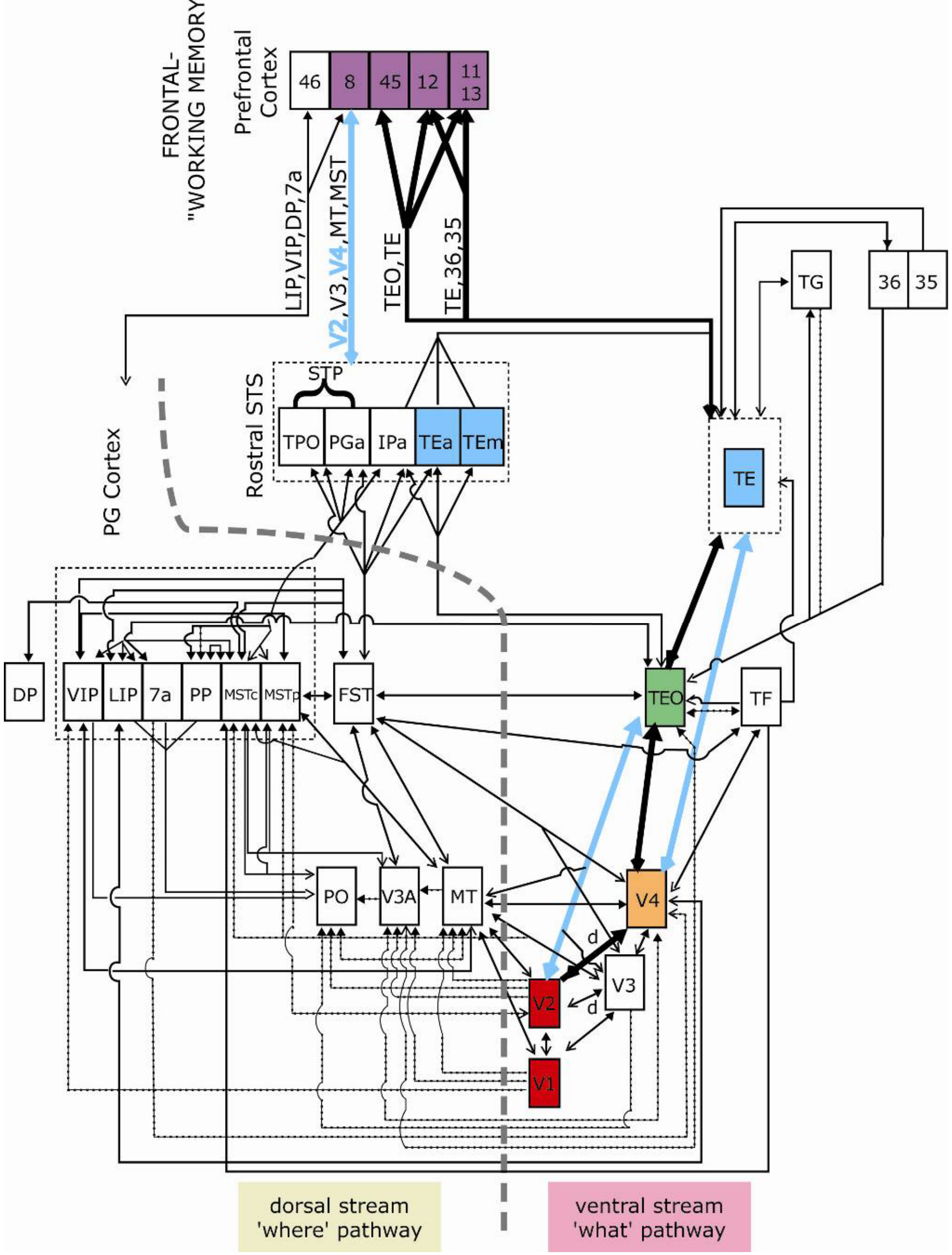
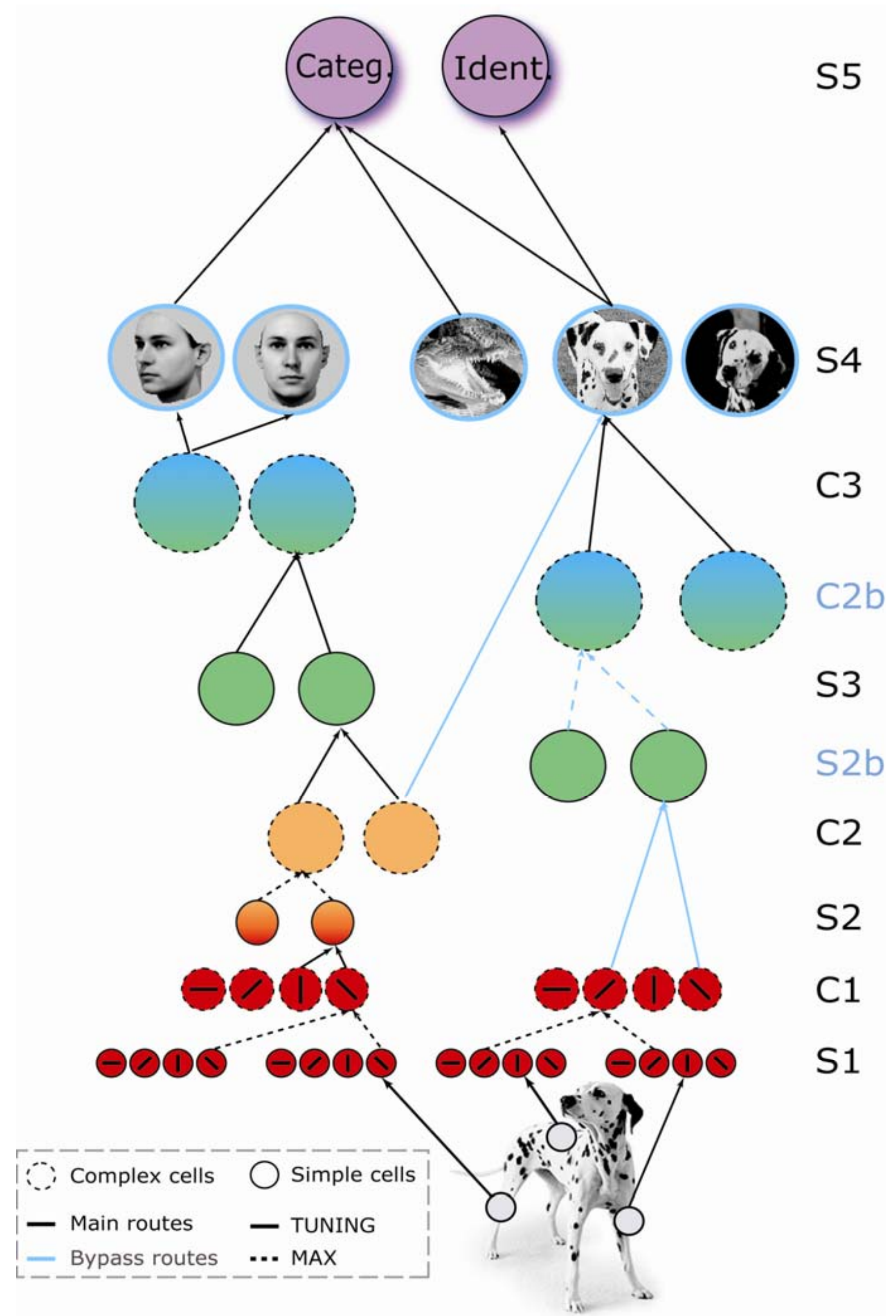
Feedforward theories of visual cortex predict human performance in rapid image categorization

Thomas Serre

Center for Biological and Computational Learning
McGovern Institute for Brain Research
Brain and Cognitive Sciences Department



Modified from (Ungerleider & VanEssen)

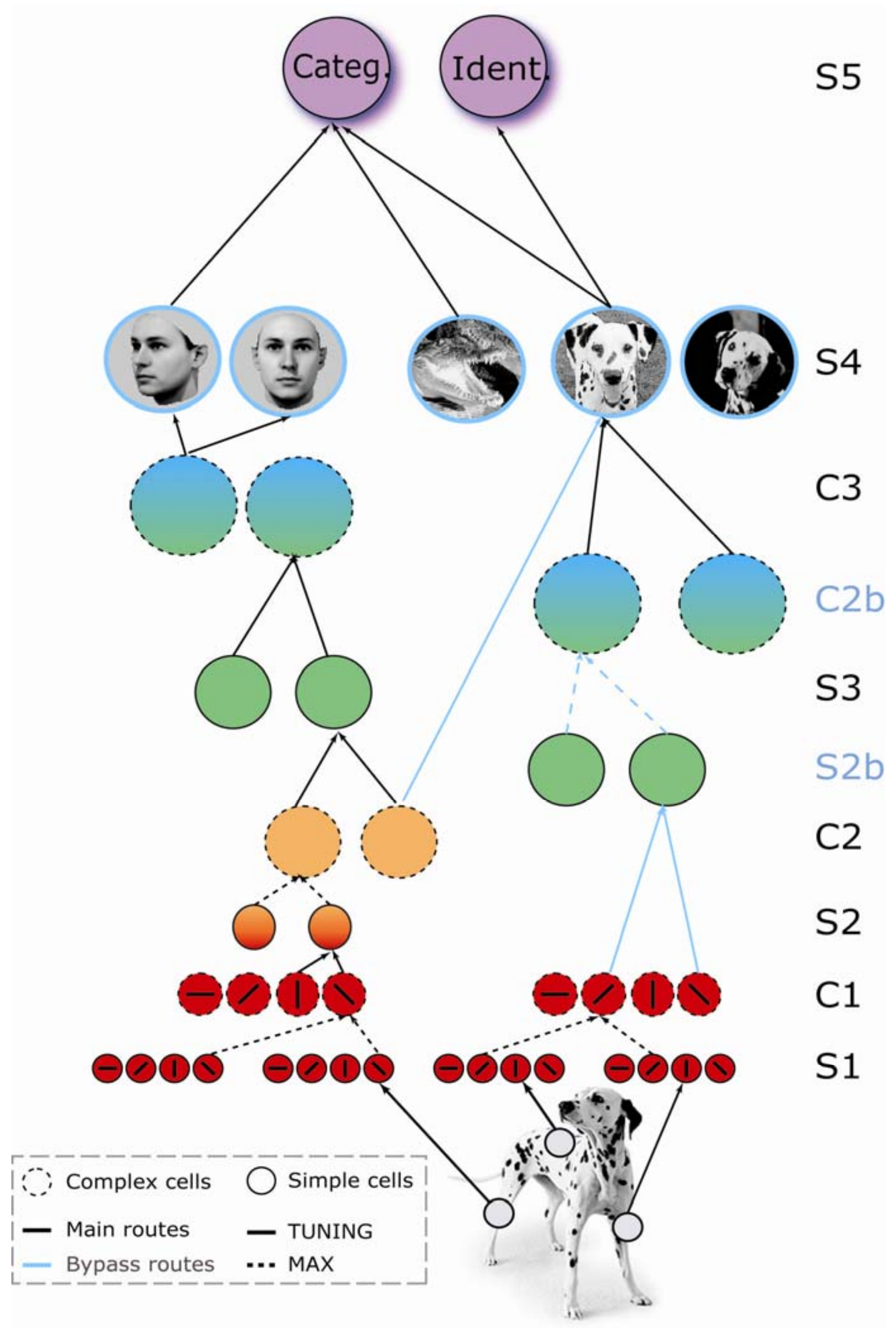


- **Builds upon previous neurobiological models**
  (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999)

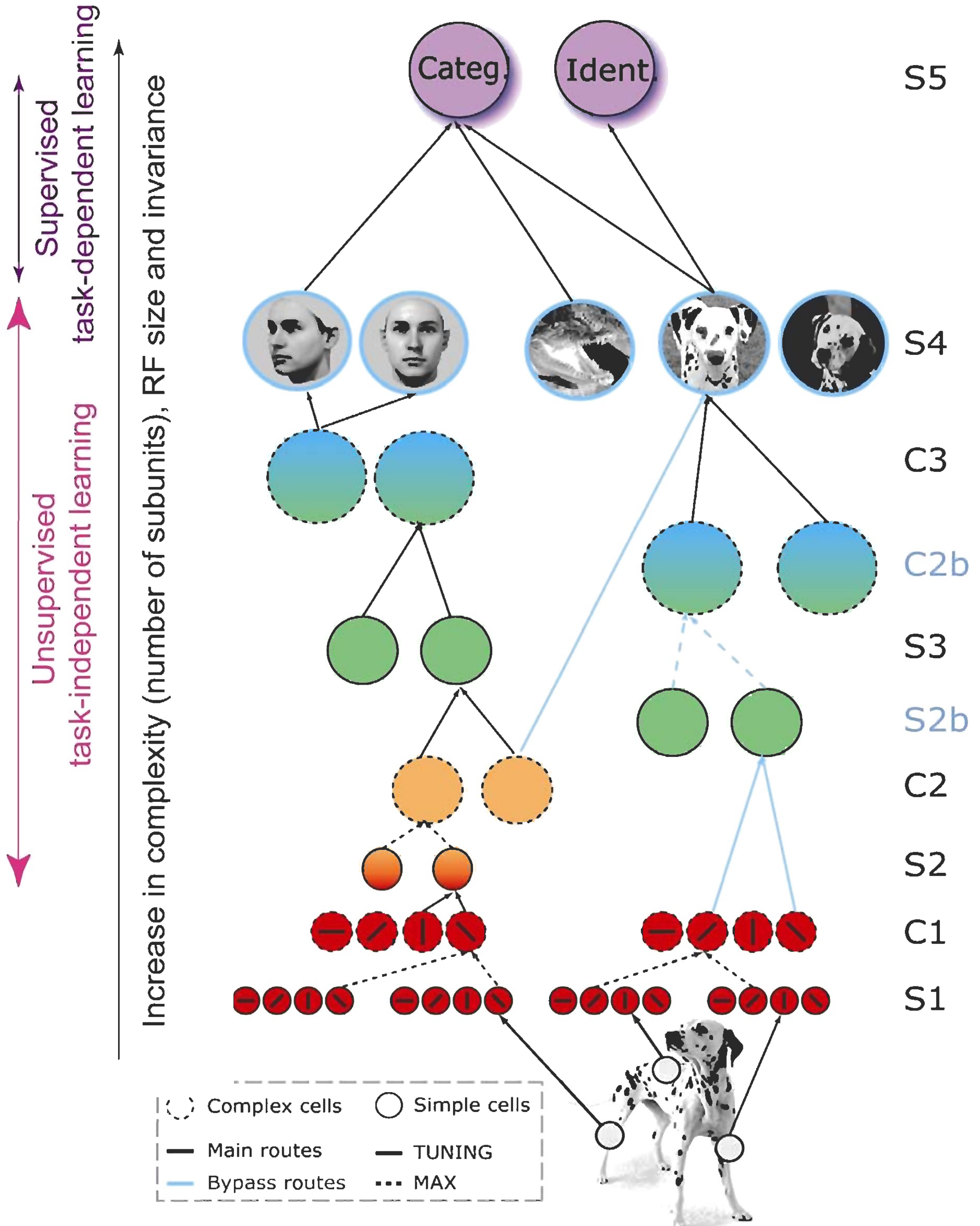
- **General class of feedforward hierarchical models of object recognition in cortex**

- **Biophysically plausible operations**

- **Predicts several properties of cortical neurons**
  (Serre, Kouh, Cadieu, Knoblich, Kreiman, Poggio, 2005)

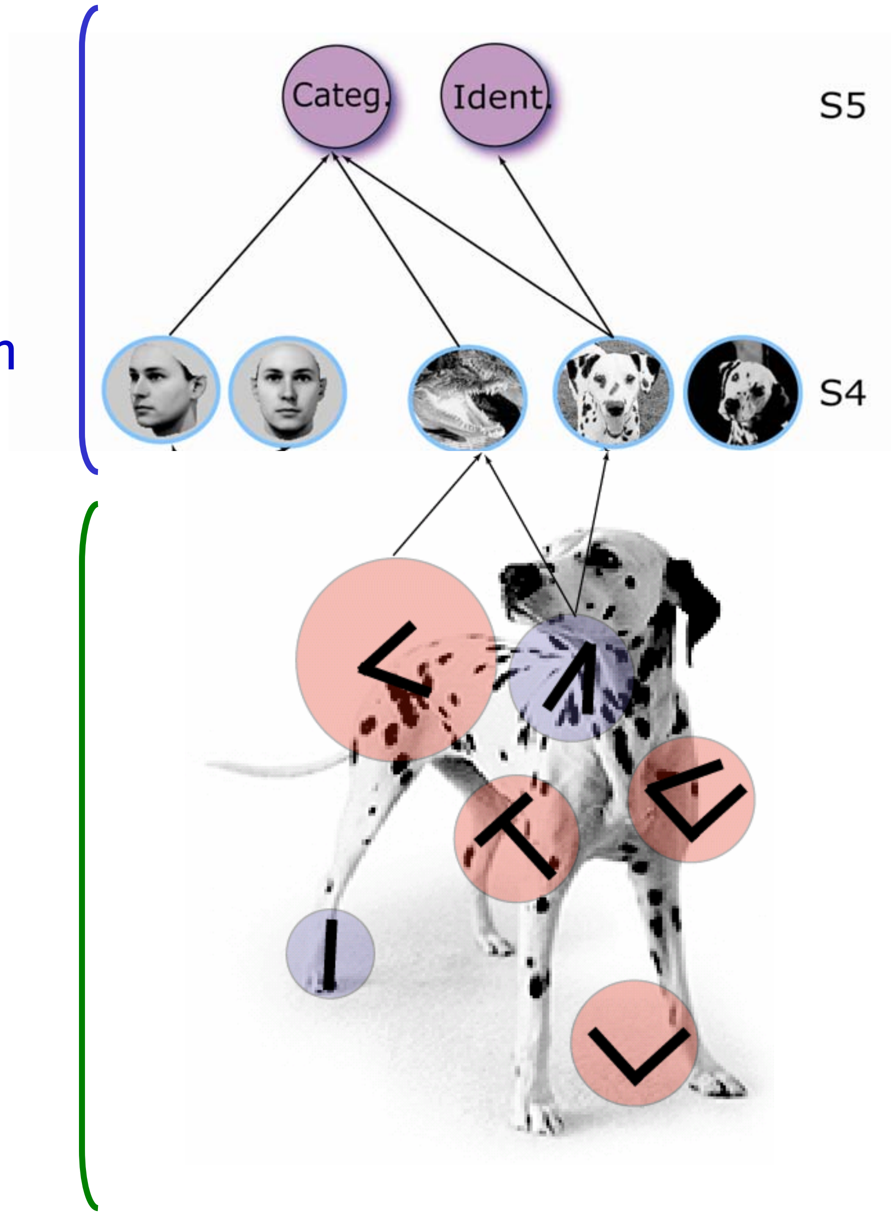


Model layers	Corresponding brain area (tentative)	RF sizes	Number units	
classifier	PFC		$1.0 \; 10^0$	
S4	AIT	$>4.4^o$	$1.5 \; 10^2$	~ 5,000 subunits
C3	PIT - AIT	$>4.4^o$	$2.5 \; 10^3$	
C2b	PIT	$>4.4^o$	$2.5 \; 10^3$	
S3	PIT	1.2^o- 3.2^o	$7.4 \; 10^4$	~ 100 subunits
S2b	V4 - PIT	0.9^o- 4.4^o	$1.0 \; 10^7$	~ 100 subunits
C2	V4	1.1^o- 3.0^o	$2.8 \; 10^5$	
S2	V2 - V4	0.6^o- 2.4^o	$1.0 \; 10^7$	~ 10 subunits
C1	V1 - V2	0.4^o- 1.6^o	$1.2 \; 10^4$	
S1	V1 - V2	0.2^o- 1.1^o	$1.6 \; 10^6$	

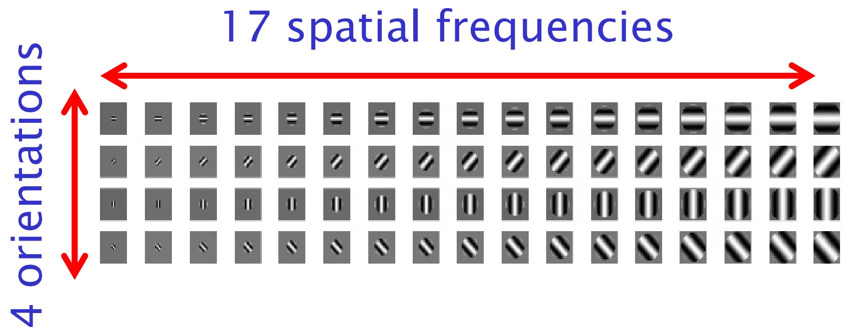


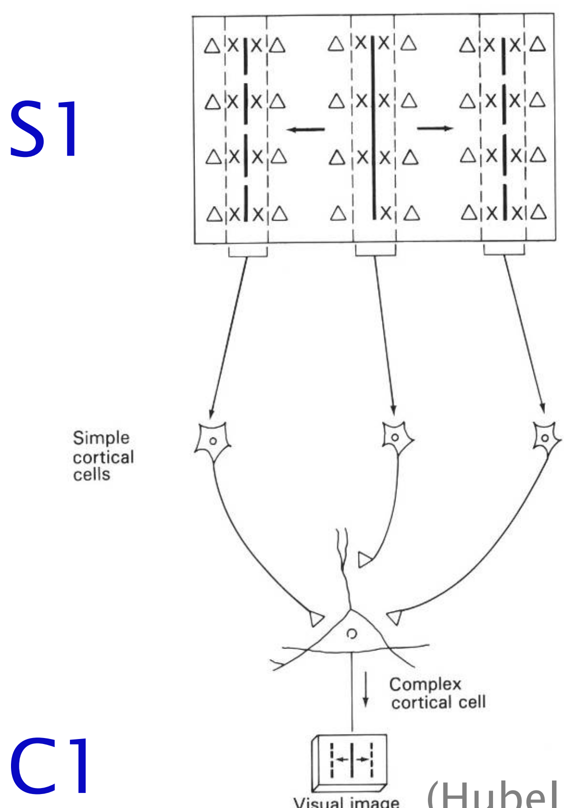
- Task-specific circuits (from IT to PFC)
  - Supervised learning
  - Linear classifier trained to minimize classification error on the training set (~ RBF net)

- Generic dictionary of shape components (from V1 to IT)
  - Unsupervised learning during developmental-like stage
  - From natural images unrelated to any categorization tasks

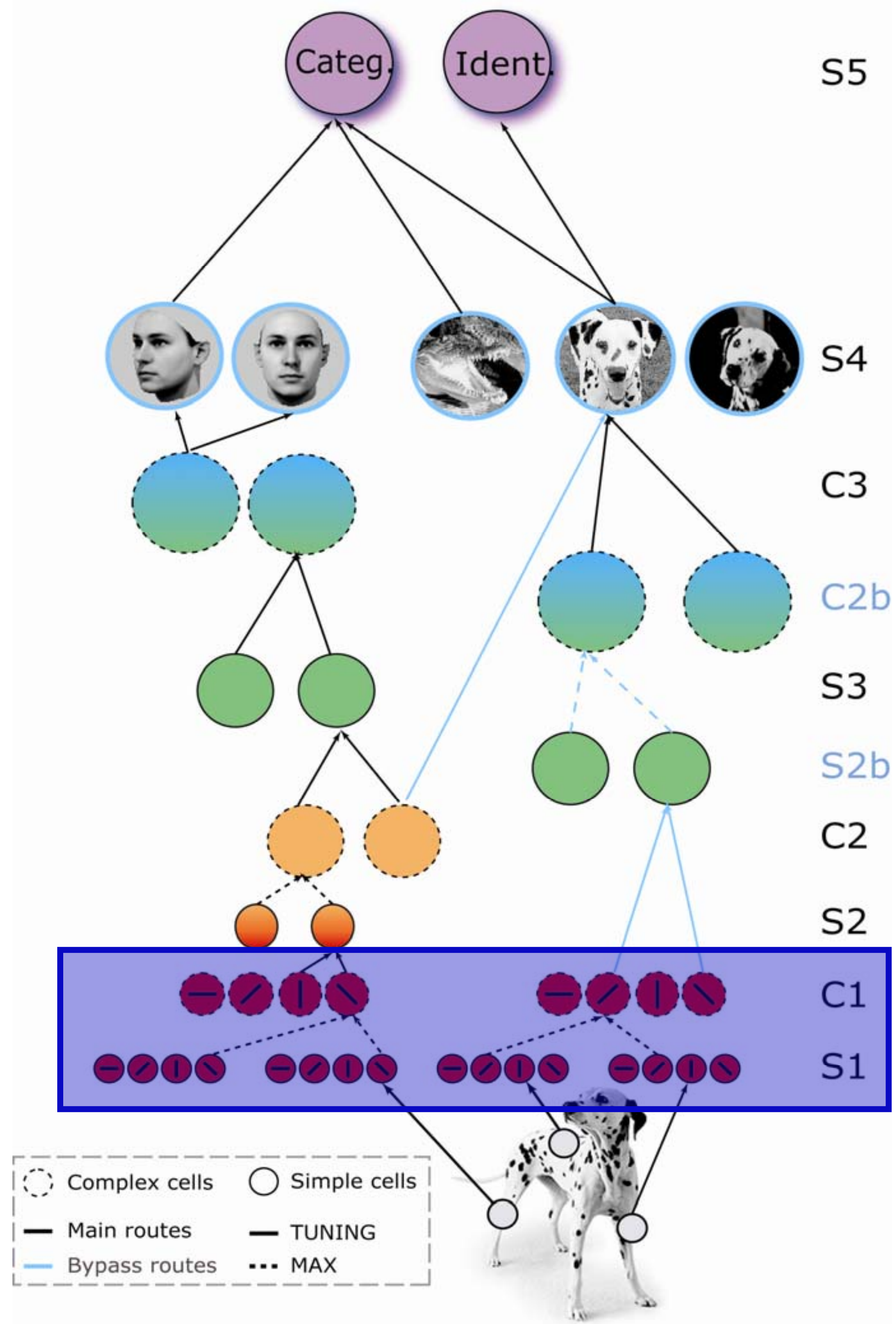


S1 and C1 units





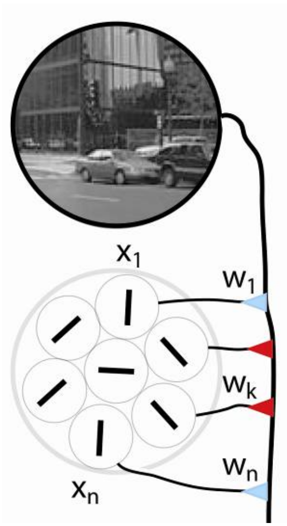
(Hubel & Wiesel, 1959)

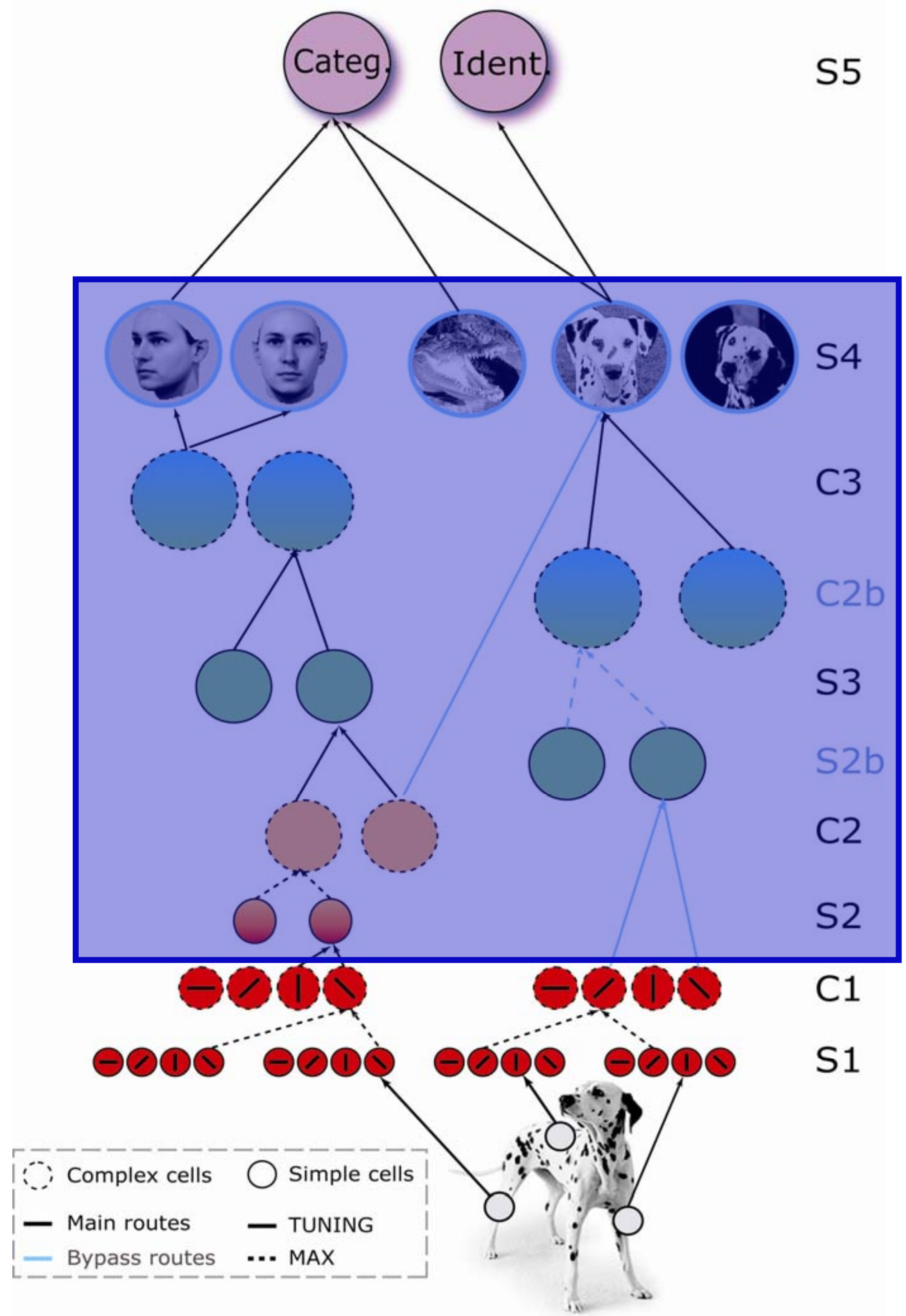


From S2 to S4

- Units are increasingly complex and invariant
- e.g, combination of V1-like complex units at different orientations

S2
unit

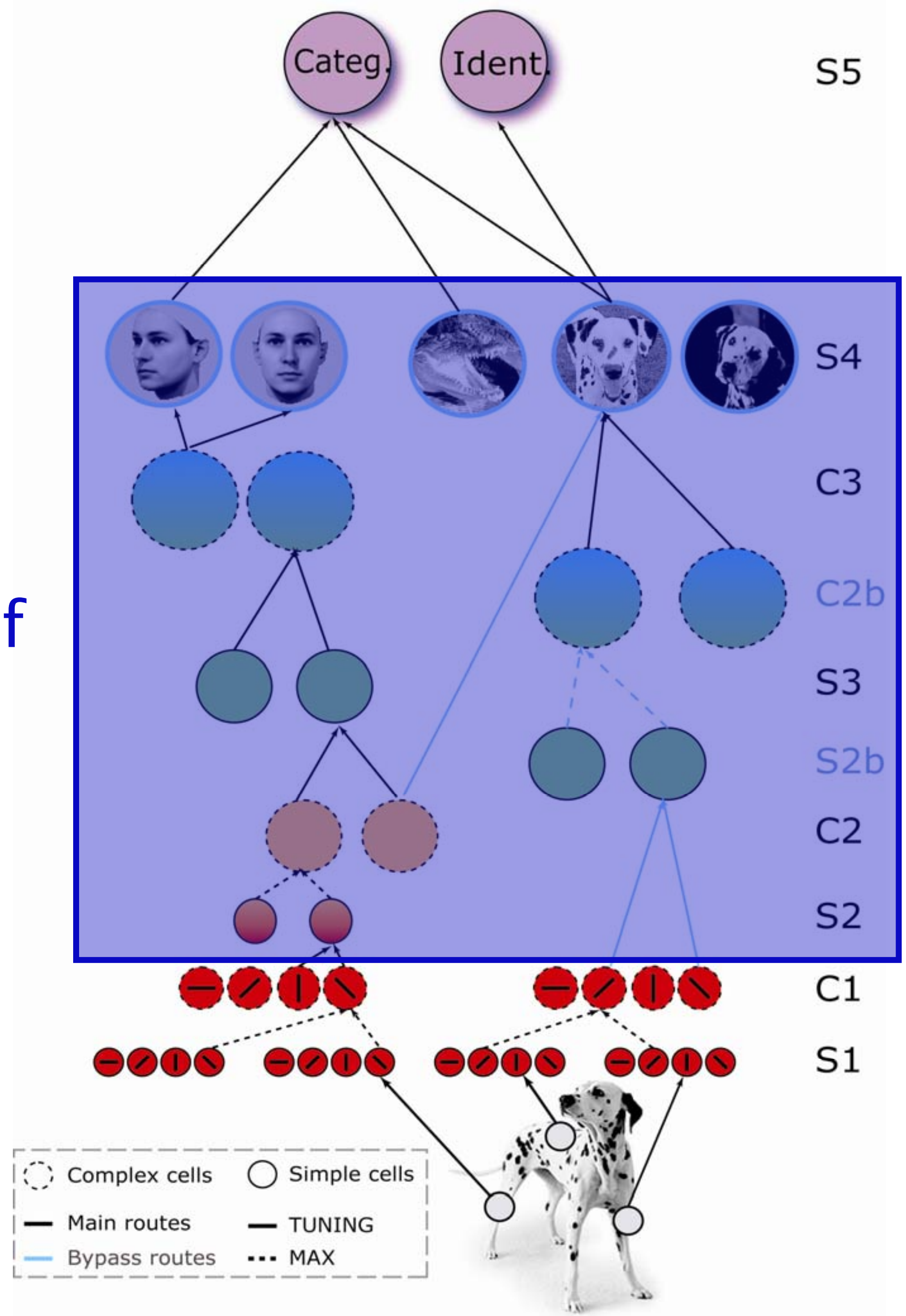




From C2 to S4

- 2,000 "features" at the C3 level ~ same number of feature columns in IT
  (Fujita et al, 1992)

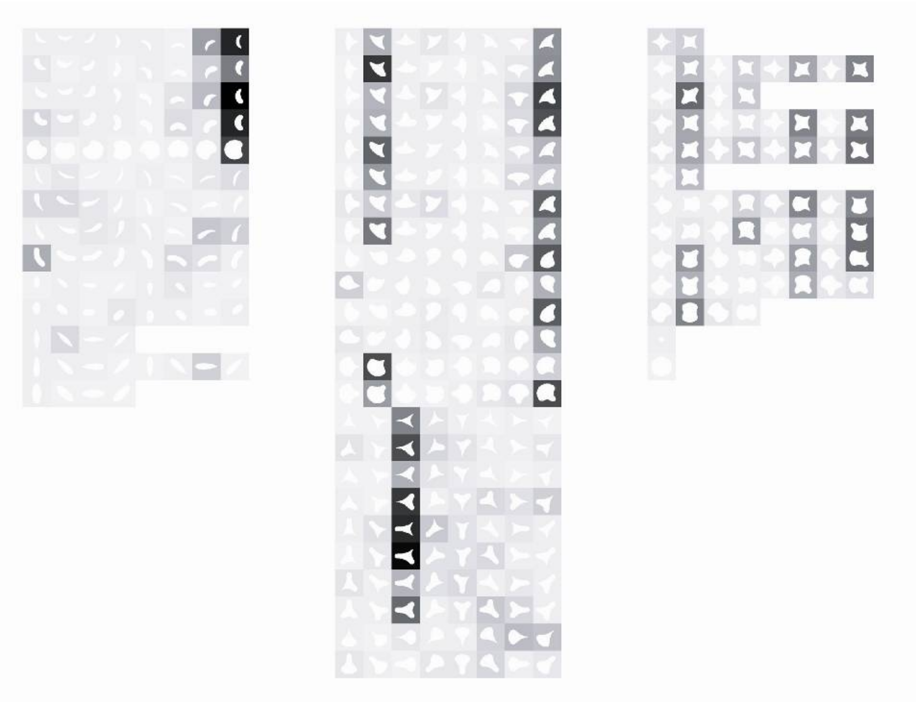
- Total ~6,000 types of features with various levels of complexity and invariance



The model predicts several properties of cortical neurons

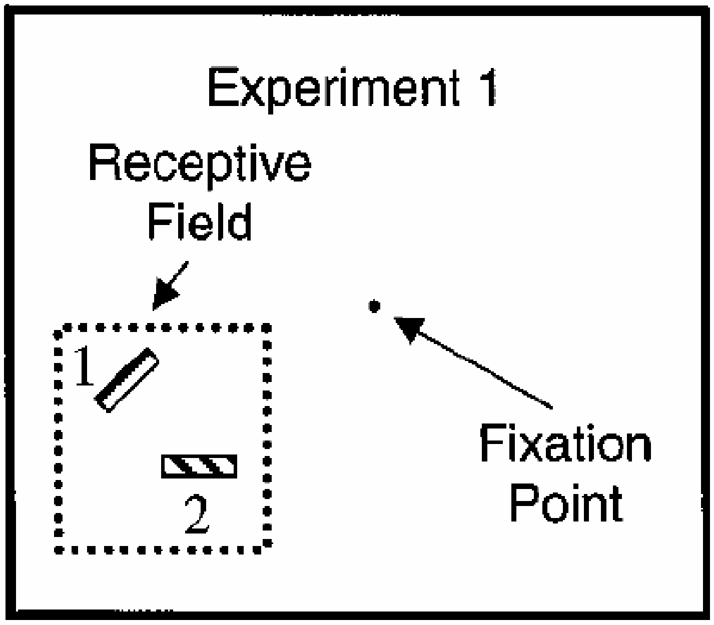
- In various cortical areas
- Examples from V4

Tuning for boundary conformation



(Pasupathy & Connor, 2001)

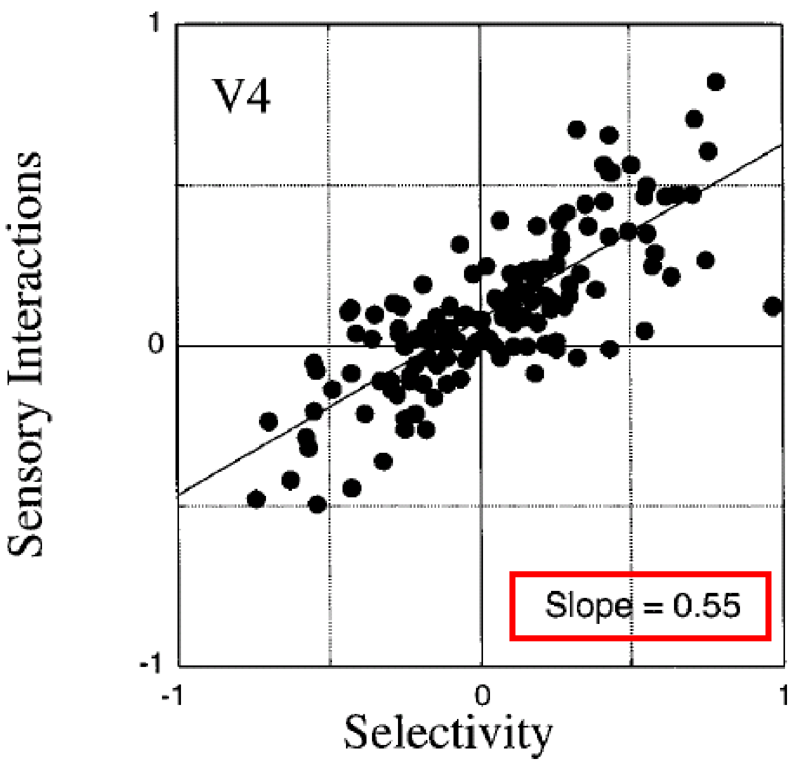
Tuning for two-bar stimuli



(Reynolds, Chelazzi and Desimone, 1999)

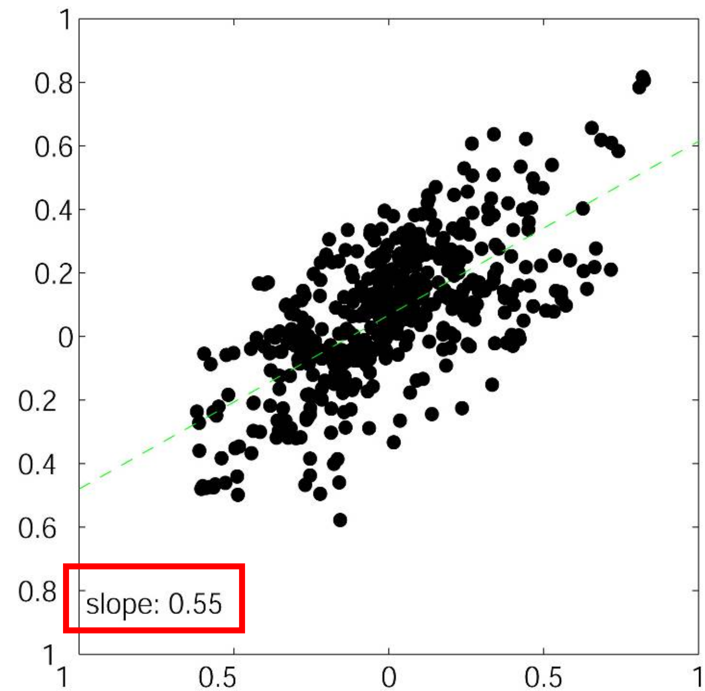
Prediction: Response of the pair is predicted to fall between the responses elicited by the stimuli alone

V4 neurons
(with attention directed away from receptive field)



(Reynolds , Chelazzi and Desimone, 1999)

C2 units



(Serre, Kouh, Cadieu, Knoblich, Kreiman and Poggio, 2005)

The model can perform complex recognition task very well

- At the level of some of the best computer vision systems
- e.g, constellation models
  (Leung et al, 1995; Burl et al, 1998; Weber et al., 2000; Fergus et al, 2003; Li et al, 2004)

rear-car airplane frontal face motorbike leaf

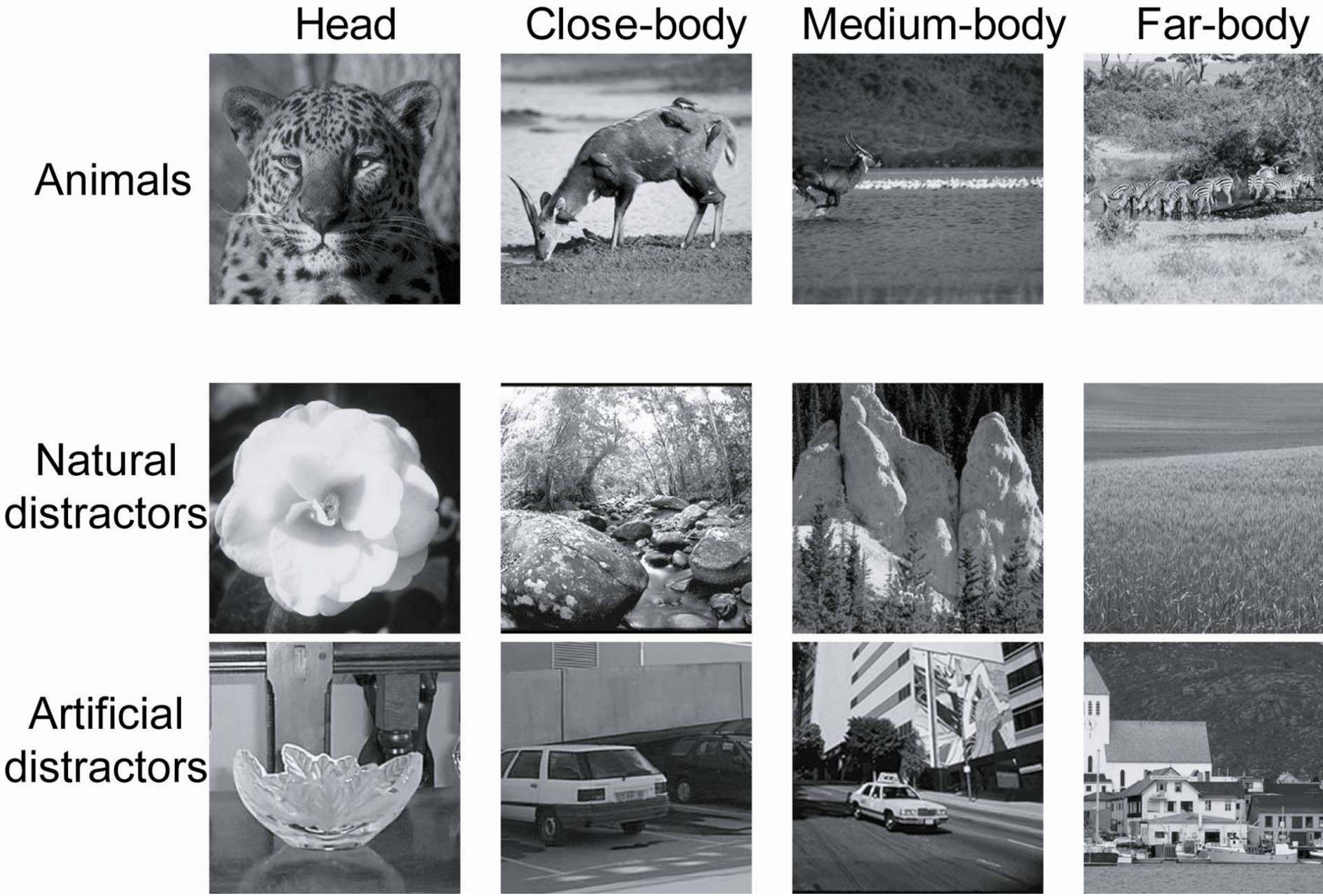
Datasets			AI systems	Model
(CalTech)	Leaves	[Weber et al., 2000b]	84.0	97.0
(CalTech)	Cars	[Fergus et al., 2003]	84.8	99.7
(CalTech)	Faces	[Fergus et al., 2003]	96.4	98.2
(CalTech)	Airplanes	[Fergus et al., 2003]	94.0	96.7
(CalTech)	Motorcycles	[Fergus et al., 2003]	95.0	98.0

How does the model compare to human observers?

Animal vs. non-animal categ.

- 1,200 stimuli (from Corel database)
- 600 animals in 4 categories:
  - Head
  - Close-body
  - Medium-body
  - Far-body and groups
- 600 matched distractors (½ art., ½ nat.) to prevent reliance on low-level cues

(Torralba & Oliva, 2003; Oliva & Torralba, in press)



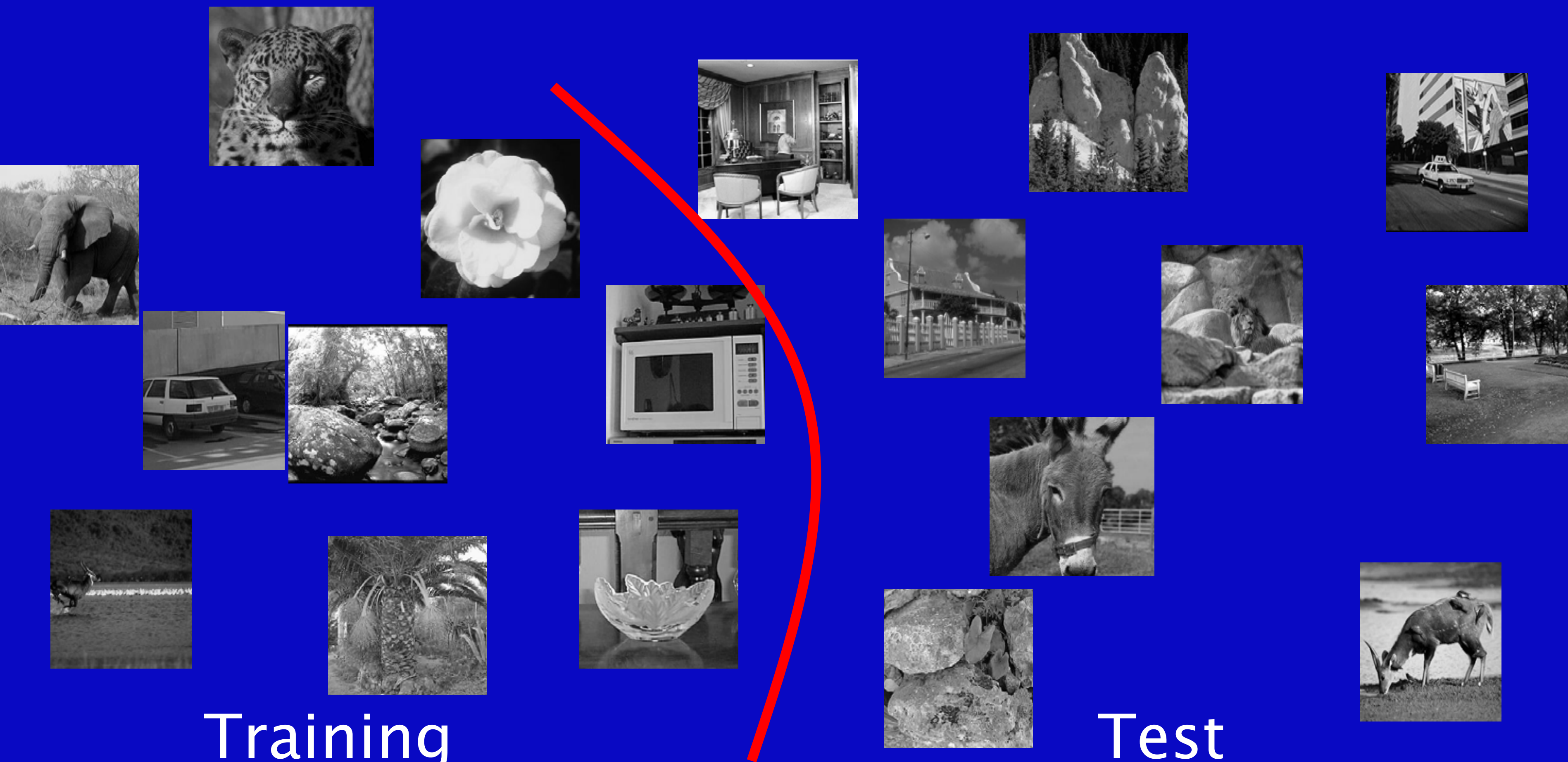
(Torralba & Oliva, 2003; Oliva & Torralba, in press)

Training and testing the model

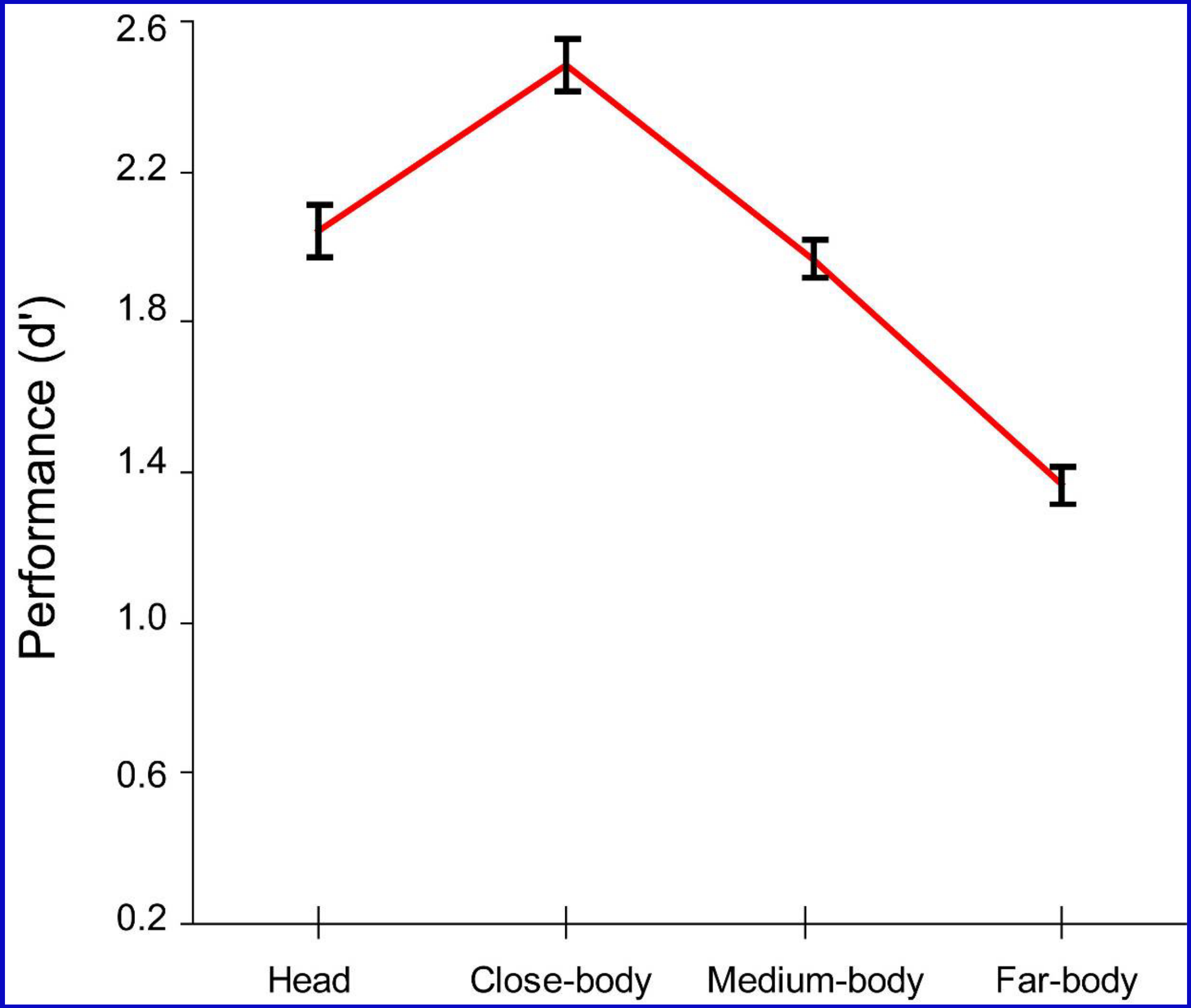
- ➢ Random splits (good estimate of expected error)
- ➢ Split 1,200 stimuli into two sets

Training the model

- Repeat 20 times
- Average model performance over all

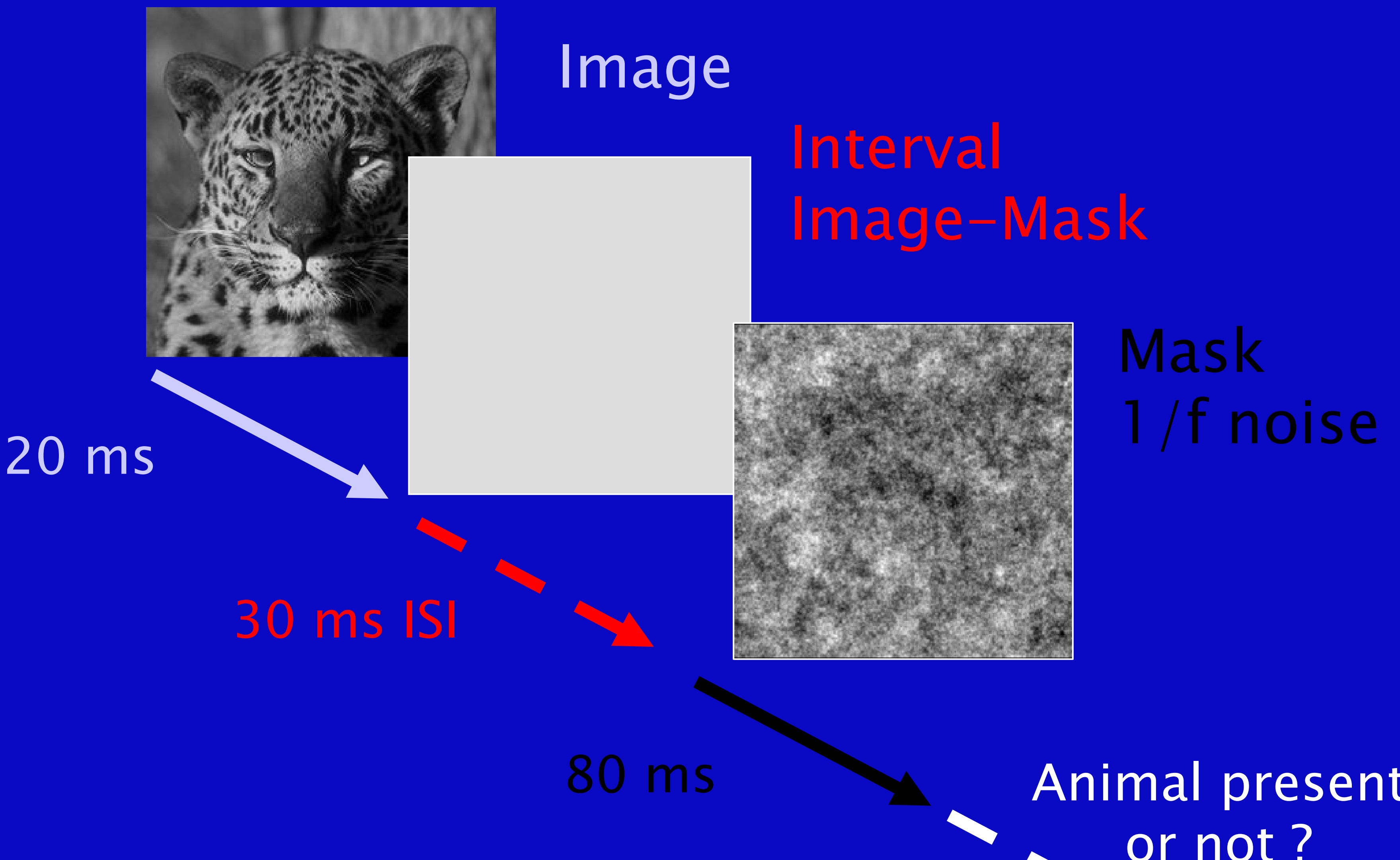


Results: Model



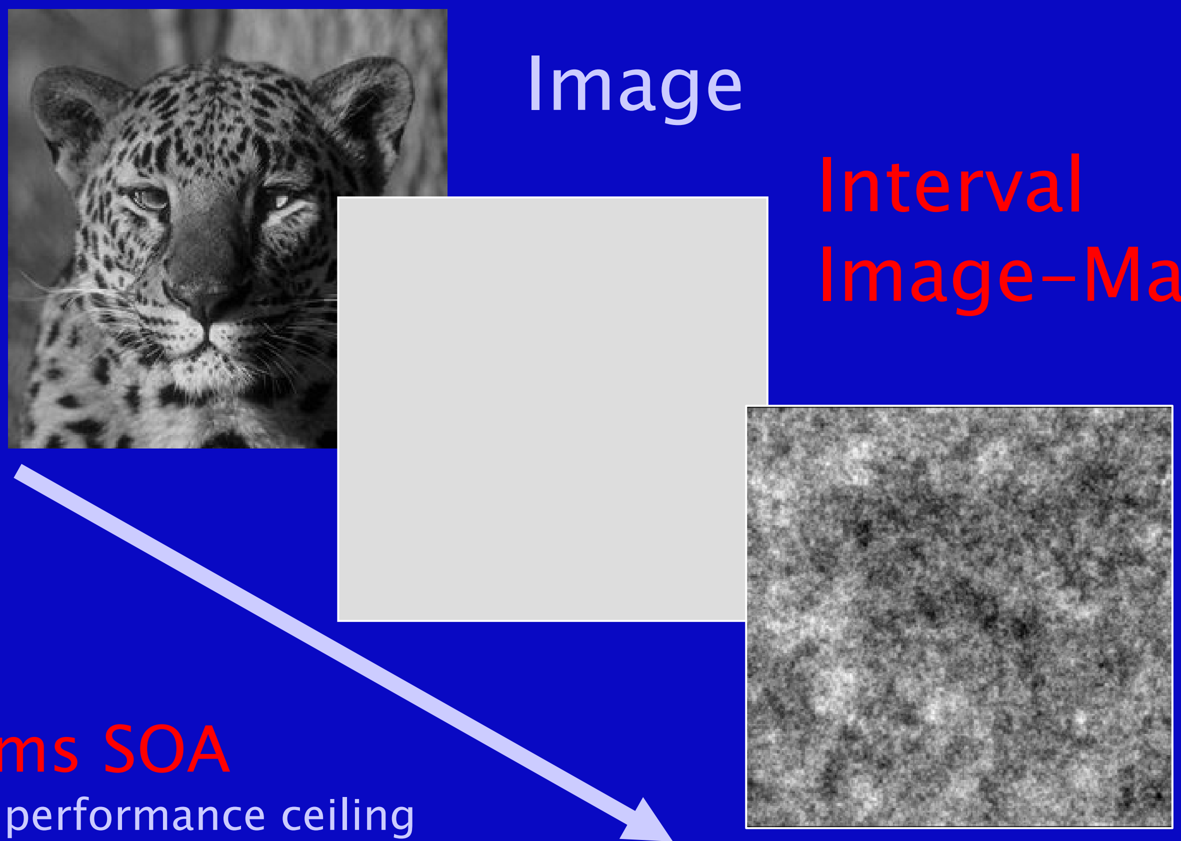
model

Rapid categorization task



(Thorpe et al, 1996; Van Rullen & Koch, 2003; Bacon-Mace et al, 2005; Oliva & Torralba, in press)

Rapid categorization task



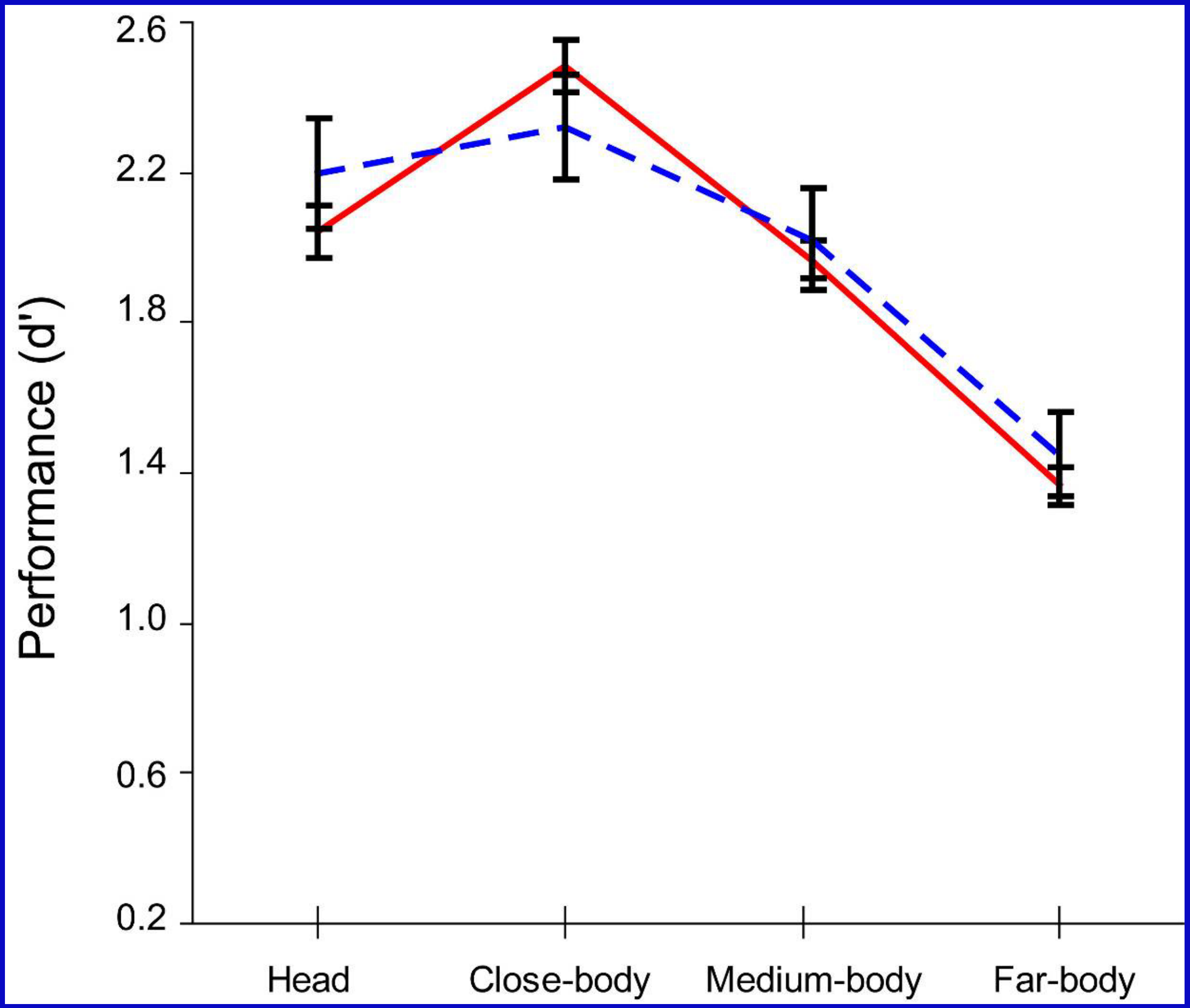
~ 50 ms SOA
close to performance ceiling
in (Bacon-Mace et al, 2005)

80 msec

Animal present
or not ?

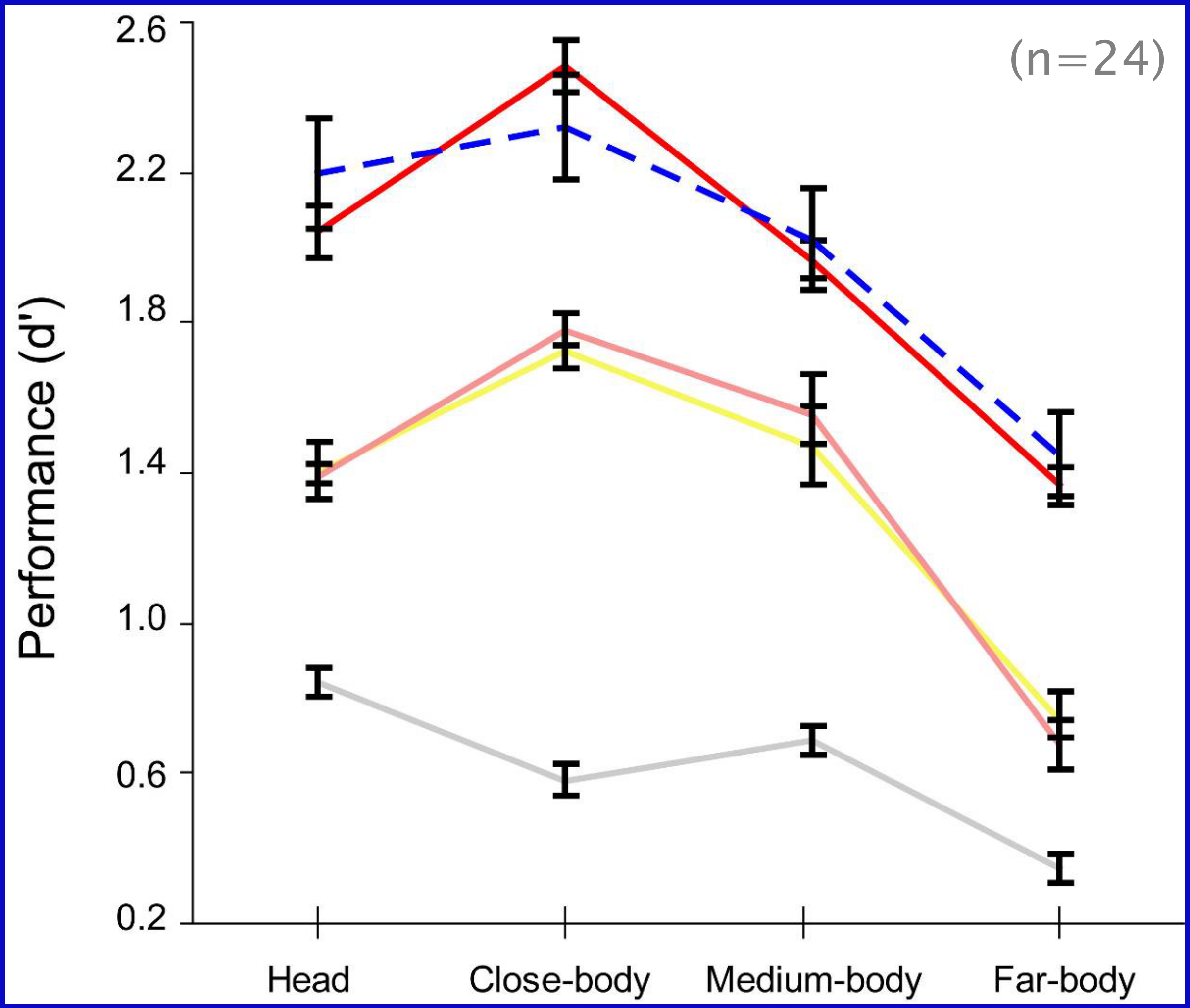
(Thorpe et al, 1996; VanRullen & Koch, 2003;
Bacon-Mace et al, 2005; Oliva & Torralba, in press)

Results: Human-observers



50 ms SOA (ISI=30 ms)
model

“Simpler” models cannot do the job



50 ms SOA (ISI=30 ms)

model

Model C1

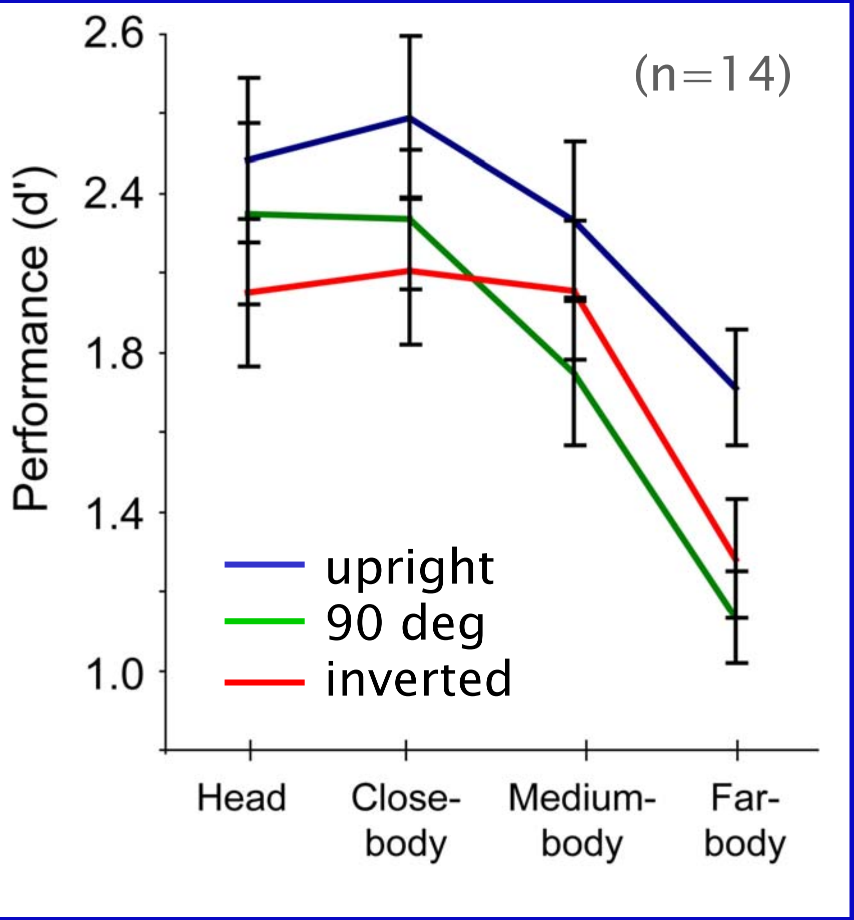
(Torralba & Oliva, 2001)

(Renninger & Malik, 2004)

(Serre, Oliva and Poggio, in prep)

Results: Image orientation

Human observers

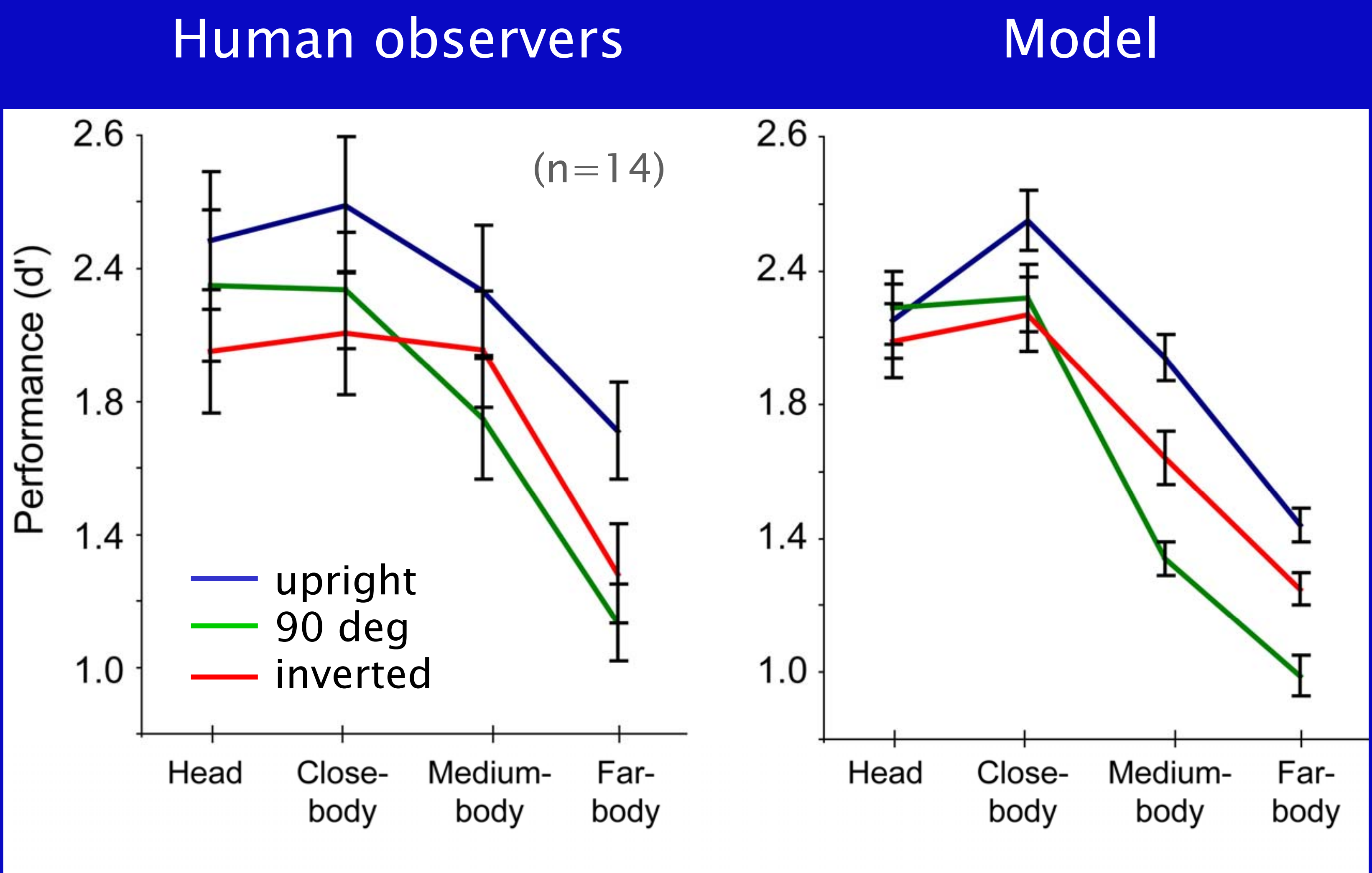


50 ms SOA (ISI=30 ms)

Robustness to image orientation is in agreement with previous results
(Rousselet et al, 2003; Guyonneau et al, ECVP 2005)

(Serre, Oliva and Poggio, in prep)

Results: Image orientation



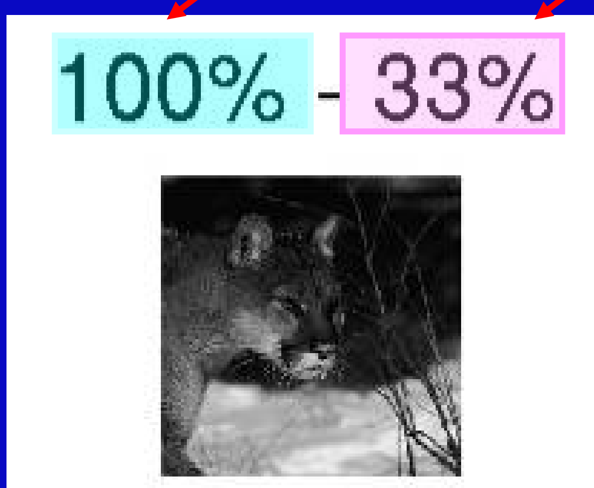
50 ms SOA (ISI=30 ms)

(Serre, Oliva and Poggio, in prep)

Detailed comparison

- For each individual image
- How many times image classified as animal:
  - For humans: across subjects
  - For model: across 20 runs

model humans



- Heads: $\rho=0.71$
- Close-body: $\rho=0.84$
- Medium-body: $\rho=0.71$
- Far-body: $\rho=0.60$

Good agreement: Correctly rejections

Good agreement: Correct detections

100% 96%

100% 96%

100% 96%

100% 96%

100% 96%

100% 96%

100% 96%

100% 96%

100% 96%

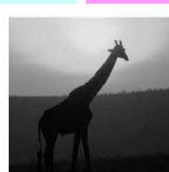
100% 96%

100% 96%

100% 96%

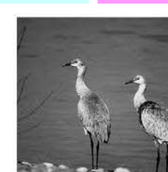
100% 96%

100% 96%

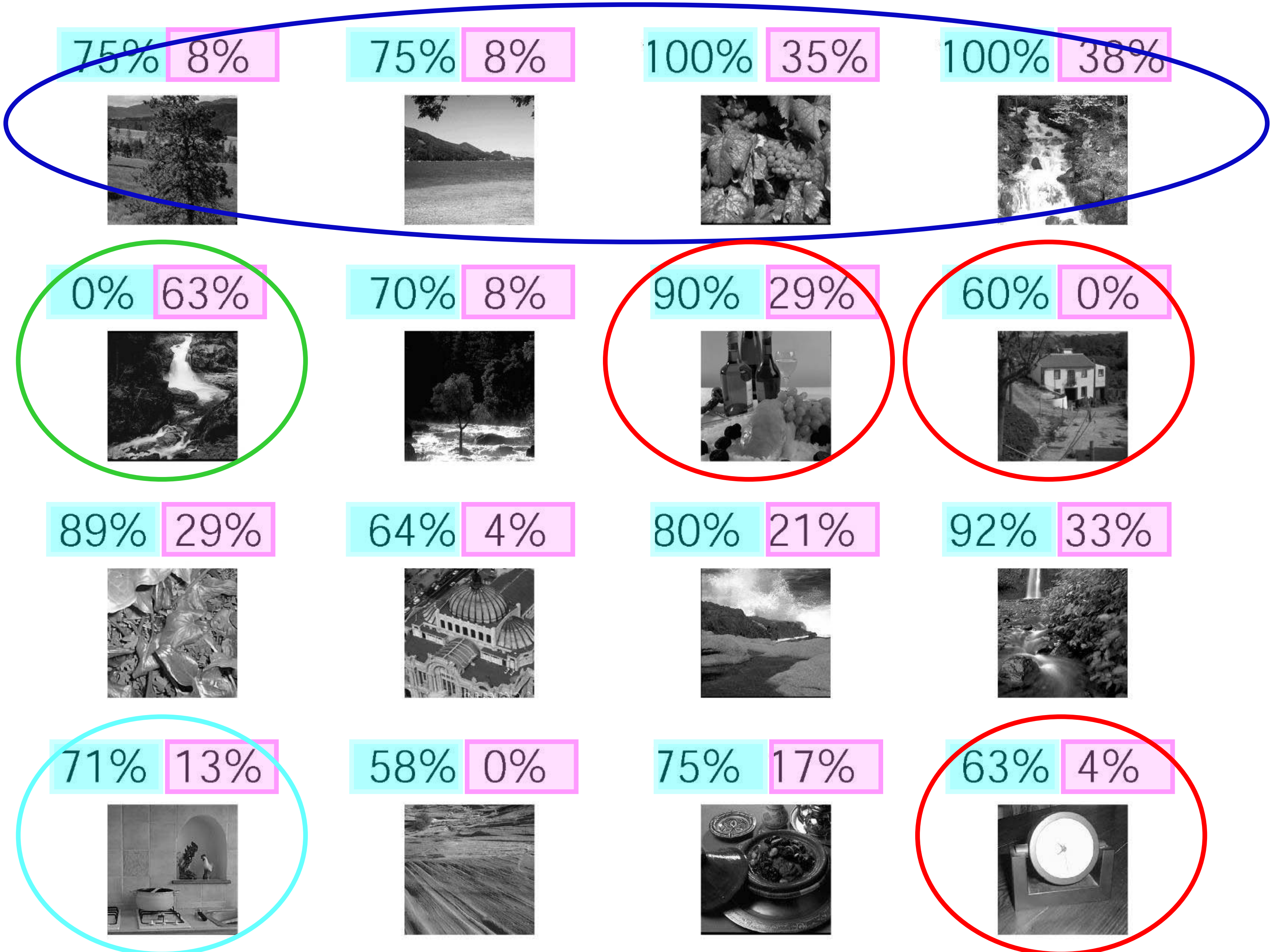


50% 54%

100% 96%



Disagreement



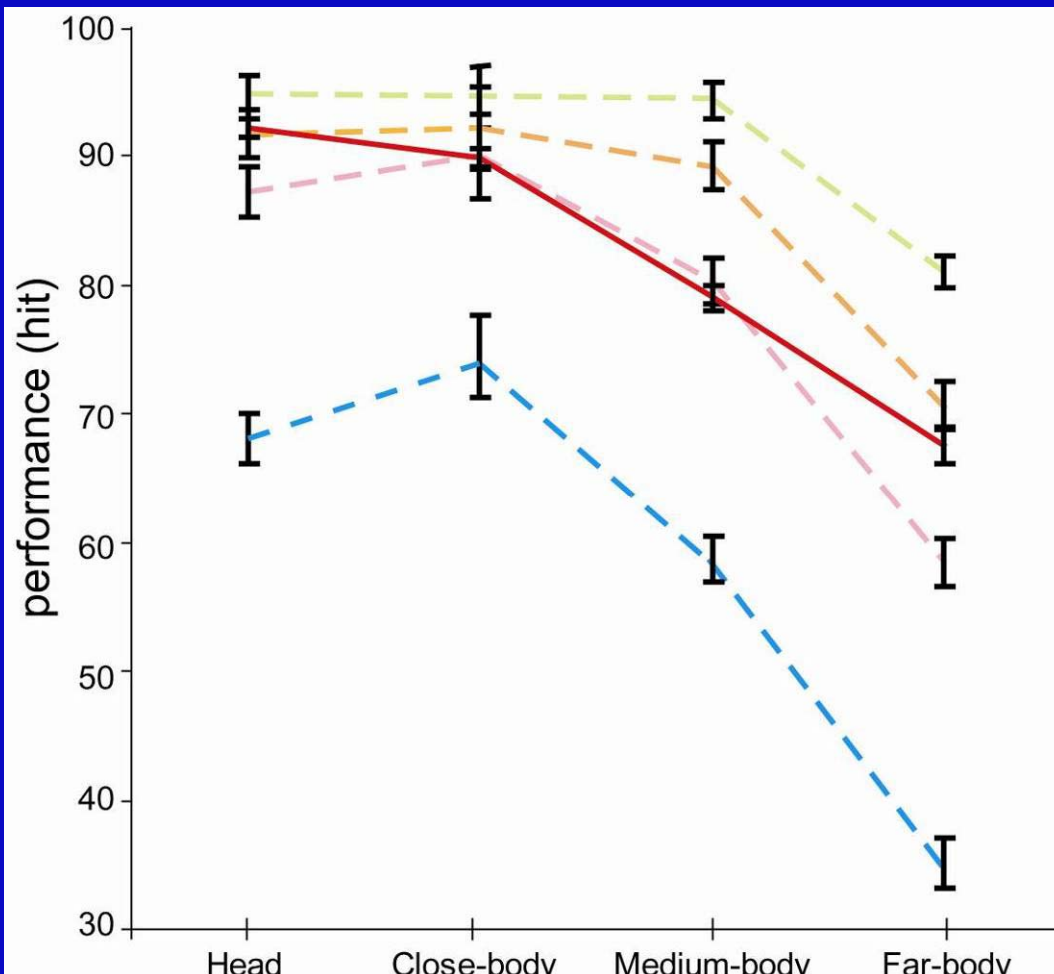
Disagreement

Discussion

- The model predicts human performance extremely well when the delay between the stimulus and the mask, i.e. the SOA is ~50 ms
- What happens for different SOAs?

Discussion

- Why should we except the model to account for human performance around 50 ms SOA?



no mask condition

80 ms SOA (ISI=60 ms)

model

50 ms SOA (ISI=30 ms)

20 ms SOA (ISI=0 ms)

(Serre, Oliva and Poggio, in prep)

Discussion

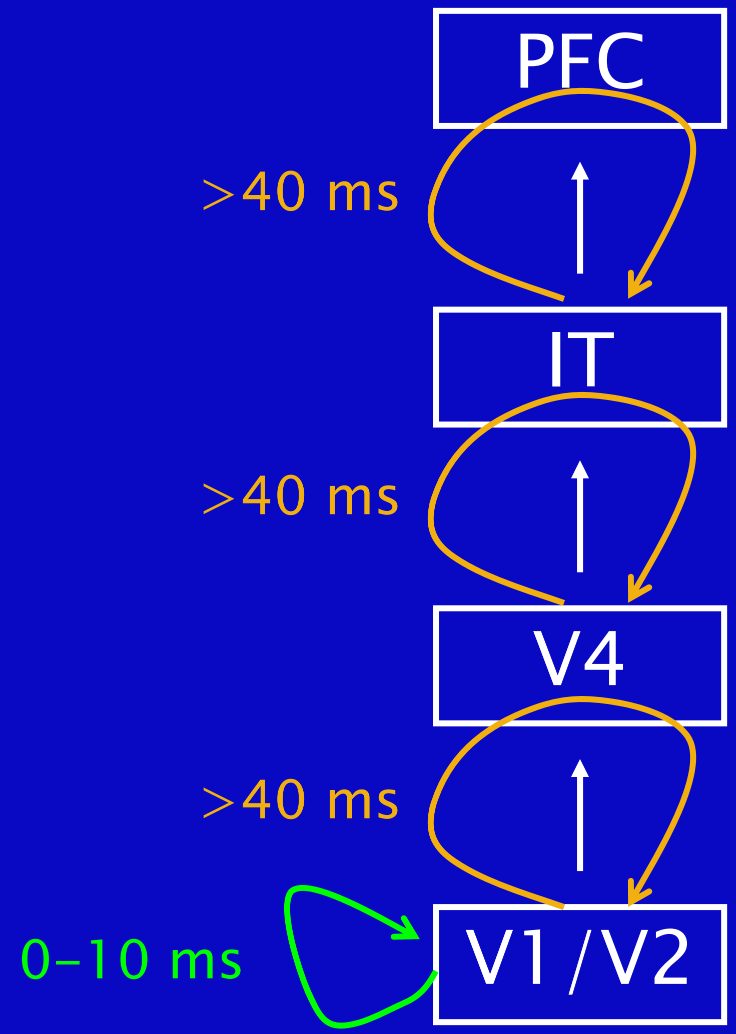
- What is so special with 50 ms SOA?
  - Possible answer:
    - ✓Nothing!!
    - ✓Mask disrupts signal integration at the neural level
    - ✓Model does not yet account for human level of performance

Discussion

- Alternative answer:
  - 50 ms is a very long time!
    - ✓Within 50 ms most of the information has already been transmitted from one stage to the next (Rolls et al, 1999; Vogels et al, 1995, Keysers et al, 2001)
    - ✓Reading out from IT (~10-20ms):
      - both object category and identity
      - largely translation and scale invariant (Hung, Kreiman, Poggio, DiCarlo, 2005)
- So what happened after the first 50 ms?

Speculation!!

- Our model is purely feedforward
  - Only local feedback loops
  - No feedback loops
- Feedback loops may already play a role for SOAs longer than 50 ms
- Discrepancy for longer SOAs may be due to the cortical back-projections



Timing estimates are for monkeys, based on (Thorpe & Fabre-Thorpe, 2001) and (Thorpe, Personal communication)

Summary

- I have described a model that is faithful to the anatomy and physiology of the ventral stream of visual cortex
- The model builds a dictionary of image features from V2 to IT which is compatible with the tuning of cortical neurons in several brain areas
- The model seems to be able to predict very well the level of performance of human observers in a rapid categorization task

Collaborators

- Aude Oliva
- Tomaso Poggio
- Other contributors
  - S. Bileschi
  - C. Cadieu
  - U. Knoblich
  - M. Kouh
  - G. Kreiman
  - M. Riesenhuber
  - L. Wolf