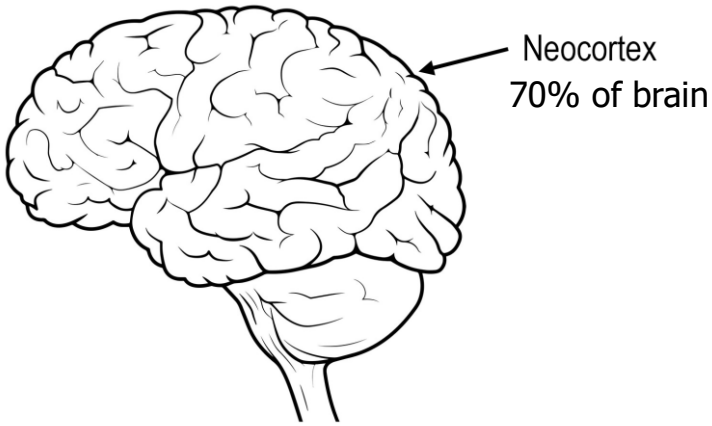
FROM BRAINS TO SILICON
APPLYING LESSONS FROM NEUROSCIENCE TO MACHINE LEARNING

JEFF HAWKINS
SUBUTAI AHMAD

Numenta has two goals

1) **Reverse engineer the neocortex**
2) **Apply what we learn to AI and ML**

Numenta



Organ of intelligence

- Sensory perception: vision, touch, hearing
- Motor: limbs, fingers, vocalization
- Language
- Abstract thought: math, science, engineering

Attributes

- Learns continuously
- Learns rapidly
- Efficient: 20 watts for brain
- Flexible: learns thousands of tasks

Today's AI is not as capable, not even close.

What we have learned, outline of talk

1) The neocortex learns a model of the world.

2) It is a distributed model.
   There are thousands of complete, yet complementary models of everything you know.
   They "vote" to reach a consensus.

3) Each cortical column is a complete modeling system.
   Columns use "reference frames" to provide structure to data and to plan movements.

4) Reference frames in the cortex are derivatives of grid cells and place cells.

The neocortex learns a model of the world

Your model of the world, the number of things you know, is huge

- Thousands of physical objects, how they look, feel, and sound
- How objects are composed other objects
- Where objects are located relative to each other
- How objects behave
- Conceptual "objects" such as math, democracy

We use the model to infer and create goal-oriented behaviors

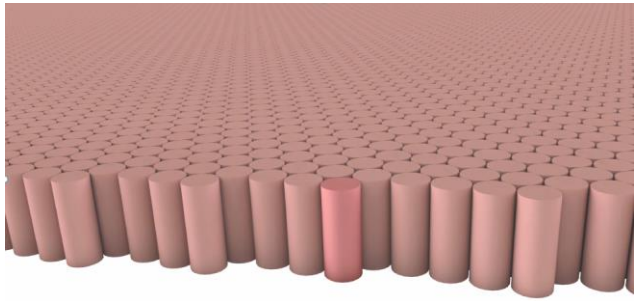
The model is predictive

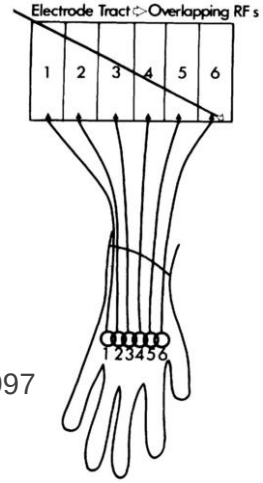
- Prediction error is the primary training signal

Q: How does the architecture of the neocortex support learning this model?

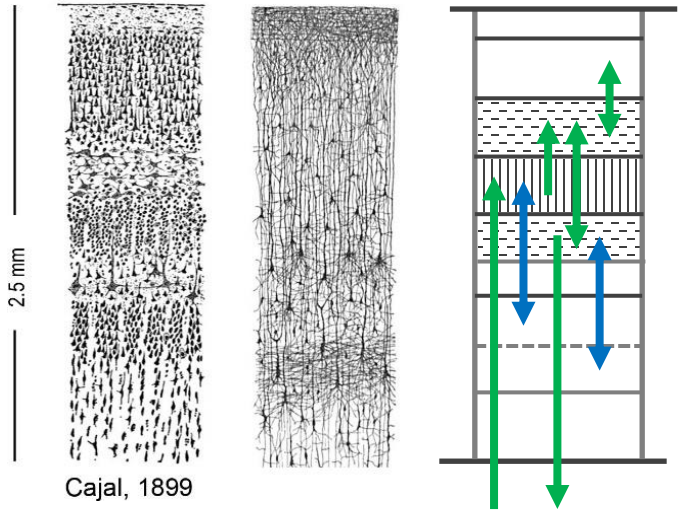
Cortical columns

Approx. 150,000 columns (1 x 2.5mm)





Mountcastle 1997



Cajal, 1899

1) **Columns are complex**
   100K neurons, 500M synapses
   Dozens of cell types, hundreds of minicolumns
   Whatever they do is also complex

2) **All columns have a motor output**

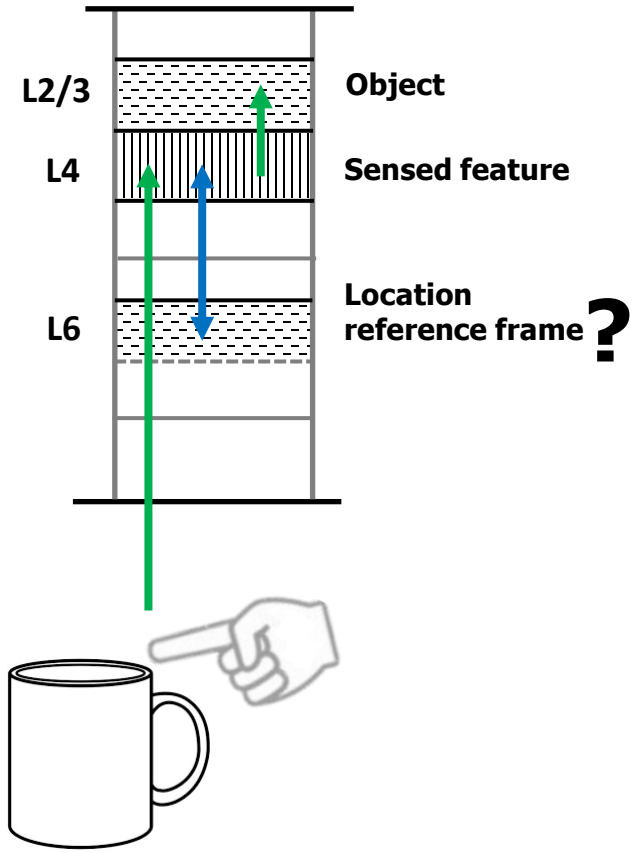
3) **All columns are remarkably similar**

Mountcastle 1979:

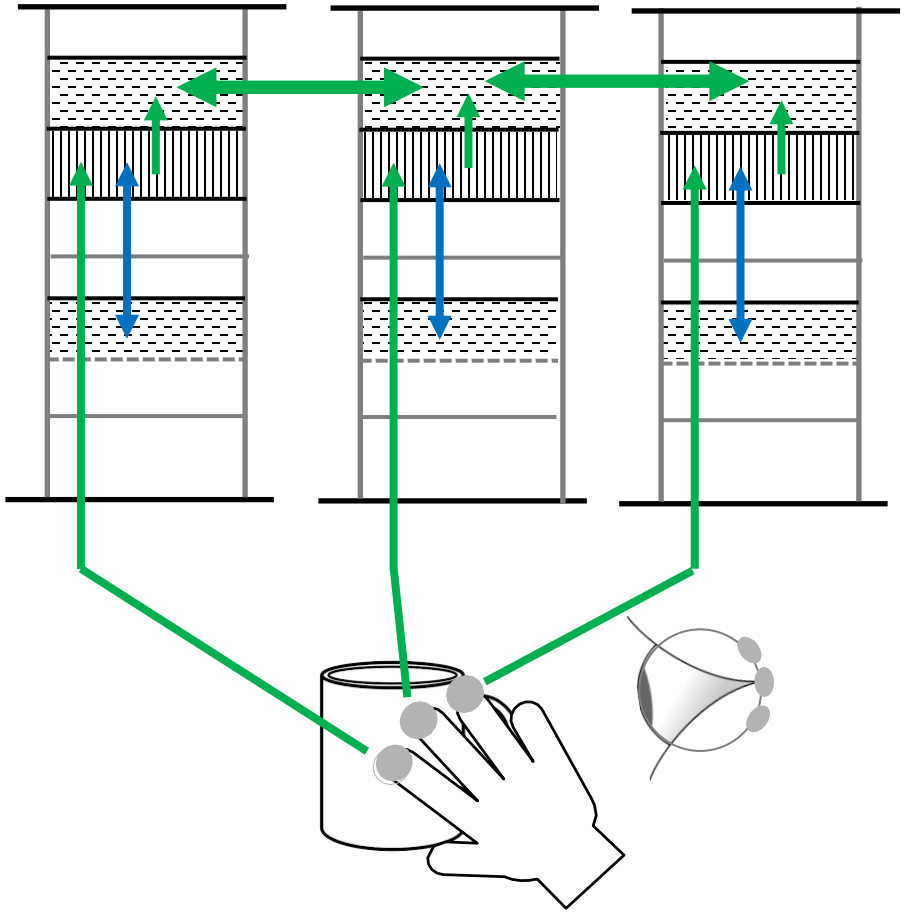
- All columns look similar because they perform the same intrinsic function.
- What a column does is determined by what it is connected to.
- Understanding what a column does will have "great generalizing significance".

Thought Experiment

"A Theory of How Columns in the Neocortex Enable Learning the Structure of the World" (Hawkins, et. al., 2017)

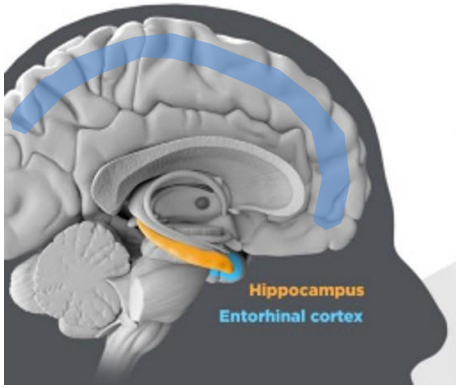


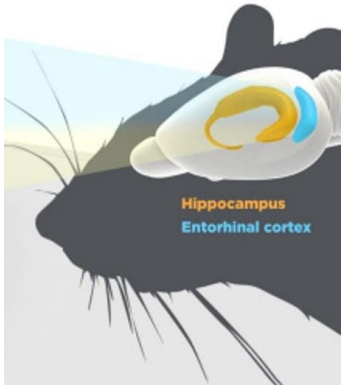
A single column learns completes models of objects by integrating features and locations over time.



Multiple columns can infer objects in a single sensation by "voting" on object identity.

Reference Frames in the Old Brain





"Grid cells" in entorhinal cortex
- Reference frames for environments

Moser, 2005

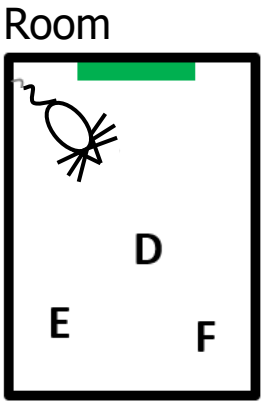
"Place cells" in hippocampus
- Sensory driven representation of location

Okeefe, 1978

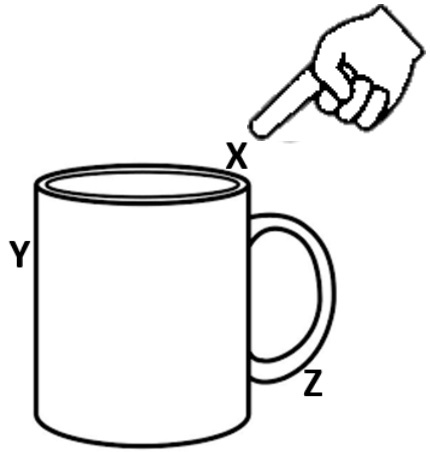
Grid and place cell equivalents exist in every cortical column
- Create reference frames for objects

Hawkins et. al., 2018
Lewis et. al., 2018
Hawkins et. al., 2017

Entorhinal Cortex



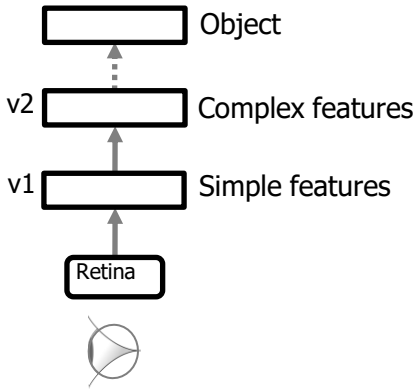
Neocortex

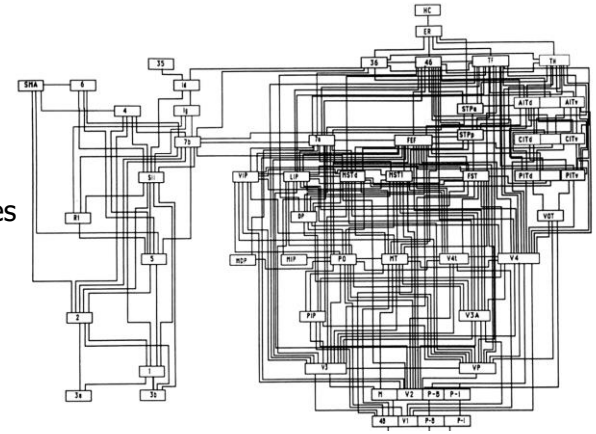


Grid cells
Represent location of body in a reference frame relative to room.

Cortical grid cells
Represent location of sensor in a reference frame relative to object.

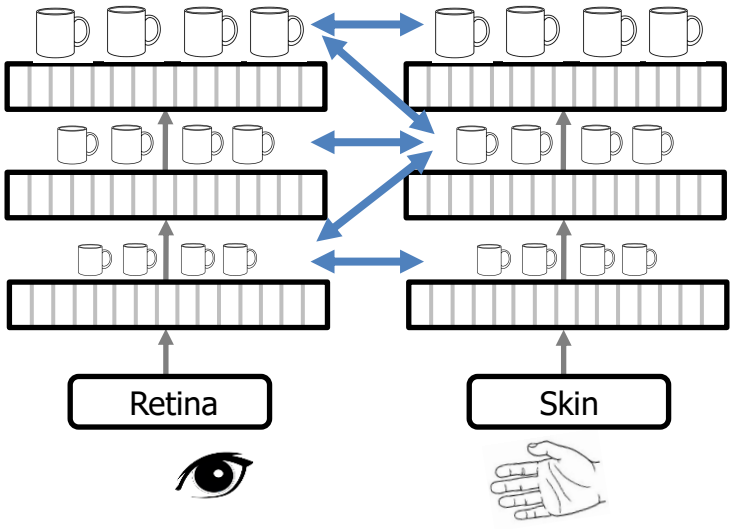
Hierarchy





Felleman, van Essen, 1991

- Most connections are not hierarchical.
- More than 40% of all possible connections exist.
- Primary and secondary regions are largest.
- Primary sensory regions exhibit multi-modal responses.



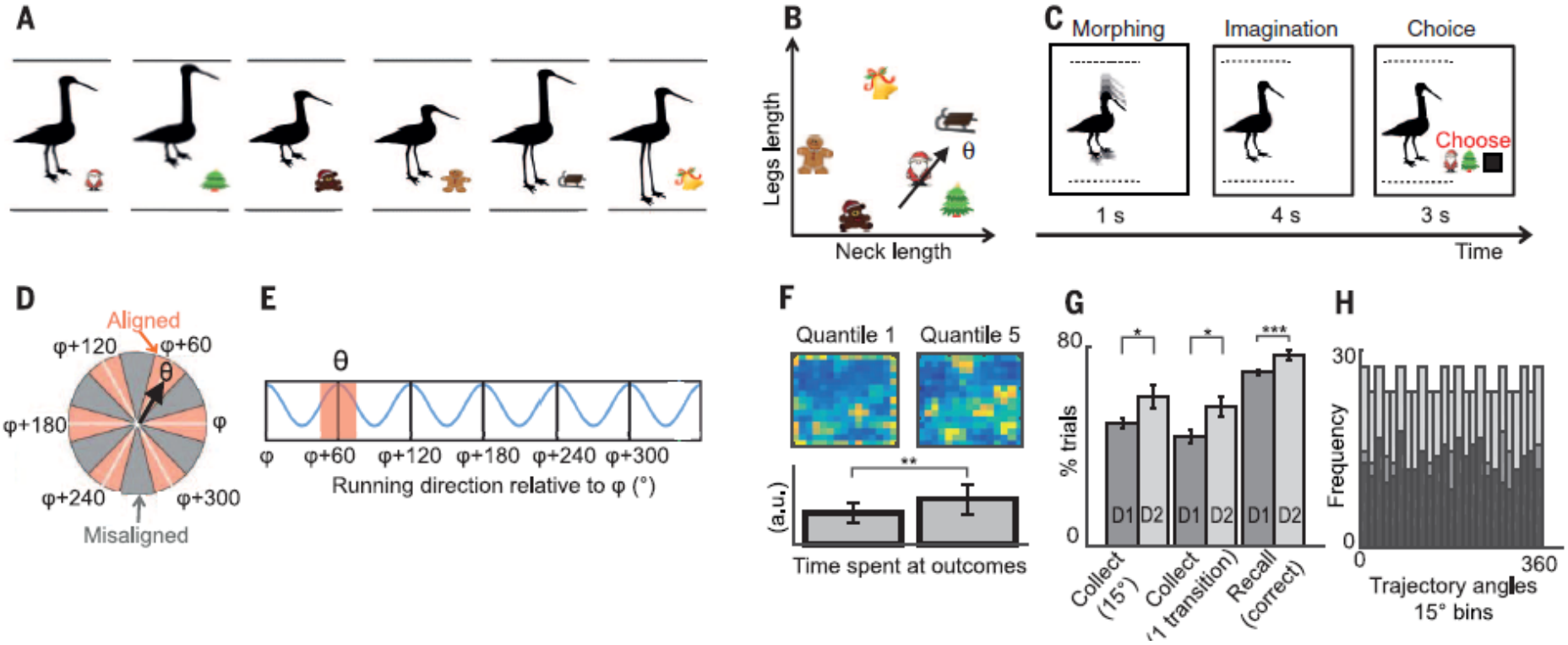
The Thousand Brains Theory of Intelligence

- There are thousands of complementary models

- Most connections are for voting. (blue)
  We are aware of the consensus

- Hierarchical connections pass complete objects

Growing empirical evidence for grid cells in the neocortex



Constantinescu, A., O'Reilly, J., Behrens, T. (2016) Organizing Conceptual Knowledge in Humans with a Gridlike Code. Science

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. Nature

Summary

Intelligence requires learning a model of the world.

1) Each cortical column is a complete sensory-motor modeling system.
2) Columns vote to reach a consensus.
3) Cortical columns use reference frames to represent knowledge.
   - objects, body, concepts

I believe true machine intelligence (AGI) must work on the same principles.

Neuroscience Papers

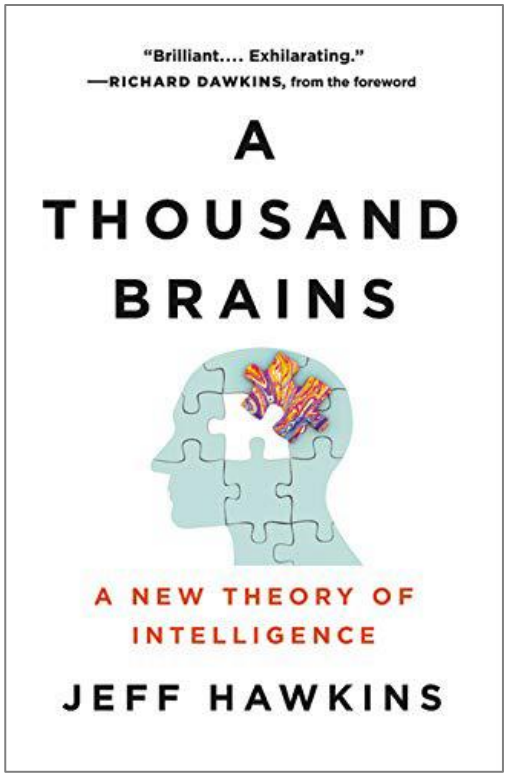
2019: A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex
2019: Locations in the Neocortex: A Theory of Sensorimotor Object Recognition Using Cortical Grid Cells
2019: Flexible Representation and Memory of Higher-Dimensional Cognitive Variables with Grid Cells
2017: A Theory of How Columns in the Neocortex Enable Learning the Structure of the World
2017: The HTM Spatial Pooler—A Neocortical Algorithm for Online Sparse Distributed Coding
2016: Why Neurons Have Thousands of Synapses, A Theory of Sequence Memory in Neocortex
2016: Continuous Online Sequence Learning with an Unsupervised Neural Network Model

See Numenta.com for annotated list of papers

Section 1: Neuroscience

Section 2: AI

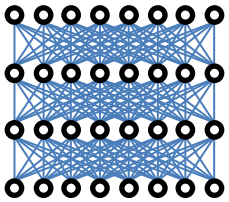
Section 3: Implications



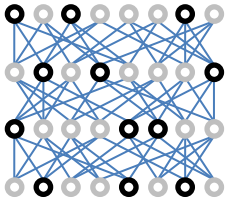
March 2, 2021

ROADMAP TO MACHINE INTELLIGENCE

CNN

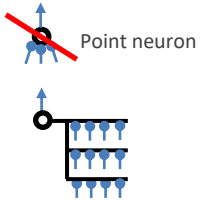


Sparsity



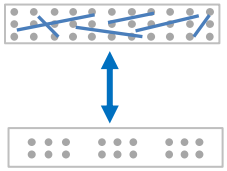
Performance
Robustness

Active Dendrites



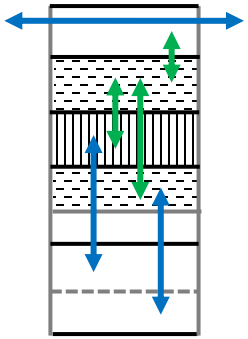
Continuous learning

Reference frames



Invariant representations
Fast learning

Model voting



The neocortex is highly sparse

Source: Prof. Hasan, Max-Planck-Institute for Research

- Neural activity and connectivity are both highly sparse
  - Only 0.5% to 2% of cells are active at any time
  - Only 1% - 5% of connections actually exist between two connected layers
  - Dynamic structural plasticity - 30% of connections change every few days

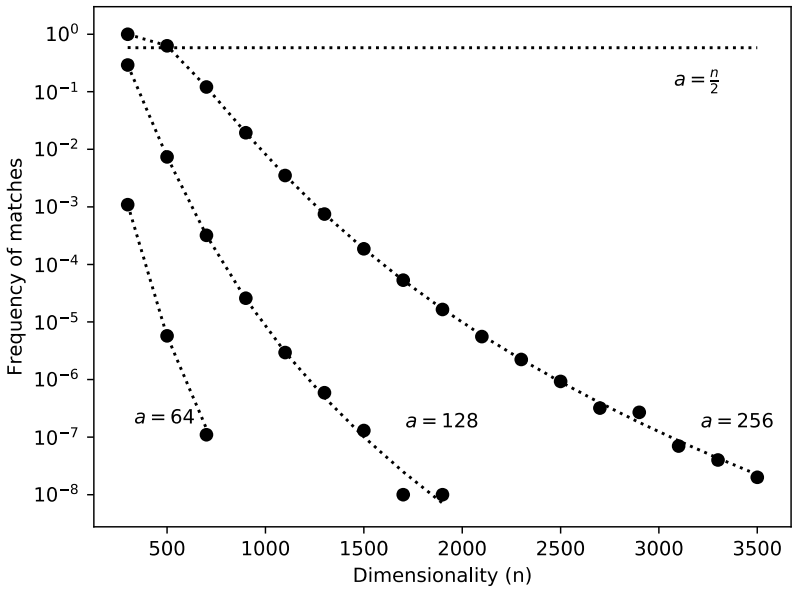
(Attwell & Laughlin, 2001; Lennie, 2003; Holmgren et al., 2003; Loewenstein, et al., 2015)

- Nothing like today's dense deep learning networks

High dimensional sparse representations

- High dimensional sparse representations are extremely robust to noise and failure
- Information content of sparse vectors increases with dimensionality, without introducing additional non-zeros
- For a given task, can reduce the number of non-zero parameters by increasing dimensionality

Sparse vector dot product: probability of false matches

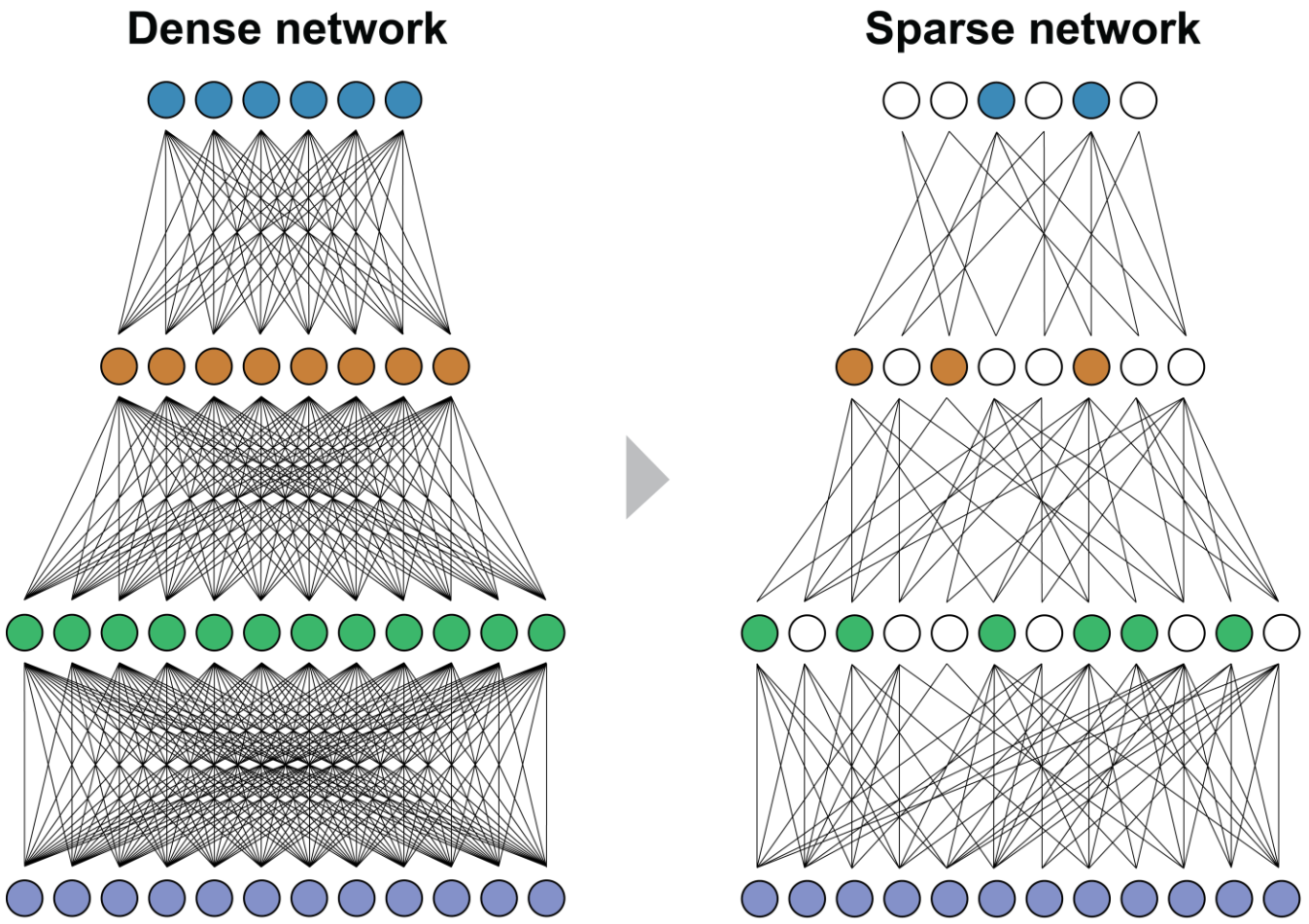


$$P(\boldsymbol{x}_i \cdot \boldsymbol{x}_j \geq \theta) = \frac{\sum_{b=\theta}^{|\boldsymbol{x}_i|} \mid \Omega^n(\boldsymbol{x}_i, b, |\boldsymbol{x}_j|) \mid}{\binom{n}{|\boldsymbol{x}_j|}}$$

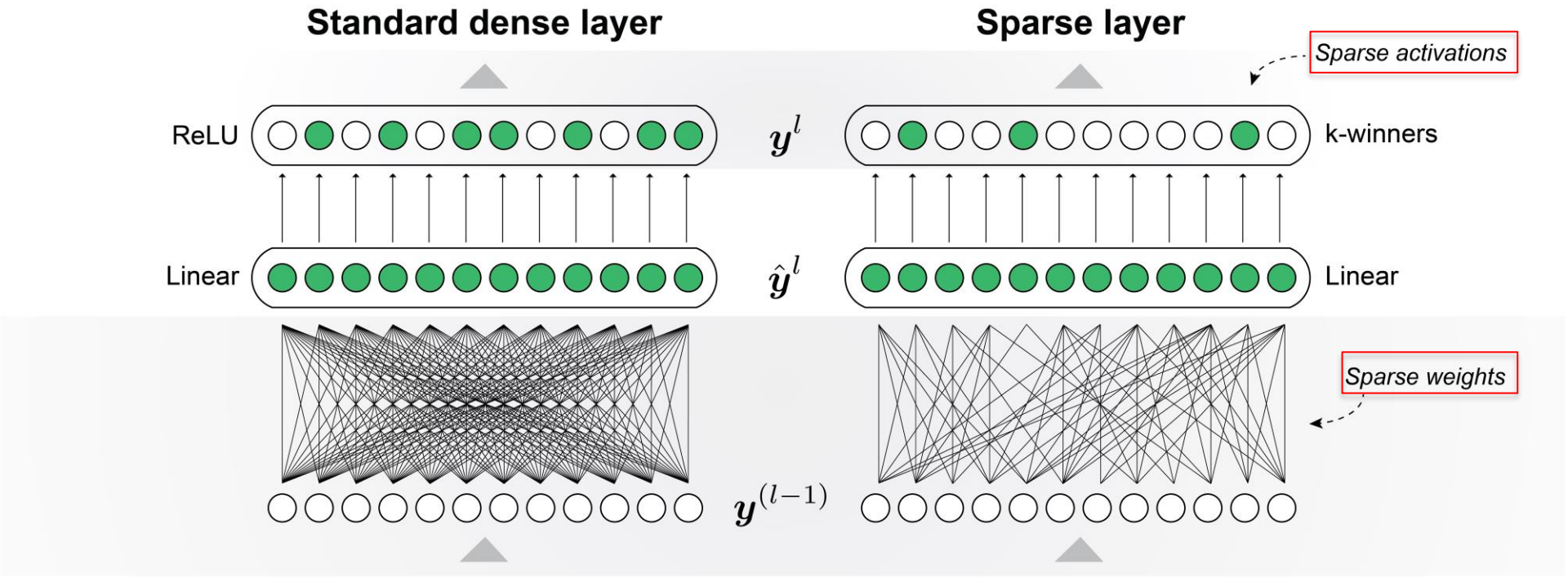
$$|\boldsymbol{x}_i| = 24, \theta = 12, a = |\boldsymbol{x}_j|$$

(How Can We Be So Dense? The Robustness of Highly Sparse Representations, Ahmad & Scheinkman, ICML Robustness Workshop 2019)

Sparse deep networks



Sparse layers



(Hawkins, Ahmad, & Dubinsky, 2011)
(Makhzani & Frey, 2015)
(Ahmad & Scheinkman, 2019)

1) Sparse weights: weight matrix is sparse, enforced via mask

2) Sparse activations: outputs of top-k units are maintained

3) An exponential boosting term favors units with low activation frequency: $b_i^l(t) = e^{\beta(\hat{a}^l - d_i^l(t))}$
This helps maximize the overall entropy of the layer.

4) Extension to sparse convolutional layers

Google speech commands dataset

Dataset of spoken commands

- One word utterances, thousands of individuals
- State of the art accuracy is 95 - 97.5% for 10 categories
- Tested robustness to white noise

Network	Mean accuracy	Mean accuracy with noise	Non-zero weights	Sparsity
Dense CNN	97.05%	31.08%	1,700,000	0%
Sparse CNN	97.03%	44.45%	160,952	90.6%

1) Networks used two sparse CNN layers + one sparse linear layer + one softmax output layer.
2) Trained with random static sparse masks

performance on FPGA

- Implemented sparse and dense networks on three Xilinx chips
  - Processing spoken words (Google Speech Commands)
  - Chips de data d emb (internet of things) appl

	Alveo U250	Zynq UltraScale+ ZCU104	Zynq UltraScale+ ZU3EG
System logic cells	1,728,000	504,000	154,000
Memory	54MB	4.75MB	0.95MB
DSP slices	12,288	1,728	360
System power	225W	60W	24W

Sparse networks: more than 50X faster

Name of chip	Network type	Throughput for single network	Speedup over dense	Number of networks on chip	Full chip throughput	Full chip speedup
Alveo U250	Dense	3,049	-	4	12,195	-
Alveo U250	Sparse	31,250	10.25	20	625,000	51.25
ZCU104	Dense	6,410	-	1	6,410	-
ZCU104	Sparse	26,667	4.16	3	80,000	12.48
ZU3EG	Dense	0	-	0	0	-
ZU3EG	Sparse	21,053	Infinite	1	21,053	Infinite

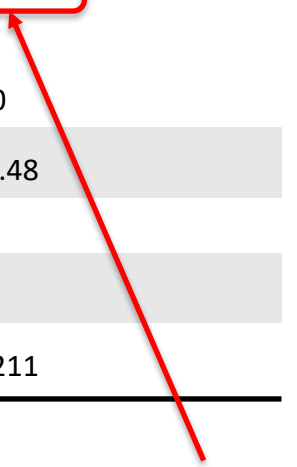
Each network is >10X faster

Dense network does not even fit on the small chip

Overall >50X throughput

Sparse networks are far more power efficient

Name of chip	Network type	System power	Words / Watt	Relative efficiency (compared to best dense network)
Alveo U250	Dense	225	54	0.507
Alveo U250	Sparse	225	2,778	26.00
ZCU104	Dense	60	107	1.0
ZCU105	Sparse	60	1,333	12.48
ZU3EG	Dense	24	0	-
ZU3EG	Sparse	24	877	8.211

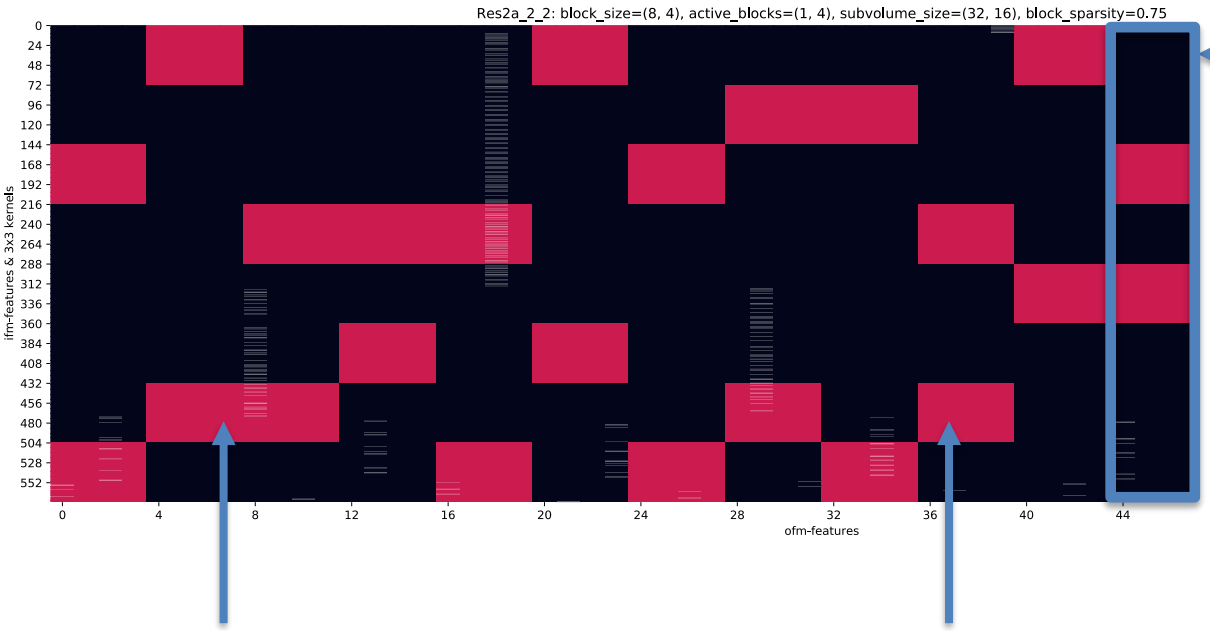


>25X efficiency

Structural plasticity for hardware

- Dynamic block sparse weights, trained using variational technique
- Training respects hardware constraints

Output features



Number of non-zero blocks per column controlled for balanced computation

During training, each block has a probability of being “on”

Weights are noisy, trained to be quantizable

$$w_{ij} = \widetilde{w}_{ij} z_{ij}$$
$$z_{ij} \sim N(1, \alpha_{ij})$$

Sparse resnet50 on imagenet

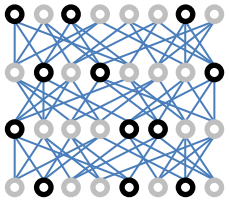
- Created a sparse version of ResNet50
  - Trained on Imagenet 1K
  - Sparsity structure trained using structural plasticity
  - Training respects optimization constraints
- Accuracy results:

Network	Sparsity	Accuracy (float32)	Accuracy (int8)	Quantization impact
MLPerf benchmark, dense	0%	76.7%	75.7%	-1.00%
NVIDIA, static sparsity	50%		76.8%	
Ours, static sparsity	75%	76.22%	74.67%	-1.55%
Ours, dynamic sparsity	75%	77.1%	76.77%	-0.33%

- FPGA implementation in process (inference only)

ROADMAP TO MACHINE INTELLIGENCE

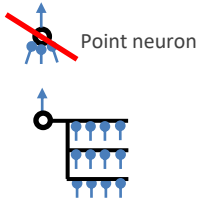
Sparsity



Performance and robustness

- Sparse activations and weights
- Structural plasticity
- Custom sparse processing logic
- 50X to 100X more efficient
- Robustness to noise

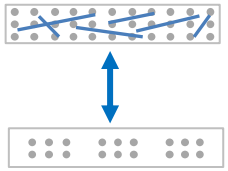
Active dendrites



Continuous self-supervised learning

- Learn new sparse patterns without disrupting existing patterns
- Fewer training passes
- Learn from prediction errors
- Far less labeled data

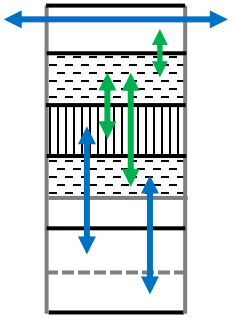
Reference frames



Invariant representations

- Much smaller training sets
- Compositional structures
- Improved generalization

Cortical columns



Common cortical algorithm

- Common repeating circuit for intelligence
- Highly scalable
- Integrated sensorimotor
- Advanced robotics

Contact:
Jeff: jhawkins@numenta.com
Subutai: sahmad@numenta.com

Twitter: @Numenta, @SubutaiAhmad
Papers: numenta.com/papers