Machine Learning for Human Vision and Language

Lecture 4:
Higher (and recurrent) visual processing

Ben Harvey

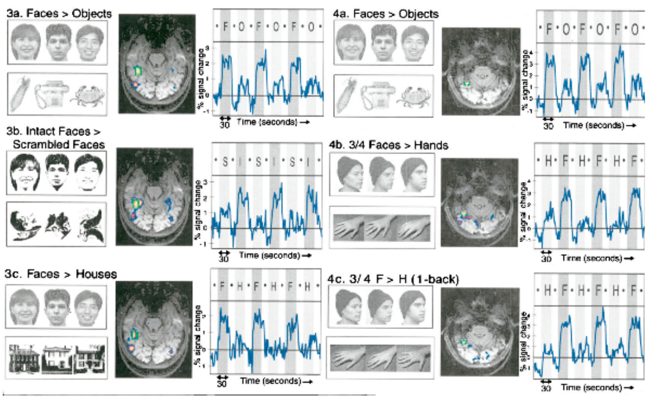
Today we are going to build on what we looked at last time, extending up to higher levels of visual processing and how these interact with lower levels.

Why study higher vision?

- Image processing is a common application of artificial deep networks
  - Particularly object recognition
- Important properties of early and higher visual processing are absent in current artificial DCNNs
- Including these properties may:
  - Give artificial DCNNs new abilities
  - Make artificial DCNNs more brain-like
  - Unite Hebbian learning and backpropagation
  - Reduce the gap between unsupervised (biological) and supervised (artificial) learning

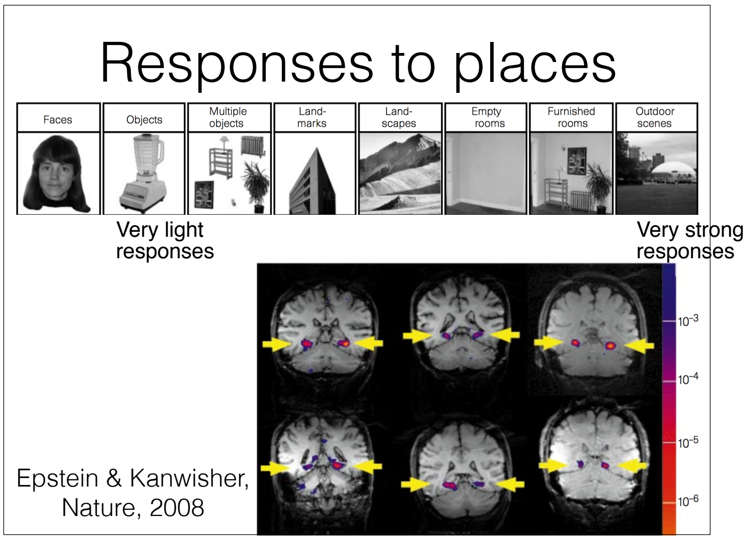
These properties are what we will look at today.

Face responses in the FFA



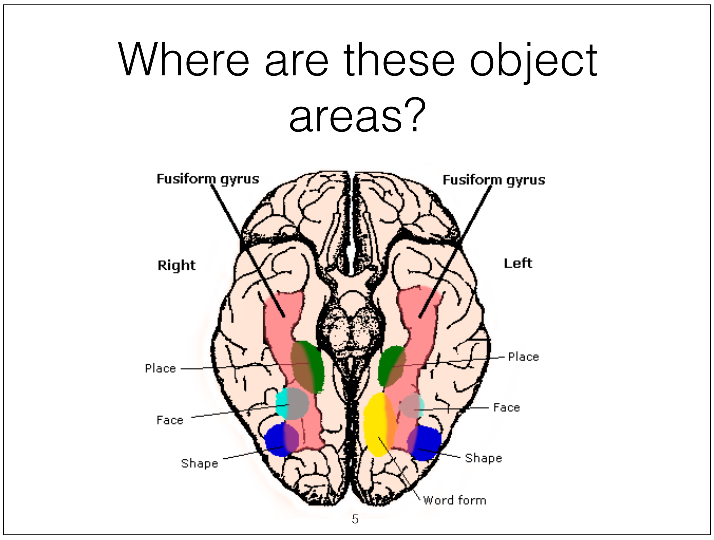
From Kanwisher et al. 1997

Last time we saw responses of human face-selective areas with fMRI.
They respond more to faces than various other object types



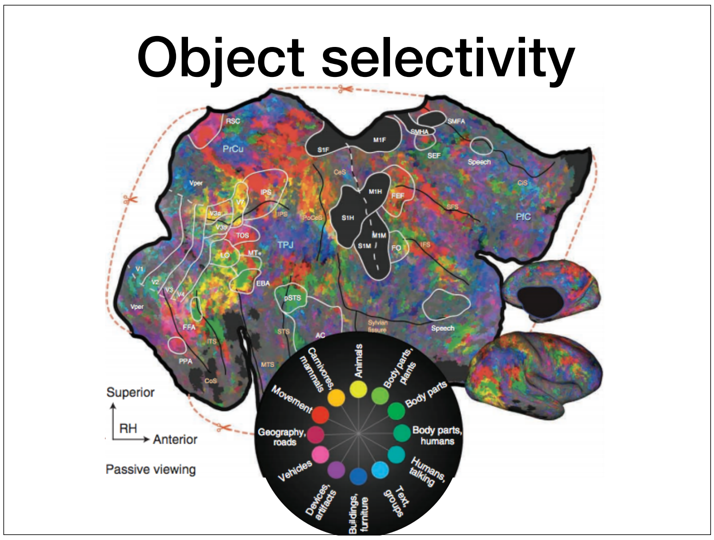
There is no specific response for male/female —> DNA is a limited bandwidth medium to share data/information

But faces are not the only type of objects that produce responses in specific brain areas
-If we show subjects in an MRI scanner pictures of objects, faces, and places, we also find an area that responds specifically to places.
-This has almost no response to faces, some response to objects, and a stronger response to places.
-Again, the images of places vary considerably, but all are clearly places
-The images of landmarks and objects look relatively similar, but the responses they produce are very different.



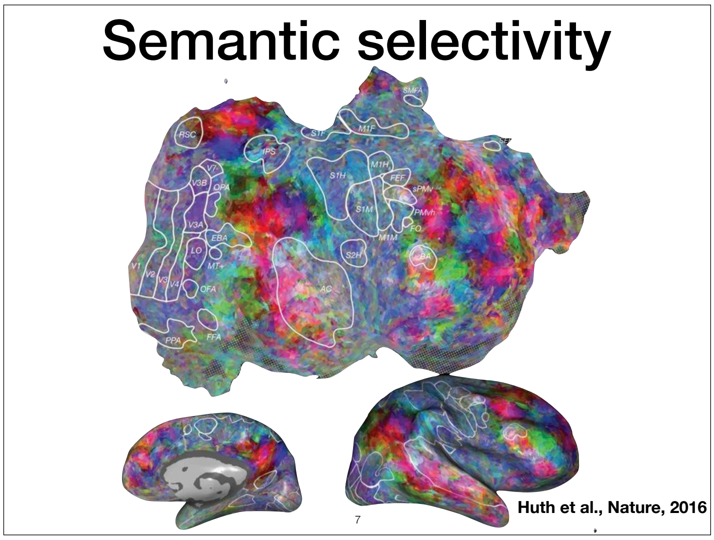
Circular as the researcher's brain is selecting the object for which it wants to find responses

So far, we have seen specific responses to faces and objects, which are quite similar.
-Both activate specific areas on the bottom of the temporal lobe (inferior temporal lobe, or IT)
(V1 is all the way at the bottom of the picture)
-Faces and places are particularly important to humans as we move around our environment a lot and are very social compared to other animals.
-We also find a 'visual word form area' that responds to written words, which are again very important for modern humans.
-Faces, places and words are major categories that experiments on object processing have focussed on.
-To some extent, this approach is limited because the experimenters have used their brain to identify a type of object that they think is a specific class, then tested whether the brain has a response to that type of object.
-As they have used their brain to choose an object type, then found a response to that in their brain, this could be seen as pretty circular. The experimenter has chosen what aspect of the scene they think is important for the brain.
-But the chosen images of faces and places differ in many ways, and any of these could drive the responses that they see.

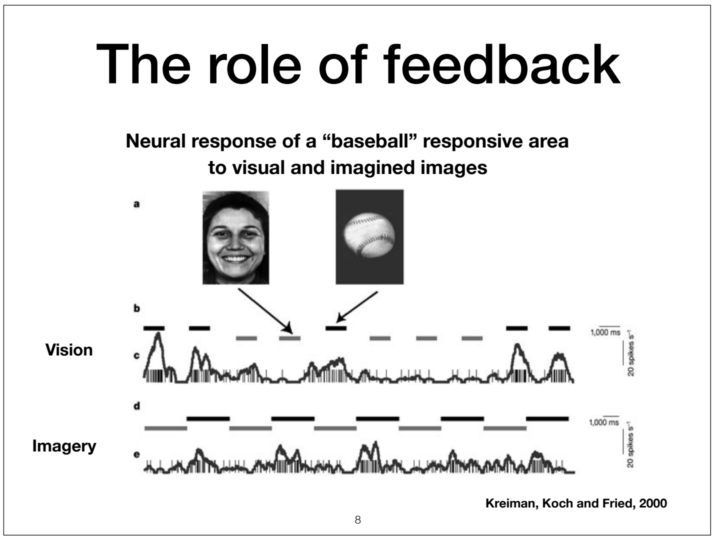


MT = visual motion —> vehicles
PPA = plants parts and …

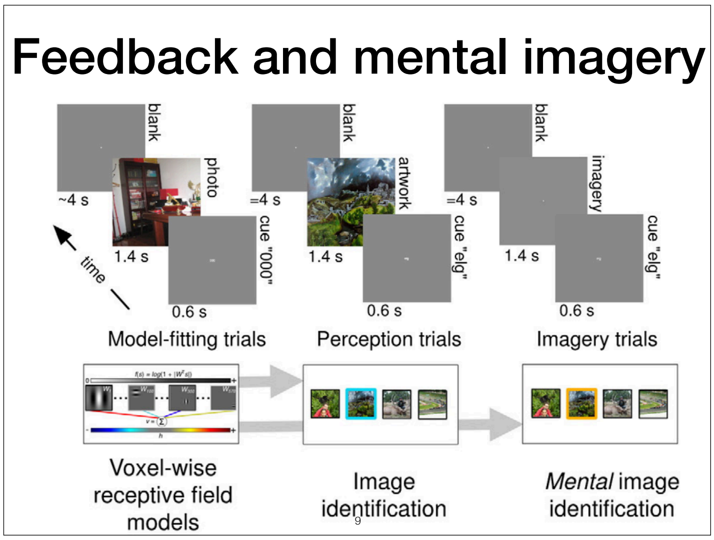
–Here, experimenters labelled all objects in natural movies, then shown these movies to a brain.
–By correlating the responses with the content of labelled movies, they have determined the object preferences of fMRI recording sites in a less biased way
UNFOLDED BRAIN, FFA, PPA
–We see a lot of brain areas responding to various object types, labelled with different colors
–So there are object-selective areas for a large range of object types.



More recently, the same group has revealed a network of areas responding to different types of semantic content in tagged narrative speech and reading.
–This is particularly important because we can't investigate linguistics well in animal models of the brain.
–But also, this is interesting because, although we have little idea how the brain processes language, we see that it reaches very similar results to processing visual objects.
–This implies deep learning mechanisms are likely to be involved in language comprehension.
–This paper is on your reading list, and you will see more about linguistic processing in the second part of the course.



So, if we show some images to an object-selective area that respond to baseballs more than faces, we see it respond more when baseballs are shown.
–However, if we ask people to alternatively imagine baseballs and faces, we also see greater activity in this area when imagining baseballs, even though no images were presented.
–So object-selective responses don't just depend on what we are seeing, but also what we are visualising or paying attention to.



Indeed, when we imagine images, feedback on image content seems to go right back to V1.

-If we show a large set of photographs to a subject in an MRI scanner, we can determine where the contrast lay in those images and compare this to when each of V1's recording sites responds.

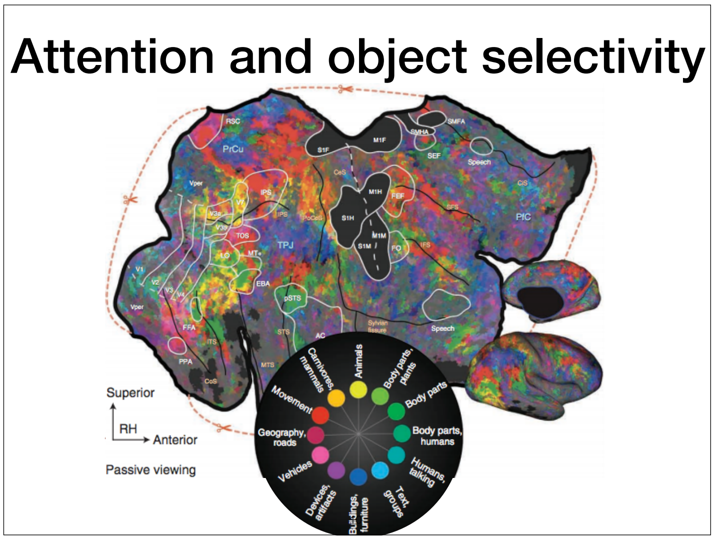
-This gives a model of which locations, orientations and spatial frequencies each recording site responds to.

-If we then show a new image to the resulting model, we can predict how each recording site should respond to that image.

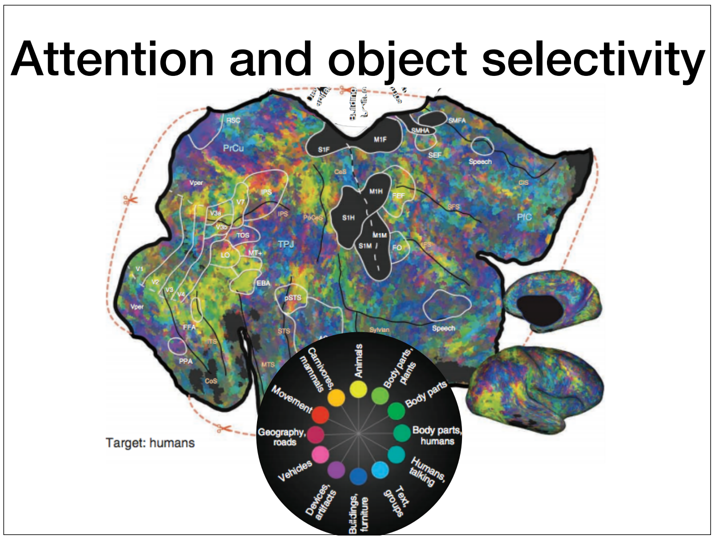
-We can then compare this prediction to the actual recordings from an unknown image, revealing which image was shown from the neural activity.

-But here, we can also ask subjects to imagine an image while showing nothing. This allows us to identify which image the subject was imagining.

-Importantly, this shows that the image representation from higher levels and from memory is imposed onto a precise retinotopically organised pattern of activity neural activity in V1.



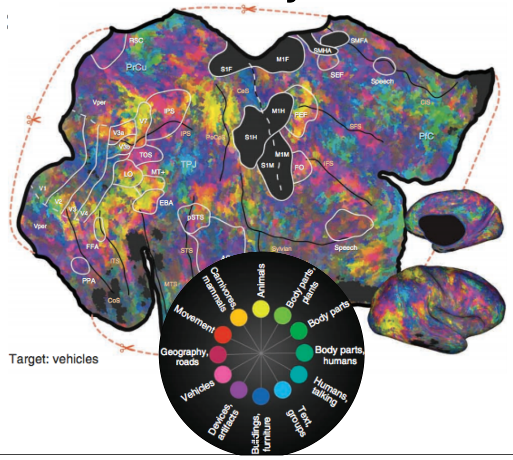
When mapping these visual object-selective responses, the subjects were just passively watching the movies



When subjects have a task to press a button when the see a human, they have to pay attention to humans.

Recording sites throughout the brain change their object selectivity, and start responding more strongly to humans and similar objects, in blue-green colours

Attention and object selectivity

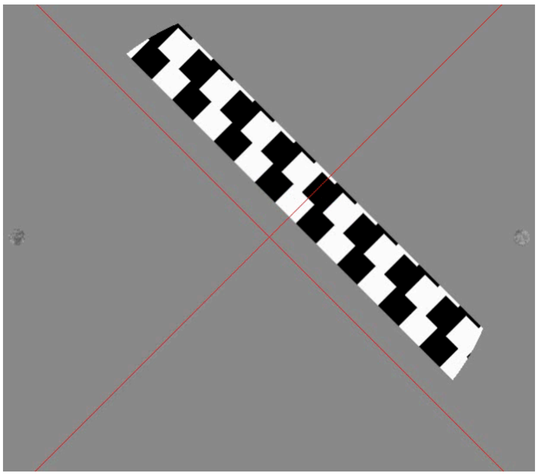


And when they are asked to respond to vehicles instead, recording sites start responding more to man-made objects like vehicles

Therefore, object selectivity throughout the brain depends strongly on what task we are doing.
A face-selective area can become a car-selective area if given a car identification task.

Activation of brain activity/ neural processing depends on what we are seeing, paying attention to or imagining

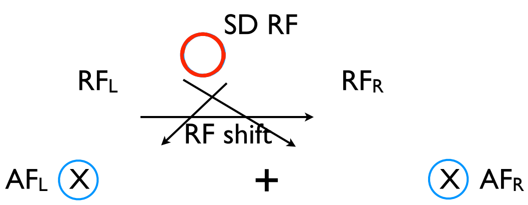
---



Similarly, attending to a specific spatial location attracts spatial receptive fields towards that location.
Here, we can use the responses to this moving bar to map out the spatial receptive fields of fMRI recording sites throughout the brain.
While doing this, we can get subjects to do a very difficult task at one of two spatial locations, the little patches on the left and right

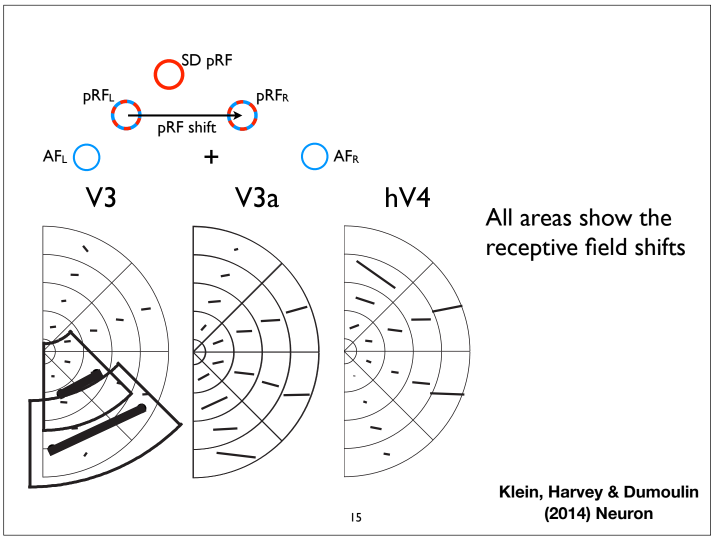
---

Attention and spatial selectivity

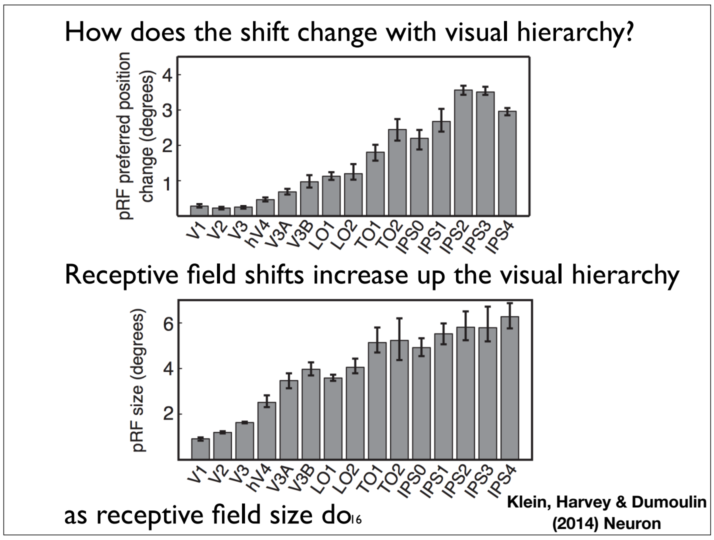


Klein, Harvey & Dumoulin (2014) Neuron

If we attend to left target, receptive fields are drawn towards that target
If we attend to the right, receptive fields are drawn to the right
We can then compare these positions to see how attention affects receptive field positions



Here we see the receptive field positions when attending left and right, with a line joining them



In later visual processing, neurons are less sensitive to stimulus position

And they can also have their stimulus positions moved more by attention

So, throughout the visual hierarchy, visual position preferences move towards the attended location

This reallocation of the visual position representation has the effect of zooming in on attended areas.

Object-selectivity, imagination and attention

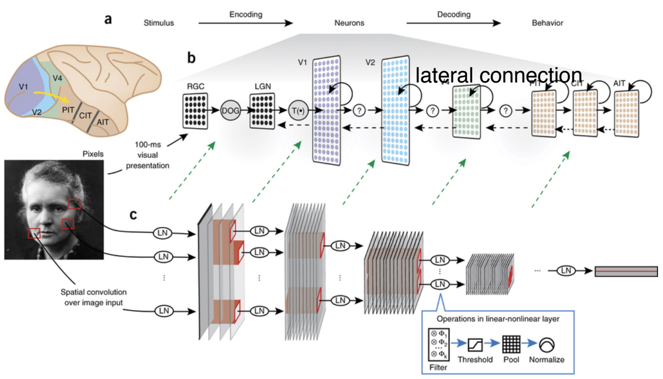
- Responses to many classes of object
  - Faces, places, words and tools are commonly examined
- Similar responses recently found for semantic content in language
  - Suggests similar processes are involved
- Responses can be driven by imagined content too
  - Gives similar responses to seen content
  - All the way back to the early image representation in V1
- Responses are drawn towards attended content
  - Object responses drawn towards attended object
  - Spatial responses drawn to attended locations

18

Hierarchic response to attention and object representation

AT END:
None of this can happen in the networks we have seen so far.
These take an input, transform it and give an output.
There is no way that imagined content or attended content can affect this process.
Imagining an object will start by activating a high level object representation, which must then be fed back to the earlier representation of the visual image.
Similarly, attention is thought to be driven by higher levels, where goals and decisions are represented, then fed back to earlier stages.

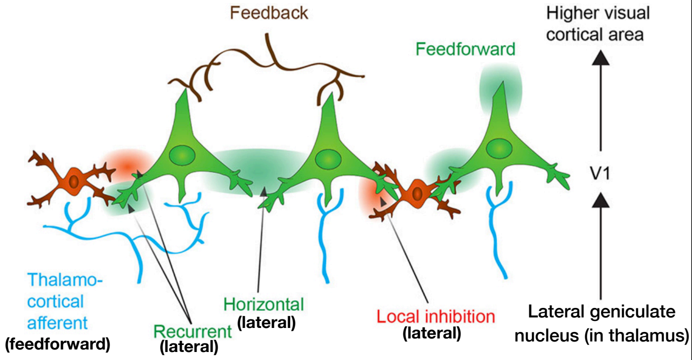
Feed-forward convolutions only?



19

Here we see that a biological network, at the top, has feedforward (or bottom-up) connections, together with feedback (or top-down) connections.
-For these connections to put a precise activity pattern into earlier areas, each of these feedback arrows must carry a similarly complex pattern of activity to the feedforward arrows that represent our convolutional filters.
-Here we also see inputs to each unit coming from other units in the same layer, called lateral connections.
-In the artificial DCNN at the bottom, both of these are missing.
-A final type of connection pathway is often found in artificial deep networks. This is where a neuron's activity depends on its own recent activity.
-In many artificial DCNNs, the input is a single static image with no temporal information, so recent activity is always zero and we can ignore it. But in networks with temporal inputs, recurrent connections are common.
-These are particularly important in linguistics, as language is a temporal sequence.

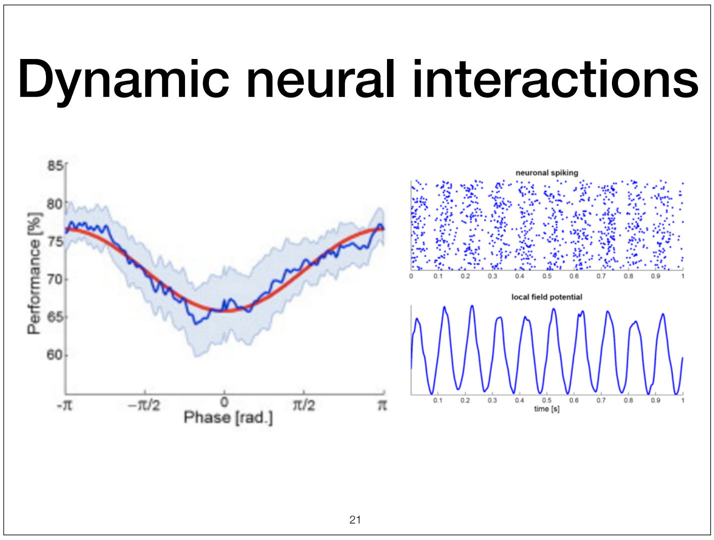
Feedback, lateral and recurrent activity in the brain



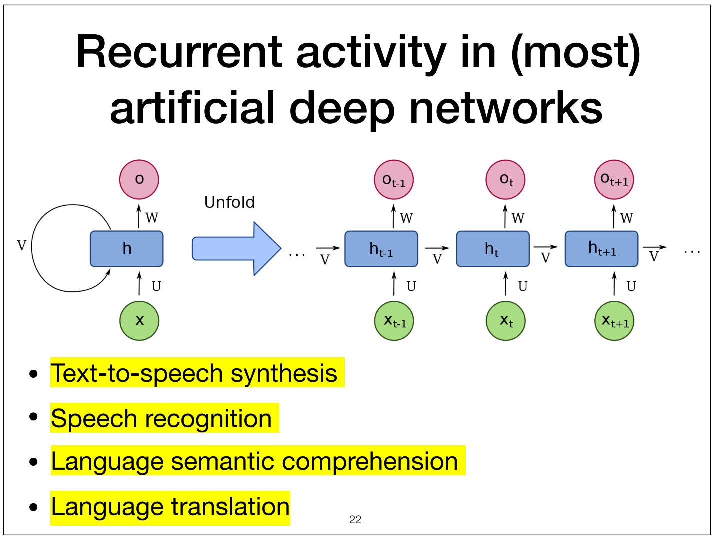
20

So far, we have only looked at feedforward connections, represented in this case as inputs to V1 from the thalamus, shown in blue at the bottom.
-After some transformation, these will then pass the image representation up to higher visual areas.
-Likewise, feedback connections from these higher areas will affect the activity of the neurons.
-And the neurons will interact with each other, through both excitatory and inhibitory lateral connections (labelled 'local inhibition' and 'horizontal').
-For example, a cell responding to a vertical orientation at one location may form an excitatory synapse with a cell responding to a vertical orientation at a nearby location, making that more likely to fire.
-Or it may form an inhibitory synapse with a cell responding to a horizontal orientation, making that less likely to fire.
-A neuron can activate other neurons in the same brain area, higher or lower areas, changing their activity.
-Changes in the activity of those neurons can in

turn activate or deactivate the first neurons (here labelled ‘recurrent’).
–Note that these neurons never form synapses with themselves, there is always another neuron forming a circuit.
–As these interactions take a moment to happen, the neuron’s previous activity will then affect its current activity.
–Indeed, all of these interactions have a temporal component, as each interaction step takes a little time.
–As a result, a feedforward network will reach it’s end state with one pass of the information through the network, as information only moves in one direction.
–But networks with interactions like we see here will change the activity in this layer depending on results of surrounding activity and higher level activity.
–The result will in turn change the surrounding activity and the higher level activity, changing their influences back and forth until an equilibrium is reached. Because of this dynamic interaction, such connections always make biological networks



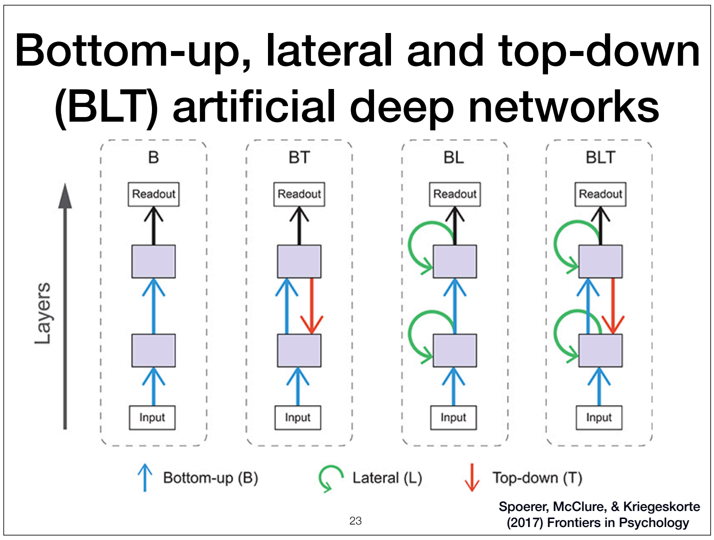
These dynamic interactions between excitatory and inhibitory neural population produce oscillations in neural population activity, which can be measured using EEG.
At oscillation peaks, excitatory population activity is highest, including the spike rate
This also affects perception. Threshold stimuli presented at an oscillation peak are more likely to be perceived than those presented at a trough



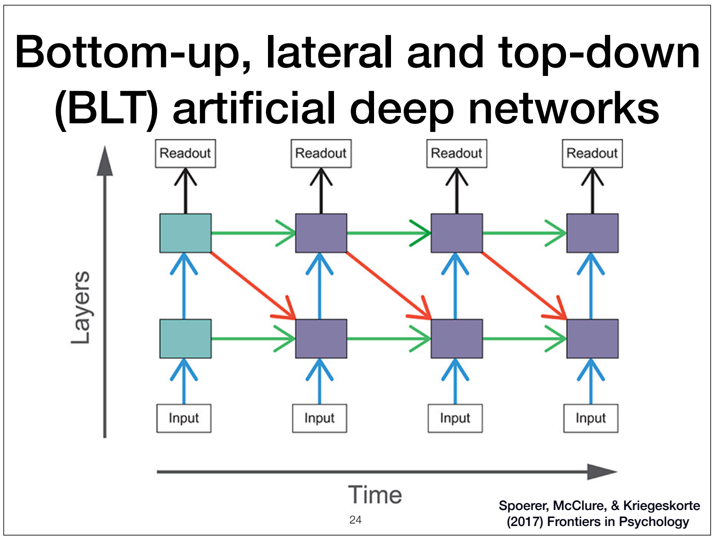
Recurrent activity is sometimes implemented in deep networks. Initially, this was a way of dealing with time.
–In a layer with recurrent activity, the activity of each unit at one time point feeds into the same unit at the next time point.
–This is useful for many tasks where temporal sequence is important
–While spatial relationships are very important for visual processing, temporal relationships are very important for linguistic processing.

–Effectively, this architecture means the unit is partially activated by the past version of itself.
–This is a simplification of recurrent activity in biological networks because biological neurons never synapse directly with themselves.
–The link between the unit’s activity at the two time points again has a weight that is a target for machine learning

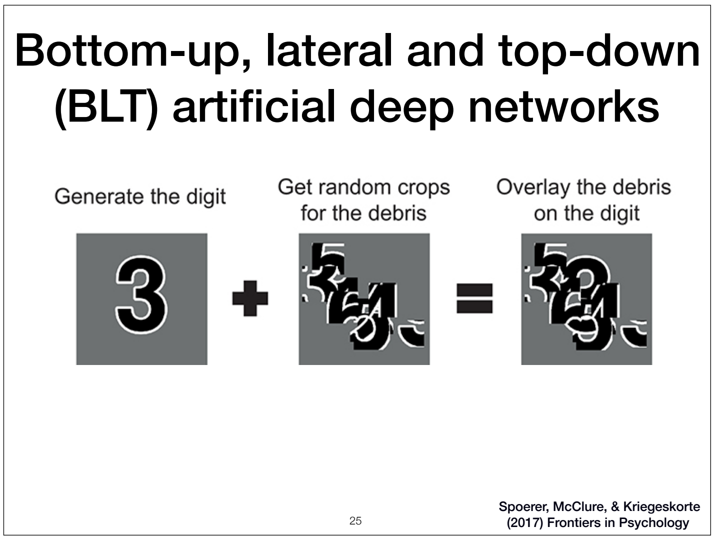
–This type of network unit is often referred to as a ‘long short-term memory’ unit
–Here, short term memory is conceptualised as maintenance of information over time that depends on the present network activity. Long-term memory, on the other hand, would be achieved through adjustment of network weights.
–Because this activity-dependent effect can last for a long time, the phrase ‘long short-term memory’ is used.



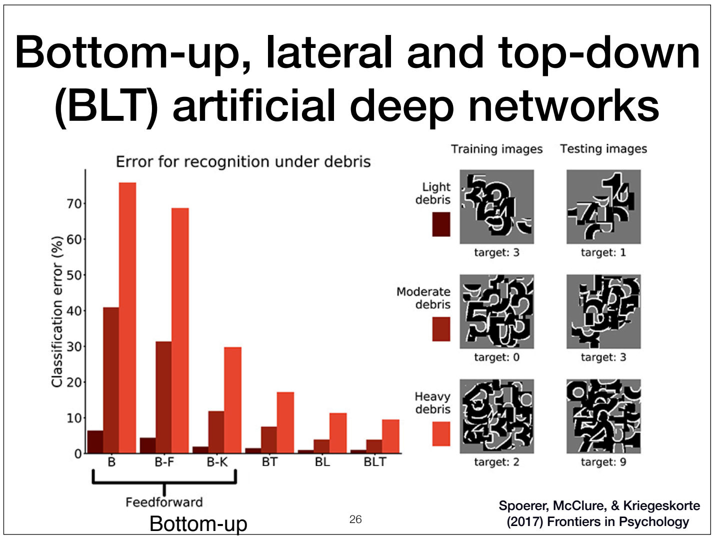
Here is a set of artificial network that incorporate the extra connections that exist in biological networks.
So far, the networks we have looked at are of type B, bottom-up connections only.
It is actually pretty straightforward to make the convolutional filter not only sample from the previous layer, but also from surrounding units in the same layer, and from the next layer up.
The convolutional filter needs to be larger, so more weights need the be learned.



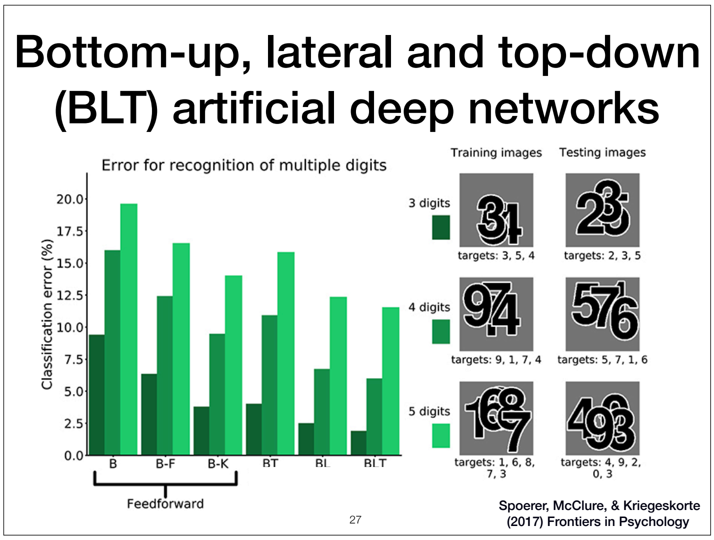
Perhaps more complex is that the network's activity will move back and forth as the results of each interaction affect the activity of the interacting elements.
Here the experimenters include a fixed number of cycles, here 4.
Of course, in the brain there are not such discrete time steps and there is no end to the interaction cycles.
Here, the input image remains the same through these cycles, while the input to the brain will always be changing



They test this on a digit recognition task.
As you have seen in the labs, this is a pretty easy task and does not require very deep or complex networks.
This gets much harder if we add some parts of other digits to our target digit, but a human can still make out the target digit well enough.



When there is relatively little junk (debris) in the way, a simple bottom-up network does pretty well (far left, dark colours, low error).
But adding a lot of debris (lighter colours) makes performance of a bottom-up network drop quickly.
In a BLT network, we see much better performance in all three levels of debris.
B-F and B-K networks are both bottom-up networks, but the filters are larger or there are more filters so that the same number of weights needs to be learned as in the more complex networks.
These also do well, but clearly the feedback and lateral influences are improving performance.



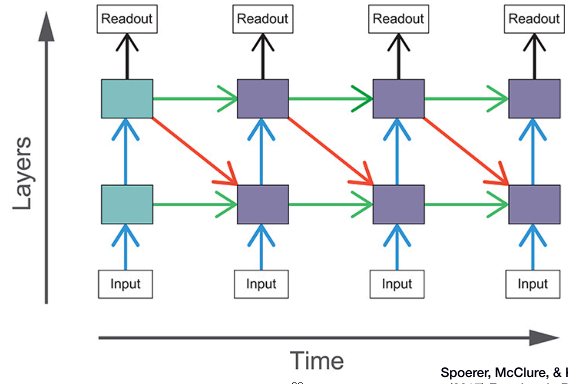
Similarly, when recognising multiple digits, we see an improvement over feed-forward only networks

Feedback, lateral and recurrent activity

- The brain relies heavily on lateral and feedback activity
  - Recurrent activity is implemented through both lateral and feedback connections
  - Results in dynamic neural oscillations
  - No fixed states of activity
- In artificial DCNNs, recurrent activity is usually implemented by a neuron activating ITSELF
  - A simple way of representing time and memory for time dependent tasks (e.g. language processing)
  - Biological neurons never synapse with themselves
- Recent experimental artificial networks are investigating lateral and top-down connections
  - Resulting interactions make the network inherently time-dependent
  - Image classification performance in difficult task improves considerably
  - More computationally intensive
    - More extensive filters
    - Multiple time steps modelled

28

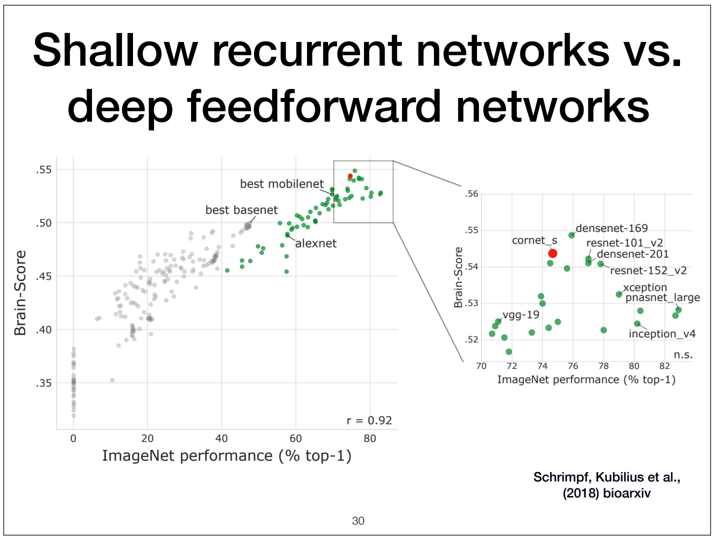
Bottom-up, lateral and top-down (BLT) artificial deep networks



Spoerer, McClure, & Kriegeskorte (2017) Frontiers in Psychology

29

In a network with recurrent connections, the activity is processed by the same layers many times. This network has only 2 hidden layers, but cycles activation patterns though each 4 times, effectively making 8 layers that responses can be processed through.
Also note that the path of information is not linear, there are longer routes and shorter routes so some of these steps can be skipped.
The layers have the same connection weights each time.
The connection weights will reach some kind of compromise between what is most effective for the first pass, the second pass, and the third and forth.
So a relatively shallow network can effectively be made much deeper using recurrent connections, though some compromises may be necessary.



Therefore, shallow recurrent networks can perform much like very deep feedforward networks.

The network ‘cornet-s’ is a 4 layer recurrent network.

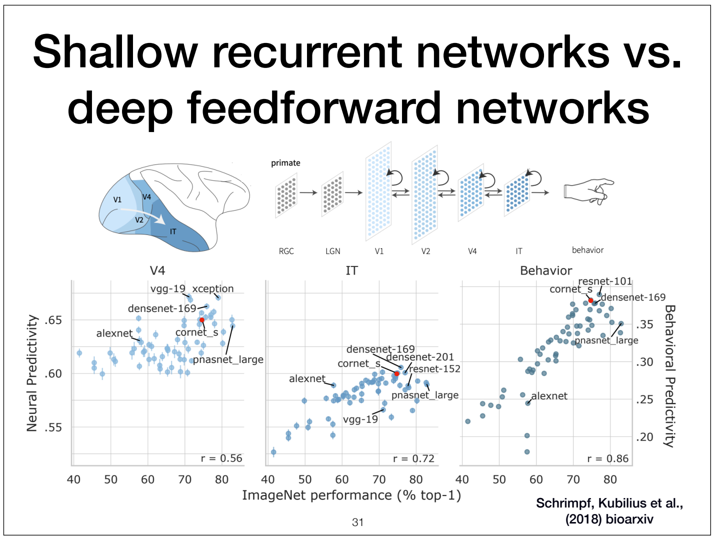
Its performance on classifying the ImageNet image set of object photographs is comparable to much deeper, 100 layer networks.

Because recurrent connections allow activity to be analysed over and over again by one layer, shallow recurrent networks can act like deeper feedforward networks.

But in the context of the massive increase in feedforward network depth, it may more useful to think that a very deep feedforward network can act like a shallow recurrent network like the brain.

But there are far fewer weights to learn, as each cycle through a layer uses the same weights.

The other measure here, Brain-Score, summarises the ability of network units follow the activity of biological neurons and human behaviour. Here cornet does better than most far deeper feedforward networks.

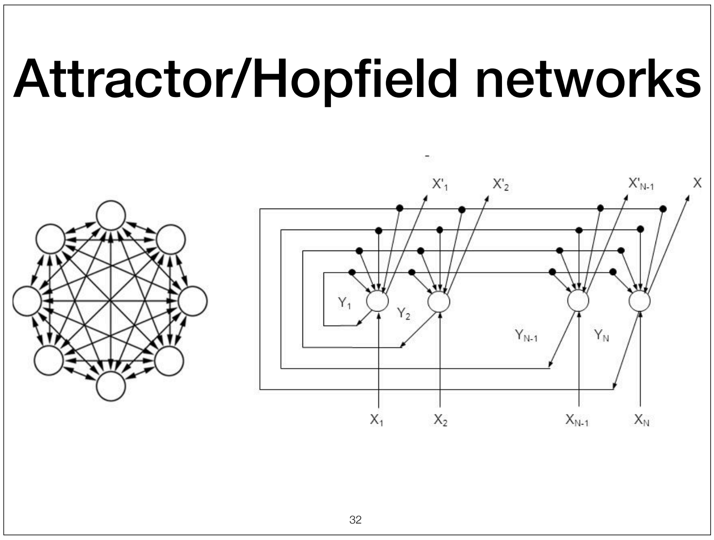


The architecture of cornet is designed to mimic that of biological neural processing’s recurrent structure, so perhaps it is not surprising that it mimics the brain’s response closely.

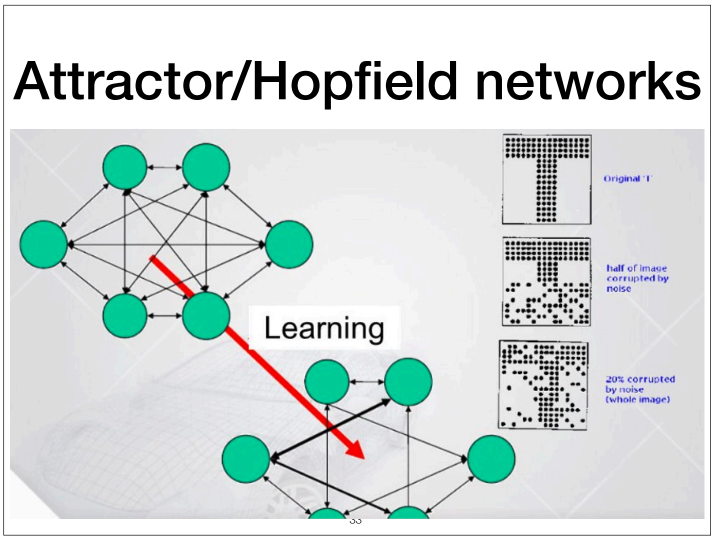
The pathway to inferior temporal cortex only contains around 4 steps, so a 100-layer network is clearly not a close match for the brain.

Many networks can predict V4 activity well, often better than cornet. By IT, only two networks perform better, each with over 100 layers. In predicting human behaviour, only one

network performs better.
As well as giving extra depth through recurrency, feedback connections seem to be carrying some useful information. What does this information look like?

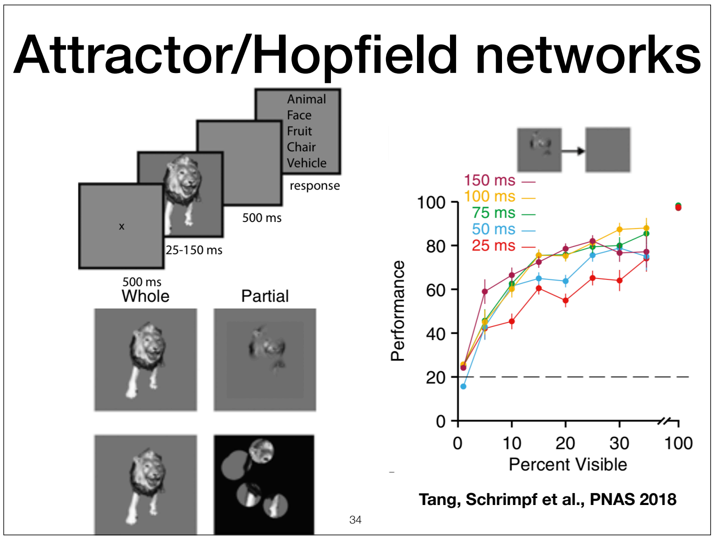


It has become clear that something very interesting is happening in the lateral (within-layer) connections in biological networks.
An incoming signal generates a pattern of neural activity in the higher layer.
Here, all of the neurons in this higher layer are connected to all others, as shown here in two common notations of the same structure.

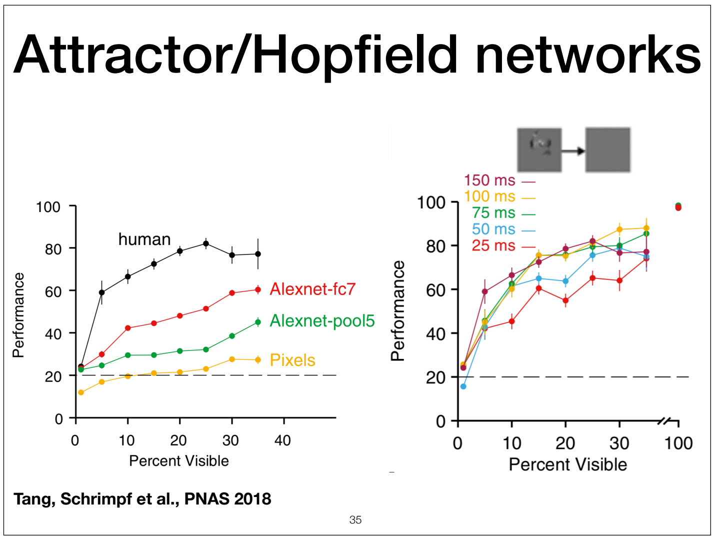


The result of this organisation is that frequently seen patterns of activity produce strong connections between the activated group of neurons, due to Hebbian learning.
After these common patterns are learned, a new incoming pattern will cause a sequence of recurrent activity that attracts the network's state towards a common pattern.
Neurons that are usually activated together in the higher layer all become active, completing the pattern of activity based on what they

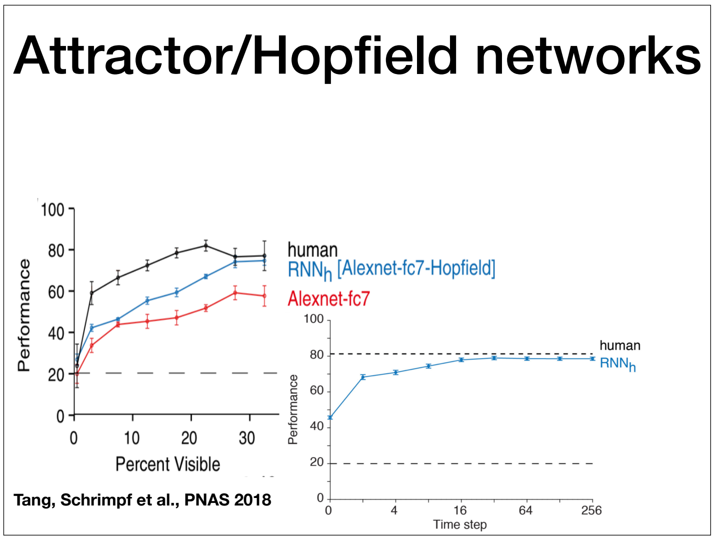
usually see.
So if an incomplete pattern comes in, it is completed based on experience.



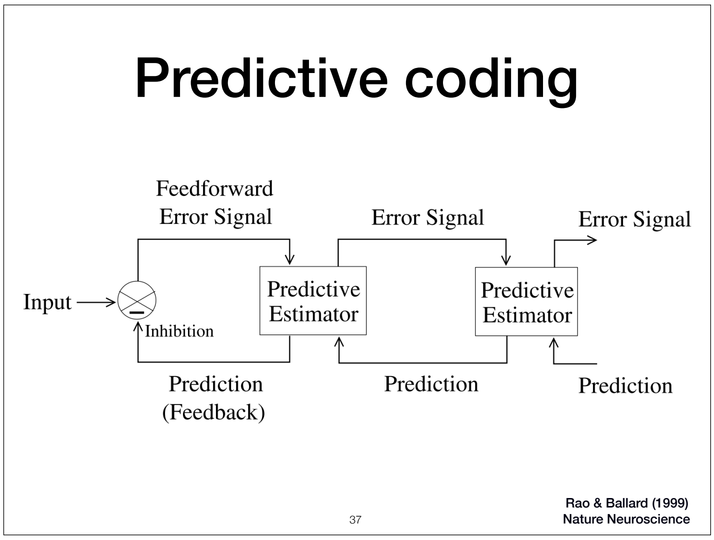
Humans are very good at object recognition, and we also do well when much of the object is hidden.
With only 30% visible and only presented for 100 ms, we are 80% correct.



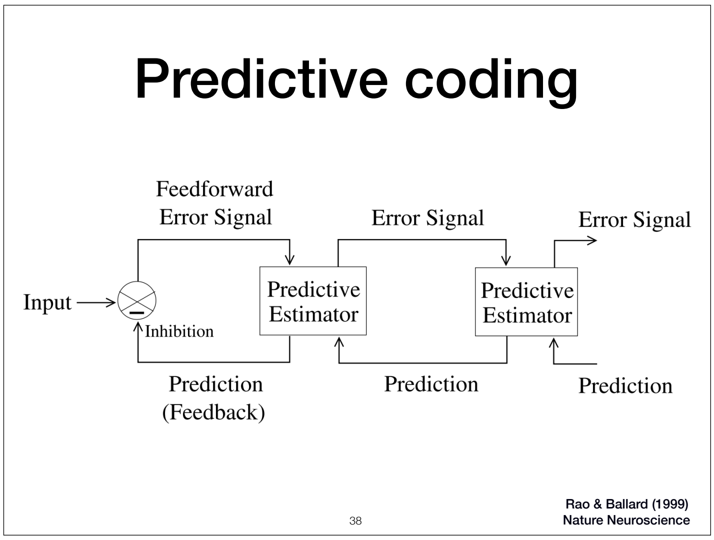
Feedforward networks don’t do that: Their performance is poorer, and increases approximately linearly as more of the object is shown.



But a recurrent Hopfield attractor network does approach human performance at classifying partial images.
If we track how this performance emerges over different time steps, it relies on this recurrent interaction.



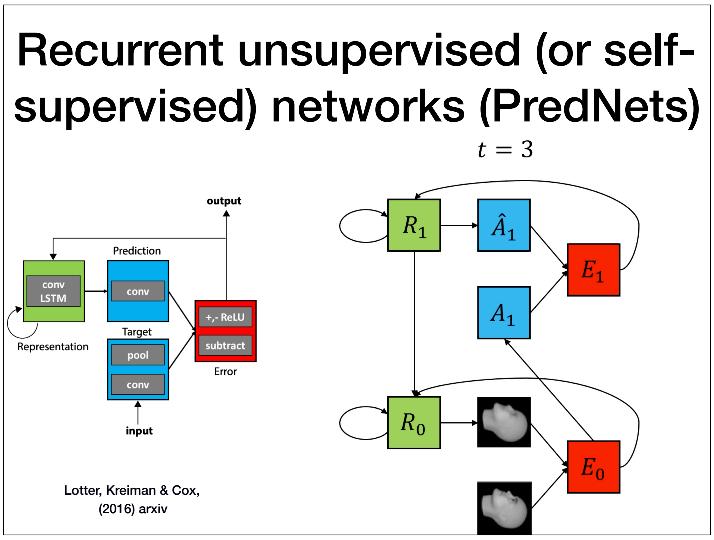
This completed pattern of activity based on experience is often described as a ‘prediction’ of what is coming in. Completing this prediction seems to be the important role of lateral connections.
This completed prediction is then fed back to the lower layer, and inhibits activity that fits with this prediction: ‘I think I recognise this, you can ignore it’.
This feedback connection therefore carries the prediction and allows it to interact with incoming signals.
Due to this specific inhibition of activity consistent with the prediction, the feedforward signal becomes only the parts that do not fit with the prediction: the prediction error.
This happens repeatedly in feedforward-feedback interactions between pairs of layers, though again it gets hard to think about beyond the first layer: it becomes the prediction error on the prediction error signal.



So the first feedforward sweep of activity encodes the incoming stimulus and learns common patterns in it, as we have already seen. But the recurrent loops of activity instead represent the error between the incoming signal and what is expected from previous learning. Because the activity in the network then responds to this error, Hebbian learning can operate on the error signal: connection weights change depending on the prediction error. This incorporates new patterns of activity into the predictions that can be made, by making strong connections for unexpected inputs. This is remarkably similar to backpropagation of error. The main distinction is that the error is based on the difference from something the network has seen before, not the difference from the expected input classification.

In other words, the error is analysed at every layer, not just determined from a final output.

In neuroscience, predictive coding and the related concept of Bayesian inference have been important concepts for the last 20 years. Predictive coding is consistent with many unexpected aspects of neural activity that are difficult to explain without it.



what is trying to do is classifying what is going to be displayed next

A new type of recurrent networks, PredNets, have recently been introduced to implement predictive coding. These are unsupervised networks: they predict future activity based on current activity and previous experience.

Let's see how this plays out in a simple 2-layer network, with layers 0 (where the input image comes in) and layer 1 (a higher layer).

The network initially has no input and so no activity. So the error state in layer 1 is effectively zero.

This feeds into a recurrent circuit of layer 1, but nothing much happens, because the activity is all zero. This feeds into the recurrent circuit of layer zero which also takes an input from the empty error circuit of layer 1 (again little happens).

The prediction at this point is an empty image, but an image of a face comes in. The empty prediction is then subtracted from the face image, and so the image is passed forward unaltered in this first feedforward sweep.

However, at this stage the network starts to generate predictions of what comes next, based on previous patterns of activity in the green attractor circuit that are similar to the incoming image. These get subtracted from the incoming image, determining the difference and so allowing the prediction to be refined by further steps.

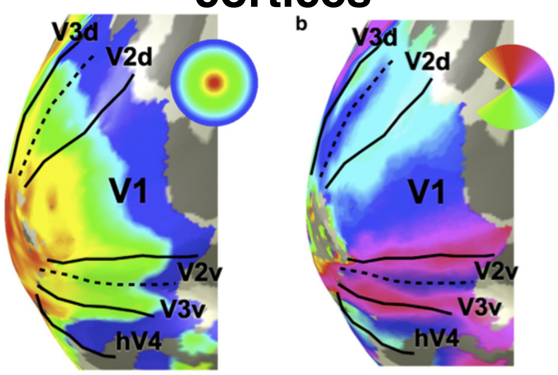
PredNets making predictions

Here we see the layer 0 (image) prediction of the next frame in driving data.
Here the network was trained on a database of driving videos, and is making predictions for videos that it has never seen.
“The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths.” Ada Lovelace

Attractor networks and predictive coding

- Recurrent connections effectively make networks deeper
    - Same analysis repeated by recurrent cycles
    - Each layer performs multiple layer operations
        - But each with the same set of weights
    - Matches neural architecture and activity more closely than deeper networks
- Attractor/Hopfield network among lateral connections
    - Attract activity patterns towards previously common states
    - Complete or predict noisy or incomplete input patterns
    - Requires multiple recurrent cycles
- Predictive coding inhibits input activity that matches predictions
    - Initial feedforward sweep that lacks a prediction
    - Network activity then signals prediction error
        - Hebbian learning follows the error signal, like backpropagation

Map organisation sensory cortices



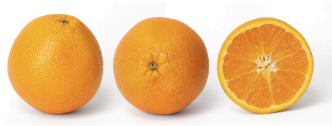
-We started the last lecture by introducing the map layout of the visual cortex.
-Here we see colors representing preferred positions in visual space, overlaid on the occipital lobe’s anatomy
-This reveals that visual position preferences change gradually across the brain, forming multiple maps of visual space on the cortical surface.
-This map organisation allows transformations of features to focus on patterns across nearby spatial location.
-Within these maps, neurons with similar responses are grouped together
-So the response of each MRI recording site, or VOXEL, can tell us a lot about the many similar neurons inside
- In fact, these maps don’t really represent space directly, but rather the position on the sensory organ, the retina.

- But outside of sensory systems like these, we still didn’t have a good idea of how the brain processes and organises information.

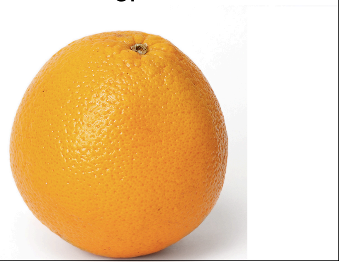
-Most of my current research aims to reveal how the brain’s cognitive functions are processed and organised.
-When we started thinking beyond primary sensory properties, it seemed quite possible that such map structure could only follow sensory organ structures, as the neural pathways from the sensory organs already follow sensory organ structures.
-Alternatively, when responses to cognitively interesting features emerge, we may also see a map organisation to allow easier communication between similarly-responding neurons and to allow patterns within those features to be analysed efficiently.

Numerical vision

- We enumerate small sets of objects (up to 6) quickly, accurately and confidently
    - Evolutionarily preserved
    - Allows decisions between greater and lesser options
- But other dimensions of quantity are also relevant in these choices



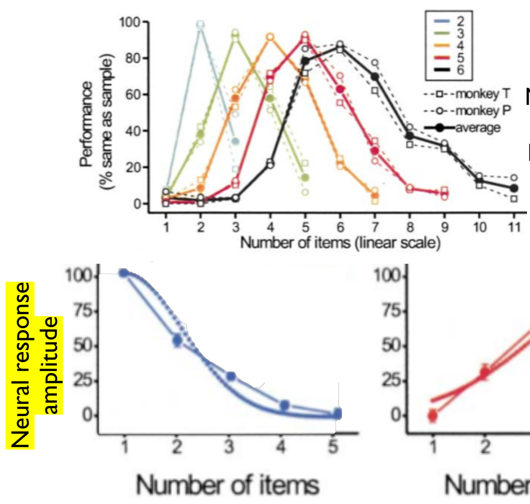
or



-Cognitive functions are appealing because they are often absent or very different in animals
-So to understand the neural processing underlying cognition, it's best to measure in humans.

-BEFORE CLICK: I started in this direction by looking at responses to numbers of visual objects, or numerosity

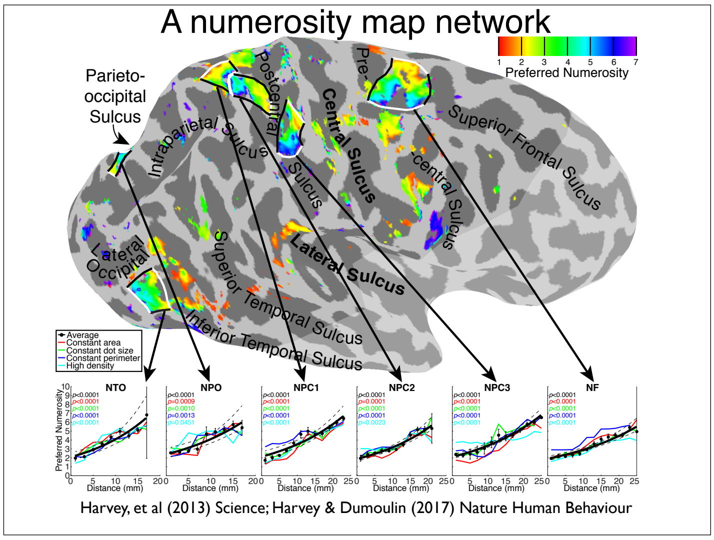
-AT END: We can easily describe the structure of quantities like numbers, size, and times: is there more stuff or less stuff
-Back in 2013, we began testing the idea that there may also be maps for quantities
-This was based on the efficiency of map structures for further analysis within convolutional networks
-We followed the straightforward prediction that the map structure would follow level of quantity: 'less stuff' represented at one end, 'more stuff' represented at the other
-This would group similarly-responding neurons just like the primary sensory maps, and importantly allow characterisation of the neural response selectivity of these grouped neural populations even at the limited spatial resolution of human fMRI



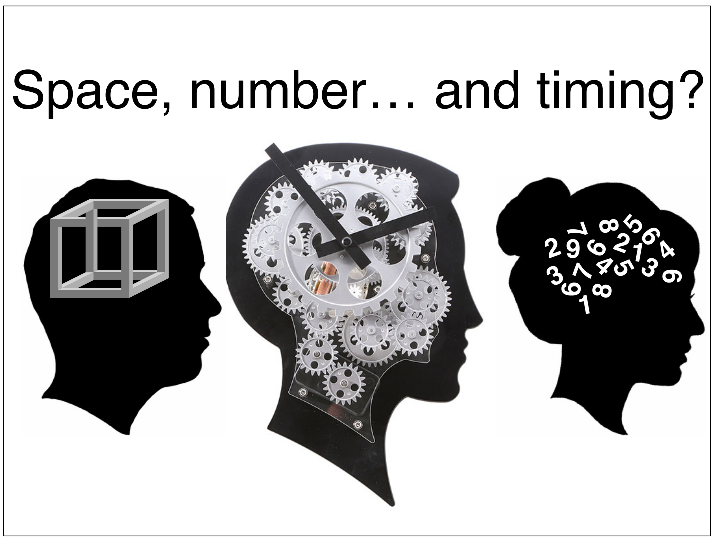
Nieder & Miller, *Science* (2002); Neuron (2003)

We already knew some interesting properties of numerosity perception and processing.
In behavioural tasks, macaques (like humans) are very accurate with small numbers
With increasing numerosity, accuracy declines.
Like the neurons responding to visual position or orientation, neurons in the brain show tuned responses to numerosity, where response amplitude decreases gradually as we move away from the preferred numerosity.
In feedforward artificial networks

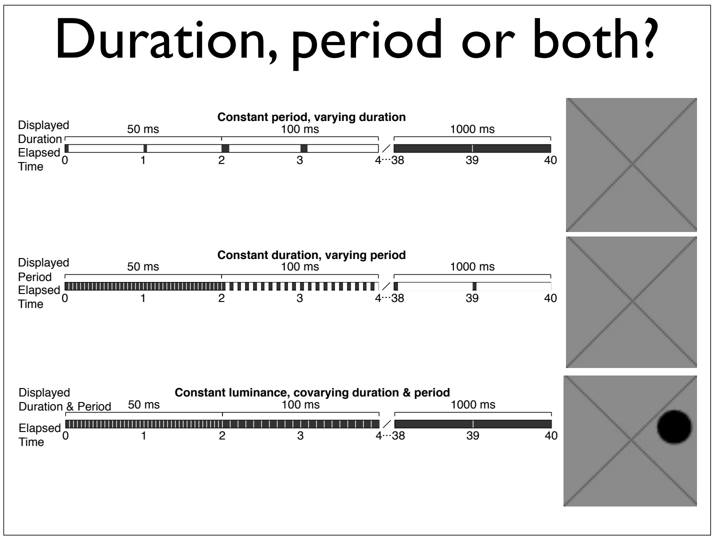
trained for object recognition, it has numerosity-selective units have also been found recently.



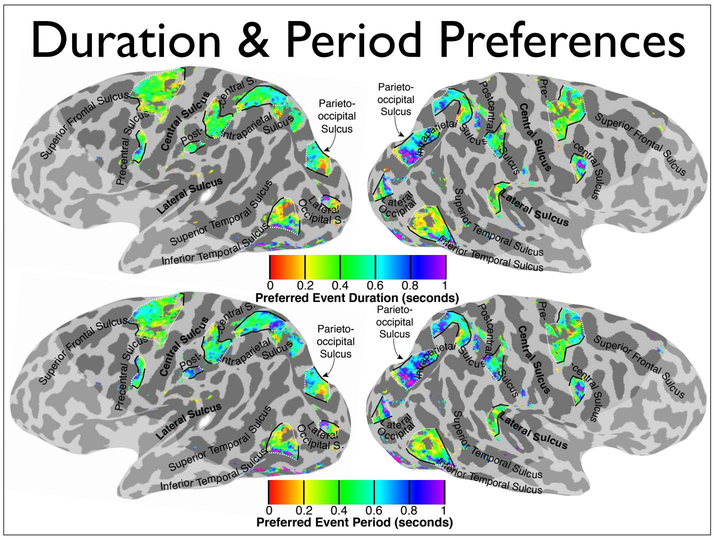
–When we measured the responses to different numerosities at each recording site in the brain, we saw the preferred numerosity of recording sites (colors) changes gradually across the cortical surface.
–So maps can emerge in the brain to represent continuous features that do not follow the structure of sensory organs.
–We are interested to test whether patterns of lateral connection weights in networks with attractor layers also predict this organisation, and I am looking for a student to test this idea in a thesis project.



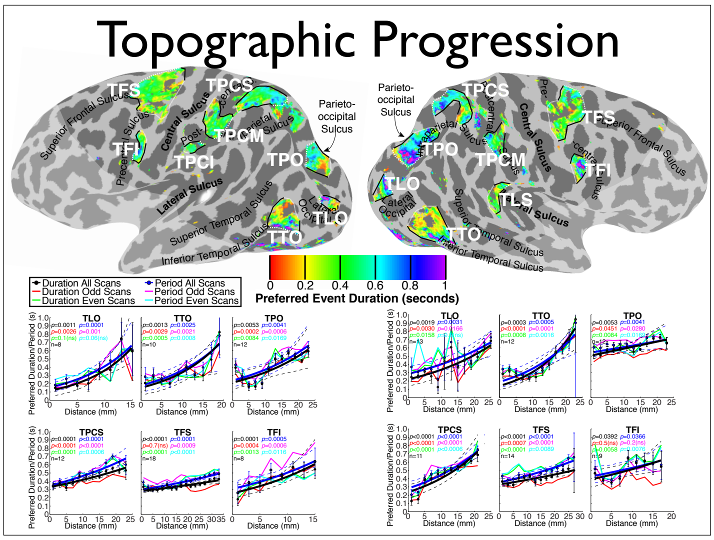
We have since extended this approach to event timing.
Following similarities in perceptual properties, we hypothesised that processing of event timing may rely on similar mechanisms to space and number processing.



We wanted to test the idea of timing-tuned neural populations using fMRI.
We gradually varied two parameters of the timing of these events.
First was the duration that a dot was on the display, the time between the event's onset and offset.
Second was the period between two dot appearances, the time between two onsets.
We also include a condition where we change event duration and period together, so there is always a dot on the screen and luminance remains constant
We then analysed the responses to these changes in timing to see how the neural populations in the brain respond to timing



If we project the preferred durations and periods on the cortical surface, we see that these areas contain a range of duration and period preferences.



Furthermore, within these areas the preferred duration and period gradually and repeatedly change across the cortical surface, forming a series of topographic maps of neurons with gradually varying preferences for event timing.
So here we are interested to test whether recurrent networks contain timing-selective units, how duration and period preferences are related, and whether their lateral connection patterns are consistent with map organisation.

Space, number and timing

- Representations of visual position, body position and auditory frequency are all spatially mapped on the cortical surface
  - Follow the structure of the sensory organ
  - Allows analysis of local features using spatially-limited convolutional filters
- Representations of object number and event timing also spatially mapped
  - Although there is no sensory organ with this structure
  - Likely to result from strong lateral connections between similarly responding neurons
- This allows analysis of the quantities by similar neural operations
  - Emergence of maps cognitive maps allows hierarchy to begin again

50

Neuroscience results are becoming more and more CS results as if we can make a computer/RNN do it then that's how our brain does it!

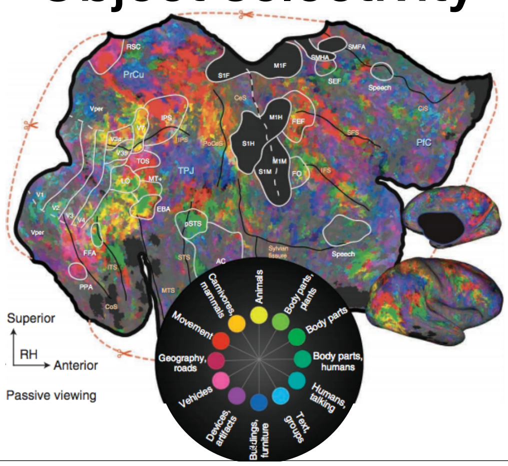
If you are interested in a thesis research project in these directions, get in touch.

DON'T STUDY

Now read this:

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron, 76(6):1210-24.

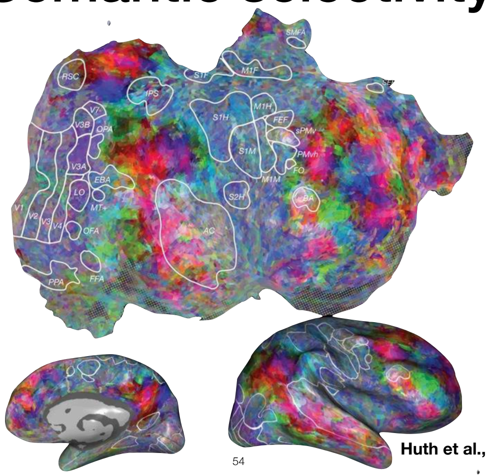
Object selectivity



Now read this:

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 532(7600):453-8.

Semantic selectivity



Huth et al., Nature, 2016

What we see for visual processing recognition is the same as language processing