Deep learning in the visual cortex

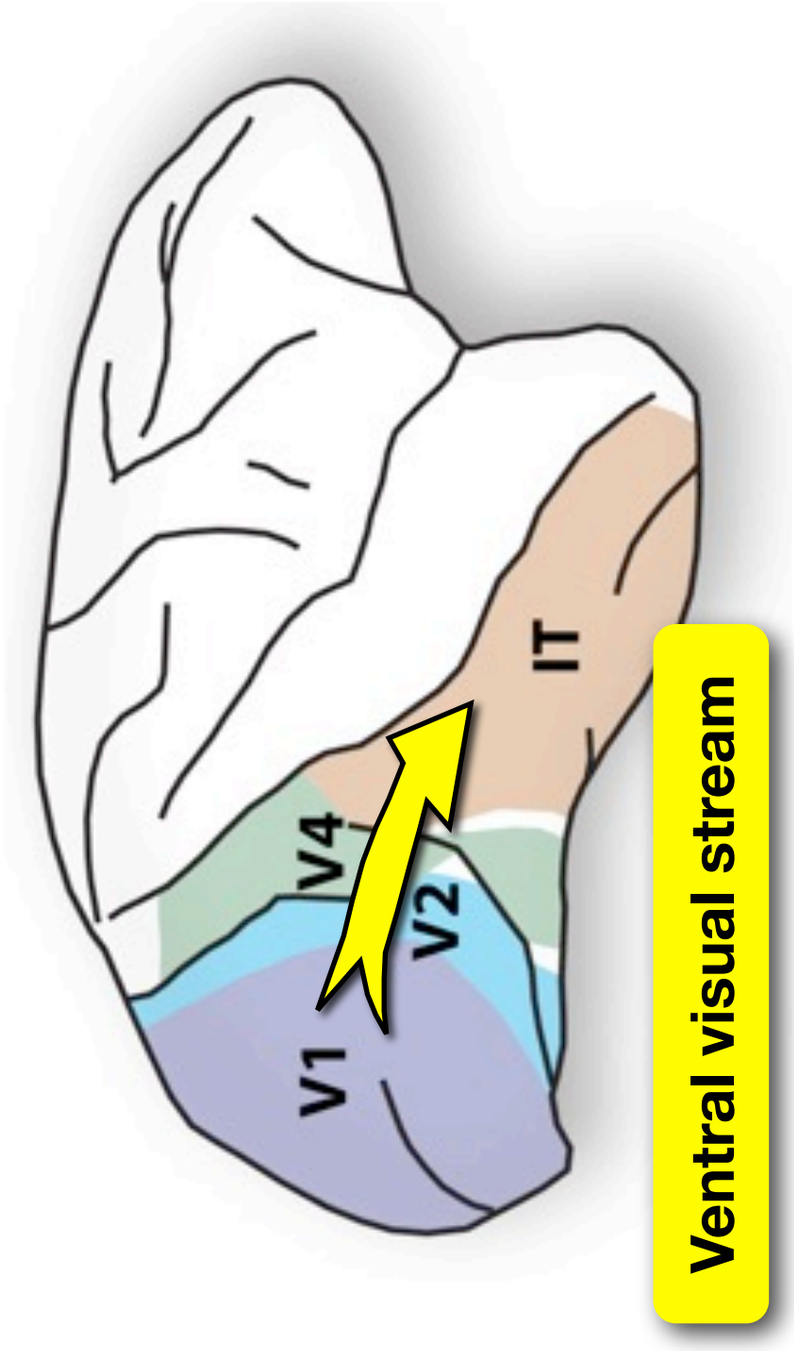
Thomas Serre

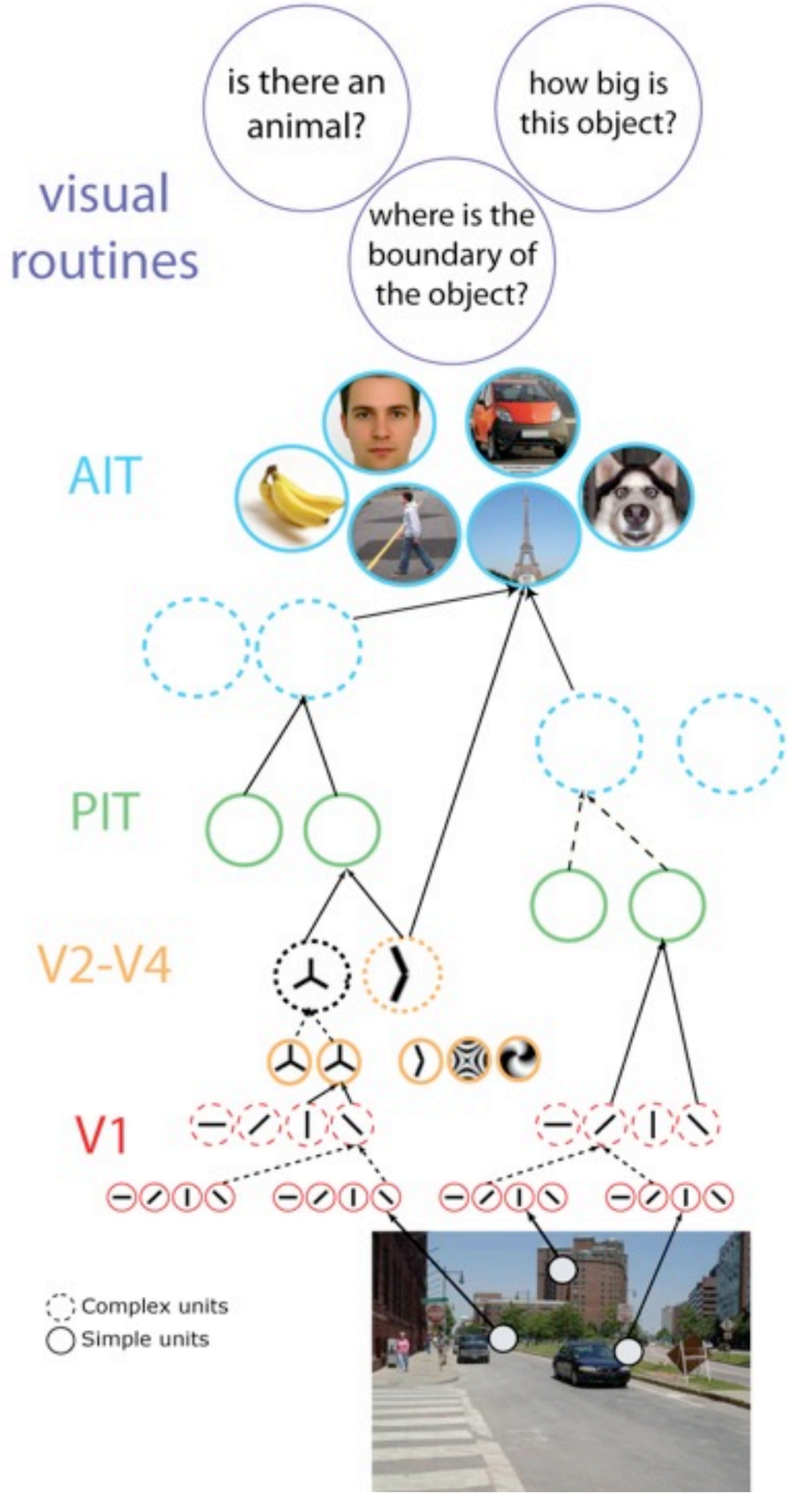
Brown University

I. Fundamentals of primate vision

II. Computational mechanisms of rapid recognition and feedforward processing

III. Beyond feedforward processing: Attentional mechanisms and cortical feedback



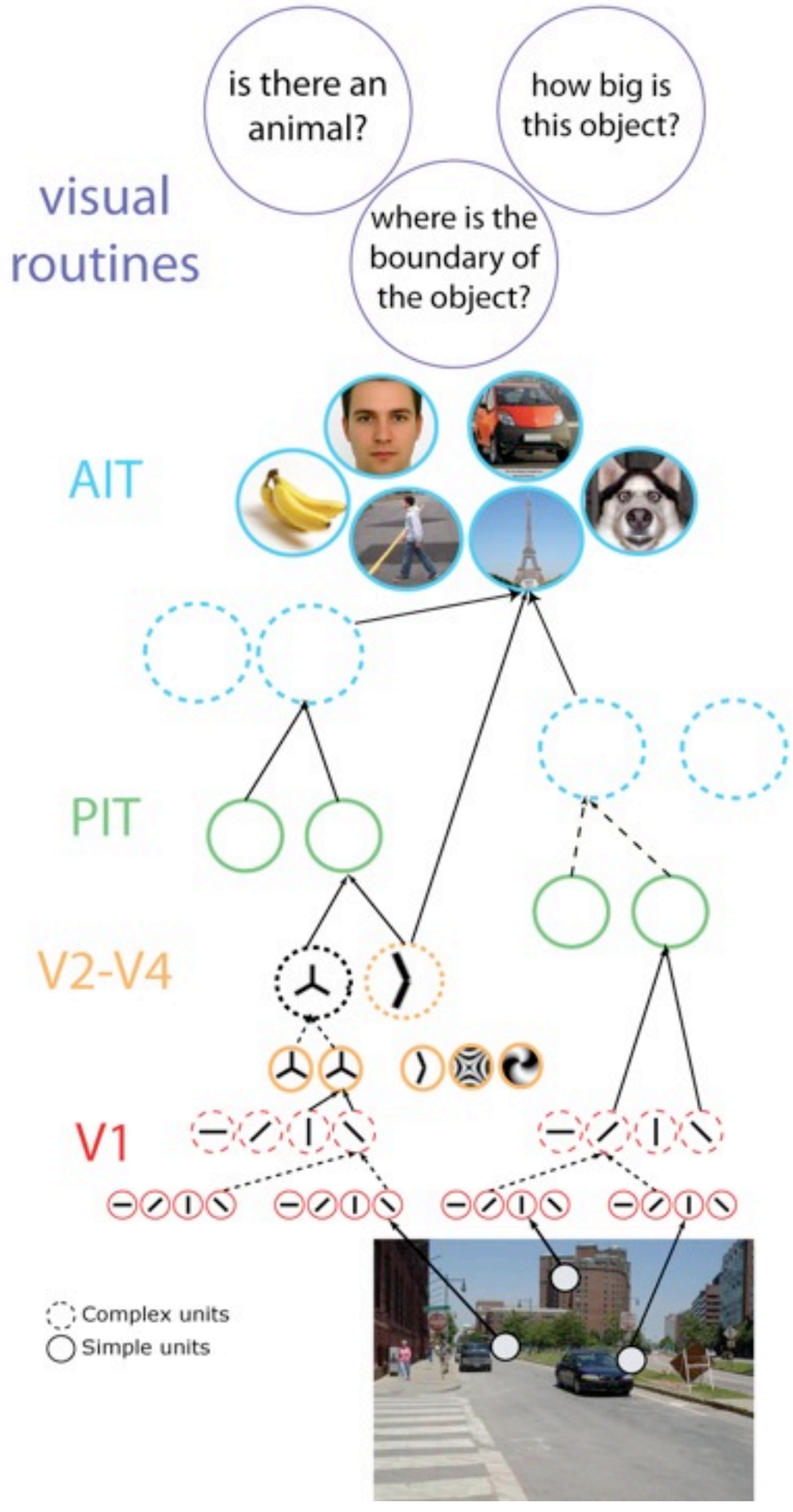


Feedforward processing

- Coarse initial base representation

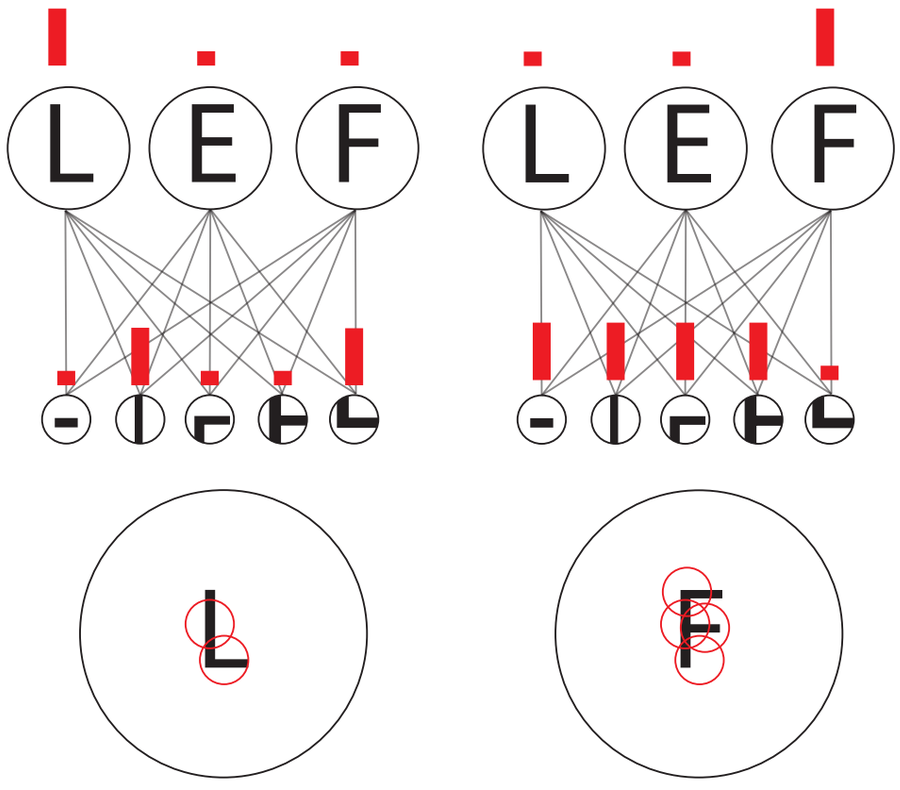
- Enables rapid object detection/recognition (‘what is there?’)
- Insufficient for object localization
- Sensitive to presence of clutter

Feedforward processing



The ‘crowding’ problem

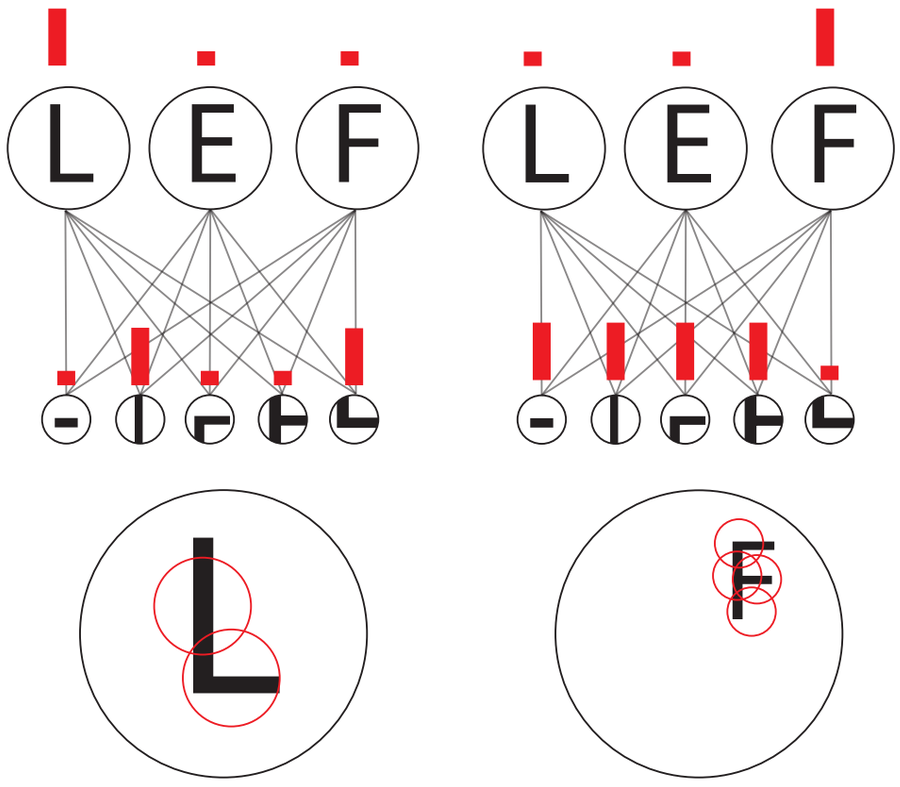
isolated objects



see also Dayan ’08; Pelli & Tillman ’08; and many others

The ‘crowding’ problem

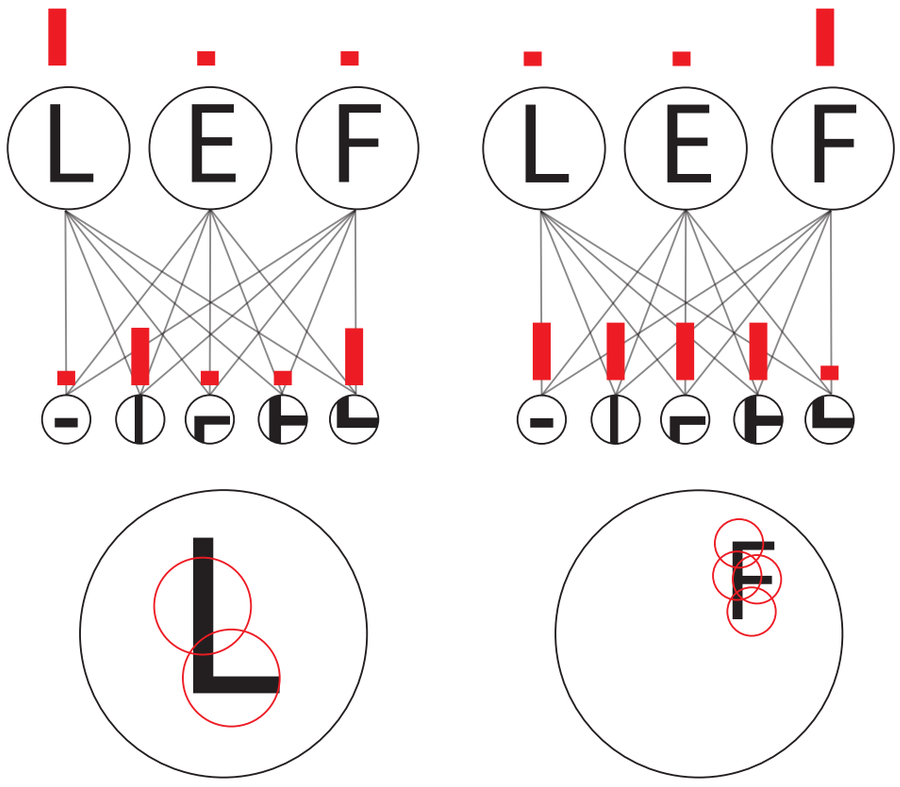
isolated objects



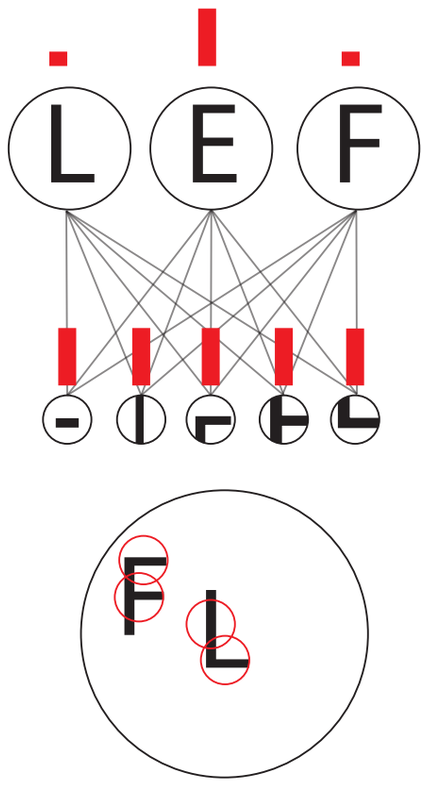
see also Dayan ’08; Pelli & Tillman ’08; and many others

The ‘crowding’ problem

isolated objects



clutter

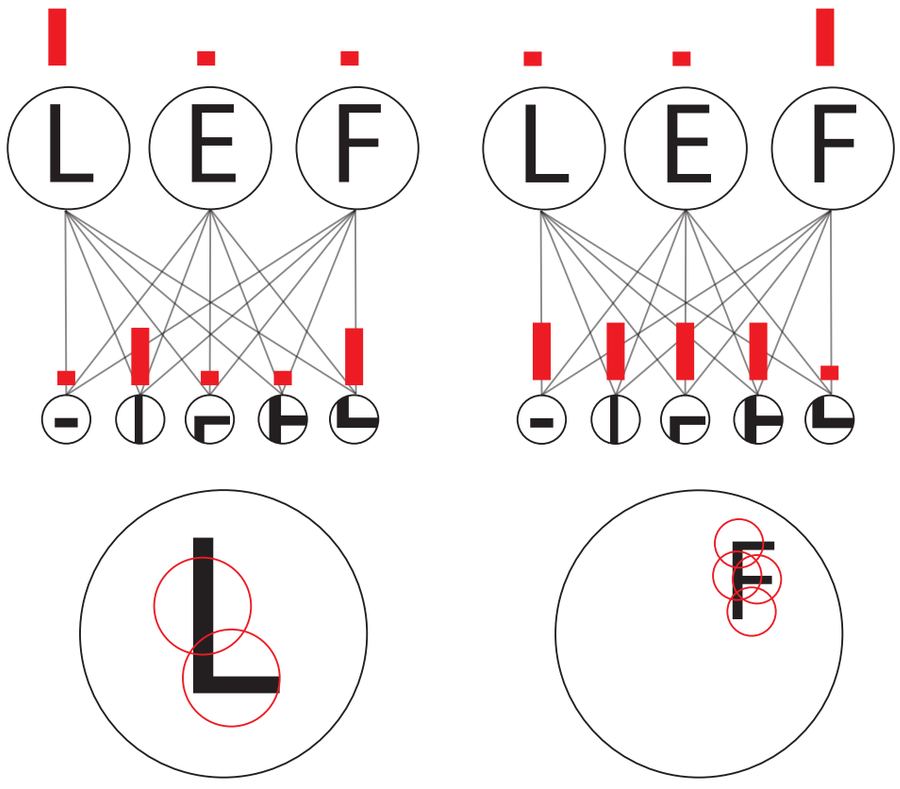


see also Dayan ’08; Pelli & Tillman ’08; and many others

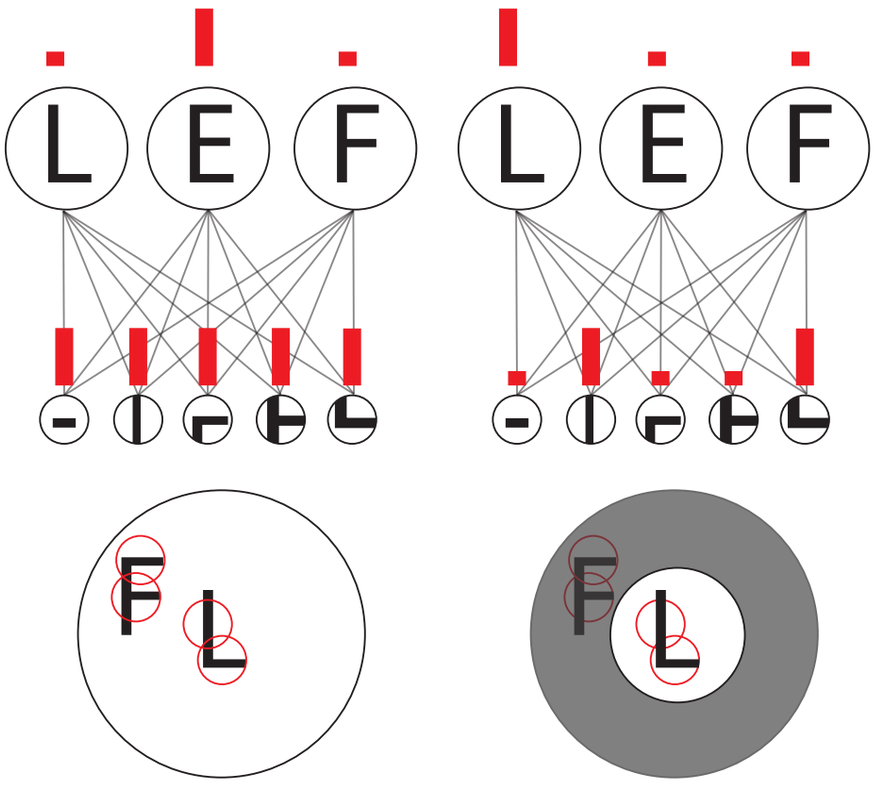
The ‘crowding’ problem

Prediction: recognition in clutter requires attentional mechanisms via cortical feedback

isolated objects



clutter



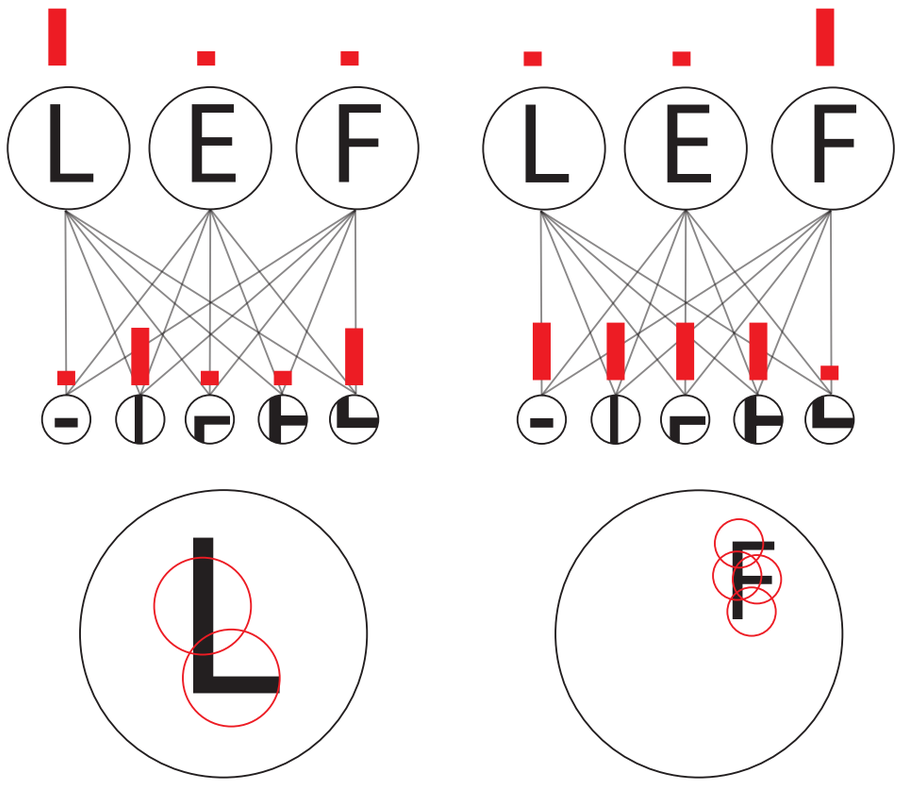
spatial attention

see also Dayan ’08; Pelli & Tillman ’08; and many others

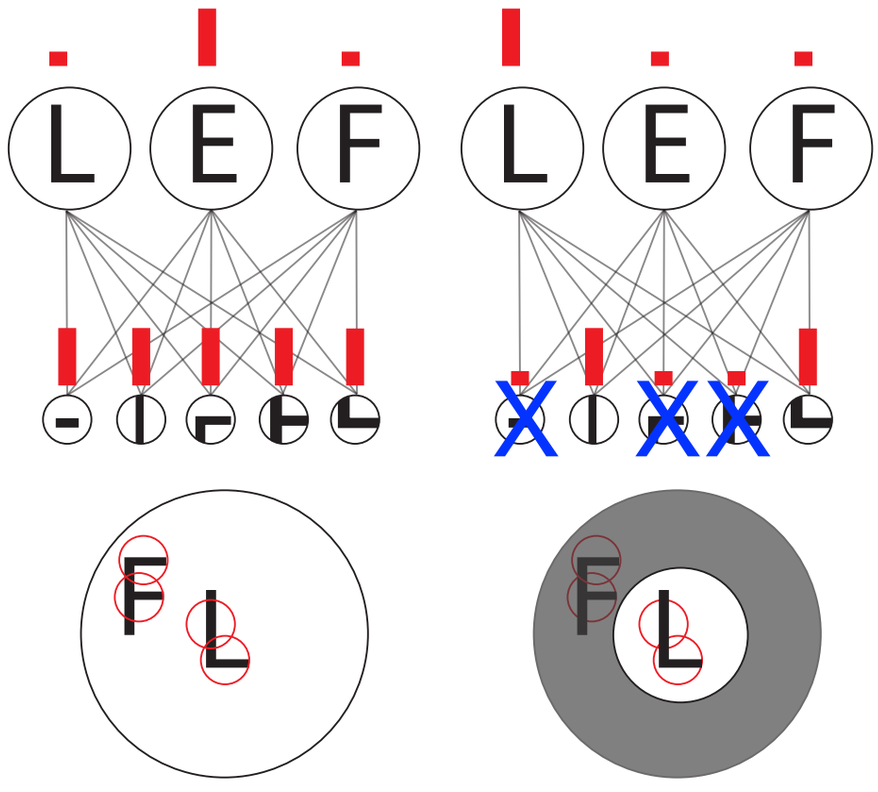
The ‘crowding’ problem

Prediction: recognition in clutter requires attentional mechanisms via cortical feedback

isolated objects



clutter



spatial attention

see also Dayan ’08; Pelli & Tillman ’08; and many others

1. Monkey electrophysiology

Read-out of inferior temporal cortex population activity: Spatial attention eliminates clutter

with Zhang, Meyers, Bichot, Poggio & Desimone

2. Computational model of integrated attention and recognition

What and where: a Bayesian attention theory of attention

with Chikkerur, Tan & Poggio

1. Monkey electrophysiology

Read-out of inferior temporal cortex population activity: Spatial attention eliminates clutter

with Zhang, Meyers, Bichot, Poggio & Desimone

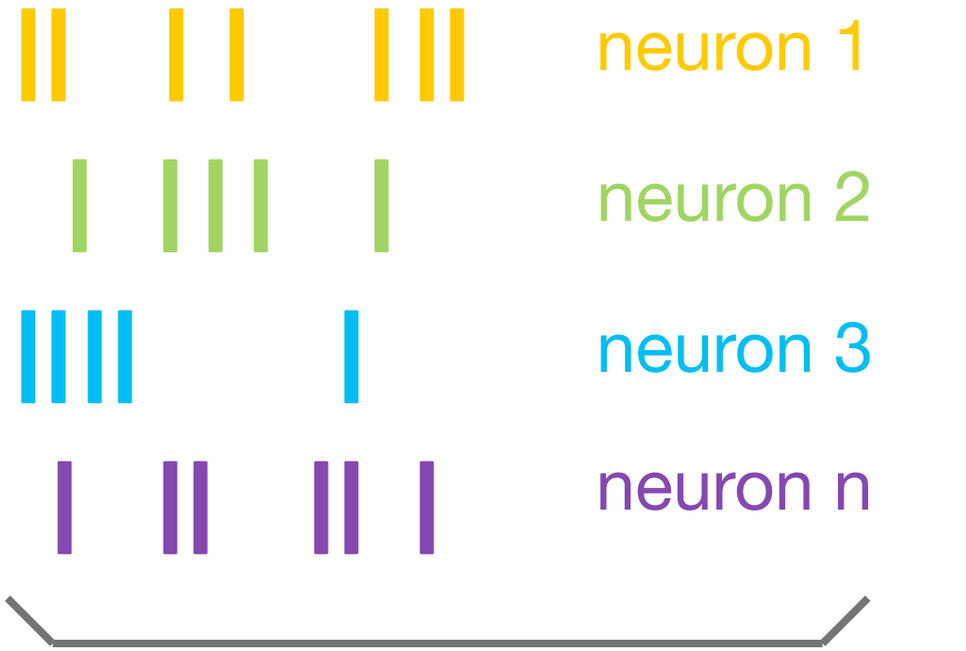
---

2. Computational model of integrated attention and recognition

What and where: a Bayesian attention theory of attention

with Chikkerur, Tan & Poggio

The ‘readout’ approach

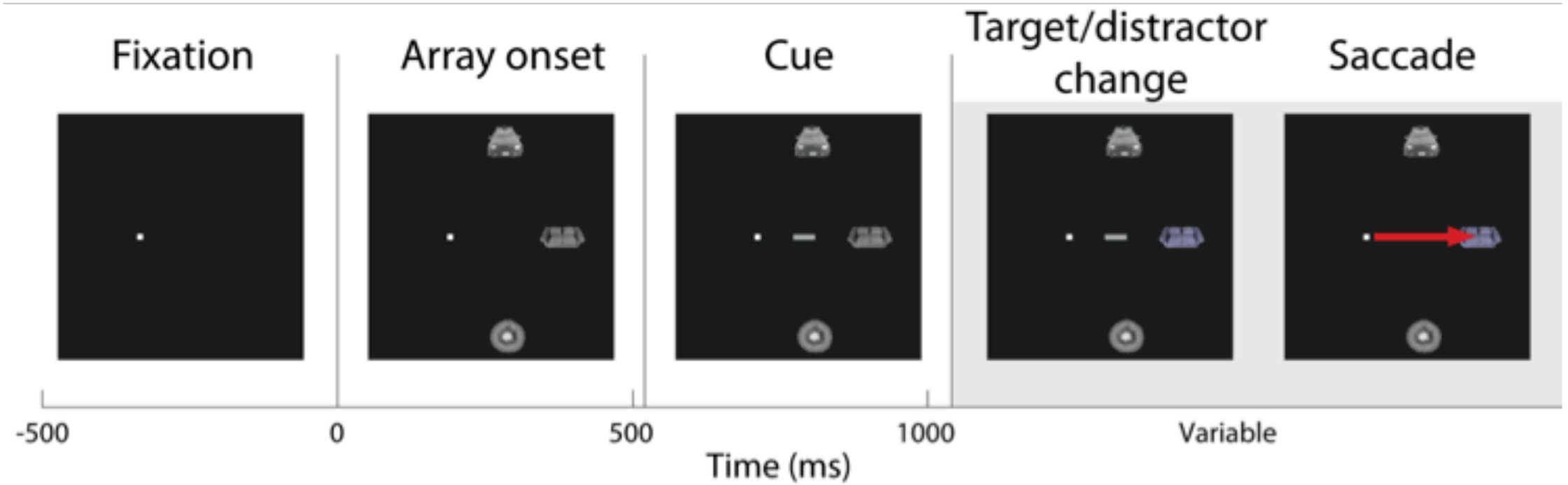


Pattern Classifier

Prediction

Zhang Meyers Bichot Serre Poggio Desimone PNAS’11

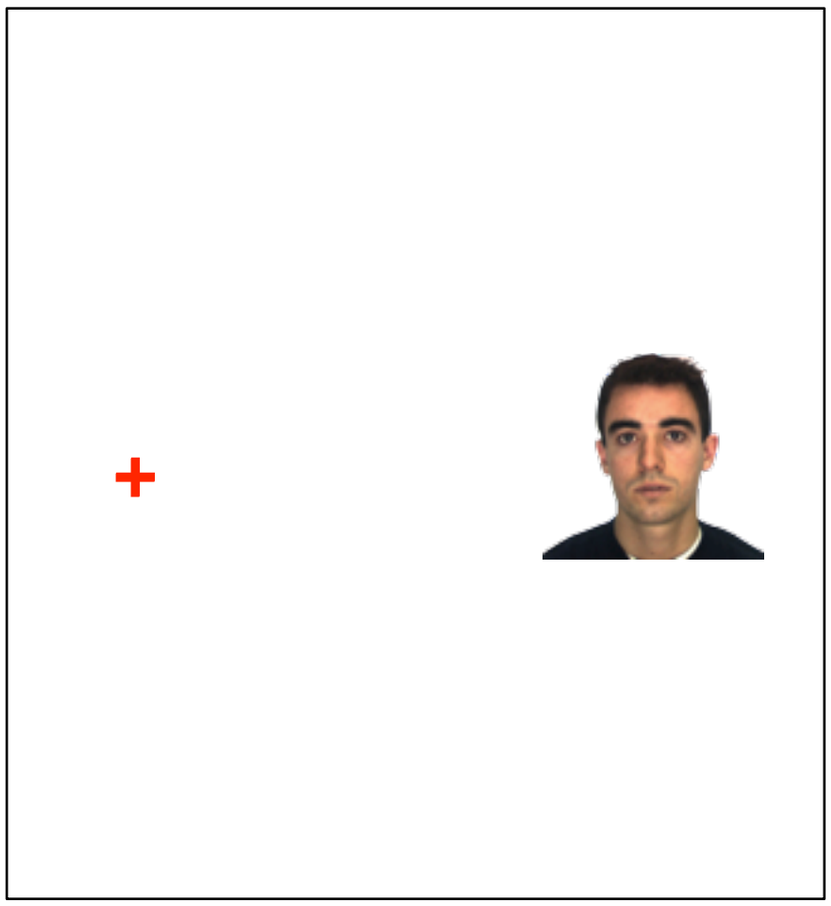
The experiment



Zhang Meyers Bichot Serre Poggio Desimone unpublished data

The experiment

train readout classifier
on isolated object



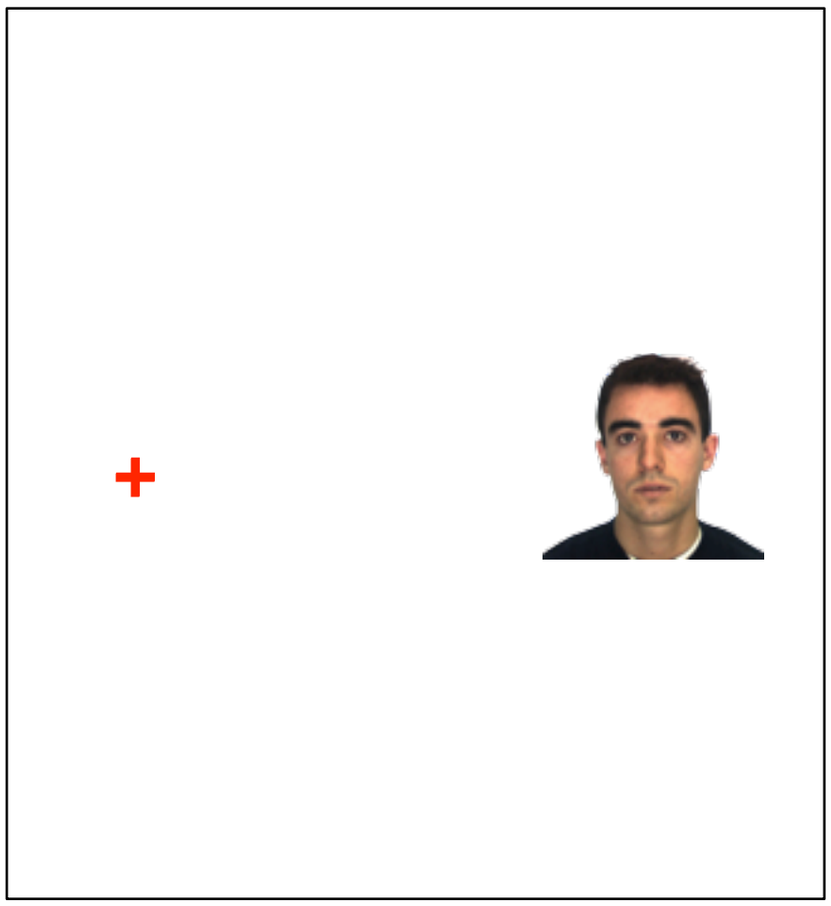
test generalization in
clutter

Zhang* Meyers* Bichot Serre Poggio Desimone in sub

The experiment

train readout classifier
on isolated object

test generalization in
clutter



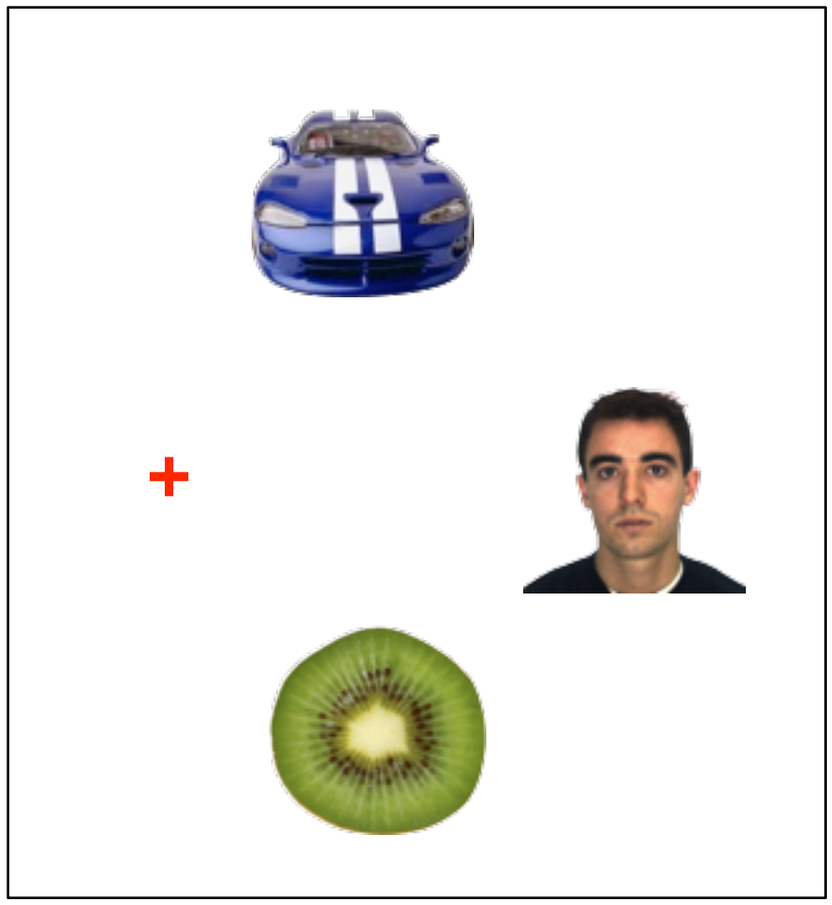
isolated

Zhang* Meyers* Bichot Serre Poggio Desimone in sub

The experiment

train readout classifier
on isolated object

test generalization in
clutter

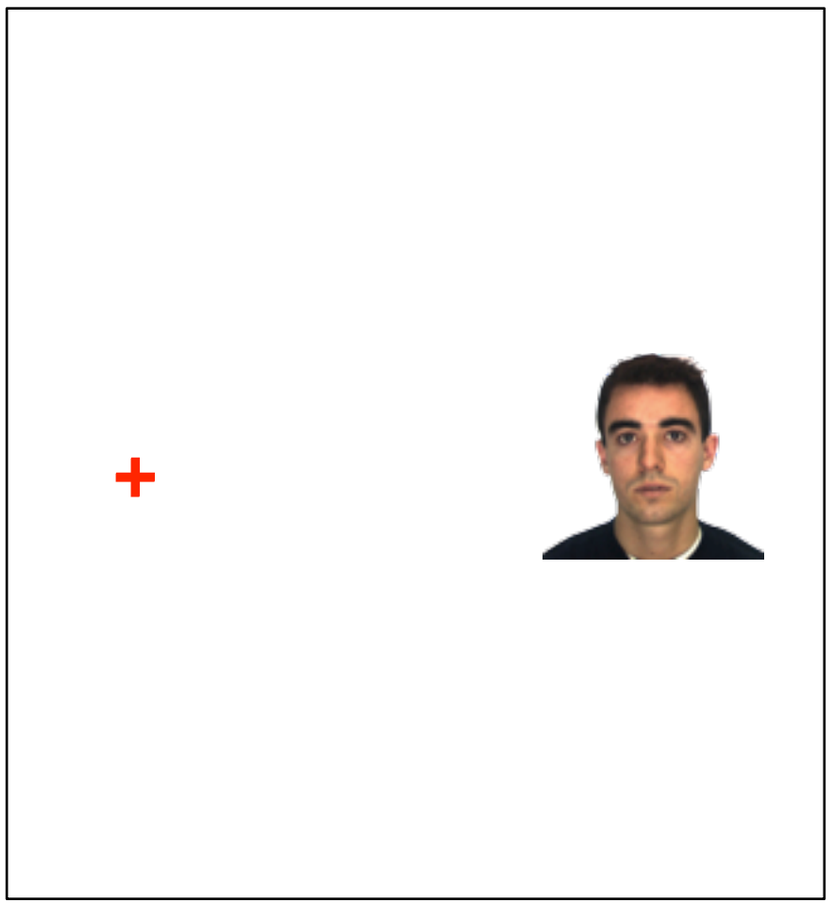


isolated

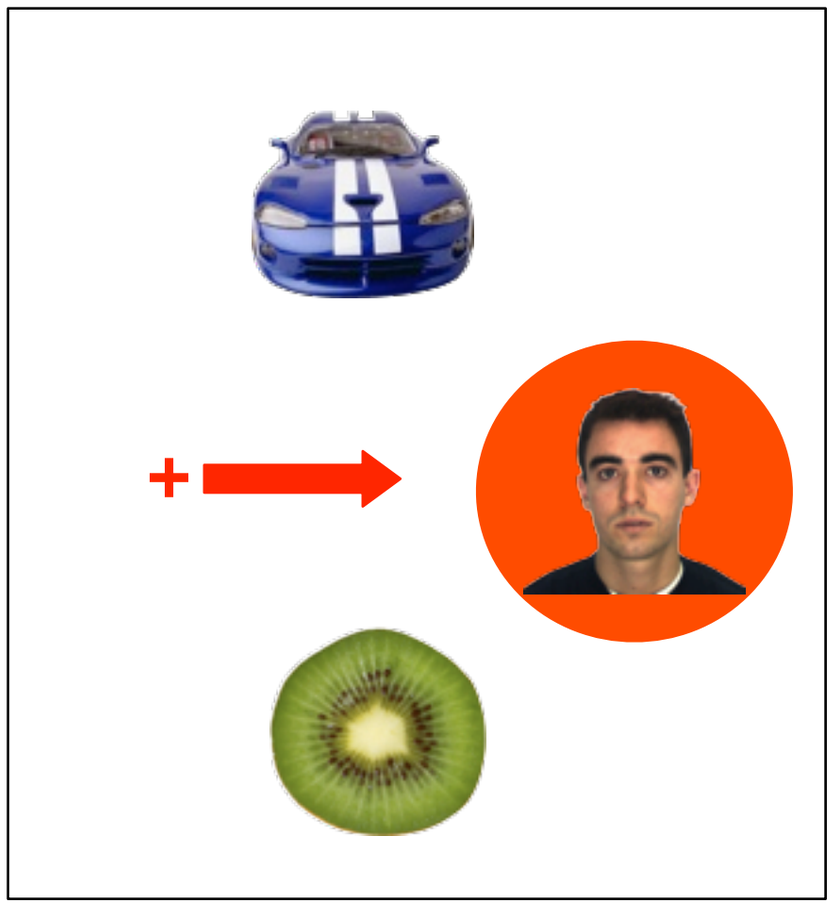
Zhang* Meyers* Bichot Serre Poggio Desimone in sub

The experiment

train readout classifier
on isolated object



test generalization in
clutter



isolated

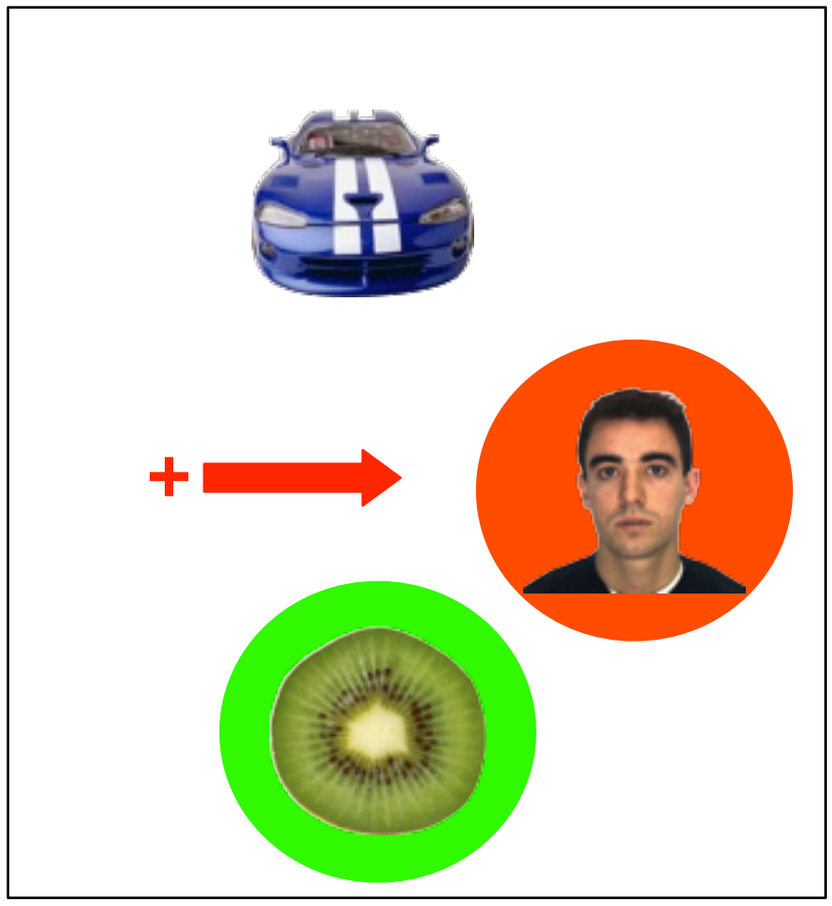
attended

Zhang* Meyers* Bichot Serre Poggio Desimone in sub

The experiment

train readout classifier
on isolated object

test generalization in
clutter



isolated

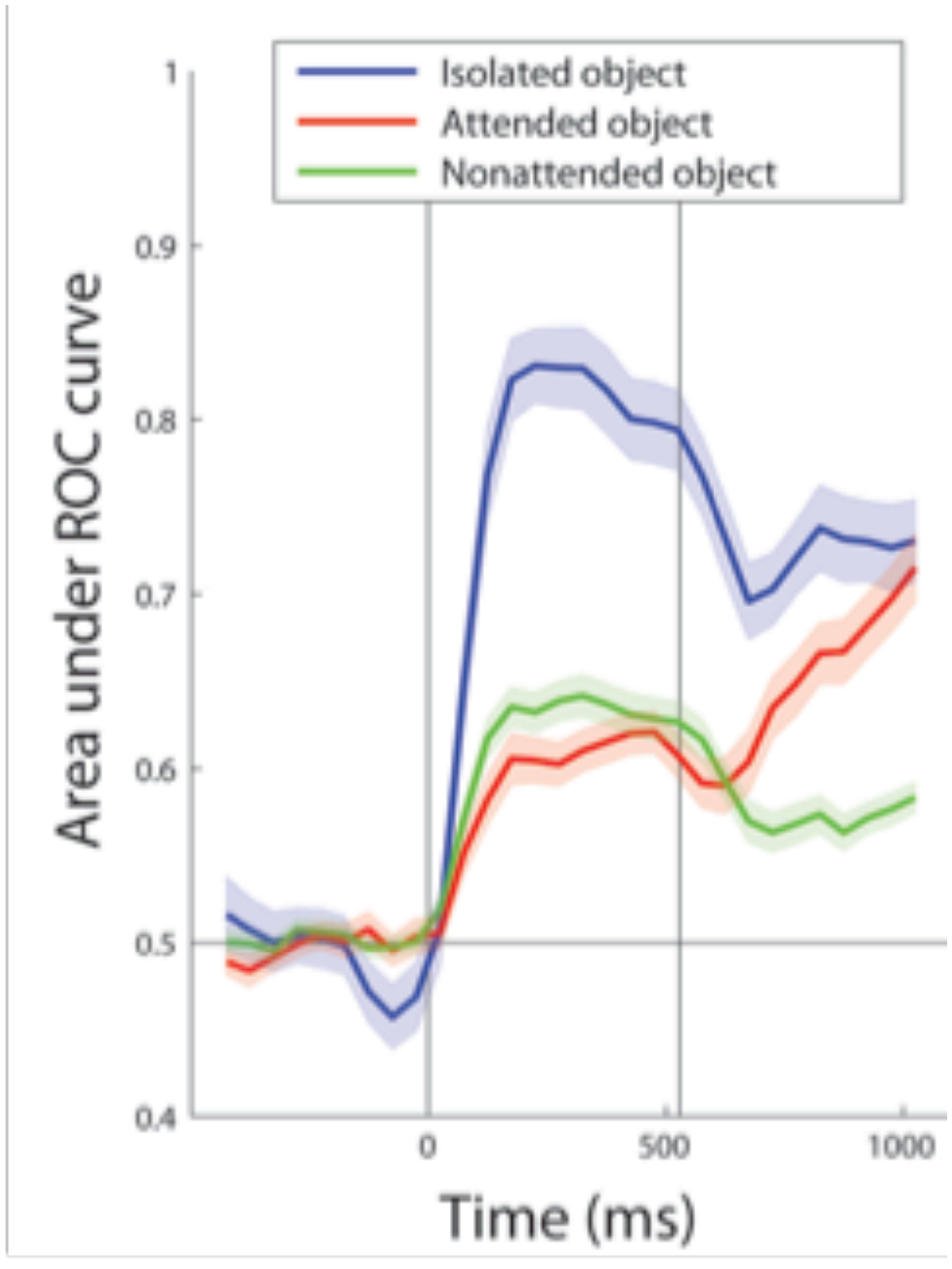
attended

Non-attended

Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

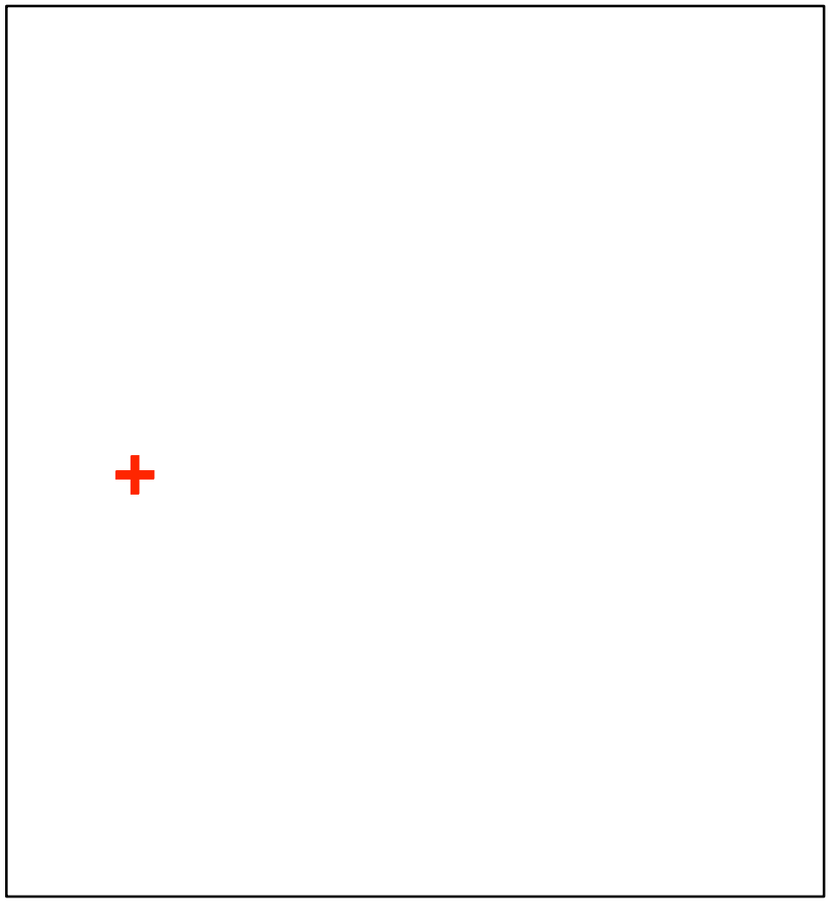
Decoding object identity

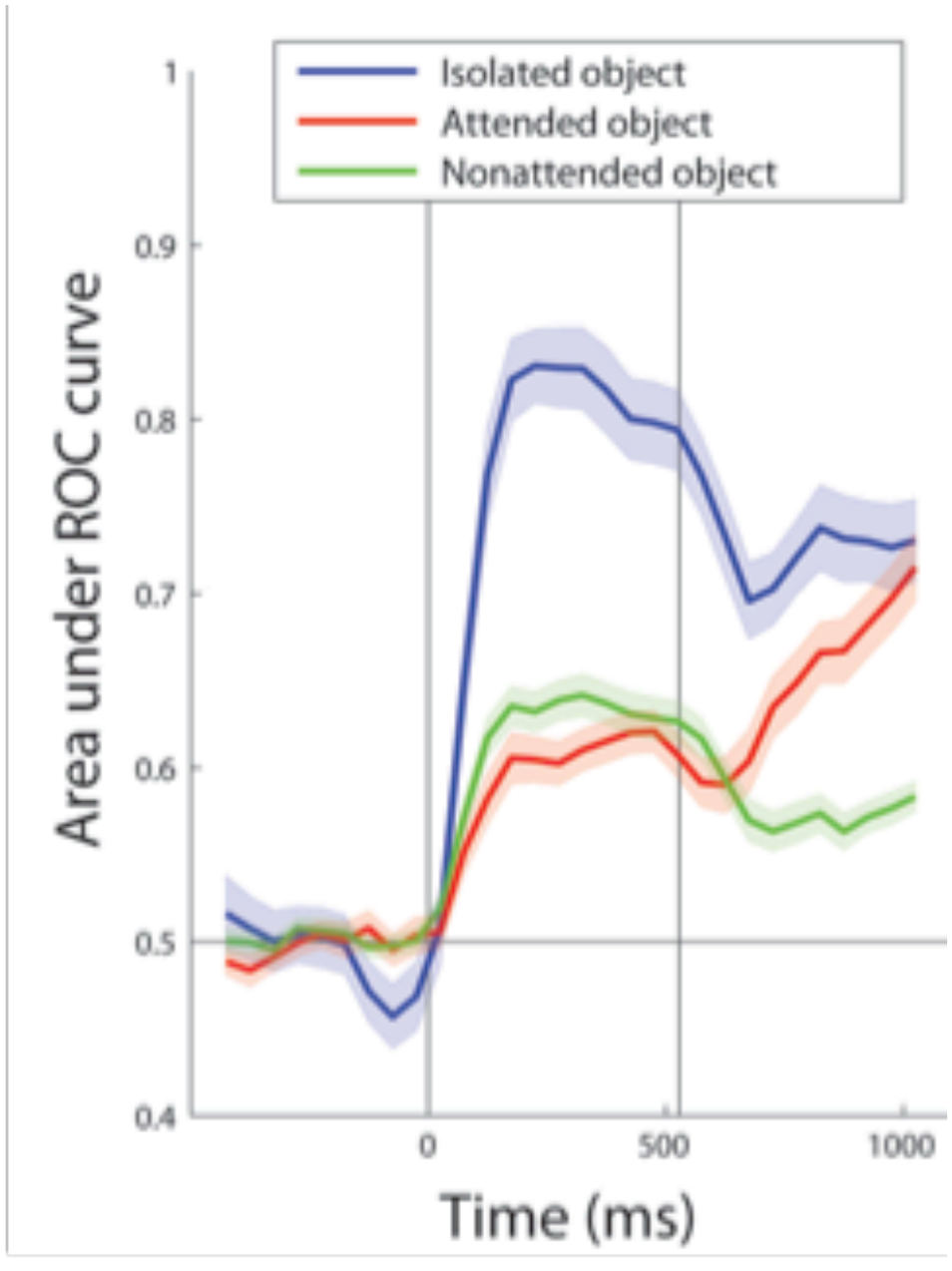


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding object identity

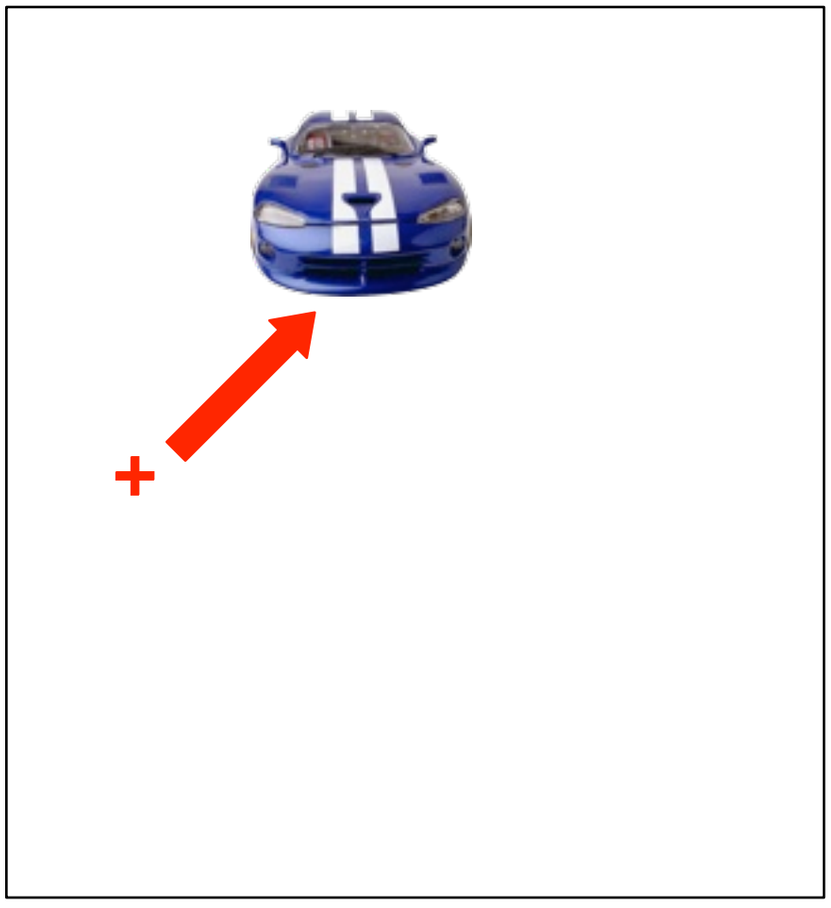


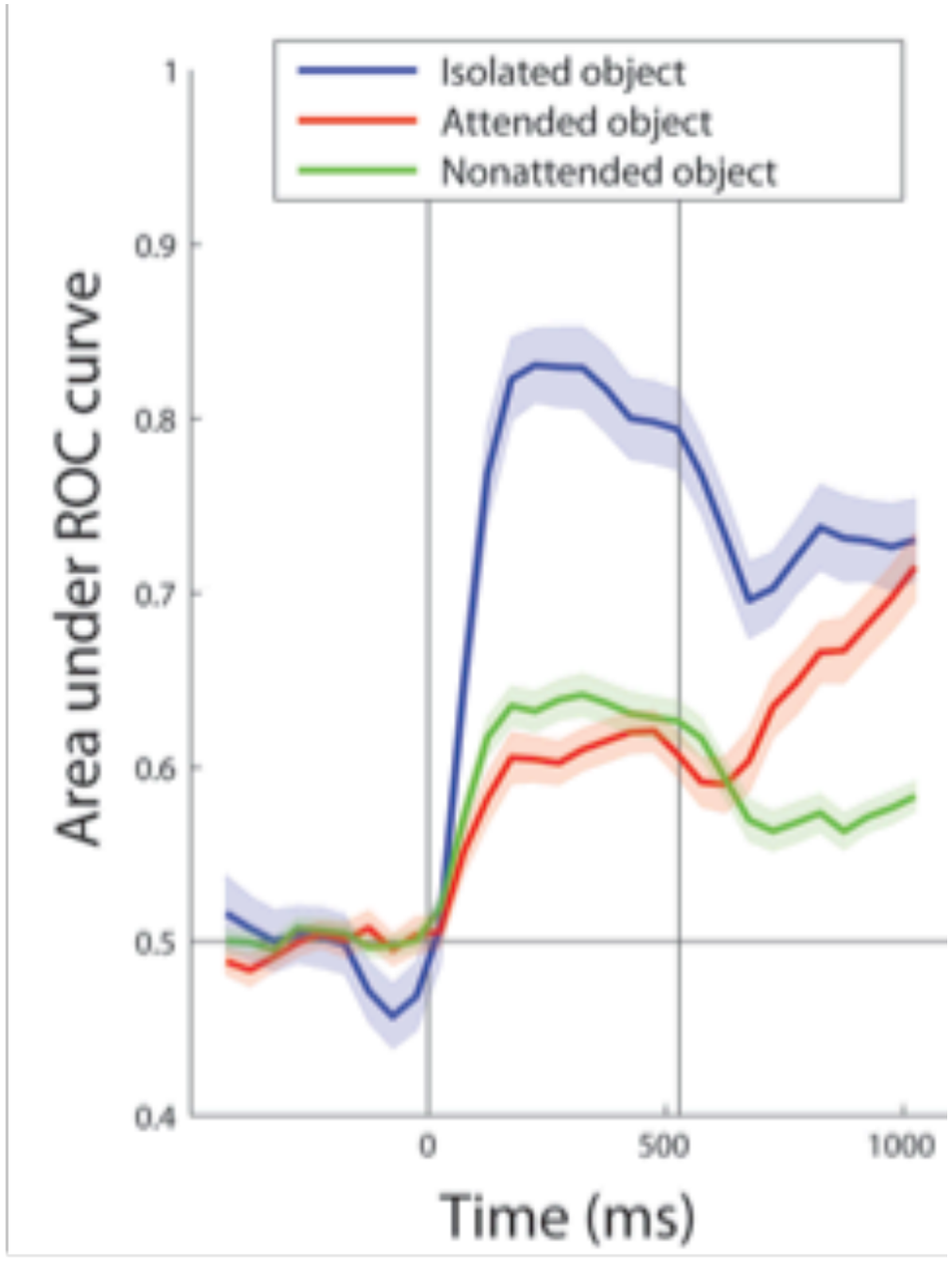


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding object identity

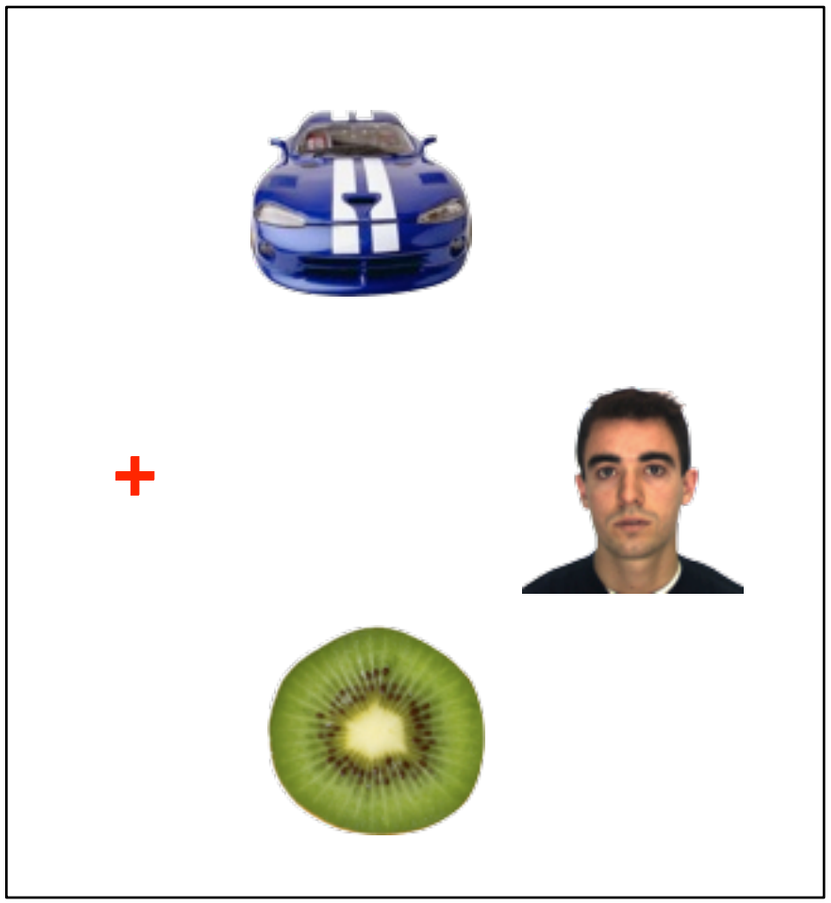


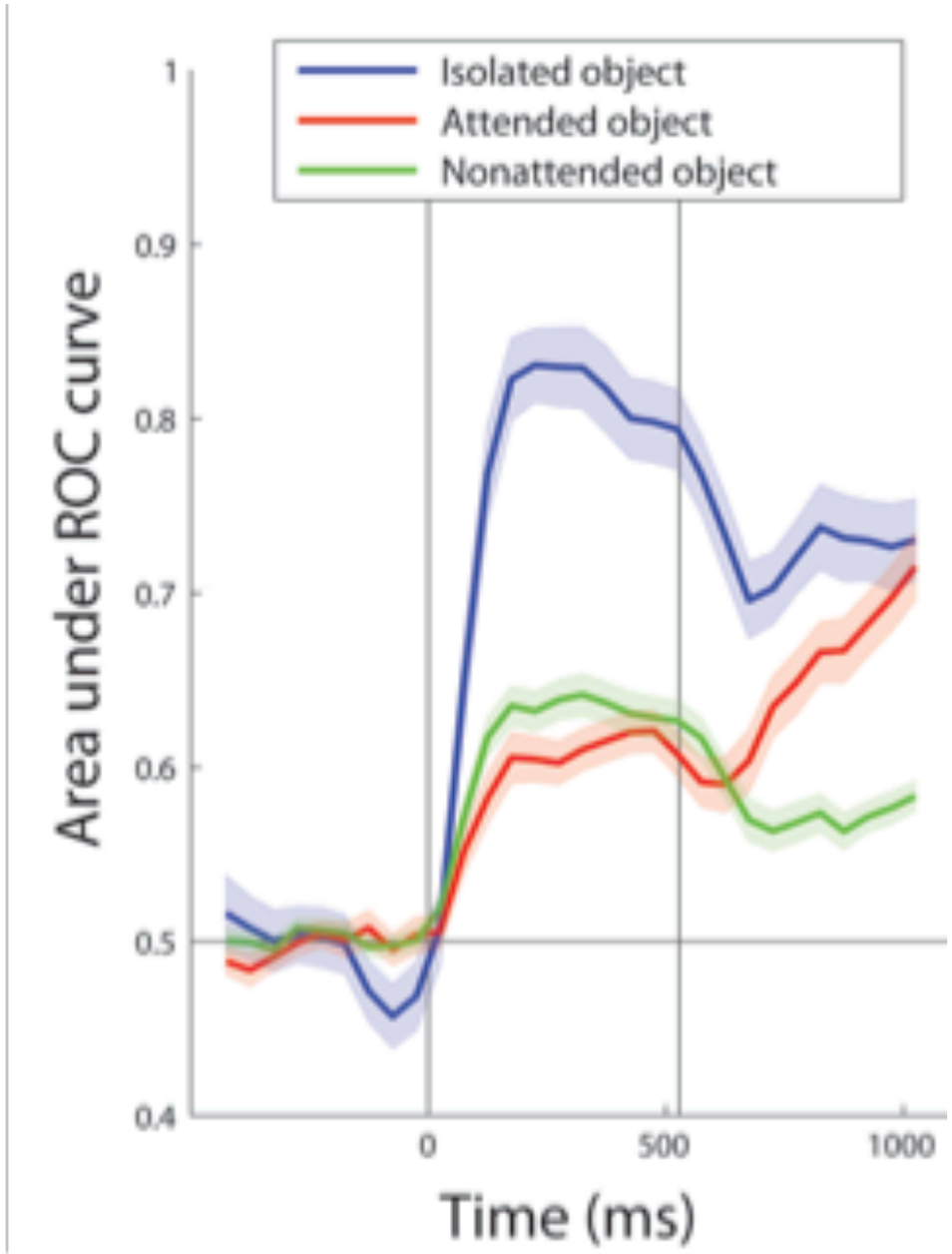


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding object identity

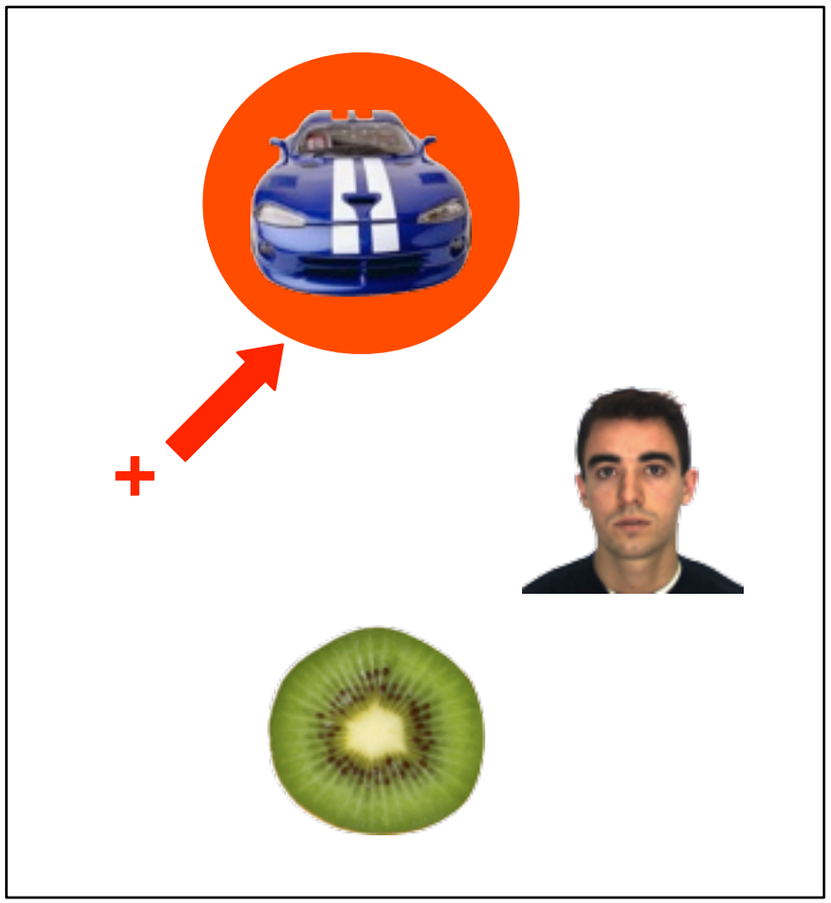


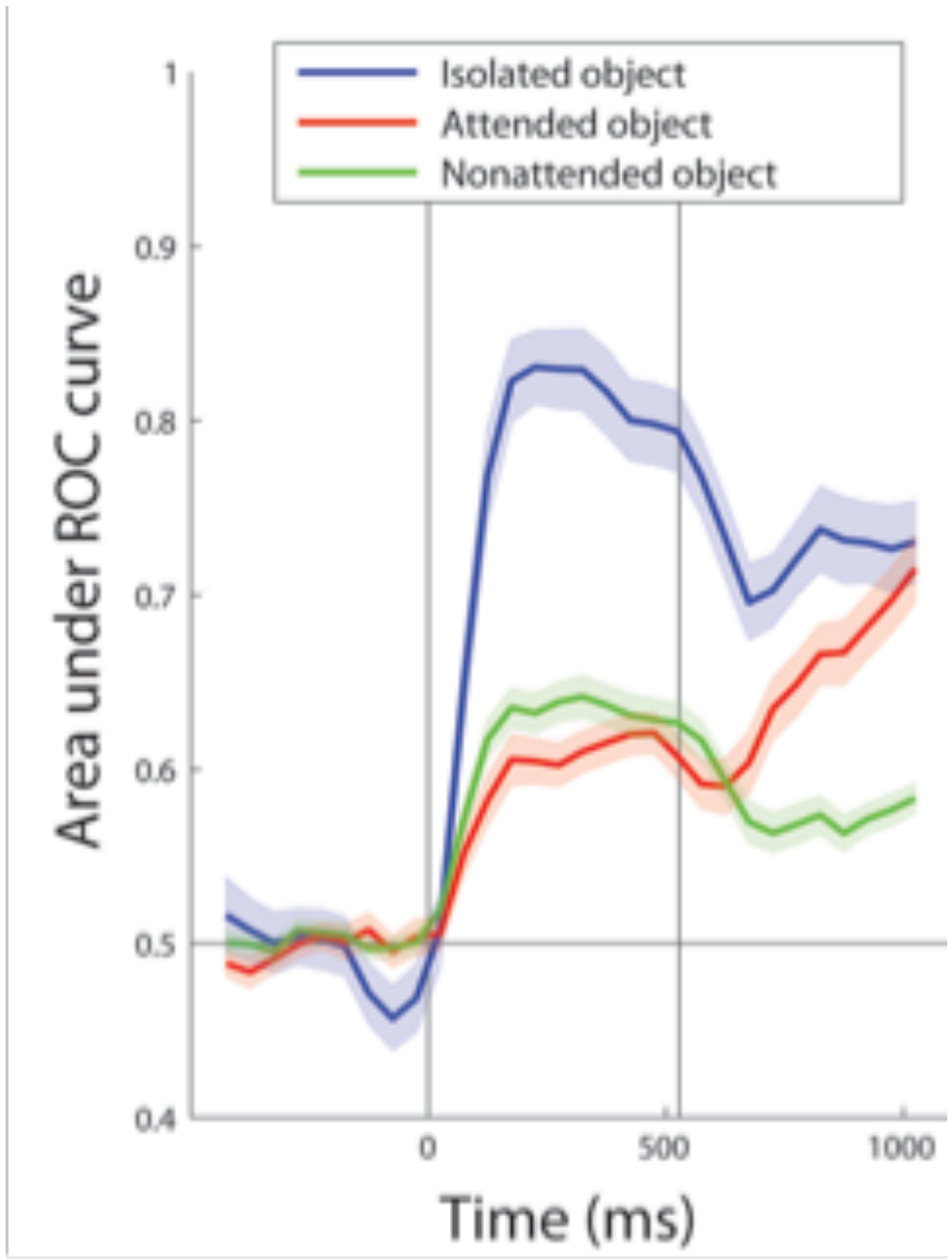


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding object identity

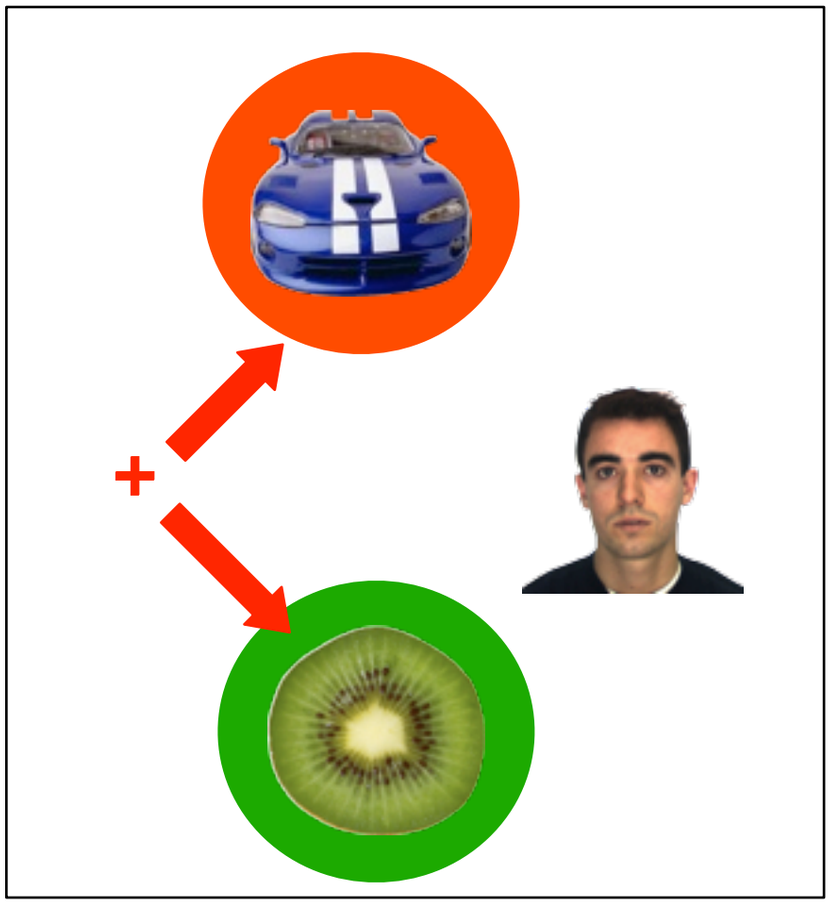


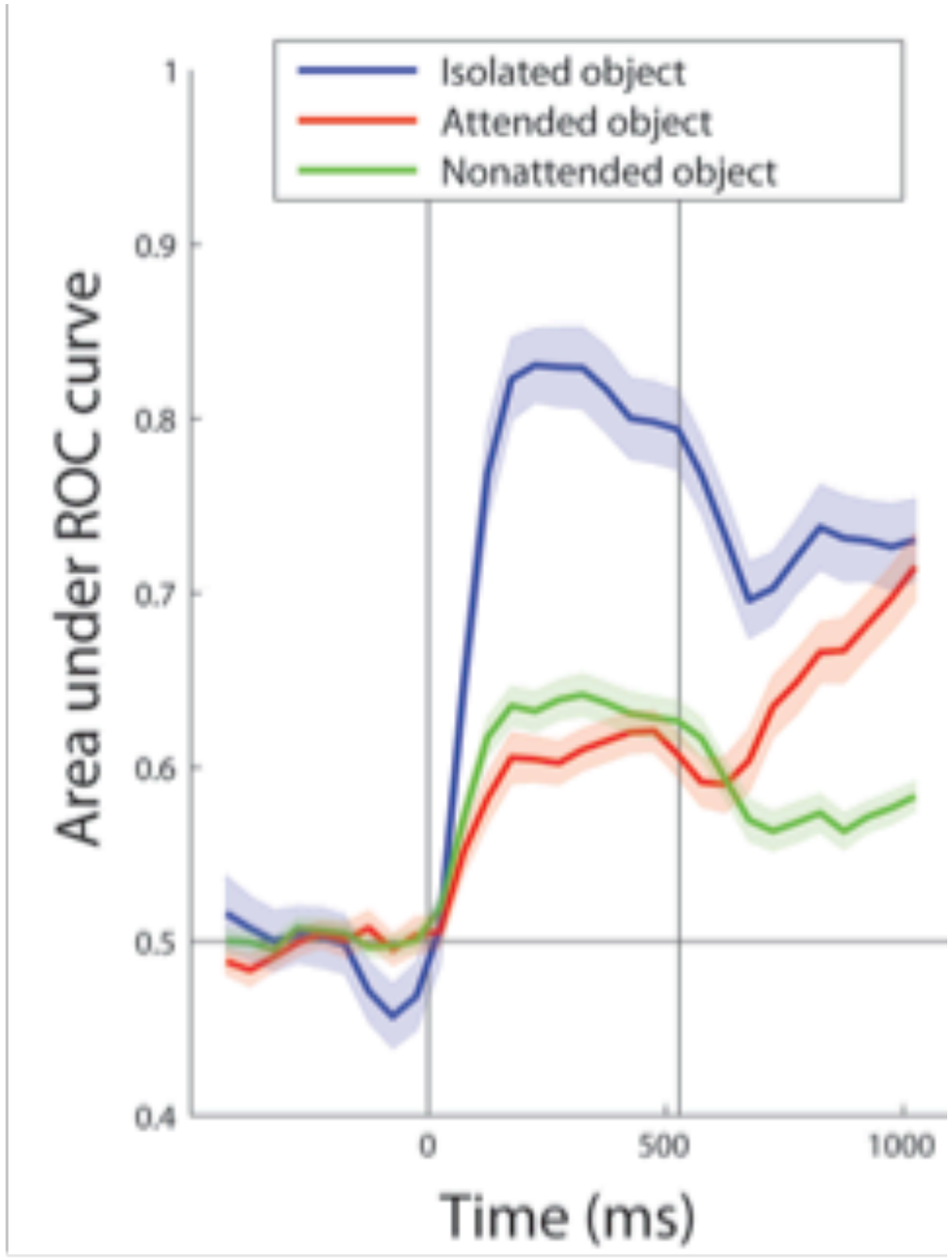


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding object identity

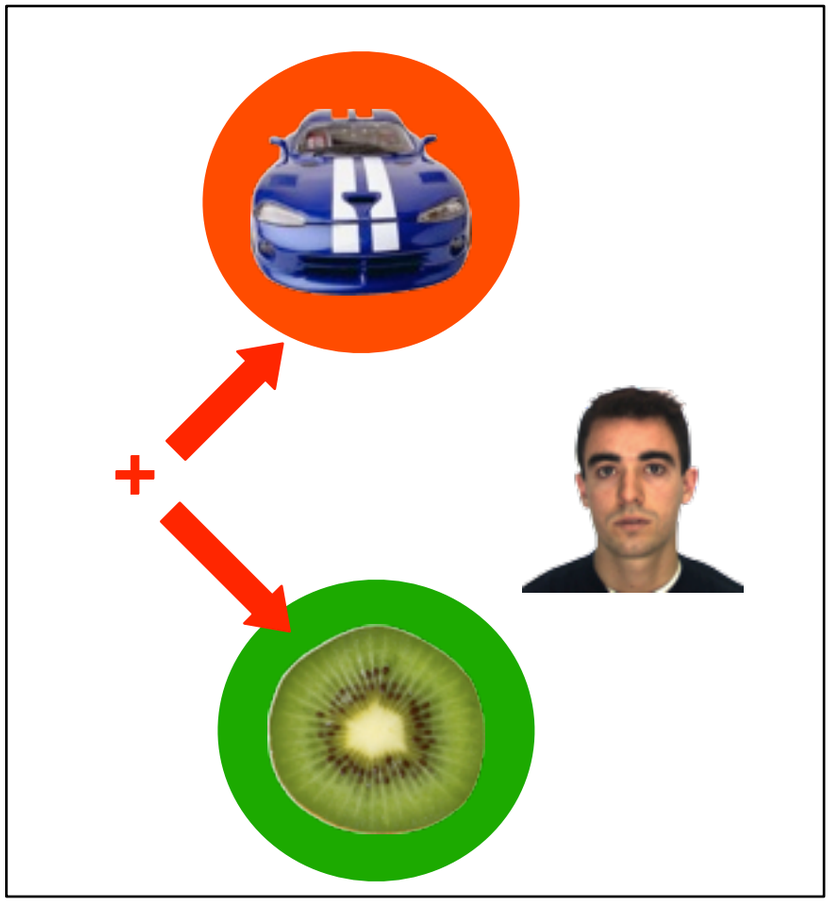


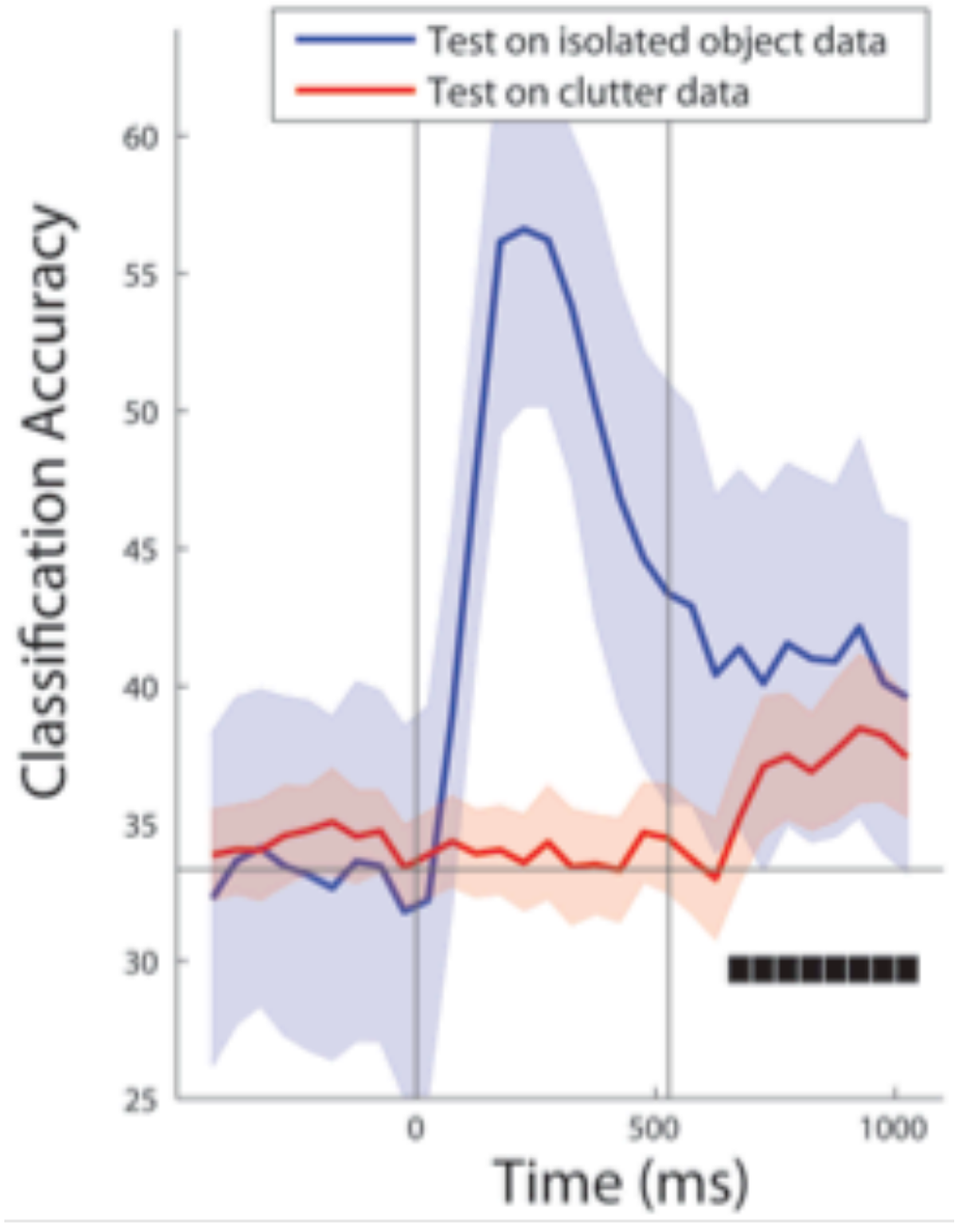


Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Effects of attention on decoding accuracy: Decoding category

Decoding target location

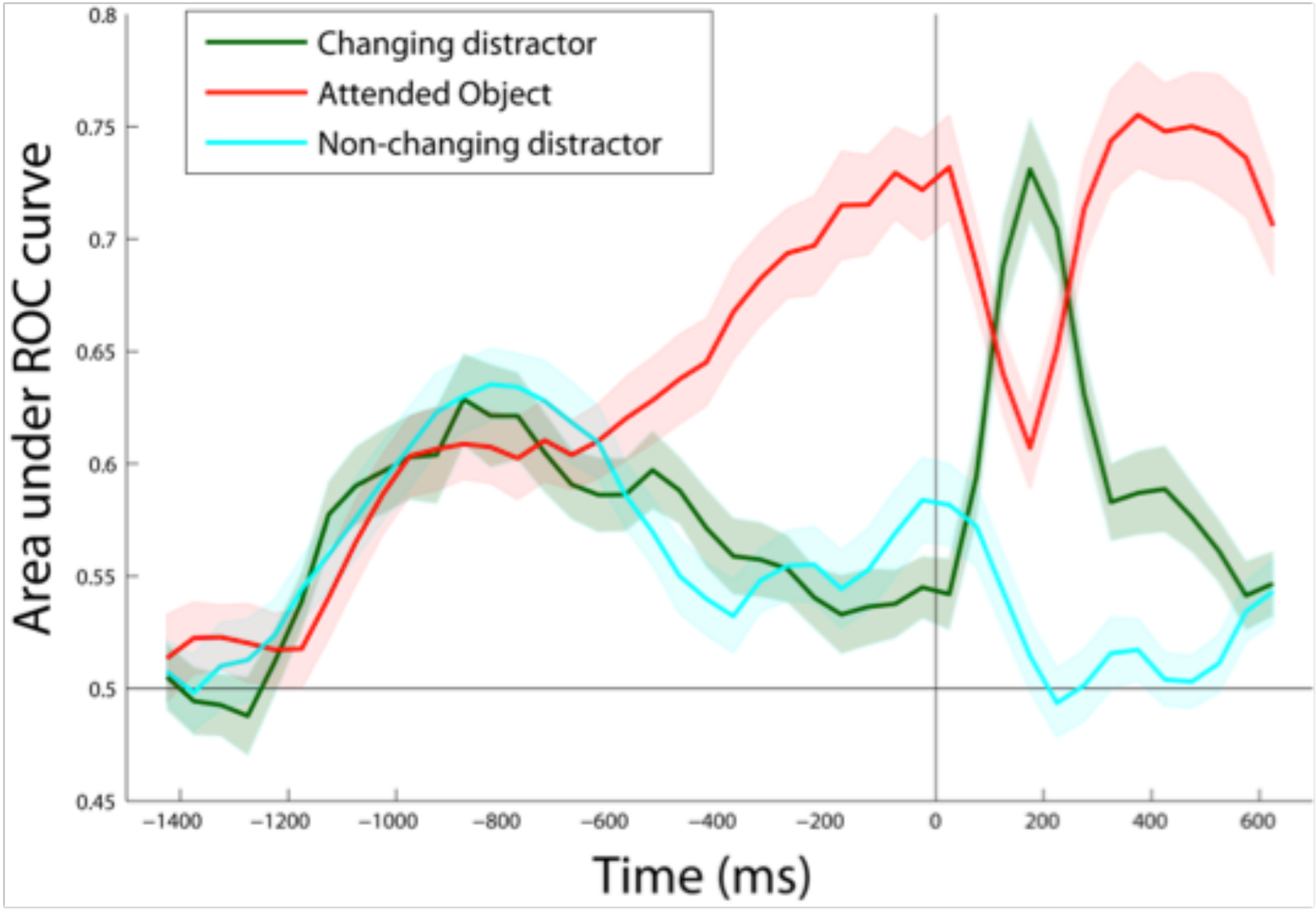




Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Changes in the salience of distractor stimuli dominate over attention related enhancements

Aligned to the time when one of the distractors underwent a change



Zhang* Meyers* Bichot Serre Poggio Desimone in sub

Summary

Summary

- Consistent with feedforward hierarchical models: In the absence of attention information about the identity of individual objects (and position) in clutter is greatly reduced relative to when objects are shown in isolation

Summary

- Consistent with feedforward hierarchical models: In the absence of attention information about the identity of individual objects (and position) in clutter is greatly reduced relative to when objects are shown in isolation
- Attention seems to restore the pattern of neural activity toward the vector representing the isolated object

Summary

- Consistent with feedforward hierarchical models: In the absence of attention information about the identity of individual objects (and position) in clutter is greatly reduced relative to when objects are shown in isolation
- Attention seems to restore the pattern of neural activity toward the vector representing the isolated object
- In spite of this nearly exclusive representation of the attended object, an increase in the salience of non-attended objects overrode these attentional enhancements

Summary

- Consistent with feedforward hierarchical models: In the absence of attention information about the identity of individual objects (and position) in clutter is greatly reduced relative to when objects are shown in isolation
- Attention seems to restore the pattern of neural activity toward the vector representing the isolated object
- In spite of this nearly exclusive representation of the attended object, an increase in the salience of non-attended objects overrode these attentional enhancements
- Results provide computational level explanation for how attention operates on neural representations to solve the problem of invariant recognition in clutter

1. Monkey electrophysiology

Read-out of inferior temporal cortex population activity: Spatial attention eliminates clutter

with Zhang, Meyers, Bichot, Poggio & Desimone

---

2. Computational model of integrated attention and recognition

What and where: a Bayesian attention theory of attention

with Chikkerur, Tan & Poggio

image measurements

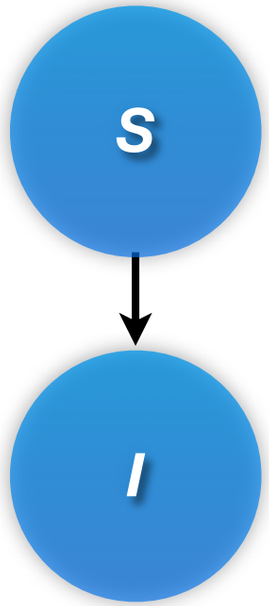
Perception as Bayesian inference

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; Yu & Dayan ’05; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

image measurements

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; Yu & Dayan ’05; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07



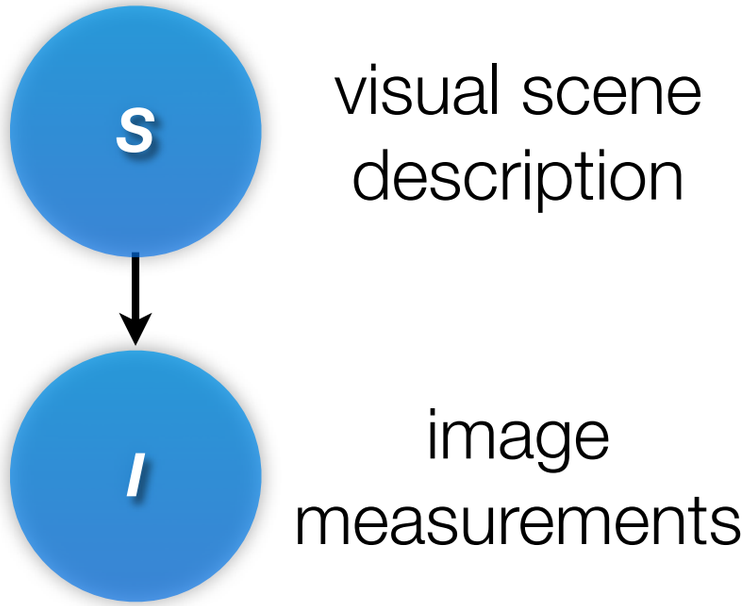
visual scene description

image measurements

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; Yu & Dayan ’05; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

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

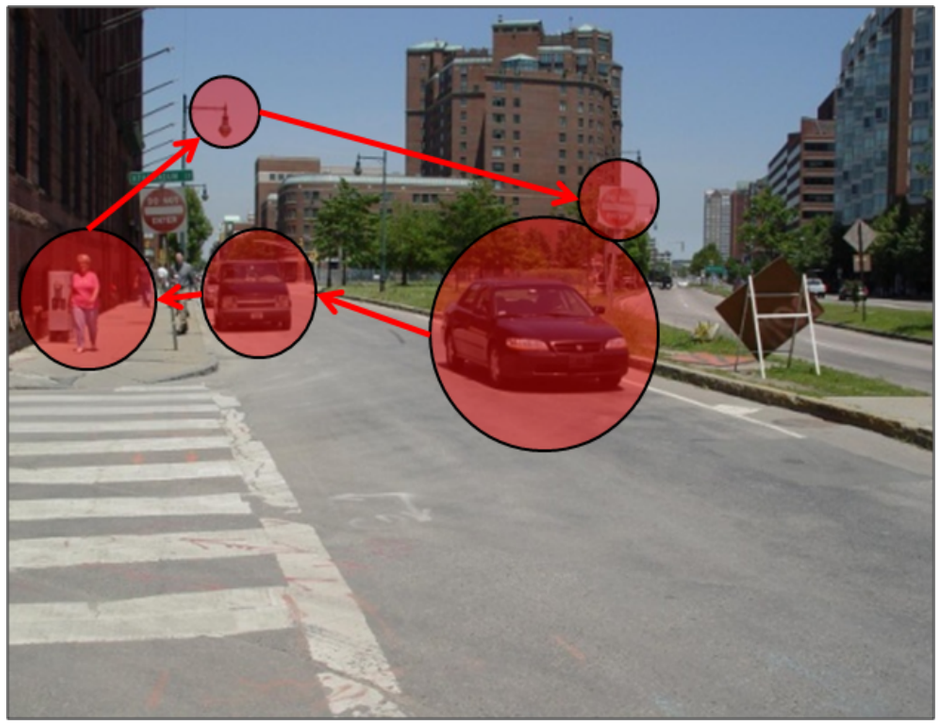


Perception as Bayesian inference

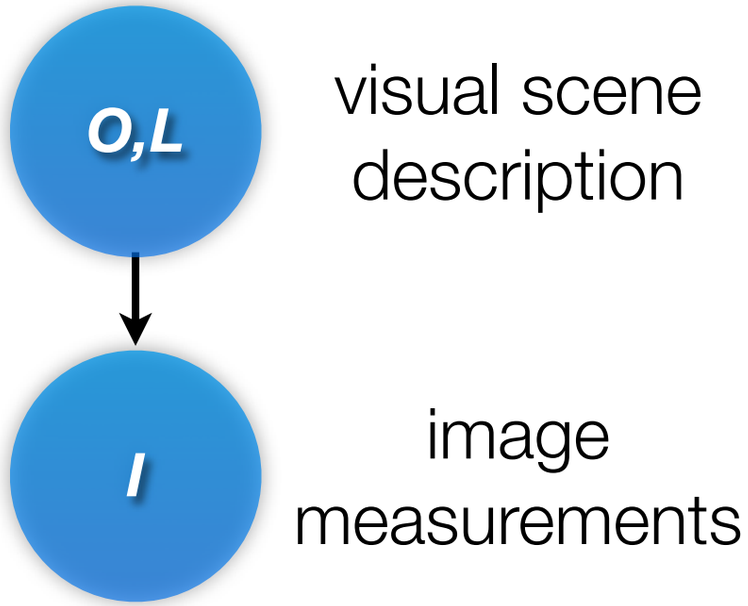
Mumford ’92; Knill & Richards ‘96; Dayan & Zemel ’99; Rao ’02 ’04; Kersten & Yuille ‘03; Kersten et al ‘04; Lee & Mumford ‘03; Dean ’05; George & Hawkins ’05 ’09; Yu & Dayan ’05; Hinton ‘07; Epshtein et al ‘08; Murray & Kreutz-Delgado ’07

Hypothesis #1

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



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



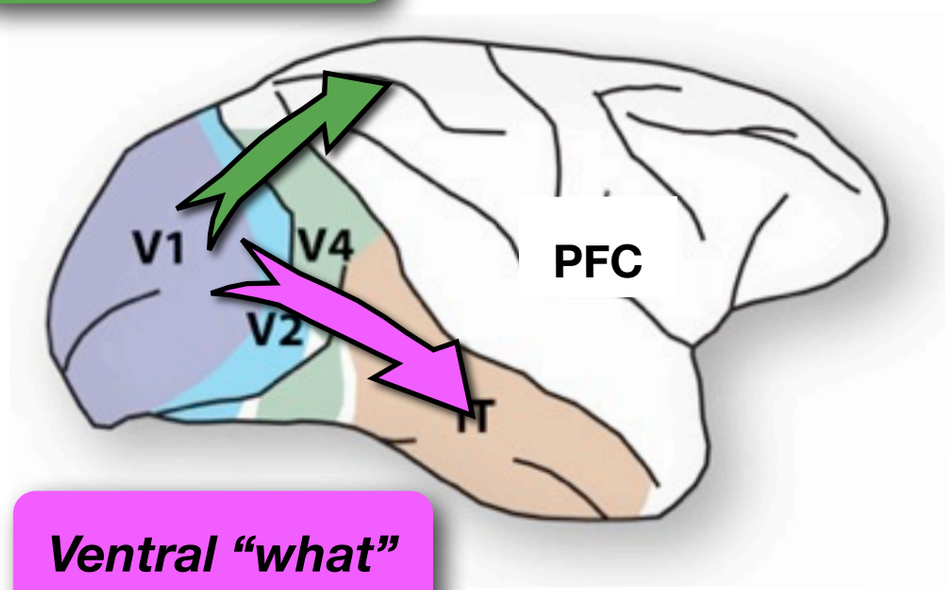
Chikkerur Serre Tan & Poggio ’10;

Hypothesis #2

Two independent streams of processing for object and location

Explicit tuning for position
(shape implicit)

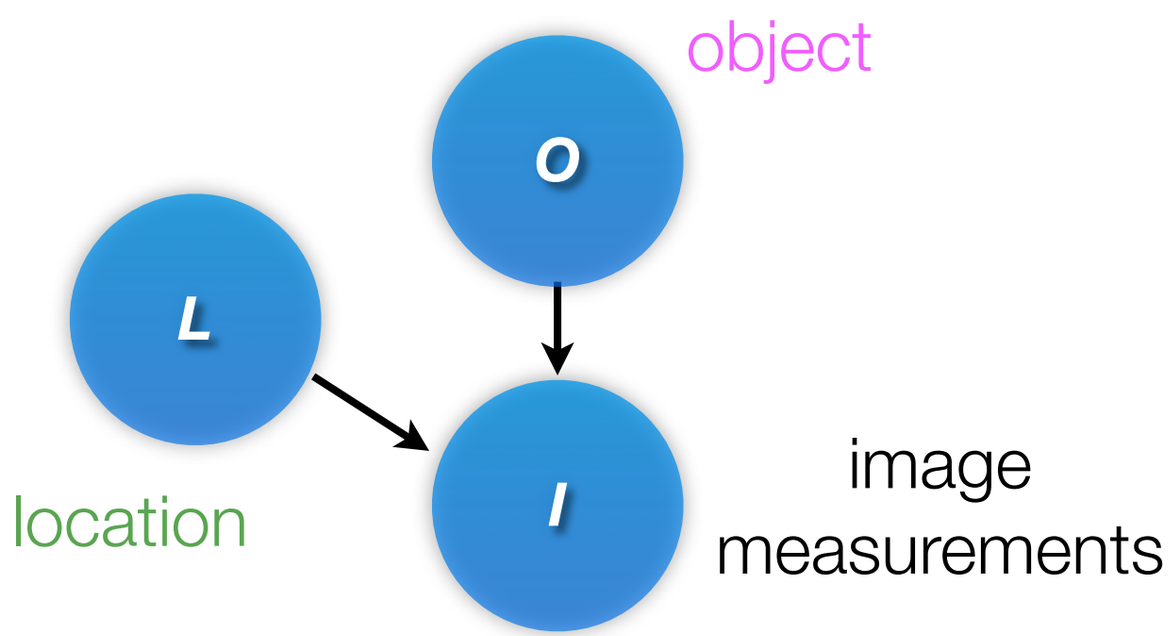
Dorsal "where" stream



Ventral "what" stream

Explicit tuning for shape
(position implicit)

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



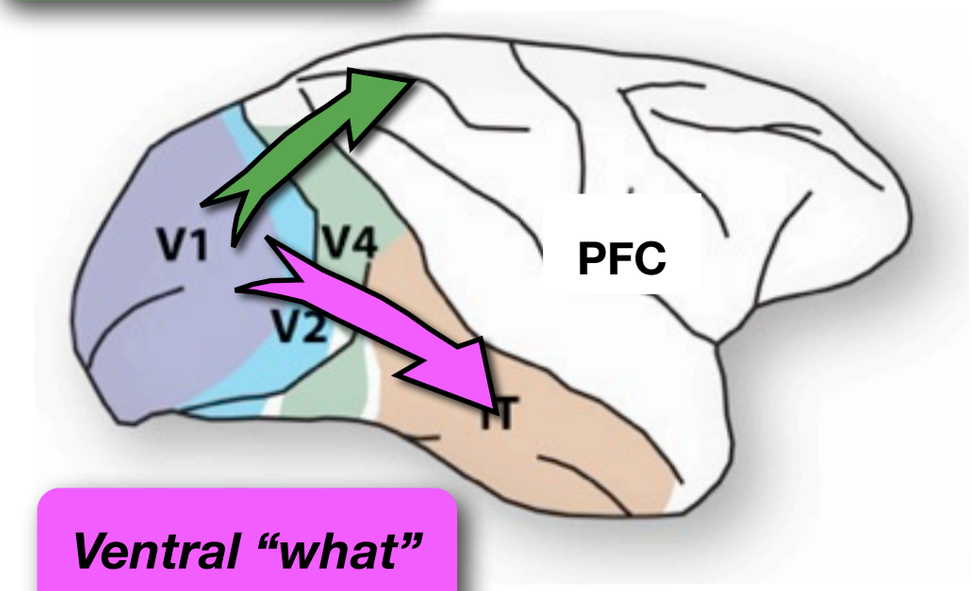
Chikkerur Serre Tan & Poggio '10;

Hypothesis #3

Objects encoded by collections of generic features (cond. ind. given an object and its location)

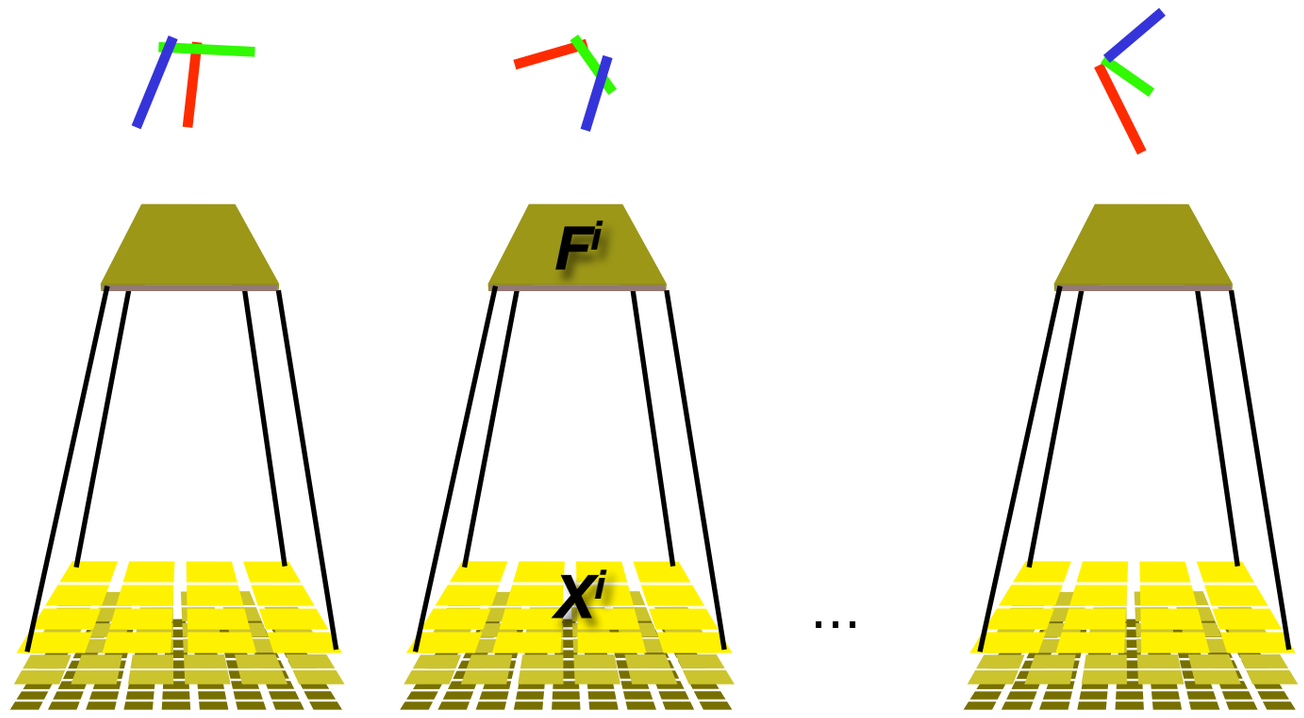
Explicit tuning for position
(shape implicit)

Dorsal "where" stream



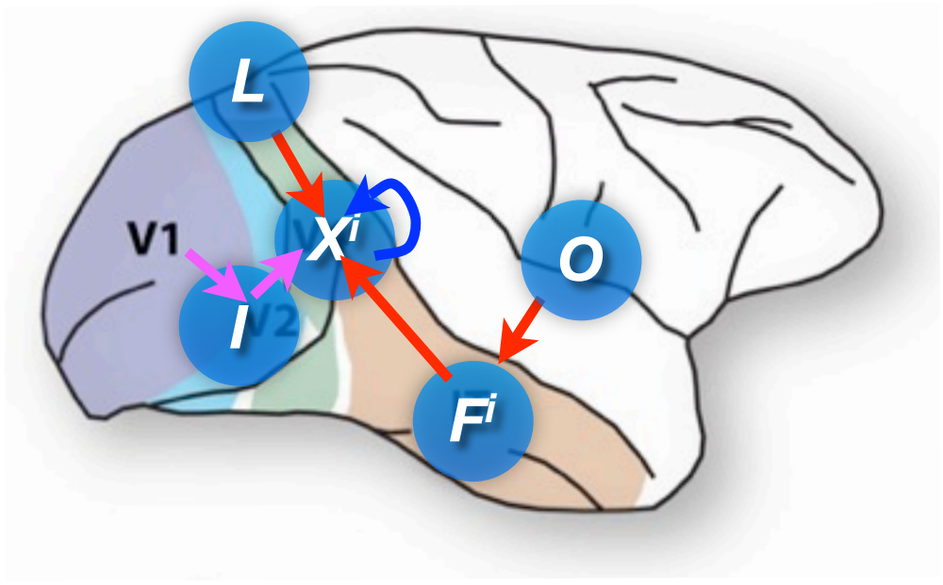
Ventral "what" stream

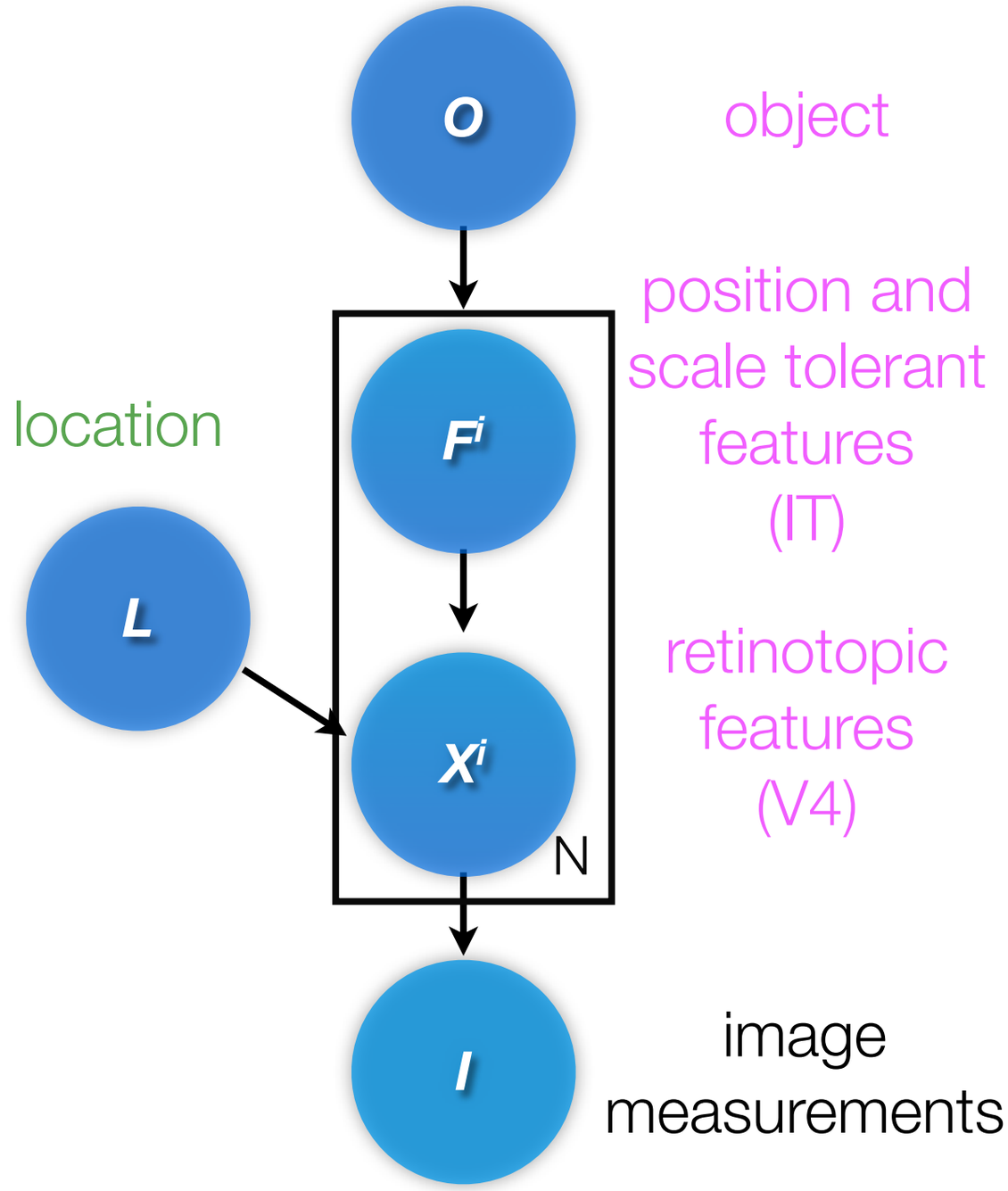
Explicit tuning for shape
(position implicit)



Chikkerur Serre Tan & Poggio '10;

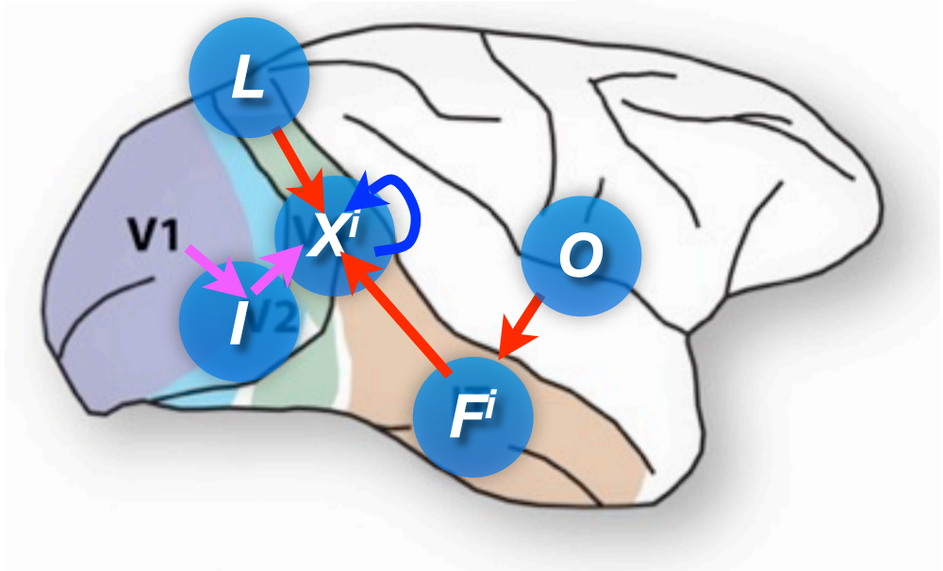
Perception as Bayesian inference



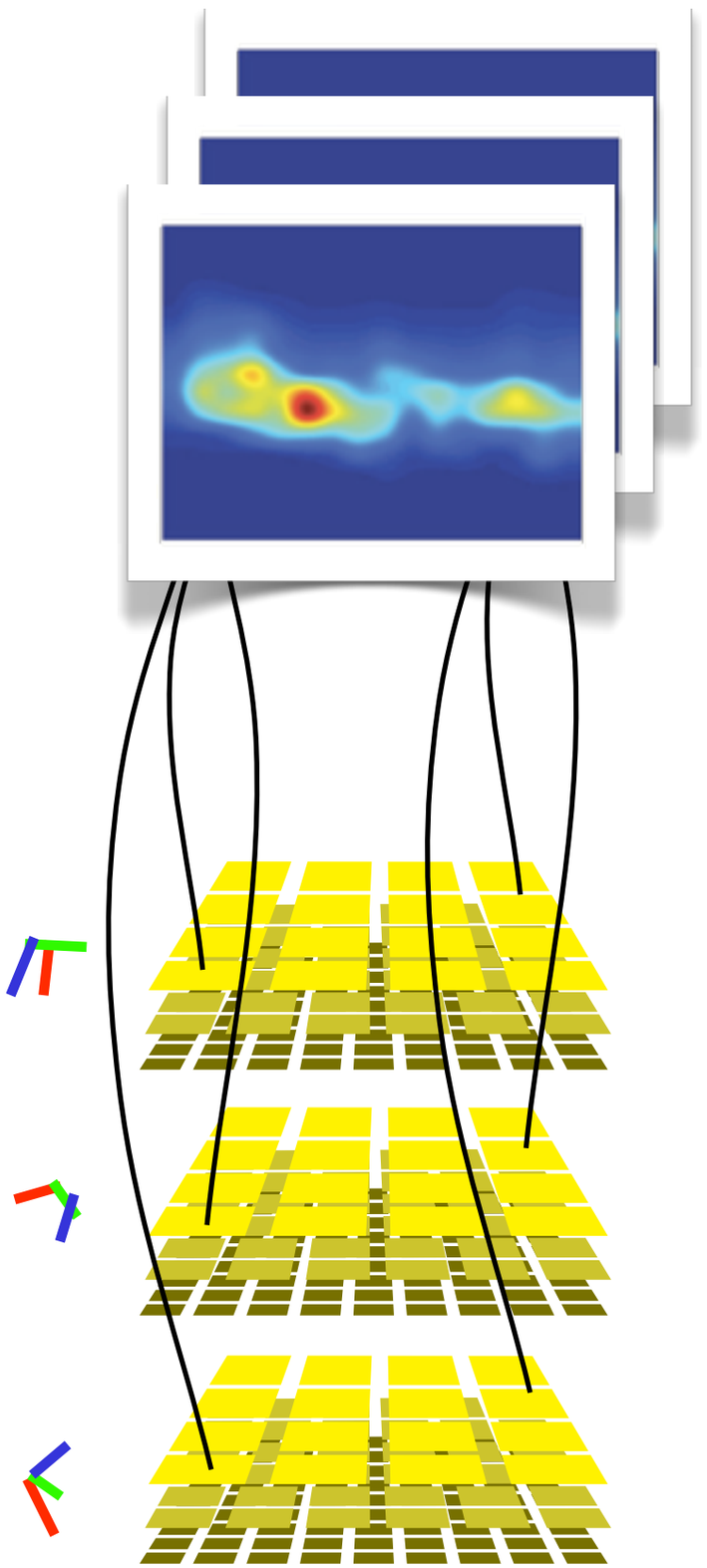


Chikkerur Serre Tan & Poggio ’10;

Perception as Bayesian inference



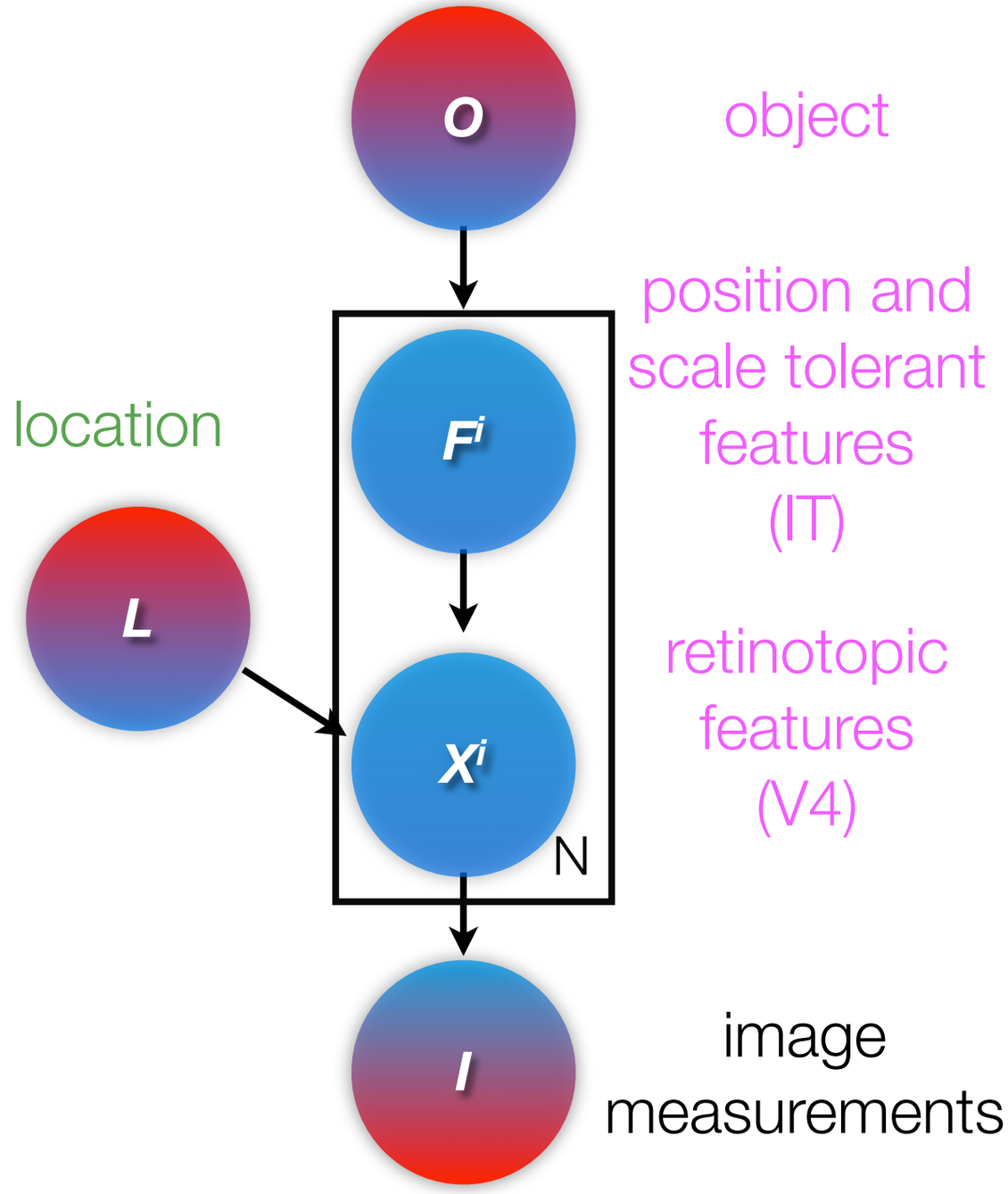
p(L|I)
saliency map



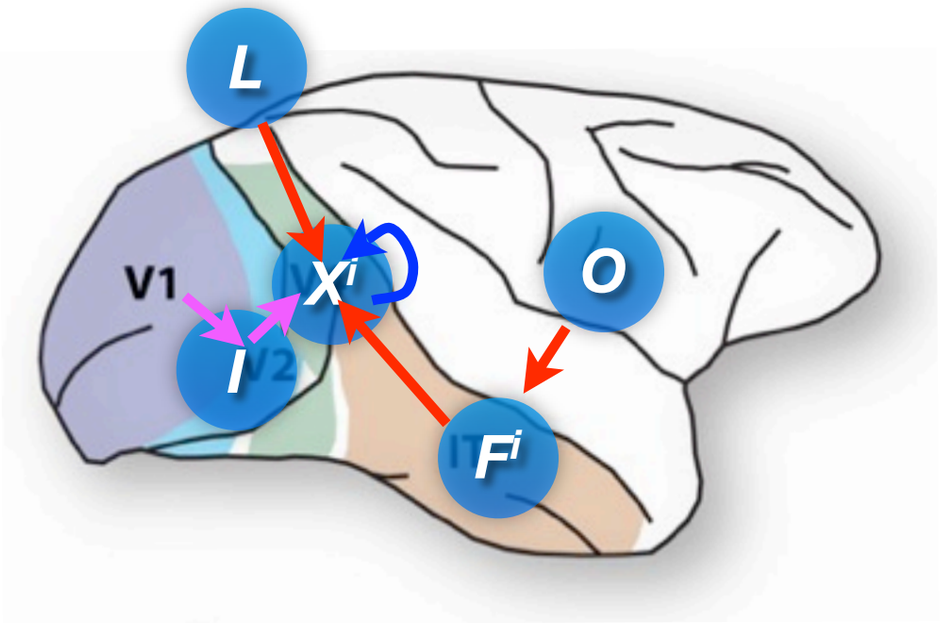
Chikkerur Serre Tan & Poggio ’10;

Perception as Bayesian inference

- ***Goal of visual perception*** is 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 objects or locations (as well as on sensory input), and doing inference over the other latent variables
- Here we used belief propagation to solve the inference problem

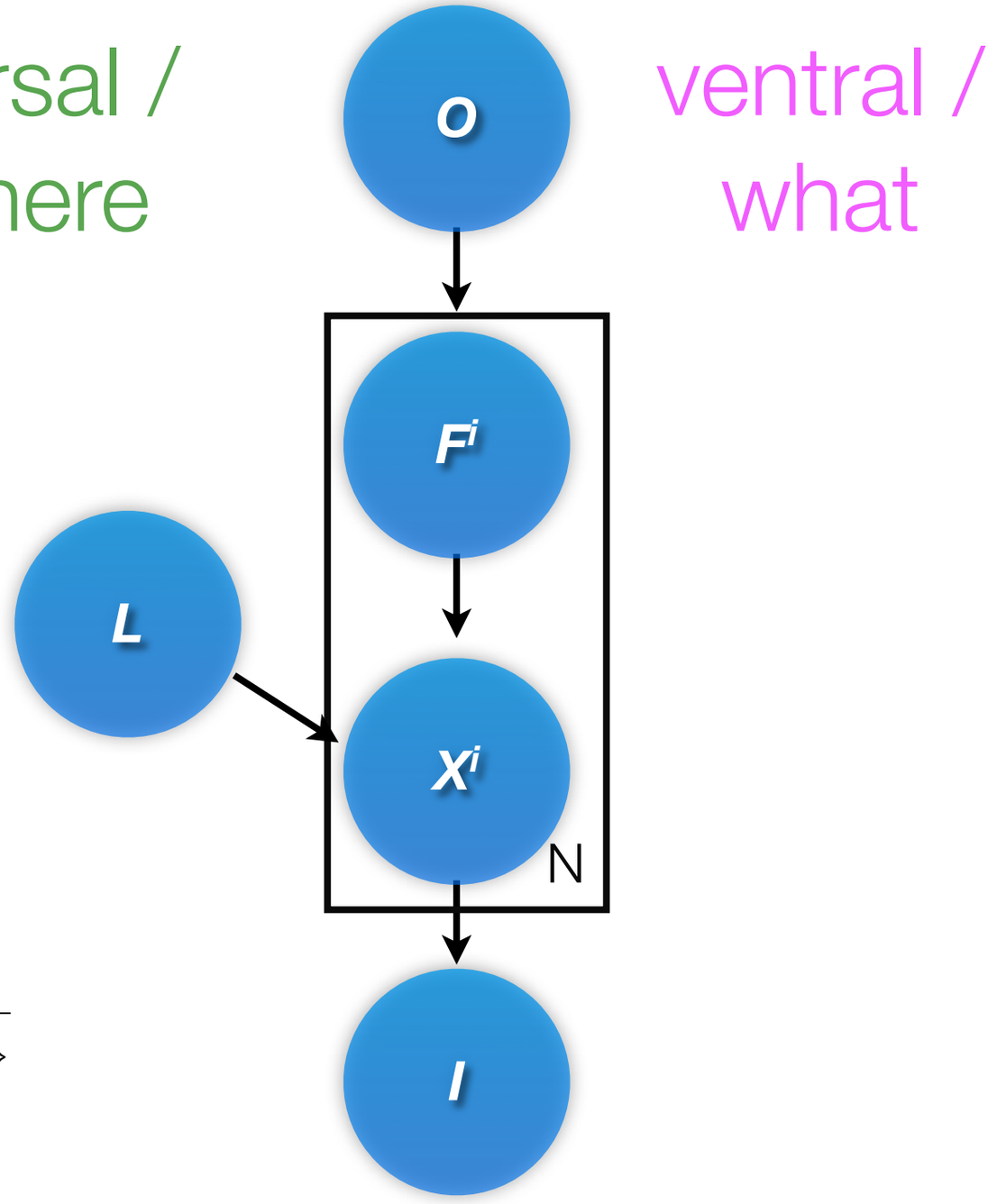


Biologically-plausible implementations of belief propagation: Zemel et al. ’98; Beck & Pouget ‘07; Deneve ’08; George ‘08; Litvak & Ullman ‘09; Rao ‘04; Steimer et al. ’09



dorsal / where

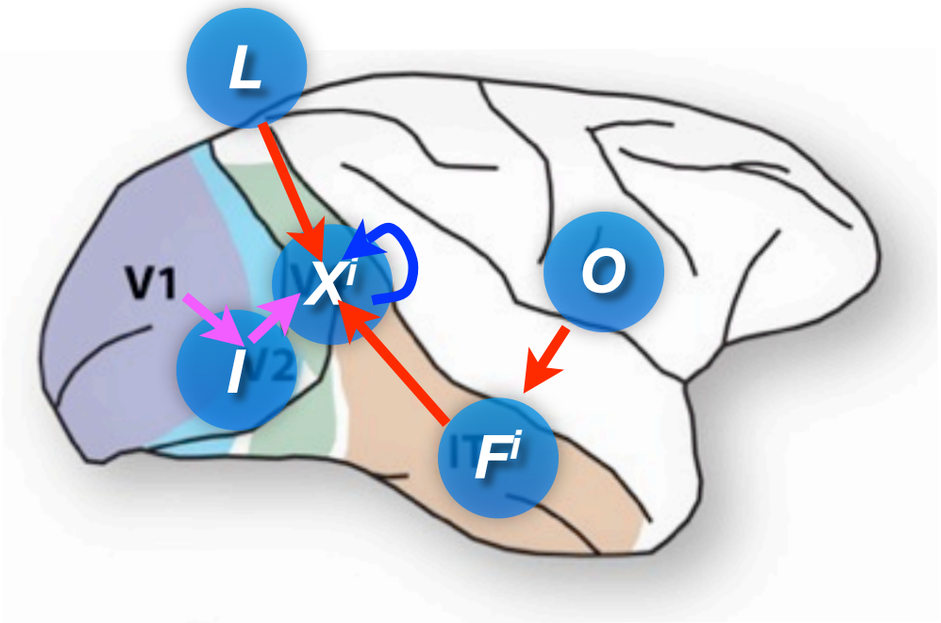
ventral / what



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

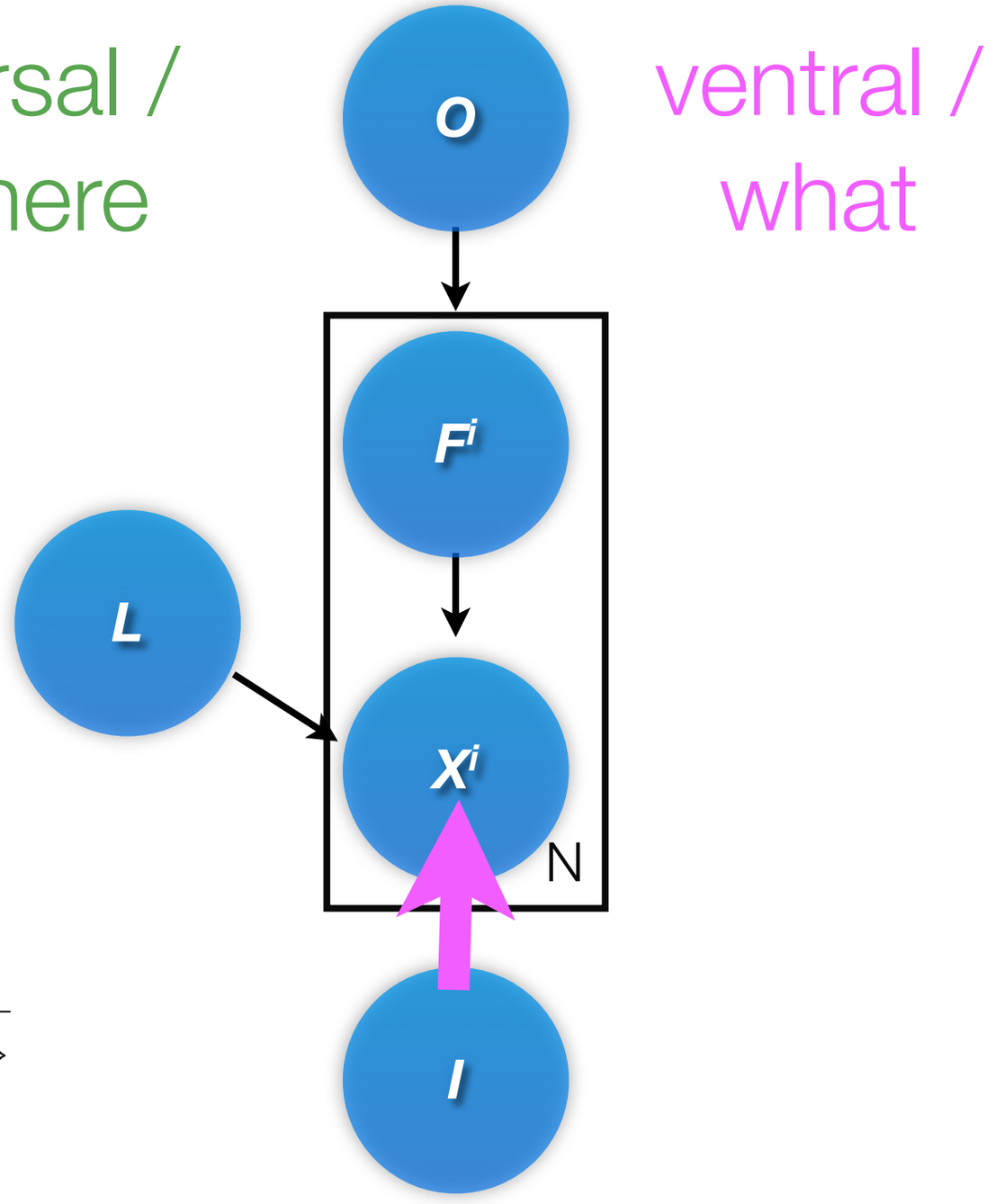
Bayesian inference and attention

Special case of the normalization model of attention by Reynolds & Heeger ’09



dorsal / where

ventral / what



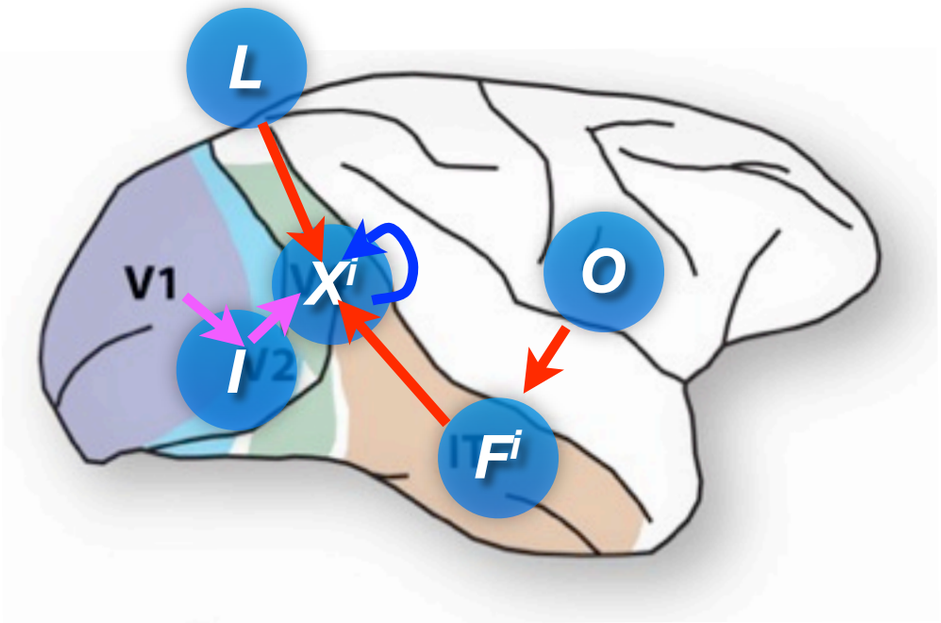
feedforward (bottom-up) sweep

feedforward input

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

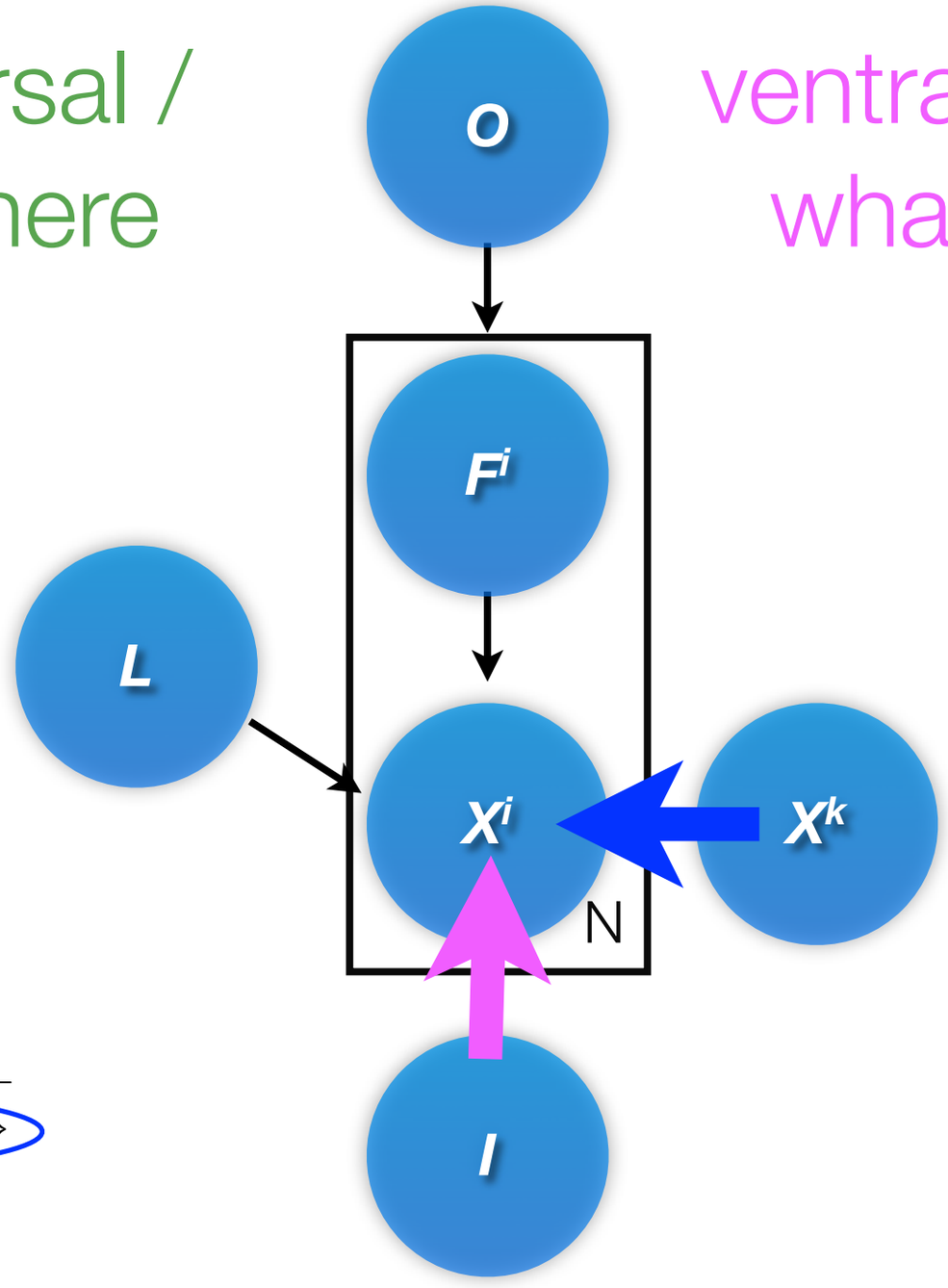
Bayesian inference and attention

Special case of the normalization model of attention by Reynolds & Heeger ’09



dorsal / where

ventral / what



feedforward (bottom-up) sweep

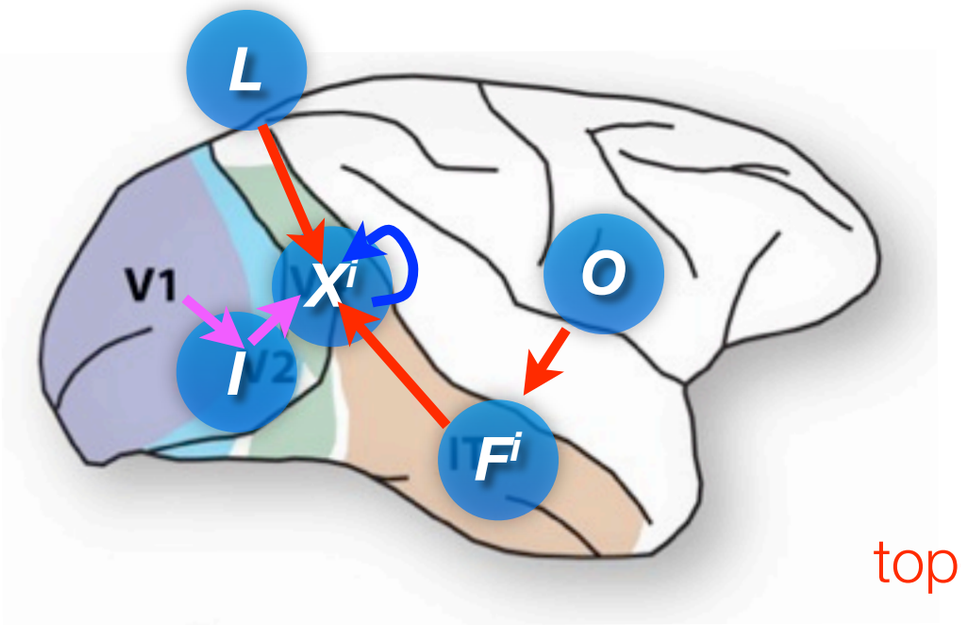
feedforward input

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

suppressive drive

Bayesian inference and attention

Special case of the normalization model of attention by Reynolds & Heeger ’09

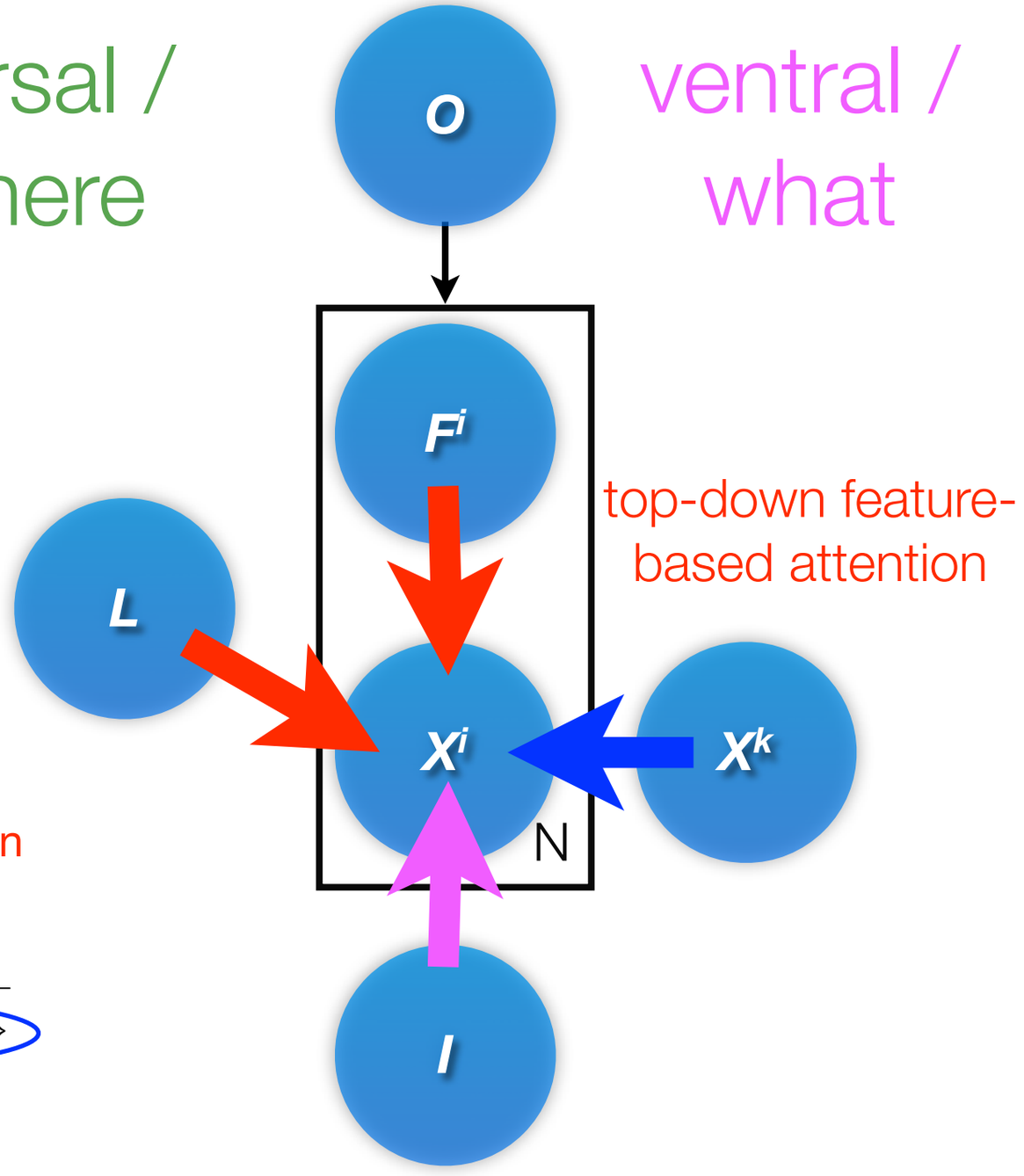


dorsal / where

ventral / what

top-down spatial attention

top-down feature-based attention



feedforward input

attentional modulation

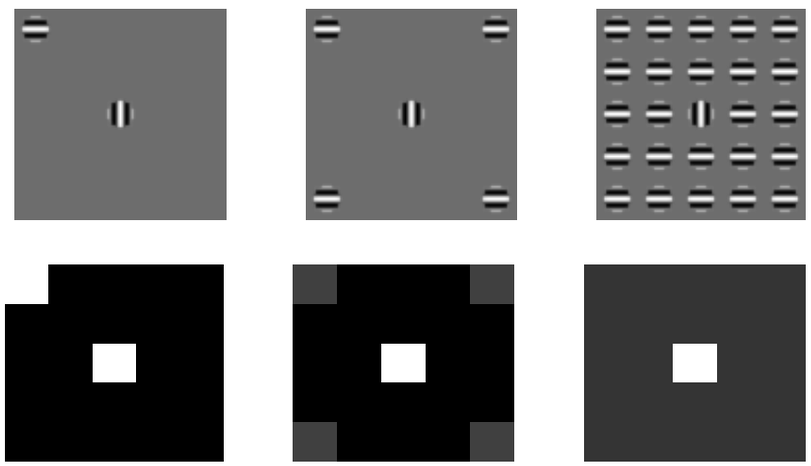
$$P(X^i|I) = \frac{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

suppressive drive

feedforward (bottom-up) sweep

Bayesian inference and attention

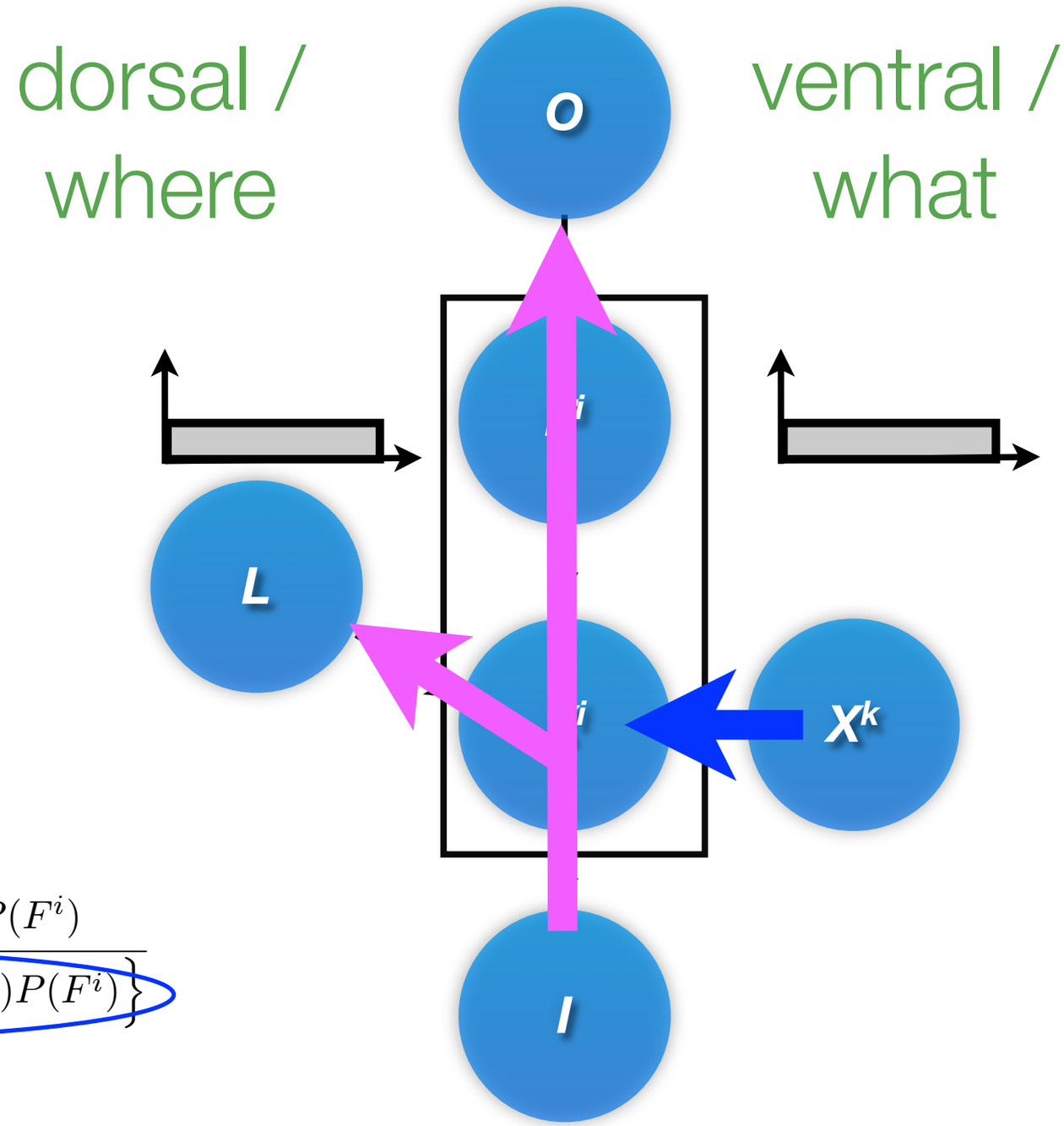
Special case of the normalization model of attention by Reynolds & Heeger ’09



feedforward input

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

suppressive drive

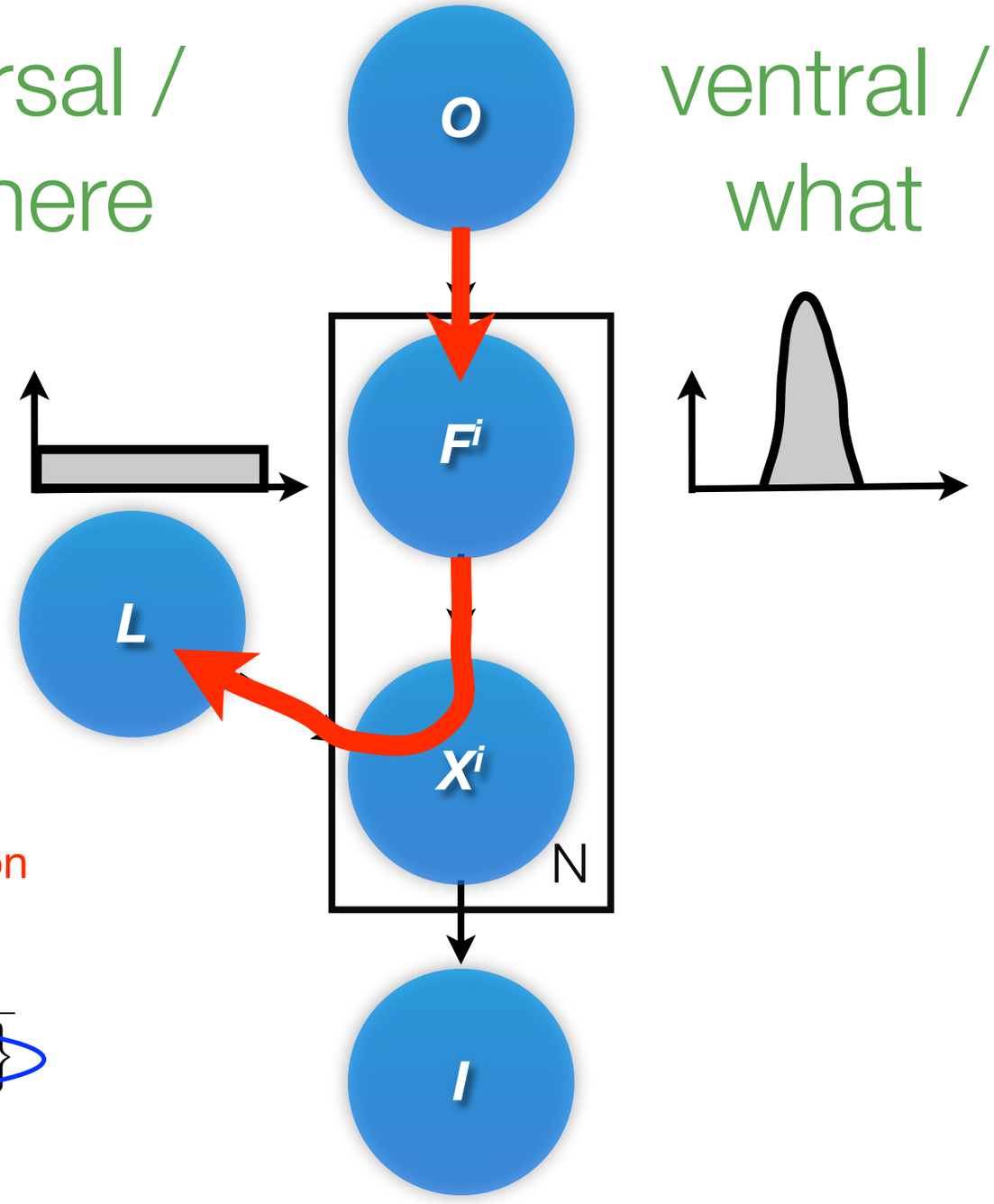


Bayesian inference and attention

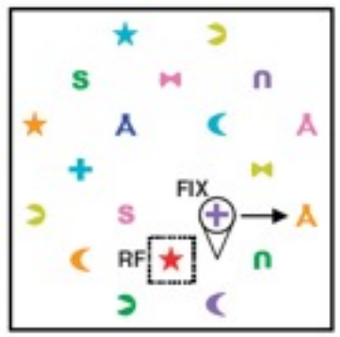
(parallel) feature-based attention

dorsal / where

ventral / what



Visual search



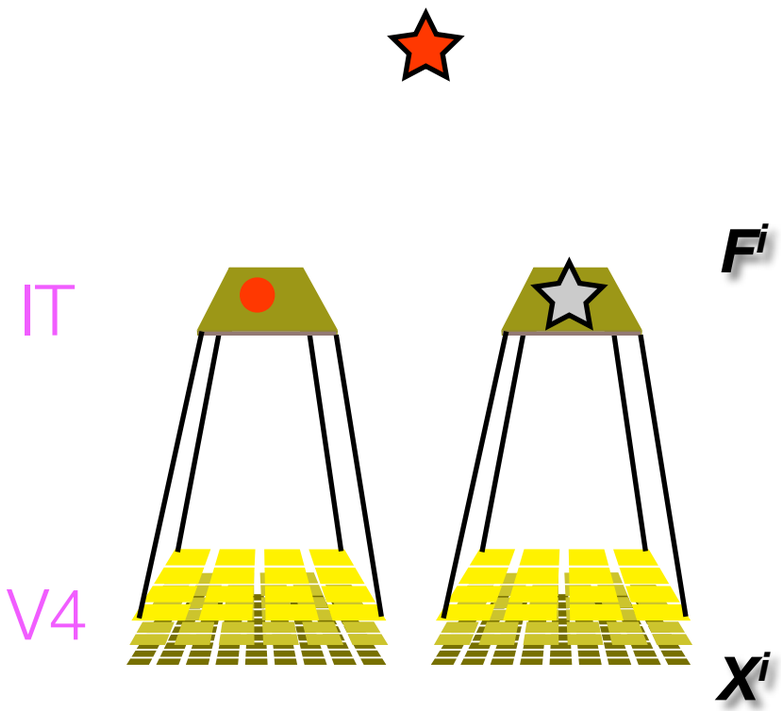
feedforward input

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

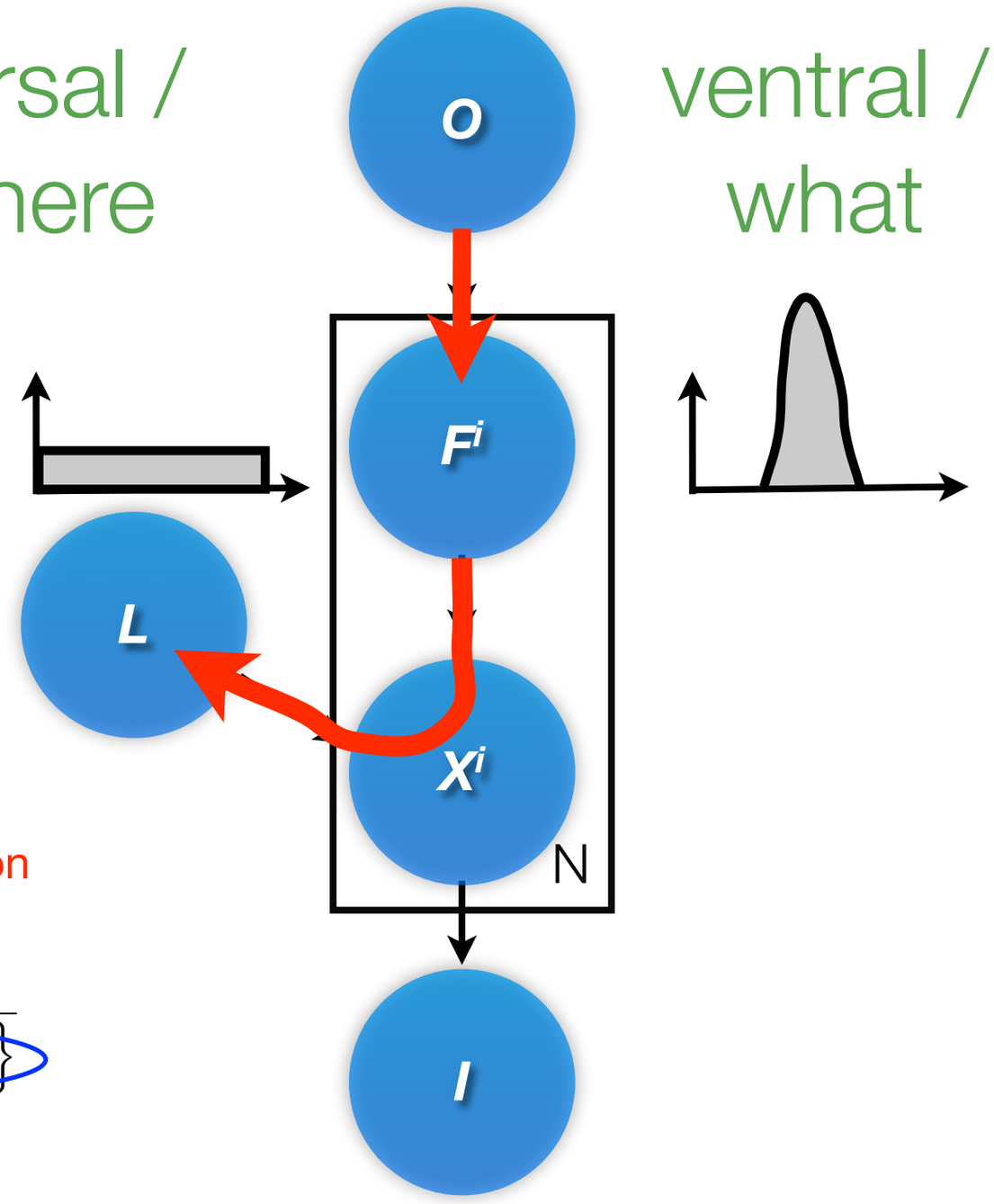


feedforward input

attentional modulation

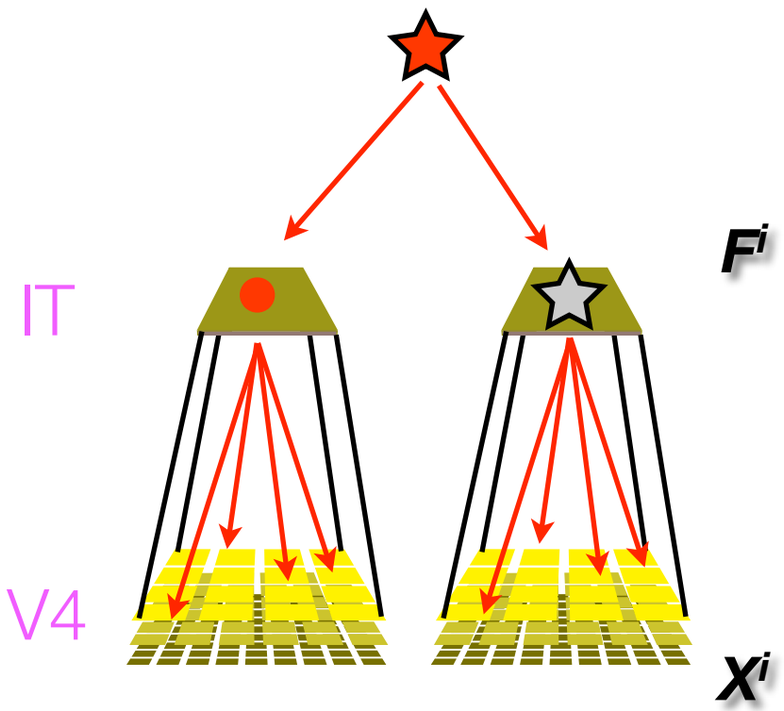
$$P(X^i|I) = \frac{P(I|X^i)\ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

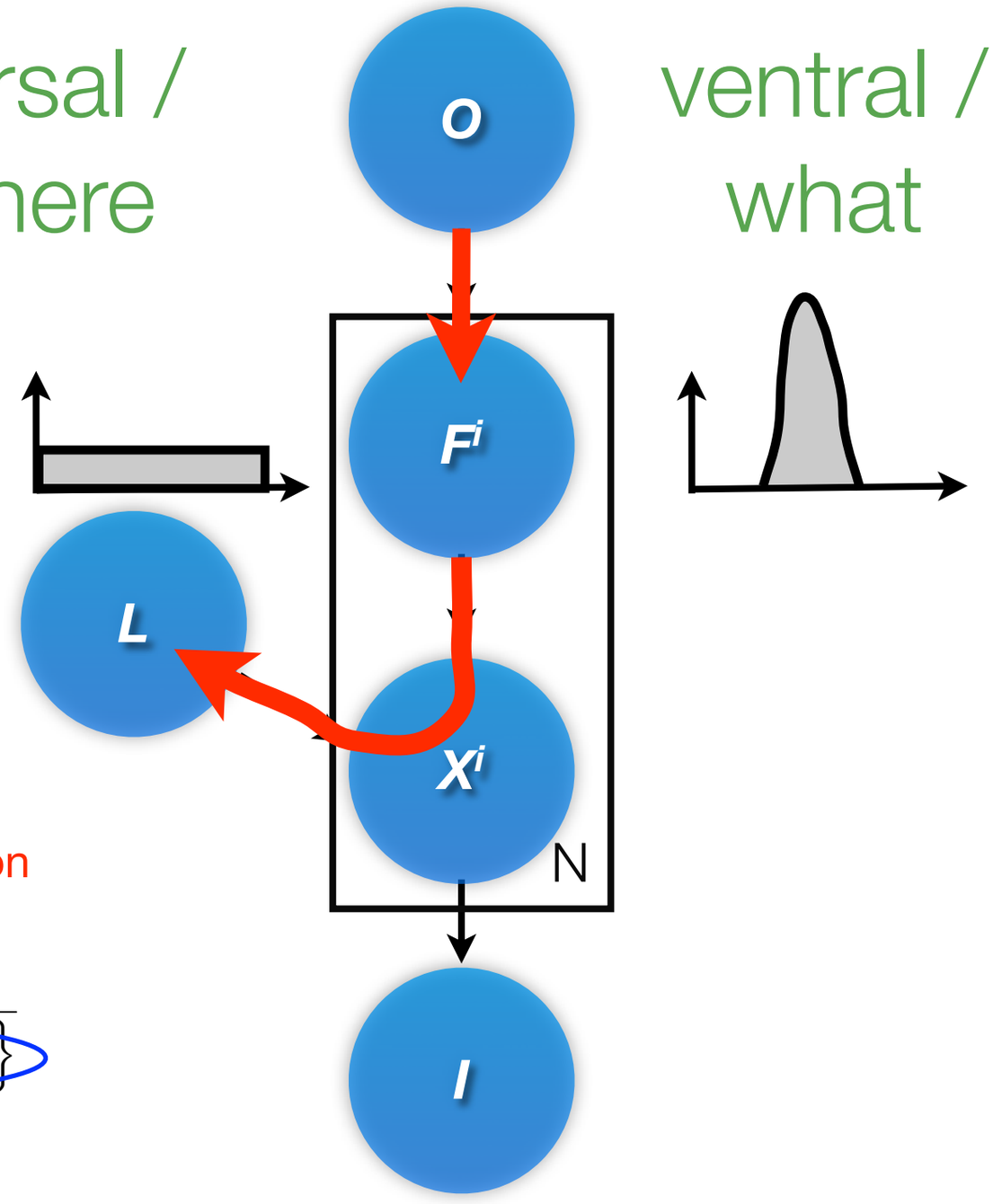
suppressive drive



Bayesian inference and attention

Consistent with data from V4 by Bichot et al ’05





feedforward input

attentional modulation

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

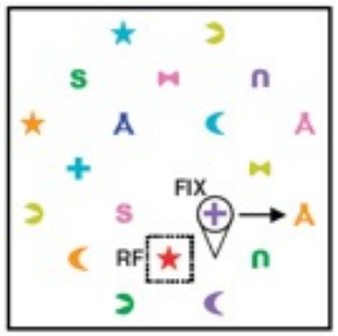
suppressive drive

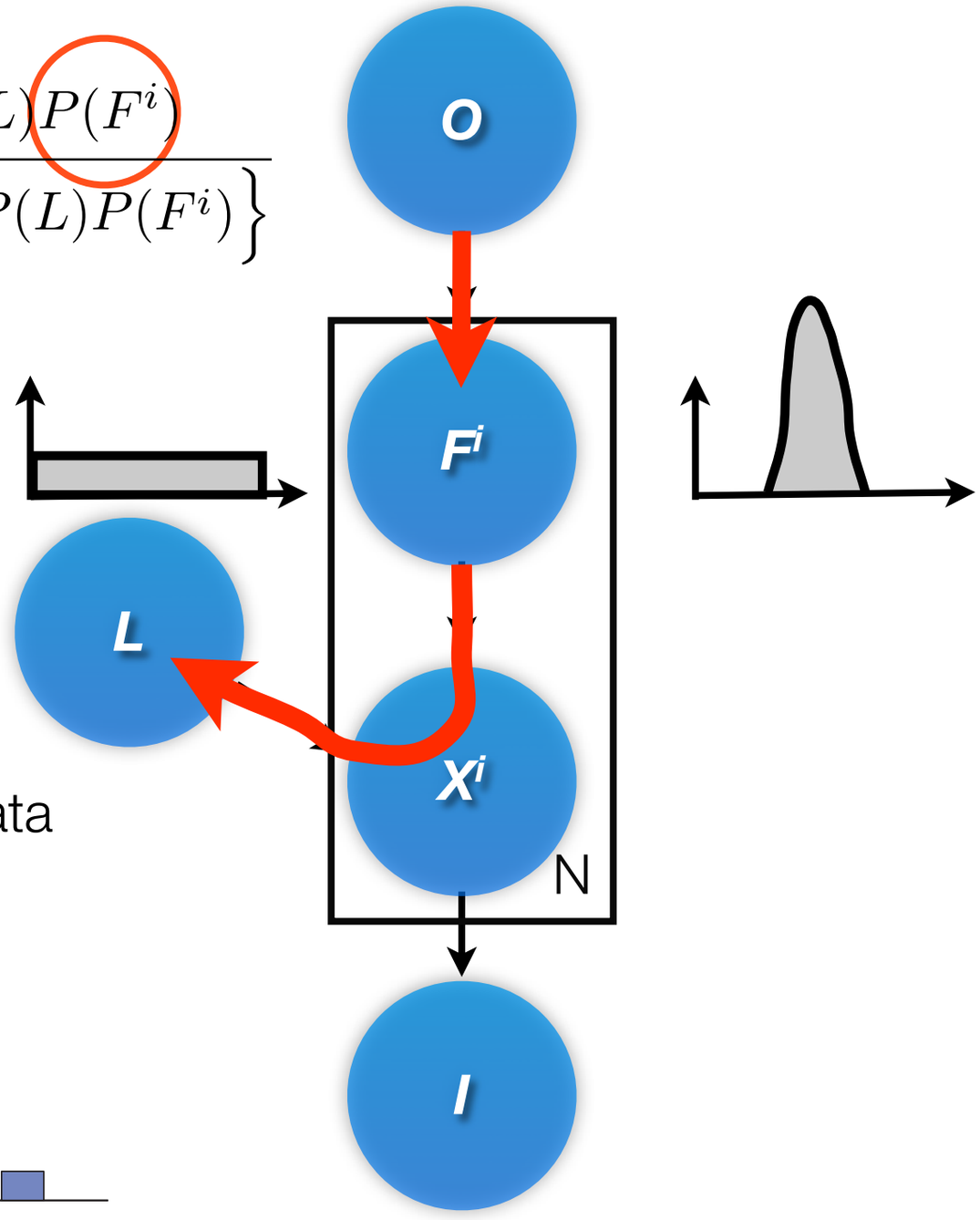
Bayesian inference and attention

Consistent with data from V4 by Bichot et al ’05

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

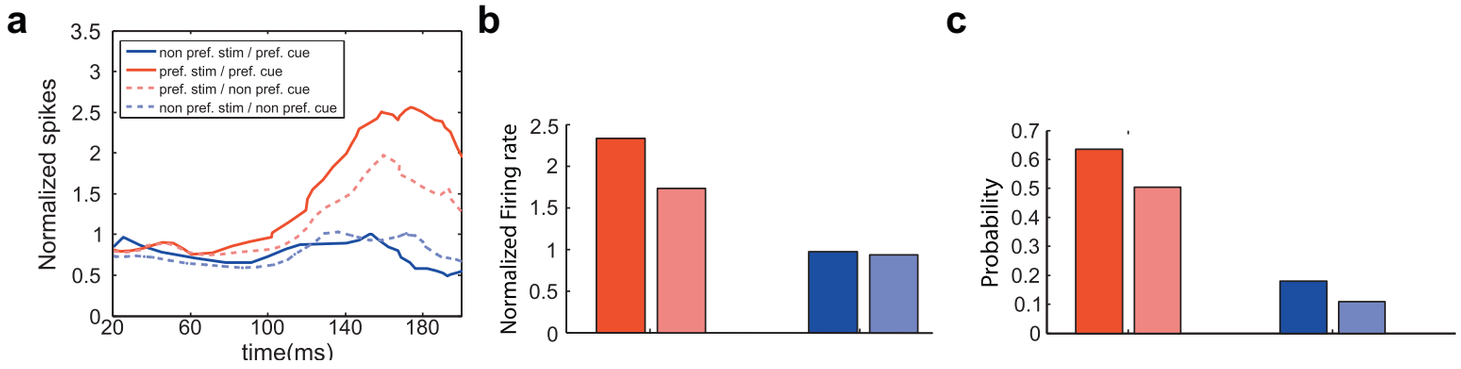
Visual search





Experimental data

Model data



Bayesian inference and attention

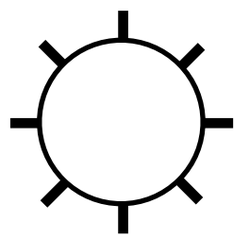
Consistent with data from V4 by Bichot et al ’05

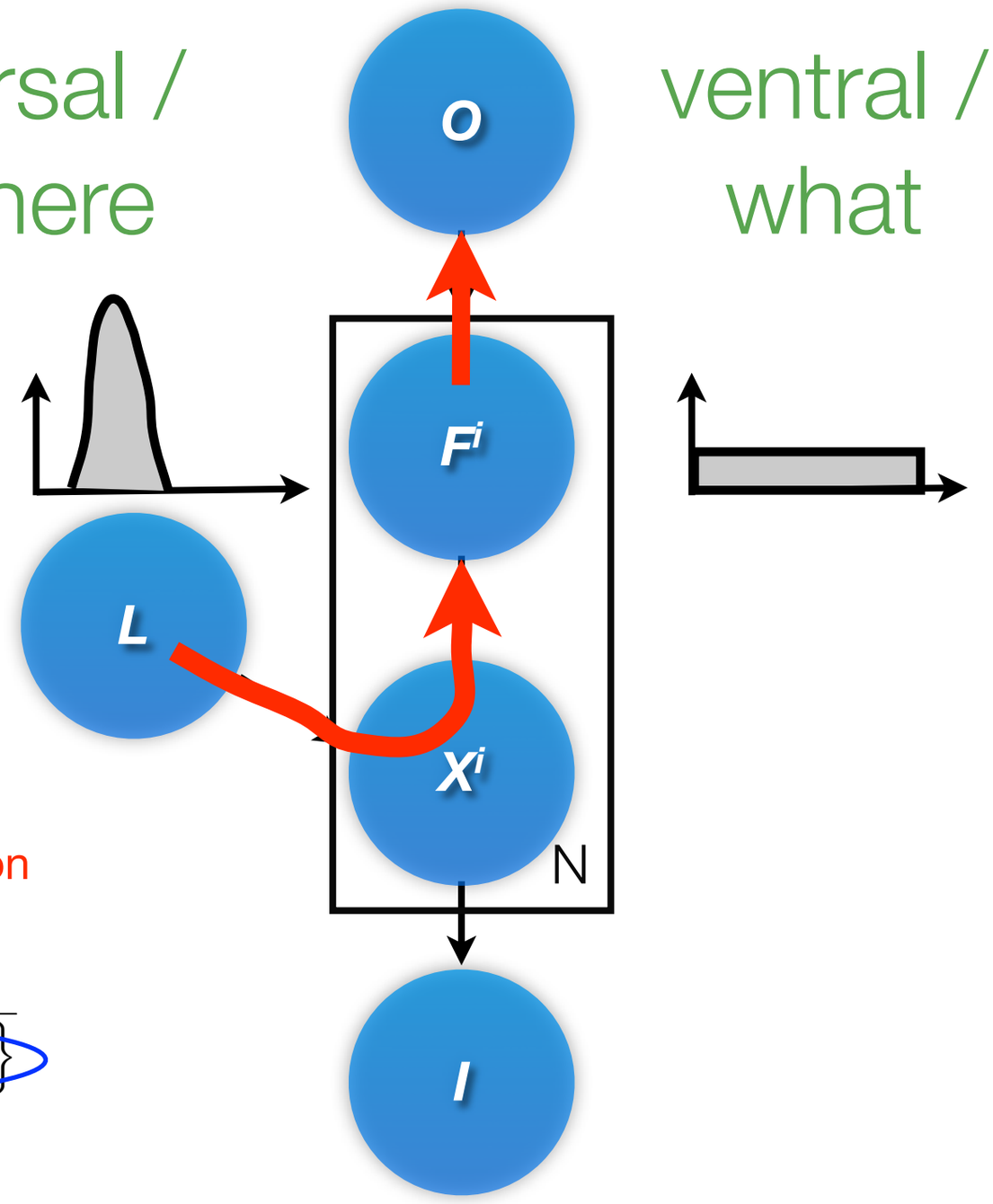
(serial) spatial attention

dorsal / where

ventral / what

Spatial cueing





feedforward input

attentional modulation

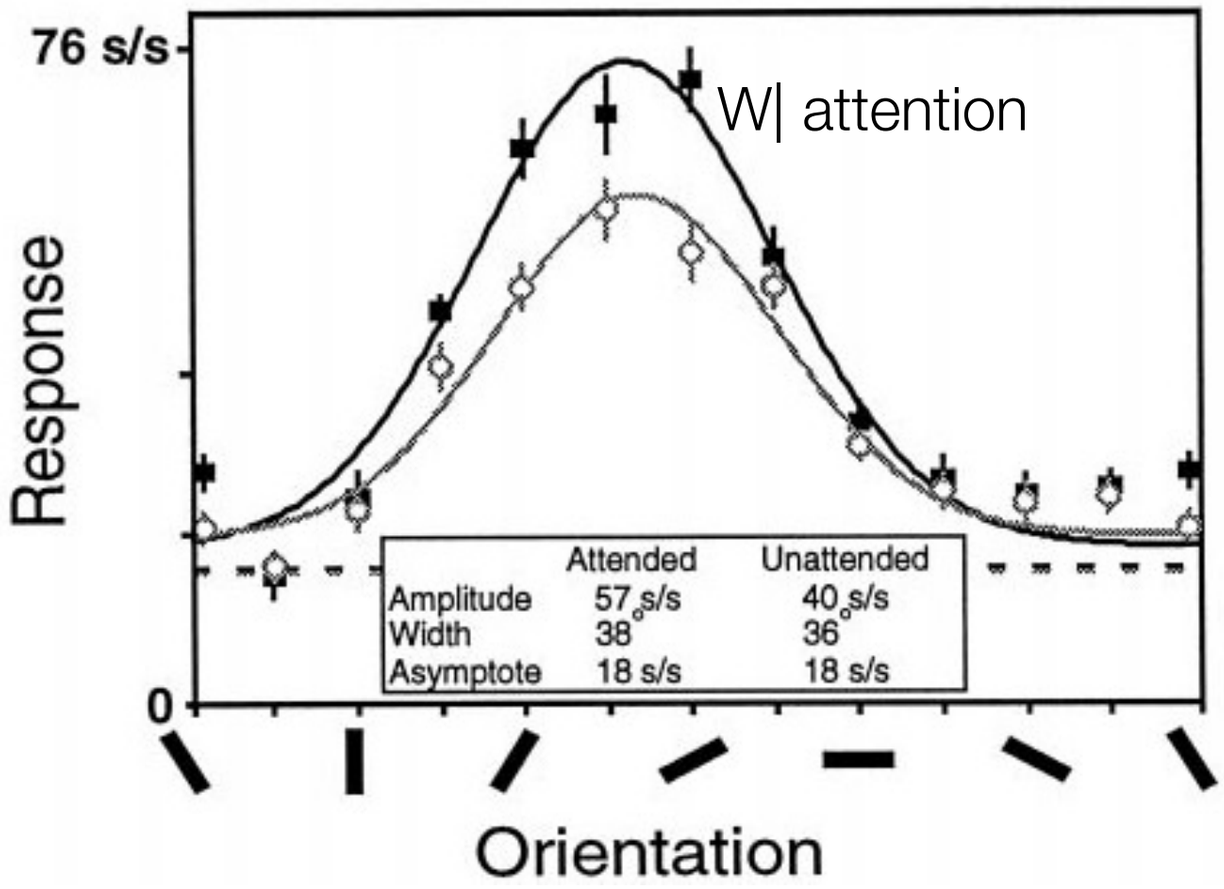
$$P(X^i|I) = \frac{P(I|X^i)\ \sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)}{\sum_{X^i}\left\{P(I|X^i)\sum_{F^i,L} P(X^i|F^i,L)P(L)P(F^i)\right\}}$$

suppressive drive

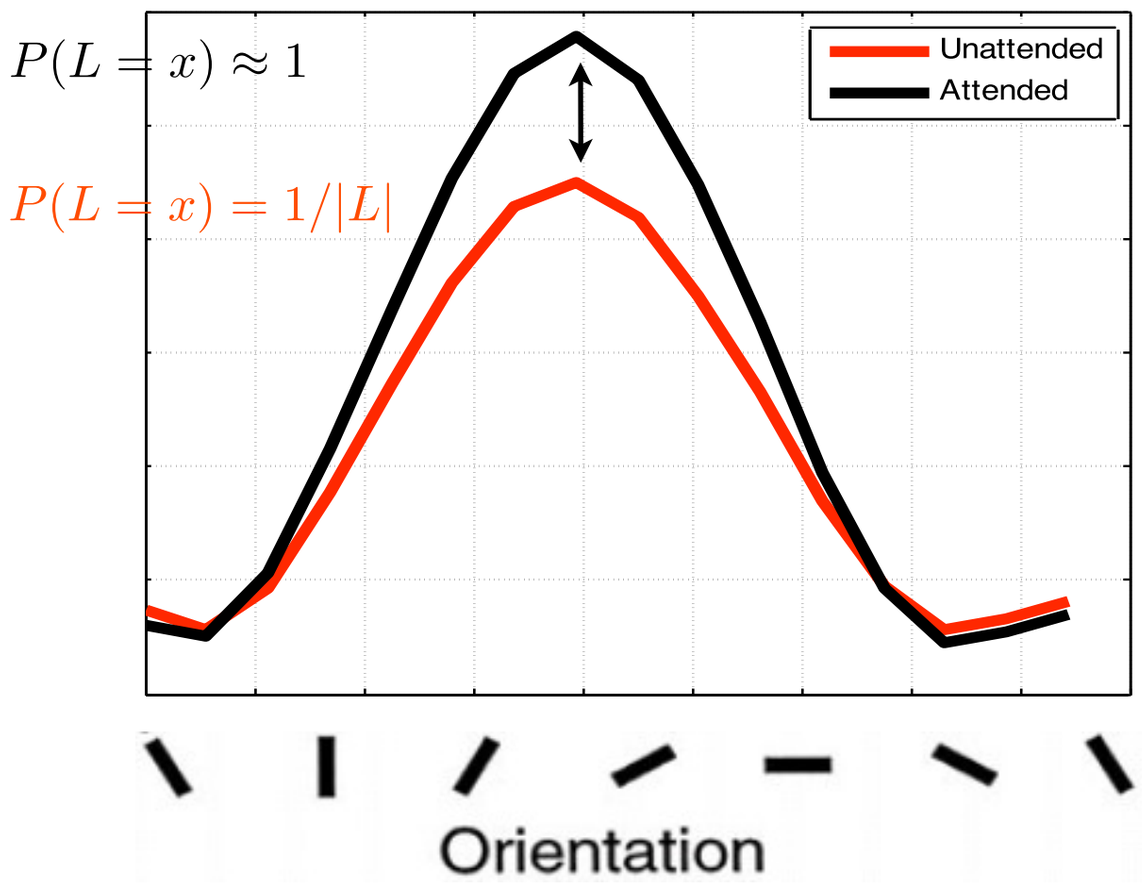
Bayesian inference and attention

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

McAdams and Maunsell ‘99



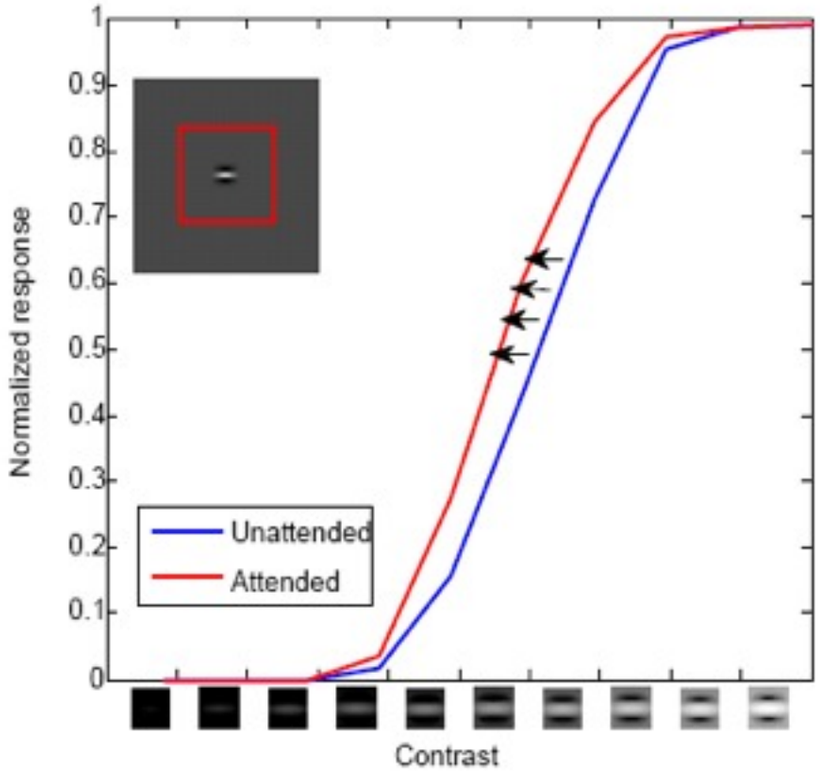
Model



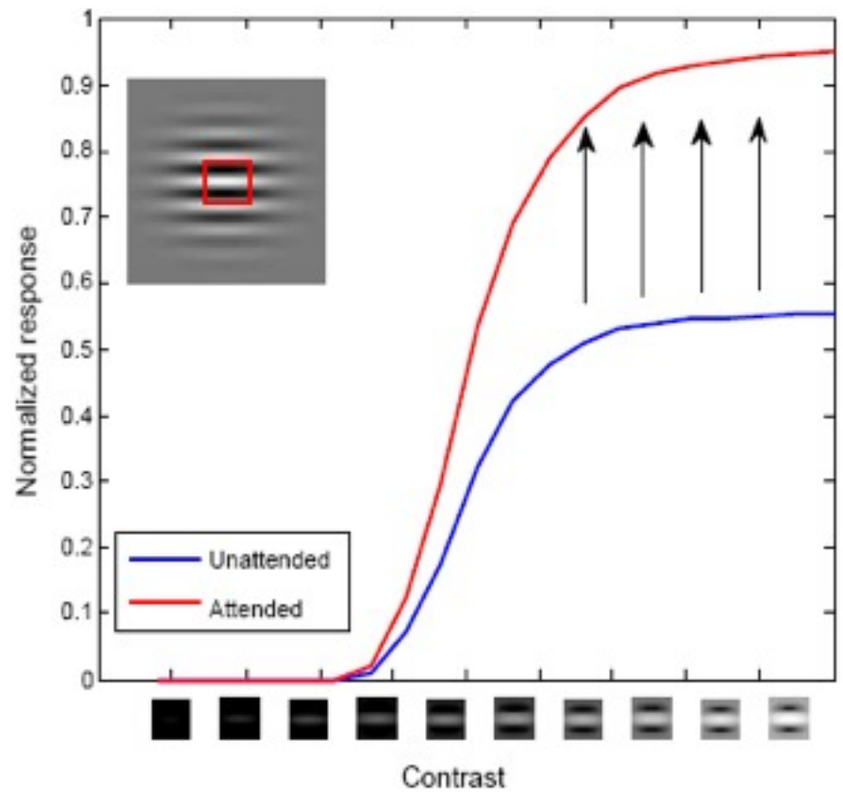
Multiplicative scaling of tuning curves by spatial attention

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

Trujillo and Treue ‘02



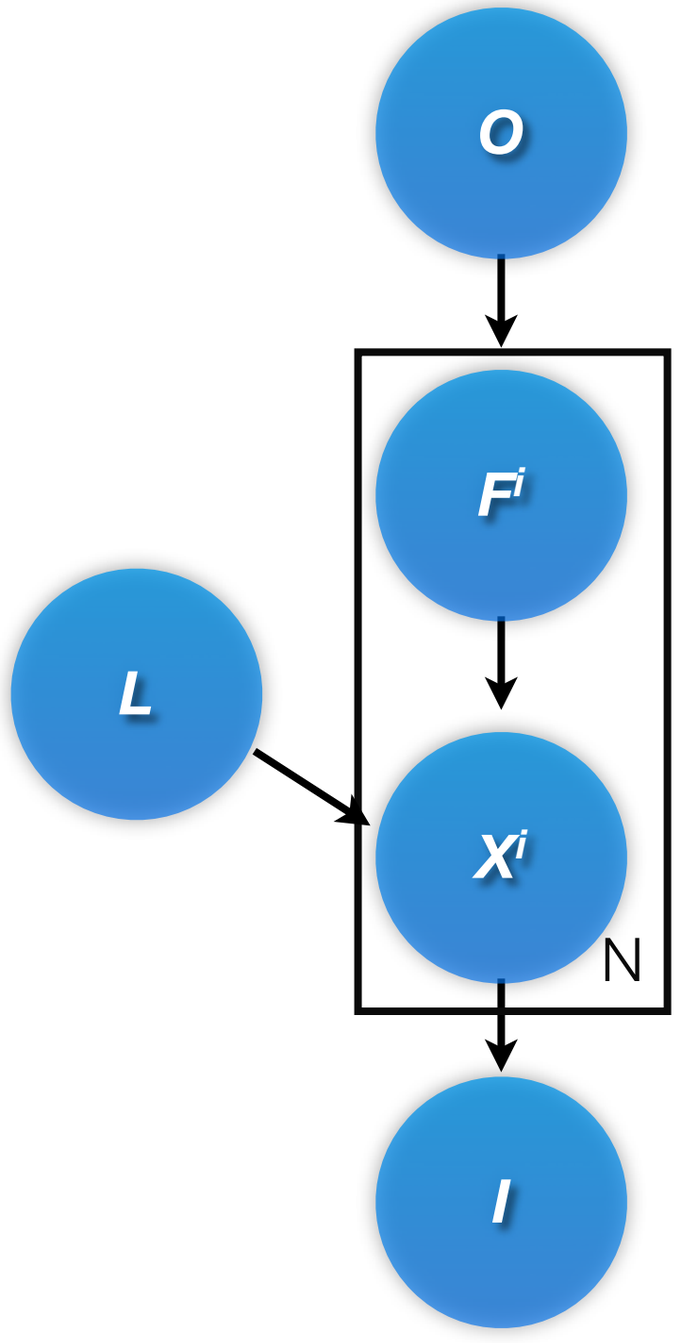
Mc Adams and Maunsell ’99



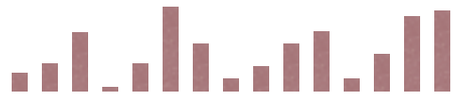
Contrast vs. response gain

Predicted by Reynolds & Heeger ’09

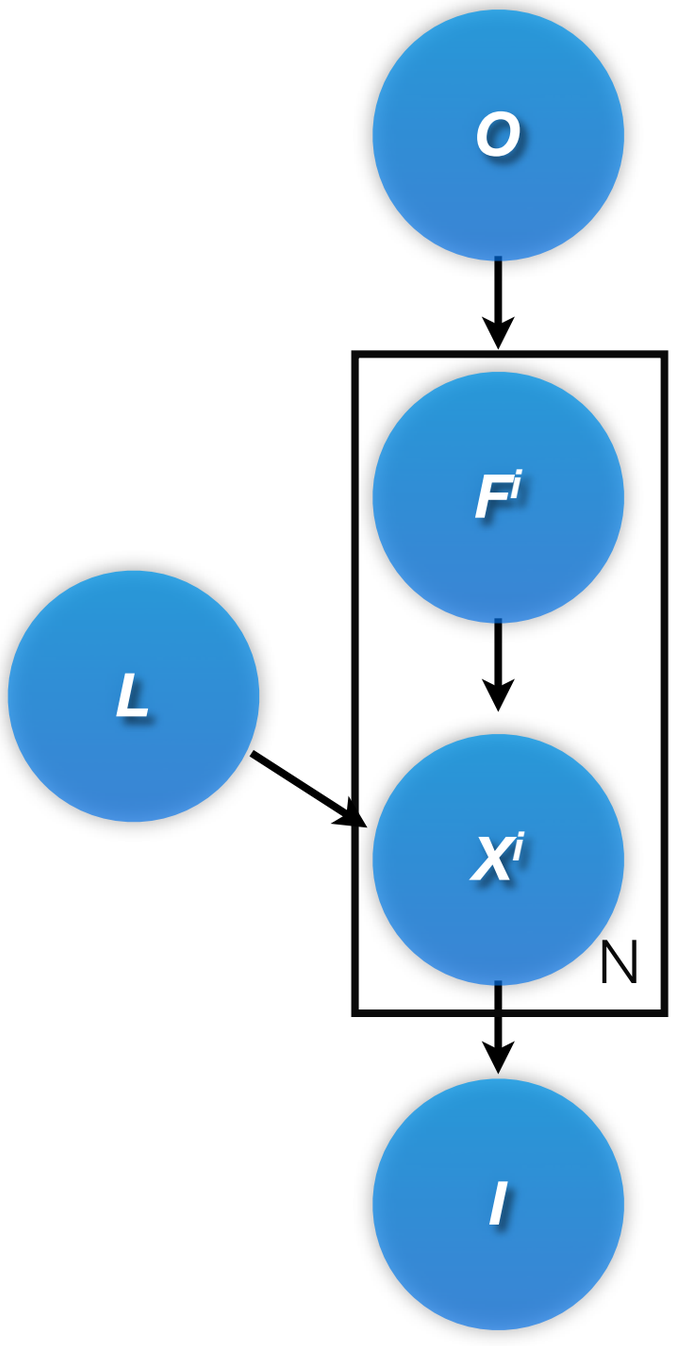
Learning to localize cars and pedestrians in street scenes



Learning to localize cars and pedestrians in street scenes

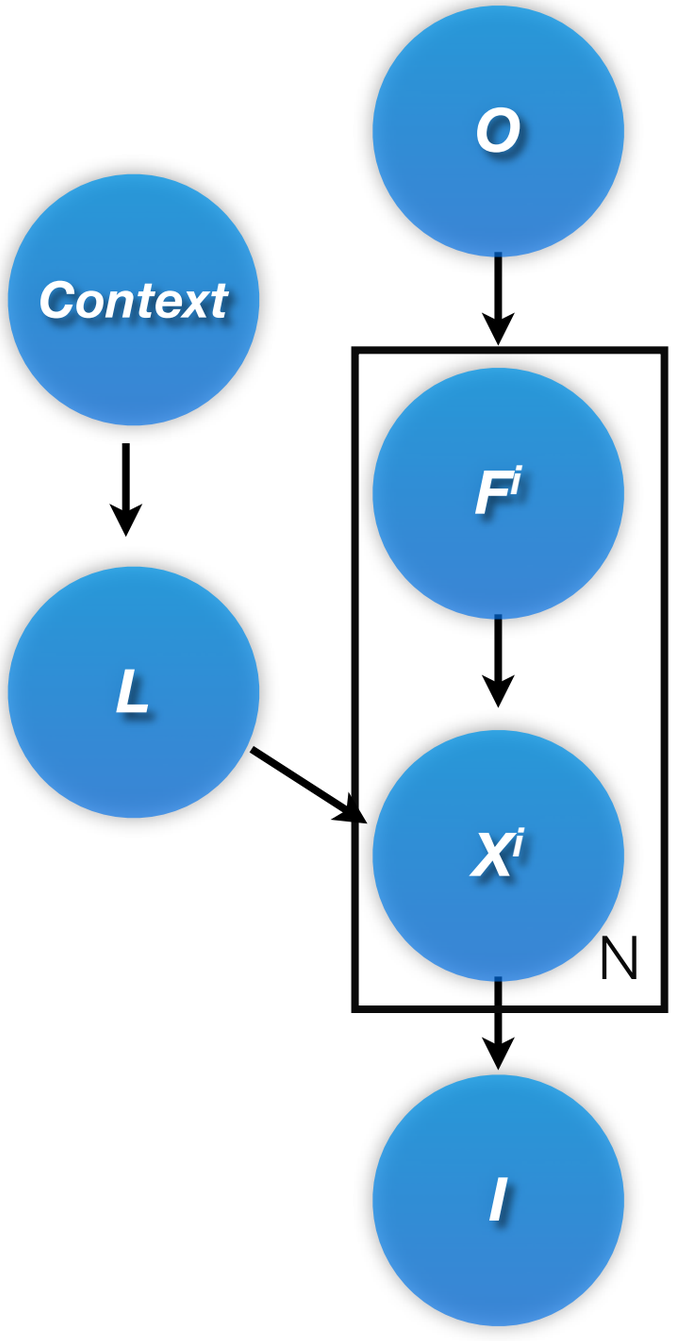


learning object priors



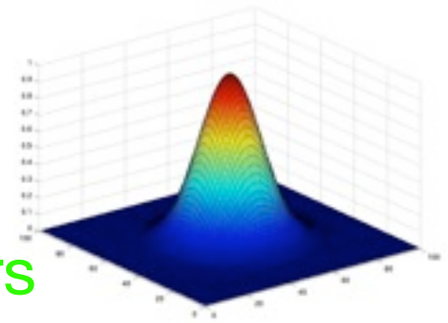
Learning to localize cars and pedestrians in street scenes

learning object priors

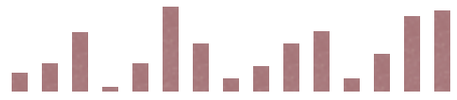


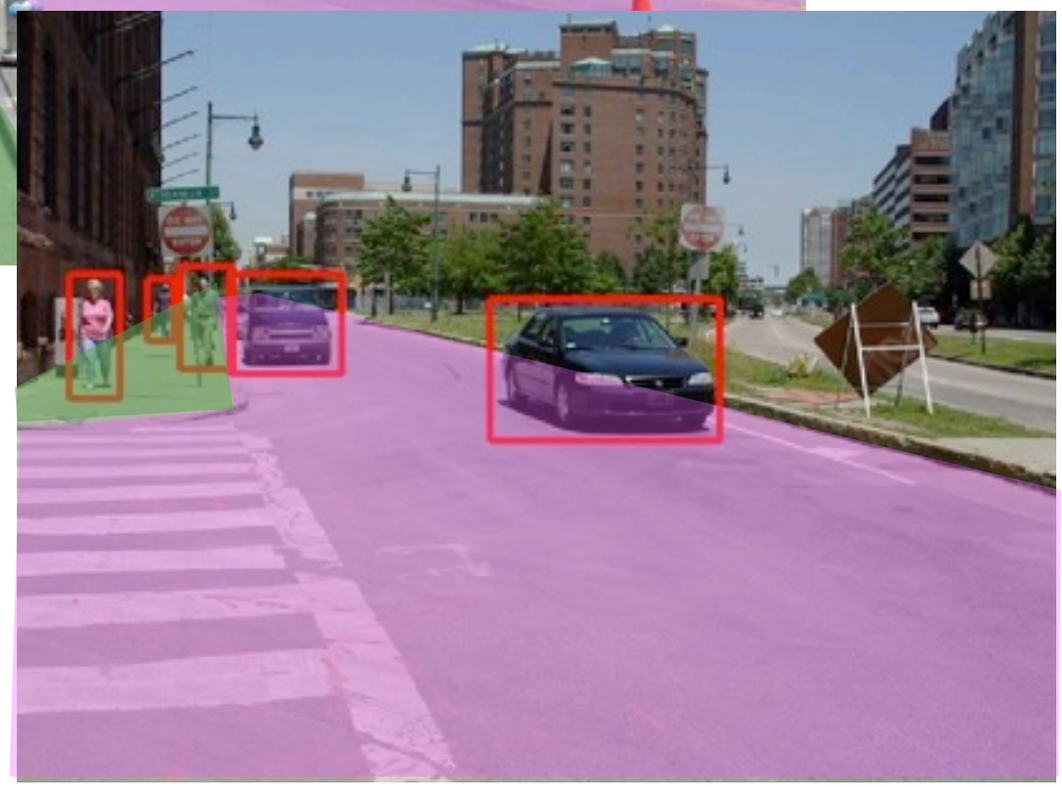
Learning to localize cars and pedestrians in street scenes

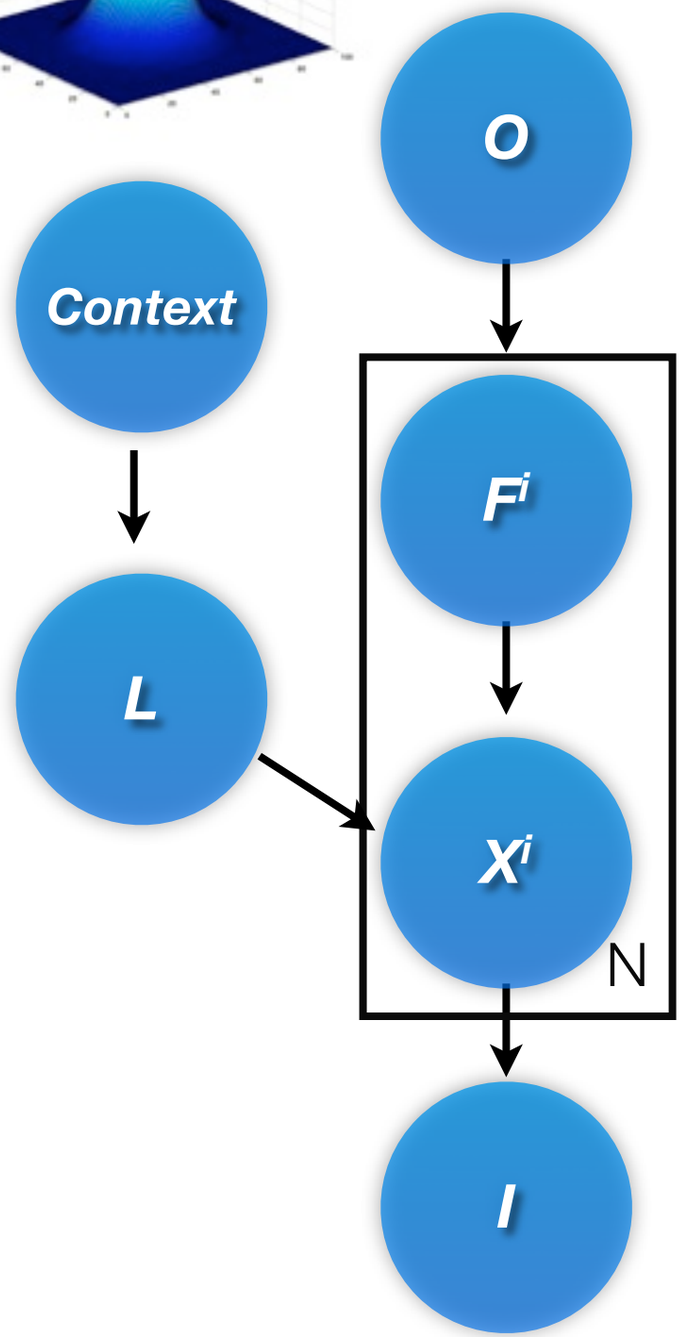
learning location priors
(global contextual cues)



learning object priors



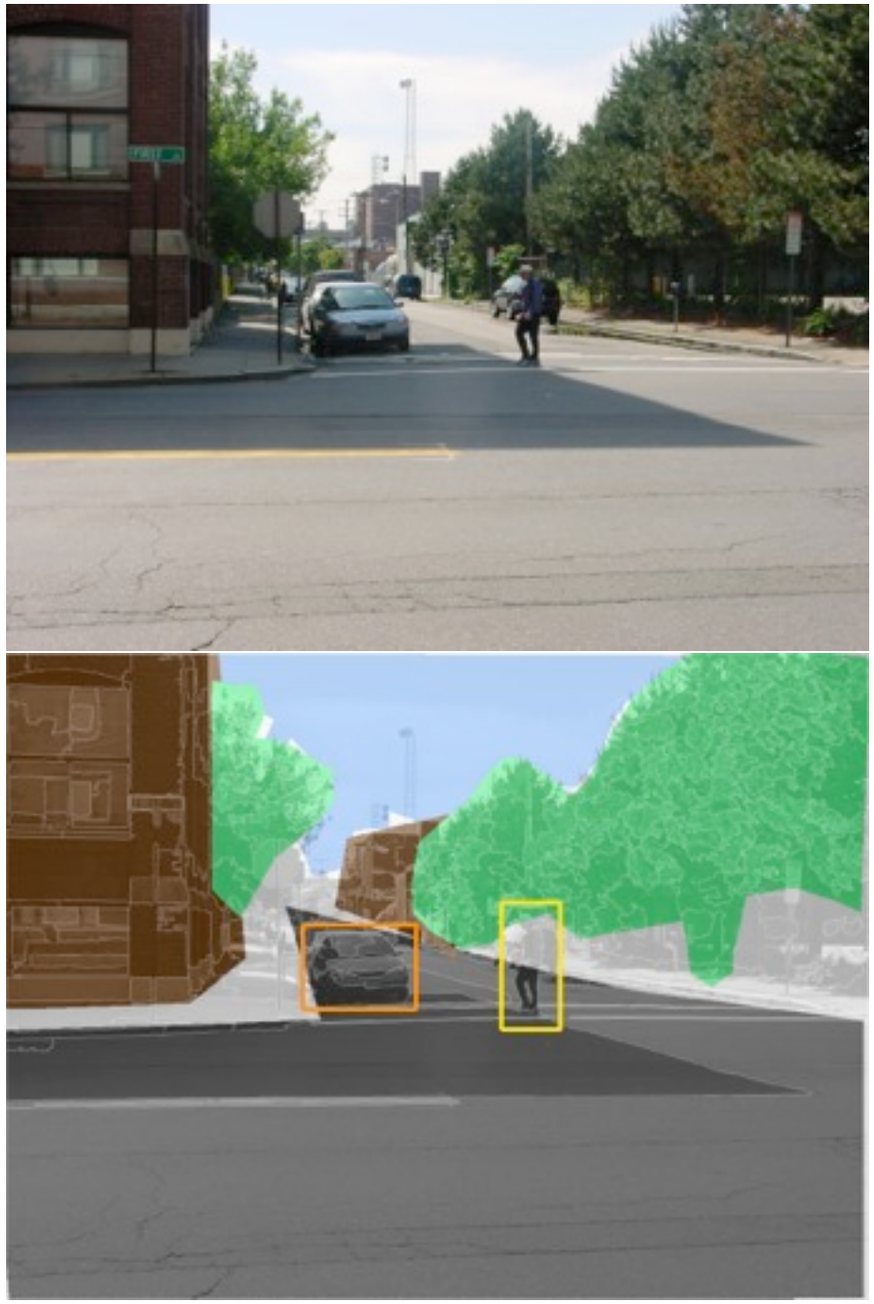




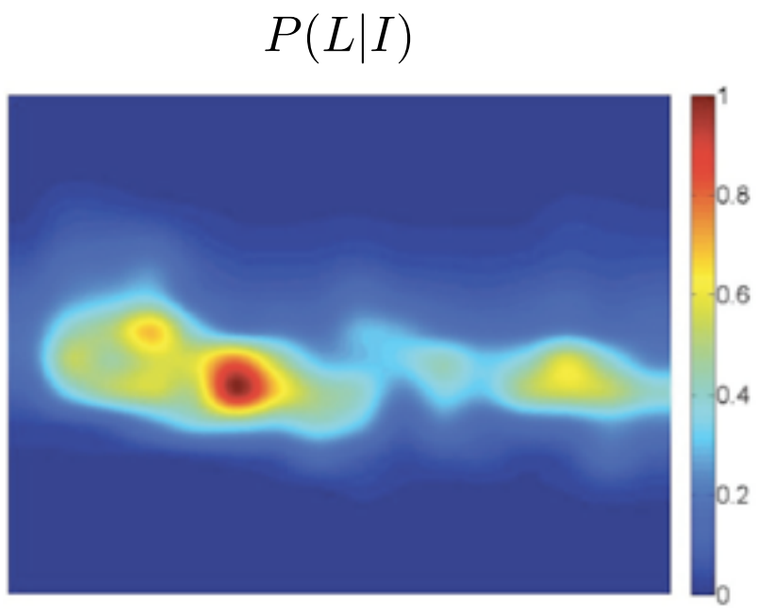
(see Torralba et al)

The experiment

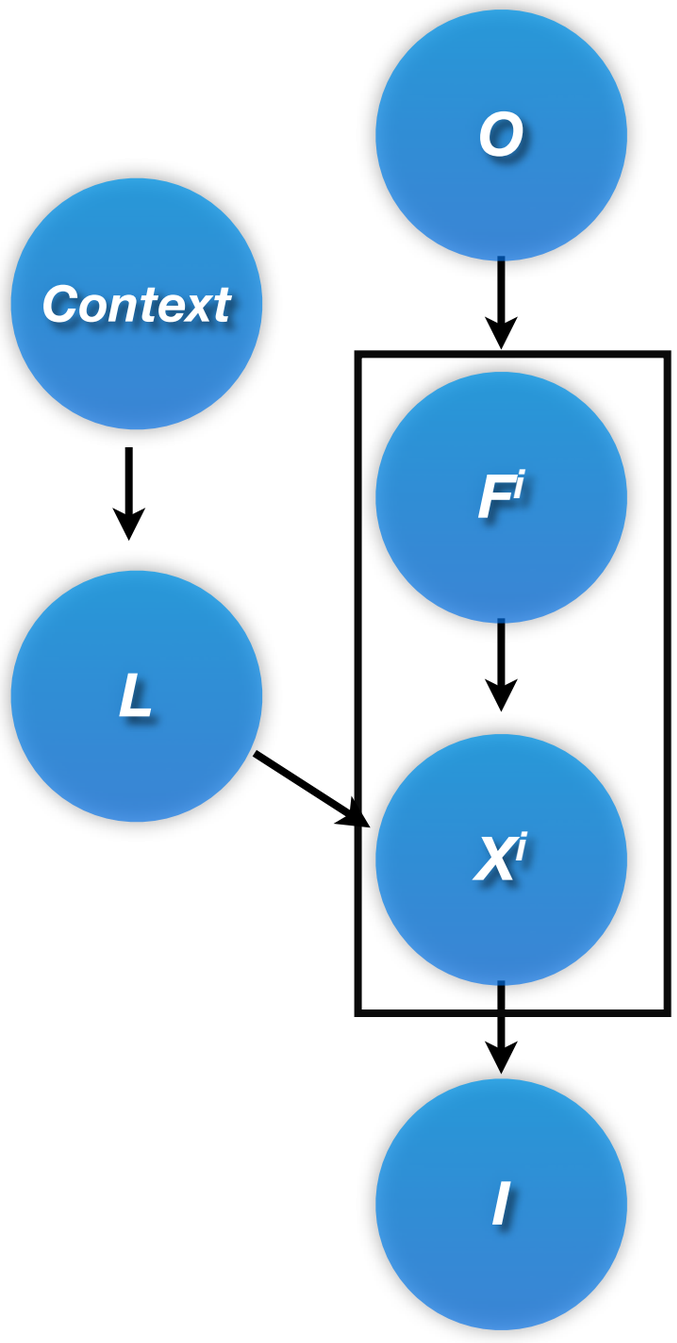
- Dataset:
  - 100 street-scenes images with cars & pedestrians and 20 without
- Experiment
  - 8 participants asked to count the number of cars/ pedestrians
  - Blocks/randomized presentations
  - Each image presented twice
- Eye movements recorded using an infra-red eye tracker
- Eye movements as proxy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



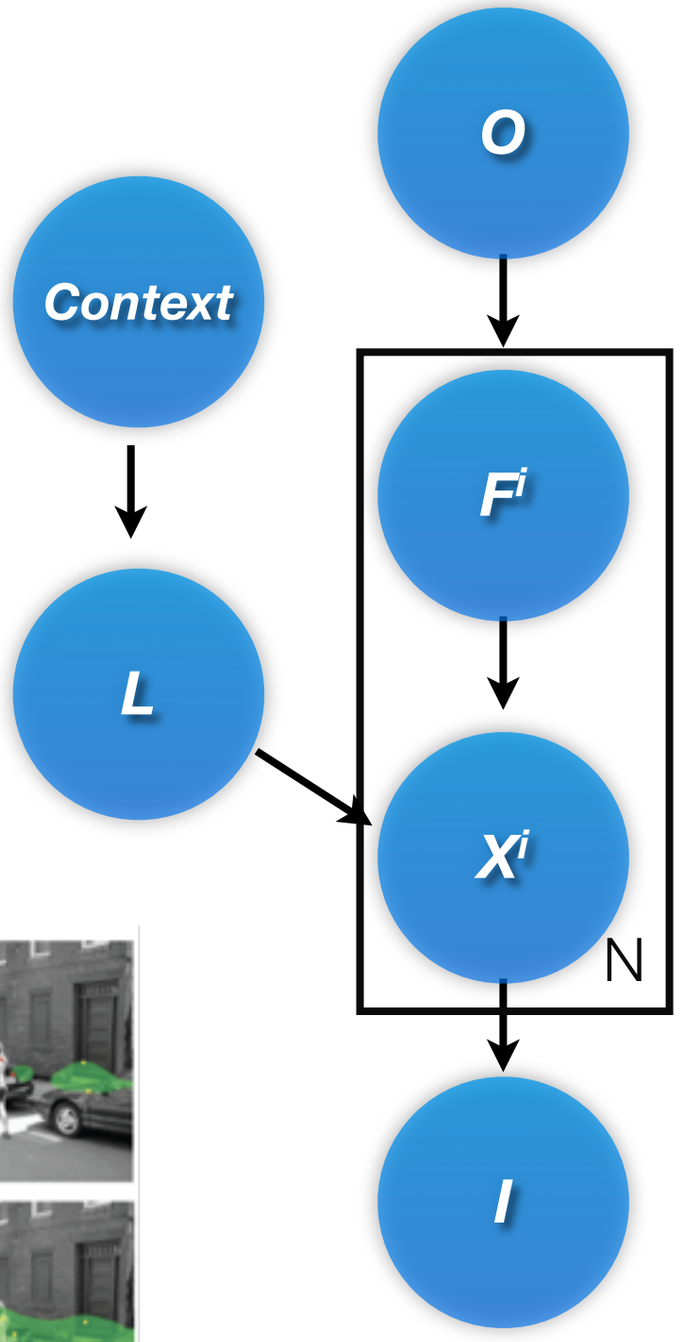
Predicting eye movements during searches for cars and pedestrians

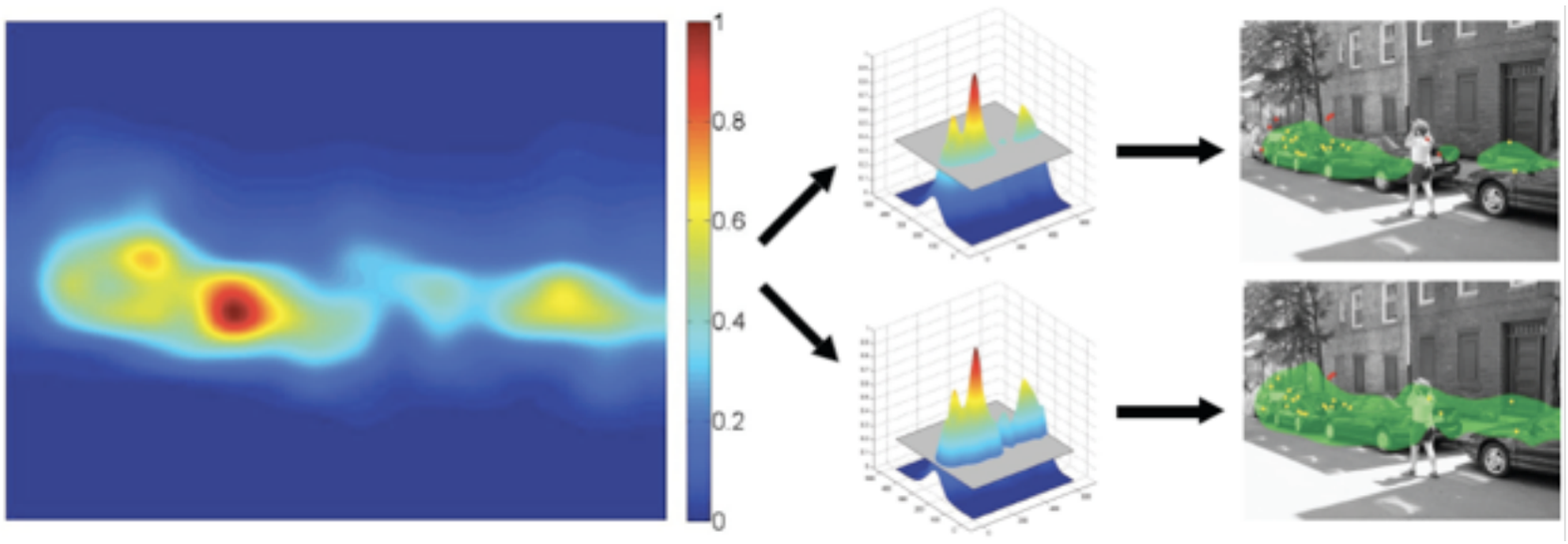


Saliency map

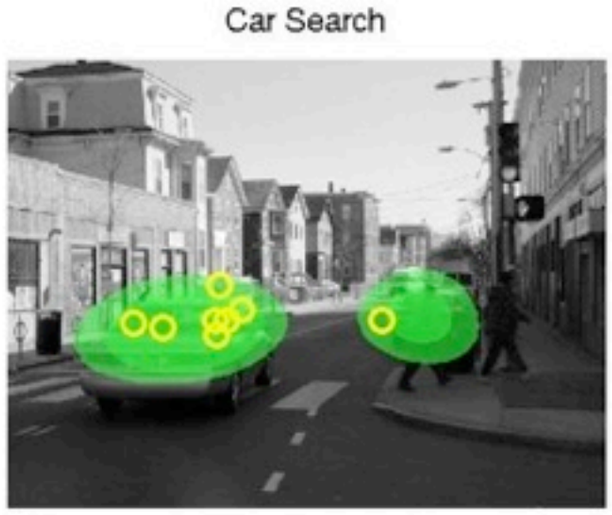


Predicting eye movements during searches for cars and pedestrians



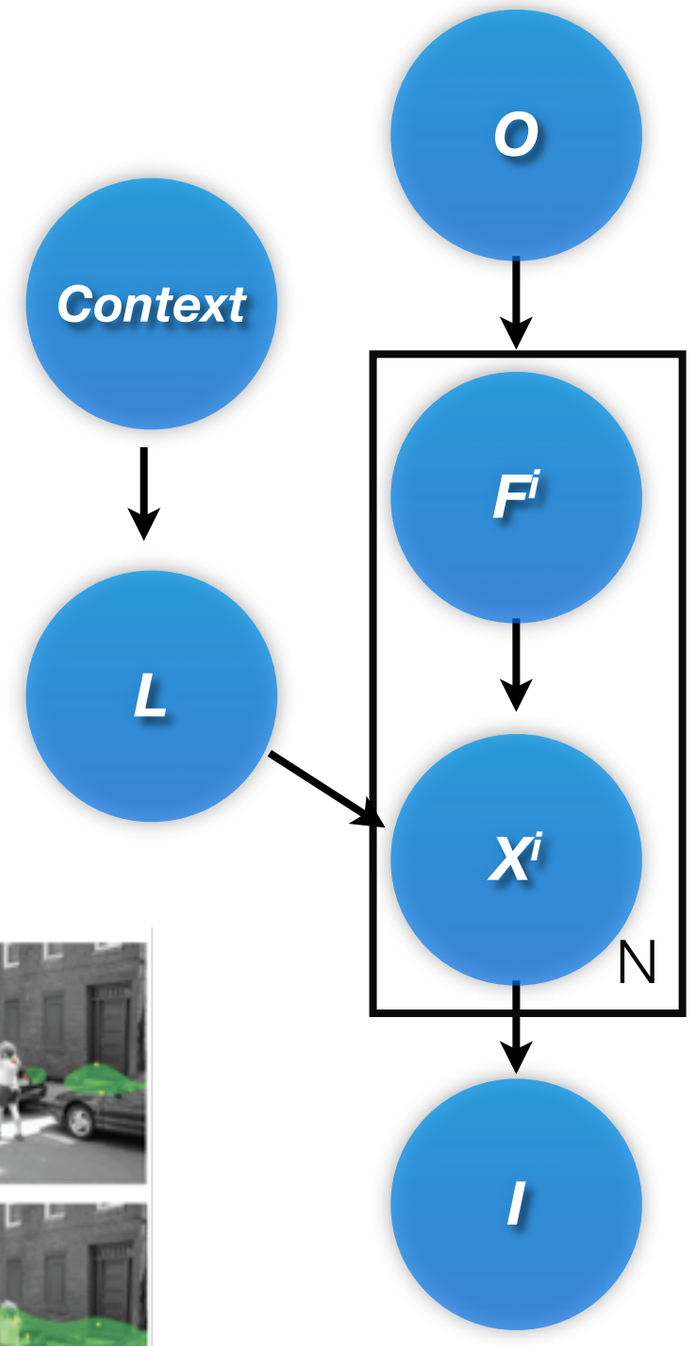


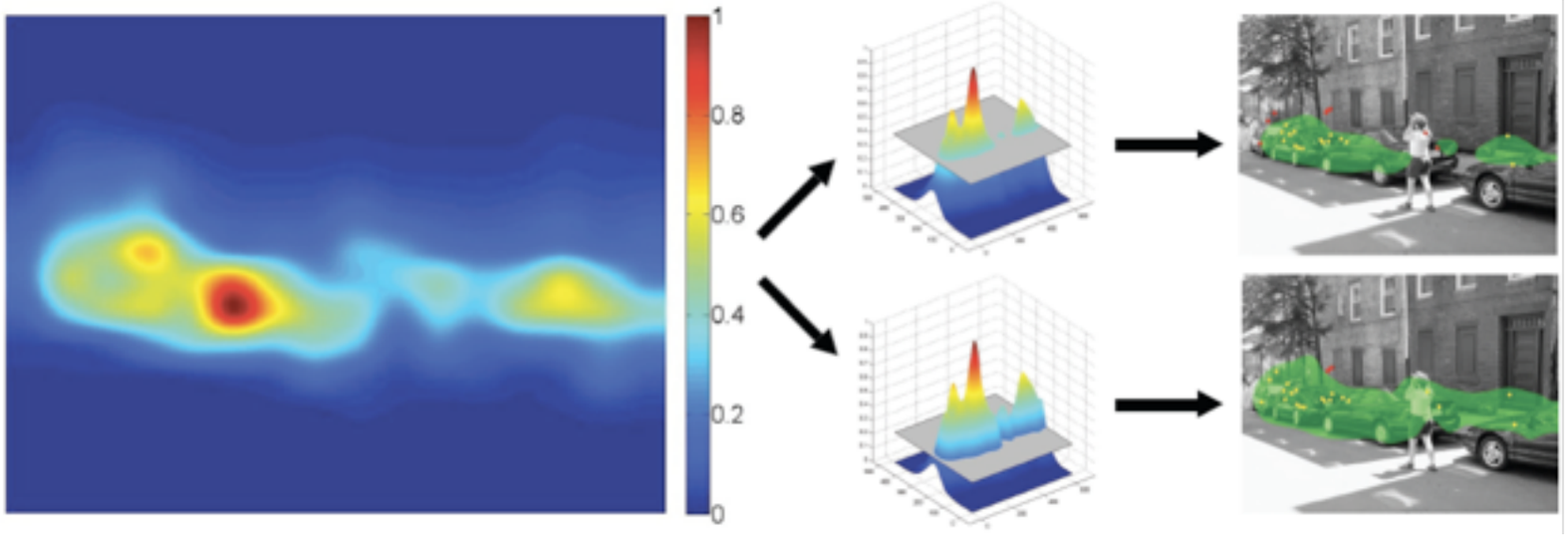

Car Search


Car Search

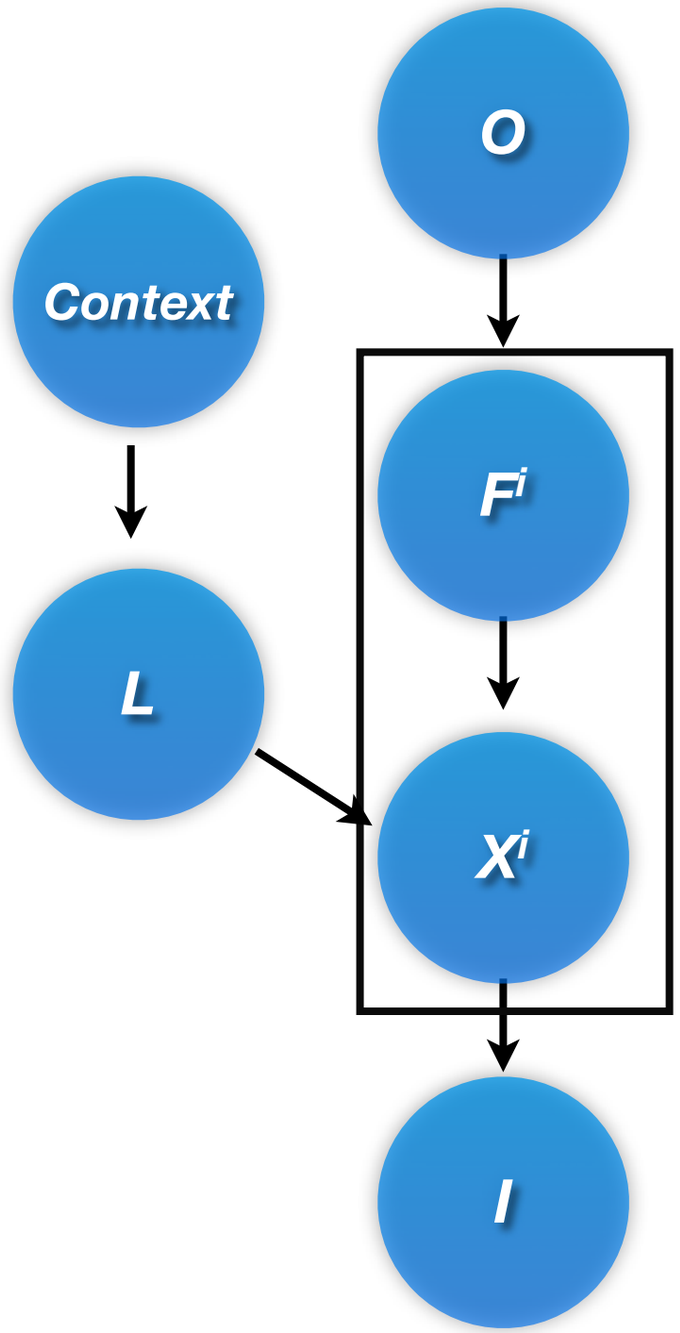

Pedestrian Search


Pedestrian Search


O
Context
L
F^i
X^i
N
I


1
0.8
0.6
0.4
0.2
0

Predicting eye movements during searches for cars and pedestrians

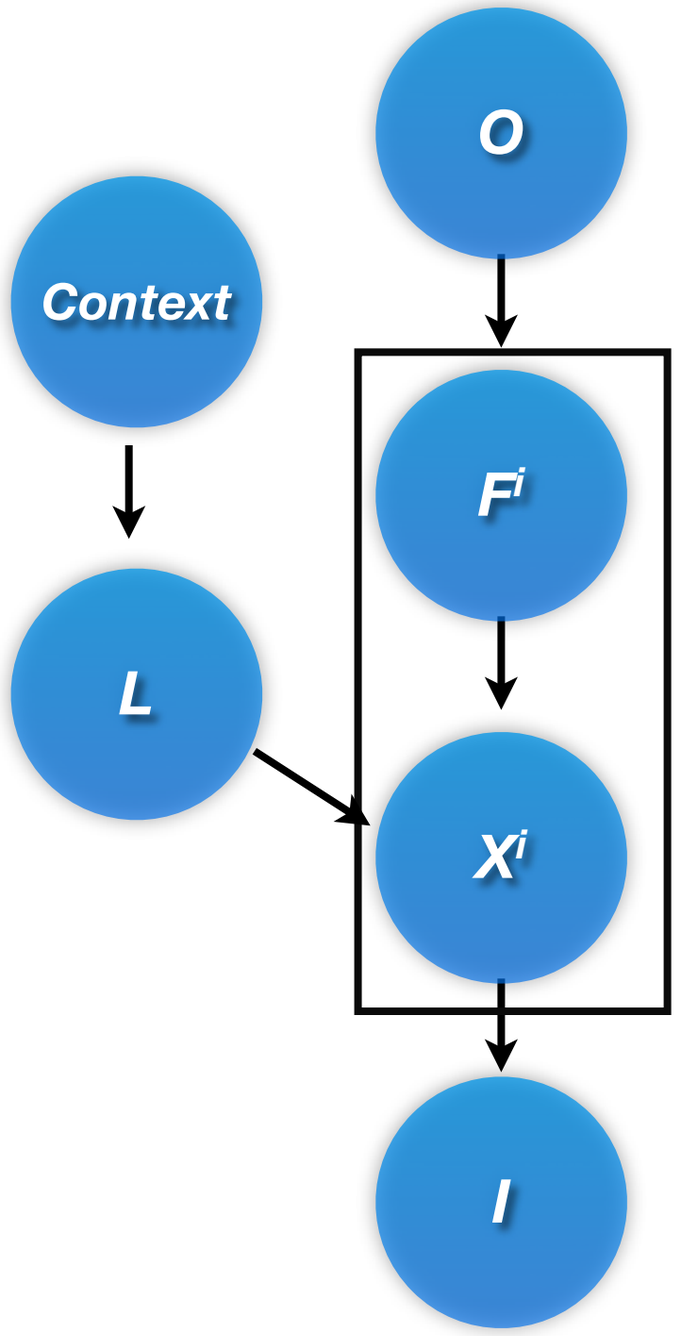


- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during searches for cars and pedestrians

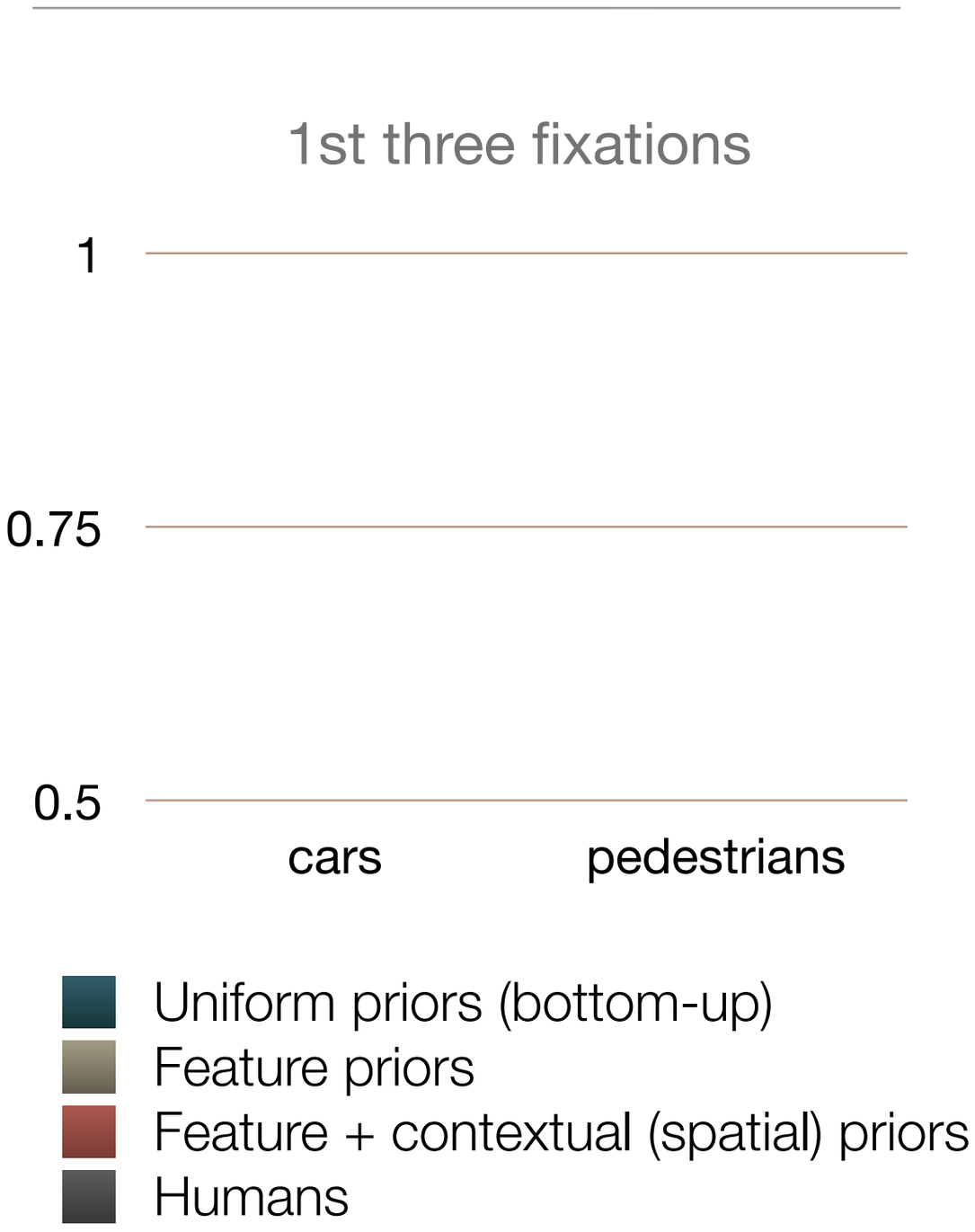
1st three fixations

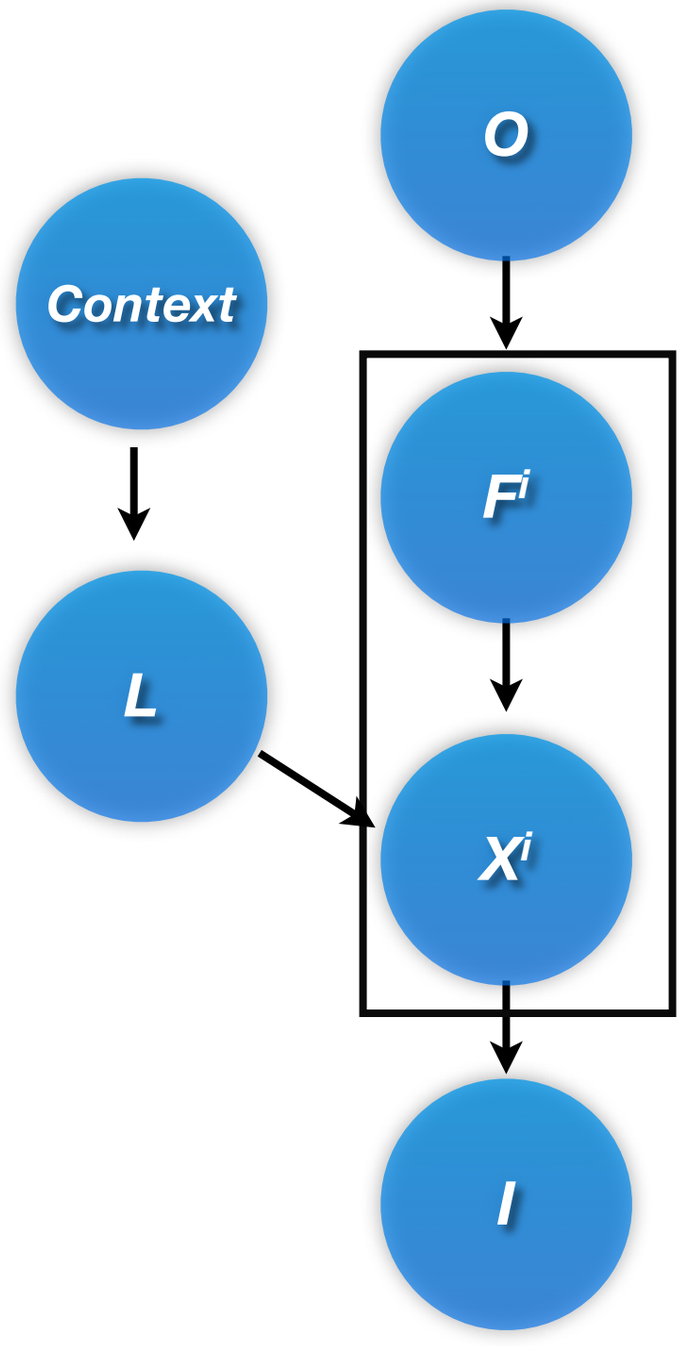


- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

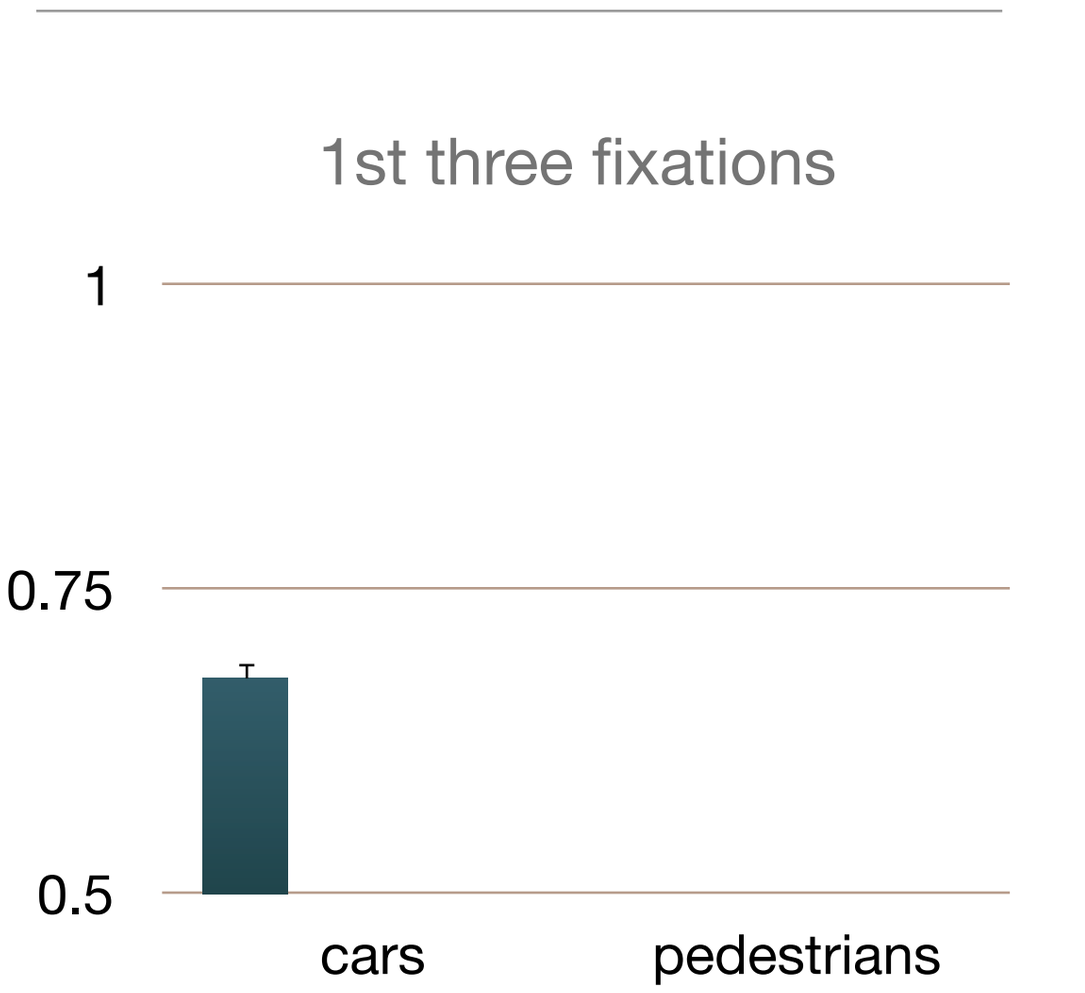
Predicting eye movements during searches for cars and pedestrians



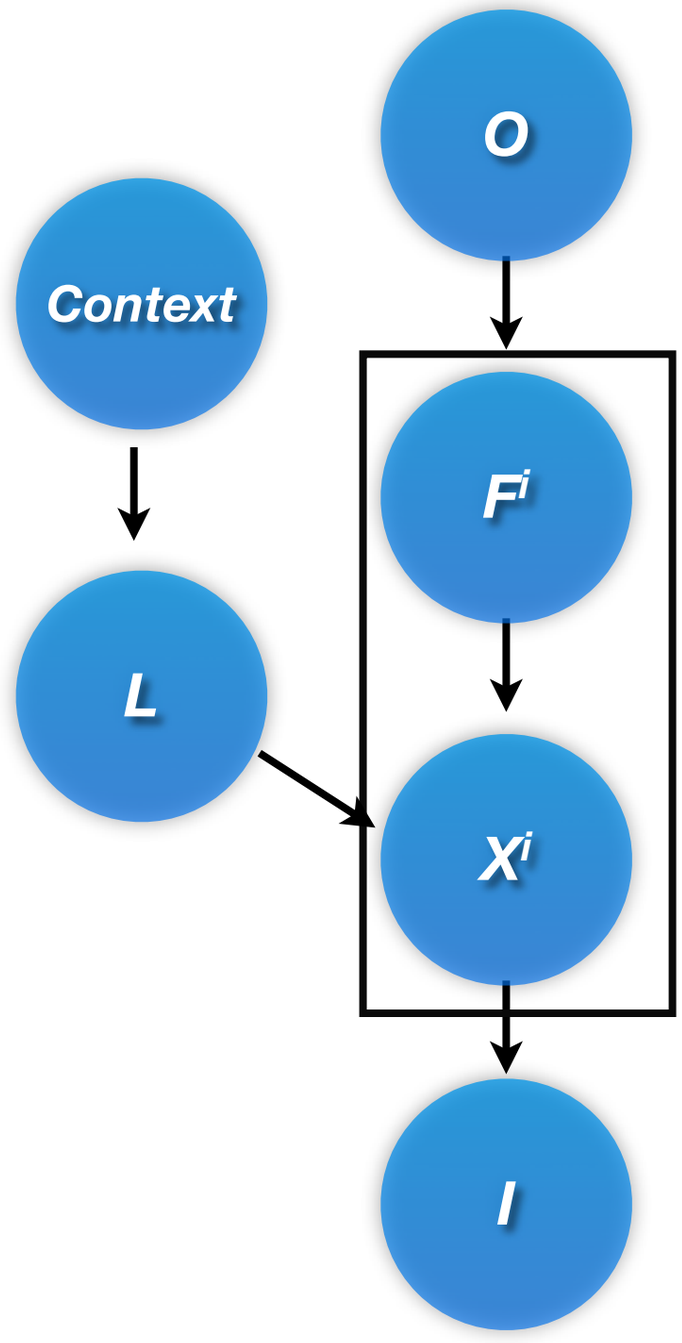


*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during searches for cars and pedestrians

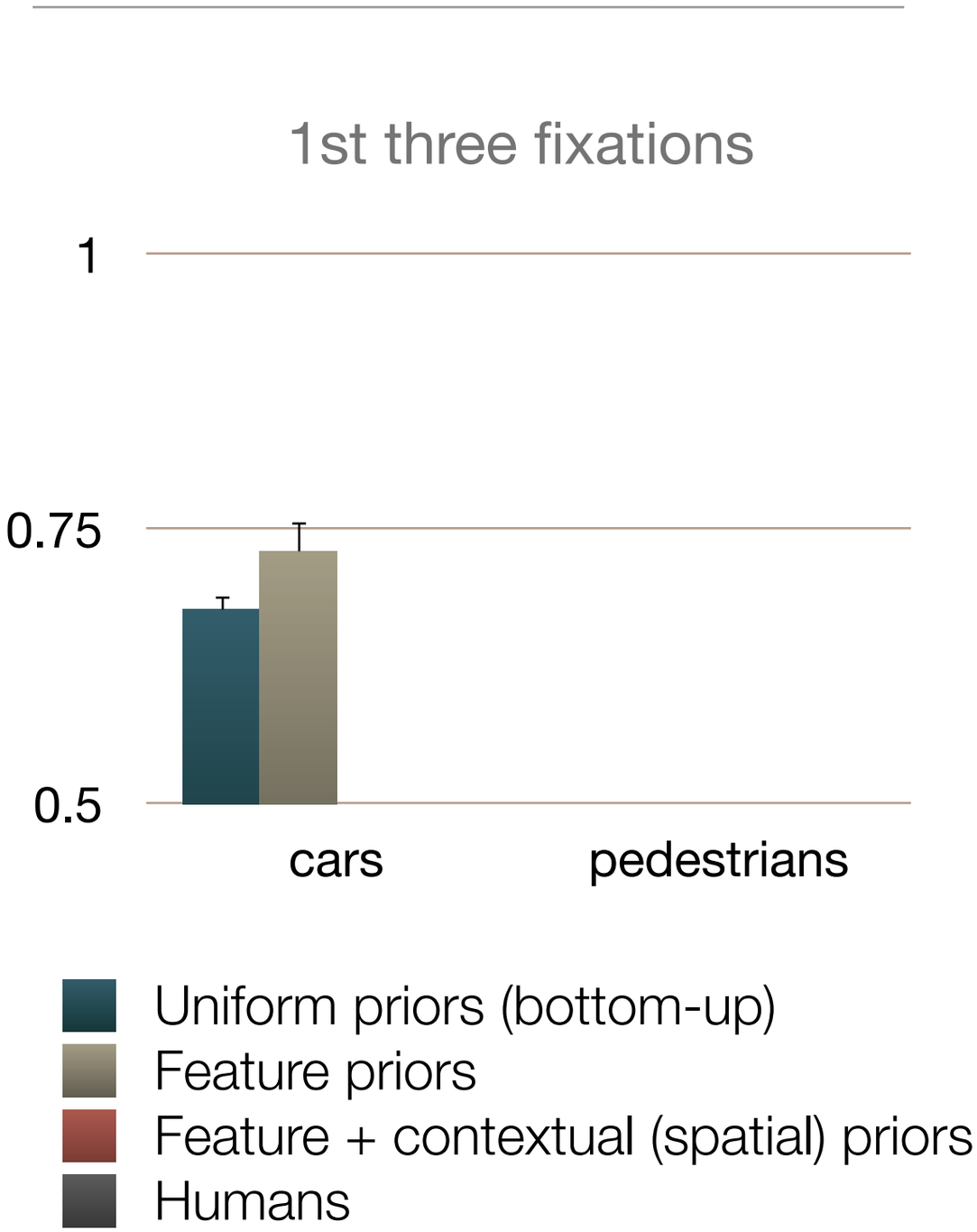


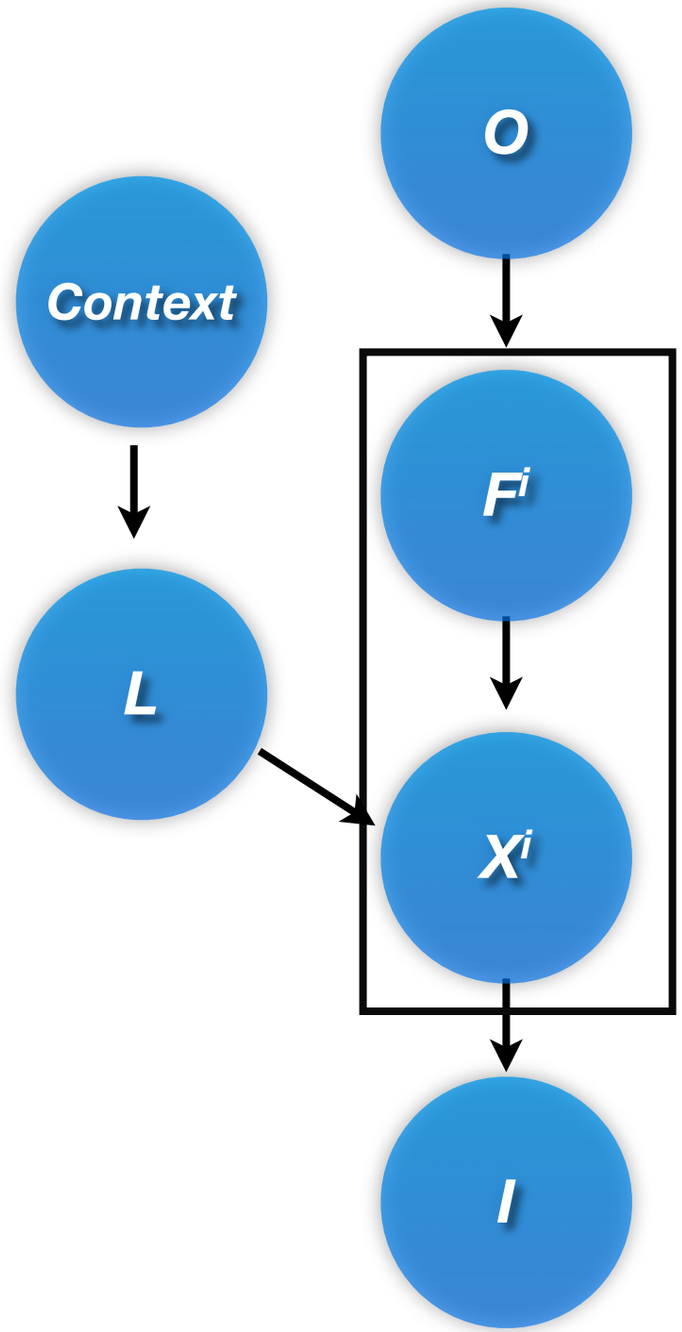
- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans



*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

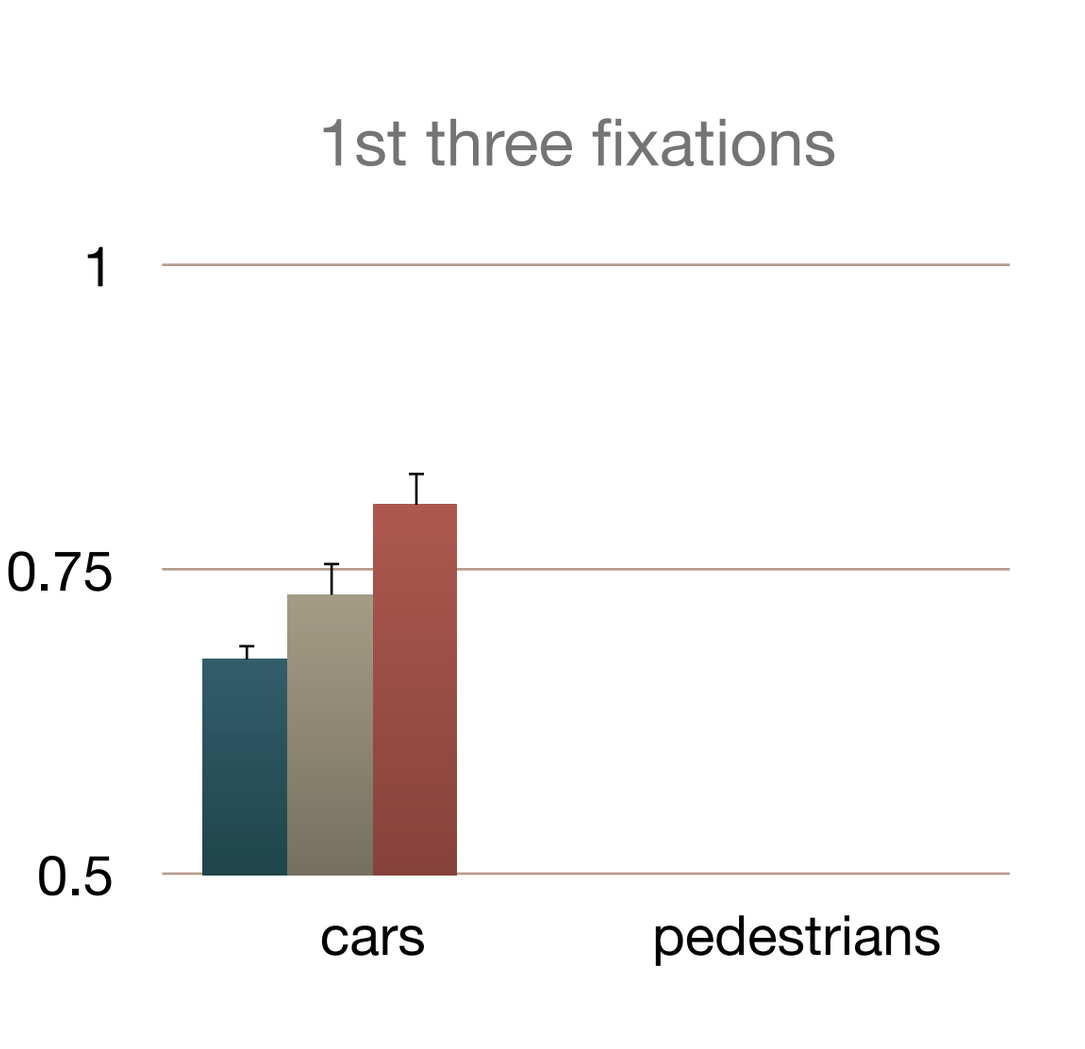
Predicting eye movements during searches for cars and pedestrians

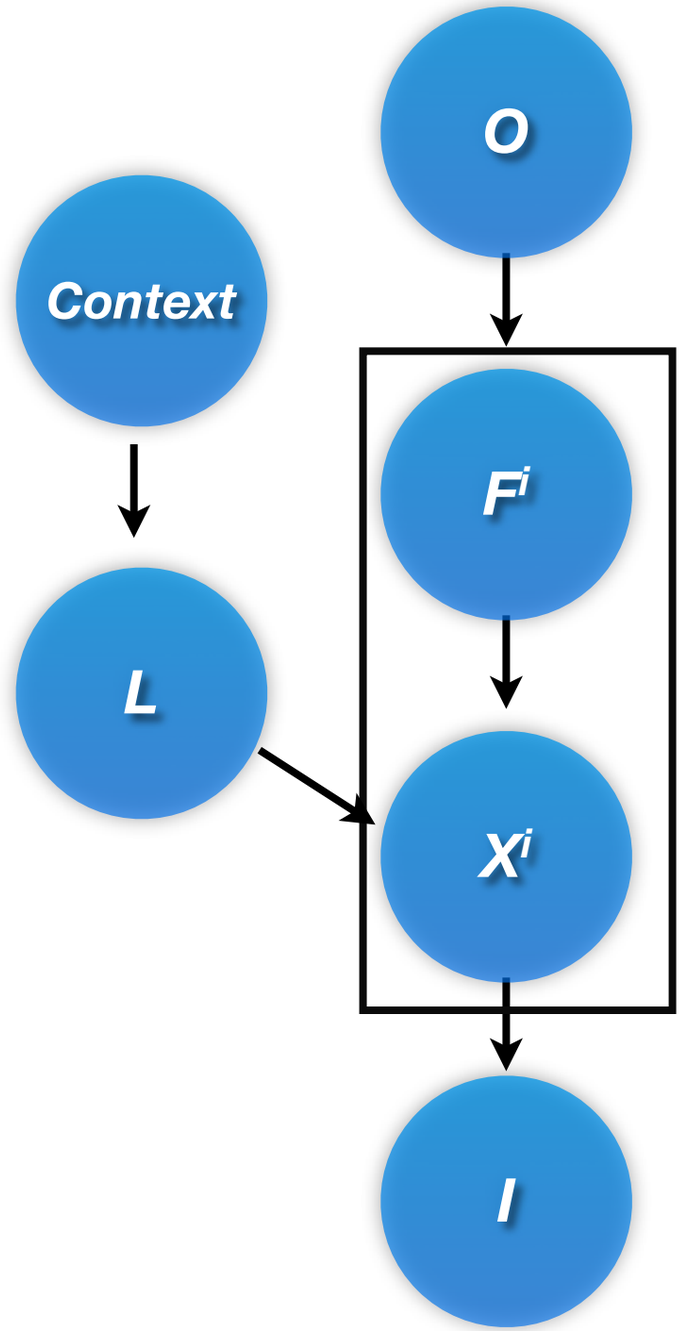




*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

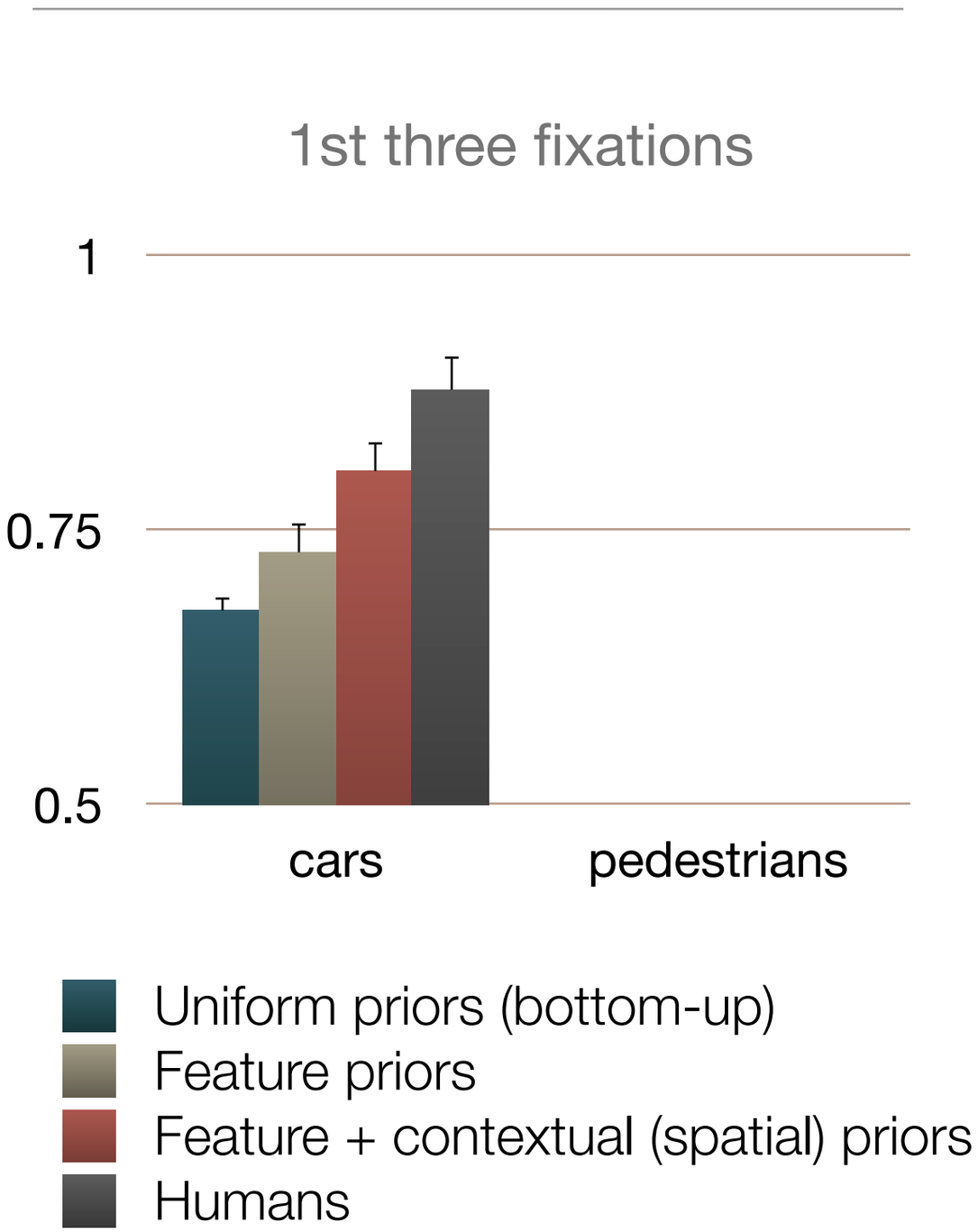
Predicting eye movements during searches for cars and pedestrians

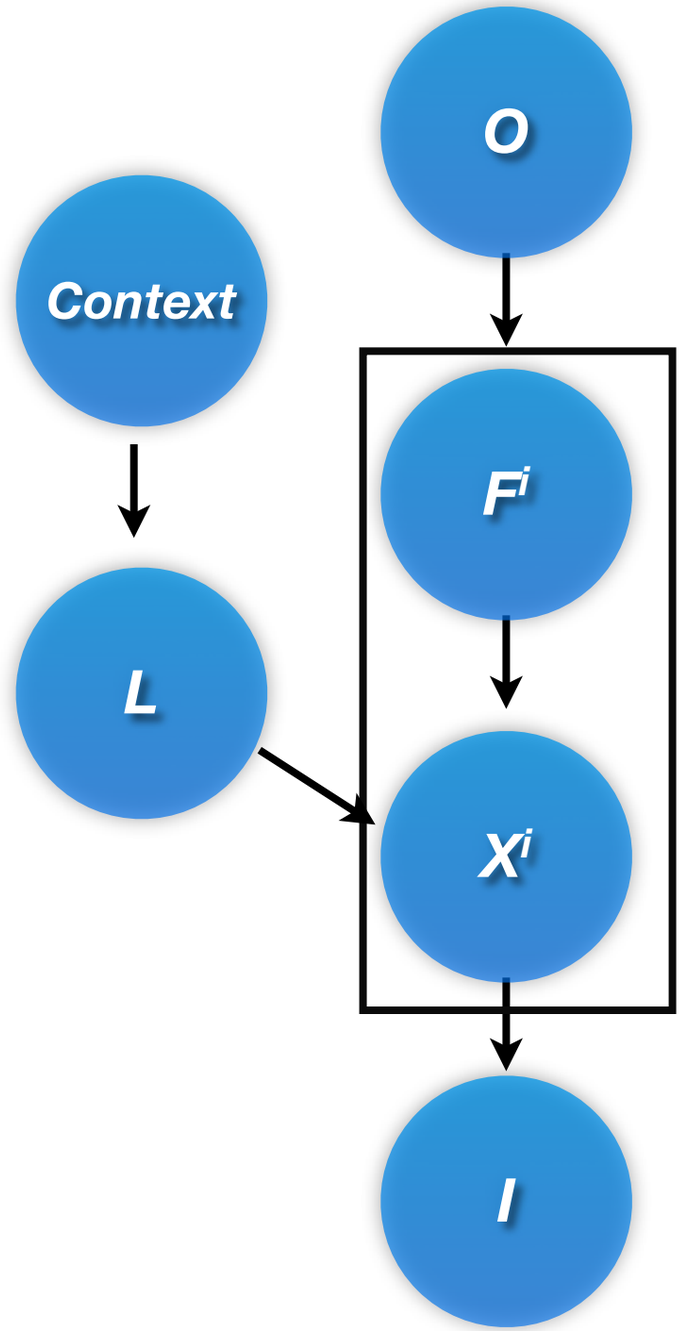




*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

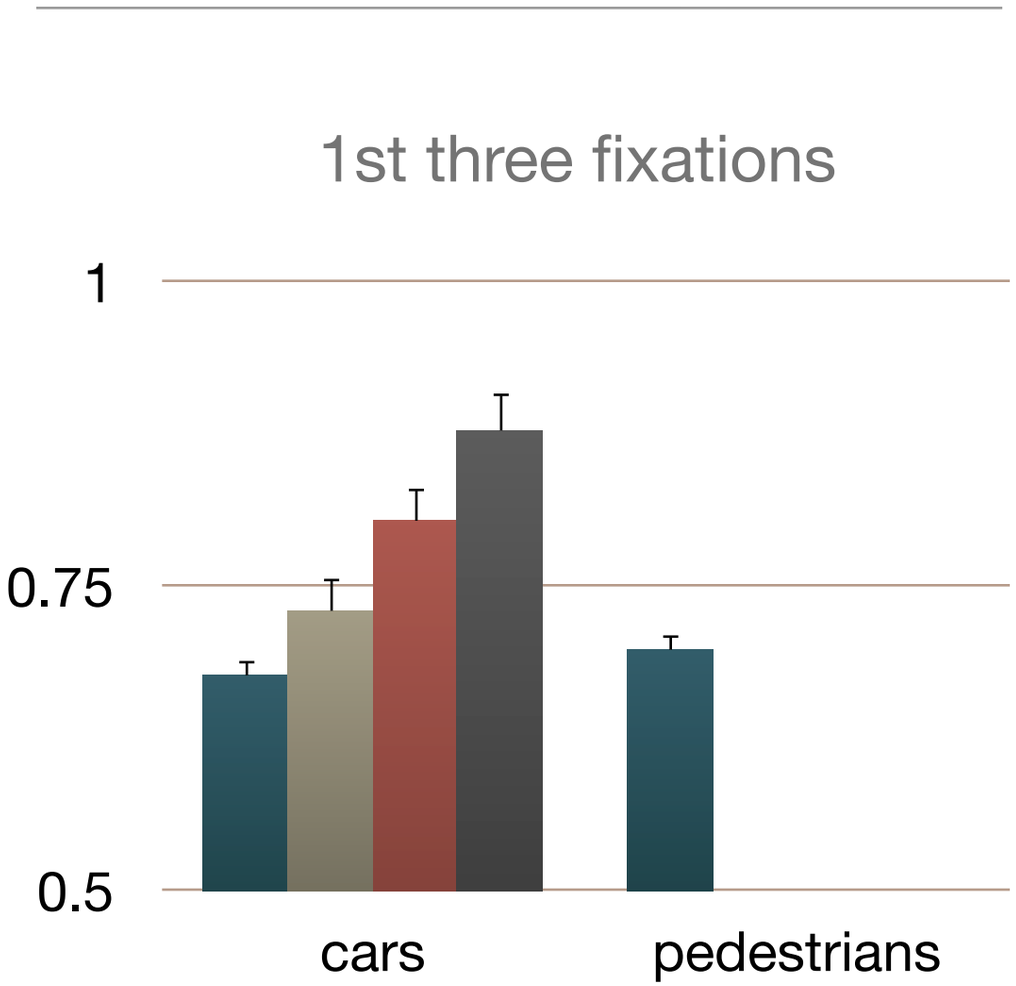
Predicting eye movements during searches for cars and pedestrians



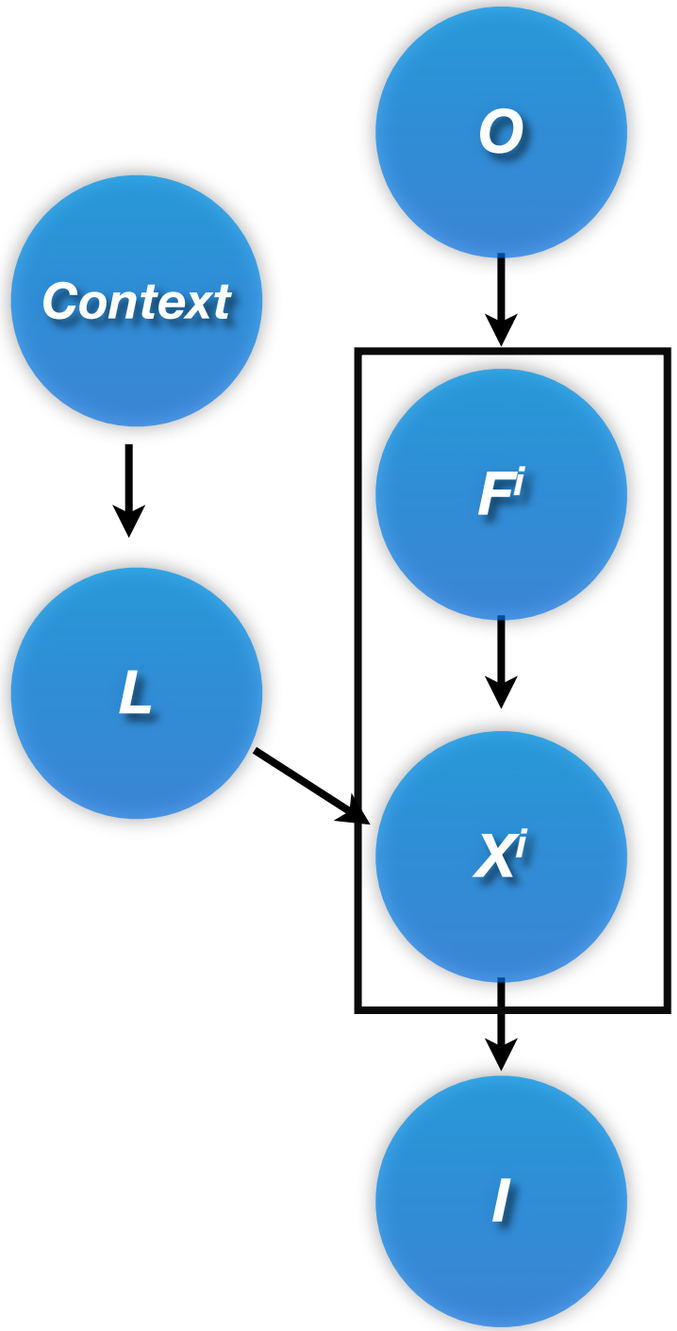


*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during searches for cars and pedestrians

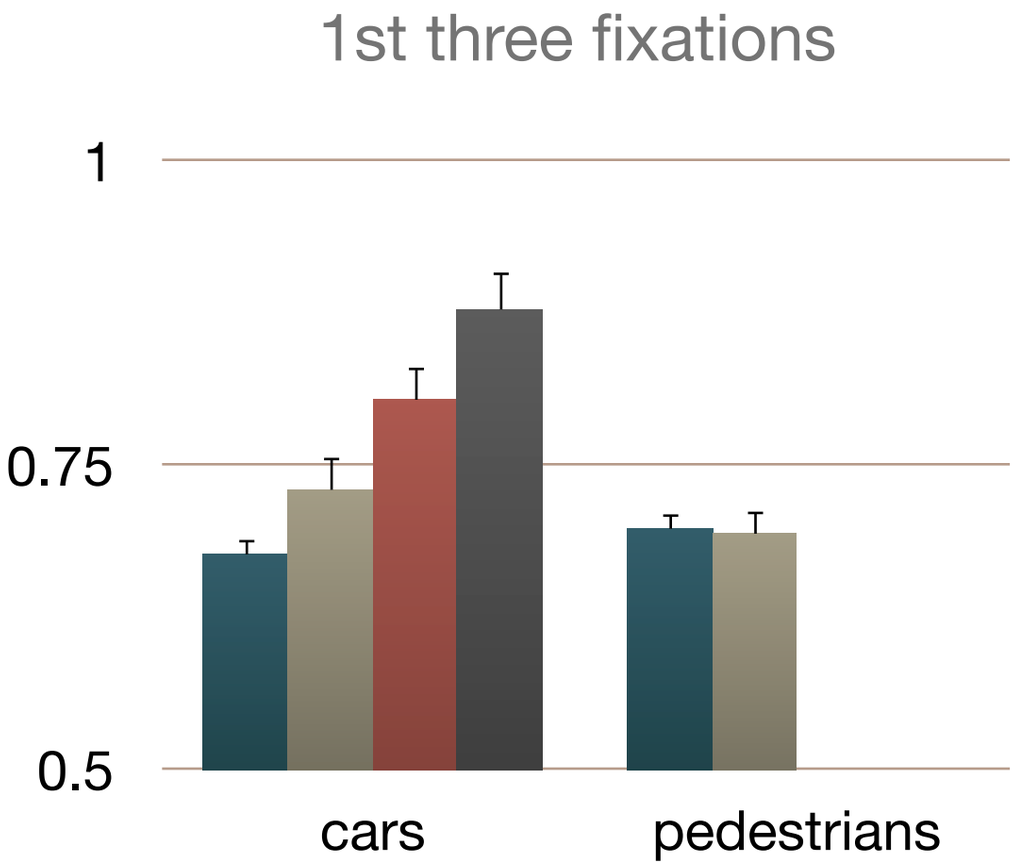


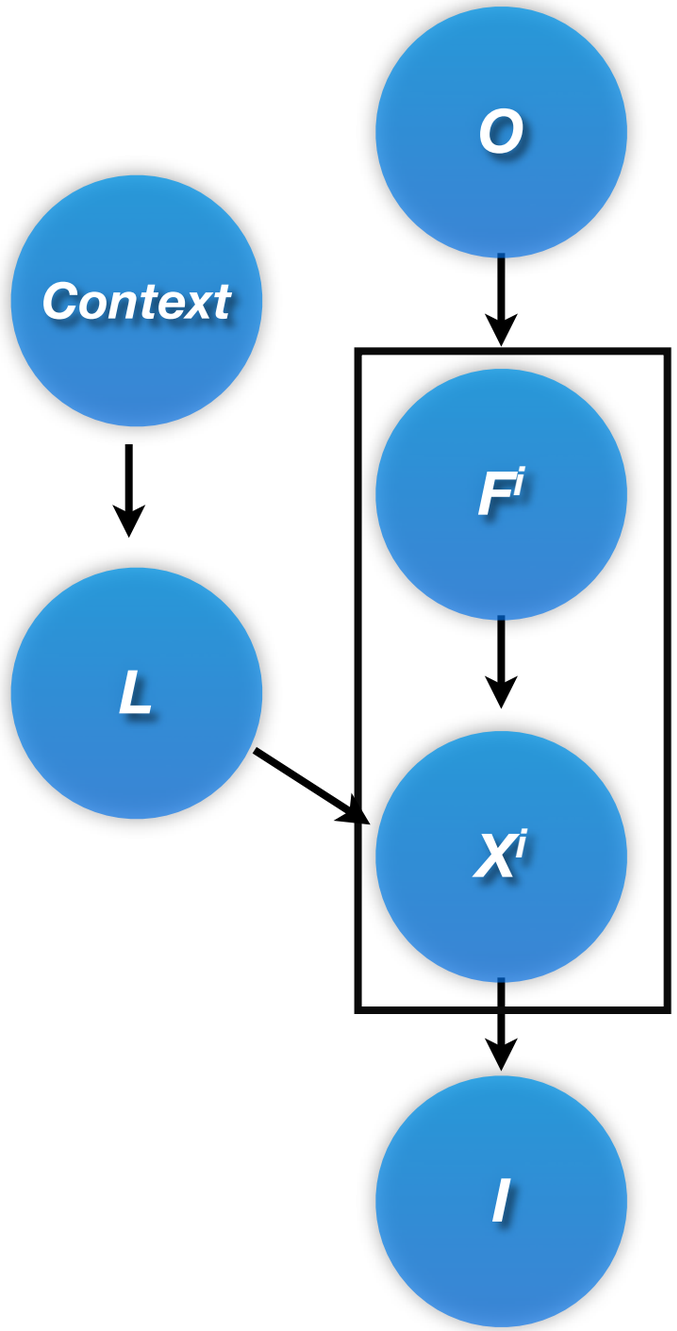
- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans



*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

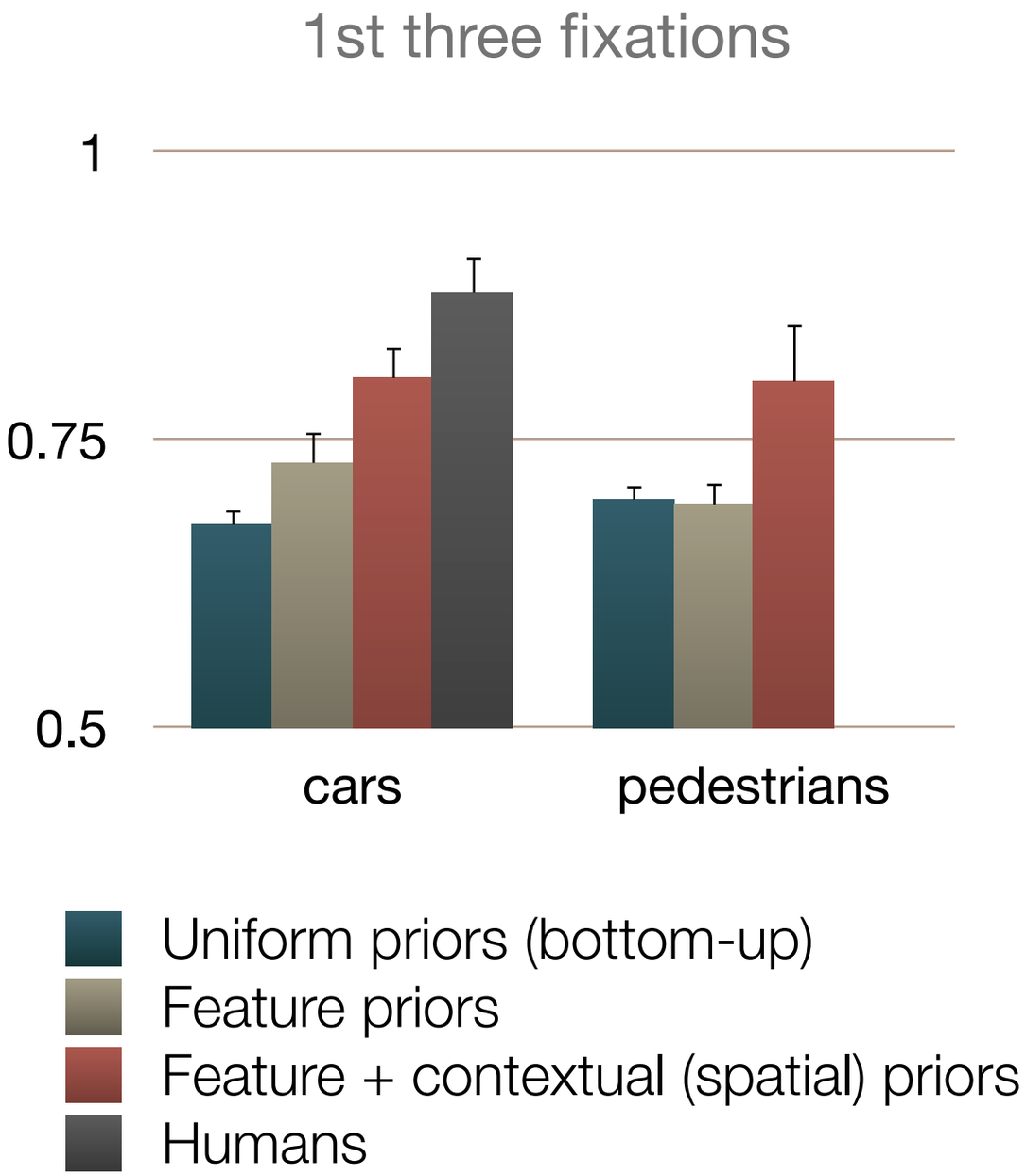
Predicting eye movements during searches for cars and pedestrians

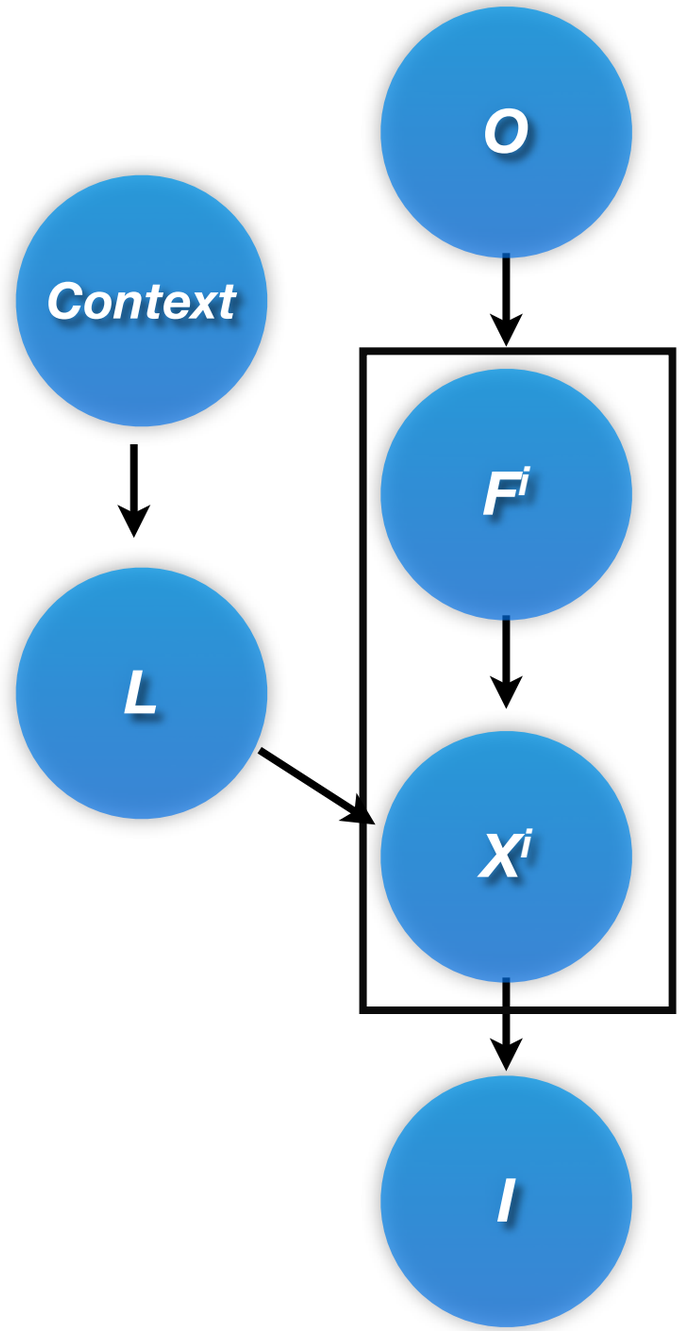




*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

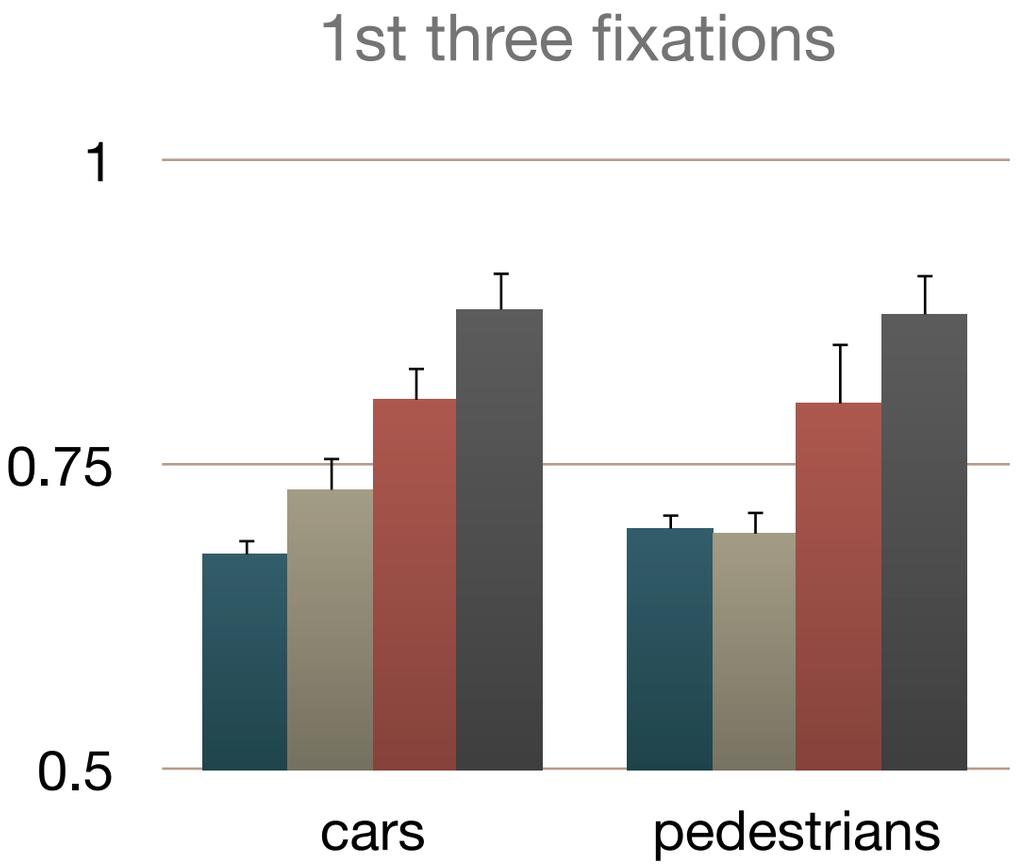
Predicting eye movements during searches for cars and pedestrians

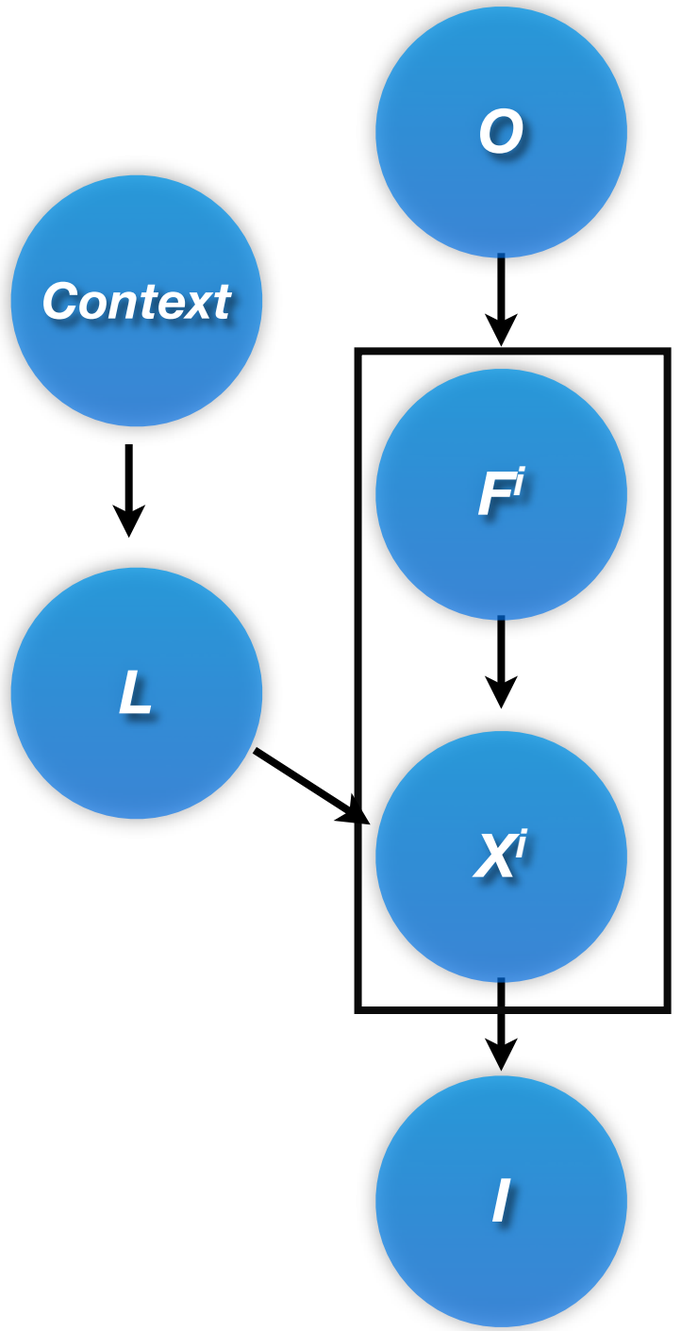




*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during searches for cars and pedestrians



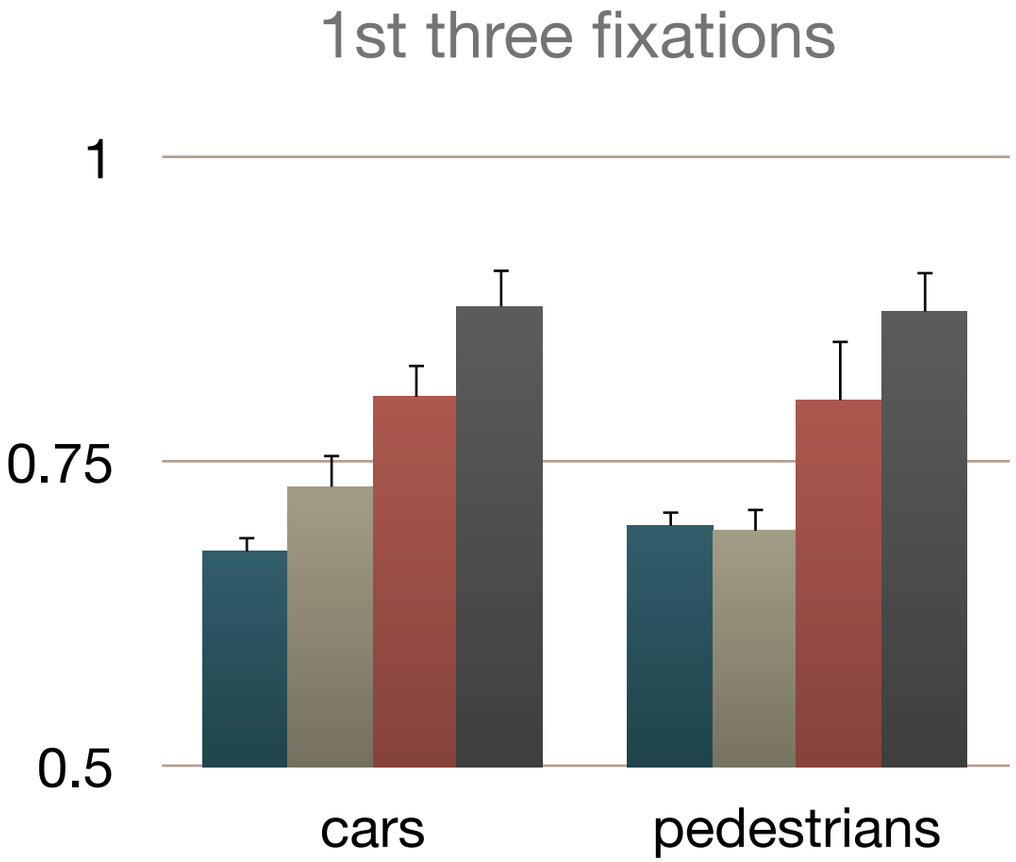


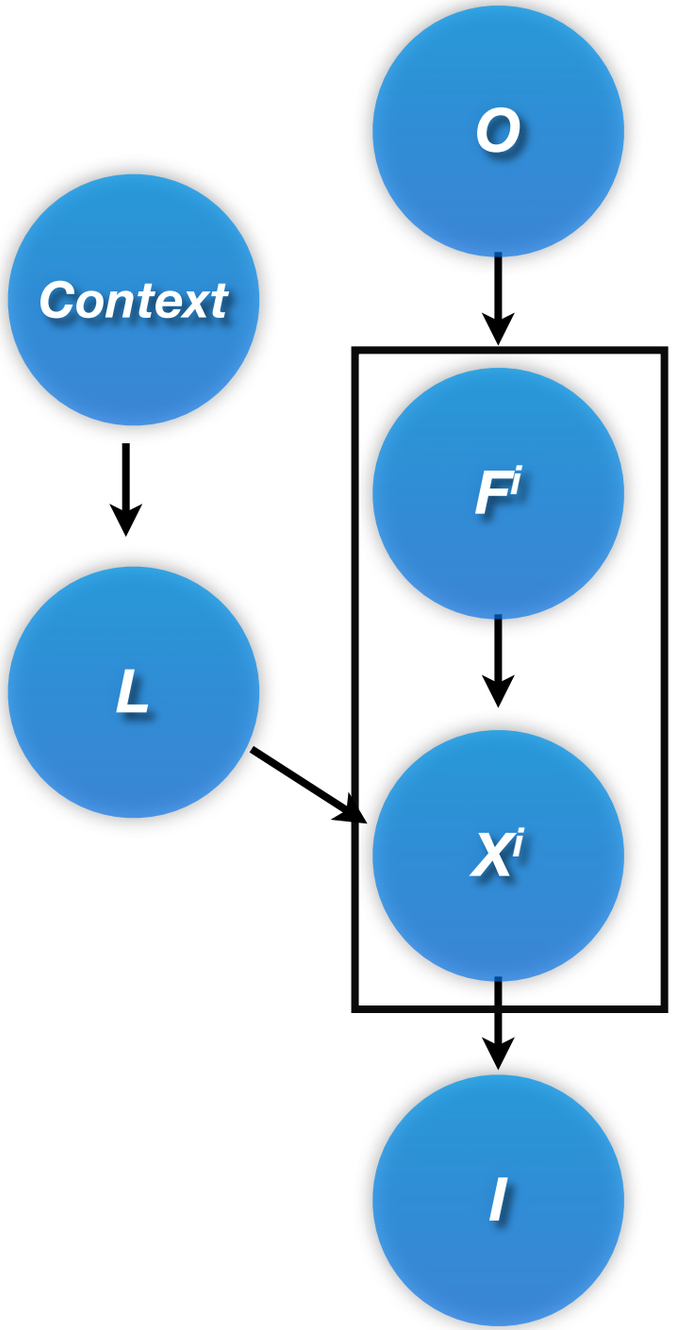
- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during searches for cars and pedestrians

Overall model accounts for 92% of inter-subject agreement!



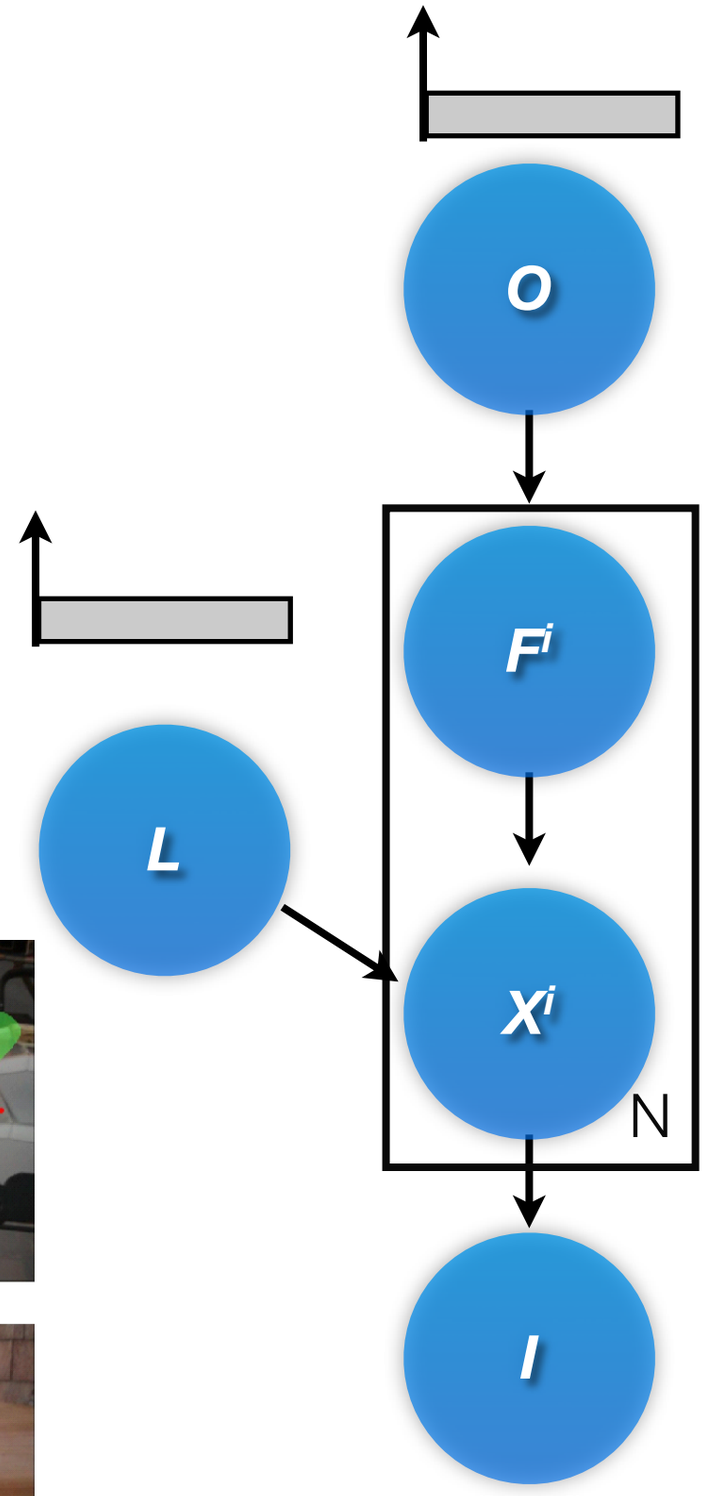


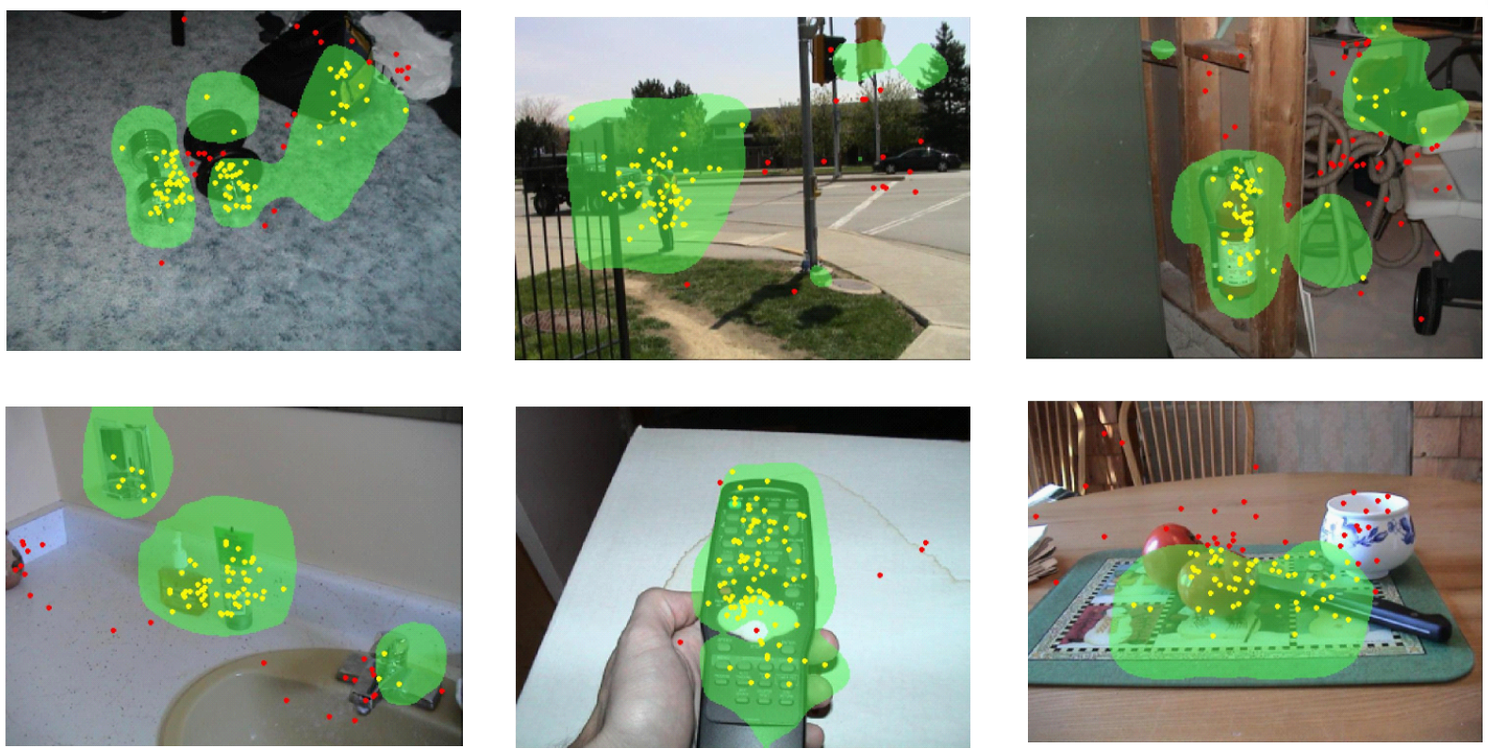
- Uniform priors (bottom-up)
- Feature priors
- Feature + contextual (spatial) priors
- Humans

*similar (independent) results by Ehinger Hidalgo Torralba & Oliva (in press)

Predicting eye movements during free viewing

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





human eye data from Bruce & Tsotsos

Summary

Summary

- Attention as part of the inference process that solves the visual recognition problem of ‘what is where’

Summary

- Attention as part of the inference process that solves the visual recognition problem of ‘what is where’
- Main goal of the visual system is to infer the identity and the position of objects in visual scenes:
  - Spatial attention emerges as a strategy to reduce the uncertainty in shape information while feature-based attention reduces the uncertainty in spatial information
  - Featural and spatial attention represent two distinct modes of a computational process solving the problem of recognizing and localizing objects, especially in difficult recognition tasks such as in cluttered natural scenes

Summary

- Attention as part of the inference process that solves the visual recognition problem of ‘what is where’
- Main goal of the visual system is to infer the identity and the position of objects in visual scenes:
  - Spatial attention emerges as a strategy to reduce the uncertainty in shape information while feature-based attention reduces the uncertainty in spatial information
  - Featural and spatial attention represent two distinct modes of a computational process solving the problem of recognizing and localizing objects, especially in difficult recognition tasks such as in cluttered natural scenes
- Model agnostic about the specific algorithm for the inference process (i.e., no claim made about the brain computing probabilities explicitly)

Two modes of vision

Two modes of vision

- Rapid bottom-up / feedforward processing during first 100-150ms of visual processing:
  - Coarse/initial base representation
  - Enables rapid object detection/recognition (‘what is there?’)

Two modes of vision

- Rapid bottom-up / feedforward processing during first 100-150ms of visual processing:
  - Coarse/initial base representation
  - Enables rapid object detection/recognition (‘what is there?’)
- Top-down / re-entrant attentional processing
  - Enables recognition in clutter
  - Enables object localization

Collaborators (MIT):

- Narcisse Bichot
- **Sharat Chikkerur**
- Bob Desimone
- **Ethan Meyers**
- Tomaso Poggio
- Cheston Tan
- **Ying Zhang**

Robert J. and Nancy D. Carney
Fund for Scientific Innovation

Acknowledgments