The brain's operating system

Gabriel Kreiman

Gabriel.kreiman@tch.harvard.edu

klab.tch.harvard.edu

Center for Brains, Minds and Machines

Charlotte Moerman

Camille Gomez

Martin Schrimpf

Richard Born

Jojo Nassi

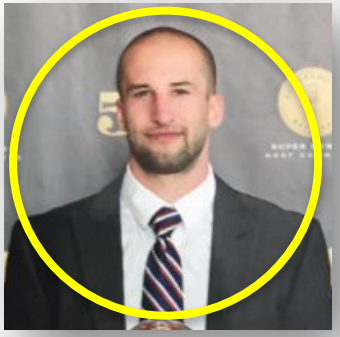
Laura Groomes

Joseph Madsen

Hanlin Tang

Thomas Miconi

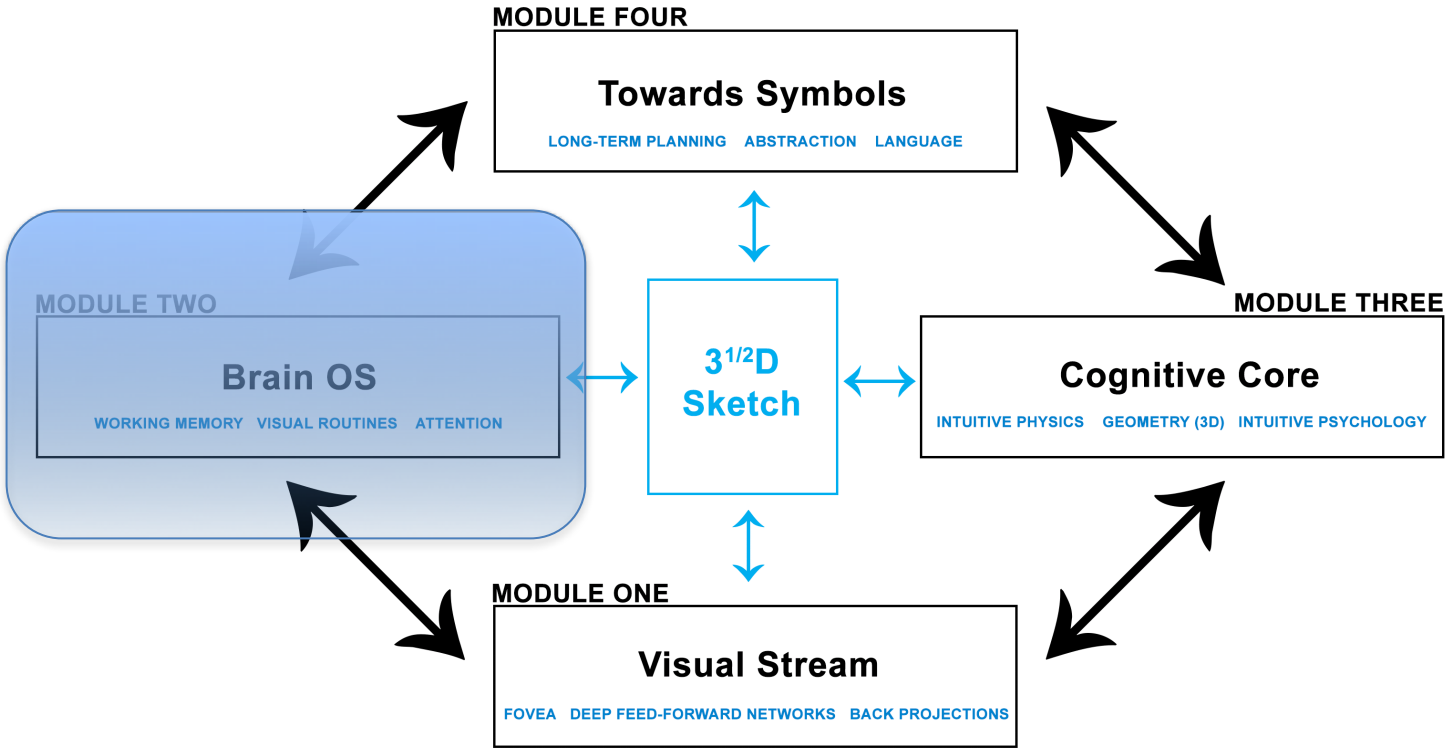
Bill Lotter



David Cox

Stan Anderson

Architecture of Visual Intelligence and its Modules



An image is worth a million words

Who are they?
What is there?
Where are they?
Where is Obama's foot?
What are they doing?
Why are they doing it?
Are they friends?
How old are they?
Who is looking at whom?
What happened before?
What will happen next?
Search for all the mirrors
How many shoes can you see?

Describe the scene

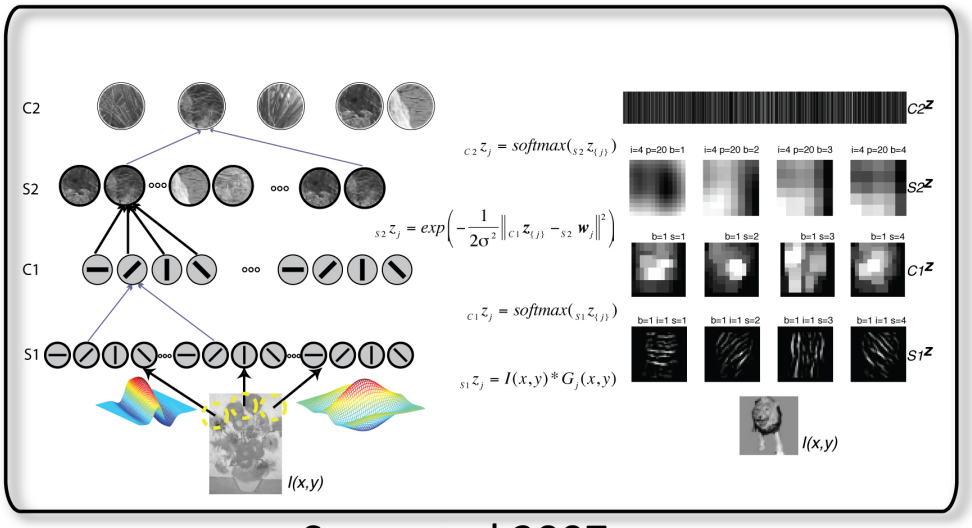
Biologically-inspired computations are powerful

Over millions of years of evolution, "interesting" solutions to difficult problems have emerged through changes in neuronal circuits

- Hardware and software that work for many decades
- Parallel computation (with serial bottlenecks)
- Reprogrammable architecture
- Low power
- Single-shot learning
- "Discover" structure in data
- Fault tolerance
- Robustness to sensory transformations
- Component interaction and integration of sensory modalities

Algorithms, solutions

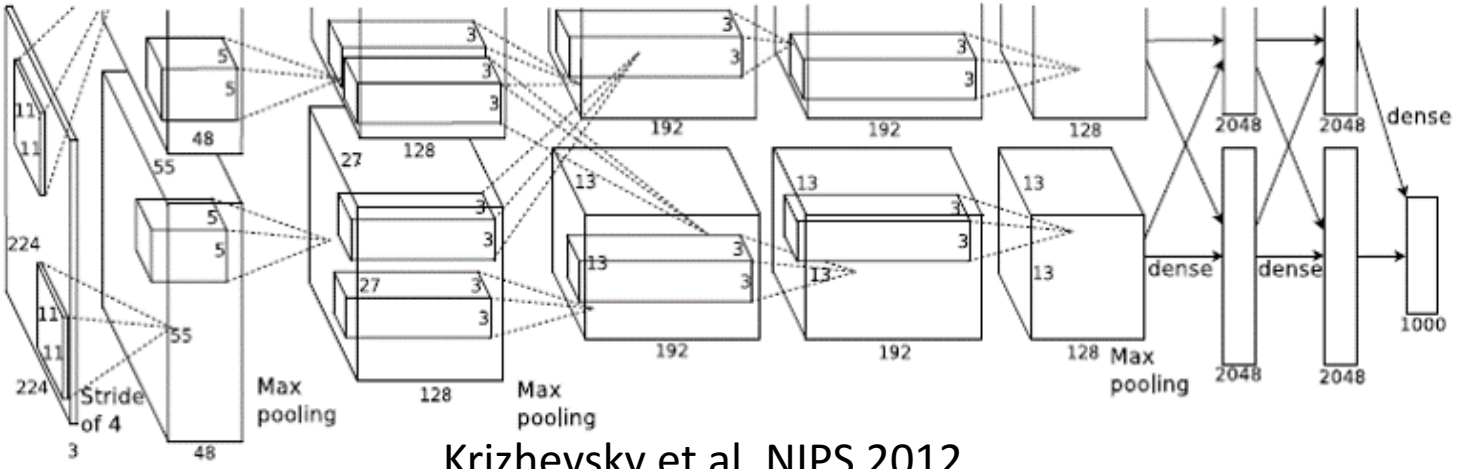
Bottom-up models of object recognition (the first ~100-150 ms)



Fukushima, Mel, Olshausen, LeCun, Riesenhuber, Rolls, DiCarlo, …

Serre et al 2007

Deep convolutional networks



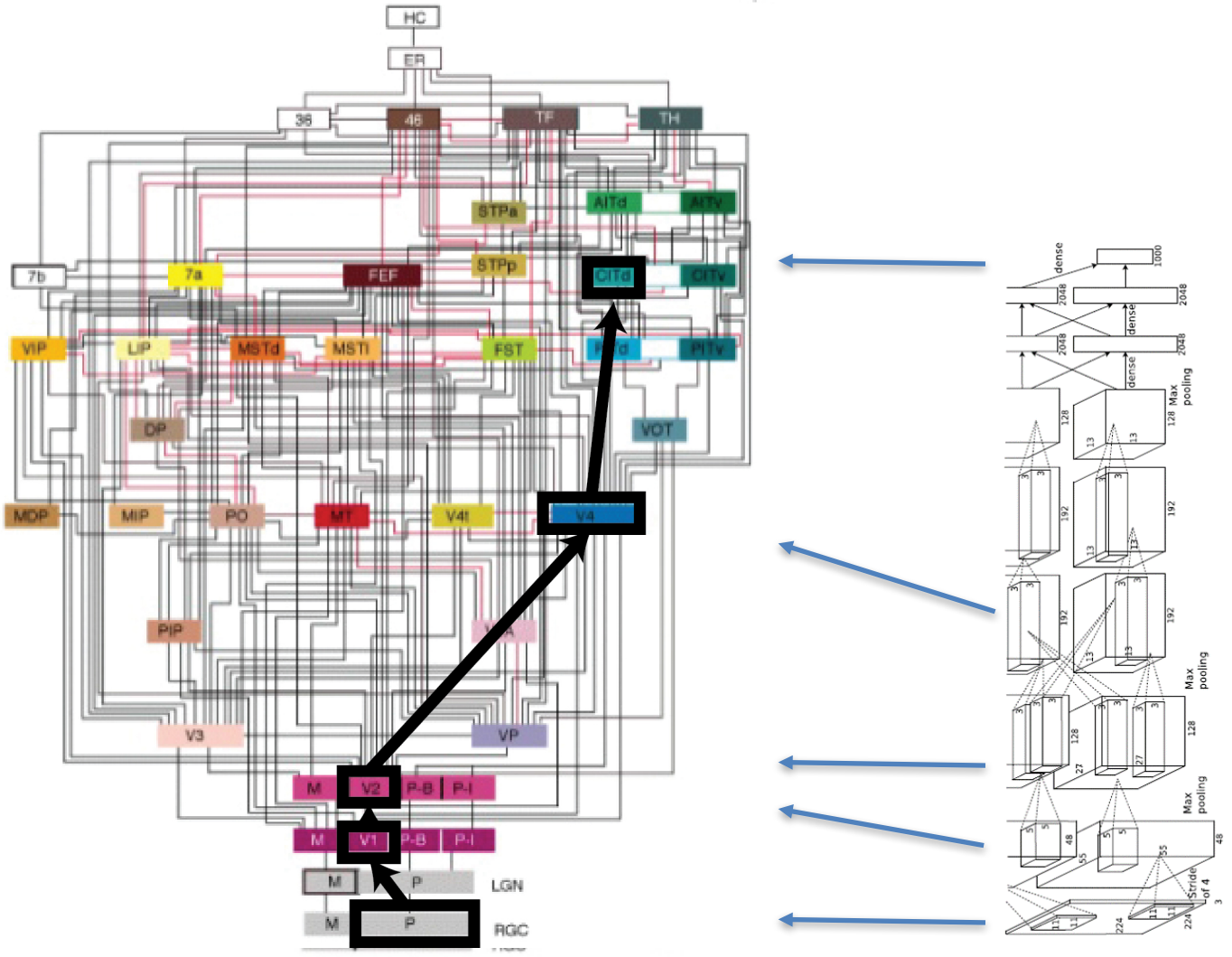
Krizhevsky et al, NIPS 2012

There is more. Much more.

A schematic diagram of visual cortex connections in macaque monkeys

State-of-the-art computer vision architectures mimic a small fraction of visual cortex

There is much more to visual cortex. And there is the rest of the brain …



Felleman and Van Essen 1991

Visual cognition: a sequence of routines

Shimon Ullman. 1984. Visual routines

- Step 1. Bottom-up representation of the environment
- Step 2. Visual routines, sequences of elementary operations → Using a fixed set of basic operations, the visual system can assemble different routines to extract an unbounded variety of shape properties and spatial relations

Visual cognition: a sequence of routines

Divide et impera

Operations

- Candidate labels for foveated region
- Inference and pattern completion
- Candidate representation of the periphery
- Select target for active sampling (eye movements)
- Determine spatial relations
- Temporal comparisons
- Store information
- Retrieve previously stored information
- Make temporal predictions

Visual cognition: a sequence of routines

What are they doing?

1. Extract initial sensory map → Call `VisualSampling`
2. Propose image gist → Call `RapidPeripheralAssessment`
3. Propose foveal objects → Call `FovealRecognition`
4. Inference from 1+2+3 → Call `PatternCompletion`
5. Temporary information storage → Call `VisualBuffer`
6. Task-dependent sampling → Call `TargetAttentionProposal`
7. Active sampling → Call `EyeMovementImplementation`
8. Detect people → Call `PeopleDetection`
9. Determine spatial relationships → Call `SpatialRelationships`
10. Repeat steps 3+4+5
11. Repeat steps 6-7
12. Repeat 8-9
13. Got answer? → Call `TaskTerminationDecision`
14. If satisfactory, answer the question → Call `TaskReport`

Visual cognition: a sequence of routines

What are they doing?

1. Extract initial sensory map → Call `initial sampling`
2. Propose image gist → Call `rapid peripheral assessment`
3. Propose foveal objects

   `[PreliminaryLabels]=FovealRecognition(SensoryInput, History)`

   i. Query V1, V2, V4, PIT, AIT from `SensoryInput`
   ii. Integrate with temporal context from `History`
   iii. Integrate with spatial context from `History`
   iv. Select specific classifier
   v. Upload information to classifier
   vi. Propose initial labels → `PreliminaryLabels`
4. Inference from 1+2+3 → Call `pattern completion`
5. Temporary information storage → `Call visual buffer`
6. Task-dependent sampling → Call `target eye movement proposal`
7. Active sampling → Call `eye movement implementation`
8. Detect people → Call `people detection`
9. Determine basic spatial relationships → Call `spatial relationships`
10. Repeat steps 3+4+5
11. Repeat steps 6-7
12. Repeat 8-9
13. Got answer → Call `task termination evaluation`
14. If satisfactory, answer the question → Call `task report`

Example problem: What are they doing?

Visual routines: Interpreting the task/goal

Language, symbols, abstractions, context, plans **[Module 4]**

Example problem: What are they doing?

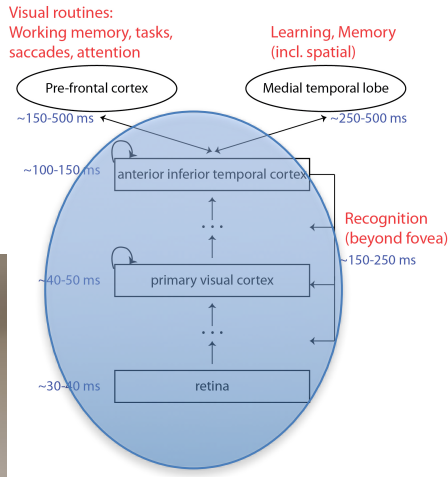
Visual routines: Initial glimpse

Detecting objects near fixation, getting the gist of a scene

[Module 1]

Ventral visual cortex

~150 ms




```
VisualSampling

PeripheralAssessment

FovealRecognition

PatternCompletion
```

Active sampling via eye movements

Eye movements

Subject 1

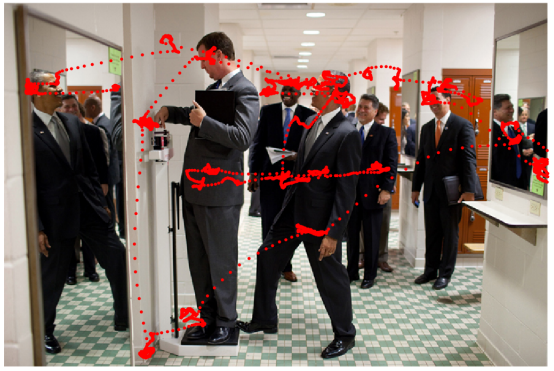
Subject 2

Subject 3

Subject 4

Subject 5

Subject 6



10 deg

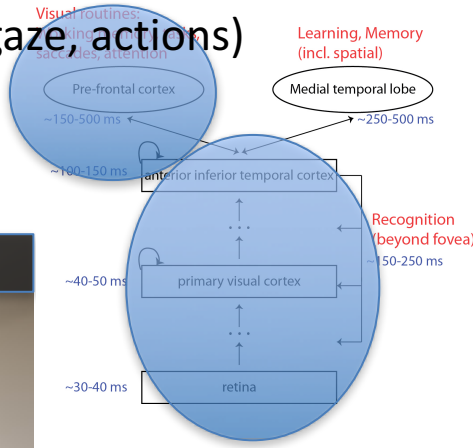
Example problem: What are they doing?

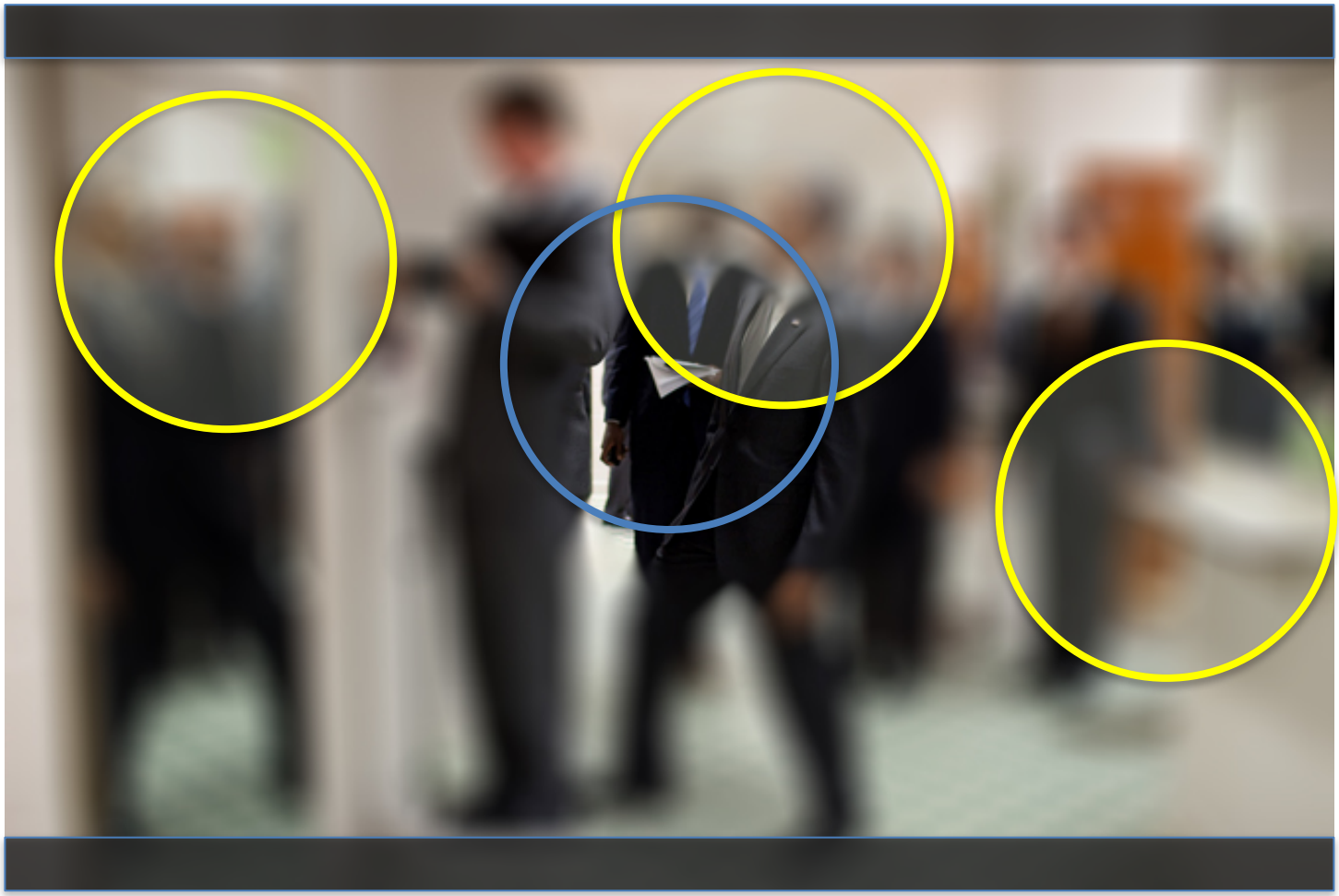
Visual routines: Where to look next?

Bottom-up saliency, Common sense knowledge, Visual search (faces, gaze, actions)

Interactions between ventral visual cortex and pre-frontal cortex

150-250 ms





`VisualBuffer`

`TargetAttention`

`EyeMovement`

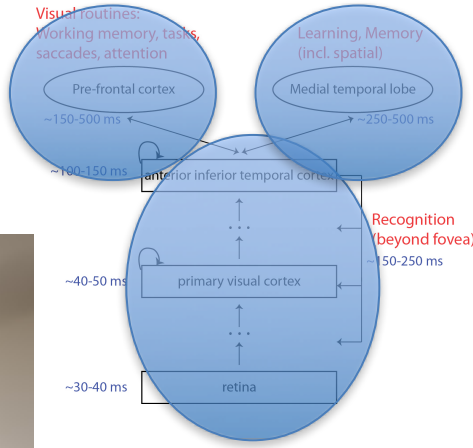
Example problem: What are they doing?

Visual routines: Recognition, Spatiotemporal relationships

Object recognition, Working memory

Interactions between ventral visual cortex, pre-frontal cortex, MTL

200-350 ms



```
PeopleDetection

SpatialRelationships

FovealRecognition

PeripheralAssessment
```

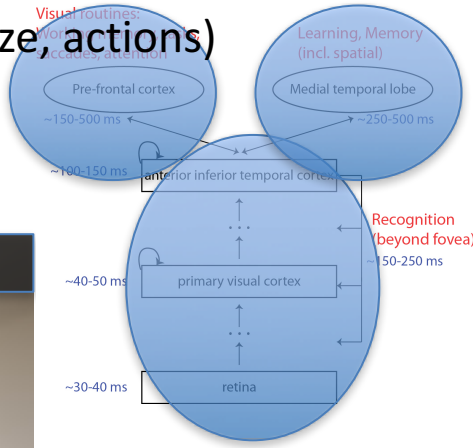
Example problem: What are they doing?

Visual routines: Where to look next?

Bottom-up saliency, Common sense knowledge, Visual search (faces, gaze, actions)

Interactions between ventral visual cortex and pre-frontal cortex

350-400 ms



```
VisualBuffer

TargetAttention

EyeMovement
```

Example problem: What are they doing?

Visual routines: decision making and output

Illusion of full image understanding
When to stop?
Cast an answer

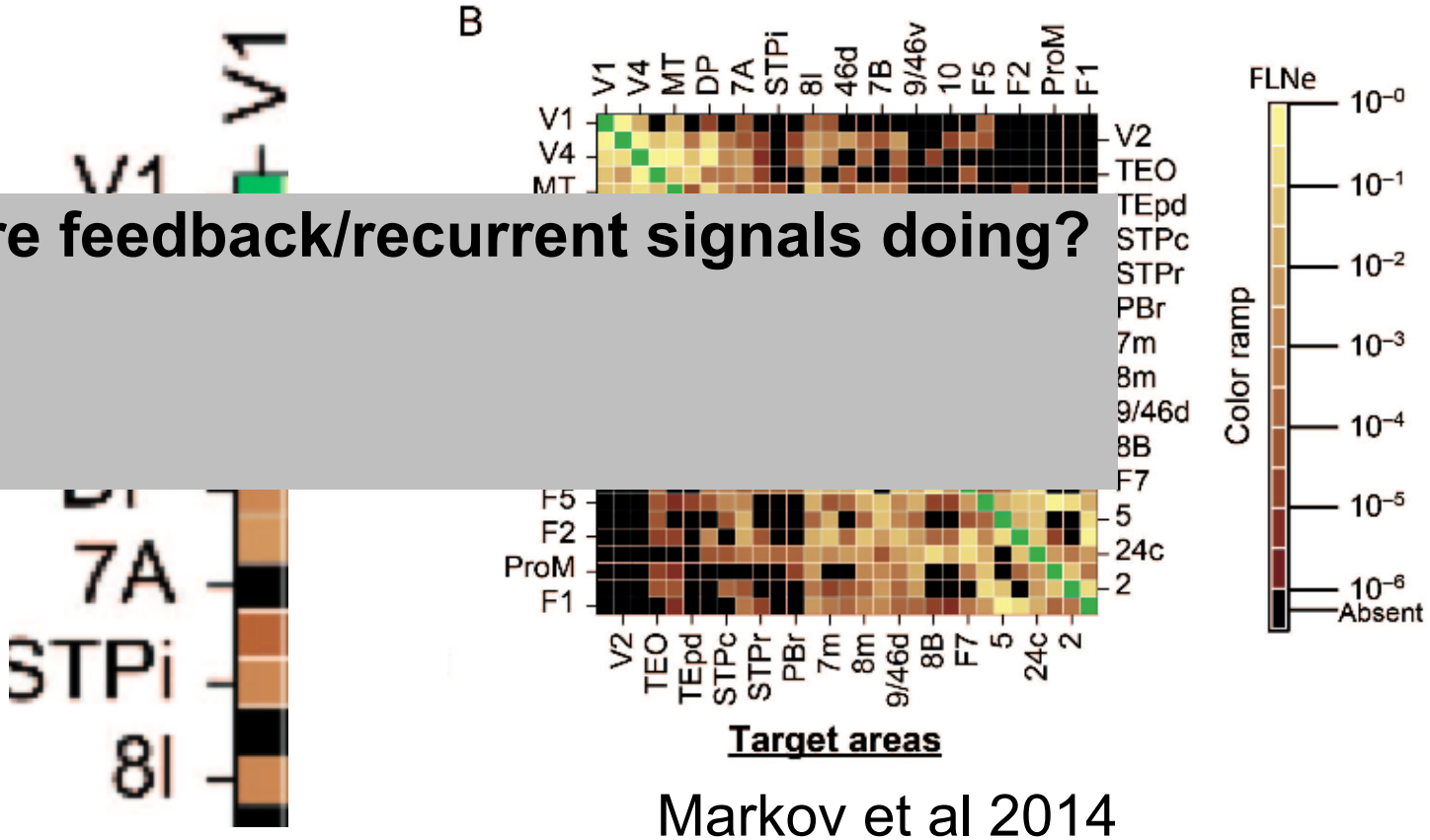
`TaskTermination`

`TaskReport`

Why are there so many feedback and recurrent connections?

There are more horizontal + top-down projections than bottom-up ones (e.g. Douglas 2004, Callaway 2004)

What are feedback/recurrent signals doing?
When?
Why?
How?



Markov et al 2014

Visual cognition: a sequence of routines

Divide et impera

Operations

- Candidate labels for foveated region
- **Inference and pattern completion**
- Candidate representation of the periphery
- Select target for active sampling (eye movements)
- Determine spatial relations
- Temporal comparisons
- Store information
- Retrieve previously stored information
- Make temporal predictions

Visual cognition: a sequence of routines

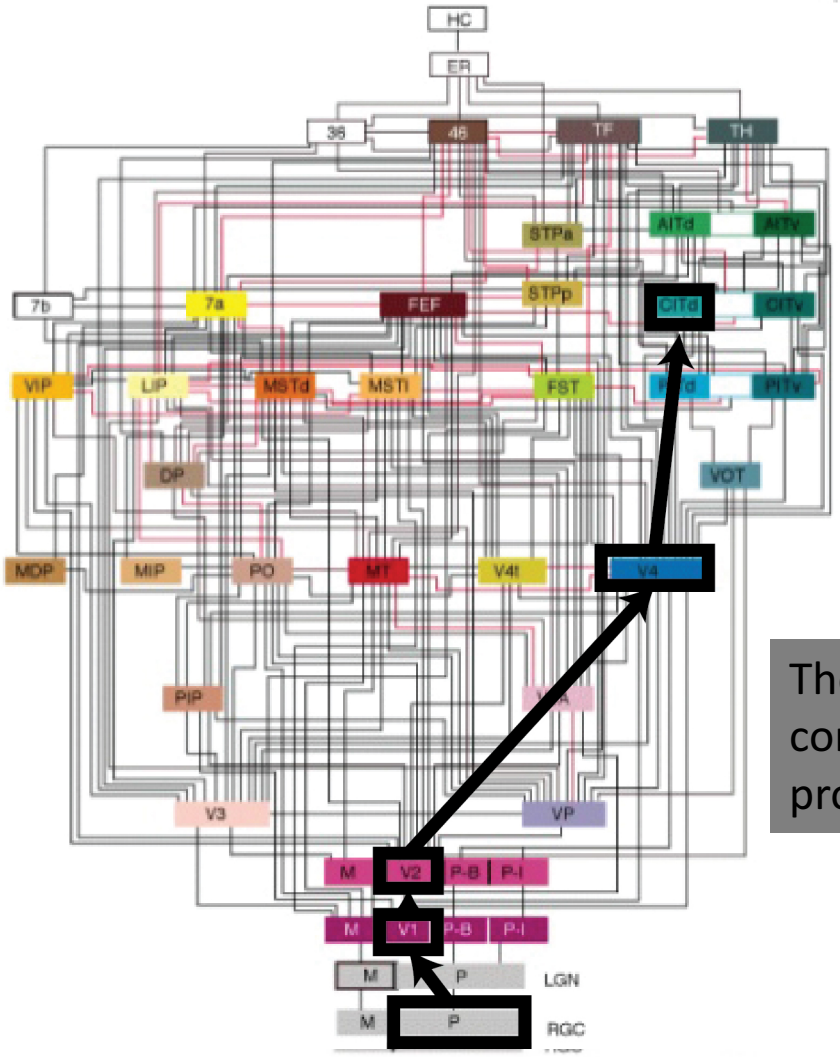
What are they doing?

1. Extract initial sensory map → Call `initial sampling`
2. Propose image gist → Call `rapid peripheral assessment`
3. Propose foveal objects → Call `foveal recognition`
4. Inference from 1+2+3 → Call `pattern completion`
5. Temporary information storage → `Call visual buffer`
6. Task-dependent sampling → Call `target eye movement proposal`
7. Active sampling → Call `eye movement implementation`
8. Detect people → Call `people detection`
9. Determine basic spatial relationships → Call `spatial relationships`
10. Repeat steps 3+4+5
11. Repeat steps 6-7
12. Repeat 8-9
13. Got answer → Call `task termination evaluation`
14. If satisfactory, answer the question → Call `task report`

There is more. Much more.

A schematic diagram of visual cortex connections in macaque monkeys

State-of-the-art computer vision architectures mimic a small fraction of visual cortex



There are recurrent connections at every processing stage

Felleman and Van Essen 1991

Pattern completion is a hallmark of intelligence

A, C, E, G, → **I**

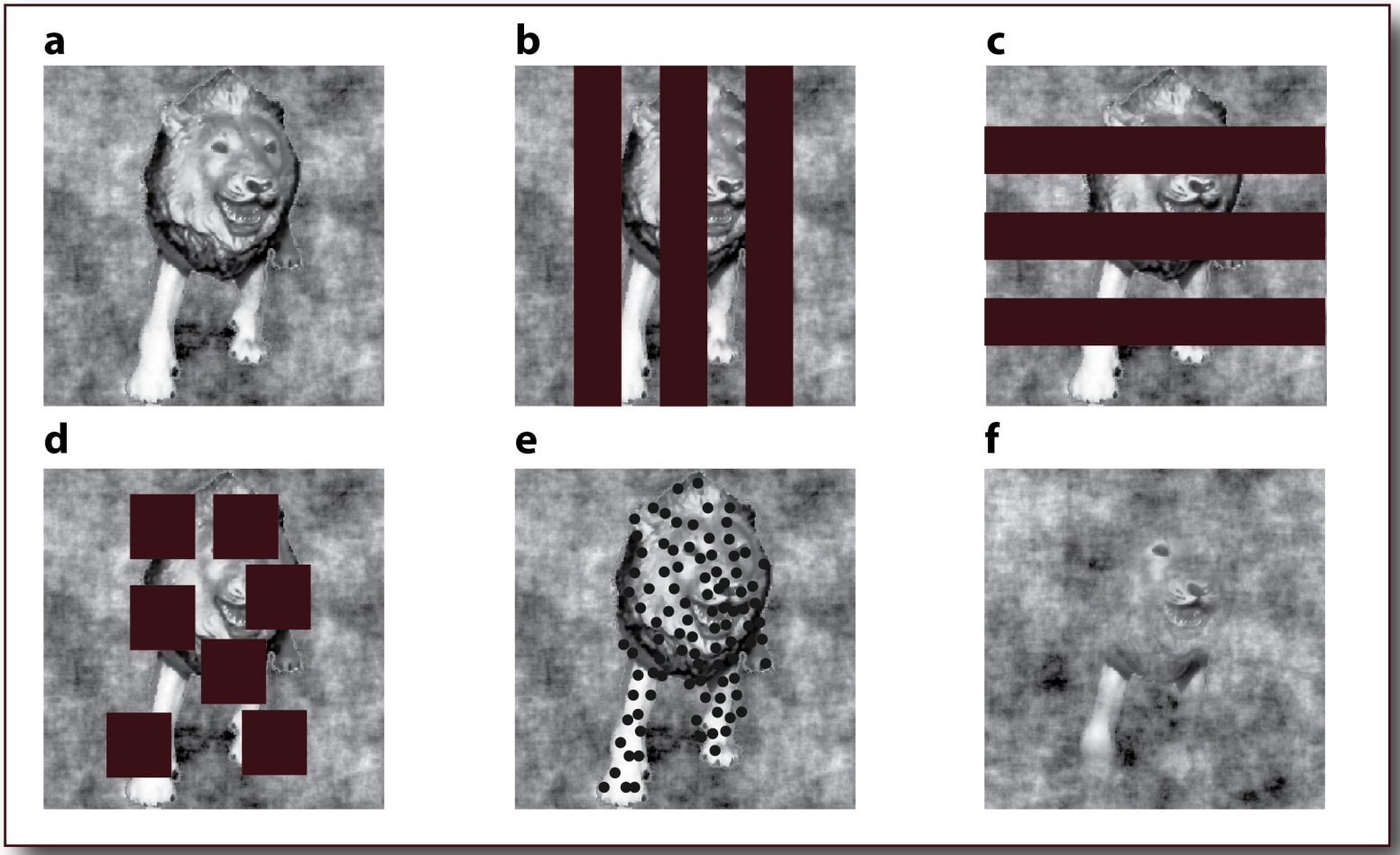
1, 2, 3, 5, 7, 11, → **13**

V-s-a- R-c-g-i-i-n → **Visual Recognition**

Even though it was raining heavily, Jonathan decided to go out without an … → **Umbrella**

Also:
Other sensory modalities
Music
Reading a story
Social interactions

Objects can be recognized from partial information

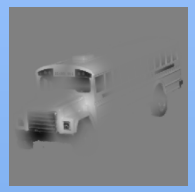


4 bubbles

Evaluating pattern completion

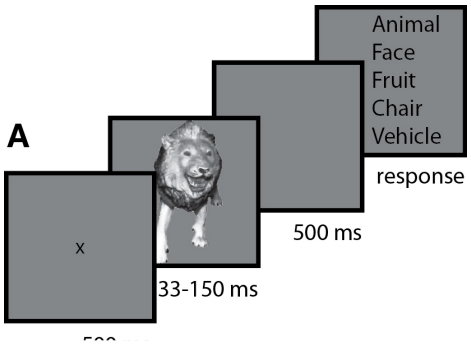
20 bubbles

10 bubbles



6 bubbles

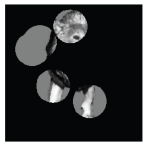
4 bubbles



C

Whole Partial

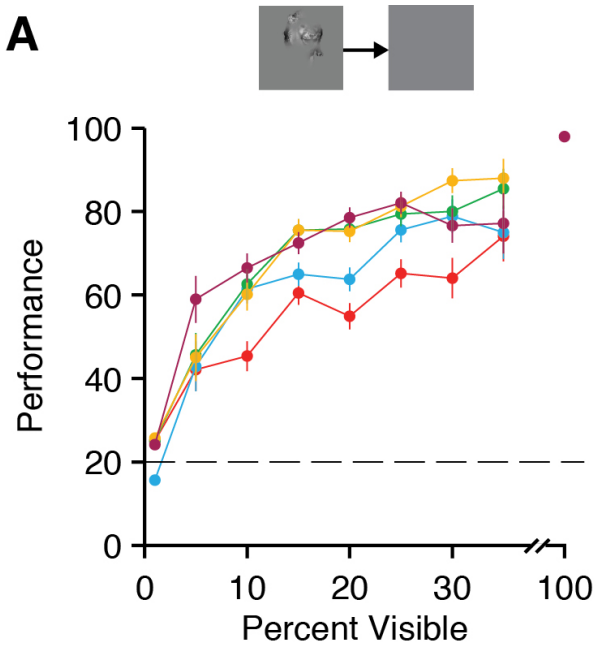
D



Occluded

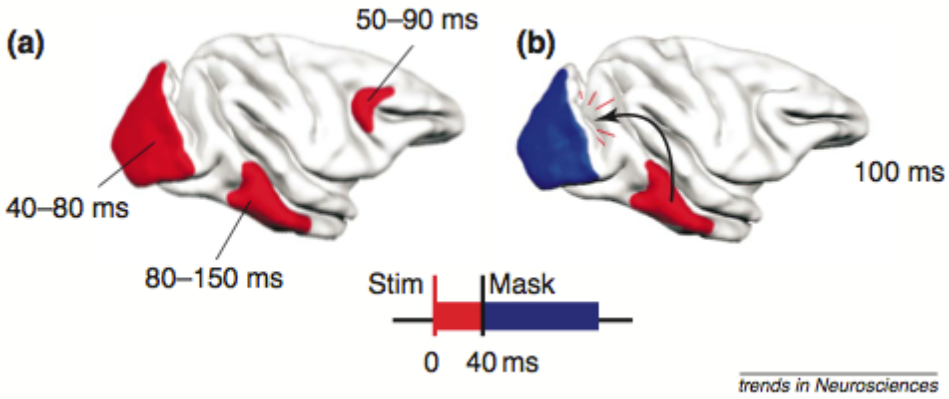
Hanlin Tang, Bill Lotter, Martin Schrimpf

Strong robustness to limited visibility



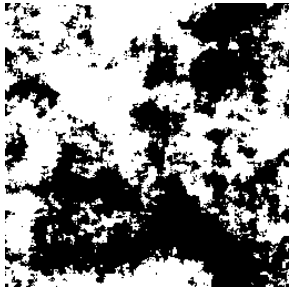
Backward masking interrupts processing (presumably of feedback/recurrent computations)

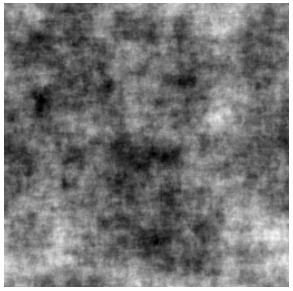
Models:



Lamme V, Roelfsema P (2000)

Masks:





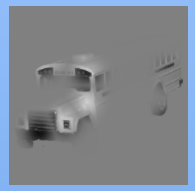
- Short delays (SOA<20ms): mask reduces visibility
- Longer delays: mask is purported to disrupt recurrent/top-down processing

V1: Bridgeman 1980, Maknik and Livinsgtone 1998, Lamme et al 2002
IT: Kovacs et al 1995, Rolls et al 1999

Evaluating pattern completion abilities

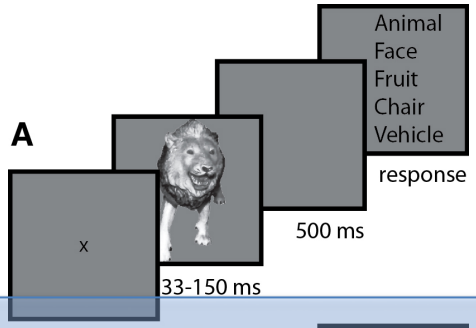
20 bubbles

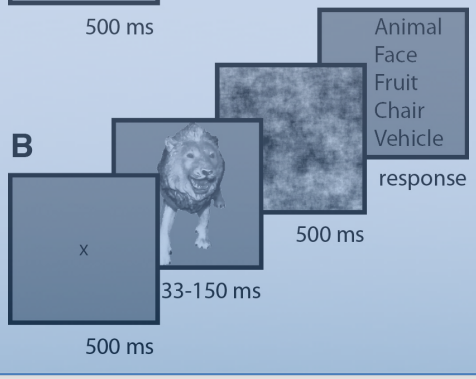
10 bubbles



6 bubbles

4 bubbles





C

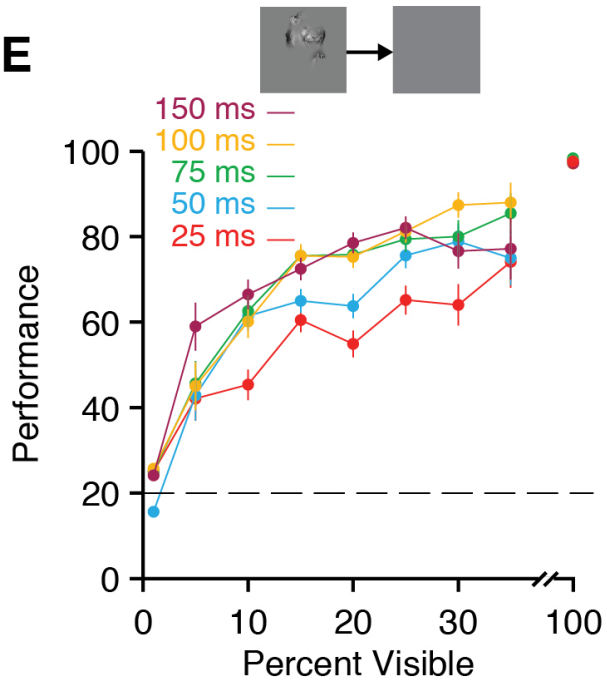
Whole

Partial

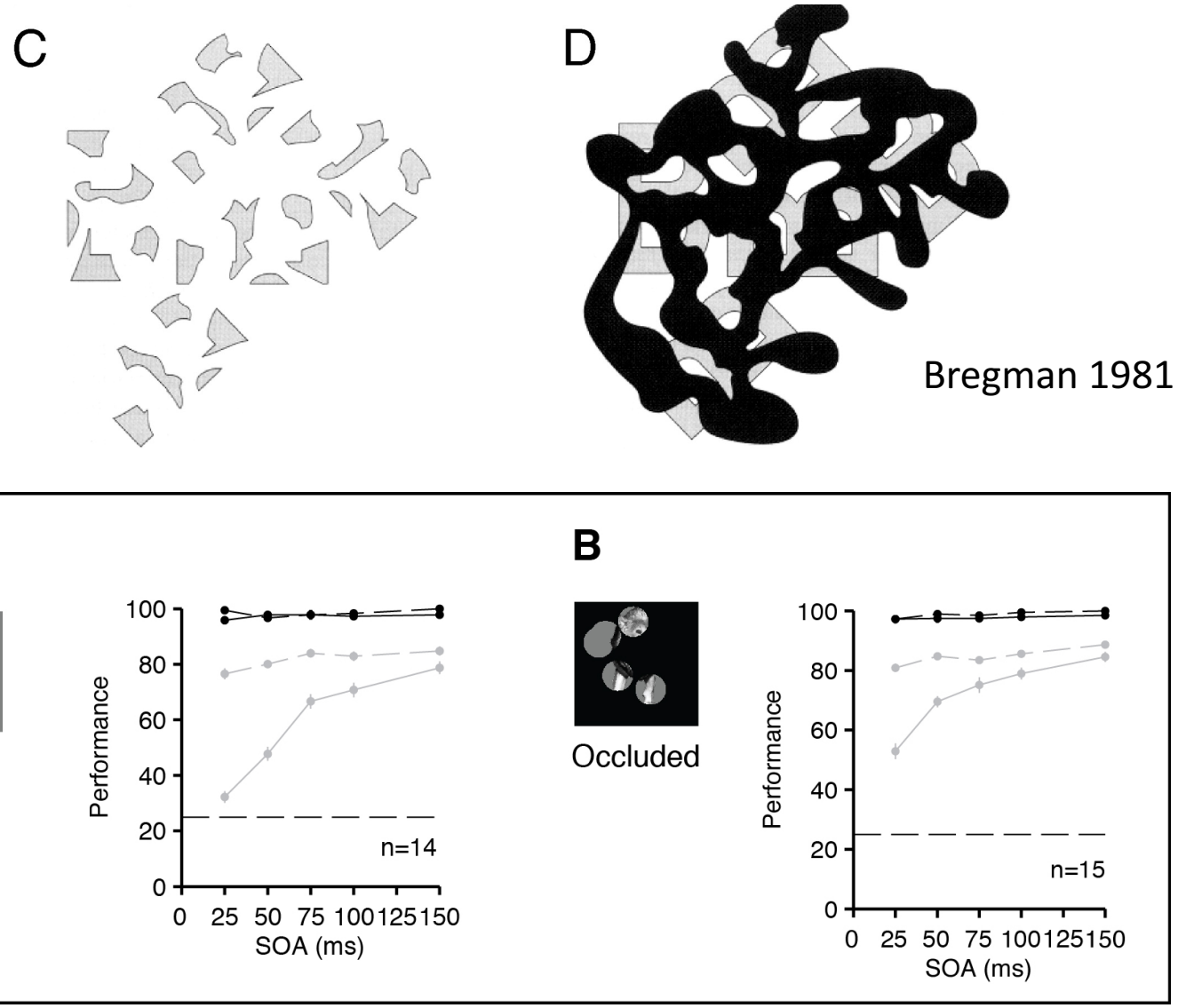
D

Occluded

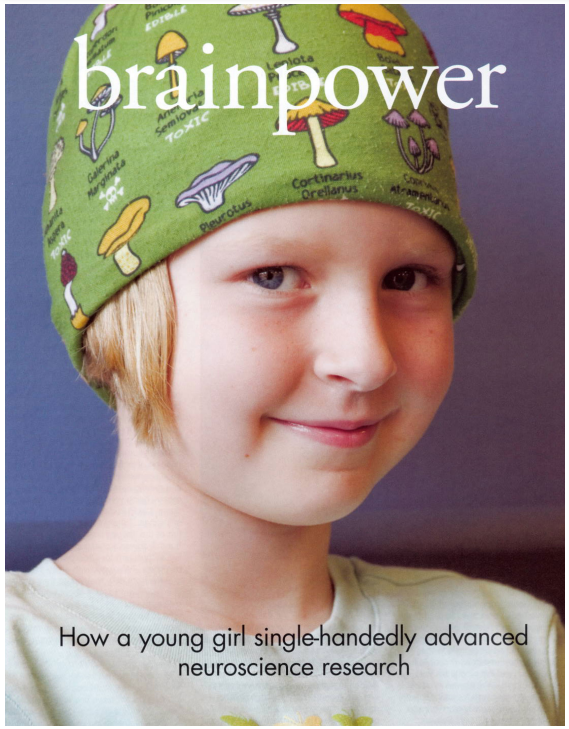
Backward masking disrupts pattern completion

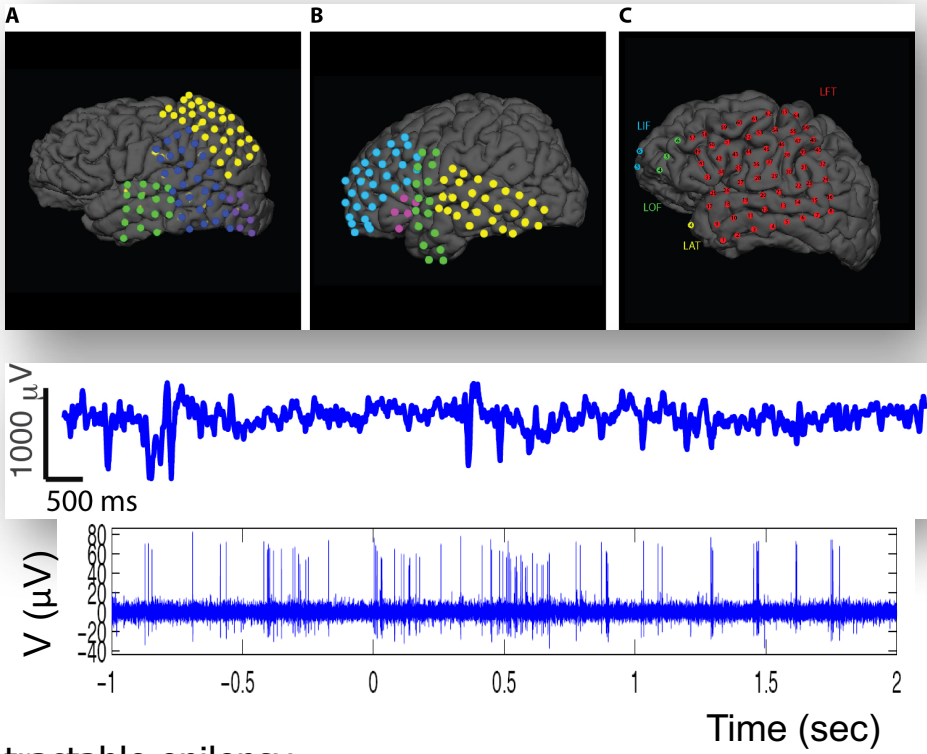


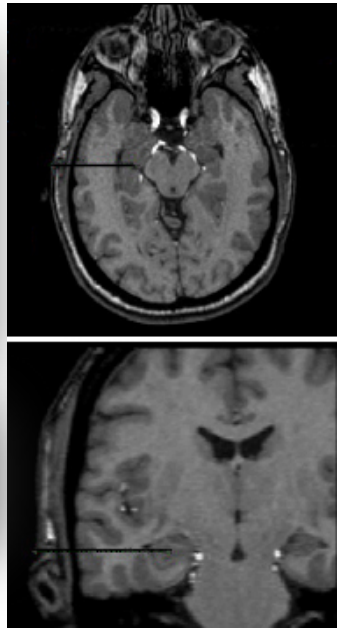
Backward masking also disrupts recognition of occluded objects



Peeking inside the human brain



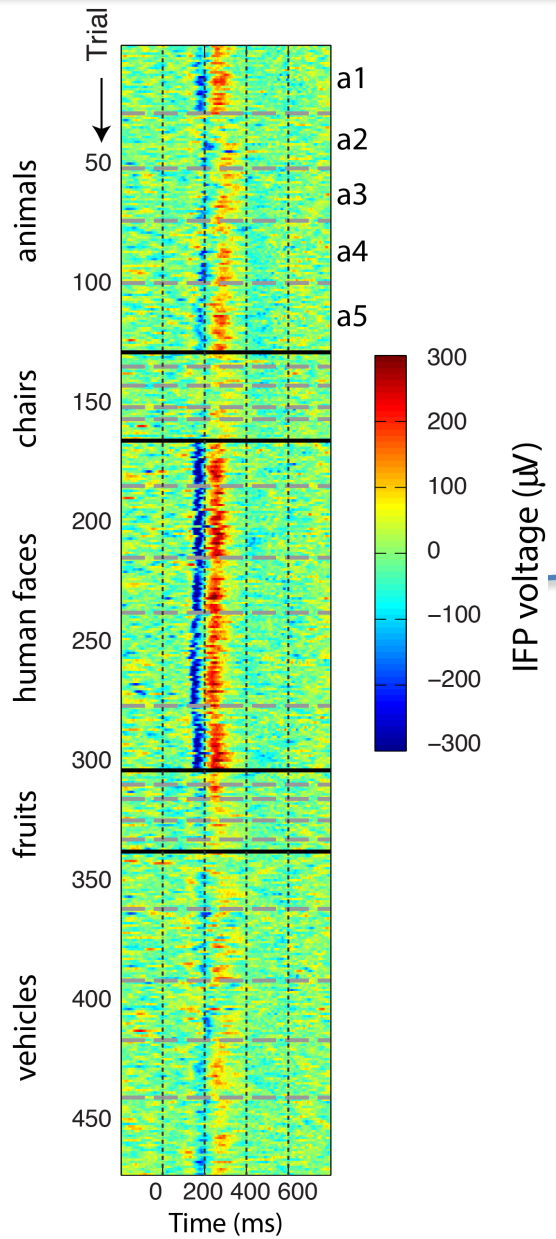


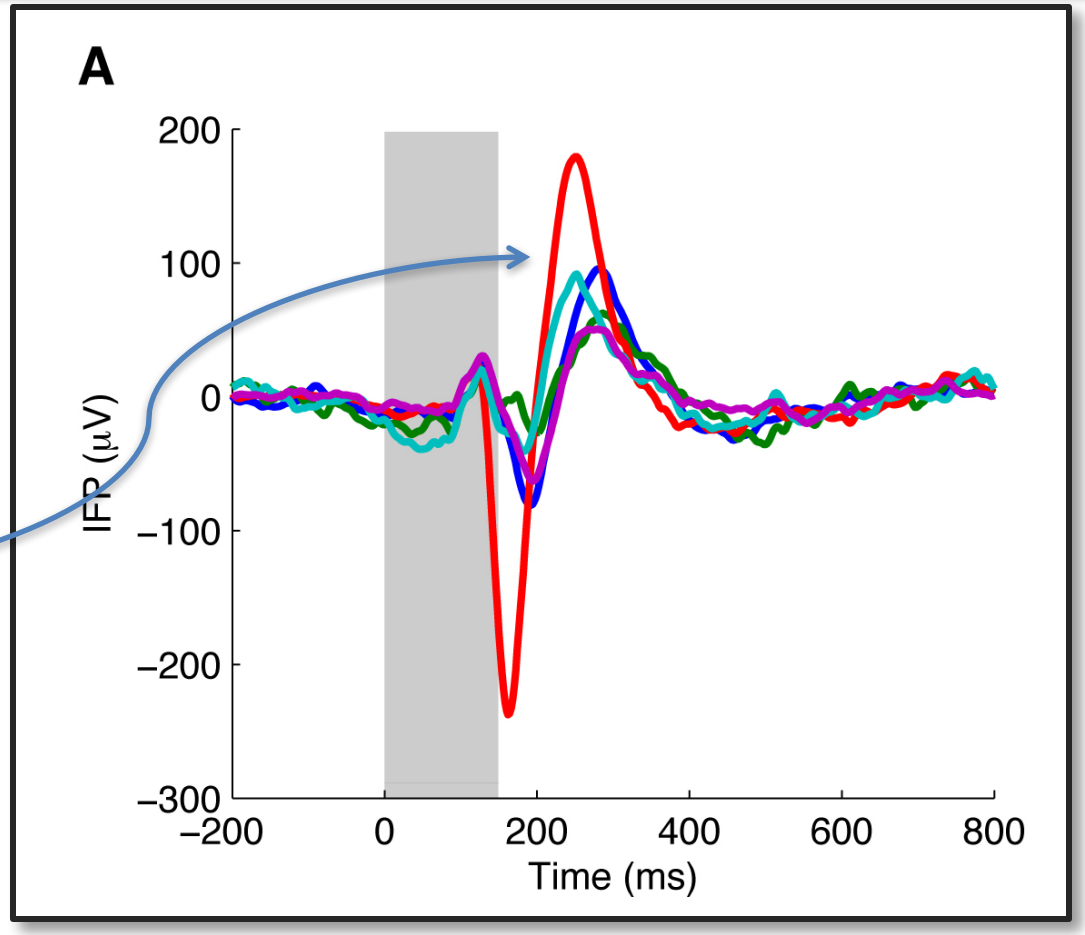


- Patients with pharmacologically intractable epilepsy
- Multiple electrodes implanted to localize seizure focus
- Patients stay in the hospital for about 7-10 days
- All experiments are approved by the Institutional Review Boards
- All testing is performed with the subjects' consent

Neurosurgeons: **William Anderson, Joseph Madsen, Itzhak Fried**

Reliable, selective and rapid responses in human inferior temporal cortex





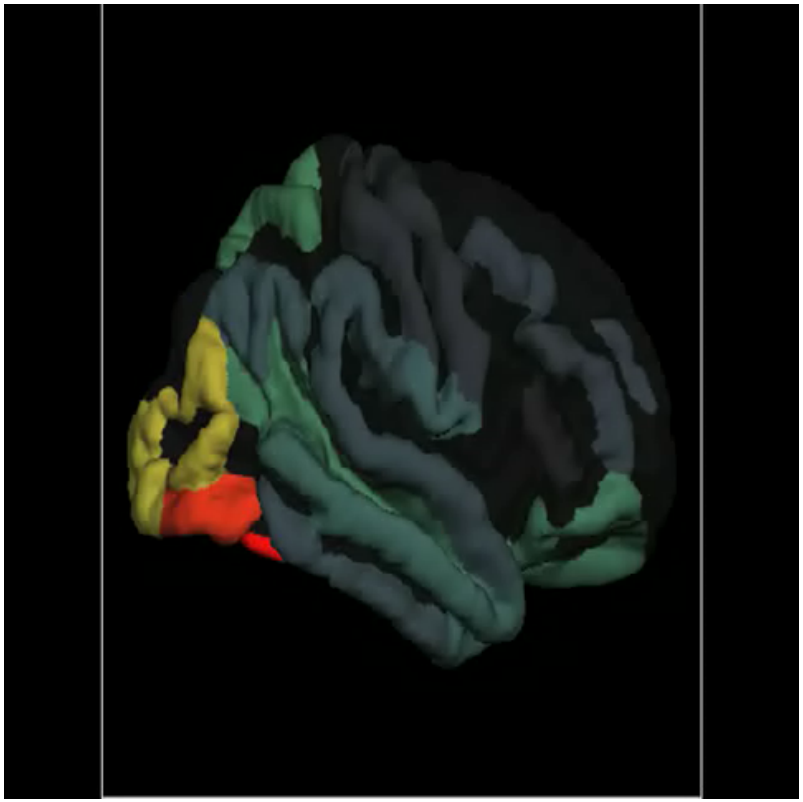
Inferior temporal gyrus

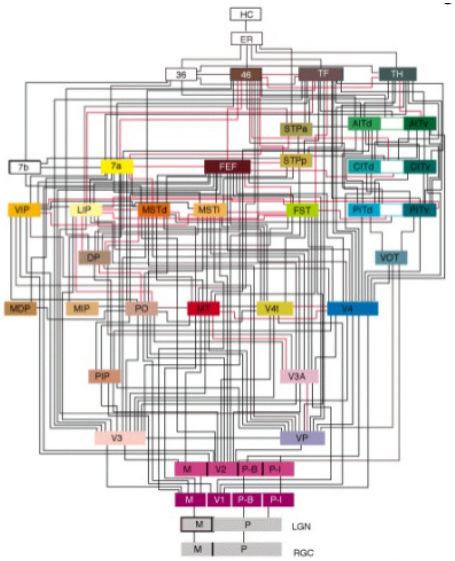
Tang et al 2014

Visual selectivity along the human ventral visual cortex

2205 electrodes
27 subjects

Main areas showing visual selectivity
Inferior-occipital gyrus
Fusiform gyrus
Medial temporal gyrus
Inferior temporal gyrus
Temporal pole

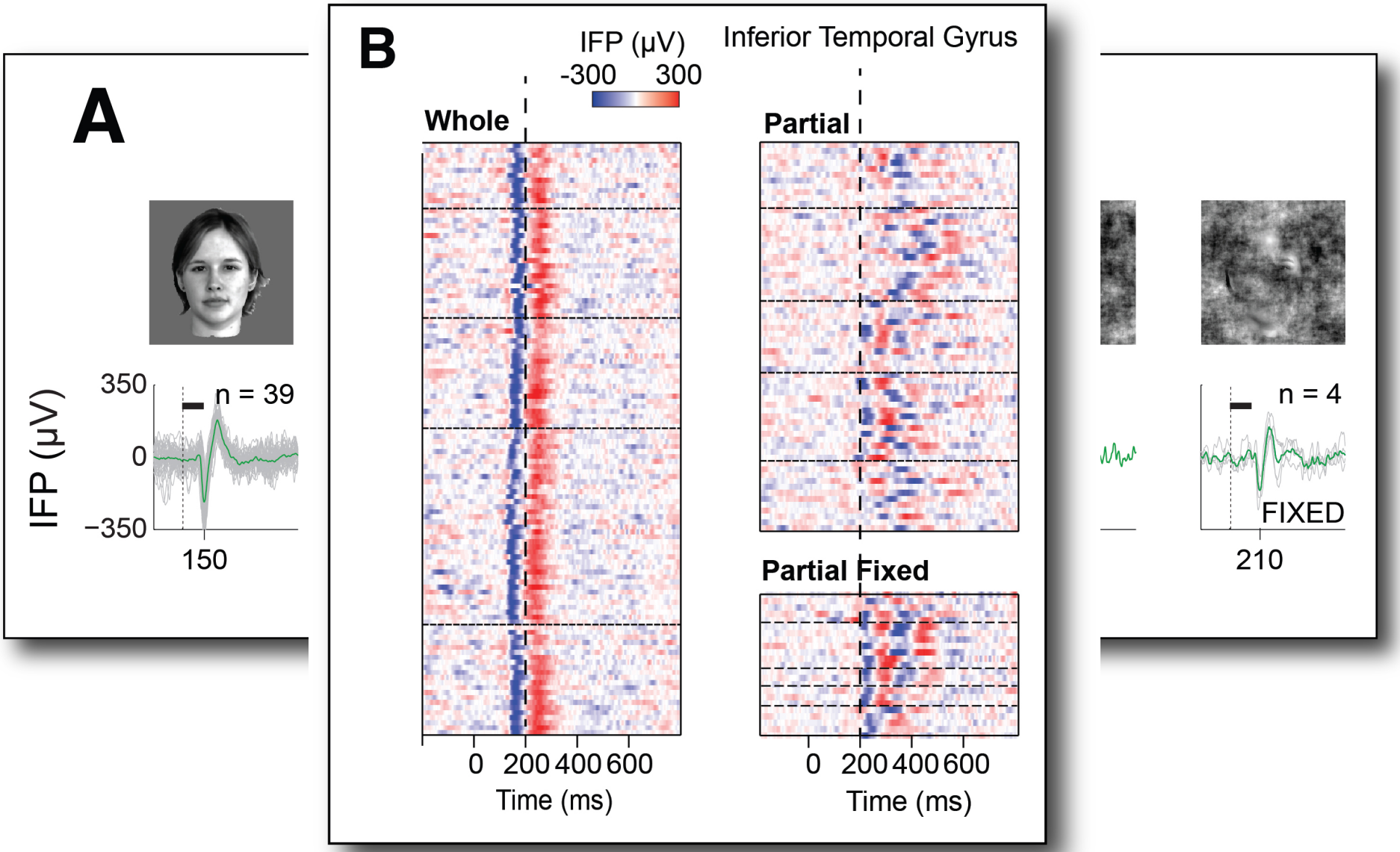




Agam, Liu, Singer, Tang

Gross, Desimone, Logothetis, Richmond, Tanaka, DiCarlo, Rolls, Connors, Ito, Vogels

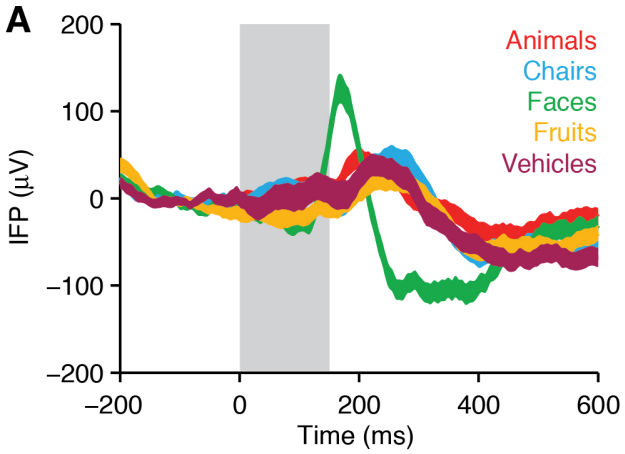
Example responses during object completion



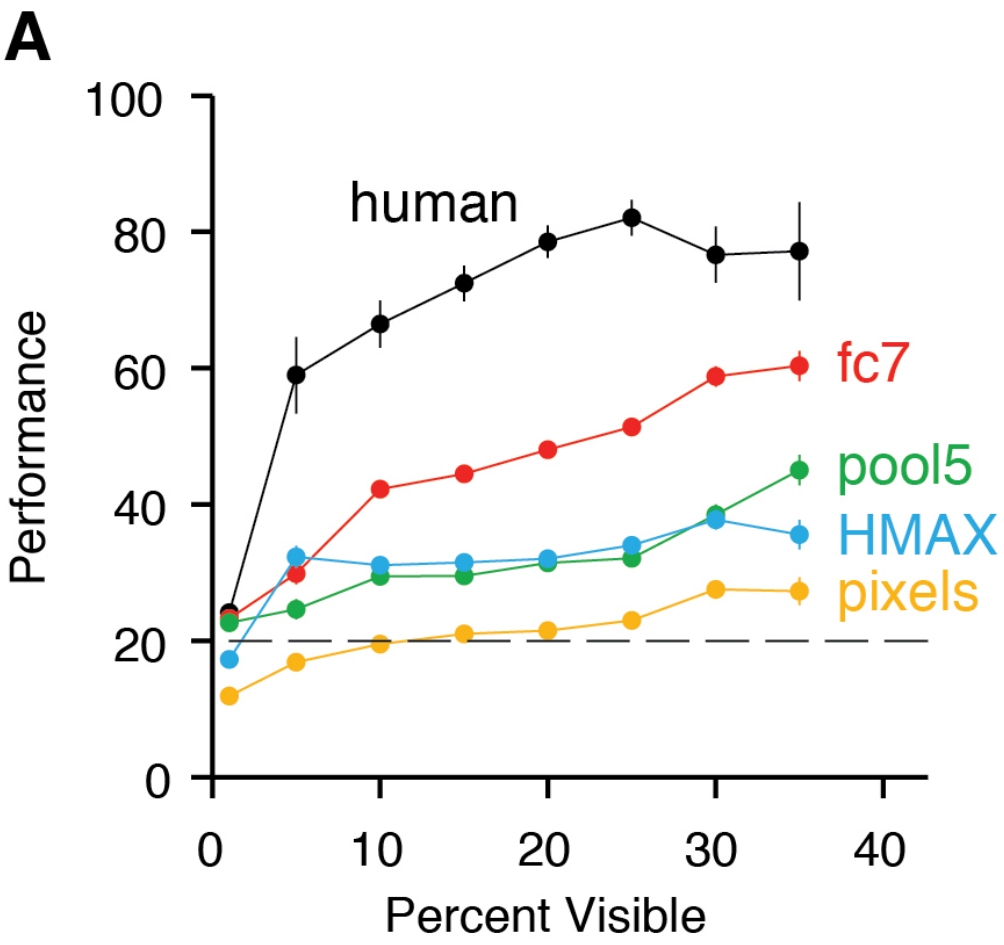
Inferior Temporal Gyrus

Tang et al, 2014

The behavioral effect of masking correlated with the neural response latency on an image-by-image basis

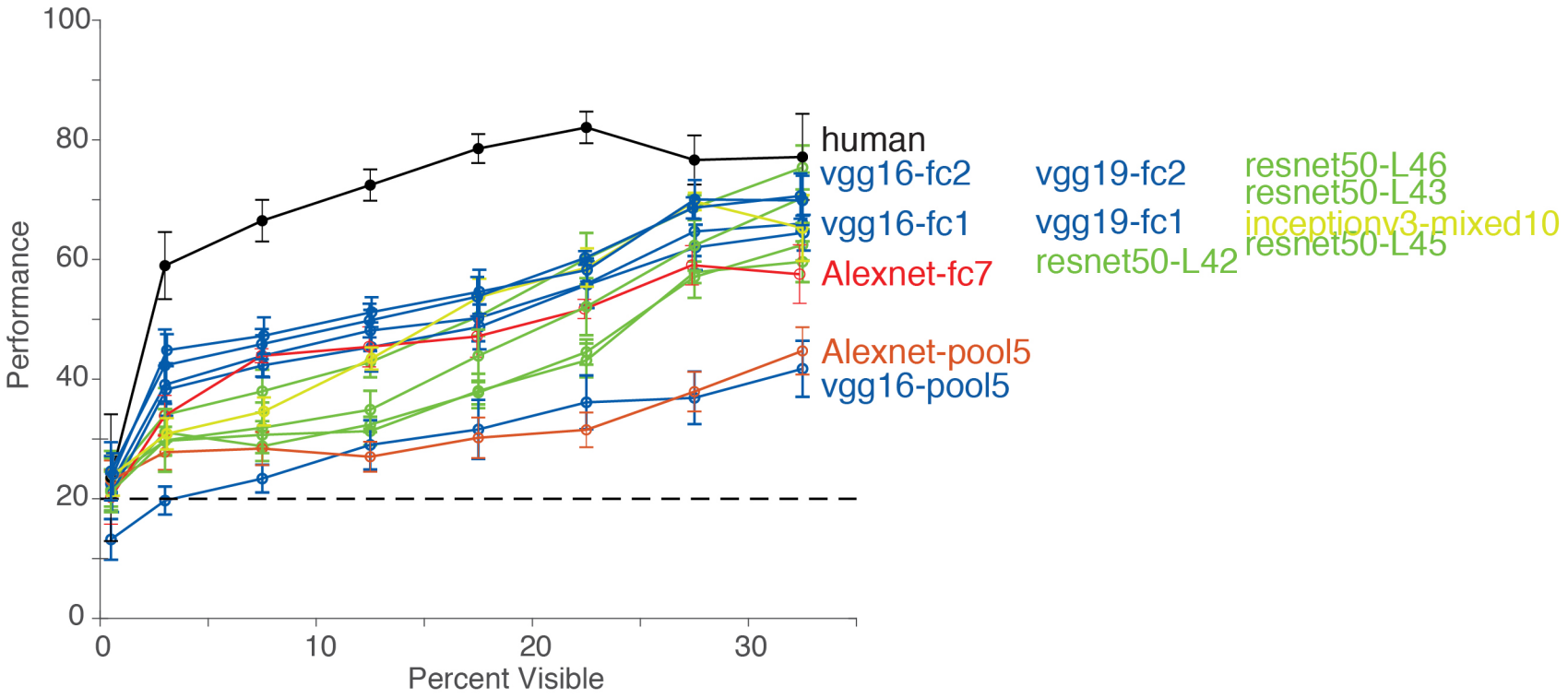


Bottom-up models significantly underperform in recognition of partial images

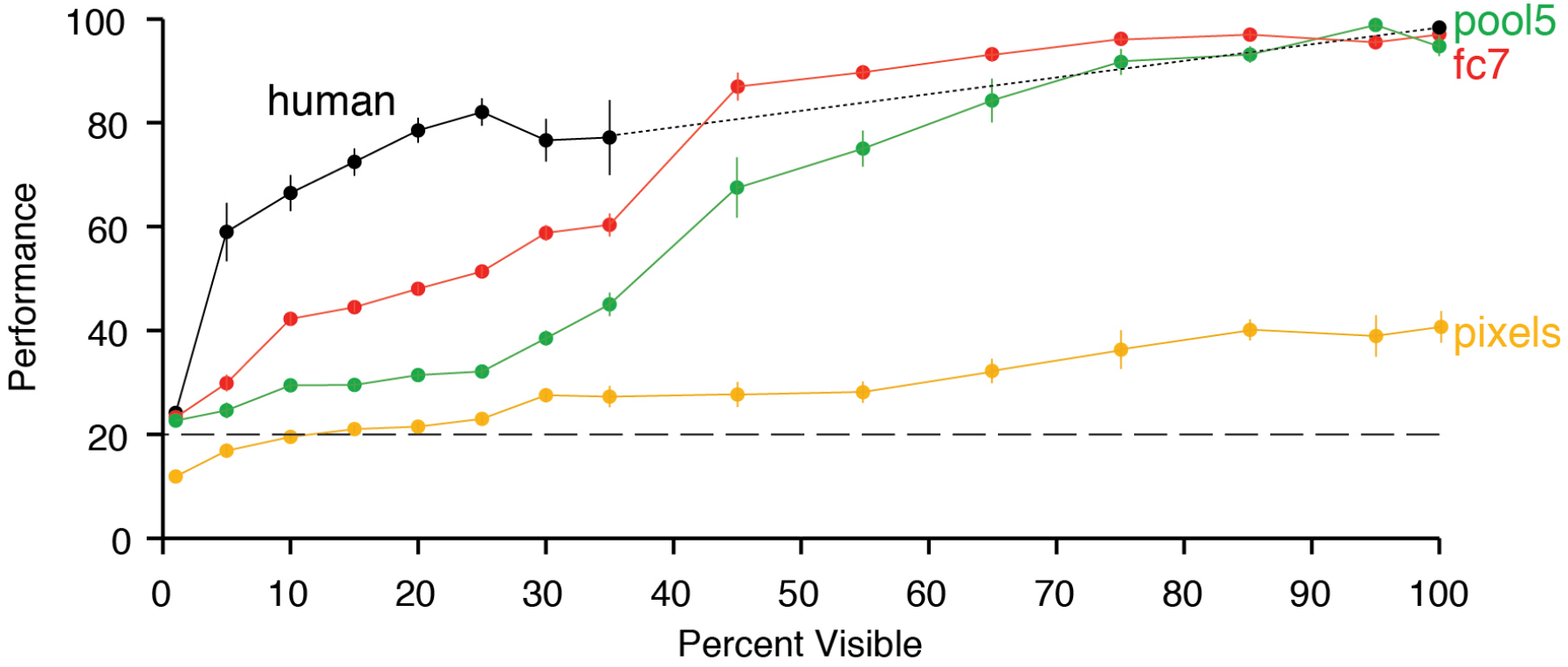


See also Pepik et al 2015, Wyatte et al 2012

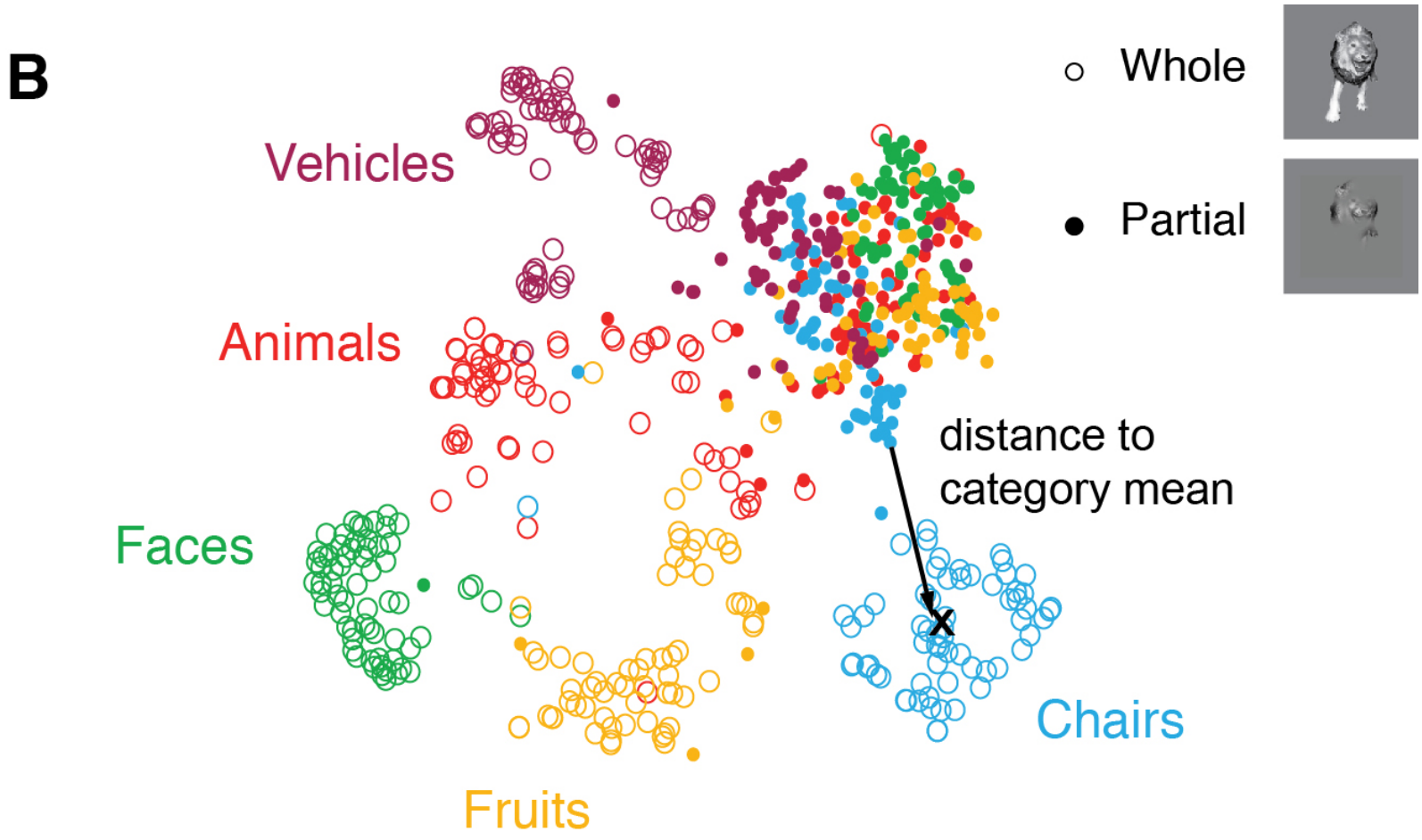
Every feed-forward model that we tested is well below humans in pattern completion



At higher levels of visibility, feed-forward models capture human performance

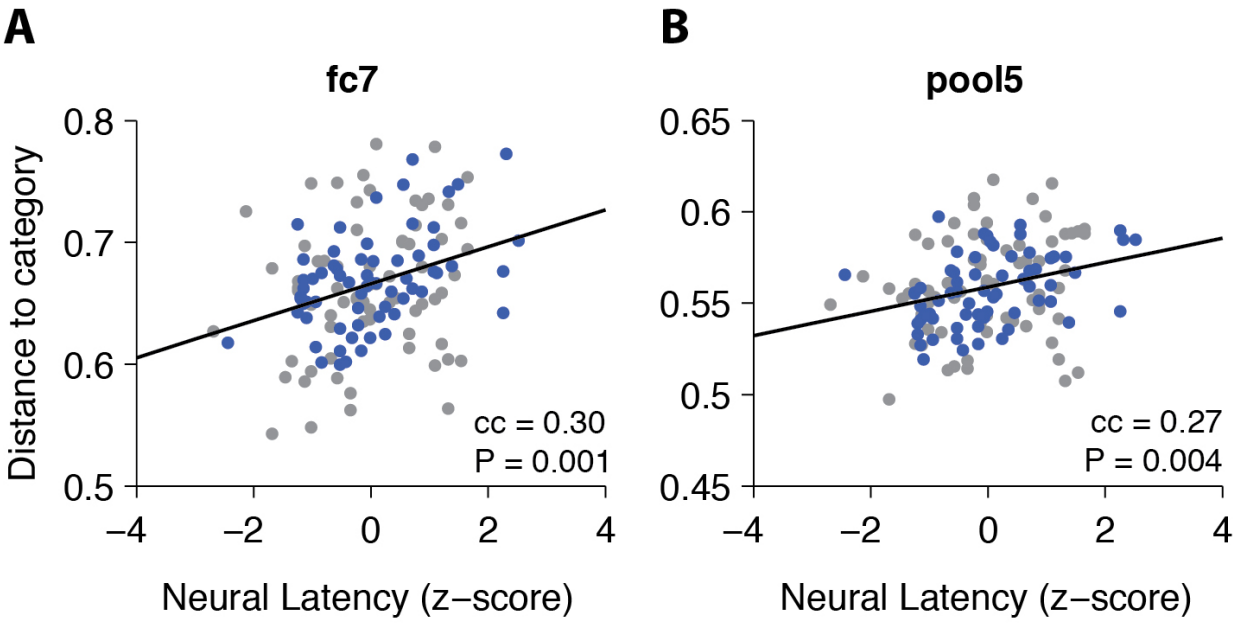


2D object representation at the top of the model hierarchy is not robust to occlusion



Stochastic neighborhood embedding. Van der Maaten 2008

The neural latency for each image was correlated with the distance to category center



Hopfield network with binary neurons

Each neuron i has two states: V_i=0 or V_i=1

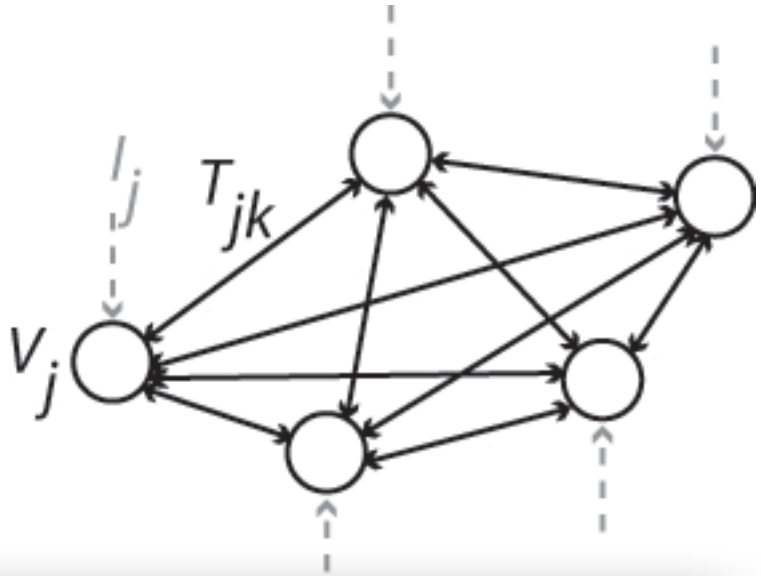
Ensemble: $\mathbf{V} = [V_1, V_2, ..., V_N]$ Note: $\mathbf{V}=\mathbf{V}(t)$

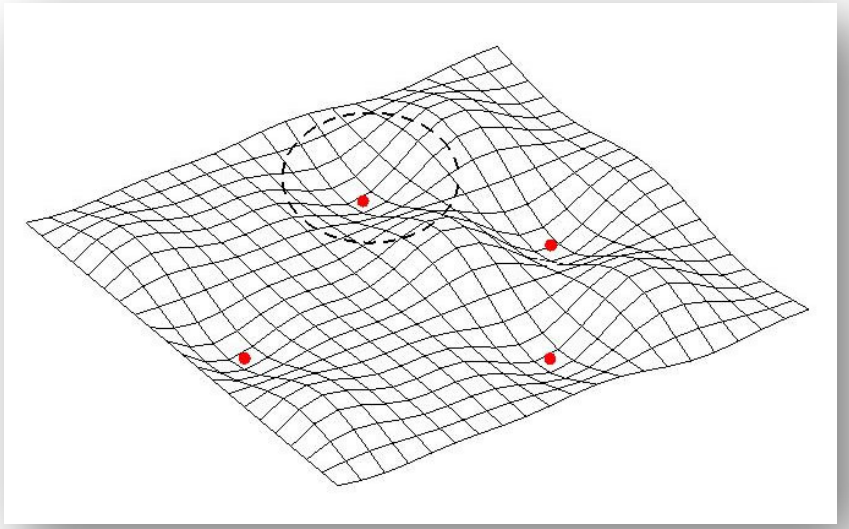
Synaptic strength: T_{ij}

If two neurons are not connected: $T_{ij}=0$

No self connections: $T_{ii}=0$

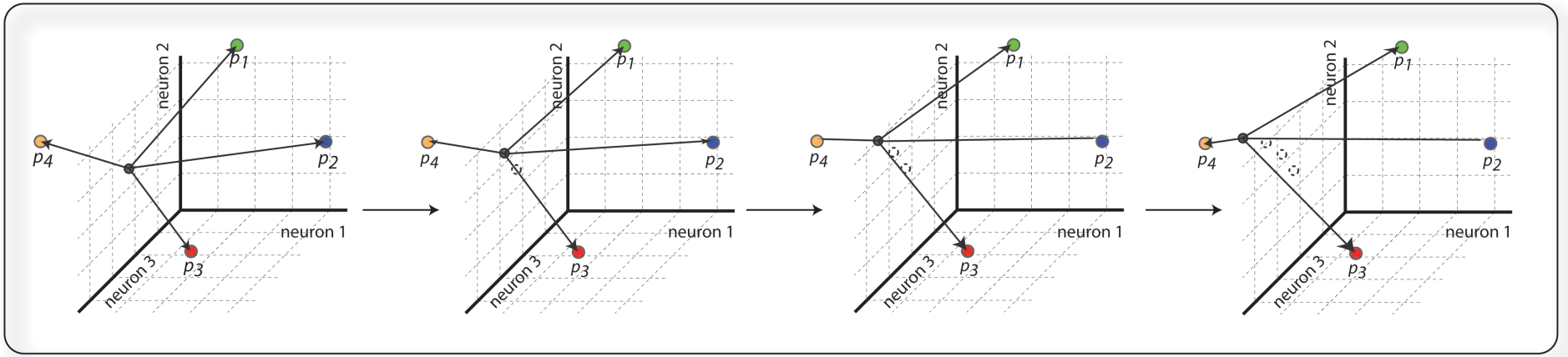
Update rule: $V_i(t)=1$ iff $\sum_j T_{ij} V_j(t) > 0$





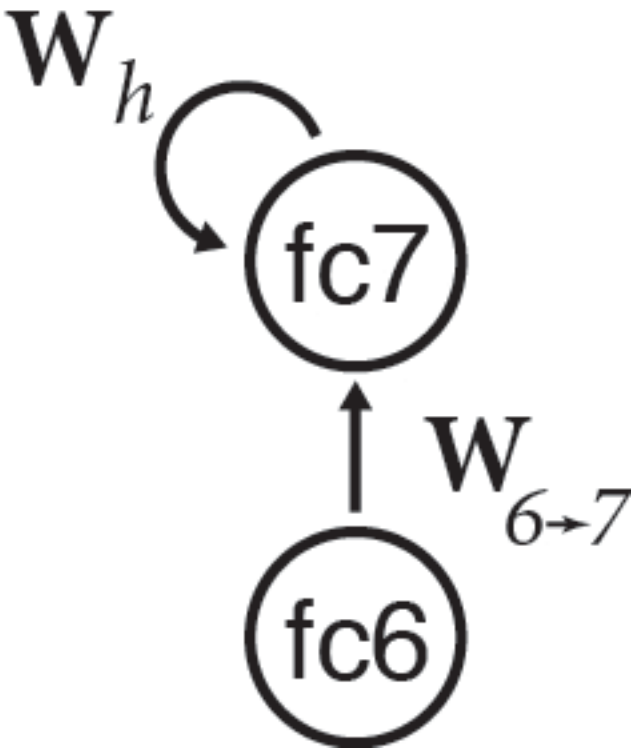
Hopfield, 1982

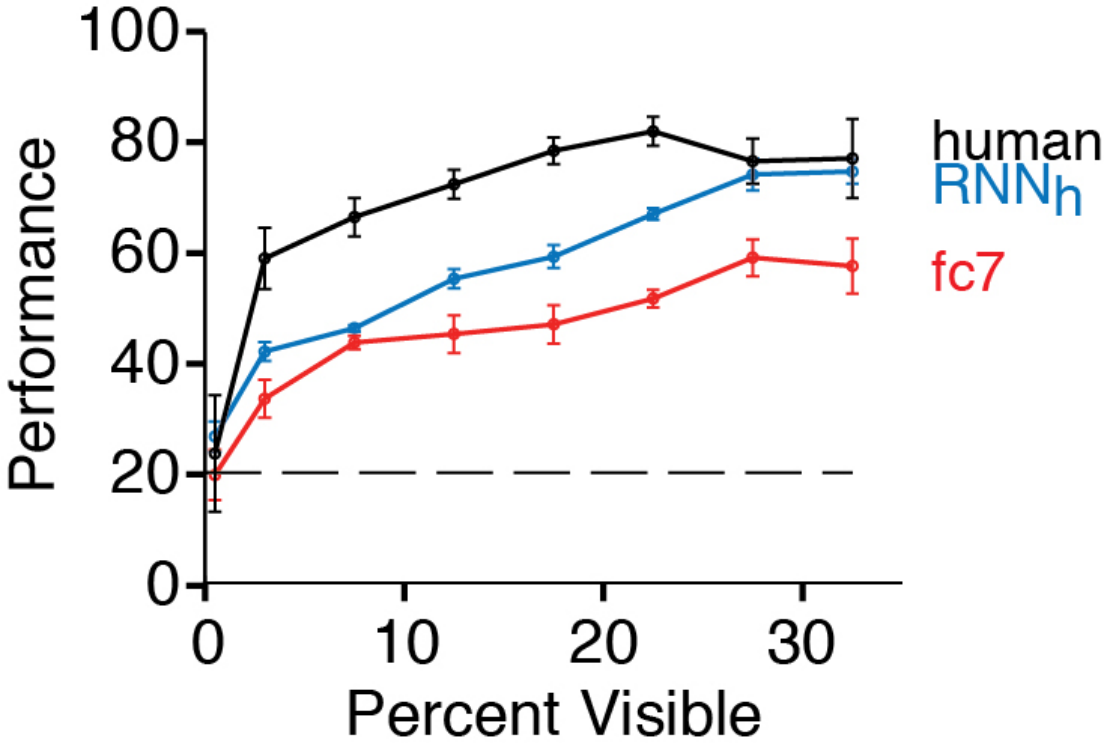
A recurrent network may ameliorate the problem of missing information



$\boldsymbol{p}$ = prototypes (fixed)

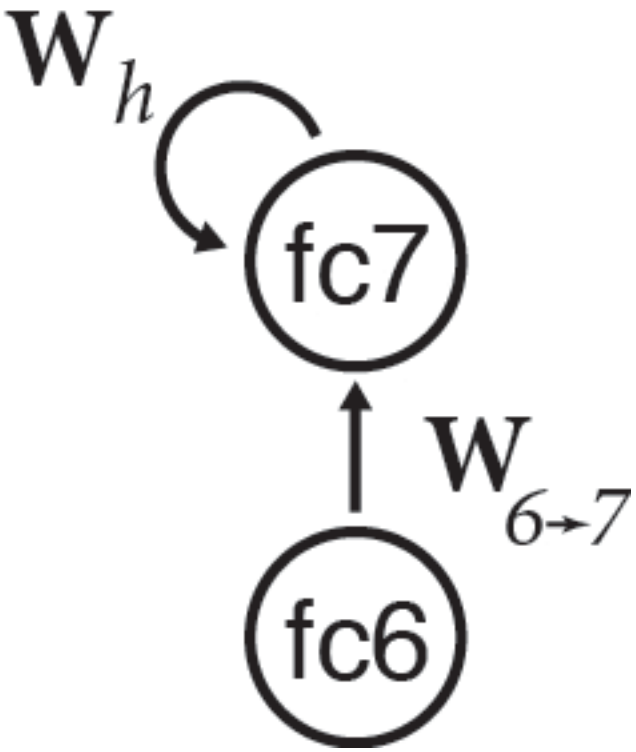
Recurrent Hopfield network (RNN_h) improves recognition performance for partial images

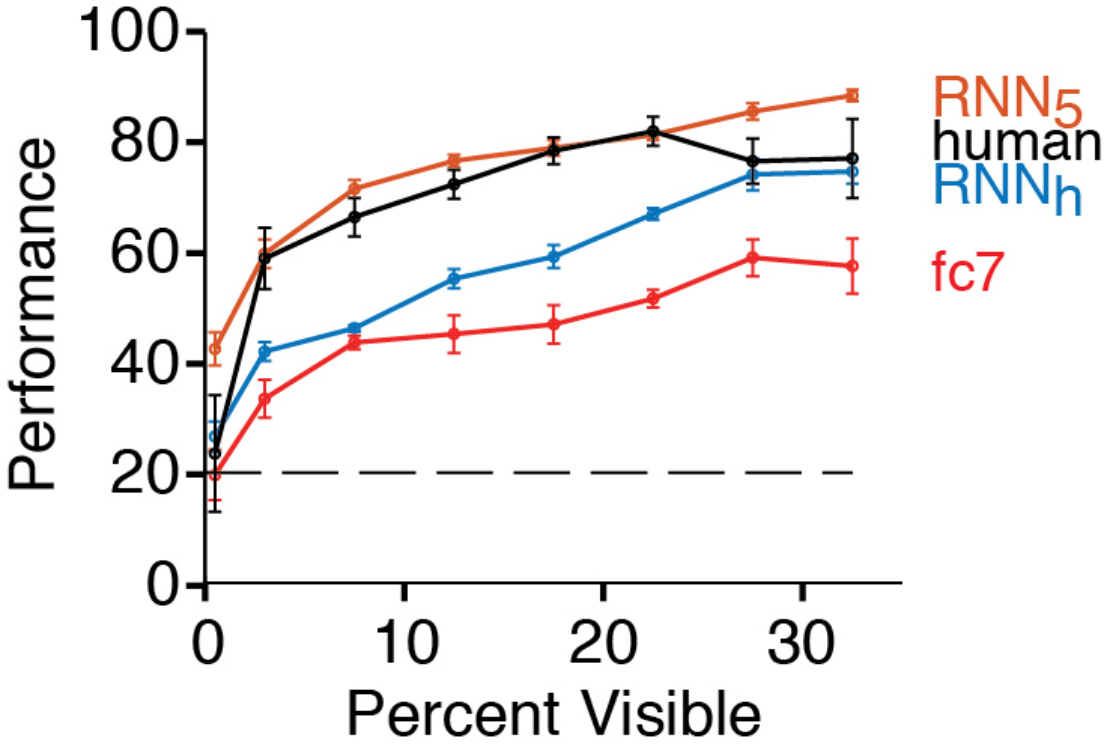




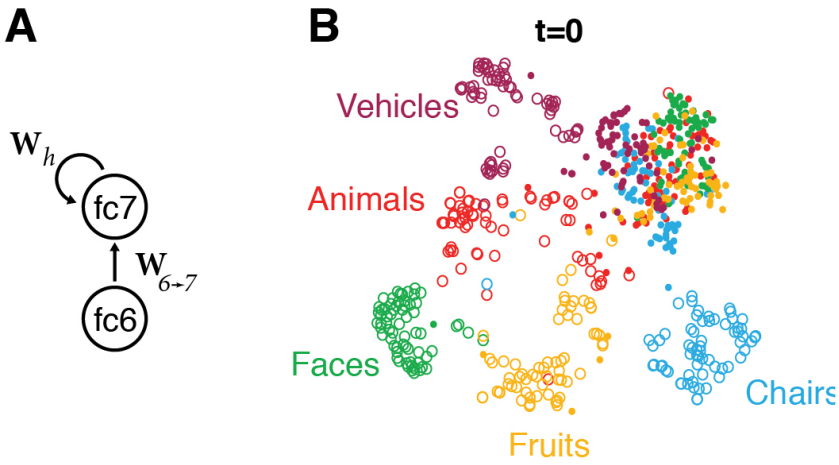
NOTE: 0 free parameters

Training with occluded objects leads to matching human performance in pattern completion

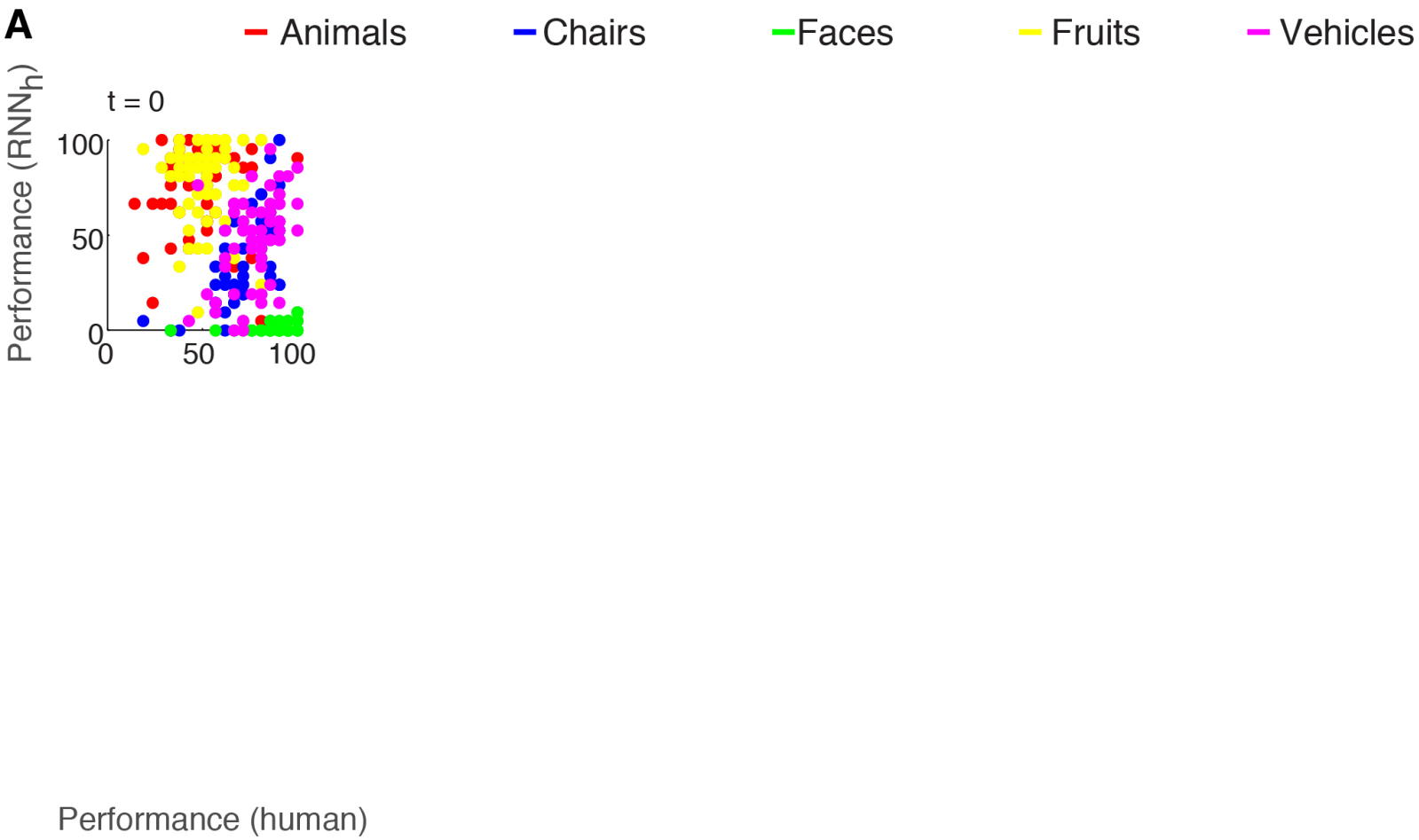




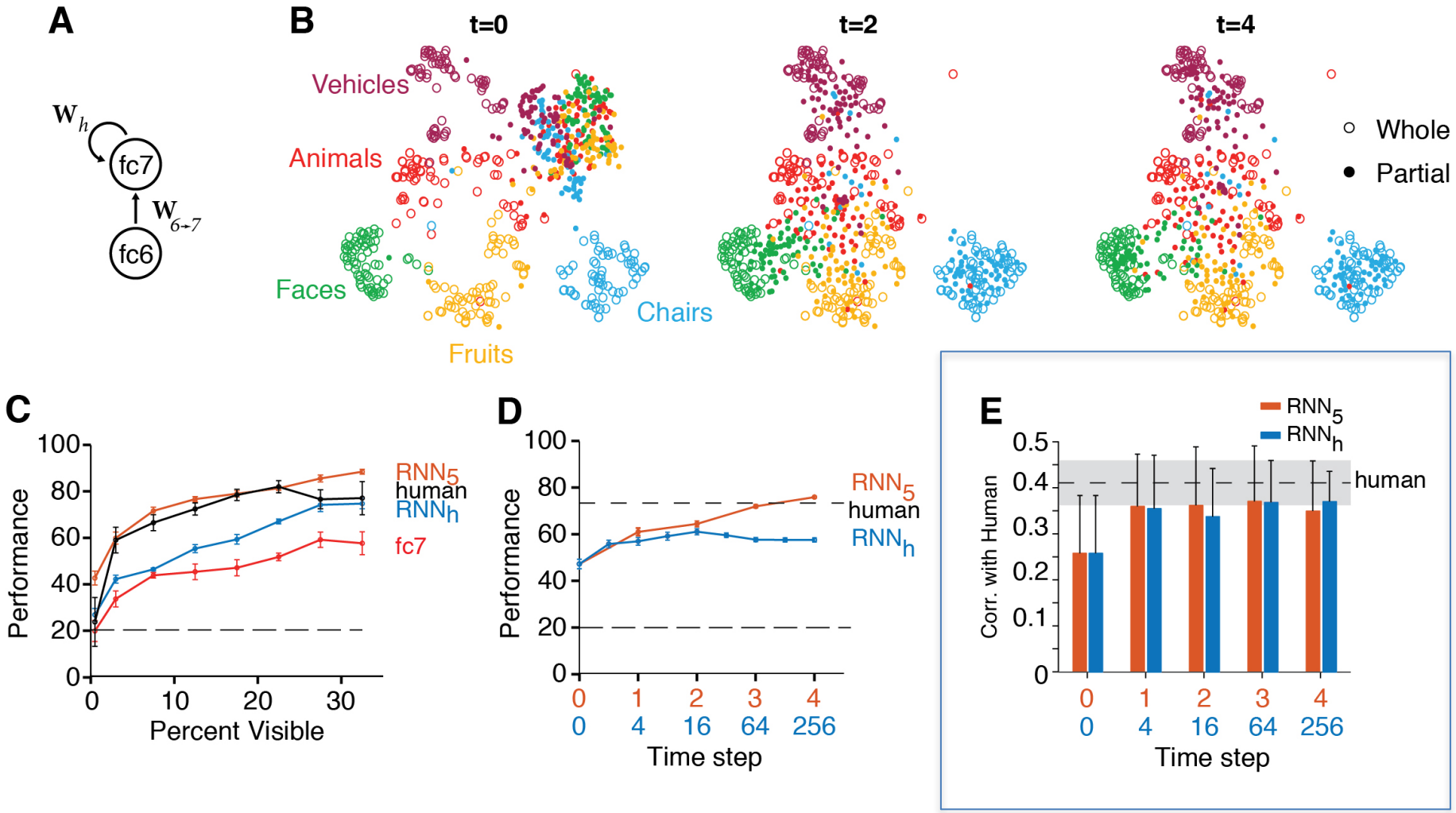
Temporal evolution in recurrent networks



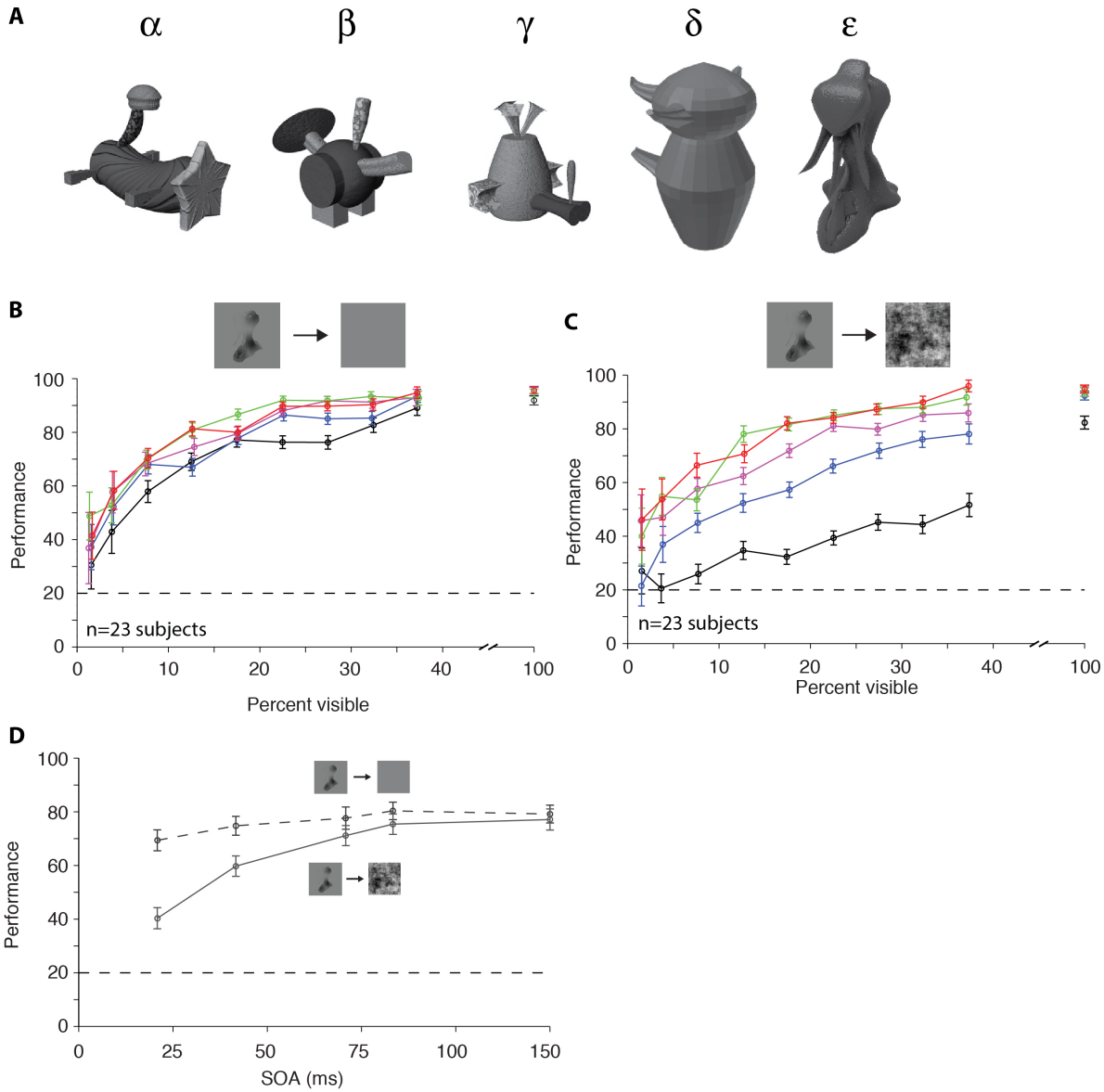
Correlation between RNN models and human performance for individual objects



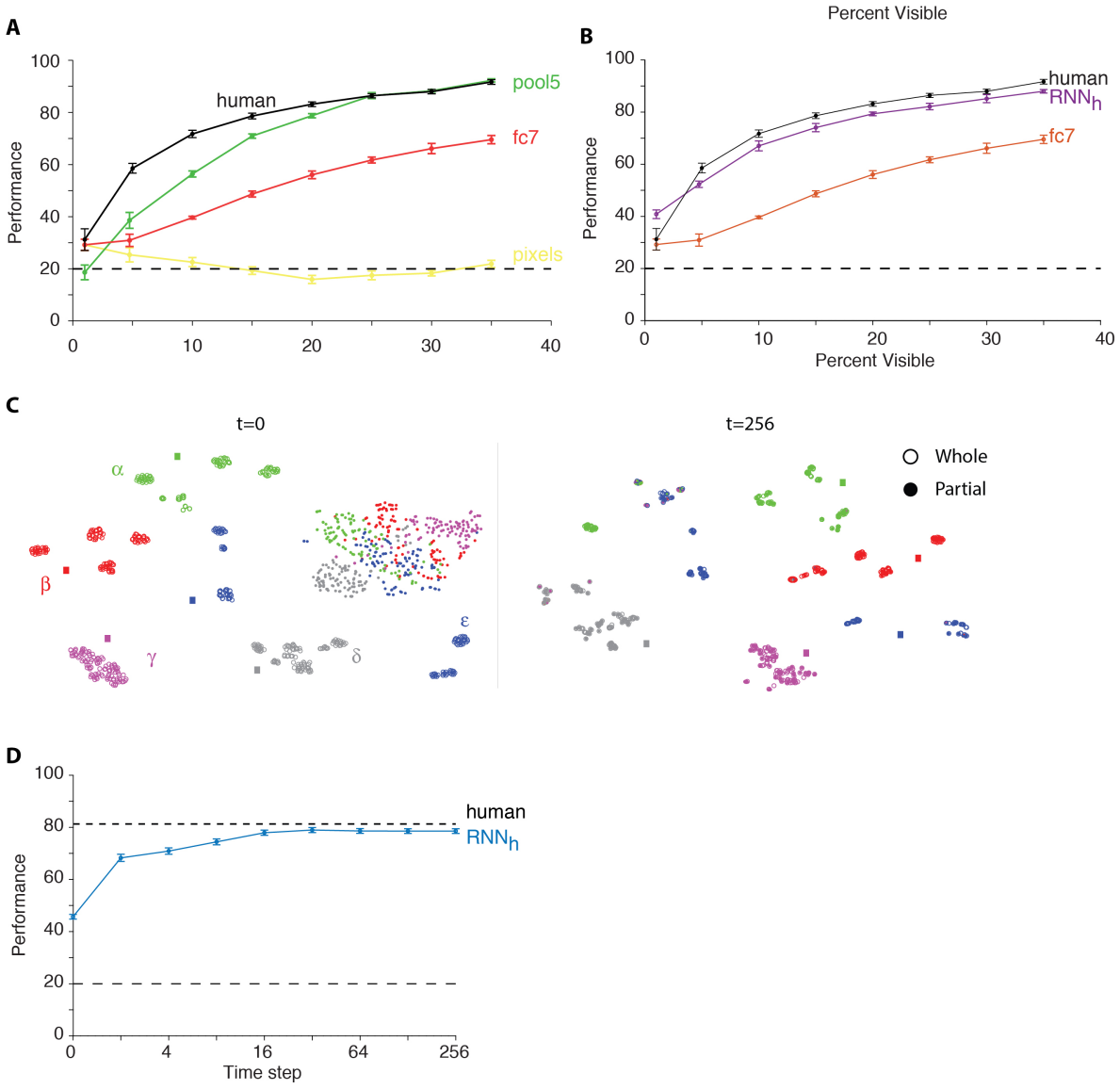
Recurrent neural networks match human performance in pattern completion



Pattern completion with novel objects



Computational model results for the novel objects



Visual cognition: a sequence of routines

Divide et impera

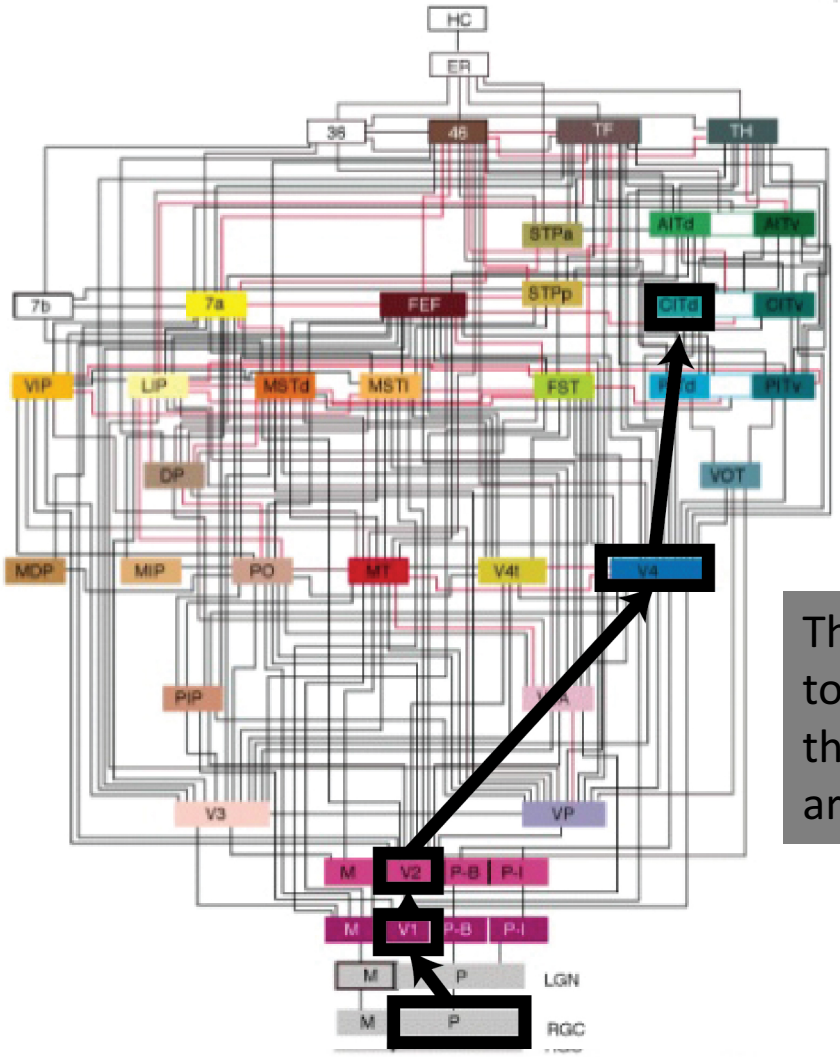
Operations

- Candidate labels for foveated region
- Inference and pattern completion
- Candidate representation of the periphery
- Select target for active sampling (eye movements)
- Determine spatial relations
- Temporal comparisons
- Store information
- Retrieve previously stored information
- **Make spatiotemporal predictions**

There is more. Much more.

A schematic diagram of visual cortex connections in macaque monkeys

State-of-the-art computer vision architectures mimic a small fraction of visual cortex

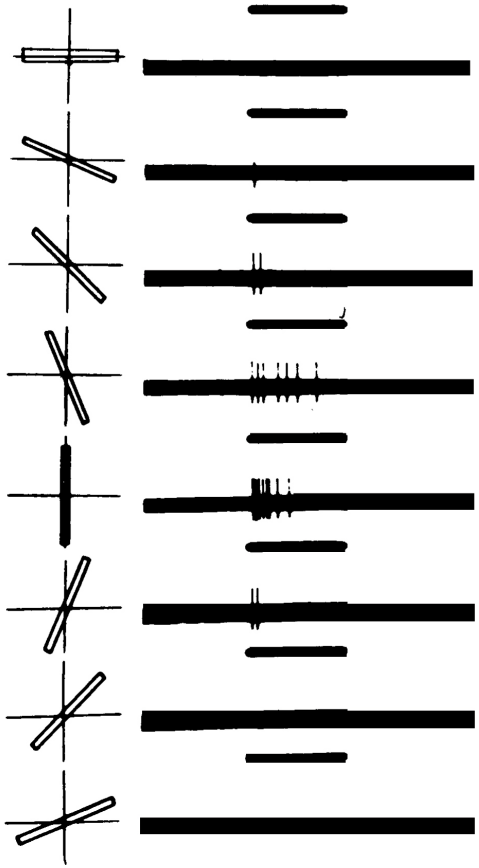


There are ubiquitous top-down connections throughout this architecture

Felleman and Van Essen 1991

Neurophysiology led the way to basic filters

Orientation selectivity



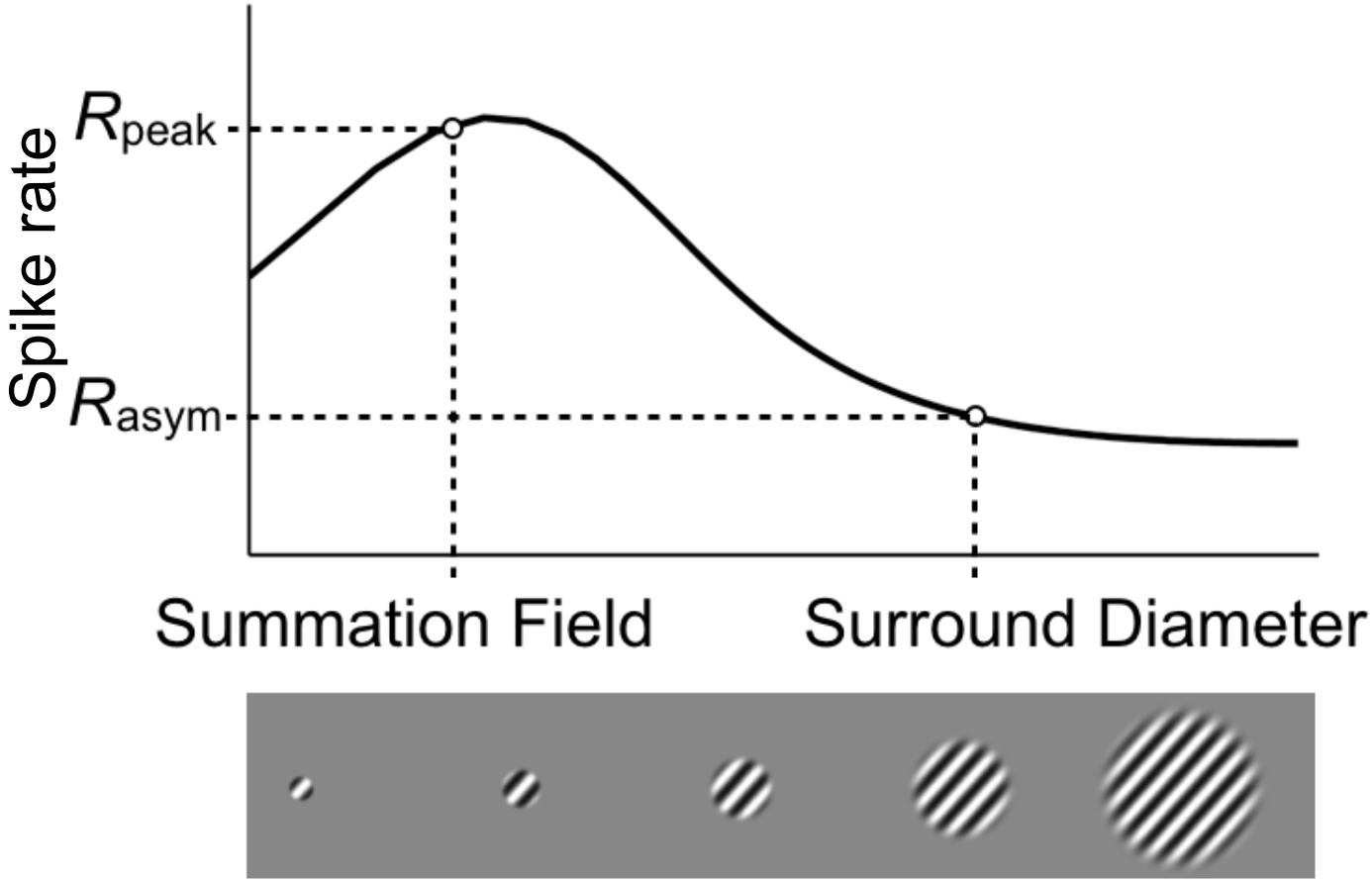
Hubel and Wiesel 1968

Hubel – Nobel Lecture

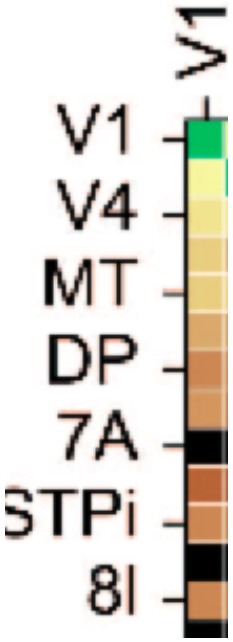
Gabor function

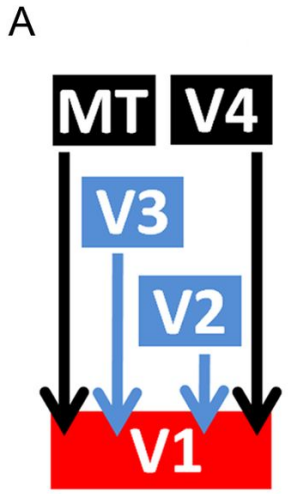
$$D(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{x^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2}\right]\cos(kx-\phi)$$

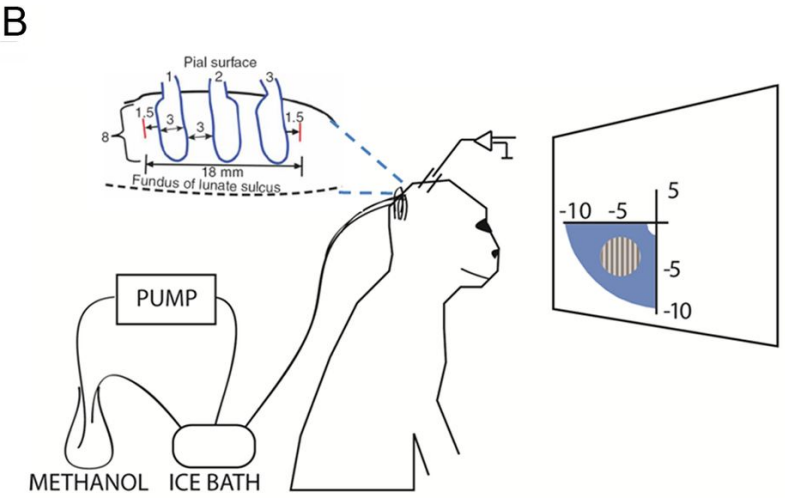
Area summation curve in V1

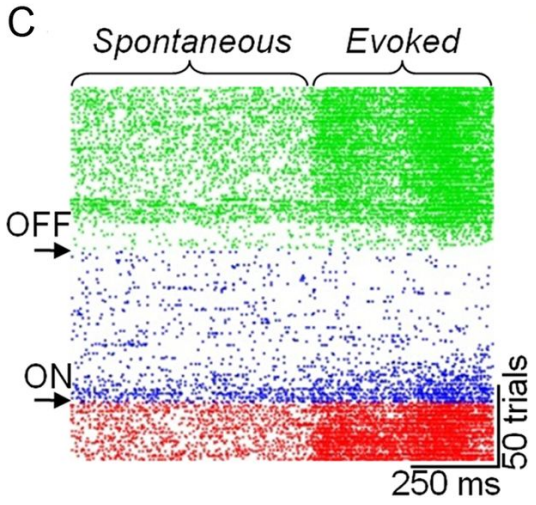


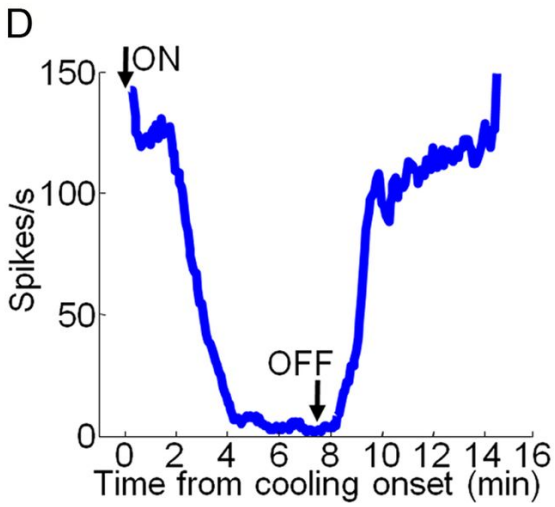
Reversible inactivation of feedback signals (from V2/V3 to V1)





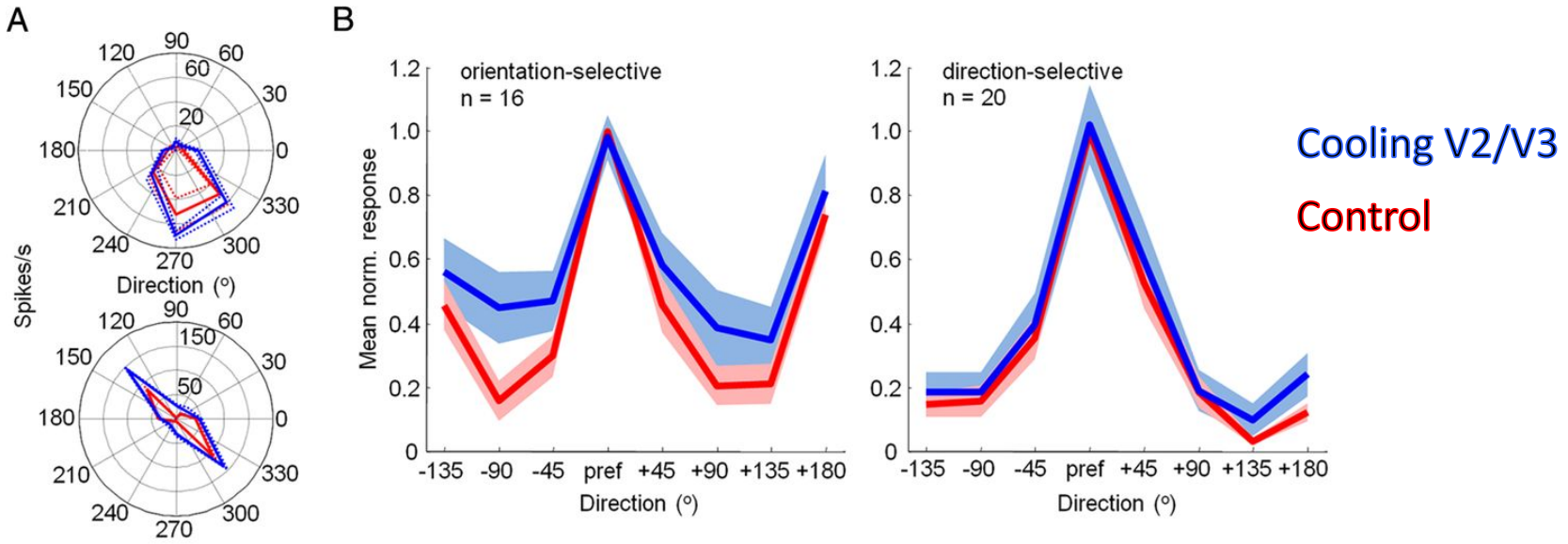






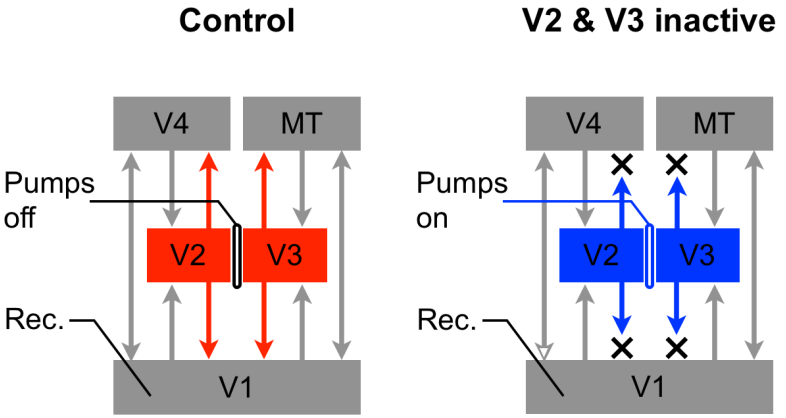
Camille Gomez-Laberge, JoJo Nassi and Richard Born

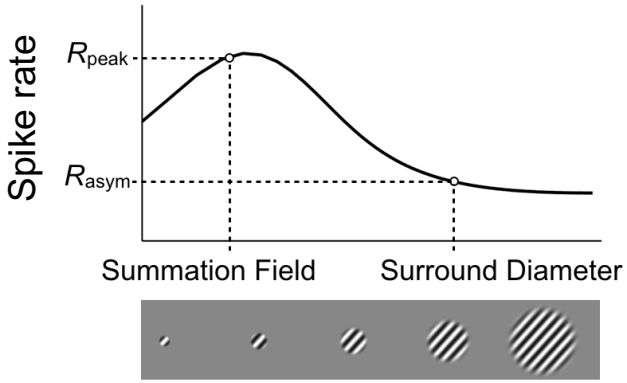
Feedback inactivation does not change orientation or direction selectivity

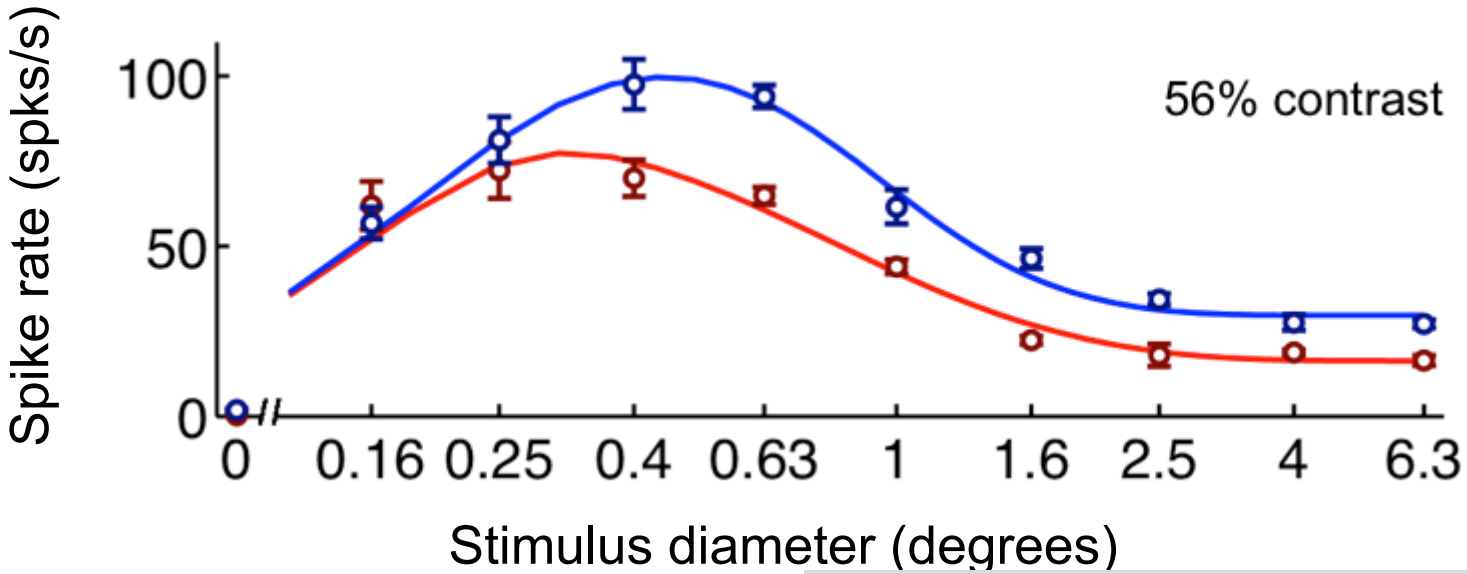


Nassi et al 2013, Gomez-Laberge et al 2014, 2016

Feedback inactivation leads to reduced surround suppression



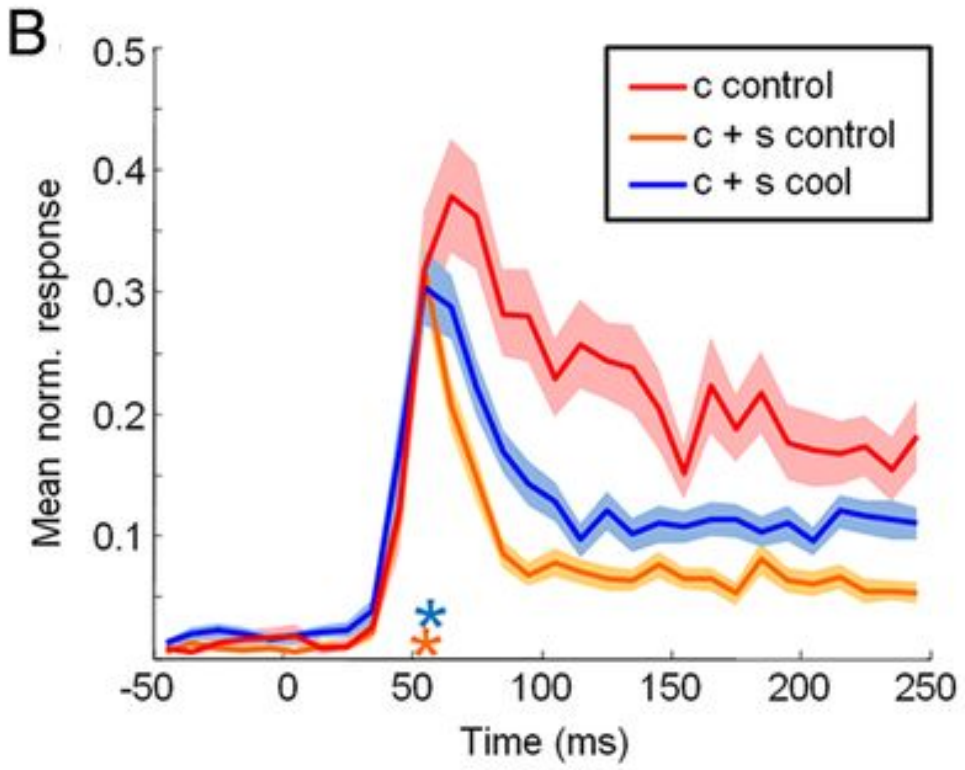




Nassi et al 2013, Gomez-Laberge et al 2014, 2016

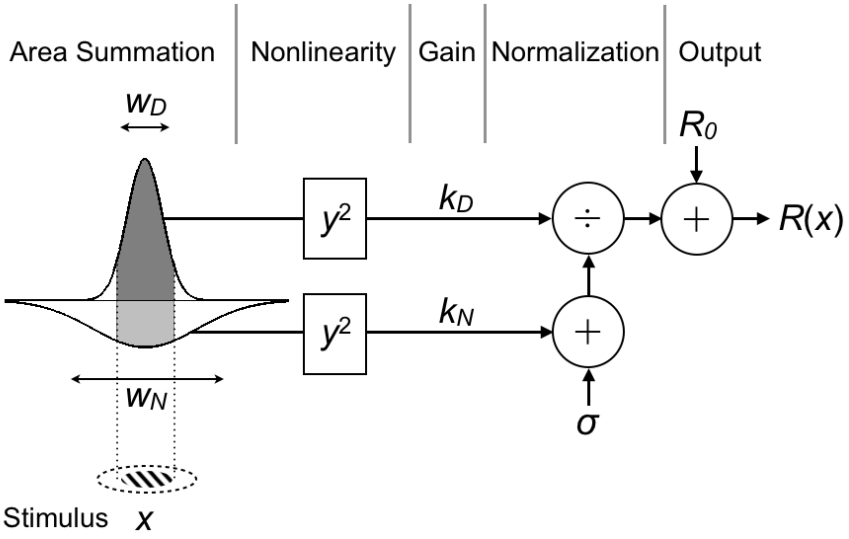
Recurrent connections also contribute to surround suppression (Adesnik et al 2012)

Feedback inactivation effects are delayed



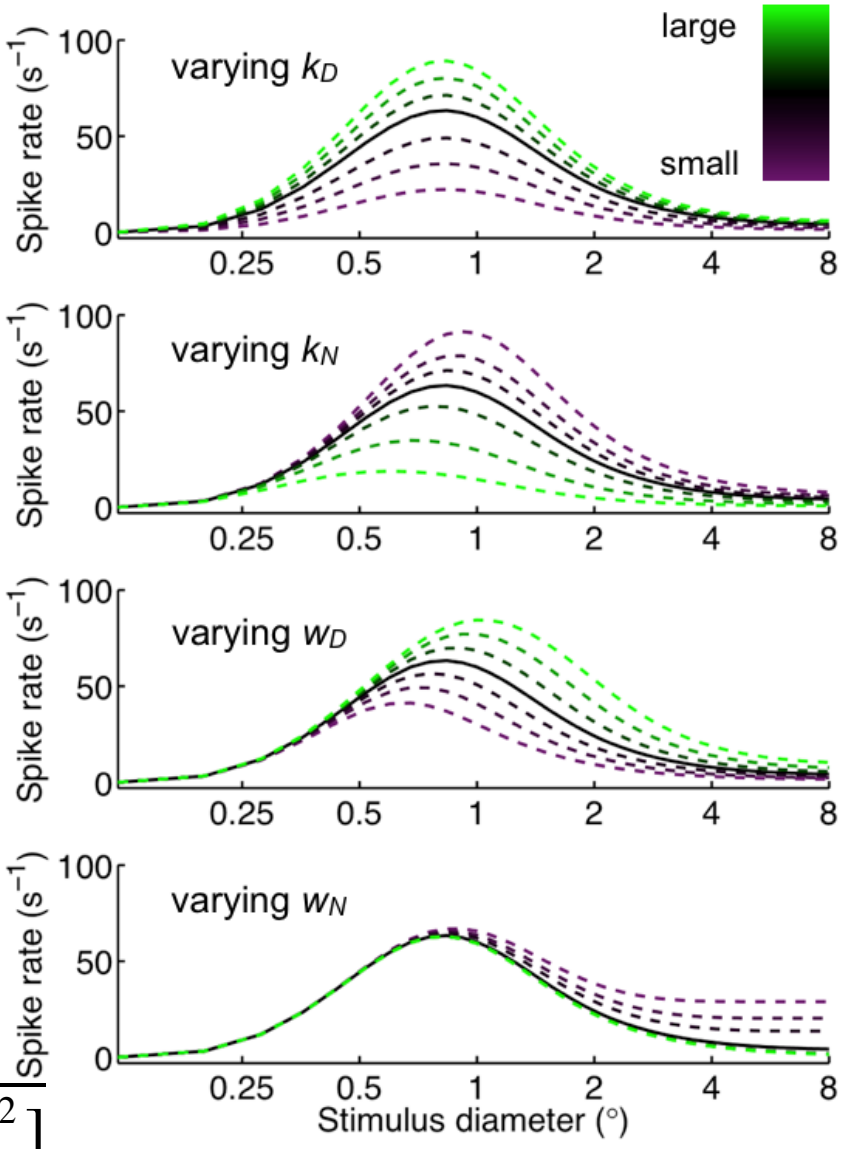
Nassi et al 2013, Gomez-Laberge et al 2014, 2016

A simple normalization model to explain area summation curves



$$R_{ROG}(x) = R_0 + \frac{D(x)}{\sigma + N(x)}$$

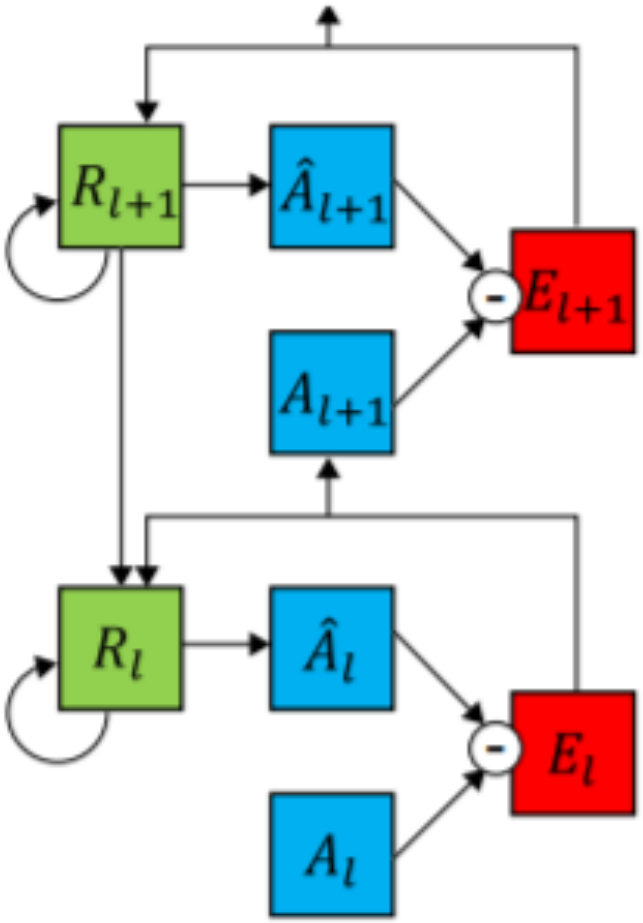
$$R_{ROG}(x) = R_0 + \frac{k_D[w_D erf(x/2w_D)^2]}{\sigma + k_N[w_N erf(x/2w_N)^2]}$$



Deep Learning Implementation of Predictive Coding

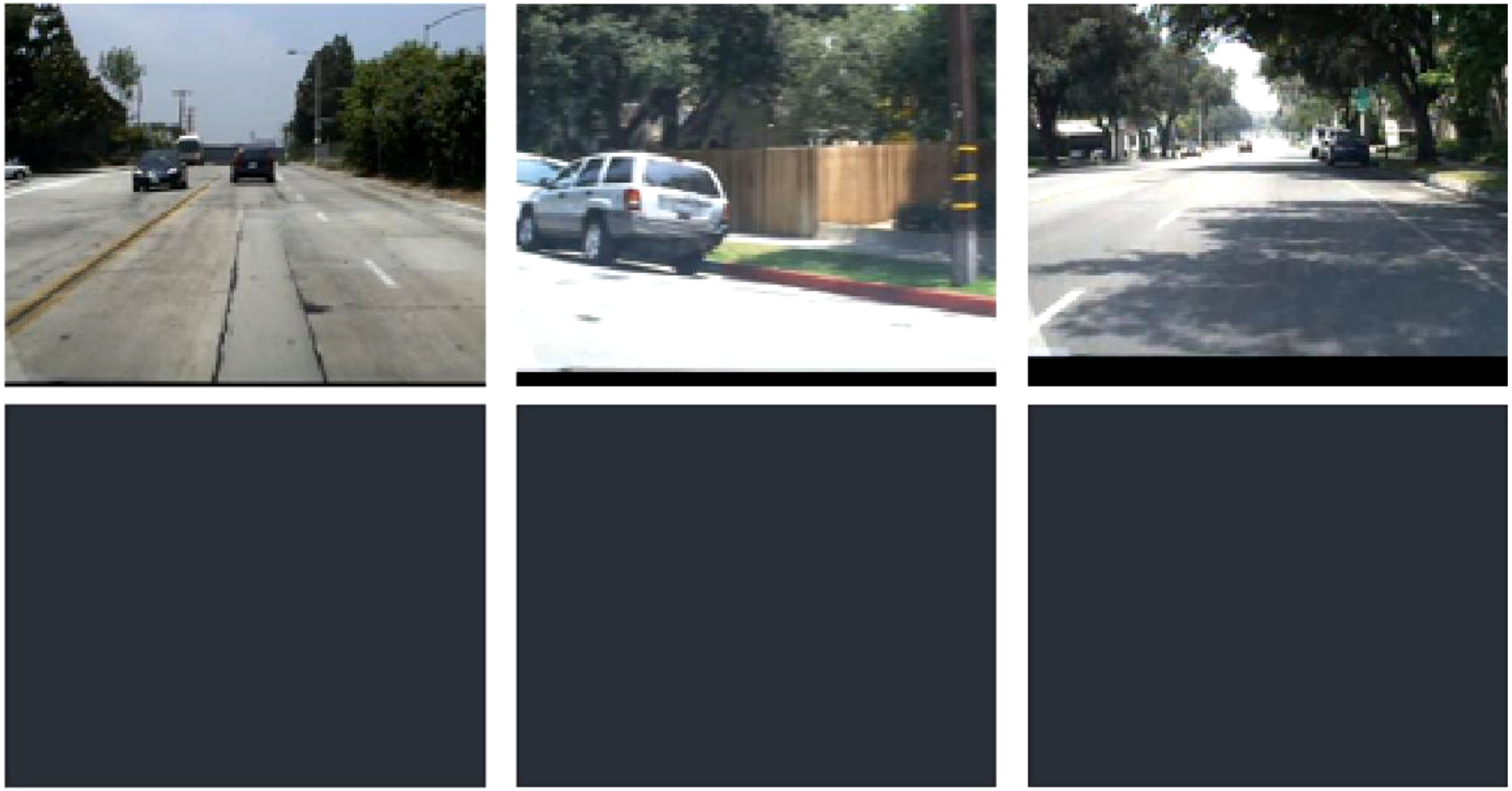
Essential elements:

- "**Representation**" neurons: hold "state of world"
- **Predictions**
- **Targets**
- "**Error**" neurons



Lotter et al 2015, 2016

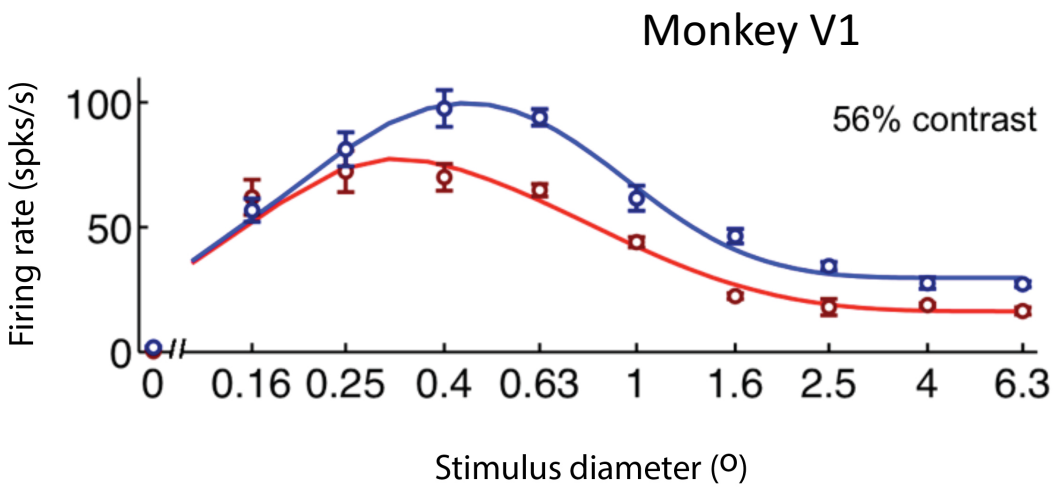
Testing the model on natural video sequences

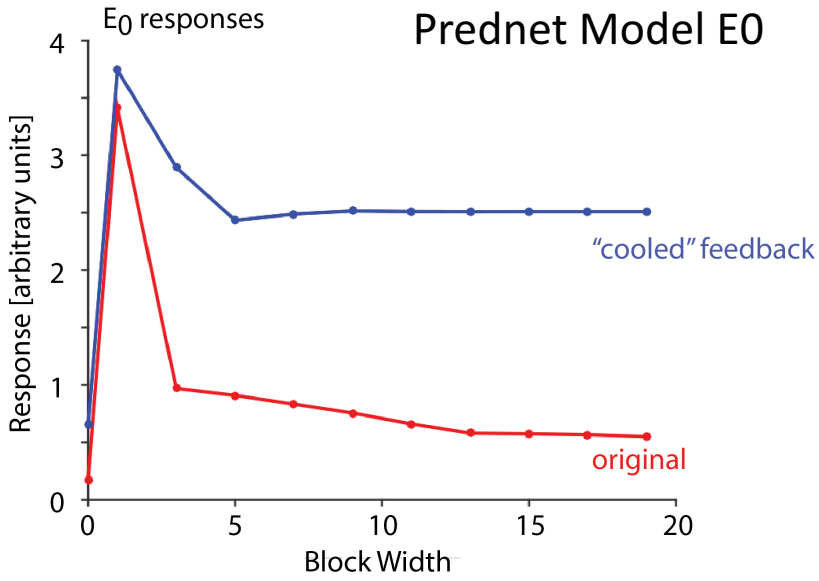


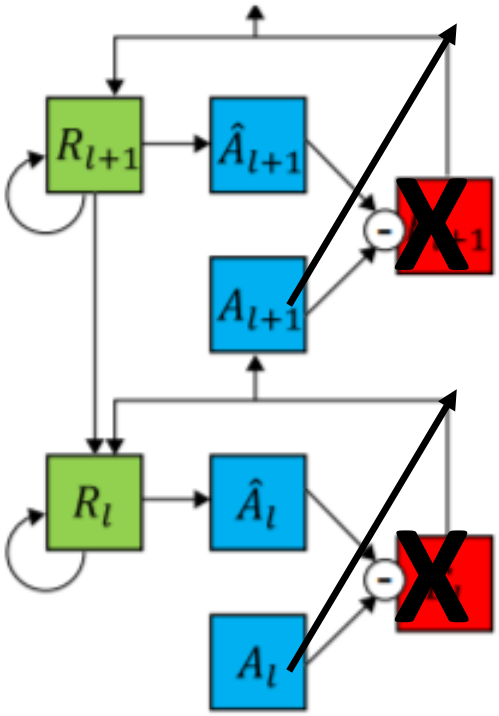
Trained on KITTI Dataset (Geiger et al. 2013)
Tested on CalTech Pedestrian Dataset (Dollar et al. 2009)

Lotter et al 2015, 2016

Removing feedback signals leads to reduced surround suppression

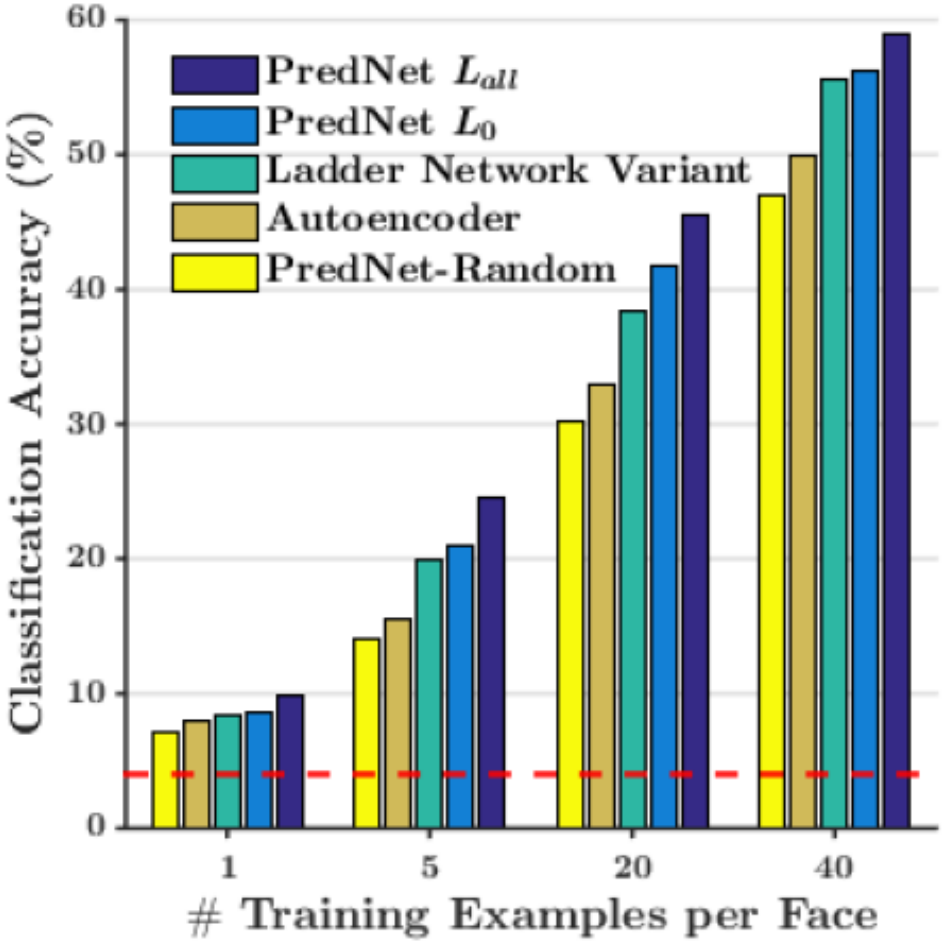






Lotter et al 2015, 2016

Training for prediction → successful image classification



Lotter et al 2015, 2016

Visual cognition: a sequence of routines

Divide et impera

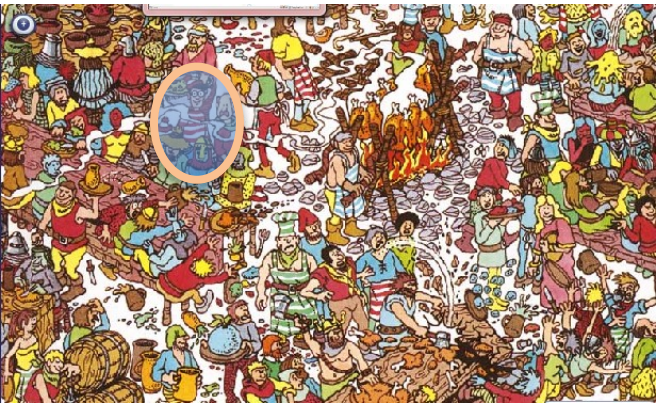
Operations

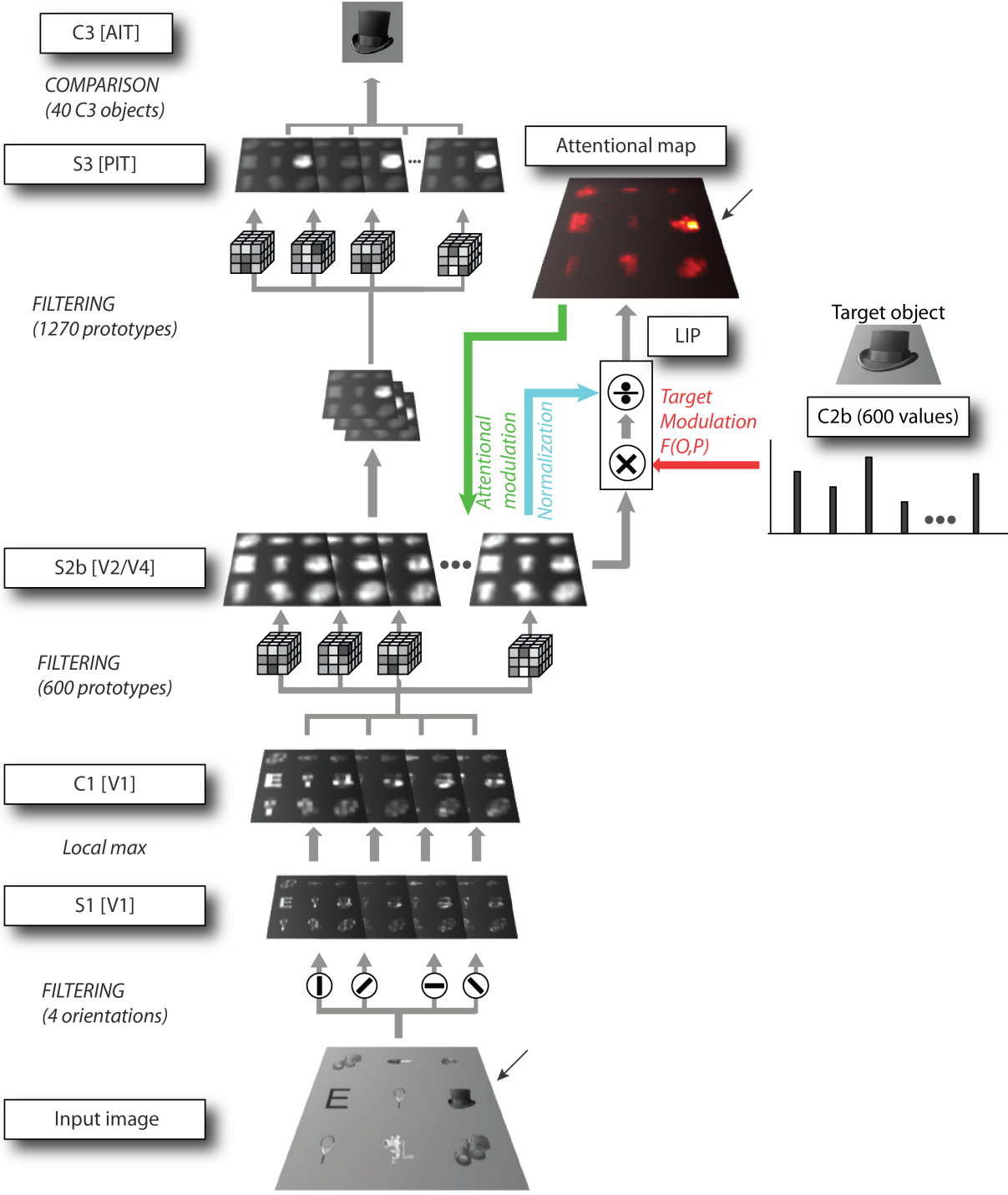
- Candidate labels for foveated region
- Inference and pattern completion
- Candidate representation of the periphery
- **Select target for active sampling (eye movements)**
- Determine spatial relations
- Temporal comparisons
- Store information
- Retrieve previously stored information
- Make spatiotemporal predictions

Visual search

Miconi et al, 2016

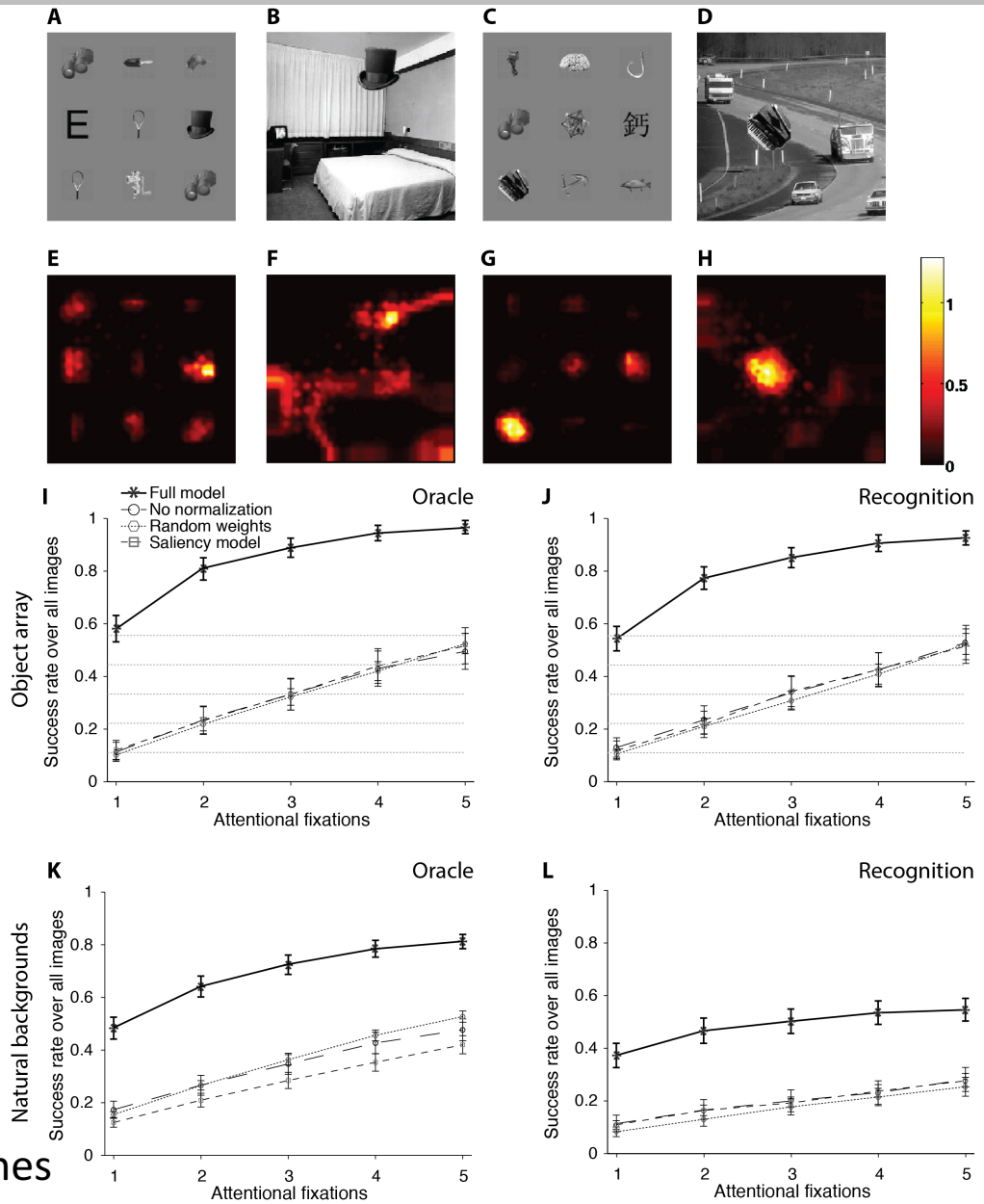
Feedback signals in visual search





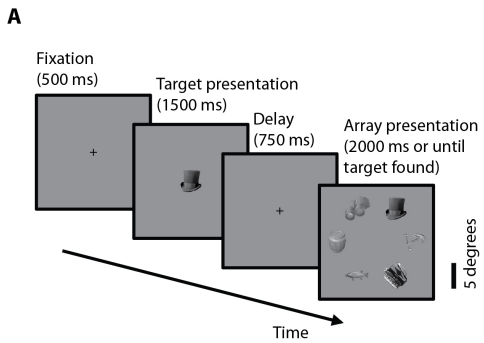
Miconi et al, 2016

The model can find objects in cluttered images



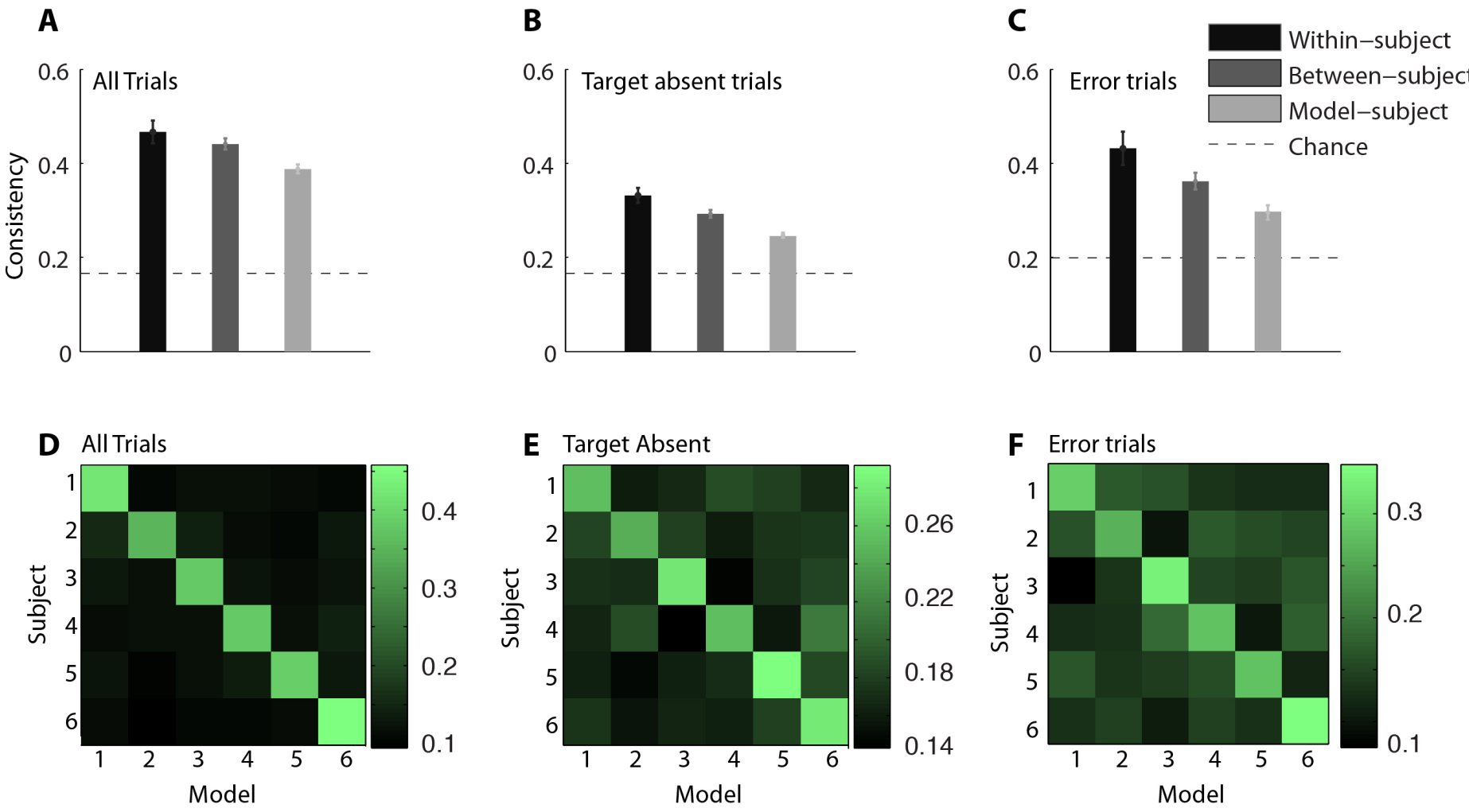
Thomas Miconi, Laura Groomes

The model's performance is comparable to human behavior

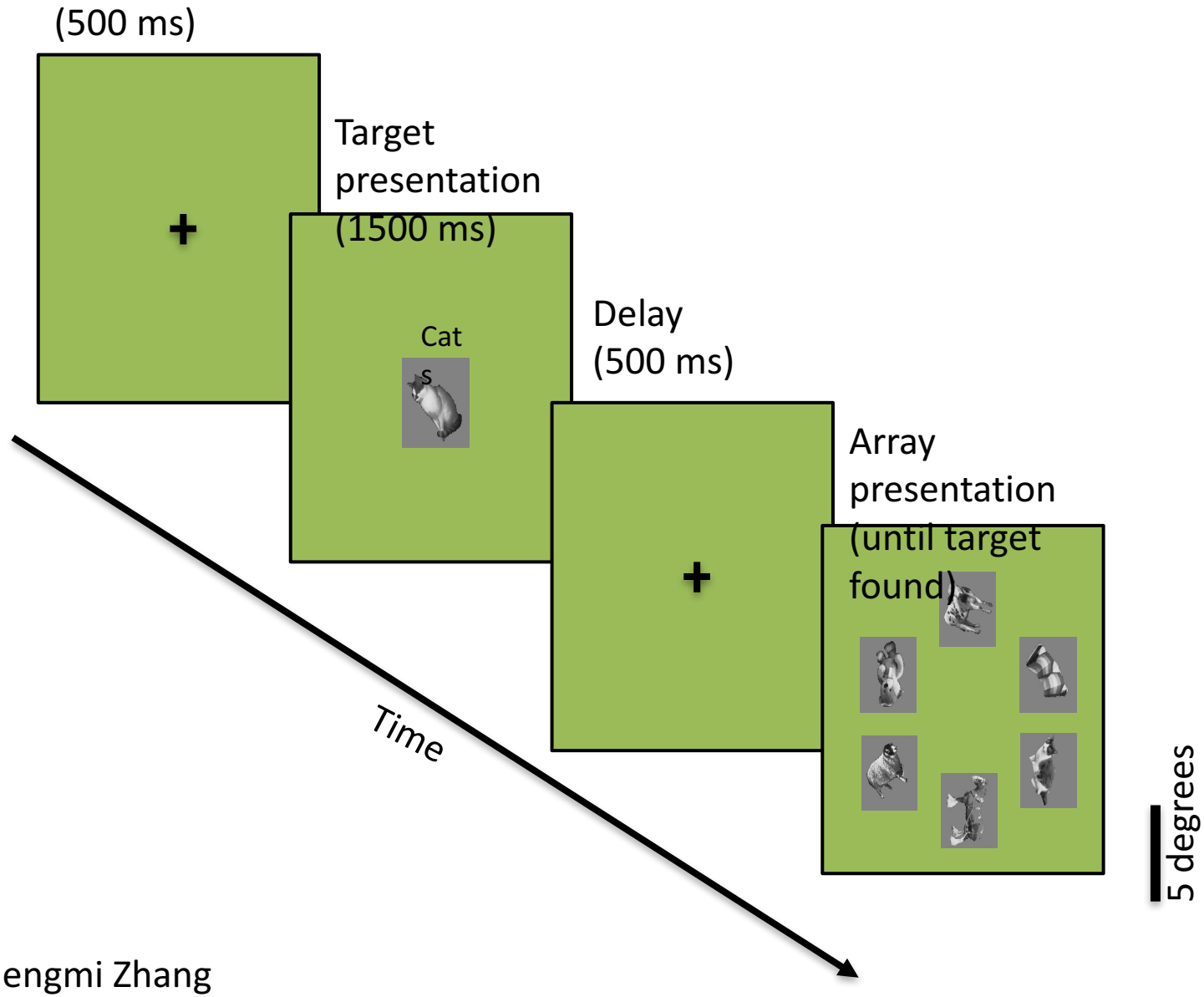


Thomas Miconi, Laura Groomes

Further comparisons between humans and model

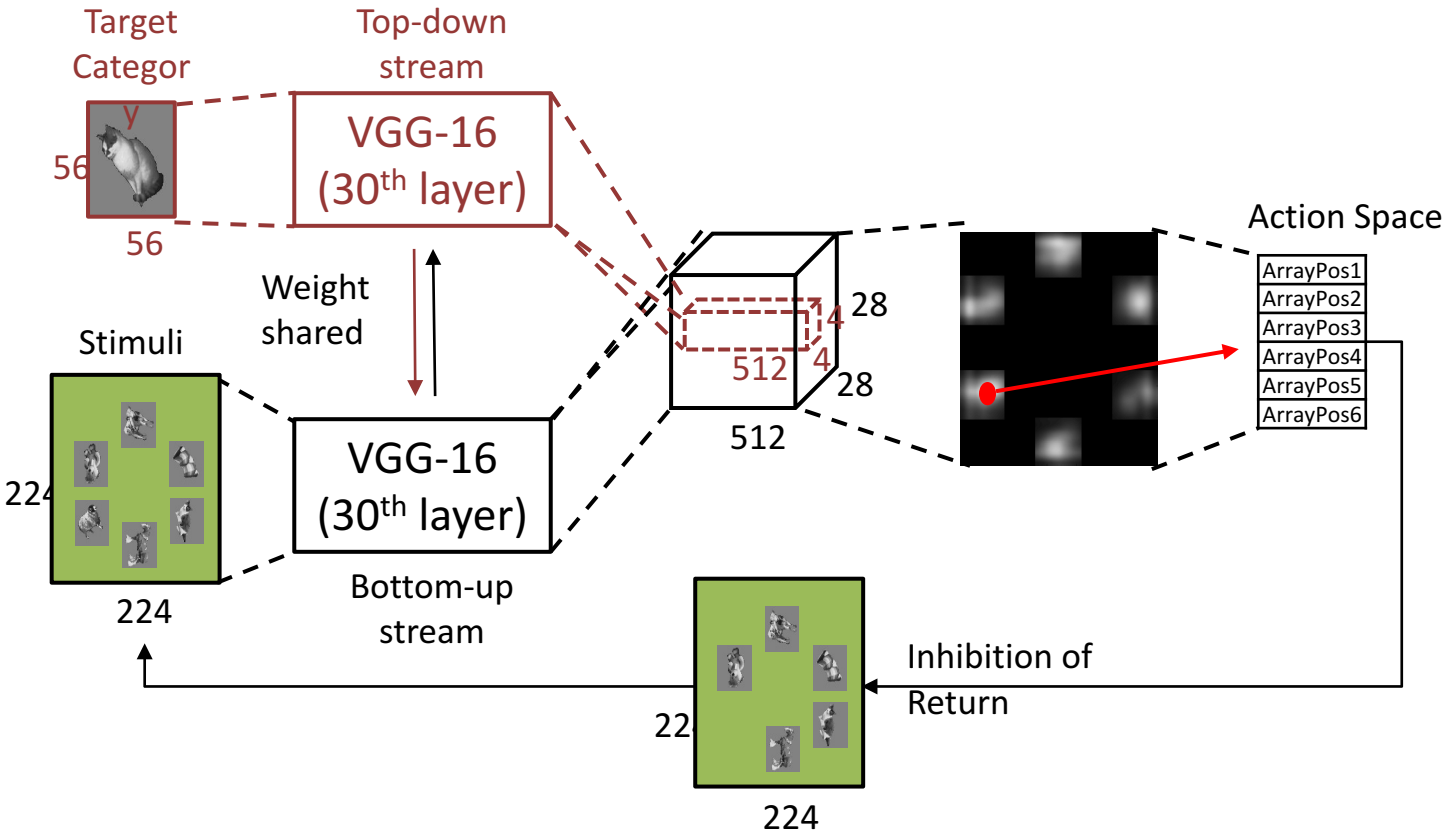


Invariance in visual search



Mengmi Zhang

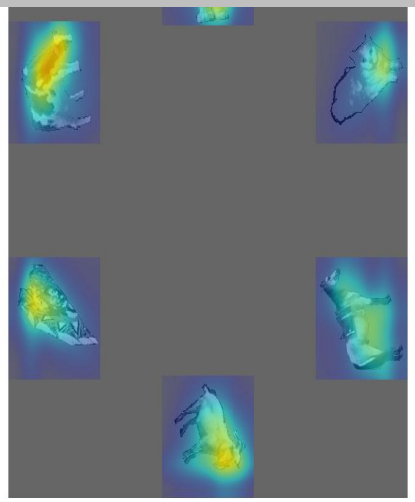
Invariance in visual search -- Model

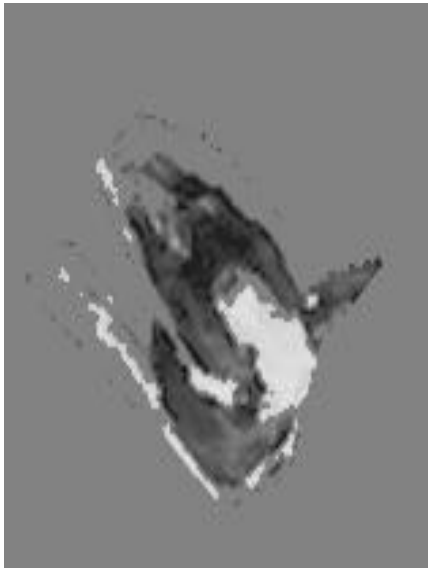
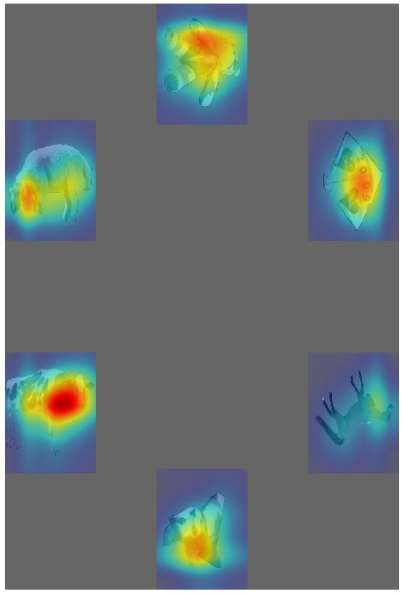


Mengmi Zhang

Invariance in visual search – Example

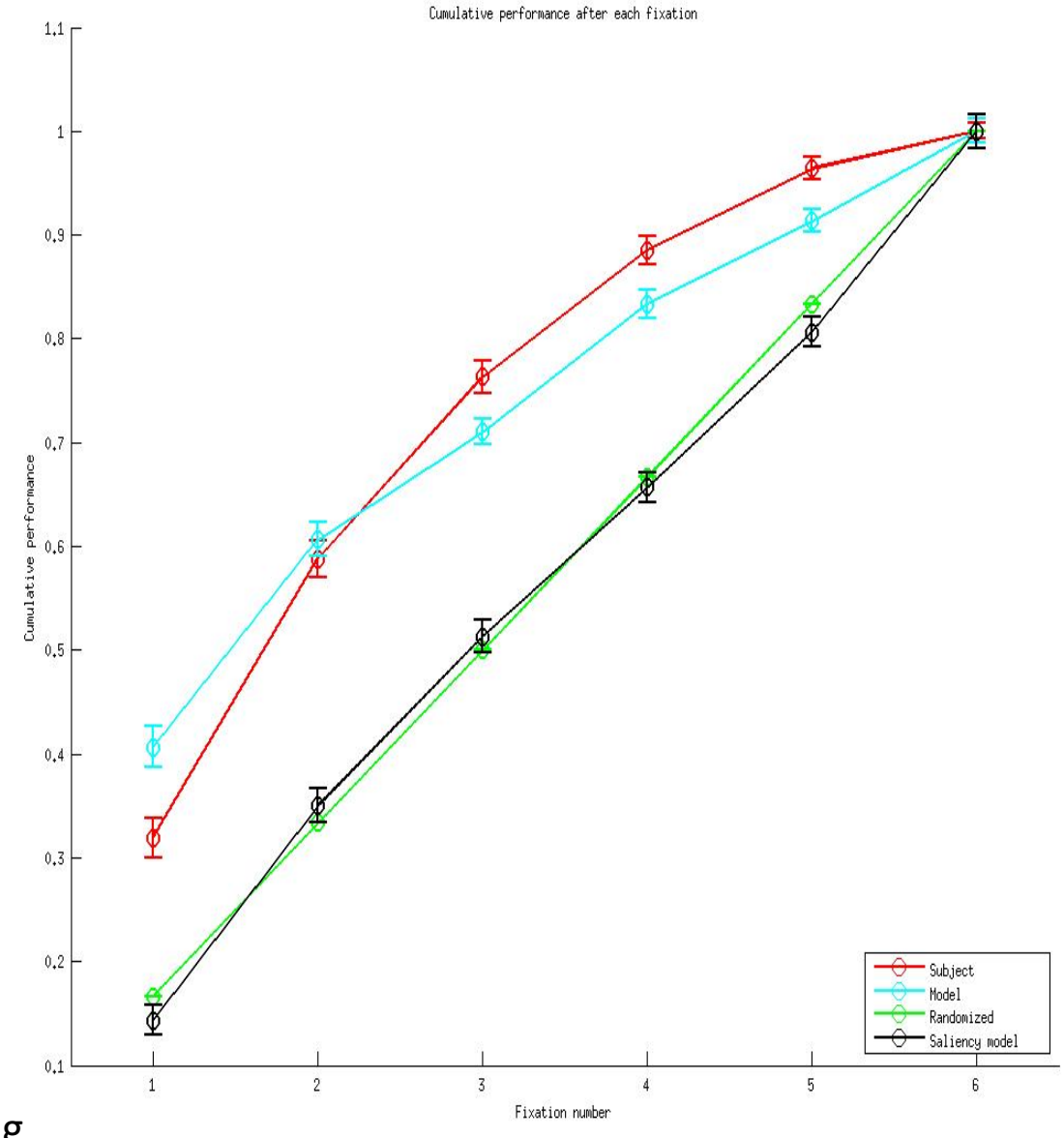




Mengmi Zhang

Invariance in visual search -- Performance



Mengmi Zhang

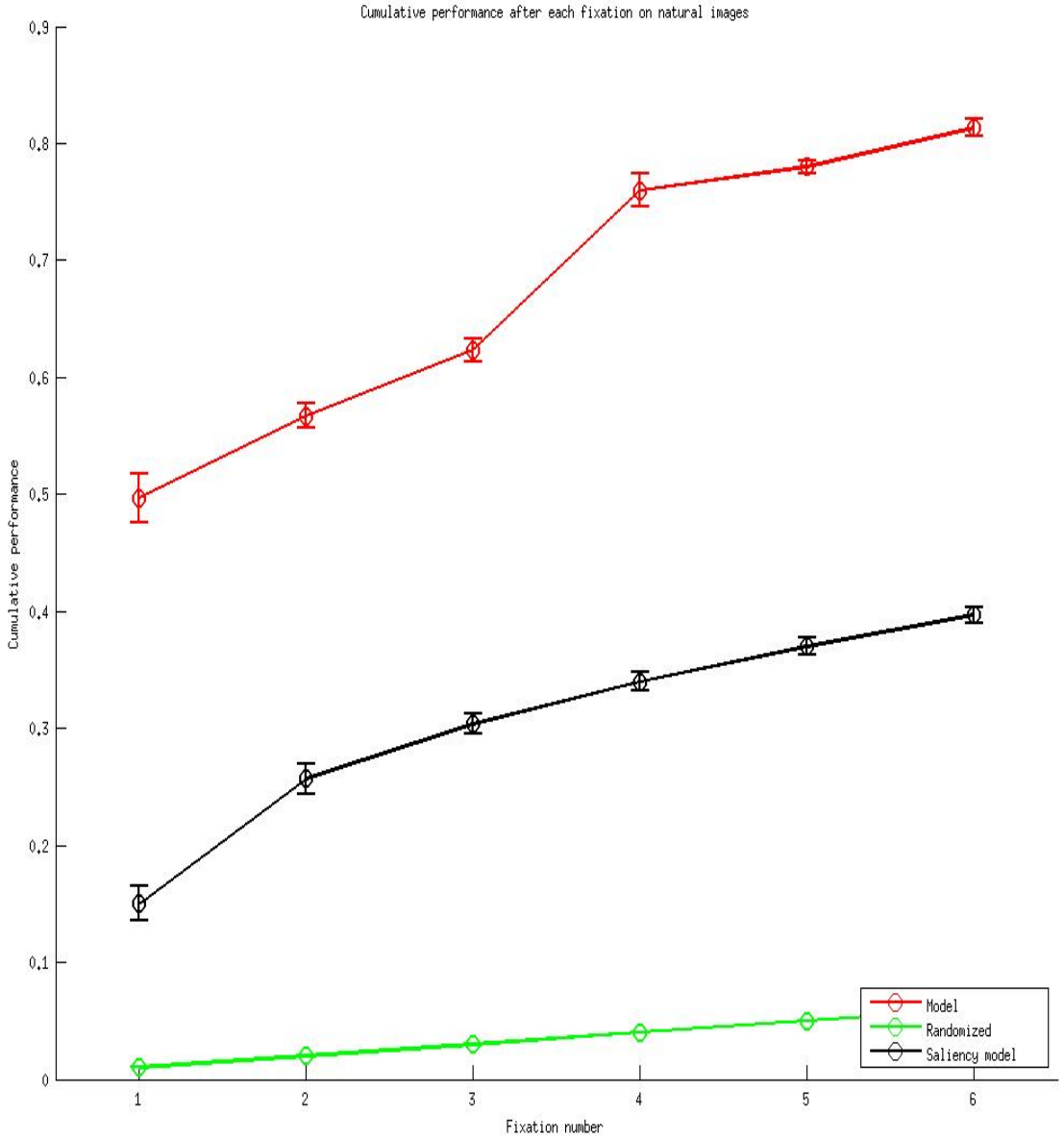
Invariance in visual search (Natural images)






Mengmi Zhang

Invariance in visual search (Natural images)



Mengmi Zhang

Visual cognition: a sequence of routines

Divide et impera

Operations

Candidate labels for foveated region

Inference and pattern completion

Candidate representation of the periphery

Select target for active sampling (eye movements)

Determine spatial relations

Temporal comparisons

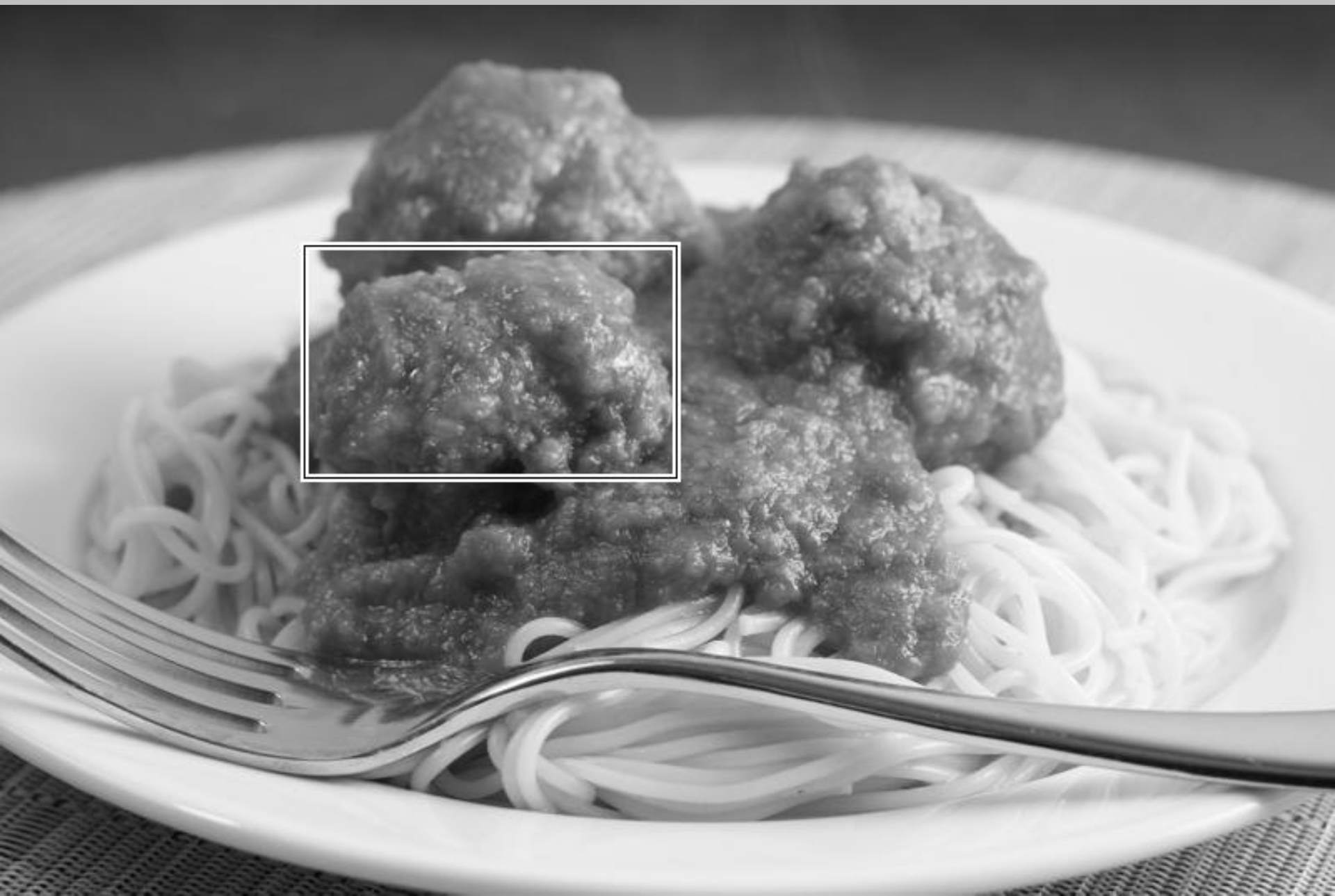
Store information

Retrieve previously stored information

Make spatiotemporal predictions

Context example 1

Context example 2



Context example 3

Context example 4

Context example 5

The brain's operating system

Gabriel Kreiman

Gabriel.kreiman@tch.harvard.edu

klab.tch.harvard.edu

Center for Brains, Minds and Machines

Charlotte Moerman

Camille Gomez

Martin Schrimpf

Richard Born

Jojo Nassi

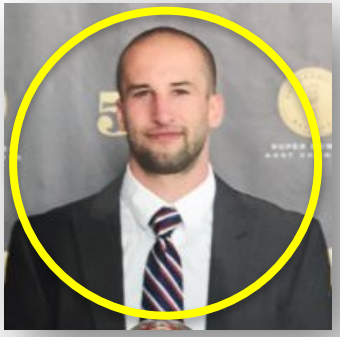
Laura Groomes

Joseph Madsen

Hanlin Tang

Thomas Miconi

Bill Lotter



David Cox

Stan Anderson

Moving forward

1. Other cues (e.g. stereo, what is in front of what, context, world knowledge)
    a. Stereo cues, what is in front of what
    b. Context
    c. World knowledge (people sit on chairs, things fall if there is no support, kitchens may have coffeemakers but not giraffes)

2. How to integrate information across space and time to understand a scene
    a. Integrating fovea and periphery
    b. Integrating information across saccades

3. The role of memory
    a. To compare sensory inputs to stored representations
    b. Working memory

Computational roles of recurrent/feedback signals

1. Pattern completion (recurrent computations)
2. Predictive coding (feedback computations)

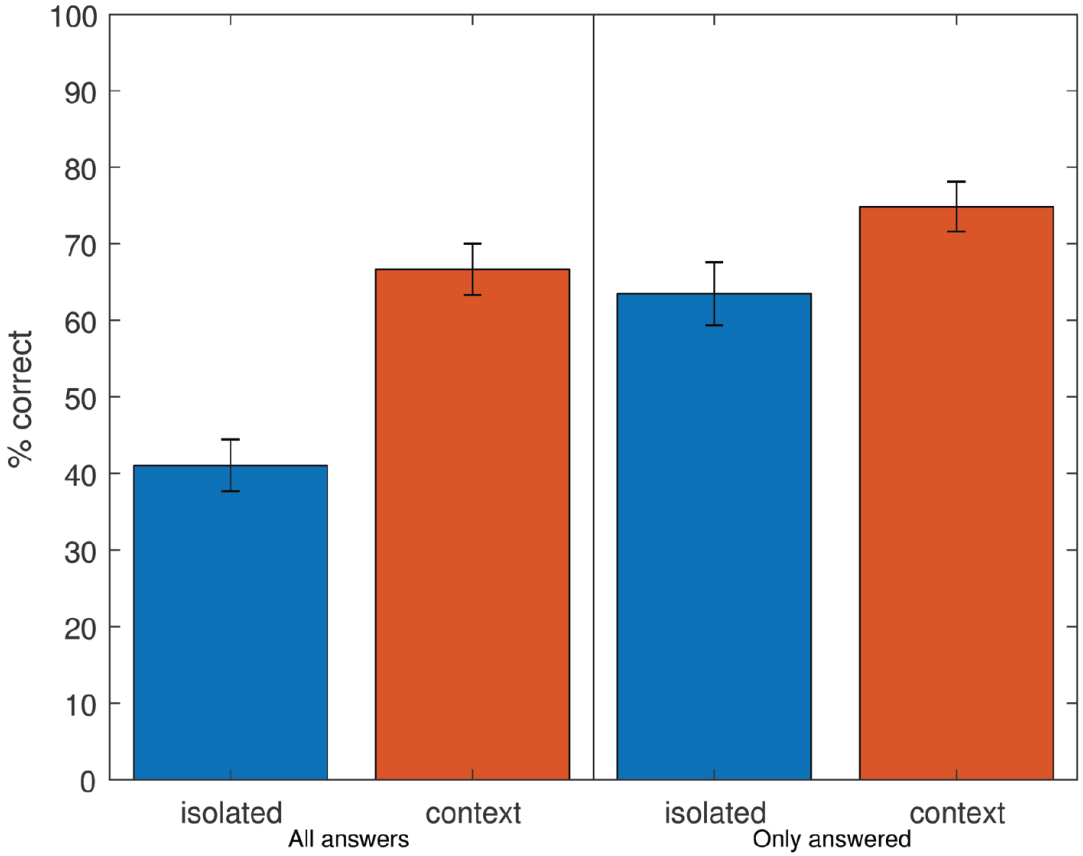
2.1 Surround suppression

2.2 Predicting the next frame in video sequences

2.3 Predicting eye movements during visual search

2.4 The role of contextual information in visual recognition

Context matters



7.3.: Human performance on objects in isolation and objects with context: considering only unmasked trials with a presentation time of 500 ms, percent correct is compared on images shown in isolation and with visual context. Responses in the context condition significantly outperformed responses in the isolated condition, both when comparing *all answers* ($41.04\% \pm 3\%$ vs $66.67\% \pm 3\%$, $p = 1.20\text{e}{-7}$) and also when analyzing *only answered* responses where an answer other than silence or "I don't know" were given ($63.50\% \pm 4\%$ vs $74.86\% \pm 3\%$, $p = 2.91\text{e}{-2}$)

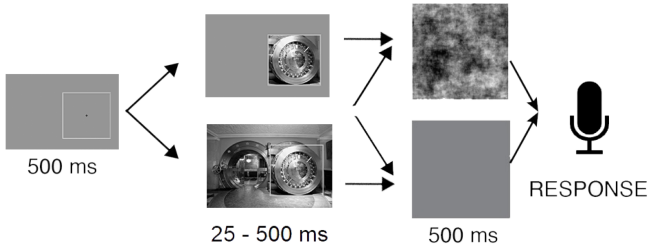


Figure 7.2.: Visual context psychophysics experiment: Fixations with the white bounding box of the bounding box were shown for 500 ms and checked with an eyetracker if possible. The object was then shown in either the isolated (top image) or the context (bottom image) condition with varying SOAs. In half of the cases, the image was followed by a noise mask (masked) and in the other half by a gray screen (unmasked) for 500 ms. Subjects were then required to verbally name the object within the white box. These responses were later transcribed by the investigators.

Martin Schrimpf

Context leads to better educated guessing

gorilla, orangatang, orangutan
bear, gorilla
bear, clown, monkey

freezer
cabinet, drawers, refrigerator
fridge

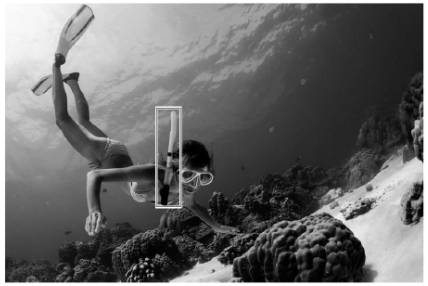
airport, carrier, fuel, lorry, refueling, tank, truck
airplane, shell
train

christmas, gift, present
pattern, sitting, striped, toy
dress, present

beans, carrots, green, vegetable
french, fries, pasta
arm

beach, sand
island, palm, tree
cruise, ocean, ship

banjo, carrier, case, guitar, instrument
case, music
couch, flower, head, sofa



snorkel, snorkle
bridge, snorkel
arm, leg, purse

Figure 7.4.: Objects with frequent errors with and without context: we picked 8 images where subjects in both conditions (context/isolated) made the most errors. The image displayed is always with context; below are the unique words from different responses, ignoring empty responses. Mturk responses are shown in the top line (black), context responses in the middle (blue) and isolated responses on the bottom (orange). Although most responses are wrong, the

Backward masking disrupts object recognition

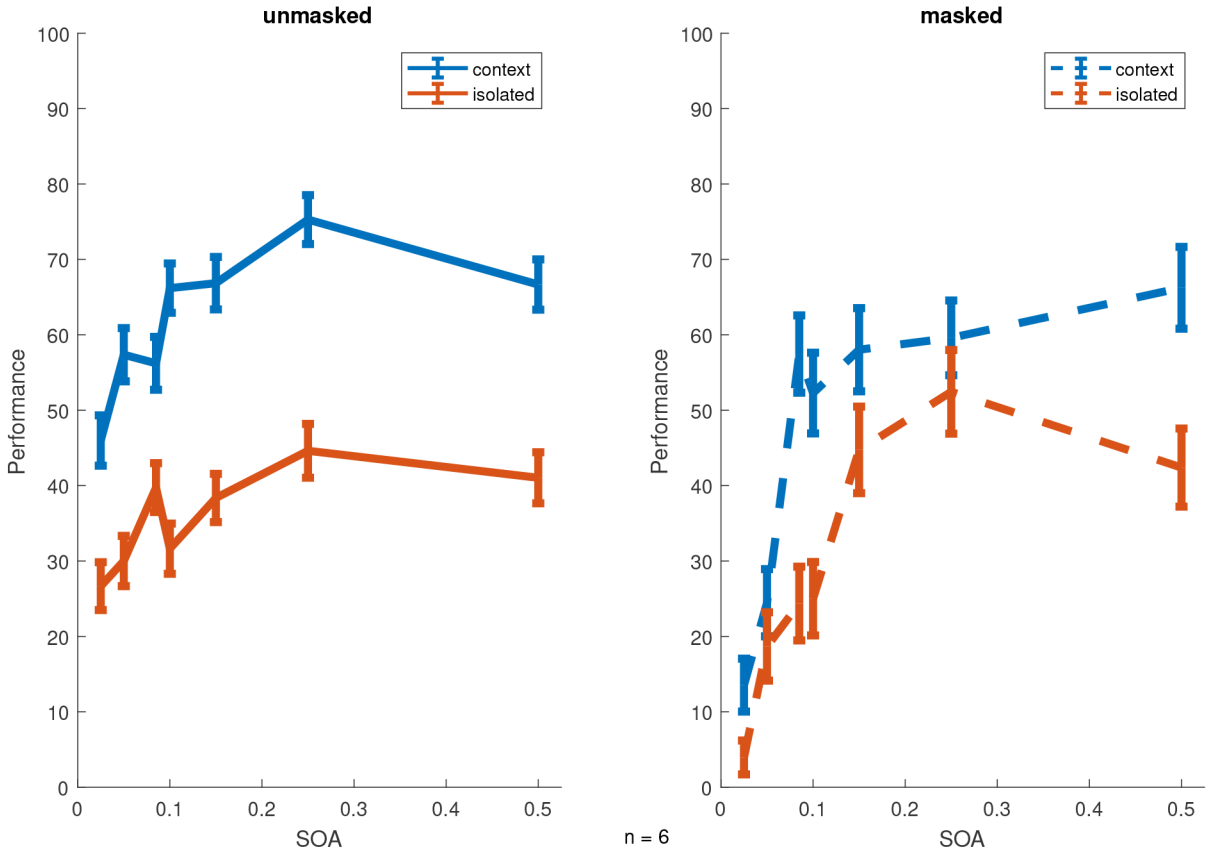
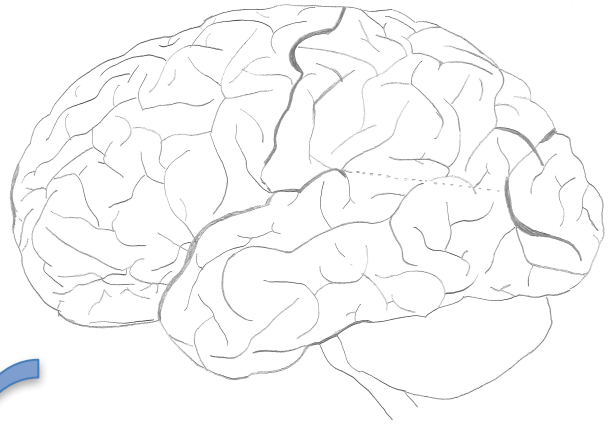
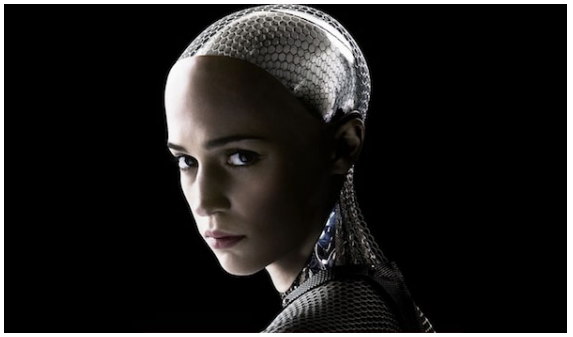


Figure 7.5.: Effects of backward masking on recognition performance at different SOAs: context performance is consistently above isolated performance (all $p << 0.05$) and context affects both conditions equally ($p = 0.29$). Backward masking significantly affects both conditions for SOA = 25 ms ($p = 1.61e{-}8$ and $p = 2.36e{-}5$ respectively) but only inconsistently above. More tests are necessary to conclude the role of recurrent computations in the context as well as in the isolated case

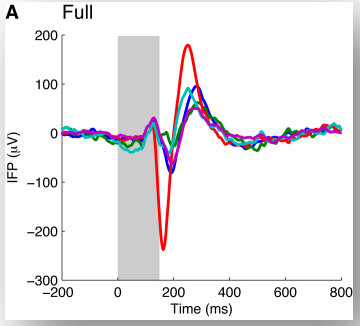
From biological codes to computational codes

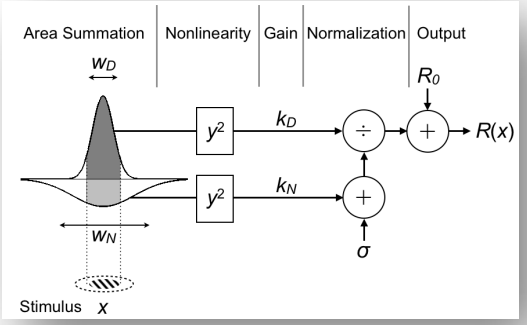




Biological codes

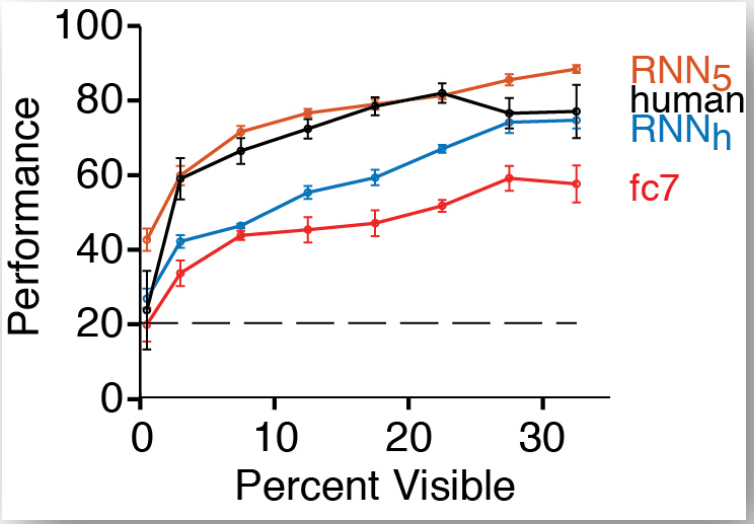
Computer codes



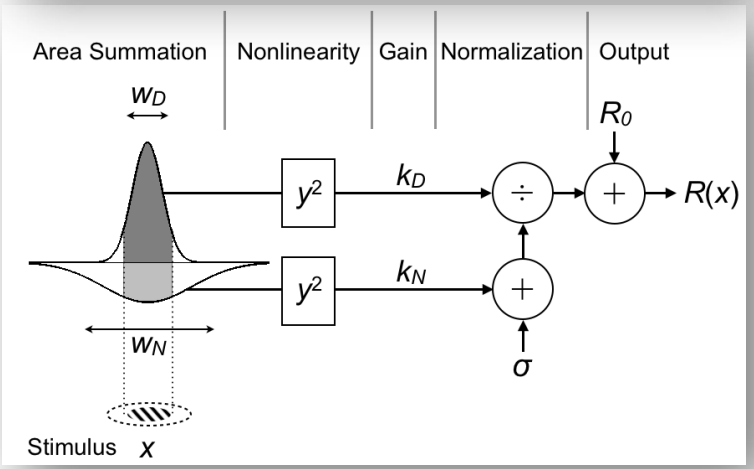


Summary

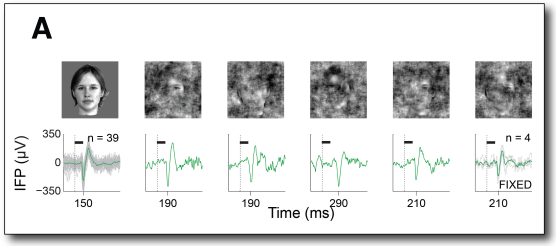
Pattern completion: Recurrent connections can help recognize heavily occluded objects and make inferences from partial information

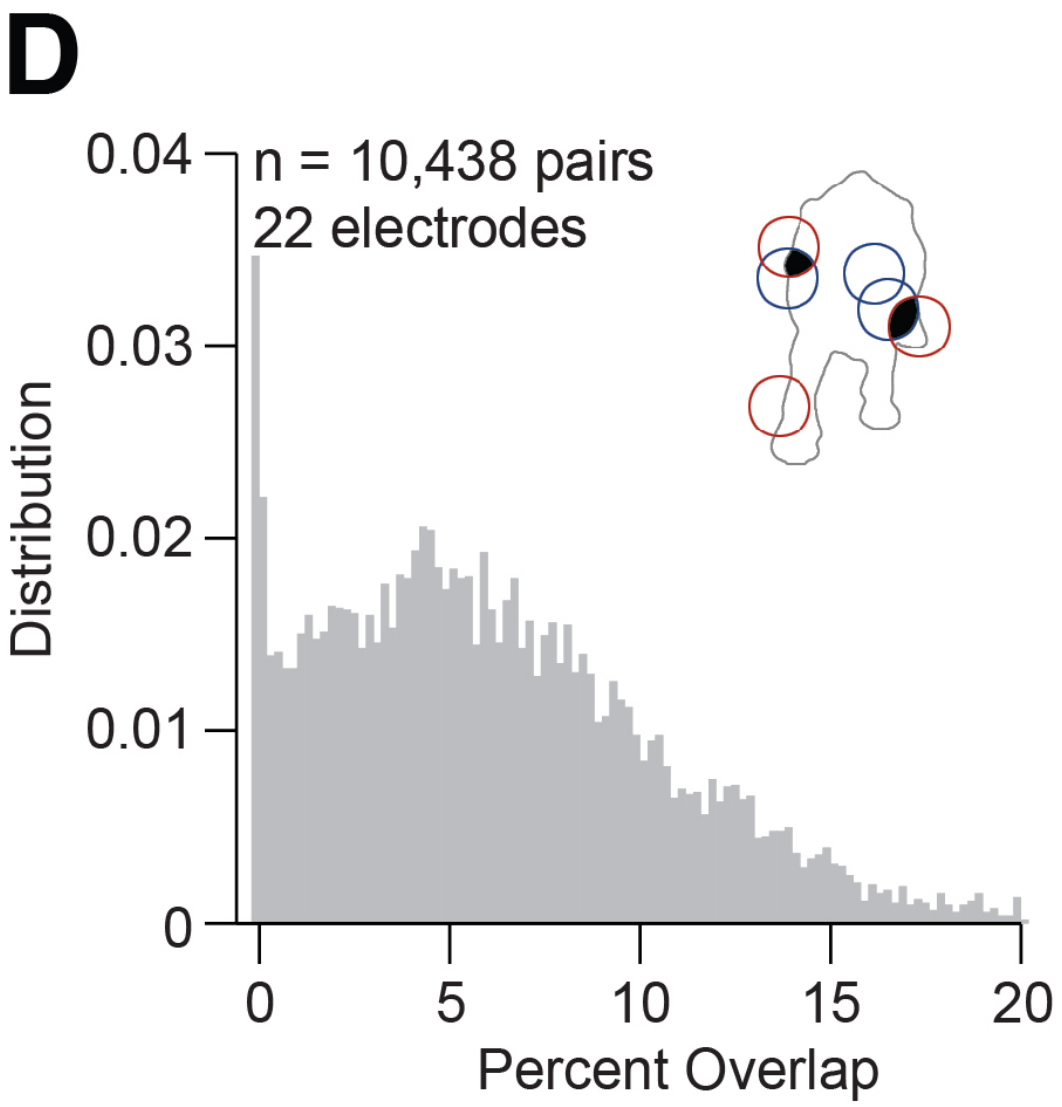


Feedback signals enhance surround suppression and may provide a signal for predictive coding that can help in unsupervised learning

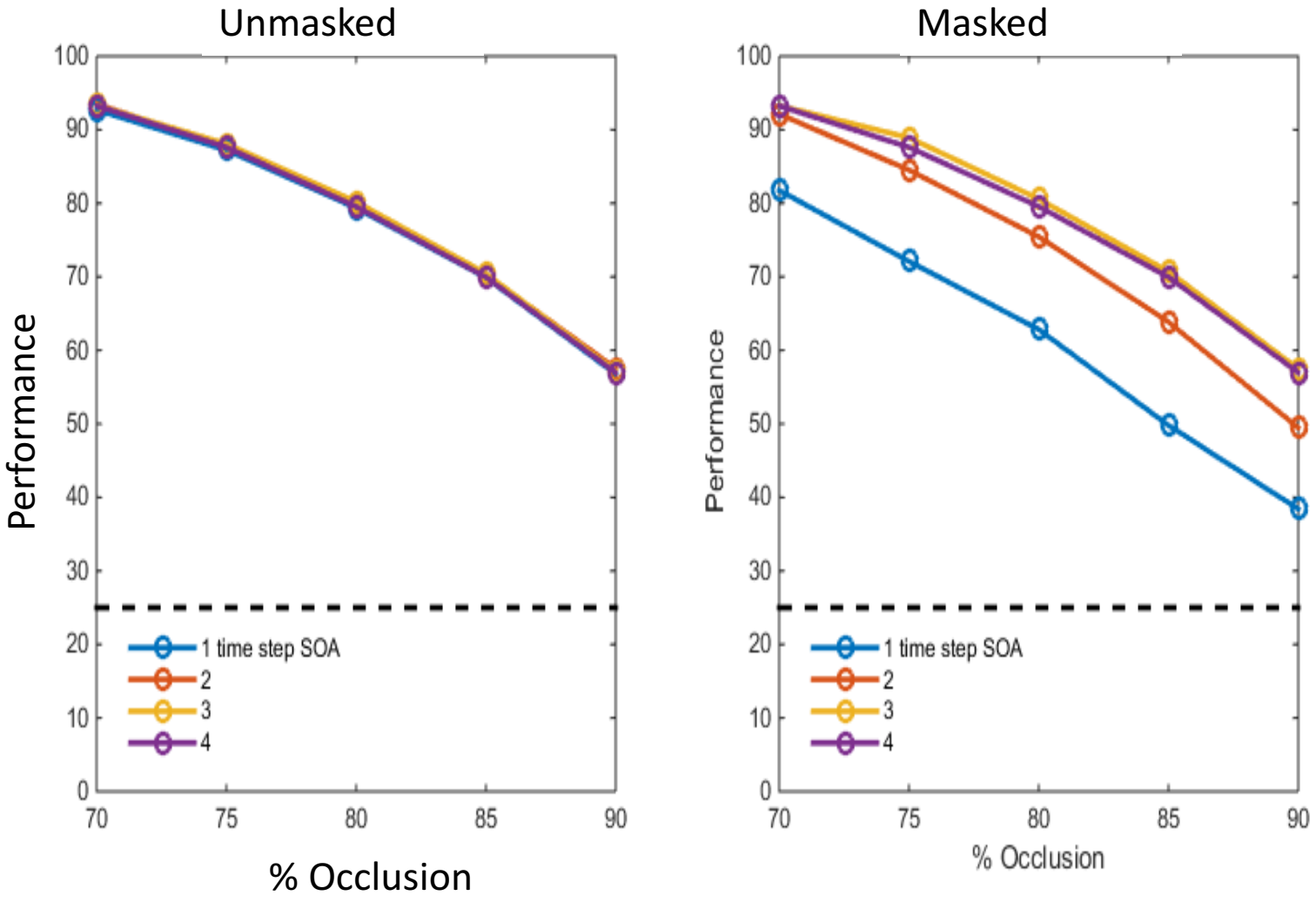


Holistic responses (?)

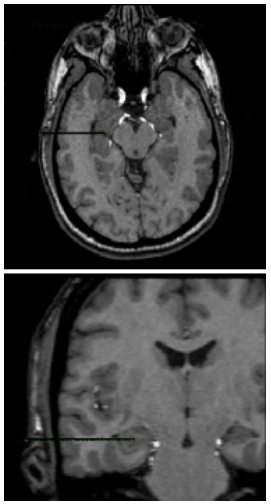


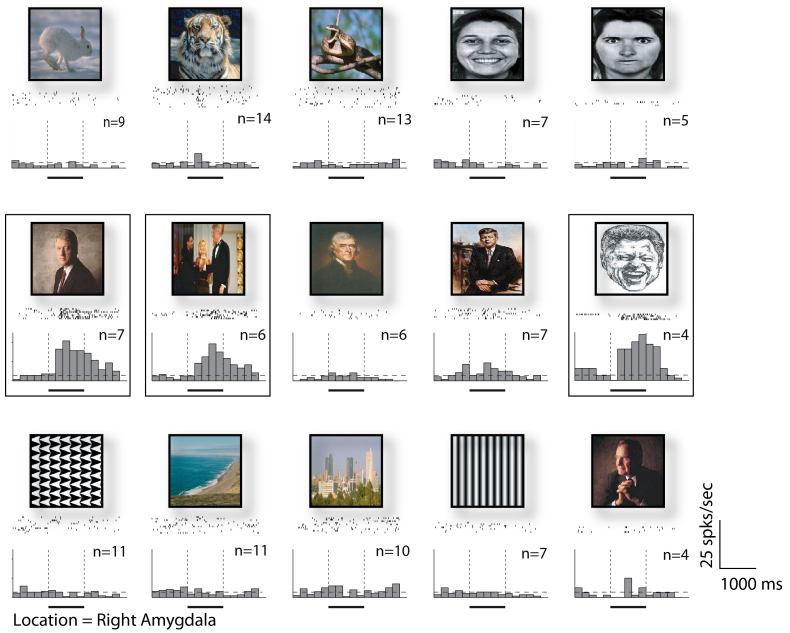


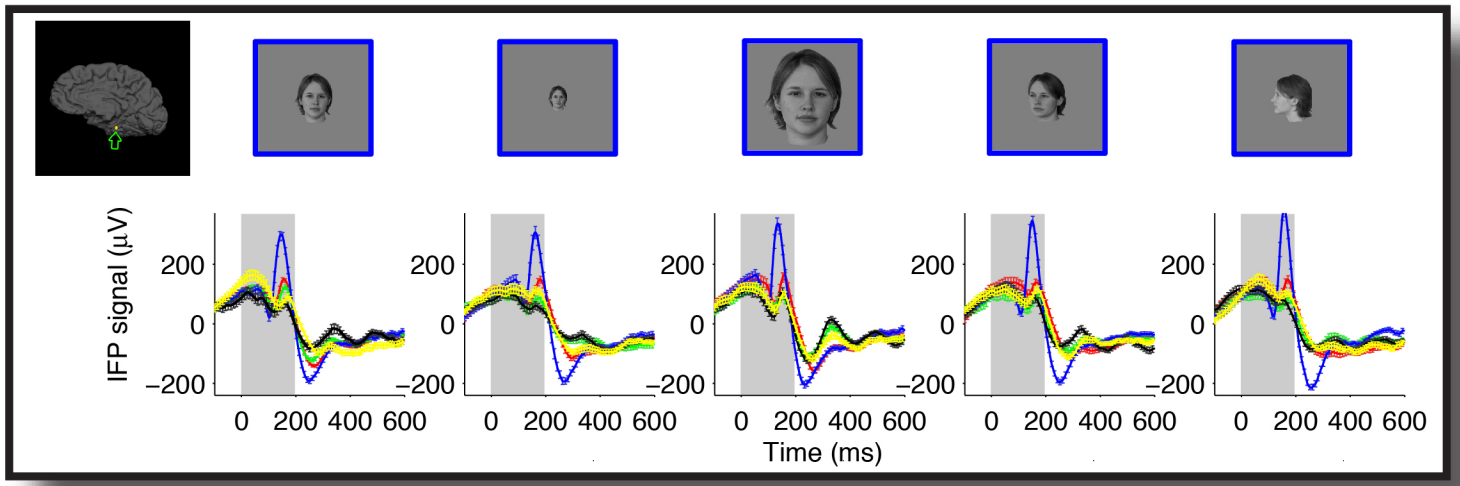
Model performance in masking experiment



We can interrogate neural circuits in the human brain







Supplementary Figure 4

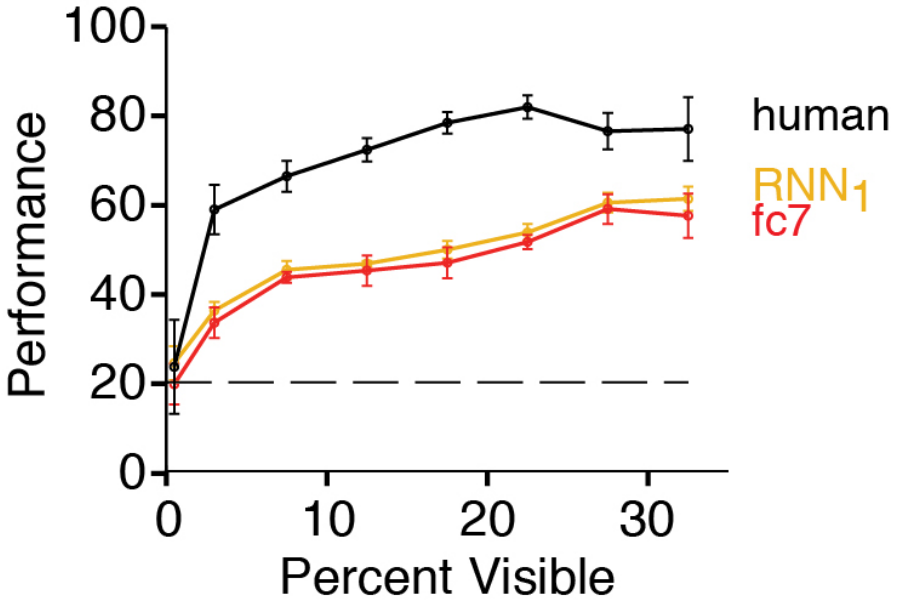


Figure S4: Performance of the RNN_1 model

Using the same format and conventions as in **Figure 6C**, we show performance as a function of object visibility for the RNN_1 model. The performance of the fc7 model and human performance are copied from **Figure 6C** for comparison purposes.

Supplementary Figure 5

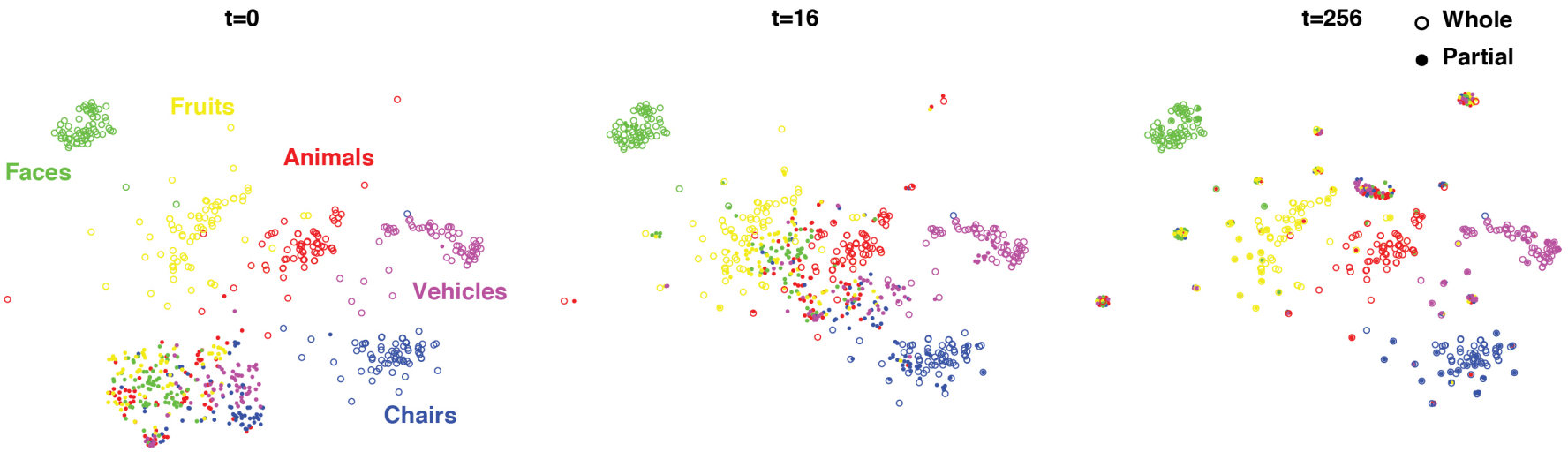
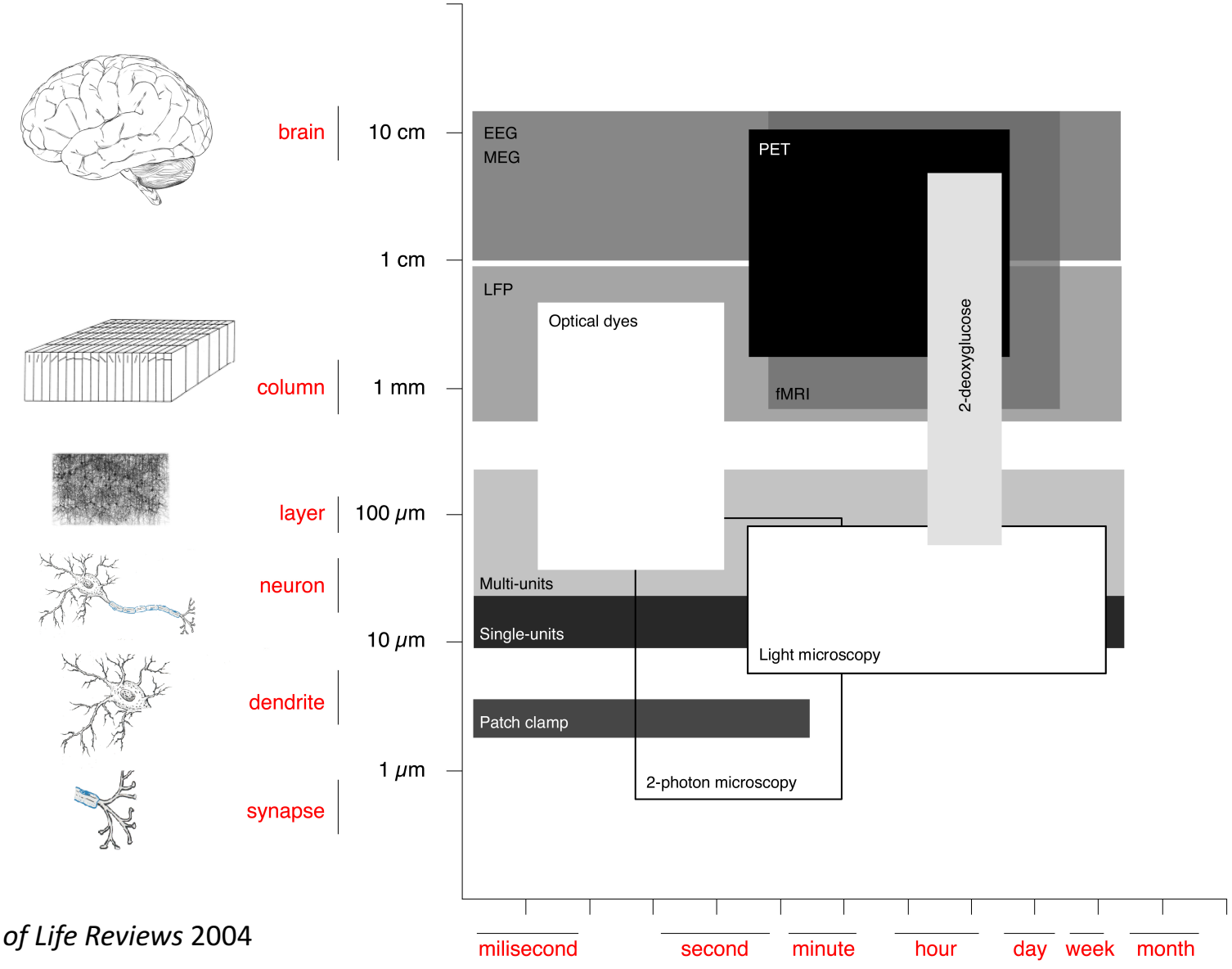


Figure S5: Dynamic evolution of the feature representation for RNN_h.

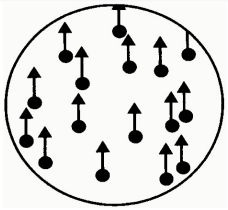
Using the same format and conventions from **Figure 6B**, this figure uses tSNE to visualize the dynamic evolution of the feature representation for the RNN_h model at 3 time points (t=0, t=16 and t=256). Over time, the representation of the partial images (filled circles) approach the correct category in the corresponding clusters for the whole images (empty circles).

Methods to study the brain at different scales

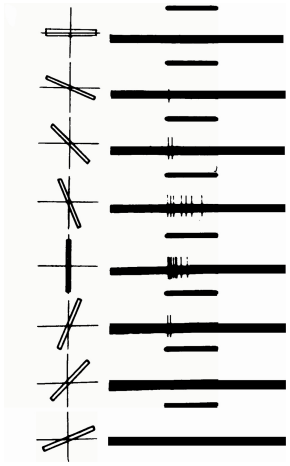


Kreiman. *Physics of Life Reviews* 2004

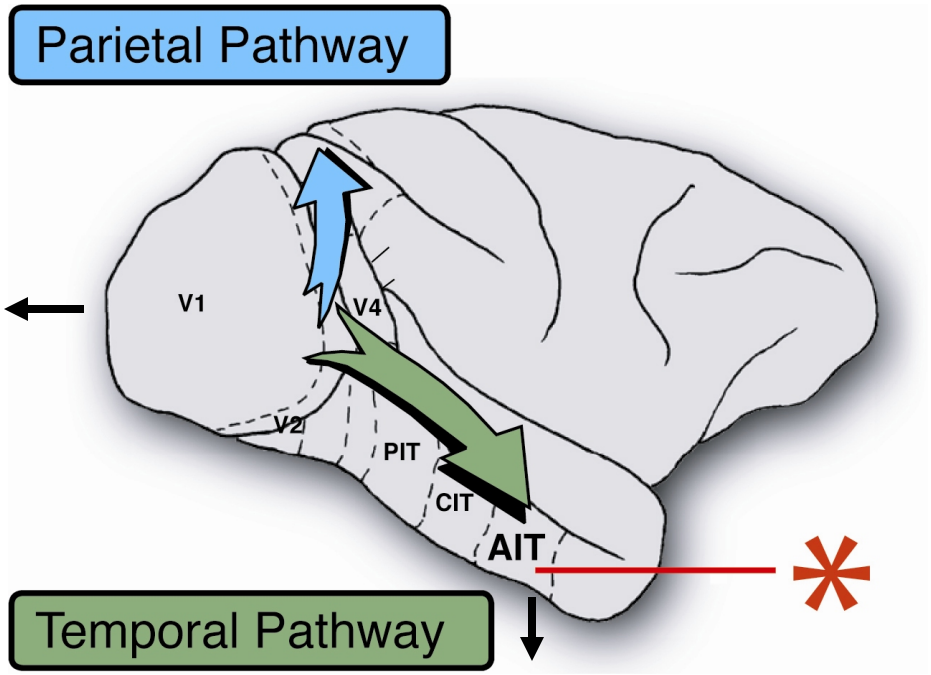
Peeking inside the brain

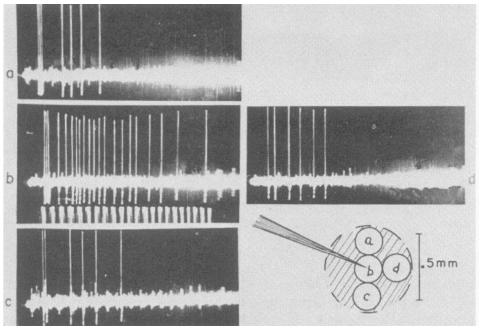


Newsome *et al* (1989)
Nature **341**:52-54

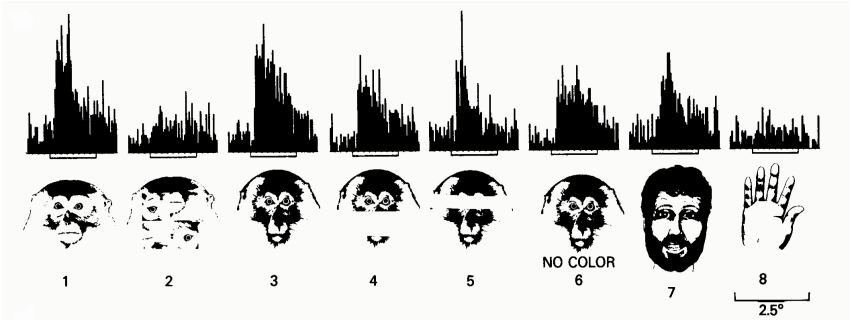


Hubel and Wiesel (1959) J. Physiol. **148**: 574-591





Kuffler, S. (1953)
J. Neurophys. **16**: 37-68

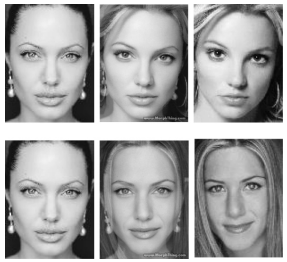


Desimone *et al* (1984)
*J. Neurosci. 4:*2051-2062

Key canons of visual recognition

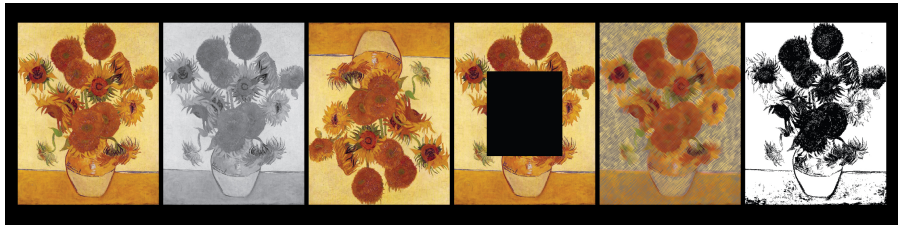
1. Selectivity

Similar,
yet distinct



ÒÓ =≈ ✚ ♩♪ ★

2. Tolerance (scale, position, etc.)

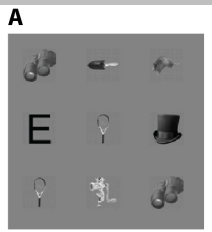


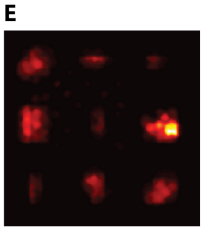
Different,
yet similar

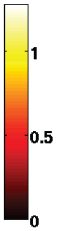
3. Speed (~100-200 ms)

4. Inference

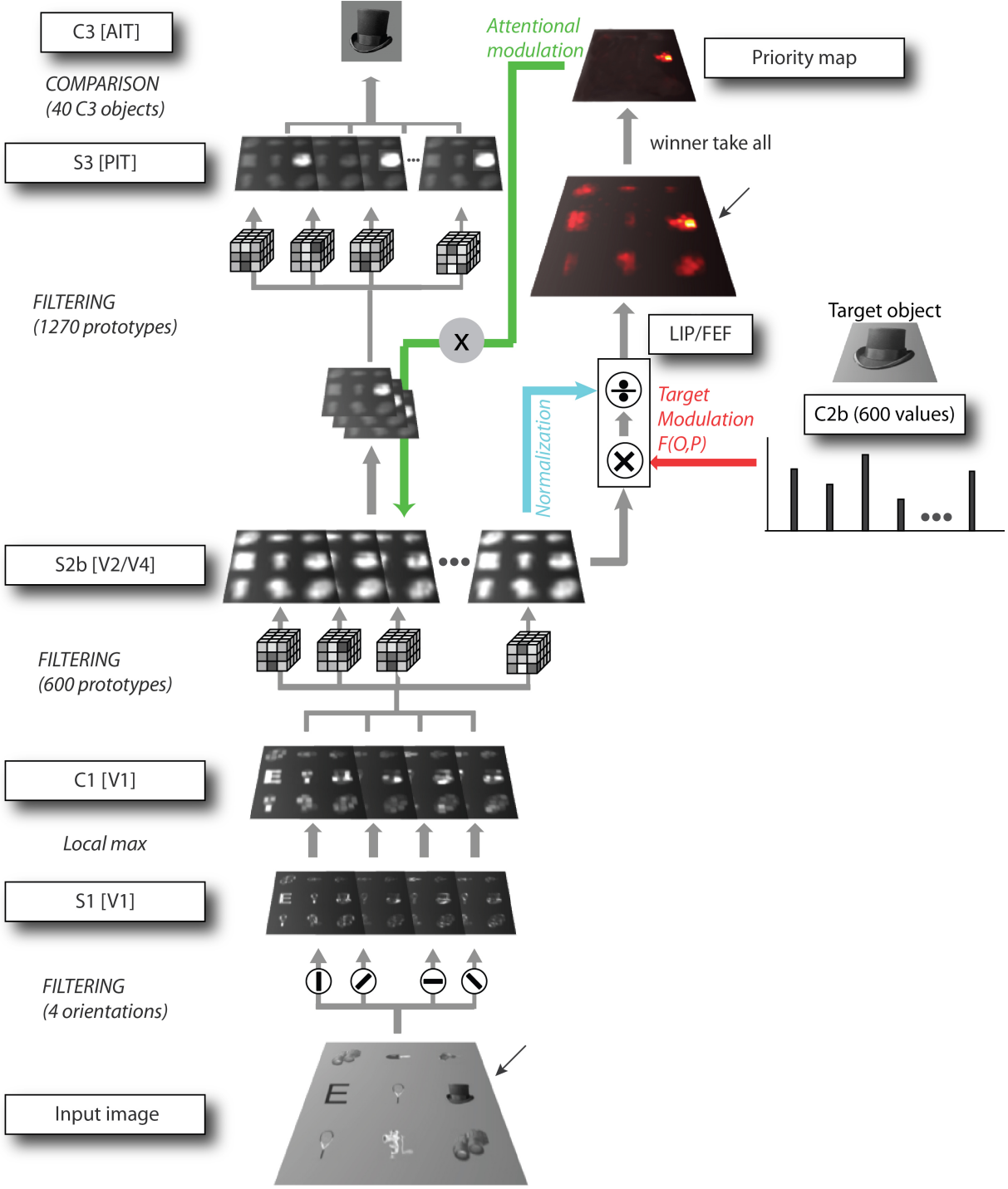
The model can find objects in cluttered images





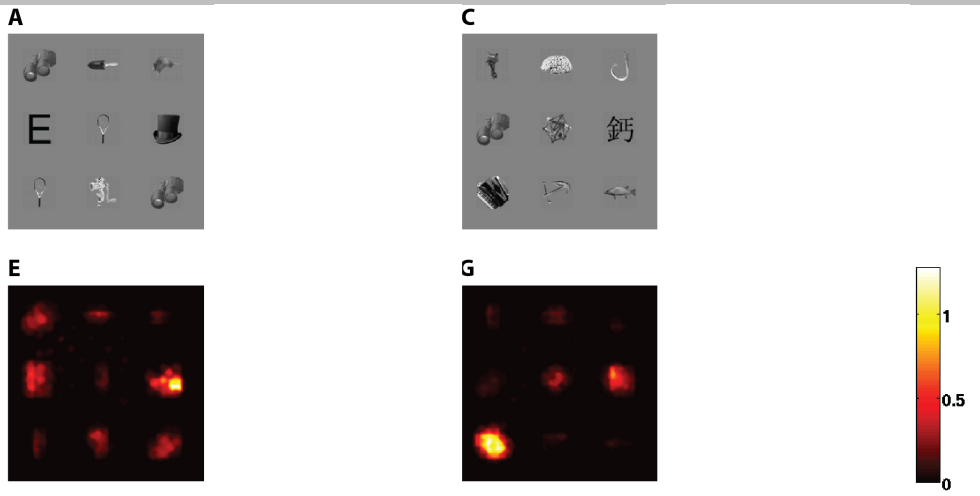


Feedback signals in visual search

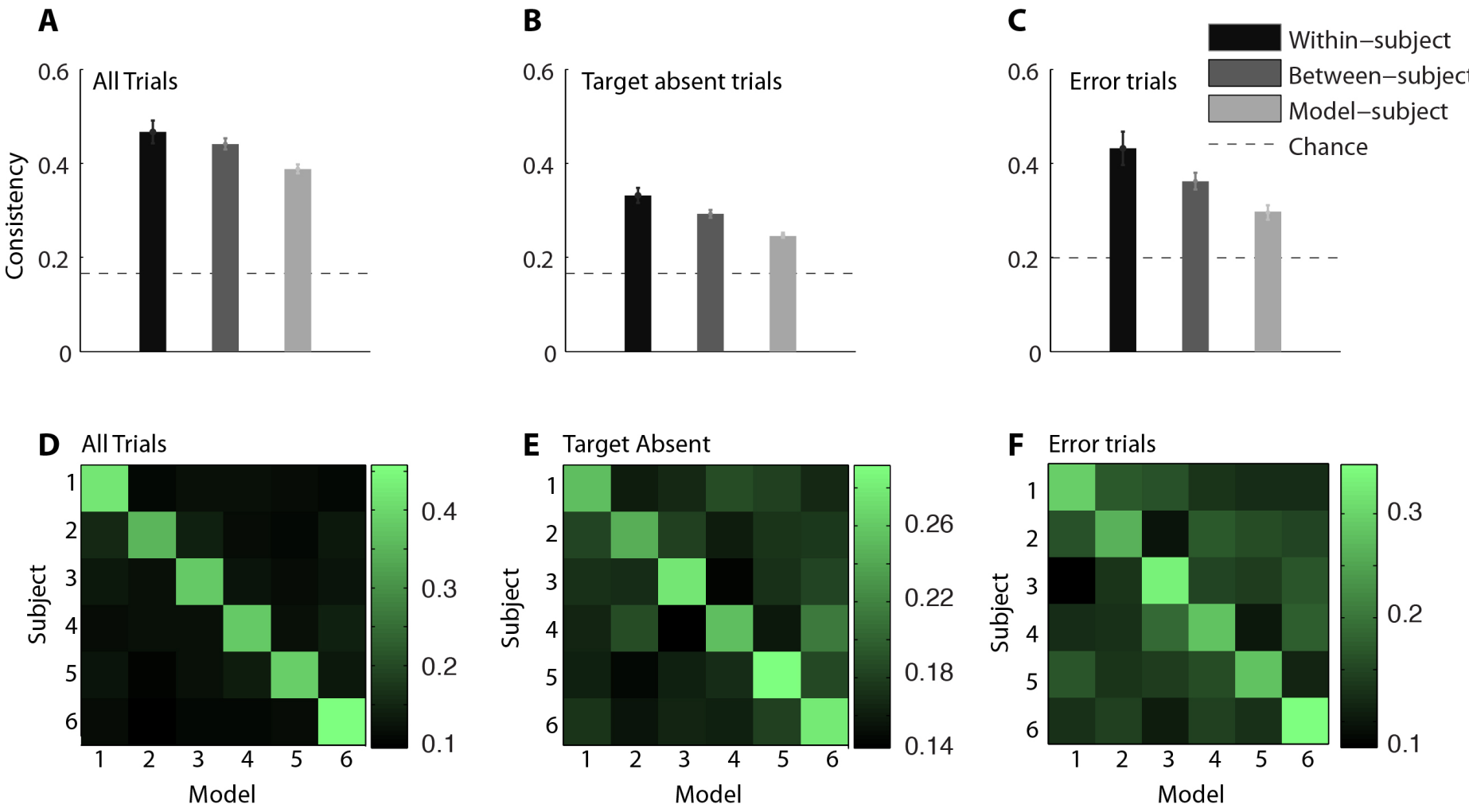


Miconi et al, 2016

The model can find objects in cluttered images



Further comparisons between humans and model

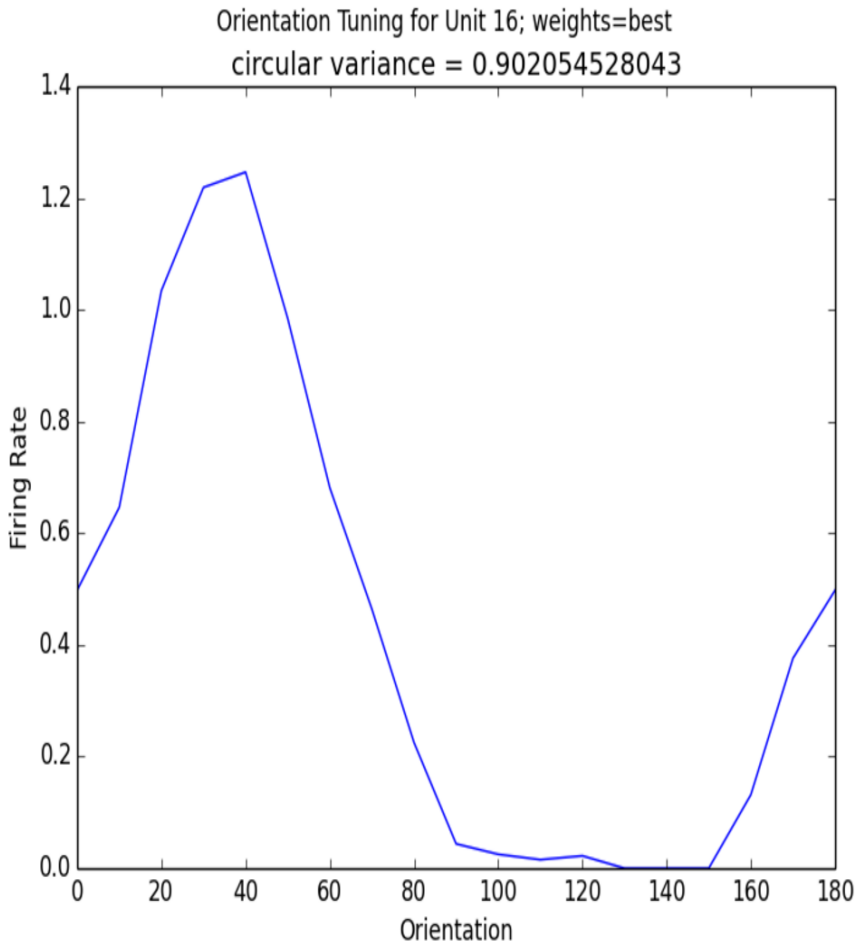


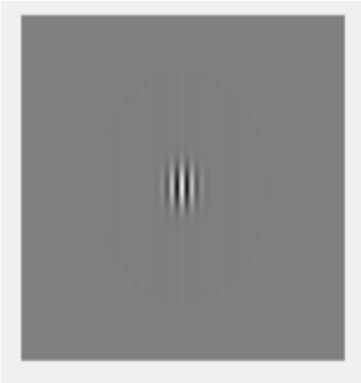
Computational roles of recurrent/feedback signals

1. Pattern completion (recurrent computations)
2. Predictive coding (feedback computations)

Image by Hanlin Tang

Looking at Gabor Tuning in Network



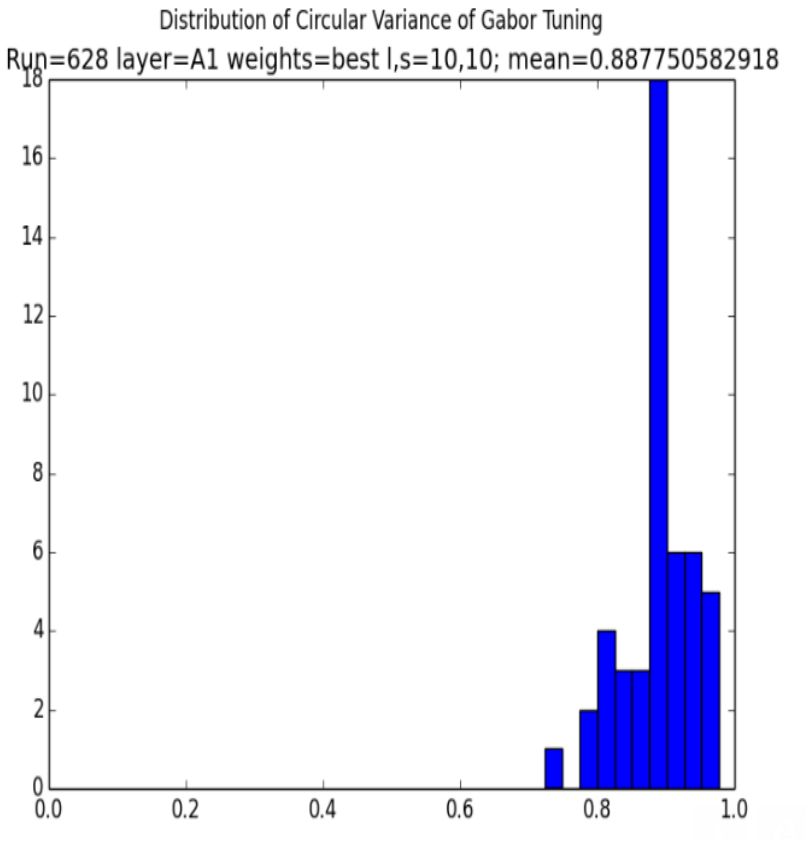


$$R = \frac{\sum_k r_k e^{i2\theta_k}}{\sum_k r_k},$$

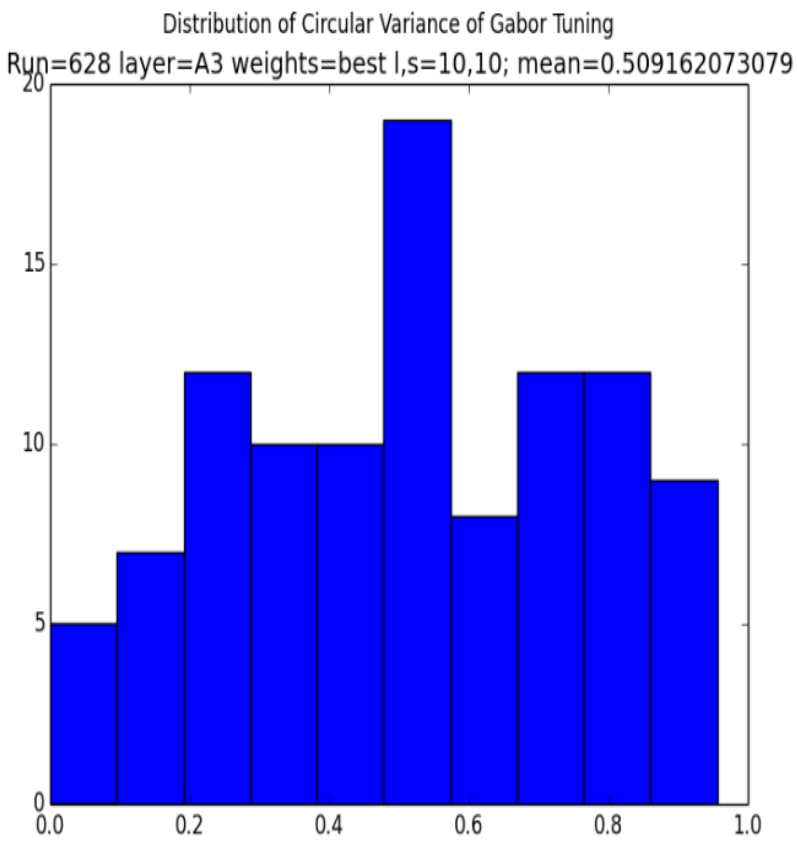
CV = 1 – R

r_k: firing rate at that angle

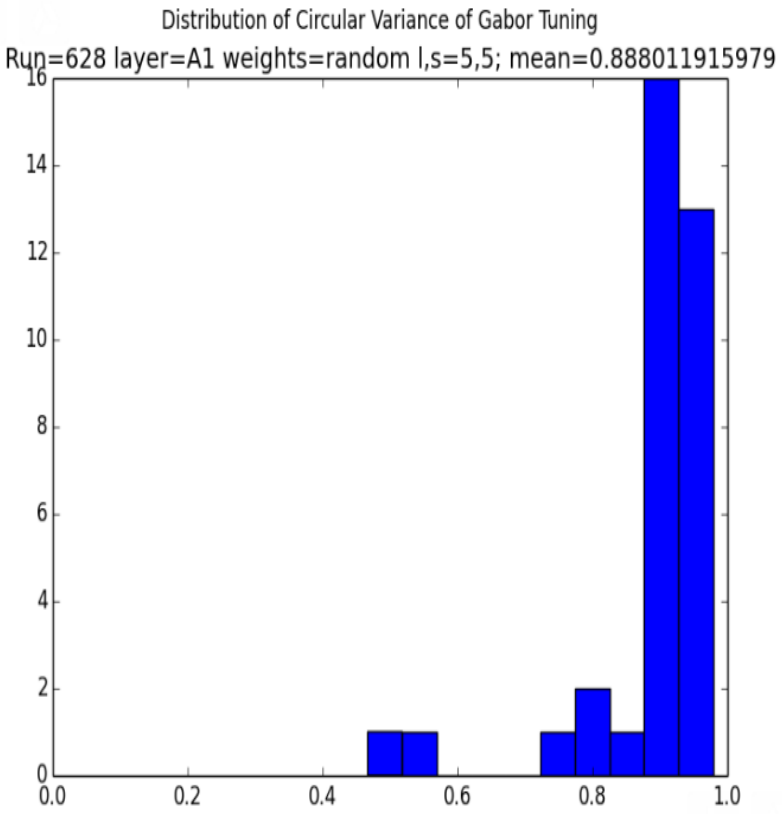
First Layer of Network

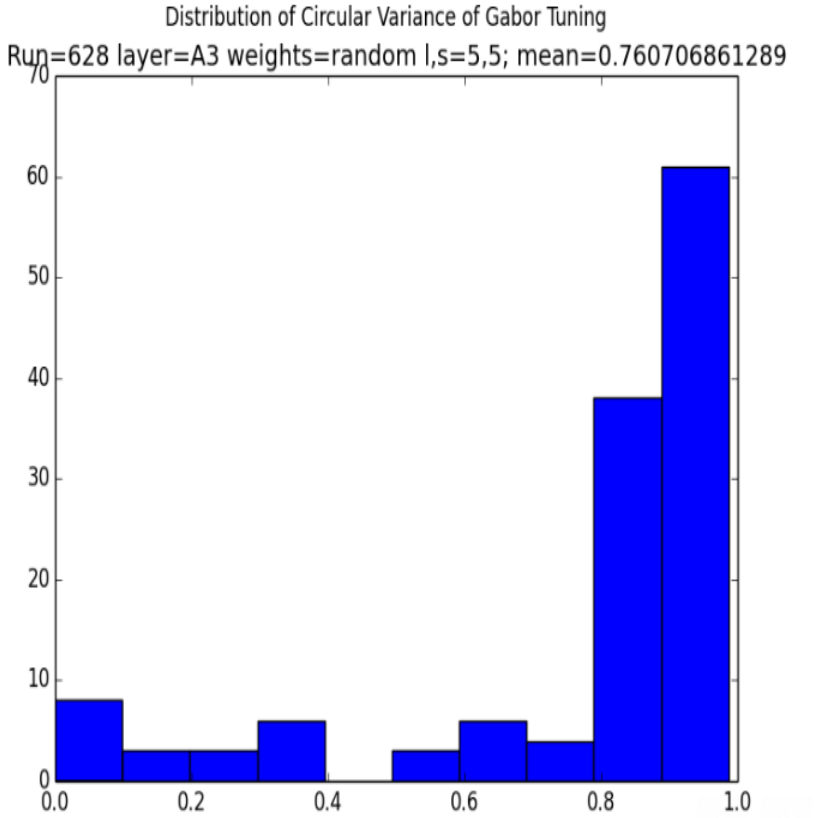


Highest Layer in Network



Compared to Random Initial Weights



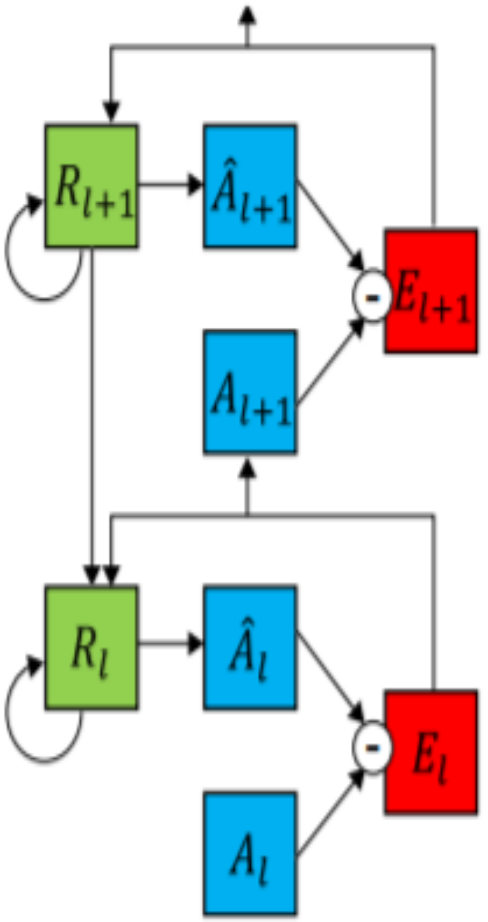


More Quantitative Evaluation

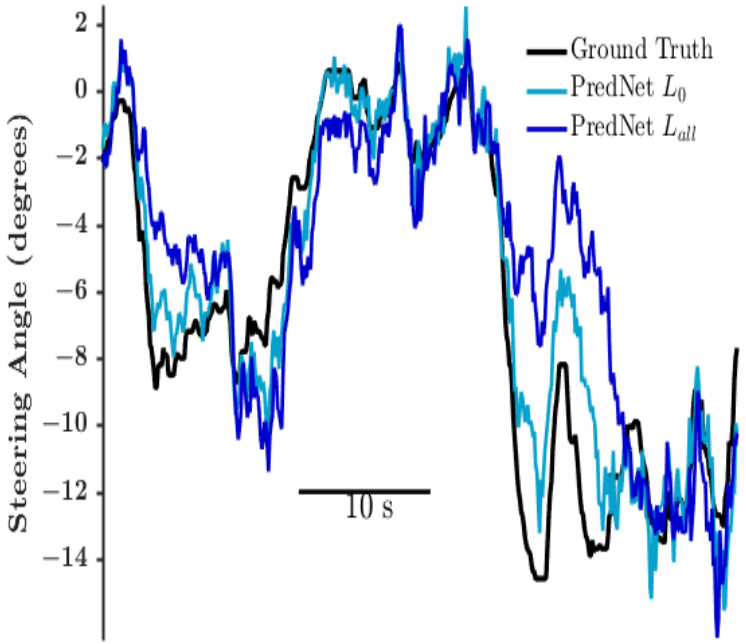
CalTech Pedestrian

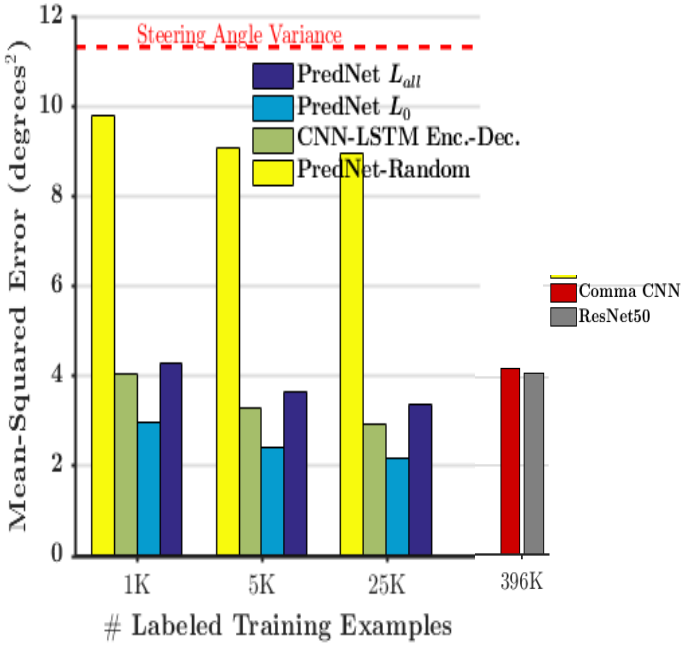
	MSE (x 10^{-3})	PSNR	SSIM
PredNet	**3.13 (3.33)**	**25.8 (25.5)**	**0.884 (0.878)**
PredNet (no E_l split)	3.20 (3.37)	25.6 (25.4)	0.883 (0.878)
CNN-LSTM Enc.-Dec.	3.67 (3.91)	25.0 (24.6)	0.865 (0.856)
CNN-LSTM Enc.-Dec. (2x A_l filts)	3.82 (3.97)	24.8 (24.6)	0.857 (0.853)
CNN-LSTM Enc.-Dec. (except pass E_0)	3.41 (3.61)	25.4 (25.1)	0.873 (0.866)
CNN-LSTM Enc.-Dec. (+/- split)	3.71 (3.84)	24.9 (24.7)	0.861 (0.857)
Copy Last Frame	7.95	20.0	0.762

- **PredNet (no E split)**: PredNet model except the error responses (E_l) are simply linear ($\hat{A}_l - A_l$) instead of being split into positive and negative rectifications.
- **CNN-LSTM Enc.-Dec. (2x A_l filts)**: CNN-LSTM Encoder-Decoder model (A_l's are passed instead of E_l's) except the number of filters in A_l is doubled. This controls for the total number of filters in the model compared to the PredNet, since the PredNet has filters to produce $\hat{A}_l$ at each layer, which is integrated into the model's feedforward response.
- **CNN-LSTM Enc.-Dec. (except pass E_0)**: CNN-LSTM Encoder-Decoder model except the error is passed at the lowest layer. All remaining layers pass the activations A_l. With training loss taken at only the lowest layer, this variation allows us to determine if the "prediction" subtraction operation in upper layers, which is essentially unconstrained and learnable in the L_0 case, aids in the model's performance.
- **CNN-LSTM Enc.-Dec. (+/- split)**: CNN-LSTM Encoder-Decoder model except the activations A_l are split into positive and negative populations before being passed to other layers in the network. This isolates the effect of the additional nonlinearity introduced by this procedure.

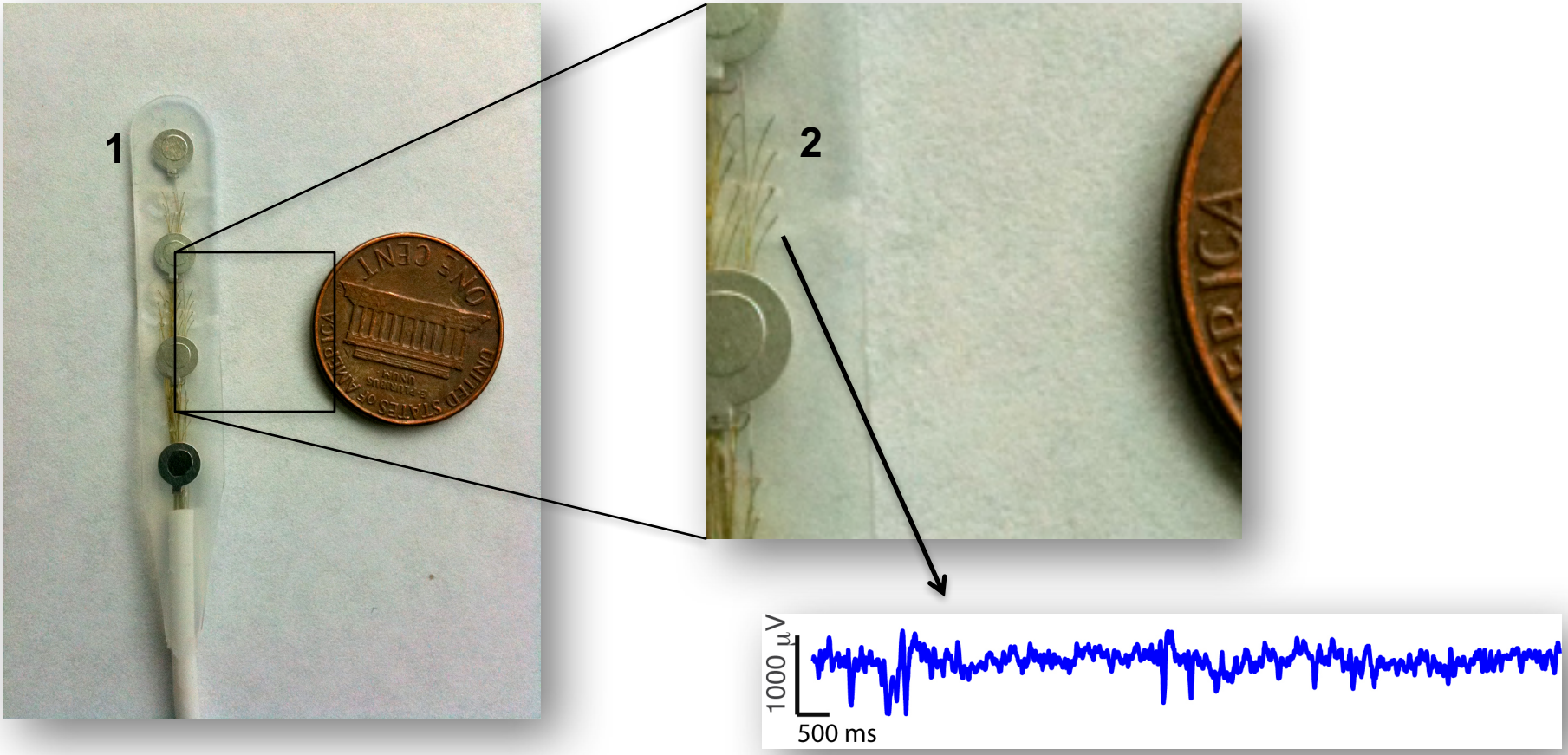


Assessing Representation: Decoding Steering Angle





Electrode types for invasive recordings



1 → Low impedance macro contacts (< 1 kΩ)

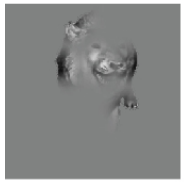
2 → High impedance microwires (~ 1MOhm)

•Subdural (occipito-temporal cortex, frontal cortex)

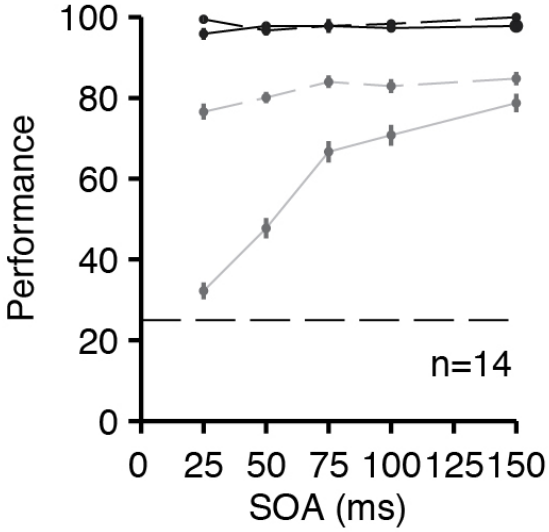
•Large coverage

Performance for occluded objects

A

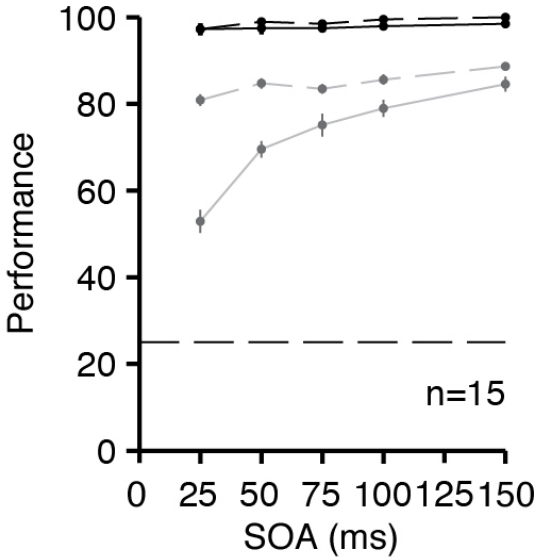


Partial

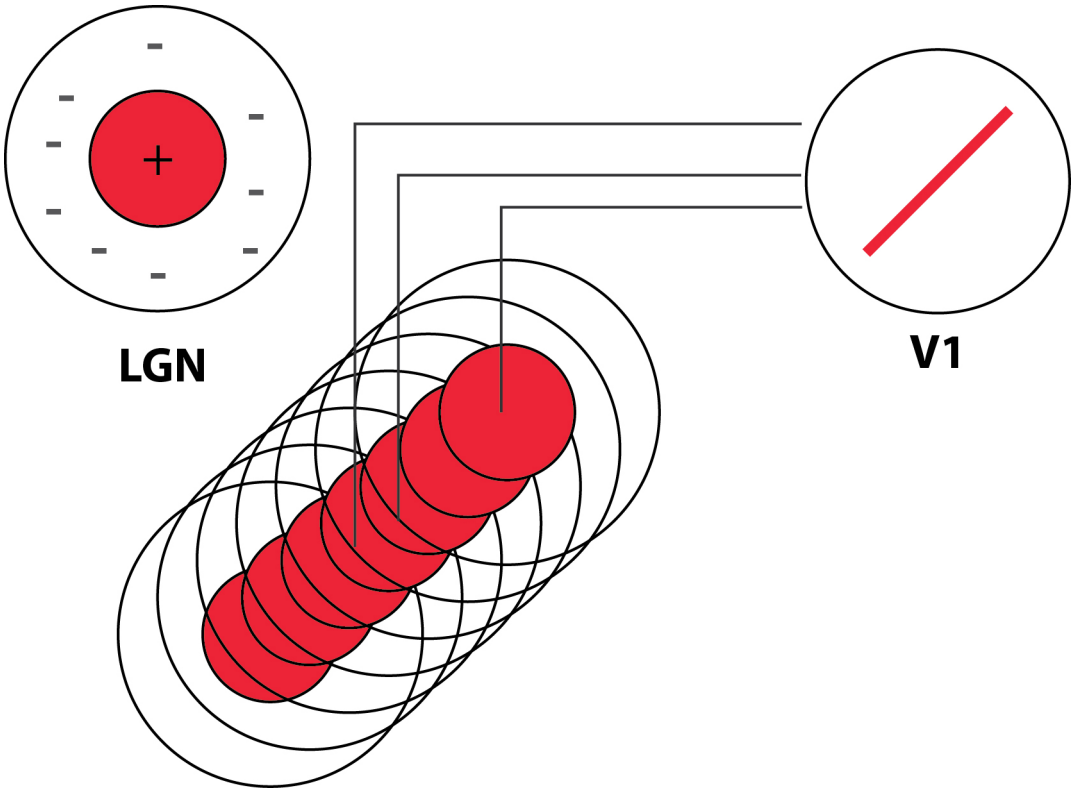


B

Occluded

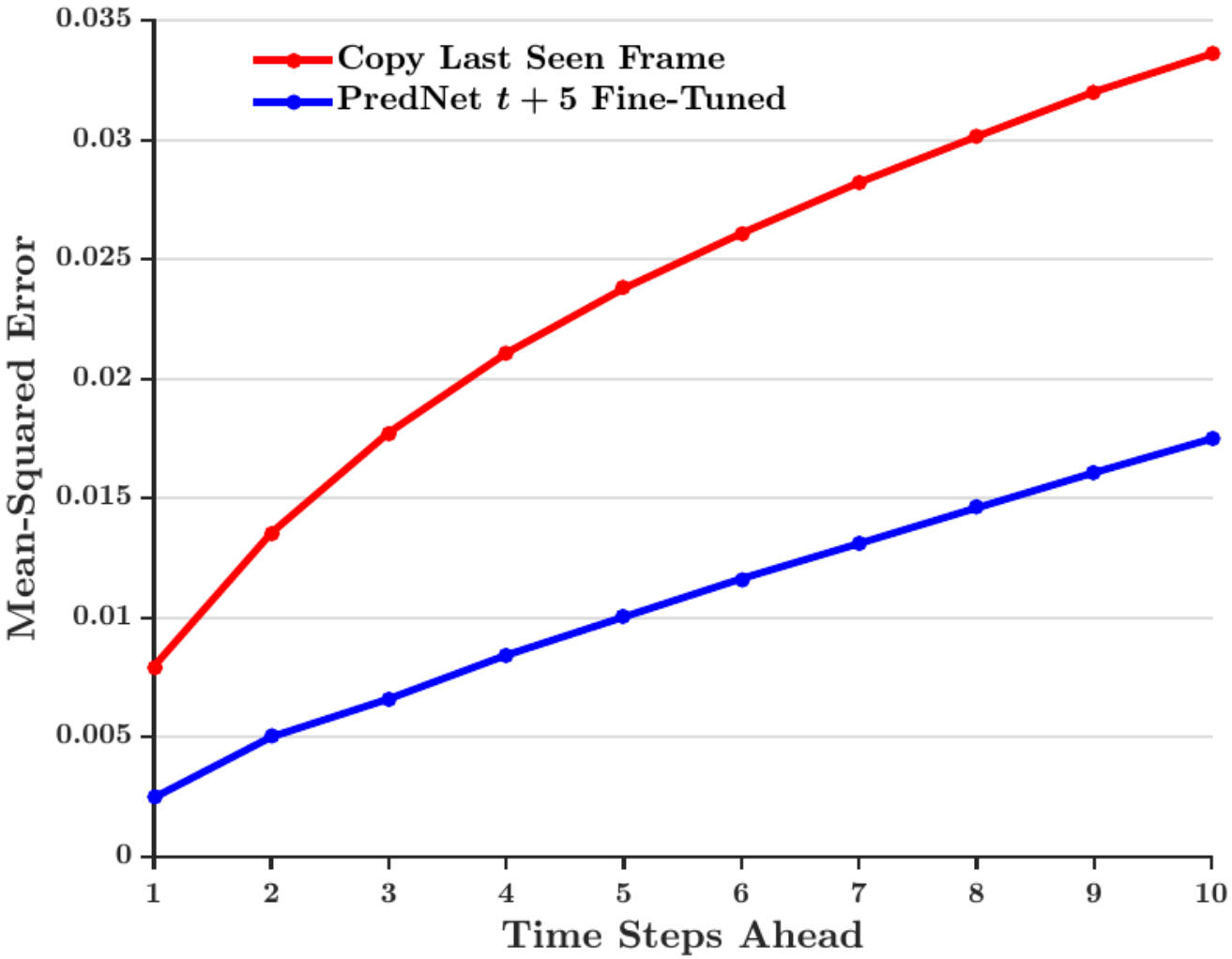


A simple model for simple cells

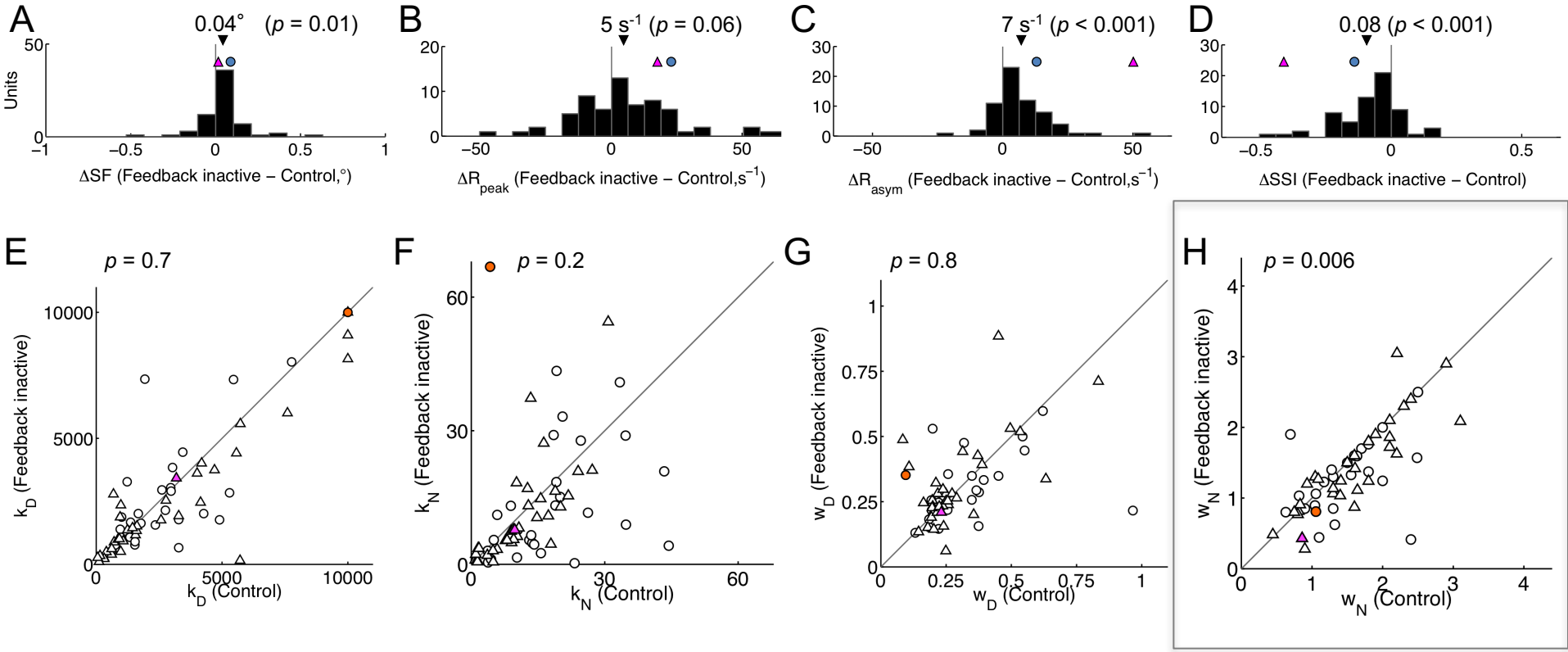


A feed-forward model for orientation selectivity in V1
(by no means the only model)

Multiple Time Step Prediction



Feedback increases the normalization width: w_N



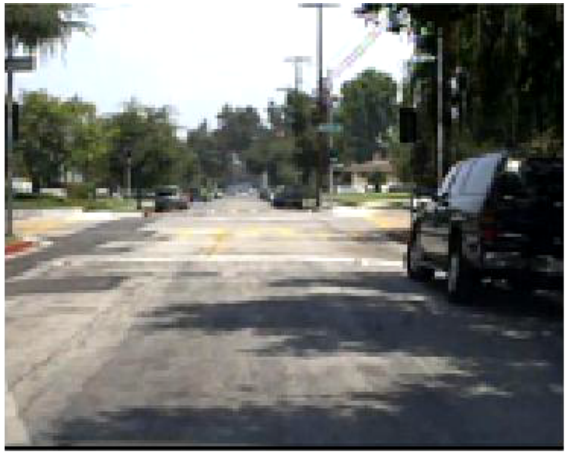
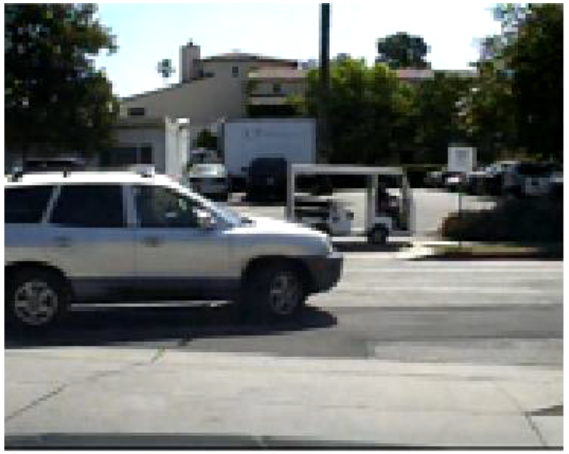
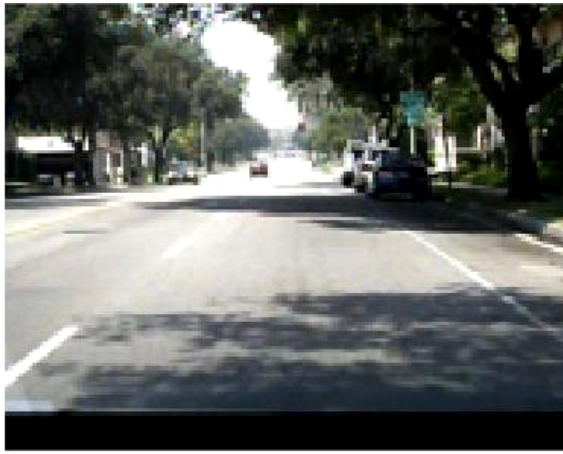
Calculation of PredNet states

Algorithm 1 Calculation of PredNet states

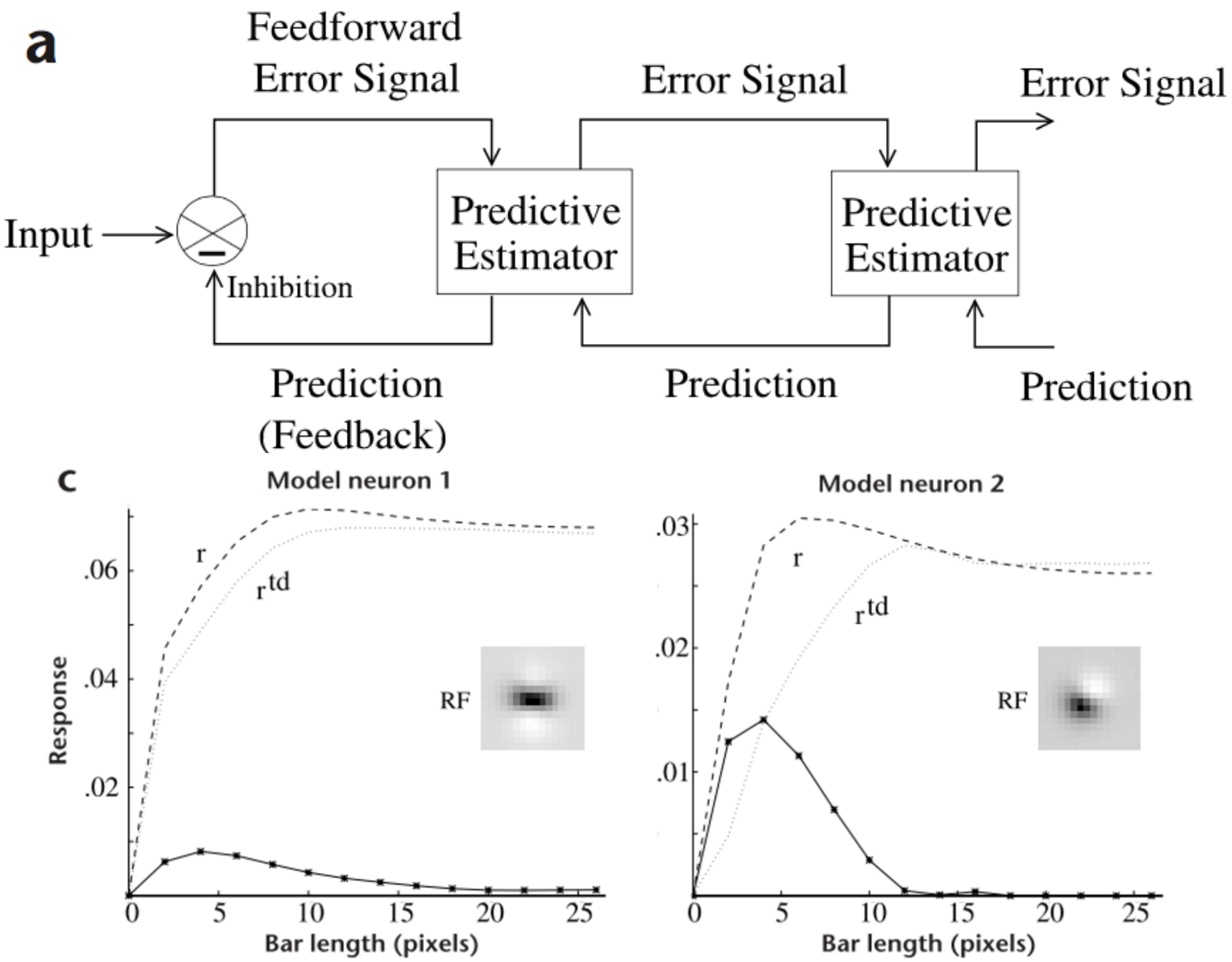
Require: x_t

$A_0^t \leftarrow x_t$
$E_l^0, R_l^0 \leftarrow 0$
for $t = 1$ **to** T **do**
  for $l = L$ **to** 0 **do** ▷ Update R_l^t states
    if $l = L$ **then**
      $R_L^t = \text{CONVLSTM}(E_L^{t-1}, R_L^{t-1})$
    else
      $R_l^t = \text{CONVLSTM}(E_l^{t-1}, R_l^{t-1}, \text{UPSAMPLE}(R_{l+1}^t))$
  for $l = 0$ **to** L **do** ▷ Update $\hat{A}_l^t, A_l^t, E_l^t$ states
    if $l = 0$ **then**
      $\hat{A}_0^t = \text{SATLU}(\text{RELU}(\text{CONV}(R_0^t)))$
    else
      $\hat{A}_l^t = \text{RELU}(\text{CONV}(R_l^t))$
    $E_l^t = [\text{RELU}(A_l^t - \hat{A}_l^t); \text{RELU}(\hat{A}_l^t - A_t^l)]$
    if $l < L$ **then**
      $A_{l+1}^t = \text{MAXPOOL}(\text{CONV}(E_t^l))$

Multiple Time Step Prediction

Predictive coding in visual cortex



Rao and Ballard, 1999