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°3

Exploiter la factorisation et les sous-espaces vectoriels pour coder l'information et communiquer entre aires cérébrales

Course 3

Exploiting factorization and vector subspaces for information encoding and inter-areal communication

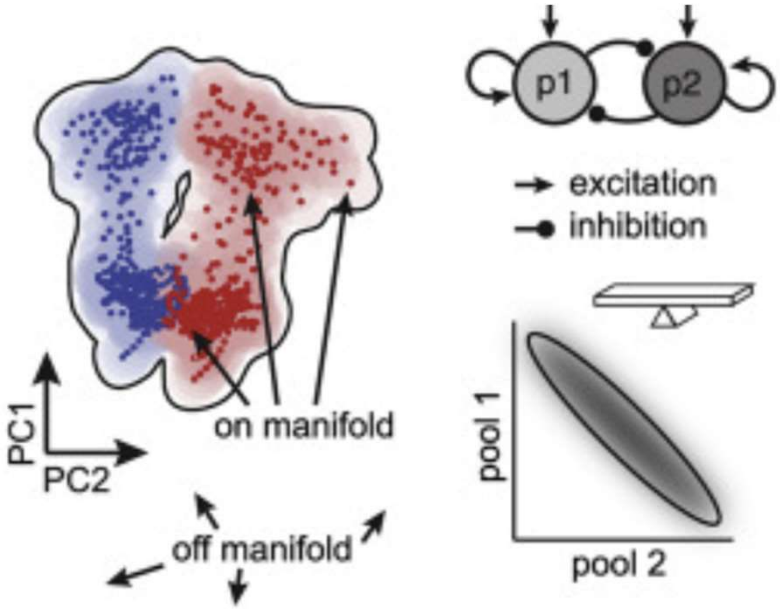
The concept of « neural manifold »

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.

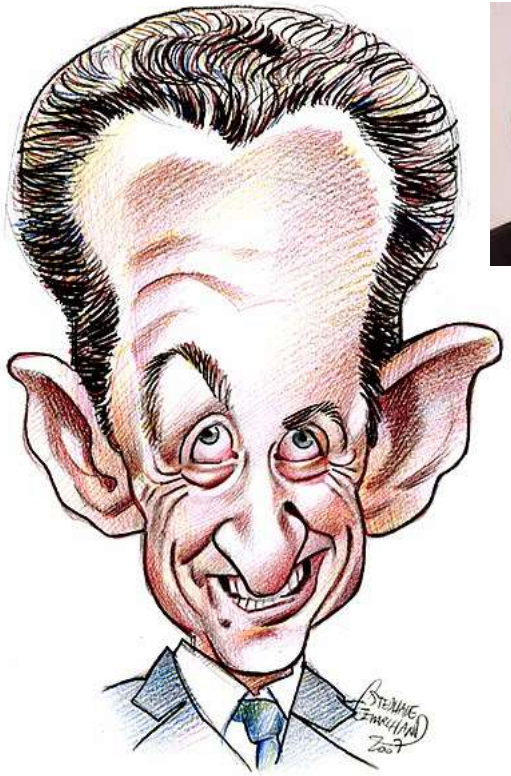
To put it another way, there is a lot of white space in our state space diagrams: neural activity tends to occupy fewer neural states than it would if each neuron made an independent, random contribution to population activity.

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

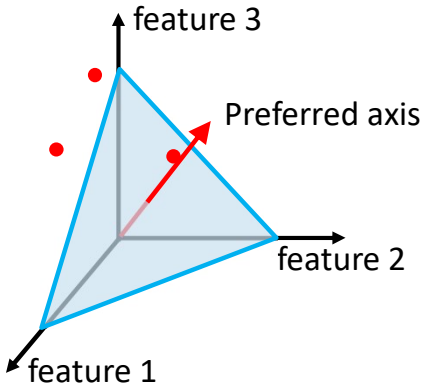


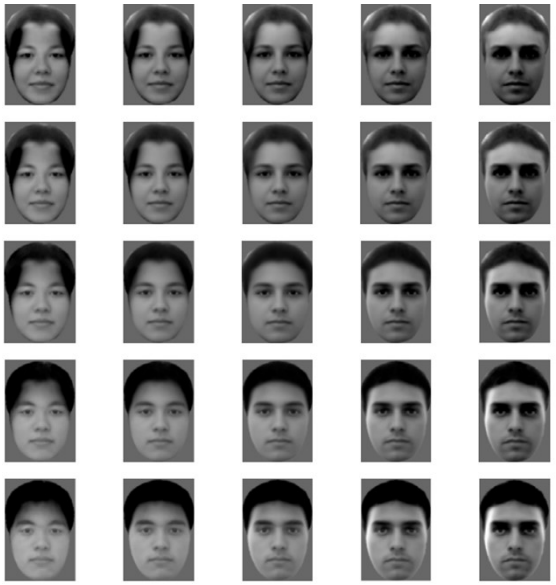
The neural representation of faces in inferotemporal face patches

A ~50-dimensional vector space, in which
each face is a vector
each neuron has a preferred axis

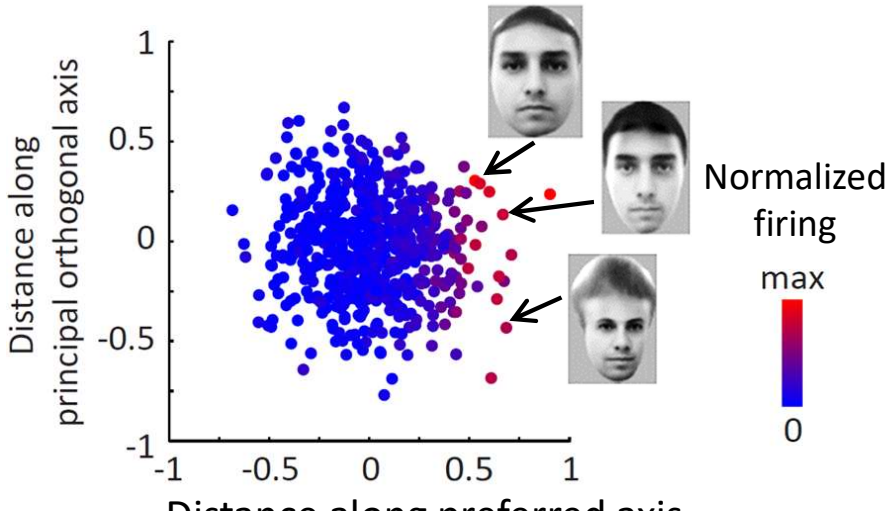


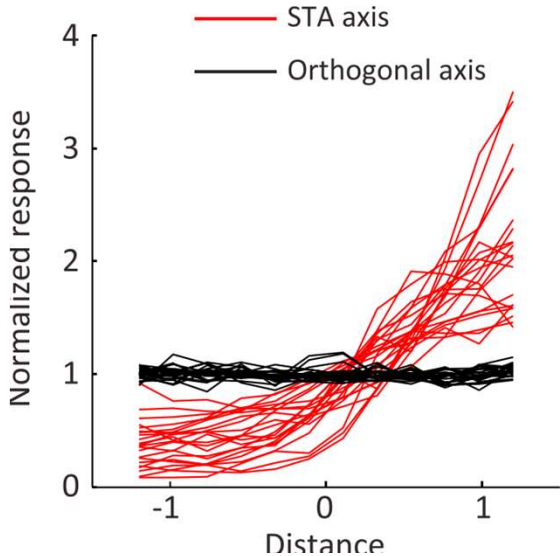
Response of each face neuron =
$s1 \cdot feature1 + s2 \cdot feature2 + \ldots s50 \cdot feature50$





Responses to 2000 faces





The brain seems to act as an auto-encoder that compresses incoming information

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

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.

The compressed, low-dimensional representation learned by a Beta-VAE provide an excellent fit to the neurons recording by Doris Tsao.

- *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.

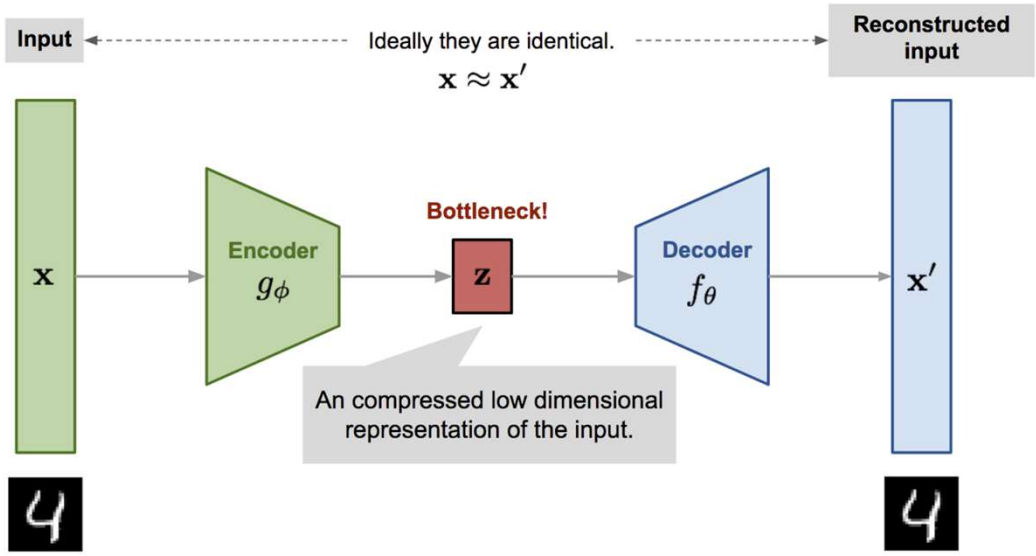


Fig. 1. Illustration of autoencoder model architecture.

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

A few questions raised by the audience

1. **Isn't is surprising that *monkey* face neurons respond to *human* faces?**

The monkeys that were tested have considerable experience with human faces.

- Is there an innate component to face recognition ? Yes: Conspec/Conlearn + responses to faces in newborns.

2. Do face vectors activate only when we perceive a face, or also when we think of a face, or imagine one?

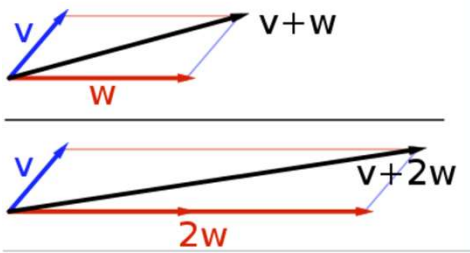
Imagination has been studied in humans (e.g. by Nancy Kanwisher) and clearly suffices to activate the FFA.

Cf Arcaro, M. J., Ponce, C., & Livingstone, M. (2020). The neurons that mistook a hat for a face. *Elife*, *9*, e53798.

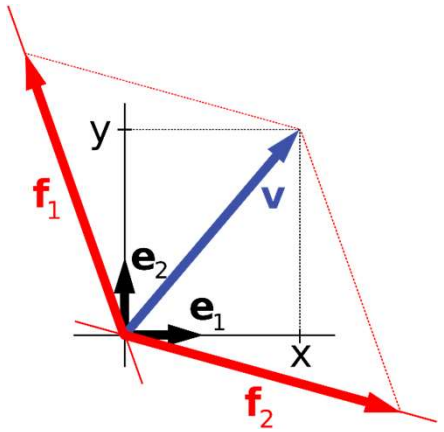
Also the FFA activity correlates with subjective perception, for instance during binocular rivalry.

A mini-glossary of some key concepts of vector spaces

Concept	Definition	Application in neuroscience

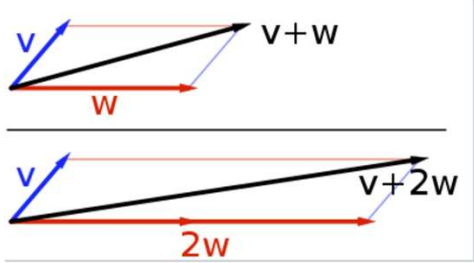


Vector addition and scalar multiplication: a vector **v** (blue) is added to another vector **w** (red, upper illustration). Below, **w** is stretched by a factor of 2, yielding the sum **v** + 2**w**.

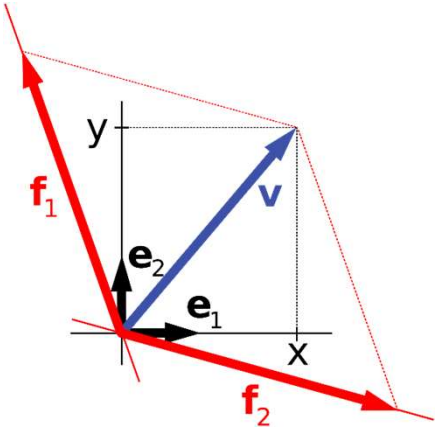


A mini-glossary of some key concepts of vector spaces

Concept	Definition	Application in neuroscience
Vector space	A vector space is a set of "vectors" with two operations: - **Addition** of two vectors to yield a third one - **Multiplication** by a scalar (a number) See https://en.wikipedia.org/wiki/Vector_space for a full list of mathematical requirements	- Neural activity of n neurons = vector in n dimensions - Addition = **superposition** of two neural assemblies - Scalar multiplication = **amplification of activity** - Average = **prototype**
Vector subspace	Each vector can be expressed by a **linear combination** of **base vectors**: $\vec{V} = \sum_{i=1}^{n} \alpha_i \, \vec{v}_i$ A subset of vectors $\vec{v}_i, \mathrm{i} = 1 \ldots \mathrm{m}$ define, by their linear combinations, a **vector subspace** of the original space.	A vector subspace can be used to encode - a **subset of the total information** (e.g. just color) - only the information that will be **communicated downstream**
Dimension	The number of base vectors in the main space (n). It can contain a subspace of lower dimension (m). For a manifold, intrinsic dimensionality is the number of coordinates needed to specify a position (e.g. 2 for a torus or a folded surface)	n = total number of neurons in the population considered m = size of a given coding subspace
Orthogonality	Vector spaces can be endowed with an **inner product (dot product or scalar product)** $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$ $\mathbf{x} \cdot \mathbf{y} = \cos(\angle(\mathbf{x}, \mathbf{y})) \cdot \lvert\mathbf{x}\rvert \cdot \lvert\mathbf{y}\rvert$	Independent encoding and decoding of multiple dimensions of a given stimulus



Vector addition and scalar multiplication: a vector $\mathbf{v}$ (blue) is added to another vector $\mathbf{w}$ (red, upper illustration). Below, $\mathbf{w}$ is stretched by a factor of 2, yielding the sum $\mathbf{v} + 2\mathbf{w}$.



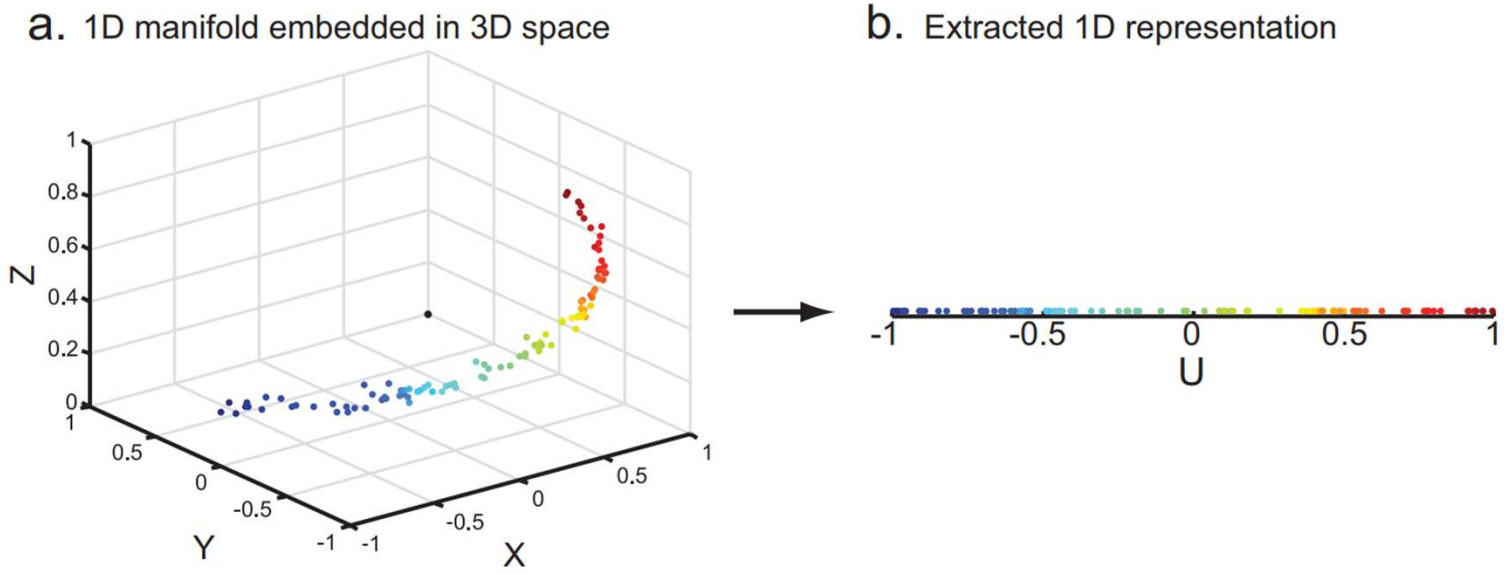
Measuring the dimensionality of a neural representation

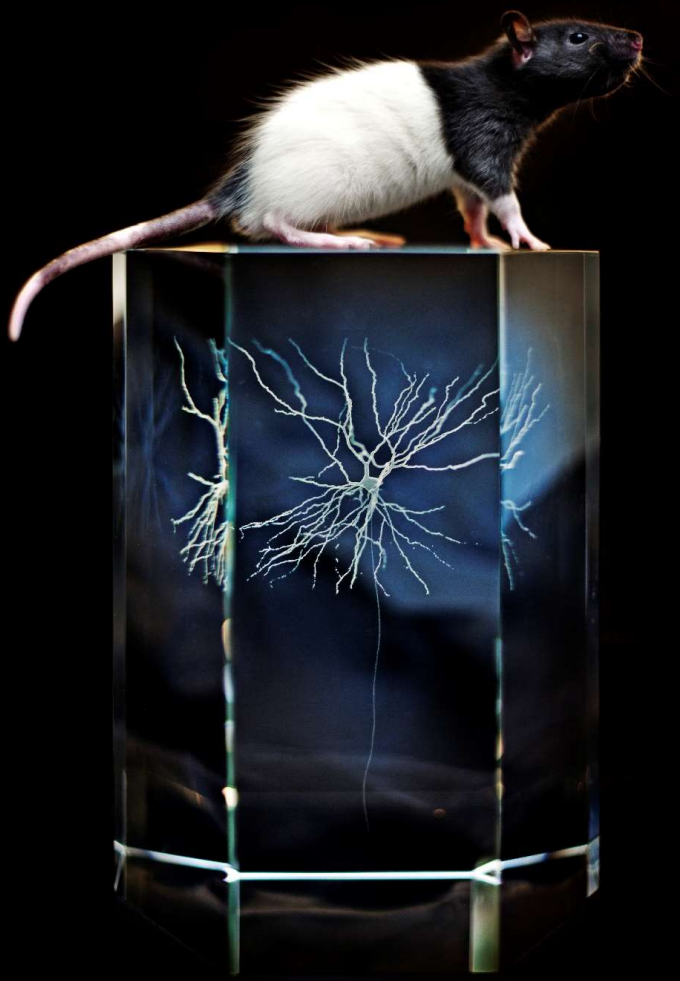
Lehky, S. R., Sereno, M. E., & Sereno, A. B. (2013). Population Coding and the Labeling Problem : Extrinsic Versus Intrinsic Representations. Neural Computation, 25(9), 2235-2264. https://doi.org/10.1162/NECO_a_00486
Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality (p. 2022.07.13.499969). bioRxiv. https://doi.org/10.1101/2022.07.13.499969

The responses of multiple neurons are generally not independent of each other.
→ The multi-neuron data "lives" in a smaller space than the full n-dimensional space of n neurons.
Here for instance:

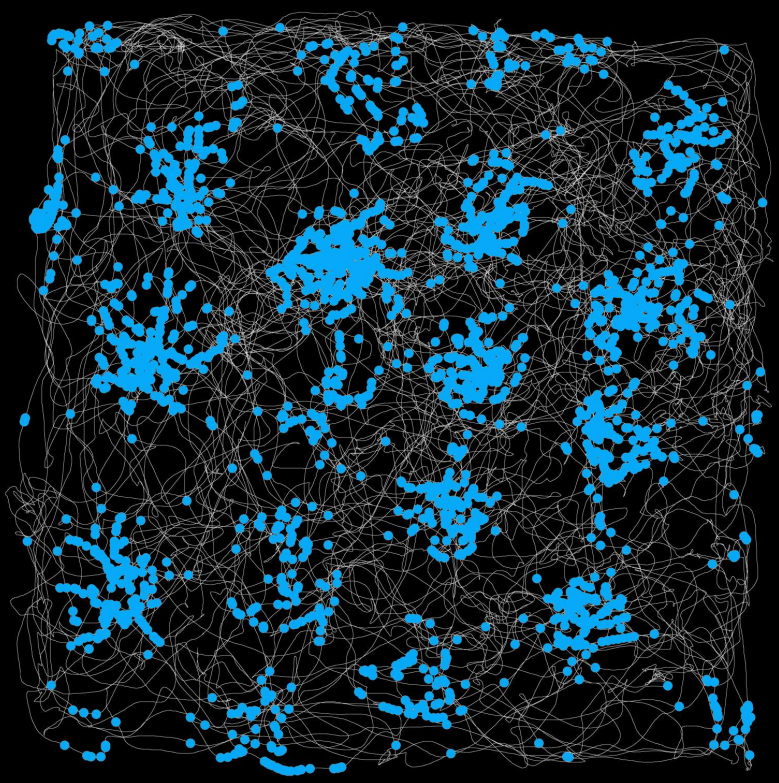
- The original neural space is 3-dimensional
- The data occupy a plane in that space (global dimension D=2)
- The data actually lie on a 1-D manifold (local dimension D=1)

Various techniques are available to "unfold" a manifold.
E.g. Multidimensional scaling, Isomap, UMAP, tSNE…





Grid cells in entorhinal cortex

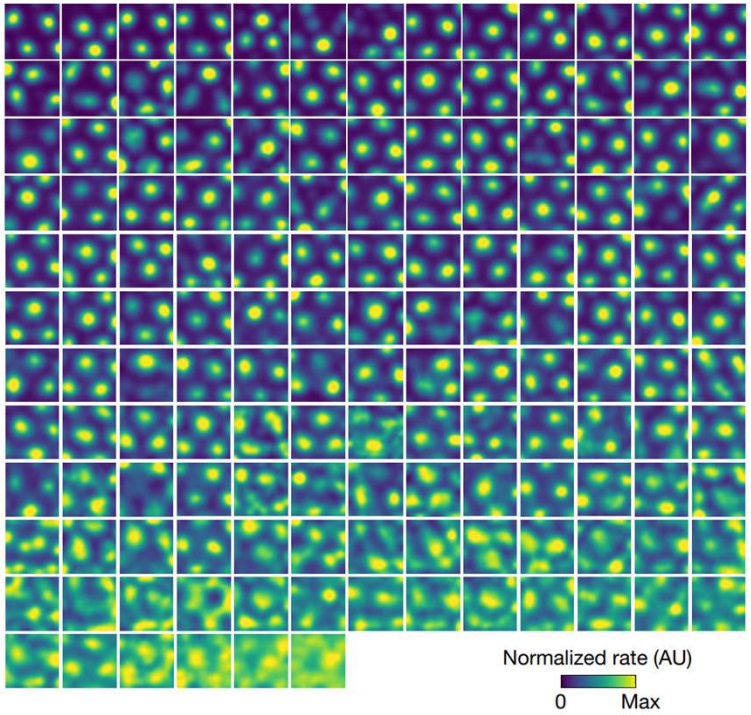


Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806. https://doi.org/10.1038/nature03721

From single grid cells to ensemble coding : A toroidal neural manifold

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., & Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, 602(7895), 123-128. https://doi.org/10.1038/s41586-021-04268-7

Recording from entorhinal cortex using NeuroPixels.
Massive increase in the number of simultaneously recorded cells.
Here, 2460 cells of which 483 are grid cells, of which 149 belong to the same "module" in entorhinal cortex.

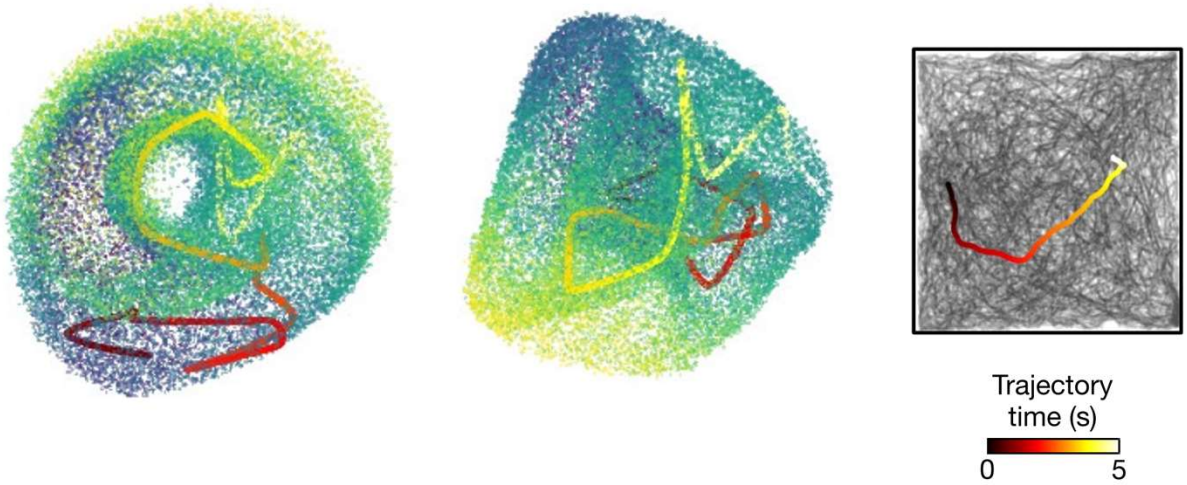


This paper is designed as a test of the "attractor network" hypothesis:

"The invariance of the correlation structure of this population code across environments and behavioural states, independent of specific sensory inputs, has pointed to intrinsic, recurrently connected continuous attractor networks (CANs) as a possible substrate of the grid pattern"

Dimensionality reduction: 3-dimensional embedding of the data, using a 6-dimensional PCA followed by UMAP (uniform manifold approximation and projection).

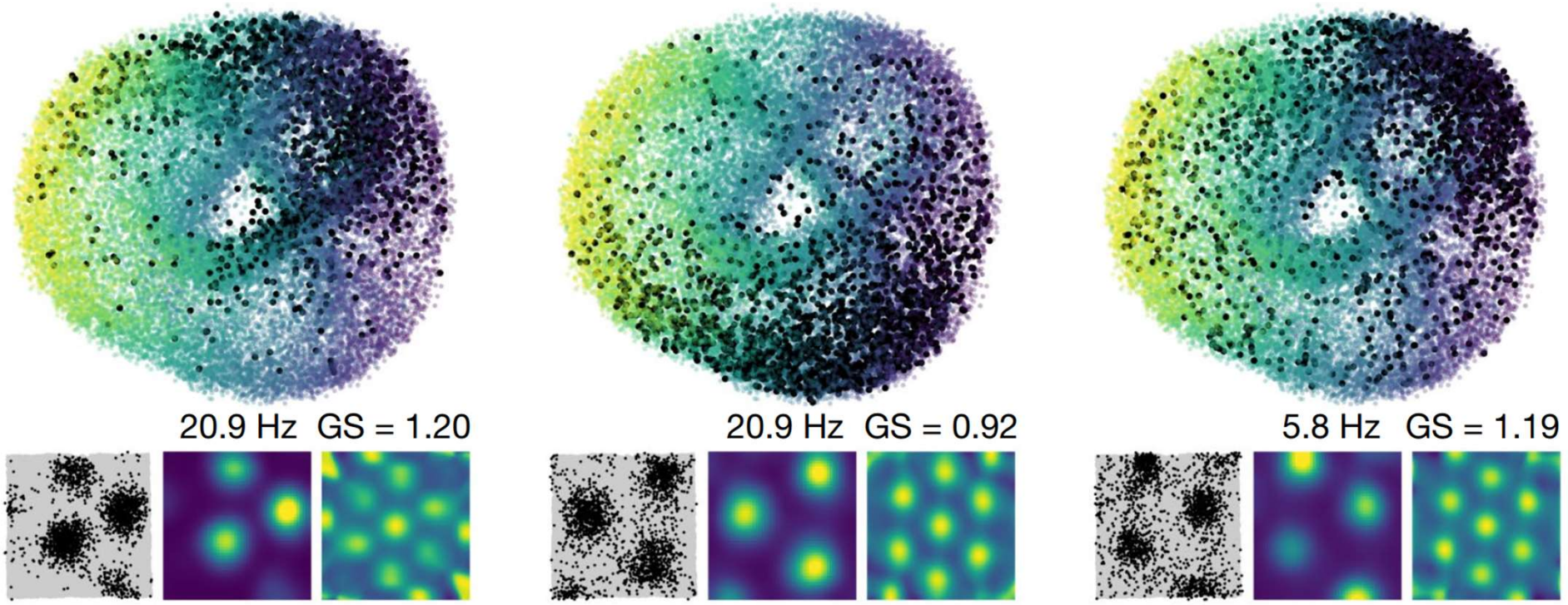
Result : a torus is visible with the naked eye!



From single grid cells to ensemble coding : A toroidal neural manifold

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., & Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, 602(7895), 123-128. https://doi.org/10.1038/s41586-021-04268-7

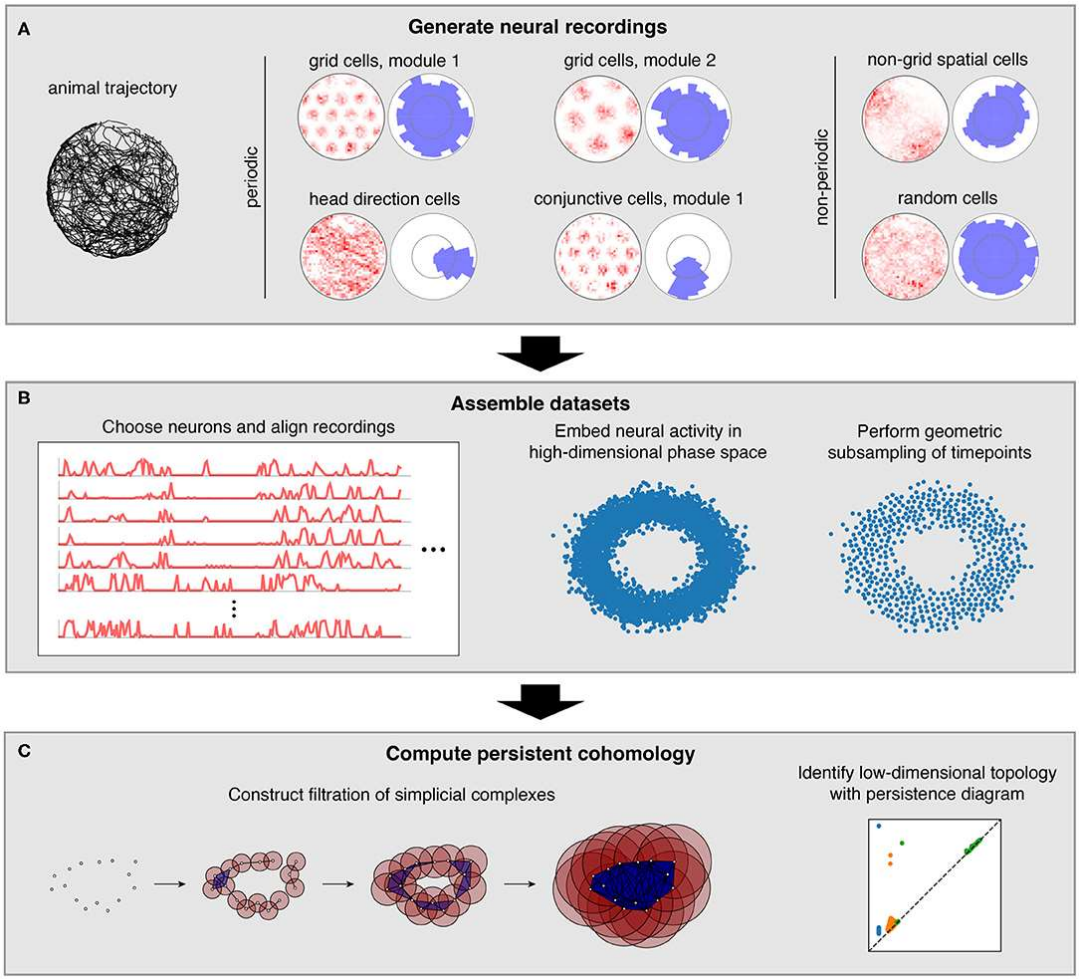
Firing of 3 grid cells, plotted on the torus – here when the animal is navigating a square room



Persistent (co)homology analysis can determine the shape of a neural manifold

Kang, L., Xu, B., & Morozov, D. (2021). Evaluating State Space Discovery by Persistent Cohomology in the Spatial Representation System. *Frontiers in Computational Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fncom.2021.616748

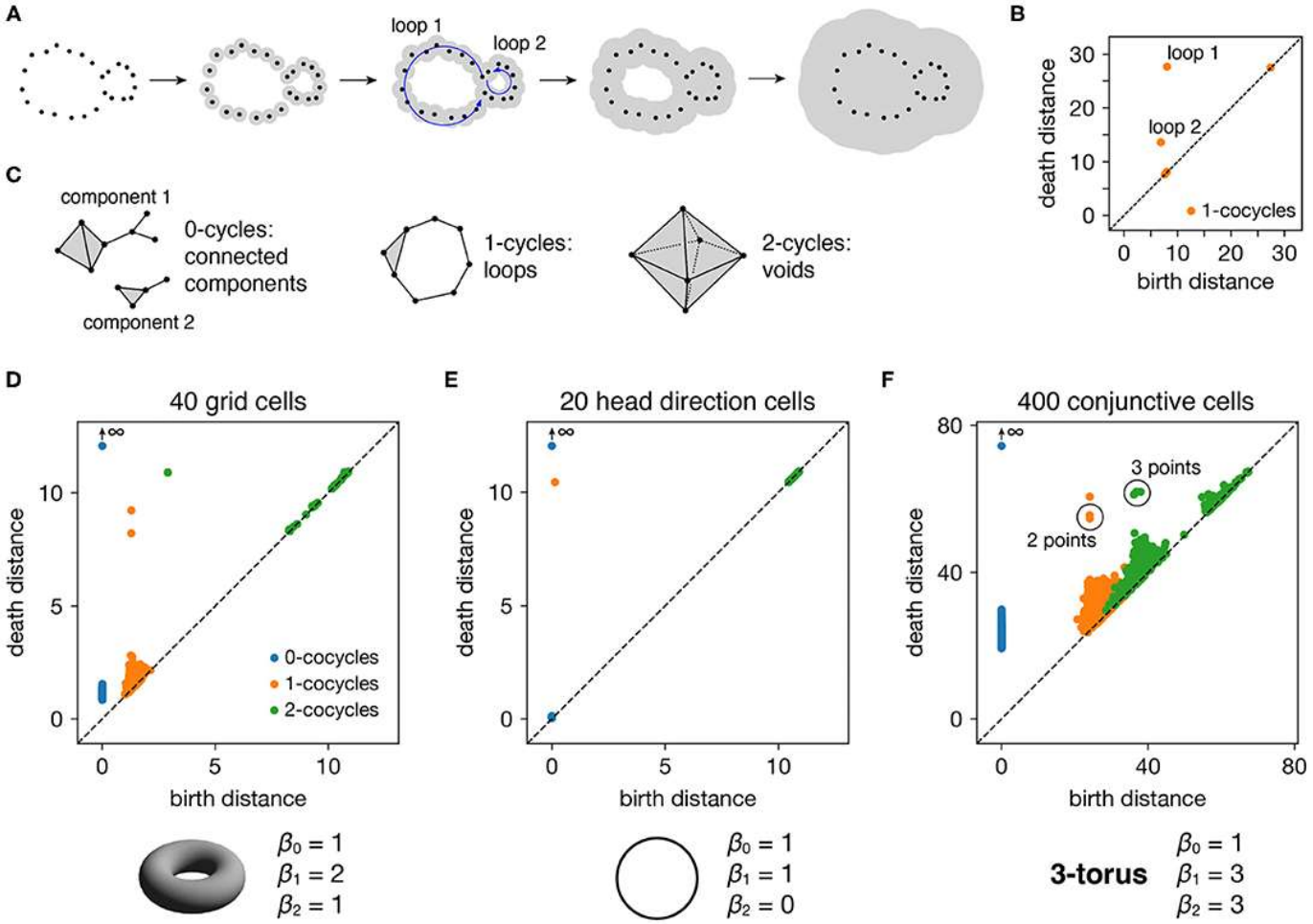
The torus interpretation can be confirmed by "persistent cohomology" analysis – "a method for computing topological features of a space at different spatial resolutions" (Wikipedia):



Persistent (co)homology analysis can determine the shape of a neural manifold

Kang, L., Xu, B., & Morozov, D. (2021). Evaluating State Space Discovery by Persistent Cohomology in the Spatial Representation System. *Frontiers in Computational Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fncom.2021.616748

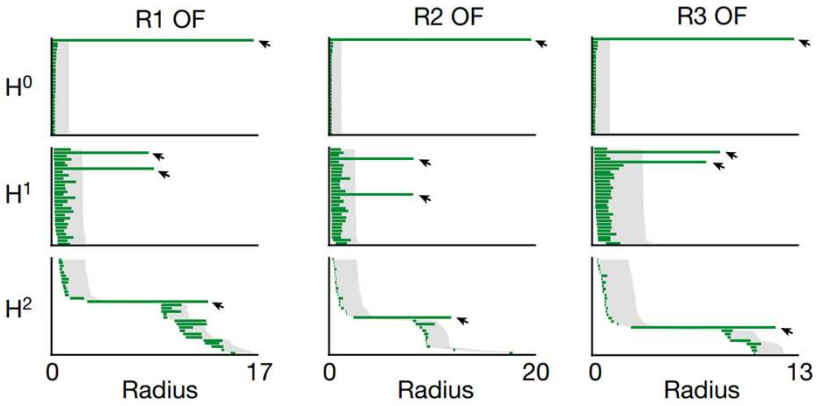
The torus interpretation can be confirmed by "persistent cohomology" analysis – "a method for computing topological features of a space at different spatial resolutions" (Wikipedia):



From single grid cells to ensemble coding : A toroidal neural manifold

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., & Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, 602(7895), 123-128. https://doi.org/10.1038/s41586-021-04268-7

The torus interpretation can be confirmed by "persistent cohomology" analysis – "a method for computing topological features of a space at different spatial resolutions" (Wikipedia):

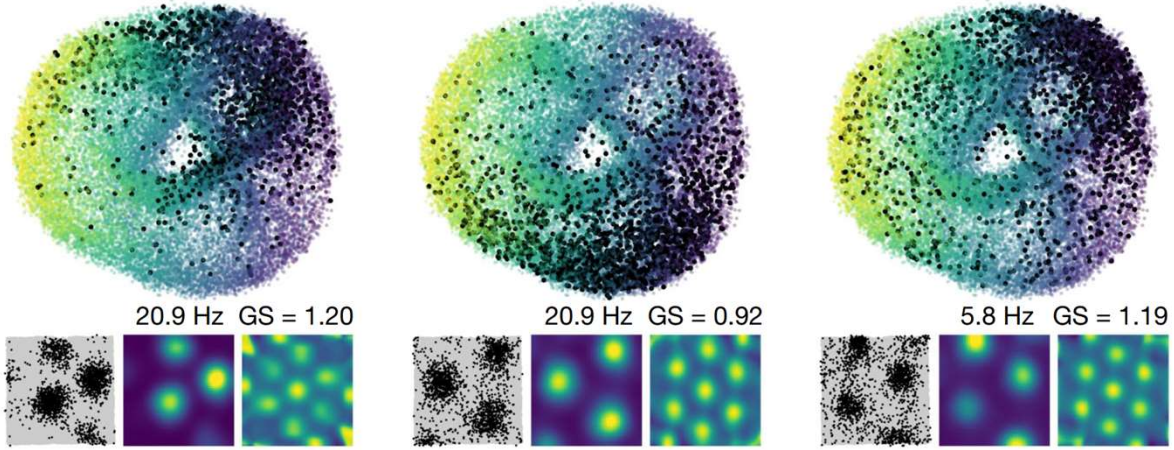


The cell firing is better predicted by its location on the torus than by the animal's physical location.
It is even present during sleep.
"the population activity in an individual grid-cell module resides on a toroidal manifold

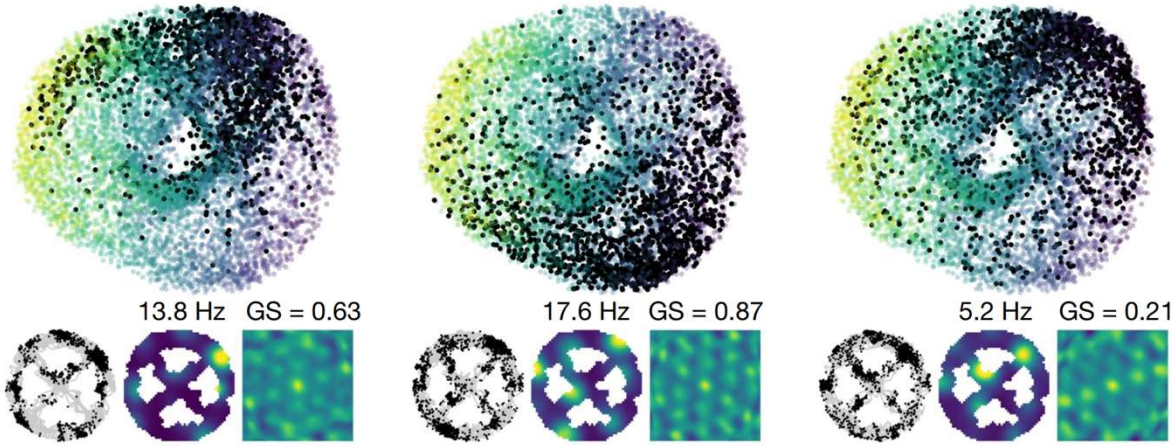
- independently of behavioural tasks and states
- and decoupled from the position of the animal in physical space."

The invariance of the correlation structure of this population code suggest an internal attractor.

Firing of 3 grid cells, plotted on the torus – here when the animal is navigating a square room



Firing of the same 3 grid cells when the animal is navigating a wheel-shaped track.



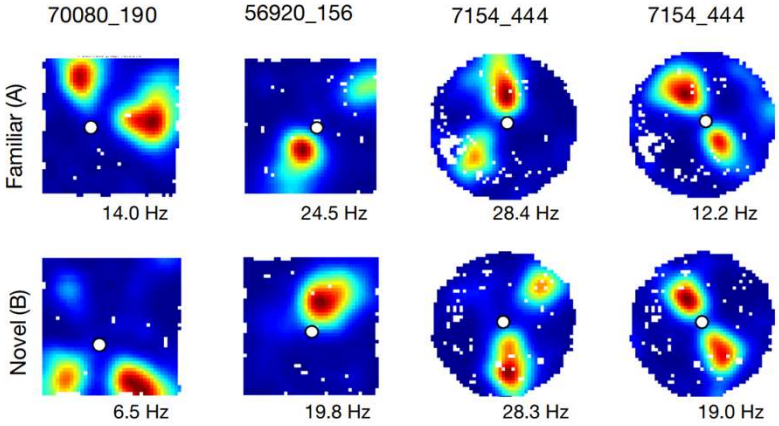
Object-vector cells: a very different form of spatial representation

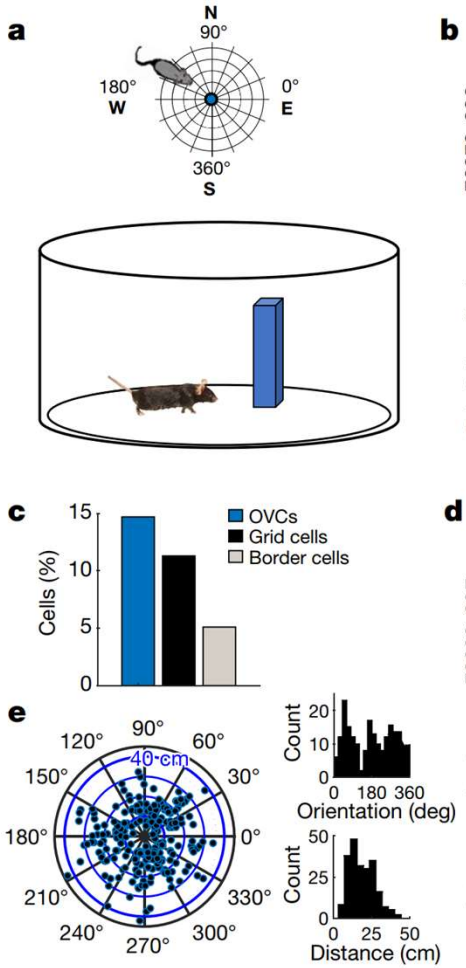
Høydal, Ø. A., Skytøen, E. R., Andersson, S. O., Moser, M.-B., & Moser, E. I. (2019). Object-vector coding in the medial entorhinal cortex. Nature, 568(7752), Art. 7752. https://doi.org/10.1038/s41586-019-1077-7

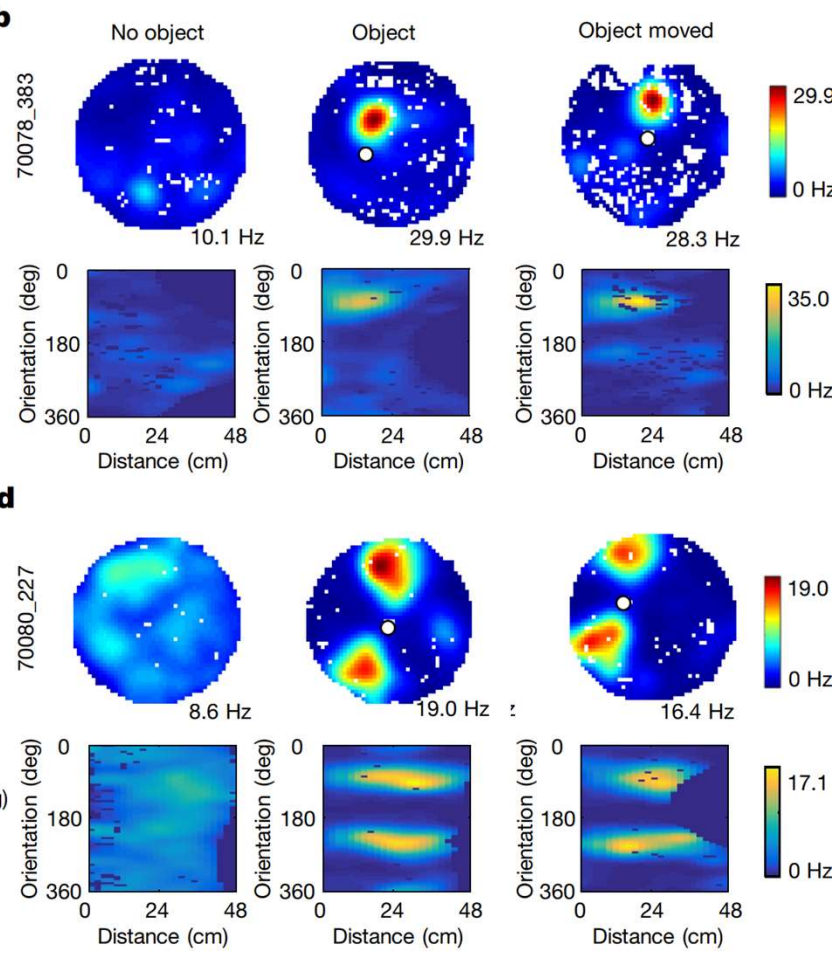
These cells fire only when the animal is at a certain distance and position relative to an object.

Unlike grid cells, they shift their firing when the object moves, but the room stays the same.

And conversely, they keep their object-vector properties regardless of the room or the identity of the object.

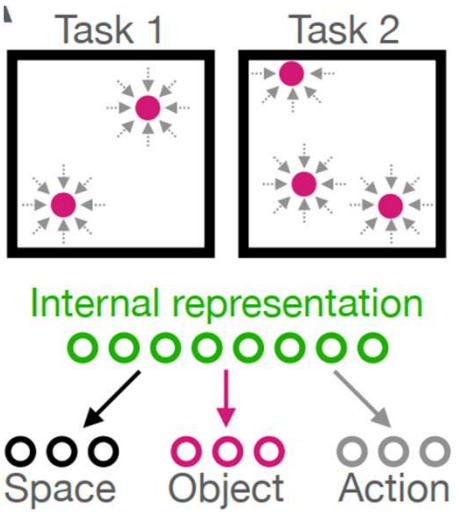






Why the separation of grid cells and object vector cells? Non-negative factorization can predict it

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. E. J. (2022). Disentangling with Biological Constraints : A Theory of Functional Cell Types (arXiv:2210.01768).

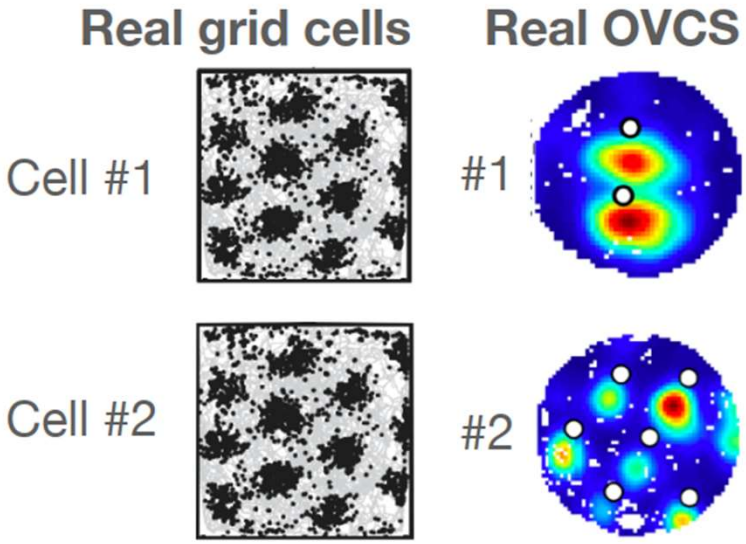


The authors train a network with a non-negative constraint (ReLu units) to perform a spatial navigation task with movable objects. Task = predict where the animal is (space), whether it will hit an object (object), and what its next action will be.
Result of the simulation : the network units separate into two “modules”:

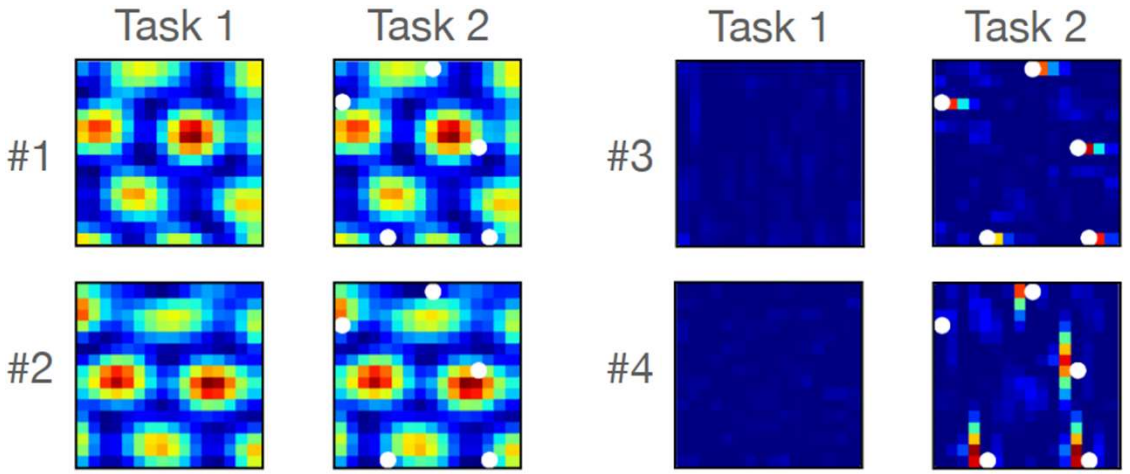
- Units that encode spatial position
- Units that encode position relative to the objects

These separate modules do not appear if the units are not ReLu, or if objects are fixed in space.

Real neurons



Simulation



Can we predict when dimensions stay entangled, and when they don't?

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. E. J. (2022). Disentangling with Biological Constraints : A Theory of Functional Cell Types (arXiv:2210.01768). arXiv. https://doi.org/10.48550/arXiv.2210.01768

"we mathematically prove that simple **biological constraints** on neurons, namely **nonnegativity and energy efficiency** in both activity and weights, promote such sought after disentangled representations by enforcing neurons to become selective for single factors of task variation."

"We demonstrate these constraints lead to disentangling in a variety of tasks and architectures, including variational autoencoders. "

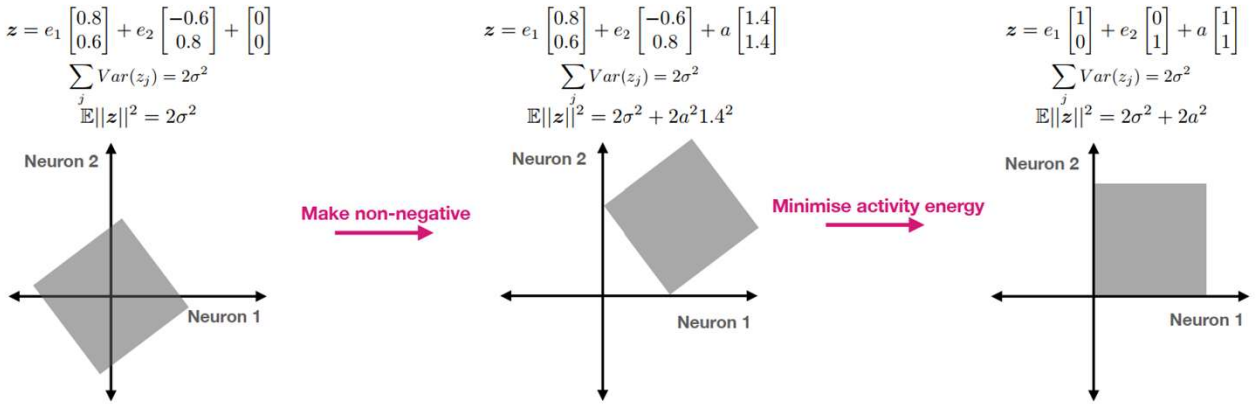
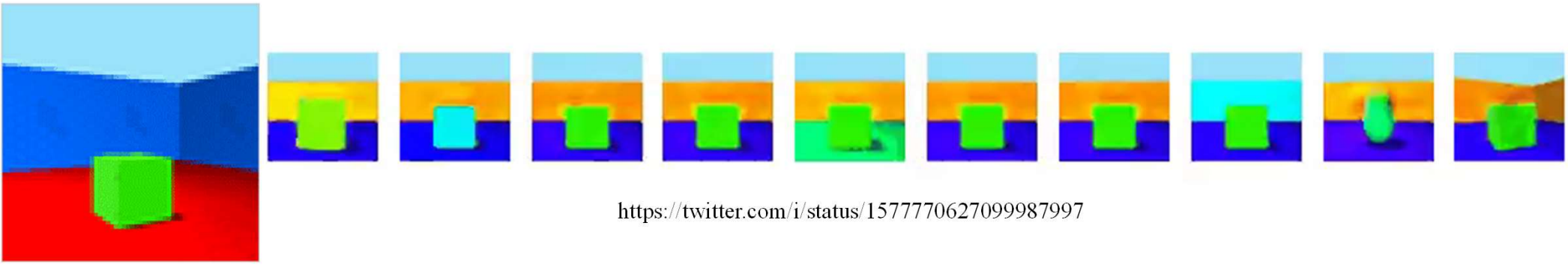


Figure 1: **Proof intuition.** Two uniformly distributed independent factors represented with two entangled neurons (left). The representation can be made nonnegative at the expense of activity

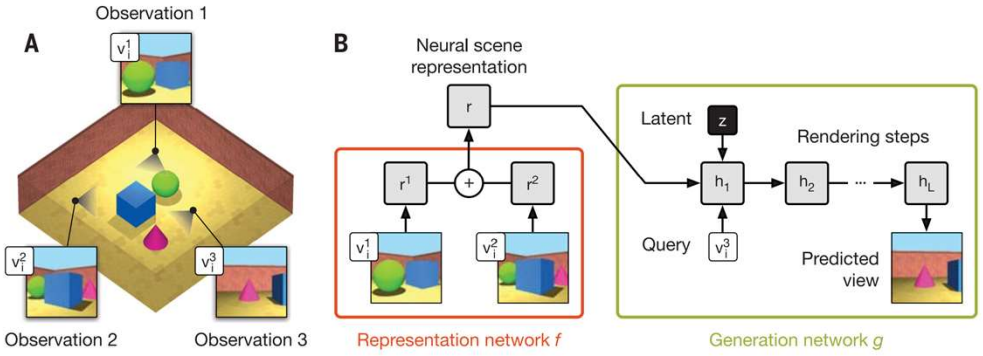
Example: Training of a beta-VAE on the Shape3D data set : individual units capture unique dimensions of variation in the data set.

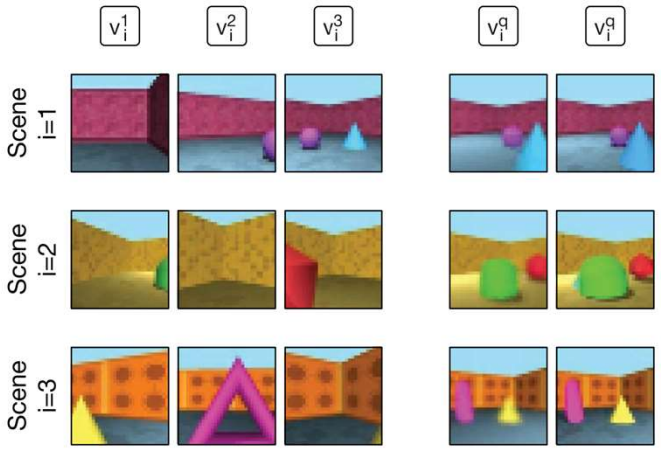


https://twitter.com/i/status/1577770627099987997

Factorized representations and vector arithmetic

Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., Reichert, D. P., Buesing, L., Weber, T., Vinyals, O., Rosenbaum, D., Rabinowitz, N., King, H., Hillier, C., Botvinick, M., … Hassabis, D. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204-1210. https://doi.org/10.1126/science.aar6170





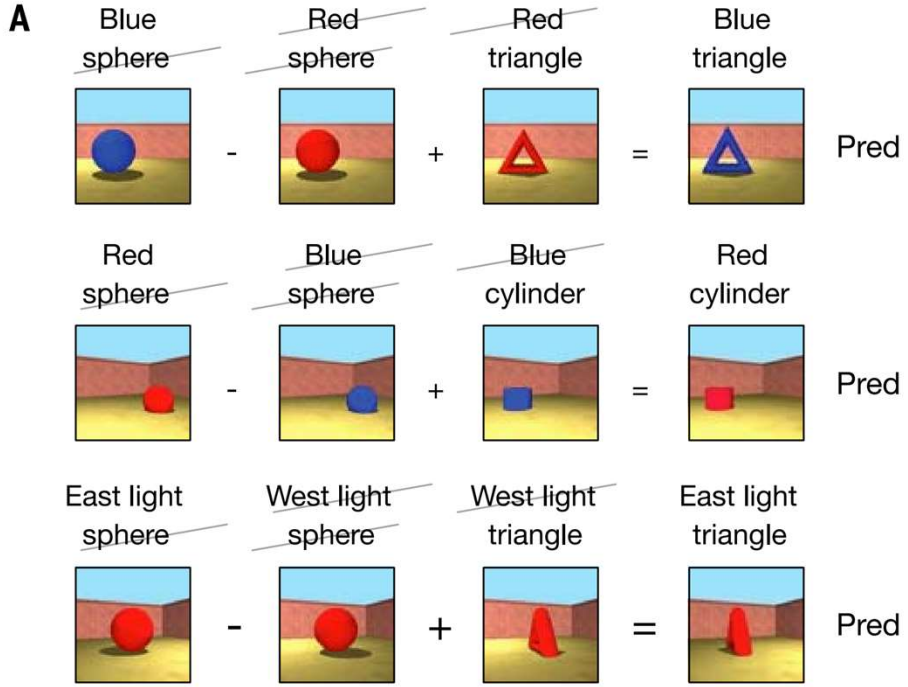
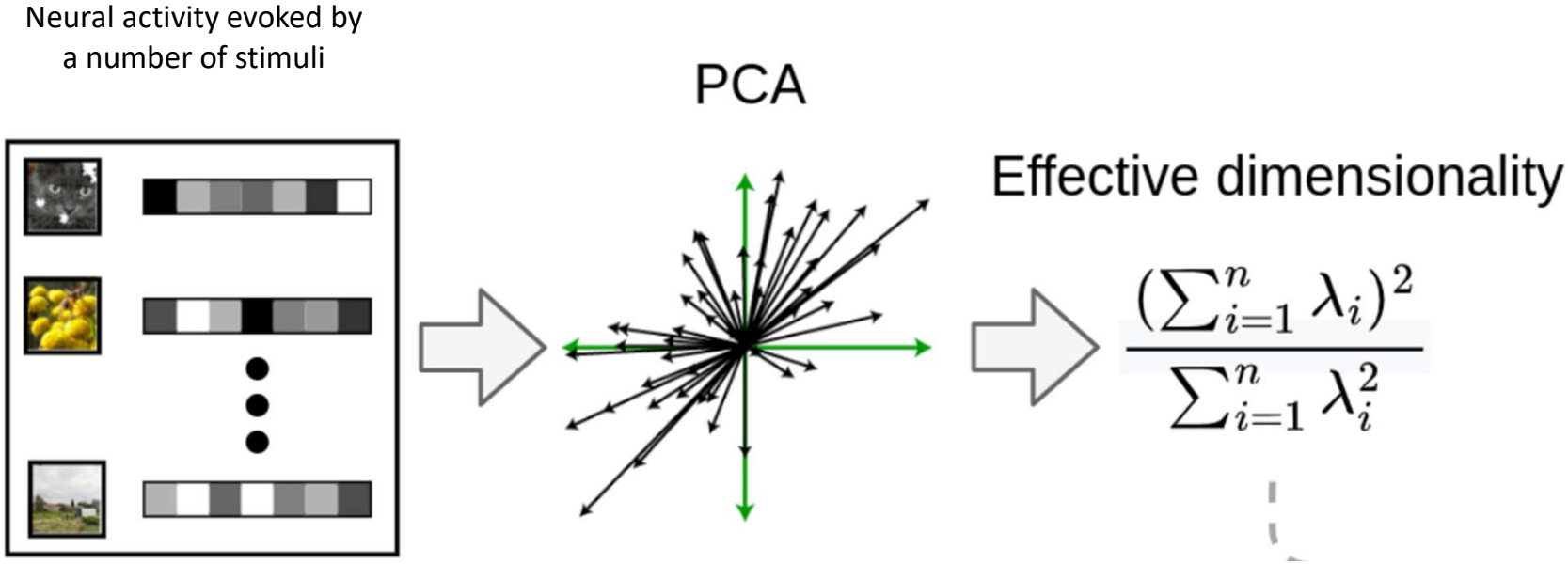


Fig. 4. Scene algebra and Bayesian surprise. (**A**) Adding and subtracting representations of related scenes enables control of object and scene properties via "scene algebra" and indicates factorization of shapes, colors, and positions. Pred, prediction. (**B**) Bayesian surprise at a new observation

Measuring the effective dimensionality of a neural representation

Lehky, S. R., Sereno, M. E., & Sereno, A. B. (2013). Population Coding and the Labeling Problem : Extrinsic Versus Intrinsic Representations. Neural Computation, 25(9), 2235-2264. https://doi.org/10.1162/NECO_a_00486
Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality (p. 2022.07.13.499969). bioRxiv. https://doi.org/10.1101/2022.07.13.499969

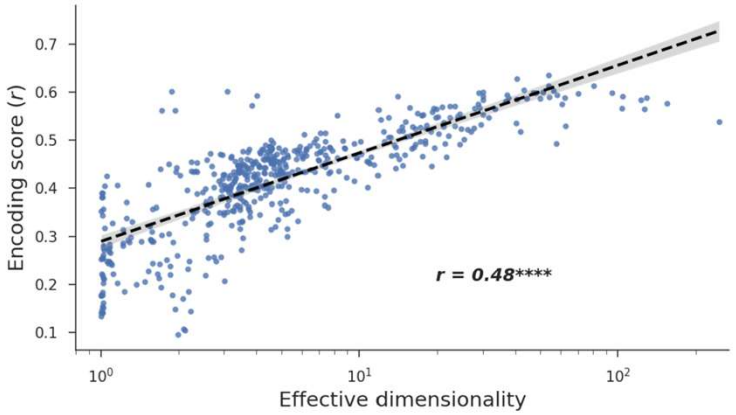
The “effective dimensionality” of a given neural representation (or CNN layer) can be measured by how fast its eigenvalues decrease.
Intuition behind this formula :
Imagine that m values are large (L), and the others negligibly small. Then ED = $(m\,L)^2 / (m\,L^2) = m$.
Thus the equation approximates the number of values that are larger than the others.



Measuring the effective dimensionality of a neural representation

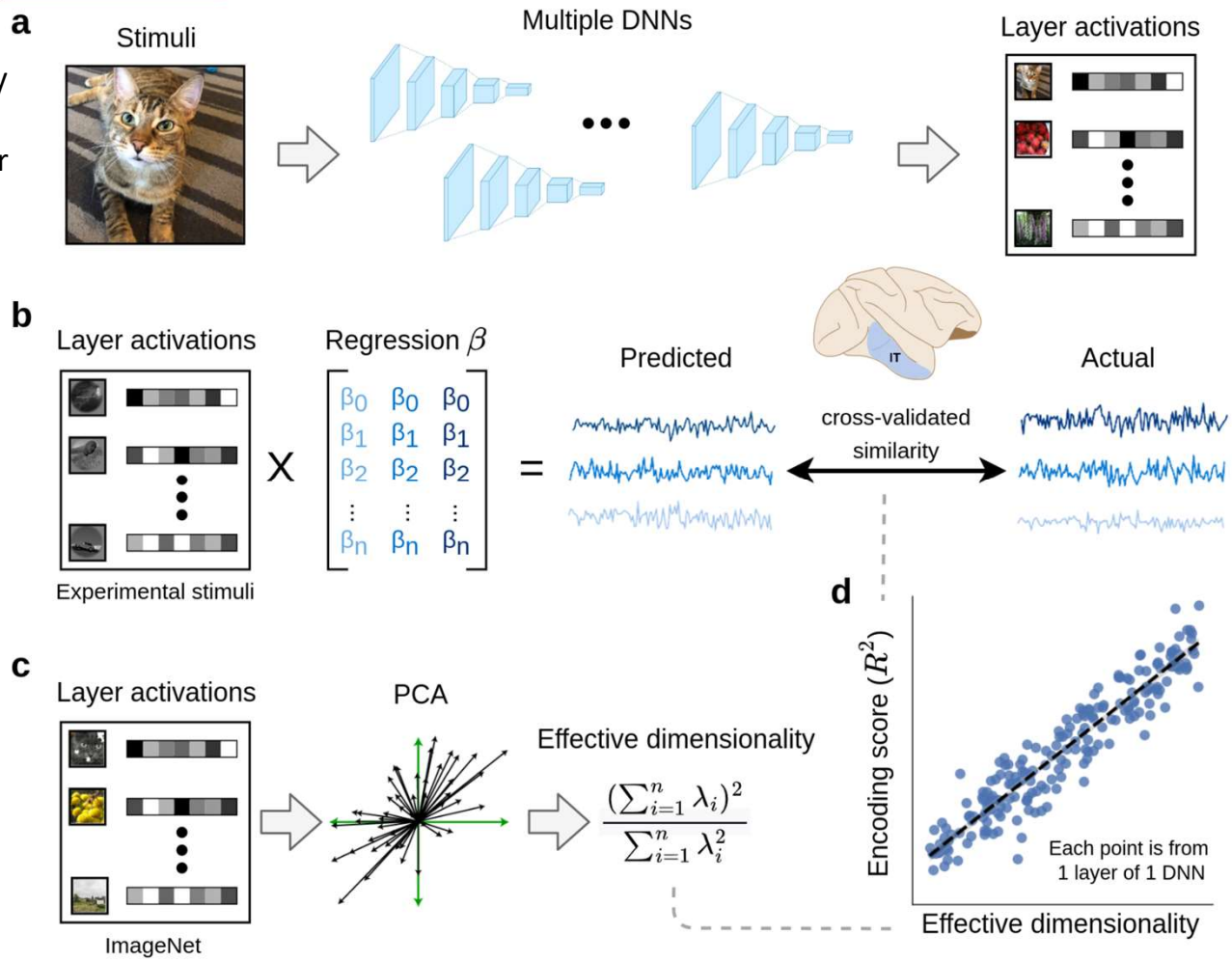
Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality (p. 2022.07.13.499969). bioRxiv. https://doi.org/10.1101/2022.07.13.499969

Measure the "effective dimensionality" of artificial neural networks (c,d) and compare it with how well they predict brain activit (b).
Networks with higher effective dimensionality are better predictors in forward modelling of IT activity.
This is also true when fitting a forward model for high-resolution fMRI data:



This is true:

- regardless of how the network is trained
- regardless of the actual size of the network (ie. for a full rank matrix, it is the dimensionality that counts) [because the R^2 is cross-validated]



Training increases the effective dimensionality of neural representations in CNNs

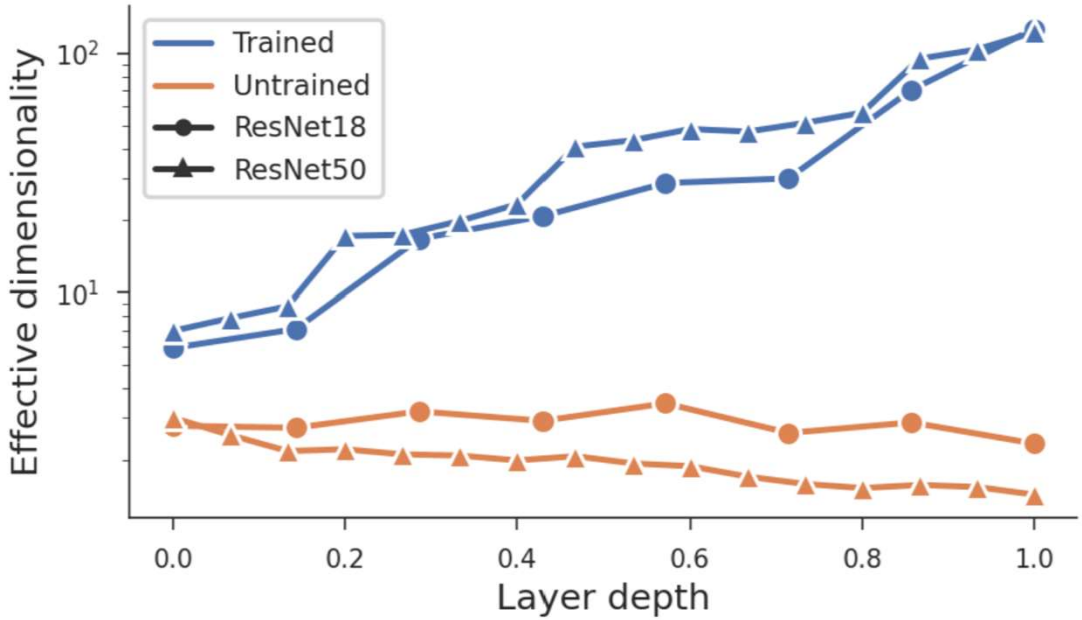
Elmoznino, E., & Bonner, M. F. (2022). High-performing neural network models of visual cortex benefit from high latent dimensionality (p. 2022.07.13.499969). bioRxiv. https://doi.org/10.1101/2022.07.13.499969

Effective dimensionality is primarily driven by learning:

It is low and decreases across successive layers for untrained networks.

It is higher and increases across successive layers when the same network is trained.

Note : dimensionality is computed after the max pooling operation of the CNN, otherwise dimensionality would be exceedingly large at lower layers.

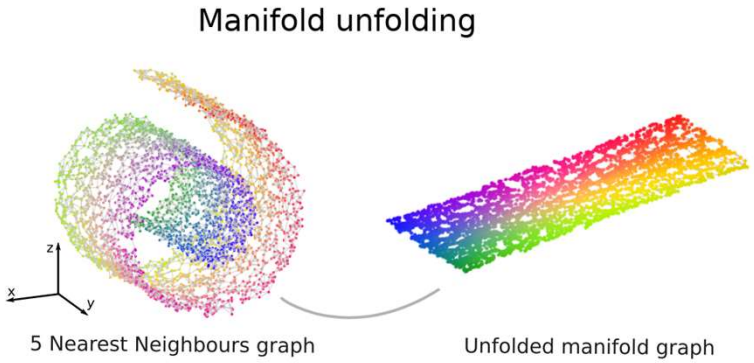


Does dimensionality only increase, or does it also decrease (compression) ?

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., & Shea-Brown, E. (2019). Dimensionality compression and expansion in Deep Neural Networks (arXiv:1906.00443). arXiv. https://doi.org/10.48550/arXiv.1906.00443

> "we apply state-of-the-art techniques for intrinsic dimensionality estimation to show that neural networks learn low-dimensional manifolds in two phases: first, dimensionality expansion driven by feature generation in initial layers, and second, dimensionality compression driven by the selection of task-relevant features in later layers."

This conclusion seems radically different from the above – but this is because they measure dimensionality differently: they attempt to estimate the "local" dimension of the manifold.



Global dimension (3) vs local dimension (2)

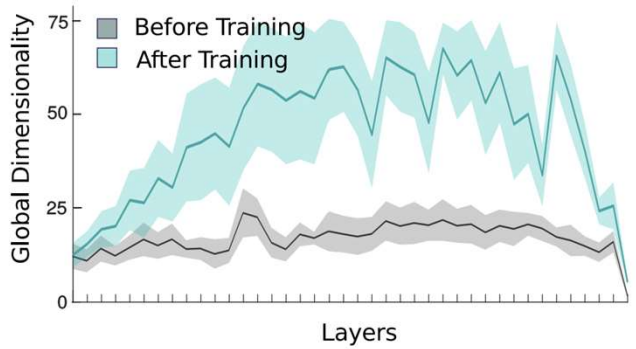
These dimensions can be estimated by evaluating the growth, as a function of a local radius *r*, of the number of neighboring points
(intuitively, for instance, this number should grow as r^2 for a surface)

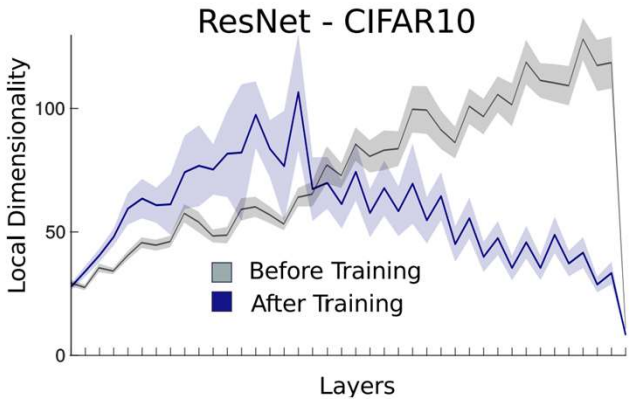
Findings :

Global dimension increases massively with training, as found by Elmoznino et al.

Local dimension, however, increases and then decreases:

- Expansion of the number of encoding features
- Compression to a small number of task-relevant features.





Fast learning of object categories

An example of very fast induction or « few-shot learning » (Tenenbaum, *Science*, 2011):
The objects in red are « tufa ».

This paradigm is well captured by Bayesian induction : Bayes rule automatically selects the smallest branch of the similarity tree that is compatible with all observations.

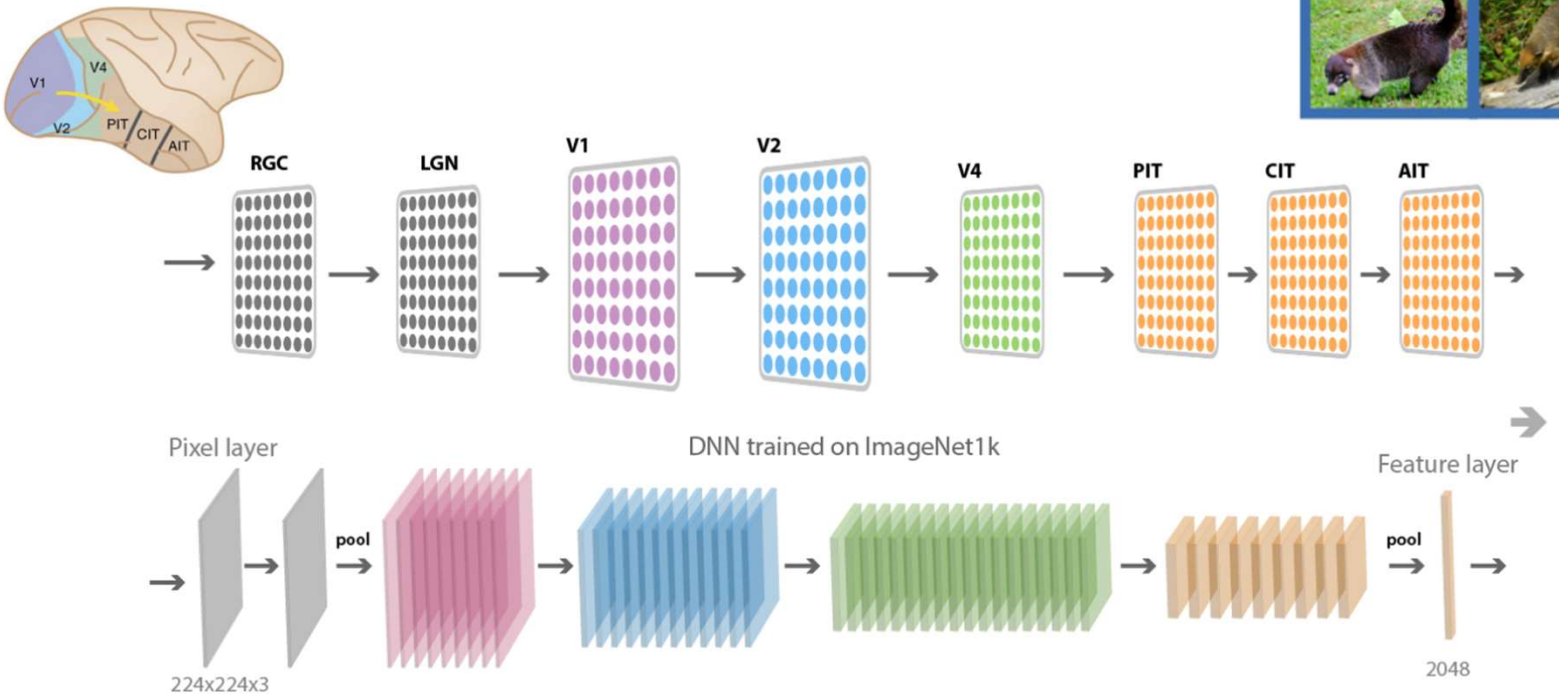
But how do subjects encode images in such a space?

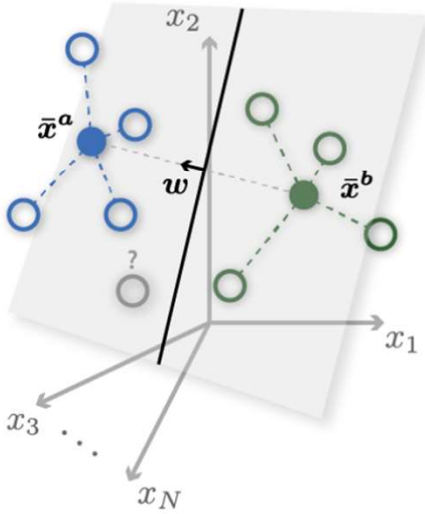
A theory of dimensionality and concept learning

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. PNAS, 119(43), e2200800119.

Here, the authors propose a general theory of “few shot learning” for image recognition.

1. Prior training has resulted in a tuned high-dimensional vector space for images, which can be used to perform one-shot or few-shot learning of new concepts.
2. Each example image (possibly 1) is encoded in this high-dimensional vector space
3. The **barycenter of examples** defines a **prototypical vector** for the new concept.
4. Classification of new images, or discrimination between two possibilities, is based on the **nearest prototype**





A theory of dimensionality and concept learning

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. PNAS, 119(43), e2200800119.

Empirical explorations of this scheme:

- Train networks on ImageNet (1000 image categories)
- Test on **binary classification** of all possible pairs of 1000 new images from ImageNet21k

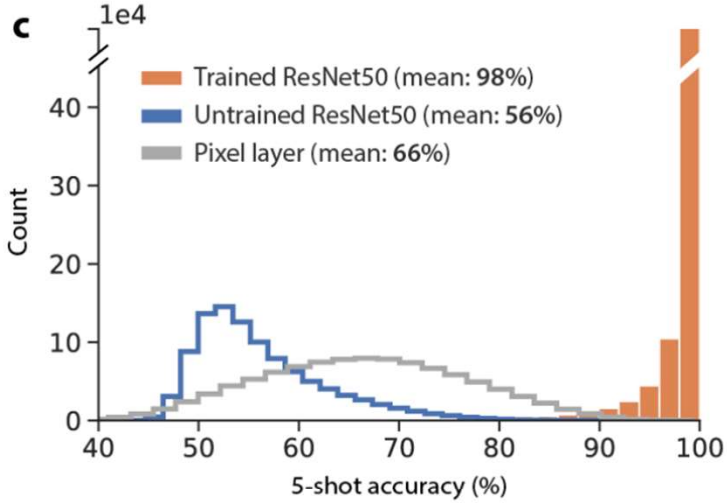
Results:

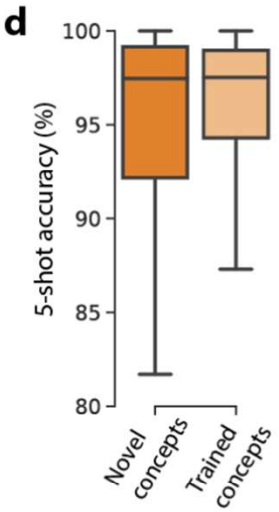
With just 5 examples, prototype learning manages to accurately classify new concepts, with an average of 98.6% correct ! (1-shot learning = 92 %)

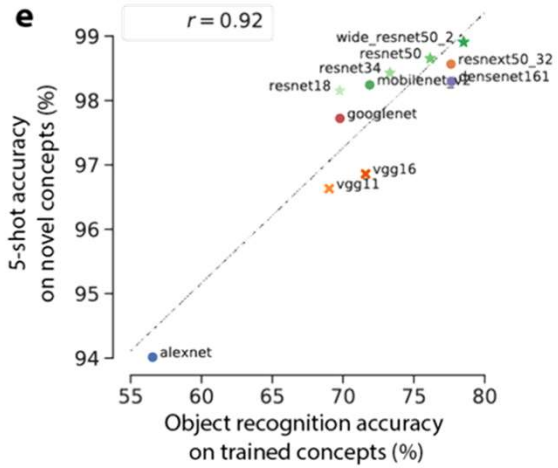
All sorts of trained networks work, and their performances are intercorrelated with each other.

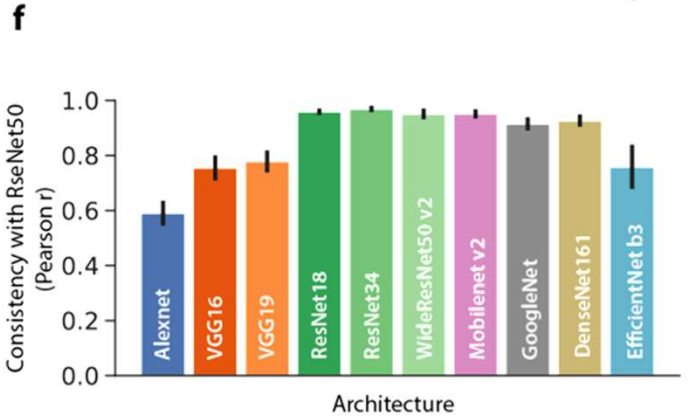
Untrained networks, however, do not perform well

→ the vector space must be tuned to pictures.









Geometry explains why some concepts are easier to discriminate than others

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. PNAS, 119(43), e2200800119.

$$\mathrm{SNR}_a = \frac{1}{2}\frac{\|\Delta \boldsymbol{x_0}\|^2 + (R_b^2 R_a^{-2} - 1)/m}{\sqrt{D_a^{-1}/m + \|\Delta \boldsymbol{x_0}\cdot \boldsymbol{U}_b\|^2/m + \|\Delta \boldsymbol{x_0}\cdot \boldsymbol{U}_a\|^2}}$$

In this vector space, the images for each new concept trace a manifold, which the authors approximate with a high-dimensional ellipsoid.

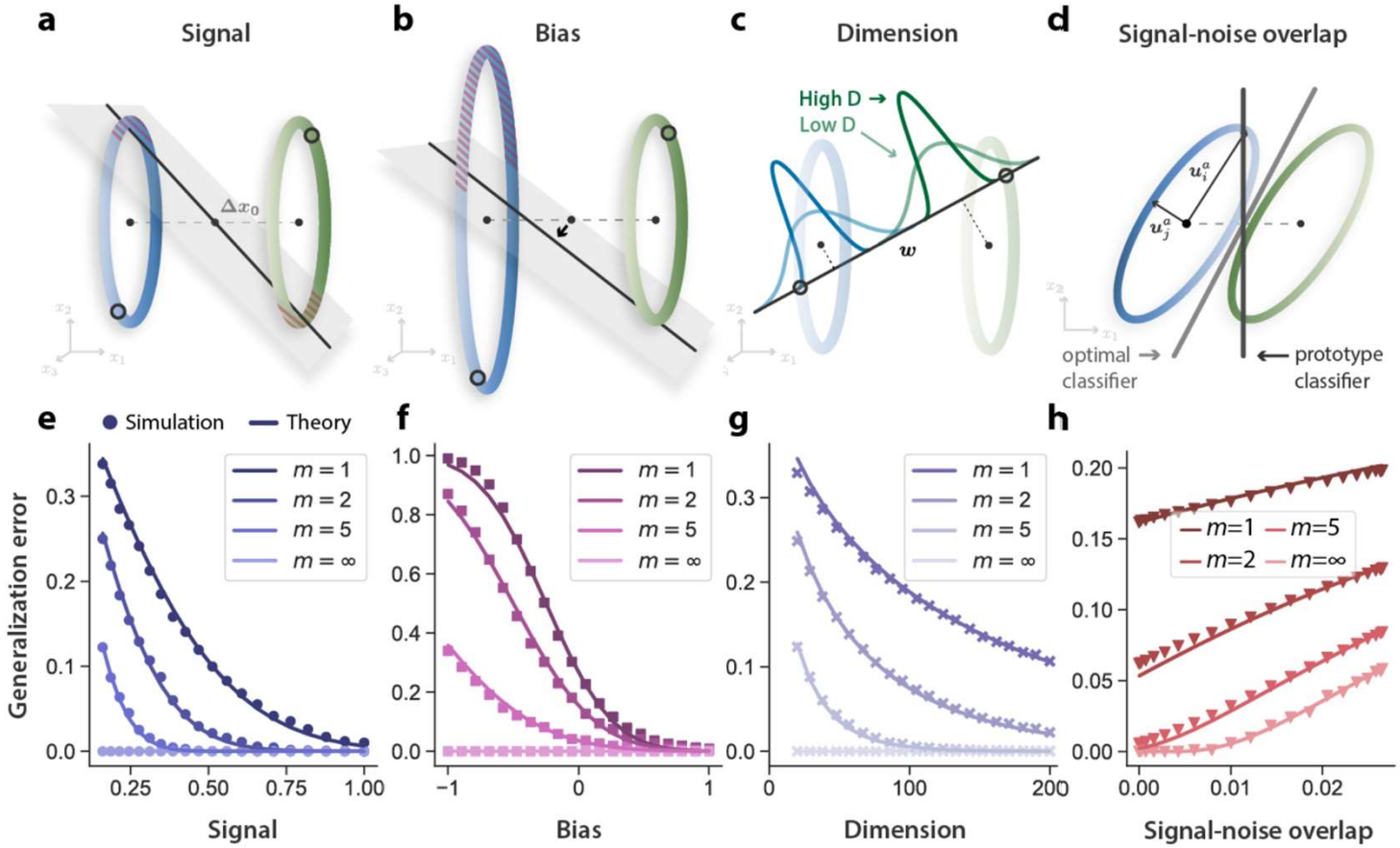
Four sources of errors in classification (m is number of training examples):

a. The **pairwise separation** between the two manifolds (signal) may be weak relative to the noise level.

b. One manifold may have a **larger variance** than the other, resulting in generalization errors and even worse-than chance performance.

c. The **dimensionality** of the manifolds may vary – and here, surprisingly, performance is better in higher dimensions (blessing of dimensionality).

d. **Noise** may vary in the same direction as the signal (the centroid separation vector). Note that here, the asymptote does not go to zero… unless a more optimal classifier is used.



The dark line shows a fit of the author's equation to simulations of ellipsoid categories

Conclusions

Considering neural responses as a **manifold**, a subspace within the huge space of potential neural responses, leads to interesting analyses:

- What is the **dimensionality** and **topology** of the neural manifold ? E.g. a torus for grid cells
- When are neural assemblies **disentangled**? During data compression, it may be advantageous to assign distinct neurons to distinct dimensions, e.g. grid cells versus object cells

In the case of faces or objects in IT cortex, the vector view leads to the following conclusions

- A **low-dimensional representation** emerges with **learning** (e.g. 50 for faces): data compression
- It can support **very fast encoding of new categories**
- The decision boundaries can explain the psychophysics of **conscious and unconscious decision making**

Next week

- Dynamics of decision making in vector spaces
- How to use vector spaces to communicate between areas

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)