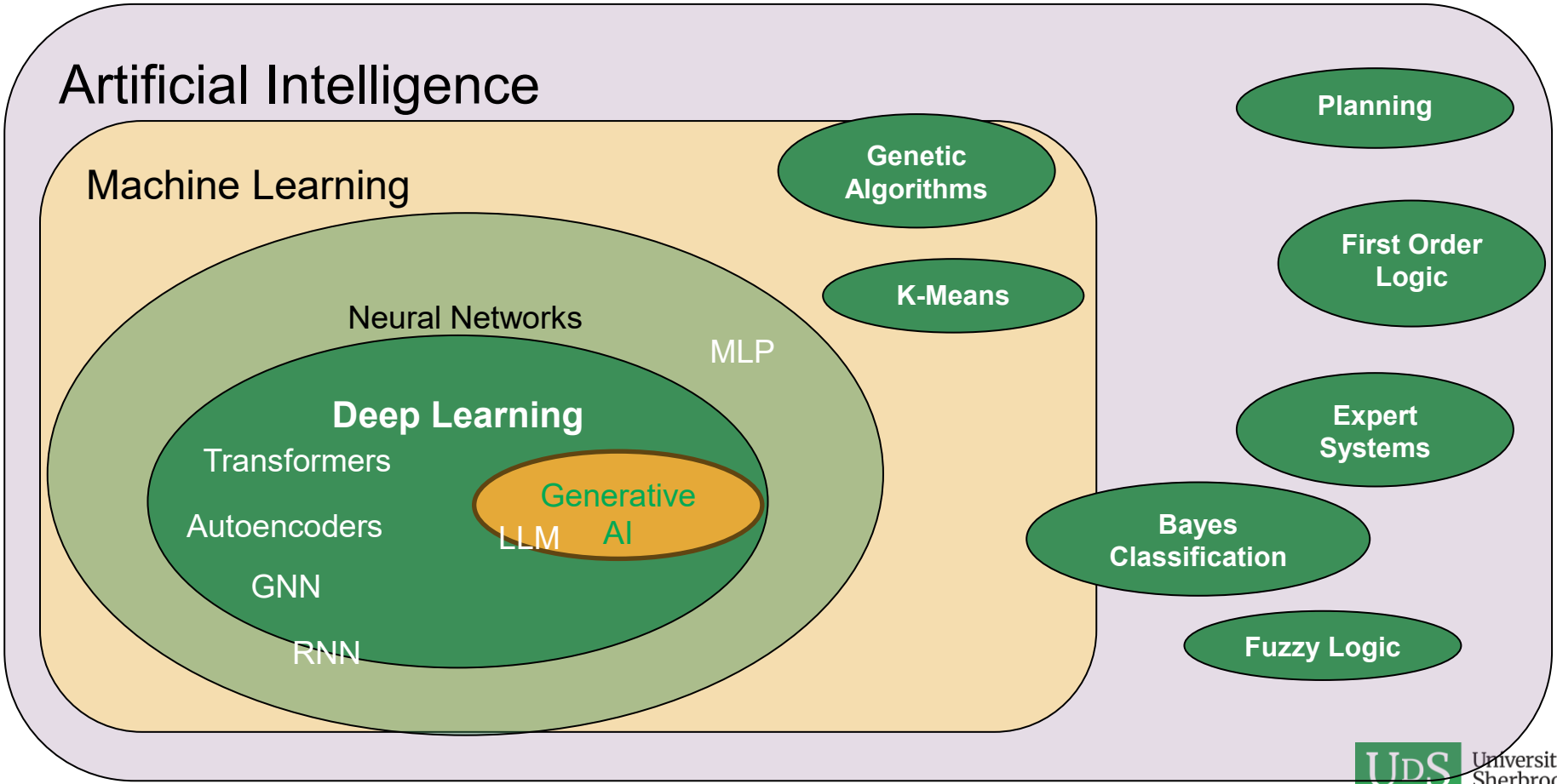
Time	Name	Instructor
9:30-12:30	Designing Artificial Intelligence for Embedded Systems	Audrey Corbeil Therrien
14:30-18:00	hls4ml	Benjamin Ramhorst
9:30-12:30	Coyote v2: Open-source Abstractions and Infrastructure for FPGAs	Benjamin Ramhorst
14:30-18:00	Super Neural Architecture Codesign Package (SNAC-Pack)	Dmitri Demler

DESIGNING AI FOR EMBEDDED SYSTEMS

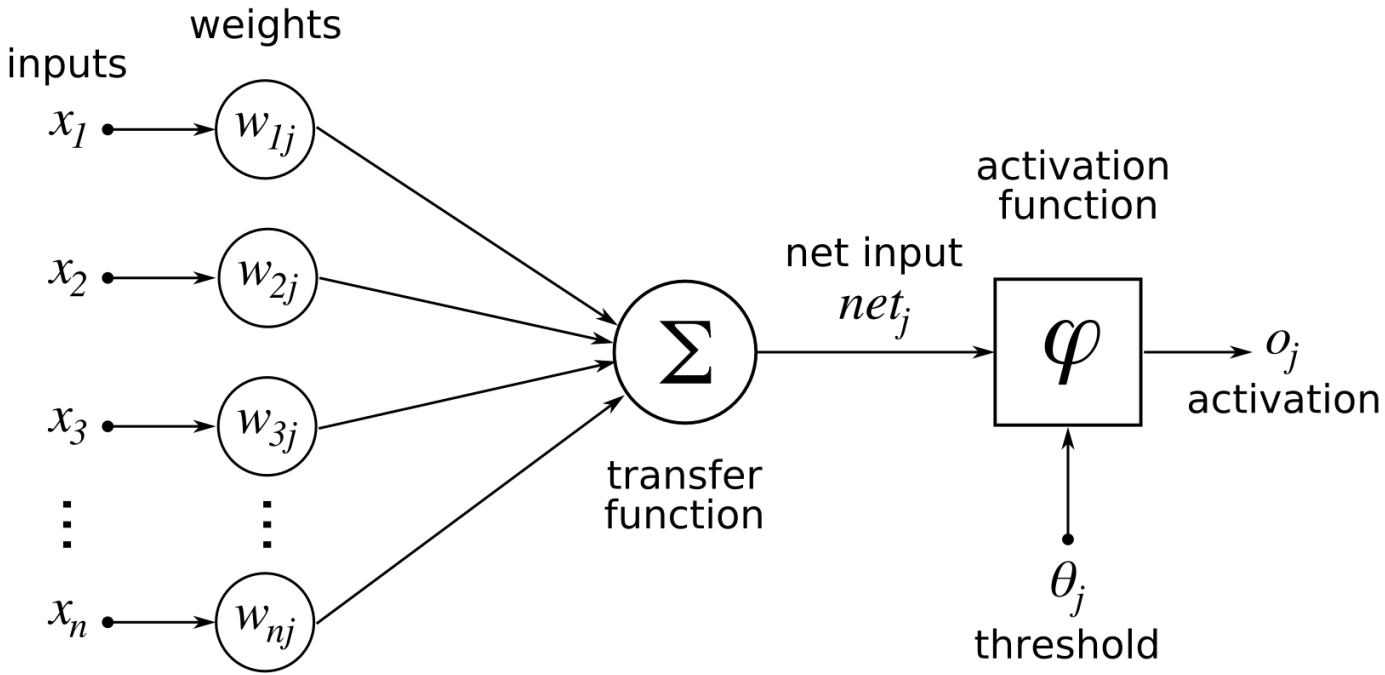
- What is AI?
- Hardware
- Co-design
- Squish
- Validation
- Overview of tools

WHAT IS AI


A3D3 Institute
FPGA/ASIC
CPU/GPU
1 TB/yr
1 PB/yr
1 EB/yr
LHC HLT
Google Cloud
LHC L1T
DUNE
LIGO
ZTF
Neuro
IceCube
Netflix 4K UHD
Streaming data rate [B/s]
Latency requirement [s]
10^{19}
10^{17}
10^{15}
10^{13}
10^{11}
10^{9}
10^{7}
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}

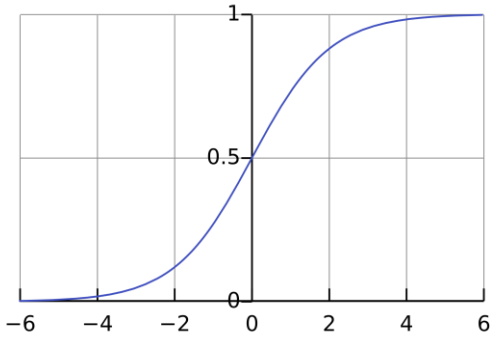

Artificial Intelligence
Machine Learning
Neural Networks
MLP
Deep Learning
Transformers
Autoencoders
GNN
RNN
Generative AI
LLM
Genetic Algorithms
K-Means
Planning
First Order Logic
Expert Systems
Bayes Classification
Fuzzy Logic

NEURON

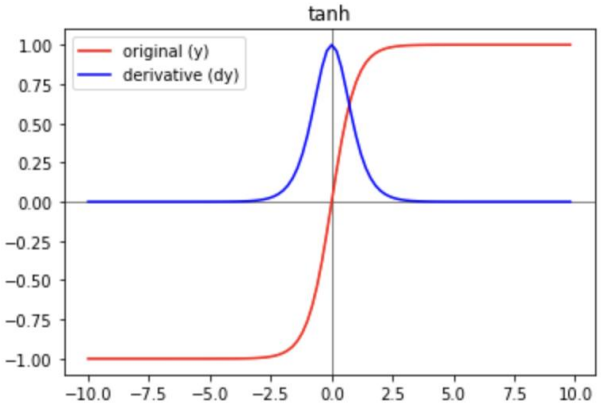


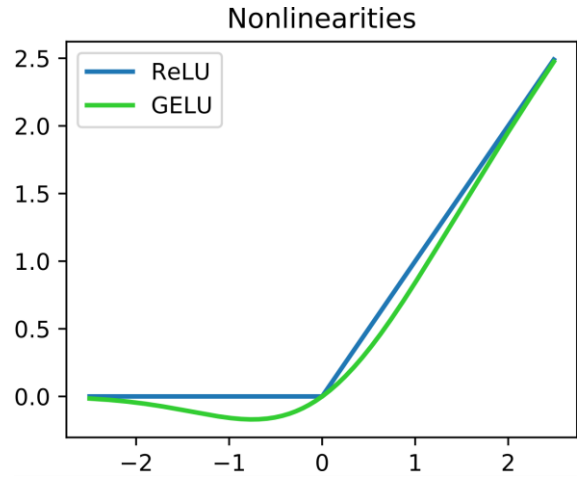
Source : Chrislb, Wikipedia, https://en.wikipedia.org/wiki/Backpropagation#/media/File:ArtificialNeuronModel_english.png

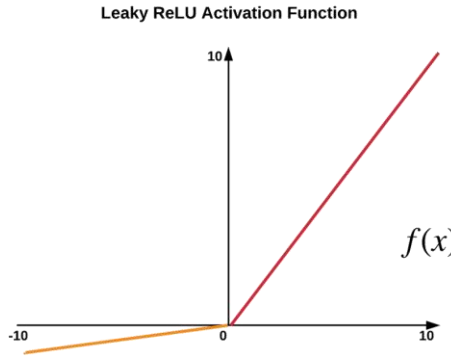
ACTIVATION FUNCTIONS



Logistic/sigmoid



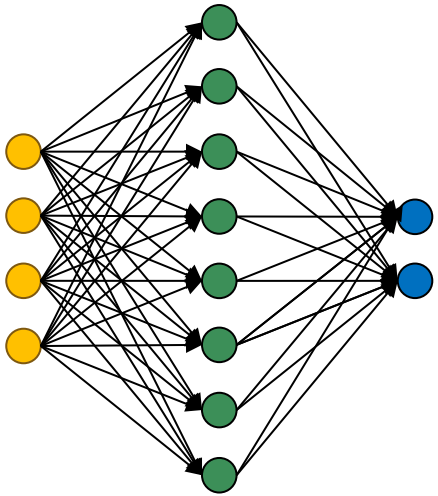




$$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

Université de Sherbrooke

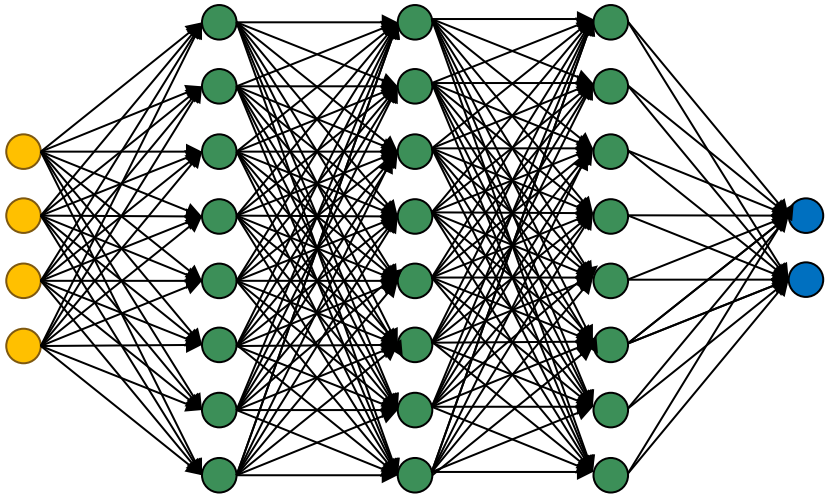
NEURAL NETWORK

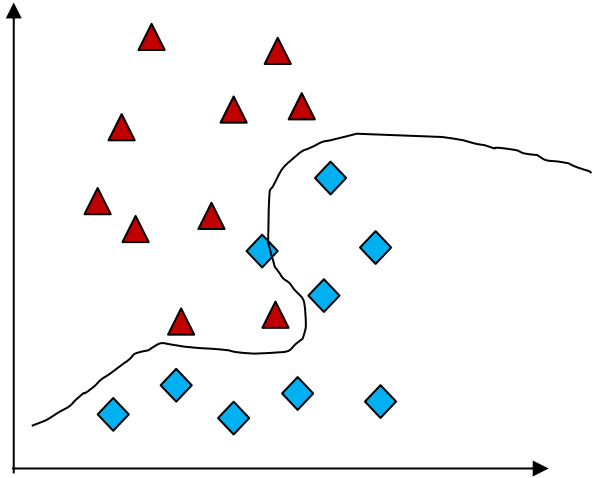


$$f(x) = \varphi(\mathbf{w} \cdot \mathbf{x} + b)$$

$$f(x) = \varphi\left(\left(\sum_{i=1}^{n} w_i x_i\right) + b\right)$$

NEURAL NETWORK





CONVOLUTIONAL NEURAL NETWORK


chocolate cookie


[unknown]


fawn smooth Chihuahua


brown coated Chihuahua


baked blueberry muffin


baked muffin


white chihuahua


beige short coated puppy


fawn smooth Chihuahua


tan smooth Chihuahua puppy


blueberry muffin


blueberry cupcakes


fawn smooth Chihuahua


three smooth Chihuahua puppies


muffin


white and black muffin

https://blog.cloudsight.ai/chihuahua-or-muffin-1bdf02ec1680

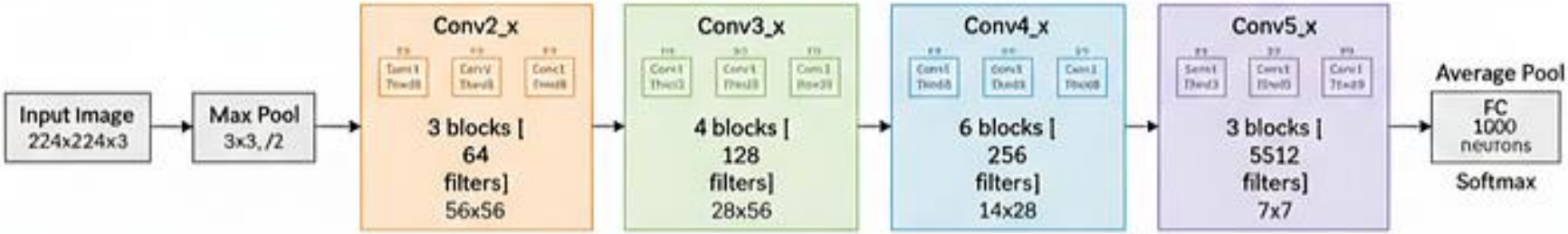
CONVOLUTIONAL NEURAL NETWORK

6	5	2	0	4	1
5	4	3	2	3	4
4	2	1	0	1	1
3	2	1	0	0	0
1	1	1	1	3	4

3	2	1
2	1	0
1	0	0

48	33	23	16
39	33	18	17
26	12	7	4

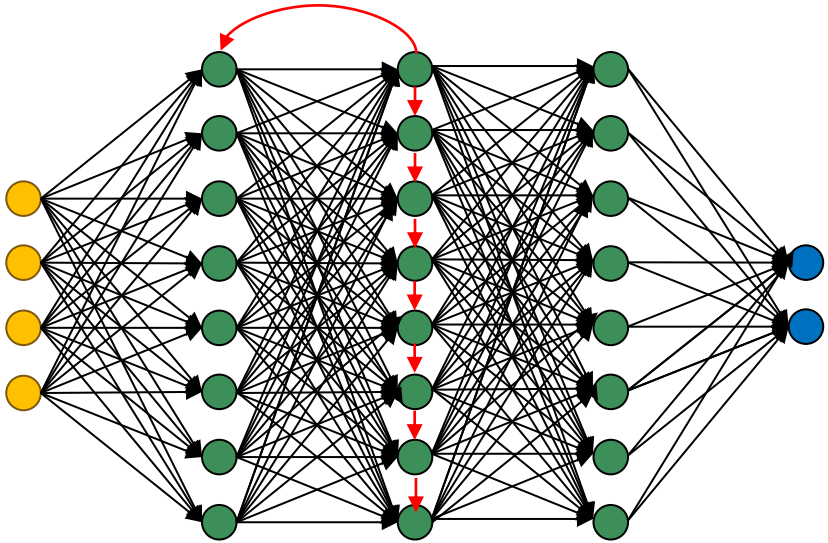
CONVOLUTIONAL NEURAL NETWORKS



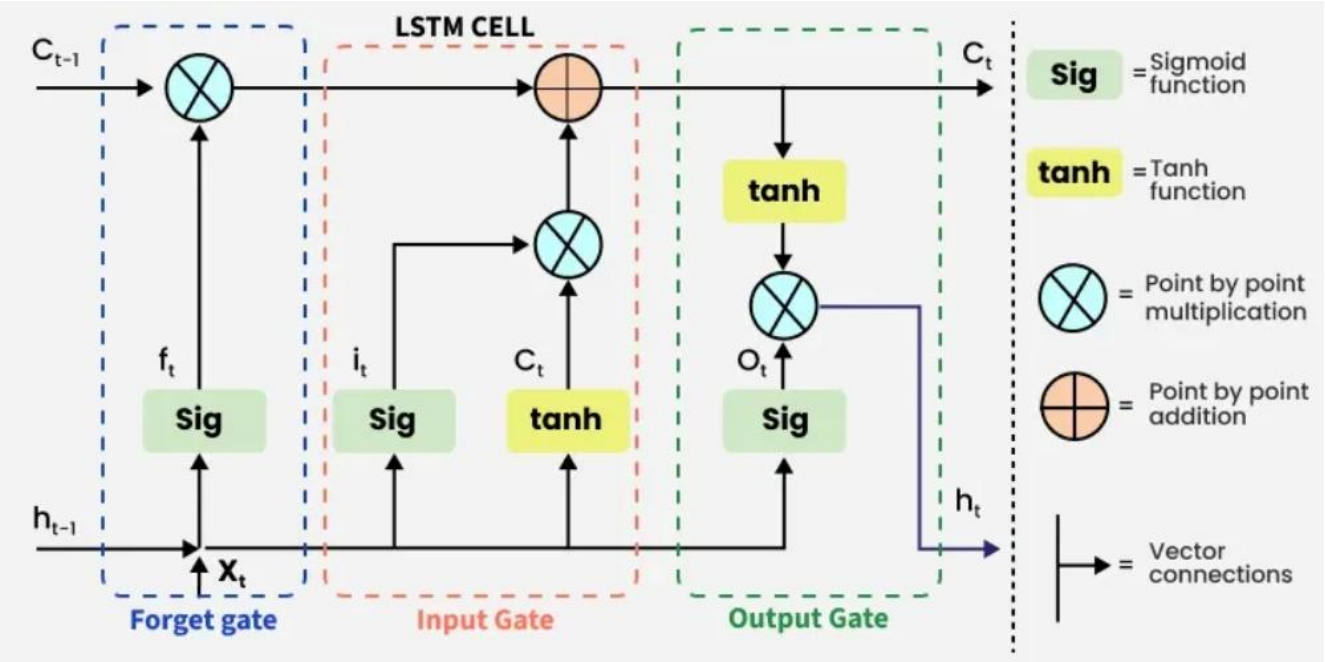
Resnet-50 - ~2016

https://medium.com/@deepvisionkararhaider/resnet-50-explained-step-by-step-the-easiest-guide-to-deep-residual-networks-7616f4f45046

RECURRENT NEURAL NETWORKS



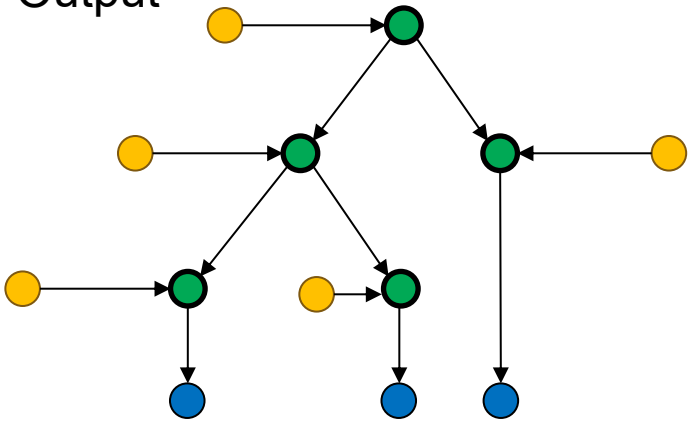
RECURRENT NEURAL NETWORKS



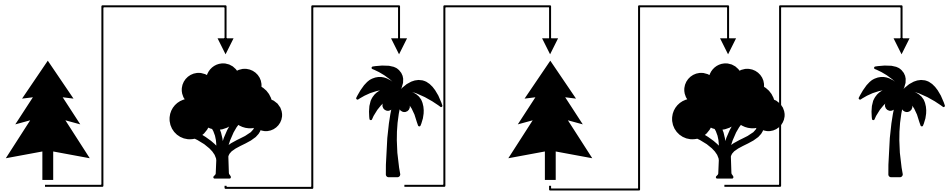
GeeksforGeeks : https://www.geeksforgeeks.org/deep-learning/deep-learning-introduction-to-long-short-term-memory/

DECISION TREE

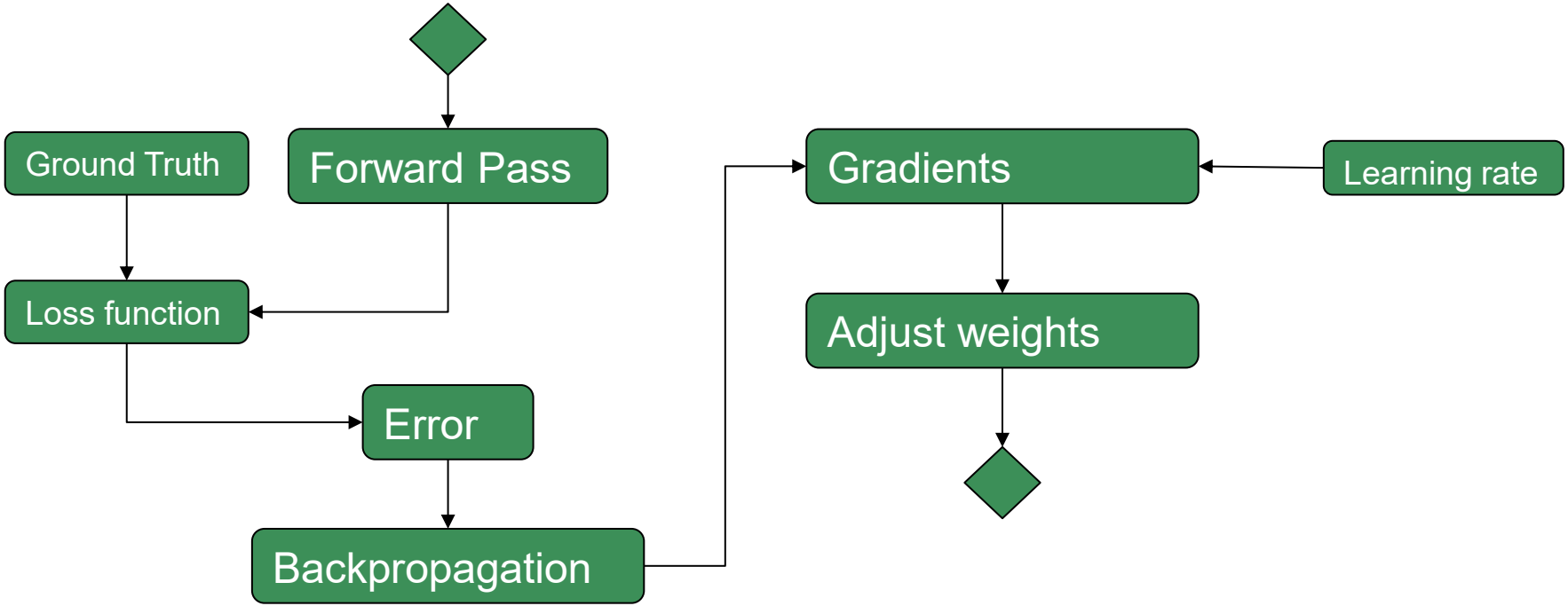
- Input
- Decision
- Output



Gradient boosting
Decision trees



TRAINING


Ground Truth
Forward Pass
Loss function
Error
Backpropagation
Gradients
Learning rate
Adjust weights

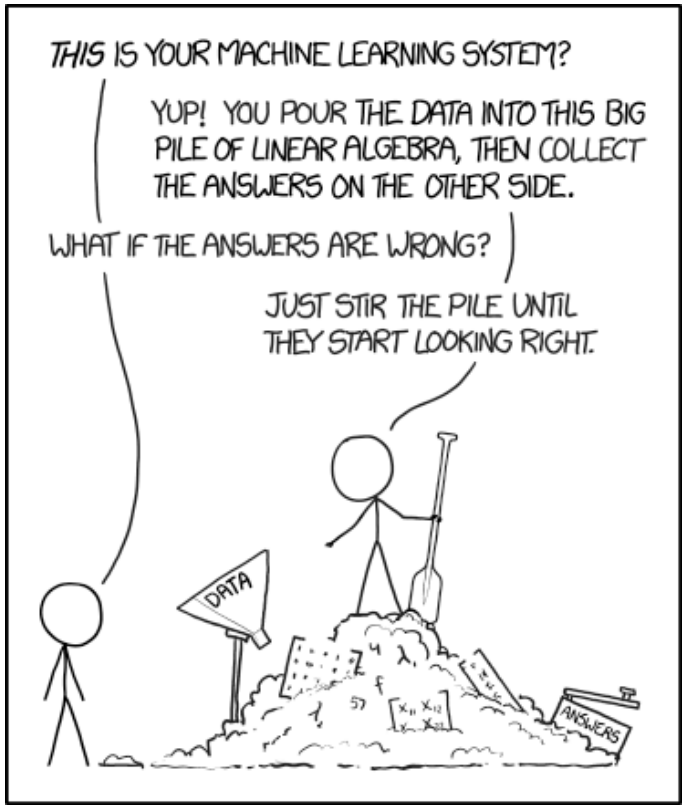
DATASETS

- Source
- Accuracy
- Authenticity
- Scope
- Usability

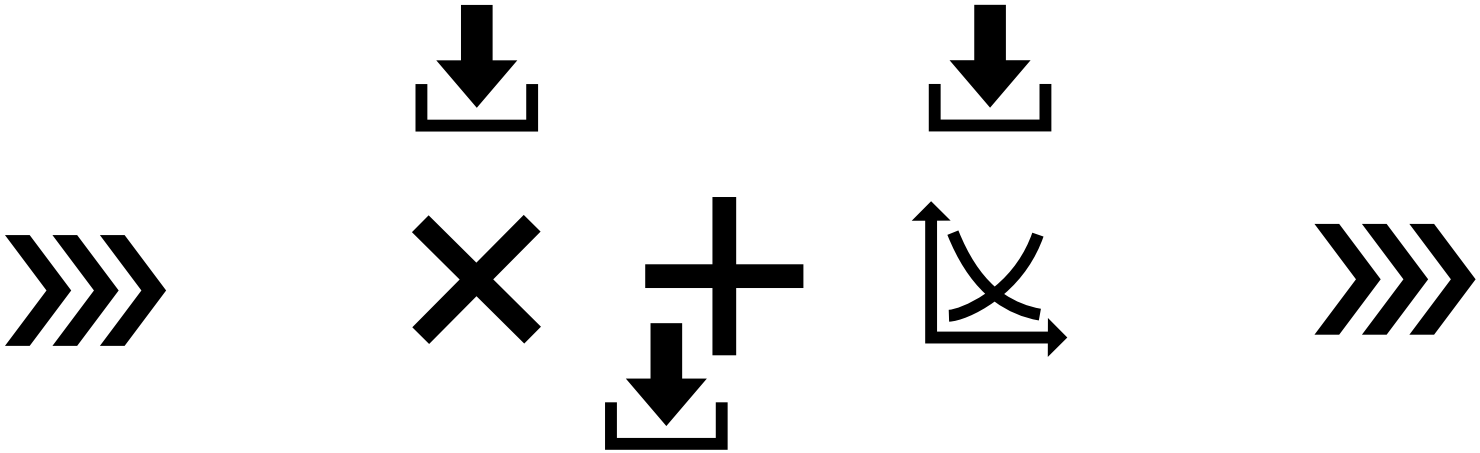
MODELS

- A model is only as good as its data
  - (at best…)

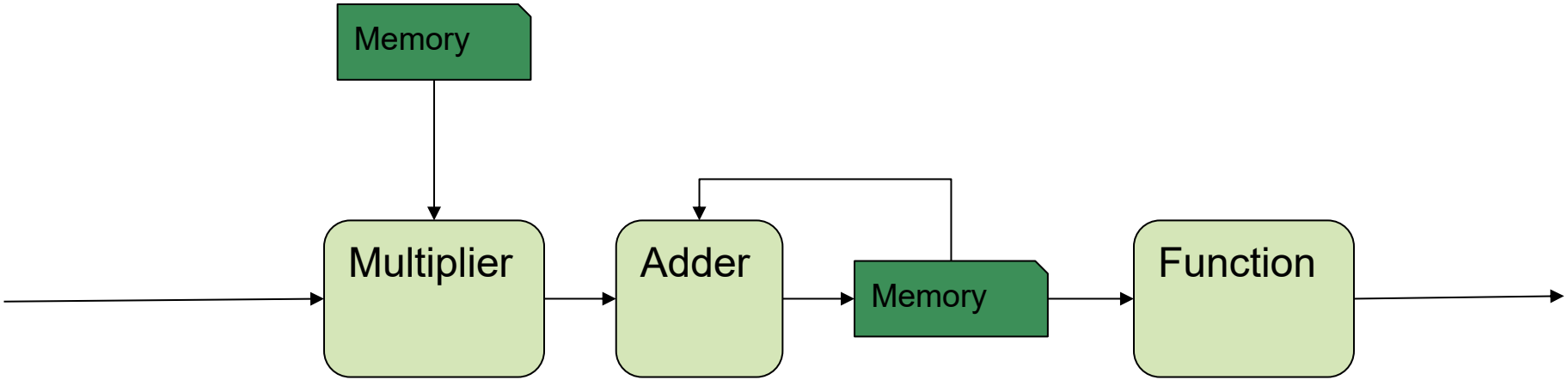
- “All models are wrong, but some are useful.”
  - George Box



OPERATIONS OF A NEURAL NETWORK

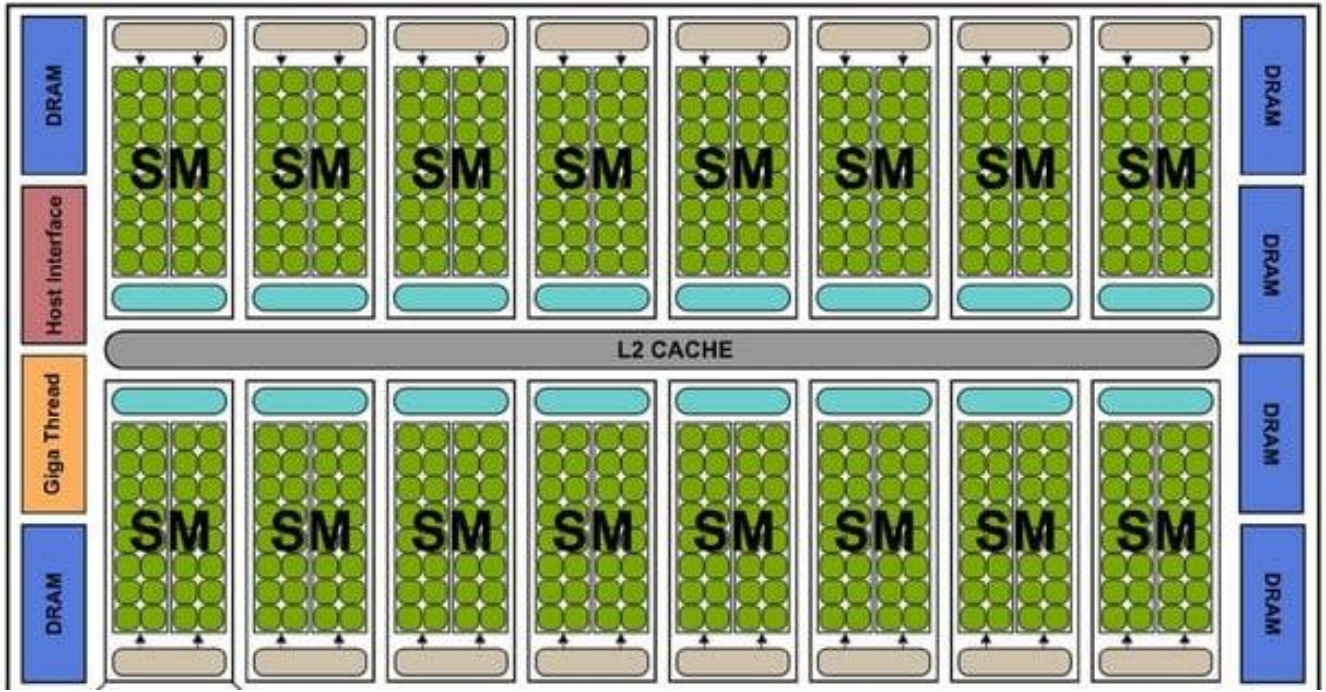


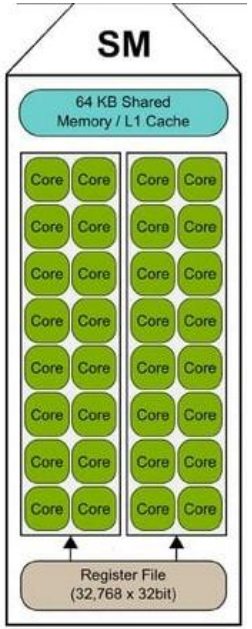
OPERATIONS OF A NEURAL NETWORK



HARDWARE

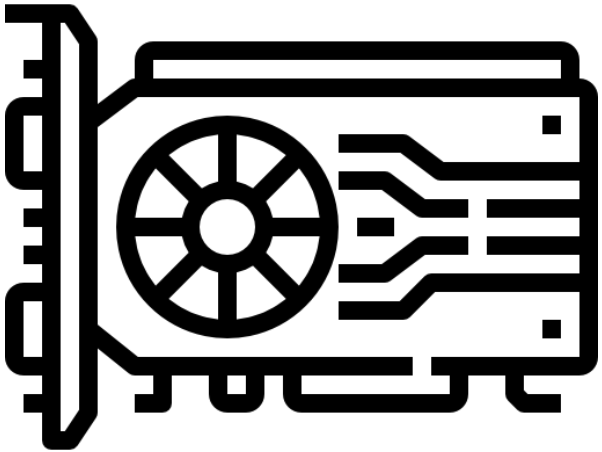
GRAPHICS PROCESSING UNIT





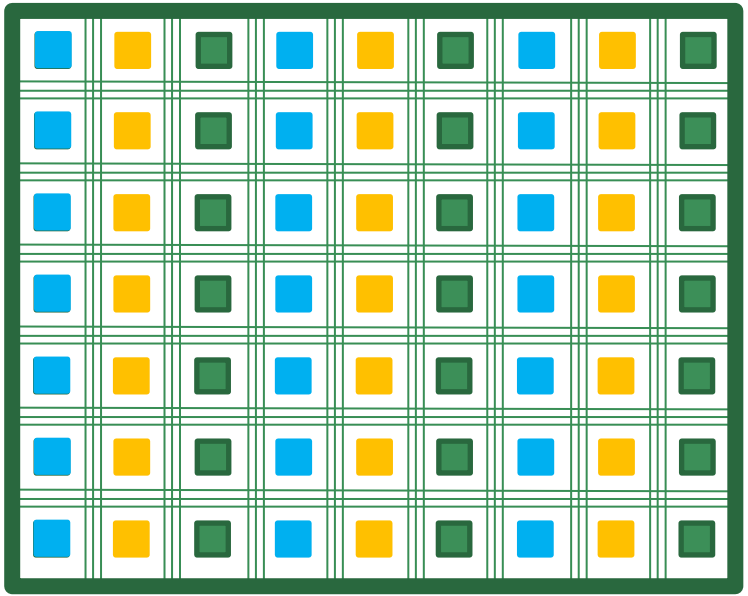
Source : Modified from https://www.researchgate.net/publication/236666656_Accelerating_Fibre_Orientation_Estimation_from_Diffusion_Weighted_Magnetic_Resonance_Imaging_Using_GPUs/figures?lo=1

GRAPHICS PROCESSING UNIT



- Highly parallel
- Flexible
- Well supported
- Limited I/O
- Large footprint
- High power usage

FPGA



Logic
Memory
Digital signal processing slices

Reconfigurable

Efficient

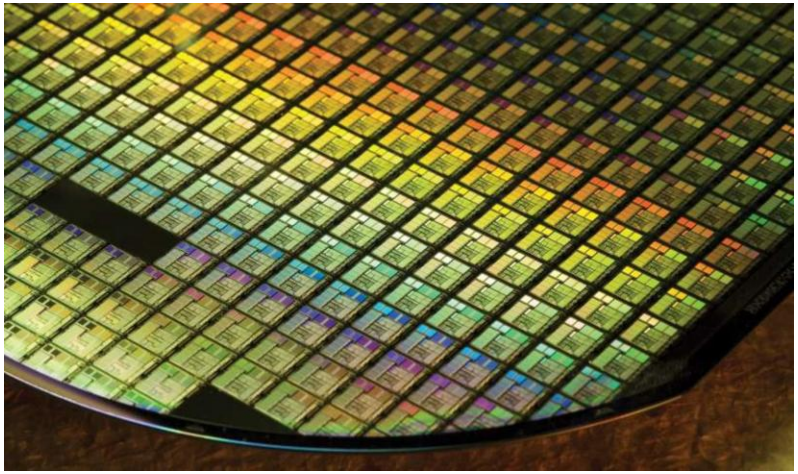
I/O capacity

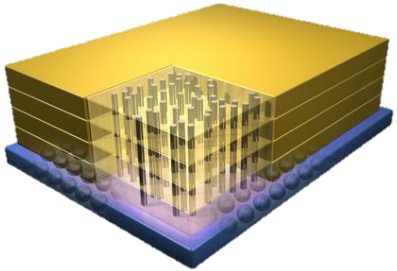
Programming

Limited clock

Limited resources

ASIC





Efficient++

Custom I/O

3DIC

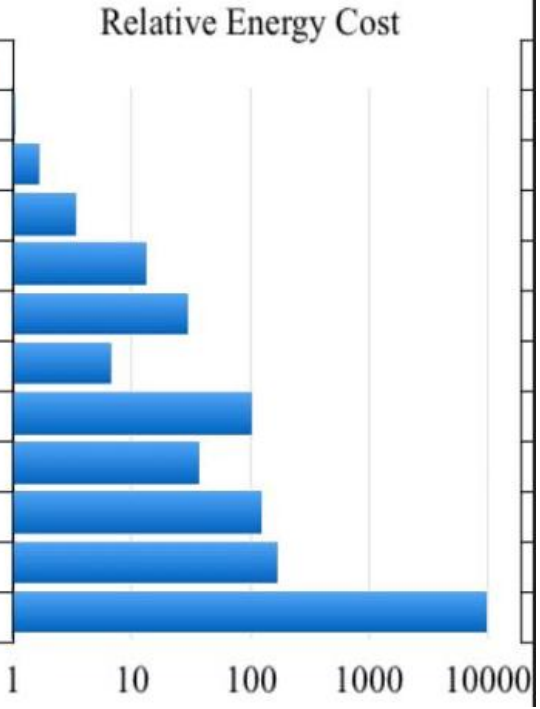
Reconfigurable

Expensive

Long design cycle

ENERGY

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

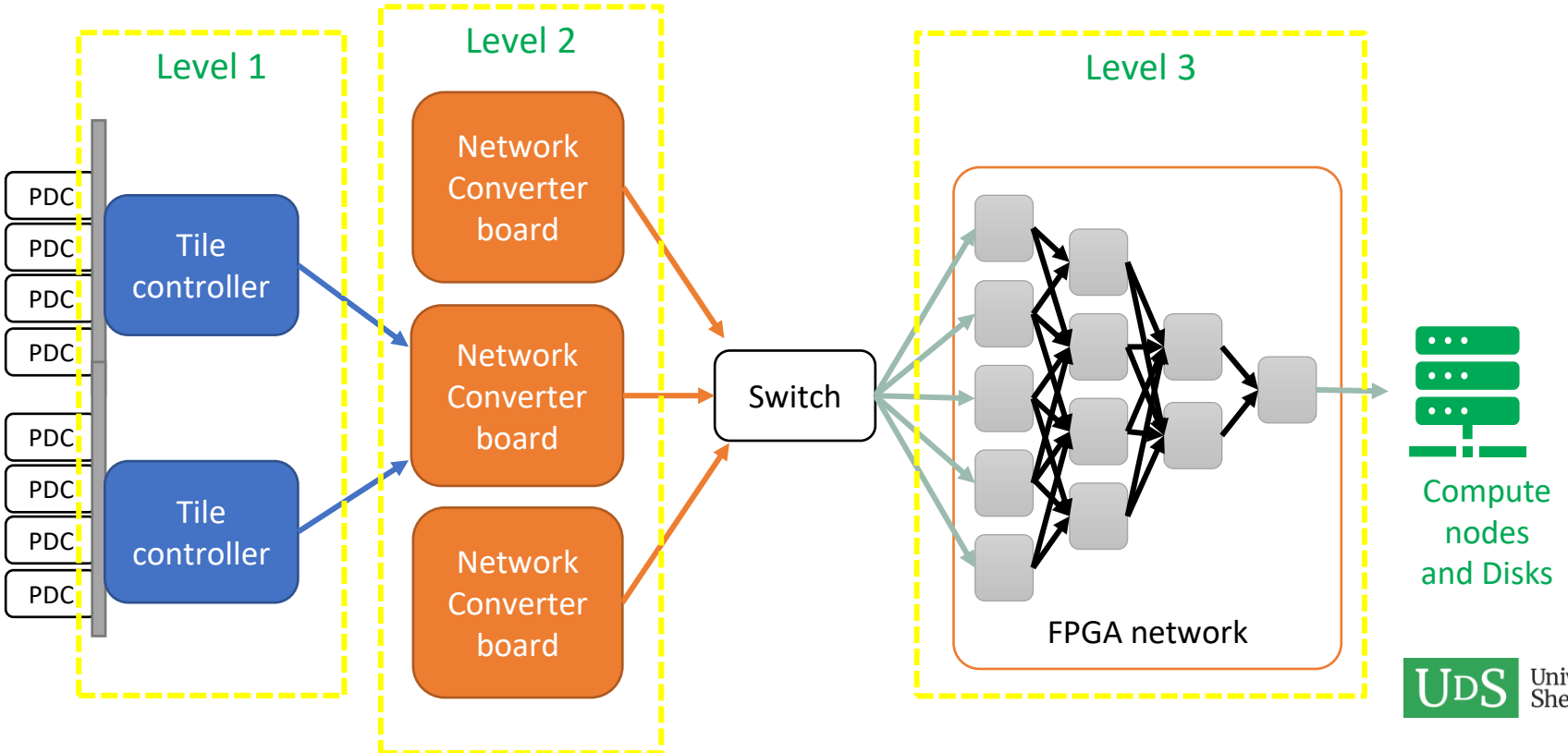


Computing's Energy Problem :10.1109/ISSCC.2014.6757323

EFPGA

- Embedding FPGA logic in an ASIC designs
  - Combine the low power and interconnectivity of ASIC with the reconfigurability of FPGA
  - More options to update the ML architecture
  - Pre-determined dimensions
  - Can be expensive depending on technology node

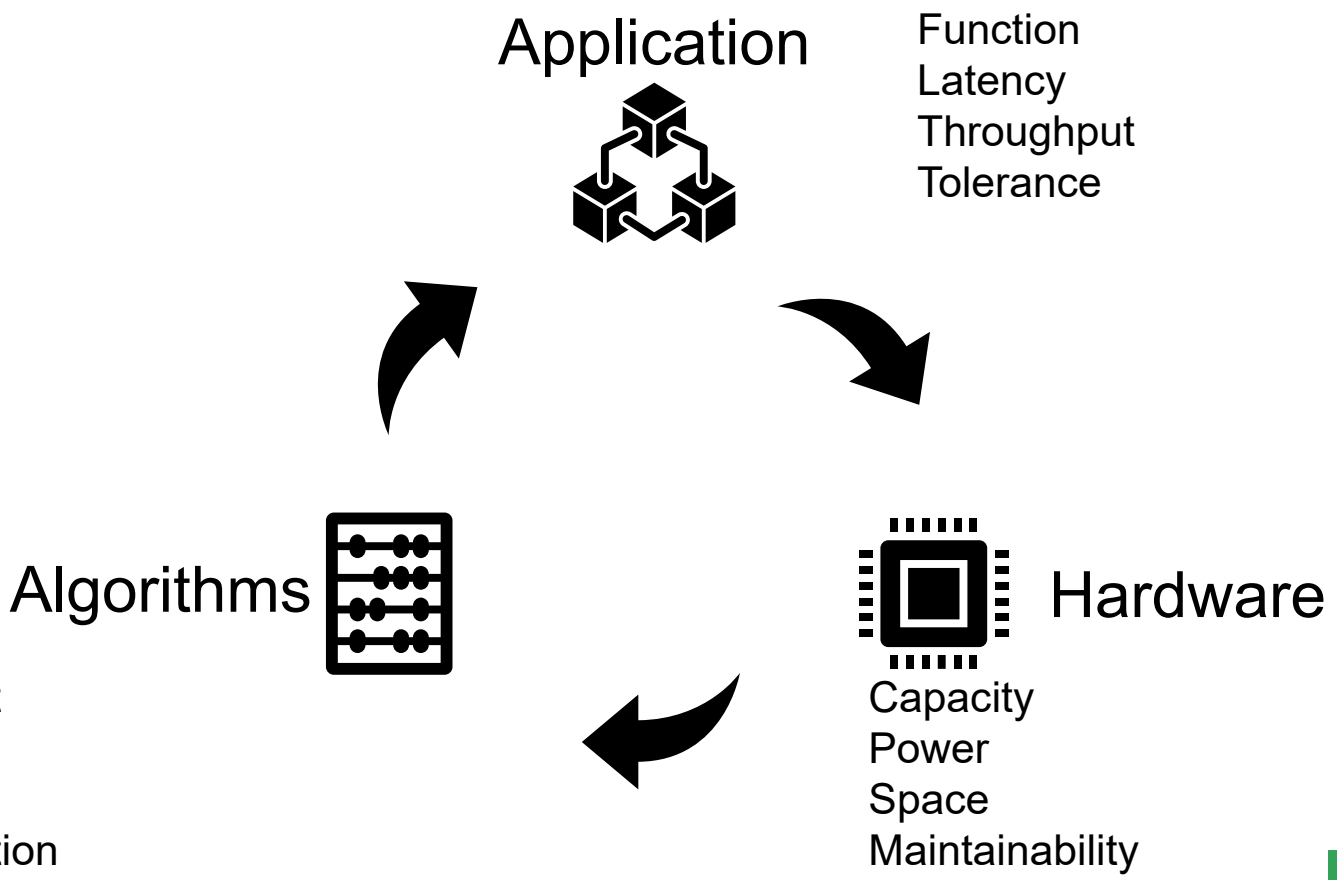
DISTRIBUTED COMPUTE



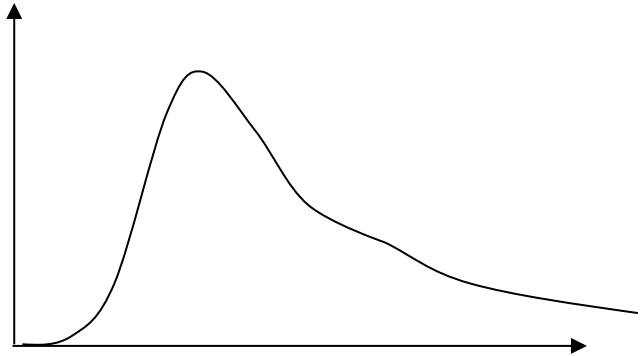
DISTRIBUTED COMPUTE

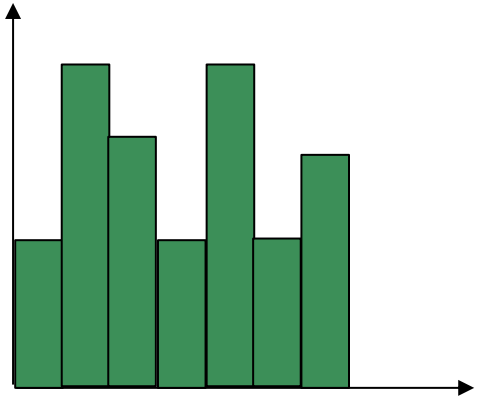
- Look for unused compute potential
  - ASIC
  - Network Switches
  - Data movers
- Some model architectures reduce data
  - Lower requirements on data links
- Tools are under developement

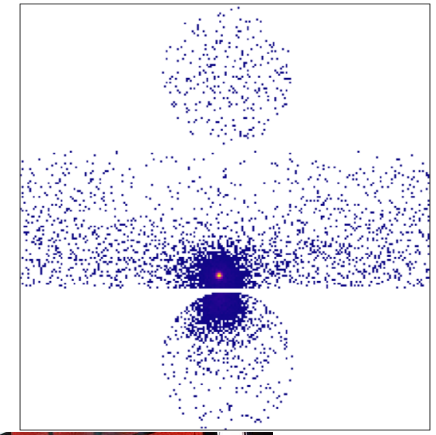
CO-DESIGN


Application
Function
Latency
Throughput
Tolerance
Hardware
Capacity
Power
Space
Maintainability
Algorithms
Data format
Complexity
Depth
Error detection

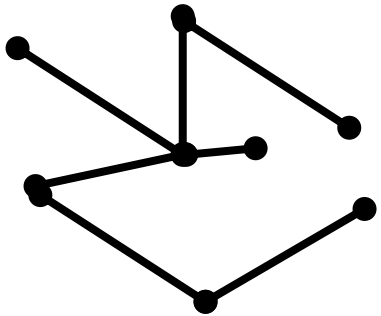
INPUT-OUTPUT

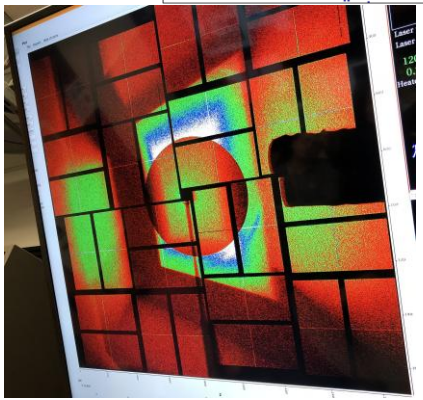






Feature	Value
Energy	412
Position	(8,6,4)
Fprompt	0,56
Direction	(1,1,6)
Color	Blue





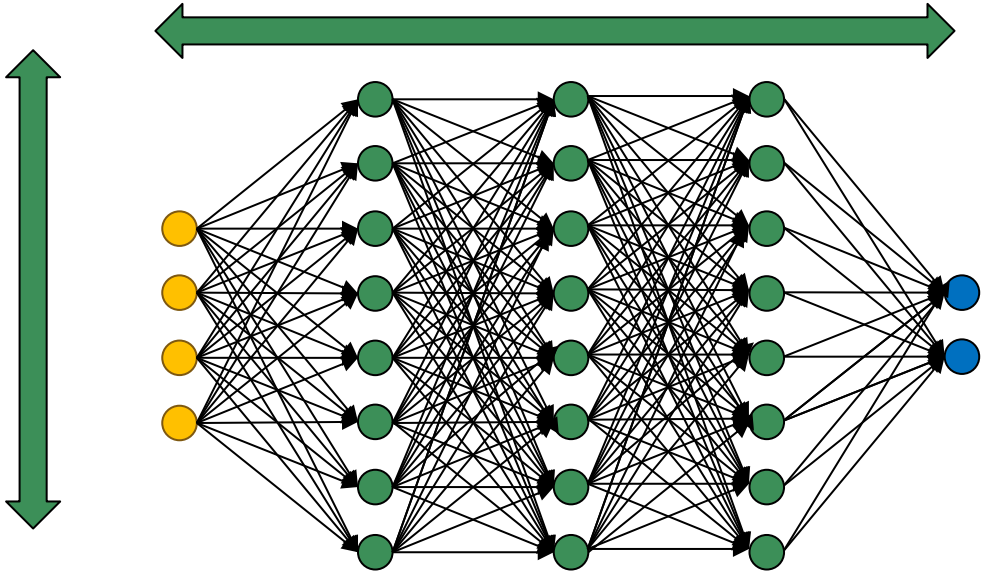
UDS Université de Sherbrooke

INPUT/OUTPUT

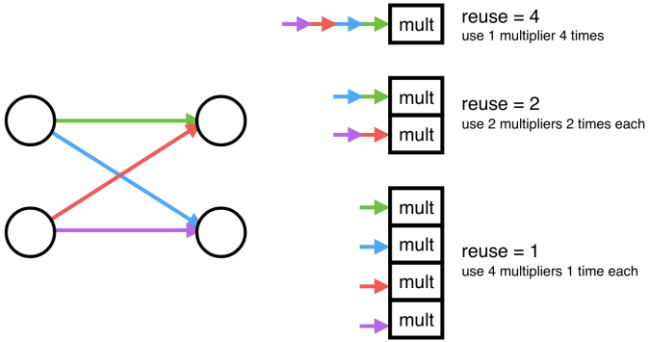
- Type
- Dynamic range
  - Normalisation
- Representation
- Transformation?

Type	Range
FP64/double	1.7E +/- 308 (fifteen digits)
FP32/float	3.4E +/- 38 (seven digits)
INT32/ long int	-2,147,483,648 to 2,147,483,647
INT16/ Short int	-32,768 to 32,767
Unsigned INT8	0 to 255

STRUCTURE/DIMENSIONS



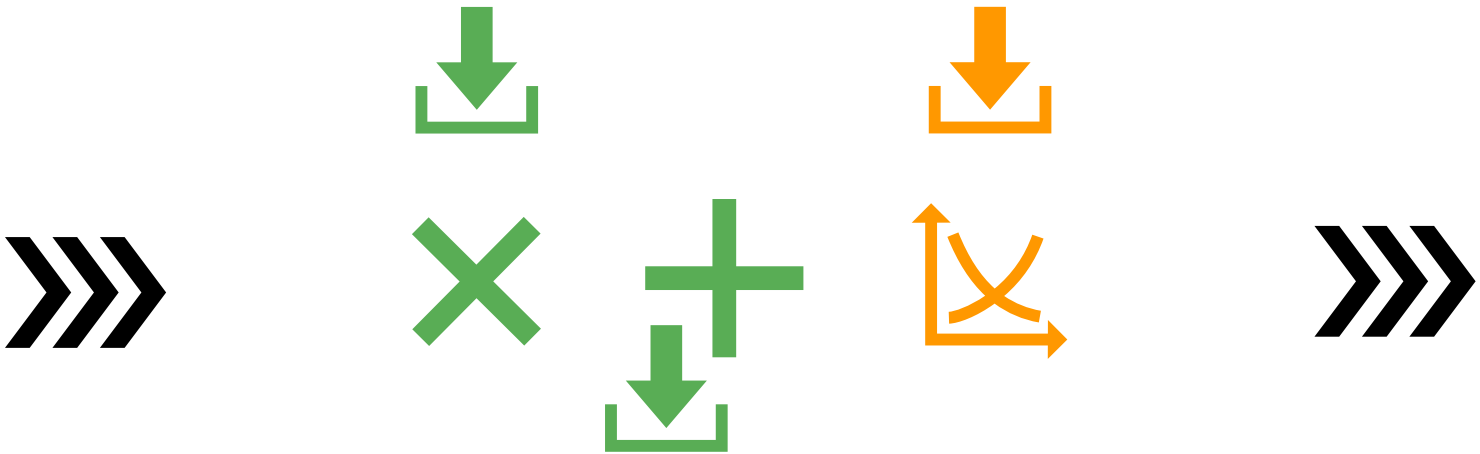
Unrolled Vs Reuse



Hls4ml
https://fastmachinelearning.org
/hls4ml/concepts.html

2,4,8,16,32,64,128.

OPERATIONS



ACTIVATION FUNCTION

- Calculation is often expensive
- Look-up table may be memory intensive
- Not all activations are implemented

- Use the same activation if you can
- Retrain with a simpler activation function

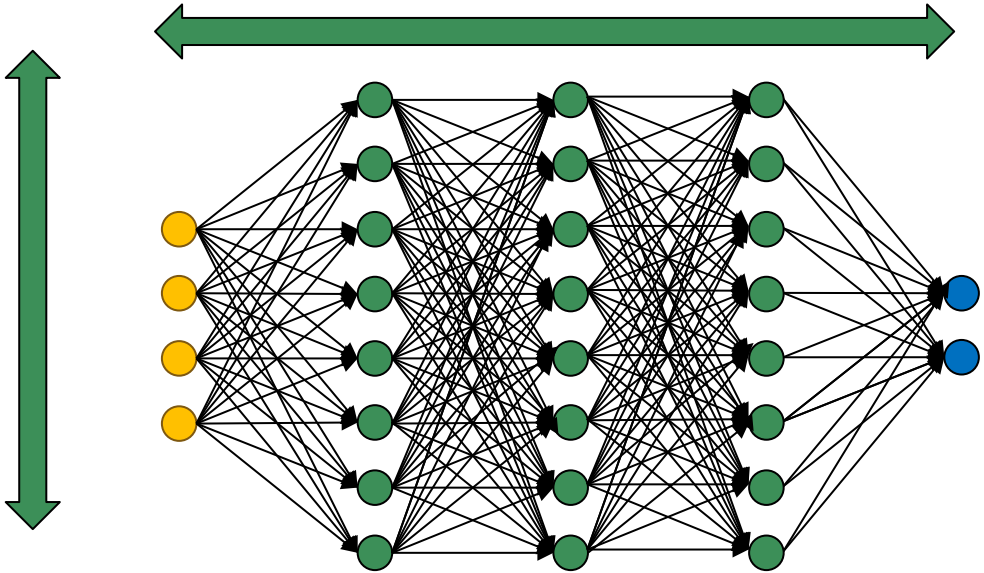
MODEL STRUCTURE

- Not all operations are implemented by tools

- Non-standard architectures can fail to synthesize, or have unexpected behavior

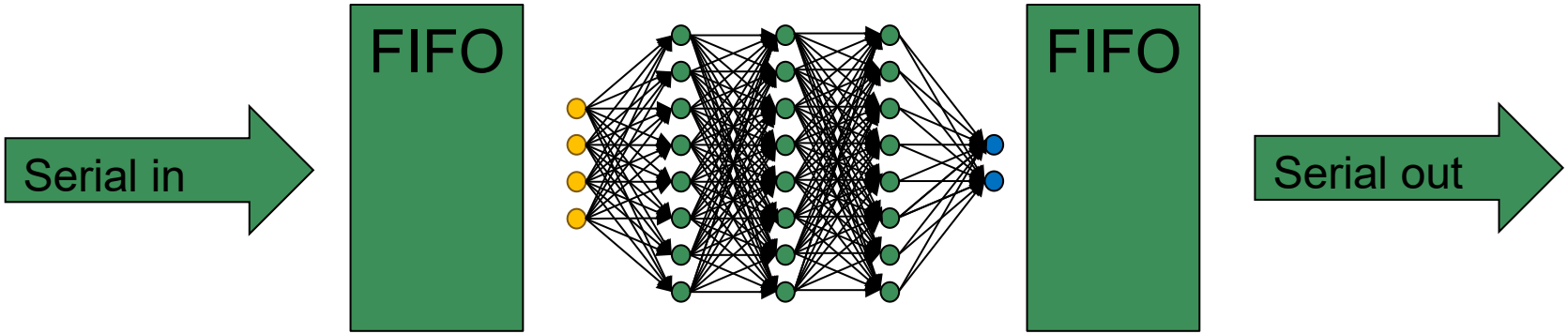
DATA FLOW

- Latency vs throughput



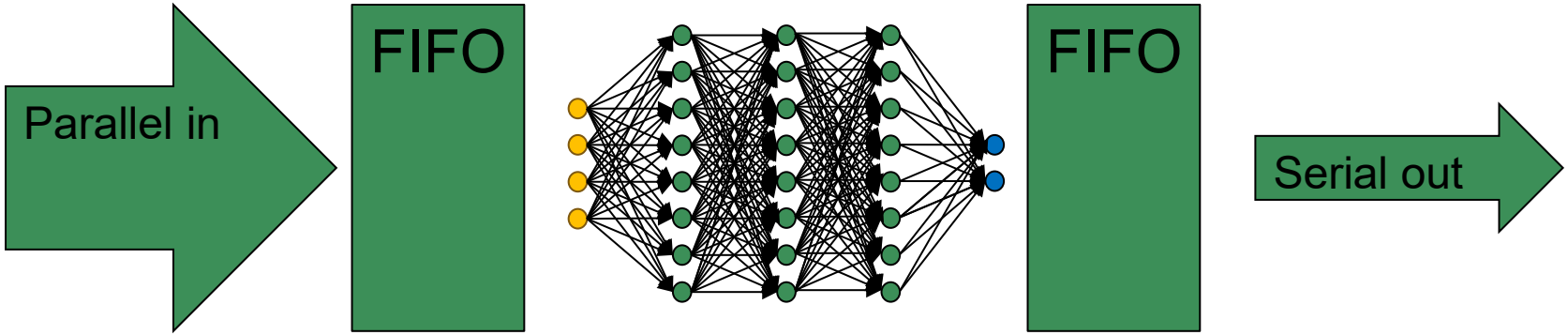
DATA FLOW

- Buffers
- Communication



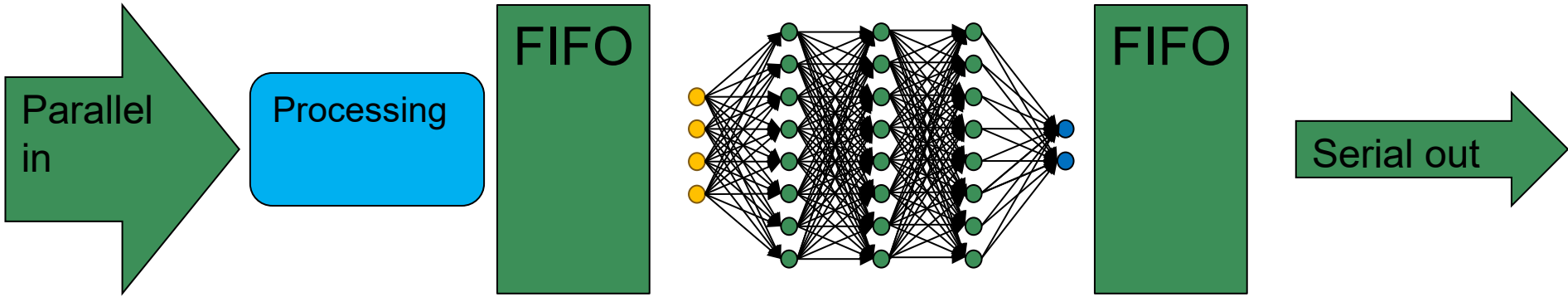
DATA FLOW

- Buffers
- Communication



DATA FLOW

- Buffers
- Communication



SUMMARY

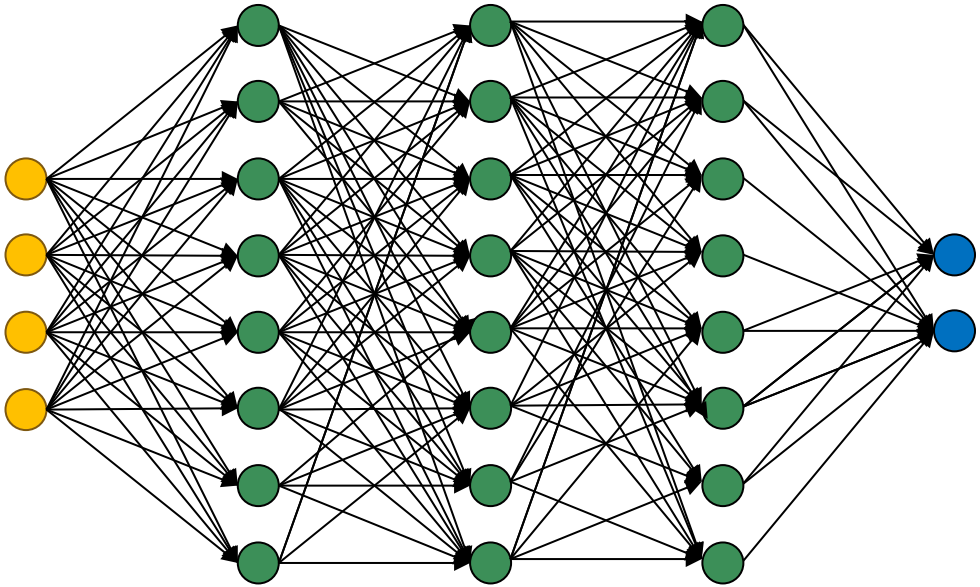
- Input-Output
- Dynamic range
- Structure/dimensions
- Operations
- Data flow requirements

- How will this work in hardware?

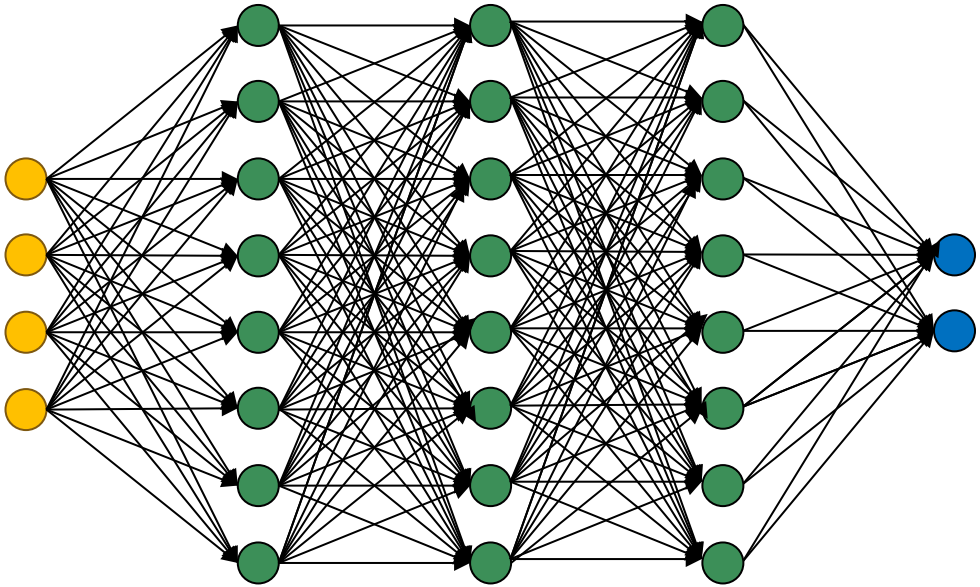
SQUISH

PRUNING

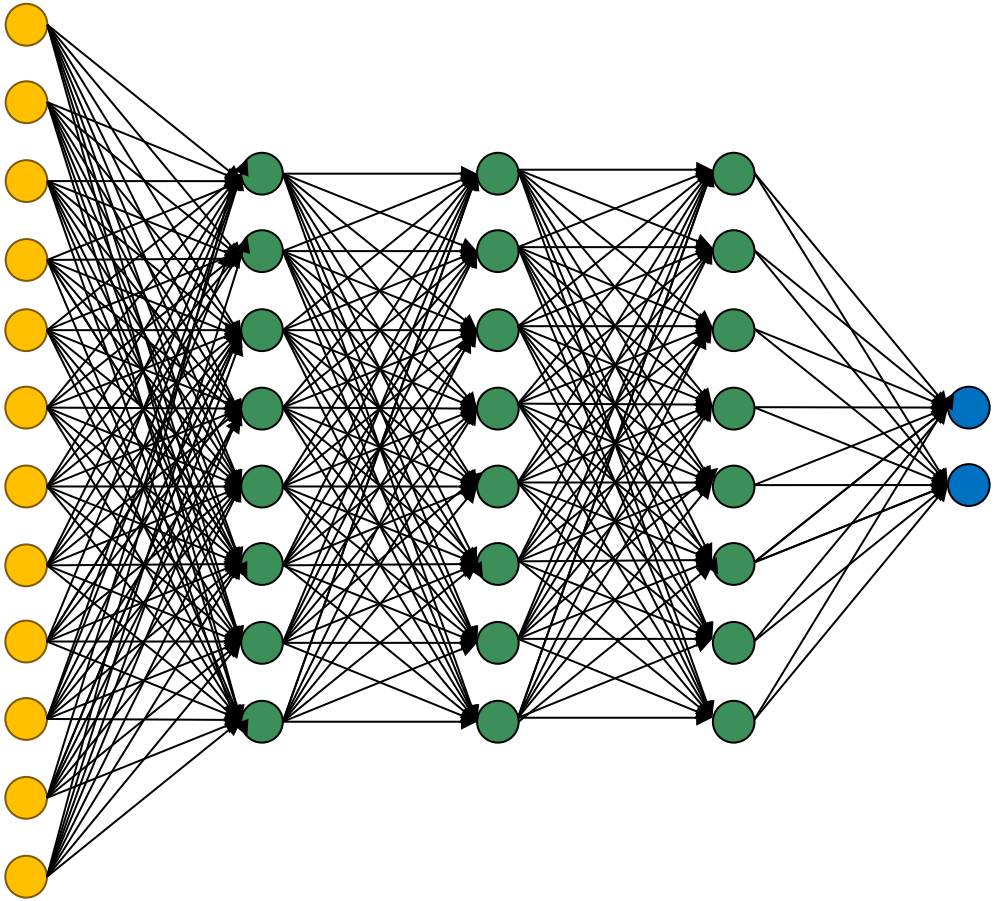
!! Logic overhead



NODES AND LAYERS

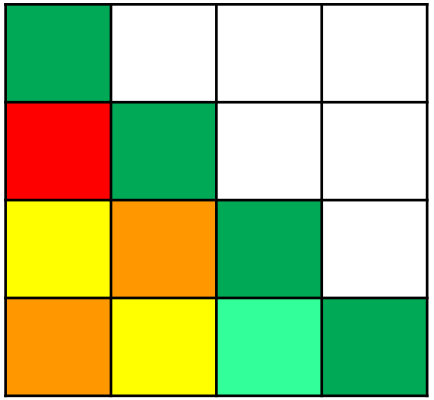


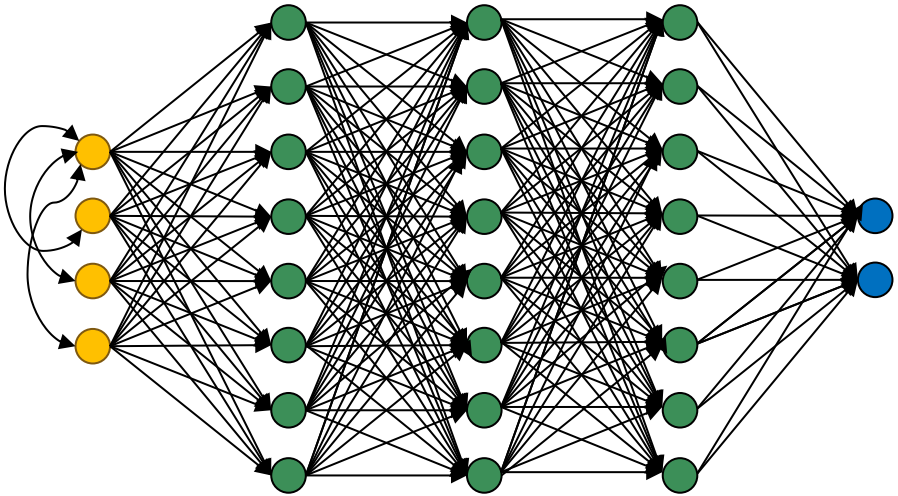
INPUTS



INPUTS – FEATURE IMPORTANCE

- Permutation
  - Swapping inputs

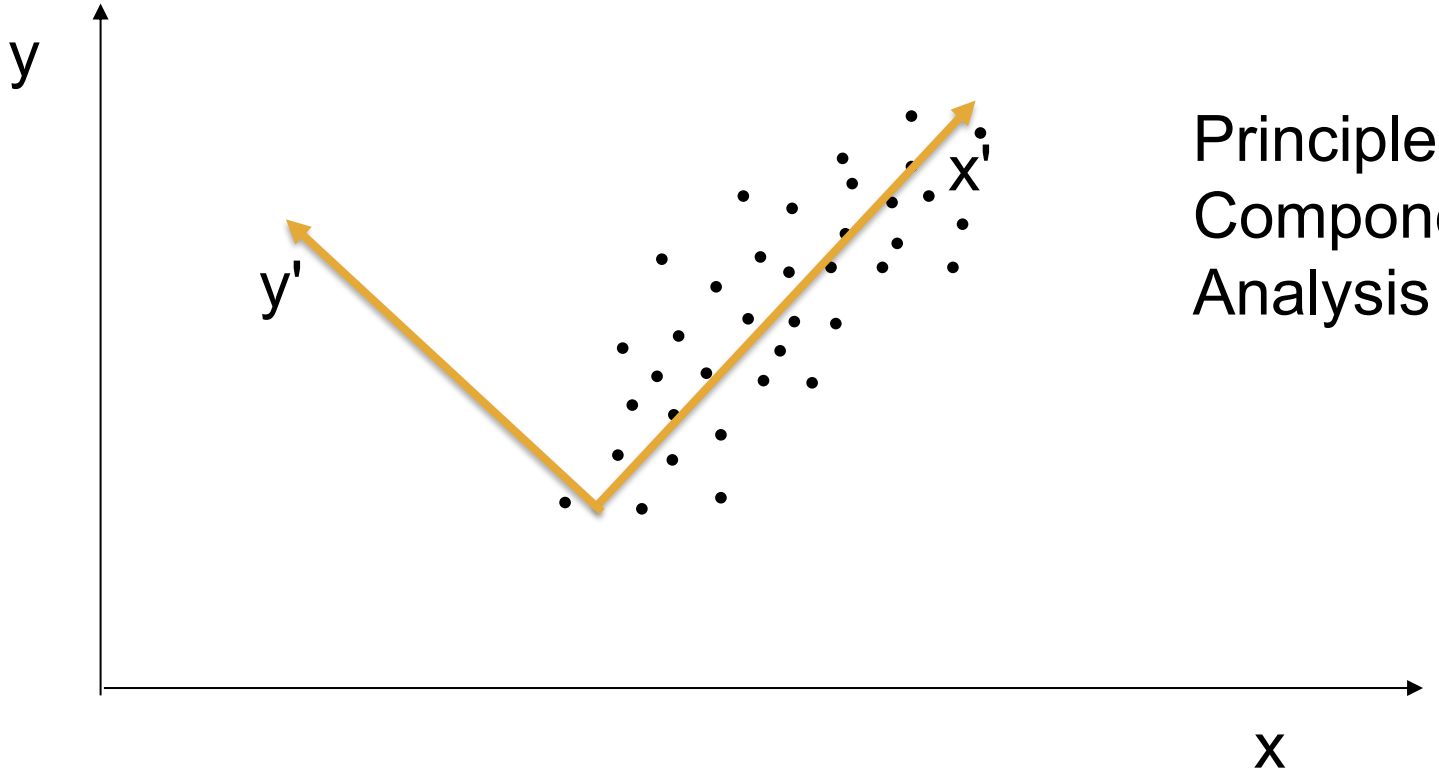




INPUTS – FEATURE IMPORTANCE

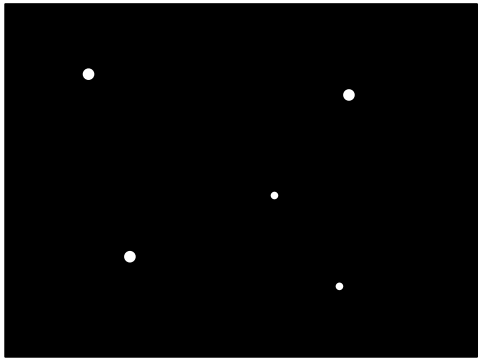
- Feature Ablation
  - Remove, retrain, compare
- SHAP (SHapley Additive exPlanations)
  - https://shap.readthedocs.io/en/latest/

INPUT DECORRELATION



Principle Component Analysis

FEATURE EXTRACTION



(15,10,3)
(10,45,3)
(30,20,1)
(40,8,1)
(42,42,3)

If f(t) > threshold
then t

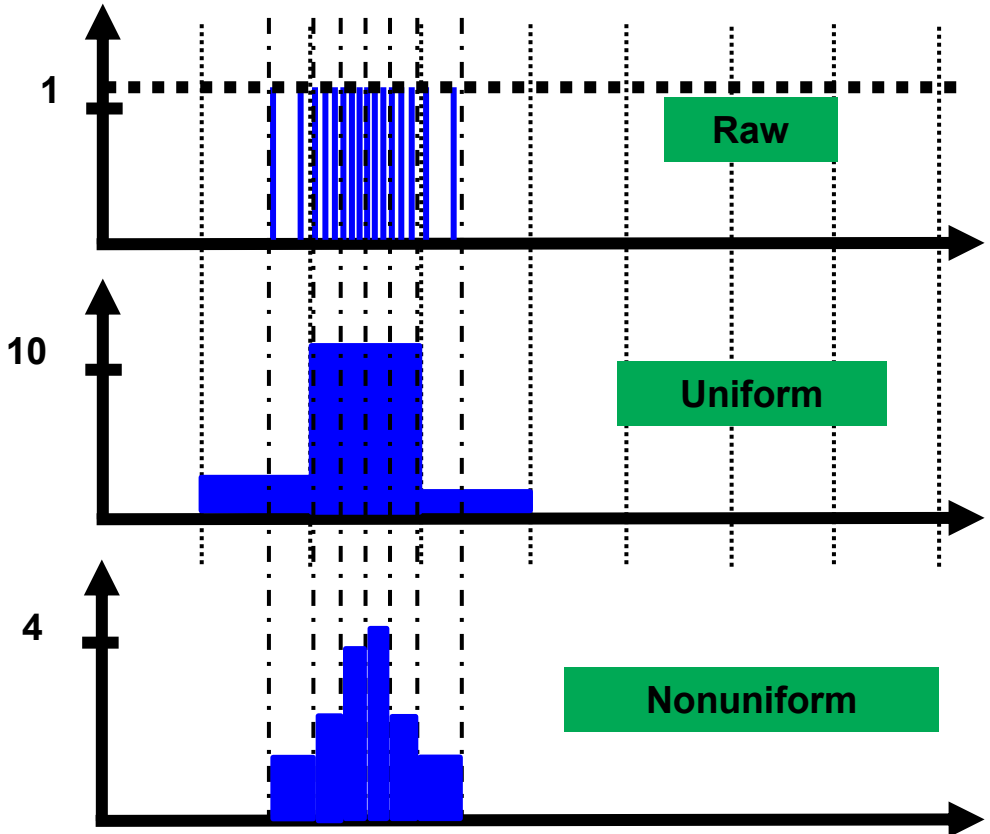
QUANTIZATION - WEIGHTS

- Reduces memory footprint
- Allows simpler multipliers/adders
- Reduces latency and power

FP32 to INT8
4x reduction in memory
20x reduction in power
Reduction in clock cycles

- Careful of dynamic range!
  - Overflow
  - Representation

QUANTIZATION - INPUTS

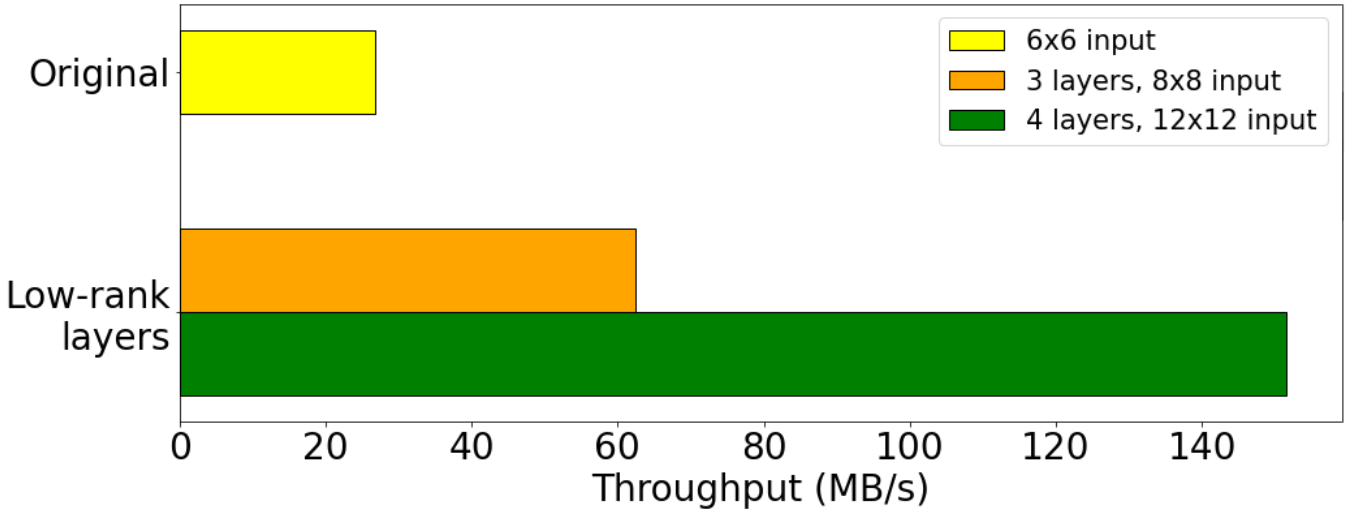


- ✓ Optimized for the data probability distribution function (PDF)
- ✓ Minimizes the quantization error
- ✓ Reduces our input shape dimensionality

https://doi.org/10.3389/fphy.2022.957128

MATRIX FACTORISATION

- Reduce the size of a layer by decomposing into smaller matrices.



MATRIX FACTORISATION

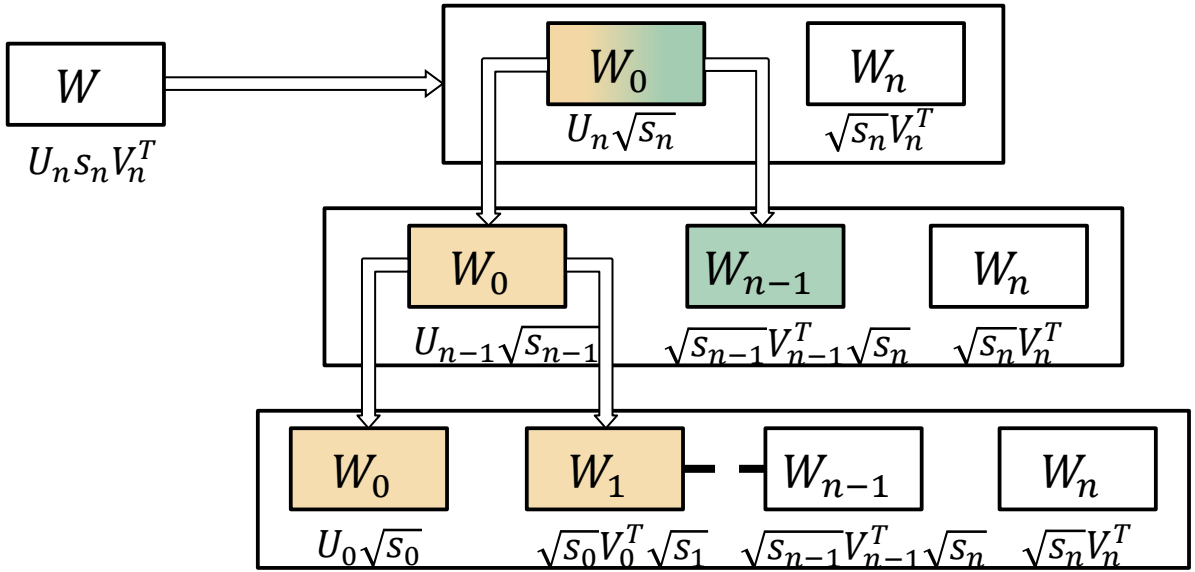
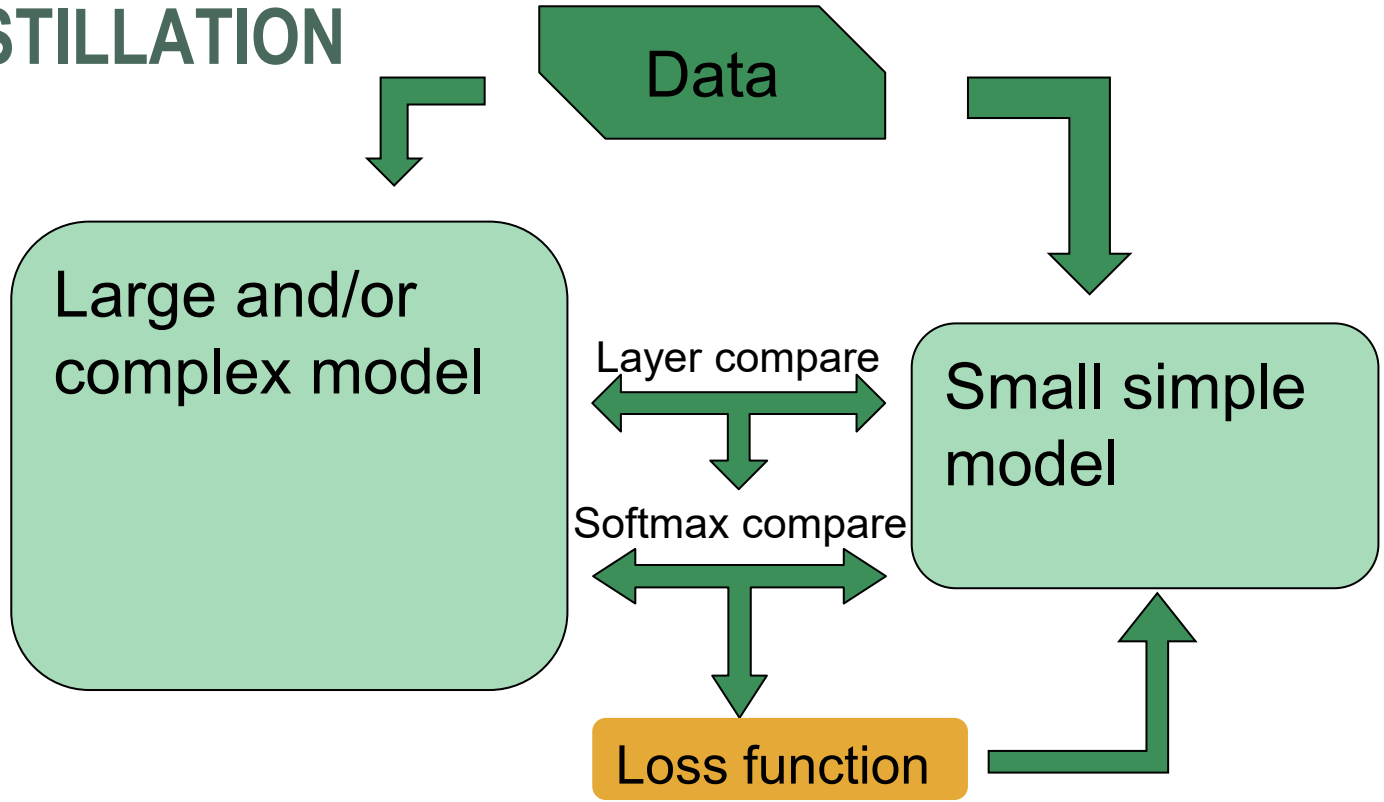


Figure 4: Decomposition of a single network layer into smaller multiple layers.

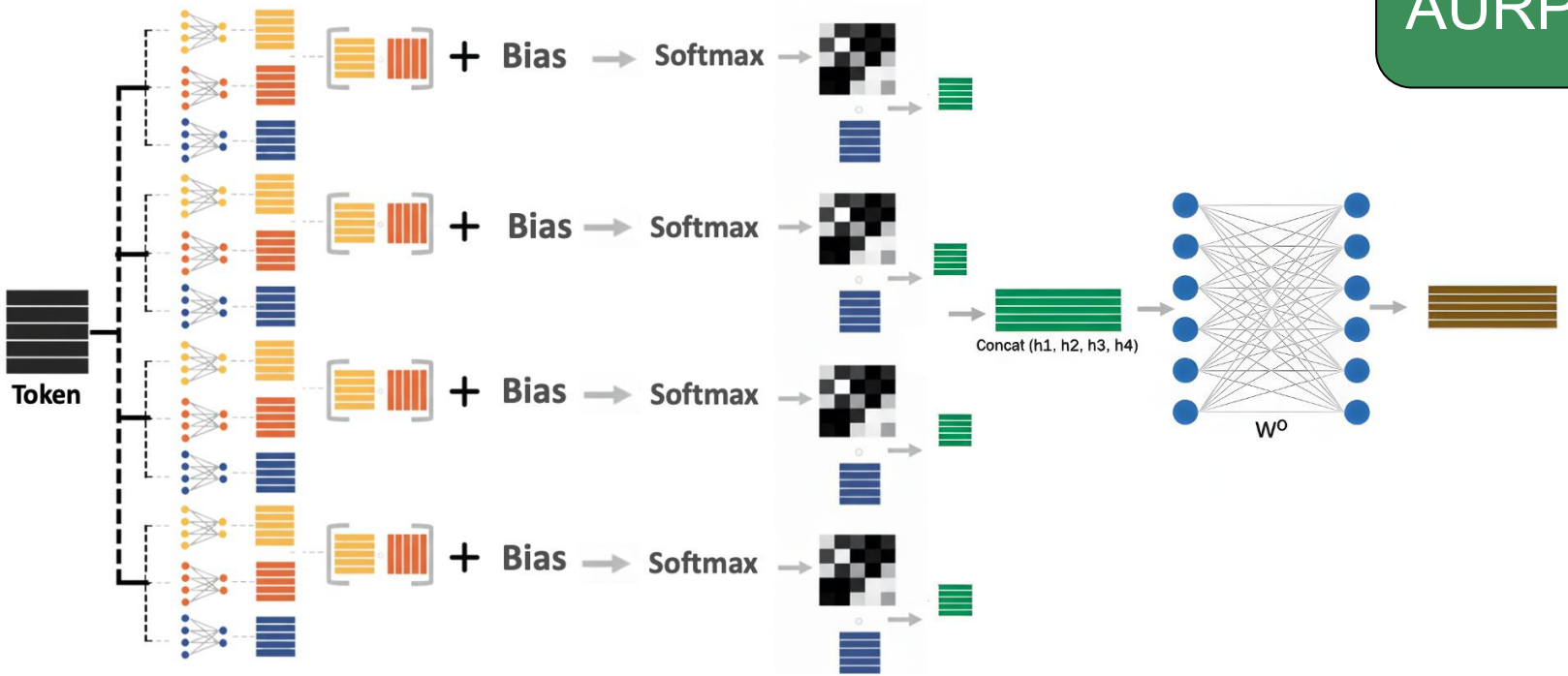
DISTILLATION



https://docs.pytorch.org/tutorials/beginner/knowledge_distillation_tutorial.html

DISTILLATION

0,99
AURPC



DISTILLATION

- Physics informed distillation + Quantization aware training

Performance	**0,99 AURPC**	**No change**
Parameters	2300	157x smaller
Memory footprint	2,3 kB	640x smaller
Latency	151 cycles	
Throughput	1,32 Mevents/s	
Usage	<40 %	

SQUISH

- Prune
  - Connections, Nodes, Layers
  - Inputs – Importance analysis
- Input decorrelation
- Feature extraction/signal processing
- Quantization
  - Weights
  - Inputs
- Matrix factorisation
- Distillation

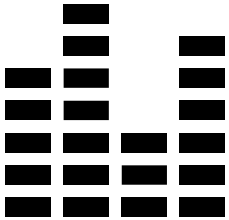
Underlined is model compression

VALIDATION

DESIGN FOR FAILURE

- How should the system fail?

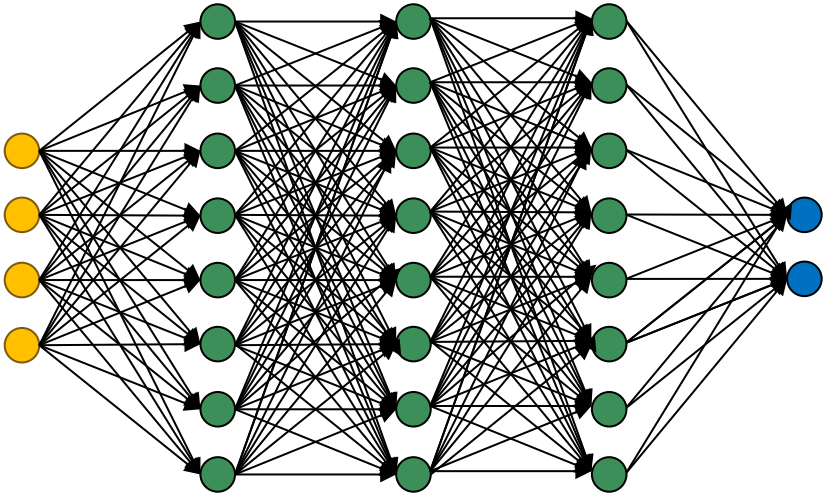
- Prediction uncertainty
- Train failure/unknown case
- Sanity check collection



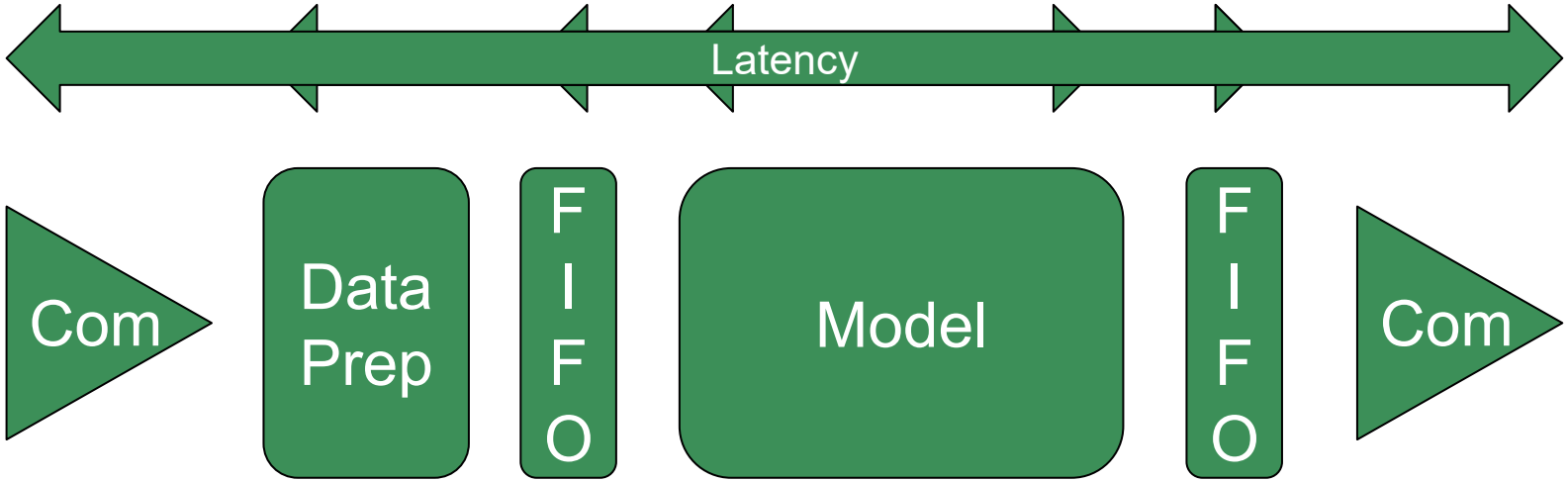
INPUT/OUTPUT VALIDATION

- Design your test set carefully
  - Good inputs, covering the entire application space
  - Extrapolation, outside the scope, but plausible
  - Bad inputs, implausible
  - Noise
  - Detector failure

MODEL STABILITY/SENSITIVITY



LATENCY



POWER

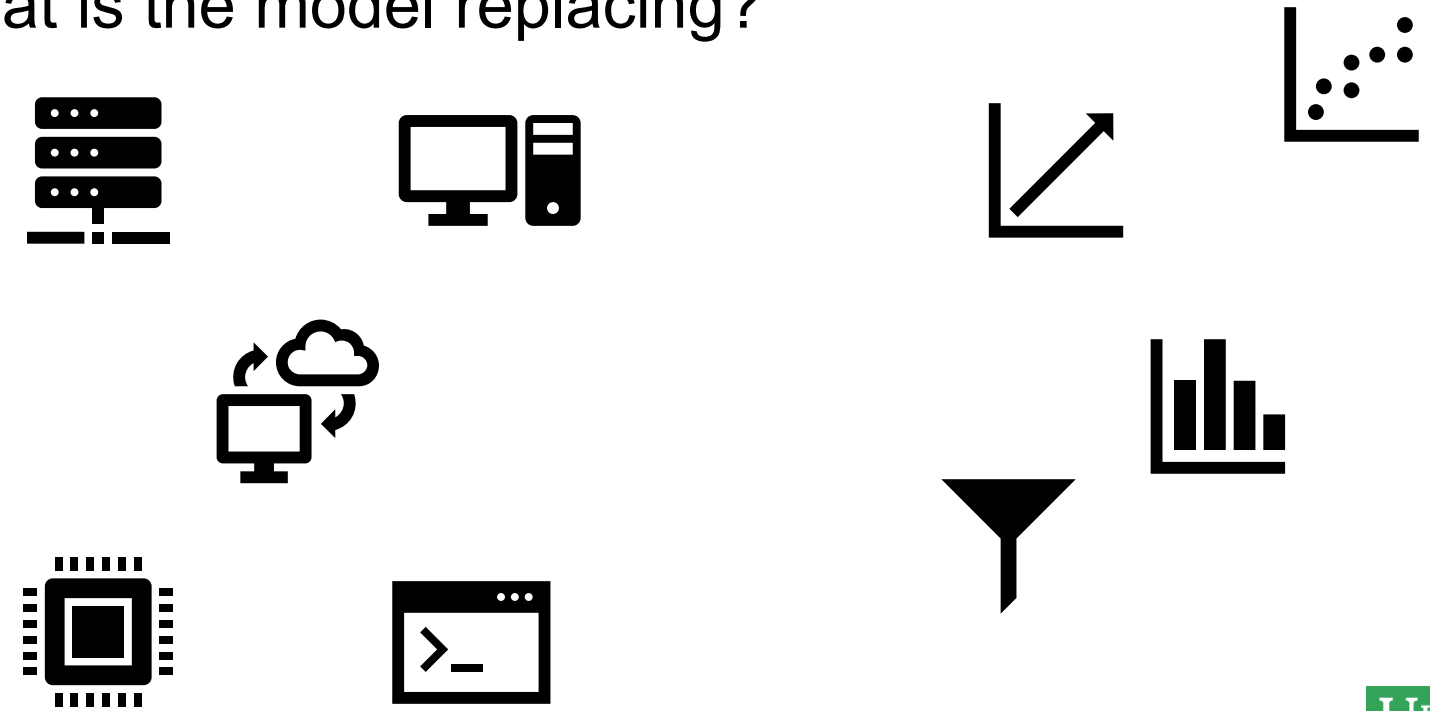
- Challenging to measure

$$\text{Idle} - \text{Full Load} = \text{Avg power}$$

$$\frac{\text{Avg power}}{\text{throughput}} = \text{Energy per inference}$$

COMPARISON

- What is the model replacing?

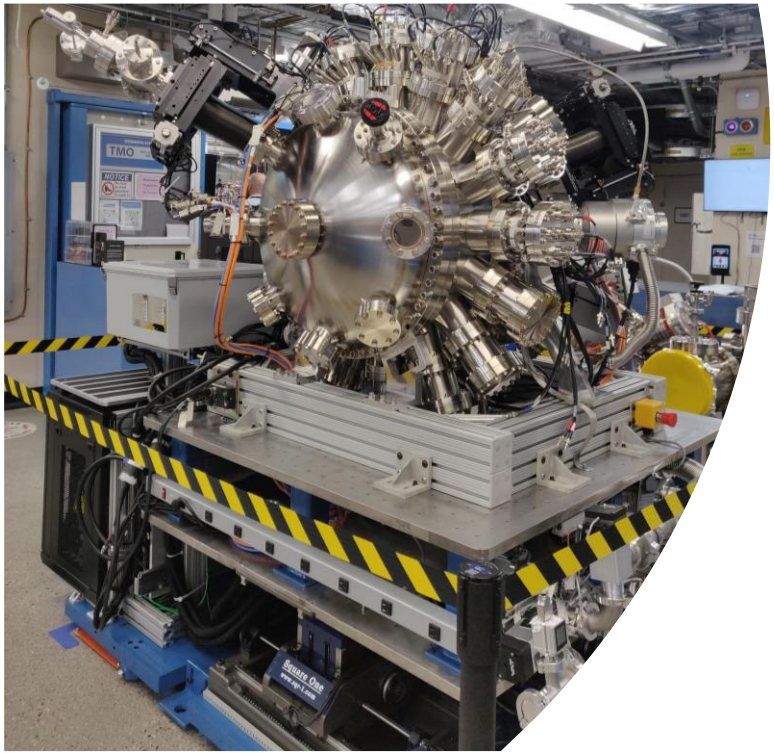


MAINTAINABILITY

- Code readability
- Documentation
- Training data archive
- Update process
  - Functionability update
  - Library compatibility updates
  - Hardware updates

EXAMPLES

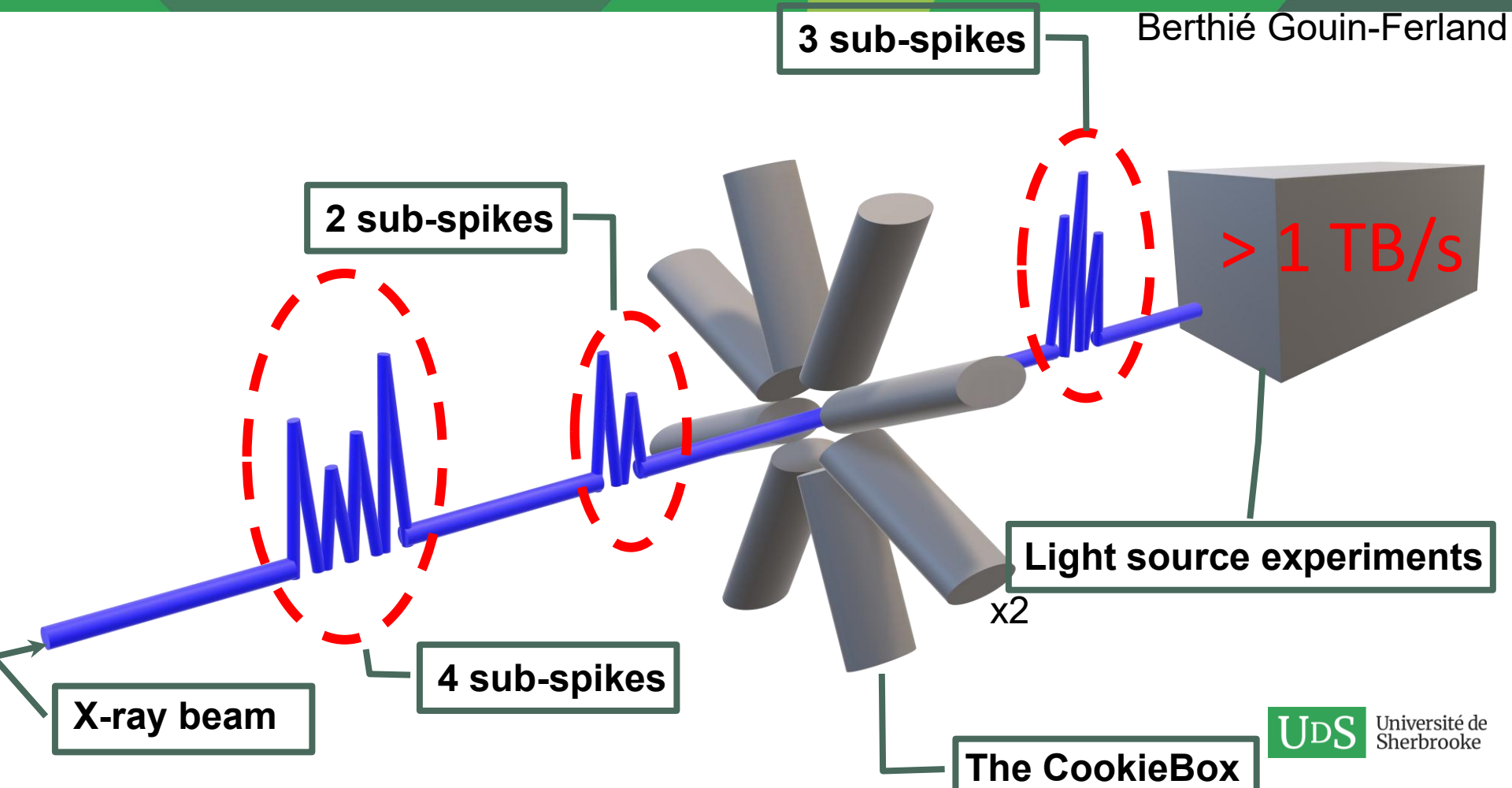
COOKIEBOX



- 16 Time-of-Flight Spectrometers
- 800 GB/s
- 1 MHz rate
- Less than 1 μs to complete analysis

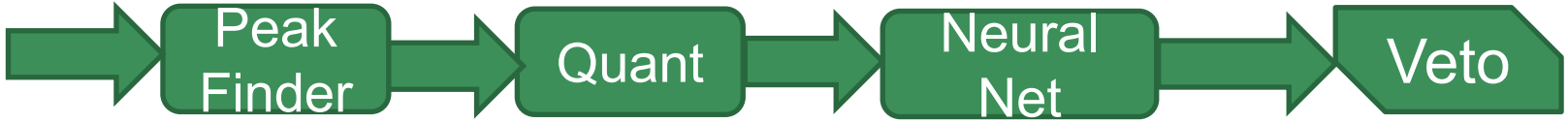
COOKIEBOX - PROBLEM

Berthié Gouin-Ferland

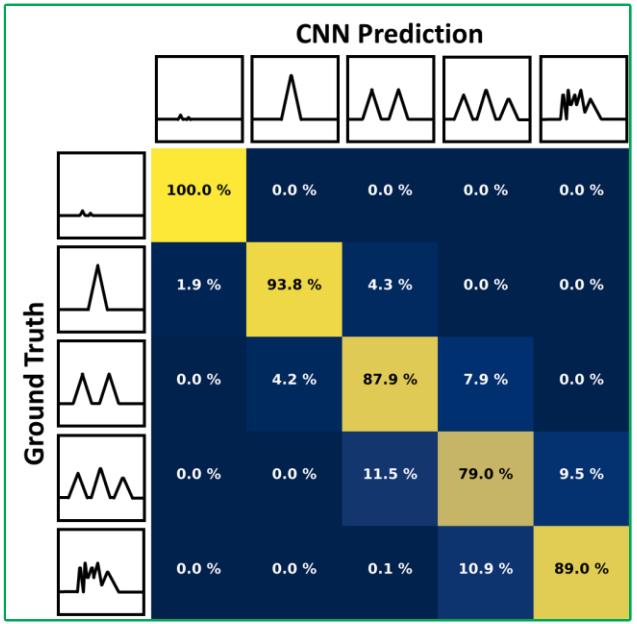


UDS Université de Sherbrooke

COOKIEBOX - ML



Layers
Conv 10x3x3
MaxPool
Conv 20x3x3
MaxPool
Conv 30x3x3
MaxPool
Fully connected 5
Fully connected 5
Softmax classification



Berthié Gouin-Ferland

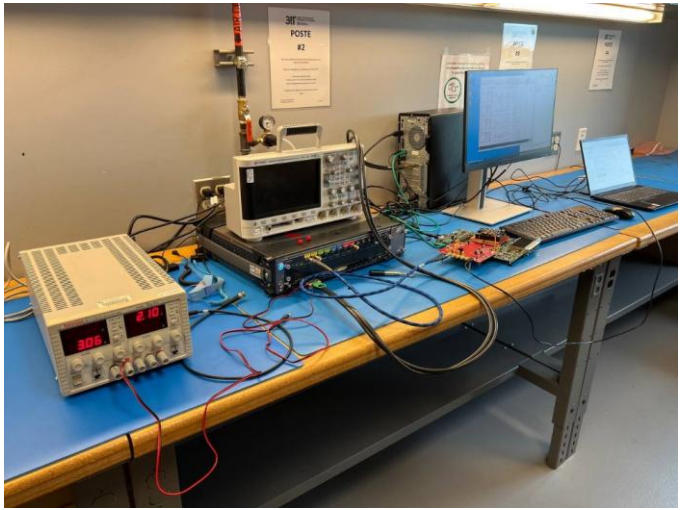
COOKIEBOX – LAB MEASUREMENTS

Berthié Gouin-Ferland Quentin Wingeiring Mehdi Rahimifar

Signal generator

Digitizer

FPGA



Complete Analysis – 400 ns

AI Module – 200 ns

CONCLUSION

CONCLUSION

Neural Networks and Decision Trees are compatible with hardware

Co-design is necessary to optimise the system

Squishing the model helps its overall performance

Validation and maintainbility are often overlooked!

Edge and distributed ML is a great tool to reduce resource usage

TOOLS

SOME TOOLS

- Hls4ml
- Coyote
- SNAC-Pack
- Conifer - https://github.com/thesps/conifer
- SNL – to be announced - https://arxiv.org/abs/2508.21739
- da4ml - https://arxiv.org/abs/2507.04535
- NeuraLUT - https://arxiv.org/abs/2504.00592

THANK YOU!