Machine Learning for Cognitive Sciences:
Principles and Applications

Simona Cocco (Physics Departement of ENS)
simona.cocco@phys.ens.fr CM &TD

Vito Dichio (Physics Departement of ENS)
vito.dichio@phys.ens.fr TD

Week 3

Recall of previous lecture

- Connection between **Entropy** in physics and Information and communication theory:
Shannon Information

- **In Physics**: Entropy S measures the logarithm of the number of relevant configurations giving the same macroscopic state (eg. number of off different configuration for the particles of liquid at a given temperature/pressure).

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

- **In communication theory S** measures the uncertainty (or missing information or average gain in information) in a communication removed after the answer are given. It is expressed in bit.

- Property of S: When $p_i=1/N$ $S=\log N$

- Connession with **Code Compression**: optimal coding (shortest code) for a language with N words: is obtained when each word is coded by a number of bit proportional to $-\log p_i$ where p_i is its frequency. The average length of the words in this language.

Plan of the lecture

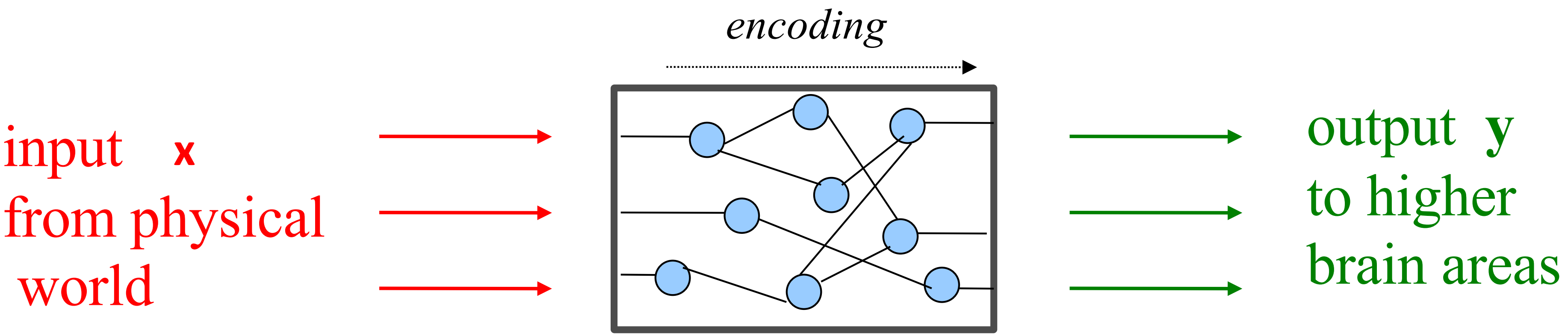
- **Mutual Information** & its relashionship with the **Entropy**

- Some applications to **Efficient Coding Theory** in neuroscience.

- **Ccross Entropy & Kullback-Leibler Divergence** and theis relashionships with Entropy, Mutual Information, and the theory of **Asymptotic Inference**: Theoretical measures on how the error in the parameters estimation decreases with the number of data.

Mutual Information & Efficient coding

Organization principle for sensory areas



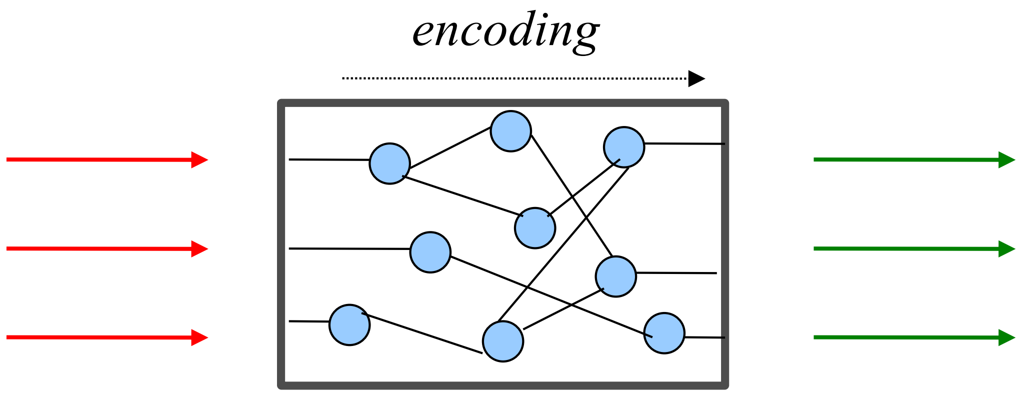
Hypothesis: encoding should maximize mutual information

$$\text{MI (input x, output y)} = \int dx \int dy\, P\,(x, y)\, log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right)$$

Mutual Information & Efficient coding

Organization principle for sensory areas

input **x**
from physical
world



output **y**
to higher
brain areas

Hypothesis: encoding should maximize mutual information

MI (input x, output y) $= \int dx \int dy\, P\,(x,y)\, log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right)$

For variables taking discrete values:

$$MI(1,2) = \sum_{x_1 x_2} p(x_1,x_2)\ \log_2\left[\frac{p(x_1,x_2)}{p_1(x_1)\,p_2(x_2)}\right]$$

Mutual information

The dependence between events (or event distributions) is characterized through the **mutual information**

$$MI(1,2) = \sum_{x_1 x_2} p(x_1, x_2) \ \log_2 \left[\frac{p(x_1, x_2)}{p_1(x_1) p_2(x_2)} \right] \qquad \textbf{in bits}$$

This quantity measures how much information one has on one variable from knowledge of the other one.

(1) Always positive (or null) (it can be shown from the convexity of the logarithm log (x)<x, see Supp. slides)

(2) It is null for independent variables

(3) Degraded through processing: $x_2 \rightarrow x_3$ then $MI(1,3) \leq MI(1,2)$

<u>Exercise :</u> *compute MI(box,cookie type)*

Mutual Information & Entropy

- The mutual information represents the average gain in information over x_1 when x_2 is known, or alternatively, over x_2 when x_1 is known

$$MI(x_1,x_2) = S[p(x_1)] - S[p(x_1|x_2)]$$

Mutual Information & Entropy

- The mutual information represents the average gain in information over x_1 when x_2 is known, or alternatively, over x_2 when x_1 is known

$$MI(x_1,x_2) = S[p(x_1)] - S[p(x_1|x_2)]$$

Conditional entropy of x_1 at given x_2

$$S[p(x_1|x_2)] = -\sum_{x_2} p(x_2) \sum_{x_1} p(x_1|x_2) \log_2 p(x_1|x_2)$$

Mutual Information & Entropy

- The mutual information represents the average gain in information over x_1 when x_2 is known, or alternatively, over x_2 when x_1 is known

$$MI(x_1,x_2) = S[p(x_1)] - S[p(x_1|x_2)] = S[p(x_2)] - S[p(x_2|x_1)]$$

Conditional entropy of x_2 at given x_1

$$S[p(x_2|x_1)] = -\sum_{x_1} p(x_1) \sum_{x_2} p(x_2|x_1) \log_2 p(x_2|x_1)$$

Mutual Information & Entropy

- The mutual information represents the average gain in information over x_1 when x_2 is known, or alternatively, over x_2 when x_1 is known

$MI(y,\theta)=$ $\quad = \; S[p(\theta)] - S[p(\theta|y)]$

0.05 bit $\quad$ 1 bit

Conditional entropy of θ at given y

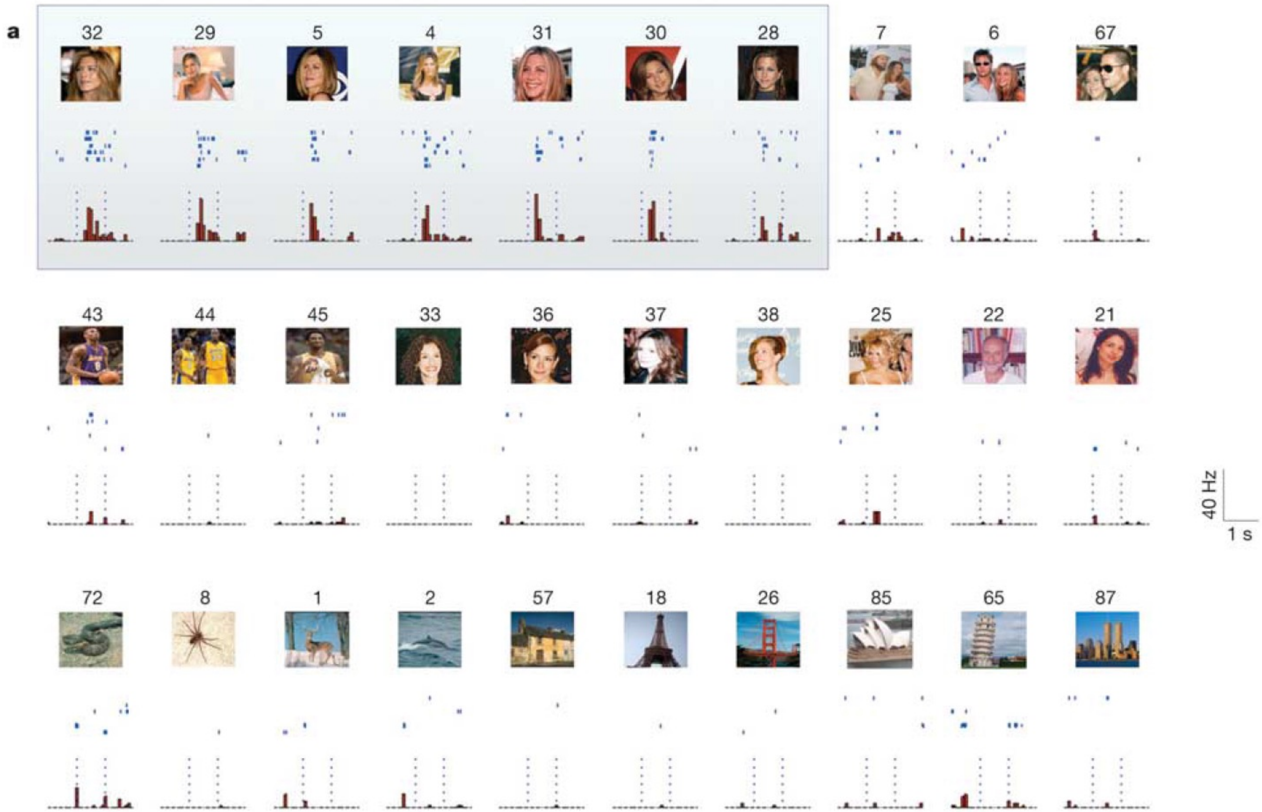
$$S[p(\theta|y)] = -\sum_y p(y) \sum_\theta p(\theta|y) \log_2 p(\theta|y)$$

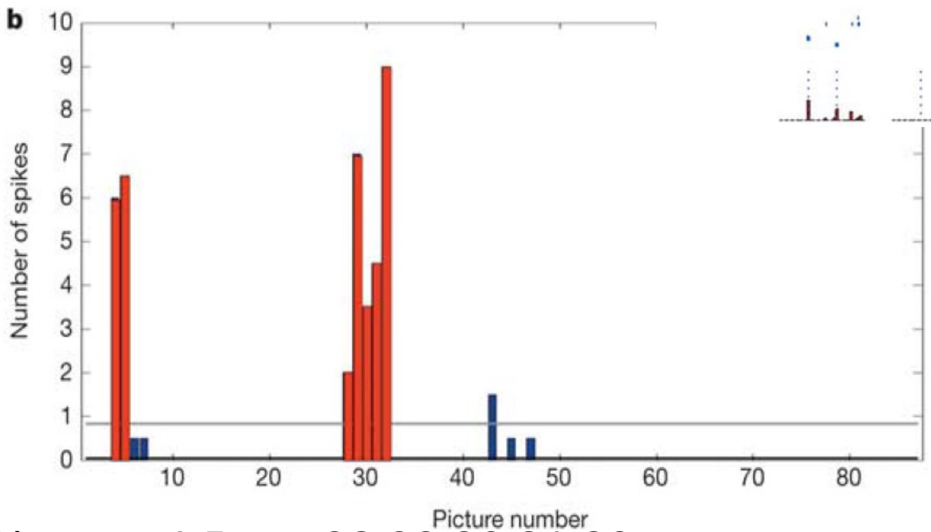
0.95 bit

- *exercises : compute the gain in information on the cookie type knowing the box & recompute them directly through the conditional entropy.*

Information conveyed by the spiking activity of a neuron about a stimulus

A single neuron in hippocampus
 activated exclusively by different
photos of the actress Jennifer Aniston





Pictures 4,5   28,29,30,31,32

nature Vol 435|23 June 2005|doi:10.1038/nature03687

LETTERS

Invariant visual representation by single neurons in the human brain

R. Quian Quiroga[1,2]†, L. Reddy[1], G. Kreiman[3], C. Koch[1] & I. Fried[2,4]

(2005)

MI= S[p(activity)]-S[p(activity|picture=JA)]

Mutual Information in Electroencephalography

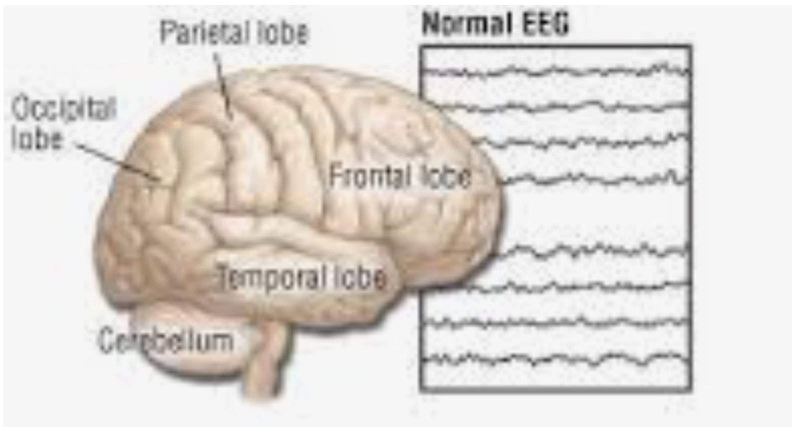
Mutual information analysis of the EEG in patients with Alzheimer's disease

Jaeseung Jeong[a,*], John C. Gore[b], Bradley S. Peterson[a]

Clinical Neurophysiology 112 (2001) 827–835

EEG obtained in 15 healthy controls and 15 AD patients age ~ 70 years

Estimate the average MI between EEG electrodes to assess the information transmission between different cortical areas in Alzheimer's disease (AD) patients .



Mutual Information in Electroencephalography

Normal subjects

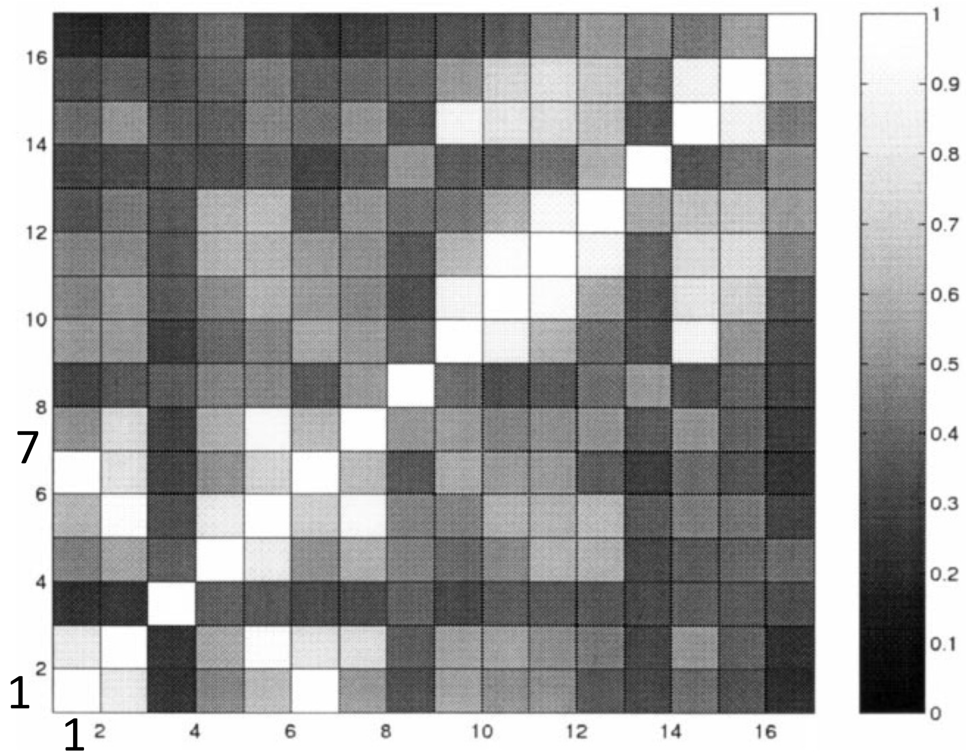


Fig. 2. Distribution of the average CMI values between all pairs of channels in the normal subjects. The numbers (1–16) correspond to F7, T3, T5, Fp1, F3, C3, P3, O1, Fp2, F4, C4, P4, O2, F8, T4, and T6.

AD patients

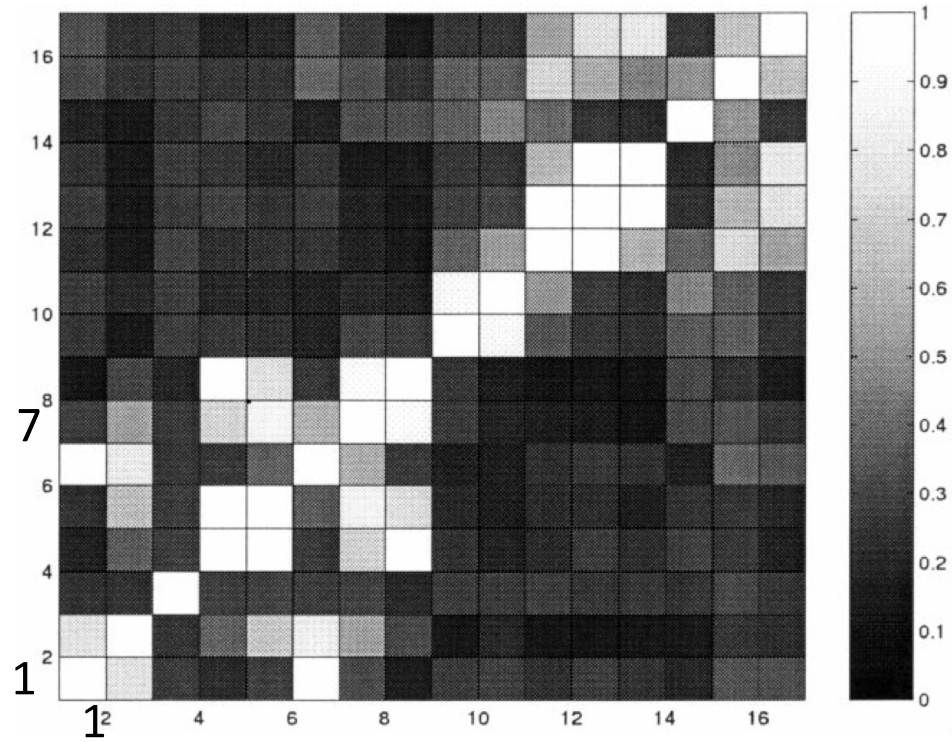
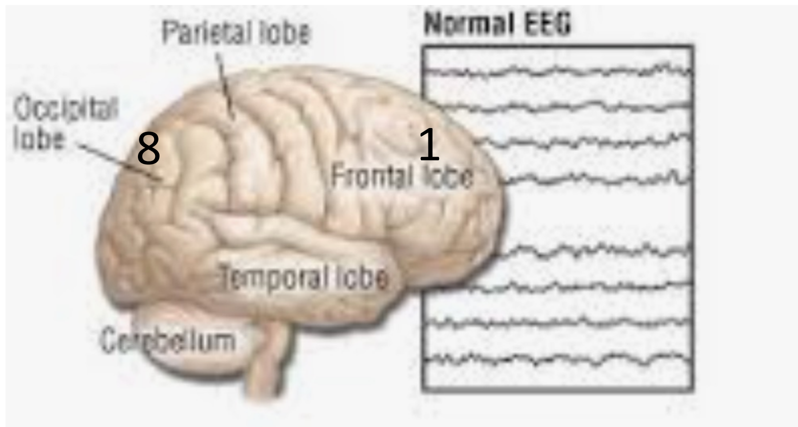


Fig. 3. Distribution of the average CMI values between all pairs of channels in AD patients.

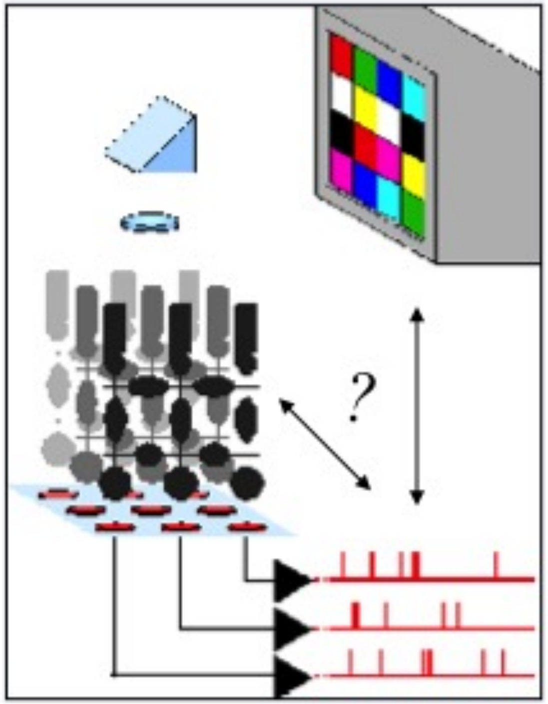
- Possible values of the EEG Discretized in 64 bins

- EEG abnormalities in AD patients : functional impairment of information



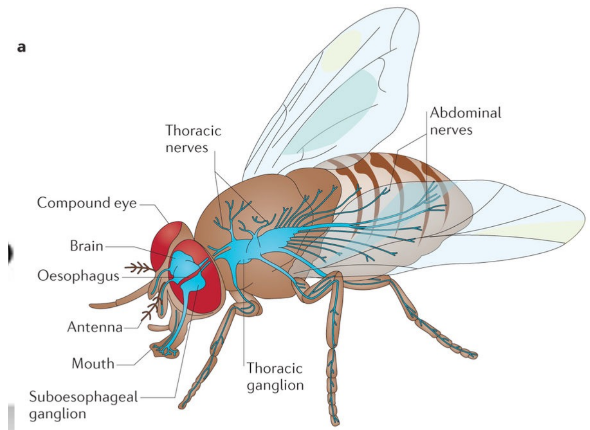
Tutorial: Mutual Information conveyed by the spiking activity of a neuron about a stimulus

In vitro recording on retina

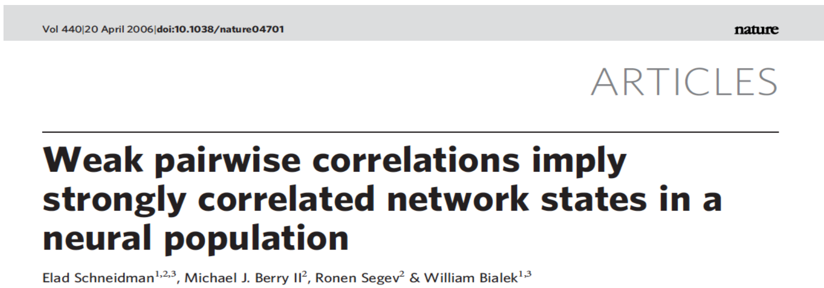


Retina Ganglion cells
K.Koch … M.A.Freed (2004)

Tomorrow tutorial:
40 ganglions, salamandra retina natural movies. Recordings last for about 1 hour



Motion sensitive neuron in the fly brain
de Ruiter van Stenevich … Bialek (2003)


Vol 440|20 April 2006|doi:10.1038/nature04701 nature

ARTICLES

Weak pairwise correlations imply strongly correlated network states in a neural population

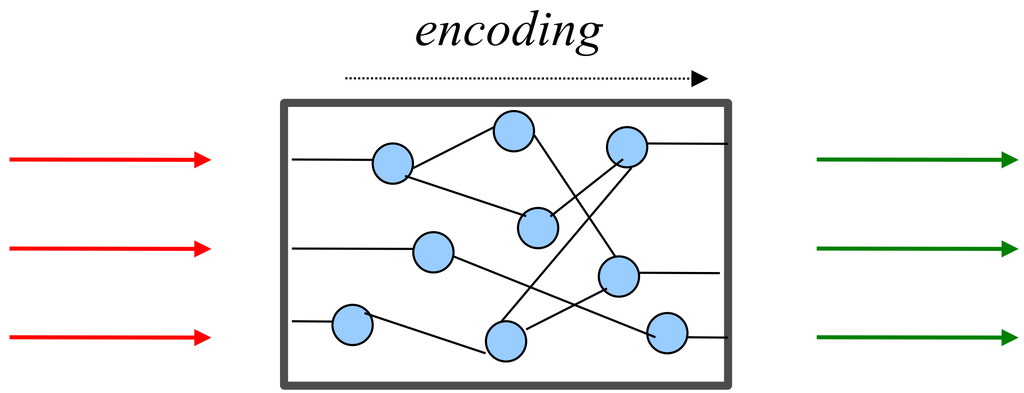
Elad Schneidman[1,2,3], Michael J. Berry II[2], Ronen Segev[2] & William Bialek[1,3]

Same movie repeated 120 times. MI(activity,t=stimulus)=S(activity)-S(activity|t=stimulus)

Efficient coding

Organization principle for sensory areas

input **x**
from physical
world



output **y**
to higher
brain areas

Hypothesis: encoding should maximize mutual information

$$\text{MI (input x, output y)} \quad = \int dx \int dy\, P\,(x, y)\; log_2 \;\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

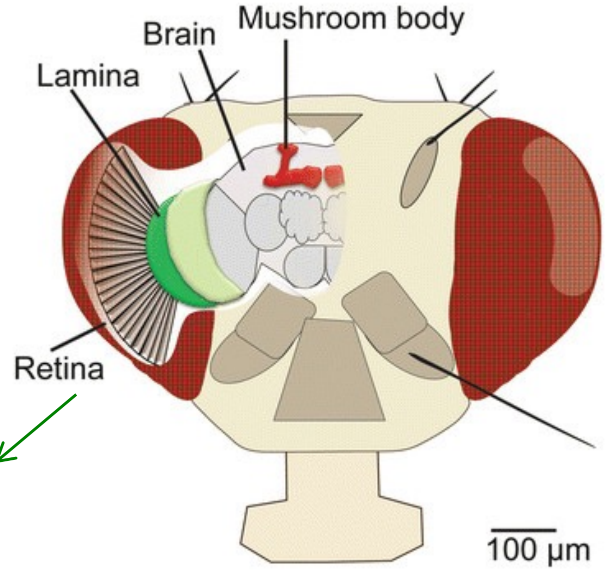
S.B. Laughlin, A Simple Coding Procedure Enhances a Neuron's Information Capacity, 1981

Efficient coding

S.B. Laughlin, A Simple Coding Procedure Enhances a Neuron's Information Capacity, 1981

Goal.
explain coding of **light intensity** by the fruit-fly large lamina monopolar cells

Intensity x
Distribution $P(x)$



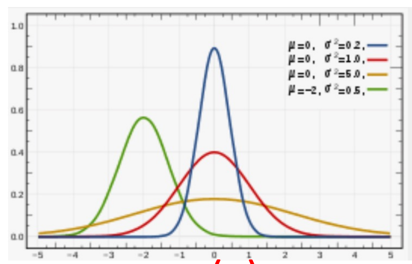
y

Ando et al., 2016

Hp: **Firing rate y** is a Gaussian variable with average value y(x) and variance (σ)

y=y(x)+ weak gaussian noise

$$P(y|x) = \frac{e^{-\frac{(y-y(x))^2}{2\ \sigma^2}}}{\sqrt{2\pi\sigma^2}}$$



y(x)

Optimal code: Maximize MI (x,y),-> Distribution *P(y)* , *Optimal y(x)*

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\, log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right) = S(P(y)) - S(P(y|x))$$

$$S\big(P(y)\big) = -\int dy\, P(y) \log P(y)$$

$$P(y|x) = \frac{e^{-\frac{(y-y(x))^2}{2\ \sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

- S(P(y|x) = $\int dx\, P(x) \int P(y|x) \log(P(y|x) =$

=$\int dx\, P(x) \int P(y|x) \left(-\frac{(y-y(x))^2}{2\sigma^2} - \frac{1}{2}\, log(2\,\pi\,\sigma^2)\right)$ = - $\frac{1}{2}$ $log_2(2\,\pi\, e\, \sigma^2)$

Entropy of a Gaussian distribution

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\, log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right) = -\int dy\, P(y) \log P(y)\ - \frac{1}{2}\ log(2\,\pi\, e\, \sigma^2)$$

Maximise the MI distribution with respect to y: Maximize the Entropy of P(y)

Argmax $_{P(y)}$ $[-\int dy\, P(y) \log P(y) + \lambda\, (\int dr\, P(y) - 1)]$

P(y)= const in the interval $[0, y_{max}]$= $\dfrac{1}{y_{max}}$

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\; log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right) = -\int dy\, P(y) \log P(y)\; - \tfrac{1}{2}\; log(2\,\pi\, e\, \sigma^2)$$

Maximise the MI distribution with respect to y: Maximize the Entropy of P(y)

Argmax $_{P(y)}$ $[-\int dy\, P(y) \log P(y) + \lambda\, (\int dr\, P(y) - 1)]$

$\Downarrow$

P(y)= const in the interval [0,y_{max}]= $\dfrac{1}{y_{max}}$

- In the hypothesis of a small gaussian noise: $y \sim y(x)$

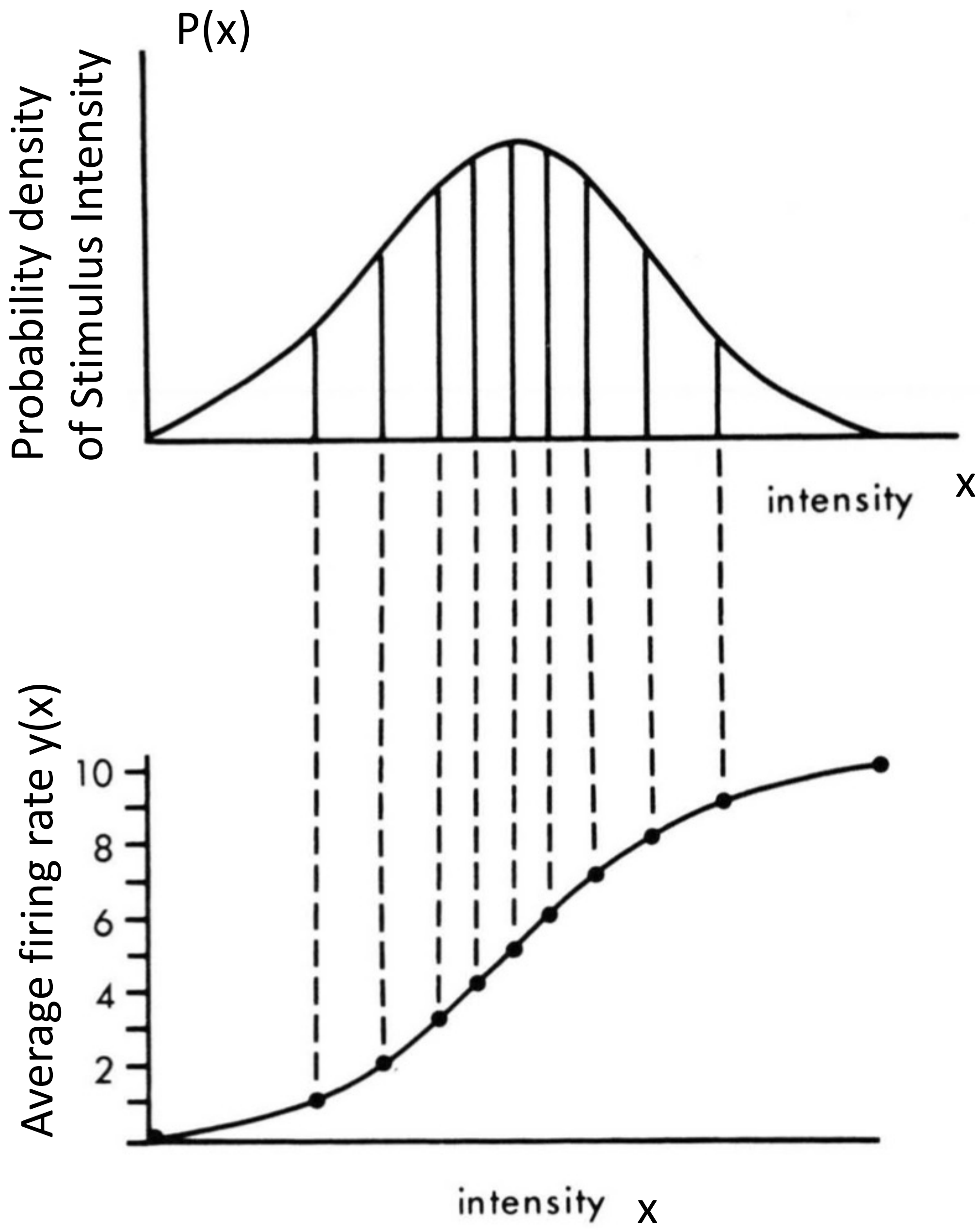
- Looking for *y(x)* such that P(y)=1/y_{max} and P(x) (stimulis intensity) is given.

By changing variable:

P(y) $\frac{dy}{dx}$=P(x) $\Rightarrow$ $\frac{y(x)}{y_{max}} = \int_{x_{min}}^{x} P\,(\mathrm{x})\mathrm{dx}$

Efficient Coding

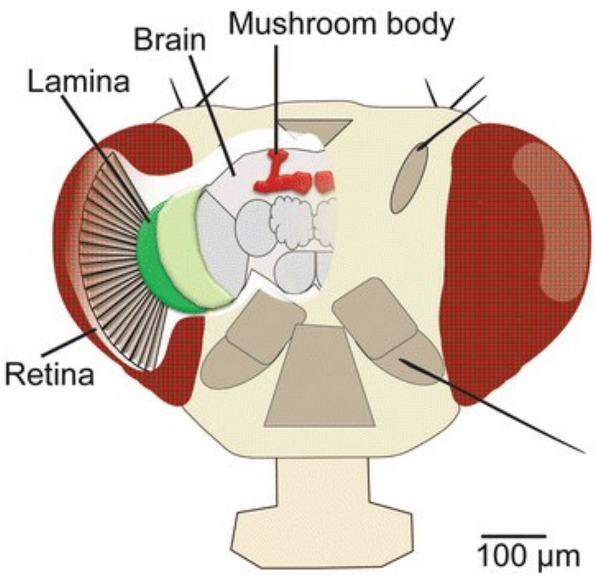
$$\frac{y(x)}{y_{max}} = \int_{x_{min}}^{x} P\,(\mathrm{x})\mathrm{dx}$$

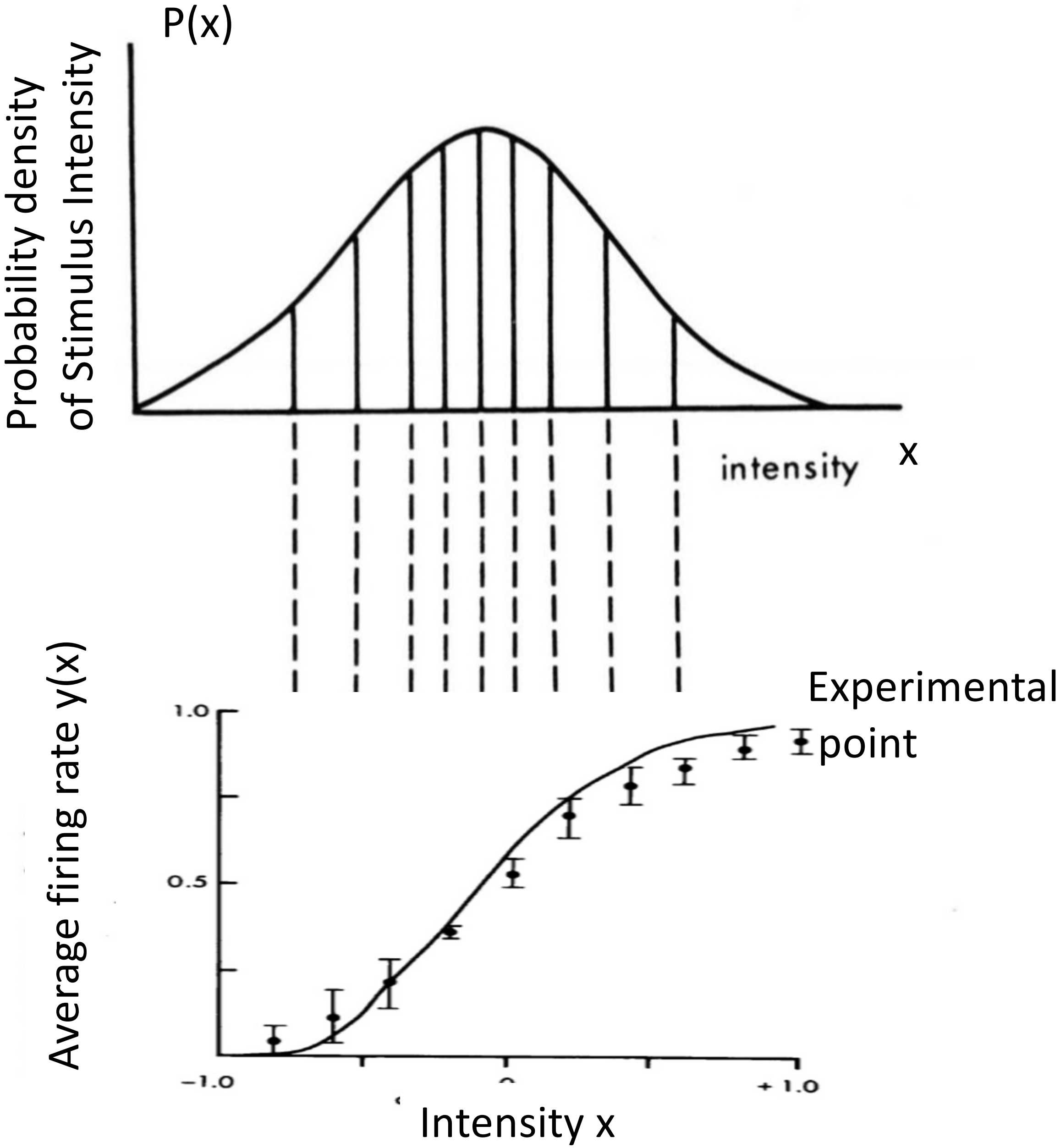


Efficient Coding

S.B. Laughlin, A Simple Coding Procedure Enhances a Neuron's Information Capacity, 1981

$$\frac{y(x)}{y_{max}} = \int_{x_{min}}^{x} P\,(\mathrm{x})\mathrm{dx}$$





Bring home message :

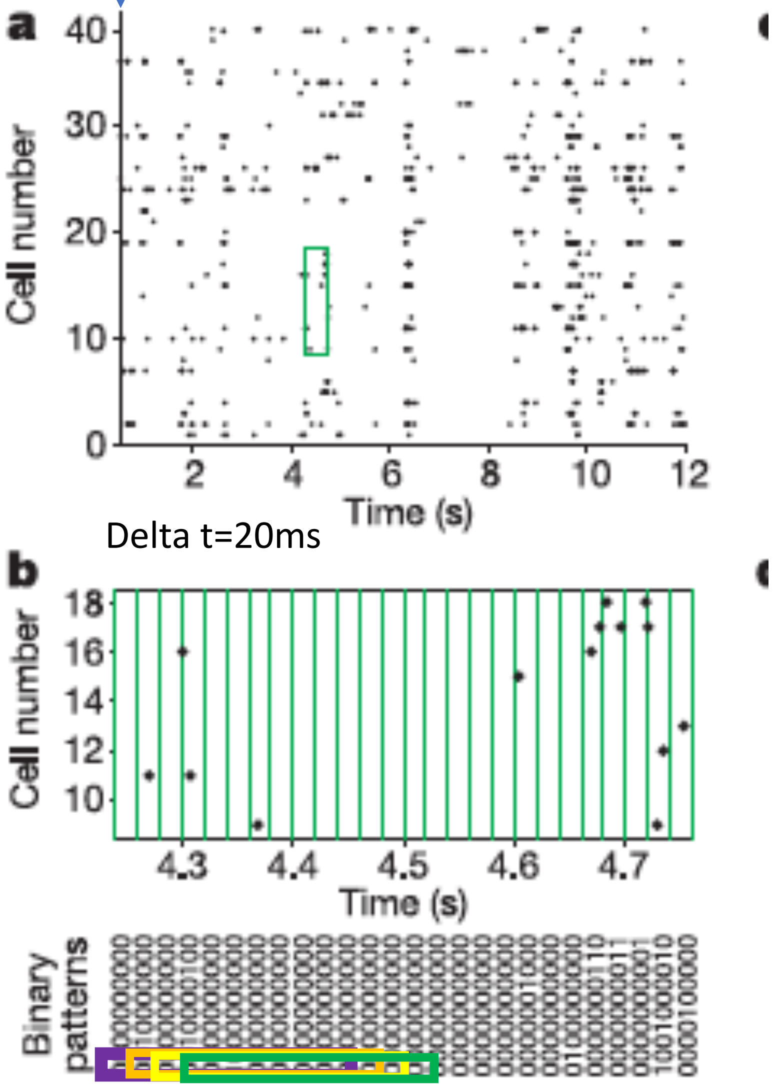
- The Mutual Information between 2 variables quantify the gain in information on one variable knowing the other one

- Efficient Coding: Maximising Mutual Information between Stimulus and firing rate

- The entropy and Asymptotic Inference: Definition of
Cross Entropy and Kullback Leibler Divergence,

Tutorial: Entropy and Information conveyed by the spiking activity of a neuron about a stimulus

movie starts ....movie end –movie starts again



120 repeats of a natual movie of 26.5 s each

Words: Patterns of 10 binary letters

0000000000

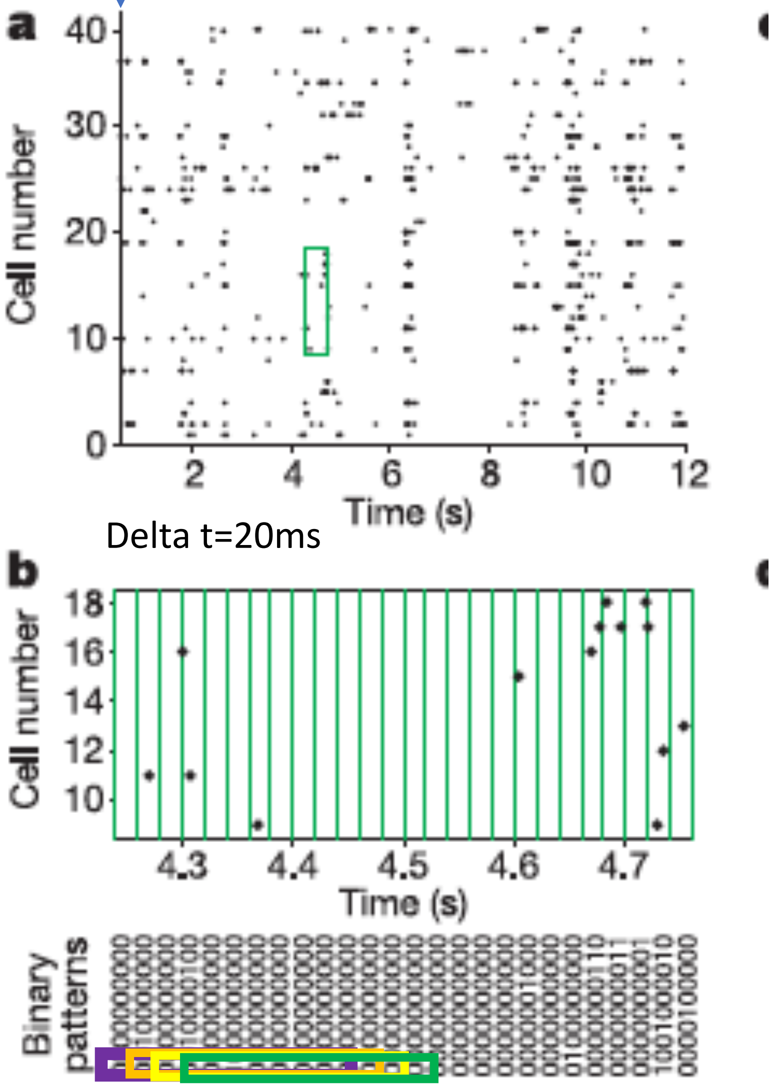
0000010000

0000001000

Tutorial: Entropy and Information conveyed by the spiking activity of a neuron about a stimulus

:

movie starts ……………………………………………………………..movie end –movie starts again



120 repeats of a natual movie of 26.5 s each

Words: Patterns of 10 binary letters

0000000000

0000010000

0000001000

- Extract the Entropy of the ‘language’ : Total Entropy
- Study How this entropy is reduced at fixed stimulus (over the 120 repetitions) =Information conveyed by the neuron about the stimulus
- How it does depend on the neuron frequency?

Supplementary slides

Mutual Information & Entropy

MI(x_1,x_2)= S[p(x_1)- S[p(x_1|x_2)] = S[p(x_2)- S[p(x_2|x_1)]

Conditional entropy of x_1 at given x_2

S[p(x_1|x_2)]= $-\sum_{x_2} p(x_2\ ) \sum_{x1} p(x_1|x_2) \log_2 (\ p(x_1|x_2)$

Excerciese compute conditional Entropy for the Boxes and Cookies given MI=0.05,

S[p(Box)]= $\log_2 2$ bit= 1 bit, S[p(cookie)]= $-\frac{3}{8}\ log_2\left(\frac{3}{8}\right) - \frac{5}{8}\ log_2\ \left(\frac{5}{8}\right) = 0.95\ bit$

S[p(Box|cookie)]= 0.95 bit S[p(cookie|box)]= 0.9 bit

Entropy of a distribution p : Notations

$$S(p) = -\sum_i p_i \log p_i :$$

$$-\sum_{y \in A} p(y) \log p(y)$$

For categorical variables

$$: -\int_A \mathrm{d}y\, p(y) \log p(y) \;.$$

For continous variables
p(y) is a probability density