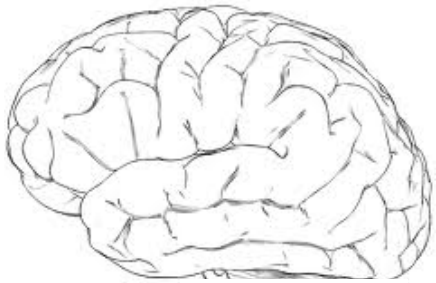
Sparsity In The Neocortex, And Its Implications For Machine Learning



Nantes Machine Learning Meetup
October 5, 2020

Subutai Ahmad
Email: sahmad@numenta.com
Twitter: @SubutaiAhmad

Numenta

Agenda

1) Neuroscience & AI
   A framework for intelligence and cortical function

2) Sparsity in the neocortex

3) Implications for Machine Learning

Numenta

Numenta

Founded in 2005,
by Jeff Hawkins and Donna Dubinsky

Mission

1) Reverse engineer the neocortex
   - biologically accurate theories
   - open access neuroscience publications

2) Apply neocortical principles to AI
   - improve current techniques
   - move toward truly intelligent systems

AI and the Brain

Neocortex
- 70% of human brain (newest)
- Organ of intelligence

- Sensory: vision, touch, hearing
- Motor: limbs, fingers, language
- Abstract: engineering, science, philosophy, etc.

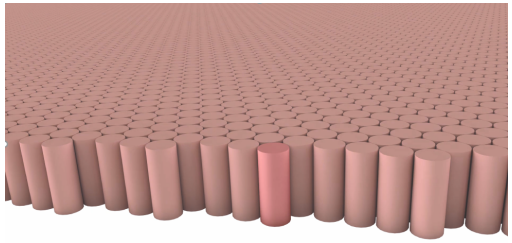
- Learns continuously and rapidly
- Robust
- Efficient (fastest processing step about 5ms)
- Low energy (20 watts for entire brain)

Today's AI and robotics work on different principles and are not remotely as capable.

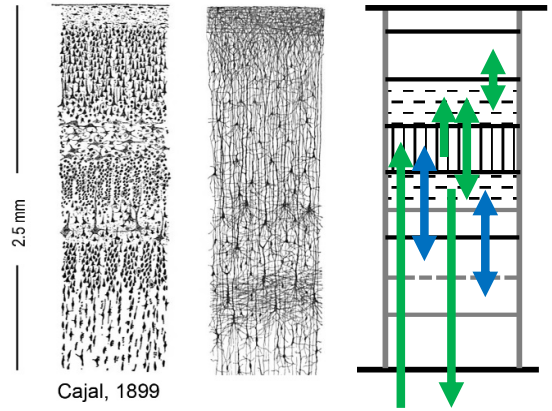
How the neocortex works has been a mystery.

The Thousand Brains Theory

(1000 meter view)



150,000 columns (1mm x 2.5mm)



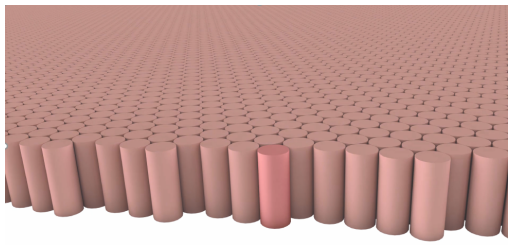
All columns (vision, touch, language, etc.) have the same complex circuitry

Mountcastle 1979:

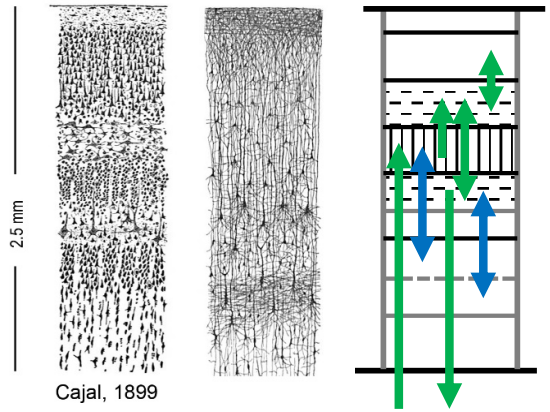
All columns look the same because they perform the same intrinsic function.

The Thousand Brains Theory

(1000 meter view)



150,000 columns (1mm x 2.5mm)



All columns (vision, touch, language, etc.) have the same complex circuitry

Numenta's discoveries (all published)

- Theory of sparsity
- New neuron model
- 2016: Columns are reference frame processors

Reference Frames

- All knowledge is stored at locations in reference frames.
- Allow columns to learn complete models through movement.
- Columns "vote" to reach a consensus.
- Allow the generation of goal-oriented behavior.
- Thinking is "moving" from location to location in reference frames. *(Hawkins et al., 2019)*

The unit of scaling for AGI and robotics is the column equivalent.

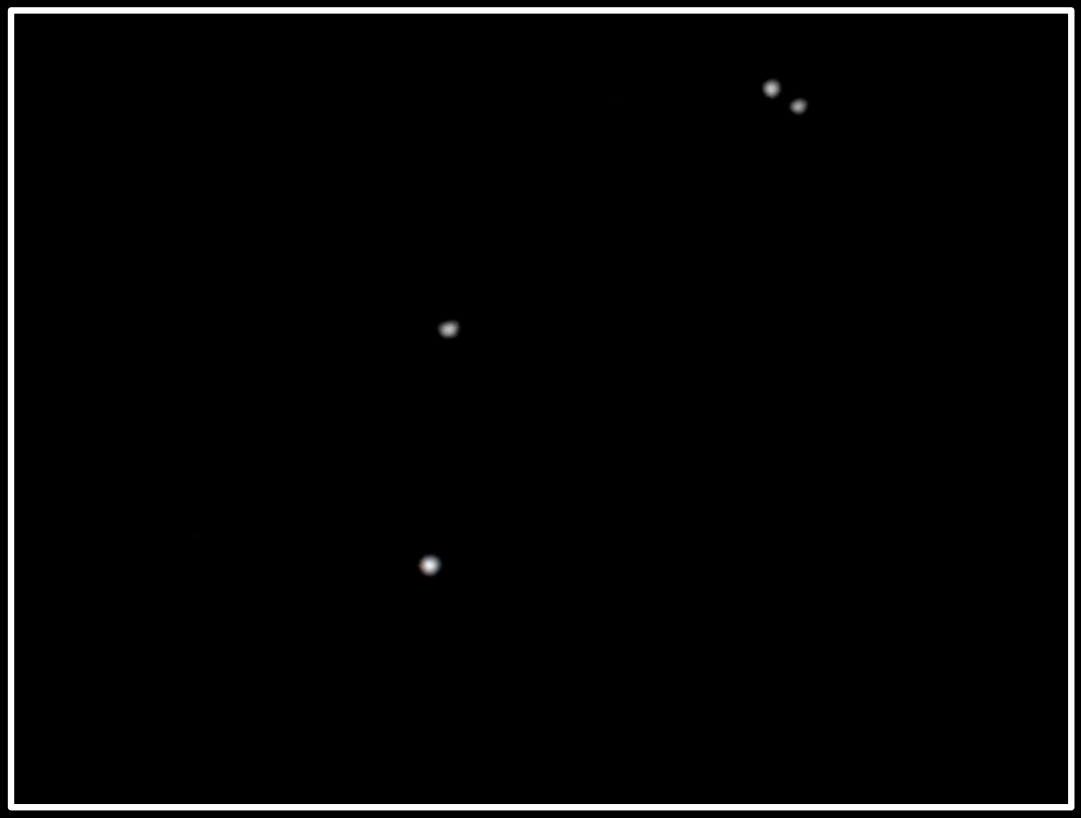
Agenda

1) Neuroscience & AI
   A framework for intelligence and cortical function

2) Sparsity in the neocortex

3) Implications for Machine Learning

Numenta



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

What exactly is "sparsity"?

"mostly missing"
sparse vector = vector with mostly zero elements

Most papers describe three types of sparsity:

1) Population sparsity
   How many neurons are active right now?
   Estimate: roughly 0.5% to 2% of cells are active at a time (Attwell & Laughlin, 2001; Lennie, 2003).

Numenta

What exactly is "sparsity"?

"mostly missing"
sparse vector = vector with mostly zero elements

Most papers describe three types of sparsity:

1) Population sparsity
   How many neurons are active right now?
   Estimate: roughly 0.5% to 2% of cells are active at a time (Attwell & Laughlin, 2001; Lennie, 2003).

2) Lifetime sparsity
   How often does a given cell fire?

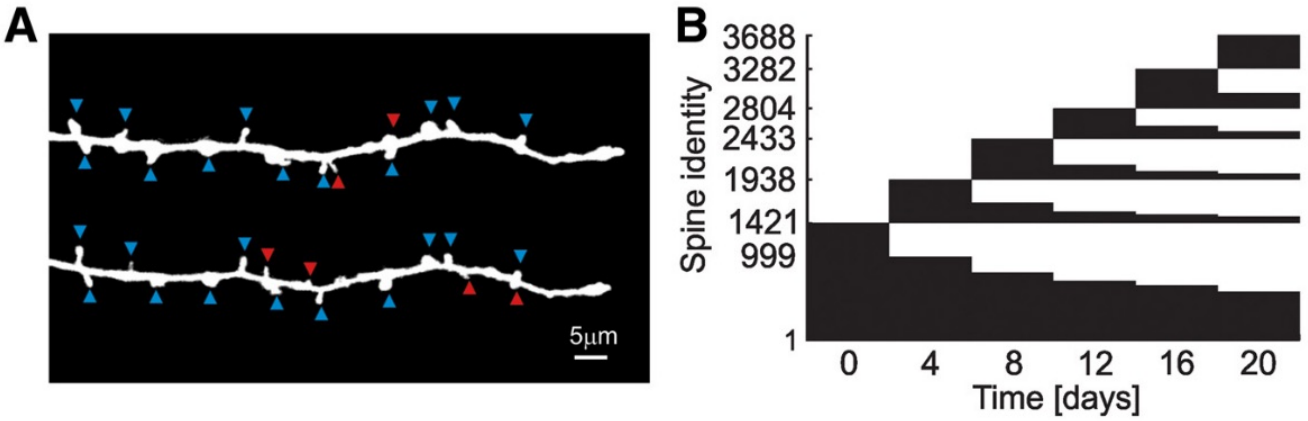
3) Connection sparsity
   If a group or layer of cells projects to another layer, what percentage are physically connected?
   Estimate: 1% - 5% of possible neuron to neuron connections exist (Holmgren et al., 2003).

Numenta

Connectivity is sparse and surprisingly dynamic

Learning involves growing and removing synapses

- Structural plasticity: network structure is dynamically altered during learning



"We observed substantial spine turnover, indicating that the architecture of the neuronal circuits in the auditory cortex is dynamic (Fig. 1B). Indeed, 31% ± 1% (SEM) of the spines in a given imaging session were not detected in the previous imaging session; and, similarly, 31 ± 1% (SEM) of the spines identified in an imaging session were no longer found in the next imaging session. *(Loewenstein, et al., 2015)*

What exactly is "sparsity"?

"mostly missing"
sparse vector = vector with mostly zero elements

Most papers describe three types of sparsity:

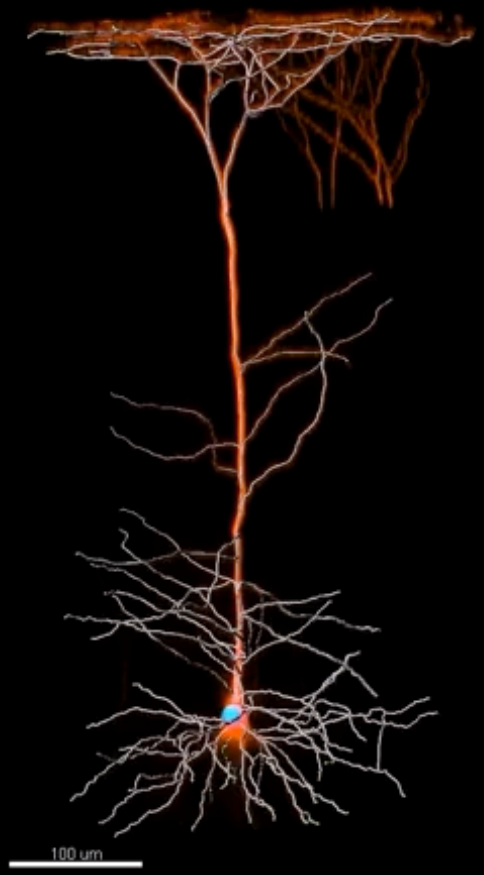
1) Population sparsity
   How many neurons are active right now?
   Estimate: roughly 0.5% to 2% of cells are active at a time (Attwell & Laughlin, 2001; Lennie, 2003).

2) Lifetime sparsity
   How often does a given cell fire?

3) Connection sparsity
   If a group or layer of cells projects to another layer, what percentage are physically connected?
   Estimate: 1% - 5% of possible neuron to neuron connections exist (Holmgren et al., 2003).

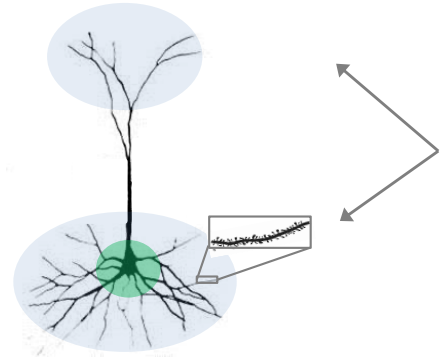
…but there's more to sparsity…

Numenta



Source: Smirnakis Lab, Baylor College of Medicine

Dendrites of cortical neurons detect sparse patterns

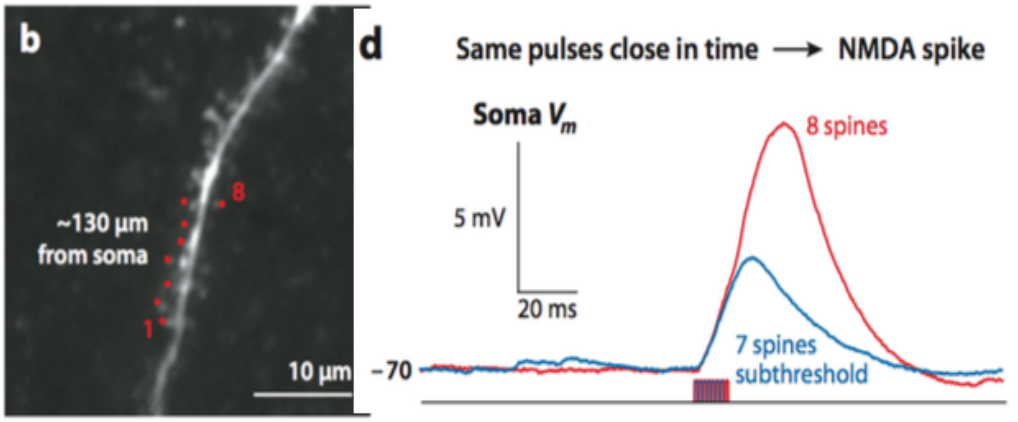


Pyramidal neuron
3K to 10K synapses

Dendrites contain dozens of independent computational segments

These segments become active with cluster of 10-20 active synapses

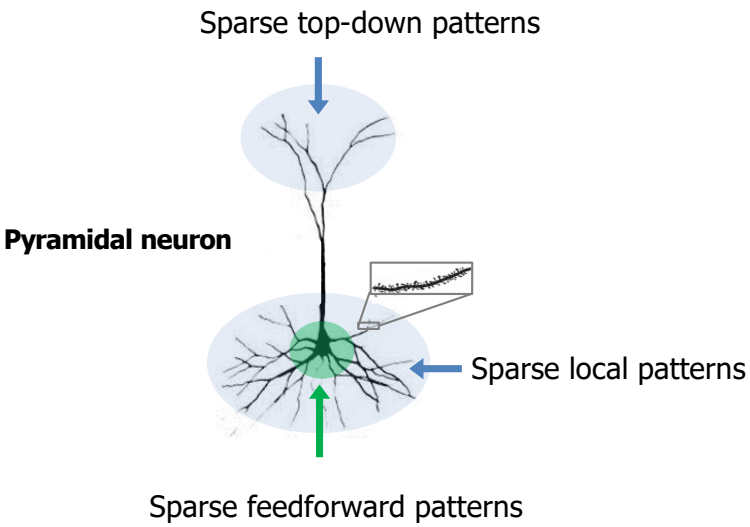
Neurons detect highly sparse patterns, in parallel



Major, Larkum and Schiller 2013

(Mel, 1992; Branco & Häusser, 2011; Schiller et al, 2000; Losonczy, 2006; Antic et al, 2010; Major et al, 2013; Spruston, 2008; Milojkovic et al, 2005, etc.)

Sparse learning: only a small part of the neuron is updated



Learning localized to dendritic segments
"Branch specific plasticity"

<u>If cell becomes active:</u>

- If there was a dendritic spike, reinforce that segment
- If there were no dendritic spikes, grow connections by subsampling cells active in the past

<u>If cell is not active:</u>

- If there was a dendritic spike, weaken the segments

(Gordon et al., 2006; Losonczy et al., 2008; Yang et al., 2014; Cichon & Gang, 2015; El-Boustani et al., 2018; Weber et al., 2016; Sander et al., 2016; Holthoff et al., 2004)

Sparsity is deeply ingrained in the neocortex

1) Dynamic population sparsity
   A small percent of neurons are active at any time
2) Lifetime sparsity
   Cells don't fire very often
3) Sparse dynamic connections
   Connected to a small percent of potential connections
4) Neurons detect sparse patterns
   Dozens of independent segments detect sparse patterns
5) Sparse learning
   Tiny percentage of synapses are updated during learning
6) Sparse synapse weights
   Highly quantized, close to binary
7) Sparse energy usage

Why are deep learning systems so dense?

Are there advantages to highly sparse systems that we can exploit?

Agenda

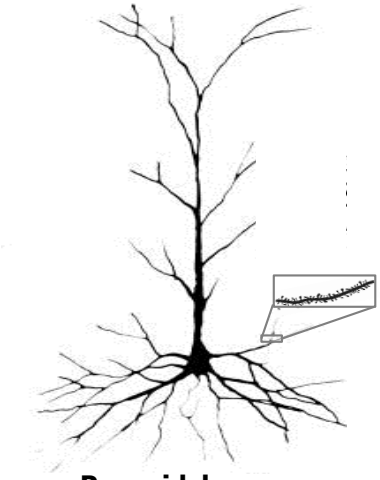
1) Neuroscience & AI
   A framework for intelligence and cortical function

2) Sparsity in the neocortex

3) Implications for Machine Learning

Numenta

Stability of highly sparse representations

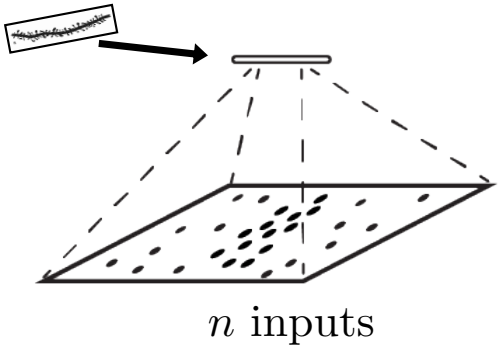


Pyramidal neuron
3K to 10K synapses

Thousands of neurons send input to any single neuron.

On each neuron, 8-20 synapses on tiny segments of dendrites can recognize patterns.

The connections are learned.



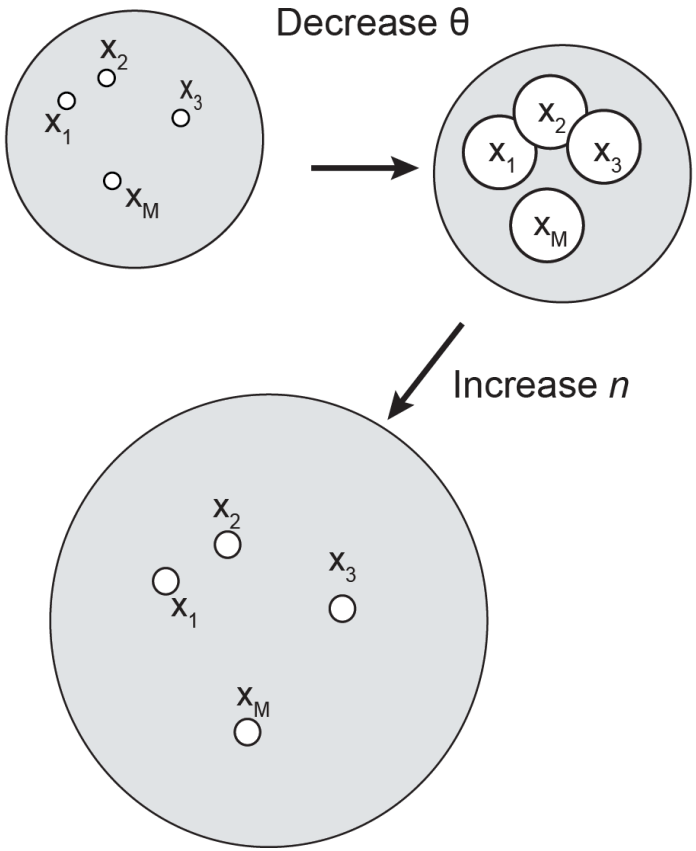
Binary sparse vector matching

$\boldsymbol{x}_i$ = connections on dendrite

$P(\boldsymbol{x}_i \cdot \boldsymbol{x}_j \geq \theta)$

$\boldsymbol{x}_j$ = input activity

Sparse high dimensional representations are highly stable



$$P(\boldsymbol{x}_i \cdot \boldsymbol{x}_j \geq \theta)$$

We can get excellent stability and robustness by reducing θ, at the cost of increased "false positives" and interference.

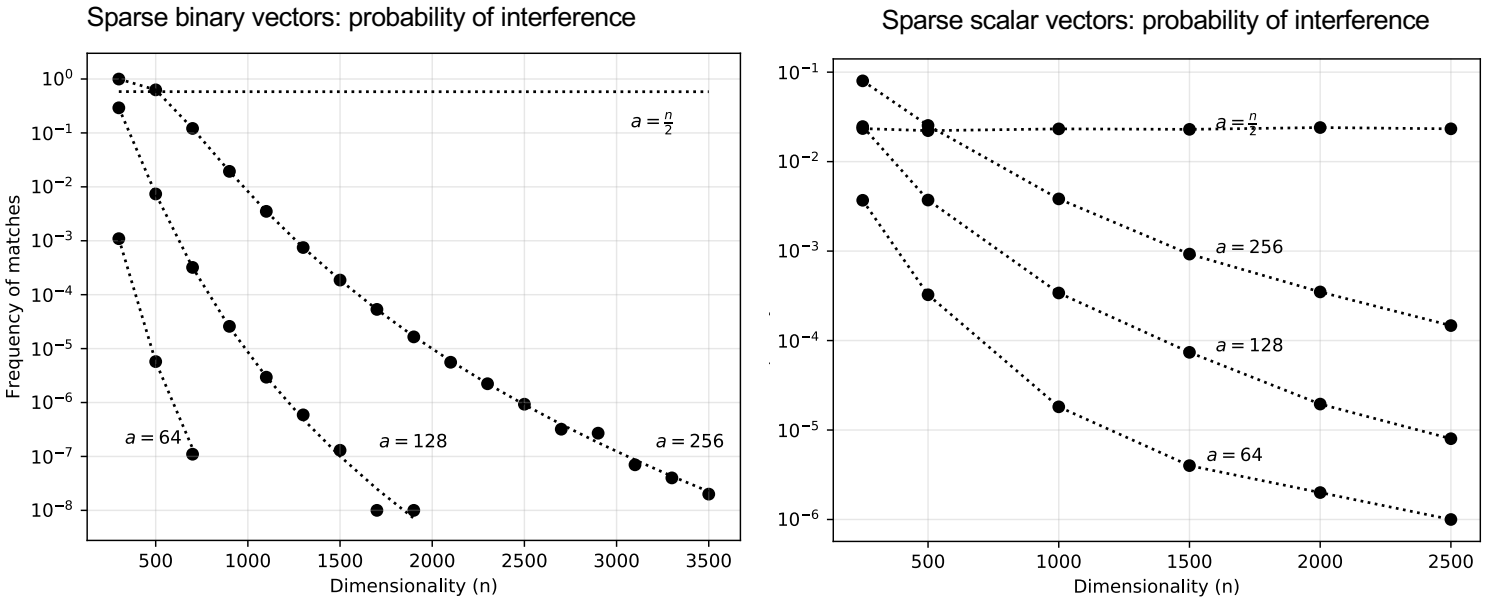
Can compute the probability of a random vector $\boldsymbol{x}_j$ matching a given $\boldsymbol{x}_i$:

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

Numerator: volume around point (white)
Denominator: full volume of space (grey)

$$|\Omega^n(\boldsymbol{x}_i, b, k)| = \binom{|\boldsymbol{x}_i|}{b}\binom{n - |\boldsymbol{x}_i|}{k - b}$$

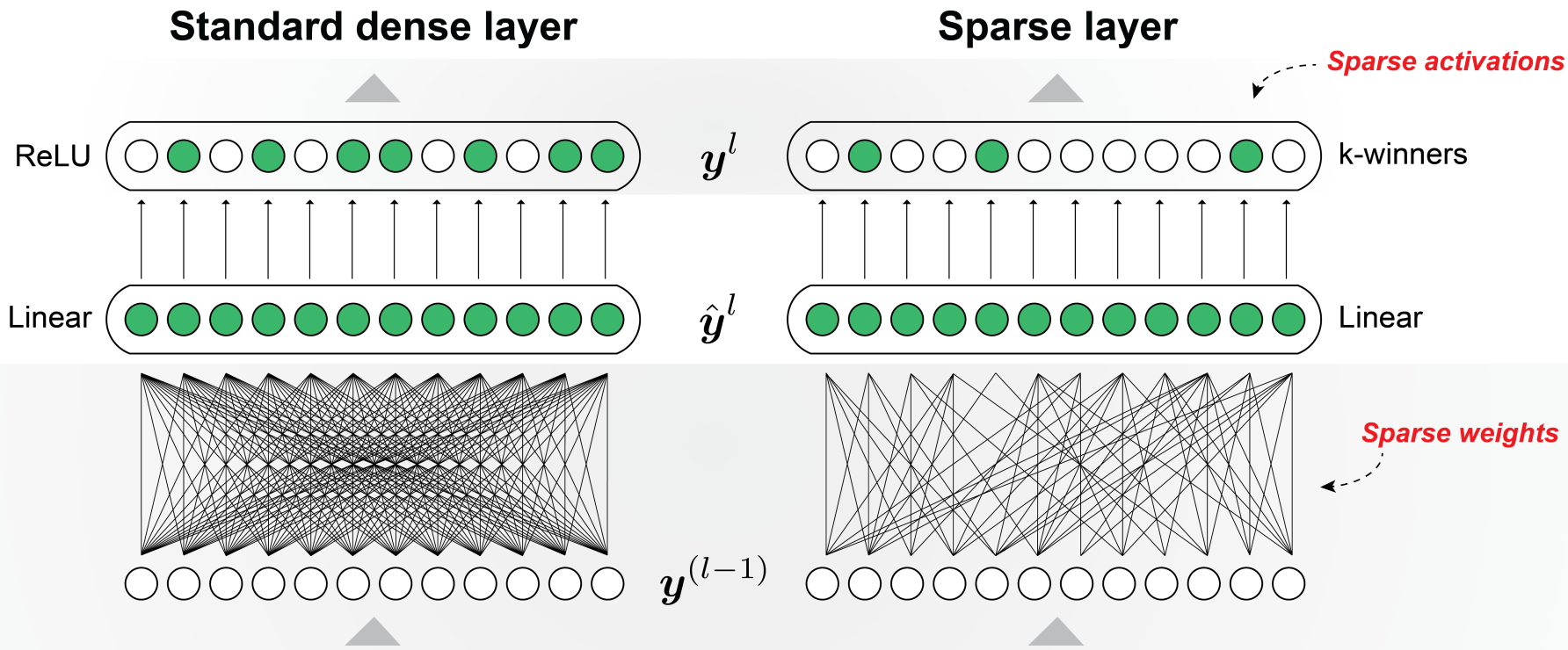
Sparse high dimensional representations are highly robust



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

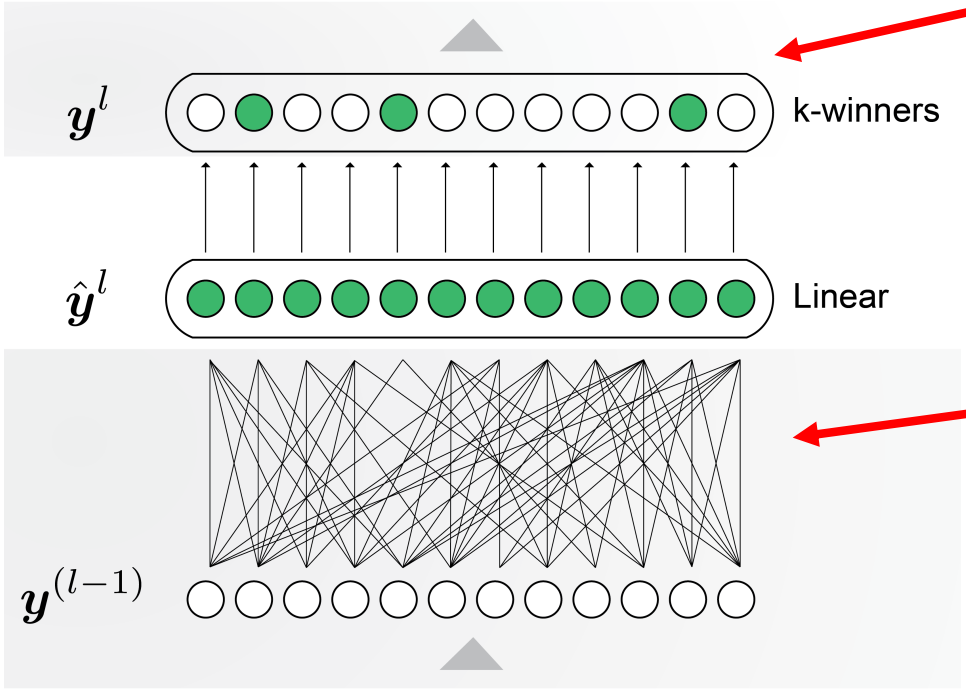
1) False positive error decreases exponentially with dimensionality with sparsity.
2) Error rates do not decrease when activity is dense (a=n/2).
3) Assume uniform random distribution of vectors.

Sparse network formulation



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

Sparse layer formulation



Sparse activations: outputs of top-k units are maintained, the rest are set to 0.

Sparse weights: weight matrix is sparse (most weights are zero, enforced via mask).

1) 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.

2) Easy extension to sparse convolutional layer

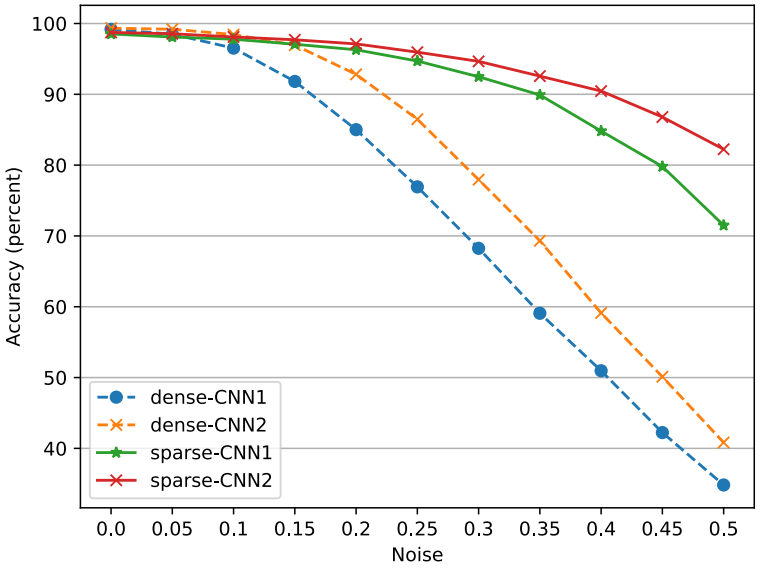
(Hawkins, Ahmad, & Dubinsky, 2011)
(Makhzani & Frey, 2015)
(Ahmad & Scheinkman, 2019)

Results: MNIST with sparse networks

MNIST

NETWORK	TEST SCORE
DENSE CNN-1	99.14 ± 0.03
DENSE CNN-2	99.31 ± 0.06
SPARSE CNN-1	98.41 ± 0.08
SPARSE CNN-2	99.09 ± 0.05

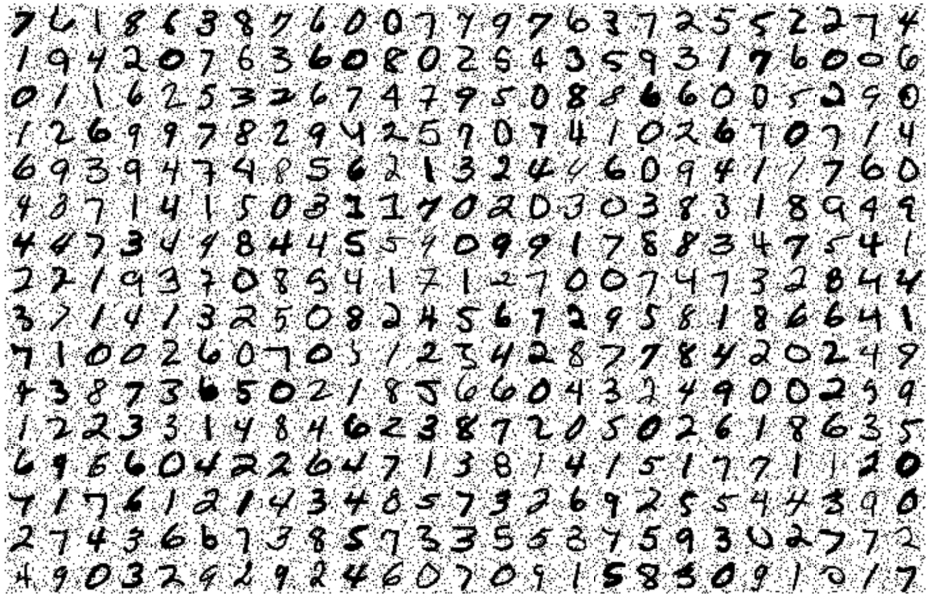
MNIST: Accuracy vs noise



1) Networks used CNN layers + one (sparse) linear layer + one softmax output layer.
2) State of the art test set accuracy is between 98.3% and 99.4% (without data augmentation)

(LeCun et al., 1988)
(Ahmad & Scheinkman, 2019)

Sparse networks are significantly better on noisy data

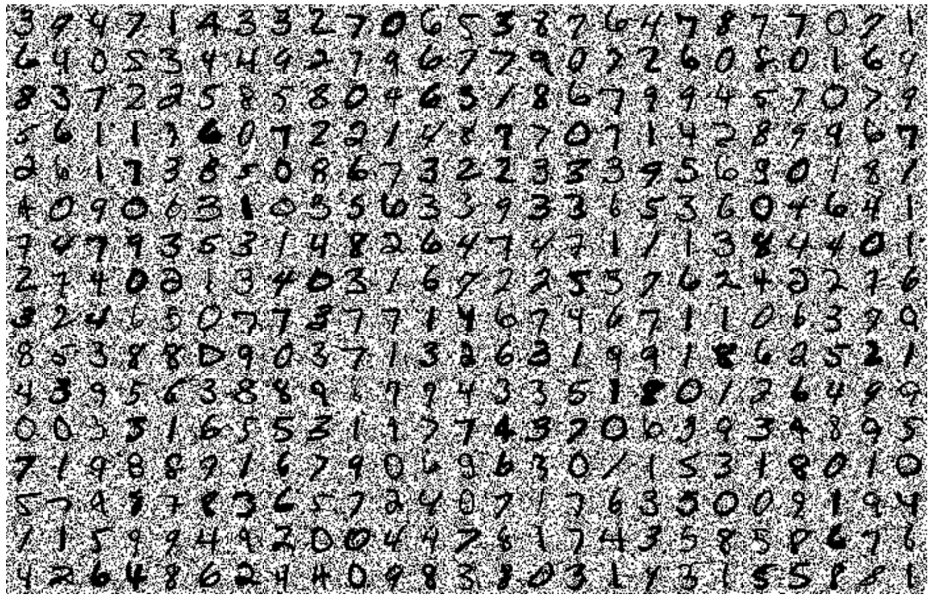


10% Noise

Dense CNN
97 %

SparseNet
98 %

Sparse networks are significantly better on noisy data



30% Noise

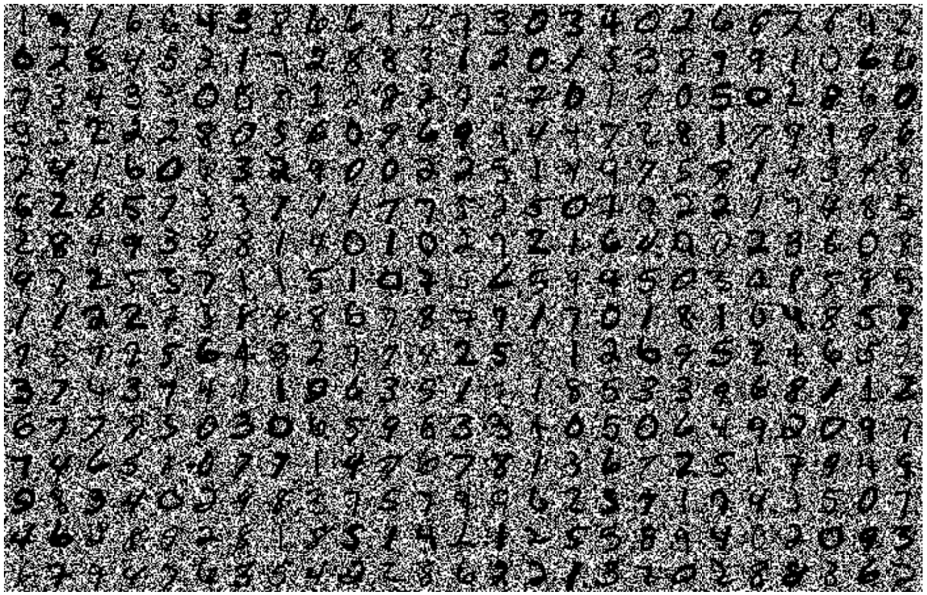
Dense CNN

64 %

SparseNet

92 %

Sparse networks are significantly better on noisy data



50% Noise

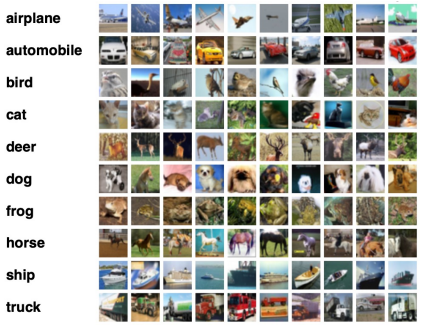
Dense CNN

34 %

SparseNet

72 %

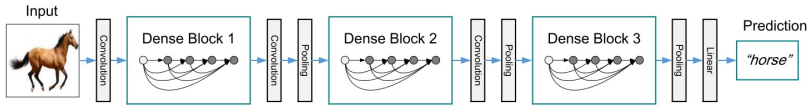
Results: CIFAR-10



NOISE	DENSENET	NOTSODENSENET	VGG19-DENSE	VGG19-SPARSE
0.0%	92.80	93.09	93.24	92.10
2.5%	86.34	87.50	85.07	86.21
5.0%	77.19	79.10	75.88	79.00
7.5%	66.22	69.52	63.60	71.34
10.0%	55.10	61.13	52.41	64.18
12.5%	45.79	52.10	42.25	56.49
15.0%	38.67	45.25	35.25	50.86
17.5%	33.03	39.60	29.37	45.00

DenseNet

- DenseNet (Huang et al., 2016), implements dense level skipping in blocks.
- NotSoDenseNet: sparse transition layers and sparse linear classification layer.



VGG

- VGG19 as in (Simonyan & Zisserman, 2014), with batch norm
- Sparse CNN layers with sparse weights, no linear layer

Results: Google Speech Commands Dataset

Dataset of spoken one word commands

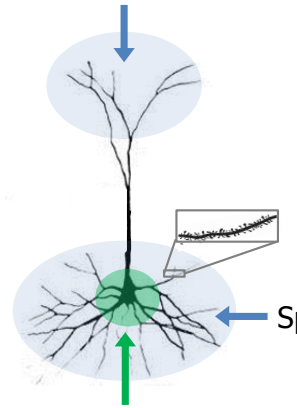
- Released by Google in 2017
- 65,000 utterances, thousands of individuals
- Harder than MNIST
- State of the art is around 95 - 97.5% for 10 categories
- Tested accuracy with white noise

Network	Test Score	Noise Score	Params
Dense CNN	97.05± 0.20	31.08± 2.46	1.7M
Sparse CNN	97.03± 0.14	44.45± 2.54	160,952
Dense Small1	96.14± 0.73	26.57± 2.39	536,008
Dense Small2	95.89± 0.51	26.29± 3.11	270,376

1) Networks used two sparse CNN layers + one sparse linear layer + one softmax output layer.
2) Batchnorm used for all hidden layers
3) Audio files were converted to 32-MFCC coefficients, with data augmentation during training.

Continuous learning with active dendrites

Sparse top-down patterns

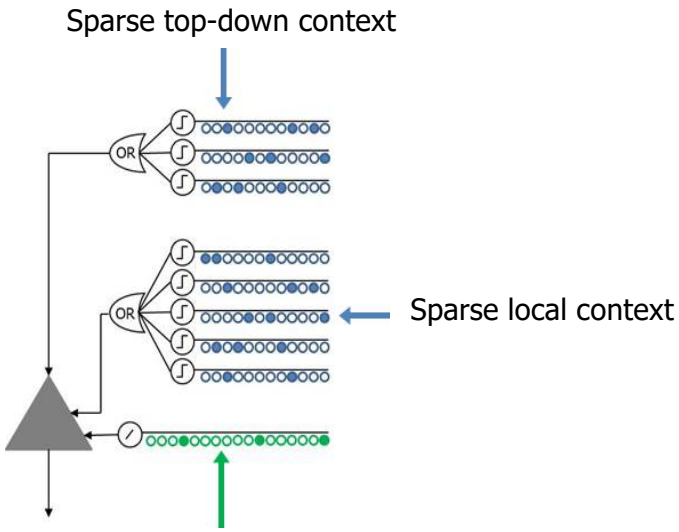


Sparse local patterns

Sparse feedforward patterns

Pyramidal neuron

Continuous learning with active dendrites



Model pyramidal neuron

(Hawkins & Ahmad, 2016)

Simple localized learning rules

When a cell becomes active:

1) If a segment detected pattern, reinforce that segment
2) If no segment detected a pattern, grow new connections on new dendritic segment

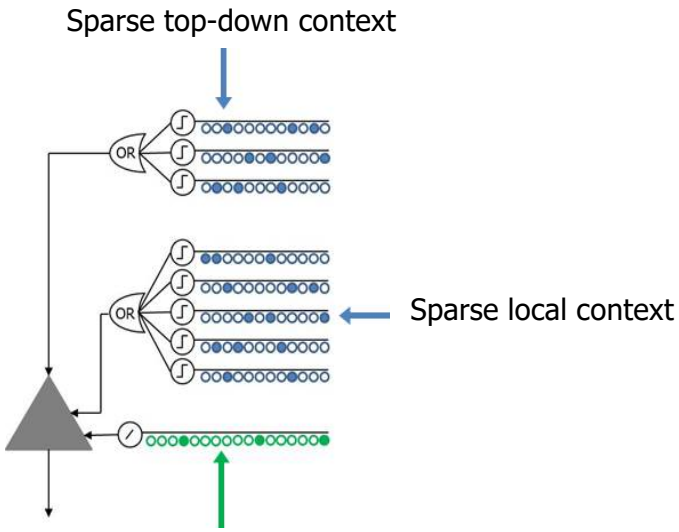
If cell did not become active:

1) If a segment detected pattern, weaken that segment

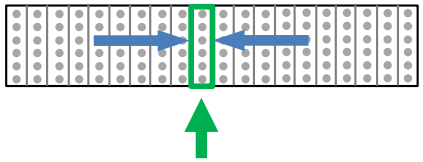
- Learning consists of growing new connections

- Neurons learn continuously but since patterns are sparse, new patterns don't interfere with old ones

Recurrent layer forms an unsupervised predictive system



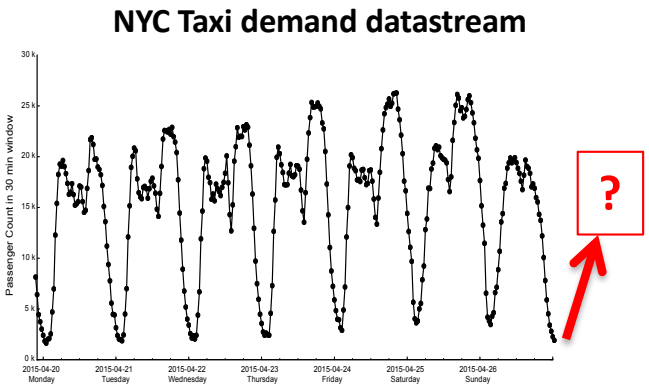
Model pyramidal neuron



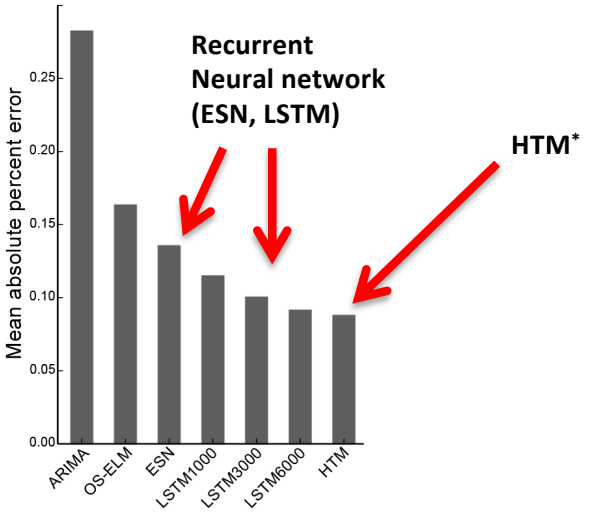
1) Associates past activity as context for current activity
2) Automatically learns from prediction errors
3) Learns continuously without forgetting past patterns
4) Can learn complex high-Markov order sequences

(Hawkins & Ahmad, 2016)

Continuous learning with streaming data sources



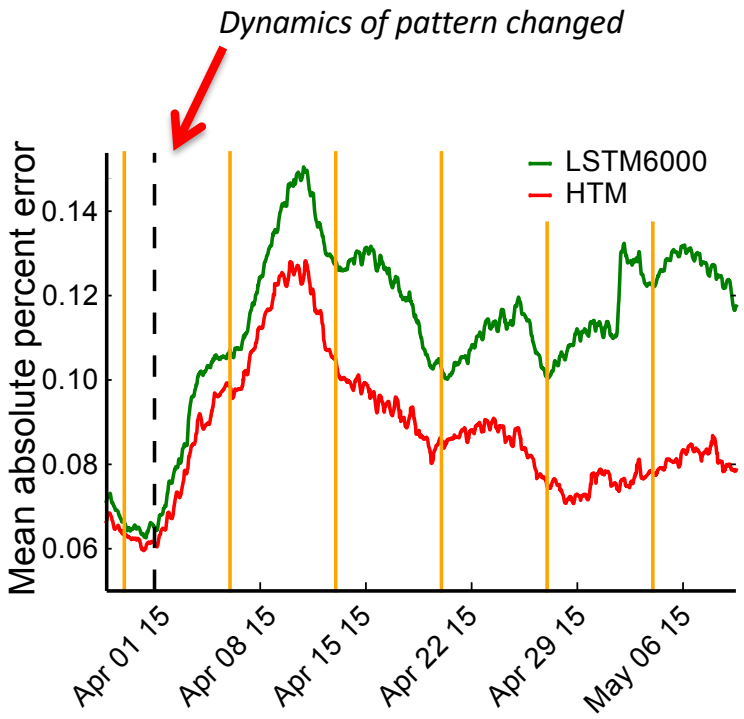
Source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml



(Cui et al, Neural Computation, Nov 2016)

HTM = Hierarchical Temporal Memory

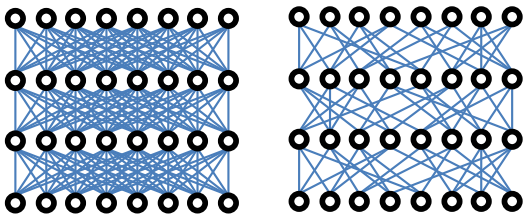
Adapts quickly to changing statistics



(Cui et al, Neural Computation, 2016)

Implications of sparsity for machine learning

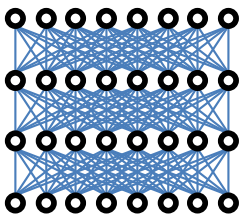
CNN

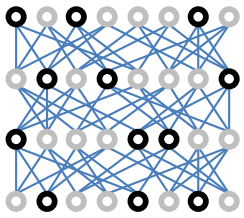


- Sparse connections

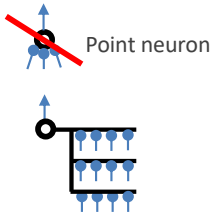
Implications of sparsity for machine learning

CNN

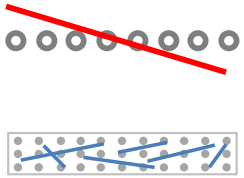




- Sparse connections
- Sparse activations



- Compartmental neurons (aka active dendrites)



- Intralaminar circuits

Improved robustness
no loss of accuracy
Large performance gains on hardware
energy – speed – size

Continuous learning
quickly reacts to changes

Unsupervised learning
predictive learning without teacher signal

Neuroscience principles inform each of these steps
Need to design new learning architectures that exploit these properties

Thanks! Questions:
sahmad@numenta.com

Code:
https://github.com/numenta/nupic.torch

Papers:
http://numenta.com/papers