Natural Scene Recognition: From Humans to Computers

Li Fei-Fei

1. Computer Science Department
2. Psychology Department

A picture is worth a thousand words.

--- Confucius

or *Printers' Ink* Ad (1921)

blue

rugged

white and red

bright

textured structure

green

porous

elongated shapes

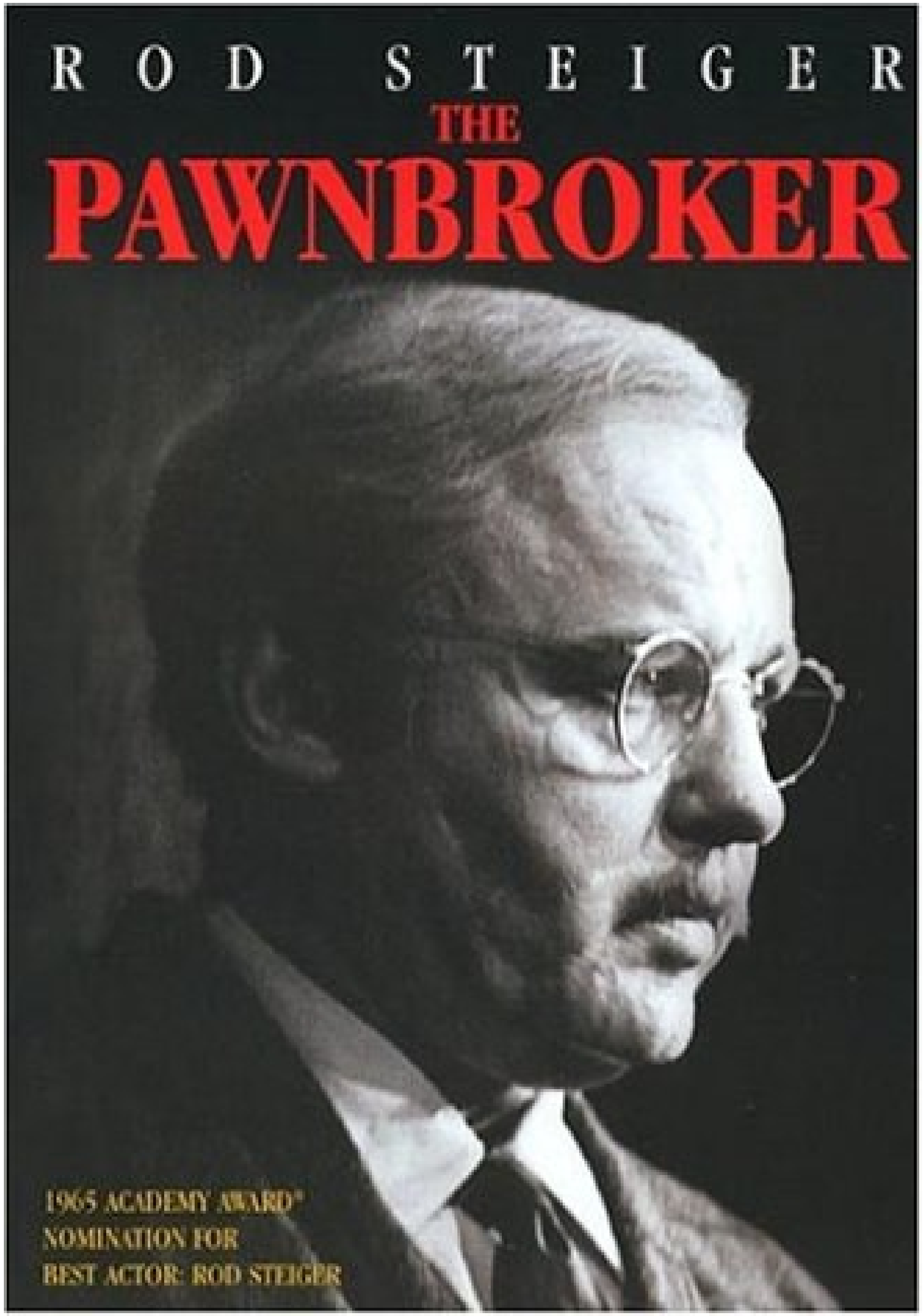
grey

- To understand human visual intelligence by via psychophysical and physiological experiments

- To build intelligent visual algorithms for machines and robots living room

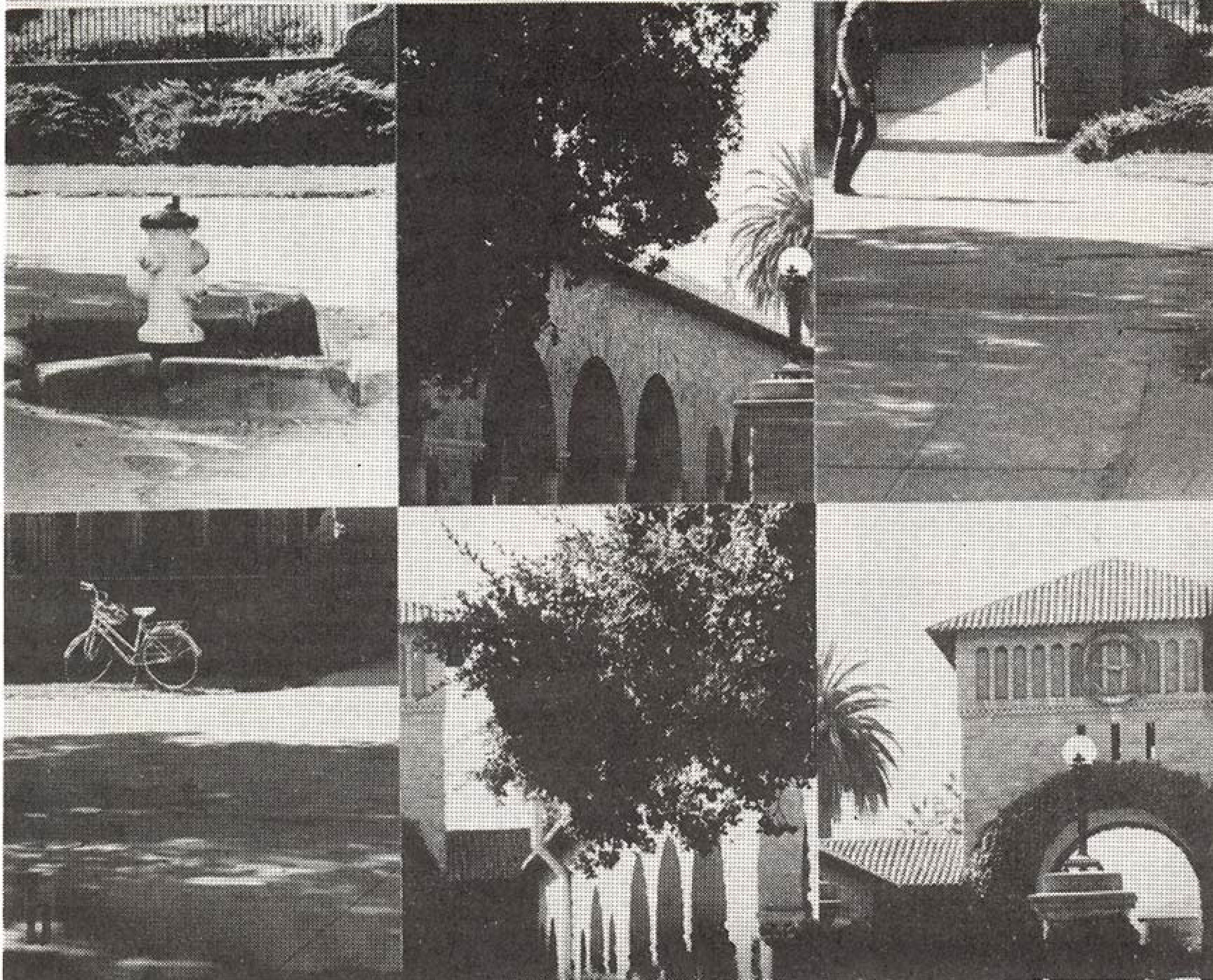
beach

city

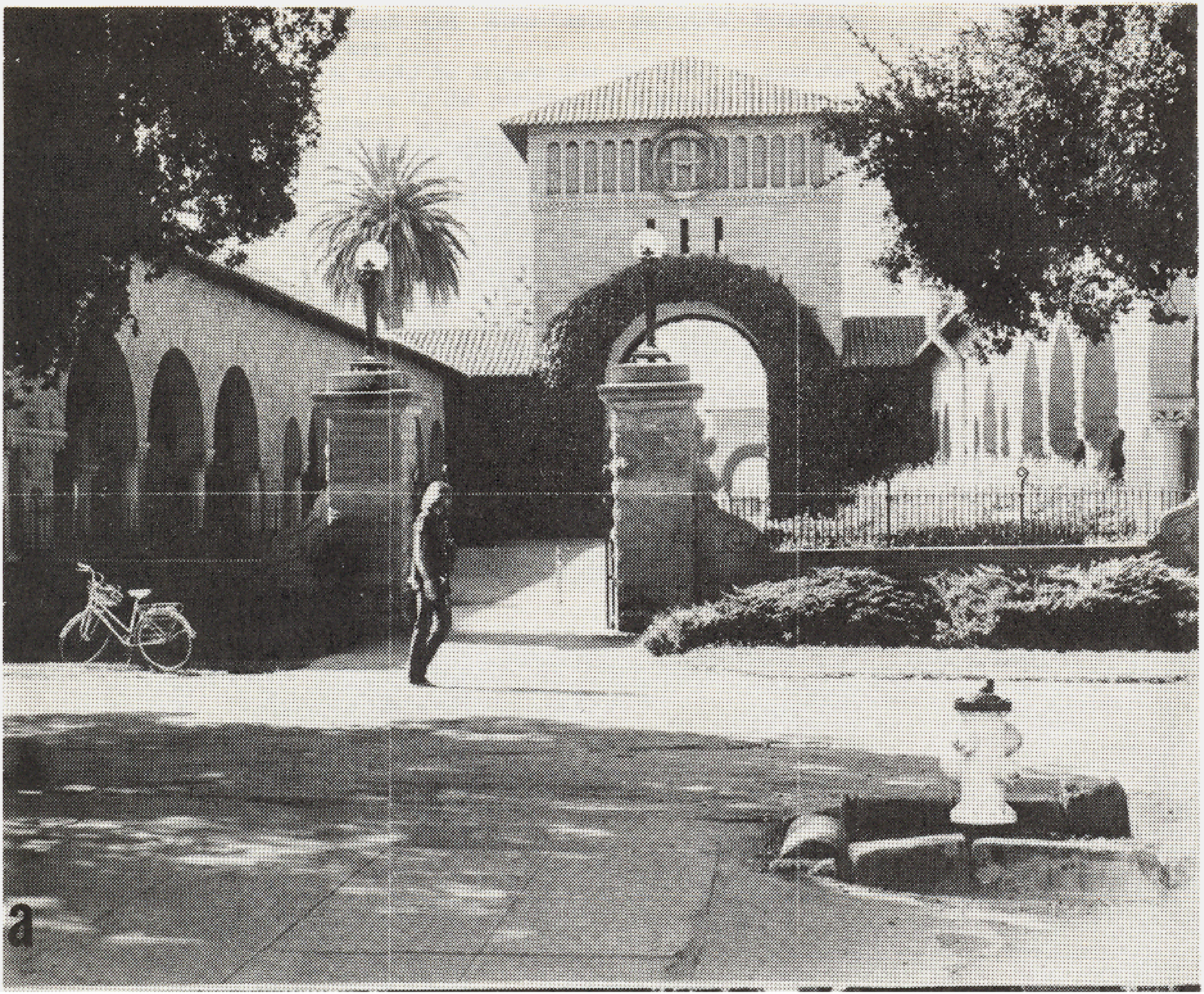

ROD STEIGER
THE
PAWNBROKER
1965 ACADEMY AWARD®
NOMINATION FOR
BEST ACTOR: ROD STEIGER

S. Lumet, 1965

Potter, Biederman, etc. 1970s

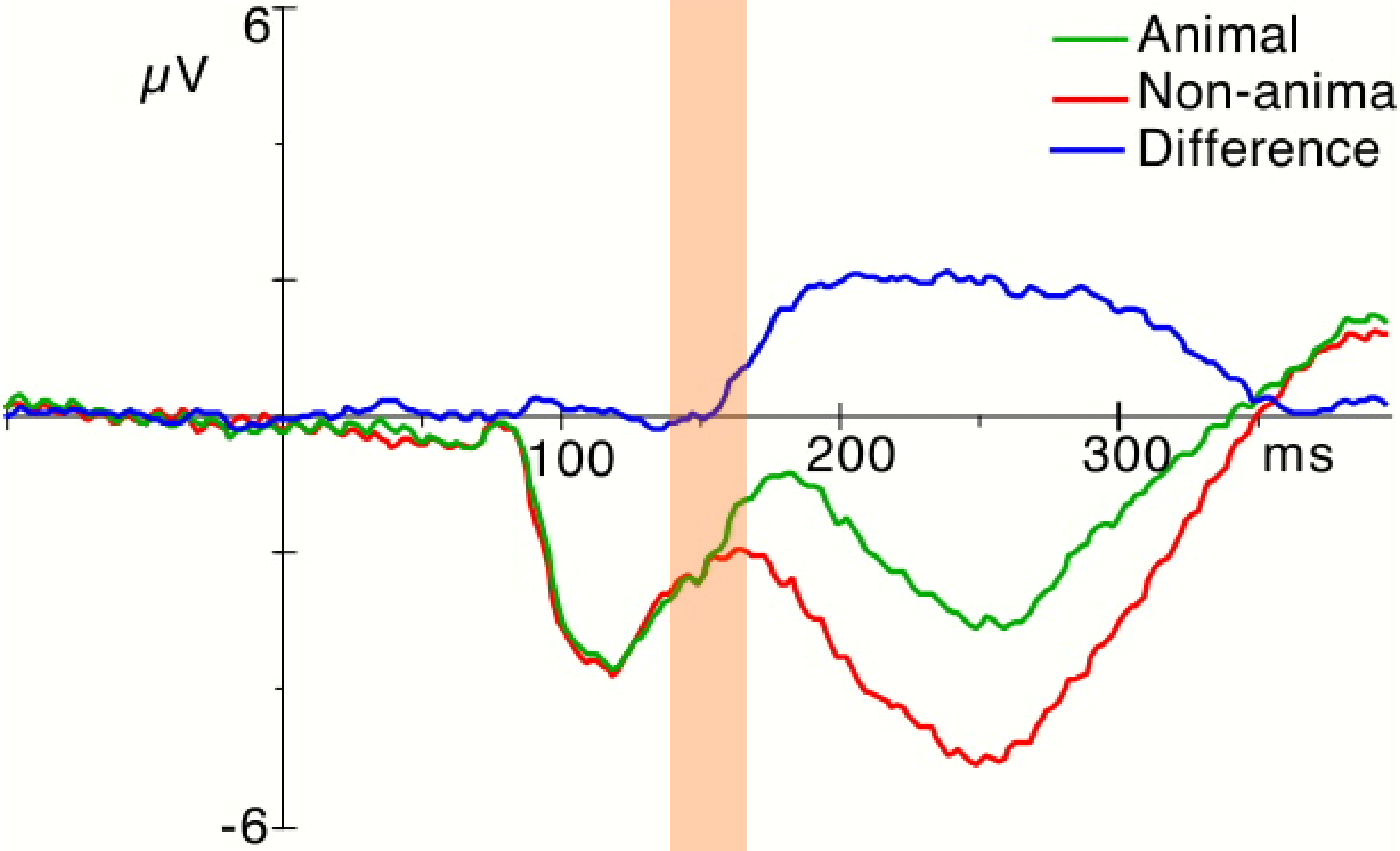


Biederman, *Science*, 1973



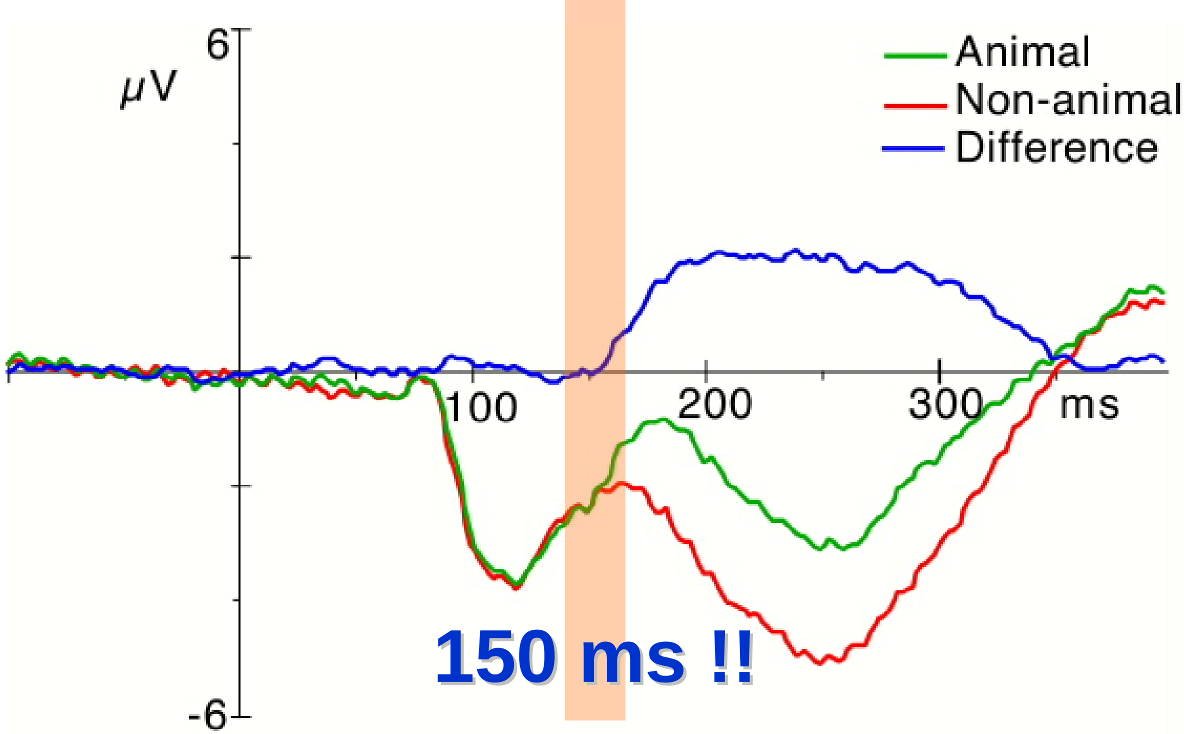
Biederman, *Science*, 1973

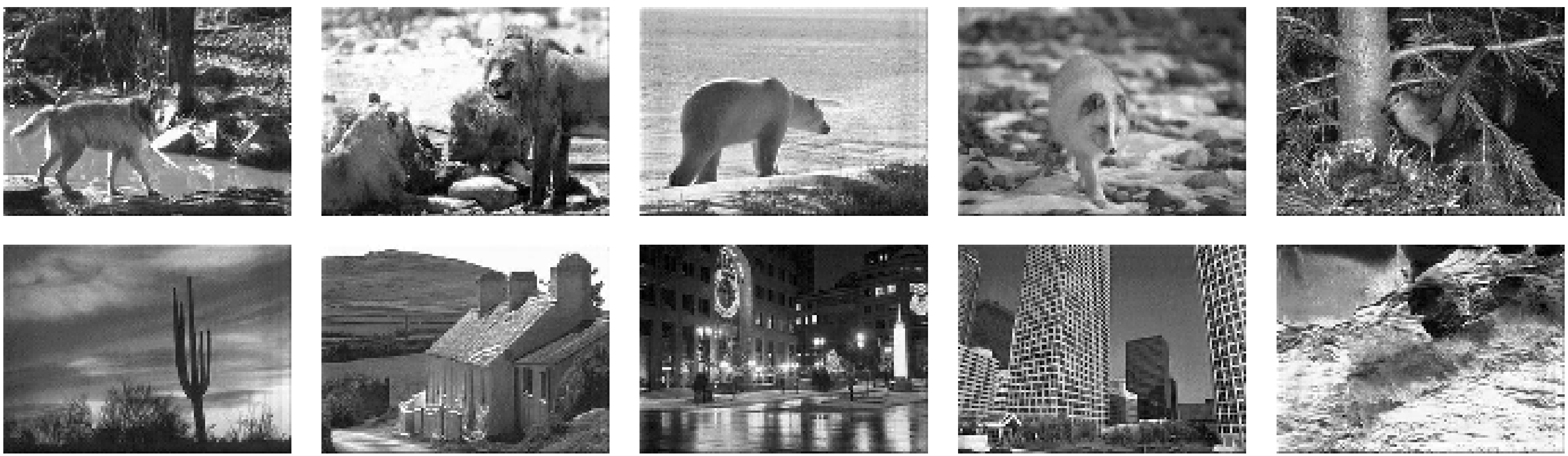
Thorpe, et al. *Nature,* 1996



150 ms !!

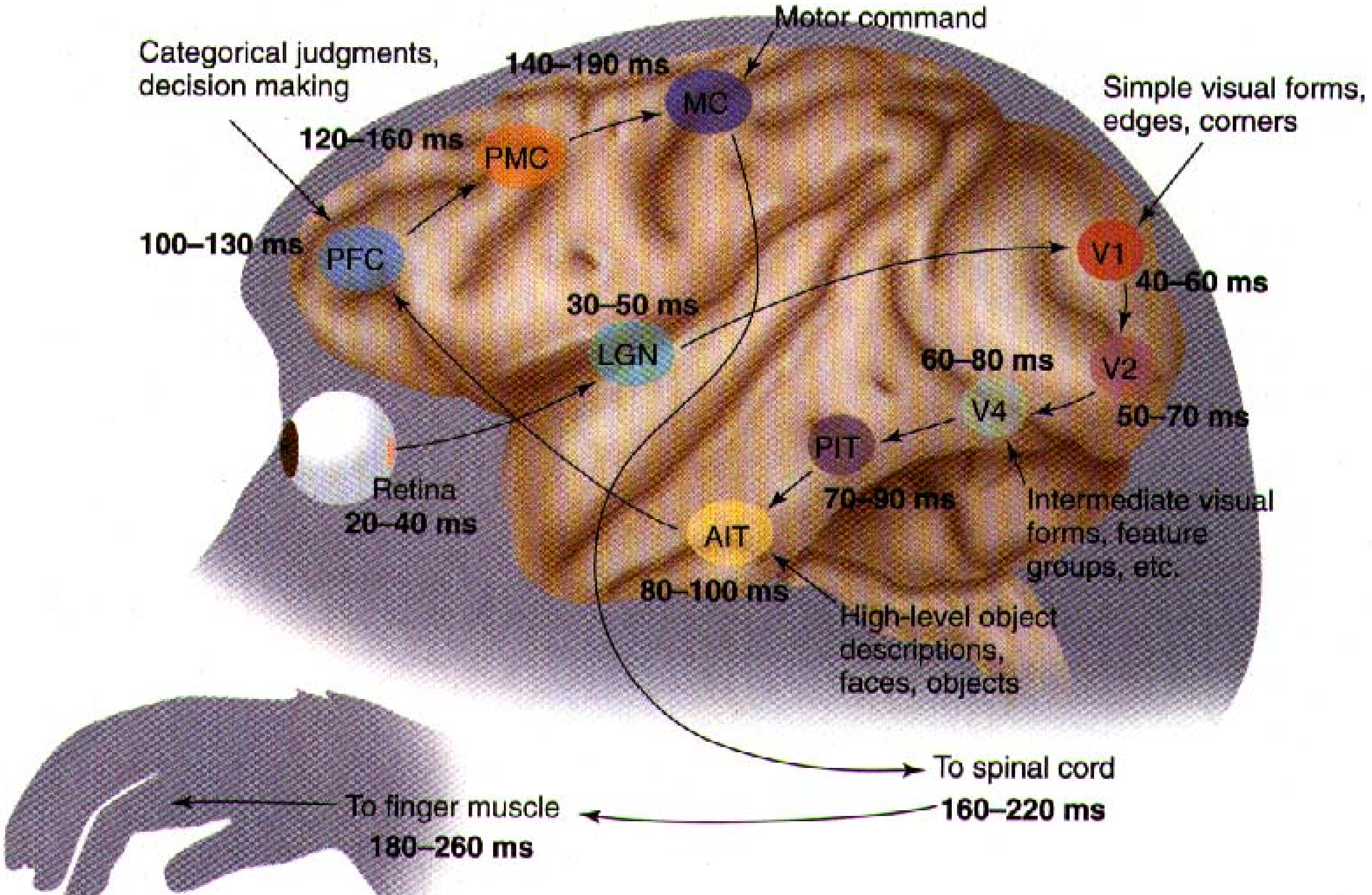
Thorpe, et al. *Nature,* 1996





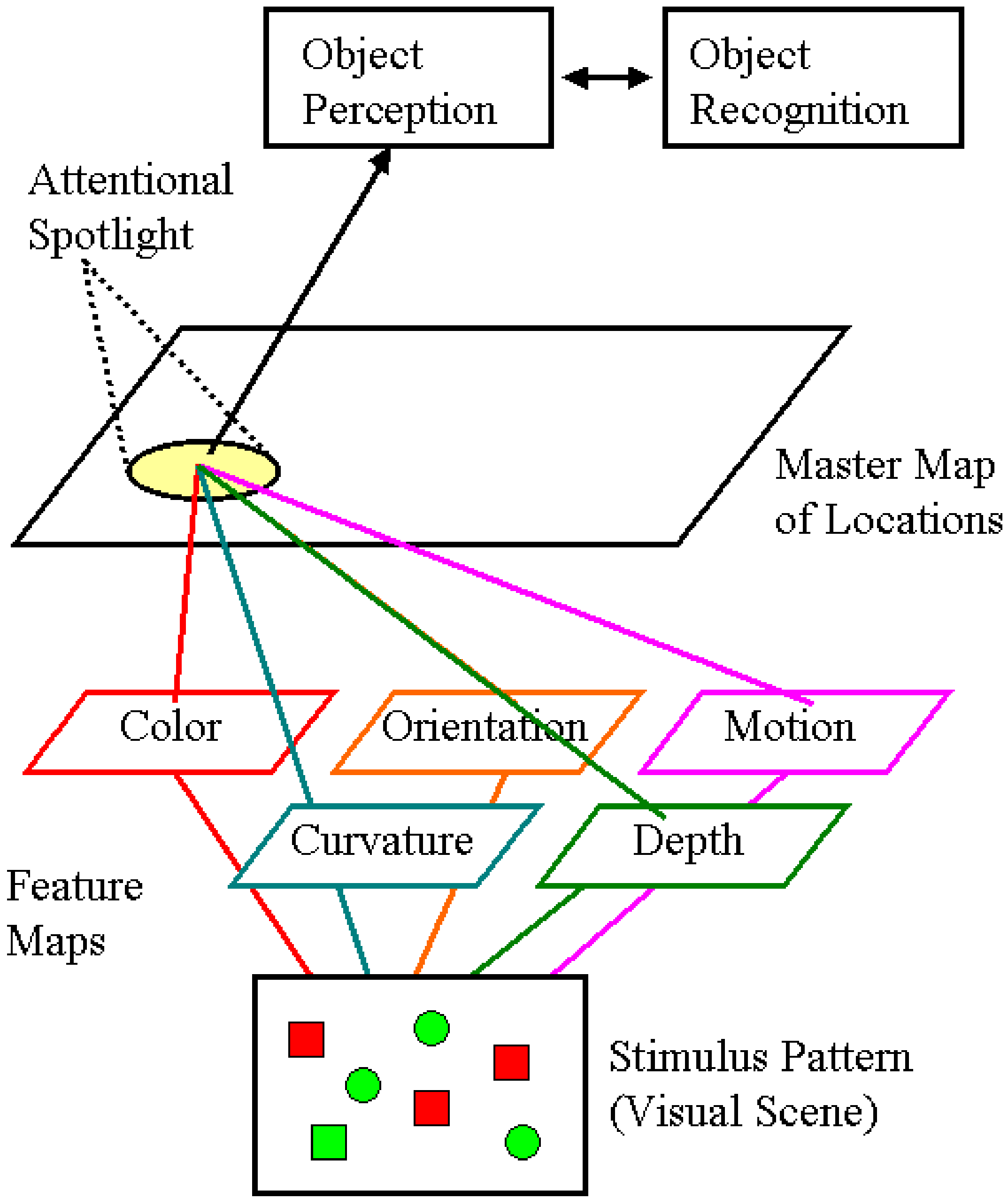
Delorme, et al. 1998

A feed-forward mechanism?



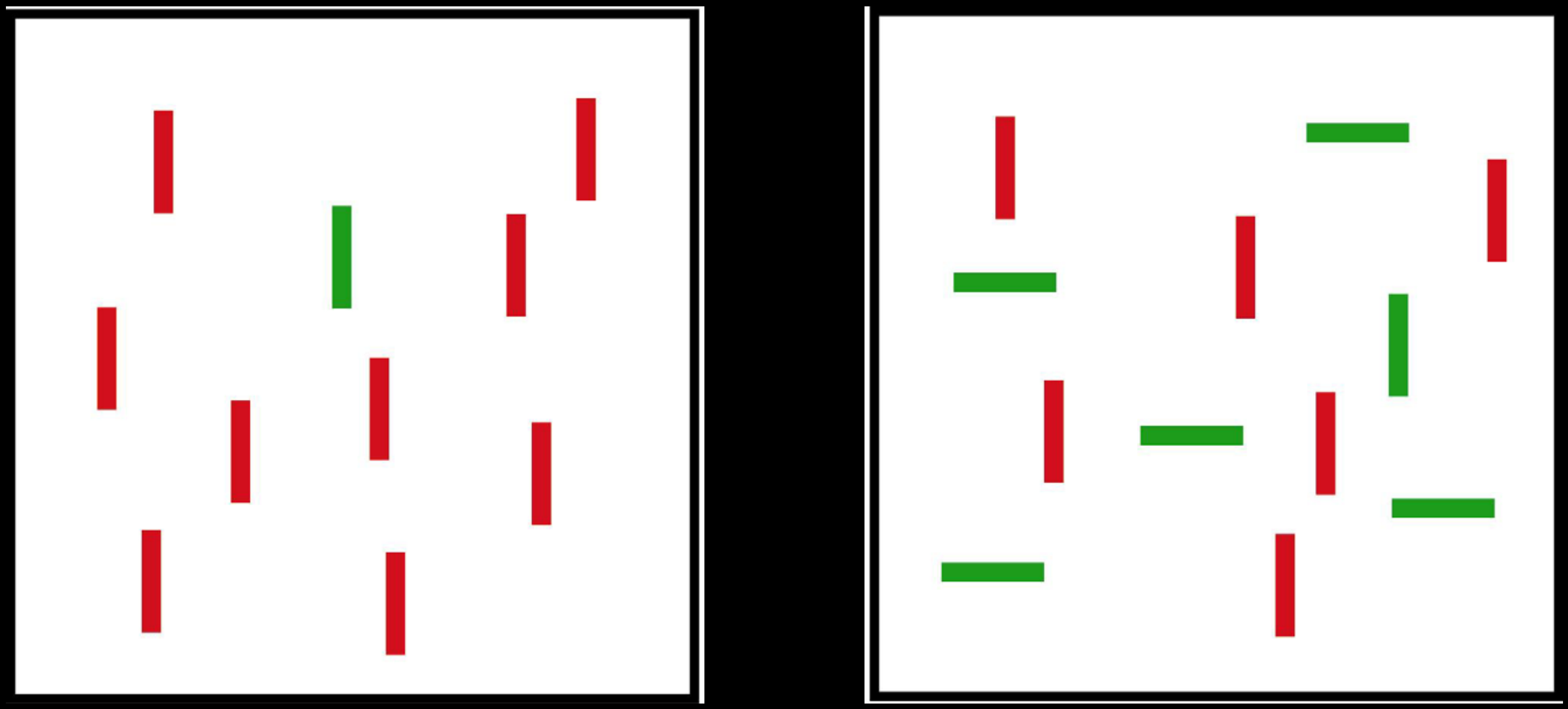
Thorpe, et al. *Science, 2001*

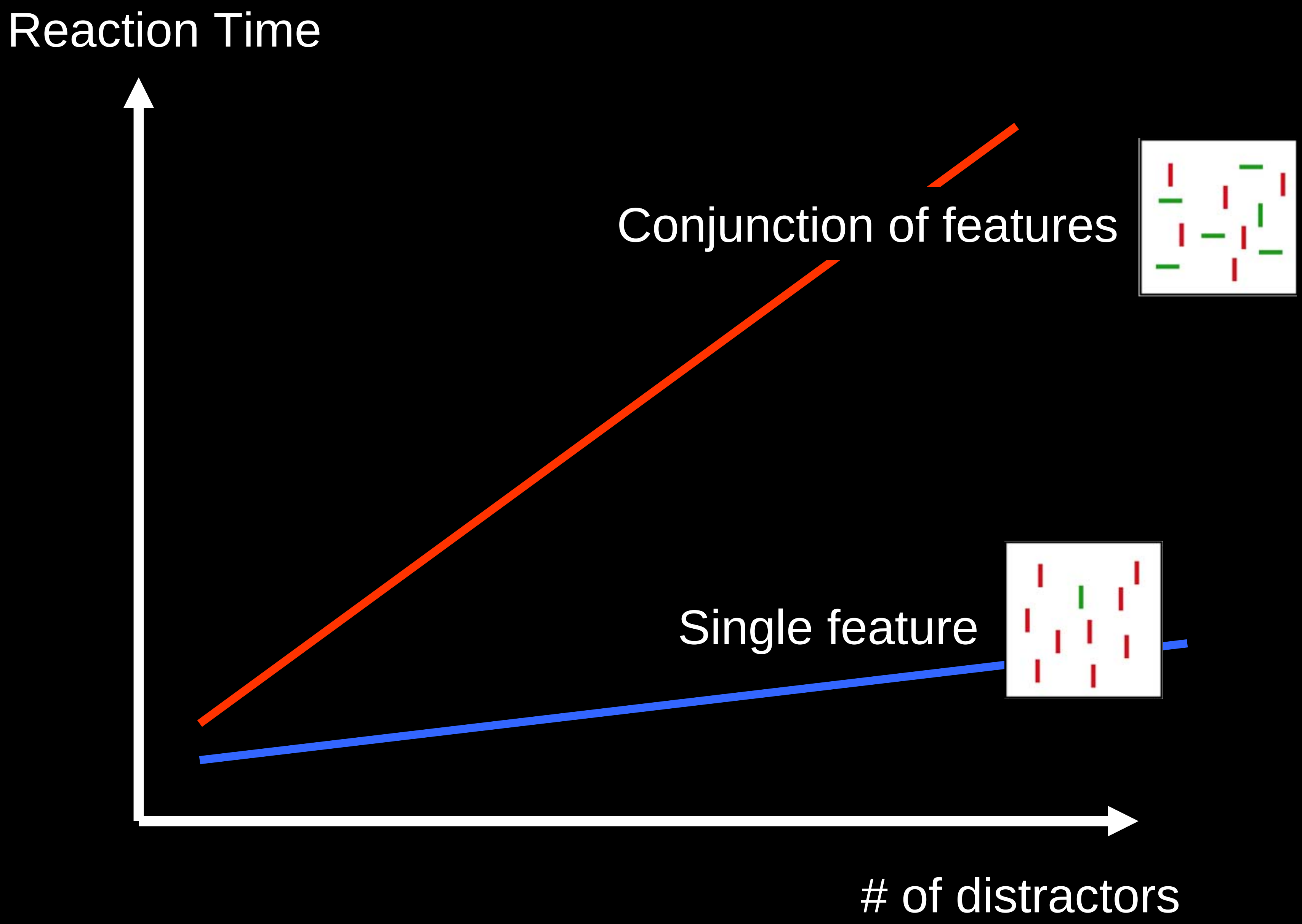
Feature Integration Theory



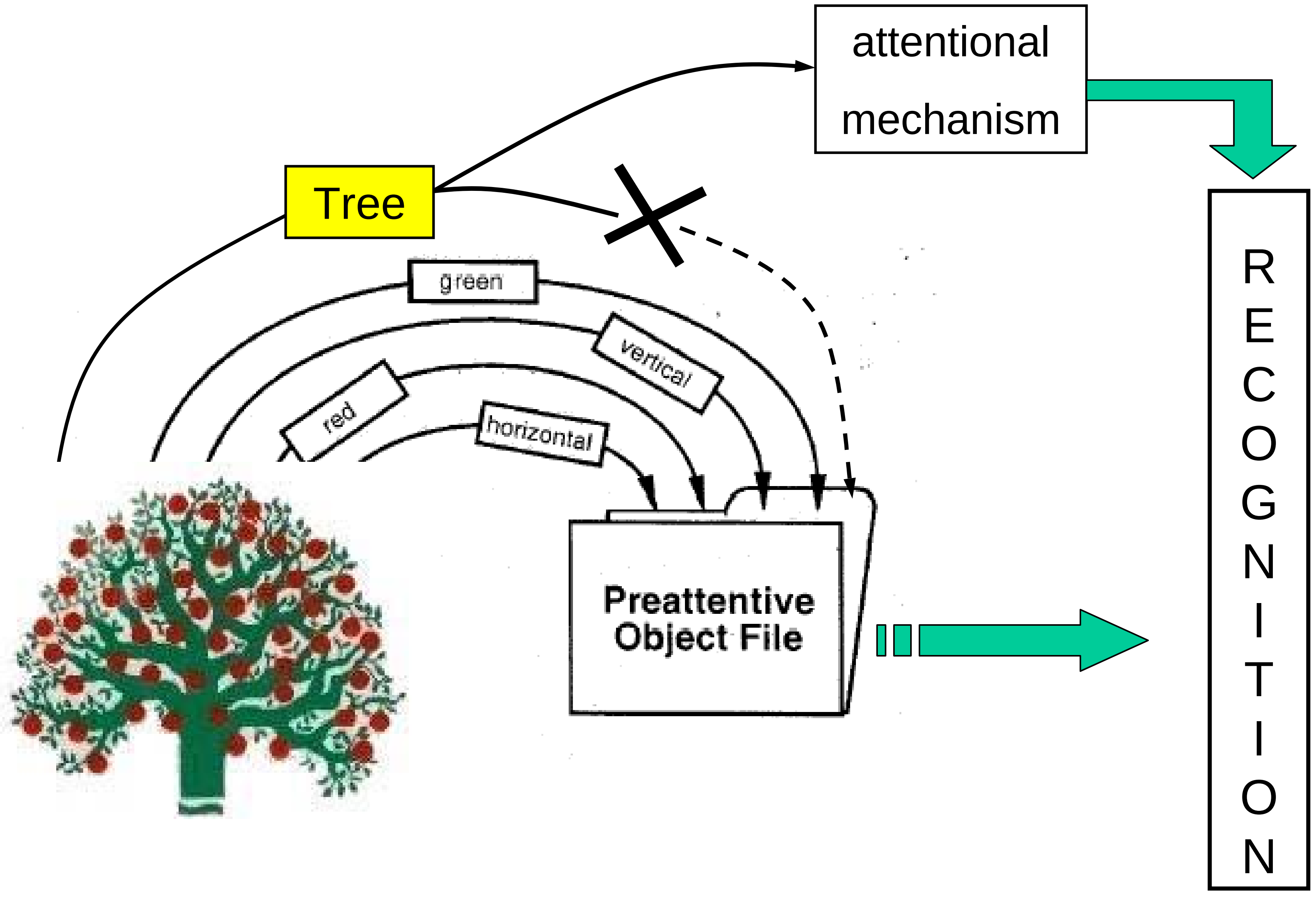
Treisman et al. 1980

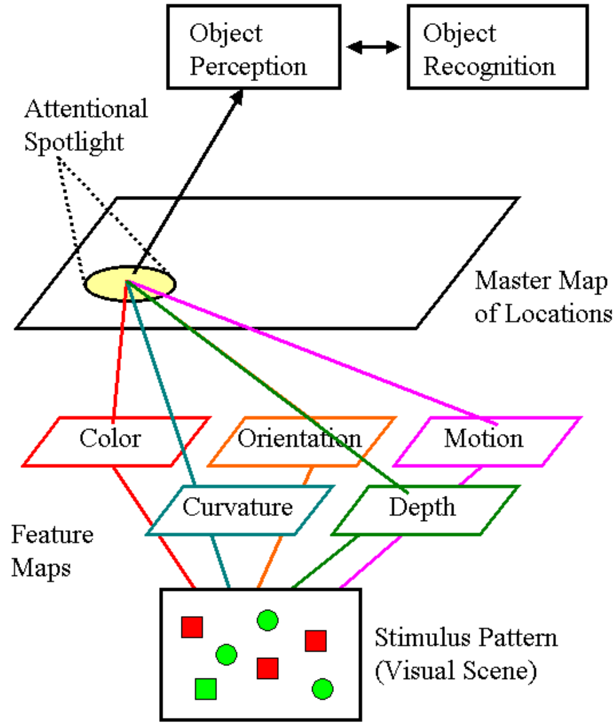
Visual Search:
find the green-vertical bar



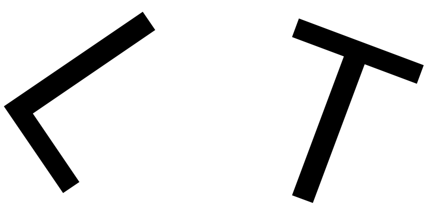


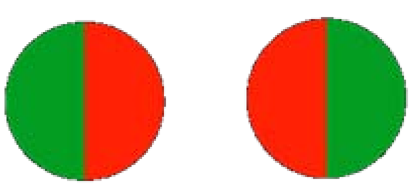
Reaction Time
Conjunction of features
Single feature
# of distractors

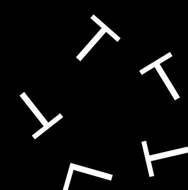

attentional
mechanism
Tree
green
vertical
red
horizontal
Preattentive
Object File
RECOGNITION

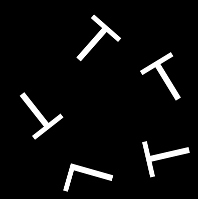


less attentional load more

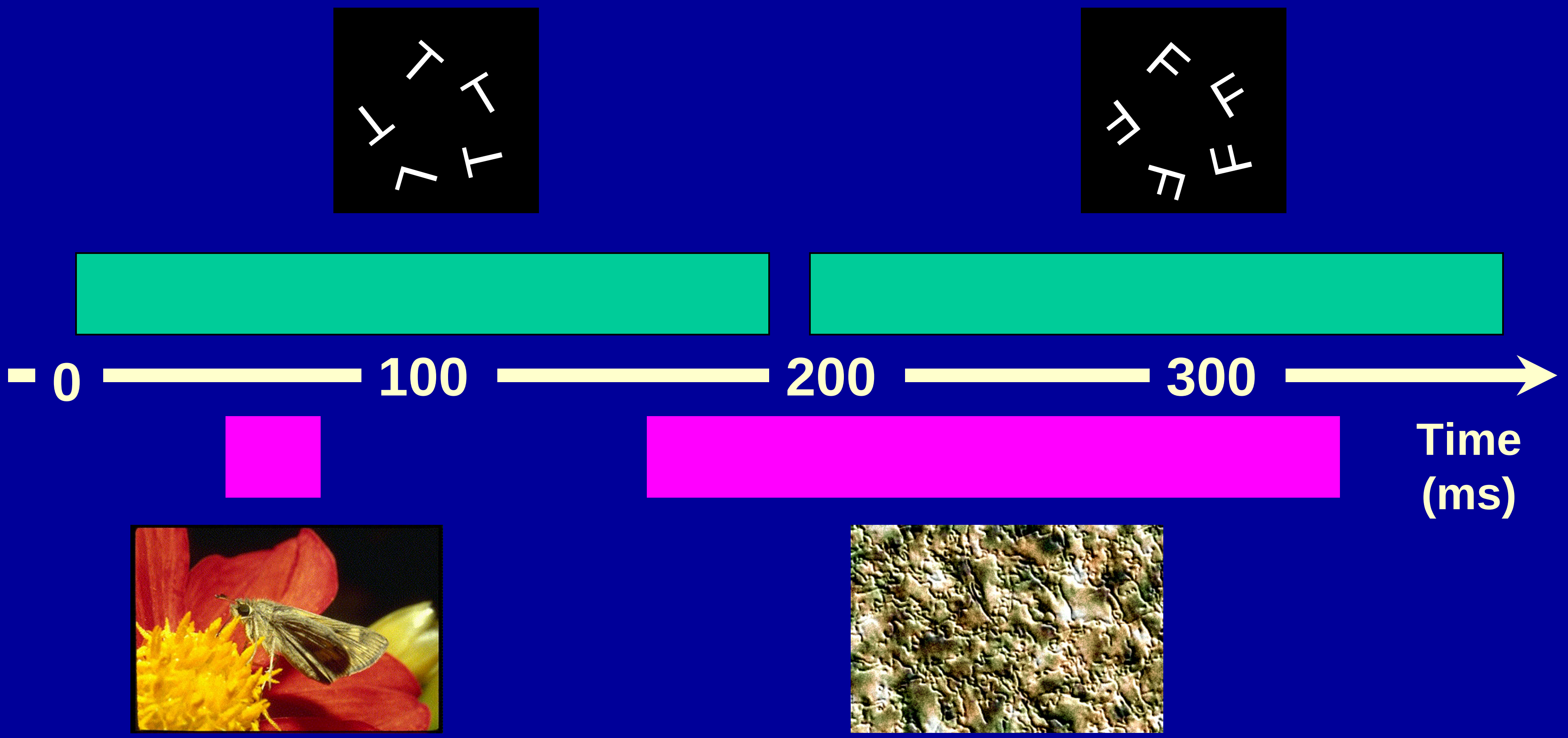




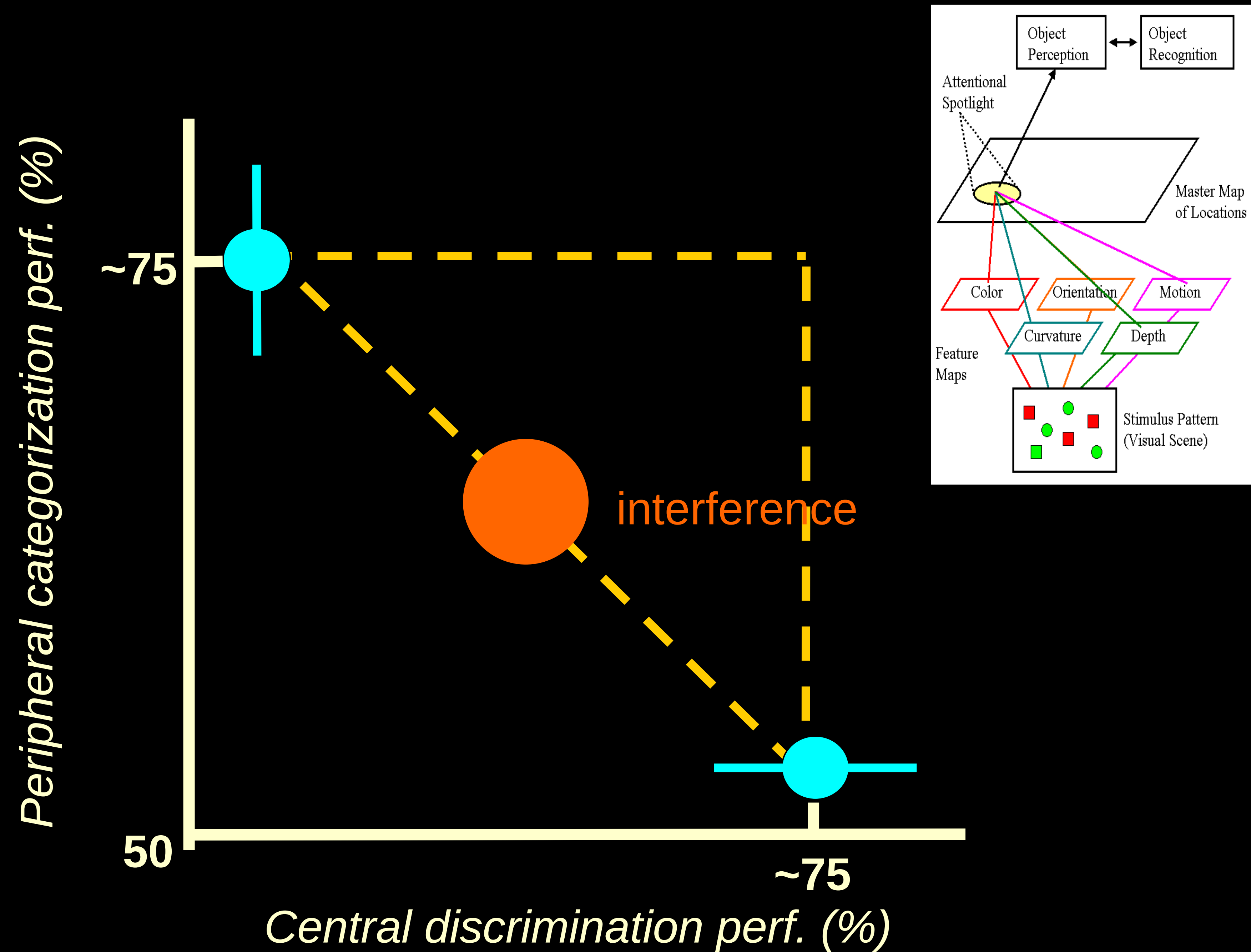




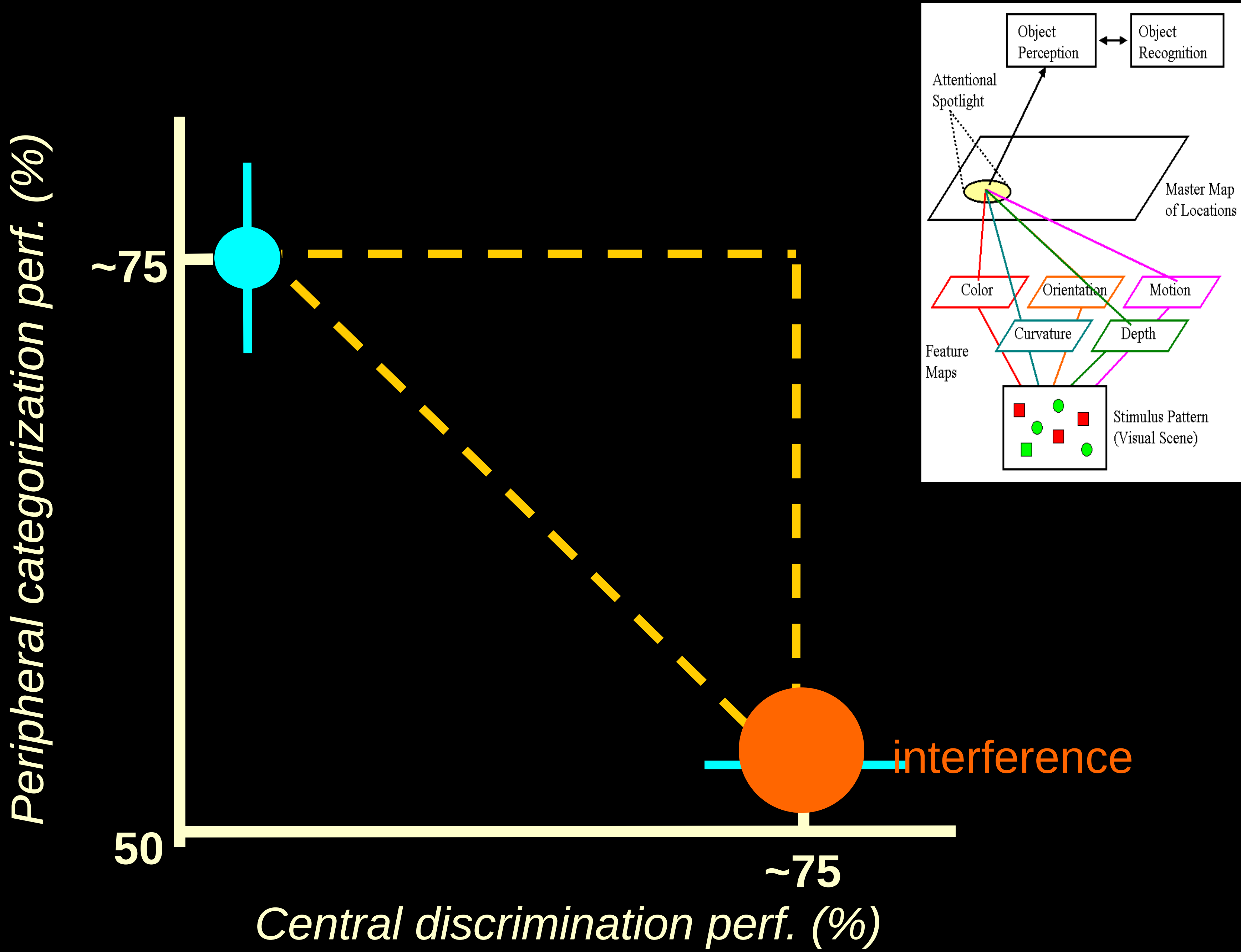
Central

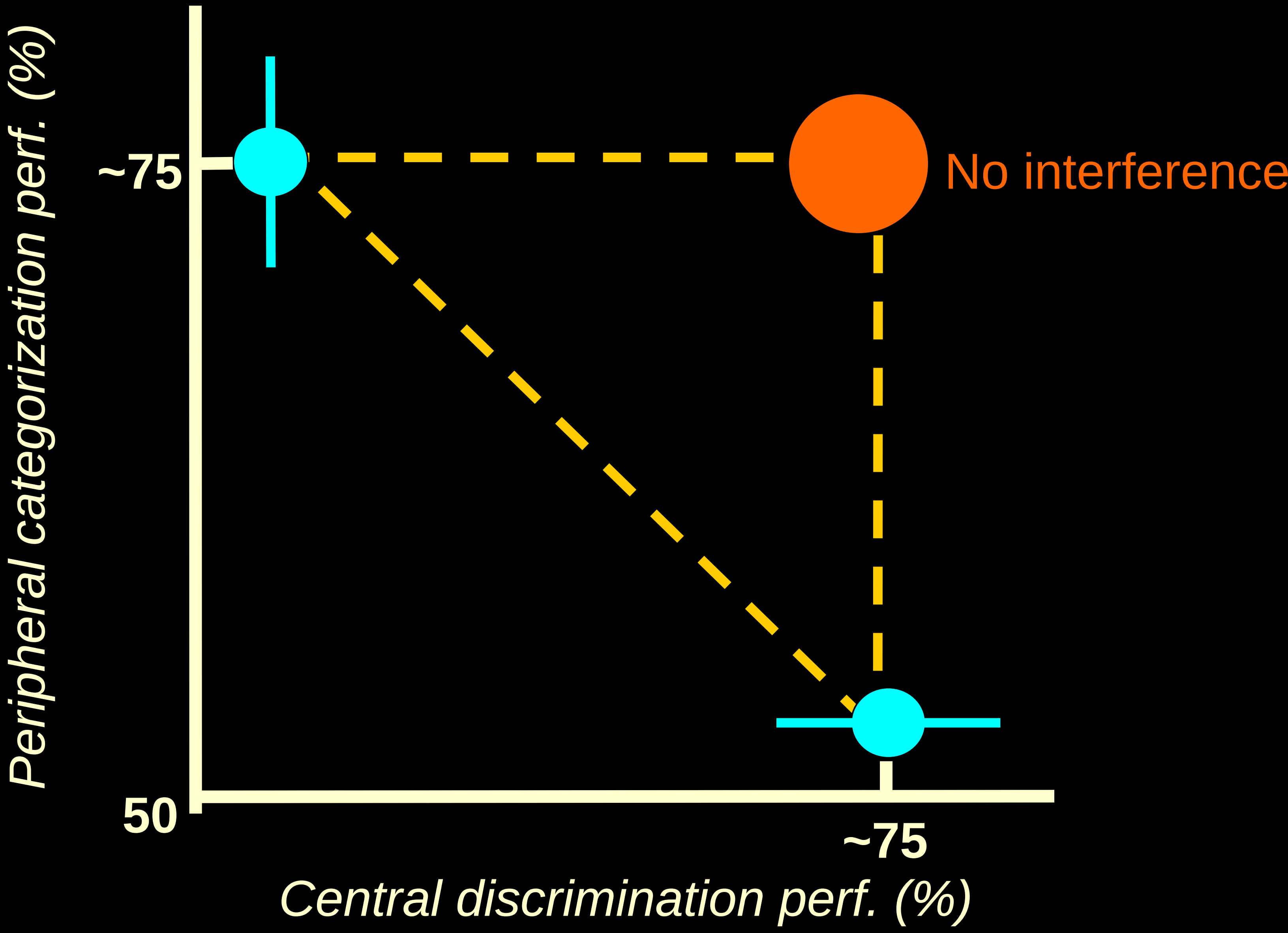


Peripheral



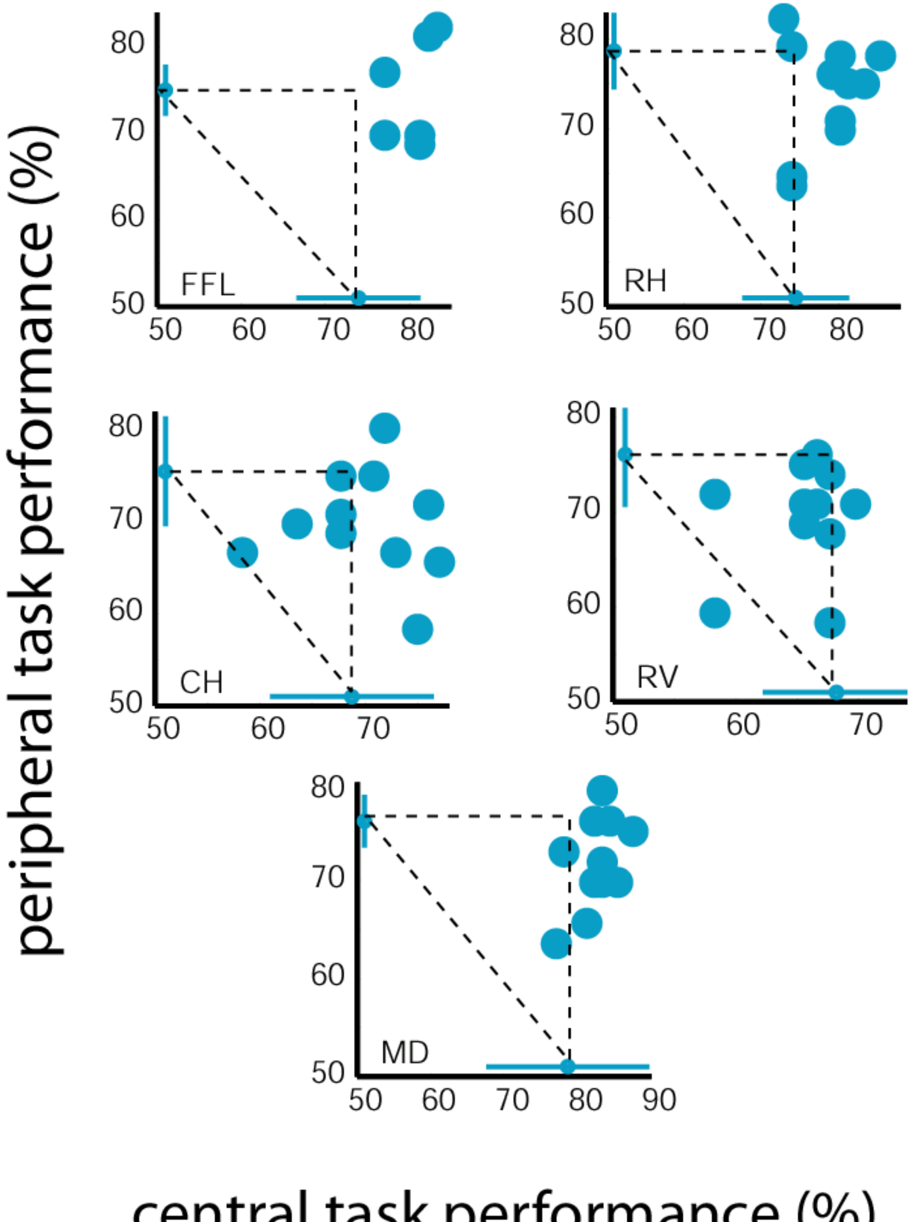
Object Perception
Object Recognition
Attentional Spotlight
Master Map of Locations
Color
Orientation
Motion
Curvature
Depth
Feature Maps
Stimulus Pattern (Visual Scene)
~75
50
~75
interference
Peripheral categorization perf. (%)
Central discrimination perf. (%)

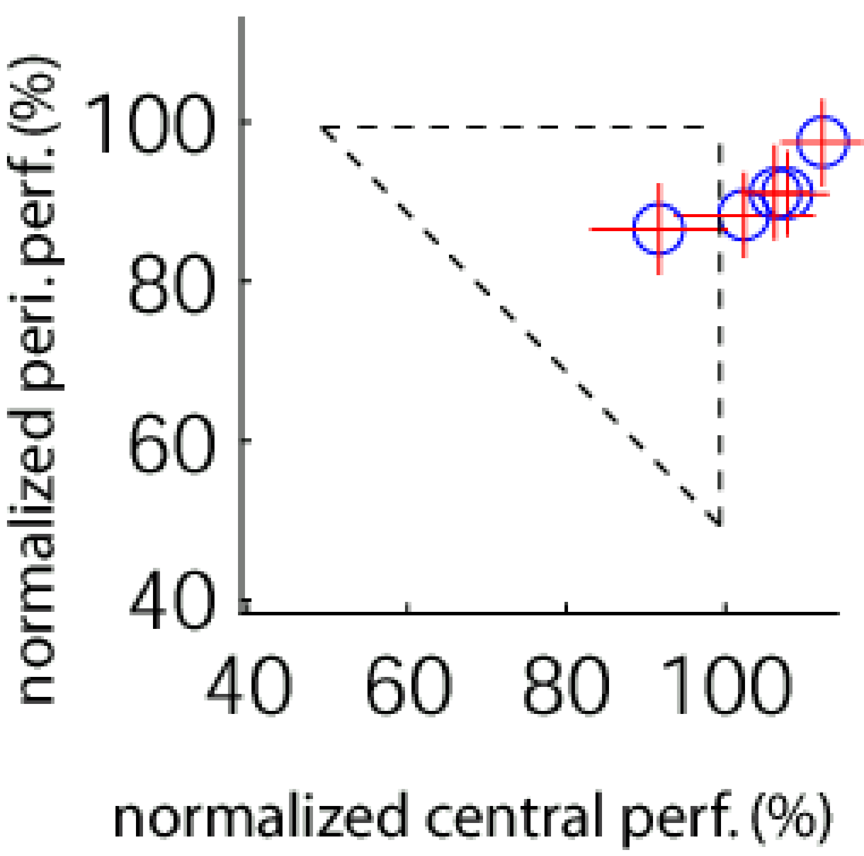

Peripheral categorization perf. (%)
~75
50
interference
~75
Central discrimination perf. (%)
Object Perception
Object Recognition
Attentional Spotlight
Master Map of Locations
Color
Orientation
Motion
Curvature
Depth
Feature Maps
Stimulus Pattern (Visual Scene)



Peripheral categorization perf. (%)
~75
50
No interference
~75
Central discrimination perf. (%)

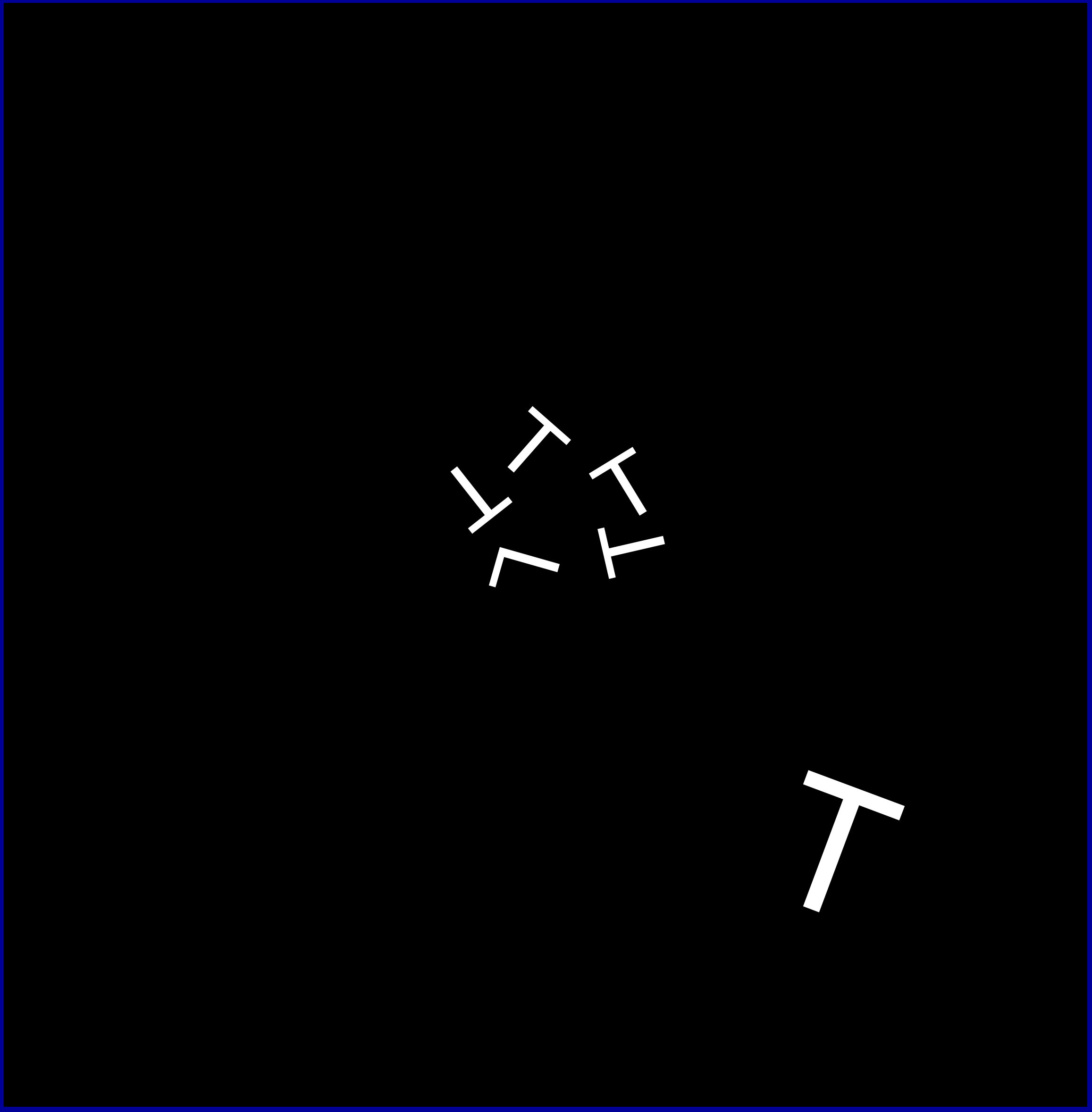
individual results



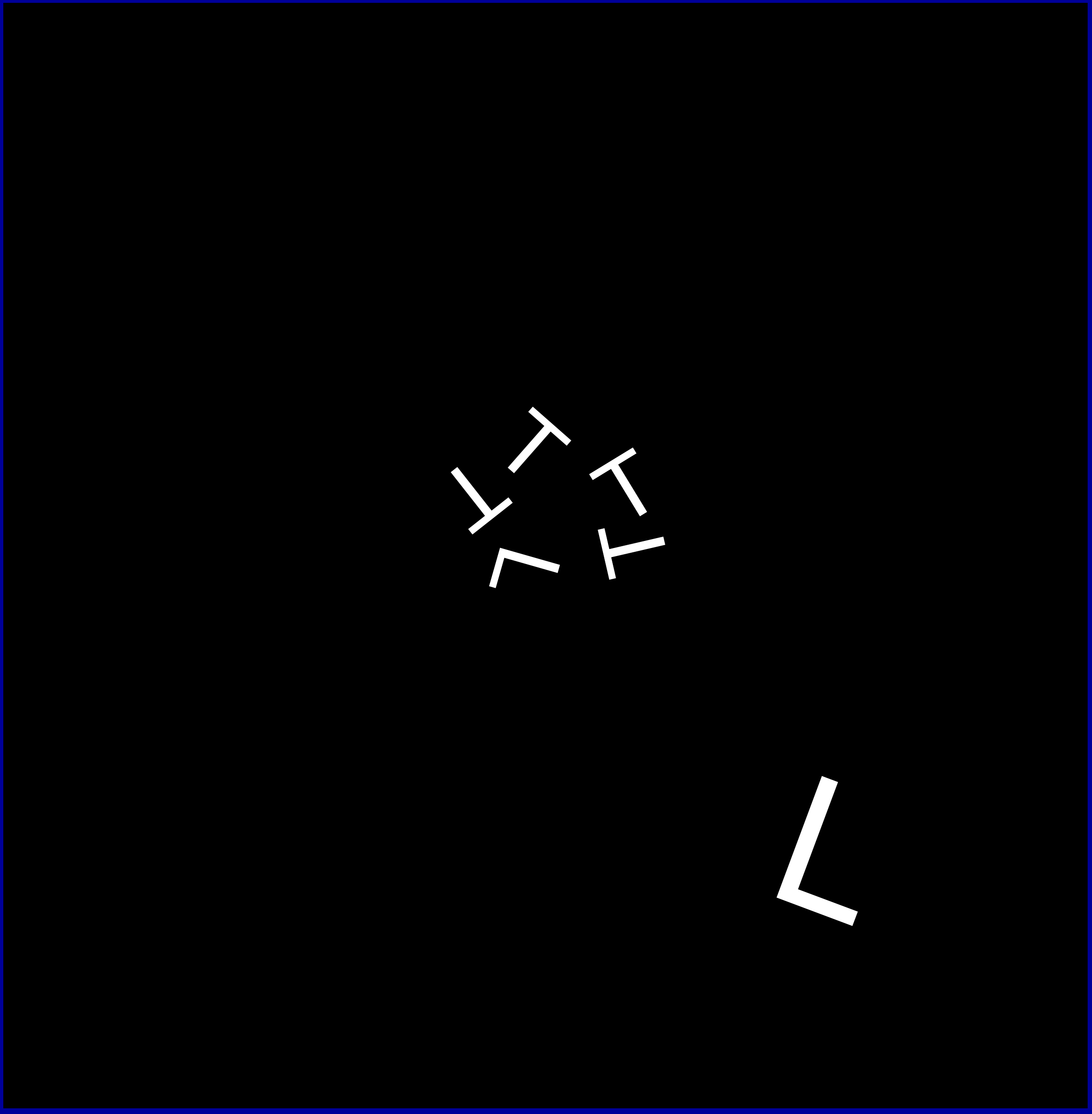


Fei-Fei et al. *PNAS*, 2002

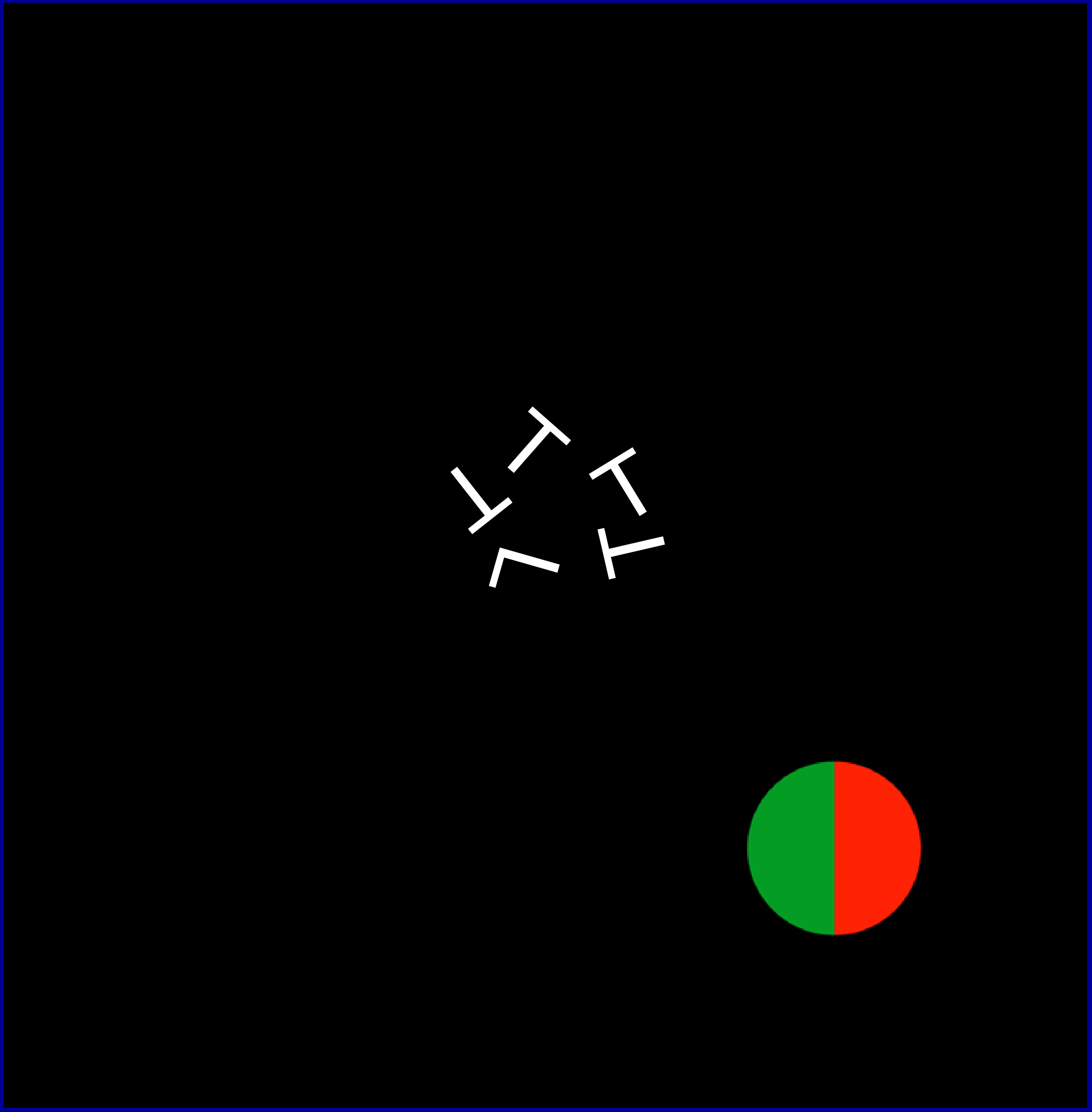
Compare to seemingly simpler tasks



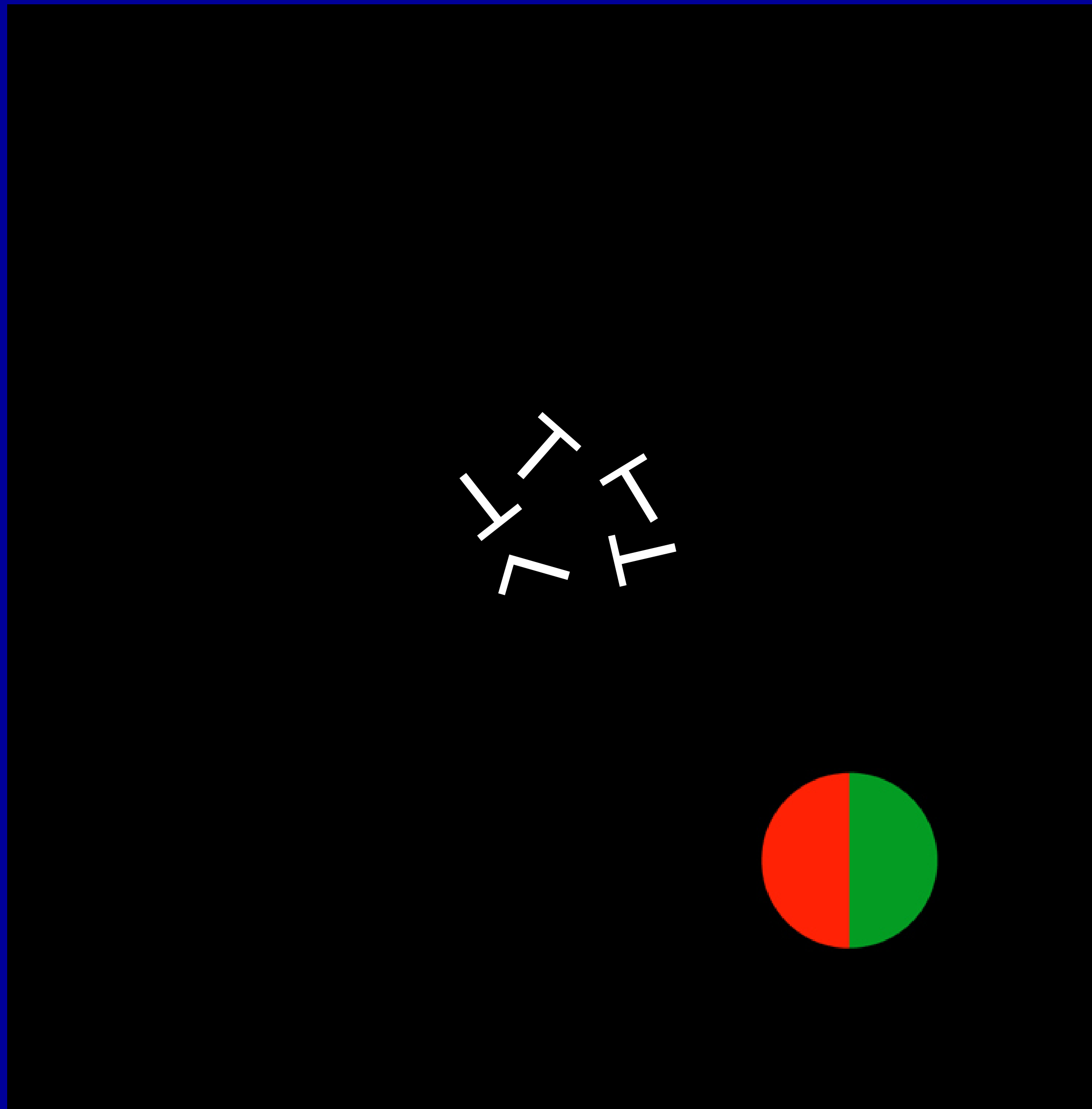
Compare to seemingly simpler tasks

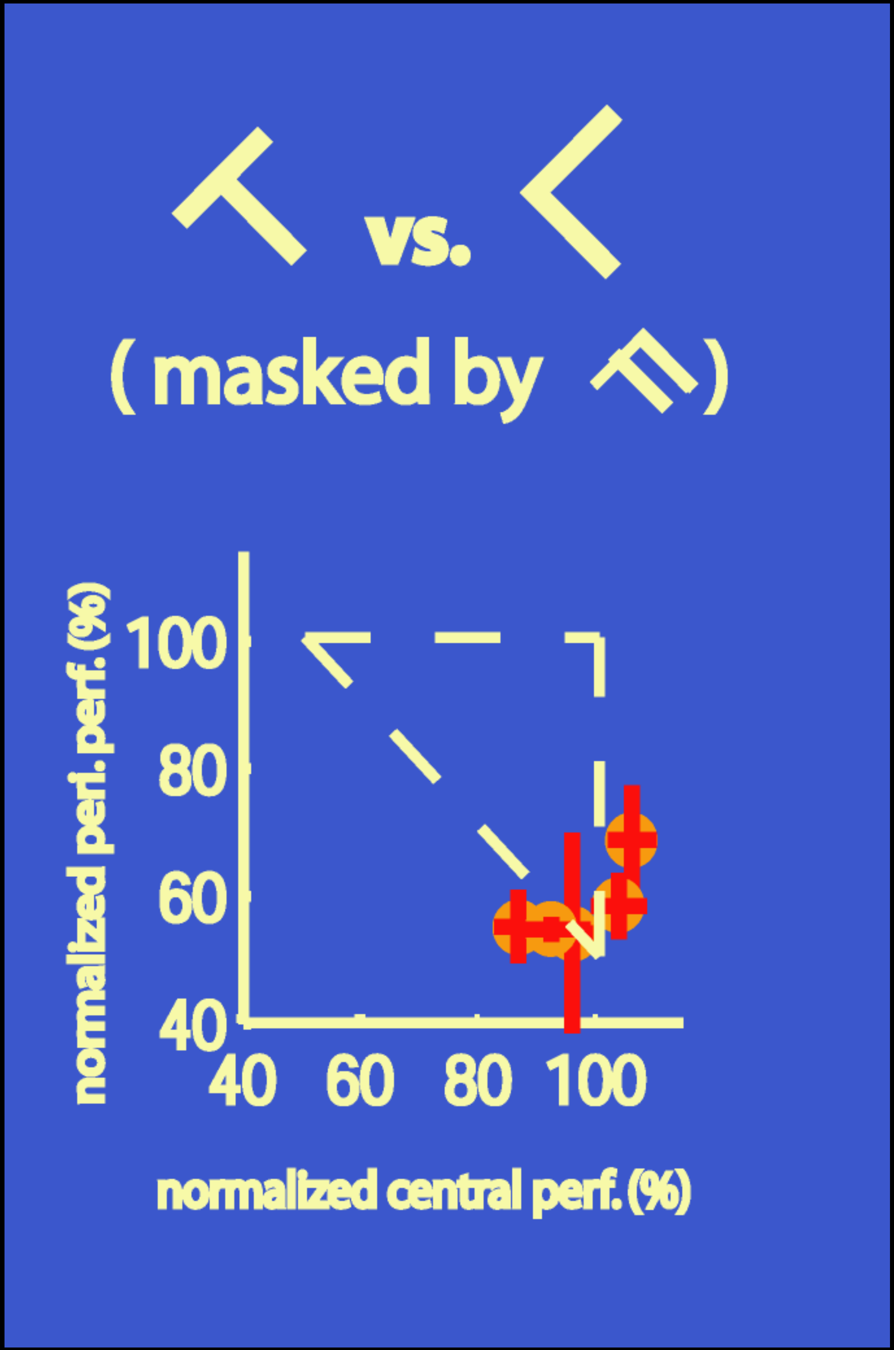


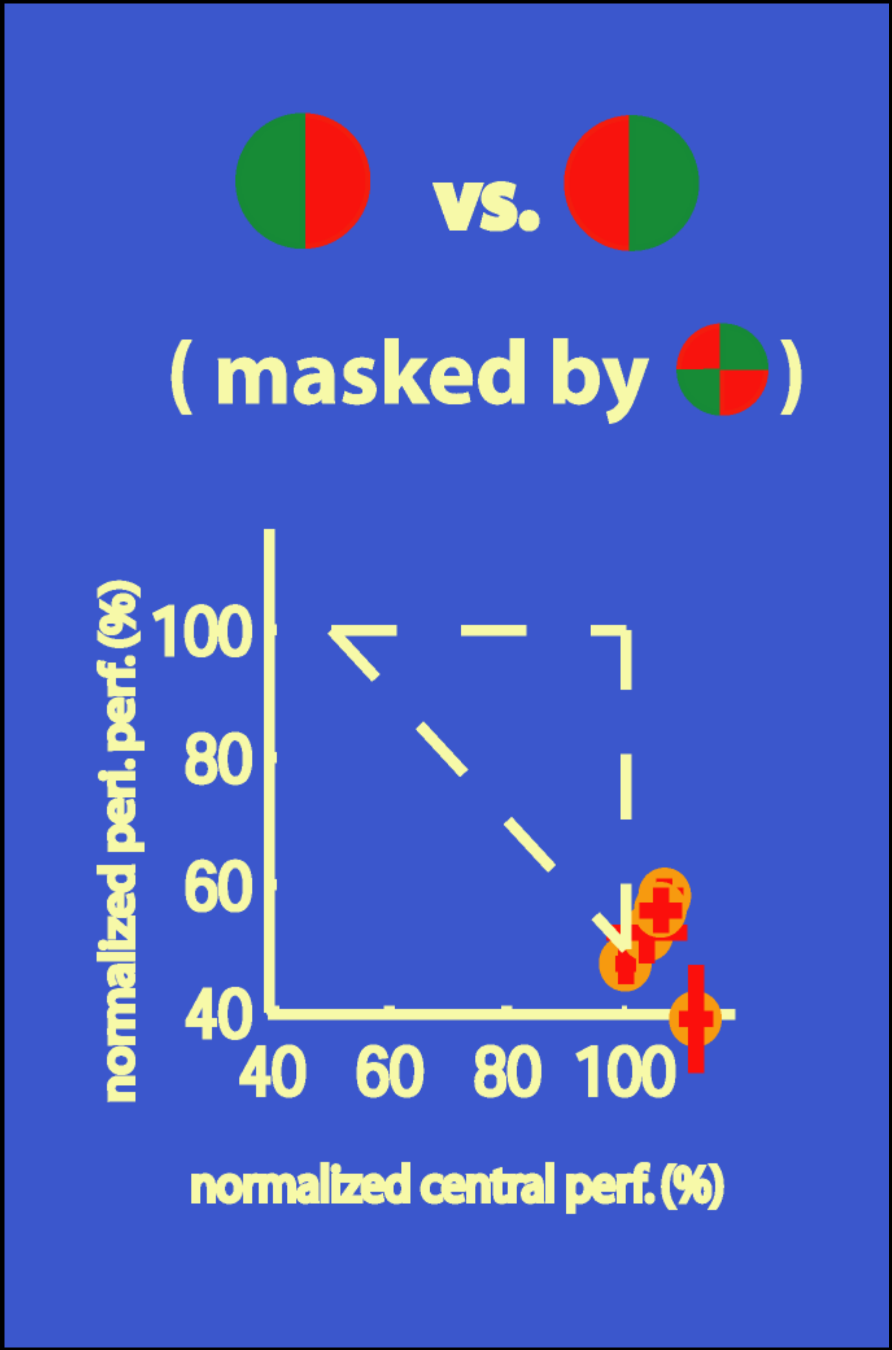
Compare to seemingly simpler tasks



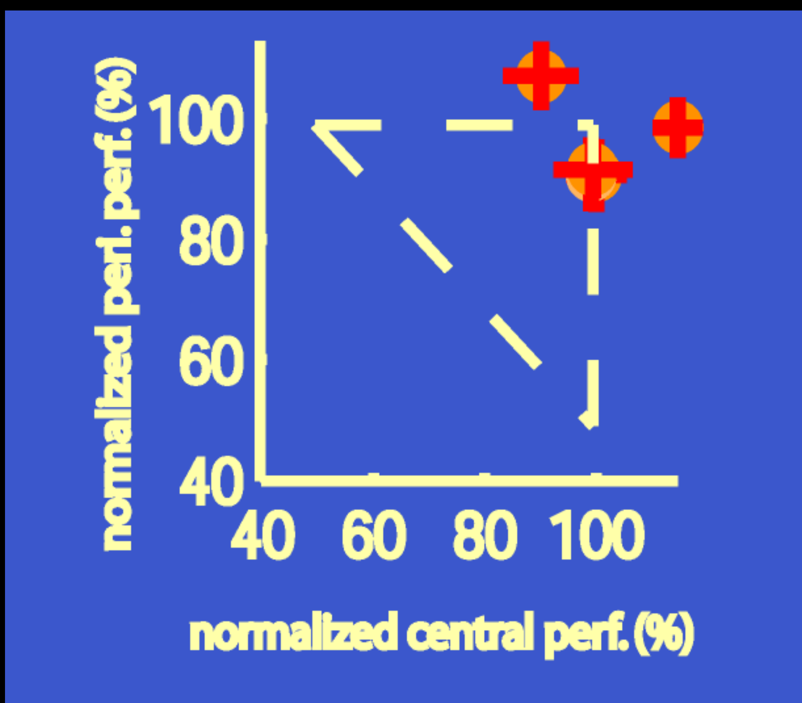
Compare to seemingly simpler tasks

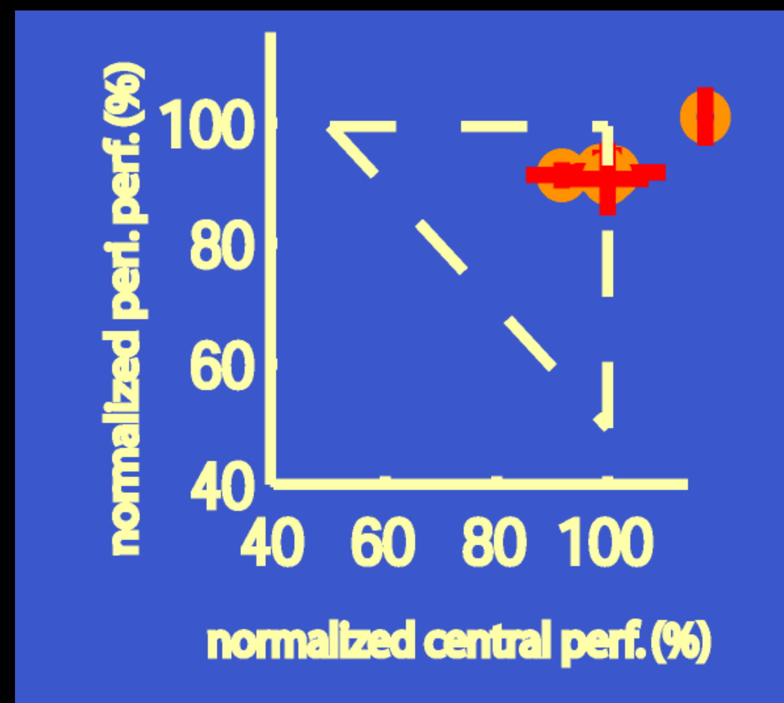



vs.
(masked by)
normalized peri. perf. (%)
100
80
60
40
40
60
80
100
normalized central perf. (%)

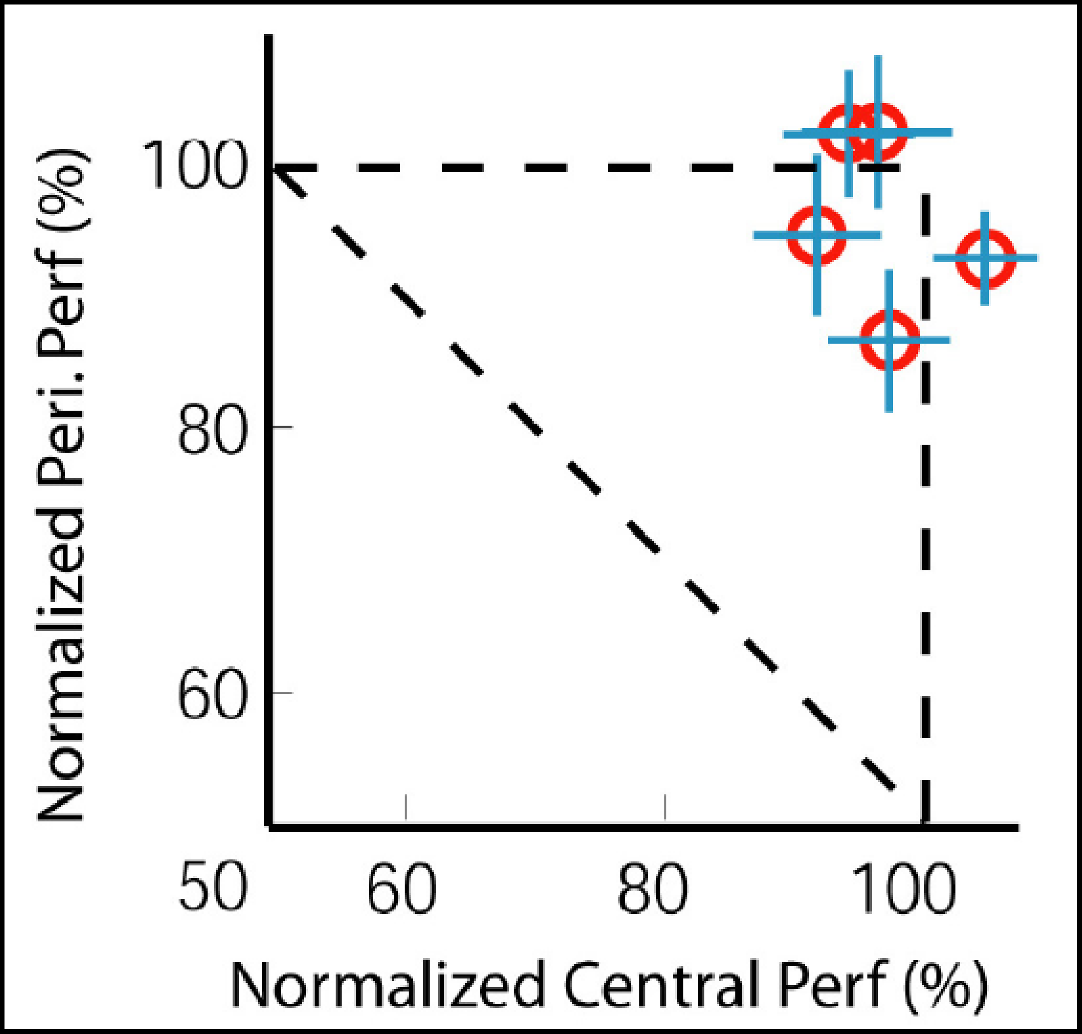

vs.
(masked by)
normalized peri. perf. (%)
100
80
60
40
40
60
80
100
normalized central perf. (%)

Are animals special?


normalized perl. perf. (%)
100
80
60
40
40
60
80
100
normalized central perf. (%)

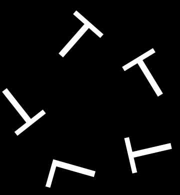

normalized perl. perf. (%)
100
80
60
40
40
60
80
100
normalized central perf. (%)

Without color…

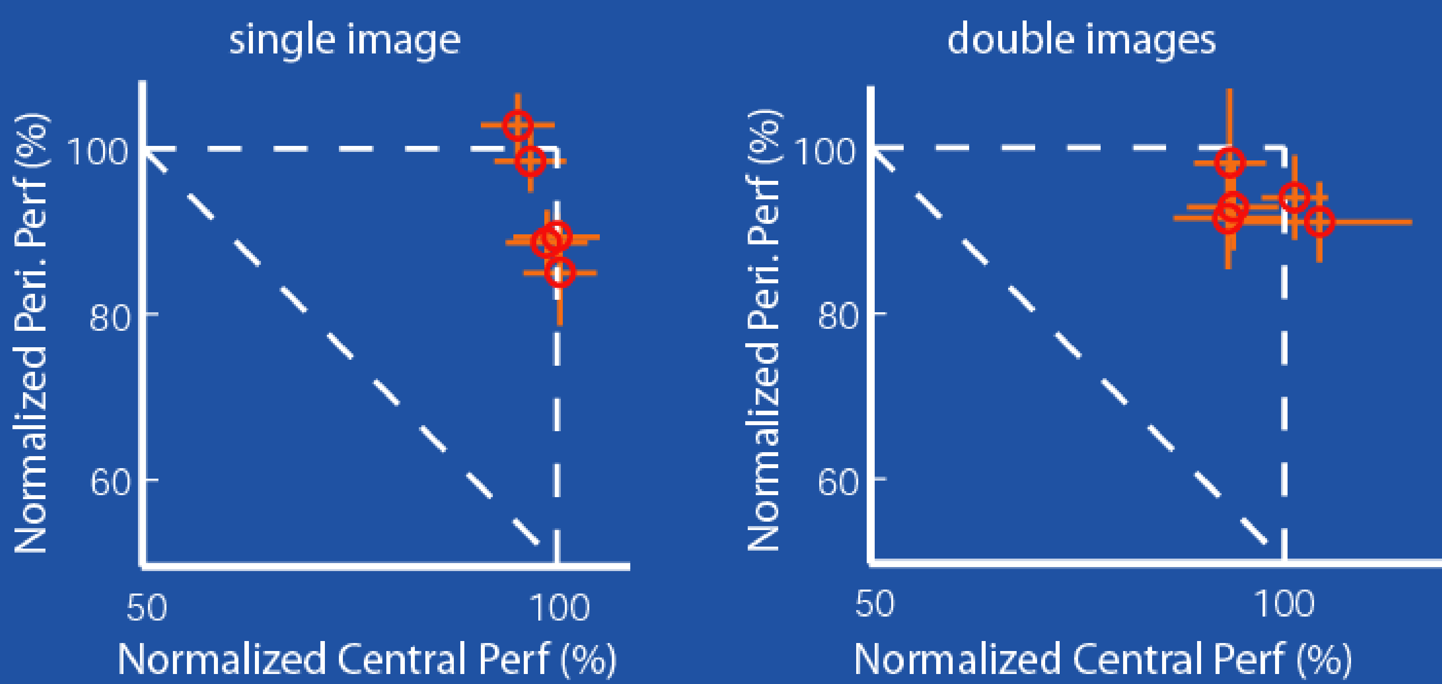


Fei-Fei et al. *Vis. Cog.*, 2004


SMITH
Middletown
Honda
343-0055
PEPSI
LIBERTY

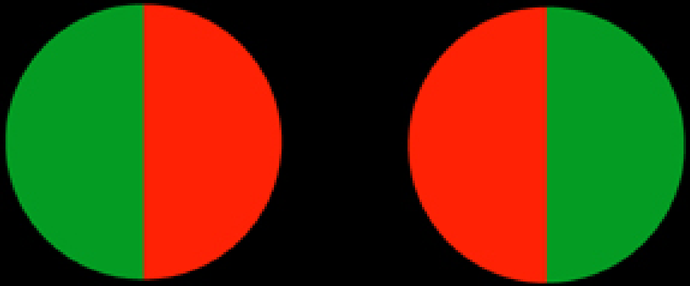


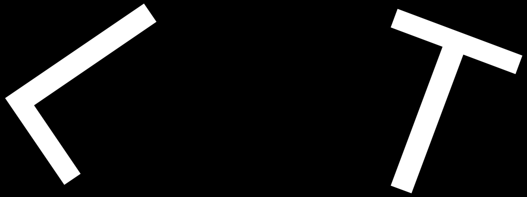
Categorization without attention: Single Image vs. Double Images



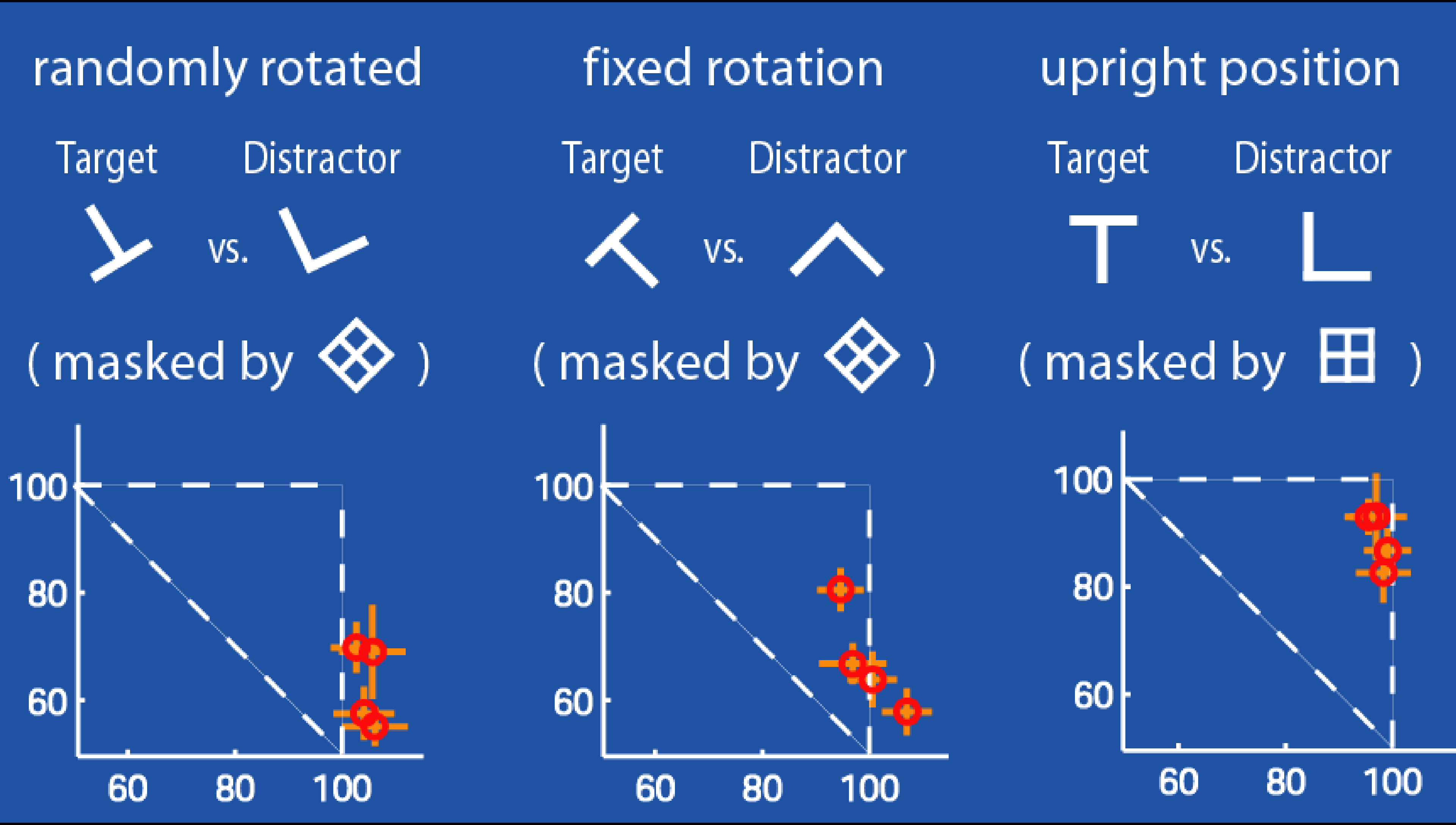
Fei-Fei et al. *Vis. Cog.*, 2004


NOAA

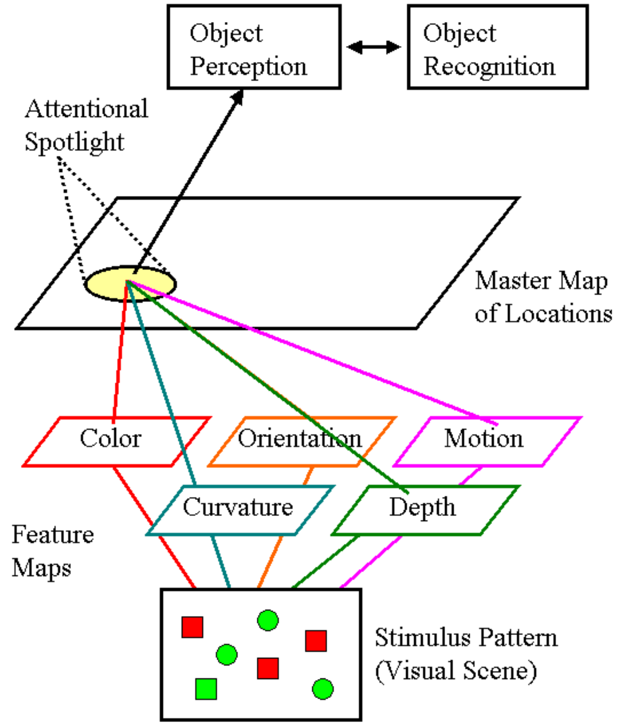




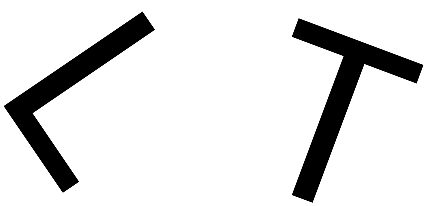
Effect of “meaningful” category

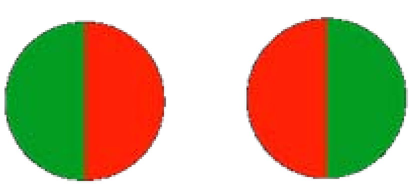


F.I.T. predicted…

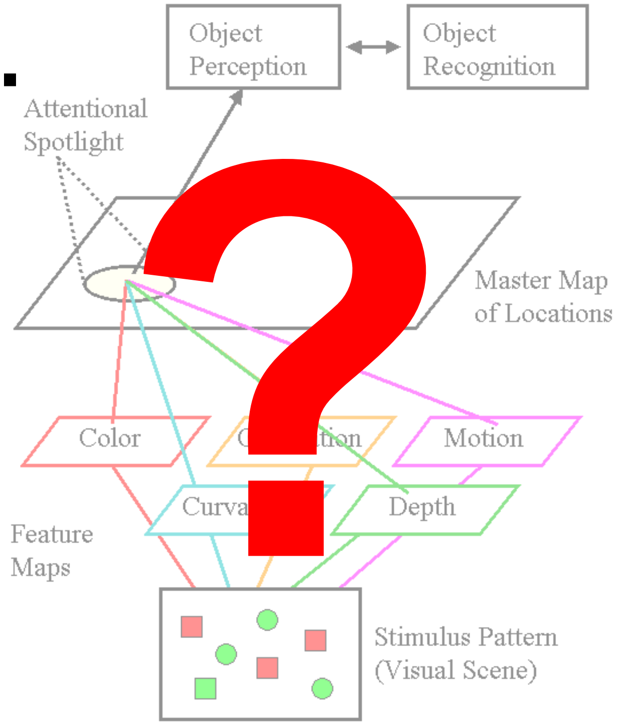


less attentional load more

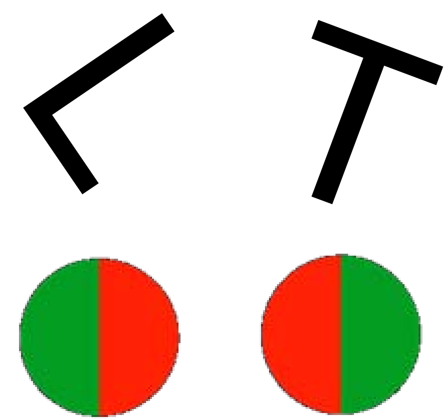




Our data shows…



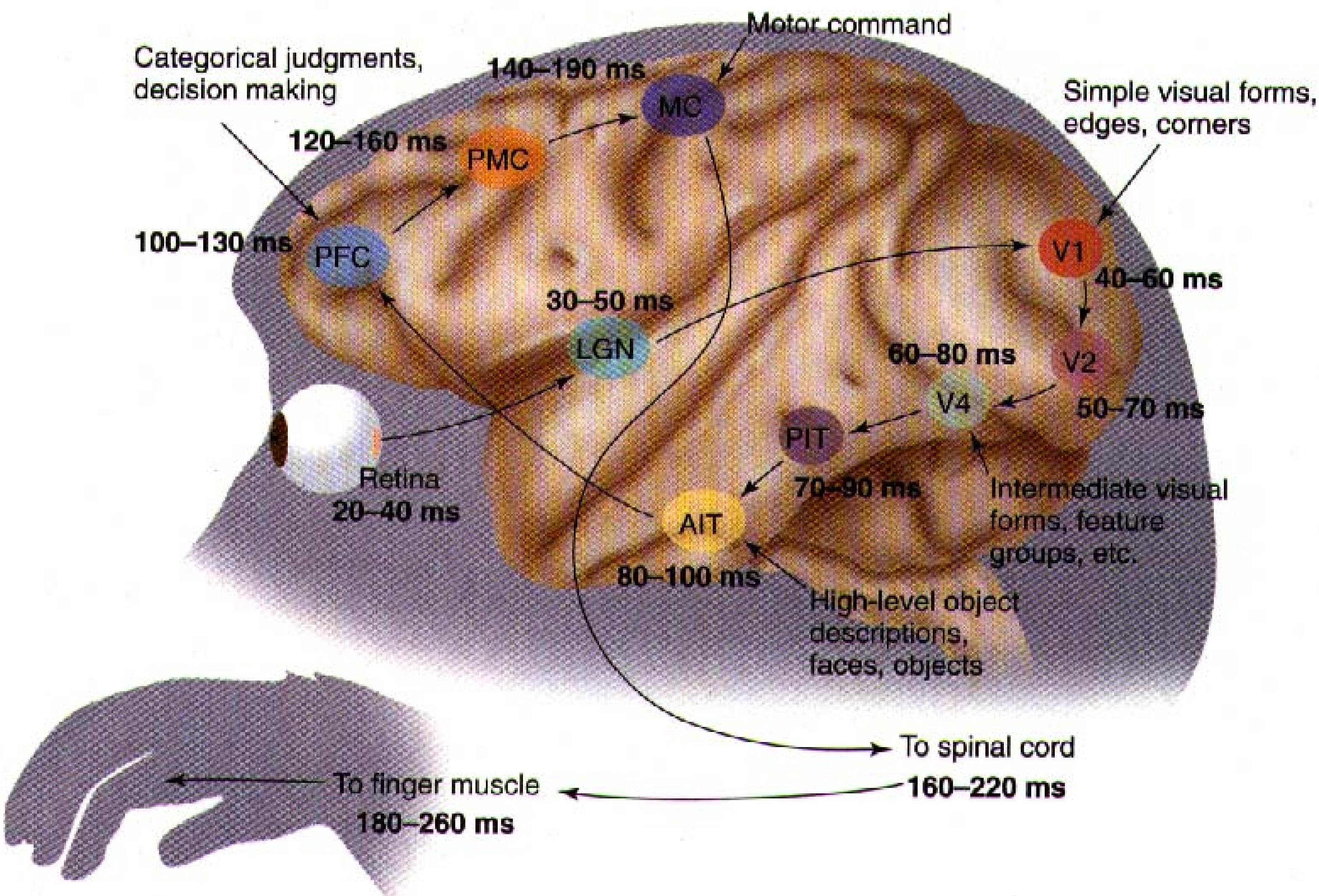
less **attentional load** **more**



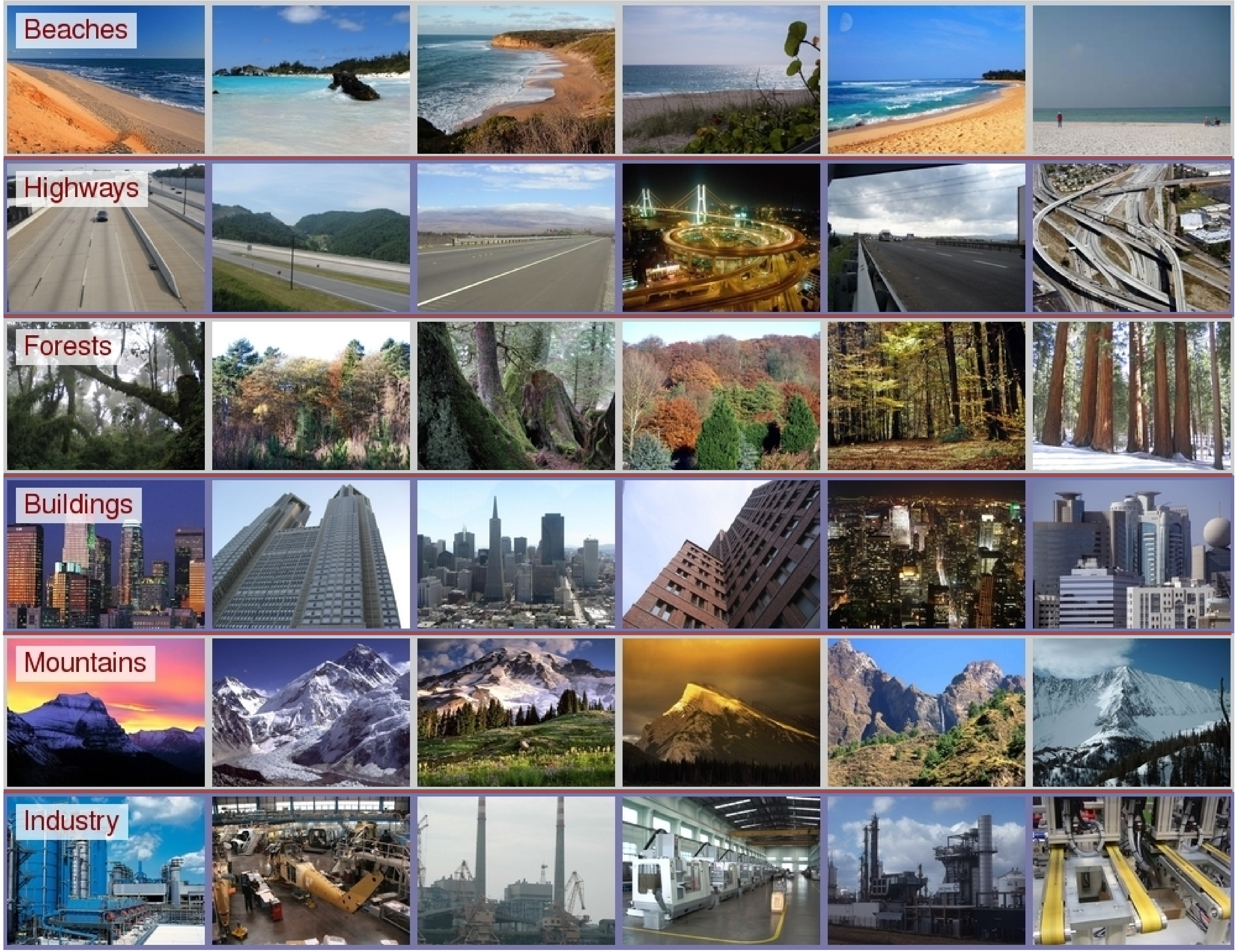
Li et al. 2002; Fei-Fei et al. 2005

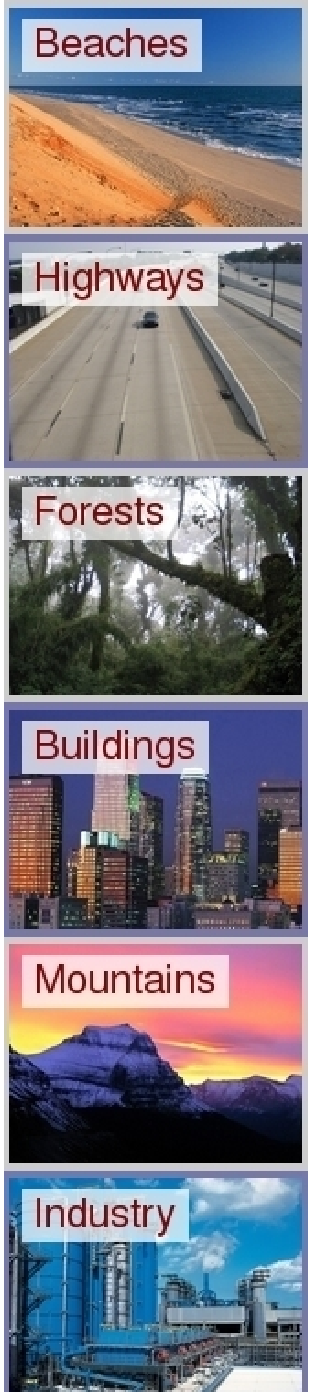
Rapid Perception of Natural Scenes

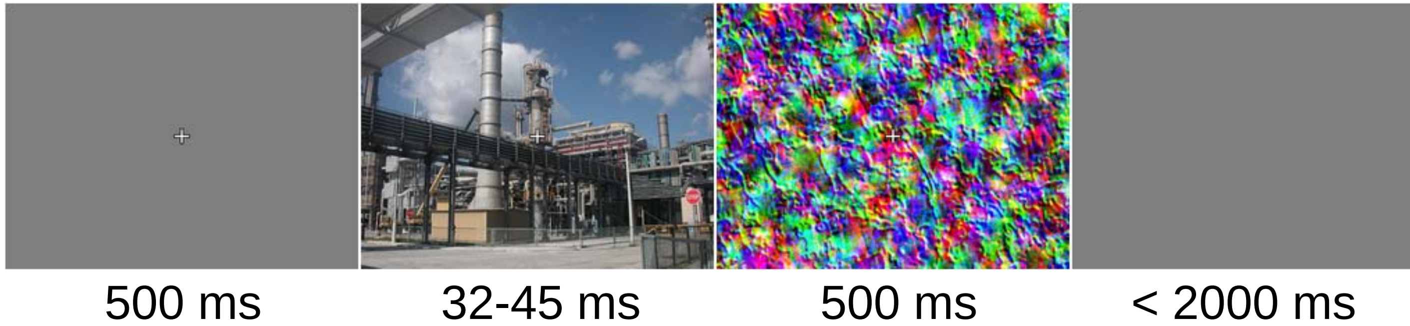
- Where/how does this happen?

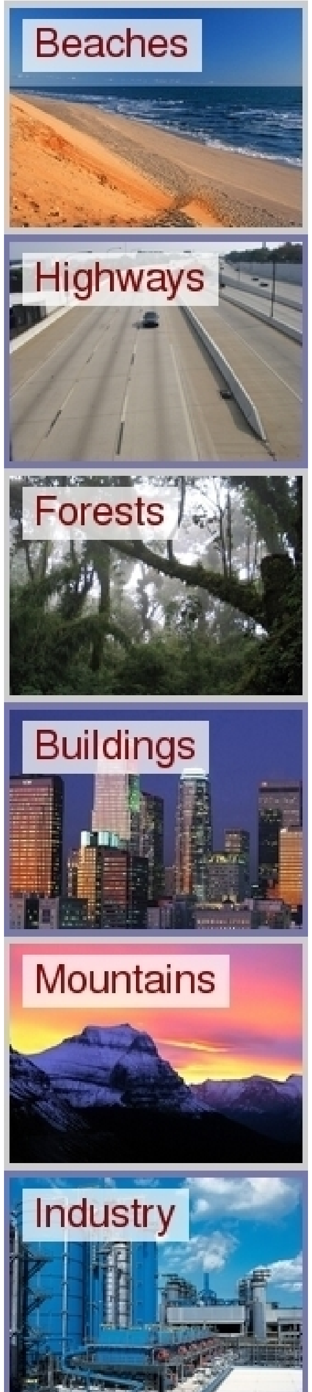


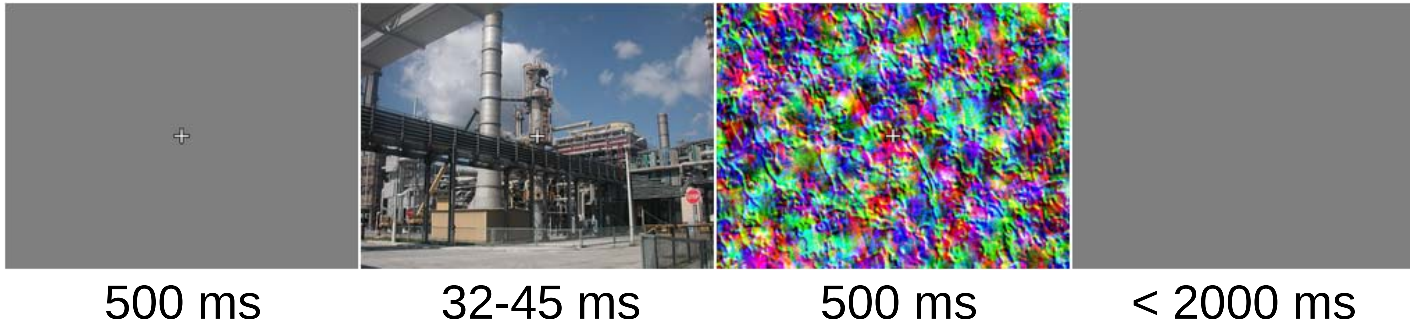
Thorpe, et al. *Science, 2001*

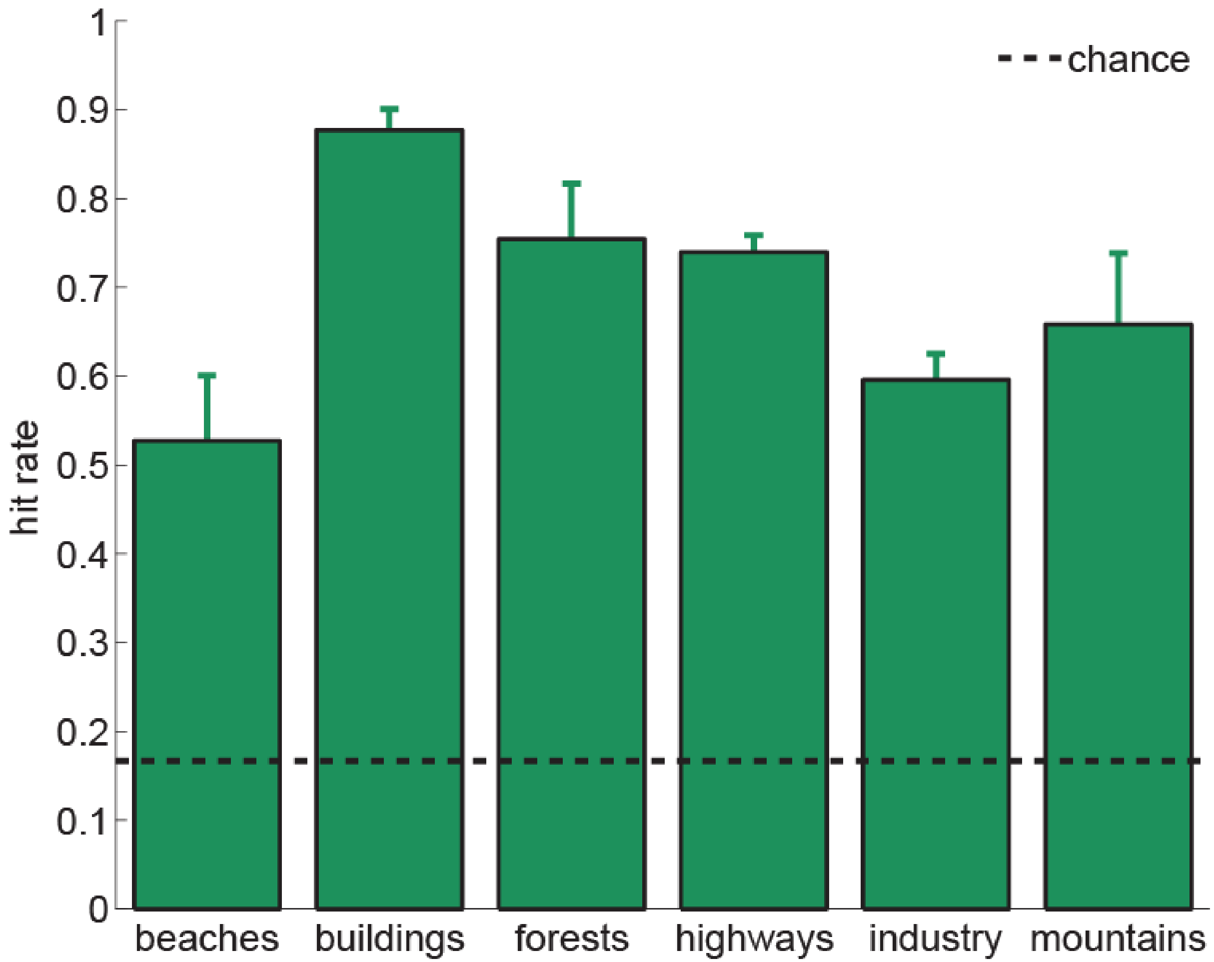

Beaches
Highways
Forests
Buildings
Mountains
Industry


Beaches
Highways
Forests
Buildings
Mountains
Industry


500 ms
32-45 ms
500 ms
< 2000 ms







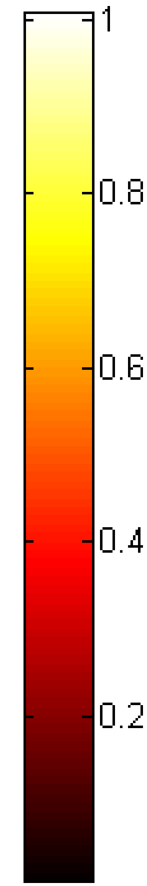
6 AFC, N = 4, error bars: s.e.m.

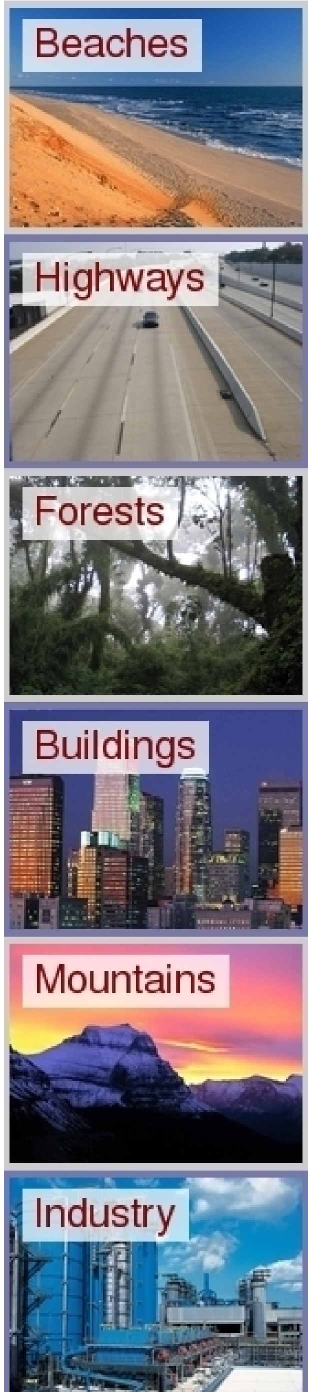
Behavioral Performance

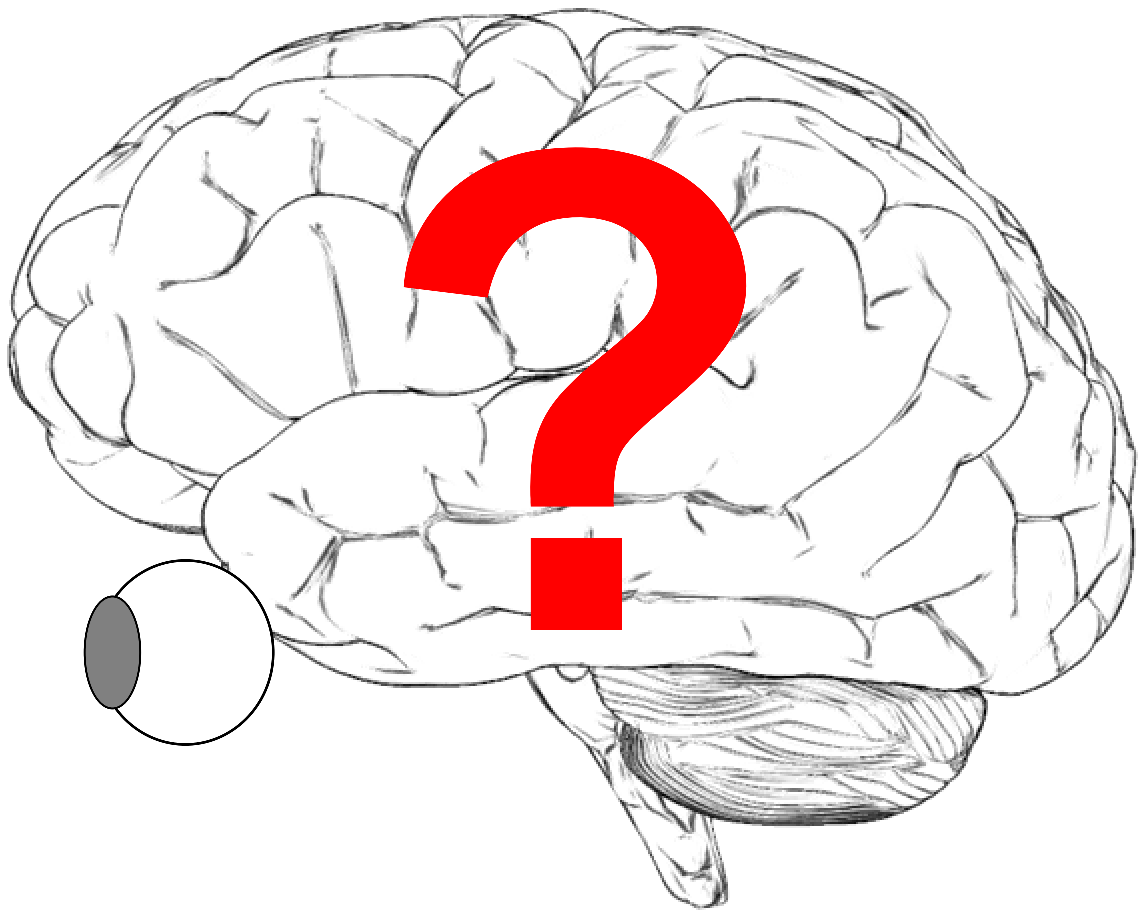
chance: 0.167

Subjects' response

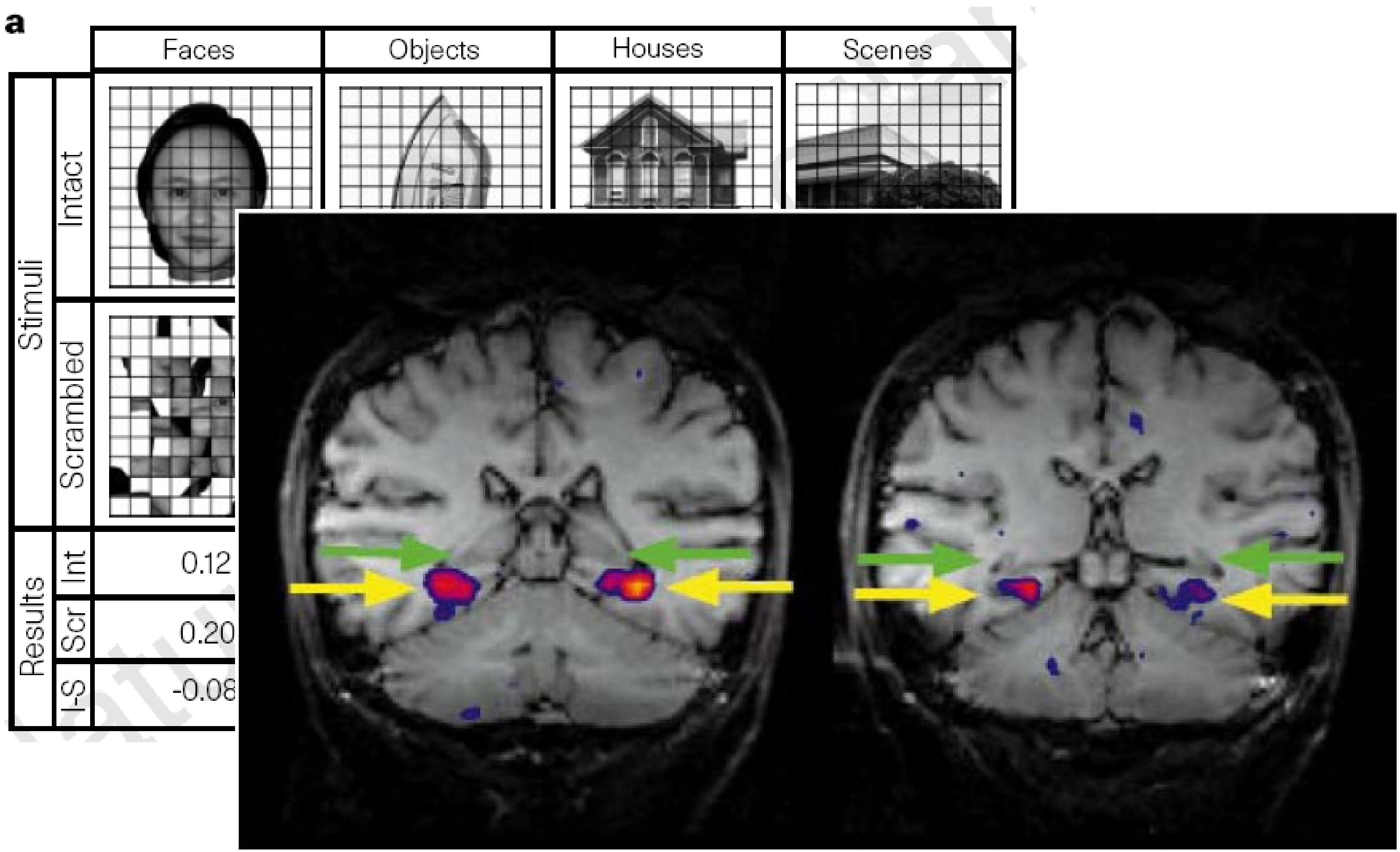
Viewed image category (ground truth)	beaches	buildings	forests	highways	industry	mountains
beaches	0.53	0.03	0.05	0.12	0.05	0.08
buildings	0.02	0.88	0.02	0.01	0.03	0.01
forests	0.03	0.05	0.75	0.01	0.02	0.05
highways	0.04	0.05	0.02	0.74	0.04	0.04
industry	0.03	0.18	0.05	0.04	0.60	0.03
mountains	0.07	0.05	0.07	0.04	0.03	0.66




Beaches
Highways
Forests
Buildings
Mountains
Industry

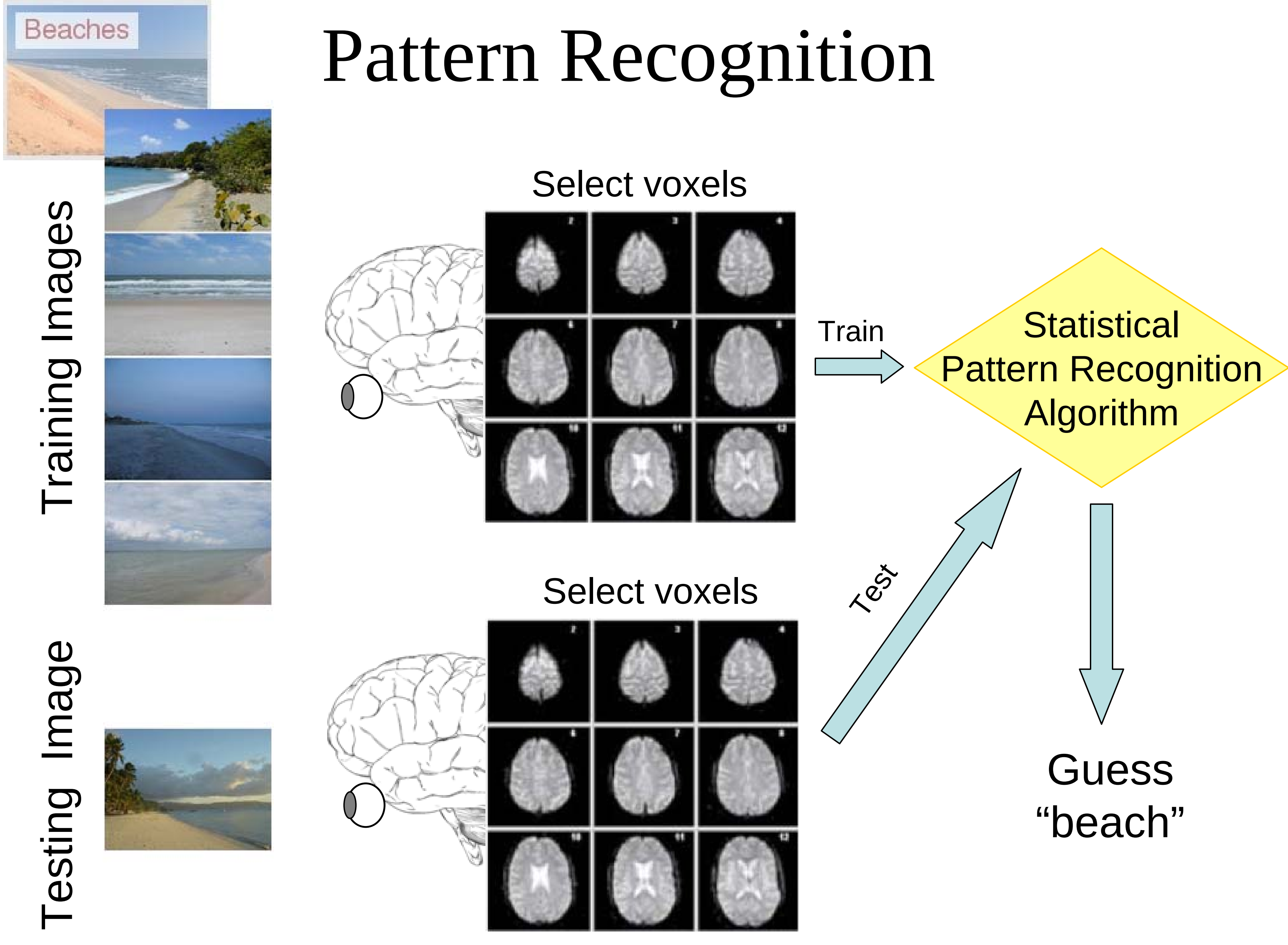

?

PPA: Parahippocampal Place Area

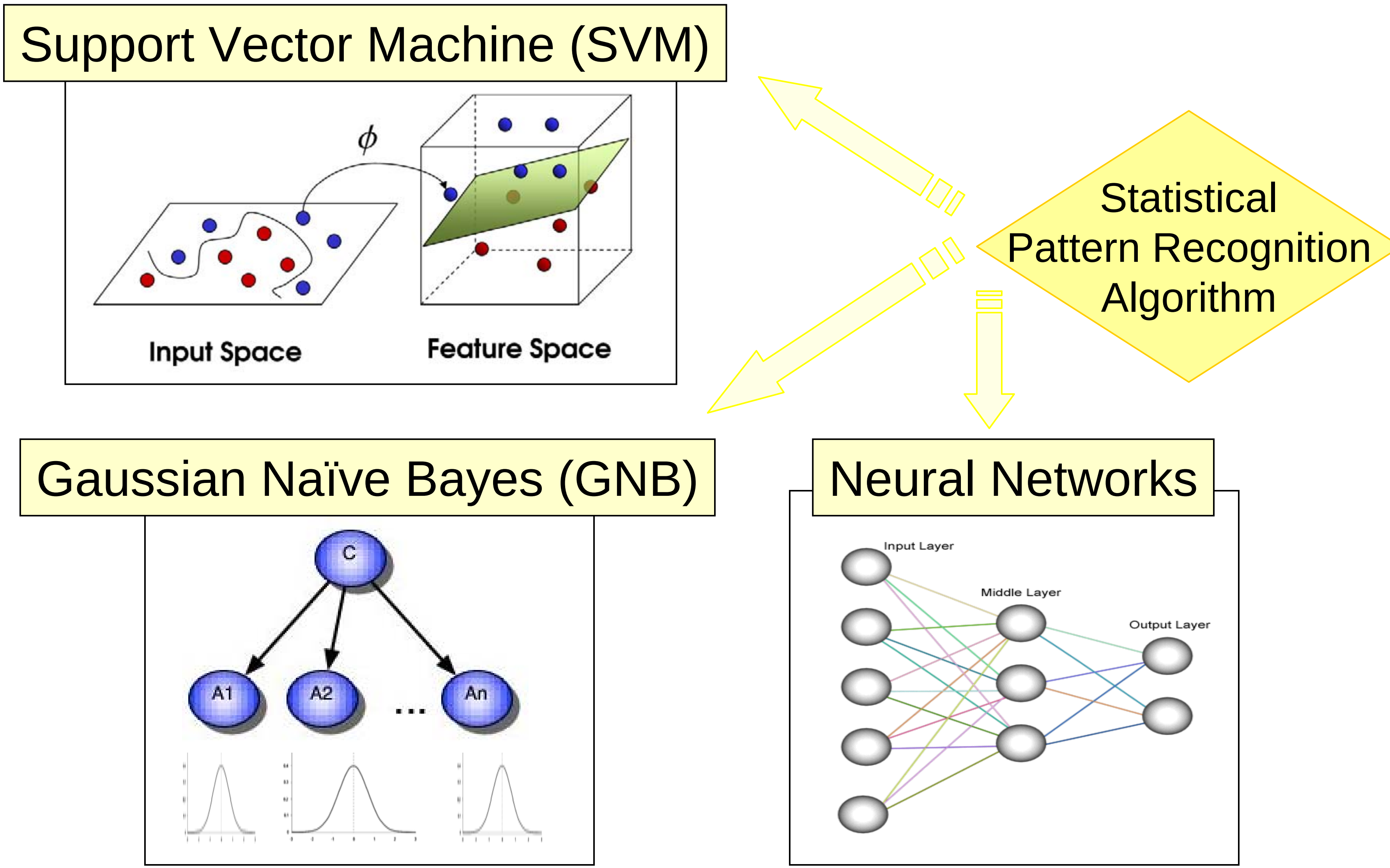


Epstein & Kanwisher, 1998

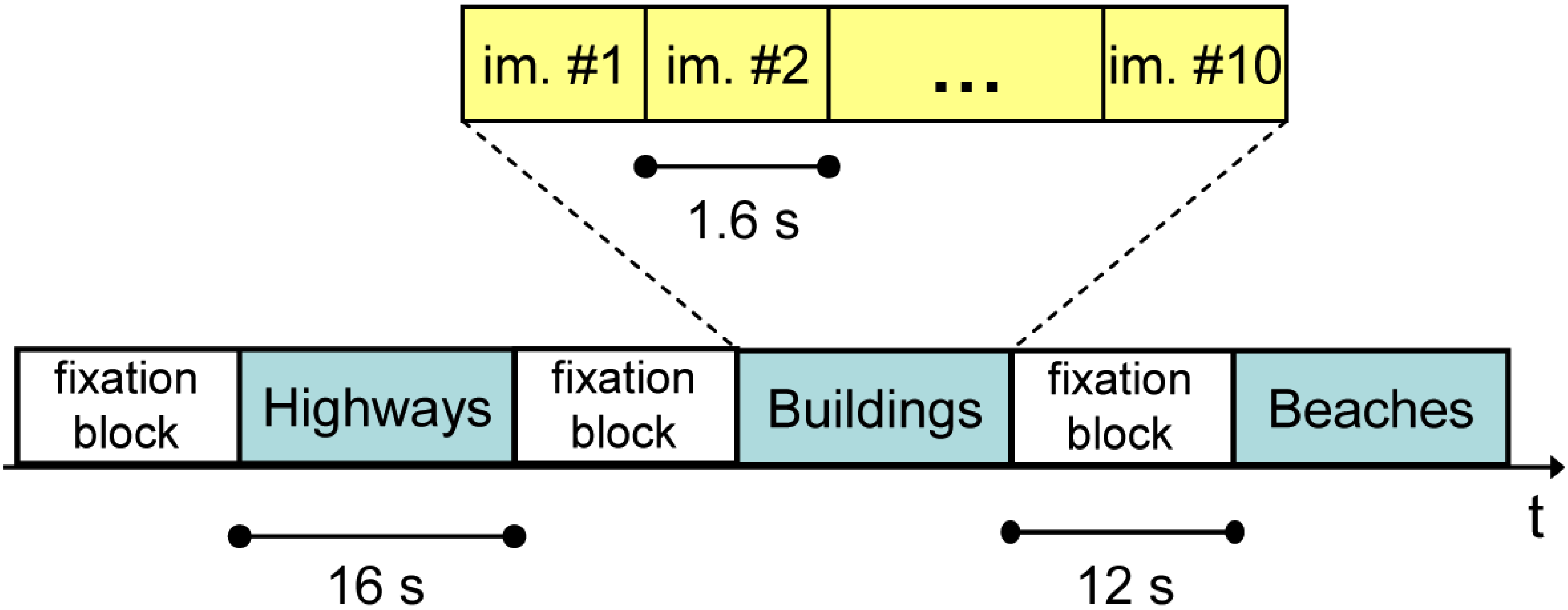
Pattern Recognition



Pattern Recognition

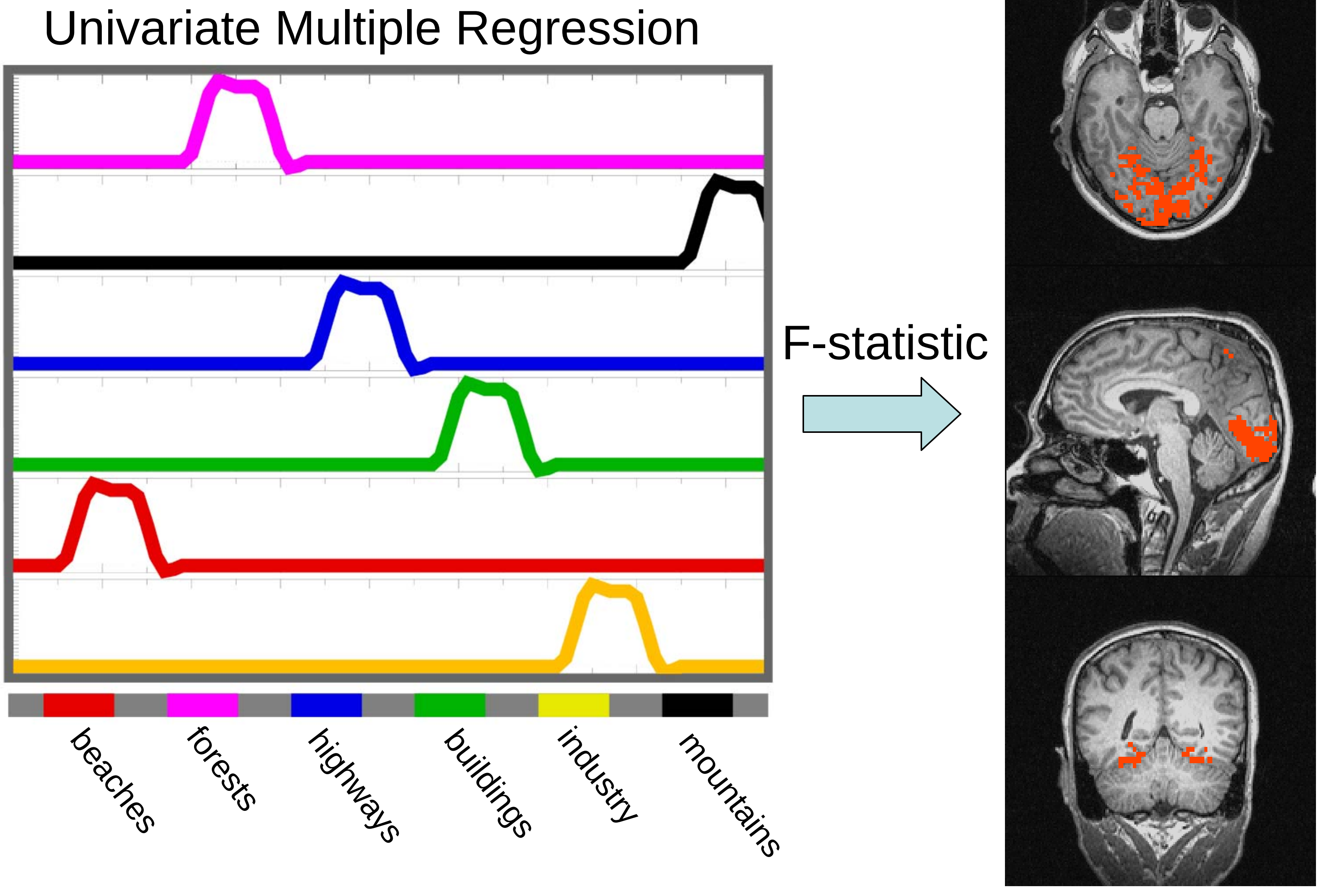


Experimental Setup (fMRI)

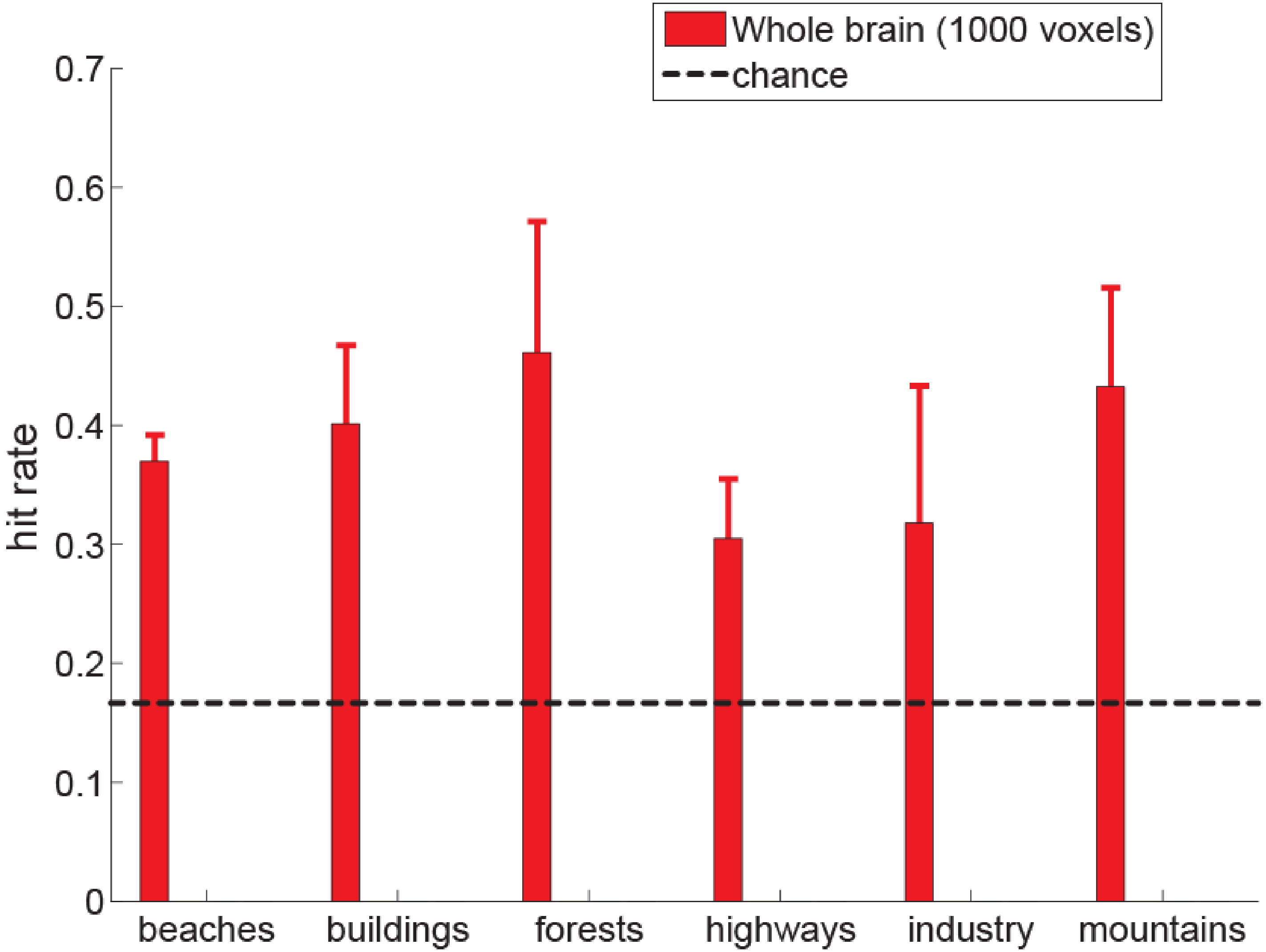


- 6 blocks per run (all 6 categories)
- 12 runs for each subject
- Alternating runs feature upright or inverted images

Voxel Selection

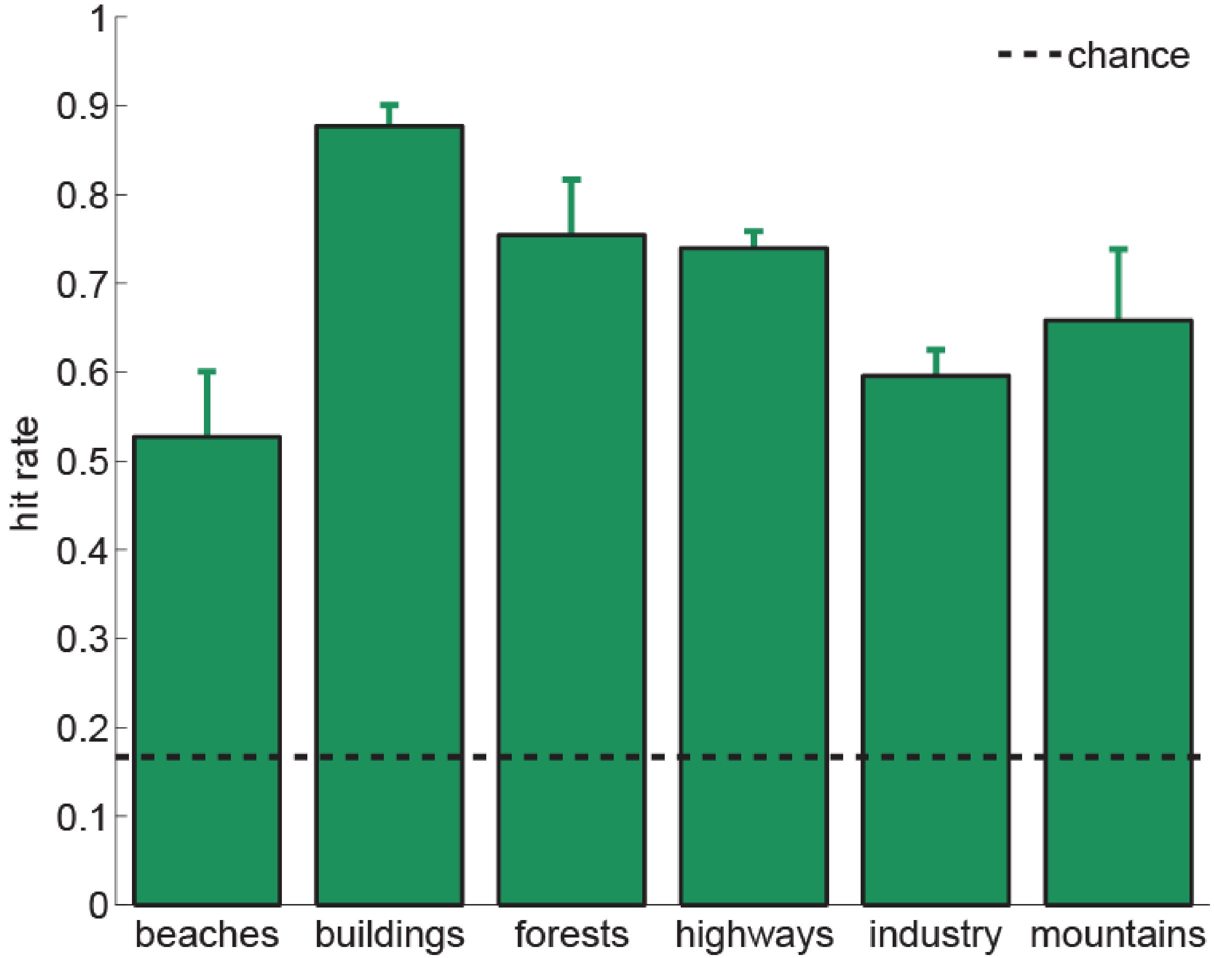


Decoding Performance

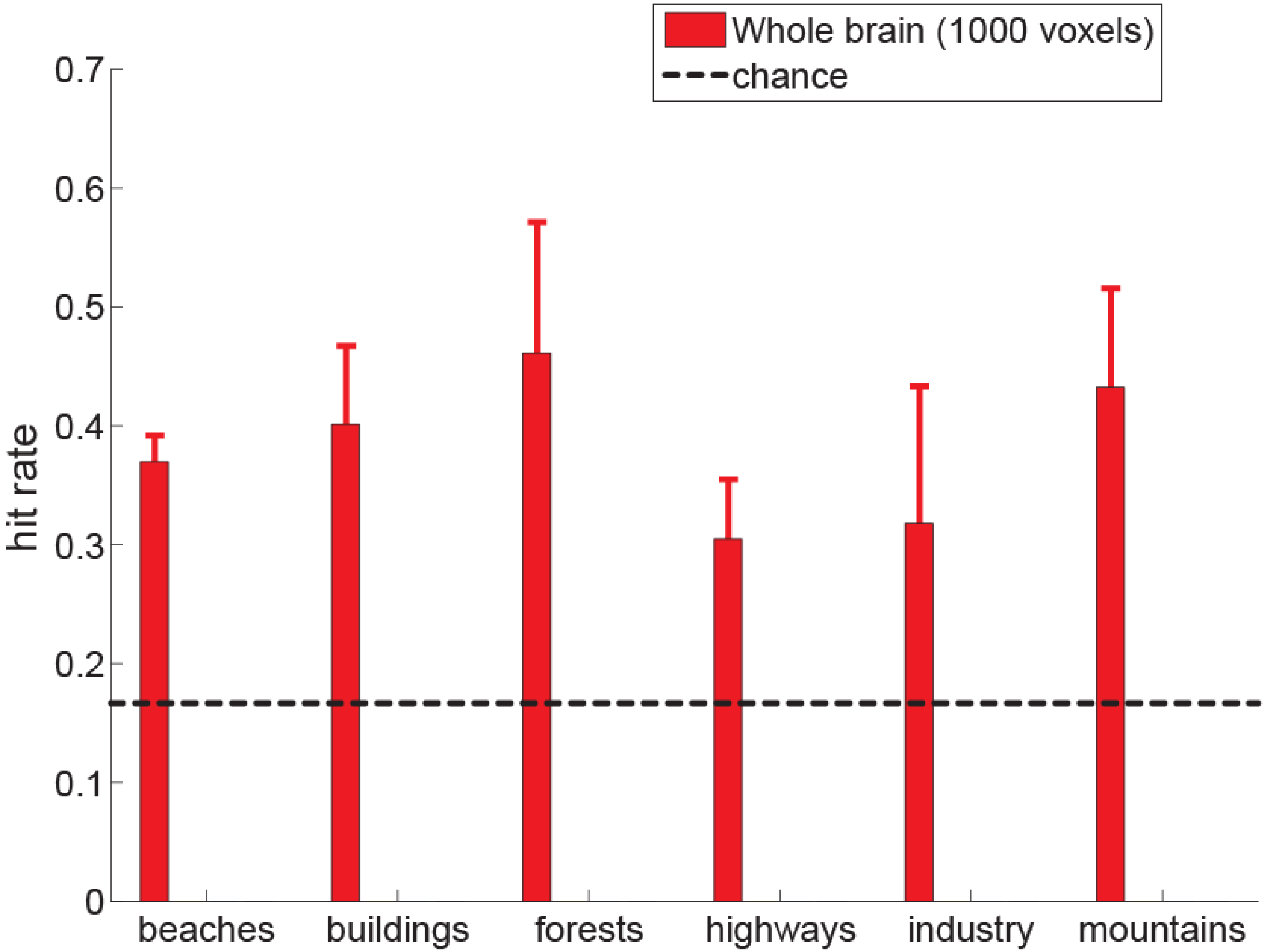


N = 4 error bars: s.e.m.

Decoding Performance



Decoding Performance



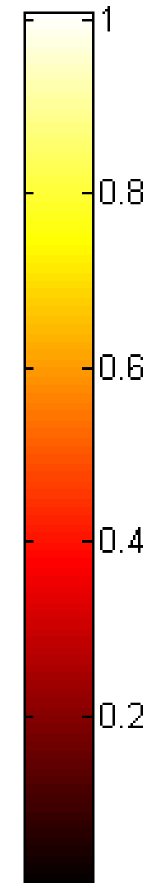
N = 4 error bars: s.e.m.

Decoding Performance

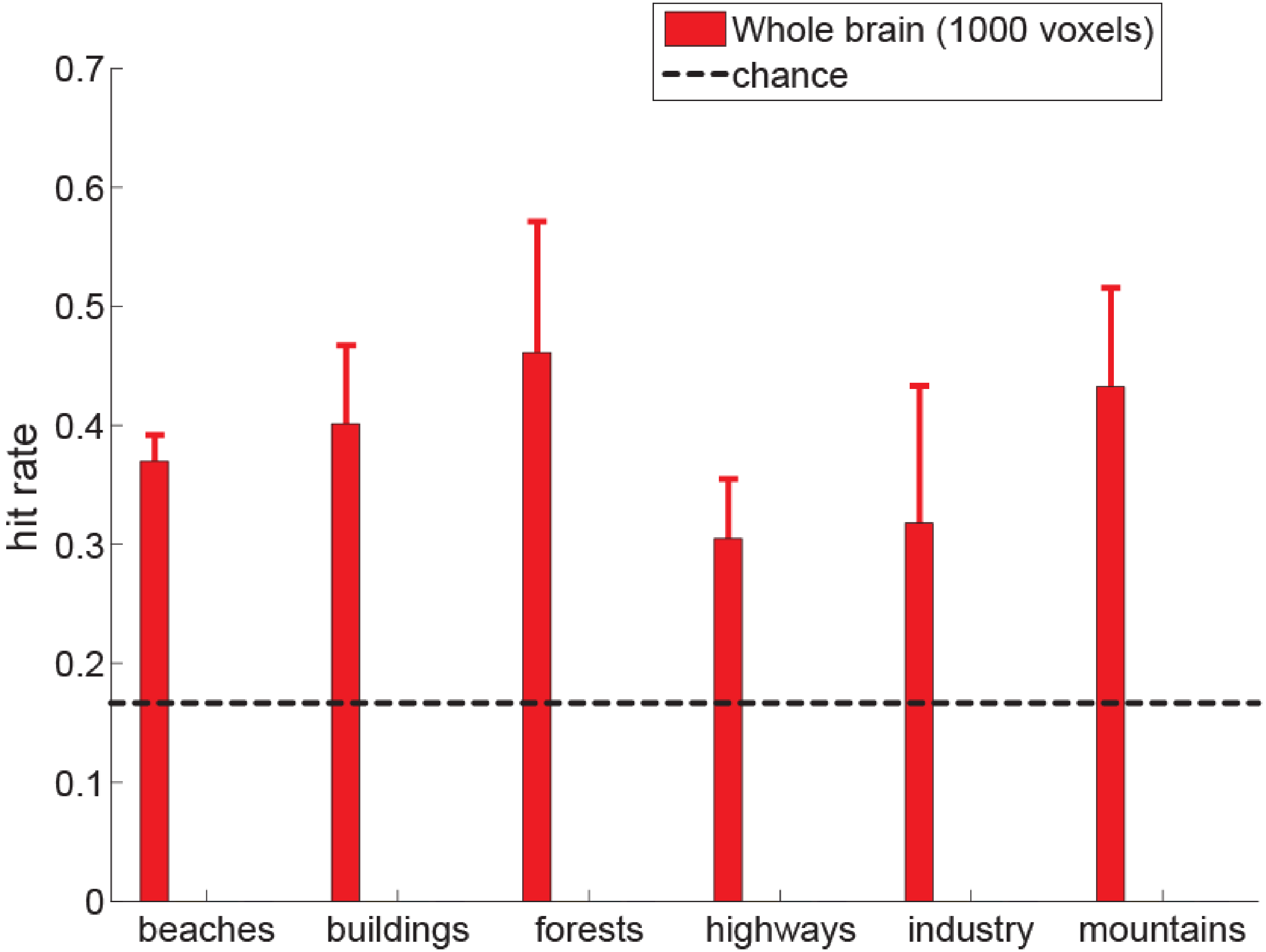
chance: 0.167

Classifier prediction

Viewed image category (ground truth)	beaches	buildings	forests	highways	industry	mountains
beaches	0.37	0.13	0.05	0.26	0.09	0.10
buildings	0.10	0.40	0.11	0.11	0.16	0.12
forests	0.08	0.15	0.46	0.05	0.11	0.15
highways	0.27	0.11	0.07	0.30	0.09	0.16
industry	0.11	0.23	0.10	0.08	0.32	0.16
mountains	0.13	0.12	0.09	0.12	0.12	0.43

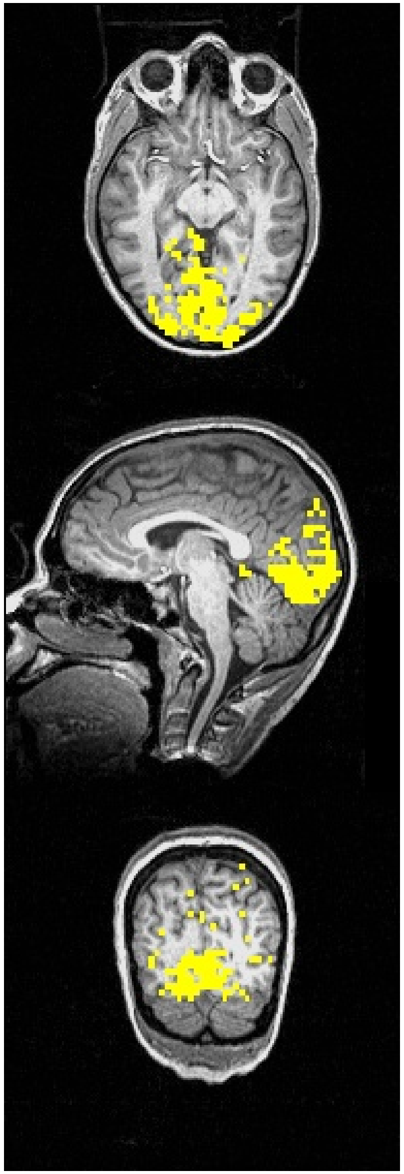


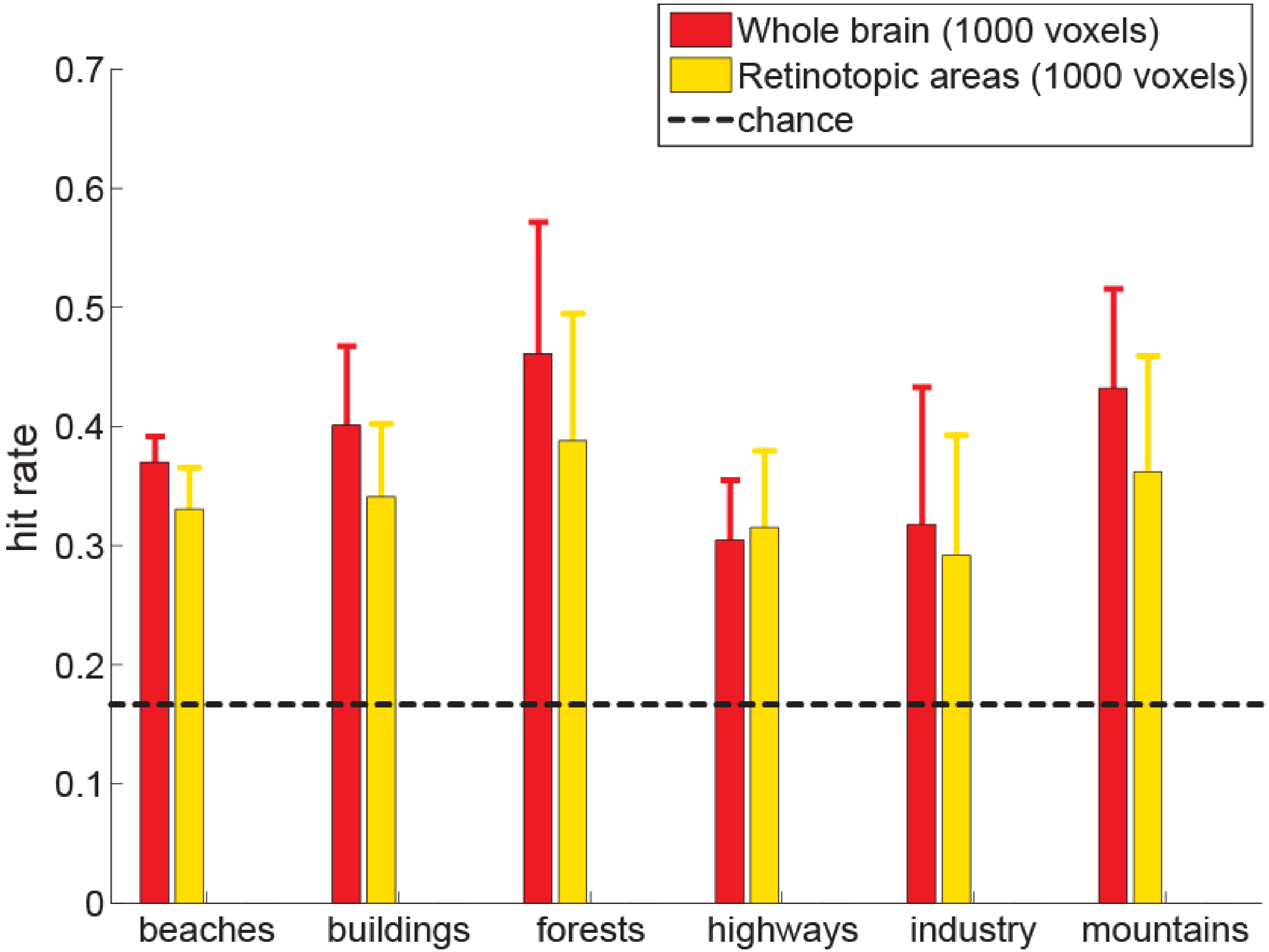
Decoding Performance



N = 4 error bars: s.e.m.

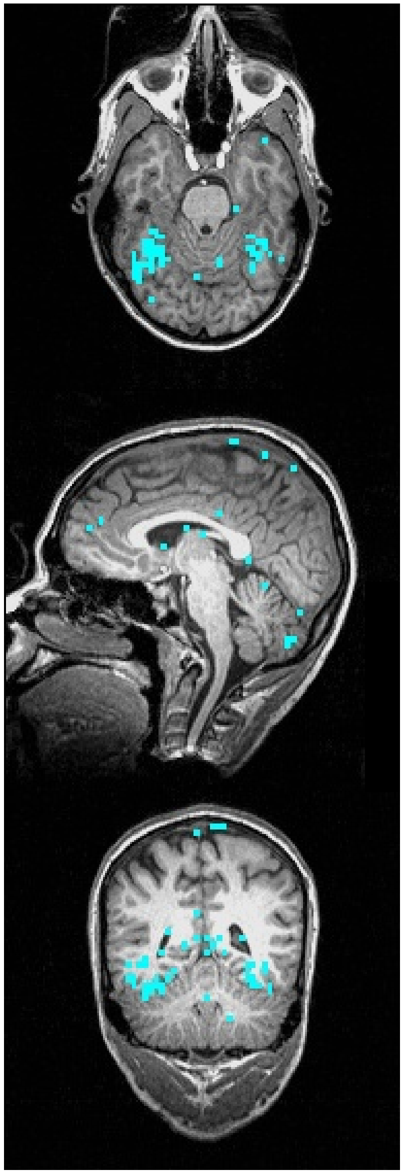
Retinotopic Areas

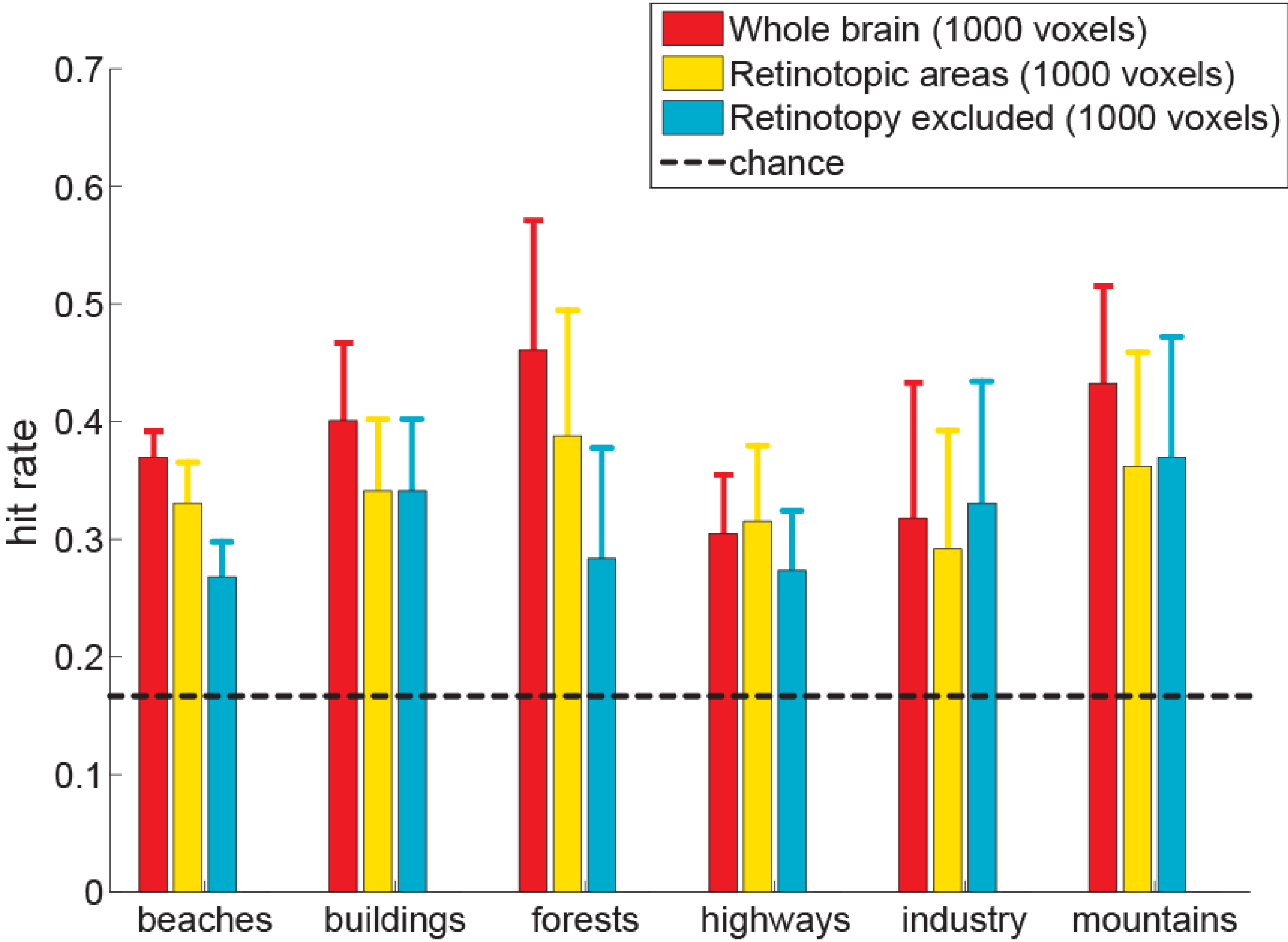




N = 4 error bars: s.e.m.

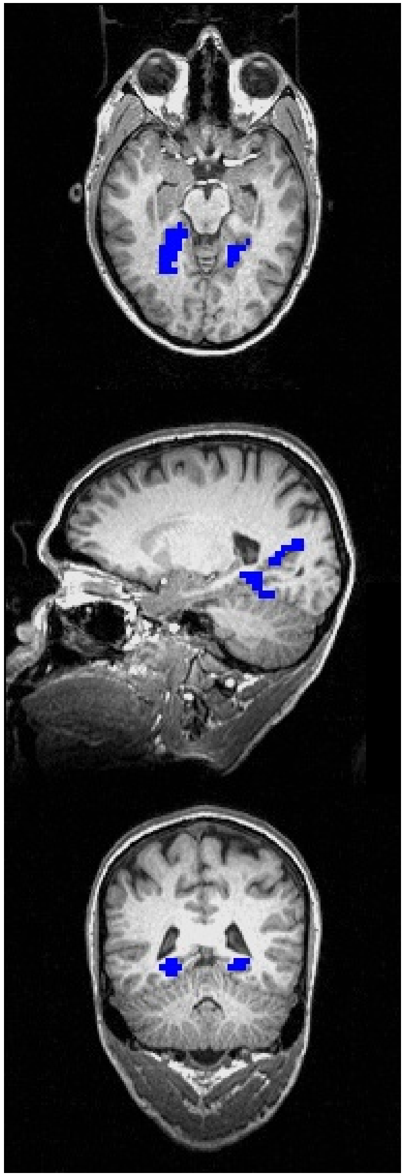
Retinotopic Areas Excluded

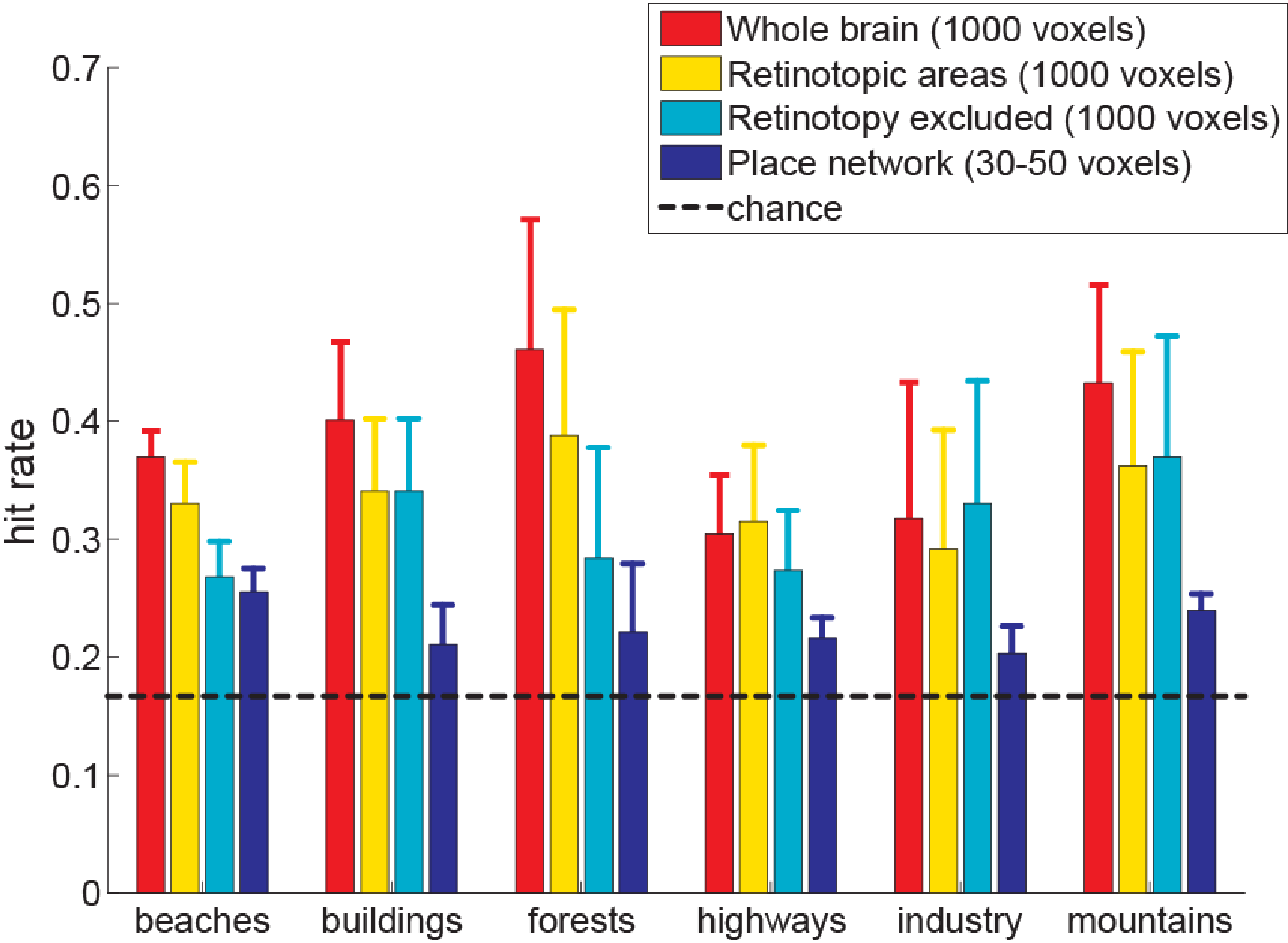




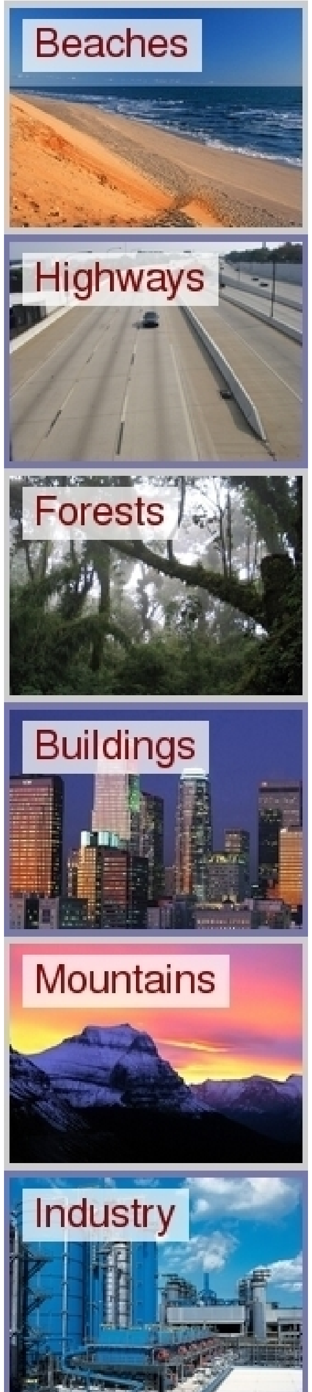
N = 4 error bars: s.e.m.

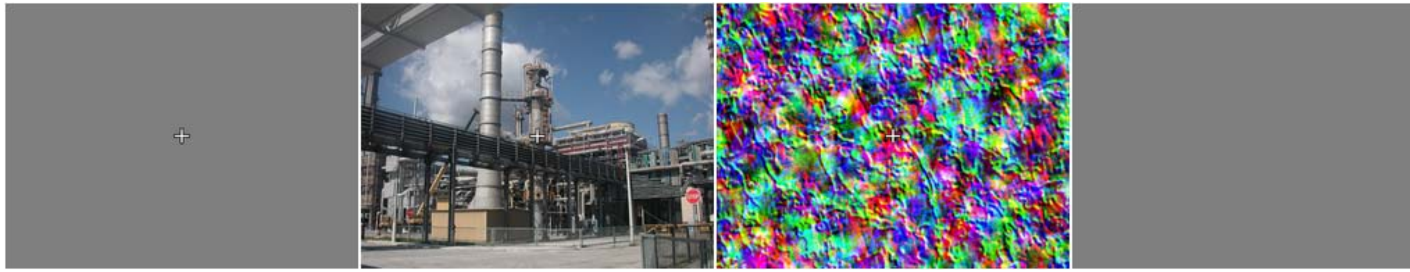
Place Network (PPA + RSC)





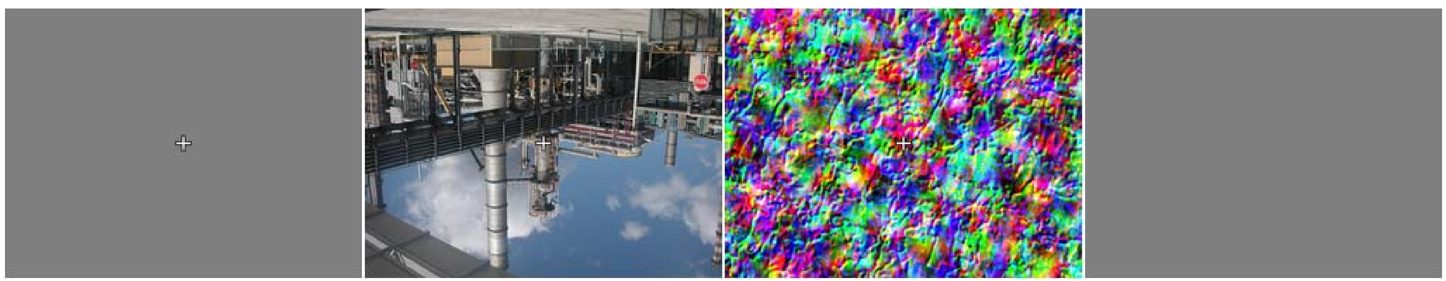
N = 4 error bars: s.e.m.





500 ms | 32-45 ms | 500 ms | < 2000 ms

Upright images

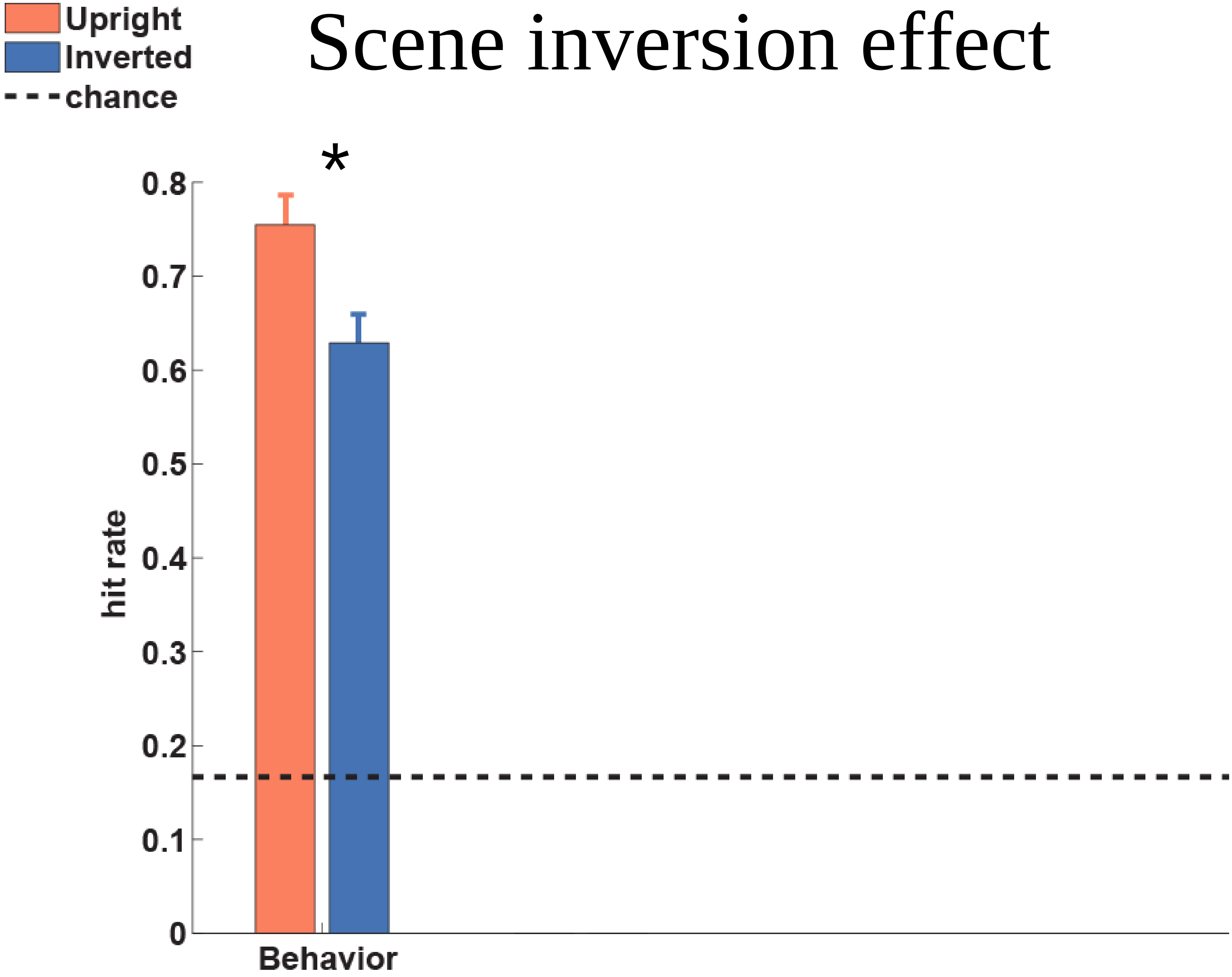


500 ms | 32-45 ms | 500 ms | < 2000 ms

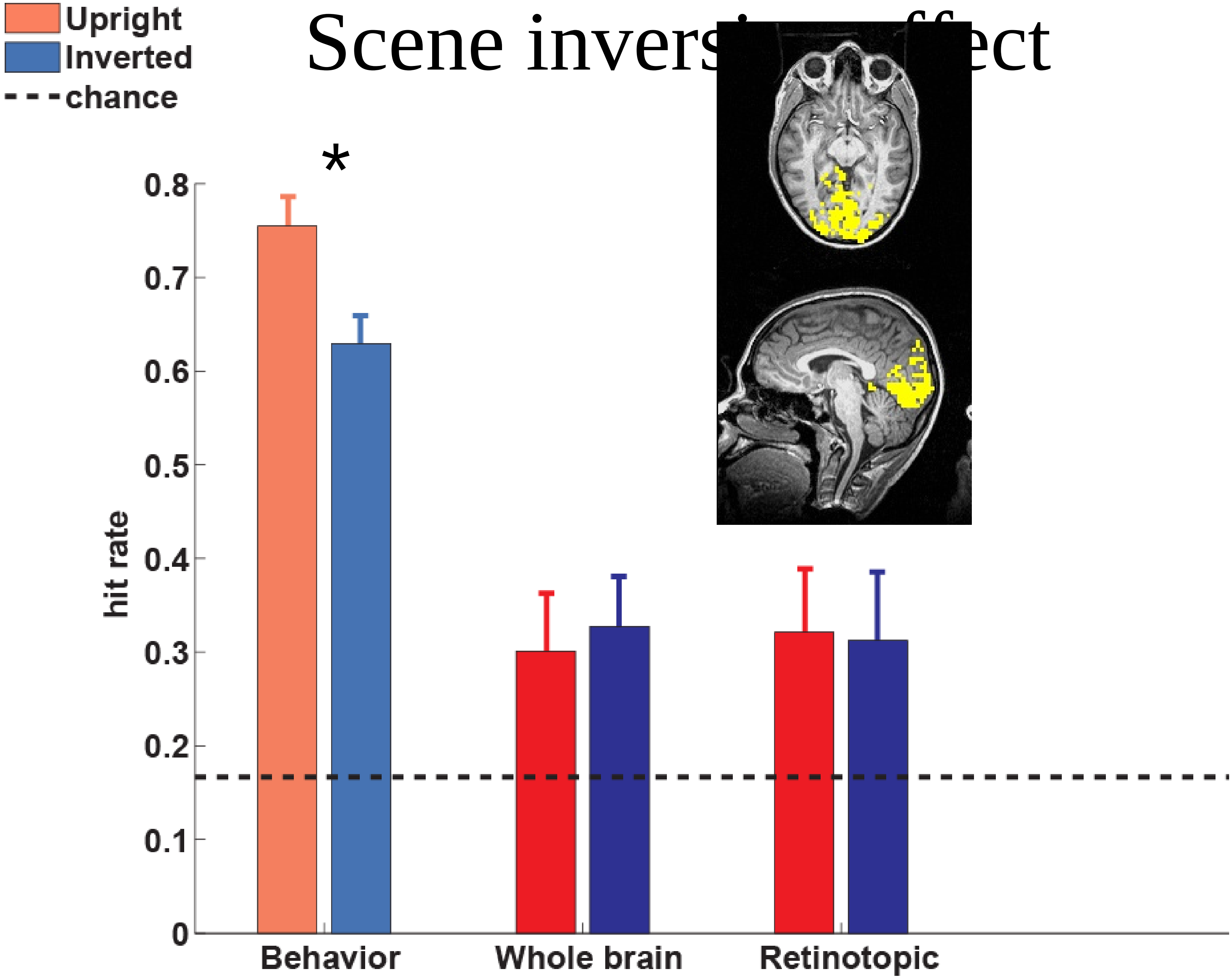
Inverted images

Training: upright only;
Testing: upright & inverted blocks intermixed

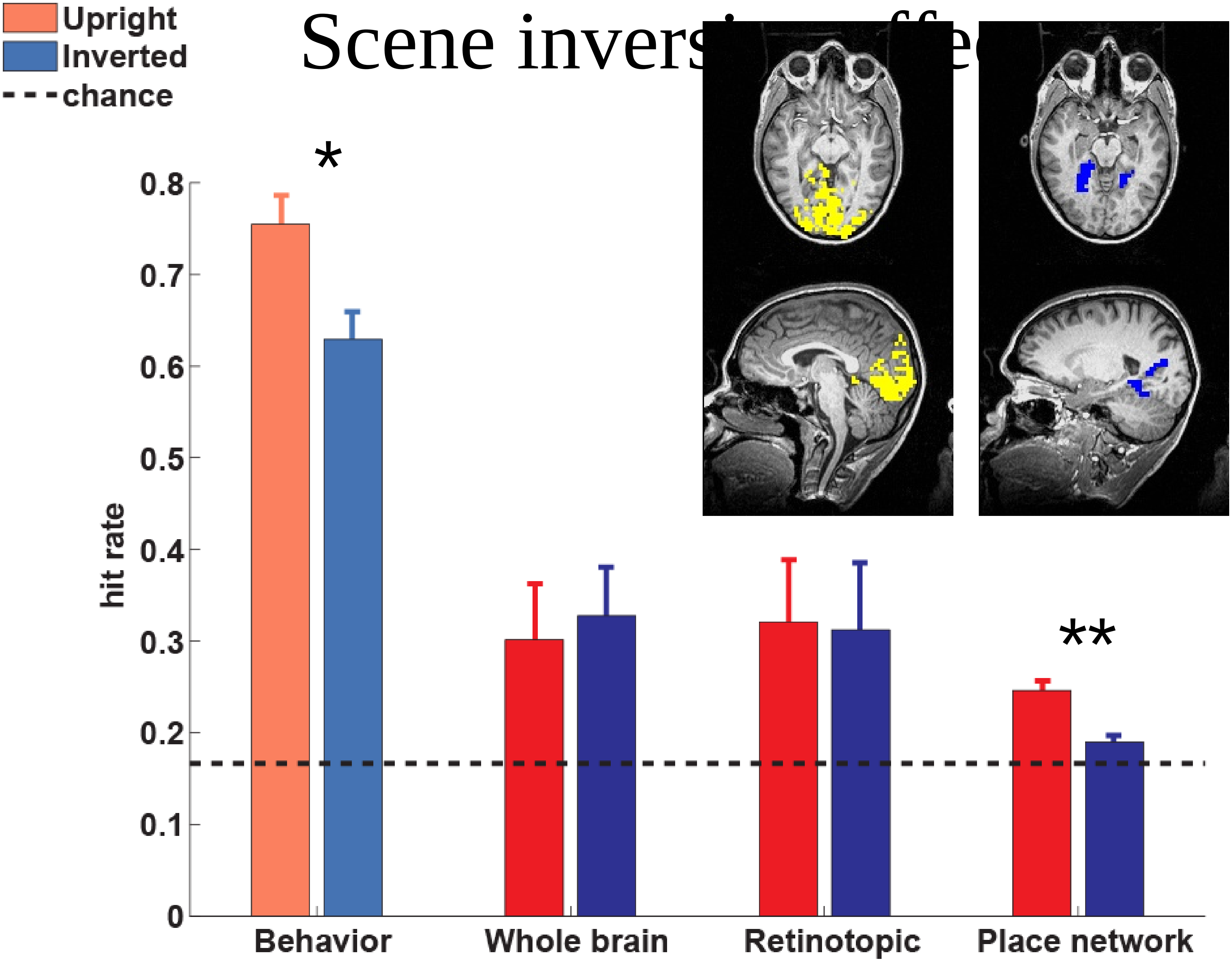
Scene inversion effect

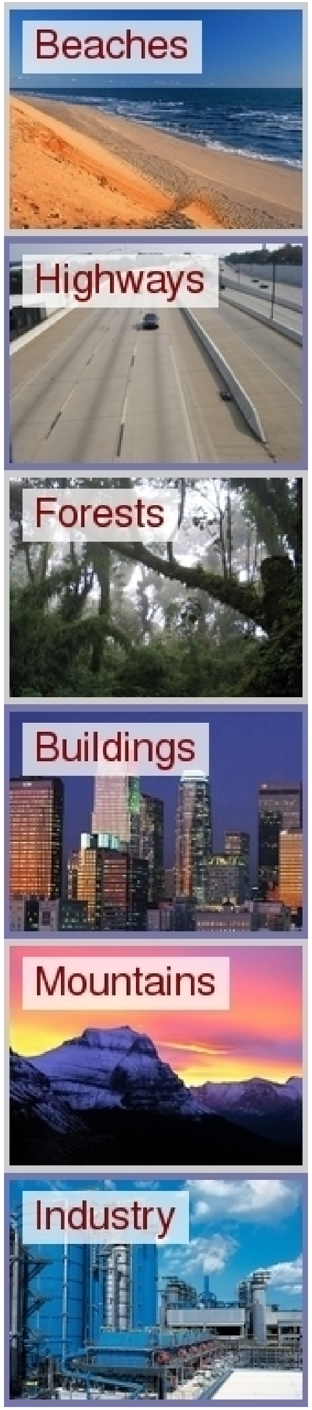


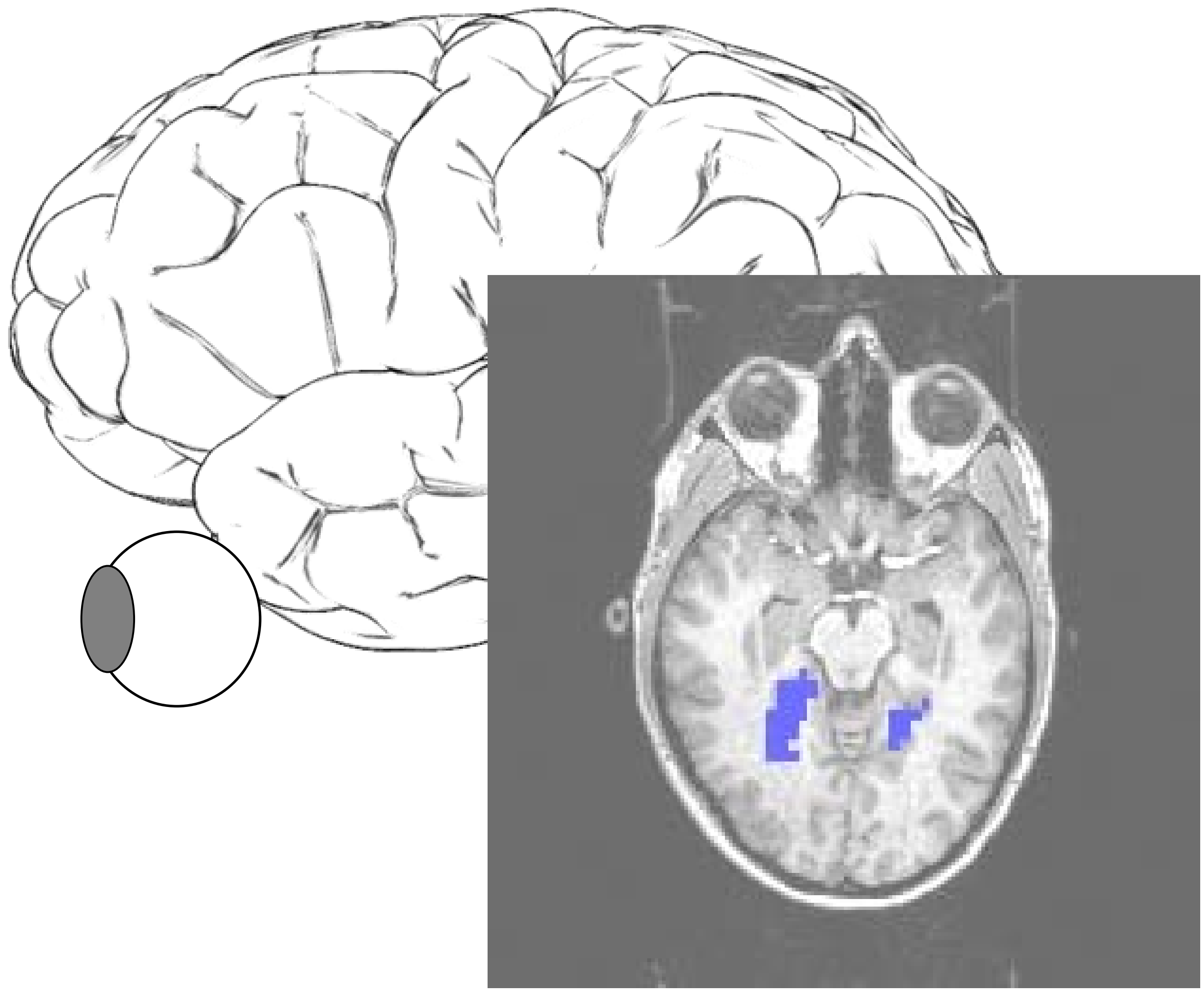
Scene invers ffect

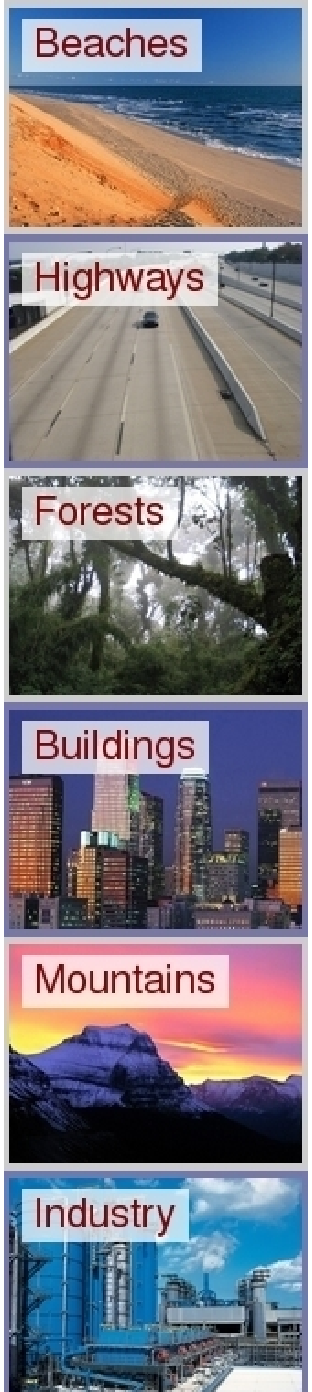


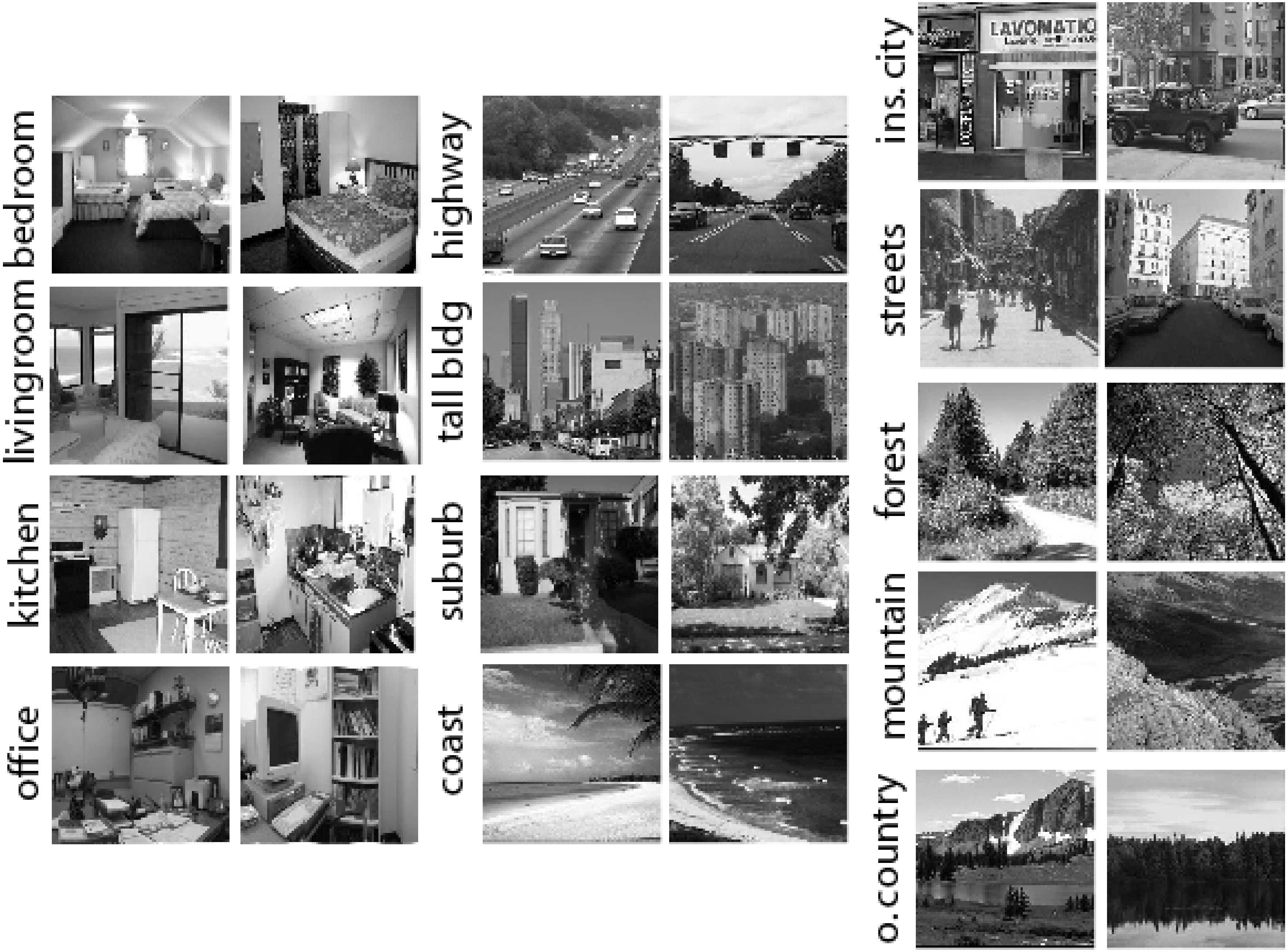
Scene invers



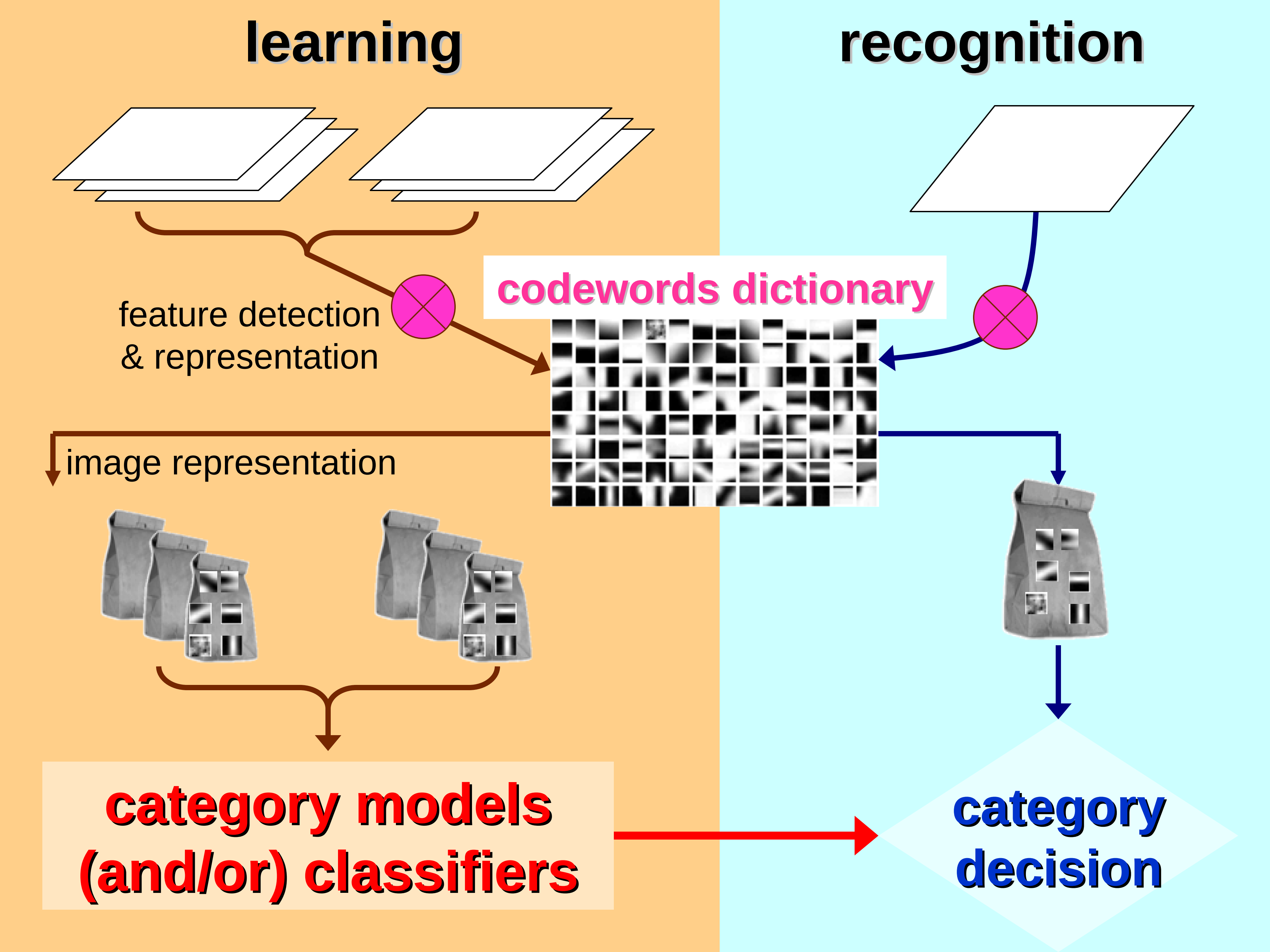

Beaches
Highways
Forests
Buildings
Mountains
Industry



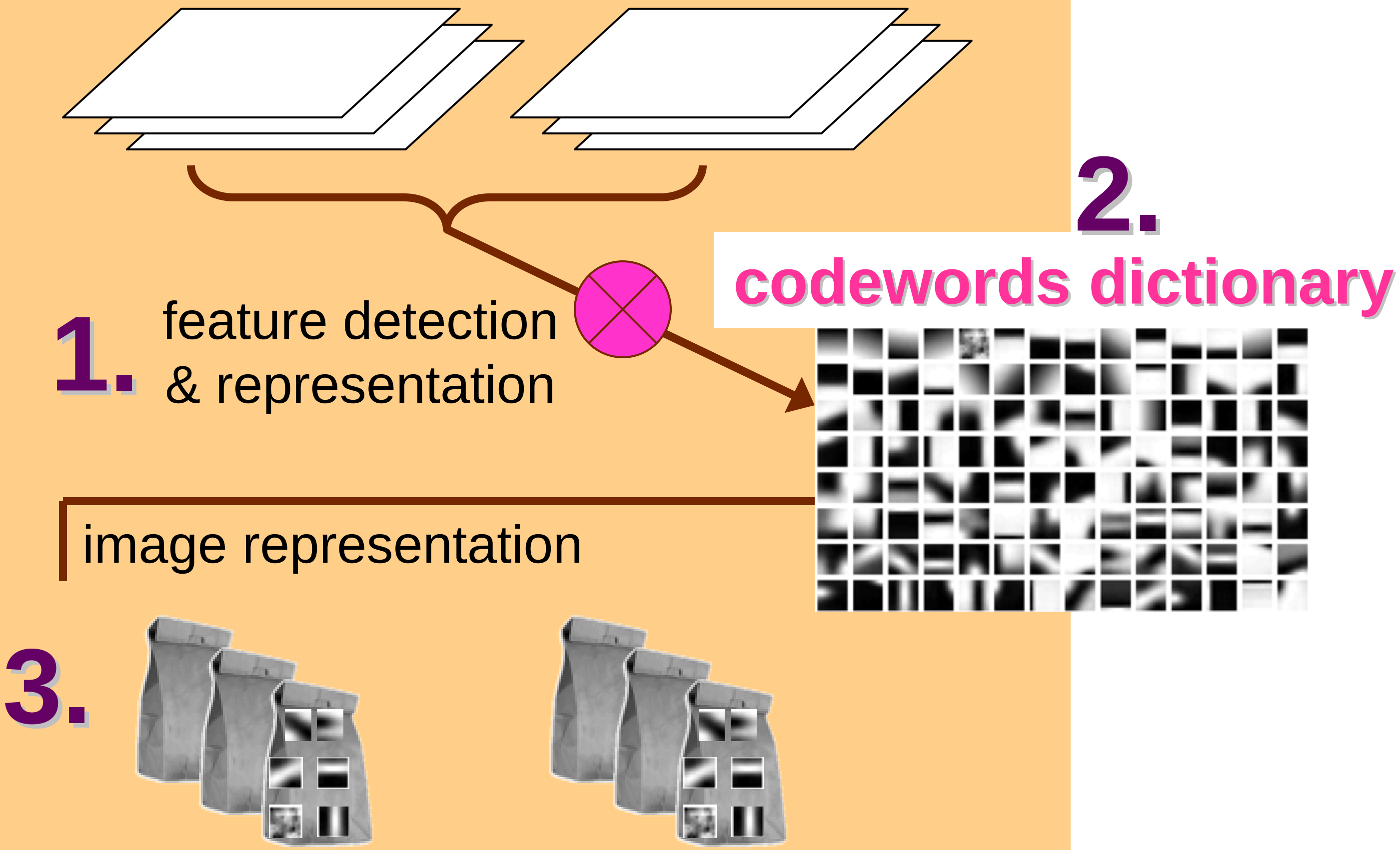

Beaches
Highways
Forests
Buildings
Mountains
Industry



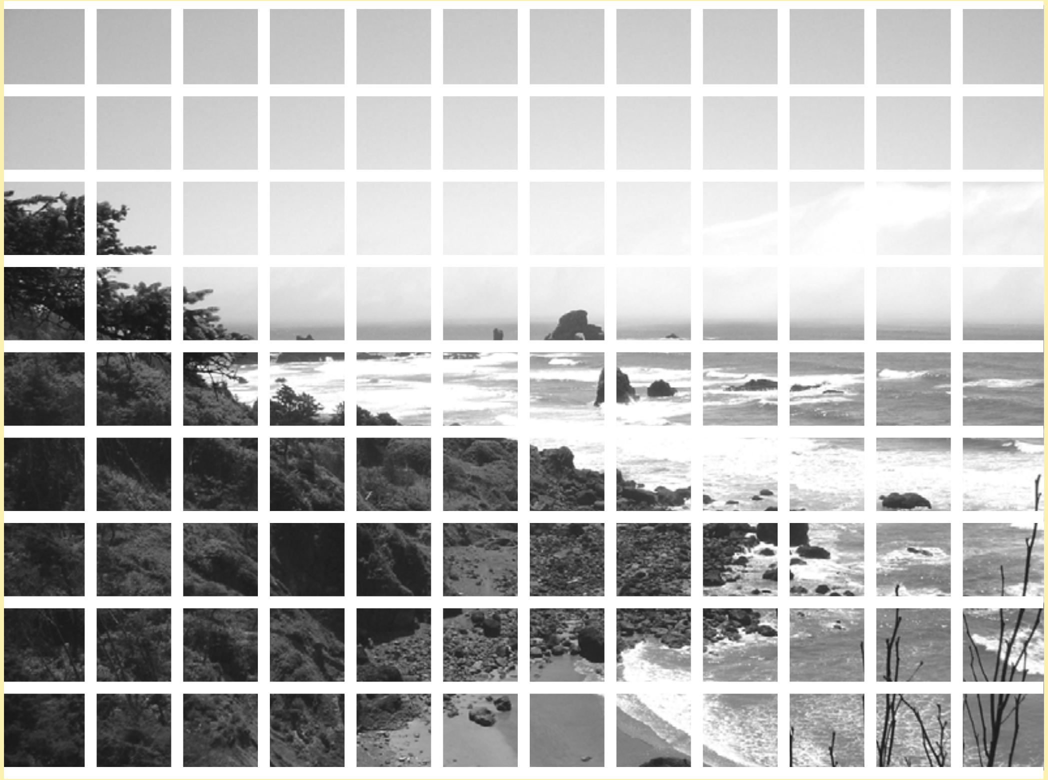
Fei-Fei & Perona, CVPR 2005


learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

Representation



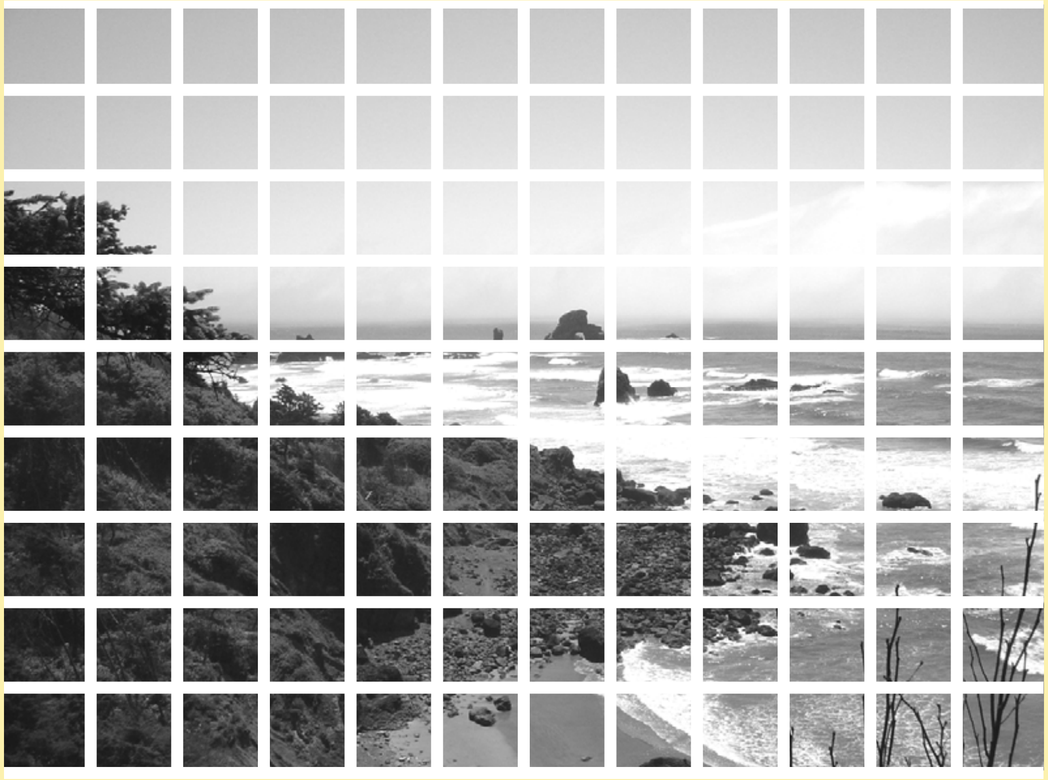
1.Feature detection and representation



extract interest points

- DoG
- Saliency detector (Kadir and Brady)
- grid

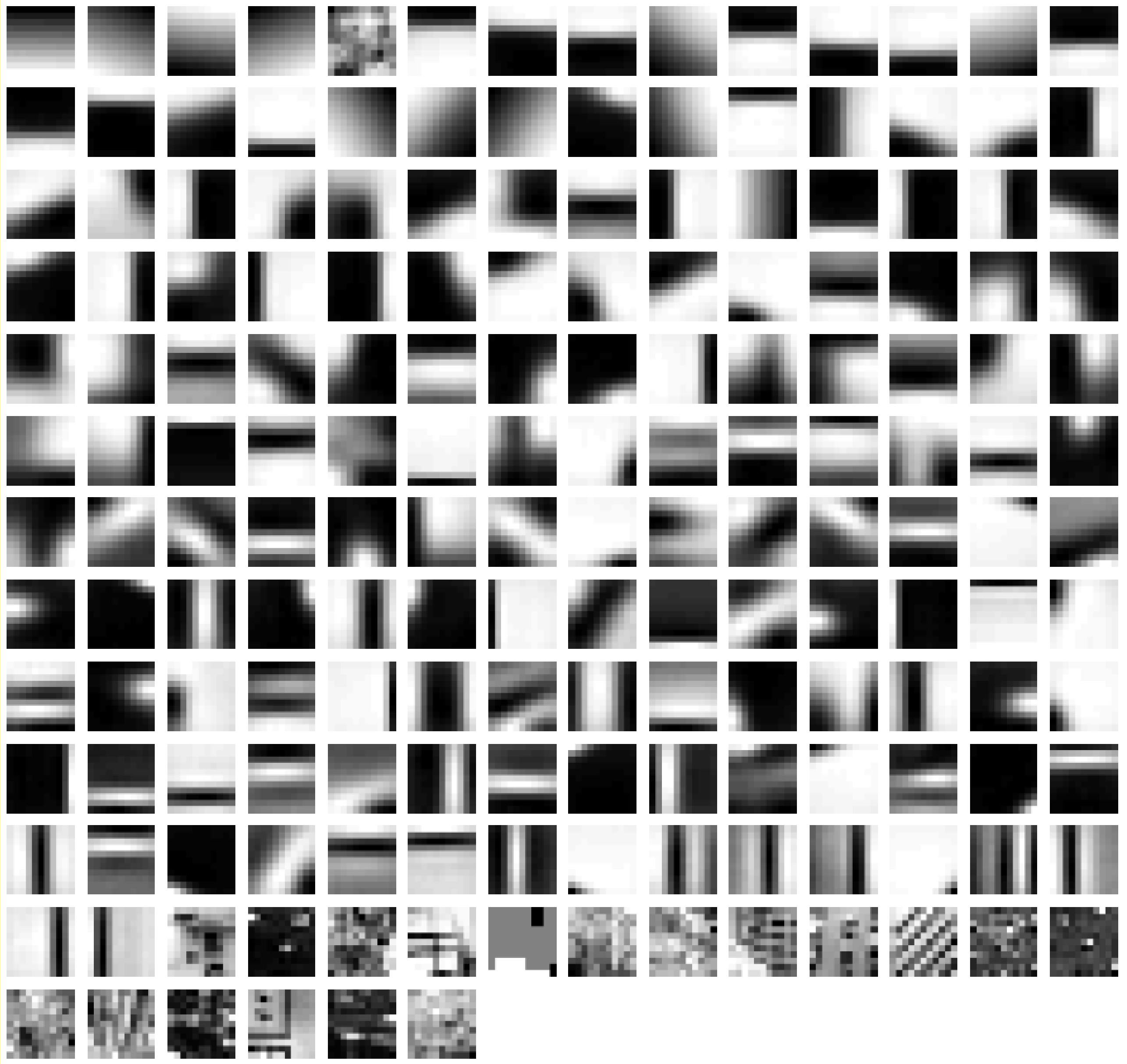
1.Feature detection and representation



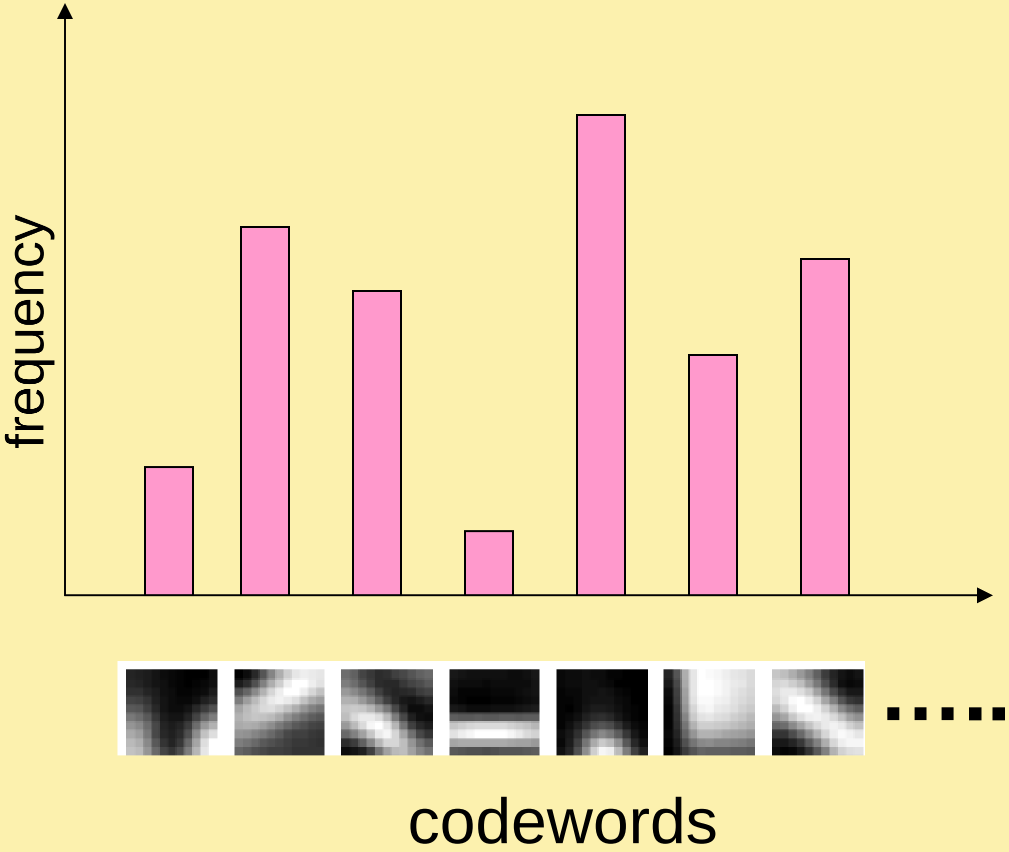
represent interest points

- SIFT (Lowe ’99)
- gray scale values

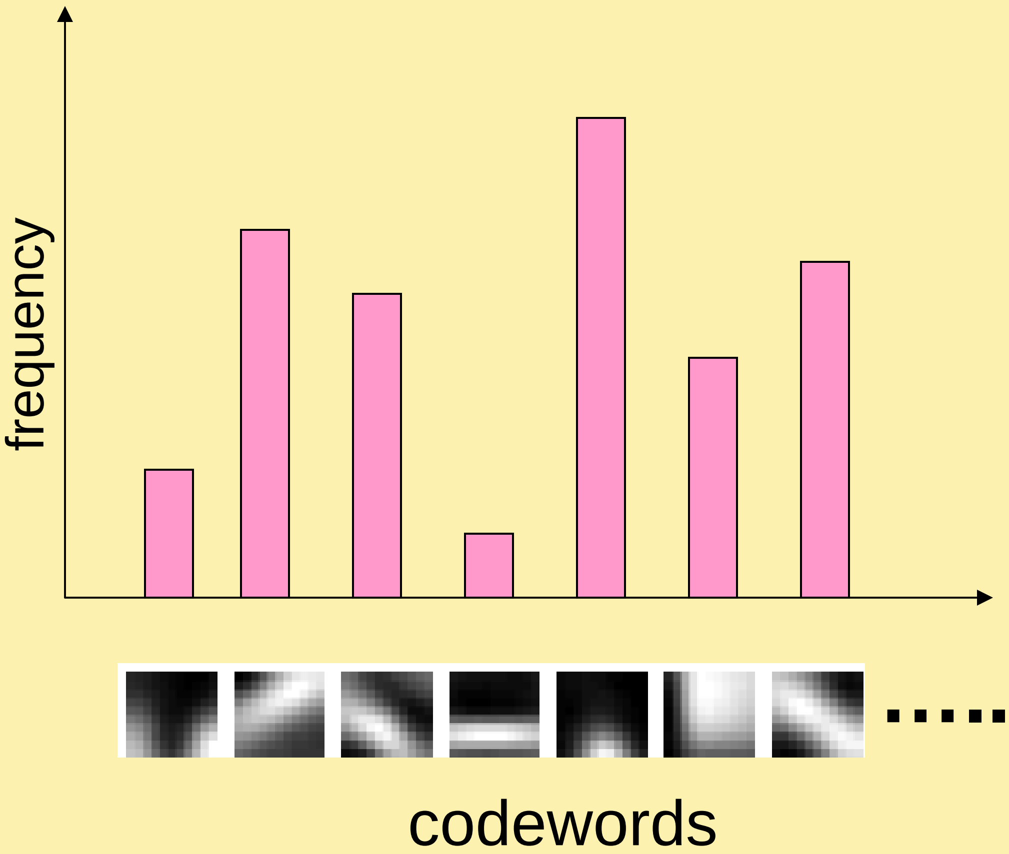
2. Codewords dictionary formation



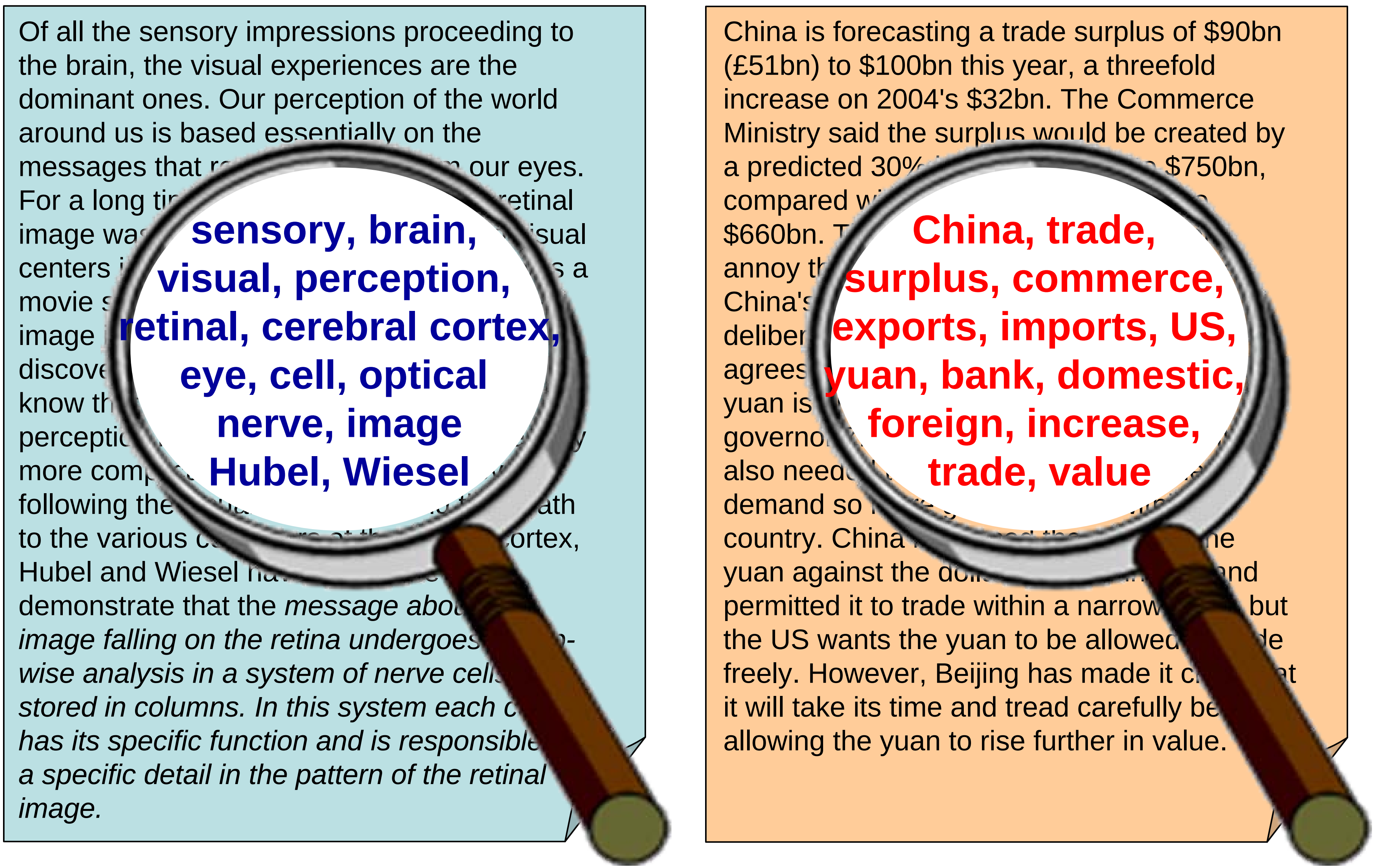
3. Image representation



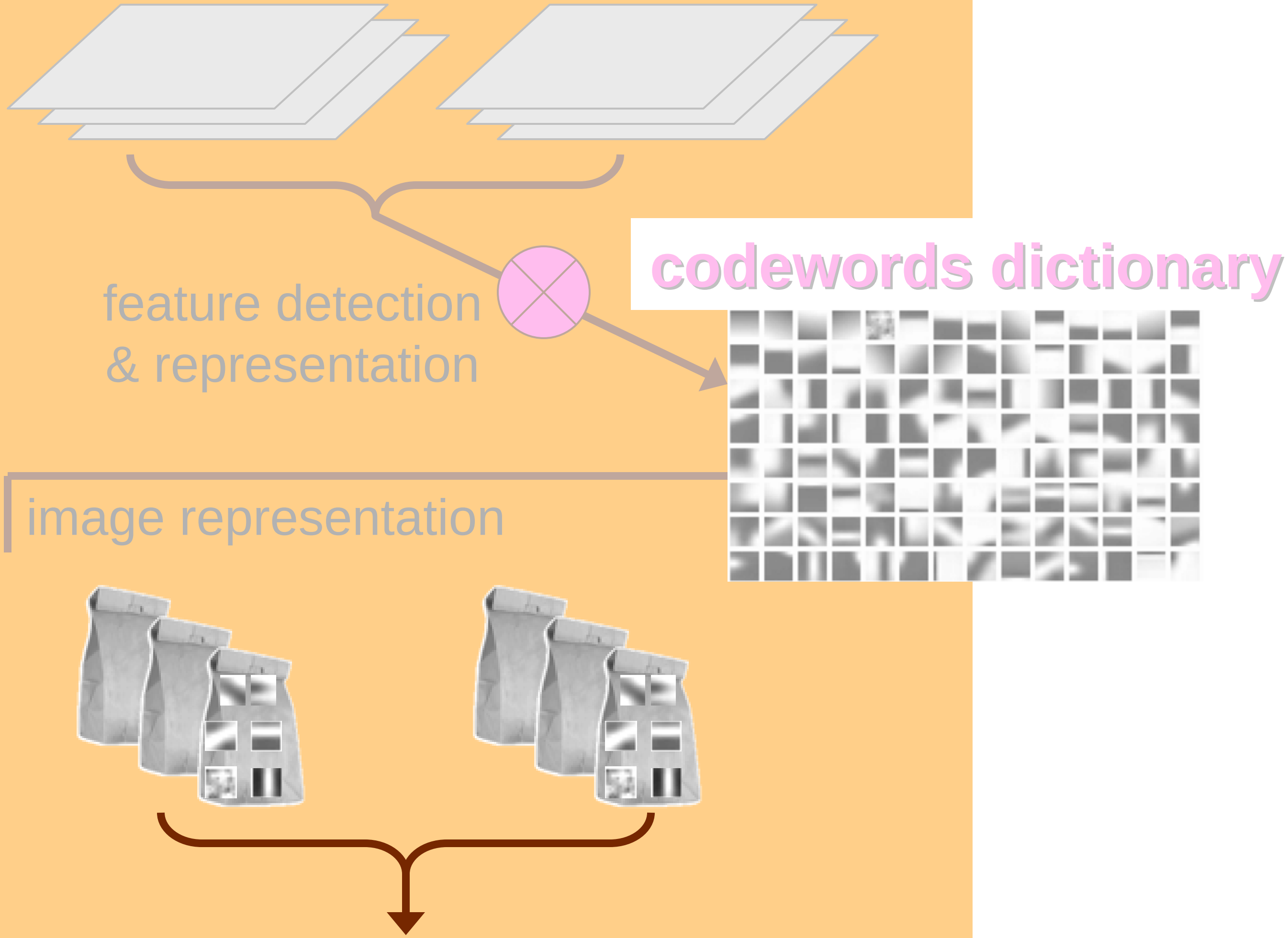
3. Image representation



Analogy to documents

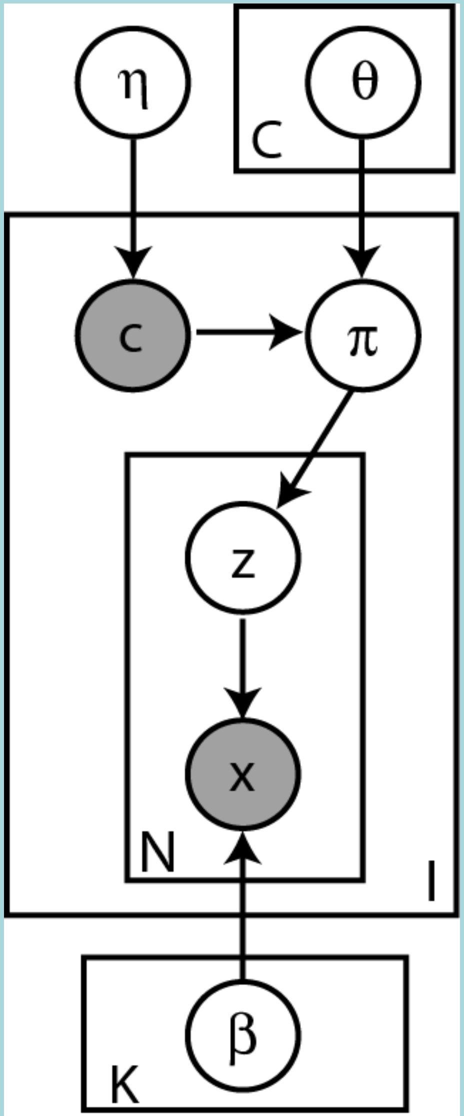


learning

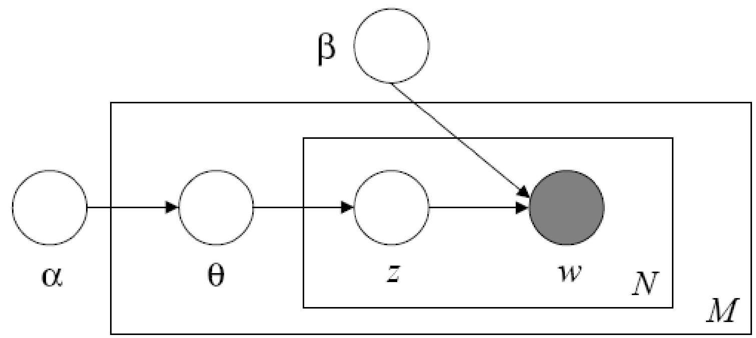


category models
(and/or) classifiers

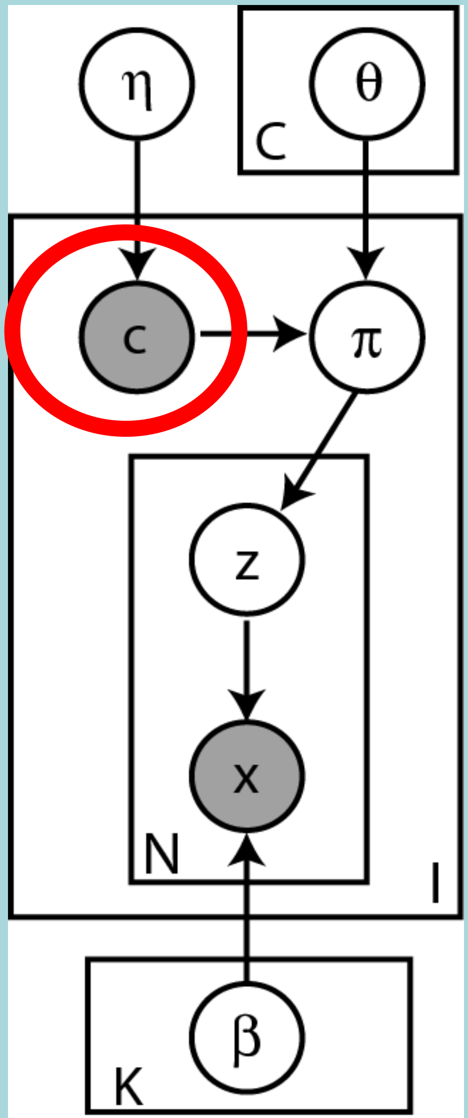
A Generative Model



LDA: Blei, Ng, & Jordan. 2003

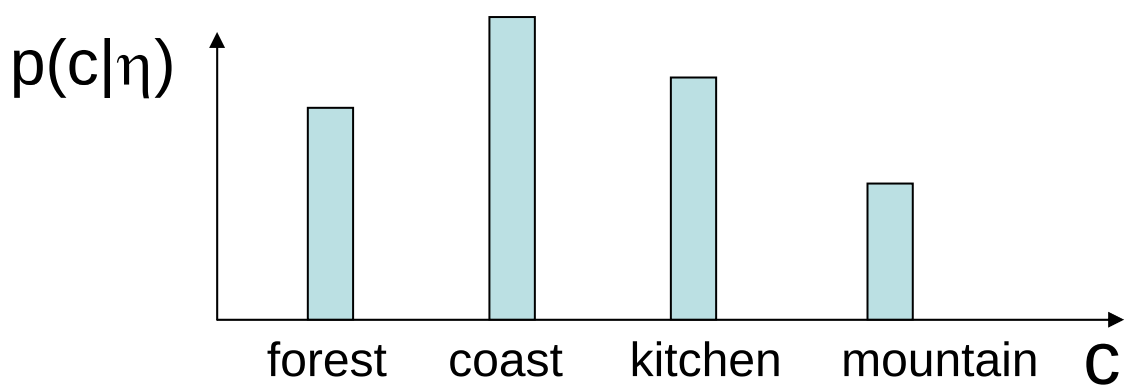


A Generative Model

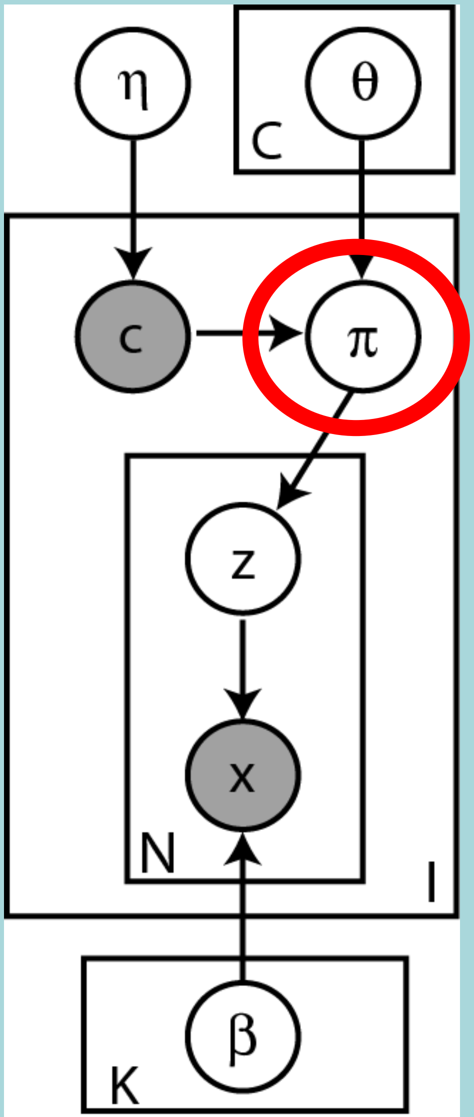


scene category

discrete variable: $c \sim p(c|\eta)$



A Generative Model

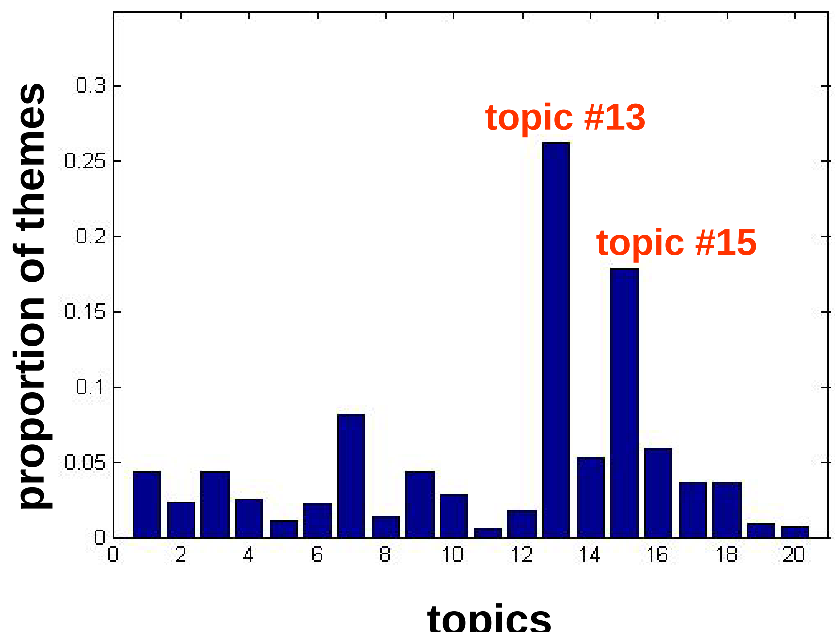


mixing parameter for the latent topics

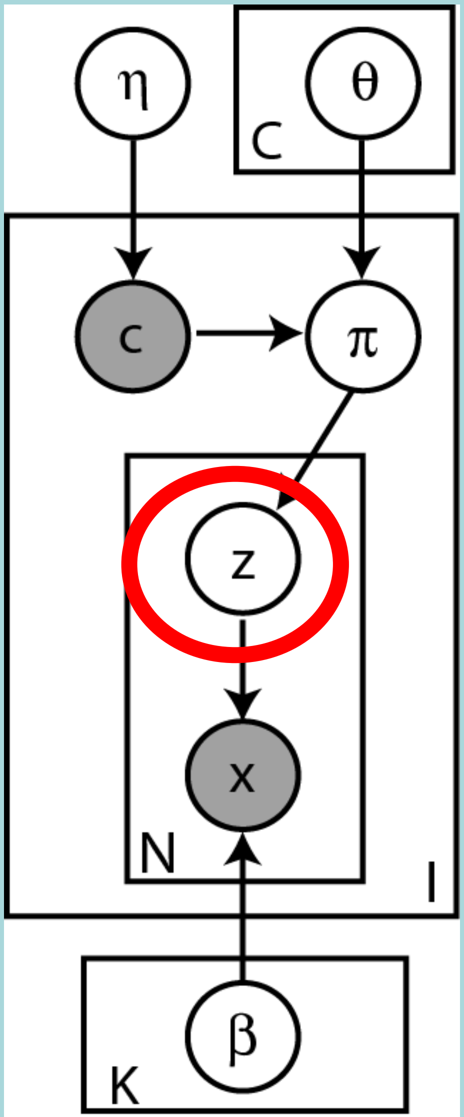
$$\pi \sim p(\pi | c, \theta)$$

$$\sim Dir(\pi | c, \theta)$$

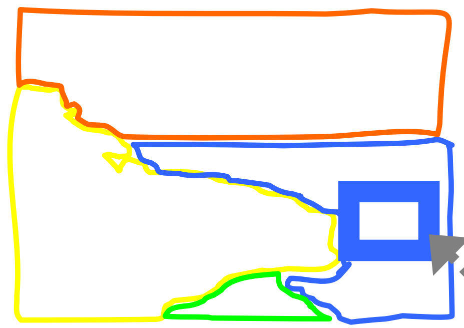
where $\sum_{k=1}^{K} \pi_k = 1$



A Generative Model



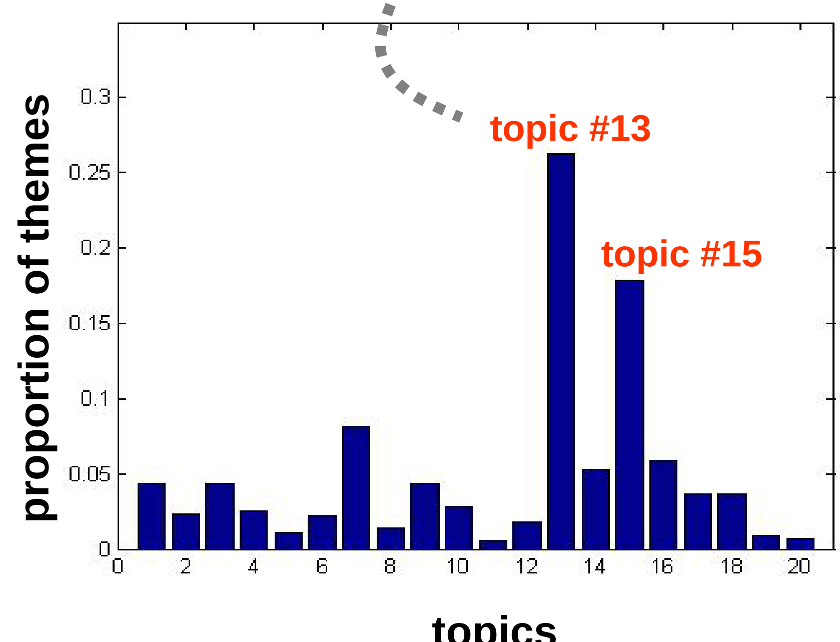
topic label



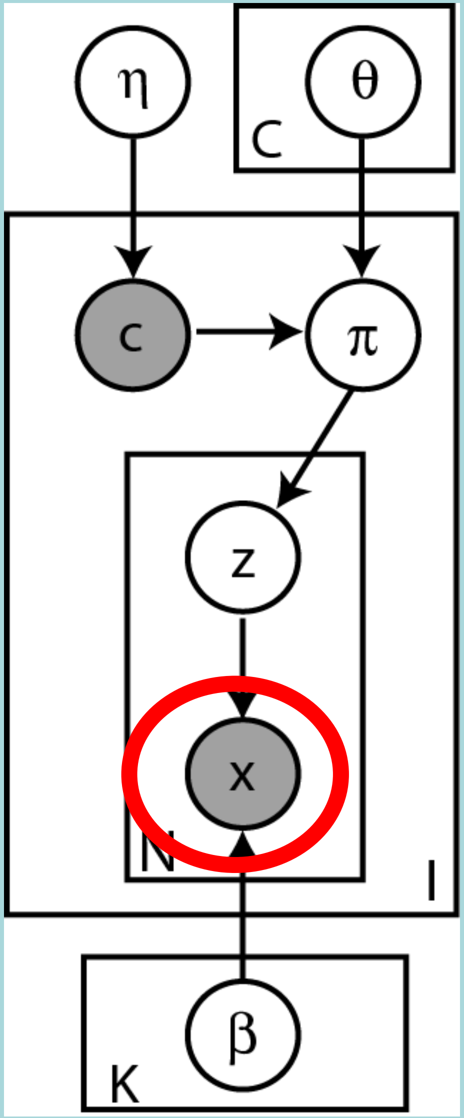
discrete variable:

$$z \sim p(z|\pi)$$

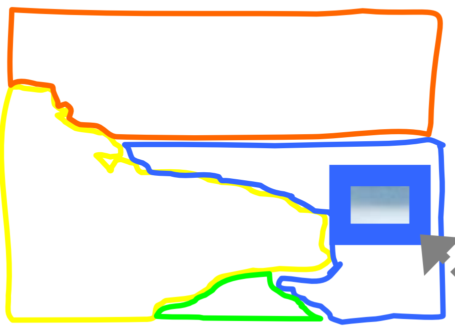
$$\sim Mult\ (z|\pi)$$



A Generative Model



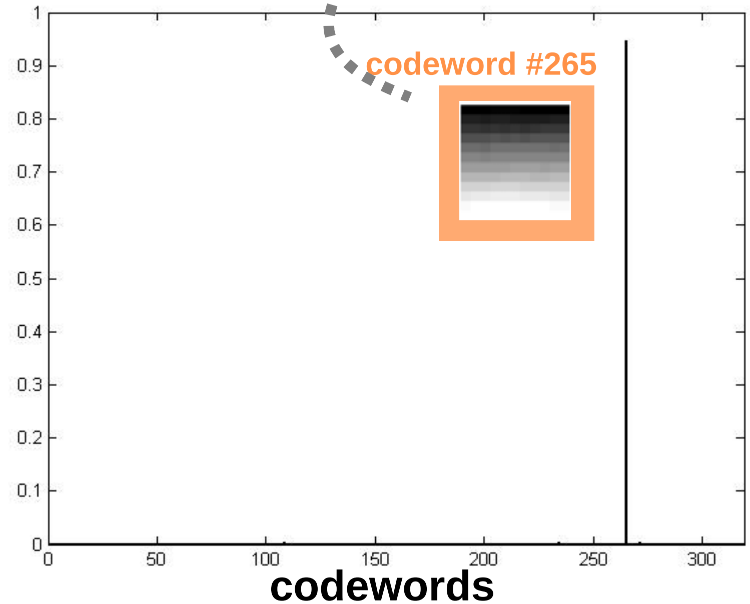
patch label



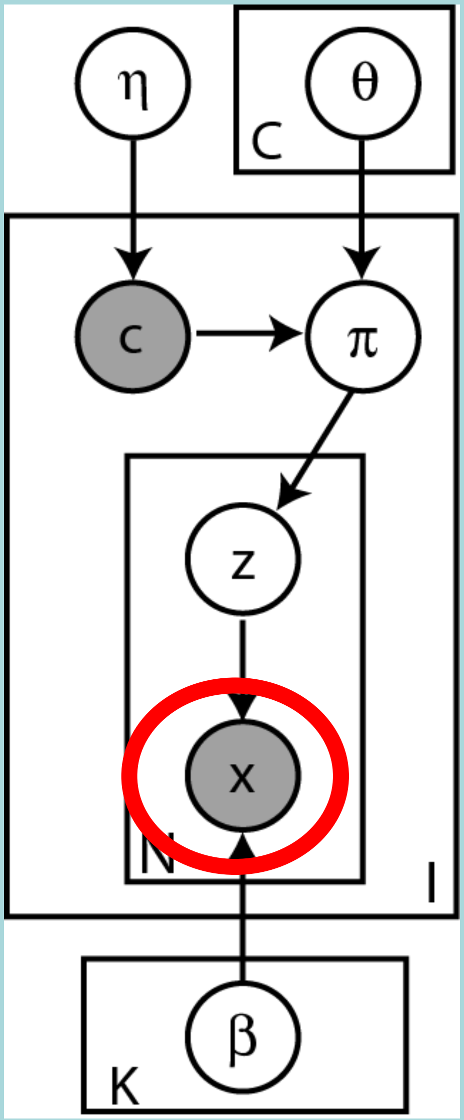
discrete variable:

$$x \sim p(x \mid z, \beta)$$
$$\sim Mult(x \mid z, \beta)$$

expected value of β given 'z=13'

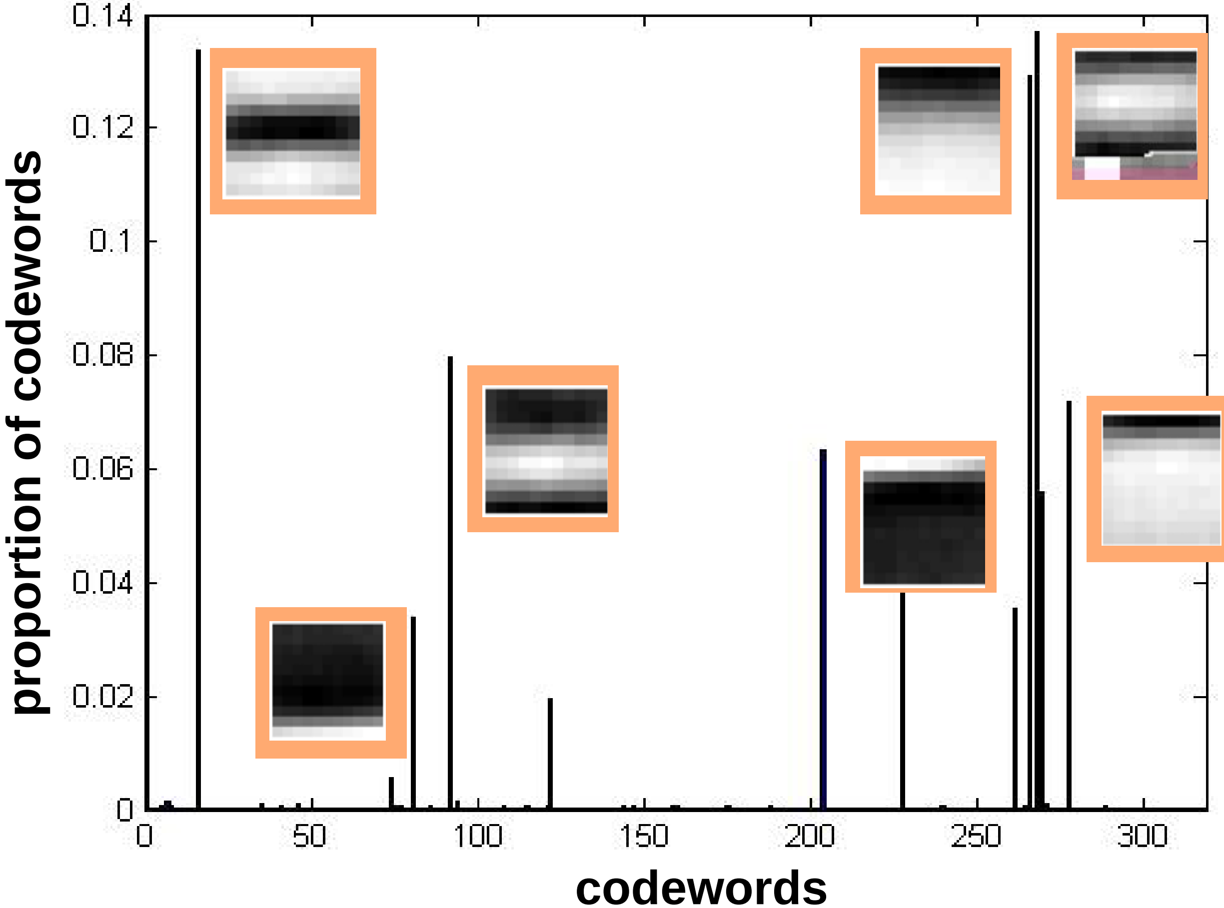


A Generative Model

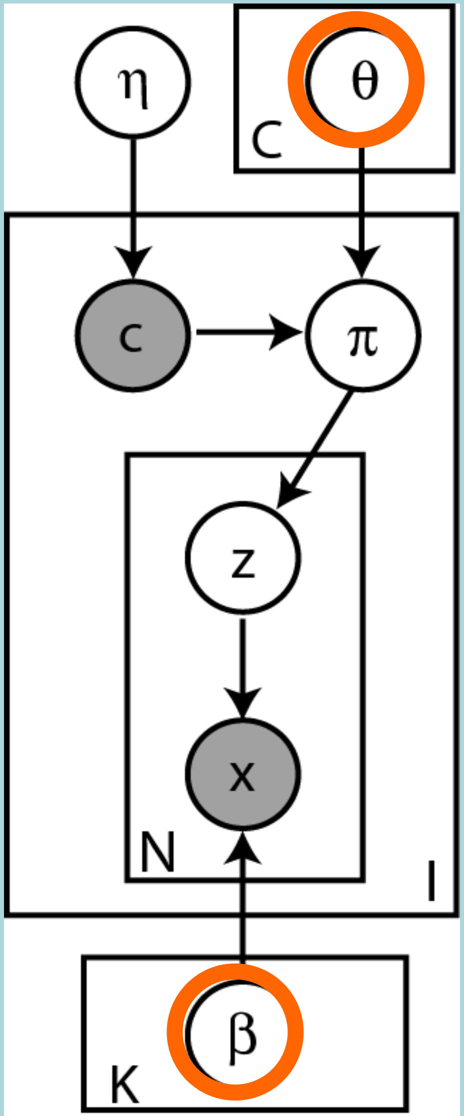


patch label

expected value of β given ‘z=15’



A Generative Model



learning

Find the 'best' θ and β

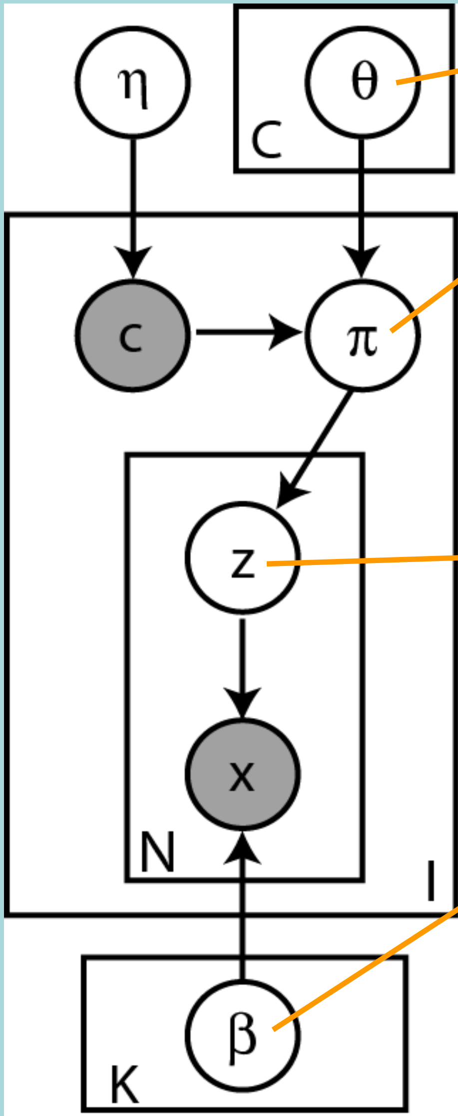
joint probability

$$p(x,z,\pi|\theta,\beta,c) = p(\pi|c,\theta)\prod_n^N p(z_n|\pi)p(x_n|z_n,\beta)$$

$$p(x|\theta,\beta,c) = \int p(\pi|c,\theta)\left(\prod_n^N \sum_{z_n} p(z_n|\pi)p(x_n|z_n,\beta)\right)d\pi$$

- exact inference is intractable
- use Variational Inference

A Generative Model



Variational Inference

Maximum Likelihood estimation (Minka 2000)

$$\gamma_{ck} = \theta^0_{ck} + \sum_n^N \left\langle \delta\left(z_n^k = 1\right)\right\rangle$$

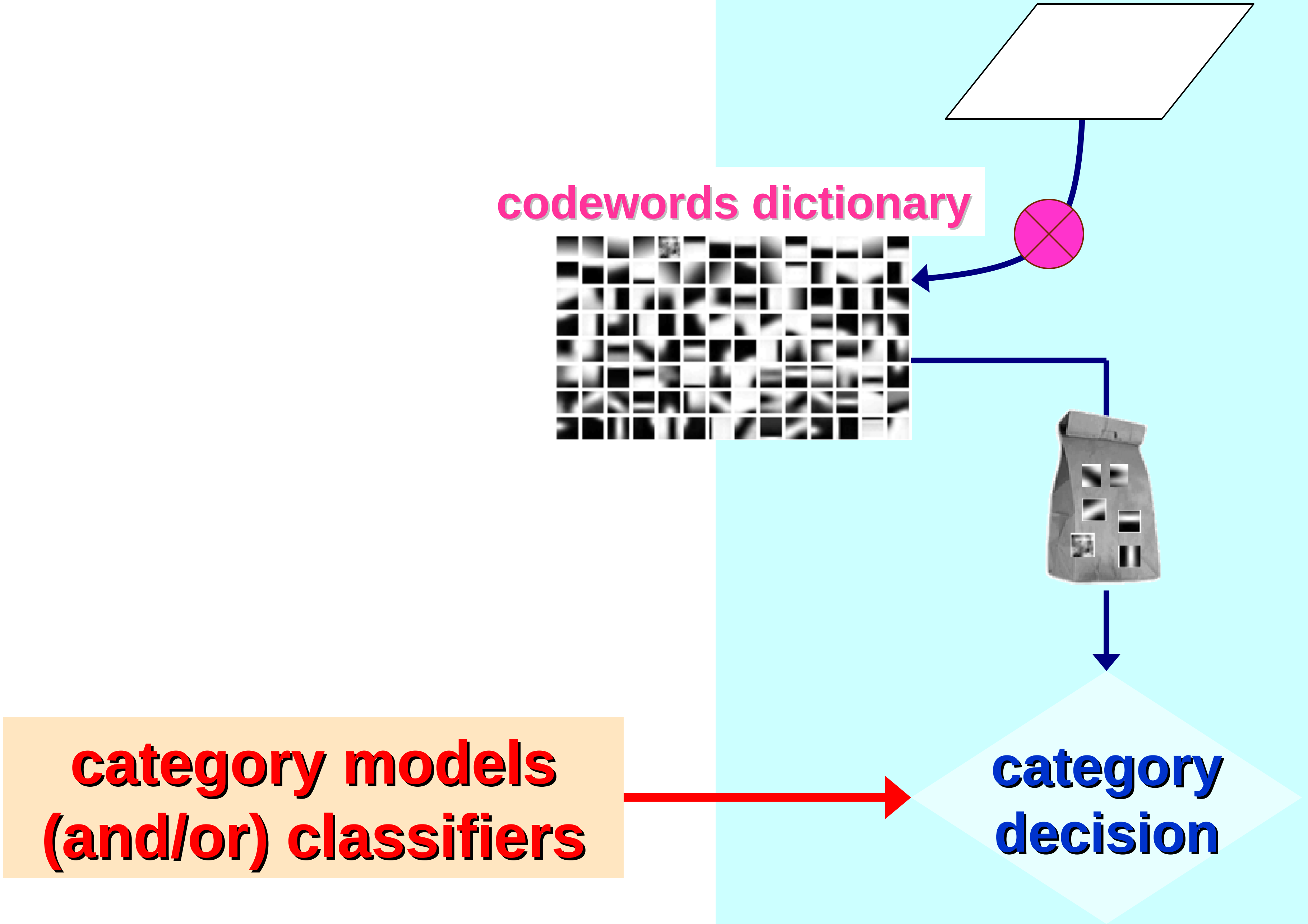
$$\left\langle \log \pi_{ck} \right\rangle = \Psi\left(\gamma_{ck}\right) - \Psi\left(\sum_k \gamma_{ck}\right)$$

$$\left\langle \delta\left(z_n^k = 1\right)\right\rangle = \exp\left\{\left\langle \log \pi_{ck}\right\rangle + \sum_t^T \left\langle \log \beta_{kt}\right\rangle \delta\left(x_n^t = 1\right)\right\}$$

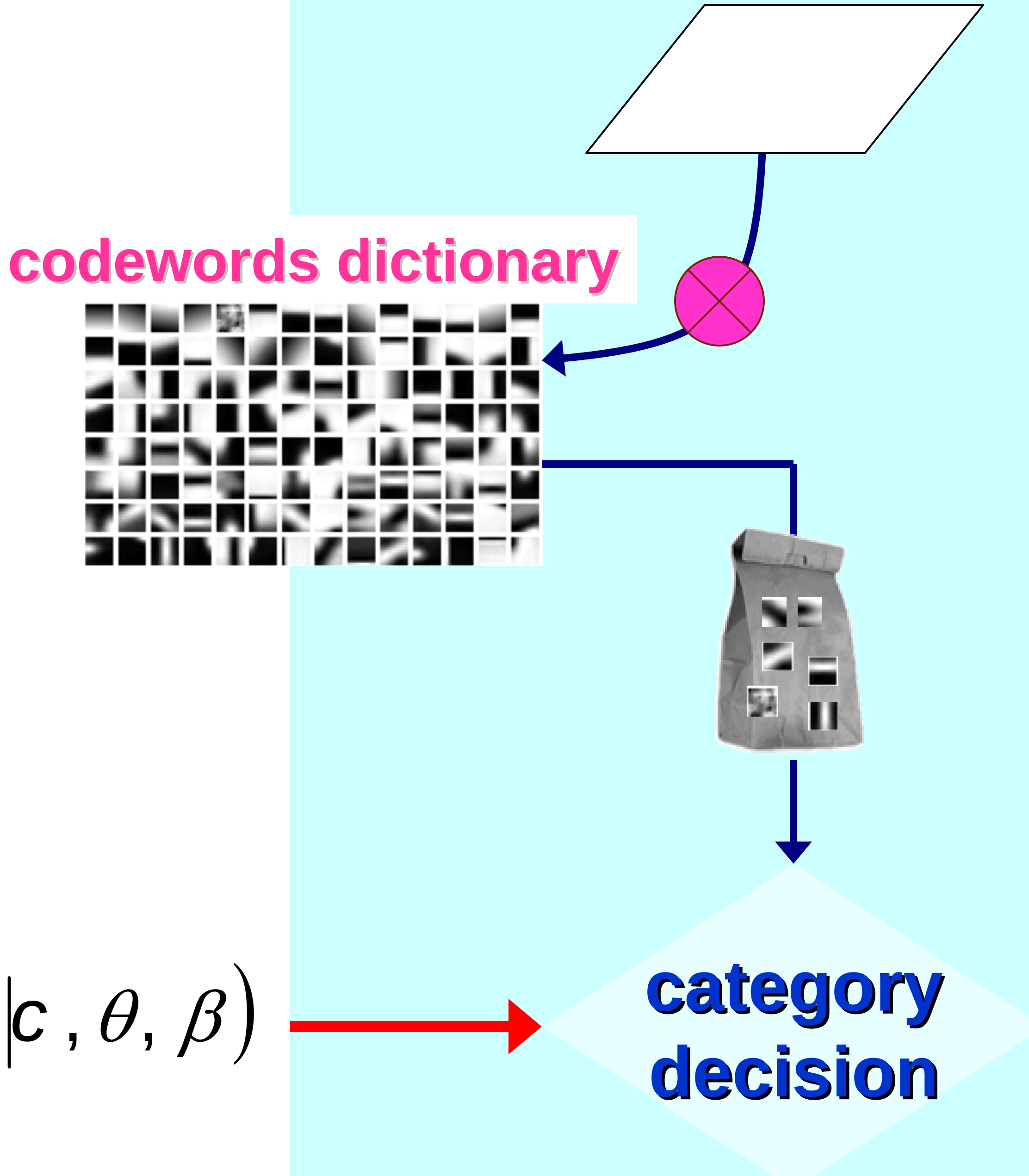
$$\xi_{kt} = \zeta^0 + \sum_i^I \sum_n^N \left\langle \delta\left(z_{i,n}^k = 1\right)\right\rangle \delta\left(x_{i,n}^t = 1\right)$$

$$\left\langle \log \beta_{kt} \right\rangle = \Psi\left(\xi_{kt}\right) - \Psi\left(\sum_t \xi_{kt}\right)$$

Recognition

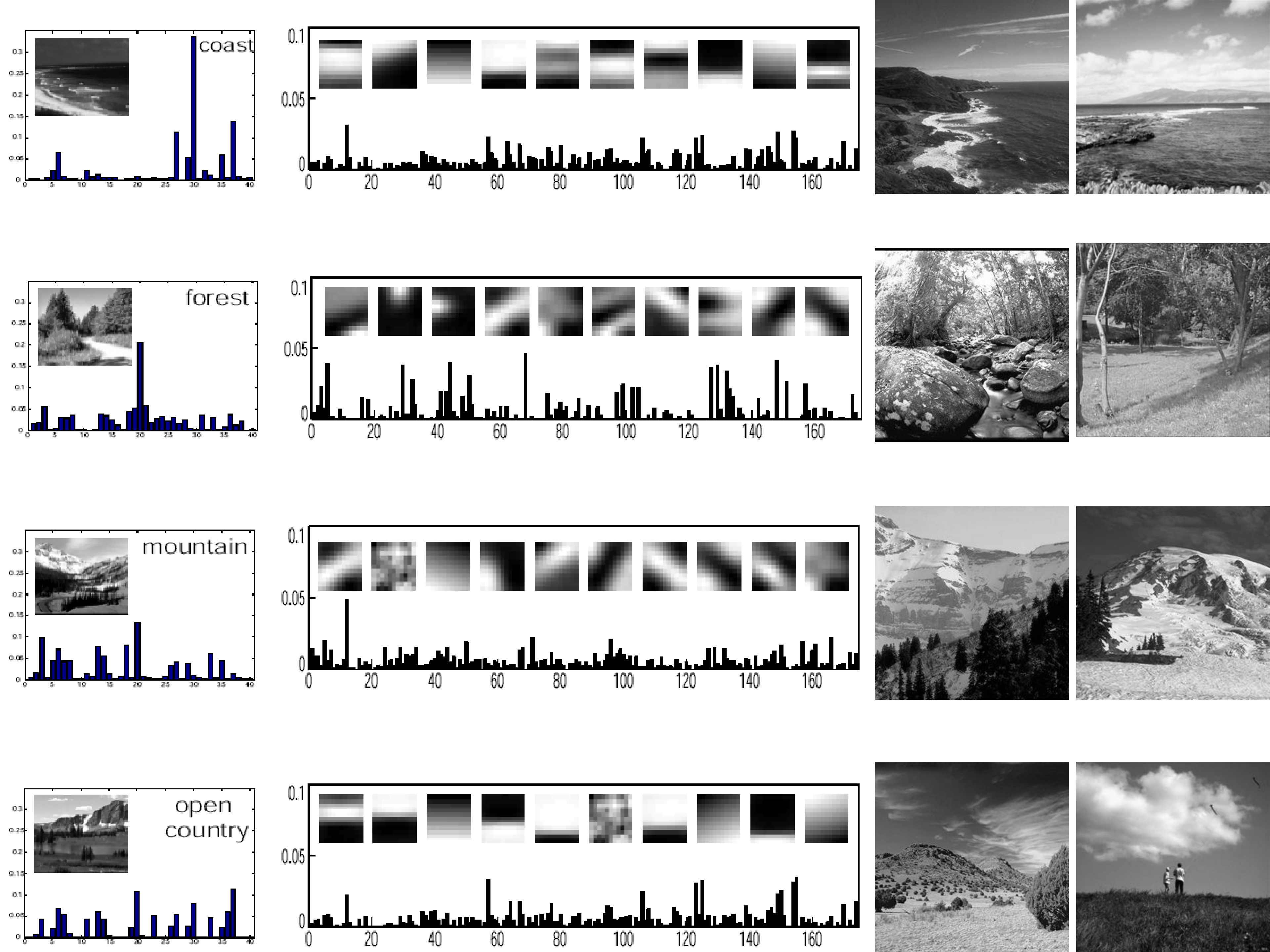


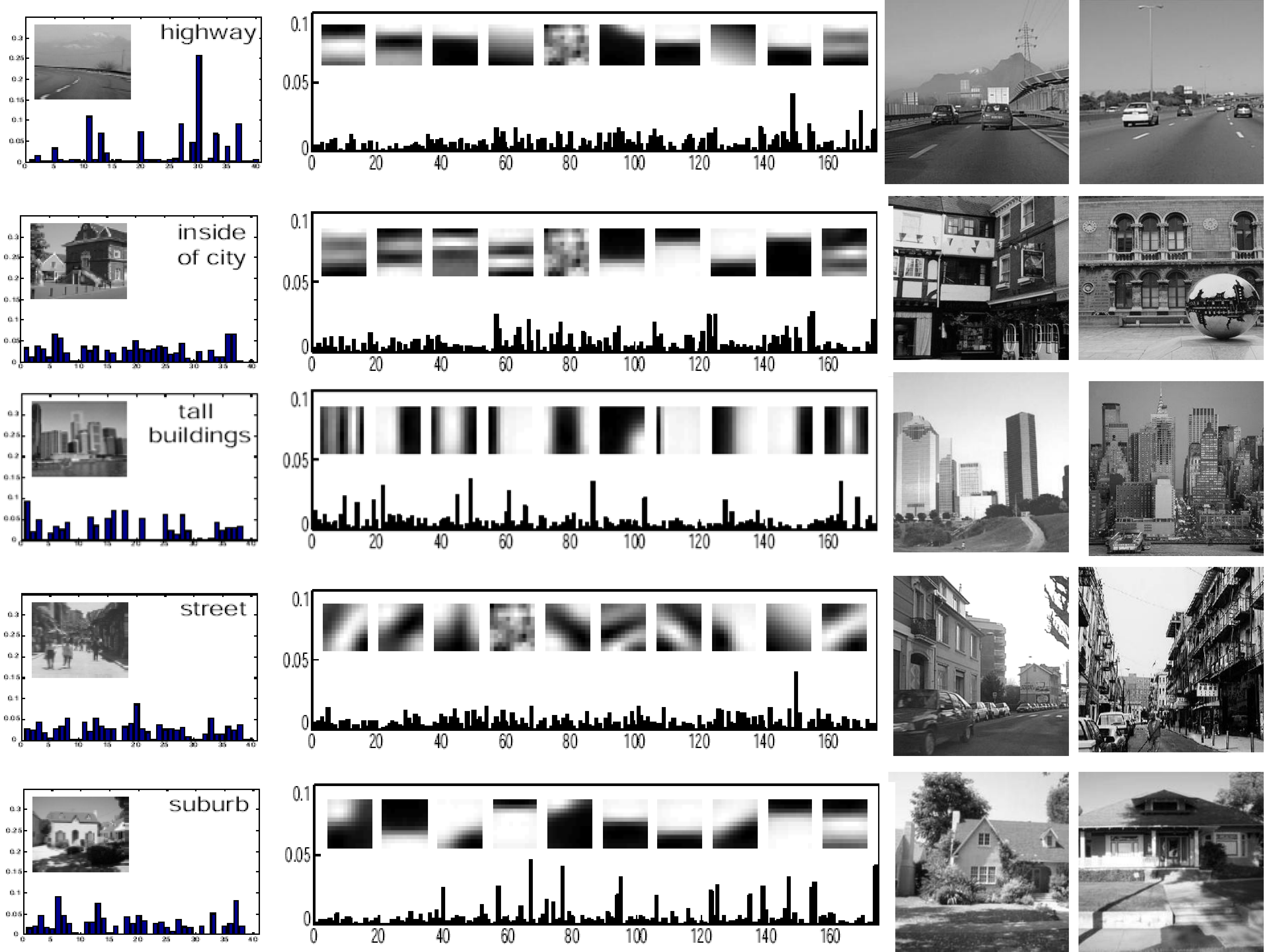
Recognition

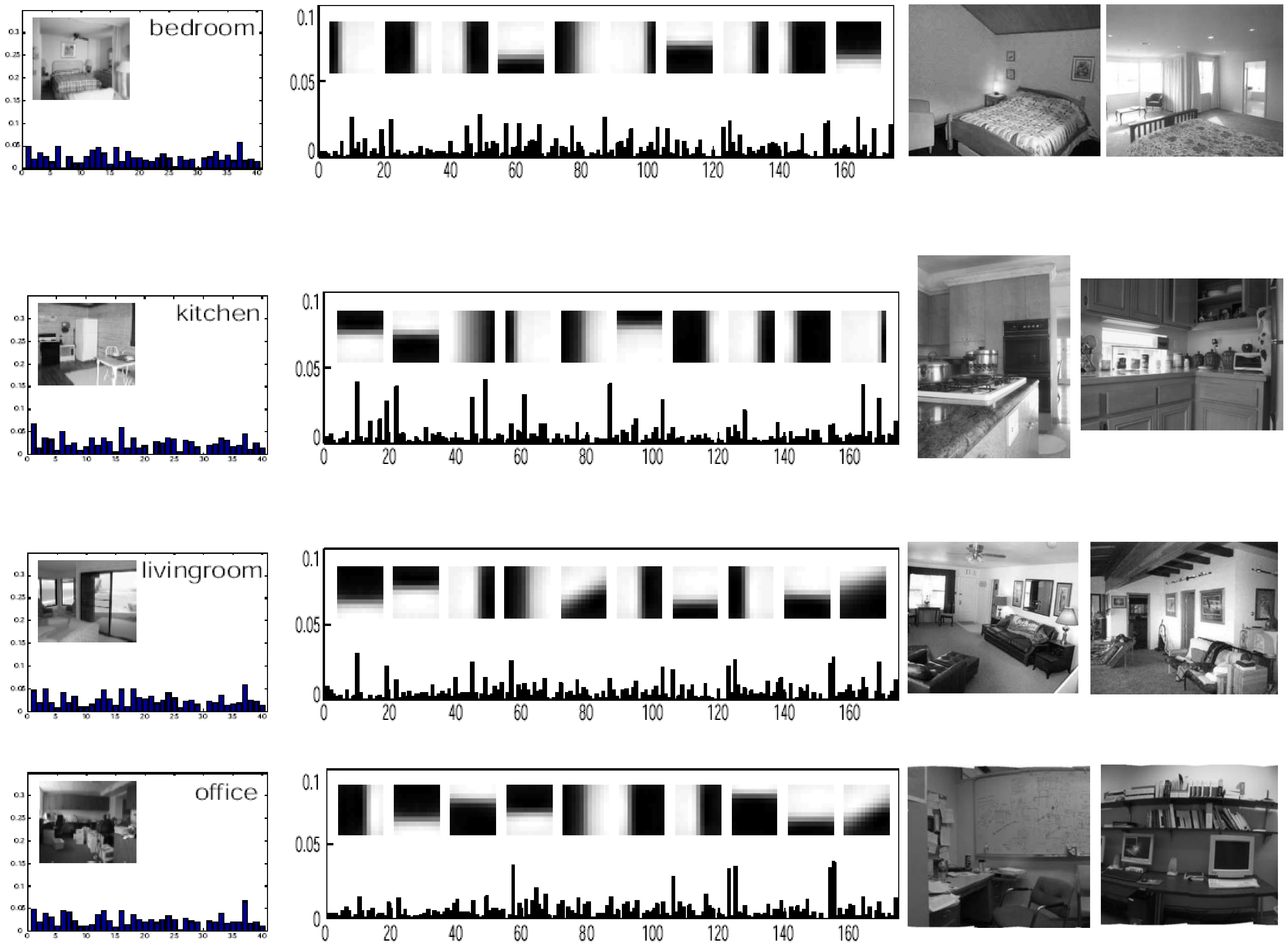


$$c = \arg\max_{c} \; p\left(x \mid c, \theta, \beta\right)$$

	highway	insidecity	tallbuildings	street	suburb	forest	coast	mountain	opencountry	bedroom	kitchen	livingroom	office
highway	**74**	2		2	2		14	4		2			
insidecity		**58**	10	6	8		4			2	6	4	2
tallbuildings		4	**76**	10				4		4		2	
street	2	4	6	**78**		2		2	2			4	
suburb					**94**					2			4
forest						**88**		12					
coast	2						**78**		20				
mountain	4		4		2	6	8	**70**	6				
opencountry	8				8	10	16	10	**48**				
bedroom	4	2	2		2	2	2	4		**28**	12	38	4
kitchen		8	2				2				**60**	14	14
livingroom		2	2	2			2	4		4	18	**56**	10
office					2		2			8	12	12	**64**

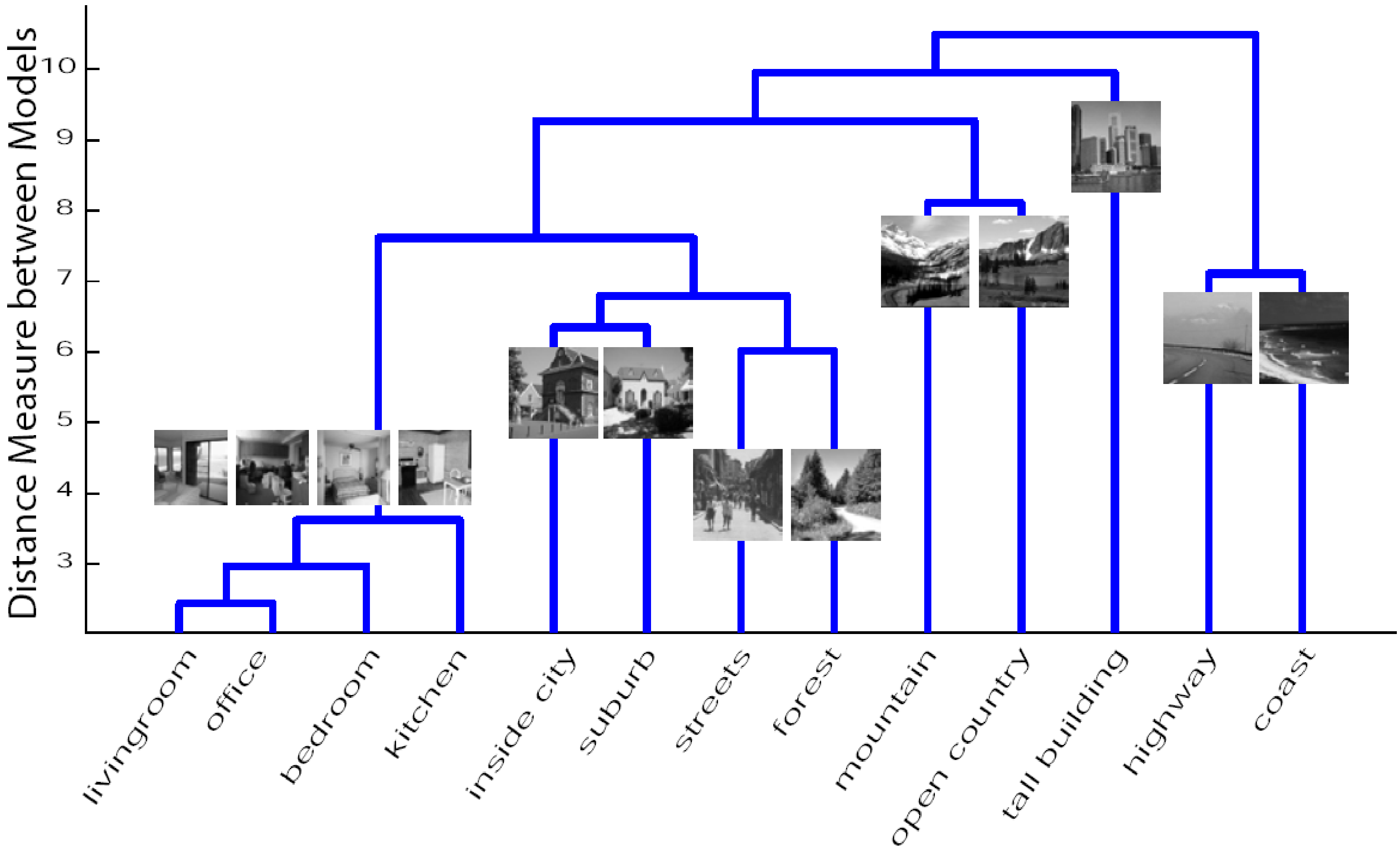

coast
forest
mountain
open
country
0.1
0.05
0
0
20
40
60
80
100
120
140
160

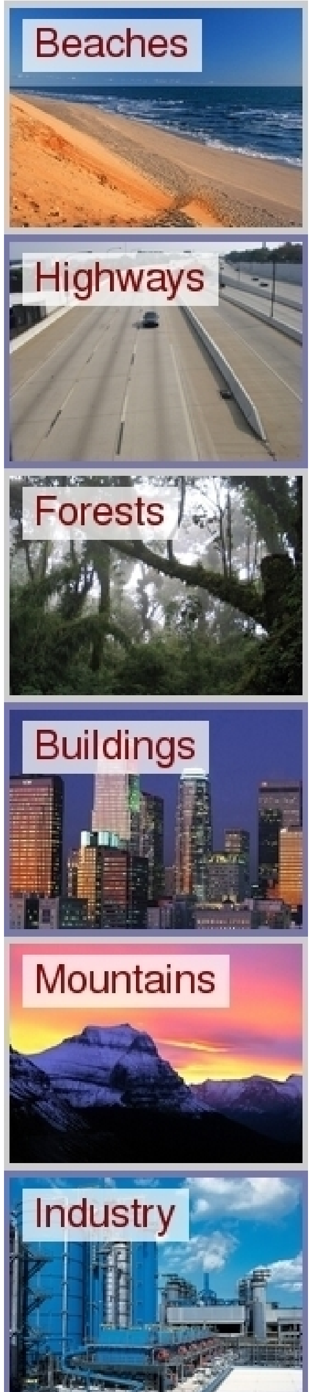

highway
inside
of city
tall
buildings
street
suburb
0.1
0.05
0
0
20
40
60
80
100
120
140
160

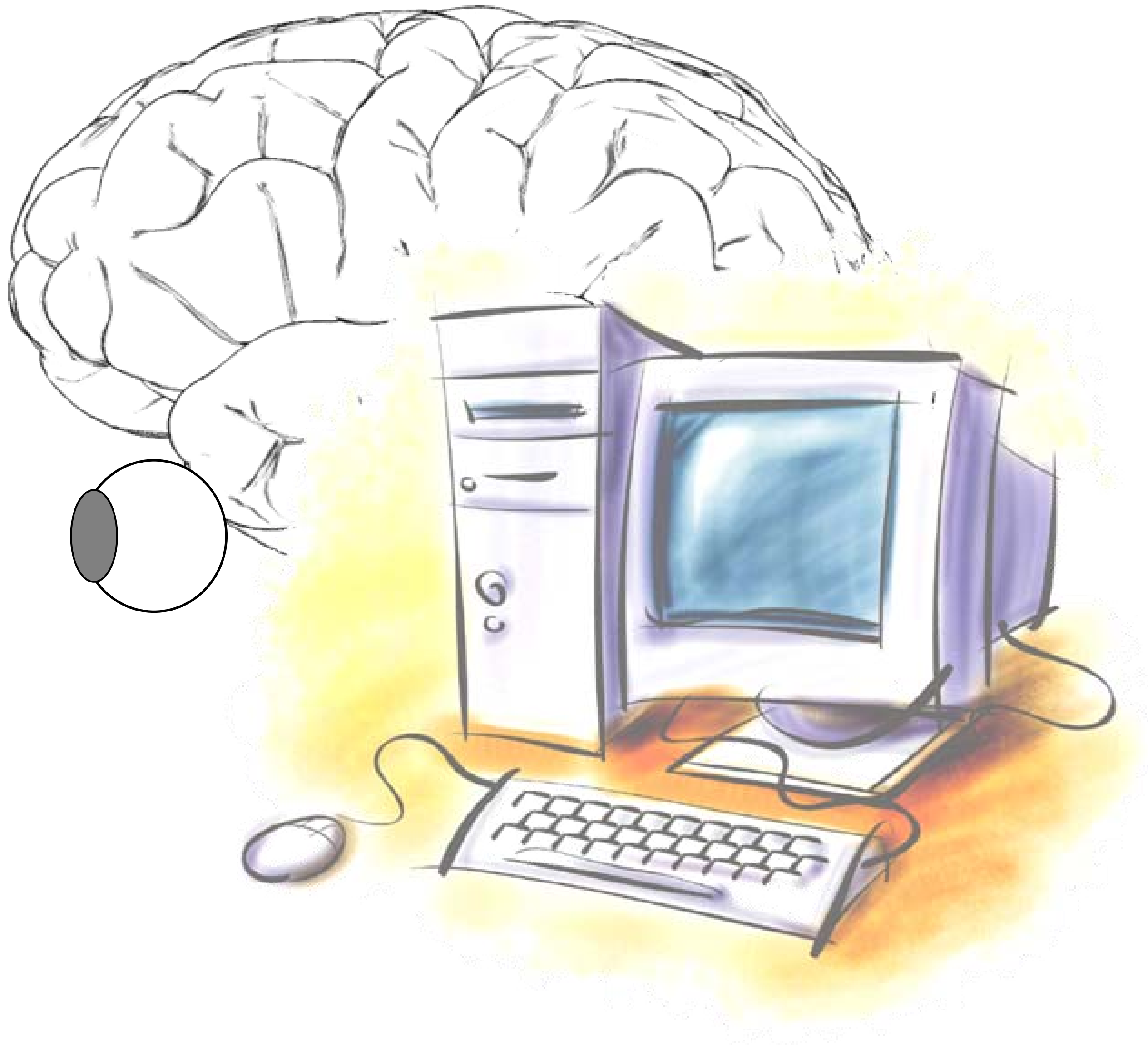


bedroom
kitchen
livingroom
office
0.1
0.05
0
0
20
40
60
80
100
120
140
160

model distance based on topic distribution




Beaches
Highways
Forests
Buildings
Mountains
Industry



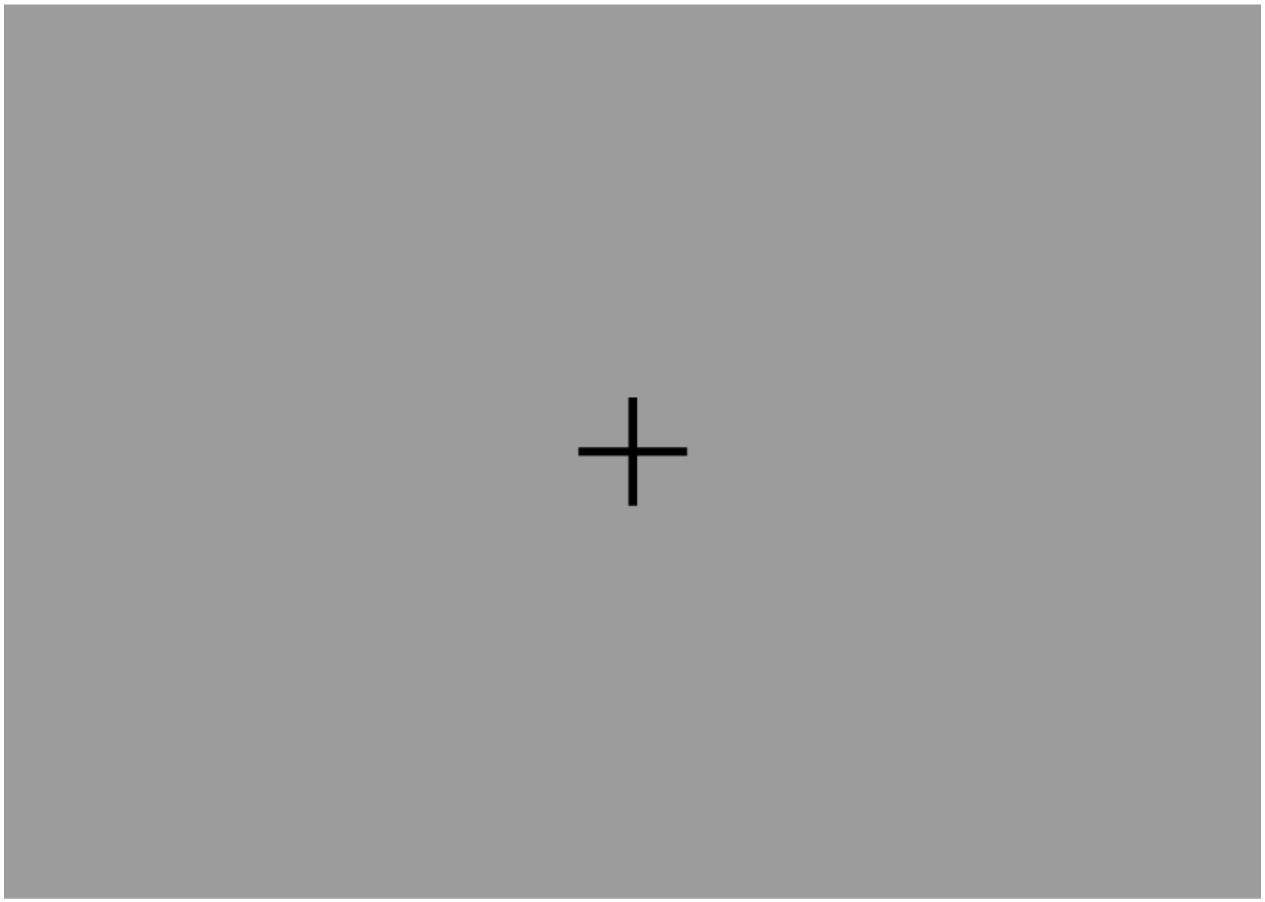
Change blindness

Rensink, O'regan, Simon, etc.

Change blindness

Rensink, O’regan, Simons, etc.

what **DO** we see in a glance?



Fei-Fei et al. JoV 2007









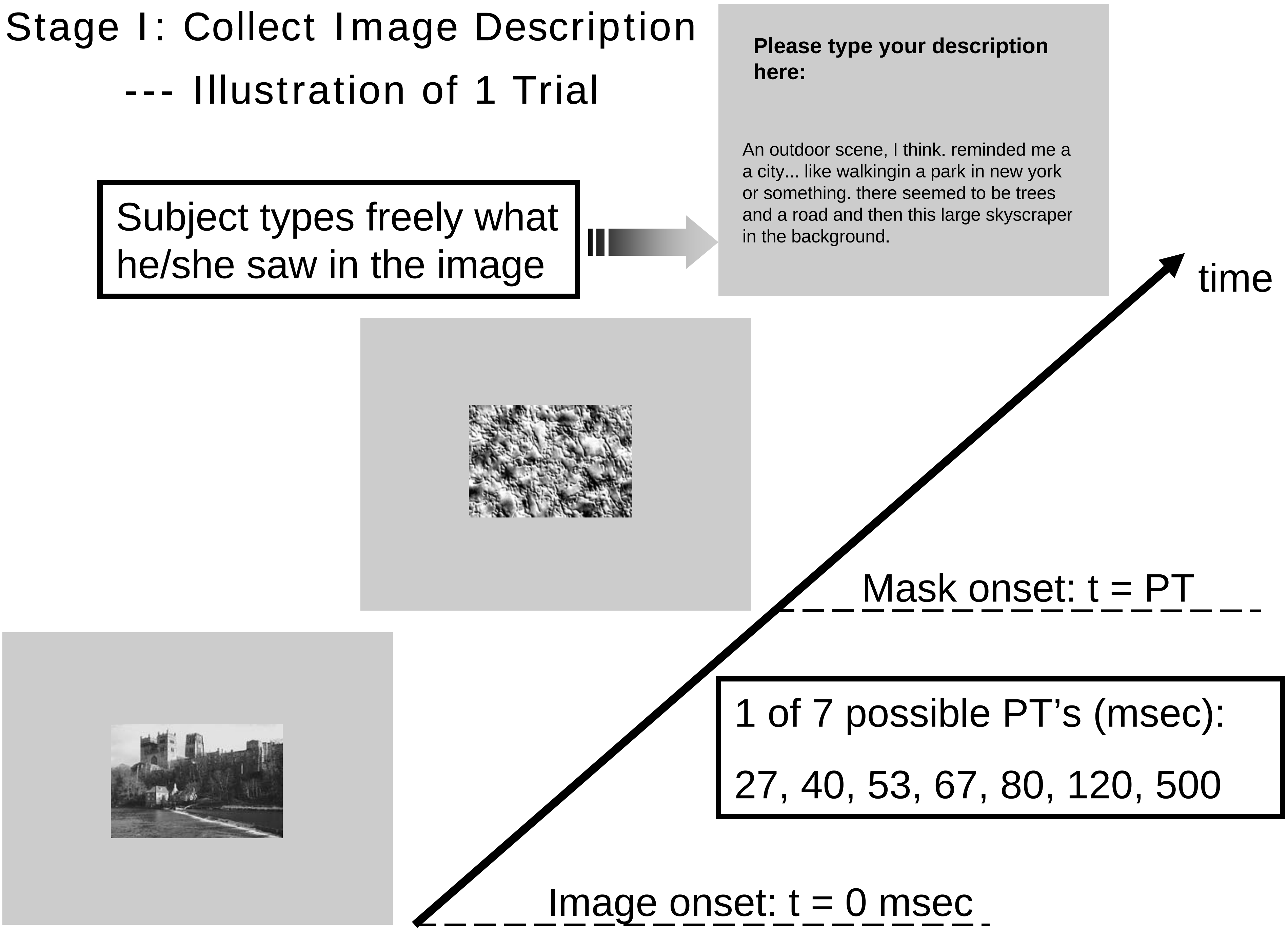










Stage I: Collect Image Description
--- Illustration of 1 Trial
Please type your description here:
An outdoor scene, I think. reminded me a a city... like walkingin a park in new york or something. there seemed to be trees and a road and then this large skyscraper in the background.
Subject types freely what he/she saw in the image
time
Mask onset: t = PT
1 of 7 possible PT's (msec):
27, 40, 53, 67, 80, 120, 500
Image onset: t = 0 msec

PT = 27ms

Couldn't see much; it was mostly dark w/ some square things, maybe furniture. (Subject: AM)

PT = 40ms

This looked like an indoor shot. Saw what looked like a large framed object (a painting?) on a white background (i.e., the wall). (Subject: RW)

PT = 500ms

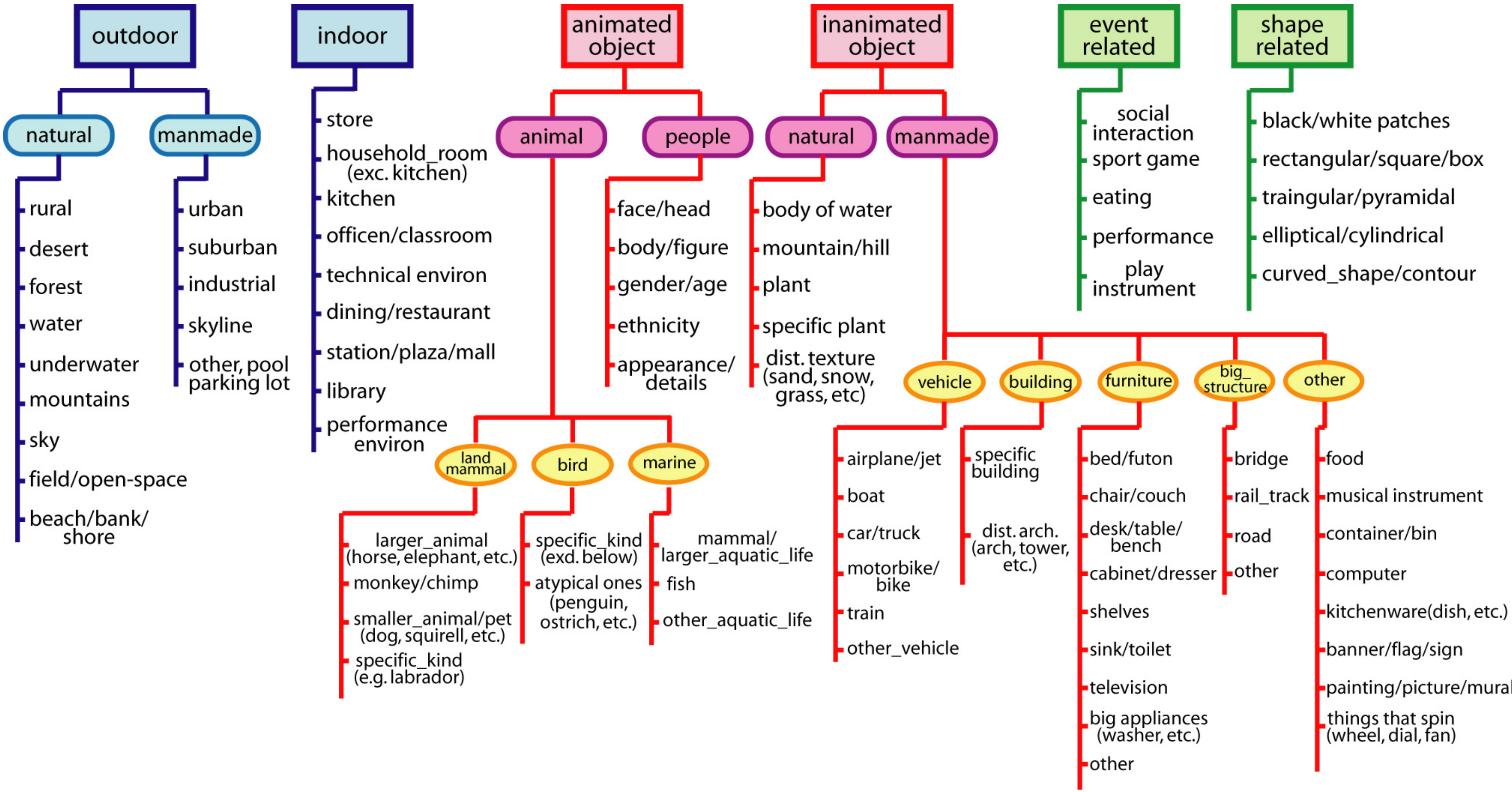
This is indoors. It's must be a rich person's house. There are many paintings on the wall. The largest painting might have a fireplace beneath it. I think the largest painting was that of a man standing erect. The room is richly decorated and it looks like one of the rooms in Mr. Darcy's house in the A&E movie Pride and Prejudice. Or maybe it more closely resembles one of the rooms where the one of the rooms in Hungtington's house (at the Huntington).

PT = 67ms

I saw the interior of a room in a house. There was a picture to the right, that was black, and possibly a table in the center. It seemed like a formal dining room. (Subject: JB)

Fei-Fei et al. JoV 2007

Response attributes??



Fei-Fei et al. *JoV, 2007*

Response No. 18 for Image No. 4

I could make out some kind of circular shapes near the bottom of the picture. These reminded me of those round life preservers that are on ships. There was also a man standing on top of some wooden structure.

CATEGORY: SENSORY/SHAPES

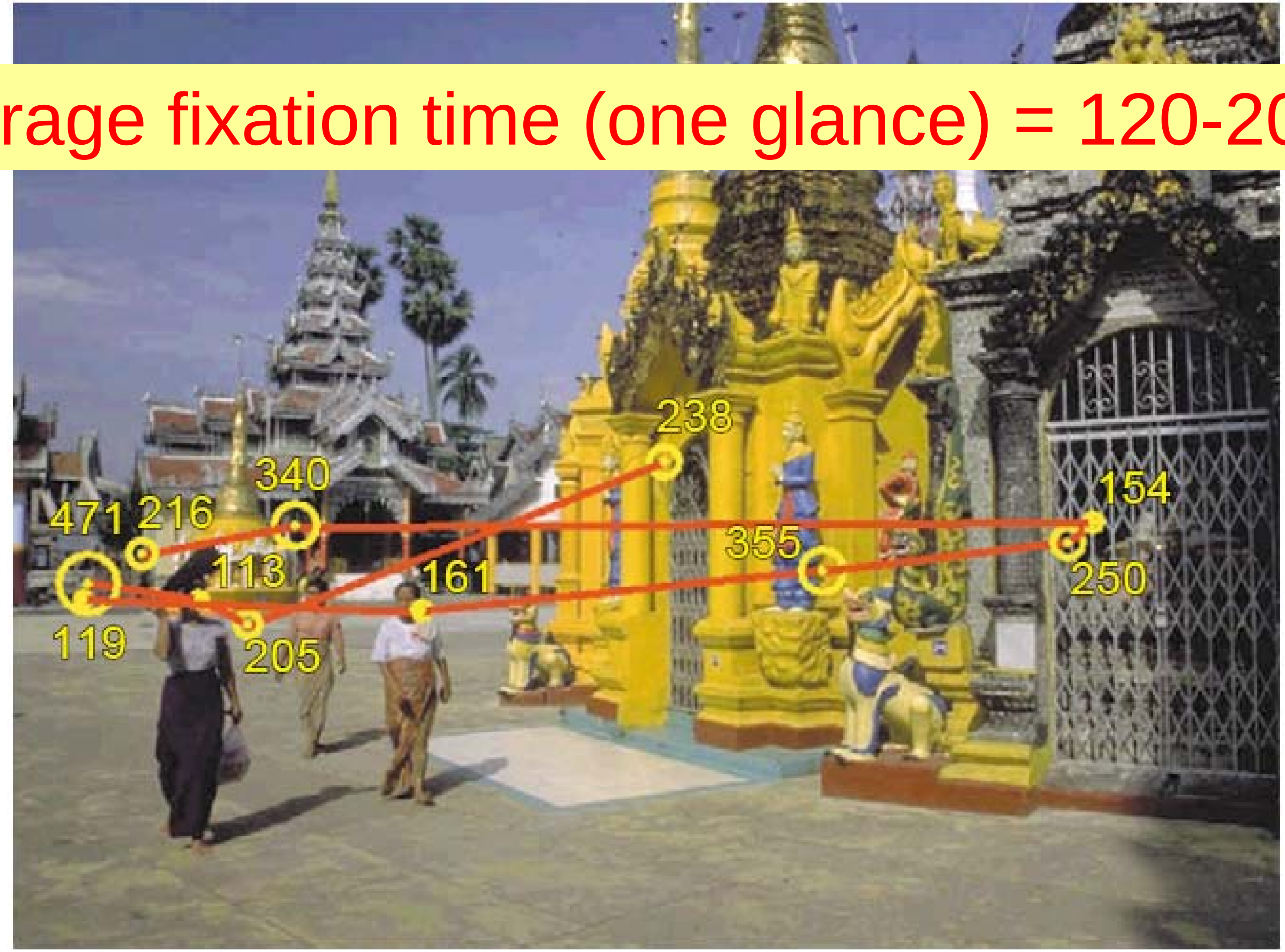
Please select one of "correct" or "incorrect" for each checked description. Click "Next>>" to continue

black/white_patches	☐ described	○ correct	○ incorrect
rectangular/square/box	☐ described	○ correct	○ incorrect
triangular/pyramidal	☐ described	○ correct	○ incorrect
elliptical/cylindrical(eg.round,blob)	☑ described	◉ correct	○ incorrect
curved_shape/contour(eg.arc,'S')	☐ described	○ correct	○ incorrect

Quit | Next Response | Next >>

What's in a glance?

Average fixation time (one glance) = 120-200ms

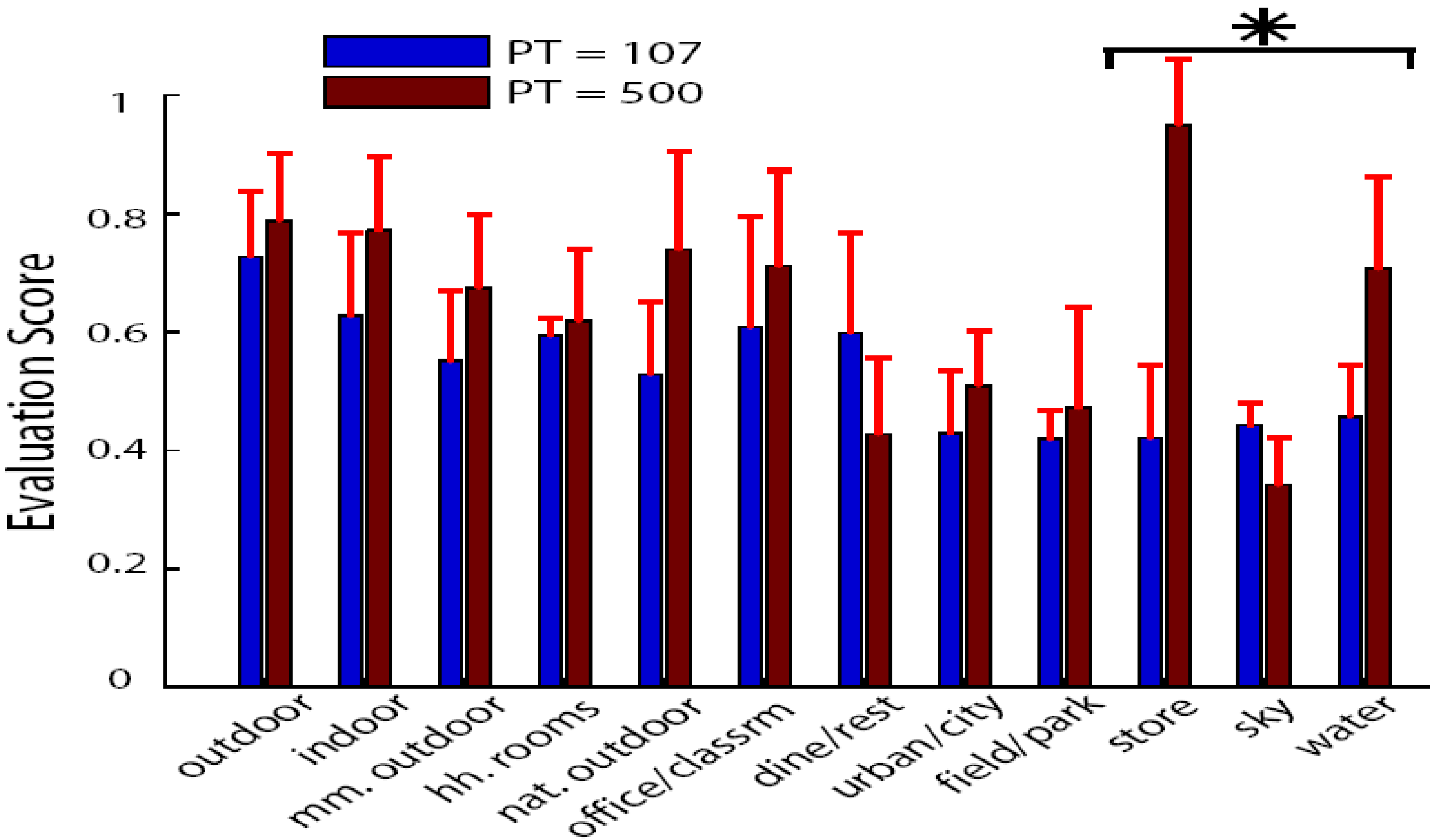


What's in a glance?

	PT = 107ms	PT = 500ms
	This is outdoors. A black, furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball. (Subject EC)	I saw a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out. (Subject JB)
	inside a house, like a living room, with chairs and sofas and tables, no ppl. (Subject HS)	A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left. (Subject RW)

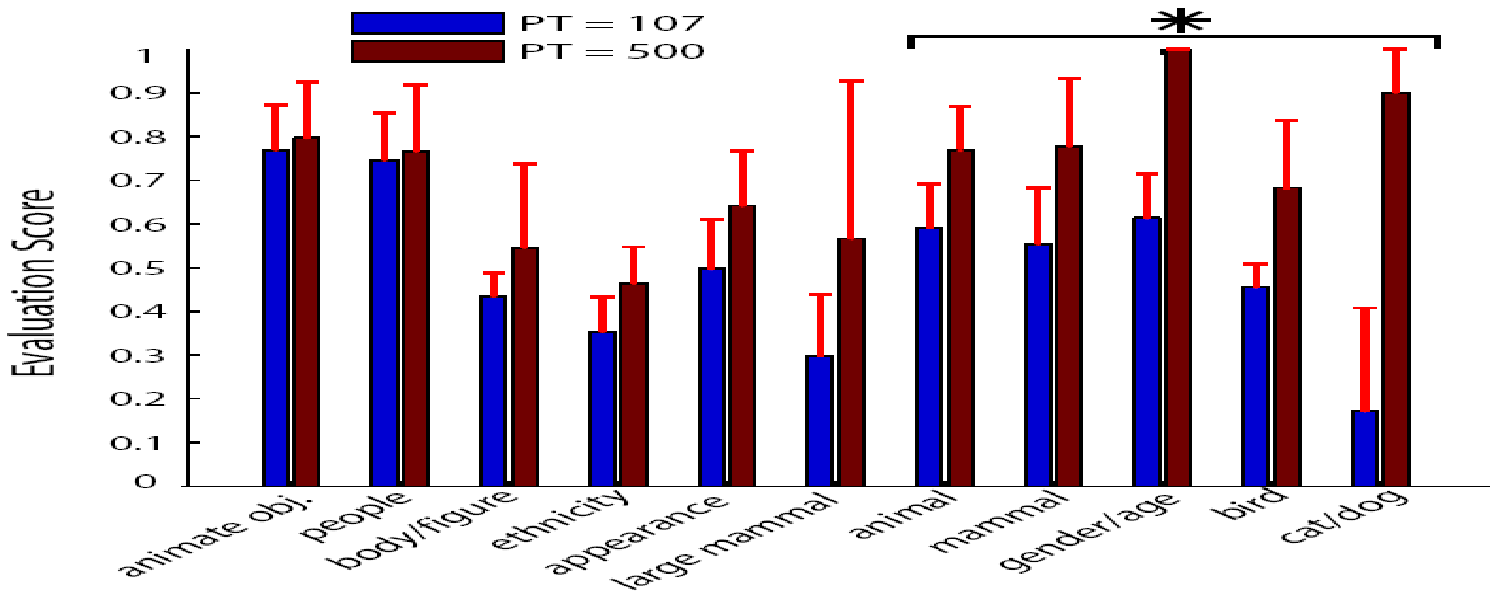
Fei-Fei et al. *JoV, 2007*

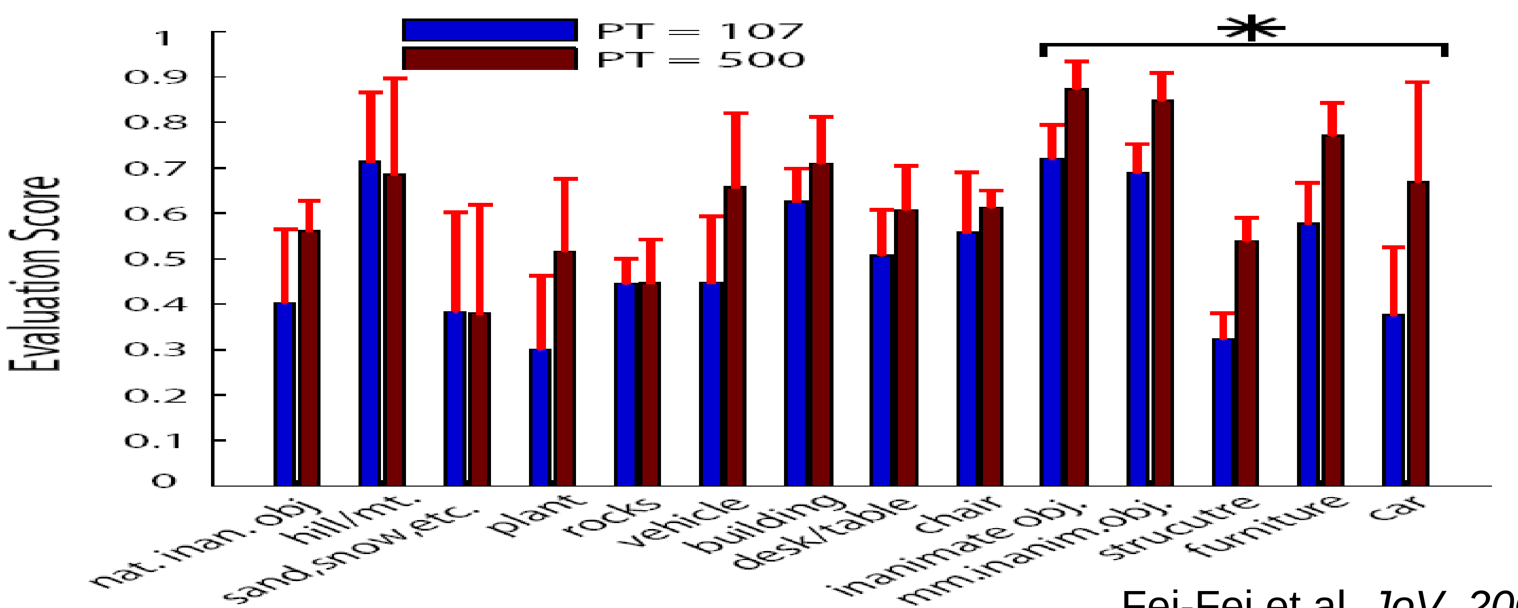
Scene level



Fei-Fei et al. *JoV, 2007*

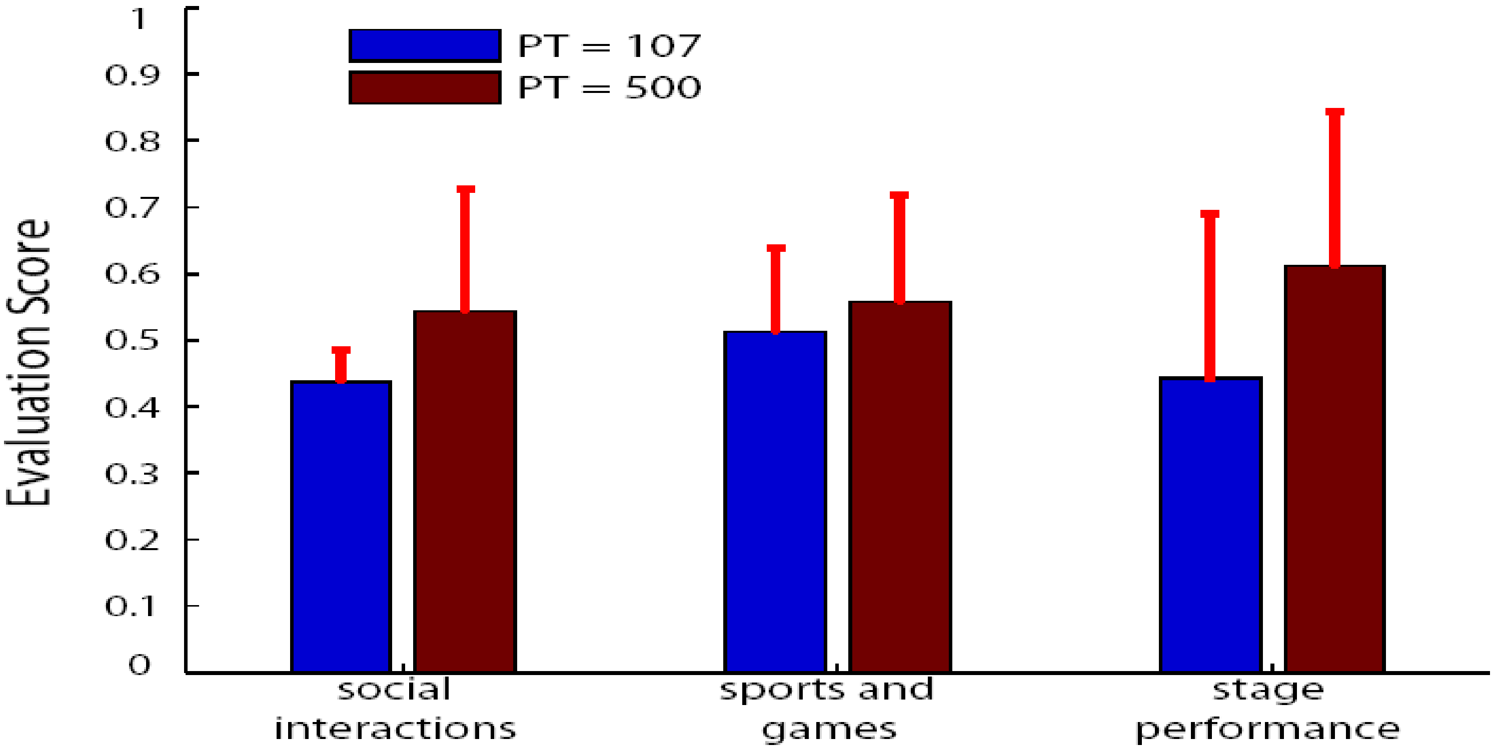
Object level





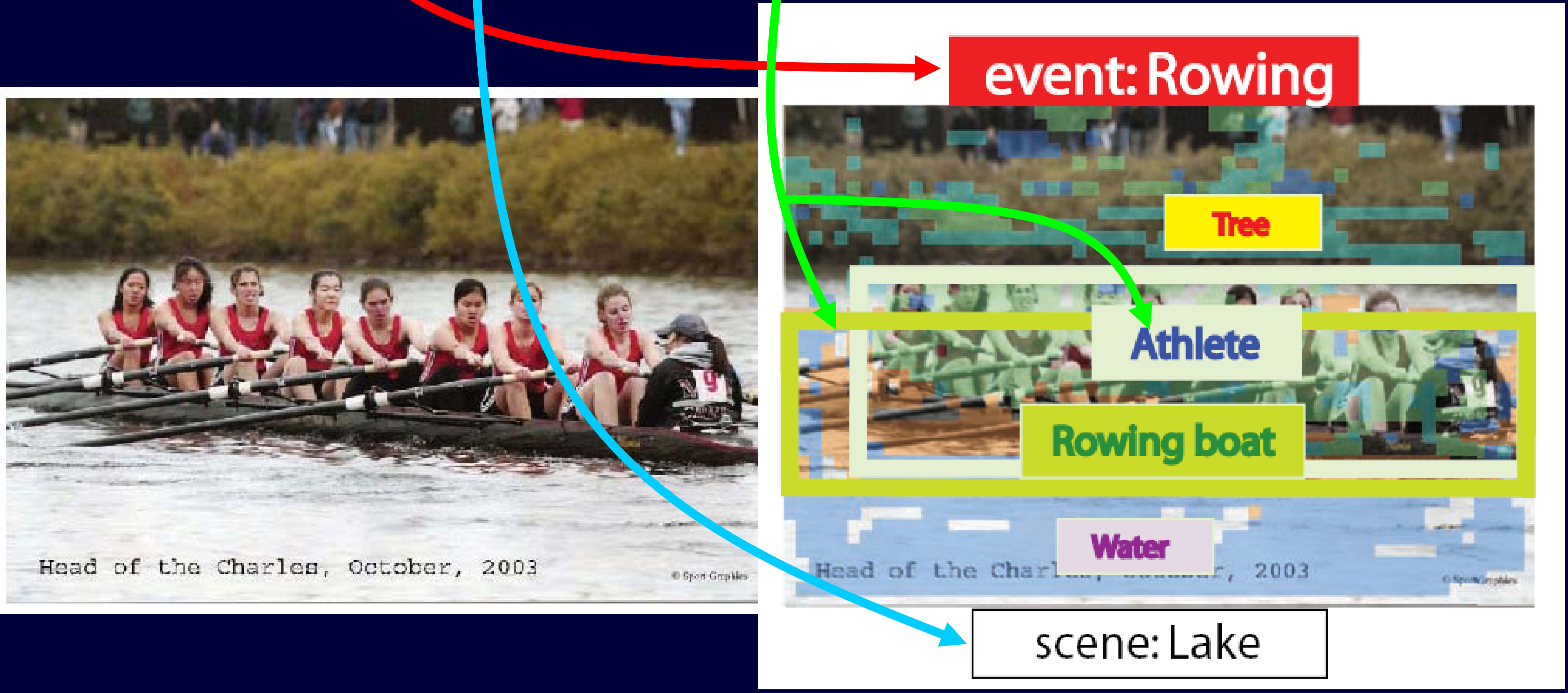
Fei-Fei et al. *JoV, 2007*

(Social) Events

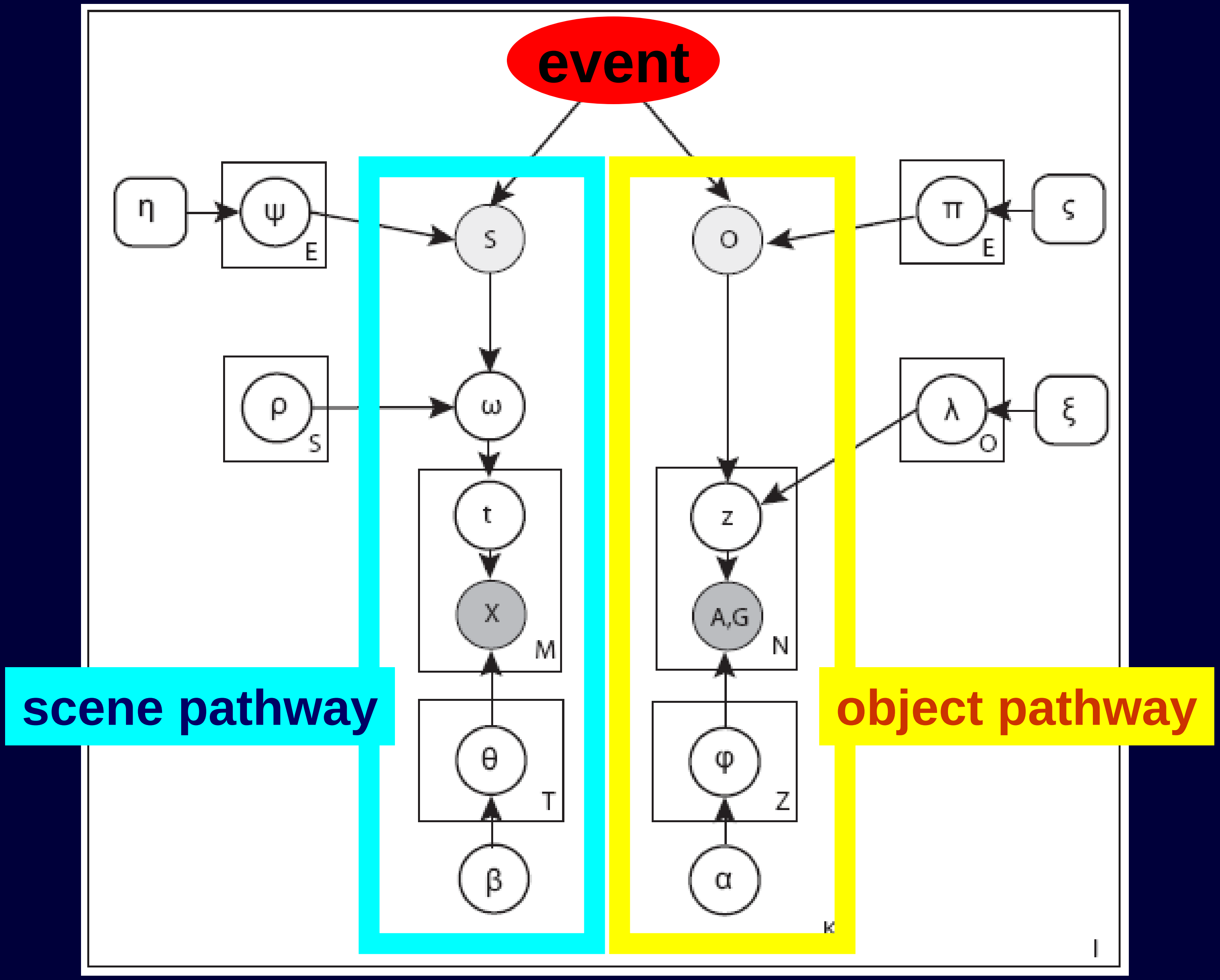


Fei-Fei et al. *JoV, 2007*

What, where and who? Classifying events by scene and object recognition

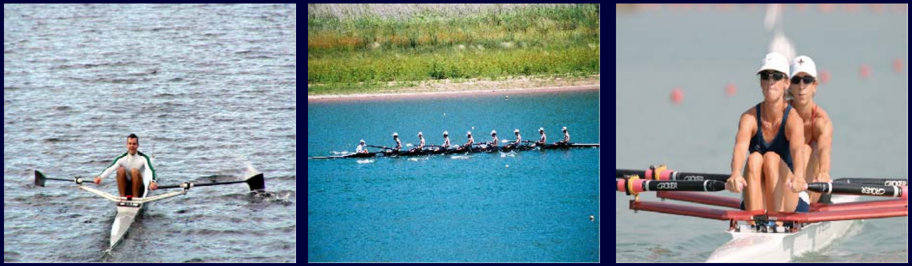


L.-J. Li & L. Fei-Fei ICCV 2007



L.-J. Li & L. Fei-Fei ICCV 2007

rowing



badminton

polo

sailing

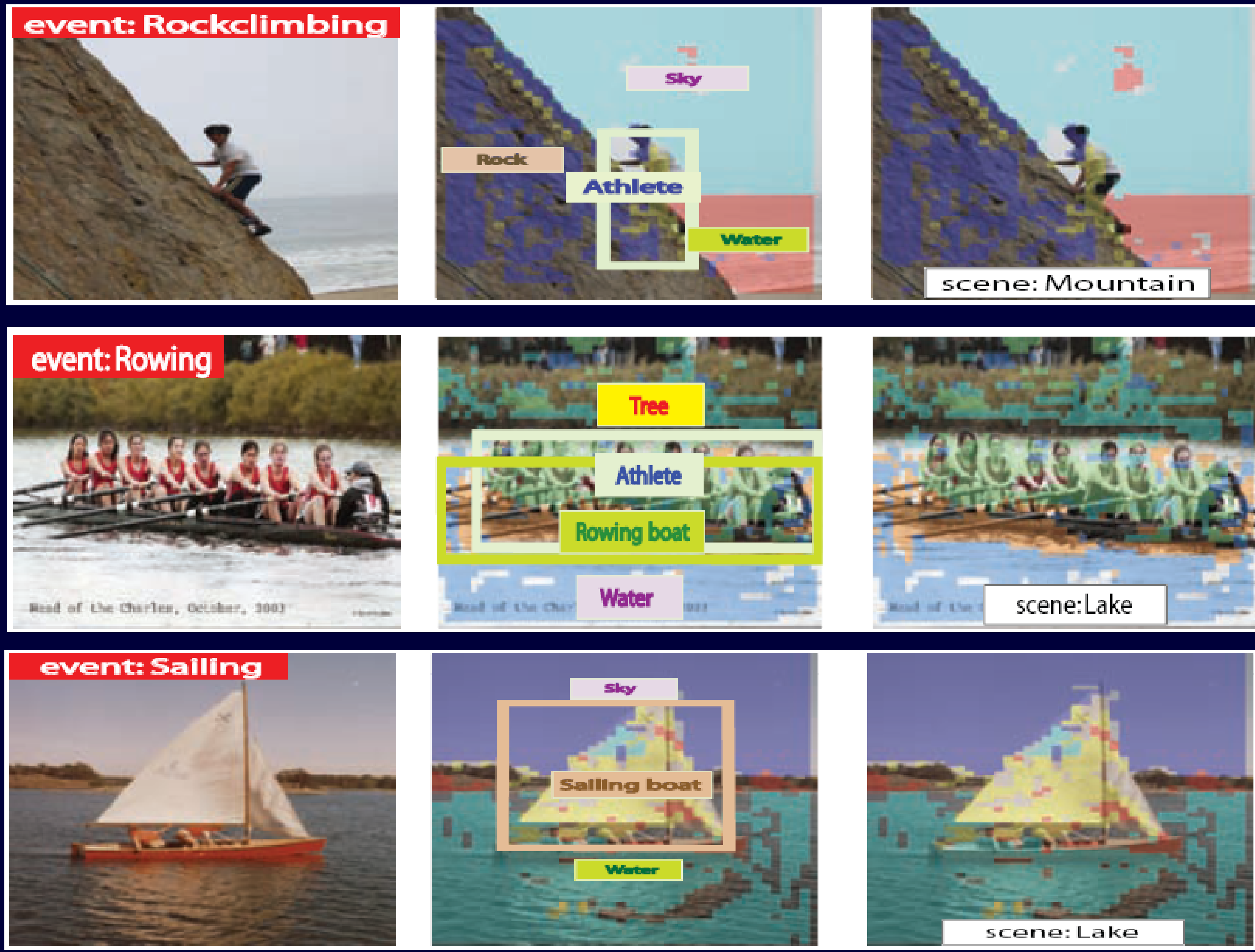
bocce

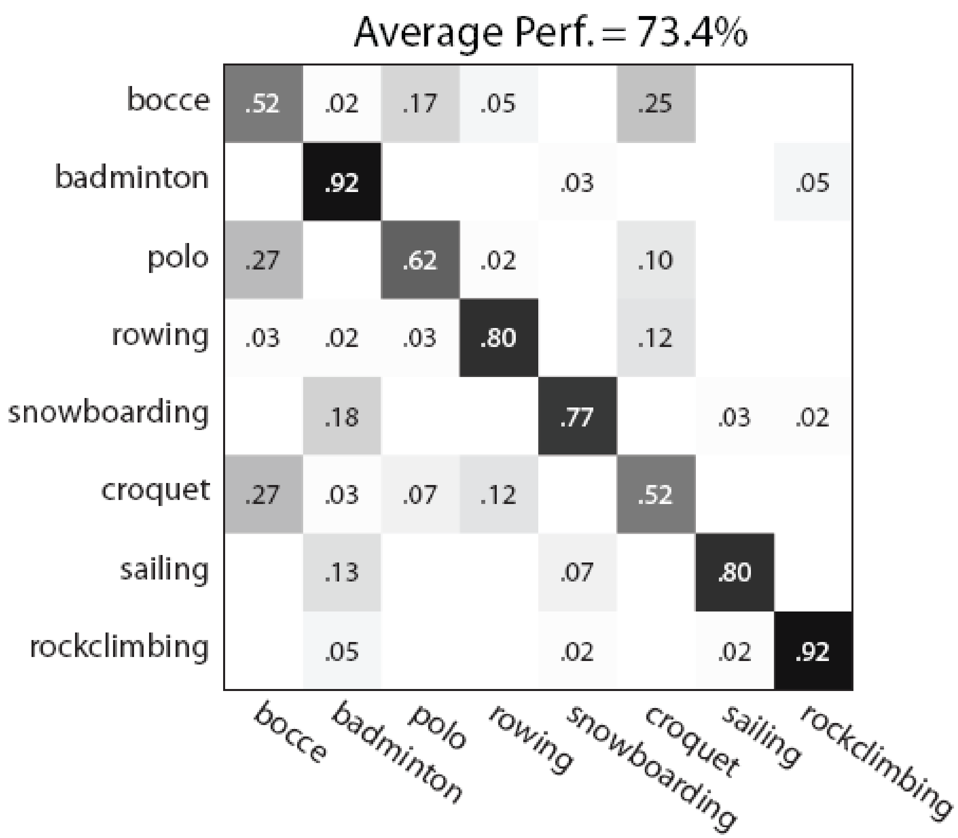
snow boarding

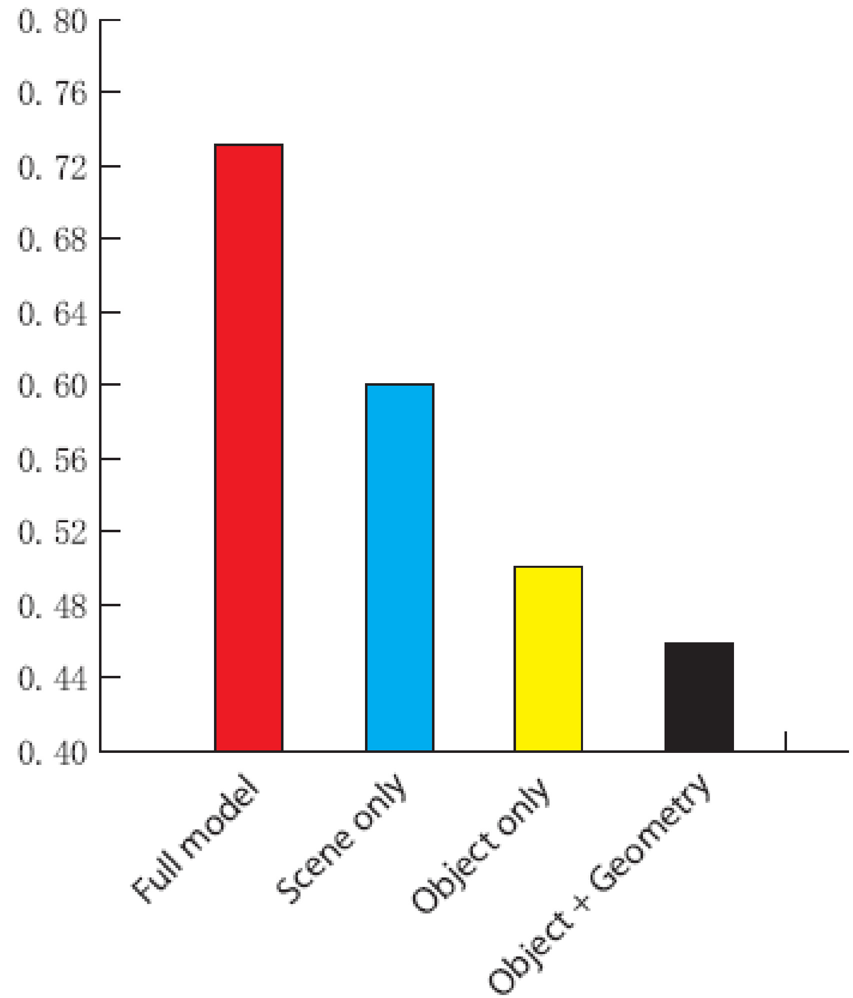
croquet

rock climbing


event: Badminton
Athlete
Floor
scene: Badminton court
event: Bocce
Athlete
Ground
scene: Bocce court
event: Croquet
Tree
Athlete
Grass
scene: Croquet court

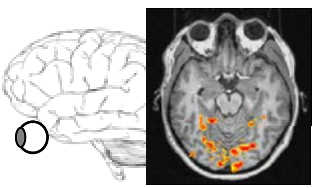

event: Rockclimbing
Sky
Rock
Athlete
Water
scene: Mountain
event: Rowing
Tree
Athlete
Rowing boat
Water
scene: Lake
event: Sailing
Sky
Sailing boat
Water
scene: Lake



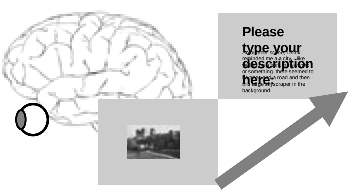


L.-J. Li & L. Fei-Fei ICCV 2007

#1: natural scene categorization entails little attention



#2: decoding the neural representation of natural scene categories

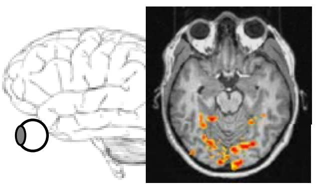


#3: what can we perceive within a glance of a scene?

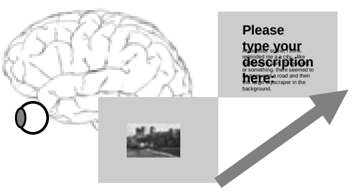
#4: Bayesian graphical models for natural scene categorization and event recognition

Thank you!

#1: natural scene categorization entails little attention (Rufin VanRullen, Pietro Perona, Christof Koch)



#2: decoding the neural representation of natural scene categories (Eamon Caddigan, Dirk Walther, Diane Beck)



#3: what can we perceive within a glance of a scene? (Asha Iyer, Pietro Perona, Christof Koch)

#4: Bayesian graphical models for natural scene categorization and event recognition (Pietro Perona, Li-Jia Li)