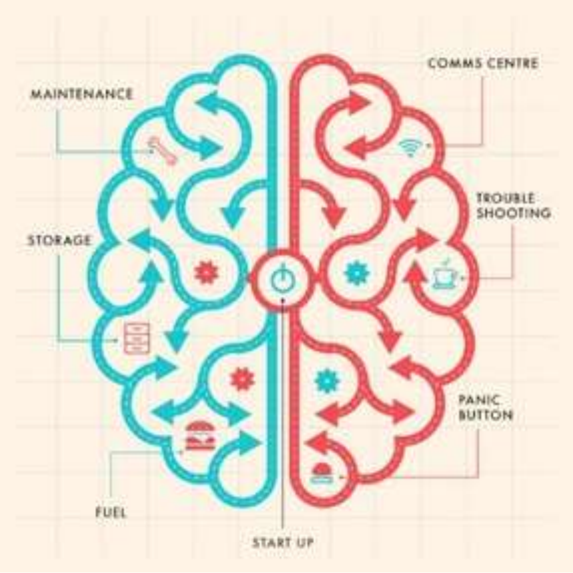
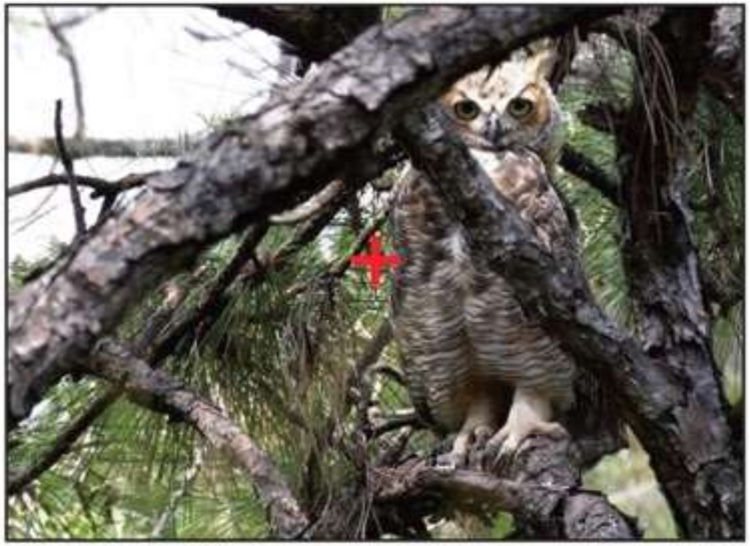
Object recognition
in ventral visual cortex and deep networks

Giuseppe di Pellegrino
Department of Psychology, University of Bologna

g.dipellegrino@unibo.it



Cognition and Neuroscience
Second cycle Degree in Artificial Intelligence – 2024/25



Visual object-related tasks

Coarse discrimination	Bird or dog?
Fine discrimination	Owl or osprey?
Position estimation	Left or right of + ?
Segmentation	In front or behind the branch?
Memory-based tasks	Familiar or novel?
Valence judgment	Threatening or pleasant?
⋮	⋮

What is vision?

What does it mean, to see?
Vision is the process of discovering what is present in the world, and where it is. (Marr, Vision, 1982)

Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information (Marr and Nishihara, 1978).

Vision dominates our perceptions and memories of the world and appears even to frame the way we think.

Vision is used not only for object recognition but also for guiding our movements.
These separate functions are mediated by at least two parallel and interacting pathways.

Vision, and more generally the brain, is a system that analyzes information (information processing device): receives inputs and transforms them into outputs.

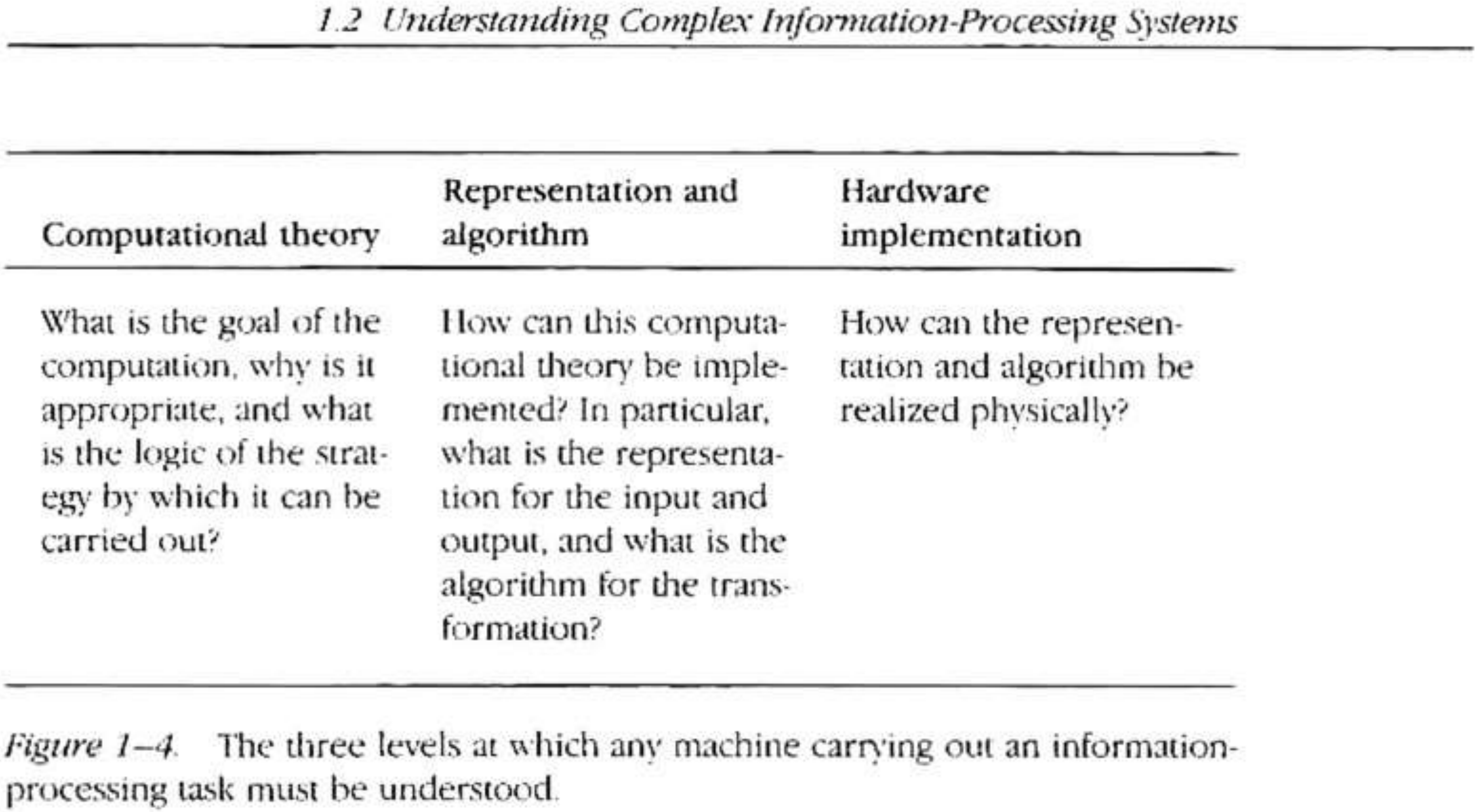
What is vision?

“What does it mean, to see? The plain man's answer (and Aristotle’s, too) would be, to know what is where, by looking. In other words, vision is the process of discovering what is present in the world, and where it is.
Vision is therefore, first and foremost, an information-processing task, but we cannot think of it just as a process.
For if we are capable of knowing what is where in the world, our brains must somehow be capable of representing this information.
The study of vision must therefore include not only the study of how to extract from images the various aspects of the world that are useful to us, but also an inquiry into the nature of the internal representations by which we capture this information and thus make it available as a basis for decisions about our thoughts and actions.
This duality the representation and the processing of information-lies at the heart of most information-processing tasks and will profoundly shape our investigation of the particular problem posed by vision”

(David Marr, Vision, 1982)

Marr’s “tri-level hypothesis”

“arr argued that to truly understand vision — biological or artificial — we have to approach it on multiple levels: computational, algorithmic, and implementational. That framework still guides how we think about perception today, whether we’re studying the human brain or training a neural net



1.2 Understanding Complex Information-Processing Systems

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Figure 1–4. The three levels at which any machine carrying out an information-processing task must be understood.

Marr, Vision (1982)

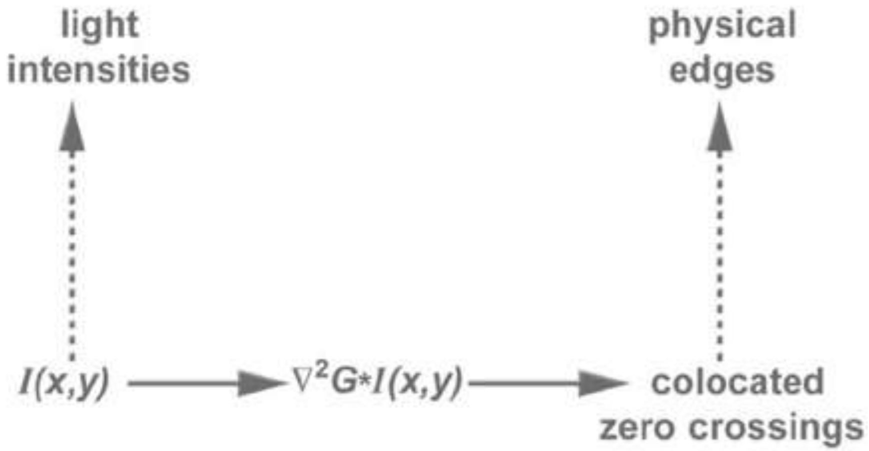


Figure 3. Formal description of edge detection. Retinal image, $I(x, y)$, is convoluted through the filtering operator ∇^2G, where G is a Gaussian and ∇^2 is a second-derivative (Laplacian) operator. Early vision processes include several filters with different Gaussian distributions, and each produces a different set of zero crossings. Intensity values are interpreted as representing light intensities in the visual field, and the colocated zero crossings are interpreted as representing edges, such as object boundaries.

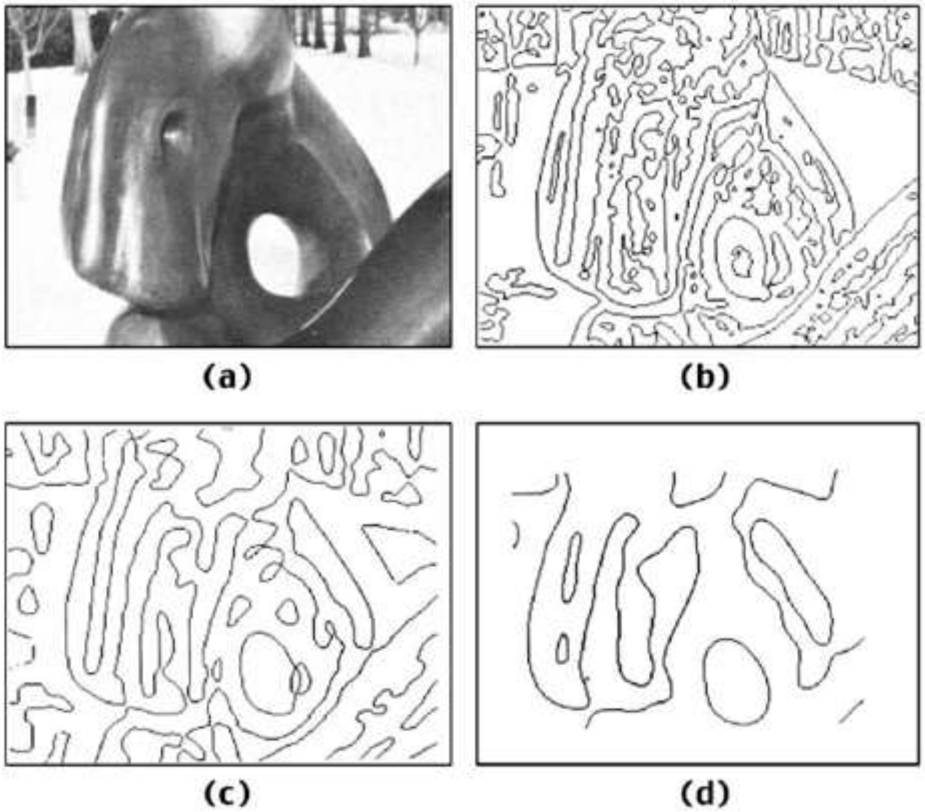
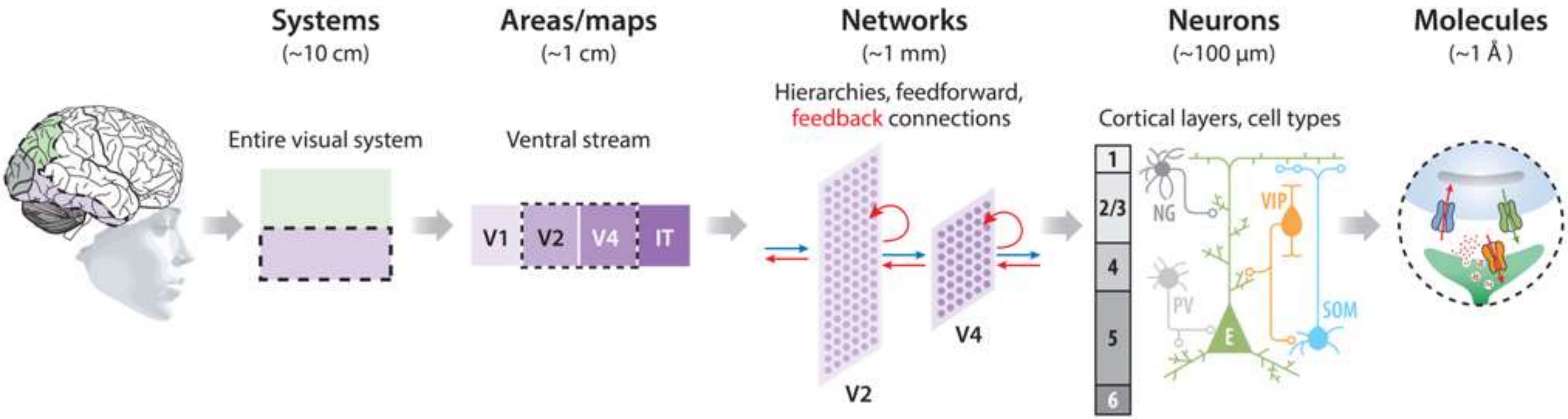


Figure 4. Different sets of zero crossings. Image (*a*) is convoluted with different-sized filters. The zero crossings obtained are shown in *b*, *c*, and *d*. Many of the fine details obtained through the smaller-sized filter (*b*) are not obtained by the larger-sized filter (*c*), but some of the zero crossings obtained in *c* do not appear in the *b*. From Marr and Hildreth (1980), 201, fig. 6. Reprinted with permission from Royal Society Publishing.

Shagrir, Philosophy of Science, 2010

Aim

Models that successfully integrate all of these levels to accurately and causally bridge from molecules to minds in visual object recognition.

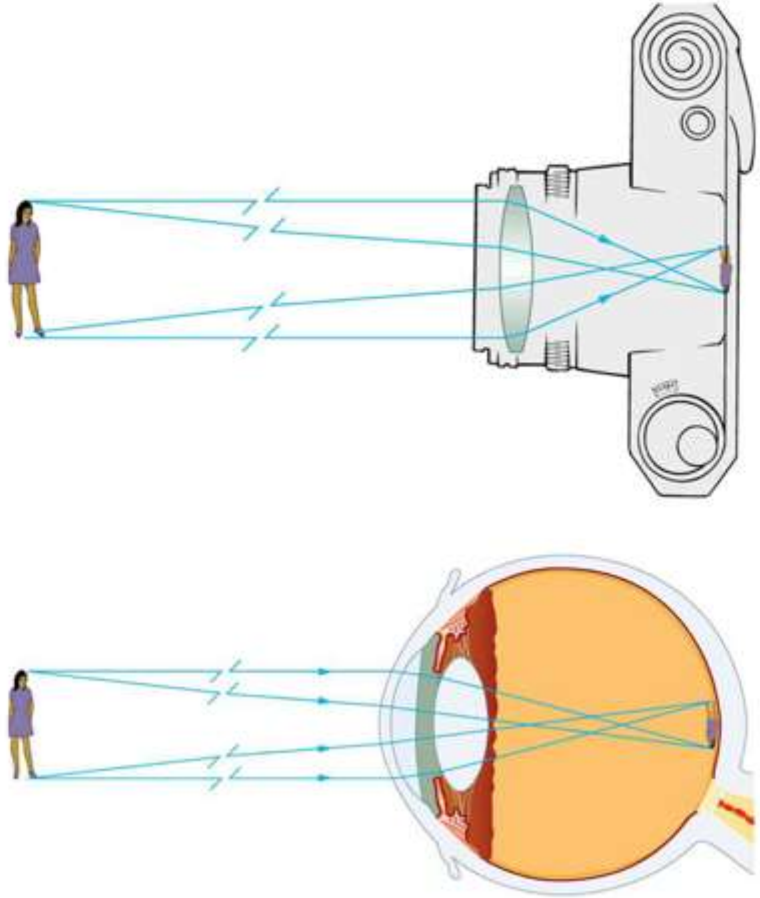


Mechanistic understanding of object recognition with increasing levels of detail.

Vision is often incorrectly compared to the operation of a camera.

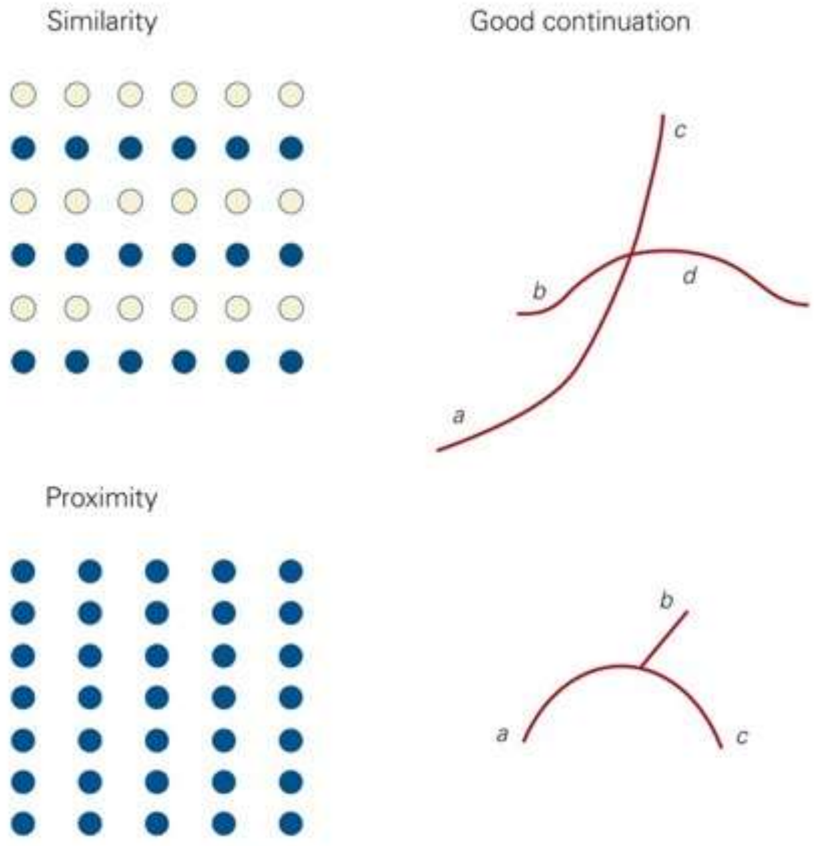
A camera simply reproduces point-by-point the light intensities in one plane of the visual field.

The visual system, in contrast, does something fundamentally different. It interprets the scene and parses it into distinct components, separating foreground from background.

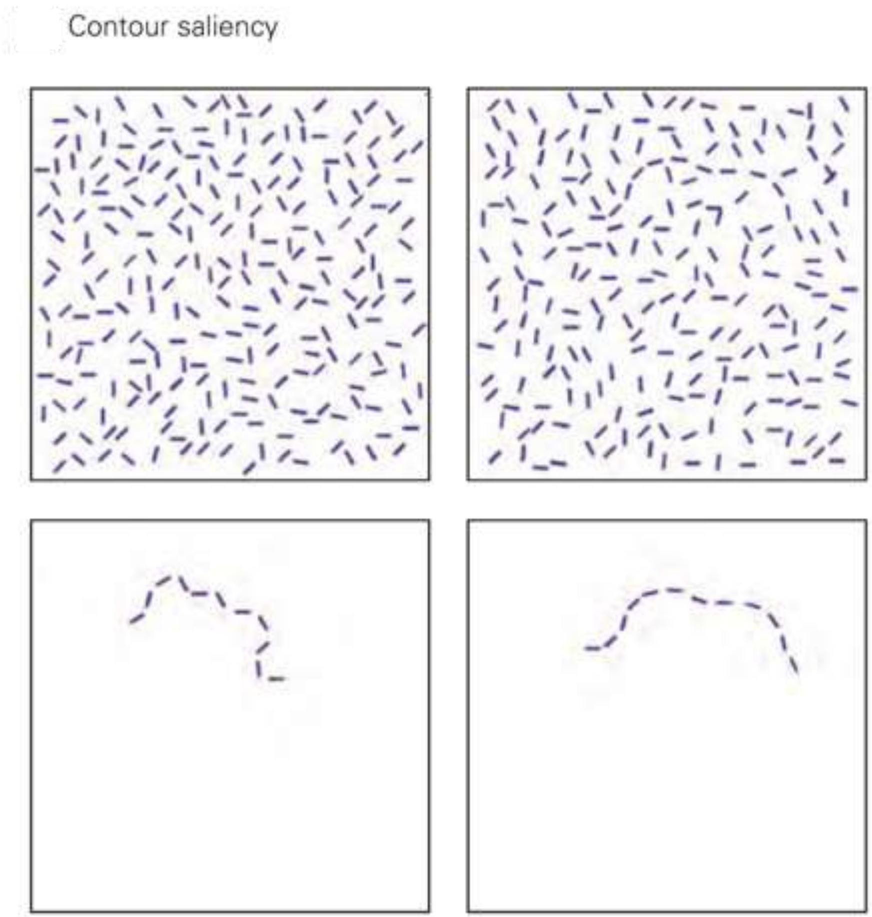


Vision is an active and bidirectional process

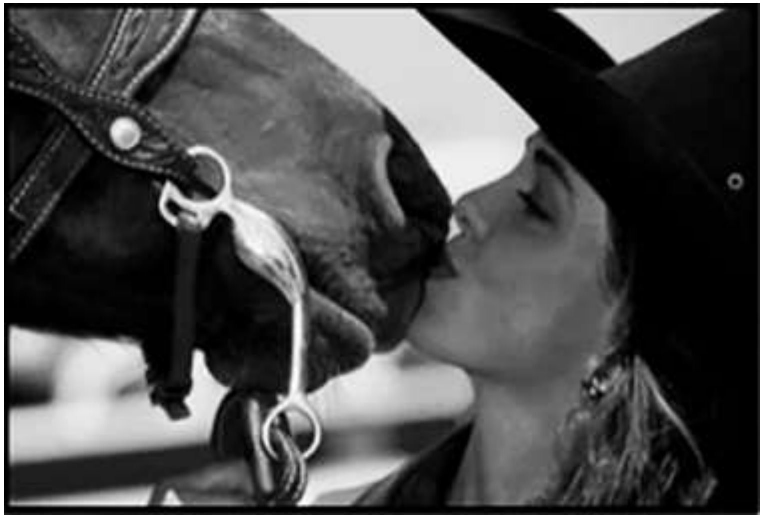
Vision is a generative process that involves more than just the information provided to the retina. The brain has a way of looking at the world, a set of expectations about the structure of the world that derives in part from experience and in part from built-in neuronal wiring.



To link the elements of a visual scene into unified percepts, the visual system relies on organizational rules such as similarity, proximity, and good continuation.



The principle of good continuation is also seen in contour saliency. On the right, a smooth contour of line elements pops out from the background, whereas the jagged contour on the left is lost in the background.



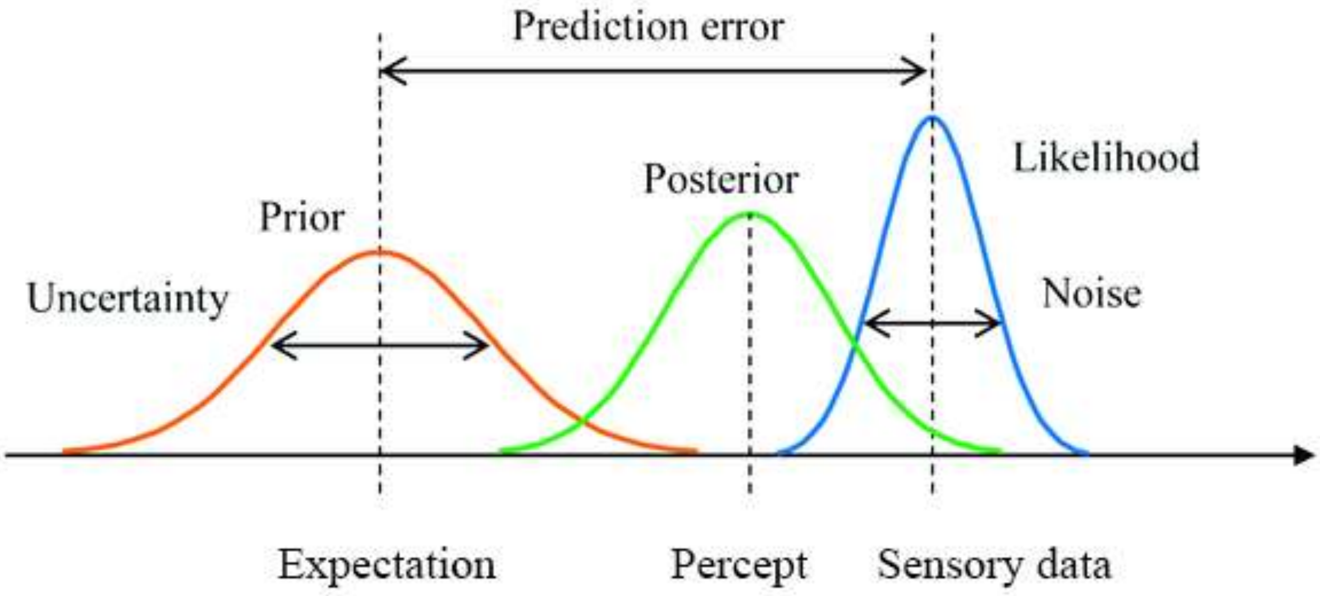
Visual priming
Higher-order representations of shape (in memory) guide lower-order processes of surface segmentation.

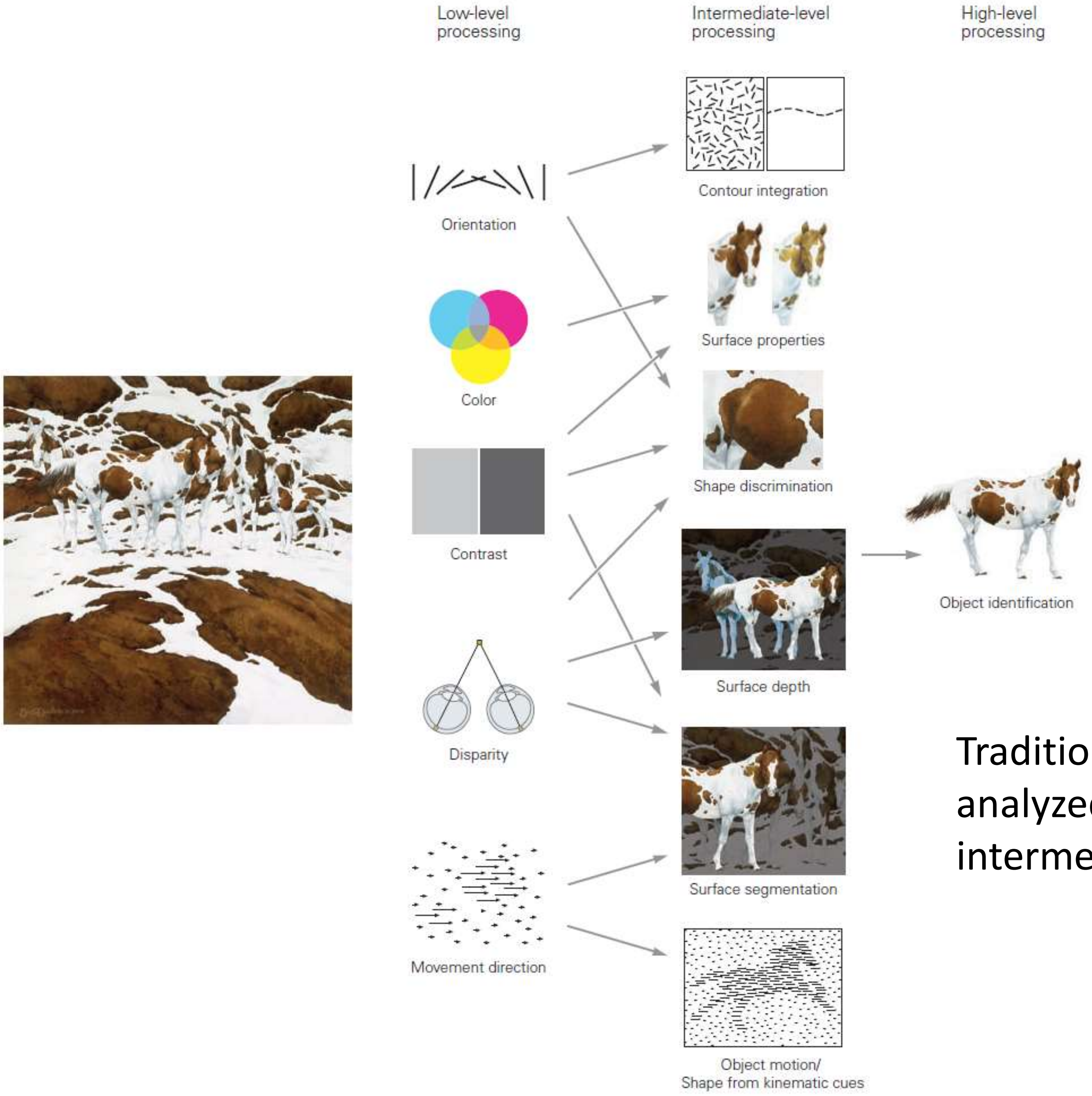
Bayesian theories treat the visual system as an ideal observer that uses prior knowledge about visual scenes and information in the image to infer the most probable interpretation of the image.

The posterior probability of a possible real–world stimulus S (i.e., percept) is proportional to the product of the prior probability of S (that is, the probability of S before receiving the stimulus I, e.g., expectation) and the likelihood (the probability of I given S, i.e., sensory data).

Is it reasonable to assume that the visual system knows the probability calculus and operates according to it?

$$p(S|I) = \frac{p(I|S)p(S)}{p(I)}$$





Traditionally, a visual scene is analyzed at three levels: low, intermediate and high.

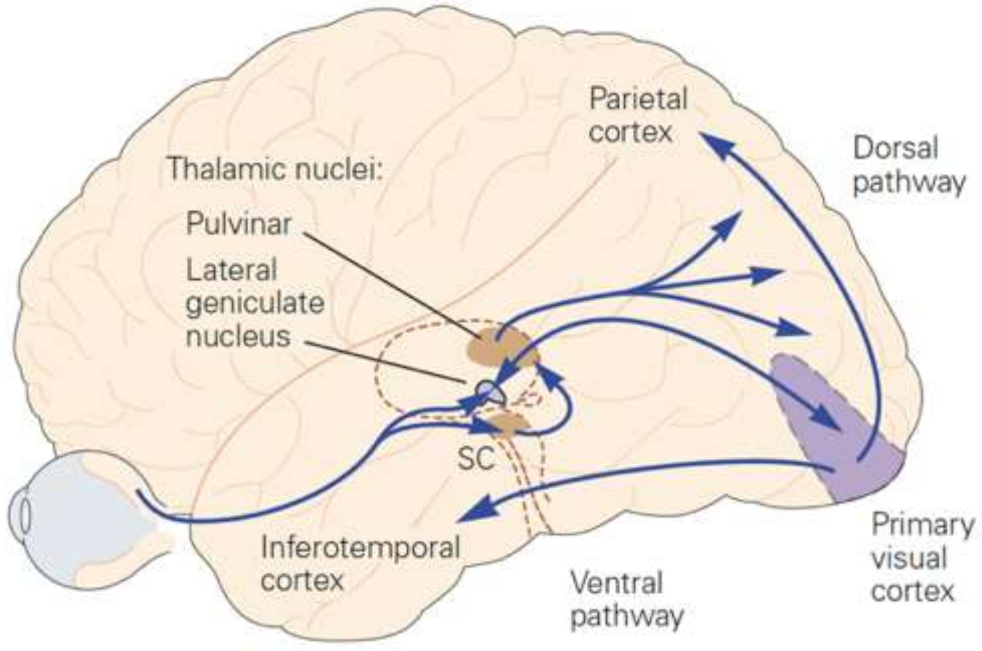
At the lowest level, visual attributes such as local contrast, orientation, color, depth and motion are processed.

Intermediate-level processing: low-level features are used to parse the visual scene.
Local orientation is integrated into global contours (contour integration); local visual features are assembled into surfaces, objects are segregated from background (surface segmentation), surface shape is identified from depth, shading and kinematic cues.

The highest level concerns object recognition.

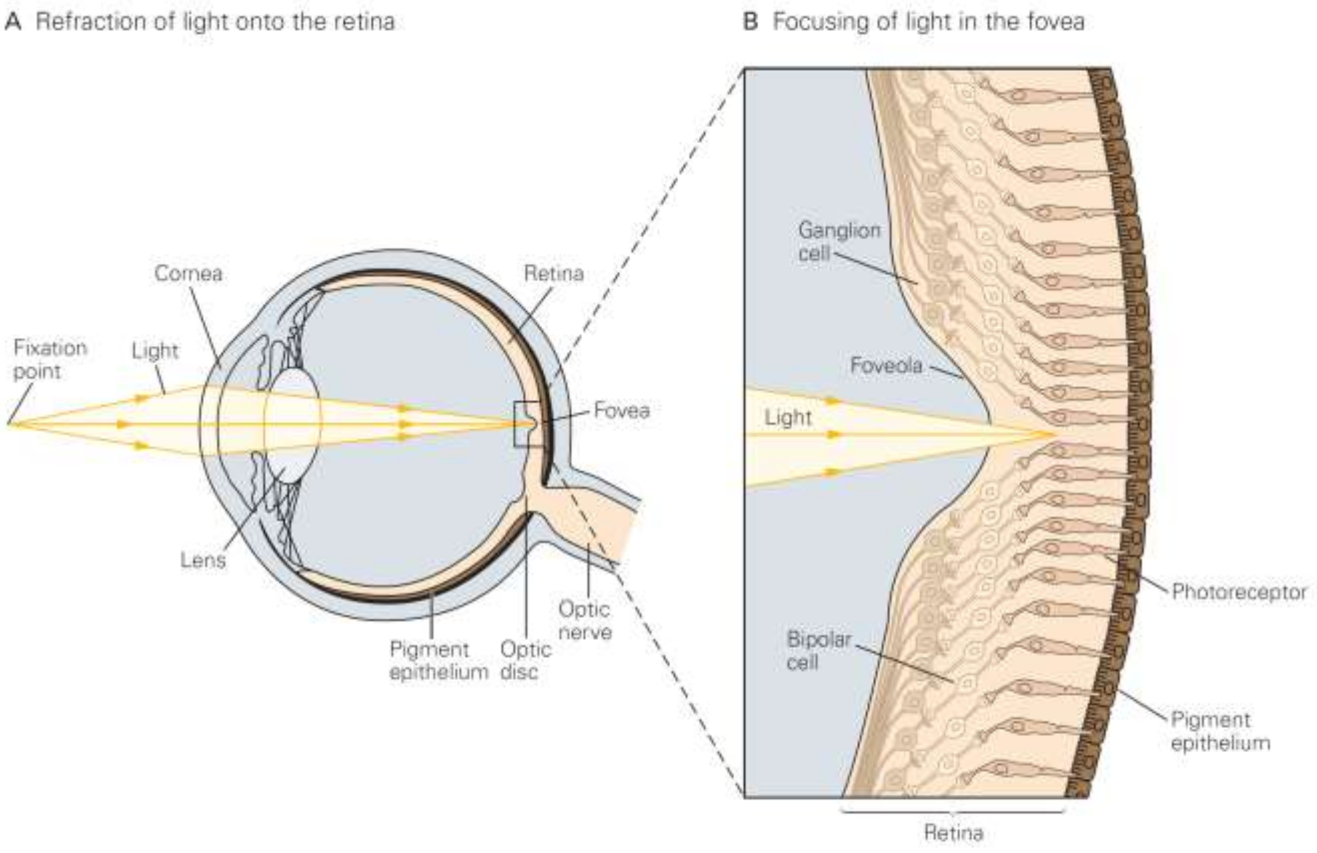
Once a scene has been analyzed by the brain and the objects have been recognized, the objects can be associated with memories of shapes and their meanings.

Visual processing is mediated by the retino-geniculo-striate pathway



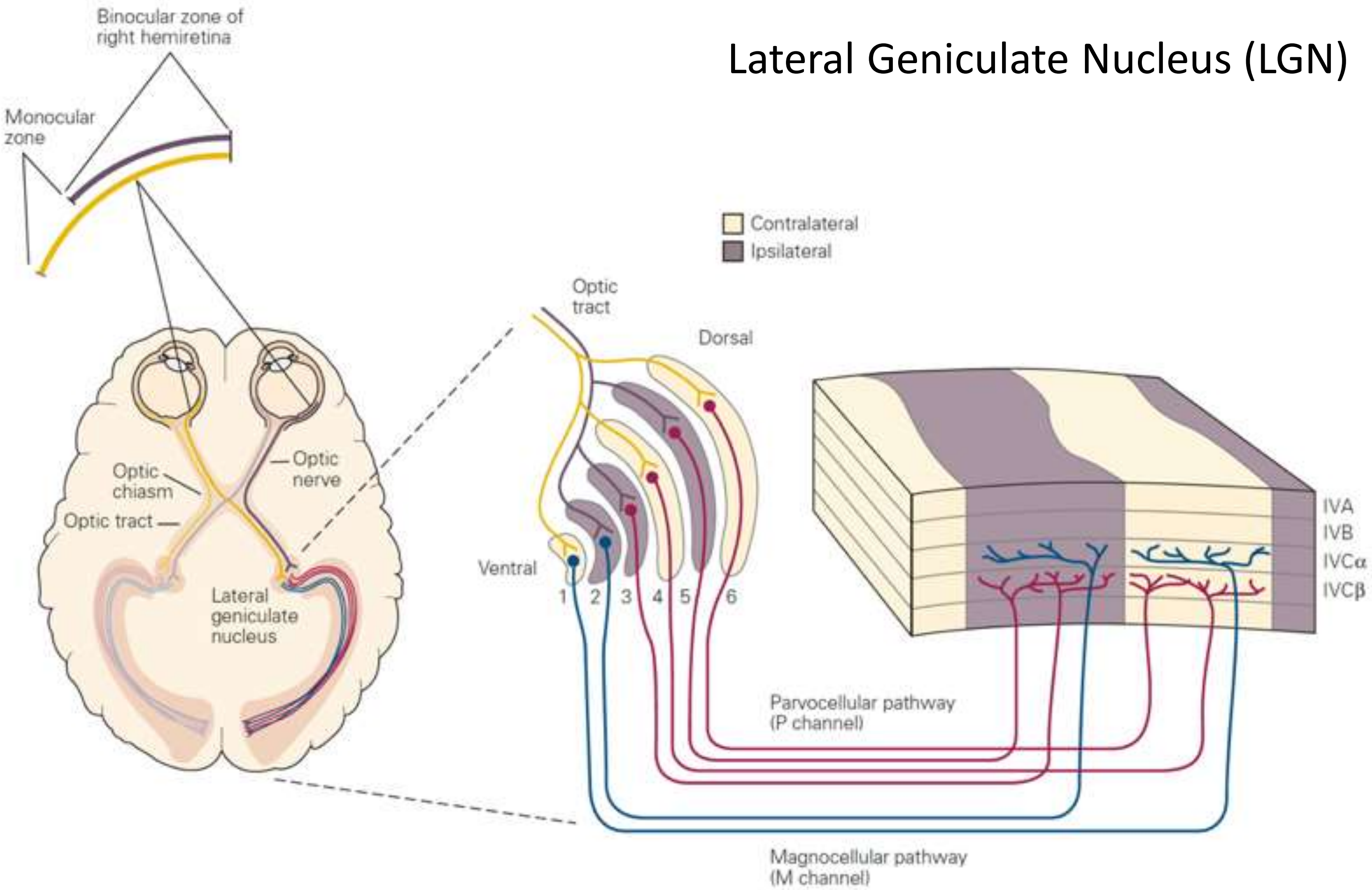
This pathway includes:

- Retina;
- Lateral geniculate nucleus (LGN) of the thalamus;
- Primary visual cortex (V1) or striate cortex;
- Extrastriate visual areas



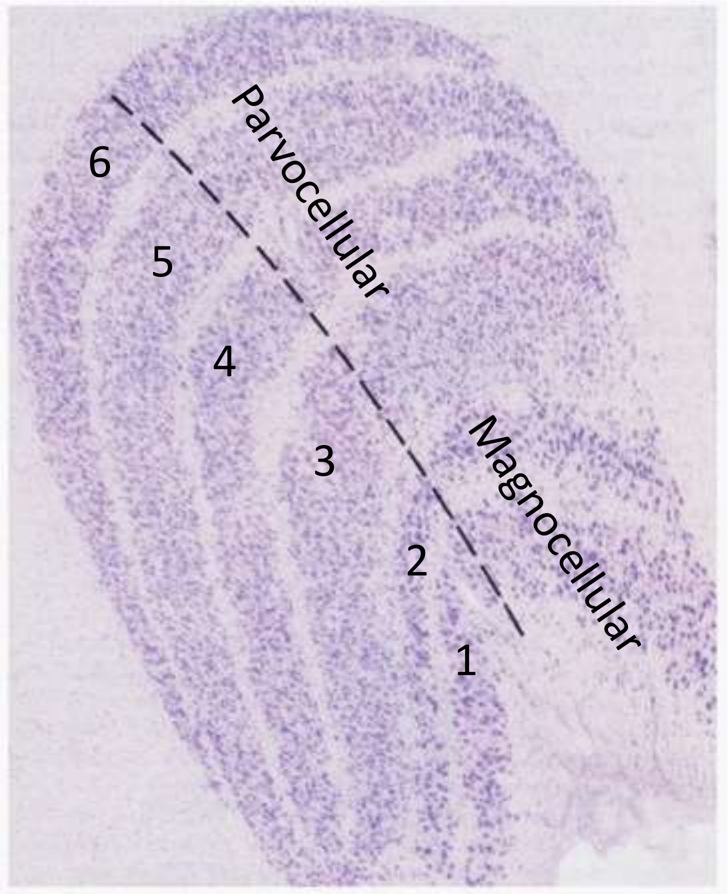
The brain’s analysis of visual scenes begins in the two retinas, which transform the visual input into neural signals, a process known as phototransduction.

Lateral Geniculate Nucleus (LGN)


Binocular zone of right hemiretina
Monocular zone
Optic chiasm
Optic nerve
Optic tract
Lateral geniculate nucleus
Contralateral
Ipsilateral
Optic tract
Dorsal
Ventral
1
2
3
4
5
6
IVA
IVB
IVCα
IVCβ
Parvocellular pathway (P channel)
Magnocellular pathway (M channel)

Beyond the optic chiasm, the axons from nasal and temporal hemiretinas carrying input from one hemifield join in the optic tract, which extends to the LGN of the thalamus.

Lateral Geniculate Nucleus (LGN)



Interspersed between the Magno and Parvo layers, are thin but dense layers called Koniocellular which receive from the K retinal ganglion cells.

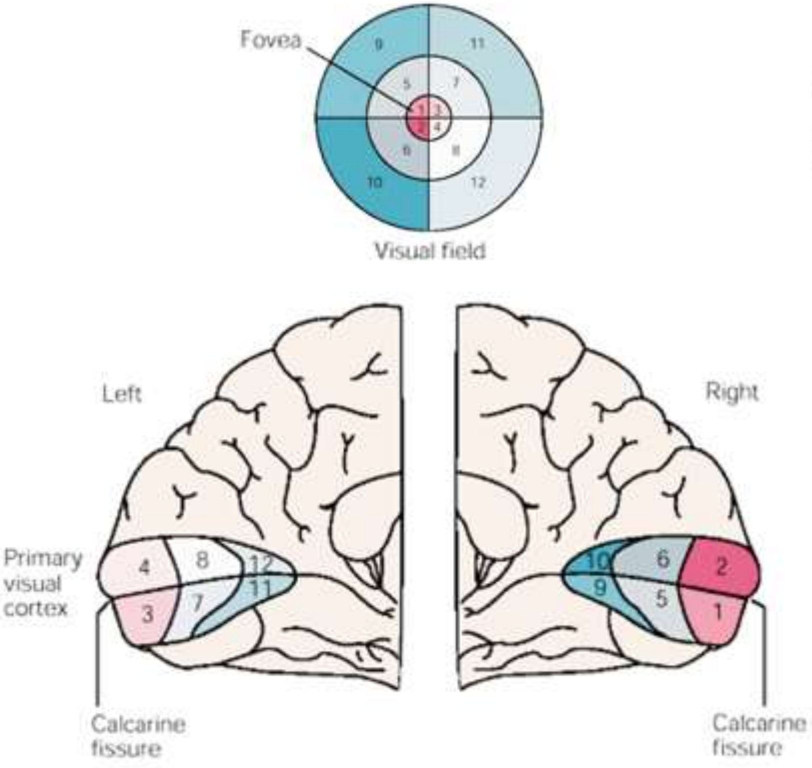
In primates, LGN is a layered structure consisting of six layers, of which two Magnocellular layers (layers 1 and 2), and four Parvocellular layers (layers 3 to 6).
Each layer receives input from either the ipsilateral eye (temporal hemiretina, layers 2, 3, 5) or the contralateral eye (nasal hemiretina, layers 1, 4, 6).

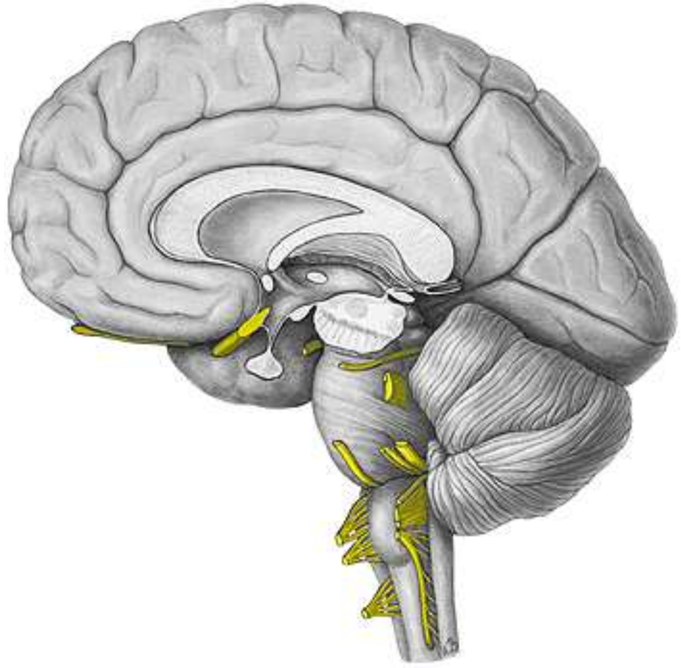
Since each layer contains a map of the contralateral hemifield, the six maps are stacked on top of each other and in spatial register.

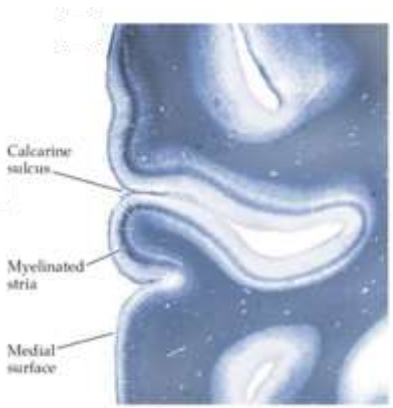
The magnocellular layers project to the IVCα layer while the Parvocellular layers project to the IVCβ layer of V1. The Koniocellular intercalated layers project to the blobs (layers 1-2) of V1.

Primary visual cortex (V1)

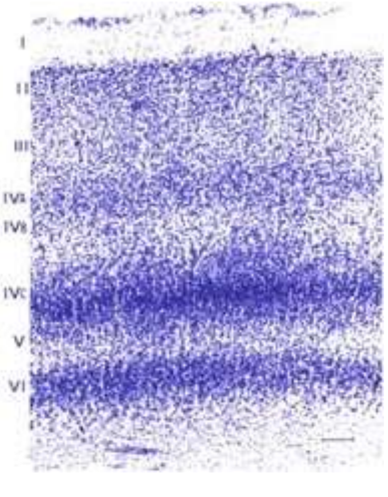
In humans, V1 (BA17) is located in the occipital portion of the brain along the calcarine fissure of the brain.
V1 constitutes the first level of cortical information processing.







Nissl stain of V1 reveals the different layers quite clearly.

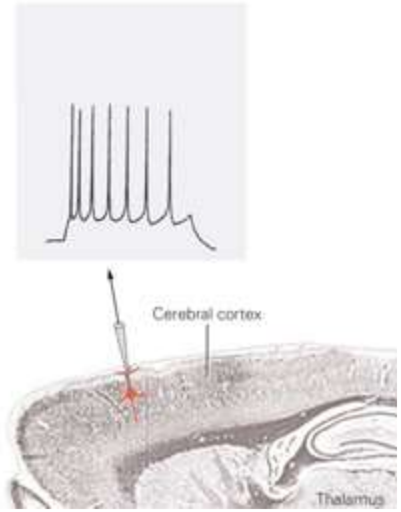


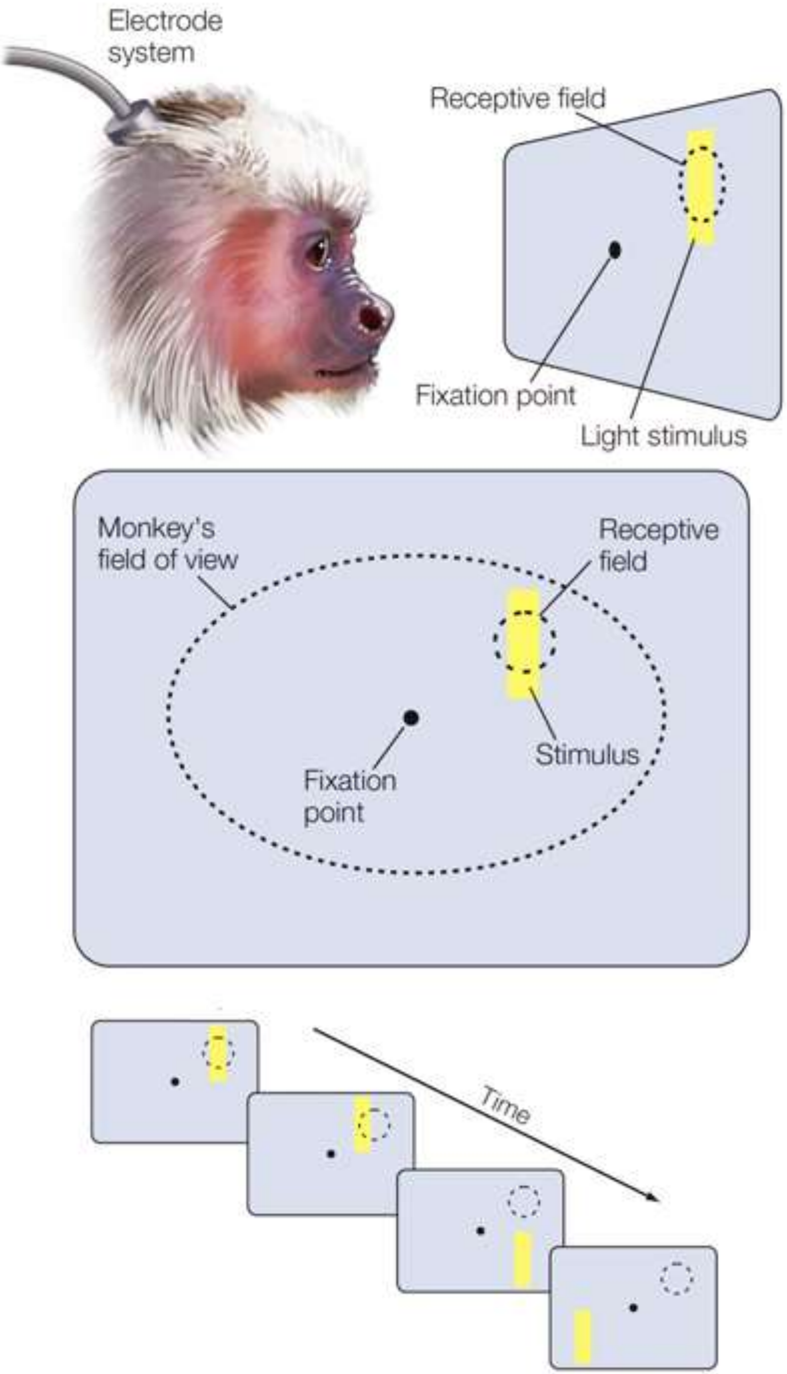
Single-cell recording

This technique allows recording signals (firing rate) from single neurons.

A fine-tipped, usually metal (platinum), electrode is inserted in the animal brain to record extracellularly change in electrical activity called action potential (AP, 1ms duration) or spike. Collected signals are appropriately amplified, filtered, viewed through an oscilloscope, and saved to a computer for offline analysis.
Since spikes are all-or-none highly stereotyped signals, most information is encoded in the brain as neuron firing rate, i.e., the number of AP in 1s.
The primary goal of single-cell recording experiments is to determine what experimental manipulations produce a consistent change in the firing rate of an isolated neuron.

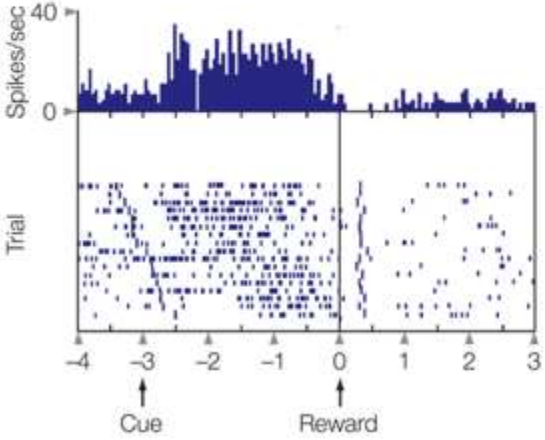
Disadvantages
– invasive
Advantages
– high spatial and temporal resolution
– differentiation between excitation and inhibition

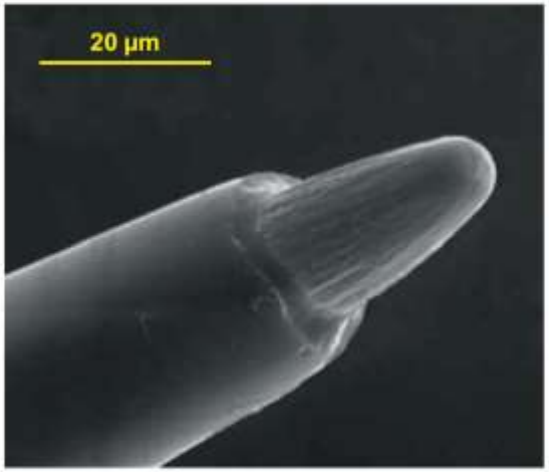




Receptive field

In early visual areas, neurons respond to stimuli in only a limited region of space.
Th is region of space is referred to as that cell’s receptive field (RF).
Monkey is required to maintain fixation, and stimuli are presented at various positions in the field of view.
The neuron below fires vigorously only when the stimulus is presented in the upper right quadrant, thus defining the upper right region as the RF for this cell.





Recording of the activity of a V1 neuron. In the receptive field a moving stimulus is presented, directed downward or upward. Note that neuron firing rate is higher when the stimulus moves up.

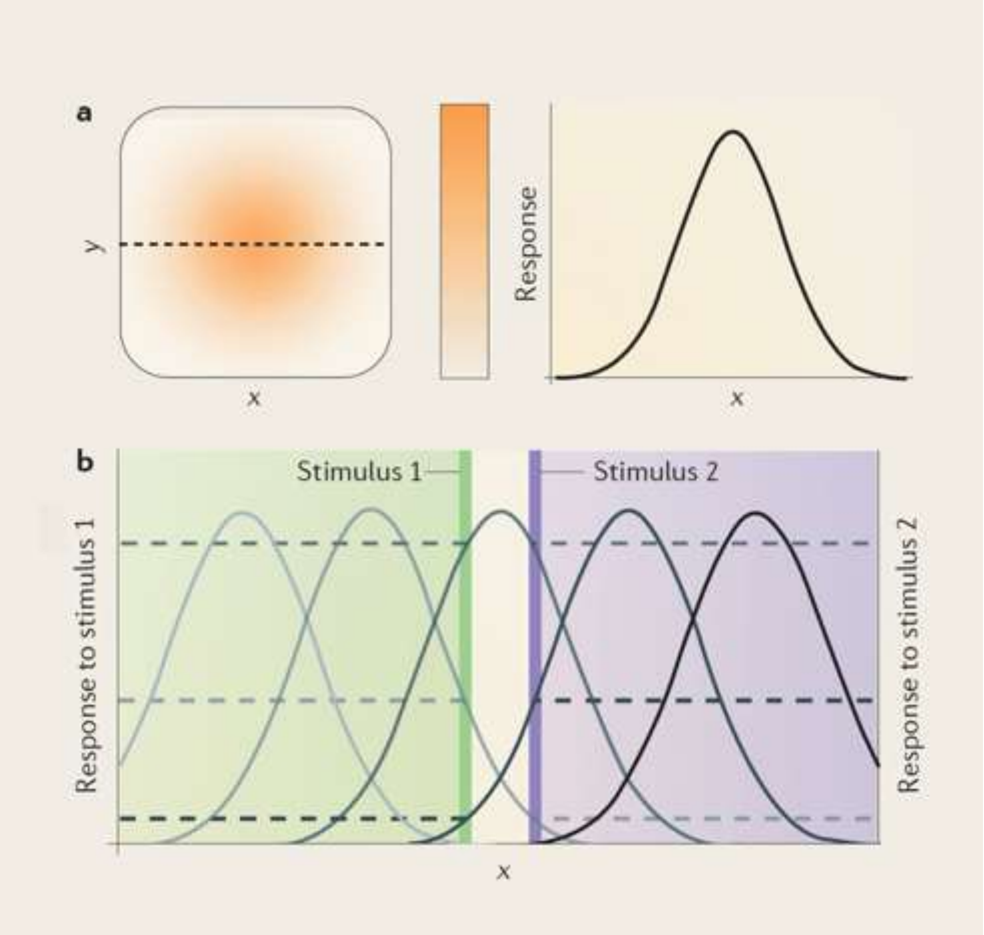
Why studying the nonhuman primate model?

First, it has human-level perceptual capabilities

Second, it allows high-spatial- and -temporal-resolution neural measurements and targeted causal perturbations to interrogate brain circuits, not feasible in humans.

Third, it has evolutionary proximity and established brain-area homologies.

Kar & DiCarlo, Ann Rev. Vis. Sci., 2024



The receptive field (RF; Charles Sherrington, 1906) of a neuron is defined as the part of stimulus space within which a stimulus elicits a response from the neuron.

In the visual system, the neuron responds to a stimulus presented in a region of space in the visual field (that is, its RF) but not to the same stimulus when it is presented outside this region.

A given stimulus typically elicits the strongest response from the centre of the RF, with the response gradually declining as the stimulus is presented further away from the centre of the RF.

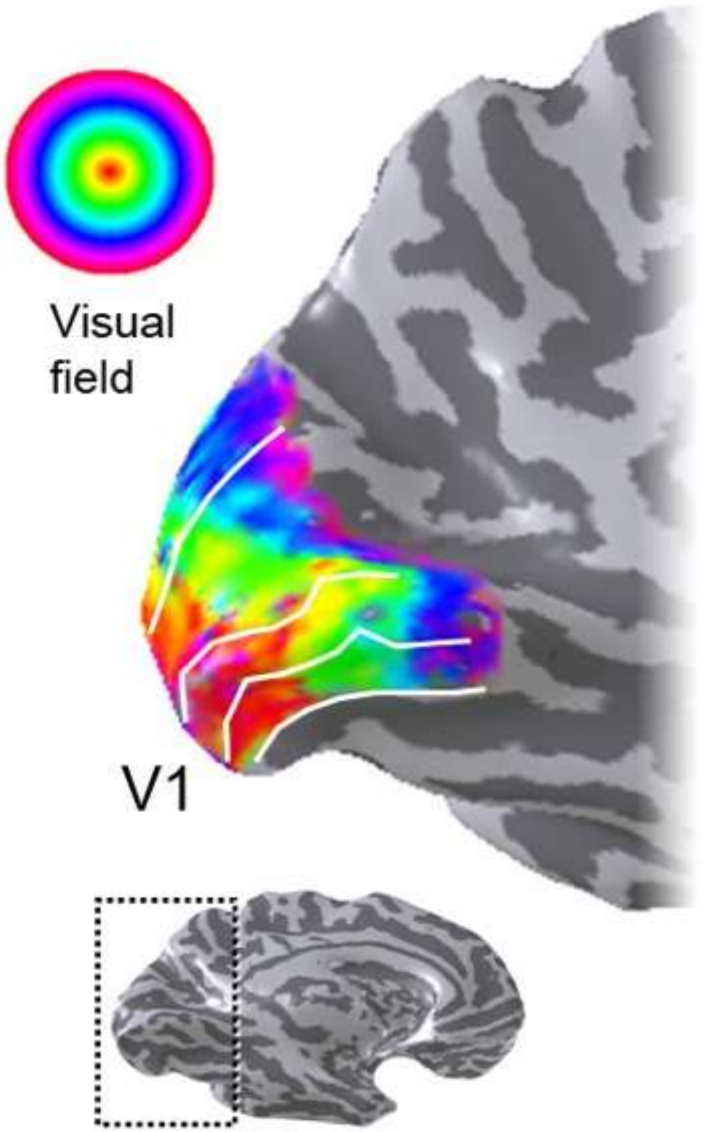
Thus, the RF can be well described by a two-dimensional Gaussian distribution.

Retinotopy

In early visual areas (e.g., V1 to V5), neuron RFs reveal an ordered organization, termed a retinotopic or visuotopic map.

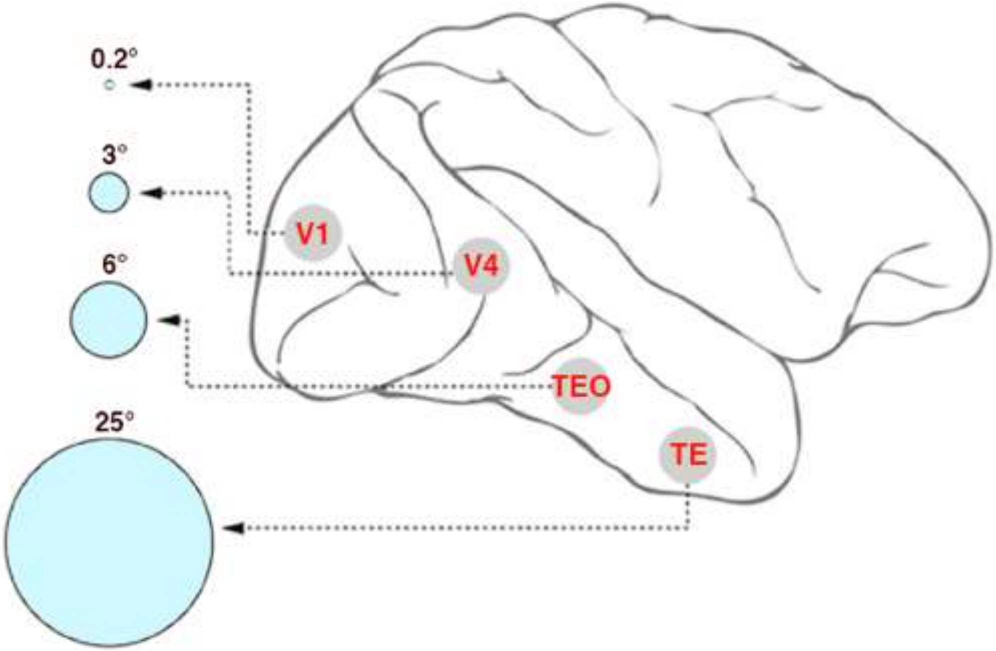
This refers to the existence of a non-random relationship between the position of neurons in the visual areas.

Neuron RFs form a 2D map of the visual field, such that neighbouring regions in the visual image (and therefore on the retinal surface) are represented by adjacent regions of the visual cortical area (i.e., orderly mapping of RF positions in retinotopic coordinates)



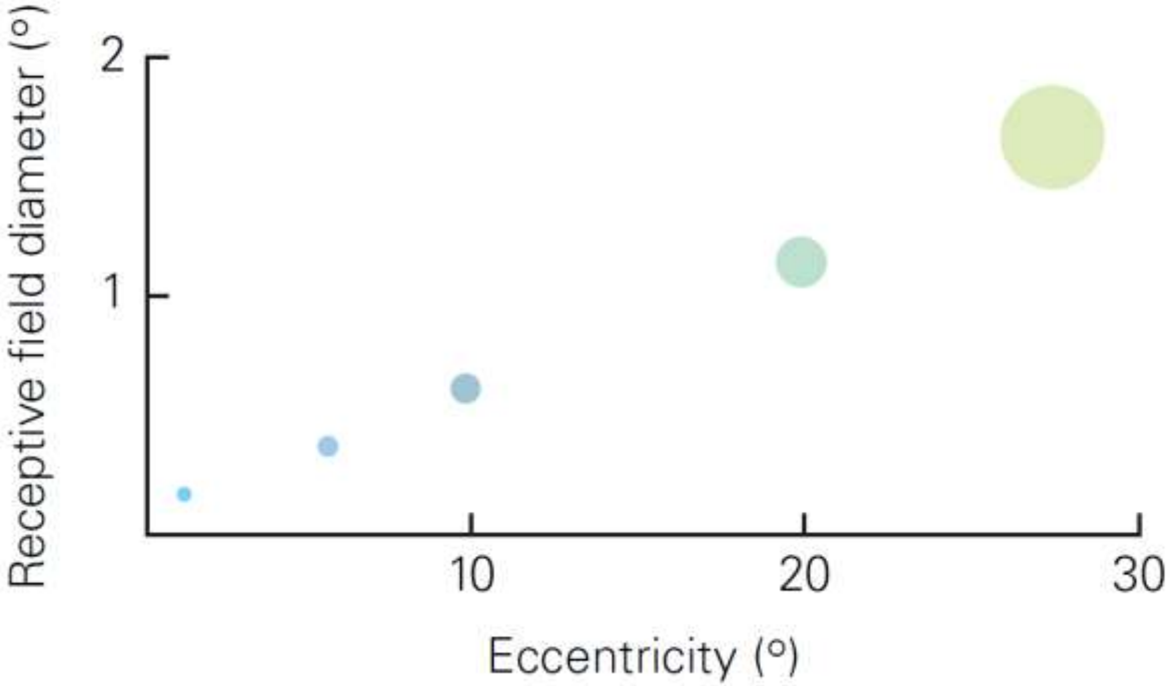
Organization of the RFs the visual system

The size of the RF changes according to the location along the visual system hierarchy



Eccentricity

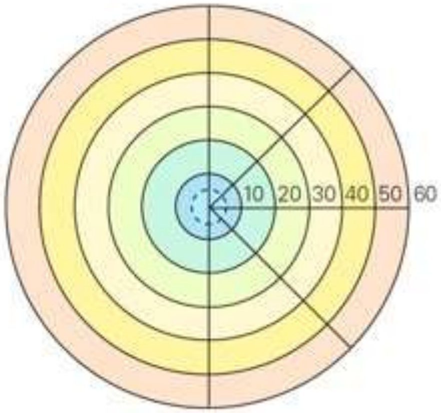
The receptive fields of the retinal ganglion cells that monitor portions of the fovea subtend about 0.1 ° (equal to 6 min of arc), while those in the visual periphery reach up to 1 ° of visual angle or more.



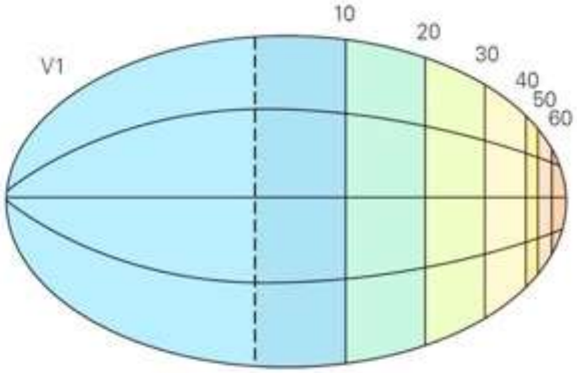
1 Arc min = 1/60 degree

Cortical magnification

The amount of cortical area devoted to each degree of the visual field, known as the magnification factor, varies with eccentricity (i.e., the neural maps of the visual field are not isometric).
In fact, the central part of the visual field controls the largest area of the cortex.
For example, in V1 more cortex is dedicated to the central 10 ° of the visual space than to everything else.



Visual field



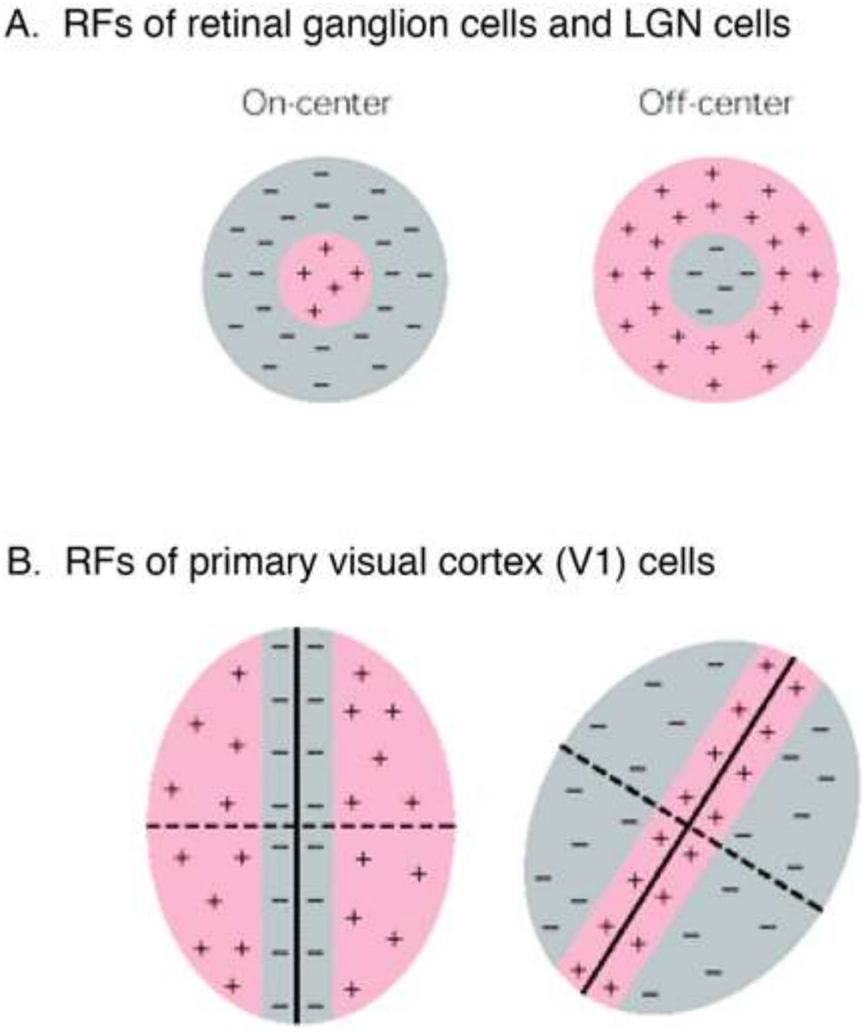
Primary visual cortex

Receptive field properties

Properties change from relay to relay along a visual pathway.

By determining these properties, one can assay the function of each relay nucleus and how visual information is progressively analysed.

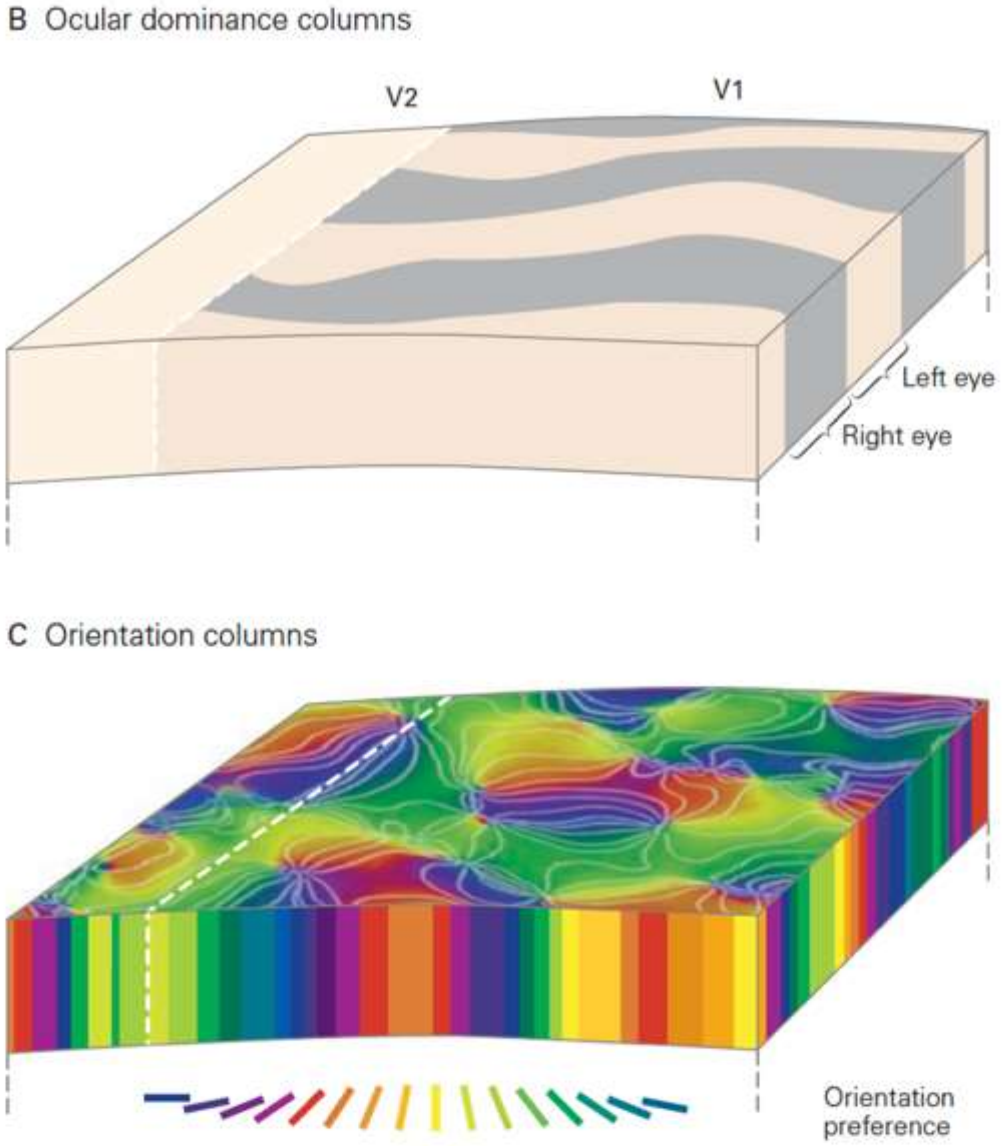
Whereas retinal ganglion cells and neurons in the LGN have concentric center-surround receptive fields, those in the primary visual cortex, although equally sensitive to contrast, analyse oriented contours (orientation selectivity).


A. RFs of retinal ganglion cells and LGN cells
On-center
Off-center
B. RFs of primary visual cortex (V1) cells

Cortical columns

In V1, neurons with similar functional properties are found close together in columns or functional modules (about 50-75μm in diameter) that extend from the cortical surface to the white matter (about 2mm high).

V1 includes specific columns for stimulus orientation, and ocular dominance.



Orientation columns

Neurons with the same orientation preference (i.e., vertical) are grouped together into orientation columns. Each column contains a few hundred cells and is 50-75μm wide;

Moving from one column to the adjacent one, orientation preference changes systematically by 10-15 °, both clockwise and counterclockwise, completing a 180 ° cycle (12 steps) every 750-1000μm.

The set of columns corresponding to a complete sequence of orientations (a period) is called a hypercolumn.

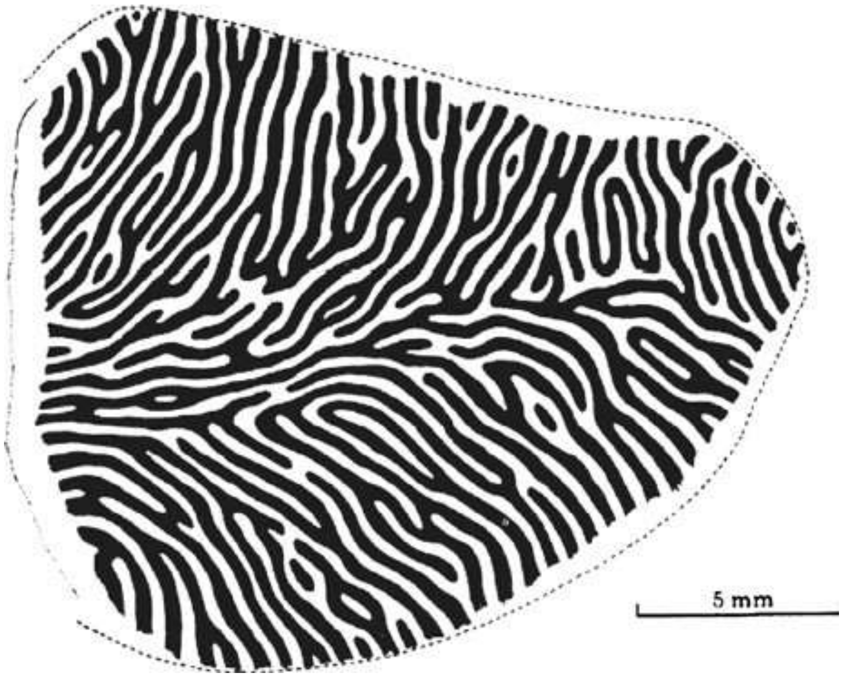
There are approximately 3-4 thousand hypercolumns, each monitoring a position of the visual field, in accordance with the retinotopic topology.

Ocular dominance columns

The ocular dominance columns group neurons that respond more vigorously to stimuli presented to one of the two eyes. They are stripes with an average width of approximately 750 µm, running tangentially for various mm.

The ocular dominance columns reflect the segregation of inputs from different layers of the LGN, which receive inputs from retinal ganglion cells located in the ipsilateral or contralateral retina.

In tangential penetrations of V1, the dominance columns of the left and right eye have been found to alternate regularly with a periodicity of 750 to 1,000µm.



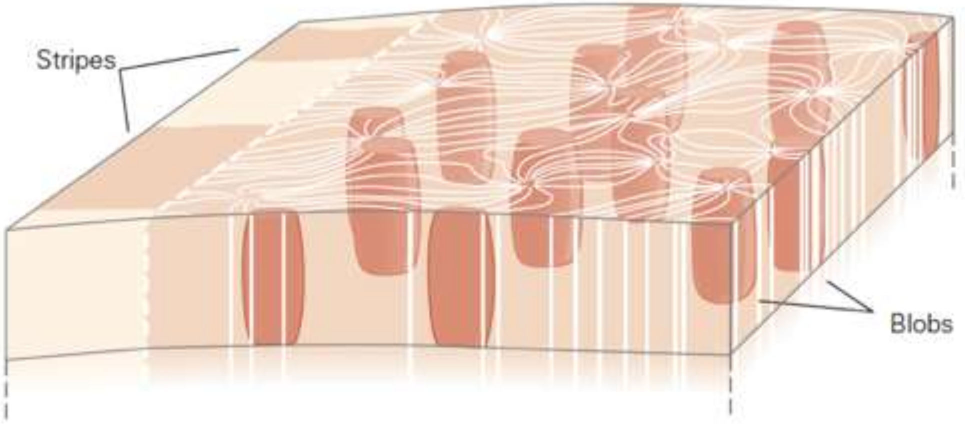
Ocular dominance columns in primary visual cortex (V1) of macaque monkey shown in tangential section. Regions receiving input from one eye are shaded black and regions receiving input from the other eye are unshaded. The dashed line signifies the border between areas V1 and V2 (taken from Hubel and Wiesel, 1977).

Blob e interblob

The columns of orientation and ocular dominance include groups of neurons that are poorly selective for orientation (they have circular receptors) but with strong preferences for the color of the stimulus.

These cell groups are located in the superficial layers (II and III) of V1. They are detectable by a specific marker for the cytochrome oxidase (CO) enzyme, which distributes in a regular pattern of regions defined as blobs (CO rich and color responsive) separated by interblob areas (CO poor and orientation responsive) .

D Blobs, interblobs (V1), and stripes (V2)

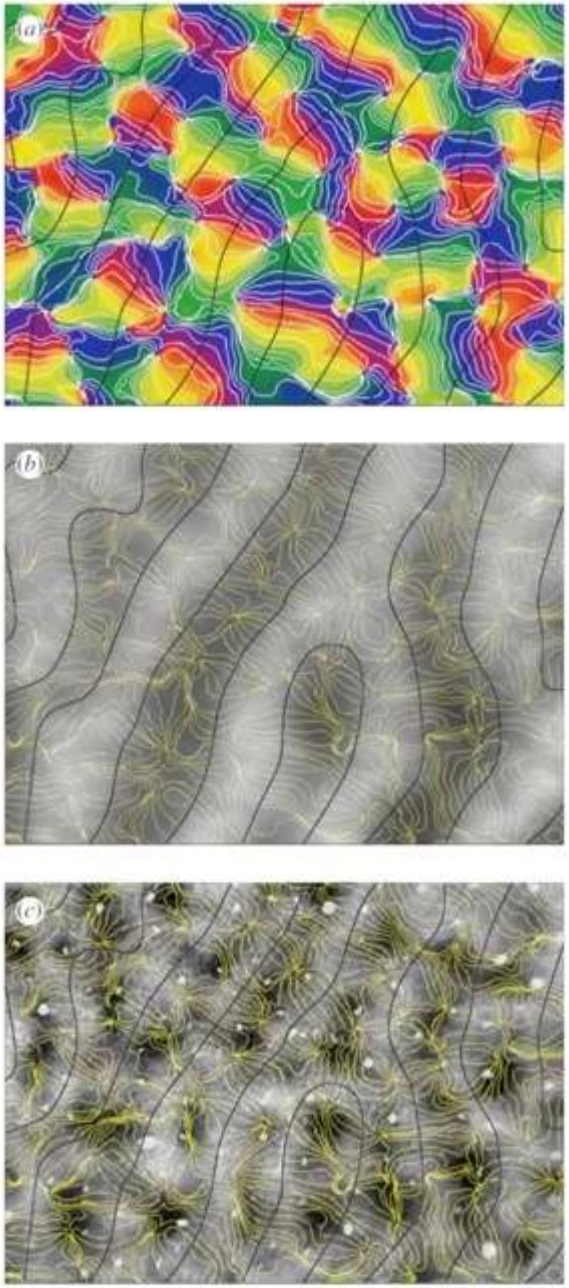


Pinwheels

More recently, optical imaging technique has enabled to visualize a surface representation of the orientation and ocular dominance columns in living animals.

The cycles of orientation columns form various structures, from parallel stripes to pinwheel-like shapes.

Sharp jumps or singularities (discontinuities) in orientation preference occur at the pinwheel centers and cause “fractures” in the orientation map.

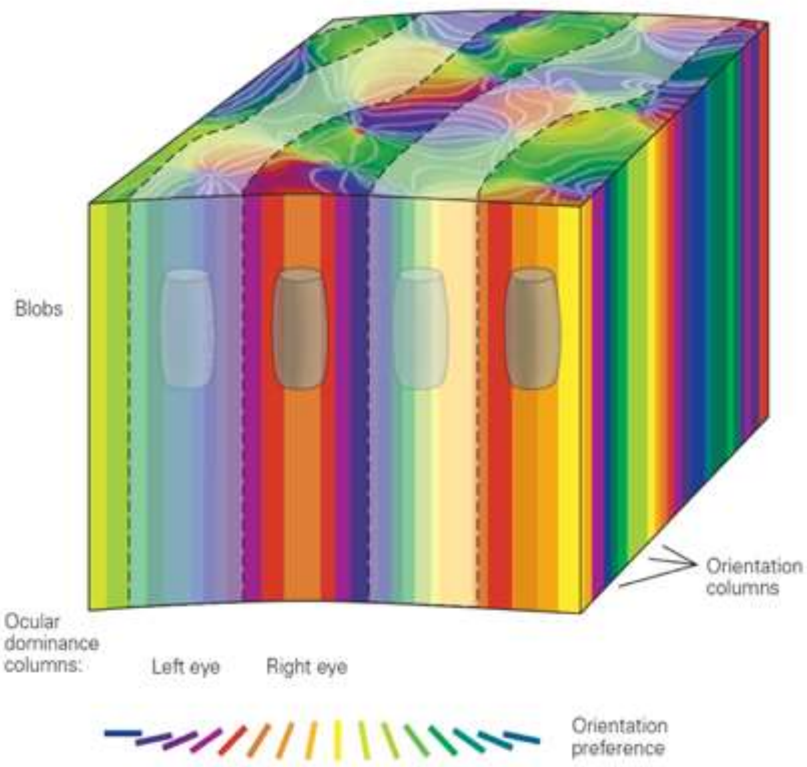


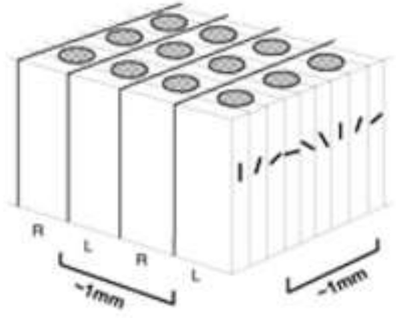
Orientation columns

Ocular dominance columns

Blobs

Ice cube model (Hubel e Wiesel, 1977)



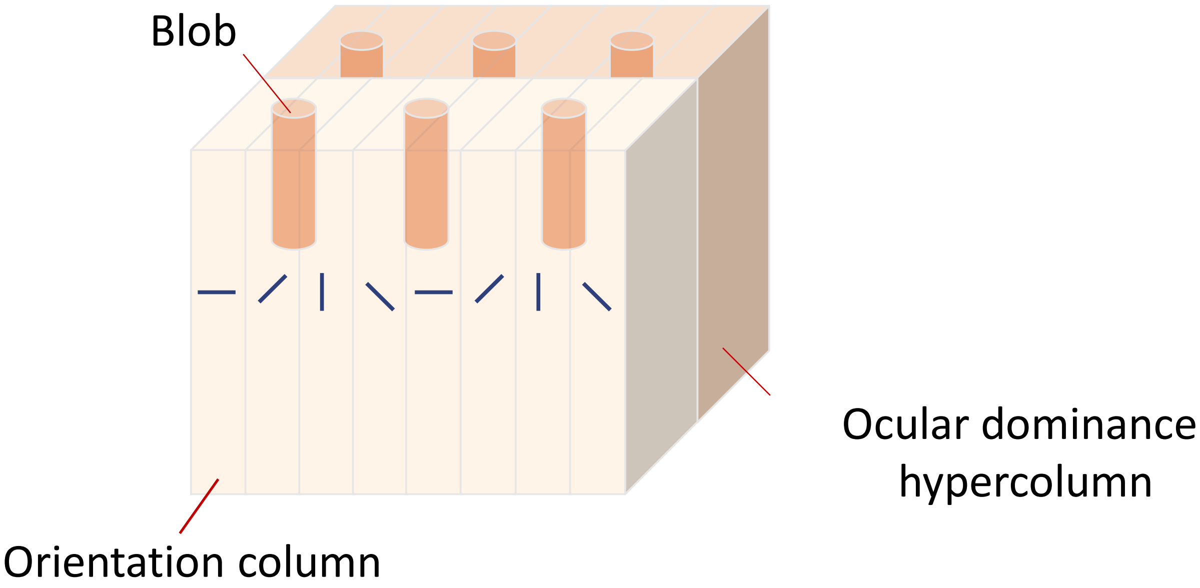


A region of cortical tissue of about 1mm contains two orientation hypercolumns (a complete cycle of selective vertical columns for orientation), one for the left eye and one for the right that alternate regularly, blob and interblob. This computational module contains all the anatomical-functional types of V1 neurons, and would be repeated thousands of times to cover the entire surface of the visual field.

The functional organization of the primary visual cortex is therefore based on two systems running orthogonally to each other:

↓ orientation system

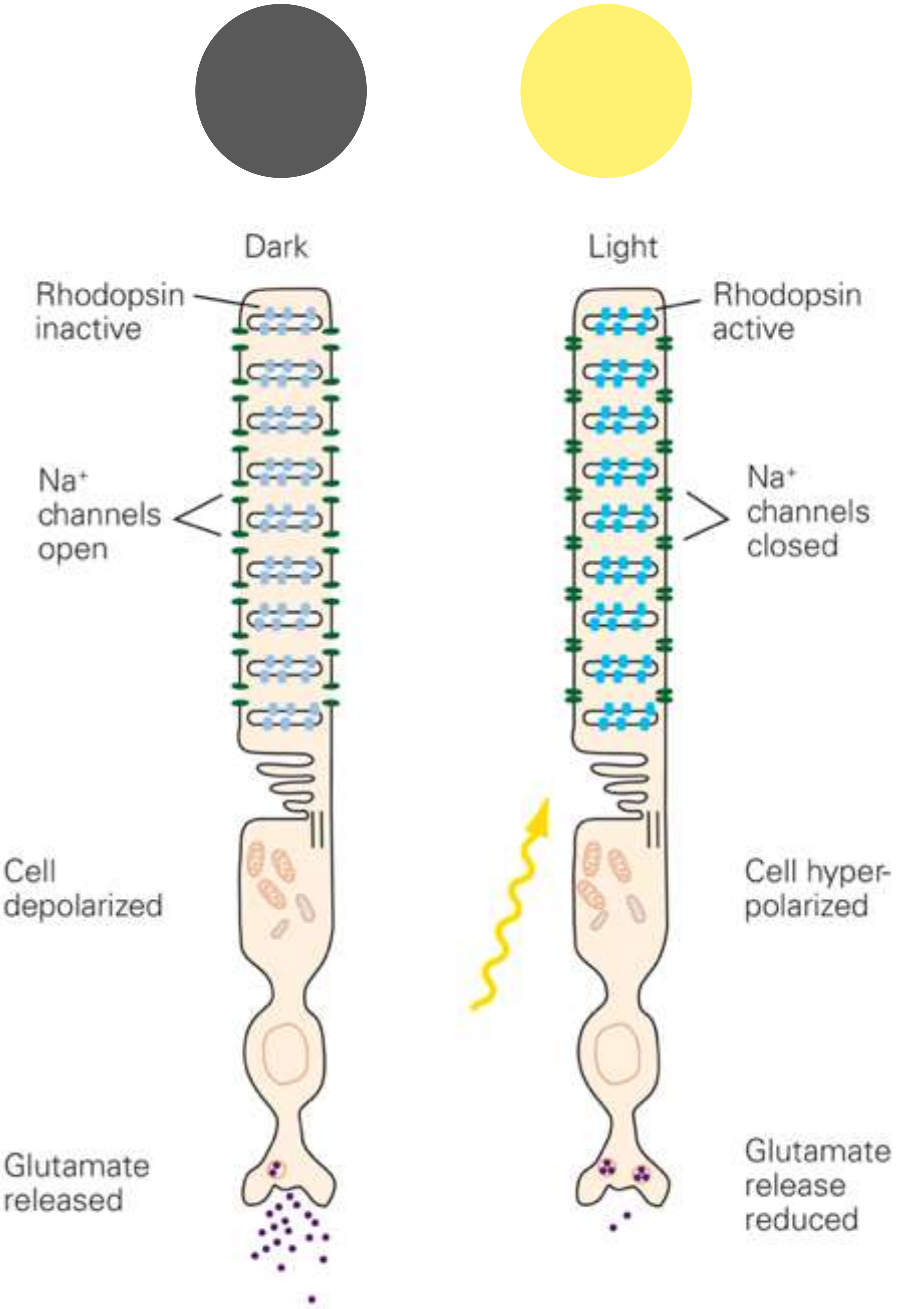
→ ocular dominance system

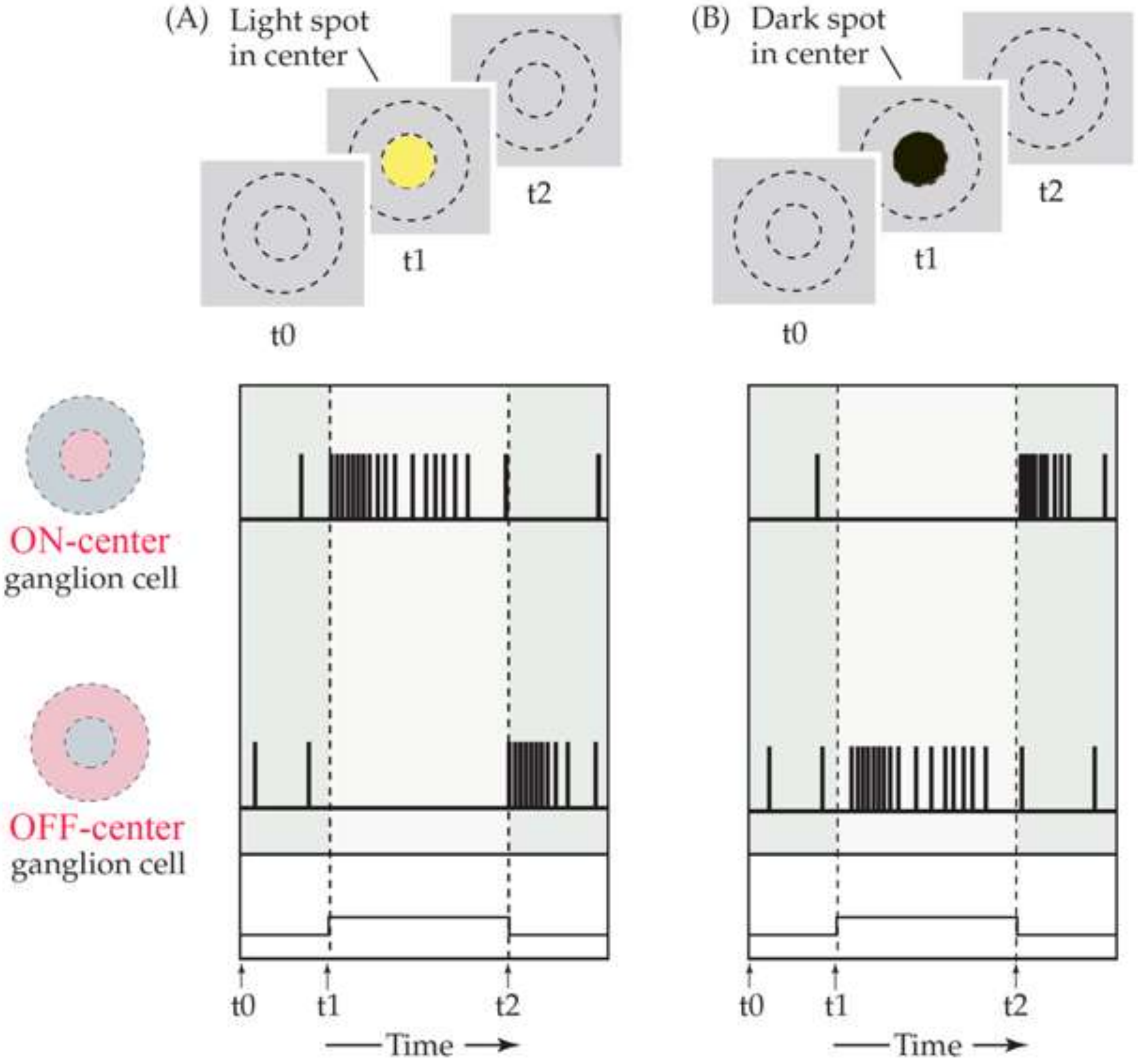


Single neurons in the visual system

Photoreceptors produce a relatively simple neural representation of the visual scene:

Neurons in the bright regions are hyperpolarized, while those in the dark regions are depolarized.





RGCs have concentric circular RFs and fall into one of two categories:

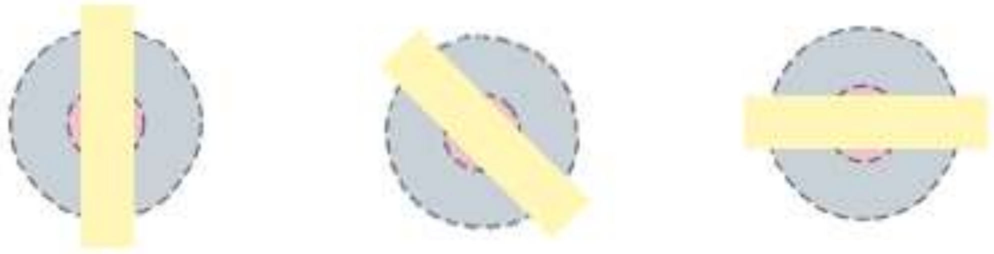
ON-center
OFF-center

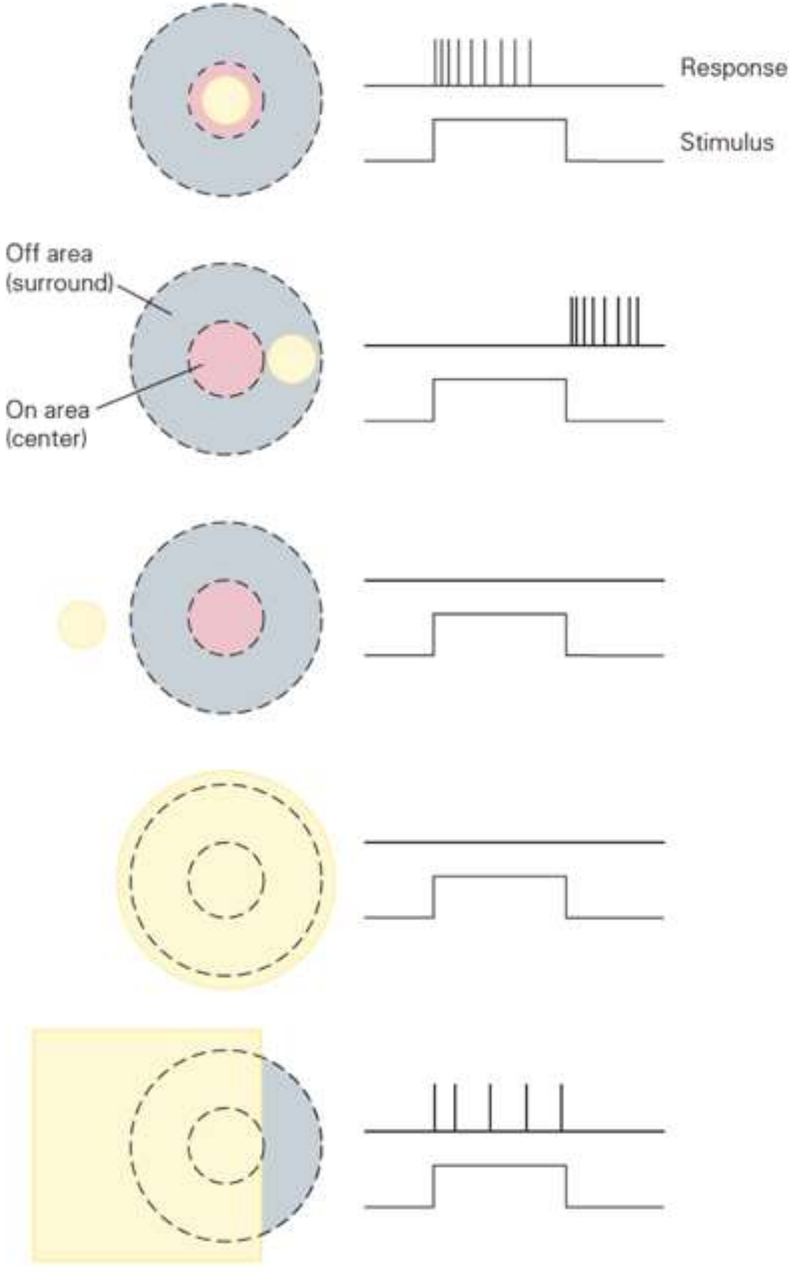
The ON-center cells discharge in response to the increase in brightness in the center of the receptive field, while the OFF-center cells discharge in response to the reduction of brightness in the center of the receptive field.

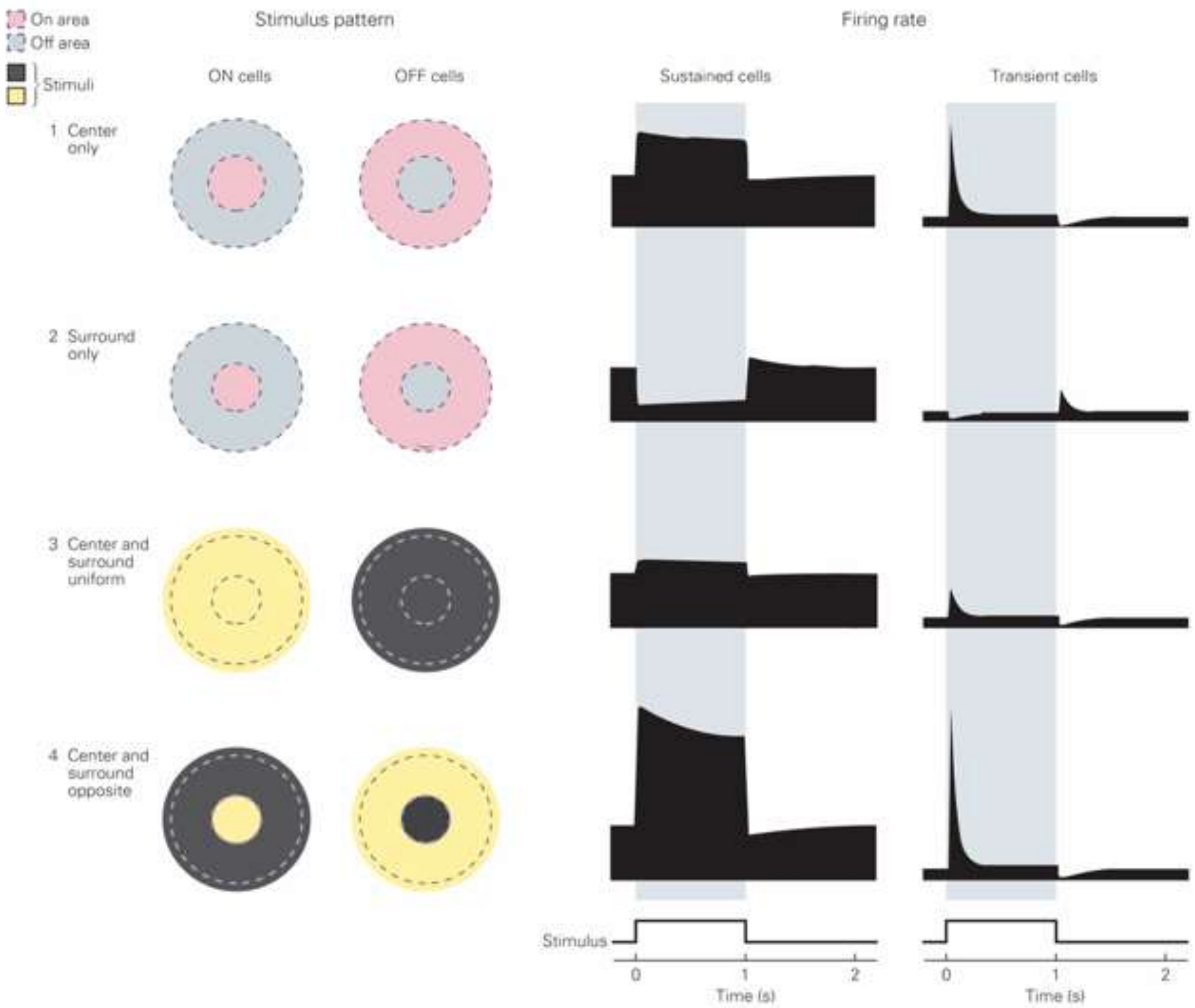
Center and surround reveal opposite response (lateral inhibition) mutually antagonistic

A uniform stimulus that activates the center and surround simultaneously causes a weak or no response.

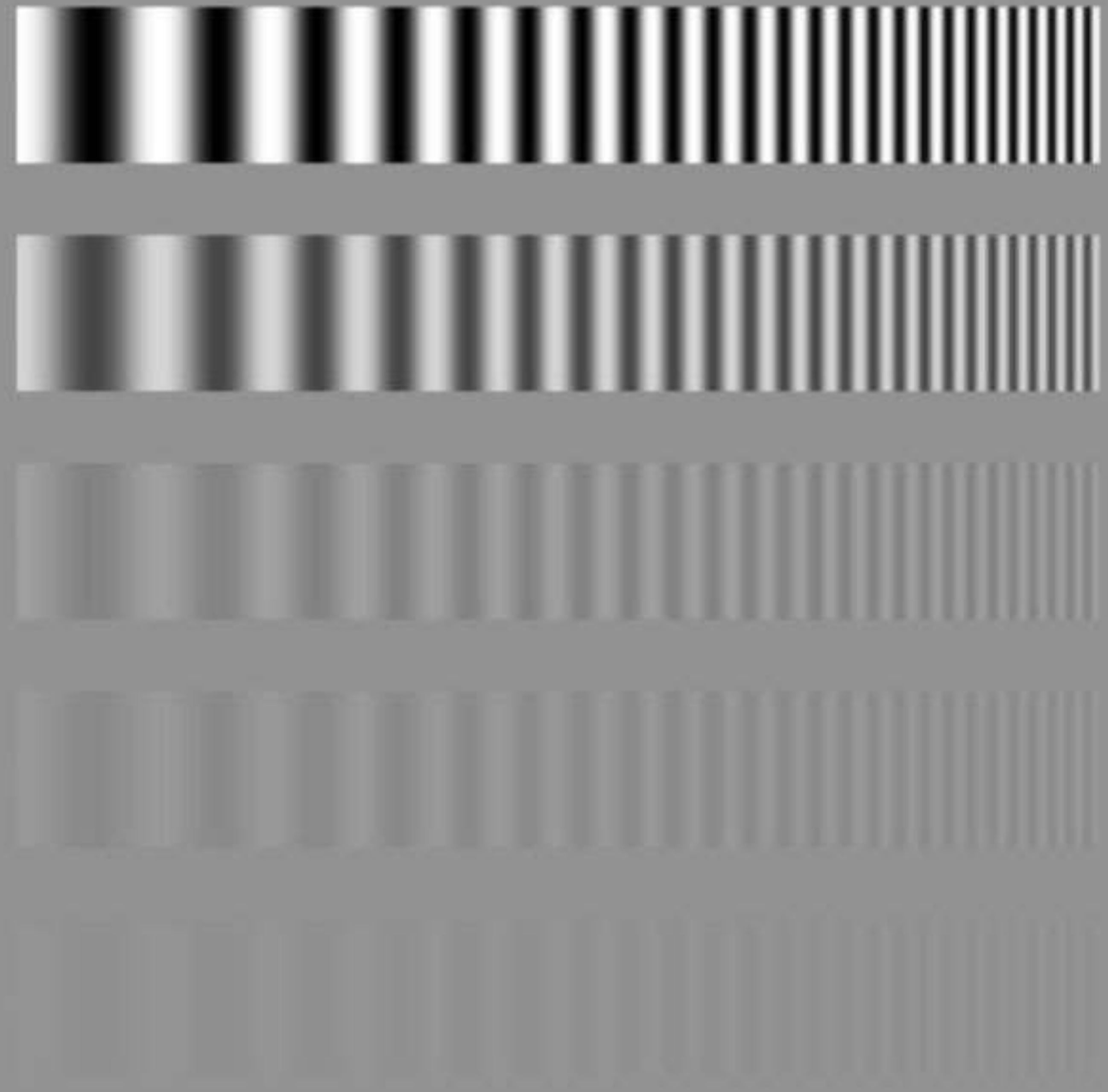
Note that ganglion cells are not selective for the orientation of lines or edges.



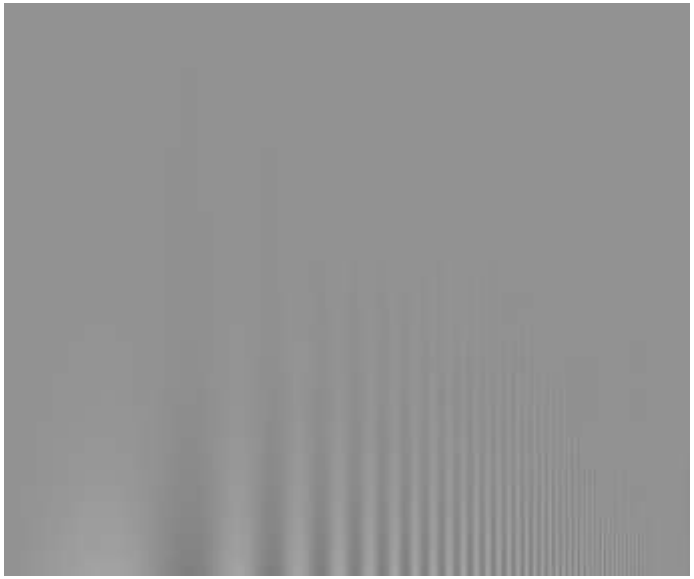


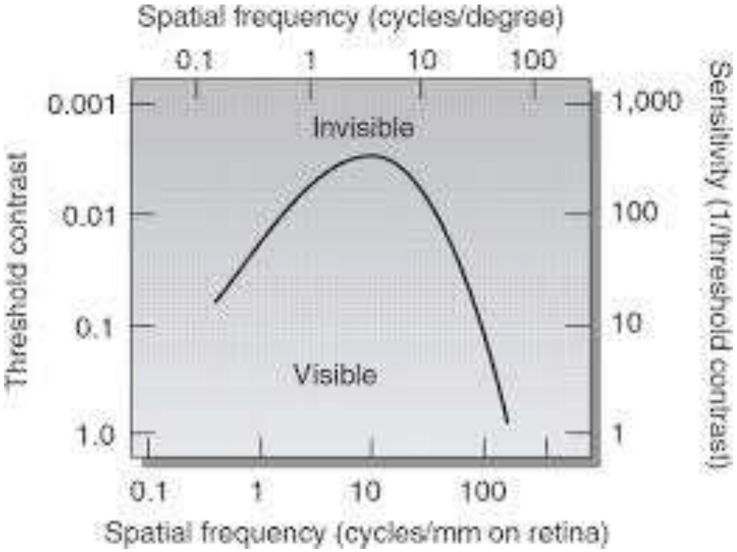


Retinal ganglion cells respond weakly to uniform illumination but strongly emphasize brightness contrasts, as these carry the most reliable visual information, unlike absolute light intensity which varies with the illumination source.



Contrast perception as a function of spatial frequency

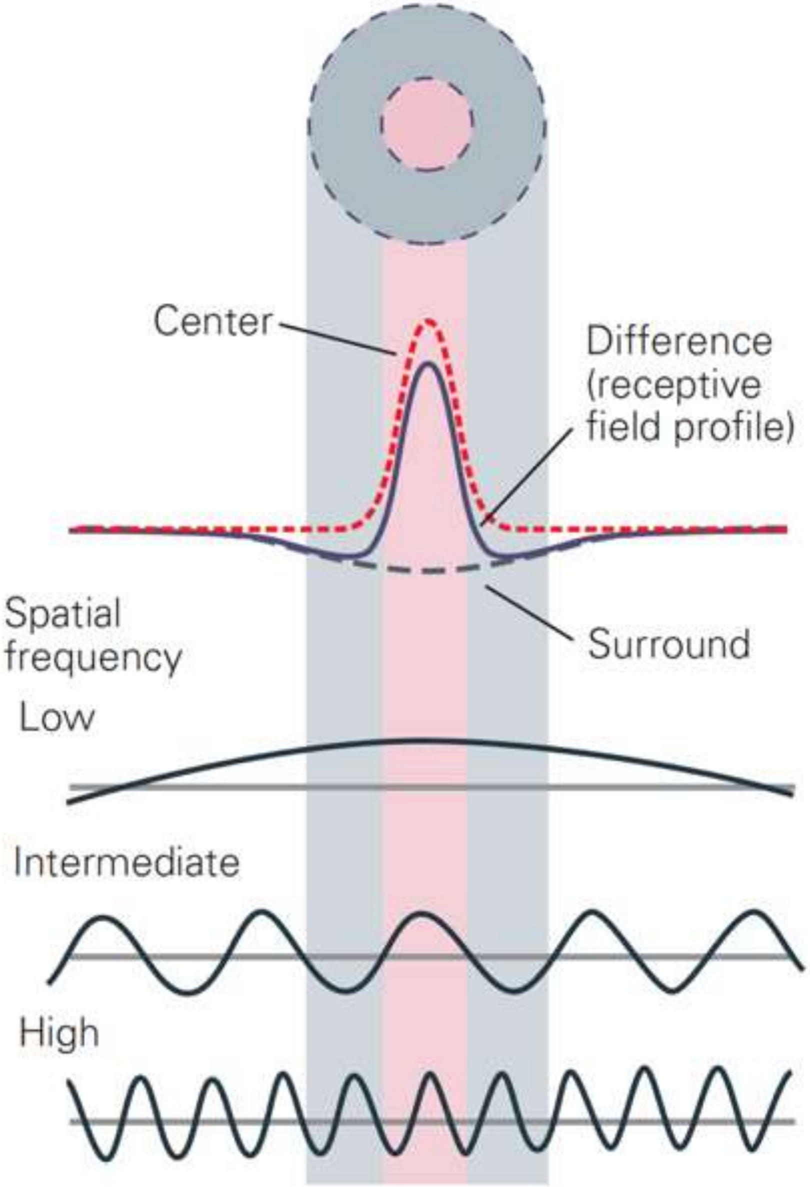


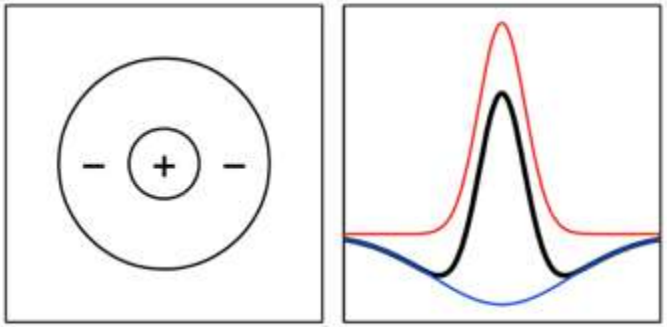

Spatial frequency (cycles/degree)
0.1
1
10
100
0.001
0.01
0.1
1.0
Threshold contrast
Invisible
Visible
1,000
100
10
1
Sensitivity (1/threshold contrast)
0.1
1
10
100
Spatial frequency (cycles/mm on retina)

The contrast sensitivity function (CSF) describes an observer's sensitivity to sinusoidal gratings as a function of their spatial frequency.

This is measured using a contrast detection experiment in which the minimum (threshold) contrast required to detect sinusoidal gratings of various spatial frequencies is determined.

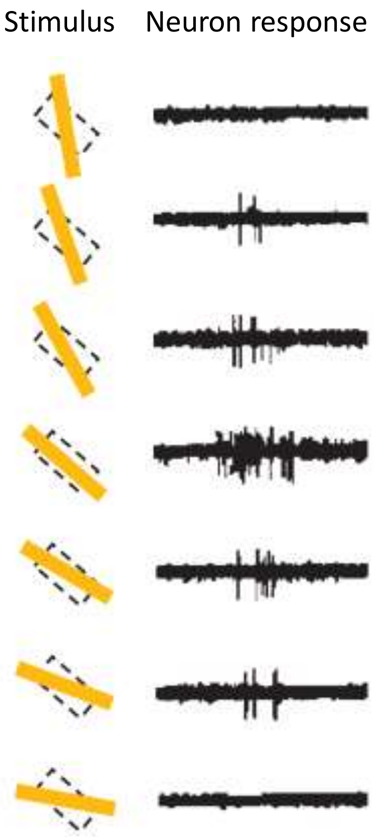
Sensitivity is defined as 1 / (threshold contrast) (so if the threshold is low, the sensitivity is high).

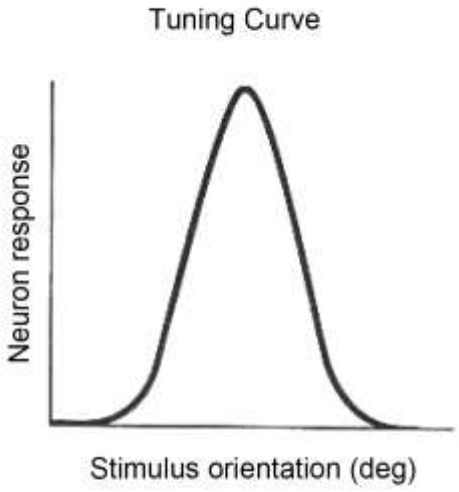


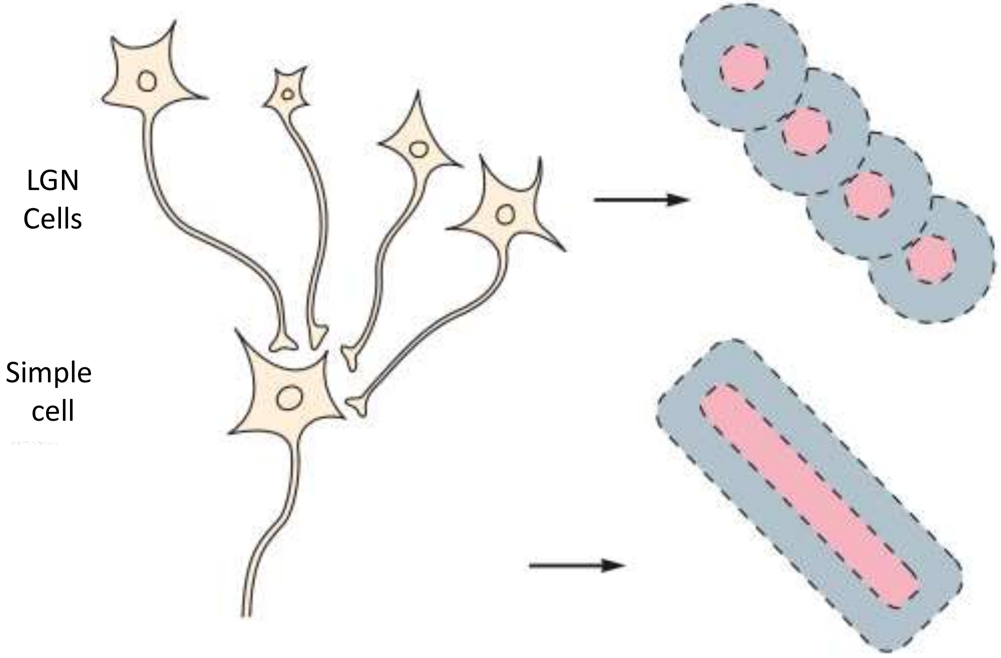


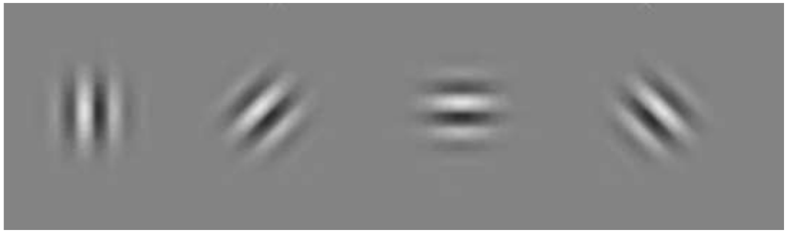
The response of the center-surround concentric RF is modeled as the difference of two Gaussian curves, a positive with a narrow base corresponding to the center of the receptive field, and a negative with a wide base, corresponding to the center and surround.

V1 – Simple cells



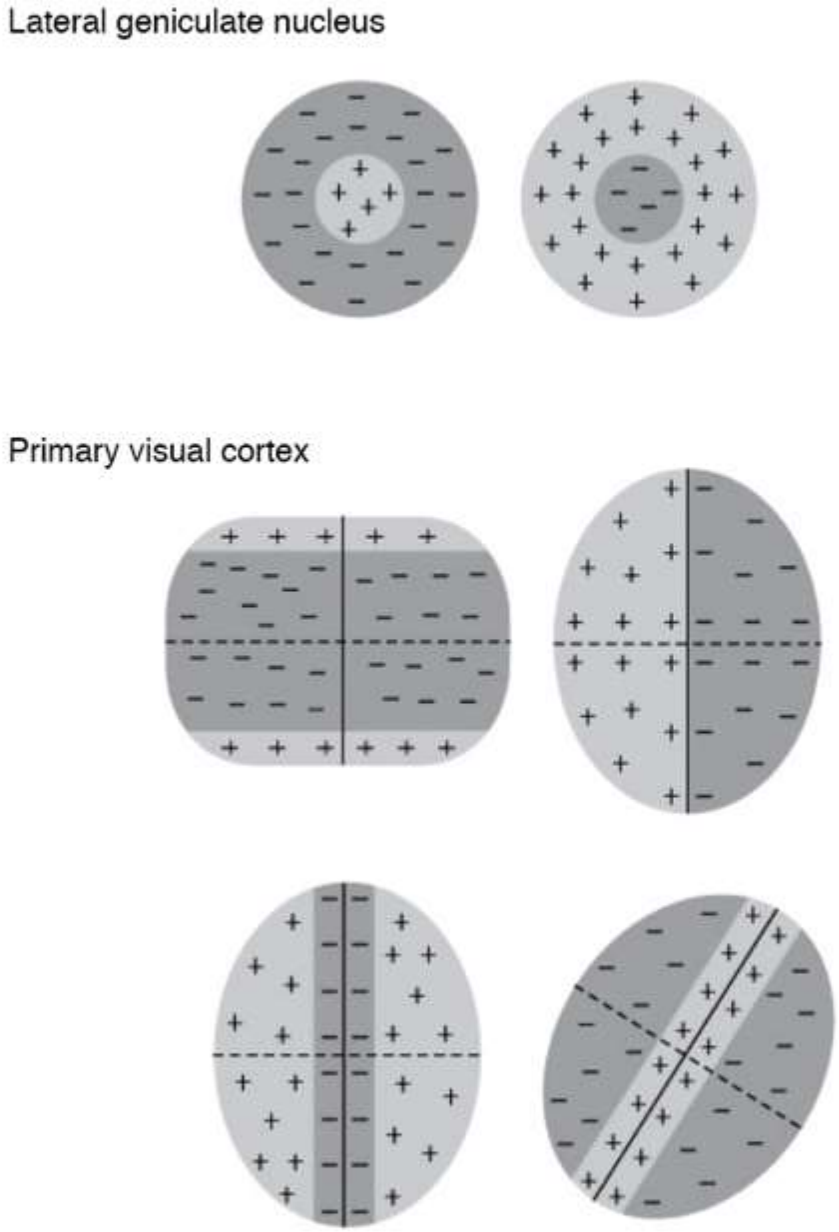






Oriented sinusoidal gratings

In V1, neurons selectively respond to oriented bars or gratings.
In simple cells, the receptive fields have separate ON and OFF areas.



Lateral geniculate nucleus

Primary visual cortex

The receptive fields of the simple cells of the primary visual cortex are different and less homogeneous than those of the ganglion cells of the retina and the LGN

Complex cells

Have rectangular receptive fields, larger than those of simple cells;

Respond to linear stimuli with specific orientation;

The position of the stimulus within the receptive field is not critical as the demarcation between on and off zones is not so clear;

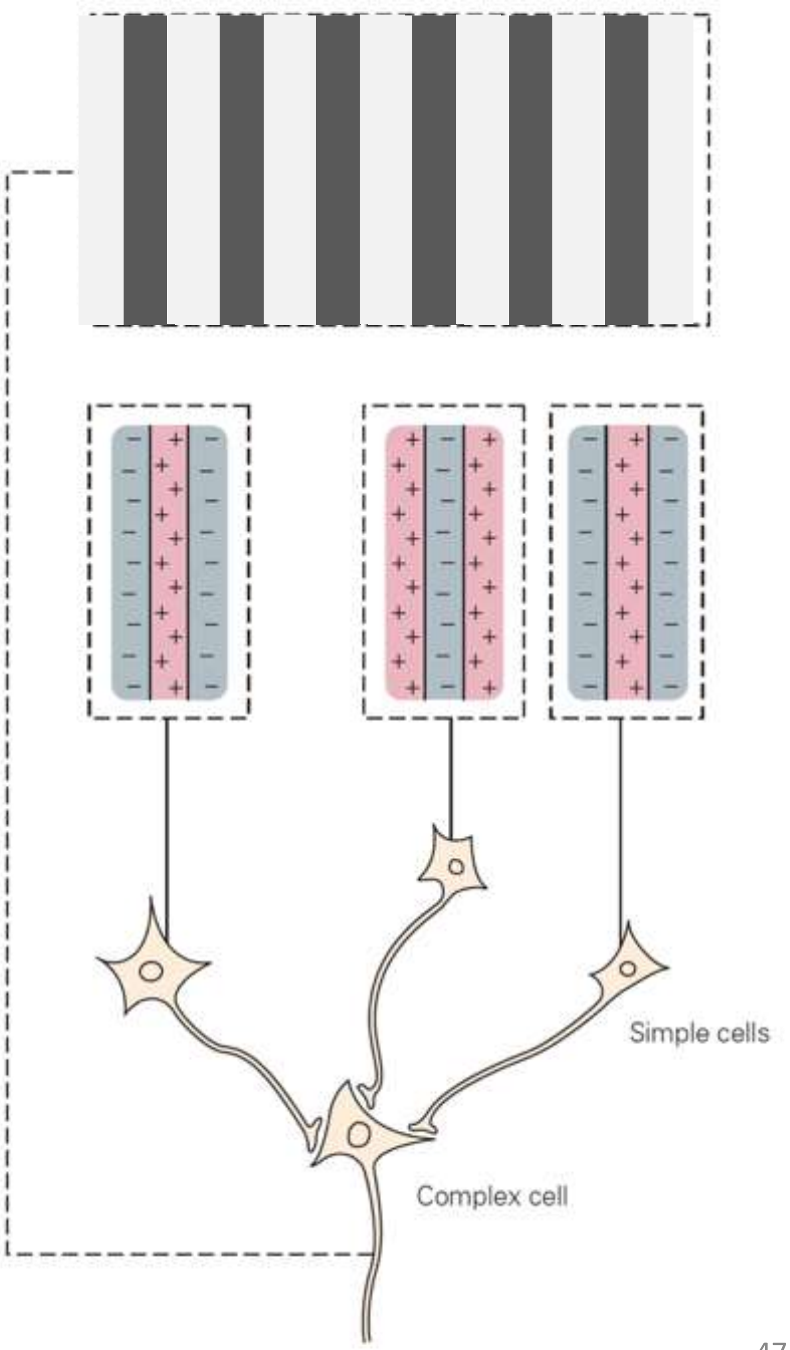
Movement of the stimulus in the receptive field is particularly effective in activating the cells;

Complex cells selectively respond to stimuli that move in particular directions;

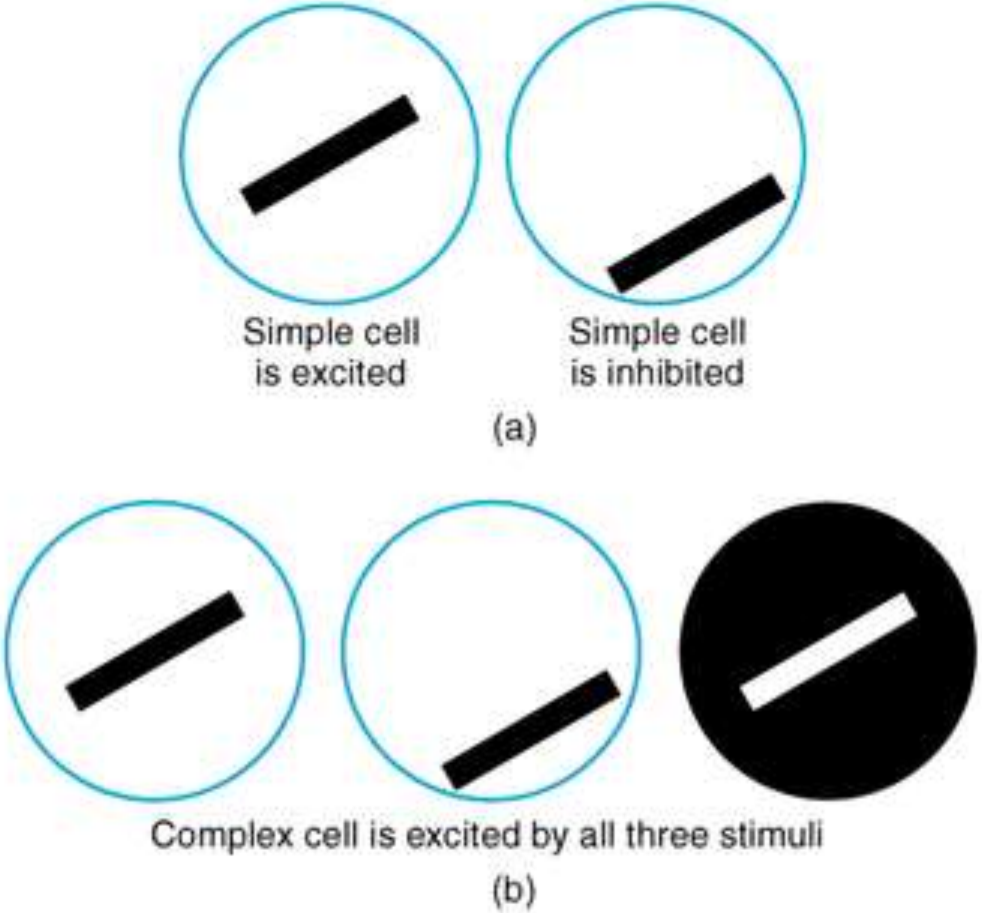
V1 – Complex cells

In complex cells, the ON and OFF regions are superimposed, i.e. each position in the receptive field responds to both white and black bars, and the cells respond when a line or edge crosses the receptive field along an axis perpendicular to the orientation of the receptive field.

This constancy in the response to variations of stimulus location in the RF is commonly called position invariance.



► Response Characteristics of Neurons to Orientation in the Primary Visual Cortex



Simple cell is excited

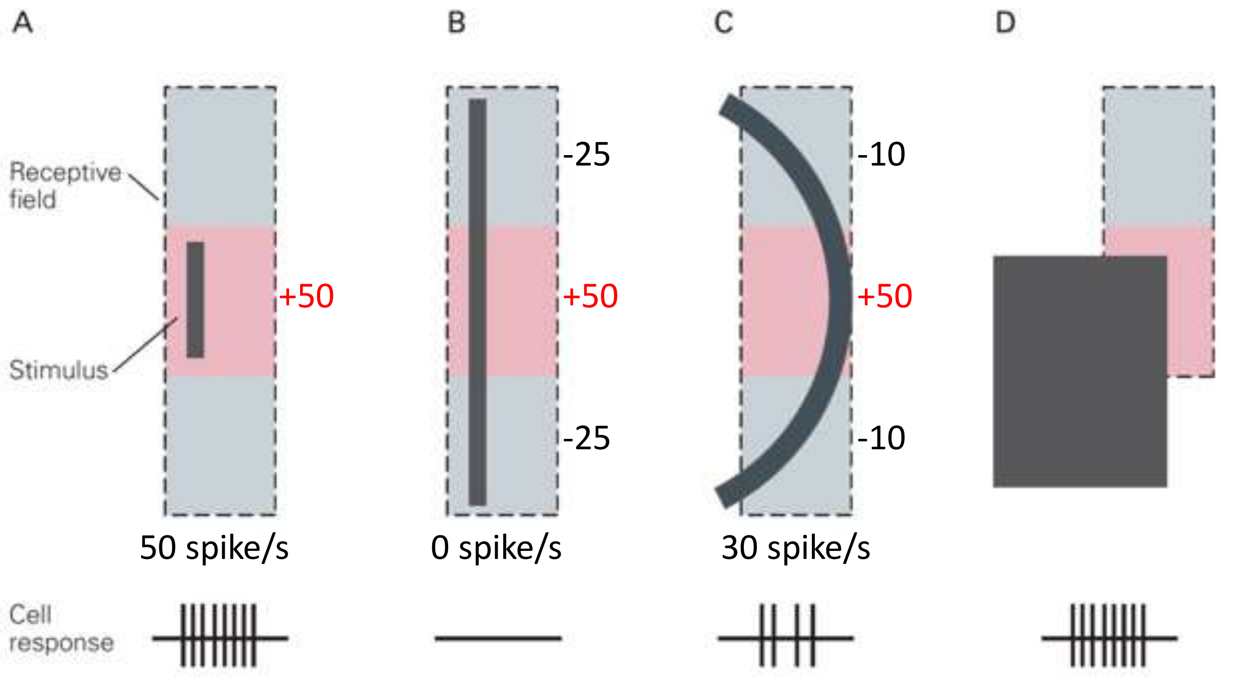
Simple cell is inhibited

(a)

Complex cell is excited by all three stimuli

(b)

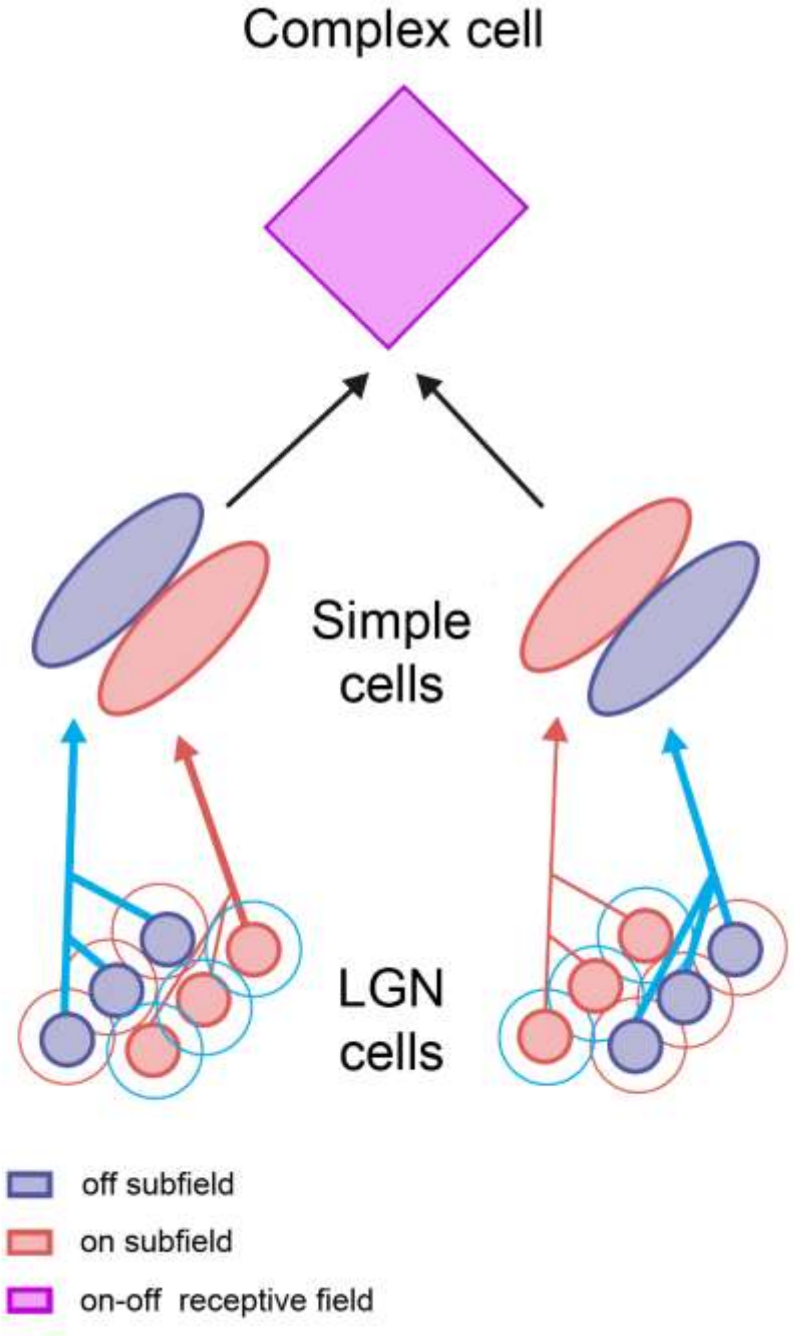
End-stopped cells



The receptive fields of some cells have a central excitatory region flanked by inhibitory regions that have the same preferential orientation.

A short linear segment (A) or a long curved line (C) will be effective in activating the neuron, but not a long straight line (B).

A neuron with a receptive field that has only one inhibitory region in addition to the excitatory region can signal the presence of angles (D).



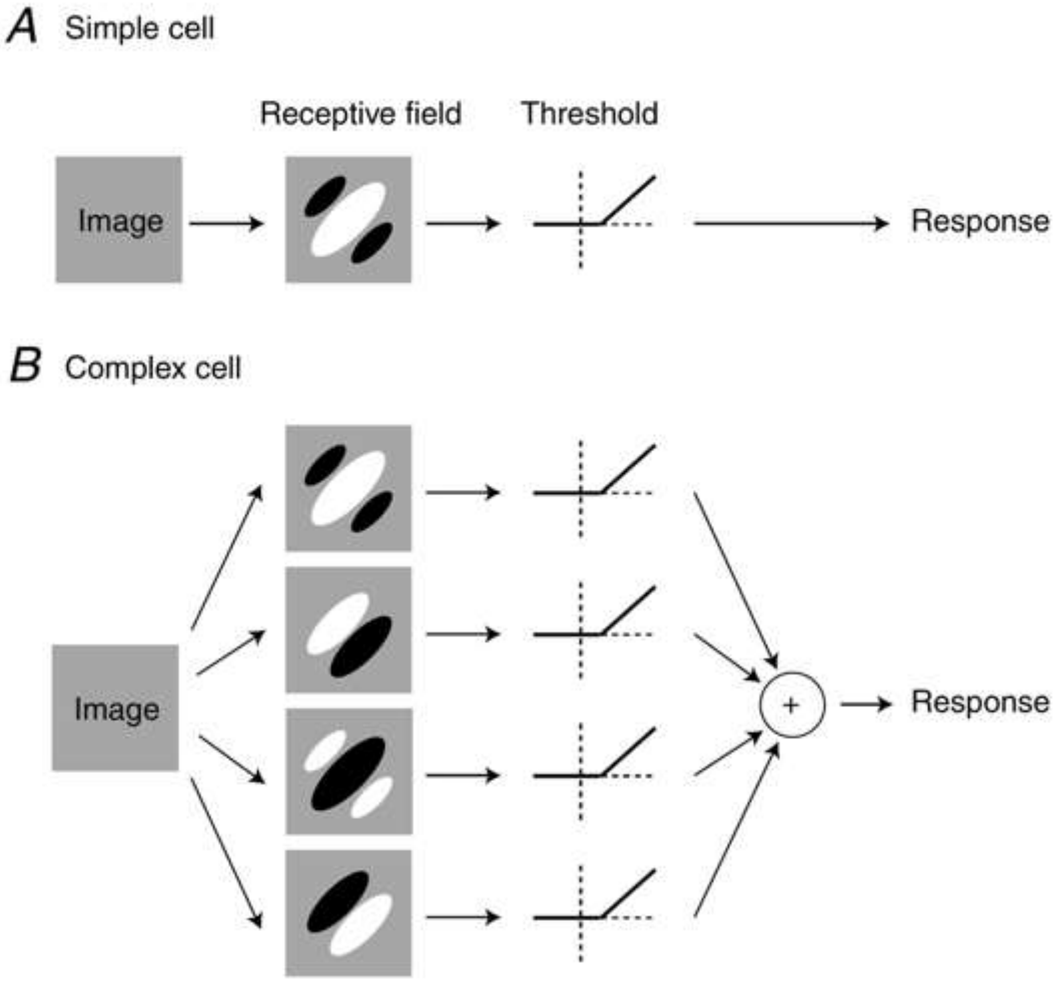
Hierarchical model of the receptive field

The processing of some characteristics of the visual images is performed through a progressive convergence of information within the visual system.

Each RF at one level of the visual processing hierarchy emerges from the convergence of inputs from many neurons of the previous level.

The size of the RF increases along the visual hierarchy.

Hubel, Wiesel, J. Neurophysiol. 1962

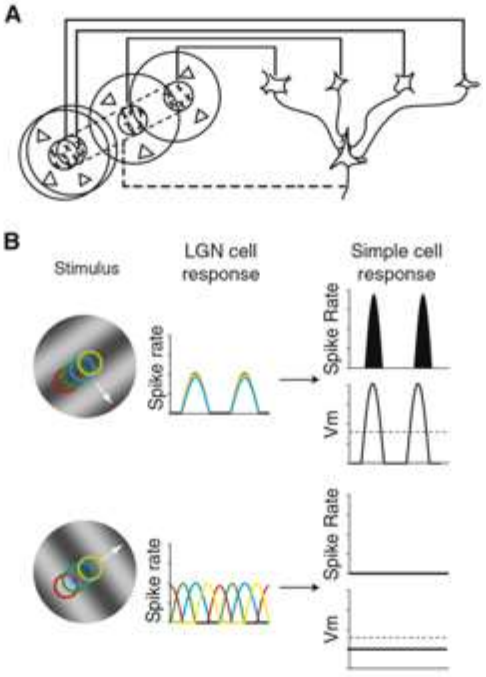


The models of simple and complex cells proposed by Movshon, Thompson and Tolhurst (Movshon et al. 1978)

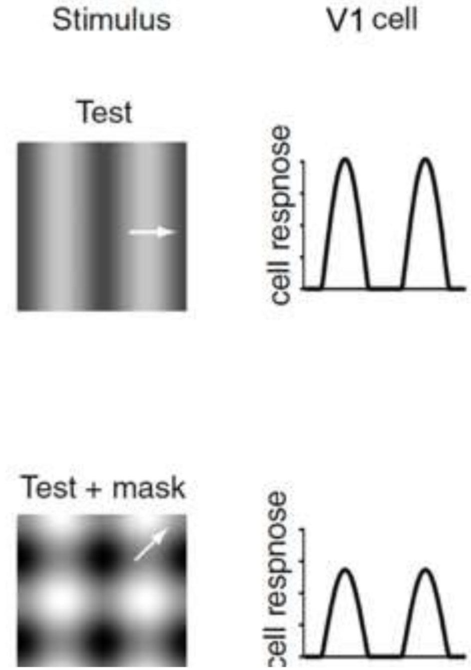
A, simple cells. The first stage is linear filtering, i. e. a weighted sum of the image intensities, with weights given by the receptive field. The second stage is rectification: only the part of the responses that is larger than a threshold is seen in the firing rate response.

B, complex cells. The first stage is linear filtering by a number of receptive fields such as those of simple cells (here we show four of them with spatial phases offset by 90 deg). The subsequent stages involve rectification, and then summation.

The feedforward model of orientation selectivity in V1



The feedforward model as originally proposed by Hubel and Wiesel (1962). Four cells from the LGN, whose receptive fields are shown to the left, synapse onto a V1 simple cell. The simple cell derives its preferred orientation from the axis of alignment of these cell receptive fields.

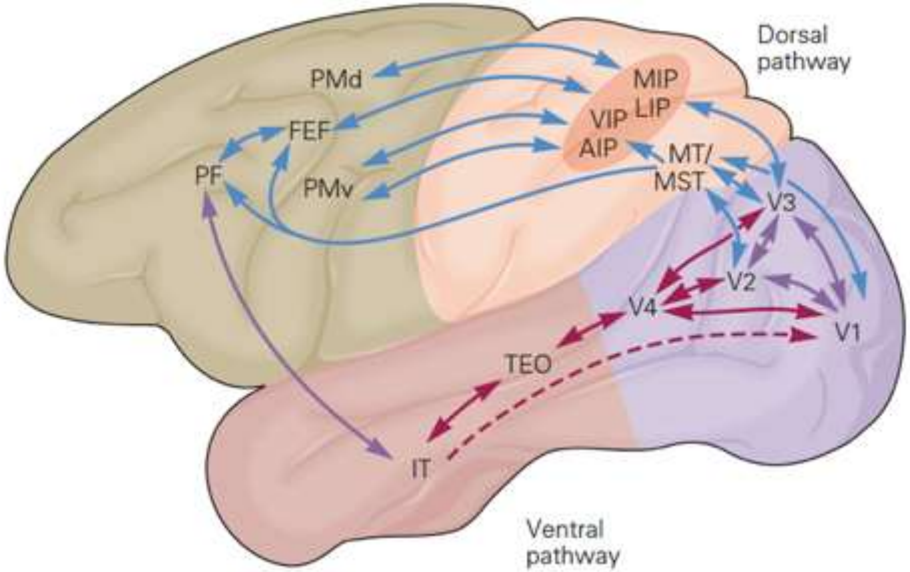


Cross-orientation suppression cobtradicts a purely feedforfward model of orientation swlectivity in V1

Beyond V1 are the extrastriate visual areas (more than 30 areas in macaques), a set of higher-order visual areas organized as neural maps of the visual field.

Visual areas are organized in two hierarchical pathways, a ventral pathway involved in object recognition and
a dorsal pathway dedicated to the use of visual information for guiding movements.

The ventral or object recognition pathway extends from V1 to the temporal lobe
The dorsal or movement-guidance pathway connects V1 with the parietal lobe and then with the frontal lobes.

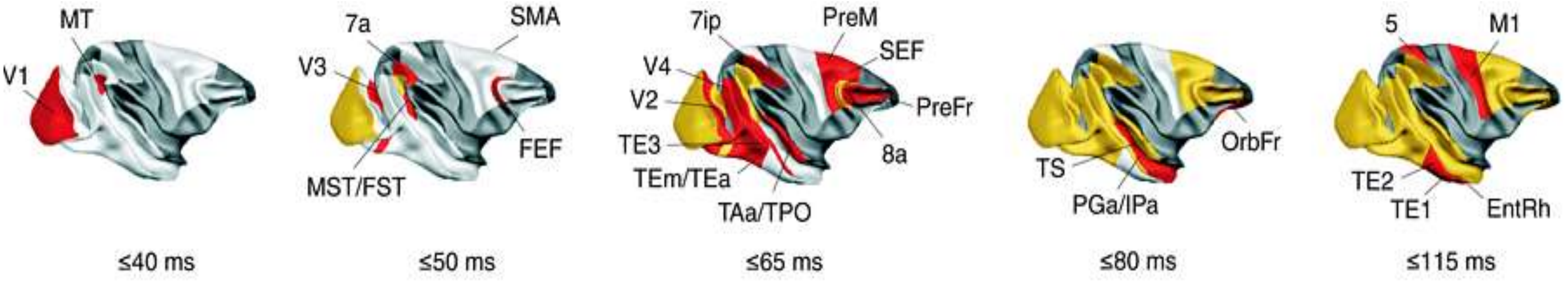


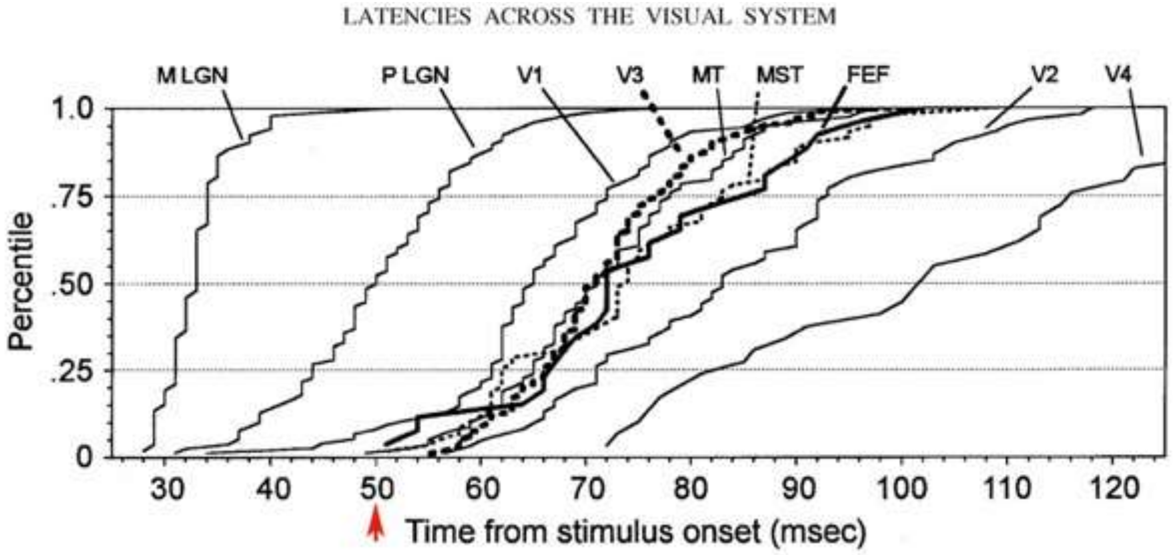
Ungerleider & Mishkin, Two cortical visual systems. 1982

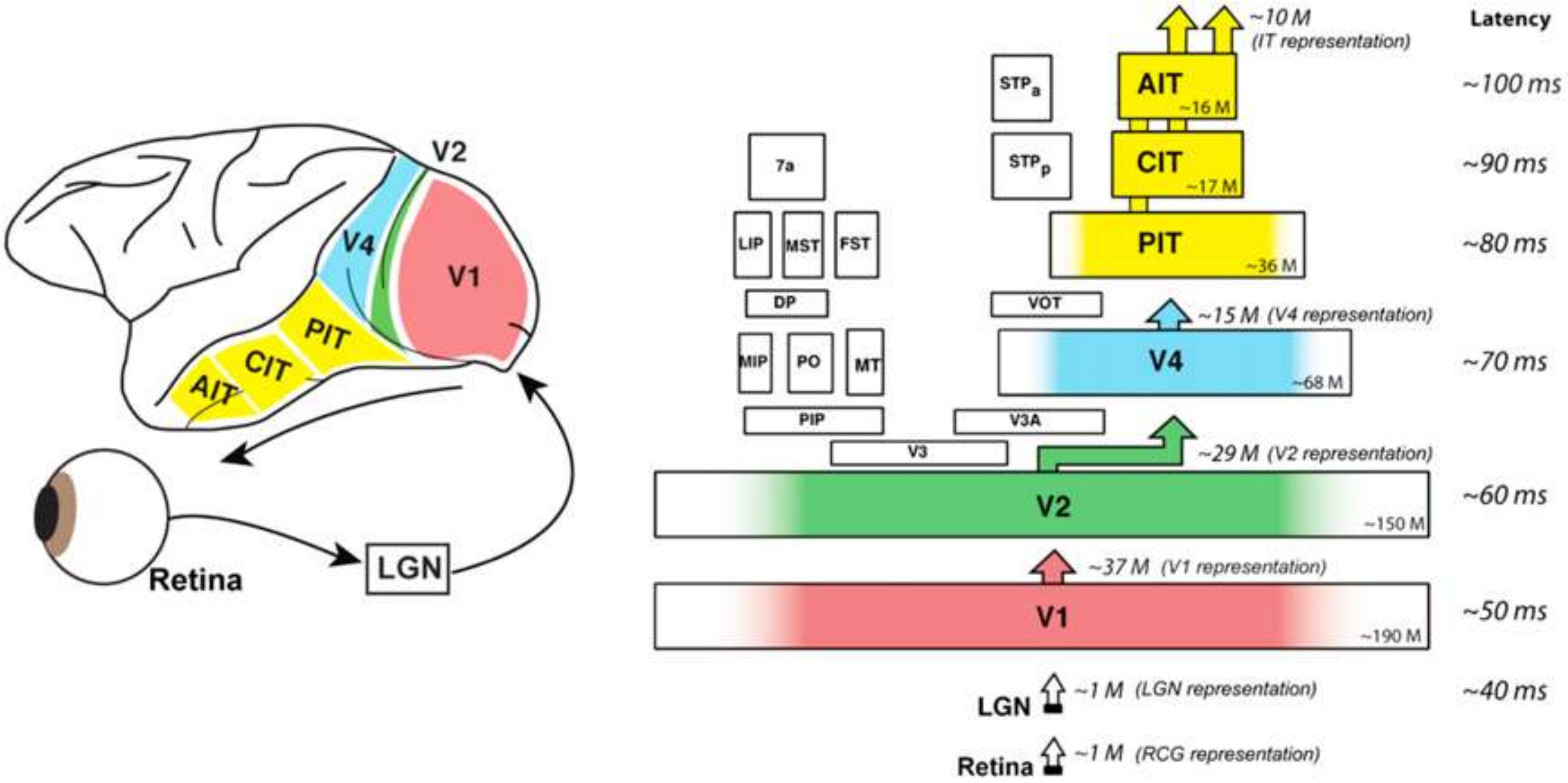
Ventral vs. Dorsal Visual Pathways

Aspect	Ventral Pathway ("What")	Dorsal Pathway ("Where/How")
Function	Object recognition, identification	Spatial awareness, movement coordination
Pathway	Occipital → Inferior Temporal Cortex	Occipital → Posterior Parietal Cortex
Main Input Type	Parvocellular (local, color)	Magnocellular (global, motion)
Processes	Shape, color, textures	Motion, depth, position
Consciousness	Linked to conscious perception	Often unconscious, guides real-time action
Damage Effects	Visual agnosia (can’t identify objects/faces)	Optic ataxia (impaired visually guided movement)
Example	Recognizing a face	Grasping a moving object

Visual information in both the ventral ("what") and dorsal ("where/how") pathways is processed hierarchically in the brain. As signals progress through each stage, neuronal responses become slower, receptive fields get larger, and stimuli become more complex—reflecting more abstract levels of processing.



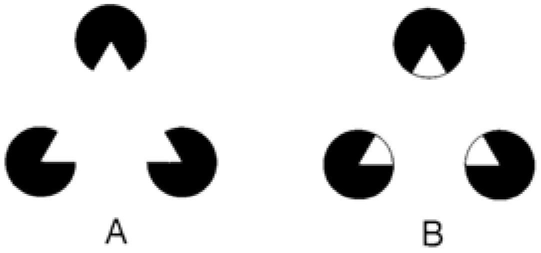




Visual area V4 is an intermediate cortical area in the ventral visual pathway.

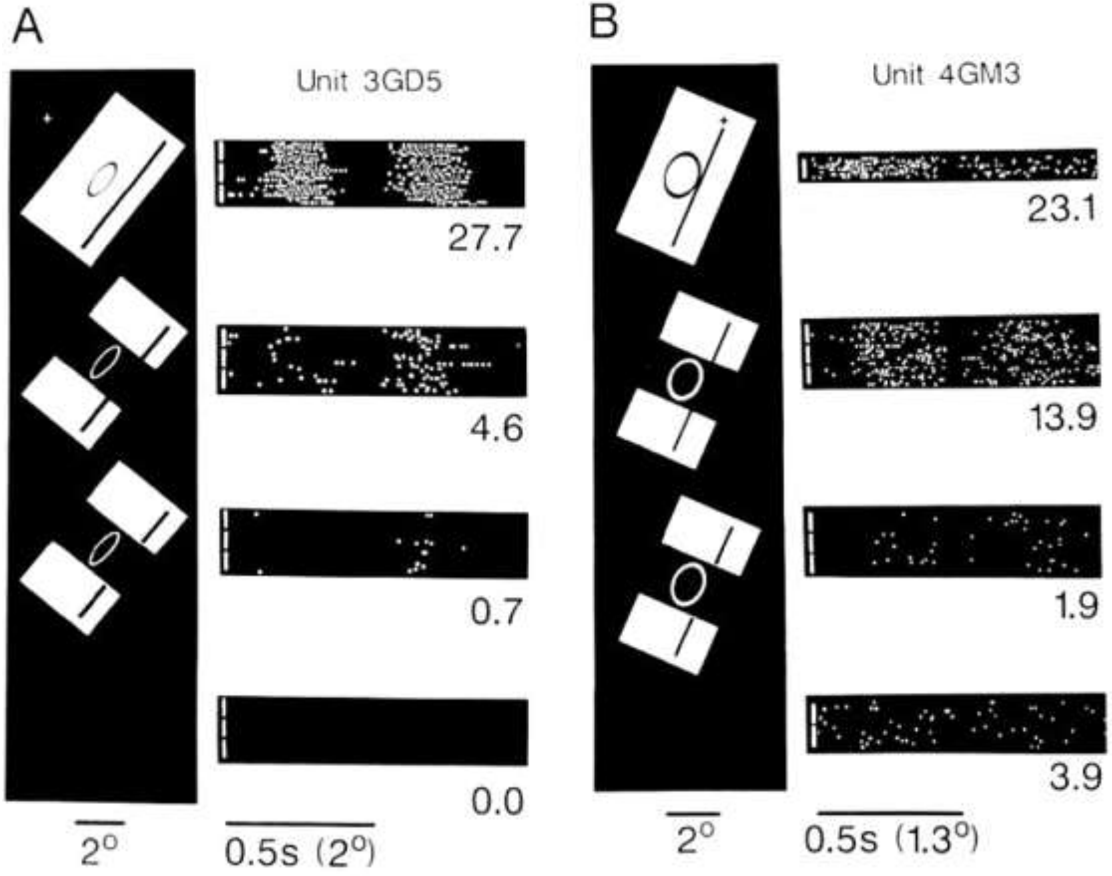
It considered to be crucial for visual object recognition and visual attention.
A primary role of V4 is to facilitate figure-ground segmentation of the visual scene, thus enabling both bottom-up and top-down visual processes.

Neural responses to illusory contours



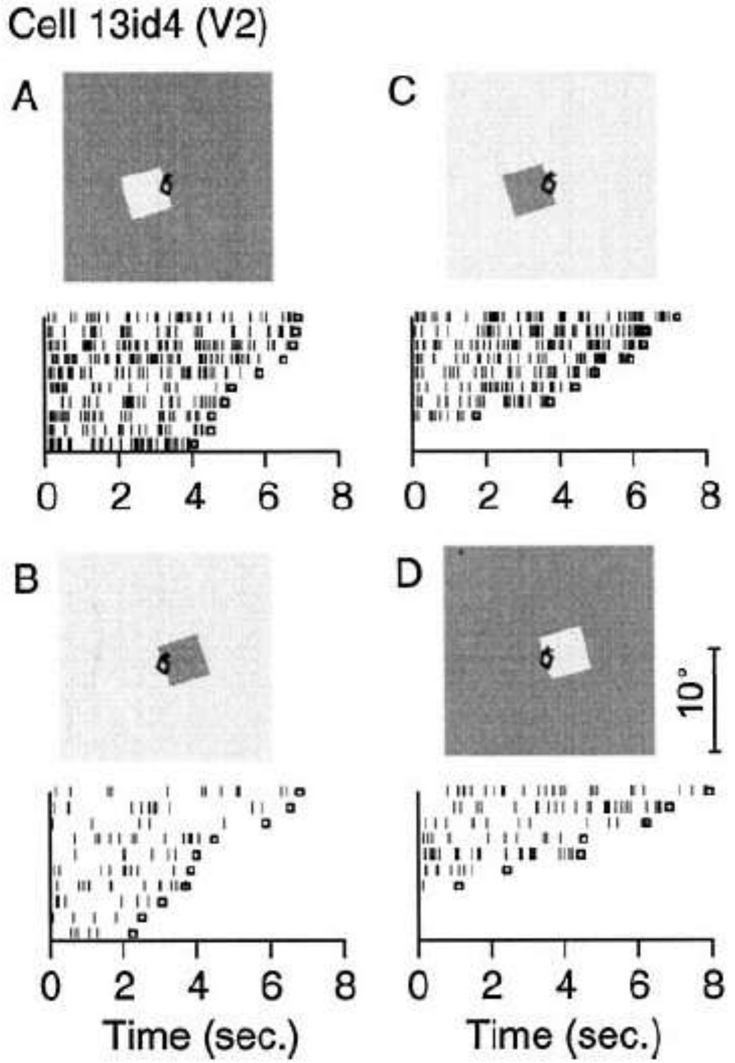
Figures in which humans perceive illusory contours evoke responses in cells of area 18 (V2) in the monkey visual cortex. Modifications that weakens the perception of contours also reduces the neuronal responses.

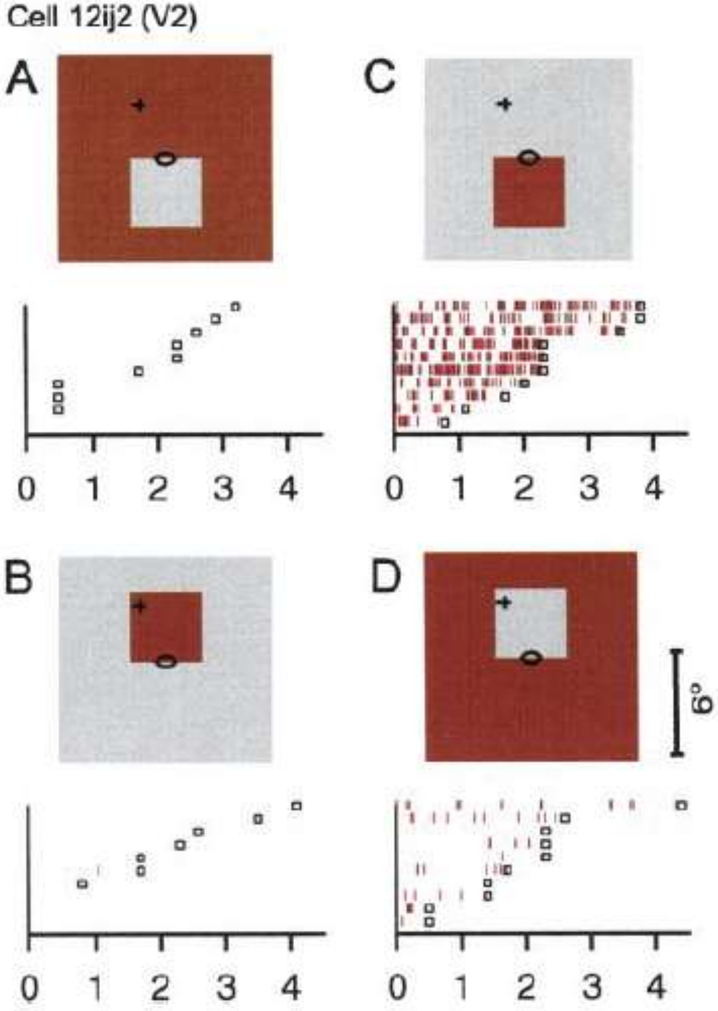
In contrast, cells in V1 (area 17) were apparently unable to see these contours.



Von der Heydt et al., Science, 1984

Coding of border ownership in monkey area V2

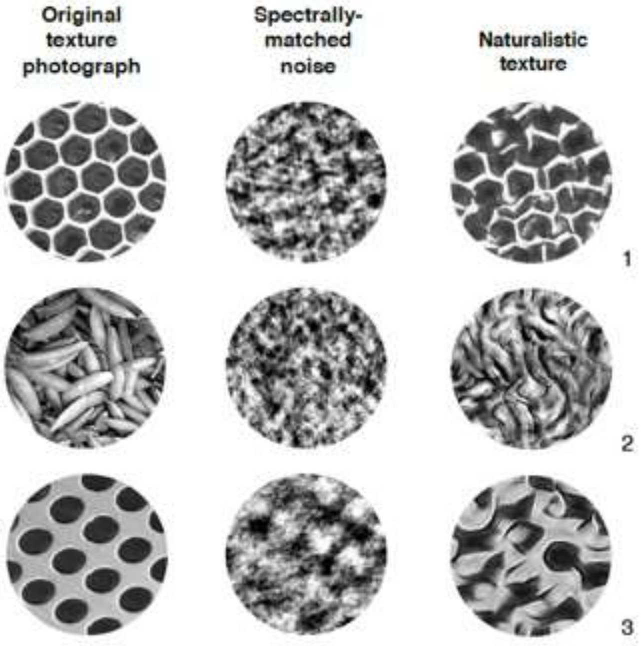


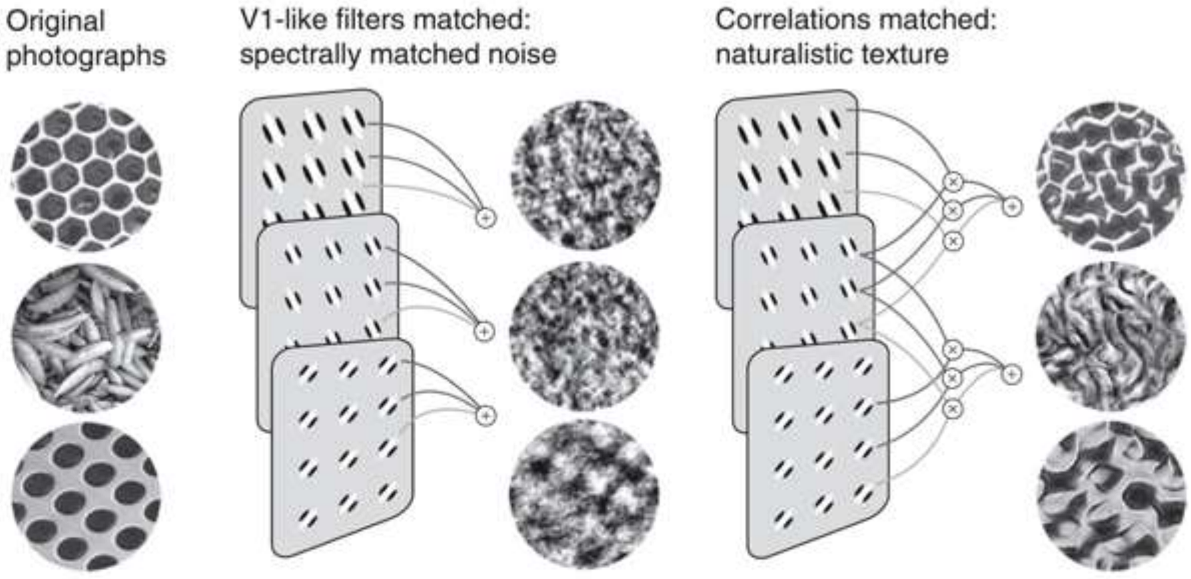


Zhou et al., J Neurosci, 2000

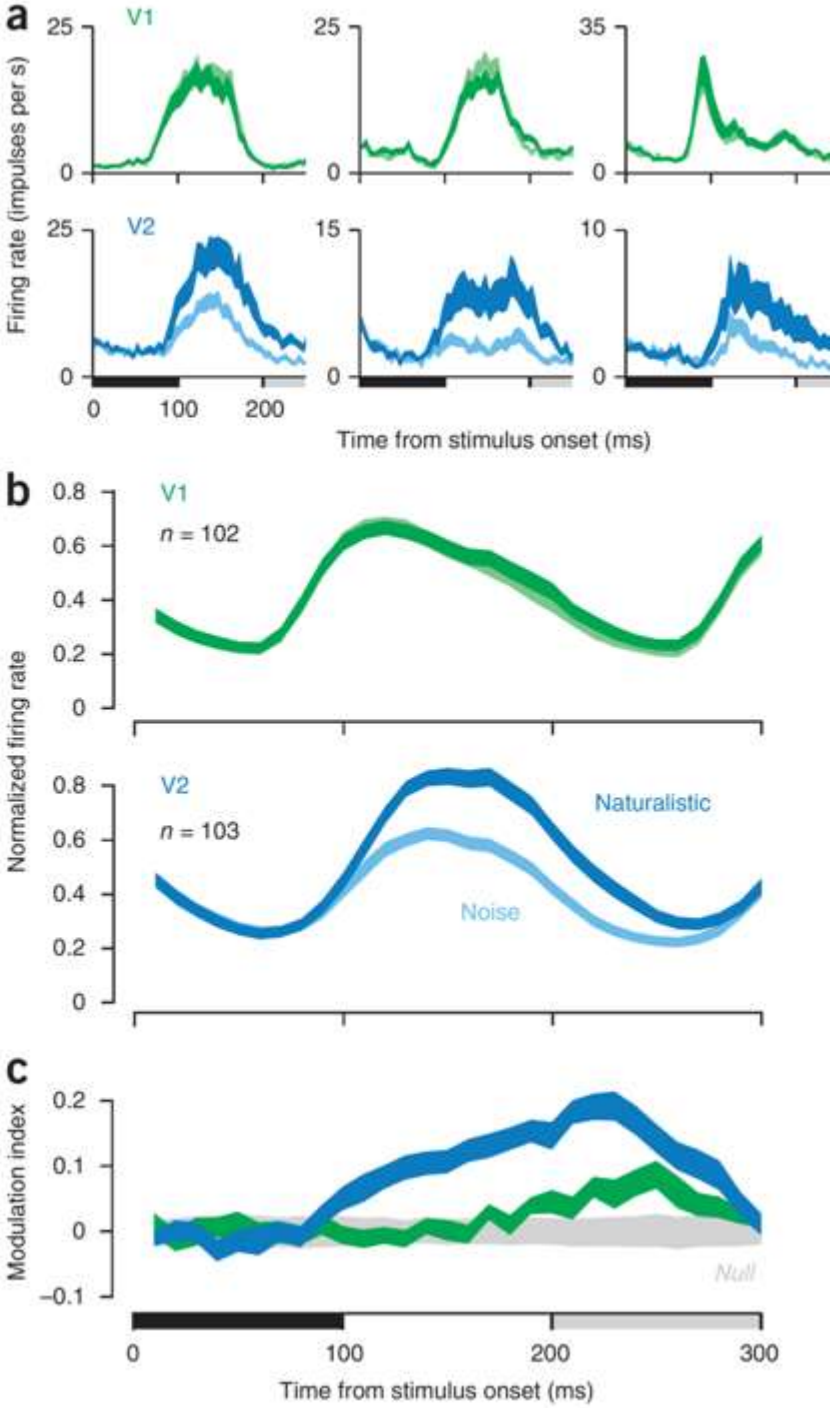
There is no generally accepted account of the function of V2, partly because no simple response properties robustly distinguish V2 neurons from those in V1.

Stimuli replicating the higher-order statistical dependencies found in natural texture images were used to stimulate macaque V1 and V2 neurons.




Original photographs
V1-like filters matched: spectrally matched noise
Correlations matched: naturalistic texture

Freedman et al., Nat. Neurosci., 2013



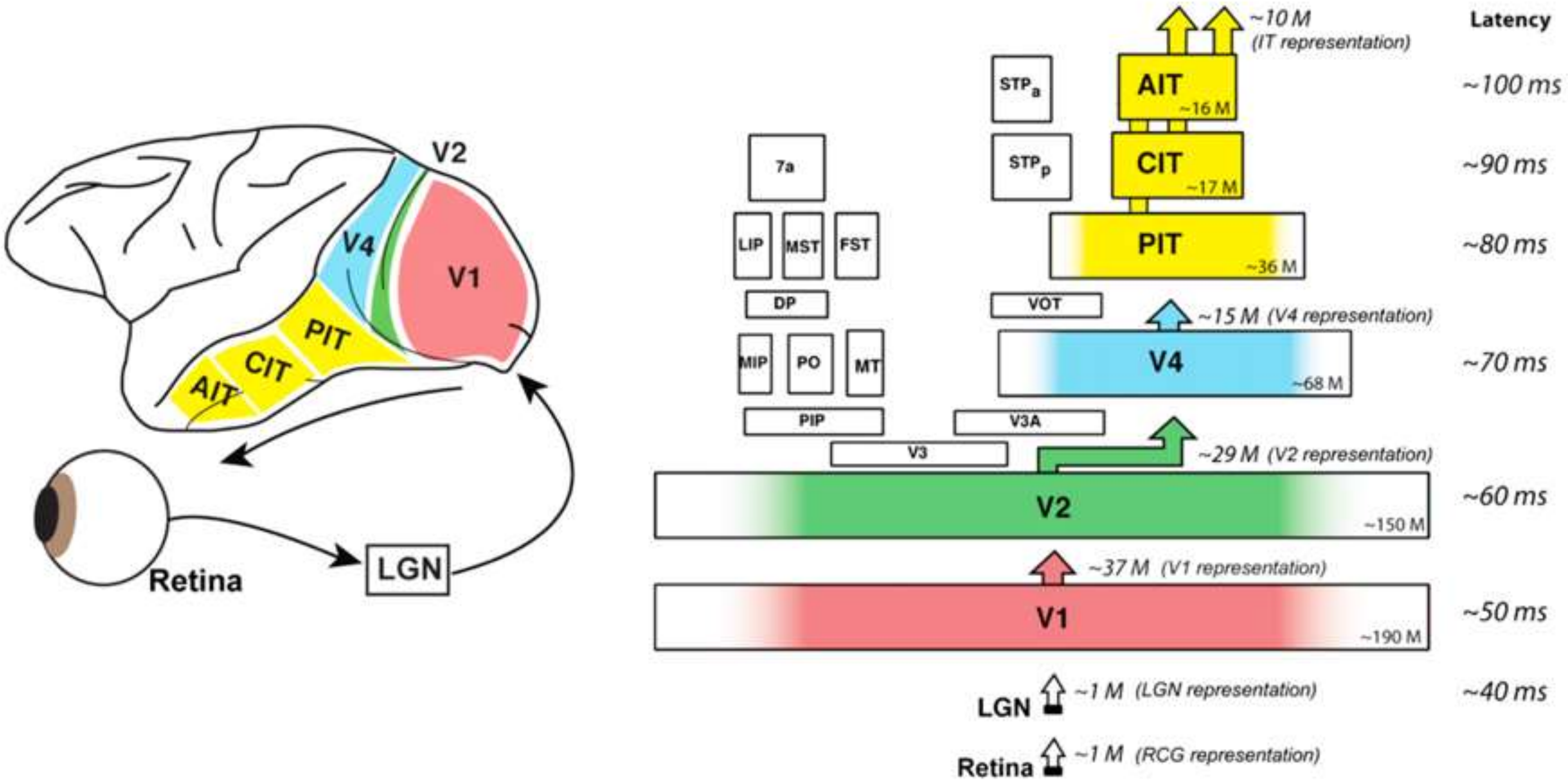
Neuronal responses to naturalistic textures differentiate V2 from V1 in macaques.

a) Time course of firing rate for three single units in V1 (green) and V2 (blue) to images of naturalistic texture (dark) and spectrally matched noise (light). Black bar indicates the presentation of the stimulus; gray bar indicates the presentation of the subsequent stimulus.

b) Time course of firing rate averaged across neurons in V1 and V2. Each neuron's firing rate was normalized by its maximum before averaging.

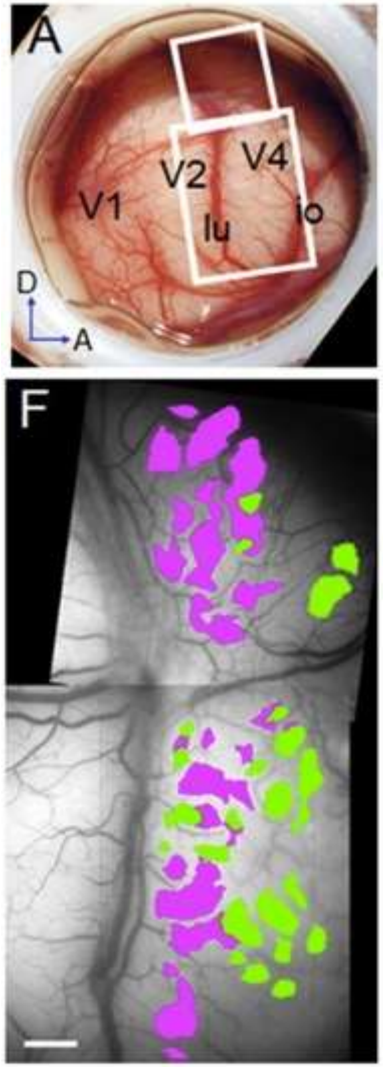
c) Modulation index, computed as the difference between the response to naturalistic and the response to noise, divided by their sum. Modulation was computed separately for each neuron and texture family, then averaged across all neurons and families.

Freedman et al., Nat. Neurosci., 2013

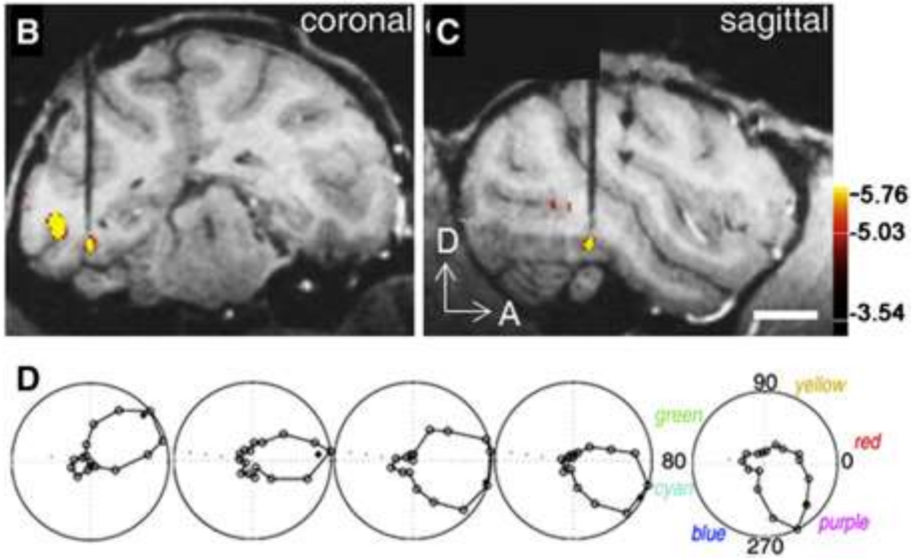


Visual area V4 is an intermediate cortical area in the ventral visual pathway.

It considered to be crucial for visual object recognition and visual attention.
A primary role of V4 is to facilitate figure-ground segmentation of the visual scene.
In other words, V4 is where a representation for “things” begins to emerge.

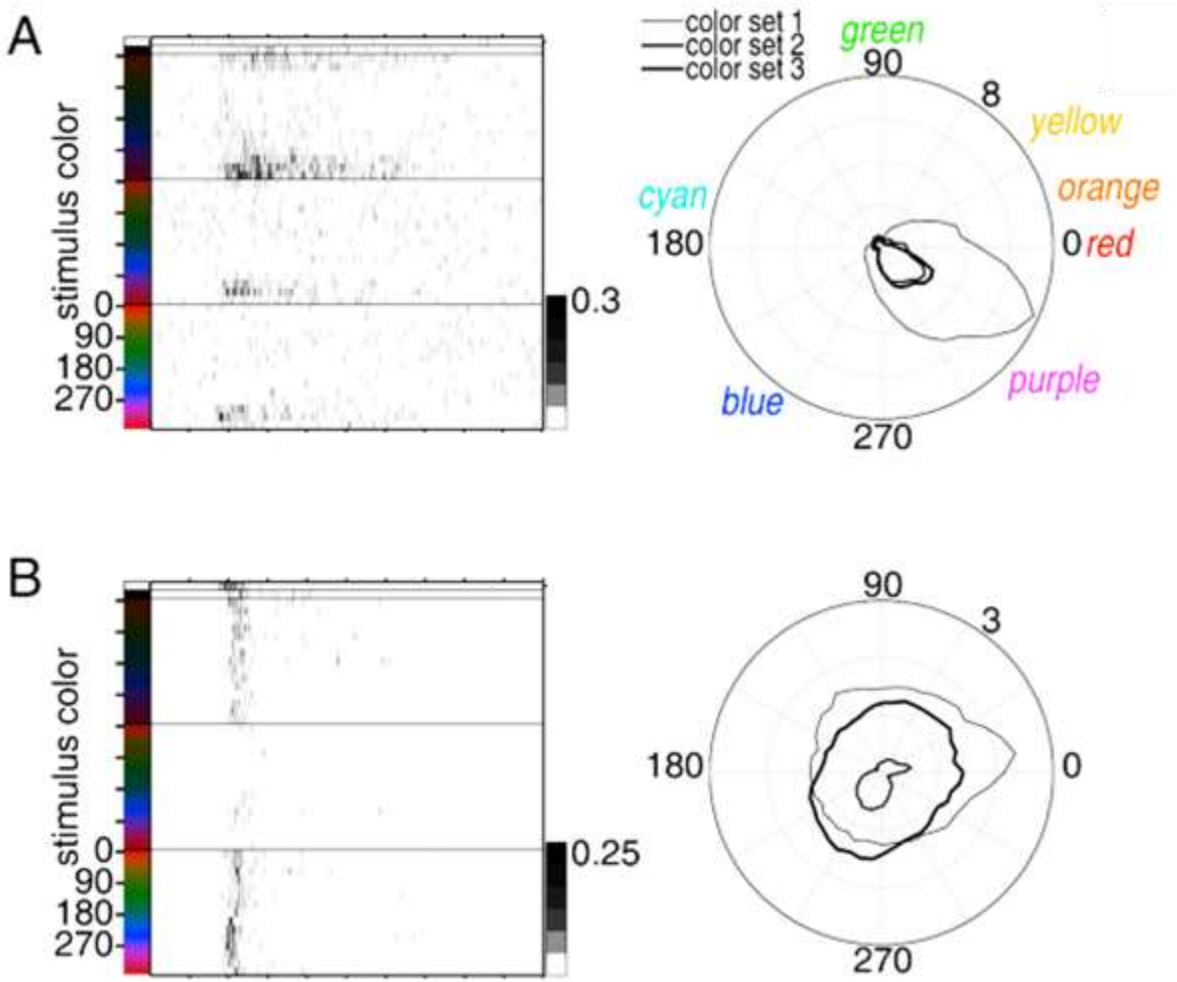


Recent advances in fMRI and optical imaging methods have revealed that V4 is not a homogenous area and may comprise segregated functional domains, dubbed globs (in analogy with the color blobs of V1), specialized for color processing, and interglobs, for orientation and shape processing



(F) illustrates segregation of color/ Lum and orientation preference bands in area V4

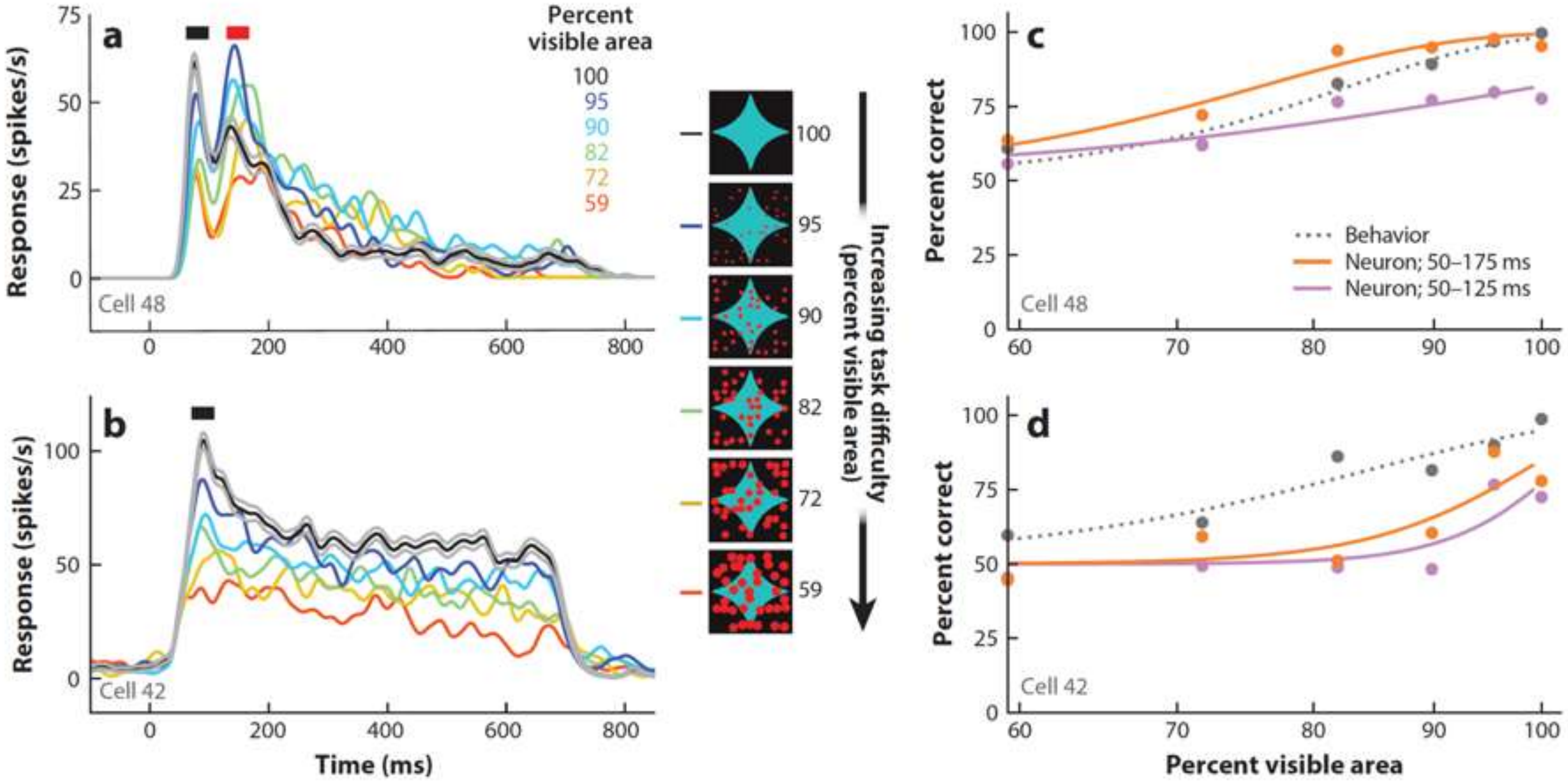
Figure shows color-response measurements for a typical cell in a glob (A), and a cell in the interglob (B) region. Left panels show time histograms to an optimally shaped bar of various colors; right panels show the color tuning in polar coordinates.



Most glob cells were excited by a specific hue, as shown by hue tuning in the polar plots.

This hue selectivity was luminance invariant.
That is, hue tuning does not change with change in luminance between stimulus and background.

Conway et al., Neuron, 2007



Responses of two example V4 neurons that exhibit different response profiles to partially occluded shape stimuli.

Pasupathy et al., Ann Rev Vis Sci, 2020

Object identification and categorization

The visual experience of the world is fundamentally centered on objects.

By visual object we mean a set of visual characteristics (e.g., visual features) grouped or joined perceptually in discrete units on the basis of the organizational principles of the Gestalt, such as proximity, similarity, closure, good continuation, good form, connection, etc.

By visual recognition we mean the ability to assign a verbal label (e.g., a name) to objects in the visual scene.

There are at least two possible object recognition tasks, distinguished by level of specificity: identification and categorization.

An object can be recognized at an individual level (e.g., a Siamese cat), or at a more general categorical level, as an object belonging to a given class (a cat, a mammal, an animal, and so on).

Object identification and categorization

The visual experience of the world is fundamentally centered on objects.

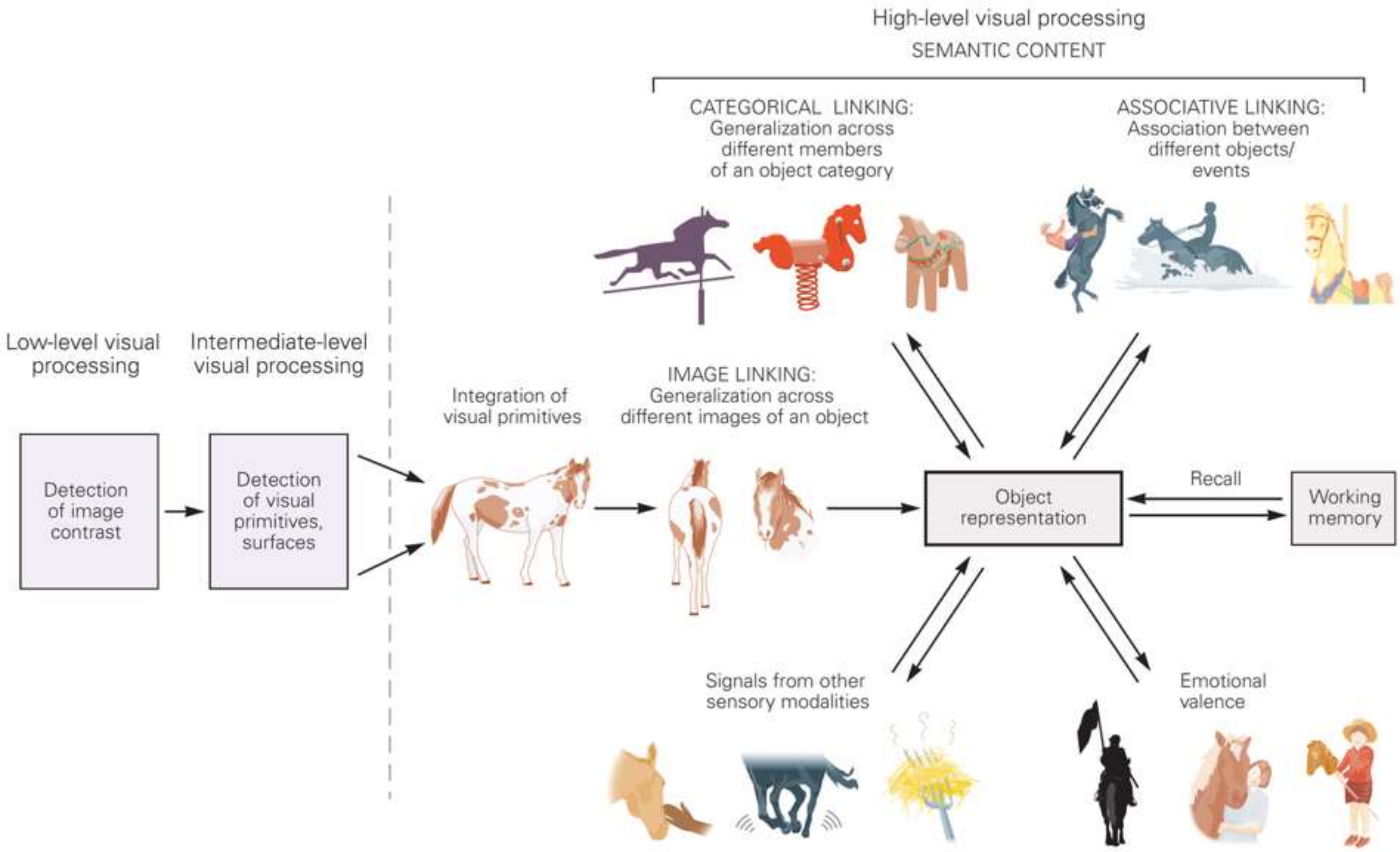
By visual object we mean a set of visual characteristics (e.g., visual features) grouped or joined perceptually in discrete units on the basis of the organizational principles of the Gestalt, such as proximity, similarity, closure, good continuation, good form, connection, etc.

By visual recognition we mean the ability to assign a verbal label (e.g., a name) to objects in the visual scene.

There are at least two possible object recognition tasks, distinguished by level of specificity: identification and categorization.

An object can be recognized at an individual level (e.g., a Siamese cat), or at a more general categorical level, as an object belonging to a given class (a cat, a mammal, an animal, and so on).

High-level visual processing is concerned with object recognition


High-level visual processing
SEMANTIC CONTENT
CATEGORICAL LINKING: Generalization across different members of an object category
ASSOCIATIVE LINKING: Association between different objects/ events
Low-level visual processing
Intermediate-level visual processing
Integration of visual primitives
IMAGE LINKING: Generalization across different images of an object
Detection of image contrast
Detection of visual primitives, surfaces
Object representation
Recall
Working memory
Signals from other sensory modalities
Emotional valence

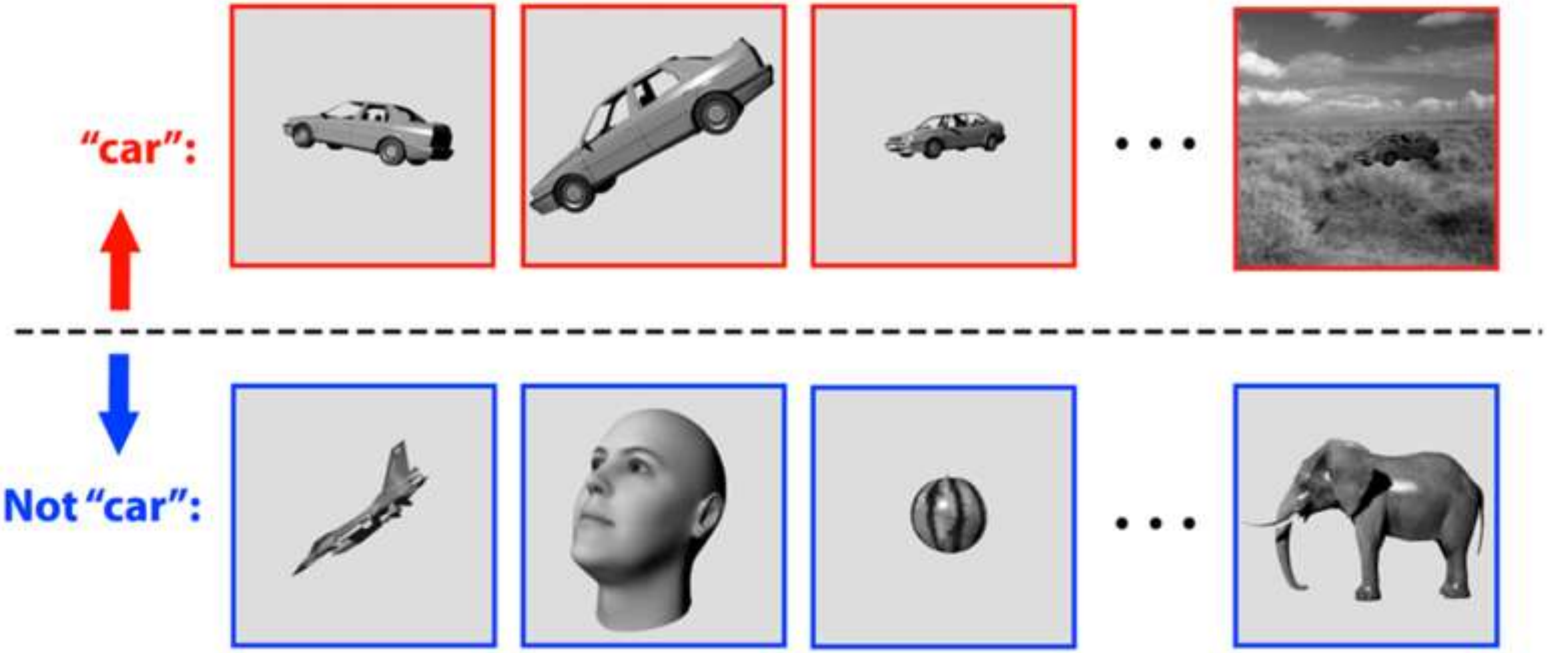
Selectivity and object constancy (or invariance)

A computational difficulty of object recognition is that it requires both:

selectivity (different responses to distinct objects, such as one face with respect to another face);

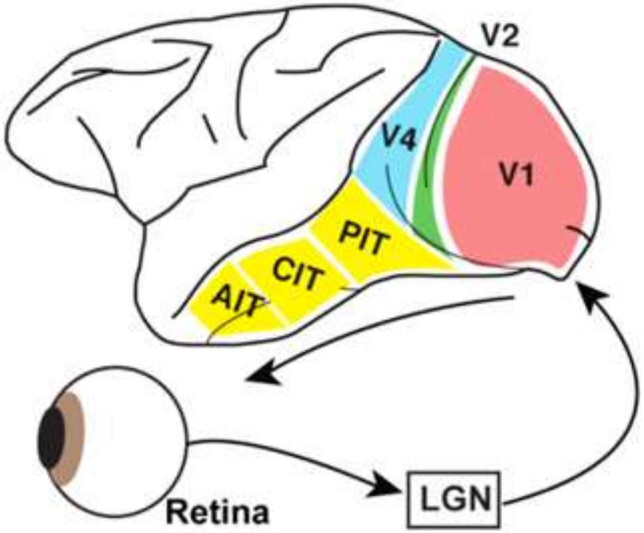
and invariance with respect to image transformations (similar responses to, for example, rotations or translations of the same face);

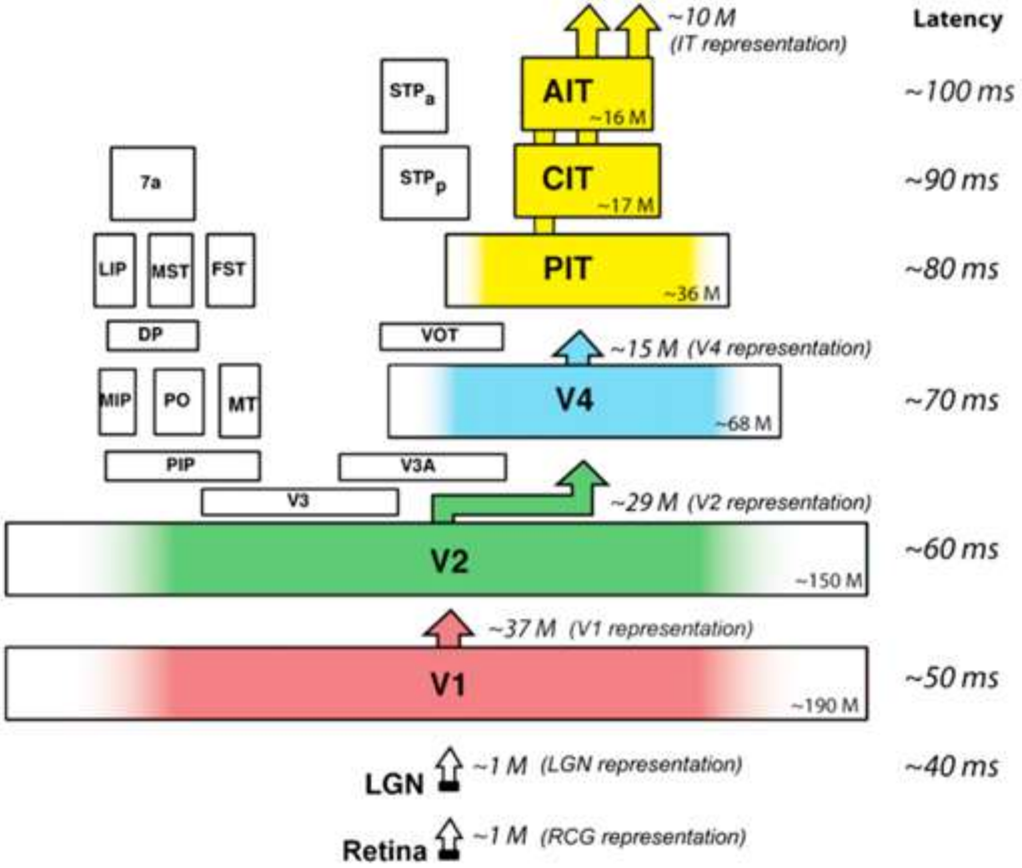
In fact, we are able to recognize the same object even when the image it projects on the retina varies considerably.



Core object recognition is the ability to rapidly (<200 ms viewing duration) discriminate a given visual object (e.g., a car, top row) from all other possible visual objects (e.g., bottom row).

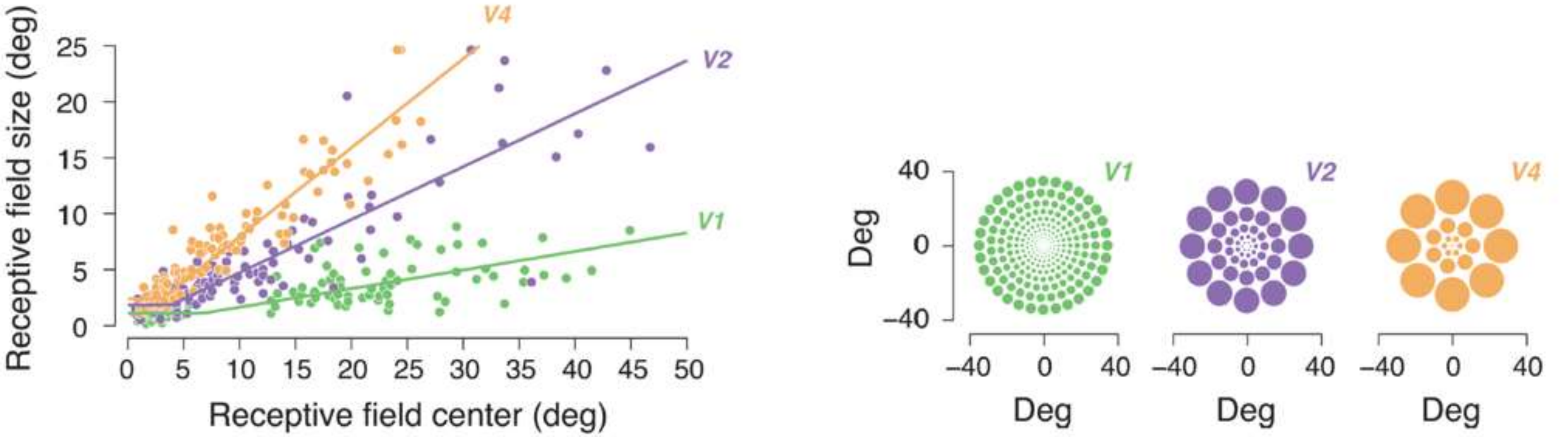
Primates perform this task remarkably well, even in the face of (identity-preserving) transformations (e.g., changes in object position, size, viewpoint, and visual context).





Studies in animals (e.g., primates) primarily implicates the inferior temporal cortex in object perception and recognition.

The inferior temporal cortex is a large region of the brain that includes at least two major functional subdivisions - the posterior area, or temporo-occipital cortex (TEO area), and the anterior area, or inferotemporal cortex (IT area).



In contrast to V1, V2, V4 and partly PIT, CIT and AIT show no clear retinotopy.

Large receptive fields (from 2 to 30 degree). Most of the receptive fields include the fovea. Foveal stimuli evoke stronger responses.

Freeman and Simoncelli, Nat. Neurosci., 2013

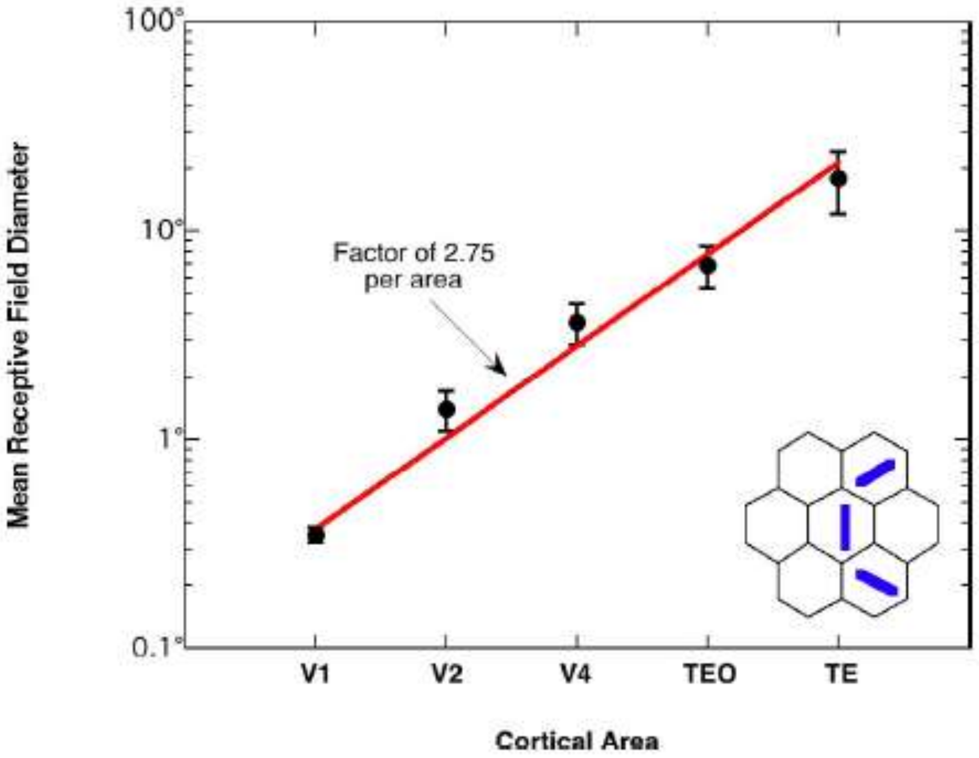
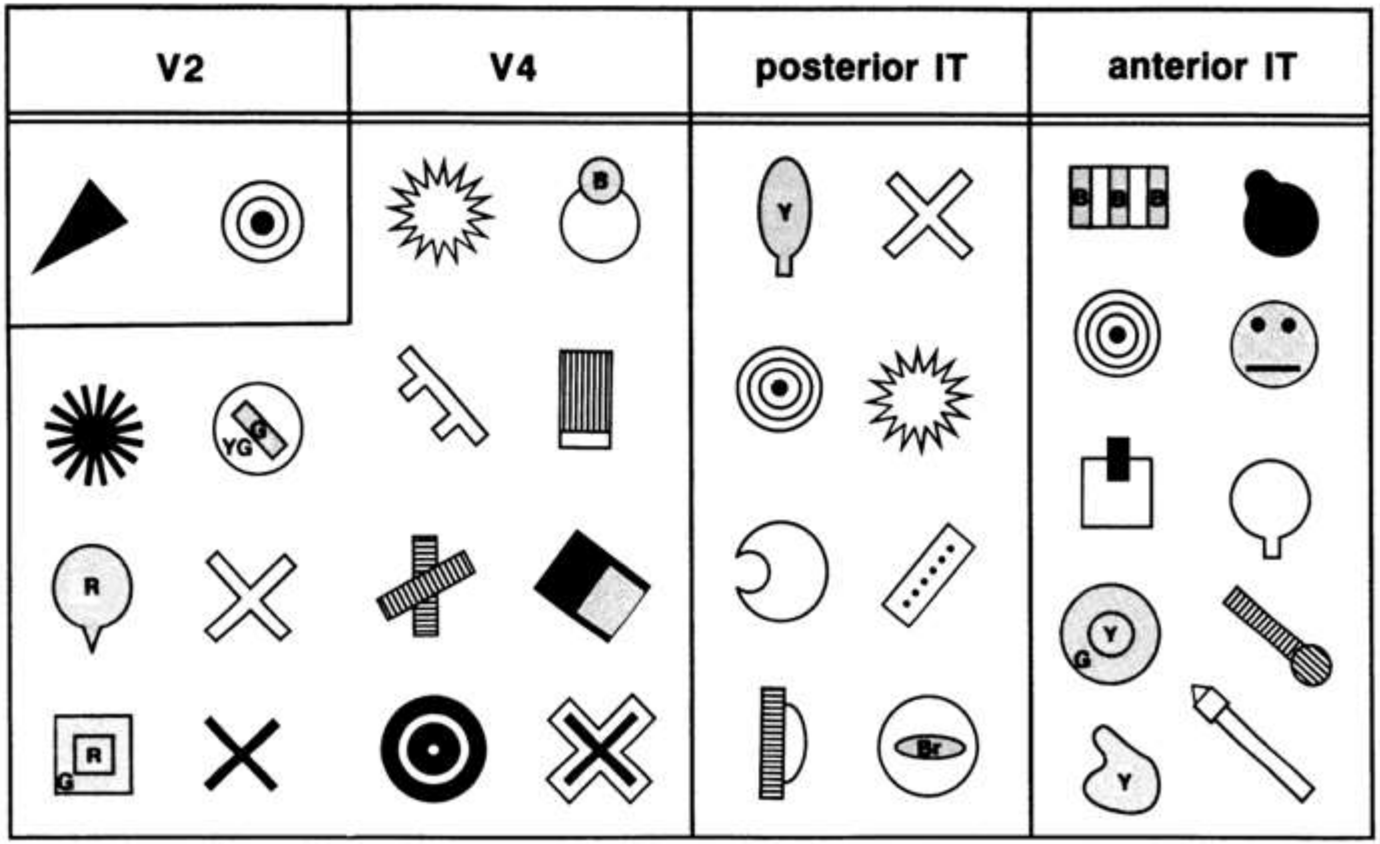


Figure 1. Mean data from four studies on receptive field diameter in successive cortical areas in the ventral pathway. Error bars represent the range of means across the four studies. The red line shows that the data can be fit by constant diameter increase of 2.75× from area to area. Inset in lower right shows an example of a curved contour represented by a combination of adjacent, oriented (orientations in blue) V1 receptive fields (individual hexagons) in a 3.0× larger diameter V2 receptive field.

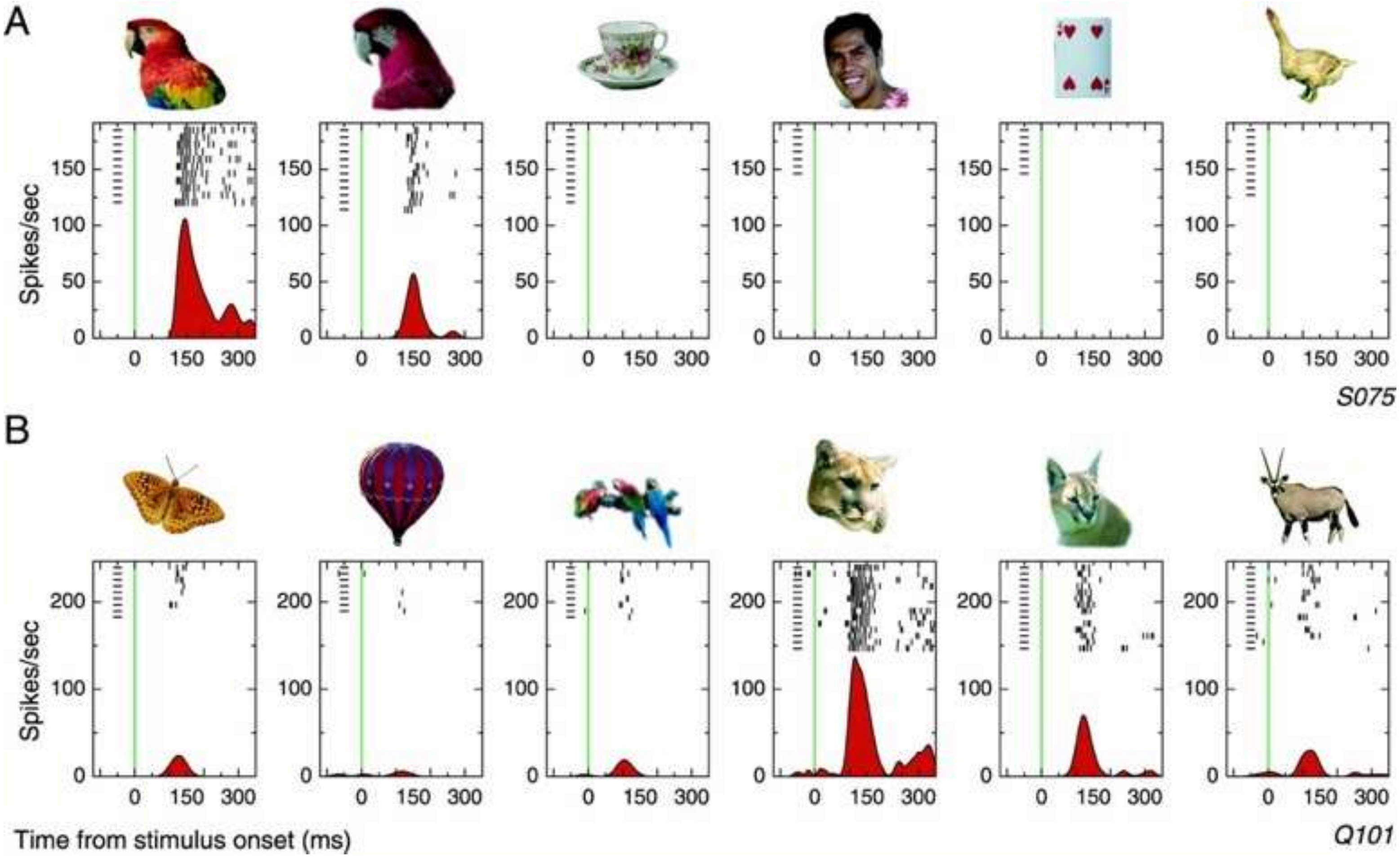
The diameter increase of a factor of about 3.0 is consistent with receptive fields one area higher in the hierarchy being constructed from combinations of nearest neighbours in its input area.

Increased complexity of effective stimuli along the ventral visual path

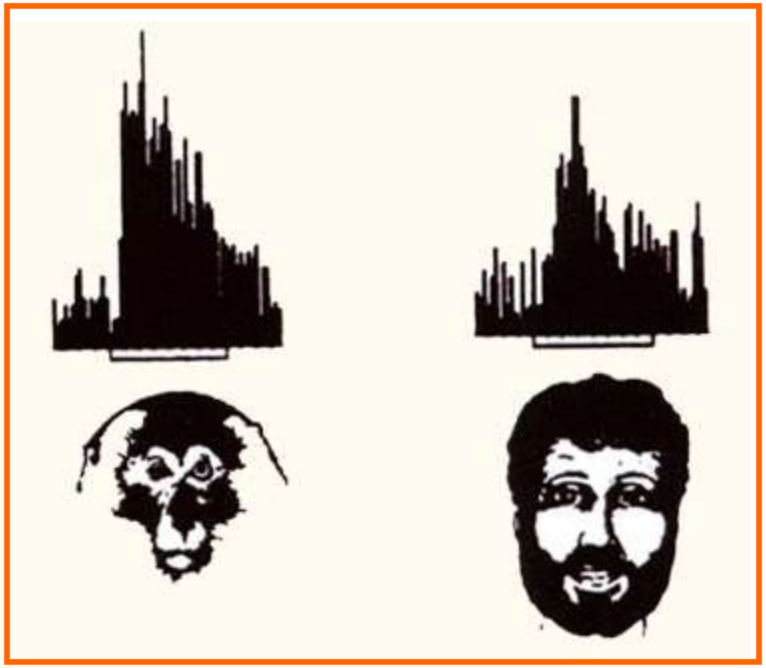


Kobatake and Tanaka, J. Neurophysiol., 1994

Neurons in the IT respond to relatively complex stimuli, often to biologically relevant objects such as human and other animal faces, hands and other parts of the body.

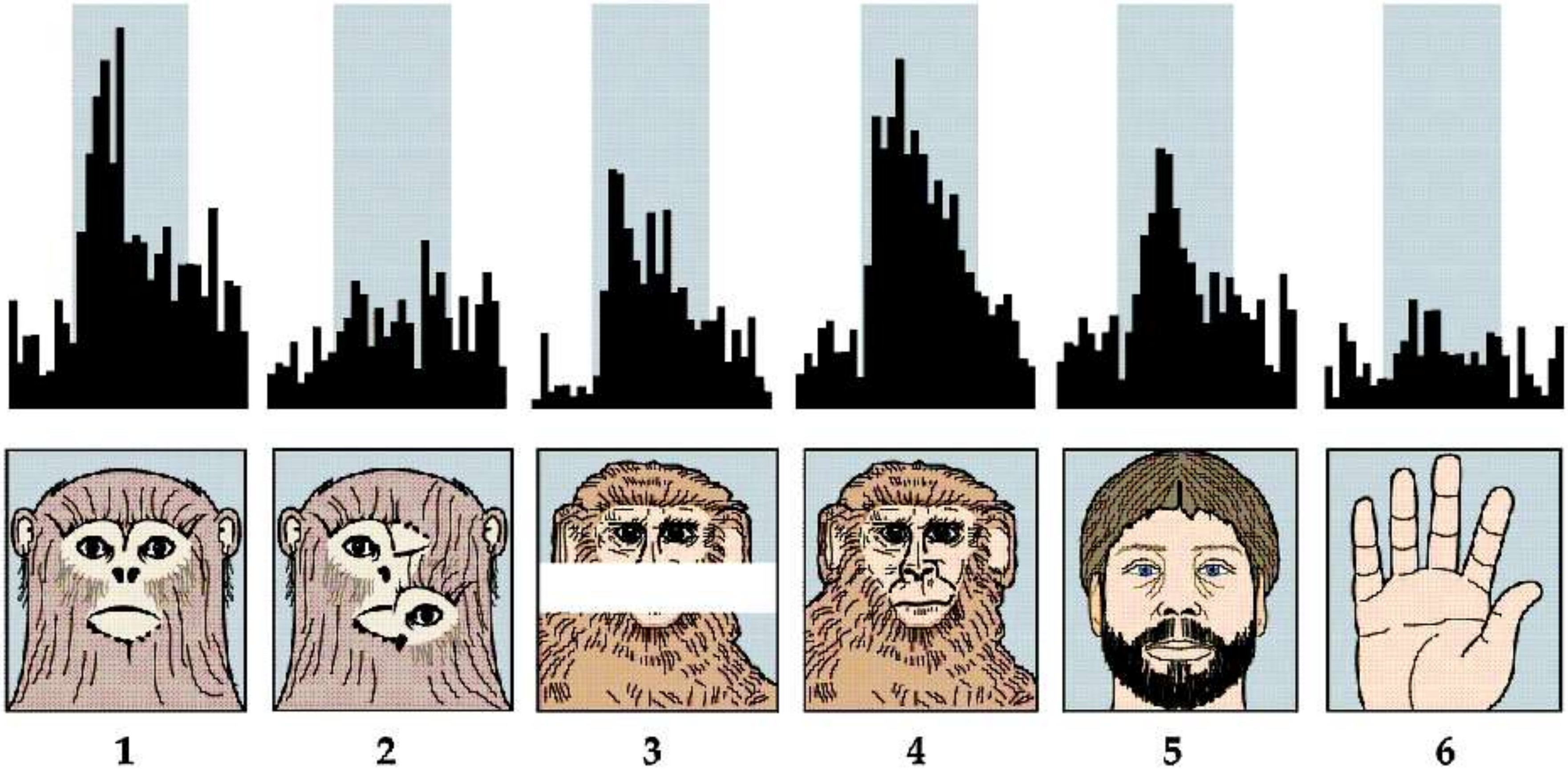


In the early 1980s, several researchers (Bruce et al., Perrett et al.) Identified in the monkey a group of IT neurons that responded selectively to faces.



Question: Are there in IT, as for faces, selective cells for the different types of objects that can be encountered in the outside world (neurons for chairs, for flowers, for cars, etc ...)?

Neuron that responds to faces: The neuron responds to faces of different species (1, 4, 5). The discharge is reduced if the elements of the face are mixed (2) or occluded (3). The neuron does not respond to other biologically relevant stimuli (5).



Desimone, Albright, Gross and Bruce, J. Neurosci, 1984

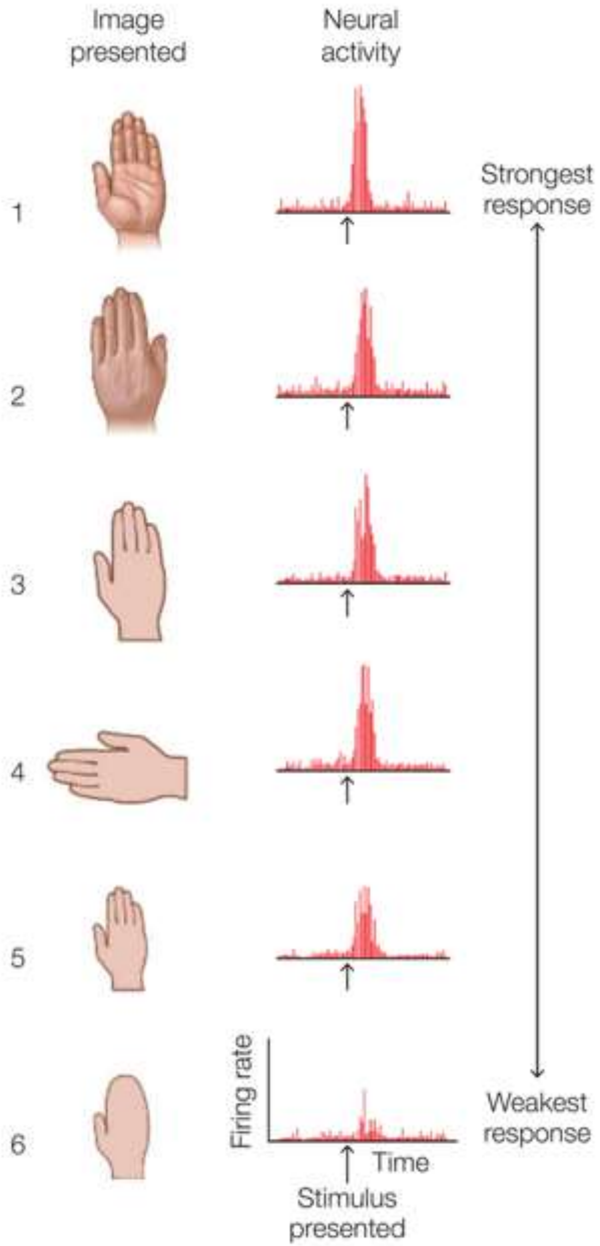
Recording of a single neuron from monkey IT cortex (Desimone et al., 1984)

This cell is activated by the vision of the human hand.

The first five images in the figure show the cell's response to various perspectives of a hand.

Activity is high regardless of hand orientation and only decreases slightly when the hand is noticeably smaller.

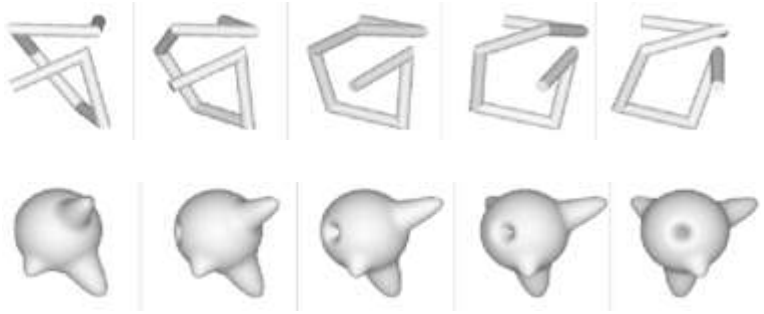
The sixth image shows that the response decreases if the stimulus has the same shape, but does not have well-defined fingers.



Hung et al., Science, 2005

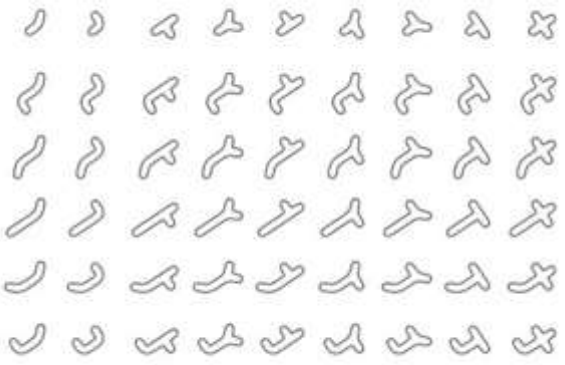
Tanaka, Science, 1993

Logothetis et al., Curr, Biol., 1995

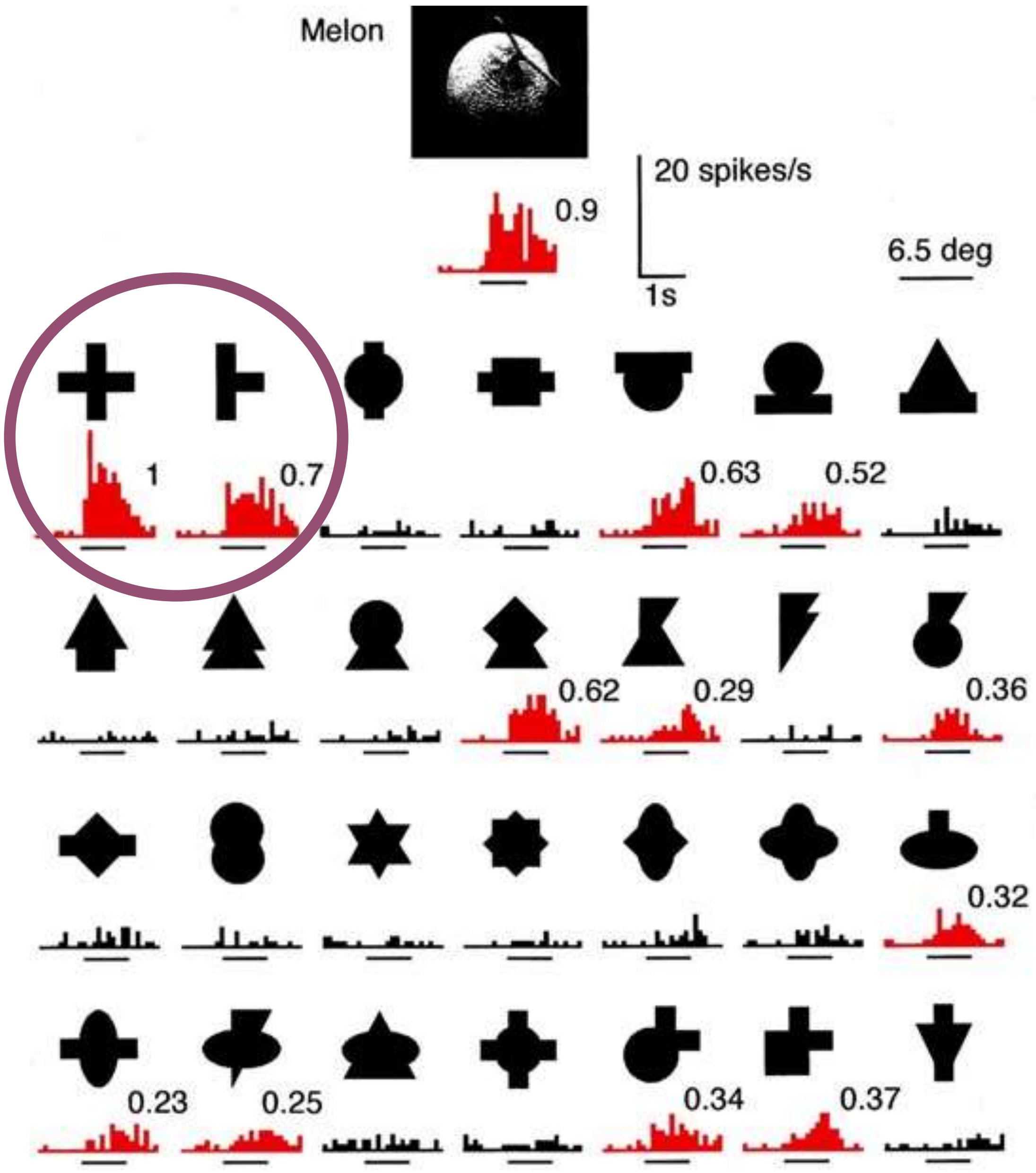


Kiani et al., J. Neurophysiol., 2007

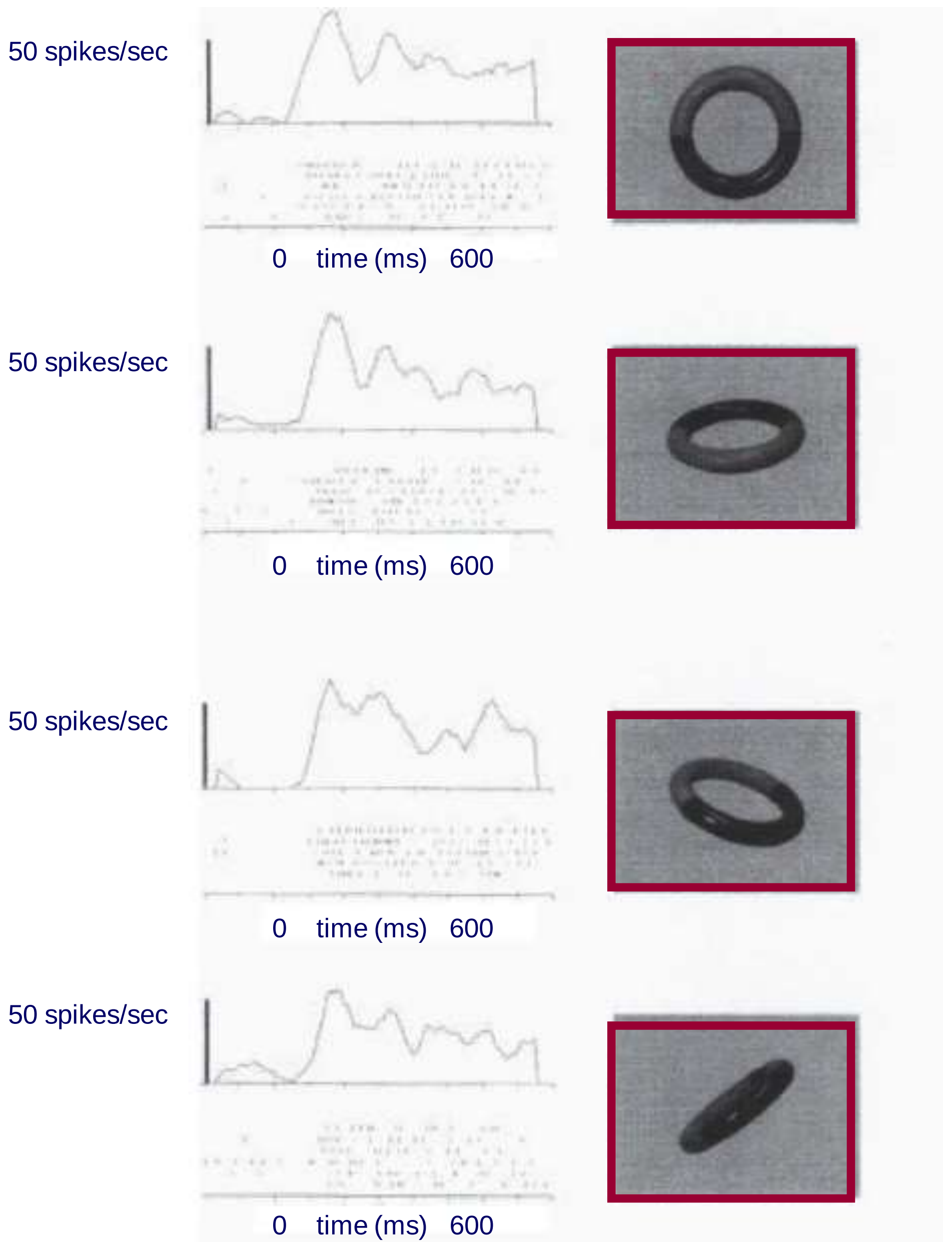
Freiwald and Tsao, Science, 2010



Brincat and Connor, Nat. Neurossci, 2007



For example, the neuron shown in the figure seems to respond to the complex image of a fruit (melon). However, a detailed study (method of progressive reduction) shows that the neuron also responds to simpler stimuli that represent the visual elements of an object to which the neuron is sensitive.



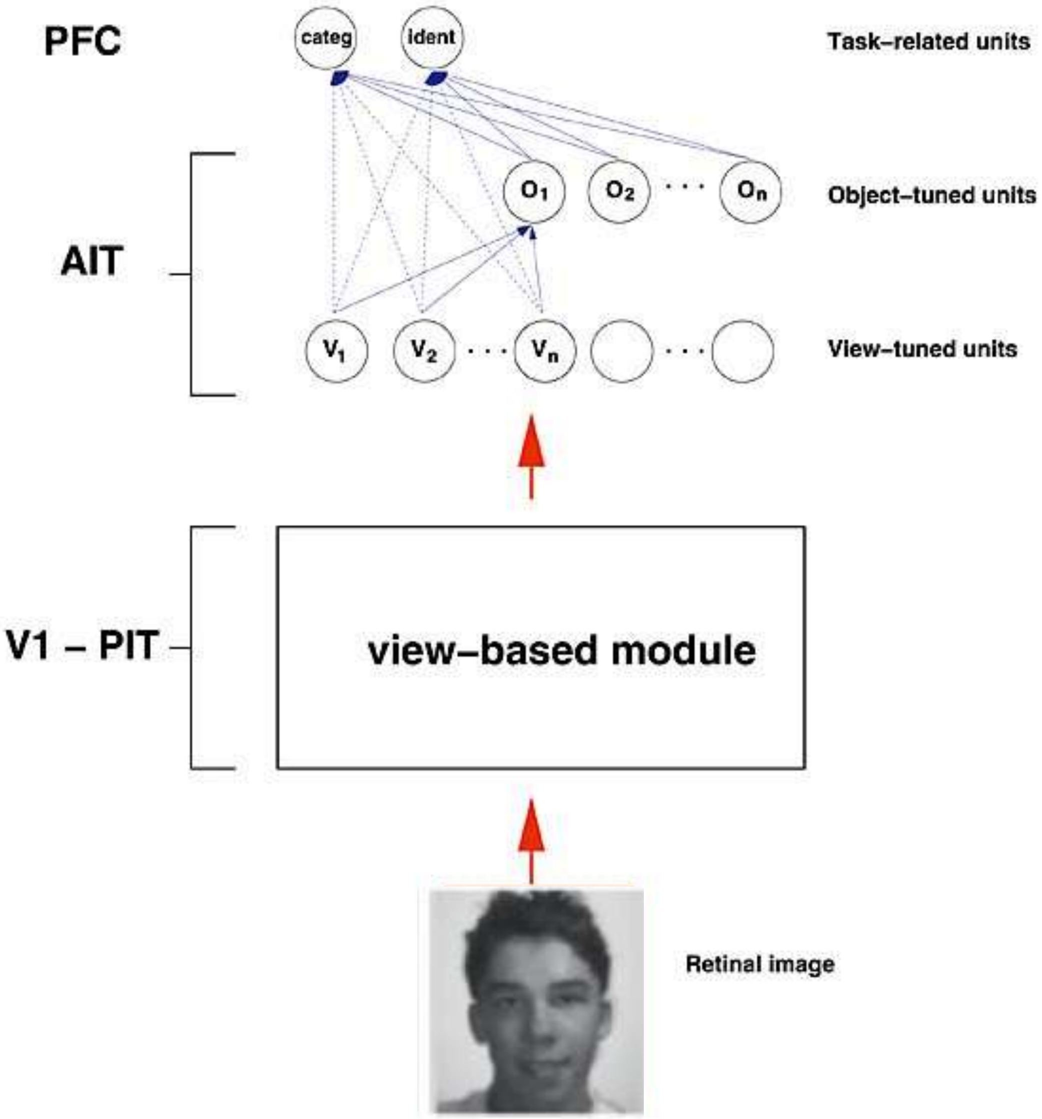
The majority of IT neurons respond to a stimulus only when it is presented from specific points of view (view-dependent responses).

Some neurons (10%) selectively respond to familiar stimuli regardless of their position with respect to the observer (view-independent responses).

These response, although rare, indicate that IT is capable of forming a (relatively abstract) representation of the object, rather than responding to one of the different forms that the object can take when its position with respect to the observer changes.

Logothetis & Sheinberf, Ann Rev Neurosci, 1996

Hierarchical model of object recognition



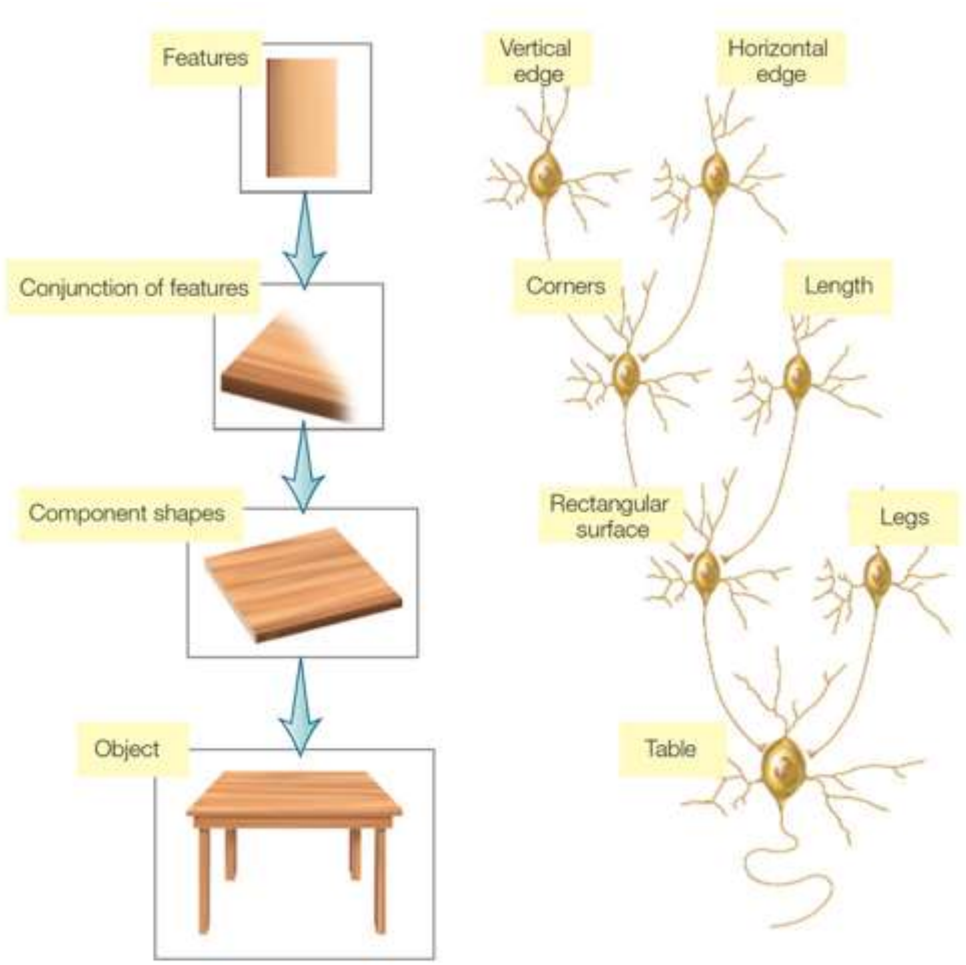
(view-invariant units)

(view-dependent units)

PFC = Prefrontal cortex
AIT = Anterior IT
PIT = Posterior IT

Riesenhuber & Poggio, Nat. Neurosci. 2000

Hierarchical model of the object recognition

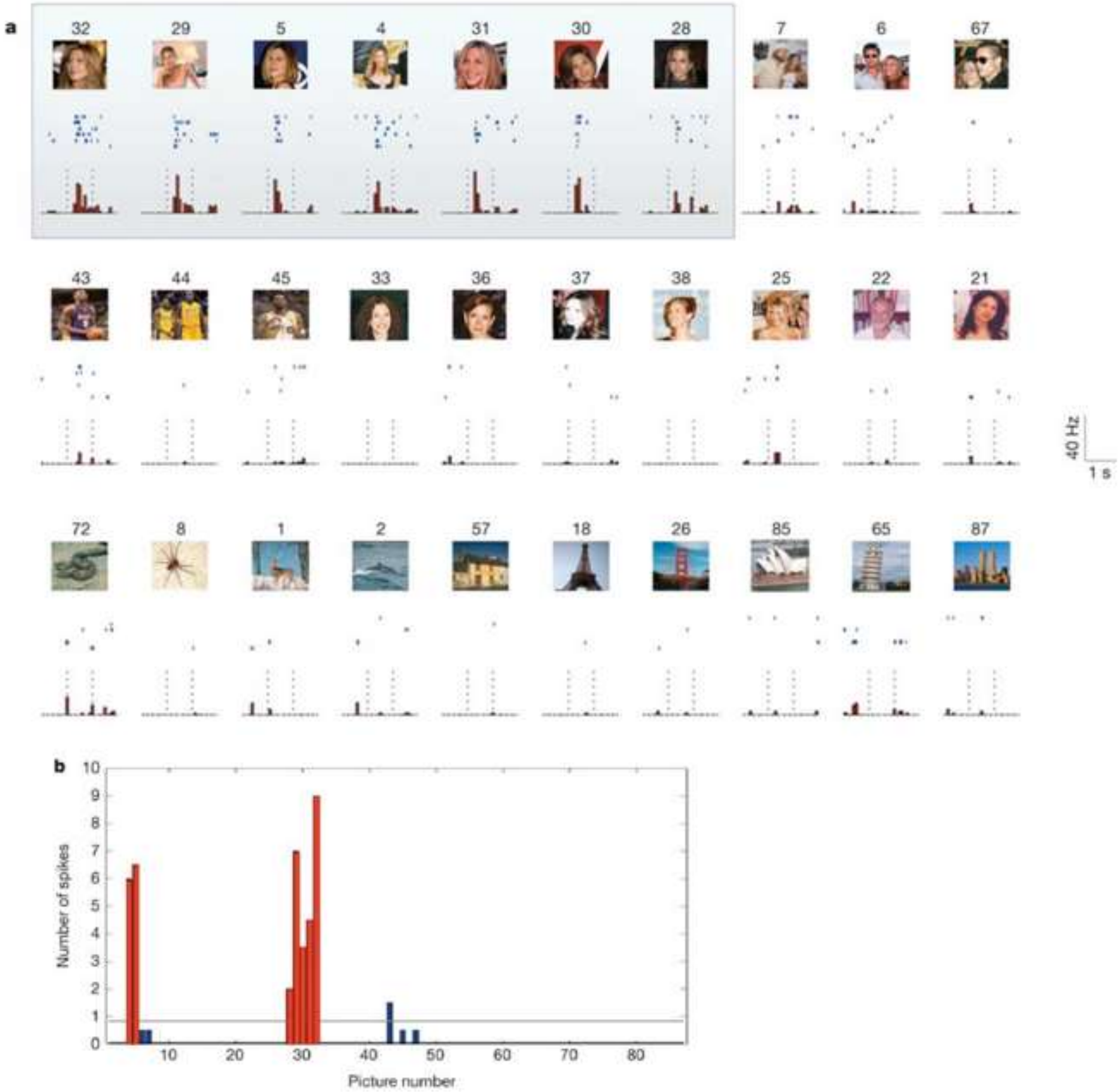


The finding that IT cells selectively respond to more complex stimuli than V1, V2 and V4 is consistent with a hierarchical model of object perception.

According to this model, each subsequent level encodes more complex combinations from the inputs of the previous level.

The type of neuron that can recognize a complex object has been called the gnostic unit, referring to the idea that the cell (or cells) signals the presence of
a complex, highly specific, and significant stimulus: that is, a known object, place or animal that has been encountered in the past.

Jennifer Aniston cell



Quiroga et al., Nature 2006

Local or distributed coding?

It is tempting to conclude that the cell represented by the activity of IT cells signal the presence of an object (a hand or face), independent of the point of view.

In this regard, the researchers coined the term 'grandmother cell’ (Lettvin, 1969) to convey the idea that people's brains may have a gnostic unity that is activated only when the grandmother comes into view.

Other Gnostic units would specialize in recognizing, for example, a blue Volkswagen or the Golden Gate Bridge.

Jerome Lettvin at MIT, 1952

Distributed code hypothesis

An alternative to the Grandmother cell hypothesis is that object recognition is the result of a distributed activation pattern on the population of IT neurons.

According to this hypothesis, recognition is due not to one unit but to the collective activation of many units.

Distributed code theories easily explain why we can recognize similarities between objects (say, a tiger and a lion) and make mistakes between visually similar objects - both objects activate many of the same neurons.

Losing some units may degrade our ability to recognize an object, but the remaining units may be enough.

Distributed code theories also explain our ability to recognize new objects. New objects have a resemblance to familiar things, and our perceptions result from activating units that represent their characteristics.

IT Cortex (Pre-ANN Era)

- IT neurons respond only to visual stimuli.
- The RF always include the fovea, that is the part of the retina most involved in the fine recognition of a visual stimulus.
- RFs tend to be large, providing the opportunity to generalize the stimulus within the receptive field, and often extend along the midline in both visual hemifields
- IT neurons encode complex characteristics of the stimulus (not simple features, such as color, form orientation, depth). A single neuron might respond strongly to a face, a hand, or a specific object category (e.g. tools, animals). Response is invariant to some changes in size, position, pose, and illumination.
- IT neurons often maintain selectivity even with partial or noisy stimuli. Suggests robustness and generalization, not pixel-perfect matching.

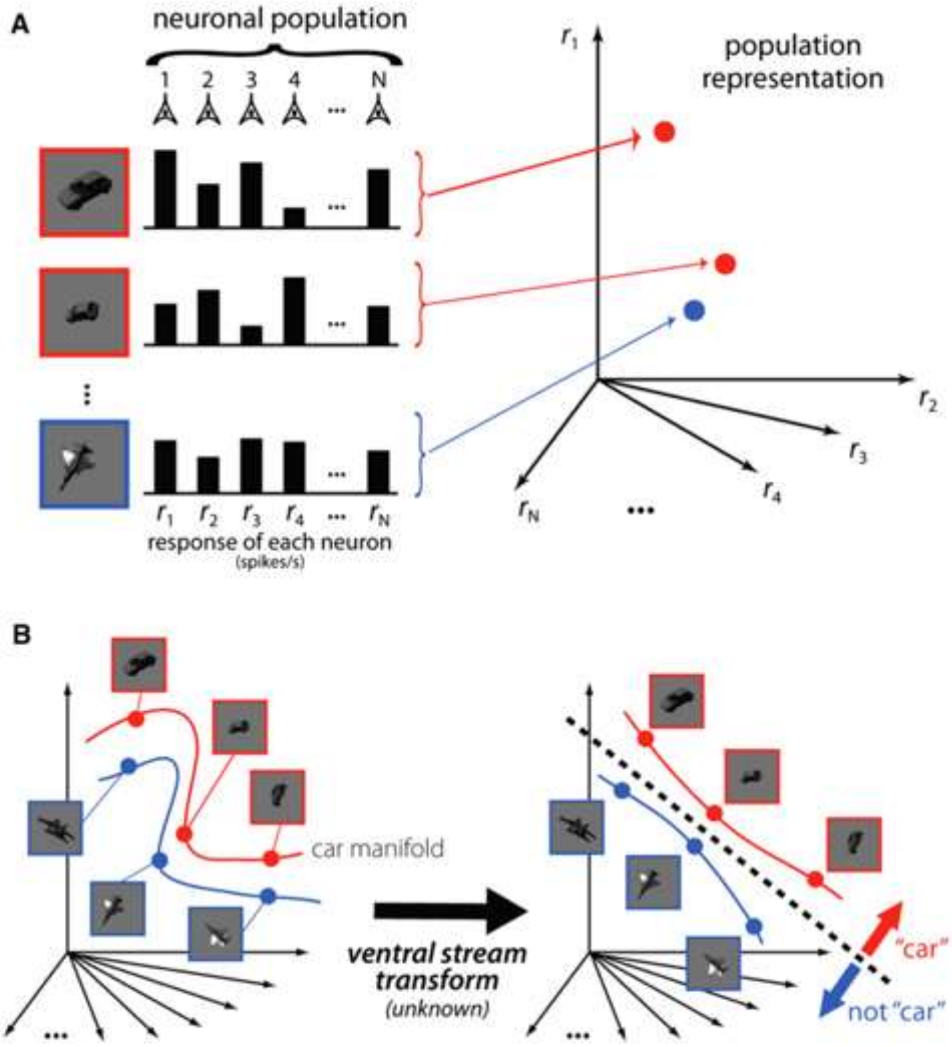
- IT neurons use a distributed representation: object identity is encoded across populations of neurons. Yet each neuron may have a sparse and relatively specific response profile—so the code is neither purely local nor fully distributed.

- Responses of IT neurons are shaped by experience and learning. Neurons can develop selectivity for new categories or become more finely tuned with exposure (e.g., training with novel shapes or objects). This supports category learning, memory, and conceptual abstraction.

- IT neurons fire ~100-150 ms after stimulus onset, showing that they're at the end of the ventral visual stream. This timing reflects a feedforward cascade through V1 → V2 → V4 → IT, with progressively more abstract representations.

- Some IT neurons correlate with object-related decisions, attention, or memory, especially in delayed match-to-sample tasks. This blurs the line between perception and cognition—IT is not just about visual features, but meaningful representations.

Ventral visual pathway gradually "untangles" information about object identity

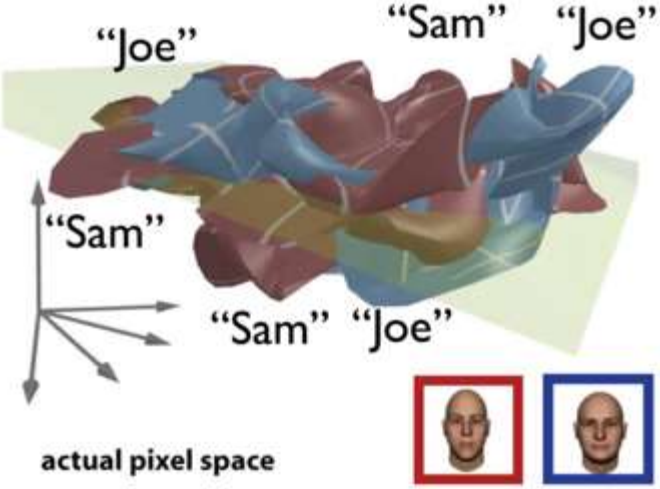


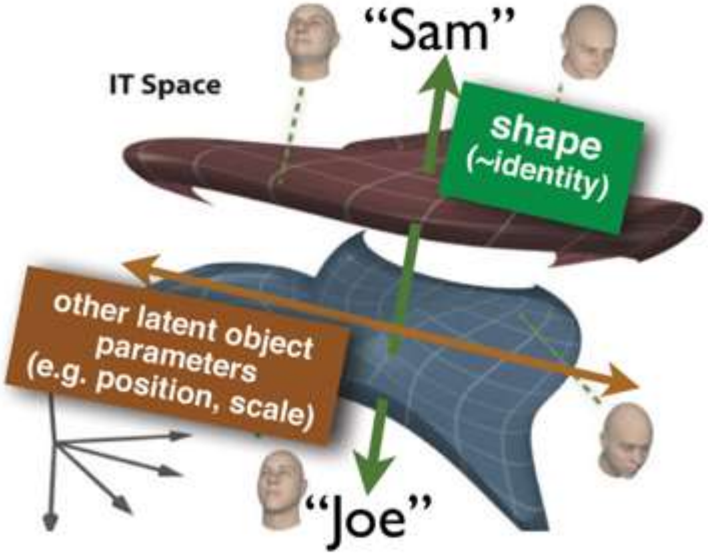
DiCarlo et al., Neuron, 2012

Response of a population of neurons to a particular view of one object can be represented by a **response vector** in a space whose dimensionality is defined by the number of neurons in the population.

When an object undergoes an identity-preserving transformation, it produces a different pattern of population activity, which corresponds to a different response vector.

Together, the response vectors corresponding to all possible identity preserving transformations define a low-dimensional surface in this high-dimensional space—an **object identity manifold.**



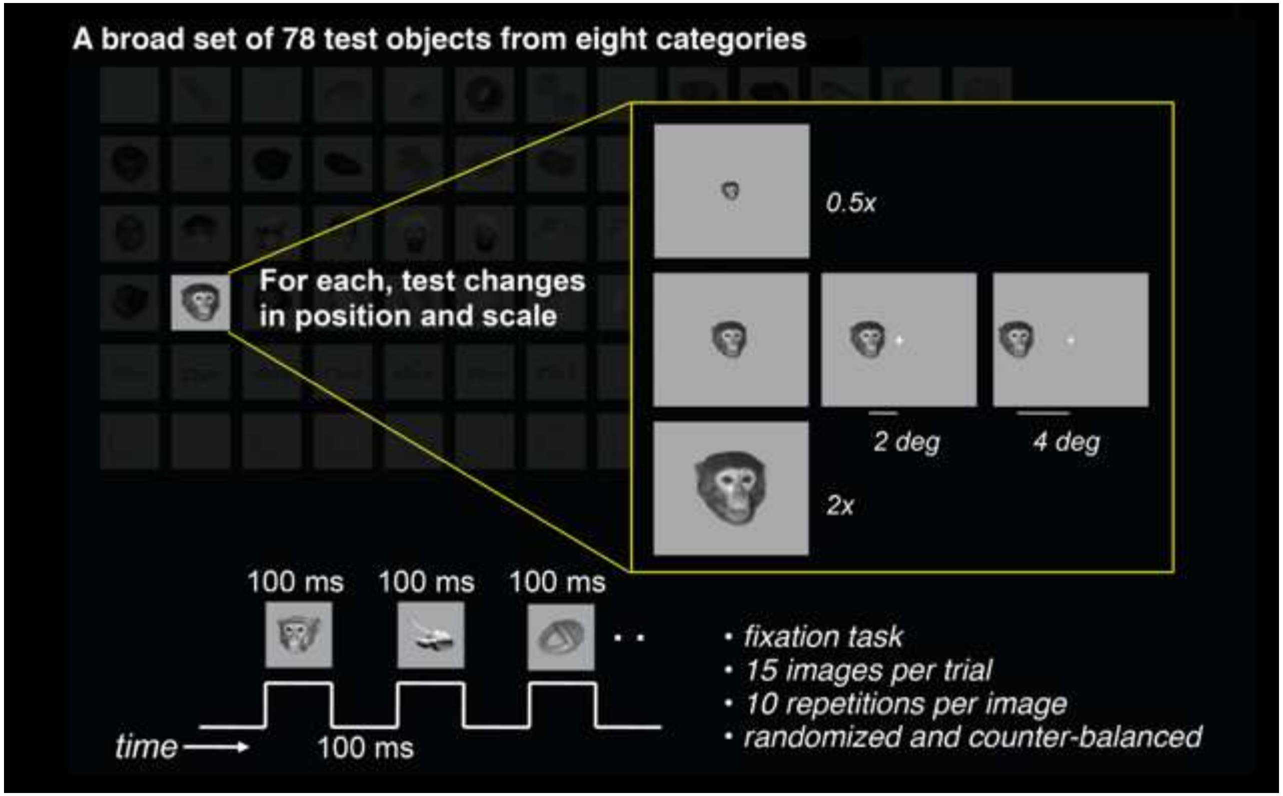


Object recognition is the ability to separate representation that contain one particular object from representation that do not.

Thus, object manifolds are thought to be gradually untangled through nonlinear selectivity and invariance computations applied at each stage of the ventral pathway.

At higher stages of visual processing, neurons tend to maintain their selectivity for objects across changes in view; this translates to manifolds that are more flat and separated (more ''untangled'').

DiCarlo et al., Neuron, 2012


A broad set of 78 test objects from eight categories
For each, test changes in position and scale
0.5x
2 deg
4 deg
2x
100 ms
100 ms
100 ms
time
100 ms
• fixation task
• 15 images per trial
• 10 repetitions per image
• randomized and counter-balanced

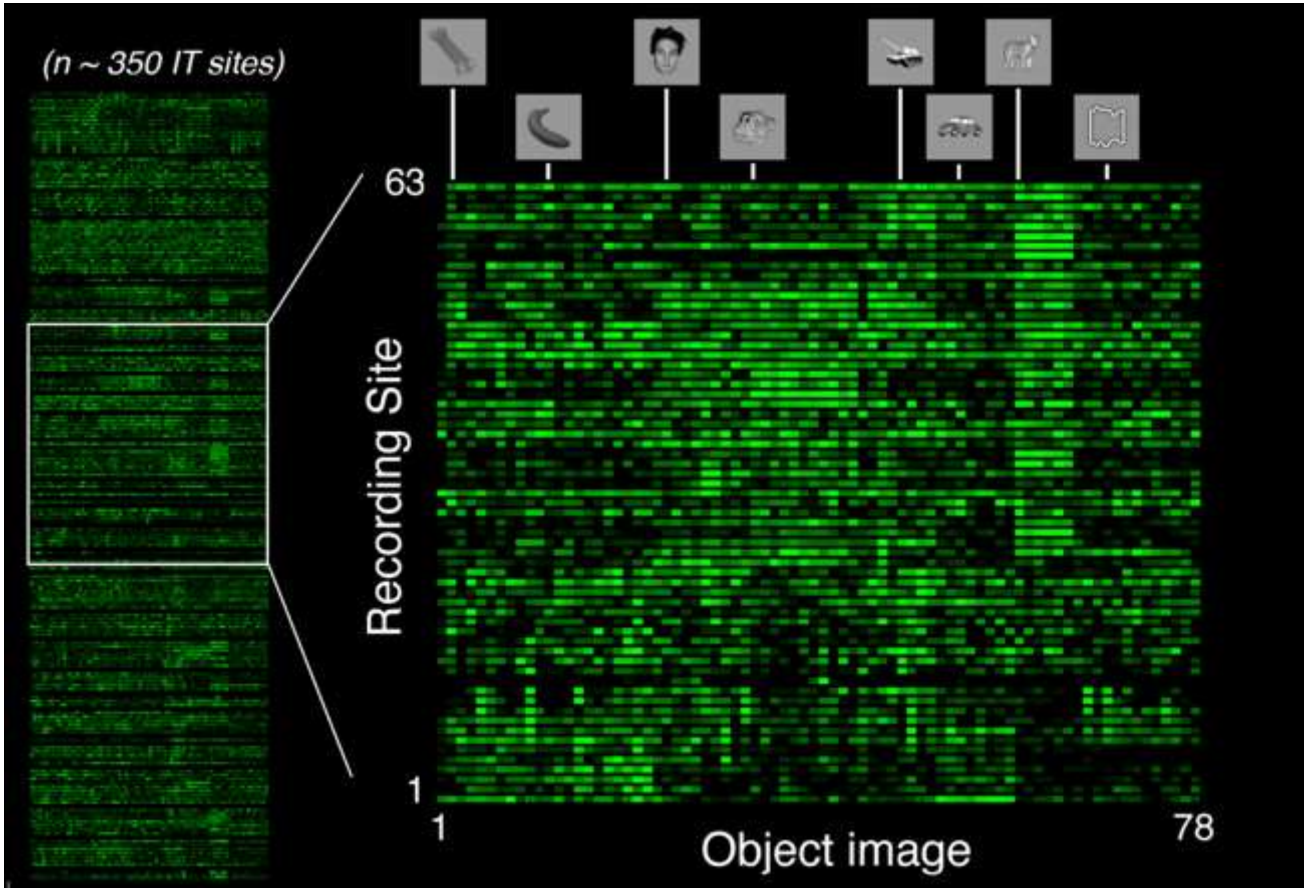
Categories: toys, food, human faces, monkey faces, hand/body, vehicles, white boxes, cats/dogs.

Readout of object Identity from IT cortex

By using a classifier-based readout technique, Hung et al (2005) investigated the neural coding of **selectivity and invariance** at the IT population level.

They showed that the activity of small neuronal populations (~ 300 units) over very short time intervals (as small as 12.5 milliseconds) contain accurate and robust information about both object ''identity'' and ''category.''

Hung et al., Science, 2005



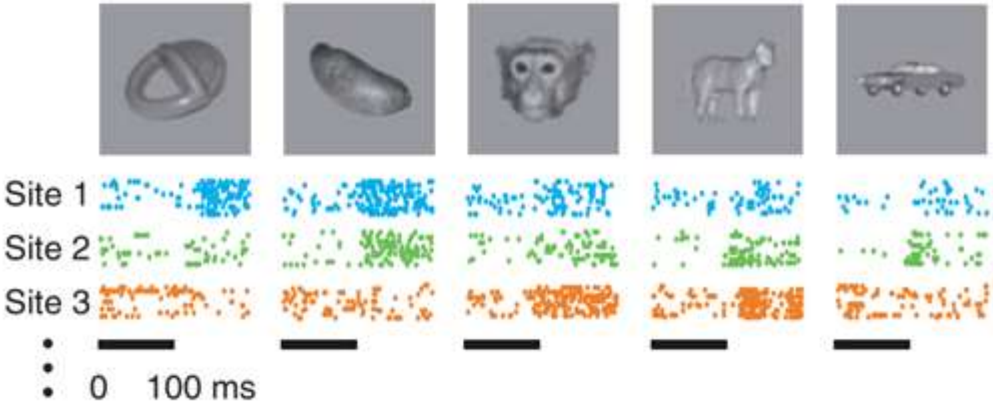
The readout technique consists of training a regularization classifier to learn the map from neuronal responses (from the independently recorded neurons) to each object label.

The input consists of the neuronal responses of IT neurons

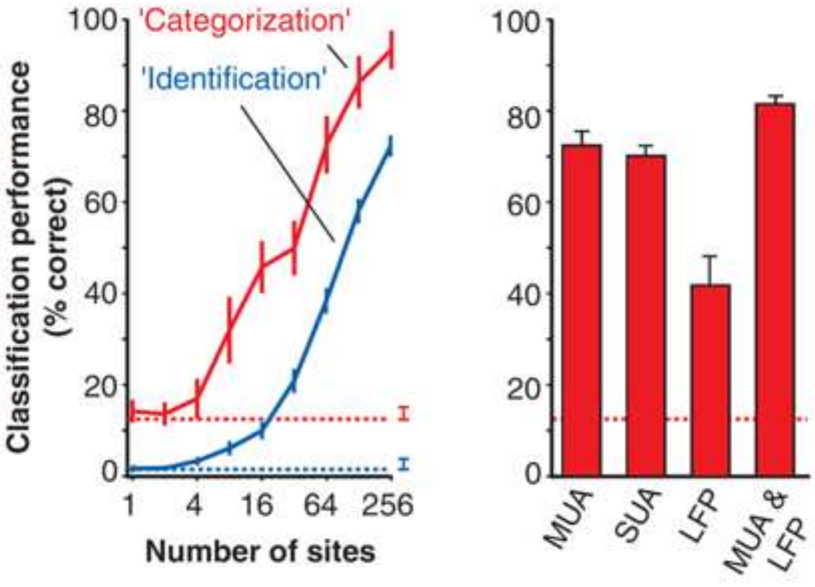
.

The activity of ~ 250 randomly selected multi- and single-unit activity in response to 78 images of different objects were recorded at 350 IT sites in two monkeys

Hung et al., Science, 2005

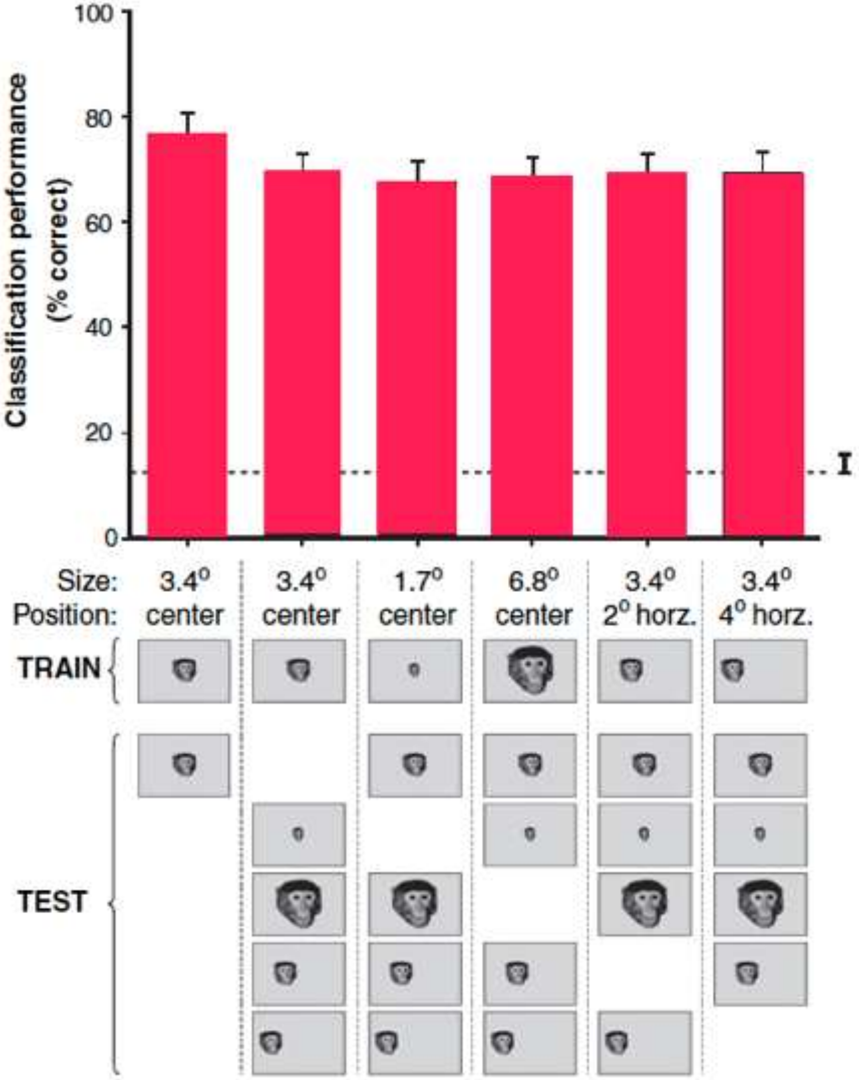


The spiking activity of 256 randomly selected multi-unit activity sites was sufficient to categorize the objects with 94% accuracy.



Classifier performance increased approximately linearly with the **logarithm of the number of sites**, which is indicative of a distributed representation in contrast to a grandmother-like representation.

Hung et al., Science, 2005



After training, the classifier was used to decode the responses to novel stimuli.

The classifier performance is maintained high over a range of object positions and scales, even for novel objects.

A one-versus-all approach was used whereby for each class of stimuli (8 classes for categorization, 78 classes for identification, 3 classes for scale and position readout), one binary classifier was trained.

Recognition performance is what downstream neurons (for instance PFC neurons) could, in theory, perform by simply computing a weighted sum of IT spikes over a short time interval (100- to 300-ms interval divided into bins of 50 ms in this case).

This is notable considering the high trial-to-trial variability of cortical neurons.

Hung et al., Science, 2005

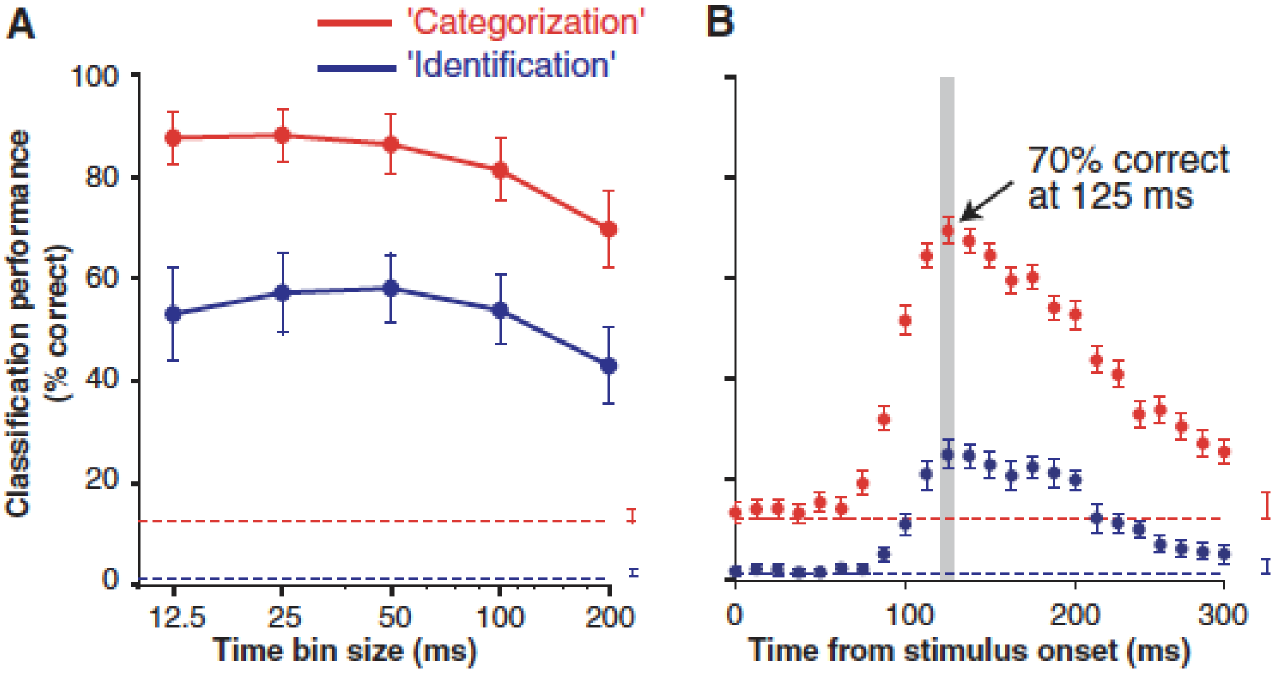
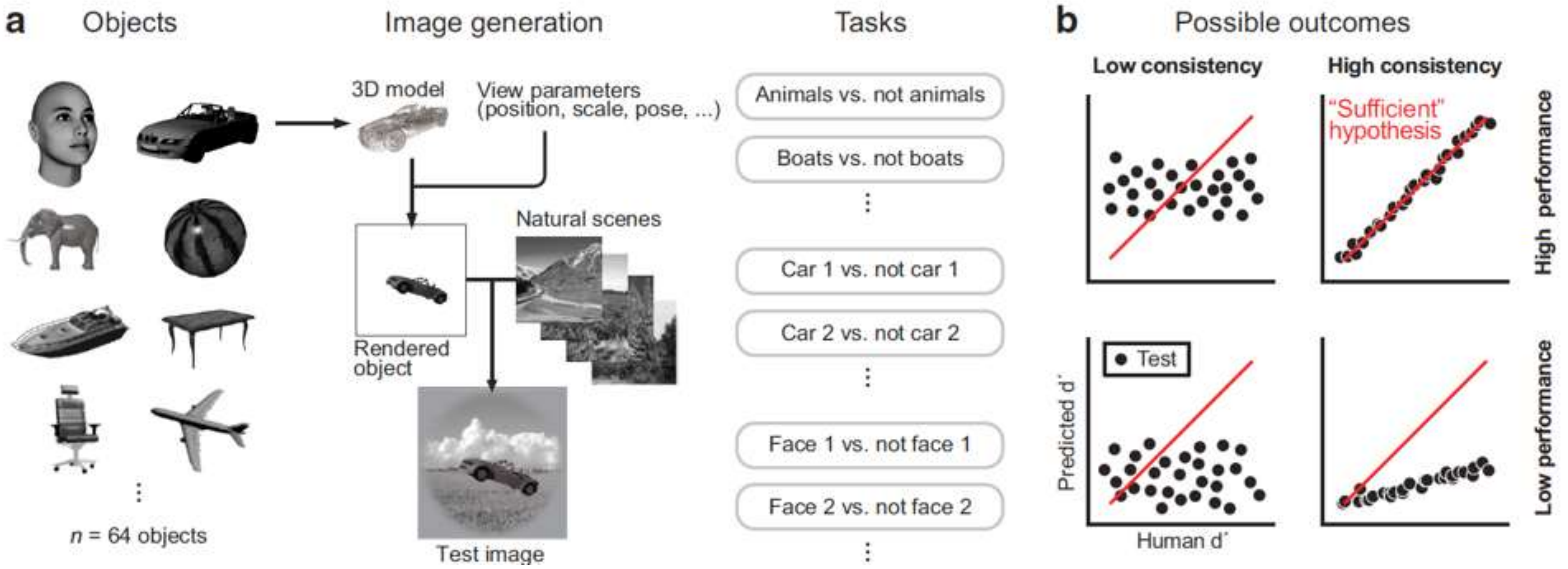


Fig. 3. Latency and time resolution of the neural code. (A) Classification performance (n = 128 sites) as a function of the bin size (12.5 to 200 ms, i.e., temporal resolution) to count spikes within the 100- to 300-ms window after stimulus onset for categorization (red) and identification (blue). The same linear classifier as in Figs. 1 and 2 was used. (B) Classification performance (n = 256 sites) using a single bin of 12.5 ms to train and test the classifier at different latencies from stimulus onset (x axis). The colors and conventions are as in Fig. 1B.

Hung et al., Science, 2005

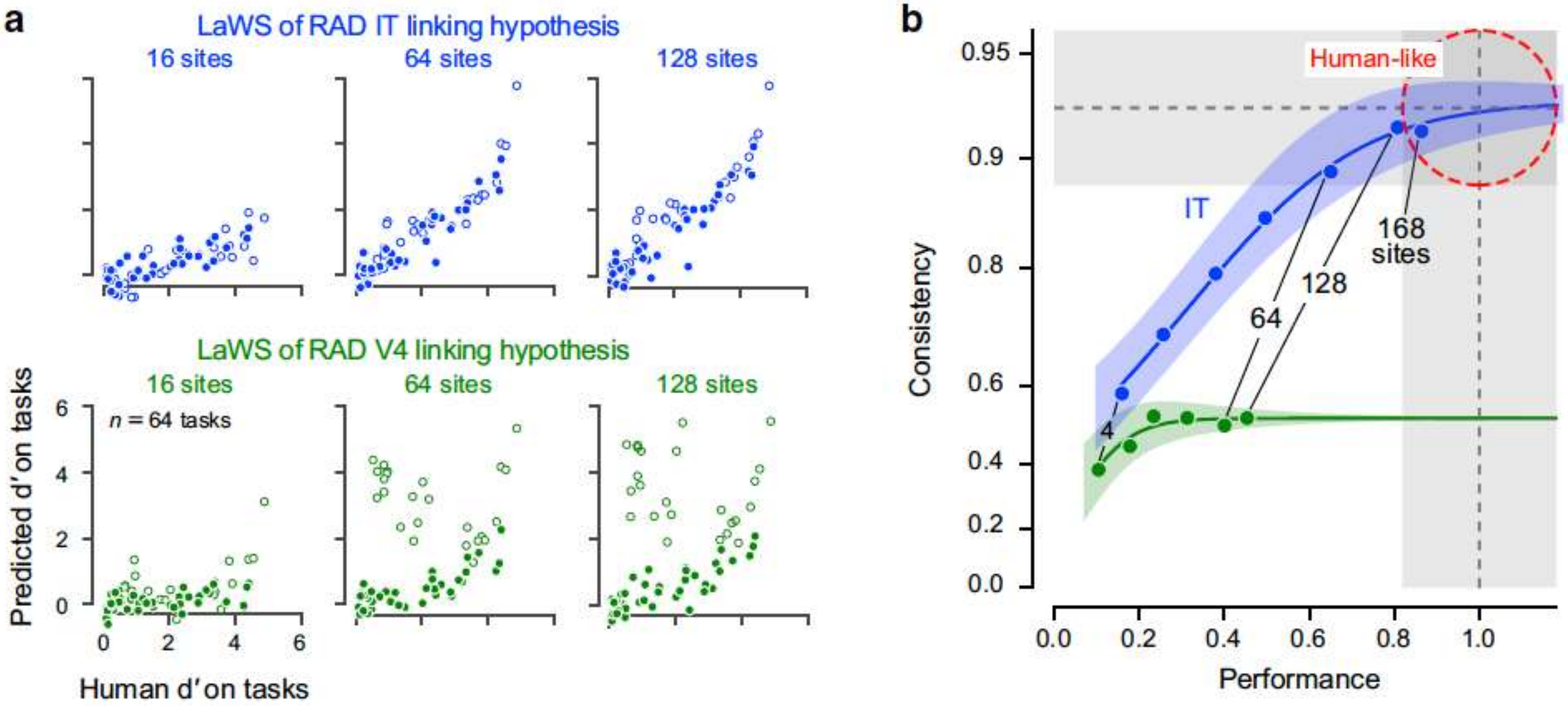
Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance



To determine whether models based on the firing rates of neurons in monkey V4 and IT cortex could predict human object recognition behavior.

They asked whether simple linear classifiers (weighted sums) trained on these neural responses can match human-level performance on object recognition tasks

Majaj et al., J. Neurosci, 2015

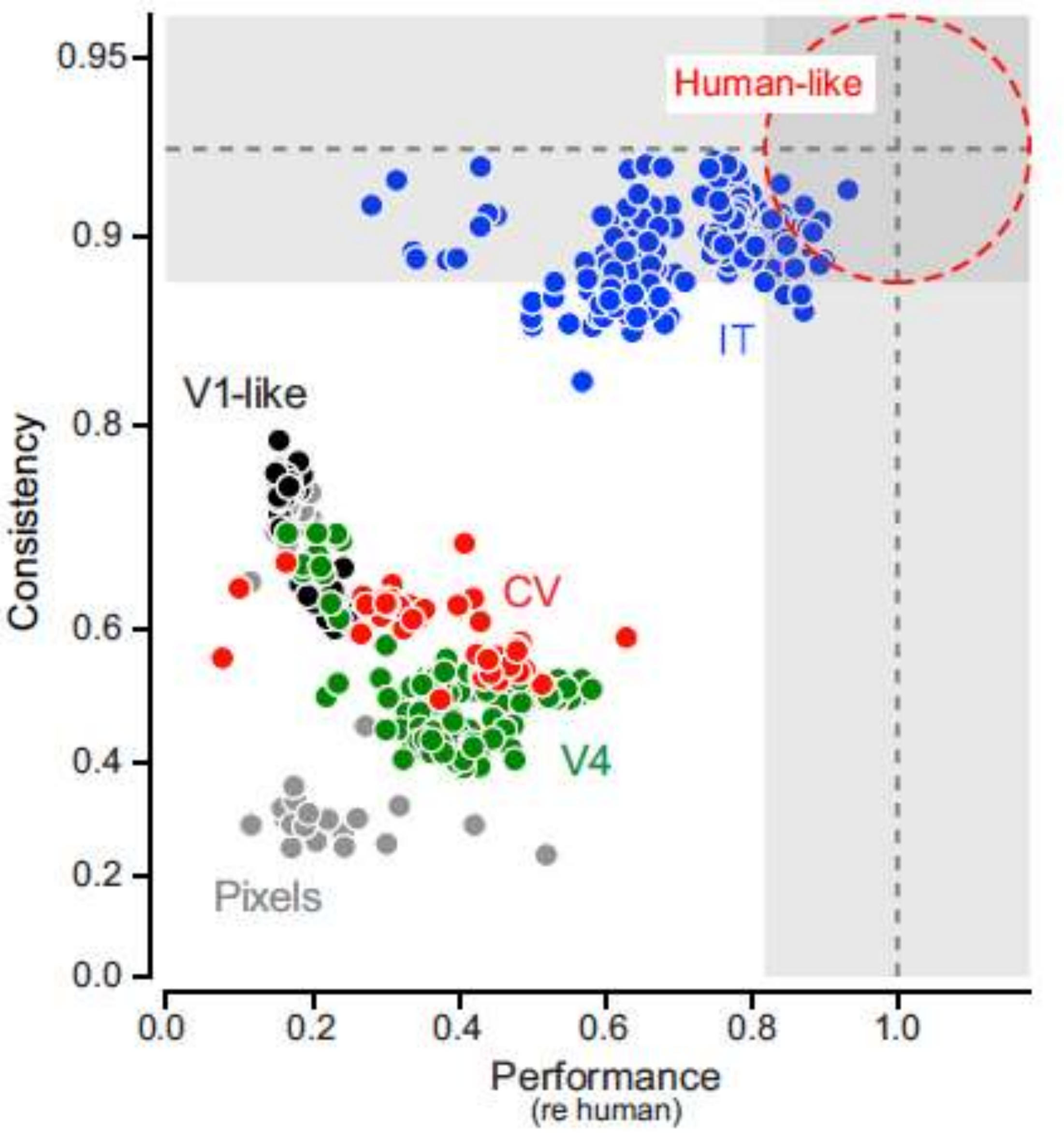


They measured how decoding accuracy improved as they added more IT neurons. As expected, more neurons = better prediction.

By contrast, adding more V4 neuron did not improve decoding accuracy.

LaWS of RAD = Linear Weighted Summation of Responses After Delay

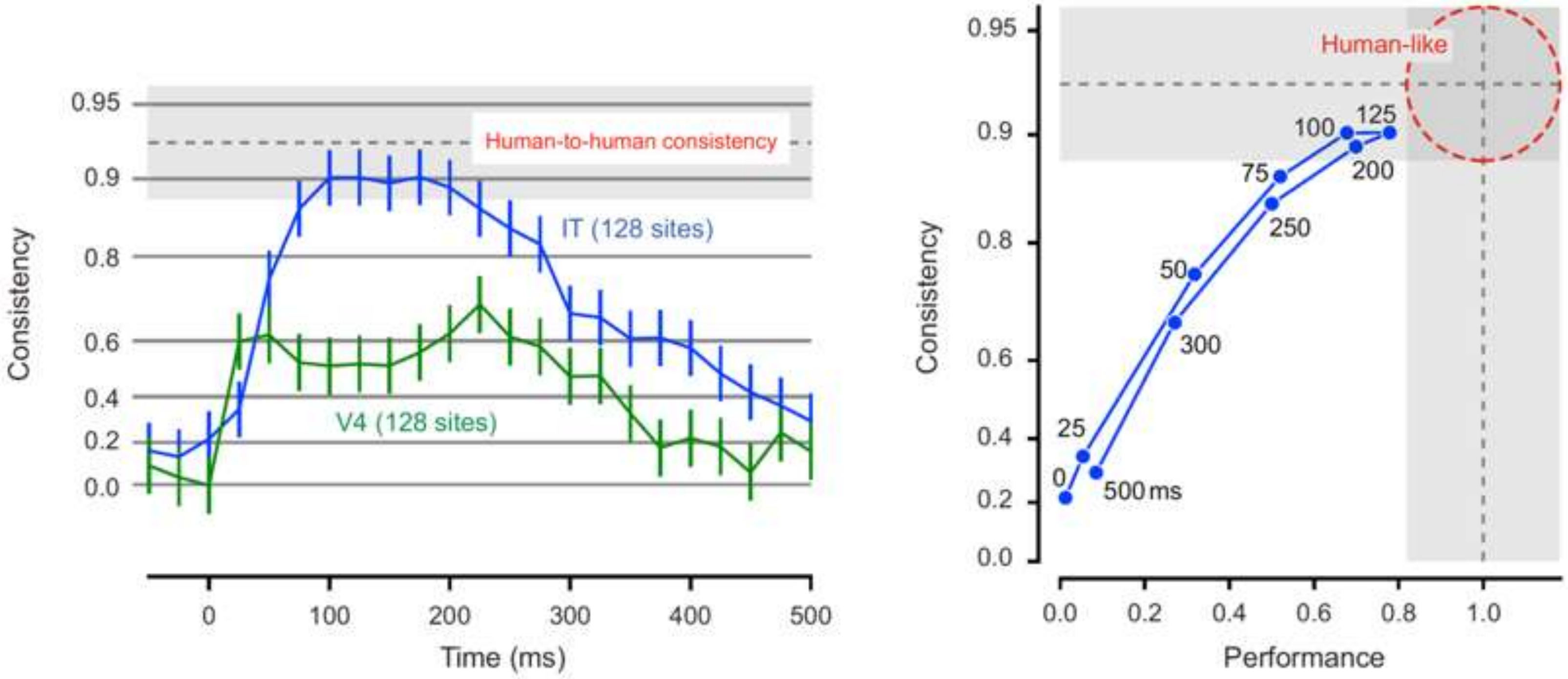
Majaj et al., J. Neurosci, 2015



Majaj et al. tested 944 different models (linking hypotheses) to see how well neural data predicts human object recognition. Each model's **consistency** (match to human patterns) and **performance** (accuracy compared to humans) were plotted.

IT-based models (in blue) showed the best results, especially when more neurons were used, often coming close to or matching human performance. The ideal models both matched human accuracy and predicted which tasks were easier or harder for humans. This highlights that only models using rich IT data can fully explain human object recognition behavior.

Majaj et al., J. Neurosci, 2015



Majaj et al. showed that decoding IT activity depends on when you measure it. Using fixed 100-ms windows from 0 to 500 ms after image onset, they found that certain time windows yield higher consistency with human behavior. The best predictions came from neural activity shortly after image onset, revealing an optimal temporal window for matching human object recognition.

Majaj et al., J. Neurosci, 2015