tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

June 26 - 28, 2023

TINY ML

www.tinyML.org

TINY ML

A novel mechanism for edge ML

Steve Furber CBE FRS FREng

ICL Professor of Computer Engineering

The University of Manchester

MANCHESTER 1824

The University of Manchester

1. Unsustainable energy use

What's wrong with AI?

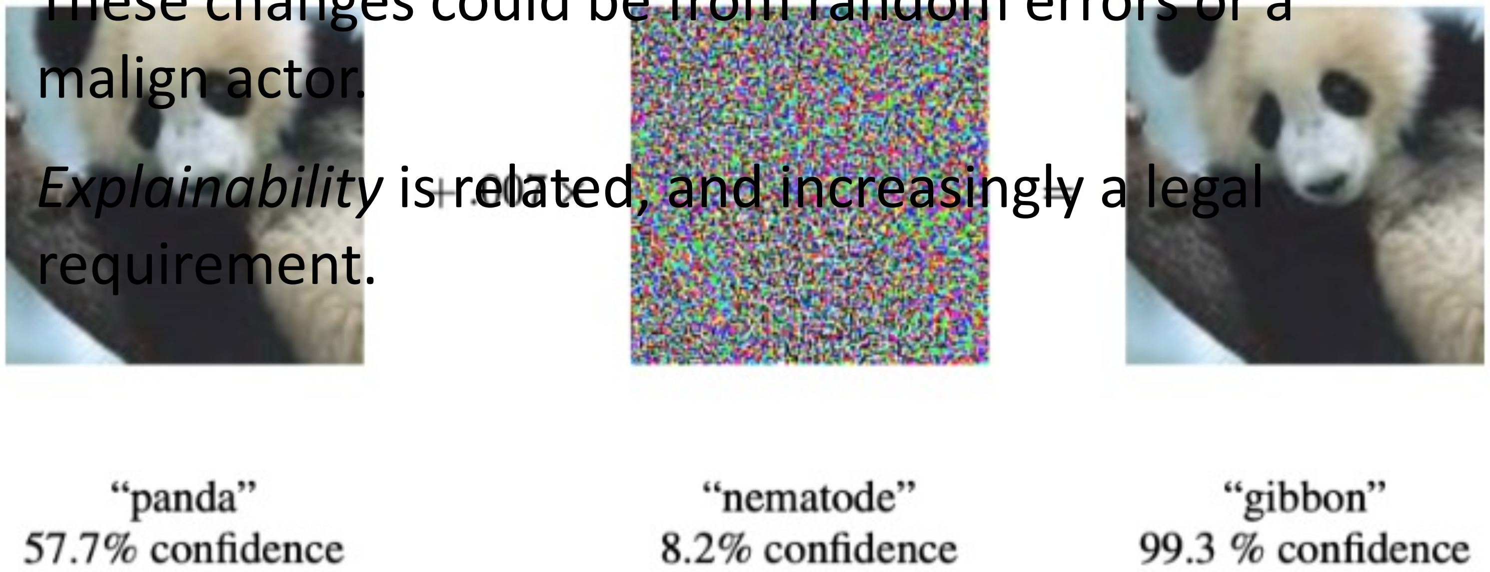
- Energy (and CO_2) needed to train latest ANNs already huge. Power stations built near the server farms.
- Cannot continue growing super-exponentially or planet will be covered with memory chips with no electricity left for anything else!
- Autonomous EVs require AI onboard using ~1W instead of current ≈100W. IoT/edge devices need to be far more economical still.
- Green legislation will make this waste unviable not just very expensive (e.g. 21 days on 2,000 GPUs for LLaMa = $4M). GPT-4 is 1000x bigger than LLaMa…

Human Brain Project

What's wrong with AI?

2. Brittle, easily fooled & hard to explain

- Vast parameter sets (100 trillion+ for GPT-4) used for current ANNs 'over-fit' the data.
- Tiny changes (e.g. to an image) can fool the system and make classification incorrect – with very high confidence.
- These changes could be from random errors or a malign actor.
- *Explainability* is related, and increasingly a legal requirement.

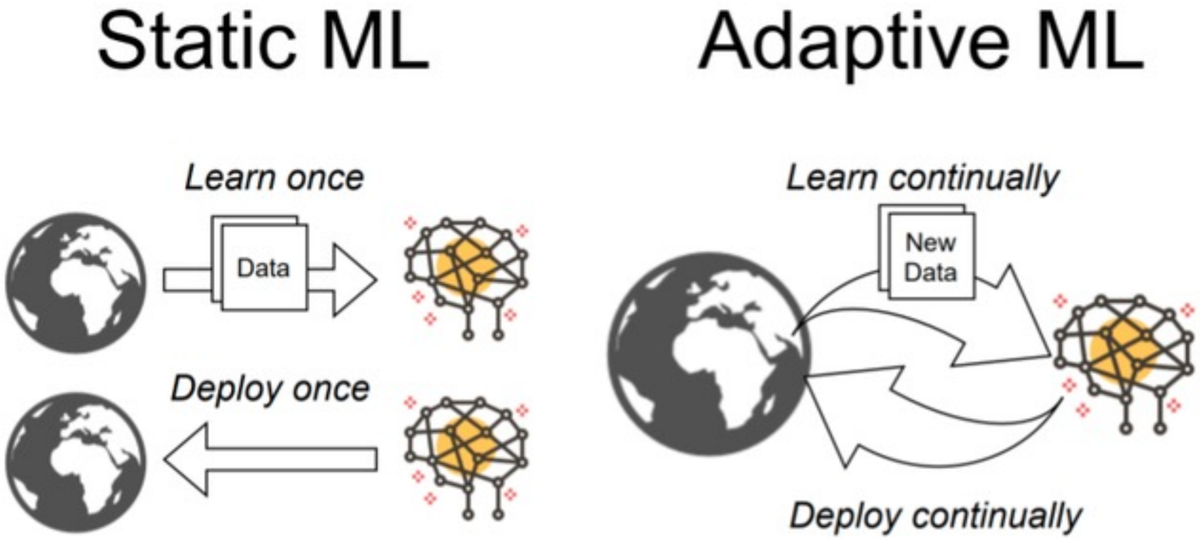


Human Brain Project

What’s wrong with AI?

3. Static & inflexible – soon outdated

- Current ANNs are trained once on fixed training set at huge cost, and the learning is frozen.
- Real world is not like that - everything is constantly changing both within and outside the system.
- Hence the growing number of research projects setting out to address this issue.

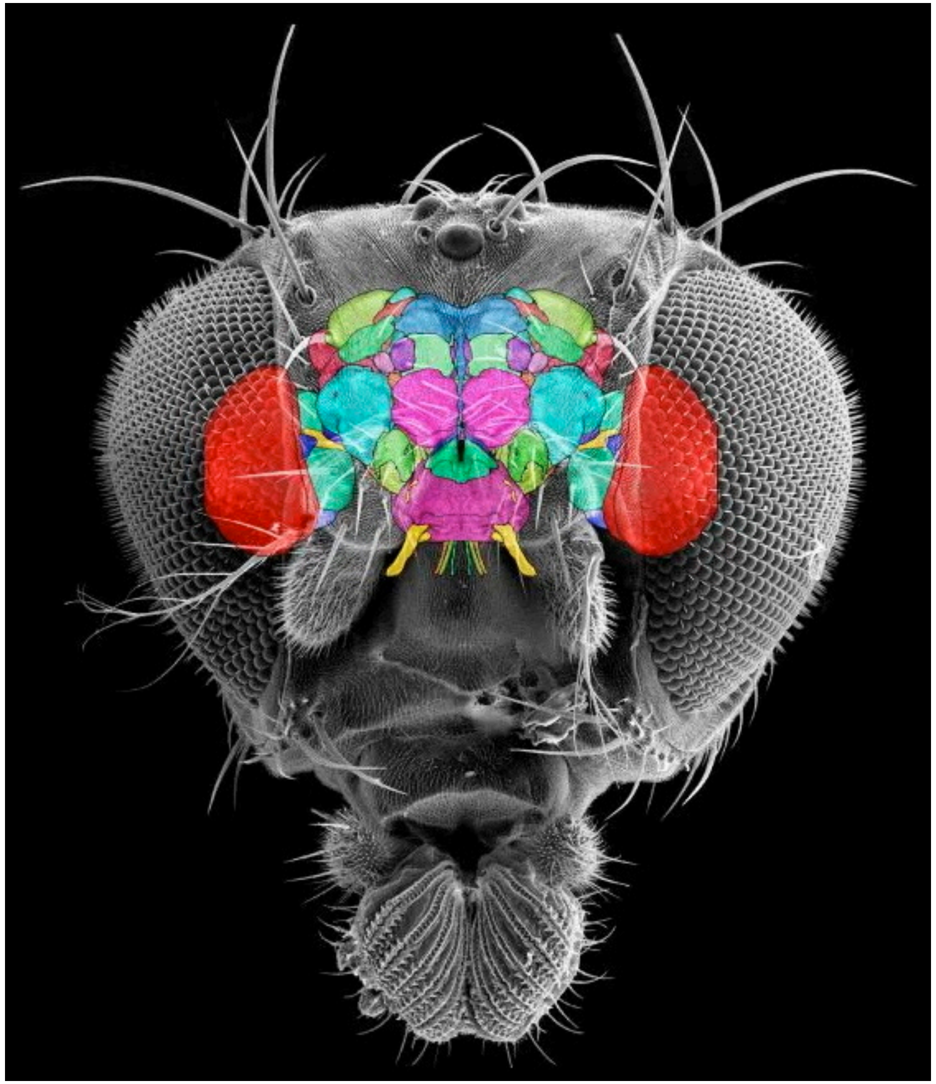


Human Brain Project

Biological systems do things differently

- Learn continuously & adapt to changes
- Use tiny amounts of energy
- Performance degrades gracefully & safely

…and so we learn from biology and create *BitBrain* to address these concerns



(image by Kei Ito *et al*)

So what's all this *BitBrain* stuff about?

An innovative working mechanism (the *SBC memory*) and surrounding infrastructure (*BitBrain*) based upon a novel synthesis of ideas from sparse coding, computational neuroscience and information theory.

Patent GB 2113341.8 was filed at the UKIPO on 17th September 2021.

- **Single-pass supervised learning** avoiding expensive computatiuons.
- **Accurate inference** that is **very robust against imperfect inputs**.
- C**ontinuous and adaptive** learning.
- **Fast and low-energy** operation on conventional & neuromorphic processors.
- Implemented on **Raspberry Pi** and **SpiNNaker**.

Hopkins, M., Fil, J., Jones, E. G. & Furber, S. (2023); ***BitBrain* and Sparse Binary Coincidence (SBC) memories: Fast, robust learning and inference for neuromorphic architectures**. *Frontiers in Neuroinformatics*

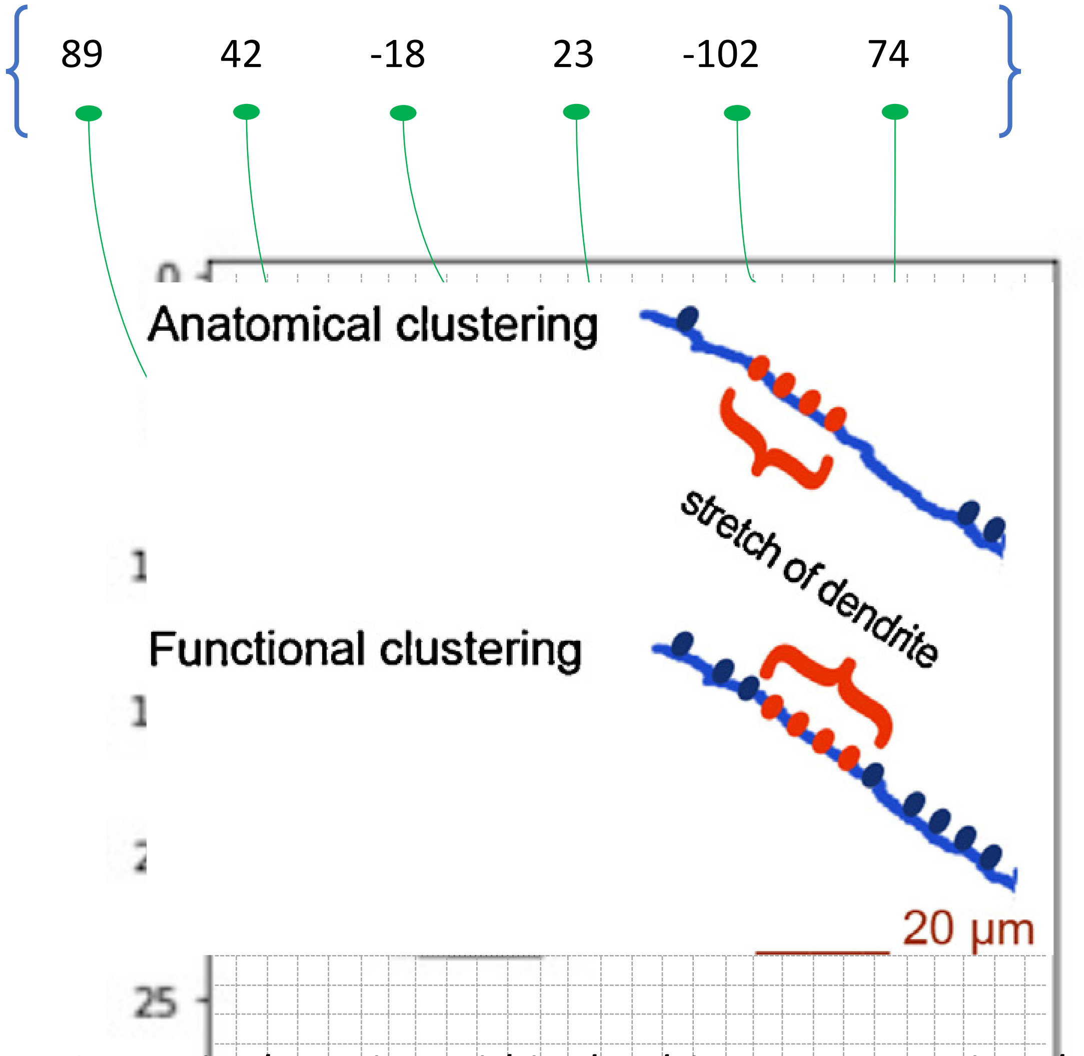
Address Decoder Elements (ADEs)

Each ADE samples a small subset of the input data, like a synaptic cluster.

An example ADE which contains multiple synapses with individual weights which can signify strength and/or longevity of connection.

The input stream can be any objects or data which are able to be coded as a vector of bits or any other scalar values i.e. almost anything! Here using a 784-vector of 8-bit values to represent a greyscale raster image.

ADE fires when the sum of the connected input values multiplied by their respective synaptic weights within an ADE reaches a threshold - which is learned homeostatically.



From *Synaptic clustering within dendrites: An emerging theory of memory formation.* Kastellakis et al (2015)

Address Decoders (ADs) accessing a 2D SBC memory

This is a 2D memory; 3D and higher (using more ADs) are also of interest.

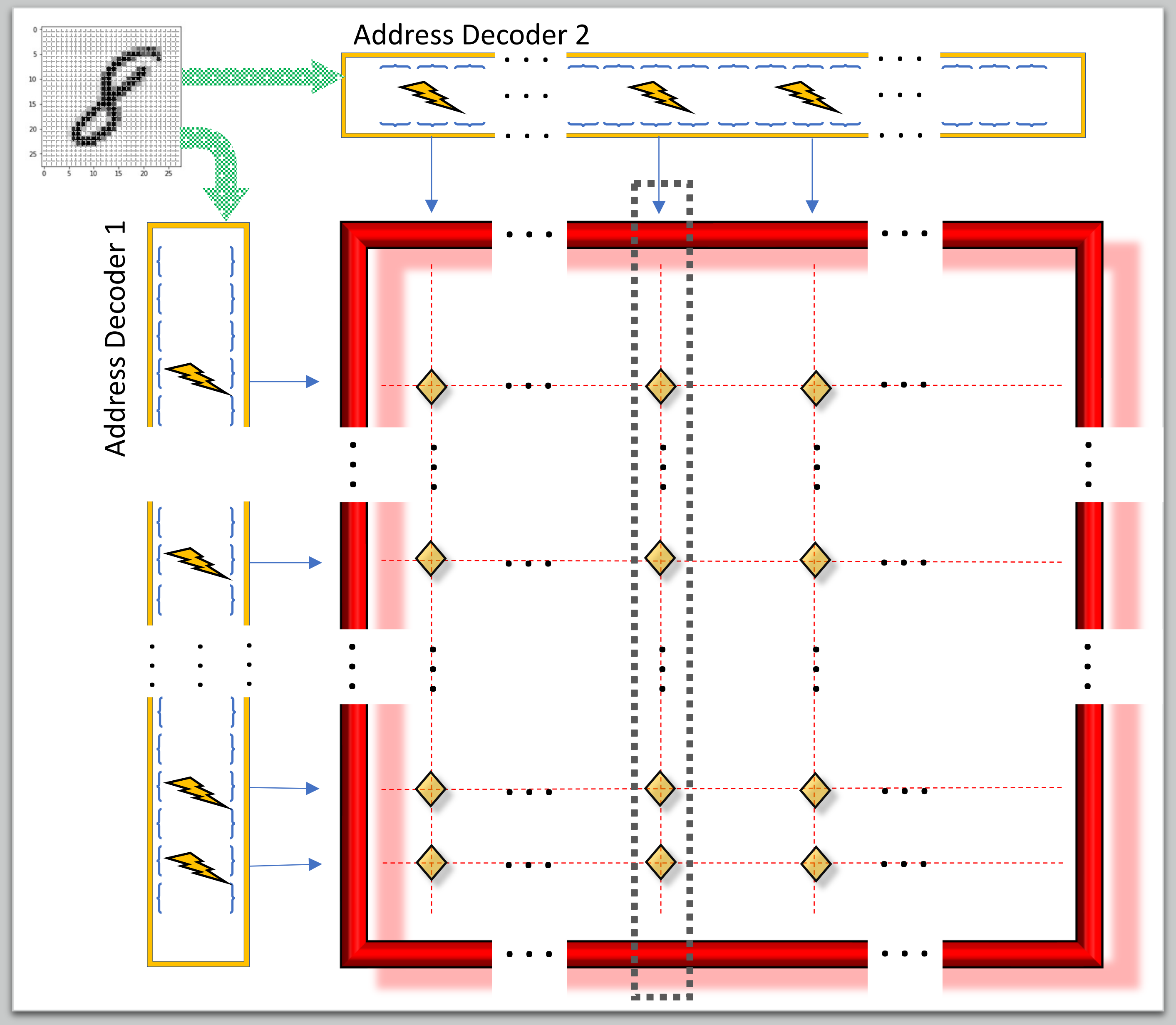
Widths of SBC match the length of ADs e.g. 1,024 elements.

depth = *f*(# classes)

◇ <- activated memory position

Activation pattern is *sparse* i.e. only a small percentage of the ADEs in each AD will fire for any given input.

Each *coincidence* of active ADEs between ADs activates a memory location that reads or writes information about the class which has activated it.



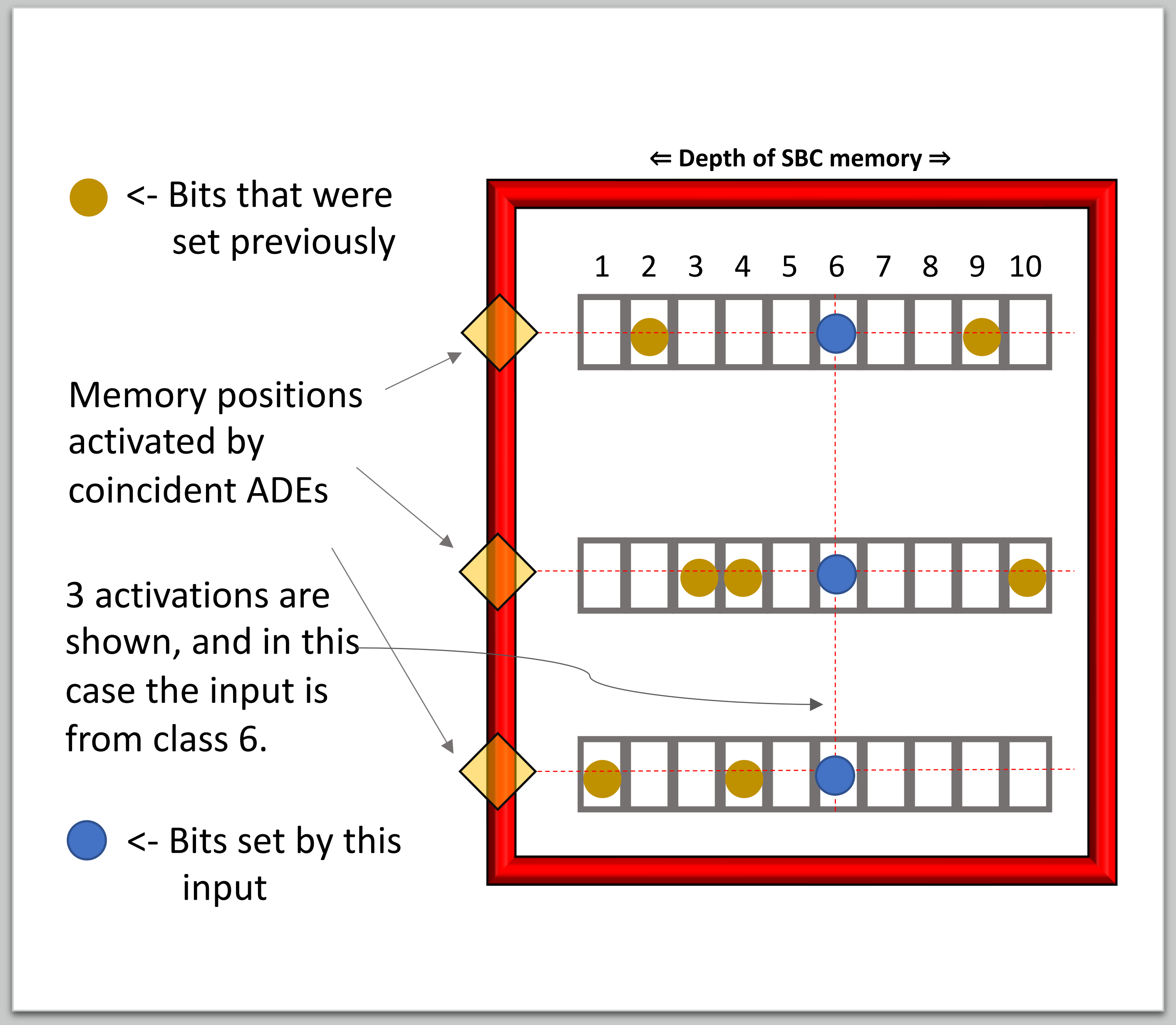
Class information held within SBC memory

'Side view' of SBC memory, showing 'depth' which varies with number of classes in problem. In this case, 10 classes with 'one-hot' coding meaning 10 bit cells per memory position.

Writing to the SBC: go to all activated memory positions & set the relevant class bits if they are not already set.

Reading from the SBC: count bits set over all activated memory positions & choose class with the highest sum.

Assumes 'one-hot' encoding. If classes are coded differently then another encoding/decoding process required.

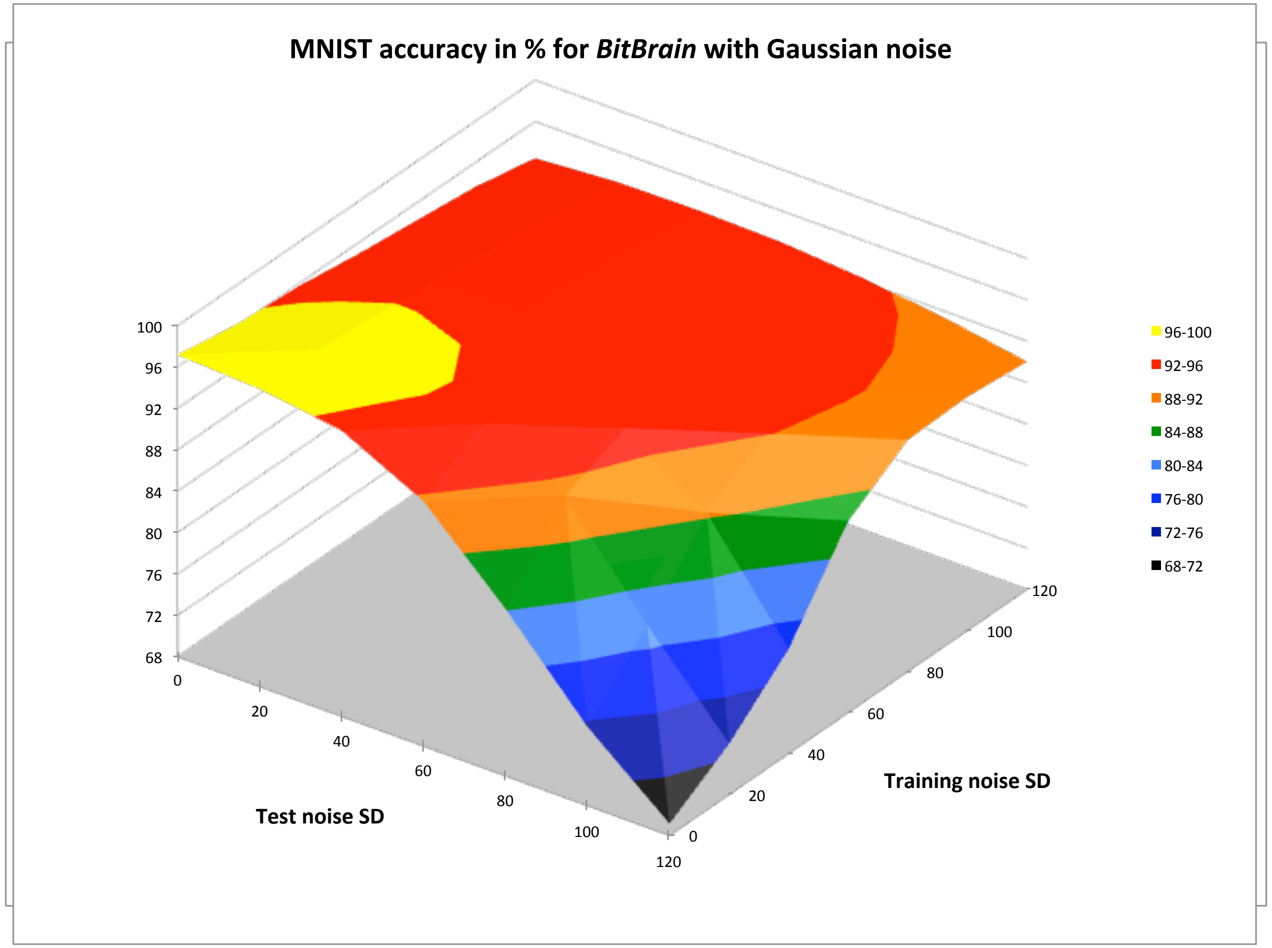


Basic results from MNIST
(10 classes balanced)

- AD lengths = 2,048
- 4x ADs with { 6, 8, 10, 12 } widths of synapses
- AD target firing ~1% per input
- synapses spatially clustered and then structural plasticity used to home in on features

10x 2D SBCs:

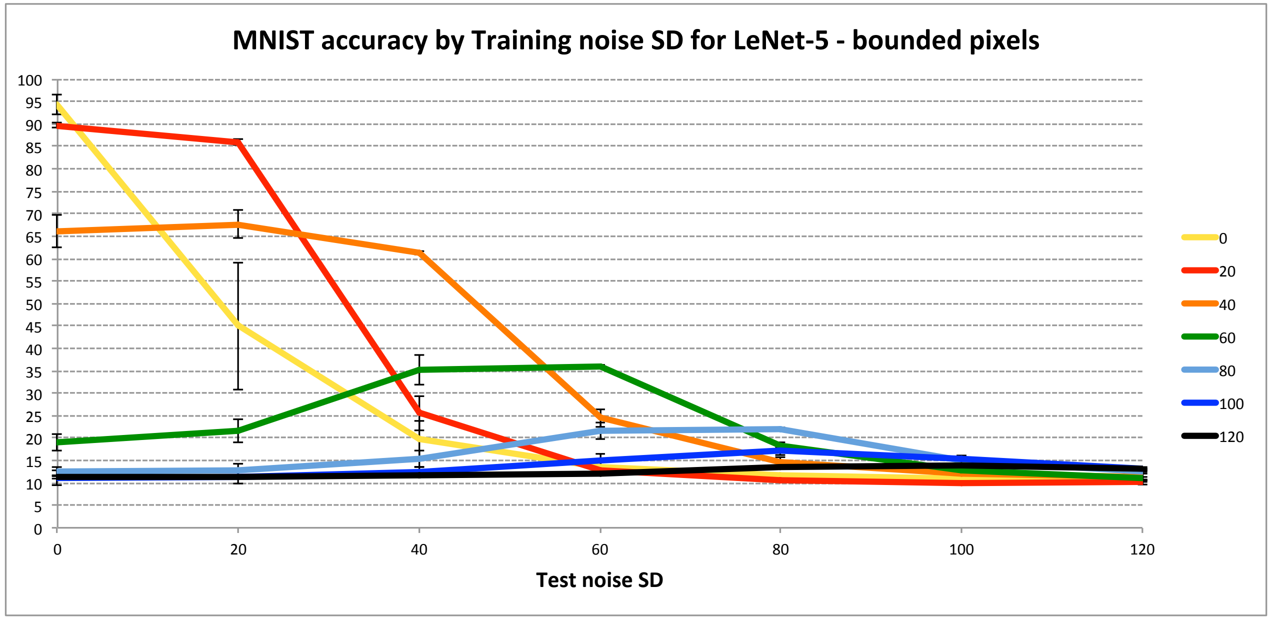
- 6x full-size between ADs
- 4x half-size within ADs
- 42MB memory for full occupancy



Human Brain Project

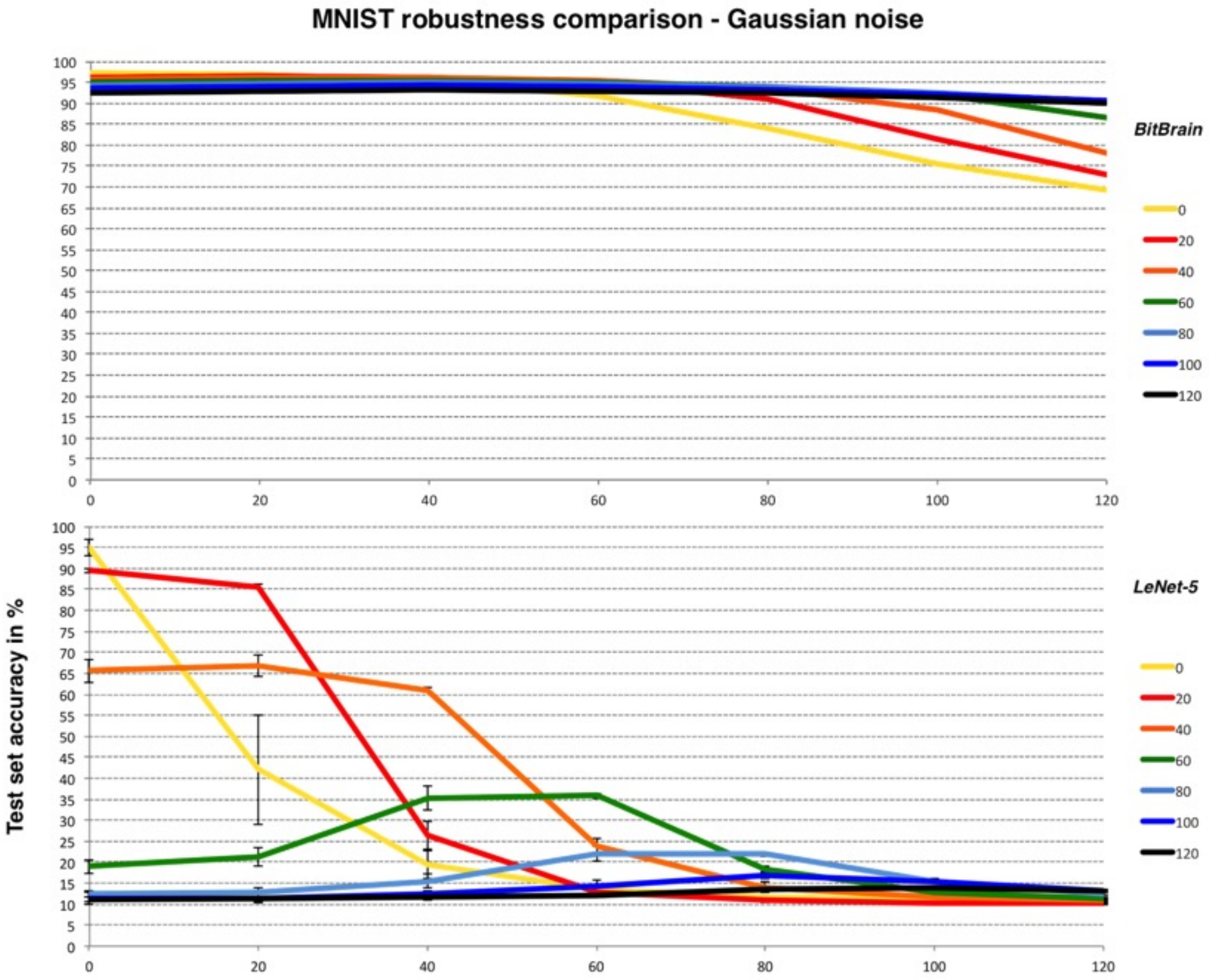
MNIST with *LeNet-5*

- Early but respected CNN designed for character recognition: https://en.wikipedia.org/wiki/LeNet
- max 100 epochs
- early stopping, 'patience' = 5
- sigmoidal activations
- 'static' noise

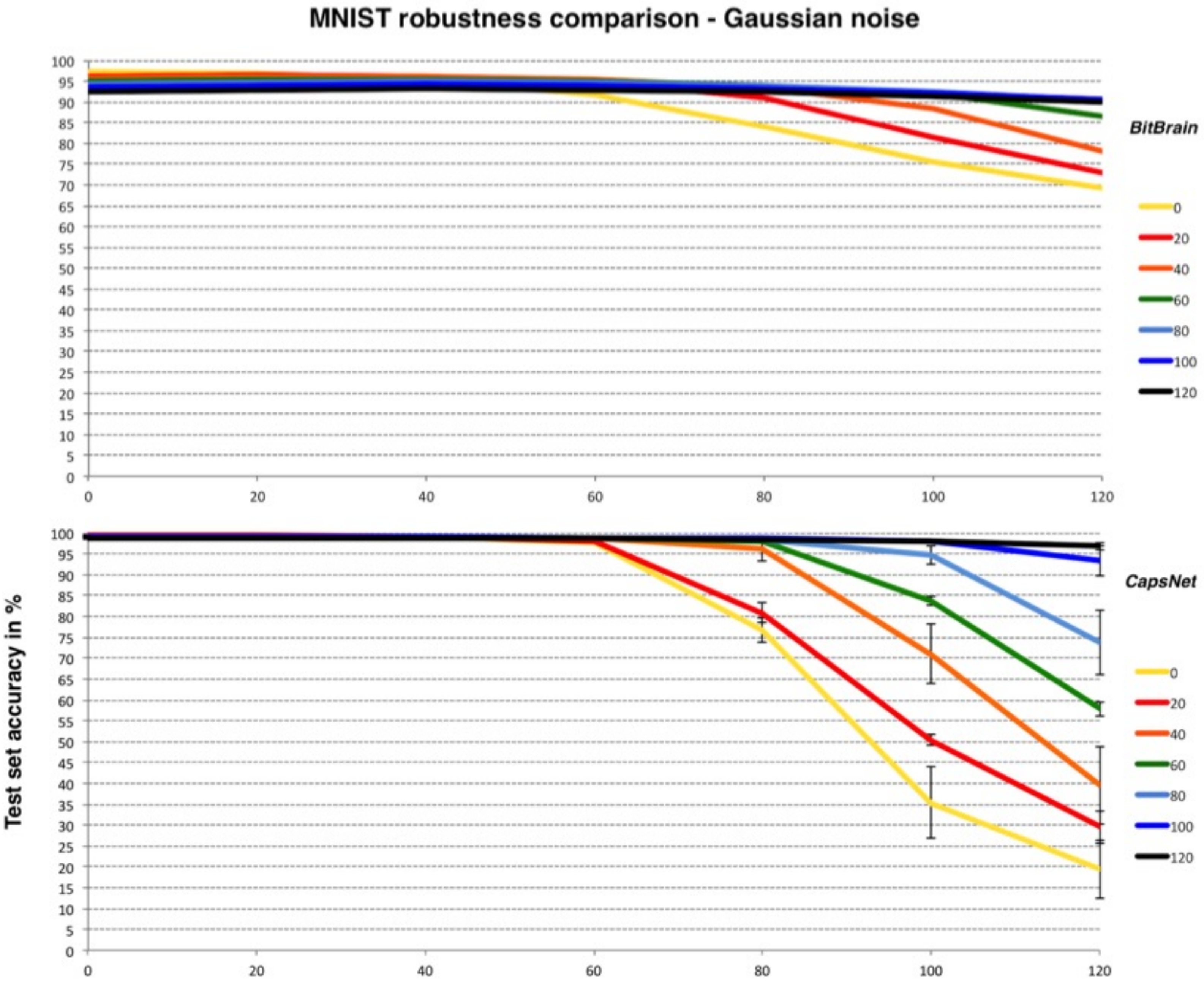


Human Brain Project

MNIST robustness comparison *BitBrain* vs *LeNet-5*



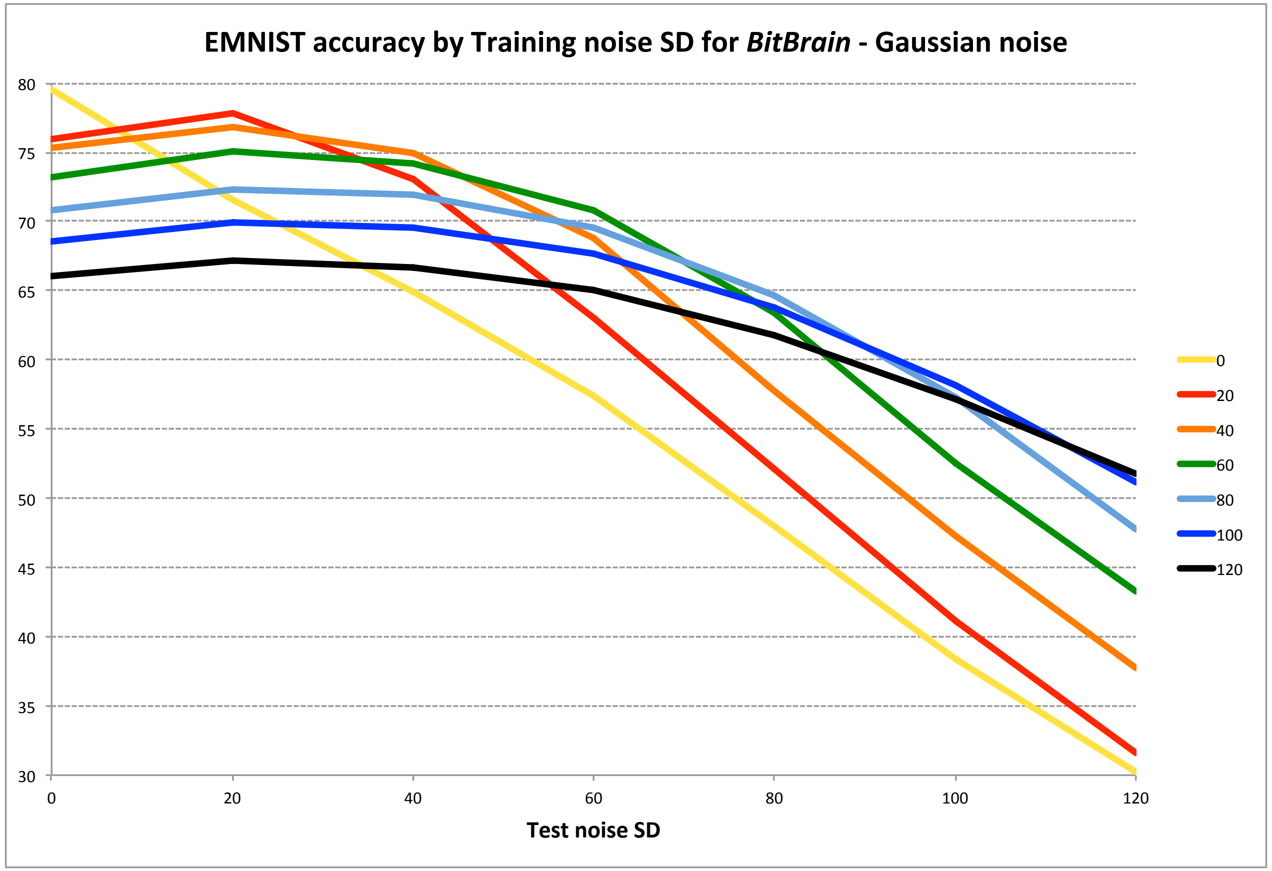
MNIST robustness comparison *BitBrain* vs *CapsNet*



Human Brain Project

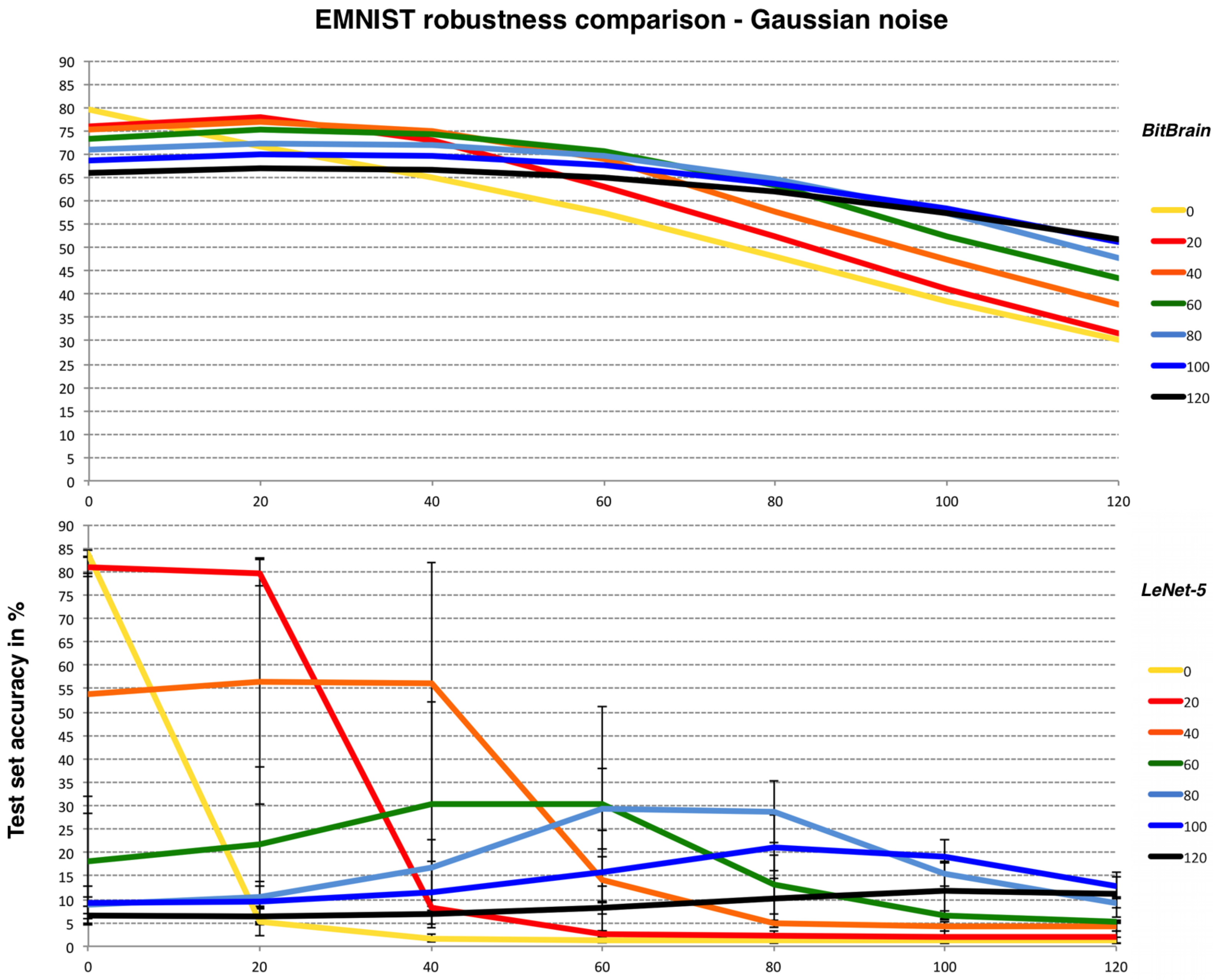
Basic results from EMNIST
(62 classes unbalanced)

- *BitBrain* setup identical to MNIST
- much harder problem
- very unbalanced
- natural class aliasing:
  { o, O, 0 }, { i, I, l, 1 }, { s, S, 5 }, { B, 8 }

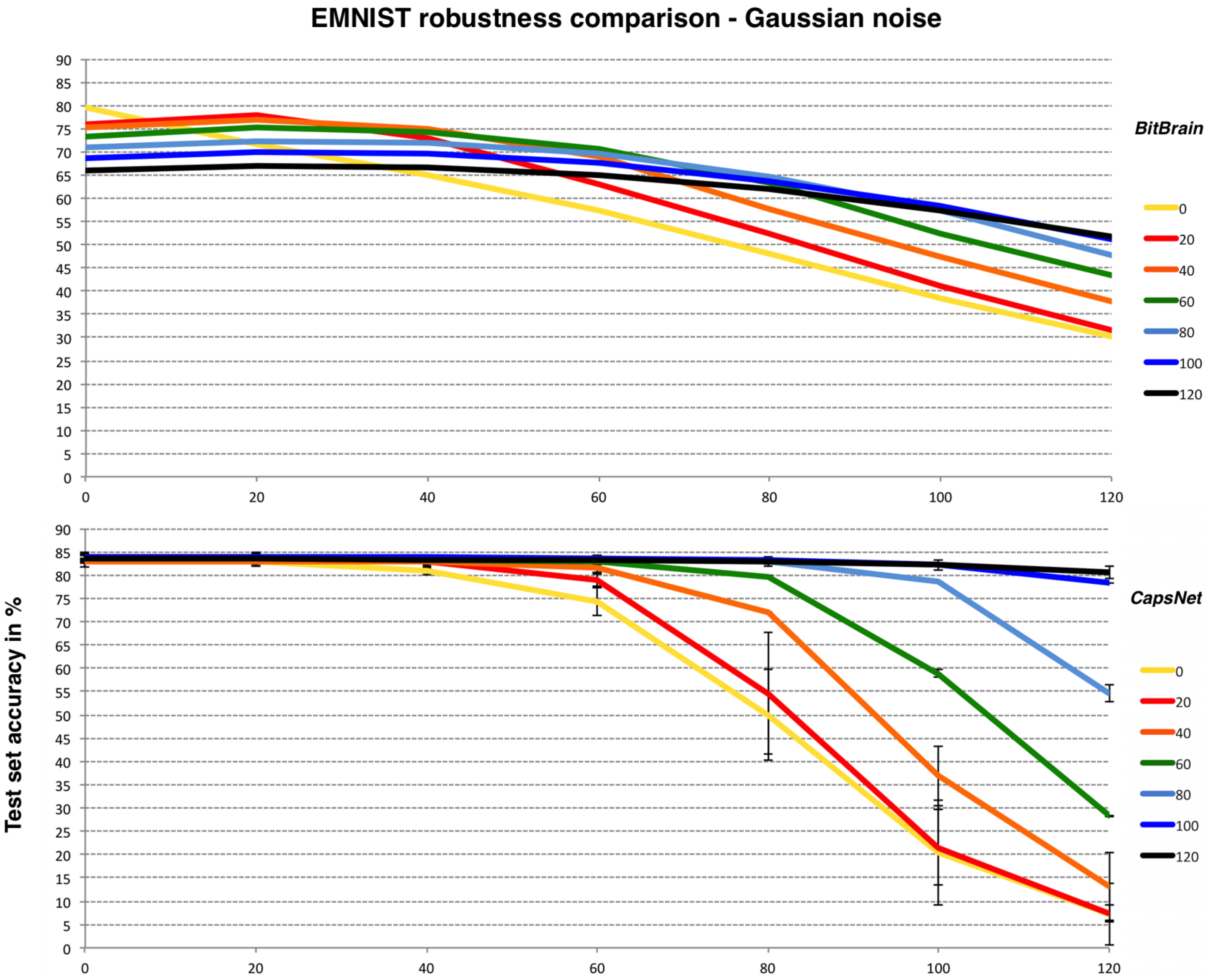


Human Brain Project

EMNIST robustness comparison *BitBrain* vs *LeNet-5*



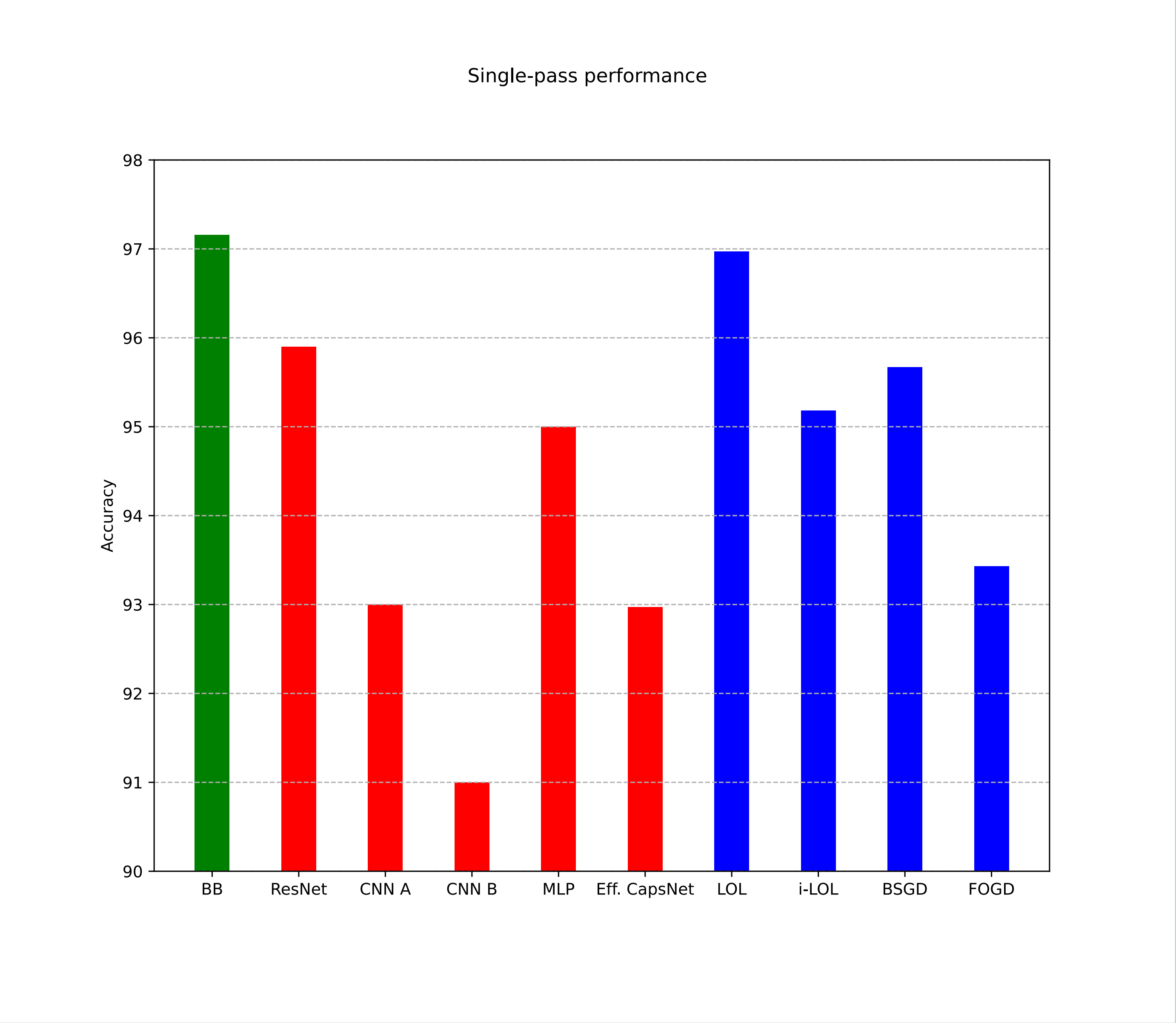
EMNIST robustness comparison *BitBrain* vs *CapsNet*



Human Brain Project

MNIST comparison with other single-pass methods - 1

- Red bars are CNNs trained for only one epoch
- Blue bars are specifically designed single-pass classification methods
- References for all methods are in our upcoming paper



Human Brain Project

MNIST comparison with other single-pass methods - 2

BitBrain compared with single-pass SVM methods for two-class problems.
- best results are in **bold**.

	libSVM	*Perceptron*	*Pegasos 1*	*Pegasos 20*	*LASVM*	*StreamSVM 1*	*StreamSVM 2*	*BitBrain*
0 vs *1*	99.52	99.47	95.06	99.48	98.82	99.34	99.71	**99.95**
8 vs *9*	96.57	95.90	69.41	90.62	90.32	84.75	94.70	**98.49**

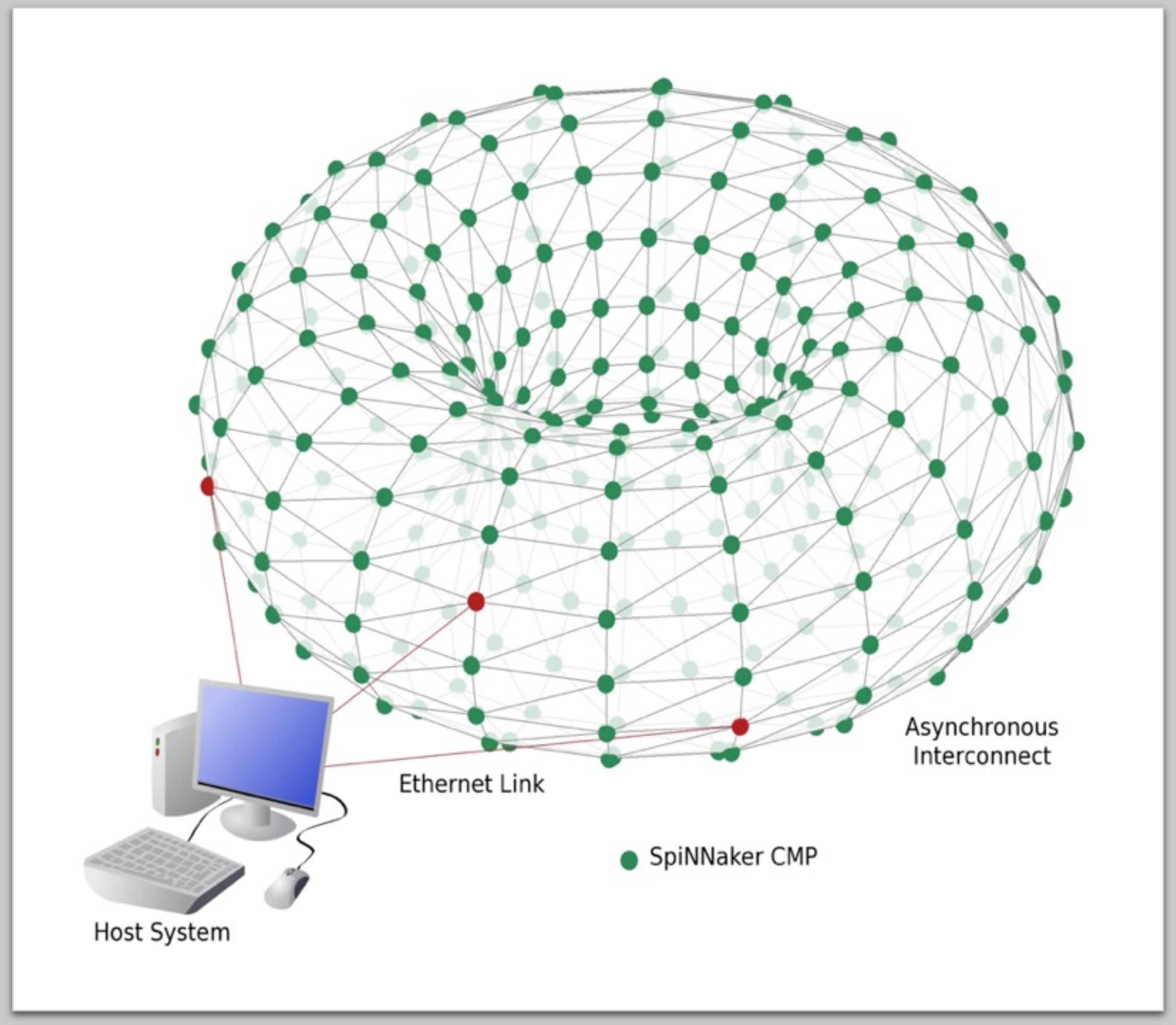
SVM results from: Piyush Rai, Hal III, and Suresh Venkatasubramanian. *Streamed learning: One-pass SVMs.*
IJCAI International Joint Conference on Artificial Intelligence, 2009.

Human Brain Project


Activities
Tk
May 26 14:12
BitBrain MNIST LIVE demo
PREDICT
PREDICT SPINN
Add noise
Center img.
TRAIN
Draw a number
CLEAR
EXIT
Normal. SBCs
SpiNN-3

SpiNNaker

- A million ARM processors in one computer
- Able to model about 1% of the human brain…
- …or 10 mice!

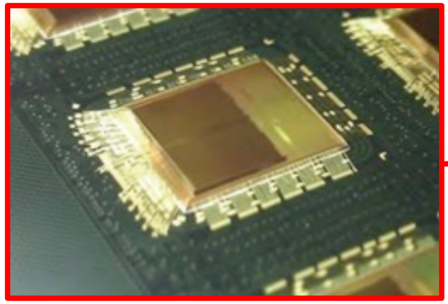


Human Brain Project

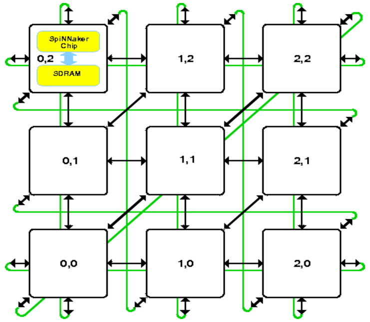
SpiNNaker
Biologically
Inspired
Massively
Parallel
Architectures

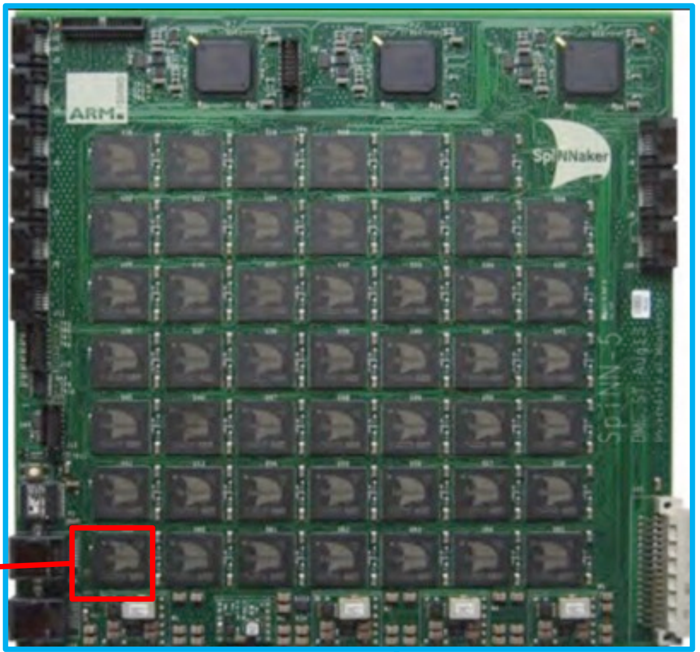
SpiNNaker machines

SpiNNaker chip
(18 ARM cores)



SpiNNaker board
(864 ARM cores)





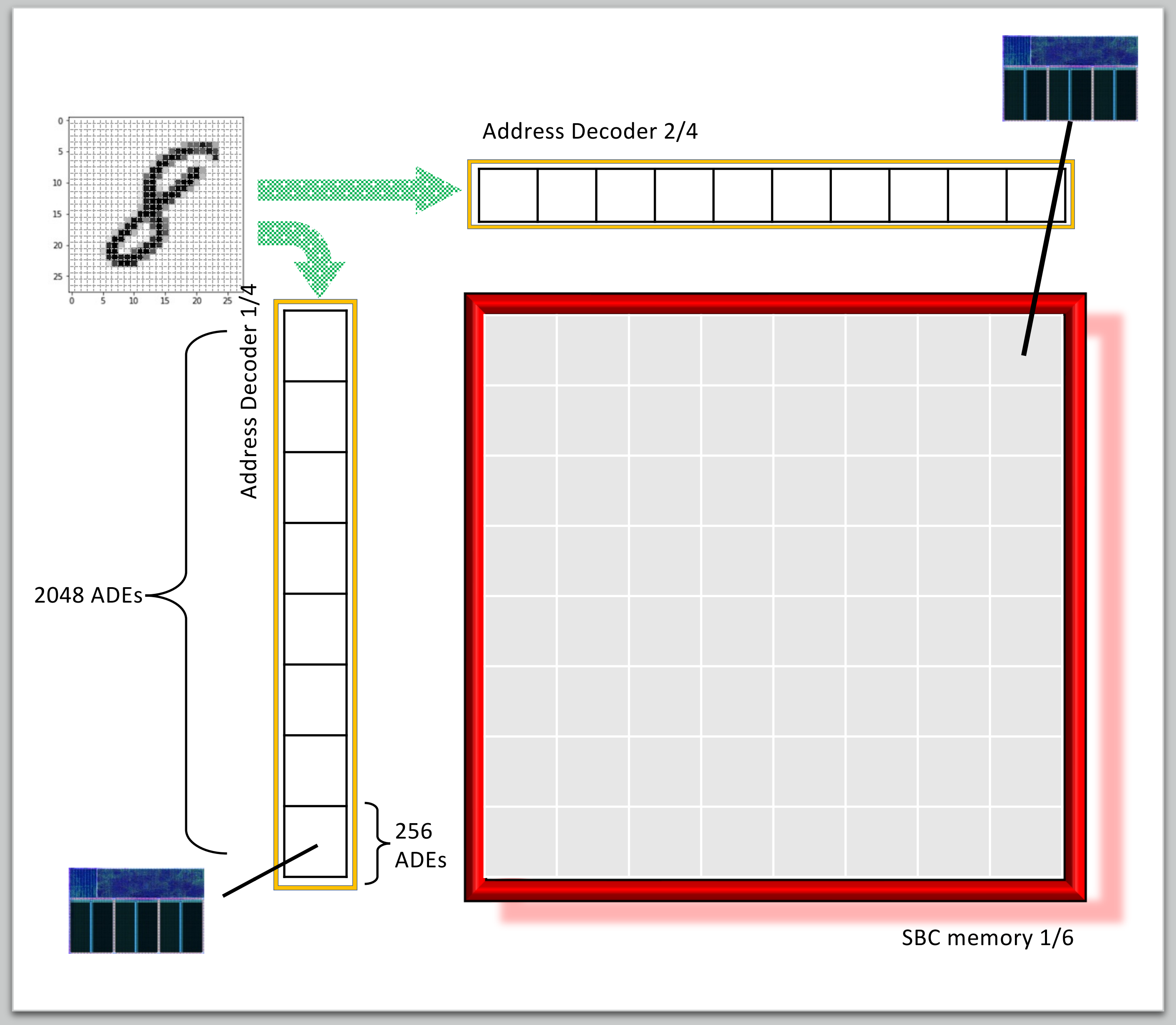
- HBP platform
  - 1M cores
  - 11 cabinets
    (including server)
- Launch 30 March 2016
  - then 500k cores
  - ~450 remote users
  - 5M SpiNNaker jobs run

Human Brain Project

SpiNNaker
Biologically
Inspired
Massively
Parallel
Architectures

Distributed implementation of *BitBrain* on *SpiNNaker*

- *BitBrain* was designed to be compatible with conventional CPUs, but also with energy-efficient, distributed computers, such as SpiNNaker.
- The algorithm can be spread among arbitrarily many cores on SpiNNaker, and thus can make full use of its inherent parallelism.



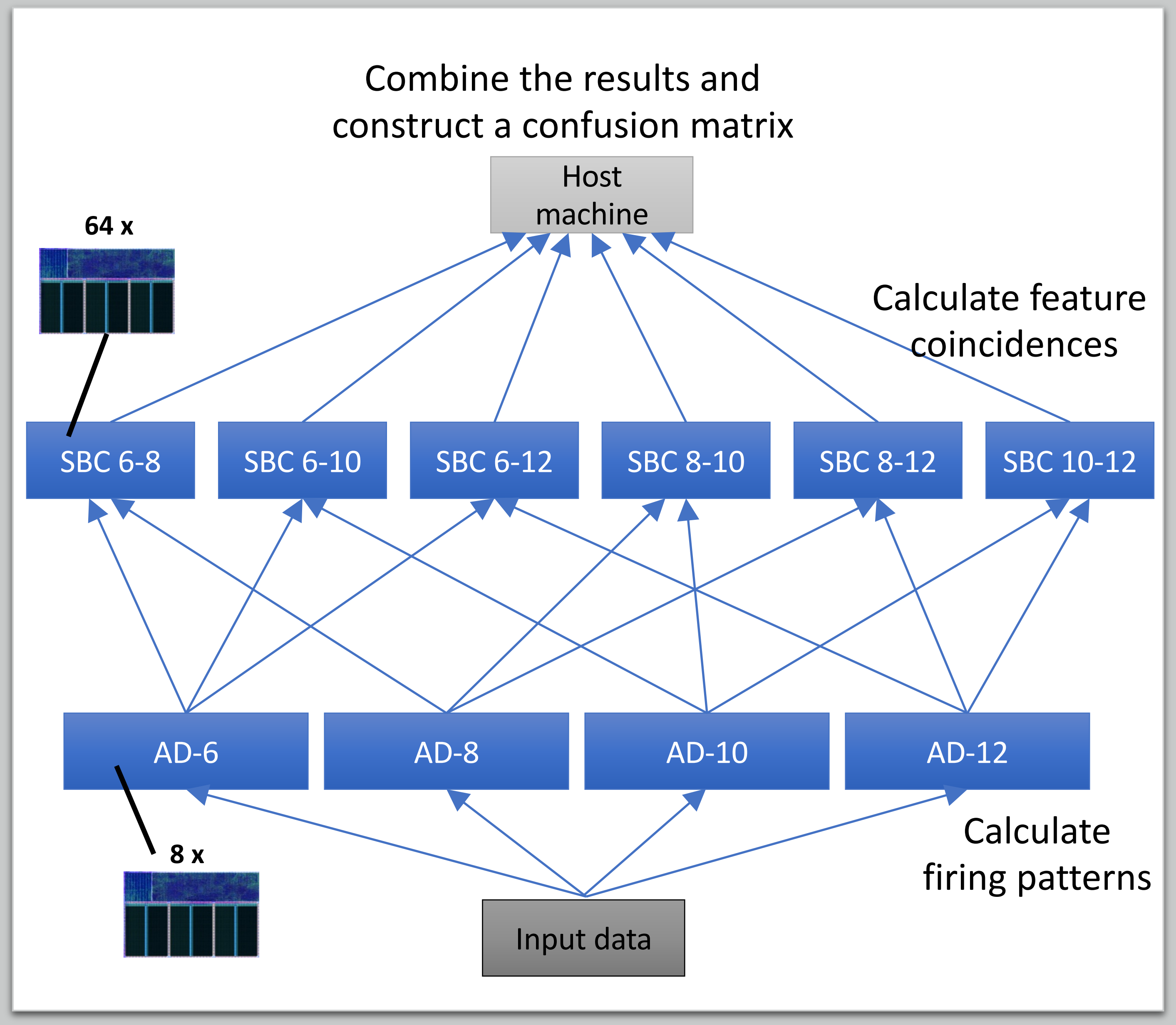
Human Brain Project

SpiNNaker
Biologically Inspired Massively Parallel Architectures

Distributed implementation of BitBrain on SpiNNaker

Example implementation:

- 4 x 8 = 32 ADE cores
- 6 x 64 = 384 SBC cores




Activities
Tk
May 26 16:31
BitBrain MNIST LIVE demo
0.04
0.02
0.00
−0.02
−0.04
Class membership
Draw a number
PREDICT
PREDICT SPINN
Add noise
Center img.
TRAIN
Normal. SBCs
CLEAR
EXIT
SpiNN-3

Conclusions

***BitBrain* status:**

- novel single-pass learning mechanism
- accurate inference – best in ‘single-pass’ class
- good robustness to imperfect inputs
- simple & energy-efficient operation (no floating-point or backpropagation)
- continuous and single-shot learning
- single-thread *BitBrain* on 3.2GHz Apple M1 gives 10k inferences in 0.42 secs
- implementations on Raspberry Pi & SpiNNaker

Improvements already investigated:

- ‘jitter’/*data augmentation* – +~1% on MNIST
- weighting of counts by occupancy – +~1% on MNIST

To do:

- more benchmarks, e.g.: CIFAR-10 & -100, German traffic sign database…
- CNN front end (in progress)
- layers of SBC memories
- application to different types of data – time series, DNA, abstract codes
- differing delays on synaptic connections for spatio-temporal patterns
- theory – connection to *kernel methods?*

SpiNNaker2 job opportunities!

SpiNNcloud

Join us:

https://spinncloud.softgarden.io/en/vacancies

Human Brain Project

SpiNNaker2
Universal Spiking Neural Network Architecture

High-level Frameworks (Neuromorphic, Machine Learning, Hybrid)

Intermediate Graph Representation

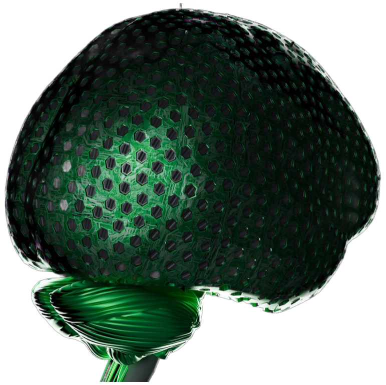
Partition and Mapping

Bare-metal Deployment

Optimization Loop

Edge-based Systems

Large-scale System



- Large NLP Models
- Brain simulation
- Highly parallel control

Sponsored by: European Innovation Council; 6G-life; Bundesministerium für Bildung und Forschung