A neuro-inspired computing module for unsupervised continual learning

Constantine Dovrolis
School of Computer Science,
Georgia Tech

James Smith

Seth Baer

Zsolt Kira

Supported by the Lifelong Learning Machines (L2M) program of DARPA/MTO

Disclaimers & Apologies

- This work is neuro-inspired but it is not neuromorphic
  - Not at the level of neurons & synapses

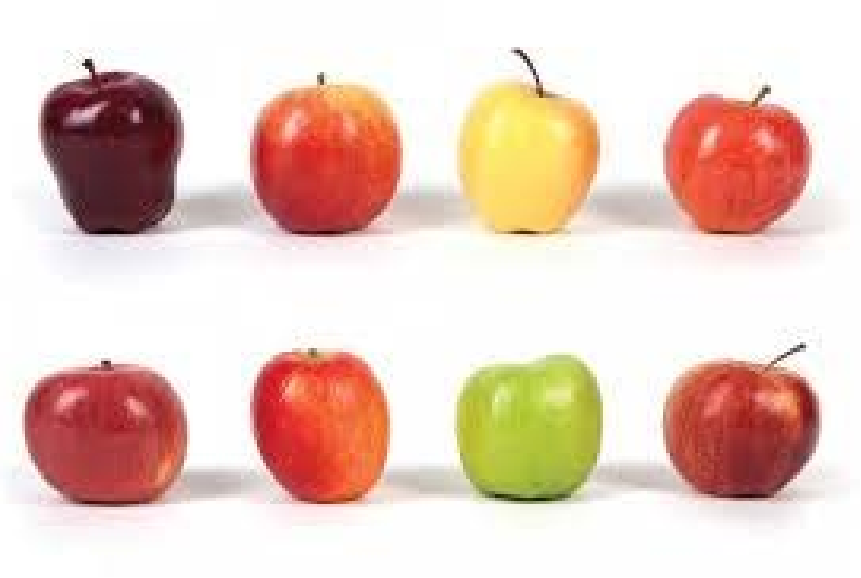
- Strong opinion: Neuroscience can provide great insights & ideas to ML/AI beyond the level of neurons and synapses
  - Network neuroscience
  - Cognitive neuroscience
  - Evolutionary neuroscience
  - Neurology & psychiatry

Unsupervised Continual Learning (UCL)

- Learn efficient data representations from stream of unlabeled data
  - No labeled data in stream
  - No specific task or reward signal given
  - Class/data distribution is non-stationary
  - No storage/replay of past inputs

Occasionally, create associations between learned representations and named concepts/classes

- Given a generalized representation, how can we label it?
  - Labeled data does not change representations



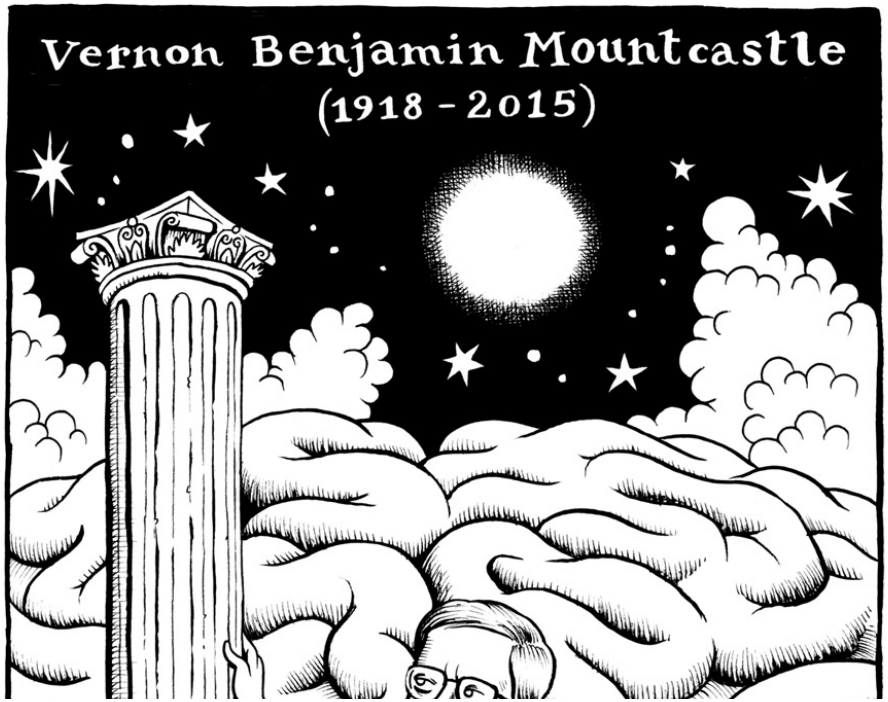
What UCL is NOT

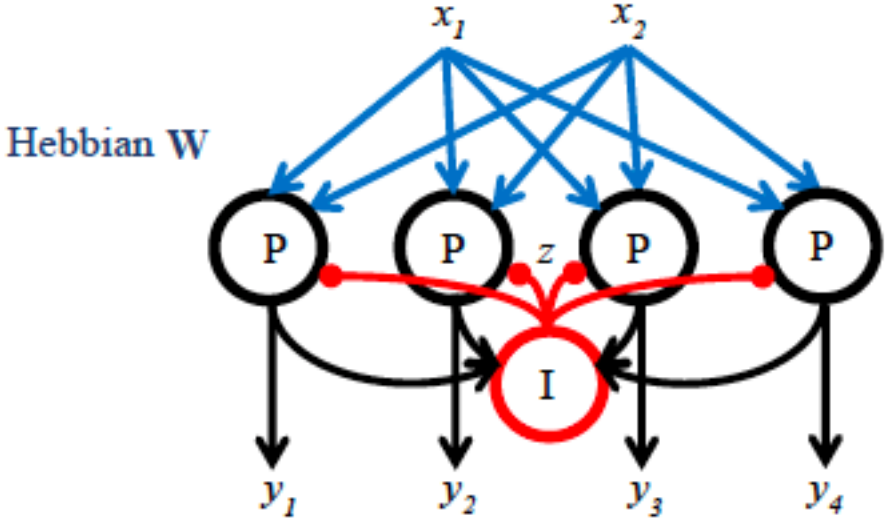
- Representation learning
  - Typically does not operate in online context
- Supervised continual learning
  - Learns continually from stream of labeled data
- Semi-supervised learning
  - Learns from both labeled and unlabeled data
- Self-supervised learning
  - Relies strictly on prediction – does not learn associations

Arguably, UCL is how natural organisms learn most of the time

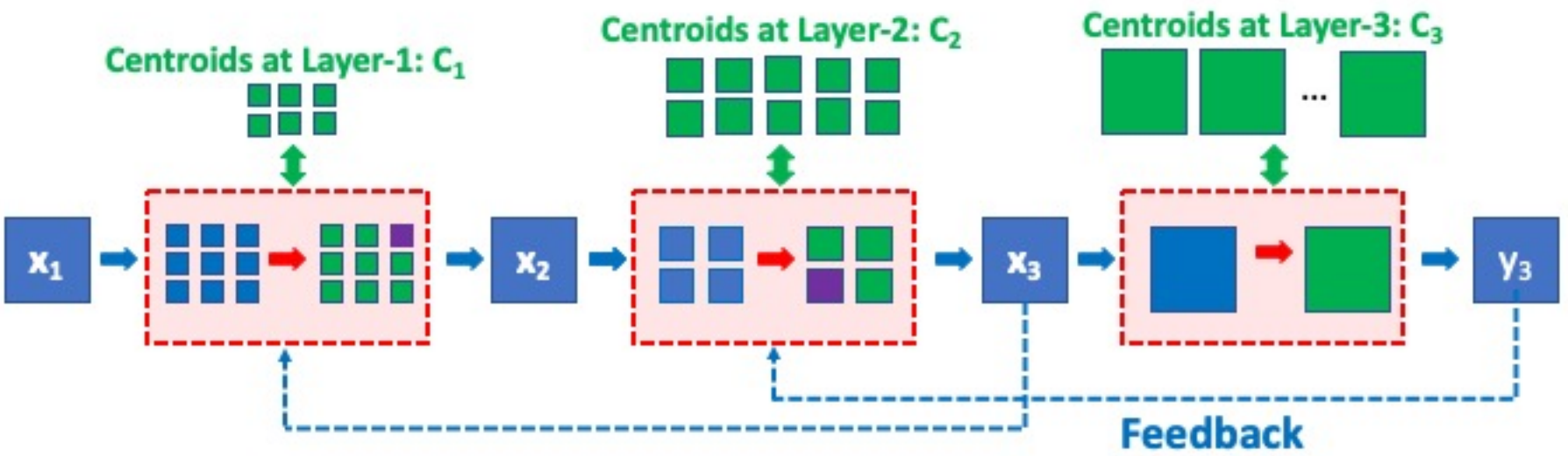
What we borrow from neuroscience

1. *Hierarchical organization* between cortical areas with feedforward feedback connections
2. *Mountcastle:* cortical columns perform an (unknown) common function across cortex
3. The *"canonical cortical circuit"* of Douglas & Martin
4. *Intra-column recurrent circuits* and role of *inhibitory neurons*
5. Friston: cortical columns perform *predictive coding*
6. Cortical columns appear to learn *"prototypes"* – see Tanaka '96 for monkey IT experiments
7. E/I recurrent circuits (similar to L4) can perform *k-means online clustering* (Pehlevan et al. '18)





Self-Trained Associative Memory (STAM) Architecture Overview



Intrinsic dimensionality reduction

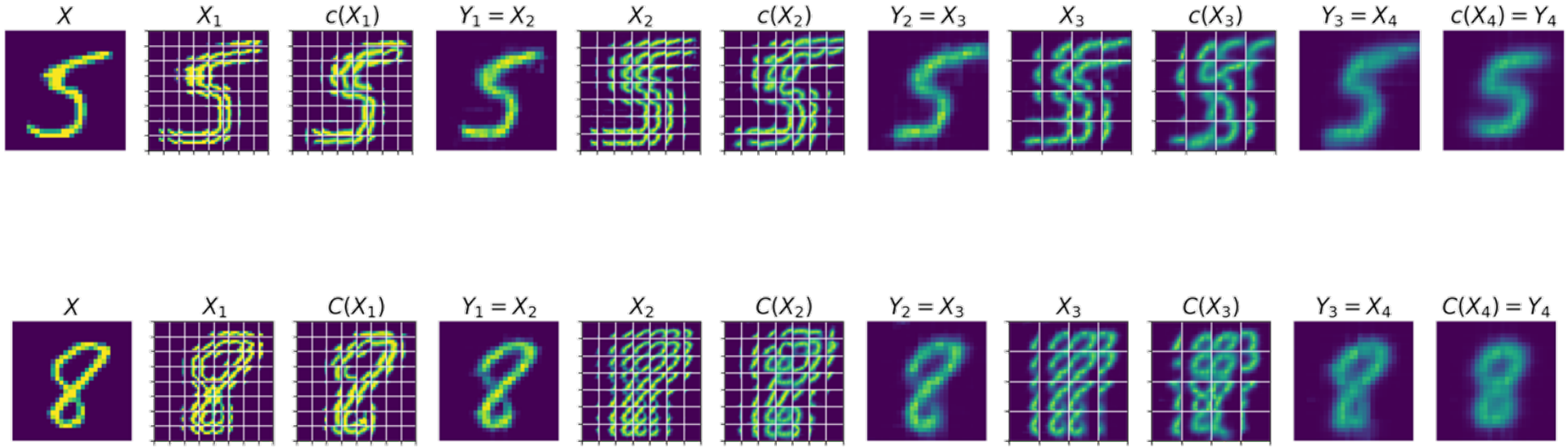
Online learning of centroids

Novelty detection

Forgetting mechanism

Prediction & feedback

Illustration of how STAMs work



- Gradual reduction of intrinsic dimensionality
  - If there are N centroids at a layer, and that layer consists of M STAMS (or RFs), the output image at that layer can take N^M possible values
  - Deeper in hierarchy: N increases, M decreases (so that N^M decreases)

Key equations

- Centroid selection:

$$c(x_{i,m}) = \arg \min_{j=1\ldots|C_i|} ||x_{i,m} - w_{i,j}||$$

- Centroid online learning:

$$w_{i,j} = \alpha\, x_{i,m} + (1-\alpha)\, w_{i,j}, \text{ when } c(x_{i,m}) = j$$

- Centroid novelty detection:

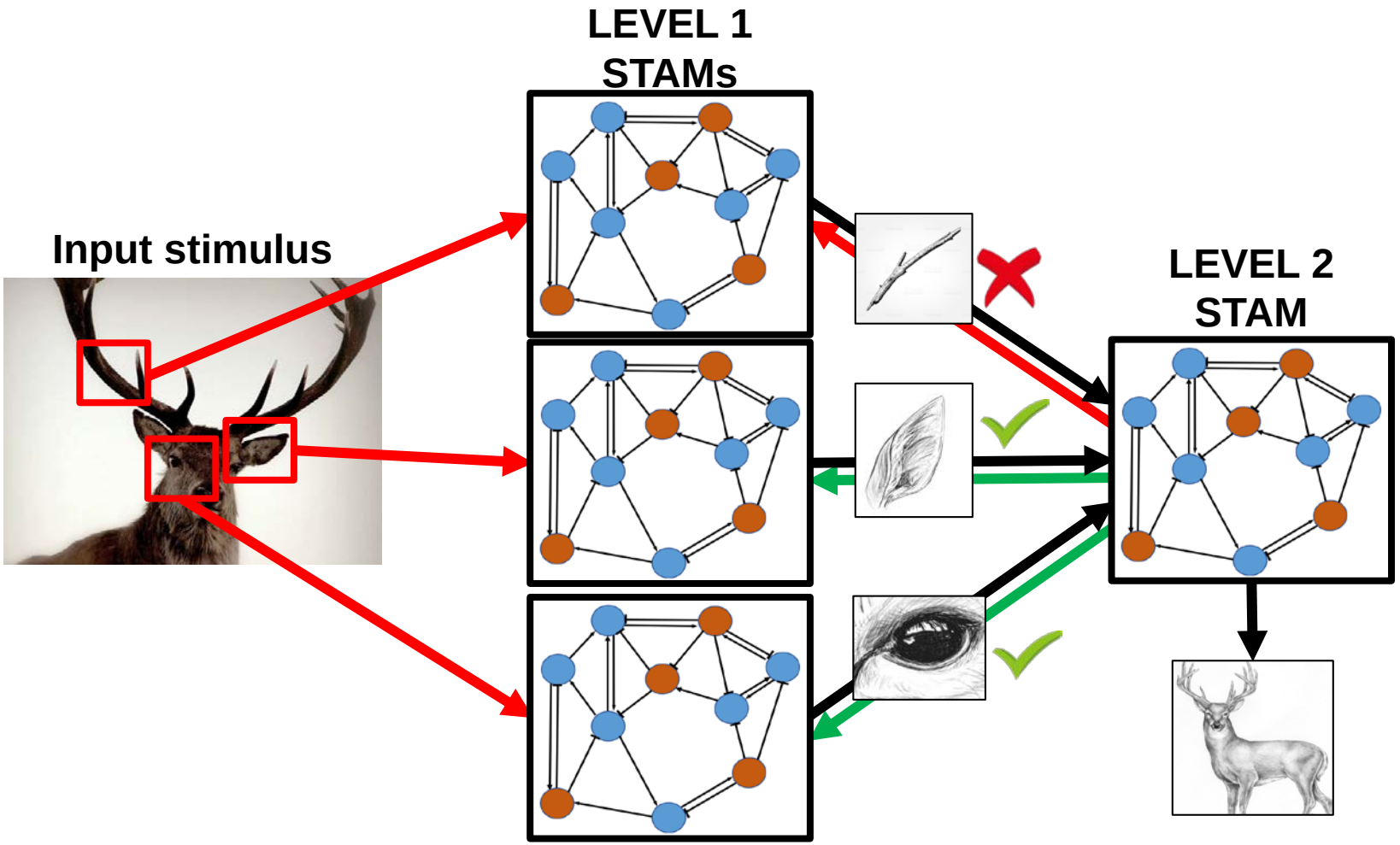
$$\mu_j = \alpha\, ||x_{i,m} - w_{i,j}|| + (1-\alpha)\, \mu_j$$

$$\hat{\sigma}_j = \alpha\, |\, ||x_{i,m} - w_{i,j}|| - \mu_j\, | + (1-\alpha)\, \hat{\sigma}_j$$

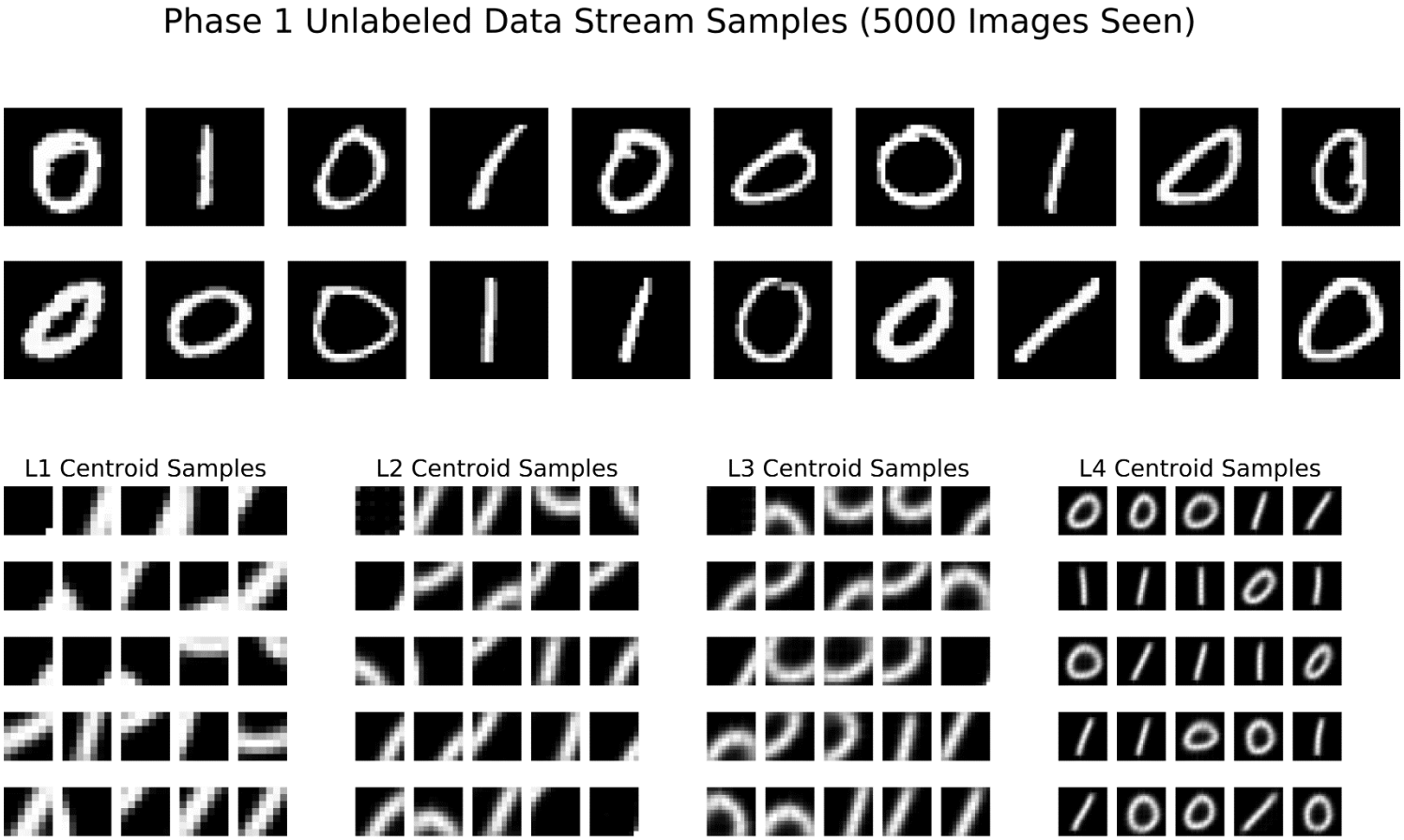
$$||x_{i,m} - w_{i,j}|| > \mu_j + 3\, \hat{\sigma}_j$$

- Centroid forgetting:
  - when capacity of N centroids at a layer is reached, forget *Least Recently Used* centroid

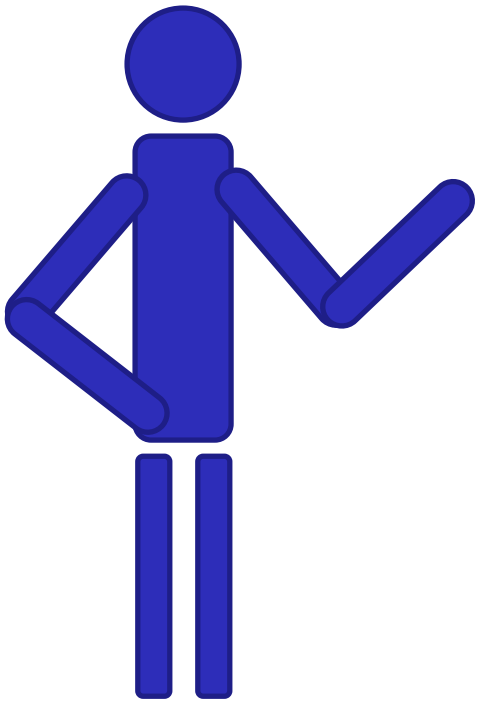
The role of top-down predictions



Phase-1: stream of 0s & 1s



Phase-1 evaluation given a couple of labeled examples



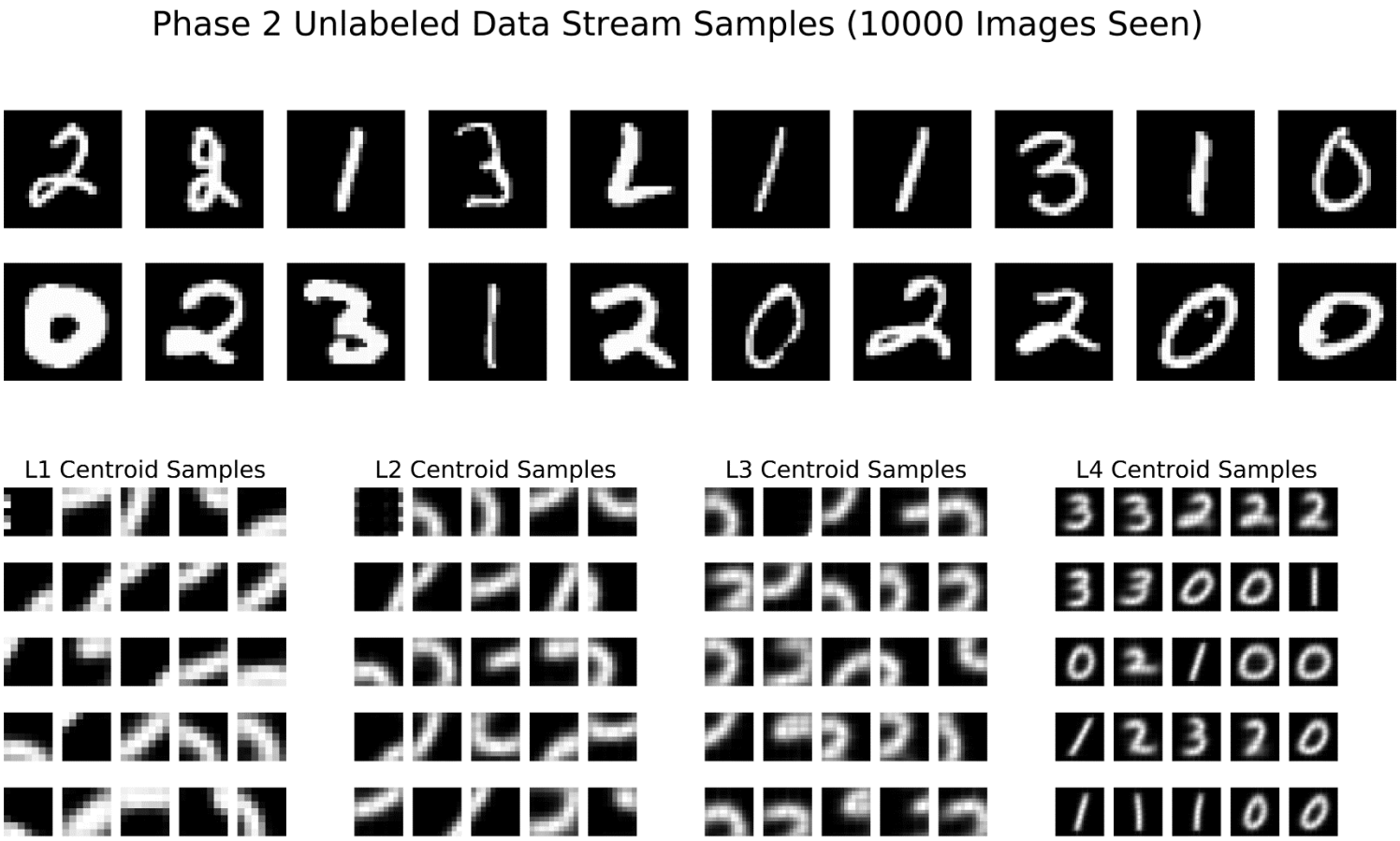
Clue:

zero one

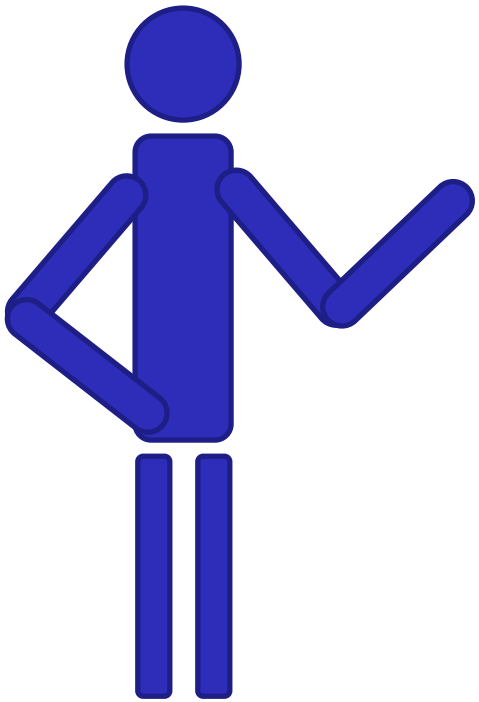
zero one

Quiz:

Phase-2: stream of {0,1,2,3}



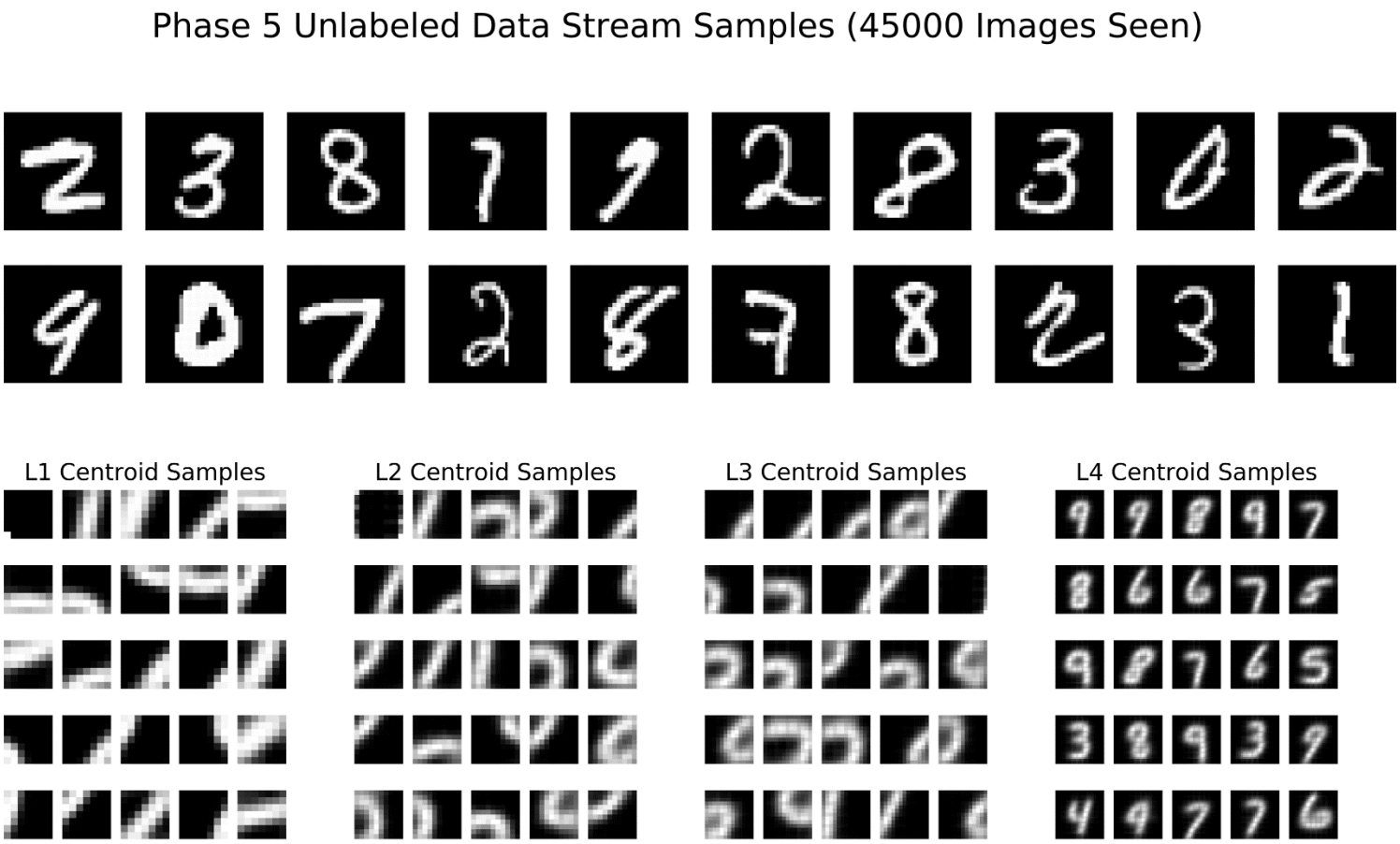
Phase-2 evaluation given a couple of labeled examples



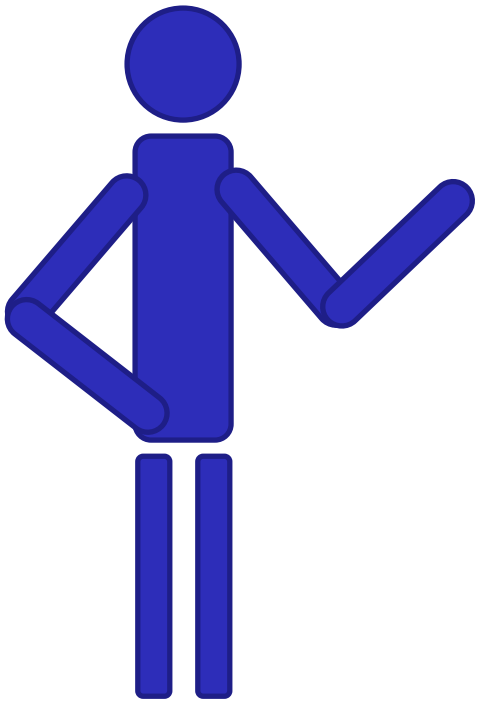
Clue:

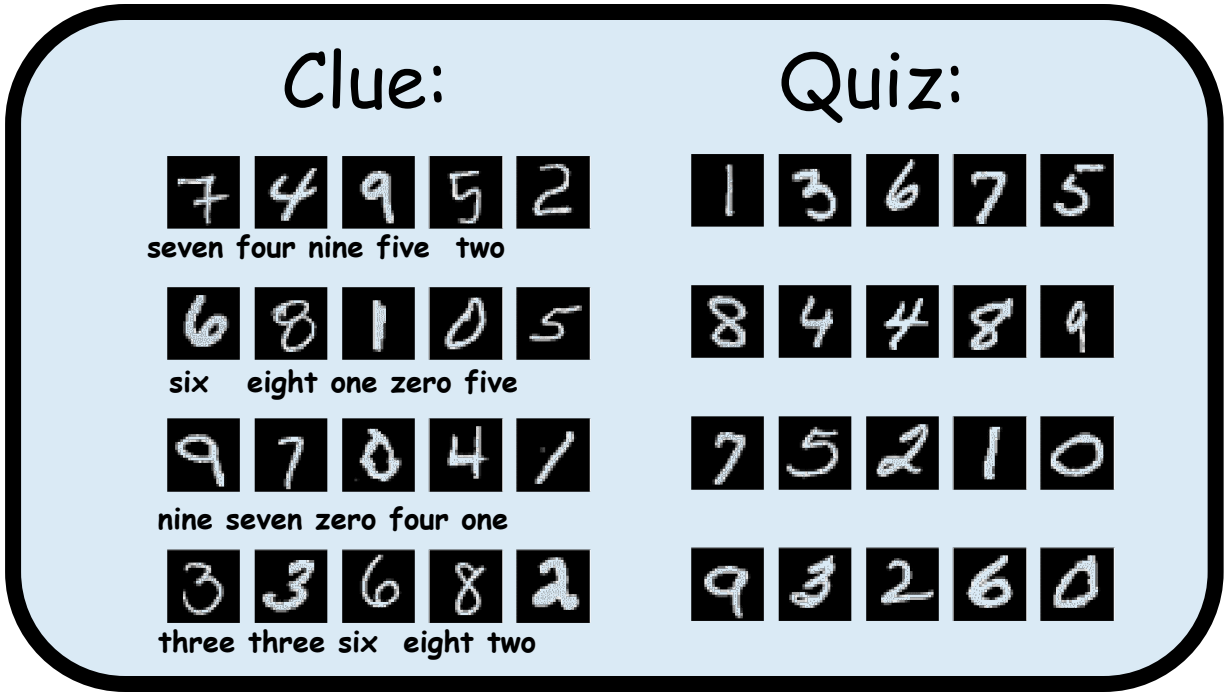
Quiz:

Phase-5: stream of {0,1,..9}

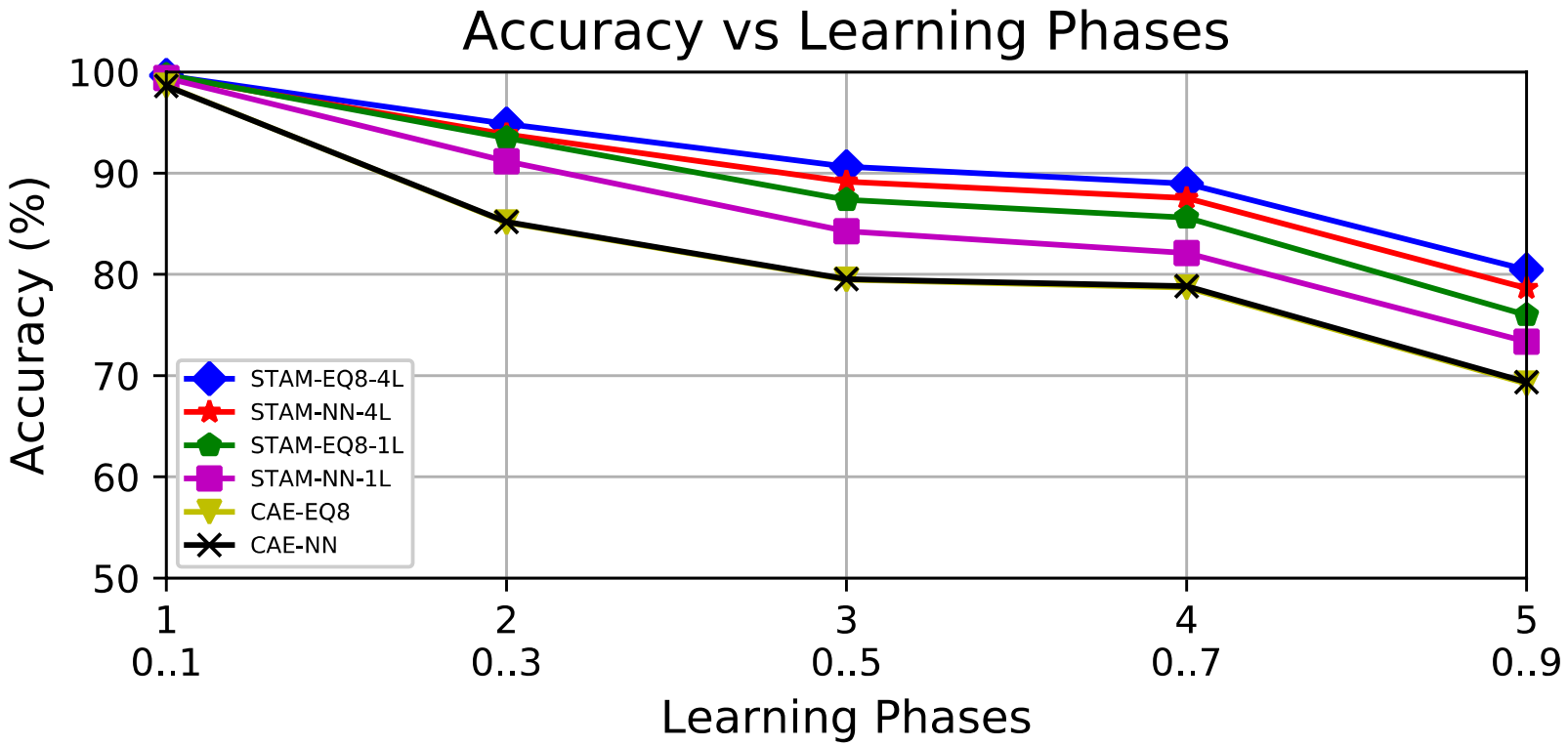


Phase-5 evaluation given a couple of labeled examples per class



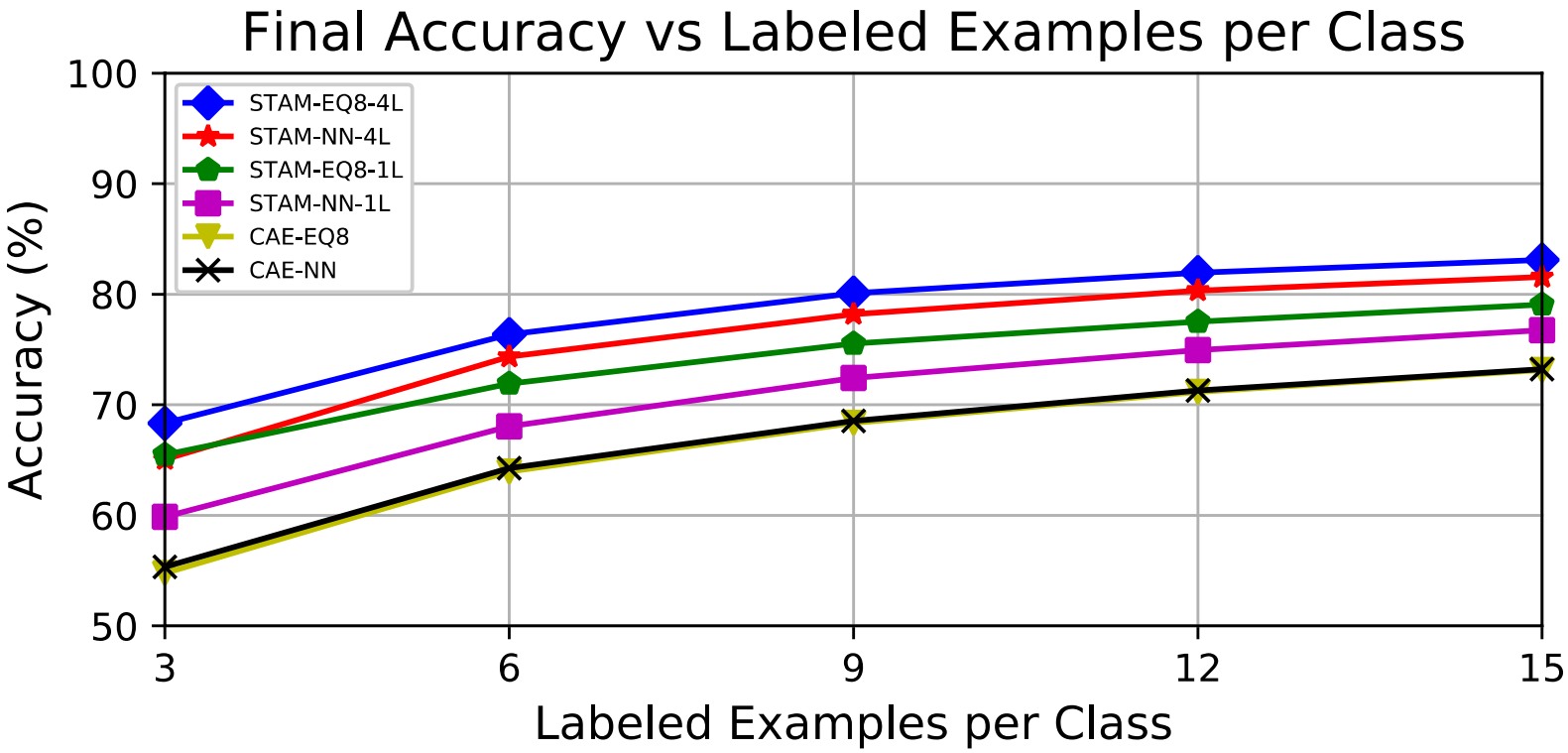


Accuracy as the system learns more classes



STAM hierarchy performs best for the entire scenario

Accuracy as function of number of labeled examples



Next Steps

- Work with natural image datasets and video
- Compare with additional baseline methods compatible with UCL problem
- Generalize classification to use centroids from any layer in hierarchy
- Apply in video and timeseries problems (where prediction and top-down connections can play major role)