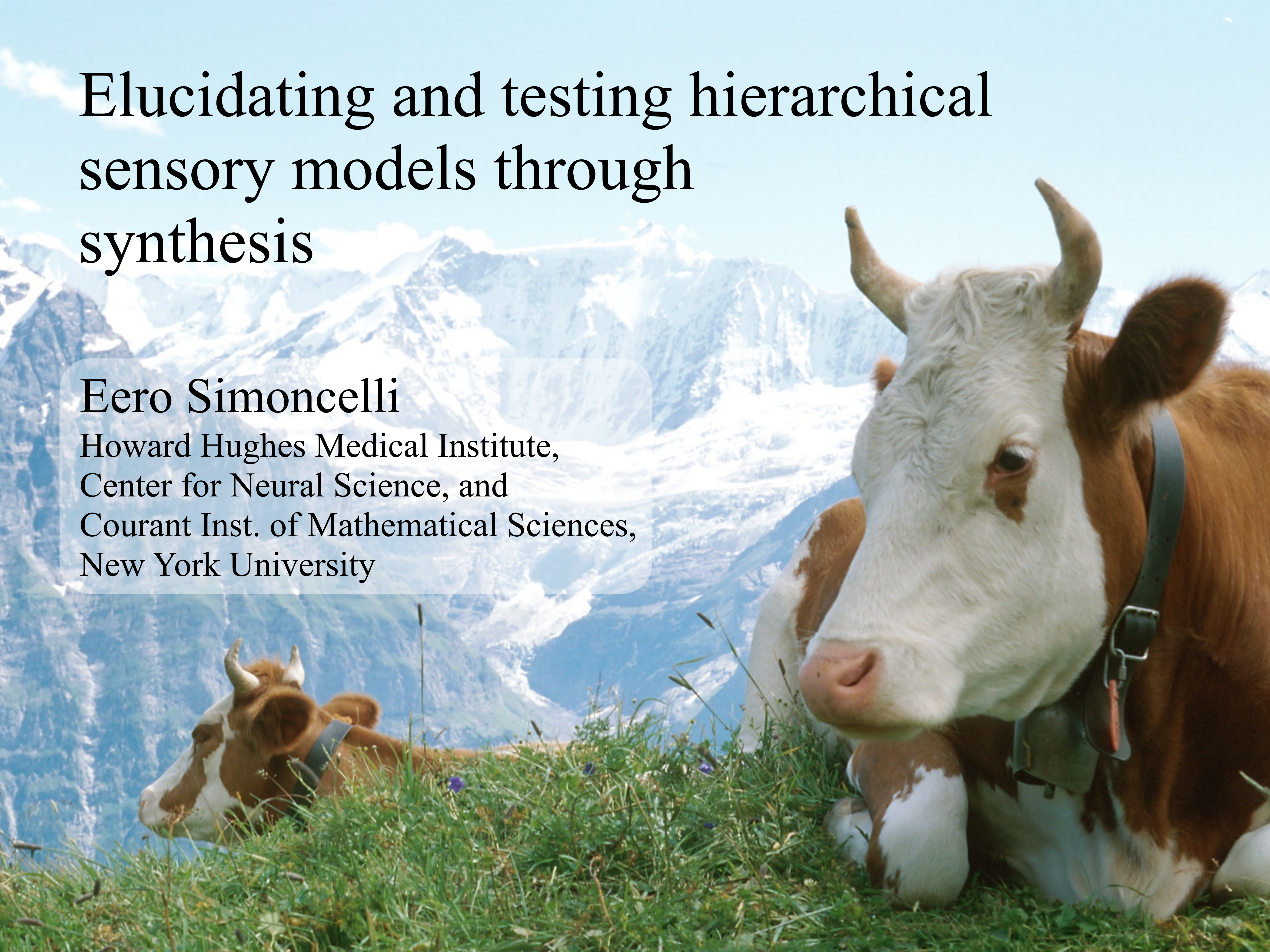
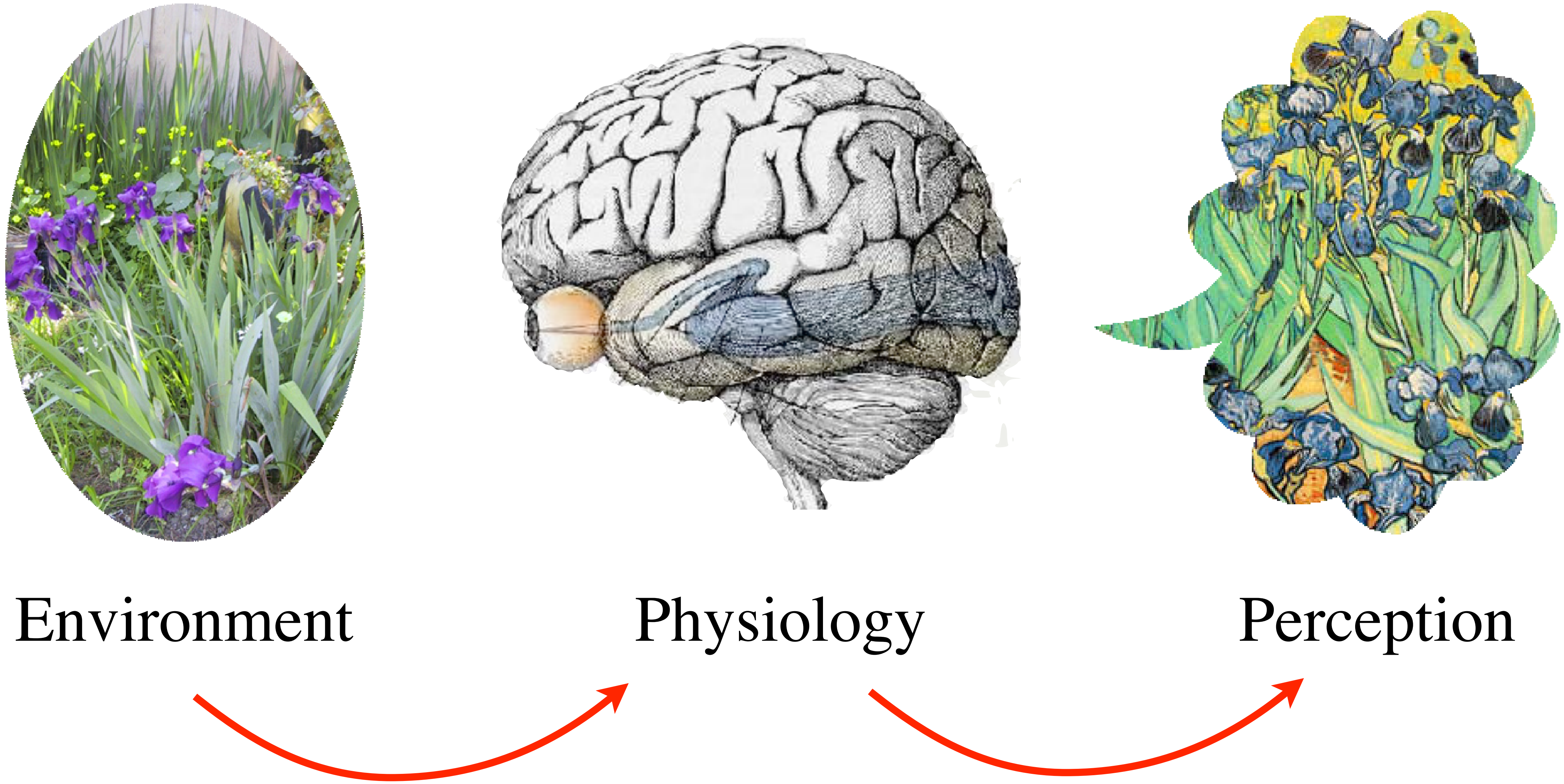
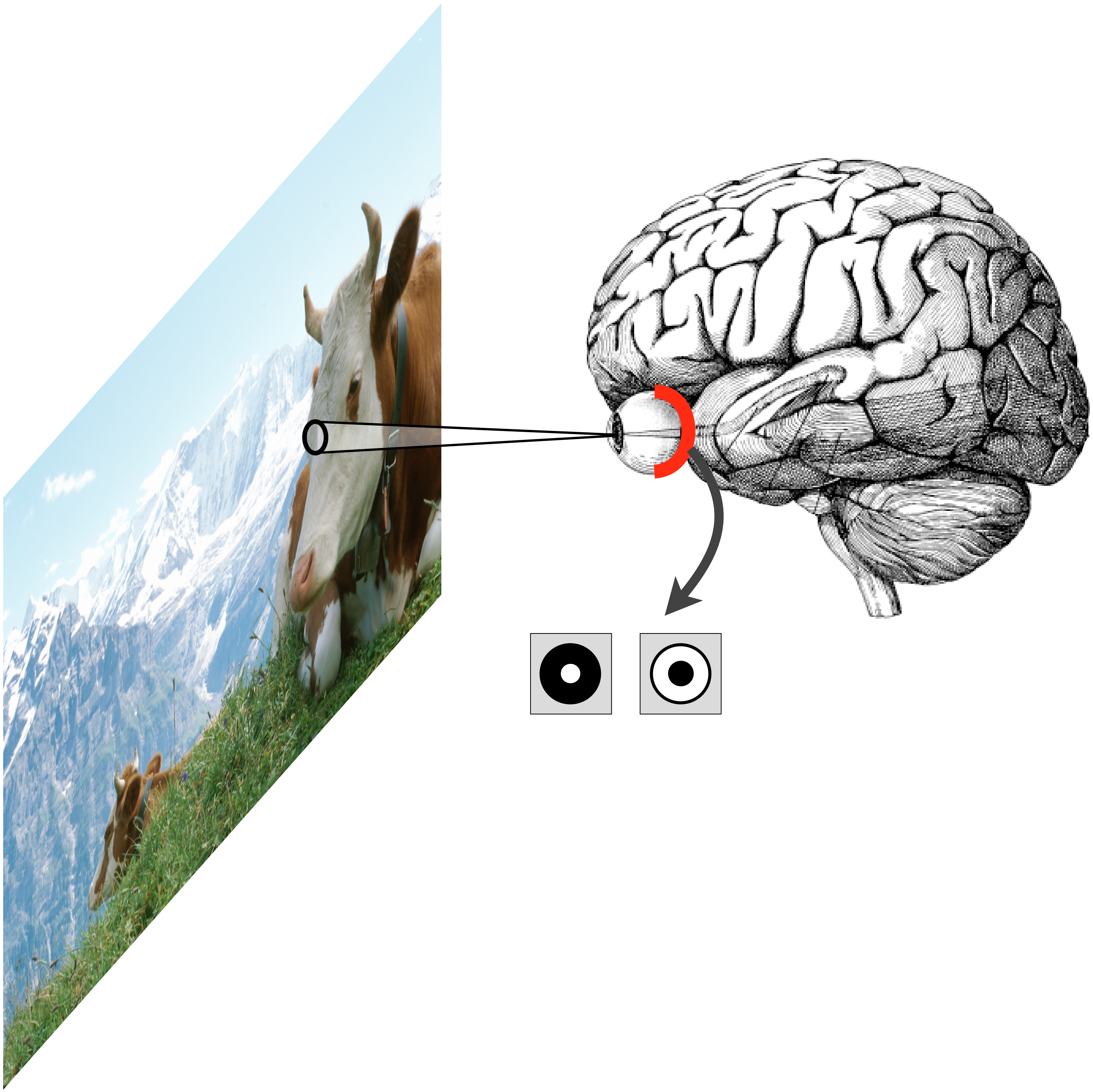
Elucidating and testing hierarchical sensory models through synthesis

Eero Simoncelli
Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Inst. of Mathematical Sciences,
New York University

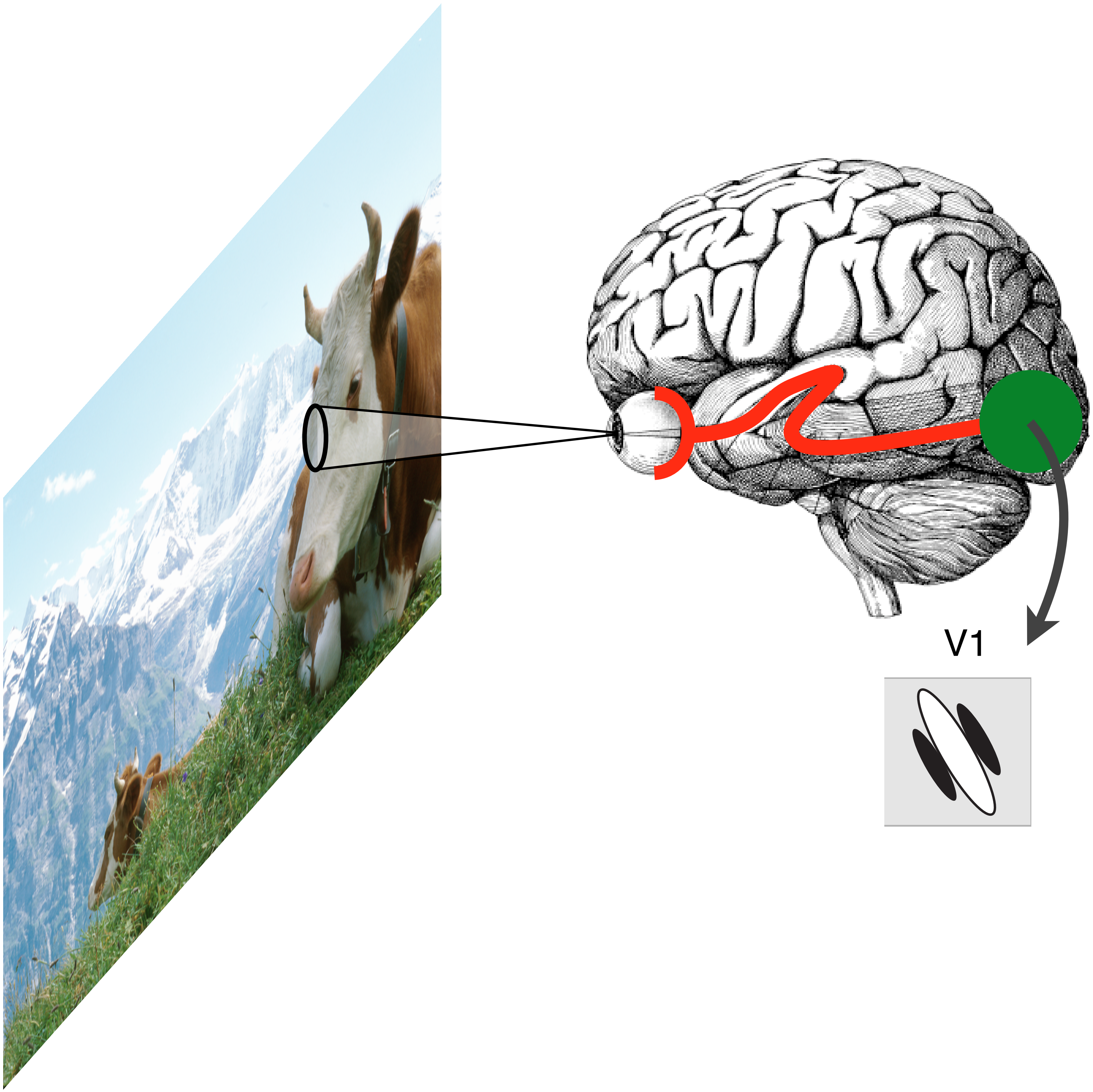


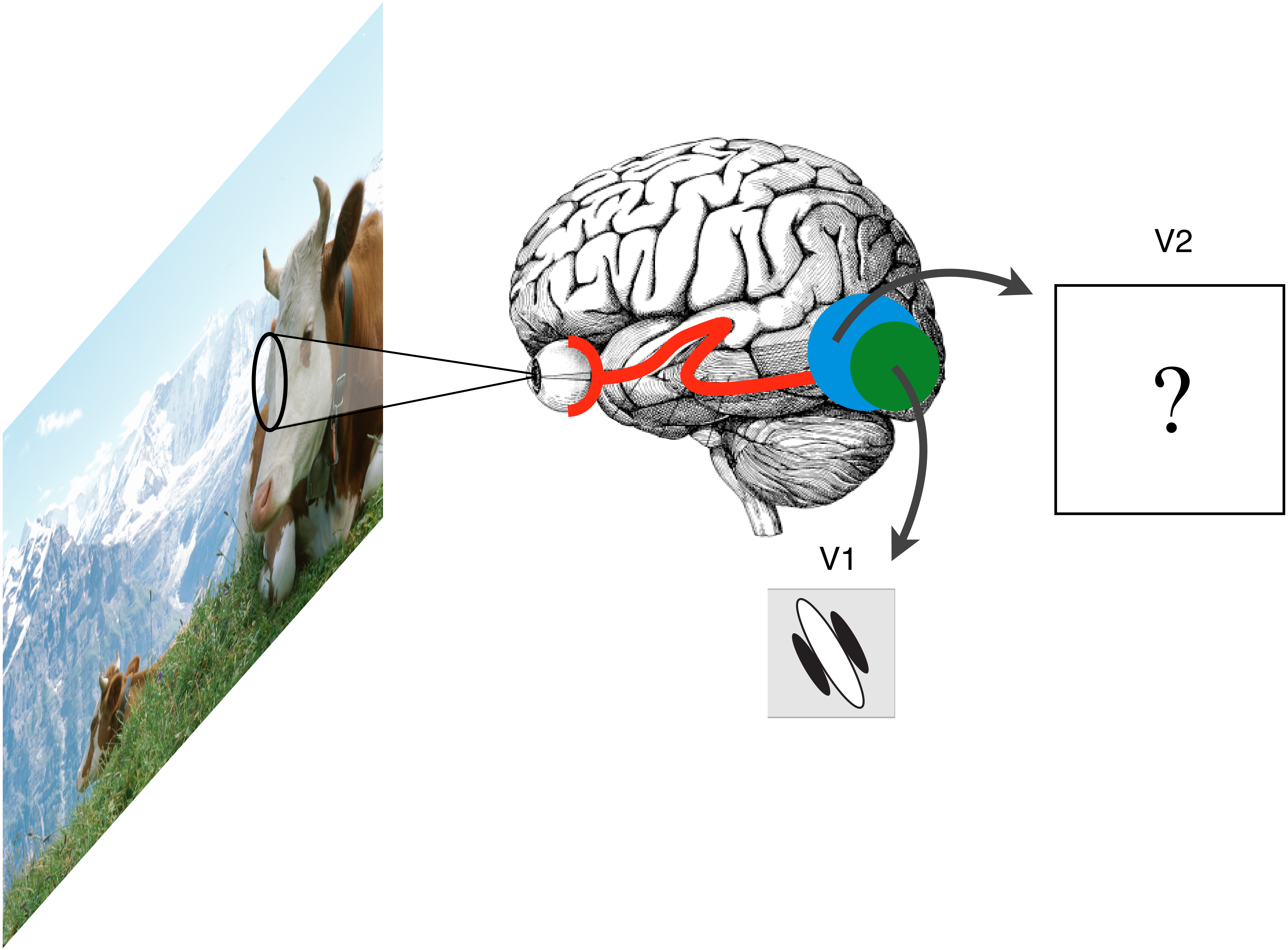


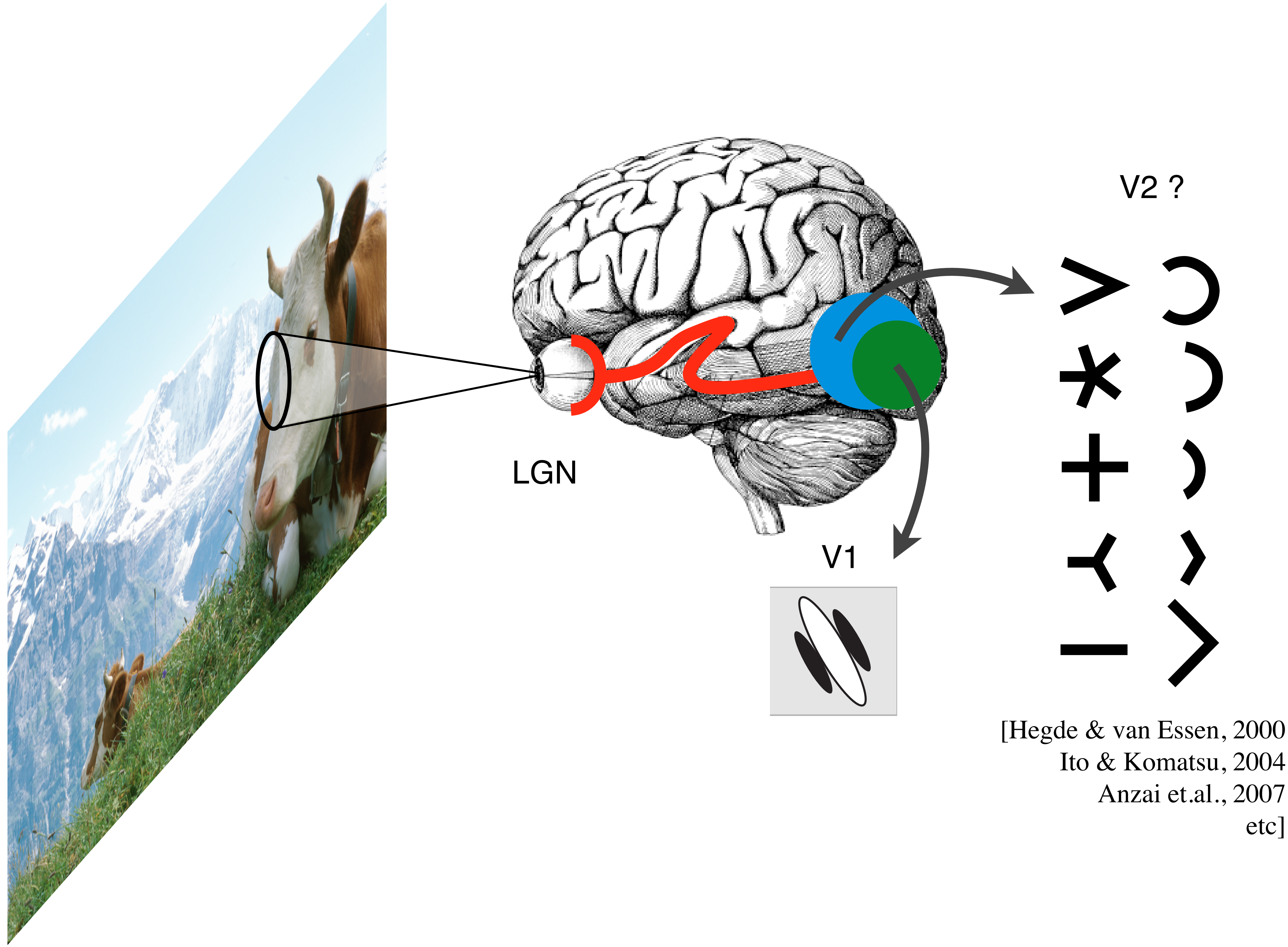
- How do populations of neurons extract/represent visual information?
- How is this matched to, or optimized for, our visual environment?
- How do these representations enable/limit perception?
- What new principles may be gleaned from these representations, and applied to engineered imaging or vision systems?



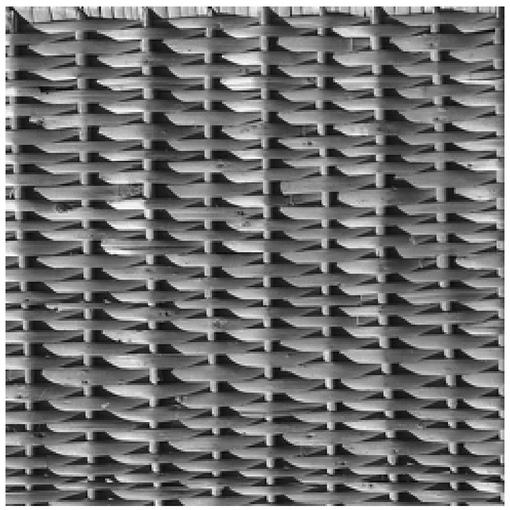
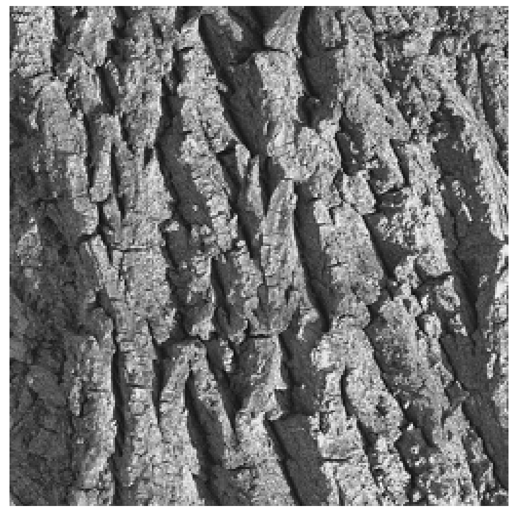
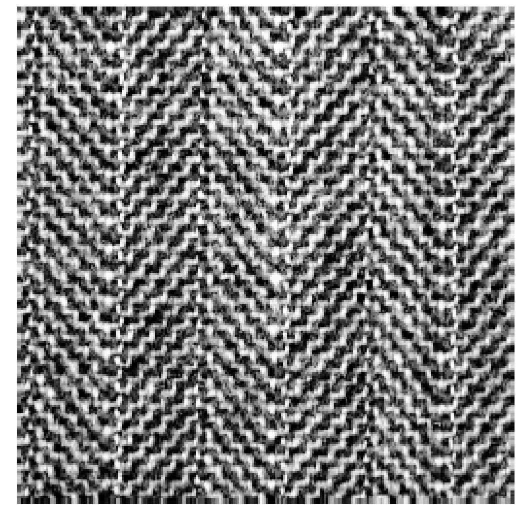
[brain figure: Hubel '95]


V1


V2
?
V1


LGN
V1
V2 ?
[Hegde & van Essen, 2000
Ito & Komatsu, 2004
Anzai et.al., 2007
etc]

Visual texture

Homogeneous, with repeated structures

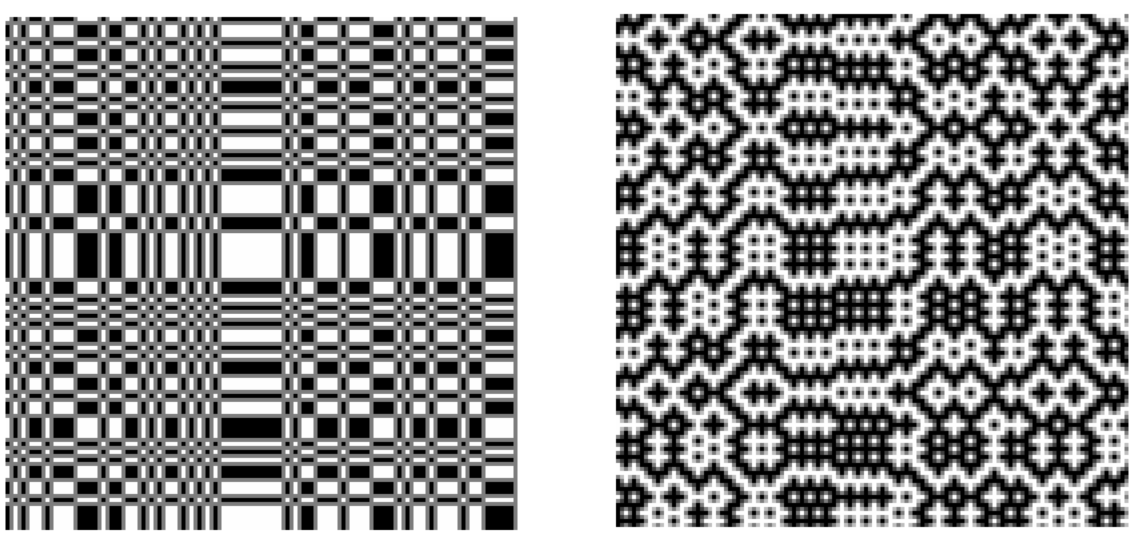
"Let us say that to the extent that visible objects are different and far apart, they are forms. To the extent that they are similar and congregated they are a texture. A man has form; a crowd has man-texture. A leaf has form; an arbor has leaf texture, and so on."

[Lettvin, 1976]

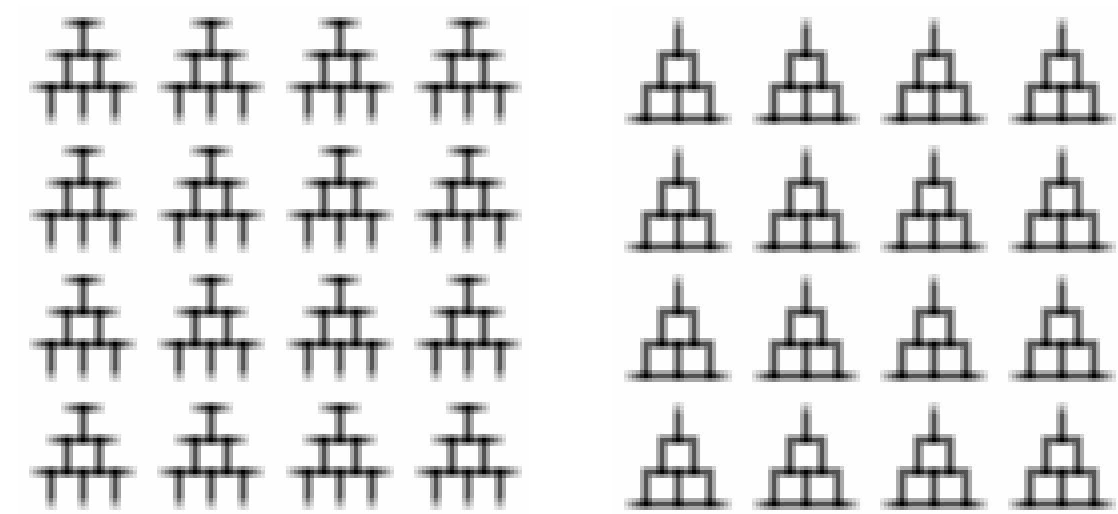
Julesz's conjecture (1962)

Two textures with identical *N*th-order pixel statistics will look the same (for some *N*).

Hand-constructed counter-examples (N=3):

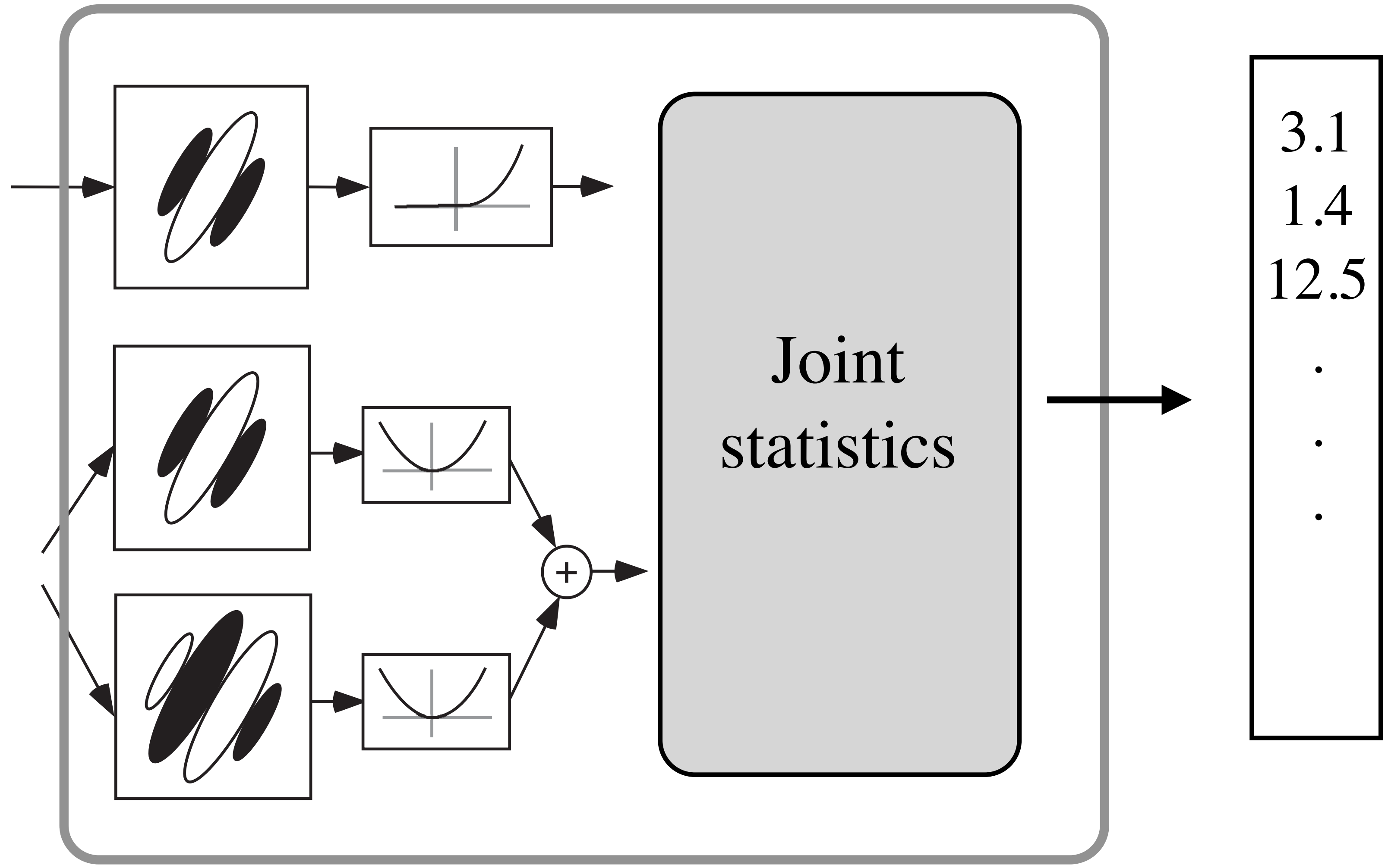


Julesz ‘78



Yellott ‘93

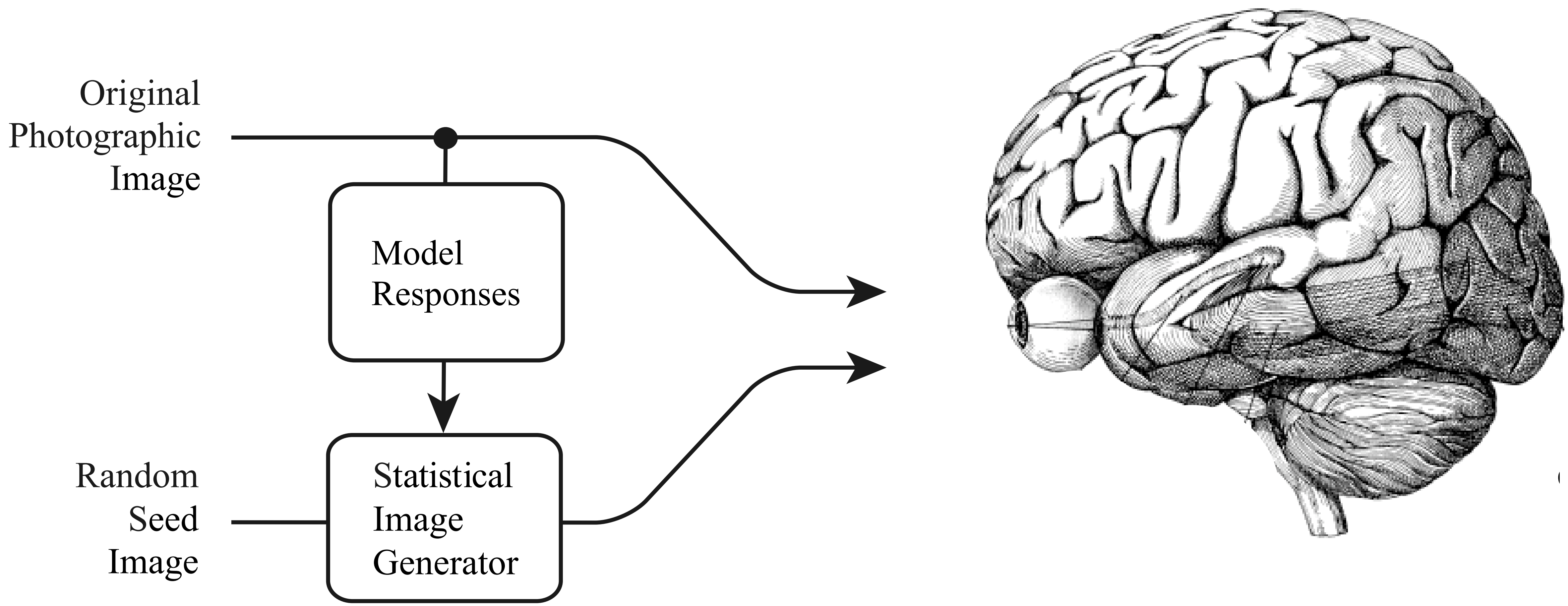
Physiologically-inspired Julesz-style texture model



Statistics: Correlations across position, orientation, scale.
... 710 measurements

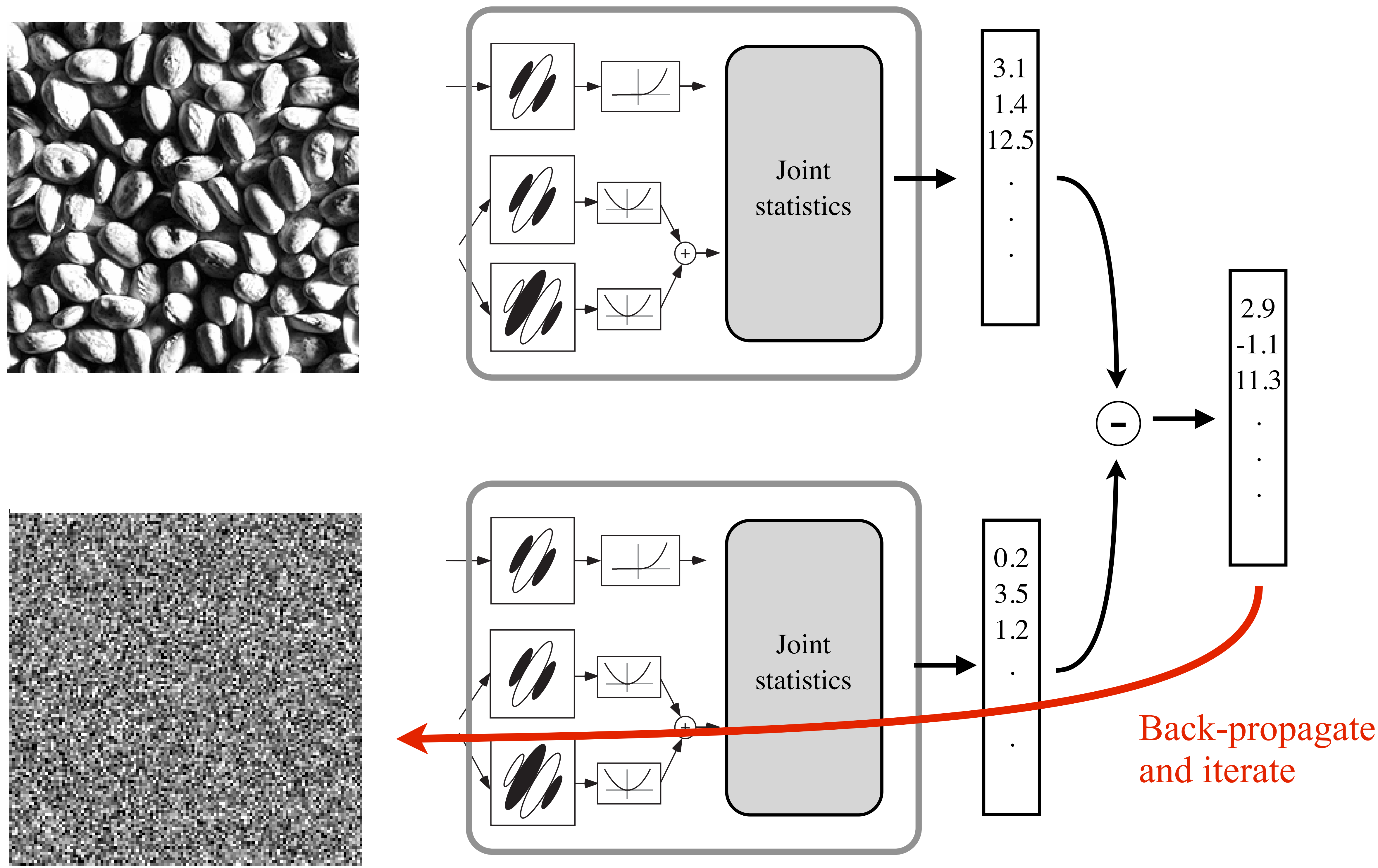
[Portilla & Simoncelli, *IJCV* 2000]

Synthesis-based experimental test

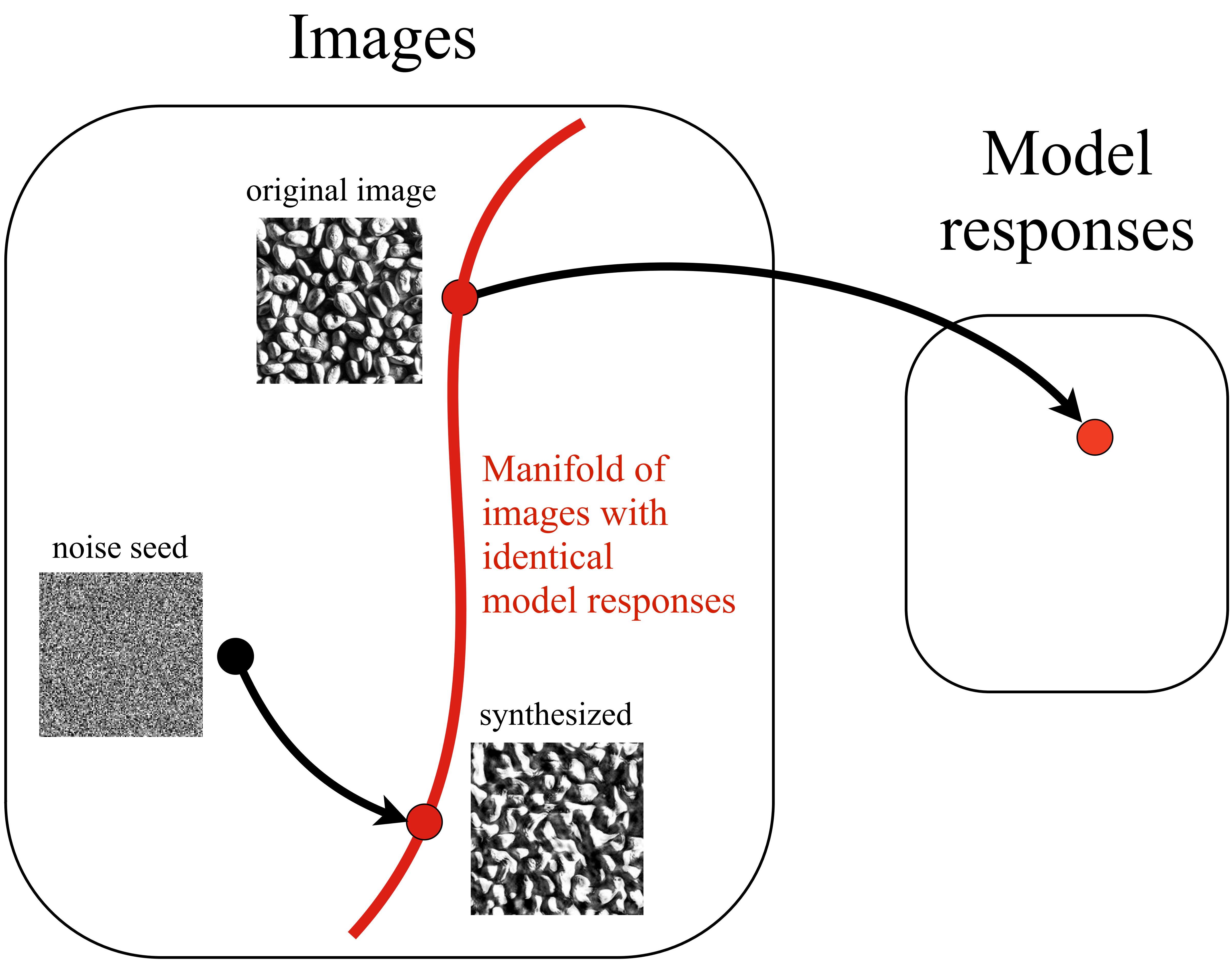


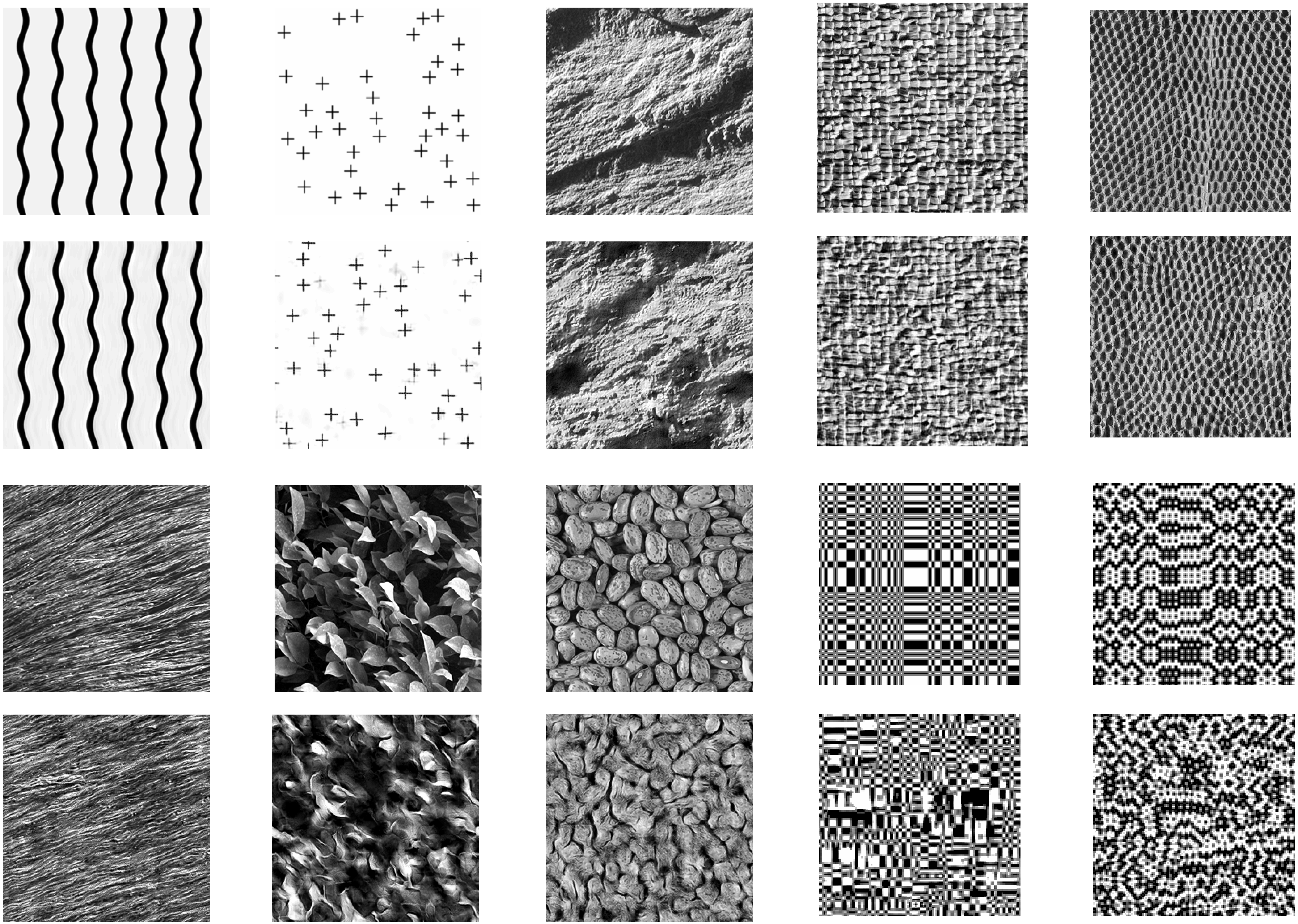
If model captures the same properties as the brain, images with identical model responses should appear identical to a human.

Texture synthesis



[Portilla & Simoncelli, *IJCV* 2000]

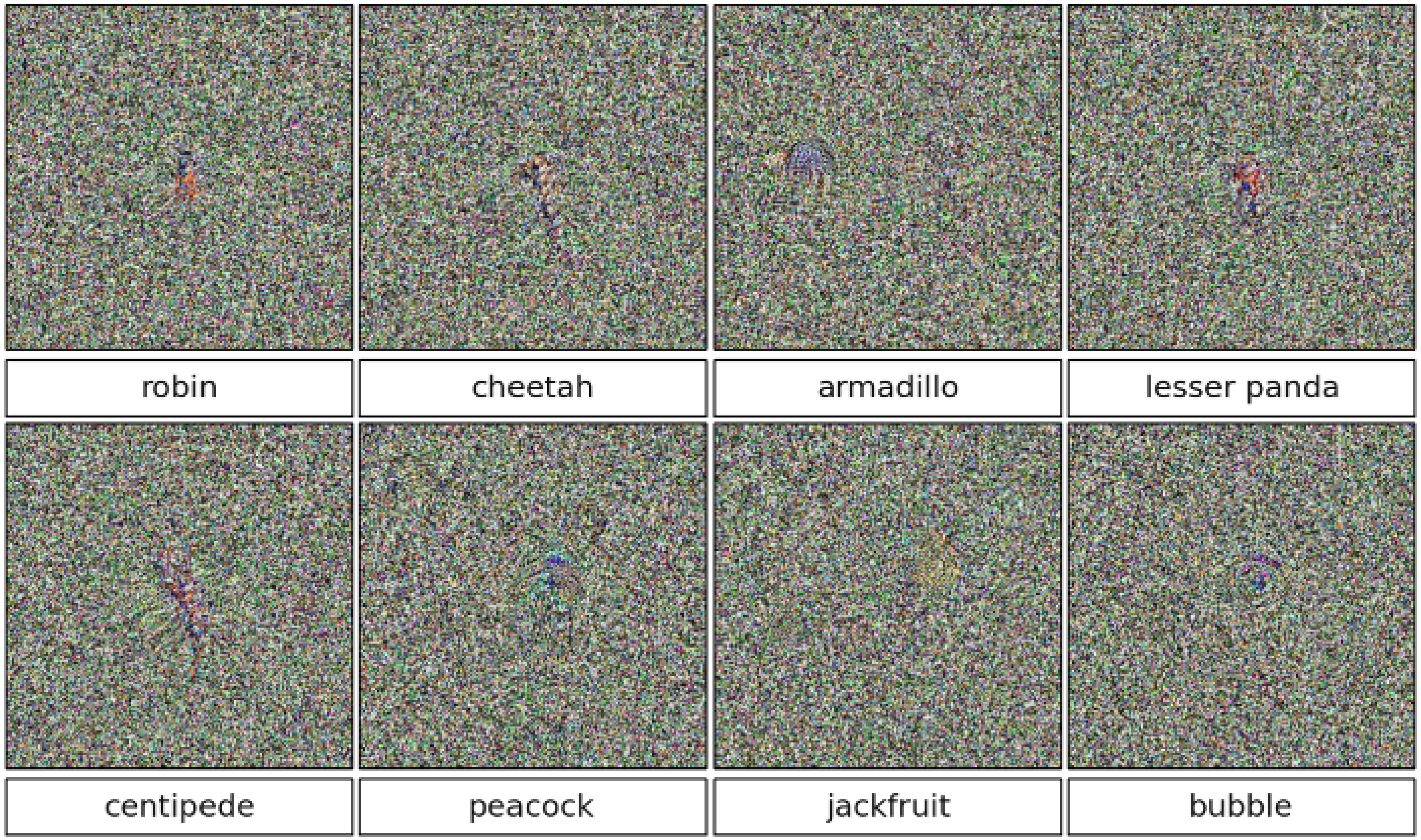

Images
Model
responses
original image
noise seed
synthesized
Manifold of
images with
identical
model responses



(more at www.cns.nyu.edu/~lcv/texture/)

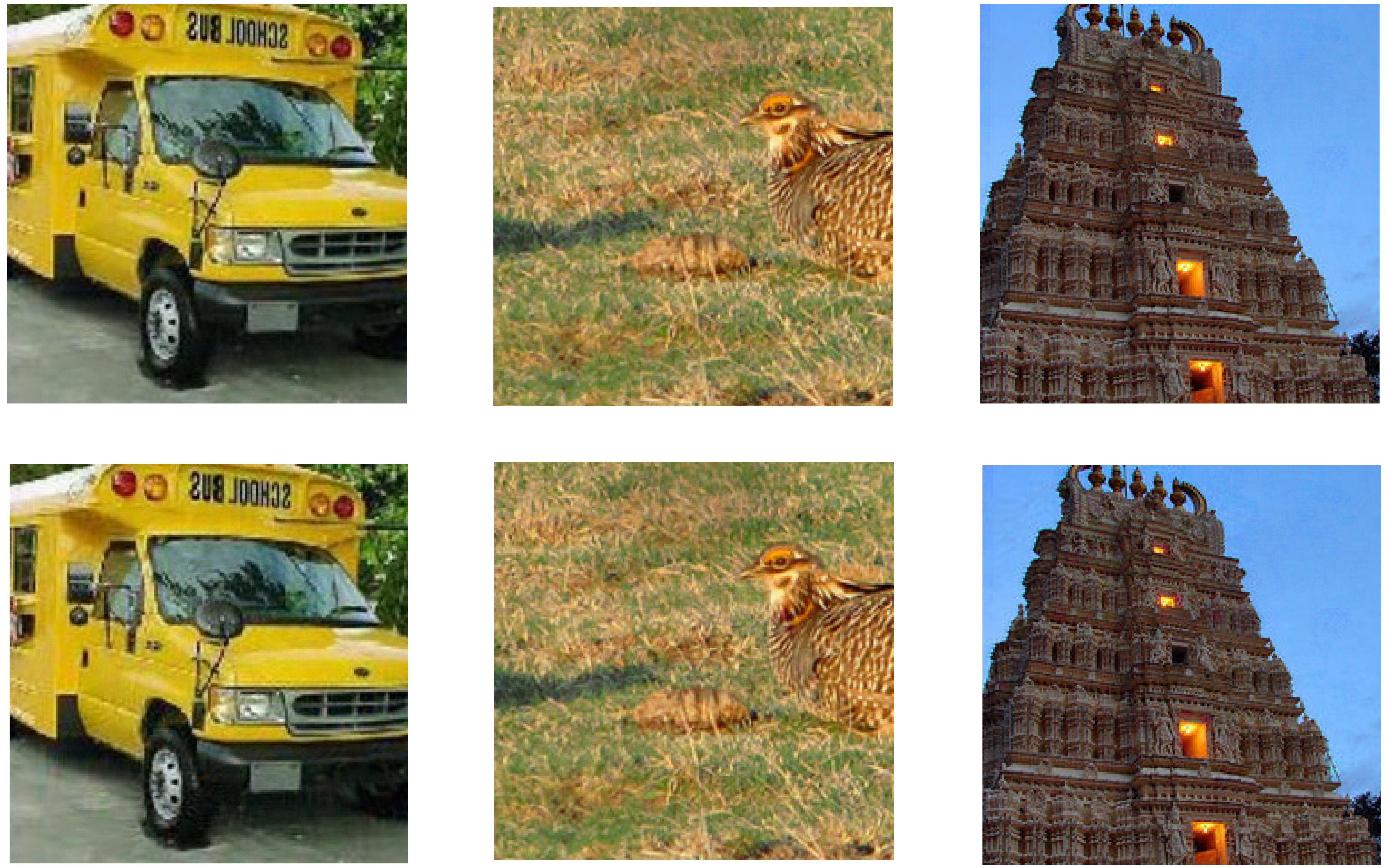
[Portilla & Simoncelli, *IJCV* 2000]

Note: analogous synthesis from deep ConvNets produces "fooling images"



Deep Neural Networks are Easily Fooled...
Nguyen, Yosinski, Clune *CVPR* 2015

Also: Adjusting an initial image to attain a deep convNet target category leads to “adversarial examples”



Initial image

Synthesized

Target category

“ostrich” “ostrich” “ostrich”

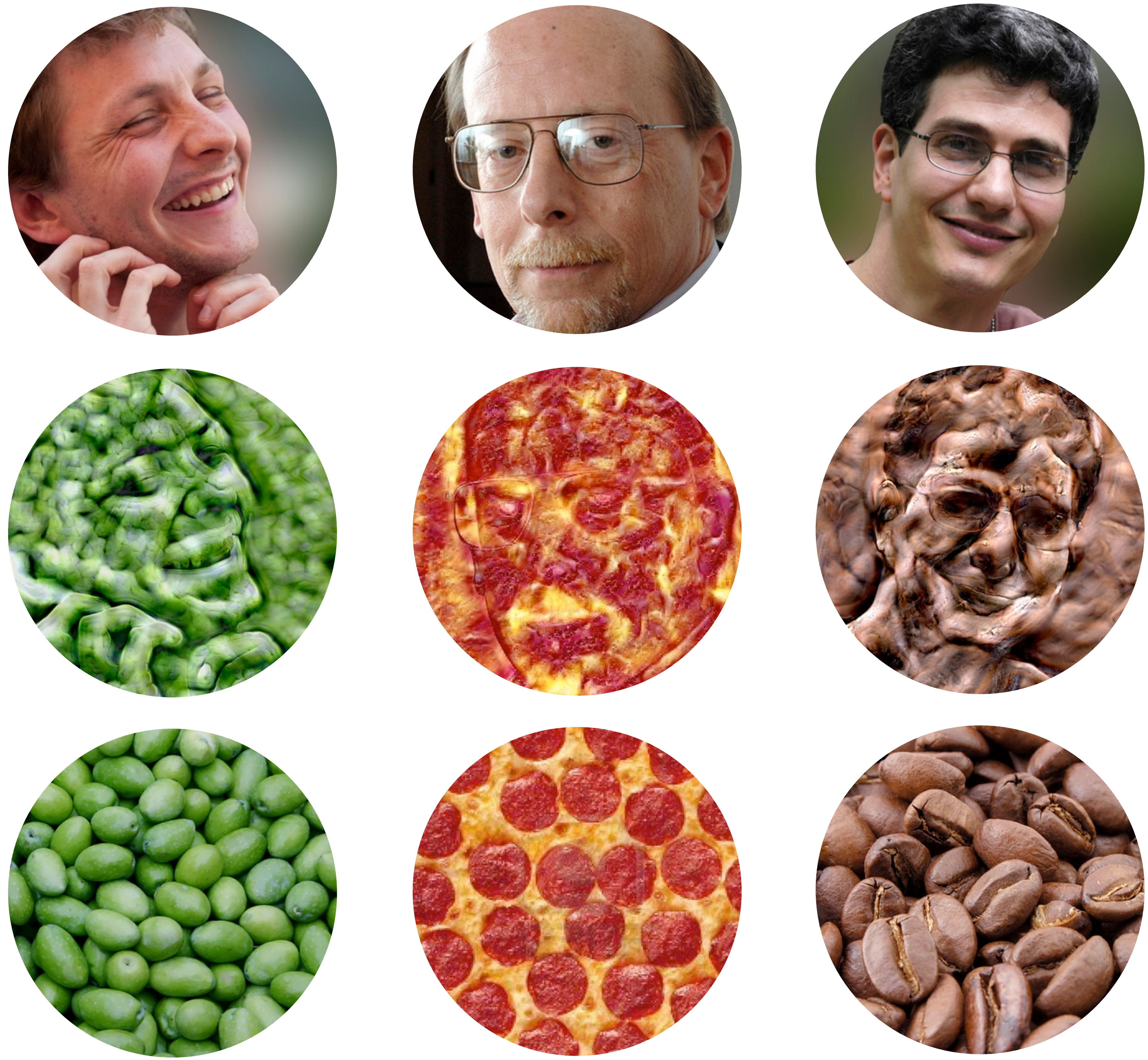
Intriguing properties of neural networks
Szegedy et. al. *arXiv* 2014

... but not for this texture model

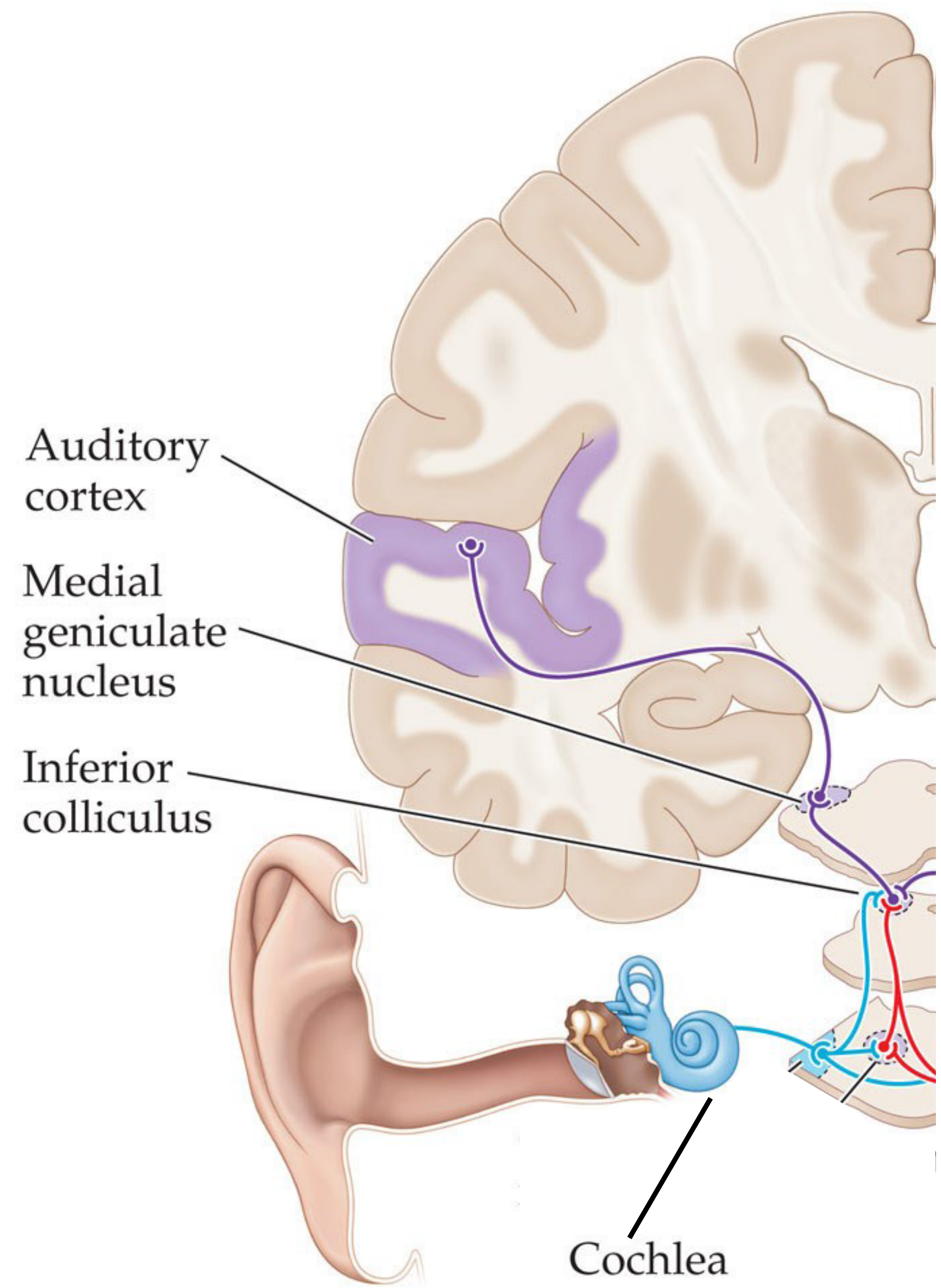
Initial image

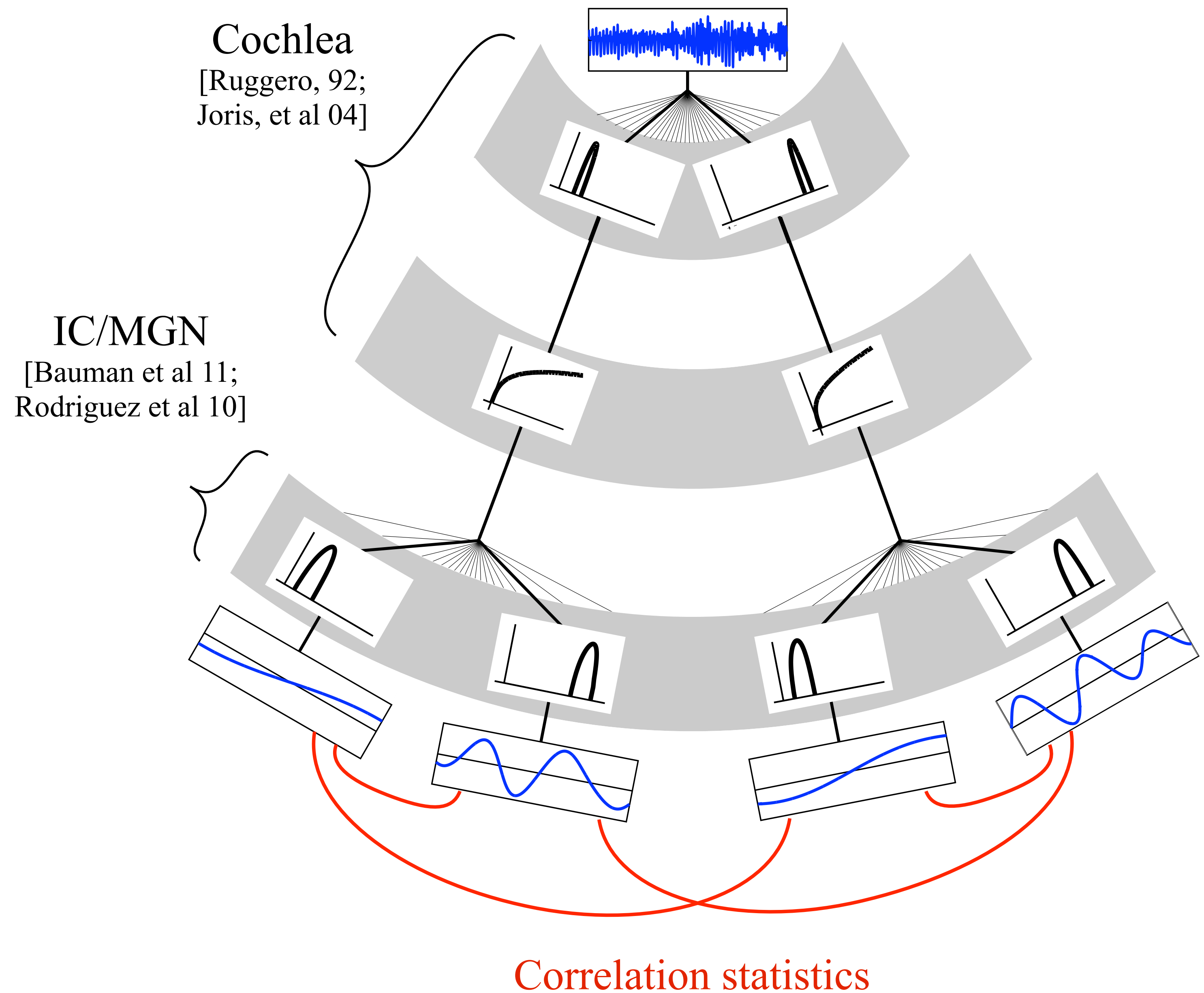
Synthesized

Target texture



Analogous auditory texture model



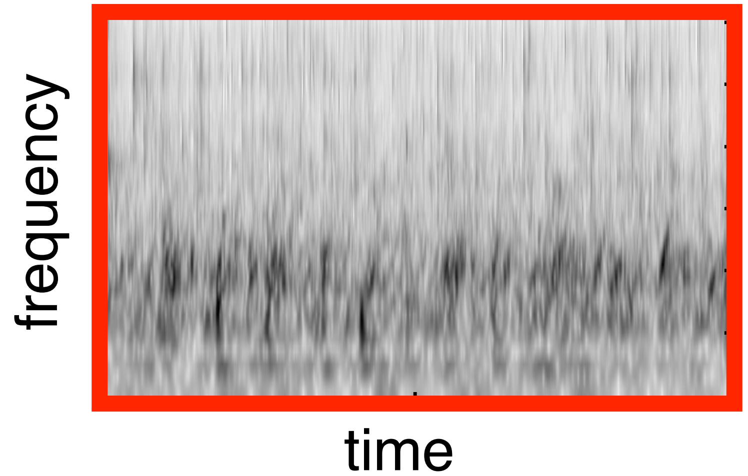


[figure: Sensation & Perception, 2015]

[McDermott & Simoncelli, *Neuron* 2011]

Auditory texture examples

original

Bubbling water: 

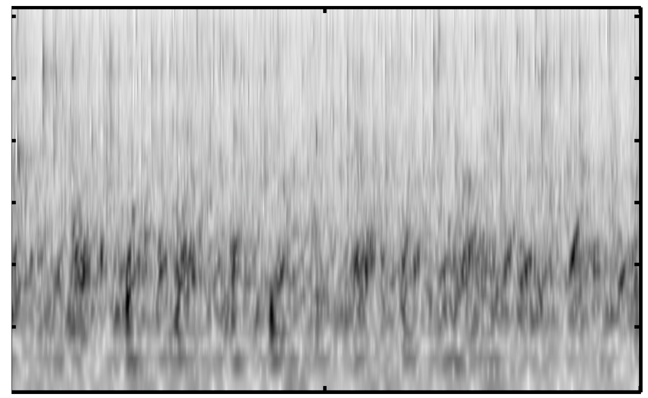
[McDermott & Simoncelli, *Neuron* 2011]

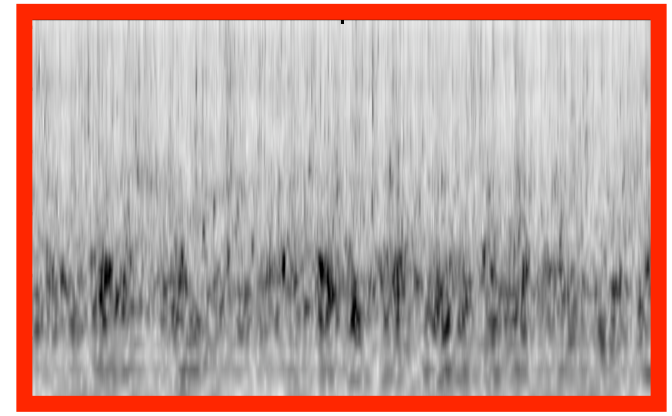
Auditory texture examples

original

synthesized

Bubbling water:





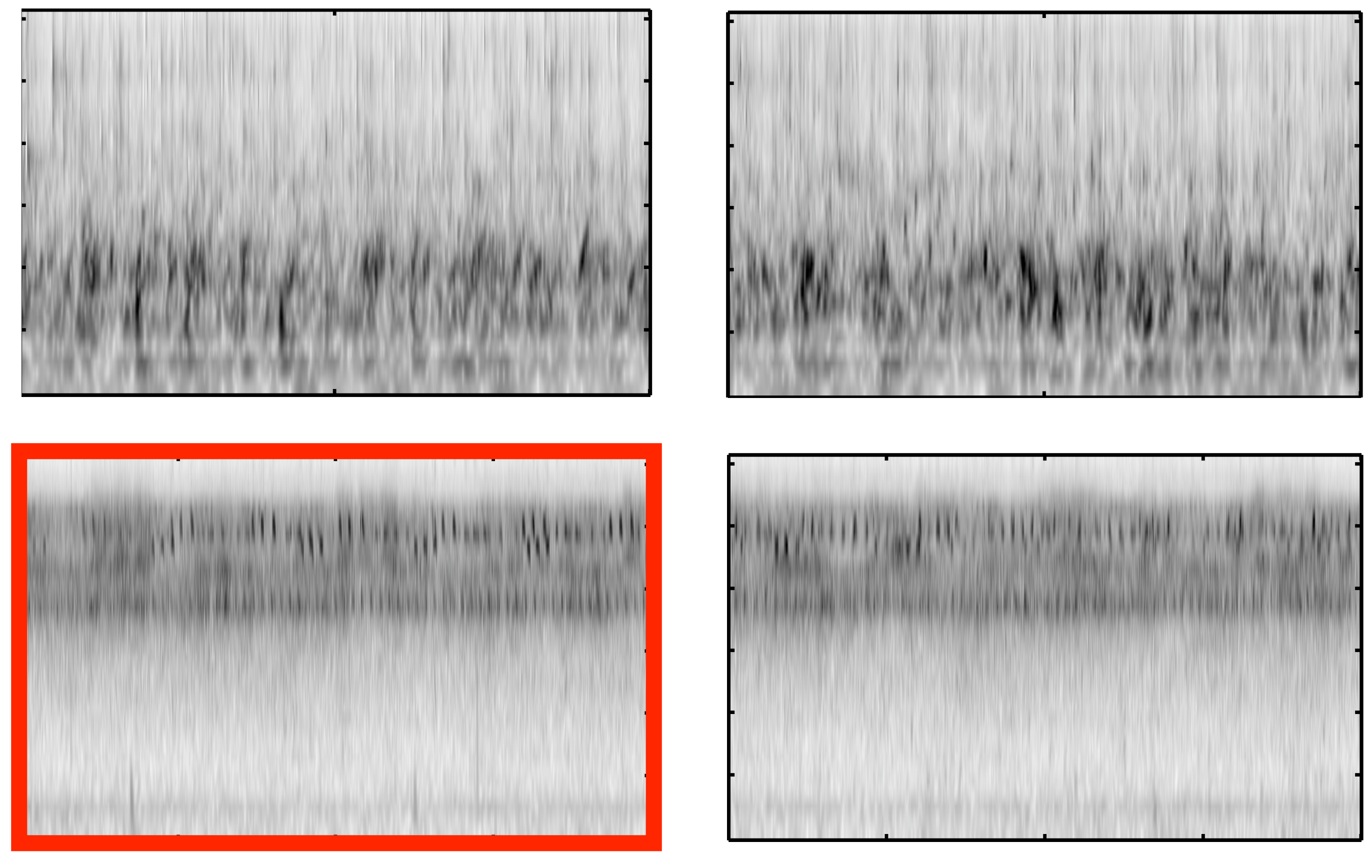
[McDermott & Simoncelli, *Neuron* 2011]

Auditory texture examples

original | synthesized

Bubbling water:

Insects:



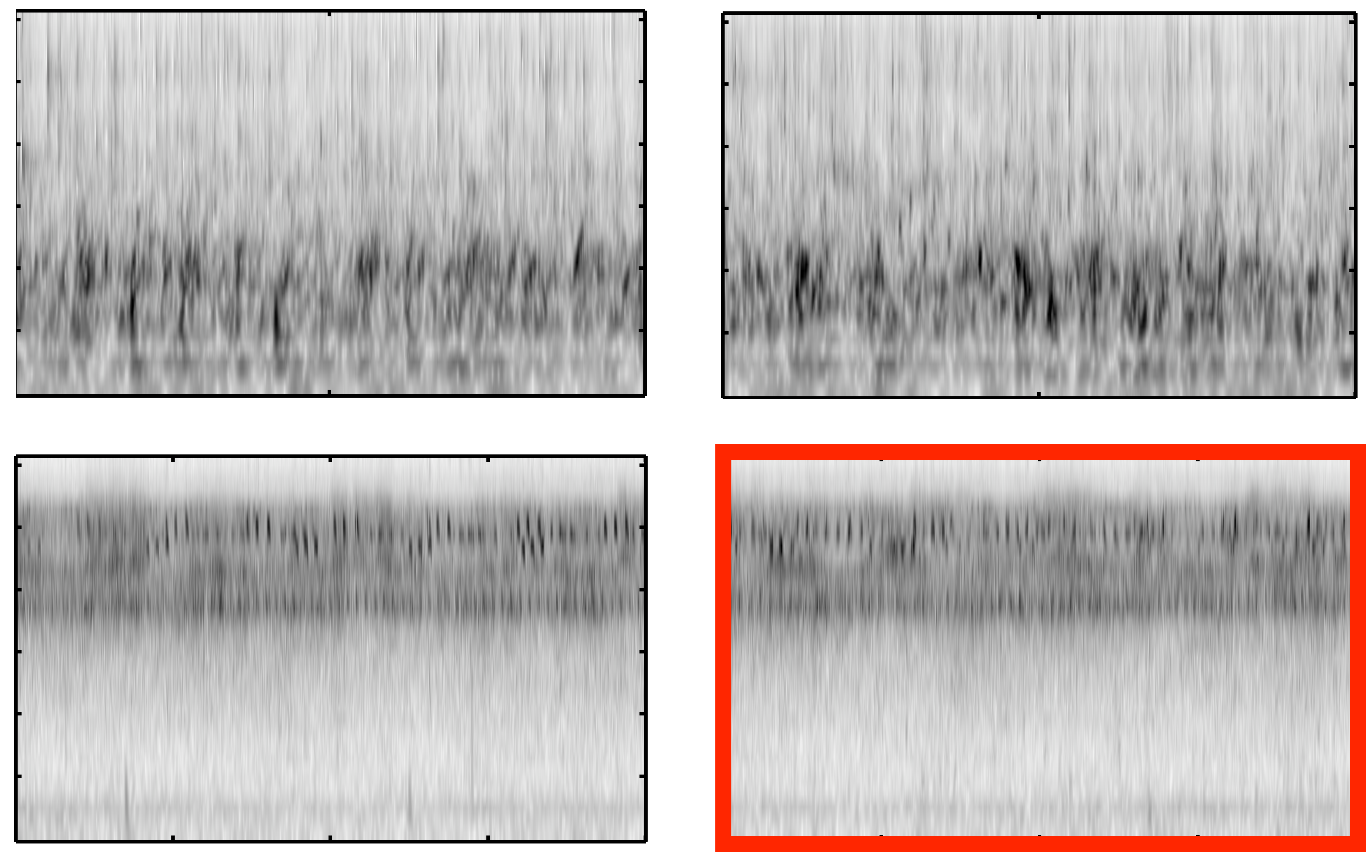
[McDermott & Simoncelli, *Neuron* 2011]

Auditory texture examples

original synthesized

Bubbling water:

Insects:



[McDermott & Simoncelli, *Neuron* 2011]

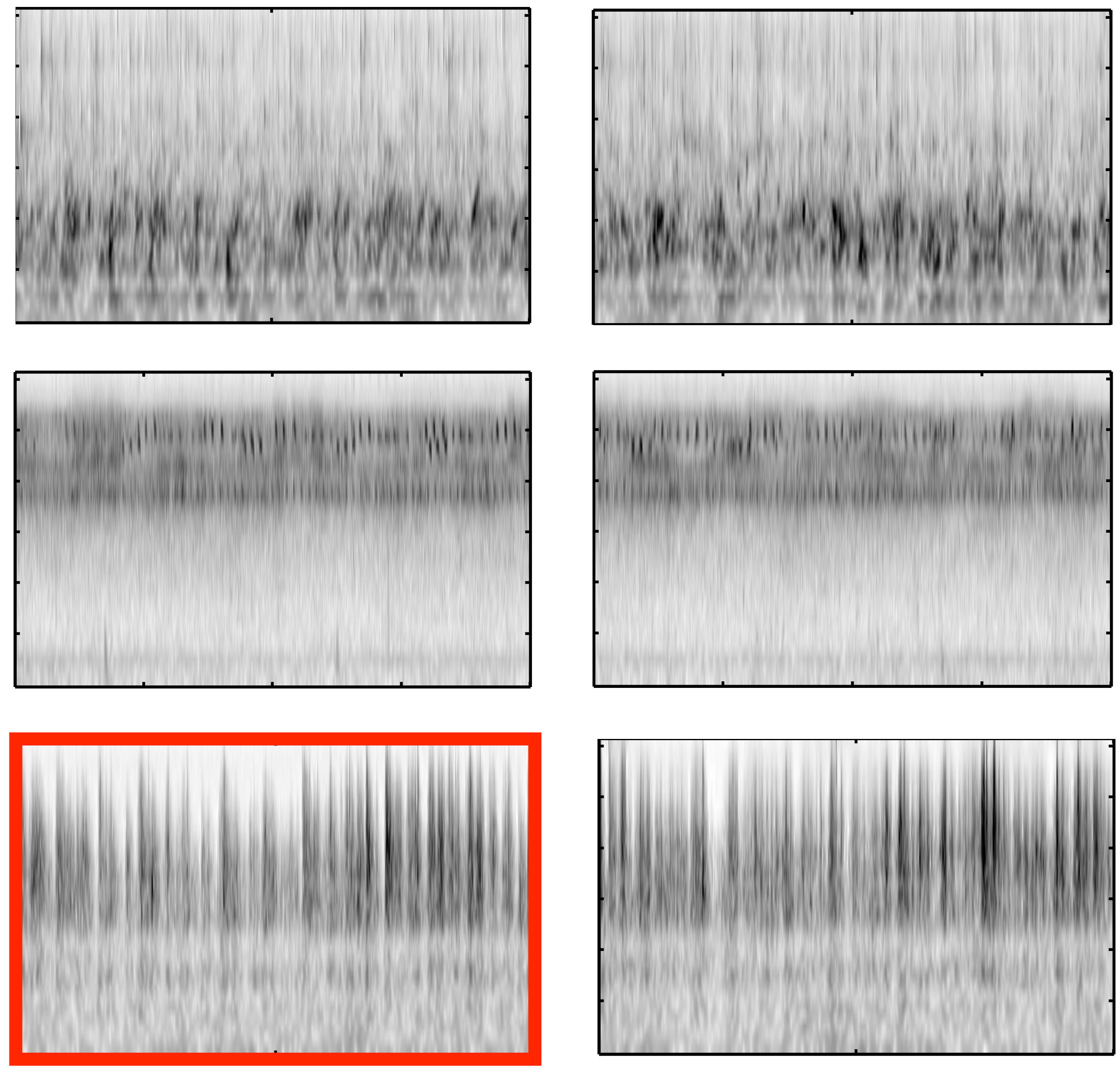
Auditory texture examples

original synthesized

Bubbling water:

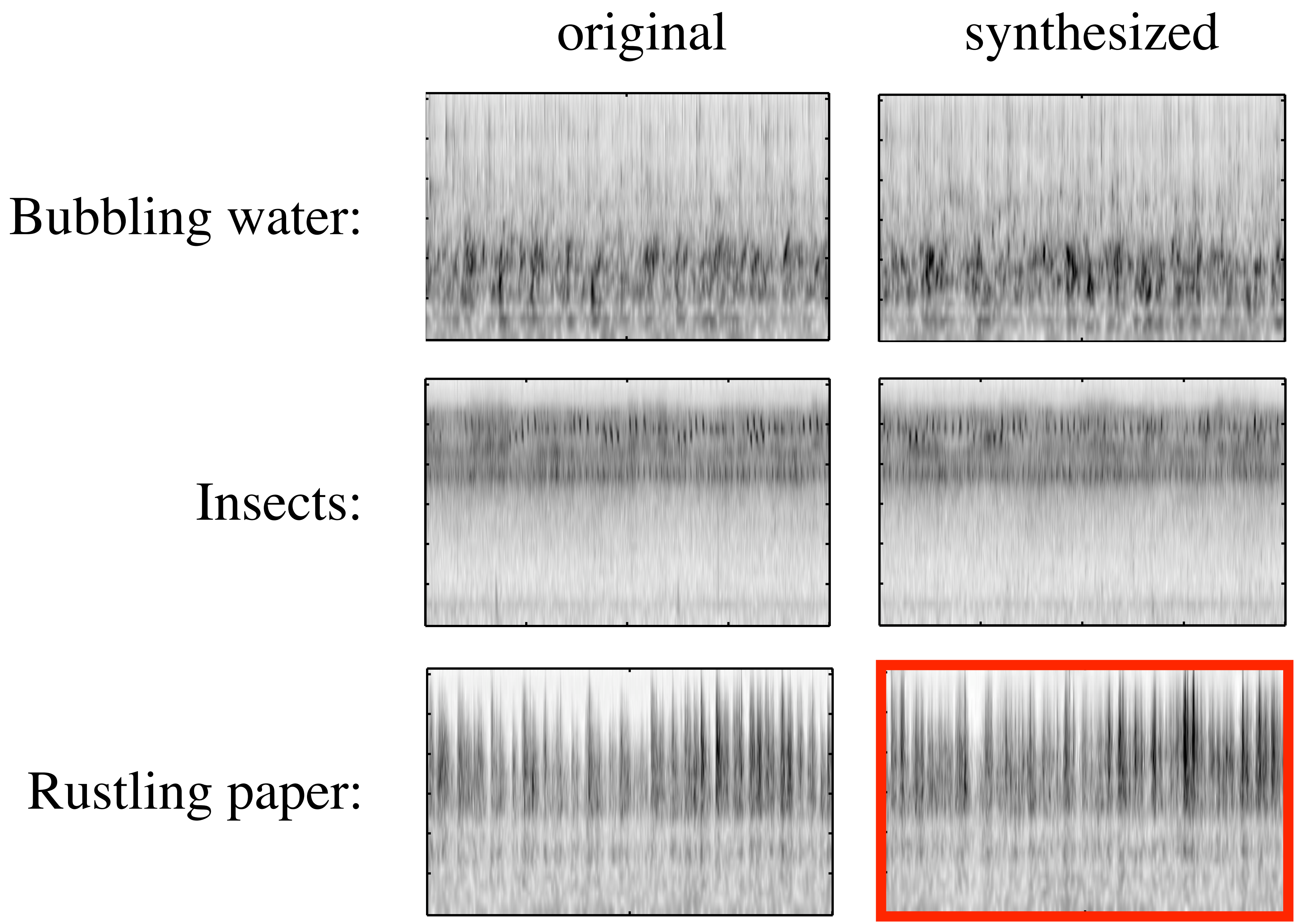
Insects:

Rustling paper:



[McDermott & Simoncelli, *Neuron* 2011]

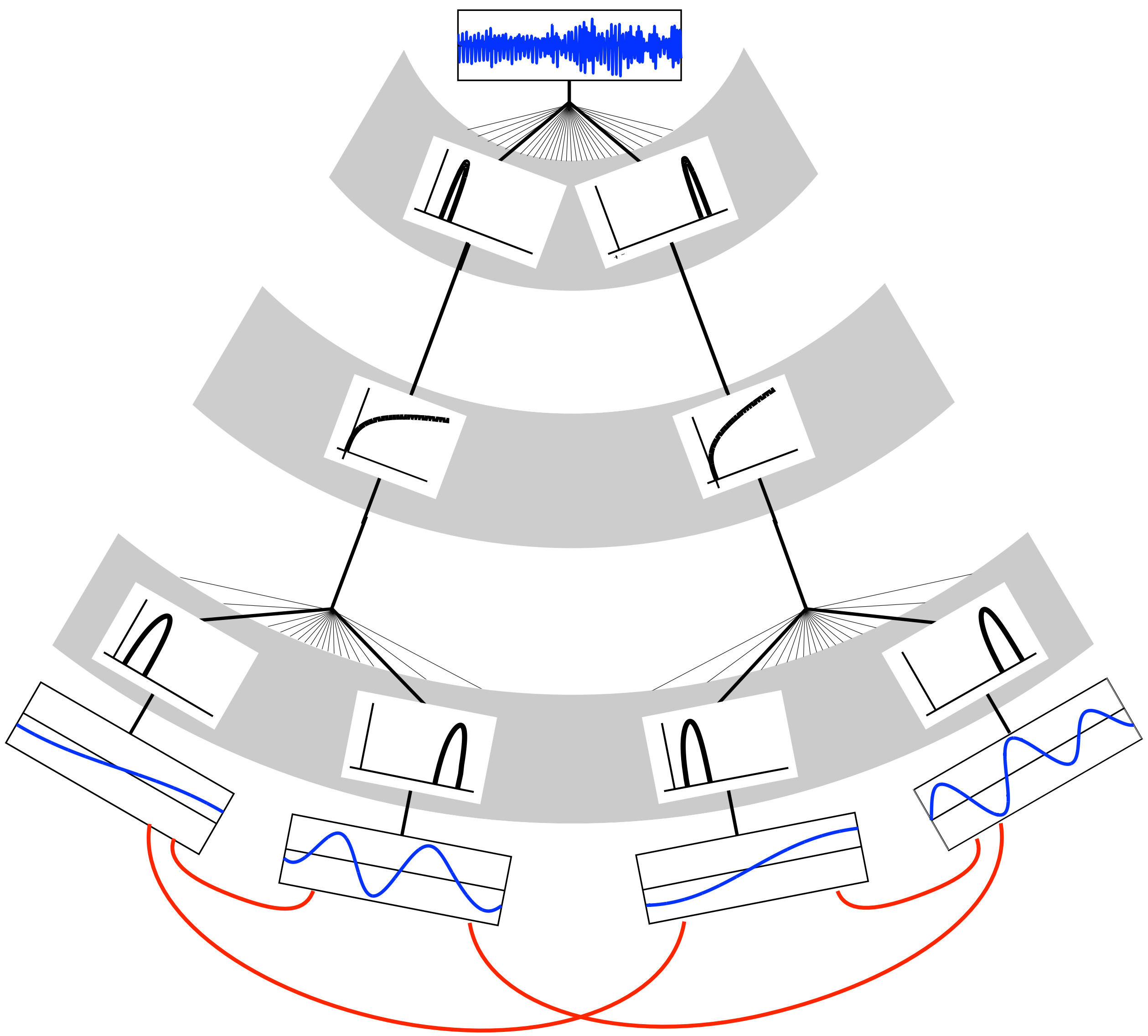
Auditory texture examples



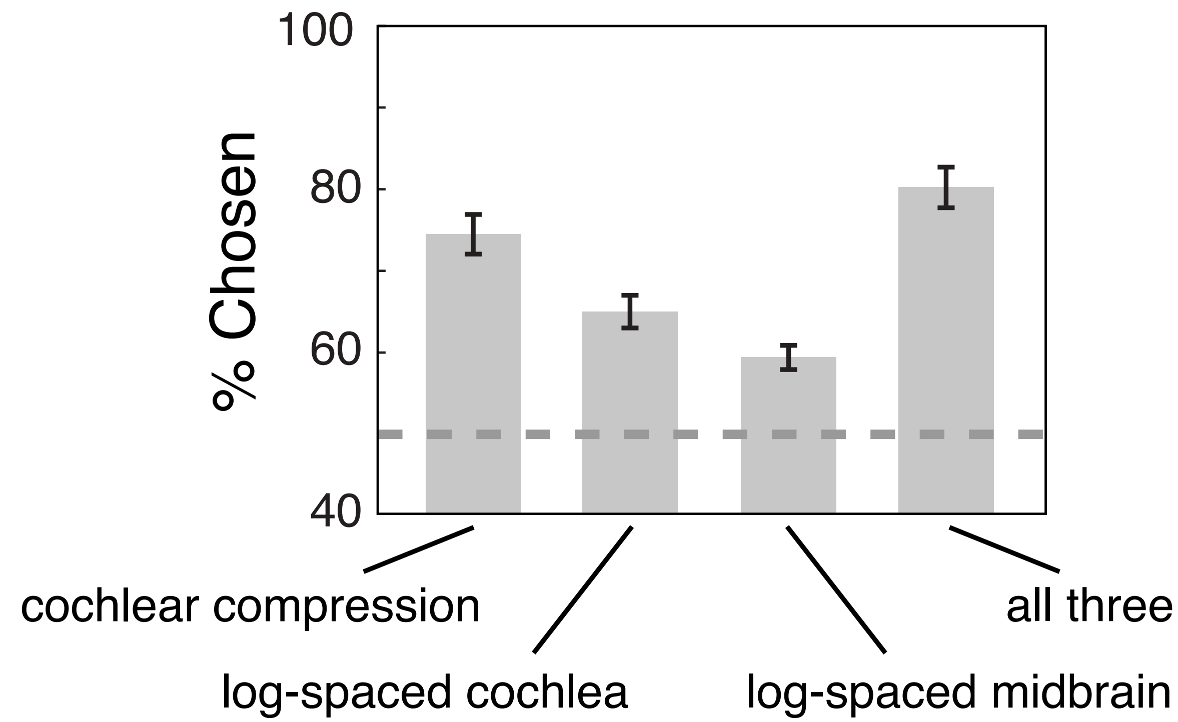
(see mcdermottlab.mit.edu/texture_examples/)

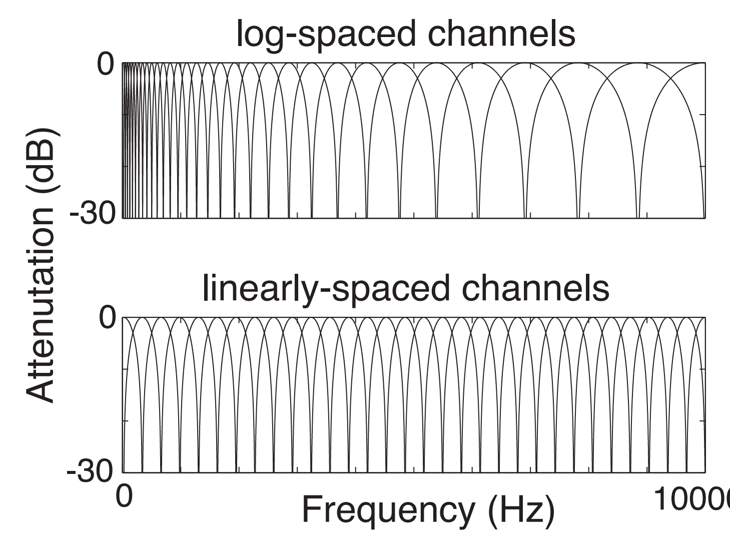
[McDermott & Simoncelli, *Neuron* 2011]

“Biological” model choices do matter



Human perception: which sounds more “real”?



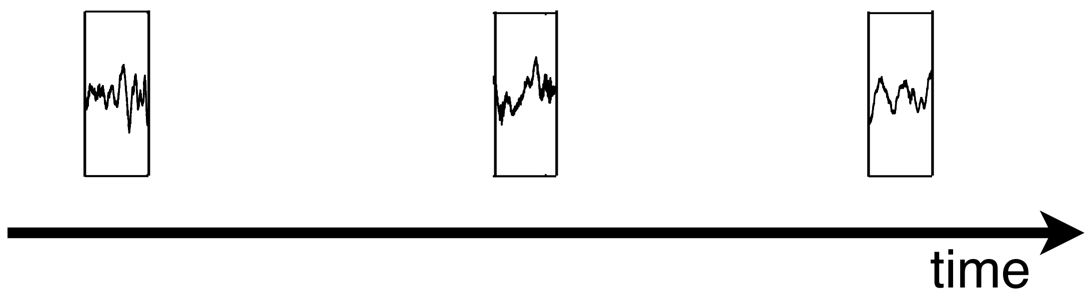


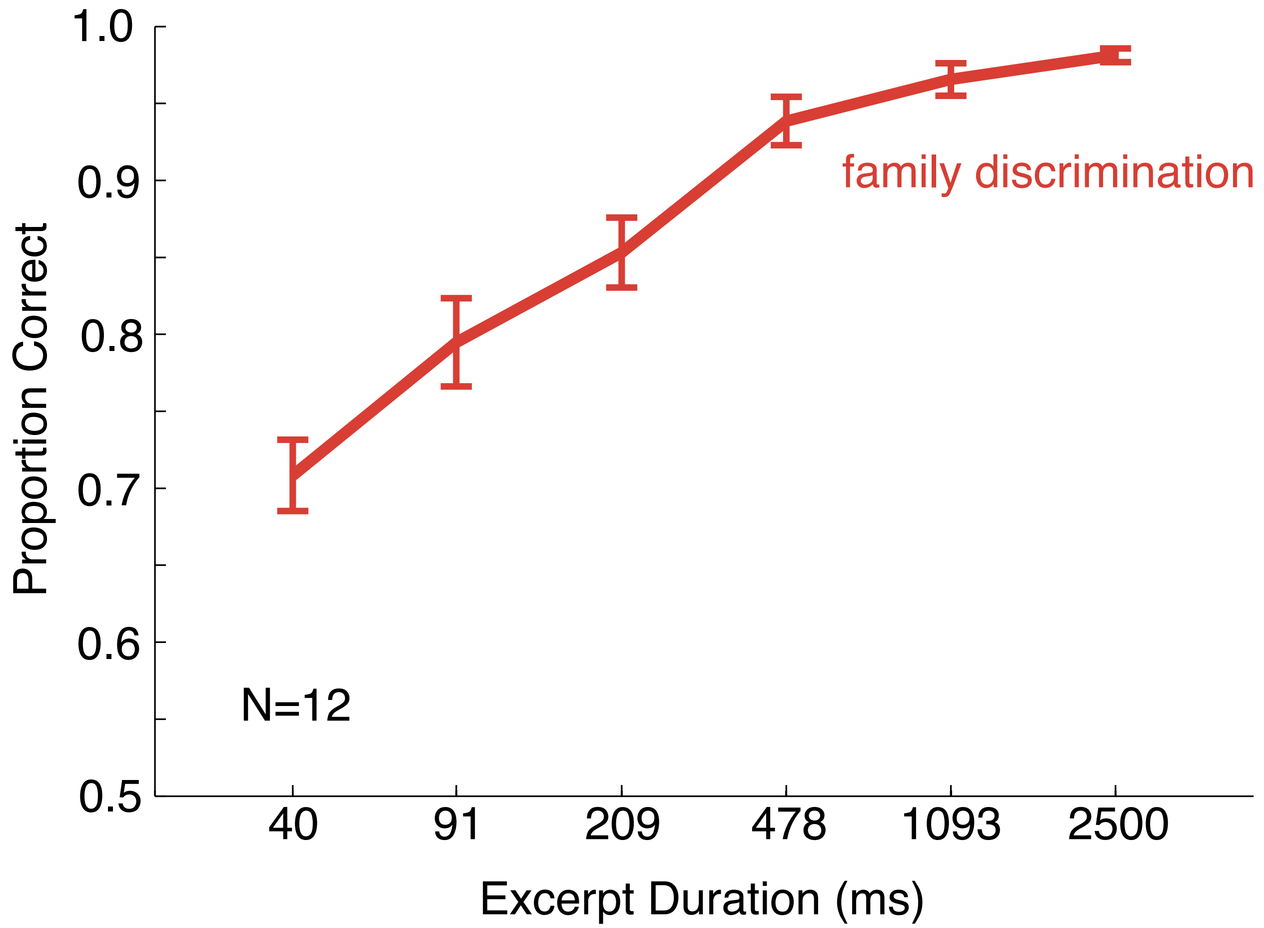
[McDermott & Simoncelli, *Neuron* 2011]

Auditory texture discrimination depends on **duration**

Family discrimination

Which sound was produced by a different source?



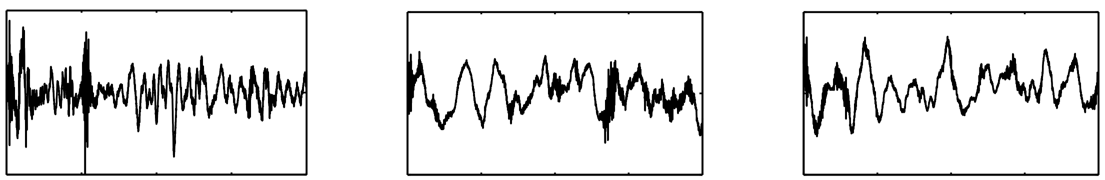


[McDermott & Simoncelli, *Nature Neuro* 2014]

Auditory texture discrimination depends on **duration**

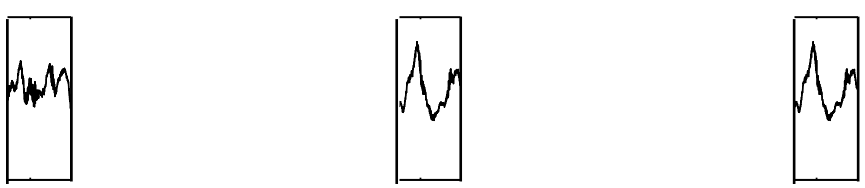
Family discrimination

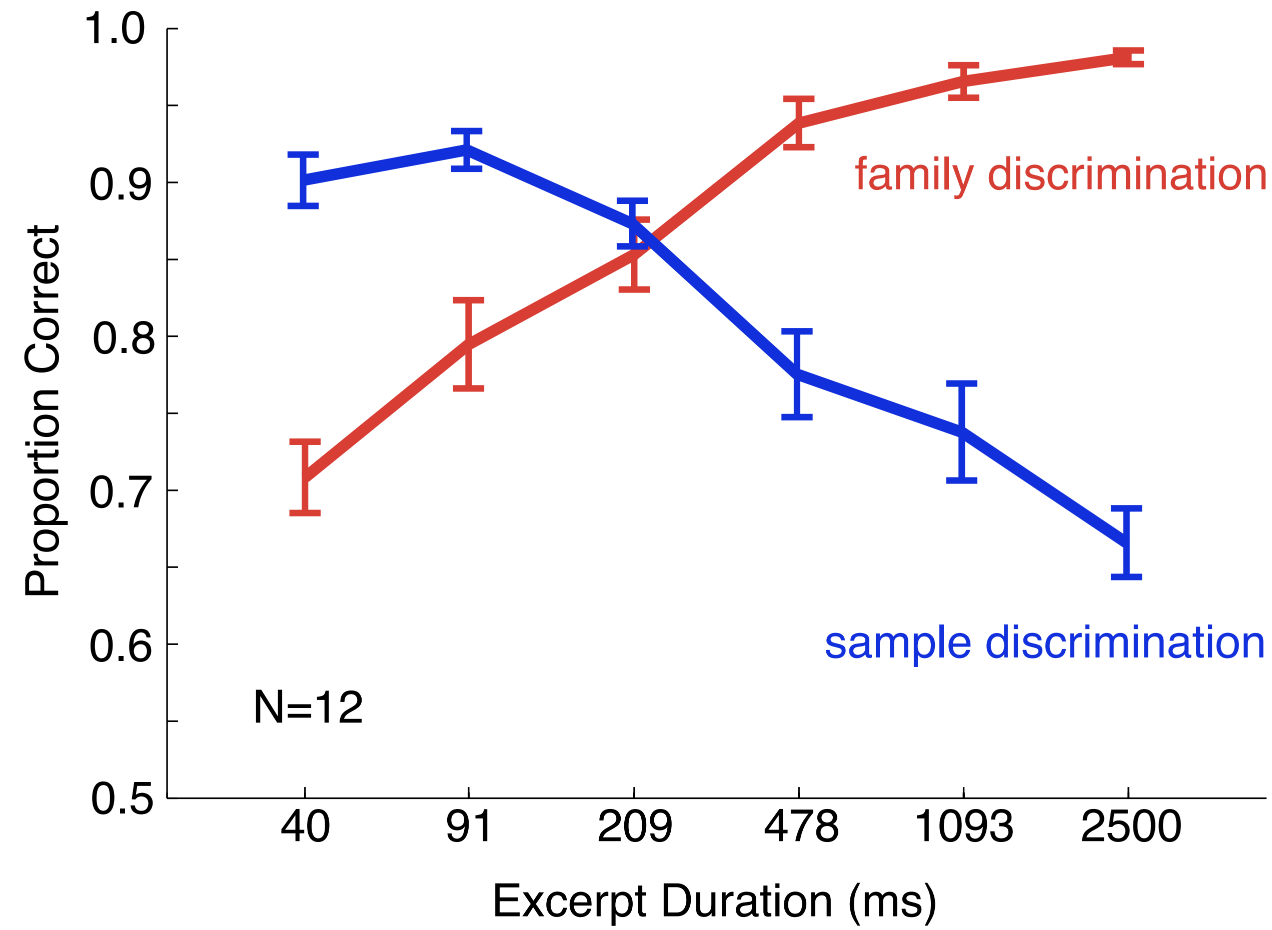
Which sound was produced by a different source?



Sample discrimination

Which sound was different from the other two?

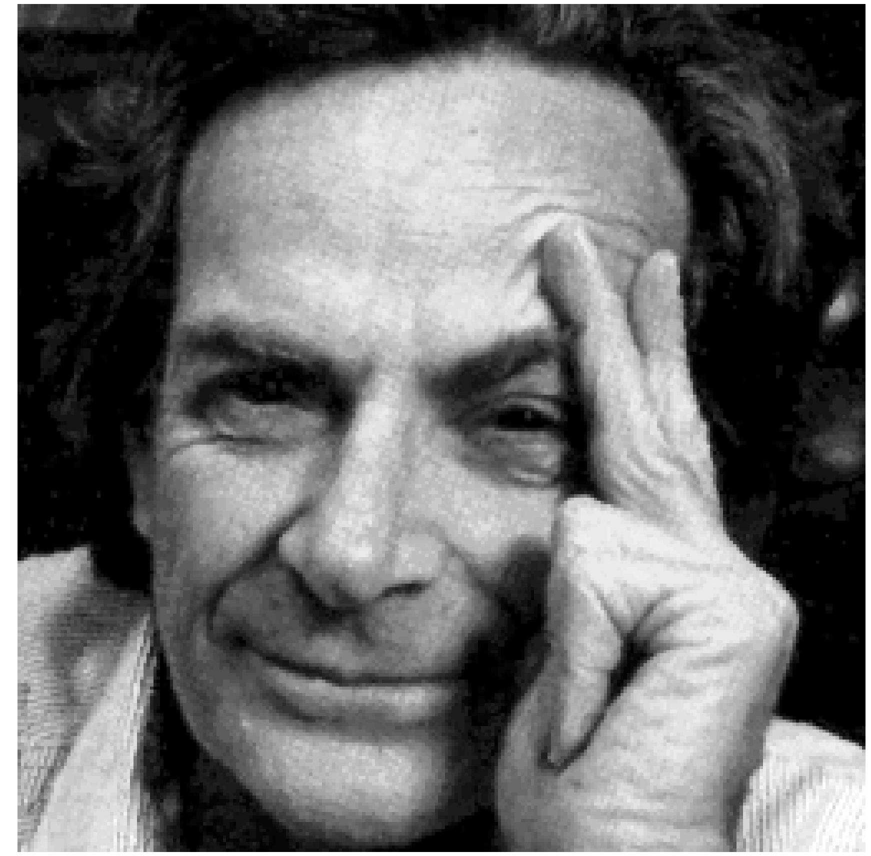




Interpretation: auditory system is forced into "summary mode" for dense long-duration stimuli

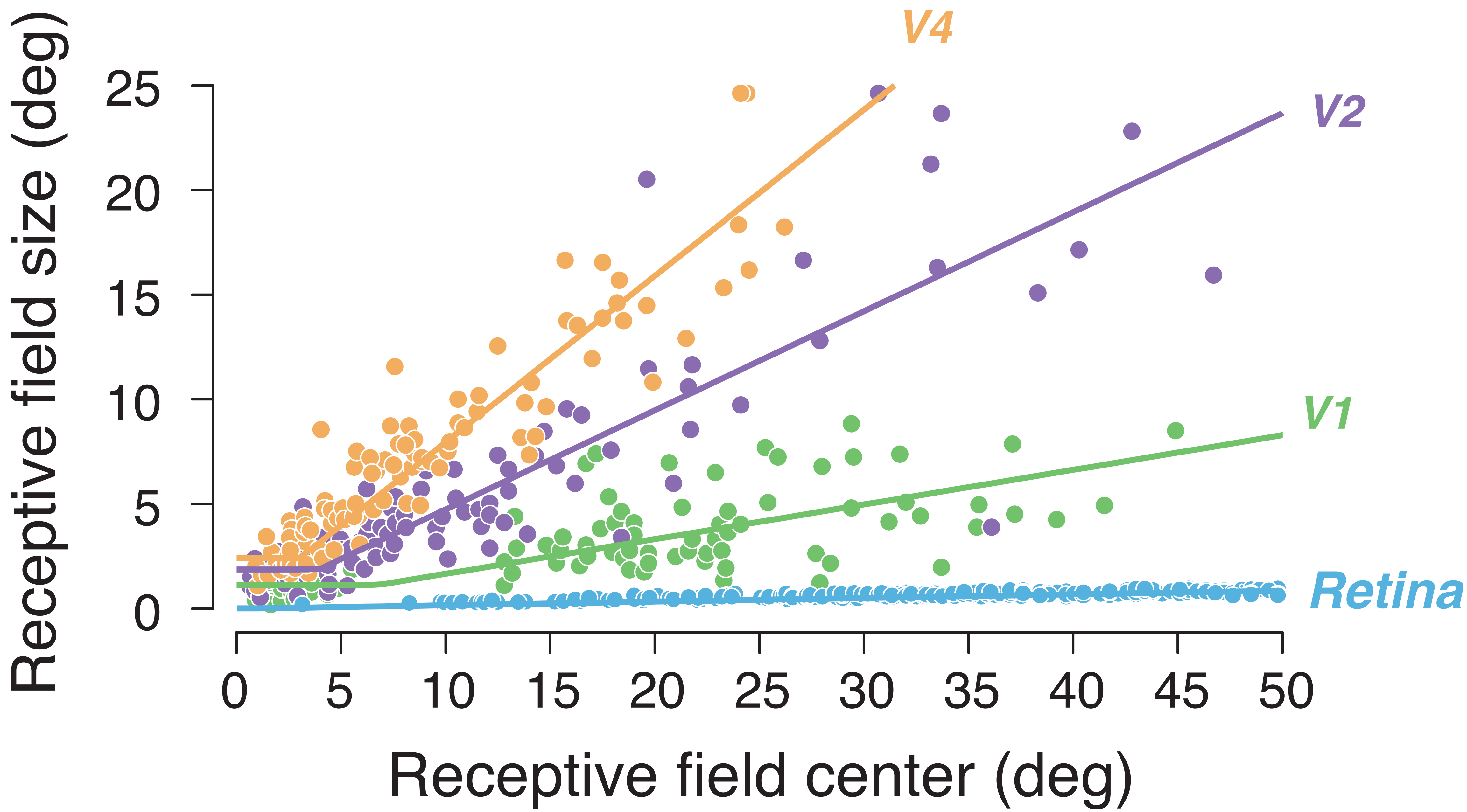
[McDermott & Simoncelli, *Nature Neuro* 2014]

What about non-texture (inhomogeneous) images?



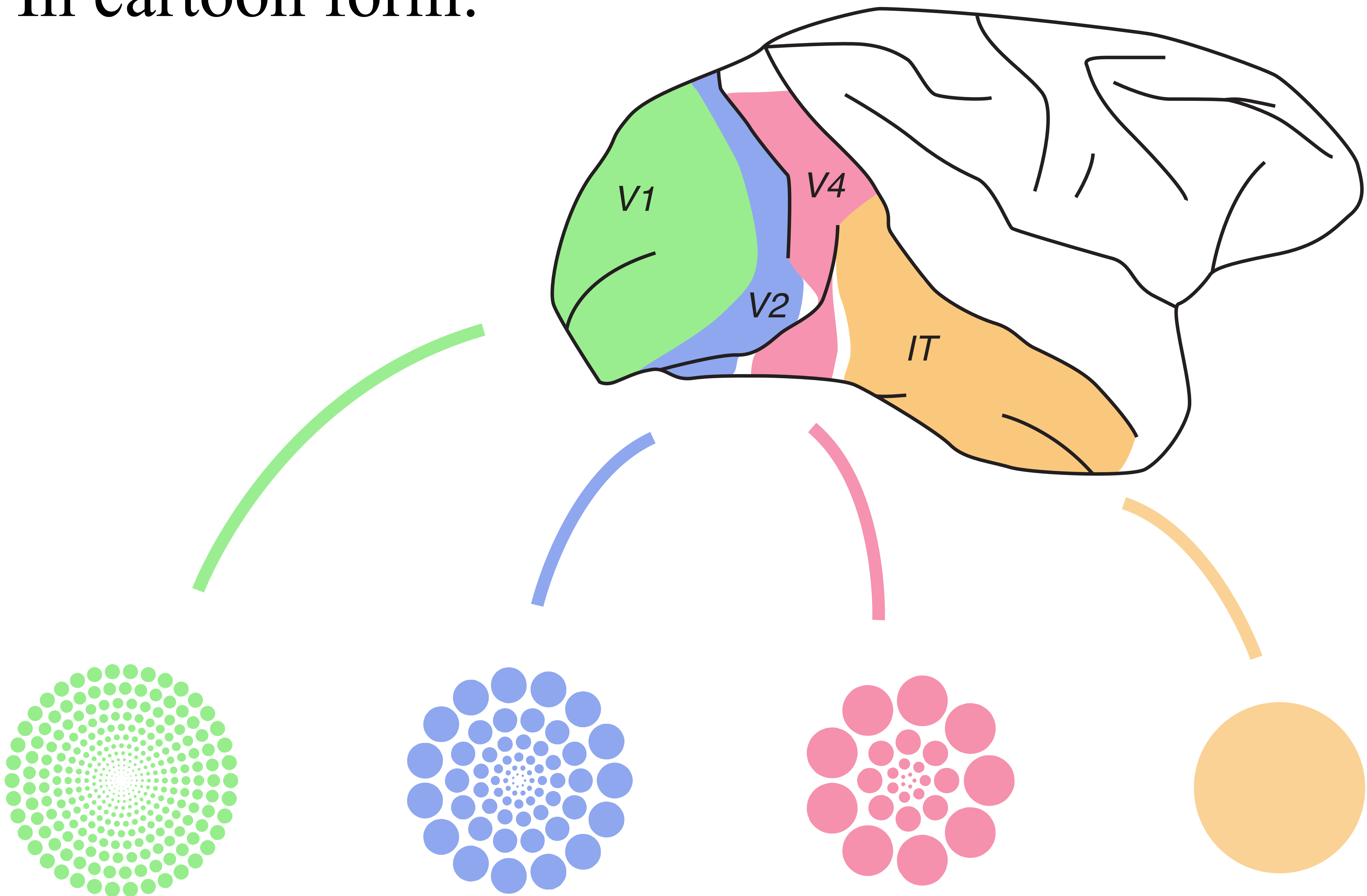
Can we make the model more physiological?

RF sizes grow with eccentricity (distance from fovea)



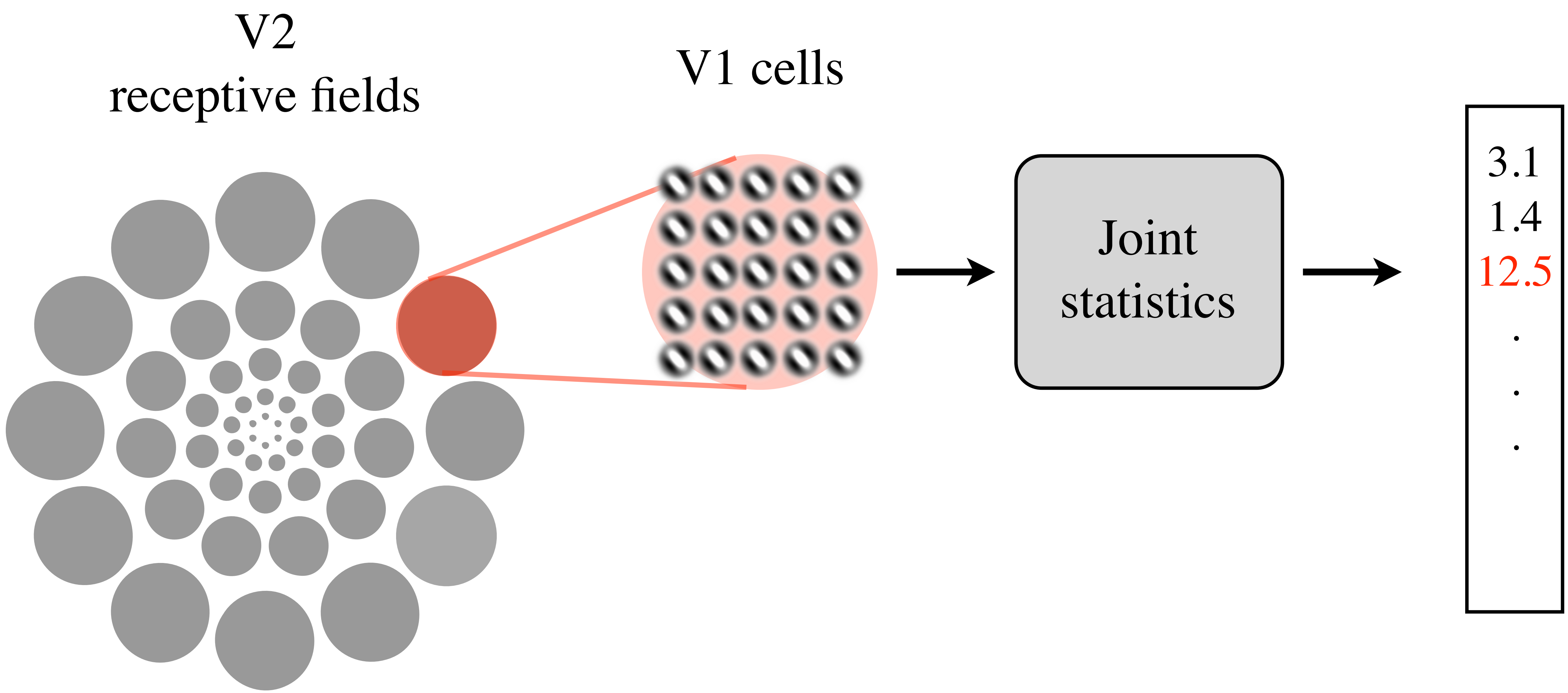
[Freeman & Simoncelli 2011,
macaque data from Gattass et. al., 1981; Gattass et. al., 1988; Perry et. al., 1984]

In cartoon form:

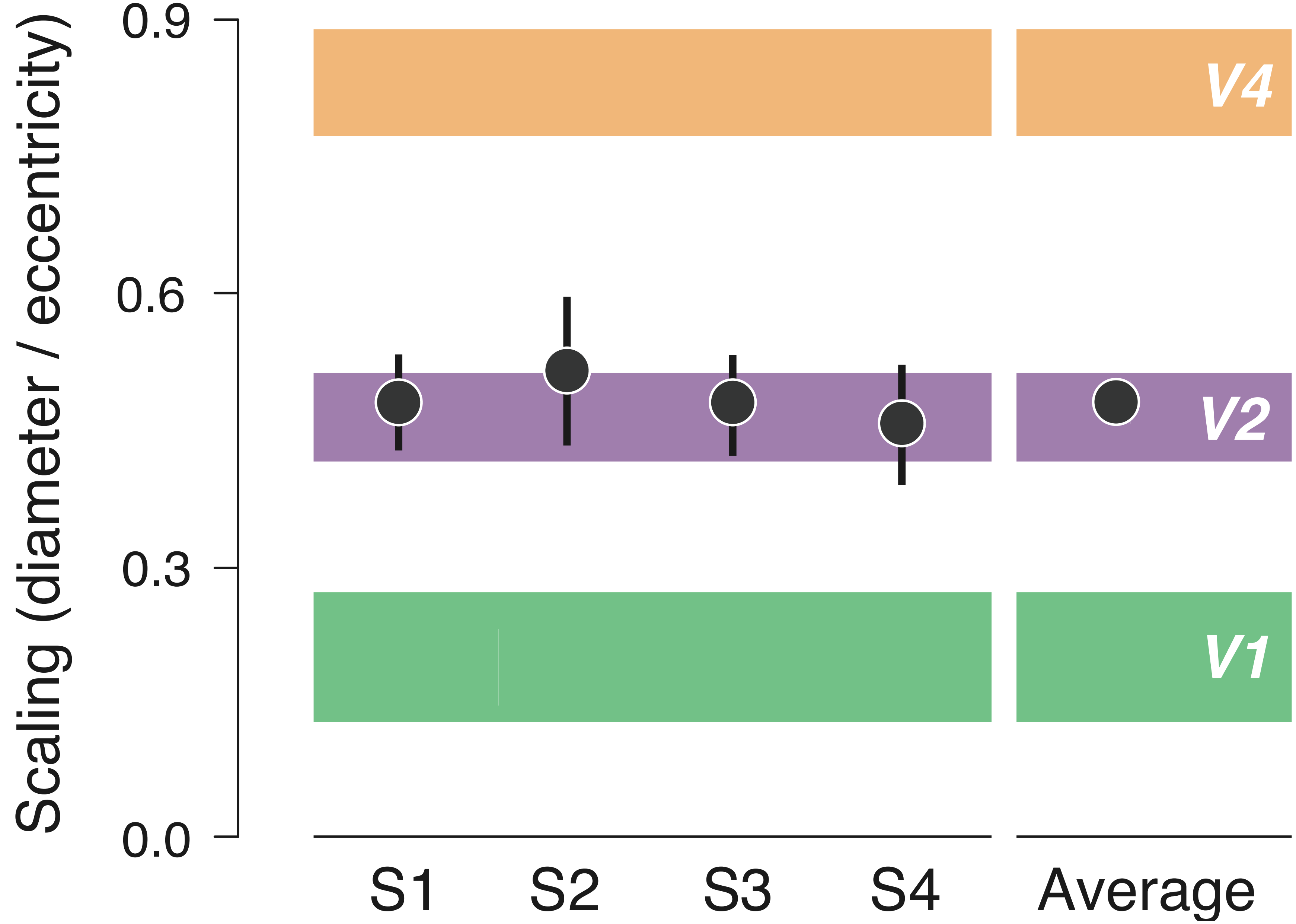


[Freeman & Simoncelli, 2011]

Local texture representation in the ventral stream



[Freeman & Simoncelli, 2011]



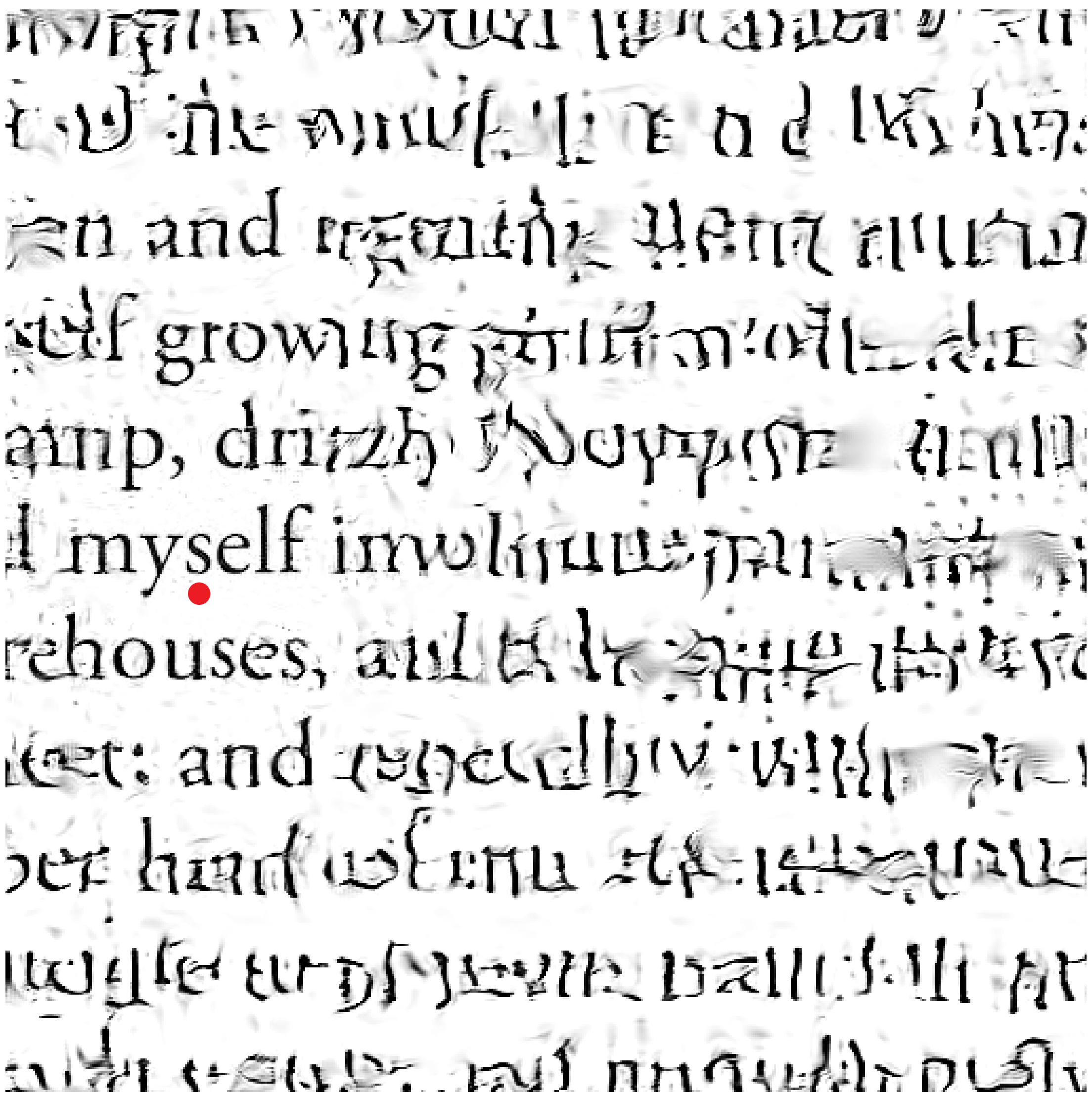
Human Perception

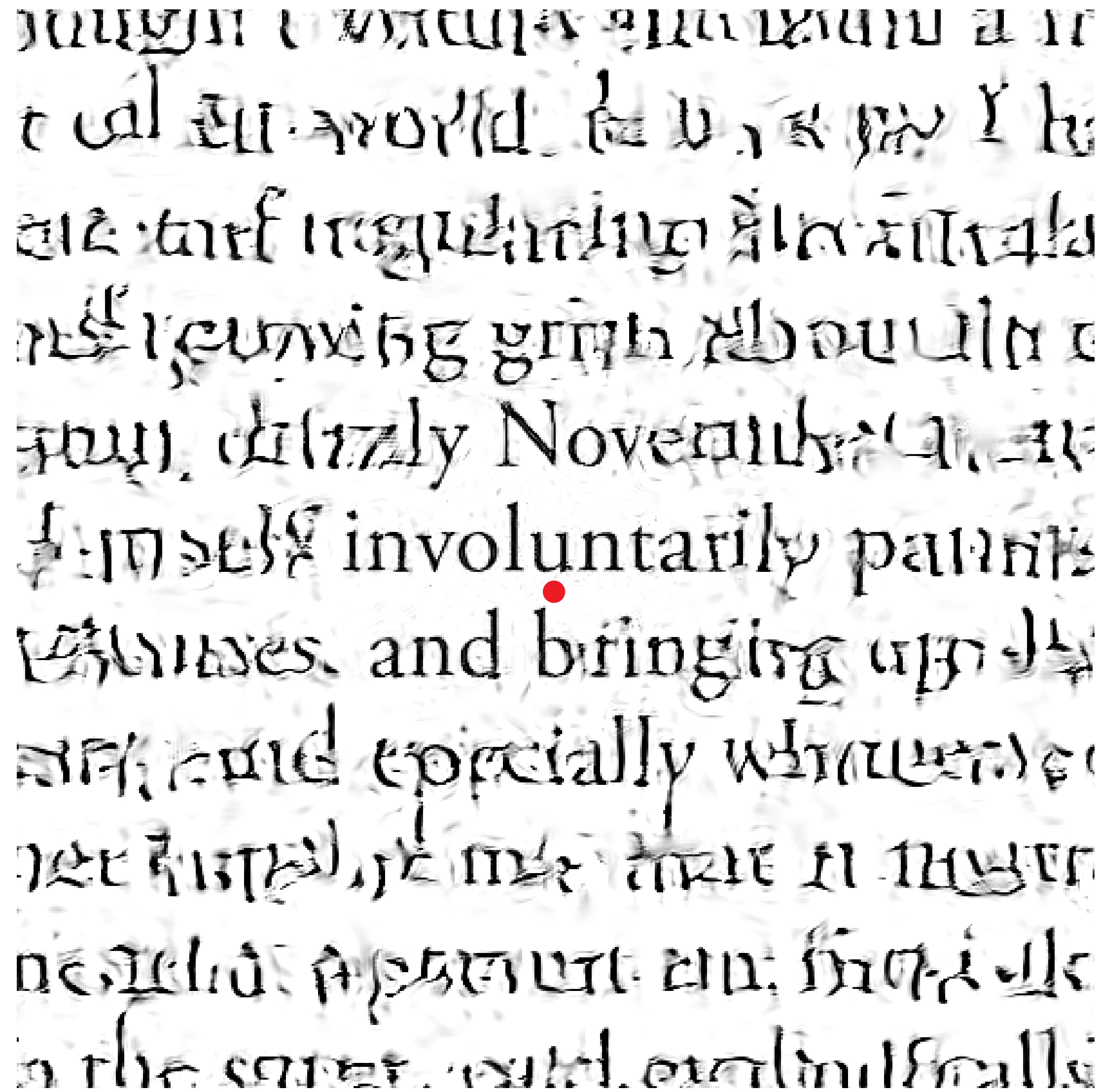
Macaque Physiology

[Allman & Kaas, 1971; Allman & Kaas, 1974; Gattass et.al., 1981; van Essen et.al., 1984; Maguire & Baizer, 1984; Burkhalter & van Essen, 1986; Gattass et.al., 1987; Desimone & Schein, 1987; Gattass et.al., 1988; Cavanaugh et. al., 2002]

[Freeman & Simoncelli, 2011]

Reading





[Freeman & Simoncelli, 2011]

Linear filtering, rectification, local statistics ....

Can this really explain all of vision?

“Perhaps texture, somewhat redefined, is the primitive stuff out of which form is constructed”

[Lettvin, 1976]

Shallow hierarchical models with biological attributes are more powerful than expected

Synthesis provides a powerful test of representation

- can be used to verify invariances
- can also be used to verify metric properties
  - distance/curvature (ICLR16)
  - perceptual quality (Cosyne17)

How can we learn representations (unsupervised)?

- compression (next talk!)

Javier Portilla

Jeremy Freeman

Corey Ziemba

Josh McDermott

hhmi
Howard Hughes Medical Institute

National Eye Institute
NATIONAL INSTITUTES OF HEALTH

NSF