Information theory and efficient coding in neural systems

Eugenio Piasini
Neural Computation Lab
International School for Advanced Studies

Outline

- Part 1: motivation – why information theory?
- Part 2: Shannon’s idea – information, compression and surprise
- Part 3: Efficient coding in neural systems
- Demo

Outline

- **Part 1: motivation – why information theory?**
- Part 2: Shannon's idea – information, compression and surprise
- Part 3: Efficient coding in neural systems
- Demo

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

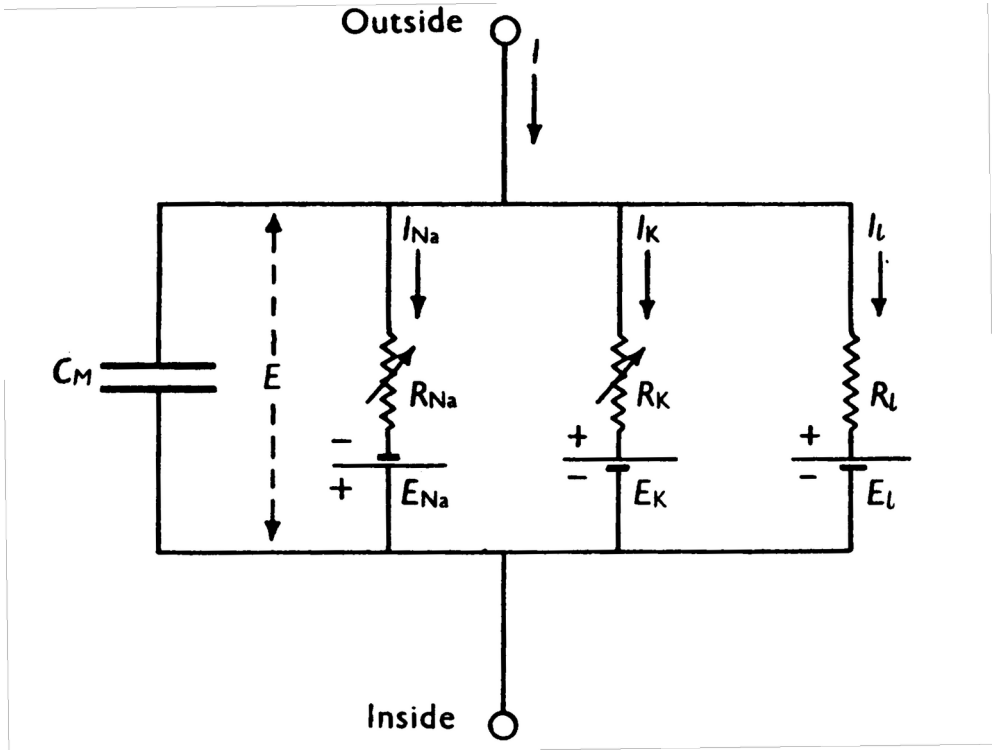
By C. E. SHANNON

$$\nabla \cdot \boldsymbol{D} = \rho$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0$$

$$\nabla \cdot \boldsymbol{B} = 0$$

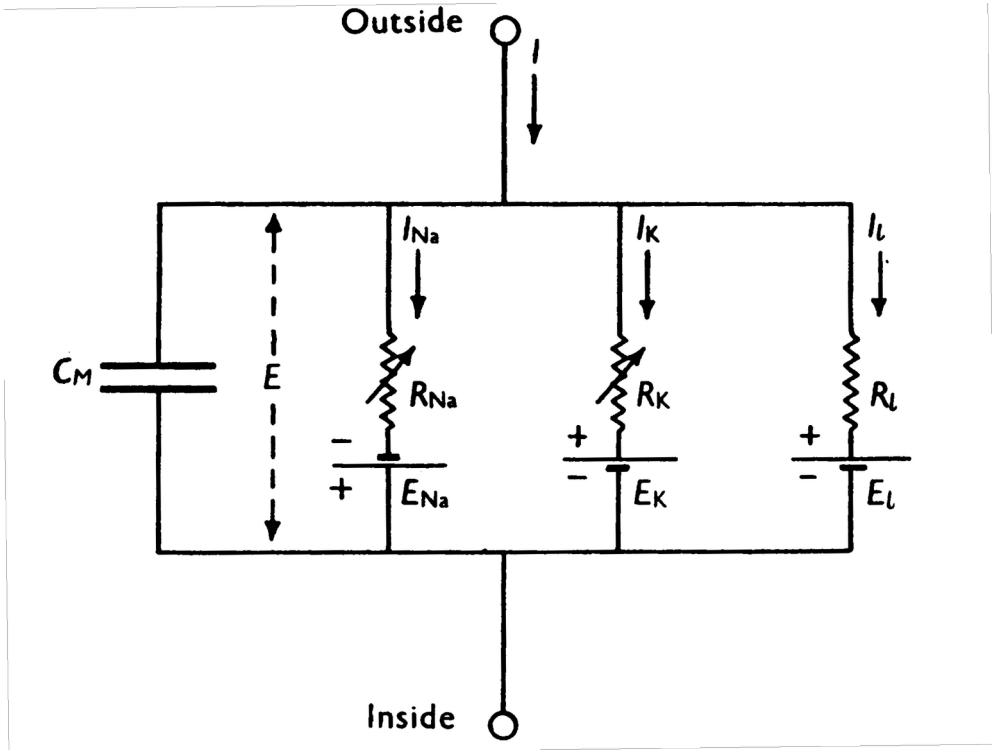


$$\nabla \cdot \boldsymbol{D} = \rho$$

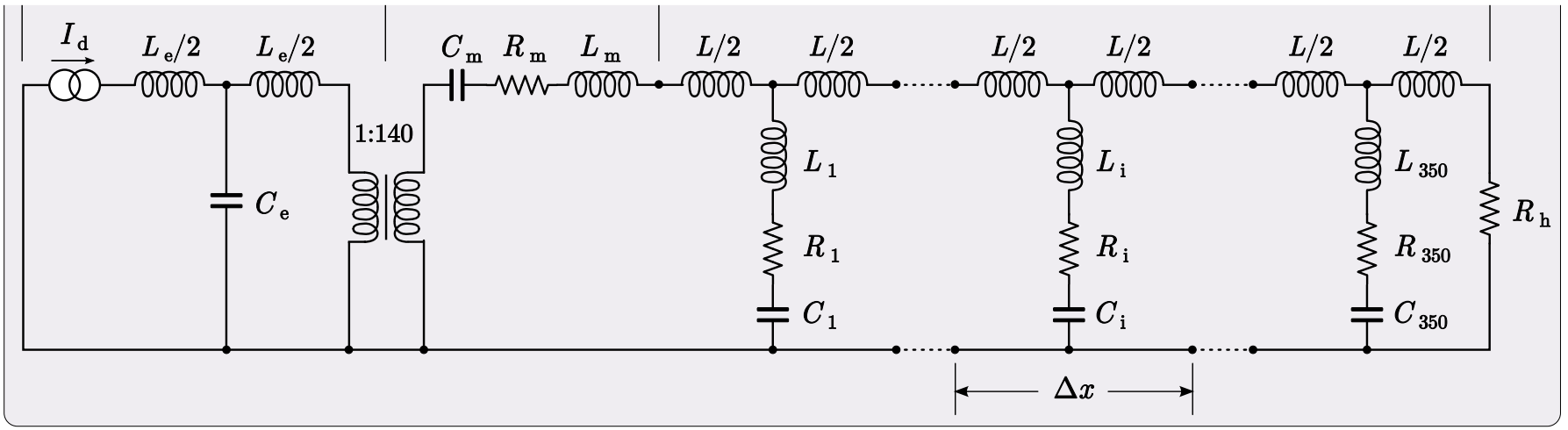
$$\nabla \times \boldsymbol{H} = \boldsymbol{J}$$

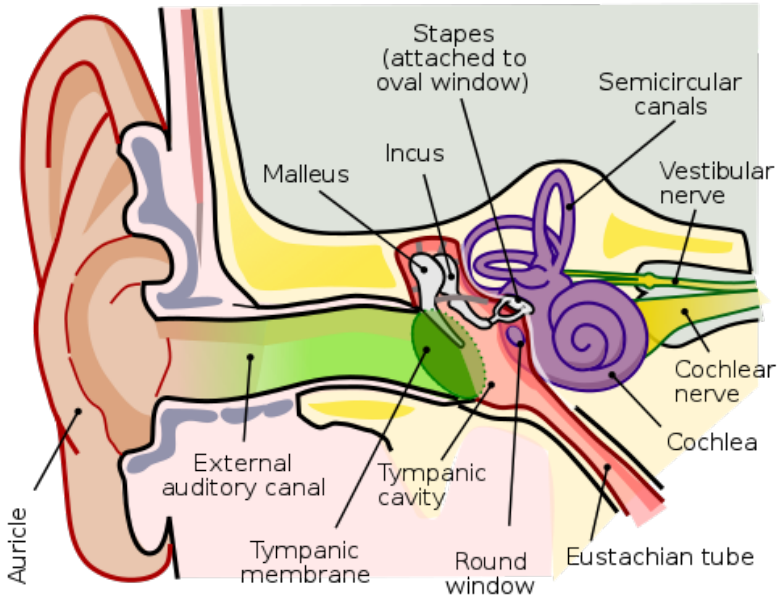
$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0$$

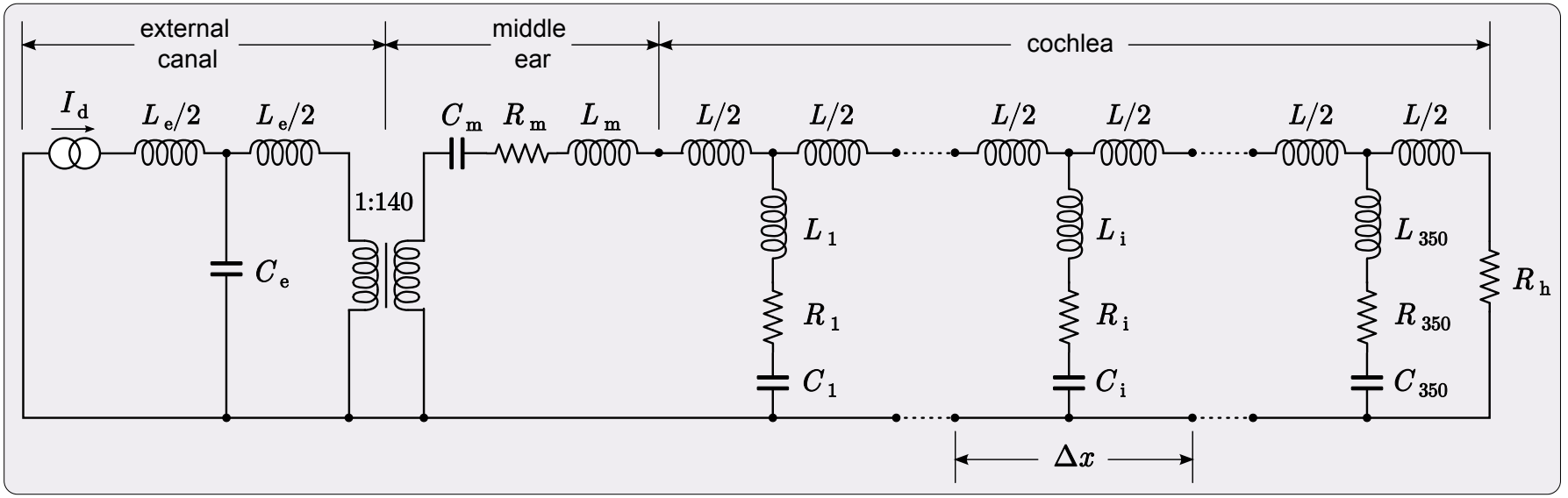
$$\nabla \cdot \boldsymbol{B} = 0$$



More general ⟷ Less general


I_{d}
$L_{\mathrm{e}}/2$
$L_{\mathrm{e}}/2$
C_{m}
R_{m}
L_{m}
$L/2$
$L/2$
$L/2$
$L/2$
$L/2$
$L/2$
1:140
C_{e}
L_1
R_1
C_1
L_{i}
R_{i}
C_{i}
L_{350}
R_{350}
C_{350}
R_{h}
Δx


Stapes
(attached to
oval window)
Semicircular
canals
Incus
Malleus
Vestibular
nerve
Cochlear
nerve
Cochlea
External
auditory canal
Tympanic
cavity
Auricle
Tympanic
membrane
Round
window
Eustachian tube

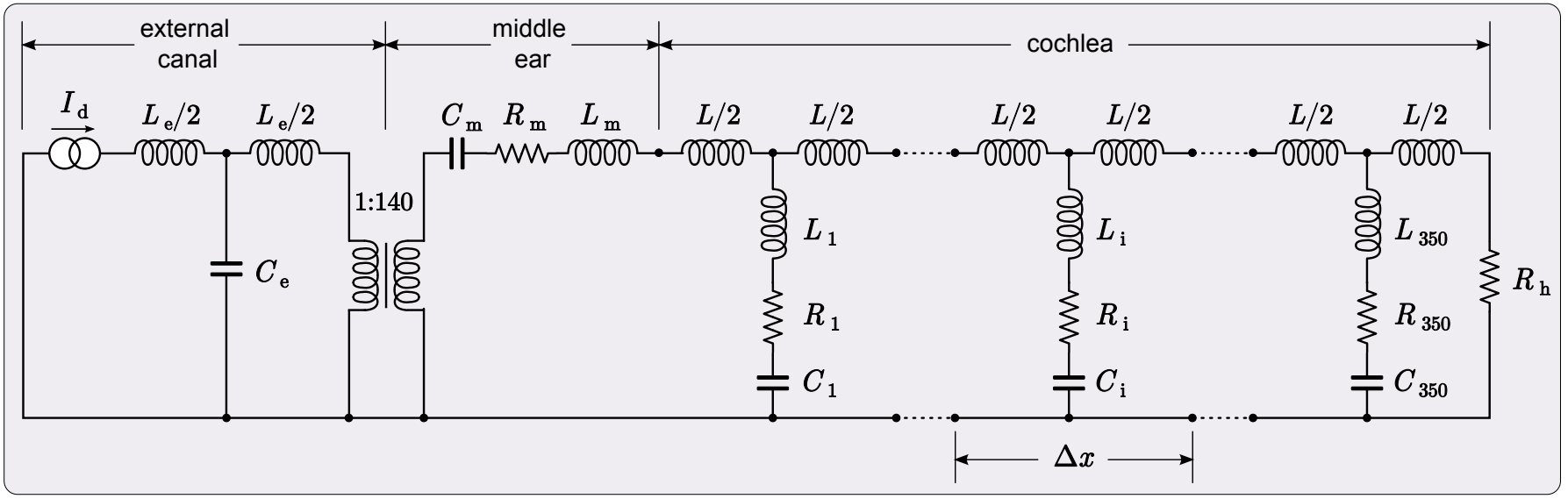

external
canal
middle
ear
cochlea
I_d
$L_e/2$
$L_e/2$
C_m
R_m
L_m
$L/2$
$L/2$
$L/2$
$L/2$
$L/2$
$L/2$
1:140
C_e
L_1
R_1
C_1
L_i
R_i
C_i
L_{350}
R_{350}
C_{350}
R_h
Δx

$$\nabla \cdot \boldsymbol{D} = \rho$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0$$

$$\nabla \cdot \boldsymbol{B} = 0$$



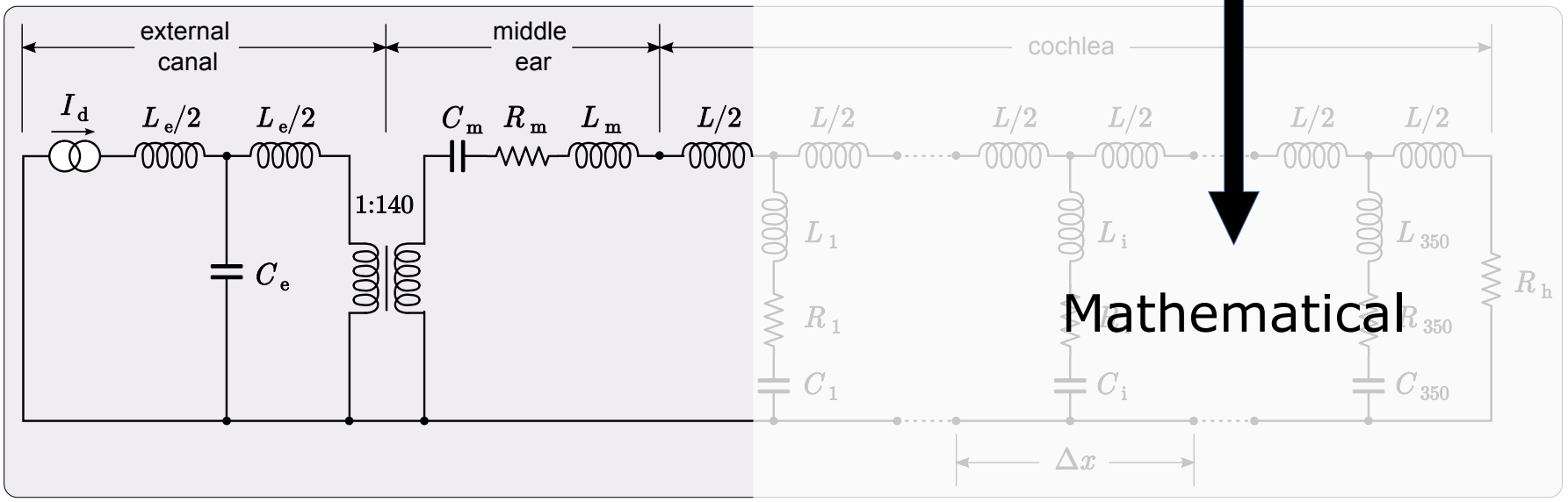
$$\nabla \cdot \boldsymbol{D} = \rho$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = 0$$

$$\nabla \cdot \boldsymbol{B} = 0$$

Physical



Mathematical

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

By C. E. SHANNON

Why information theory in neuroscience?

- *How much* is being encoded (as opposed to *what* or *how*)
  - Model-independent
  - Data-type agnostic
  - Not limited to linear interactions
  - Produces results on a universal, meaningful scale (bits)
- Measures of encoding efficiency
- Possible foundation for theoretical understanding of design

Outline

- Part 1: motivation – why information theory?
- **Part 2: Shannon's idea – information, compression and surprise**
- Part 3: Efficient coding in neural systems
- Demo

Real-world data sources are *redundant*

Real-world data sources are *redundant*

Mr. and Mr*. Dursley,*of number *our, Privet D*ive, were proud to *ay that t*ey were p*rfectly n*rmal, tha*k you very much. Th*y were t*e last p*ople you*d expect*to be in*olved in anyt*ing str*nge or m*ysterio*s, beca*se they*just di*n’t hold w*th suc* nonse*se.

Mr. D*rsley*was t*e dir*ctor *f a f*rm ca*led Grunnings, whic* mad* dri*ls. *e wa* a b*g, b*efy *an w*th h*rdly*any nec*, a*tho*gh *e d*d h*ve * ve*y l*rge*mus*ach*. M*s. Du*sl*y *as*th*n *nd*bl*nd* a*d *ad*ne*rl* t*ic* t*e *su*l a*o*n* *f*n*c*,*w*i*h*c*m* *n*v*r* *s*f*l*a* *h* *p*n* *o m**h**f**e**t**e**r**i** **e**g**d** **n**s**s**i** ** **e n***h***s***h***u***e***h***a***a***s***c***e***u***y***d*** t**** ****i****h****w****o****e****y****w****.

Real-world data sources are *redundant*

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

Redundancy, compression and surprise

Redundancy, compression and surprise

0000000000…

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is “surprising” and tells us something we didn’t know!

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is “surprising” and tells us something we didn’t know!
→ surprise is informative

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is “surprising” and tells us something we didn’t know!
→ surprise is informative

31124011223444010 3...

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is “surprising” and tells us something we didn’t know!
→ surprise is informative

31124011223444010 3...

Increasing the range of things that can happen increases surprise, and increases information

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

0110101110111000101…

Each new digit is “surprising” and tells us something we didn’t know!
→ surprise is informative

31124011223444010З...

Increasing the range of things that can happen increases surprise, and increases information

011011001111011….

Redundancy, compression and surprise

0000000000…

Always 0. Each new “0” tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is “surprising” and tells us something we didn’t know!
→ surprise is informative

31124011223444010 3...

Increasing the range of things that can happen increases surprise,
and increases information

011011001111011….

Ones always come in pair so the second one tells us nothing (redundancy).
We could replace 11→2 and we would shorten (compress) the string!

Redundancy, compression and surprise

0000000000…

Always 0. Each new "0" tells us nothing!

1111111111…

Same thing!

01101011101110001010…

Each new digit is "surprising" and tells us something we didn't know!
→ surprise is informative

31124011223444010 3...

Increasing the range of things that can happen increases surprise,
and increases information

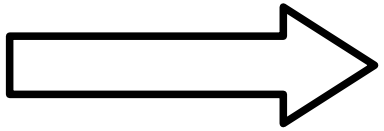
011011001111011….

Ones always come in pair so the second one tells us nothing (redundancy).
We could replace 11→2 and we would shorten (compress) the string!

We would "learn" (= be surprised) the most if data looked completely random!
So… randomness is informative?

Redundancy, compression and surprise

$x_1\ x_2\ x_3\ x_4\ \ldots$



Q1: What is the information content of the message?
Q2: How much space will the message take on disk?

How do we measure surprise?

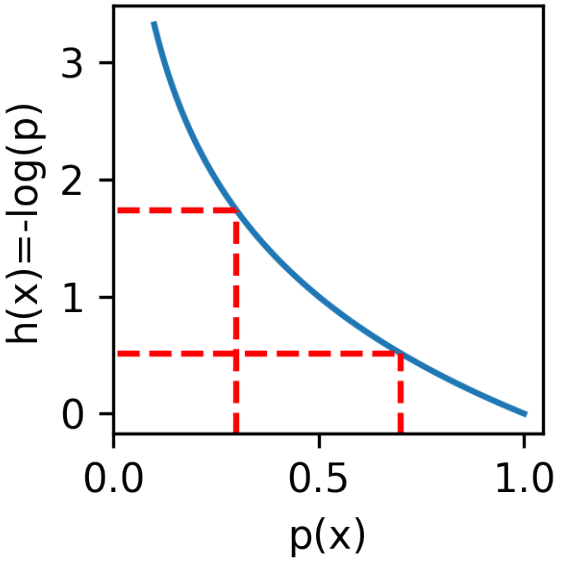
x = 0 or 1

p(0) = 0.3
p(1) = 0.7

Ex. message: 11011111111000111011…

Shannon's idea: a good measure of surprise for event x is

$$h(x) = \log_2 \frac{1}{p(x)}$$ (Shannon's information content)



How do we measure surprise?

x = 0 or 1

p(0) = 0.3
p(1) = 0.7

Ex. message: 11011111111000111011…

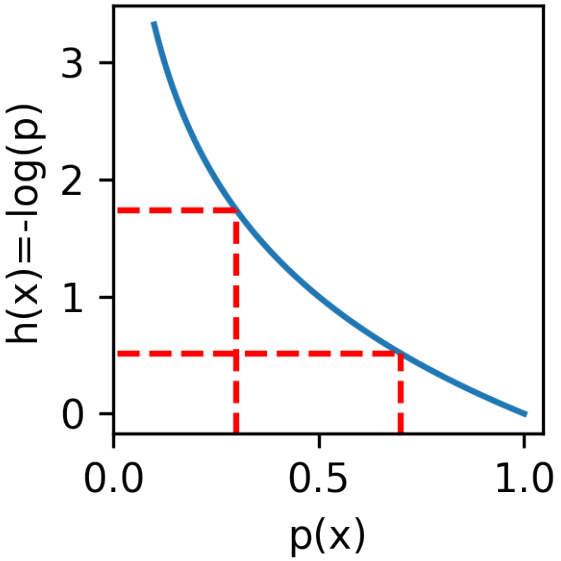
Shannon's idea: a good measure of surprise for event x is

$$h(x) = \log_2 \frac{1}{p(x)}$$ (Shannon's information content)

Properties:

1. h(x) is biggest for improbable (=surprising) outcomes:
   h(0.3) = 1.73 > 0.51 = h(0.7)

2. h is **additive** for independent random variables:
   if p(x,y) = p(x)p(y),
   then h((x,y)) = -log(p(x,y)) = -log(p(x)) -log(p(y))
   = h(x) + h(y)



Entropy: average information content

Information content: h(x) = log(1/p(x)) = - log p(x)

But x is a random variable! What is its expected value?

$$H[X] = \sum_x p(x)h(x) = -\sum_x p(x)\log p(x)$$

Claim:

1. H is the correct measure of information
2. H is the compressed file length we should aspire to, in binary digits (bits)

Entropy: average information content

Information content: h(x) = log(1/p(x)) = - log p(x)

But x is a random variable! What is its expected value?

$$H[X] = \sum_x p(x)h(x) = -\sum_x p(x)\log p(x)$$

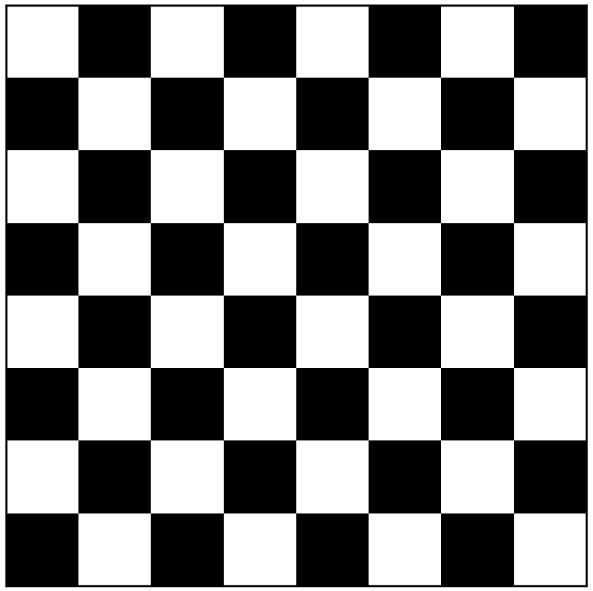
Claim:

1. H is the correct measure of information
2. H is the compressed file length we should aspire to, in binary digits (bits)

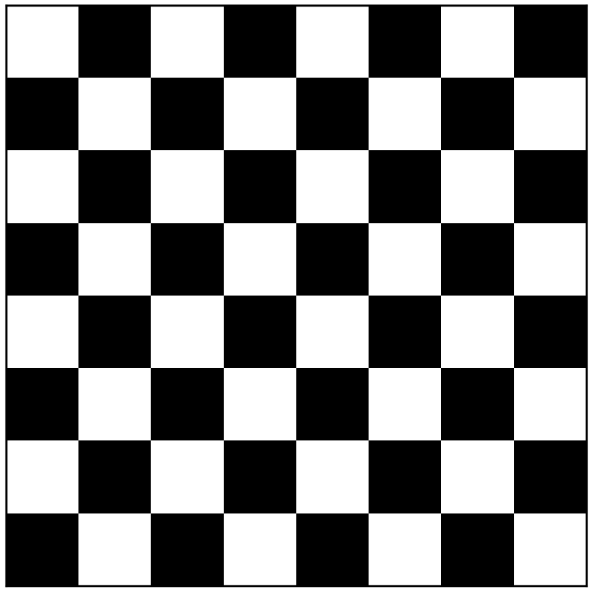
Example:

If p(x=0)=0.3, p(x=1)=0.7, then H[X]=0.88 bit

→ if we have a message of 1000 characters, we should be able to compress it down to about 880 binary digits without losing information.

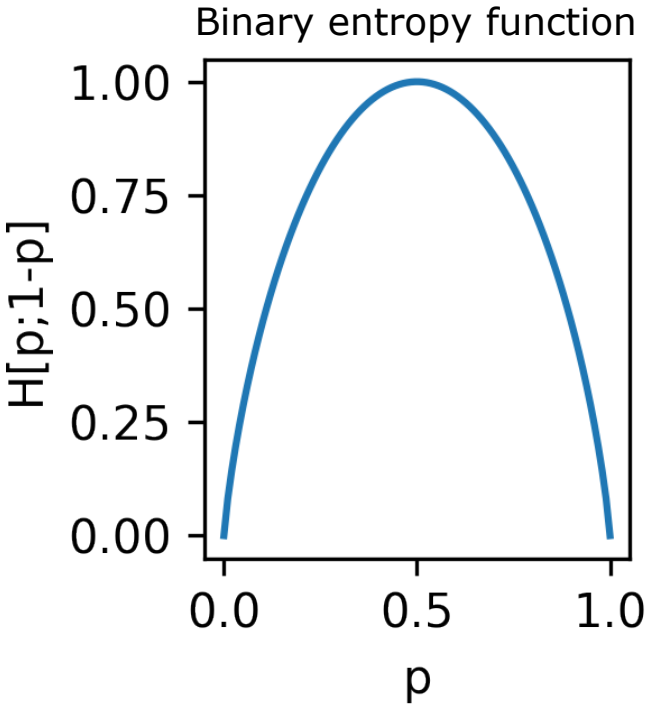


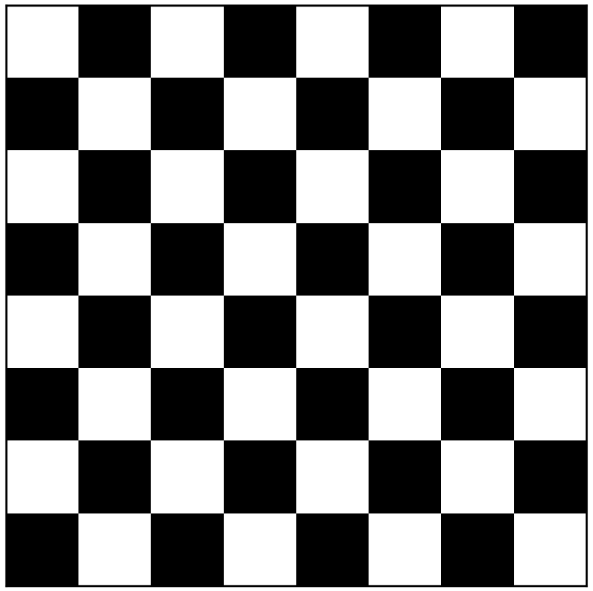
Q: Imagine having to guess a position on this 8x8 checkerboard. What is the minimum number of yes/no questions required? What is the information content of the answers? (remember $h(x)=-\log_2[p(x)]$)



Q: Imagine having to guess a position on this 8x8 checkerboard. What is the minimum number of yes/no questions required? What is the information content of the answers? (remember $h(x)=-\log_2[p(x)]$)

A: the optimal answers are those that maximise the information content. This happens for p(yes)=p(no)=1/2. **Uniform distributions maximise the entropy!**

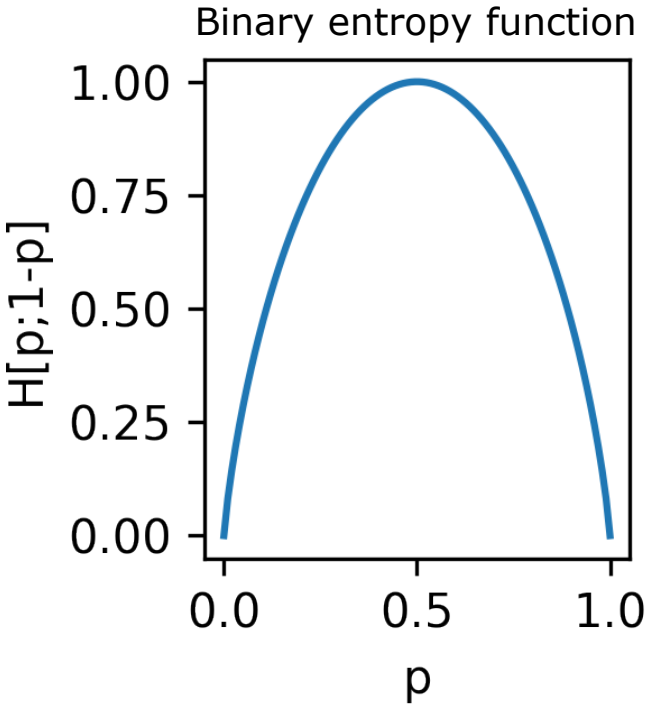


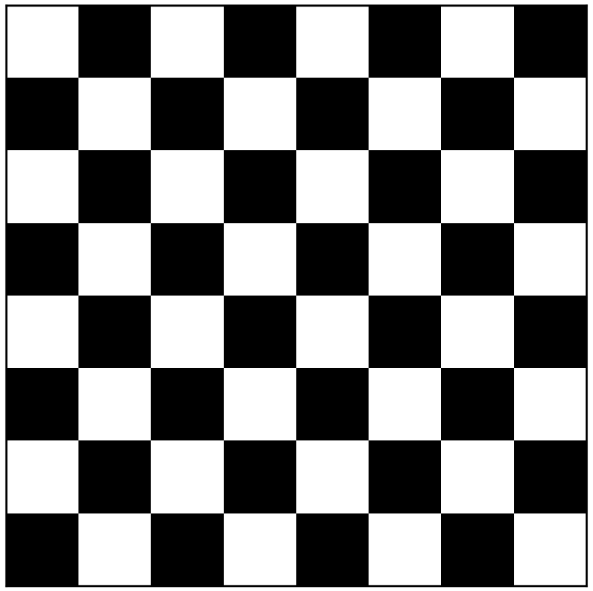


Q: Imagine having to guess a position on this 8x8 checkerboard. What is the minimum number of yes/no questions required? What is the information content of the answers? (remember $h(x)=-\log_2[p(x)]$)

A: the optimal answers are those that maximise the information content. This happens for p(yes)=p(no)=1/2. **Uniform distributions maximise the entropy!**

Each question has h(x)=log(2)=1 regardless of the outcome. Therefore, the total information content is **6 bit**.



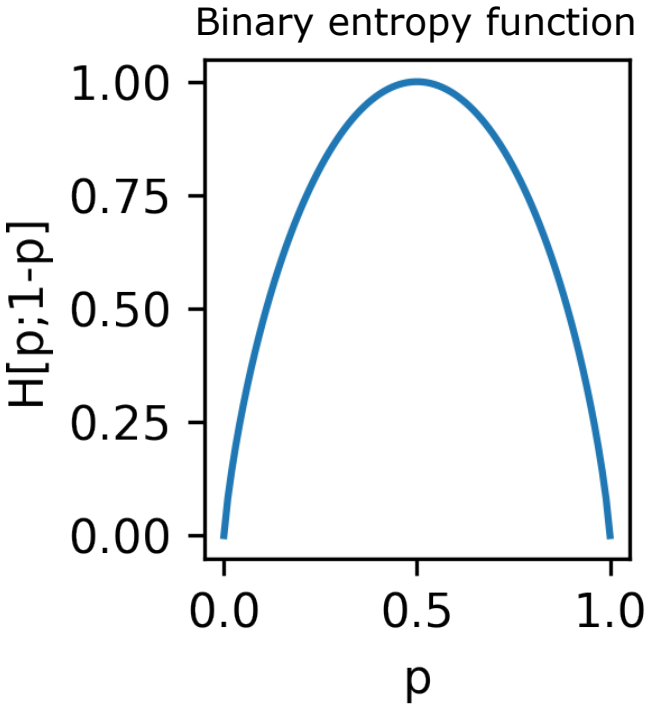


Q: Imagine having to guess a position on this 8x8 checkerboard. What is the minimum number of yes/no questions required? What is the information content of the answers? (remember $h(x)=-\log_2[p(x)]$)

A: the optimal answers are those that maximise the information content. This happens for p(yes)=p(no)=1/2. **Uniform distributions maximise the entropy!**

Each question has h(x)=log(2)=1 regardless of the outcome. Therefore, the total information content is **6 bit**.

But what about other strategies?



Outline

- Part 1: motivation – why information theory?
- Part 2: Shannon's idea – information, compression and surprise
- **Part 3: Efficient coding in neural systems**
- Demo

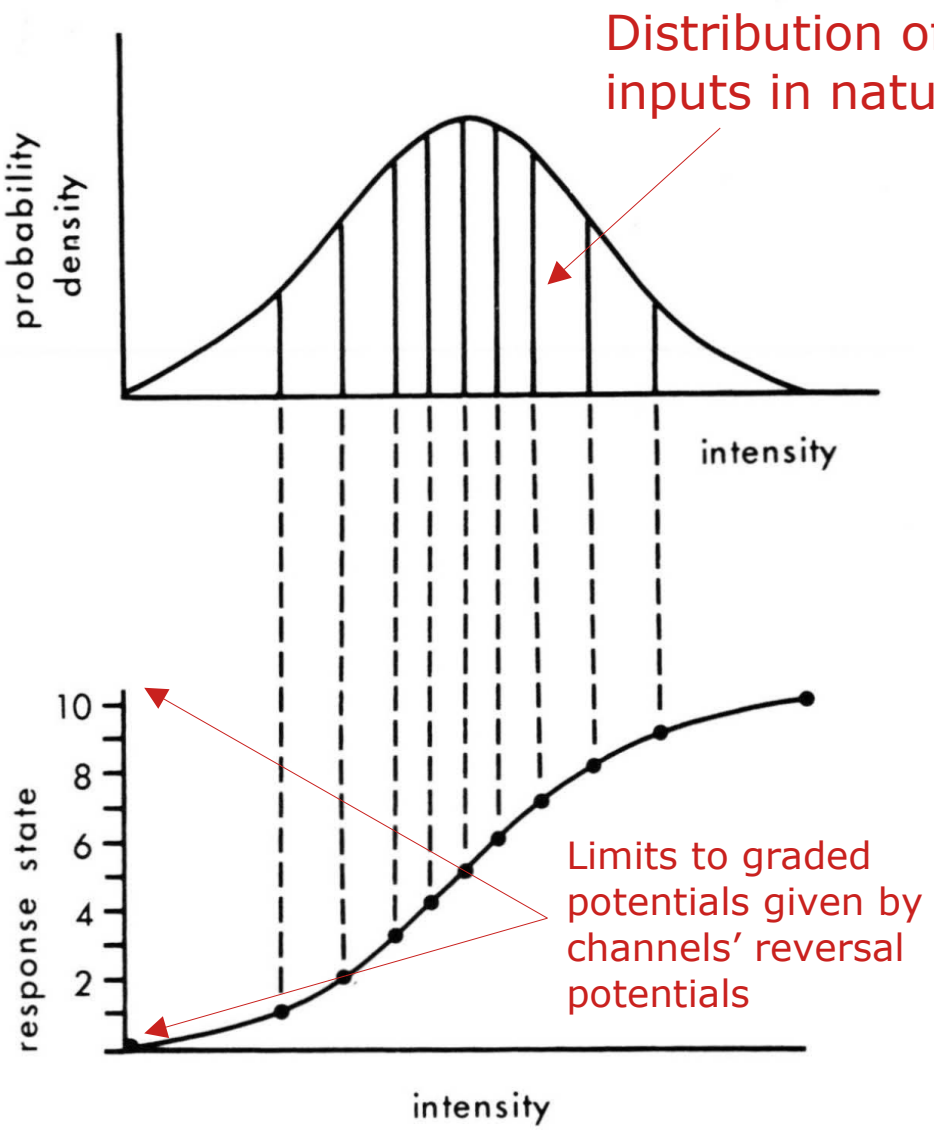
So far, in a nutshell

Entropy: $$H[X] = \sum_x p(x)h(x) = -\sum_x p(x)\log p(x)$$

1. The entropy is a reasonable measure of information content (it's the only one, really, but we don't have time to show it)

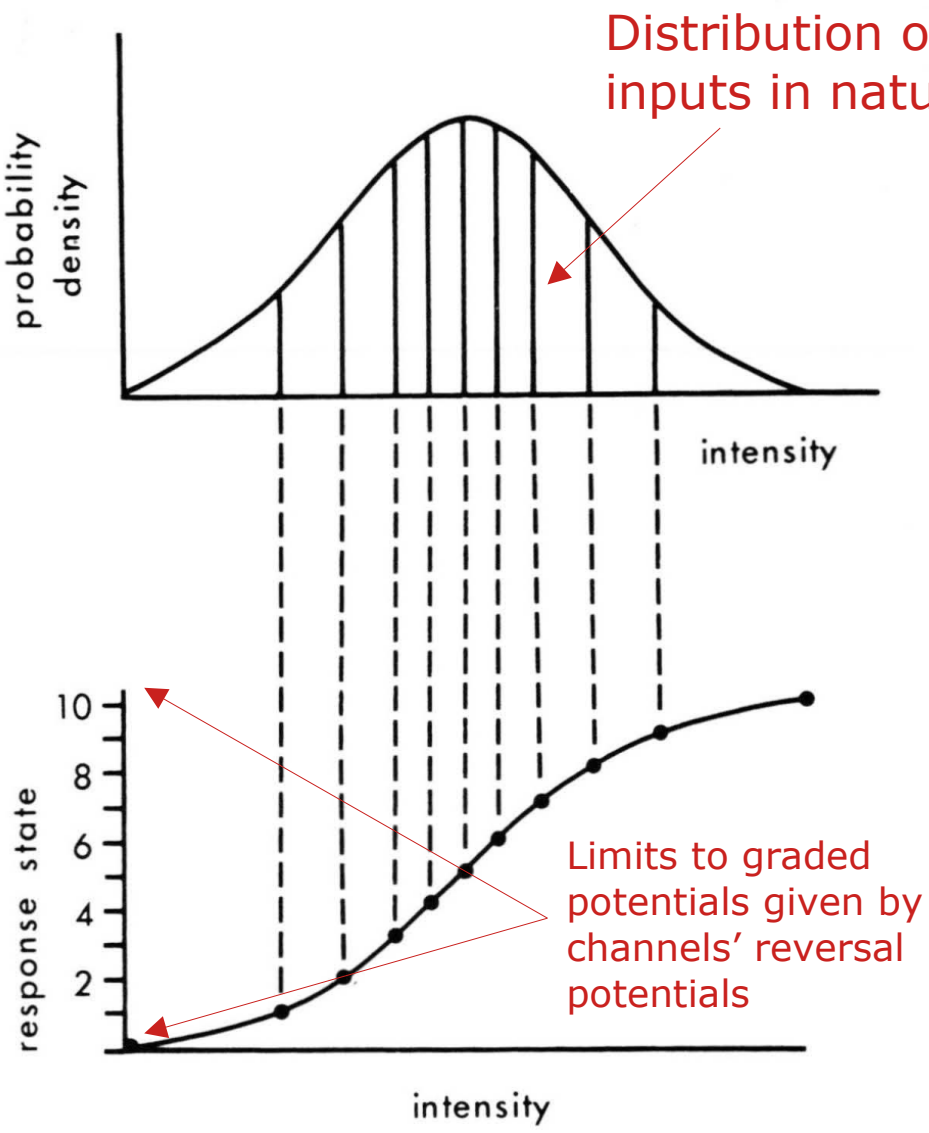
2. Entropy is maximised by uniform distributions

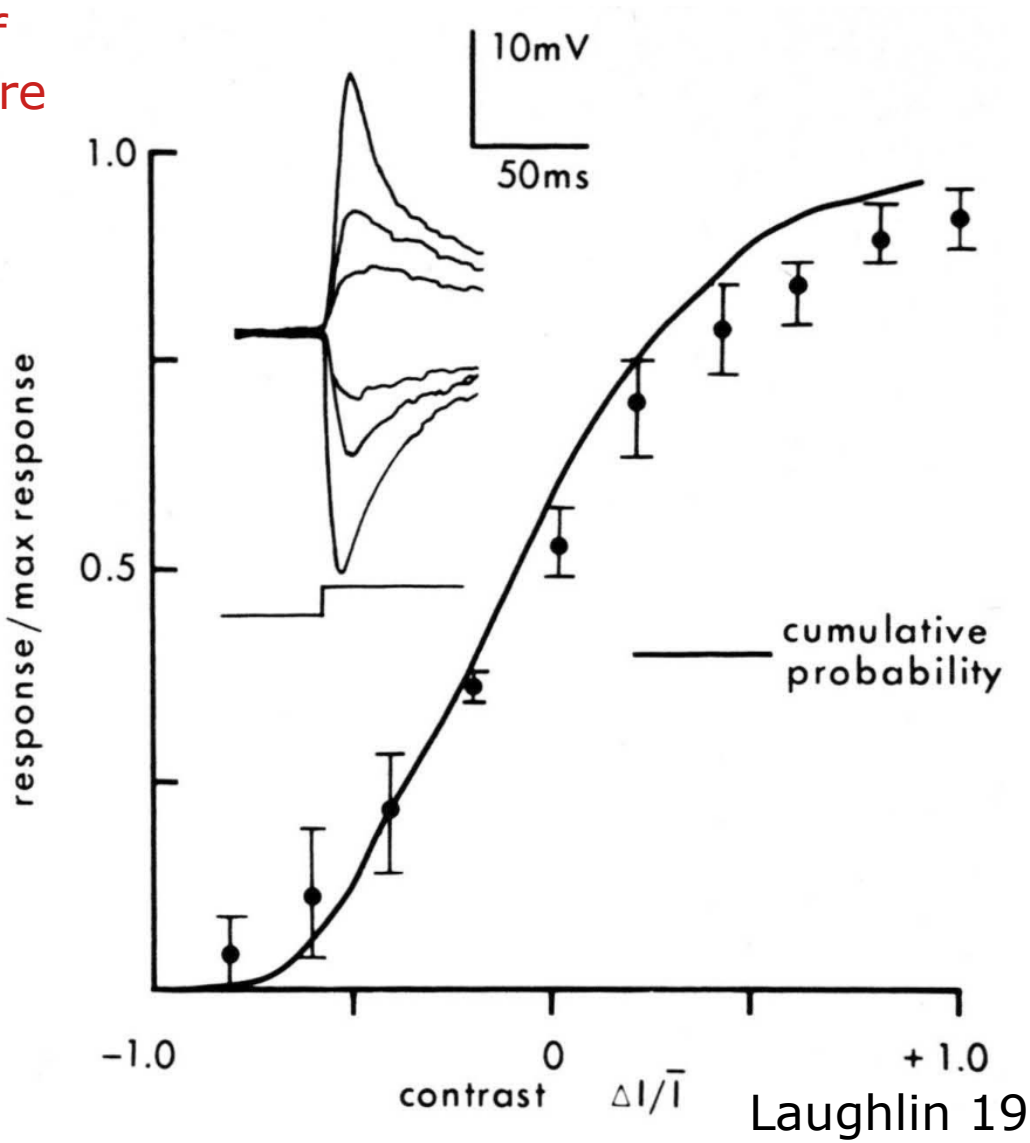
Large monopolar cell (blowfly compound eye)



Laughlin 1981

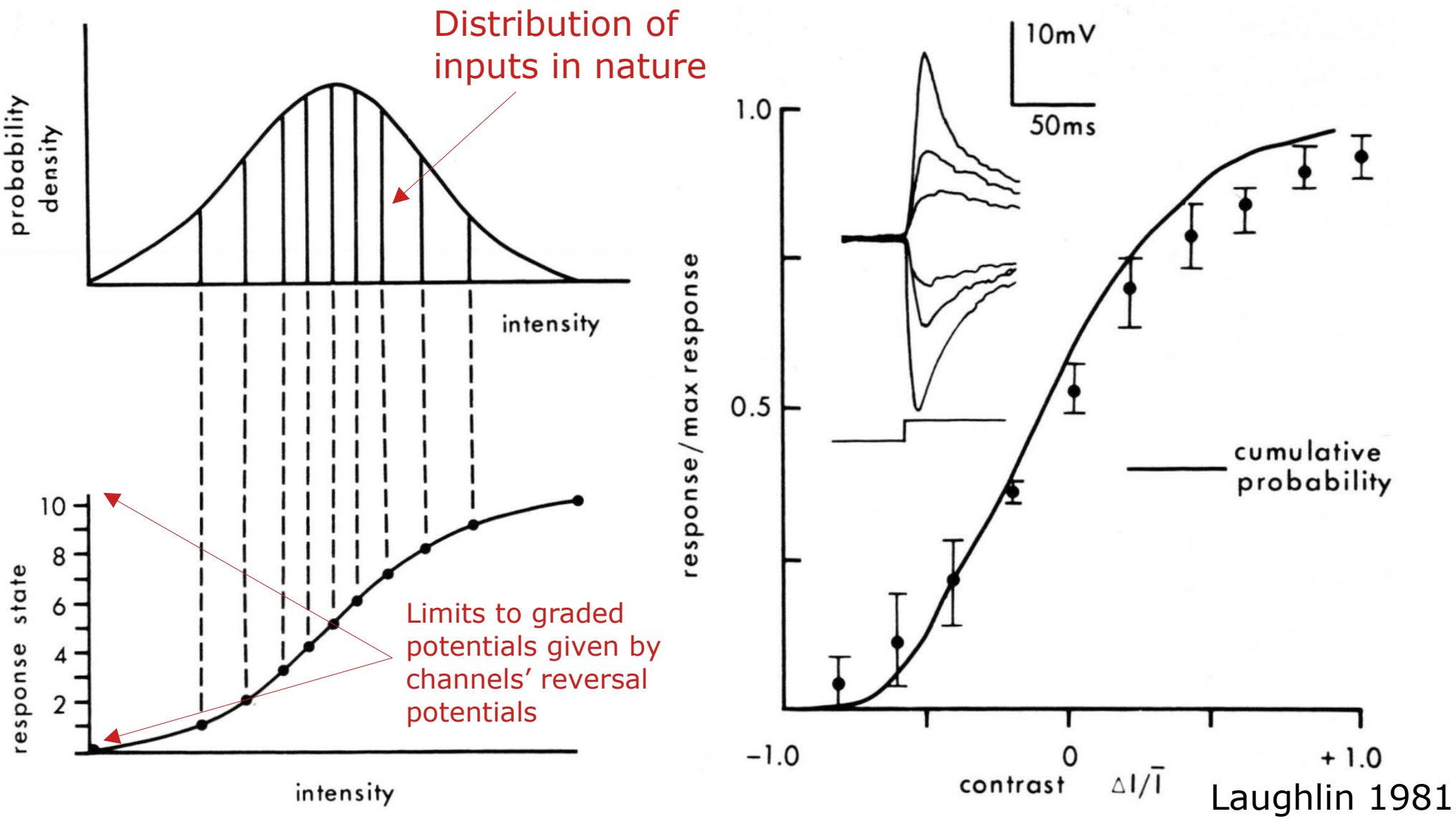
Large monopolar cell (blowfly compound eye)





Laughlin 1981

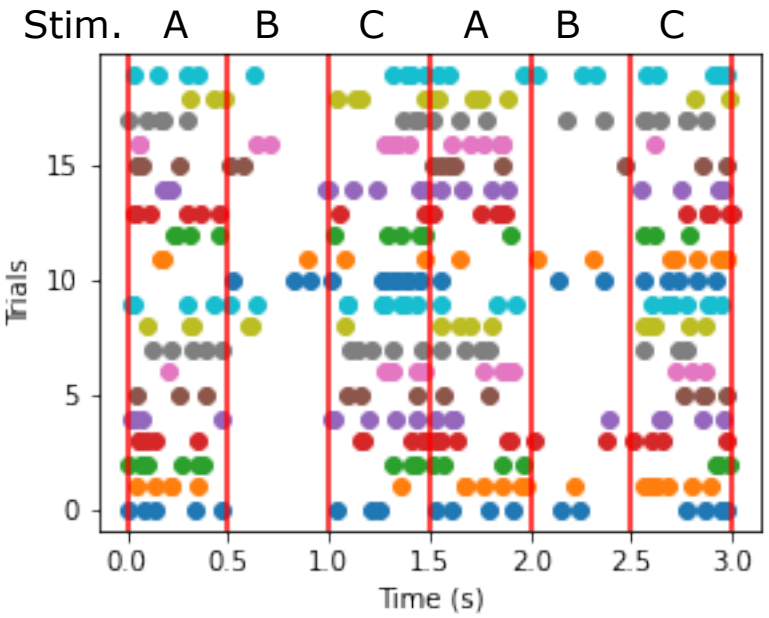
Large monopolar cell (blowfly compound eye)



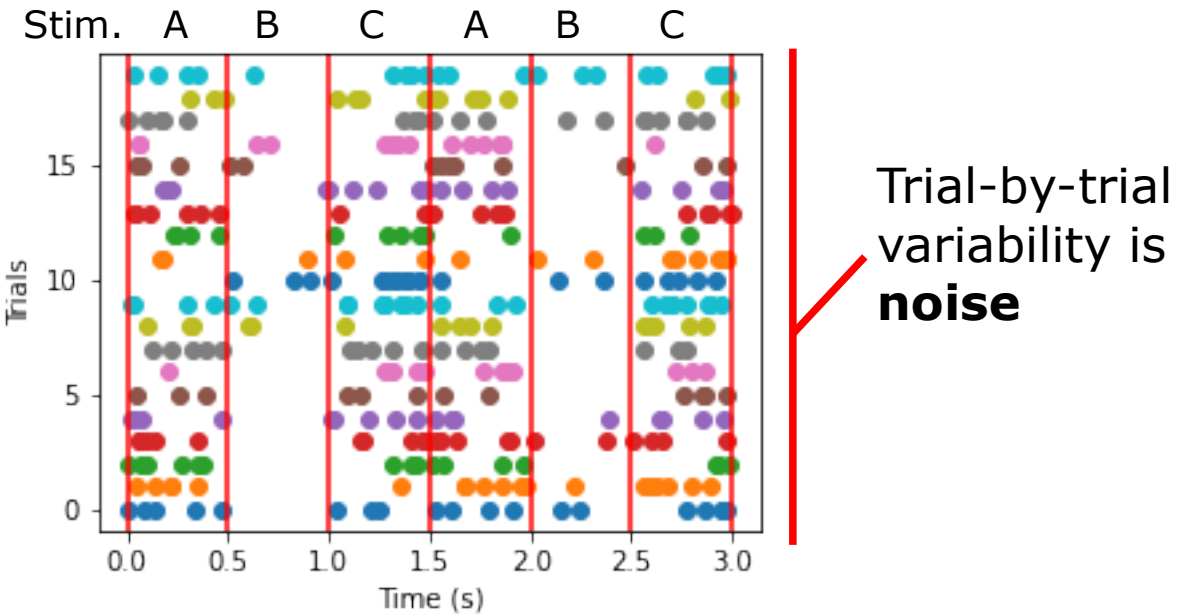
Laughlin 1981

Efficient coding: **the behavior of neurons and circuits is adapted to the statistics of the environment**

What about noise?

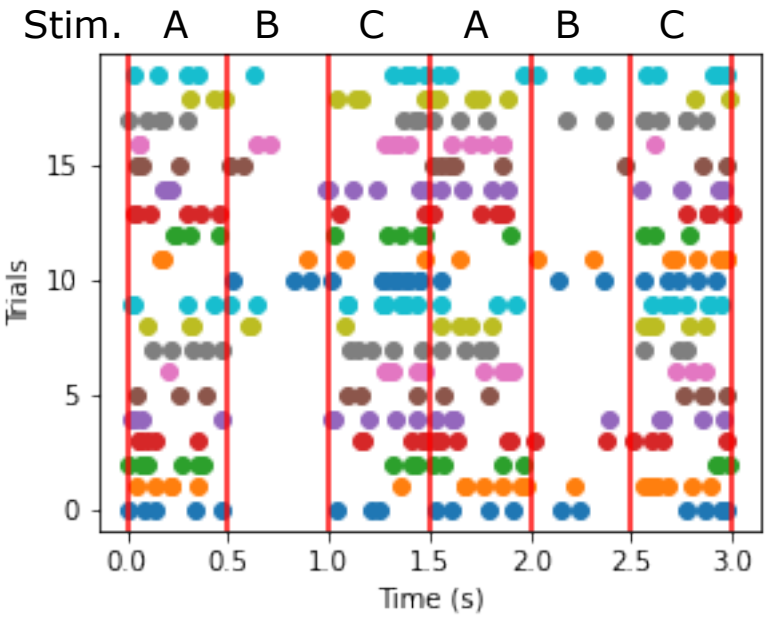


What about noise?



Trial-by-trial variability is **noise**

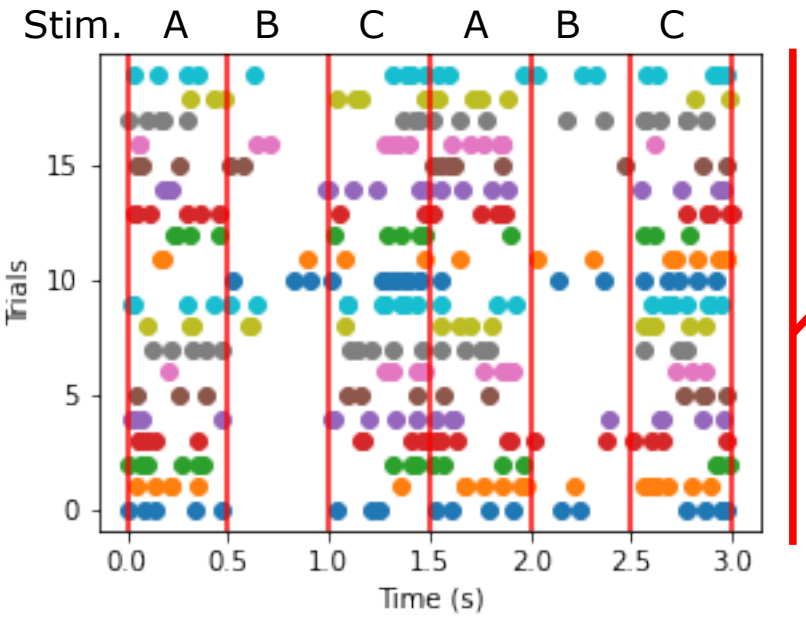
What about noise?



Trial-by-trial variability is **noise**

Variability associated with the (time-varying) stimulus is **information**

What about noise?



Trial-by-trial variability is **noise**

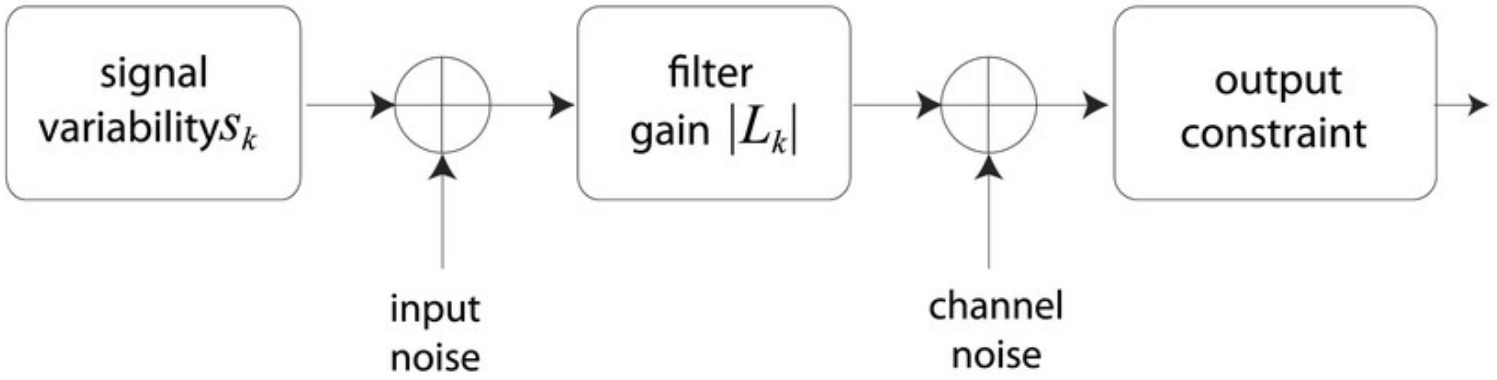
Variability associated with the (time-varying) stimulus is **information**

In presence of noise, information is defined as the total variability (surprise) of the neural activity R minus the amount of variability that persists at fixed stimulus S, and is therefore due to the noise.

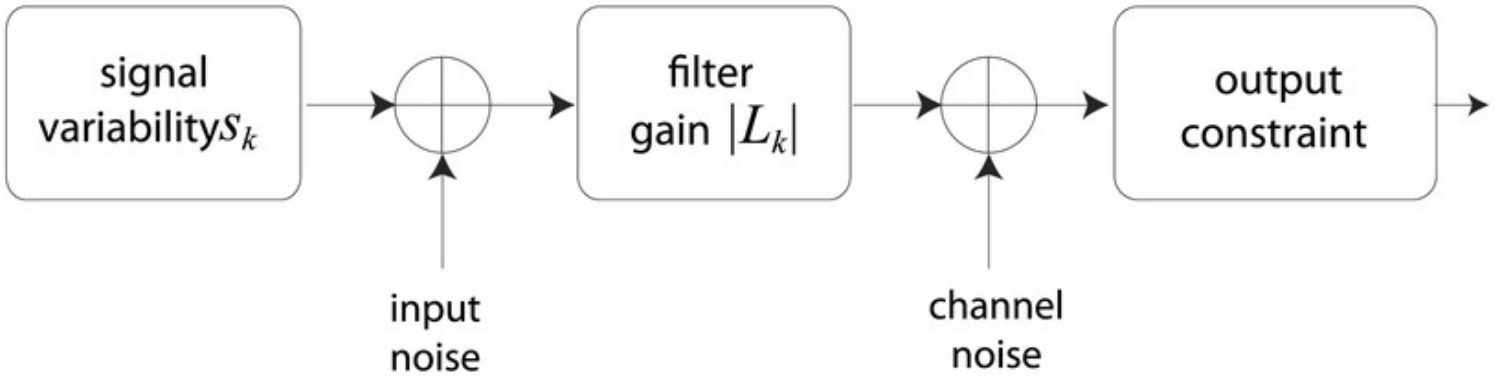
$$I[S:R] = H[R] - \sum_s p(s) H[R|S=s] = H[R] - H[R|S]$$

Shannon's **mutual information**

Optimal gain control in presence of noise



Optimal gain control in presence of noise



Biol. Cybern. 68, 23–29 (1992)

Biological Cybernetics

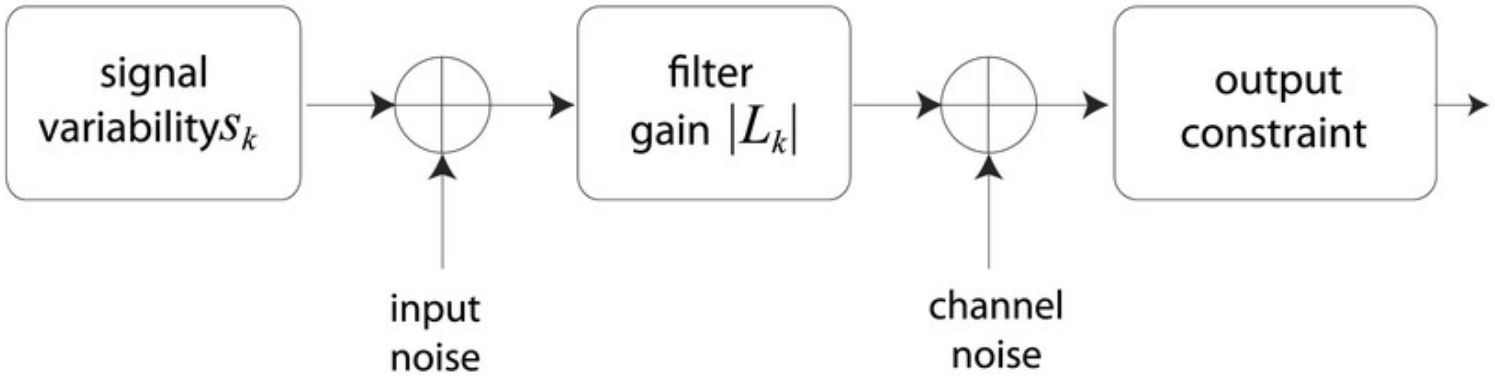
A theory of maximizing sensory information

J. H. van Hateren

Department of Biophysics, University of Groningen, Westersingel 34, 9718 CM Groningen, The Netherlands

Received April 2, 1992/Accepted in revised form May 22, 1992

Optimal gain control in presence of noise



Biol. Cybern. 68, 23–29 (1992)

Biological Cybernetics

A theory of maximizing sensory information

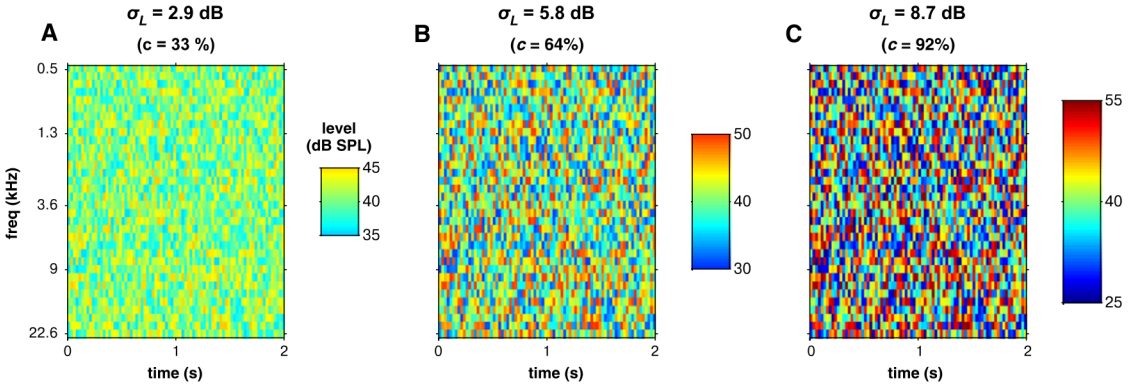
J. H. van Hateren

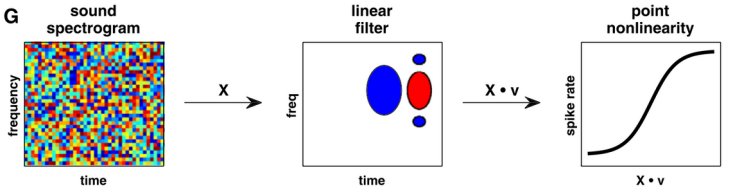
Department of Biophysics, University of Groningen, Westersingel 34, 9718 CM Groningen, The Netherlands

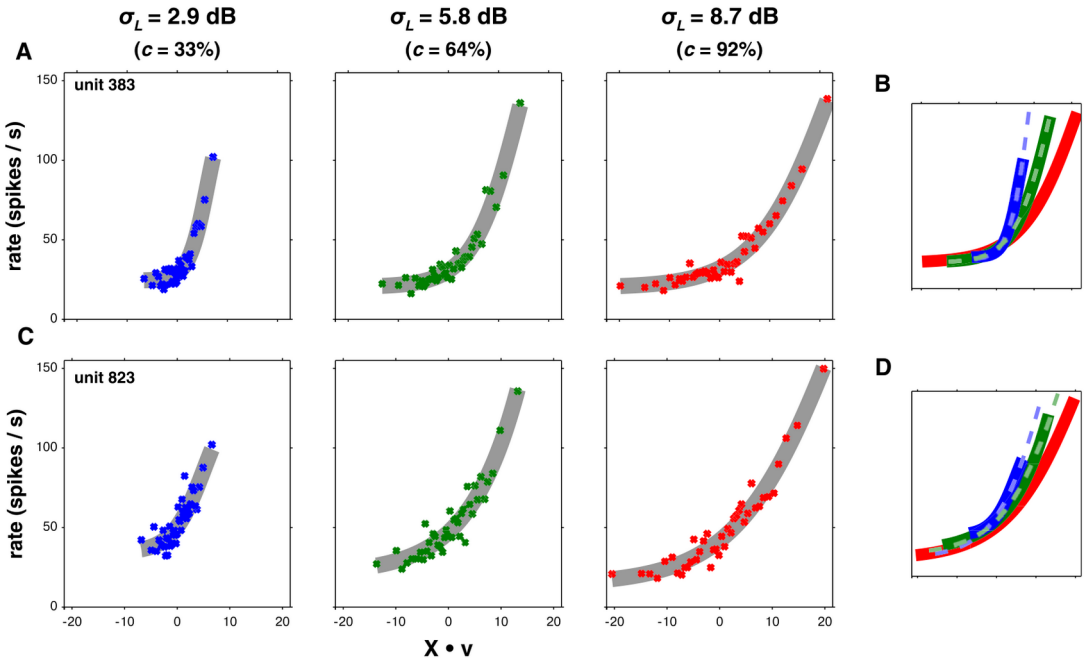
Received April 2, 1992/Accepted in revised form May 22, 1992

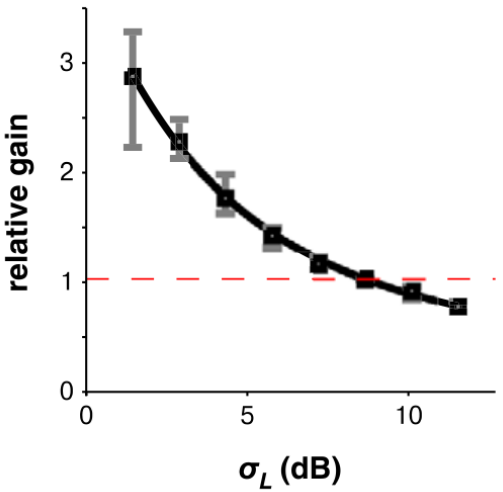
→ when the output noise/bandwidth limitation is the main constraint, neural gain should **decrease** as stimulus variability **increases.**

Contrast gain control





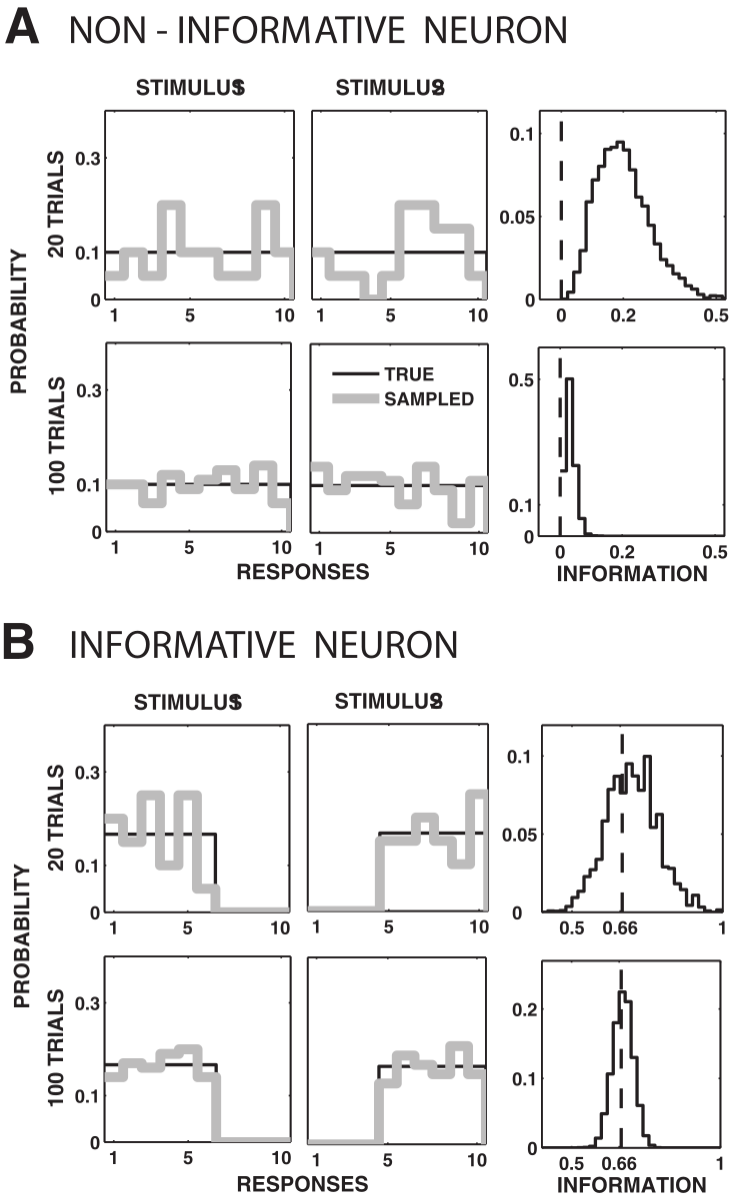




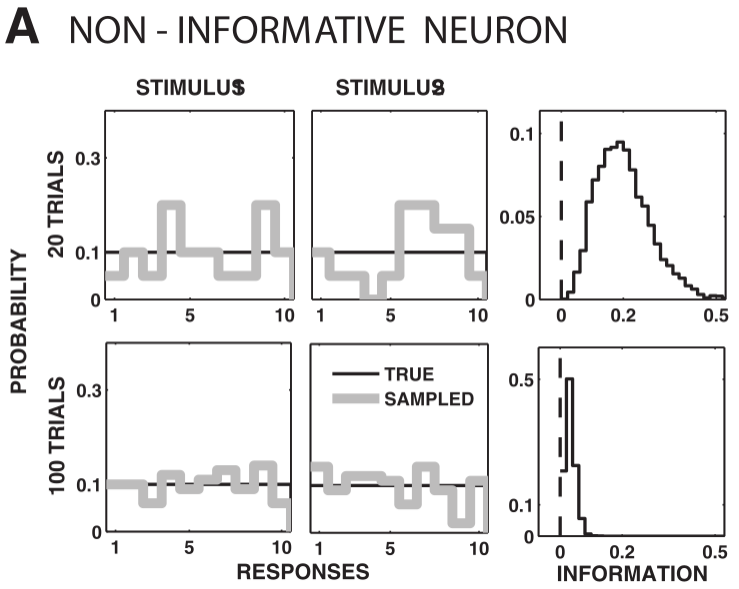
(recorded from A1 + AAF in ferrets)

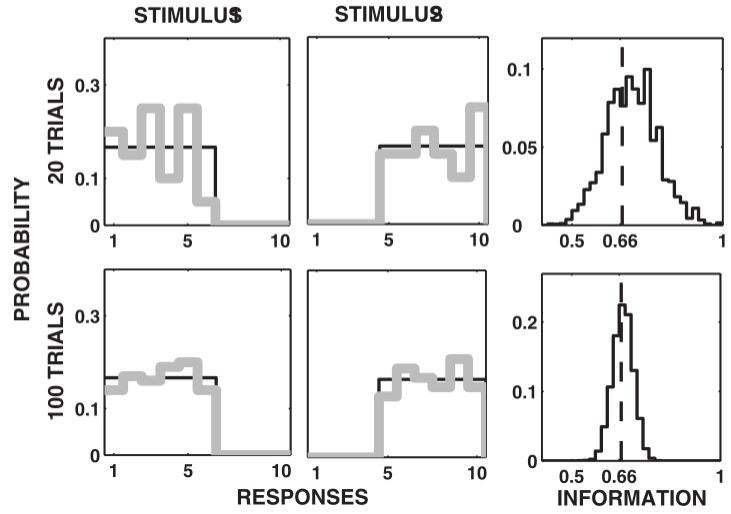
Rabinowitz et al, Neuron 2011

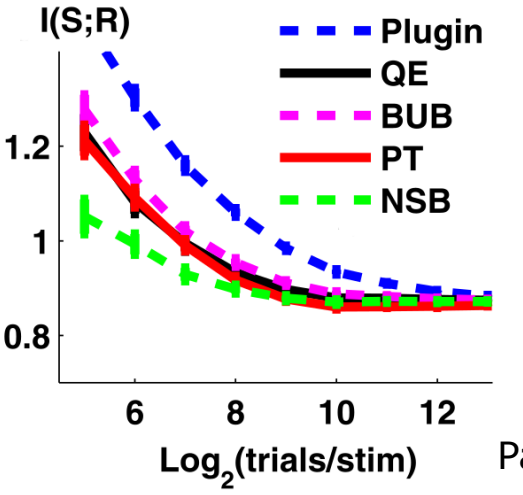
Limited sampling bias for entropy and information



Limited sampling bias for entropy and information

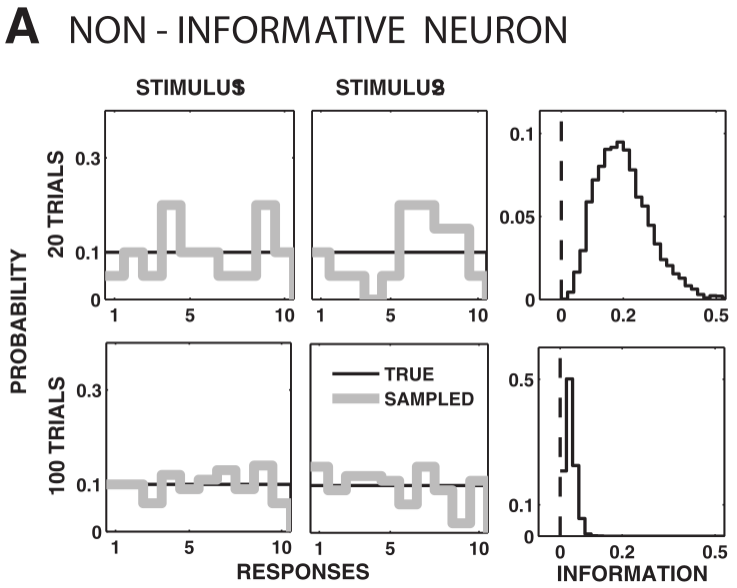


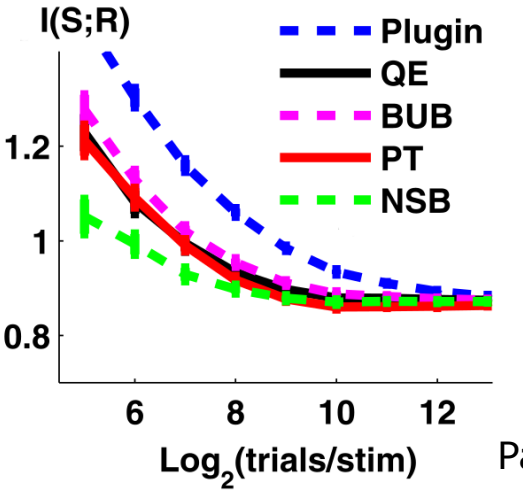




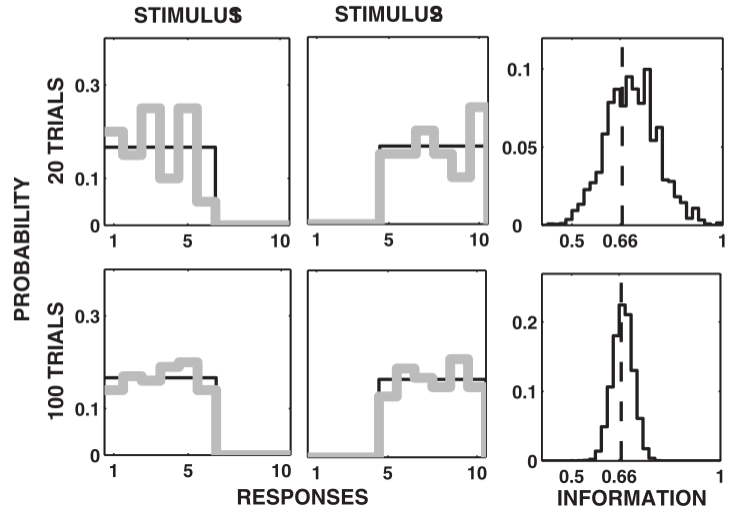
Panzeri et al 2007

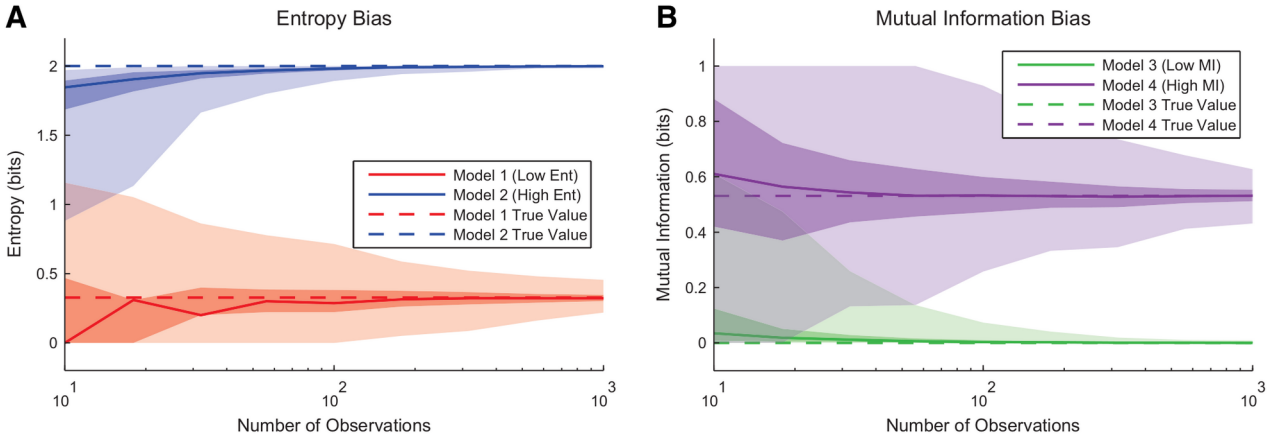
Limited sampling bias for entropy and information





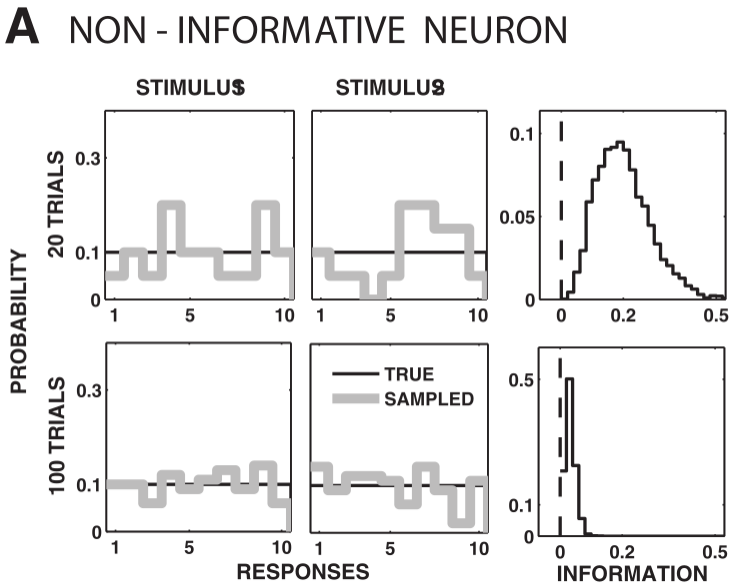
Panzeri et al 2007

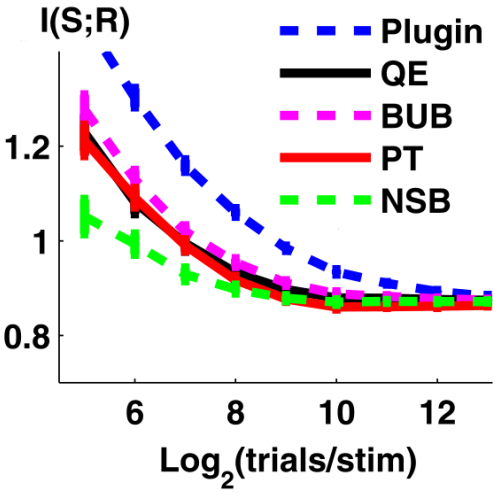




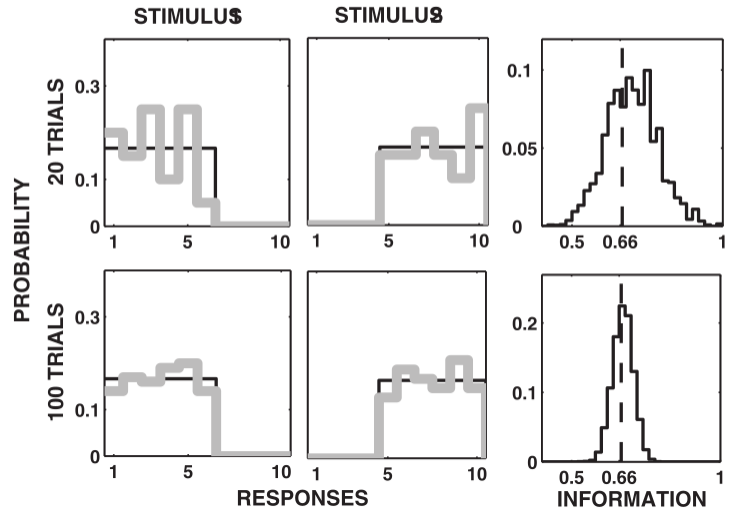
Timme et al 2018

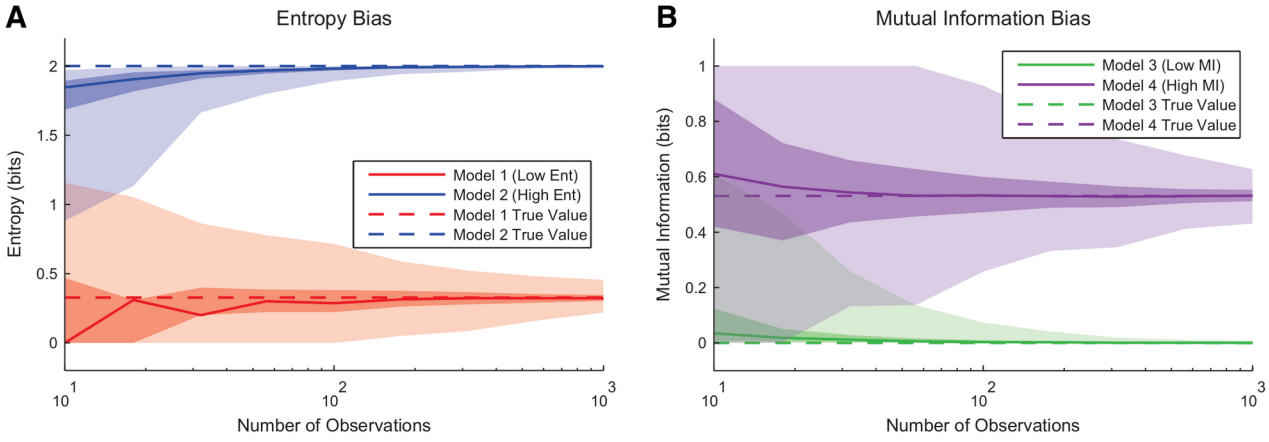
Limited sampling bias for entropy and information





Panzeri et al 2007





Timme et al 2018

(see reference list for details on methods)

Outline

- Part 1: motivation – why information theory?
- Part 2: Shannon's idea – information, compression and surprise
- Part 3: Efficient coding in neural systems
- **Demo: information estimation and optimal gain control in a model neuron**

Summary

- Information theory is a mathematical framework that brings together information, communication, randomness and predictability
- Entropy is a measure of information content that meets certain reasonable criteria (additivity for independent variables etc)
- It also turns out to be the minimum number of yes/no questions we need, on average, to identify the output of a random source
- Mutual information I(S;R) is the decrease in uncertainty in S when you learn about R, or the variability in R that is not due to the noise in S→R
- In neuroscience, popular applications of information theory are:
  - To measure the information content that neural activity carries about stimuli, behavior or other neural signals
  - As a foundation to a class of principles ("efficient coding" etc) that attempt to reach theoretical understanding of neural function by positing adaptation to the statistics of the natural environment (via evolution, development and learning) under constraints from physics and biology.

References

Textbooks:

- **MacKay. *Information Theory, Inference and Learning Algorithms* (2003).** A great introduction to information theory, where the connections to coding theory and statistical inference are explored in depth. The PDF is available online. Also check out the online lectures.
- **Rieke, Warland, R Van Stevenick, Bialek. *Spikes: exploring the neural code* (1997).** Introductory textbook on neural coding.
- Cover and Thomas. *Elements of Information Theory* (2nd ed 2006). Standard reference on information theory.
- Pierce. *An Introduction to Information Theory* (2nd ed 1980). A very enjoyable introduction to information theory written for a general audience. Still interesting even though quite old now.

Papers (good reviews):

- Panzeri et al. *Correcting for the Sampling Bias Problem in Spike Train Information Measures*. Journal of Neurophysiology 2007. On the finite sampling problem.
- Averbeck et al. *Neural correlations, population coding and computation.* Nature Reviews Neuroscience 2006. On the interplay of cross-cell correlations and information content in a neural population.
- Quian Quiroga and Panzeri. *Extracting information from neuronal populations: information theory and decoding approaches.* Nature Reviews Neuroscience 2009. Overview and comparison of applications of information theory and decoding approaches in neuroscience.
- Dimitrov et al. *Information theory in neuroscience.* J Comp Neurosc 2011.
- Timme et al. *A Tutorial for Information Theory in Neuroscience.* eNeuro 2018.

Open source software:

- pyentropy (python) - github.com/robince/pyentropy
- dit: Discrete Information Theory (python) - docs.dit.io
- Neuroscience Information Theory Toolbox (matlab) - github.com/nmtimme/Neuroscience-Information-Theory-Toolbox

Thanks