Bottom-up and top-down processing in visual perception

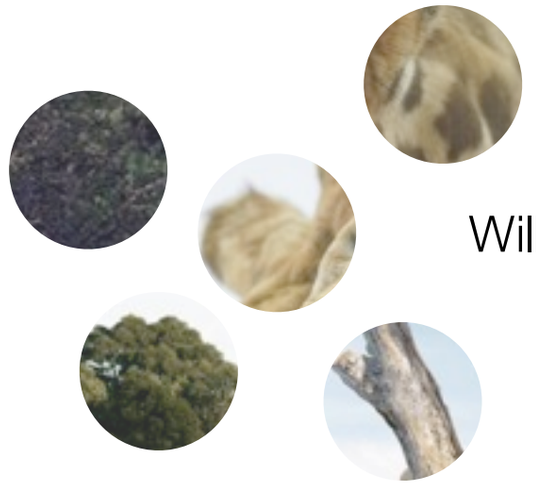
Thomas Serre

Brown University

Department of Cognitive & Linguistic Sciences
Brown Institute for Brain Science
Center for Vision Research

The ‘what’ problem

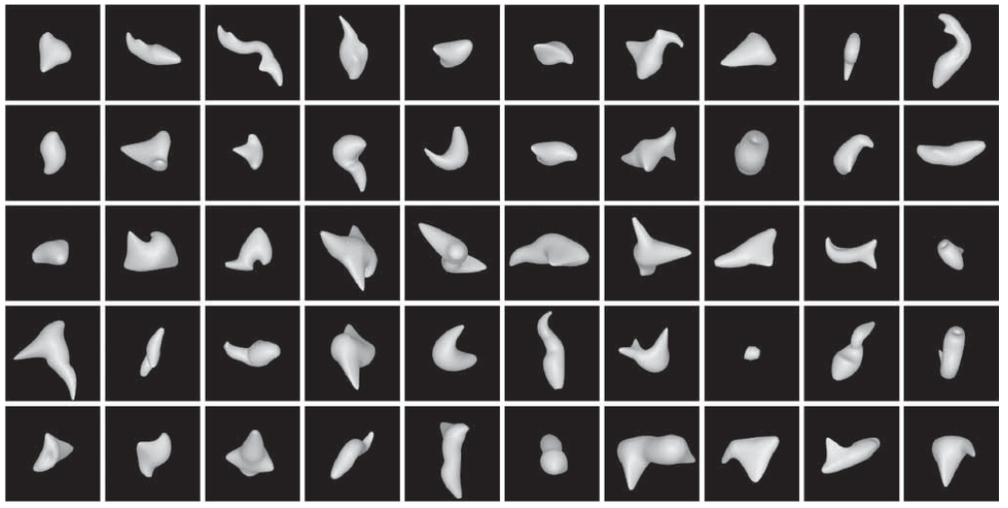
monkey electrophysiology



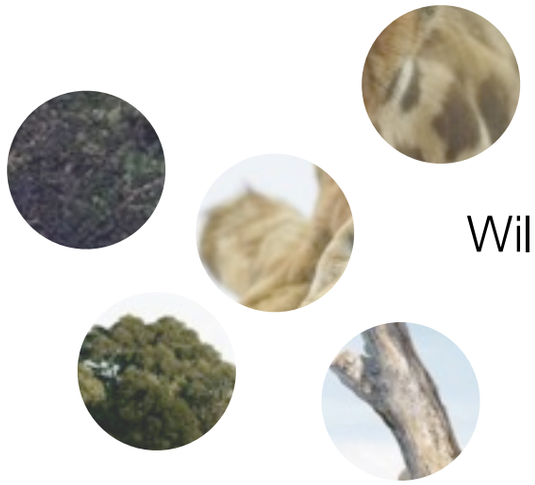
Willmore Prenger & Gallant ’10

The ‘what’ problem

monkey electrophysiology



Yamane Carlson Bowman Wang & Connor ’08

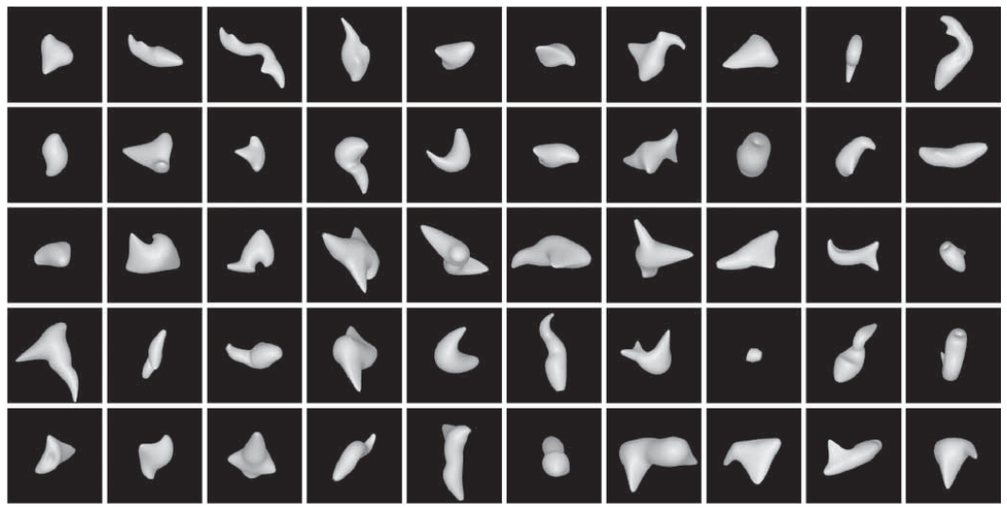


Willmore Prenger & Gallant ’10

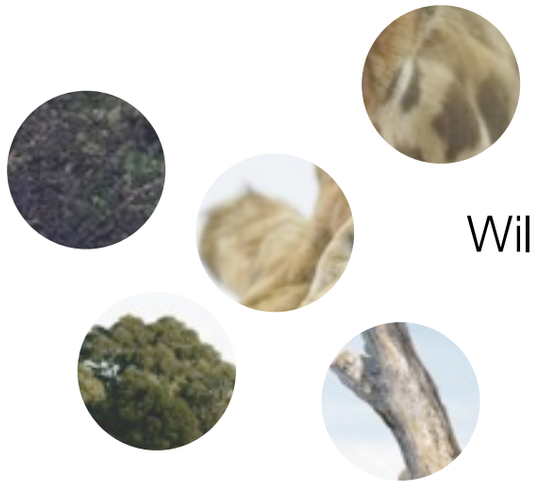
The ‘what’ problem

monkey electrophysiology

Hung Kreiman et al ’05



Yamane Carlson Bowman Wang & Connor ’08



Willmore Prenger & Gallant ’10

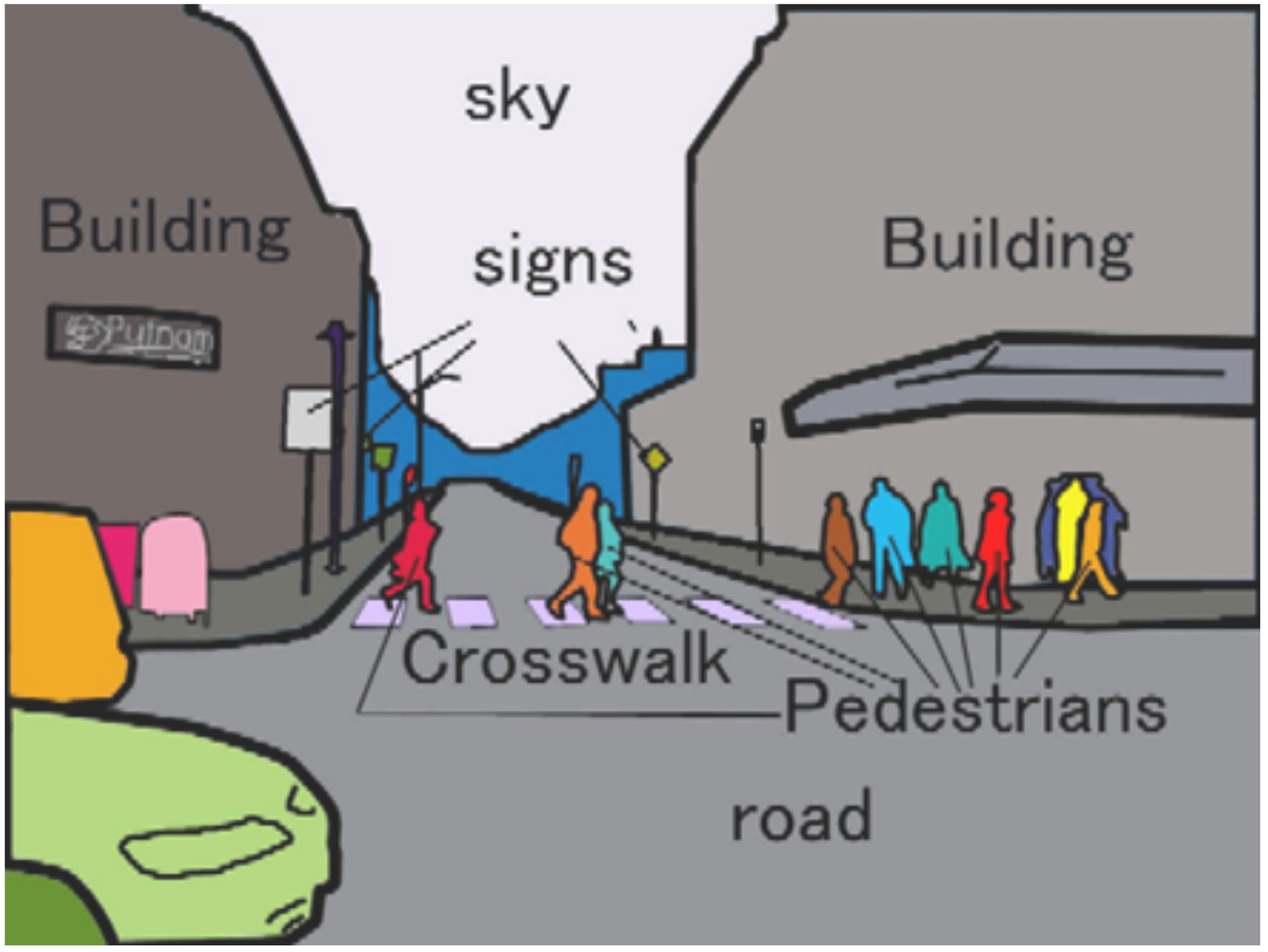
The ‘what’ problem

monkey electrophysiology

The ‘what’ problem

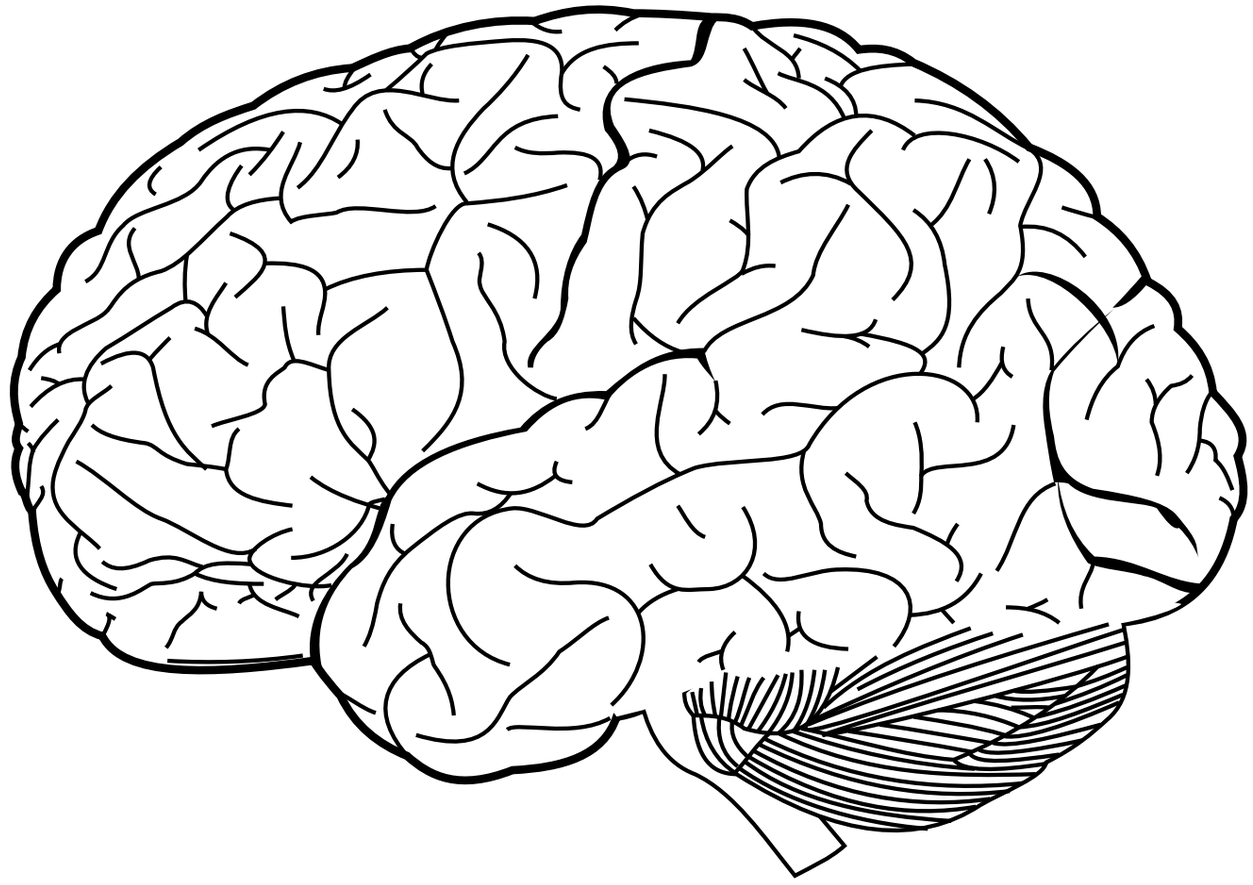
human psychophysics

Thorpe et al ’96; VanRullen & Thorpe ’01; Fei-Fei et al ’02 ’05; Evans & Treisman ’05; Serre Oliva & Poggio ’07



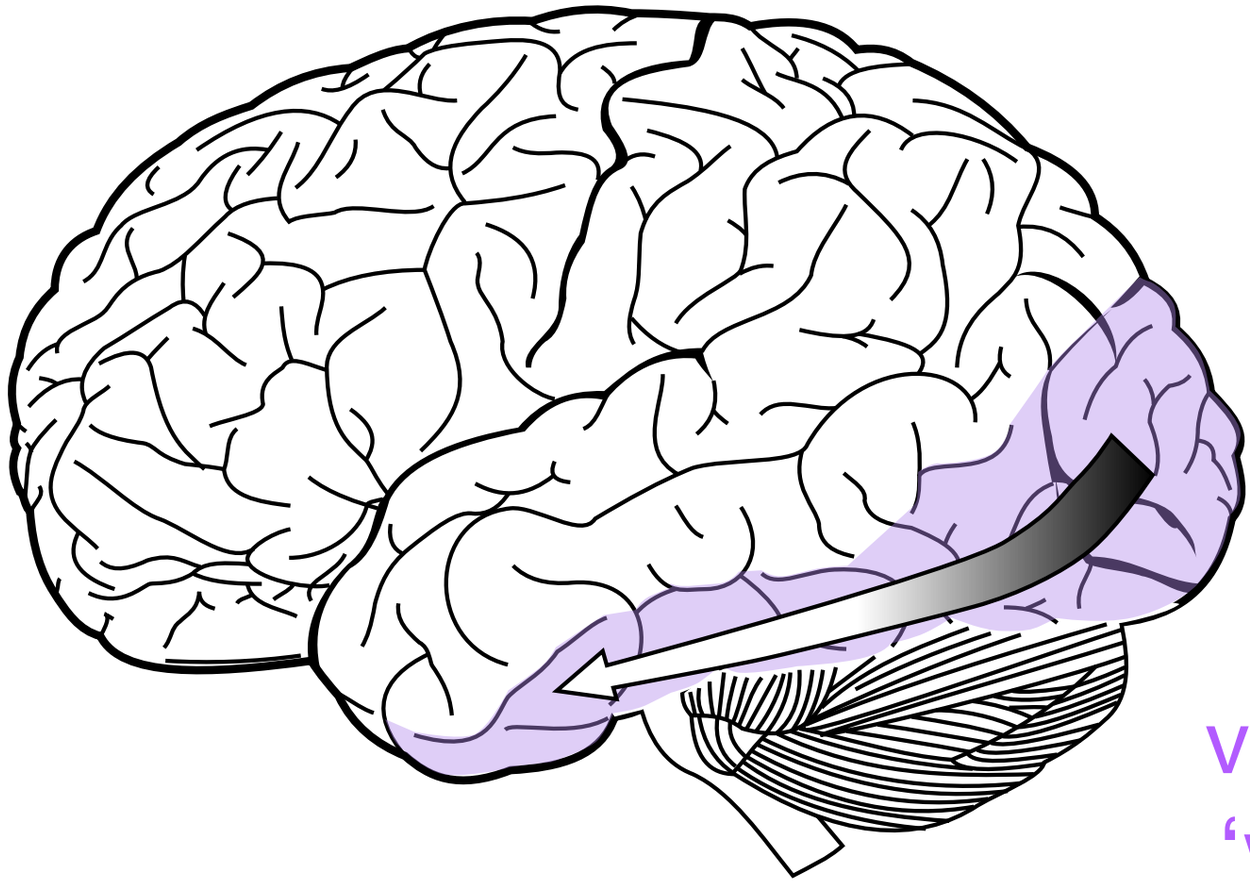
Vision as ‘knowing what is where’

Aristotle; Marr ’82



‘What’ and ‘where’ cortical pathways

Ungerleider & Mishkin ‘84

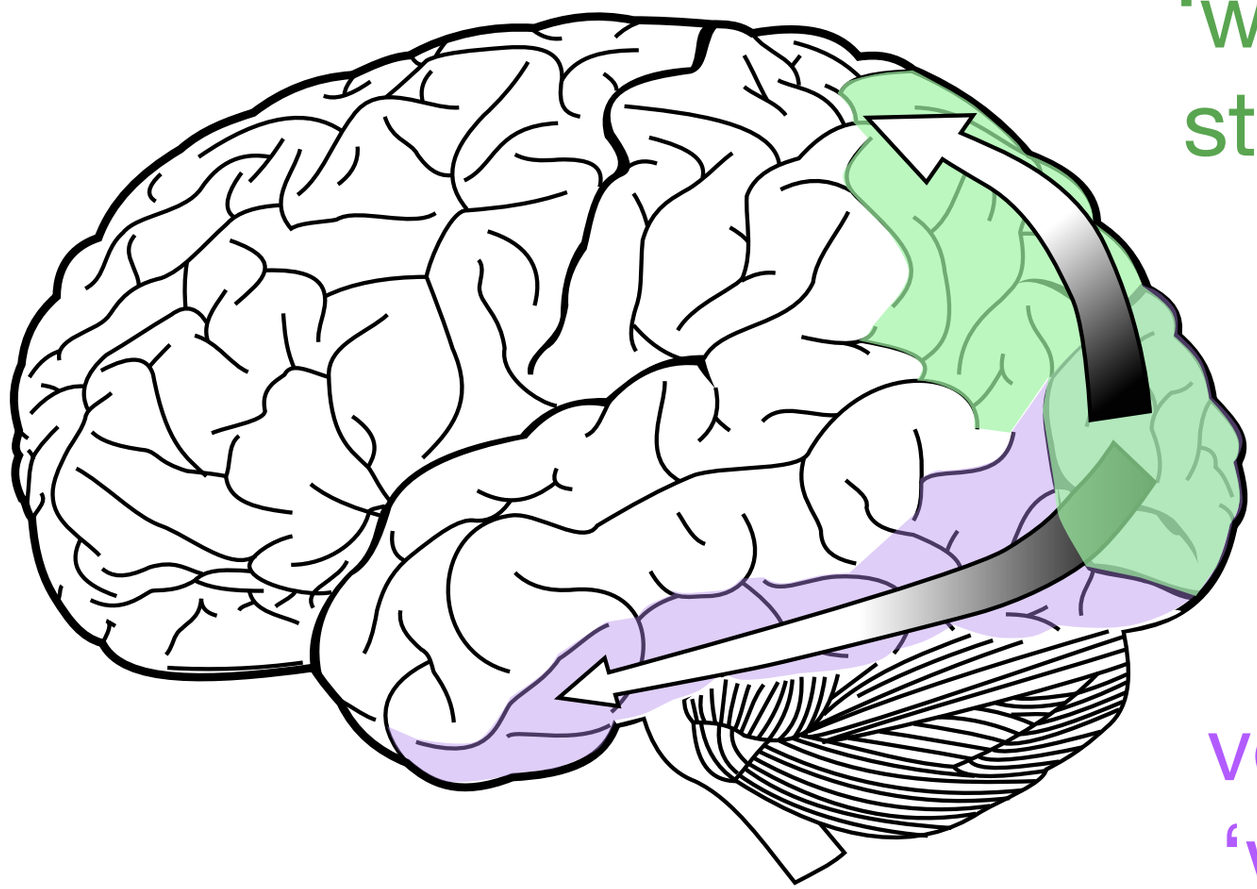


ventral
‘what‘
stream

‘What’ and ‘where’
cortical pathways

Ungerleider & Mishkin ‘84

dorsal
‘where’
stream

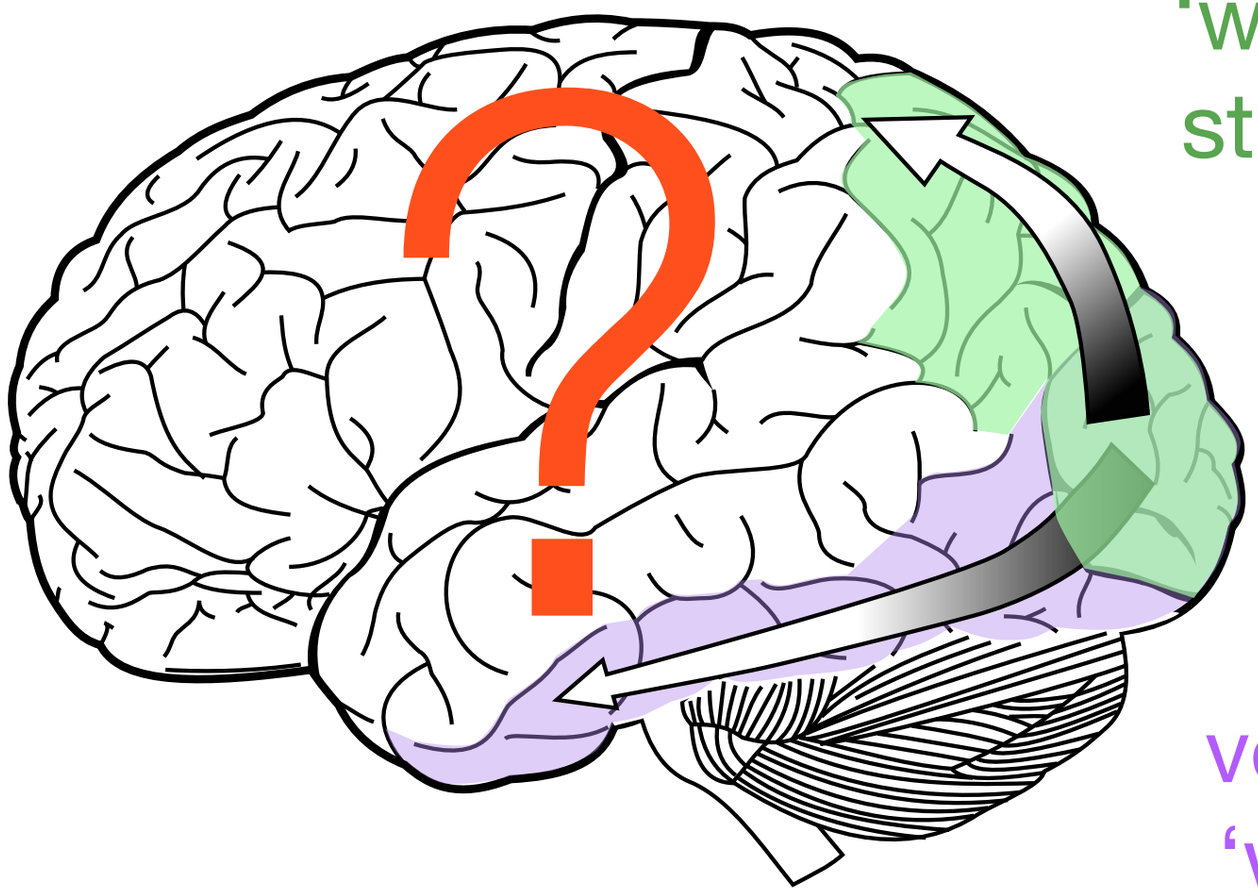


ventral
‘what‘
stream

‘What’ and ‘where’
cortical pathways

Ungerleider & Mishkin ‘84

How does the visual system combine information about the identity and location of objects?



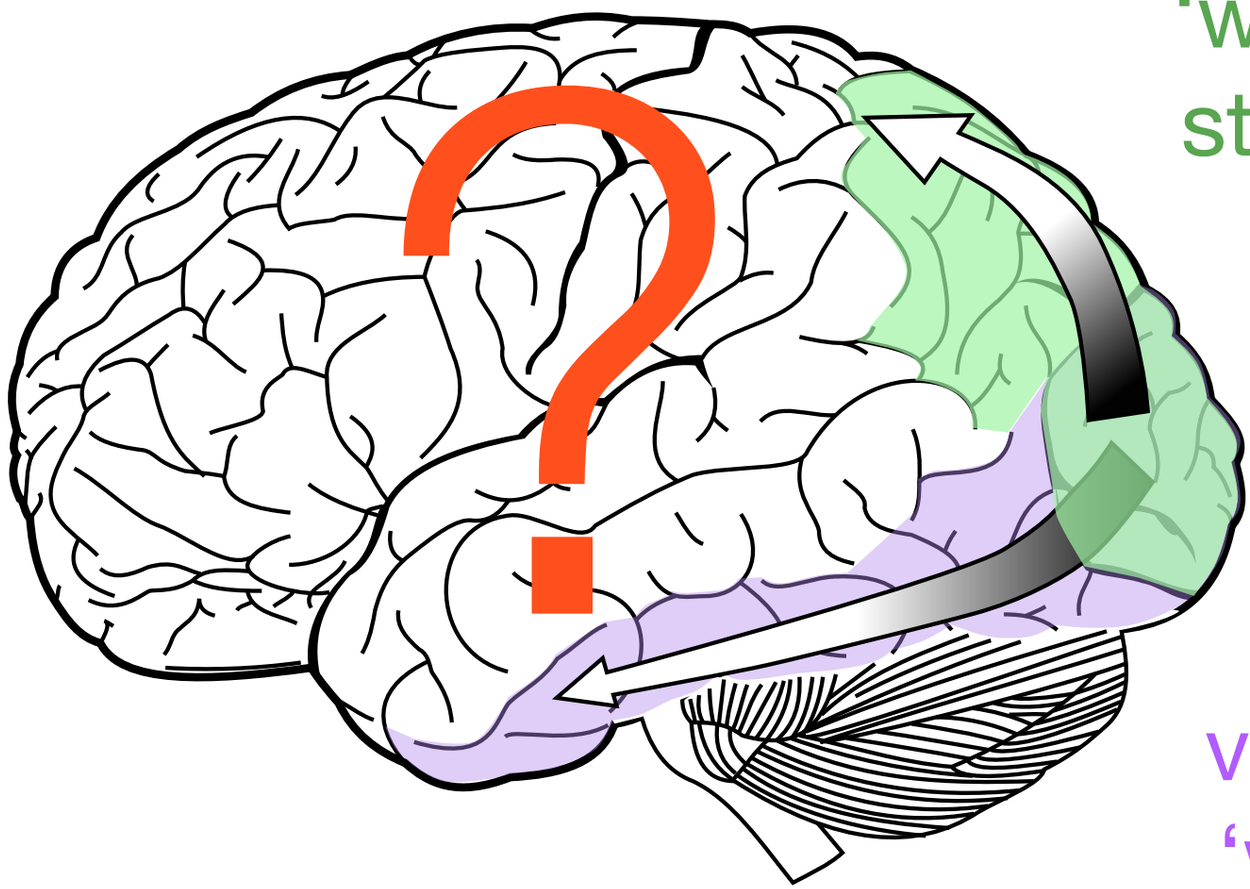
‘What’ and ‘where’ cortical pathways

Ungerleider & Mishkin ‘84

How does the visual system combine information about the identity and location of objects?

➡ Central thesis: visual attention

(see also Van Der Velde and De Kamps ‘01; Deco and Rolls ‘04)



‘What’ and ‘where’ cortical pathways

Ungerleider & Mishkin ‘84

Part I: A ‘Bayesian’ theory of attention that solves the ‘what is where’ problem

- Predictions for neurophysiology and human eye movement data

- Sharat Chikkerur
- Cheston Tan
- Tomaso Poggio

Part II: Readout of IT population activity

- Attention eliminates clutter

- Ying Zhang
- Ethan Meyers
- Narcisse Bichot
- Tomaso Poggio
- Bob Desimone

Part I: A ‘Bayesian’ theory of attention that solves the ‘what is where’ problem

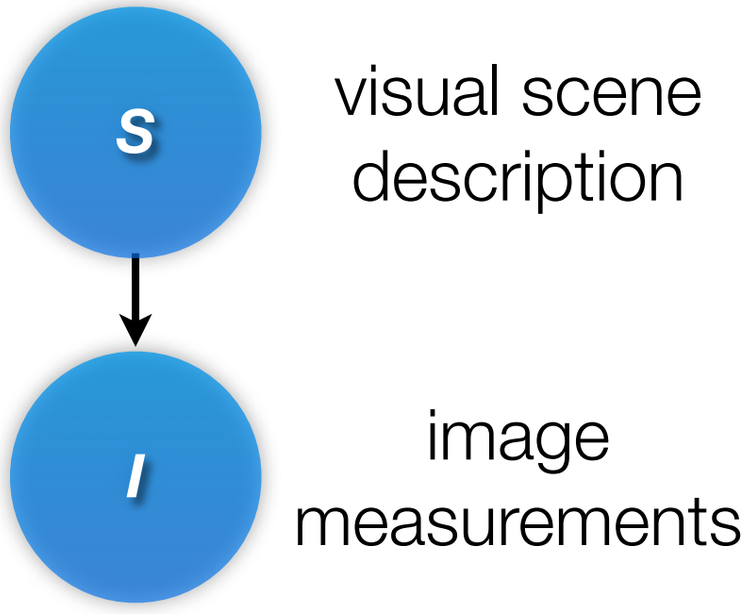
- Predictions for neurophysiology and human eye movement data

- Sharat Chikkerur
- Cheston Tan
- Tomaso Poggio

Part II: Readout of IT population activity

- Attention eliminates clutter

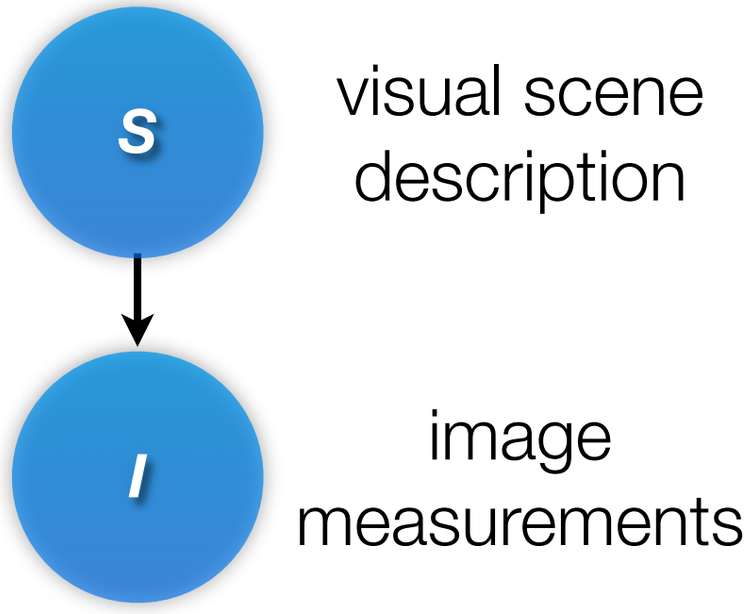
- Ying Zhang
- Ethan Meyers
- Narcisse Bichot
- Tomaso Poggio
- Bob Desimone



Perception as Bayesian inference

Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

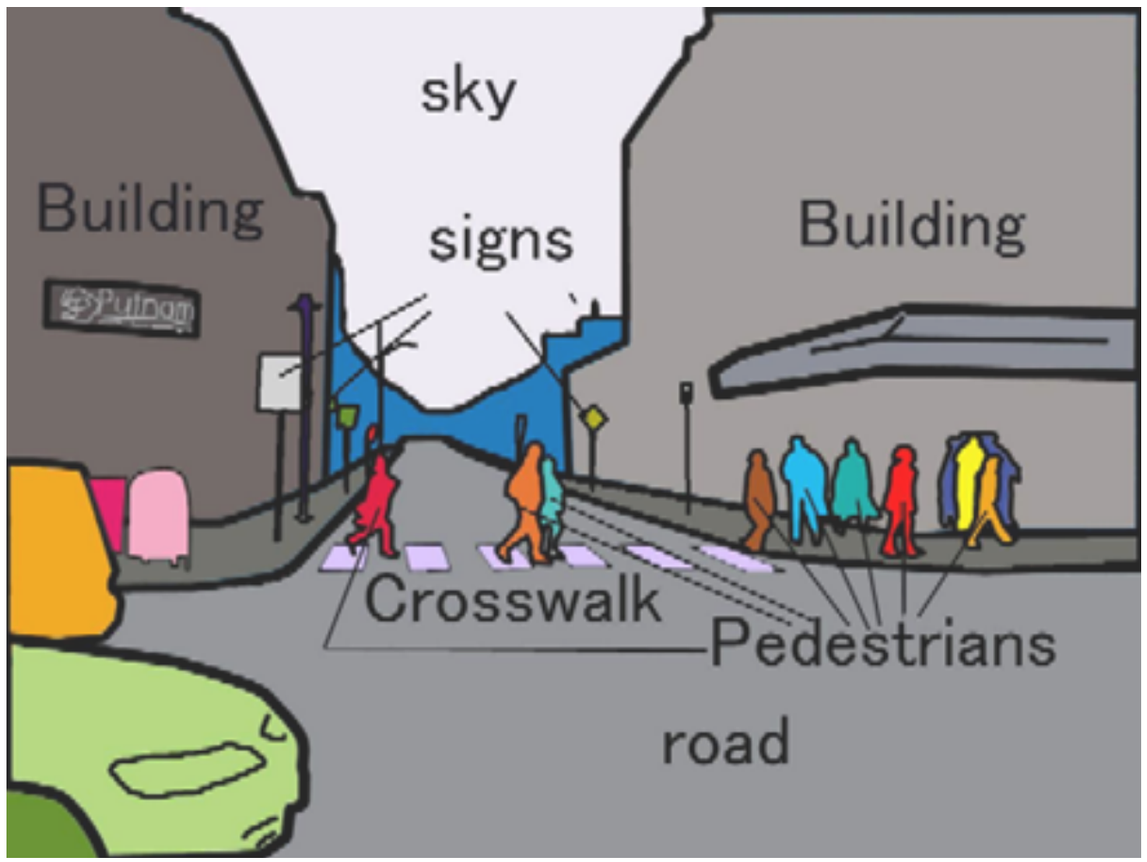
$$P(S|I) \propto P(I|S)P(S)$$

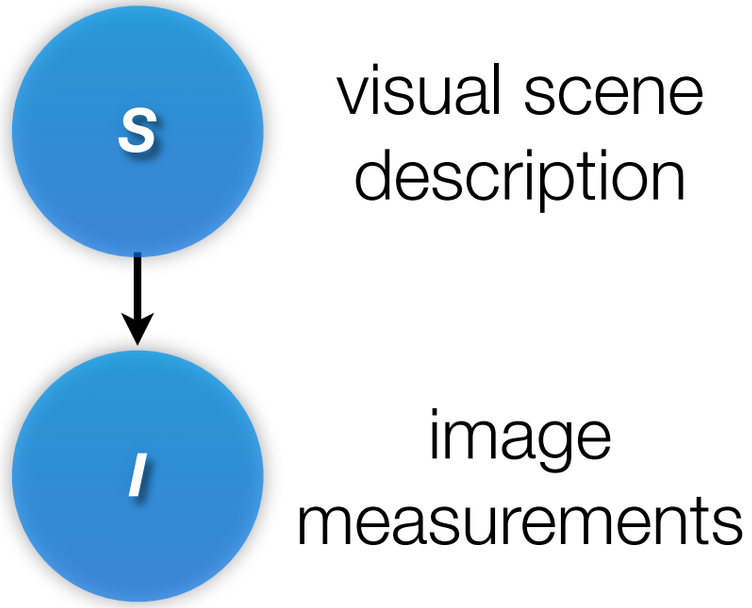


Perception as Bayesian inference

Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

$$P(S|I) \propto P(I|S)P(S)$$





Perception as Bayesian inference

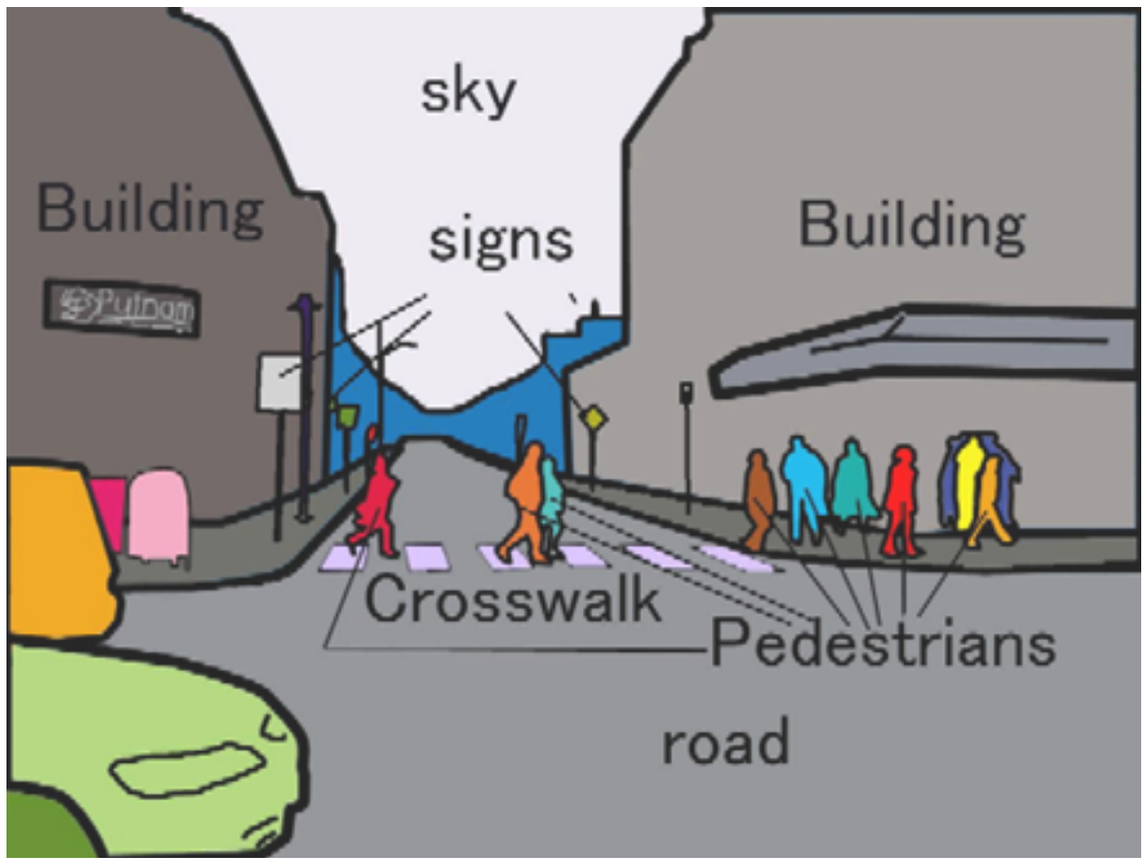
Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

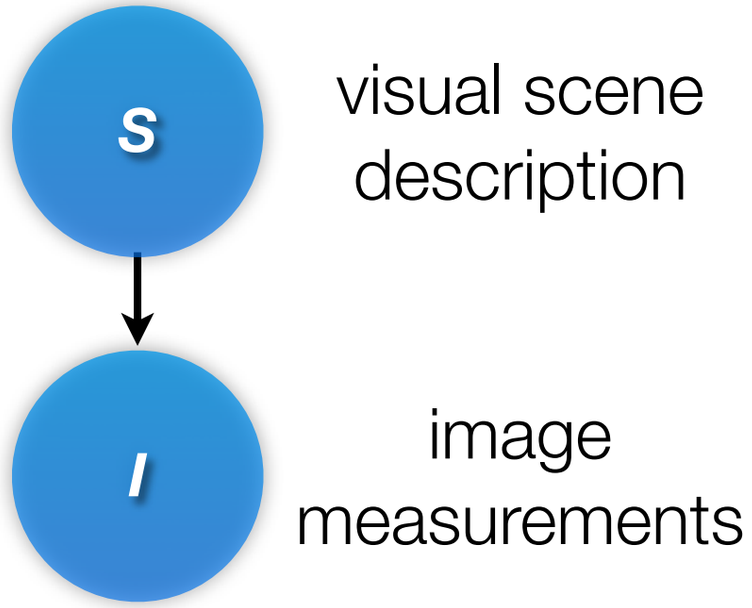
object identities

object locations

$$S = \{O_1, O_2, \cdots, O_n, L_1, L_2, \cdots, L_n\}$$

$$P(S|I) \propto P(I|S)P(S)$$





Perception as Bayesian inference

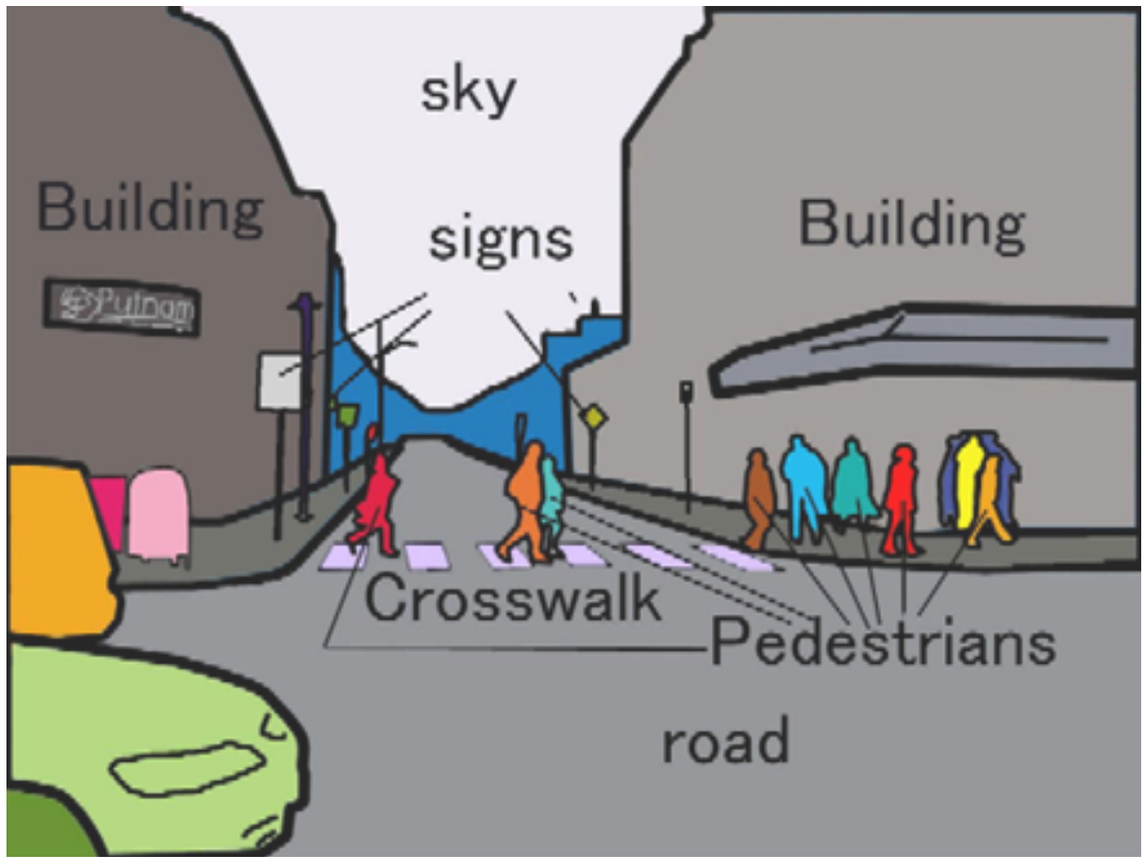
Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

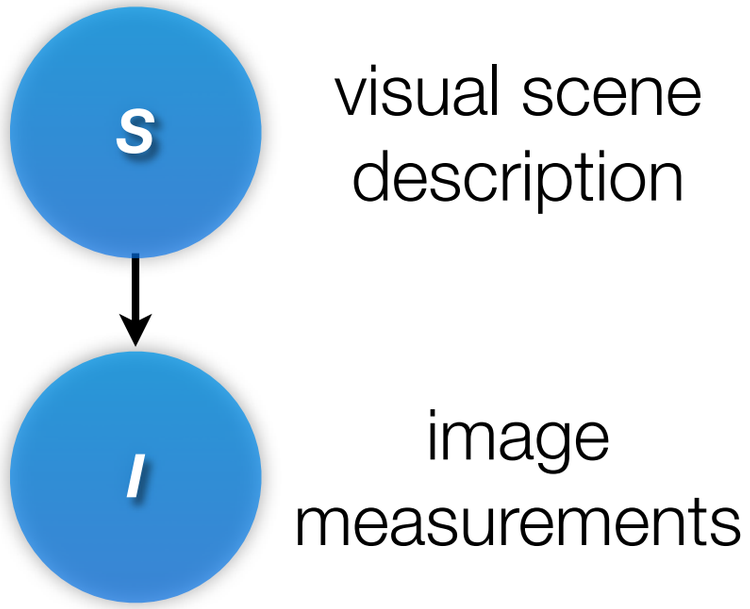
object identities

object locations

$$S = \{O_1, O_2, \cdots, O_n, L_1, L_2, \cdots, L_n\}$$

$$P(S, I) = P(O_1, L_1, O_2, L_2, \cdots, L_n, L_n, I)$$

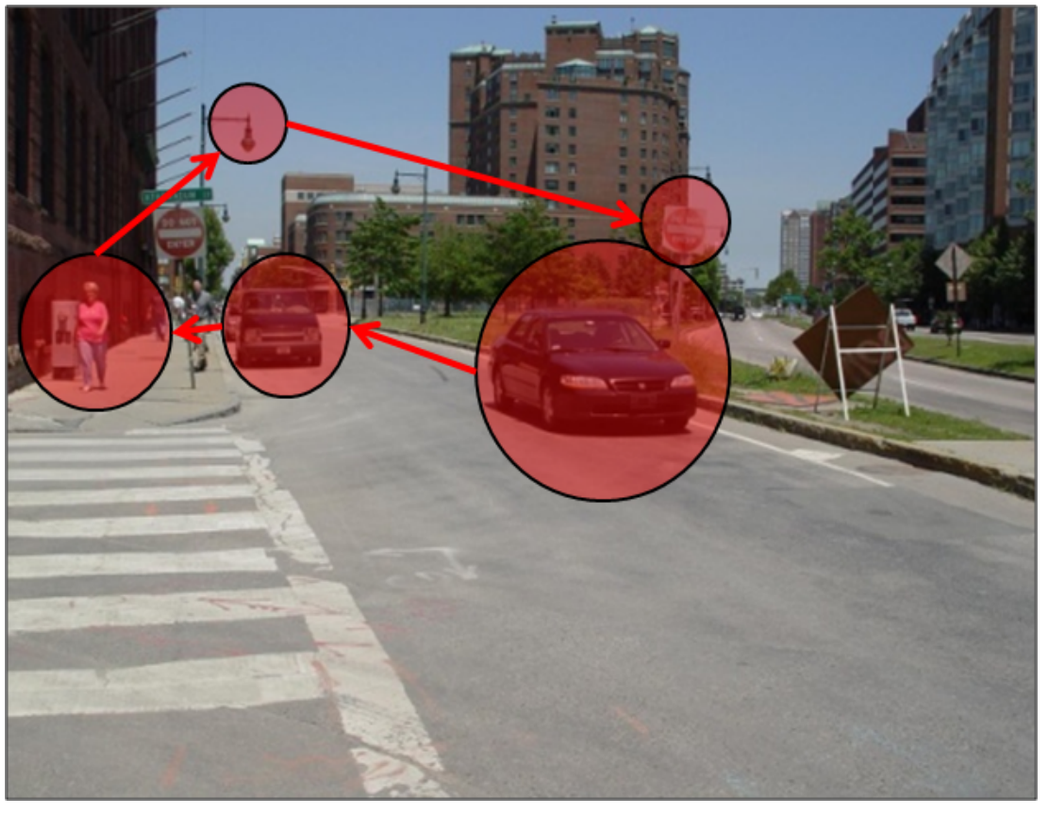




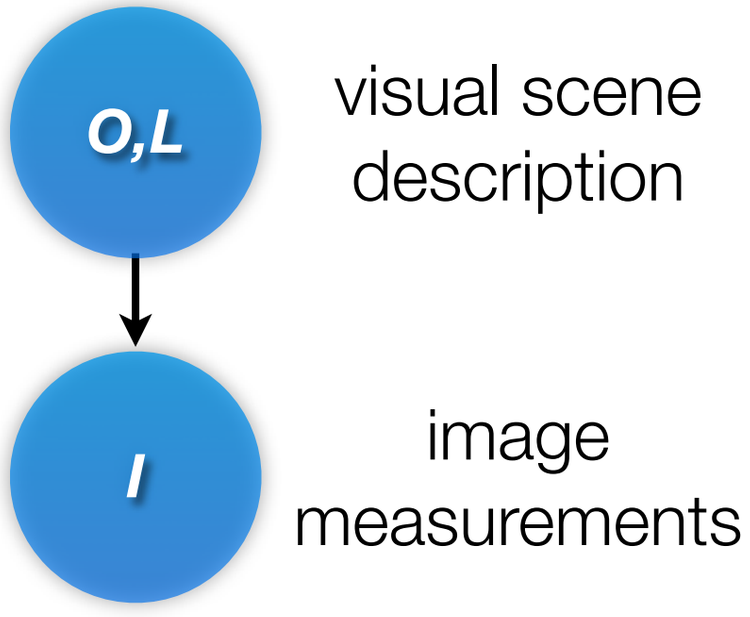
Perception as Bayesian inference

Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

- To recognize and localize objects in the scene, the visual system selects objects, one object at a time



$$P(O, L, I)$$

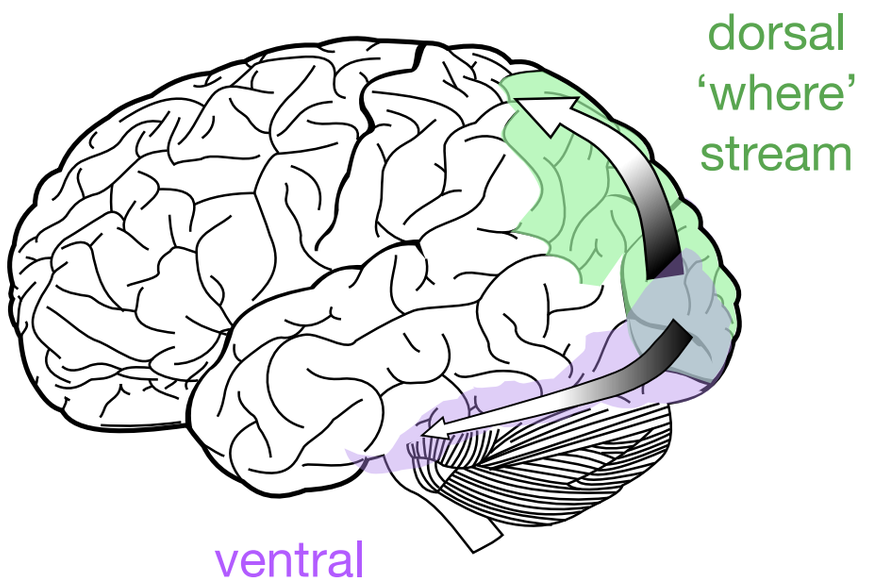


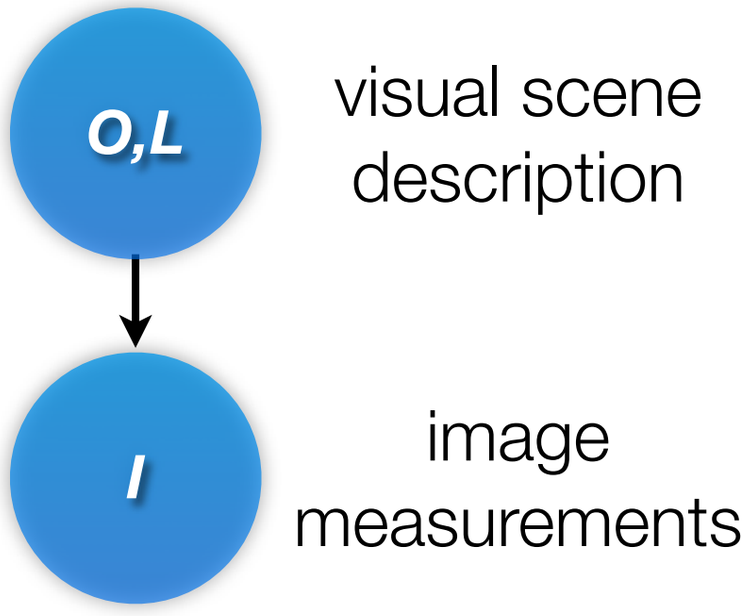
Assumption #1: Attentional spotlight

Broadbent ’52 ‘54; Treisman ‘60; Treisman & Gelade ‘80; Duncan & Desimone ‘95; Wolfe ‘97; and many others

- Object location *L* and identity *O* are independent

$$P(O, L, I)$$

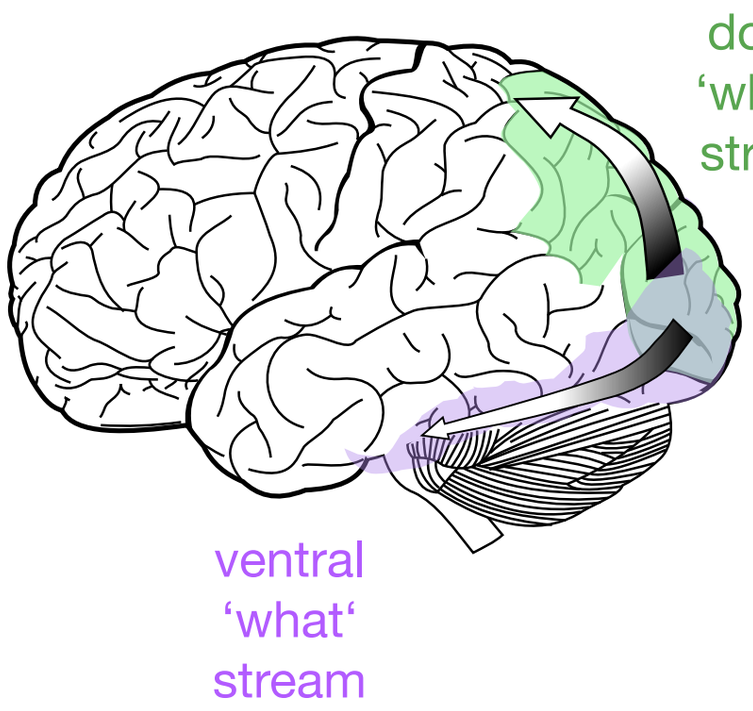


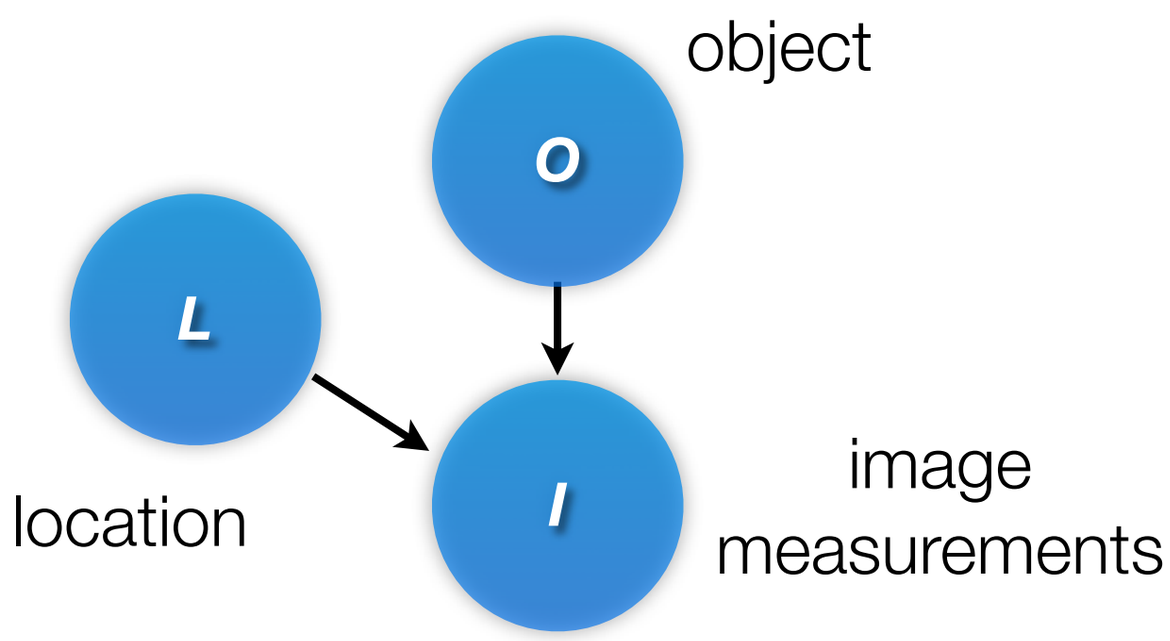


Assumption #2: 'what' and 'where' pathways

- Object location *L* and identity *O* are independent

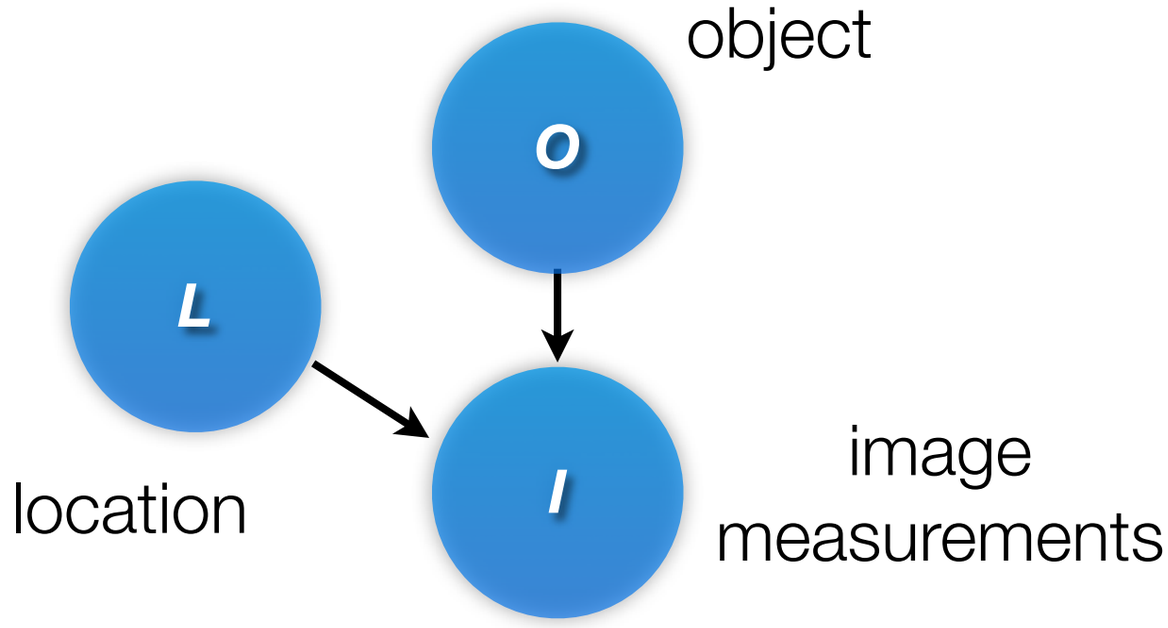
$$P(O, L, I) = P(O)P(L)P(I|L, O)$$





Assumption #2: 'what' and 'where' pathways

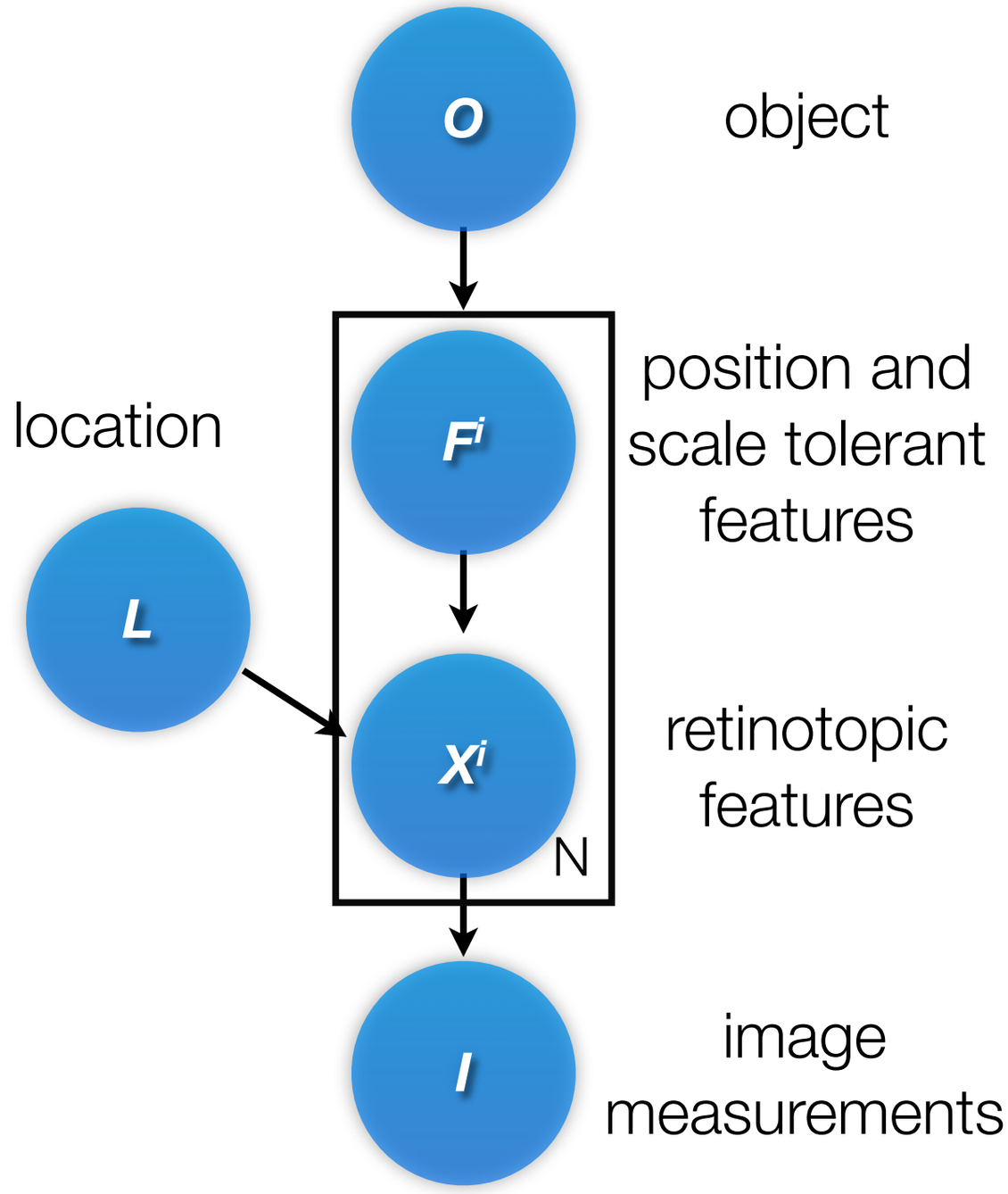
- Objects encoded by collections of ‘universal’ features (of intermediate complexity)
  - Either present or absent
  - Conditionally independent given the object and its location



Assumption #3: universal features

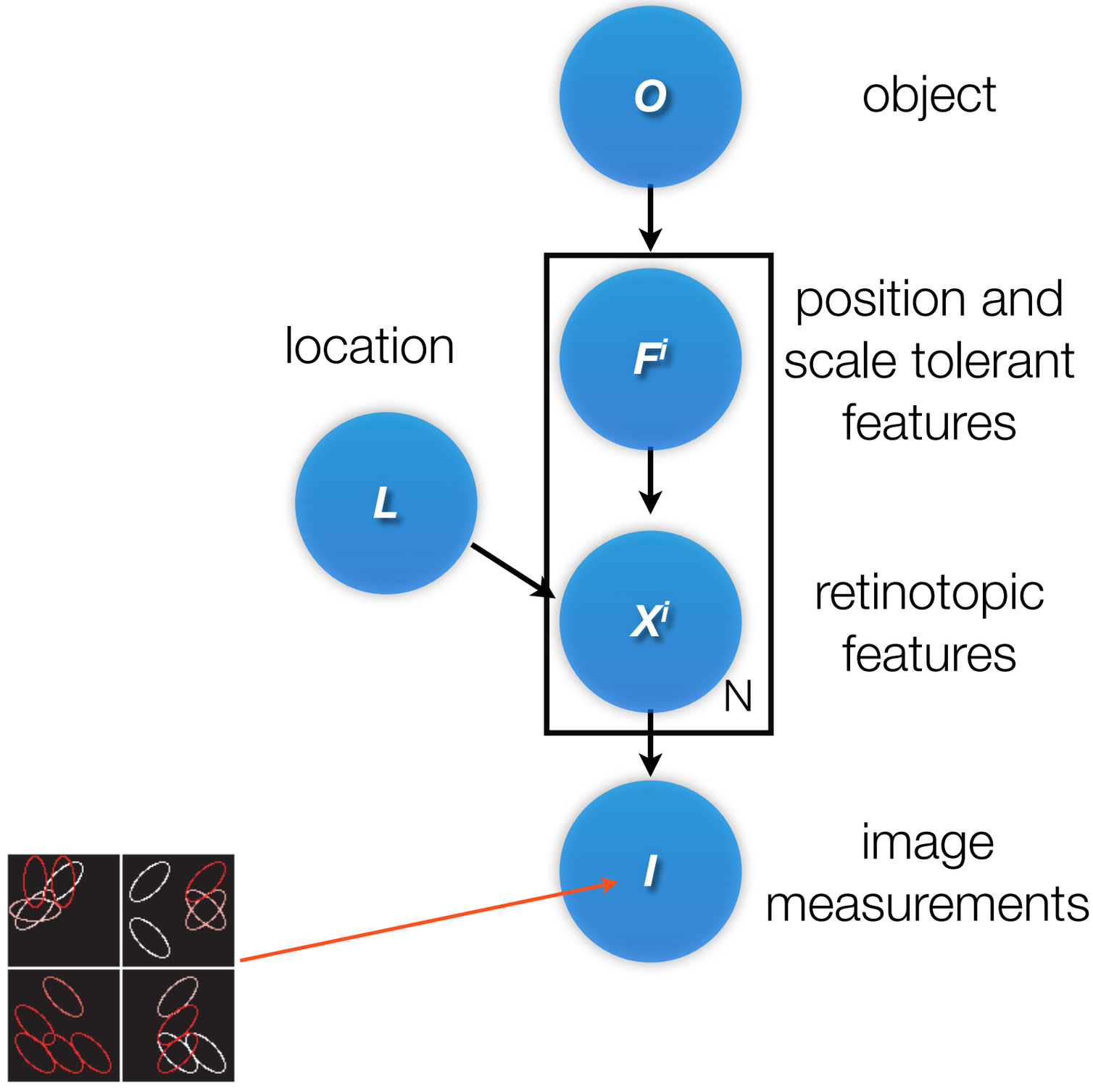
see Riesenhuber & Poggio ’99; Serre et al ’05 ’07

- Objects encoded by collections of ‘universal’ features (of intermediate complexity)
  - Either present or absent
  - Conditionally independent given the object and its location



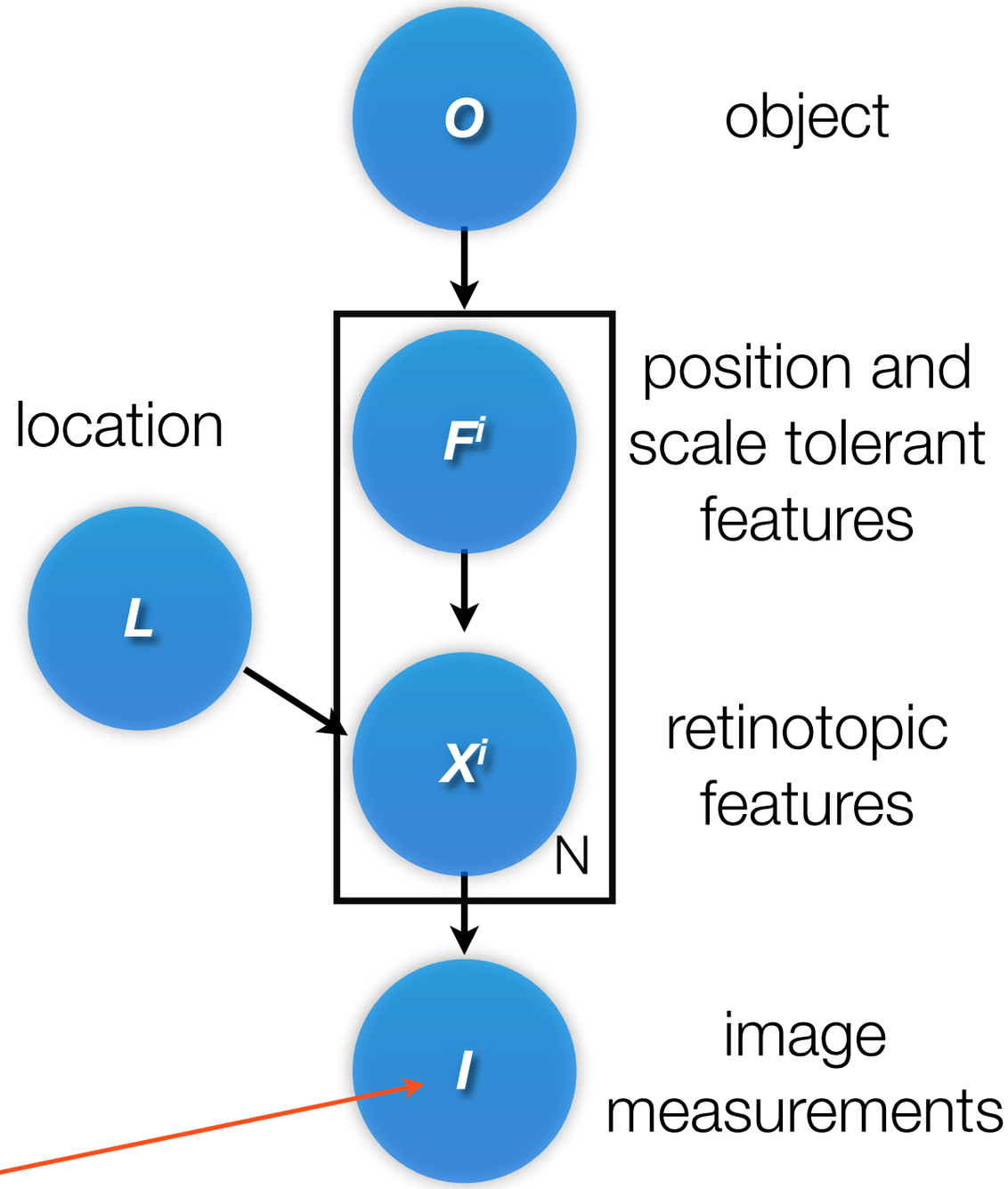
Assumption #3: universal features

see Riesenhuber & Poggio ’99; Serre et al ’05 ’07



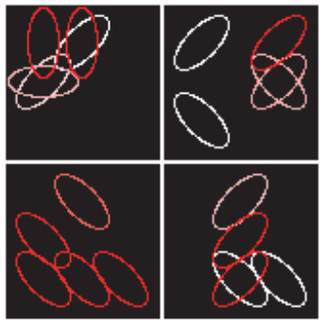
Assumption #3: universal features

see Riesenhuber & Poggio ’99; Serre et al ’05 ’07



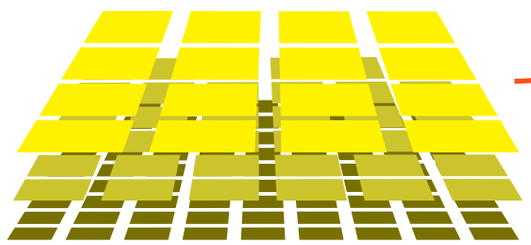
V2/V4

Serre et al ’05
David et al ’06
Cadieu et al ’07
Willmore et al ’10



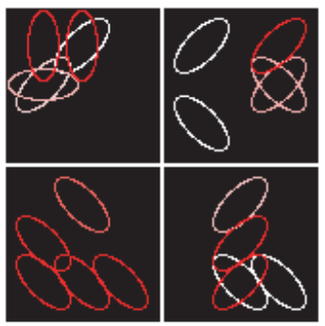
Assumption #3: universal features

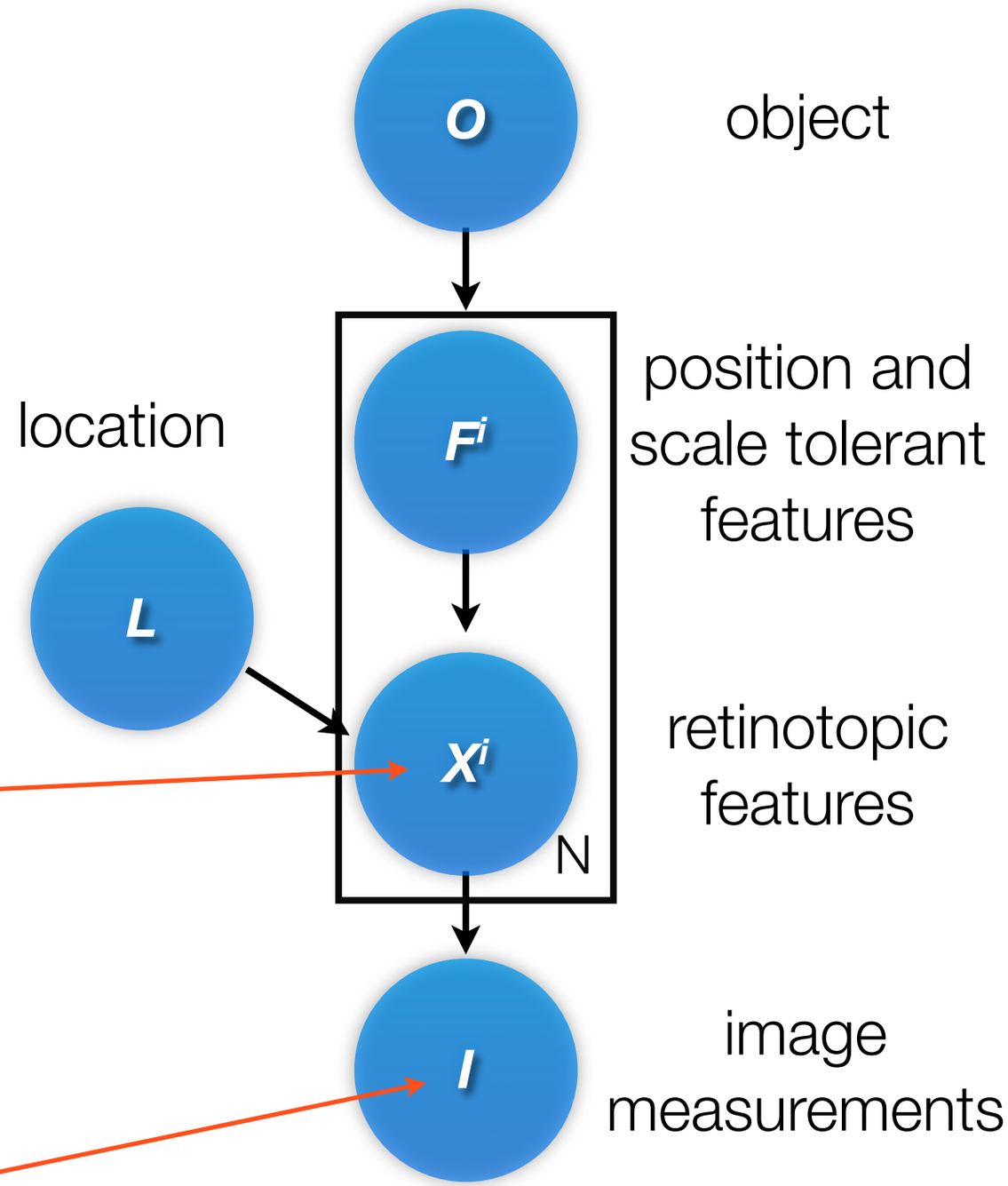
see Riesenhuber & Poggio ’99; Serre et al ’05 ’07



V2/V4

Serre et al ’05
David et al ’06
Cadieu et al ’07
Willmore et al ’10



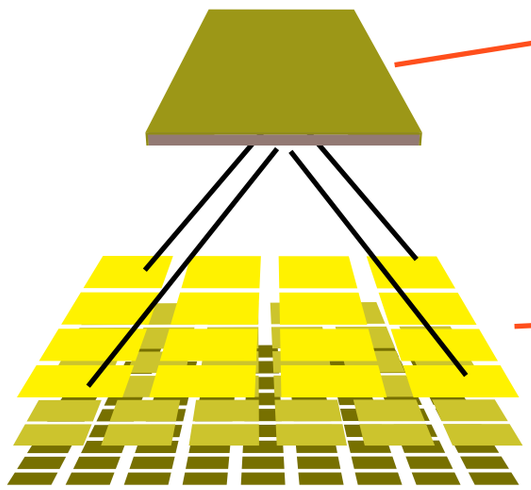


Assumption #3: universal features

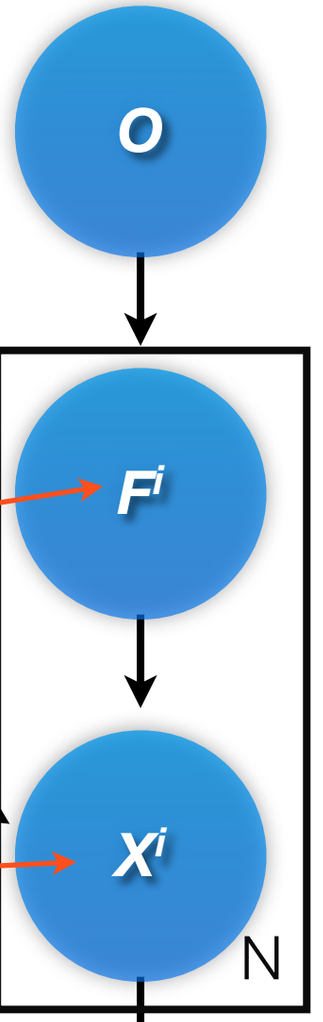
see Riesenhuber & Poggio ’99; Serre et al ’05 ’07

IT

Serre et al ’07



location



object

position and scale tolerant features

retinotopic features

V2/V4

Serre et al ’05
David et al ’06
Cadieu et al ’07
Willmore et al ’10

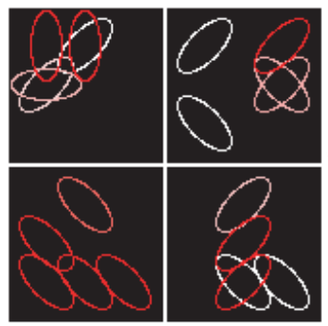


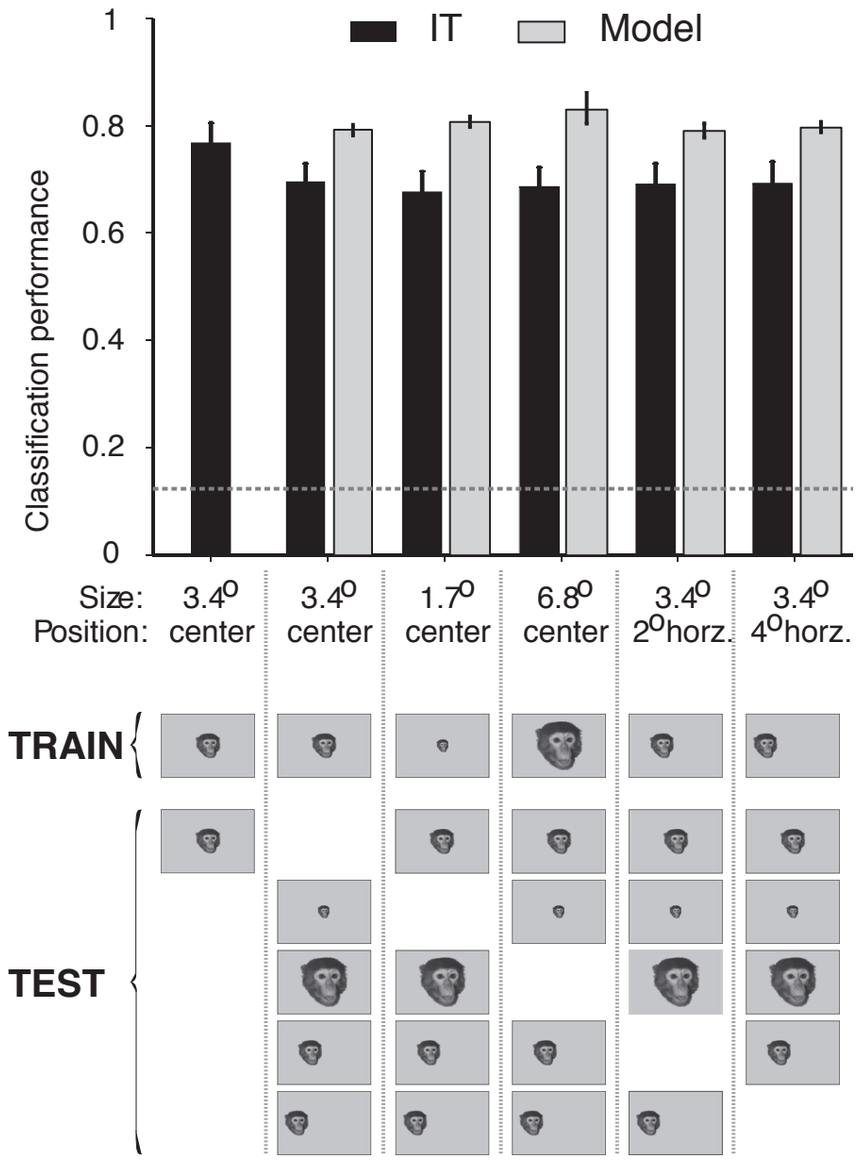
image measurements

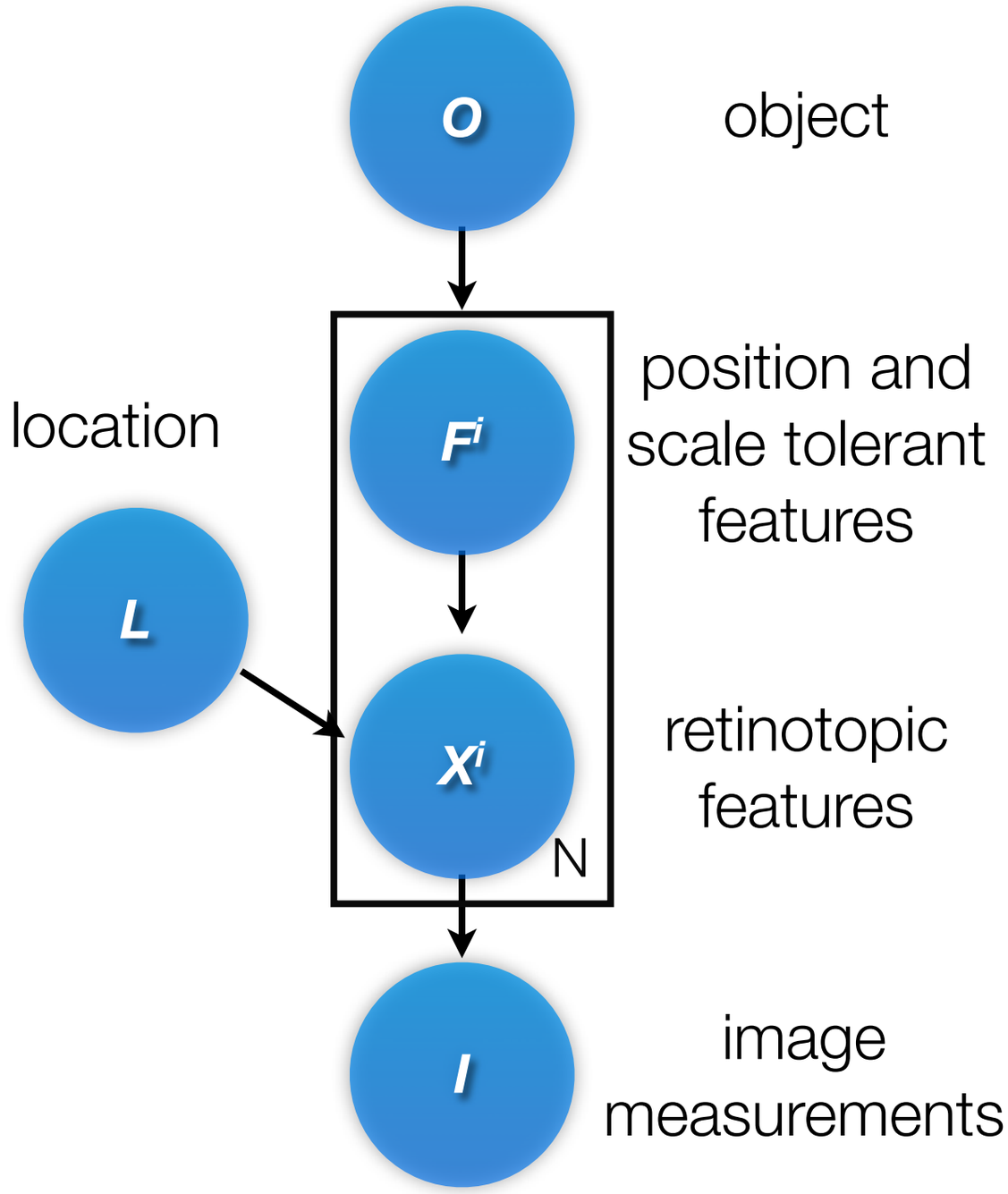
Assumption #3: universal features

see Riesenhuber & Poggio ’99; Serre et al ’05 ’07

Predicting IT readout

Serre et al '07



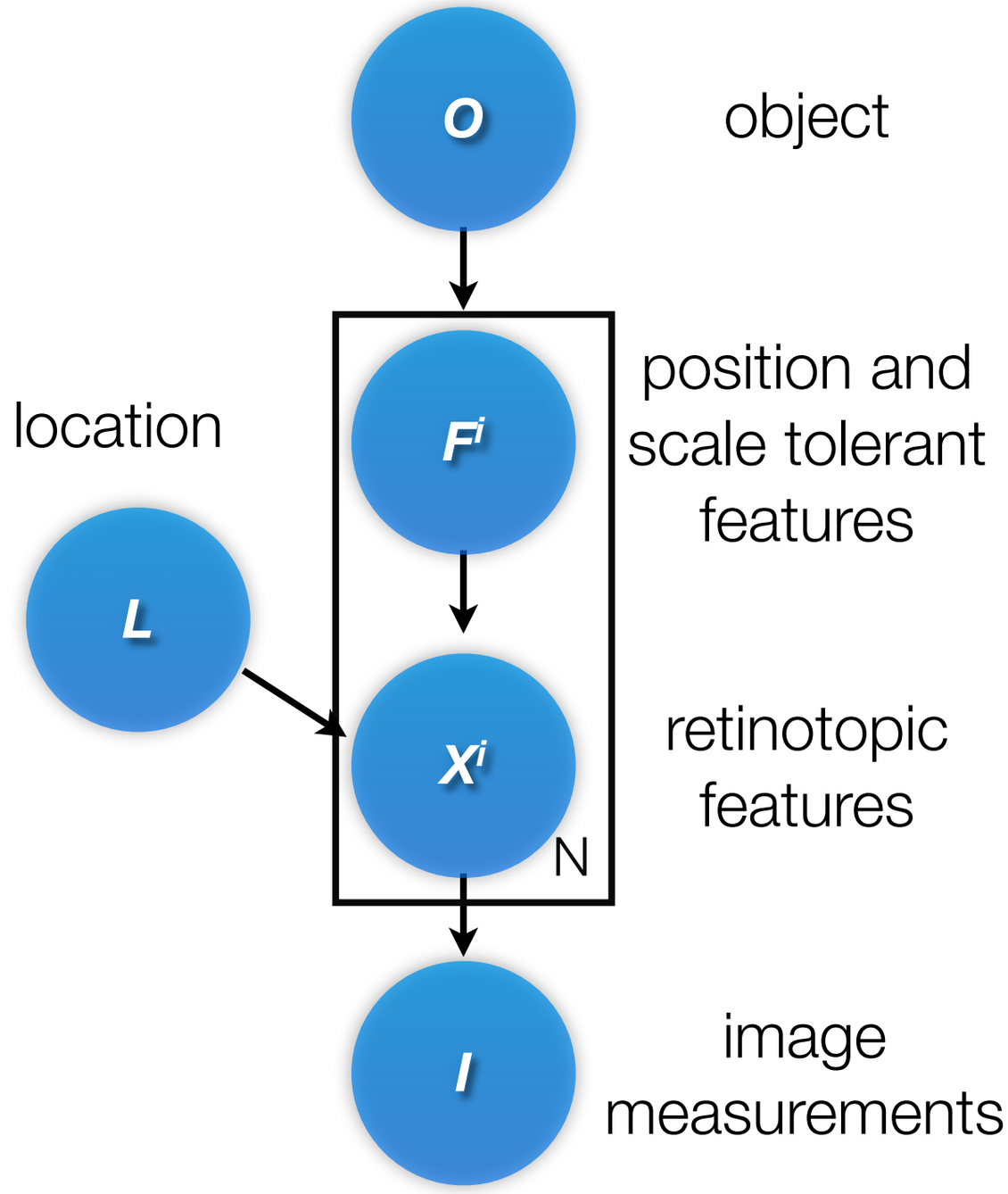


Bayesian inference and attention

Model: Serre et al '05
Experimental data: Hung* Kreiman*et al '05

behavior

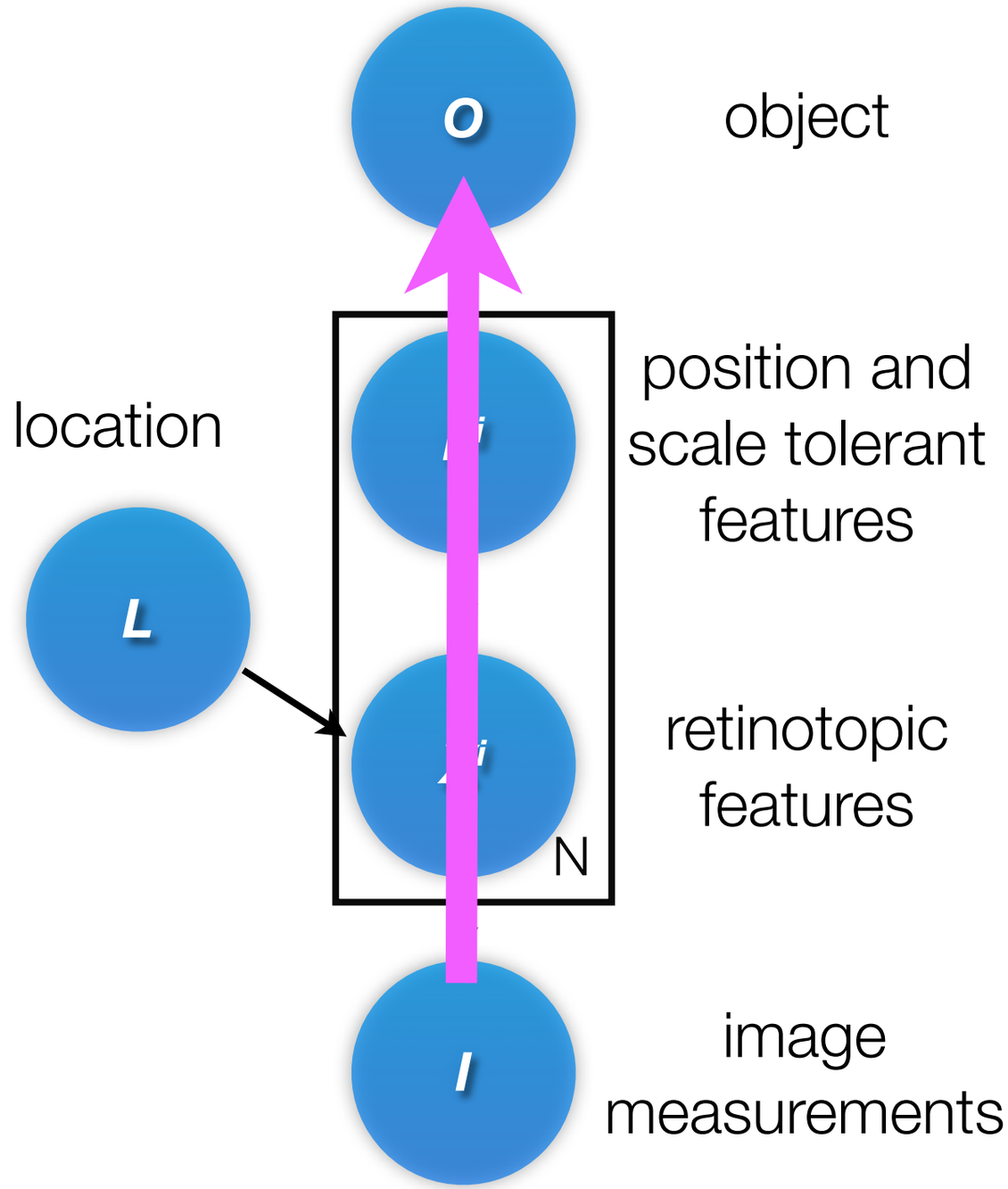
Serre Oliva & Poggio ’07



Bayesian inference and attention

behavior

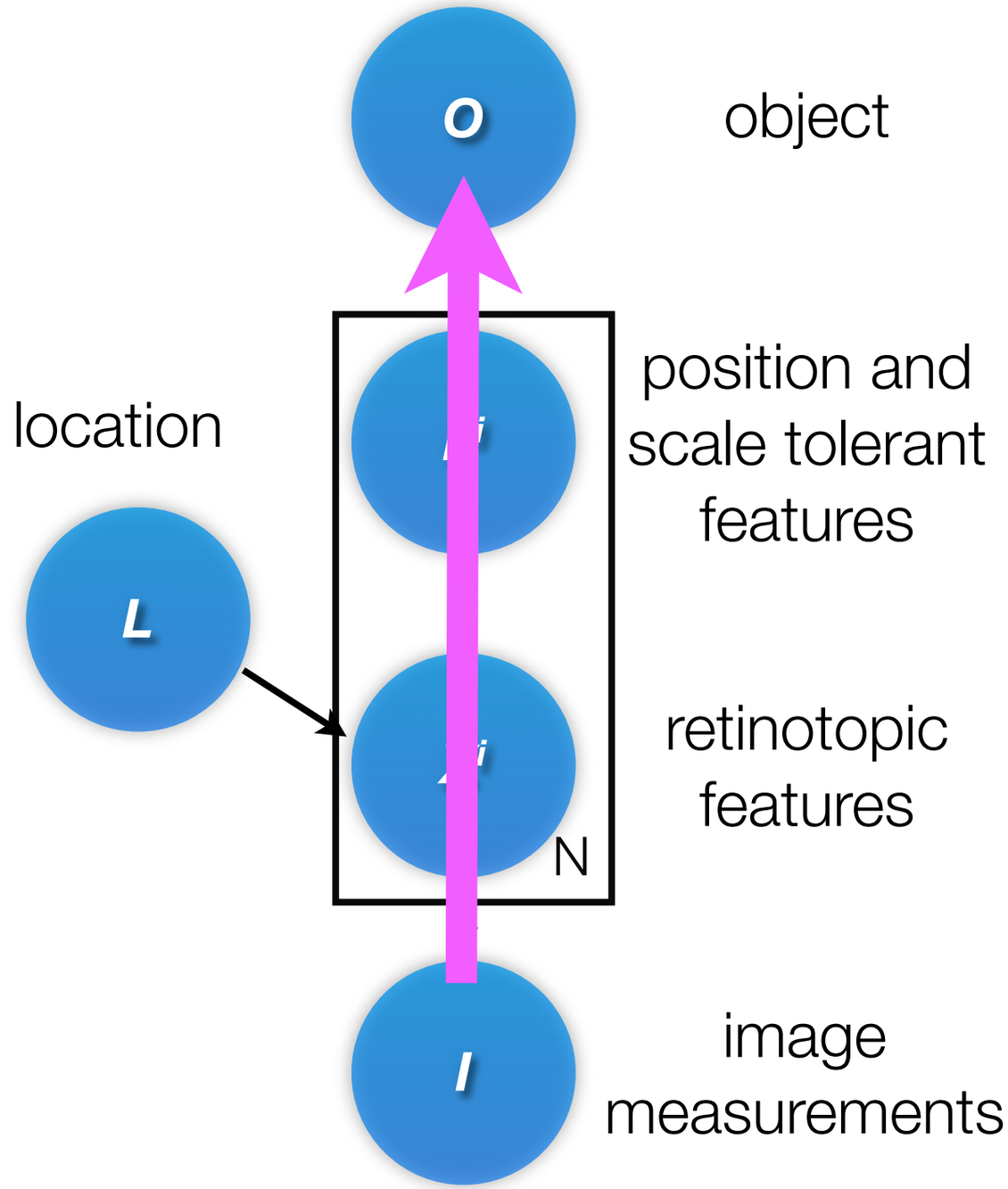
Serre Oliva & Poggio ’07



Bayesian inference and attention

behavior

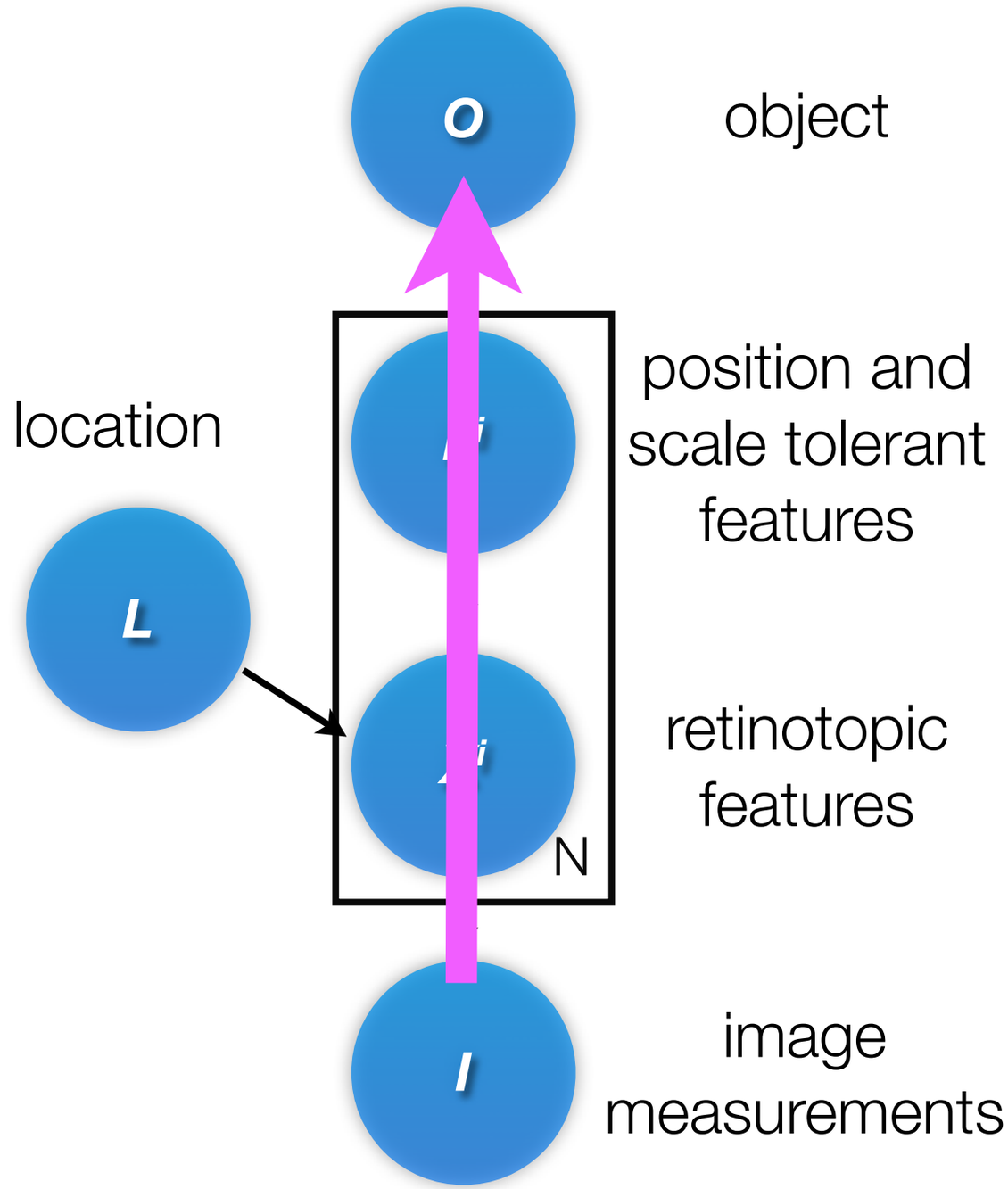
Serre Oliva & Poggio ’07



Bayesian inference and attention

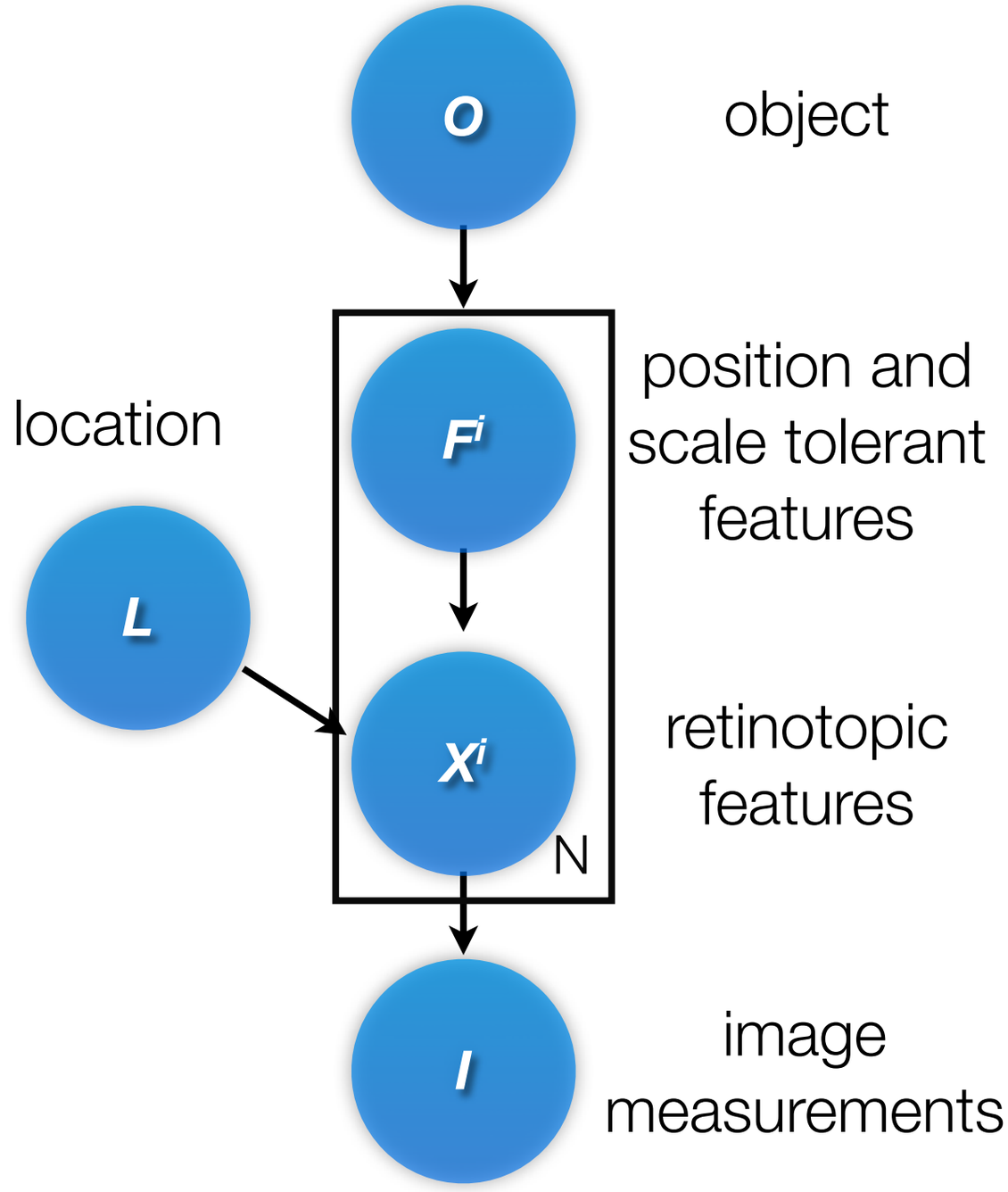
behavior

Serre Oliva & Poggio ’07

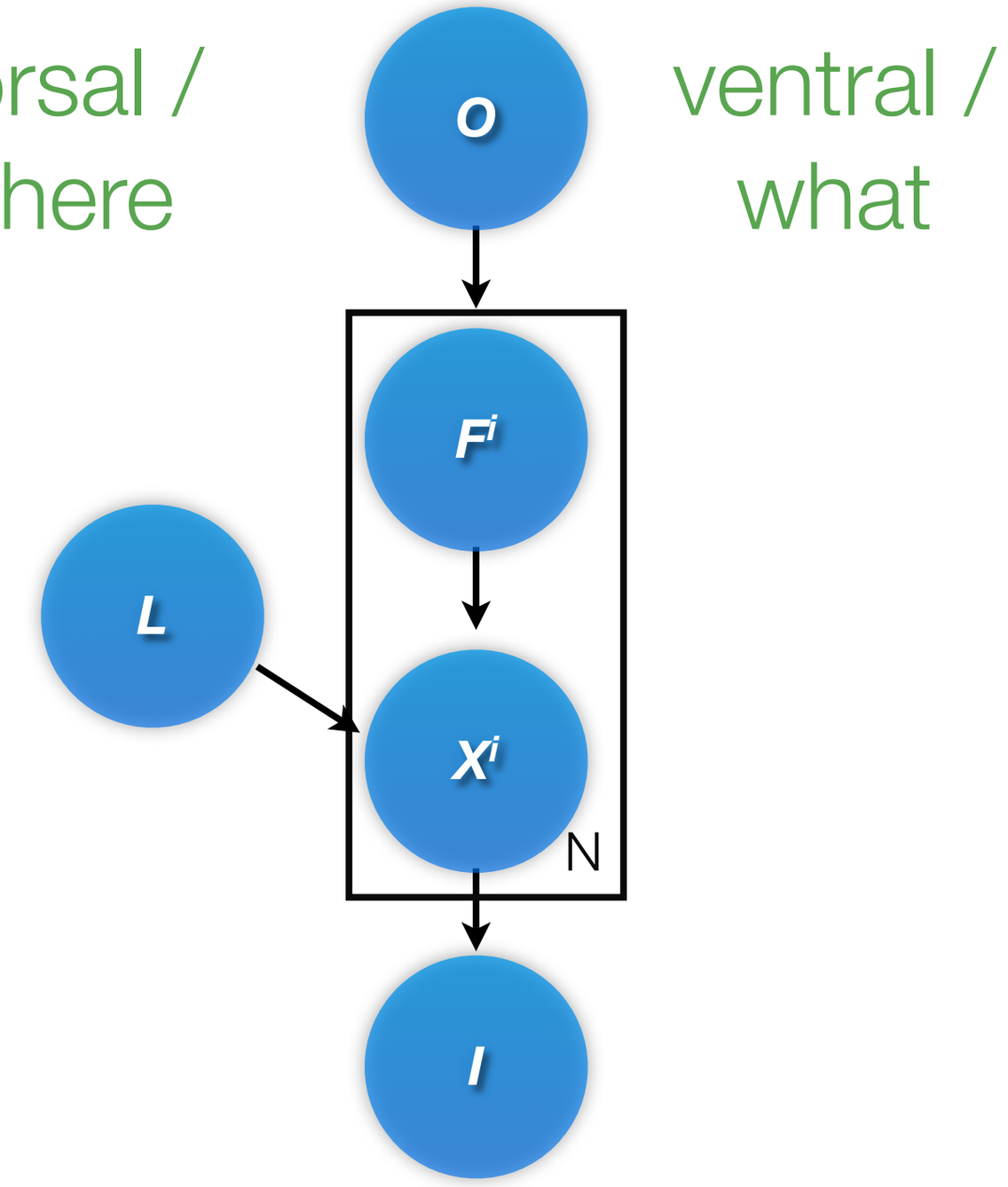


Bayesian inference and attention

- ***Goal of visual perception:*** to estimate posterior probabilities of visual features, objects and their locations in an image
- Attention corresponds to conditioning on high-level latent variables representing particular features or locations (as well as on sensory input), and doing inference over the other latent variables



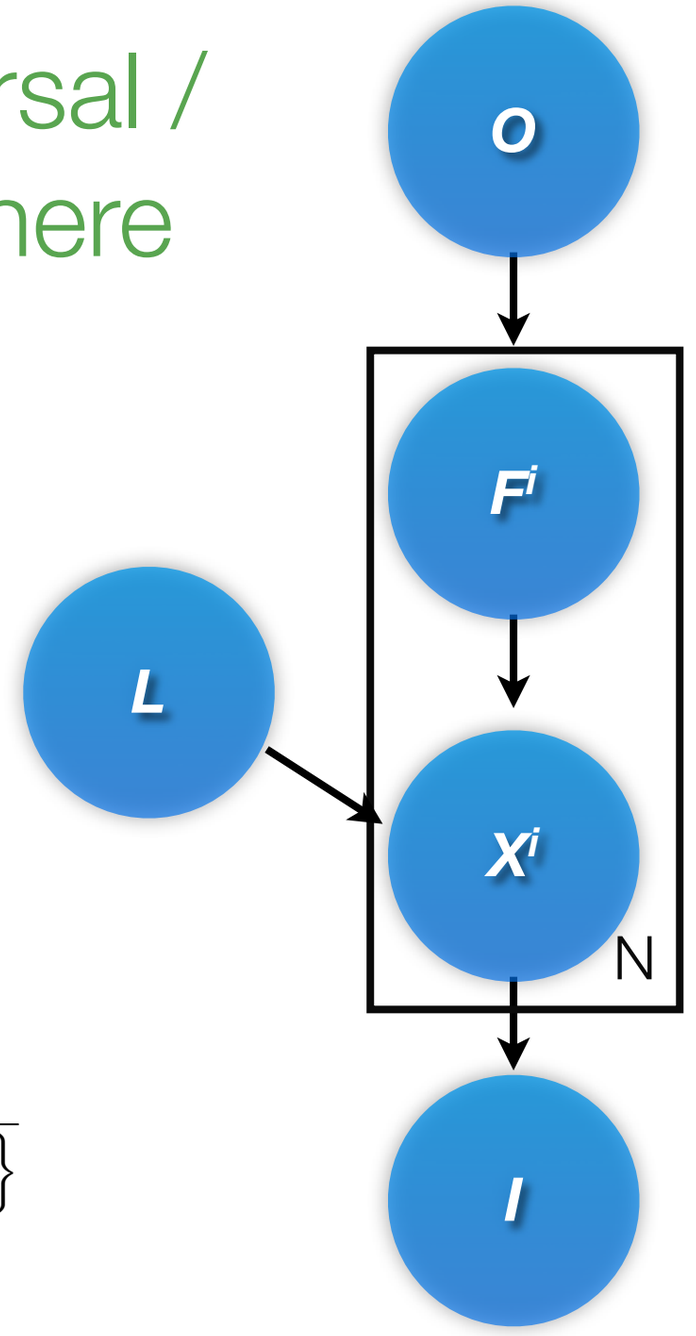
Bayesian inference and attention



Bayesian inference and attention

dorsal / where

ventral / what



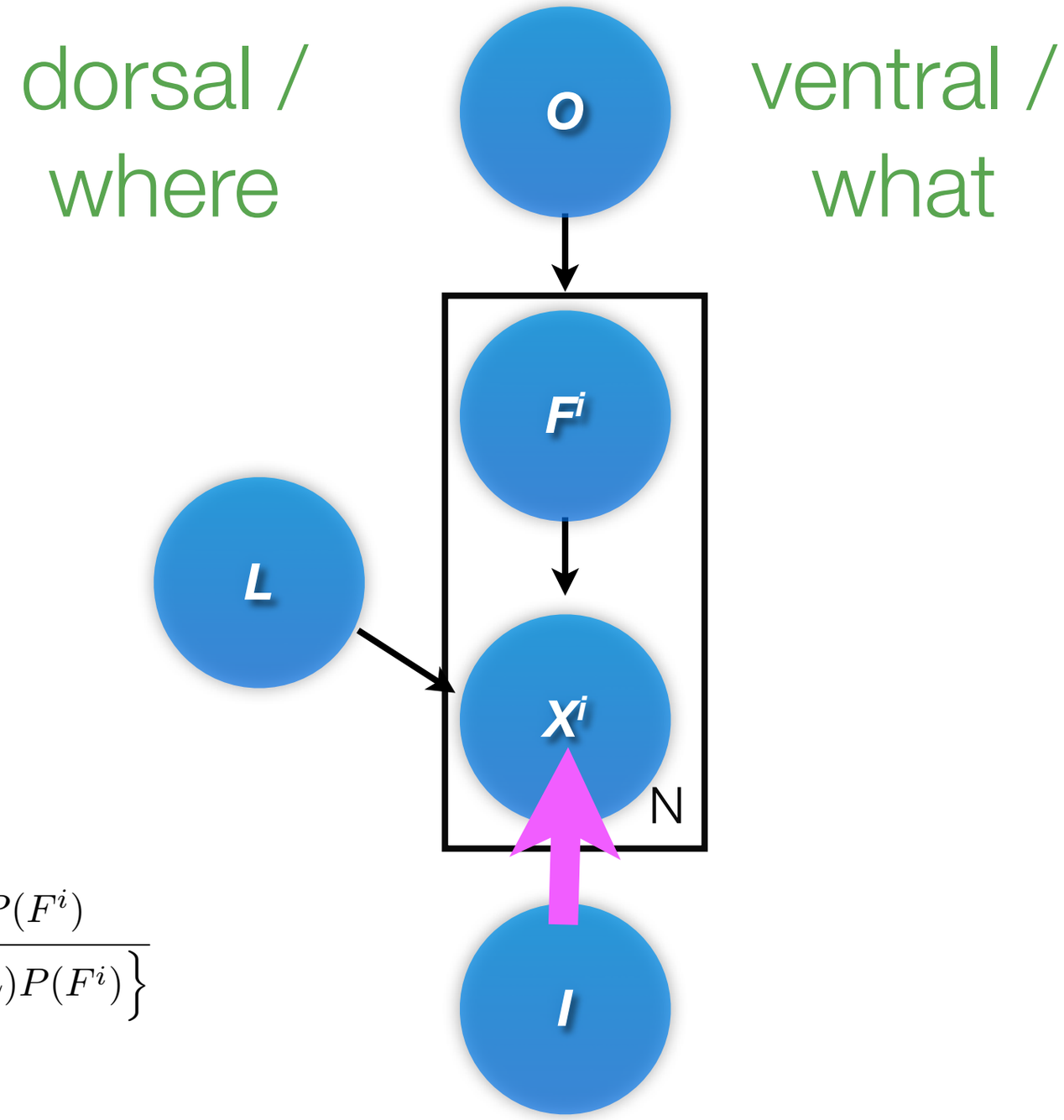
$$P(X^i|I) \quad = \quad \frac{P(I|X^i) \ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i} \left\{ P(I|X^i) \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i) \right\}}$$

Bayesian inference and attention

feedforward input

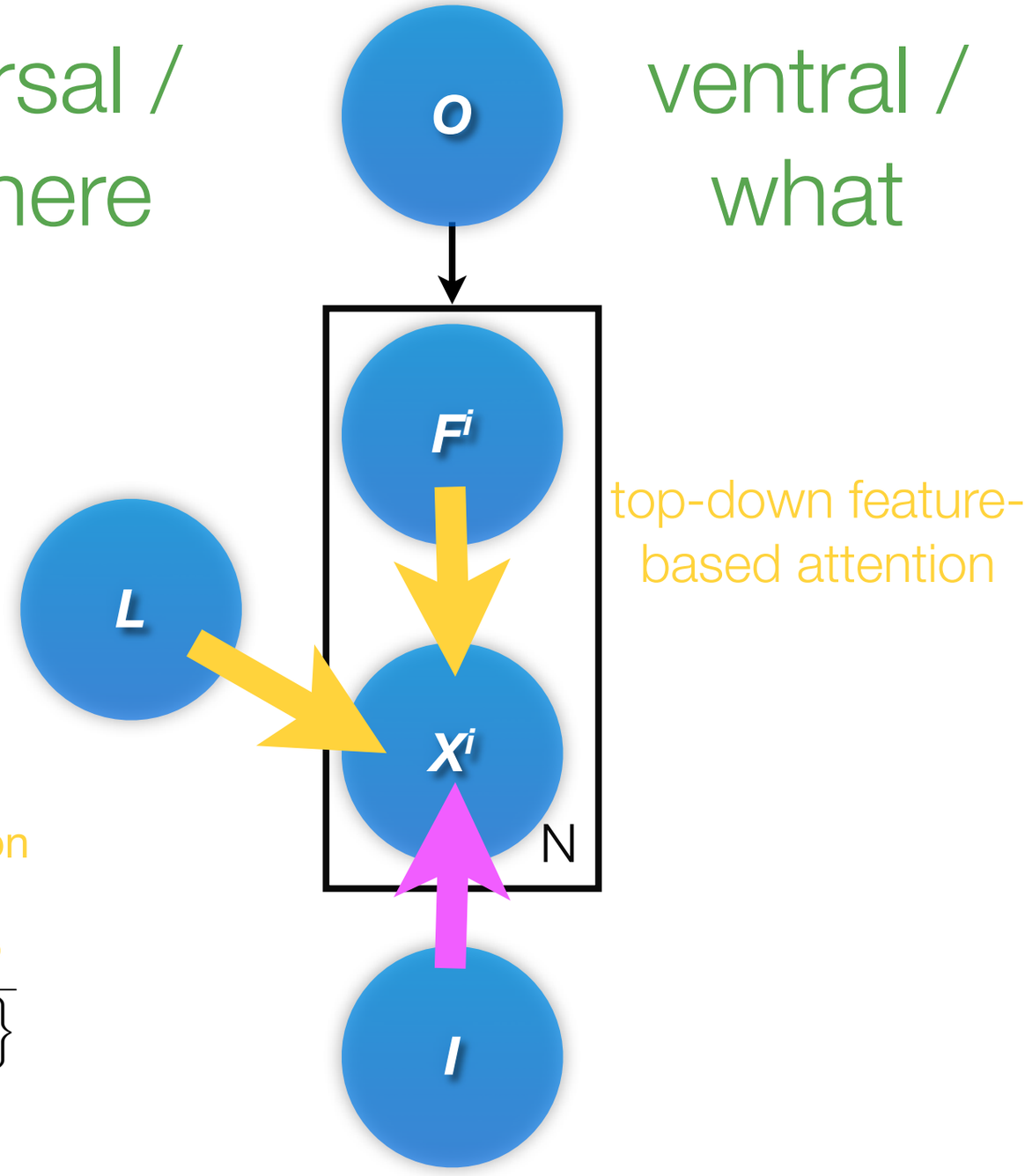
$$P(X^i|I) \quad = \quad \frac{P(I|X^i) \ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i} \left\{ P(I|X^i) \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i) \right\}}$$

Bayesian inference and attention



dorsal / where

ventral / what



top-down spatial attention

top-down feature-based attention

feedforward input

attentional modulation

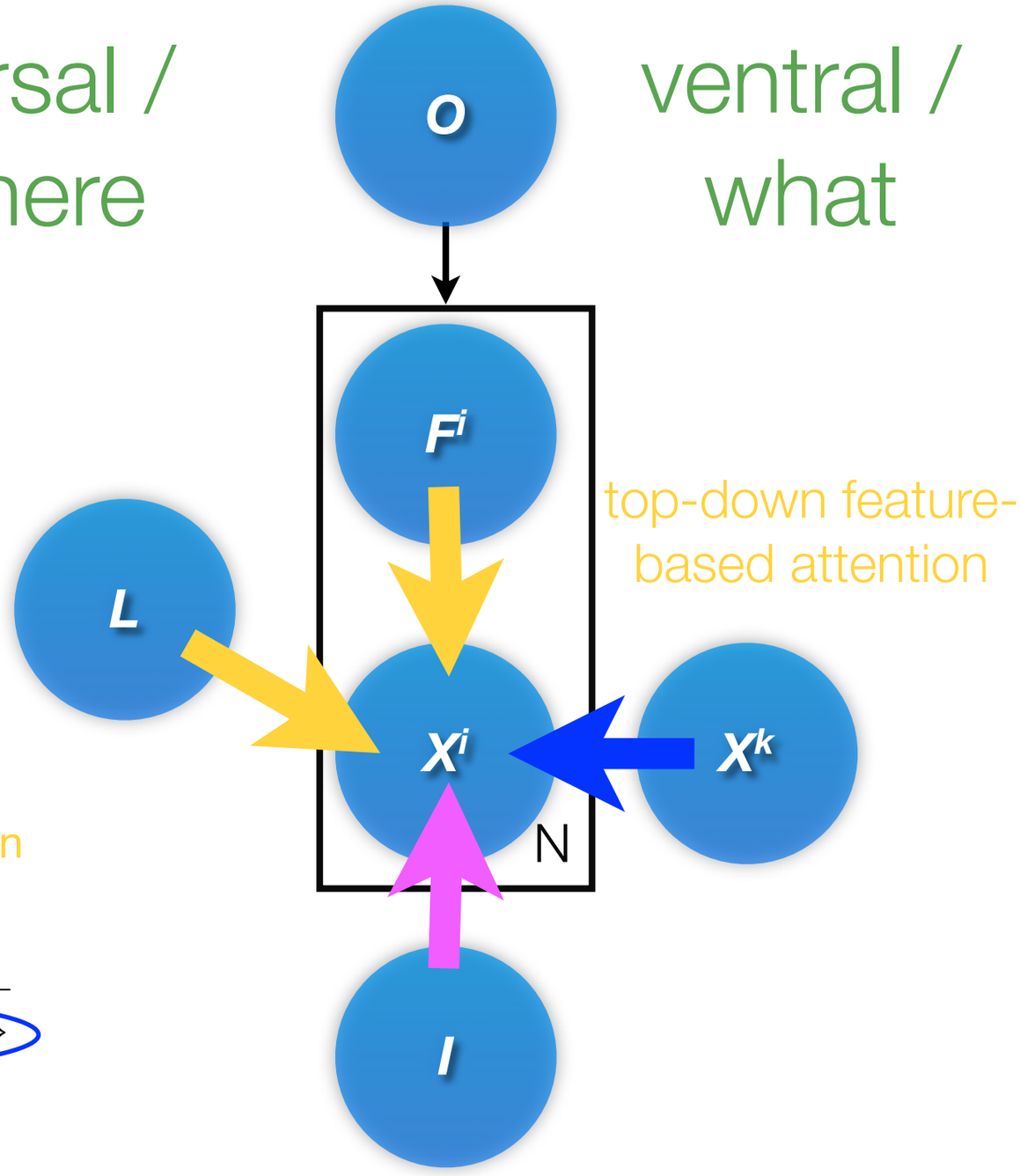
$$P(X^i|I) = \frac{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

feedforward (bottom-up) sweep

Bayesian inference and attention

dorsal / where

ventral / what



top-down spatial attention

top-down feature-based attention

feedforward (bottom-up) sweep

feedforward input

attentional modulation

$$P(X^i|I) = \frac{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

suppressive drive

Bayesian inference and attention

Neuron
Review

The Normalization Model of Attention

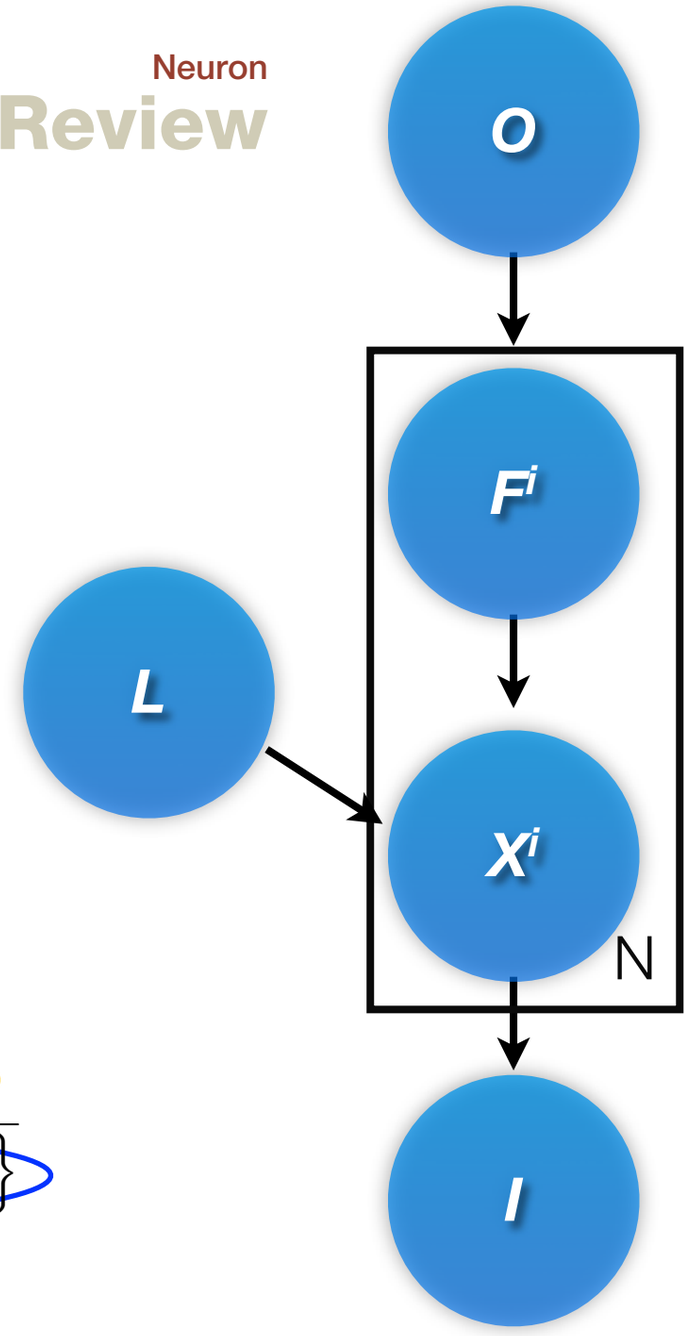
John H. Reynolds[1,*] and David J. Heeger[2]
[1]Salk Institute for Biological Studies, La Jolla, CA 92037-1099, USA
[2]Department of Psychology and Center for Neural Science, New York University, New York, NY 10003, USA
*Correspondence: reynolds@salk.edu
DOI 10.1016/j.neuron.2009.01.002

$$R(x,\theta) = \frac{A(x,\theta)E(x,\theta)}{S(x,\theta)+\sigma}$$

E A

$$P(X^i|I) = \frac{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

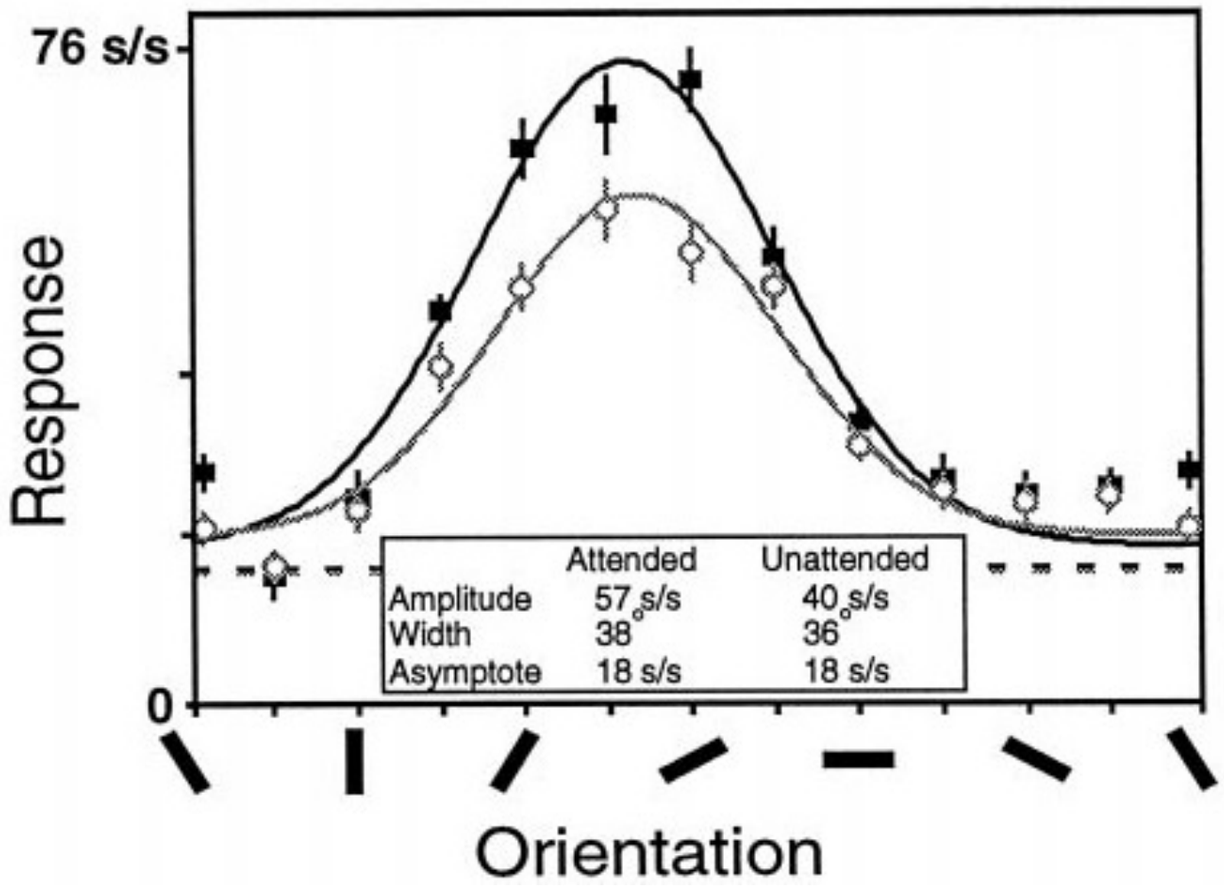
S



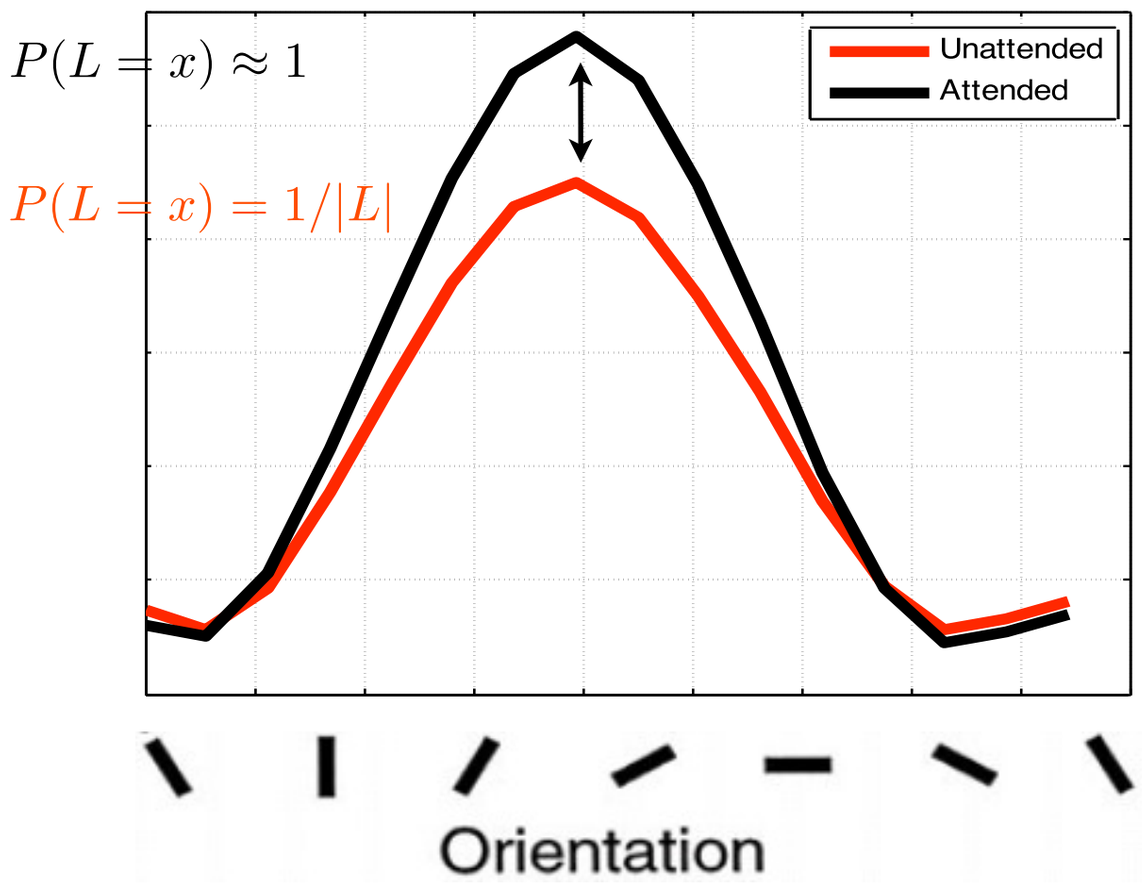
Bayesian inference and attention

$$P(X^i = x|I) \quad \propto \quad \sum_{F^i, L} P(X^i = x|F^i, L) P(I|X^i) P(F^i) P(L)$$

McAdams and Maunsell ‘99



Model

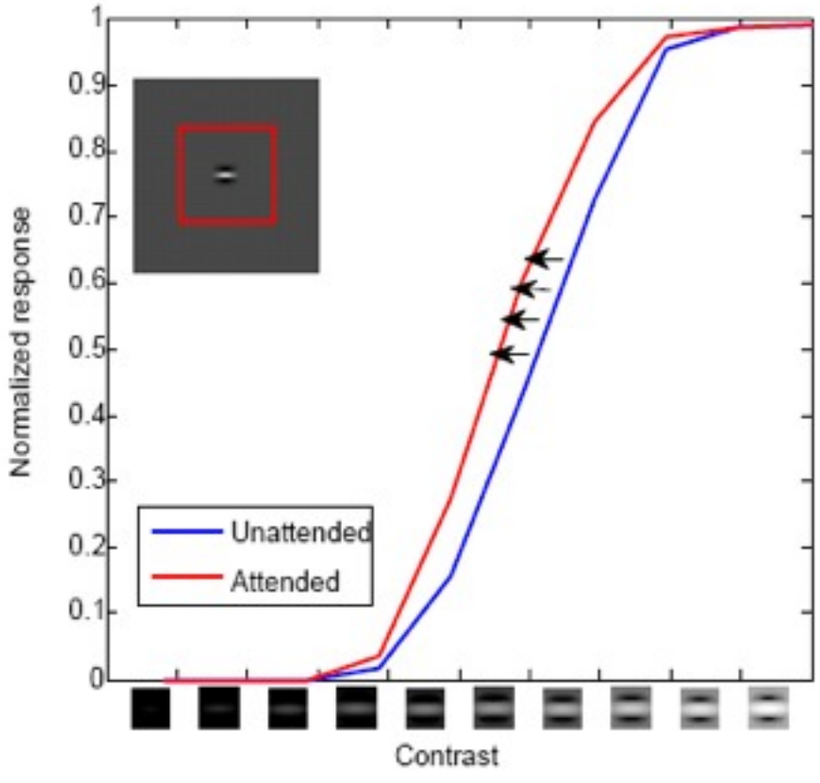


Multiplicative scaling of tuning curves by spatial attention

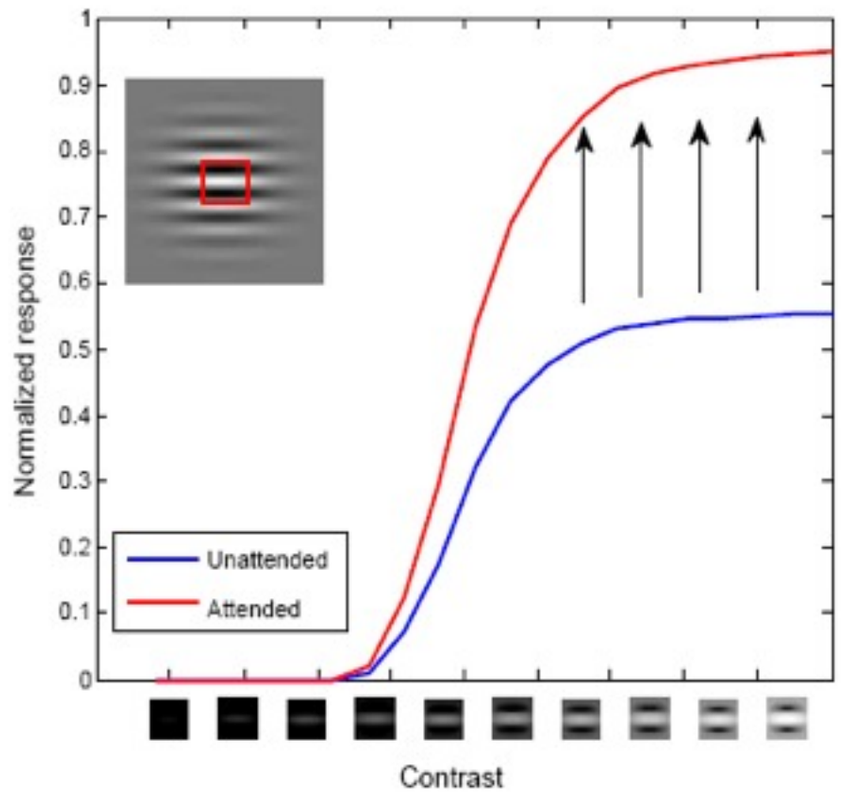
see also Heeger & Reynolds ’09

$$P(X^i|I) \quad = \quad \frac{P(I|X^i) \ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

Trujillo and Treue ‘02



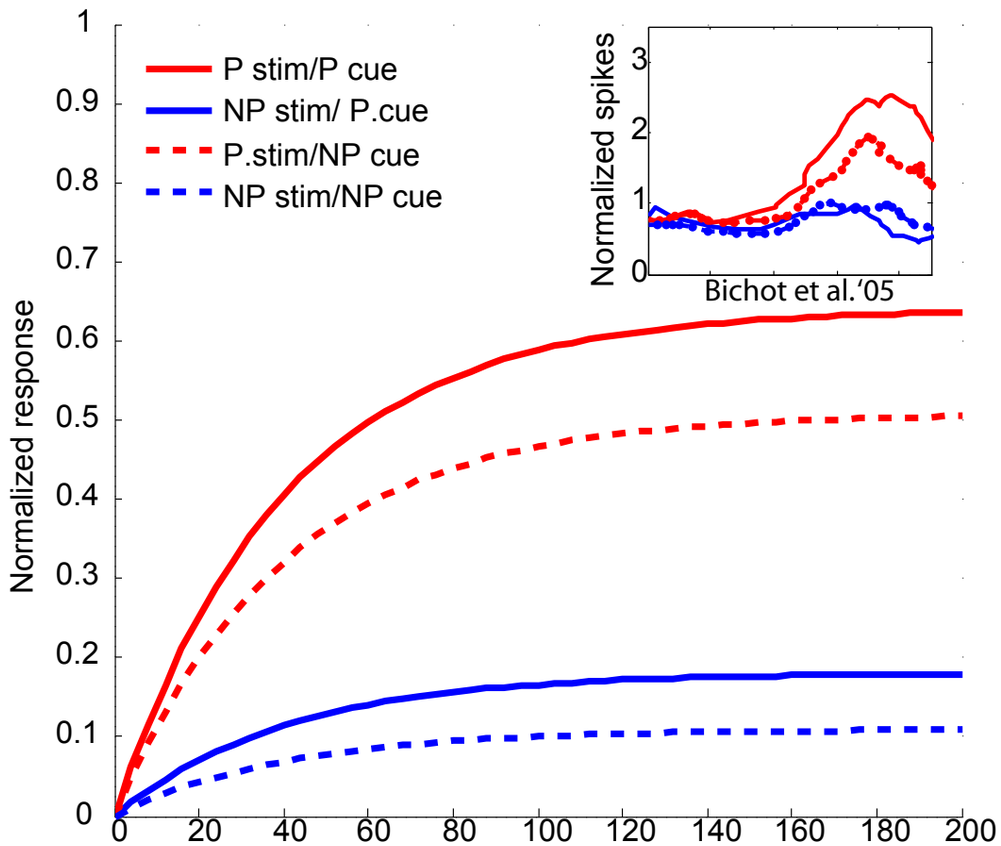
Mc Adams and Maunsell ’99

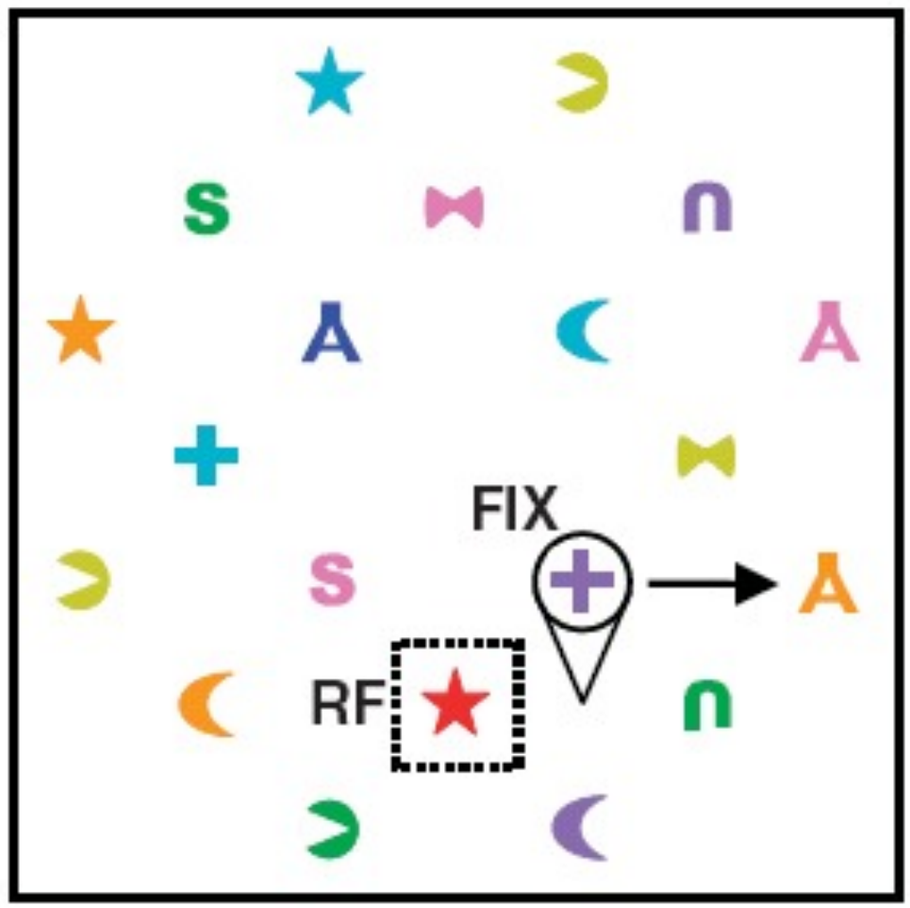


Contrast vs. response gain

see also Heeger & Reynolds ’09

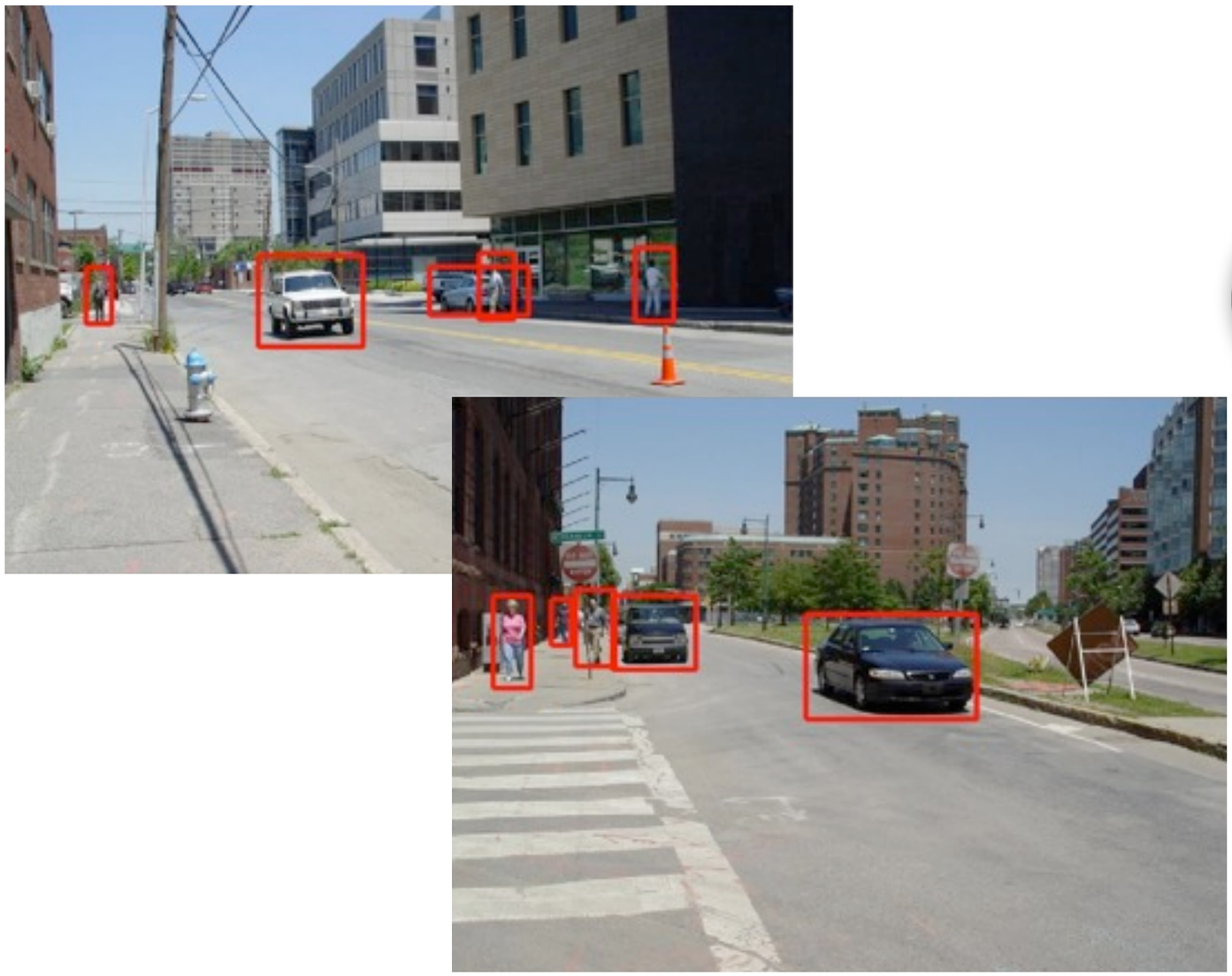
$$P(X^i = x|I) \quad \propto \quad \sum_{F^i, L} P(X^i = x|F^i, L) P(I|X^i) P(F^i) P(L)$$

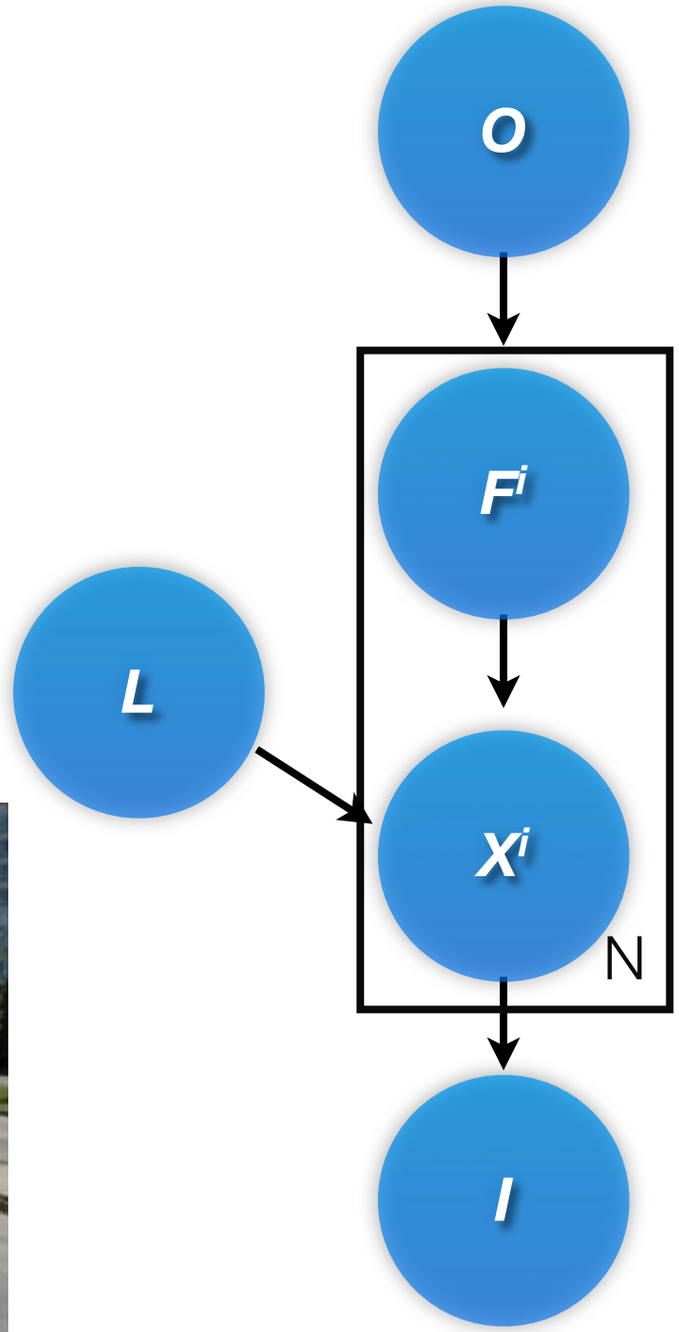




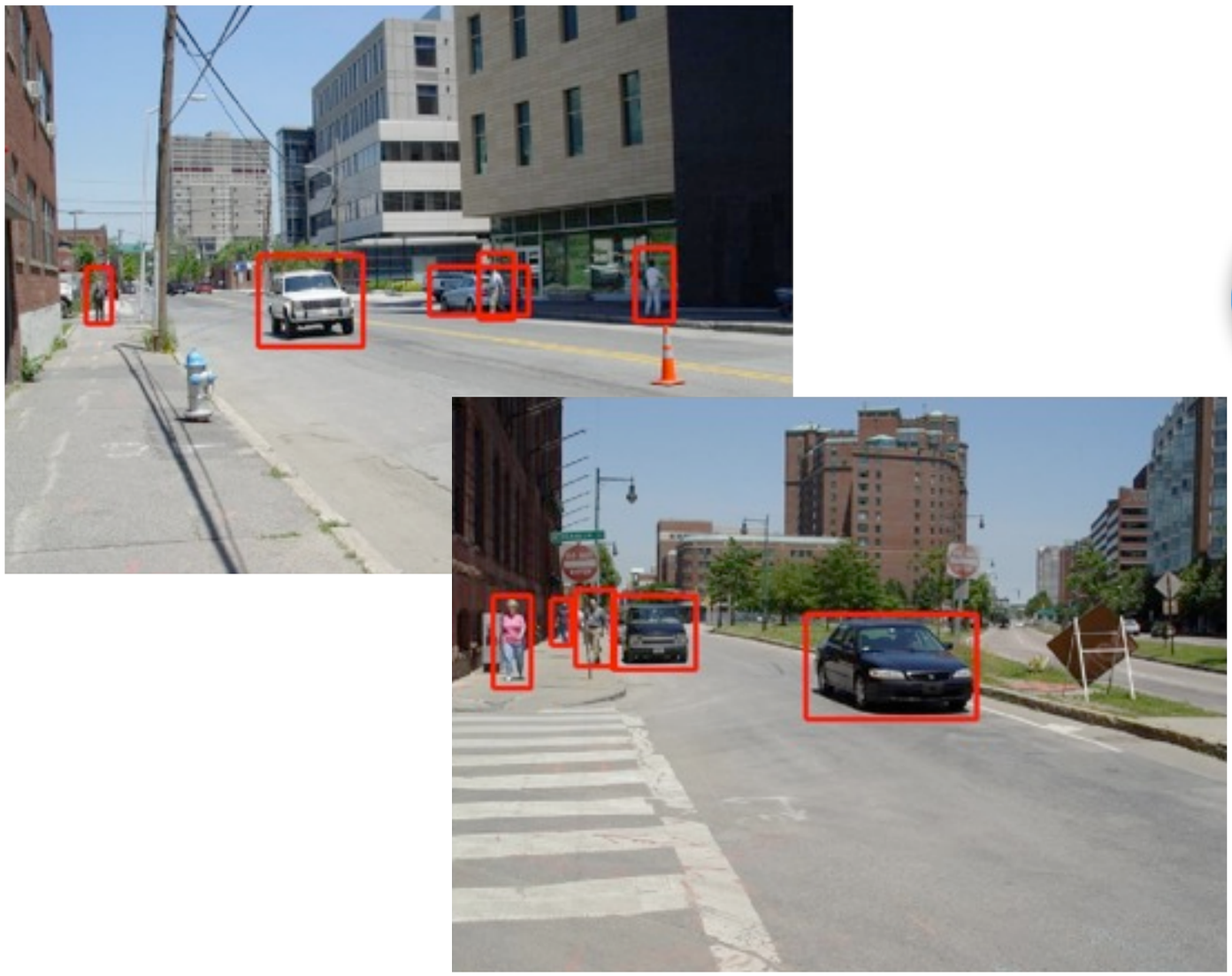
Feature-based attention

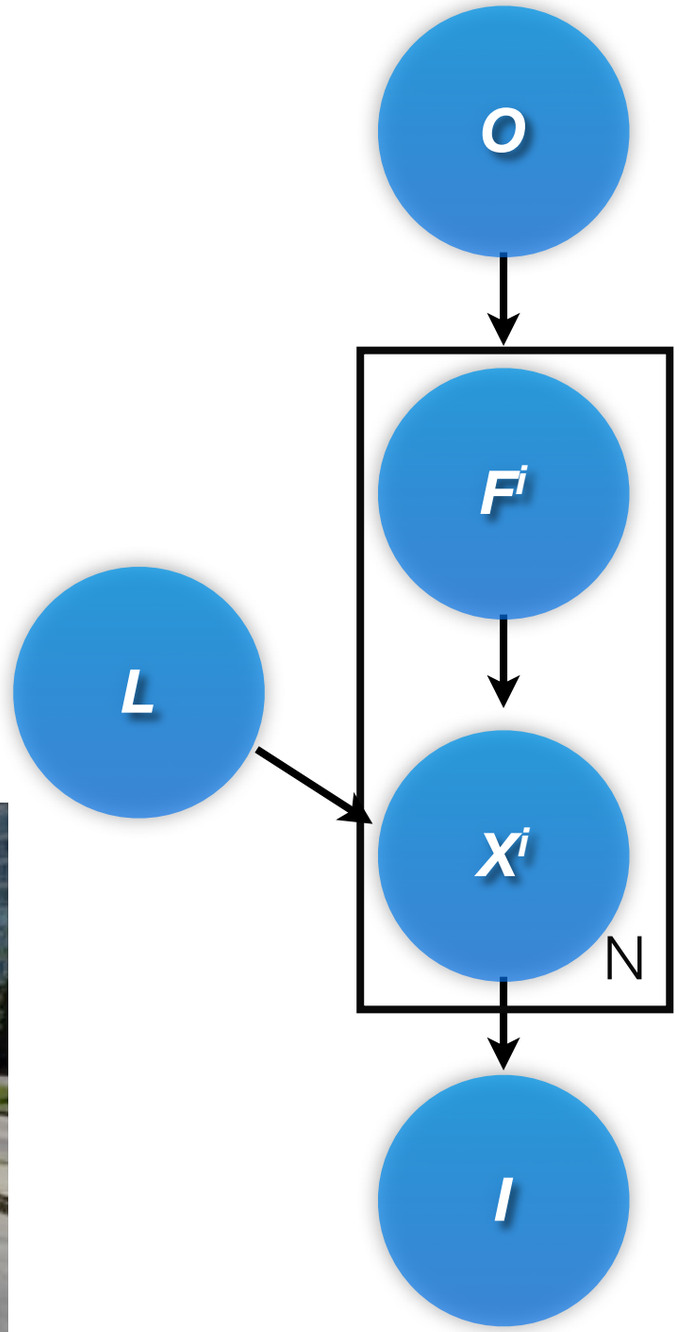
neural data from Bichot et al '05





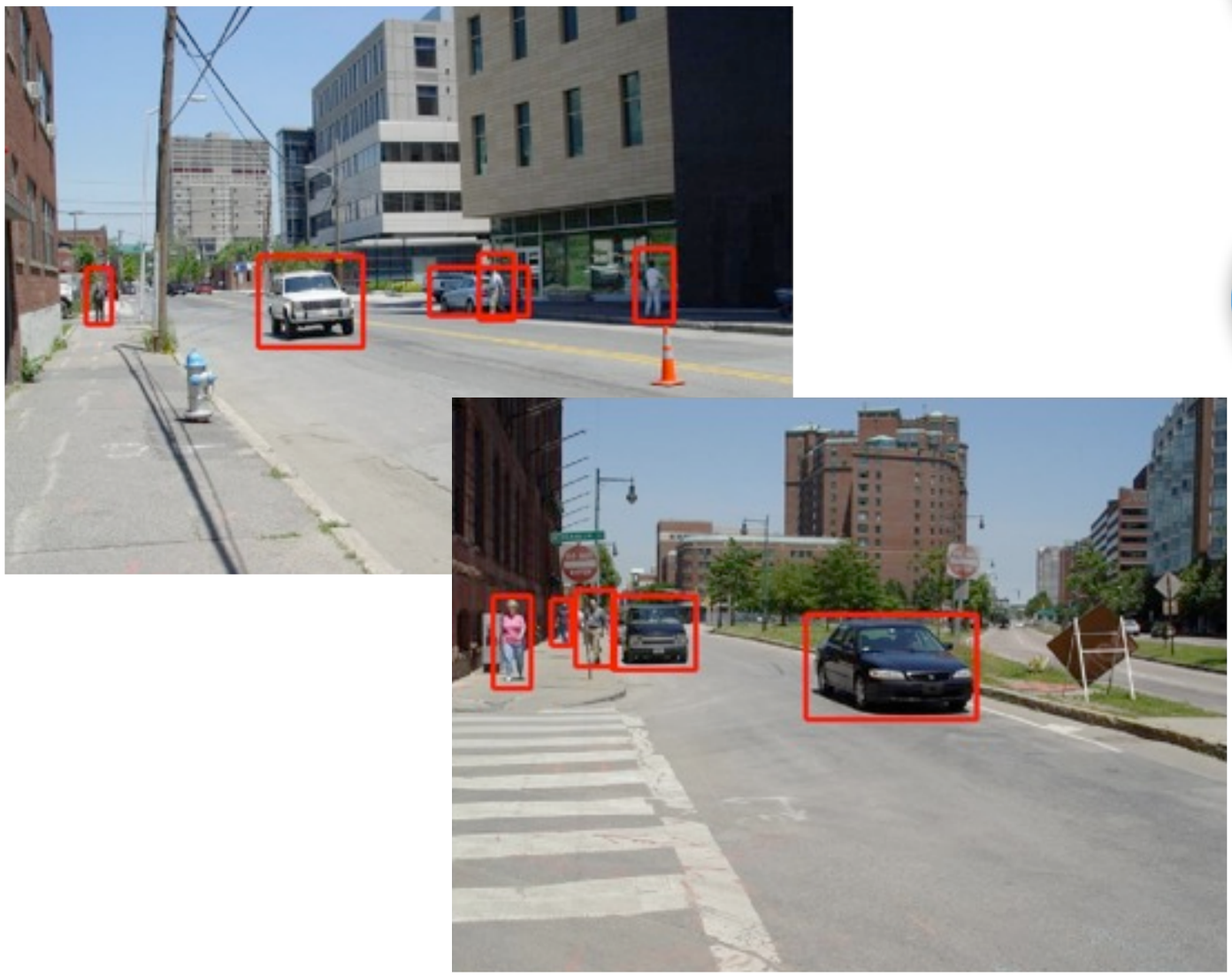
Learning to localize cars and pedestrians

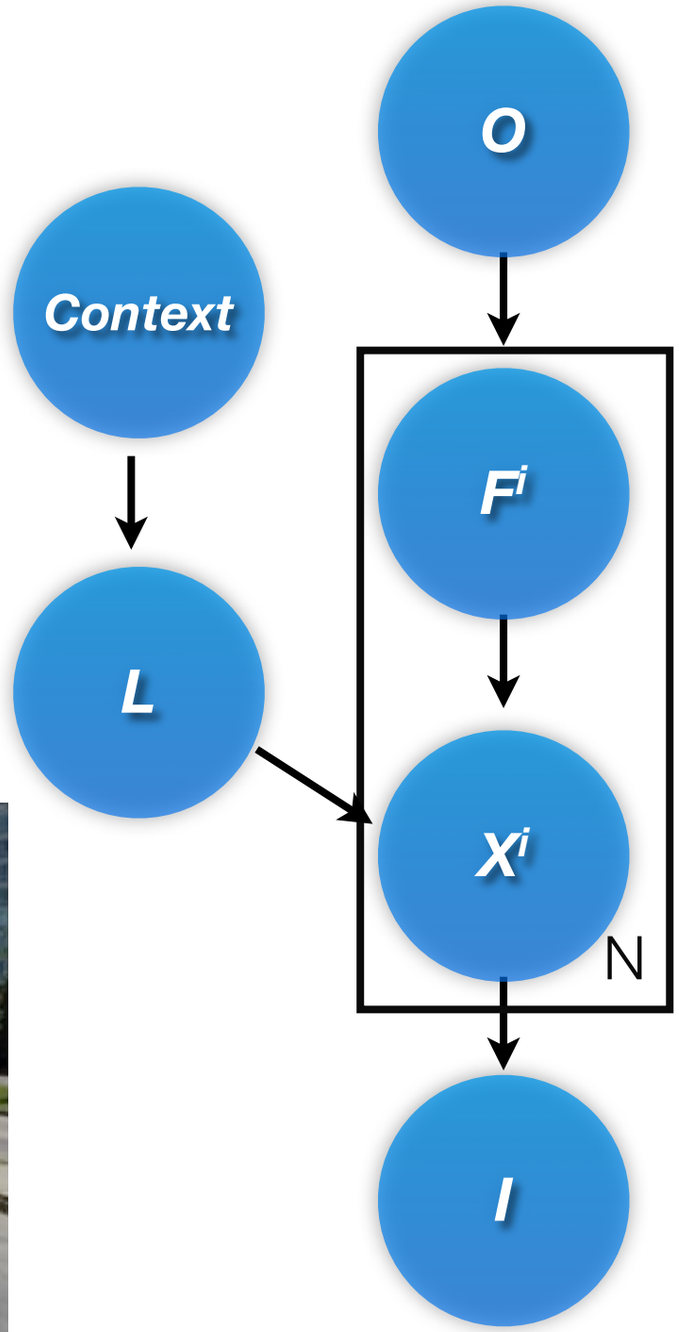




learning object priors

Learning to localize cars and pedestrians

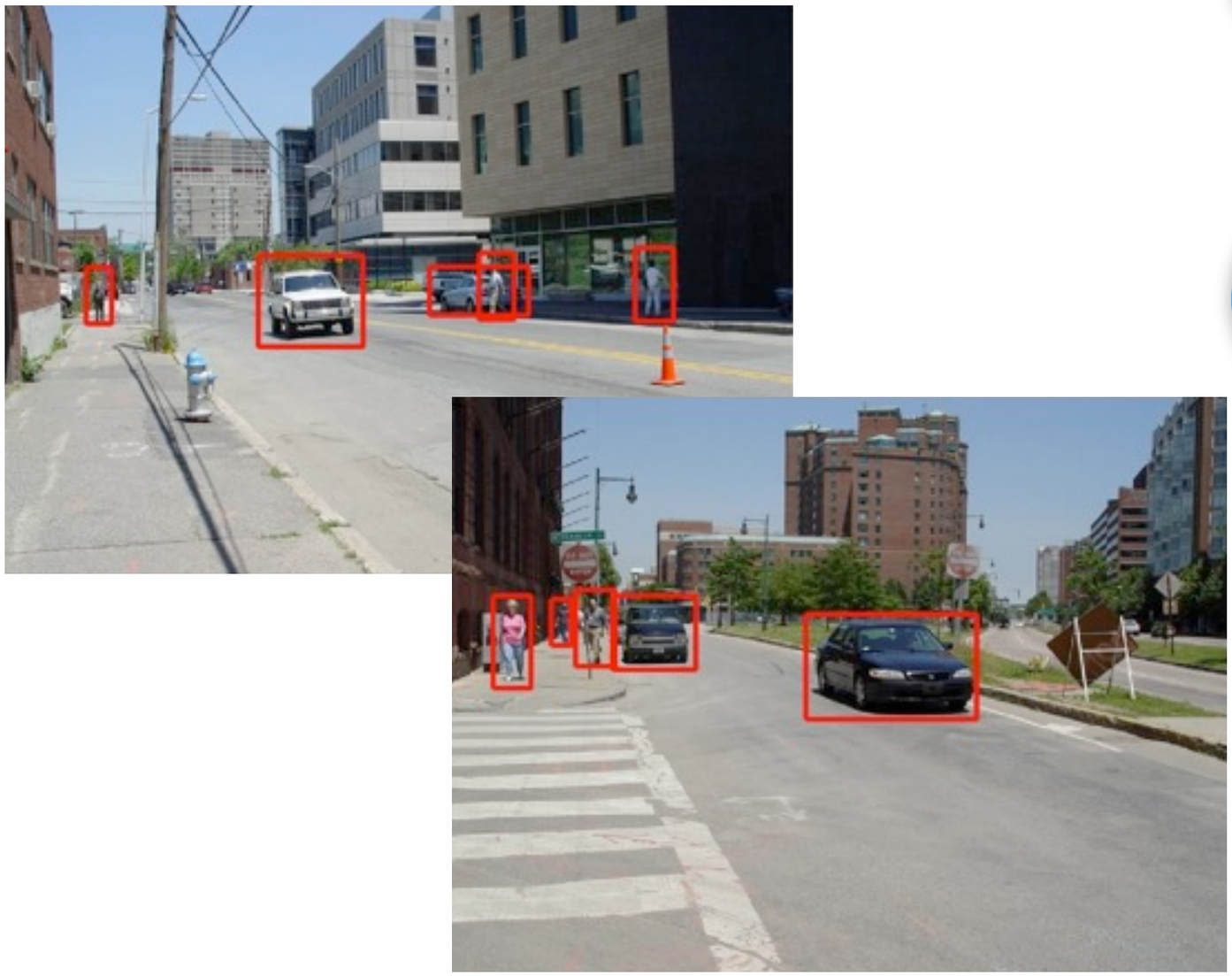


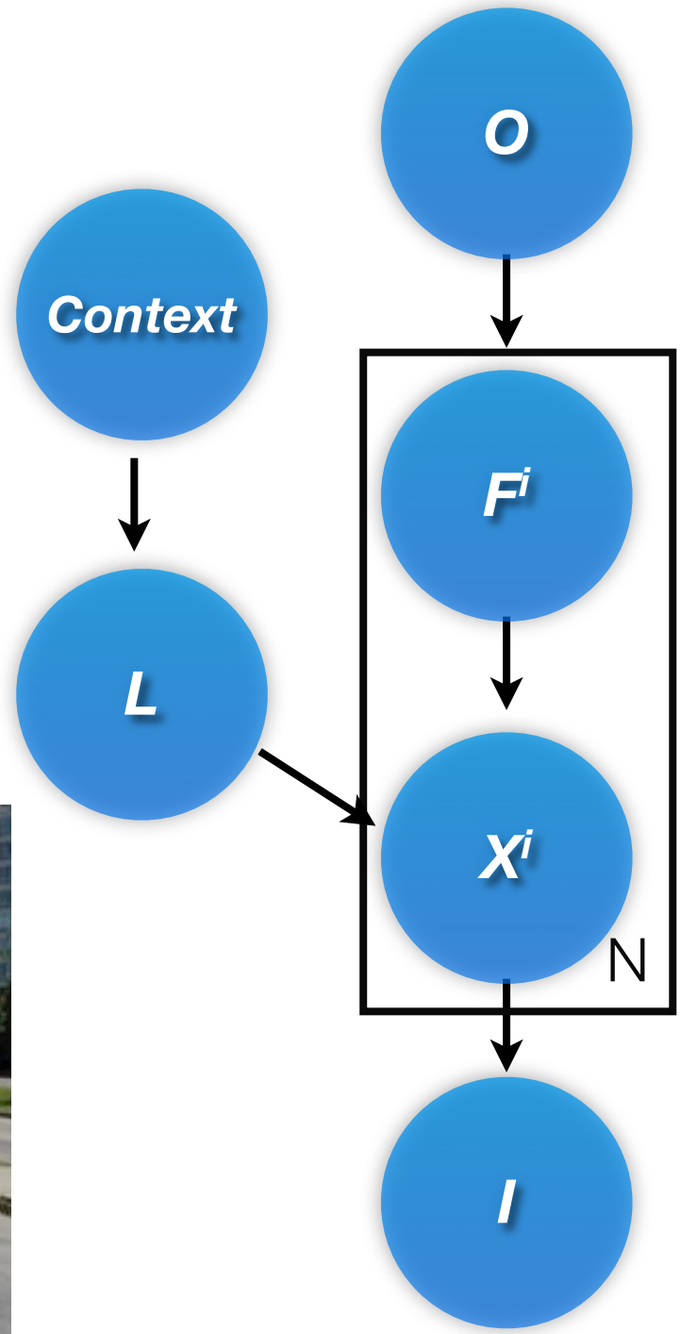


learning object priors

Learning to localize cars and pedestrians

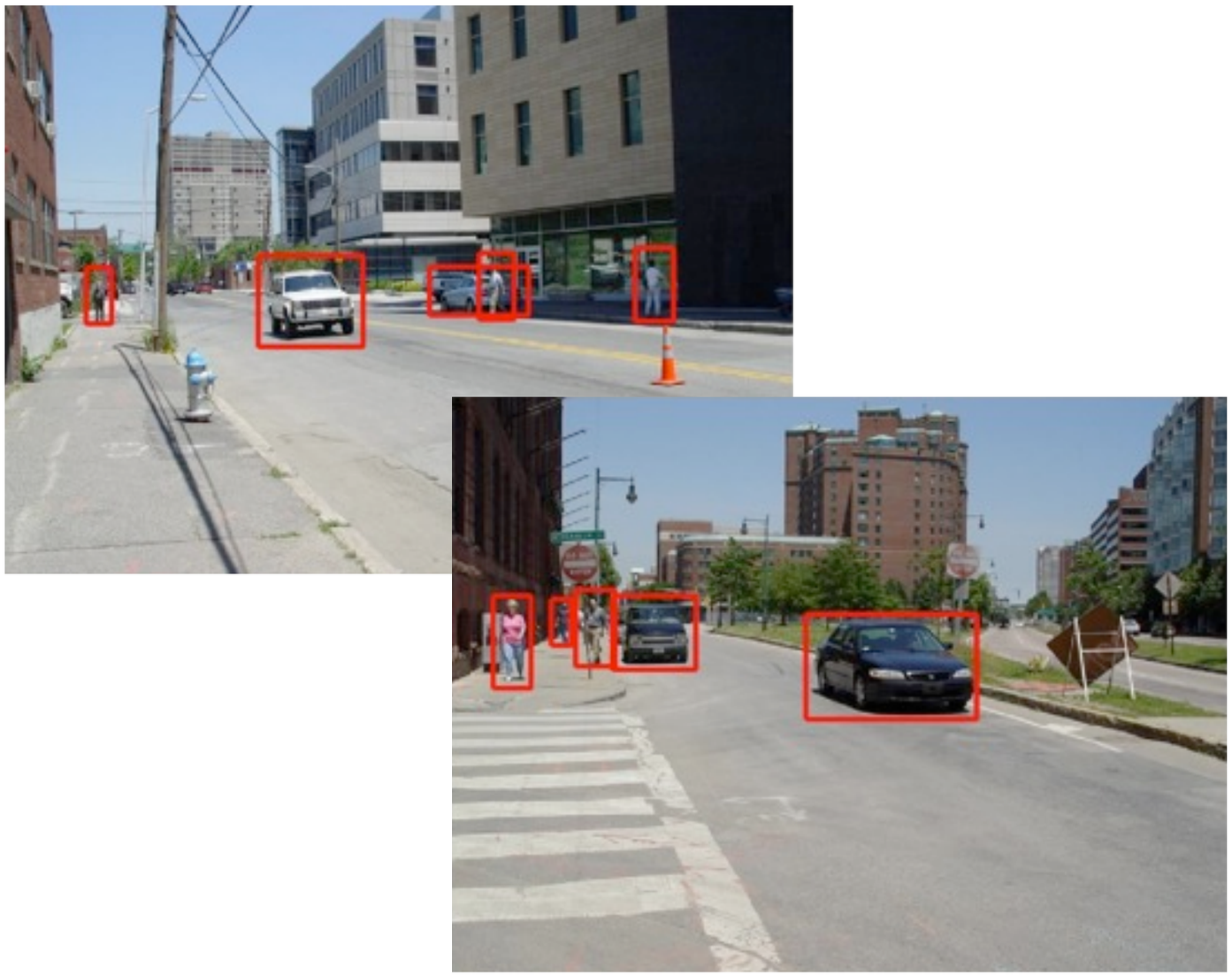
learning location priors
(global contextual cues,
see Torralba, Oliva et al)





learning object
priors

Learning to localize cars and pedestrians



The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

1st three fixations

Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

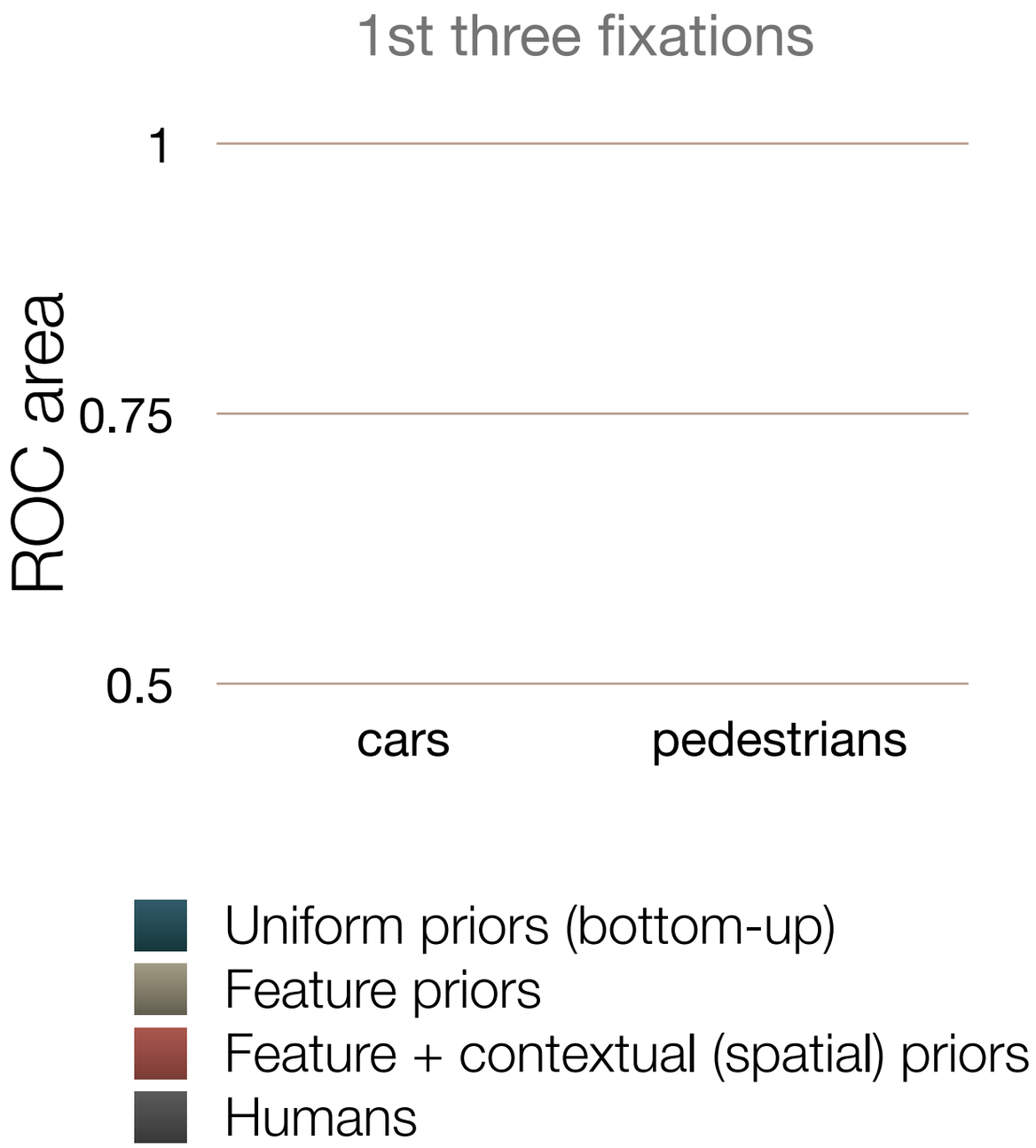
1st three fixations

ROC area

Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

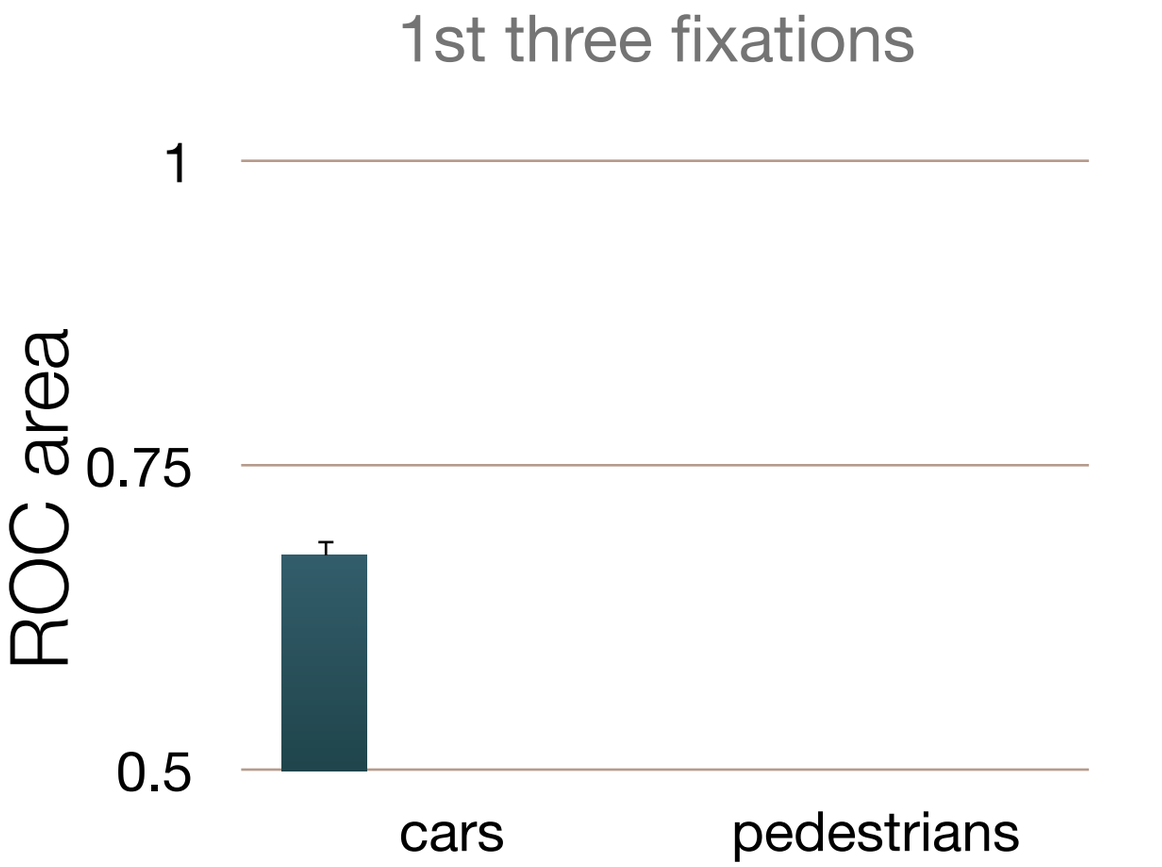
The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



The experiment

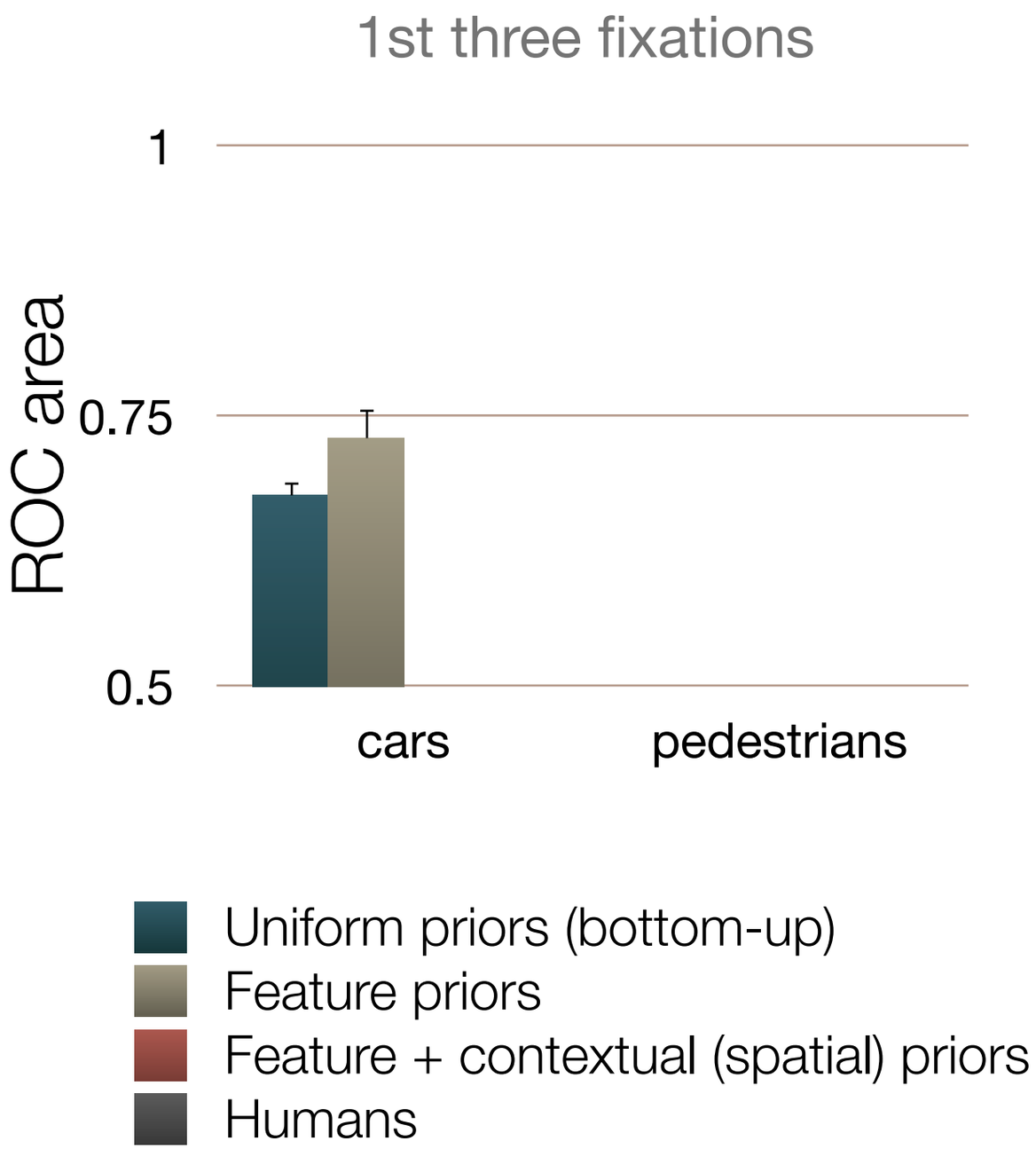
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

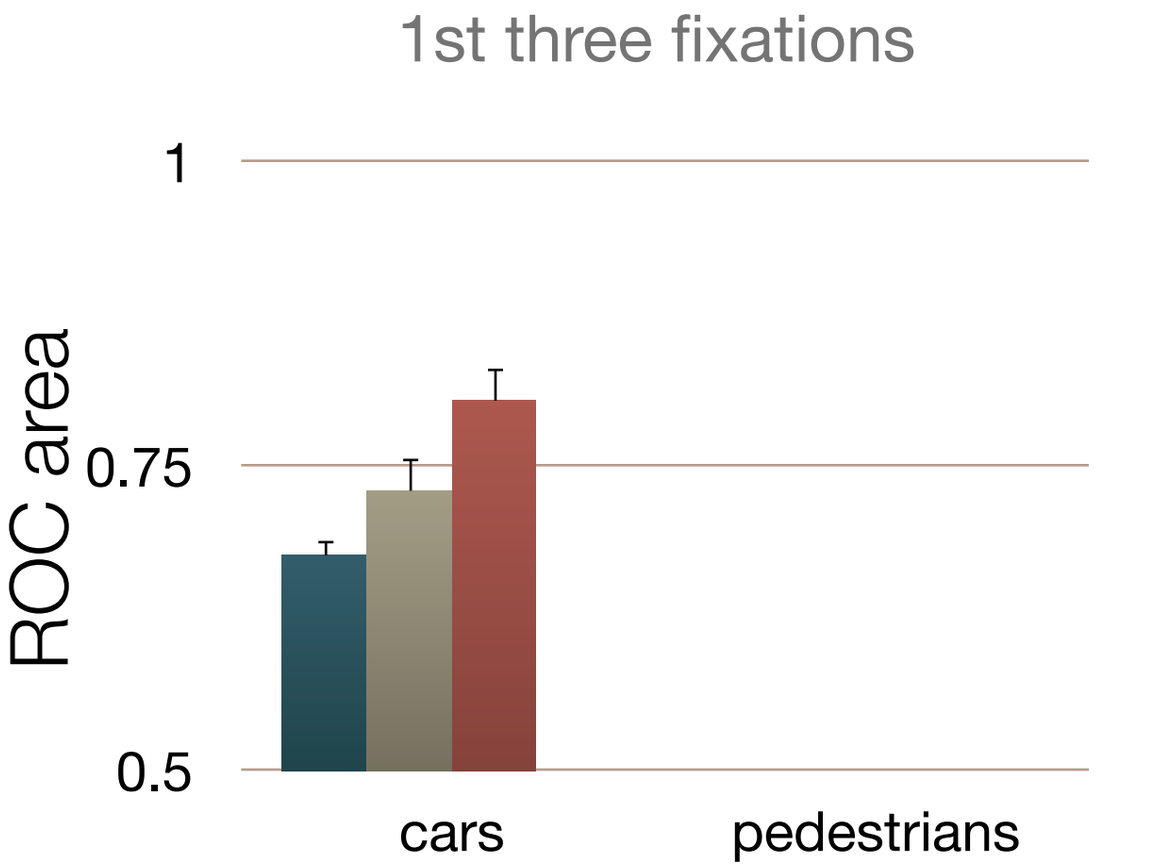
The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



The experiment

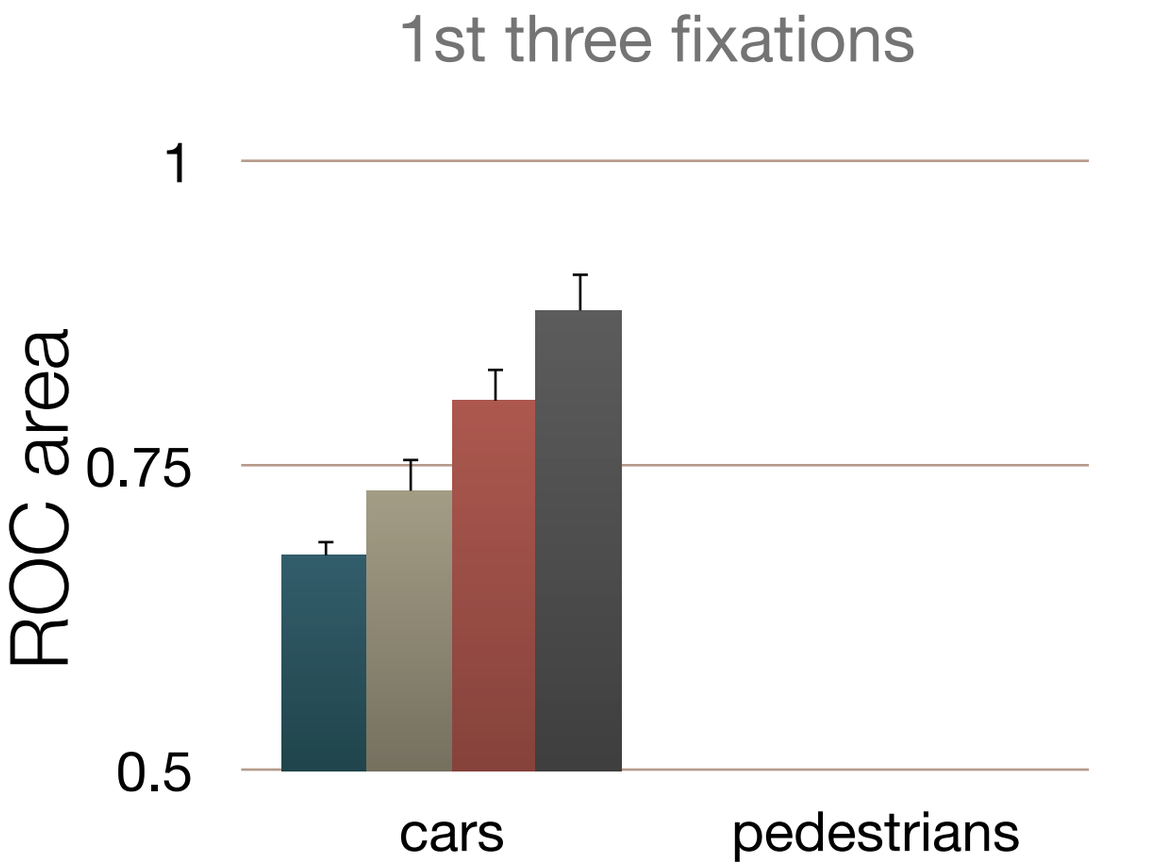
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

The experiment

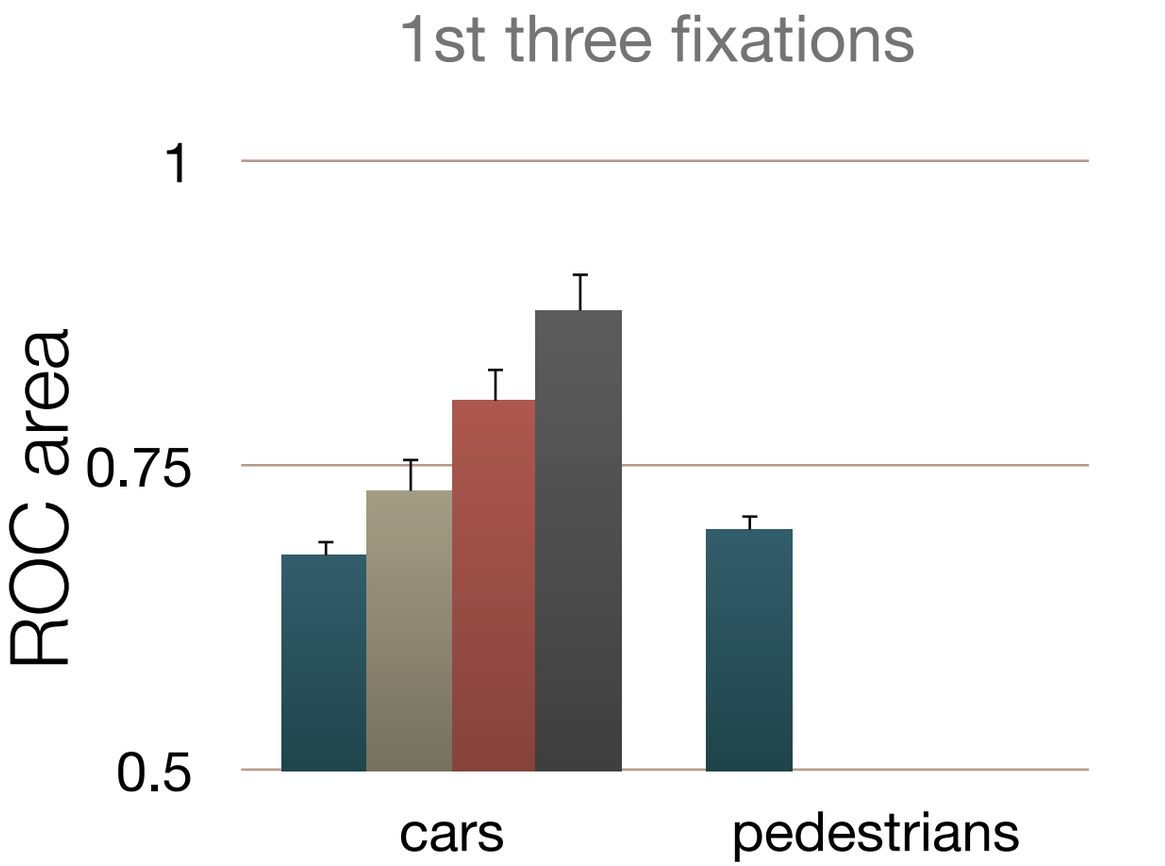
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

The experiment

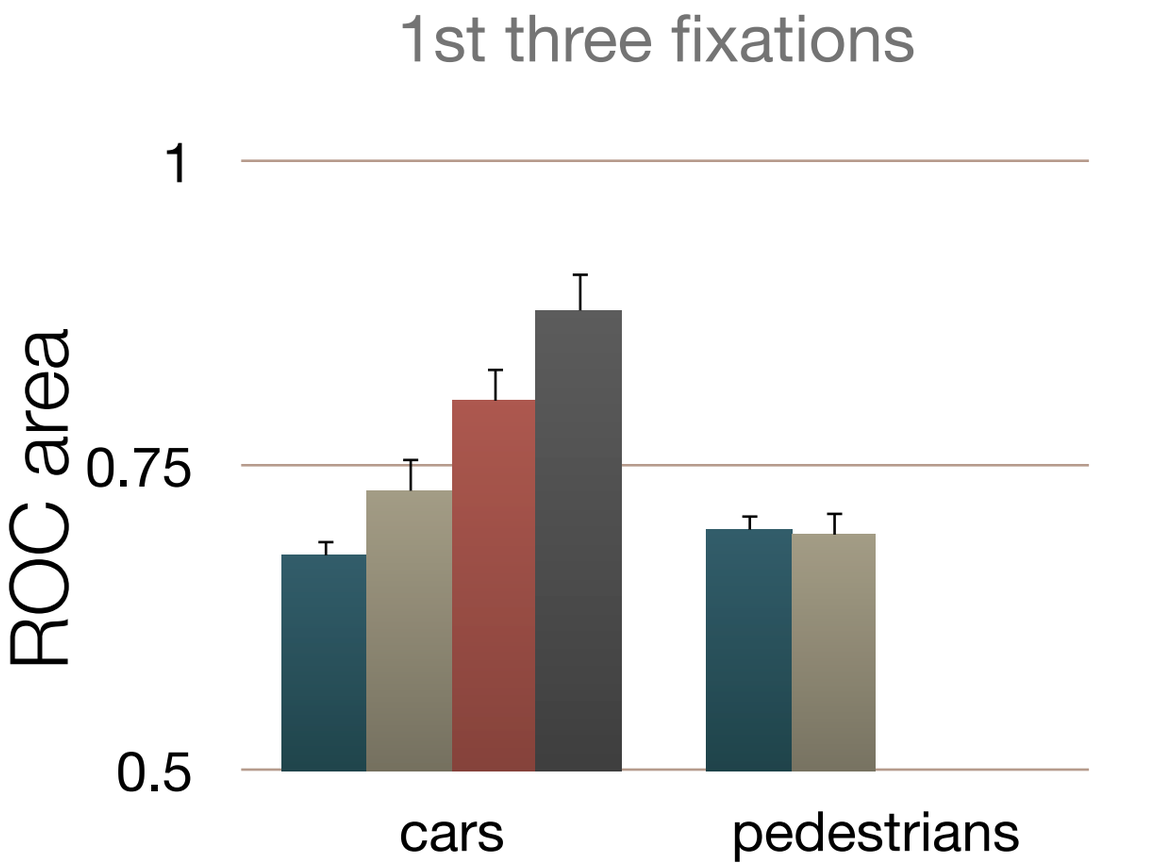
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

The experiment

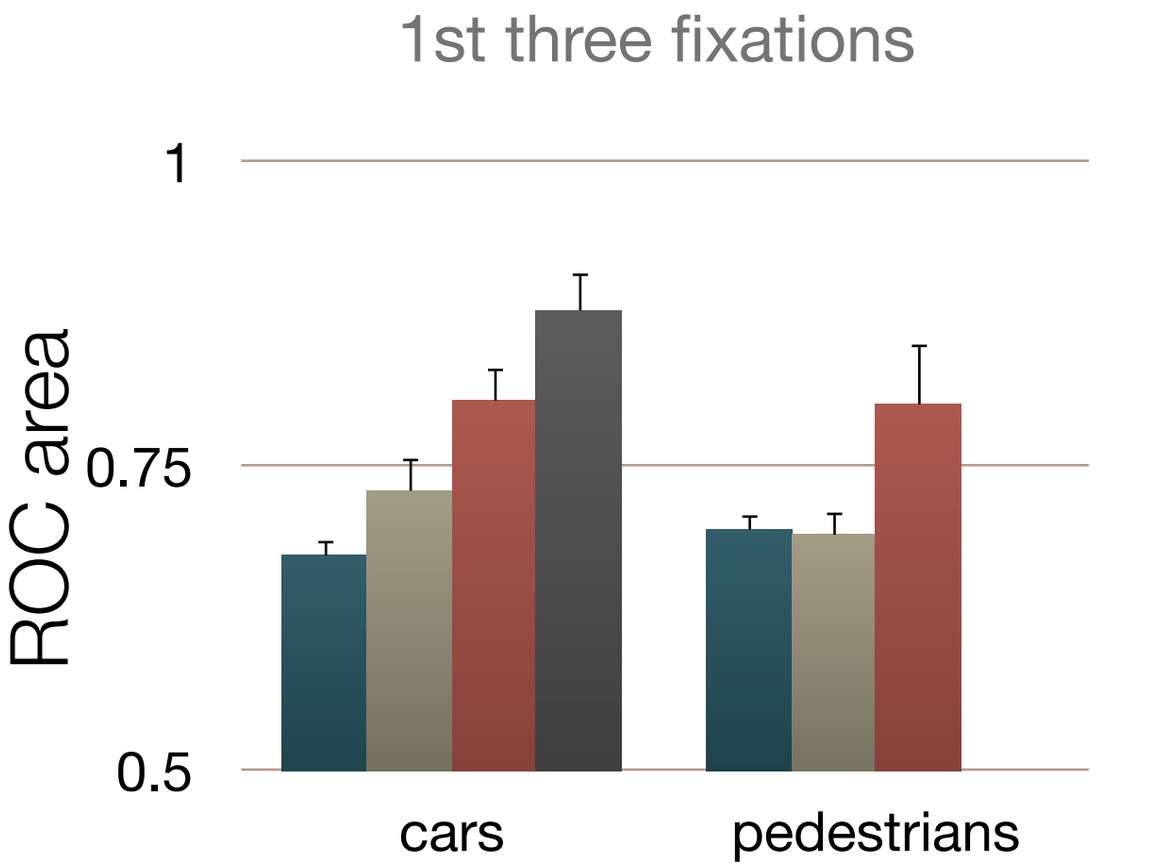
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

The experiment

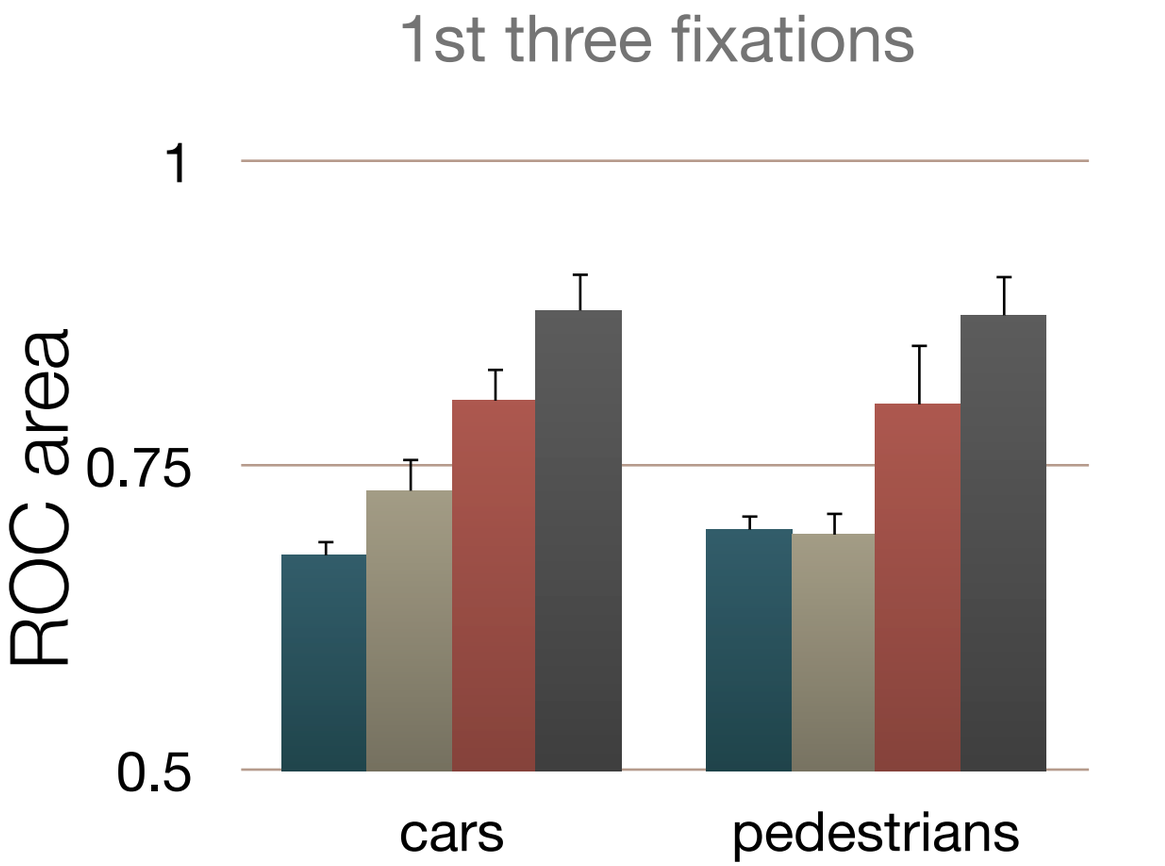
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker



Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

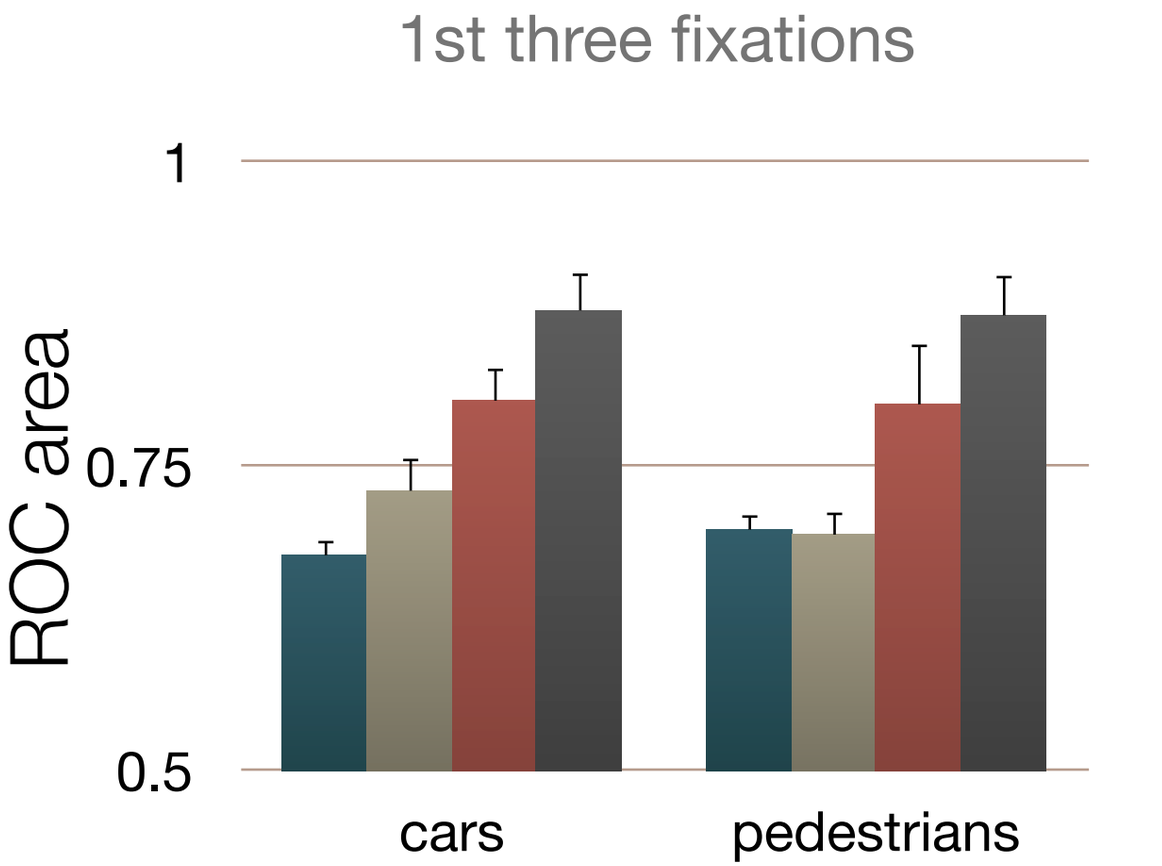


Uniform priors (bottom-up)
Feature priors
Feature + contextual (spatial) priors
Humans

The experiment

- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

Explains 92% of the inter-subject agreement!



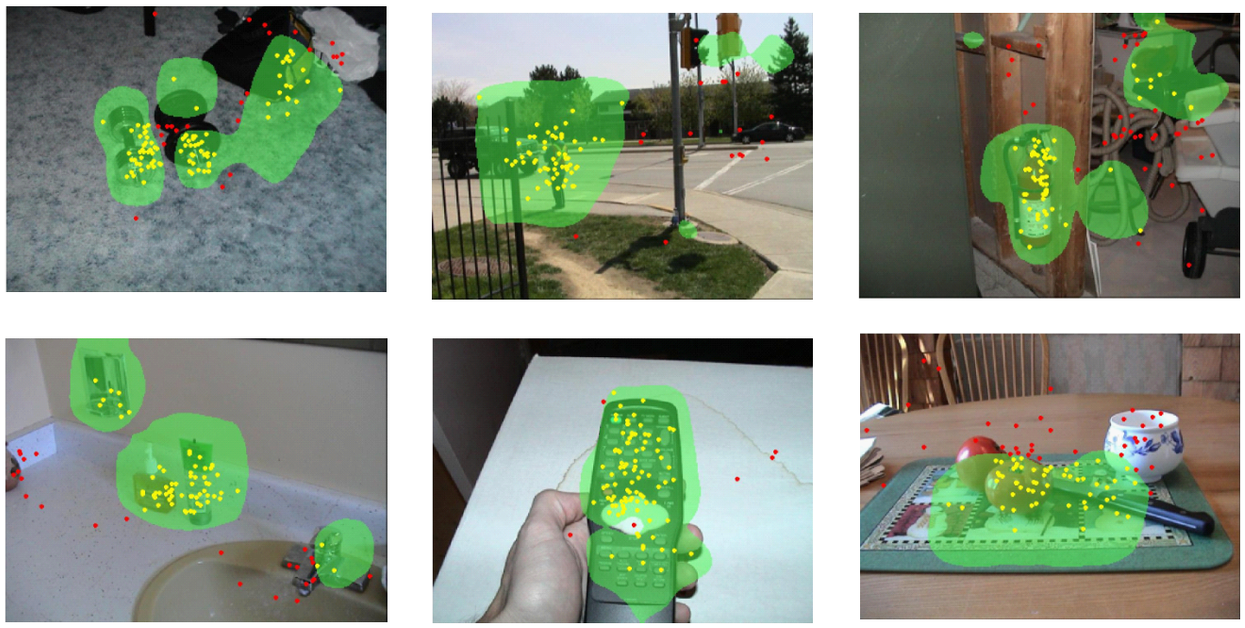
- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

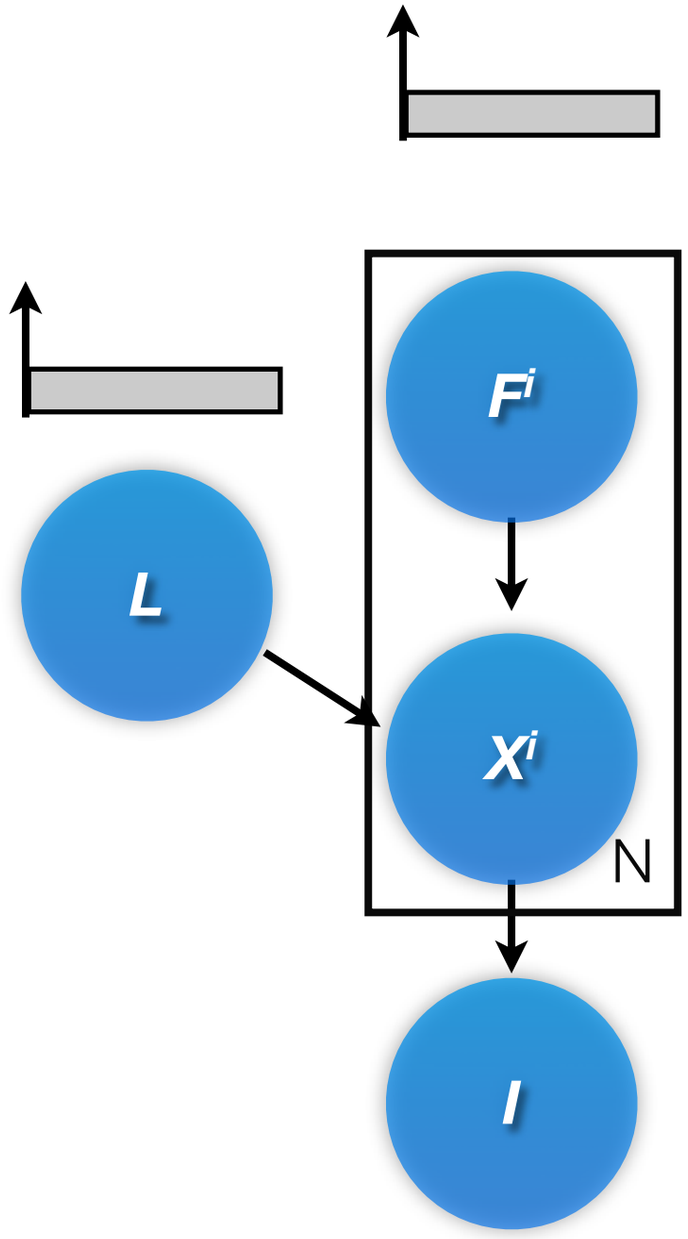
- Eye movements as proxy for attention
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/pedestrians
  - block design for cars and pedestrians
  - eye movements recorded using an infra-red eye tracker

The experiment

*similar (independent) results by Ehinger Hidalgo Torralba & Oliva ’10

Method	ROC area
Bruce & Tsotos ’06	72.8%
Itti et al ’01	72.7%
Proposed	**77.9%**





Bottom-up saliency and free-viewing

human eye data from Bruce & Tsotsos

Summary I

Summary I

- ***Goal of vision:***
  - To solve the problem of ’what is where’ problem

Summary I

- ***Goal of vision:***
  - To solve the problem of 'what is where' problem
- ***Key assumptions:***
  - 'Attentional spotlight' → recognition done sequentially, one object at a time
  - 'What' and 'where' independent pathways

Summary I

- ***Goal of vision:***
  - To solve the problem of 'what is where' problem
- ***Key assumptions:***
  - 'Attentional spotlight' → recognition done sequentially, one object at a time
  - 'What' and 'where' independent pathways
- Attention as the inference process implemented by the interaction between ventral and dorsal areas
  - Integrates bottom-up and top-down (feature-based and context-based) attentional mechanisms
  - Seems consistent with known physiology

Summary I

- ***Goal of vision:***
  - To solve the problem of 'what is where' problem
- ***Key assumptions:***
  - 'Attentional spotlight' → recognition done sequentially, one object at a time
  - 'What' and 'where' independent pathways
- Attention as the inference process implemented by the interaction between ventral and dorsal areas
  - Integrates bottom-up and top-down (feature-based and context-based) attentional mechanisms
  - Seems consistent with known physiology
- Main attentional effects in the presence of clutters
  - Spatial attention reduces uncertainty over location and improves object recognition performance over first bottom-up (feedforward) pass

Part I: A ‘Bayesian’ theory of attention that solves the ‘what is where’ problem

- Predictions for neurophysiology and human eye movement data

- Sharat Chikkerur
- Cheston Tan
- Tomaso Poggio

Part II: Readout of IT population activity

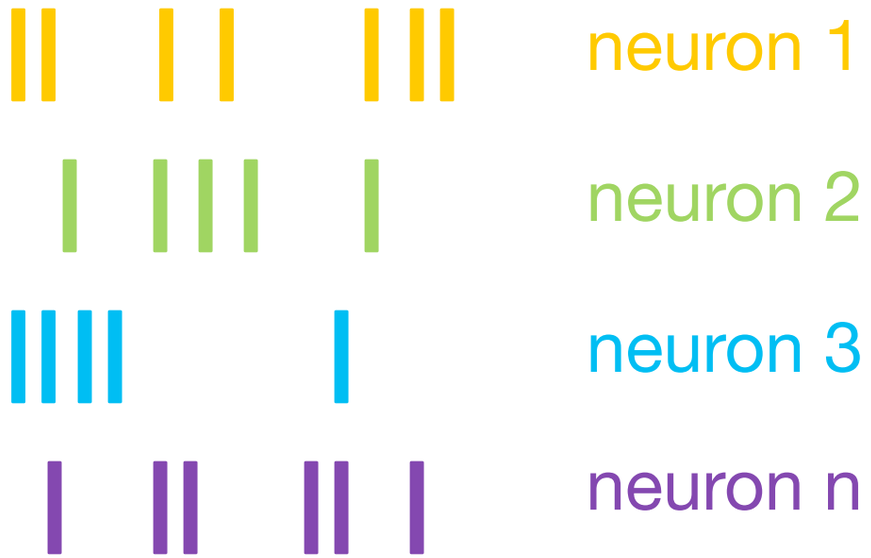
- Attention eliminates clutter

- Ying Zhang
- Ethan Meyers
- Narcisse Bichot
- Tomaso Poggio
- Bob Desimone

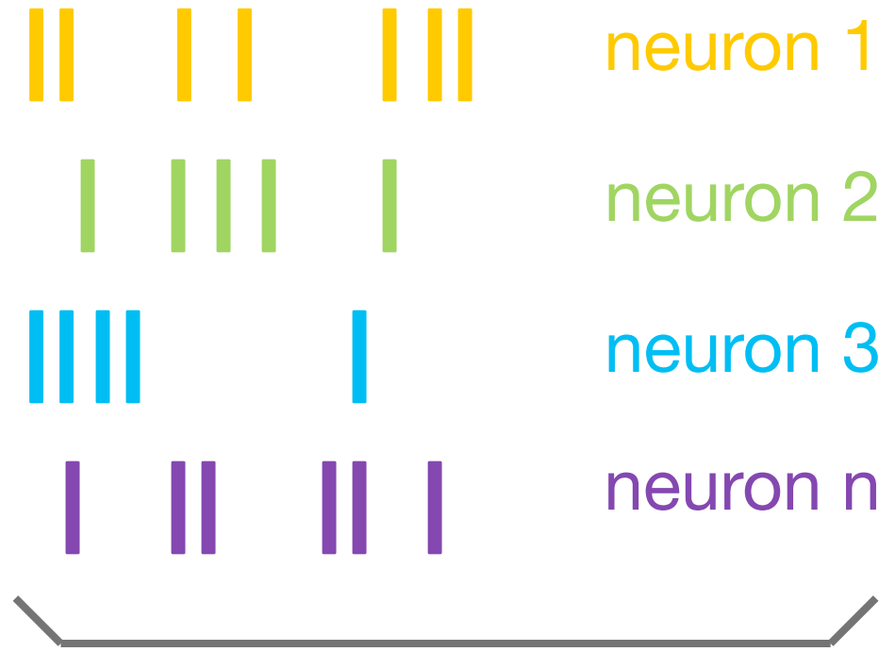
The ‘readout’ approach

The ‘readout’ approach

The ‘readout’ approach

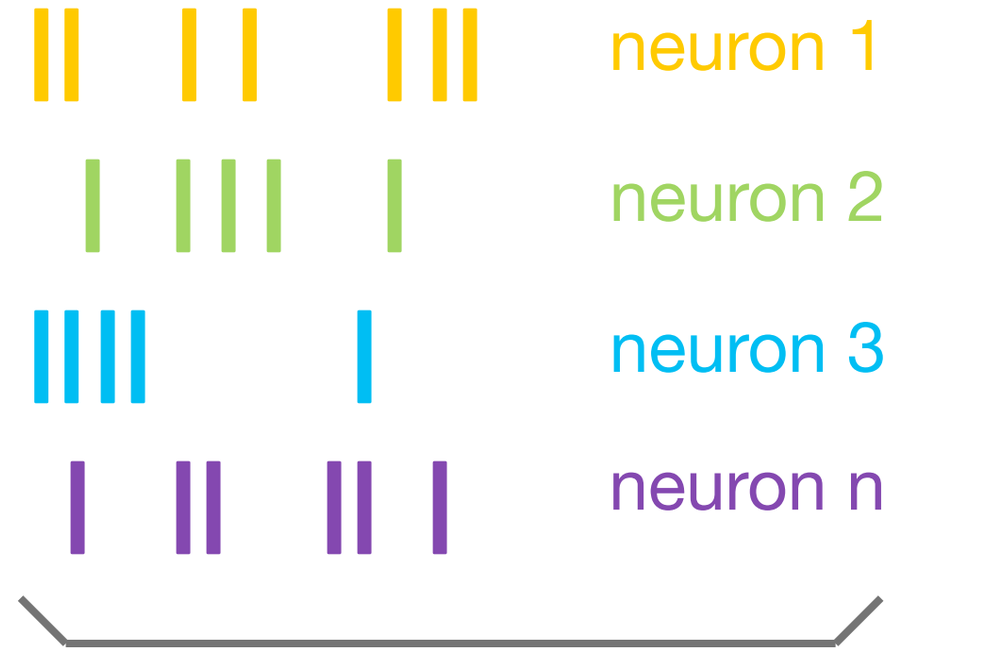


The ‘readout’ approach



Pattern Classifier

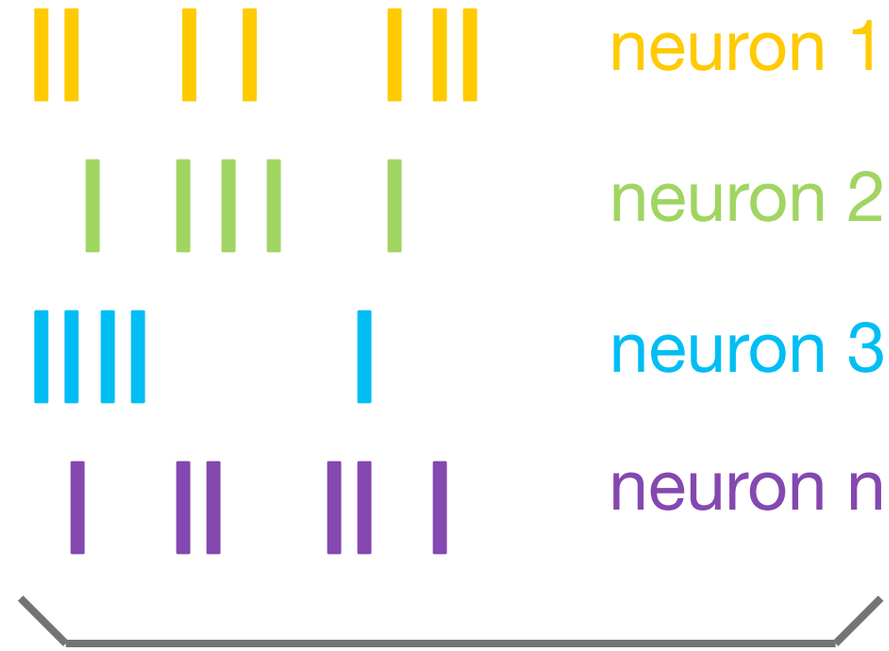
The ‘readout’ approach



Pattern Classifier

Prediction

The ‘readout’ approach



“Correct”

Prediction

Pattern Classifier

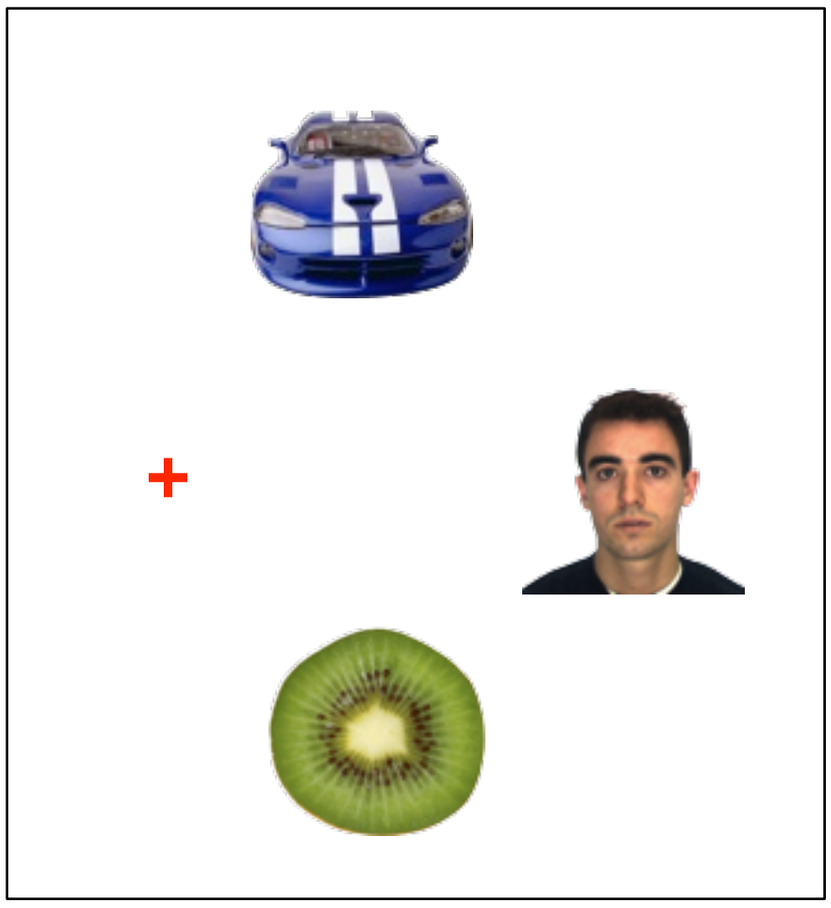
IT readout improves with attention

train readout classifier
on isolated object

IT readout improves with attention

train readout classifier on isolated object

test generalization in clutter



IT readout improves with attention

train readout classifier on isolated object

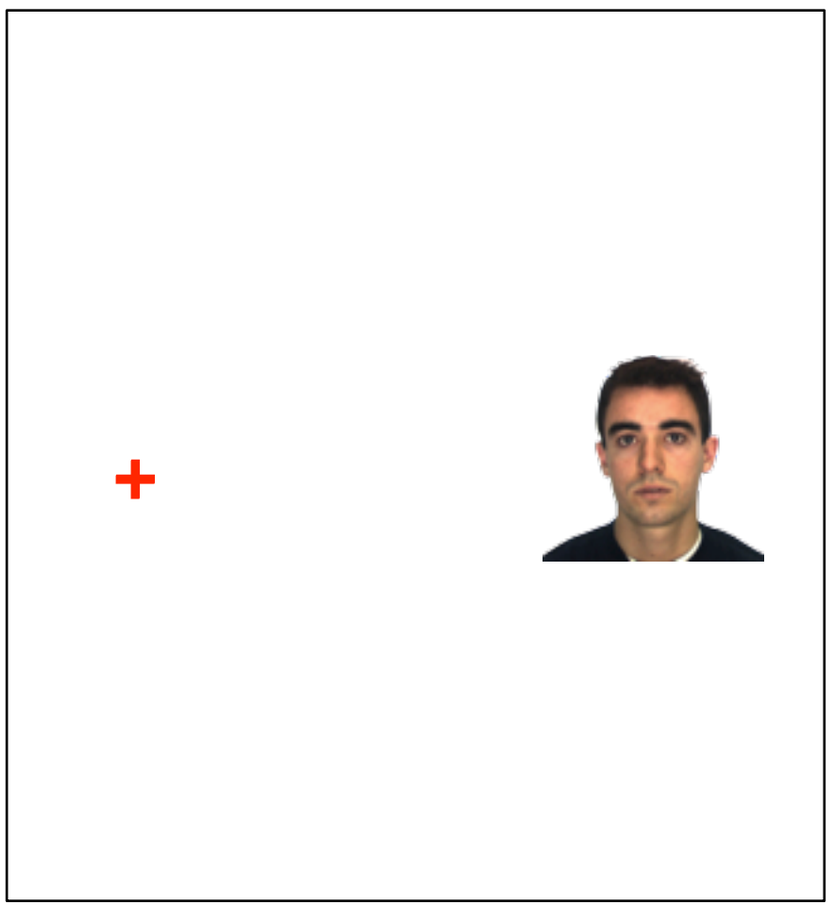
test generalization in clutter

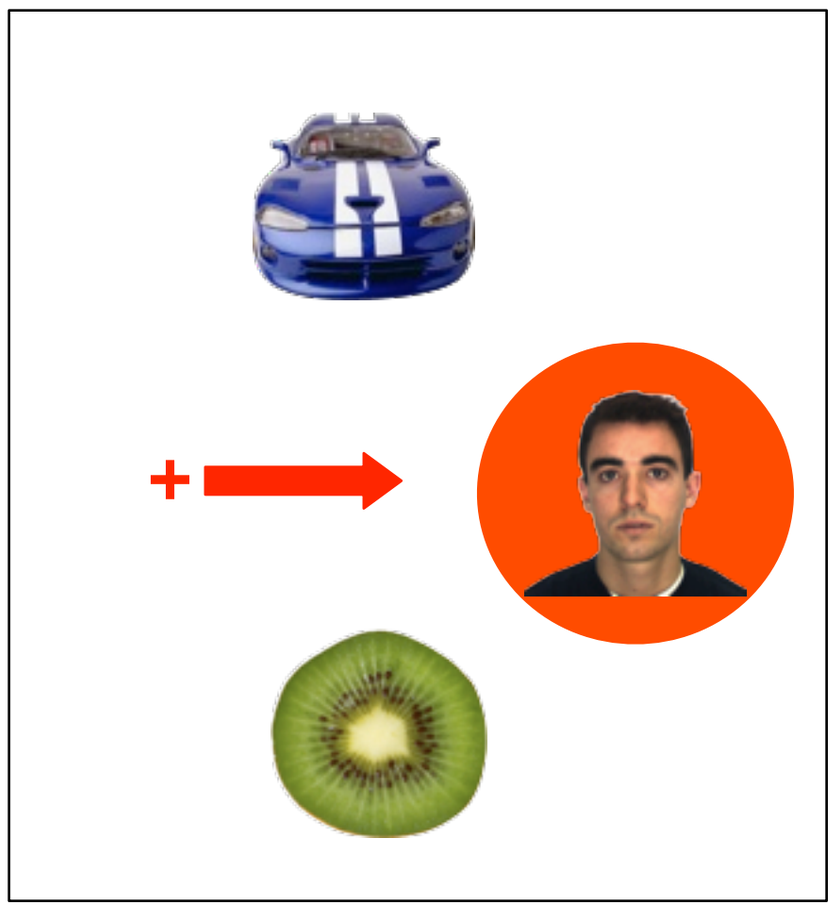
unattended

IT readout improves with attention

train readout classifier on isolated object

test generalization in clutter





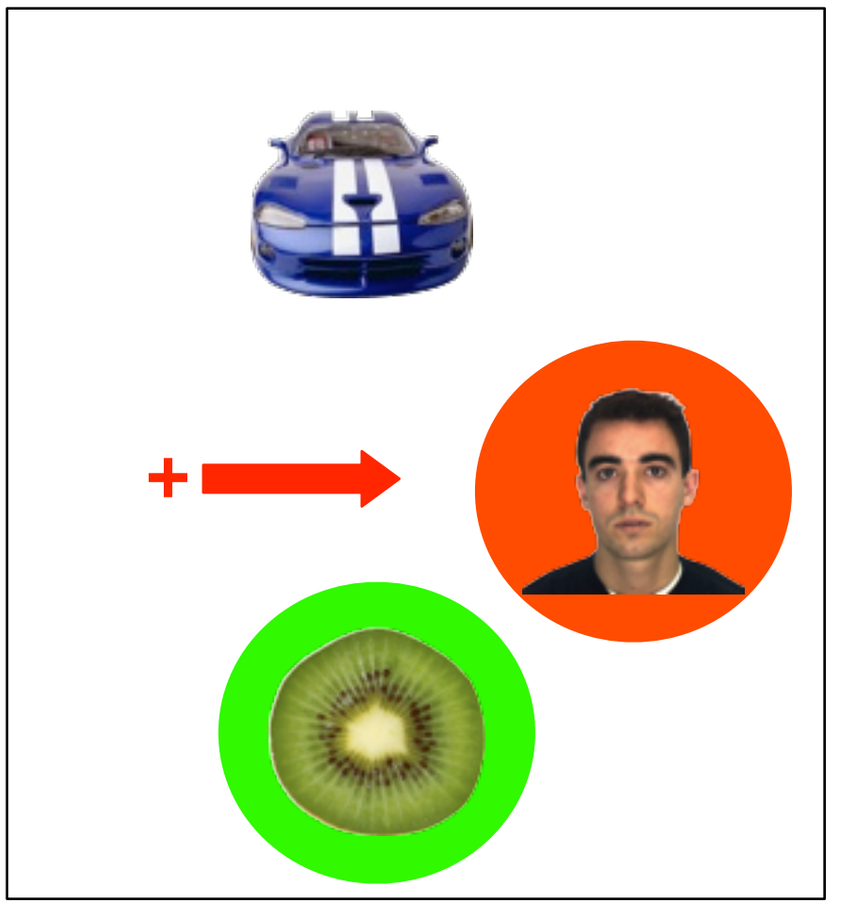
unattended

attended

IT readout improves with attention

train readout classifier on isolated object

test generalization in clutter

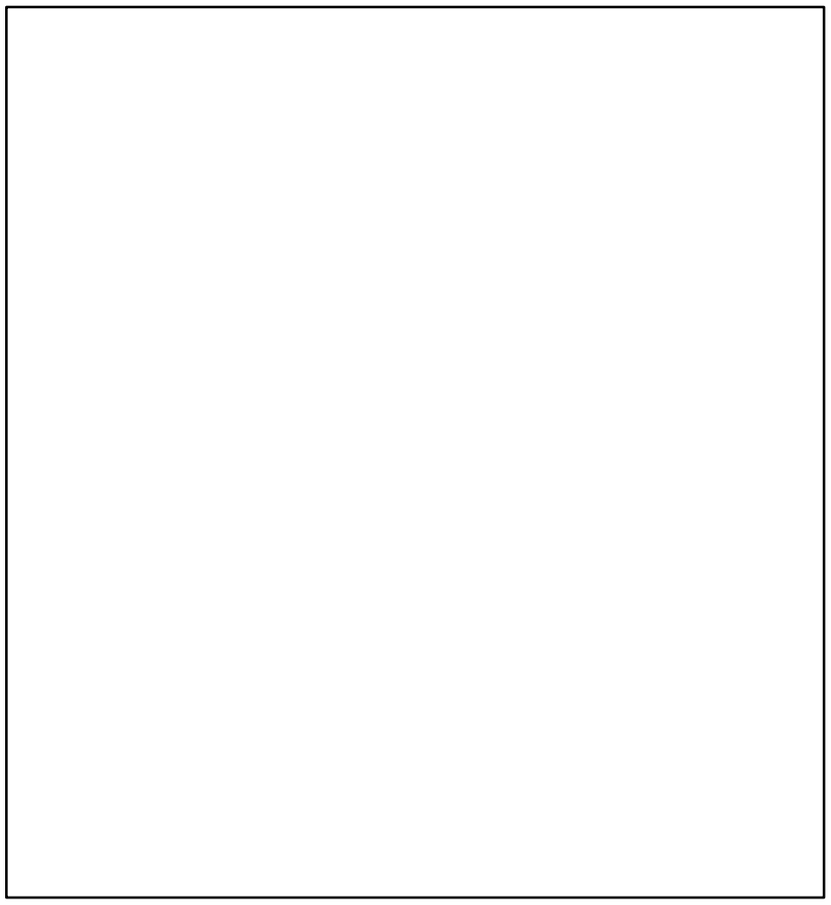


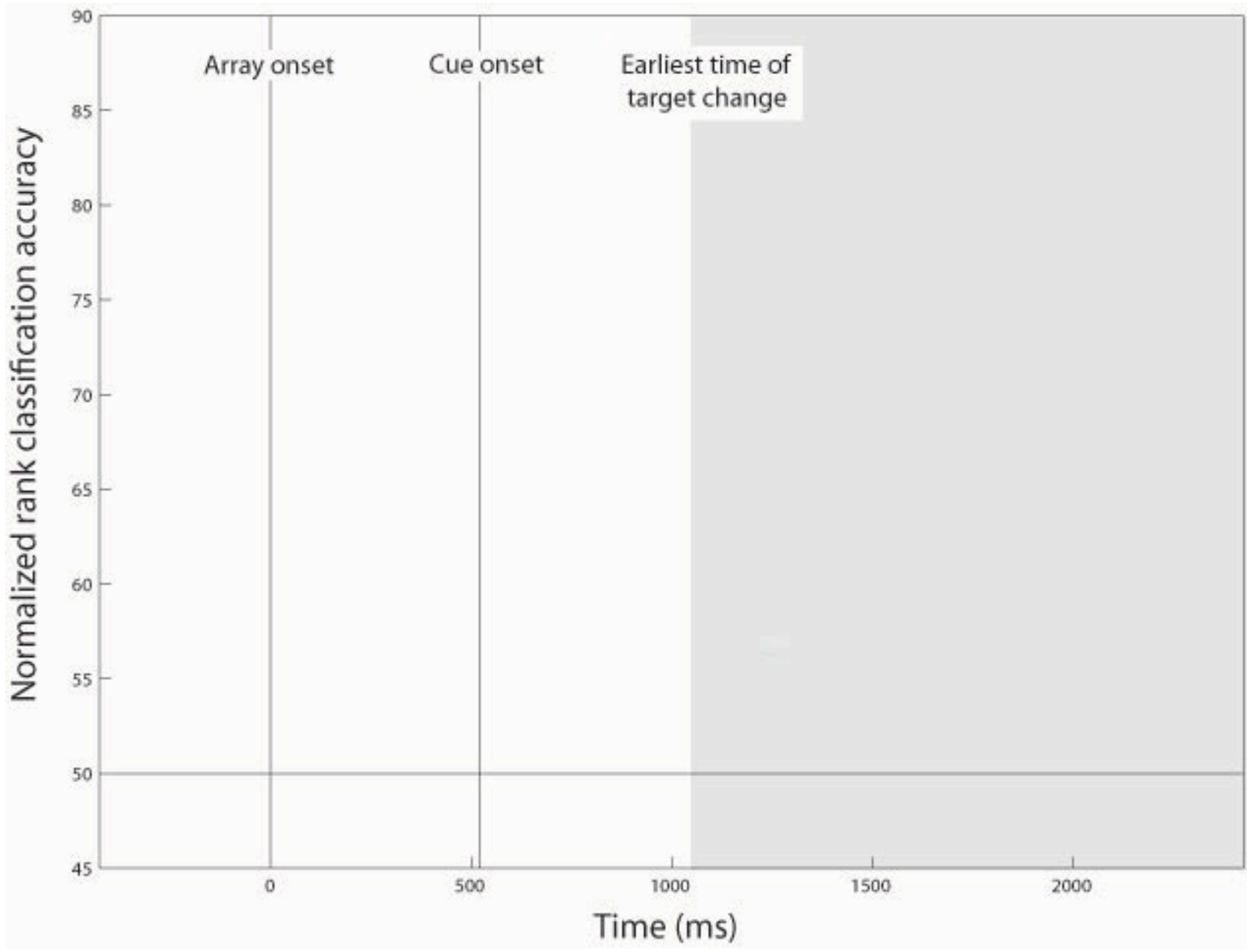
unattended

attended

attend away

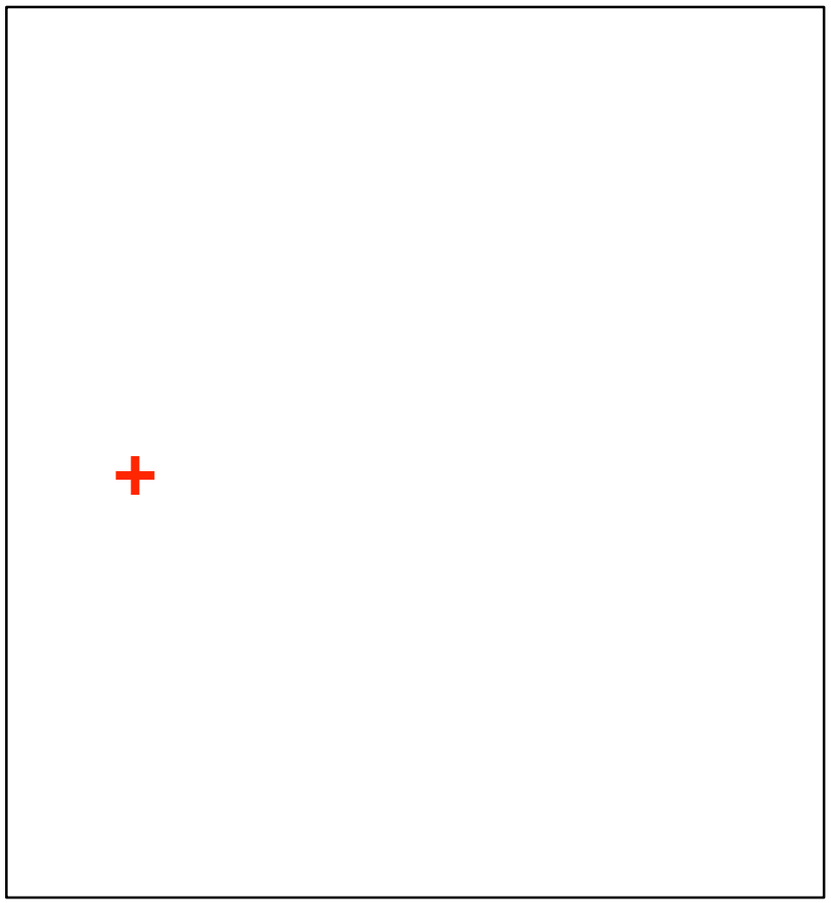
IT readout improves with attention

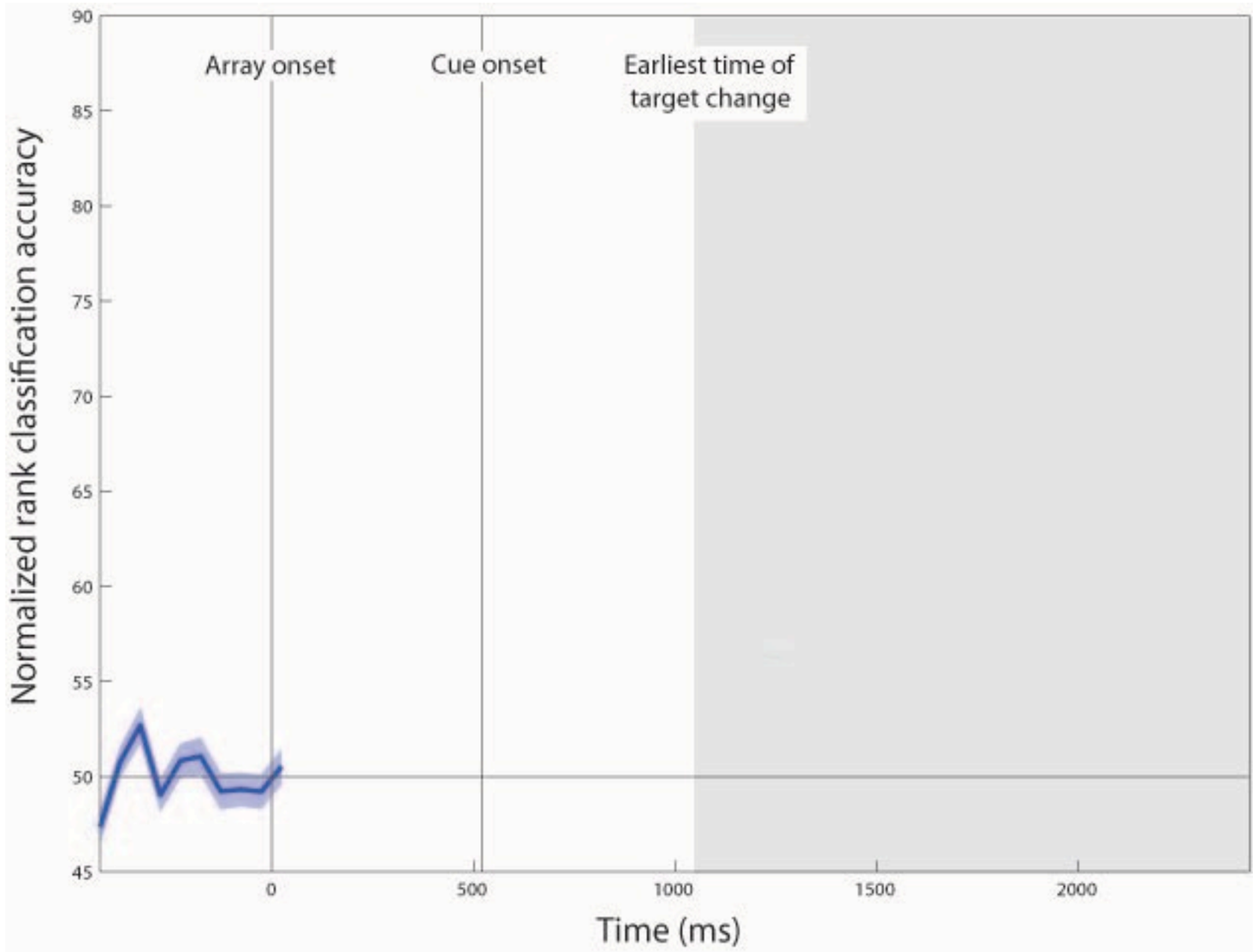




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

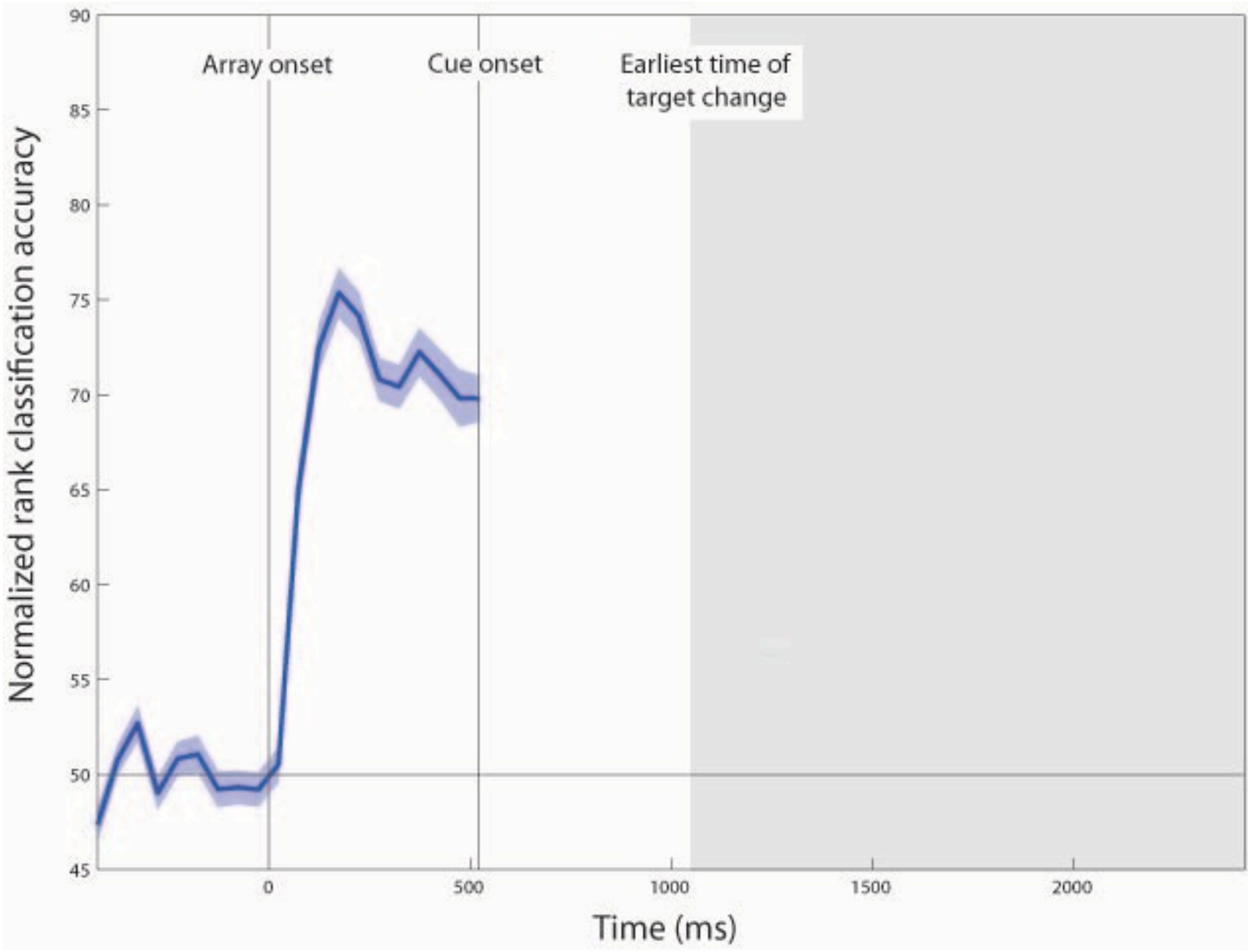
IT readout improves with attention





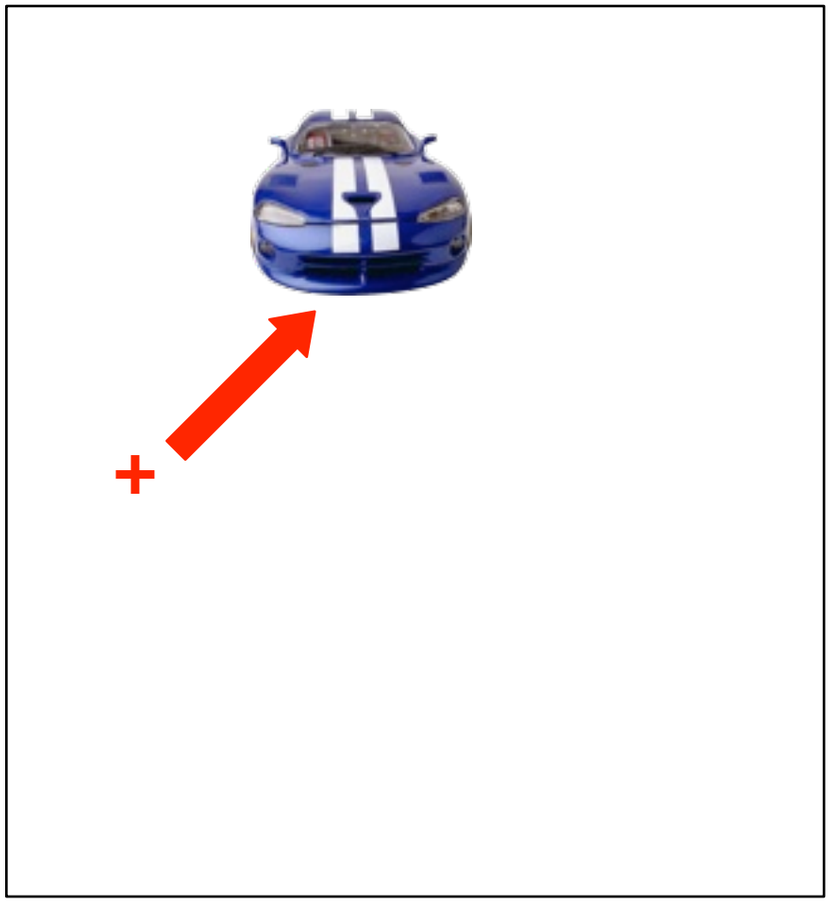
Zhang Meyers Bichot Serre Poggio Desimone unpublished data

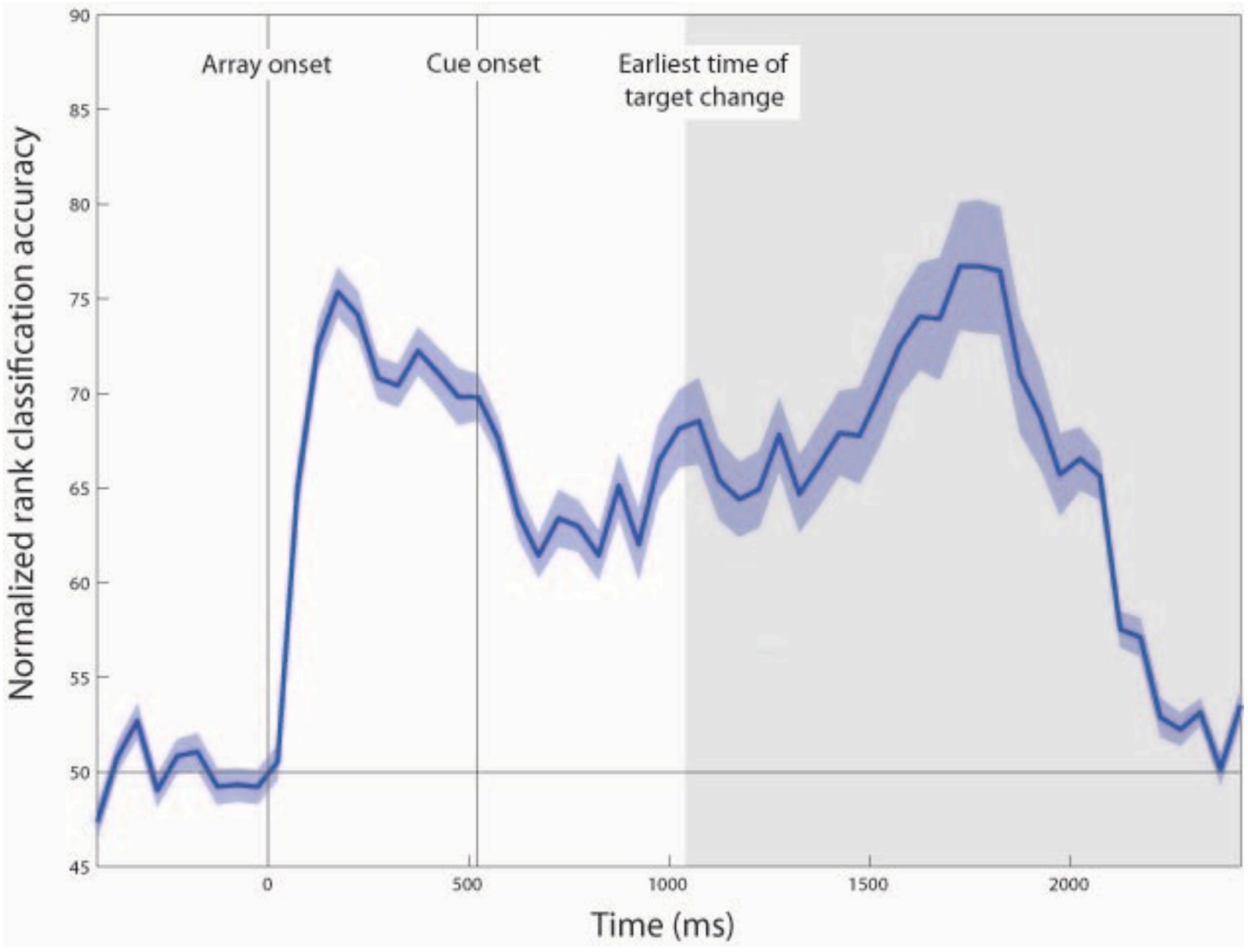
IT readout improves with attention



Zhang Meyers Bichot Serre Poggio Desimone unpublished data

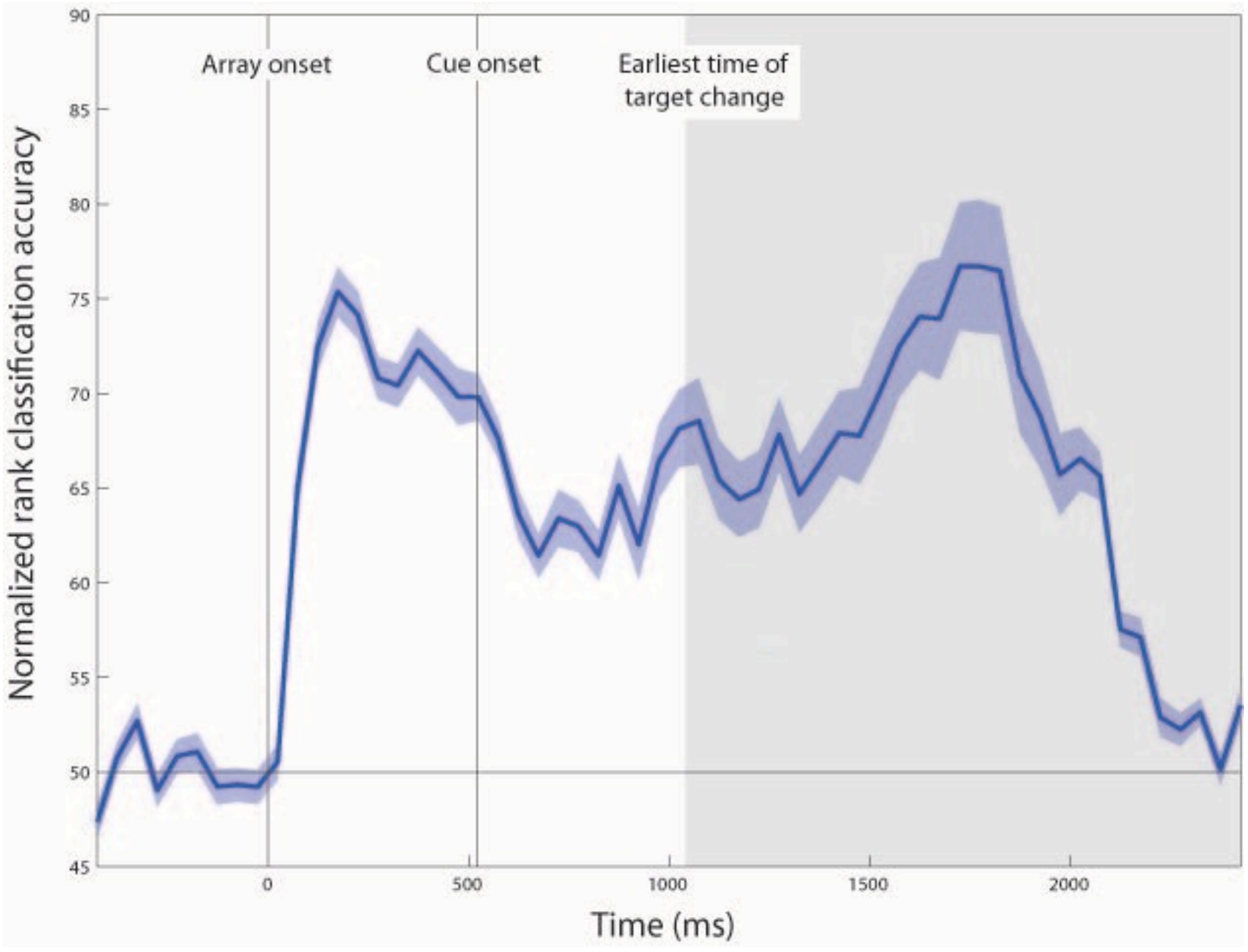
IT readout improves with attention





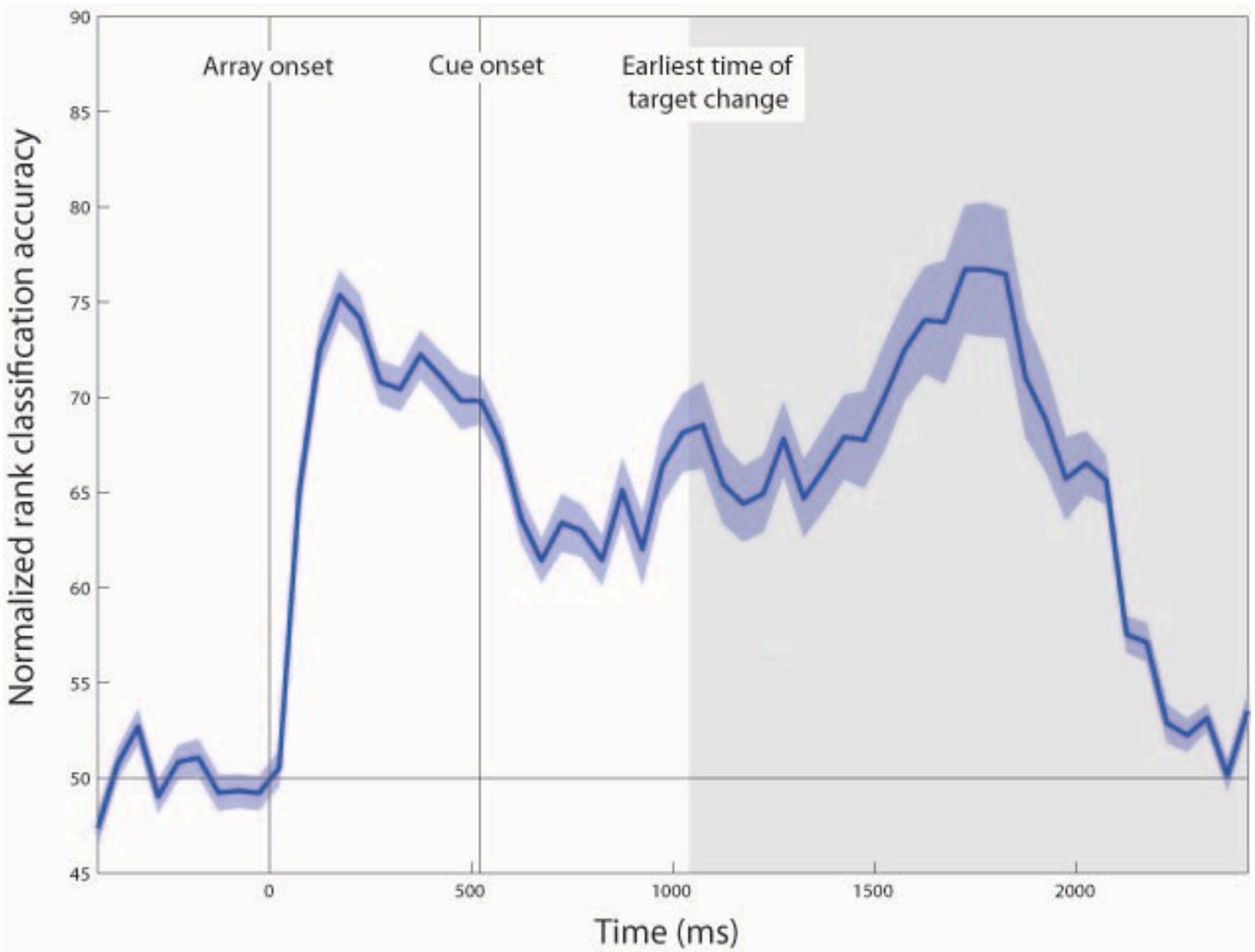
Zhang Meyers Bichot Serre Poggio Desimone unpublished data

IT readout improves with attention



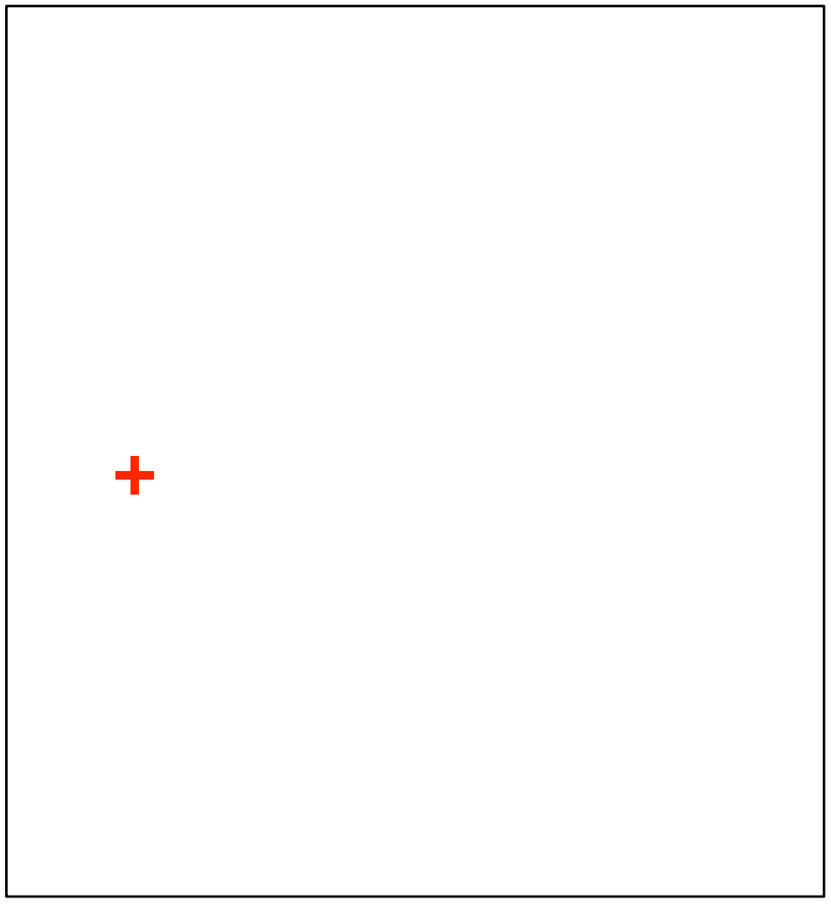
Zhang Meyers Bichot Serre Poggio Desimone unpublished data

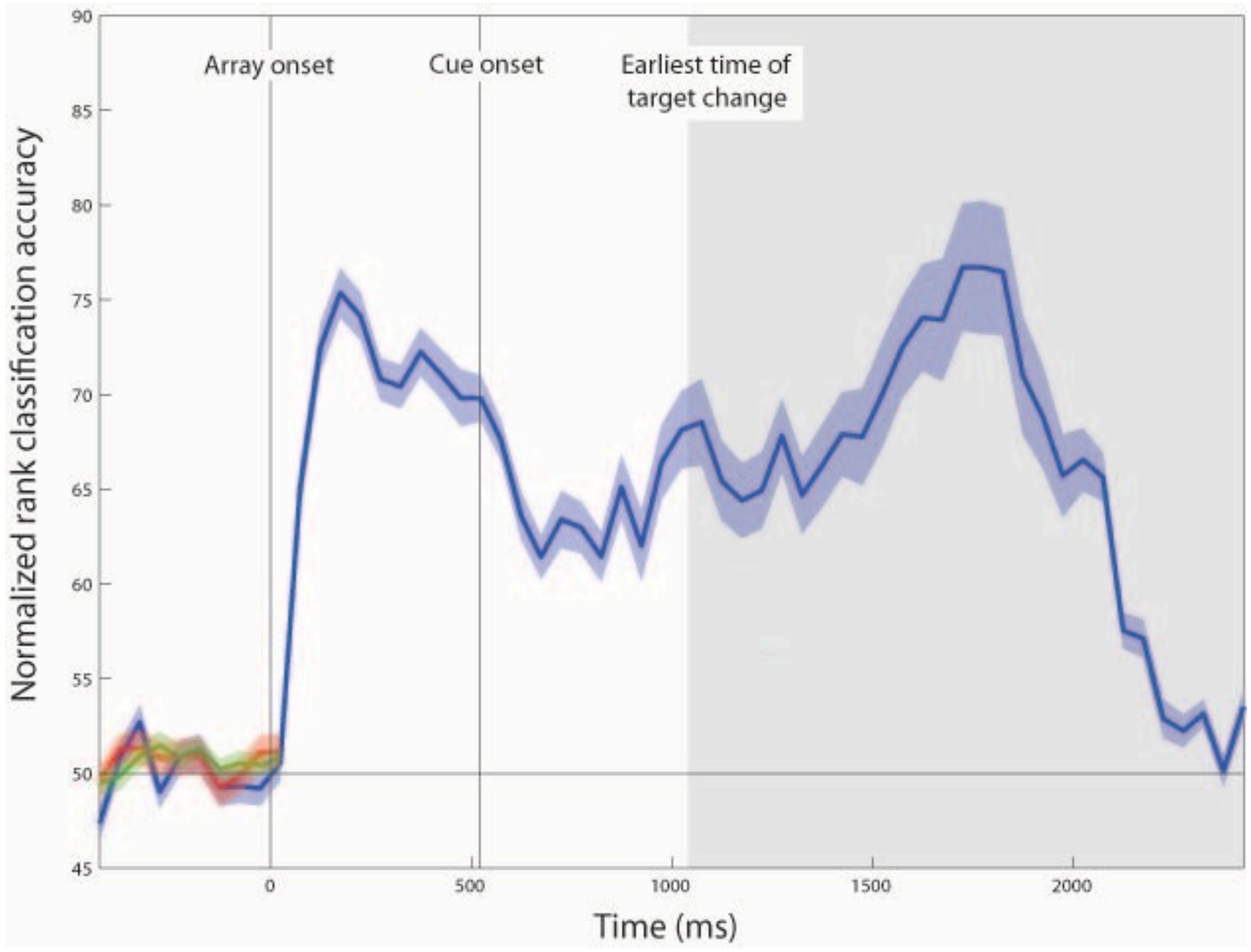
IT readout improves with attention



Zhang Meyers Bichot Serre Poggio Desimone unpublished data

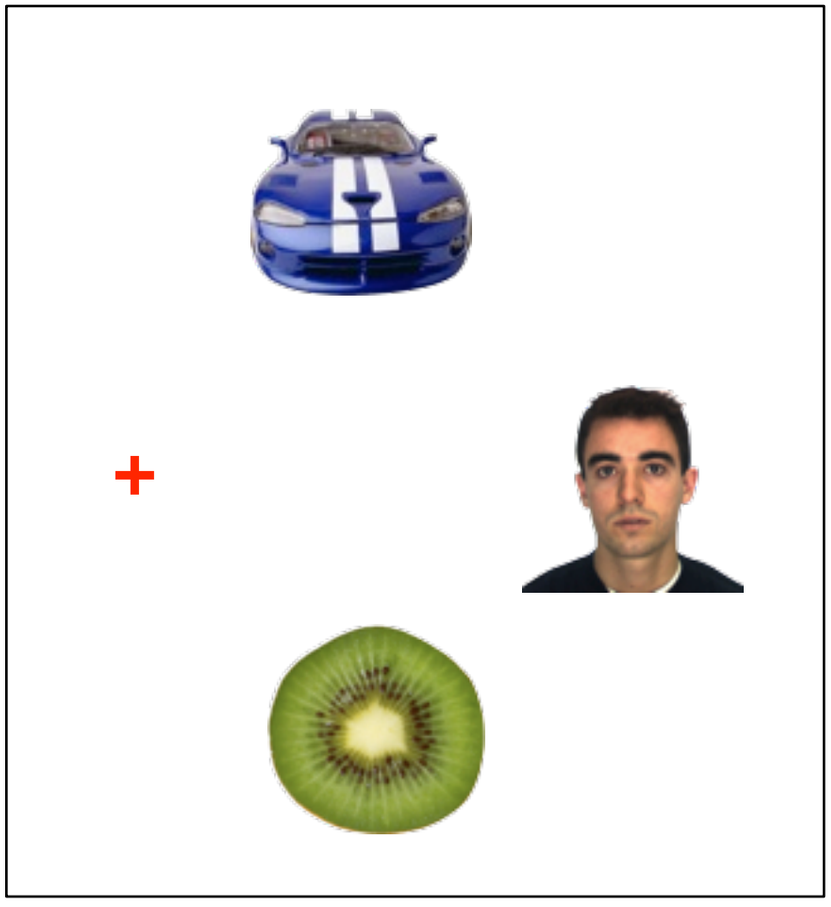
IT readout improves with attention

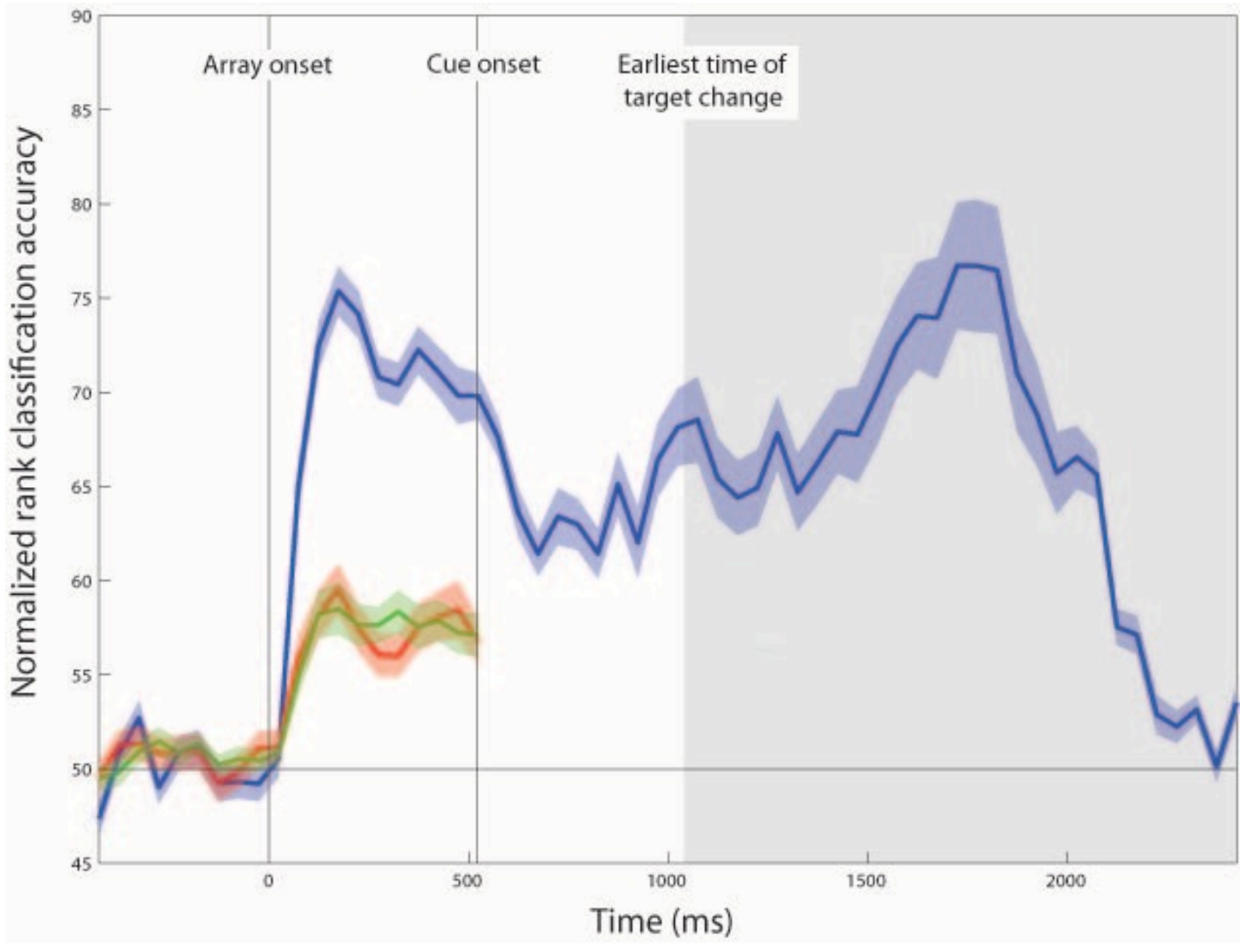




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

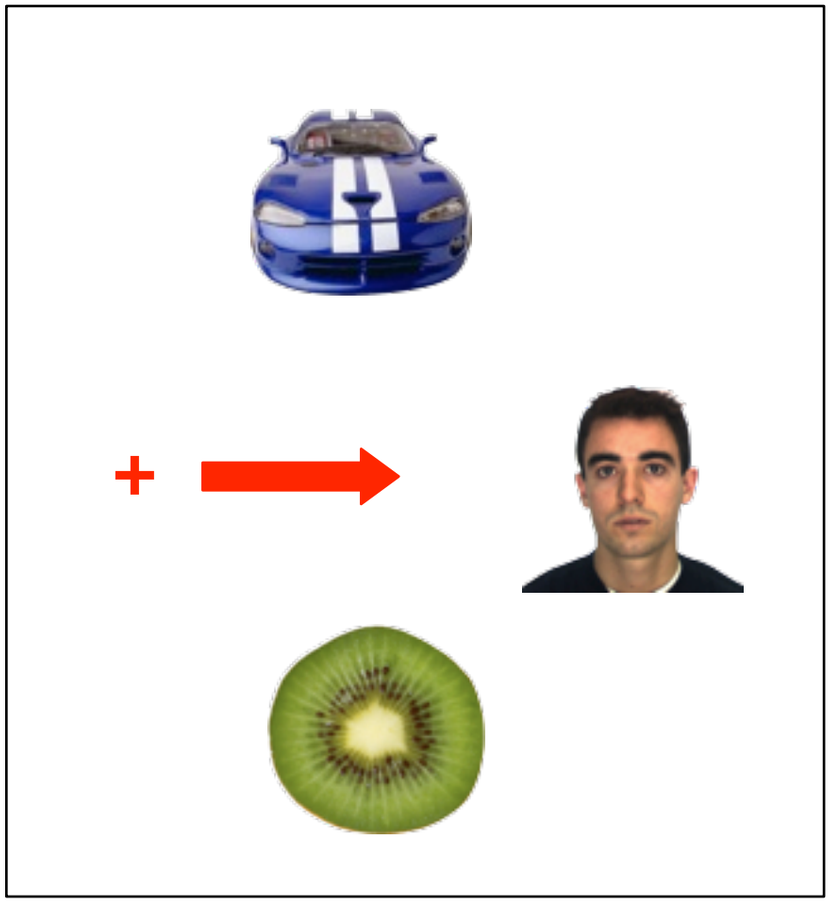
IT readout improves with attention

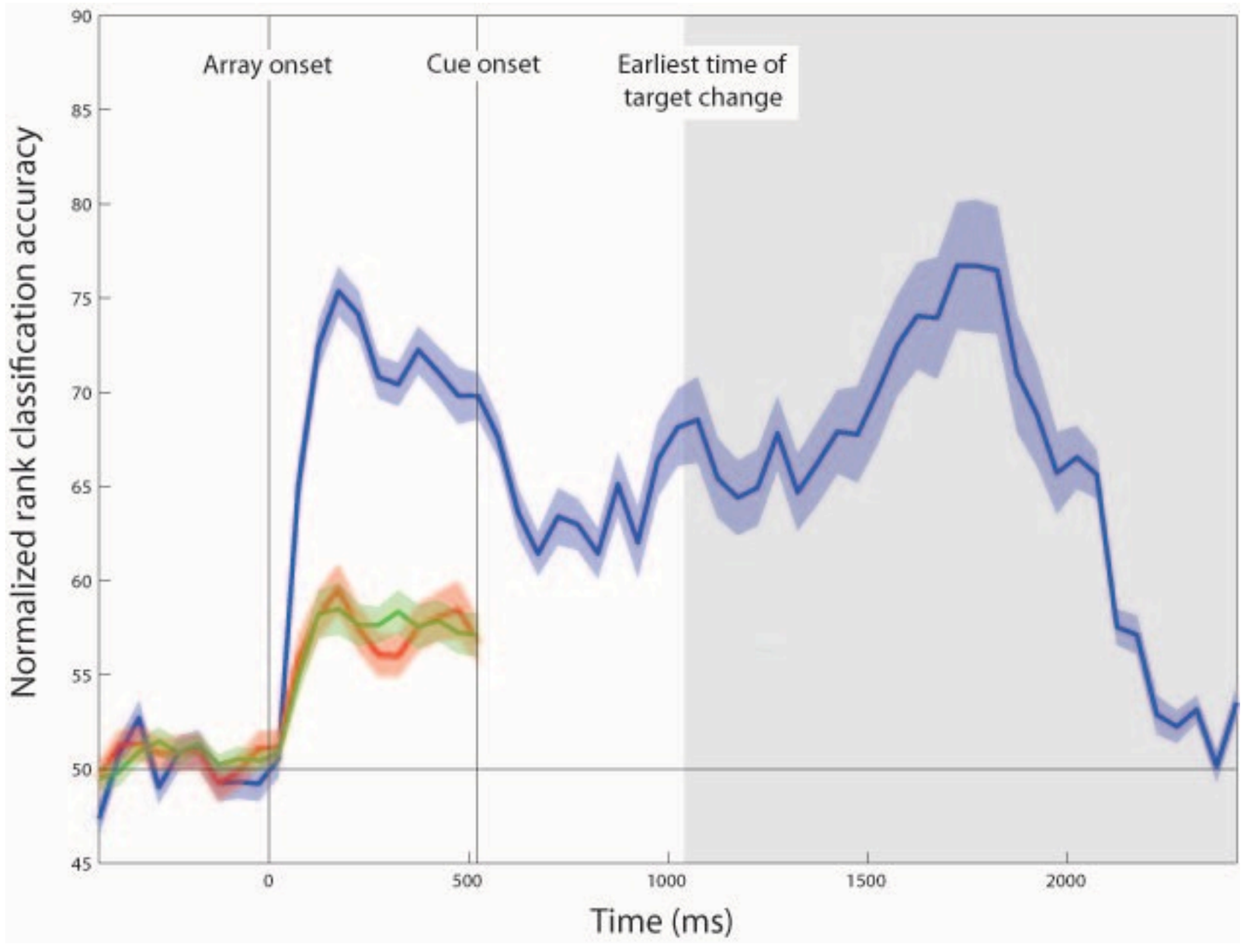




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

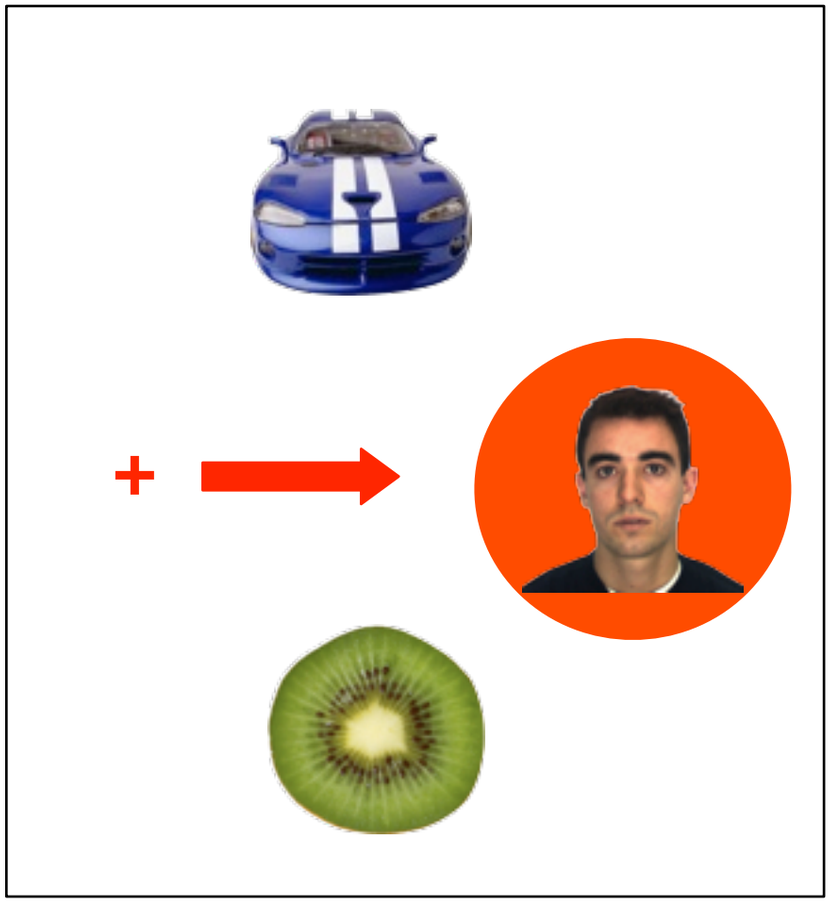
IT readout improves with attention

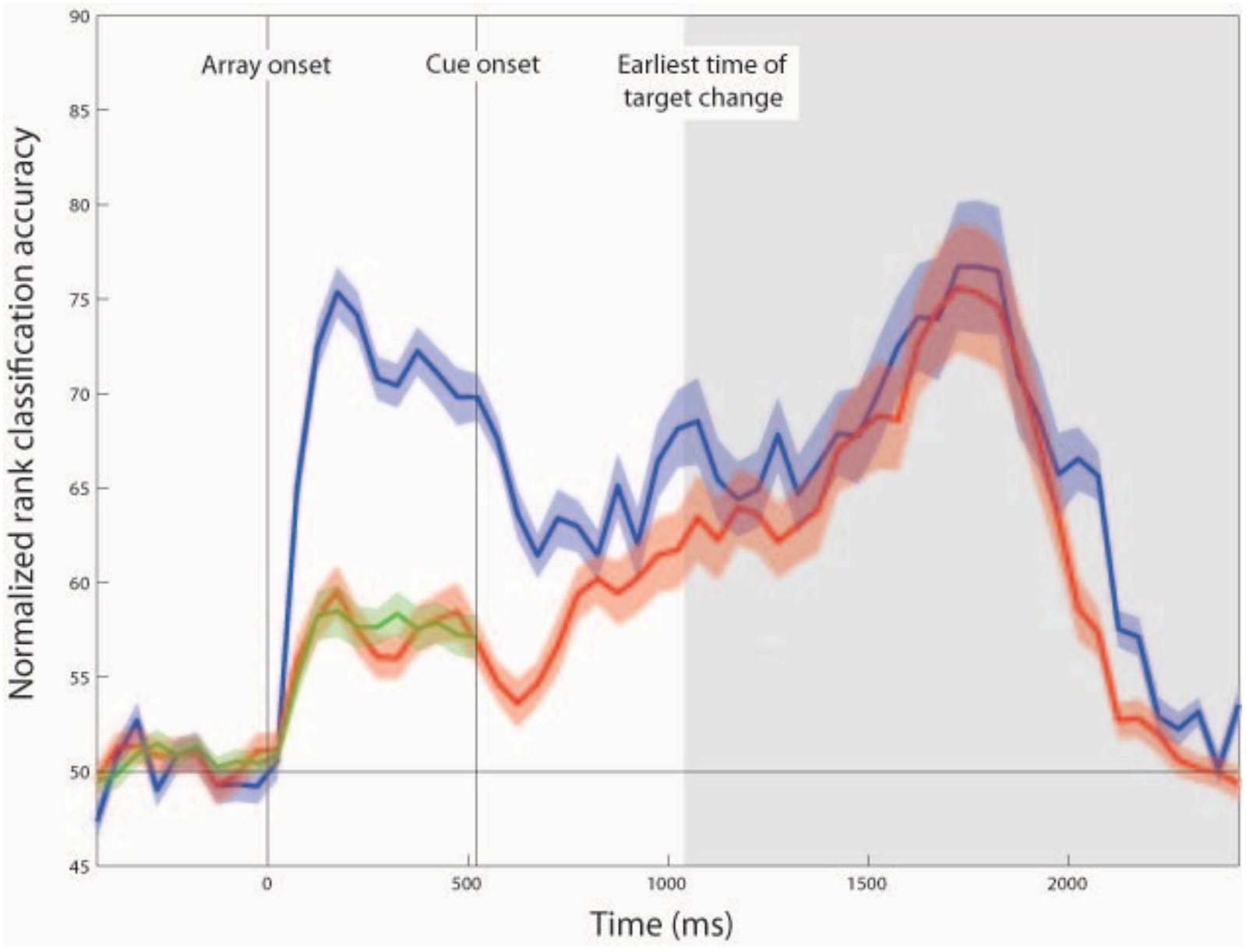




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

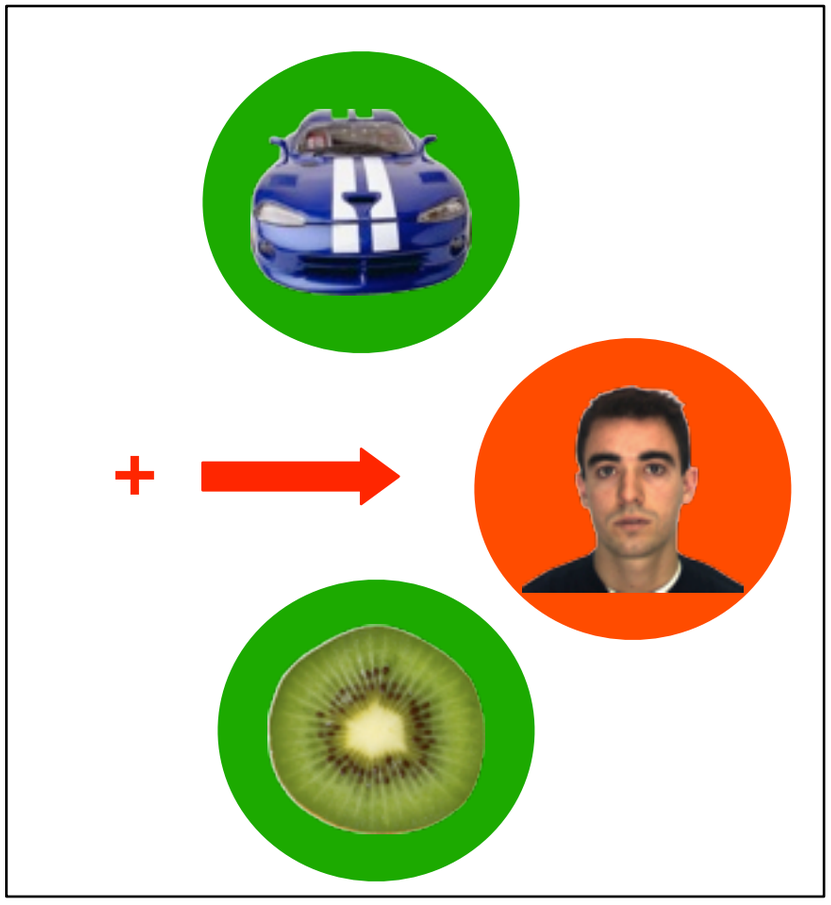
IT readout improves with attention

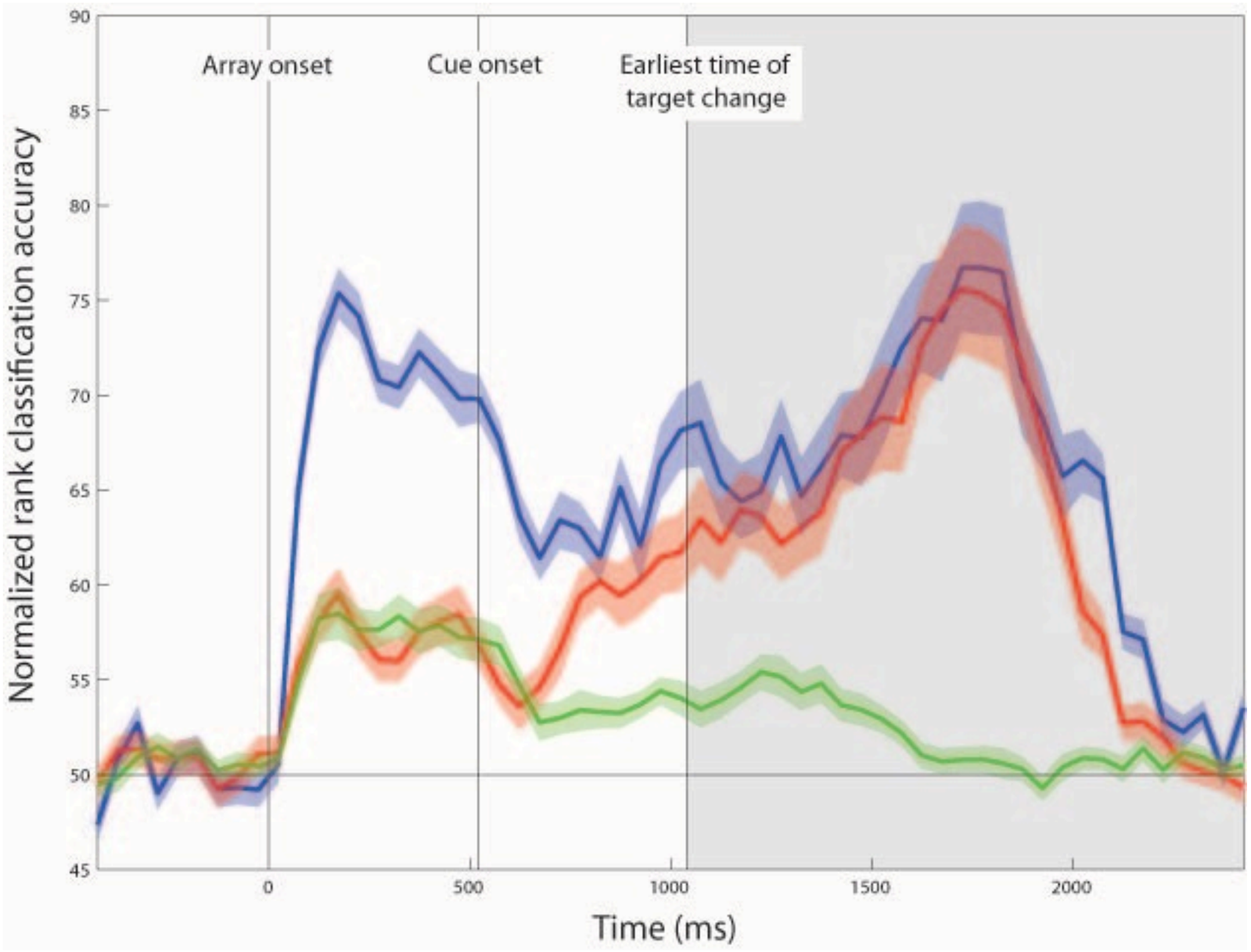




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

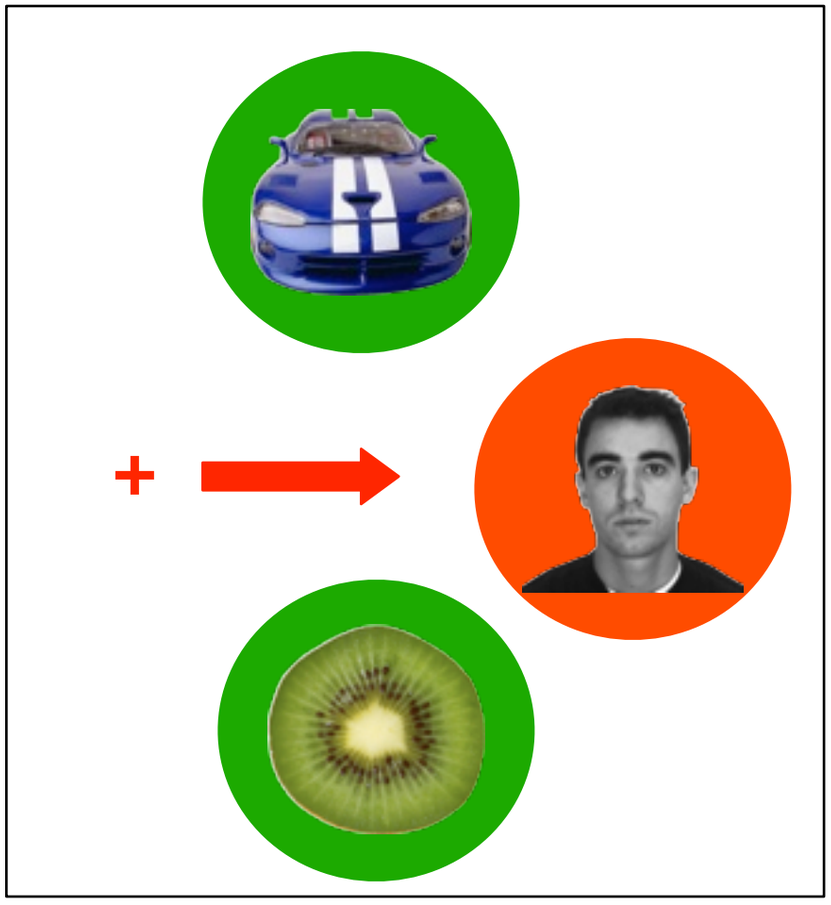
IT readout improves with attention

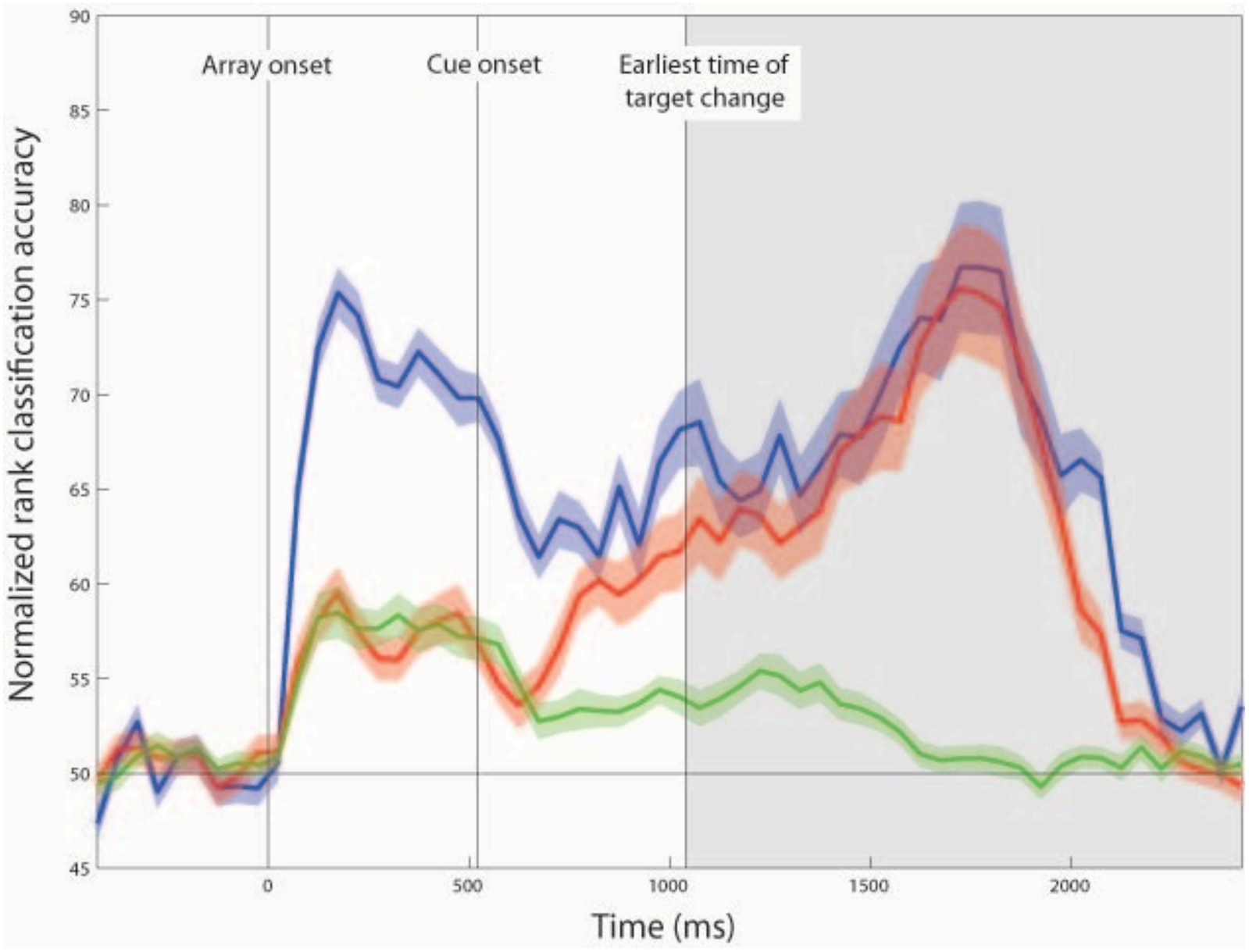




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

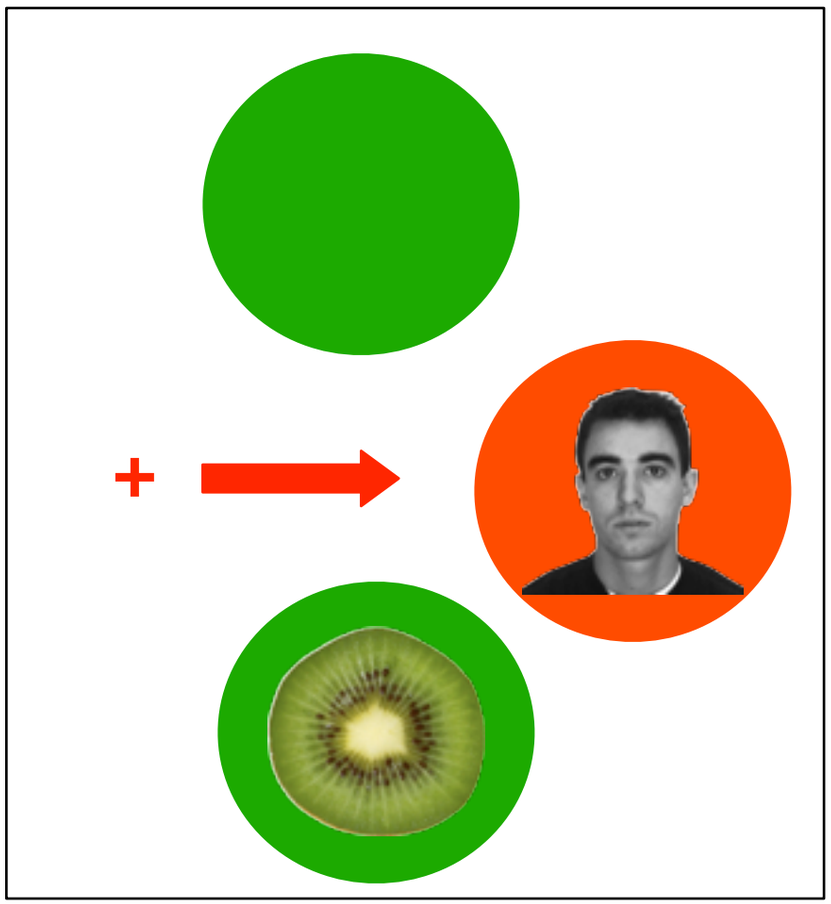
IT readout improves with attention

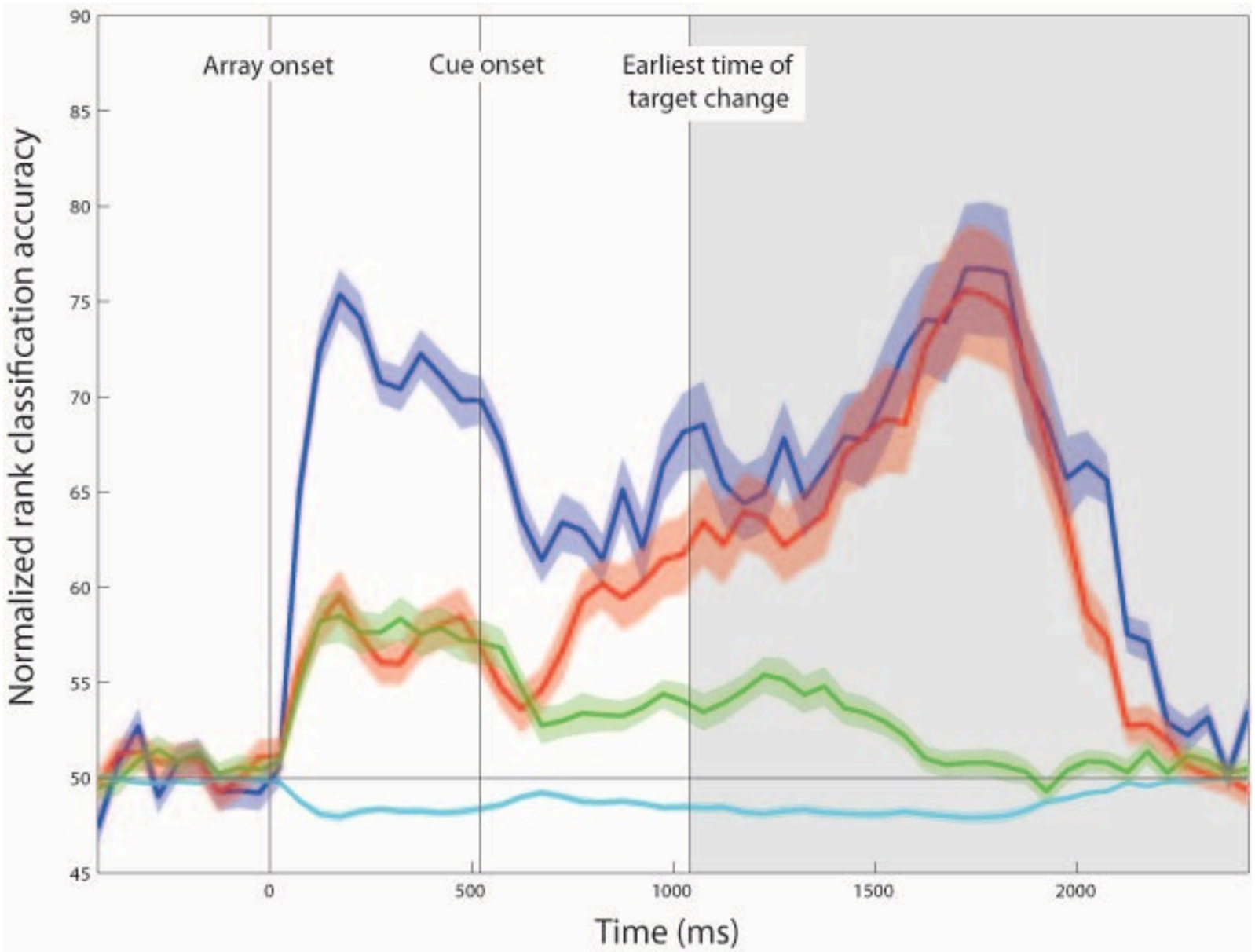




Zhang Meyers Bichot Serre Poggio Desimone unpublished data

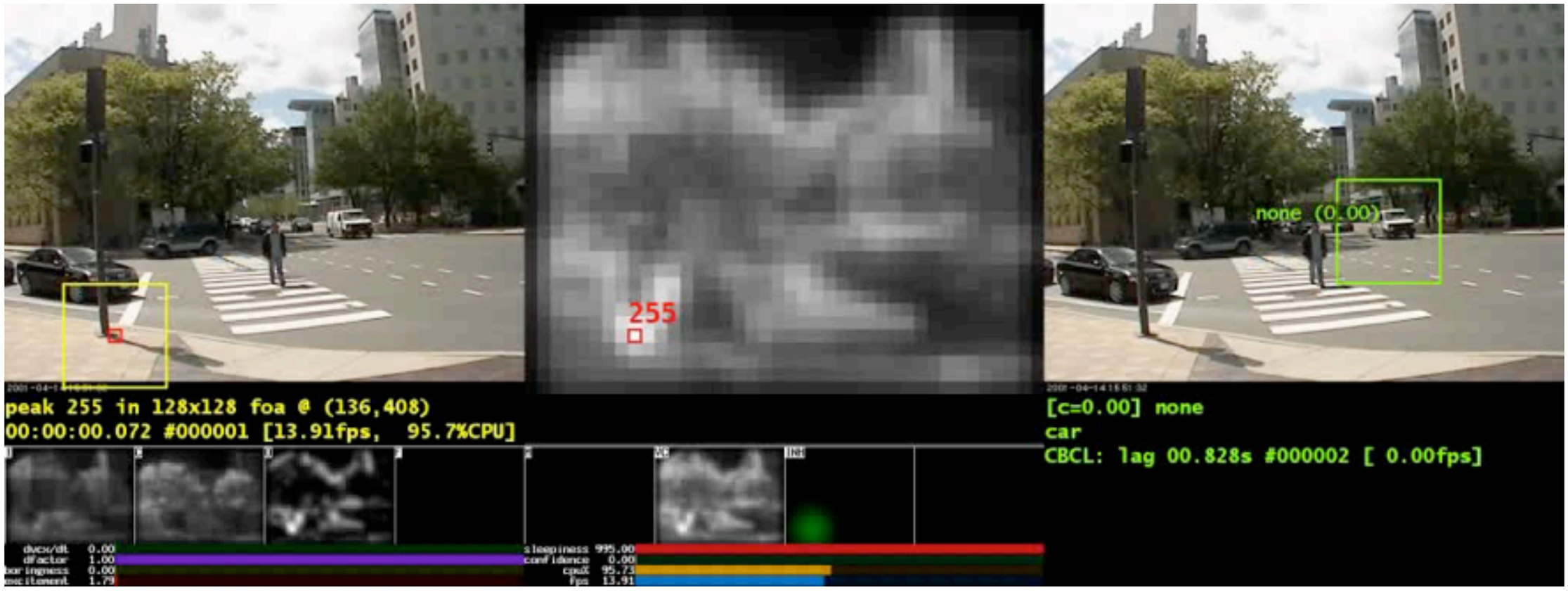
IT readout improves with attention





Zhang Meyers Bichot Serre Poggio Desimone unpublished data

- Robert Desimone (MIT)
- Christof Koch (CalTech)
- Laurent Itti (USC)
- Tomaso Poggio (MIT)
- David Sheinberg (Brown)



Thank you!