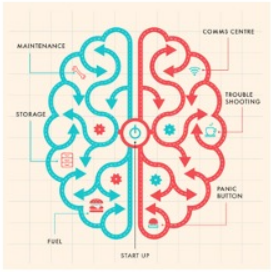
Object recognition
in ventral visual cortex and deep networks



Giuseppe di Pellegrino
Department of Psychology, University of Bologna

g.dipellegrino@unibo.it

Cognition and Neuroscience
Second cycle Degree in Artificial Intelligence – 2032/24

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).

1

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.

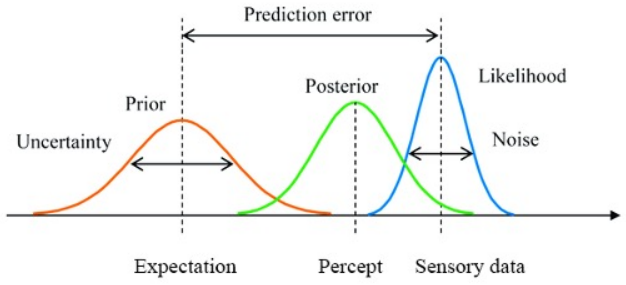
2

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).

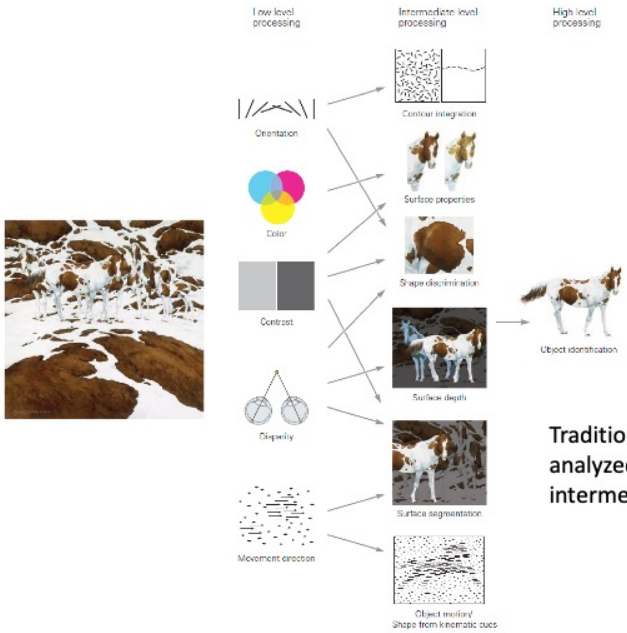
Is it reasonable to assume that the visual system knows the probability calculus and operates according to it?

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



Prior probability distributions in typical applications of the Bayesian strategy represent knowledge of the regularities governing object shapes, constituent materials, and illumination, and likelihood distributions represent knowledge of how images are formed through projection on the retina. Some examples of prior knowledge are that solids are more likely to be convex than concave and that the light source is above the viewer.

The more ambiguous the image – the greater the influence of prior knowledge in yielding a nonambiguous percept.
Some perceptions may be more data-driven, others more prior knowledge driven.



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.

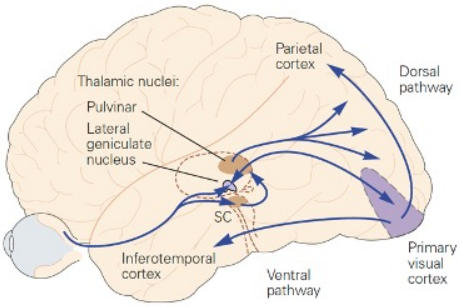
Intermediate-level processing: low-level features are used to parse the visual scene.
Local orientation is integrated into global contours (contour integration); local visual features are assembled into surfaces, objects are segregated from background (surface segmentation), surface shape is identified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

A visual scene comprises many thousands of line segments and local surface patches.
Intermediate-level visual processing is concerned with determining which boundaries and surfaces belong to specific objects and which are part of the background

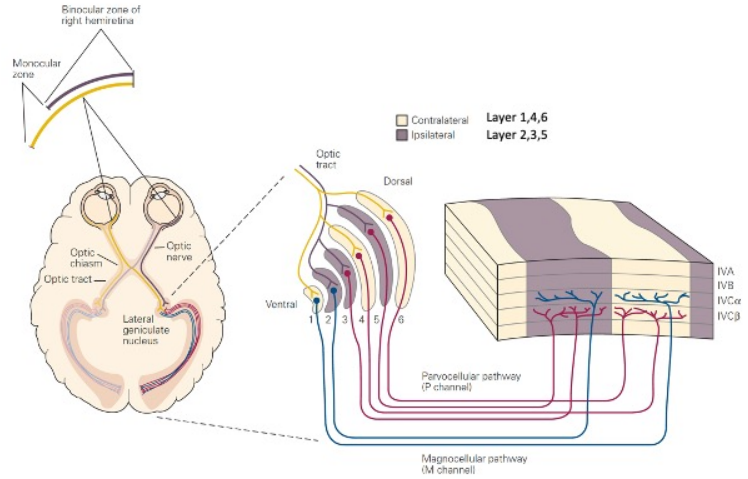
Visual processing is mediated by the retino-geniculo-striate pathway



This pathway includes:

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

6



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.

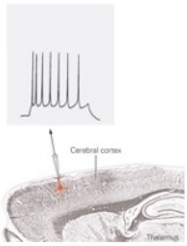
7

Single-cell recording

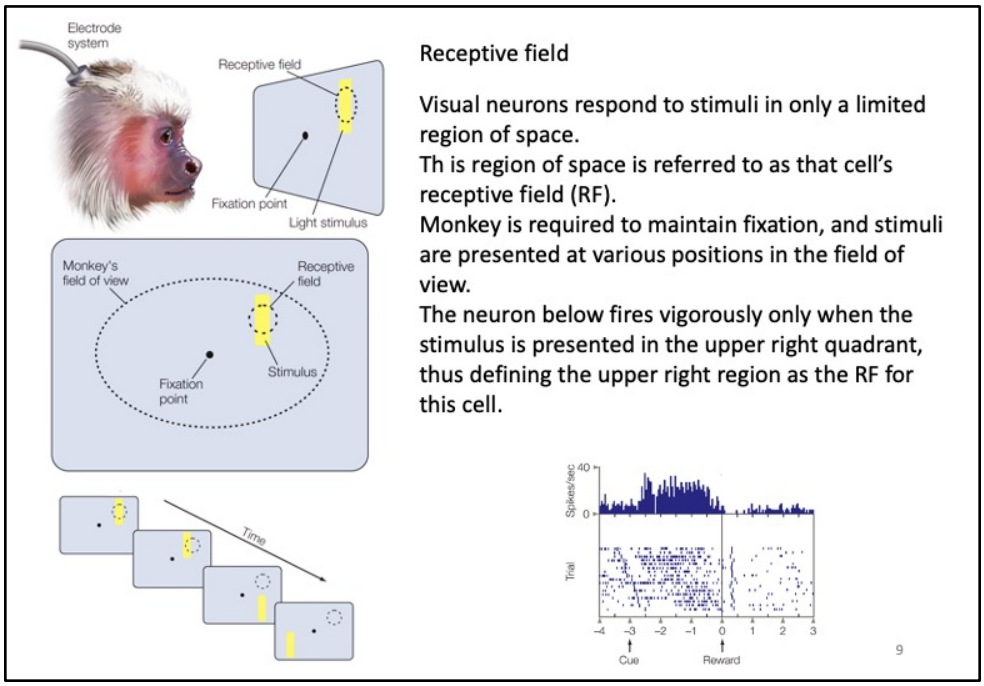
This technique allows recording signals (firing rate) from single neurons.

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.

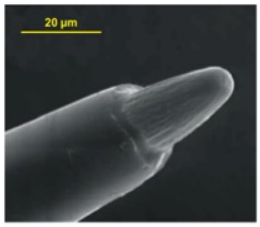
Disadvantages
– invasive
Advantages
– high spatial and temporal resolution
– differentiation between excitation and inhibition



8

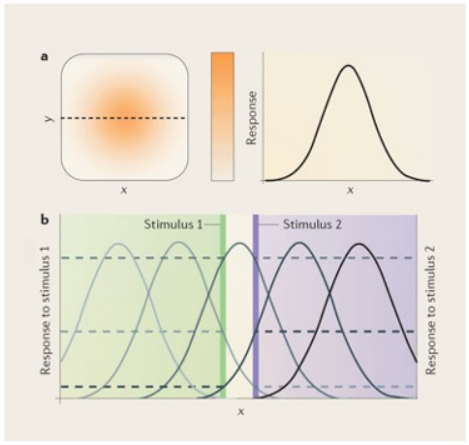


The concept of RF was introduced in 1906 by Charles Sherrington. The receptive field is a characteristic of all neurons and, in vision, indicates the region of the visual scene where the stimulus must fall to excite or inhibit the neuron being studied.



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.

10



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.

11

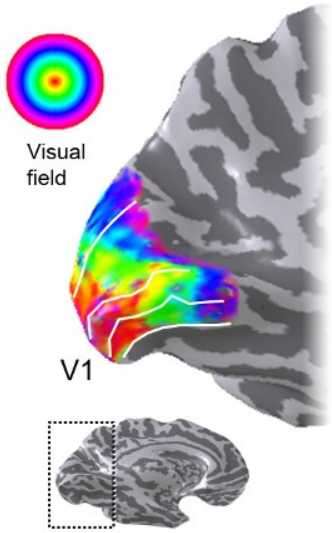
Figure b, RF profiles of a set of five neurons with different RF locations. Horizontal dashed lines indicate the response of these five example neurons to two stimuli at nearby locations (vertical green and purple lines). Both stimuli fall into the same RF (middle grey curve), but they stimulate neurons with neighbouring RFs differently so that the population can resolve the two locations even though a single neuron cannot. In addition, the size of the RF determines the neuron's spatial frequency tuning: the smaller the RF, the higher the spatial frequency it can resolve.

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.

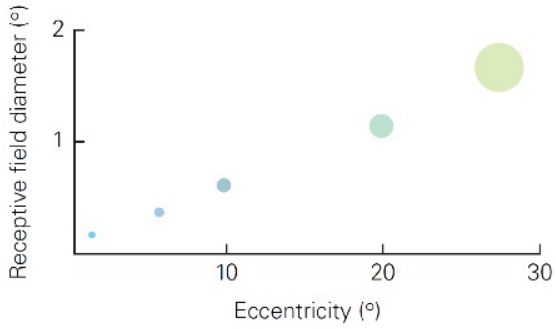
Neuron RFs form a 2D map of the visual field, such that neighbouring regions in the visual image (and therefore on the retinal surface) are represented by adjacent regions of the visual cortical area (i.e., orderly mapping of RF positions in retinotopic coordinates)



12

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

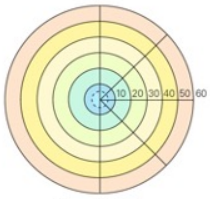
13

The amount of cortex devoted to one degree of viewing angle changes with eccentricity.

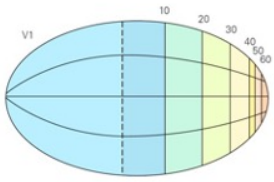
Accordingly, more cortical space is dedicated to the central part of the visual field, where the receptive fields are smaller and densely packed and the visual system has the highest spatial resolution.

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

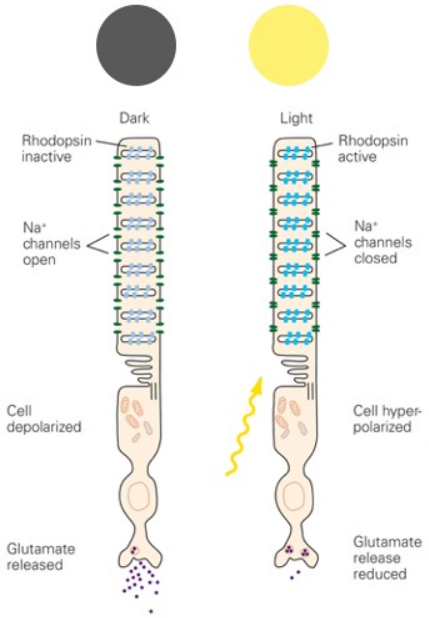


Primary visual cortex

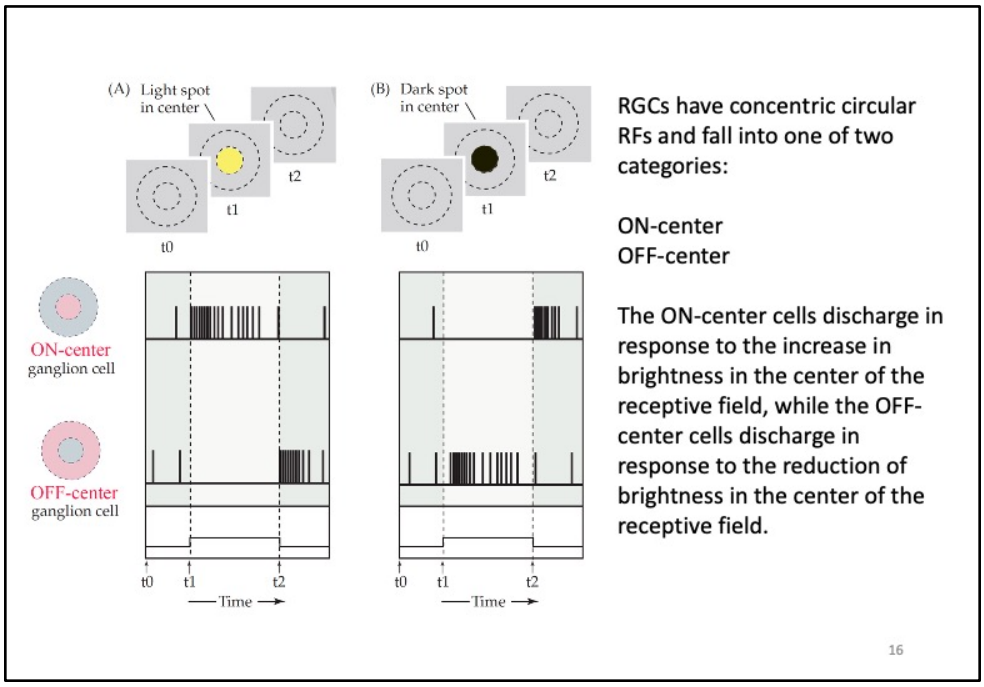
14

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.



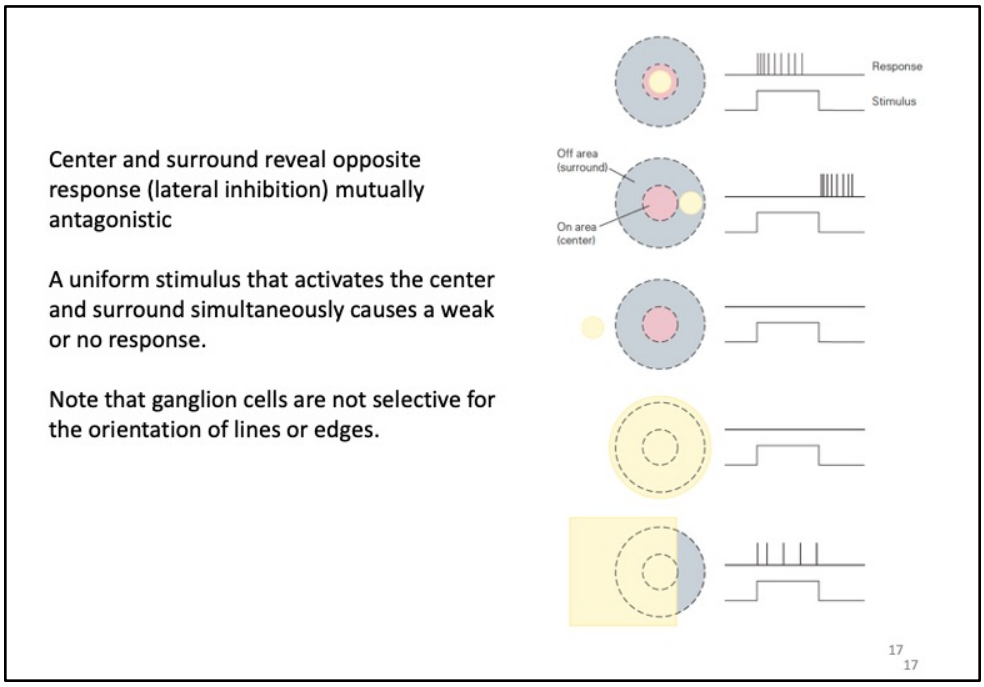
15



ON-Center ganglion cells are excited by a light stimulus in the center of the receptive field;
OFF-Center ganglion cells are excited by a dark stimulus in the center of the receptive field;

Note that the firing rate of ON-center ganglion cells increases soon after the dark stimulus disappears;

Similarly, the discharge rate of OFF-center ganglion cells increases soon after the disappearance of the light stimulus;

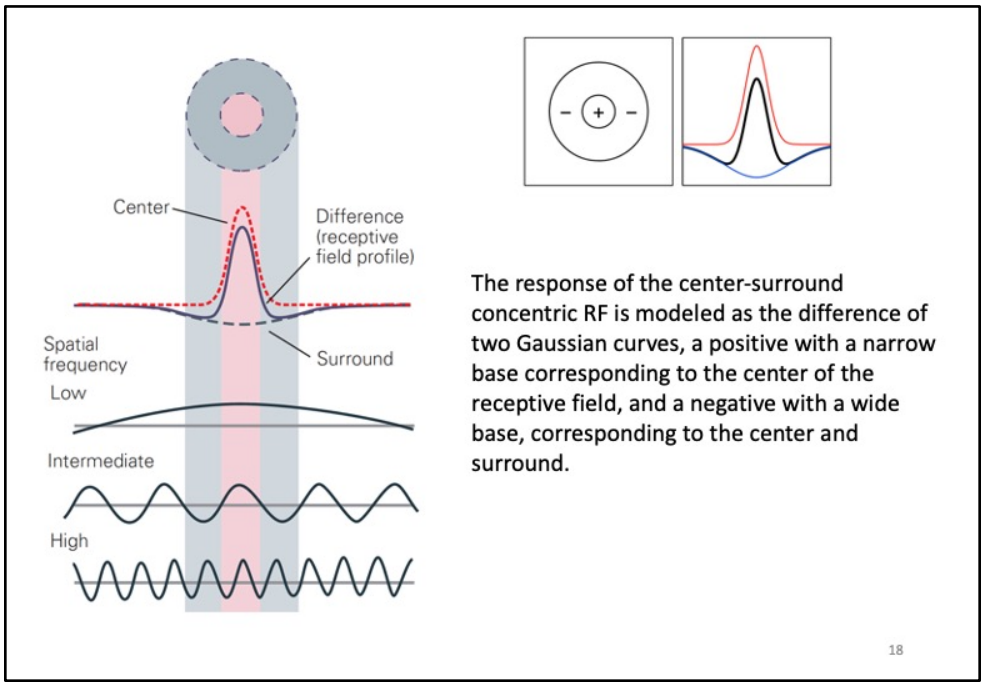


The retinal ganglion cells have an organization of the receptive field with two concentric circular areas with opposite and antagonistic response.

In ON-center cells, the illumination of the central part of the receptive field causes an excitatory response, i.e. an increase in the discharge of the cell, while the illumination of the surrounding part of the receptive field causes an inhibitory response (mechanism of lateral inhibition).

The OFF-center cells are instead organized in the opposite way: the illumination of the surrounding area causes an excitatory response, while the illumination of the central part of the receptive field causes an inhibitory response.

The simultaneous illumination (or darkness) of the center and surround does not evoke a variation in the discharge frequency.

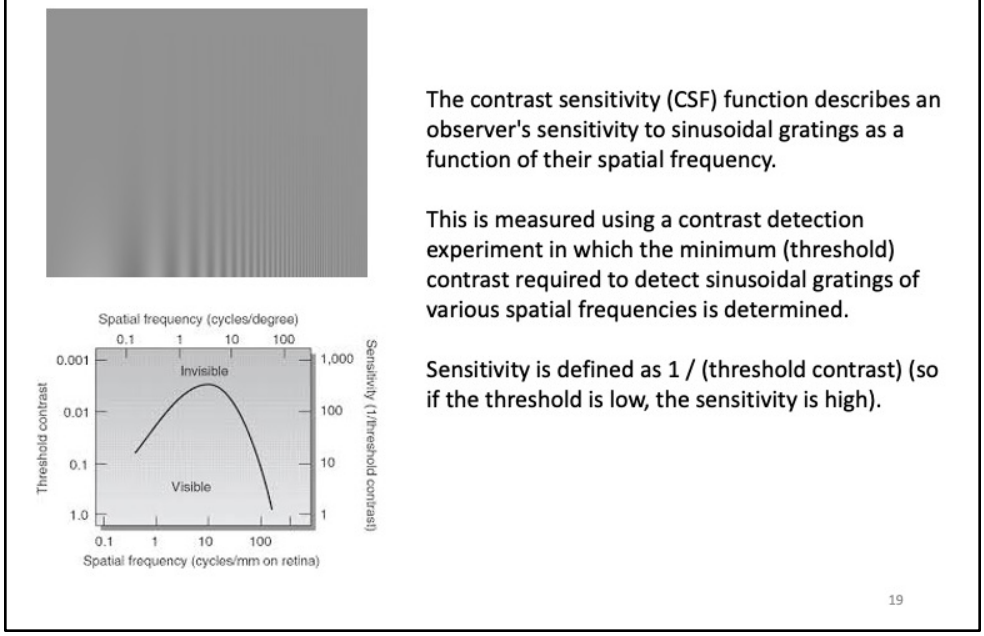


In humans, if sinusoidal gratings are used, sensitivity is greater for spatial frequencies around 5-8 cycles / visual degree, and is attenuated both for higher frequencies (up to acuity around 30-50 cycles / degree) and for frequencies less than 1 cycle / degree.

Multiplying the profile of the grating stimulus (intensity vs position) with the profile of the receptive field (sensitivity vs position) and integrating over all space calculates the stimulus strength delivered by a particular grating.

In day light, contrast sensitivity declines sharply at high spatial frequencies, with an absolute threshold at approximately 50 cycles per degree.

Interestingly, sensitivity also declines at low spatial frequencies. The attenuation at low frequencies reflects the inhibitory and antagonistic action of the periphery (surround) of the receptive fields of the retina, geniculate and cortex.
Patterns with a frequency of approximately 5 cycles per degree are most visible.
The visual system is said to have band-pass behavior because it rejects all but a band of spatial frequencies.

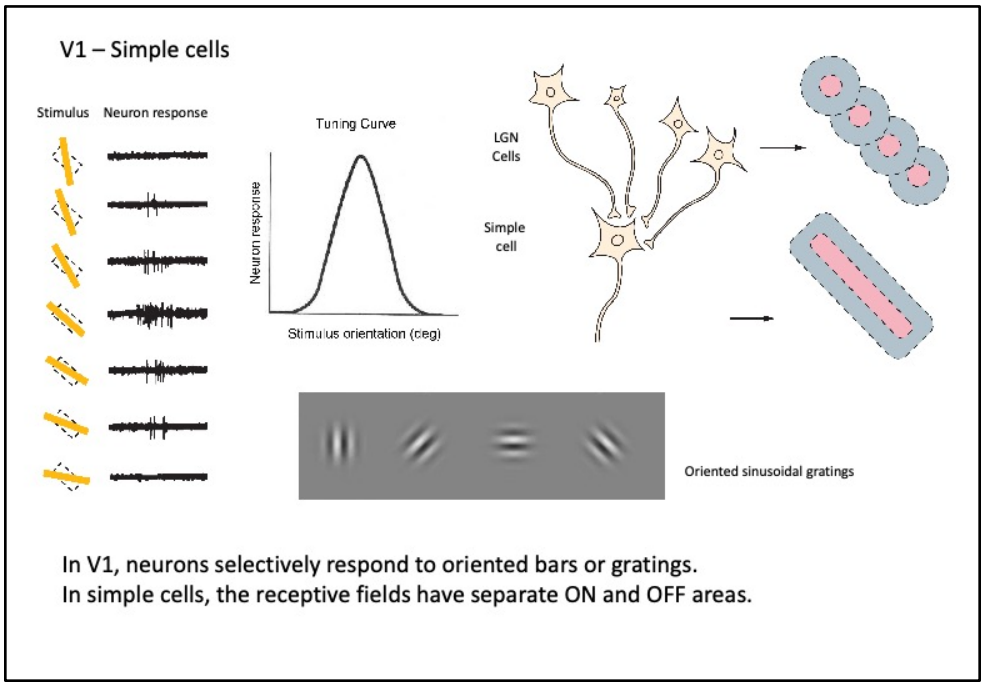


Humans are more sensitive to an intermediate range of spatial frequencies (about 4-6 cycles / degree) and less sensitive to both lower and higher space frequencies.

Gratings with a frequency of about 5 cycles per degree are the most visible. The visual system is said to have bandpass behavior because it rejects everything but a narrow band of spatial frequencies.

In the figure above, the stimulus contrast increases from top to bottom, while the spatial frequency increases from left to right.

The central bars in the figure (medium spatial frequency) are visible even at low contrast, while the wide bars and narrow bars are visible only at high contrast.



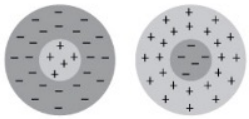
Neurons in area V1 are classically divided into two types: simple and complex (Hubel and Wiesel, 1959).
Neurons have elongated RFs and respond to a narrow range of orientations.
Different neurons respond optimally to distinct orientations (orientation tuning curve).
Example of a neuron in area V1 that selectively responds to lines that adapt to the orientation of its receptive field.
This selectivity is the first step in the brain's analysis of the shape of an object.

The orientation of the receptive field is thought to result from the alignment of the center-surround circular receptive fields of different LGN cells.

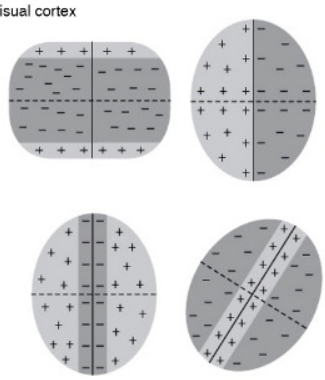
In the monkey, the neurons of the LGN have non-oriented circular receptive fields.
However, projections of adjacent LGN cells onto a simple cell create a receptive field with a specific orientation.

Simple cells respond well to sinusoidal gratings (Gabor patches) of specific spatial frequencies and phases.

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;

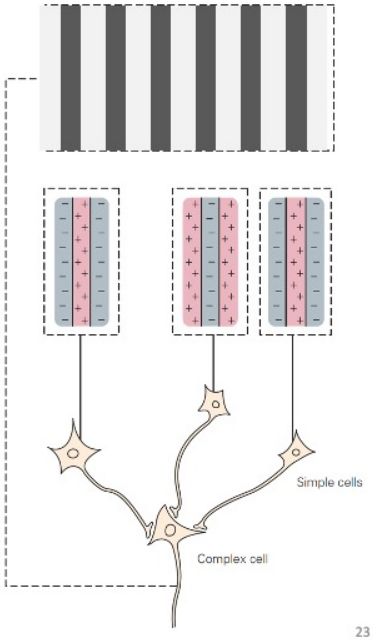
Movement of the stimulus in the receptive field is particularly effective in activating the cells;

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.



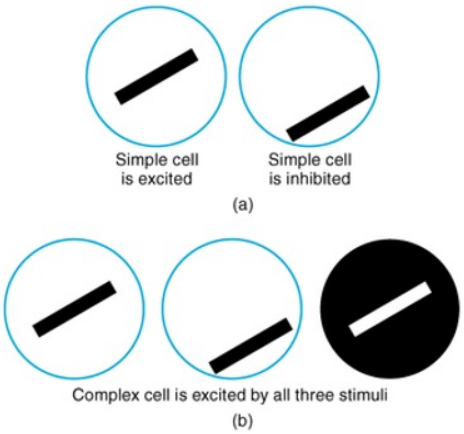
23

Complex cells are less selective for the position of the stimulus in the receptive field

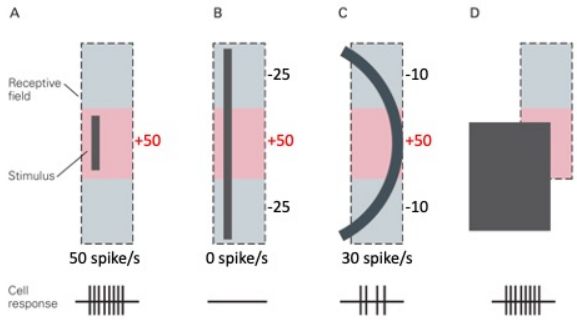
The receptive field has no defined ON and OFF regions and responds similarly to light (on a dark background) or dark (on a light background) stimuli in all positions of the receptive field.

They are activated as a linear oriented stimulus crosses their receptive fields in one direction.

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



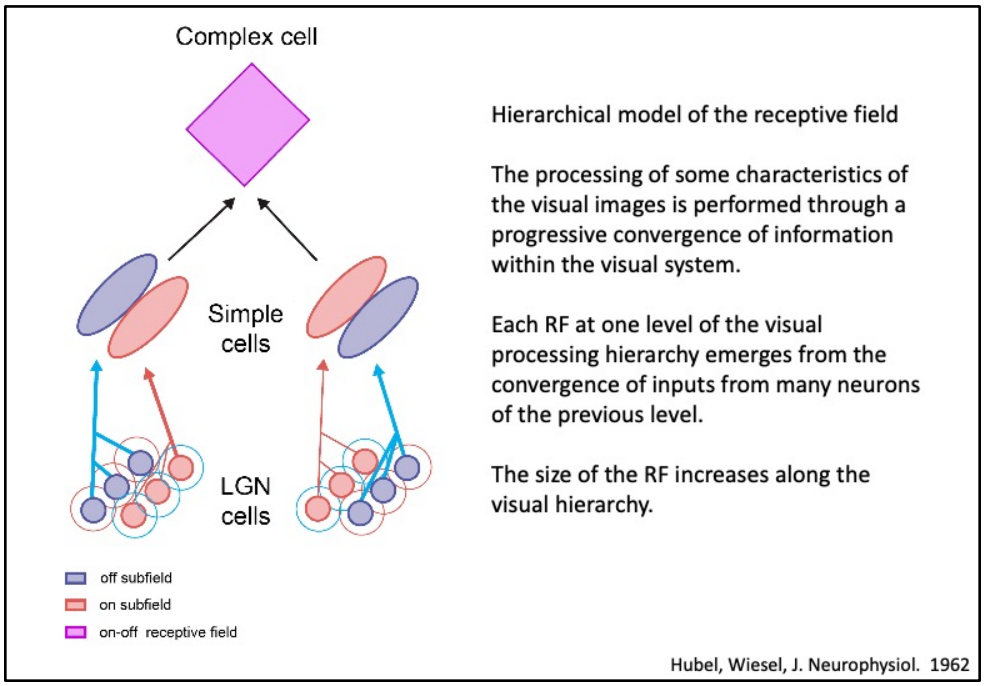
End-stopped cells



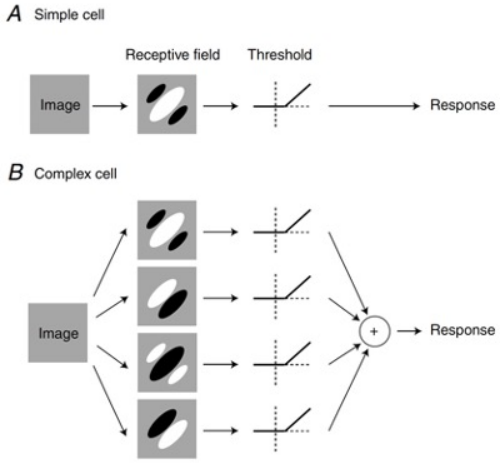
The receptive fields of some cells have a central excitatory region flanked by inhibitory regions that have the same preferential orientation.
A short linear segment (A) or a long curved line (C) will be effective in activating the neuron, but not a long straight line (B).
A neuron with a receptive field that has only one inhibitory region in addition to the excitatory region can signal the presence of angles (D).

Respond better to linear stimuli of a certain length, or that have an end that does not extend beyond a specific portion of the cell's receptive field.
End-stopped may serve to detect angles (“angle-detectors”) or curved lines of visual images.

ON and OFF regions of the RF have the same preferred orientation (vertical, in the neuron illustrated in the figure).
Therefore, the inhibitory effect is greater if the same oriented contour is presented both in the ON and OFF regions.
A short linear segment (A), or a long curved line (C) will be effective in activating the neuron, because excitation will be greater than inhibition.
On the contrary, a long straight line (B) will not be effective, because excitation will be canceled by the inhibitory effect.



According to the hierarchical model (Hubel and Wiesel, 1962), simple cell receptive fields are constructed from the convergence of geniculate inputs with receptive fields aligned in the visual space.
In turn, complex receptive fields arise from the convergence of simple cells with similar orientation preferences.



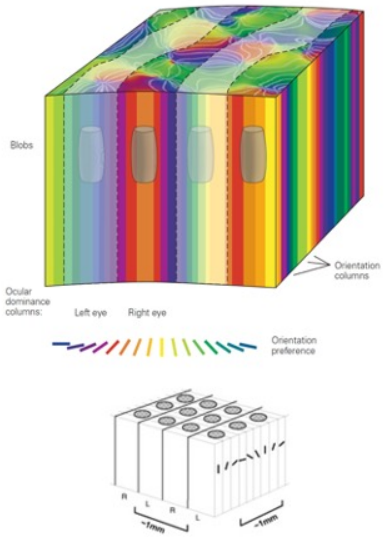
The models of simple and complex cells proposed by Movshon, Thompson and Tolhurst (Movshon et al. 1978)

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.

27

Ice cube model (Hubel e Wiesel, 1977)



A region of cortical tissue of about 1mm contains two orientation hypercolumns (each representing 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.

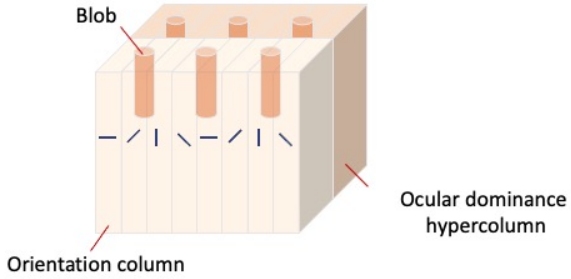
This computational module contains all the anatomical-functional types of V1 neurons, and would be repeated thousands of times to cover the entire surface of the visual field.

28

However, it remains unclear what advantage, if any, is conveyed by this form of columnar segregation.
One candidate function for cortical columns is the minimization of connection lengths and processing time, which could be evolutionarily important;

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

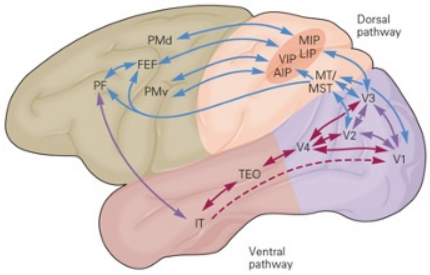
orientation system
ocular dominance system



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.

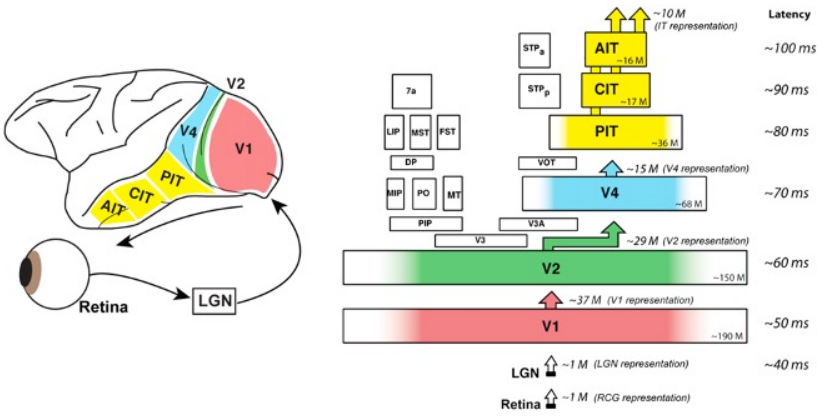


Ungerleider & Mishkin, Two cortical visual systems. 1982

The dorsal and ventral pathways are highly interconnected so that information is shared.

For example, stimulus movement information in the dorsal pathway (area V5) can contribute to object recognition through kinematic cues. Information about movements in space derived from areas in the dorsal pathway is therefore important for the perception of object shape and is fed into the ventral pathway.

Note: all connections between areas in the ventral and dorsal pathways are reciprocal: each area sends information to the areas from which it receives input.
Reciprocity is an important feature of connectivity between cortical areas



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.

31

In the macaque monkey, V4 is located on the prelunate gyrus and in the depths of the lunate and superior temporal sulci and extends to the surface of the temporal-occipital gyrus.

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).

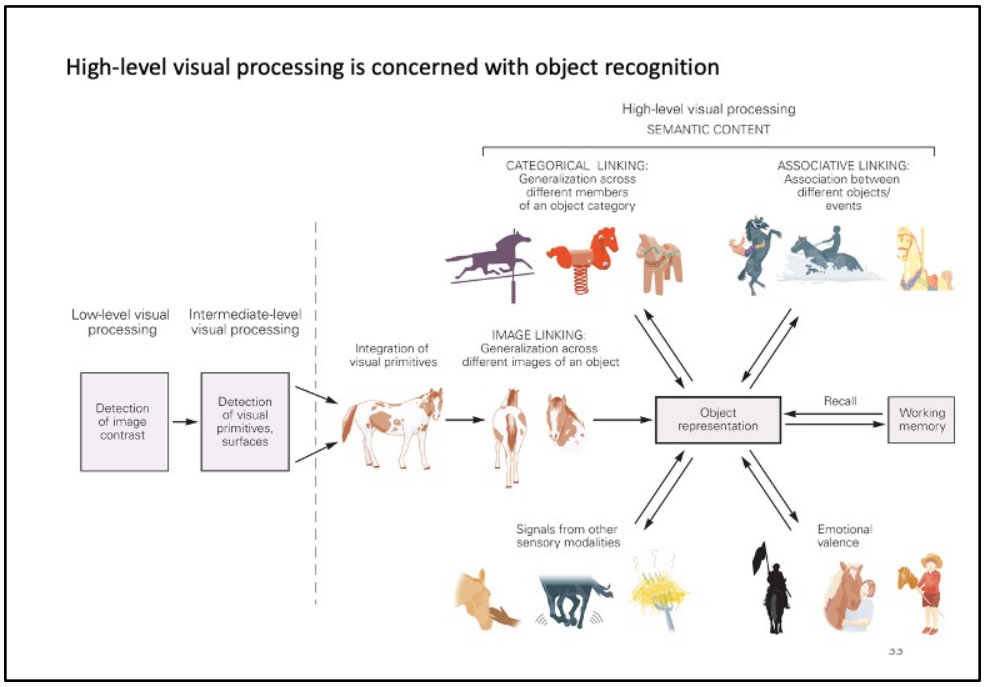
32

It is quite simple for computer vision techniques to identify (rather than categorize) objects.

On the contrary, for the human vision the task of identification (compared to categorization) is more difficult.
Categorization
A category exists whenever two or more distinct objects or events are treated equivalently.

For example, when distinct objects or events are labeled with the same name, or when the same action is performed on different objects.
Although the stimuli are distinct, organisms do not treat them uniquely; but they respond on the basis of past experience and categorization.

In this sense, categorization can be considered one of the most basic functions of living beings (Mervis and Rosch, 1981)



We effortlessly and rapidly (100-200ms) detect and classify objects from among tens of thousands of possibilities despite the tremendous variation in appearance that each object produces on our eyes.

Our daily activities (e.g., finding food, social interaction, selecting tools, reading, etc.), and thus our survival, depend on our accurate and rapid extraction of object identity from the patterns of photons on our retinae.

The fact that half of the nonhuman primate neocortex is devoted to visual processing speaks to the computational complexity of object recognition.

Object recognition involves integration of visual features extracted at earlier stages in the visual pathways.

This integration requires generalization across different retinal images of an object, as well as generalization across different members of an object category.

The representation also incorporates information from other sensory modalities, attaches emotional valence, and associates the object with the memory of other objects or events.

Selectivity and object constancy (or invariance)

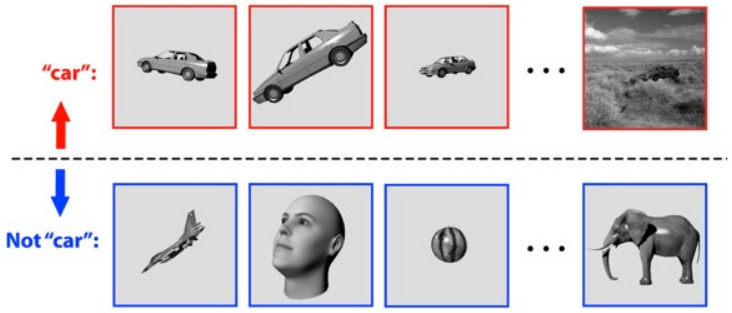
A computational difficulty of object recognition is that it requires both:

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.

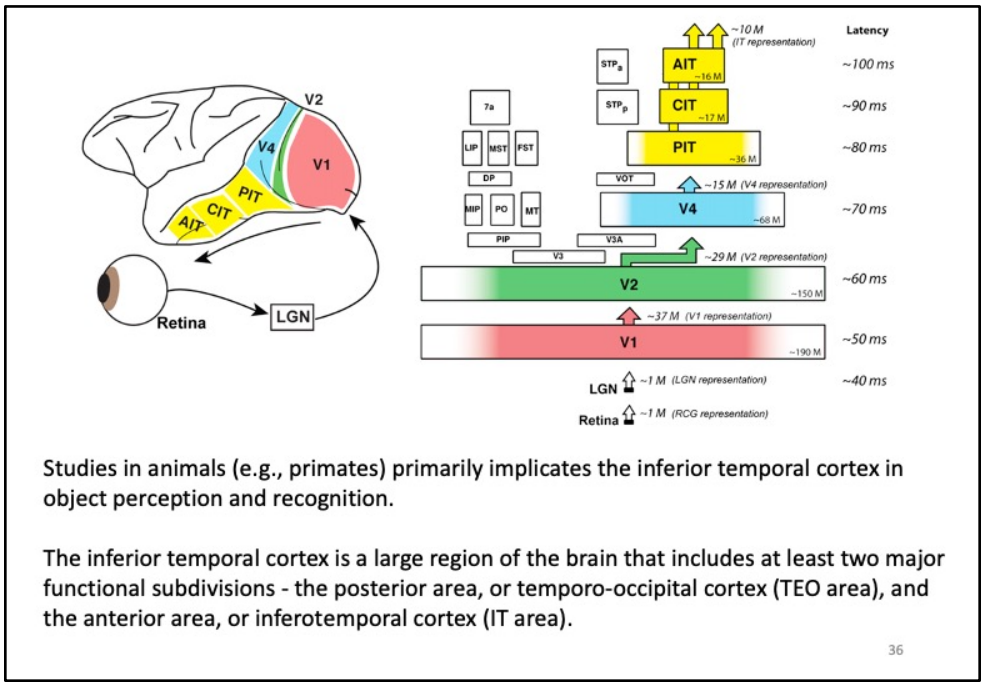
34



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).

35



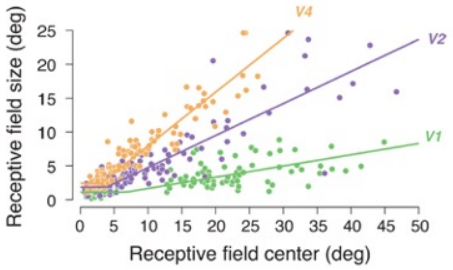
Studies in animals (e.g., primates) primarily implicates the inferior temporal cortex in object perception and recognition.

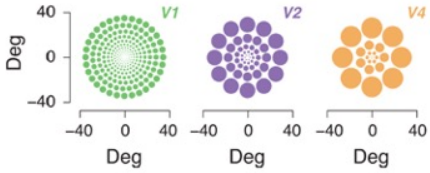
The inferior temporal cortex is a large region of the brain that includes at least two major functional subdivisions - the posterior area, or temporo-occipital cortex (TEO area), and the anterior area, or inferotemporal cortex (IT area).

36

Area V1,V2 and V4 are located in the occipital lobe;

Area TEO (TEmporal-Occipital) and IT (InferoTemporal) are located in the temporal lobe;





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

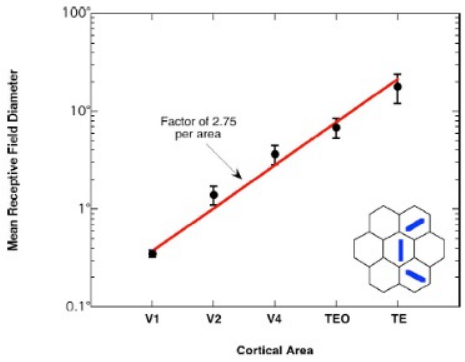
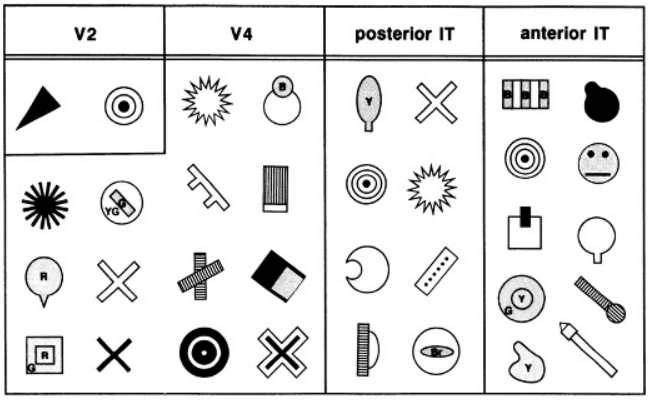


Figure 1. Mean data from four studies on receptive field diameter in successive cortical areas in the ventral pathway. Error bars represent the range of means across the four studies. The red line shows that the data can be fit by constant diameter increase of 2.75× from area to area. Inset in lower 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 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.

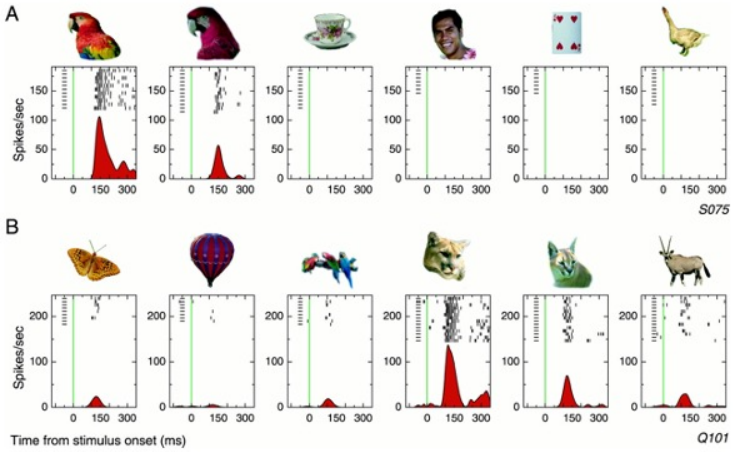
38

Increased complexity of effective stimuli along the ventral visual path

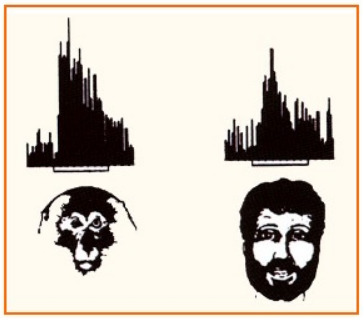


Kobatake and Tanaka, J. Neurophysiol., 1994

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.

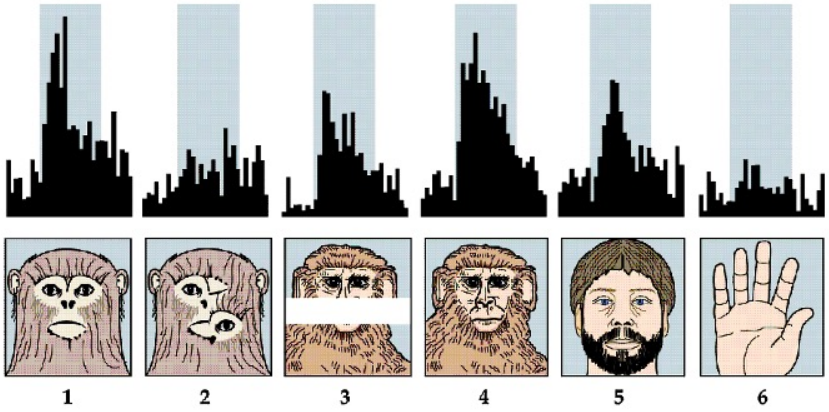


In the early 1980s, several researchers (Bruce et al., Perrett et al.) Identified in the monkey a group of IT neurons that responded selectively to faces.



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

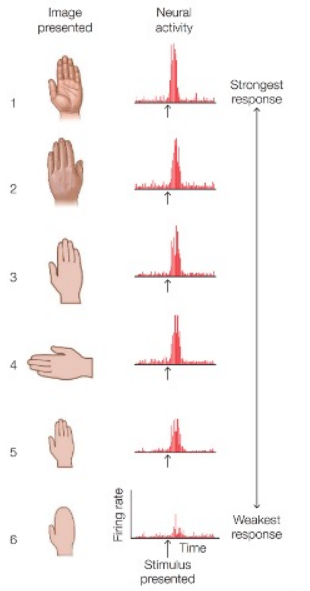
Recording of a single neuron from monkey IT cortex (Desimone et al., 1984)

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

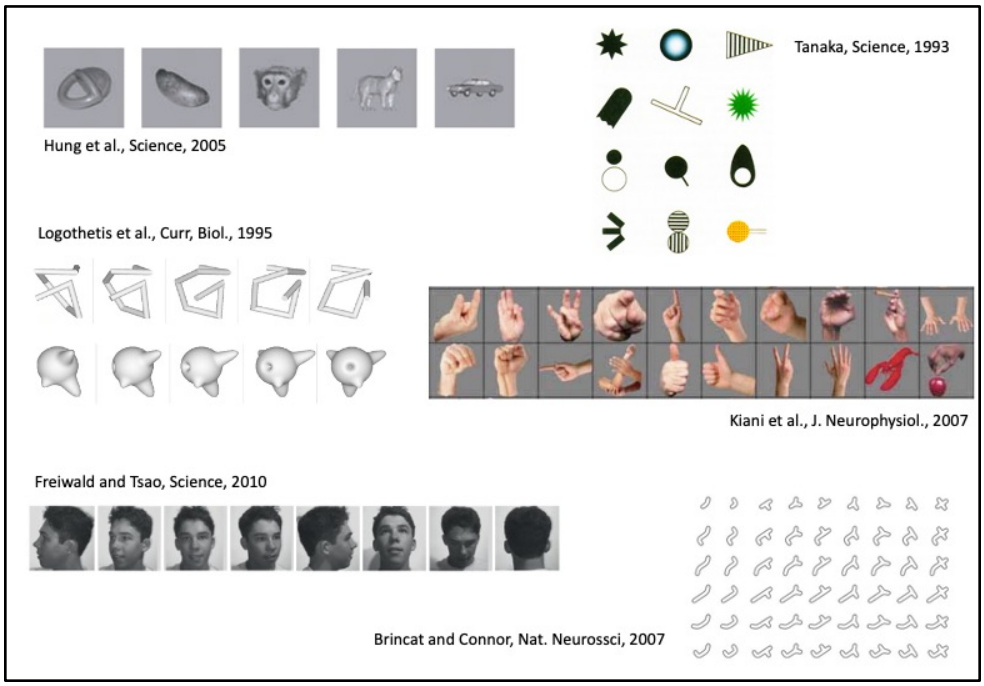
The first five images in the figure show the cell's response to various perspectives of a hand.

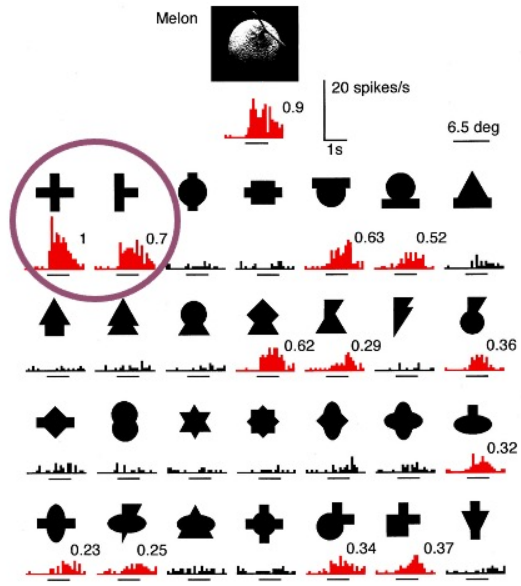
Activity is high regardless of hand orientation and only decreases slightly when the hand is noticeably smaller.

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



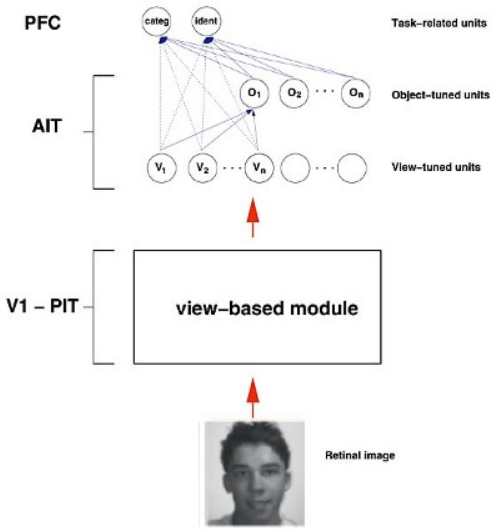
43


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.

Object recognition

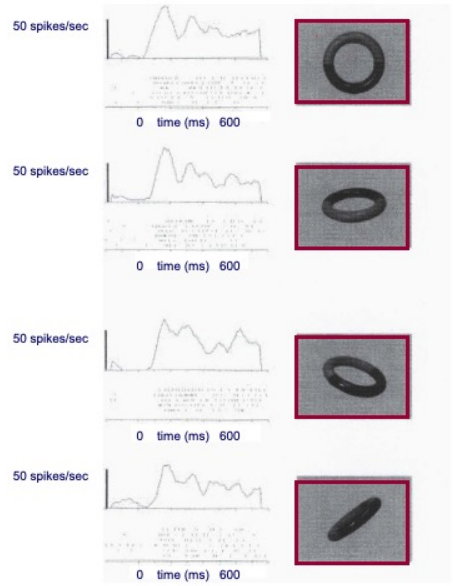


(view-invariant units)

(view-dependent units)

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

Riesenhuber & Poggio, Nat. Neurosci. 2000

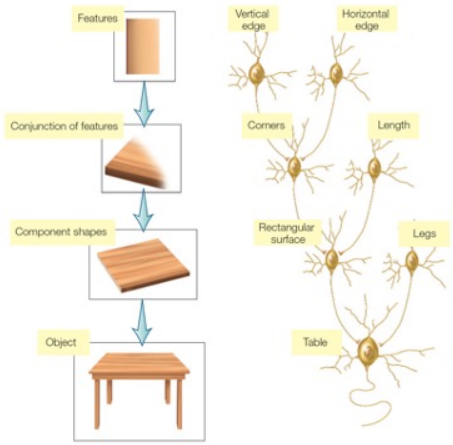


The majority of IT neurons respond to a stimulus only when it is presented from specific points of view (view-dependent responses).

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.

Hierarchical model of the object recognition

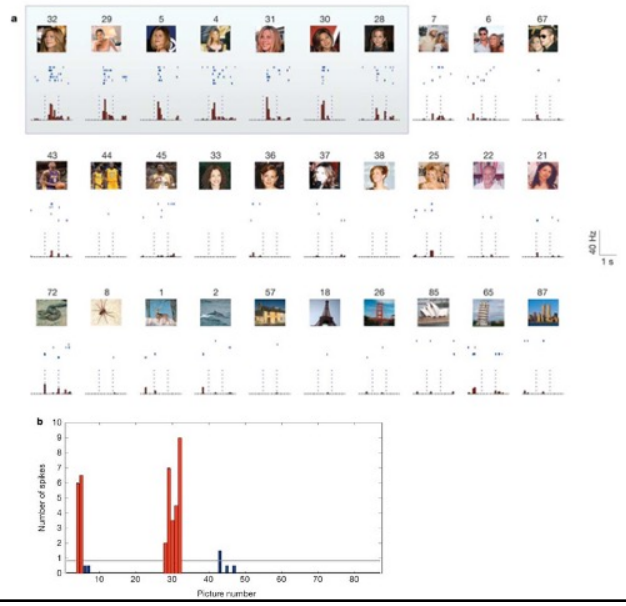


The finding that IT cells selectively respond to more complex stimuli than V1, V2 and V4 is consistent with a hierarchical model of object perception.

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' 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.

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.

51

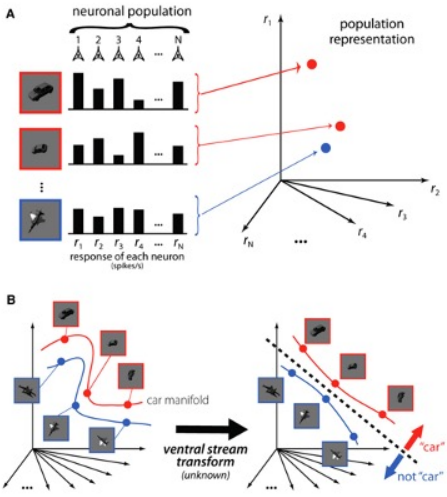
The results of the studies on single neurons of the temporal lobe are in agreement with the theories of the distributed code of object recognition.

Although it is surprising that some cells are selective for complex objects, the selectivity is almost always relative, not absolute.

- IT neurons respond only to visual stimuli.
- The receptive fields always include the fovea, that is the part of the retina most involved in the fine recognition of a visual stimulus.
- The receptive fields 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, thus joining the two halves of the space for the first time. This property depends on the interhemispheric connections through the splenium of the corpus callosum and the anterior commissure.
- IT neurons encode complex characteristics of the stimulus (not simple features, such as color, form orientation, depth).

IT neuron selectivity often appears somewhat arbitrary.
A single IT neuron could, for example, respond vigorously to a crescent of a particular color and texture.
Cells with such selectivity likely provide inputs to higher-order neurons that respond to specific objects.

Ventral visual pathway gradually "untangles" information about object identity



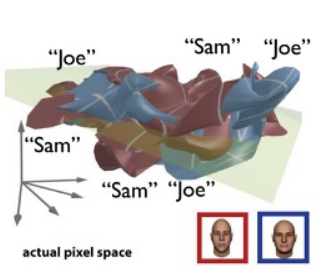
Response of a population of neurons to a particular view of one object can be represented by a response vector in a space whose dimensionality is defined by the number of neurons in the population.

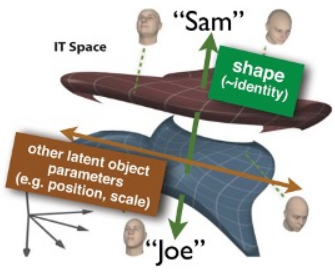
When an object undergoes an identity-preserving transformation, it produces a different pattern of population activity, which corresponds to a different response vector.

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.

53

DiCarlo et al., Neuron, 2012





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.

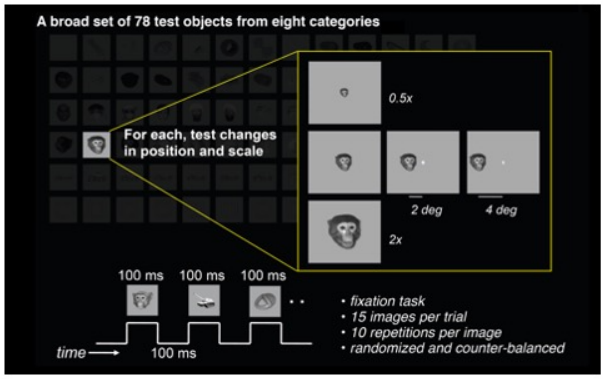
At higher stages of visual processing, neurons tend to maintain their selectivity for objects across changes in view; this translates to manifolds that are more flat and separated (more "untangled").

DiCarlo et al., Neuron, 2012

54

For neurons with small receptive fields that are activated by simple light patterns, such as retinal ganglion cells and V1, each object manifold will be highly curved.

Moreover, the manifolds corresponding to different objects will be ''tangled'' together,



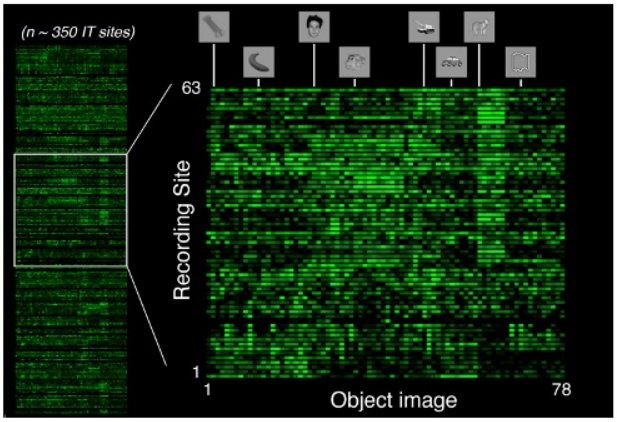
Categories: toys, food, human faces, monkey faces, hand/body, vehicles, white boxes, cats/dogs.

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

55



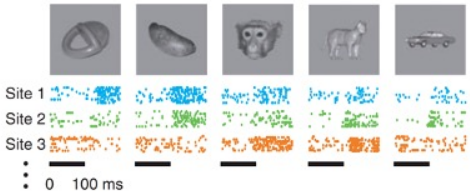
The readout technique consists of training a regularization classifier to learn the map from neuronal responses (from the independently recorded neurons) to each object label.

The input consists of the neuronal responses of IT neurons

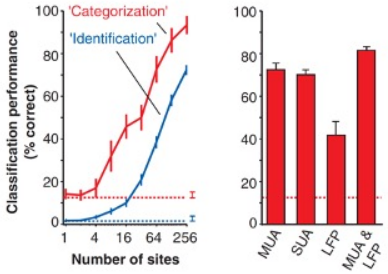
The activity of ~ 250 randomly selected multi- and single-unit activity in response to 78 images of different objects were recorded at 350 IT sites in two monkeys

Hung et al., Science, 2005

56



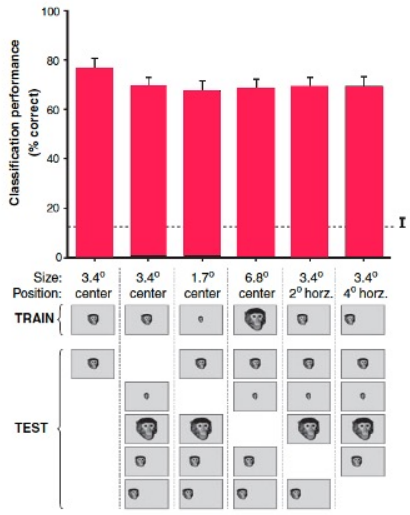
The spiking activity of 256 randomly selected multi-unit activity sites was sufficient to categorize the objects with 94% accuracy.



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

57



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.

58

Hung et al., Science, 2005

Objects could be reliably categorized and identified (with less than 10% reduction in performance) even when transformed (spatially shifted or scaled),
although the classifier only saw each object at one particular scale and position during training.

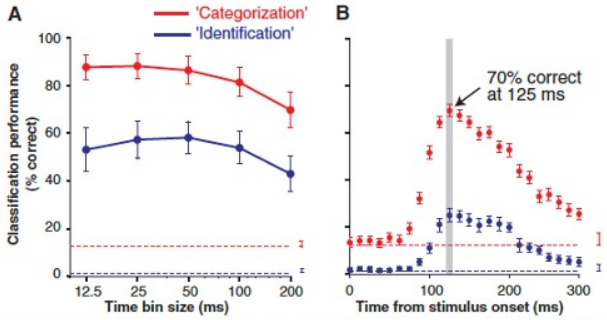


Fig. 3. Latency and time resolution of the neural code. (A) Classification performance (n = 128 sites) as a function of the bin size (12.5 to 200 ms, i.e., temporal resolution) to count spikes within the 100- to 300-ms window after stimulus onset for categorization (red) and identification (blue). The 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

59


60

Explaining the neural encoding in higher ventral areas remains an open question in systems neuroscience

The space of possible transformations that the brains neural networks could potentially compute (from retinal input to behavioral output) is vast.

To understand visual object recognition, one possibility is to test whether state-of-the-art neural networks from AI can be used as simulacre of true biological transformation.

Yamins et al., PNAS , 2014

61

Combining computational and electrophysiology techniques (single-unit recordings), Yamins et al. (2014) explored a wide range of biologically plausible hierarchical neural network models and then assessed them against measured IT and V4 neural response data, as well as human performance data.

The idea of this approach is to
first optimize network parameters for performance on an challenging object recognition task,
once network parameters have been fixed, compare networks to neural data.

Showed a strong correlation between a model's performance on difficult object recognition task and its ability to predict individual neural unit responses in V4 and V4, the top two layers of the ventral visual hierarchy.

62

Models of higher ventral areas should

- provide information useful to support behavioral task (equalling the decoding capacity of IT object recognition tasks neurons)
- mappable (i.e., layers correspond to distinct regions within the visual system, such as V1, V2, V4, IT).-
- neurally predictive (at the single-unit level, and at the neural population level)

Combining computational and electrophysiology techniques (single-unit recordings), Yamins et al. (2014) explored a wide range of biologically plausible hierarchical neural network models and then assessed them against measured IT and V4 neural response data, as well as human performance data.

Yamins et al., PNAS , 2014

63

Combining computational and electrophysiology techniques (single-unit recordings), Yamins et al. (2014) explored a wide range of biologically plausible hierarchical neural network models and then assessed them against measured IT and V4 neural response data, as well as human performance data.

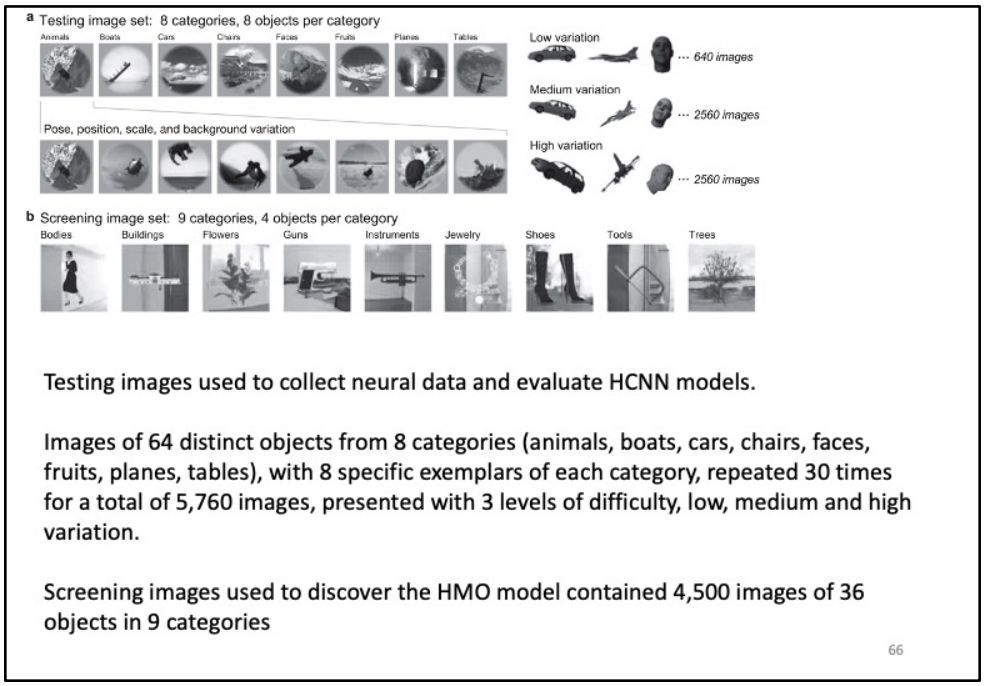
The idea of this approach is to
first optimize network parameters for performance on an challenging object recognition task, once network parameters have been fixed, compare networks to neural data.

64

Data collection
Collected neural data from V4 and IT (128 units in V4, and 168 units in IT) in two macaque monkeys, assessed human behavior, and tested HCNN models on a common image set.

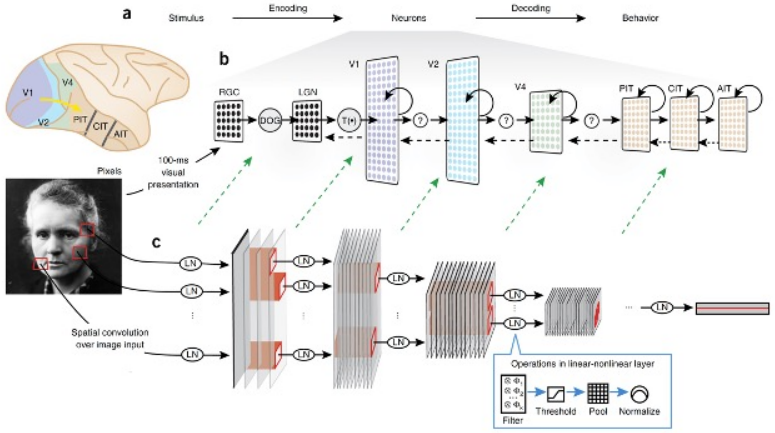
Fixating animals were presented in the center of screen (8° visual angle) with images in pseudorandom order, each for 100 ms, followed by a 100-ms gray blank period with no image.
For each image and electrode, firing rates were by averaging spike counts in the period 70–170 ms after stimulus presentation.
Final neuron output responses were obtained for each image and site by averaging over image repetitions.

65



Although the overall natural statistics of the screening images were roughly similar to those of the testing set, the specific content (semantic category) was quite different. Moreover, different camera, lighting and noise conditions, and a different rendering software package, were used.

Hierarchical convolutional neural networks (HCNNs)
HCNNs are good candidates for models of the ventral visual pathway and have achieved near-human-level performance on challenging object categorization tasks. HCNNs are multilayer neural networks, arranged in series, each of whose layers performing linear-nonlinear (LN) transformation on the input data, analogous to the transformation produced in the ventral stream.



These operations include

(i) filtering, a linear operation that takes the dot product of local patches in the input stimulus with a set of templates;

(ii) activation, a pointwise nonlinearity—typically either a rectified linear threshold or a sigmoid;

(iii) pooling, a nonlinear aggregation operation—typically the mean or maximum of local values;

(iv) divisive normalization, correcting output values to a standard range;

68

Models were drawn from a large parameter space of HCNN
Any given HCNN is characterized by the following:

discrete architectural parameters, including the number of layers, for each layer, discrete parameters specifying the number of filter templates; the local radius of each filtering, pooling and normalization operation; the pooling type; threshold values, and other choices required by the specific HCNN implementation;

continuous filter parameters, specifying the filter weights of convolutional and fully connected layers.

Let Nk denote the space of stacked networks (N) of depth k, the study used HCNN of depth 3 or less.

Models were selected for evaluation (measuring categorization performance and IT neural predictivity for each model) by one of three procedures:
random sampling of the models from the parameter space N3 (n = 2,016, green dots);

searched models that maximized performance on the high-variation categorization task (n = 2,043, blue dots);

searched models that maximized for IT neural predictivity (n = 1,876, orange dots).

69

Constructing a high-performing (mixture) HCNN model

A hierarchical modular optimization (HMO) procedure was used to create high-performing architecture that contains combinations (e.g., mixtures) of deeper CNN networks.

Algorithmically, HMO is analogous to an adaptive boosting procedure (Freund and Schapire, 1995).

Since the networks are convolutional, they can be combined by aligning the module output layers along the spatial convolutional dimension.

The diameter increase 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.

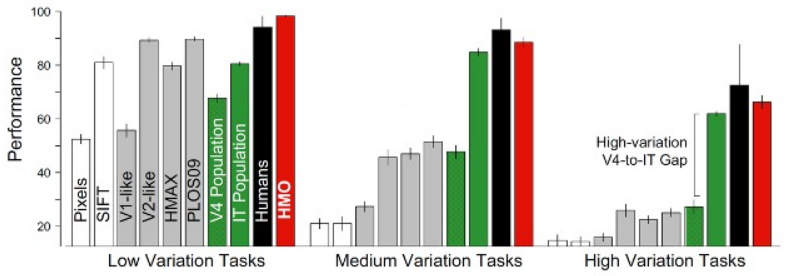
70

Categorization performance

Object recognition performance was assessed by training linear SVM classifiers with regularization on model and neural output.

For models, the output features on each stimulus are defined as the set of scalar values for each top-level model unit when evaluated on that stimulus.

For neuronal sites, the output features are defined as the vector of scalar firing rates for each unit, as is typical in neural decoding studies.

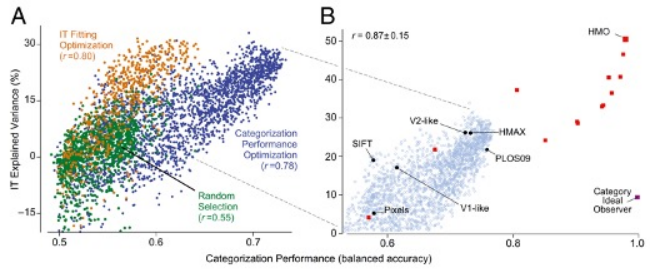


71

Models were selected for evaluation (measuring categorization performance and IT neural predictivity for each model) by one of three procedures:

- random sampling of the models from the parameter space N3 (n = 2,016, green dots);
- searched models that maximized performance on the high-variation categorization task (n = 2,043, blue dots);
- searched models that maximized for IT neural predictivity (n = 1,876, orange dots).

72



Performance/IT-predictivity correlation

The x axis shows performance (balanced accuracy, chance is 0.5) of the model output features on a high-variation categorization task; the y axis shows the median single
site IT explained variance percentage (n=168 sites) of that model. Each dot corresponds to a distinct model selected from a large family of HCNN.

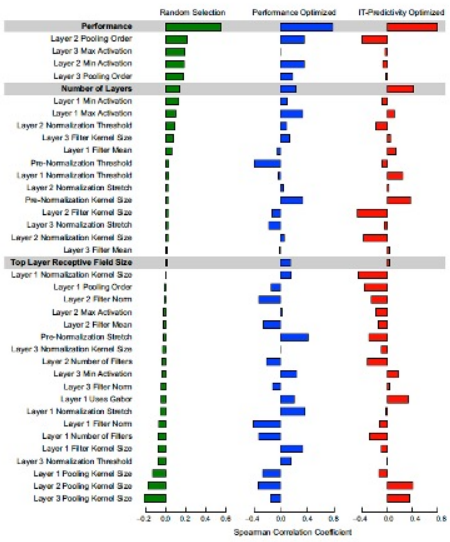
73

Performance was significantly correlated with neural predictivity in all cases.

Models that performed better on the categorization task were also more likely to produce outputs more closely aligned to IT neural responses.

Thus, although the Hierarchical Linear-Nonlinear (HLN) hypothesis (i.e., higher level neurons (e.g., IT) output a linear weighting of inputs from intermediate-level (e.g., V4) neurons, followed by simple additional nonlinearities) is consistent with a broad spectrum of particular neural network architectures, specific parameter choices have a

large effect on a given model’s recognition performance and neural predictivity.



No individual model parameters correlated nearly as strongly with IT predictivity as performance, indicating that the performance/IT predictivity correlation cannot be explained by simpler mechanistic considerations (e.g., receptive field size of the top layer).

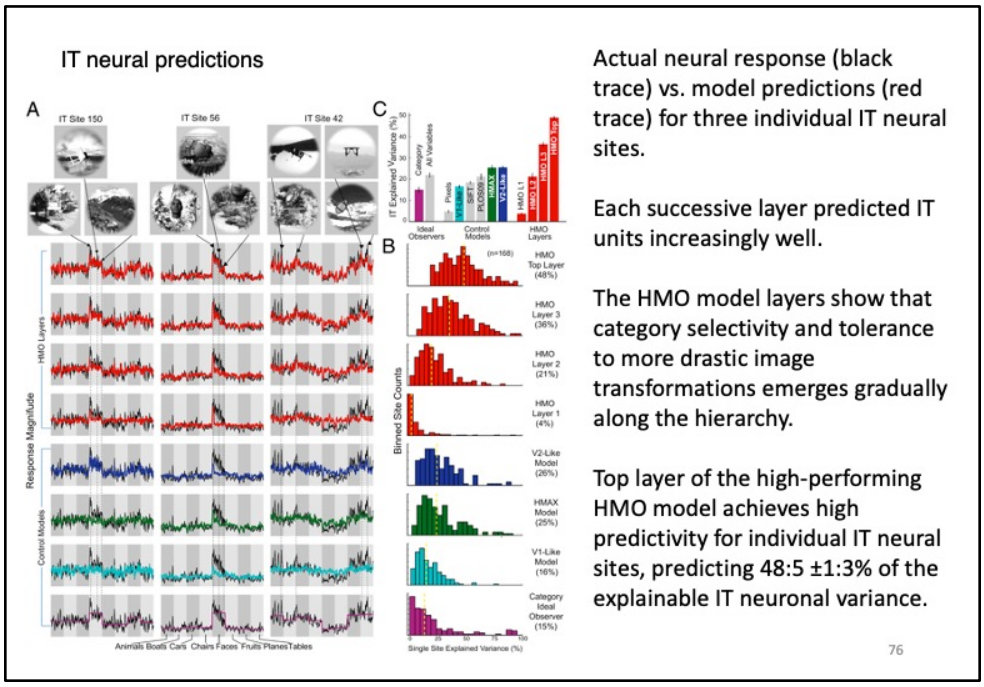
Correlation of model parameters with IT predictivity for the three types of model selection

Predicting neural responses in individual IT and V4 neural sites.

To assess model's ability to predict a given neuron output, a standard linear regression methodology was used, in which each neuron site is modeled as combination of model output.

Briefly, a partial least squares (PLS) regression procedure was used to determine weightings of top-level model outputs which best fit a given neurons' output on a randomly chosen subset of the testing images. The percentage of explained variance was then computed on a per-site basis using the r2 prediction value for that site

75

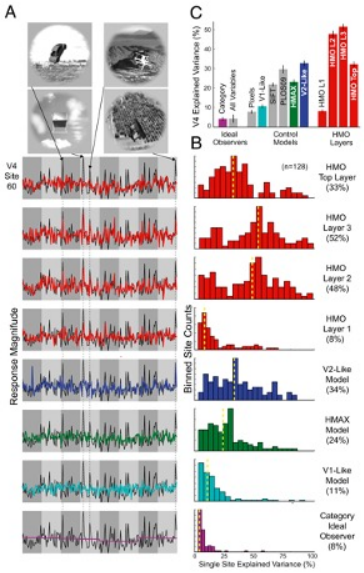


The x axis in each plot shows 1,600 test images sorted first by category identity (8 stimulus categories) and then by variation amount, with more drastic image transformations toward the right within each category block. The y axis represents the prediction/response magnitude of the neural site for each test image (those not used to train the model).

In B, Distributions of model explained variance percentage (r^2), over the population of all measured IT sites (n = 168).

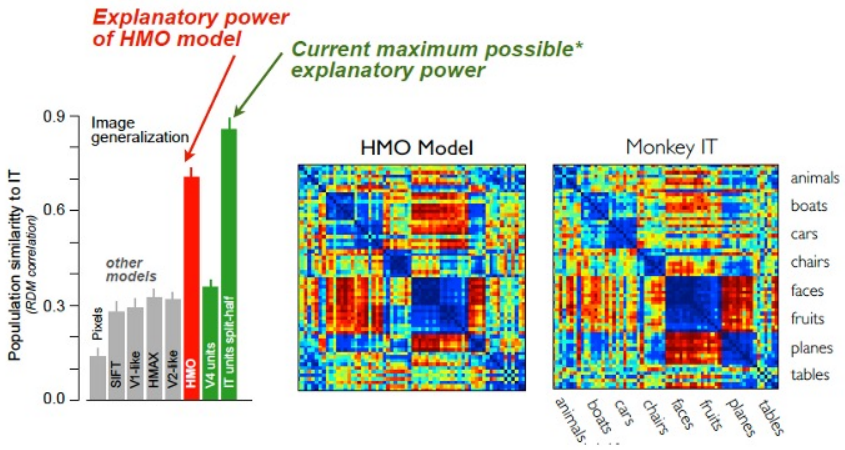
In C, comparison of IT neural explained variance percentage for various models. Bar height shows median
explained variance, taken over all predicted IT units.

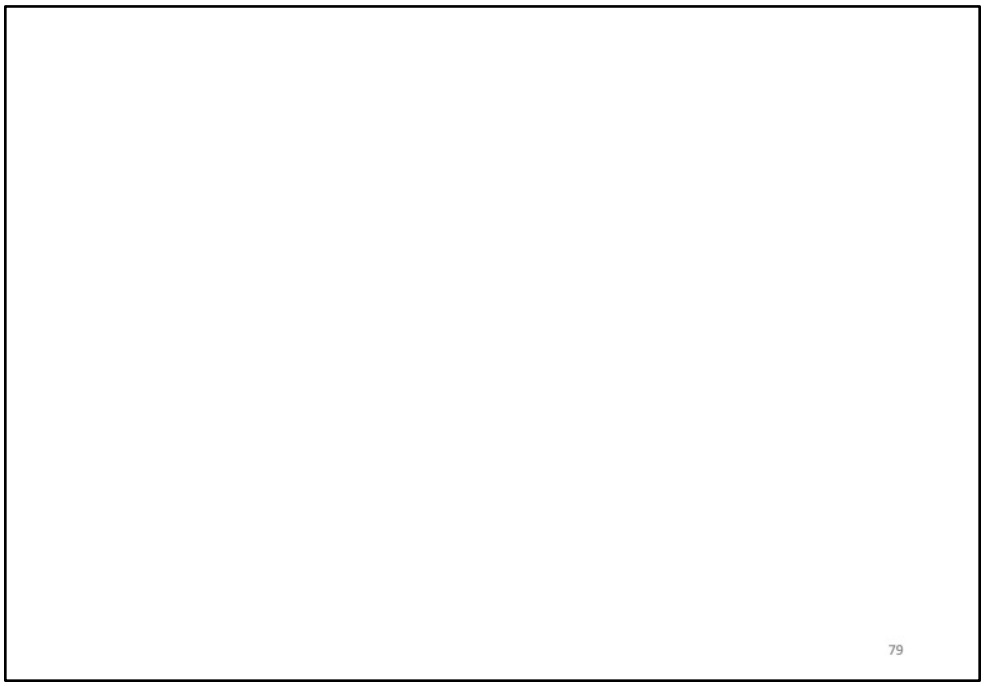
V4 neural predictions

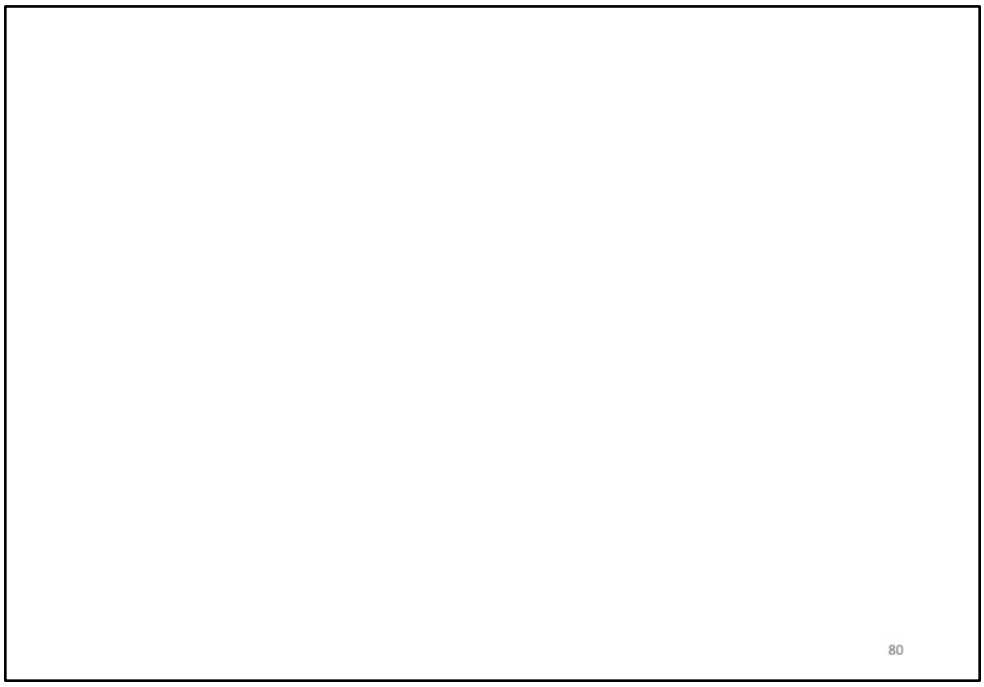


HMO model's penultimate layer is highly predictive of V4 neural responses (51:7± 2:3% explained V4 variance), providing a significantly better match to V4 than either the model's top or bottom layers.

Representation Dissimilarity Matrices (Kriegeskorte, 2008)




79


80