# Biologically inspired learning in Anns and Visual cortex

# Class 1

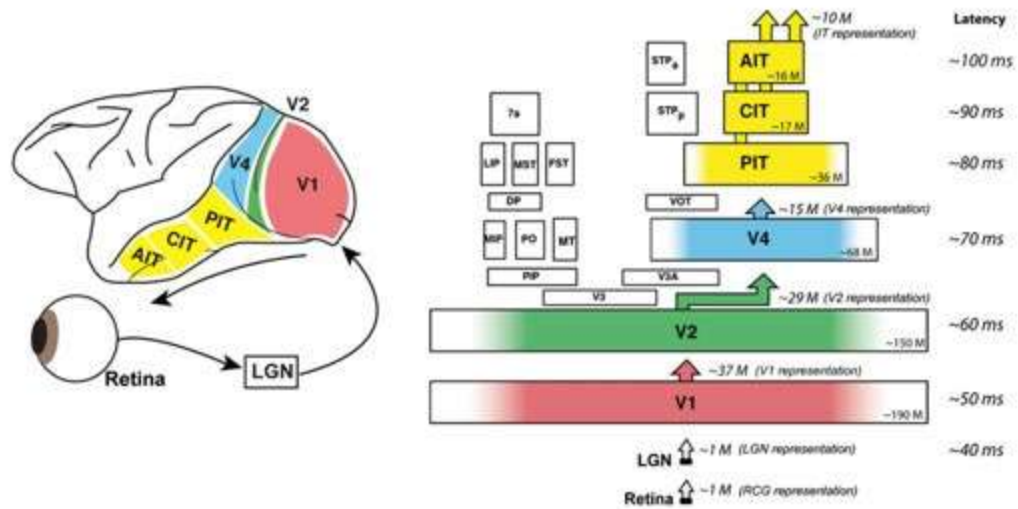
Invariance and sample complexity. Invariant and selective representations.
Neural motivation and neurally plausible algorithms. A group perspective.
Invariant and selective representation using kernel methods.

To survive an animal's brain needs to figure out how to recognize a predator in complex scenes having seen him a (very) few times.

How does the brain extract meaningful information from a wide variety of sensory input to achieve **recognition from few examples of new objects**? Invariant and selective representations.

# Invariant representations in visual cortex?

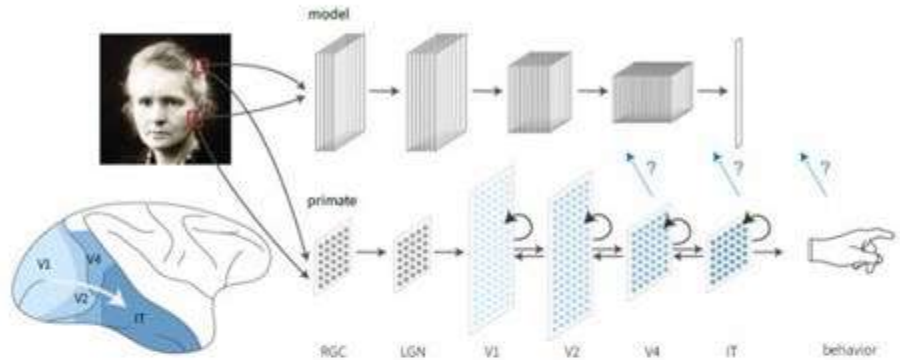
Despite object variability, mammals can rapidly recognize objects with no effort.



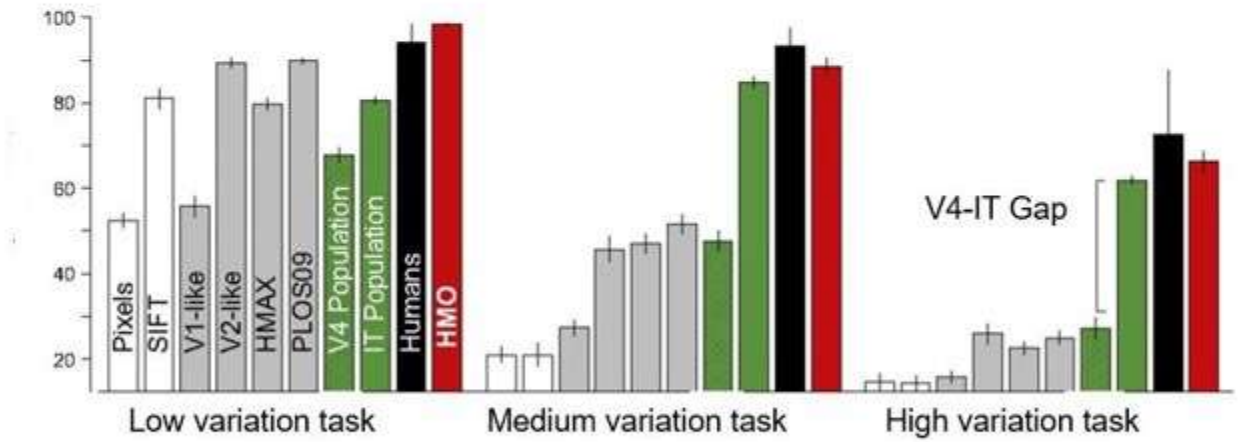

"How Does the Brain Solve Visual Object Recognition?", DiCarlo, Neuron, 2012

Experimental evidence points to the hypothesis that the ventral stream **progressively builds invariance to nuisance transformations and selectivity to object identity.** However the underlying neuronal mechanism has yet been not elucidated.

# Invariant representations in ANNs?



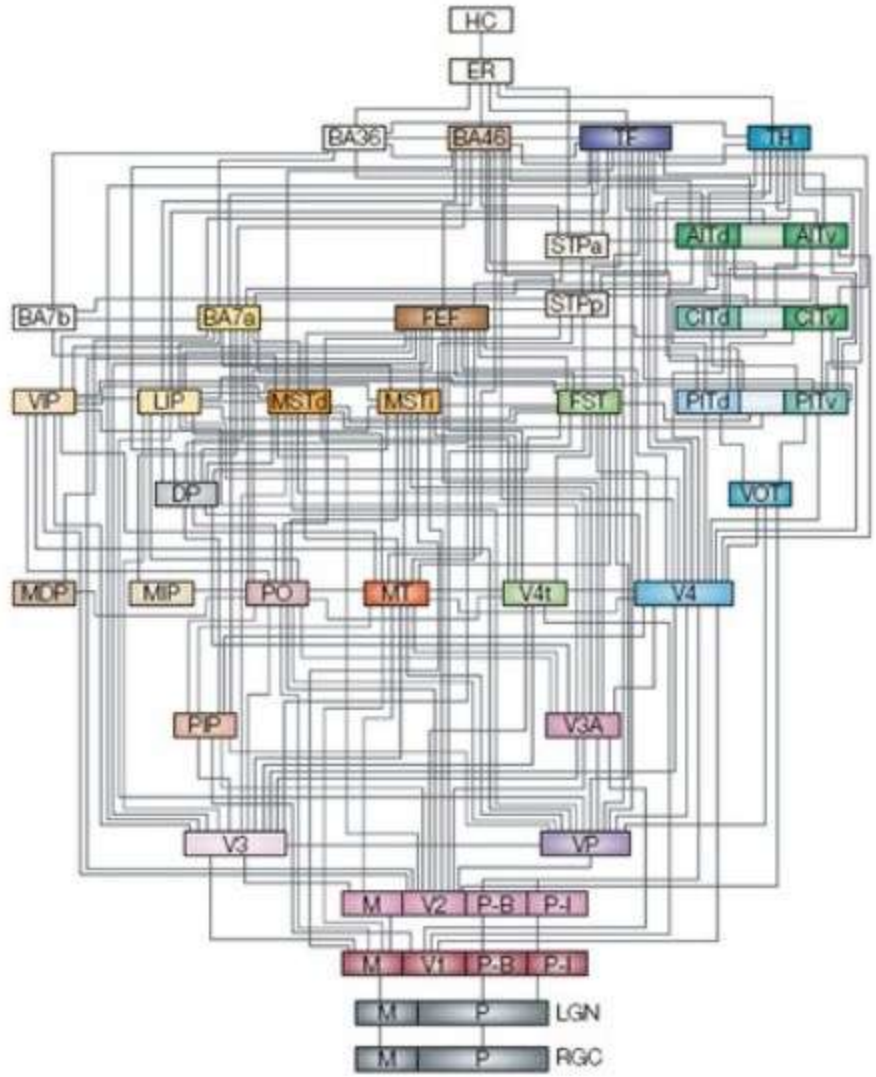

DiCarlo, PNAS, 2021
Nat. Neur., 2016
NIPS, 2013




ANNs are **very good predictors** of neural activity and are capable of dealing with nuisances transformations. However if and how they mimic brain computations is not clear.

How does the brain rapidly extract meaningful information from a wide variety in the sensory input?

# How does the brain organize such information?



# Importance of ANNs in Neuroscience

2021 Special Issue on AI and Brain Science: Perspective

## Natural and Artificial Intelligence: A brief introduction to the interplay between AI and neuroscience research

Tom Macpherson [a], Anne Churchland [b], Terry Sejnowski [c,d], James DiCarlo [e], Yukiyasu Kamitani [f,g], Hidehiko Takahashi [h], Takatoshi Hikida [a,*]

[a] Laboratory for Advanced Brain Functions, Institute for Protein Research, Osaka University, Osaka, Japan
[b] Cold Spring Harbor Laboratory, Neuroscience, Cold Spring Harbor, NY, USA
[c] Computational Neurobiology Laboratory, Salk Institute for Biological Studies, CA, USA
[d] Division of Biological Sciences, University of California San Diego, CA, USA
[e] Brain and Cognitive Sciences, Massachusetts Institute of Technology, MA, USA
[f] Department of Neuroinformatics, ATR Computational Neuroscience Laboratories, Kyoto, Japan
[g] Graduate School of Informatics, Kyoto University, Kyoto, Japan
[h] Department of Psychiatry and Behavioral Sciences, Tokyo Medical and Dental University Graduate School, Tokyo, Japan

ARTICLE INFO

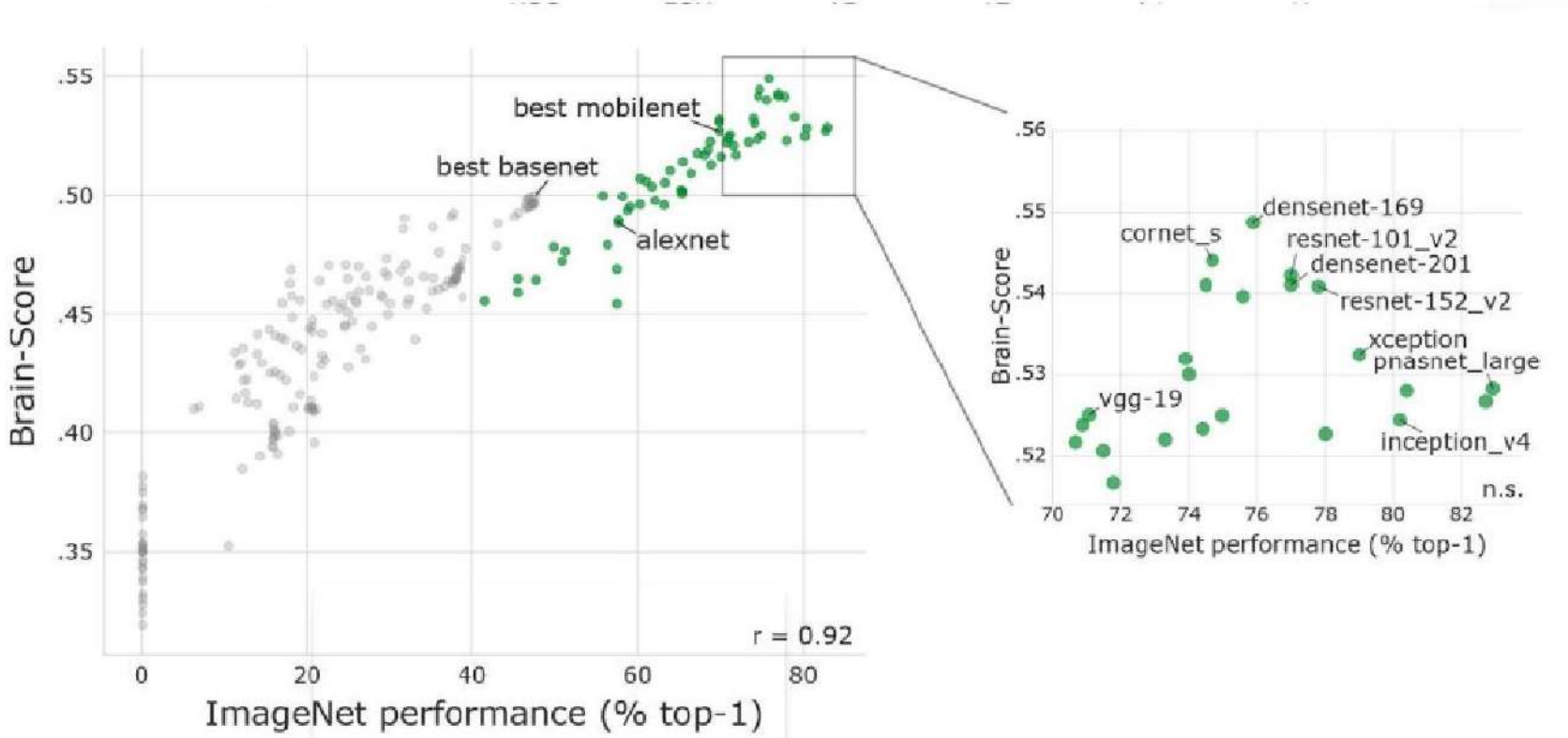
*Article history:*
Available online 28 September 2021

*Keywords:*
Artificial intelligence
Neuroscience
Neural imaging
Visual processing
Working memory
Computational psychiatry

ABSTRACT

Neuroscience and artificial intelligence (AI) share a long history of collaboration. Advances in neuroscience, alongside huge leaps in computer processing power over the last few decades, have given rise to a new generation of *in silico* neural networks inspired by the architecture of the brain. These AI systems are now capable of many of the advanced perceptual and cognitive abilities of biological systems, including object recognition and decision making. Moreover, AI is now increasingly being employed as a tool for neuroscience research and is transforming our understanding of brain functions. In particular, deep learning has been used to model how convolutional layers and recurrent connections in the brain's cerebral cortex control important functions, including visual processing, memory, and motor control. Excitingly, the use of neuroscience-inspired AI also holds great promise for understanding how changes in brain networks result in psychopathologies, and could even be utilized in treatment regimes. Here we discuss recent advancements in four areas in which the relationship between neuroscience and AI has led to major advancements in the field; (1) AI models of working memory, (2) AI visual processing, (3) AI analysis of big neuroscience datasets, and (4) computational psychiatry.

# Which ANN?: Brain score



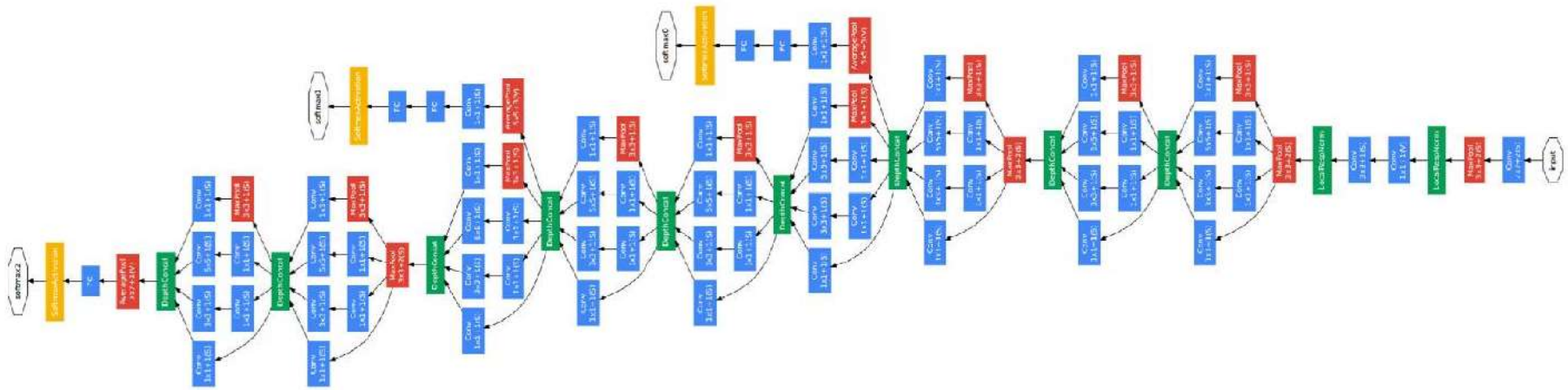

"Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" M. Schrimpfet al. 2020, *https://www.brain-score.org/*

Score based on: neural activity matching and behavioural tasks (recognition etc.).

# Brain score: competition winners

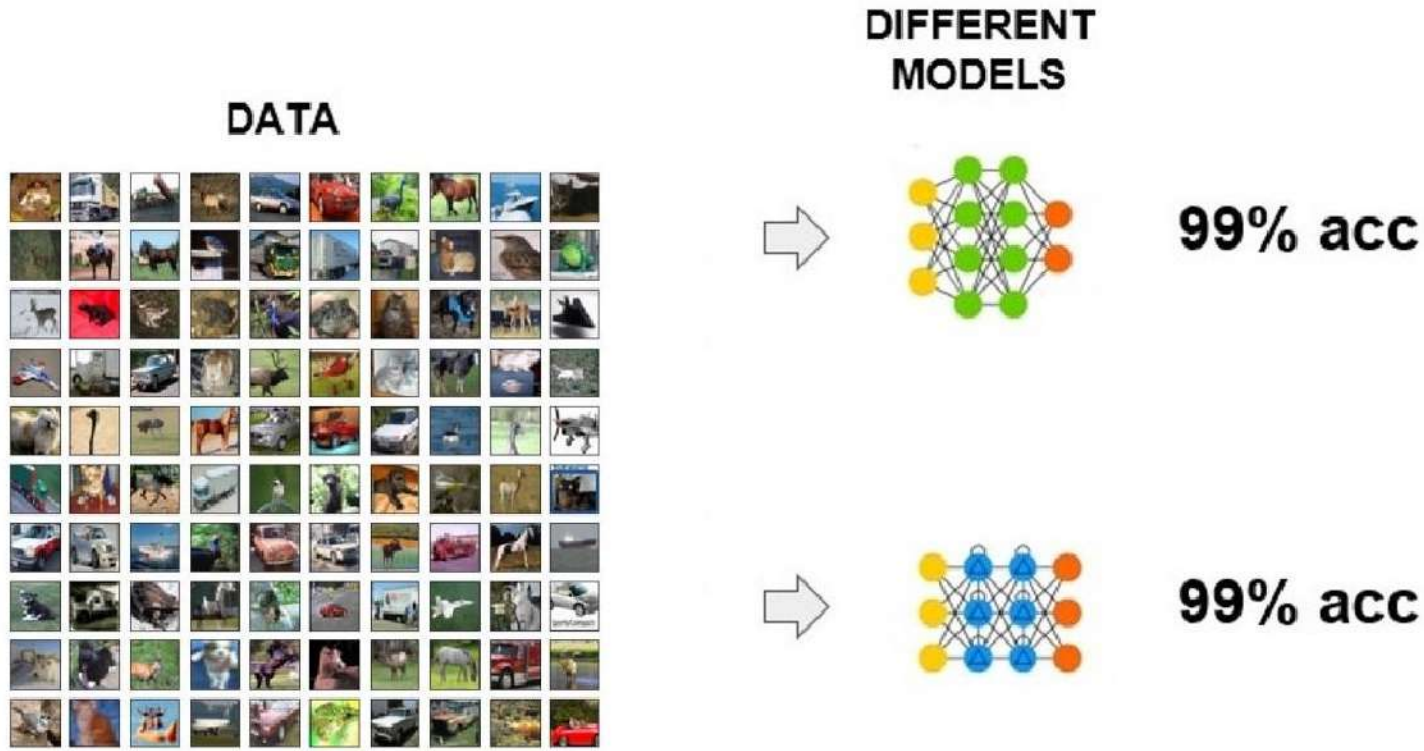
| Rank | Model<br>submitted by | average | V1 23 benchmarks | V2 1 benchmark | V4 4 benchmarks | IT 5 benchmarks | behavior 1 benchmark | engineering 6 benchmarks |
|---|---|---|---|---|---|---|---|---|
| 1 | effnetb1_cutmixpatch_augmix_robust32_avge4e7_manylayers_324x288<br>Alexander Riedel | .495 | .588 | .360 | .481 | .412 | .652 | .277 |
| 2 | voneresnet-50-robust<br>Tiago Marques | .471 | .531 | .391 | .471 | .417 | .545 | X |
| 3 | custom_model_cv_18_dagger_408<br>William Berrios | .467 | .493 | .342 | .514 | .425 | .562 | .368 |
| 4 | resnet50_finetune_cutmix_e3_robust_linf8255_e0_247x234<br>Alexander Riedel | .466 | .584 | .362 | .472 | .364 | .549 | .227 |
| 5 | effnetb1_cutmix_augmix_sam_e1_5avg_424x377<br>Alexander Riedel | .463 | .482 | .291 | .499 | .381 | .664 | .216 |
| 6 | resnet50_finetune_cutmix_AVGe2e3_robust_linf8255_e0_247x234<br>Alexander Riedel | .462 | .584 | .360 | .464 | .368 | .536 | .232 |
| 7 | voneresnet-50-non_stochastic<br>Tiago Marques | .461 | .589 | .326 | .484 | .398 | .530 | .446 |
| 8 | resnet-50-robust<br>Joel Dapello | .459 | .594 | .365 | .437 | .386 | .515 | .307 |
| 9 | resnet-152_v1<br>Brain-Score Team | .458 | .524 | .338 | .498 | .397 | .533 | .480 |

# Perfect score network



Does this architecture and neuronal connectivity/activity reflects somehow that of the brain?

# Multiple solutions for the same computational problem



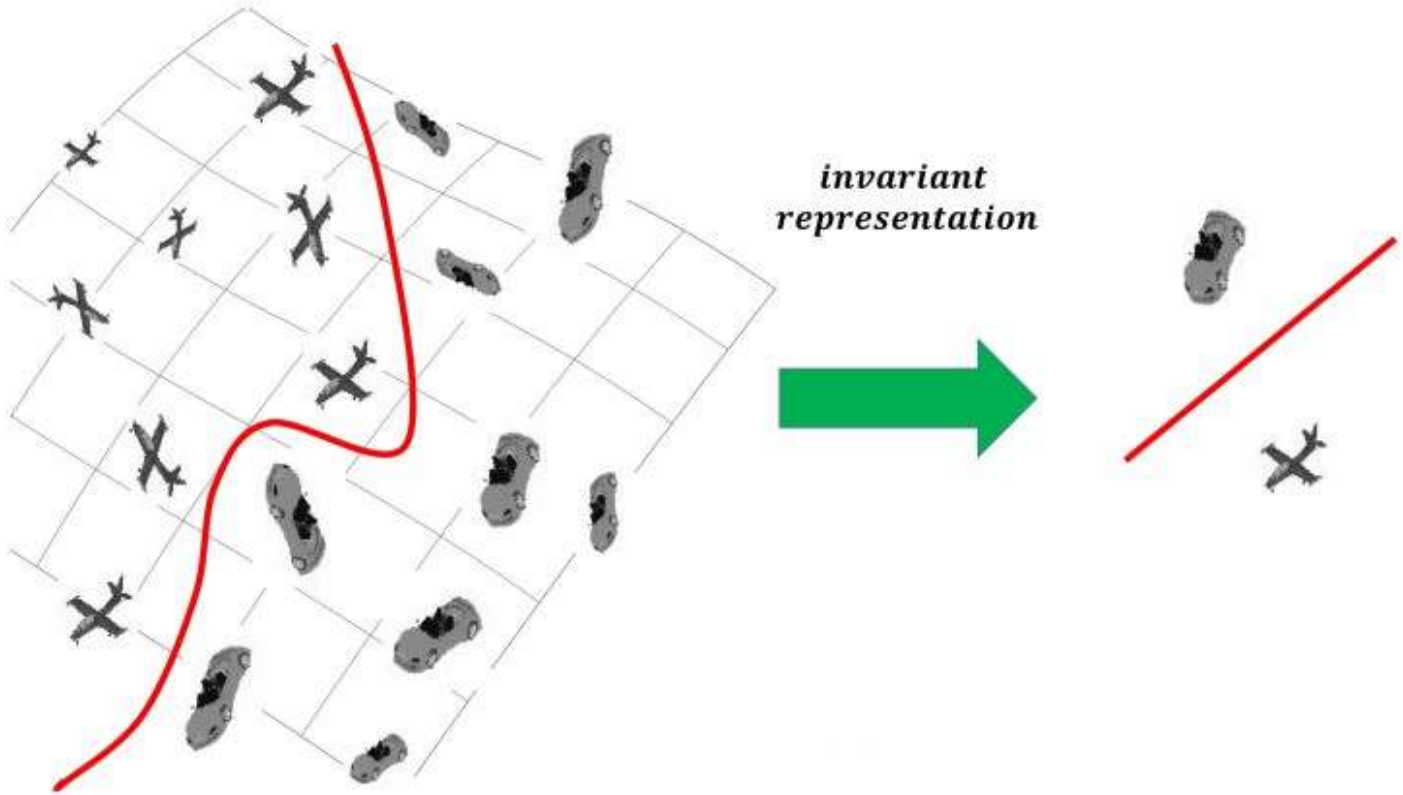

How is the network choosing a particular solution?

# Approach: invariant representations

Invariant recognition has been studied for decades, however the underlying neuronal mechanism has been not jet elucidated.

I argue that the brain has built a representation invariant to viewpoint and selective for object identity. Such representation allows rapid recognition.

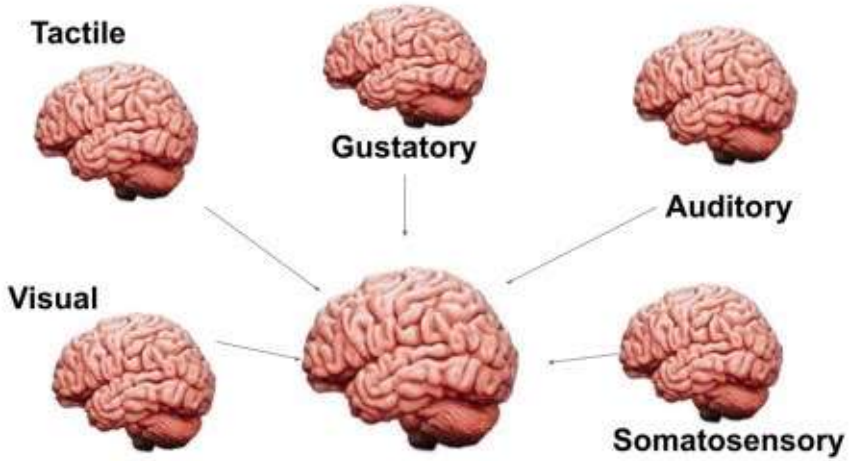
# Why is invariant coding useful?




Once an invariant representation is learned, learning to recognize new objects is much easier provided they preserve their identity under similar transformations: this is key in fast recognition.

# Other examples of perceptual persistence

In general:object identity $\rightarrow$ "sensory objects"; viewpoint $\rightarrow$ transformations.



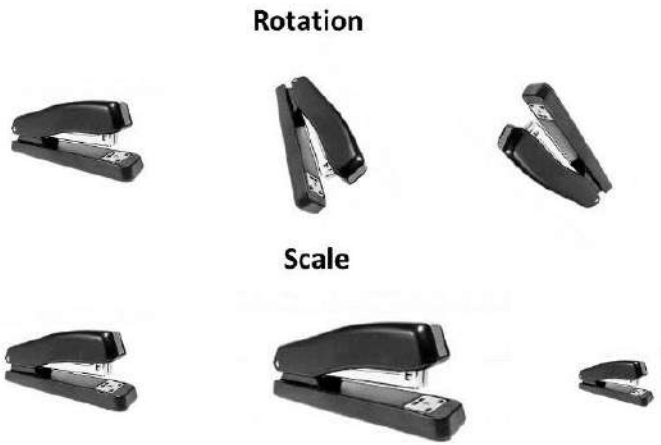

**Q**: How can the brain learn representations that are invariant to most frequent transformations that preserve the sensory object identity?

**A**: My work shows that the brain learns statistical regularities of the sensory input and use them to build invariant representations through **invariant detectors**.

# Invariance w.r.t. to what?: group transformations

Group: set $\mathcal{G}$ with composition/multiplication operation satisfying:

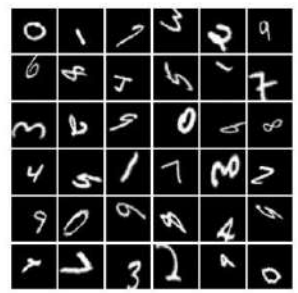
1. closure: $gg^{'} \in \mathcal{G}, \ \forall \ g, g^{'} \in \mathcal{G}$
2. associativity: $(gg^{'})g^{''} = g(g^{'}g^{''}) \in \mathcal{G}, \ \forall \ g, g^{'}, g^{''} \in \mathcal{G}$
3. identity: there exists $Id \in \mathcal{G}$ such that $Idg = gId = g \ \forall \ g \in \mathcal{G}$
4. invertibility: $\forall \, g \in \mathcal{G}$ there exists $g^{-1} \in G \ : \ gg^{-1} = Id.$



# Examples of Transformations/Groups

- Compact: rotation
- Locally compact: translation, scaling (multiplication), affine
  e.g., Translation group: linear operator $T_\tau : \mathcal{X} \to \mathcal{X}$

$$T_\tau x(p) = x(p - \tau), \quad \forall p, \tau \in \mathbb{R}, x \in \mathcal{X}$$



- Non-locally compact: diffeomorphisms, local/global deformations
  e.g. given a smooth map $d : \mathbb{R} \to \mathbb{R}$: linear operator $D_d : \mathcal{X} \to \mathcal{X}$

$$D_d x(p) = x(d(p)), \quad \forall p \in \mathbb{R}, x \in \mathcal{X}$$

- Non-group transformations
  - e.g. 3D rotations in depth, illumination.

# Defining invariance w.r.t. group transformations of stimuli: equivalence classes

Let $\mathcal{G}$ a group. Let $X$ set is a collection of orbits of stimuli:

$$O_t = \{t' \mid t' = gt,\ g \in \mathcal{G},\ t \in \mathcal{X}\}, \qquad X = (O_{t_1}, \cdots, O_{t_M})$$

The group $\mathcal{G}$ induces a **partition on the data**:

$$\boxed{x' \sim x \ \Leftrightarrow\ x' = gx, \quad \exists g \in \mathcal{G}}$$

# Invariant and Selective Representations

Let $\Phi : \mathcal{X} \to \mathcal{F}$.

Invariance:

$$x' \sim x \Rightarrow \Phi(x') = \Phi(x), \quad \forall x, x' \in \mathcal{X}$$

$$x' = gx \Rightarrow \Phi(x') = \Phi(x), \quad \forall x, x' \in \mathcal{X}$$

(trivial invariant representations are possible, e.g. constant function)

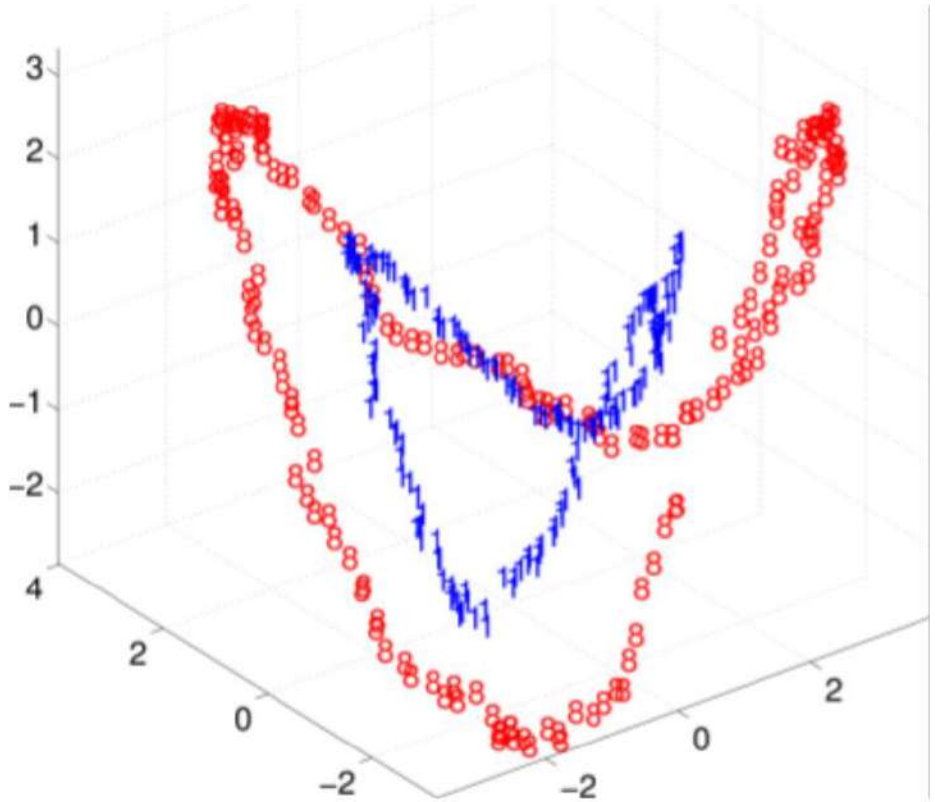
Selectivity:

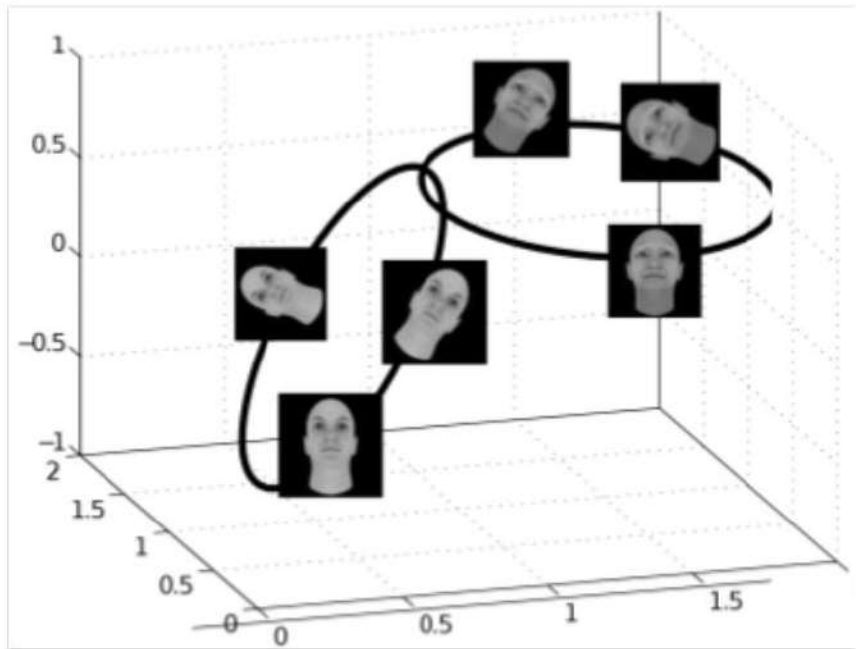
$$x' \sim x \Leftarrow \Phi(x') = \Phi(x), \quad \forall x, x' \in \mathcal{X}$$

$$x = gx' \Leftarrow \Phi(x) = \Phi(x'), \quad \forall x, x' \in \mathcal{X}$$

How to build a representation that is invariant and selective?

# Orbits are unique and invariant




If we have a representation of orbits it defines and invariant and selective representations of signals/stimuli/images.

# Invariant representations: Haar measure

**Example (Translation group/Lebesgue measure)**

$$d(T_{s'}s) = d(s - s') = ds$$

Translation invariance of the integral:

$$\int f(s)ds = \int f(s - s')ds$$

Key observation: a (locally) compact group has associated an **invariant Haar measure** $dg$ i.e.

$$d(gg') = dg, \quad \forall g' \in \mathcal{G}$$

# Building invariants by group averaging (1)

Haar invariance allows to build invariants in the form of group averages:

$$I_f = \int dg f(g) = \int dg f(ge) \equiv I_f(e)$$

In fact

$$\begin{aligned} I_f(\bar{g}) = & \int dg f(g\bar{g}) \\ = & \int d(\hat{g}\bar{g}^{-1}) f(\hat{g}) \qquad \hat{g} = g\bar{g}, \quad g = \hat{g}\bar{g}^{-1} \\ = & \int d\hat{g} f(\hat{g}), \quad d(\hat{g}\bar{g}^{-1}) = d\hat{g} \\ = & \int d\hat{g} f(\hat{g}) = I_f(e) \end{aligned}$$

where we used: 1) reparameterization of group elements 2) Haar invariance 3) group closure under composition.

How can we build from this observation an invariant representation for a stimulus?

# Choosing a specific representation

We choose $f$ as:

$$f_{x,t}(g) = \eta \langle gx, t \rangle$$

where $\eta : \mathbb{R} \to \mathbb{R}$ and $t \in \mathcal{X}$.

This choice is good for:

- Biological plausibility (we will see later)
- Allows to build an invariant representation of $x$

# Invariant representations by group averaging (2)

**Theorem**

*Invariance* $\mu_t : \mathcal{X} \to \mathbb{R}$ *defined as*

$$\mu_t(x) = \int dg \;\; \eta \left\langle gx, t \right\rangle, \quad x, t \in \mathcal{X}$$

*is invariant w.r.t. transformations* $x \to \bar{g}x, \;\; \forall \bar{g} \in \mathcal{G}$

Indeed

$$\begin{aligned}
\mu_t(\bar{g}x) &= \int dg \; \eta \left\langle g\bar{g}x, t \right\rangle \\
&= \int dg \; \eta \left\langle \hat{g}x, t \right\rangle \qquad \hat{g} = g\bar{g}, \quad g = \hat{g}\bar{g}^{-1} \\
&= \int d\hat{g} \; \eta \left\langle \hat{g}x, t \right\rangle, \quad d(\hat{g}\bar{g}^{-1}) = d\hat{g} \\
&= \mu_t(x)
\end{aligned}$$

# Enough projections ensure selectivity

If we have enough $t$-projections we can discriminate signals s.t. $x \not\sim x'$

Define a metric given $K$ templates/projections:

$$D_K(x, x') = \frac{1}{K} \sum_{k=1}^{K} |\mu_{t_k}(x) - \mu_{t_k}(x')|.$$

A simple concentration inequality shows that:

$$|D_K(x, x') - D_\infty(x, x')| \leq \epsilon, \quad \text{with probability } 1 - \delta^2$$

if $K > (\frac{2}{c}\epsilon^{-2}) \ln(\frac{Q}{\delta})$ with $Q$ the number of equivalence classes.

This can be interpreted as a random projections/Johnson-Lindenstrauss type theorem but on 1) On quotient space w.r.t. $\mathcal{G}$. 2) With non-linear projections.

$$\mu_t(x) = \int dg \;\; \eta \langle gx, t \rangle$$

Do we need the full orbit of $x$?

# Transferability and sample complexity

$\mu_t$ uses full orbit of $x$ to build an invariant representation. However:

$$\langle gx, t \rangle = \langle x, g^{-1}t \rangle, \forall x, t \in \mathcal{X}$$

Then we have mathematical equivalence of

- $\mu_t$ calculated having $O_x$
- $\mu_t$ calculated having $O_t$

i.e.

$$\mu_t(x) = \int dg\, \eta \langle gx, t \rangle = \int dg\, \eta \langle x, g^{-1}w \rangle = \int dg\, \eta \langle x, gt \rangle$$

We need only one example of $x$ and a set of transformed templates (memorized or learned) to have an invariant representation of $x$.

# Main result

## Theorem (Invariance and Selectivity for Finite Templates)

*For $Q$ orbits of $\mathcal{G}$ in $\mathcal{X}$, with $K > (\frac{2}{c}\epsilon^{-2}) \ln(\frac{Q}{\delta})$ templates, $\epsilon, \delta > 0$, the representation*

$$\Phi(x) = (\mu_1(x), \cdots, \mu_K(x))$$

*is*

- **Invariant** *i.e.* $\Phi(gx) = \Phi(x),\ \forall g \in G$
- **(Almost) selective** *i.e.* $|D_K(x, x') - D_\infty(x, x')| \leq \epsilon$ *with probability* $1 - \delta^2$.

In other words:

$$\Phi(x) = \Phi(x') \iff x \sim x'$$

# Algorithmic remarks

$$\mu_t(x) = \int dg\eta \langle x, gt \rangle$$

- We consider a finite group: $\mathcal{G} = \{g_1, \cdots, g_N\}$.
- Activation function: step functions, sigmoids, max, moments, with different thresholds e.g.
$$\eta_b \langle gx, t \rangle = \sigma(\langle x, gt \rangle - b)$$
- Templates are generic (not reflecting data distribution) and finite $\mathcal{T} = \{t^k\}_{k=1}^K$.

$$\mu_b^k(x) = \sum_i \sigma(\langle g_i x, t_k \rangle - b)$$

# Algorithmic steps

Template sampling $t \in \mathcal{T} = \{t^k\}_{k=1}^K$

Projections on transforming templates $\{\langle x, g_j t^k \rangle\}, j = 1, \ldots, |G|$

Pooling (sum) using $B$ non-linear functions (e.g., sigmoids)

$$\mu_b^k(x) = \sum_{j=1}^{|G|} \eta_b \langle x, g_j t^k \rangle, t^k \in \mathcal{T}, g_j \in G, b = 1 \ldots B$$

Signature from components

$$\Phi(x) = (\{\mu_1^1(x)\}, \{\mu_2^1(x)\}, \ldots, \{\mu_N^K(x)\}) \in \mathbb{R}^{N \times |\mathcal{T}|}$$

Number of templates (to separate $|\hat{Y}|$ classes with probability $1 - \delta^2$)

$$K \geq \frac{2}{c\epsilon^2} \log \frac{|\hat{Y}|}{\delta}$$

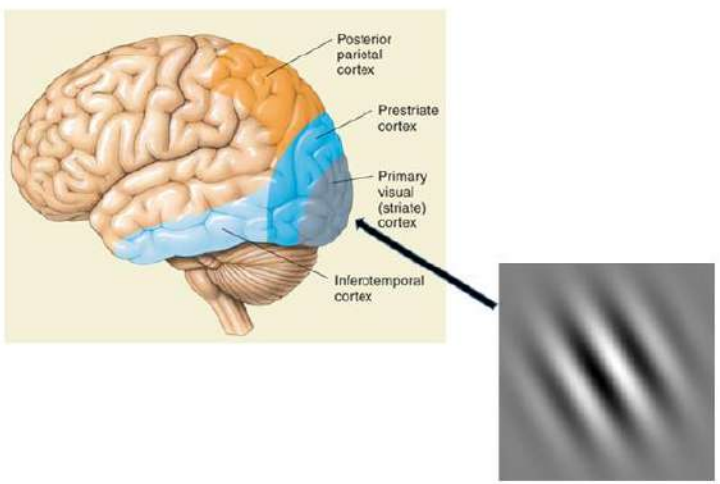
# What if the group is locally compact? approximate invariance and Gabor templates

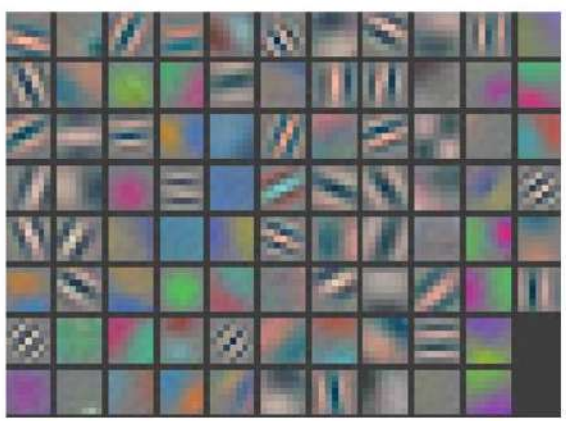
Suppose we want to implement translation and scale invariance

$$\mu_t(x) = \sum_{\lambda,y} \sigma \langle x, D_\lambda T_y t \rangle$$

Which is the best choice for $t = t^*$ to guarantee approximate invariance?

$$t^* = \arg\min_{t,\forall\ \lambda,y} |\mu_t(D_\lambda T_y x) - \mu_t(x)|$$

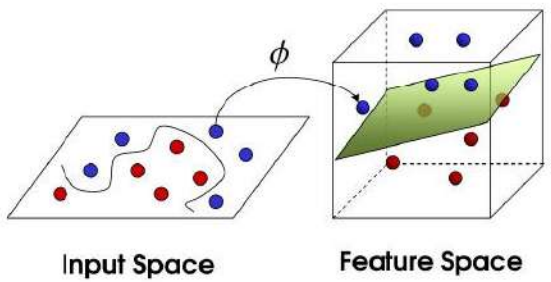


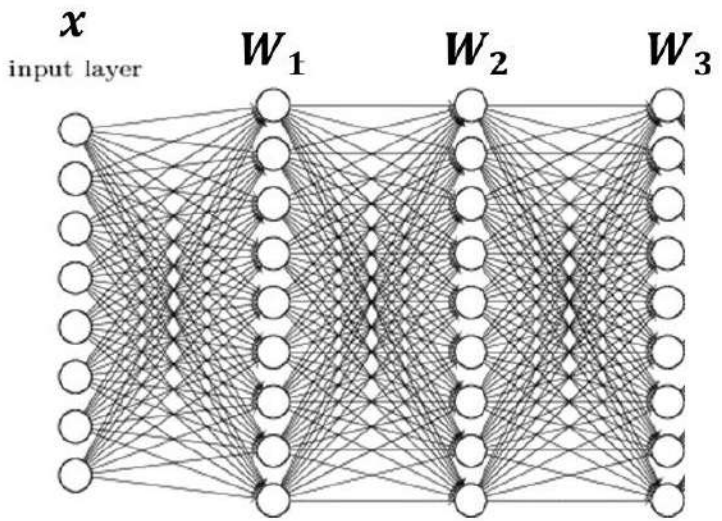



# Alternative and more ANN oriented demonstration

# How is this translated in the context of deep networks?

Deep networks: 'smart' way to parameterize functions.



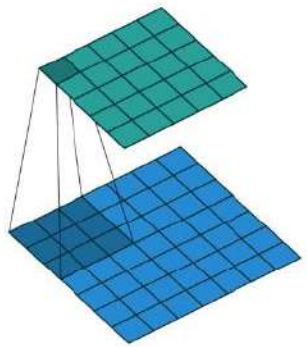

$$x \in \mathbb{R}^d, \quad \Phi : \mathbb{R}^d \to \mathbb{R}^L, \quad \Phi \equiv \Phi_L \circ \cdots \circ \Phi_1$$

$$\Phi_1(x) = \sigma \langle W_1, x \rangle$$
$$\Phi_\ell(x) = \sigma \langle W_\ell, \Phi_{\ell-1}(x) \rangle$$

$\sigma : \mathbb{R} \to \mathbb{R}$, nowadays $\sigma(\cdot) = max(0, \cdot) \equiv |\cdot|_+ \ (w \equiv t)$.

# Constraints on network operations/topology

$$W = \begin{bmatrix} w_1 & w_d & w_1 & \dots & w_2 \\ w_2 & w_1 & w_2 & \dots & w_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_d & w_{d-1} & w_{d-3} & \dots & w_1 \end{bmatrix} =$$

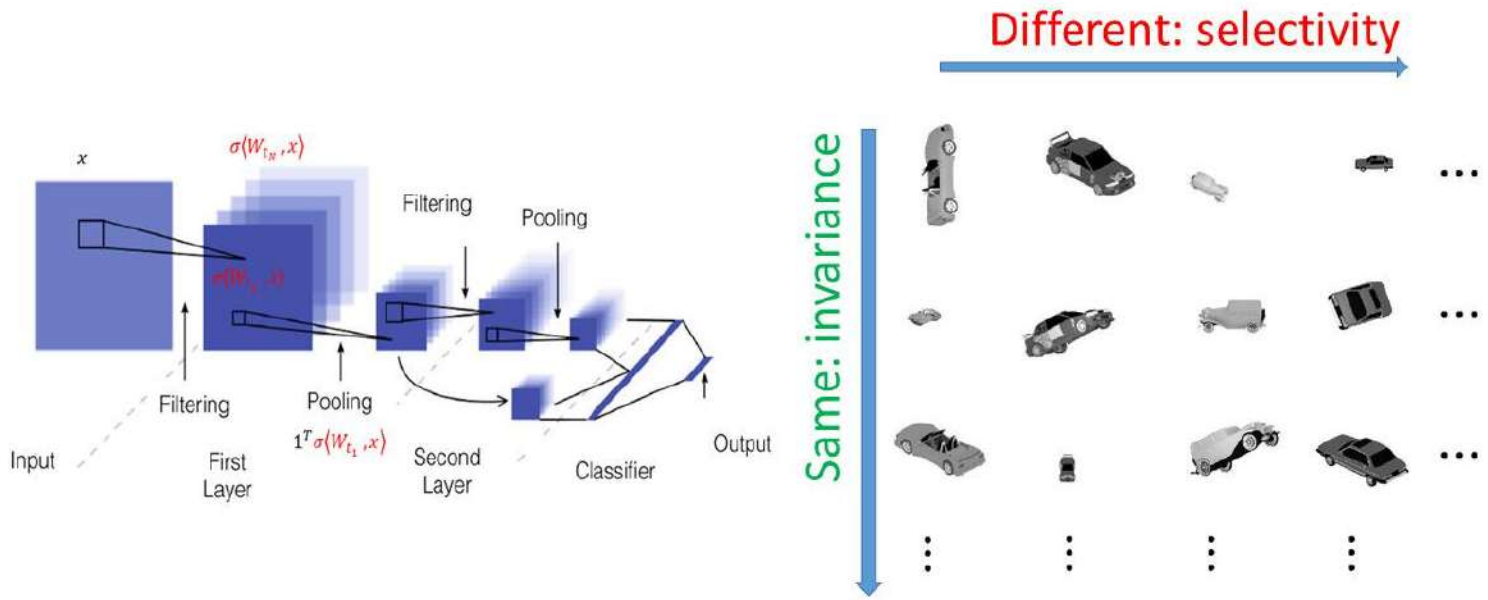


an orbit w.r.t. the cyclic group $\mathcal{G}$. This is equivalent to a convolution $x * w = W^T x$.

**Remark**

*This construction can be extended to convolutions w.r.t. other groups.*

# Invariant selective Data Representation




## Theorem

*Suppose data are generated by $\mathcal{G}$. A CNN with convolutions w.r.t. $\mathcal{G}$ is implementing a data representation $\Phi$ that is invariant and selective i.e.* $\mathbf{x} \sim \mathbf{x}' \Leftrightarrow \mathbf{\Phi}(\mathbf{x}) = \mathbf{\Phi}(\mathbf{x}')$ *with (one layer)* $\Phi : \mathbb{R}^d \to \mathbb{R}^p$ *and:*

$$\Phi_w(x) = \sum_i |(x *_{\mathcal{G}} w)_i|_+ = \sum_i |(W^T x)_i|_+ = \sum_i |\langle x, g_i w \rangle|_+$$

# Orbits and probability distribution: some definitions

Consider the map $x \to gx$ where $g$ is the random variable. This defines a probability distribution $x \to P_x(g)$ the probability distribution supported on each orbit

$$O_x = \{x' \mid x' = gx,\ g \in \mathcal{G}\}.$$

How do we compare different distributions $P_x, P_{x'}$? We use arbitrary (continuous) test functions $f \in C_c(\mathcal{X})$ and the use the following definition

$$I(f, P_x) \equiv \int_{O_x} f(y) dP_x(y) = \int_{\mathcal{X}} f(y) dP_x(y) = \int_{\mathcal{G}} f(gx) dg,\ \ \forall\ f \in C_c(\mathcal{X})$$

with $dg$ the Haar measure on the group. Then $P_x = P_{x'}$ iff $I(f, P_x) = I(f, P_{x'})$ for all $f$.

# Demonstration steps

We sketch the demonstration in 3 steps:

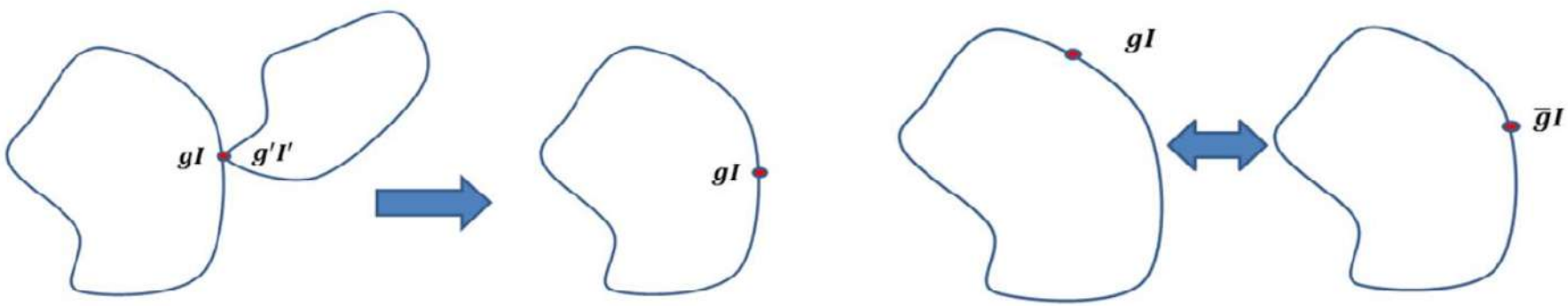
1. $x \sim x' \iff O_x = O_{x'}$
2. $O_x = O_{x'} \iff P_x = P_{x'}$
3. $P_x = P_{x'} \iff \Phi(x) = \Phi(x')$

where $\Phi$ is the CNN representation.

# Equivalence classes and orbits

$$x \sim x' \iff O_x = O_{x'}$$

1. $x \sim x' \Rightarrow x = gx' \Rightarrow O_x = O_{gx'} = O_{x'}$
2. $O_x = O_{x'} \Rightarrow x = gx' \Rightarrow x \sim x'$



# Equivalence Orbits-probability distribution

**Theorem**

$O_x = O_{x'} \Leftrightarrow P_x = P_{x'}$

$\Rightarrow$: If $O_x = O_{x'} \Rightarrow\ x = hx', \exists h \in \mathcal{G}$

Then using the definition

$$\int_{O_x} f(y)dP_x(y) = \int_{\mathcal{G}} f(gx)dg = \int_{\mathcal{G}} f(ghx')dg$$
$$= \int_{\mathcal{G}} f(gx')dg = \int_{O_{x'}} f(y)dP_{x'}(y)$$

for all $f$ and thus $P_x = P_{x'}$.

$\Leftarrow$: If $P_x = P_{x'} \Rightarrow\ supp(P_x) = supp(P_{x'}) \Rightarrow\ O_x = O_{x'}$.

# Mean probability embedding (3)

## Theorem (Probability mean embedding, Smola et al. '07)

*Let* $x \to P_x(g)$ *induced by* $\mathcal{G}$ *and* $\Psi(x)$ *be a "rich enough" representation* [a]. *Then*

$$P_x = P_{x'} \quad \Leftrightarrow \quad \Phi(x) = \int dg\ \Psi(gx) = \int dg\ \Psi(gx') = \Phi(x')$$

*i.e. the distribution is one to one with the representation map.*

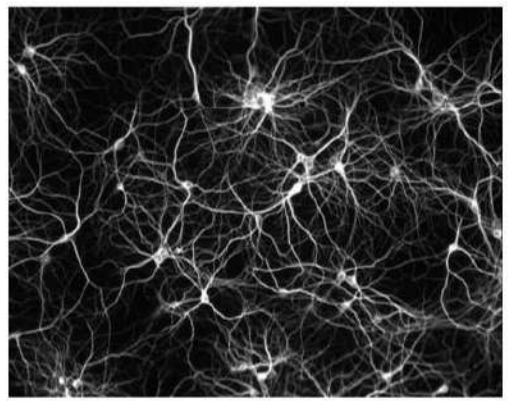
*In the case of CNNs* $(\Psi(g_i x))_{w,i} = |\langle x, g_i w \rangle|_+ = |(x * w)_i|_+$ *and*

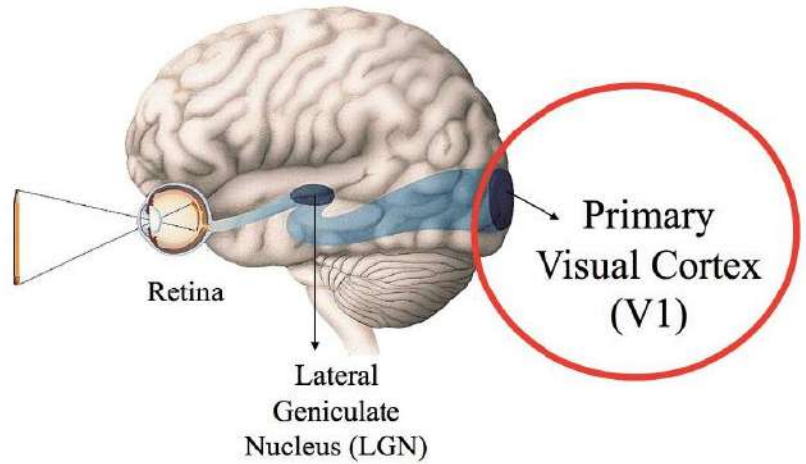
$$(\Phi(x))_w = \sum_i |\langle W_t, x \rangle_i|_+ = \sum_i |\langle g_i w, x \rangle|_+$$

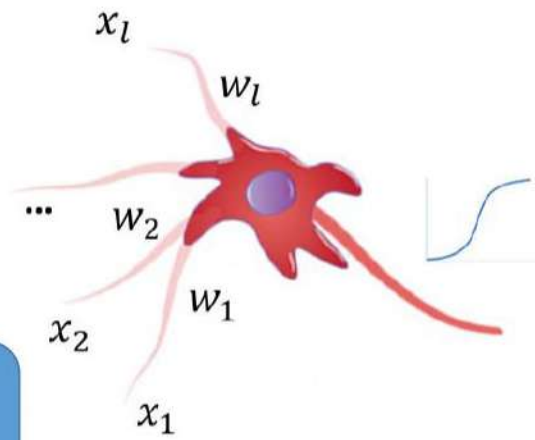
[a] The corresponding kernel universal, i.e. dense in the space of continuous functions

# Neuronal interpretation of $\mu_t$: Simple-complex model of visual cortex

# Single neuron operation



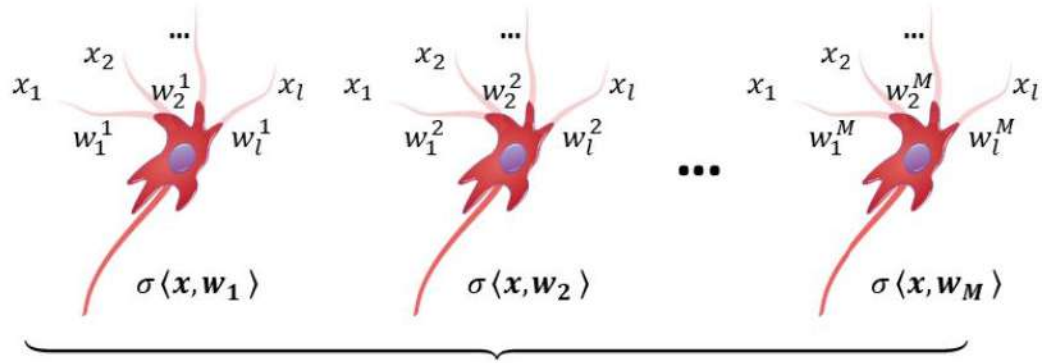


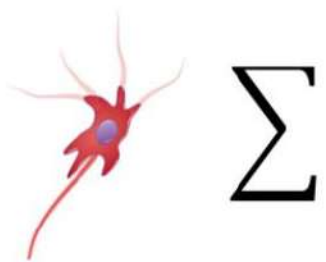




$\sigma\langle x, w\rangle$

- Scalar product
- Threshold nonlinearity

# Aggregating single neurons responses: Hubel Wiesel model



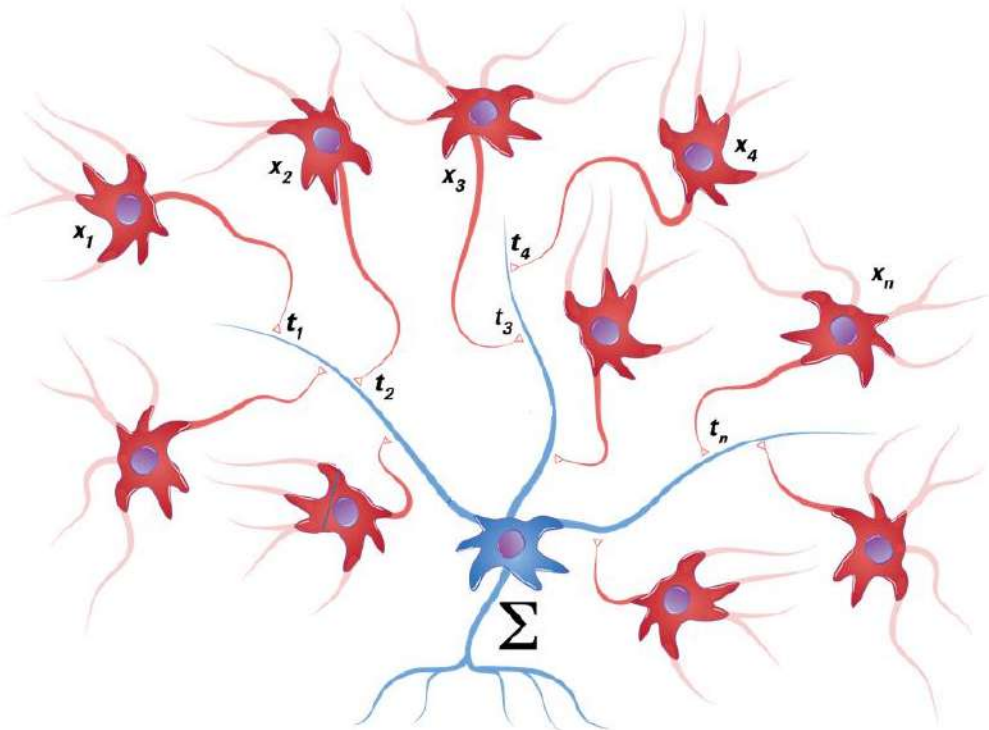



$$\sum$$

$$\mu_w(x) = \sum_{i=1}^{M} \sigma\langle x, w_i \rangle$$

This module is often called in visual cortex *Simple-Complex* and $\mu(x)$ is a neural representation of the input $x$.

# Translating our results

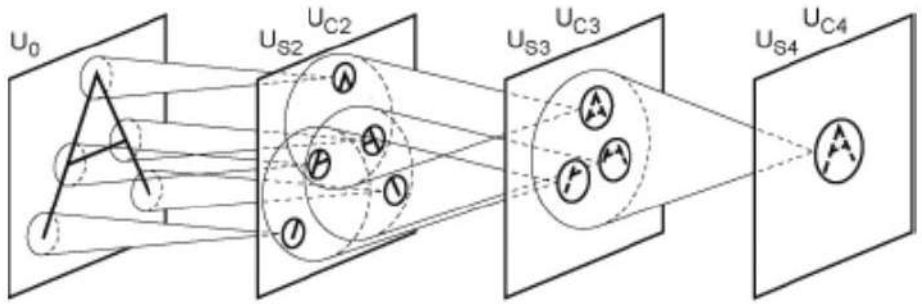



Acascade of simple and complex cells is implementing an invariant and selective representation.

# How the representation compare?

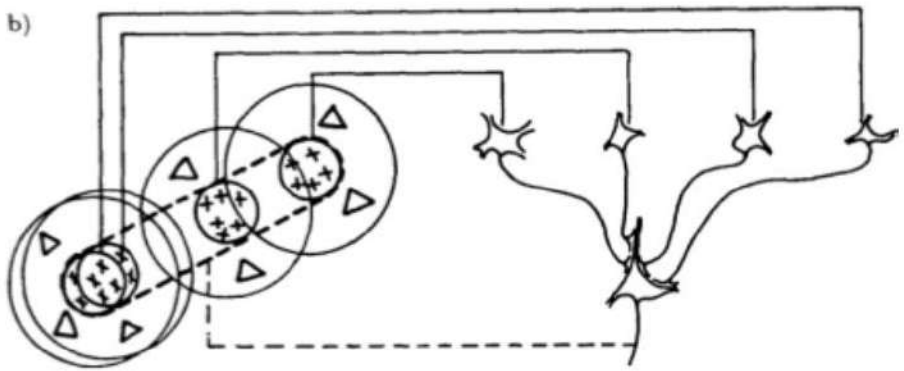
Convolutional networks: *(Fukushima, LeCun, Poggio)*

- Convolution filters: $\{\langle x, gw \rangle\}$
- Pooling: $\sum_g \sigma \langle x, gw \rangle$




V1 in visual cortex: *(Hubel and Wiesel)*

- Simple cells: $\{\langle x, gw \rangle\}$
- Complex cells: $\sum_g \sigma \langle x, gw \rangle$



# Robustness to more general transformations and locality

Let $s$ a $C^\infty$ transformation depending on $\Theta = (\theta_1, \ldots, \theta_P)$ parameters.
Expanding, e.g., around e.g. $0$:

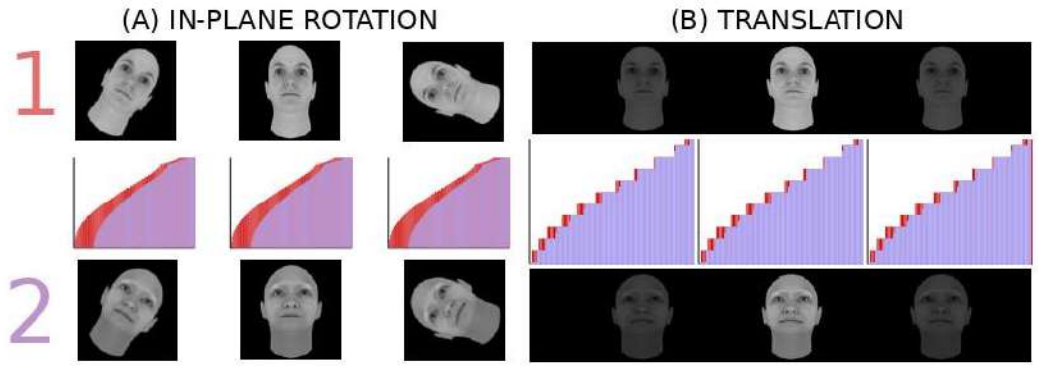
$$s(x, \Theta) = s(x, \mathbf{0}) + \sum_{i=1}^{P} \frac{\partial s(x, \Theta)}{\partial \theta_i} \theta_i + o(\|\Theta\|^2) = x + \sum_{i=1}^{P} \theta_i L_{\theta_i}(x) + o(\|\Theta\|^2)$$

where $L_{\theta_i}$ are the infinitesimal generators and Therefore locally

$$g(\Theta) = \exp(\theta_1 L_{\theta_1} + \theta_2 L_{\theta_2} + \cdots + \theta_P L_{\theta_P}).$$

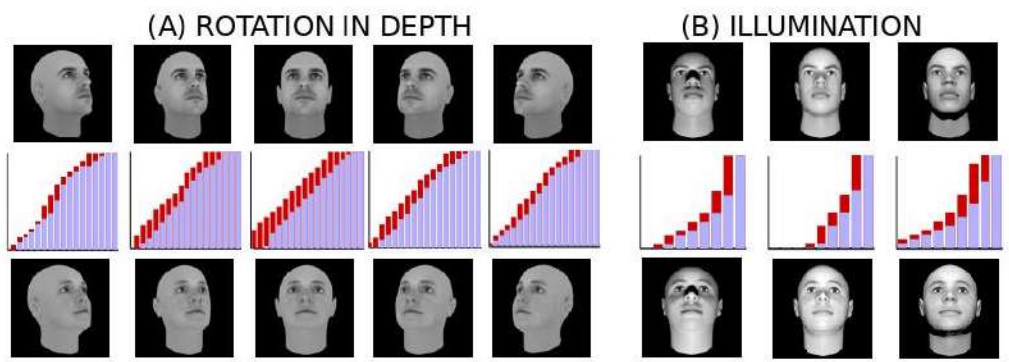
The local learned weights will be **orbits w.r.t. the local group approximating the non-group global transformation**.

# Beyond groups




$$\overline{\Phi}^t(x) = \sum_i |\langle g_i t, x \rangle|_+$$

As a processing procedure the principles extend beyond compact group transformations (Leibo et al. '15)



# Selectivity: probabilistic argument

(For compact groups) a probability distribution $P_x$ on $O_x$ is defined by measure $dg$ via random variable $g \mapsto gx$

**Theorem (Orbit-probability distribution equivalence)**

$$(x \sim x' \iff) \; O_x = O'_x \iff P_x \sim P_{x'}$$

"1D" probability distributions $P_x^t$ are defined by projections $gx \mapsto \langle gx, t \rangle$

**Theorem (Cramer-Wold)**

*Probability distributions are uniquely determined by all 1D projections*

$$P_x \equiv P_{x'} \iff P_x^t \equiv P_{x'}^t, \quad \forall t \in \mathbb{S}^d$$

1D projection can be characterized by its CDF $P_x^t \mapsto \int dg H(b - \langle x, gt \rangle)$

# Learning - Transferability

Unitary groups: $\langle gx, t\rangle = \langle x, g^{-1}t\rangle, \forall x, t \in \mathcal{X}$

$$\mu(x) = \int \eta\left(\langle gx, t\rangle\right) dg = \int \eta\left(\langle x, g^{-1}t\rangle\right) = \int \eta\left(\langle x, gt\rangle\right) dg$$

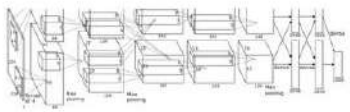
Memory-based learning:

- View based/augmentation: Need $\{gx\}, \forall g \in \mathcal{G}$ $(\forall x)$
- Memory-based: Store $\{gt\}, \forall g \in \mathcal{G}$ (once)
- Trading time with space requirements
- *Biological learning, learning from sequences, CNNs*

# Human like perspective of invariant representations

**Few examples**

**Many examples**



Learning from few examples is more human like.

# Class 2

- Learning invariant and selective representations: hebbian learning, simple cells, complex cells.

- Example: mirror symmetric tuning of neurons in the face patch.

- Class-specific invariance.

# Selective and invariant signature (recap)

Signature:

$$\mu_t(x) = \int dg\eta \langle x, gt \rangle$$

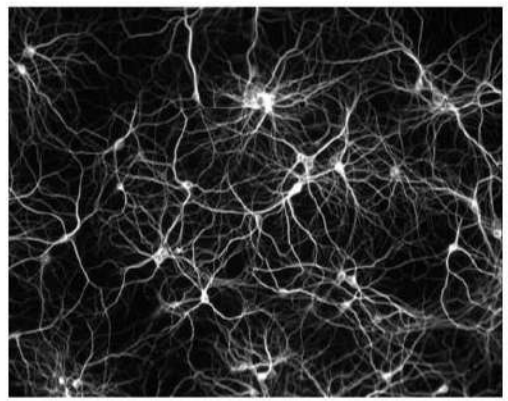
Or its computational-friendly form

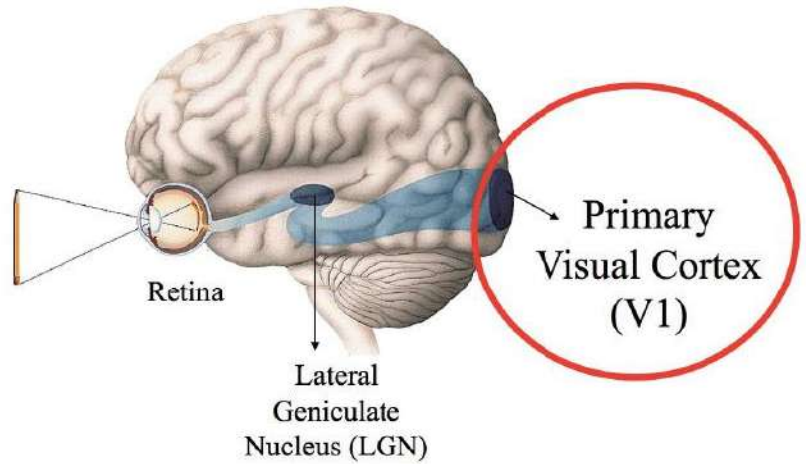
$$\mu_b^k(x) = \sum_i \sigma(\langle x, g_i t_k \rangle - b)$$

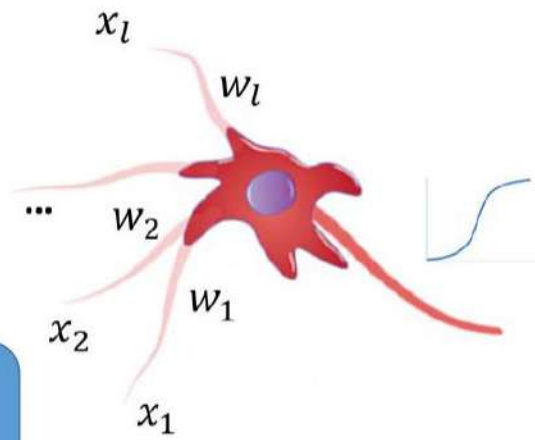
The signature is invariant and selective.

# Neuronal interpretation of $\mu_t$: Simple-complex model of visual cortex

# Single neuron operation



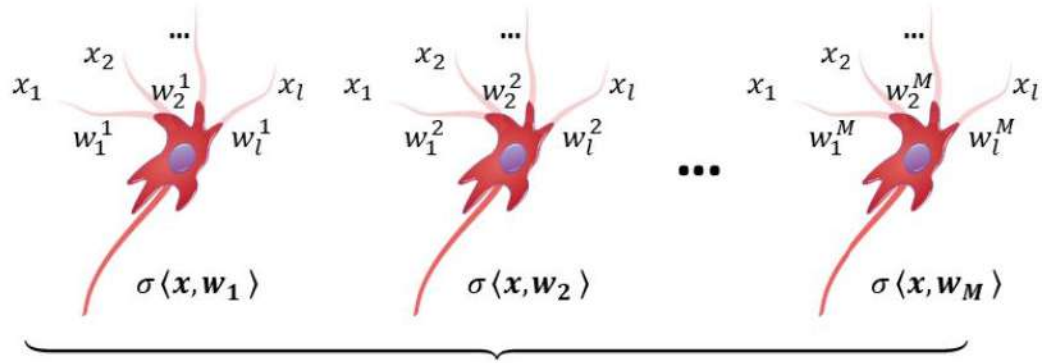


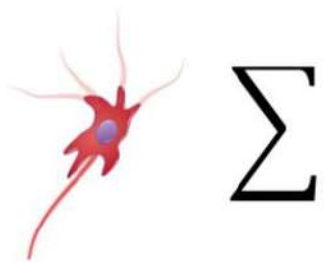




$\sigma\langle x, w\rangle$

- Scalar product
- Threshold nonlinearity

# Primary visual cortex computations: Hubel Wiesel model



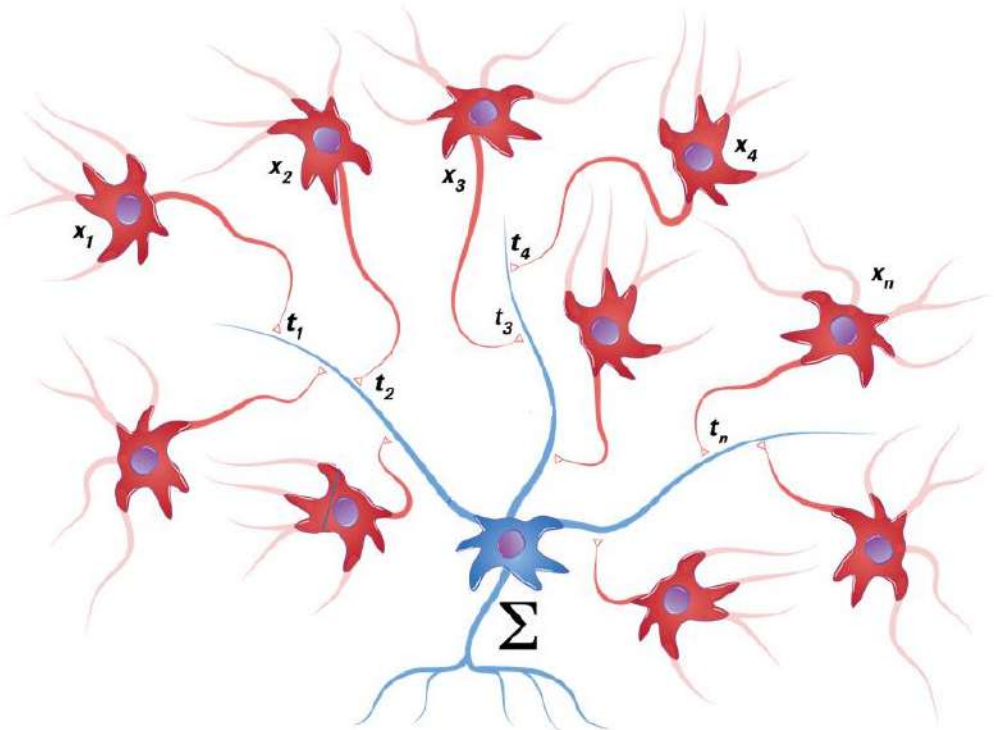



$$\sum$$

$$\mu_w(x) = \sum_{i=1}^{M} \sigma\langle x, w_i \rangle$$

This module is often called in visual cortex *Simple-Complex* and $\mu(x)$ is a neural representation of the input $x$.

# Translating our results




When $w_i = g_i w$ a cascade of simple and complex cells is implementing an invariant and selective representation.

# Learning invariant and selective representations: hebbian learning

# From previous class: stimuli hypothesis

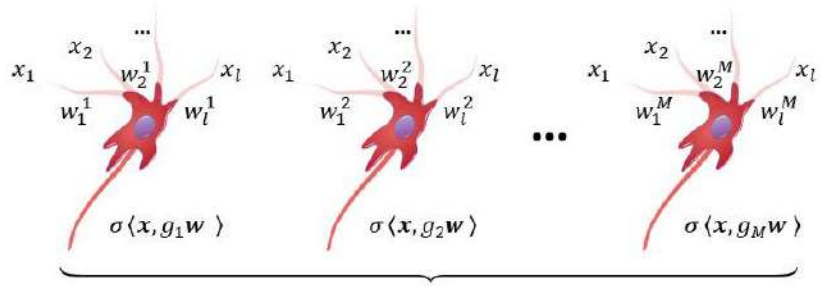
Let $\mathcal{G}$ a group. Let $X$ set is a collection of orbits of stimuli:

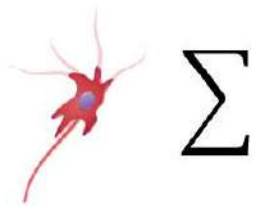
$$O_t = \{t' \mid t' = gt,\ g \in \mathcal{G},\ t \in \mathcal{X}\}, \qquad X = (O_{t_1}, \cdots, O_{t_M})$$

The group $\mathcal{G}$ induces a **partition on the data**:

$$\boxed{x' \sim x \ \Leftrightarrow\ x' = gx, \quad \exists g \in \mathcal{G}}$$

# Learning question



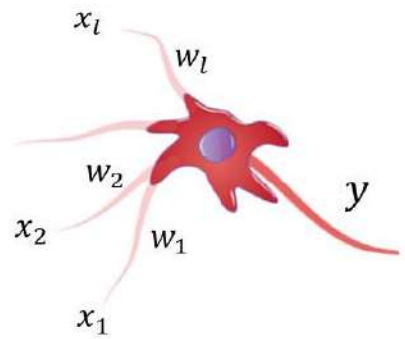
$$\sum$$

$$\mu_w(x) = \sum_{i=1}^{M} \sigma\langle x, g_i w\rangle$$

How can we learn $g_i w$? How can we learn the correct pooling over orbits?

# Learning hypothesis: Hebbian learning of simple neurons synaptic weights

Neurons updates their weights during stimuli presentation. We consider a weights-normalized version of the hebbian rule $\Delta \mathbf{w} = \eta y_t x$

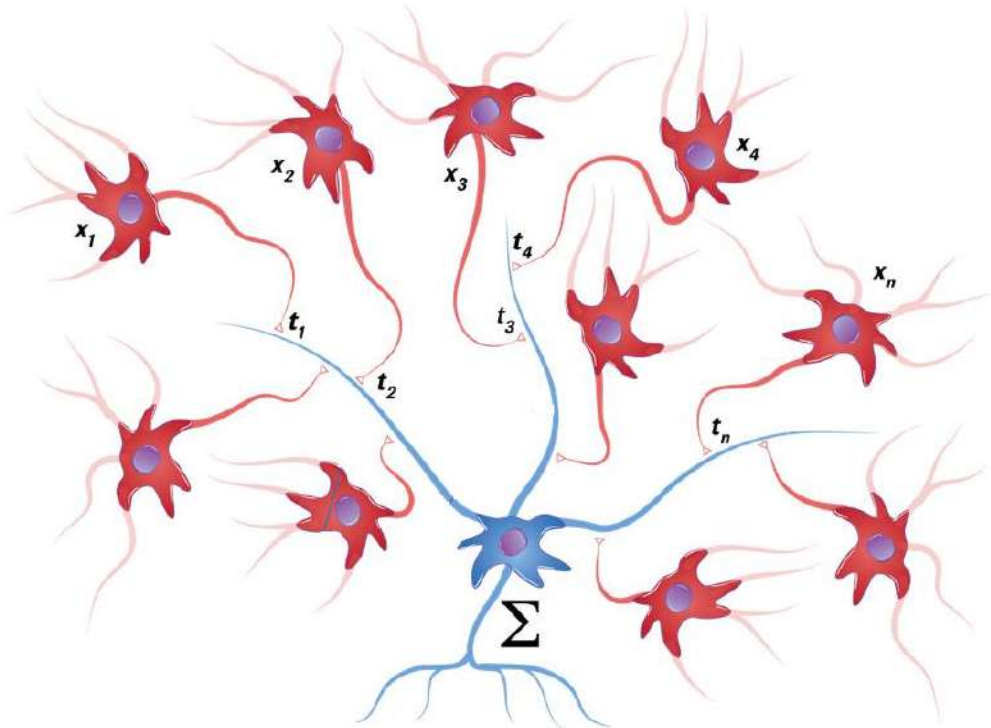



$$y_t = \langle \mathbf{x}_t, \mathbf{w}_t \rangle, \qquad \Delta \mathbf{w} = \eta y_t (\mathbf{x}_t - y_t \mathbf{x}_t)$$

$$XX^T \mathbf{w} = \lambda_{max} \mathbf{w}$$

**Simple cells weights converge to the first eigenvector of the covariance of the stimuli.**

# Question




Can a simple-complex module, $\mu_w(x)$, under the hypotheses above, learn invariant/selective properties w.r.t. $\mathcal{G}$?

# Stimuli + learning: simple neurons weights

**Theorem**

*The weights of simple cells converge to linear combinations of elements of an orbit of $\mathcal{G}$*

Simple proof:

- $X = \{g_1 t_1, \cdots, g_{|\mathcal{G}|} t_1, \cdots, g_1 t_Q, \cdots, g_{|\mathcal{G}|} t_Q\}$

$$C = XX^T = \sum_i g_i T T^T g_i^T, \quad [C, g] = Cg - gC = 0$$

and

$$Cw = \lambda w \;\; \Rightarrow \;\; Cgw = gCw = \lambda gw, \; \forall \, g \; \in \mathcal{G}, \; w \in E_\lambda$$

- In particular

$$\mathbf{E_{max}} = \mathbf{span}(\mathbf{O_w}), \;\; \forall w \in E_{max}.$$

# Extension: equivariant plasticity rules

The result is true for a much broader class of plasticity rules those deriving from a Loss function of the form:

$$\mathcal{L}(X, W) = \sum_{i,j} f(\sigma \langle w_i, x_j \rangle), \qquad f \in \mathcal{C}^1.$$

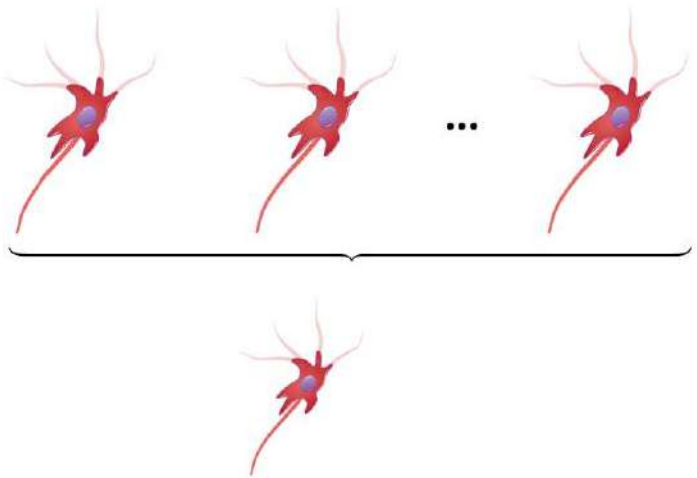
We have that the associated plasticity rule is equivariant w.r.t. $\mathcal{G}$ transformations:

$$\nabla_{w_i} \mathcal{L}(\{w_1, \cdots, g w_i, \cdots, w_N\}, X) = g \nabla_{w_i} \mathcal{L}(\{w_1, \cdots, w_i, \cdots, w_N\}, X).$$

If $\mathbf{w}$ is solution so is $g_i \mathbf{w}$.

# Which simple cells is a complex cell aggregating?

Suppose simple weights are learned.



$$O_x = \{x, g_2 x, \cdots, g_M x\}$$
$$\bar{w} = \arg\max_w \langle w, x \rangle$$
$$g_2 \bar{w} = \arg\max_w \langle w, g_2 x \rangle$$
$$\langle x, \bar{w} \rangle = \langle gx, g\bar{w} \rangle$$

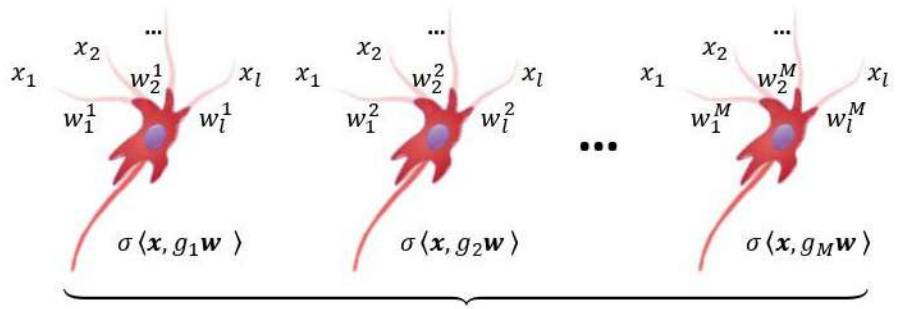
The principle is Hebbian: "fire together link together".

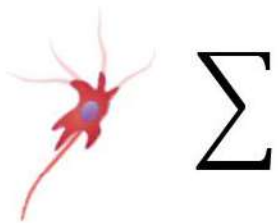
**Theorem (Complex cells pooling and invariance)**

*A complex cell learns to aggregates over simple cells whose weights form an orbit with respect to the group $\mathcal{G}$ i.e. $w^i = g_i w$.*

# Summarizing:

1. The symmetries in the stimuli are "inherited" by the weights of the simple cells: $\mathbf{E_{max}} = \mathbf{span}(\mathbf{O_w}), \ \ \forall w \in E_{max}.$
2. The set of simple cells aggregated by a complex cell have weights that are orbits i.e. $\mathbf{w^i} = \mathbf{g_i w}.$



$$\sum$$

$$\mu_w(x) = \sum_{i=1}^{M} \sigma\langle x, g_i w\rangle$$

# Simple-complex modulus equivariant-invariant properties

- Simple cells are permutation-equivariant to $g \in \mathcal{G}$ transformations.

$$\sigma(W^T g x) = P_g \sigma(W^T x), \quad W = (g_1 w, \cdots, g_{|\mathcal{G}|} w)$$

- Complex cells are invariant to $g \in \mathcal{G}$ transformations.

$$\mu_{\mathbf{w}}(\mathbf{x}) = \sum_i \sigma \langle x, g_i w \rangle = \mu_{\mathbf{w}}(\mathbf{gx}) \quad \forall\, g \in \mathcal{G}$$

Same properties hold for a wild class of aggregating functions (e.g. max) and pointwise non-linearities.

# What about selectivity/discriminability?

- Invariance:

$$x' \sim_{\mathcal{G}} x \Rightarrow \mu(x') = \mu(x), \quad \forall x, x' \in X$$

  Elements within an orbit have the same $\mu$.

- Selectivity:

$$x' \sim_{\mathcal{G}} x \Leftarrow \mu(x') = \mu(x), \quad \forall x, x' \in X$$

  Elements belonging to different orbits have different $\mu$.

**We want to collapse elements within orbits and tear apart different orbits, i.e. build a proper quotient space.**

# Discriminability: importance of the non-linearity

The information loss can be recovered considering a family of non-linearities e.g. $\{\sigma_z(\cdot) \equiv H(\cdot - z),\ z \in \mathbb{R}\}$.

$$c_z(x) \equiv (c(x))_z = \sum_{i=1}^{|\mathcal{G}|} H(\langle x, g_i w \rangle - z), \quad z \in \mathbb{R}.$$

We have:

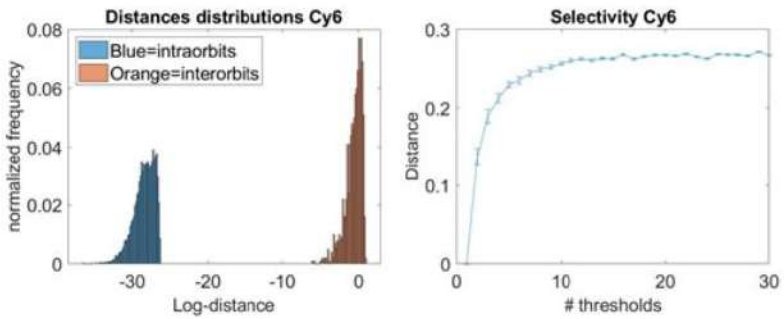
## Theorem (Complex cells selectivity)

*Let $x, x' \in \mathbb{R}^d$ be two input and $c(x), c(x')$ their complex cells responses. Then the distance defined as*

$$dist(x, x') := \|c(x) - c(x')\|_{\ell_2}.$$

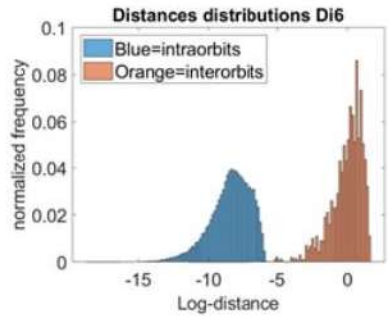
*is zero iff $x \sim x'$.*

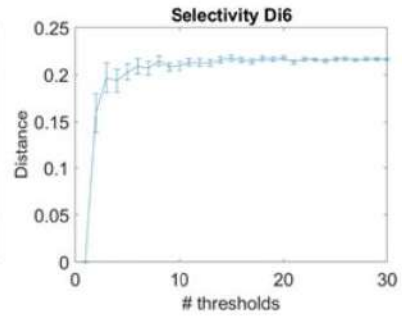
# Some simulations

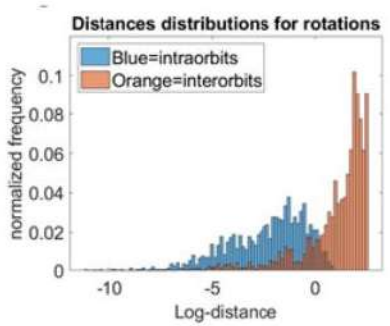


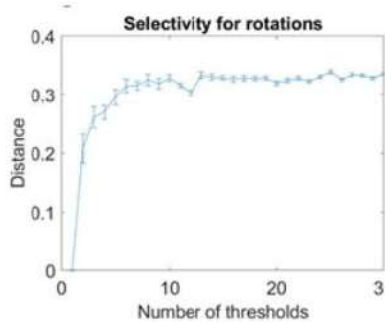






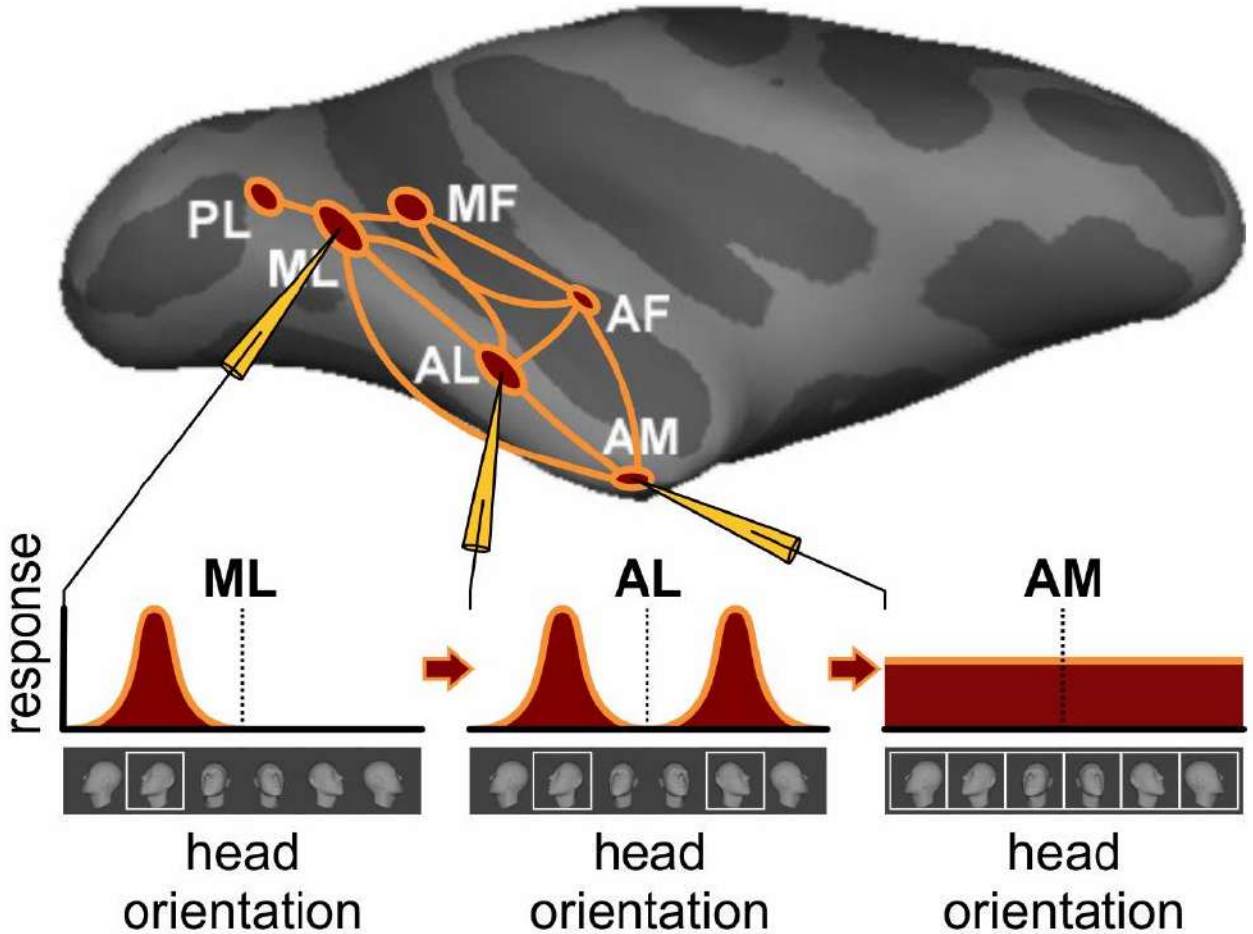



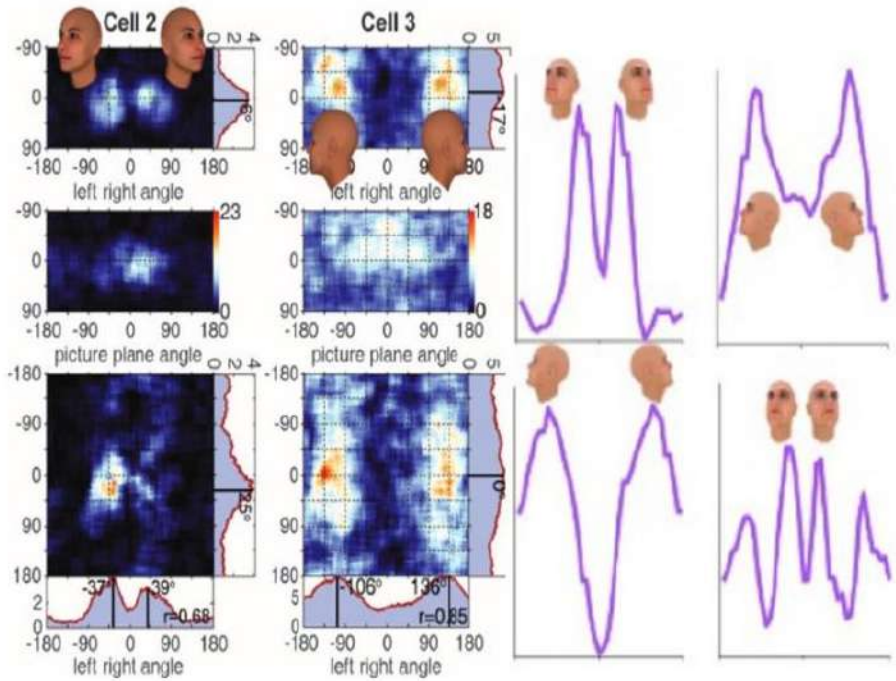

- $X$ generated by: 1) Cyclic group, 2) Dihedral group, 3) Finite group of rotations on natural image patches.
- Plots: Intra and inter orbit distances for $N$ orbits. Discriminability vs number of thresholds.

# An application of the theory: Mirror symmetric tuned cells in the face patch and their invariance properties

# Mirror symmetric tuned cells face patch (1)



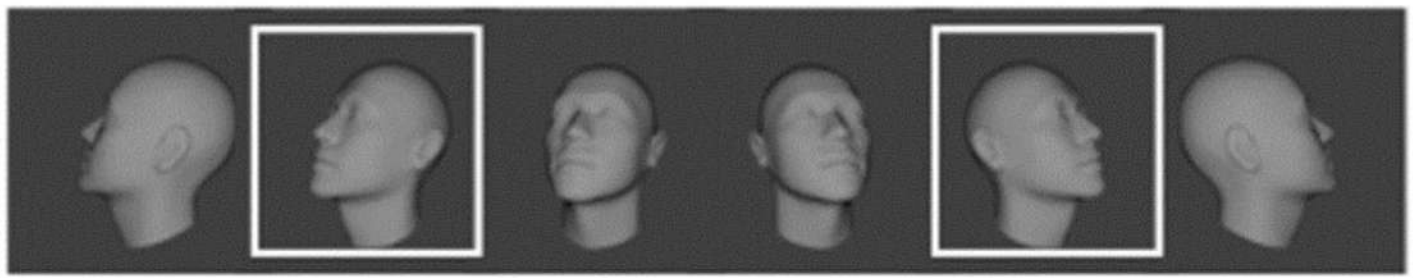
# Mirror symmetric tuned cells face patch (2)




Freiwald, Tsao, Science, 2010

# Mirror symmetry

Observation:



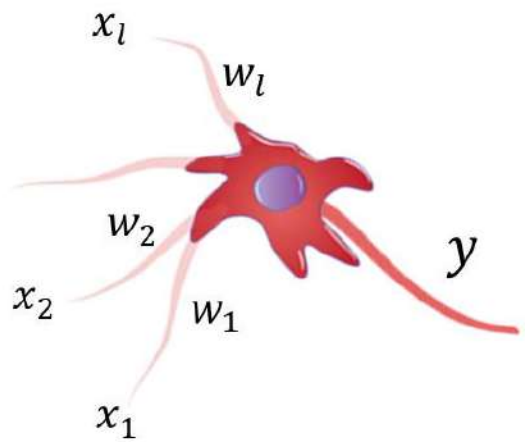
$f_i$ $Rf_i$

The input is a collection of orbits w.r.t. the reflection group $\mathcal{G} = \{e, R\}$:

$$D = \{(f_1, Rf_1), \cdots, (f_N, Rf_N)\}$$

# Oja's learning on orbits

Oja's learning:




$$w_i(t+1) = w_i(t) + \lambda y(x) x_i$$
$$||w|| = 1$$

$$w \to first\ PCA(X)$$

The covariance matrix of $D = \{(f_1, Rf_1), \cdots, (f_N, Rf_N)\}$ is

$$C = D^T D = FF^T + RFF^T R^T$$

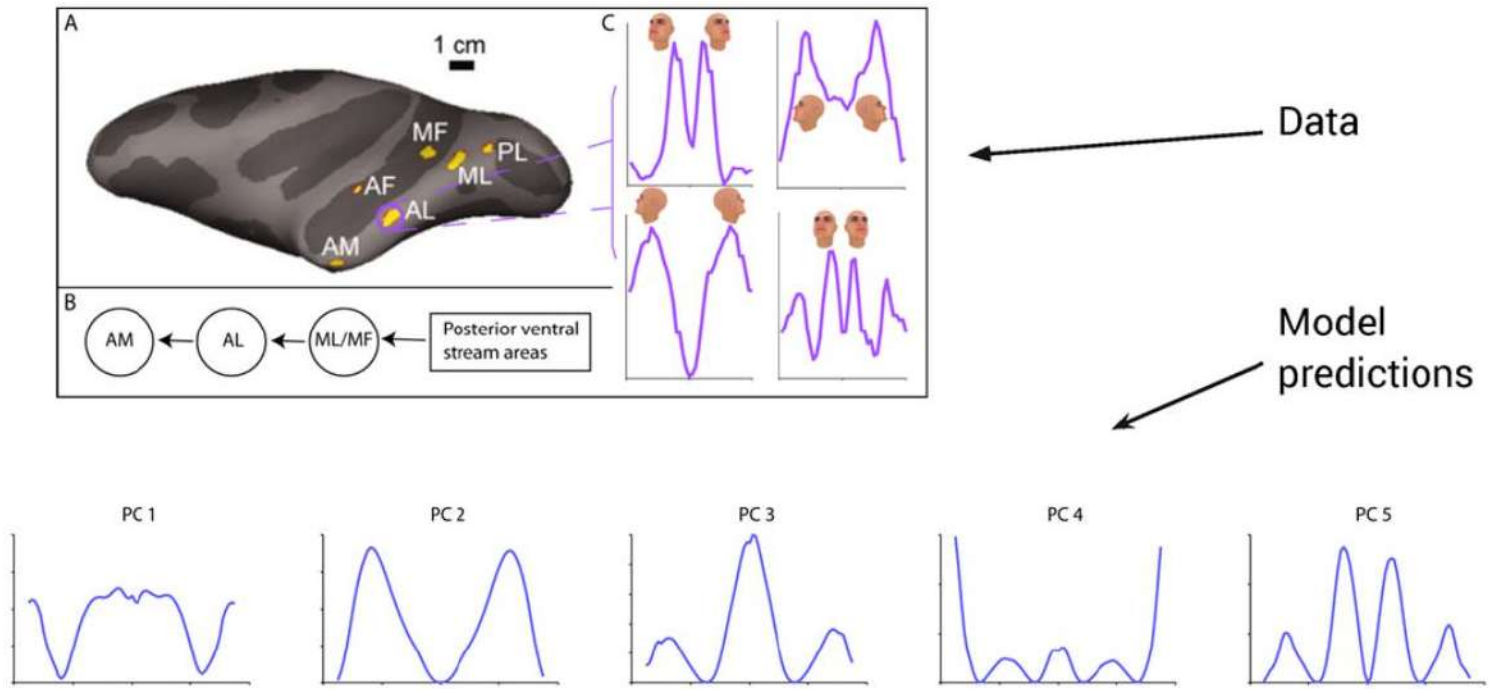
with $F = \{f, \cdots, f_N\}$ and we have

$$[C, R] = 0$$

i.e. *the eigenfunctions are such that* $Rw = w$ since $Rw$ and $w$ are eigenfunctions with the same eigenvalue.

# Mirror symmetric neuronal receptive fields

Calculating the eigenfunctions of the covariance matrix:




**View-invariance implies mirror symmetric orientation tuning curves**