Computational Model of the Cerebral Cortex that Performs Sparse Coding Using a Bayesian Network and Self-Organizing Maps

ICONIP 2010
2010-11-22

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

Department of Computer Science
The University of Tokyo, Japan
Haruo Hosoya

Parameter Learning of a Cerebral Cortex Model based on a Bayesian Network

AMBN2010
2010-11-18,19

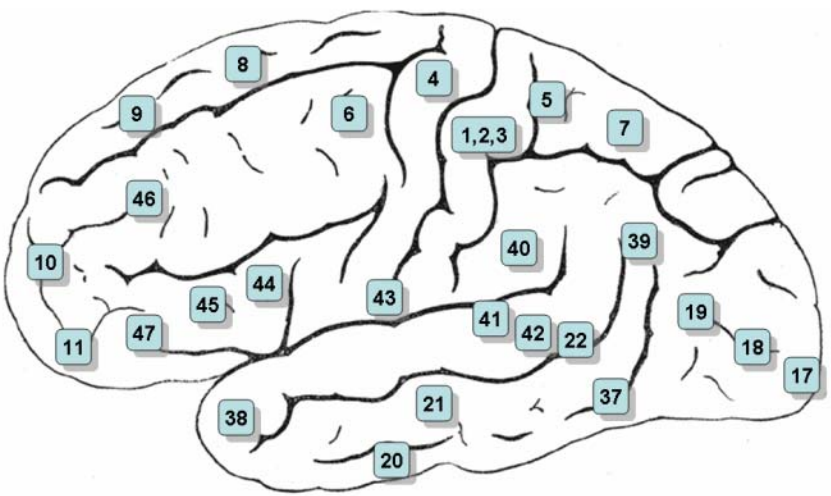
National Institute of Advanced Industrial Science and Technology(AIST)
Yuuji Ichisugi

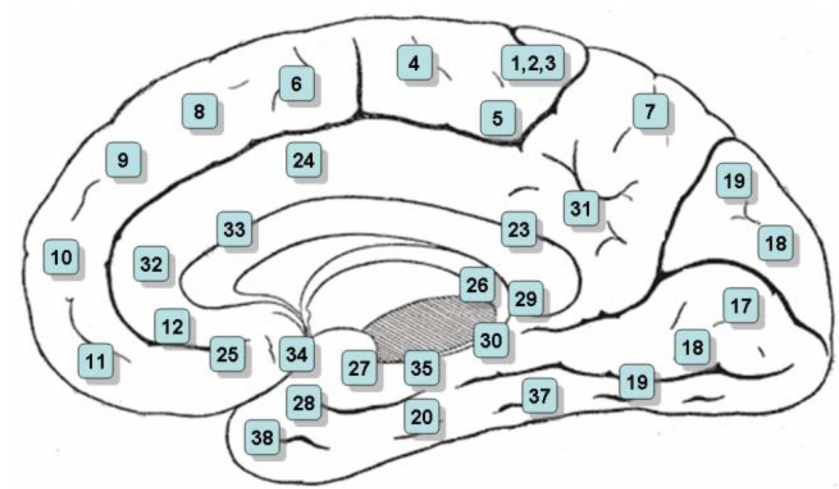
Abstract

- Some computational neuroscientists have begun to understand that the **Bayesian network** is the essential mechanism of the cerebral cortex.
- We propose a biologically plausible computational model that unifies a **Bayesian network model** and **sparse-coding model**.
- This model is an extension of our previous BESOM model [Ichisugi 2007].

Cerebral cortex

- Realizes human's intelligence.
- The principle of the cortex has not been revealed yet.





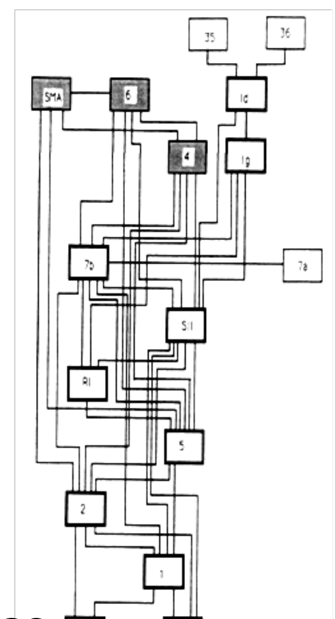
from http://en.wikipedia.org/wiki/Brodmann_area

Areas in cerebral cortex

- Each area has its own function.
  - Visual area, Motor area, Language area, etc.
- Bidirectional connection between areas.

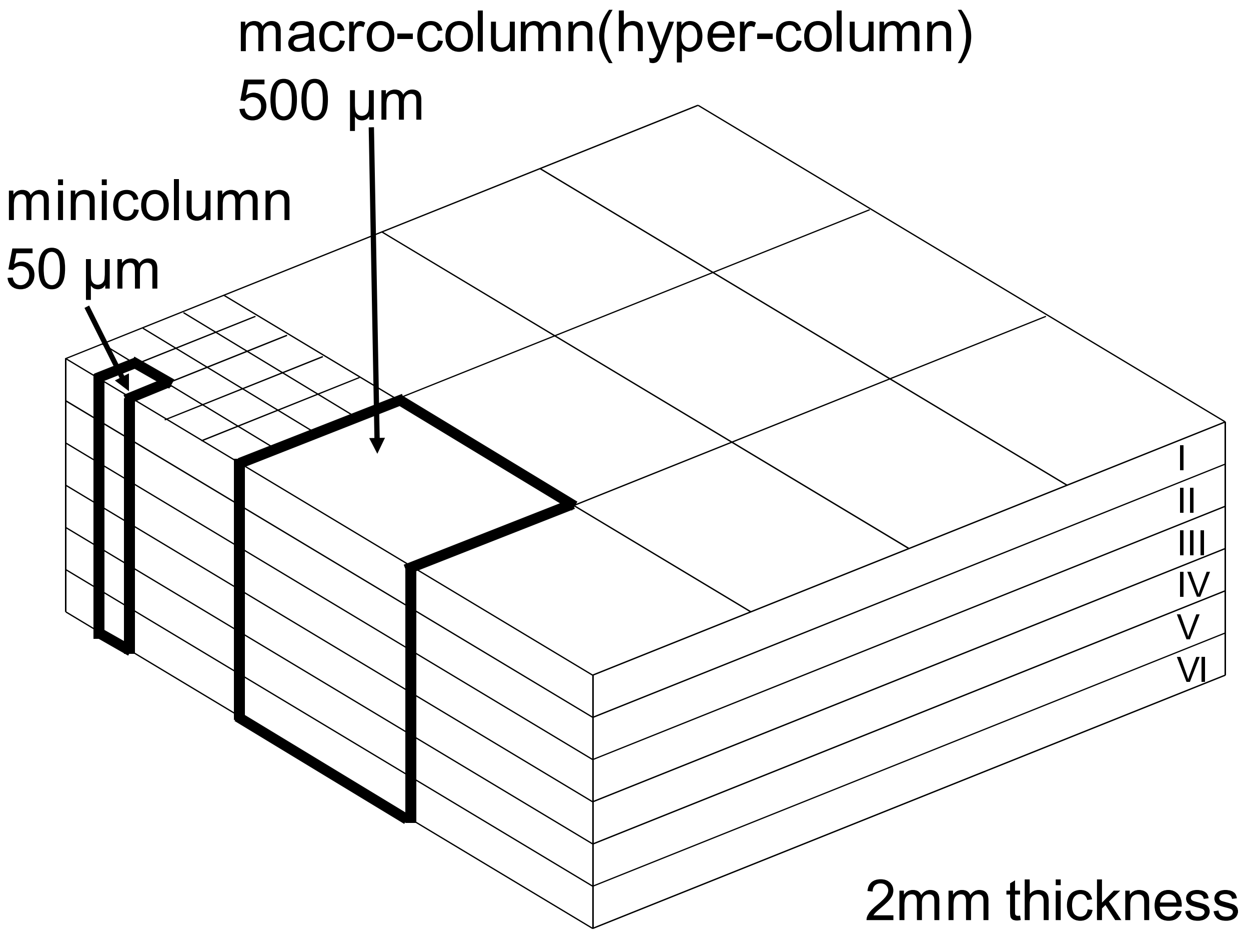
Visual areas

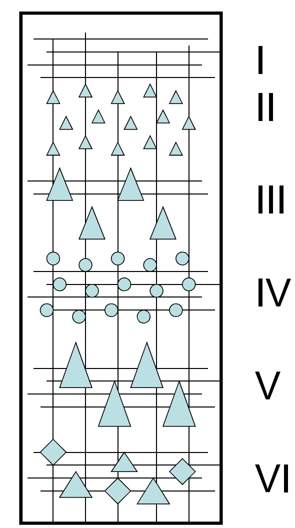
Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the Primate Cerebral Cortex
Cerebral Cortex 1991 1: 1-47



Motor areas

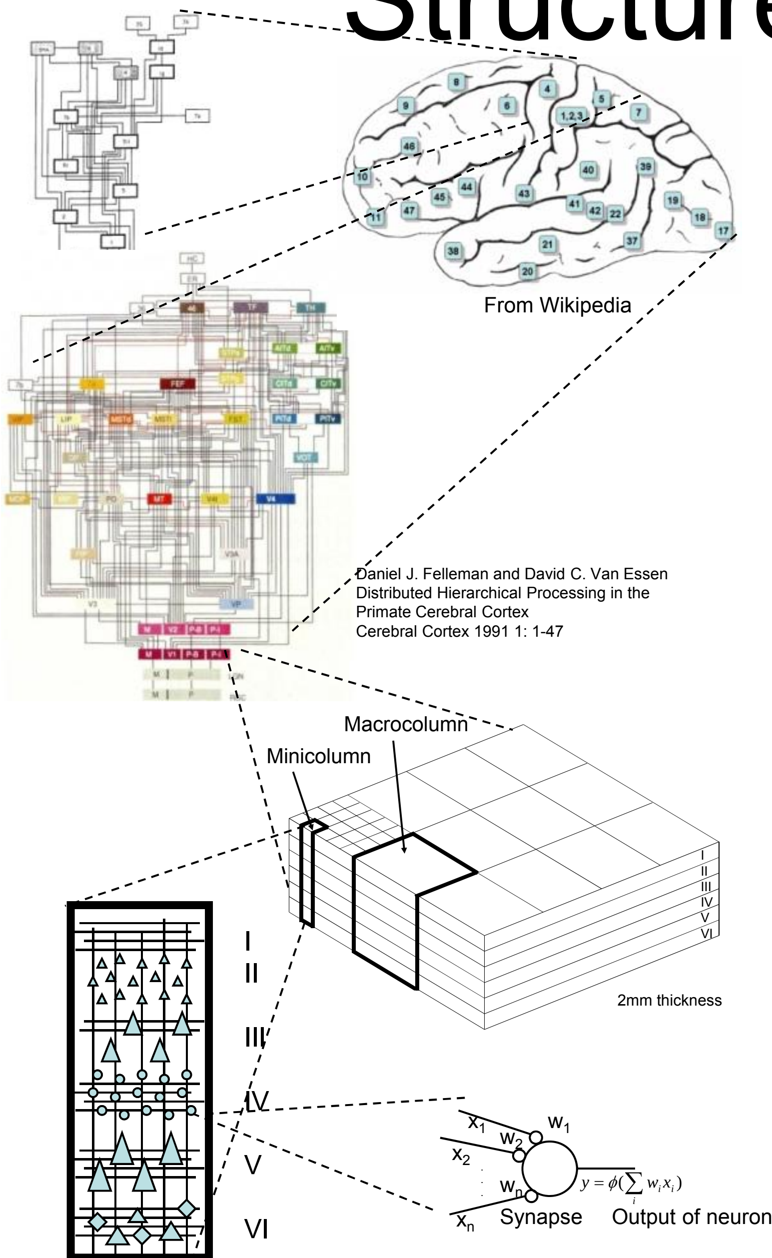
Column and 6-layer structure of cerebral cortex





6-layer structure

Structure of cerebral cortex



- 10^2 areas
- 10^6 macrocolumns
- 10^8 minicolumns
- 10^10 neurons
- 10^14 synapses

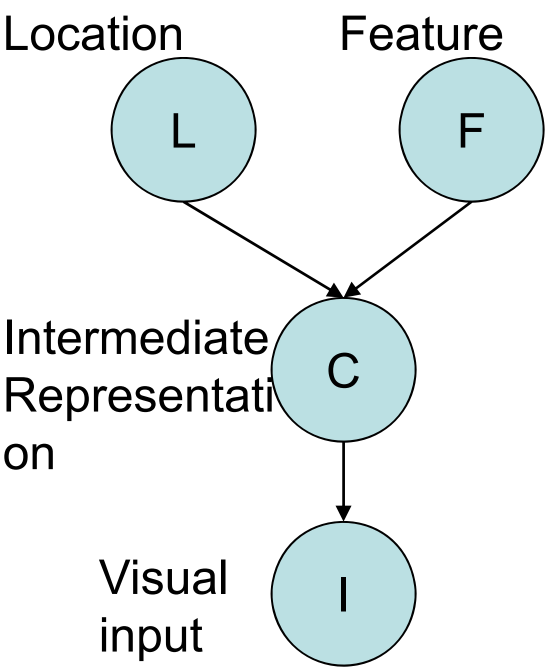
Cerebral cortex models based on Bayesian networks

- [Lee and Mumford 2003]
- [George and Hawkins 2005]
- [Rao 2005]
- [Ichisugi 2007]
- [Rohrbein, Eggert and Korner 2008]
- [Hosoya 2009]
- [Litvak and Ullman 2009]
- [Chikkerur, Serre, Tan and Poggio 2010]

– These models try to explain the **essential mechanism of cortex**, as opposed to previous phenomenological models.

Rao's model [Rao 2005]

- Bayesian network model to explain electrophysiological phenomena.



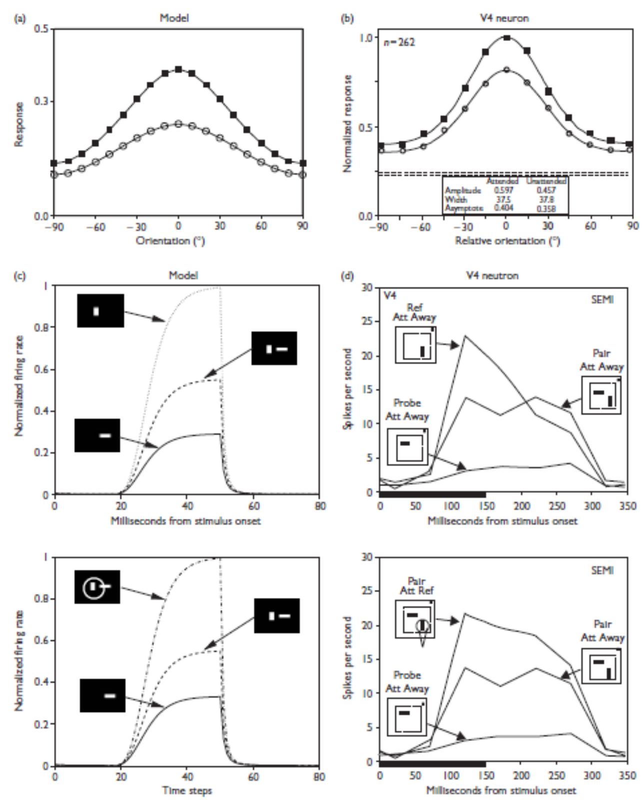


Fig. 2 Multiplicative modulation and response restoration. (a) Orientation-tuning curve of a feature-coding model neuron with a preferred stimulus orientation of 0° with (filled squares) and without (unfilled circles) attention. (b) Normalized orientation-tuning curves for a population of V4 neurons with (filled squares) and without attention (unfilled circles) (reproduced from [3], copyright 1999 by the Society of Neuroscience). (c) Top panel: the three line plots represent the vertical feature-coding neuron's response to a vertical bar ('reference', Ref), a horizontal bar at a different position ('probe'), and both bars presented simultaneously ('pair'). In each case, the input lasted 30 time steps, beginning at time step 20. Bottom panel: when 'attention' (Att) (depicted as a white oval) is focused on the vertical bar, the firing rate for the pair stimulus approximates the firing rate obtained for the reference alone. (d) Top panel: responses from a V4 neuron without attention (reproduced from [4], copyright 1999 by the Society of Neuroscience). Bottom panel: responses from the same neuron when attending to the vertical bar (see condition Pair Att Ref) (reproduced from [4], copyright 1999 by the Society for Neuroscience).

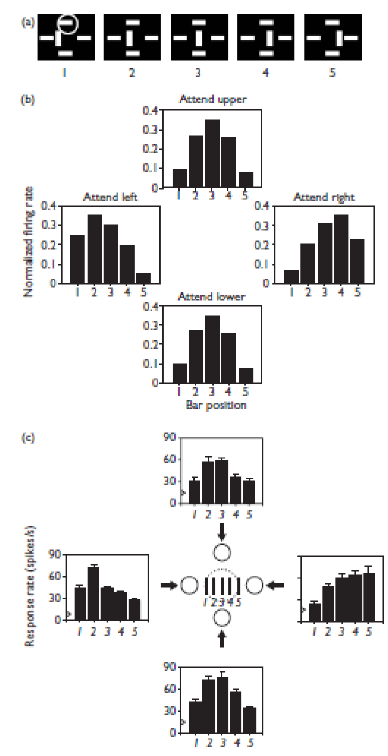
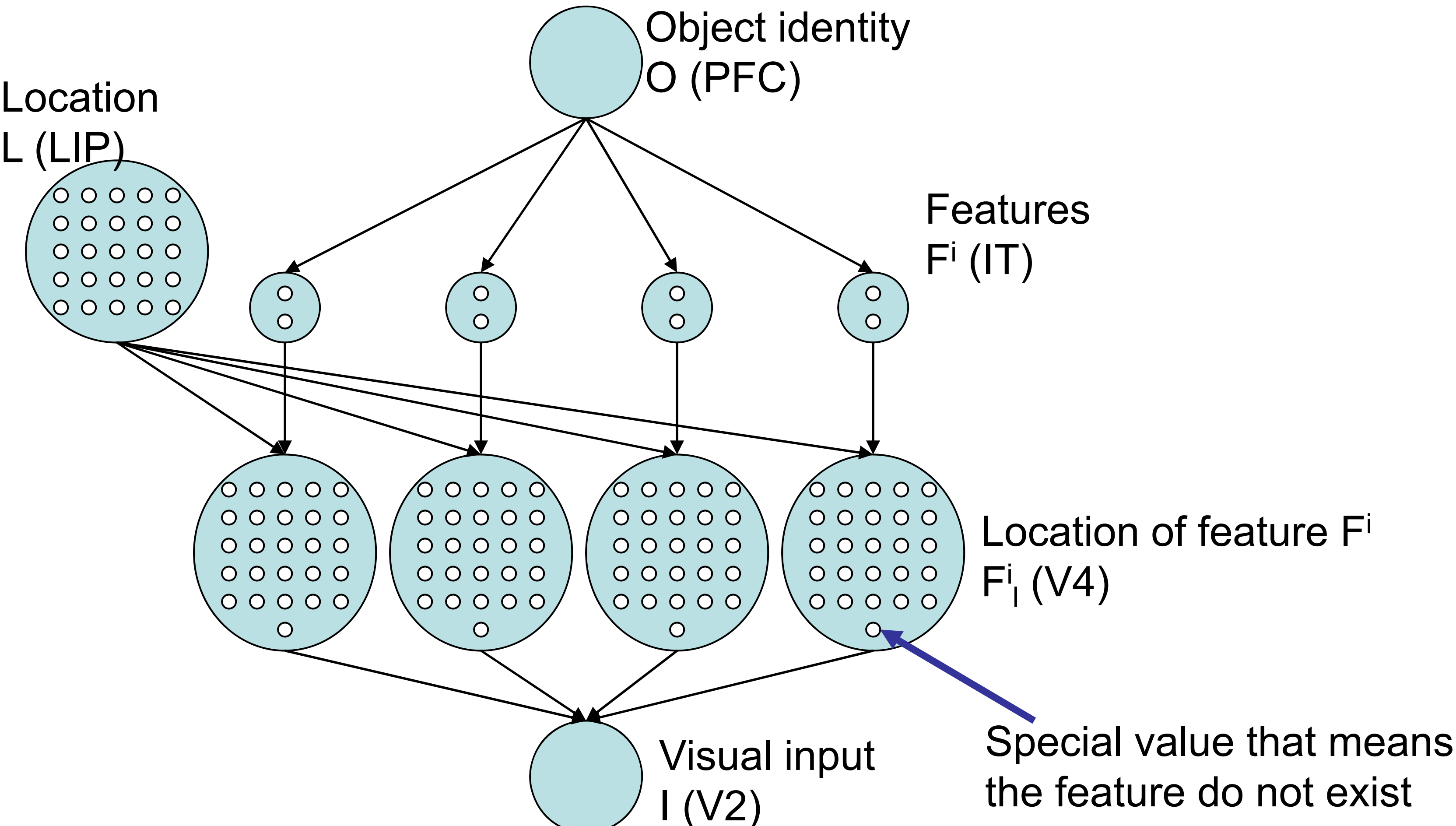


Fig. 3 Influence of attention on neighboring spatial locations. (a) Example trial based on Connor et al.'s experiments [5] showing five images, each containing four horizontal bars and one vertical bar. Attention was focused on a horizontal bar (e.g. upper bar, circled) while the vertical bar's position was varied. (b) Responses of the vertical feature-coding model neuron. Each plot shows five responses, one for each location of the vertical bar, as attention was focused on the upper, lower, left, or right horizontal bar. (c) Responses from a V4 neuron (reproduced from [5], copyright 1997 by the Society for Neuroscience).

R. Rao. Bayesian inference and attention in the visual cortex. Neuroreport 16(16), 1843-1848, 2005.

Chikkerur 's model[Chikkerur et al. 2010]

Explains electrophysiological and psychophysical phenomena.

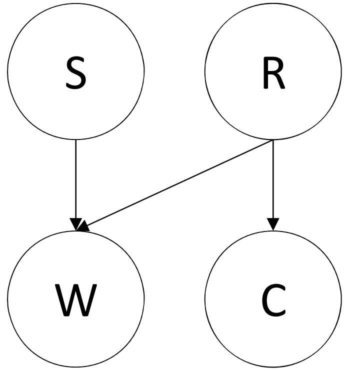


Chikkerur, S., T. Serre, C. Tan and T. Poggio,
What and Where: A Bayesian Inference Theory of Attention,
Vision Research, 2010.

What is Bayesian network?

– **Very efficient and expressive data structure for probabilistic knowledge.**
  - If a joint probability table can be factored into small **conditional probability tables (CPTs)**, time and space complexity will decrease.

ex.: $P(S,W,R,C) = P(W \mid S,R)P(C \mid R)P(S)P(R)$



CPTs

P(S=yes)
0.2

P(R=yes)
0.02

S	R	P(W=yes|S,R)
no	no	0.12
no	yes	0.8
yes	no	0.9
yes	yes	0.98

R	P(C=yes|R)
no	0.3
yes	0.995

Size 4+2+1+1=8

Similarities between Cerebral Cortex and Bayesian network

- Asymmetric and bidirectional connections between lower and higher areas.
- Local and asynchronous communications.
- Non negative values.
- Normalization of values.
- Hebb's learning rule.
- Context dependent recognition.
- Behavior based on Bayesian Statistics.

Precise correspondence between Bayesian networks and anatomical characteristics

[Ichisugi 2007]

Belief propagation algorithm [Pearl 1988]

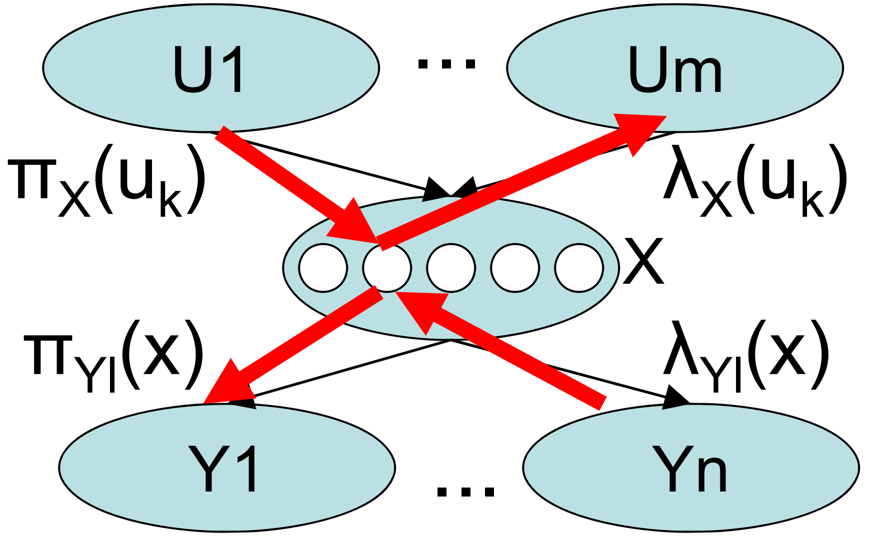
$$BEL(x) = \alpha\lambda(x)\pi(x)$$

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

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

$$\pi_{Y_l}(x) = \beta_1\pi(x)\prod_{j\neq l}\lambda_{Y_j}(x)$$

$$\lambda_X(u_k) = \beta_2\sum_x \lambda(x) \sum_{u_1,\cdots,u_m/u_k} P(x|u_1,\cdots,u_m)\prod_{i\neq k}\pi_X(u_i)$$



It's hard to be implemented by neurons.

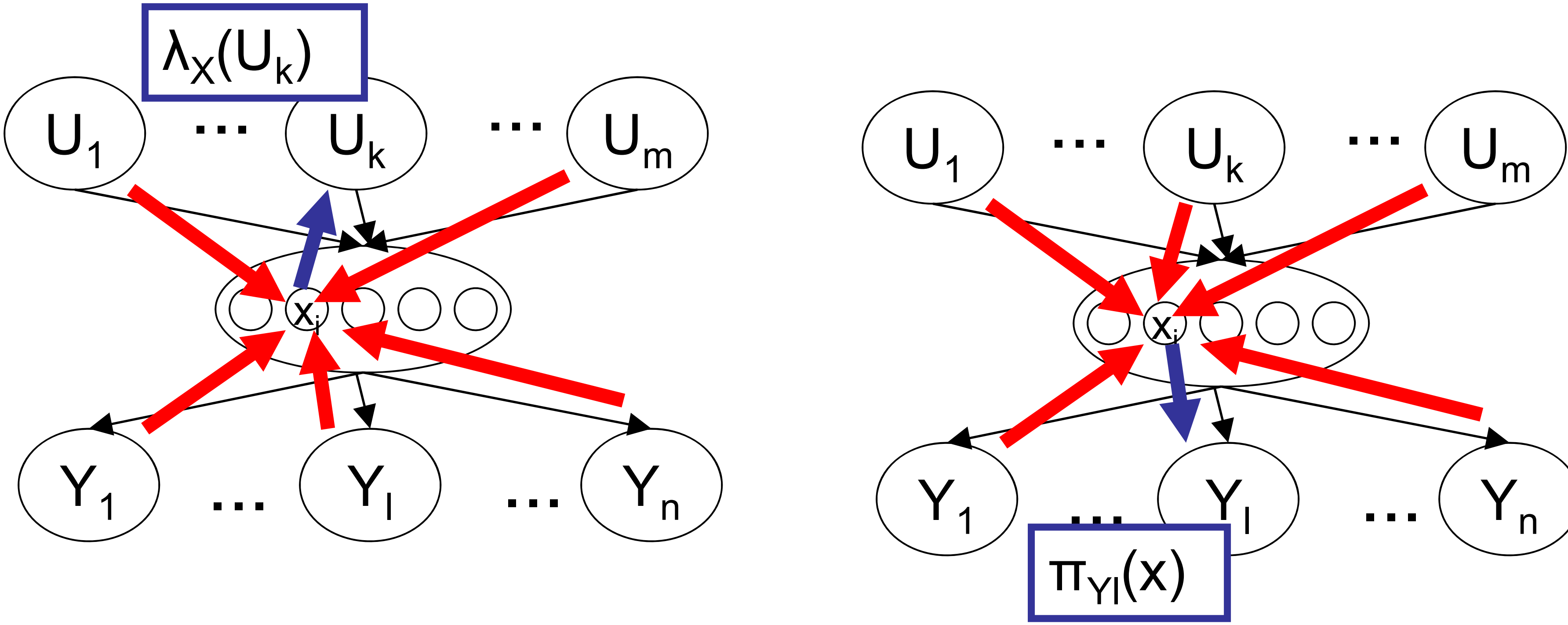
In order to approximate

- Assumption 1: Linear sum CPT model:
  - Qualitatively similar to noisy-OR model [Pearl 1988]

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

- Assumption 2: Nodes have many parent and child nodes.

Messages of BP exclude information from their target

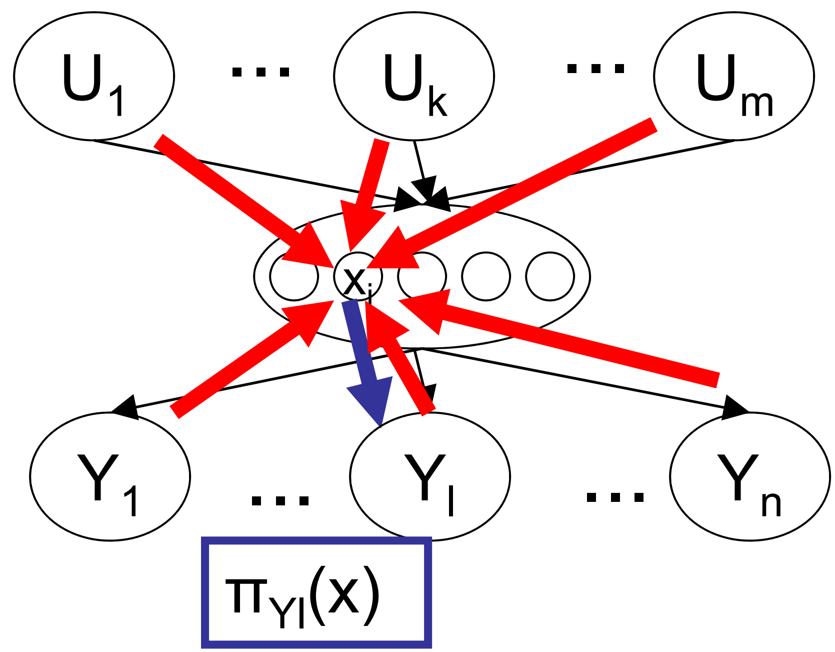


- If there are many parents and children, these information may be included.

Example: $\pi_{Y_l}(x)$ approximation

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

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



Approx. Belief Propagation[Ichisugi 2007]

$$\mathbf{l}_{XY}^{t+1} = \mathbf{z}_Y^t + \mathbf{W}_{XY}\mathbf{o}_Y^t$$

$$\mathbf{o}_X^{t+1} = \prod_{Y \in children(X)}^{\otimes} \mathbf{l}_{XY}^{t+1}$$

$$\mathbf{k}_{UX}^{t+1} = \mathbf{W}_{UX}^T \mathbf{b}_U^t$$

$$\mathbf{p}_X^{t+1} = \sum_{U \in parents(X)} \mathbf{k}_{UX}^{t+1}$$

$$\mathbf{r}_X^{t+1} = \mathbf{o}_X^{t+1} \otimes \mathbf{p}_X^{t+1}$$

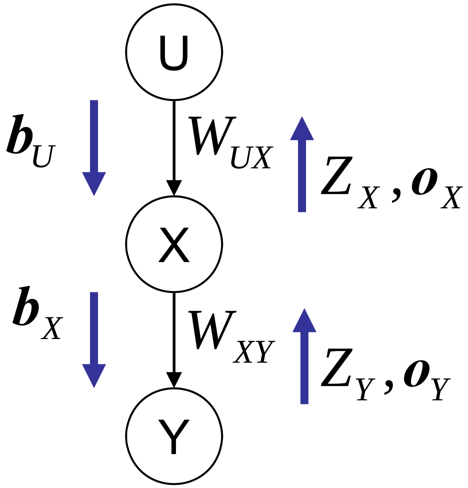
$$Z_X^{t+1} = \sum_i (\mathbf{r}_X^{t+1})_i \; (= \left\|\mathbf{r}_X^{t+1}\right\|_1 = \mathbf{o}_X^{t+1} \bullet \mathbf{p}_X^{t+1})$$

$$\mathbf{z}_X^{t+1} = (Z_X^{t+1}, Z_X^{t+1}, \cdots, Z_X^{t+1})^T$$

$$\mathbf{b}_X^{t+1} = (1/Z_X^{t+1})\mathbf{r}_X^{t+1}$$

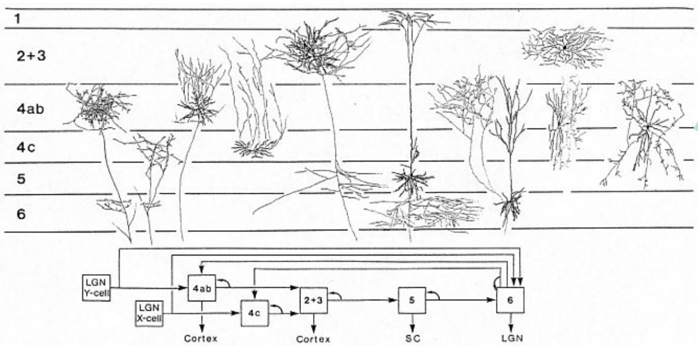
where $\mathbf{x} \otimes \mathbf{y} = (x_1 y_1, x_2 y_2, \cdots, x_n y_n)^T$

Approximates Pearl's algorithm[Pearl 1988] with some appropriate assumptions.

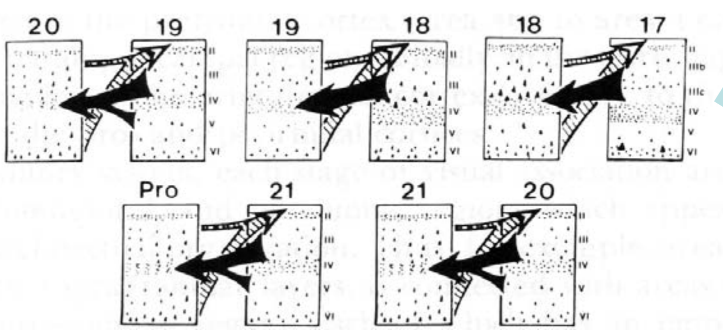


- Easy to be implemented by neurons.
- Linear time complexity in sparse network.

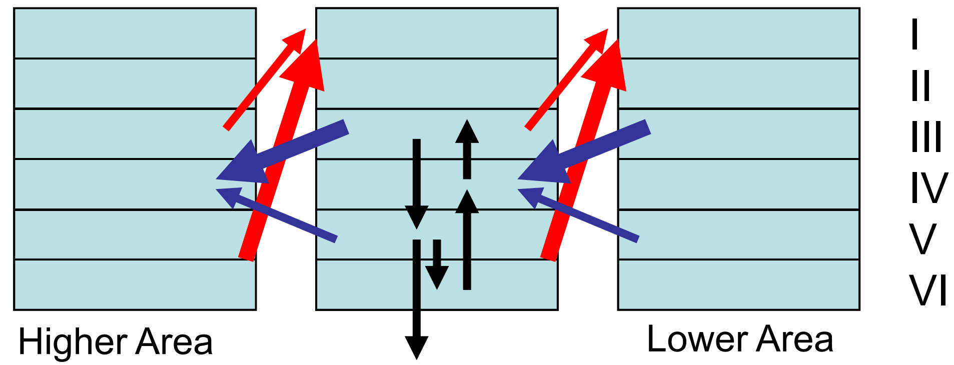
Connections between cortical layers



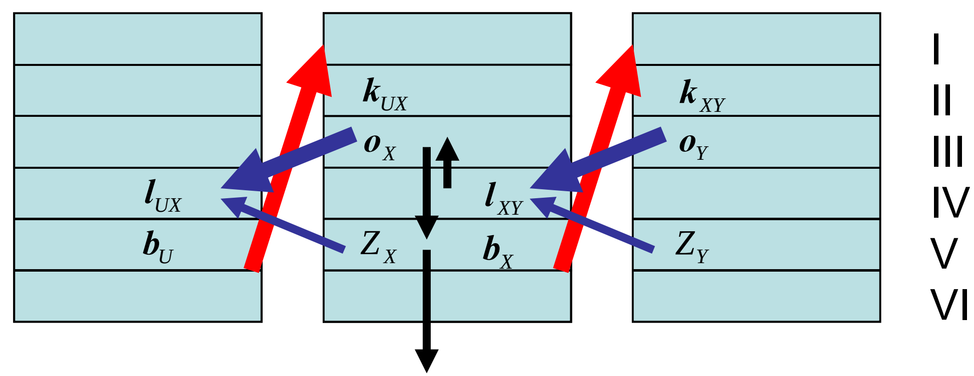
[Gilbert 1983]



[Pandya and Yeterian 1985]



Anatomical structure

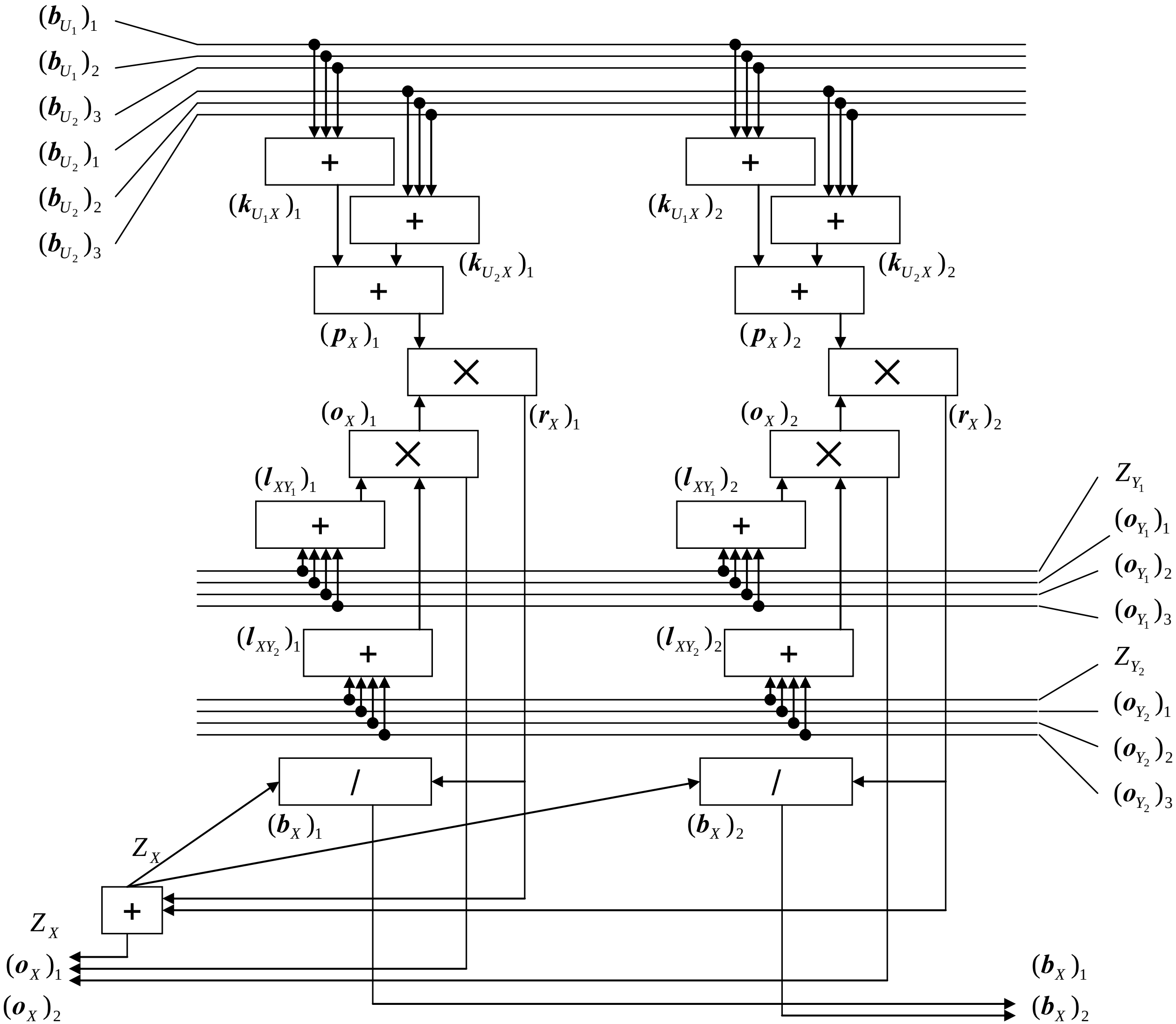


This model

Gilbert, C.D., Microcircuitry of the visual-cortex, Annual review of neuroscience, 6: 217-247, 1983.

Pandya, D.N. and Yeterian, E.H., Architecture and connections of cortical association areas. In: Peters A, Jones EG, eds. Cerebral Cortex (Vol. 4): Association and Auditory Cortices. New York: Plenum Press, 3-61, 1985.

Detailed structure in columns



I

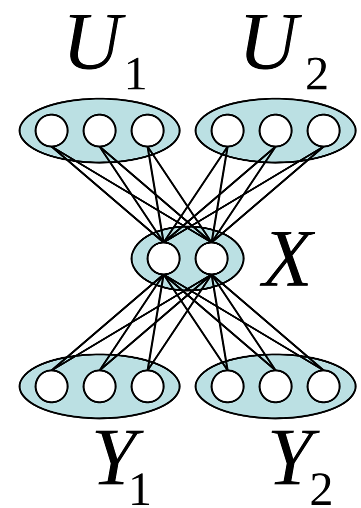
II

III

IV

V

VI

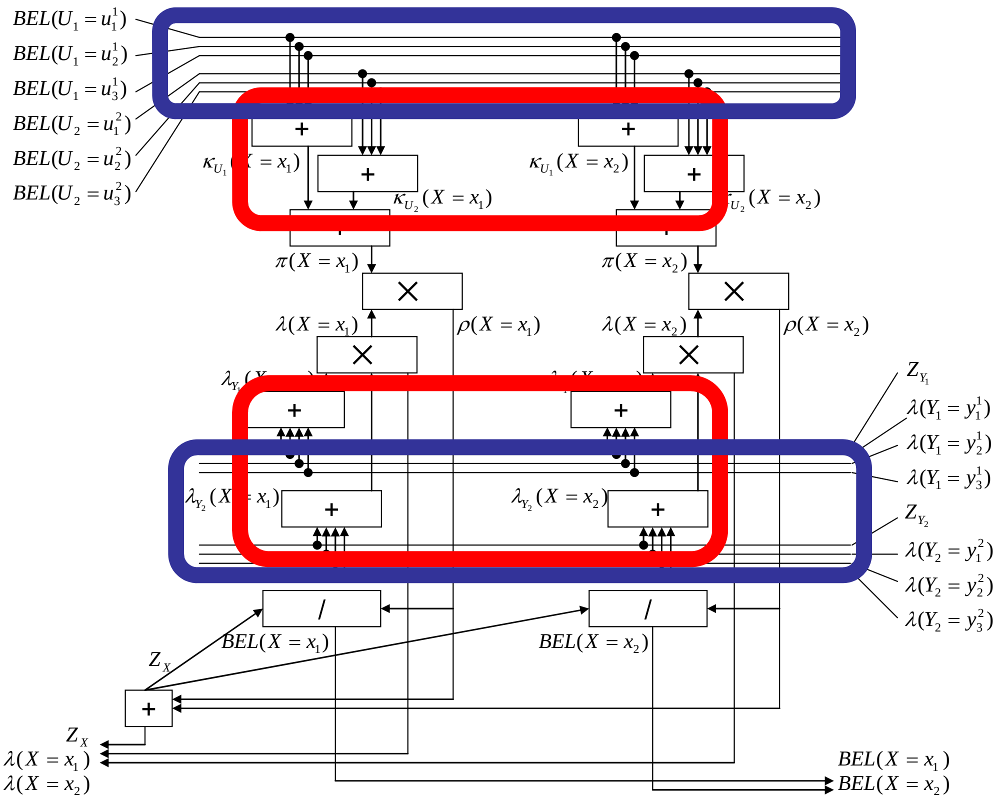


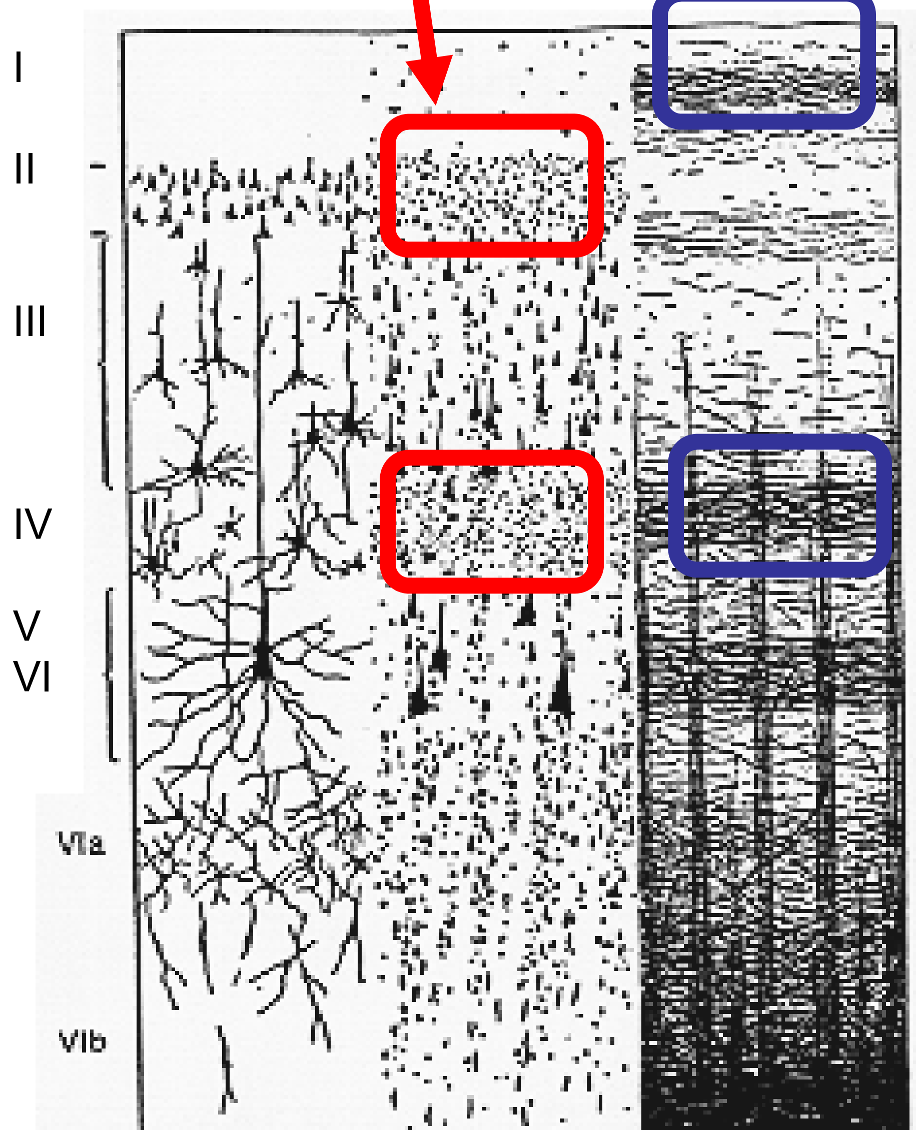
The left circuit calculates values of two units, x1 and x2, in node X in the above network.

Column structure of cortex

Horizontal fibers in layer 1, 4

Many cells in layer 2, 4

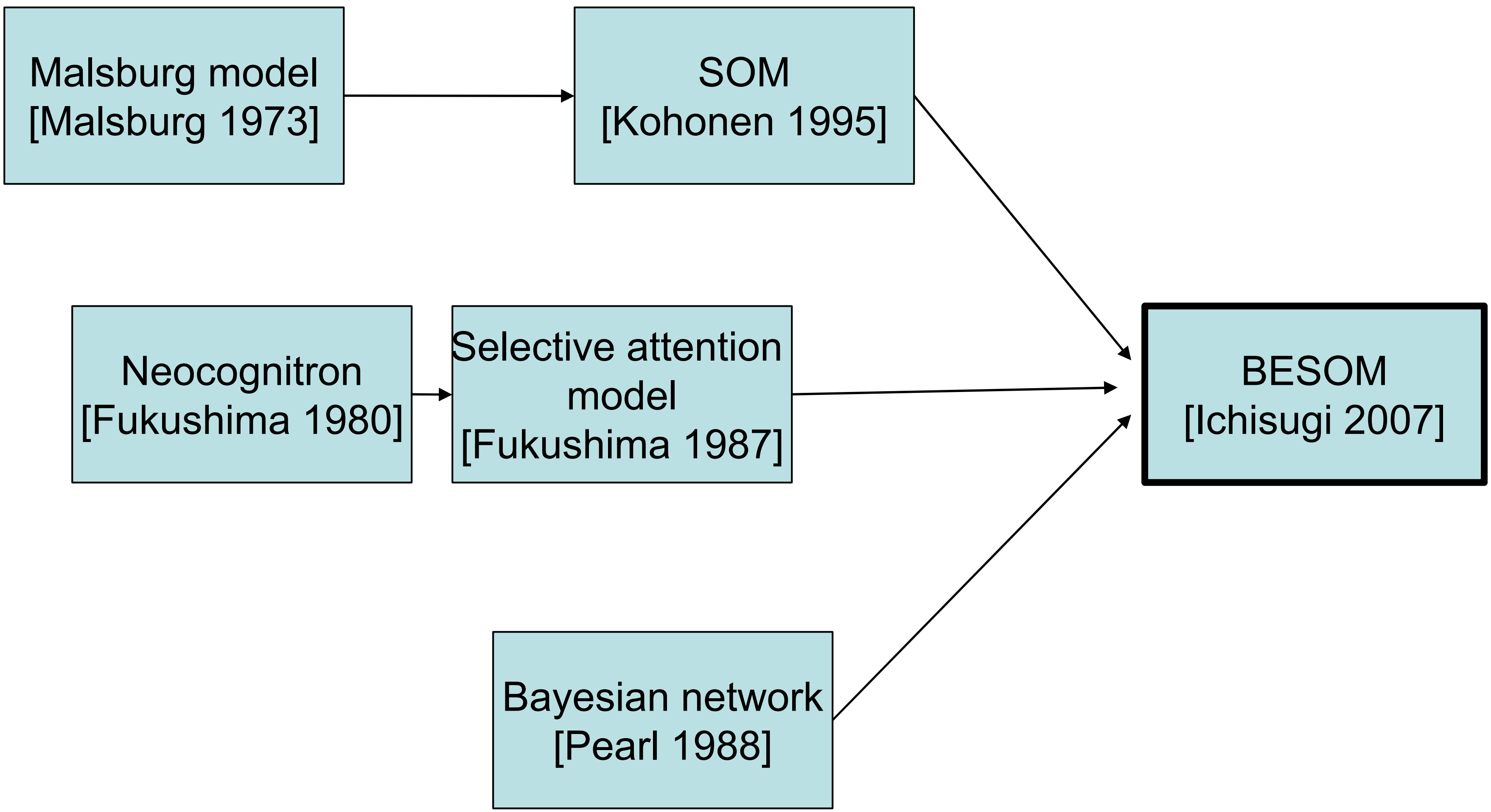




K. Brodmann, Vergleichende Lokalisation der Grosshirnrinde. in: ihren Prinzipien dargestellt auf Grund des Zellenbaues,. J.A. Barth, Leipzig, 1909.

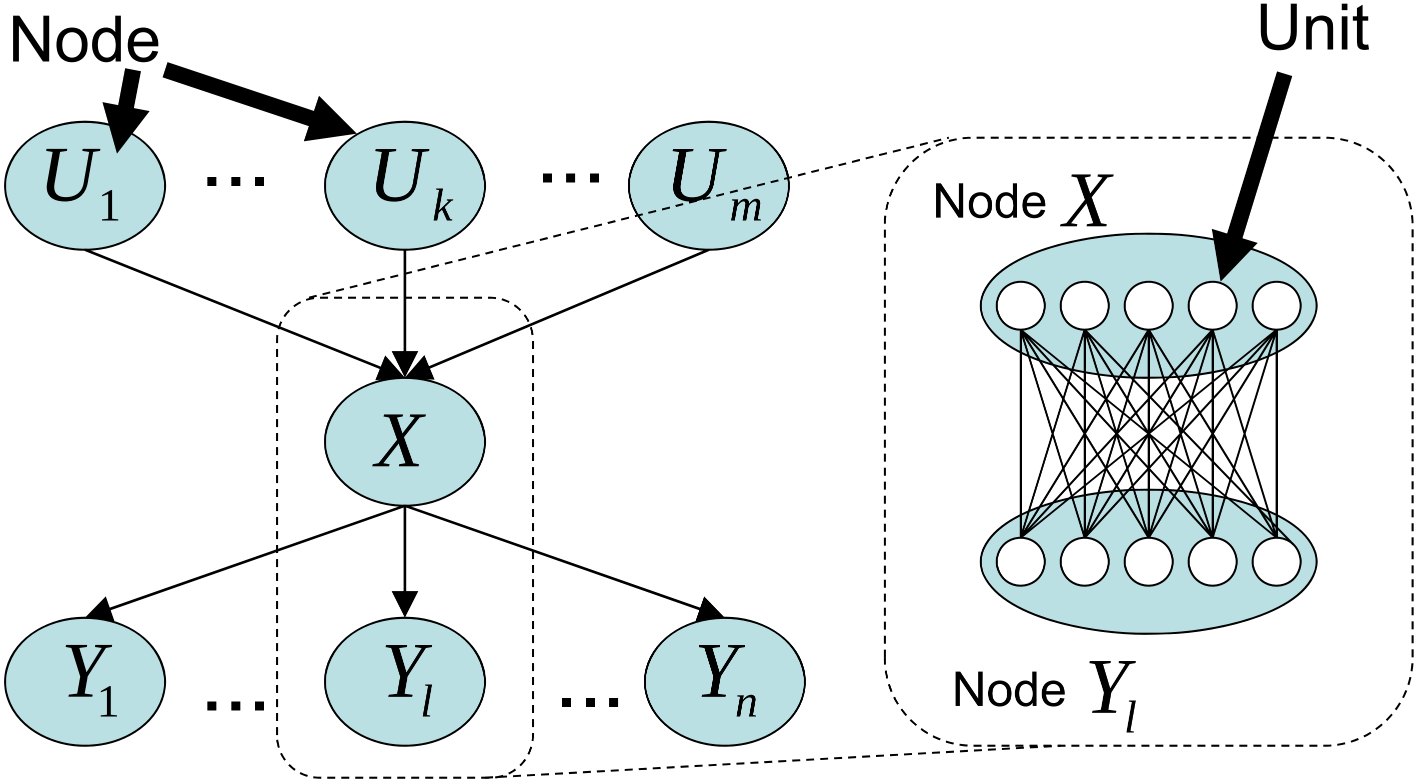
This figure is taken from the following Web page.
http://web.sc.itc.keio.ac.jp/anatomy/brodal/chapter12.html

BESOM model [Ichisugi 2007] unifies some previous models and Bayesian network

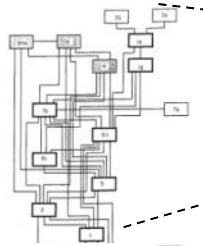


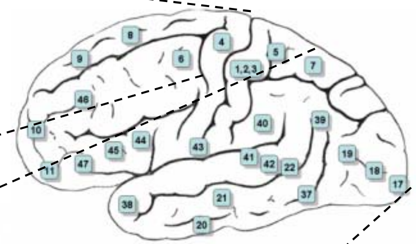
BESOM (BidirEctional SOM)

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

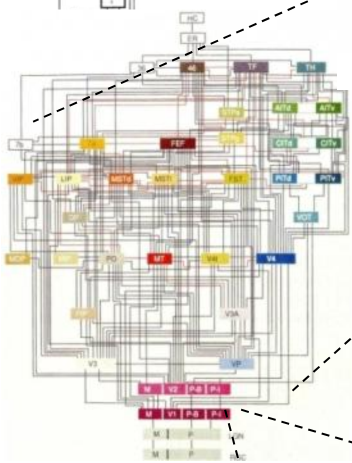


Brain and BESOM model [Ichisugi 2007]





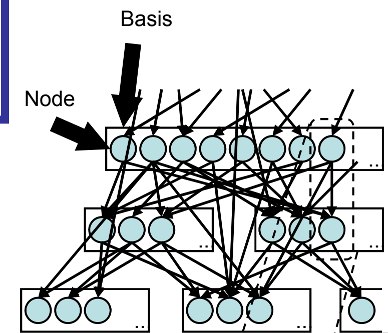
Frmo Wikipedia

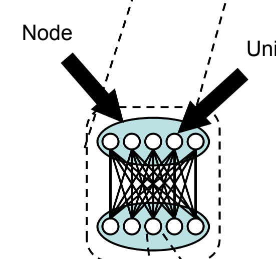


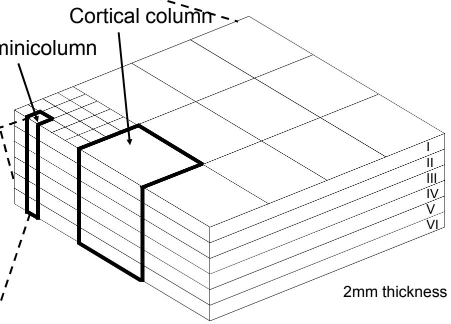
Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the
Primate Cerebral Cortex
Cerebral Cortex 1991 1: 1-47

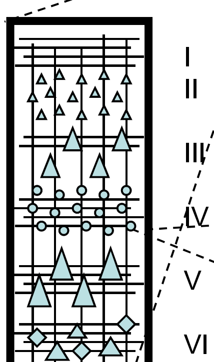
The cerebral cortex is a Bayesian network of 10^6 nodes with 10^2 states.

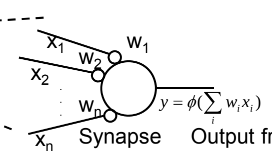
Brain	BESOM
Cerebral cortex	BESOM network
Area hierarchy	Basis hierarchy
Area	Basis
Cortical column	Node
Minicolumn	Unit
Neuron	Variable
Synapse	Weight of connection

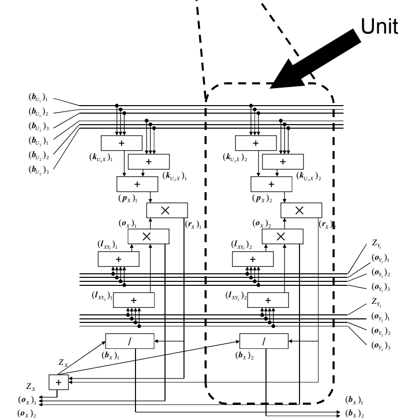




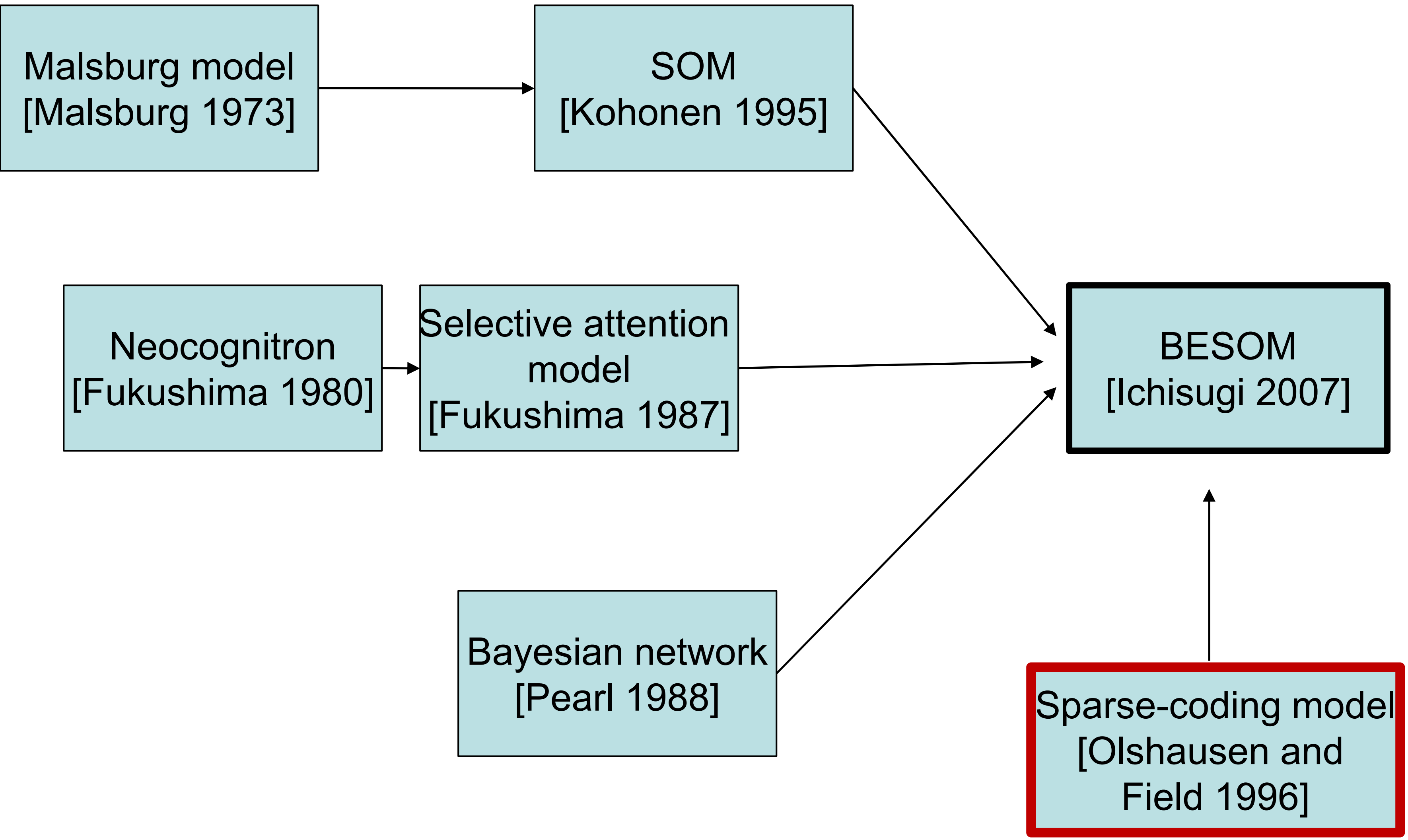








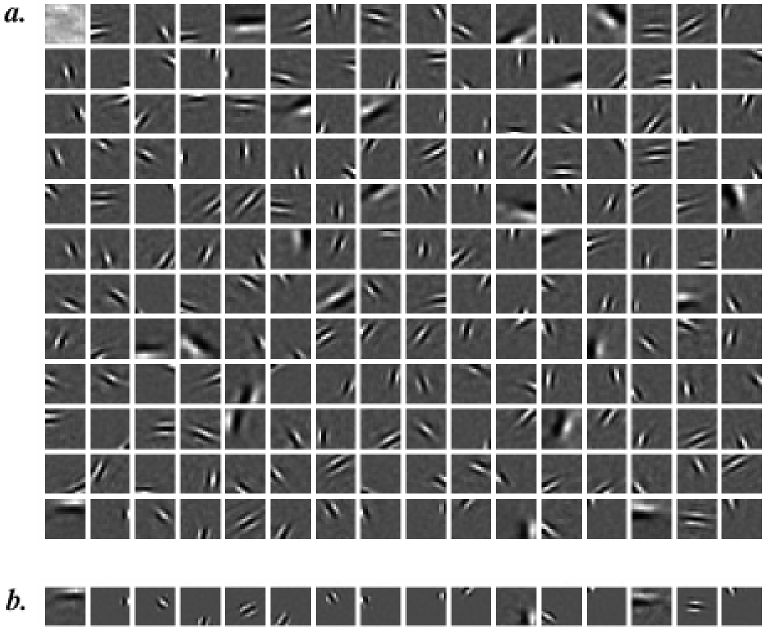
BESOM model and Sparse-coding model



Sparse-coding[Olshausen and Field 1996]

- Sparse-coding of natural images reproduces **orientation selectivity** of V1 simple-cells.

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$$



$$E = \| \mathbf{x}(t) - \mathbf{W}^T \mathbf{y}(t) \|^2 + \sum_{i=1}^{m} | y_i(t) |$$

"Emergence of simple-cell receptive field properties by learning a sparse code for natural images".
Bruno A. Olshausen and David J. Field
Nature, 381:607-609 (1996)

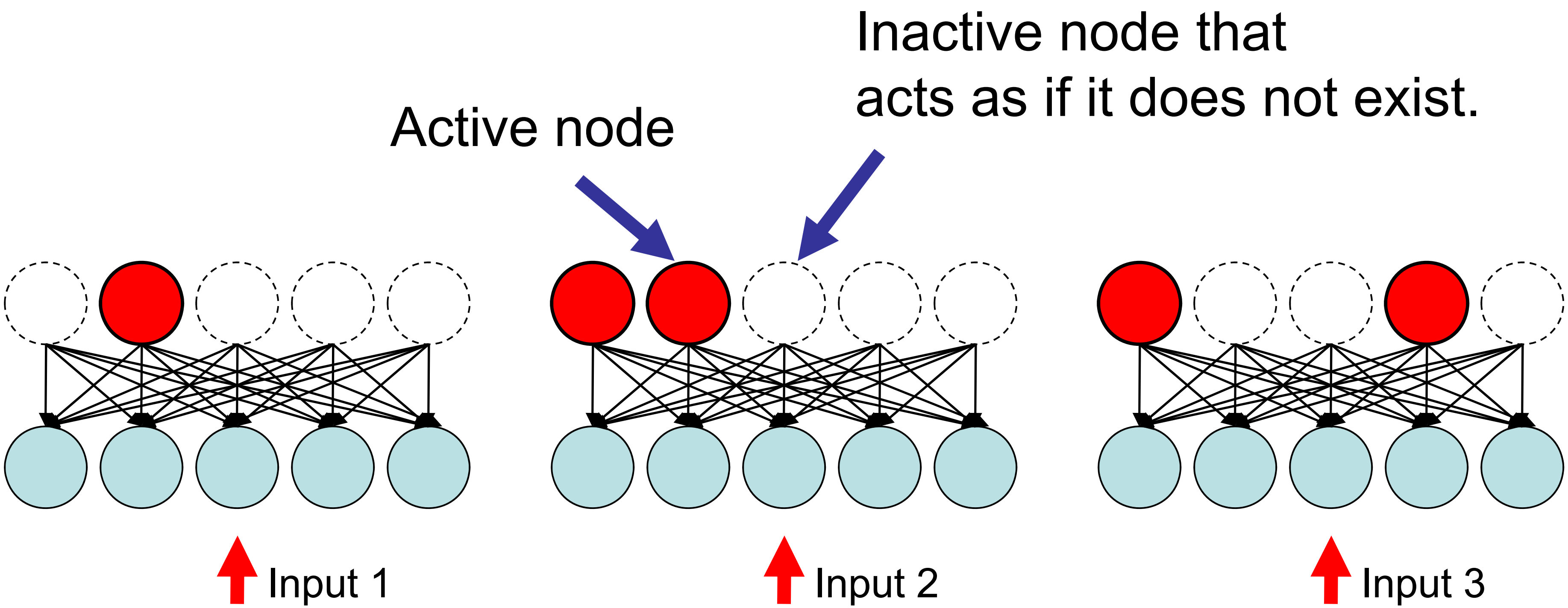
This figure is copied from:
http://watanabe-www.pi.titech.ac.jp/~shiga/ppt/ICA.ppt

What is sparse-coding ?

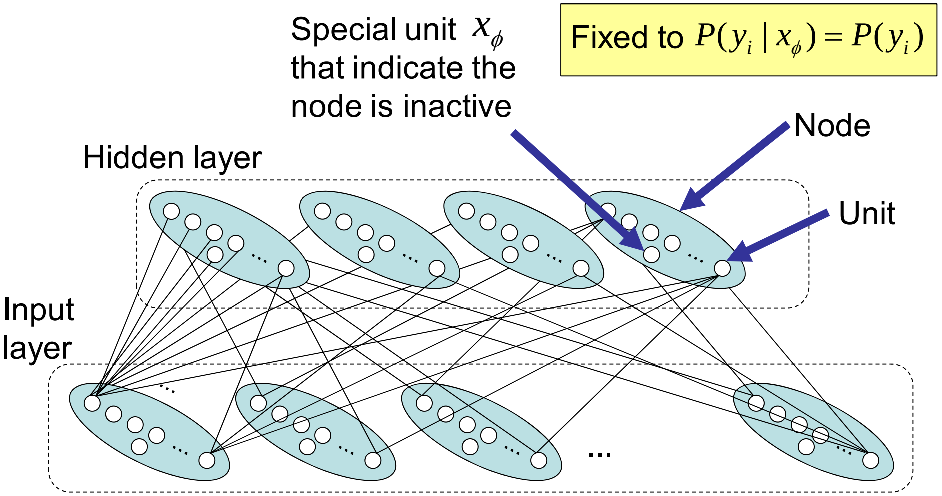
- A kind of **unsupervised learning** whose goal is to express inputs using small number of basis vectors.
- **Computational merits**:
  – Data compression.
  – Avoids "curse of dimension."
  – Blind source separation.
- **Biological merits**:
  – Saves energy, synapse maintenance cost.

Idea of sparse-coding by BESOM

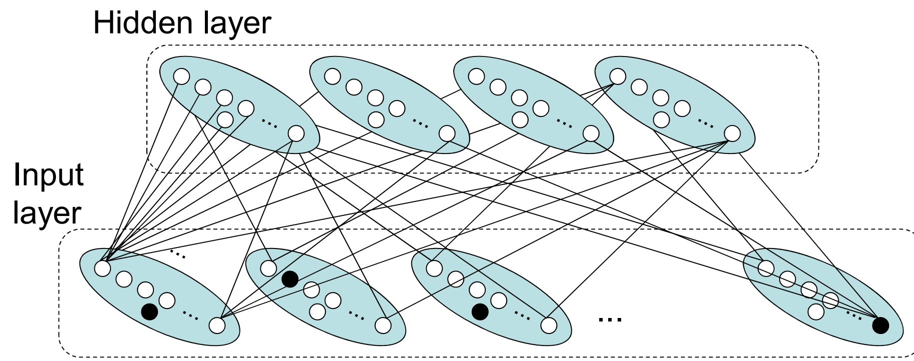
- Nodes may become "**inactive**" state.
- Only small number of nodes become active.



BESOM network for sparse coding

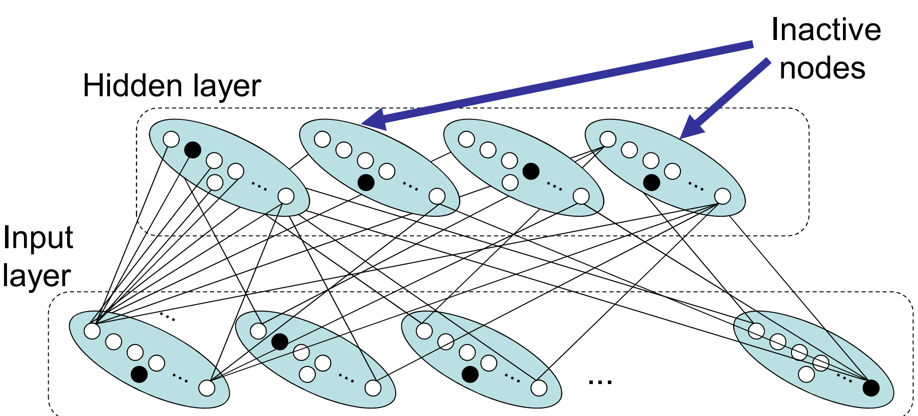


Input



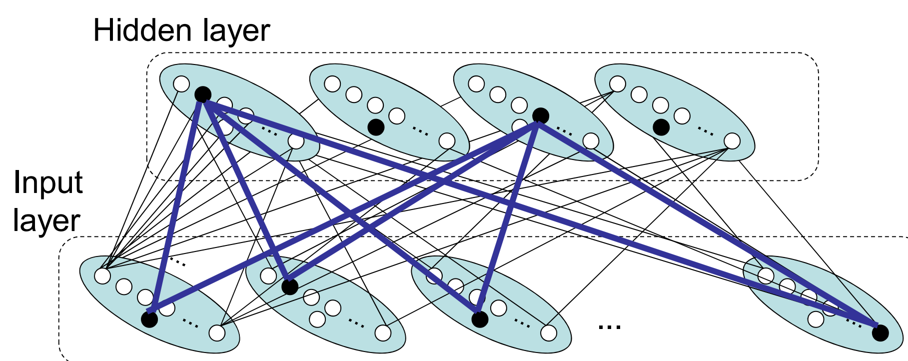
Input (observed data) is given at the lowest layer.

Recognition



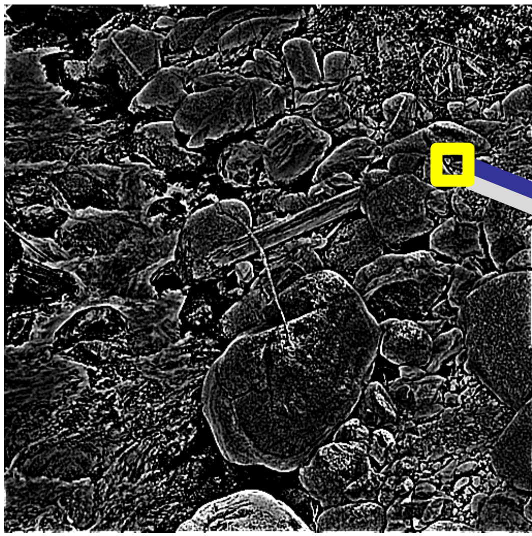
Calculate MPE with "**inactive bias**."

Learning

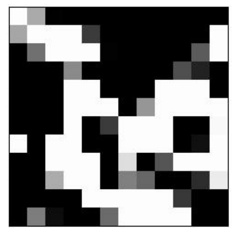


Increase the connection weights for **active** units.

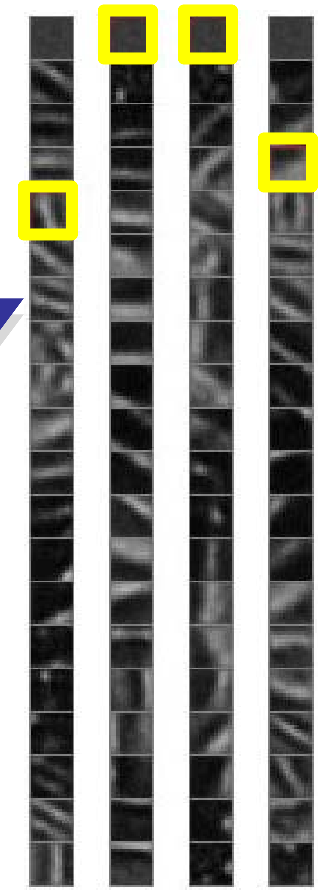
Result of learning natural images



Preprocessed natural image.



Randomly select
a 12x12 pixel image.



Winner units
learns the
input image
(with
neighborhood
learning)

Approximate the input
with linear sum of about
zero to two basis images

We used the image database provided by Olshausen
http://redwood.berkeley.edu/bruno/sparsenet/

Summary of BESOM sparse-coding

- A special value is introduced to each node that means the node is "inactive."
- Two cerebral cortex models, Bayesian network model and Sparse-coding model, can be unified to a single model.
  - The learning algorithm does not break the theoretical framework of Bayesian networks.
- Learned basis images show orientation selectivity, as in the primary visual area.

Take home message

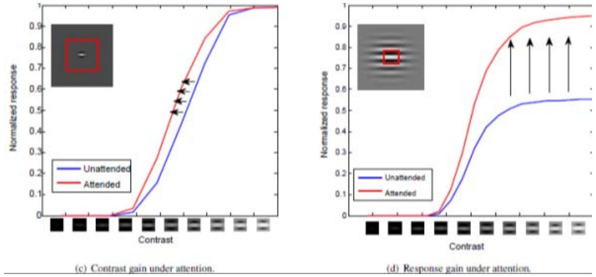
- **The cerebral cortex is a Bayesian network.**
    - However, most neuroscientist do not know what Bayesian networks are.
- Many cerebral models are being integrated into one universal model based on Bayesian networks.
    - – Such a model will become the core technology for reproducing human-like high intelligence.
- **Computational neuroscience needs Bayesian network experts!**

Additional slides

Reproducing contrast responses

[Reynolds and Heeger 2009] model clearly reproduces very complicated electrophysiological phenomena, however, it contains a mysterious constant σ .

$$R(x,\theta) = \frac{A(x,\theta)E(x,\theta)}{S(x,\theta) + \underline{\sigma}}$$



The normalization model of attention
(Reynolds JH, Heeger DJ, Neuron. 2009 Jan 29;61(2):168-85)

On the other hand, [Chikkerur et al. 2010] model reproduces some of these phenomena very naturally.

$$P(F_l^i \mid I) = \frac{P(I \mid F_l^i)P(F_l^i)}{\sum_{F_l^i} P(I \mid F_l^i)P(F_l^i)}$$

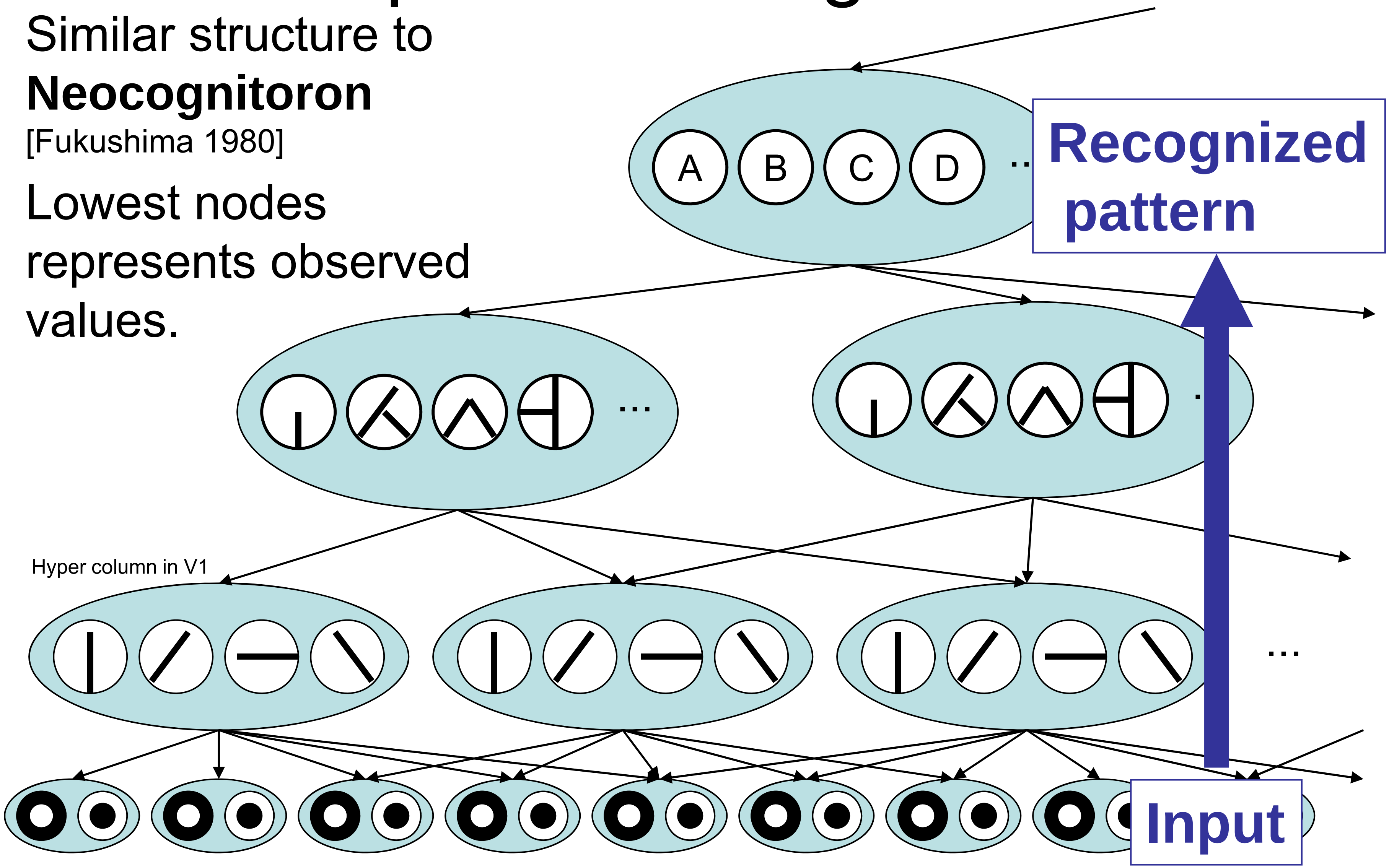
Just applies Bayes' rule.

$$= \frac{P(I \mid F_l^i)P(F_l^i)}{\sum_{i=1}^{L} P(I \mid F_l^i = i)P(F_l^i = i) + \underline{P(I \mid F_l^i = 0)P(F_l^i = 0)}}$$

This term may cause the same effect as σ .

BESOM may be used for pattern recognition

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



Simple formalization of BESOM model

- Objective of learning:
  - Calculate MAP estimator of the parameter θ assuming each input $\mathbf{i}(t)$ at time t is generated from i.i.d.

$$\theta^* = \arg\max_{\theta} \left[\prod_{i=1}^{t} P(\mathbf{i}(i) \mid \theta) \right] P(\theta)$$

$$= \arg\max_{\theta} \left[\prod_{i=1}^{t} \sum_{\mathbf{h}} P(\mathbf{h}, \mathbf{i}(i) \mid \theta) \right] P(\theta)$$

Recognition and Learning steps

Recognition step: $\widehat{\mathbf{h}}(t) = \arg\max_{\mathbf{h}} P(\mathbf{h}, \mathbf{i}(t) \mid \theta(t))$

Learning step: $\theta(t+1) = \arg\max_{\theta} \left[\prod_{i=1}^{t} P(\widehat{\mathbf{h}}(i), \mathbf{i}(i) \mid \theta) \right] P(\theta)$

$P(\mathbf{h}, \mathbf{i} \mid \theta)$: Probabilistic model. (Bayesian network.)

$P(\theta)$: Innate knowledge about the parameter, such as **sparseness**.

$\widehat{\mathbf{h}}(t)$: States of cortical columns. 10^6 dim.

$\mathbf{i}(t)$: Input vector. 10^4 dim.?

$\theta(t)$: All weights of variable synapses. 10^16 dim.

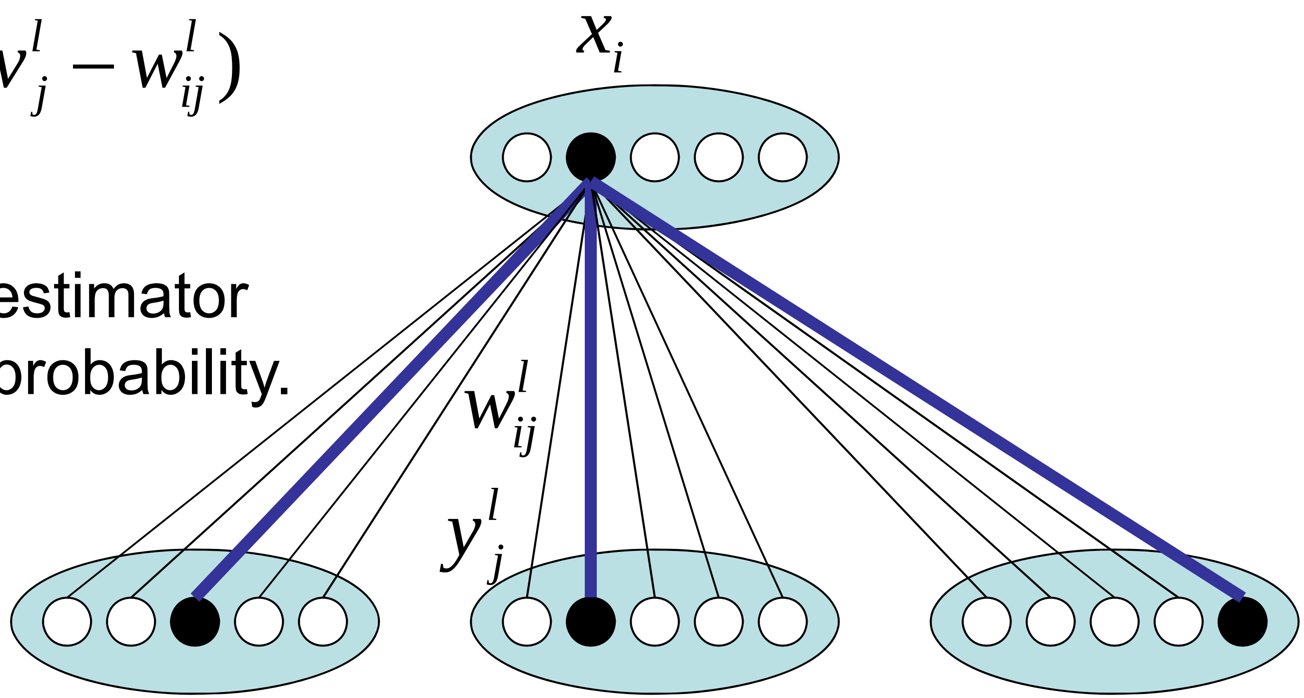
Learning conditional probabilities with Hebb's rule[Ichisugi 2007]

Learning rule for unit x_i :

$$w_{ij}^{l} \leftarrow w_{ij}^{l} + \alpha_i (v_j^{l} - w_{ij}^{l})$$

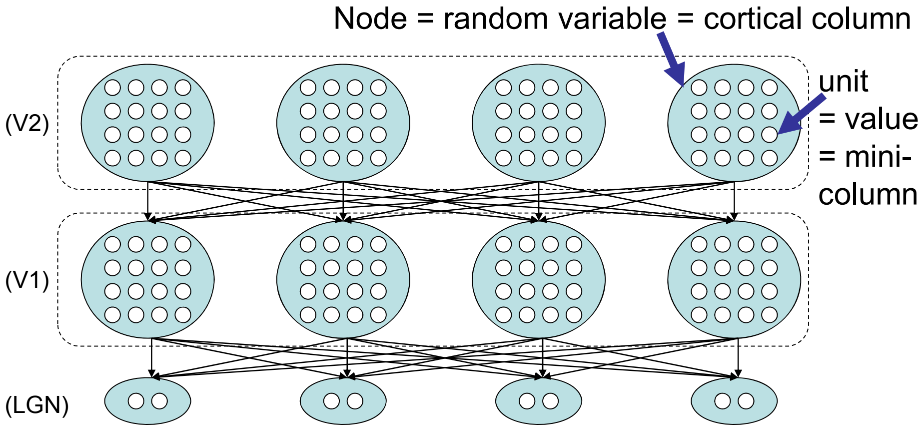
The weight is ML estimator of the conditional probability.

$$w_{ij}^{l} = P(y_j^{l} \mid x_i)$$



v^1=(0, 0, 1, 0, 0) v^2=(0, 1, 0, 0, 0) v^3=(0, 0, 0, 0, 1)

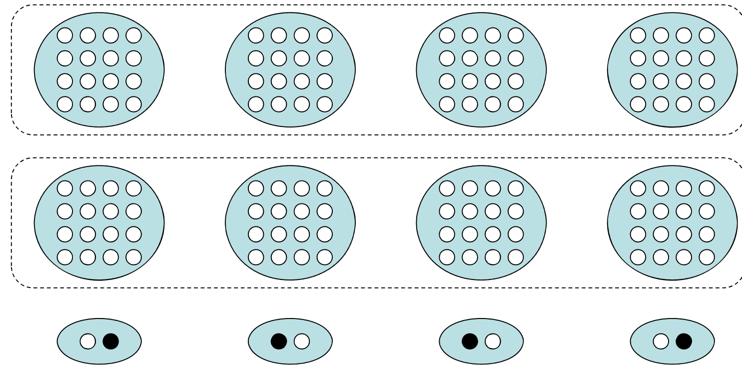
Structure of BESOM network



No connections in each layer.
Fully connected between different layers.

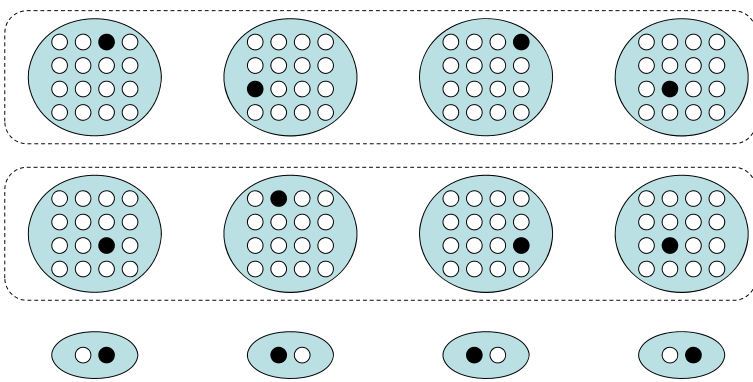
Connection weights
= CPT
= synapase weights

Input



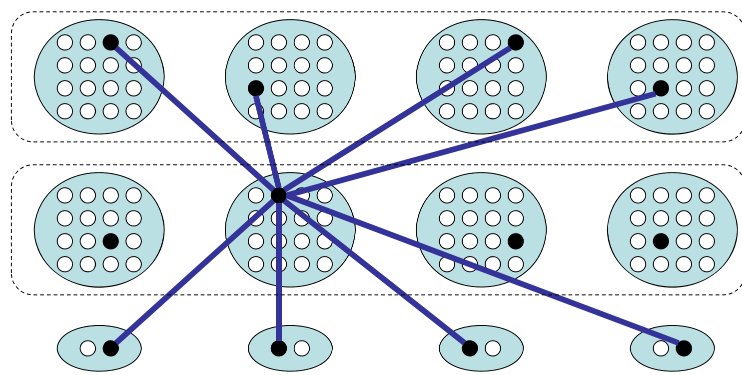
Input (observed data) is given at the lowest layer.

Recognition



Find the values of hidden variables
with the highest posterior probability.
(MPE: most probable explanation)

Learning

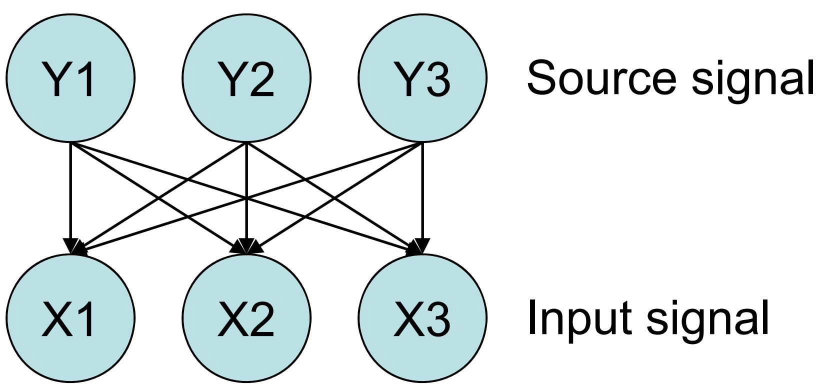


Increase the connection weights between active units
(mini-columns) and decrease the other weights.

How the network structure of Bayesian net is learned ?
- My speculation -

Generative model can be acquired by ICA

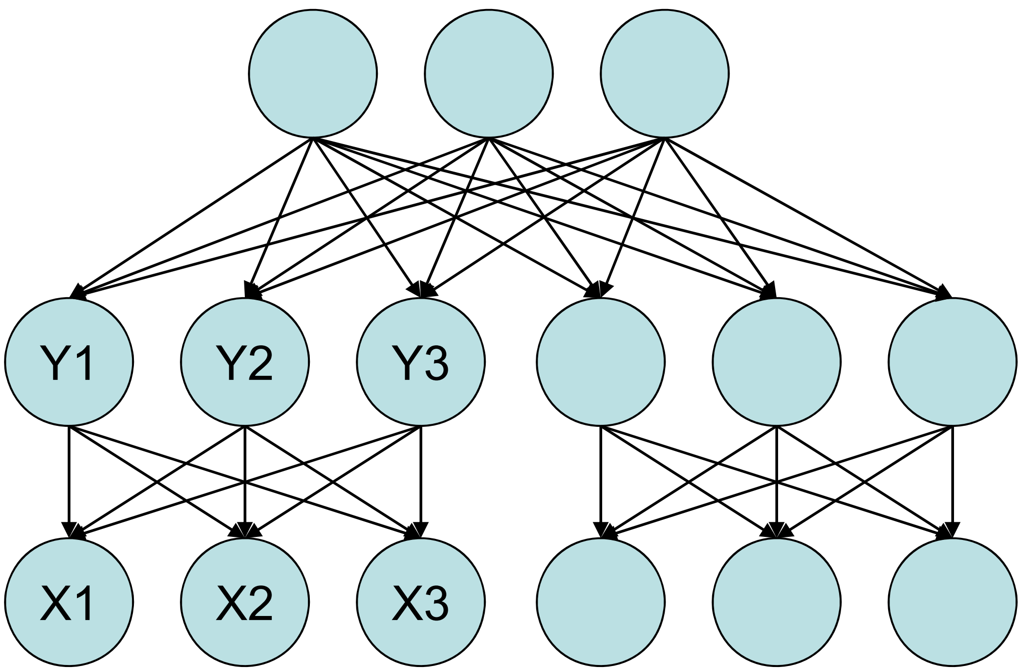
- In other words, ICA may acquire two-layered Bayesian network structure.



$x=f(y)$

Hierarchical generative model

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

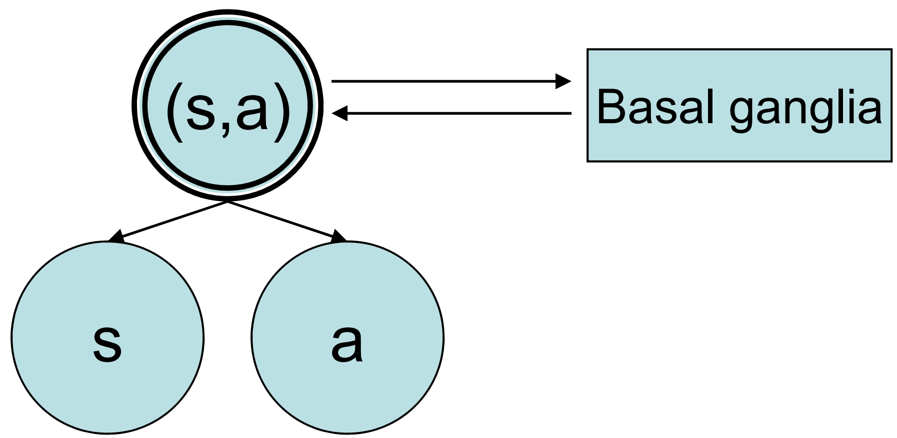


Other retinal areas or modalities.

How about motor areas ?

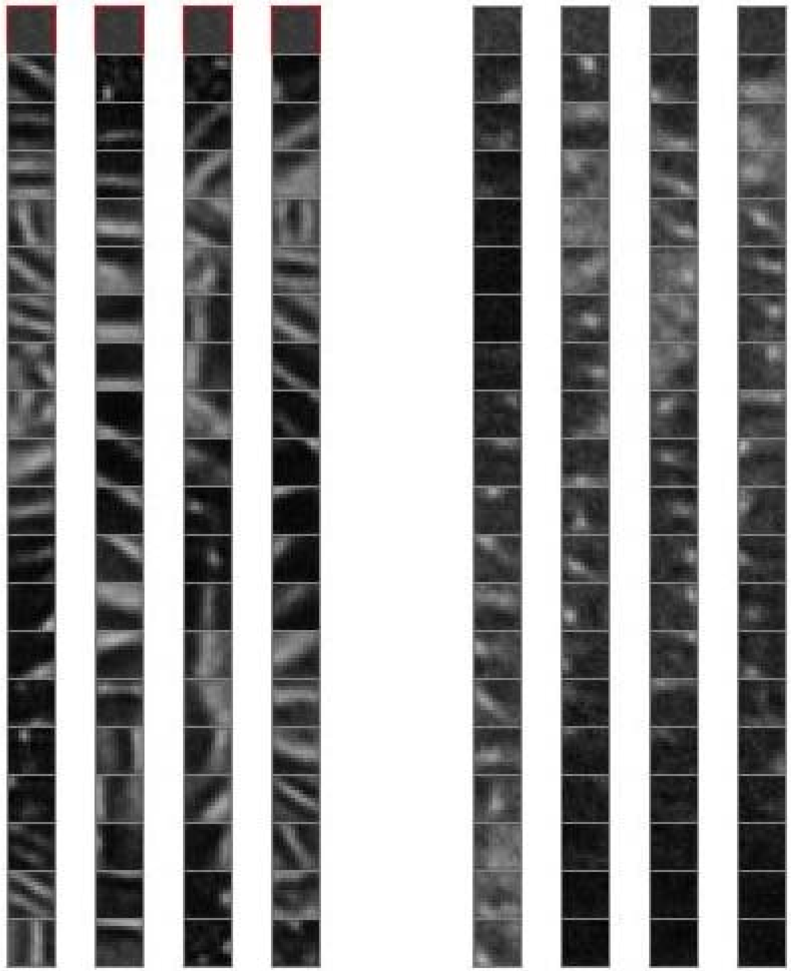
Reinforcement learning in motor areas

- 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.

Effect of Sparseness



If no sparseness, orientation selectivity of basis images become weak because every base image becomes close to the mean image of input images.

Number of used units to approximate input image.

0 - 2 units

4 units
(No sparseness)

Learning natural images

- Input:
  - We extracted image patch with 7x7=49 pixels from a random position.
  - Then, we gave the pixel intensities in the image patch to the 49 binary input nodes $(Y^l \in \{0,1\}, l = 1 \cdots 49)$.
    - For example, for intensity 0.2, the value was set to 1 with probability 0.2.
- Visualization of CPT:
  - 7x7 CPT elements of $P(Y^l = 1 | X = x_i)$ are visualized as the brightness of 7x7 pixels.

Brain is now understandable

- because of remarkable progress of computer science and neuroscience in recent 20 years.
  - Maturity of AI and machine learning technology
    - Bayesian network [Pearl 1988]
    - Reinforcement learning [Sutton 1998]
    - Independent Component Analysis [Hyvarinen 2001]
  - Important findings of neuroscience
    - **Sparse-coding at primary visual area [Olshausen 1996]**
    - Reinforcement learning at basal ganglia [Schultz 1997]
    - **Bayesian network models of cerebral cortex [Rao 2005] etc.**