Hierarchical Sparse Distributed Representations of Sequence Recall and Recognition

Gerard Rinkus
Volen Center for Complex Systems
Brandeis University

Acknowledgement: John Lisman & Eve Marder NIH Training Grant
Jeff Hawkins & Redwood Neuroscience Institute

The Goal

A biologically plausible model that:

1. **Can learn sequences with single trials**
2. **Uses sparse distributed representations for both individual sequence items and sequences**
3. **Can *recall* individual sequences as well as *recognize* novel sequences**
4. **Exhibits robust invariance to non-uniform time warping**
5. **Consists of a single monolithic substrate in which are stored both the traces of individual sequences and the higher-order statistics of the set of sequences**

- **There is no current model possessing all of these properties**

Outline

Problem Description

- Sequences (spatiotemporal patterns)

Model Description

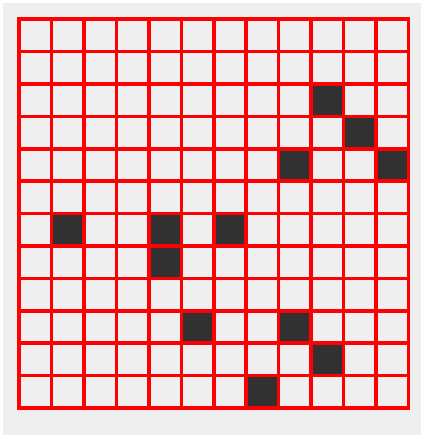
- Sparse distributed (combinatorial) coding
- Matching between expectation and actual inputs
- Progressive persistence
- Hierarchality
- Multi-stage matching ('back-off') for time invariance

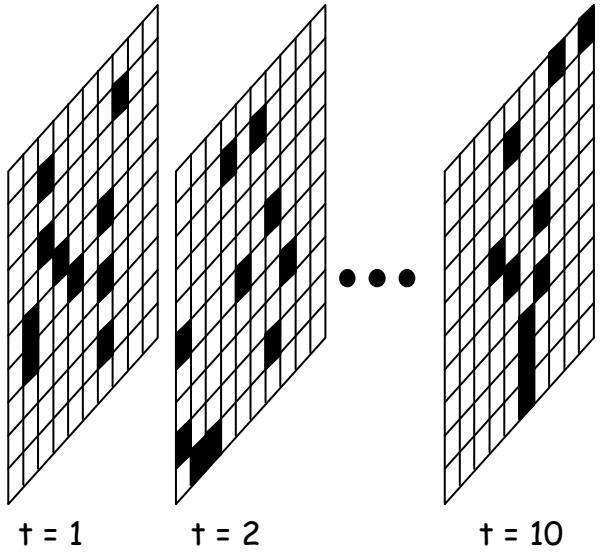
Simulation Results

The Problem: Learning, Recalling, Recognizing Sequences (Spatiotemporal Patterns)

For example:

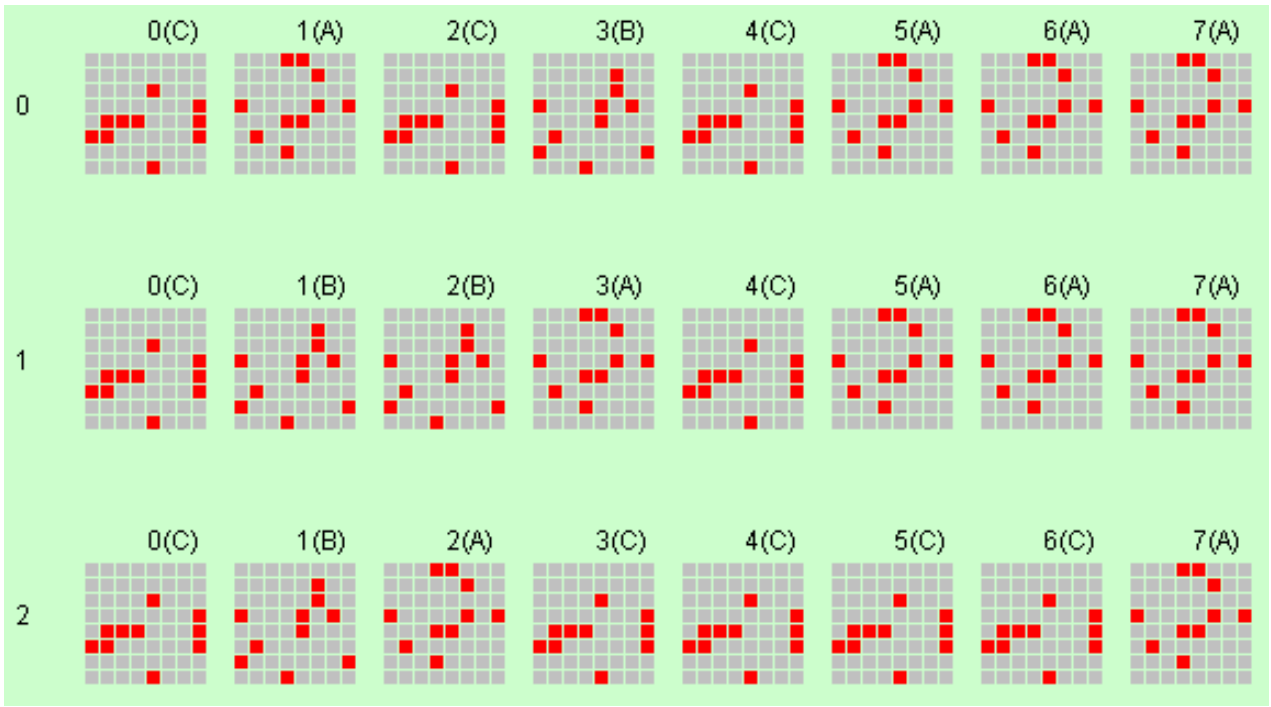
- Input layer (e.g., 'retina') is a 10 x 10 binary feature array.
- M = total number of input features = 100
- S = number of features per time slice = 10
- Features chosen at random on each slice





Complex Sequences

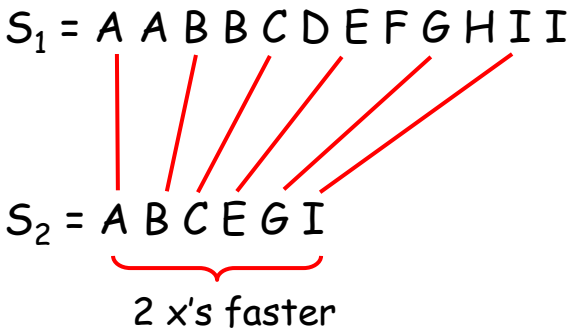
- Any Sequence model must handle complex sequences

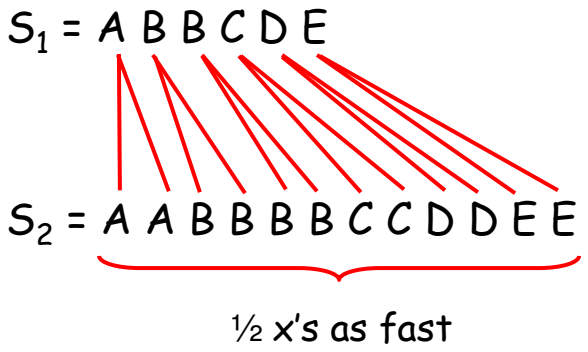


Time-Invariant Recognition

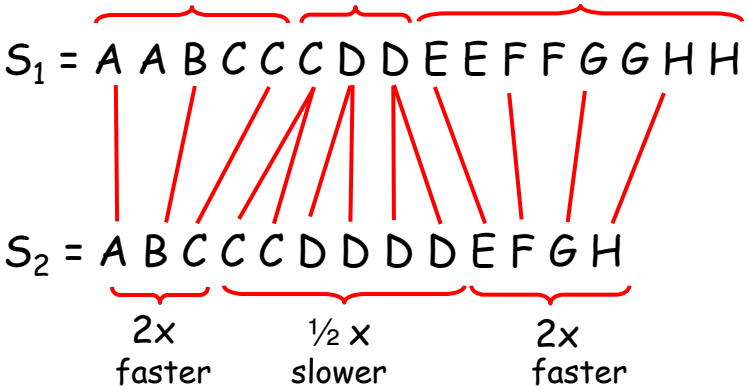
- *Uniform time warping*

If trained on S_1 then it will recognize S_2.

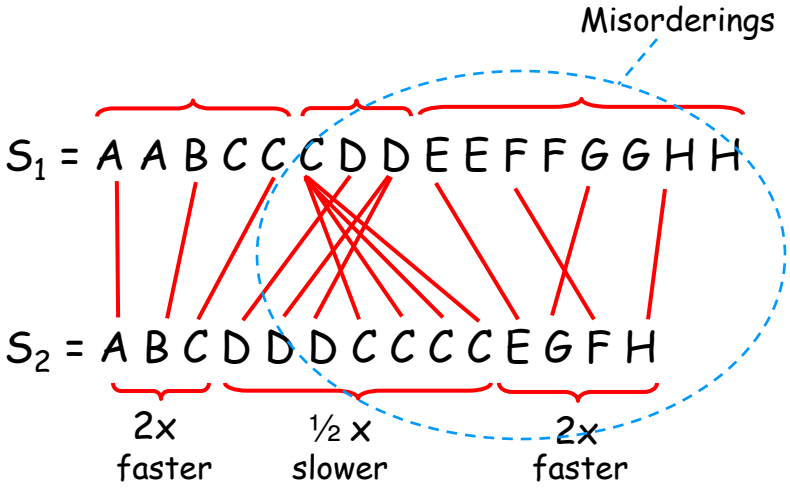




- *Non-uniform time warping*

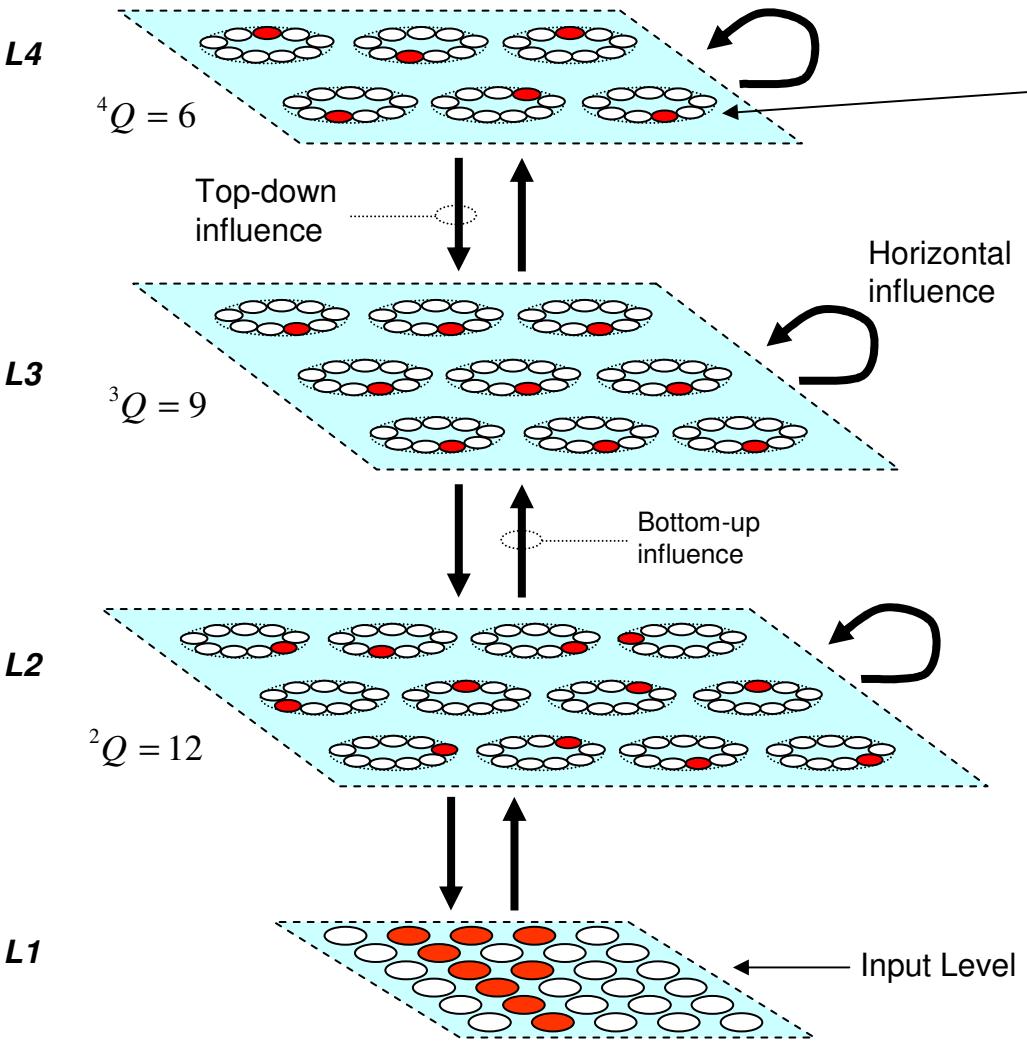


Should Pass



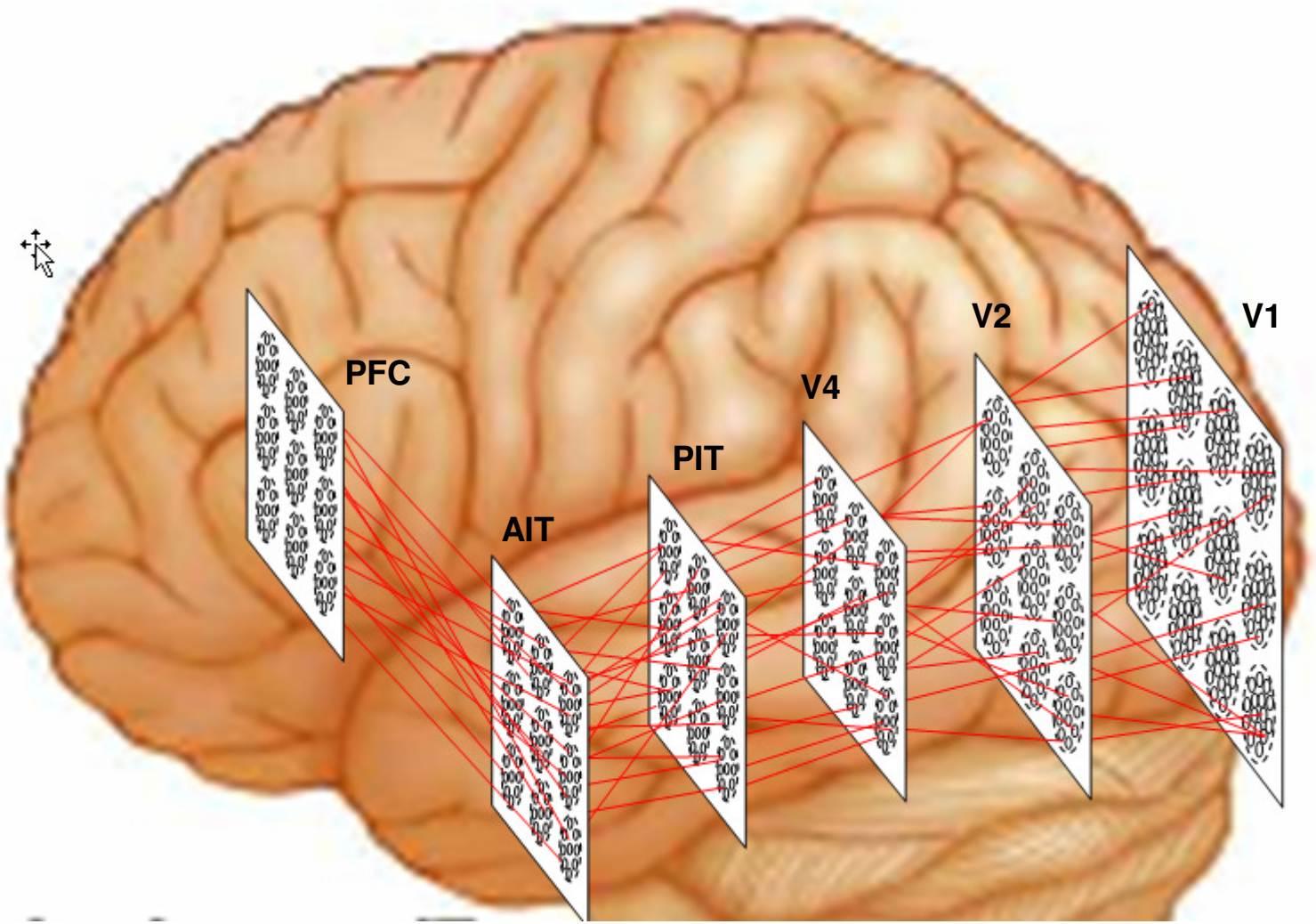
Should Fail

Some Model Properties

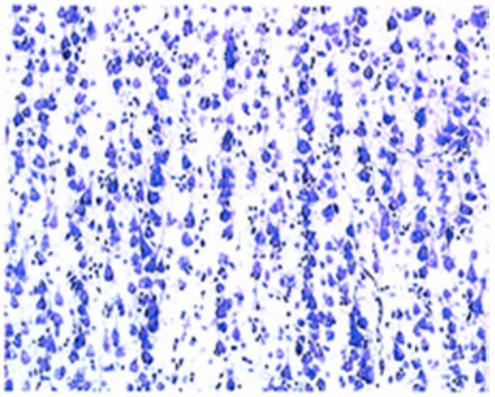


- Winner-take-all Competitive module (CM)
- On each time step, the model computes winners based on the TD, H and BU inputs.
- ^JQ is the number of CMs in the J^{th} level.
  - Codes at all levels are normalized
- Black arrows represent synaptic projections, which are all complete (all-to-all) in the simulations reported here.
- Cells at higher levels have longer persistences
  - When they become active, they tend to stay active for a longer period.
- Unsupervised Hebbian learning

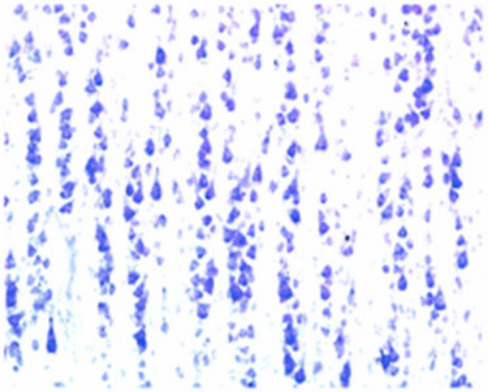
Proposed Correspondence to Brain



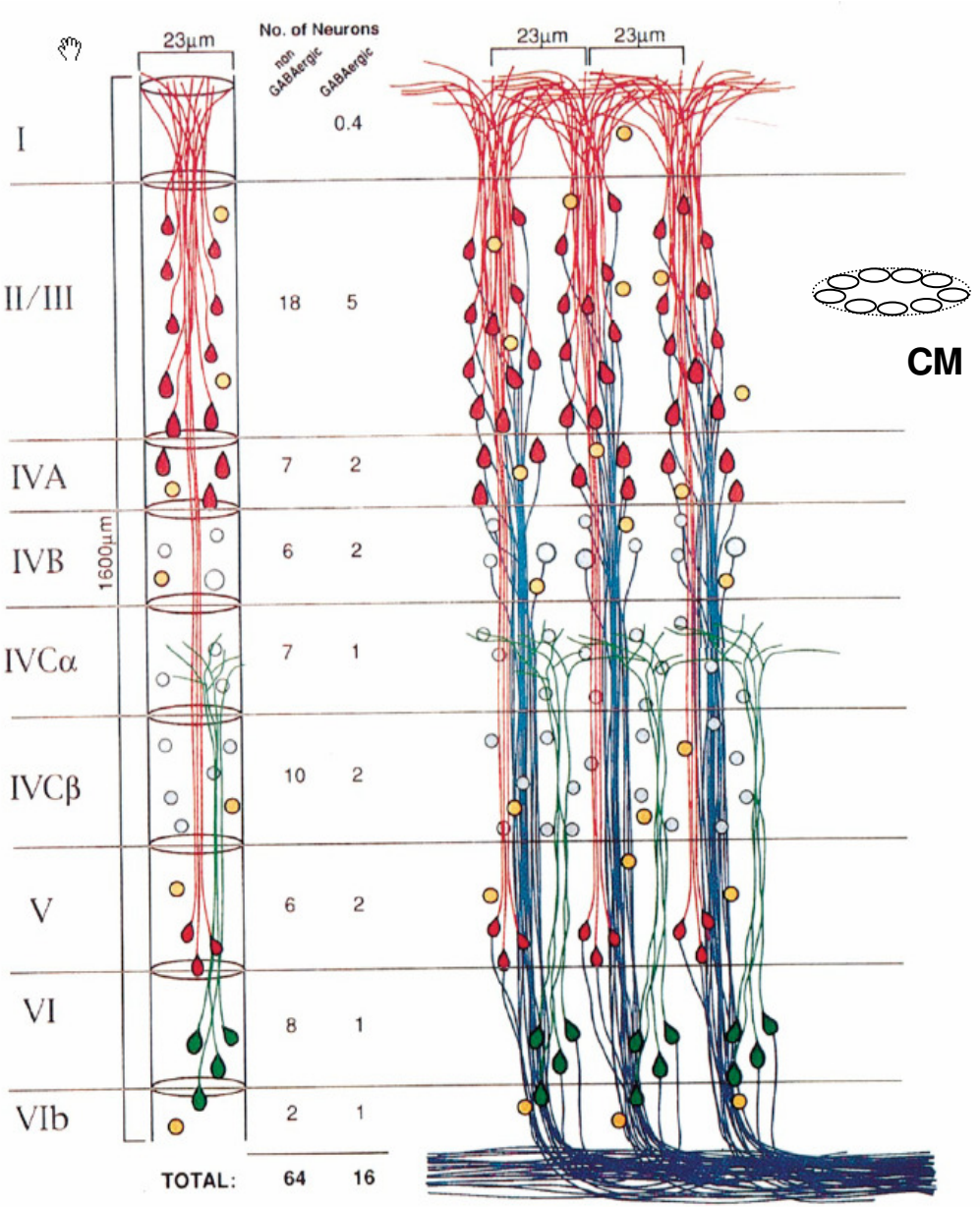
Correspondence to Minicolumns



9 year old human.



67 year old human.



- Photos from James A. Anderson Talk on the Ersatz Brain Project
  www.cog.brown.edu/research/ErsatzBrainGroup/EBP_ICCI05_SparseTalkPPT.ppt

- From Peters and Sethares, 1996

Persistence Data

- 30+ years of working memory experiments showing that cells in various cortical areas (PFC, IT, EC) can remain firing at a high rate for many seconds, given a particular stimulus.
  - Fuster & Alexander, 1971
  - Miyashita & Chang, 1988

- Using a simple visual evoked potential paradigm (no task or reward involved), Uusitalo et al. (1996) found that *activation lifetimes* (persistences) lasted longer at progressively higher levels

Uusitalo et al. (1996) Persistence Data

Location	Latency (ms) (order)	τ (s)
Subject 1		
V1	90 (1)	0.19 ± 0.03
V1 or V2	120 (2)	0.28 ± 0.04
Left V2	140 (3)	0.43 ± 0.07
Right V2/V3	180 (4)	0.15 ± 0.01
Right prefrontal sulcus	180 (5)	11 ± 2
Left supratemporal sulcus	220 (6)	13 ± 1
Right postcentral sulcus	220 (7)	28 ± 5
Left parieto-occipito-temporal junction	260 (8)	14 ± 1
Right postcentral sulcus	280 (9)	11 ± 1
Right superior frontal sulcus	330 (10)	19 ± 2
Subject 2		
V1	90 (1)	0.28 ± 0.05
V2	120 (2)	0.64 ± 0.09
Right prefrontal sulcus	180 (3)	21 ± 2
Right postcentral sulcus	230 (4)	7 ± 1
Right postcentral sulcus	270 (5)	13 ± 1
Left postcentral sulcus	320 (6)	10 ± 2
Left parieto-occipital border	440 (7)	24 ± 2

- Magnetic source imaging (MSI) of evoked responses to brief (70 msec) checkerboard stimulus
- They clustered cortical areas into two groups based on lifetimes
  - Occipital: 0.1—0.6 sec.
  - Temporal, Parietal, & Frontal: 7—30 sec.

The Model: Basic Operation

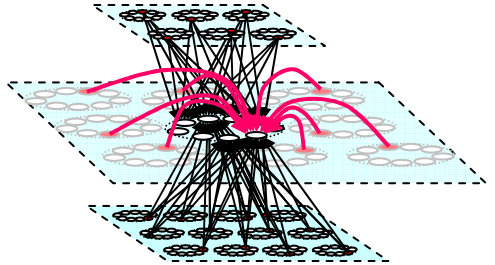
On every time step and at every level:

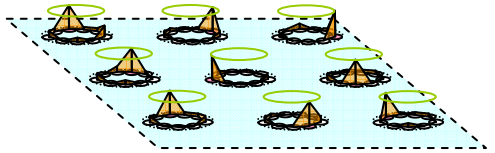
1. Every cell computes the degree of match, χ, between *its* bottom-up, top-down, and horizontal input vectors.

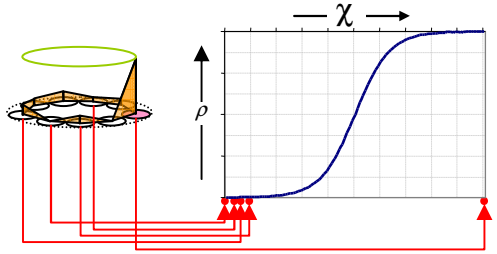
2. The overall degree of match, ${}^{J}G$, for an entire level, J, is computed.
   - ${}^{J}G$ is the average of the maximal χ values in all of J's CMs.

3. Within each CM, the vector of χ values is transformed into a probability distribution (ρ values).
   - The shape of the ρ distribution depends on ${}^{J}G$.

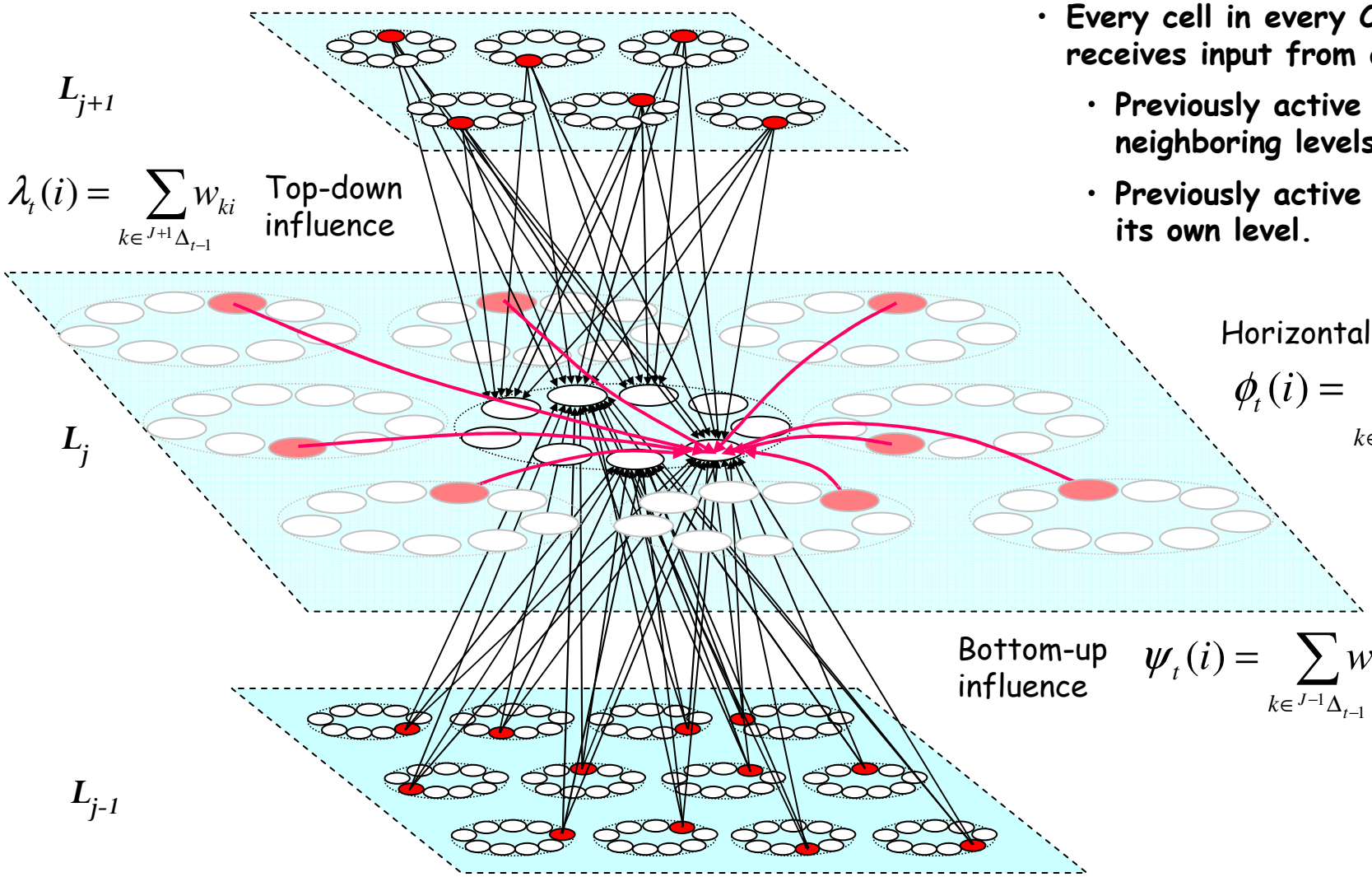
4. Within each CM, a winner is chosen on the basis of the ρ distribution.





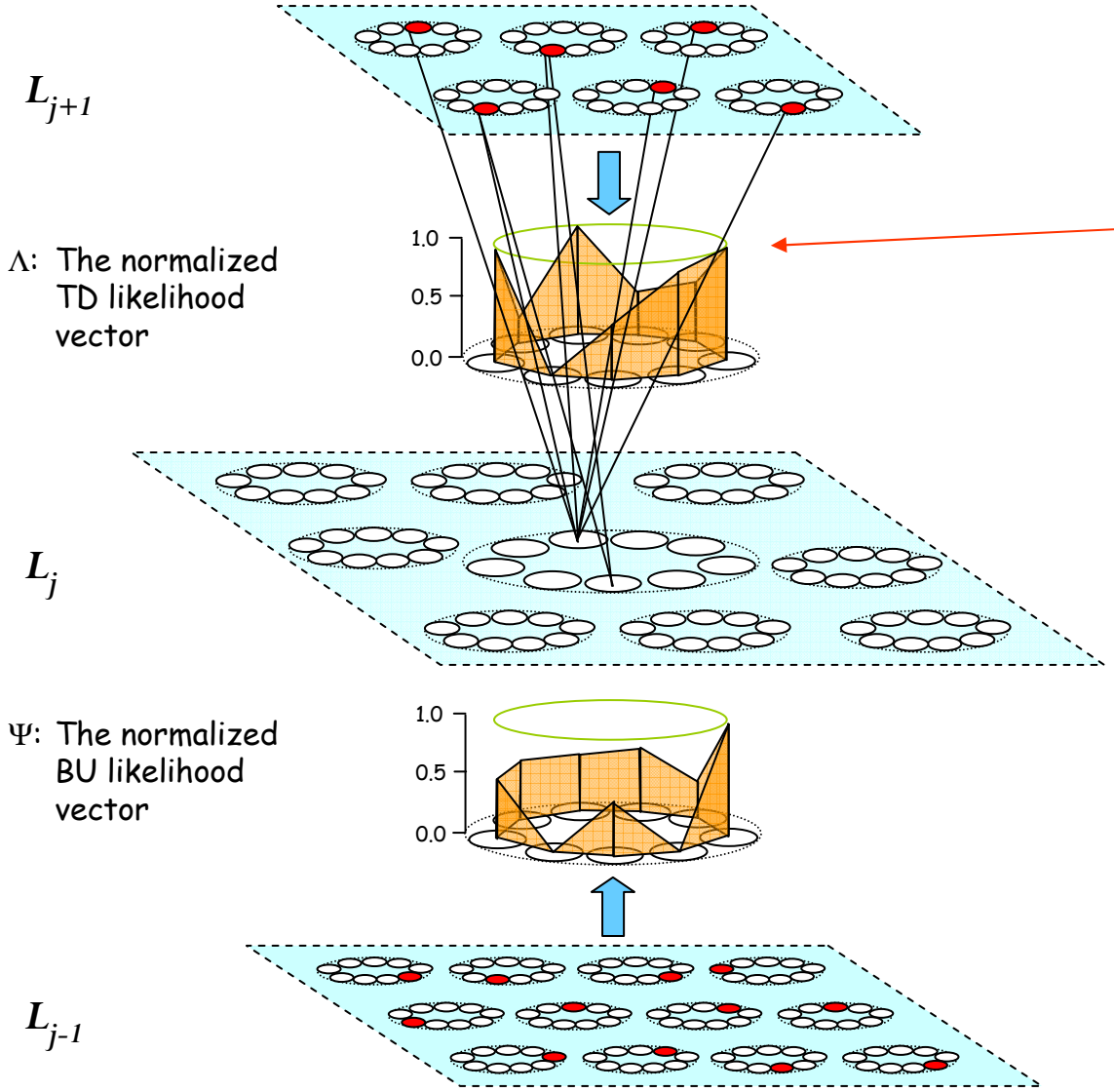


Computing Degree of Match I: Sum Raw Inputs



- **Every cell in every CM receives input from all:**
  - **Previously active cells in neighboring levels.**
  - **Previously active cells in its own level.**

Computing Degree of Match II: Normalize Raw Inputs



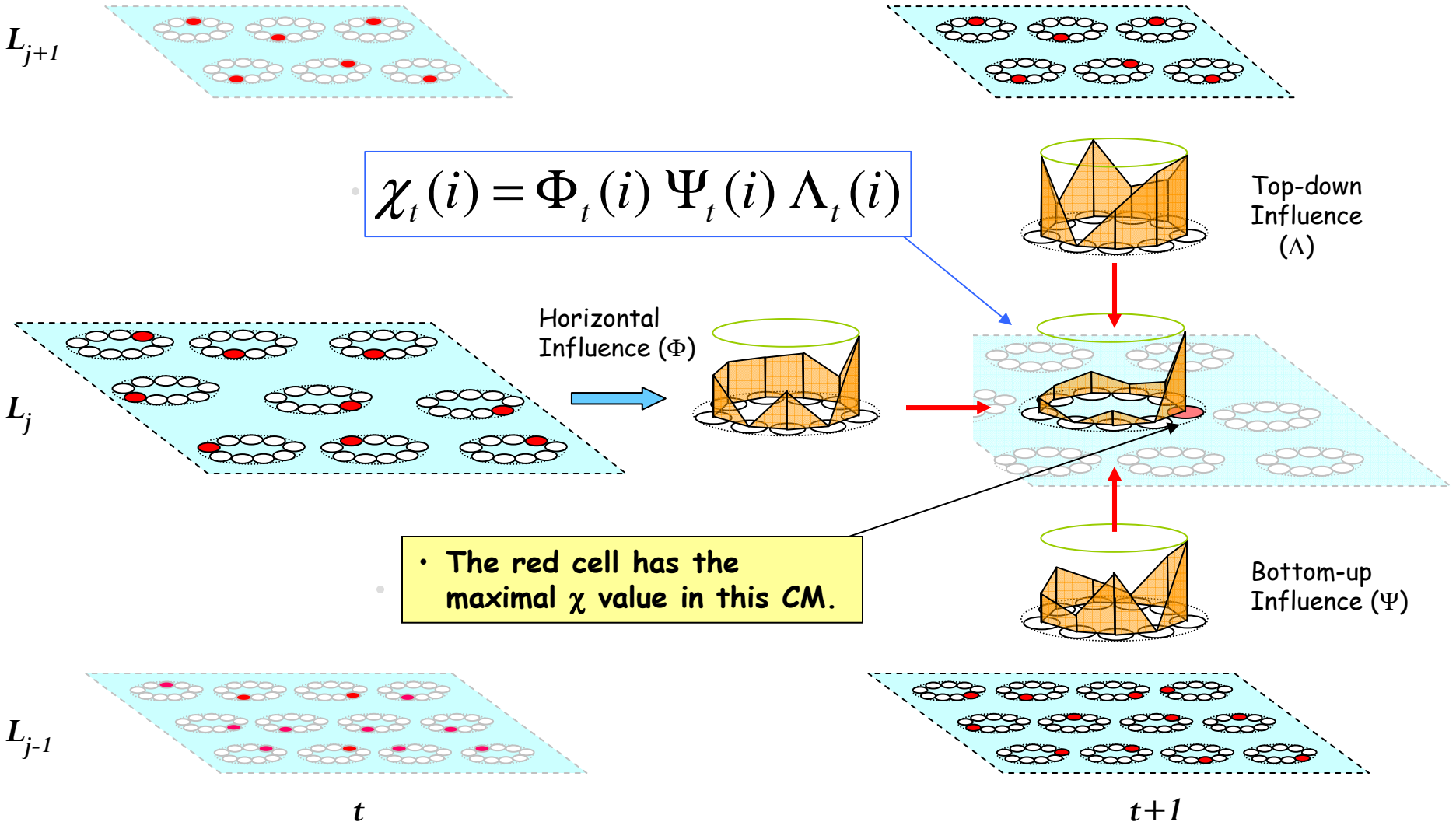
- **Green circle represents the maximal *normalized* TD input.**
- **The orange 'crown' gives a visual depiction of the *relative* TD *likelihoods* over the cells of the CM.**

$$ {}^{J}\Lambda_t(i) = \frac{\lambda_t(i)}{{}^{J+1}Q} $$ TD

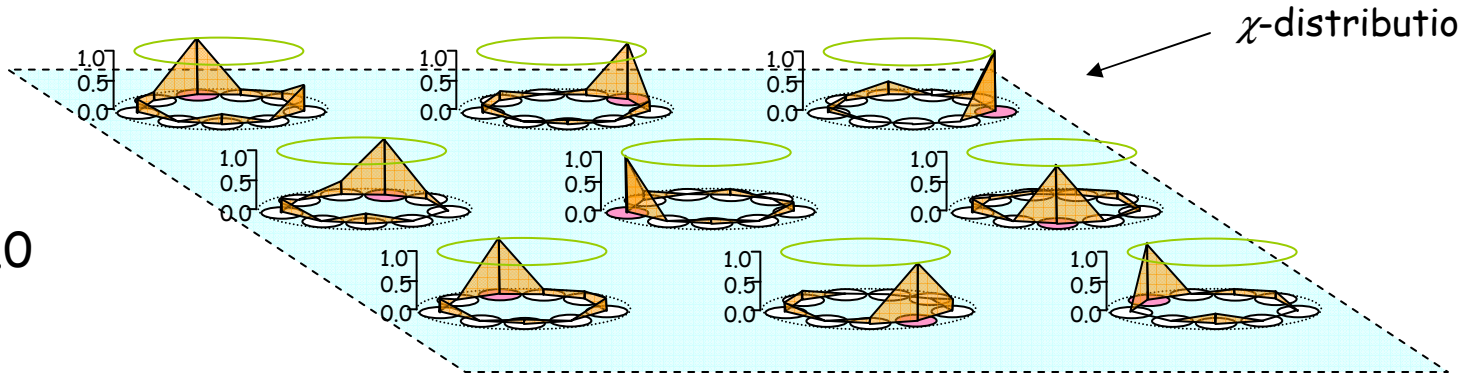
$$ {}^{J}\Phi_t(i) = \frac{\phi_t(i)}{{}^{J}Q} $$ H

$$ {}^{J}\Psi_t(i) = \frac{\psi_t(i)}{{}^{J-1}Q} $$ BU

Computing Degree of Match III: Combining TD, H, and BU Evidence



G = the Average of the Maximal χ Values Across a Whole Level

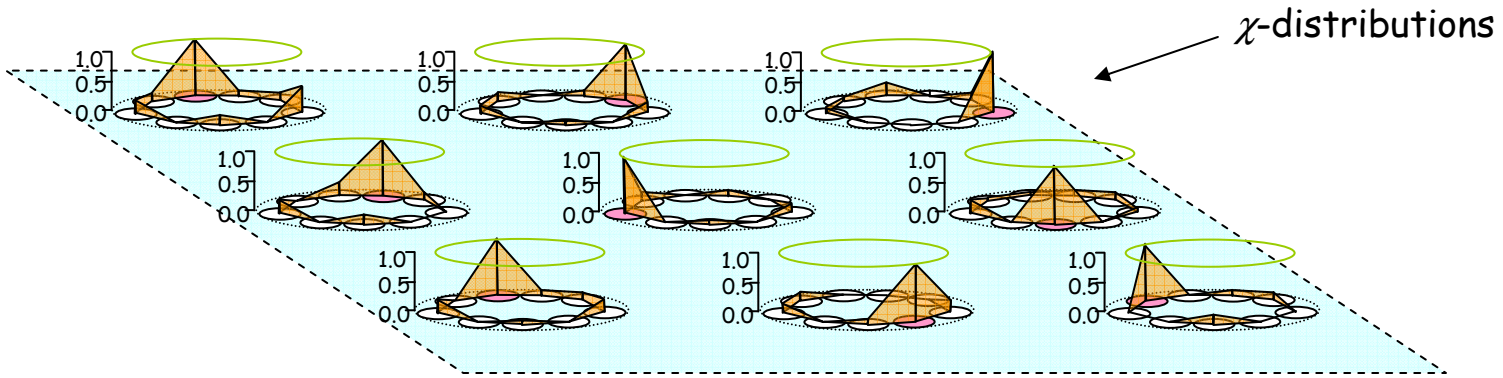


$G_t \approx 1.0$

- Find max χ value in each CM $\quad \bar{\chi}_t(z) = \max_{k \in CM_z} \chi_t(i)$
- Take the average $\quad {}^{j}G_t = \dfrac{\sum_{z=1}^{{}^{j}Q} \bar{\chi}_t(z)}{{}^{j}Q}$

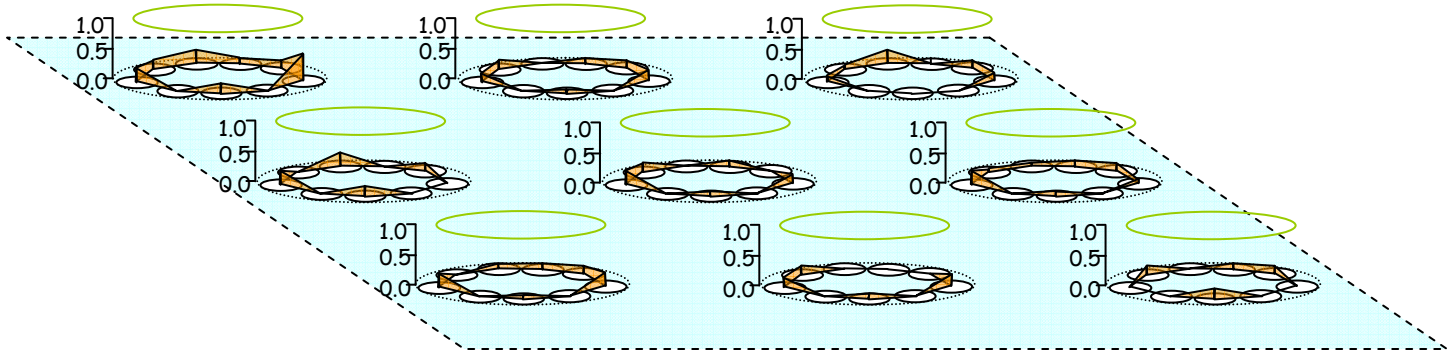
$G \approx 1$ means Familiar; $G \approx 0$ means Unfamiliar

$G_t \approx 1.0$



- This means that the model finds the current moment *familiar* given *all* of its prior experience.

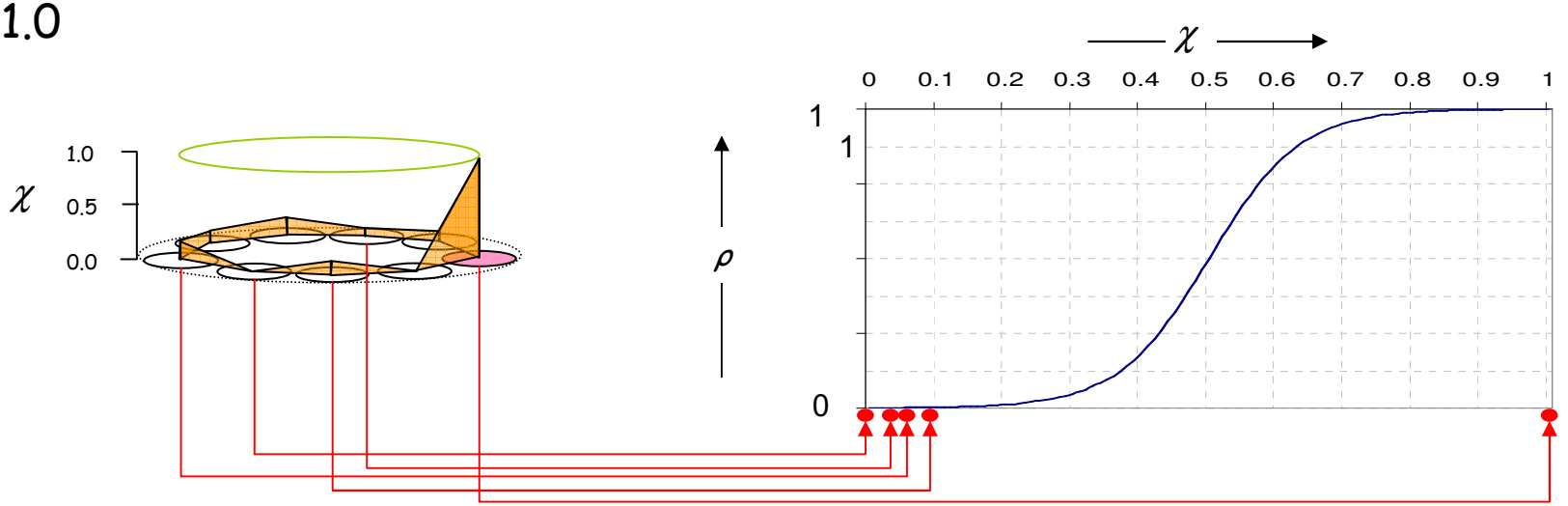
$G_t \approx 0.0$



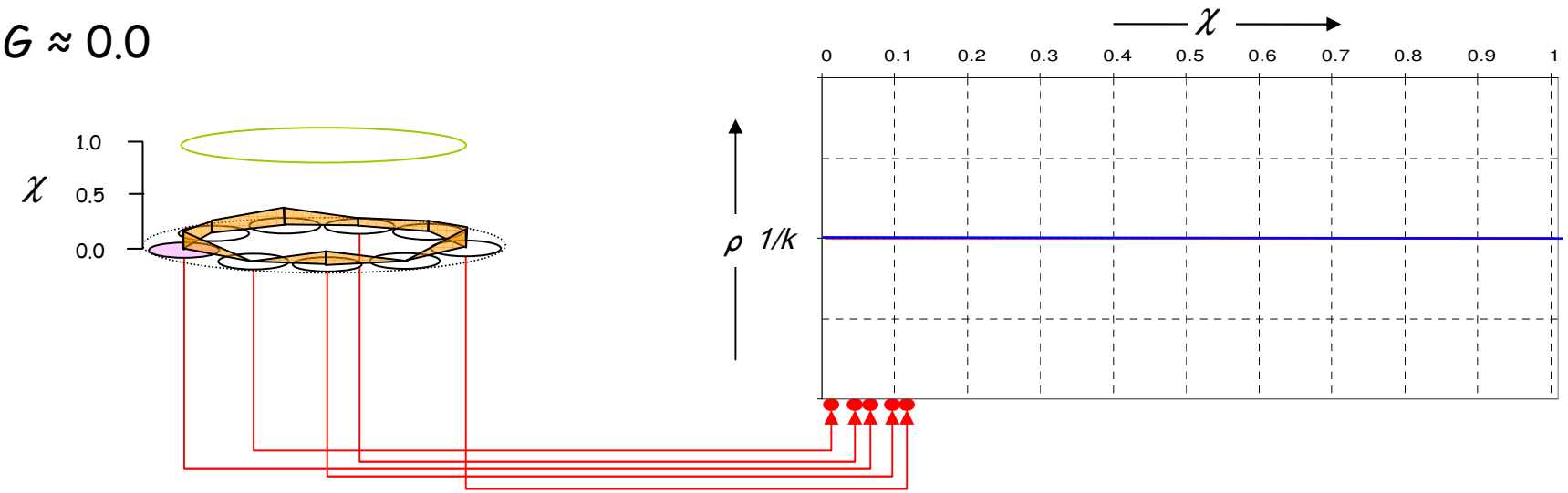
- This means that the model finds the current moment almost completely *unfamiliar (novel)* given its past experience.

Mapping χ-values to ρ-values

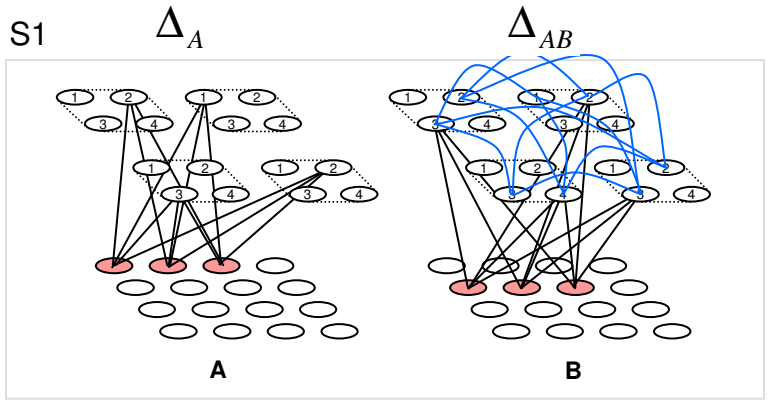
$G \approx 1.0$



$G \approx 0.0$



Spatiotemporal Similarity Maps to Code Overlap



- Suppose a 2-layer network's first learning trial is the 2-item sequence, S1 = [AB].

- Item A presents at t = 1
  All weights are zero → all χ values are 0 $\chi_i = \Psi_i$
  → G = 0
  → All cells equally likely to be chosen winner
- → A random L2 code, Δ_A, is chosen
- → BU weights increased

- Item B presents at t = 2
  Since A ∩ B = ∅, and since all H-wts are still 0, all χ values are 0 $\chi_i = \Phi_i \Psi_i$
  → G = 0
  → All cells equally likely to be chosen winner
- → A random L2 code, Δ_{AB}, is chosen
- → BU an H weights increased

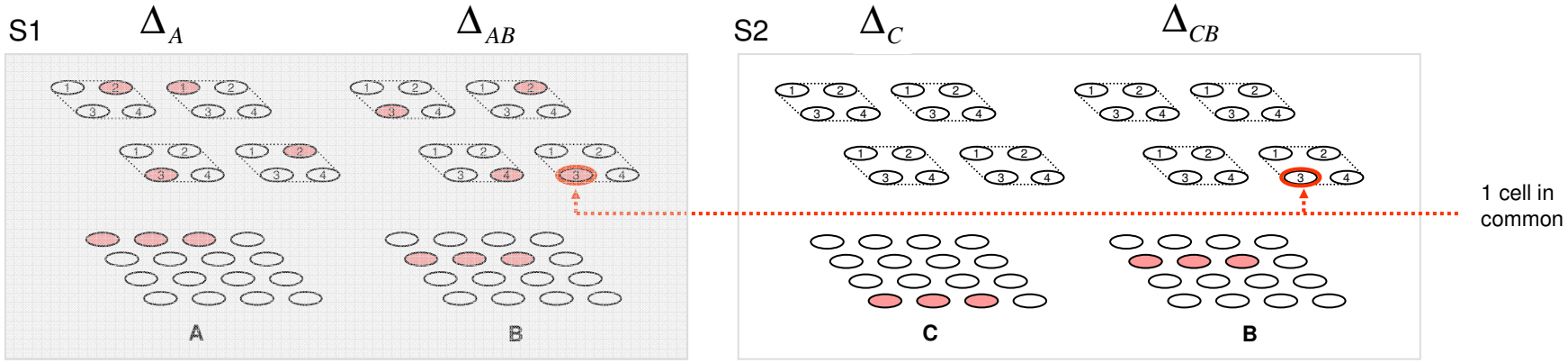
Raw BU input $\psi_t(i) = \sum_{k \in L1} x_k w_{ki}$

Normalized BU input $\Psi_t(i) = \frac{\psi_t(i)}{{}^{L1}Q}$

Assumption: L1 activity is normalized; all input items have 3 active features

Spatiotemporal Similarity Maps to Code Overlap

Suppose the network is next presented with S2 = [CB]



More Generally, when G ≈ 0, the average overlap between the code that is chosen and ALL preexisting codes is minimized.

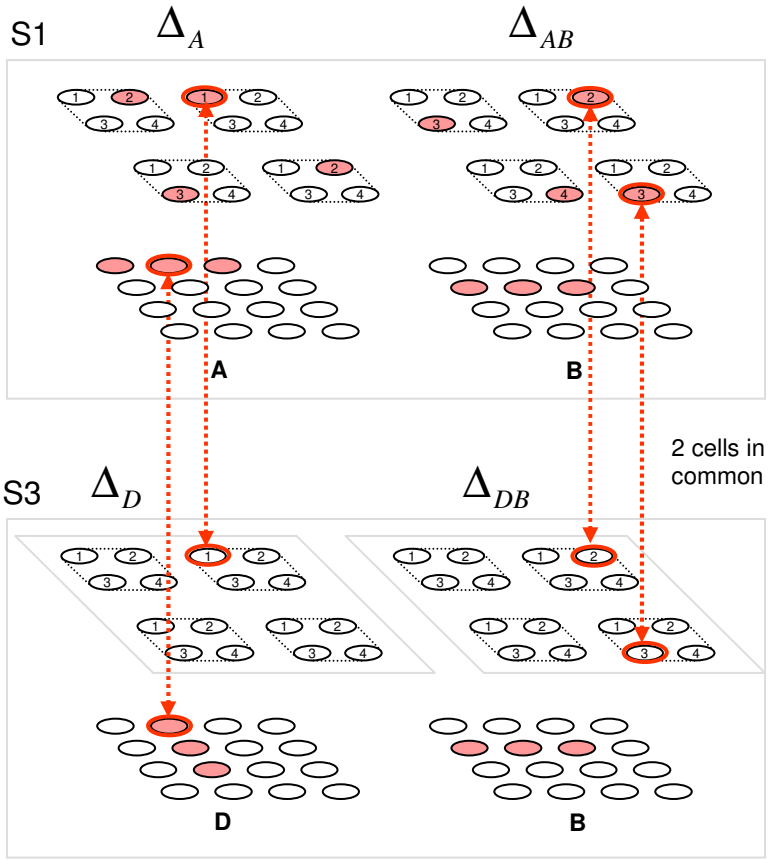
$$\chi_i = \Phi_i \Psi_i$$

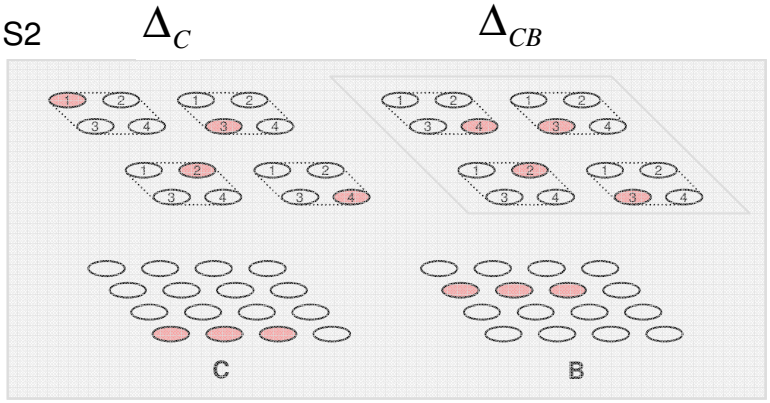
- Item C presents at t = 1
- All BU weights from item C are zero → all χ values are 0
- → G = 0 → a random L2 code, Δ_C, is chosen
- Assume $\Delta_C \cap \Delta_A = \varnothing$ and $\Delta_C \cap \Delta_{AB} = \varnothing$
- → BU weights increased

- Item B presents at t = 2
- The set of cells in Δ_{AB} will have maximal BU support (Ψ = 1.0)
- However, since $\Delta_C \cap \Delta_A = \varnothing$, the H support for the cells of Δ_{AB} is zero.
- → all χ values are 0 → G = 0 → highly novel L2 code, Δ_{CB}, is chosen
- Assume that due to chance, $\Delta_{CB} \cap \Delta_{AB} = 1$.
- → BU and H weights increased

Spatiotemporal Similarity Maps to Code Overlap

Suppose the network is next presented with S3 = [DB]





- Item D presents at t = 1
- Since D ∩ A = 1, χ values for cells in Δ_A will be 0.33 → G = 0.33
  → less noise will be added into the choice of winners
- → Assume $\Delta_D \cap \Delta_A = 1$

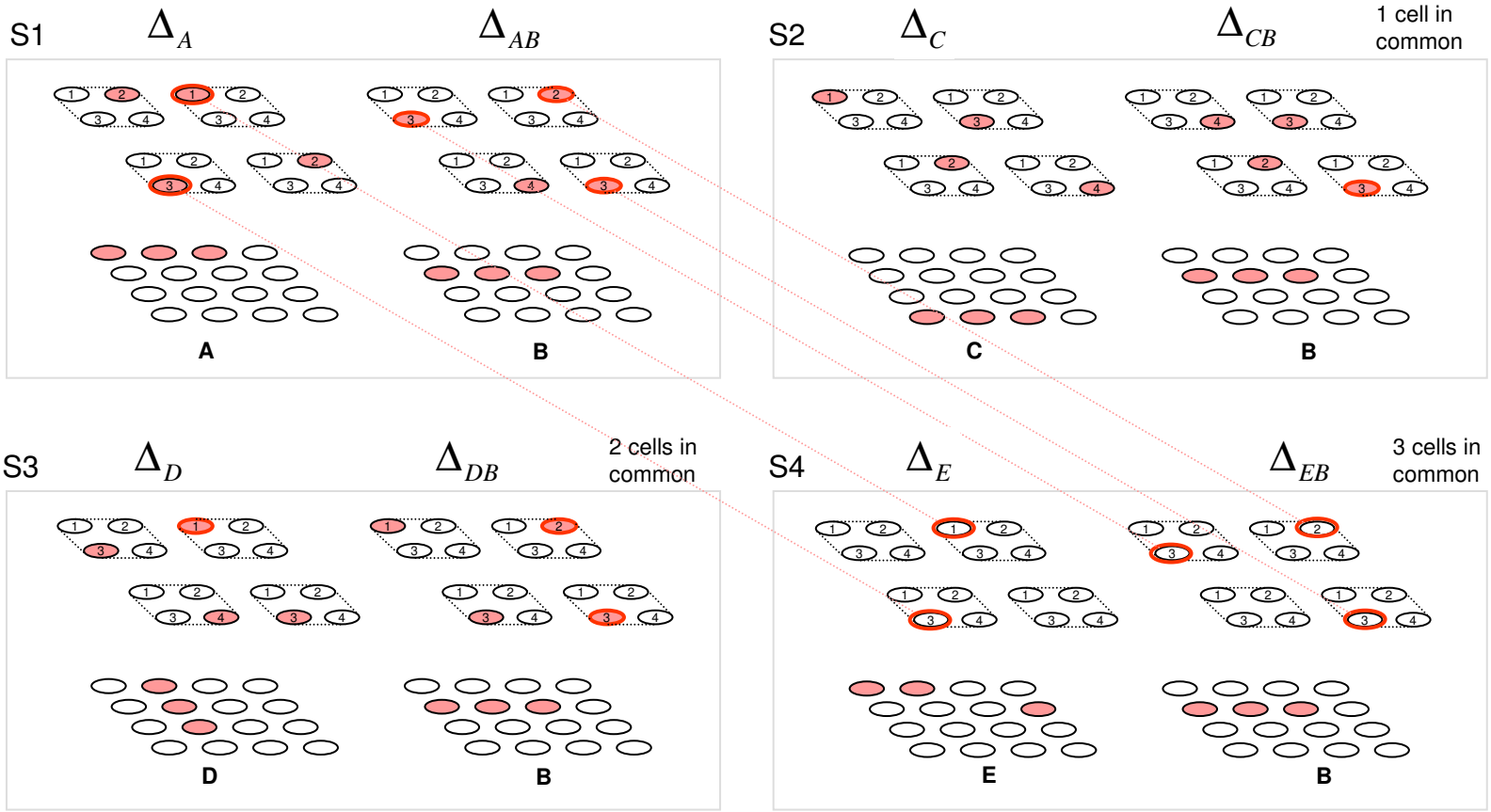
- Item B presents at t = 2
- The set of cells in Δ_{AB} will have maximal BU support ($\Psi_i = 1.0$)
- Furthermore, the cell common to Δ_A and Δ_D → the cells of Δ_{AB} will have non-zero H support; Φ values of 0.33
  → non-zero χ values (favoring the cells of Δ_{AB}) → non-zero G
- → Assume the finally chosen code, Δ_{DB} ,has two cells in common with Δ_{AB}

- **The greater spatiotemporal similarity of [DB] and [AB] than of [CB] and [AB]**
  → SIM(Δ_{DB}, Δ_{AB}) > SIM(Δ_{CB}, Δ_{AB})

- cf. Spatter Codes (Kanerva; 1998)
  Context-dependent Thinning (Rachkovskij & Kussel; 2000, 2001)

Spatiotemporal Similarity Maps to Code Overlap

Consider a final sequence S4 = [EB], which has even greater similarity to [AB]

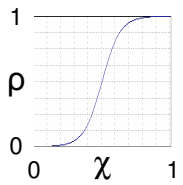


- Item E presents at t = 1
  Since E ∩ A = 2, χ values for cells in Δ_A will be 0.66
- → G = 0.66 → less noise will be added
  → Assume $\Delta_E \cap \Delta_A = 2$

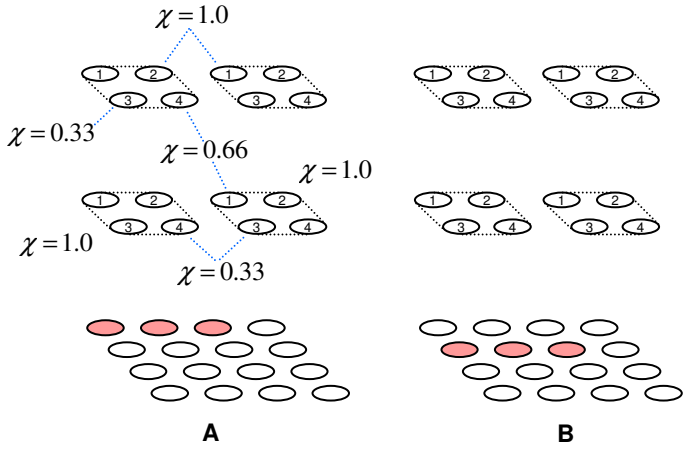
- Item B presents at t = 2
  Higher overlap of Δ_A and Δ_E → higher χ values for cells in Δ_{AB}
- → higher G than for S3 (at t=2) → less noise will be added
  → Assume $\Delta_{EB} \cap \Delta_{AB} = 3$

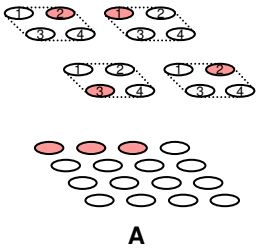
Re-presentation of a Known Sequence

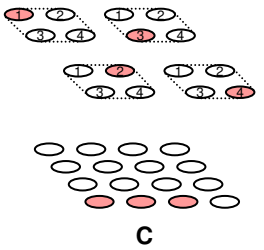
$$\chi_i = \Psi_i$$

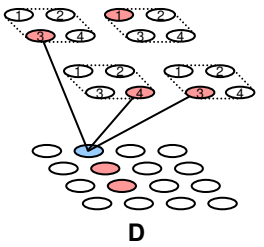


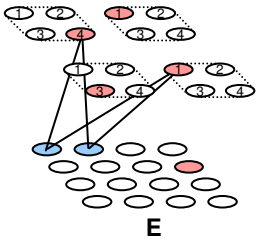
- Suppose we present [AB] to the model again.
- → At t = 1, four cells have χ values of 1.0,
  three have χ = 0.33,
  two have χ = 0.66
- → G = 1.0
- → no noise added to winner selection
- → Δ_A is reinstated with probability very close to one.





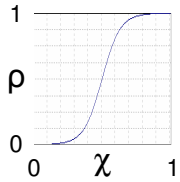




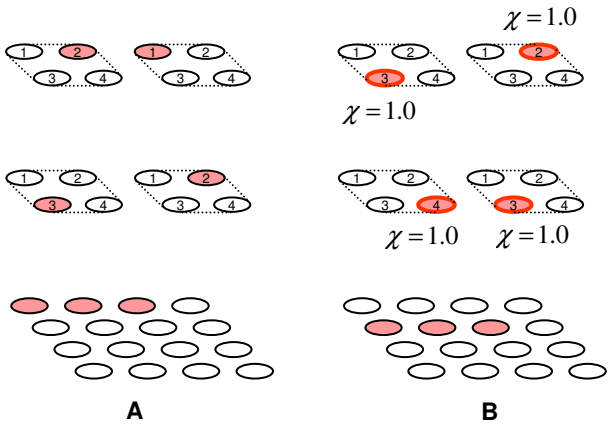


Re-presentation of a Known Sequence

$$\chi_i = \Phi_i \Psi_i$$



- → At t = 2, four cells have χ values of 1.0.
- → G = 1.0
- → no noise added to winner selection
- → Δ_{AB} is reinstated with probability very close to one.

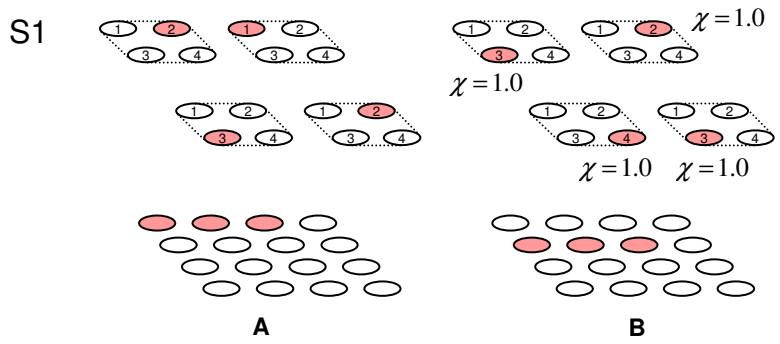


Code-specific G

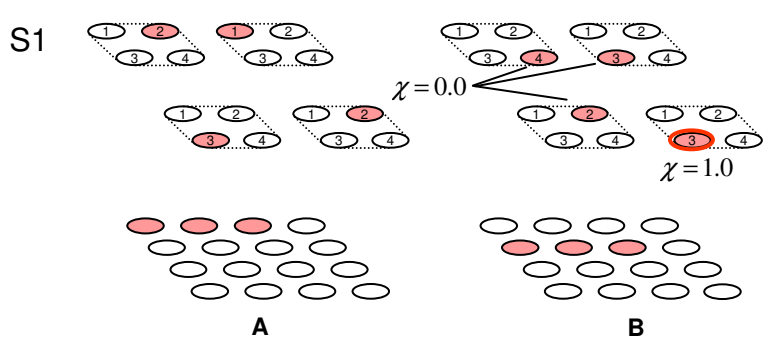
- Define the *code-specific* G, G(Δ_i), to be the average of the χ values for the cells comprising a particular code Δ_i.

$$\chi_i = \Phi_i \Psi_i$$

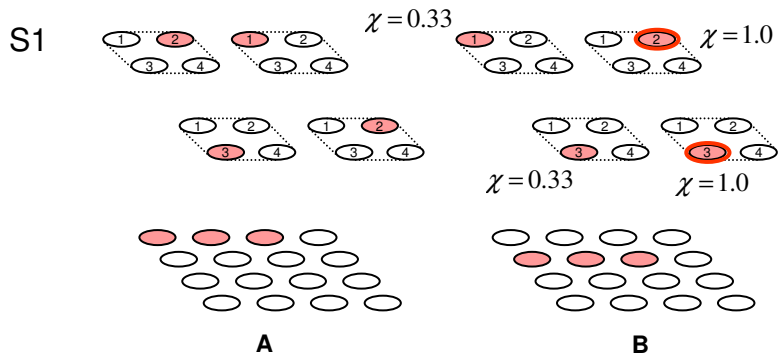
- $G(\Delta_{AB}) = \frac{4.0}{4} = 1.00$



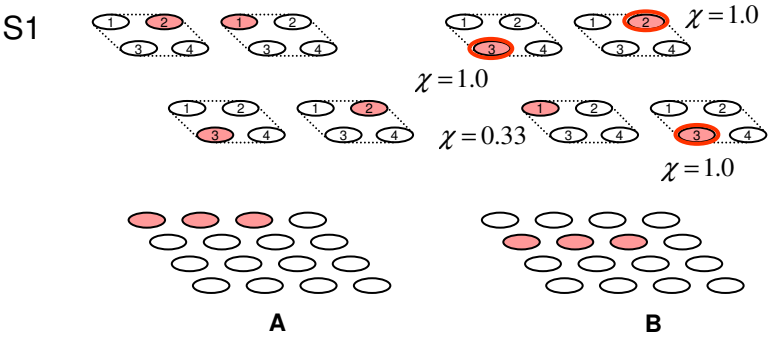
- $G(\Delta_{CB}) = \frac{1.0}{4} = 0.25$



- $G(\Delta_{DB}) = \frac{2.66}{4} = 0.665$

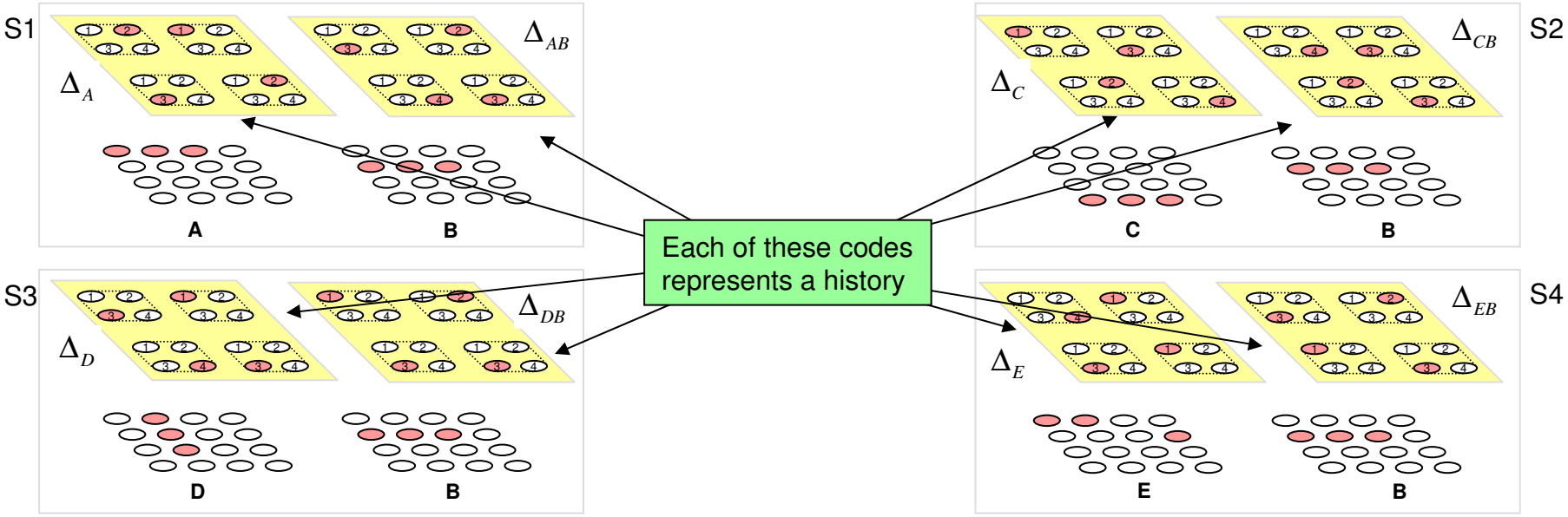


- $G(\Delta_{EB}) = \frac{3.33}{4} = 0.832$



The Code-specific G is a Spatiotemporal Similarity Metric

There have been eight spatiotemporal contexts represented in the network's 'life'



- Simply taking the average of the maximal χ values over the layer's CMs computes the maximal degree of match between the current spatiotemporal context and ALL previously experienced spatiotemporal contexts (histories)
- The subsequent act of pushing the values through the nonlinearity effectively CHOOSES which one of the set of previously experienced spatiotemporal contexts (hypotheses) is that closest match.
- THE KICKER: No aspect of the algorithm depends on the number of sequences processed
-

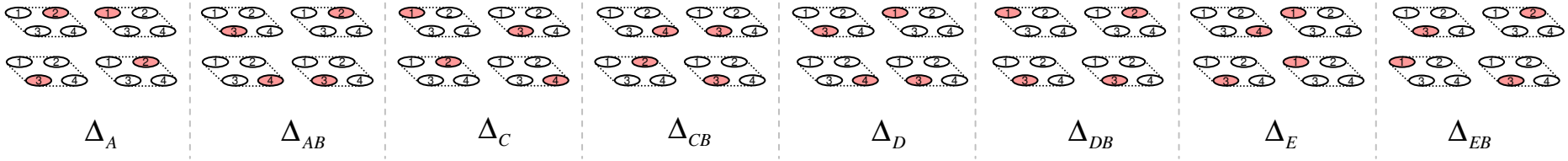
O(1) sequence comparison

with O(1) sequence learning !!!

The Normalized Code-Specific G is the Posterior Probability over the Set of all Hypotheses

All eight spatiotemporal contexts (hypotheses) that have existed during the network's entire history

- The code-specific G's during a re-presentation of [AB] computed right after B has been presented at t = 2



	Δ_A	Δ_{AB}	Δ_C	Δ_{CB}	Δ_D	Δ_{DB}	Δ_E	Δ_{EB}
$G(\Delta_A)$	**0.082**	**1.0**	**0.082**	**0.25**	**0.75**	**0.66**	**0.082**	**0.83**

- **The G the model actually uses "finds" the right answer**

$$ {}^{j}G_t = \frac{\sum_{z=1}^{{}^{j}Q} \overline{\chi}_t(z)}{{}^{j}Q} $$

$$ \overline{\chi}_t(z) = \max_{k \in CM_z} \chi_t(i) $$

$$ \chi_i = \Phi_i \Psi_i $$

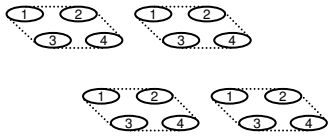
Likelihood **Evidence**

Scalability

- **THE L2 CODES BEING CHOSEN DURING LEARNING REPRESENT SPECIFIC SPATIOTEMPORAL CONTEXTS.**

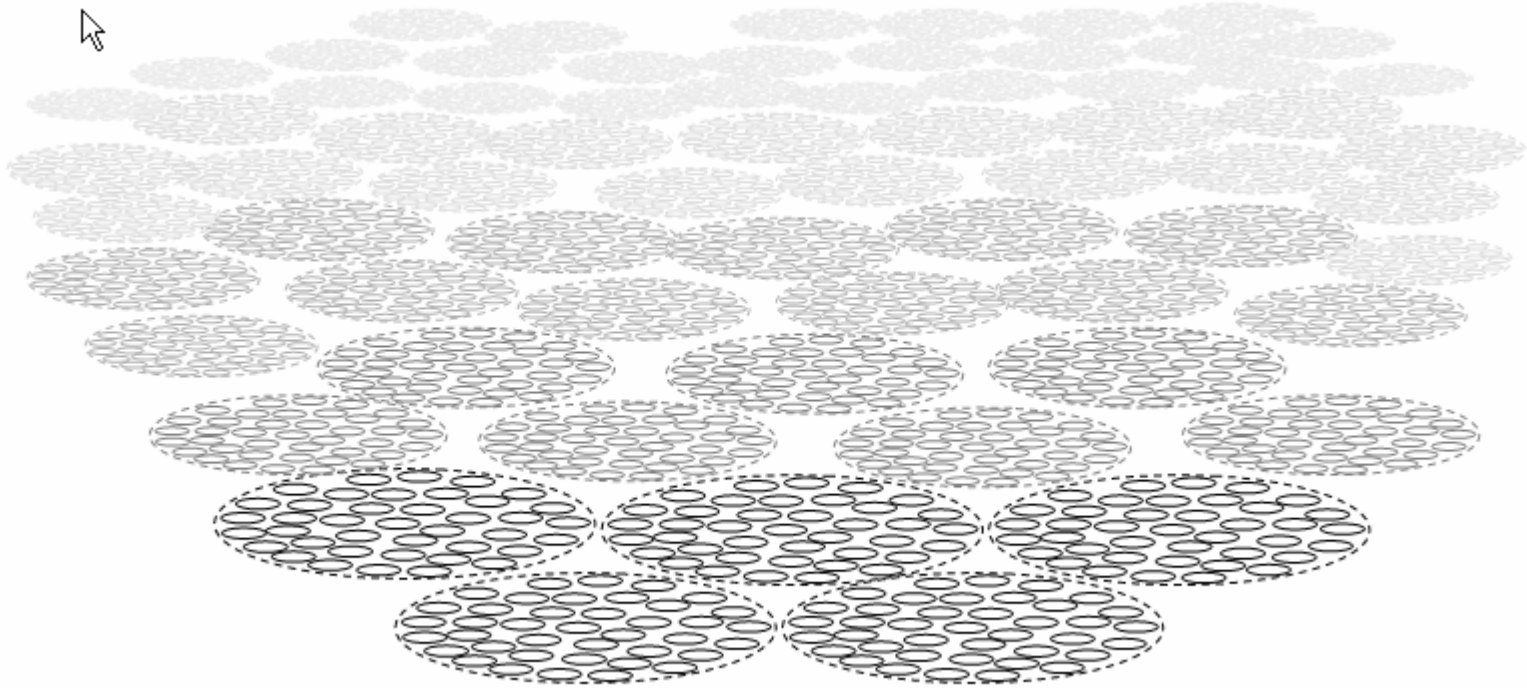
QUESTION: A new code for every spatiotemporal event experienced over the life of an organism !?

ANSWER: 1) An astronomical number of codes are available.



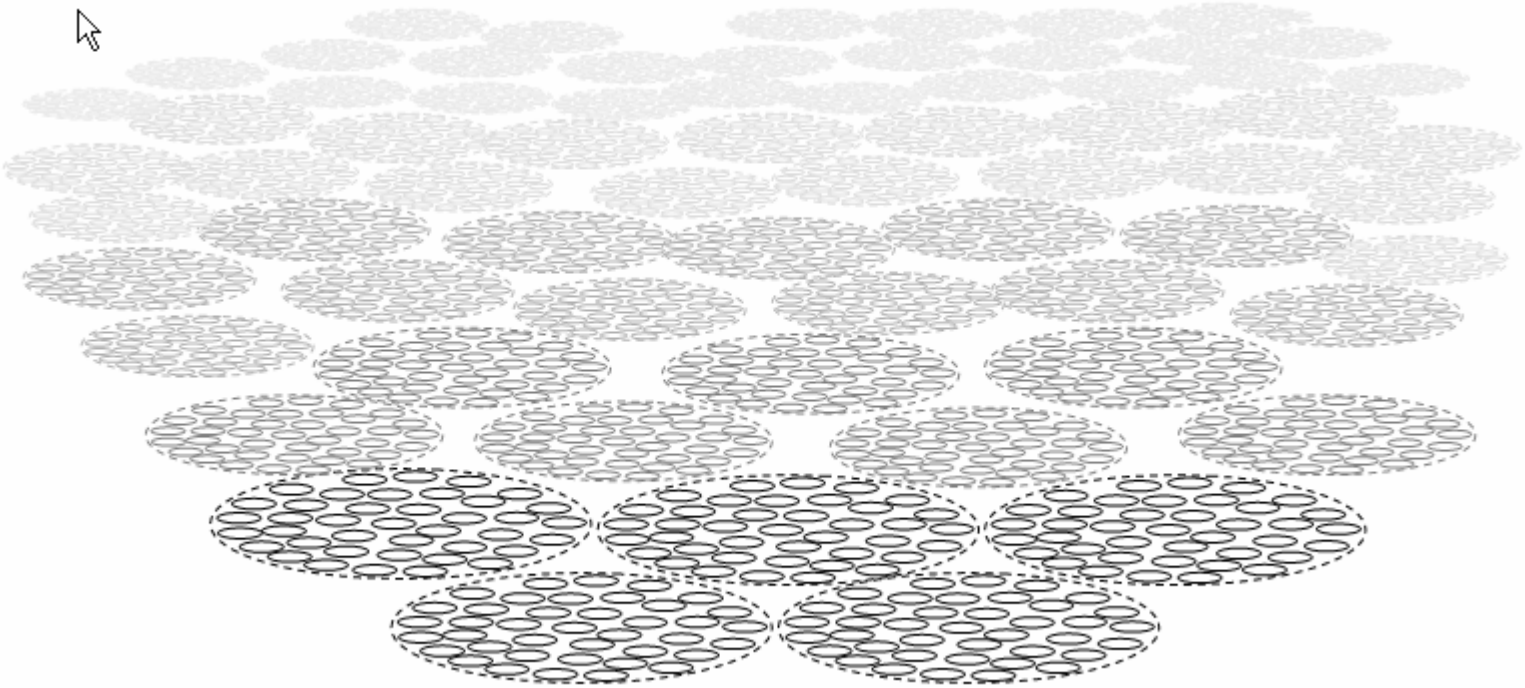
Toy example: $4^4 = 256$ unique L2 codes.

Hypercolumn: ~40^{70} unique codes



Scalability

- There may be 40^{70} unique codes over a hypercolumn
- But another important question: how many codes are there such that no two of them intersect at more than Z CMs?
    - Unsolved problem. It is sometimes referred to as the *central problem of coding theory*.
    - IT"S A LARGE NUMBER?
- Provides an answer to the binding problem. cf. O'Reilly, Busby & Soto (2003)



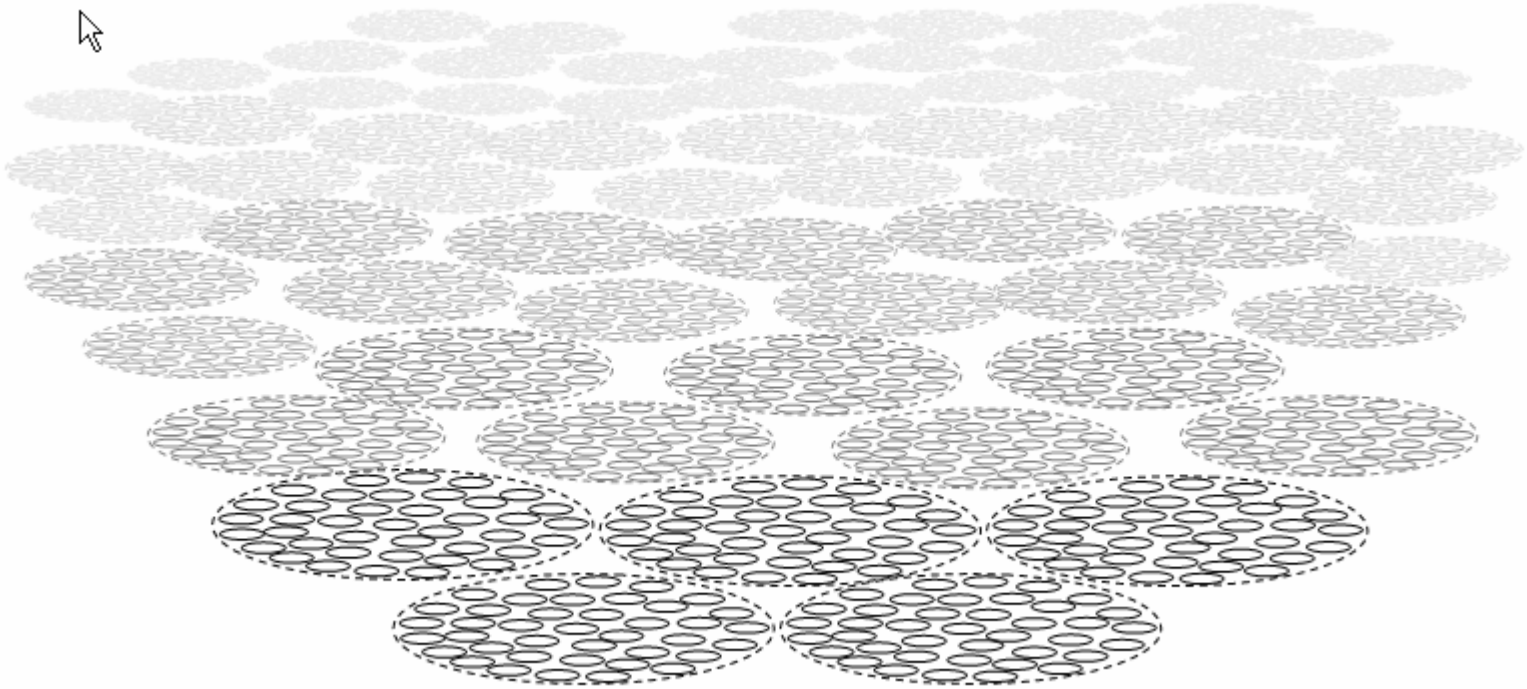
Scalability

QUESTION: A new code for every spatiotemporal event experienced over the life of an organism !?

ANSWER: 2) The model does *NOT* choose completely unique codes for every spatiotemporal event.

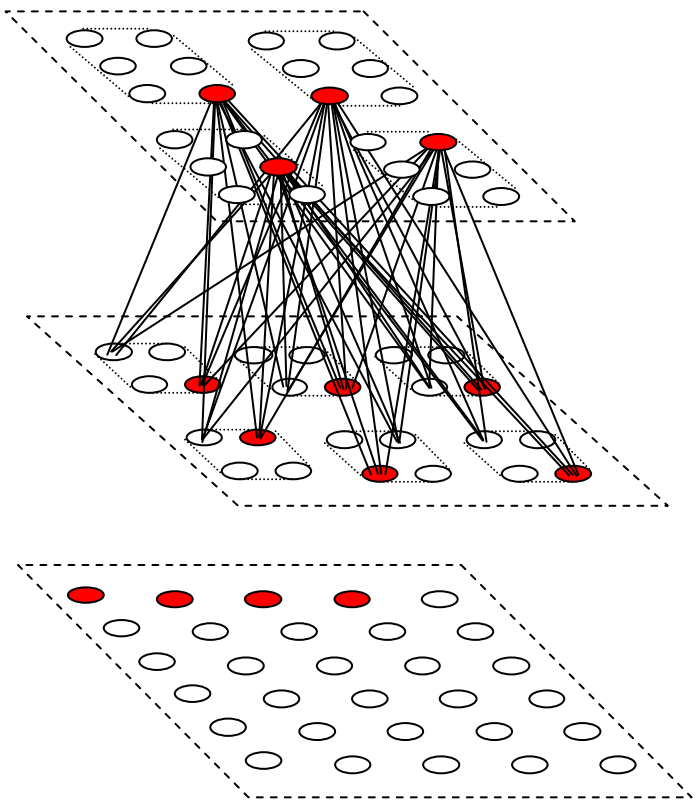
The codes chosen overlap with preexisting codes depending on the similarity of the events.

3) The hierarchical coding scheme, which allows any given L_J code to have different meanings depending on which overlying (L_{J+1}) code is active further abets scalability.



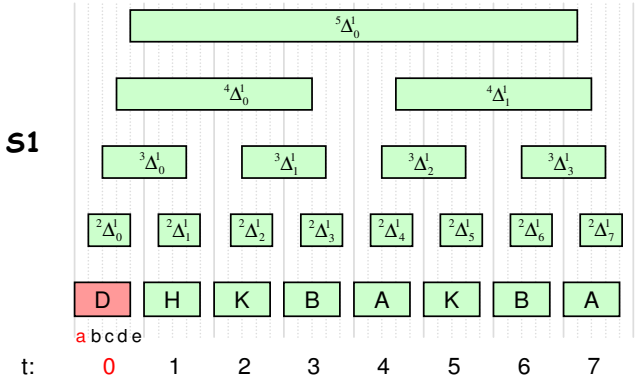
What about Persistence?

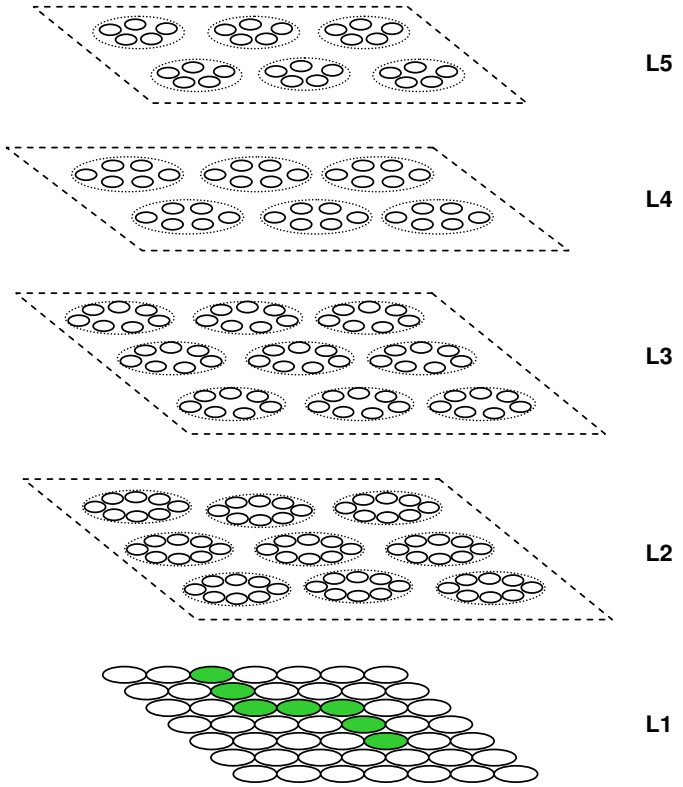
- During learning, persistence *overrules* the χ and ρ computations
  - Needed to allow/cause higher level representations to remain active during multiple successive lower-level representations.
- During recognition, the χ and ρ computations can overrule persistence
  - In order to do time-invariant recognition, representations at higher levels must be able to:
    - Turn off earlier than they did during the learning trial (to recognize speeded-up portions of a sequence)
    - Remain on longer than they did during the learning trial (to recognize slowed-down portions of a sequence)
- During Recall, persistence is honored



Processing Diagram

Learning: Embedding of a Hierarchical Spatiotemporal Trace





- The processing diagram helps illustrate the persistence property
- (click on the figure to open sub-presentation. Use up/down arrow keys to step through sub-presentation. Hit 'esc' key to exit sub-pres.)

Simulation Results

The experiments:

- use single-trial learning
- tested recall
- tested recognition

1. Illustration of time invariance mechanism
2. Small set of long non-complex sequences

Why Different Versions of *G* ?

Learned

S1: [A B C D E F]

S2: [D H K B A D]

S3: [A A D B A E]

Testing

[D H K B

…if we only compute ${}_{U}G$, then all three of the L2 codes for B will be equally active.

S1: ${}^{2}\Delta^{1}_{A}$ **${}^{2}\Delta^{1}_{B}$** ${}^{2}\Delta^{1}_{C}$ ${}^{2}\Delta^{1}_{D}$ ${}^{2}\Delta^{1}_{E}$ ${}^{2}\Delta^{1}_{F}$

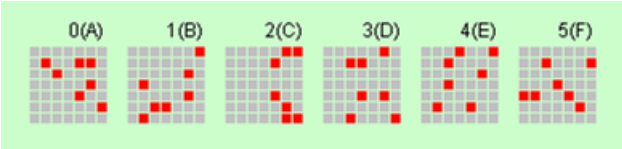
S2: ${}^{2}\Delta^{2}_{D}$ ${}^{2}\Delta^{2}_{H}$ ${}^{2}\Delta^{2}_{K}$ **${}^{2}\Delta^{2}_{B}$** ${}^{2}\Delta^{2}_{A}$ ${}^{2}\Delta^{2}_{D}$

S3: ${}^{2}\Delta^{3}_{A}$ ${}^{2}\Delta^{3}_{A}$ ${}^{2}\Delta^{3}_{D}$ **${}^{2}\Delta^{3}_{B}$** ${}^{2}\Delta^{3}_{A}$ ${}^{2}\Delta^{3}_{E}$

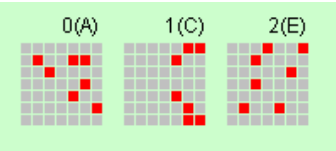
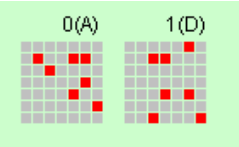
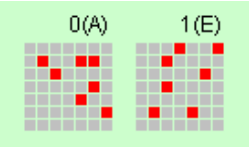
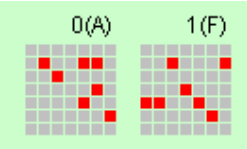
- The model should always try the most stringent version of *G*, ${}_{HUD}G$, first.
- It should then *back off* to progressively less stringent versions.

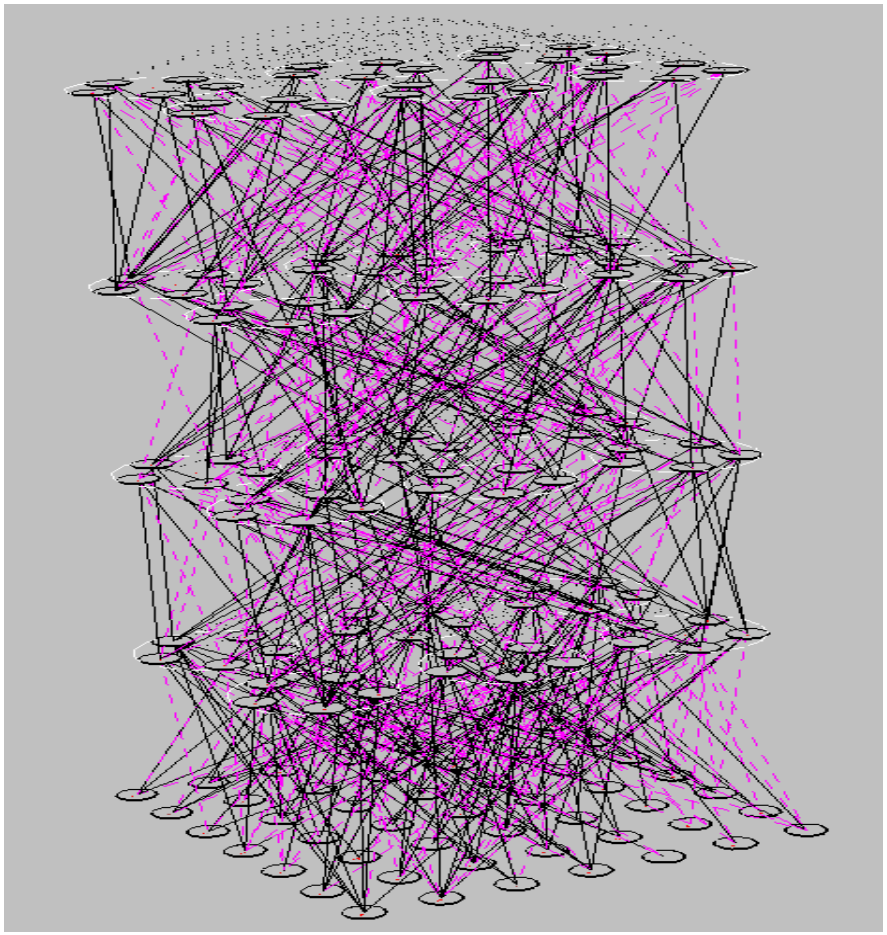
Experiment 1: Time Invariant Recognition

Present sequence



Recognition Tests

	Speed-up
	2x
	3x
	4x
	5x



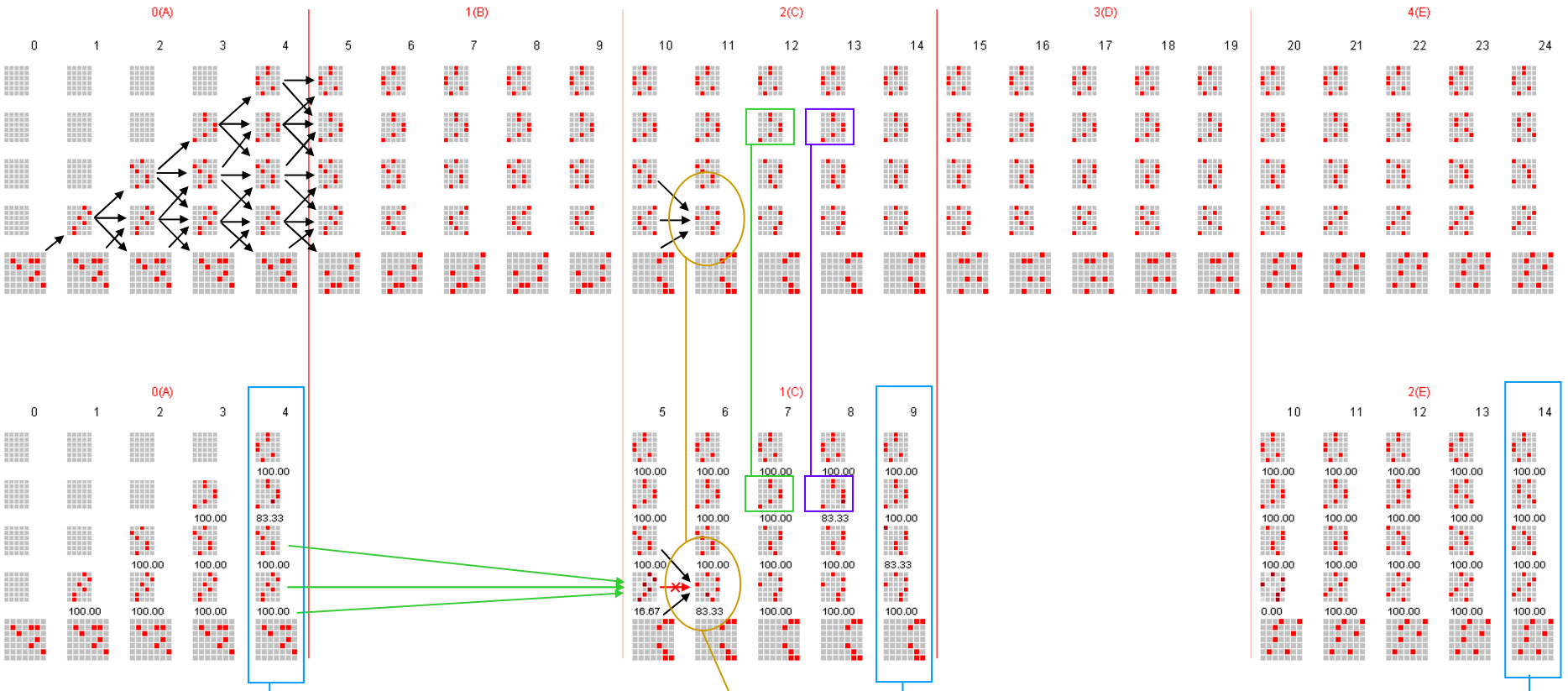
+ H-wts (Set)	+ U-wts (Set)	+ D-wts (Set)
30 (04.00%)	72 (08.00%)	NA
90 (12.00%)	108 (12.00%)	72 (08.00%)
150 (20.00%)	212 (23.56%)	138 (15.33%)
303 (40.40%)	241 (16.39%)	258 (28.67%)
NA	NA	241 (16.39%)

Level	# H-wts	# U-wts	# D-wts	Total wts
4	750	900	0	1,650
3	750	900	900	2,550
2	750	900	900	2,550
1	750	1,470	900	3,120
0	0	0	1,470	1,470

Learning & Recognition Traces: Speed-up: 2x

- Arrows depict movement of signals
- Accuracy measured by comparing corresponding (vertically aligned) codes

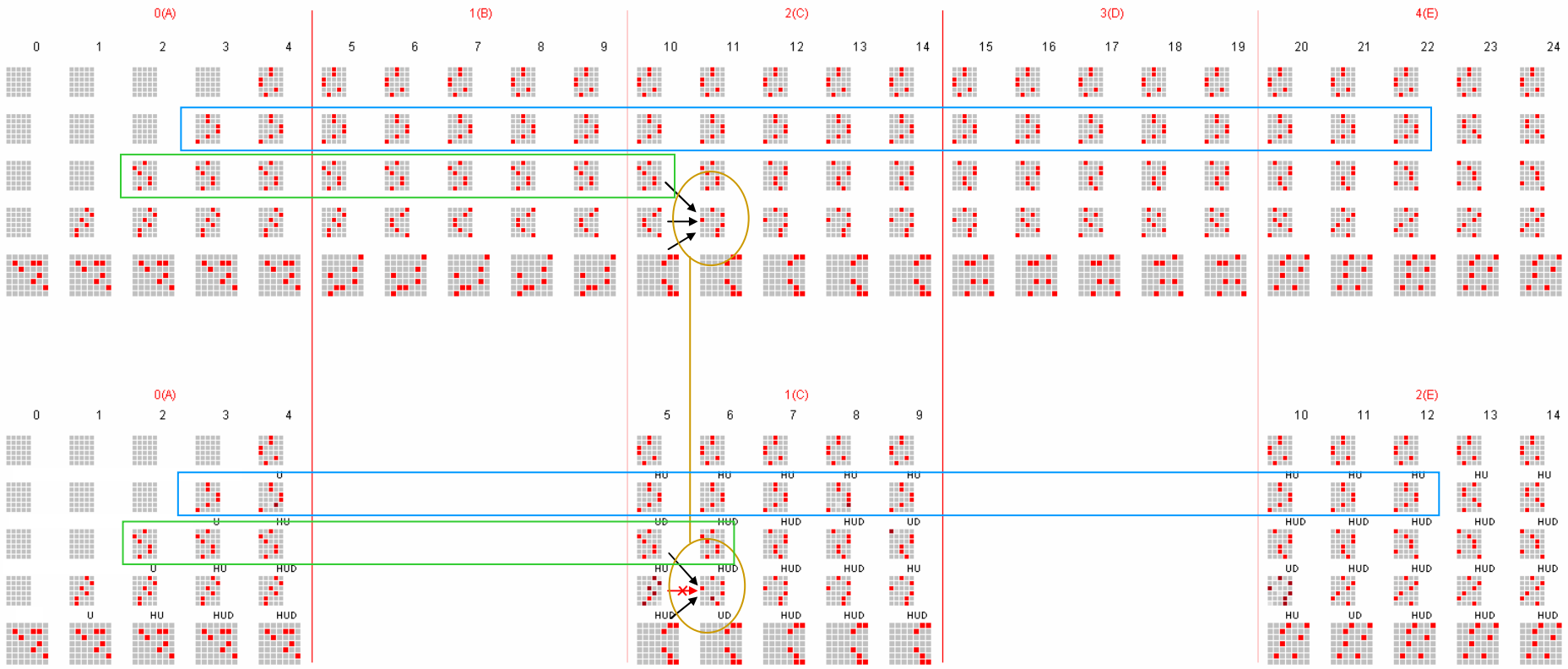
$$Acc. = \frac{Correct}{Q}$$



- 3-way match fails, so back off to 2-way match
- Accuracy measure only considers the final sub-step of each item.
- Green arrows indicate the same neural context.

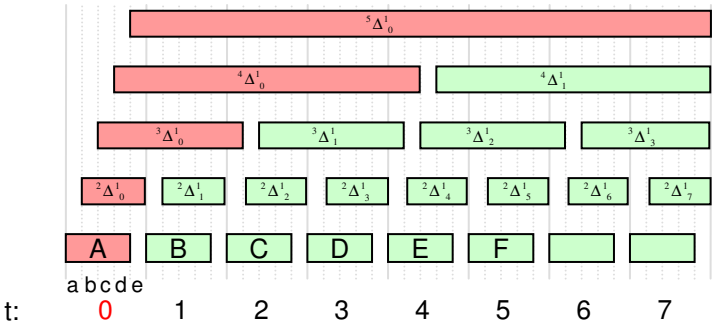
Learning & Recognition Traces: Speed-up: 2x

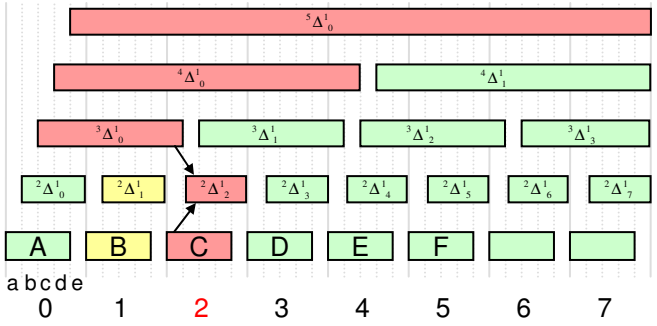
- The first L3 code lapses much earlier than it did during learning, so that the trace as a whole keeps pace with the sped-up input.
  Same for L4

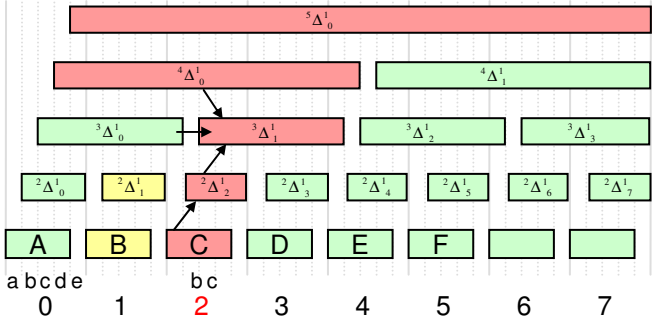


- 3-way HUD match is zero because of the time warping
- 2-way UD match is high because it ignores the H input.
- The less stringent UD match can be perfect for any L2 code that ever occurred under a given L3 code.
- This implements a tolerance to local time warping (including decimation)

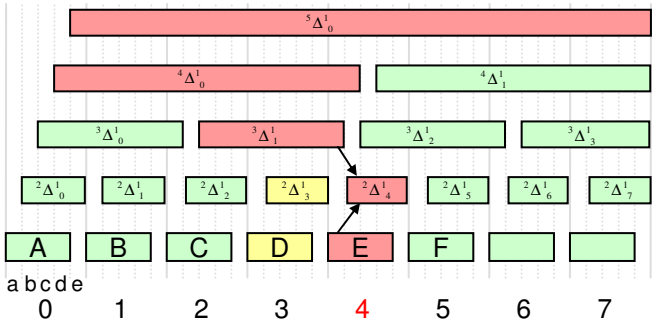
What's happening: Processing diagram

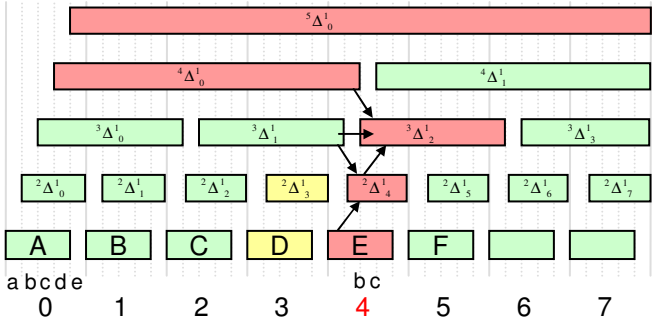




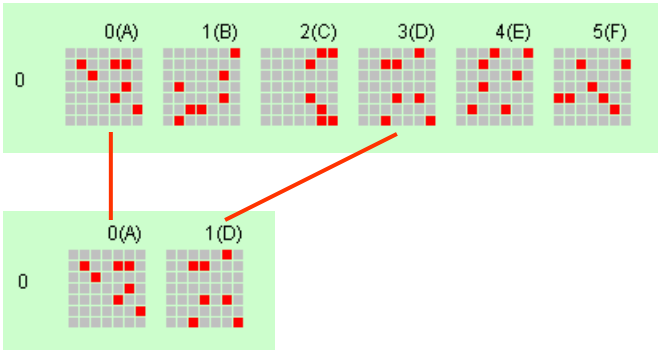


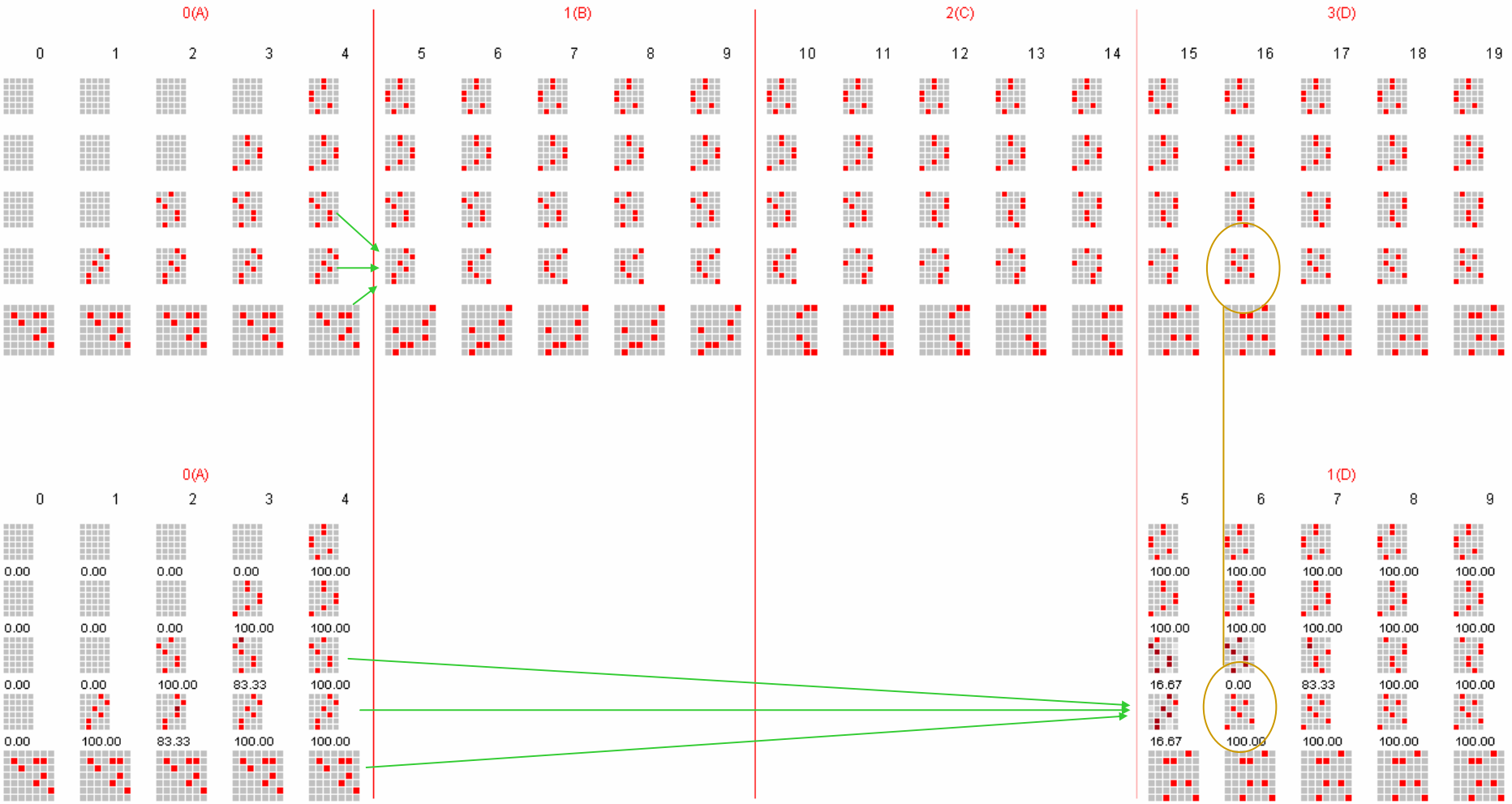
Didn't occur



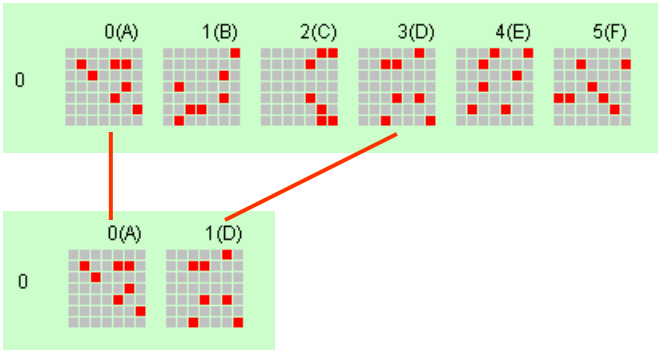


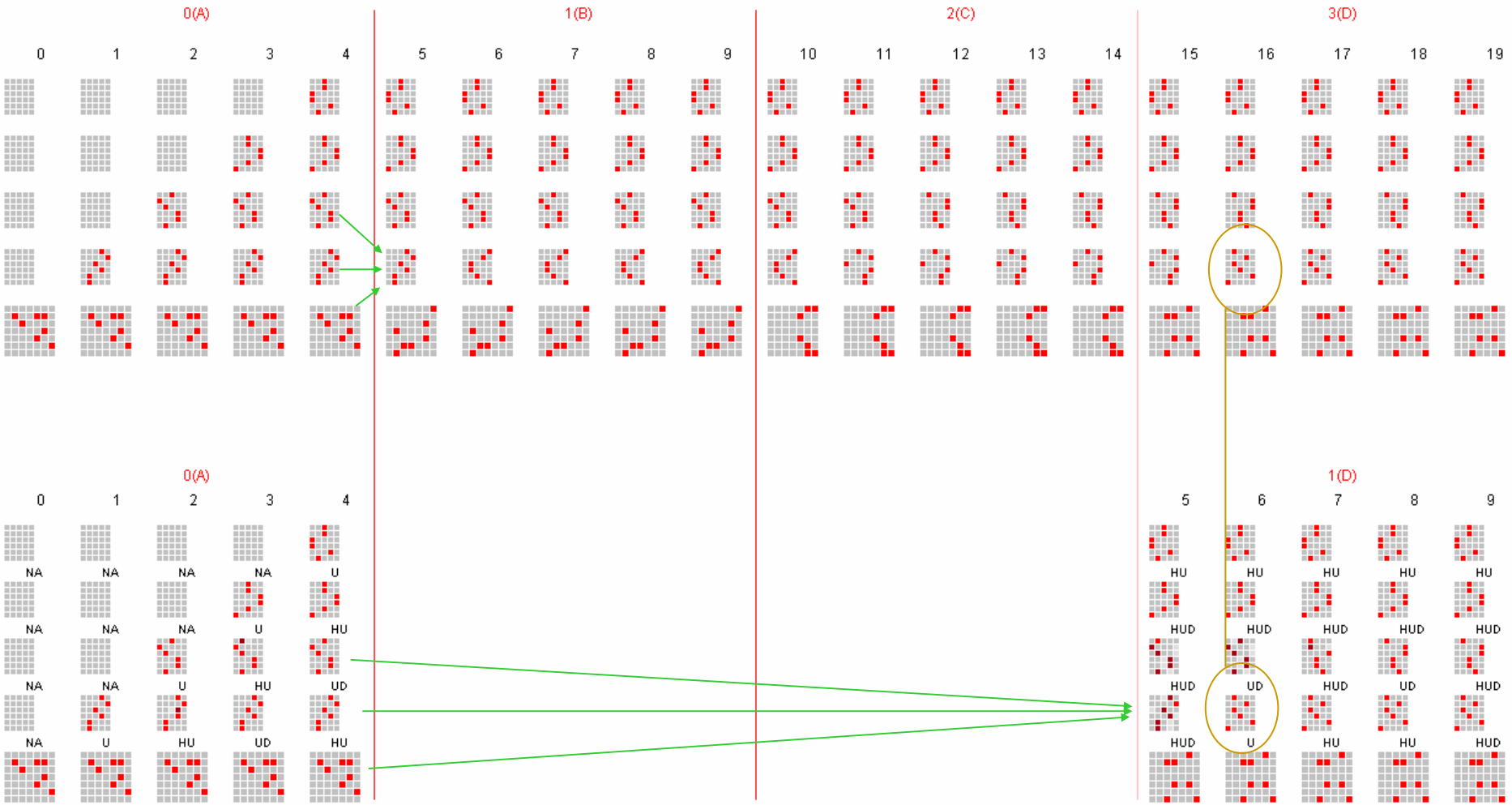
Speed-up: 3x


0(A)
1(B)
2(C)
3(D)
4(E)
5(F)
0
0(A)
1(D)
0

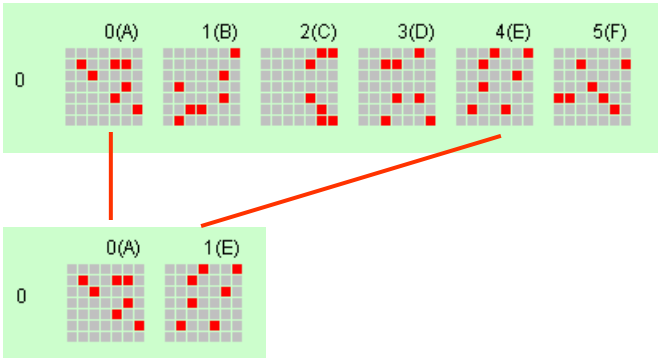

0(A)
1(B)
2(C)
3(D)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0(A)
1(D)
0 1 2 3 4
5 6 7 8 9
0.00 0.00 0.00 0.00 100.00
0.00 0.00 0.00 100.00 100.00
0.00 0.00 100.00 83.33 100.00
0.00 100.00 83.33 100.00 100.00
100.00 100.00 100.00 100.00 100.00
100.00 100.00 100.00 100.00 100.00
16.67 0.00 83.33 100.00 100.00
16.67 100.00 100.00 100.00 100.00

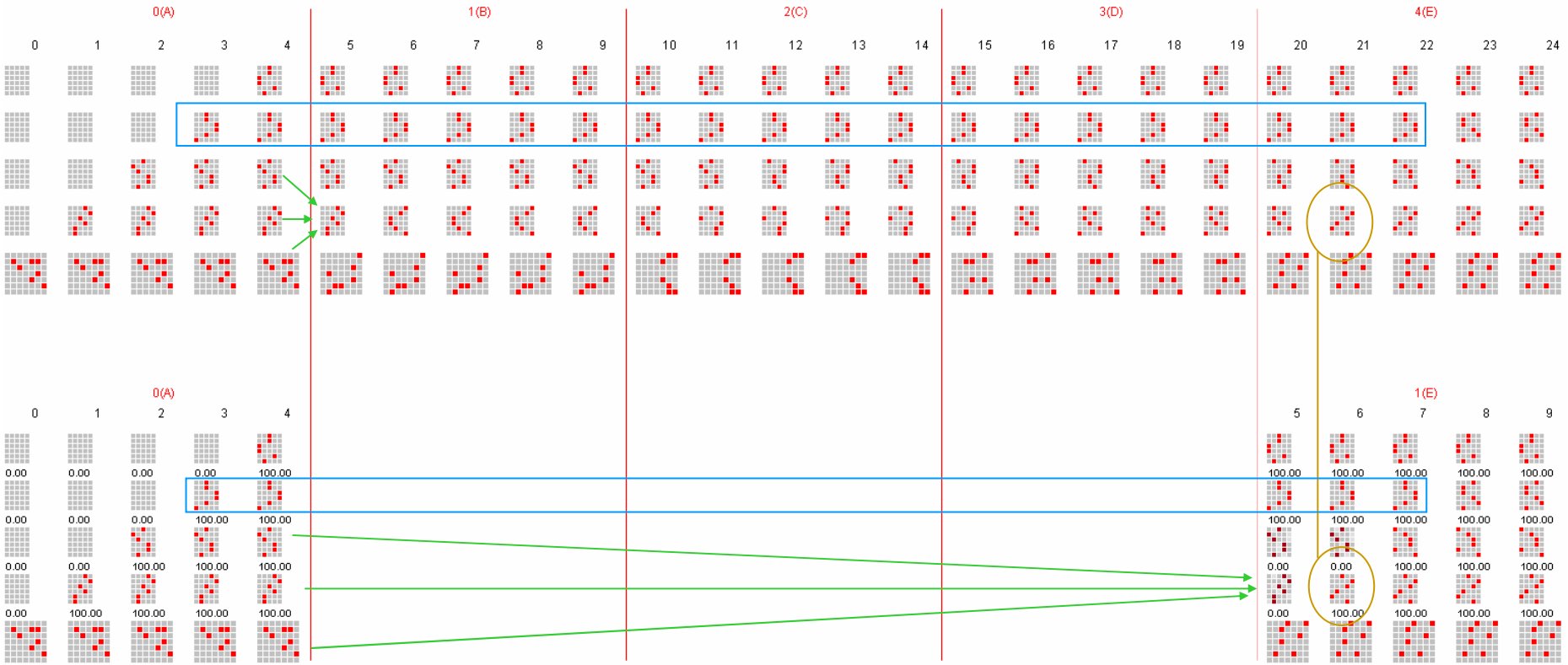
Speed-up: 3x


0(A)
1(B)
2(C)
3(D)
4(E)
5(F)
0
0(A)
1(D)
0

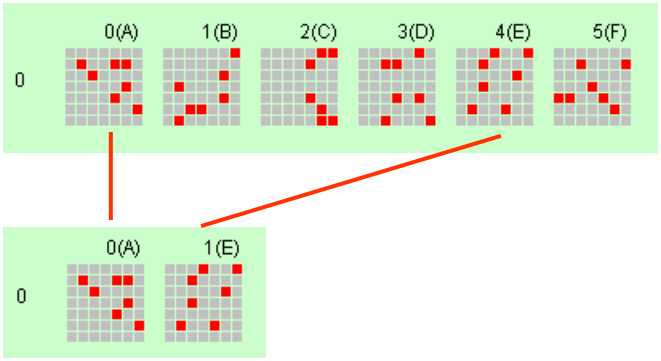

0(A)
1(B)
2(C)
3(D)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
0(A)
1(D)
0 1 2 3 4
NA NA NA NA U
NA NA NA U HU
NA NA U HU UD
NA U HU UD HU
5 6 7 8 9
HU HU HU HU HU
HUD HUD HUD HUD HUD
HUD UD HUD UD HUD
HUD U HU HU HUD

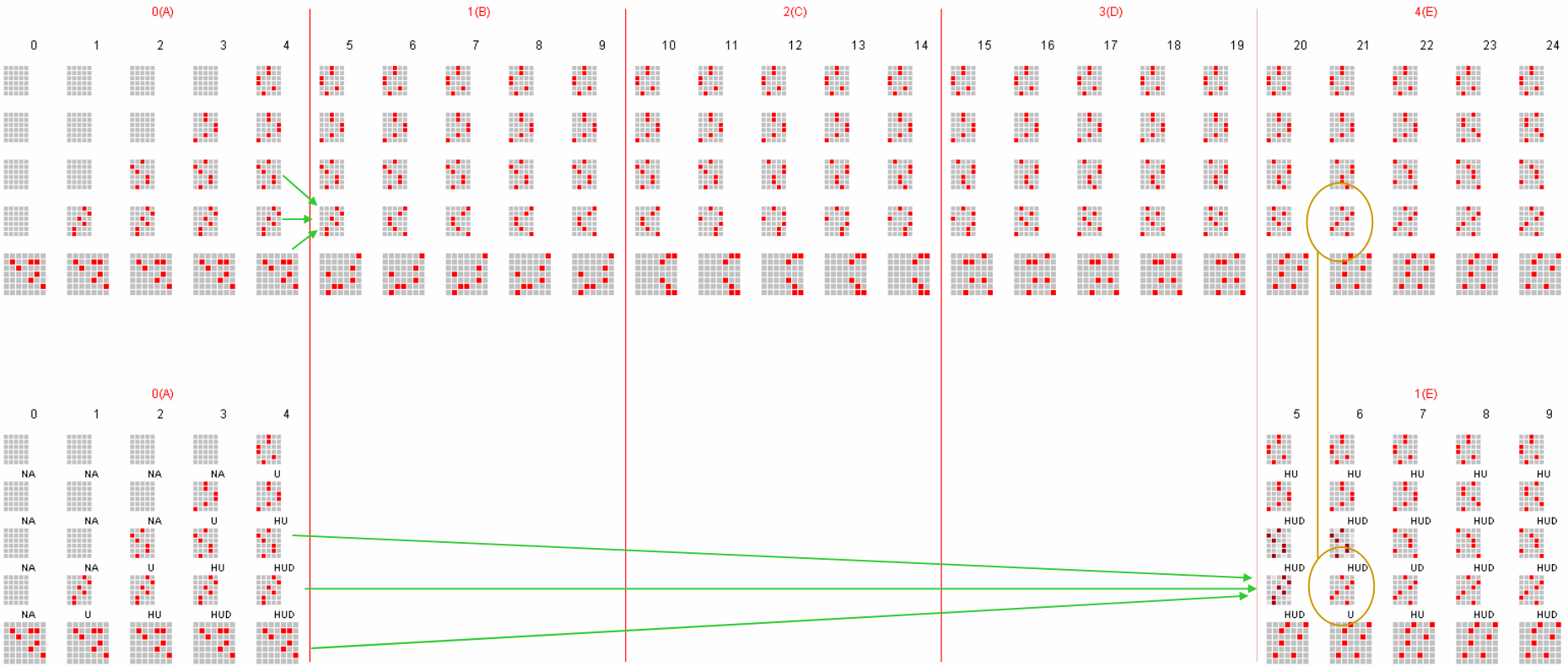
Speed-up: 4x



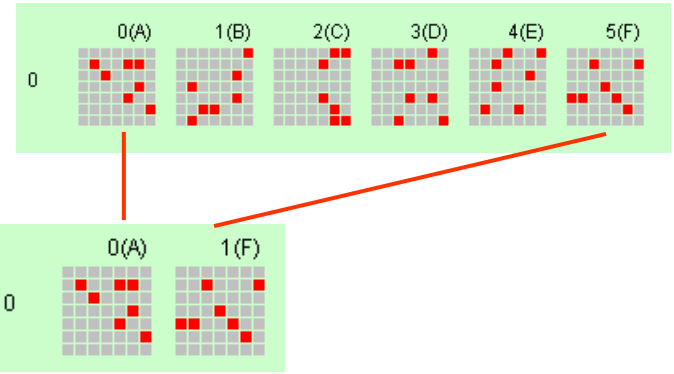


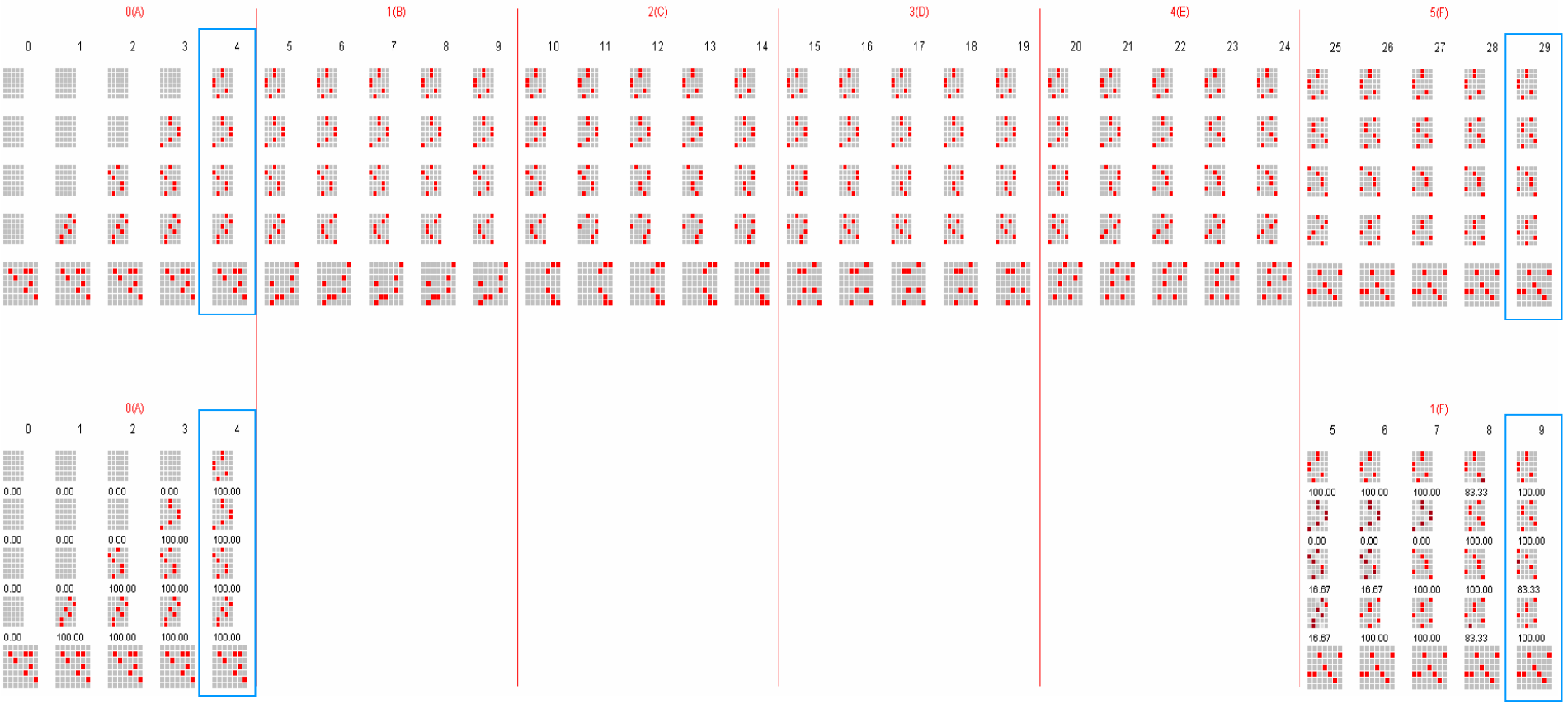
Speed-up: 4x


0(A)
1(B)
2(C)
3(D)
4(E)
5(F)
0
0(A)
1(E)
0

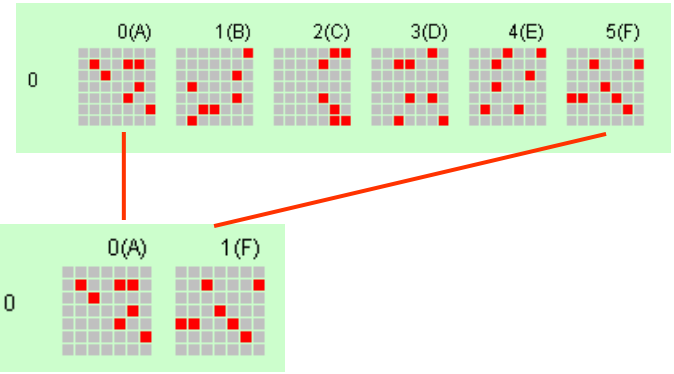

0(A)
1(B)
2(C)
3(D)
4(E)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
0(A)
1(E)
0 1 2 3 4
NA NA NA NA U
NA NA NA U HU
NA NA U HU HUD
NA U HU HUD HUD
5 6 7 8 9
HU HU HU HU HU
HUD HUD HUD HUD HUD
HUD HUD UD HUD HUD
HUD U HU HUD HUD

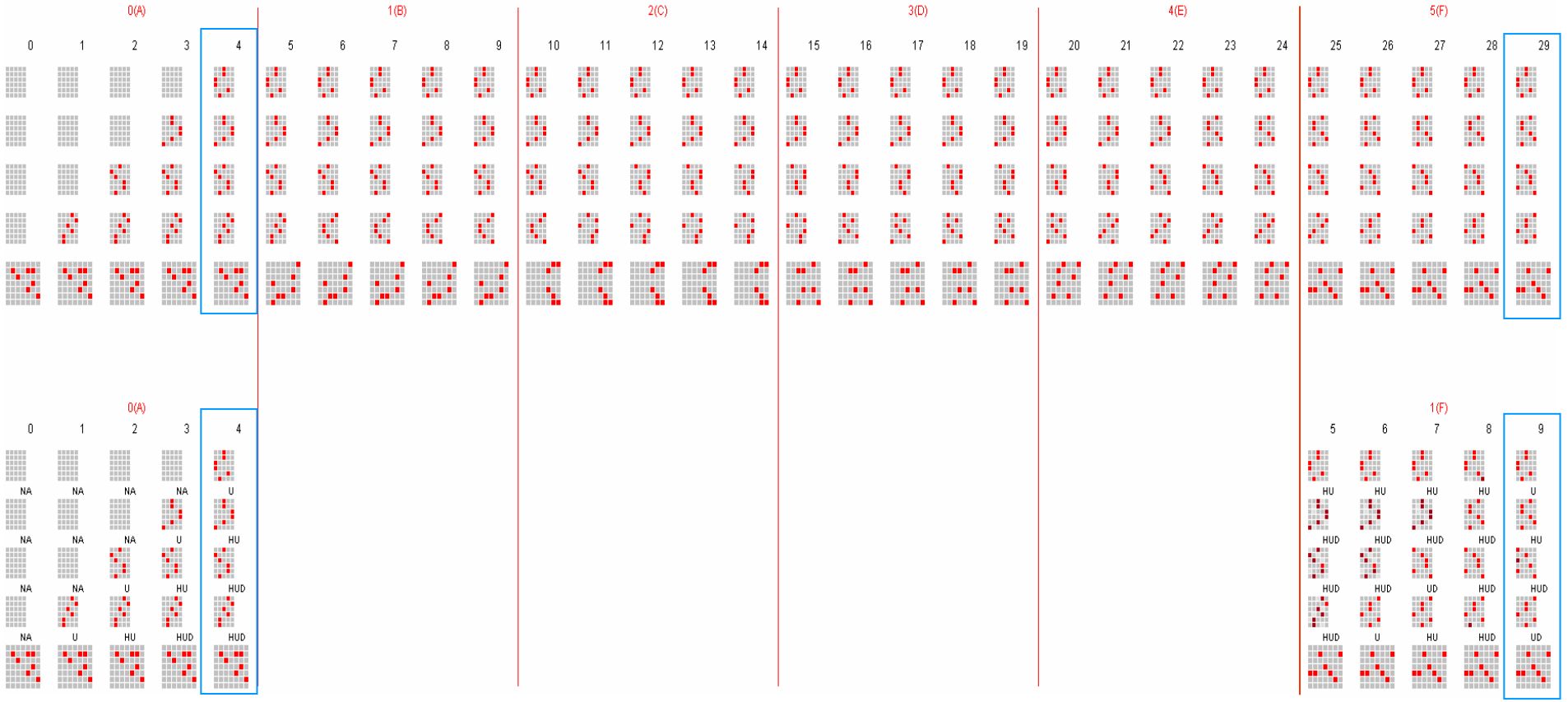
Speed-up: 5x





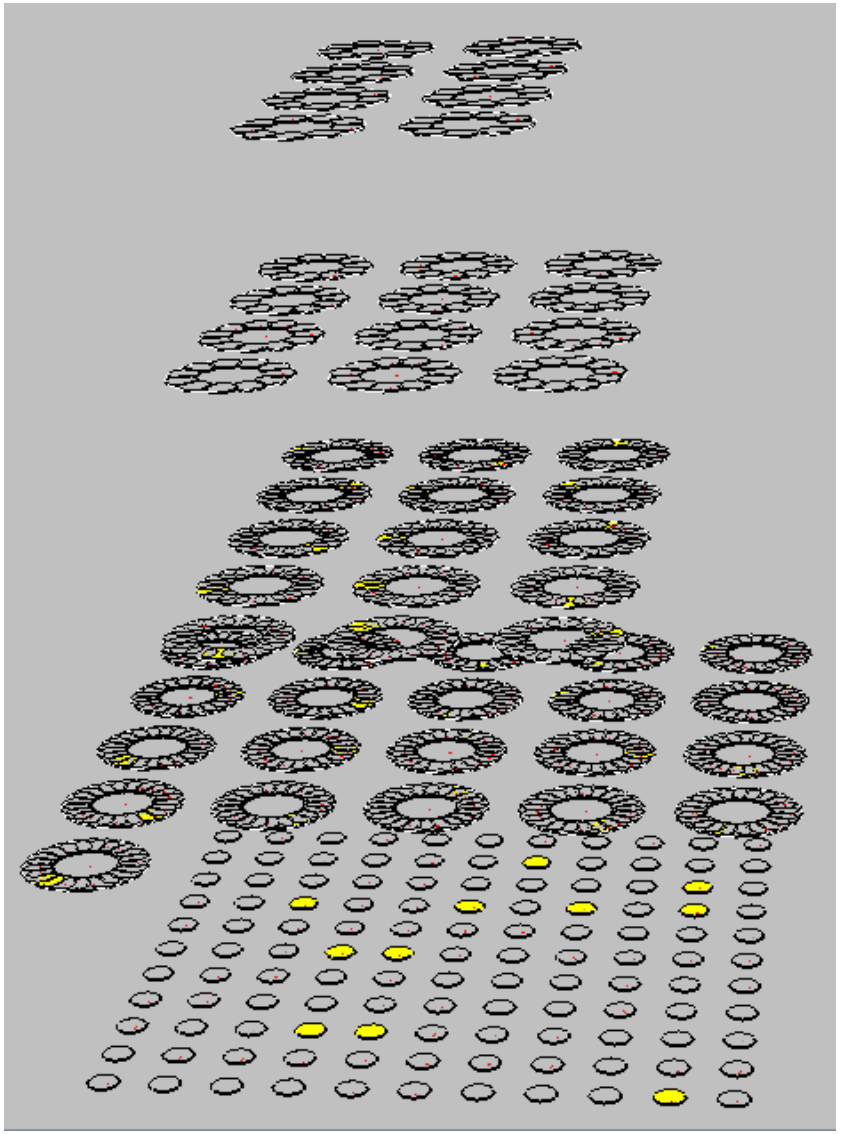
Speed-up: 5x





Experiment 2 – Long Sequences

		Exp 2
Number of Episodes		5
Number of Items		30
Features per Item		11 / 121
Total Information (bits)		8065.75
Total Weights		660,258
Bits per Weight		0.012
Recall (%)	5	100.00
	4	100.00
	3	100.00
	2	100.00
	1	99.33
Recognition (%)	5	93.07
	4	93.31
	3	93.88
	2	92.88
	1	NA
Warp Rates		2.5, 1.25, 2.0, -1.75, 1.25, 2.25
Spatial Degradation		~9% (1/11)



Training and Test Data: Experiment 2

~9% noise

- Warp regions

Original Training Episodes

Recognition Test Episodes

Summary I

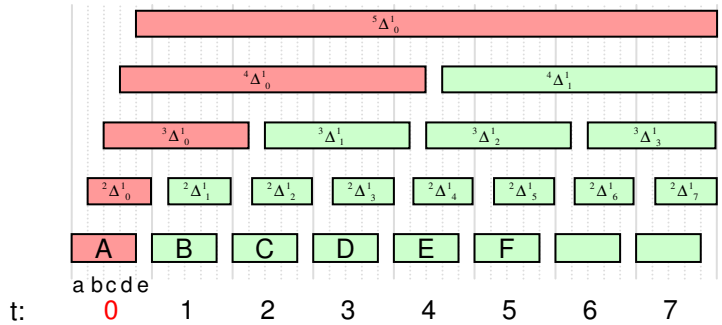
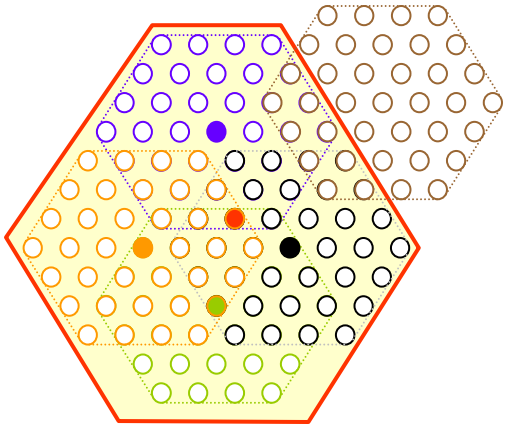
1. I've described a biologically plausible, unsupervised sequence memory
2. The use of sparse distributed binary representations allows:
   - similarity to be represented by degree of overlap between sets; a discrete measure.
   - high storage capacity (in principle).
3. O(1) sequence comparison, given O(1) learning
4. Hierarchality and progressive persistence:
   - enforce temporal chunking
   - yield robust non-uniform time invariance
5. Results showing:
   - recall of individual exemplars, AND
   - recognition of noisy and time-warped sequences
   - EVEN THOUGH all that were explicitly stored were the individual exemplars (one trial each) proves that the model is storing and using higher-order statistics (similarity structure) of the input set
     - Thus, a monolithic substrate storing both episodic and generic (semantic) memory

Summary II

- **Consistent with much neurological data**
  - Sparse activity
    - Olshausen & Field (1996, 2004), Hahnloser et al. (2002)
  - CMs = minicolumns ?
  - Exhibits many desirable properties while assuming only binary activations and binary weights
  - Stable, repeatable spatiotemporal Traces
    - Yuste
  - Memory traces are stable from their first occurrence.
    - McNaughton, Wilson
  - Longer persistences have been shown for cortical cells in deeper regions, e.g. PFC
    - 30+ years of working memory studies,
    - Uusitalo et al. data (evoked potentials)

Future Work

- 'Same prefix' problem
- Event-based updating of representations, as opposed to clocked updating of representations
- Temporal Segmentation Problem



- Multi-minicolumn competition to achieve:
  - sparser activity
  - another 'meta' level of statistical structure over the set of representations
  - could this level of structure correspond to the hypercolumn?
  - more realistic connectivity rates
- Generalization to explaining other invariances, e.g., viewpoint-invariant recognition.