Machine learning 2: biologically inspired learning

Dr. Ben Harvey
Experimental Psychology

b.m.harvey@uu.nl

The fMRI experiments I described in the last class aimed to discover patterns in large, complex fMRI data sets, or to test specific hypotheses about patterns in these data.
One problem here is that the resulting models don't show how the brain is processing the inputs, only that there is information in a particular area or that one thing is represented more or less similarly to another.

If we want to model a system, what do we already know?

- We have a large, complex data set, but do we:
  - Know nothing about it
    - Easy to find patterns, difficult to interpret
  - Have a specific hypothesis we want to test
    - Harder to set up learning model, but easier to interpret
  - Already know much relevant information
    - Should consider this in our learning model
    - Likely to constrain model inputs and structure
    - More powerful (statistically)

The experiments we look at today aim to reveal mechanisms of human neural processing, rather than just where we see information in the pattern of responses. These are two very different aims, and depend on what we already know.
In the case of the fMRI in the visual cortex, we know a lot about the system we are trying to model, so it is useful to constrain our learning model's structure using this information.

Biologically inspired analysis methods

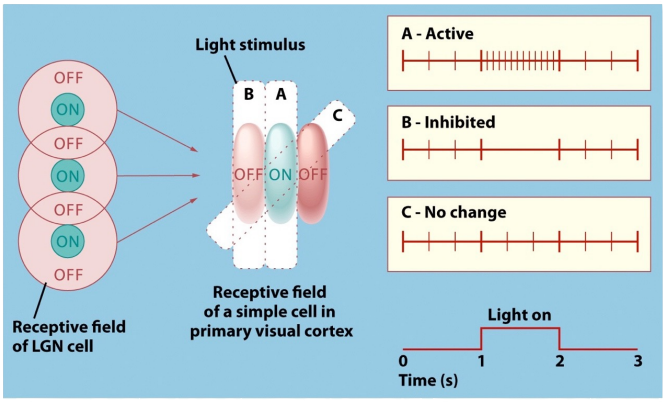
- Methods from class 1 use statistical similarity without taking into account the properties of the system (brain)
- Other approaches include known properties of the system
  - Organise the machine learning model following these properties
  - Learn the parameters of that model
  - i.e. how information is encoded in the biological system (here, the brain)
- Then predict how the brain will respond to a new input based on the parameters of that input

At END: One important distinction here is that the previous class looked at decoding models: can we make a model that predicts what responses we expect to see in the test set?
The models we look at today are called encoding models: they model how the brain encodes information. I find this a much richer, deeper goal.
To do this, it makes sense to include what we already know about how the brain encodes information.
An encoding model can also predict the responses we would expect to see in the test set: an encoding model can be used for decoding. However, a decoding model CANNOT really show how there brain encodes information.

Today's Topics

- Visual space encoding models
- Object number encoding models
- Event timing encoding models
- Object identity encoding models
  - Deep convolutional neural network models

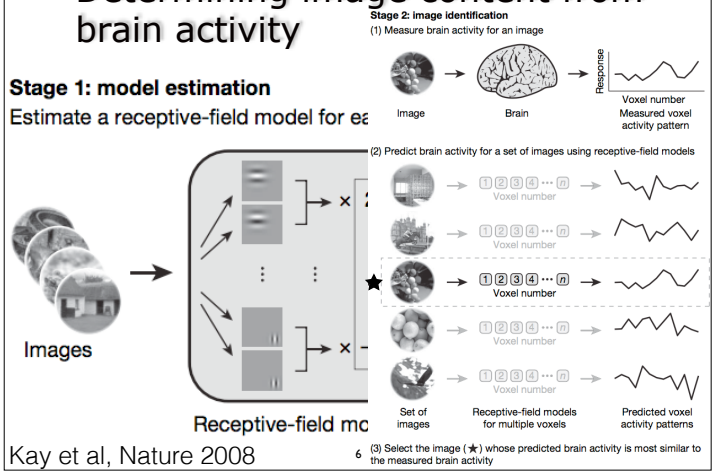
Neural responses to edge orientation



5

You may remember these orientation-selective cells from the last class.
These respond to edges with a particular position and orientation, and both position and orientation affect the responses of fMRI recording sites in early visual cortex.

Determining image content from brain activity



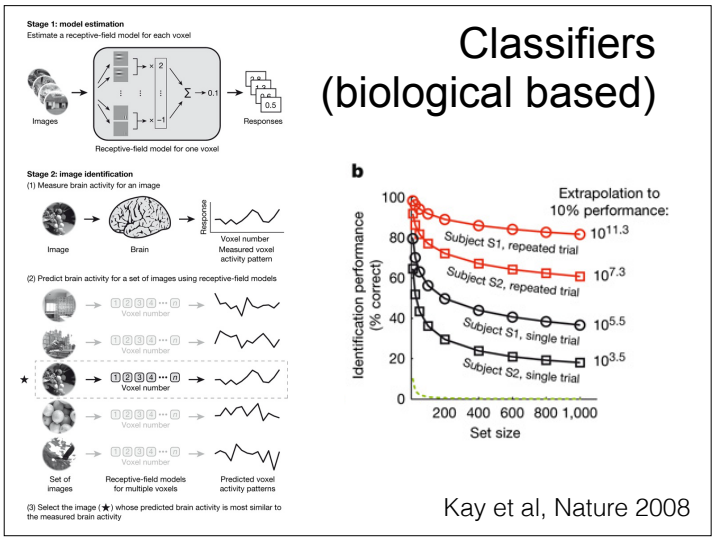
Kay et al, Nature 2008

6

By looking at the brain activity in primary visual cortex, we can make a model of each recording site's preferred position and orientation.
We do this by determining where the edges in an image lie, when the image was shown, and when the recording site responded.
By looking at these factors for a large set of images, the model gains information about which combinations of orientation and position produce a response in each recording site.
Specifically, each orientation and position is given a 'weight' for each recording site, and the weight is learned from the responses to training set images. (CHECK FAMILIARITY WITH THIS CONCEPT).
So here the machine learning approach aims to learn the parameters of a specific model of the type of neural responses we know

describes the early visual cortex's responses accurately.

From the resulting pattern of weights, the researcher can then predict the pattern of responses we would expect for a new image, from an independent test set that has not been shown before.
We can then compare this prediction to the observed pattern.
The best fit is most likely to be the image the subject was viewing.



Classifiers (biological based)

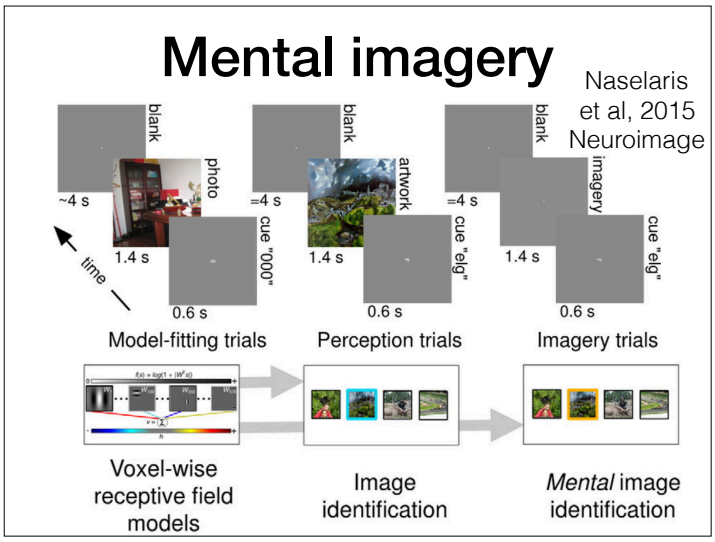
Kay et al, Nature 2008

This approach identifies the displayed image far above chance (shown by the green line at the bottom of the right graph).
Performance is still far above chance when the image is only shown once, the black lines are for single trials.

Note here that the images are chosen from a set of possible candidates.
It is not possible to reconstruct what the image looked like, only to compare the pattern of responses against some predictions for a set of candidate images.
This is a limitation, but we can use an exceptionally large set of candidate images, for example 'all the images on the internet'.

Because the performance at candidate image identification is still high from single presentations of an image, we can use very even do this for movie clips.
We can make the candidate set a very large set of movies, for example every two seconds of footage from Youtube.
For each of these clips, we can make a predicted pattern of fMRI responses, by looking at the positions and orientations of the contrast they contain.
We can then see which movie clips best fit the observed pattern and average them together to predict what the displayed clip is likely to have looked like.

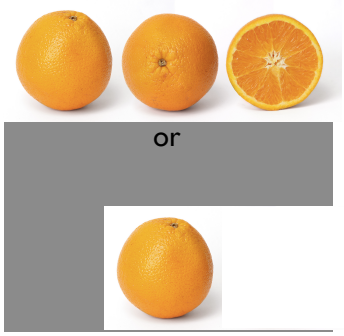
Here, we have implicitly given information about what movies look like, together with an explicit model of how neurons respond.
Therefore, faces generally appear white and the sky appears blue, although the measurements have no information about color.
-Remember here that simple machine learning could determine what orientation people were looking at, and that was exciting and impressive.
-Including known biological properties of the system quickly allowed researchers to reconstruct movies of what people were looking at from the activity of the same areas.



You may remember that MVPA decoding could identify the orientation that people were imagining as well as what they were seeing
-Using the same approach, we can first train our response model on trials where we present visual images.
-As before, this allows the model to identify which image of a candidate set the subject was seeing.
-We can then ask subjects to imagine one of these 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 the neural activity in the early visual cortex.

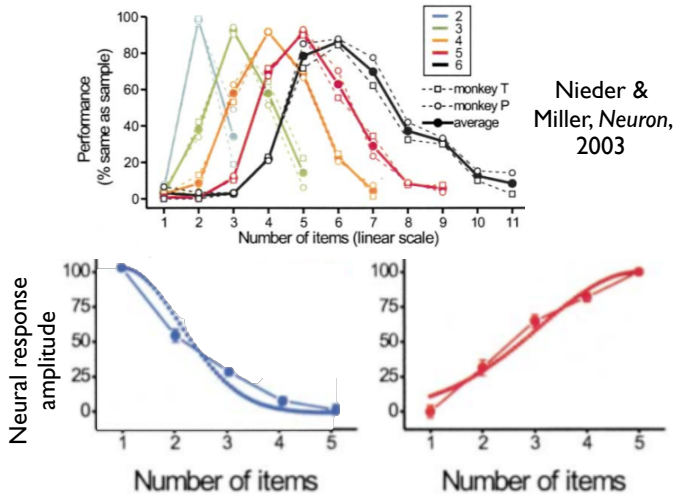
Number processing

- We enumerate small sets of objects (up to ~5) quickly, accurately and confidently
  - Evolutionarily preserved
  - Allows decisions between greater and lesser options



-We can therefore derive some good models of simple visual processing from responses to a training set of natural images, by including a biological model whose parameters we want to estimate. The resulting models make accurate predictions about responses to new images in the test set.
-This gives us a lot of information about simple visual processing, but now let's look at a more cognitive function, number processing

---

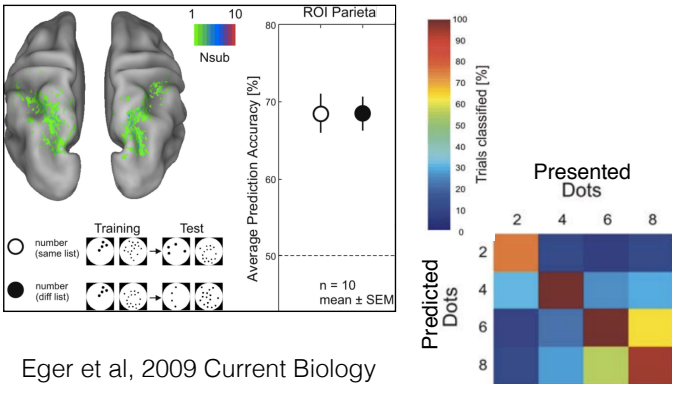


Nieder & Miller, *Neuron*, 2003

So, if we want to investigate the representation of numbers with a similar encoding model, what do we know about the brain's response to numbers of objects?
In behavioral experiments, monkeys (like humans) are very accurate with small numbers. With increasing numerosity, accuracy declines
In monkeys, we find neural responses to numerosity: responses decrease with distance from preferred numerosity. Broader tuning for higher numerosity preference

---

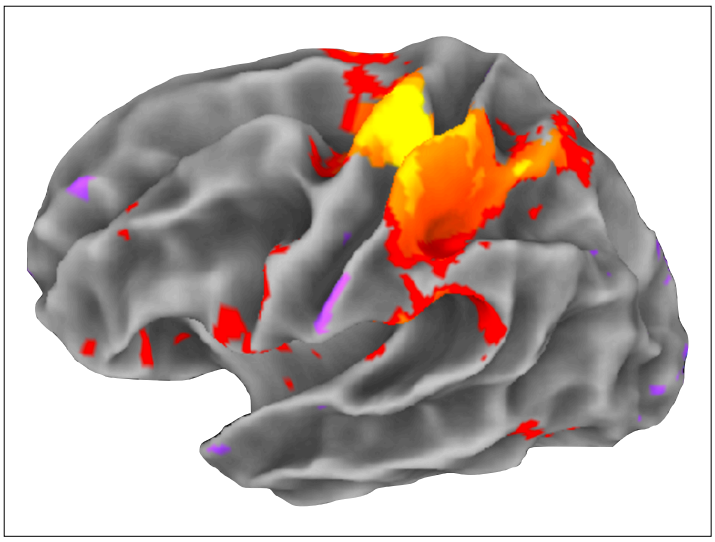
Number representations with SVM classifiers



Eger et al, 2009 Current Biology

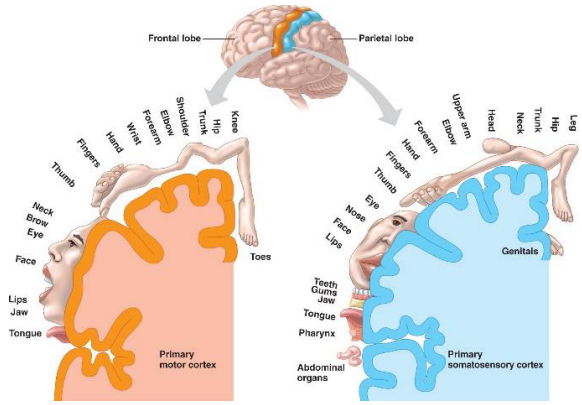
Here we see an SVM classifier in the parietal cortex can decode which number was shown with about 70% accuracy
If we also look at the mistakes made, this shows us that errors normally choose for a nearby number, not a far number
These errors suggest there is some relationship between neural activity responding to nearby numbers: specifically, representational dissimilarity decreases with numerical distance.

So this shows us there is some spatial pattern in responses to numbers in these areas. But what is this pattern?



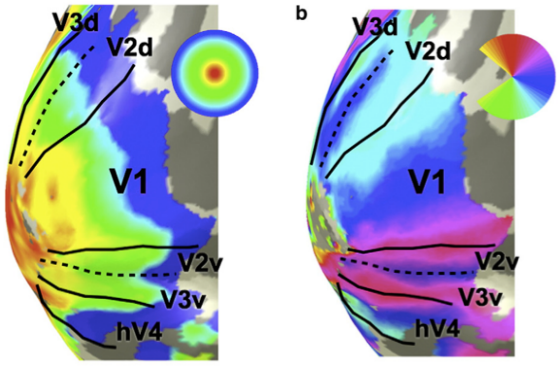
We had some ideas about this from other systems in the brain
Here we see the areas that are active when we move the right hand.
This is a large area of the brain, and knowing the location of these responses doesn't tell us much about what is happening in this area, or how it is organised.

Map organisation sensory & motor cortices



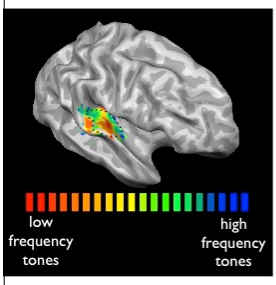
We have already seen than the somatosensory and motor areas of the brain map the body's organisation onto the cortical surface.

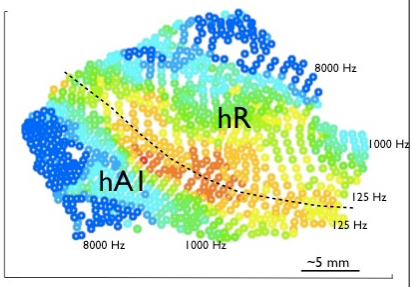
Map organisation sensory cortices



We see something similar in the brain's 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, again forming multiple maps of visual space on the cortical surface

Map organisation sensory cortices





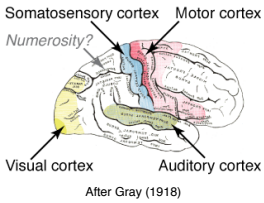
- This doesn't just work for position though. In the auditory cortex, the map reflects auditory frequency
- The colors here represent the frequency that gives the largest response at every recording site

- In fact, these maps don't really represent space directly, but rather the position on the sensory organ.
- For the human brain, visual field position is better understood as a location on the retina, while responses to auditory frequency are responses for locations on the cochlea, where different locations vibrate at different frequencies.

-Within these maps, neurons with similar responses are grouped together, and this leads to map organisation

-But outside of sensory systems like these, we still didn't have a good idea of how the brain processes and organises information.
-Often, we only know which large areas are activated by a particular stimulus class or task, or which large areas contain information about these.
-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.

Biological models of number processing

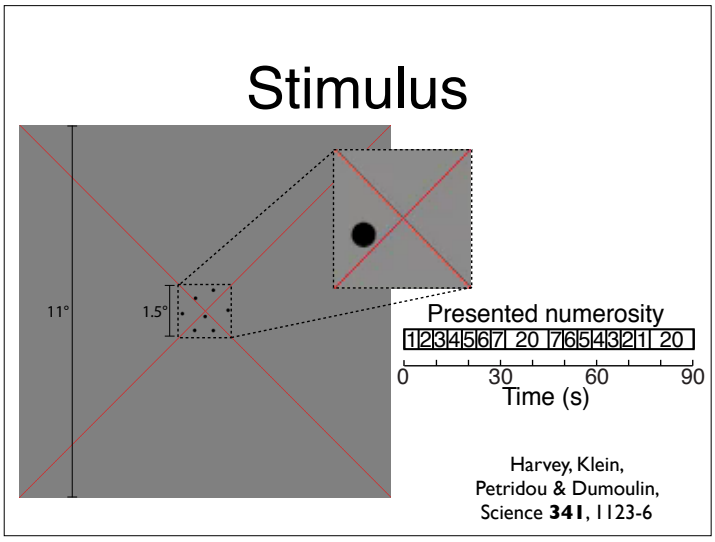


After Gray (1918)

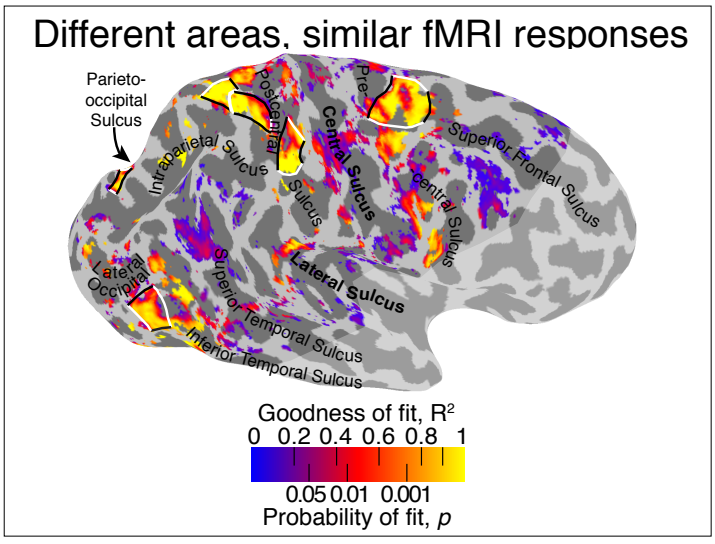
- Neurons with similar properties interact
  - Allows useful computations
- Organisational consequences
  - Topographic maps (retinotopy, tonotopy, somatotopy)
- Are responses to numerosity also mapped onto the brain's surface?
  - Using fMRI (7T) & model-based analyses

This is an important question for neuroscience.
We don't have sensory organs with numerical structure, so maps organised by numerosity preferences would demonstrate that the same organisation can be found in cognitive processing systems.
-Maps like these group similarly-responding neurons, so that the responses of a large group of neurons can tell us a lot about the individual neurons they contain. This would allow us to study the

computational properties of cognitive processing even at the limited spatial resolution of fMRI.
-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.

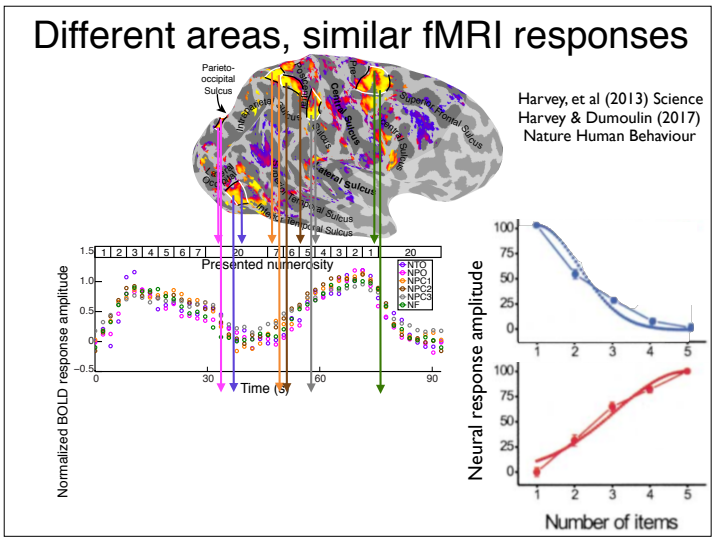


We tackled this question by showing visual stimuli that slowly changed in numerosity (object number) and relating the timing of responses to the timing of the presented numerosity.

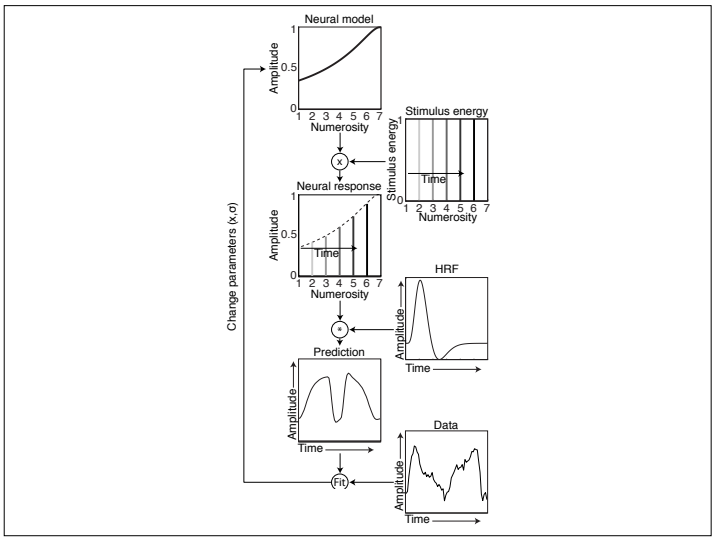


We see large responses to changing numerosity in several specific areas, and little response elsewhere

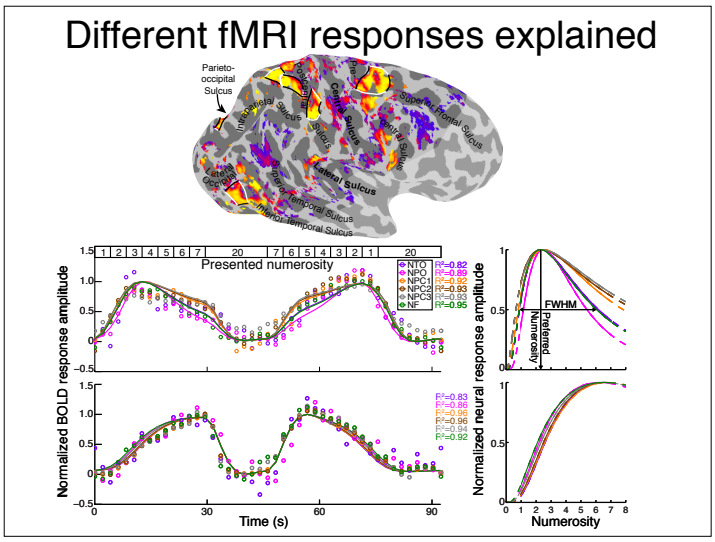
So this already shows us WHERE neurons are responding to object number (numerosity).
But this experimental design allows further analyses to show HOW these sites are responding



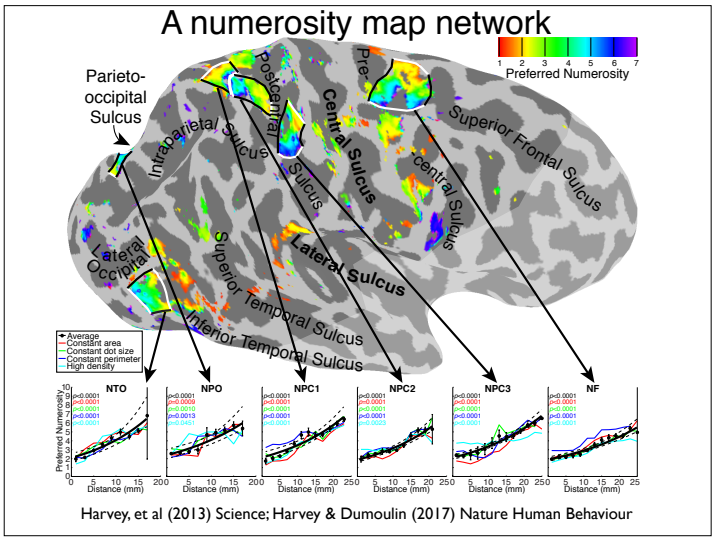
If we look within any of these areas, we see some very different types of response
In the upper trace locations, we see large responses just after presenting small numerosities.
In the lower trace locations, very close by, we see large responses following large numerosities.
So, we want to model these responses. We know that these areas are likely to be tuned for numerosity, with response amplitude peaking at a particular numerosity and dropping with distance from that numerosity.
Therefore, we will make models of tuned responses, and learn each recording site's tuning parameters.



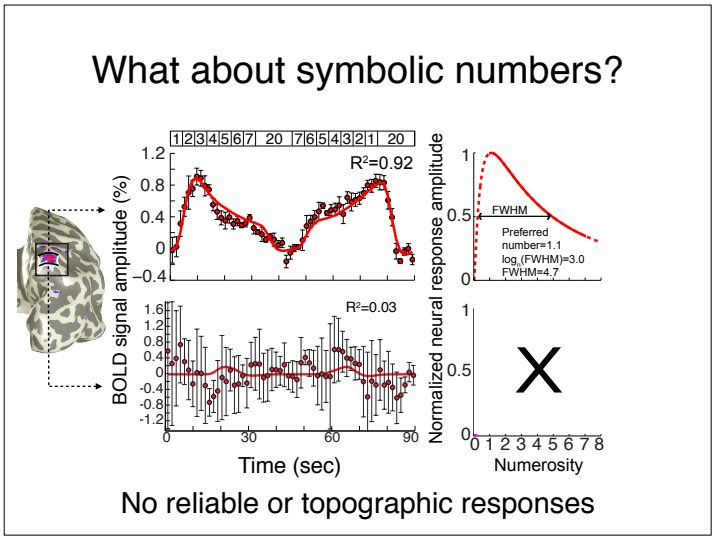
We describe this response using a model of the population neural response, together with the stimulus time course.
Neural model has two parameters: preferred numerosity and tuning width
We took this model from the known response properties of numerosity-selective neurons in monkeys.
We examine the overlap of the neural model with the stimulus sequence to predict the neural response time course if this neural model saw this stimulus sequence.
We then convolve this with an HRF to predict the fMRI response time course we would expect from a recording site with this tuning.
BAD MATCH between prediction and measurement
CHANGE MODEL RESPONSE SELECTIVITY, GOOD MODEL FIT
So again our model is using a biologically inspired model and computing the best fitting model parameters



This reveals the numerosity tuning model parameters that predict the timing of the responses we see.



-In each of these areas, the preferred numerosity of recording sites (colors) changes gradually and highly significantly across the cortical surface
-We repeated this on different days using stimuli that control for low-level, non-numerical features that sometimes change with numerosity, the colored lines.
-This is an informal test of model performance: model parameters are similar in on different recording sessions with different relationships between numerosity and the visual image
-So this already shows some information about how numbers are processed: neurons with similar response preferences are grouped together to allow easy interactions

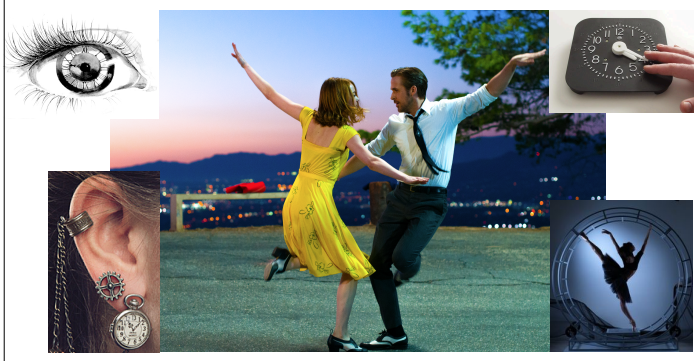


We can also use the same design to look for responses to symbolic numbers, written numerals.
If we see similar responses, this would tell us that written numbers are represented similarly to the number of objects we are seeing.
This is not possible to study in animals because they do not understand written numbers as humans do.

AT END: So this tells us that written numbers are not represented the same way as numbers of objects.

It seems these neural populations are responding to the visual impression of numbers of items, rather than the cognitive concept of number.

Timing is everything!



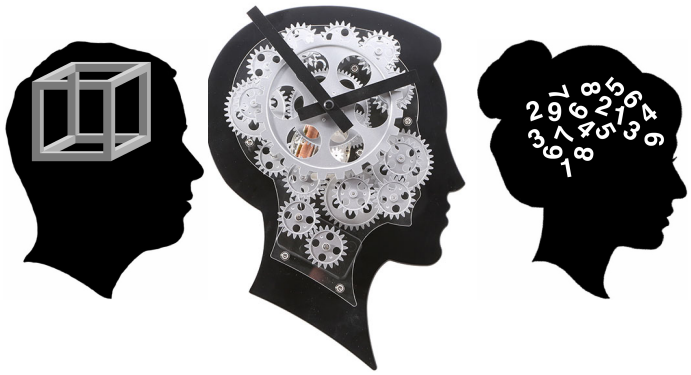
When we need to understand and interact with the world, timing is very important.
Humans combine timing from our senses of vision, hearing and touch into a unified temporal perception,
And make skilled movements that follow this timing.
Such timing perception is most relevant for short events: most dynamic events last under 1 second

---

Timing in the brain?

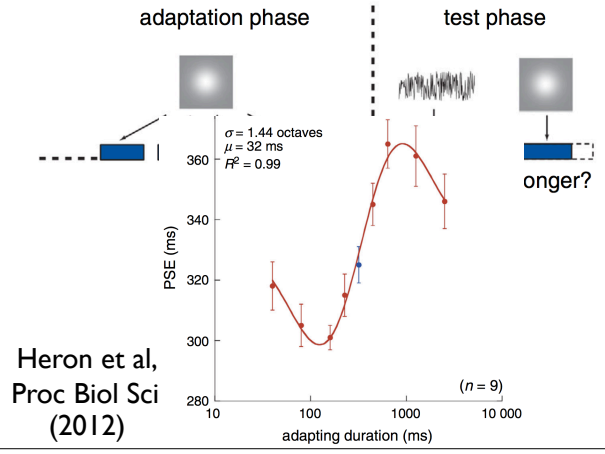
Despite the central role of timing in perception, multi-sensory integration and action planning, the mechanisms of temporal processing in the brain are only beginning be understood.

Space, number… and timing?



Following similarities in perceptual properties, we hypothesised that processing of event timing may rely on similar mechanisms to space and number processing.
We already knew that visual space and object number are represented as tuned responses organised into topographic maps.

Psychophysical duration adaptation

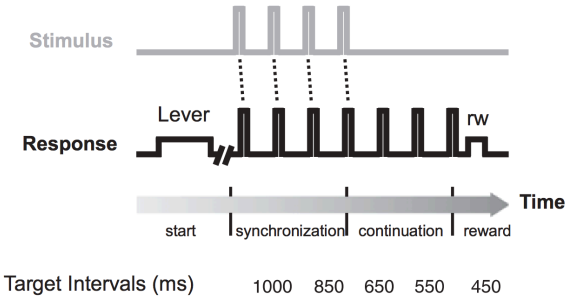


Heron et al, Proc Biol Sci (2012)

Some recent findings suggest there may be similar tuned neural responses for visual event timing. First, if humans are adapted to repetitive presentation of visual events with a specific duration, this repels the perceived duration of subsequently presented events within a nearby duration range. Repulsive aftereffects like this are thought to reflect neural tuning to the adapted property.

Neurophysiology of motor action timing

Synchronization-Continuation Task



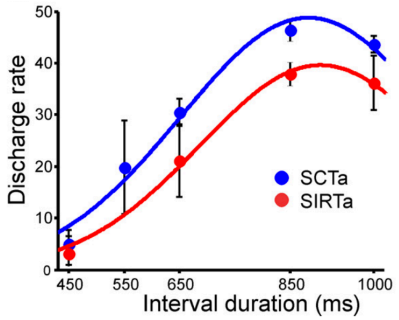
Target Intervals (ms) 1000 850 650 550 450

Merchant et al., PNAS (2011); J Neurosci (2013)

Also, in the supplementary motor area, there is tuning for the timing of motor events.
Macaques can be trained to continue repetitively tapping a lever after synchronising their movements to a repeating visual flash or auditory beep

Neurophysiology of motor action timing

Absolute motor timing cells

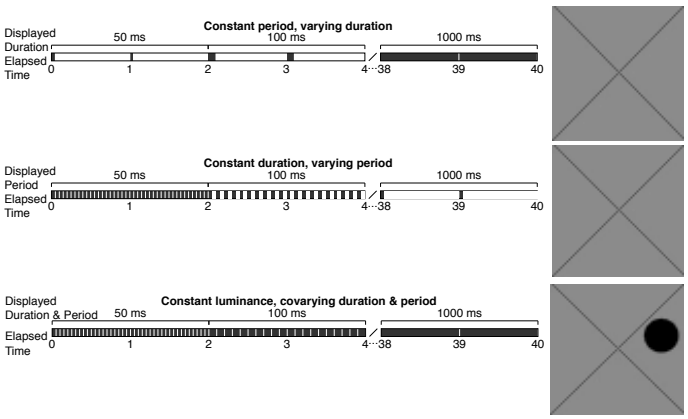


Merchant et al., PNAS (2011); J Neurosci (2013)

This reveals tuned responses to the interval between their movements.
This is also tuning for event timing, but for the period between movements rather than the duration of the movements

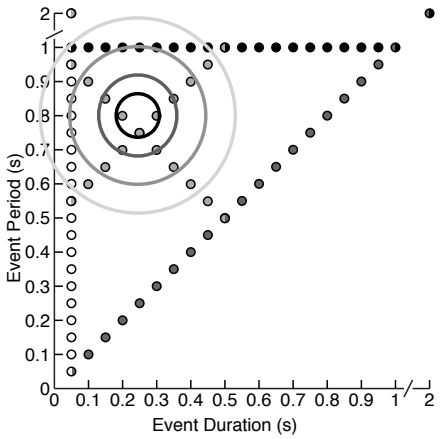
---

Mapping timing selectivity



We wanted to extend the idea of timing-tuned neural populations using 7T 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.
The subject's task was simply to press a button when a white dot was presented instead of a black dot.
No timing judgements were required.

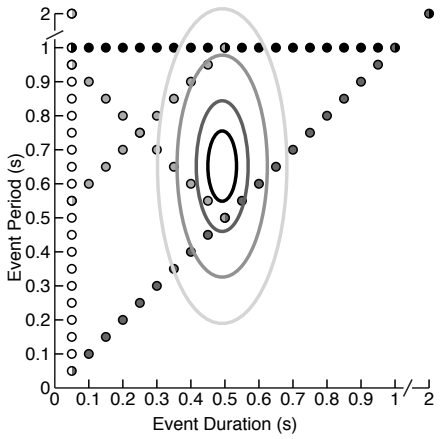
Tuning for visual timing



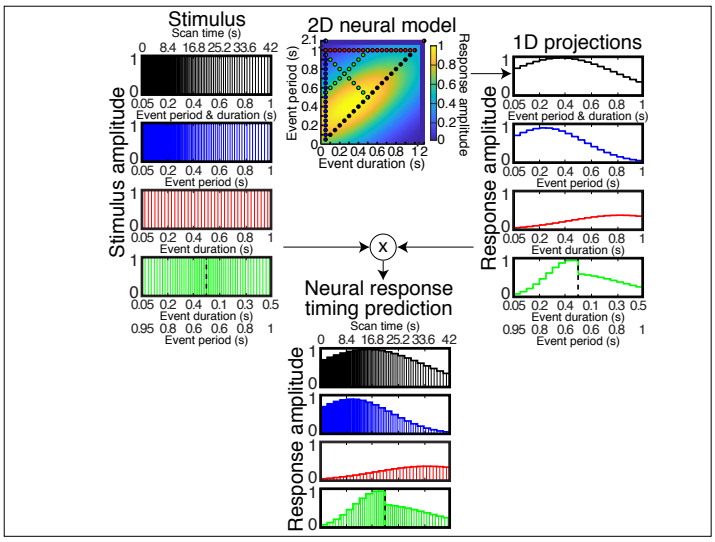
We can see any repetitive event's timing as a point in a 2-dimensional space of event duration vs event period.
Here we see the conditions where we change duration only, period only, or both.
We can then fit a single 2-dimensional Gaussian tuning function to capture the responses to all these timings in each voxel.
To better constrain our tuning function, we include a fourth set of stimulus timings to fill in the middle of this stimulus space.

The nature of this tuning function can tell us a lot.
There may be no relationship between preferred durations and periods.
The preferred duration and may be correlated with the preferred period

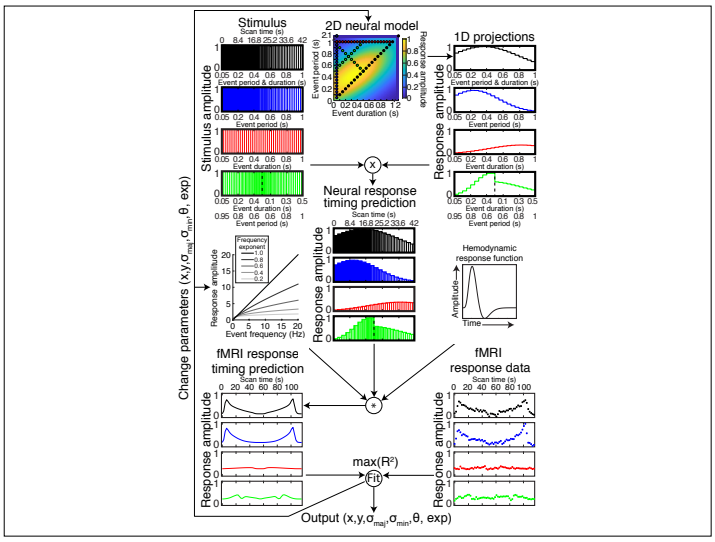
Tuning for visual timing



The tuning function may be elongated on one dimension, perhaps supporting better discriminability of one of these timing dimensions.
Or it may be angulated, such that the preferred event duration depends on event's period and vice versa



We use an extension of the population receptive field modelling approach to fit these tuning functions for event timing.
We take a large set of candidate tuning functions in this duration vs period space.
We quantify each candidate tuning function at each presented duration and period in each of the four presented stimulus progressions.
We determine when each event ended in the stimulus
And use this to predict a neural response amplitude time course that reflects when the event happened, how often an event happened, and its amplitude in this candidate tuning function

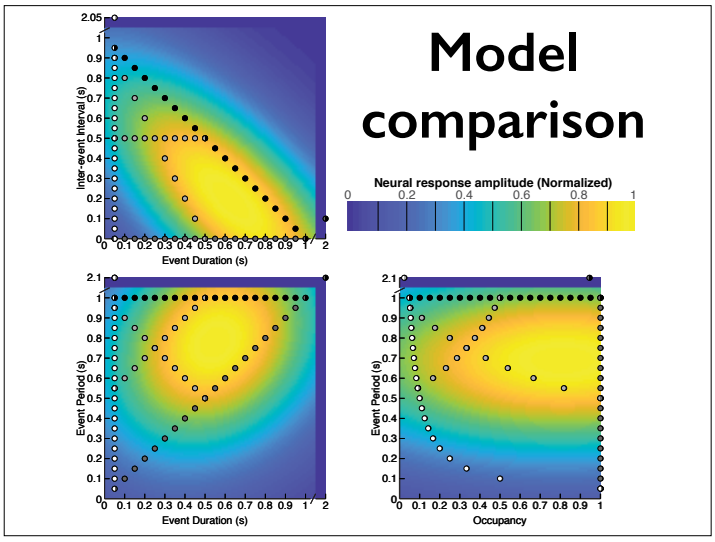


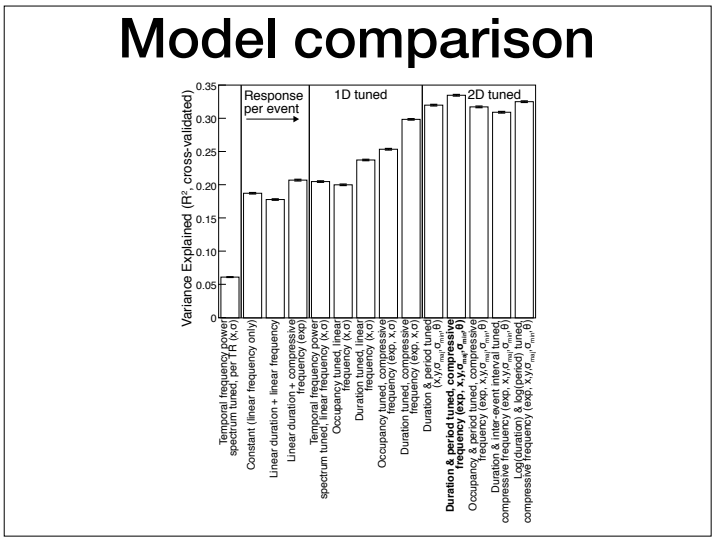
We also fit a factor that captures sub-additive accumulation of response amplitude with increasing frequency.
This allows this type of model to capture responses to events with different frequencies in the same experiment, which regular fMRI analyses don't allow.
The other trick here is that we have short events, but we change the timing of these events only very slowly.
Together, these two tricks allow us to investigate timing with fMRI, despite the poor temporal

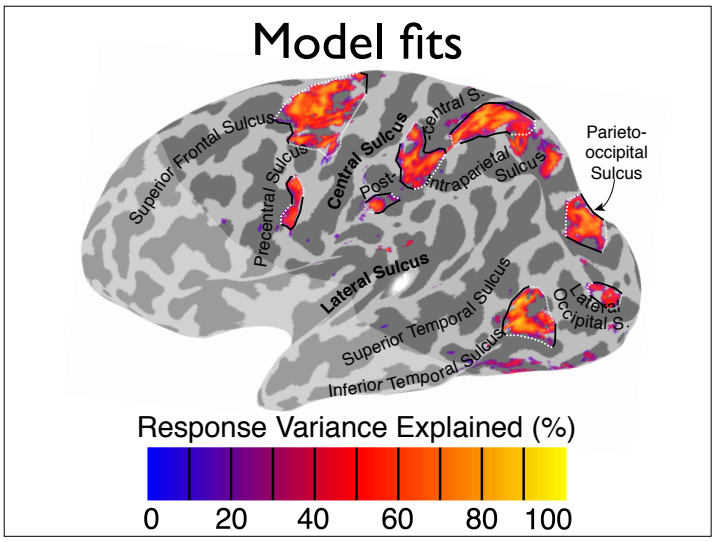
resolution. This makes this whole experiment possible.

We then convolve this neural response prediction with a hemodynamic response function to convert it to a predicted fMRI response time course.
We compare this prediction to the recorded fMRI data and find the tuning function parameters that best predict this data in each voxel

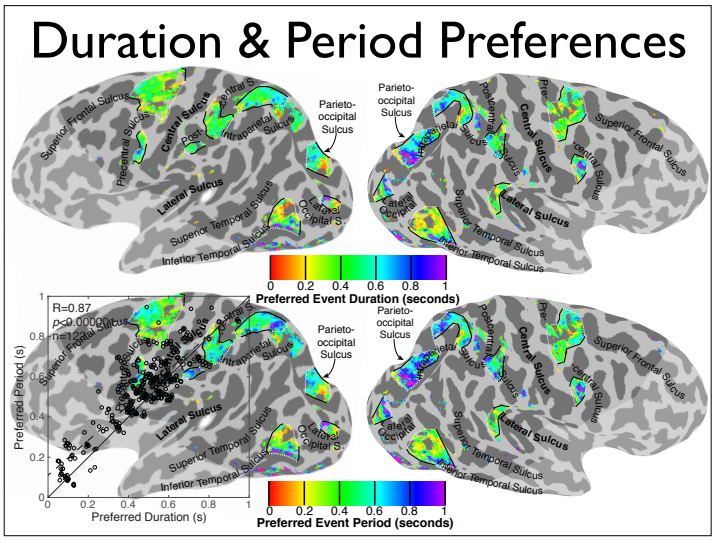
---





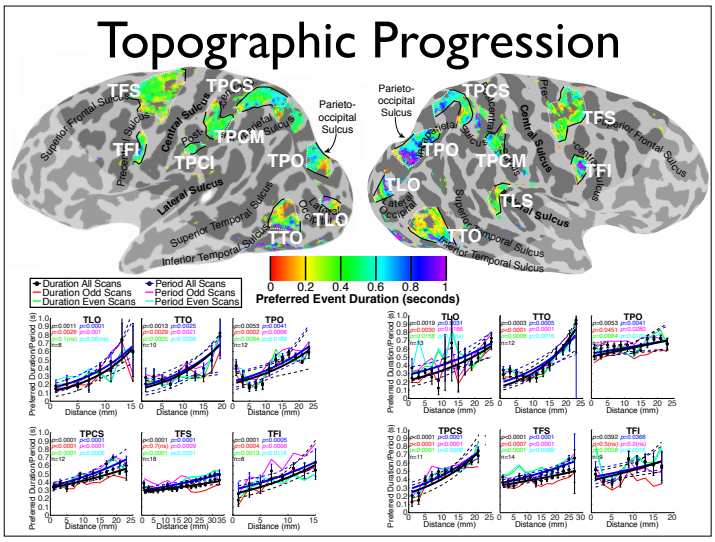


We see several areas where such models capture the responses well.



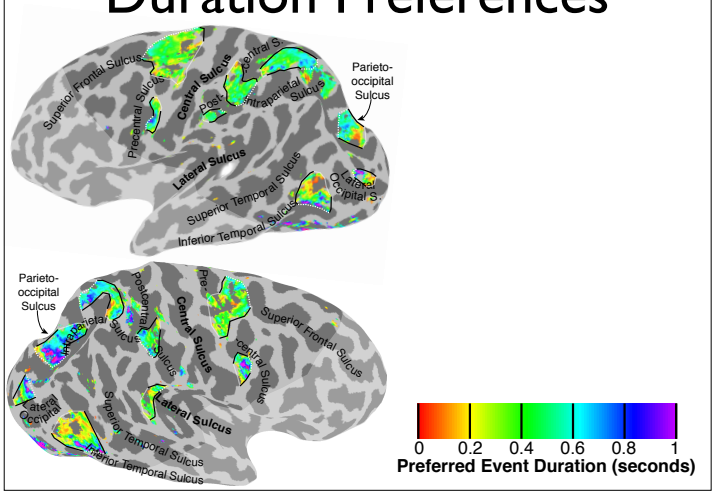
If we project the preferred durations and periods fit by the model on the cortical surface, we see that these areas contain voxels with a range of duration and period preferences.
And we see that within these areas the duration and period preferences are strongly correlated,
They consistently fall just above the unity line, so duration preferences are consistently a little shorter than period preferences

---



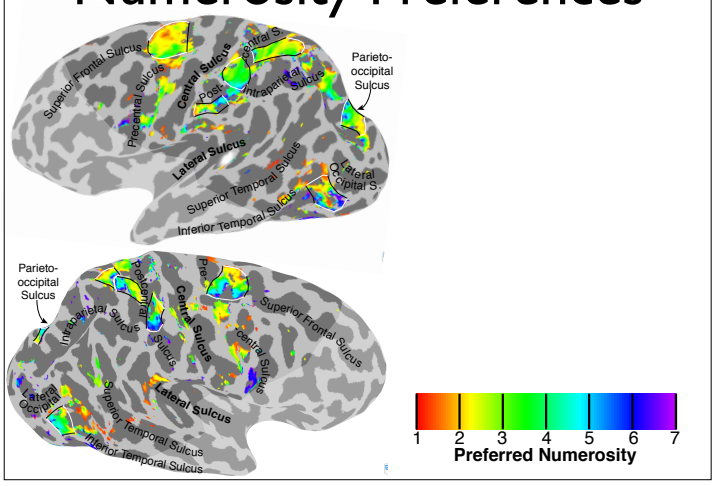
Furthermore, within these areas the preferred duration and period gradually and repeatably change across the cortical surface, forming a series of topographic maps of event timing
We demonstrate the statistical significance of this progression by permutation testing on split halves of the acquired data

Duration Preferences



The areas showing tuned responses to event timing largely overlap with areas showing tuned responses to numerosity

Numerosity Preferences



This approach relies primarily on knowing some properties of the system we are investigating
- Tuned responses to numerosity and motor event timing in monkeys

And using this to find properties that we are interested in
- Does map organisation extend from sensory systems to cognitive systems?

As a result, the machine learning needed is very simple, and comes down to fitting tuning parameters by checking the predictions each candidate parameter set makes

The bigger picture

- "Cells that fire together wire together" - Hebb's postulate
- Neural components should be arranged in a way to make the volume of wire in the brain as small as possible - Minimal wiring hypothesis
  - "Cells the wire together lie together"
- So, local populations should contain neurons that respond similarly
  - And response preferences should change gradually
- Tuned response preferences and their organisation should be accessible at the spatial scale of fMRI
  - Not just for space, number and time

However, this approach is widely applicable to cognitive neuroscience because the brain often shows tuned responses to a particular stimulus parameter.

We are very good at reading faces

We are so good with faces that this 3-second clip gives us a lot of information: age, gender, complex emotional state etc.

We are very good at reading faces

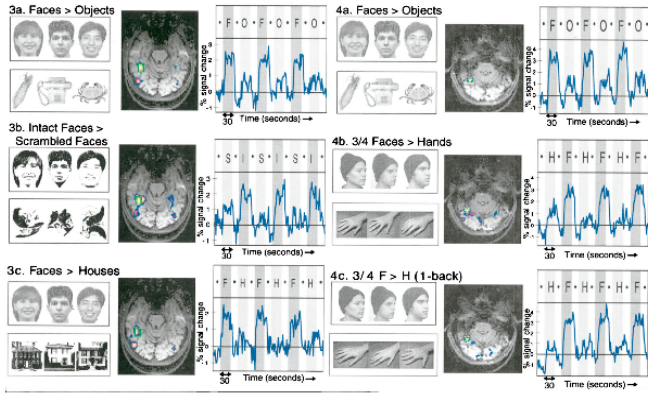
Similarly, in this cat movie, we also perceive the animal's thoughts from its facial expressions

---

Object-selective responses

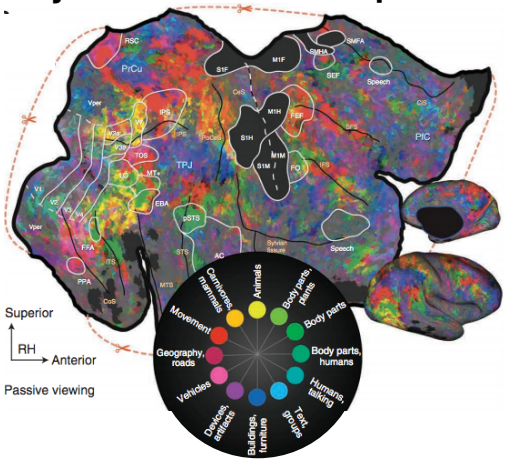
But recognition of faces and other objects has been one of the hardest cognitive processes to understand.
Somewhere along the line, researchers started finding cells that respond to particular classes of object, such as faces

Face-selectivity in fMRI



We also see brain areas that respond when we show faces, but respond less to other object classes.
The cells involved are grouped together in specific areas of the cortex.
Because we can see clear differences between response amplitudes to faces and other objects, it is straightforward to make a classifier than can distinguish these responses.

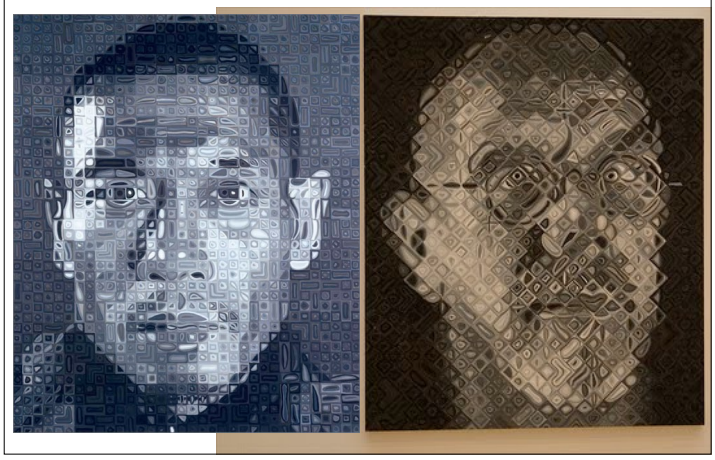
Object selective responses



As we have already seen, there are lot of brain areas responding to various object types, labelled here with different colors.
However, it has been very hard to understand how such responses arise.
Some very exciting progress has been made here recently, by combining advanced machine learning with the known properties of the early visual system.

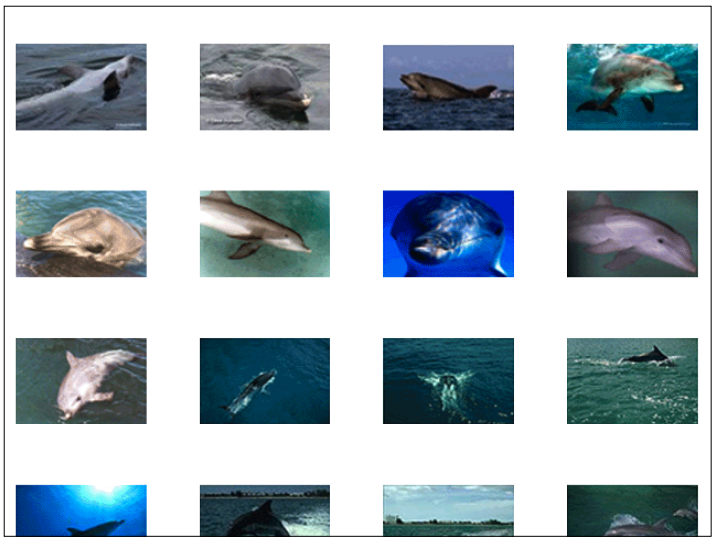
We understand V1's role in form processing as an oriented edge detector.
It has even been very hard to get a useful description of what differs between the properties of V1 and the next processing layer, V2.

Integrating edges to make objects



We understand V1's role in form processing as an oriented edge detector.
It has even been very hard to get a useful description of what differs between the properties of V1 and the next processing layer, V2.

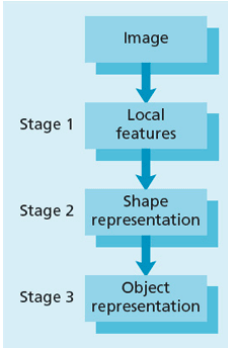
One of the hardest questions for vision science has been how the brain goes from simple, orientation responses to object responses



This is difficult because we can recognise objects regardless of position, size and viewing angle
So there is remarkably little relationship between an object's identity and the image it produces on the retina

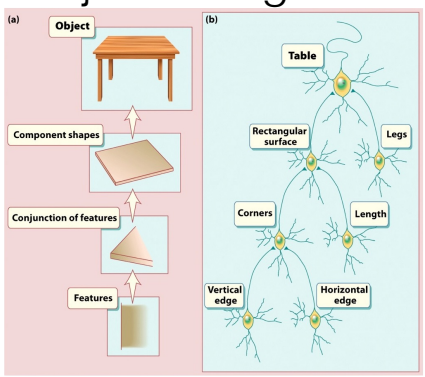
---

The 20th century view of object recognition



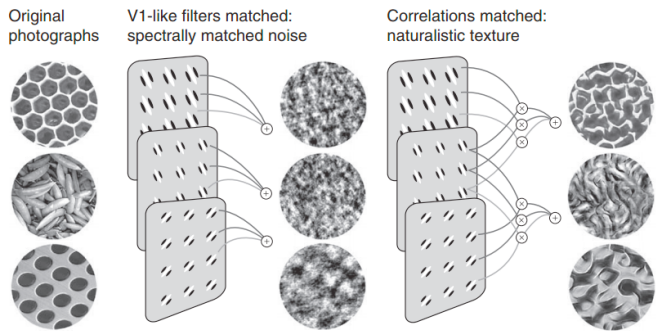
- Stage 1 builds a piecemeal representation of local image properties.
- Stage 2 builds a representation of larger-scale shapes and surfaces.
- Stage 3 matches shapes and surfaces with stored object representations-recognition.

The 20th century view of object recognition



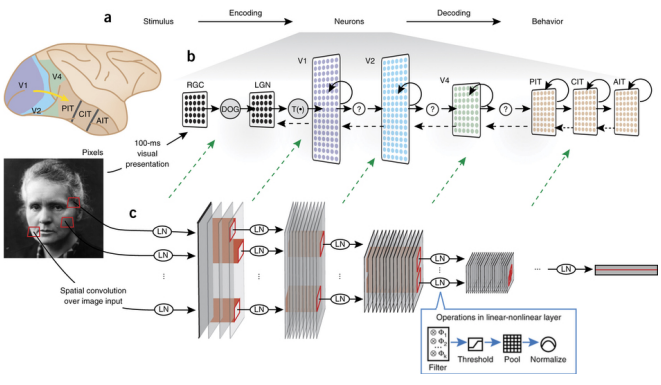
So we might think that our object representations are built from combinations of feature representations
Indeed, objects are built from parts, so simplified parts that we recognise from all angles might let us build a object viewpoint-independent object representation.
Unfortunately, we don't find neurons that look like the middle stages of this network.

What do later visual areas do? (2013 version)



Beyond V1, it has even been hard to figure out what V2 does. Generally, it responds very similarly to V1 when presented with oriented edges.
-Because V1 only responds to the spatial frequencies and orientations in its input, we can make images with the same distribution of spatial frequency and orientation as a natural image, and V1 will respond equally well to both.
-However, V2 responds more strongly to natural images than these noise patterns.
-To make similar responses in V2, the image needs to have a similar pattern of local correlations of orientation as the natural image.
-This appears to be because such patterns of local correlations are common in natural images that have trained the filters linking V1 to V2 i.e. it responds when inputs have the correlations that are common in the real world. By V2, if it fires together, it wires together.
-Orientation-selective responses in V1 were first discovered in 1959. It took until 2013 to reveal how these were transformed by V2. Before 2013, researchers tried to explain V2's responses with reference to features in the input image, rather than patterns in V1's output.
-Beyond V2, we still don't have a good feeling for what drives responses, but we now understand it is likely to be a feature transformation from the outputs of previous network layers.

Deep convolutional neural networks



DCNNs try to imitate this structure over many levels
The neurons in each visual area look very similar in their connections to other neurons.
So DCNNs suppose that, after an initial edge detection, each layer performs similar analyses of the activity in the previous network layer.
Specifically, each layer learns patterns of responses across a small area's representation in the previous layer
So DCNNs work only where there are meaningful relationships

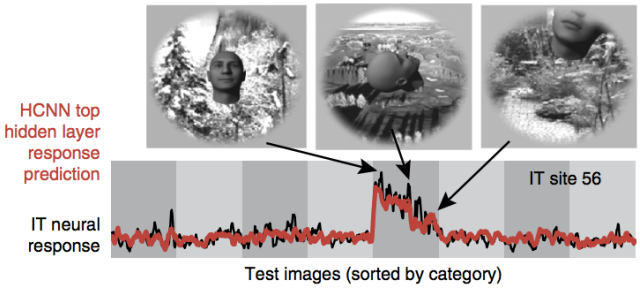
between neighbouring inputs, like neighbouring spatial locations or neighbouring times

Even after two layers, it becomes difficult for humans to understand what a 'neuron' (or unit) will respond to.

In the case of DCNNs, the network follows biological knowledge, but is very complex with lots of weights to learn.
The learning therefore becomes very computationally intensive, but can do impressively human-like things.

A proper understanding of how deep networks work and the biological systems they imitate is beyond this class, but is covered in Machine Learning for Human Vision and Language, my class in period 1.

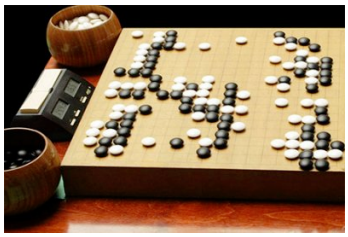
Emergence of object selective responses



Here we see a unit in the top layer of this DCNN after supervised learning of a large set of natural images

The network finds patterns and statistical structure in this natural image training set
The unit here (red line) responds to faces regardless of their position, size, orientation.
The pattern of responses closely follows how face-response neurons in monkeys respond (black line)

Other DCNNs



DCNNs are the state of the art in machine object recognition, so are used by google to label images on the internet. This has obvious applications in effective image search and in targeting advertising.
The game of Go has been exceptionally difficult for computers to solve, as there is an effectively unlimited tree of possible moves. However, humans play much better, and many players describe having an intuition of what strategy will work and what strategy won't.
DCNNs trained on previous games of Go seem to develop a similar understanding, and are now the world's best Go players.

Inverting an object recognition DCNN

http://thepsychreport.com/technology-2/googles-psychedelic-art-this-is-your-computer-brain-on-drugs/

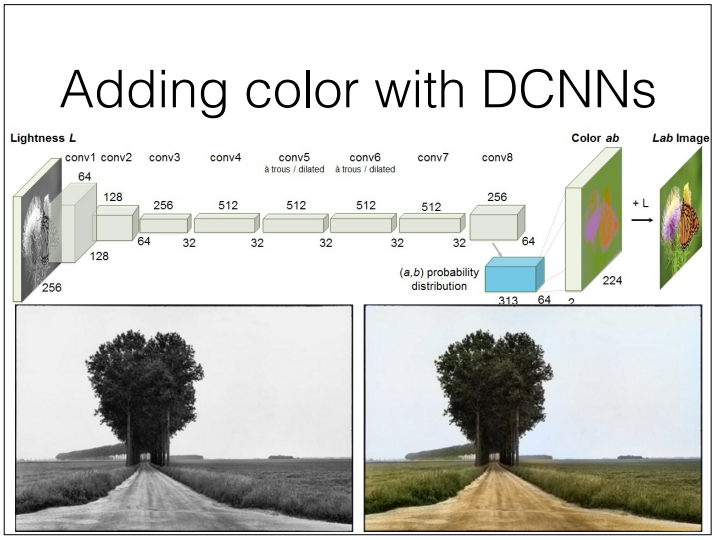
DCNNs can be surprisingly 'creative'.
Here a DCNN was trained on images from the internet to tag objects.
To do this, the DCNN developed increasingly complex representations of objects in the images, over several network layers
The researchers then 'inverted' the network, activating different layers and their connections back to the original points in the image.
This imposed the features that each layer detected onto the original image. Here they use the example image of George Seurat's Sunday on La Grande Jatte
Successive layers find increasingly complex relationships between points in the original image, following correlations and patterns found in natural images

---

Inverting an object recognition DCNN

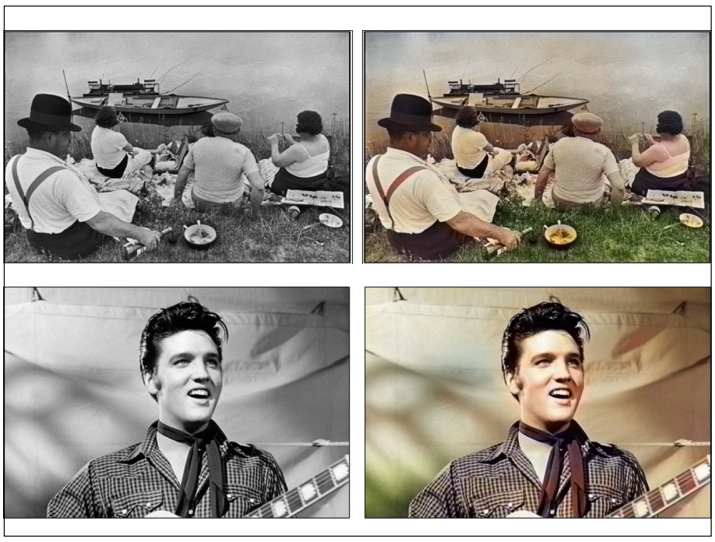
This network has been trained on images from the internet, which contains many pictures of animals
Therefore the higher level units have a lot of animal representations.
So these 'inverted' networks impose the response preferences of higher-level layers on the network's representation of the input image.
Much of how we see the world also follows our previous experience, so we don't have to process each image so extensively: the brain also constrains object recognition problems using established neural network weights.
We currently believe that hallucinogenic drugs act by amplifying these feedback signals, biasing our perception towards previous experience.
Many psychedelic drug users comment that these images look much like hallucinations they have experienced

Here the DCNN was trained on images by a particular artist.
The DCNN determined the spatial correlation patterns in these images, and how these patterns differed from normal photographs.
This style was then applied to new images (the top left photo) by passing the image through the resulting DCNN.



Not all DCNNs have the same structure
Network architecture can be tailored to the task
This network uses a late-stage layer to learn how colours are distributed through images, effectively doing object recognition first (gray boxes) then associating each object with a particular colour.
When passed a new image with no colour, it can determine the likely colours at each pixel

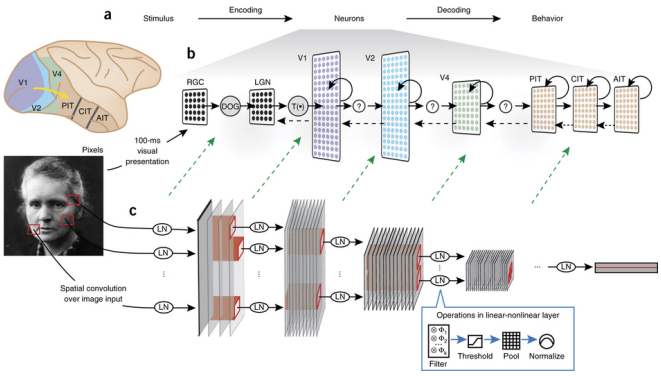
---



Here we can see some more examples of network-colorised images
The results are certainly pretty believable. Grass, trees and people are recognised and coloured correctly
Where the colours are harder to believe, a human could typically not confidently guess what the colour should be.
In a business application of this approach, a deep learning system has recently been applied to colour manga comics automatically, saving a huge amount of boring

work for humans.

Limitations of DCNNs



The human visual system has connections within each layer (i.e. curved arrows, horizontal connections within a visual field map).
Higher/later brain areas also have send information back to earlier areas (dashed black arrows), modifying their responses, while DCNN's don't: most current DCNNs only feed information forward.
Both of these are problems of computational load: it just takes a lot of computation
However, in the last couple of years we have started to see models with these connections included. These begin to do some really impressive, intelligent things, and currently look like the next step in deep network models.

Conclusions

- Including known system properties in machine learning increases explanatory power
  - Model structure known -> Learn model parameters
  - Allows image identification
- Where training set and model structure is simple, machine learning can be reduced to model fitting
  - Simpler, faster, widely applicable within some systems
- DCNNs have complex structure, with multi-level correlations and weights
  - Early stages often simple, biologically-inspired
  - Complex correlations (patterns within patterns) hard for humans to understand
  - BUT can achieve impressively human-like tasks and even appear creative
- DCNN network weights take extensive training and computational power
  - Currently limits DCNNs to feedforward architecture, though we know the brain also has feedback and horizontal (recurrent) connections
  - DCNNs beginning to implement recurrent structures, giving even more abilities