NX-414: Brain-like computation and intelligence

Martin Schrimpf

Lecture 2, February 26th

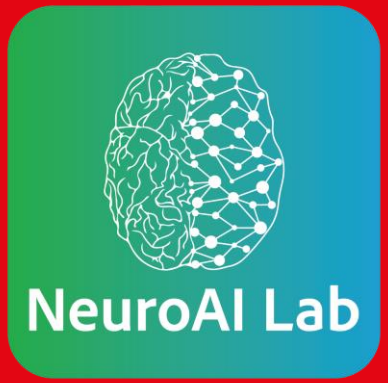
Slide credit:
- Created by A Mathis
- 2025: Modified by M Schrimpf

EPFL

Brief Intro

- Assistant Professor at EPFL since 6/2023
- Educational background in CS/ML
- PhD in neuro department

No formal office hours, but please email me if forum/TAs cannot help with a question.



go.epfl.ch/NeuroAI

Research group on computational digital-twin models of the brain.

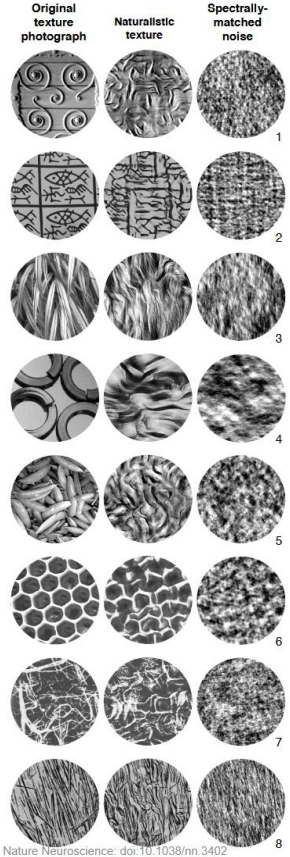
Focus on vision and language.

Co-running the Brain-Score platform.

Learning objectives today

- Normative models as a way to describe neural activity
- Population vectors to interpret neural activity
- Classic models and analyses
  - Hebbian learning, PCA, Oja's rule
  - Gabor filters
  - Sparse coding

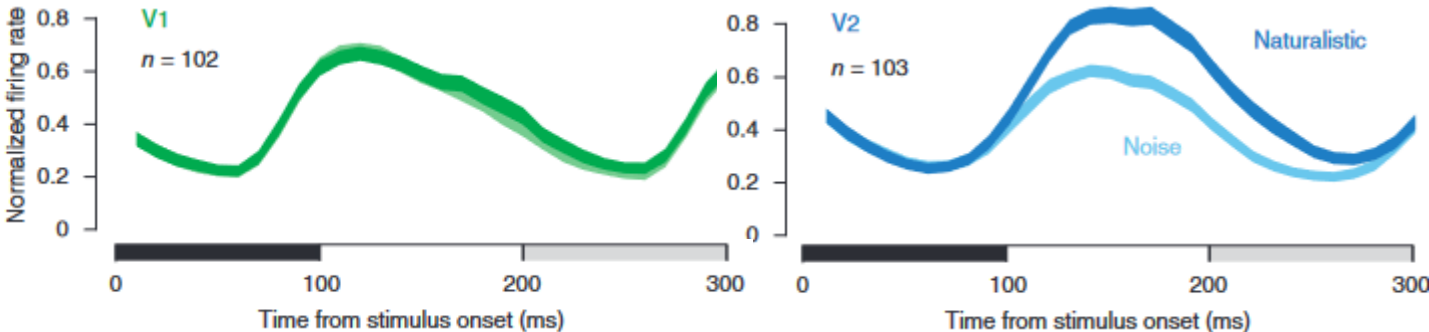
What is the right language to describe the brain's mechanisms?



A functional and perceptual signature of the second visual area in primates

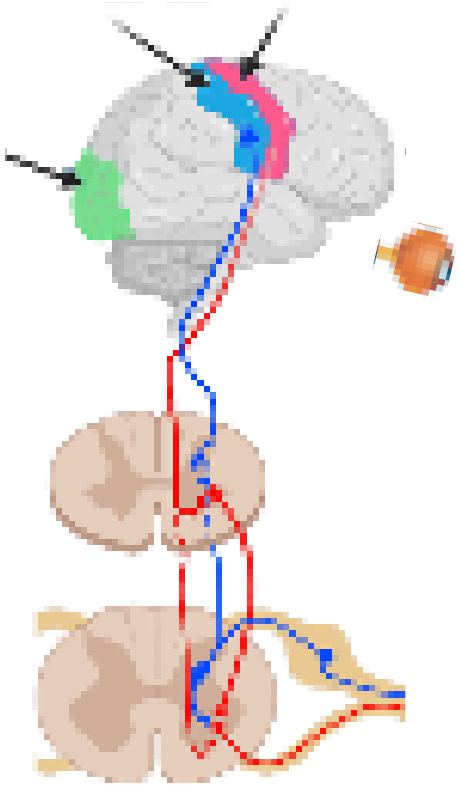
Jeremy Freeman[1,5,7], Corey M Ziemba[1,5], David J Heeger[1,2], Eero P Simoncelli[1–4,6] & J Anthony Movshon[1,2,6]

There is no generally accepted account of the function of the second visual cortical area (V2), partly because no simple response properties robustly distinguish V2 neurons from those in primary visual cortex (V1). We constructed synthetic stimuli replicating the higher-order statistical dependencies found in natural texture images and used them to stimulate macaque V1 and V2 neurons. Most V2 cells responded more vigorously to these textures than to control stimuli lacking naturalistic structure; V1 cells did not. Functional magnetic resonance imaging (fMRI) measurements in humans revealed differences between V1 and V2 that paralleled the neuronal measurements. The ability of human observers to detect naturalistic structure in different types of texture was well predicted by the strength of neuronal and fMRI responses in V2 but not in V1. Together, these results reveal a particular functional role for V2 in the representation of natural image structure.



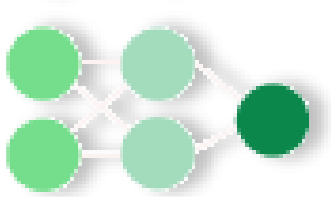
Nature Neuroscience: doi:10.1038/nn.3402

Normative frameworks



Information theoretic

e.g. sparse coding, redundancy reduction, mutual information …



Utilitarian

e.g. recognize objects, chase prey, navigate …

Reminder: Cramer-Rao inequality and Fisher information

For any biased estimator it holds that

$$\sigma_{est}^2(x) \geq \frac{(1 + b'_{est}(x))^2}{I(x)}$$

Note that for unbiased estimators, we have

$$\sigma_{est}^2(x) \geq \frac{1}{I(x)}$$

With Fisher information defined as

$$I(x) = \int p(k|x) \left(-\frac{\partial^2 \ln(p(k|x))}{\partial x^2} \right) dk$$

Well-known estimators/decoders

Stimulus x
e.g. image $\in \mathbb{R}^D$

Evoked response k
e.g. firing rate $\in \mathbb{R}^N$ for N neurons

Estimator: $x \mapsto k$
(also "encoder, predictor")
Decoder: $k \mapsto x$

Maximum likelihood estimator (MLE):

$$x_{MLE}(k) = argmax_x P(k|x)$$

Informed by only likelihood.
no "domain knowledge"

Maximum a posteriori (MAP) estimator:

$$x_{MAP}(k) = argmax_x P(k|x)P(x)$$

Informed by likelihood & prior
(P of stimulus being presented)

A simple decoder

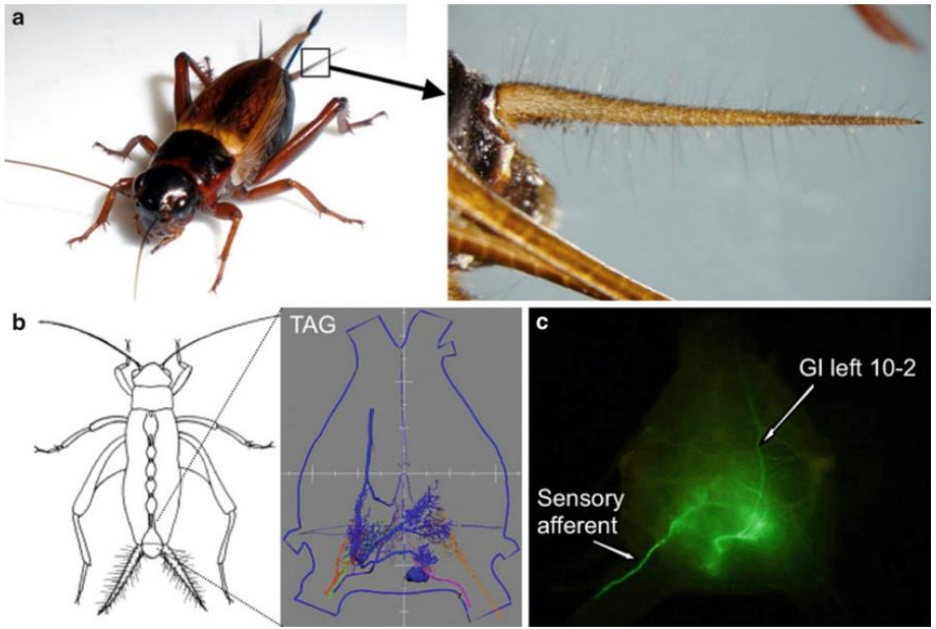
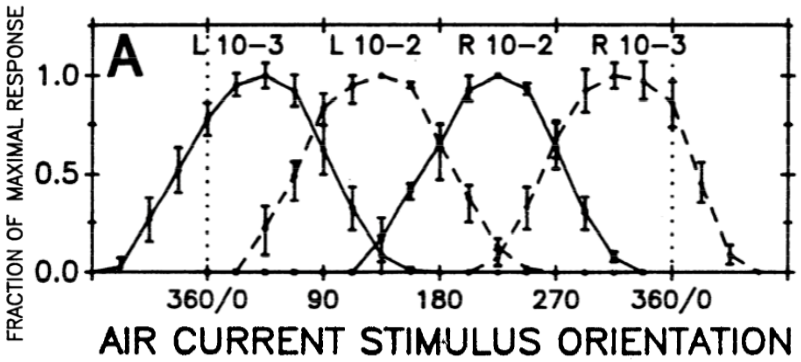


Fig. 1 The cricket cercal system. (**a**) *Gryllus bimaculatus* (female). The cerci are the two antenna-like structures, covered with fine hairs, extending from the rear of the abdomen. (**b**) Computer reconstruction of the outline (*blue*) of a terminal abdominal ganglion (TAG) with several reconstructed nerve cells. A reconstruction of a single identified giant interneuron (GI right 10-3) is shown in *blue*, and several filiform sensory afferent arbors are shown in other colors. (**c**) Micrograph of the GI left 10-2 and filiform afferent stained with fluorescent dye

Ogawa & Miller, Encyclopedia of Comp Neuro 2013

Normalized responses of the four giant interneurons



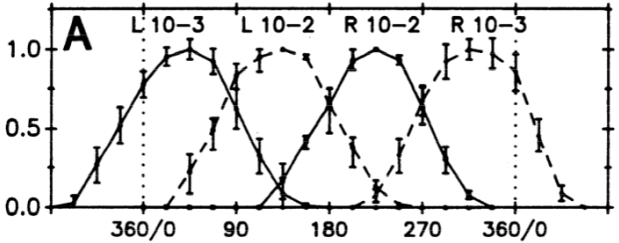
Each of the four giant interneurons codes for a particular air current direction

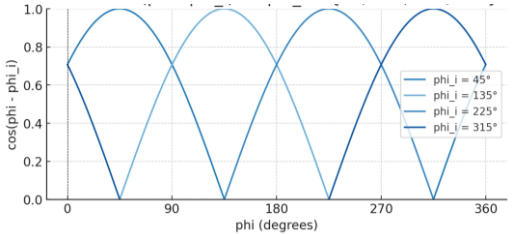
Miller, Jacobs, Theunissen, J Neurophysiology 1991

An example decoder: Population vectors

Those tuning curves can be summarized as a half-wave rectified cosine:

$$\left(\frac{f(\varphi)}{f_{max}}\right)_i = [cos(\varphi - \varphi_i)]_+$$





Let's represent the wind direction as a vector v instead of the angle. This is conveniently the same as:

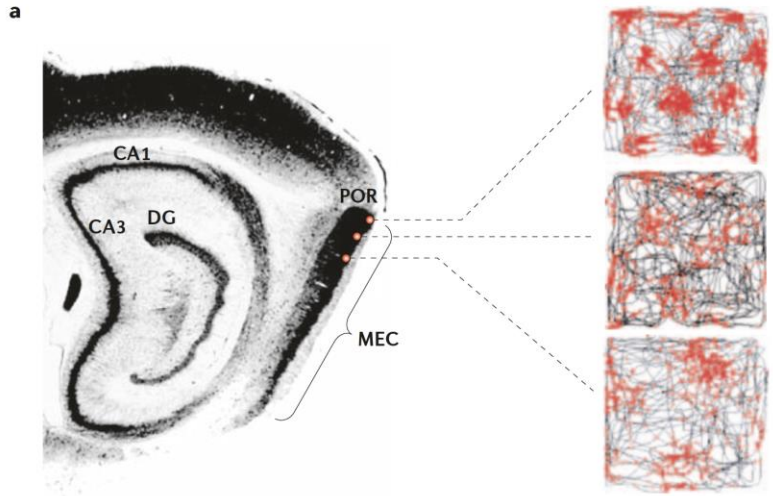
$$\left(\frac{f(\varphi)}{f_{max}}\right)_i = [v \cdot \mathrm{v}_i]_+$$

Given evoked responses from our 4 interneurons, we can then estimate the wind direction as:

$$v_{est} = \sum_i \left(\frac{k}{f_{max}}\right)_i \cdot \mathrm{v}_i$$

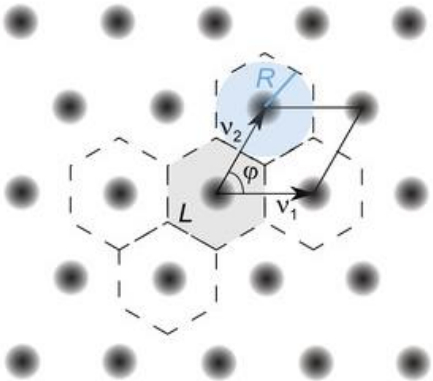
This is called the *population vector* and works quite well.

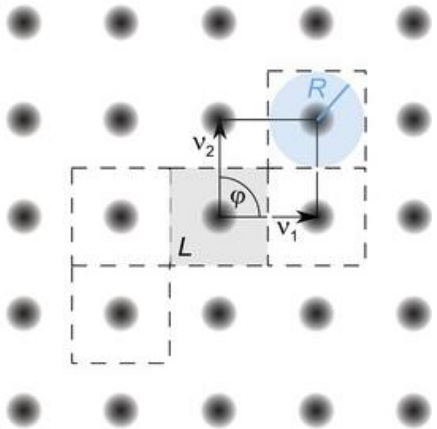
Hexagonal activity patterns of grid cells are optimal



McNaugthon et al., Nature Review Neuroscience 2006

Densest packing provides highest Fisher information.
Maximize spatial resolution/precision in 2D environments.





Mathis et al., eLife 2015

Limitations of using Fisher information to study the neural code

- This works best for parametric cases
  (i.e. with analytical formulas for the encoding model)
- Fisher information is local
- Only quantifies information of neural response, analysis only

Efficient coding hypothesis

The nervous system should exploit the statistical regularities in the sensory data

e.g., sparsity: instead of representing correlated features, encode decorrelated features.

e.g., predictive coding: instead of absolute information, encode only errors.

Barlow, 1961
Attneave 1954

Unsupervised learning with classic methods

Consider data $X = \{x^{(t)}\}_{1 \le t \le T}$ for vectors $x^{(t)} \in \mathbb{R}^D$ .We use superscript for enumerating them, as we want to use subscript x_i to denote the *i*th element. We will usually omit the superscript.

For a concrete example think, e.g., about flattened images of original dimension $\sqrt{D} \times \sqrt{D}$ for some square D for a large number of images T.

Here are and in the following $\langle \ . \ \rangle$ averages over the omitted superscript t. Assuming the data have zero mean, i.e. $\langle x_i \rangle = 0$, then:

$$\langle x_i^{(t)} \rangle = 1/T \sum_{t=1}^{T} x_i^{(t)}$$

Let’s assume there are pairwise correlations present, i.e.:

$$c_{ij} := \langle x_i, x_j \rangle \neq 0$$

Because the data have zero mean, $\langle x_i \rangle = 0$, this means that they are statistically dependent.

That’s because otherwise $\langle x_i, x_j \rangle = \langle x_i \rangle \langle x_j \rangle = 0$, which is not possible.

Reminder: Principal component analysis (PCA)

The goal of PCA is to find coordinates e_i for R^D such that:

$$y_i = e_i \cdot x \text{ and } \langle y_i, y_j \rangle = 0$$

By definition the coordinates are orthonormal:

$$\langle e_i, e_j \rangle = 0 \text{ if } i \neq j \text{ and } |e_i| = 1.$$

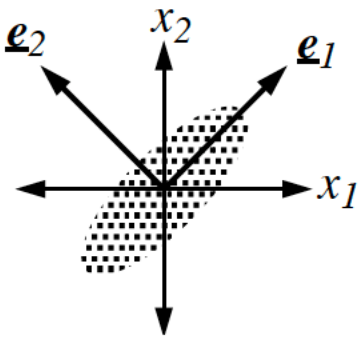
Let's assume the e_i are ordered by the variance of y_i such that:

$$\langle y_1^2 \rangle \geq \langle y_2^2 \rangle \geq \cdots \langle y_D^2 \rangle$$

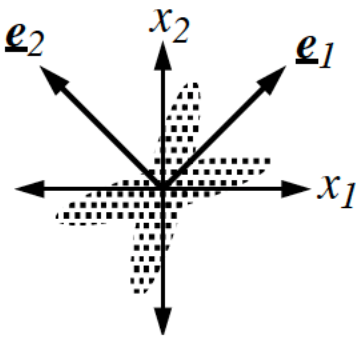
Gaussian

non-Gaussian

a.



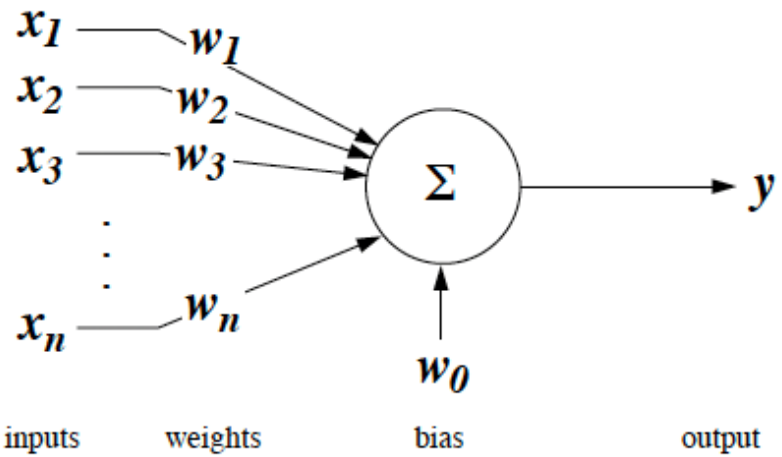
b.



Adapted from Olshausen

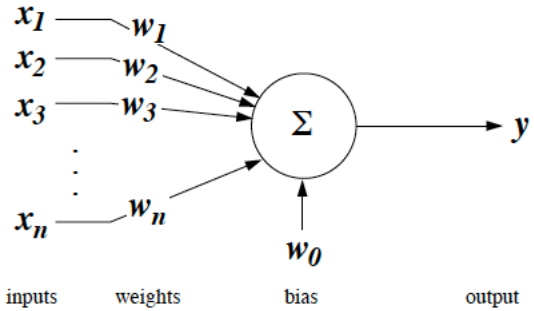
Linear Hebbian learning

Donald Hebb: “What fires together, wires together.”



Donald Hebb (1904 –1985)
Wikipedia

Linear Hebbian learning



Donald Hebb: “What fires together, wires together.”

Thus, for a linear neuron:

$$y = \sum_i w_i \, x_i$$

each weight should change proportionally to the correlation of y and x_i, i.e.:

$$w'_i \propto \langle y, x_i \rangle$$

Linear Hebbian learning

$$w'_i \propto \langle y, x_i \rangle$$

Let's rewrite this based on the activity of the neuron:

$$w'_i \propto \langle \sum_j w_j \, x_j, x_i \rangle \approx \sum_j w_j \, \langle x_j, x_i \rangle$$

Thus, for covariance matrix $C = (c_{i,j})$ we get the following differential equation that governs the evolution of the weights:

$$w' = Cw$$

$$w' = Cw$$

But how will the weights change?

Let's consider the trivial 1D case first:

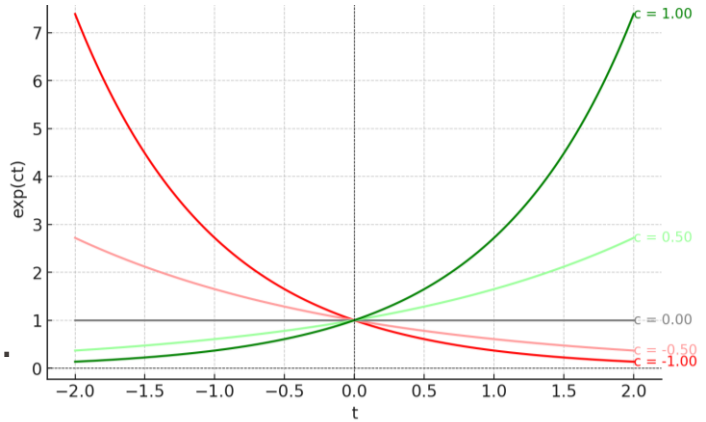
$$w' = cw$$

This first-order, linear ordinary differential equation is solved by:

$w(t) = w(0) \cdot \exp(ct)$.

So, the weight will increase exponentially ($c > 0$), or go to zero ($c < 0$).

The magnitude of the constant c determines the speed.



As C is symmetric ($c_{i,j} = c_{j,i}$), we can decompose it into:

$$C = \mathrm{U} D \mathrm{U}^T$$

with $\mathrm{U} = (e_1, \ldots, e_D)$ being an orthonormal matrix
such that $< e_i, e_j >= \delta_{i,j}$ for Kronecker delta

and

$D = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$.

Geometrically, U and U^T are rotation matrices and D a scaling matrix.

The e_i are **eigenvectors** with eigenvalue λ_i ,i.e. $C e_i = \lambda_i e_i$

Let's change the coordinates: $v = \mathrm{U}^T \mathrm{w}$

$$v' = Dv$$

Then:

$$v(t) = \exp(Dt)v(0) = \left(\exp(\lambda_i t)v_i(0)\right)_i$$

So just like in the 1D case, each component will grow/~~shrink~~ exponentially (matrix is positive def., so no shrinking).

Qualitatively, if $\lambda_1 > \lambda_i$,then quickly this component will dominate all others.

So w will approximately grow along e_1 -- i.e. the direction of the first principal component.

Shown differently, one can find for some constants $\mathrm{k_i}$:

$\mathrm{w}(t) = \mathrm{k}_1 \exp(\lambda_1 t) e_1 + \mathrm{k}_2 \exp(\lambda_2 t) e_2 + \ldots + \mathrm{k}_\mathrm{D} \exp(\lambda_\mathrm{D} t) e_\mathrm{D}$

Thus, $< \mathrm{w}(t)/|w|, e_i > \to 0$ for $i \neq 1$ and $< \mathrm{w}(t)/|w|, e_i > \to 1$ for $i = 1$

Linear Hebbian learning

We have seen that for $\mathbf{w}' \propto \mathbf{C}\mathbf{w}$ the weights grow (indefinitely) along the direction of the eigenvector of the covariance matrix $\mathbf{C}$. In other words, y will compute the projection of the data onto the first principal component.

To constrain the growth, one can modify Hebb's rule…

For more details, see **Neural Dynamics** by W. Gerstner et al.
https://neuronaldynamics.epfl.ch/online/Ch19.S3.html

Oja's rule

$$w' = \langle y(x - yw) \rangle$$

We note $\langle y(x - yw) \rangle = \langle yx \rangle - \langle y^2 \rangle w$. The first term is Hebb's rule and the second one constrains growth.

What is the equilibrium solution, $w' = 0$?

$Cw =< y^2 > w$. Thus, by design w will be an eigenvector of C.

Since

$$< y^2 >= w^T C w = w^T < y^2 > w =< y^2 > |w|^2$$

we find $|w| = 1$, indeed one can prove that it is the **strongest eigenvector**.

Learning multiple eigenvectors: Sanger's rule

Consider a system of m neurons, then the following rule:

$$w'_i = \langle y_i \left(x - \sum_{j \leq i} y_j \, w_j \right) \rangle$$

learns the first m principal components.

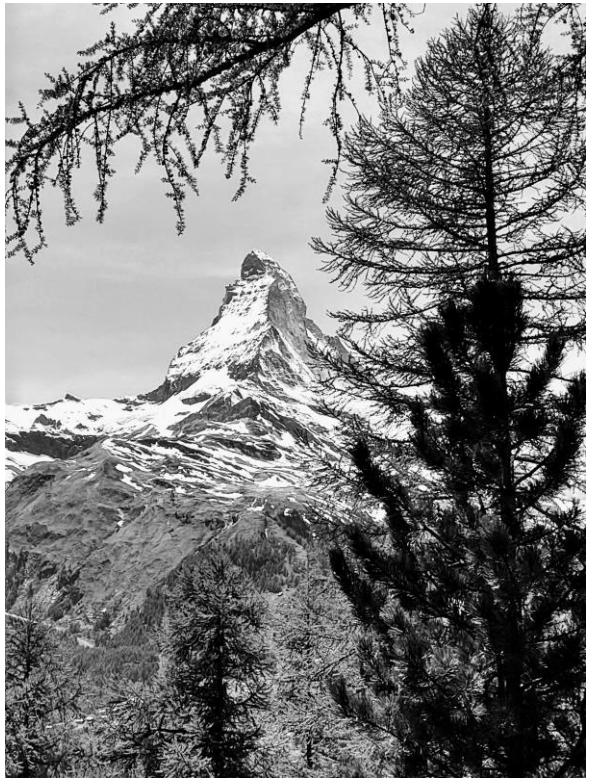
Intuitively, this works as one removes the explained variance of the largest principal components and then is left with Oja's rule. E.g., let's assume $w_1 \to e_1$, then:

$$w'_2 = \langle y_2 (x - y_1 e_1 - y_2 w_2) \rangle$$

So w_2 will point in the direction of the largest principal component of $X = \{x^{(t)} - \langle x^{(t)}, e_1 \rangle\}_{1 \leq t \leq T}$. That's of course e_2.

Let’s look at a concrete example, where the data is given by local patches from images. That’s something the visual system cares about….

How should natural images be represented?



Matterhorn
Wikipedia

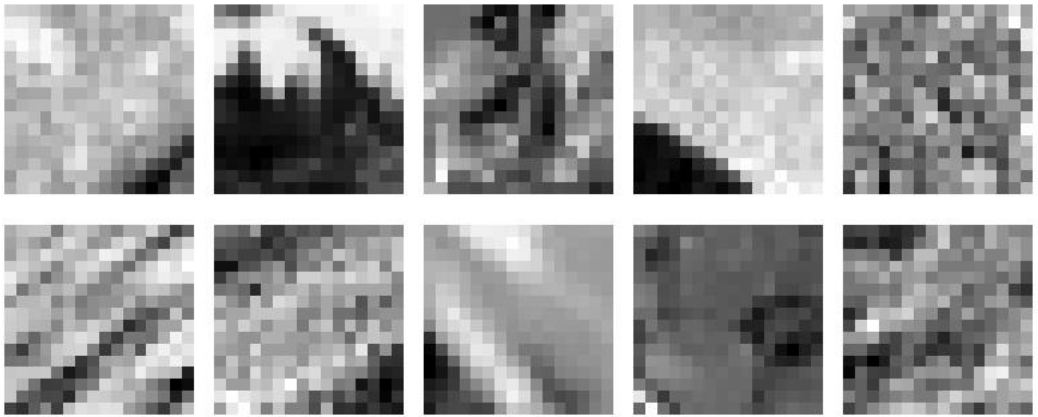
Let's assume that images are represented by linear superpositions of (*not necessarily*) orthogonal basis functions:

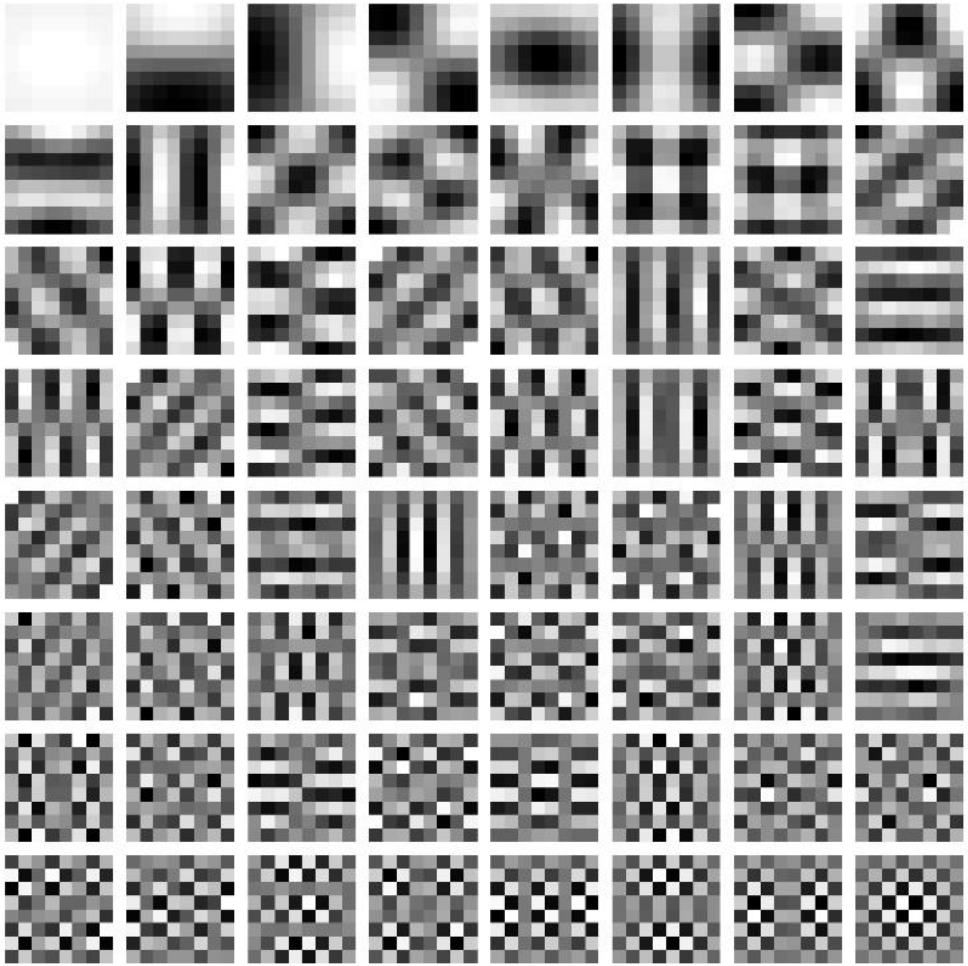
$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \epsilon(x,y)$$

$$E = \frac{1}{2}|I - \Phi a|^2$$

What are the principal components of naturalistic images?

Extract patches from natural images, e.g. 16 x 16 pixels





Learned weights with PCA (Sanger's rule)

Olshausen & Field, 1996 Nature

Orientation selectivity

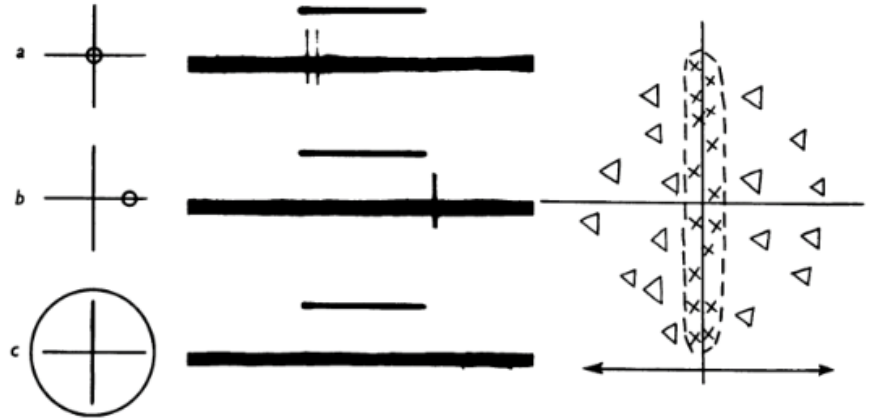


Fig. 2. Responses of a unit to stimulation with circular spots of light. Receptive field located in area centralis of contralateral eye. (This unit could also be activated by the ipsilateral eye.) *a*, 1° spot in the centre region; *b*, same spot displaced 3° to the right; *c*, 8° spot covering entire receptive field. Stimulus and background intensities and conventions as in Fig. 1. Scale, 6°.

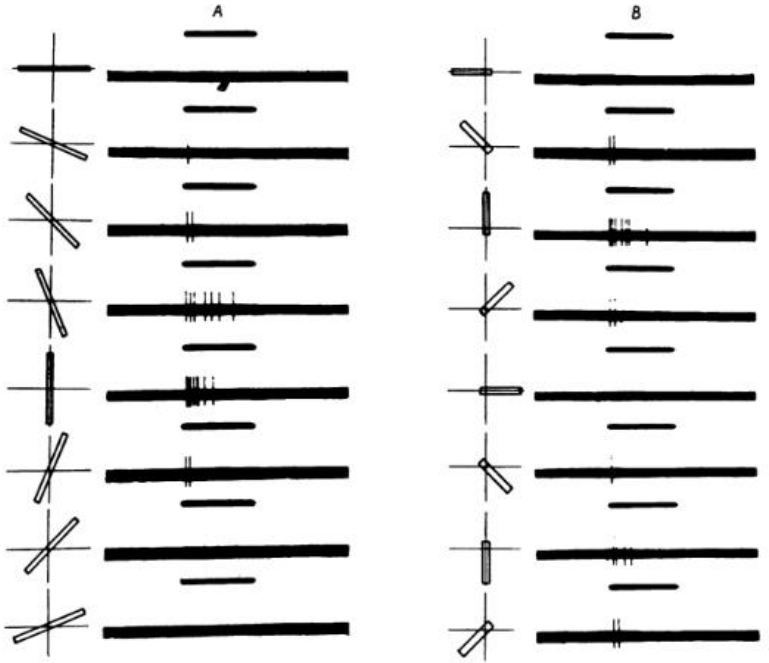


Fig. 3. Same unit as in Fig. 2. *A*, responses to shining a rectangular light spot, 1° × 8°; centre of slit superimposed on centre of receptive field; successive stimuli rotated clockwise, as shown to left of figure. *B*, responses to a 1° × 5° slit oriented in various directions, with one end always covering the centre of the receptive field: note that this central region evoked responses when stimulated alone (Fig. 2*a*). Stimulus and background intensities as in Fig. 1; stimulus duration 1 sec.

Hubel &Wiesel 1959

Gabor function

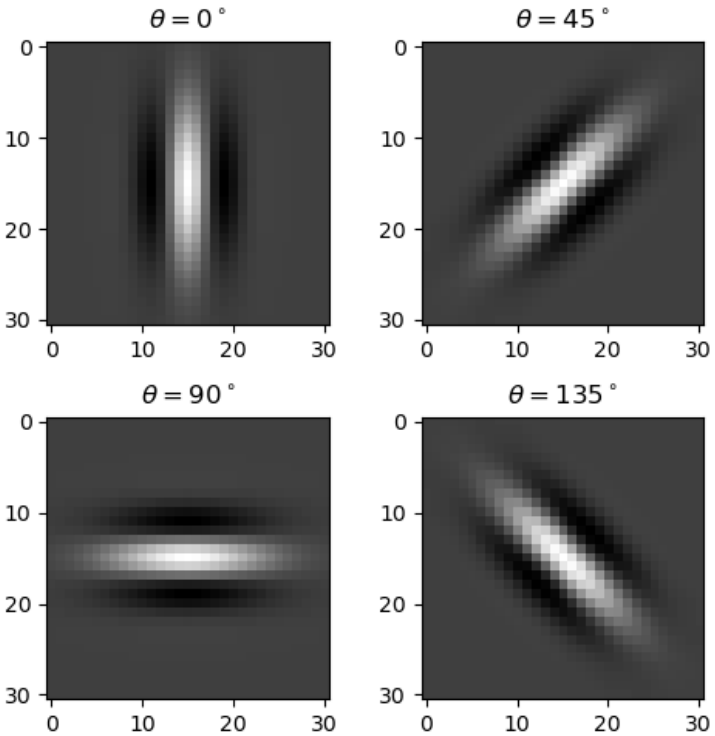
A popular mathematical approximation of the spatial receptive field of a simple cell is given by the Gabor function:

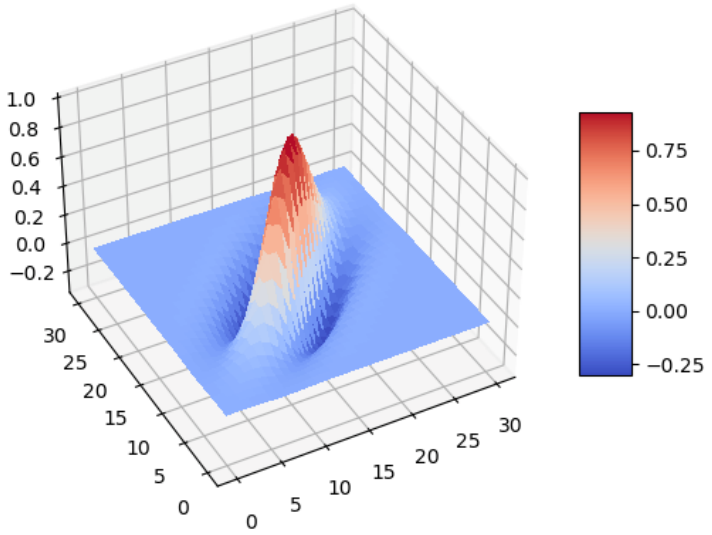
$$\mathrm{f}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cdot \cos(kx - \varphi)$$

with:

- σ_x, σ_y determine the spatial receptive field in x and y direction, respectively
- k is the preferred spatial frequency (i.e. the spacing of light/dark bars that give maximal response)
- φ is the preferred spatial phase
- for simplicity this function has coordinates so that border of On/Off region are parallel to y-axis

Example visualization (see exercises)





Single units and sensation: a neuron doctrine for perceptual psychology?

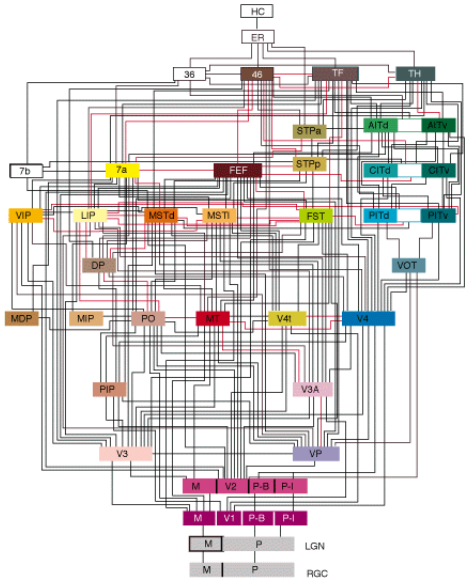
1. To understand nervous function one needs to look at interactions at a cellular level, rather than either a more macroscopic or microscopic level, because **behaviour depends upon the organized pattern of these intercellular interactions.**

2. The **sensory system is organized to achieve as complete a representation of the sensory stimulus as possible with the minimum number of active neurons**.

….

Barlow, 1972 Perception

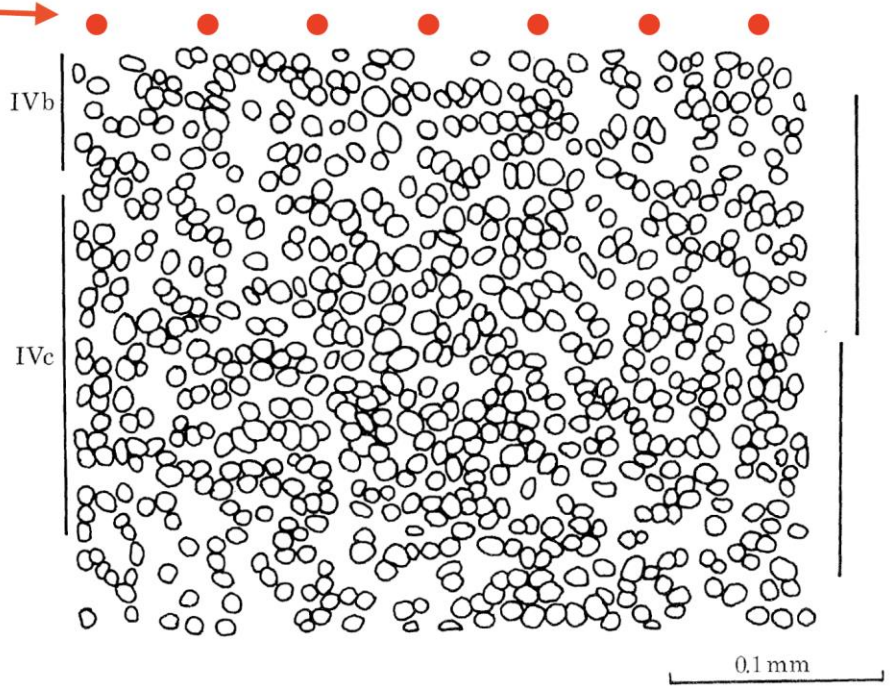
V1 is highly overcomplete



Felleman and Van Essen *Cerebral Cortex* 1991

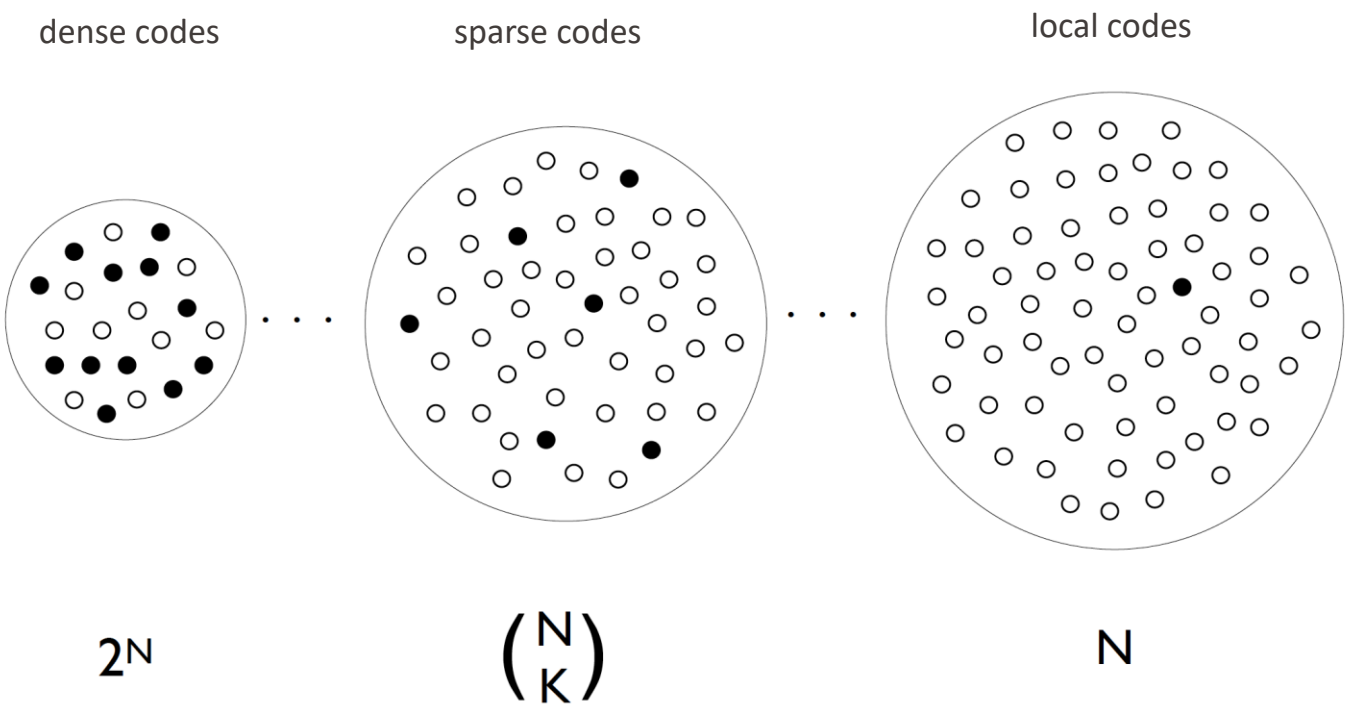
LGN afferents

layer 4 cortex



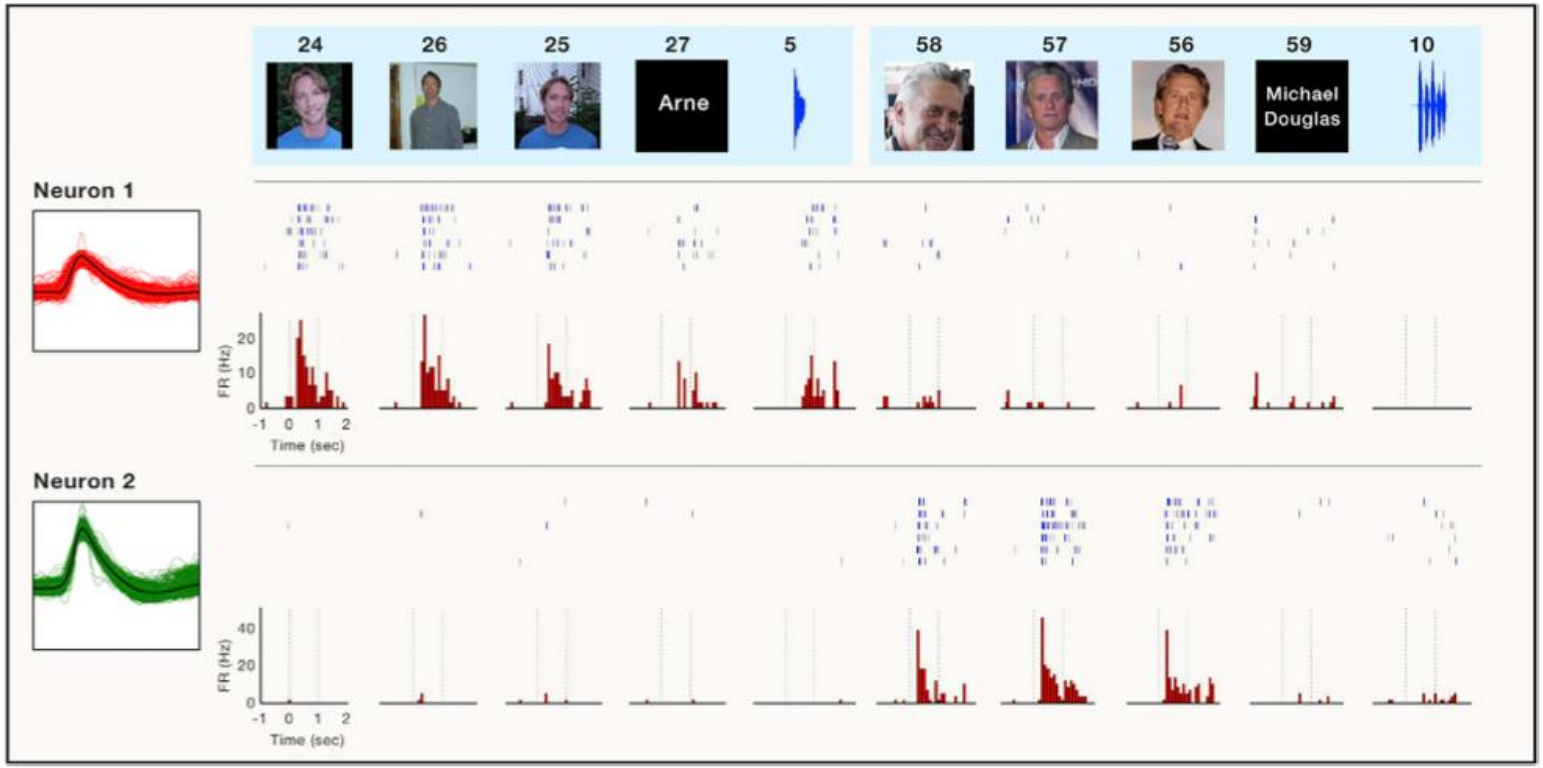
Barlow, 1981

Different types of coding schemes



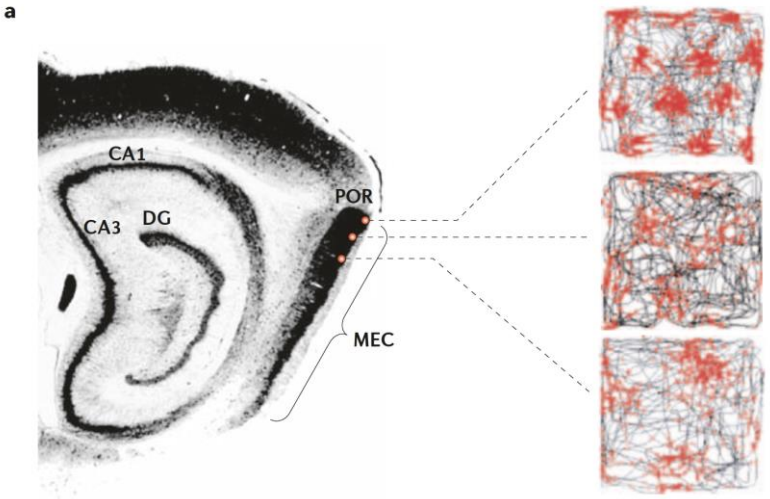
Reminder: an example grandmother cell

Concept cells



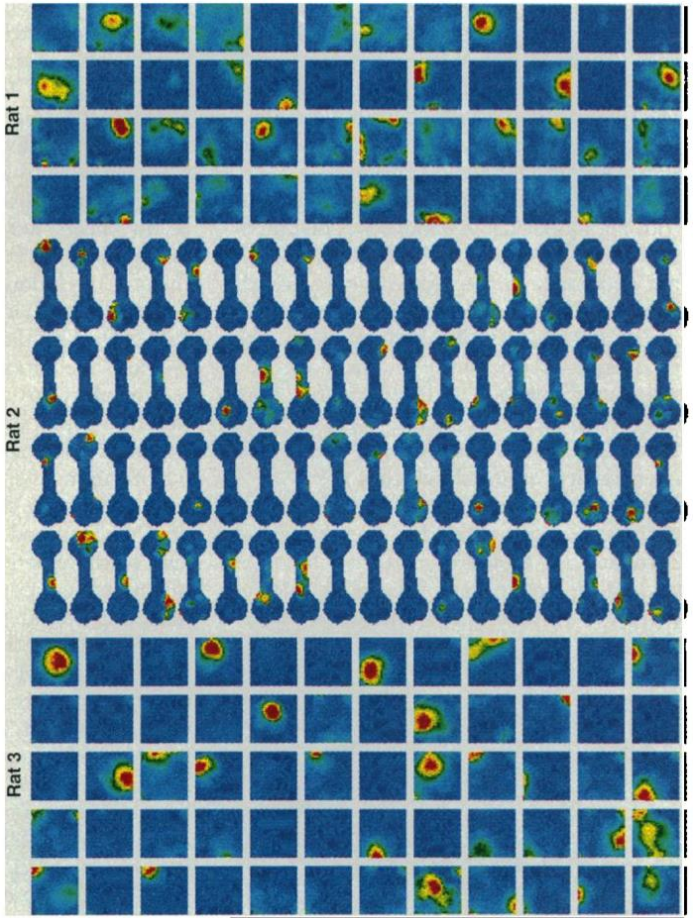
Quiroga, Cell 2019

Reminder: grid cells provide a more “dense” code

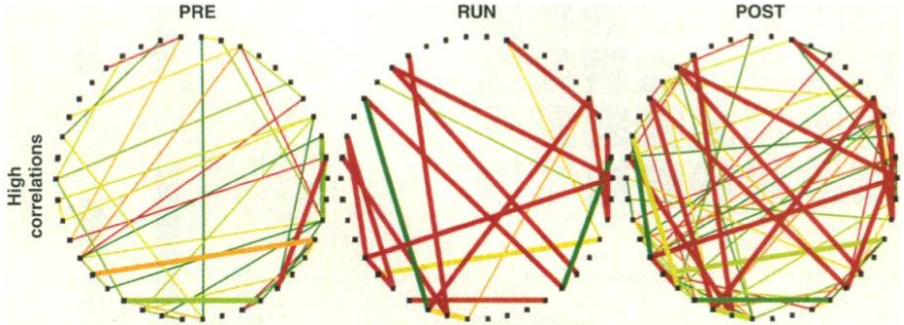


McNaugthon et al., Nature Review Neuroscience 2006

Reminder: place code is a sparse code



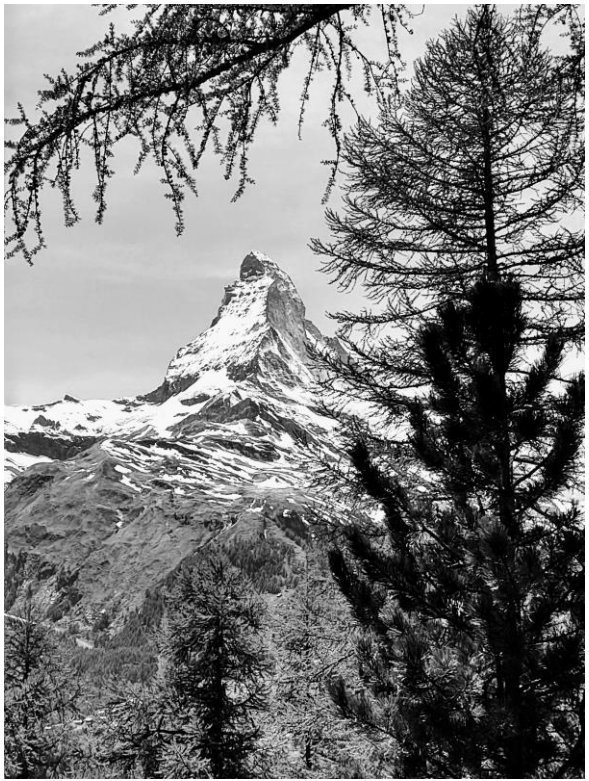
Experience creates correlated replay during sleep!



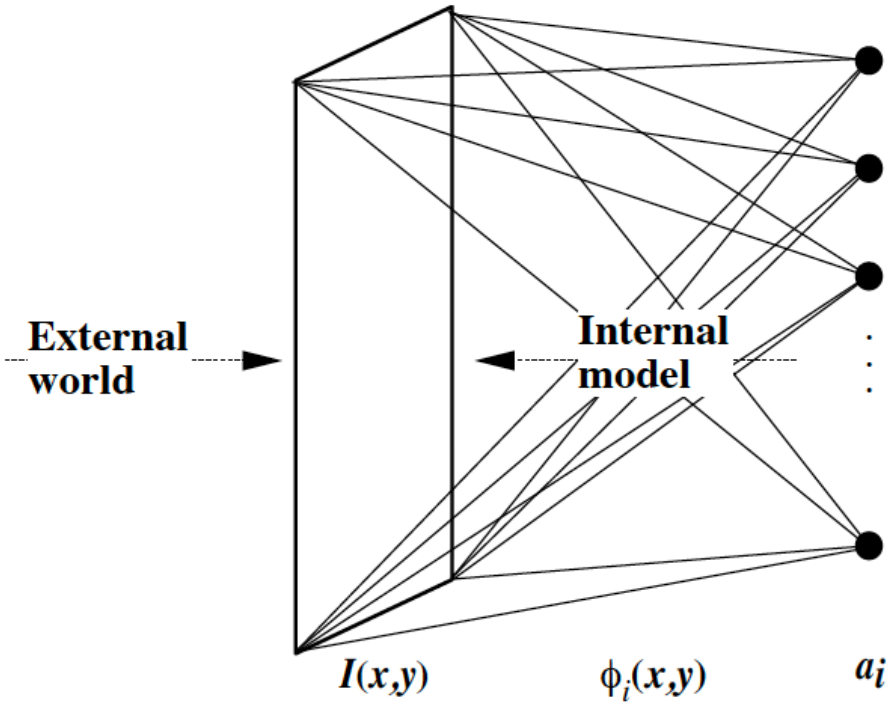
Wilson and McNaughton, Science 1994

Sparse coding

$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \epsilon(x,y)$$



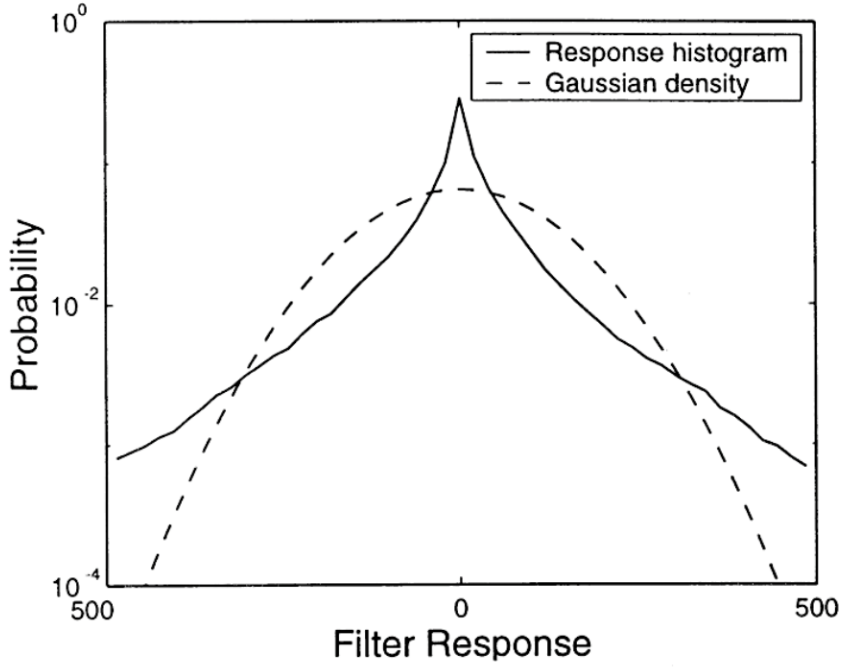
Matterhorn
Wikipedia



Sparse activity!

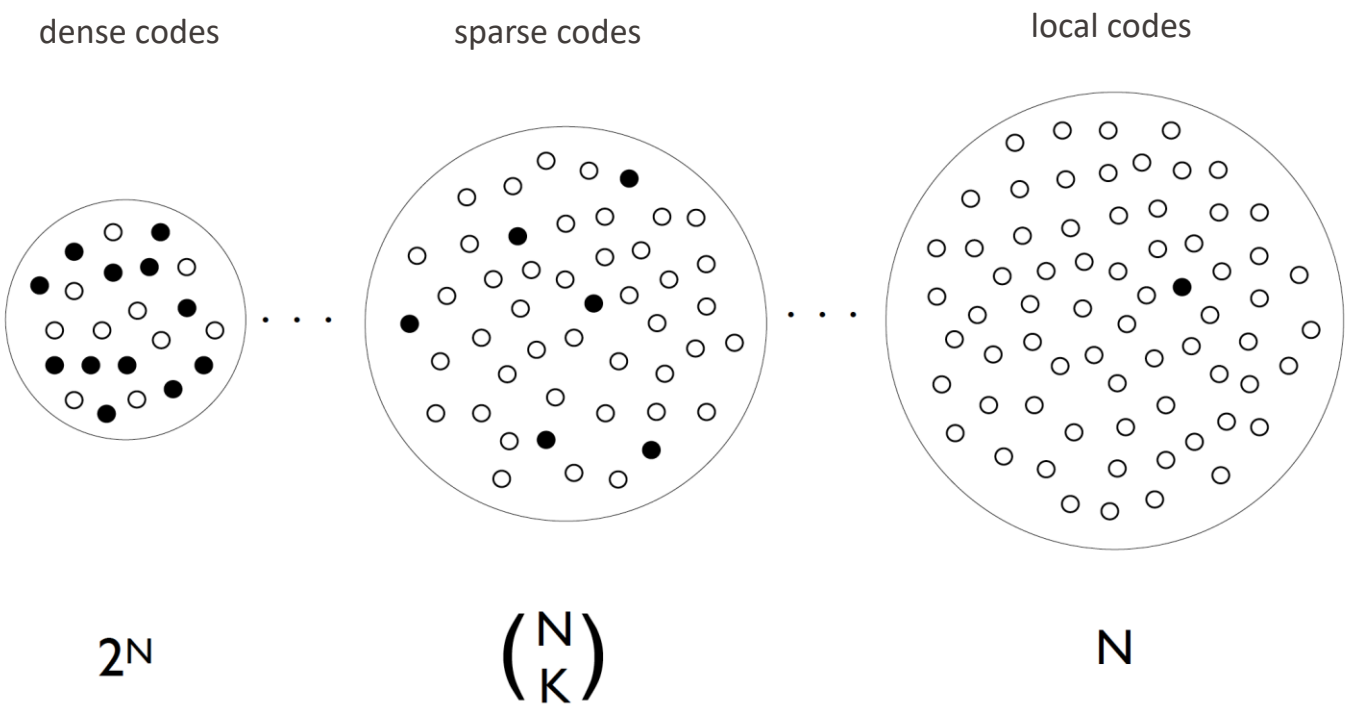
Olshausen & Field, 1996 Nature

Gabor-filters create sparse activations



Olshausen & Field, 1996 Nature

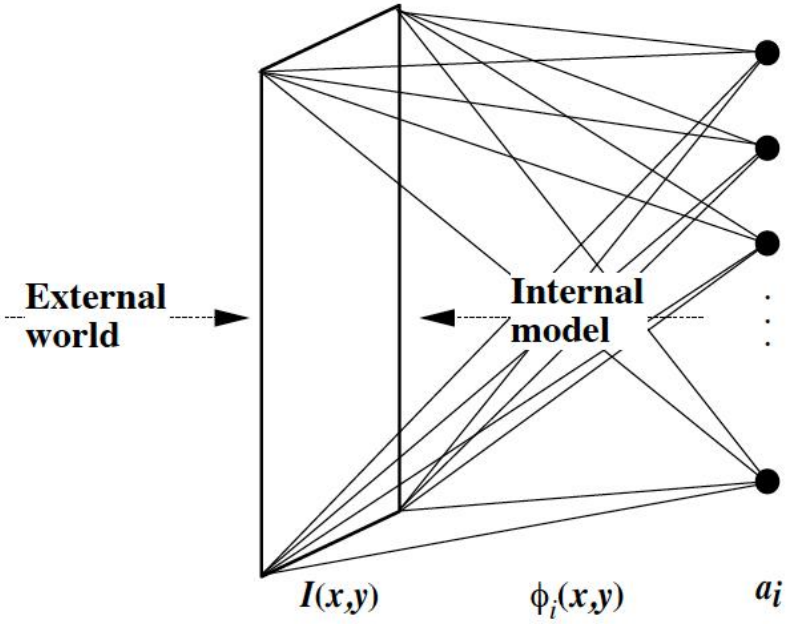
Different types of coding schemes

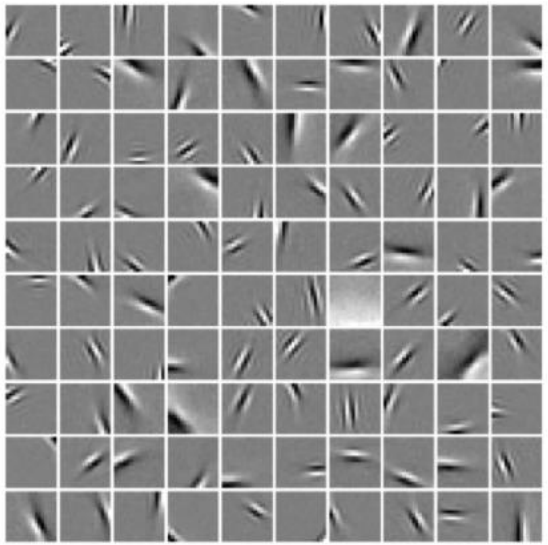


Energy function

$$E = \underbrace{\frac{1}{2}|I - \Phi a|^2}_{\text{Preserve information}} + \underbrace{\lambda \sum_i C(a_i)}_{\text{Sparse activity}}$$

Sparse coding model of V1





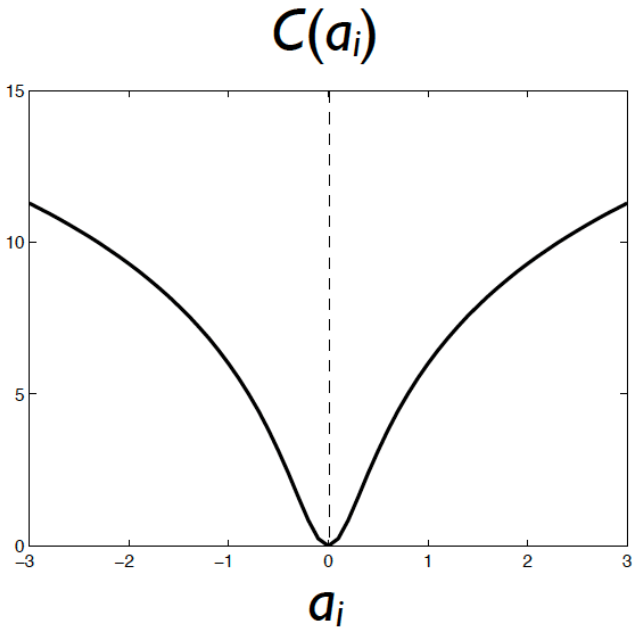
$\phi_i(x, y)$

$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \epsilon(x,y)$$

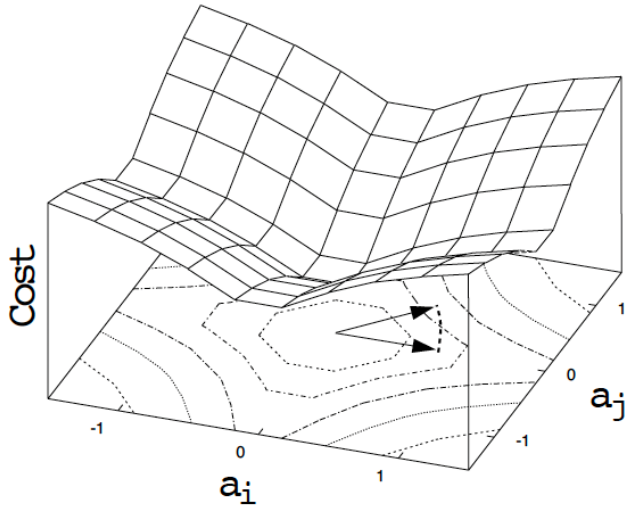
Olshausen & Field, 1996 Nature

Cost function for sparsity

$$E = \frac{1}{2}|I - \Phi a|^2 + \lambda \sum_i C\,(a_i)$$



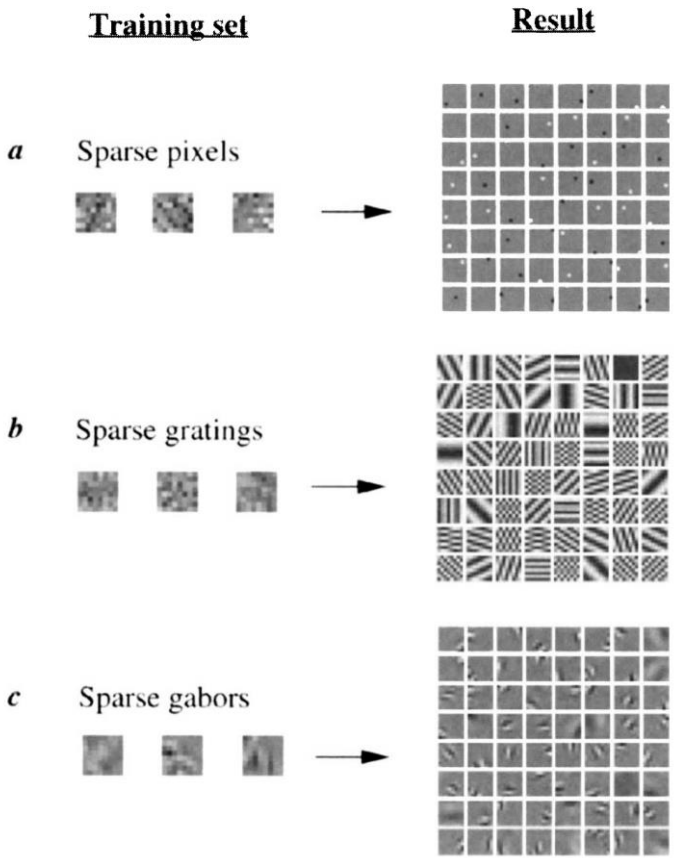
$$C(a_i) = log\left(1 + a_i^2\right)$$



Note: other cost functions are possible, e.g. |a|

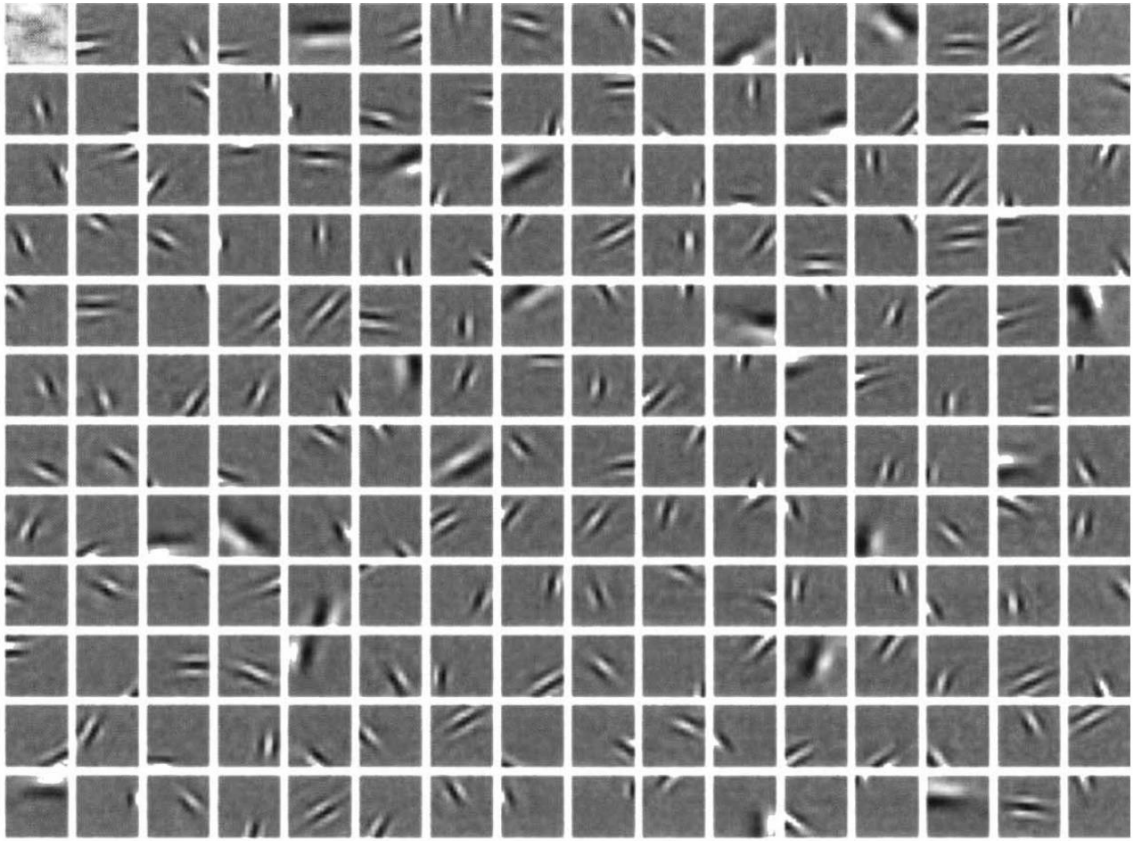
Olshausen & Field, 1996 Nature

Sparse coding recovers latent structure of data



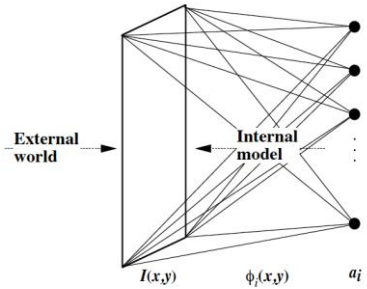
Olshausen & Field, 1996 Nature

Results of training sparse coding model on 16 x 16 patches



Olshausen & Field, 1996 Nature

Coefficients are computed via gradient descent



$$\tau a'_i = -\frac{\partial E}{\partial a_i}$$

Olshausen & Field, 1996 Nature

Coefficients are computed via gradient descent

$$\tau a'_i = -\frac{\partial E}{\partial a_i} = b_i - \sum_{j \neq i} G_{i,j}\, a_j - f_\lambda(a_i)$$

$$b_i = \sum_{x,y} \phi_i\,(x,y) I(x,y)$$

$$G_{i,j} = \sum_{x,y} \phi_i\,(x,y) \phi_j(x,y)$$

$$f_\lambda(a_i) = a_i + \lambda C'(a_i)$$

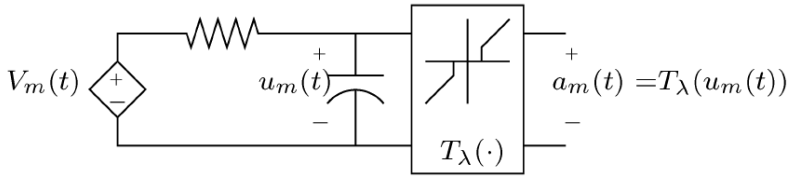
Olshausen & Field, 1996 Nature

Learning rule for the basis functions

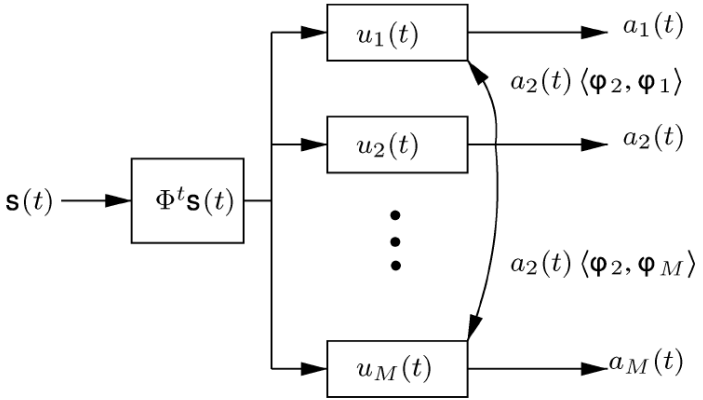
$$\Delta\phi_i = -\eta \frac{\partial E}{\partial \phi_i} = (I - \Phi\hat{a})\hat{a}_i$$

Olshausen & Field, 1996 Nature

Coefficients can be computed by leaky integrators and lateral inhibition



$$V_m(t) = \langle \varphi_m, \mathsf{s}(t) \rangle - \sum_{n \neq m} T_\lambda(u_n(t)) \langle \varphi_m, \varphi_n \rangle$$

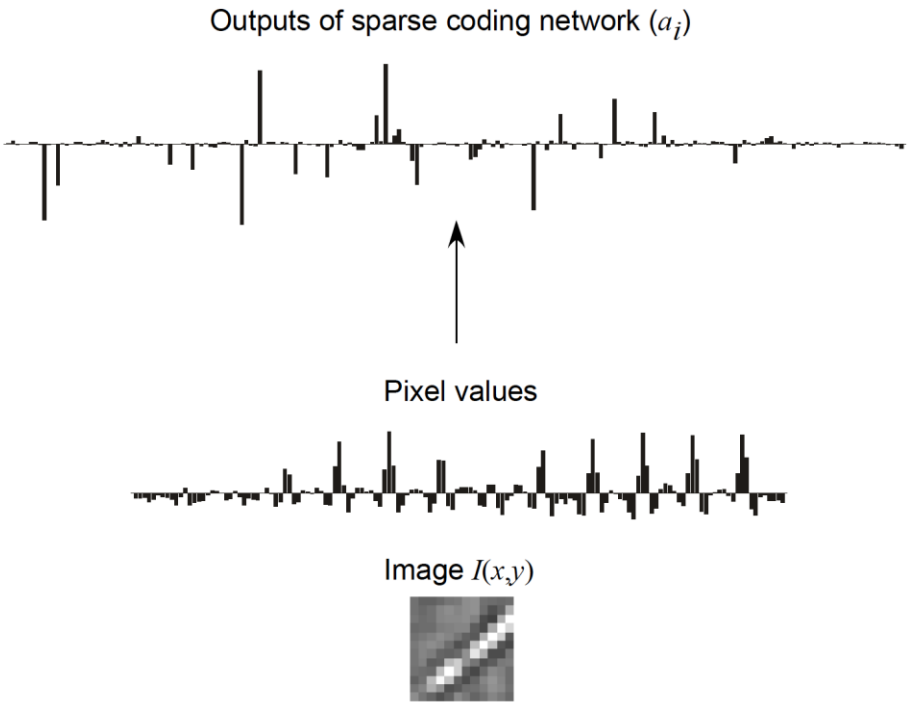


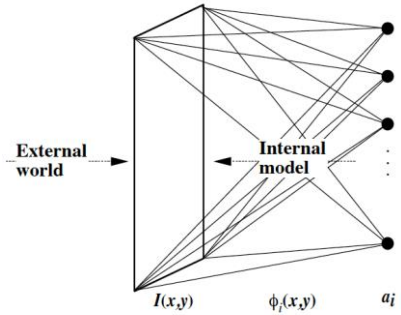
$$\dot{\boldsymbol{u}}(t) = f(\boldsymbol{u}(t)) = \frac{1}{\tau}\left[\boldsymbol{b}(t) - \boldsymbol{u}(t) - \left(\Phi^t \Phi - I\right) \boldsymbol{a}(t)\right],$$

$$\boldsymbol{a}(t) = T_\lambda(\boldsymbol{u}(t)).$$

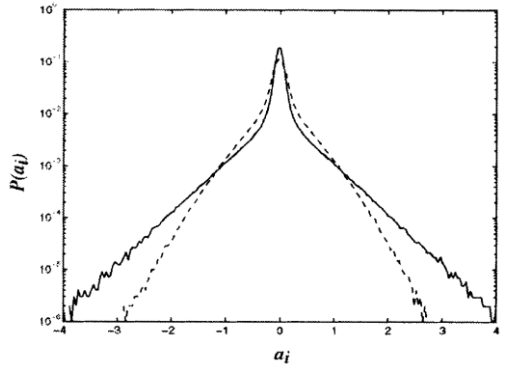
Rozell et al., 2008 Neural Computation

Sparse coding model of V1





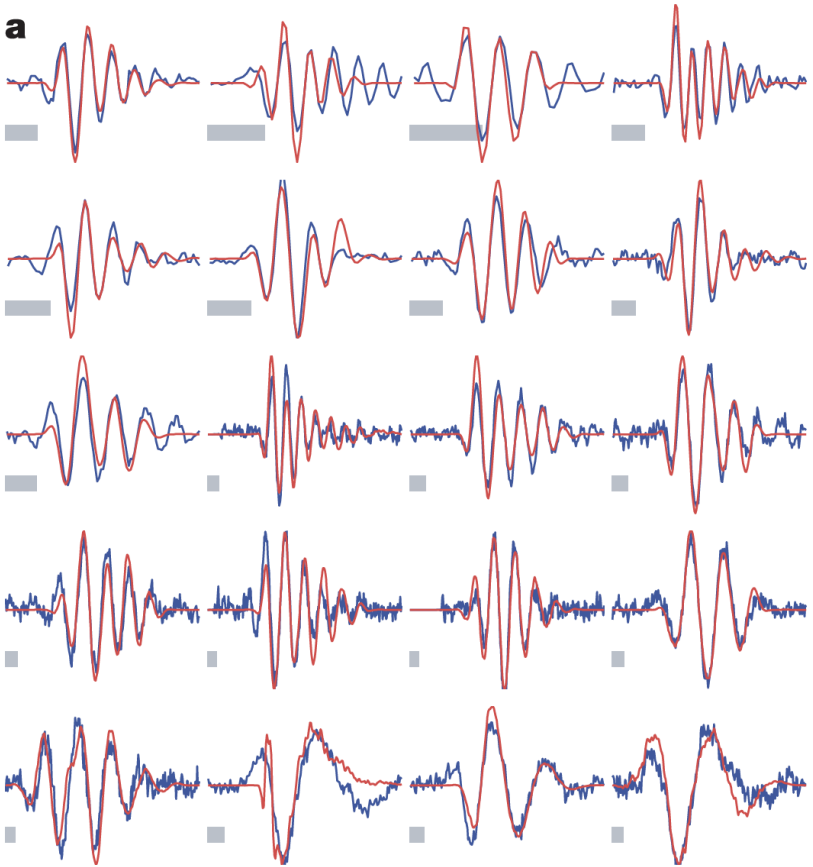
Activity becomes sparser over training:



- Olshausen & Field, 2004 Current Opin. In Neurobio

Olshausen & Field, 1996 Nature

Efficient codes for natural sounds



$$x(t) = \sum\sum s_i^m \phi_m \left(t - \tau_i^m\right) + \varepsilon(t)$$

Predicted auditory kernels

Experimental auditory kernels from cat auditory nerve

Smith & Lewicki, 2006 Nature

Evidence for sparse coding has been found in many different sensory areas

- Mushroom body, locust (Laurent)
- HVC, zebra finch (Fee)
- Auditory cortex, mouse (DeWeese & Zador)
- Hippocampus, rat/primate (Thomson & Best; Skaggs)
- Barrel cortex, rat (Brecht)
- …

Olshausen & Field, 2004 Current Opin. In Neurobio

Limitations

- These simple models have been most effective in describing early sensory responses (e.g., primary visual/auditory cortex). They are often less effective the further you move away from sensory input.

- Because these models described a lot of neural properties at the time, computational neuroscience has largely ascribed to focusing on simplicity and localized mechanisms, with the hope that these principles can eventually be scaled/assembled into a holistic computational description of the brain.

- Personal opinion: the sole focus on simplicity and individual mechanisms is holding our field back. To make sense of a system as complex of the brain, we need to embrace the complexity in the models we build and gain intuition at a higher level of abstraction.

Take-home messages

- Representation learning with unsupervised tasks provides a normative framework for studying tuning curves in the nervous system
- Oja's rule converges to the strongest eigenvector of the data
- PCA on natural image patches does not provide a localized, oriented representation (and thus does not resemble primary visual cortex V1)
- Sparse coding predicts properties across a wide range of sensory areas incl. V1

Exercises: you will implement Olshausen's model!