Adaptive Resonance Theory (ART) Networks

Proposed by
Profs. Stephen Grossberg &
Gail Carpenter
Boston University

ART Network Varieties

- Several varieties:
  - ART1: Discrete (e.g. binary) patterns
  - ART2: Continuous patterns …
  - ARTMAP: supervised learning
  - Fuzzy Art: Fuzzy version of ART1

- All work by **competitive learning** and a type of **clustering**, represented by stored prototypes, with other nuances

Stability-Plasticity Dilemma

- **stability**: Recognized patterns should be insensitive to noise.
- **plasticity**: System should be capable of learning new patterns.
- The conflict between these is one of the things that ART tries to resolve.

ART Map vs. BackProp

	Predictive ART	Back Propagation
supervised	yes	yes
self-organizing	yes	no
real-time	yes	no
self-stabilizing	yes	no
learning:	fast or slow match	slow mismatch

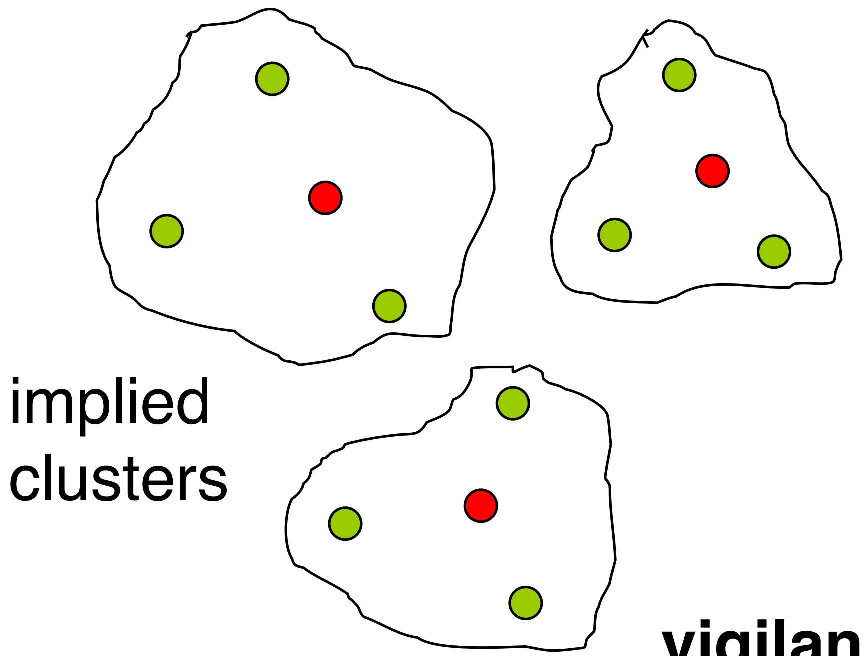
Some ART Applications

- Disease identification (HMC Math Clinic)
- Tech support email automation (text similarity) (HMC CS Clinic)
- Satellite data anomaly detection (HMC CS Clinic)
- Music recognition
- Distinguishing poisonous vs. edible mushrooms
- Modeling biological neural processes

ART Networks

- Originally biologically motivated by an ODE model
- Models short- and long-term memory
- Combine supervised and unsupervised (competitive, clustering)
- Dynamically create new categories, controllable by an attention parameter called “vigilance”

Competitive Learning in ART

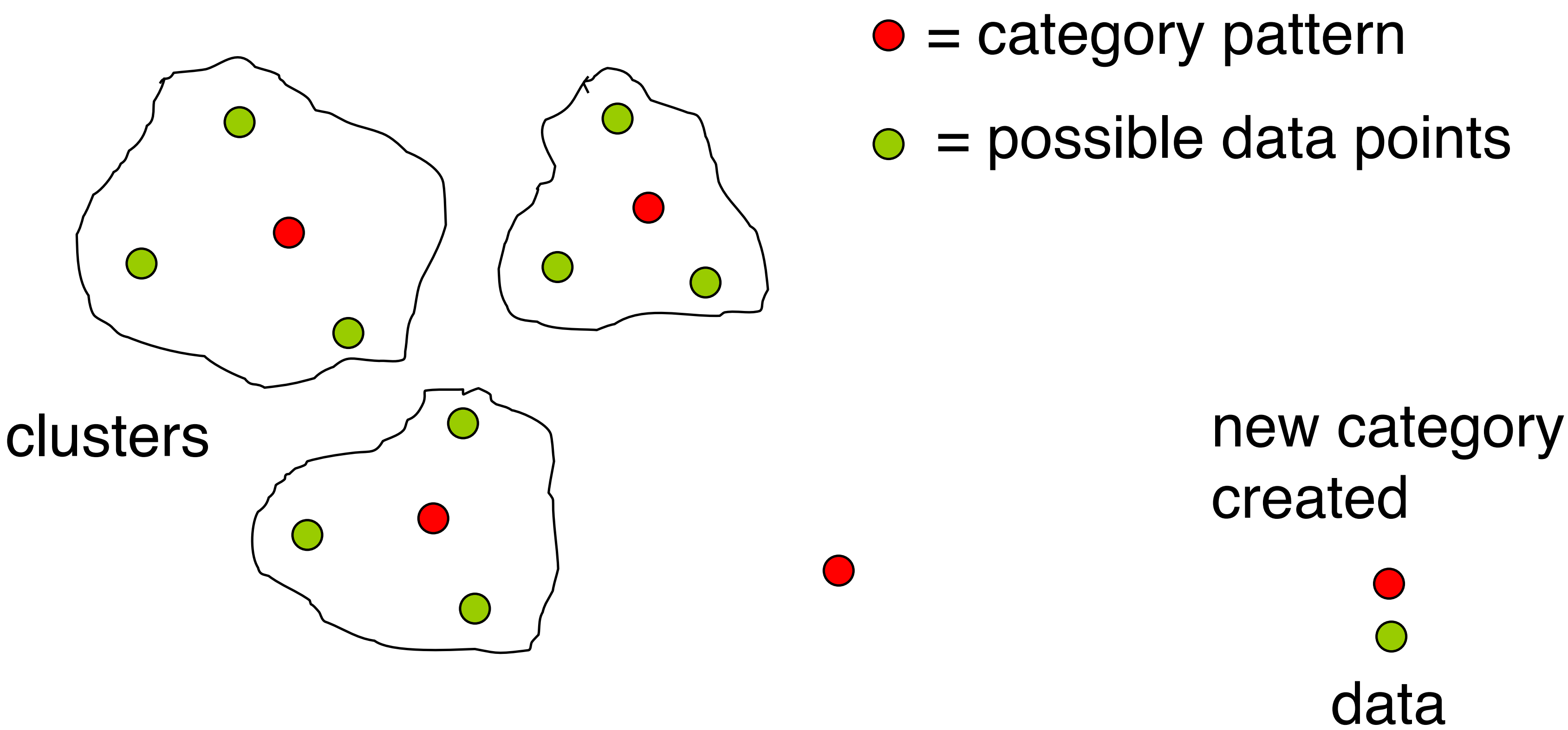


● = category pattern

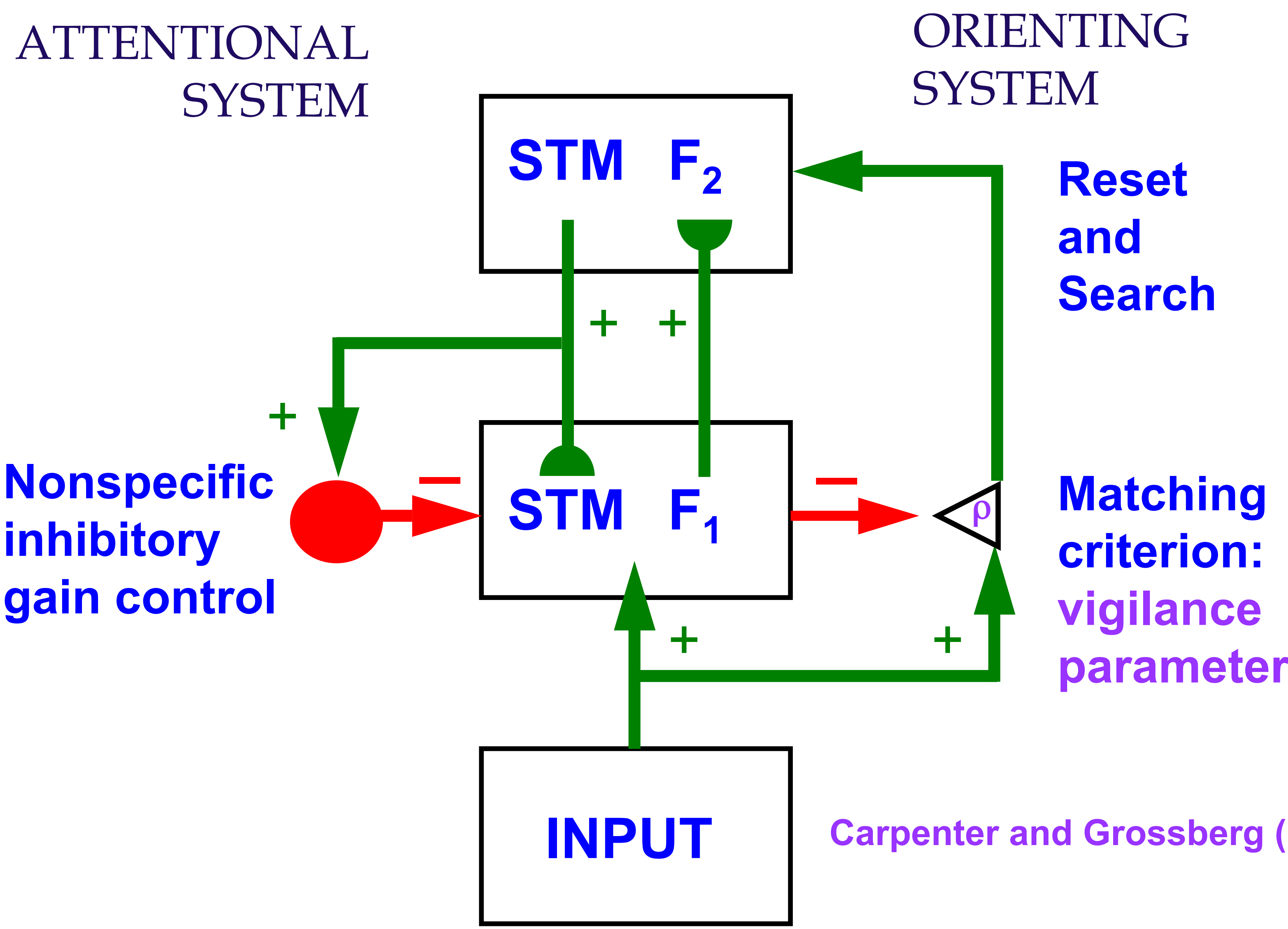
● = possible data points

vigilance threshold:
nearest prototype
too far from data
to be considered the correct category

Competitive Learning in ART

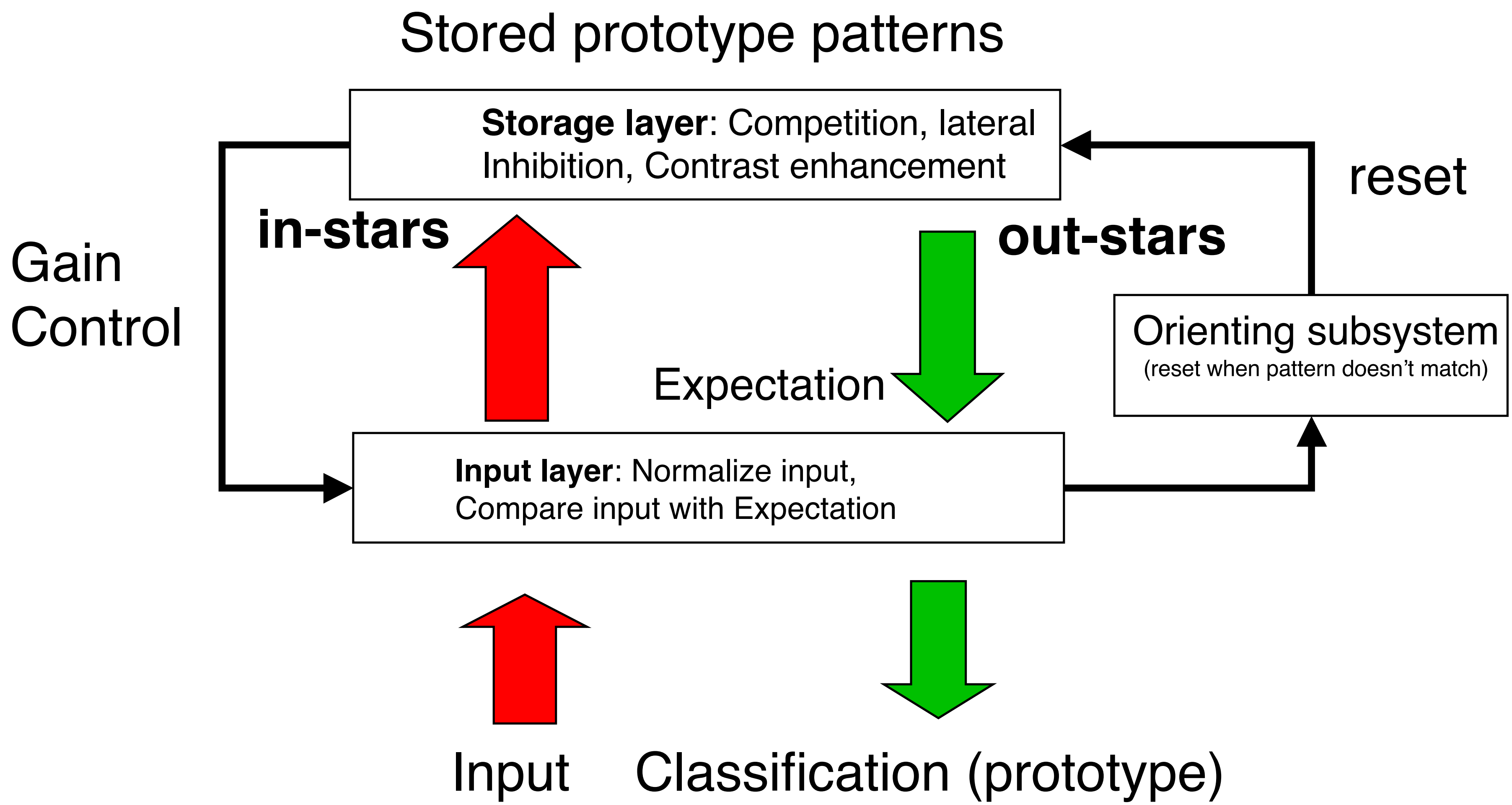


ART 1 MODEL



Carpenter and Grossberg (1987)

ART: instars and outstars



ART HYPOTHESIS TESTING AND LEARNING CYCLE

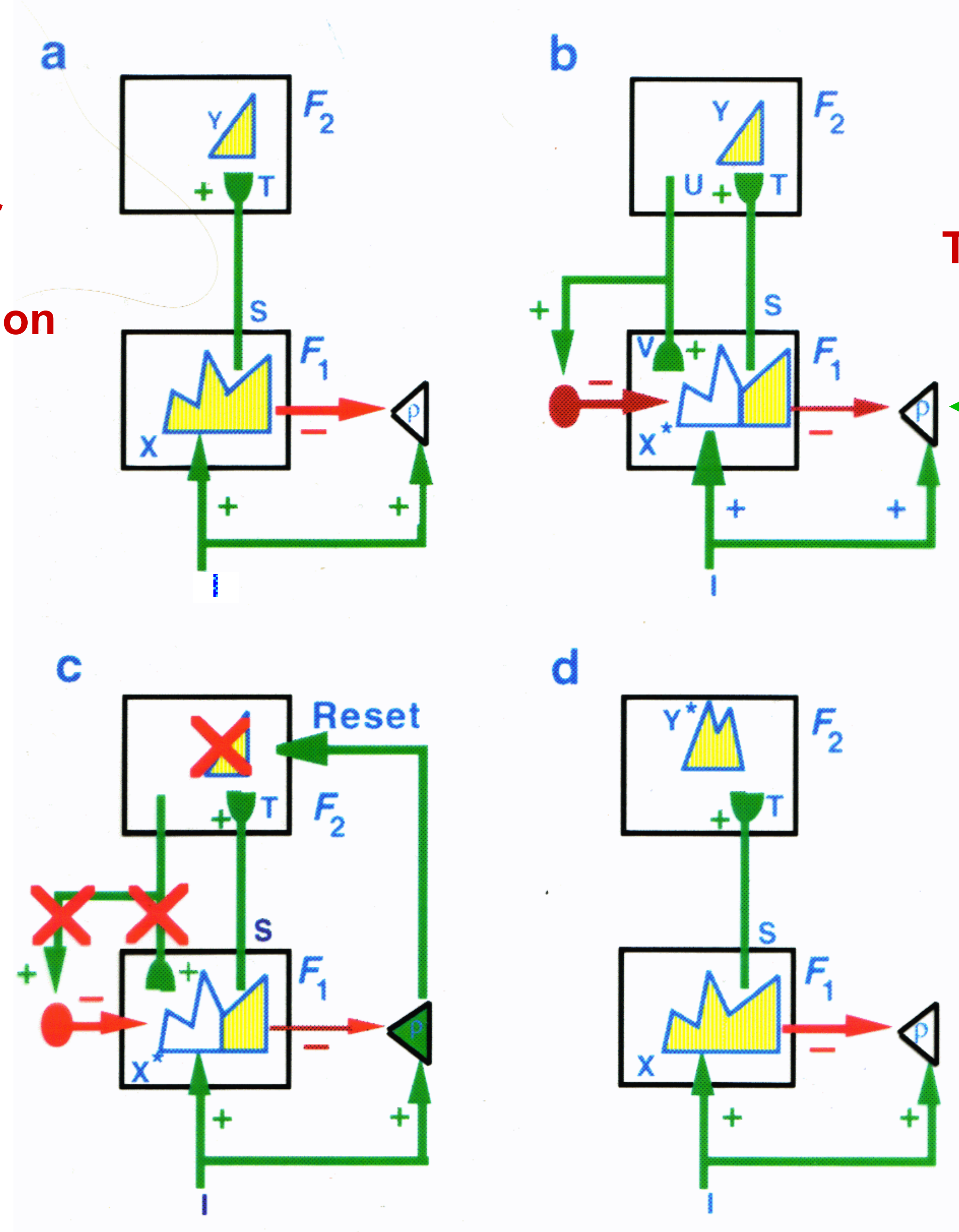
Choose category, or symbolic representation

Test hypothesis

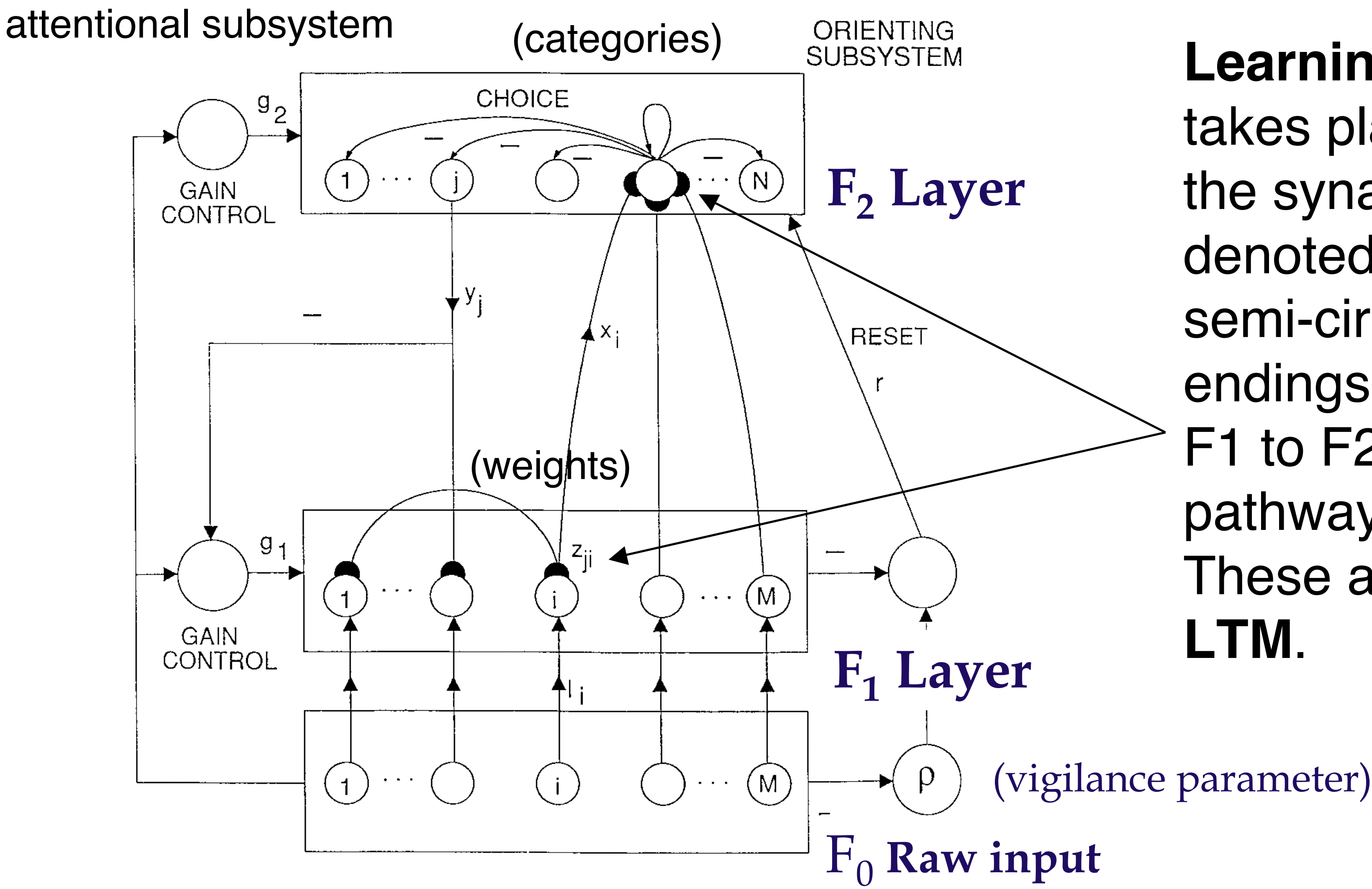
Vigilance

Mismatch reset

Choose another category



ART1 Detail



Learning takes place at the synapses denoted by semi-circular endings in the F1 to F2 pathways. These are **LTM**.

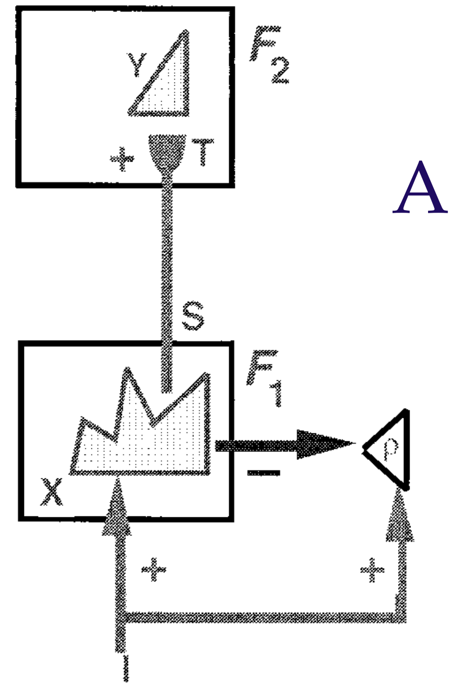
ART recognition search in more detail (1 of 4)
[Grossberg and Merrill (1996)]

1. The input pattern I is instated across the feature detectors at level F_1 as a short-term memory (STM) activity pattern **X** represented by the hatched pattern across F_1.

2. Input I also **nonspecifically** activates the orienting subsystem A

3. Pattern X both inhibits A and generates the output pattern S from F_1.

4. Pattern S is multiplied by long term memory (LTM) traces and added at F_2 nodes to form the input pattern T, which activates the STM pattern **Y** across the categories coded at level F_2.

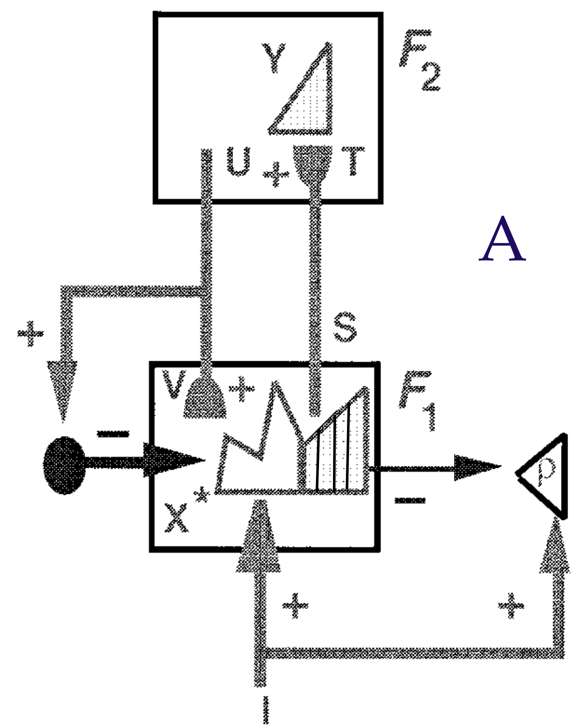


ART recognition search detail (2 of 4)

5. Pattern Y generates the top-down output pattern U, which is multiplied by top-down LTM traces and added at F_1 nodes to form the **prototype pattern V** that encodes the **learned expectation** of the active F_2 nodes.

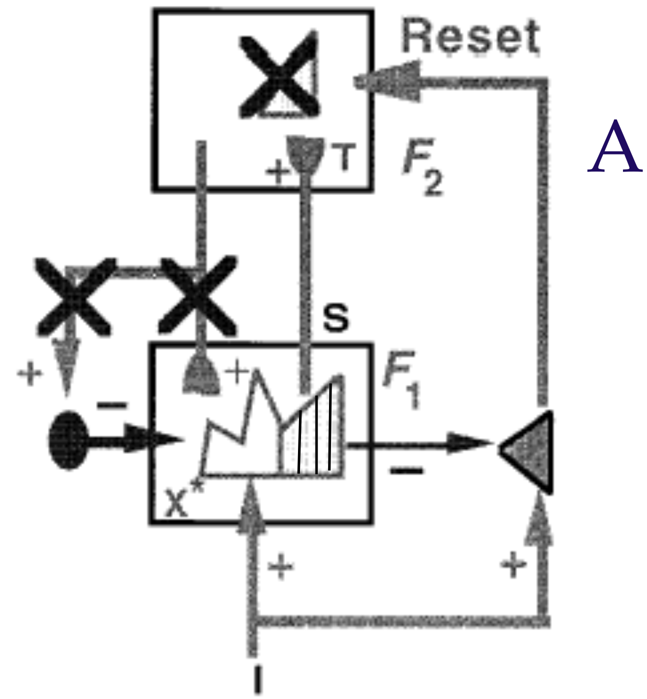
6. If **V mismatches I** at F_1, then **a new STM pattern** X* is generated at F_1. X* is represented by the **hatched pattern** and includes the features of I that are **confirmed** by V. Inactivated nodes corresponding to unconfirmed features of X are unhatched.

7. The reduction in total STM activity, which occurs when X is transformed into X* causes a **decrease in the total inhibition** from F_1 to A.



ART recognition search detail (3 of 4)

8. If inhibition decreases sufficiently [vigilance condition not met], **A** releases at a nonspecific arousal wave to F_2, which **resets** the STM pattern Y at F_2.



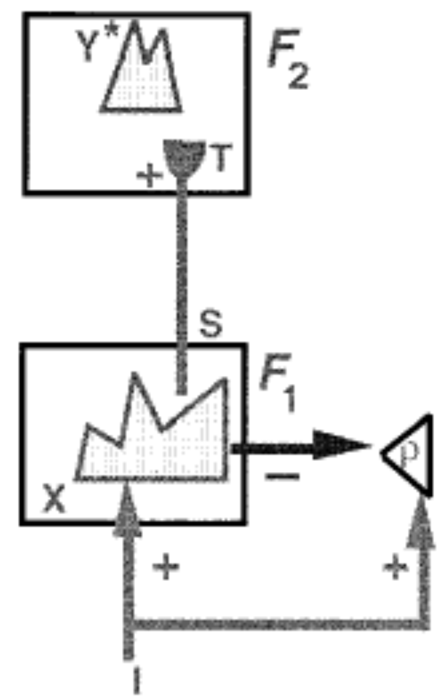
ART recognition search detail (4 of 4)

10. After Y is inhibited, its top-down prototype signal is eliminated, and **X can be reinstated** at F_1.

11. **Enduring traces** of the prior reset lead X to activate a **different** STM pattern Y* at F_2.

12. If the top-down prototype due to Y* also mismatches I at F_1, then the **search** for an appropriate F_2 code continues ***until a more appropriate*** F_2 representation is selected.

13. Then an attentive **resonance** develops and learning of the attended data is initiated.



ART Learning

- The **resonant state**, rather than bottom-up activation, drives the learning process (hence "adaptive resonance" theory):
  - "resonance" = **mutual reinforcement** between input and storage layers
  - "adaptive" = weights are adjusted when **resonance** occurs
- ART systems **learn prototypes** rather than exemplars because the attendant **feature vector *X****** rather than the input **exemplar itself** is learned.

ART1 Viewpoints

- Two kinds of explanations:
  - neural - as previously presented
  - algorithmic

- The first is more complicated, since it involves neural explanations for the control aspects of the algorithmic approach.

ART *Algorithmic* View

- Input pattern presented to input layer.
- Storage layer indicates tentative hypothetical classification.
- Input layer decides if hypothetical is **close enough**; if so, done.
- If not, storage layer indicates alternate hypothesis.
- The above two steps are repeated until the hypothetical classification is accepted.
- *All* hypotheses could be rejected; in this case, a **new** class is created in the storage layer.

Algorithm ART1:

Initialize each $t_{\ell,j}(0) = 1$, $b_{j,\ell}(0) = \frac{1}{n+1}$;

while the network has not stabilized, do

1. **Let** A **contain all nodes;**
2. **For a randomly chosen input vector** x **, compute** $y_j = b_j \cdot x$ **for each** $j \in A$.
3. **repeat**

   (a) **Let** j^* **be a node in** A **with largest** y_j.

   (b) **Compute** $s* = (s_1^*, \ldots, s_n^*)$ **where** $s_\ell^* = t_{\ell,j^*}\ x_\ell$;

   (c) **If** $\frac{\sum_{\ell=1}^{n}\ s_\ell^*}{\sum_{\ell=1}^{n} x_\ell} \leq \rho$ **then remove** j^* **from set** A

   else associate x **with node** j^* **and update weights:**

$$b_{j^*,\ell}(\text{new}) \ = \frac{t_{\ell,j^*}(\text{old})\ x_\ell}{0.5 + \sum_{\ell=1}^{n}\ t_{\ell,j^*}(\text{old})x_\ell}$$

$$t_{\ell,j^*}(\text{new}) \ = t_{\ell,j^*}(\text{old})x_\ell$$

   until A **is empty or** x **is associated with some node;**

4. **If** A **is empty, create new node with weight vector** x;

end-while.

ART1 Issues

- Subset-Superset dilemma:
  - If one pattern is **contained in** another, then a given input may have the same inner product with two different prototypes.
  - Resolvable by **normalizing** the prototypes.

Example of Subset/Superset Dilemma

Suppose that $\mathbf{W}^{1:2} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ i.e. the prototypes are ${}_1\mathbf{w}^{1:2} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ ${}_2\mathbf{w}^{1:2} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Then ${}_1\mathbf{w}^{1:2}$ is a subset of ${}_2\mathbf{w}^{1:2}$, because ${}_2\mathbf{w}^{1:2}$ has a 1 wherever ${}_1\mathbf{w}^{1:2}$ has a 1.

If the output of layer 1 is $\mathbf{a}^1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ then the input to Layer 2 will be

$$\mathbf{W}^{1:2}\mathbf{a}^1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Both prototype vectors have the **same inner product** with $\mathbf{a}^1$, even though the first prototype is identical to $\mathbf{a}^1$ and the second prototype is not.

Subset/Superset Resolution

Normalize the prototype patterns.

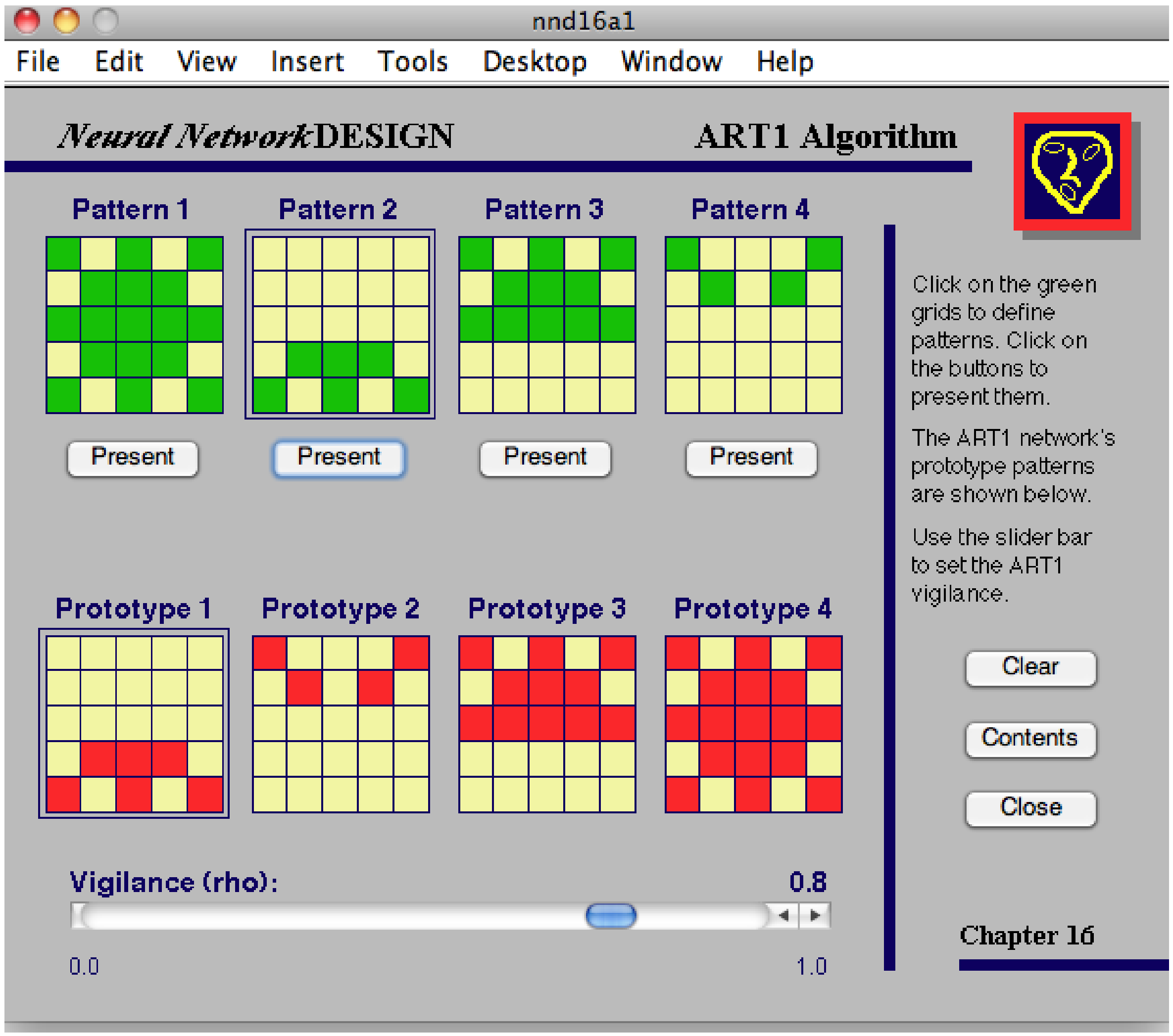
$$\mathbf{W}^{1:2} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

$$\mathbf{W}^{1:2}\mathbf{a}^{1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{2}{3} \end{bmatrix}$$

Now we have the desired result; the first prototype has the larger inner product with the input.

Matlab ART1 Demo (nnd16a1)

Increasing **vigilance** causes the network to be **more selective**, introducing a new prototype when the match is weak.

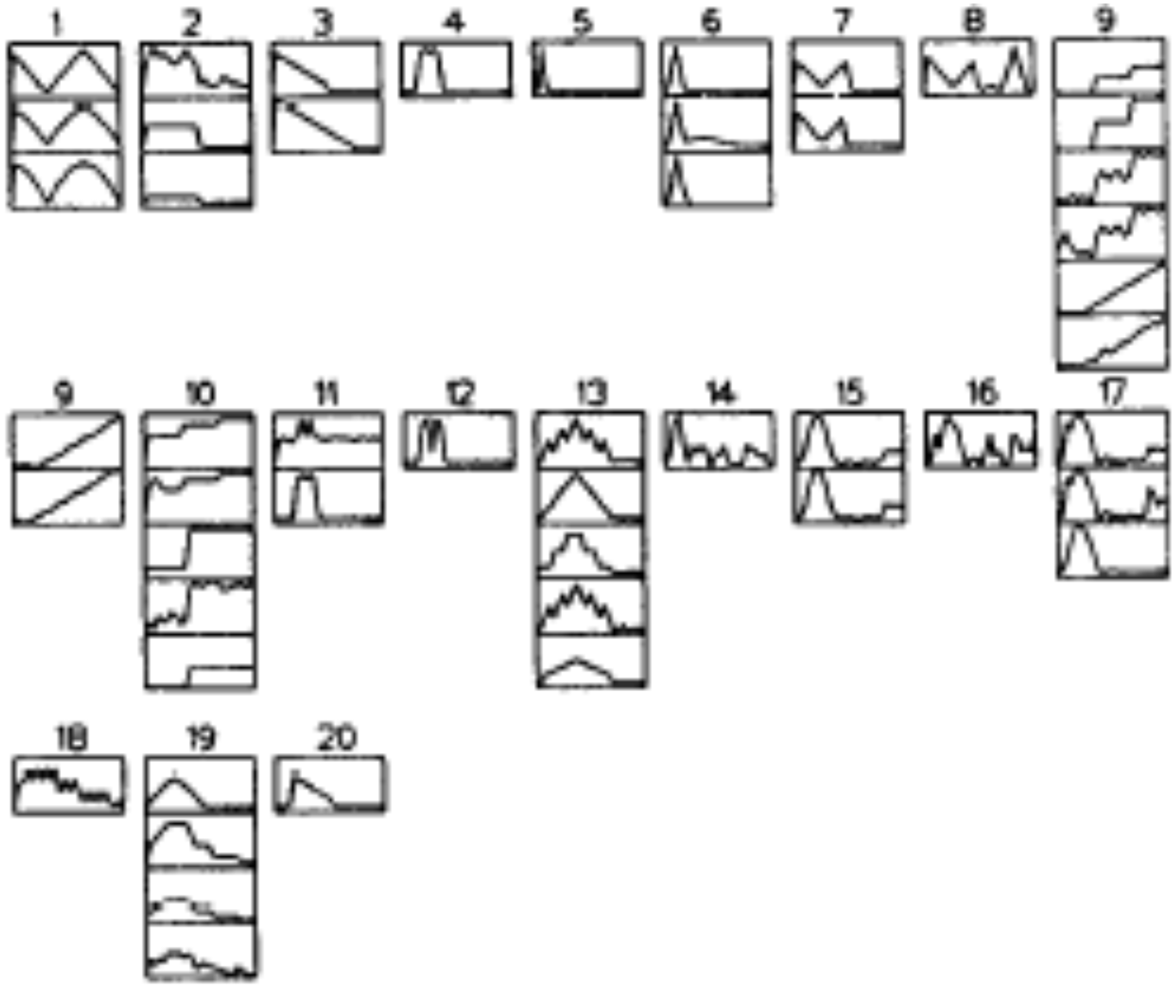


Try different patterns

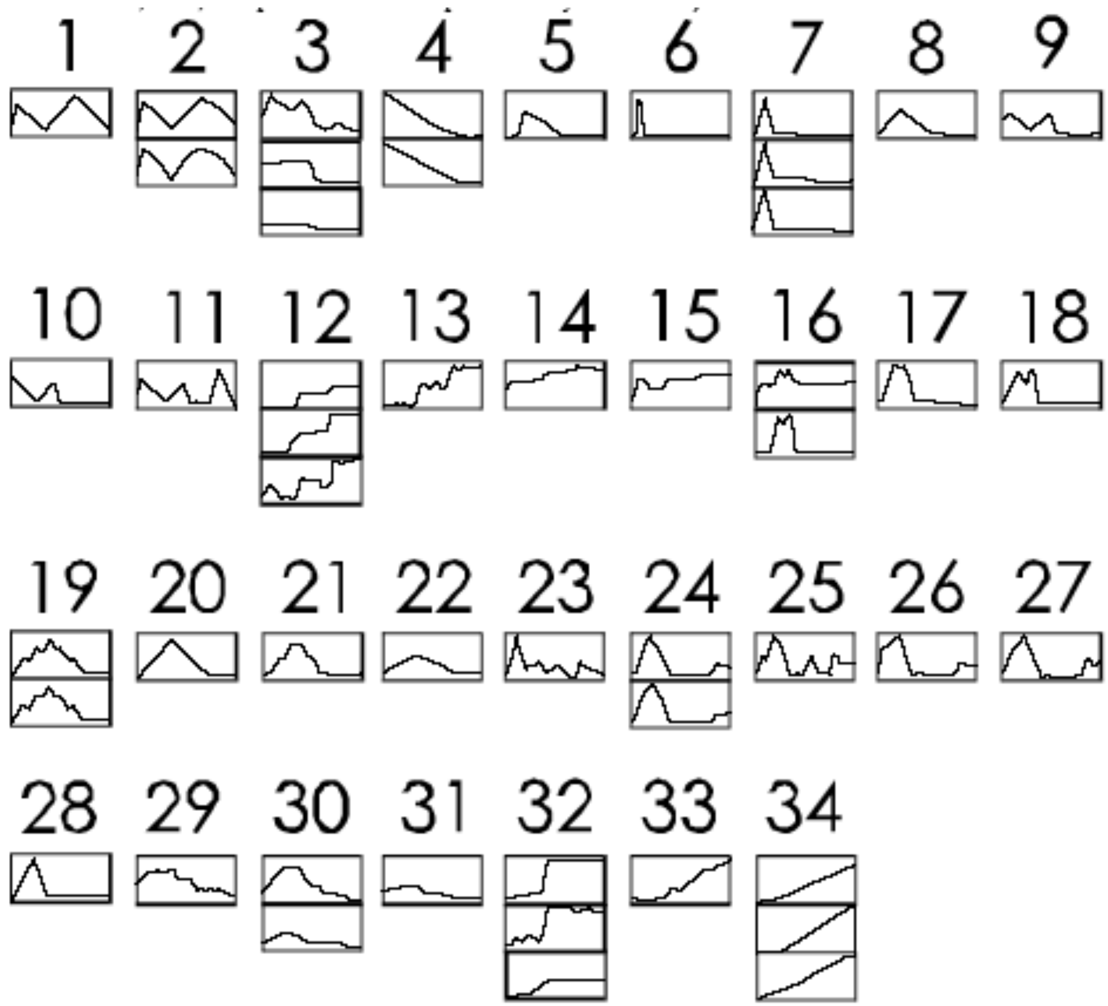
ART2

- ART2 allows continuous-valued patterns
- whereas ART1 is limited to discrete-valued ones

ART2 Clustering (at low vigilance)



ART2 Clustering (at higher vigilance)



Fuzzy ART Supplants ART2

- ART2 was evidently too complicated.
- Fuzzy-ART, based on Fuzzy Logic, was its replacement.
- Fuzzy Logic uses the min operator as conjunction. Values are continuous between 0 and 1, rather than discrete.

ART 1 (BINARY)	FUZZY ART (ANALOG)
CATEGORY CHOICE	
$T_j = \frac{\lvert \mathbf{I} \cap \mathbf{w}_j \rvert}{\alpha + \lvert \mathbf{w}_j \rvert}$	$T_j = \frac{\lvert \mathbf{I} \wedge \mathbf{w}_j \rvert}{\alpha + \lvert \mathbf{w}_j \rvert}$
MATCH CRITERION	
$\frac{\lvert \mathbf{I} \cap \mathbf{w} \rvert}{\lvert \mathbf{I} \rvert} \geq \rho$	$\frac{\lvert \mathbf{I} \wedge \mathbf{w} \rvert}{\lvert \mathbf{I} \rvert} \geq \rho$
FAST LEARNING	
$\mathbf{w}_j^{(\text{new})} = \mathbf{I} \cap \mathbf{w}_j^{(\text{old})}$	$\mathbf{w}_j^{(\text{new})} = \mathbf{I} \wedge \mathbf{w}_j^{(\text{old})}$
$\cap$ = logical AND intersection	$\wedge$ = fuzzy AND minimum

Fig. 2. Comparison of ART 1 and fuzzy ART.

Complement Encoding

An optional feature of Fuzzy ART is complement coding, a means of incorporating the absence of features into pattern classifications, which helps prevent inefficient and unnecessary category proliferation.

Complement encoding augments each input data vector with the complement of that vector
(1-x is the complement of x).

Complement encoding can also achieve the same effect as normalization in preventing the subset/superset dilemma.

ART Maps

- ARTMAP, also known as Predictive ART, combines two slightly modified ART1 or ART2 units into a supervised learning structure where
  - the first unit takes the **input data** and
  - the second unit takes the **correct output data**,

  then makes the **minimum possible adjustment of the vigilance parameter** in the first unit in order to make the correct classification.

- **Fuzzy ART Map** does this with Fuzzy ART units, which subsumes both ART1 and ART2.

Original article, 1992 IEEE Trans. on Neural Nets

Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps

Gail A. Carpenter, Stephen Grossberg, Natalya Markuzon, John H. Reynolds, and David B. Rosen, *Student Member, IEEE*

Supervised Learning in an ARTMAP

During supervised learning, ART_a receives a stream $\{\boldsymbol{a}^{(p)}\}$ of input patterns, and ART_b receives a stream $\{\boldsymbol{b}^{(p)}\}$ of input patterns, where $\boldsymbol{b}^{(p)}$ is the correct prediction given $\boldsymbol{a}^{(p)}$. These modules are linked by an associative learning network and an internal controller that ensures autonomous system operation in real time. The controller is designed to create the minimal number of ART_a recognition categories, or "hidden units," needed to meet accuracy criteria. It does this by realizing a minimax learning rule that enables an ARTMAP system to learn quickly, efficiently, and accurately as it conjointly *minimizes* predictive error and *maximizes* predictive generalization. This scheme automatically links predictive success to category size on a trial-by-trial basis using only local operations. It works by increasing the vigilance parameter ρ_a of ART_a by the minimal amount needed to correct a predictive error at ART_b.

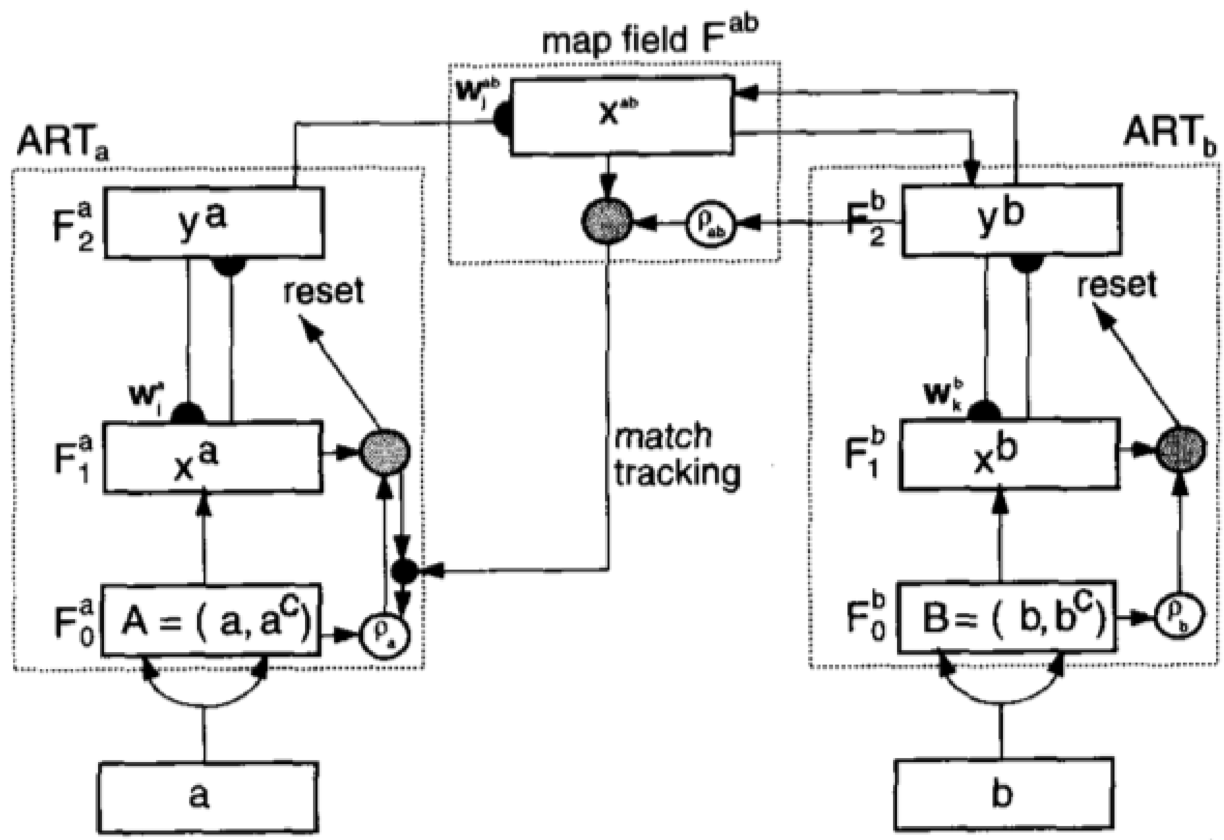


Fig. 1. Fuzzy ARTMAP architecture. The ART_a complement coding preprocessor transforms the M_a vector $\boldsymbol{a}$ into the $2M_a$ vector $\boldsymbol{A} = (\boldsymbol{a}, \boldsymbol{a}^c)$ at the ART_a field $F_0^a \cdot \boldsymbol{A}$ is the input vector to the ART_a field F_1^a. Similarly, the input to F_1^b is the $2M_b$ vector $(\boldsymbol{b}, \boldsymbol{b}^c)$. When a prediction by ART_a is disconfirmed at ART_b, inhibition of map field activation induces the match tracking process. Match tracking raises the ART_a vigilance (ρ_a) to just above the F_1^a to F_0^a match ratio $|\boldsymbol{x}^a|/|\boldsymbol{A}|$. This triggers an ART_a search which leads to activation of either an ART_a category that correctly predicts $\boldsymbol{b}$ or to a previously uncommitted ART_a category node.

Parameters: Fuzzy ART dynamics are determined by a choice parameter $\alpha > 0$; a learning rate parameter $\beta \in [0,1]$; and a vigilance parameter $\rho \in [0,1]$.

Category Choice: For each input $\boldsymbol{I}$ and F_2 node j, the *choice function*, T_j, is defined by

$$T_j(\boldsymbol{I}) = \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{\alpha + |\boldsymbol{w}_j|}, \tag{2}$$

where the fuzzy AND [7] operator $\wedge$ is defined by

$$(\boldsymbol{p} \wedge \boldsymbol{q})_i \equiv \min(p_i, q_i) \tag{3}$$

and where the norm $|\cdot|$ is defined by

$$|\boldsymbol{p}| \equiv \sum_{i=1}^{M} |p_i| \tag{4}$$

for any M-dimensional vectors $\boldsymbol{p}$ and $\boldsymbol{q}$. For notational simplicity, $T_j(\boldsymbol{I})$ in (2) is often written as T_j when the input $\boldsymbol{I}$ is fixed.

The system is said to make a *category choice* when at most one F_2 node can become active at a given time. The category choice is indexed by J, where

$$T_J = \max\{T_j : j = 1 \cdots N\}. \tag{5}$$

If more than one T_j is maximal, the category j with the smallest index is chosen. In particular, nodes become committed in order $j = 1, 2, 3, \cdots$. When the Jth category is chosen, $y_J = 1$; and $y_j = 0$ for $j \neq J$. In a choice system, the F_1 activity vector $\boldsymbol{x}$ obeys the equation

$$\boldsymbol{x} = \begin{cases} \boldsymbol{I} & \text{if } F_2 \text{ is inactive} \\ \boldsymbol{I} \wedge \boldsymbol{w}_J & \text{if the } J\text{th } F_2 \text{ node is chosen.} \end{cases} \tag{6}$$

Resonance or Reset: *Resonance* occurs if the *match function*, $|\boldsymbol{I} \wedge \boldsymbol{w}_J|/|\boldsymbol{I}|$ of the chosen category meets the vigilance criterion:

$$\frac{|\boldsymbol{I} \wedge \boldsymbol{w}_J|}{|\boldsymbol{I}|} \geq \rho; \tag{7}$$

that is, by (6), when the Jth category is chosen, resonance occurs if

$$|\boldsymbol{x}| = |\boldsymbol{I} \wedge \boldsymbol{w}_J| \geq \rho|\boldsymbol{I}|. \tag{8}$$

Learning then ensues, as defined below. *Mismatch reset* occurs if

$$\frac{|\boldsymbol{I} \wedge \boldsymbol{w}_J|}{|\boldsymbol{I}|} < \rho; \tag{9}$$

that is, if

$$|\boldsymbol{x}| = |\boldsymbol{I} \wedge \boldsymbol{w}_J| < \rho|\boldsymbol{I}|. \tag{10}$$

Then the value of the choice function T_J is set to 0 for the duration of the input presentation to prevent the persistent selection of the same category during search. A new index J is then chosen, by (5). The search process continues until the chosen J satisfies (7).

Learning: Once search ends, the weight vector $\boldsymbol{w}_J$ is updated according to the equation

$$\boldsymbol{w}_J^{(\text{new})} = \beta\left(\boldsymbol{I} \wedge \boldsymbol{w}_J^{(\text{old})}\right) + (1-\beta)\boldsymbol{w}_J^{(\text{old})}. \qquad (11)$$

Fast learning corresponds to setting $\beta = 1$. The learning law used in the EACH system of Salzberg [11]–[13] is equivalent to (11) in the fast-learn limit with the complement coding option described below.

Fast-Commit Slow-Recode Option: For efficient coding of noisy input sets, it is useful to set $\beta = 1$ when J is an uncommitted node, and then to take $\beta < 1$ after the category is committed. Then $\boldsymbol{w}_J^{(\text{new})} = \boldsymbol{I}$ the first time category J becomes active. Moore [21] introduced the learning law (11), with fast commitment and slow recoding, to investigate a variety of generalized ART 1 models. Some of these models are similar to fuzzy ART, but none includes the complement coding option. Moore described a category proliferation problem that can occur in certain analog ART systems when a large number of inputs erode the norm of weight vectors. Complement coding solves this problem.

ART Critique

It has been noted that results of ART1 and Fuzzy ART **depend critically upon the order** in which the training data are processed.

The effect can be reduced to some extent by using a slower learning rate, but is present regardless of the size of the input data set.

In other words, ART1 and Fuzzy ART estimates do not possess the statistical property of consistency.

Reference: Sarle, Warren S. (1995), *Why Statisticians Should Not FART*

ART and Other Models

- At some point, Grossberg related ARTMAPs to SOMs.
- There would seem to be a connection between ART and the GNG (Growing Neural Gas) model, as both have methods for introducing new units. See *Constructive Feedforward ART Clustering Networks*, Part II Andrea Baraldi and Ethem Alpaydın, IEEE Trans. Neural Networks, May 2002. (FOSART = Fully Self-Organizing Simplified ART).
- Survey on Self-Generating Networks (PhD Proposal)
- What about ART vs. Deep Belief Networks?

FOSART (Fully Self-Organizing Simplified ART) Comparisons

1) unlike GNG and SOM, FOSART tries to minimize a quantization (sum-of-squares) error via a soft-to-hard competitive model transition.
2) Unlike Fuzzy ART, the system requires no complement coding of the input data.
3) Unlike SOM and NG, FOSART requires no randomization of the initial template vectors.
4) Unlike SOM and NG, the system requires no *a priori* knowledge of the size of the network.
5) Unlike SOM, the system requires no *a priori* knowledge of the topology of the network.
6) Unlike SOM, NG, and Fuzzy ART, FOSART explicitly deals with lateral connections.
7) Unlike GNG, FOSART attempts to address all constraints required to make the CHR guarantee perfect topology-preserving mapping in the sense proposed in [8].
8) Unlike parameters of SOM and NG, FOSART parameters are not affected by outliers which are instead mapped onto noise categories.
9) Unlike Fuzzy ART, the system is capable of removing noise categories to avoid overfitting.
10) Unlike Fuzzy ART, FOSART is competitive with other clustering models found in the literature when the Iris data set is clustered with three reference vectors [51].

Extensions and Implementations

- MATLAB Central - File detail - Simplified Fuzzy ARTMAP Neural Network
- Adaptive Resonance Theory Microchips: Circuit Design Techniques, Springer Verlag, 1998