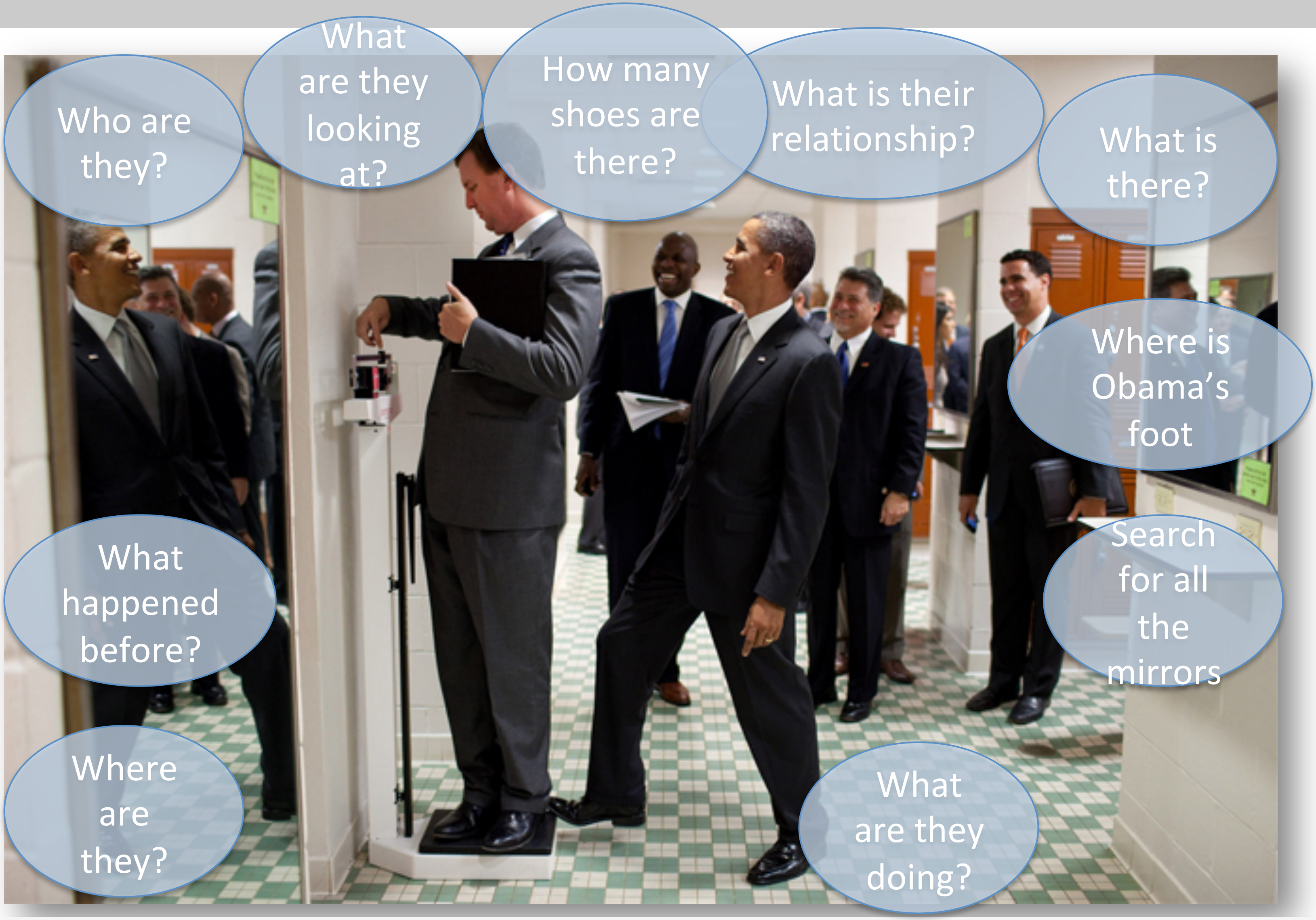
http://klab.tch.harvard.edu

The brain's operating system

Gabriel Kreiman

An image is worth a million words



Many apps: clinical image understanding, security, self-driving vehicles, intelligent image search, automatic video interpretation, … UNDERSTANDING BRAIN COMPUTATIONS!

Caption bots: not too bad, not too good

I am not really confident, but I think it's a group of people standing next to person in a suit and tie.

How did I do?

Visual cognition: a sequence of routines*

Divide et impera

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 Routines (Shimon Ullman)

Visual cognition: a sequence of routines* and subroutines

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`

* Visual Routines (Shimon Ullman)

Visual cognition: a sequence of routines*

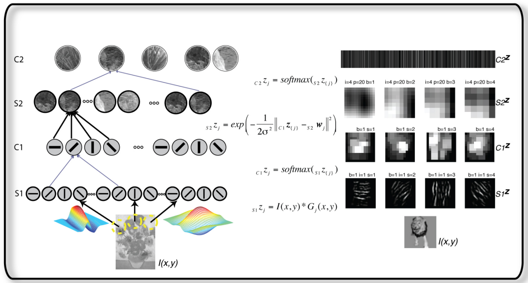
Divide et impera

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`

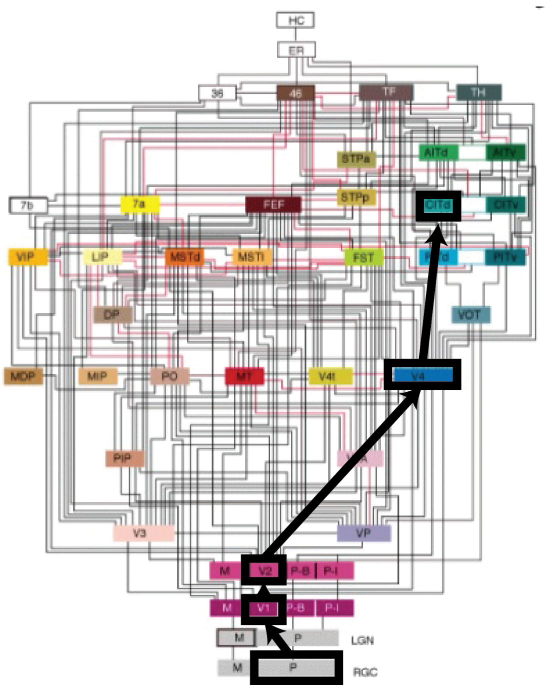
* Visual Routines (Shimon Ullman)

Deep convolutional networks achieve high performance in image labeling tasks

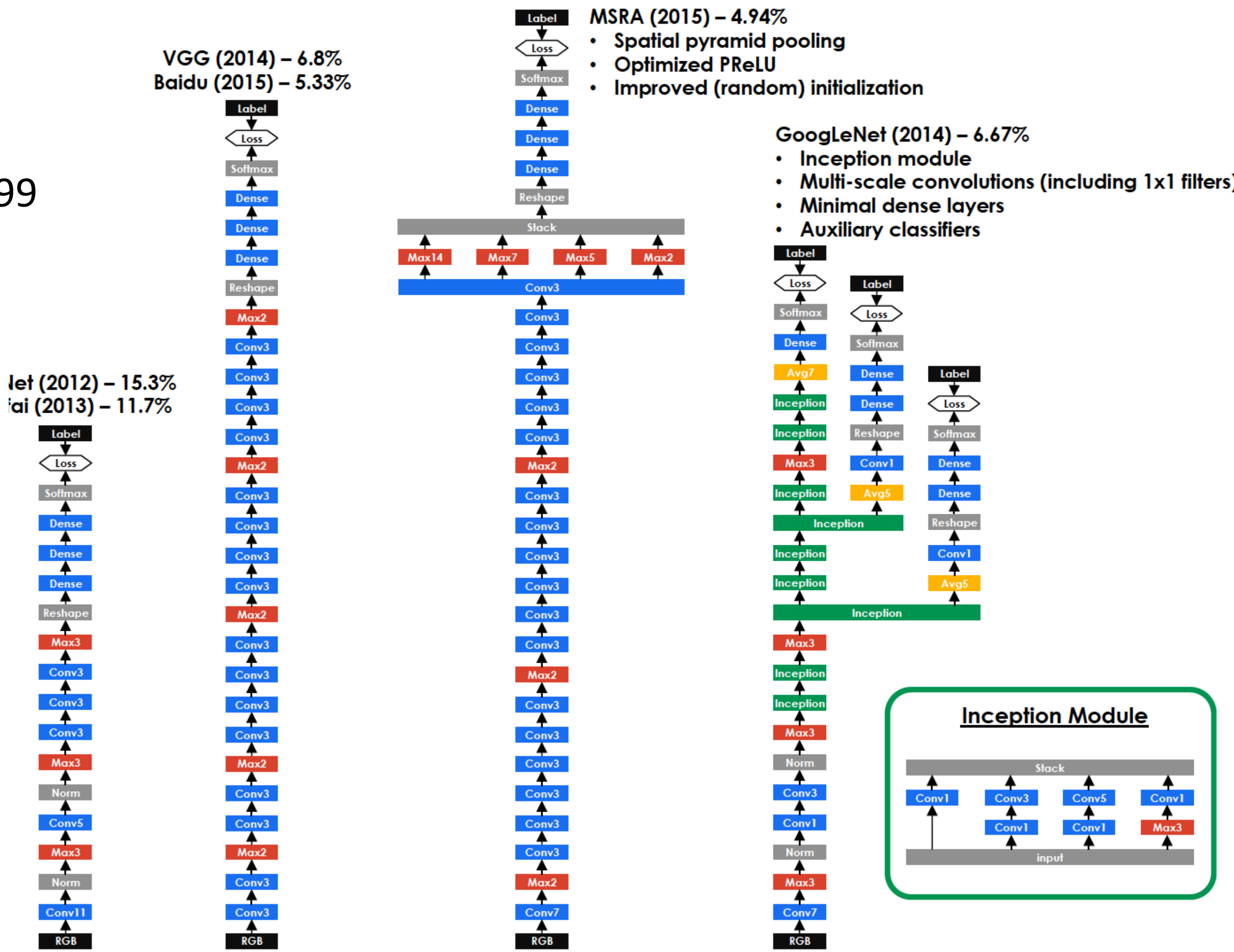
A first-order approximation to ventral visual cortex function during initial ~150 ms of processing



Riesenhuber and Poggio 1999
Serre et al 2007



Felleman and
Van Essen 1991



Marr-Poggio's three levels of explanation

1. Computational: → [Psychophysics] *What the problem is and how well animals solve it*

2. Algorithmic: → [Model] *Plausible sequence of operations to solve the problem*

3. Implementation: → [Neurophysiology] *Biological mechanisms by which animals solve the problem*

Visual cognition: a sequence of routines*

Divide et impera

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? →
14. If satisfactory, answer the question →

Mengmi Zhang Martin Schrimpf Eric Wu

High-resolution fovea, low-resolution periphery

Context example 1

Demo: eccentricity-dependent changes in resolution

Learning Scene Gist with Convolutional Neural Networks to Improve Object Recognition

Kevin Wu*
Computational Science and Engineering,
Harvard University
kevin_wu@g.harvard.edu

Eric Wu*
Computational Science and Engineering,
Harvard University
eric_wu@g.harvard.edu

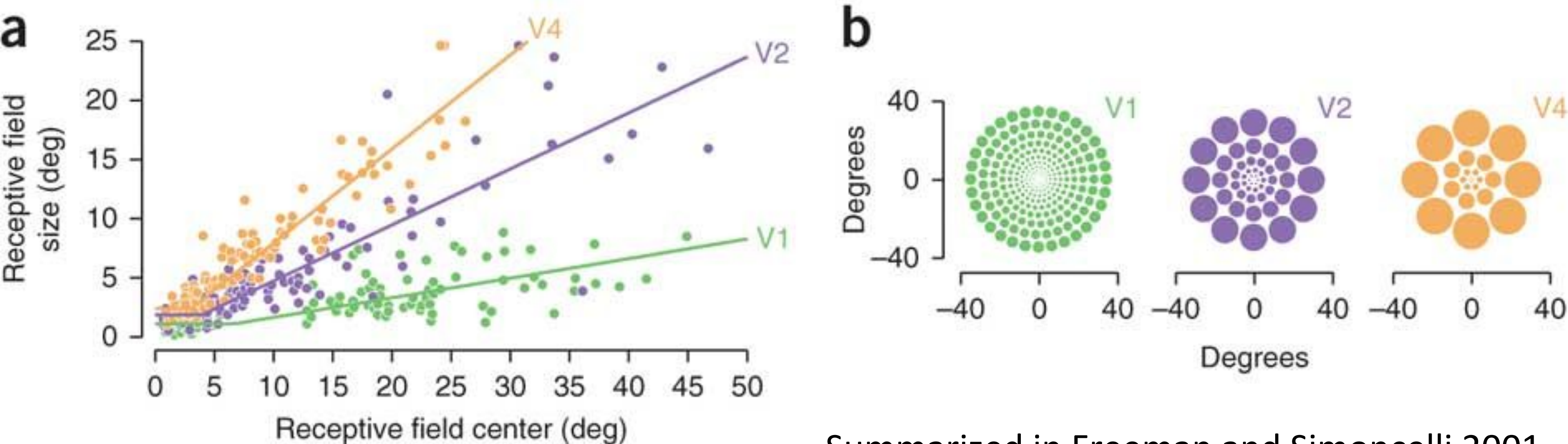
Gabriel Kreiman
Children's Hospital, Harvard Medical School
gabriel.kreiman@tch.harvard.edu

***Abstract*—Advancements in convolutional neural networks (CNNs) have made significant strides toward achieving high performance levels on multiple object recognition tasks. While some approaches utilize information from the entire scene to propose regions of interest, the task of interpreting a particular region or object is still performed independently of other objects and features in the image. Here we demonstrate that a scene's 'gist' can significantly contribute to how well humans can recognize objects. These findings are consistent with the notion that humans foveate on an object and incorporate information from the periphery to aid in recognition. We use a biologically inspired two-part convolutional neural network ('GistNet') that models the fovea and periphery to provide a proof-of-principle**

interplay between foveal and peripheral information may enable faster recognition of objects within a scene with a significantly reduced number of cells.

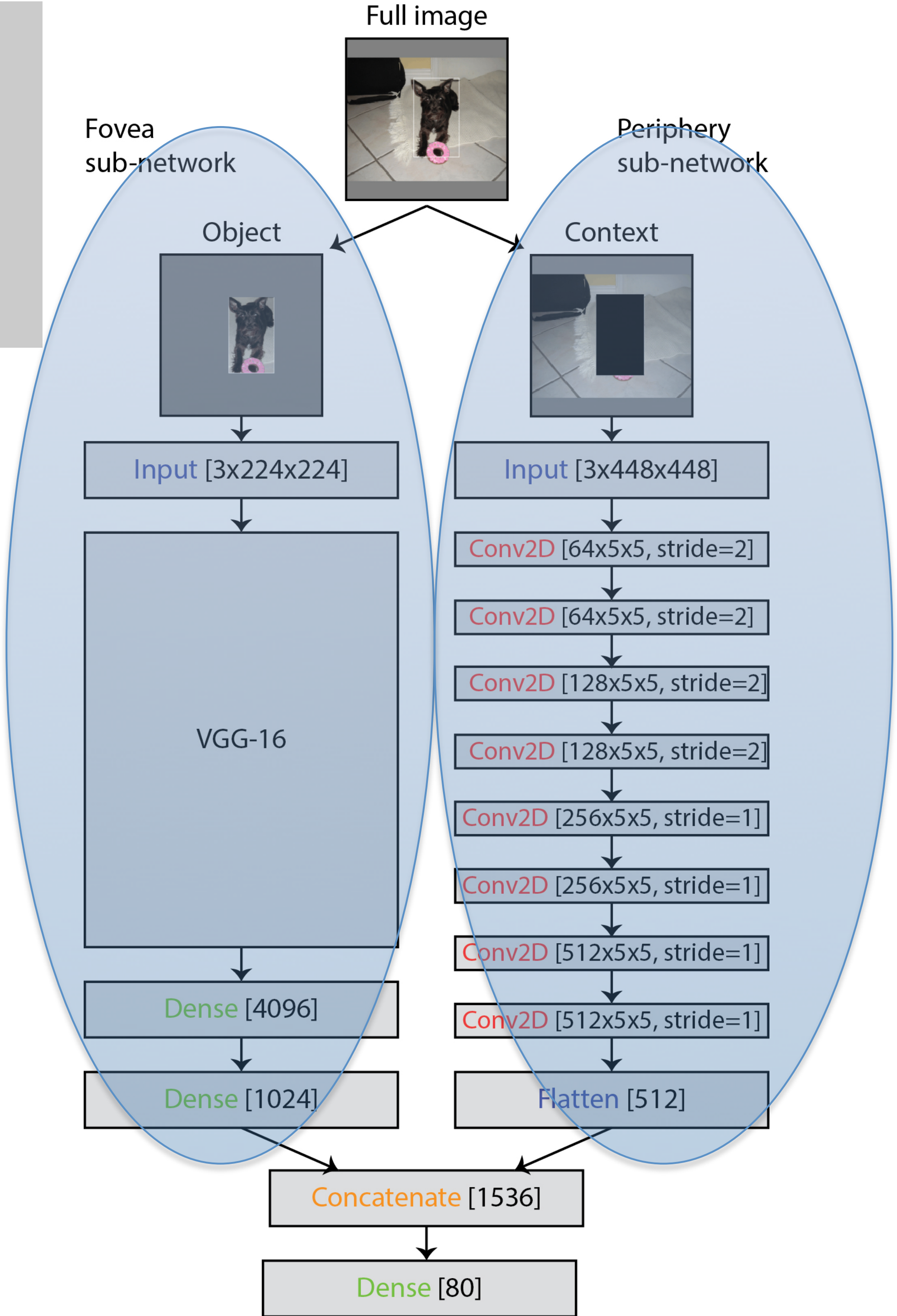
State-of-the-art computer vision architectures like Mask R-CNN [8] mirror elements of active sampling via sequential foveation by creating region proposals on the image, followed by object recognition in each region. Those region proposals cut down on the cost of having to perform classifications on the entire image. Yet, these models lack critical components of contextual information provided by interactions between the fovea and the periphery which are characteristic of human

Eccentricity-dependent receptive field sizes



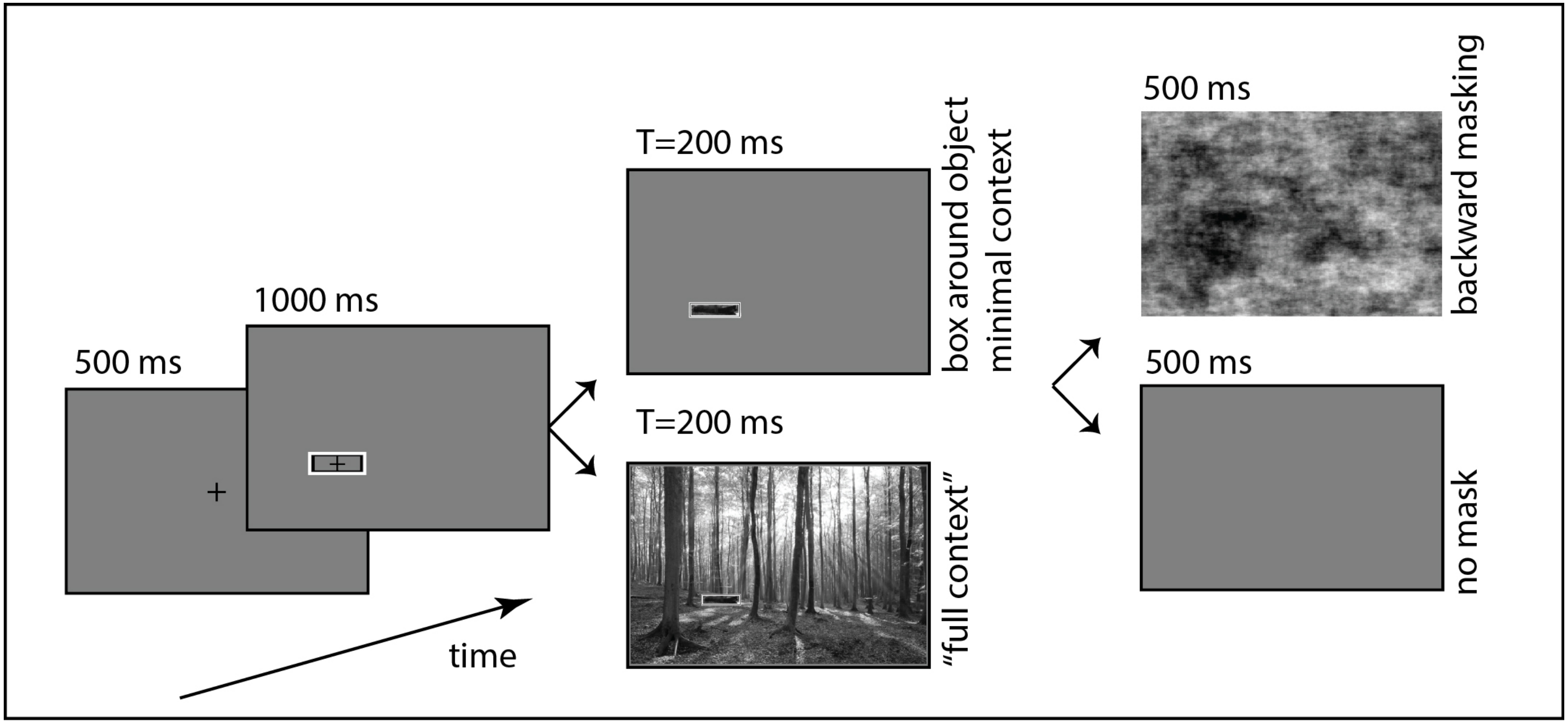
Summarized in Freeman and Simoncelli 2001

GistNet: Fovea+Periphery subnetwork

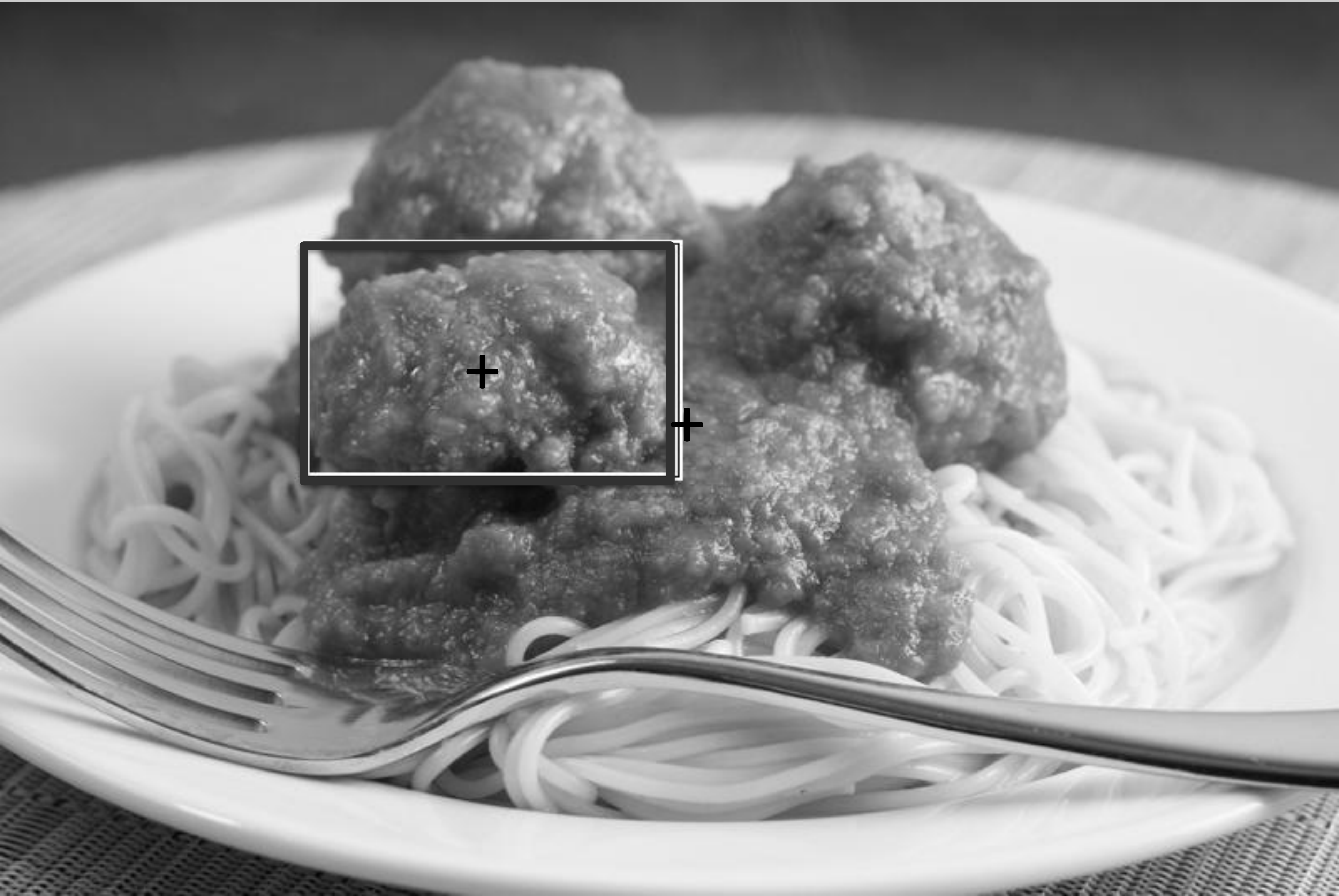


Wu et al, 2018

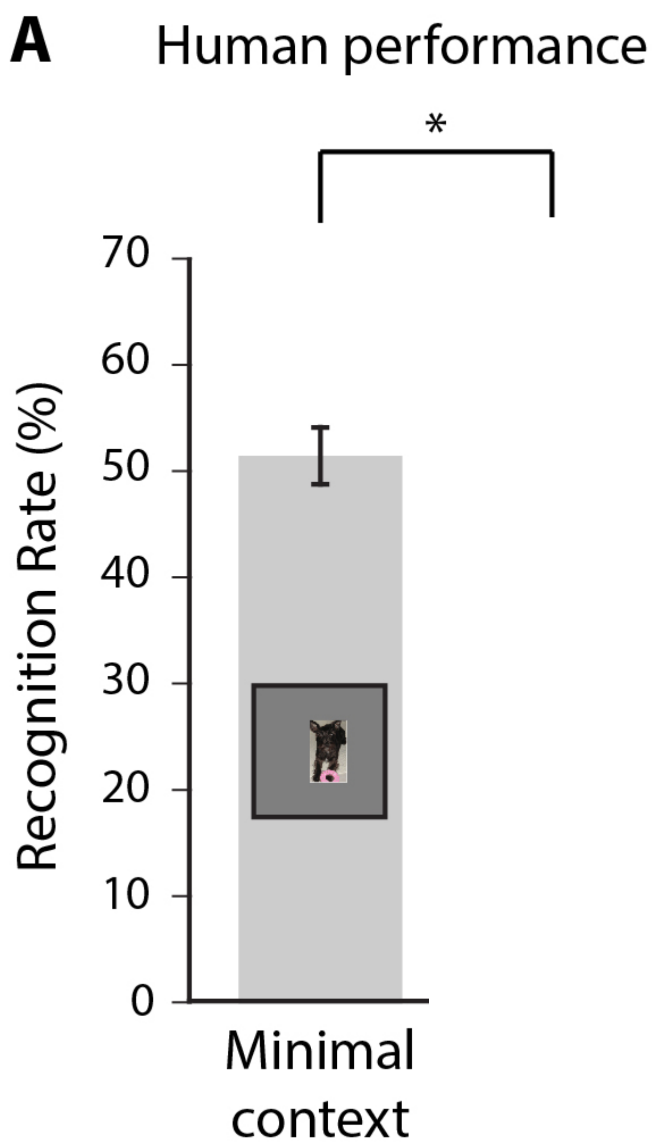
Contextual gist: Experiment setup



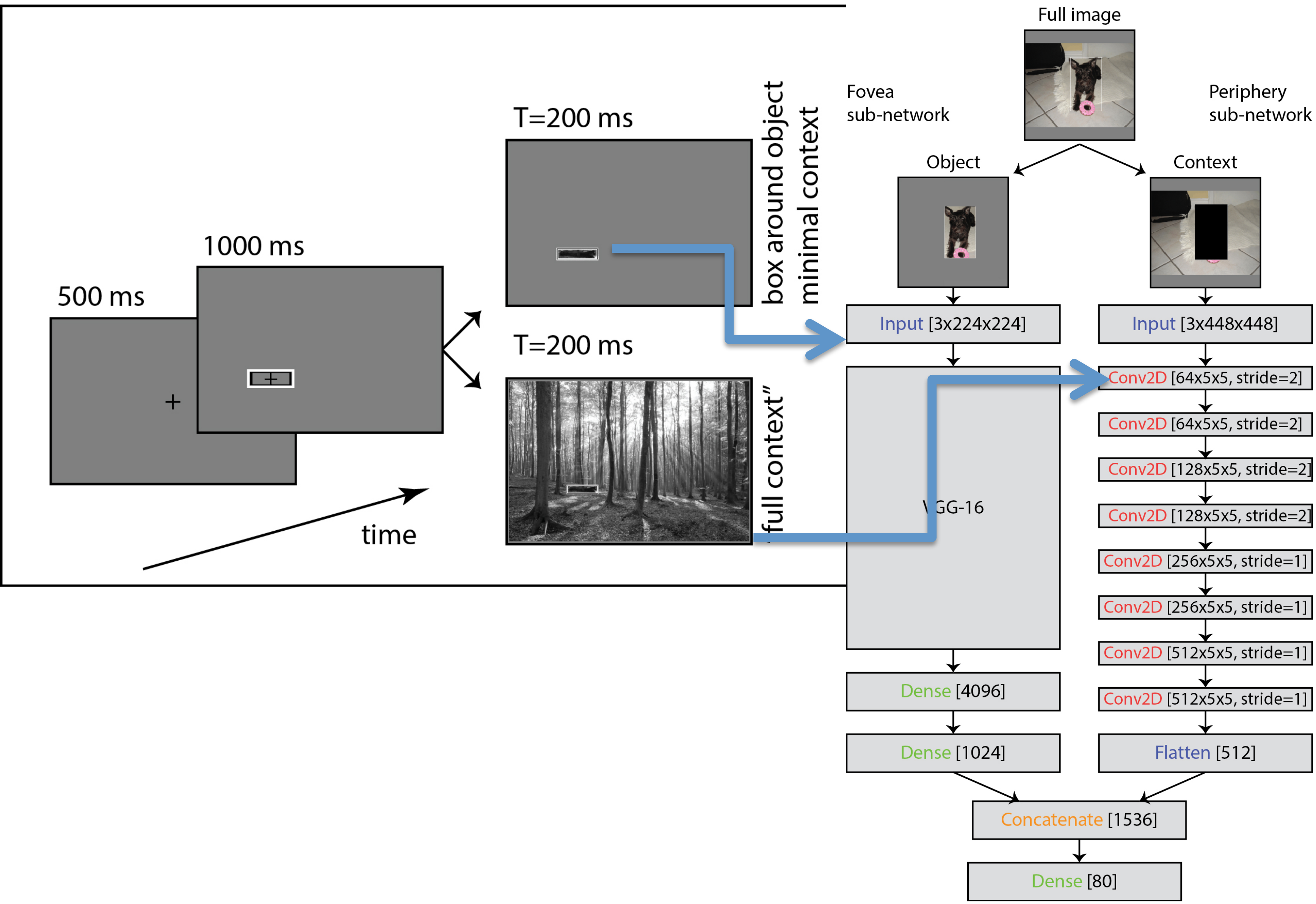
Context example 2



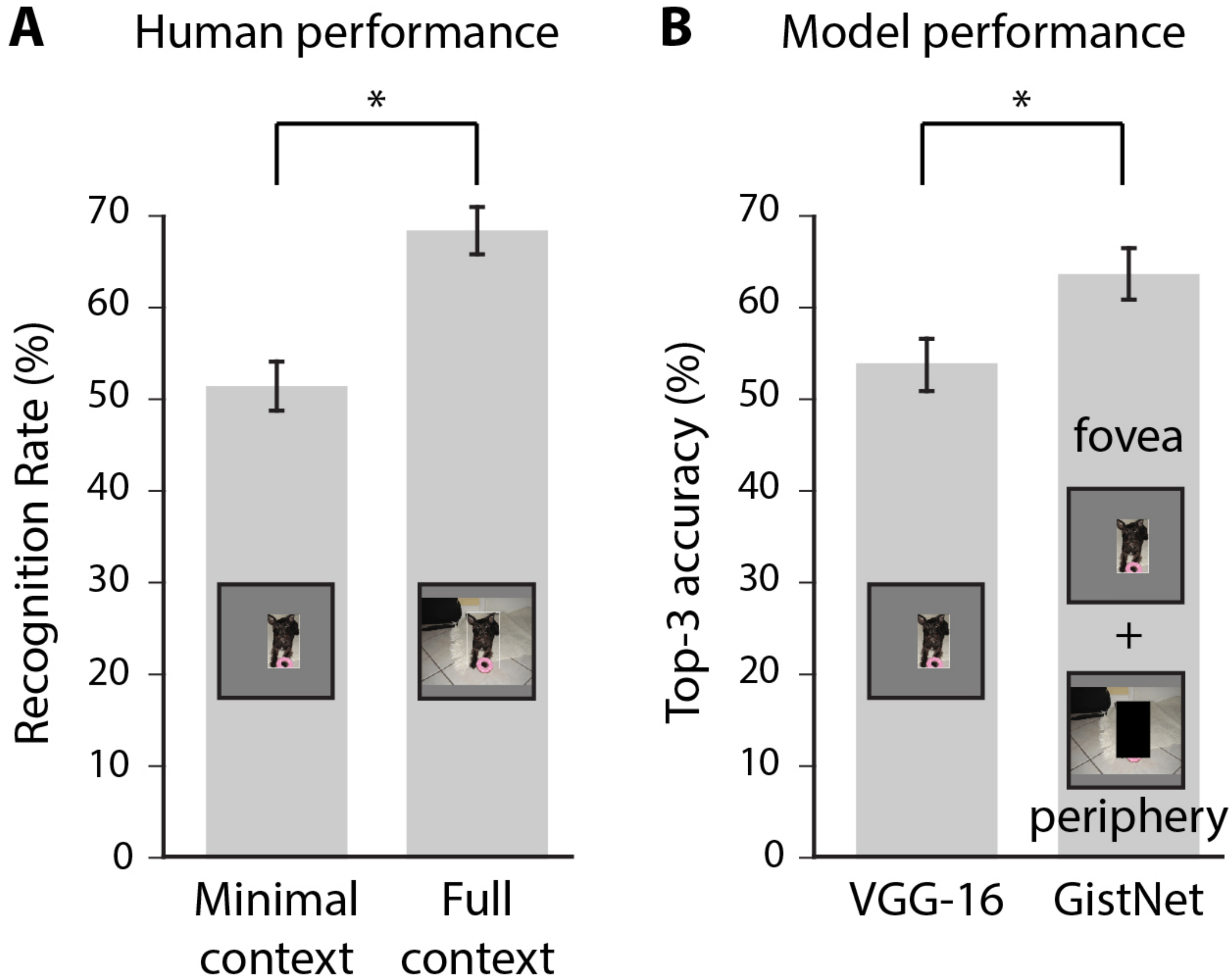
Spatial context improves object recognition



Contextual gist: Experiment setup



Spatial context improves object recognition



Interim summary 1

(Spatial) contextual information can help visual object recognition

First order effect:

- Rapid [effects observed with ~100 ms exposure]
- Low-resolution [blurred context helps too]
- Gist-like information [initial effects do not require detailed object identification]
- Bottom-up [can be approximated by simple bottom-up model]

There is much more to context:

Neurophysiological mechanisms
High-level statistical regularities
Temporal context
Multiple fixations

Visual cognition: a sequence of routines*

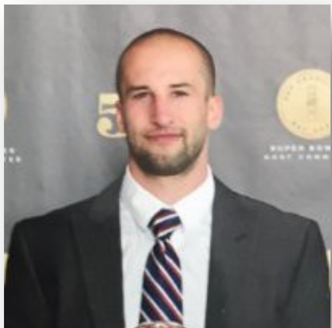
Divide et impera

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 Routines (Shimon Ullman)

Deep Learning Implementation of Predictive Coding

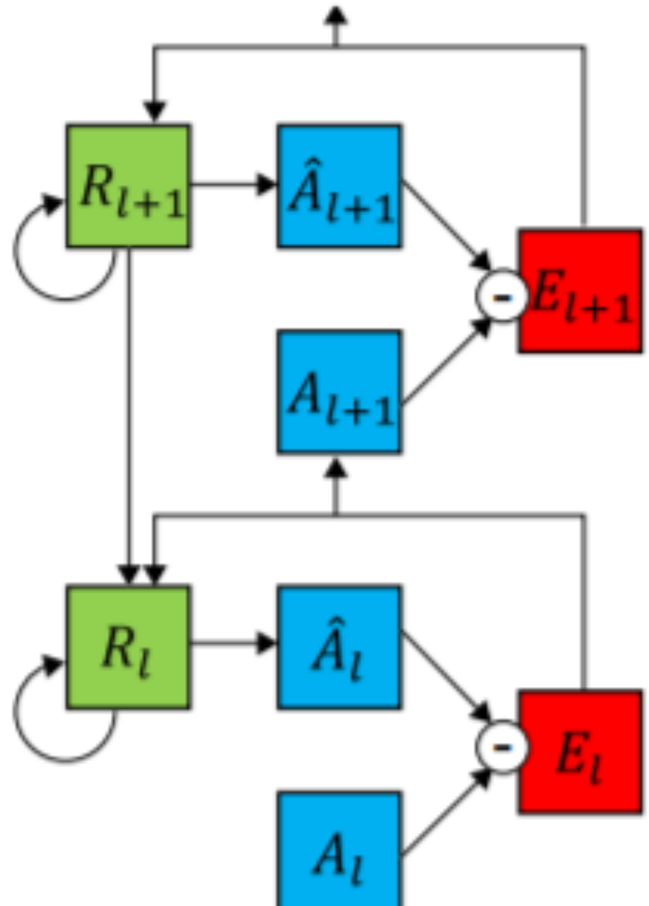
Bill Lotter



David Cox

Essential elements

- "**Representatio**
  neurons: hold '
  world"
- **Predictions**
- **Targets**
- "**Error**" neuron



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)

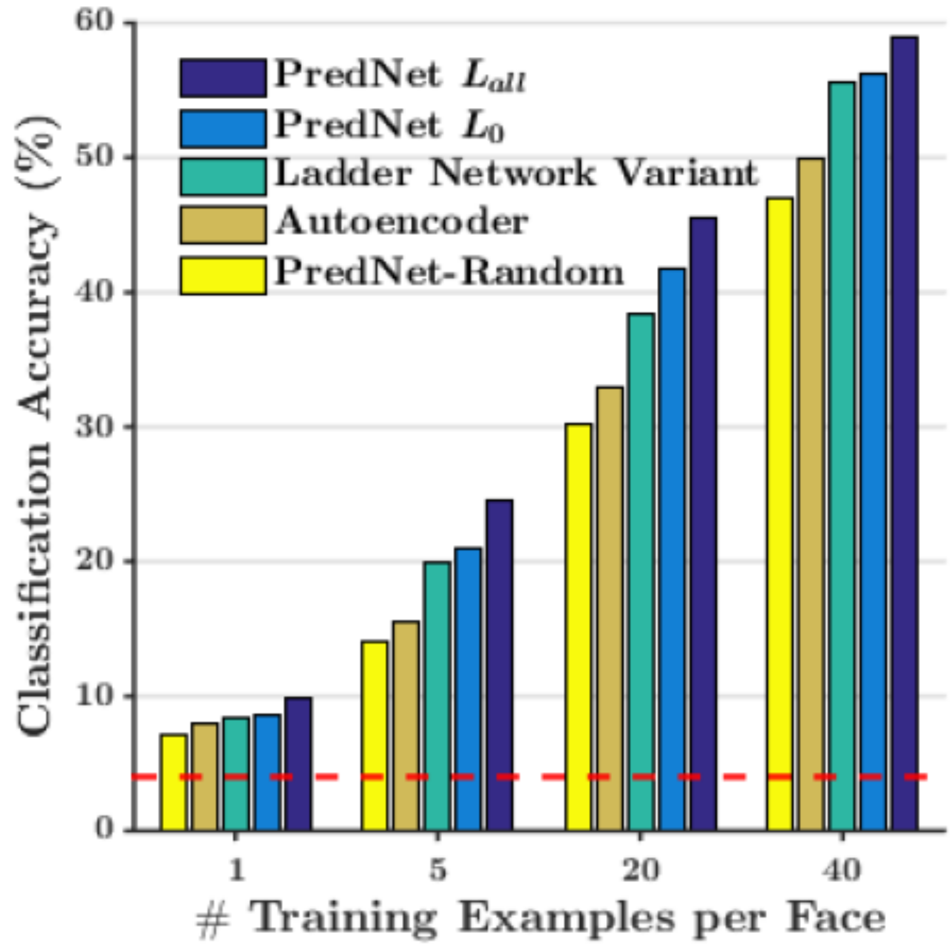
Lotter et al 2015, 2016

Training for prediction → successful image classification

Face recognition

20 faces

Training with few examples per face



Lotter et al 2015, 2016

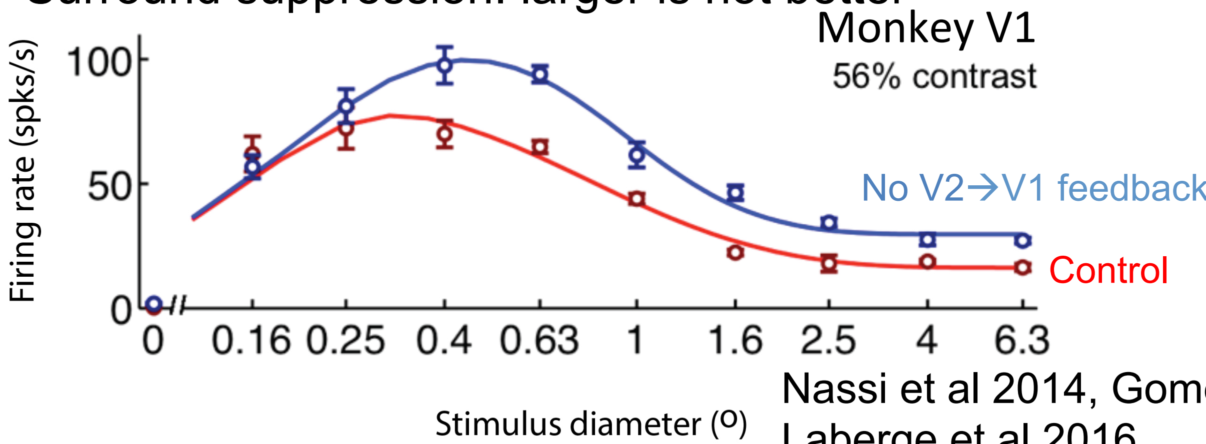
A model trained to predict video frames can reproduce many neurophysiological properties!

- On/Off temporal dynamics (e.g., Schmolesky et al, 1998)
- End-stopping and length suppression (e.g., Hubel and Wiesel, 1968)
- Sequence learning effects in visual cortex (e.g., Meyer and Olson 2011)
- Norm-based coding of faces (Leopold et al, 2006)
- Illusory contours (Lee and Nguyen 2001)
- Flash-lag effect (Khoei et al 2017)

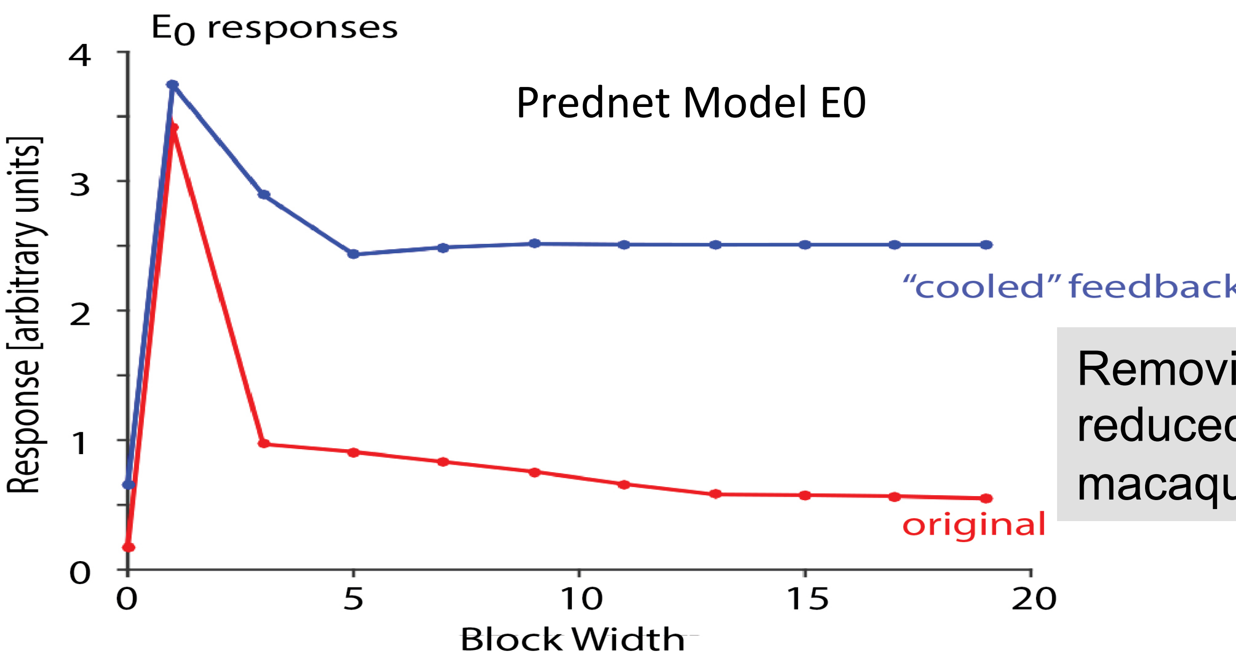
Lotter et al 2018

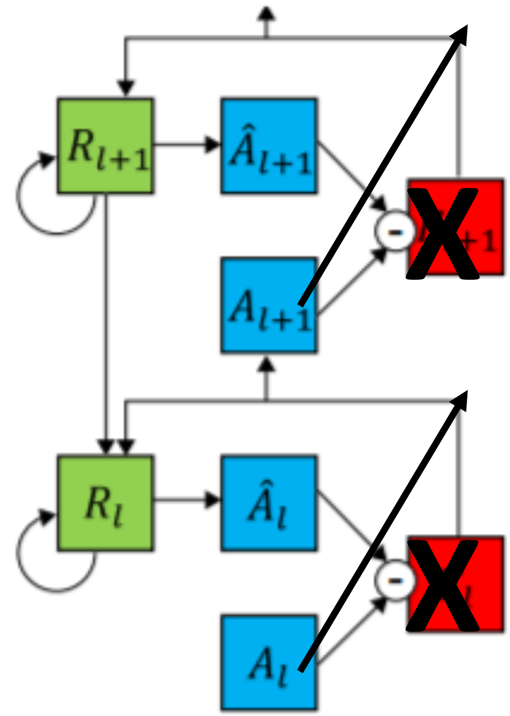
The unsupervised model can predict neural response properties

Surround suppression: larger is not better



Nassi et al 2014, Gomez-Laberge et al 2016





Removing feedback signals leads to reduced surround suppression in macaque V1 and in the model

Lotter et al 2015, 2016

Visual cognition: a sequence of routines*

Divide et impera

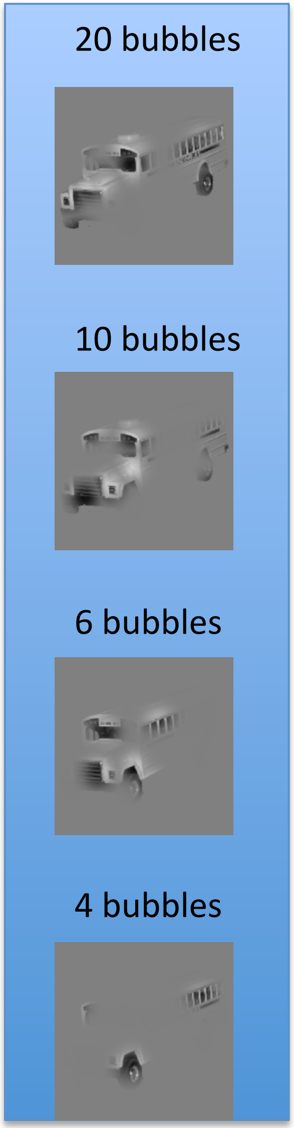
1. Extract initial sensory map → Call `VisualSampling`
2. Propose image gist → Call `RapidPeripheralAssessment`
3. Propose foveal objects → Call `FovealRecognition`
4. **<u>Inference from 1+2+3 → Call `PatternCompletion`</u>**
5. Temporary information storage → Call `VisualBuffer`
6. Task-dependent sampling → 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?
14. If satisfactory, answer the qu

Bill Lotter

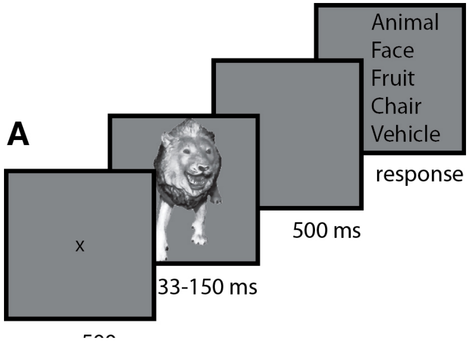
Martin Schrimpf

Hanlin Tang

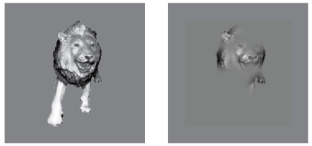
Evaluating pattern completion



A



C



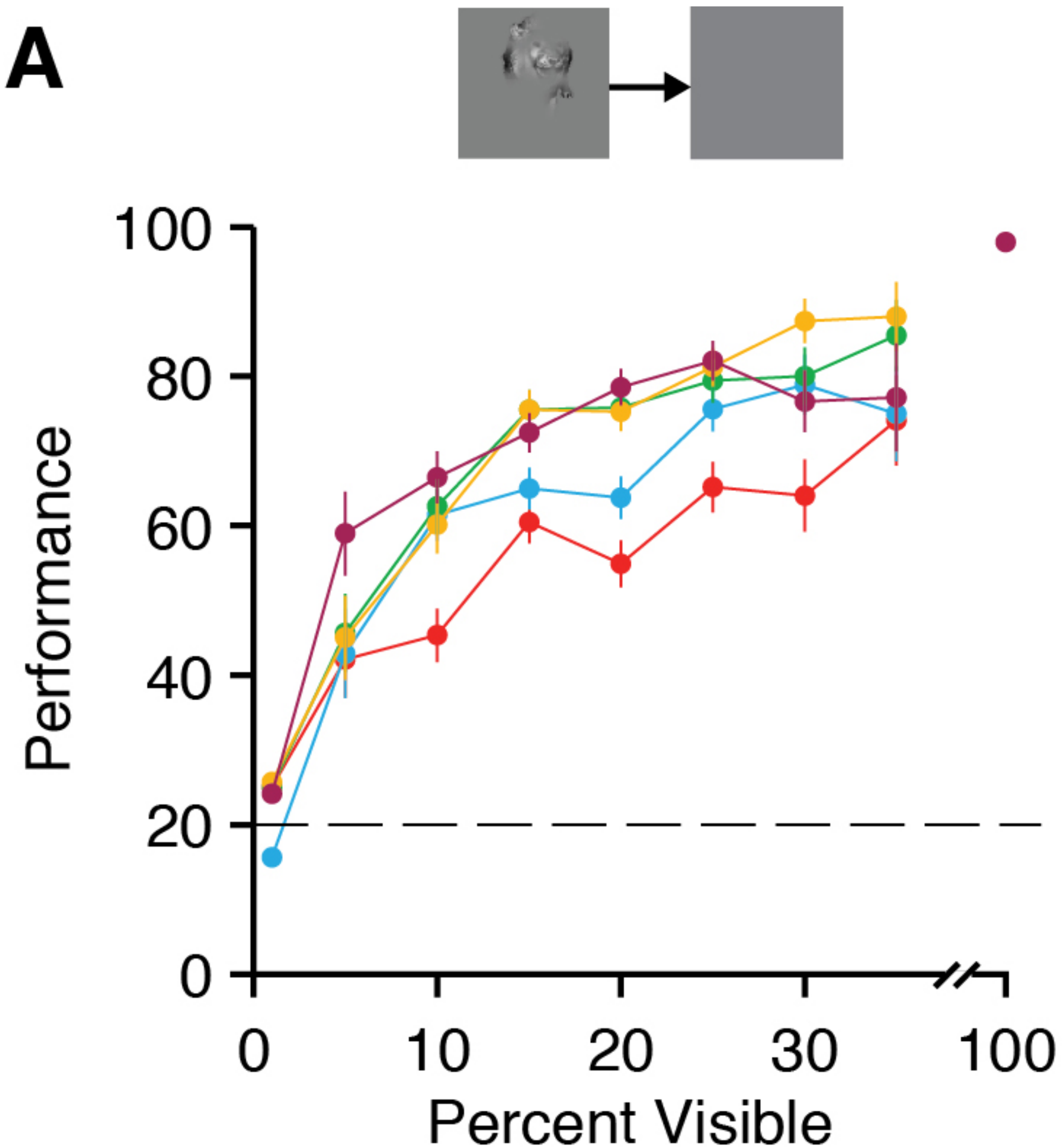
Whole Partial

D

Occluded

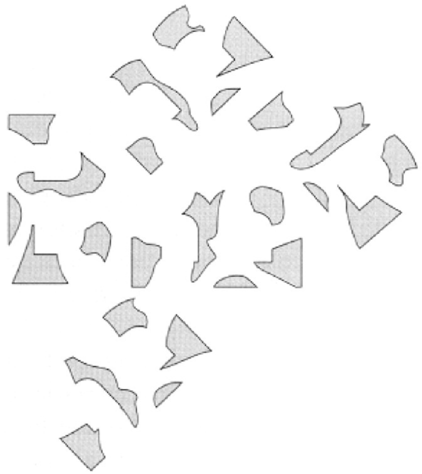
Hanlin Tang, Bill Lotter, Martin Schrimpf
Tang et al 2018

Strong robustness to limited visibility

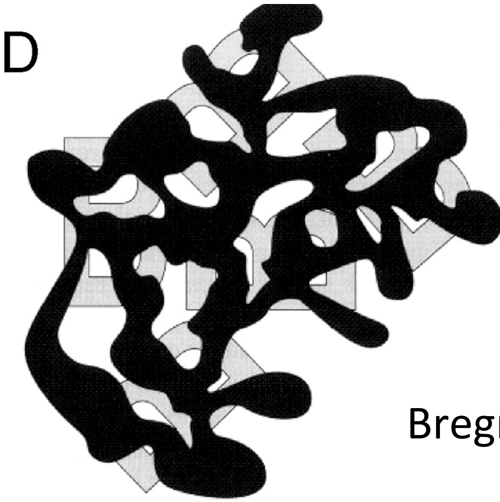


Backward masking also disrupts recognition of occluded objects

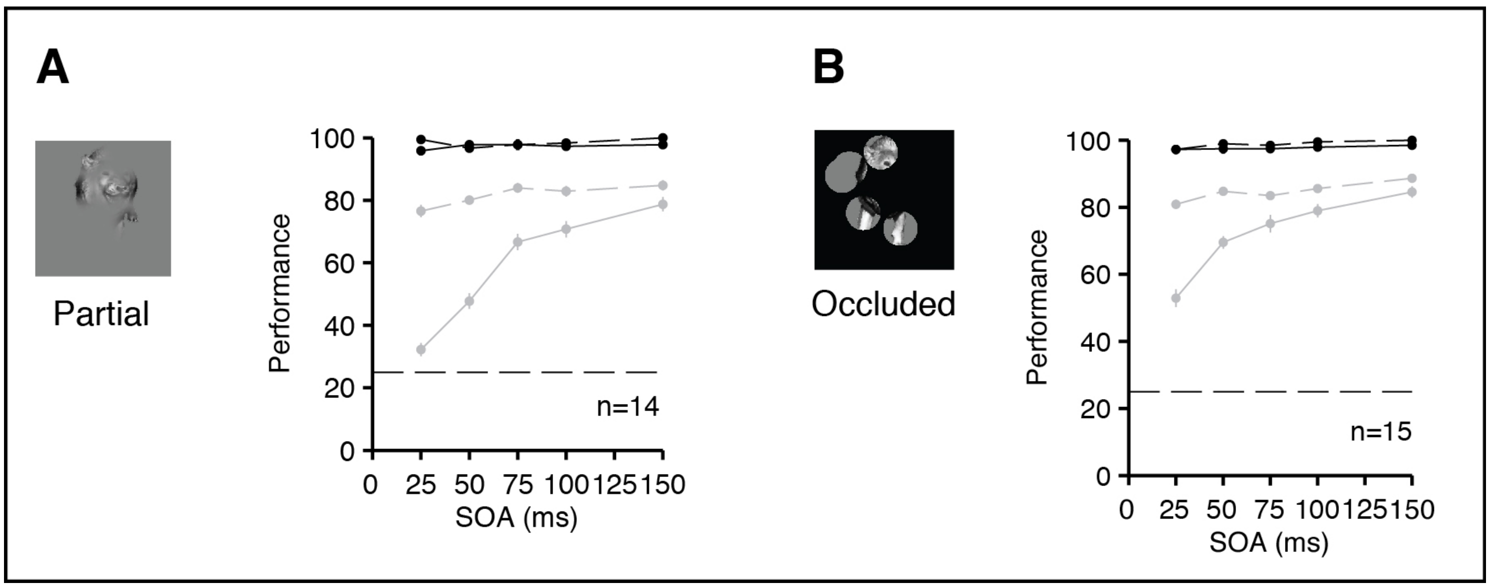
C



D

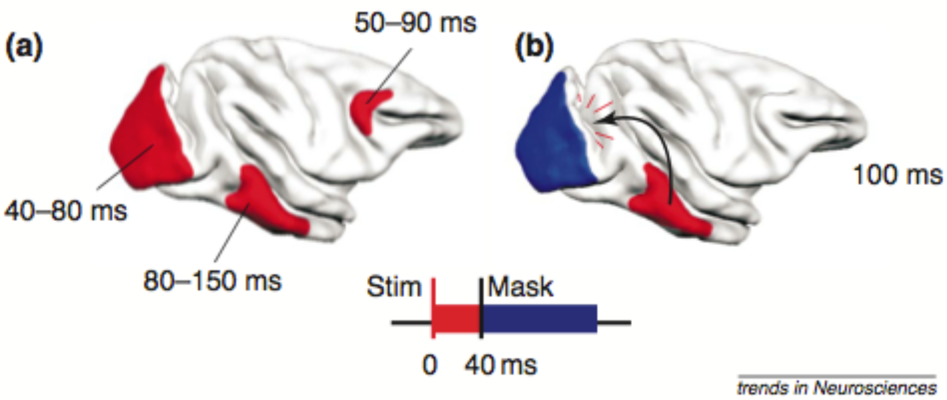


Bregman 1981



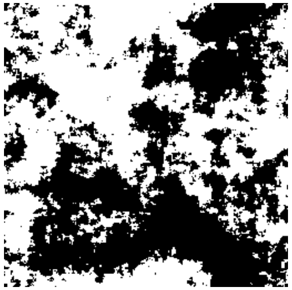
Backward masking interrupts processing (presumably of feedback/recurrent computations)

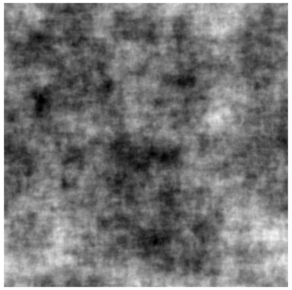
Models:



Lamme V, Roelfsema P (2000)

Masks:

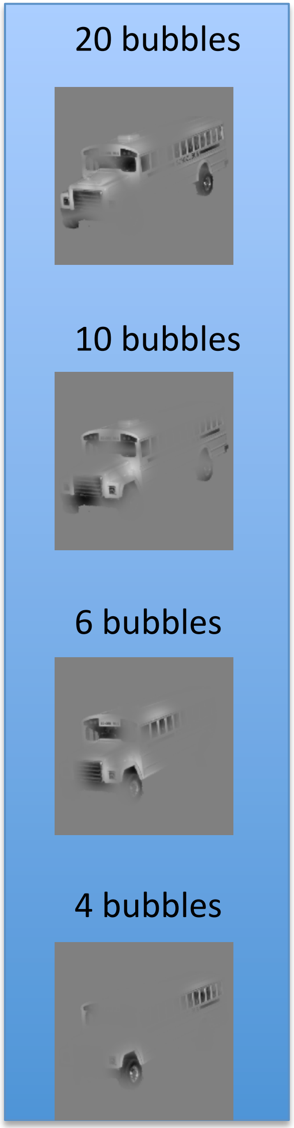


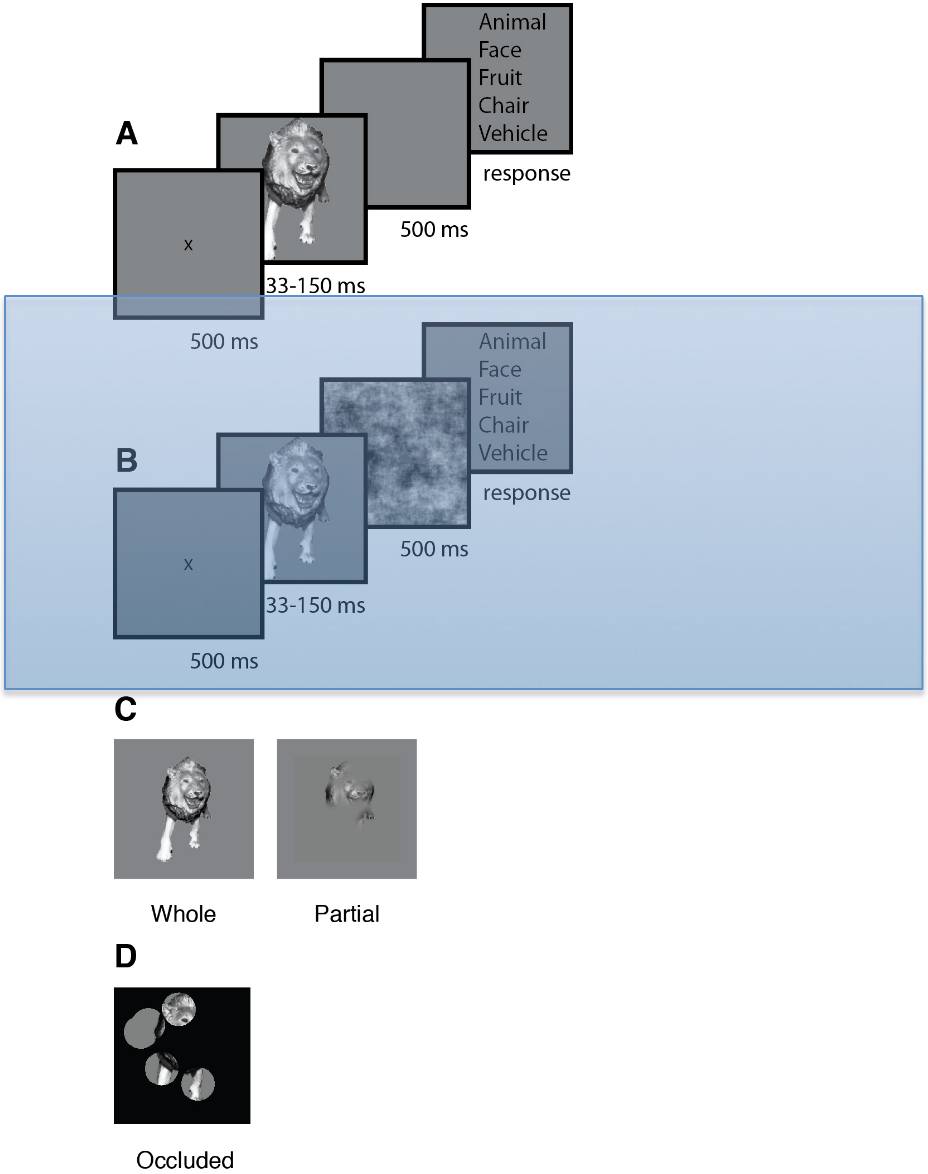


- Short delays (SOA<20ms): mask reduces visibility
- Longer delays: mask is purported to disrupt recurrent/top-down processing

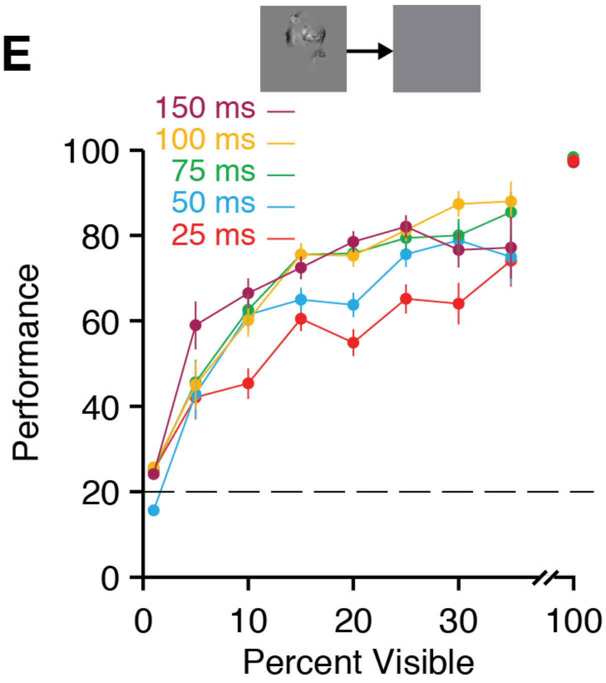
V1: Bridgeman 1980, Maknik and Livinsgtone 1998, Lamme et al 2002
IT: Kovacs et al 1995, Rolls et al 1999

Evaluating pattern completion abilities

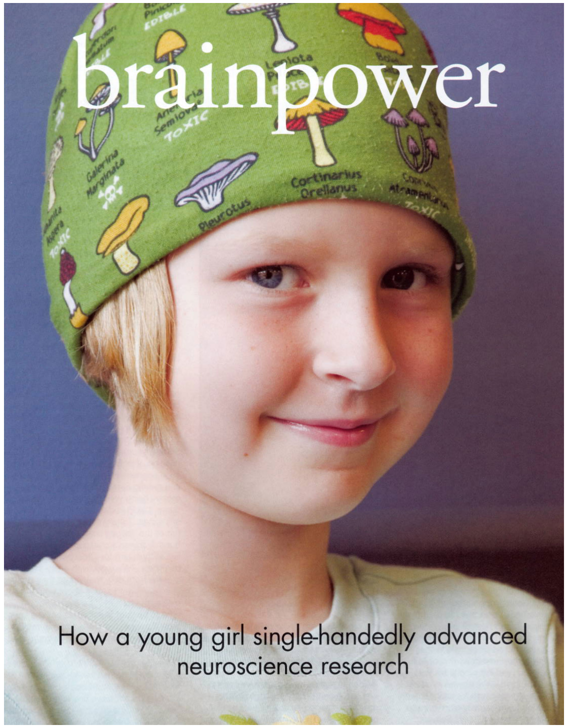


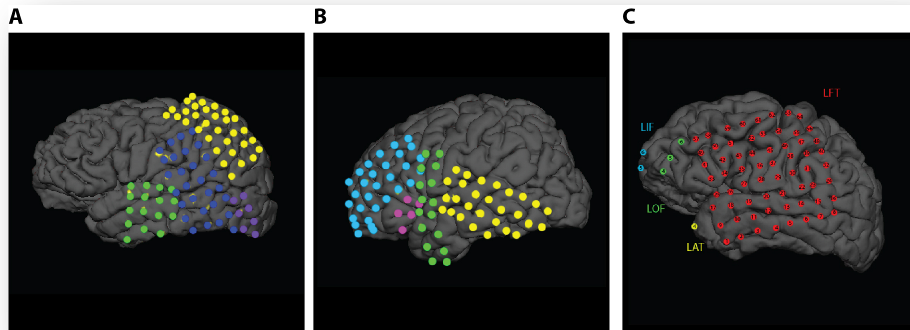


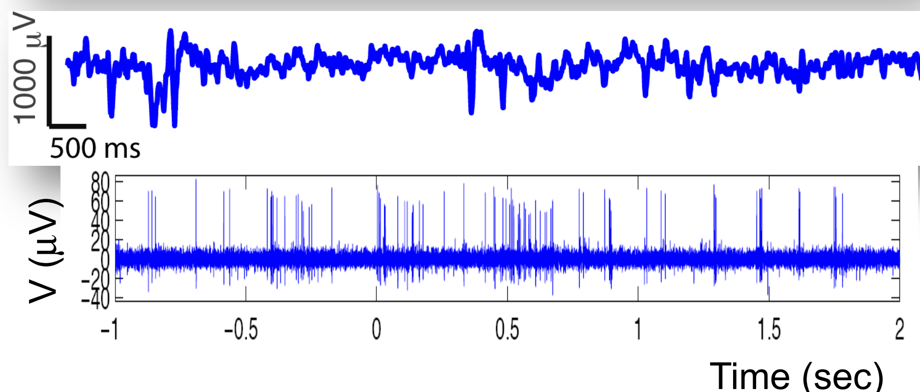
Backward masking disrupts pattern completion

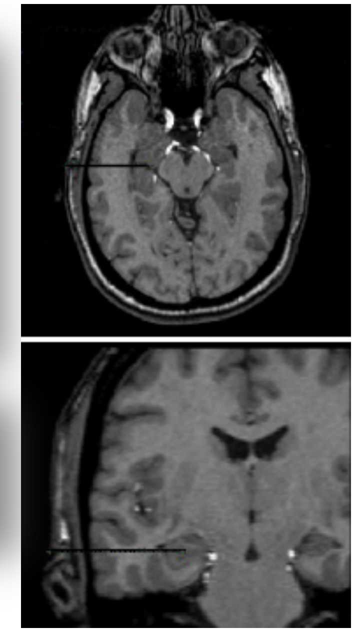


Peeking inside the human brain





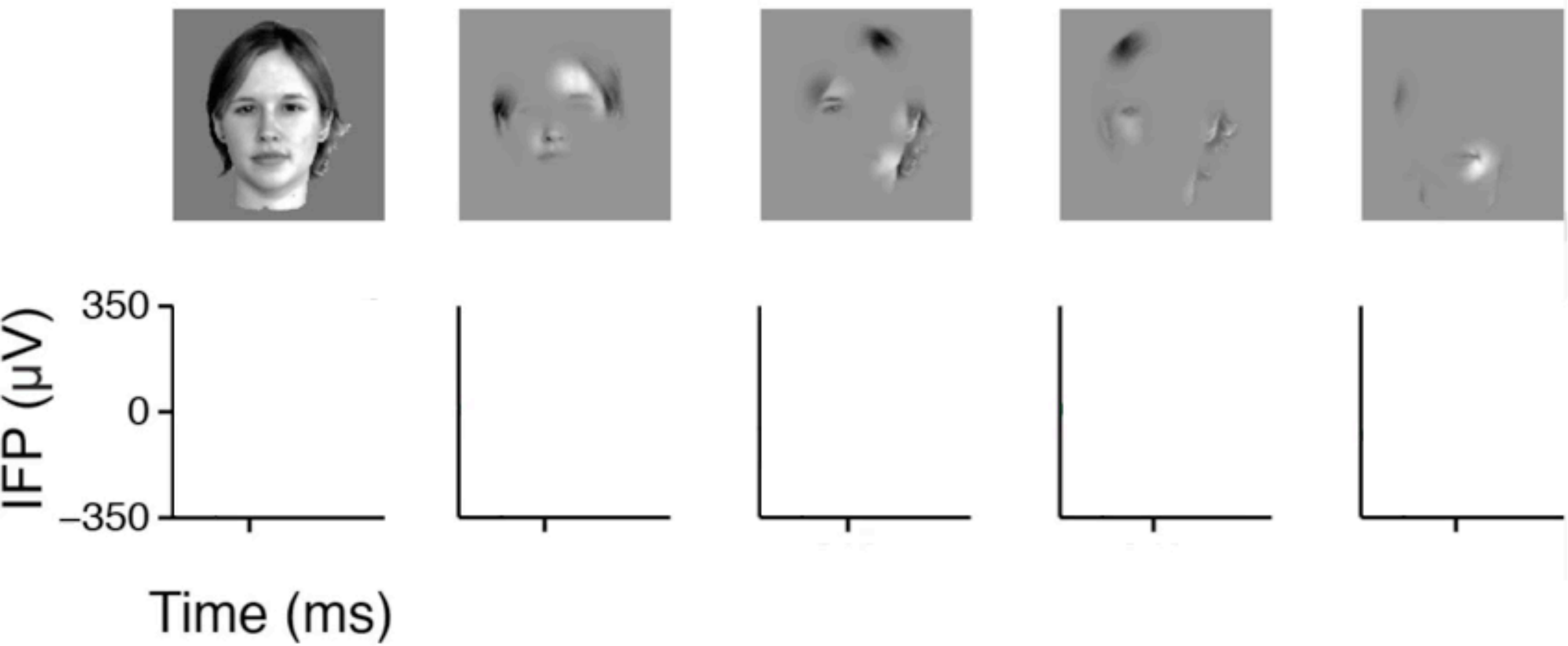




- Patients with pharmacologically intractable epilepsy
- Multiple electrodes implanted to localize seizure focus
- Patients stay in the hospital for about 7-10 days
- All experiments are approved by the Institutional Review Boards
- All testing is performed with the subjects' consent

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

Neural responses to partial objects are delayed



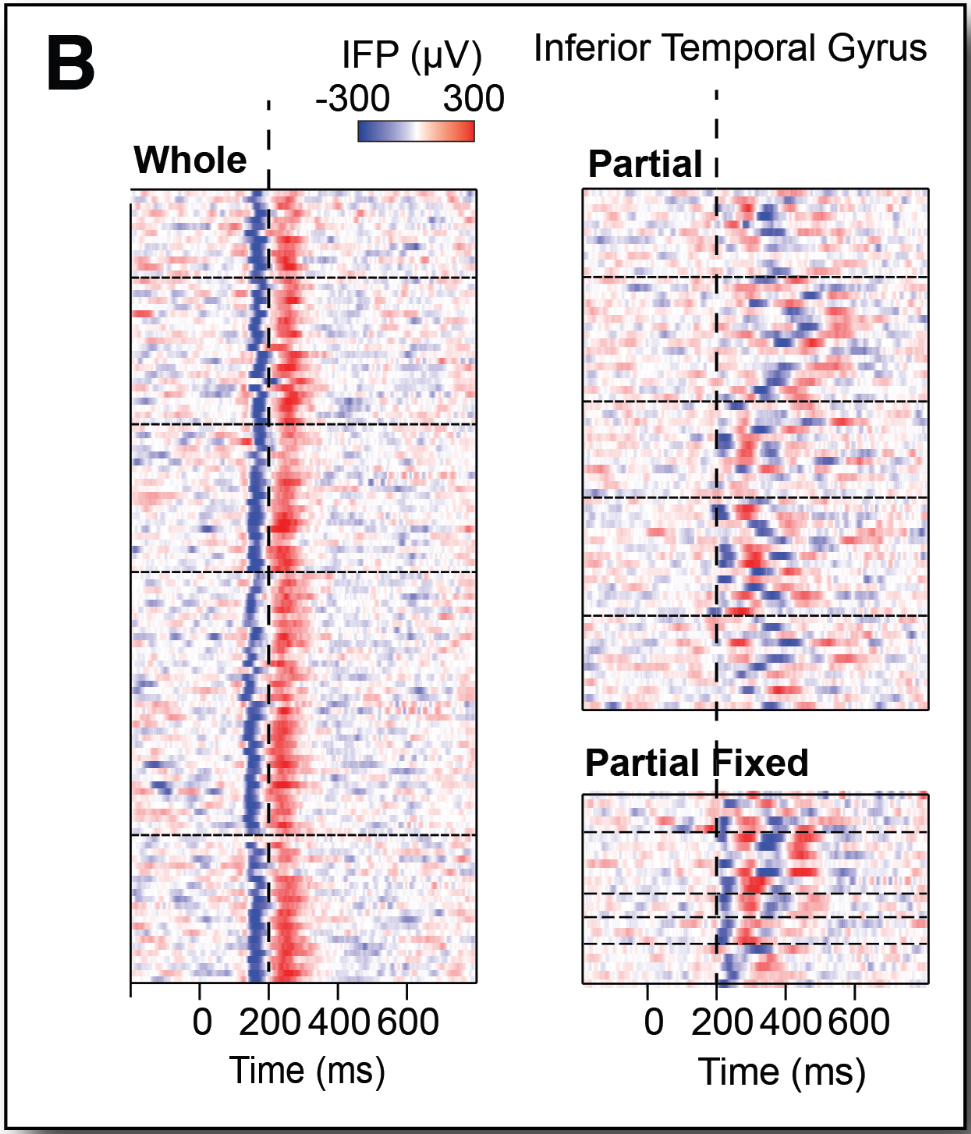
Tang et al, 2014, 2018

See also: Pasupathy lab, eLife 2017

Inferior Temporal Gyrus

Macaque IT and PFC

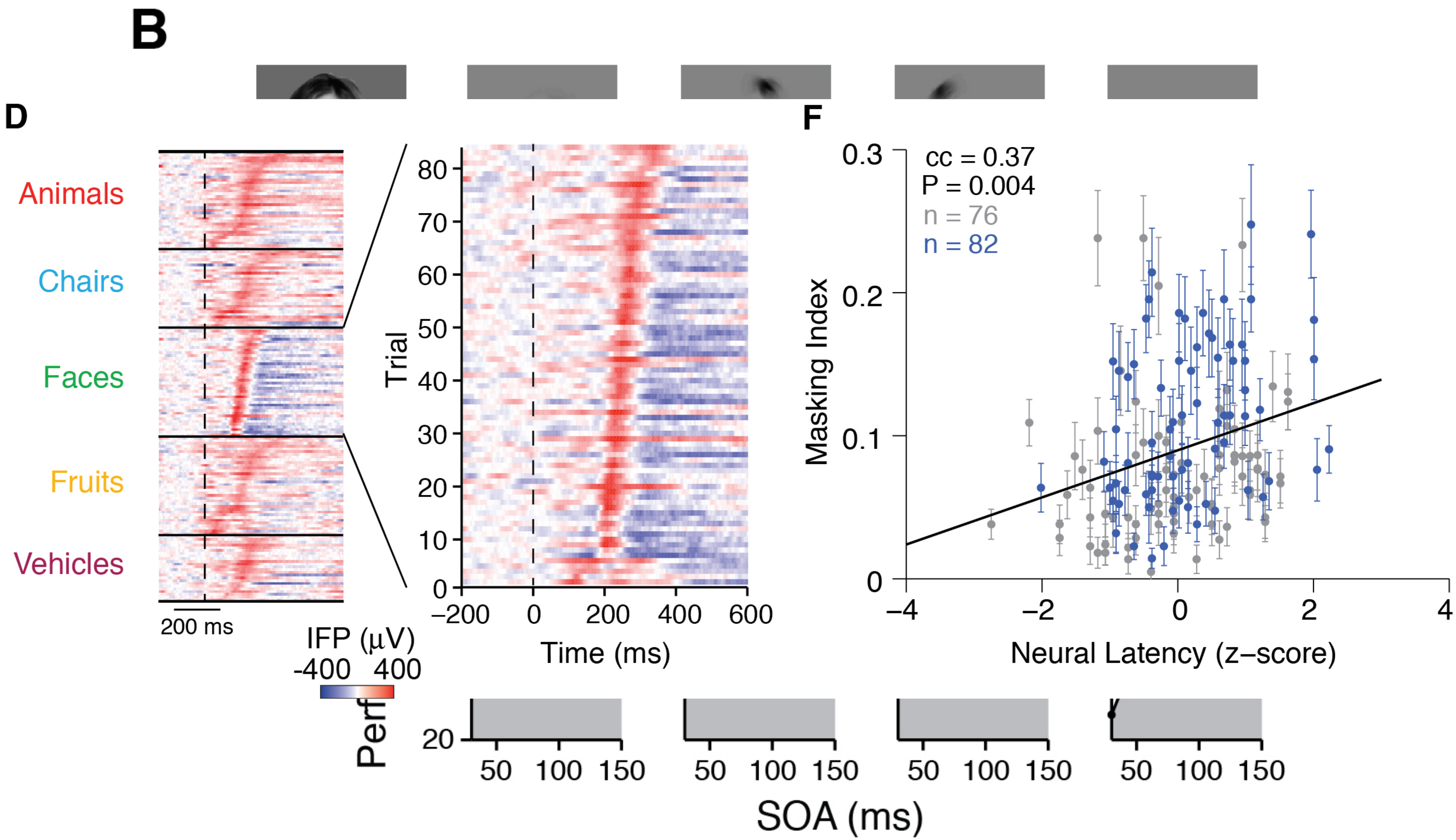
Neural responses to partial objects are delayed



Inferior Temporal Gyrus

Tang et al, 2014, 2018

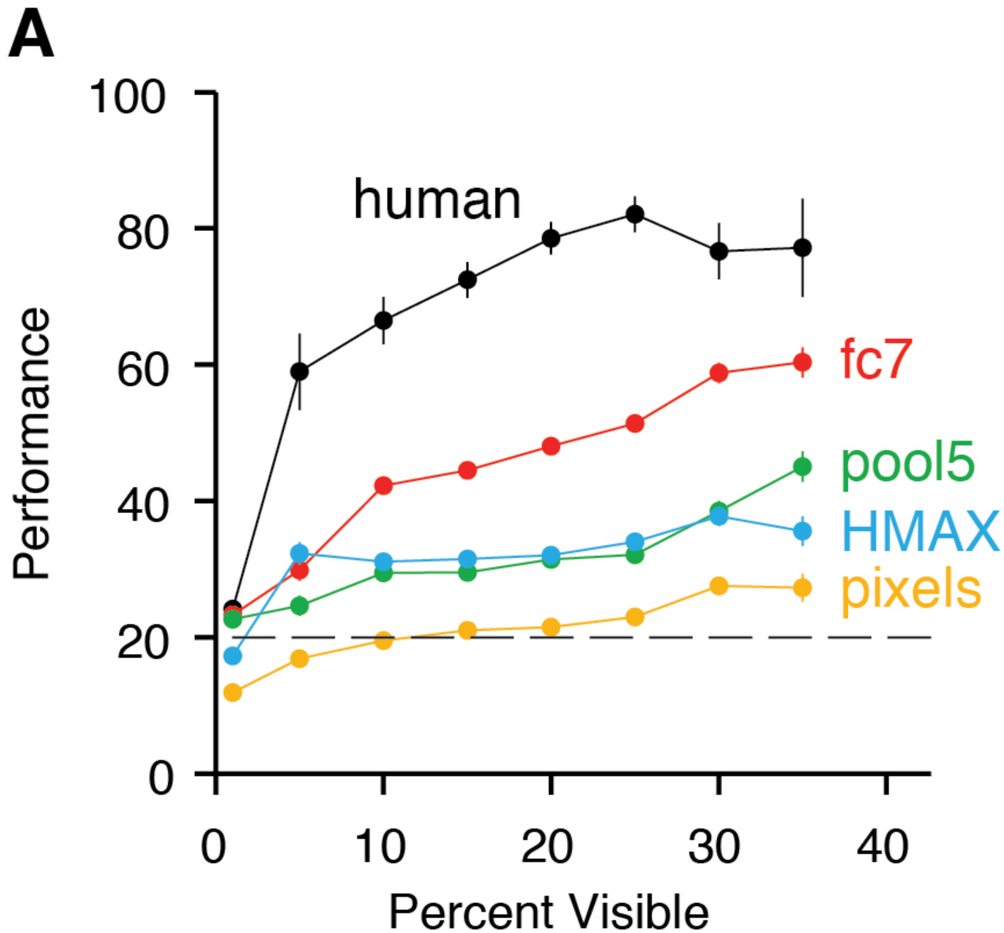
The effects of backward masking are correlated with the neural delays



Inferior Temporal Gyrus

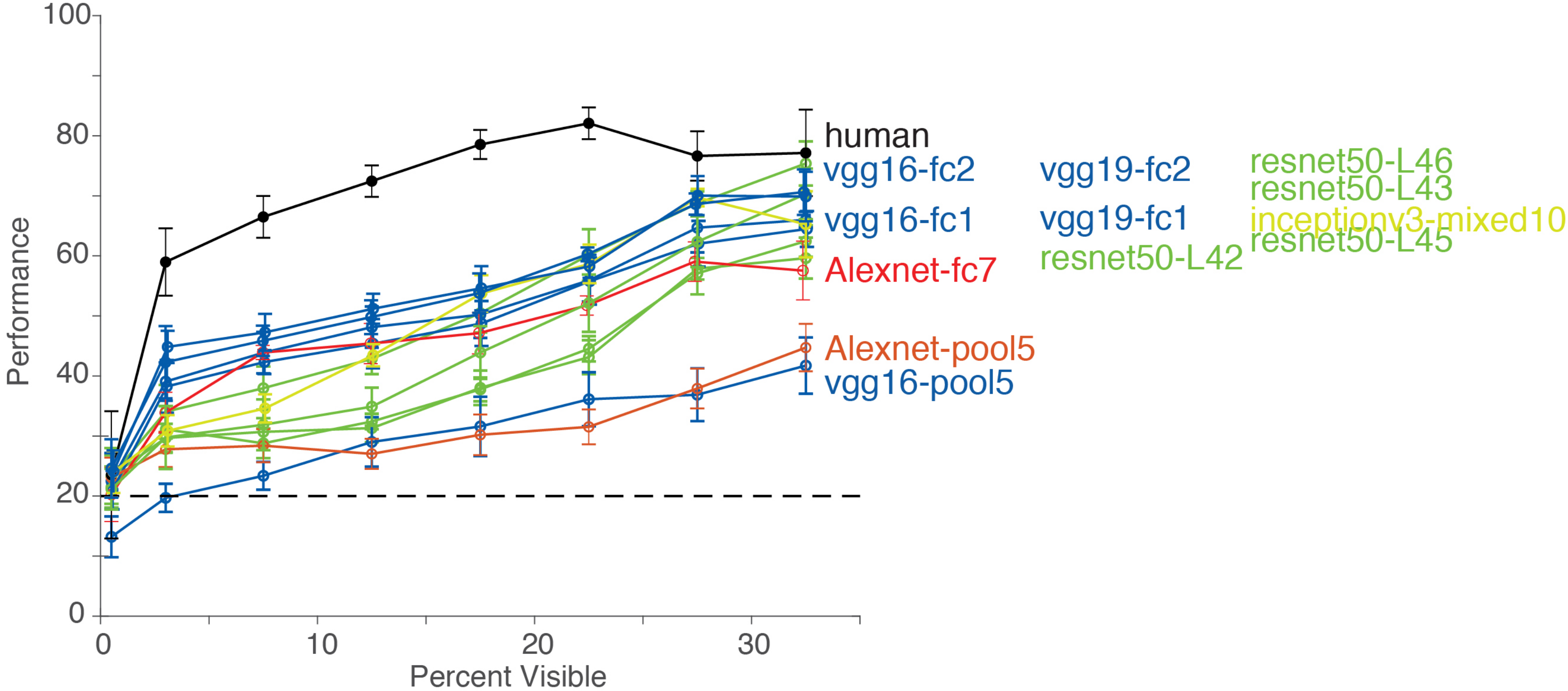
Tang et al, 2014, 2018

Bottom-up models significantly underperform in recognition of partial images

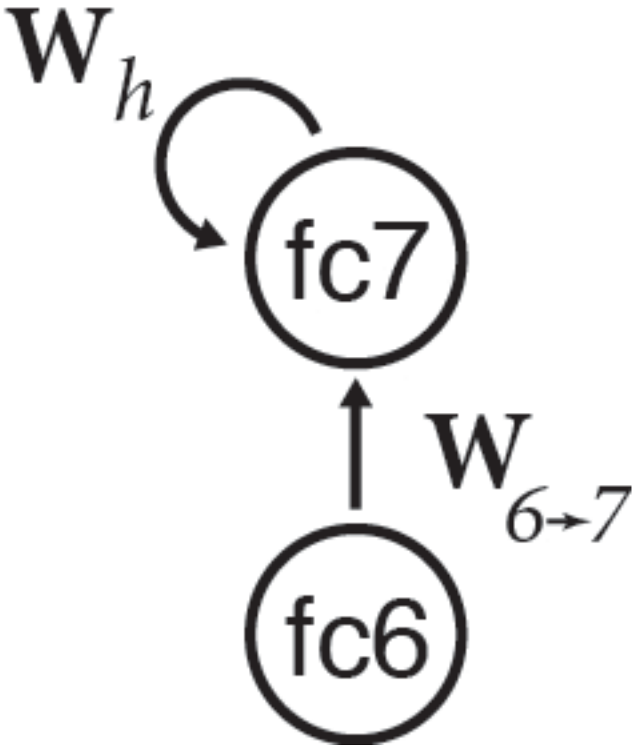


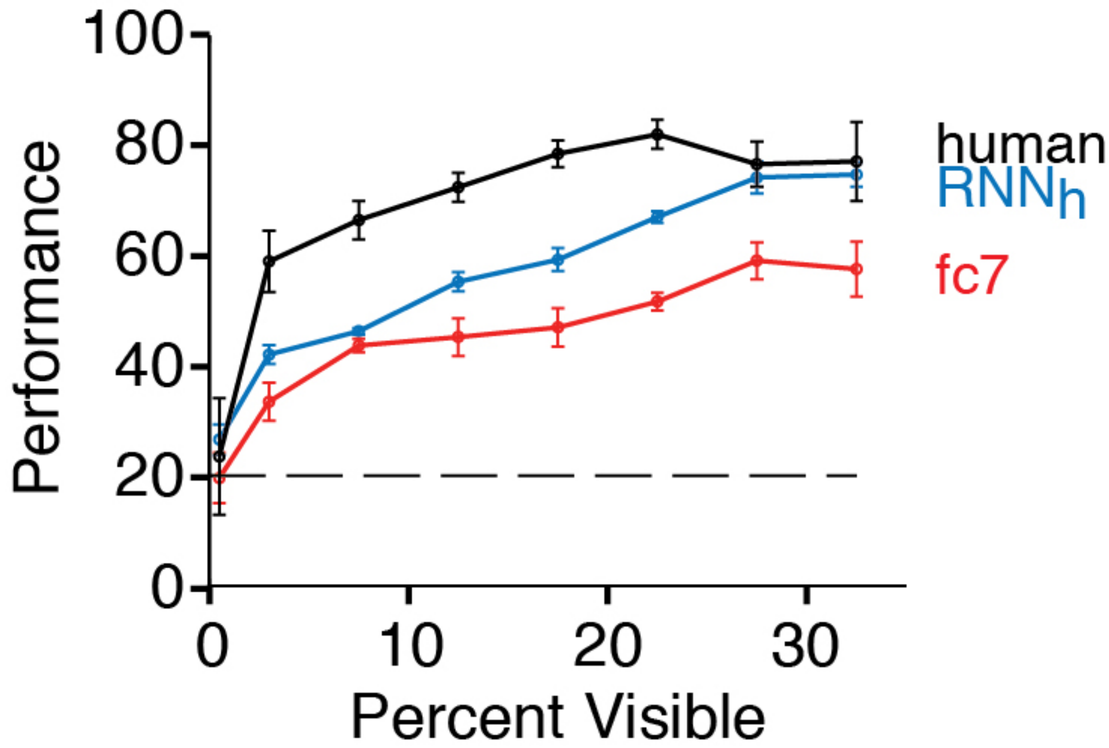
See also Pepik et al 2015, Wyatte et al 2012

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



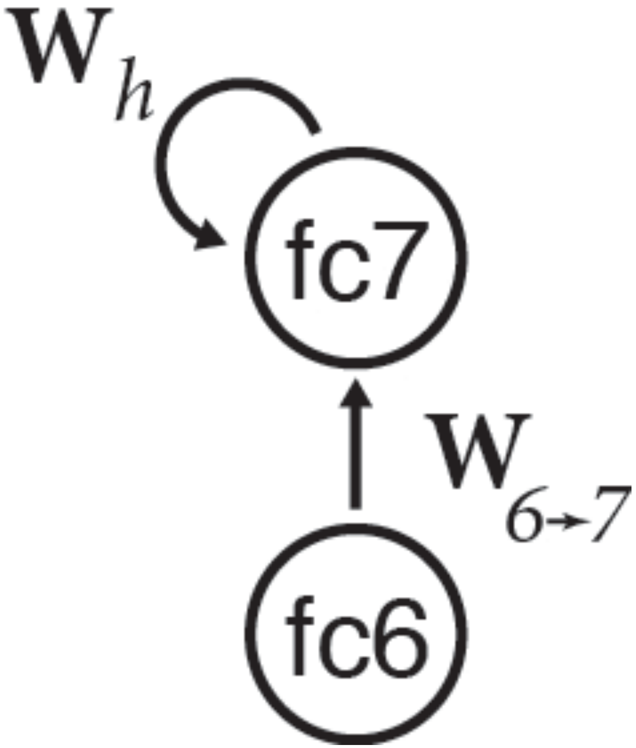
Recurrent Hopfield network (RNN_h) improves recognition performance for partial images

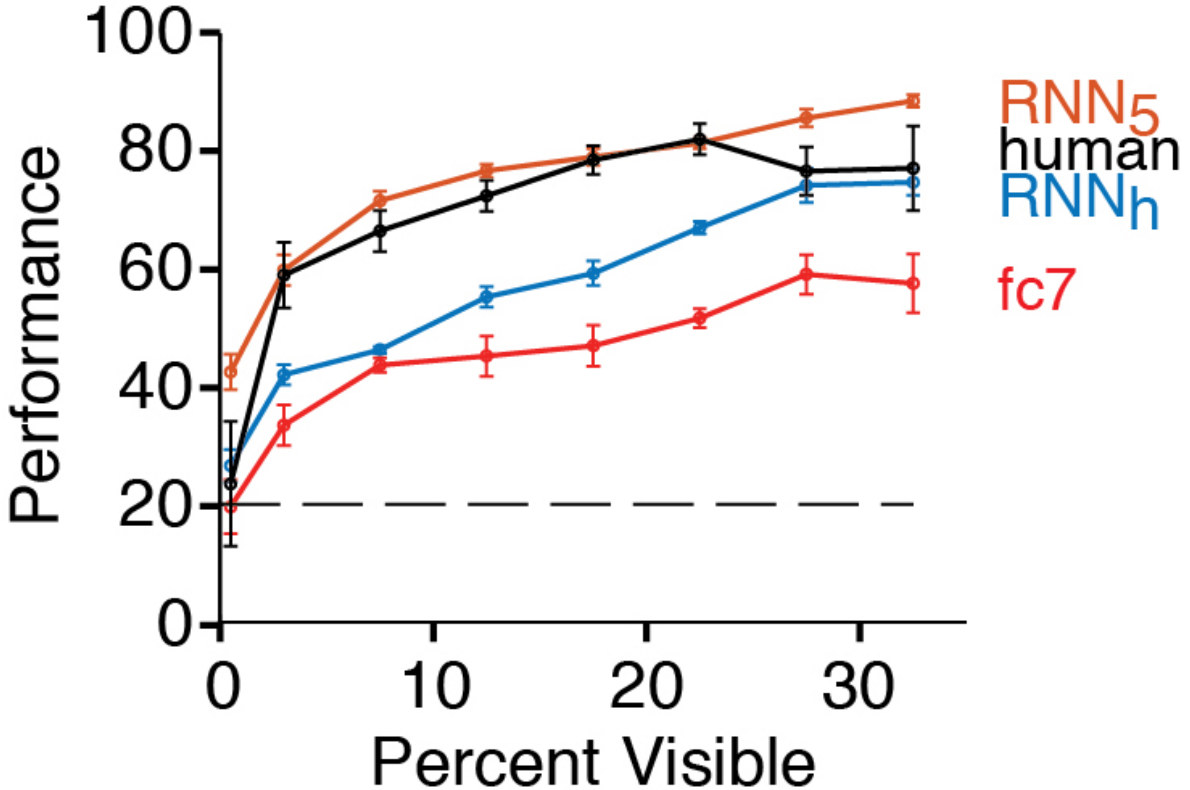




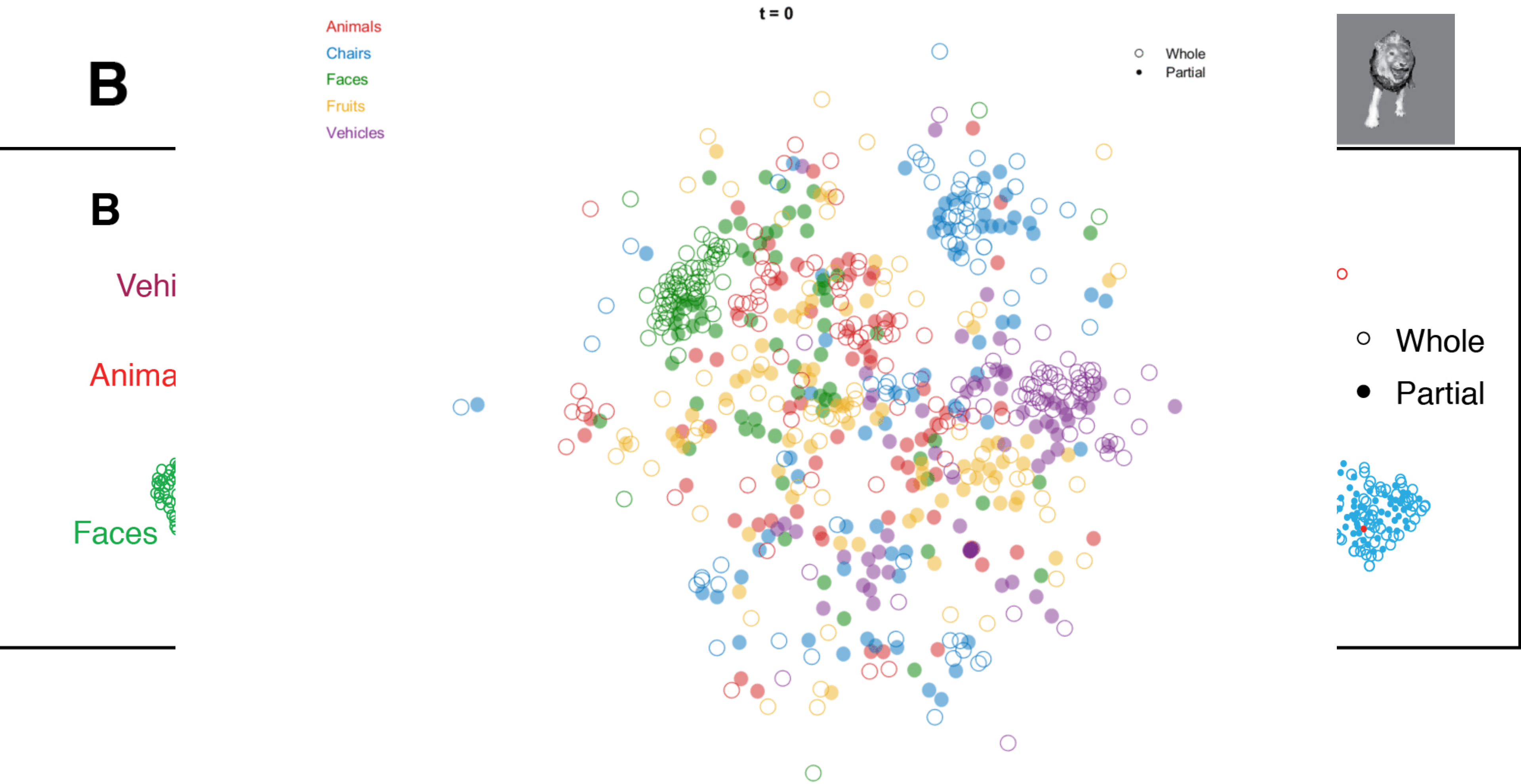
NOTE: 0 free parameters

Training the recurrent connections with partial objects yields higher performance

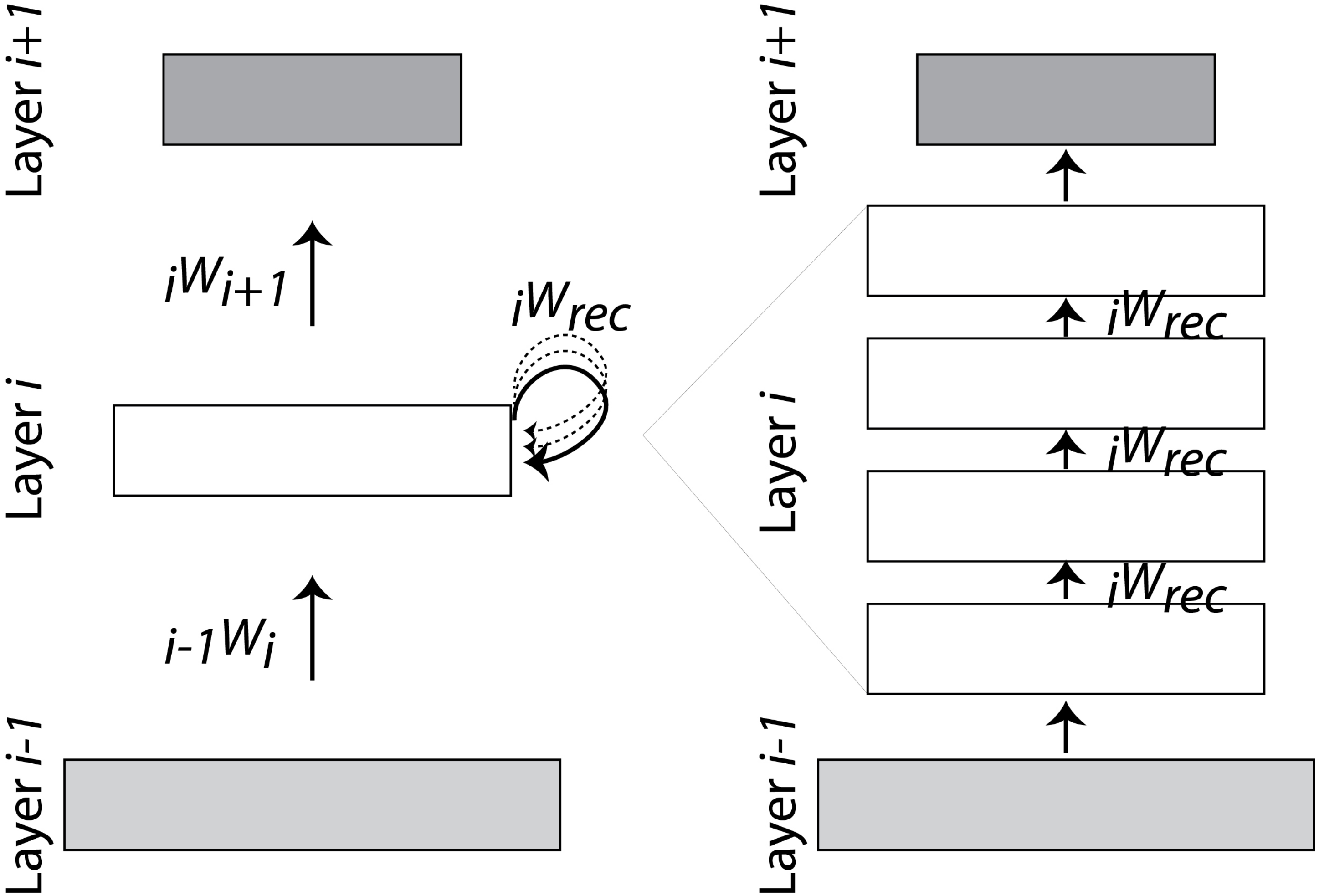




Recurrent computations bring the representation of partial objects towards the whole objects



Why recurrent connections?



1. Fewer units
2. Fewer weights
3. **Flexible number of computations**

Interim summary 2

Visual recognition is robust to heavy occlusion

Robustness impaired by backward masking with SOA<50 ms

Physiological delays of ~50 ms in visually selective signals along the ventral visual stream (humans/monkeys)

State-of-the-art bottom-up models fail to capture robustness to occlusion

Proof-of-principle model to solve pattern completion:

- Recurrent network in top layer
- Attractor-like dynamics
- 0 free parameters

There is much more to pattern completion: top-down signals, 3D cues, context

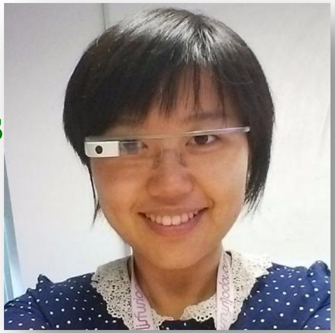
Eye movements are critical for scene understanding

Visual cognition: a sequence of routines*

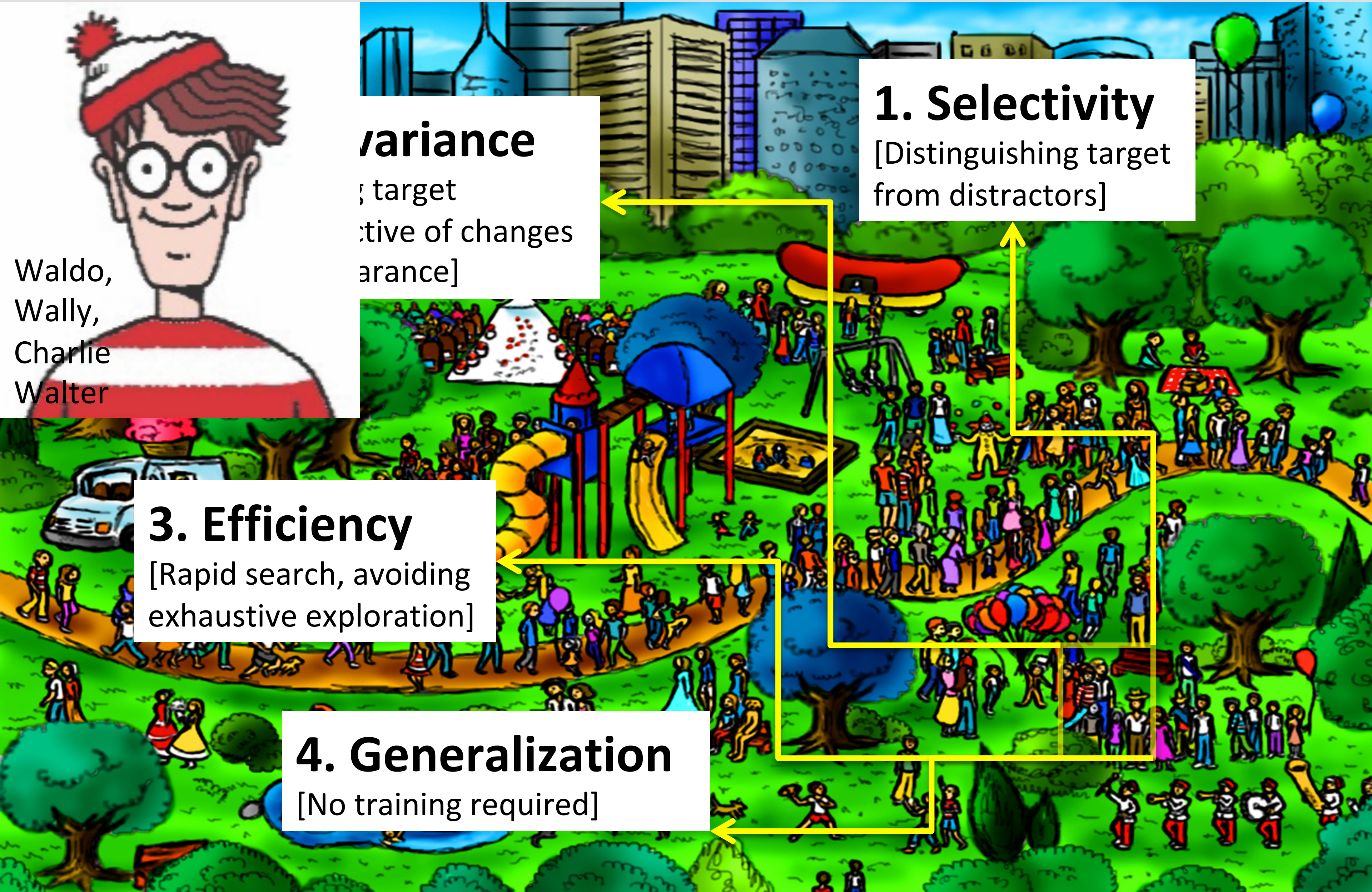
Divide et impera

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 `TaskTerminationDecis`
14. If satisfactory, answer the question → Call `TaskReport`

Mengmi Zhang

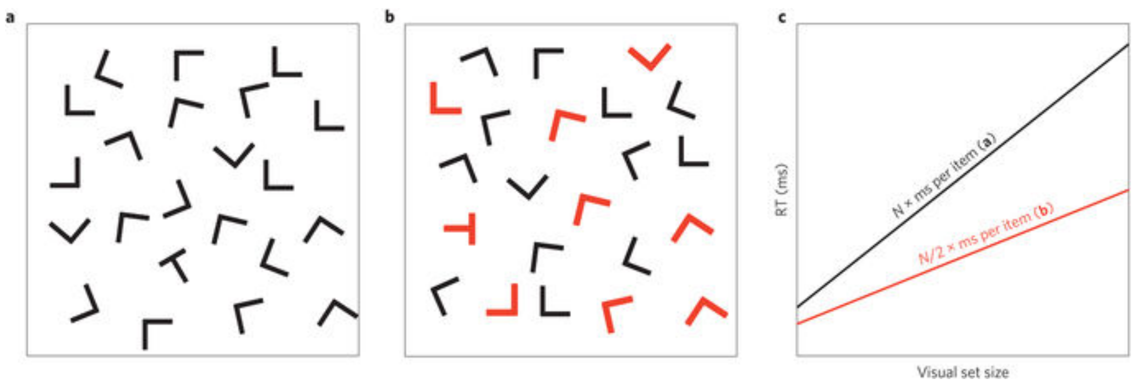


Four key properties of visual search

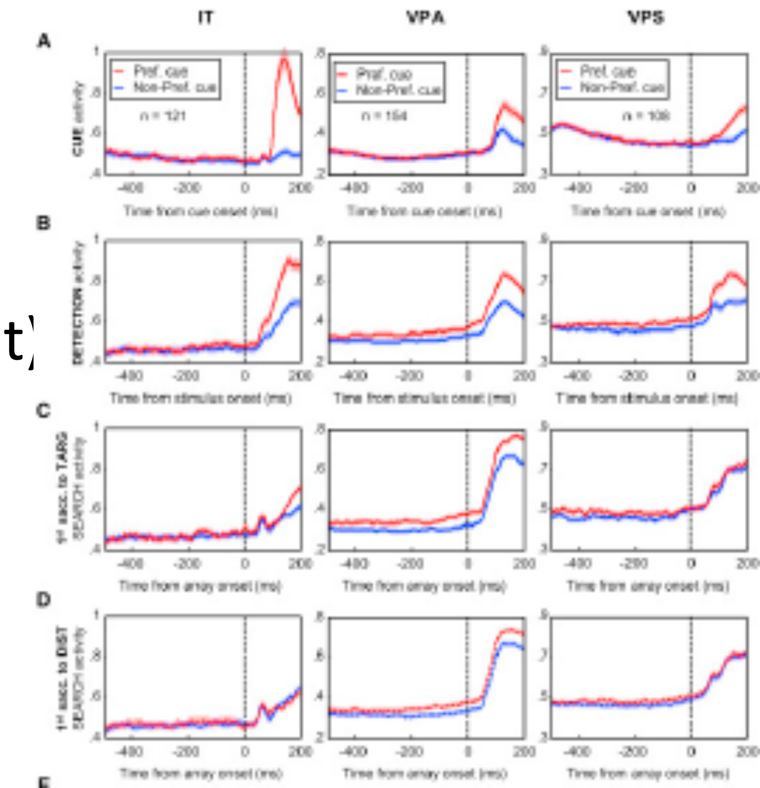


On the shoulders of giants

1. Human psychophysics: mostly identical target search (no invariance)



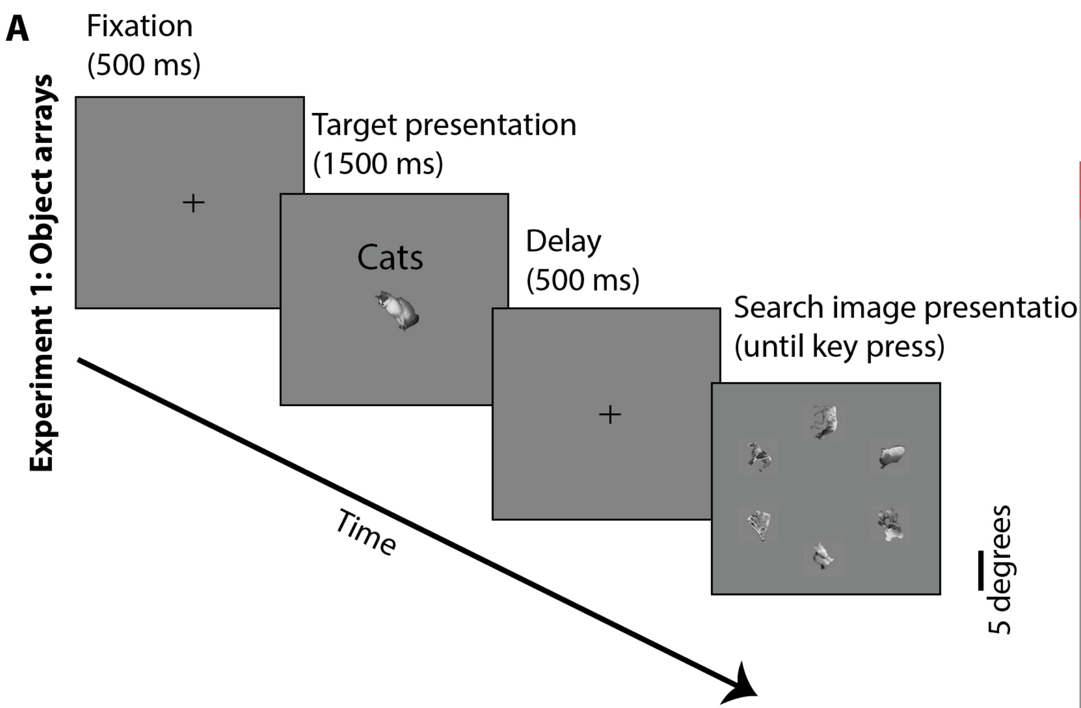
2. Neurophysiology: no invariance, no generalization

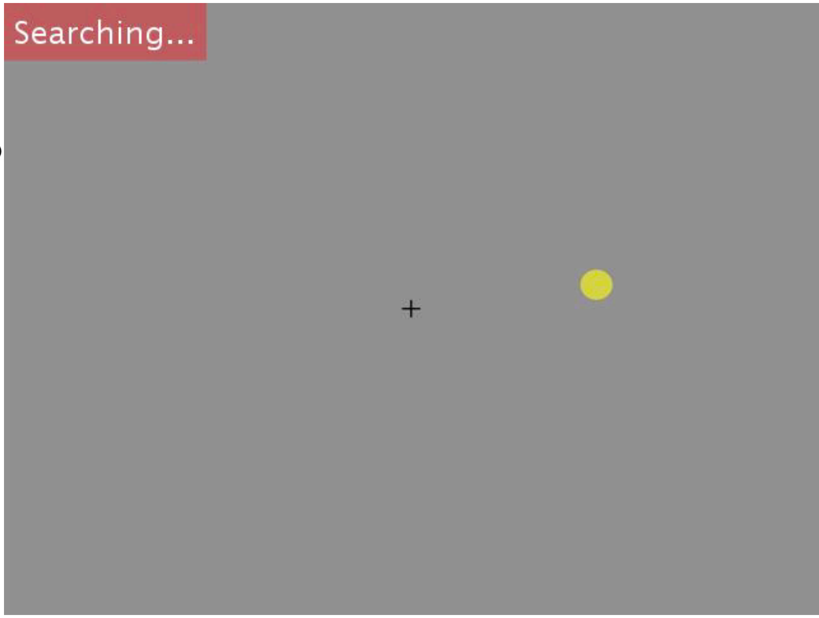


3. Computer vision: object detection via massive training (not zero-shot nor efficient)

Selectivity, invariance, efficiency, generalization

Three increasingly more complex tasks



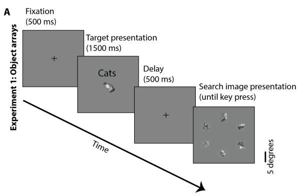


Zhang et al, 2018

Three increasingly more complex tasks

Visual search consists of a rapid sequence of saccades

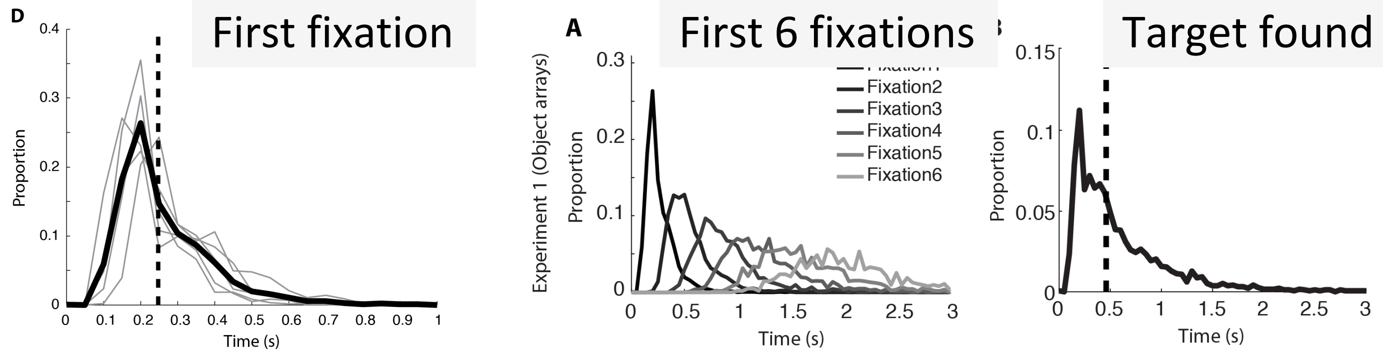
Experiment 1



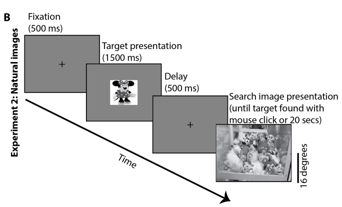
First fixation

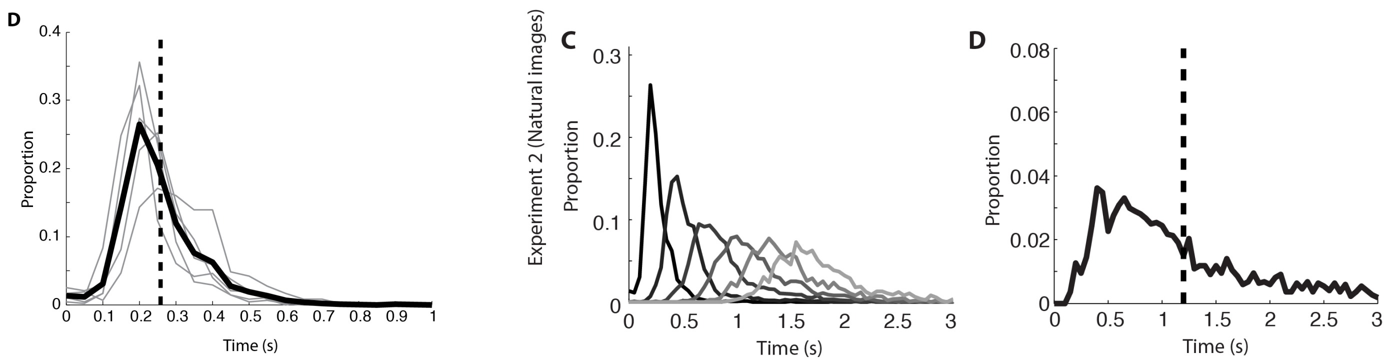
First 6 fixations

Target found

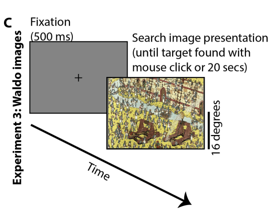


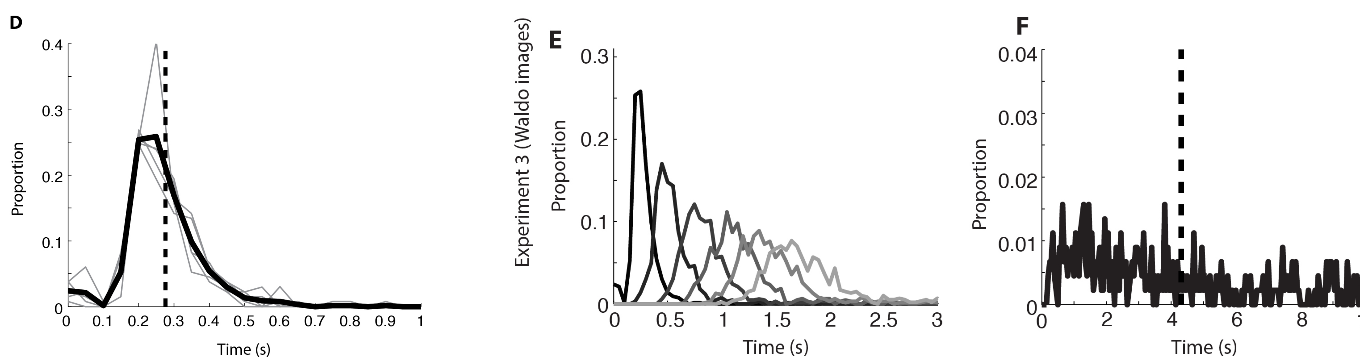
Experiment 2



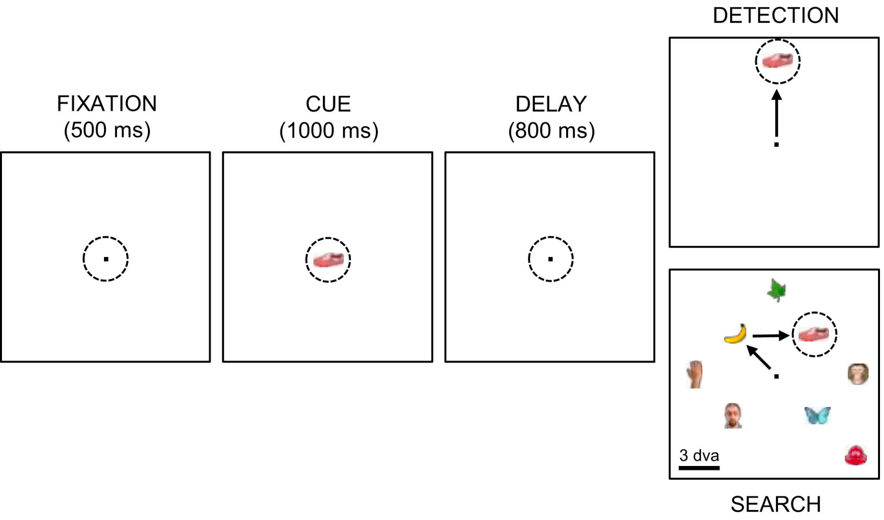


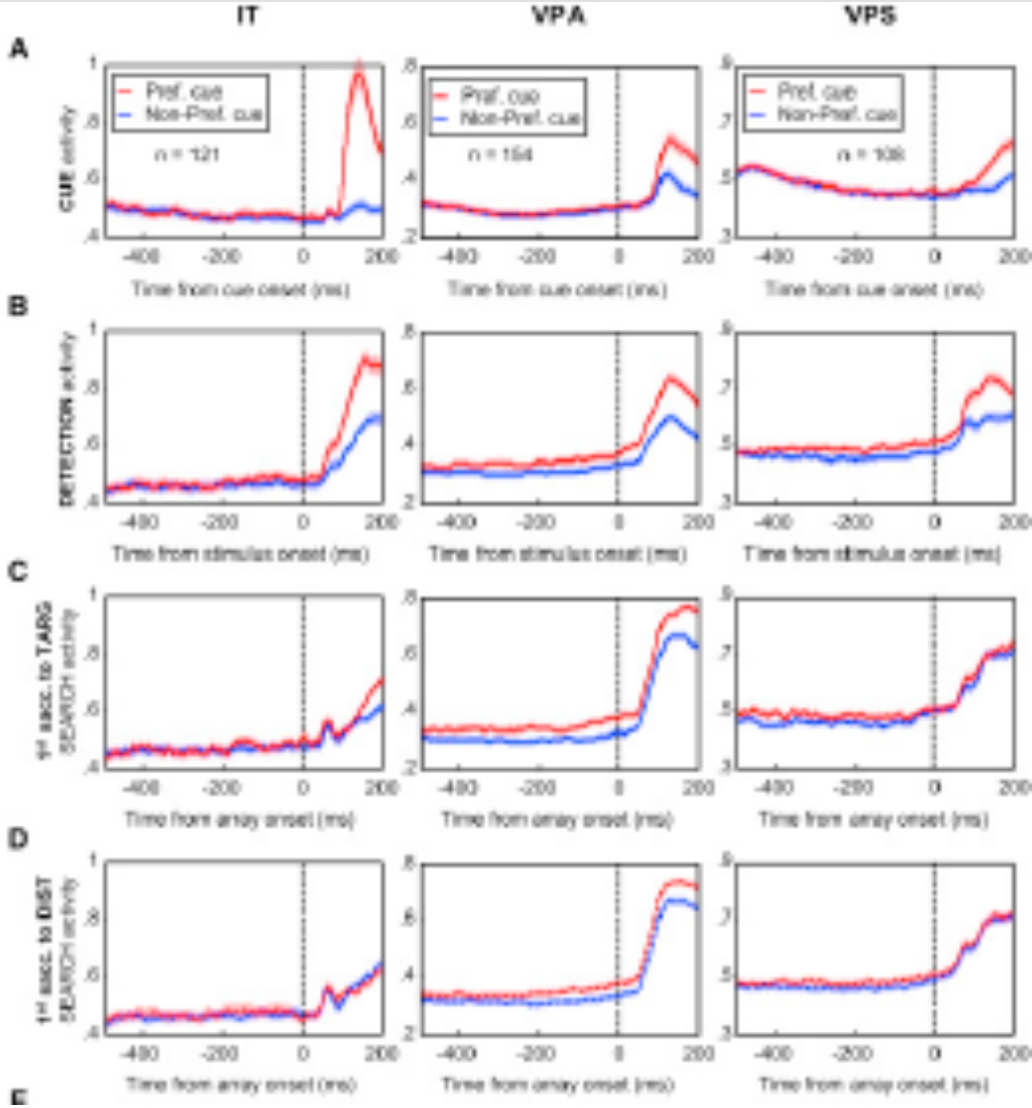
Experiment 3





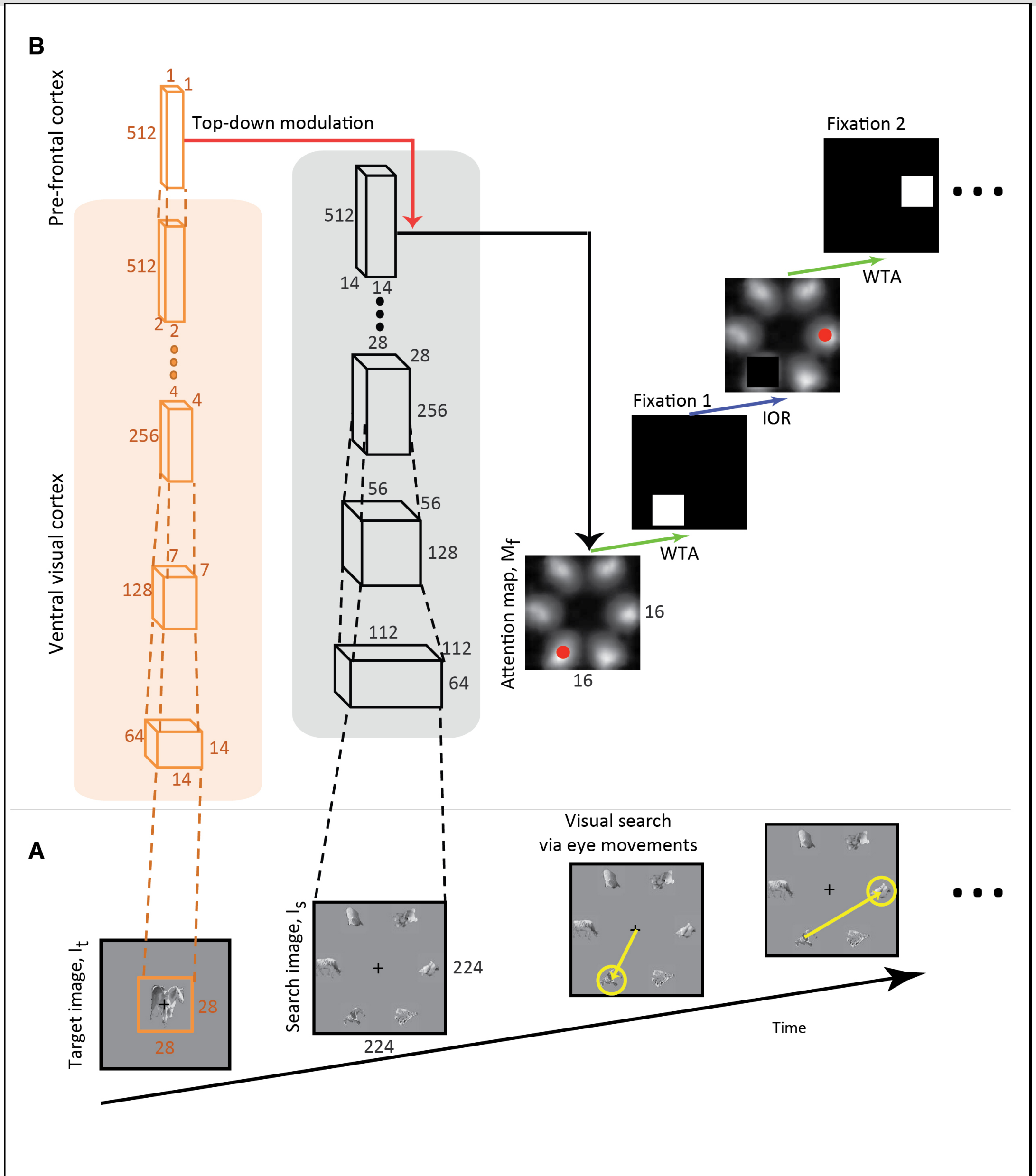
Neural mechanisms of attention modulation





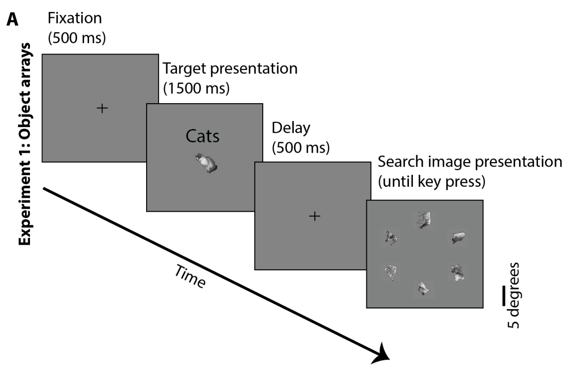
Bichot and Desimone 2015

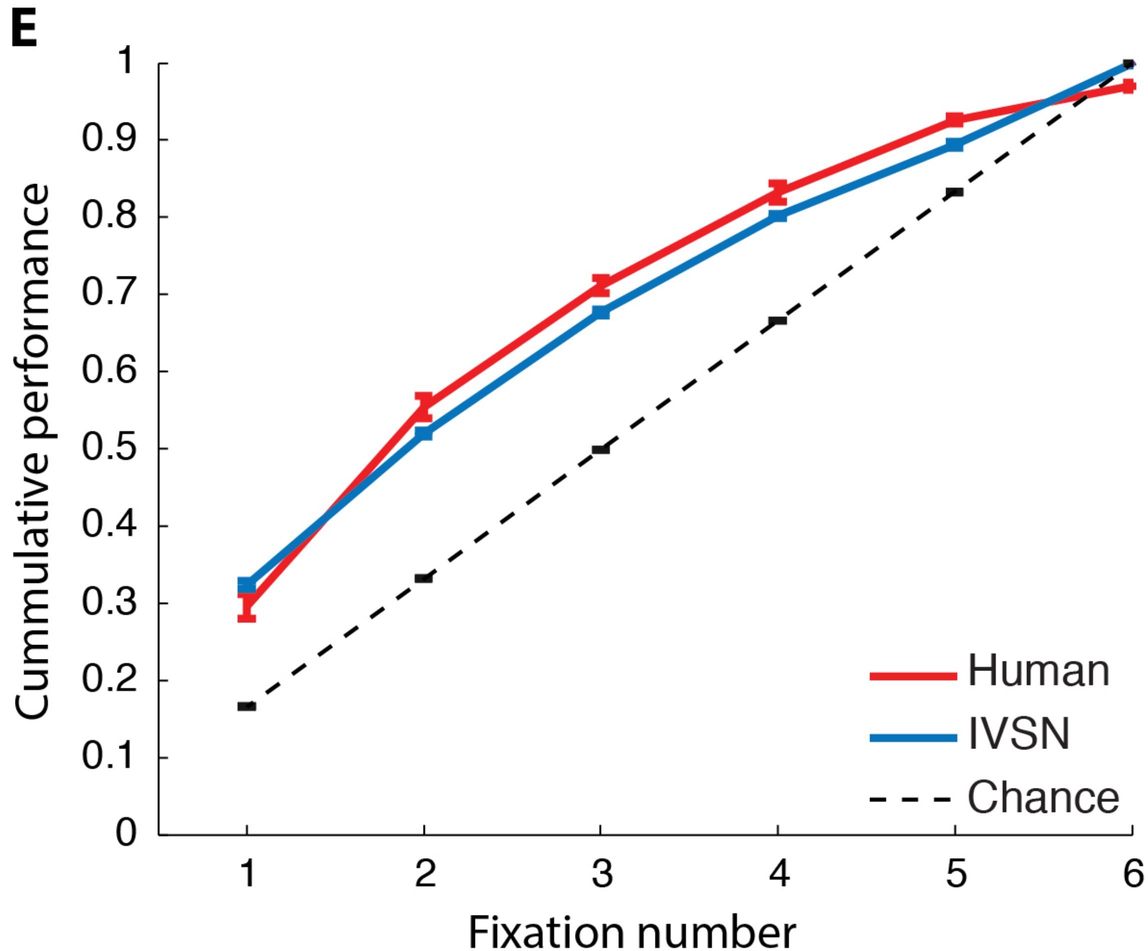
Invariant Visual Search Network (IVSN)



VGG16 (Simonyan et al 2014)
Neural circuit for visual search: e.g. Bichot et al (2015)
Neural circuitry along ventral visual cortex: e.g., Connor (2007)

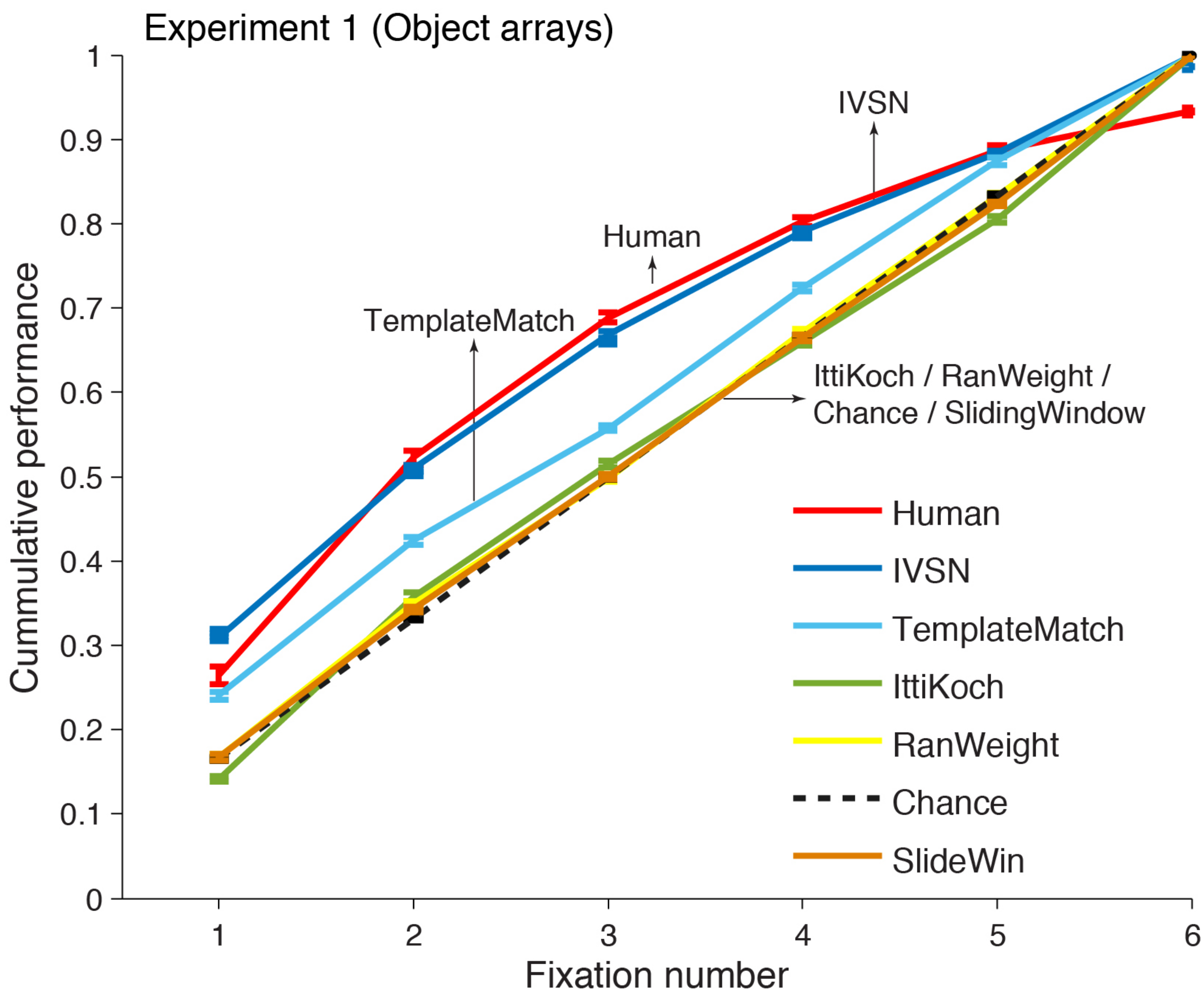
Experiment 1: Object arrays



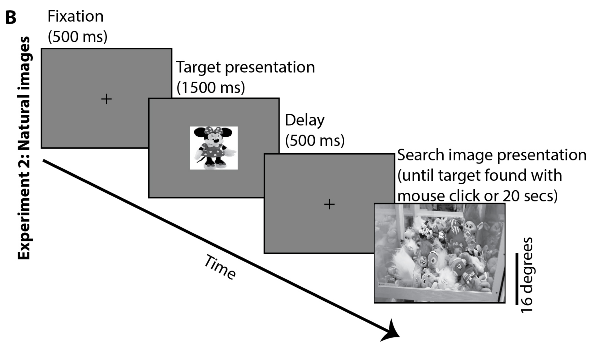


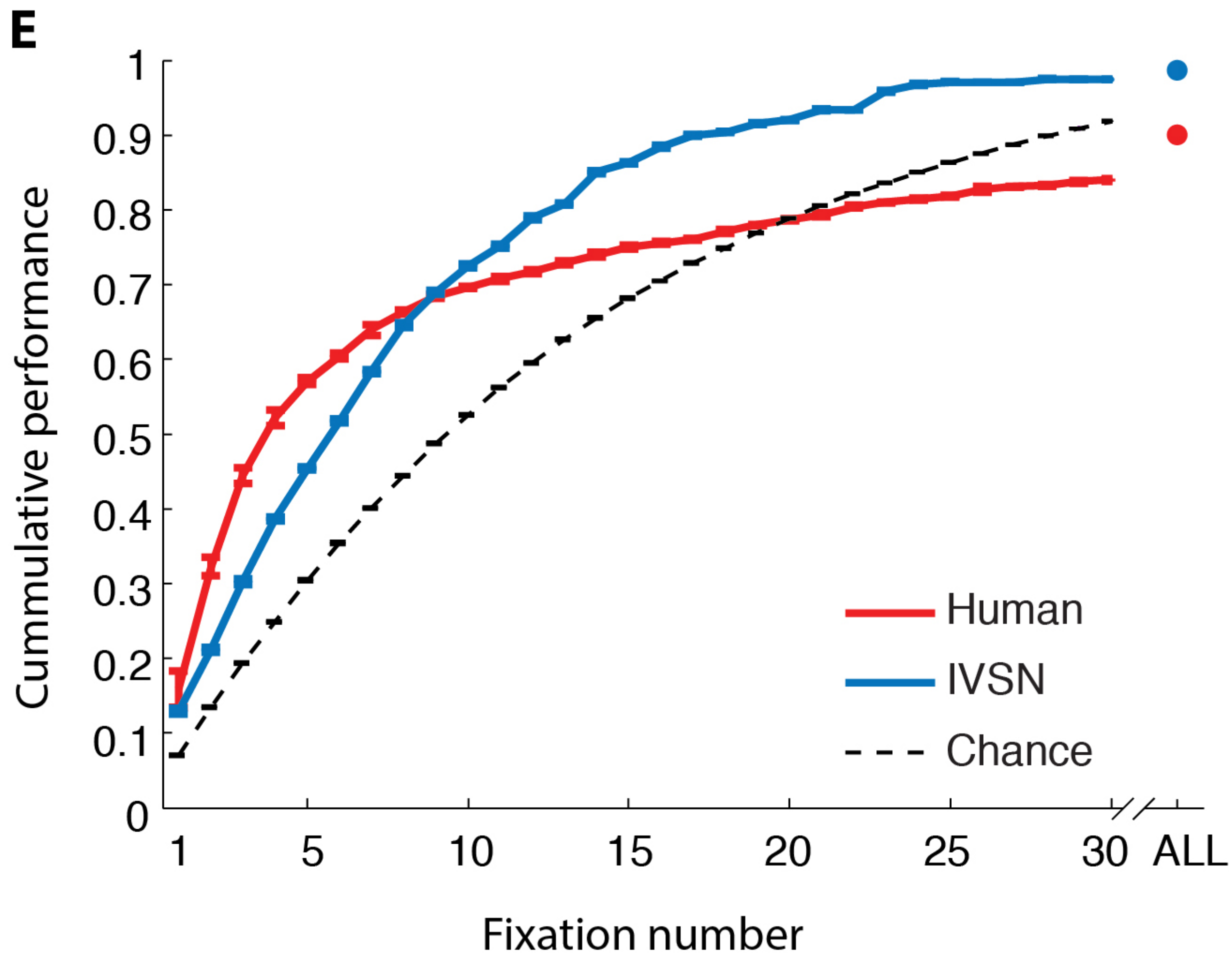
Comparison with null models

Experiment 1: Object arrays

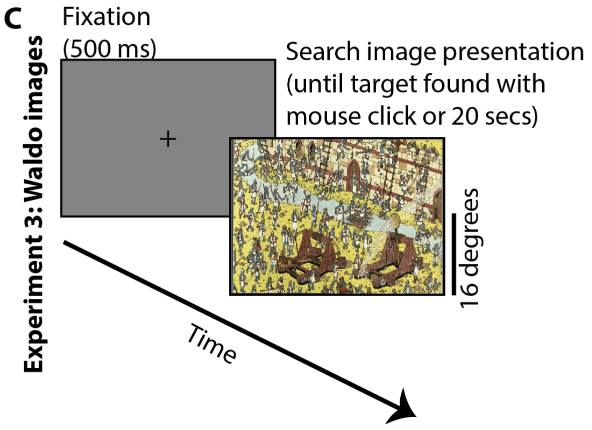


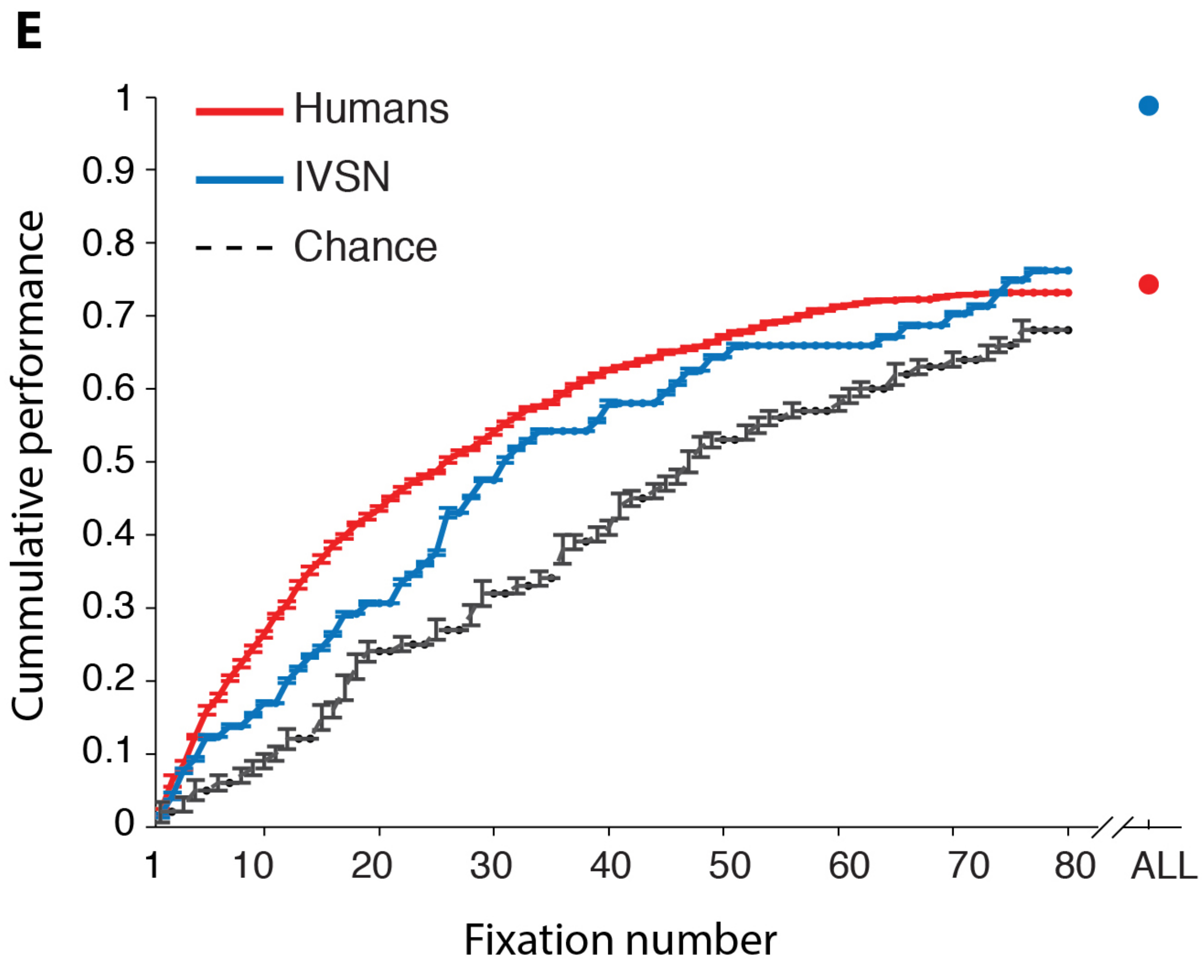
Experiment 2: Natural images





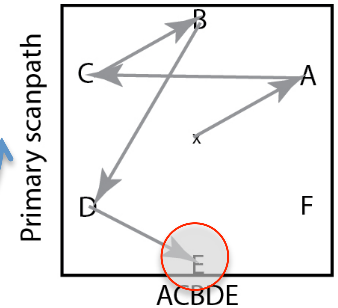
Experiment 3: Waldo images

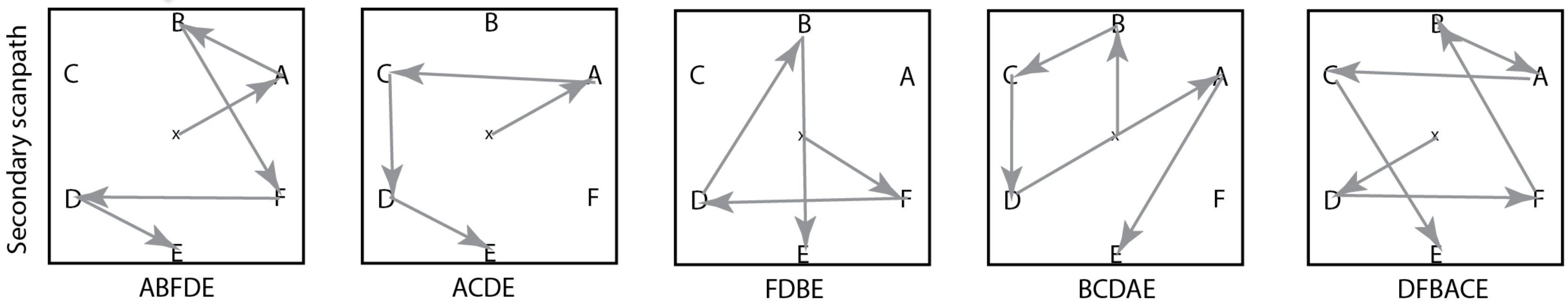




Trial-by-trial comparisons

Same #fixations,
different sequences

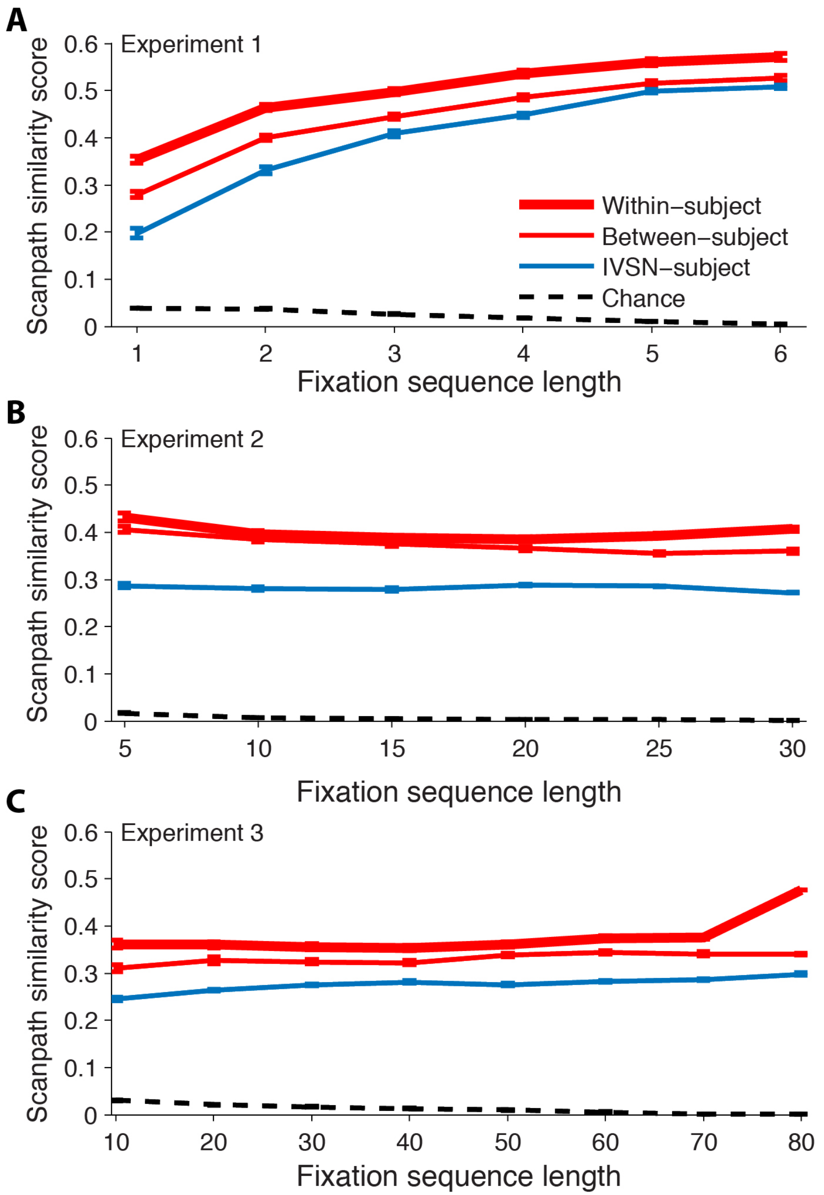




	ABFDE	ACDE	FDBE	BCDAE	DFBACE
Δ(Number of fixations)	0	1	1	0	-1
Scanpath score	0.75	0.75	0.50	0.75	0.40

Same #fixations,
Left is more similar to primary

Trial-by-trial comparisons, scanpath



Revisiting model assumptions

1. Recognition (no oracle!). *IVSN with recognition shows worse performance and is closer to humans*

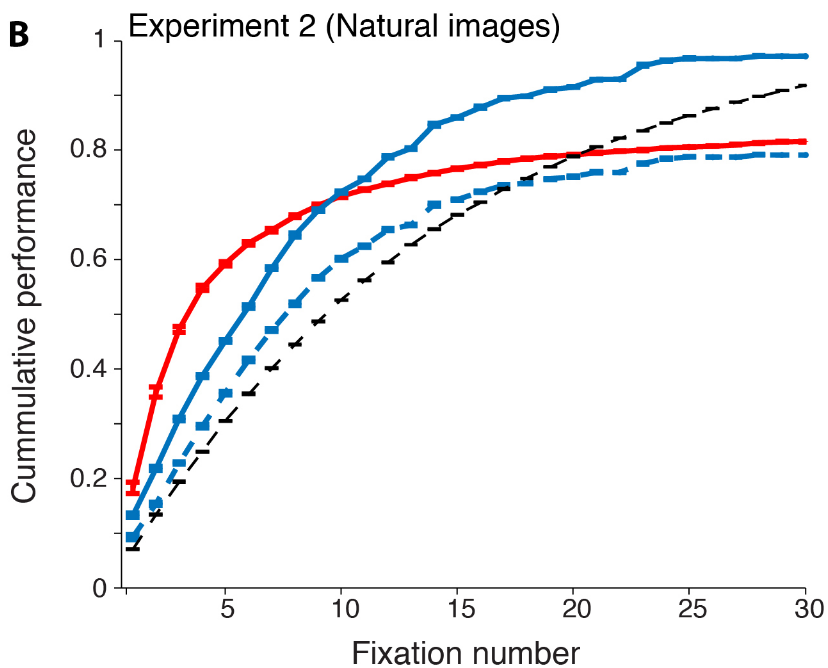
2. Finite inhibition of return. *IVSN with finite memory shows worse performance and is closer to humans*

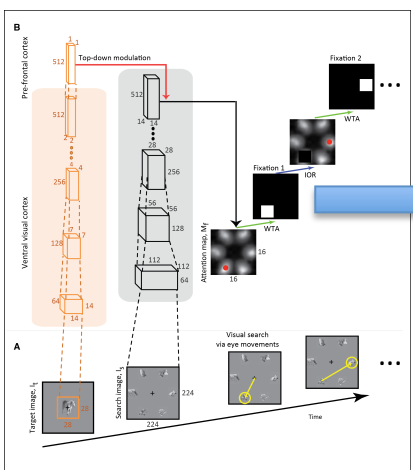
3. Restricted saccade size. *IVSN matching human saccade sizes shows the same performance*

4. Different top-down layers. *Top-down modulation can occur at multiple levels (probably all of them!)*

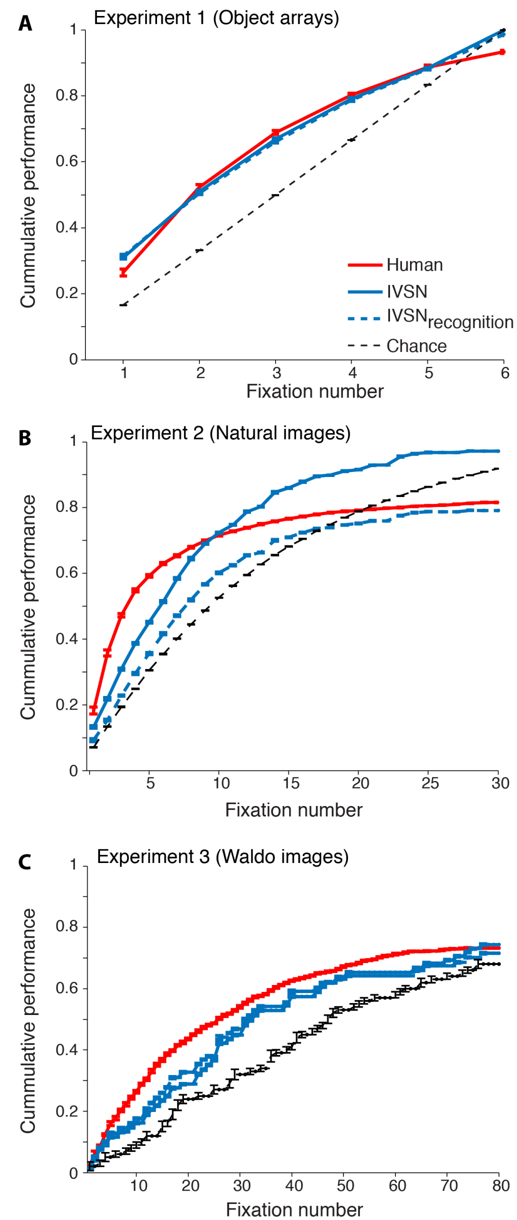
5. Other architectures. *Other “ventral visual cortex” architectures work just as well.*

Relaxing model assumptions: No oracle

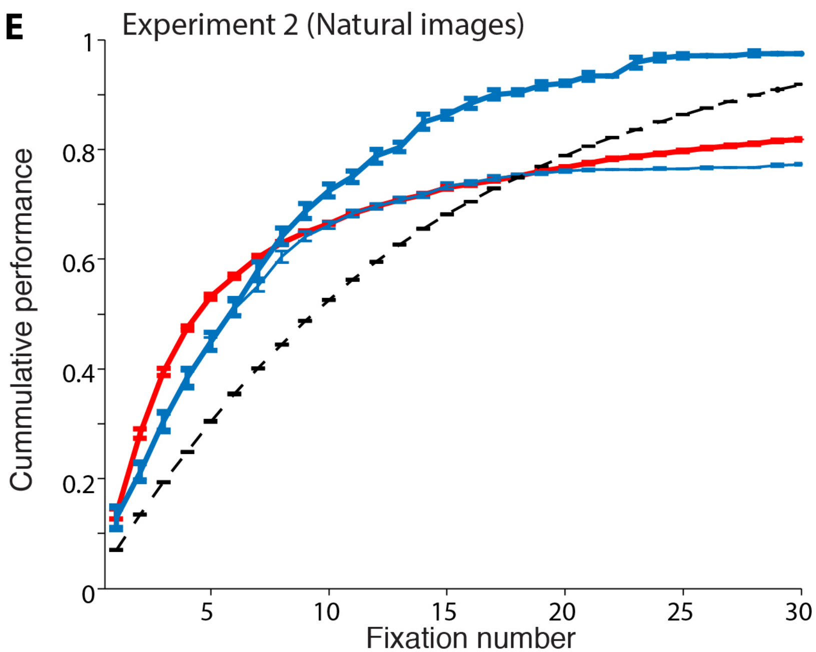


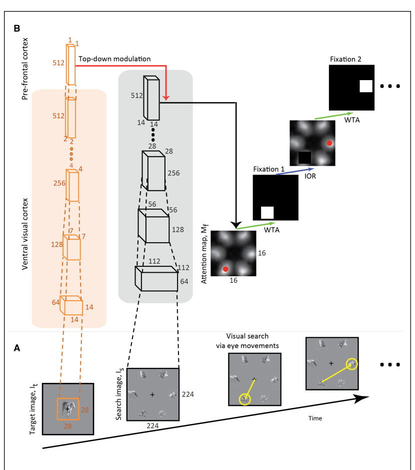


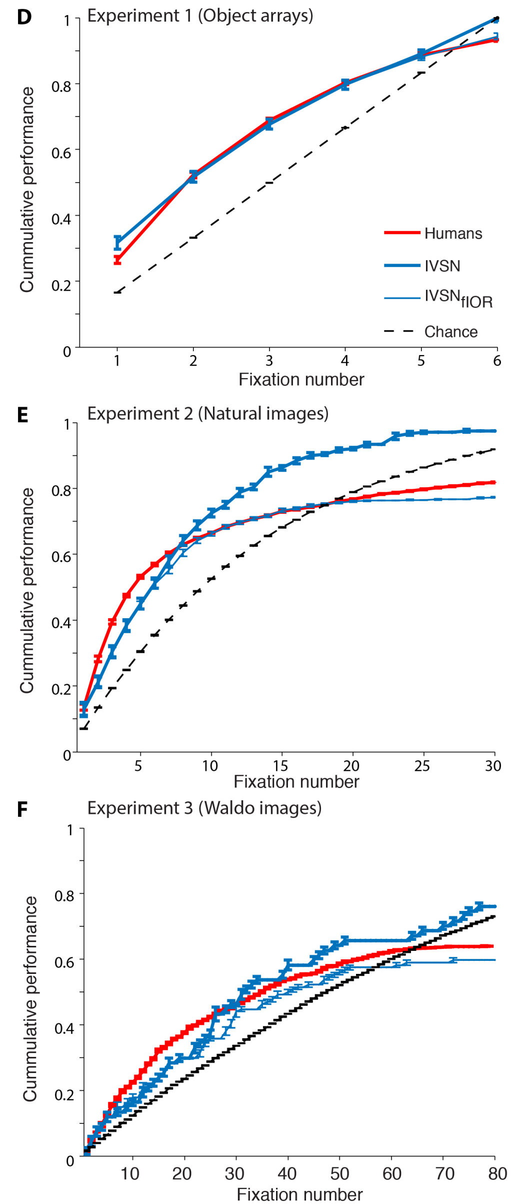
Recognition



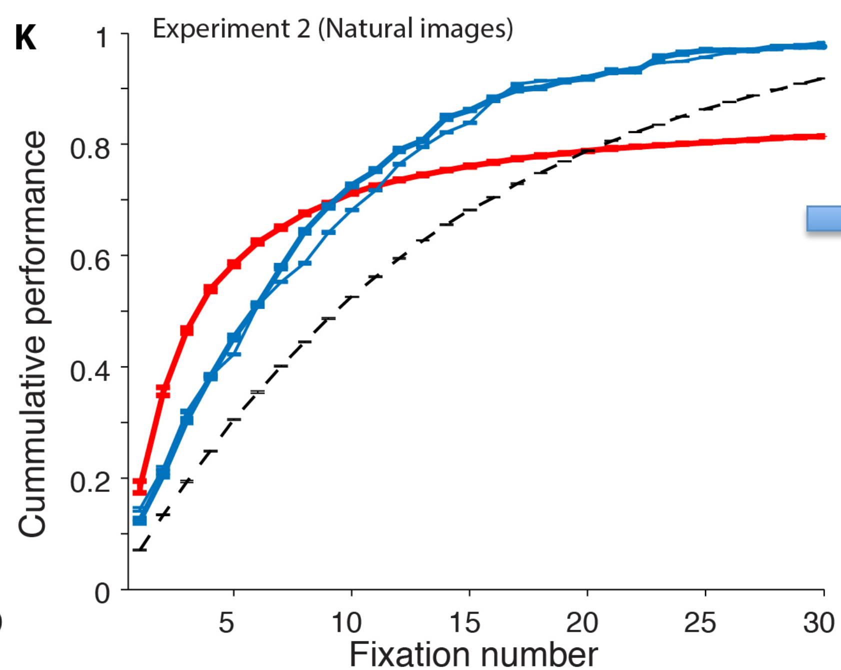
Relaxing model assumptions: Finite inhibition of return

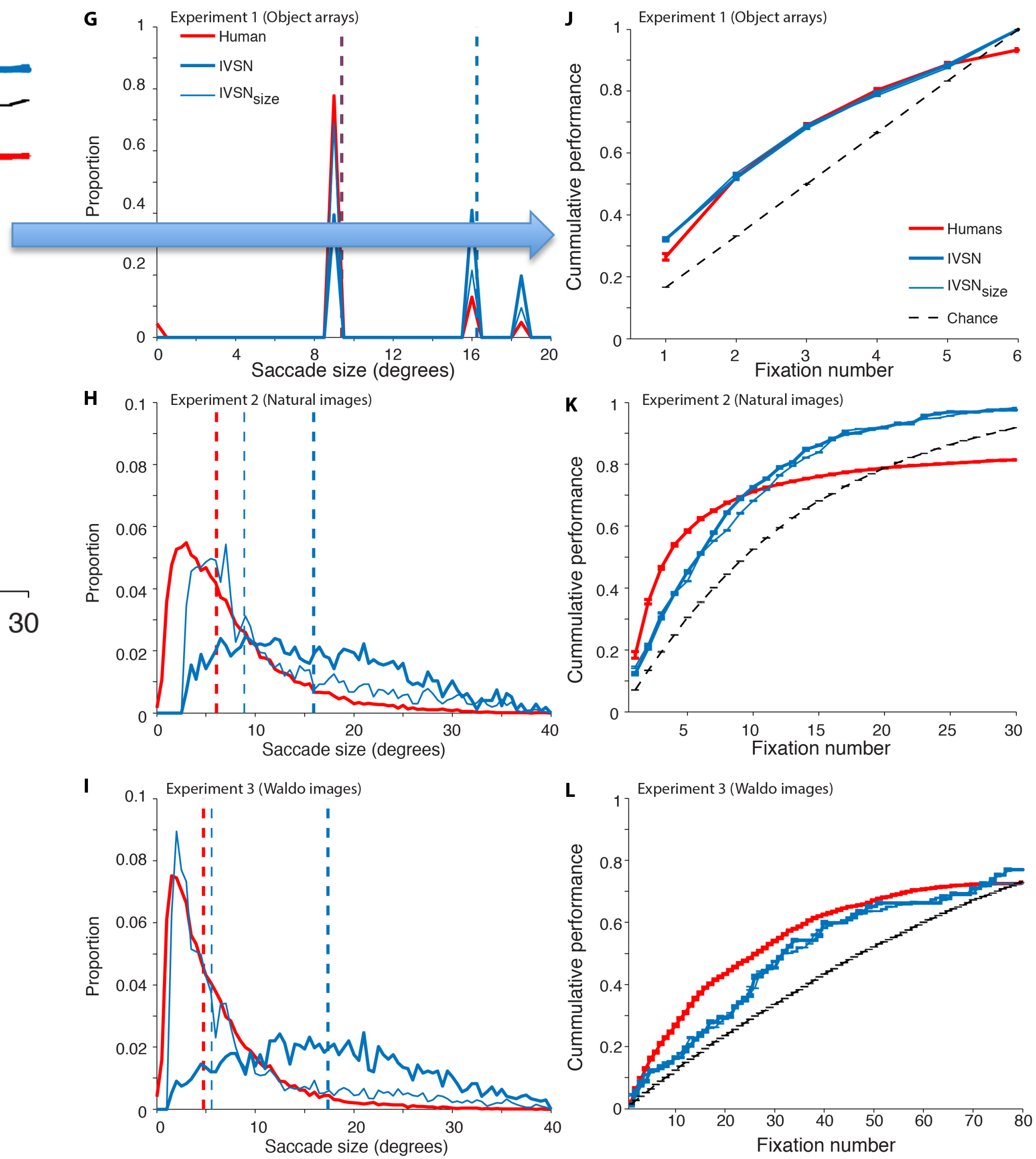




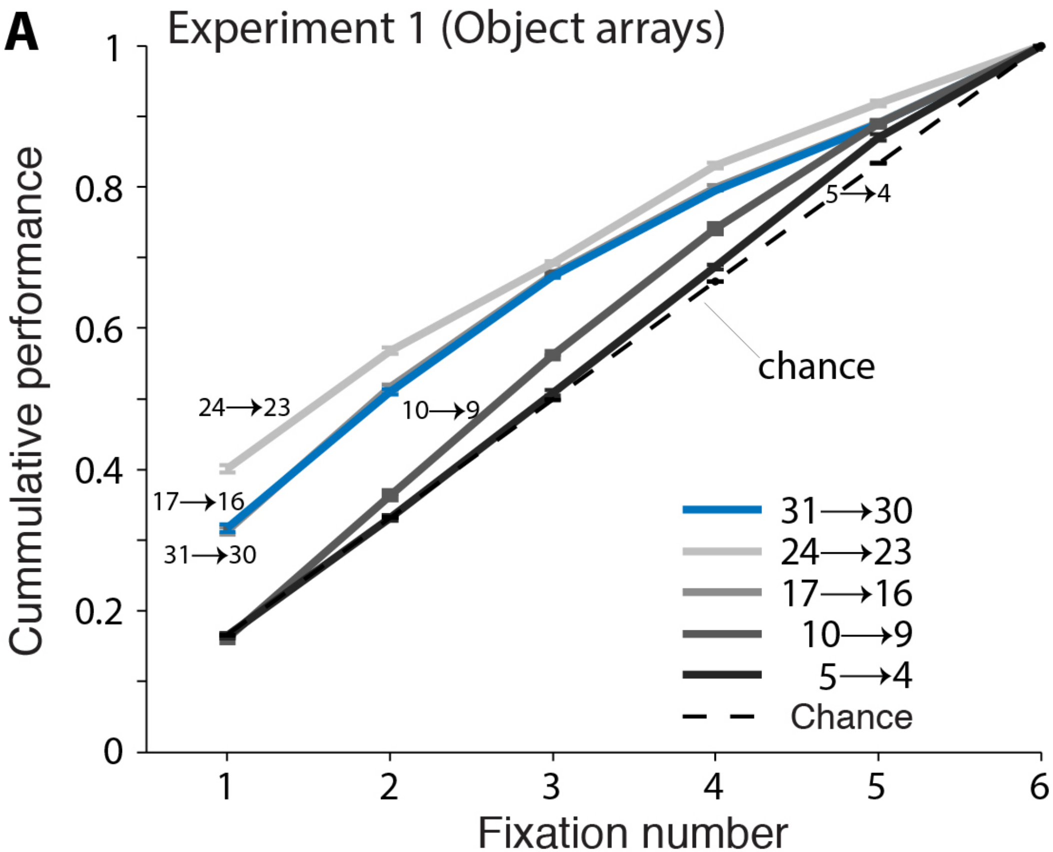


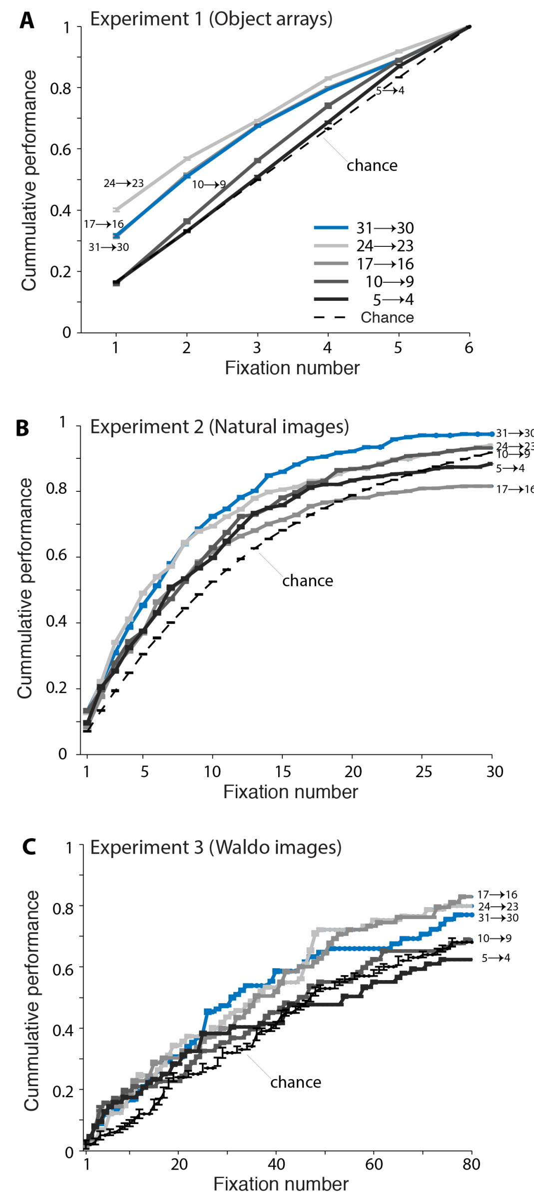
Relaxing model assumptions: Small saccade sizes



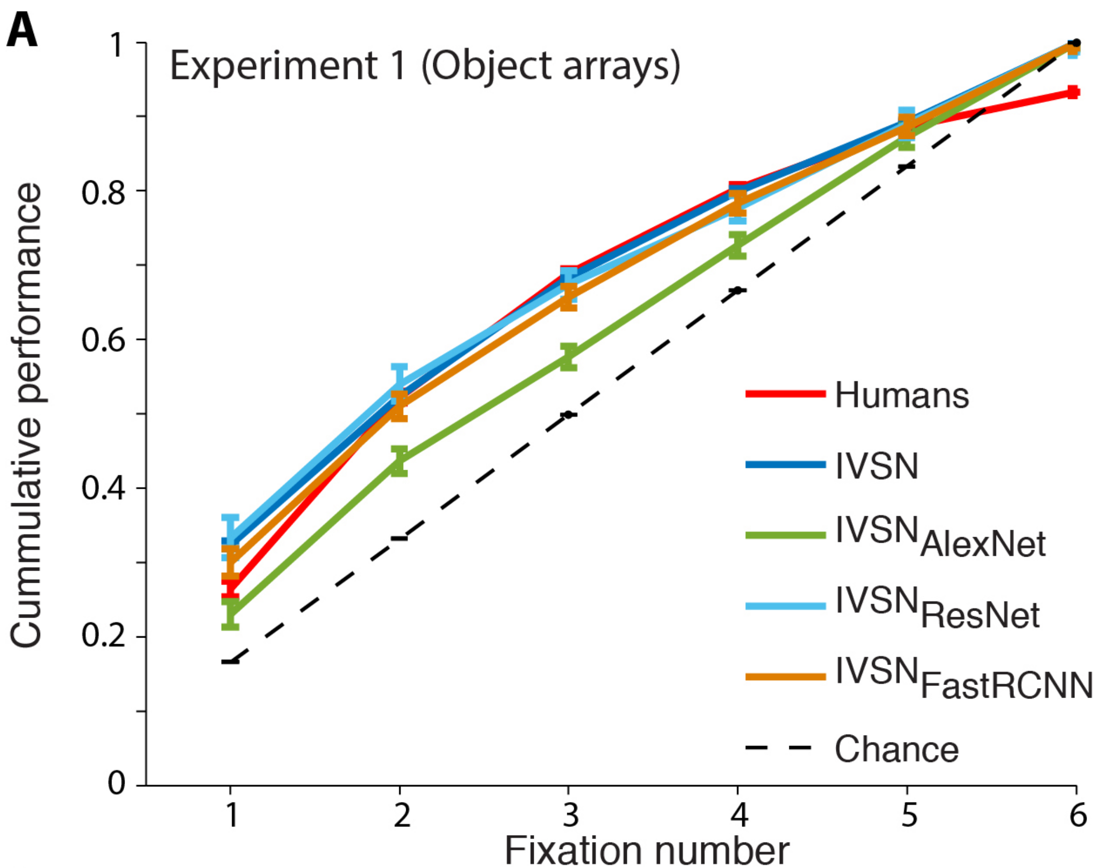


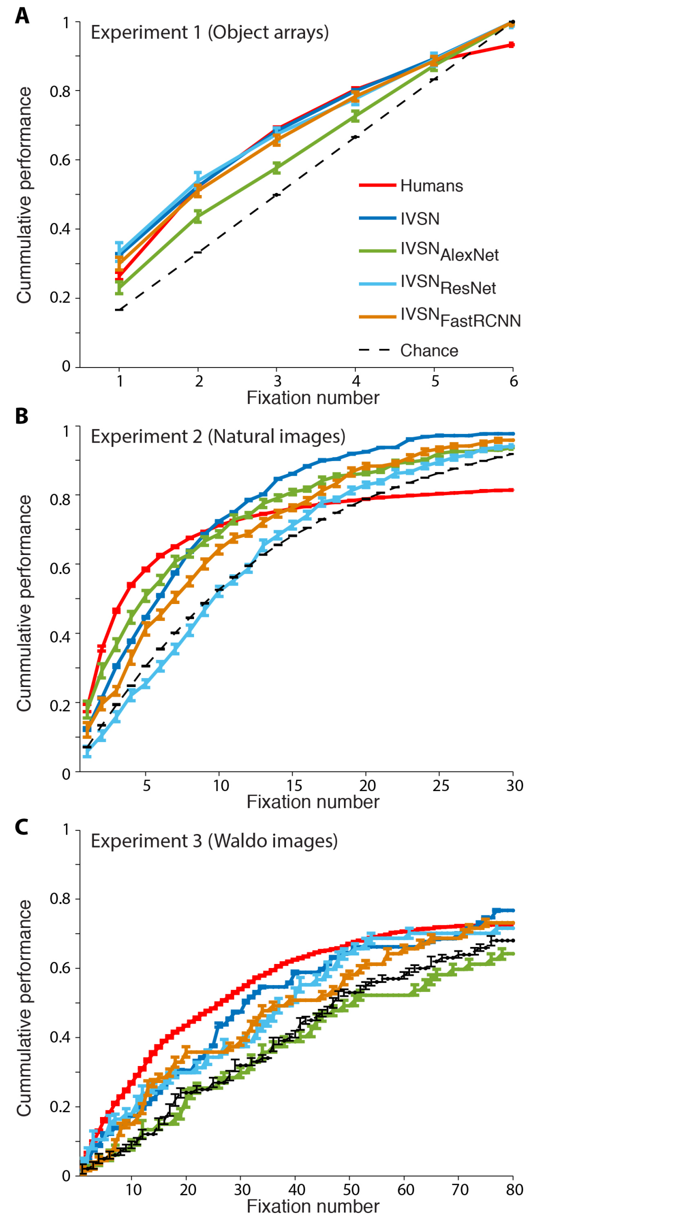
Top-down modulation at different levels (mostly) works as well





Other “ventral visual cortex” architectures (mostly) work as well





Interim summary 3

Humans show the 4 key properties of visual search: selectivity, invariance, efficiency, generalization

Invariant Visual Search Network (IVSN) model:

- 0 free parameters
- Neurobiologically inspired architecture
- Target-dependent feature-based top-down signals
- First-order approximation to human visual search (number of fixations, cumulative performance, spatiotemporal pattern of fixations)

There is much more to visual search: high-level contextual information, recognition, temporal integration, memory

Philosophical remarks

Showing that a model can be (over)trained to

perform a certain task (Computer Vision)

match human behavior (Cognitive Science)

post-dict neural data (Neuroscience)

is "necessary" but not sufficient

We need to explain computation, algorithms and hardware (Marr/Poggio)

Working hypothesis

1. VisualSampling
2. RapidPeripheralAssessment
3. FovealRecognition
4. PatternCompletion
5. VisualBuffer
6. TargetAttentionProposal
7. EyeMovementImplementation
8. PeopleDetection
9. SpatialRelationships
10. Repeat steps 3+4+5
11. Repeat steps 6-7
12. Repeat 8-9
13. TaskTerminationDecision
14. TaskReport

- Need to put all the routines together and flexibly call them for each task
- List of routines probably not exhaustive
- We will need high level world knowledge

http://klab.tch.harvard.edu

The brain's operating system

Gabriel Kreiman