Cours 2022-2023:

Quel code neural pour les représentations mentales?
Vector codes and the geometry of mental representations

Stanislas Dehaene
Chaire de Psychologie Cognitive Expérimentale

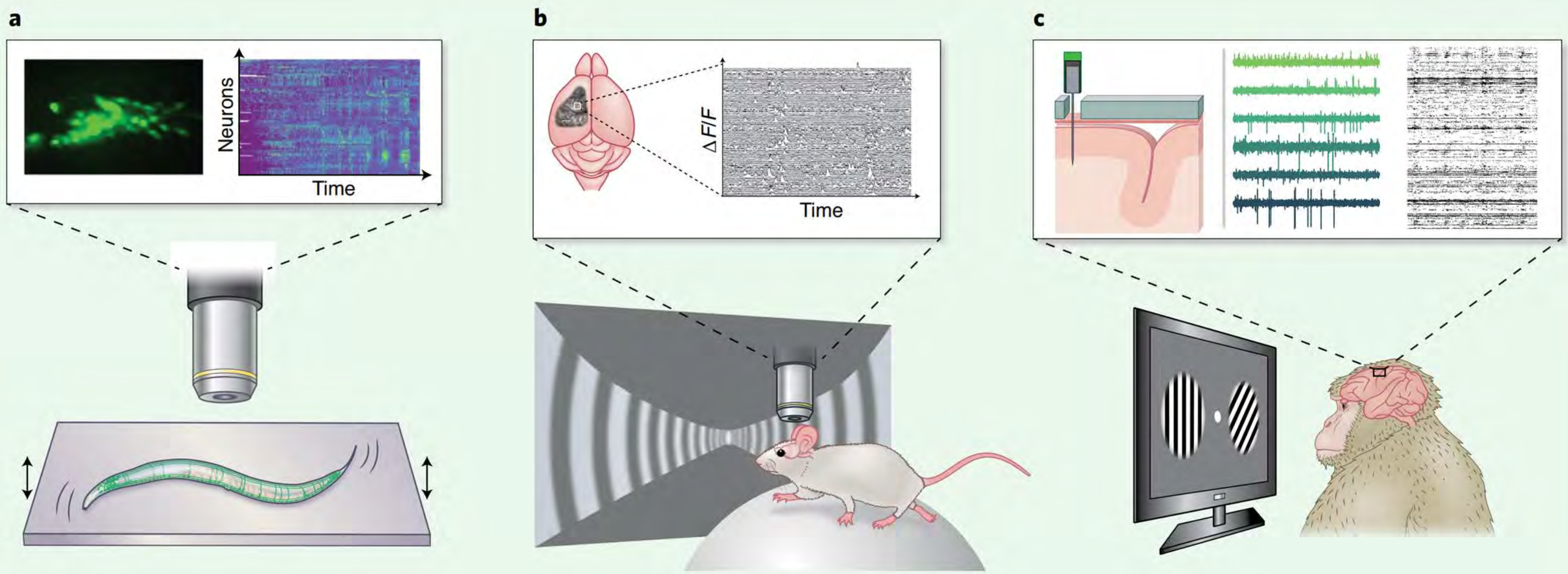
Cours n°2

Géométrie des représentations visuelles: chaque visage est un vecteur

Geometry of visual representations : every face is a vector

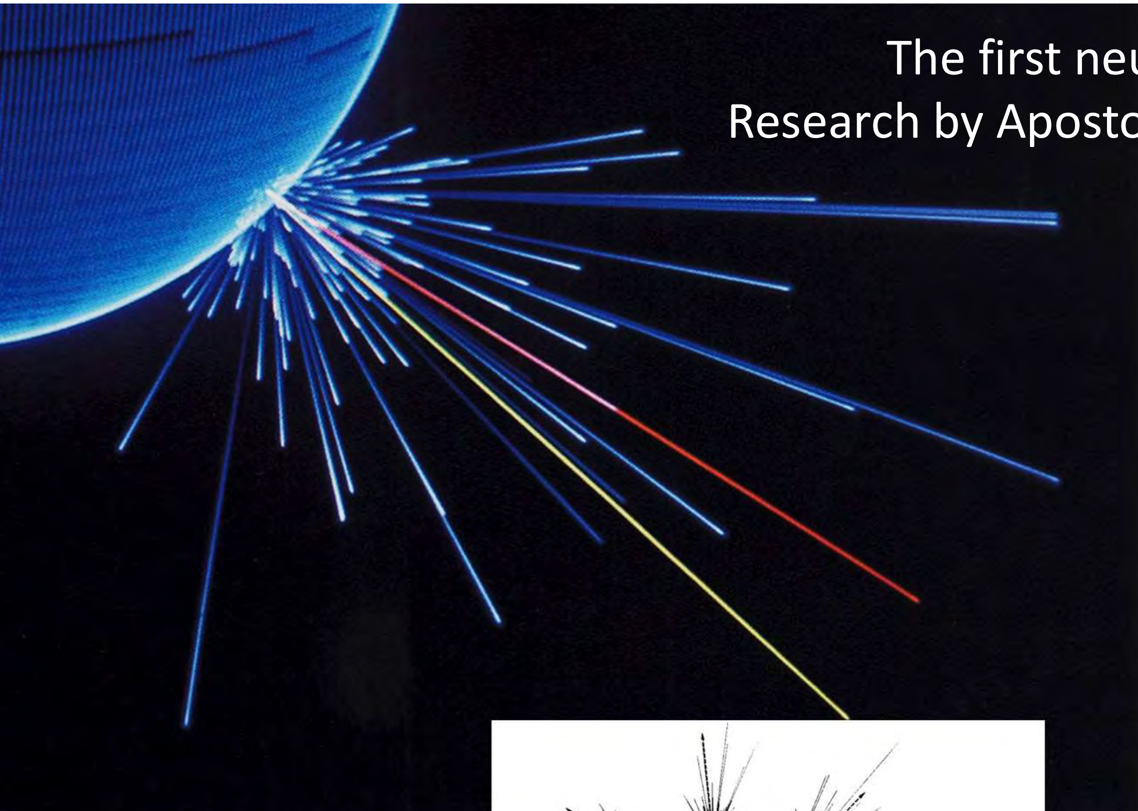
Large-scale neurophysiology is compatible with awake, behaving animals.

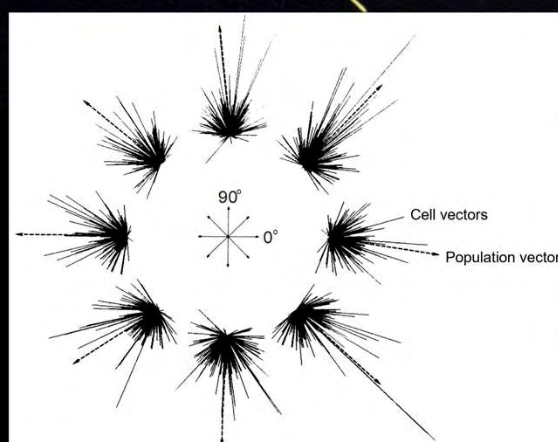
Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, 1-9. https://doi.org/10.1038/s41593-021-00980-9

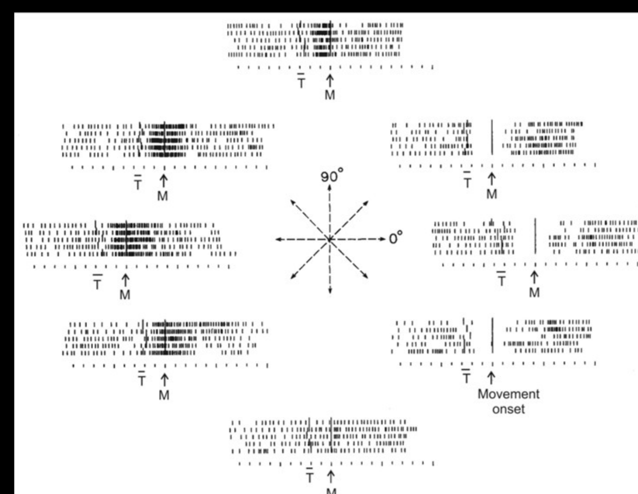


Large-scale neural recordings in behaving animals. **a**, Confocal microscopy of all neurons in *C. elegans* as it freely moves on an adjustable platform. Adapted from ref. [123] under a CC-BY license. **b**, Mesoscope two-photon imaging of the cortical surface while a mouse moves through a virtual reality by running on a ball. Adapted from ref. [124] under a CC-BY license. **c**, High-density electrophysiological recordings using Neuropixels probes, while a monkey performs a psychophysical decision-making task.

The first neuronal vector code:
Research by Apostolos Georgopoulos (1980's)







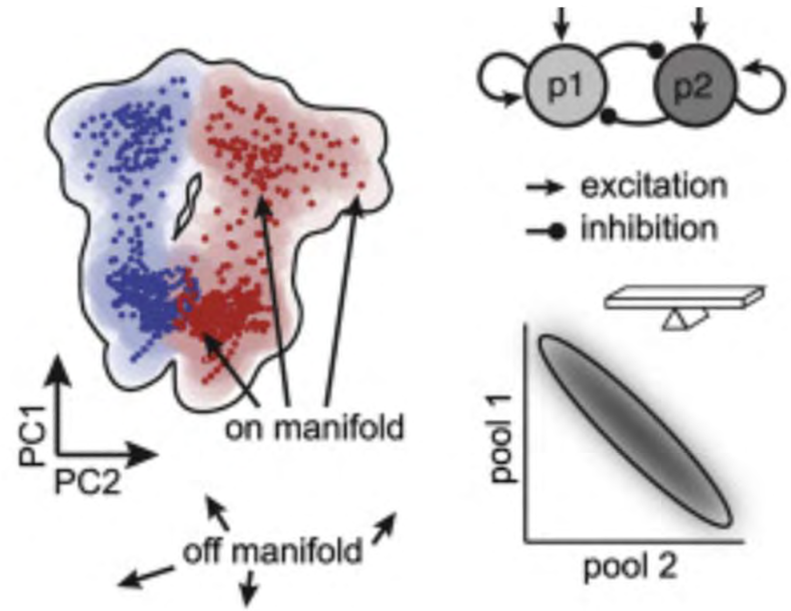
The concept of « neural manifold »

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*. https://doi.org/10.1016/j.neuron.2021.07.011

"Because activity of neurons tends to be correlated with each other, because the wiring between neurons constrains what patterns of neural activity are possible, neural states often only vary along a small number of dimensions in the neural subspace.

To put it another way, there is a lot of white space in our state space diagrams: neural activity tends to occupy fewer neural states than it would if each neuron made an independent, random contribution to population activity.

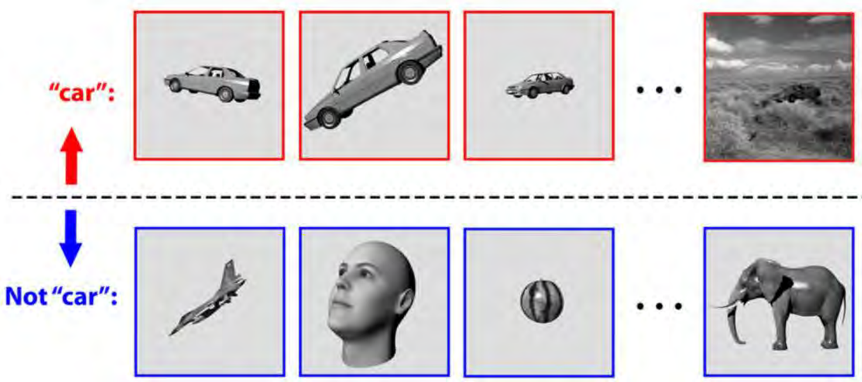
The part of the neural state space that contains the states that we observe is called the neural manifold"



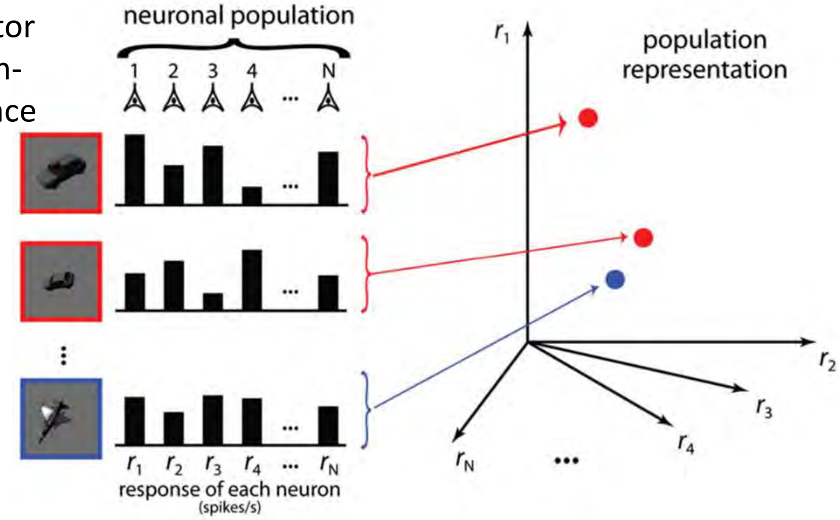
The successive stages of processing in visual cortex may « untangle » a manifold

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. Trends in Cognitive Sciences, 11(8), 333-341.
DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? Neuron, 73(3), 415-434.

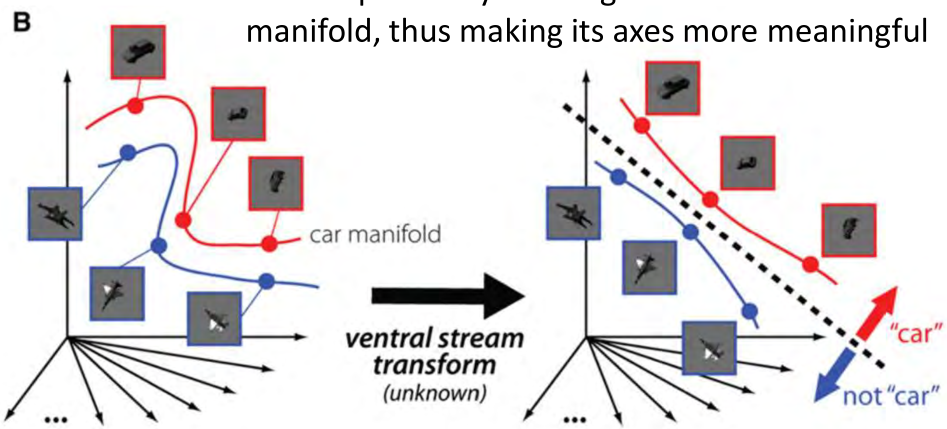
The core problem of **invariant** object recognition

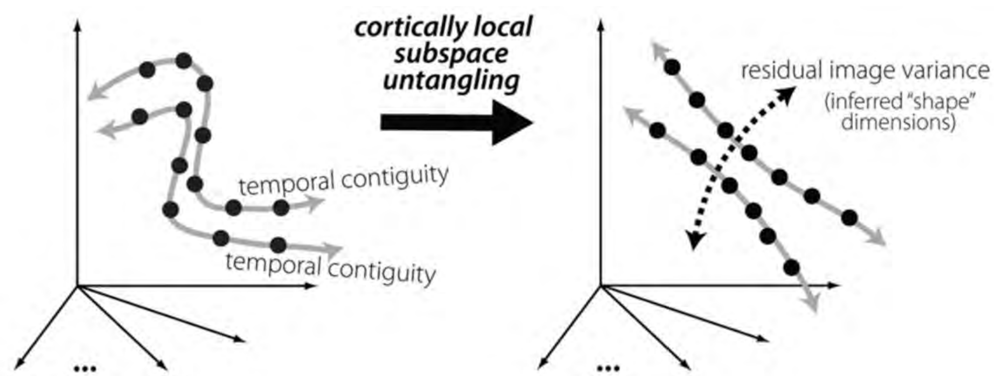


Principle of vector coding in a high-dimensional space



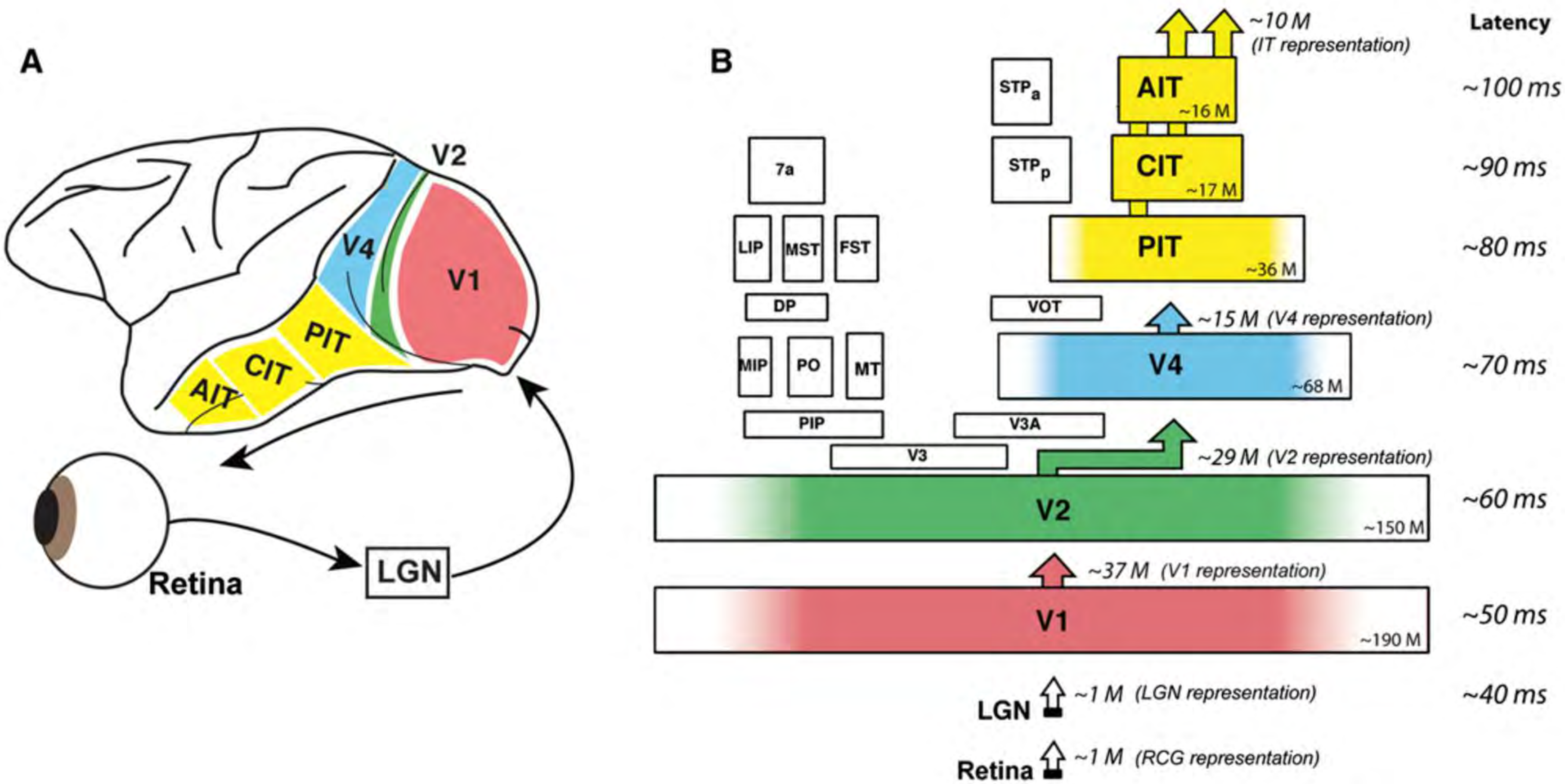
The visual space may "untangle" or "unfold" the neural manifold, thus making its axes more meaningful





The successive stages of processing in visual cortex may « untangle » a manifold

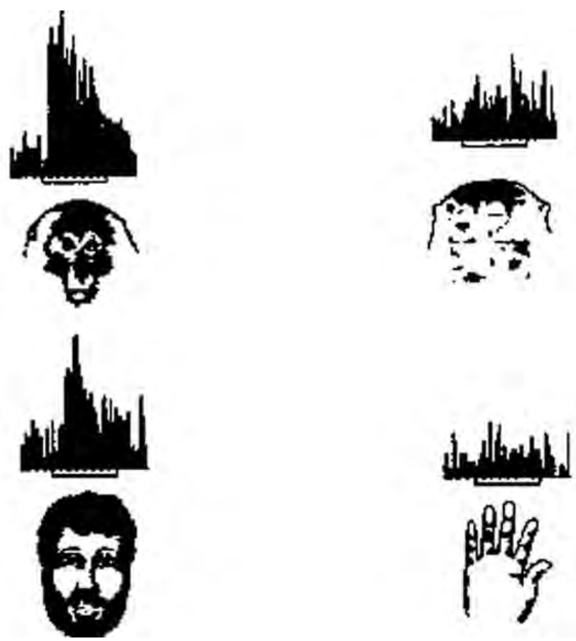
DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. Trends in Cognitive Sciences, 11(8), 333-341.
DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? Neuron, 73(3), 415-434.



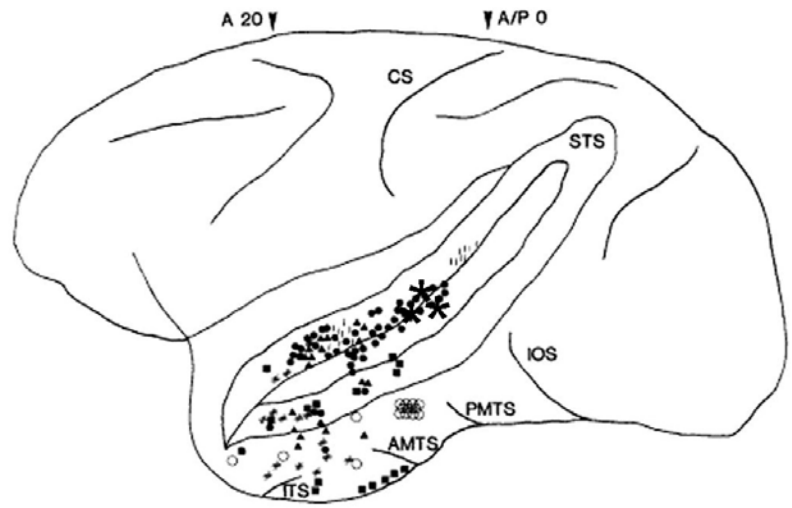
Question for today: Does infero-temporal cortex contain a disentangled vector space for visual recognition?

With many thanks to Professor Doris Tsao (UC Berkeley)

Face coding in the ventral visual pathway



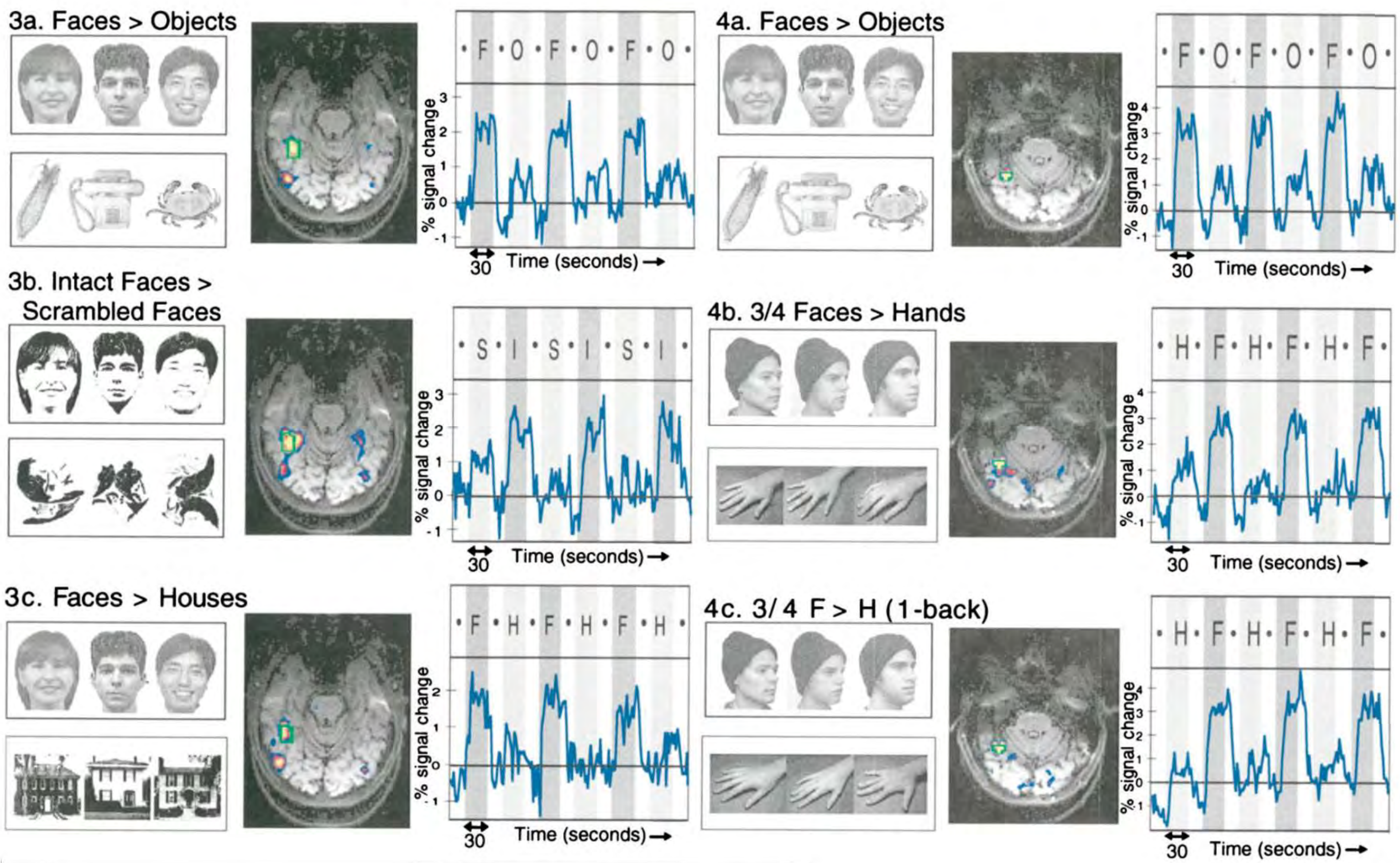
Bruce et al., 1981

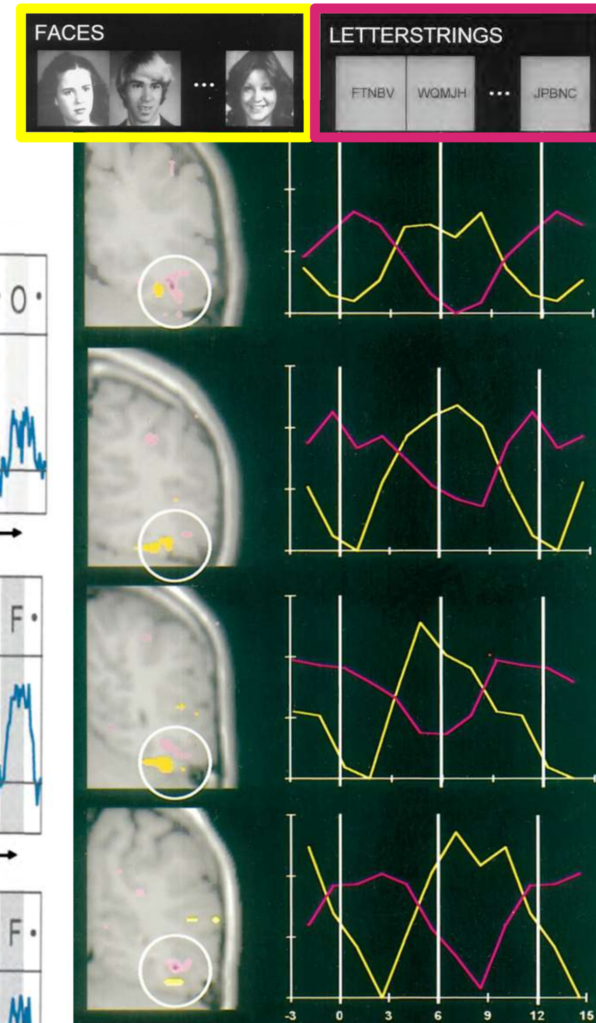


- ● Perrett (1982, 1985, 1987)
- ▲ Rolls (1984)
- ■ Yamane (1988)
- * Desimone (1984)
- # Hasselmo (1989)
- | Harries (1991)
- o Tanaka (1991)

The discovery of the human fusiform face area

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area : A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302-4311.

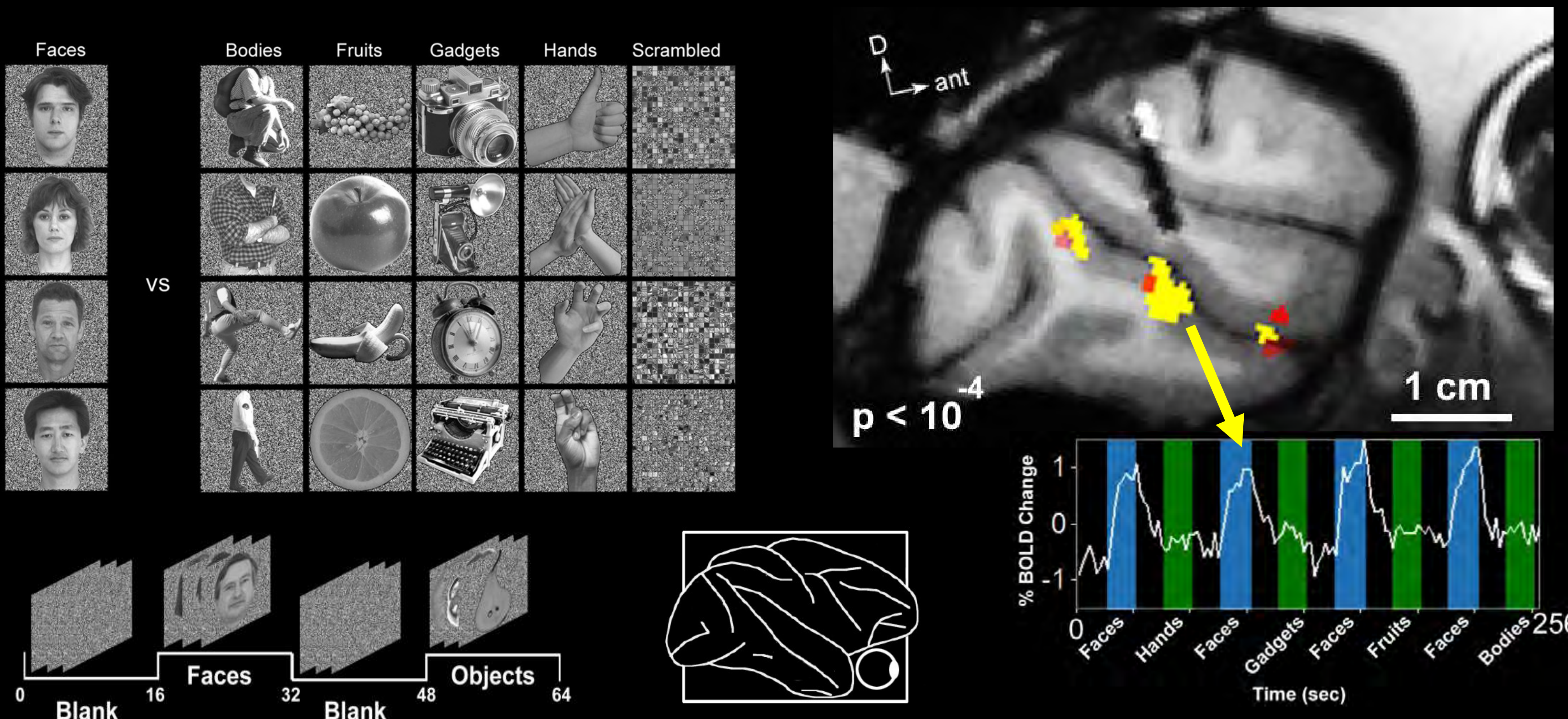




Puce, A., Allison, T., Asgari, M., Gore, J. C., & McCarthy, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings, and textures : A functional magnetic resonance imaging study. *Journal of Neuroscience*, *16*, 5205-5215.

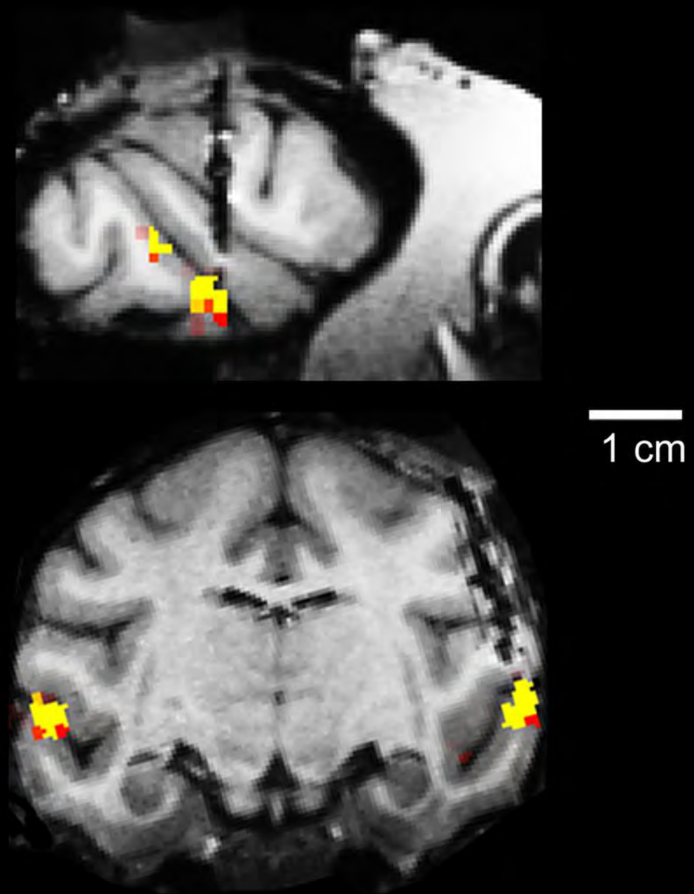
The discovery of monkey face patches

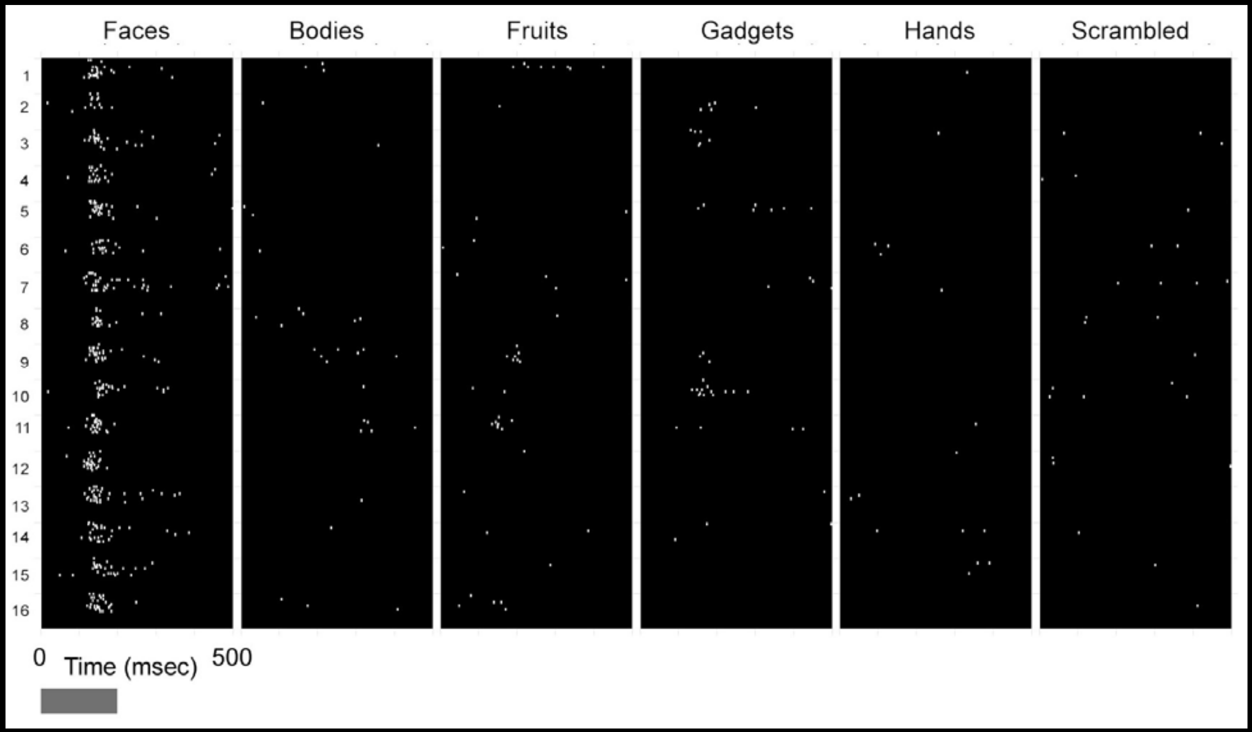
Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. Nat Neurosci, 6(9), 989-995. https://doi.org/10.1038/nn1111



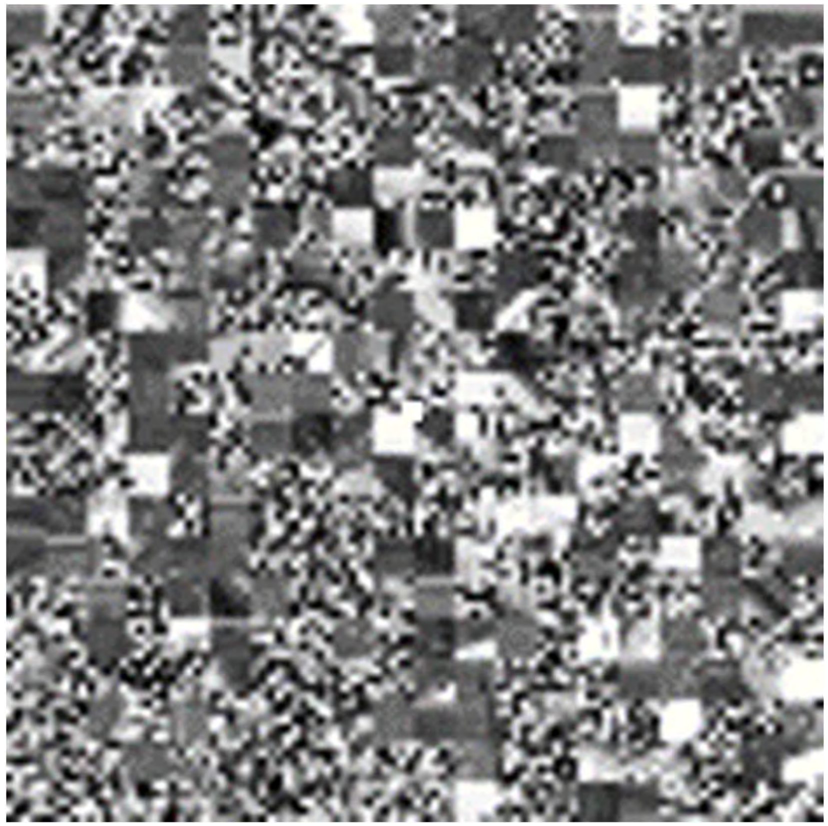
Face patches contain a majority of neurons exquisitely tuned to faces

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. Science, 311(5761), 670-674.





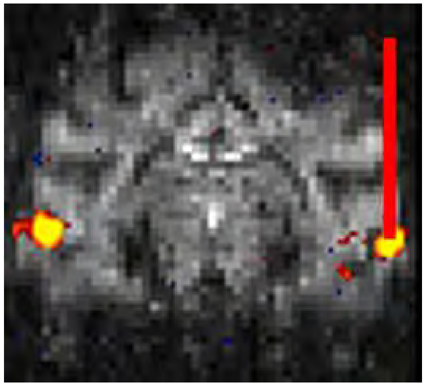
Example of a single unit recorded from ML

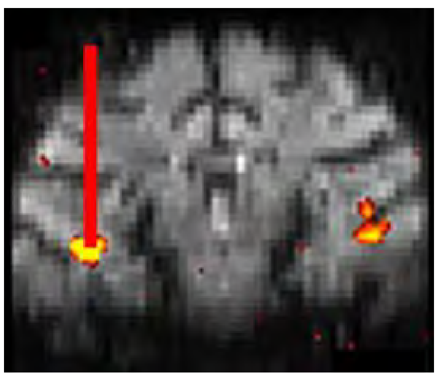


With special thanks to Professor Doris Tsao

Face patches contain a majority of neurons exquisitely tuned to faces

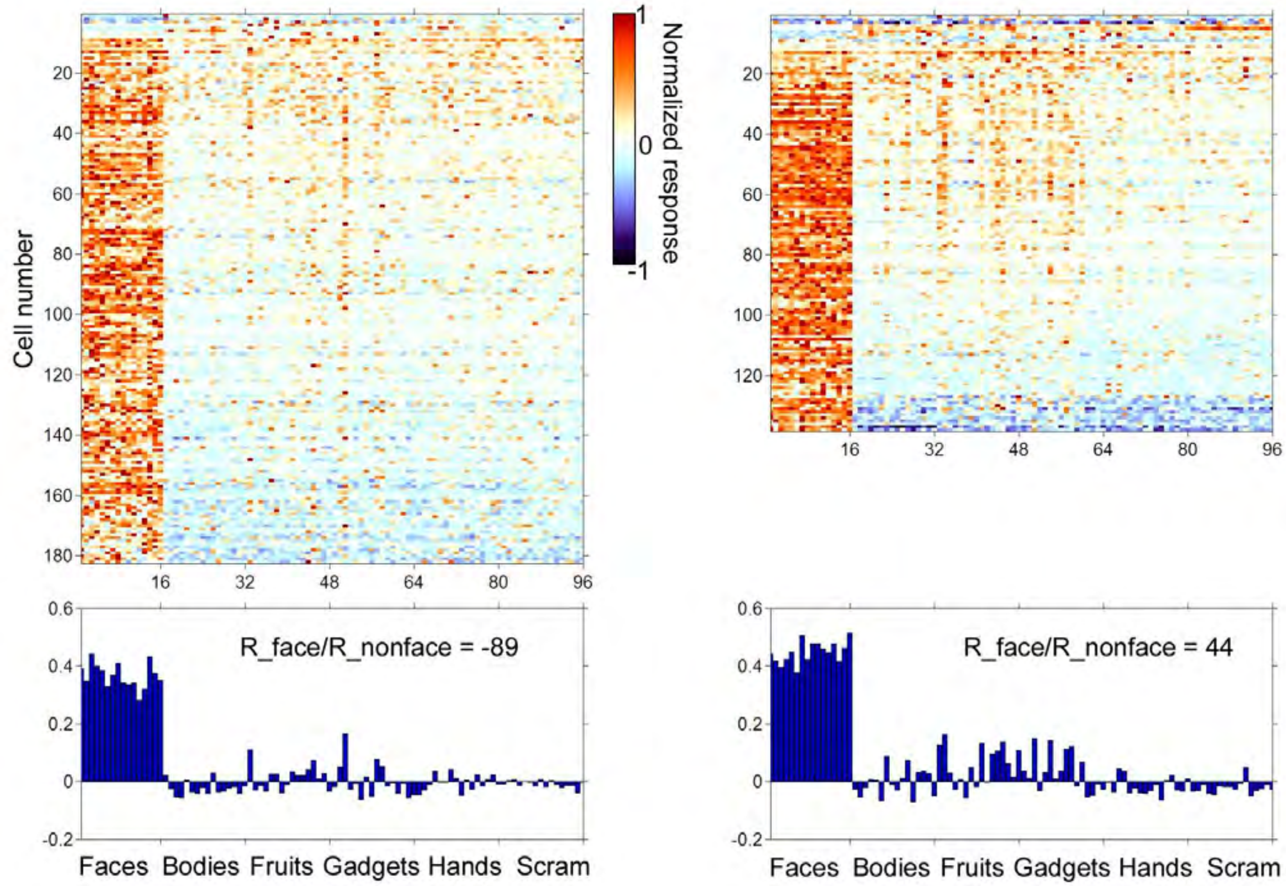
Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. Science, 311(5761), 670-674.





Monkey 1 (182 visually responsive cells /241)

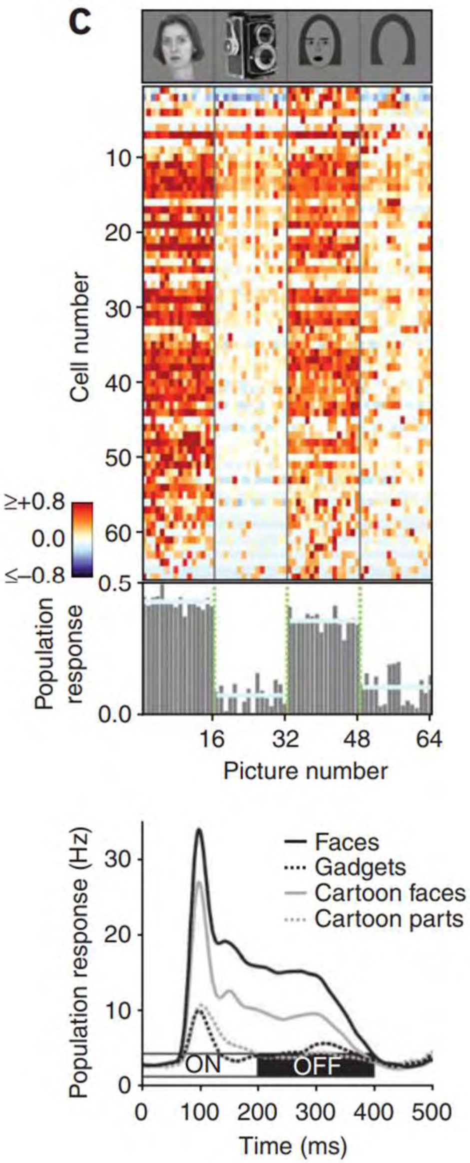
Monkey 2 (138/164)



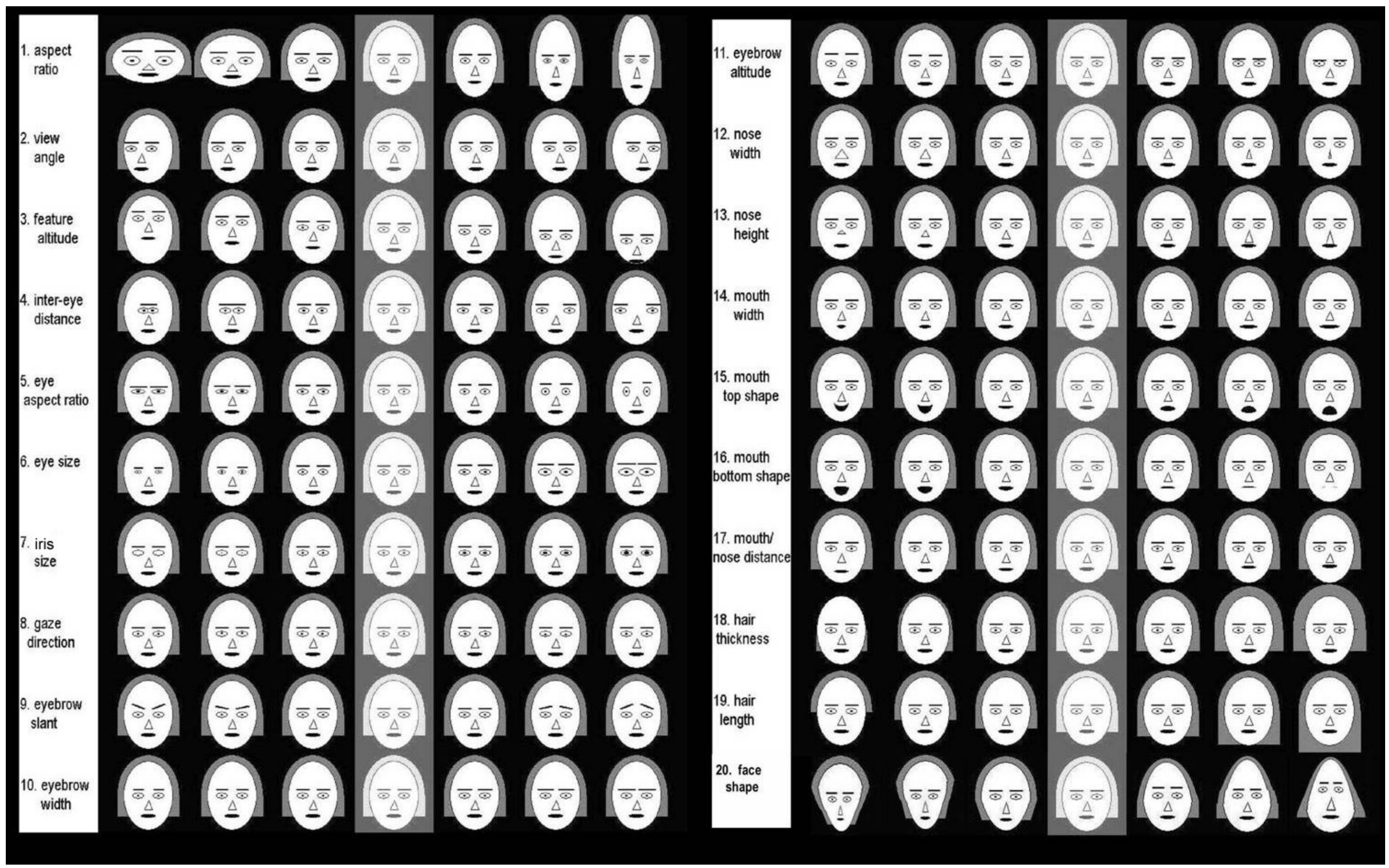
How are individual faces coded ? A study with cartoon faces

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nat Neurosci, 12(9), 1187-1196. https://doi.org/10.1038/nn.2363

Face cells respond well to cartoon faces



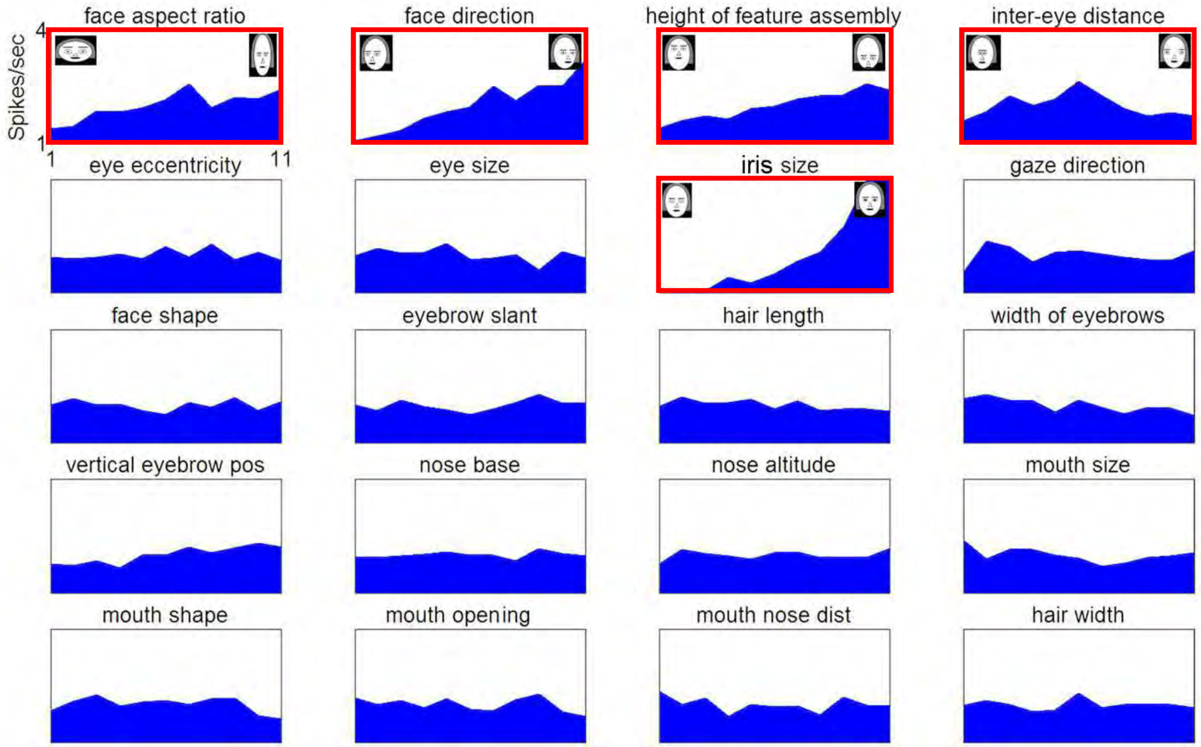
… allowing to systematically study various parameters of face variation



How are individual faces coded ? A study with cartoon faces

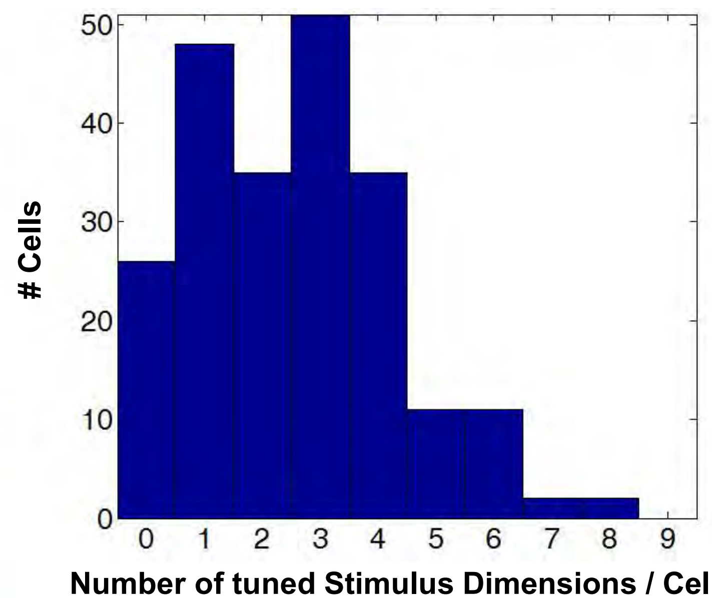
Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nat Neurosci, 12(9), 1187-1196. https://doi.org/10.1038/nn.2363

Tuning to Features of an Example Face Cell



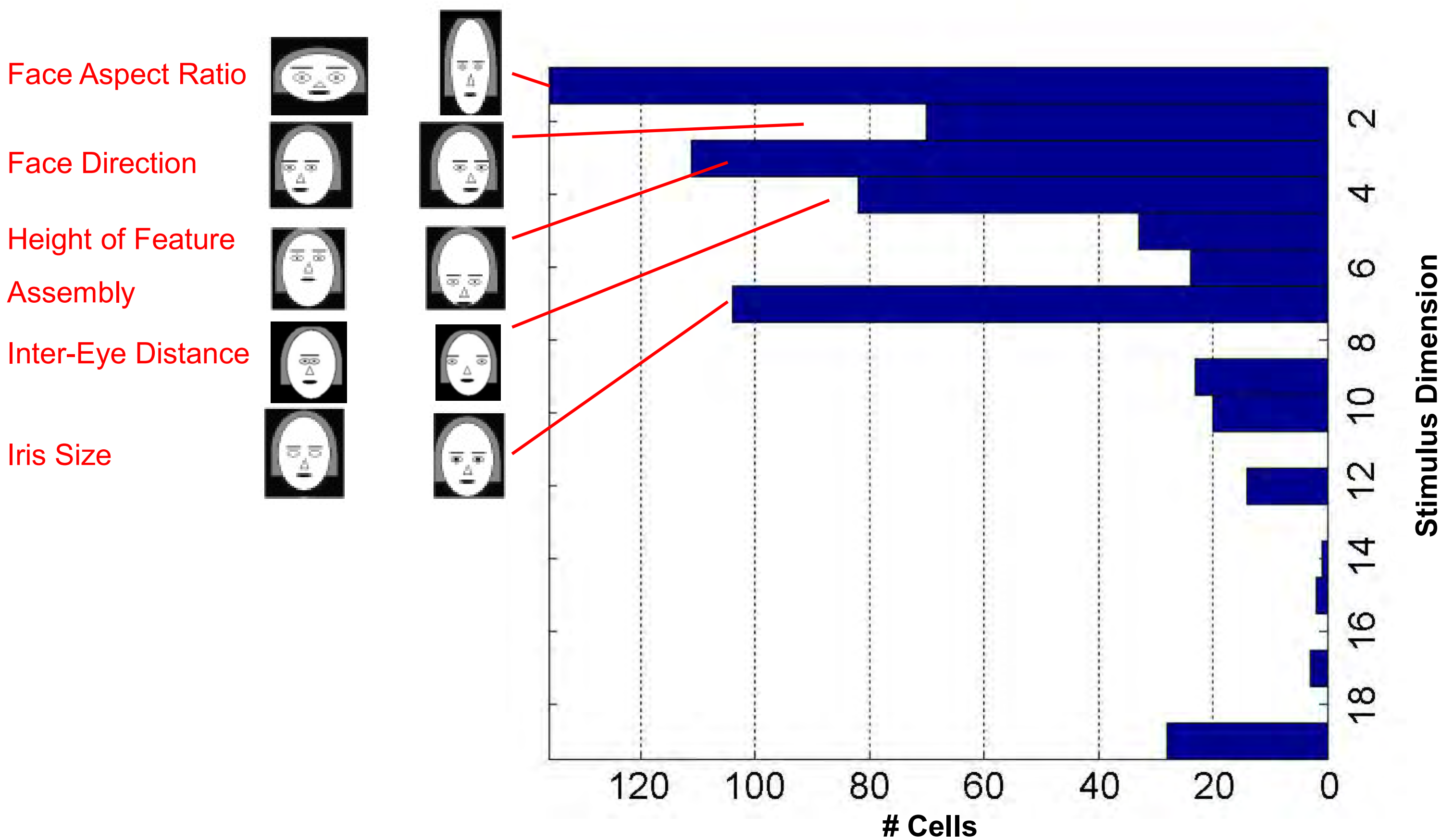
88% of face cells are tuned to some dimension.

Each cell is tuned to an average of 2.88 dimensions

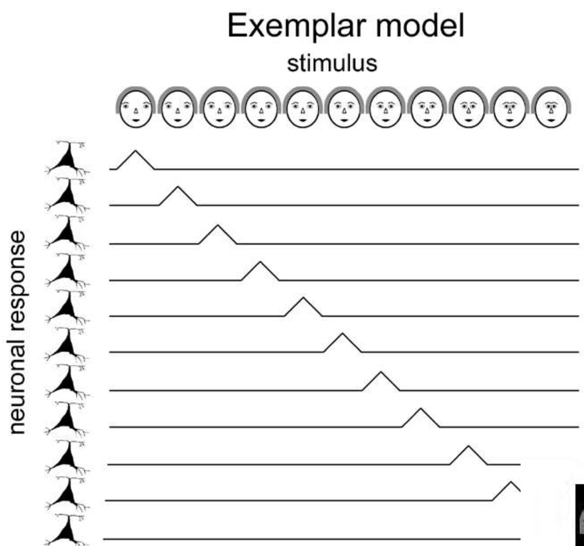


Which dimensions are face cells tuned to ?

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nat Neurosci, 12(9), 1187-1196. https://doi.org/10.1038/nn.2363

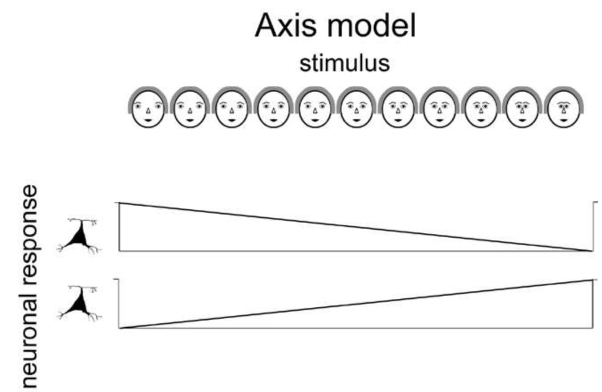


How are the units tuned? Monotonic coding of the main axes

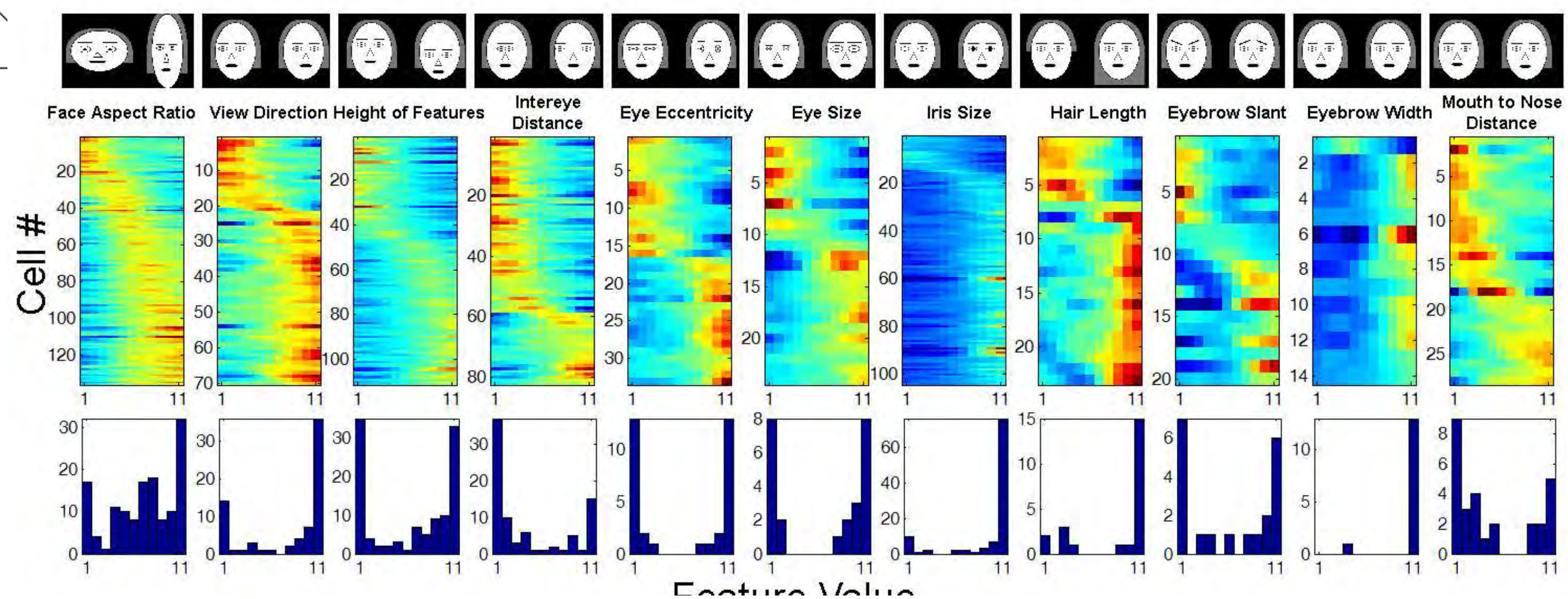


Two models can be opposed:

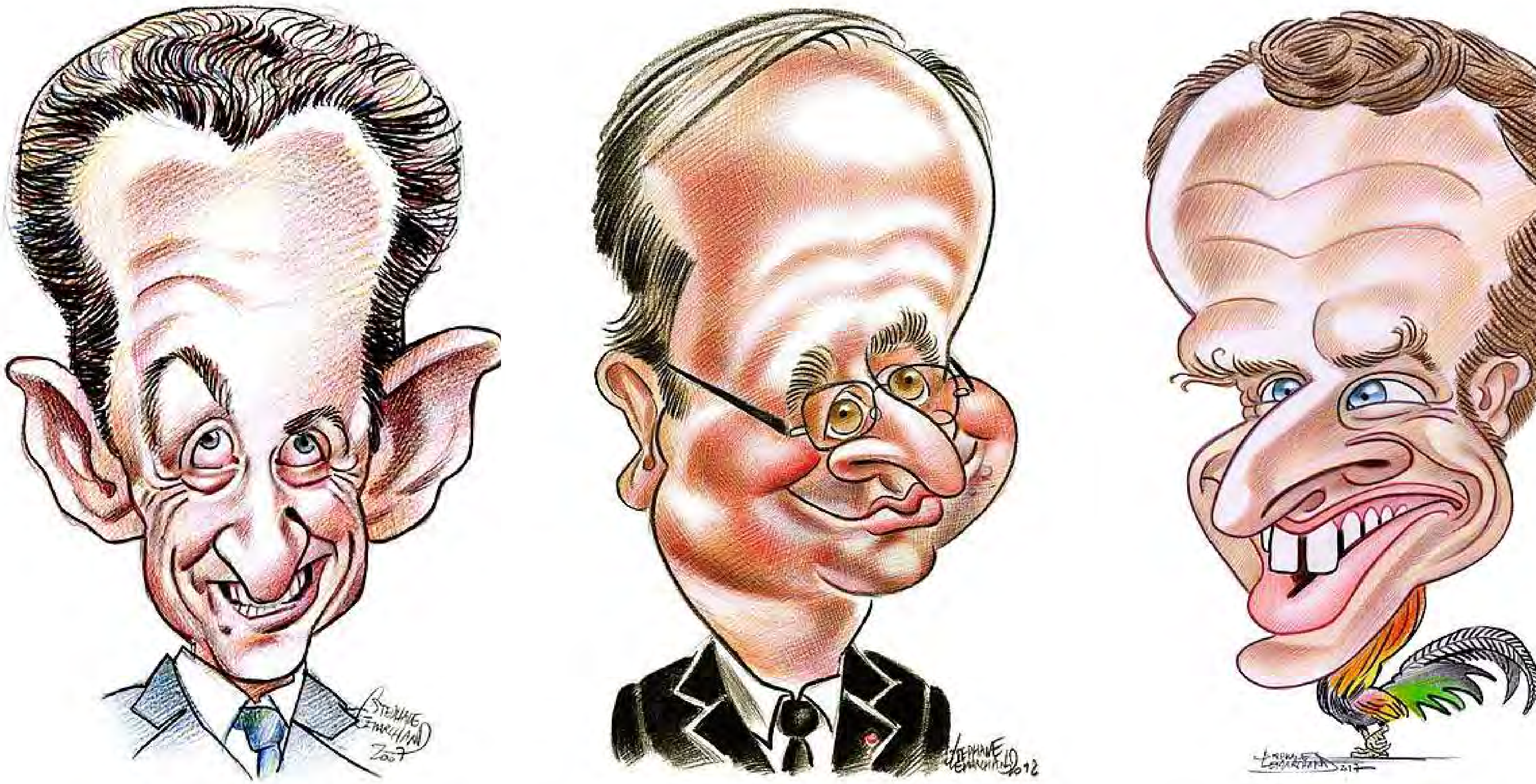
1. Cells tuned to a specific exemplar value of the parameter
2. Cells monotonically tuned to a specific **axis of the face space**



Answer:
Most cells respond in a monotonic manner…
to such an extent that they fire maximally to a stimulus that cannot occur in reality !



We recognize (and enjoy!) caricatures because our face neurons encode axes of face space and can respond beyond their normal values.



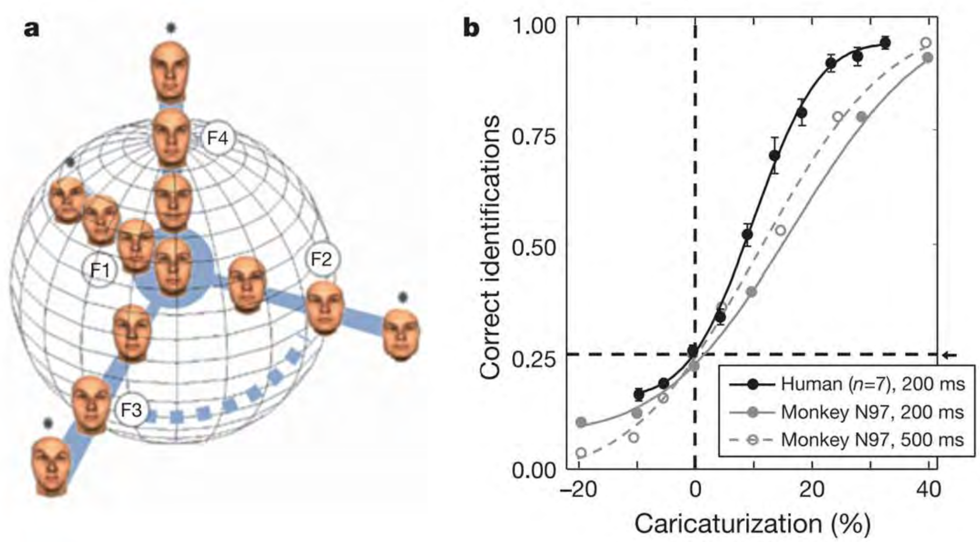
By Stéphane Lemarchand, Caricaturiste (from Wikimedia)

Caricatures activate monkey anterior inferotemporal neurons more than the original faces

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), Art. 7102. https://doi.org/10.1038/nature04951

Caricatures are **super-stimuli** for the brain's face neurons:

- They are increasingly easier to recognize as the amount of caricaturization increases (100% = normal face, >100% = caricature)
- They induce increasingly greater firing in face neurons



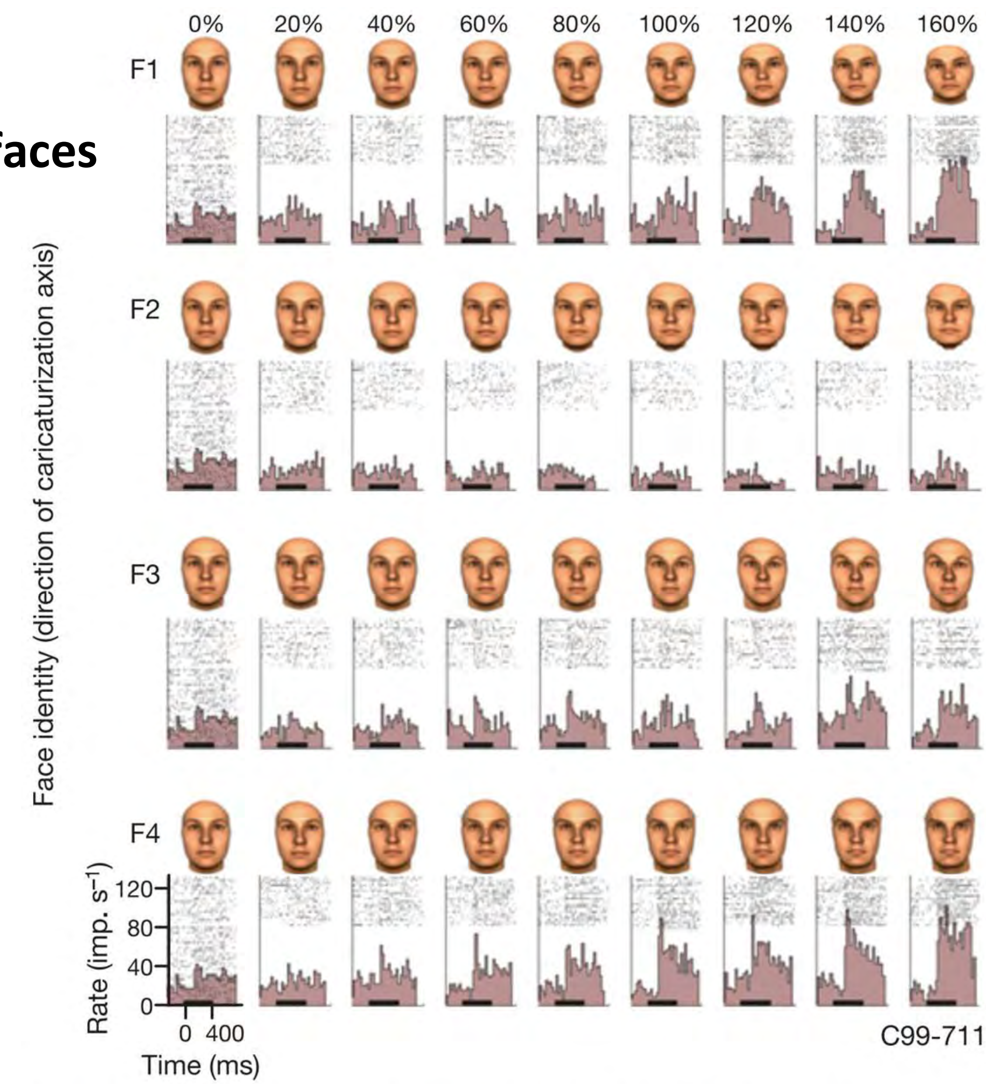


Figure 2 | Responses of a single neuron (C99-117) to faces along axes of caricaturization. Each panel corresponds to the rasters and peristimulus

A vector space for face recognition: an idea from computer and cognitive sciences

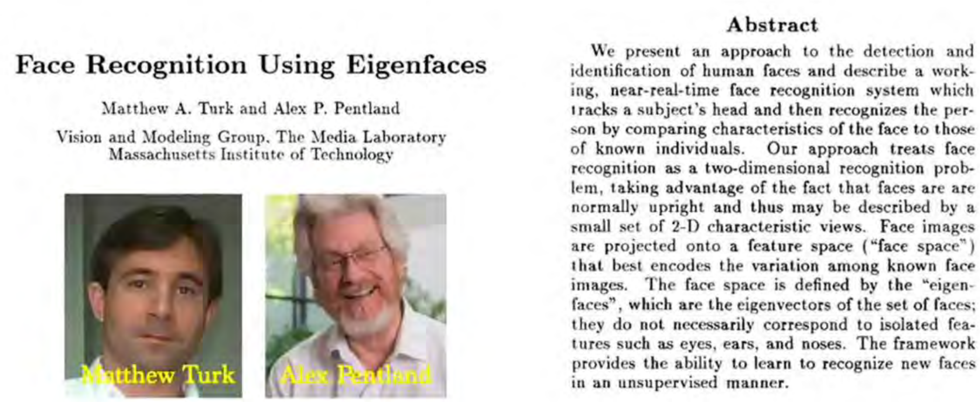
Sirovich, L. & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America A. 4 (3): 519–524.
Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3(1), 71-86. https://doi.org/10.1162/jocn.1991.3.1.71

Doris Tsao et al.'s research suggest that the middle face patch may act as a **vector space for faces :**

- The neural code is distributed over many cells
- Each cell encodes a few axes of variation of faces

→ The firing of various cells may encode the **coordinates of a given face in the space**

The idea of a face space has a long history in computer and cognitive science:
For automatic face recognition: it's a good idea to **compress the face space** : start from many pictures of faces, and identify their main axes of variation to discover a **basis set for faces.**
https://en.wikipedia.org/wiki/Eigenface : "The eigenvectors are derived from the covariance matrix of the probability distribution over the high-dimensional vector space of face images. The eigenfaces form a basis set of all images used to construct the covariance matrix".
In other words, apply **principal component analysis** to discover the main axes.
Each individual face is characterized by its coordinates on those axes.
Each face is a **weighted sum** of canonical face variations



Face Recognition Using Eigenfaces

Matthew A. Turk and Alex P. Pentland

Vision and Modeling Group, The Media Laboratory
Massachusetts Institute of Technology

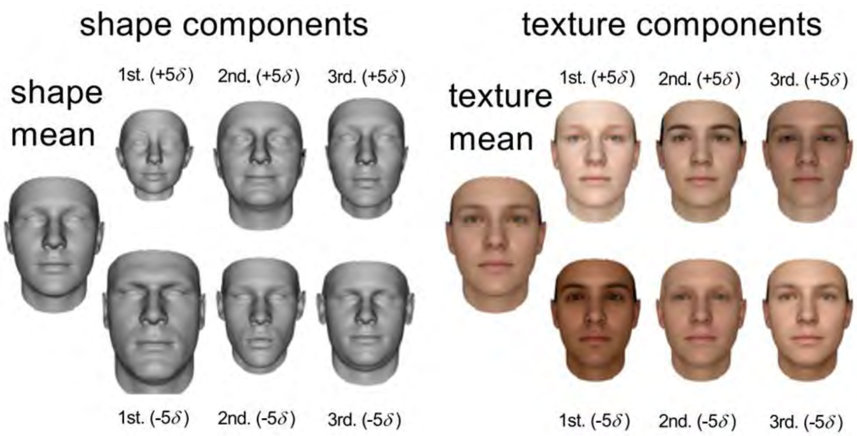
Abstract

We present an approach to the detection and identification of human faces and describe a working, near-real-time face recognition system which tracks a subject's head and then recognizes the person by comparing characteristics of the face to those of known individuals. Our approach treats face recognition as a two-dimensional recognition problem, taking advantage of the fact that faces are are normally upright and thus may be described by a small set of 2-D characteristic views. Face images are projected onto a feature space ("face space") that best encodes the variation among known face images. The face space is defined by the "eigenfaces", which are the eigenvectors of the set of faces; they do not necessarily correspond to isolated features such as eyes, ears, and noses. The framework provides the ability to learn to recognize new faces in an unsupervised manner.

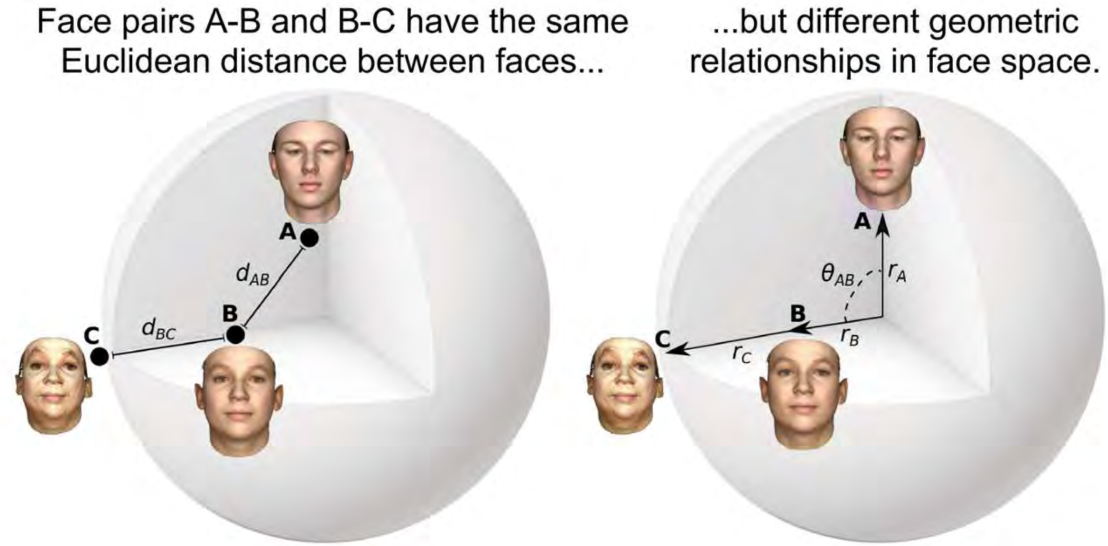
Distance in face space predicts perceptual similarity

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. Proceedings of the National Academy of Sciences, 119(27), e2115047119.

The Basel Face Model (BFM) is a formal 3D face space, with distinct axes for shape and for texture
(here showing the first 3 principal components)



Does the **Euclidean distance** between two points in face space have any psychological reality? Does it predict human similarity judgments ?

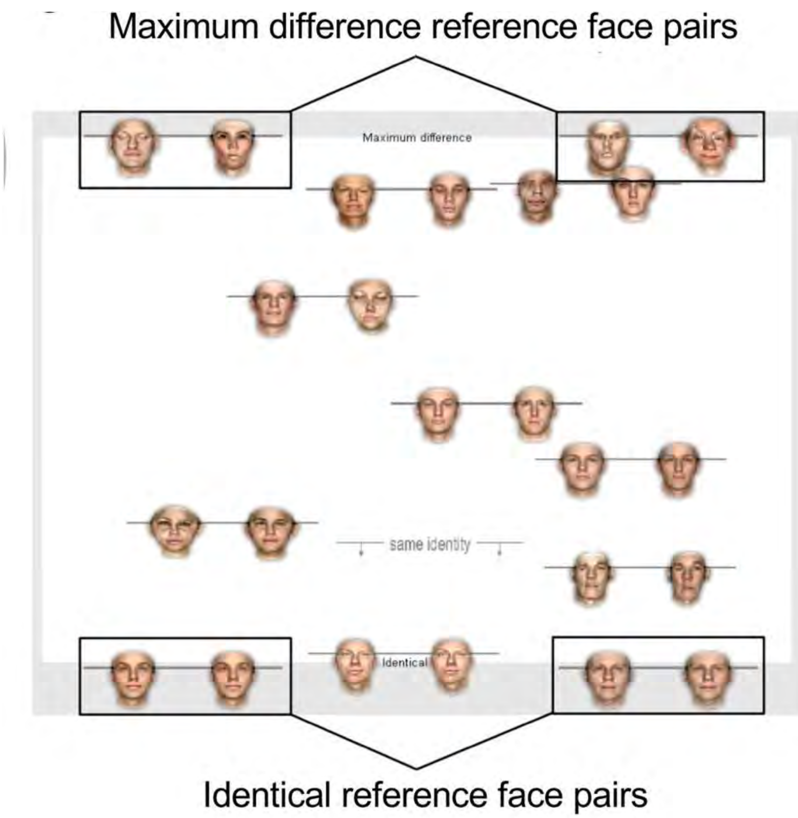


232 pairs of faces were carefully selected to span the face space and test Euclidean distance as opposed to other parameters such as angle or radius.
The predictions of various neural network models of face recognition are also tested.

Distance in face space predicts perceptual similarity

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. Proceedings of the National Academy of Sciences, 119(27), e2115047119.

Similarities for each pair are derived from a simple task of placing 8 randomly selected face pairs on a vertical axis from "identical" (bottom) to "maximum difference" (top)



Results: Dissimilarity ratings increase as a sigmoidal function of Euclidean distance.

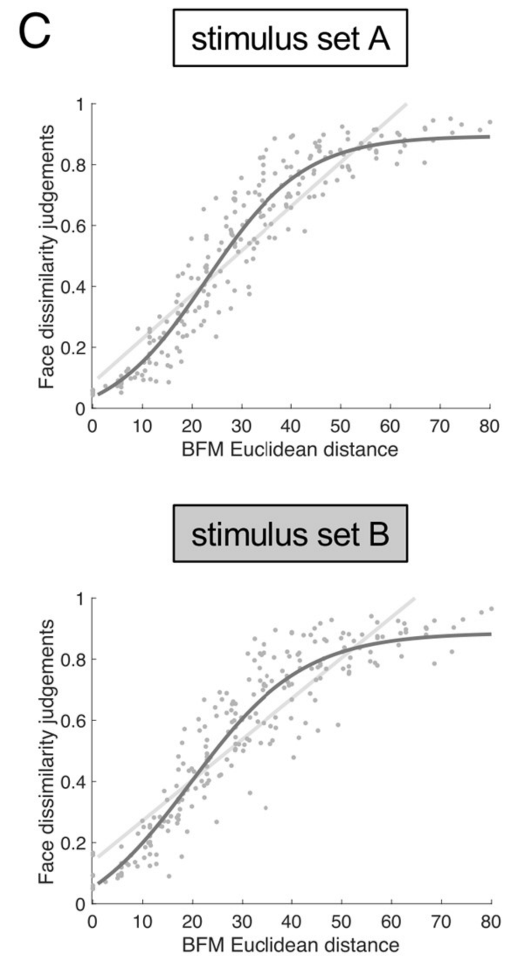
Euclidean distance is a marginally better predictor than angle or radius.

"Dissimilarity judgments are approximately **isotropic** in BFM face space" (whose axes are normally by the standard deviation of faces)

→ The direction of the pair of face factors can be rotated in any direction, without changing the similarity judgments.

Conclusion : Humans may have compiled statistics of the main axes of face variation (in their environment) and use them to encode any face.

→ Could explain the other-race effect.



Neural vectors for faces? Testing the face space at the neural level

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Taking **seriously** the idea that infero-temporal neurons implement a face space leads to many predictions:

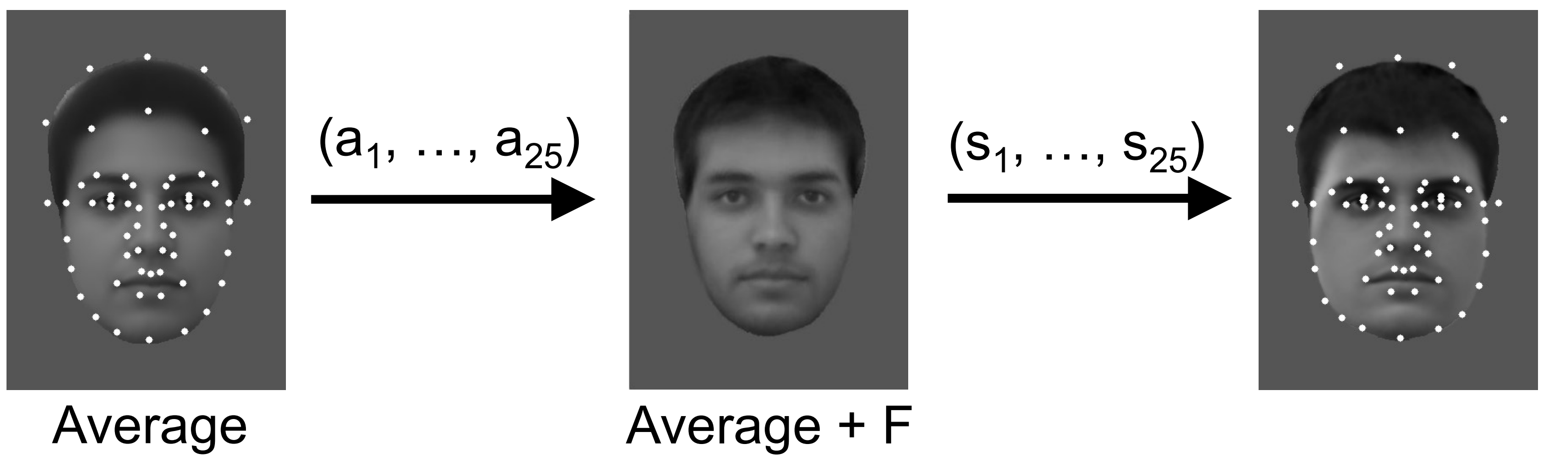
- There should be a small number of axes (forming a **basis set**) that define the space and its **dimensionality**
- Each face should correspond to a point in this vector space (and vice-versa)
- Each neuron should represent a direction in this vector space
    - Each neuron should possess a preferred axis
    - Each neuron should fire monotonically, **only** to the variations of faces along a certain axis (to the dot product of the face with its preferred axis)
    - Variations along perpendicular directions should have no effect

- If the dimensionality is small, then a small number of neurons (forming a new **basis** set) should suffice to characterize any given face
- And therefore, it should be possible to tell, from the votes of a sufficient number of neurons, which face was seen.

A parametrization of faces by their shape and appearance

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

$$F = (\underbrace{s_1, \ldots, s_{25}}_{\text{Shape}}, \underbrace{a_1, \ldots, a_{25}}_{\text{Appearance}})$$



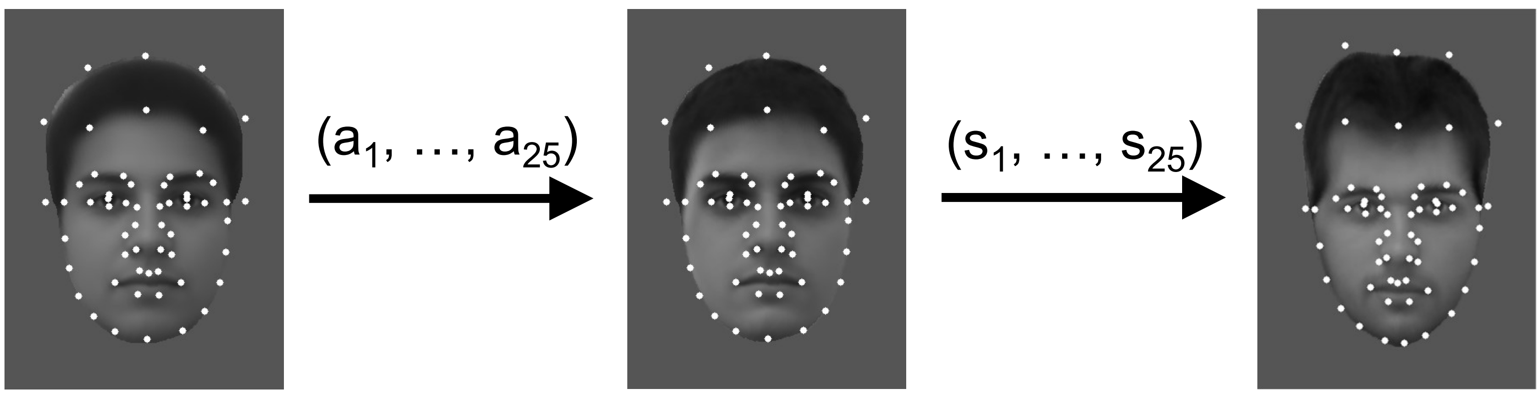
Changing appearance

Cootes & Taylor, 2001
Chang & Tsao, Cell 2017

A parametrization of faces by their shape and appearance

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

$$F = (\underbrace{s_1, \ldots, s_{25}}_{\text{Shape}}, \underbrace{a_1, \ldots, a_{25}}_{\text{Appearance}})$$



Average

Average + F

Changing shape

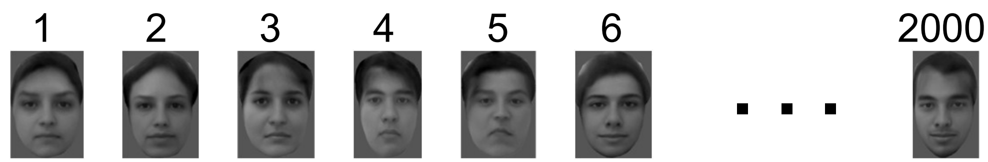
Cootes & Taylor, 2001
Chang & Tsao, Cell 2017

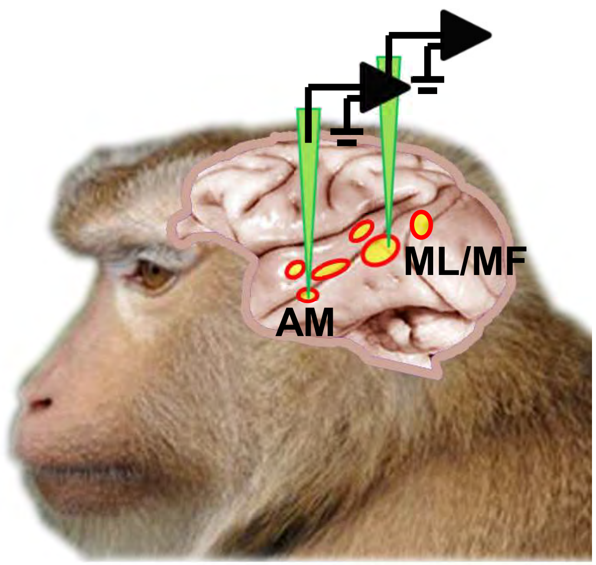
Quantifying the response of each neuron to faces covering the entire space

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

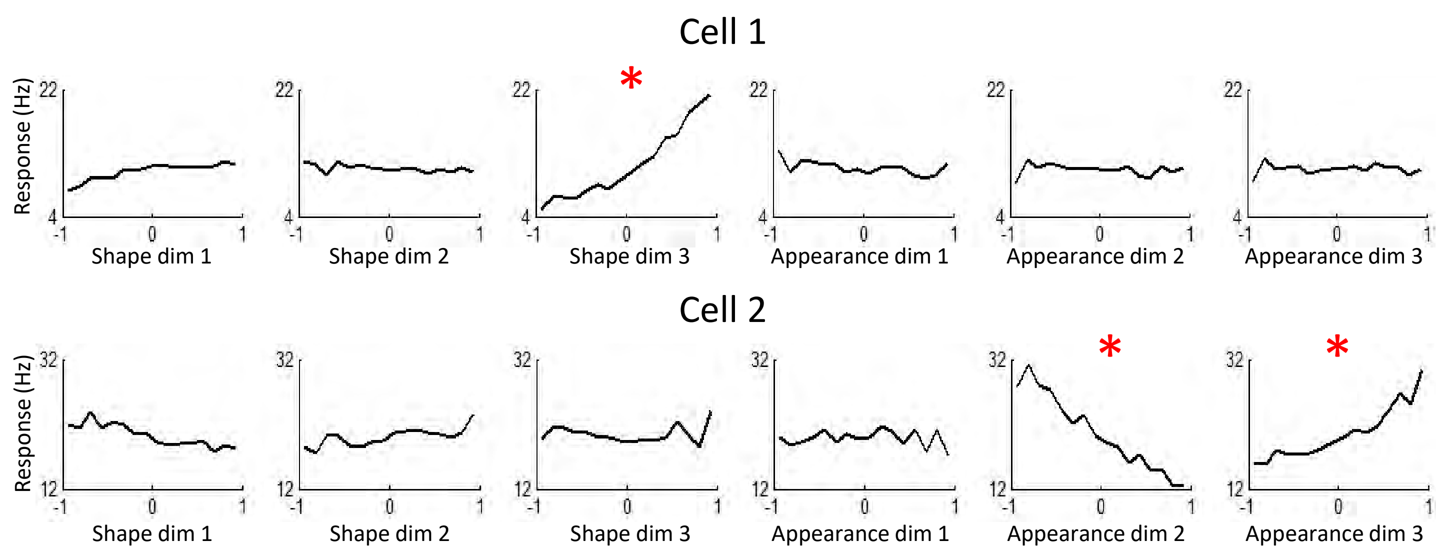
Experiment: “bombard” the monkey with 2000 faces, whose precise coordinates in the 25+25 face space are known

50-d face space



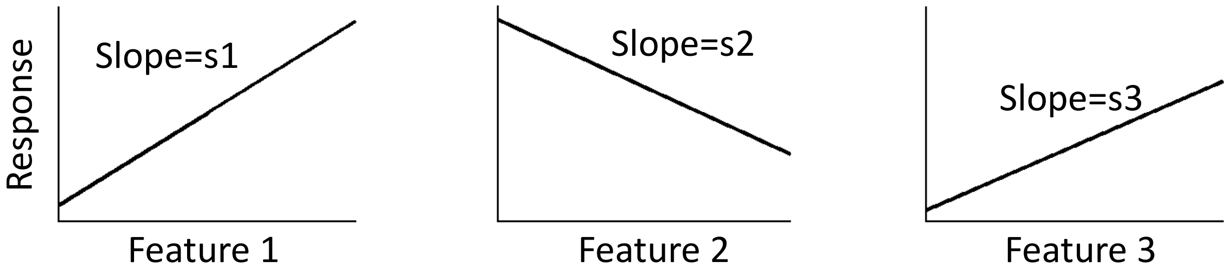


Characterize, for each neuron, the axes of variation to which it is responding.
Patch AM contains cells primarily tuned to appearance, patch ML/MF to shape



Prediction 1. If the relation betwen faces and neurons is linear, it can be inverted !

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011



response of cell = $s1 \cdot feature1 + s2 \cdot feature2 + \cdots\ s50 \cdot feature50$

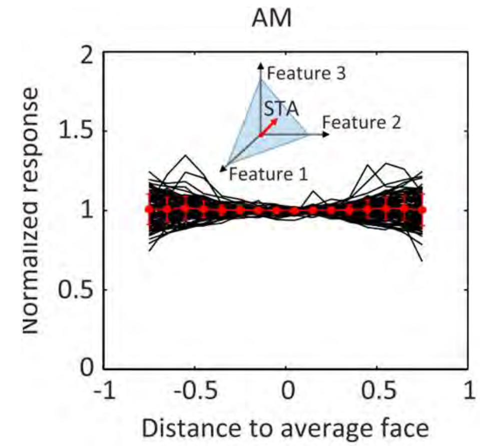
$$\begin{bmatrix} response_{cell\ 1} \\ response_{cell\ 2} \\ \vdots \\ response_{cell\ N} \end{bmatrix} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,50} \\ s_{2,1} & \cdots & s_{2,50} \\ \vdots & & \vdots \\ s_{N,1} & \cdots & s_{N,50} \end{bmatrix} \begin{bmatrix} feature_1 \\ feature_2 \\ \vdots \\ feature_{50} \end{bmatrix}$$

$$\begin{bmatrix} feature_1 \\ feature_2 \\ \vdots \\ feature_{50} \end{bmatrix} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,50} \\ s_{2,1} & \cdots & s_{2,50} \\ \vdots & & \vdots \\ s_{N,1} & \cdots & s_{N,50} \end{bmatrix}^{-1} \begin{bmatrix} response_{cell\ 1} \\ response_{cell\ 2} \\ \vdots \\ response_{cell\ N} \end{bmatrix}$$

50 face features = weight matrix * response of face cells

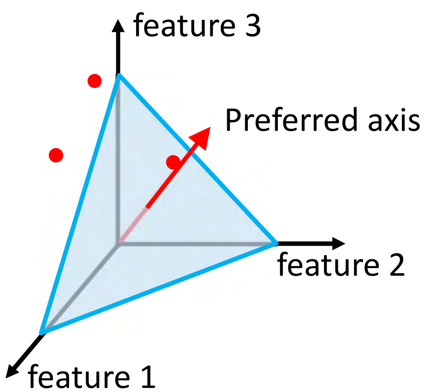
Prediction 2. Face metamers

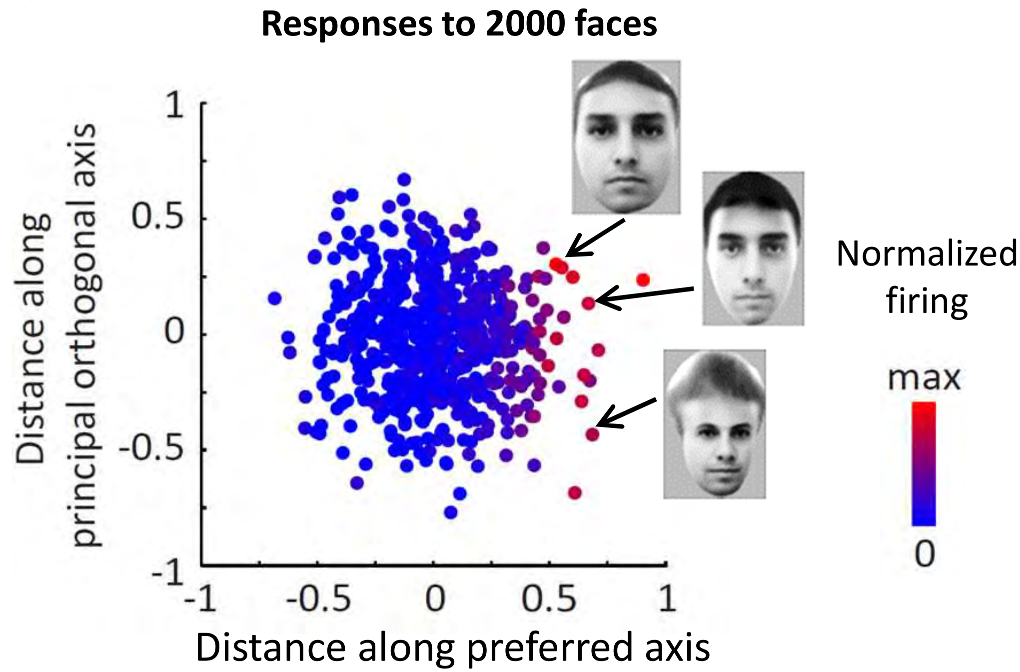
Each face has a **preferred vector** or **preferred axis**.
The model predicts that firing should only vary with the size of a face's projection on that preferred axis.
Therefore, all faces in a place perpendicular to that axis (blue plane) should be **metamers –** they should induce exactly the same firing in the chosen cell, even though they look very different.
Result: neurons do no show much, if any, variation in their firing for faces that vary along axes orthogonal to the preferred one.
On those axes, the response curve is decidedly non-Gaussian. This finding is incompatible with exemplar models, but predicted by the vector space model.
The tuning is also better with this particular face space than with Eigenfaces computed from the raw images.



Response of each face neuron =

$$\color{red}{s1}\cdot feature1+\color{red}{s2}\cdot feature2+\cdots\color{red}{s50}\cdot feature50$$

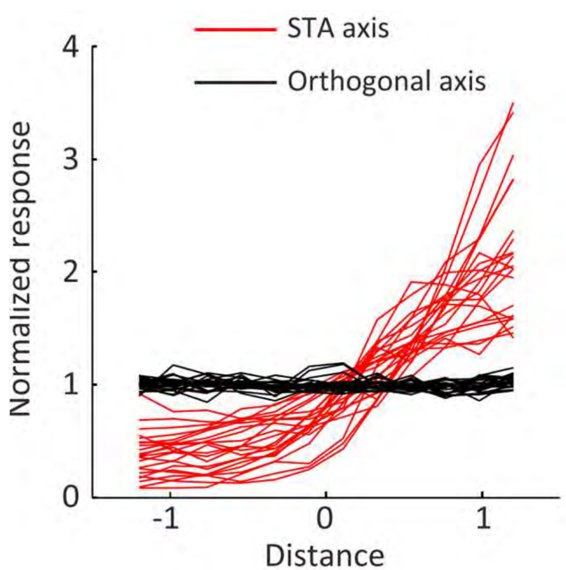


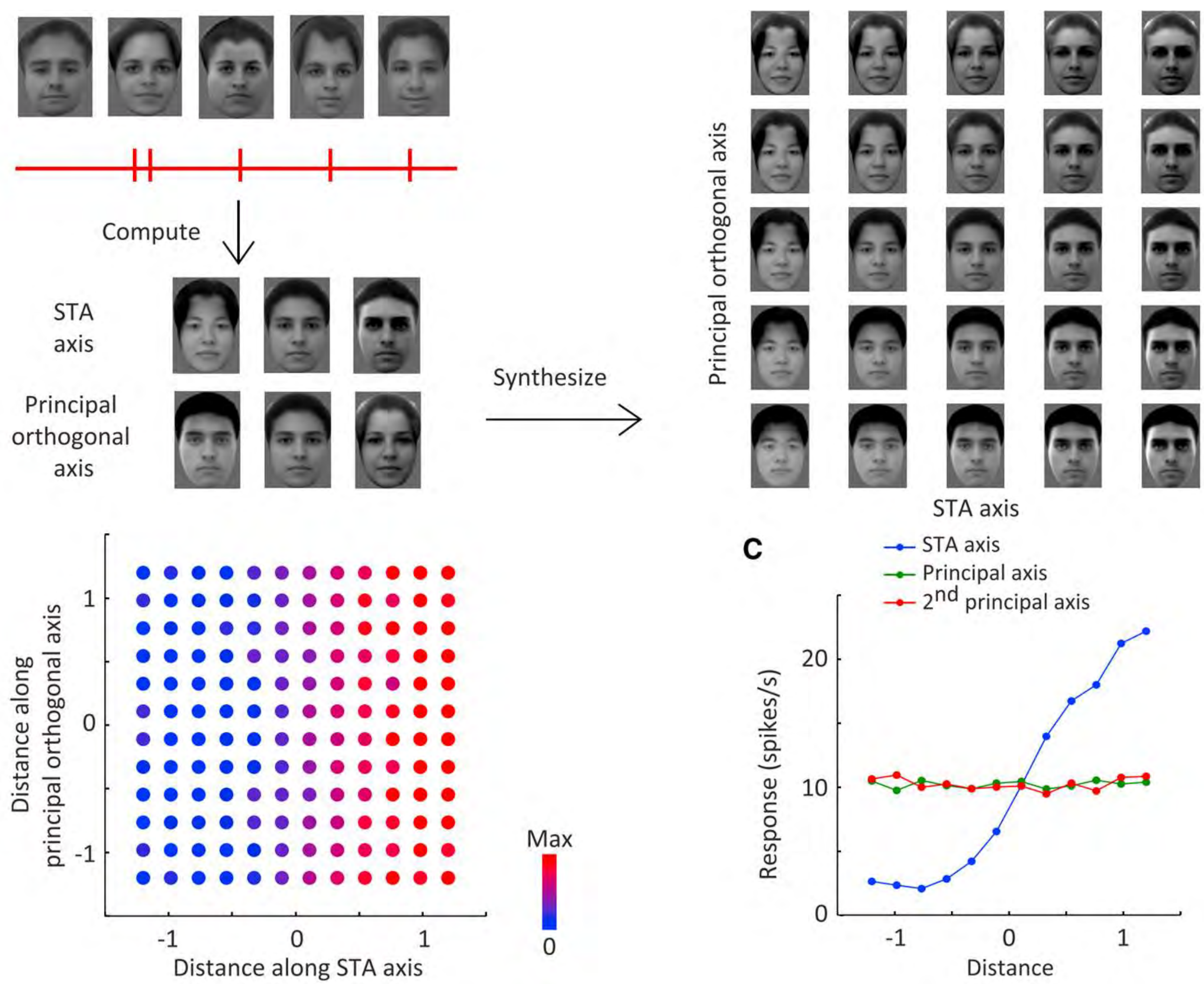


Prediction 2. Face metamers

This finding is so important that the authors replicated it with a dedicated experiment:
For each cell
- Find the principal axis
- Find the main orthogonal axis.
- Generate 144 faces (12x12) that span both axes
- Test the variation along either axis.

The results show firing variations only along the principal axis! And a completely flat curve along the other axis.





Auto-encoders : an even better predictor of neural responses to faces

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), 6456. https://doi.org/10.1038/s41467-021-26751-5

Can the dimensions of face variations be discovered automatically?
An auto-encoder is an artificial neural network that performs **dimensionality reduction.**
It is similar in logic to principal component analysis, but uses several non-linear stages to discover a **multidimensional compressed representation** that suffices to reconstruct the input.
A **beta variational autoencoder (Beta-VAE)** has an additional term that forces individual representational units to encode semantically meaningful dimensions.
« If each variable in the inferred latent representation is only sensitive to one single generative factor and relatively invariant to other factors, we will say this representation is disentangled or factorized.”
Explanations :
https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

- *Encoder* network: It translates the original high-dimension input into the latent low-dimensional code. The input size is larger than the output size.
- *Decoder* network: The decoder network recovers the data from the code, likely with larger and larger output layers.

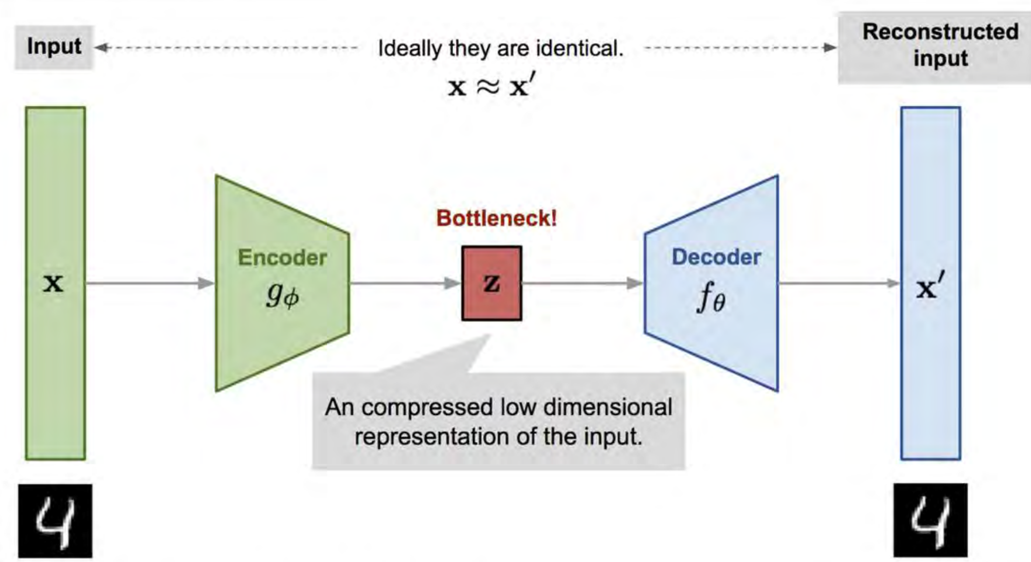


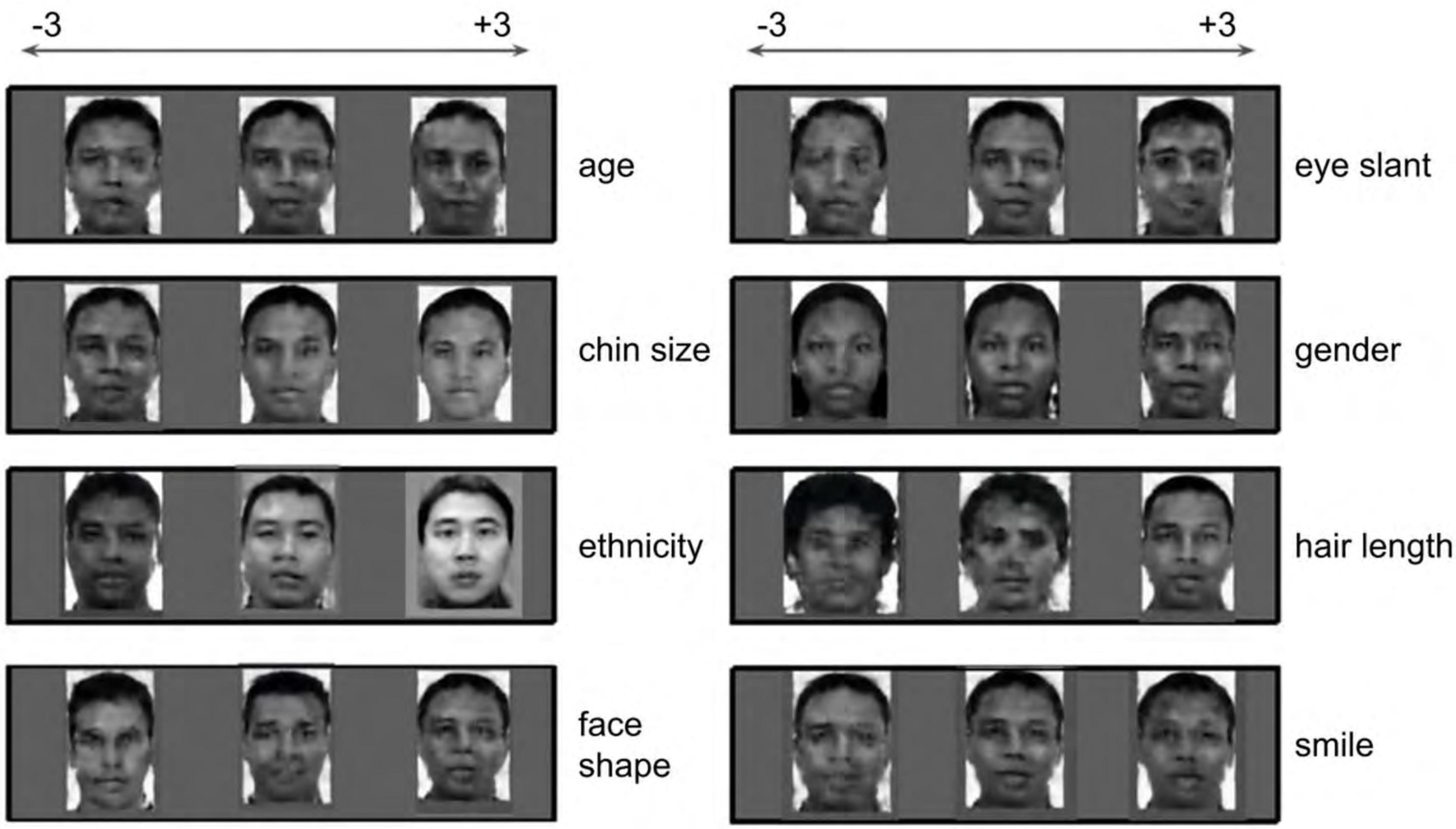
Fig. 1. Illustration of autoencoder model architecture.

The encoder network essentially accomplishes the dimensionality reduction, just like how we would use Principal Component Analysis (PCA) or Matrix Factorization (MF) for. In addition, the

Auto-encoders : an even better predictor of neural responses to faces

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), 6456. https://doi.org/10.1038/s41467-021-26751-5

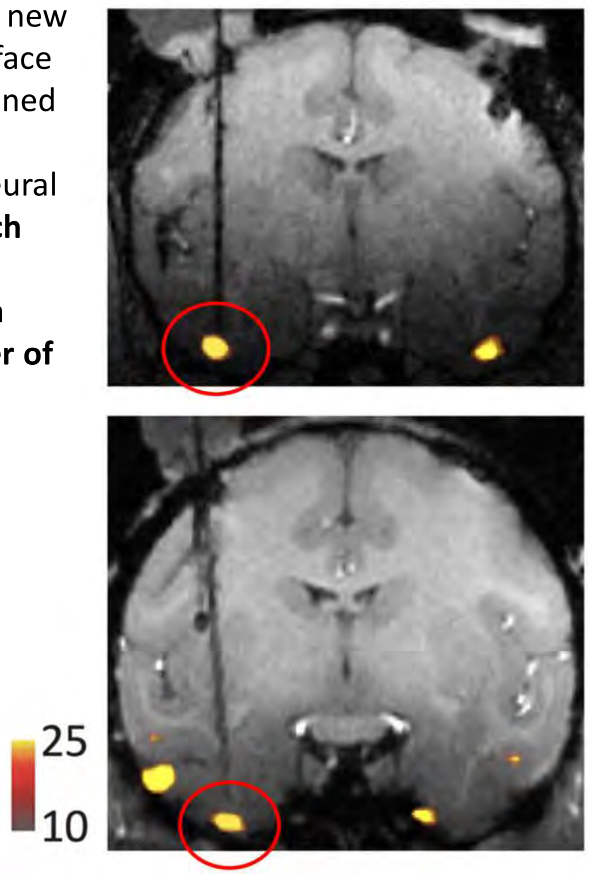
Higgins et al. used a beta-VEA to disentangle 2100 images of faces, and found that the dimensions were meaningful:

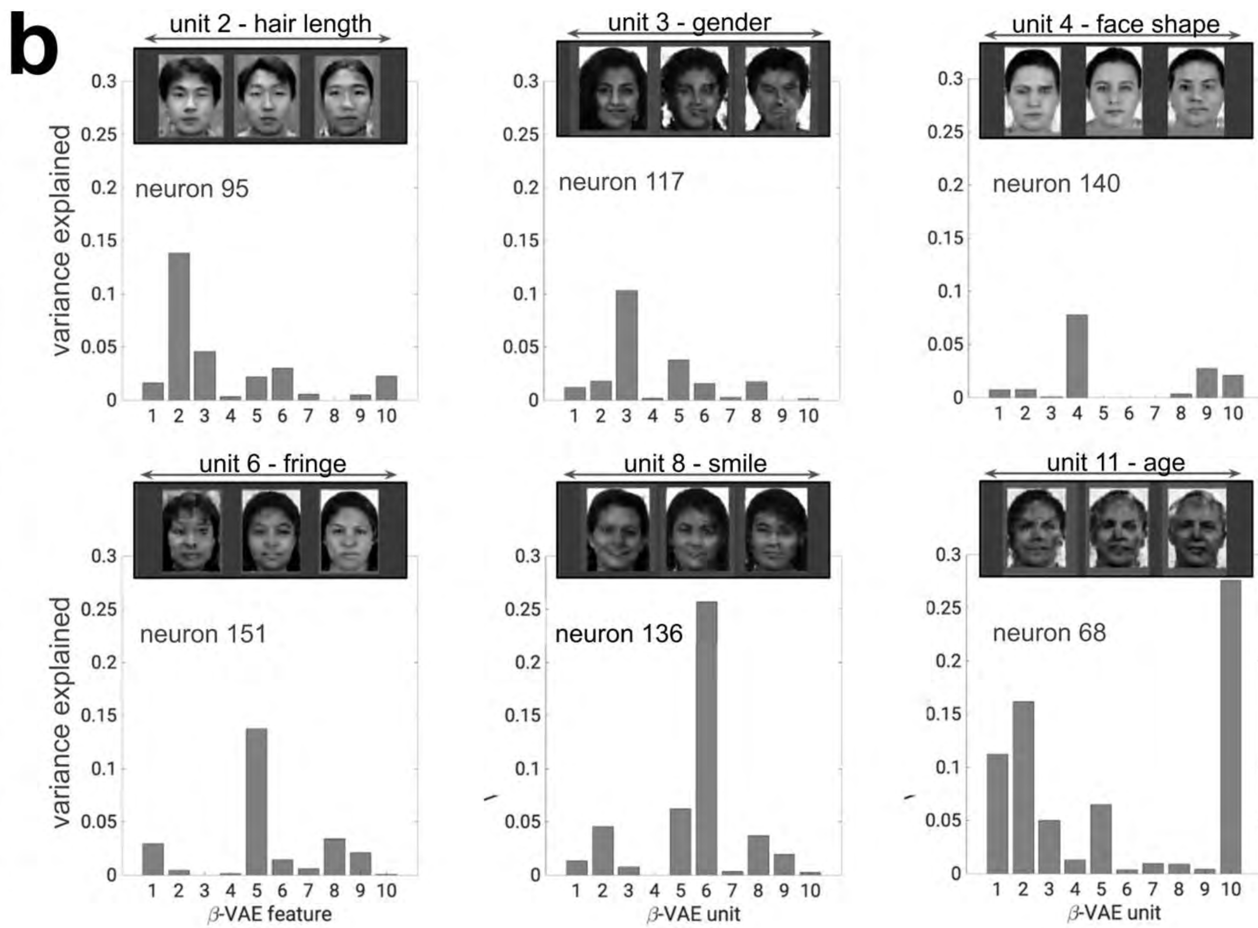


Auto-encoders : an even better predictor of neural responses to faces

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), 6456. https://doi.org/10.1038/s41467-021-26751-5

Crucially, these new dimensions of face variation explained better the variations in neural firing – and **each neuron now responded to a smaller number of dimensions**, typically 1 (as quantified by a measure of « alignment »).



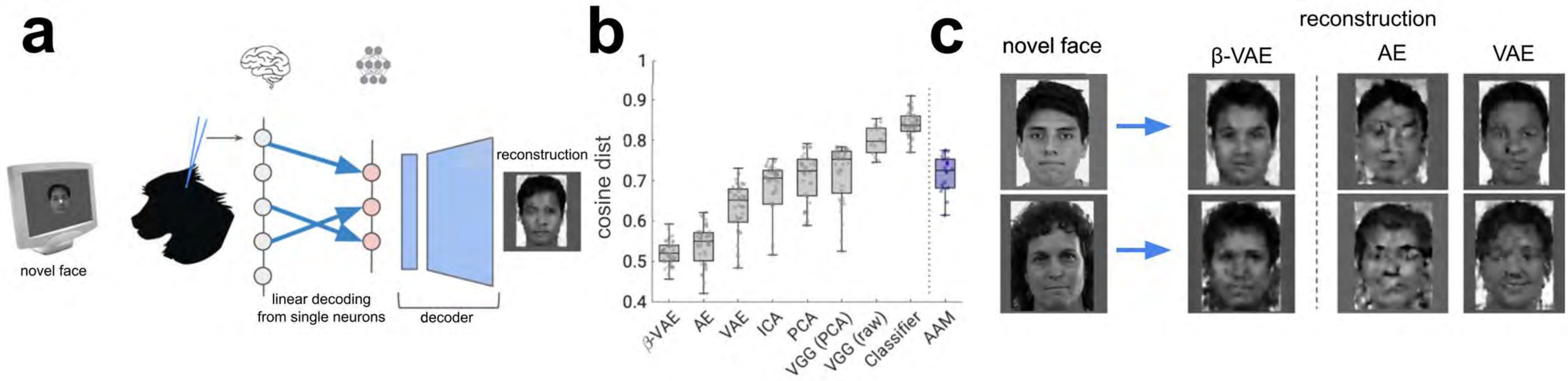


Auto-encoders : an even better predictor of neural responses to faces

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), 6456. https://doi.org/10.1038/s41467-021-26751-5

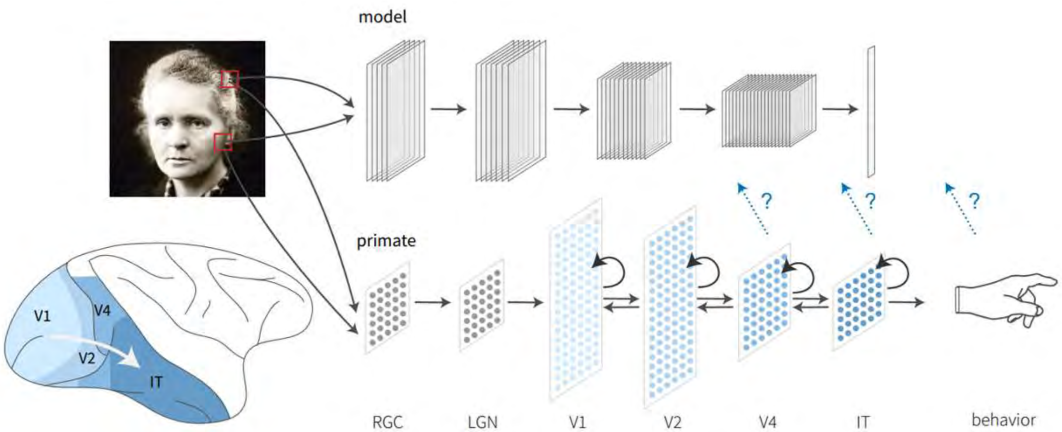
The beta-VAE also provides a **better basis set** for reconstructing a novel face.
(Here AAM = Active Appearance Model = Chang & Tsao's previous model).

Thus, during learning, infero-temporal cortex may disentangle the representation of visual objects such as faces, resulting in a compressed representation using a low-dimensional and orthogonal vector space.



Comparing artificial neural networks and real neuronal recordings

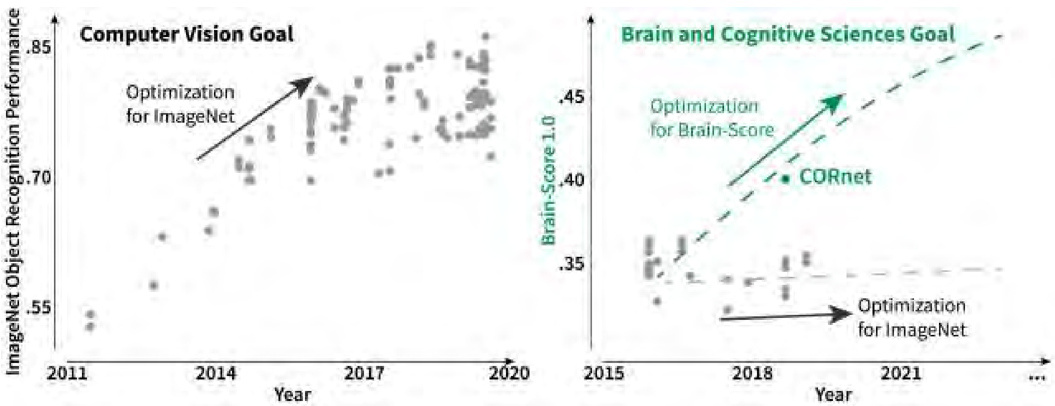
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score : Which Artificial Neural Network for Object Recognition is most Brain-Like? BioRxiv, 407007. https://doi.org/10.1101/407007
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. Neuron, 108(3), 413-423. https://doi.org/10.1016/j.neuron.2020.07.040

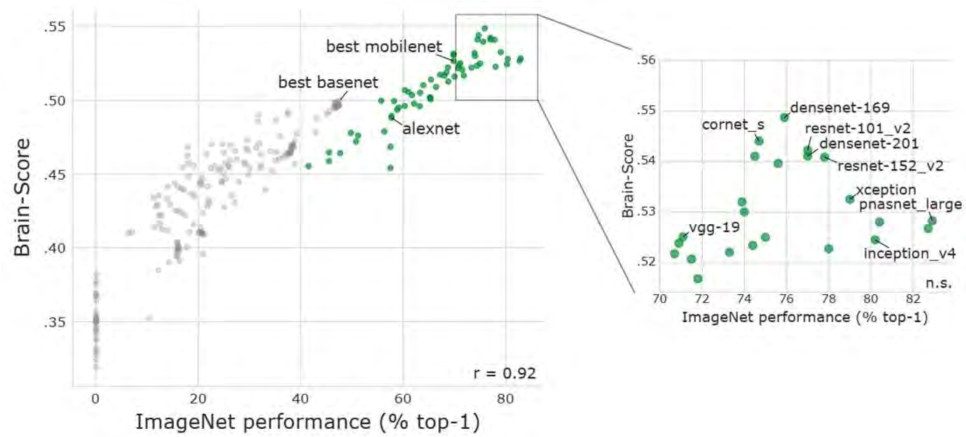


Artificial neural networks can be used as predictors of neural data, for instance using multiple regression or partial least squares.
→ Definition of a Brain-Score: how good is the fit?

The best-performing AI models are not necessarily the ones with the highest “brain score”.

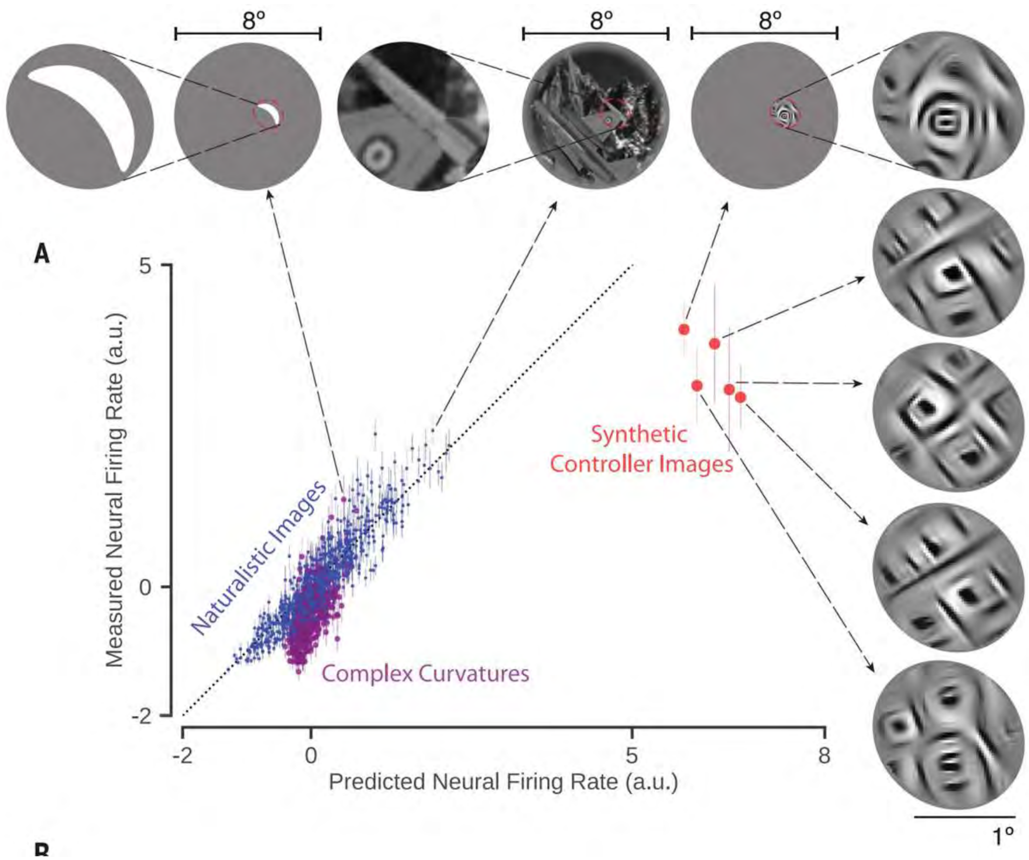
Empirically, one finds a correlation only up to a certain point, above which model architecture and parameters must be specifically optimized to fit the brain.





Artificial neural networks can help discover the “most exciting image” for real neurons

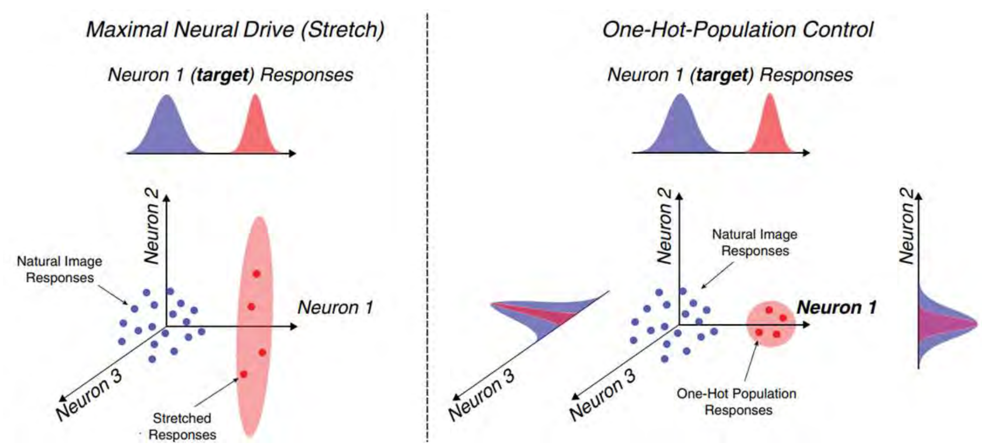
- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436.
- Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. Cell, 177(4), 999-1009.e10. https://doi.org/10.1016/j.cell.2019.04.005
- Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. Nature Neuroscience, 22(12), 2060-2065. https://doi.org/10.1038/s41593-019-0517-x



Once a single unit or a neural population has been fit to a neural network (using a first set of training images), classical AI techniques can be used to optimize the image to achieve a certain predicted output.

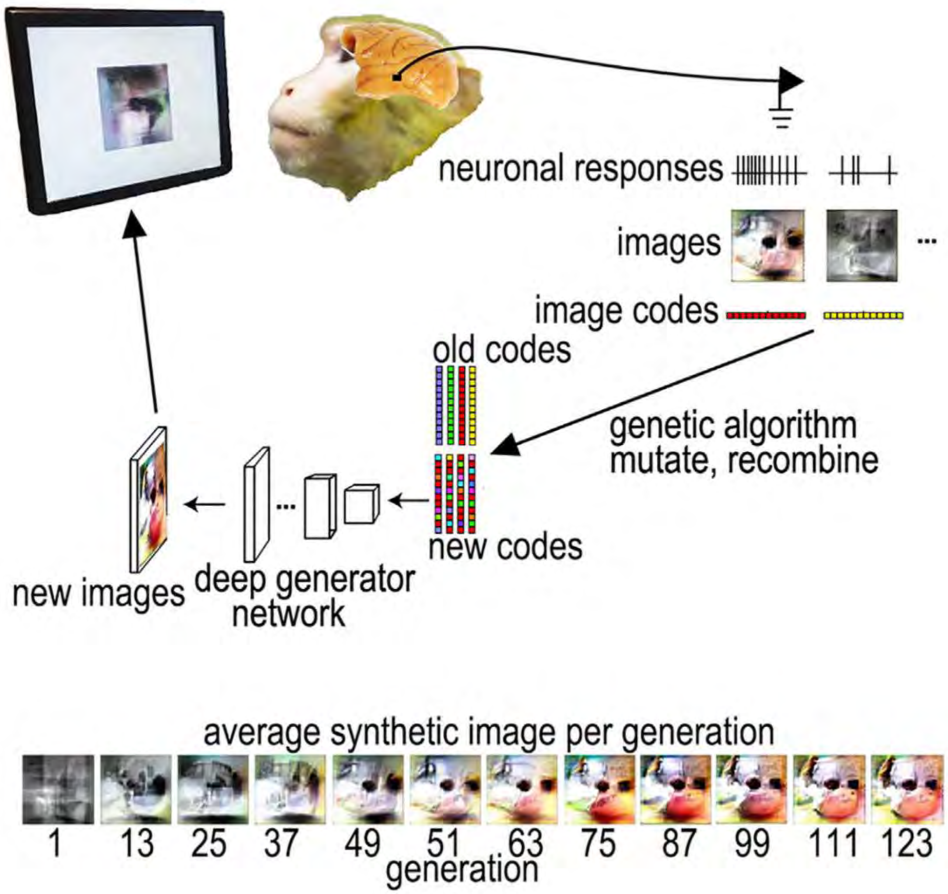
Bashivan et al.: A ConvNet correctly predicted 89% of the explainable (image-driven) variance in the neural responses of 107 V4 neurons.

Bashivan et al. then used gradient descent to find new pictures that were predicted to increase the firing of a given cell, either non-specifically (stretch) or specifically (one-hot control)… it worked!

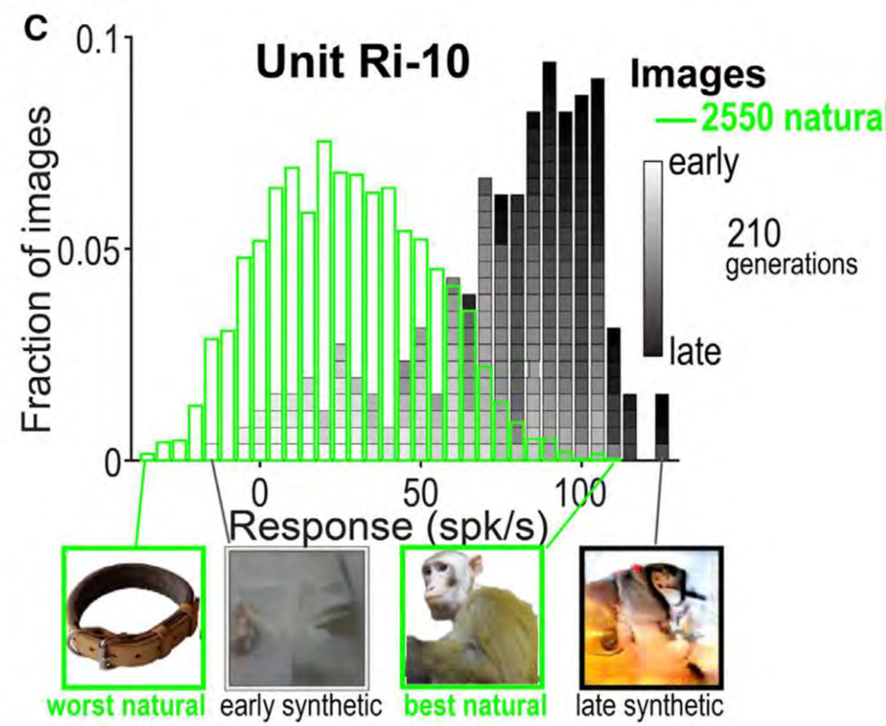


Artificial neural networks can help discover the “most exciting image” for real neurons

- Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. Cell, 177(4), 999-1009.e10. https://doi.org/10.1016/j.cell.2019.04.005

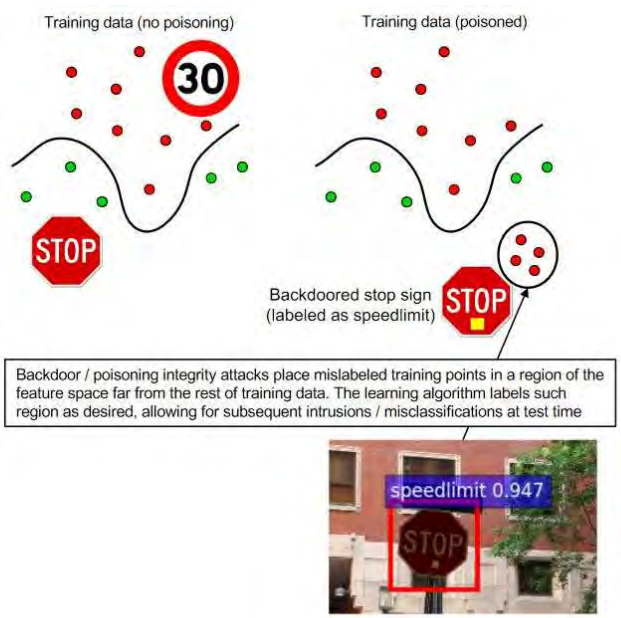


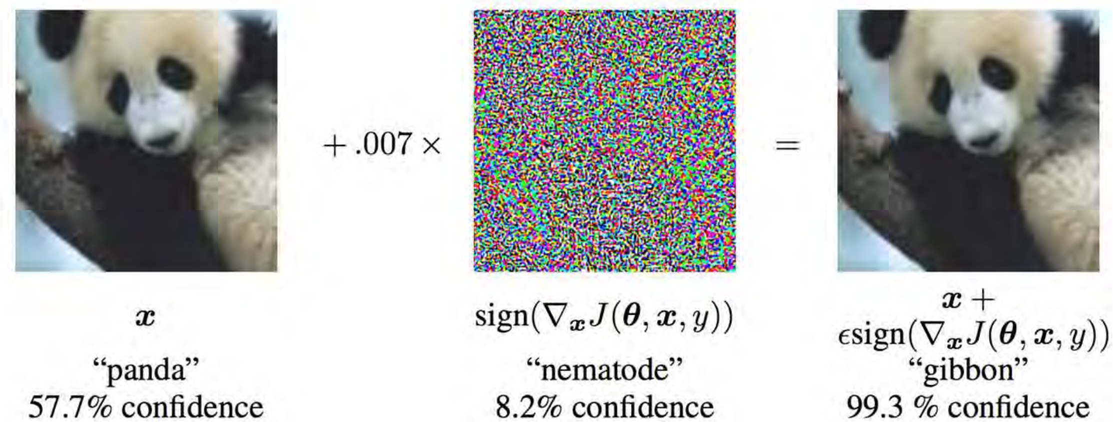
Ponce et al. use a similar idea to discover new super-stimuli (or “most exciting images”) for a given cell in infero-temporal cortex.
But… those new images do not necessarily improve our comprehension of what IT cortex does.

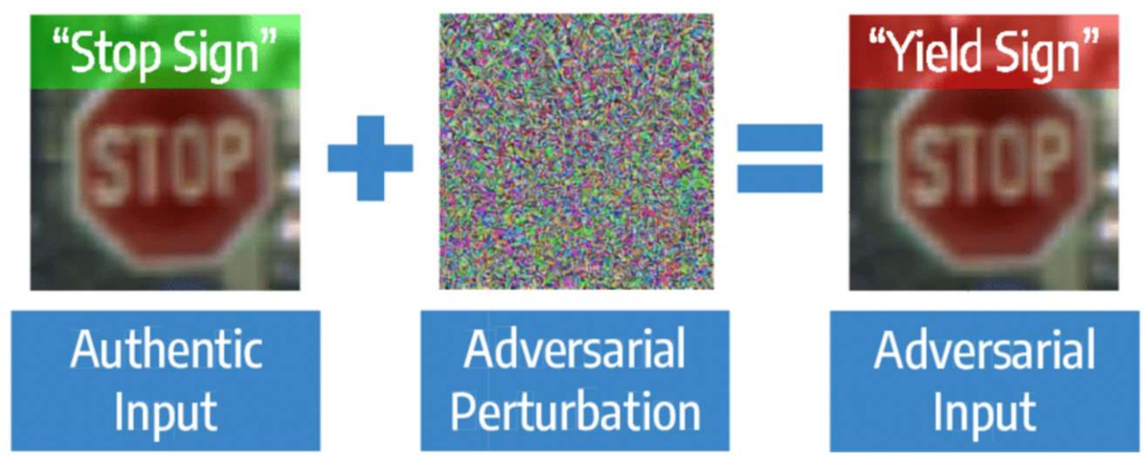


Sensibility to adversarial attacks in brains and machines

Artificial neural networks are sensitive to **adversarial attacks :** from the detailed knowledge of the network connections, it is possible to minimally (differentially) change the image to drastically change the output.
This is called a **white box attack**, but it is also possible to perform "**black box attacks**" – designing pictures that are mislabeled by not just one network, but many networks, even with different architectures, that are trained with the same data set.
It is also possible to "poison" the training data set.







Sensibility to adversarial attacks in brains and machines

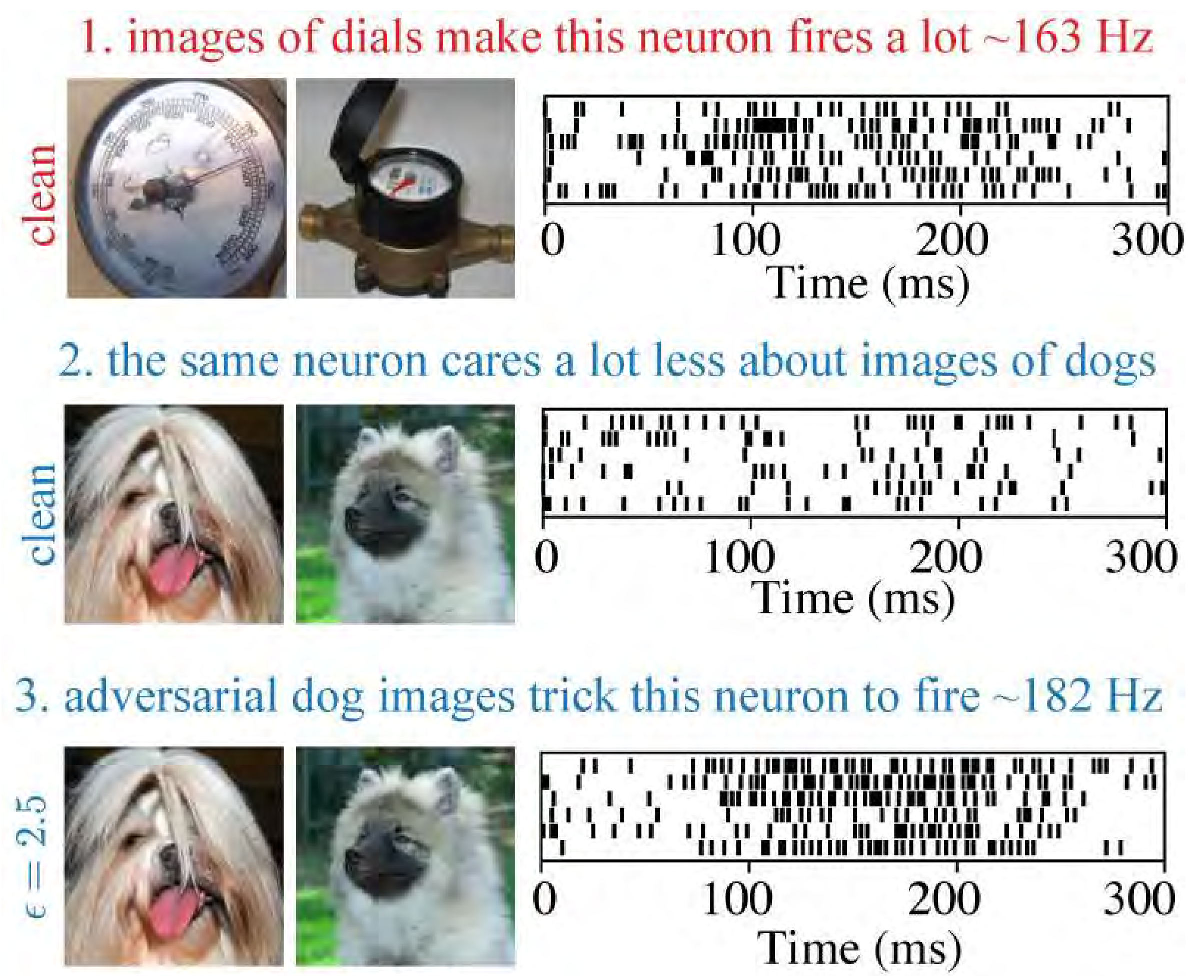
Guo, C., Lee, M. J., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & DiCarlo, J. J. (2022). *Adversarially trained neural representations may already be as robust as corresponding biological neural representations* (arXiv:2206.11228). arXiv. https://doi.org/10.48550/arXiv.2206.11228

We typically assume that real biological neural networks cannot be "attacked" in the same way as artificial ones.

Is this true, however?

Guo et al. find that they can easily find, for every neuron, images that are minimally (almost imperceptibly) changed and make the neuron fire to a seemingly unrelated category.

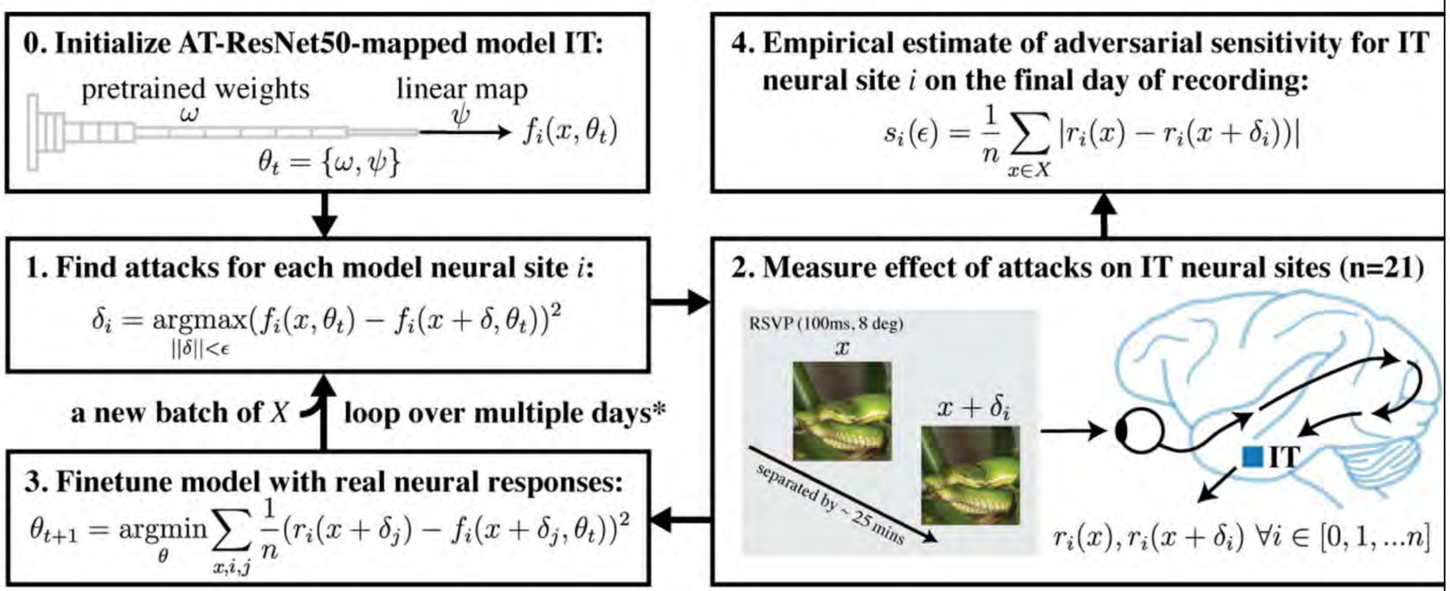
→ It is even possible to find "super-stimuli" that make the neuron fire much more.



Sensibility to adversarial attacks in brains and machines

Guo, C., Lee, M. J., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & DiCarlo, J. J. (2022). *Adversarially trained neural representations may already be as robust as corresponding biological neural representations* (arXiv:2206.11228). arXiv. https://doi.org/10.48550/arXiv.2206.11228

How is this possible?
Use a model (here ResNet) to predict the responses of real monkey infero-temporal neurons *i.*
Use the model to generate a plausible adversarial attack.
Check how good it is.
Improve the attack over several days.
The result is much better than just adding Gaussian noise or linearly mixing two images.



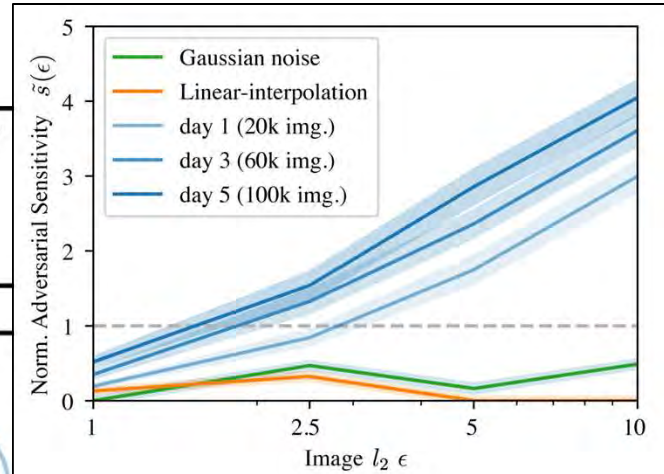


Figure 5. Normalized adversarial sensitivity lower bound improves over a number of days (blue) and is better than model-free perturbations such as Gaussian noise (green) or Linear-interpolation (orange). Sensitivity curve from day 0 where no mapping has occurred is not shown. Model-free measurements are measured as normalized root-mean-squared values which upper bounds the mean absolute differences.

Sensibility to adversarial attacks in brains and machines

Guo, C., Lee, M. J., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & DiCarlo, J. J. (2022). *Adversarially trained neural representations may already be as robust as corresponding biological neural representations* (arXiv:2206.11228). arXiv. https://doi.org/10.48550/arXiv.2206.11228

The authors propose a nice way to quantify the minimal amount of image change needed to drastically change the response
Sensitivity is expressed in standard deviations of the response over a large range of normal ("clean") images.
Results: IT neurons are less sensitive than an artificial neural network, but **more** sensitive than adversarially trained networks (AT).
The effects can be large: 1 to 3 standard deviations, even for very small image changes – thus creating "super-stimuli" for the neuron.
It is possible that we overestimate the robustness of our vision – or that downstream areas perform further error correction.

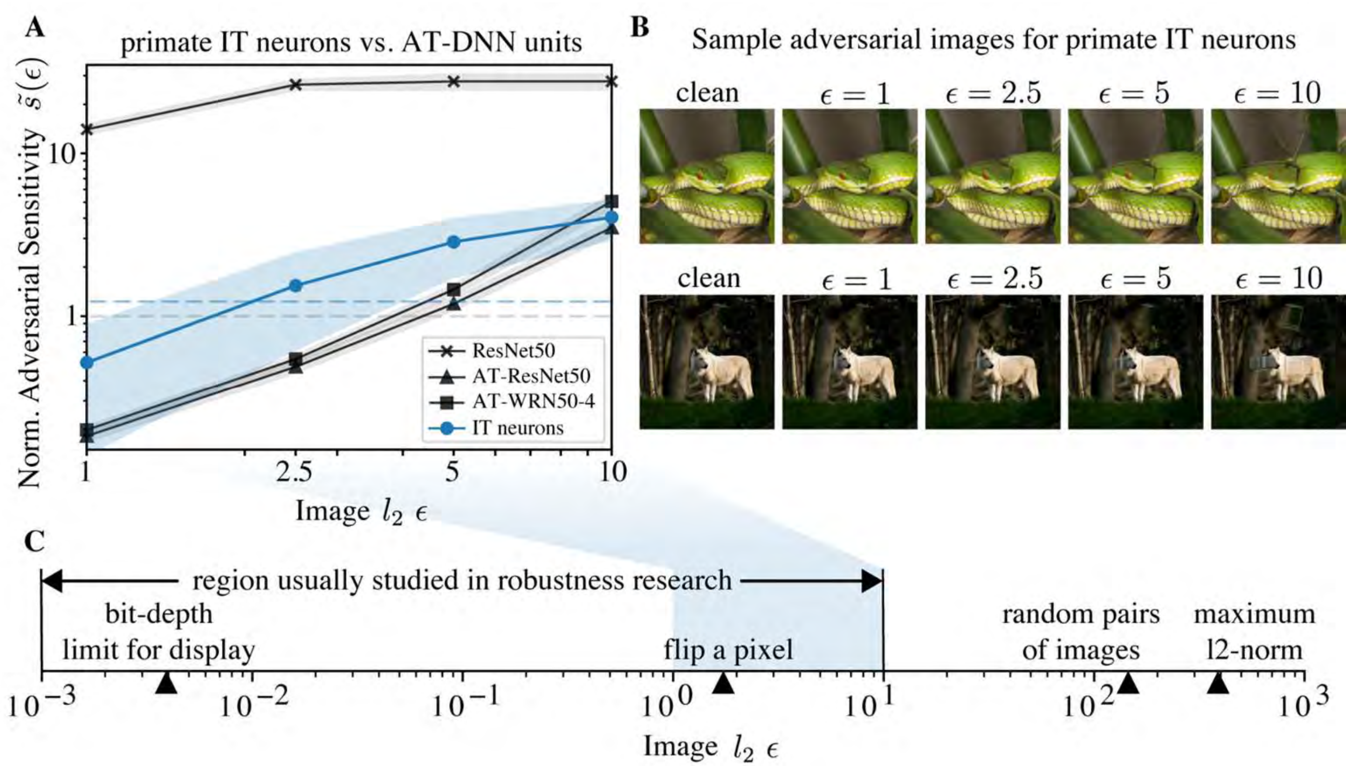
Our primary goal was to measure the sensitivity of the response of individual IT sites to worst-case local pixel perturbations of visual stimuli. For each neural site i, we measure its response $r_i(x)$ to clean images $x \sim \mathcal{D}$, where $\mathcal{D}$ is the ImageNet training set (Deng et al., 2009). We define the i^{th} site's image-specific neuronal adversarial sensitivity as the maximal observed movement that $r_i(x)$ makes under an l_2-norm bounded image perturbation δ which is ϵ away from the original image x:

$$s_i(x, \epsilon) = \max_{||\delta||<\epsilon} |r_i(x) - r_i(x+\delta)| \qquad (1)$$

Marginalizing the image distribution D, we define the i^{th} neural site's absolute adversarial sensitivity as:

$$s_i(\epsilon) = E_{x\sim D}[s_i(x, \epsilon)] \qquad (2)$$

To be able to compare individual IT neural sites with individual units in artificial neural networks, we use a dimensionless normalized adversarial sensitivity measure $\tilde{s}_i(\epsilon) = s_i(\epsilon)/\sigma_i$, where σ_i is the standard deviation of



Conclusions

At least for faces, the infero-temporal cortex (IT) in monkeys seems to be organized as a **vector space** :

- about ~50 axes sufficing to determine the identity of a face
- Each cell cares only about a few of those axes
    - Perhaps even one, if the axes are determined by an auto-encoder
- It is therefore possible to speak of “orthogonal faces” and “face metamers”
- IT cells for object recognition can be nicely modeled by artificial convolutional neural networks, to such an extent that such networks help
    - Find better stimuli for each cell
    - Find adversarial attacks

Vendredi 6 Janvier

COURS : **Vecteurs neuronaux ou cellules grand-mère : les représentations mentales sont-elles localisées ou distribuées ?**

SÉMINAIRE : **L’intelligence artificielle peut-elle modéliser le langage mathématique ?** — François Charton (FAIR Paris)

Vendredi 13 Janvier

COURS : **Géométrie des représentations visuelles : chaque visage est un vecteur**

SÉMINAIRE : **Commonsense Physical Reasoning in man and machine** — Ernest Davis (NYU, par zoom)

Vendredi 20 Janvier

COURS: **Exploiter la factorisation et les sous-espaces vectoriels pour coder l’information et communiquer entre aires cérébrales**

SÉMINAIRE : **Number symbols in the brain and mind** — Daniel Ansari (University of Ontario)

Vendredi 27 Janvier

COURS : **Comment prendre une décision ou faire des calculs avec des vecteurs dynamiques?**

SÉMINAIRE : **Comment se développent les réseaux cérébraux associés aux concepts mathématiques ?** — Marie Amalric (Université de Trento, Italie)

Vendredi 3 Février

COURS : **La représentation vectorielle des mots et des concepts**

SÉMINAIRE : **Les succès et les nouveaux défis de l’intelligence artificielle en mathématiques** — Léon Bottou (FAIR, New York)

Vendredi 10 Février

COURS : **La représentation vectorielle du langage : Comment représenter une phrase ?**

SÉMINAIRE : **Intuitions of mathematics and their refinement with age and education** — Manuela Piazza (Université de Trento, Italie)