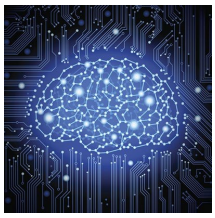
Machine learning: principles & applications

Dr. Ben Harvey
Experimental Psychology

b.m.harvey@uu.nl

Why study machine learning?

- A very powerful data analysis method
  - Gives methods to interpret complex data
  - Increasingly widely used in scientific research
- A broadly applicable method to study data
  - Large data sets are increasing common due to advances in IT ('big data')
  - Many applications to business and technology



Tech applications

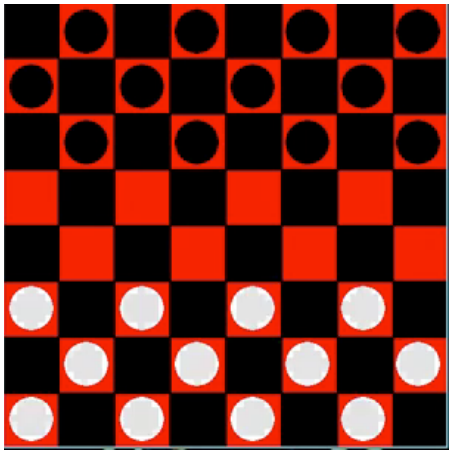
NETFLIX

amazon

What is machine learning?

- The field of science that 'gives computers the ability to learn without being explicitly programmed' (Arthur Samuel, 1959)
-

So this field has been around for a long time
Samuel showed a machine learning algorithm many games of checkers (draughts) to learn which board positions are likely to lead to a win. The only programmed rule was the way the pieces are allowed to move.



Here we see the black player (a machine learning program) consistently beating the white player (which moves randomly)

What is machine learning?

- The field of science that 'gives computers the ability to learn without being explicitly programmed' (Arthur Samuel, 1959)
- 'A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.' (Tom Mitchell, 1998)
  - Expressed in organisational terms, not cognitive terms

But Samuel's definition of machine learning uses the word 'learn' without any definitions
It also uses 'computer', which is a synonym of 'machine'
Mitchell provided a more formal definition
By describing a fundamental operation in terms of inputs and outputs, this definition avoids suggesting the machine can think like a human.
Alan Turing: "Can machines think?"->"Can machines do what we (as thinking entities) can do?"

Learning to filter spam

- 'A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.' (Tom Mitchell, 1998)
- **Task:** 'Classify which emails are wanted (not spam) vs unwanted (spam)'
- **Experience:** Watching humans labels emails (training set)
- **Performance:** The proportion of new emails (test set) classified correctly

By the end of these classes

- You should be able to:
  - Describe general principles of machine learning
  - Recognise and classify different types of machine learning approach
  - Describe the principles behind some example machine learning approaches, their advantages and disadvantages
  - Describe some example applications of machine learning to fMRI data

Supervised and unsupervised machine learning

- Supervised learning: feed in example inputs and desired output (training set).
  - Goal is a rule that maps **new** inputs (test set) to correct outputs
  - Can have discrete class outputs (classification task, like a spam filter) or continuous outputs (regression task, examples later)
- Unsupervised learning: No desired outputs are given.
  - Goal is to find structure in the inputs (like clustering data into groups)
- Reinforcement learning (intermediate): Given output of a complex task like driving (did I crash?) or game play (did I win?).
  - Not given information about how close process is to the goal

Supervised learning is impossible where we don't know (or have) a desired output (training data)

Relationship to AI & stats

- Machine learning began as a field of AI, but became unpopular by the 1980's
  - Focus on expert systems & agents
- Returned in the 1990's to solve practical problems away from AI
  - Using methods and models from statistics and probability theory
  - Now back at the cutting edge of AI
- Much in common with data mining and KDD (knowledge discovery and data mining)
  - Machine learning generally tries to reproduce known knowledge or to known outcomes (supervised learning possible)
  - Data mining/KDD aims to discover unknown knowledge (unsupervised learning only)

Model assessment

- In supervised learning, test the proportion of test set that is classified correctly
  - 'Correct' answer must be know by experimenter
  - Test set and training set must be independent (cross-validation)
- Split or holdout cross-validation method
  - Typically 2/3 training and 1/3 test
  - Best to re-train the model with different splits to demonstrate
- Leave-one-out cross-validation
  - Train model on N-1 example test set
  - Test on example N (test set)
  - Repeat with (all) different examples left out and used as test
- Correct and incorrect proportions (accuracy), sensitivity & specificity (formal discriminability)

Unsupervised learning can't be tested the same way, because the 'correct' answer is unknown

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)

Because vision is the best-understood neural system, most of the pioneering work here is done in visual processing, just like machine vision is far better developed than computer sound or touch analysis or machine cognition

Relationship to statistics

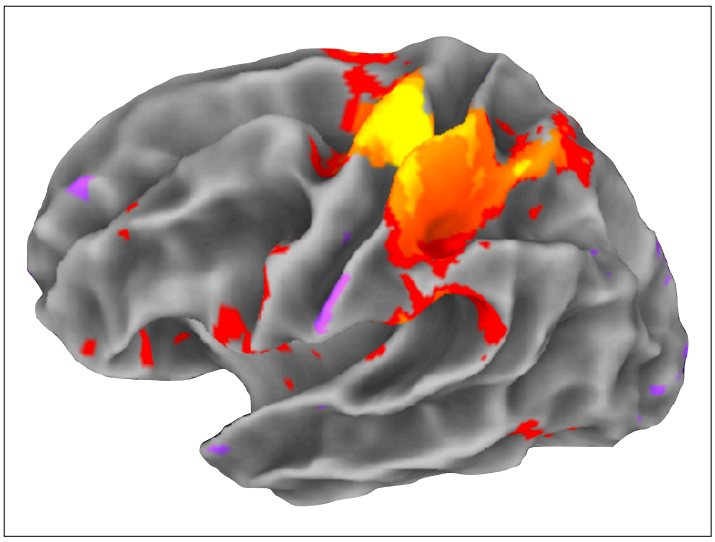
- Machine learning can capture very complex relationships between input states and outputs
- But it is also VERY sensitive to simpler relationships
- One input variable can be enough to classify output above chance
  - A T-test does this, no machine learning needed
  - Machine learning (for classification) tells us there is some relationship, a T-test tells us what relationship
- One input variable's magnitude can be sufficient to predict the magnitude of an output variable
  - Correlations do this (for regression), and also tell us what the relationship is. Preferable to machine learning.
- Two input variable states can be enough to predict/classify output
  - 2-way ANOVA with interaction terms (classification), bivariate regression (regression) or ANCOVA (combined classification & regression)
- Any small number of variables can be enough
  - N-way ANOVA, multivariate regression, complex Bayesian modelling and other advanced statistics. All reveal what the relationship is.
- Machine learning can do all of these, but often acts as a 'black box'

So machine learning can be used as 'lazy statistics'.
It rarely identifies what the relationship is between inputs and outputs.
But it provides a very flexible single method to reveal some relationship.
This can be very useful for engineering, where we don't need to know why something works.
But it is very limiting for scientific research, where we want to understand the system we are studying
Statistics are still important to determine if the classification performance you see is significant.
This can get pretty complex, so rather than avoiding statistics, good machine learning relies on very advanced statistics

So, now let's look at some examples of machine learning in action to see how these advantages and disadvantages work in practice.

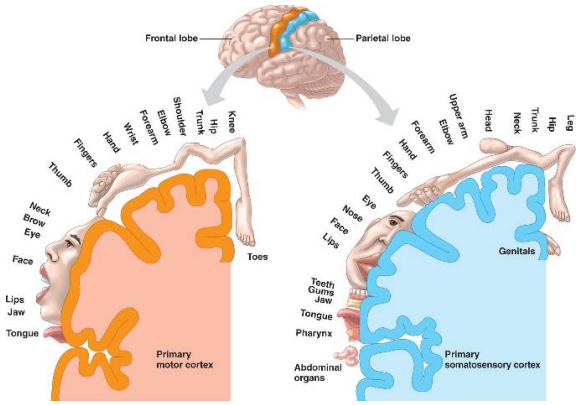
The story so far…

- Machine learning is a widely applicable method to:
  - Demonstrate a relationship between two groups of data (supervised, e.g. map inputs and expected outputs)
  - OR find patterns in complex data sets (unsupervised)
- Should consider what we already know about the system
  - Do we already know something that can inform our model?
  - How much unknown do we need to capture?
- With simpler data sets or fewer unknowns, traditional statistics is often a better choice
  - If there are a few unknowns to capture, stats will show how these are related to the results
  - ML typically just shows SOME relationship exists



Many of the examples I show here come from functional MRI of the human brain, my own field.
Machine learning and modelling are well suited to fMRI data, because it takes very large numbers of measurements (millions) at a time, and repeats these at many points in time.
This is pretty big data: an hour of MRI data collection is a couple of gigabytes.
You may think of fMRI as showing which brain area increases its activity when we perform a particular task or view a particular stimulus.
Here we see the areas that are active when we move the right hand.
Specifically, we see the T-statistic magnitude as different colours. These are areas that give a significantly different response (output) when we change between moving and not moving the hand (the input to this analysis).
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.
Here I will describe fMRI experiments that don't do this, but instead aim to discover patterns in the data or to reveal mechanisms of human neural processing.
These are two very different aims, and choosing which to pursue depends on what we already know about the response we want to characterise.

Map organisation sensory & motor cortices



If we look more closely in this area, we can see that its detailed structure is far more complex than a big blob lighting up.
We are seeing the activity of the hand area in the motor cortex, driving the movement, and the hand area of the somatosensory cortex responding to the resulting sensations.
Both of these areas are organised as maps: the body's surface is spatially mapped onto the brain's surface.
So it seems likely we can reveal more than just which blob lights up.

Computational neuroimaging

"Prior to fMRI, neuroimaging primarily tested hypotheses about the localization of function by asking whether two stimuli (or tasks) caused statistically different signals within the brain. The SNR of fMRI is large enough to measure the size [and other characteristics] of differences, not just their presence or absence. For this reason, hypotheses about neural computations, beyond localization, can be framed and tested."

Brian Wandell (1999)

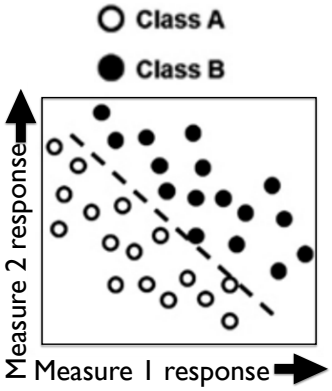
Serge Dumoulin (2008)

17

In the Netherlands, this idea was pioneered by Brian's student, Serge Dumoulin
Because vision is the best-understood neural system, most of the pioneering work here is done in visual processing, just like machine vision is far better developed than computer sound or touch analysis or machine cognition.
So many of the applications we see here focus on vision.

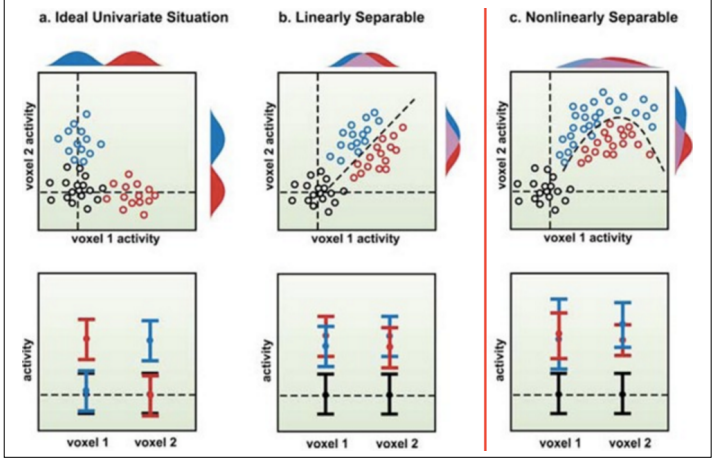
Support vector machines (SVM)

- Very common supervised approach for classification & regression
- Given training set of inputs and classifications, builds a model that correctly classifies new inputs into these classes
- Can classify based on which side of a decision boundary (line, plane or hyperplane) the new input falls
- Widely used in fMRI analysis (multi-voxel pattern analysis, MVPA)



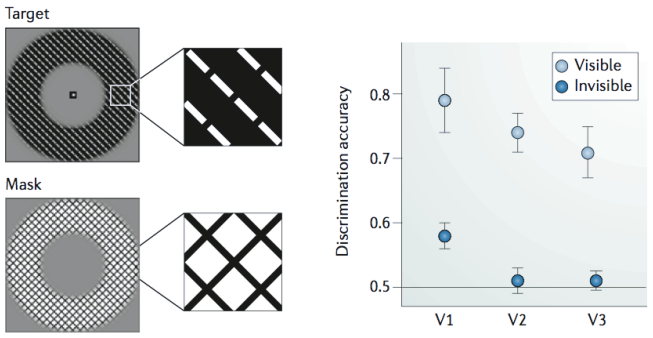
When both responses change in the same way, their average response gives the same information as the pattern: one class gives a high response and one gives a low response
In MRI, this is common, because an area may generally increase or decrease its responses.
In this case, statistics is required, but no learning, because the PATTERN has no extra information

Support vector machines (SVM)



Non-linear decision boundaries are rarely used. They tend to over-fit data and capture non-repeatable (noise) patterns.

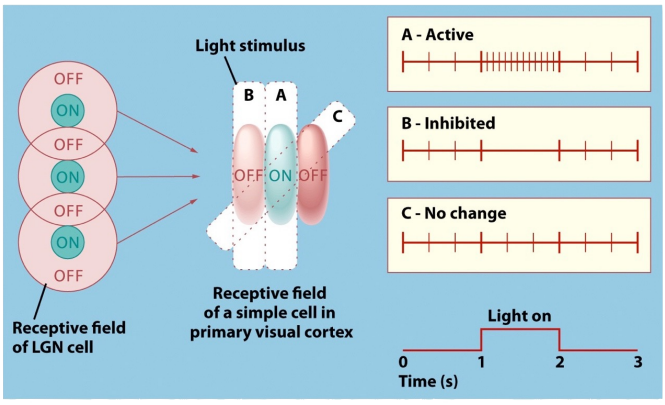
Classifiers



Kamitani and Tong (2005), Haynes and Rees (2005; 2006)

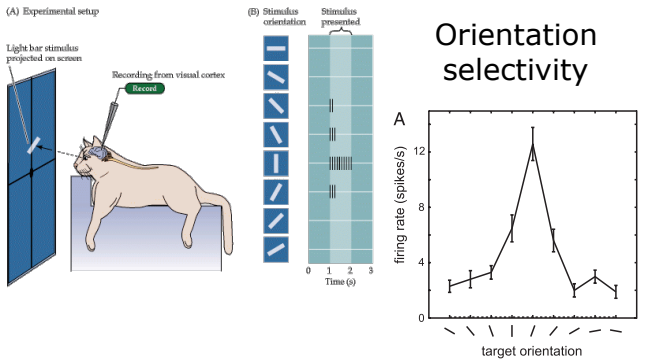
So let's look at a specific example of SVM use: decoding image orientation
Using relatively large (3mm) voxels, we can decode which orientation was shown i.e. we can learn the relationship between voxel activity pattern and orientation, then determine which was orientation was shown by looking at the voxel activity.
This also works for imagined orientation: we can determine which orientation a subject imagined from their brain activity
This certainly tells us that V1's responses have some information about orientation.
The problem here is that we don't know what information the classifier used: it distinguished orientations based on V1's entire activity.
In fact, we already knew that V1's responses have some information about orientation.

Neural responses to edge orientation



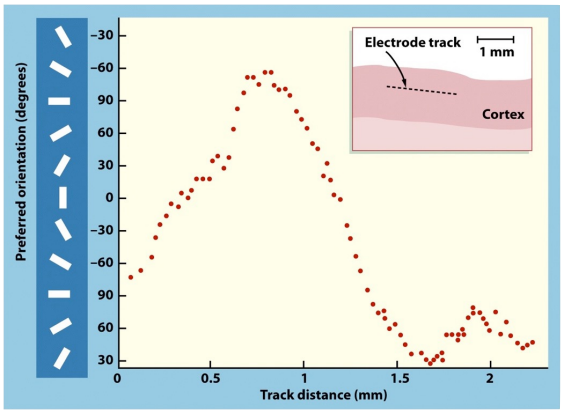
The neurons here have an excitatory centre region and suppressive surround, which detects change in the image, or edges
By integrating the responses on 2 or 3 such cells, the visual system determines the orientation (direction) of the edge
Early visual cortex contains cells that respond only when edges have a particular orientation

Neural responses to edge orientation



Orientation selectivity

Orientation selective neurons

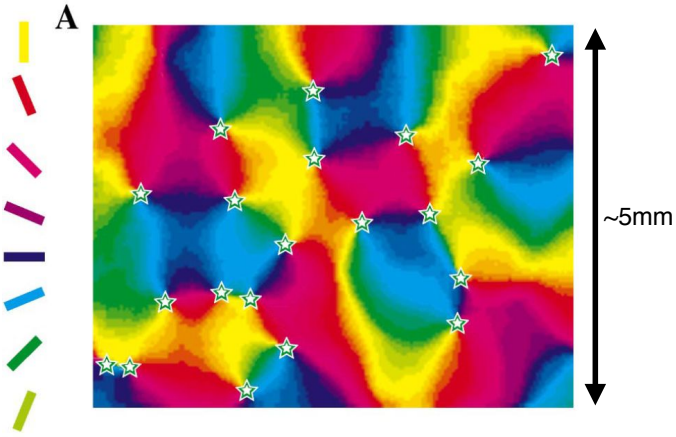


23

As we move along the cortical surface, preferred orientations gradually change.
Notice the scale here: a complete range of orientations passes in around 2mm

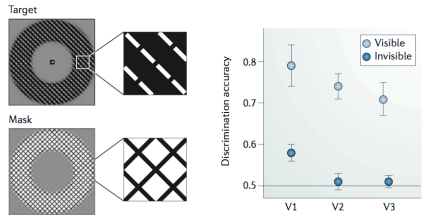
---

Orientation 'pinwheels'

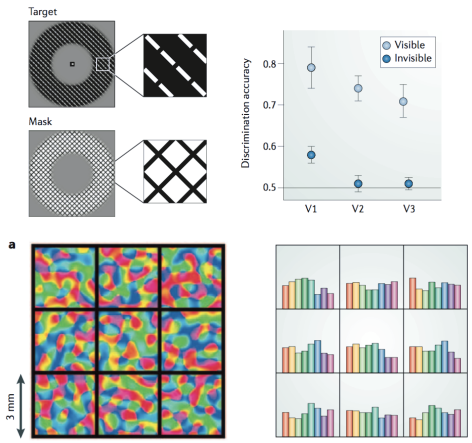


24

If we look at a small piece of cortex, we see gradual changes in orientation preferences



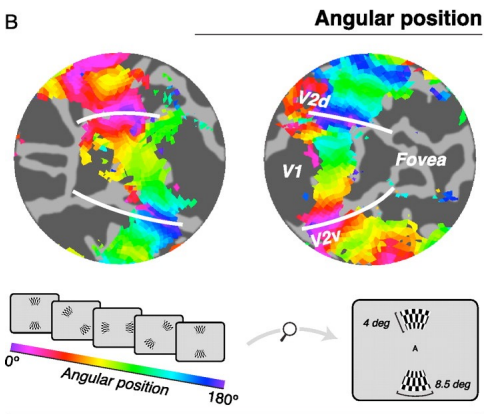
Kamitani and Tong (2005), Haynes and Rees (2005; 2006)



Kamitani and Tong (2005), Haynes and Rees (2005; 2006)

The initial explanation proposed for this ability to decode orientation at a coarse spatial scale was that each voxel contained a slightly different balance of orientation preferences. This would lead to slight differences in the fine-scale voxel activation

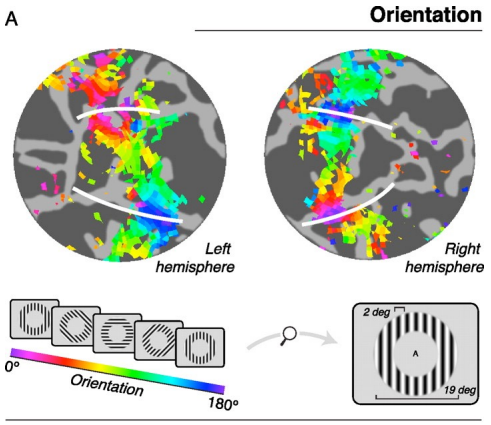
Orientation bias



27

But we also know some more about this system, which questions this interpretation.
Neurons in the human visual cortex each respond to stimuli at a specific place in the visual field (retina)
These 'preferred positions' change gradually across the brain's surface, forming maps of the visual field on the brain's surface.
Because nearby neurons all respond to similar positions, a larger MRI recording site (voxel) also has a preferred position to which it responds

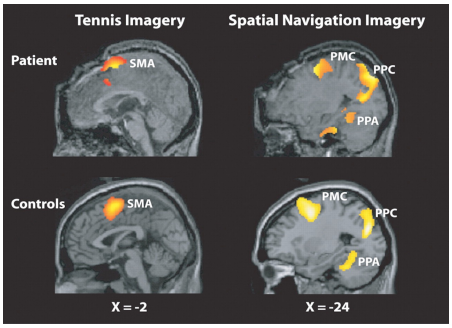
Orientation bias



28

But if we change orientation while stimulating all positions, we see very similar large-scale topographic maps of orientation
Each voxel has more orientation preferences pointing towards the central visual field
So this large-scale bias, just like the fine column-level bias, would allow orientations decoding
Ten years later, scientists still argue about the origin of this decoding performance
Using more computational methods, it is becoming clear that both of these sources of information contribute.
Also, if a human observer can tell the difference between two stimuli, we can be certain that there is some location in their brain that holds some information about stimulus state.
So is it really important to show that the brain contains some information about orientation?

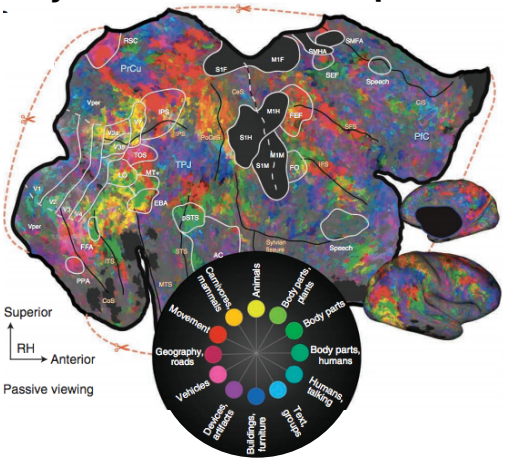
Applications: Vegetative state (coma)



Owen et al, Science, 2006

Regardless of the source of the information, MVPA is widely applicable
Here, we see brain activity when a subject and a patient are asked to imagine playing tennis, compared to when imagining exploring a city.
These make clear activations in specific places, though the researchers used MVPA to classify the patterns as it allows very fast classification of each trial.
But the patient group here was in a prolonged vegetative state, a coma in which they were assumed to be unaware of their surroundings.
This shows that they could understand questions and give a binary response through their brain activity.

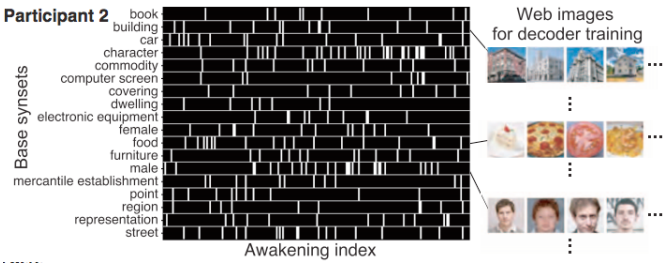
Object selective responses



Beyond the early visual cortex, we find many brain areas that respond to presentation of particular types of object.
Here, experimenters have tagged the object preferences of fMRI recording sites throughout the brain by correlating their responses with the content of labelled movies.
We see a lot of brain areas responding to various object types, labelled with different colors.
This is not really a machine learning approach, but some advanced statistics to efficiently map out which object types are correlated with responses in each area.

Classifying object categories

- If we show awake humans in fMRI scanners categorized visual images, we can associate a pattern of brain activity with each category

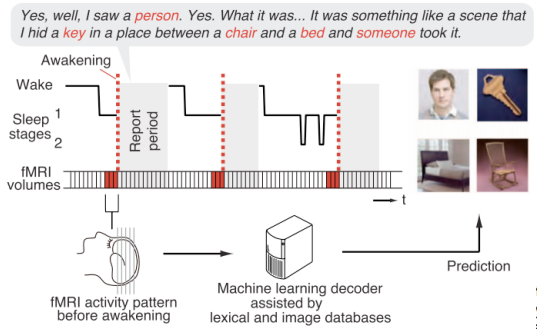


31

But if we have an idea that doesn't need to know which area responds to each specific object type, MVPA decoders are an efficient way to summarise the brain's responses.

Classifying dream categories

- We can then wake them up when dreaming and ask what they were dreaming about

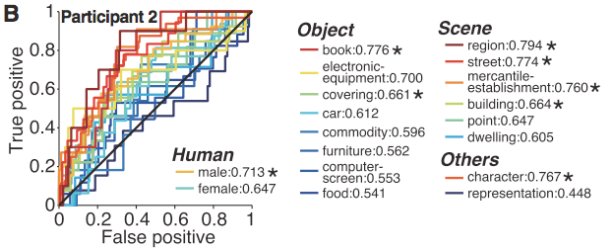


32

32

Classifying dream categories

- And we can predict, better than by chance, what they will say by examining their brain activity while dreaming



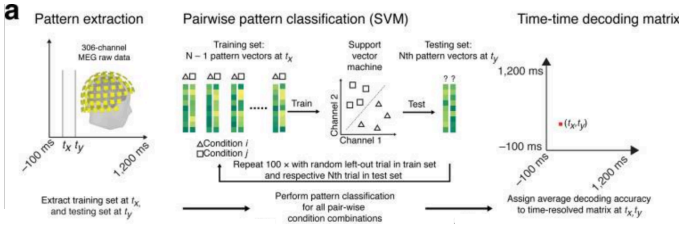
- However, this approach looks at object selective areas, and cannot reconstruct the images or determine what the dream might have looked like

33

MVPA classification with SVMs

- Widely used, simple method to find patterns in large, complex fMRI data sets
- After learning patterns associated with different stimuli, algorithm can predict which (unknown) stimulus likely produced a response
- Demonstrates that the brain are contains 'information' about this stimulus type
  - Even when there is no change in average response amplitude
  - (also sensitive to average amplitude changes, regular stats)
- However, no information about WHAT information is present (i.e. what was used to achieve the decoding)
  - Therefore prone to false positive results and unclear interpretation

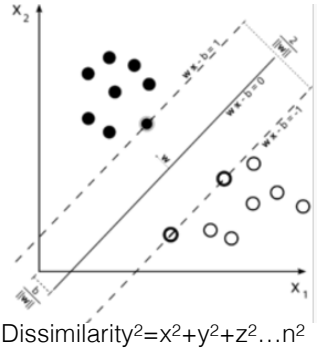
MVPA cross-classification



This is unsupervised

Representational similarity analysis (RSA)

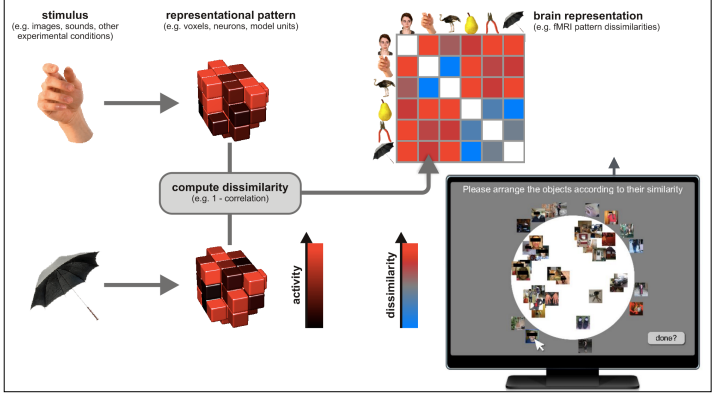
- An unsupervised regression approach with many similarities to SVM classification
  - SVM determines which side of a decision boundary each input falls
  - RSA determines how far each stimulus lies from each other (no boundary, decision)
- 'Distance' here is distance in a representational space
  - FMRI voxel activations
  - EEG electrode responses
  - Individual neuron firing rates
  - Human performance or subjective similarity ratings
  - Computational model units
  - Combinations or comparisons of these
- Distances are easy to compute: Pythagoras theorem
- or (more simply) dissimilarity = 1 - correlation



Dissimilarity^2=x^2+y^2+z^2…n^2

Your lab assignment is based on this technique, and implements the approaches on the following slides.

Representational similarity analysis (RSA)

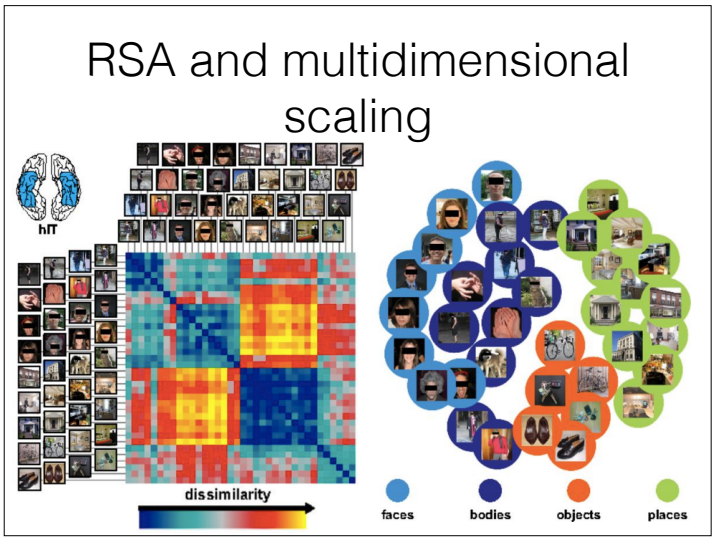


Here we see a little group of fMRI voxels that changes its pattern of responses when shown different types of objects.
We can summarise how different, or dissimilar, these response pattern are as one minus the correlation coefficient.
This can be repeated for a large set of object pairs to give a representational dissimilarity matrix, summarising all differences between response patterns for this stimulus set.

One great thing about RSA is we can examine and compare patterns of responses measured by many different methods.
If we also ask the same subjects to arrange objects by their perceived similarity, for example.
If the resulting representational dissimilarity matrices for a brain area's activity and perceptual judgements are similar, this suggests that the brain area's activity may underlie the perception of similarity between these objects.

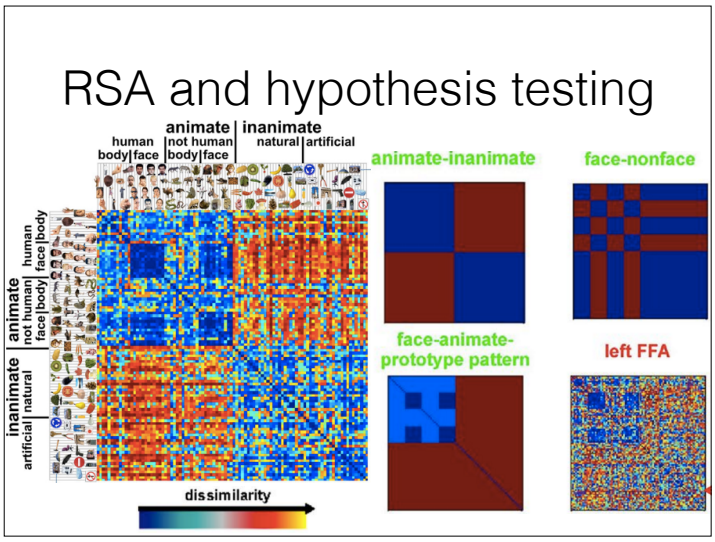
Another important property of RSA is that we can also express hypotheses about what produces a

pattern of responses using a representational dissimilarity matrix.
For example, we can predict what we would see if responses depended on the similarity of image appearance at the pixel level. We can test whether this matches the measured representational dissimilarity matrix.
Or we can test the hypothesis that this area simply acts as a face detector.



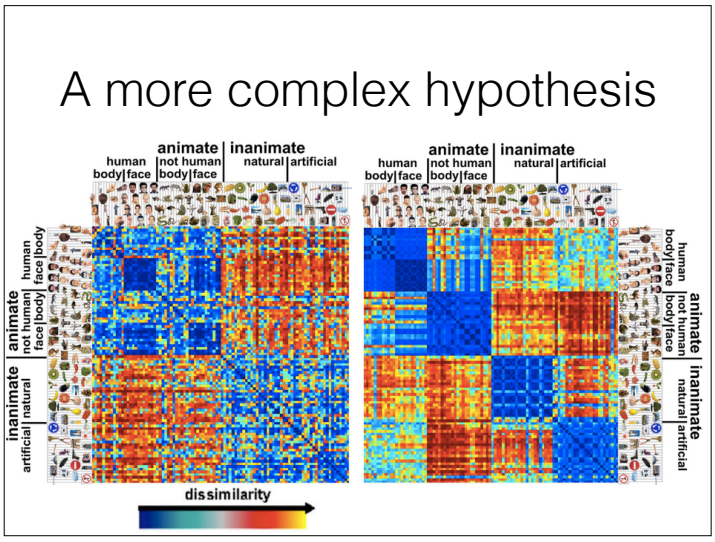
RSA dissimilarity matrices can be hard to read
Along the diagonal, we see the high similarity between repeated responses to the same image (blues)
We see (perhaps) four clusters of relatively similar images
We can use a complex analysis, multidimensional scaling, to project these dissimilarity 'distances' to two perpendicular dimensions
Here, it is easier to see what has clustered together and which clusters are more similar to others
For example, faces and body parts are represented more similarly

---



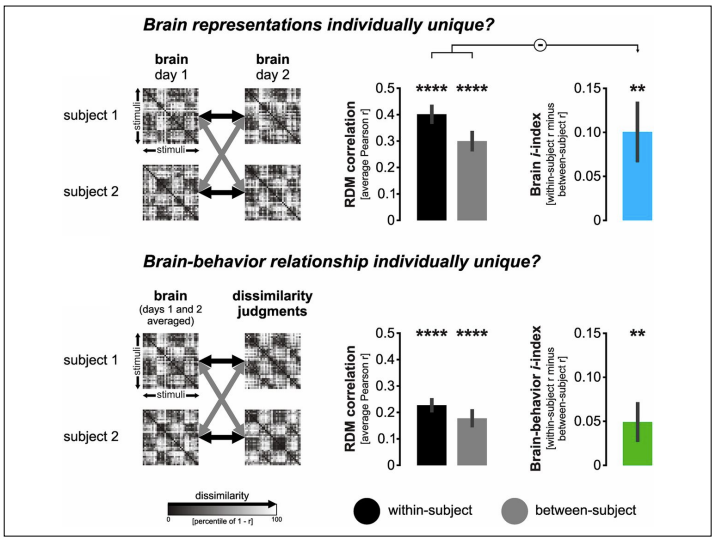
To test a hypothesis, we can first compare the dissimilarity matrix predicted by our hypothesis against our measured matrix
Here we see different ways to split the images shown
We can hypothesise that all animate objects are represented similarly, and all inanimate objects are represented similarly
Or we can hypothesise that all faces are represented similarly, and all non-faces are represented similarly.
Importantly in these dissimilarity matrices, we are not characterising

whether an object is animate or is a face.
We are instead characterising whether two objects have the SAME STATE of animacy or faceness. Two animate objects are similar, two inanimate objects are ALSO similar, one animate object and one inanimate object are dissimilar.



Here we hypothesise that these differences in neural representation underlie differences in perceived similarity
So we correlate behavioural judgements of similarity to the similarity pattern of neural representations
This will show when (or where in the brain) neural representations are consistent with behavioural patterns

---



Here we see how this can be used to show that similarity patterns are more closely related when they are measured twice from the same brain than when measured from different brains (top).
This shows that there is pattern of similarities that is specific to each person.
We also see that similarity patterns from fMRI and behavioural measures are more closely related if they come from the same person rather than different people.
This shows that these brain similarity patterns predict each person's perception.

Why RSA?

- Reveals inherent structure in a complex, high-dimensional data
  - i.e. data with many independent measurements
  - Determines what is important in determining patterns of response
  - Such data is common in neuroscience
- Easy to compare neural/behavioural representations across very different measurement methods

By the end of these classes

- You should be able to:
  - Describe general principles of machine learning
  - Recognise and classify different types of machine learning approach
  - Describe the principles behind some example machine learning approaches, their advantages and disadvantages
  - Describe some example applications of machine learning to fMRI data
  - Choose which machine learning approaches to apply to some example problems, which inputs to use, and which results to expect
  - Implement and apply a representational similarity analysis (RSA) to different types of data
  - Use simple statistics to test data results against different models of information representation/processing in the brain

The remaining learning goals of these classes will be the focus of your lab assignment on RSA