Lecture 7. Unsupervised learning

Overview

- Hebbian postulate
- Introducing competition through normalisation
- Topographic map formation and ocular dominance.

Hebb's postulate

- Supervised learning requires a teacher to provide error signals; how far can the nervous system get on its own?
- "When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949).
- AKA "cells that fire together wire together" or "out of sync lose the link".
- Detect correlations [over some small time window, say 50ms] between firing of cells.

$$\Delta w_j = \epsilon y x_j \qquad \text{where } y = f(\mathbf{w} \cdot \mathbf{x})$$

Correlation and covariance matrices

Often we average over inputs $\mathbf{x}$, assuming inputs change quicker than synaptic weights $\mathbf{w}$.
Correlation matrix $\mathbf{Q}_{ij} = \langle x_i x_j \rangle$ or $\mathbf{Q} = \langle \mathbf{x}\mathbf{x}^T \rangle$.

Covariance matrix $\mathbf{C}_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$

Properties:

1. Real, symmetric matrix $\Rightarrow$ N orthogonal real eigenvectors.

2. Positive semi-definite: for any input $\mathbf{u}$, $\mathbf{u}^T\mathbf{C}\mathbf{u} \geq 0$.

3. All eigenvalues of a positive semidefinite matrix are non-negative.

Hebbian rule

Since we assume inputs change quicker than synaptic weights $\mathbf{w}$:

$$\begin{aligned}
\tau_w \frac{\mathrm{d}w_j}{\mathrm{d}t} &= \langle y x_j \rangle, \qquad y = \sum_i w_i x_i \\
&= \langle \sum_i w_i x_i x_j \rangle \\
&= \sum_i w_i \langle x_i x_j \rangle = \sum_i \mathbf{C}_{ji} w_i \\
\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} &= \mathbf{C}\mathbf{w}
\end{aligned}$$

Variants on Hebbian rule

$$\text{activation rule} \qquad y = \mathbf{w} \cdot \mathbf{x}$$

$$\text{Hebb rule} \qquad \tau \frac{\mathrm{d}w_j}{\mathrm{d}t} = y x_j$$

$$\text{equivalently, for discrete update} \qquad \Delta w_j = \epsilon y x_j$$

This rule is unstable; for positive inputs and weights we get only growth of connections (examine $\frac{\mathrm{d}|\mathbf{w}|^2}{\mathrm{d}t}$). Can introduce decay of connections by thresholds either on input or output:

$$\text{postsynaptic threshold} \qquad \tau \frac{\mathrm{d}w_j}{\mathrm{d}t} = (y - \theta_y) x_j \qquad \theta_y = \langle y \rangle$$

$$\text{presynaptic threshold} \qquad \tau \frac{\mathrm{d}w_j}{\mathrm{d}t} = y(x_j - \theta_{x_j}) \qquad \theta_{x_j} = \langle x_j \rangle$$

However, these rules are still unstable.

Normalisation

Hebbian-based learning rules alone are unstable. Need some way to keep weights within bounds and introduce **competition**. Approaches:

1. Enforce limits on individual weights, e.g. [0,1].
2. Renormalise weights periodically to **rigidly** satisfy some constraint ($\sum_j w_j = K$ or $\sum_j w_j^2 = K$).

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = y\mathbf{x} - [\frac{y(\mathbf{n}\cdot\mathbf{x})}{\mathbf{n}\cdot\mathbf{n}}]\mathbf{n} \qquad \mathbf{n} = \text{vector of 1s}$$

subtractive normalisation: sub. k off each weight. $\frac{\mathrm{d}\sum_i w_i}{\mathrm{d}t} = \frac{\mathrm{d}(\mathbf{n}\cdot\mathbf{w})}{\mathrm{d}t} = 0$.

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = y\mathbf{x} - [\frac{y(\mathbf{n}\cdot\mathbf{x})}{\mathbf{n}\cdot\mathbf{w}}]\mathbf{w}$$

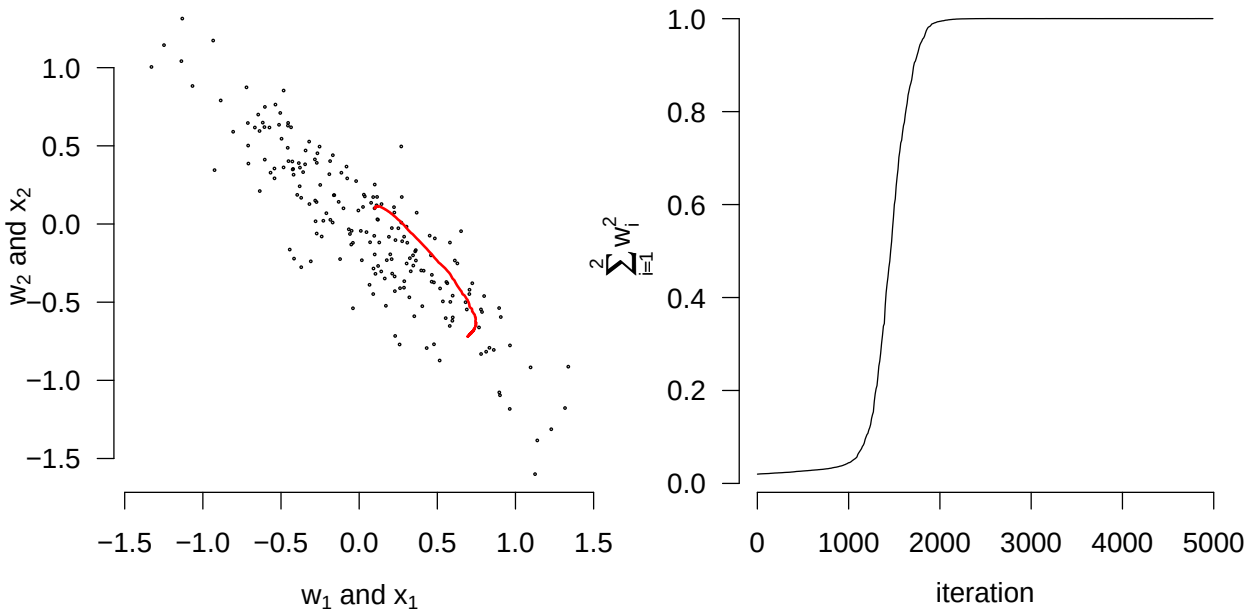
divisive normalisation: divide every weight by k.

Subtractive and divisive normalisation have different geometrical effects.

Oja rule as principal component analysis extractor

Add terms to learning rule so that constraints are **dynamically** enforced:

$$\tau \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = y\mathbf{x} - \alpha y^2 \mathbf{w}$$



At s.s. $\mathbf{w}$ is maximal eigenvector of $\mathbf{Q}$.

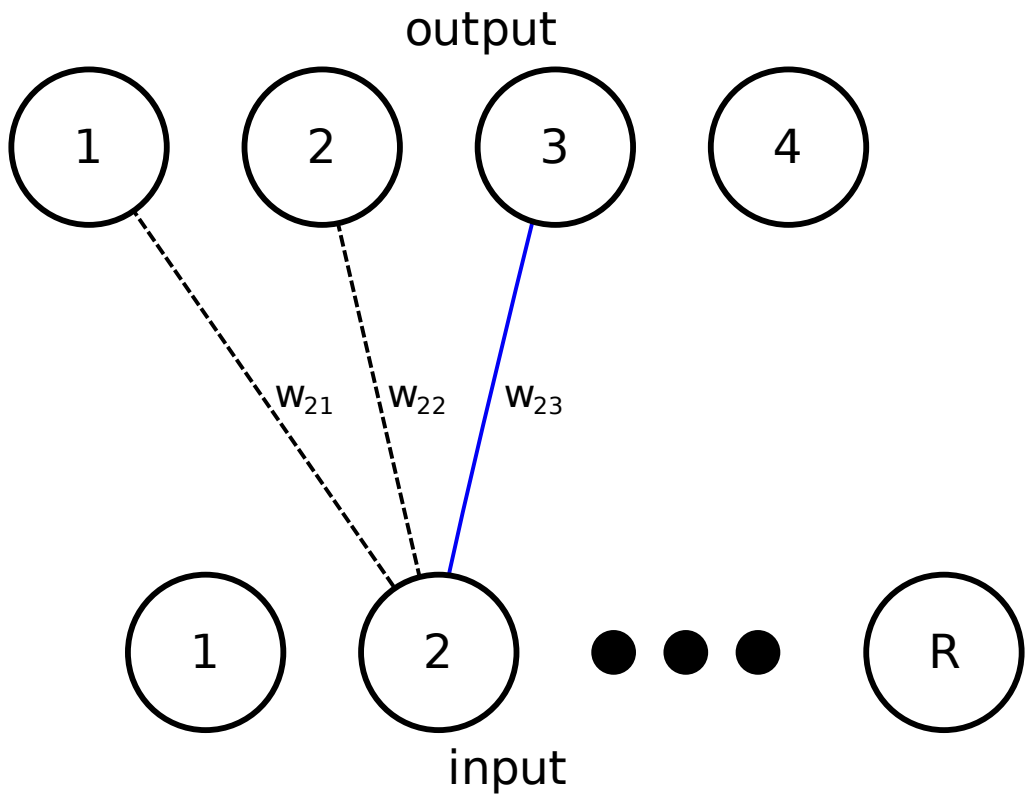
This is Principal components analysis (PCA). Rule finds vector such that projection onto that vector maximises the variance of the responses.

(nth PCA = nth largest eigenvector of correlation matrix of inputs.)

Extracting multiple principal components

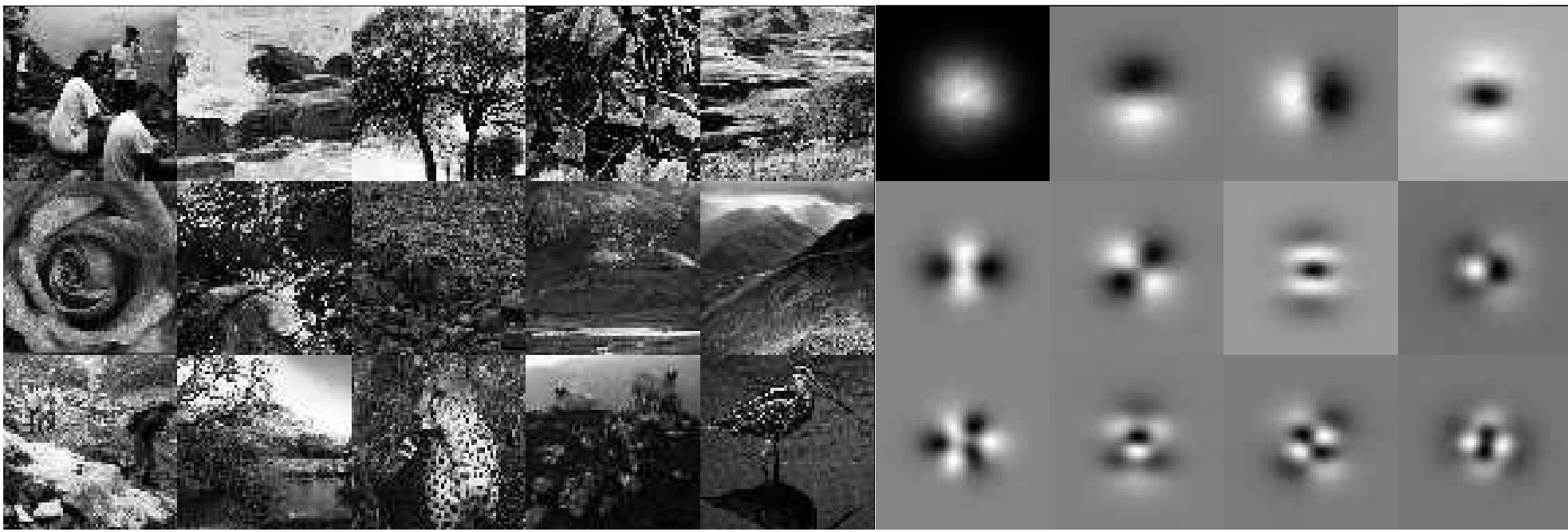
What happens when we wire up multiple output neurons and use the Oja rule? Sanger rule (1989):

$$\Delta w_{ij} = \epsilon y_j (x_i - \sum_{k=1}^{j} y_k w_{ik})$$



Non-local update rule but reliable extraction of PCs in order.

Orientation selective receptive fields (Hancock et al. 1991)



Eigenvectors of correlation matrices predict development

$$\tau_w \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = \mathbf{Q}\mathbf{w}$$

Following approach with dynamical systems, we can rewrite weight growth in terms of eigenvectors.

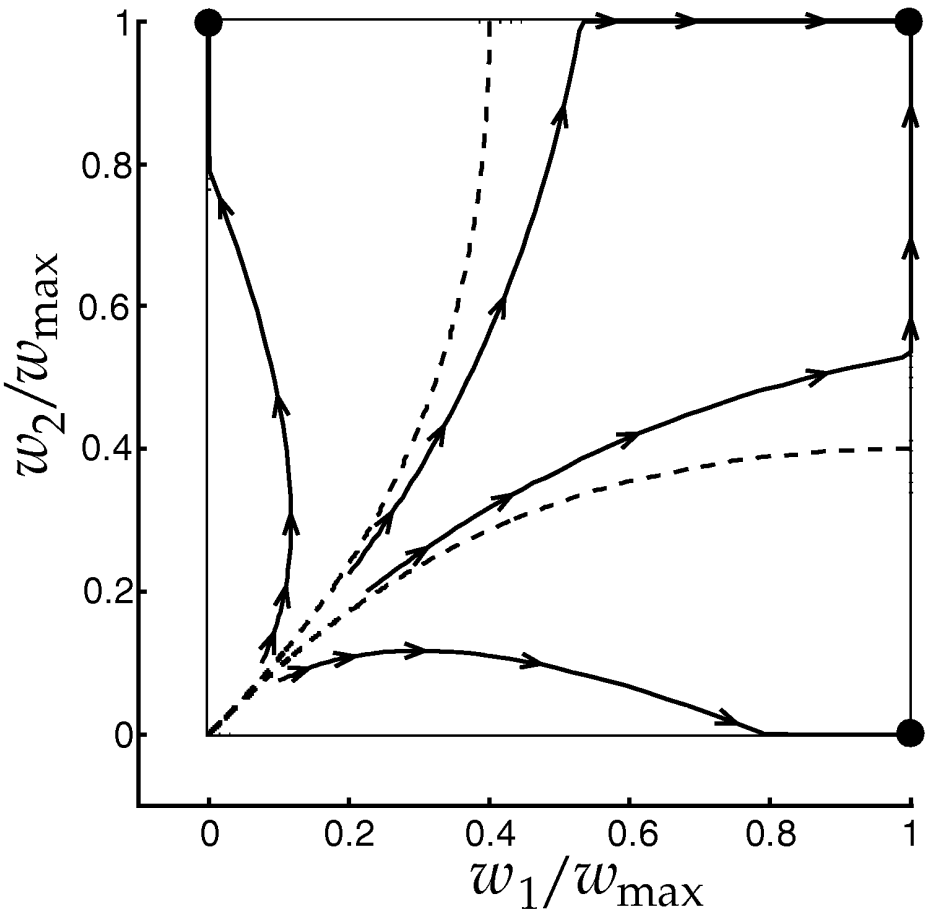
$$\mathbf{w}(t) = \sum_{i=1}^{N} \exp\left(\frac{\lambda_i t}{\tau_w}\right)(\mathbf{w}(0)\cdot\mathbf{e}_i)\mathbf{e}_i$$

Since all eigenvalues are non-negative, all grow as long as $\mathbf{e}_i \cdot \mathbf{w}(0) \neq 0$. At large times, growth dominated by $\mathbf{e}_1$ s.t. $\mathbf{w} \propto \mathbf{e}_1$.

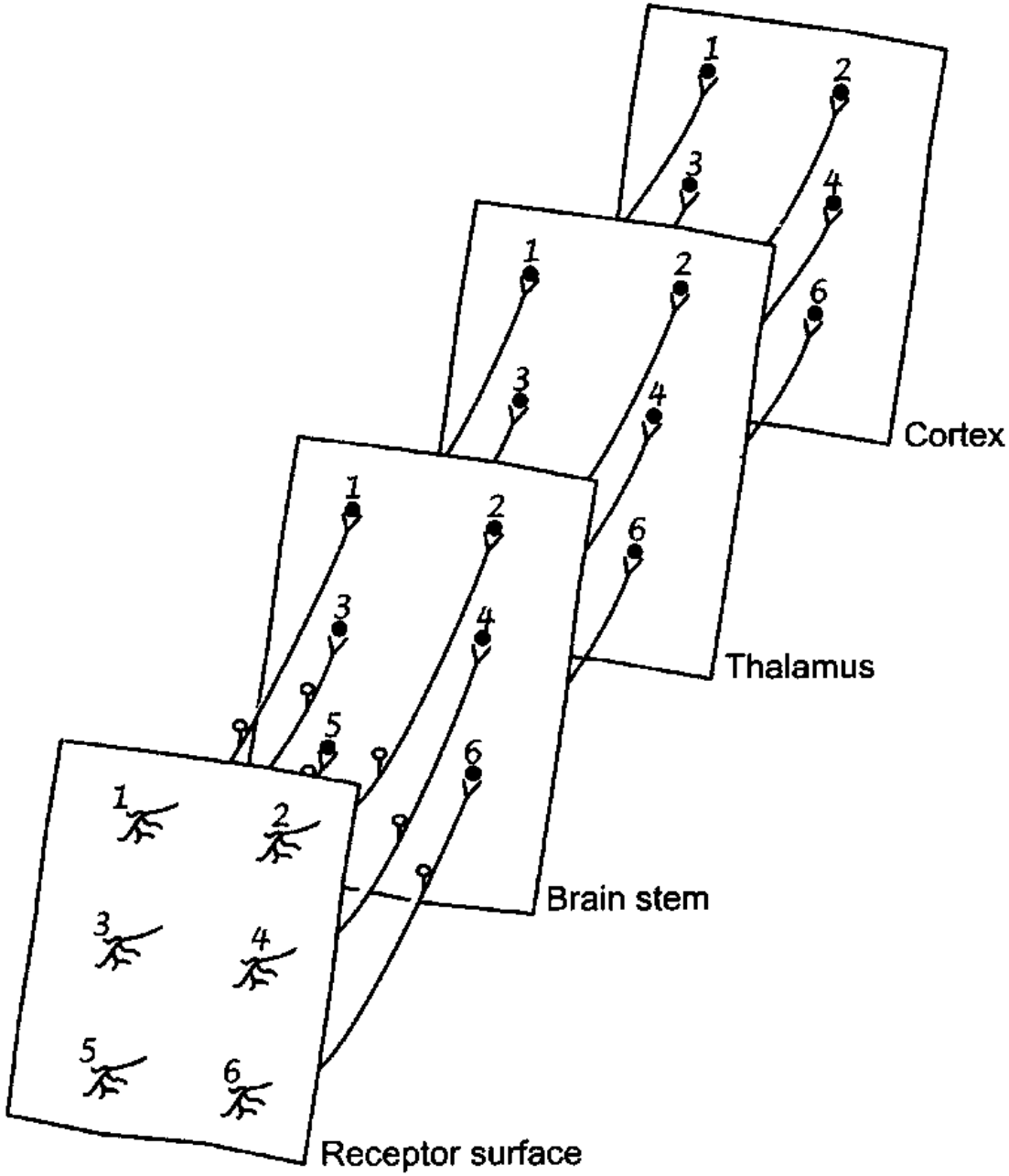
What about **saturation constraints**?

Effect of saturation limits upon development (TN Fig 8.3)

$$\mathbf{Q} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix} \qquad \mathbf{e}_1 = \qquad , \lambda_1 =$$

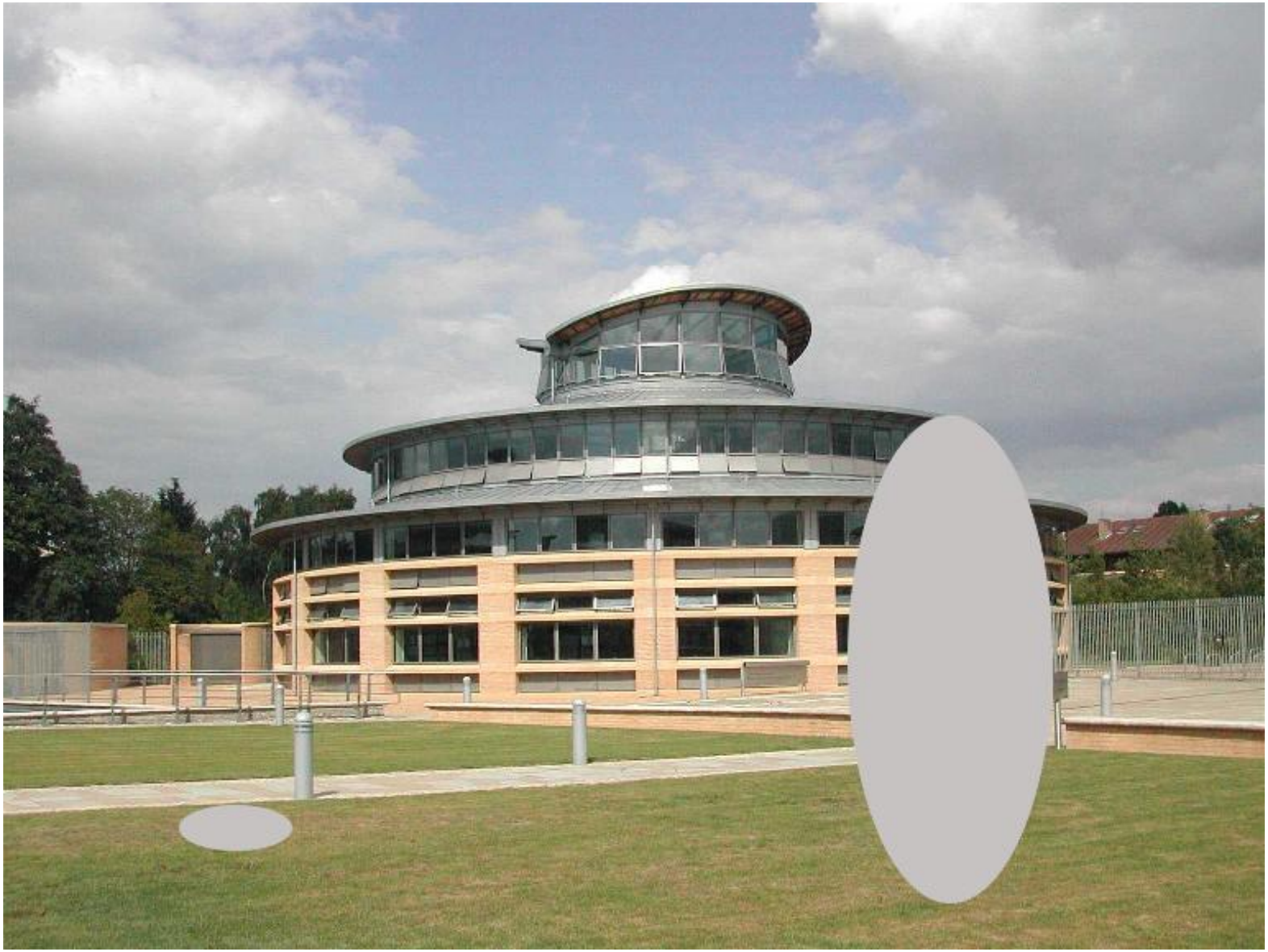


Topographic maps (Kaas & Cantina, 2002)



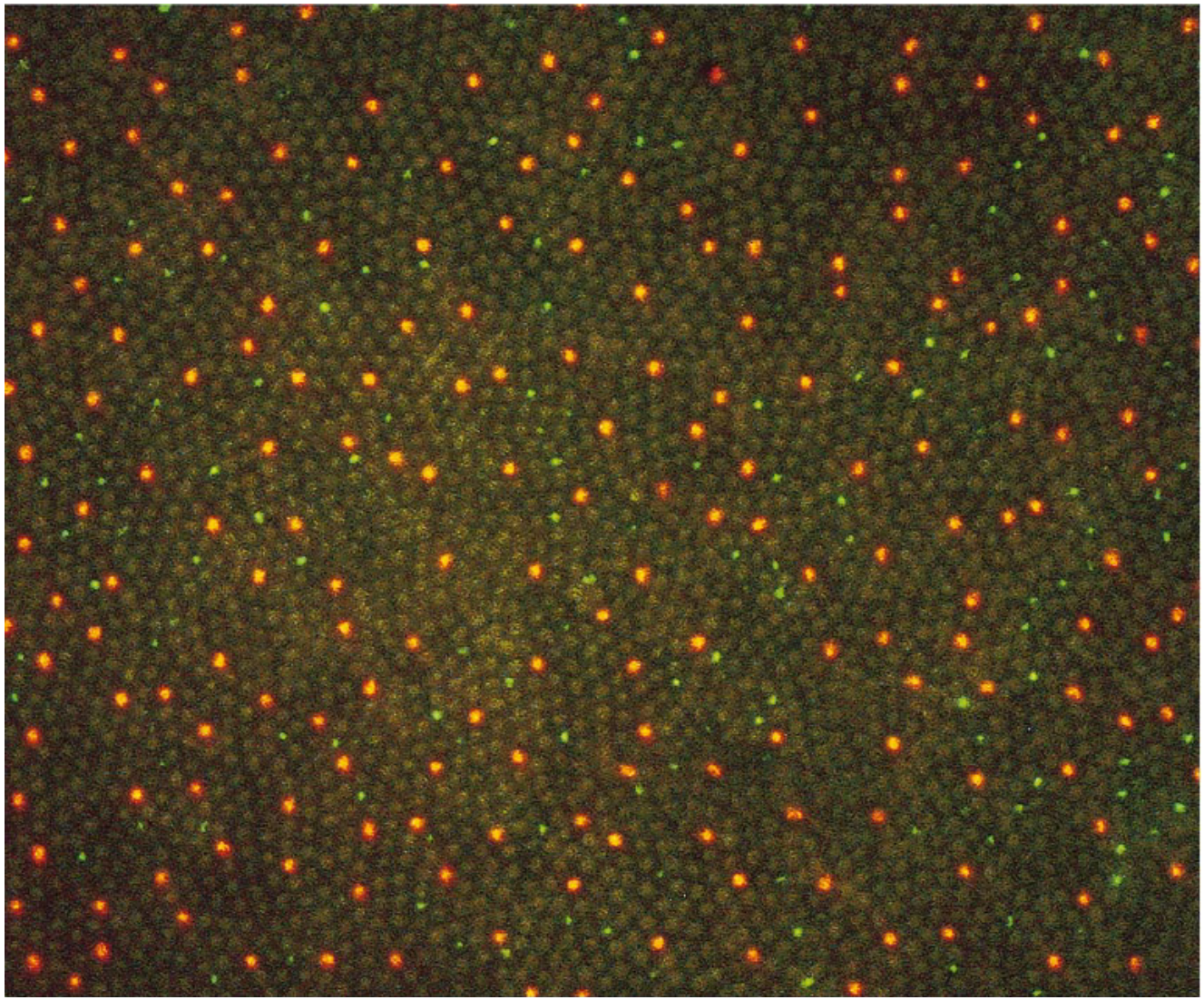
Processing the visual scene

Processing the visual scene



Processing the visual scene

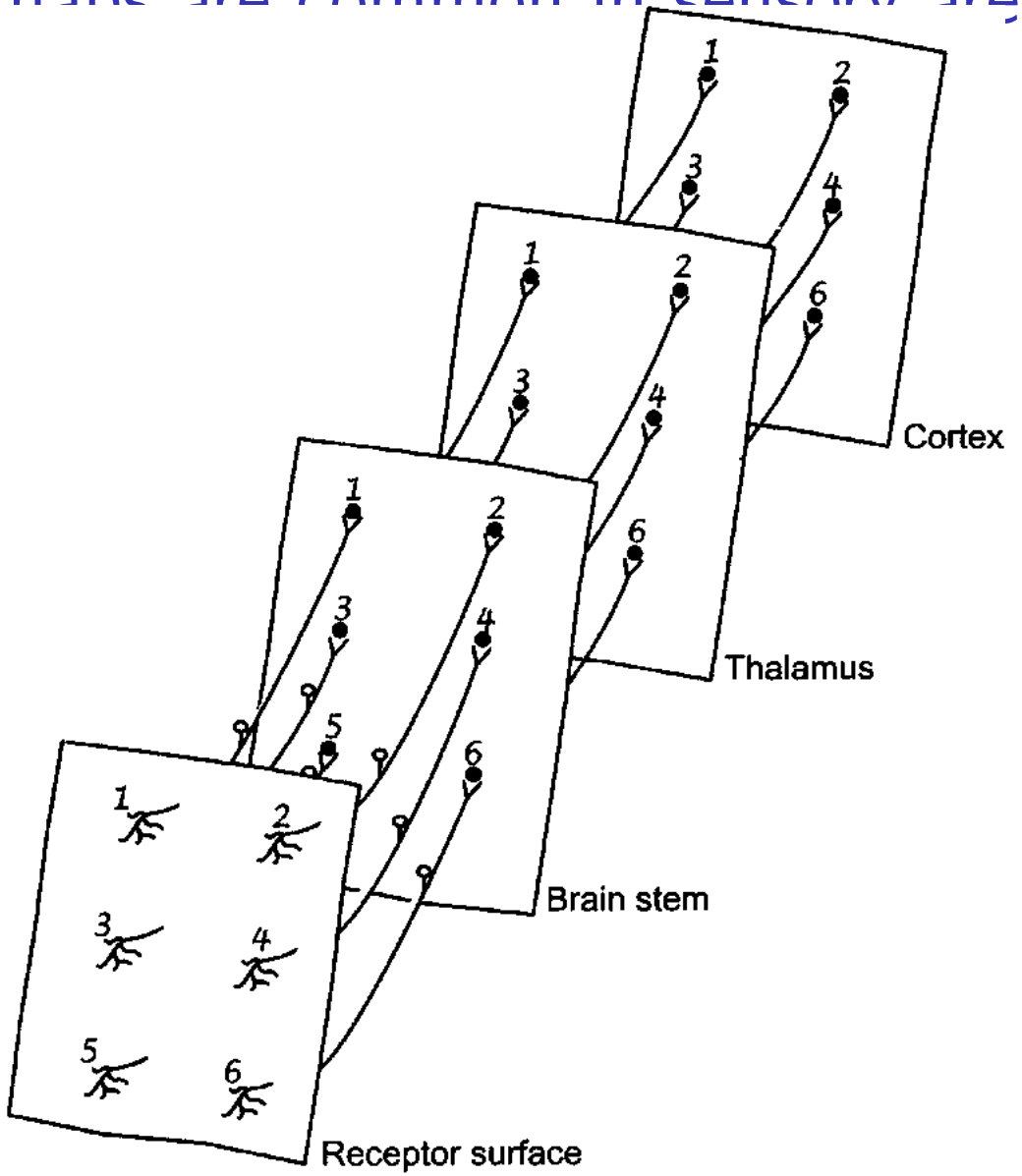
Retinal photoreceptor distribution (Ground squirrel; Galli-Resta et al., 1999)



Green = rods; orange = S cones; unlabelled = M cones.

Field of view: ≈ 400 μm wide.

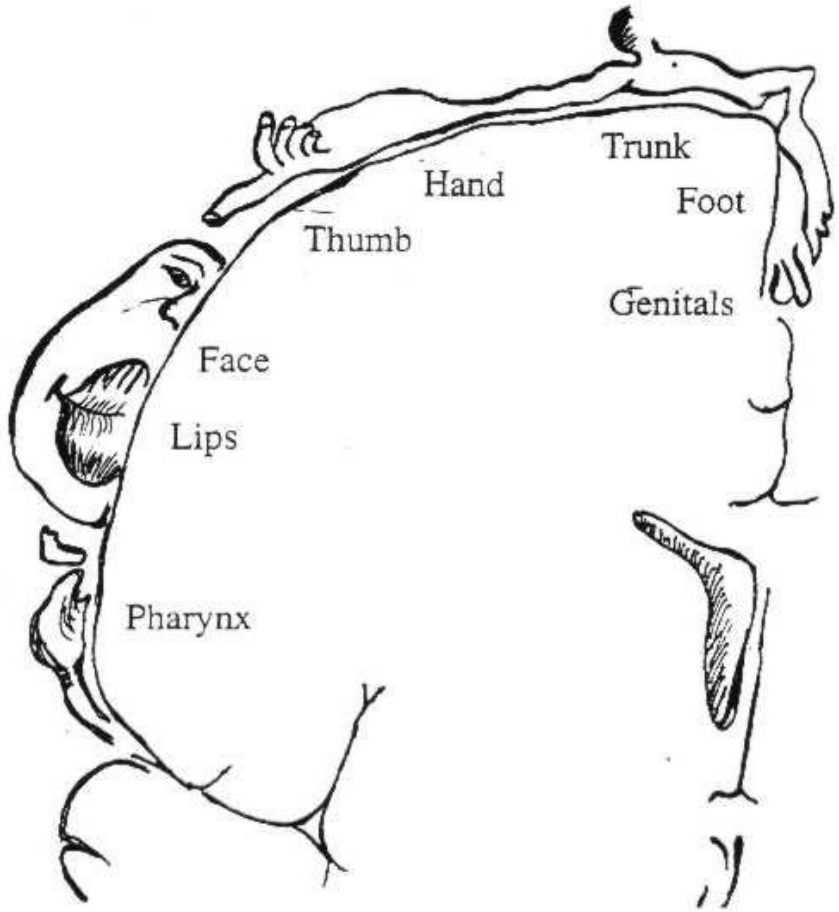
Topographic maps are common in sensory areas



Kaas & Catania, 2002

Somatosensory maps

Broadly topographic, with over-representation of hands and face.

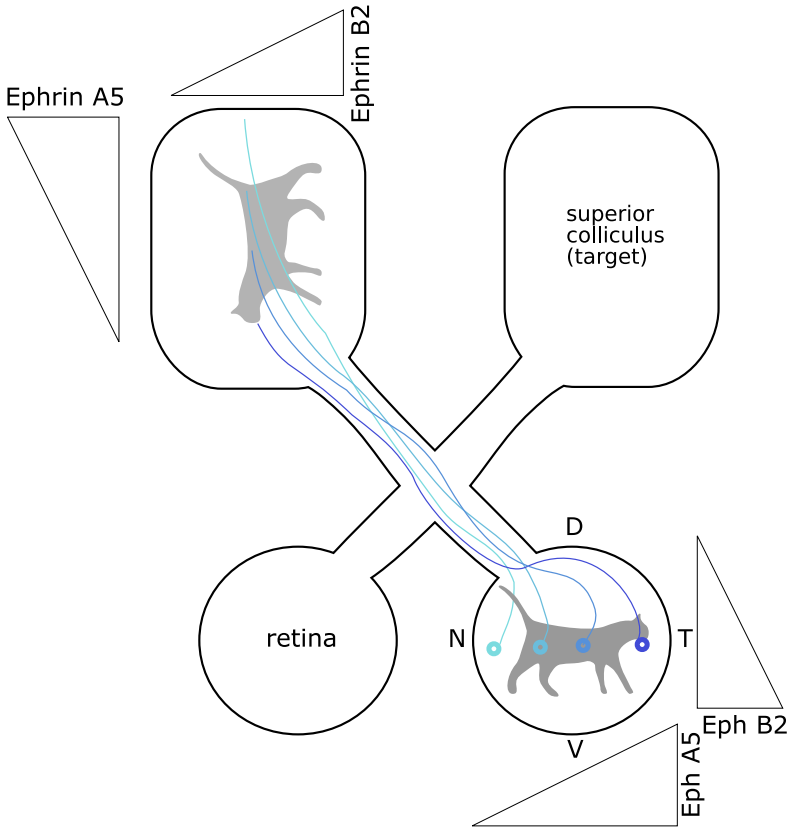


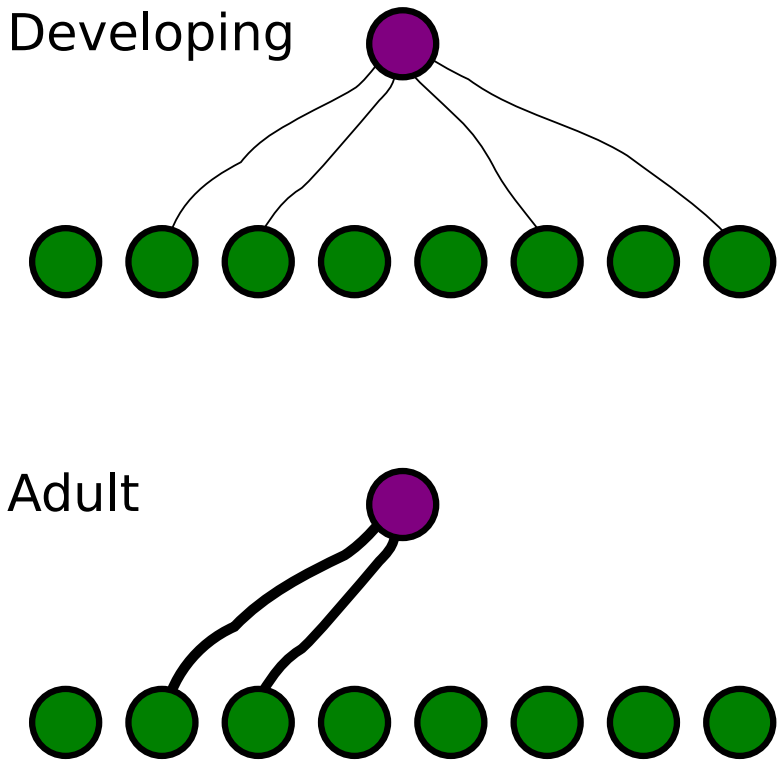
http://cogsci.bme.hu/~ikovacs/latas2005/prepI_4_2_files/fig5.jpg

http://faculty.washington.edu/chudler/flash/hom.html

Two major principles of sensory map formation

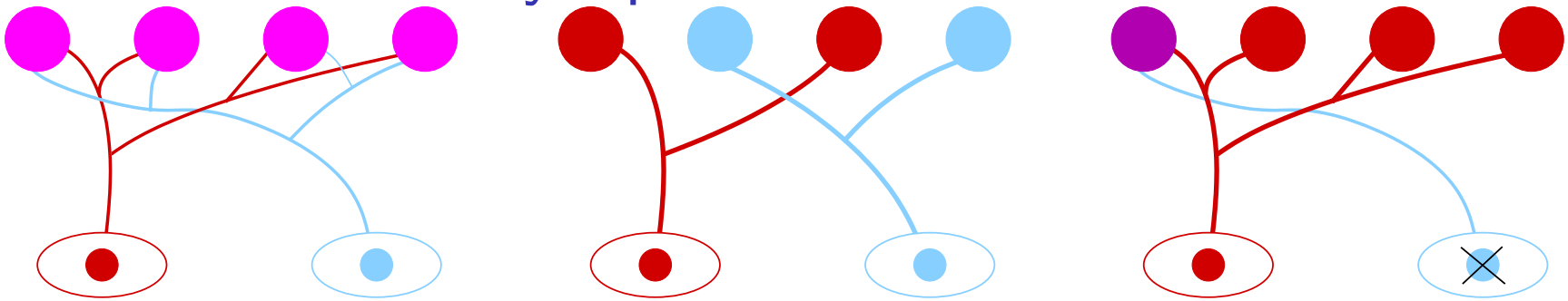
1. Activity-independent processes to establish overall connectivity. Molecular gradients.
2. Activity-dependent processes to refine local connectivity.



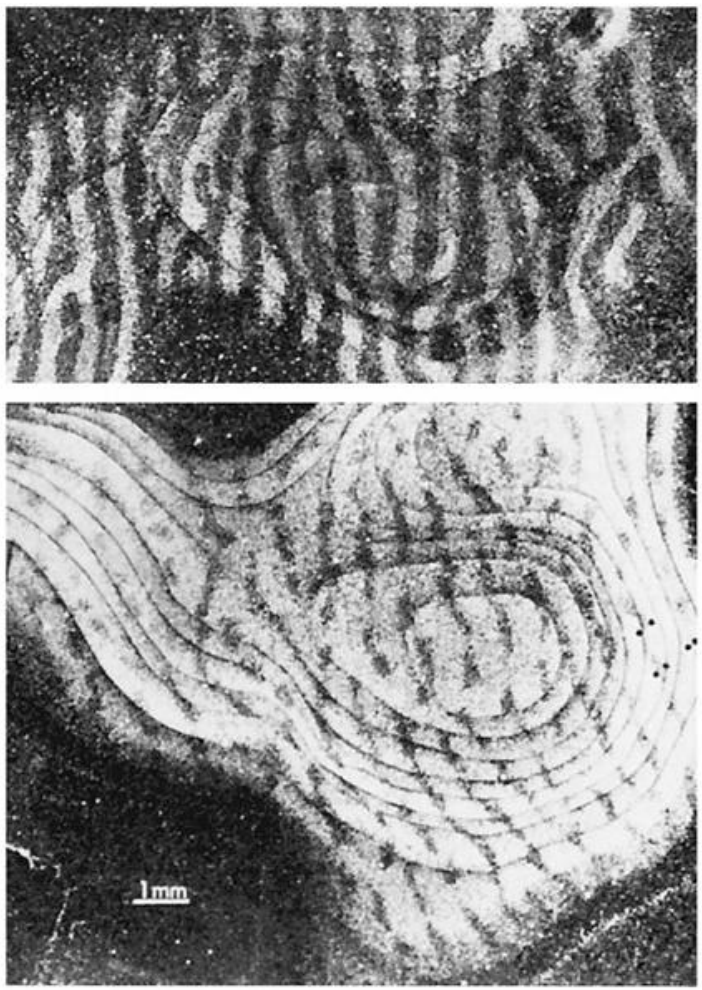


Interaction between two currently hot debated. Models may help us understand relative contributions of each mechanism.

Ocular dominance: synapse elimination and neural activity

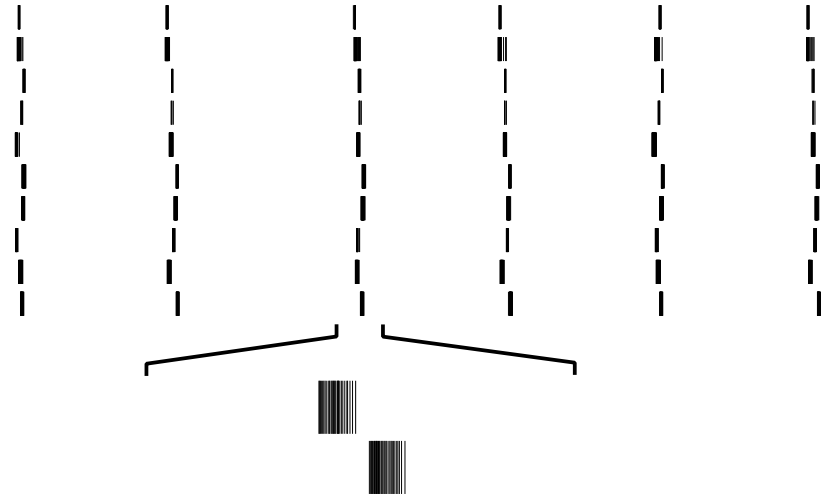


- Correlated activity between pre- and postsynaptic cells thought to drive refinement of connections (Hebb).
- Competition (e.g. for limited resources) driving development.
- Synapse elimination before eye-opening ...

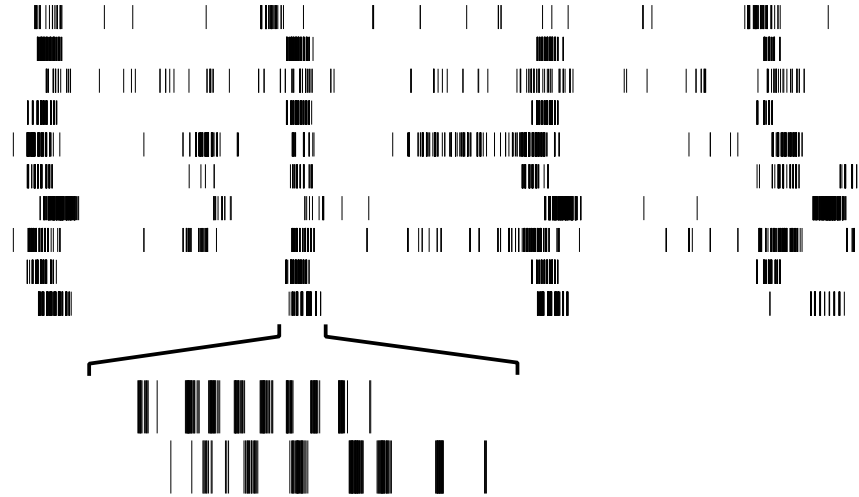


Normal development: spike trains

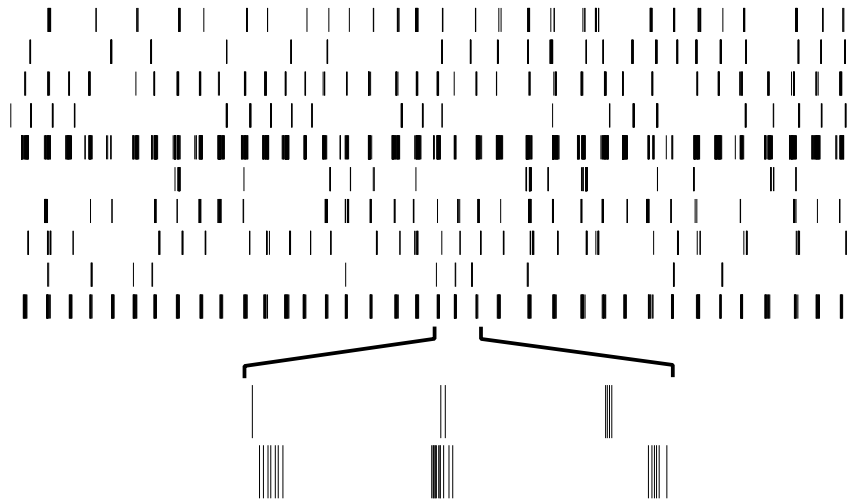
P9



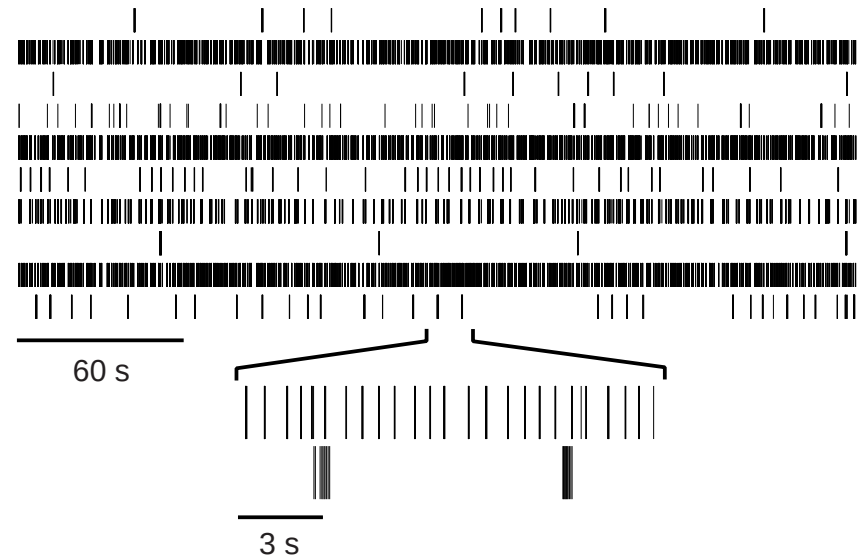
P15



P11



6 wk

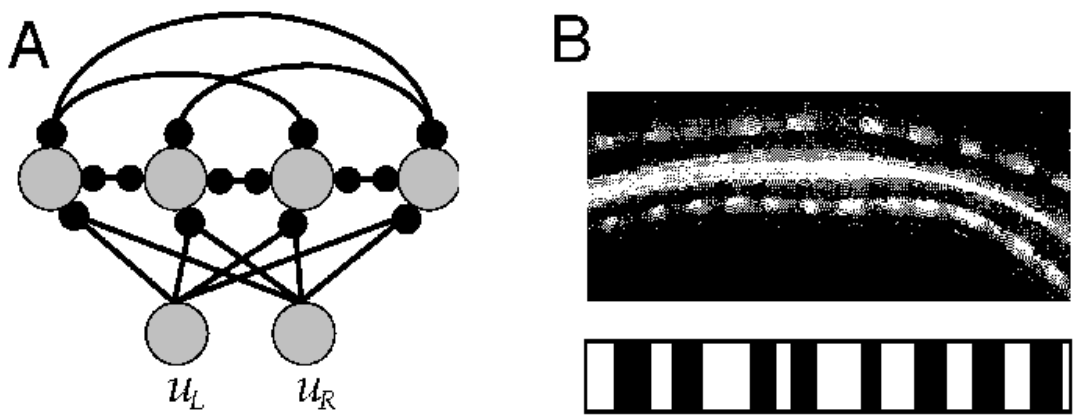


Analysis of ocular dominance (OD) formation

For one cortical output neuron, can we predict outcome?

For many cortical output neurons, can we get anything other than "salt-and-pepper" patterns? (Assume one RGC per eye.)
Follow analysis as with single cell (Assignment 2).

$$\tau_w \frac{\mathrm{d}\mathbf{w}_-}{\mathrm{d}t} = (q_s - q_d)\mathbf{K}\mathbf{w}_- \qquad \mathbf{K} = (\mathbf{I} - \mathbf{M})^{-1}$$



Goodhill (1993) model of OD and topographic map development

Key model combining the development of ocular dominance and topography.

Architecture: two eyes (each 16x16 pixels)and one cortical sheet (32x32 neurons

Initial connections random, but include topographic bias to ensure overall map layout is consistent.

Modelling visual inputs

To make one pair of left/right inputs.

- For each eye, generate 0/1 randomly for each pixel. Then, blur (convolve) activity independently in each eye.
- To introduce correlations between the eyes, find the pixel at the same position in each eye and combine them to a degree h:

$$x_i^{\text{left}} \leftarrow (1-h)x_i^{\text{left}} + hx_i^{\text{right}}$$

$h = 0.0 \Rightarrow$ activity remains independent between eyes.
$h = 0.5 \Rightarrow$ activity same in each eye.

Many pairs of inputs created.

Output unit activation and weight update

- Activation:

$$y_j = \sum_i w_{ij}^{\text{left}} x_i^{\text{left}} + \sum_i w_{ij}^{\text{right}} x_i^{\text{right}}$$

- "Winner take all": find output unit g with highest firing rate; update weights of output units physically close to g:

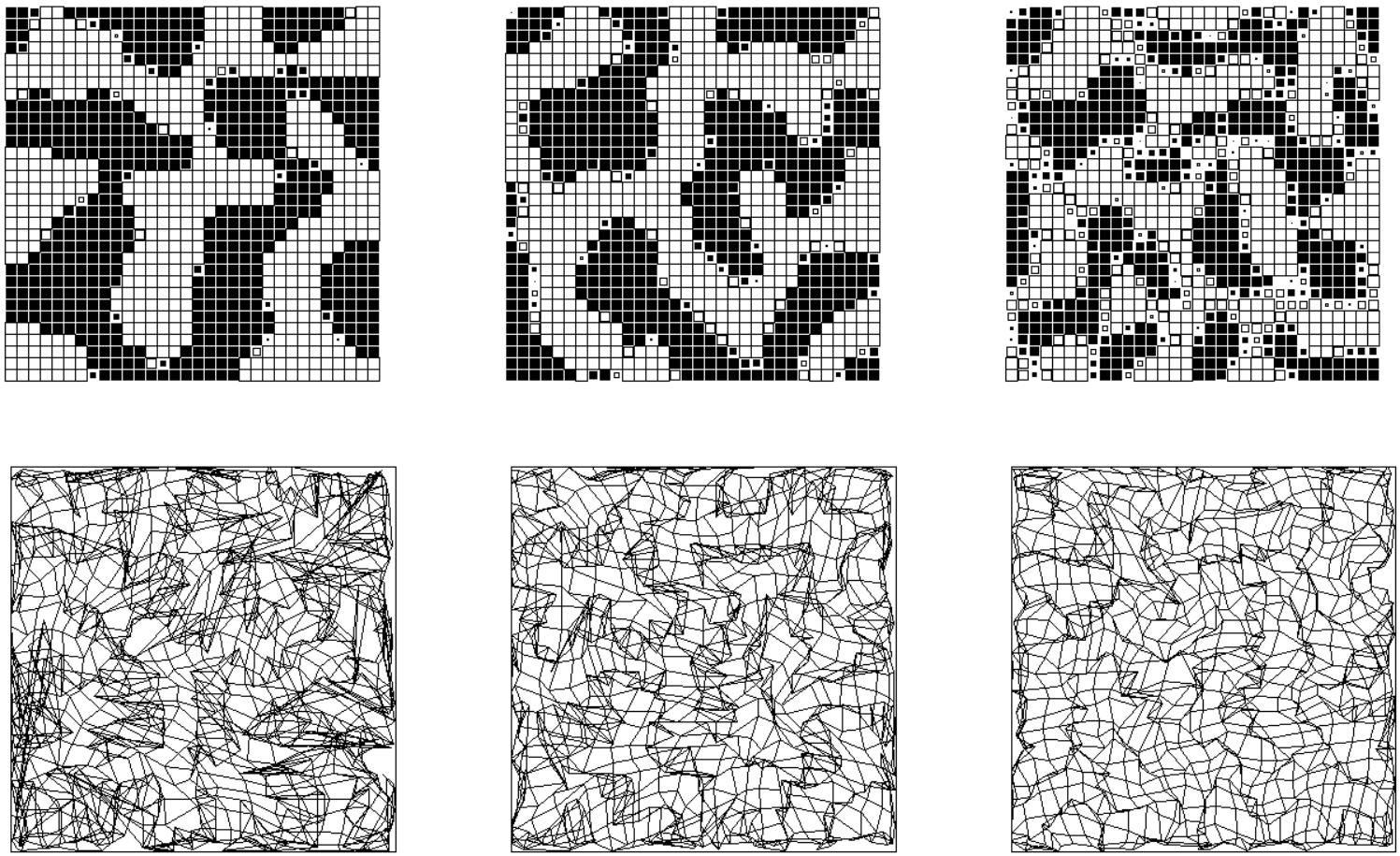
$$g =_j y_j$$
$$\Delta w_{ij}^{\text{L/R}} = \alpha x_i^{\text{L/R}} s(j, g)$$

  $s(j, g)$ Gaussian function of distance between unit j and winner g.

- Postsynaptic subtractive normalisation used.

Ocular dominance and topography

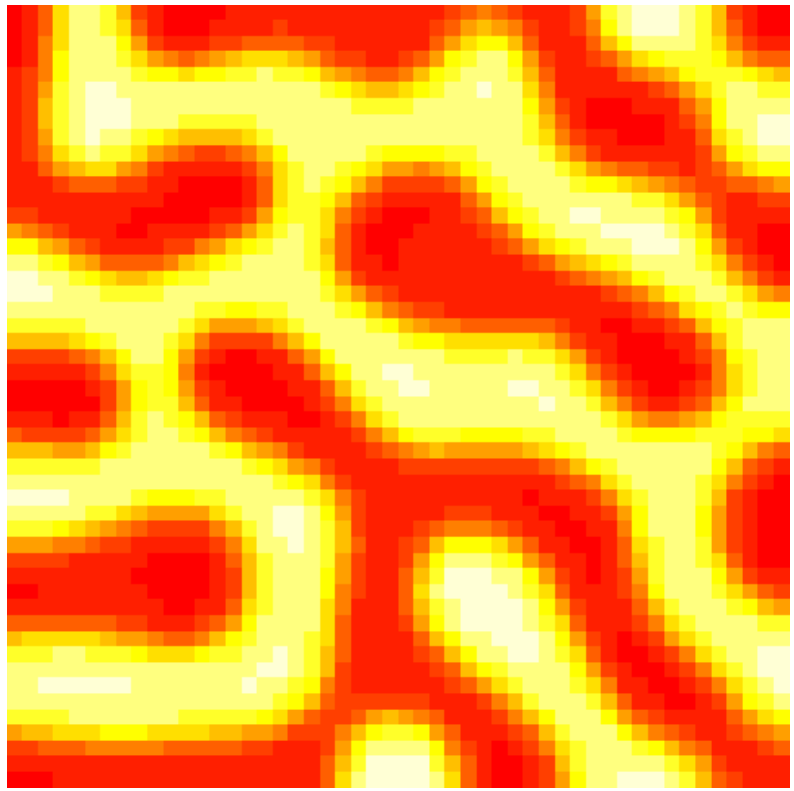
$h = 0.0$ $h = 0.1$ $h = 0.2$



Key prediction: stronger between-eye correlations $\Rightarrow$ narrower stripes.

Reaction-diffusion systems can also make stripes

$$\frac{\partial u}{\partial t} = f(u, v) + d_u \frac{\partial^2 u}{\partial x^2}$$
$$\frac{\partial v}{\partial t} = g(u, v) + d_v \frac{\partial^2 v}{\partial x^2}$$



Same outcomes does not imply same mechanism!

Feature-based models

Each element of an input vector codes for a feature of a stimulus, rather than e.g. a 2-d vector of pixel intensities.

$$\mathbf{u} = (x, y, o, a\cos\theta, a\sin\theta)^T$$

where (x, y) is the centre of mass of retinal activity, o is the ocularity [-1,1] and (a, θ) indicate orientation selectivity.
This allows two $N \times N$ images to be concisely represented by a low-dimensional vector.

Weight modification in feature-based models

No need for normalisation, since weights become similar to input vectors.
e.g. elastic net methodology:
"softmax" output of cortical unit a:

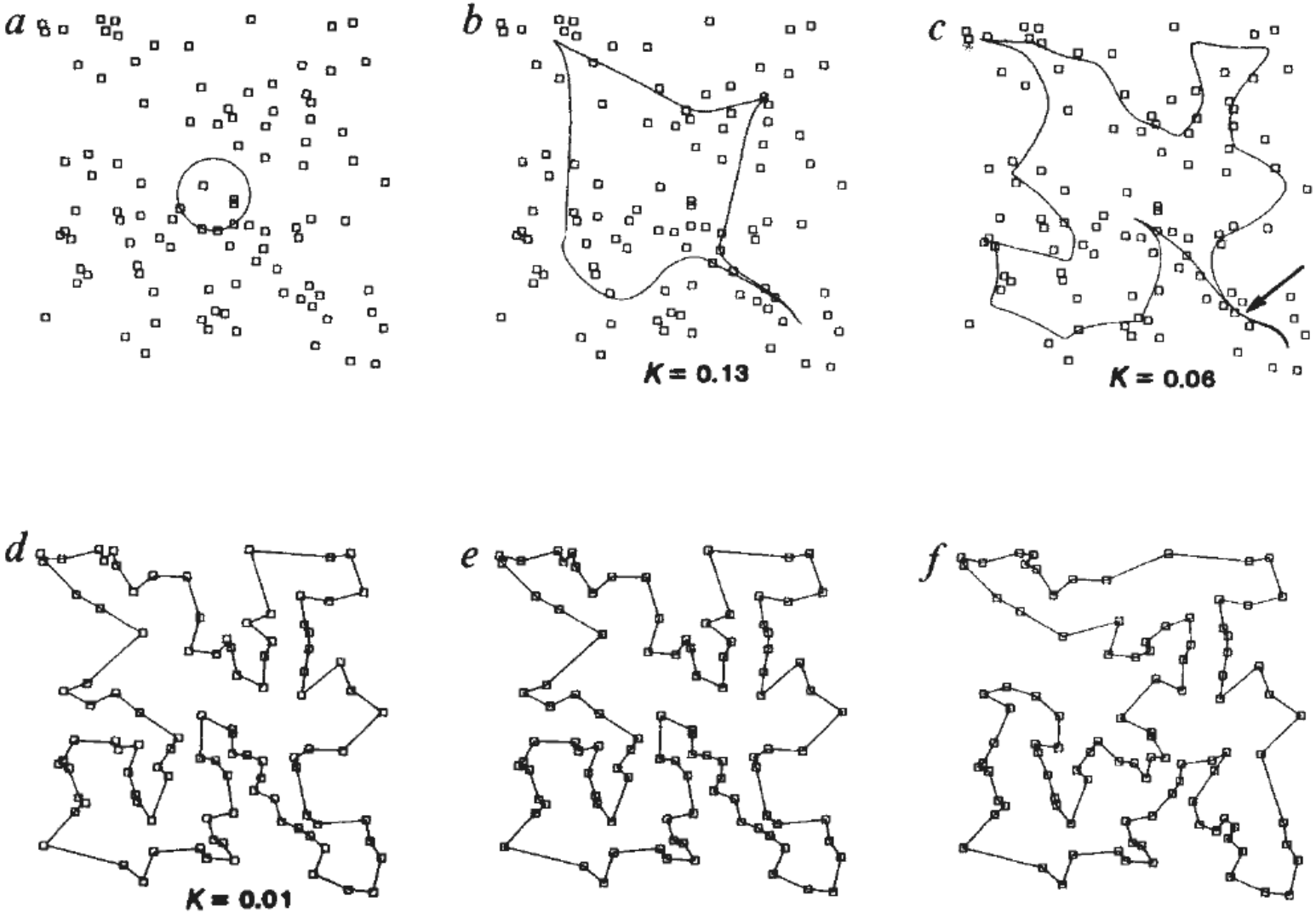
$$x_a = \exp\left(-\sum_b (u_b - W_{ab})^2/(2\sigma_b^2)\right)$$

$$v_a = \frac{x_a}{\sum_{a'} x_{a'}}$$

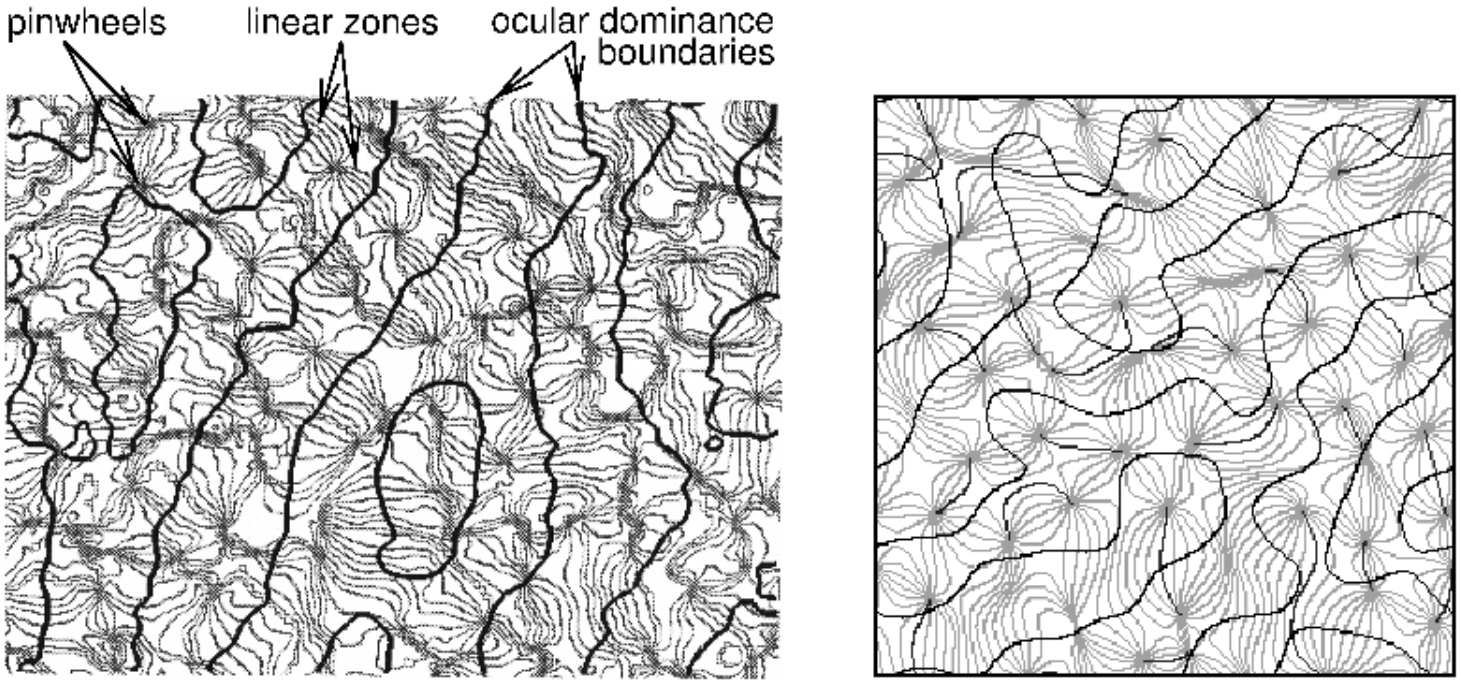
$$\tau_w \frac{\mathrm{d}W_{ab}}{\mathrm{d}t} = \langle v_a(u_b - W_{ab})\rangle + \beta \sum_{a' \in \text{neighs}(a)} (W_{a'b} - W_{ab})$$

Self organising map (SOM) works in similar way, but restricts weight updates to a "winner" and nearby neighbours.

Elastic net: TSP results for 100 cities (Durbin & Willshaw, 1987)



Joint orientation domain and ocular dominance column development



Left: macaque iso-orientation lines in grey; ocular dominance borders in black. Pinwheels often occur near OD centres, and linear zones often perpendicular to OD. Right: feature-map model output.

Summary

- Hebbian learning must be augmented by some form of competition.
- Normalisation enforces constraints.
- Subtractive normalisation often required for OD formation.
- Map formation.
- Reading: TN Chapter 8; Goodhill (1993).