# The cerebral cortex model that self-organizes conditional probability tables and executes belief propagation

IJCNN 2007 #1065

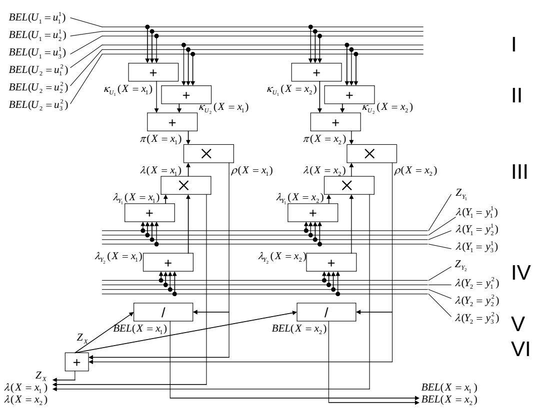
National Institute
of Advanced Industrial Science
and Technology(AIST), Japan

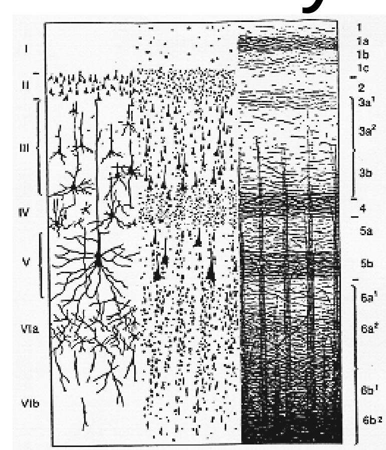
Yuuji Ichisugi

2007-08-09

# Outline

- A model of cerebral cortex, BESOM model, is explained.
- BESOM unifies self-organizing map and Bayesian network.
- This model matches main anatomical structure of cerebral cortex very well.

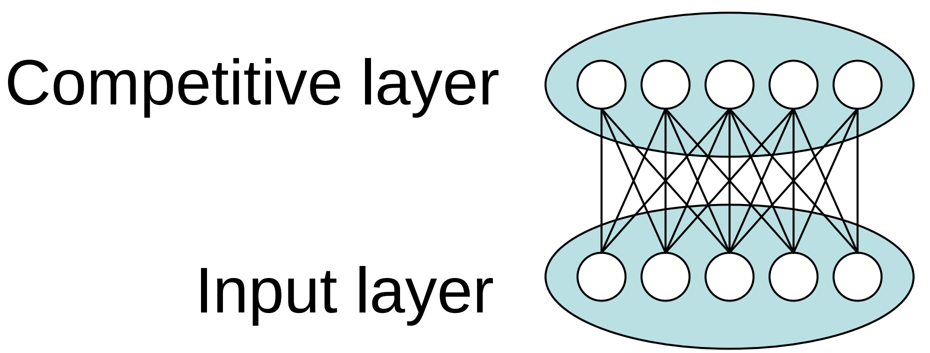


# BESOM architecture

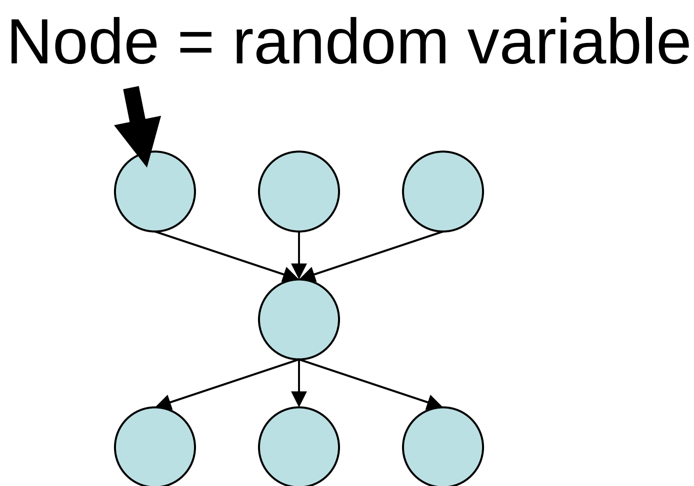
# Self-organizing map (SOM)

- Unsupervised learning algorithm.
- Competitive learning and neighborhood learning.
- Clustering high-dimensional input.



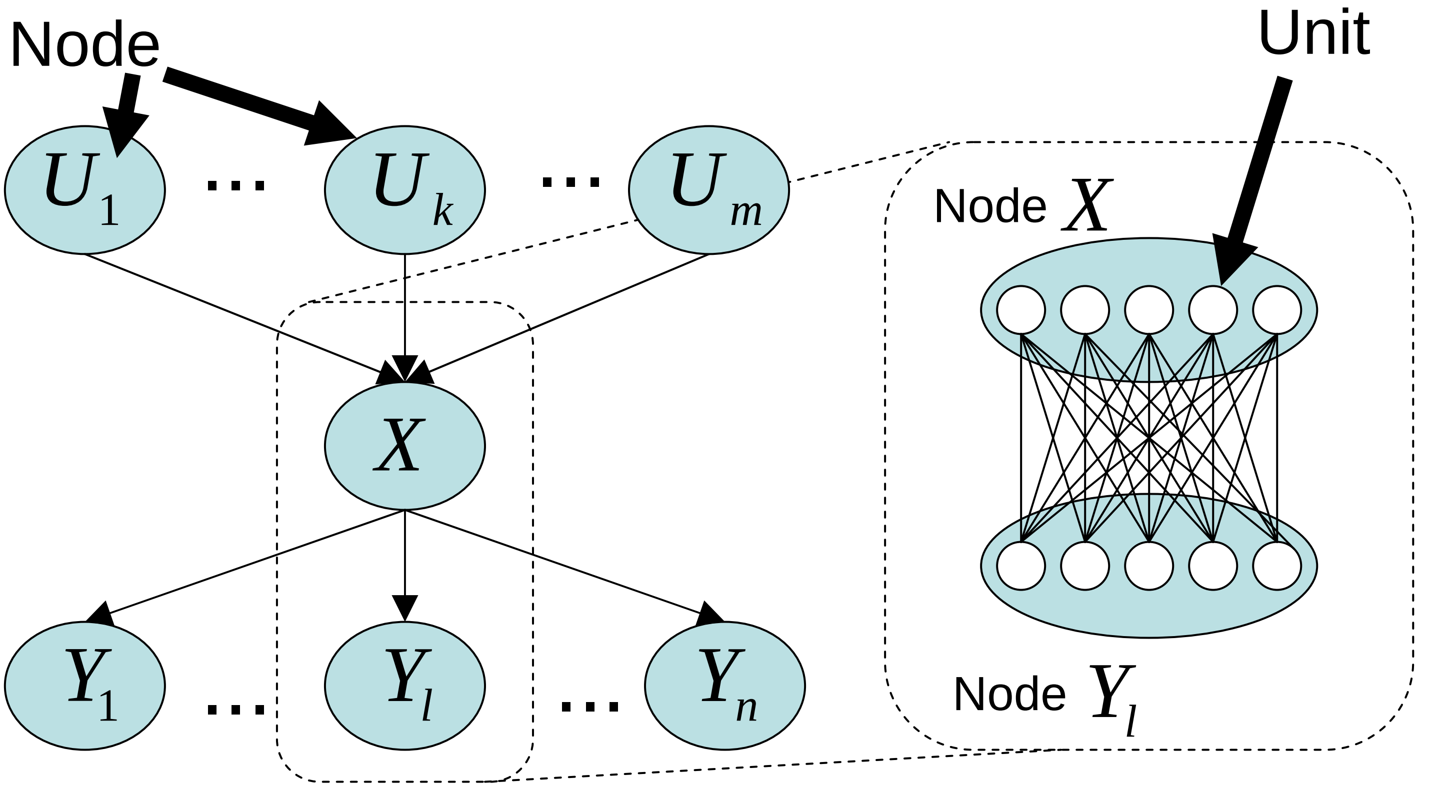
# Bayesian network

- A graphical model that represents a directed acyclic graph of causal relations between random variables.
- Belief propagation algorithm to estimate each value of node.



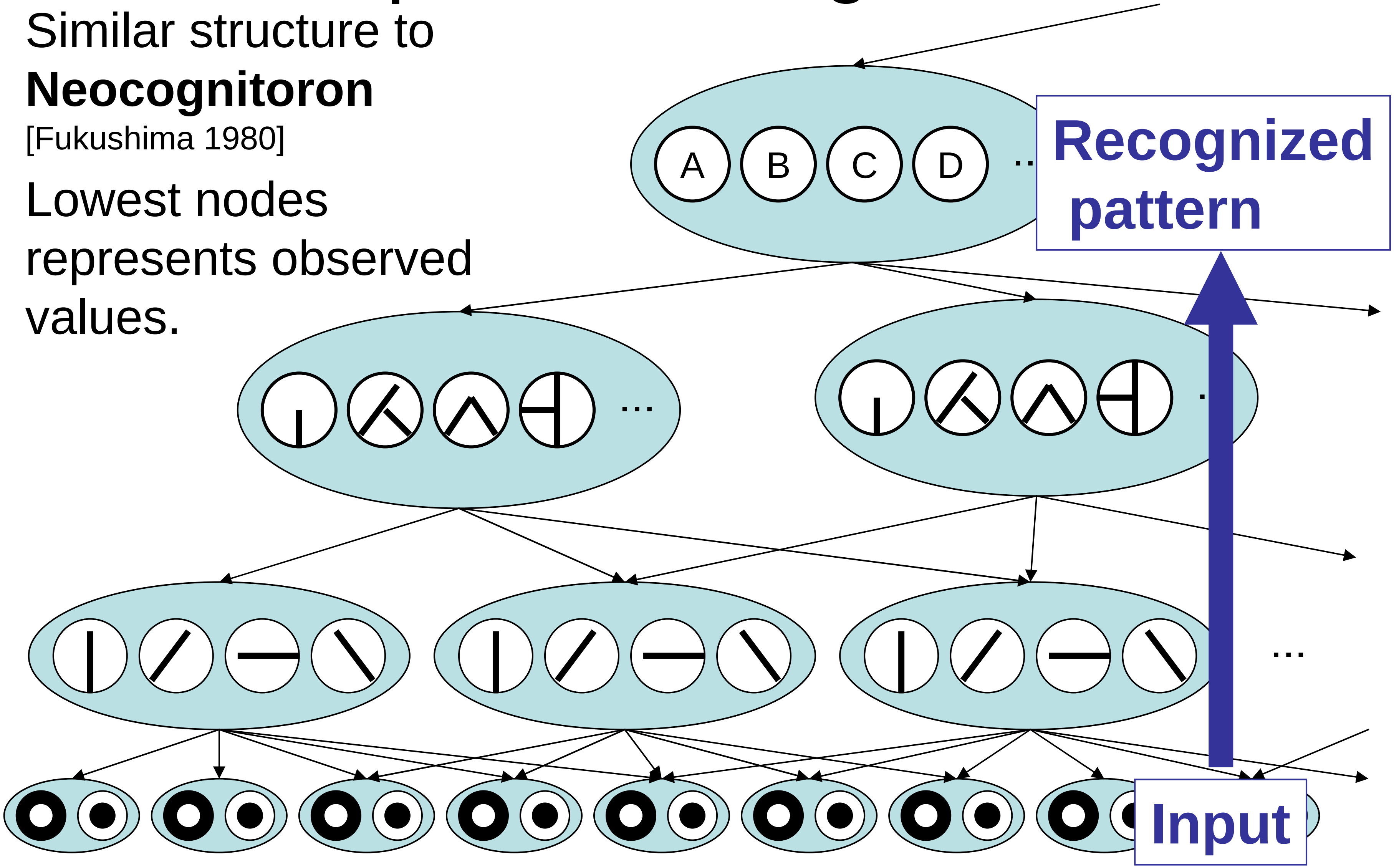
# BESOM (BidirEctional SOM)

- Each node is a competitive layer of a SOM.
- Each unit represents a value of the random variable.



# BESOM may be used for pattern recognition

- Similar structure to **Neocognitoron** [Fukushima 1980]
- Lowest nodes represents observed values.



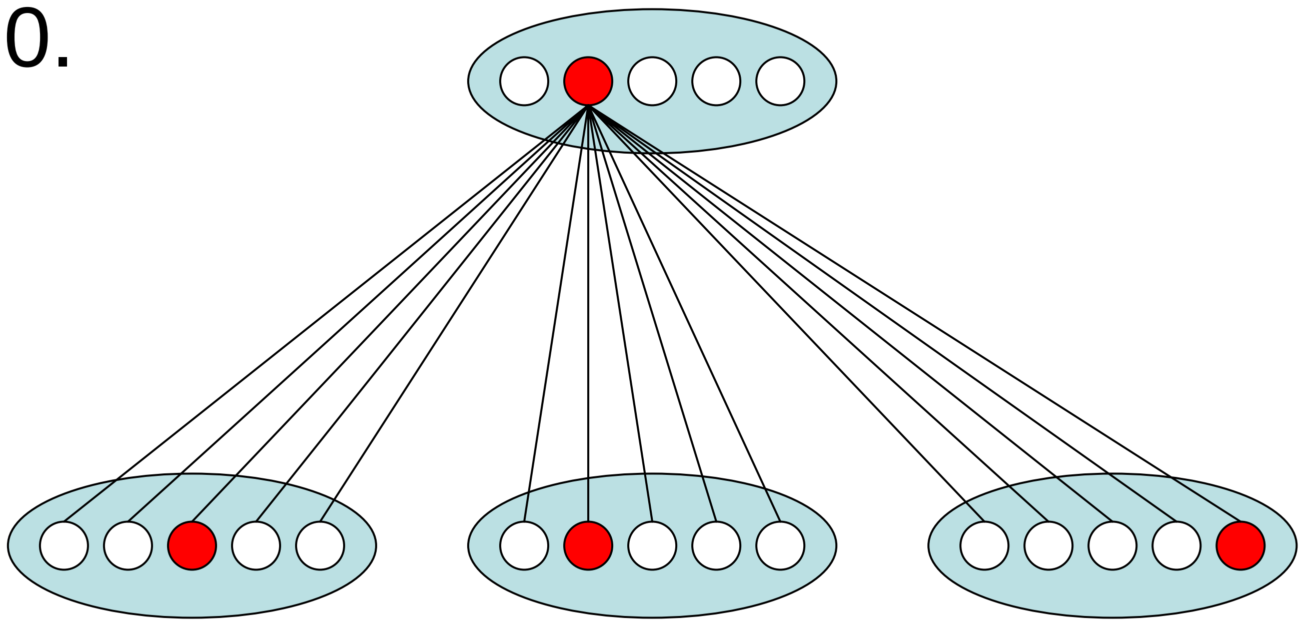
# Learning step and Recognition step

- BESOM repeats two steps alternately.
  - Learning step
    - Each node works as a SOM's competitive layer.
    - Clustering input from its child nodes.
    - Learning results are conditional probabilities, used at the next recognition step.
  - Recognition step
    - Network of nodes works as a Bayesian network.
    - Values of nodes are estimated by approx. loopy BP.
    - Results of estimation are used as input to SOMs at the next learning step.

# Learning step

# Example of input vector to a SOM

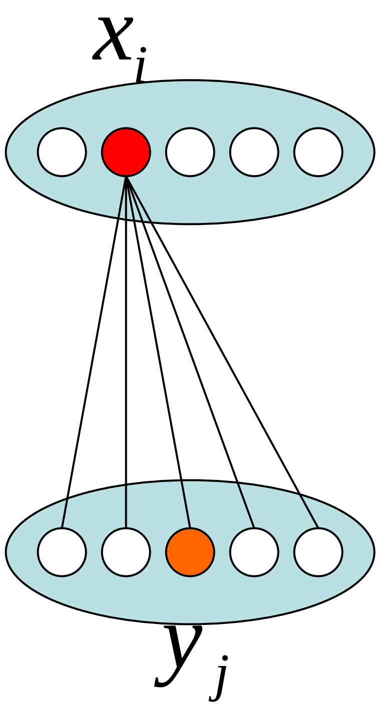
- Units corresponding to the maximum a posteriori (MAP) estimates send 1, others send 0.



$v = (0, 0, 1, 0, 0, \quad 0, 1, 0, 0, 0, \quad 0, 0, 0, 0, 1)^T$

Each red unit is a MAP estimate.

# Learning rule




- Make winner's reference vector $w_{ij}$ close to input vector.
  - with appropriate neighborhood learning.

$$w_{ij} \leftarrow (1-\alpha) w_{ij} + \alpha v_j$$

v = (0, 0, 1, 0, 0)$^T$

- Learned weight $w_{ij}$ means a conditional probability:

$$w_{ij} = P(Y = y_j \mid X = x_i)$$

# Recognition step

# Belief propagation algorithm [Pearl 1988]

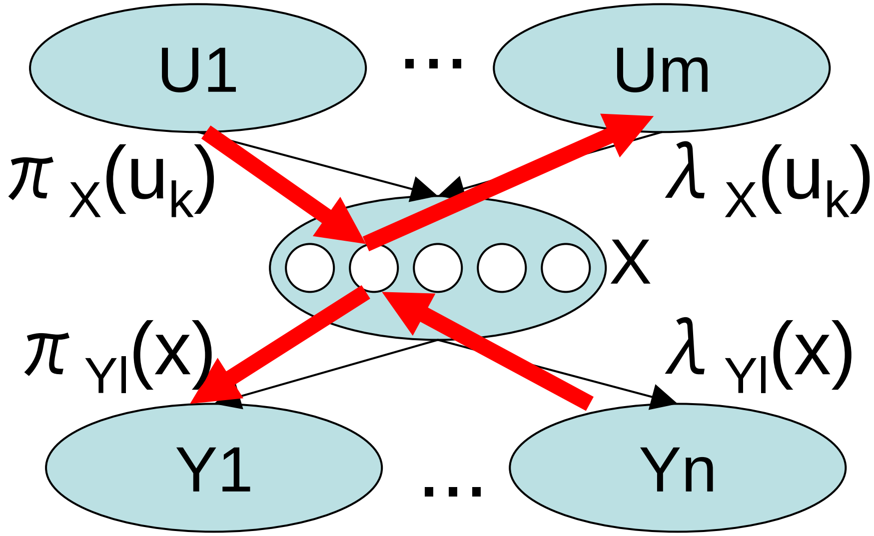
$$BEL(x) = \alpha \lambda(x) \pi(x)$$

$$\pi(x) = \sum_{u_1,\cdots,u_m} P(x|u_1,\cdots,u_m) \prod_k \pi_X(u_k)$$

$$\lambda(x) = \prod_l \lambda_{Y_l}(x)$$

$$\pi_{Y_l}(x) = \pi(x) \prod_{j \neq l} \lambda_{Y_j}(x)$$

$$\lambda_X(u_k) = \sum_x \lambda(x) \sum_{u_1,\cdots,u_m / u_k} P(x|u_1,\cdots,u_m) \prod_{i \neq k} \pi_X(u_i)$$




**It's hard to be implemented by neurons.**

# In order to approximate

- Assumption 1: CPTs (conditional probability tables) can be approximated as follows:

$$P(X \mid U_1, \cdots, U_m) \approx \sum_{i=1}^{m} P(X \mid U_i)$$

  - Reason: Smoothing, Sparseness, Infomax, ... Accurate analysis is future work.
- Assumption 2: Nodes have many parent and child nodes.

# Approximate belief propagation

$$\lambda_{Y_l}^{t+1}(x) = Z_{Y_l}^{t} + \sum_{y_l} \lambda^t(y_l) P(y_l \mid x)$$

$$\lambda^{t+1}(x) = \prod_{l=1}^{n} \lambda_{Y_l}^{t}(x)$$

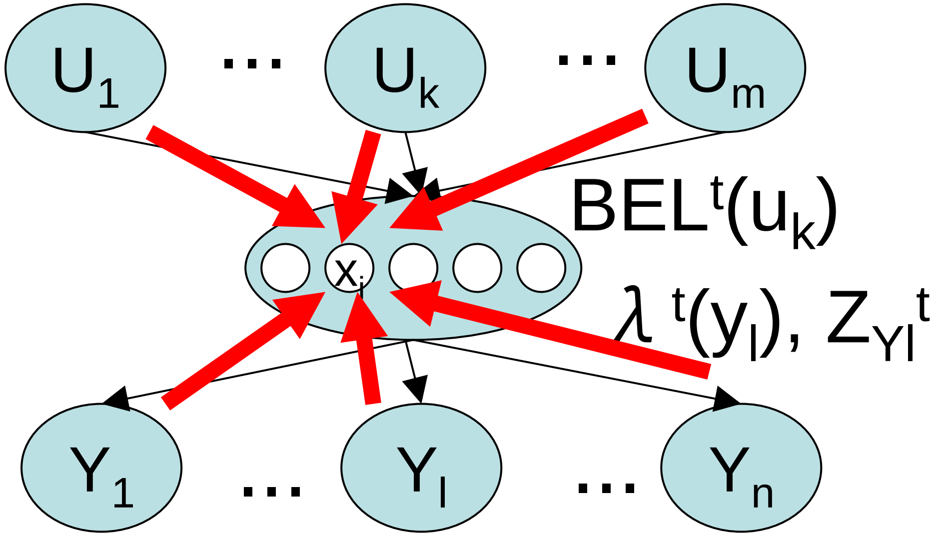
$$\kappa_{U_k}^{t+1}(x) = \sum_{u_k} P(x \mid u_k) BEL^t(u_k)$$

$$\pi^{t+1}(x) = \sum_{k=1}^{m} \kappa_{U_k}^{t+1}(x)$$

$$\rho^{t+1}(x) = \lambda^{t+1}(x) \pi^{t+1}(x)$$

$$Z_X^{t+1} = \sum_{x} \rho^{t+1}(x)$$

$$BEL^{t+1}(x) = \rho^{t+1}(x) / Z_X^{t+1}$$



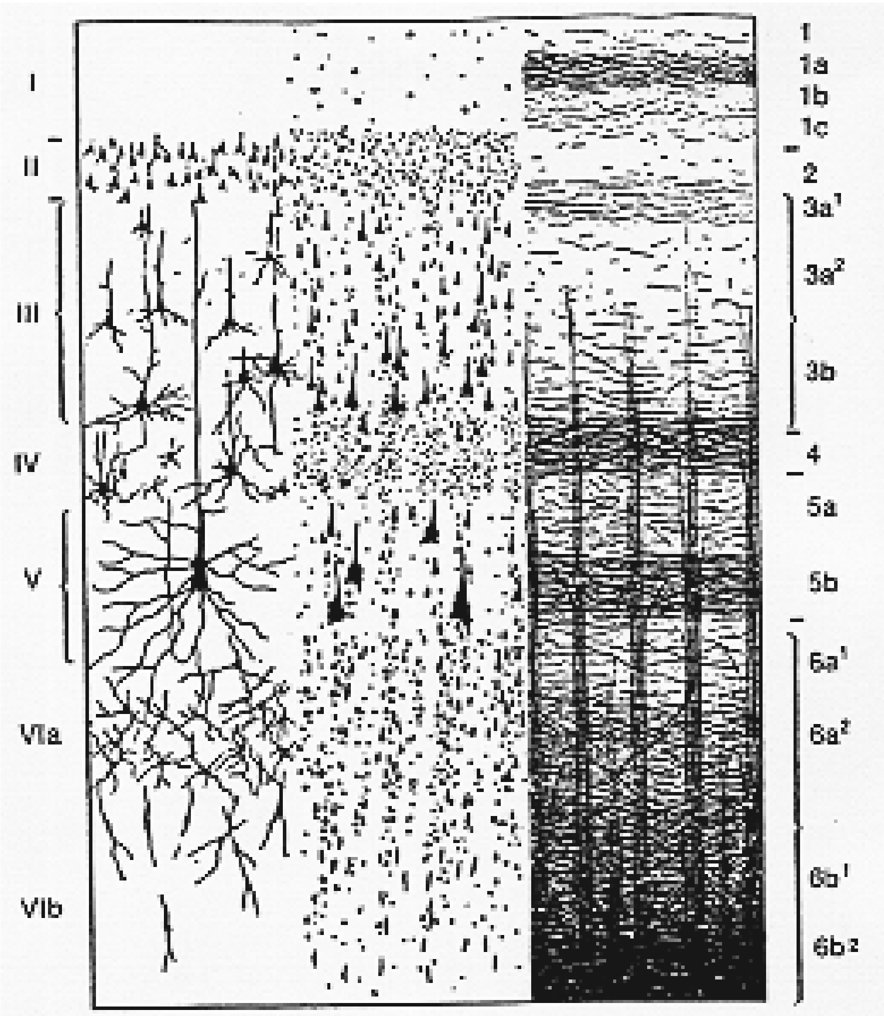

**Easy to be implemented by neurons.**

# Space and Time complexity

- With assumption of sparseness of input vectors of each SOM and sparseness of edges, each node requires
  - O(s) space for the CPT,
  - O(log s) time for parallel execution of one step of approx. BP,
    where ***s*** is a number of units in each node.
- This means the system is **scalable**.
  - It is **a necessary condition for brain models**, the large-scale information processor.

# The correspondence of the approx. algorithm to the anatomical structure

# Six-layer structure of cerebral cortex

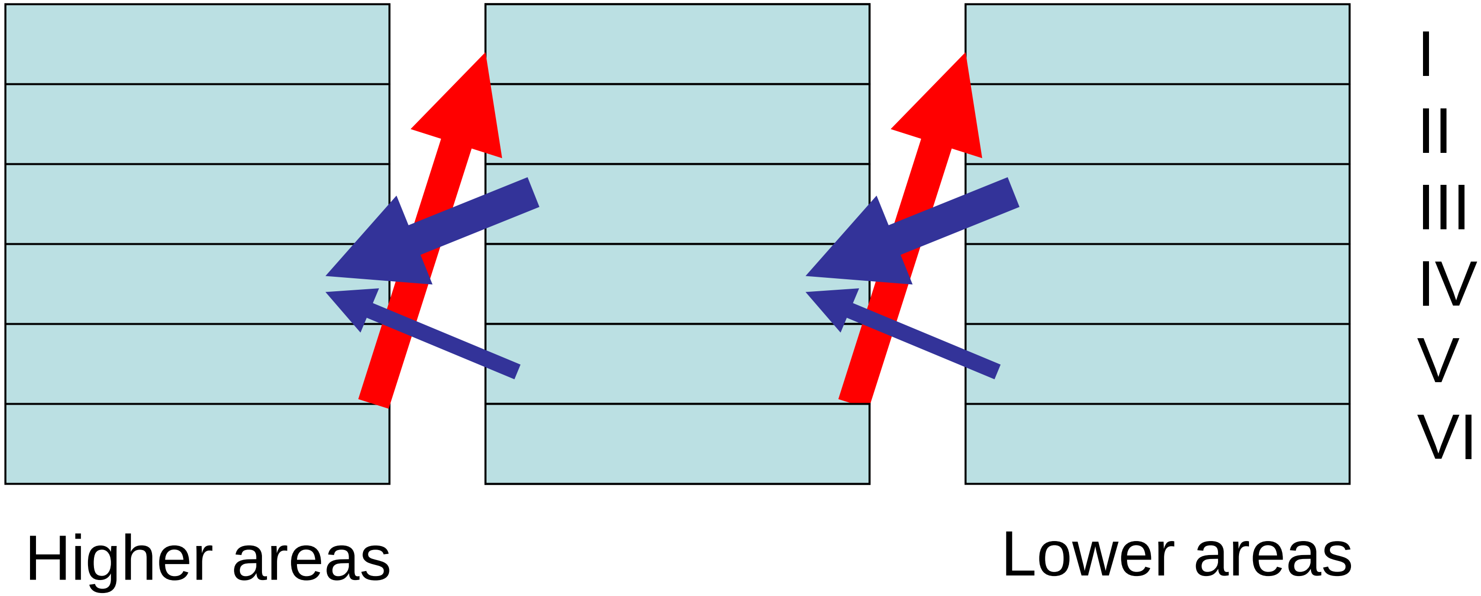



K. Brodmann, Vergleichende Lokalisation der Grosshirnrinde. in: ihren Prinzipien dargestellt auf Grund des Zellenbaues,. J.A. Barth, Leipzig, 1909.

This figure is taken from the following Web page.
http://web.sc.itc.keio.ac.jp/anatomy/brodal/chapter12.html

# Connection rule [Pandya and Yeterian 1985]

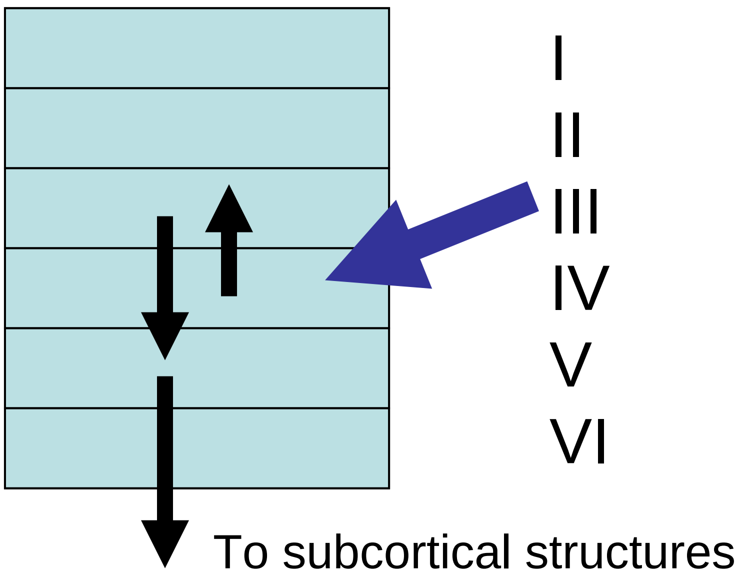
- **Bottom up** connections from layer 3 to 4.
  - Some connections from layer 5 to 4.
- **Top down** connections from layer 5/6 to 1.
  - (A few connections from layer 3 to 1.)



# Information flow in cortex [Gilbert 1983]

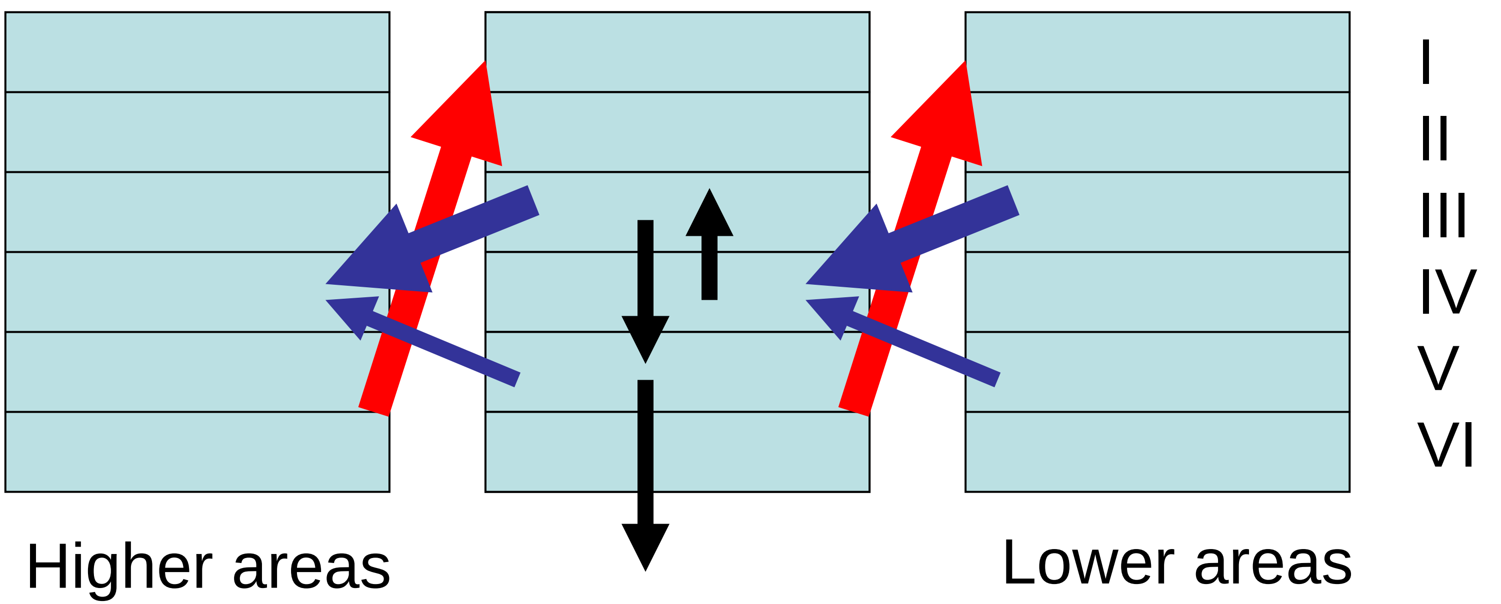
- Lower areas - 4 - 2/3 - 5 - Subcortical structures

  (And flow for recurrent input, 5 - 6 - 4. )



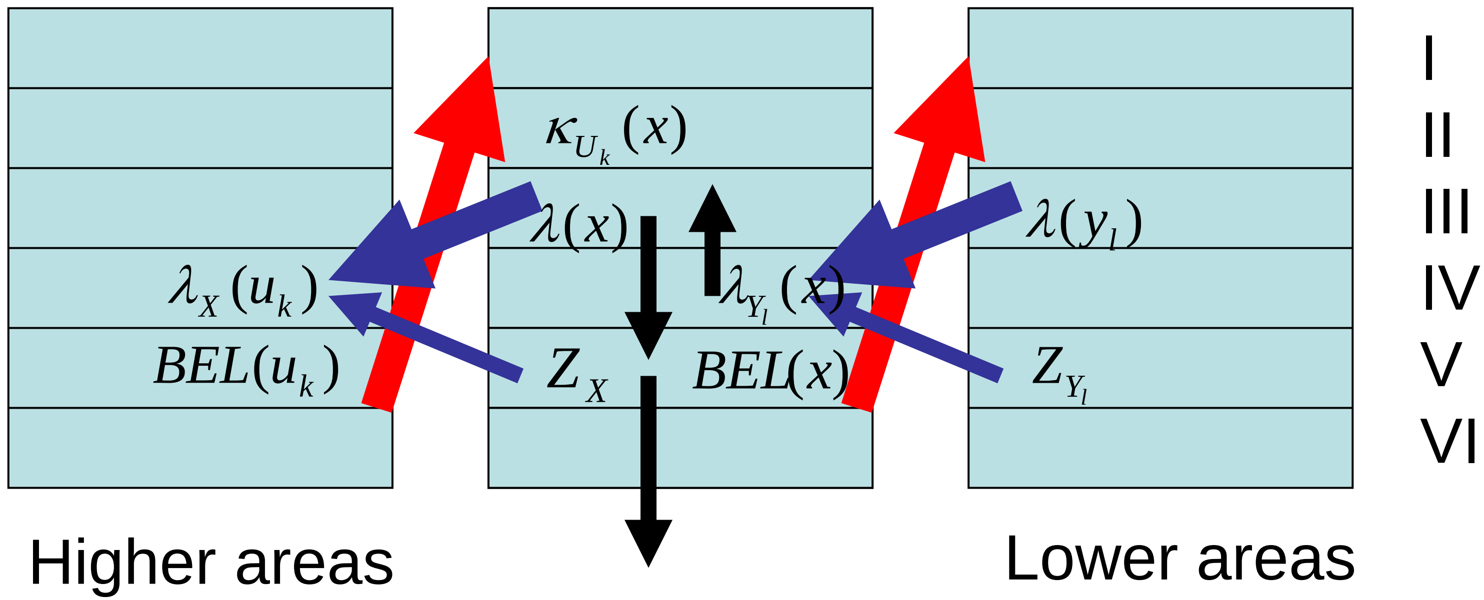
# Very strange structure

- Intermediate results, in layer 3, are sent to the higher areas.
- Final results, in layer 5, are sent back to the lower areas.

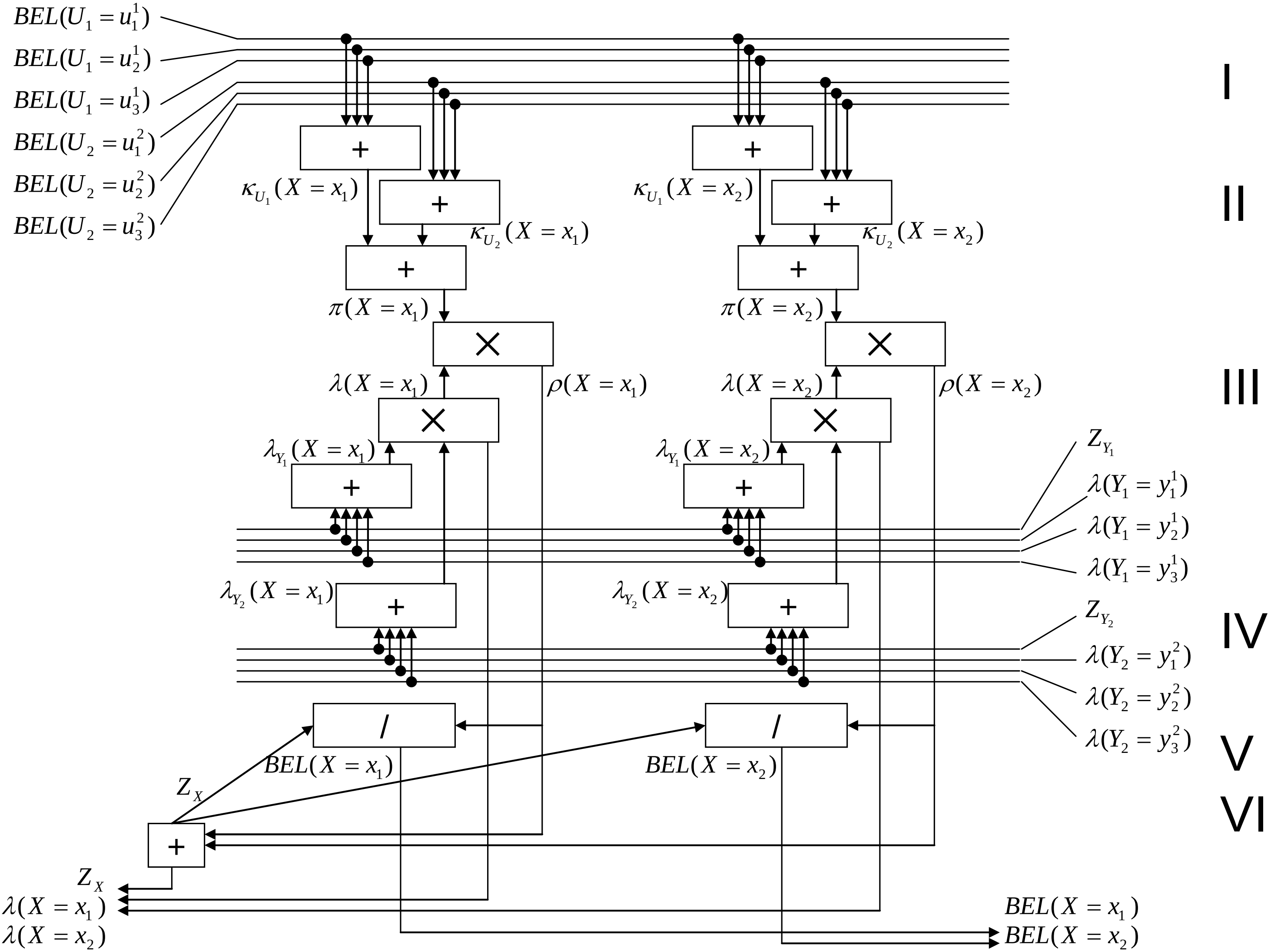


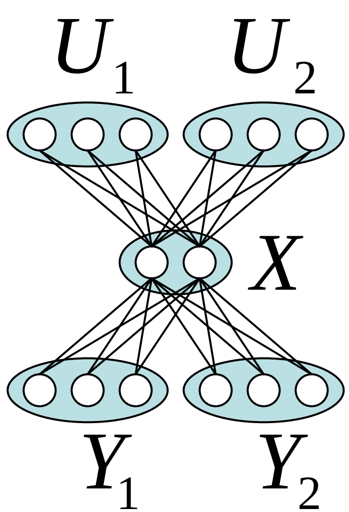
# Assigning Variables of approx. BP

- Surprisingly, **all five communication variables** can be successfully assigned to these layers.



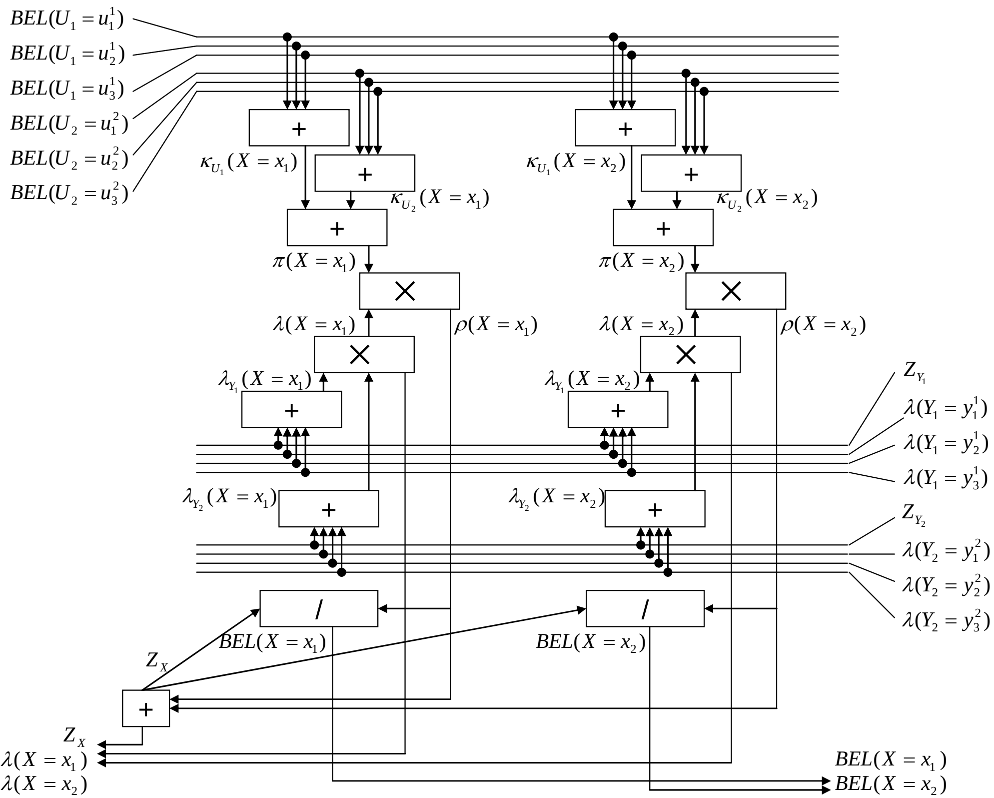
# Detailed structure in columns

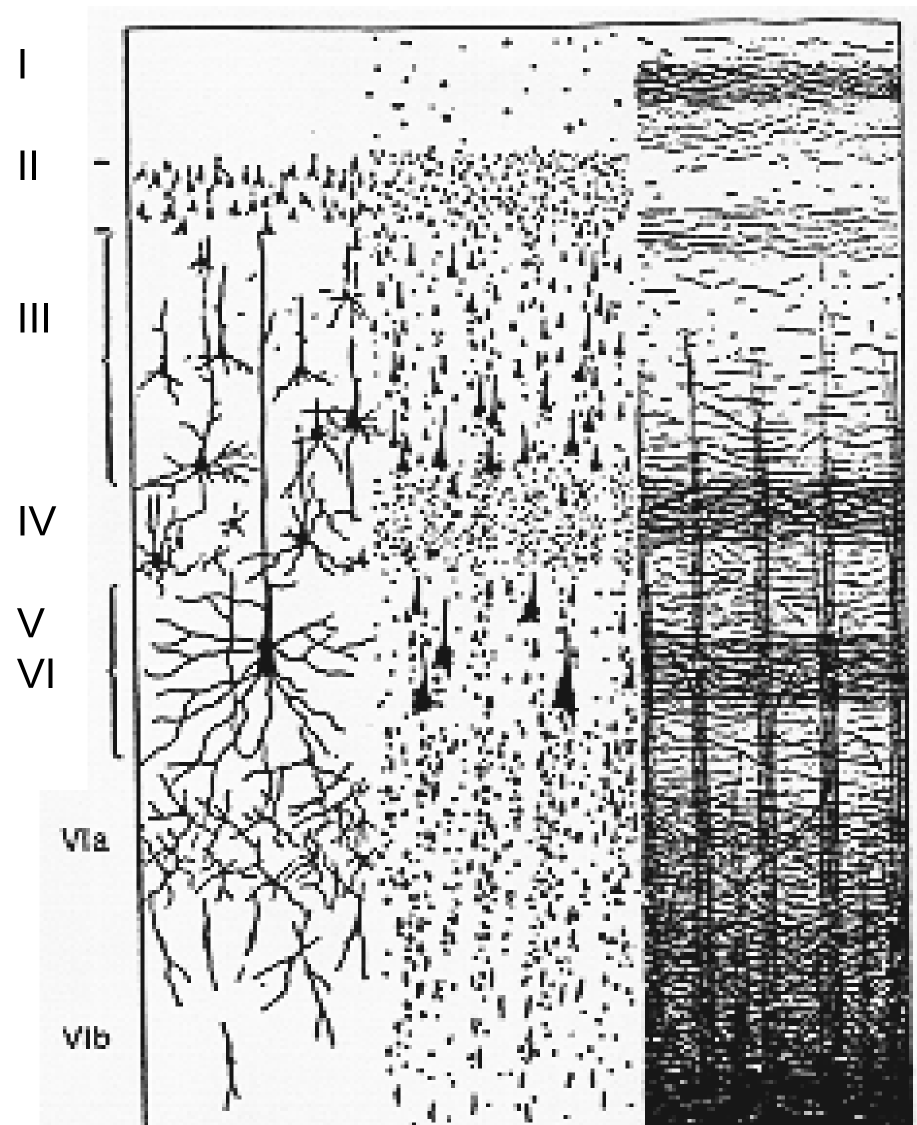


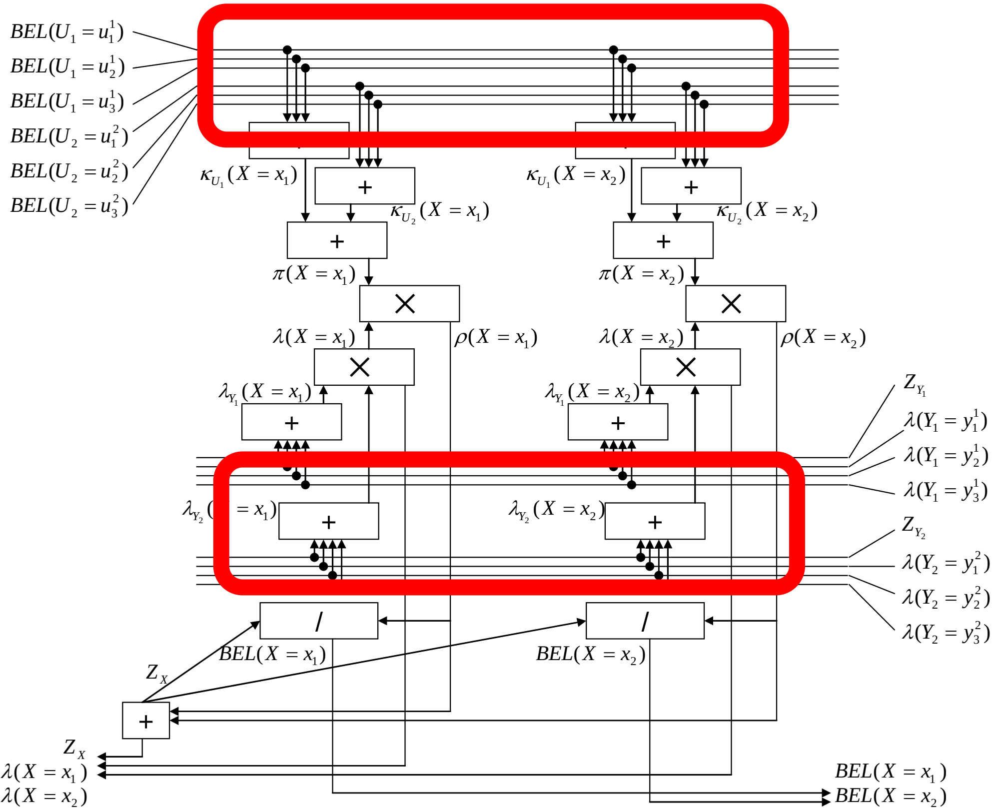

The left circuit calculates values of two units, x1 and x2, in node X in the above network.

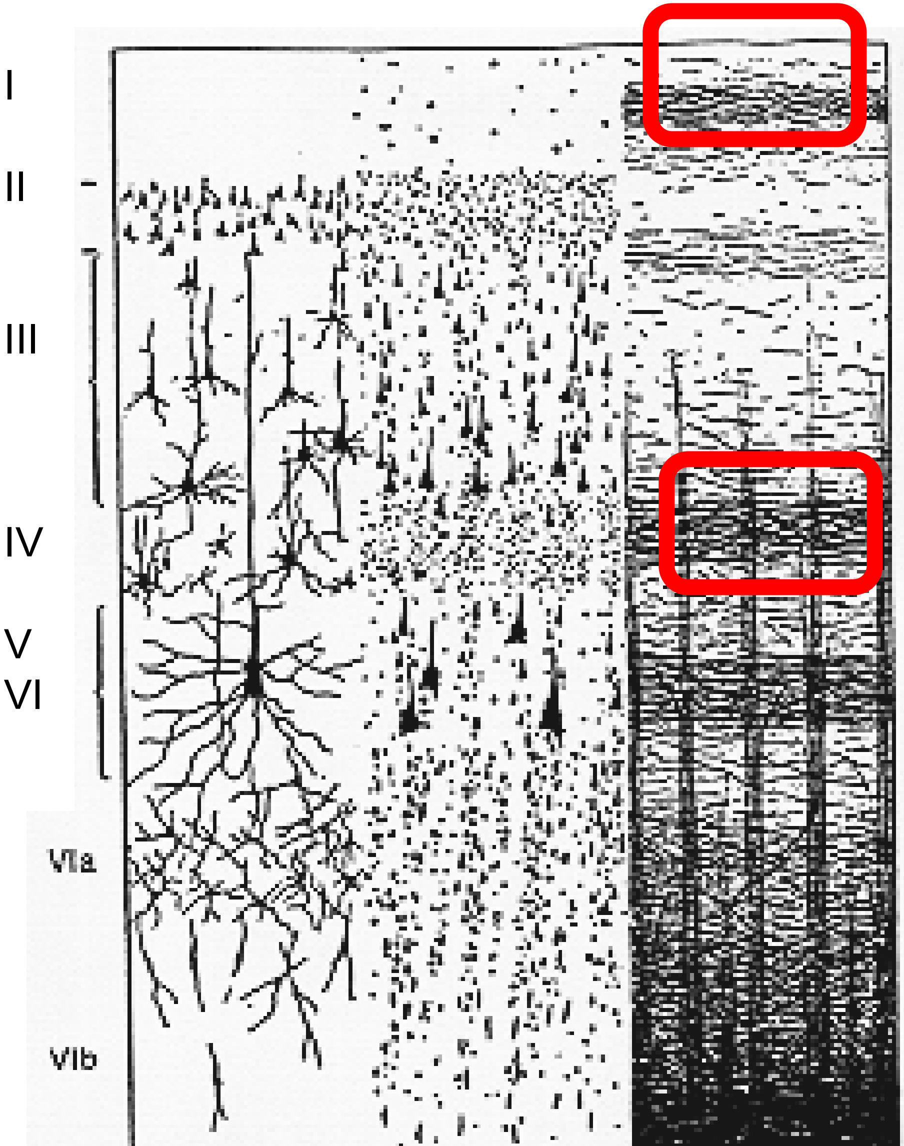
# The structure matches anatomical structure



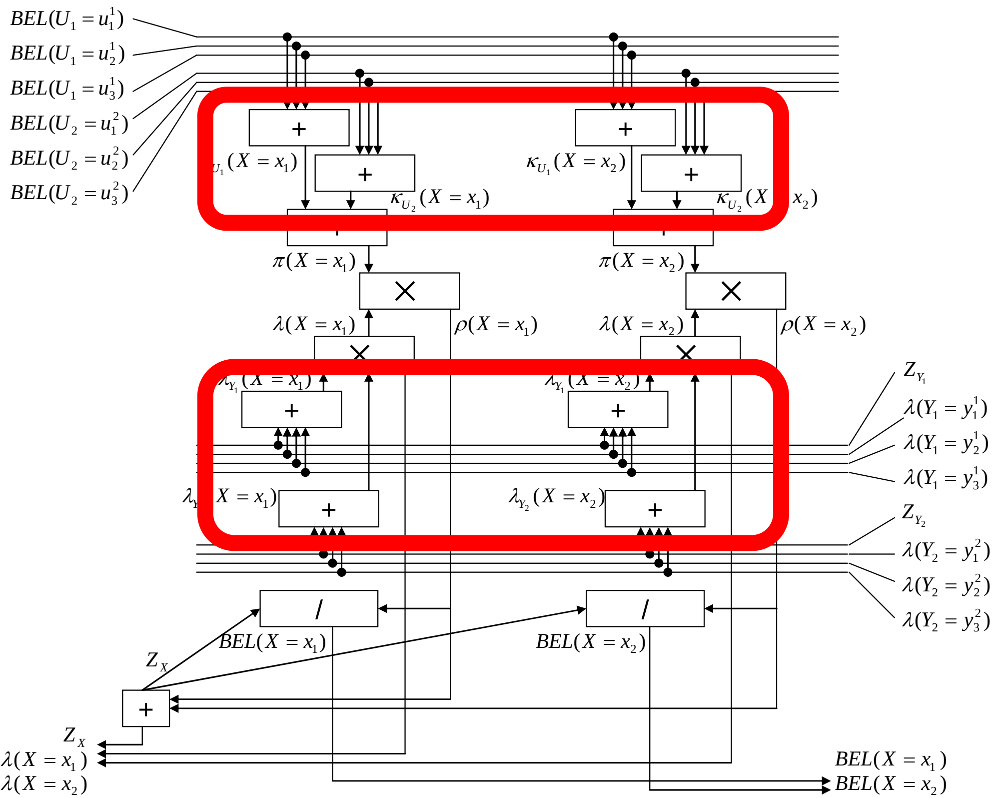



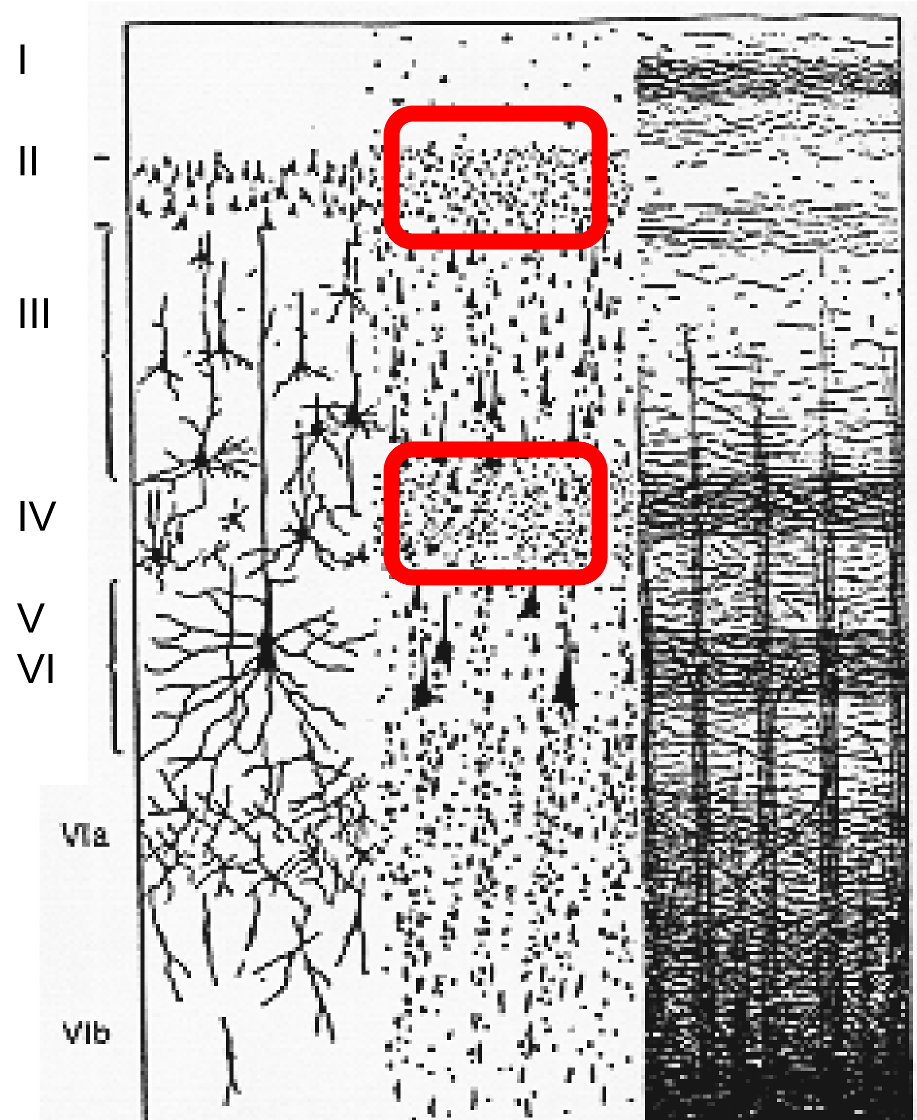
# Horizontal fibers in layer 1 and 4



# Small cells in layer 2 and 4

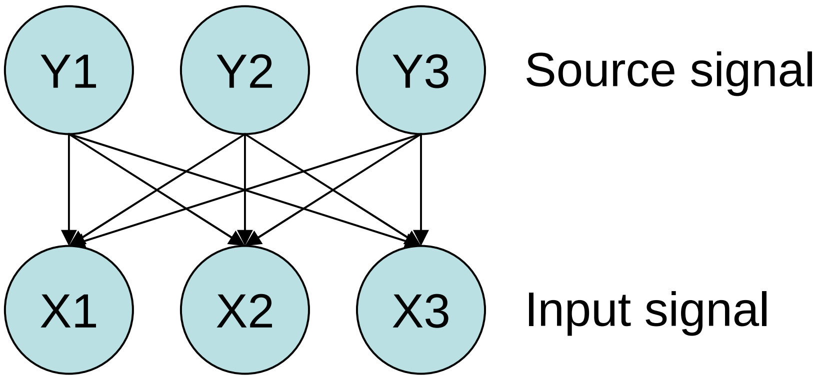


# How the network structure is learned ?
# - My speculation -

# Generative model can be acquired by ICA

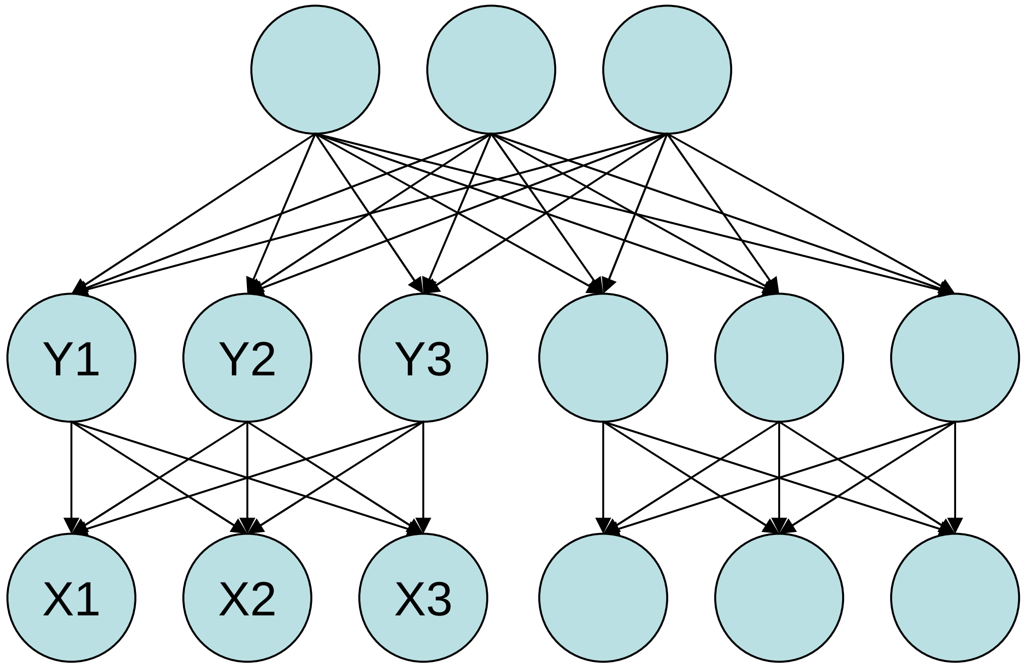
- In other words, ICA may acquire two-layered Bayesian network structure.




$$x=Ay$$

# Hierarchical generative model

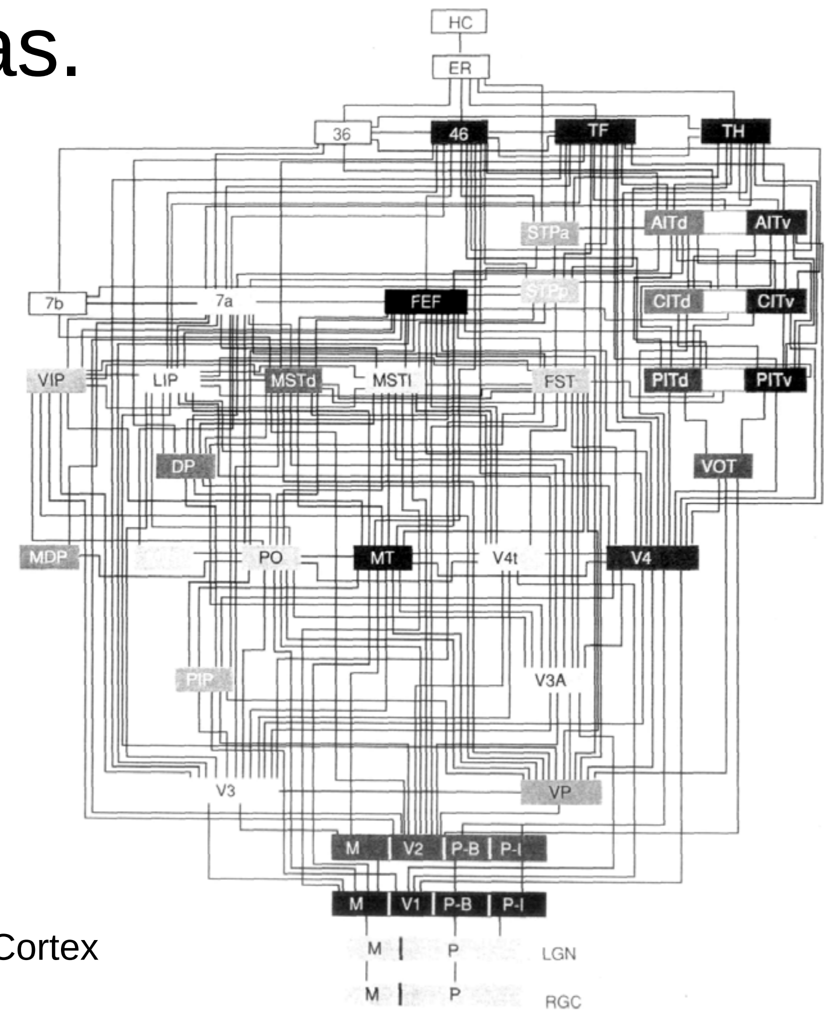
- Hierarchical ICA may acquire multi-layered Bayesian network structure.




Other retinal areas or modalities.

# How ICA corresponds to the anatomical structure ?

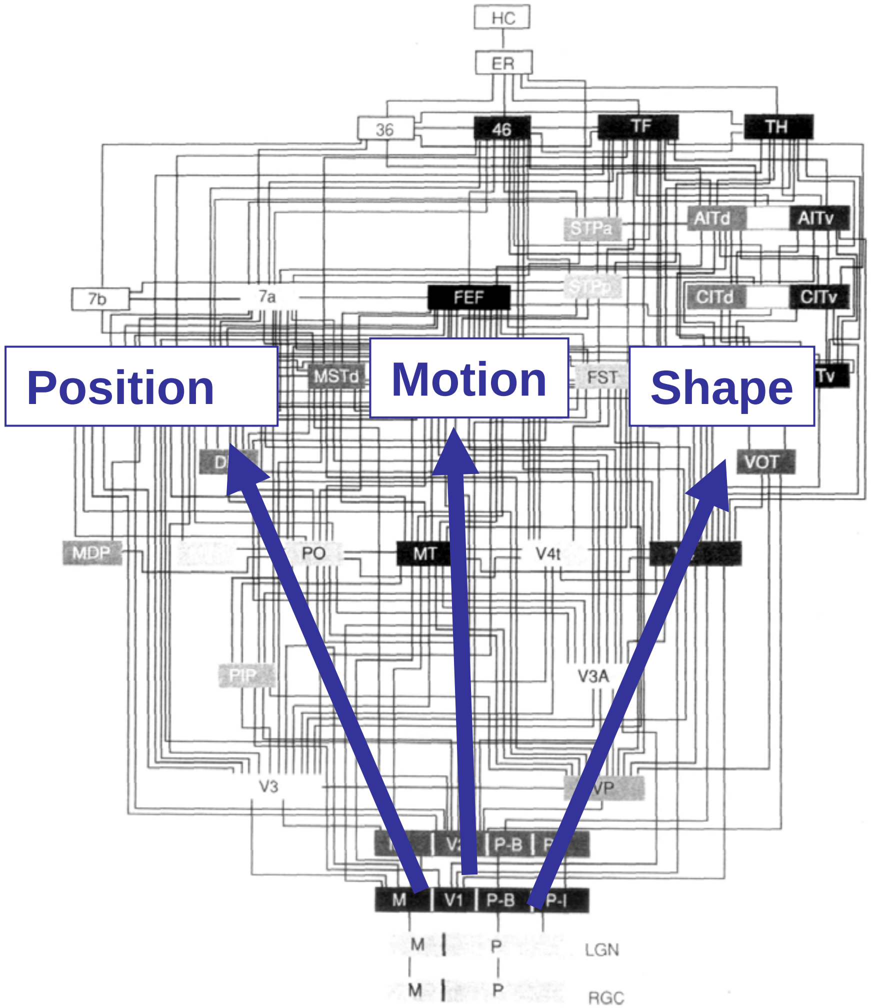
- Network of visual areas.




Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the Primate Cerebral Cortex
Cerebral Cortex 1991 1: 1-47

# Corresponding to the anatomical structure

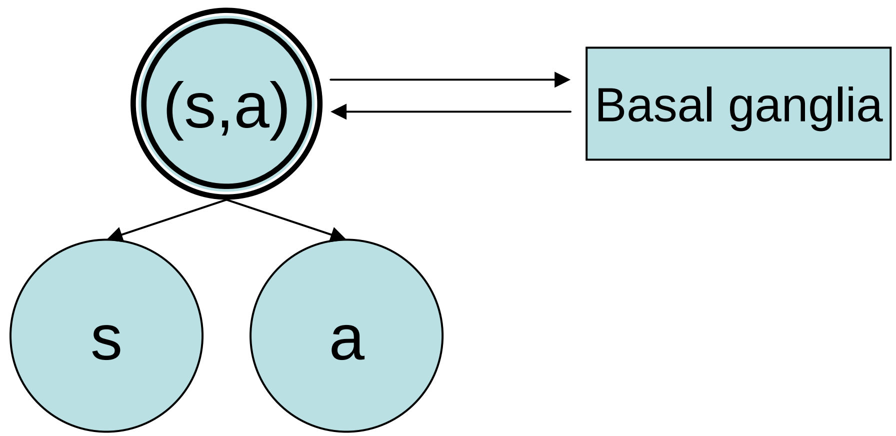
- Position, motion and shape are **independent**.
- Horizontal connections may be for ICA.



# How about motor areas ?

# Reinforcement learning in motor areas

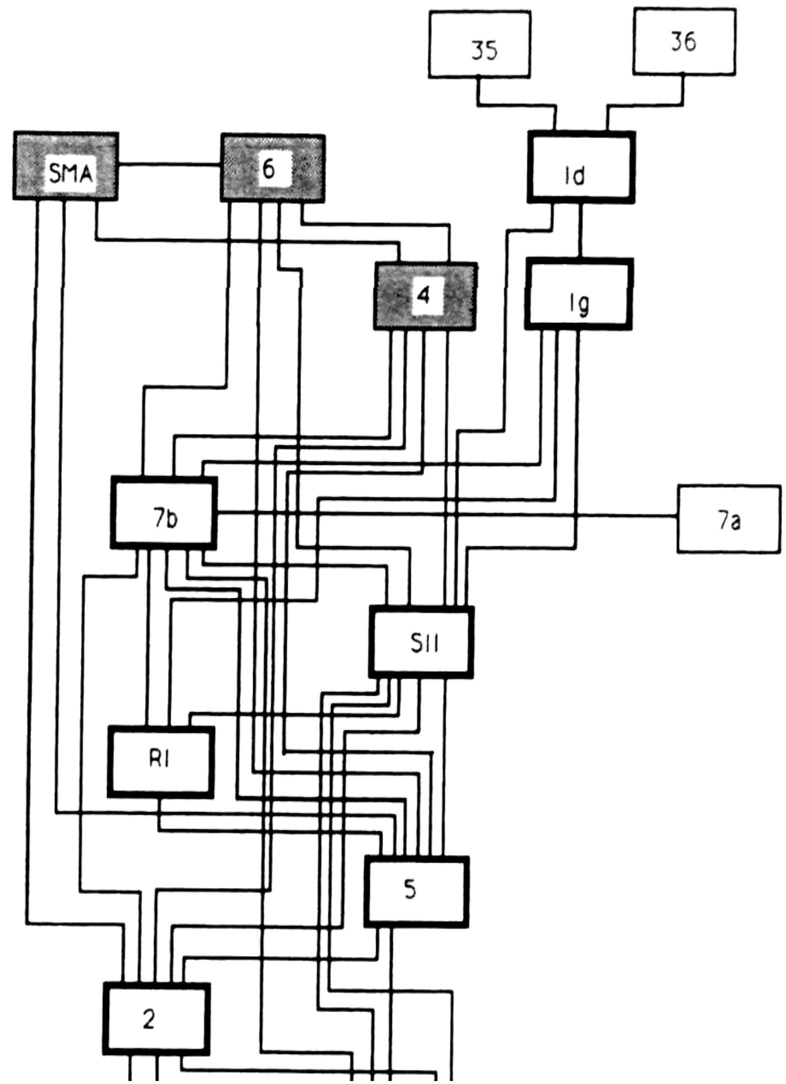
- Nodes acquire state-action pairs. State values are learned by synapses connect to basal ganglia.




- Matches the anatomical structure: cortico-basal ganglia loop.
  - This interpretation is an extension of Doya's model:
    K. Doya, Complementary roles of basal ganglia and cerebellum in learning and motor control, Current Opinion in Neurobiology 10 (6): 732-739 Dec 2000.

# Network of motor areas

- Each motor area certainly has child nodes representing state and action.
  - (M1's action node is spinal cord.)




Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the Primate
Cerebral Cortex 1991 1: 1-47

# Future work

- Sophistication and computer simulation of this model.
  - Need slight modification of approximate BP.
  - Determine concrete algorithms of SOM and ICA.
- Some other important mechanisms.
  - **Selective attention** mechanisms in order to avoid the **curse of dimensionality**.
  - **Novelty detection**, that may be an online **cross-validation** mechanism in order to maximize generalization ability.

# Summary of BESOM model

- Each hyper column in cortex is a node (a random variable) of a Bayesian network.
  - Learning step:
    - Each CPT is self-organized by a SOM.
    - Network structure is self-organized by ICA.
  - Recognition step:
    - State estimation is done by approximate loopy BP algorithm.
- The approx. BP algorithm matches main anatomical structure of cortex.
  - Hopefully, explains main function of cortex.

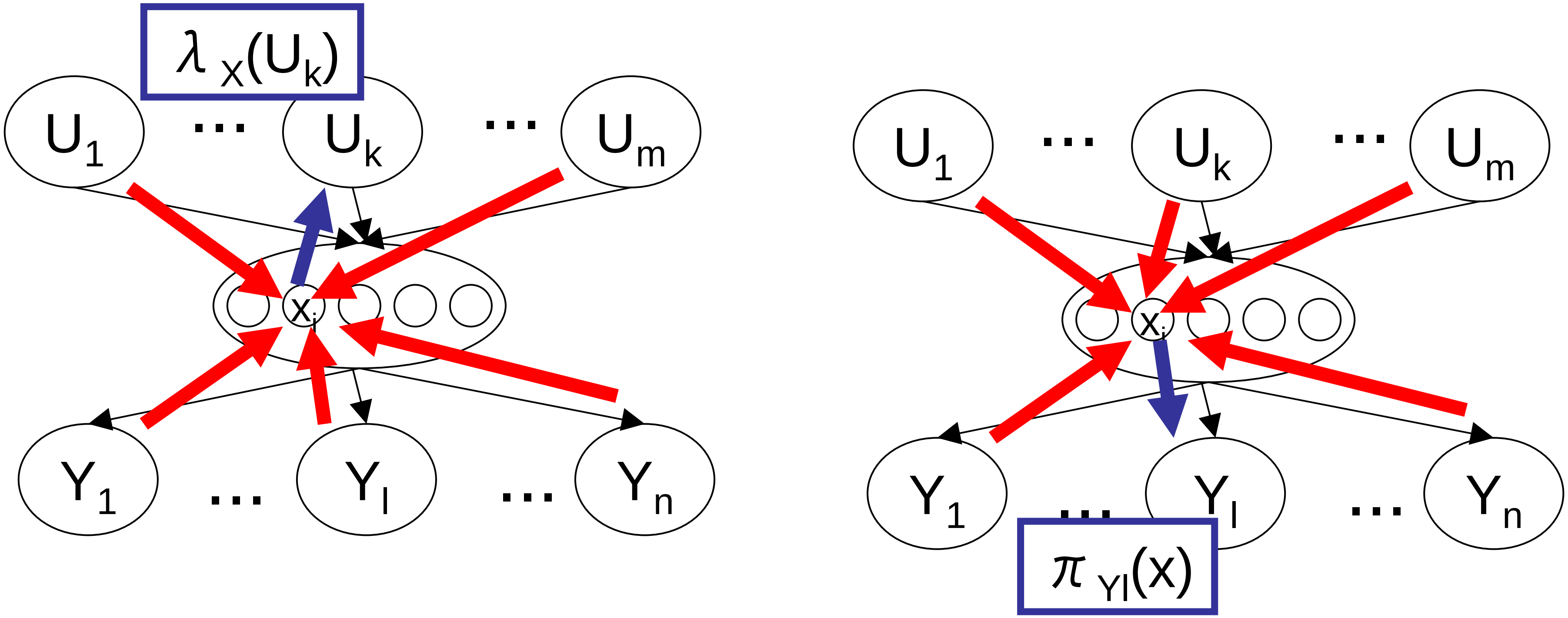
# Supplemental Slides

# Other neuroscience knowledge consistent with BESOM model

- Most information processing is done within the column.
- Output of simple cells in layer 4 are linear sum of input from LGN. Complex cells are nonlinear.
- Neurons in a column represent similar information, independent of their depth.
  - because observation and prediction matches after the learning converged.

# Information from message receiver

- In BP algorithm, messages exclude them.

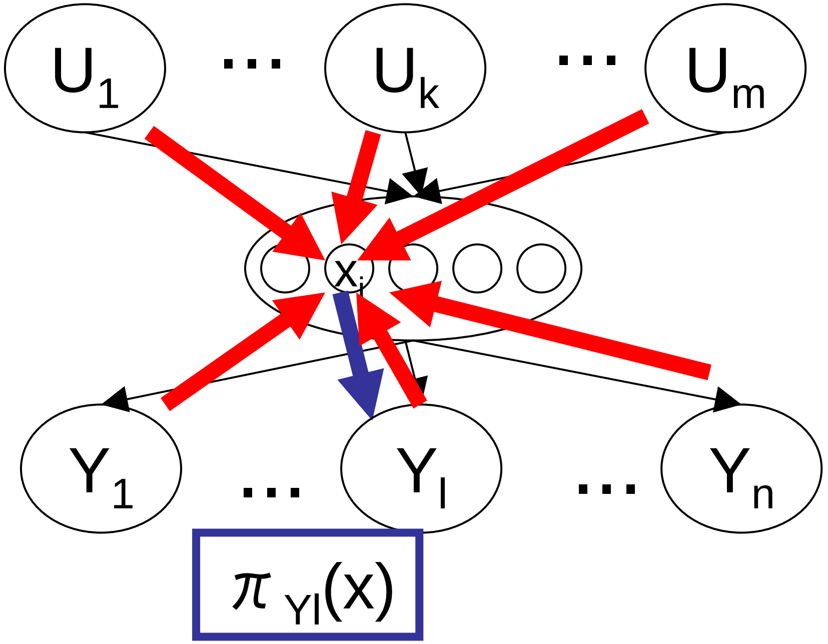



- If there are many parents and children, these information may be included.

# Example: $\pi_{Yl}(x)$ approximation

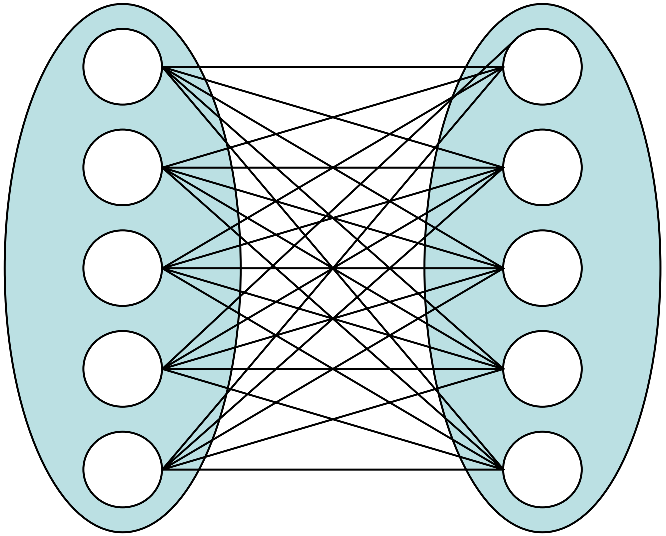
- An message $\pi_{Yl}(x)$ from node X to node $Y_l$ may include information $\lambda_{Yl}(x)$ from $Y_l$ .

$$
\begin{aligned}
&\pi_{Y_l}(x) \\
&= \pi(x)\prod_{j\neq l}\lambda_{Y_j}(x) \\
&\approx \pi(x)\prod_{j}\lambda_{Y_j}(x) \\
&= \lambda(x)\pi(x)
\end{aligned}
$$



# A possible ICA mechanism

- Two random variables may become independent if connected with anti-Hebb synapses.



cf. Naoki Oshiro, Koji Kurata, Tetsuhiko Yamamoto,
"A self-organizing model of place cells with grid-structured receptive fields", Artificial Life and Robotics, Vol.11, No.1, pp.48--51, 2007