Natural Image Statistics and Neural Representations

Michael Lewicki

Center for the Neural Basis of Cognition &
Department of Computer Science

Carnegie Mellon University

Outline

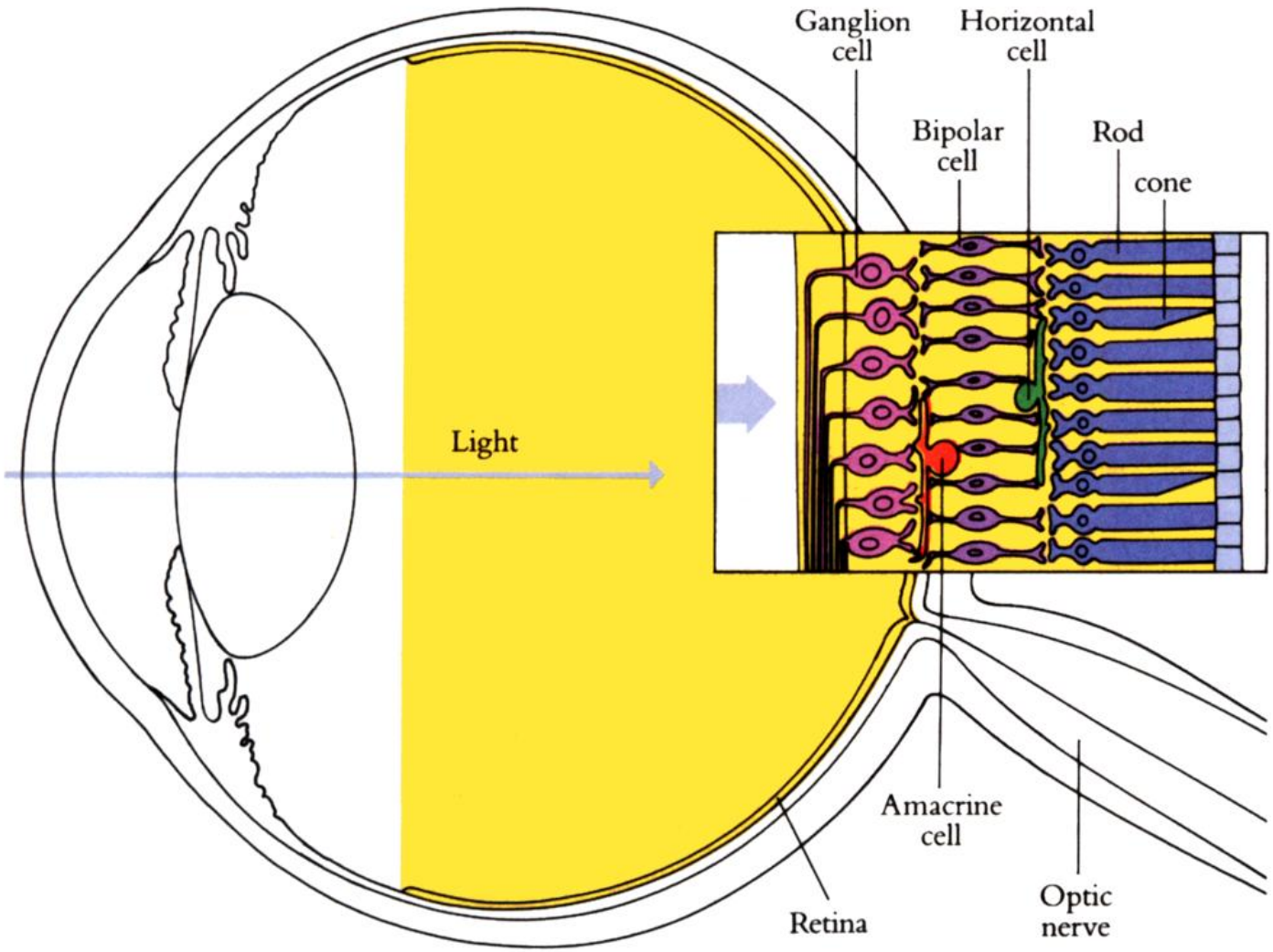
1. Information theory review, coding of 1D signals

2. Sparse coding, ICA, coding of natural images and motion

3. Representing images with noisy neural populations

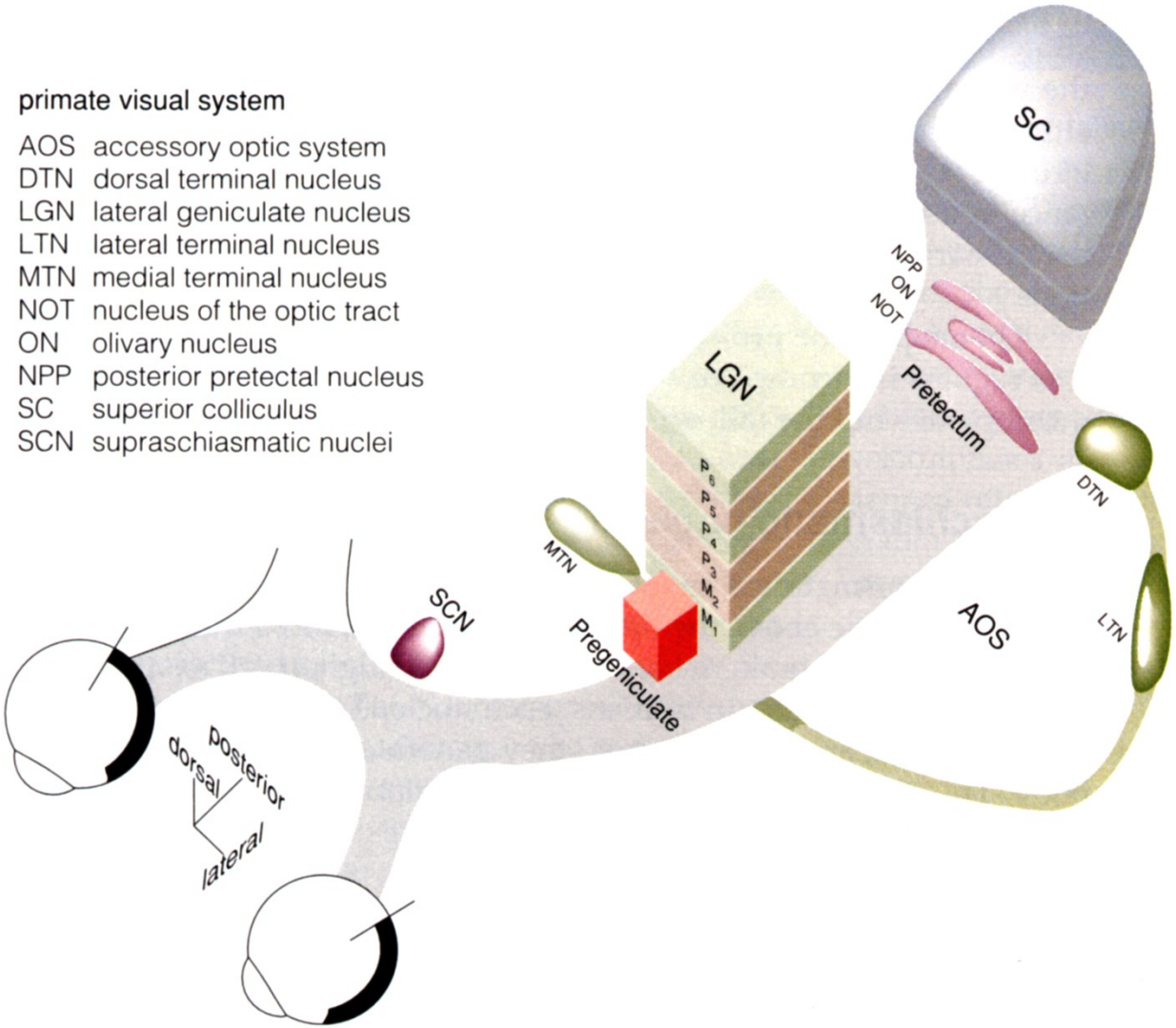
4. Learning higher-order image structure

Visual Coding



- What are the computational problems of visual coding?
- What signal should be sent out of the retina?
- How do we approach this theoretically?

After the retina...



- at least 23 distinct neural pathways out of the retina
- some receive from a single type of retinal cell, some from many, one eye, both...
- there is no simple function division

Why is it like this?

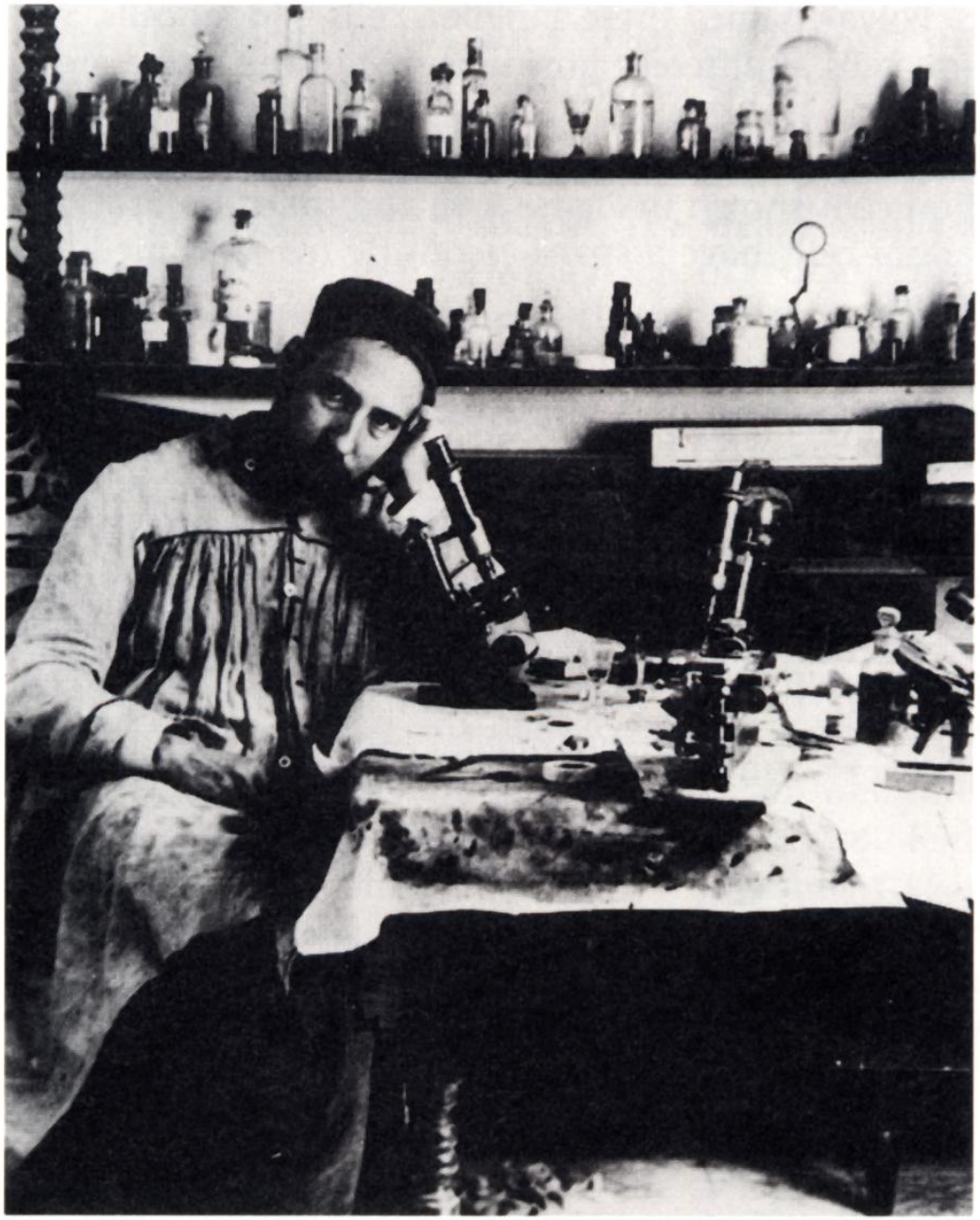
Evolutionary viewpoint:

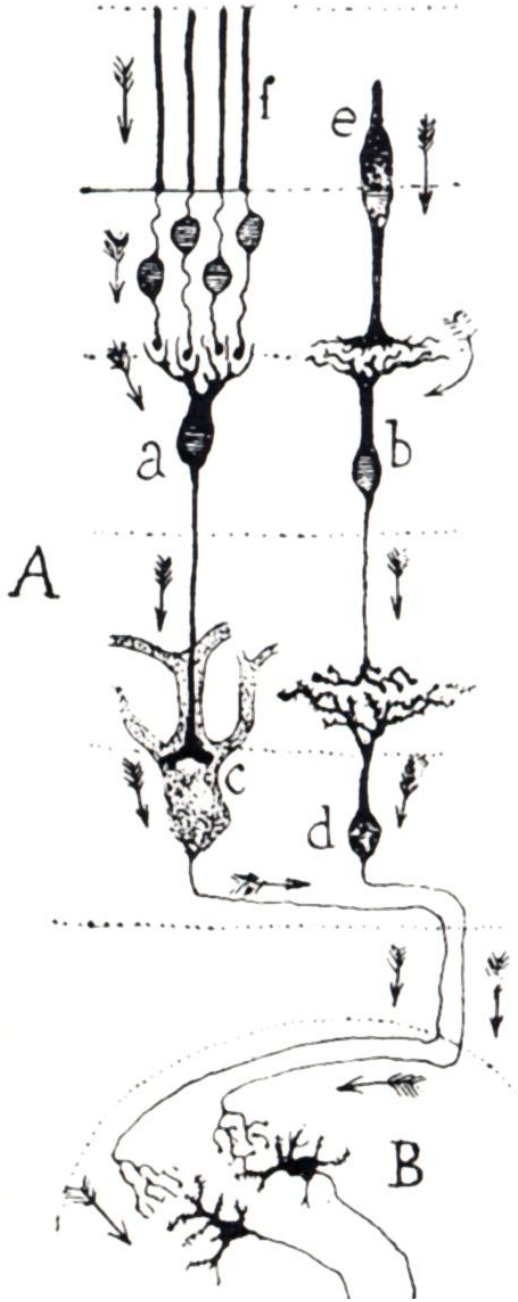
- success depends on whole organism and cooperation of areas and cell types
- there is no opportunity to "redesign", functions simply pile up
  - "layers and layers of goo"
  - "not engineering, but tinkering"
- there are few "clean" functional divisions, i.e. there are not distinct channels for color or motion

Types of optical systems

- **Suprachiasmatic nucleus**: generate the circadian rythm
- **Accessory optic system**: helps stabilize retinal image during head movement
- **Superior colliculus**: integrates visual and auditory information together with head movements, directs eyes to regions of interest
- **Pretectum**: plays role in adjusting size of pupil to changes in light intensity, and in tracking large moving objects
- **Pregeniculate**: function unknown, but cells are responsive to ambient light level
- **lateral geniculate**: main "relay" to visual cortex; contains 6 distinct layers, each with 2 sublayers. Organization is very complex and cells have a wide range of sensitivities including contrast, color, and motion.

Where is this headed?





A theoretical approach

- Look at the system from a function perspective:
  What problems does it need to solve?

- abstract from the details, make predictions from theoretical principles

- You can only have data after you do your theory.

- Models are bottom-up, theories are top-down.

- *What are the relevant principles?*

Information theory: a short introduction

Entropy:

- measure of irreducible signal complexity
- lower bound on how much a signal can be compressed without loss

Information of symbol w:

$$I(w) \equiv -\log_2 P(w)$$

For a random variable X, with probability $P(x)$, the entropy is the average amount of information obtained by observing x:

$$H(X) = \sum_x P(x)I(x) = -\sum_x P(x)\log_2 P(x)$$

- H only depends on the probabilty, not value
- Gives lower bound on average bits per code word.

Average coding cost for a message of length L (assuming independence) is

$$LH(X) \text{bits.}$$

Example

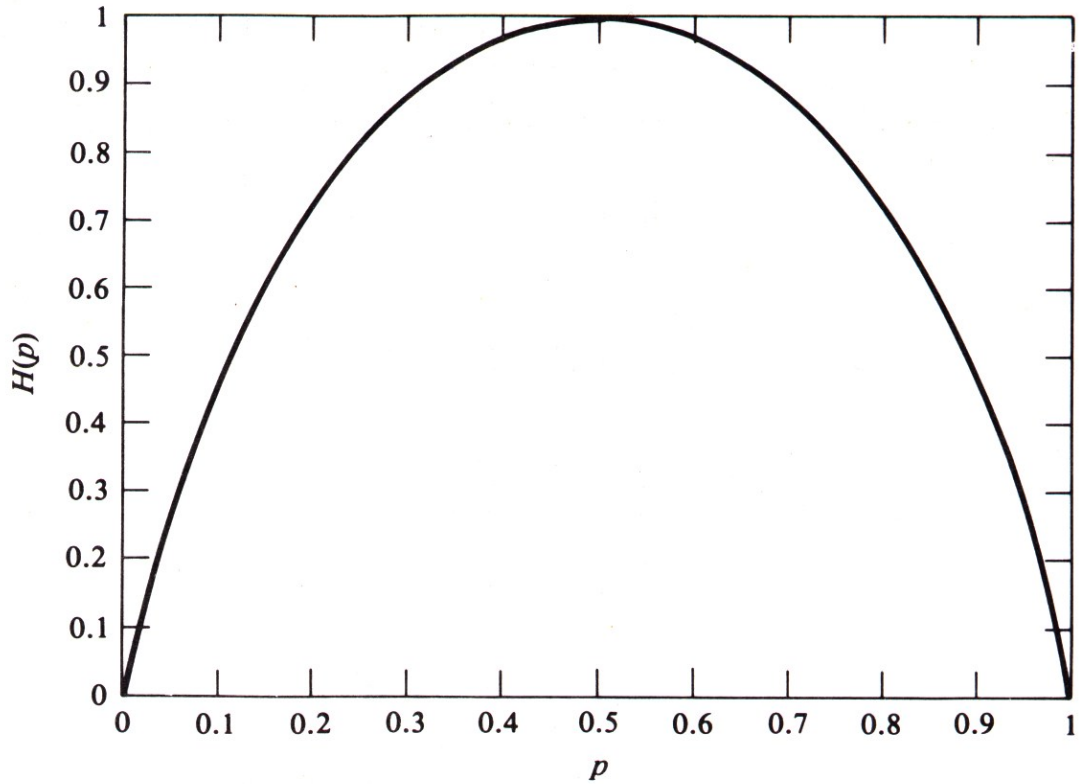


Figure 2.1. $H(p)$ versus p.

A single random variable X with $X = 1$ with probability p and $X = 0$ with probability $1 - p$. Note that $H(p)$ is 1 bit when $p = 1/2$.

Capacity

Capacity is the maximum amount of information per symbol:

$$C = \log_2 N$$

Maximum is when all N symbols have equal probability.

- English: $C = \log_2 27 = 4.73$ bits/letter
- Image: $8 \times 256 \times 256$ for 8 bit 256^2 image.

Actual entropy, i.e. the irreducible part, is much less.
Why?

Redundancy

Redundancy is a measure of (in)efficiency or actual entropy relative to capacity:

$$\mathcal{R} = 1 - H(x)/C$$

Capacity is maximum when

- code words (symbols) have equal frequency
- no inter-symbol redundancy

Examples

- English: letter probs not equal, letters not indep.
- Images: pixel value probs not equal, pixels not indep.

Example of symbols in english: A-Z and space

1. *Zero-order approximation.* (The symbols are independent and equiprobable.)

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ

   FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. *First-order approximation.* (The symbols are independent. Frequency of letters matches English text.)

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI

   ALHENHTTPA OOBTTVA NAH BRL

3. *Second-order approximation.* (The frequency of pairs of letters matches English text.)

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY

   ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO

   TIZIN ANDY TOBE SEACE CTISBE

4. *Third-order approximation.* (The frequency of triplets of letters matches English text.)

   IN NO IST LAT WHEY CRATICT FROURE BERS GROCID

   PONDENOME OF DEMONSTURES OF THE REPTAGIN IS

   REGOACTIONA OF CRE

A fourth order approximation

5. *Fourth-order approximation.* (The frequency of quadruplets of letters matches English text. Each letter depends on the previous three letters. This sentence is from Lucky's book, *Silicon Dreams* [183].)

   THE GENERATED JOB PROVIDUAL BETTER TRAND THE

   DISPLAYED CODE, ABOVERY UPONDULTS WELL THE

   CODERST IN THESTICAL IT DO HOCK BOTHE MERG.

   (INSTATES CONS ERATION. NEVER ANY OF PUBLE AND TO

   THEORY. EVENTIAL CALLEGAND TO ELAST BENERATED IN

   WITH PIES AS IS WITH THE)

   Instead of continuing with the letter models, we jump to word models.

Note that as the order is increased:

- entropy decreases: $H_0 = 4.76$ bits, $H_1 = 4.03$ bits, and $H_4 = 2.8$ bits/char
- variables, i.e. $P(c_i|c_{i-1}, c_{i-2}, \ldots, c_{i-k})$, specify more specific structure
- generated samples look more like real English

This is an example of the relationship between efficient coding and representation of signal structure.

The same model can also be applied to words

6. *First-order word model.* (The words are chosen independently but with frequencies as in English.)

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME

   CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO

   OF TO EXPERT GRAY COME TO FURNISHES THE LINE

   MESSAGE HAD BE THESE.

7. *Second-order word model.* (The word transition probabilities match English text.)

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH

   WRITER THAT THE CHARACTER OF THIS POINT IS

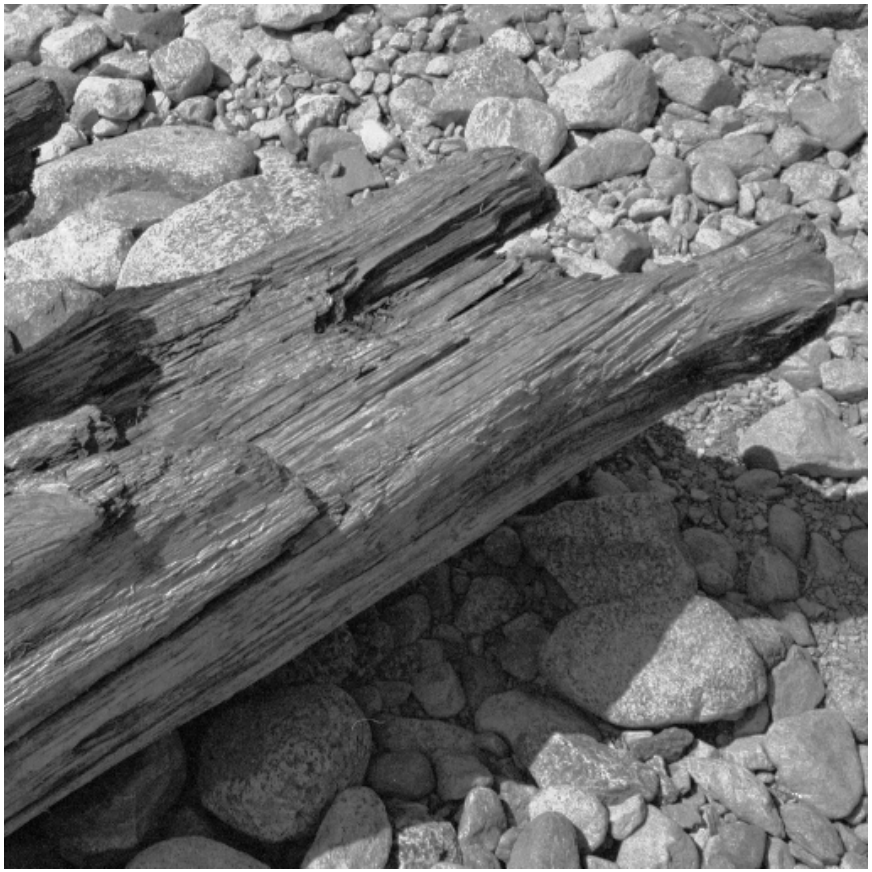
   THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE

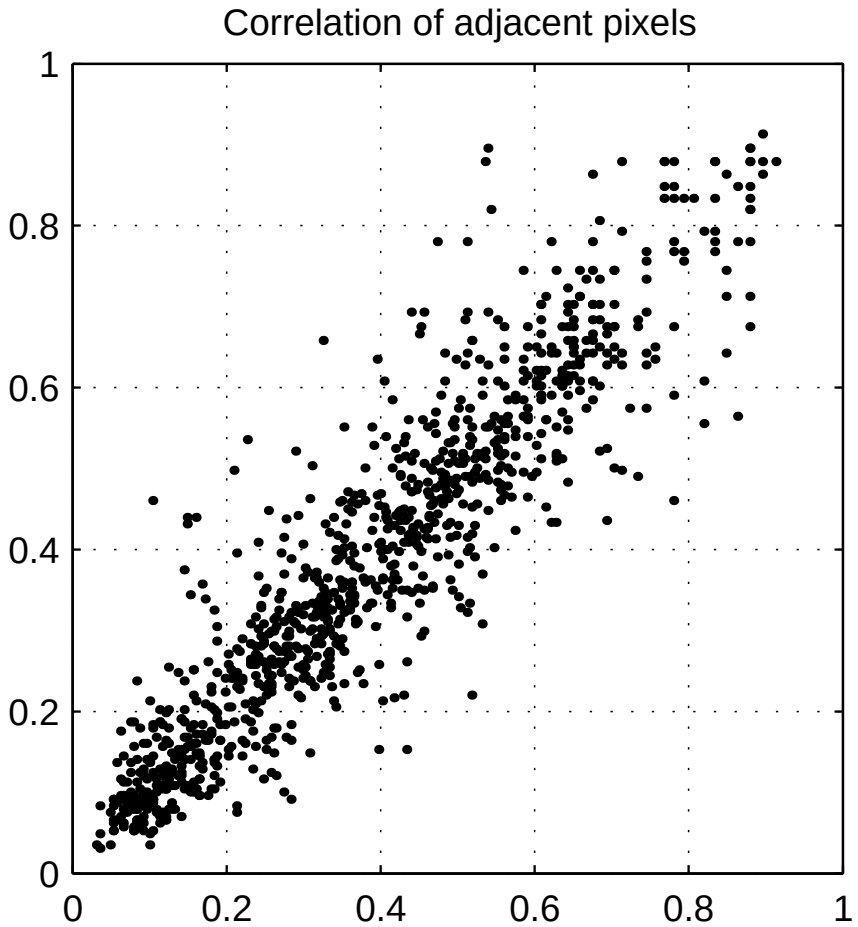
   TIME OF WHO EVER TOLD THE PROBLEM FOR AN

   UNEXPECTED

Specifying higher-order word models is problemlatic because the number of variables increases as N^k, where N is the number of words (e.g. 50,000 in English) and k is the order of the model.

A general approach to coding: redundancy reduction





Why reduce redundancy?

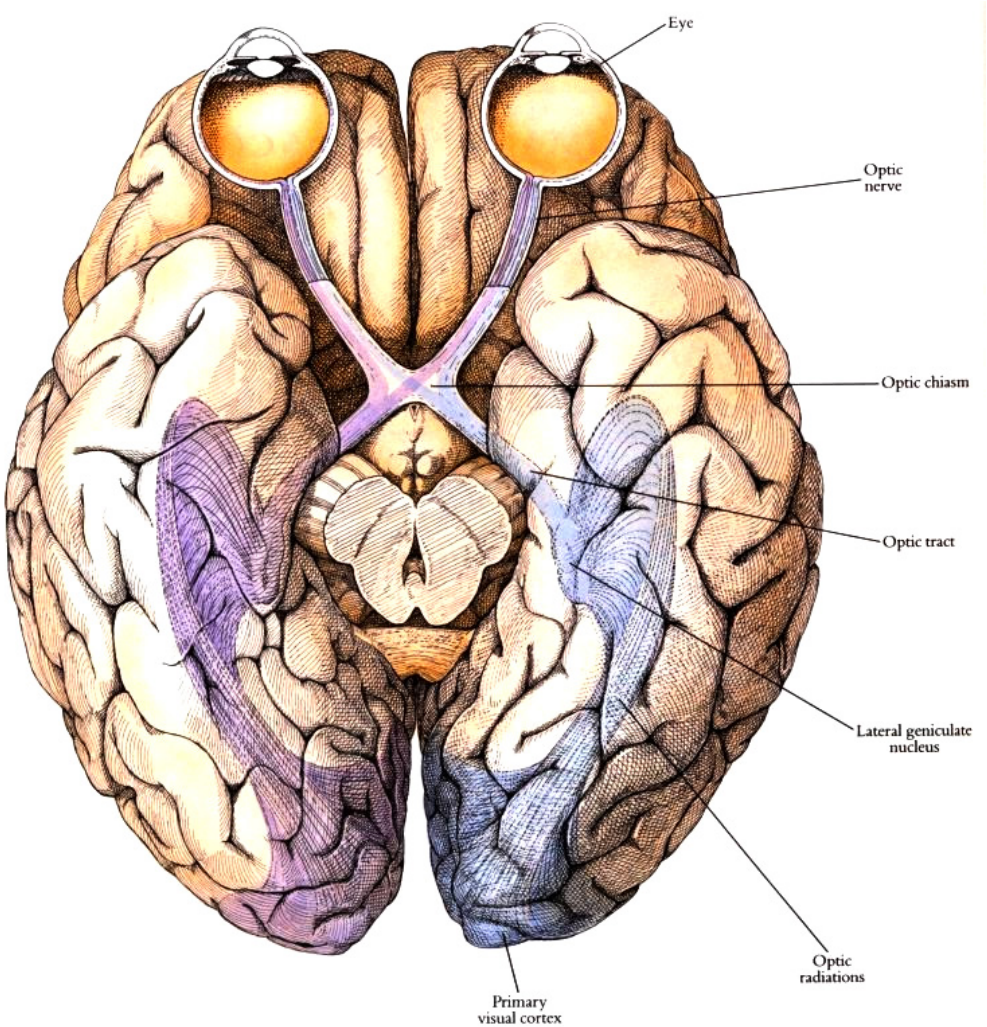
This is equivalent to efficient coding.

Why code efficiently?

Information bottleneck:

- restriction on information flow rate
  - channel capacity
  - computational bottleneck
  - $5 \times 10^6 \rightarrow 40 - 50$ bits/sec
- need even probabilities for associative learning
  - easy to calc joint probs for independent vars
- facilitate pattern recognition
  - independent features are more informative
  - better sensory codes could simply further processing

The bottleneck in vision



- Eyes must move ⇒ small, thin "cord"
- 100 million photoreceptors → 1 million optic nerve fibers
- Fovea already provides a great reduction in amount of information
- How do we reliably transmit the important visual information?

A little more information theory

- $H(X)$ is a measure of how much information it takes on average to describe random variable X.

- If we know $p(X)$, we can calculate entropy or optimize the model for the data, but what if we don't know $p(X)$ and can only approximate it, e.g. with $q(X)$?

- How many bits does this inaccuracy cost us?

Relative Entropy

- The *relative entropy* $D(p||q)$ is a measure of the inefficiency of assuming distribution q when the true distribution is p.

- If we knew p we could construct code with average code word length $H(p)$.

- If we assume q, the best average code length we can achieve is $H(p) + D(p||q)$

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- $D(p||q) = 0 \iff p = q$

- This is also called the *Kullback Leibler divergence*

- It is not called a distance, becase it is not symmetric and does not satisfy the triangle inequality.

Information theoretic viewpoint

Use Shannon's source coding theorm.

$$\begin{aligned}\mathcal{L} = E[l(X)] &\geq \sum_x p(x) \log \frac{1}{q(x)} \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} \\ &= D_{KL}(p\|q) + H(p)\end{aligned}$$

D_{KL} is the *Kullback-Leibler* divergence.
If model density $q(x)$ equals true density $p(x)$ then $D_{KL} = 0$.
$\Rightarrow q(x)$ gives *lower bound* on average code length.

greater coding efficiency $\Leftrightarrow$ *more learned structure*

Principle

Good codes capture the statistical distribution of sensory patterns.

How do we descibe the distribution?

Contrast response function in the fly eye (Laughlin, 1981)

- fly LMC (large monopolar cells) – interneuron in compound eye
- output is graded potential

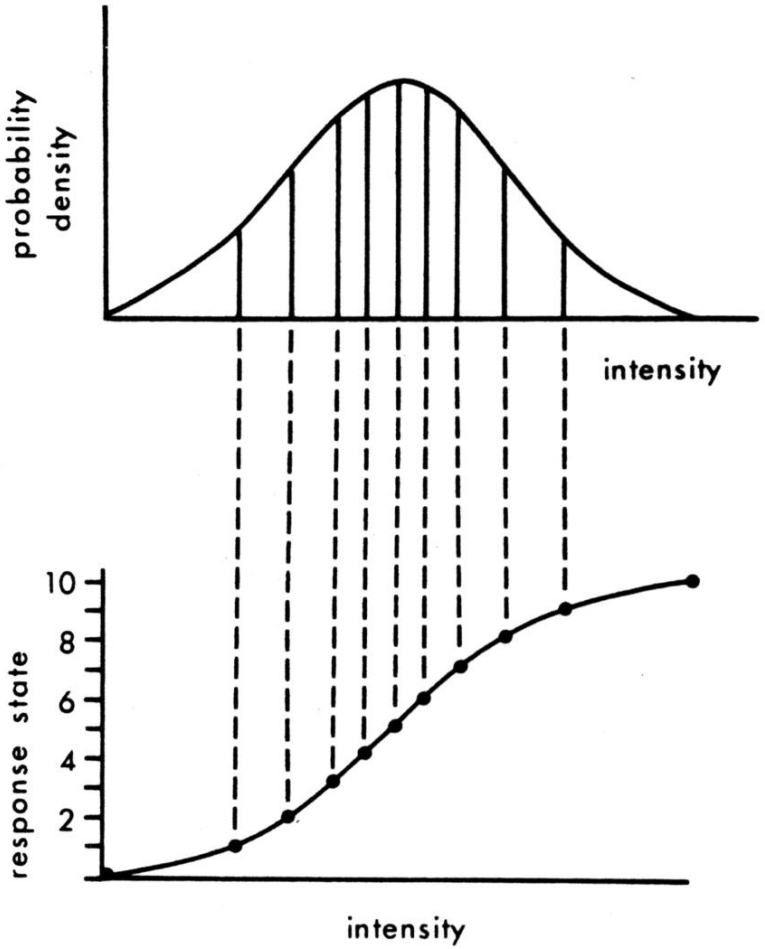
How to set sensitivity?

- too high $\Rightarrow$ response saturated

- too low $\Rightarrow$ range under utilized

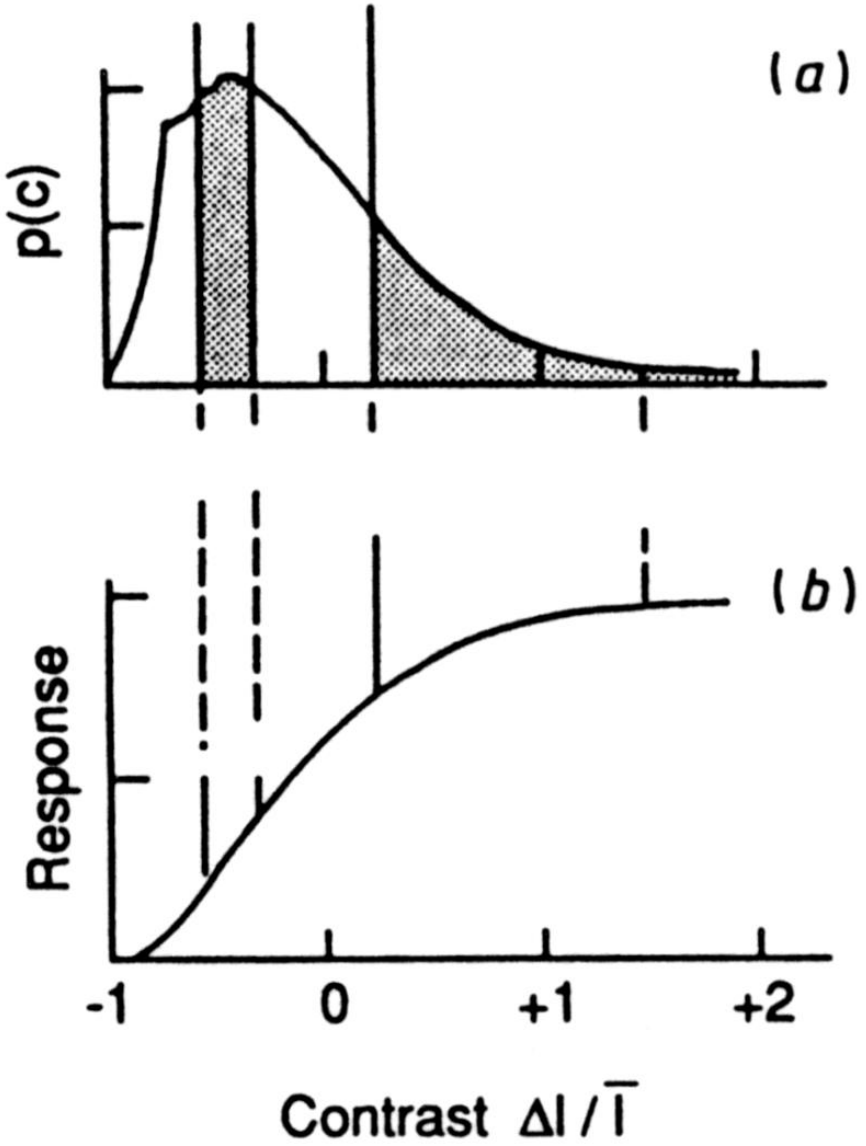
Idea: predict contrast reponse function using information theory.

Maximizing information transfer with limited channel capacity



- inputs follow given distribution
- transform so that output levels are used with equal frequency
- each response state has equal area ($\Rightarrow$ equal probability)
- continuum limit is cumulative pdf of input distribution

Another example with different statistics



Mathematical form is as cumulative probability. For $y = g(c)$

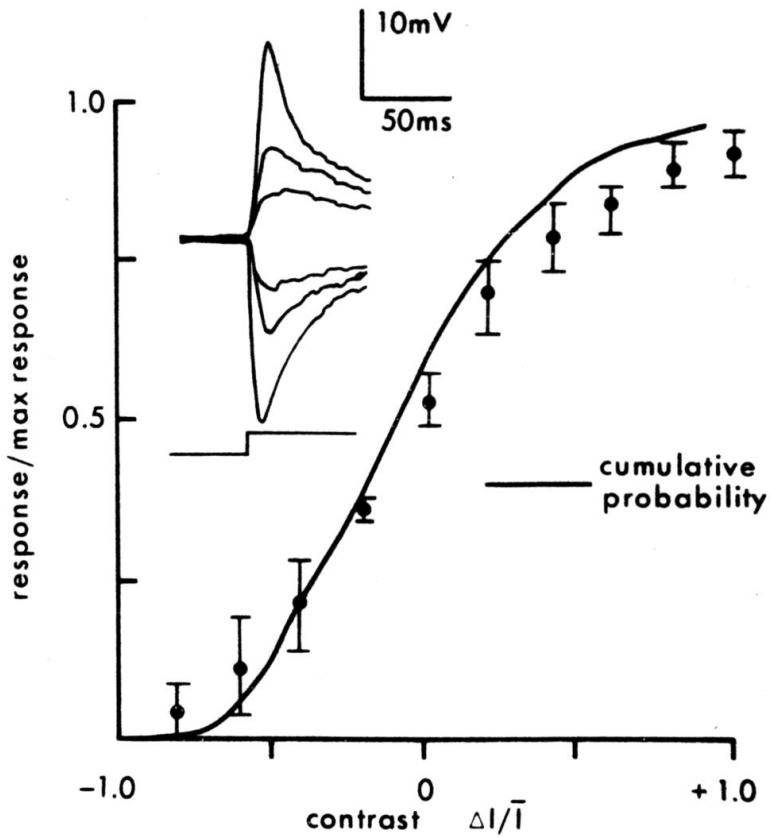
$$\frac{y}{y_{max}} = \int_{c_{min}}^{c} P(c')dc'$$

Testing the theory

Laughlin 1981:

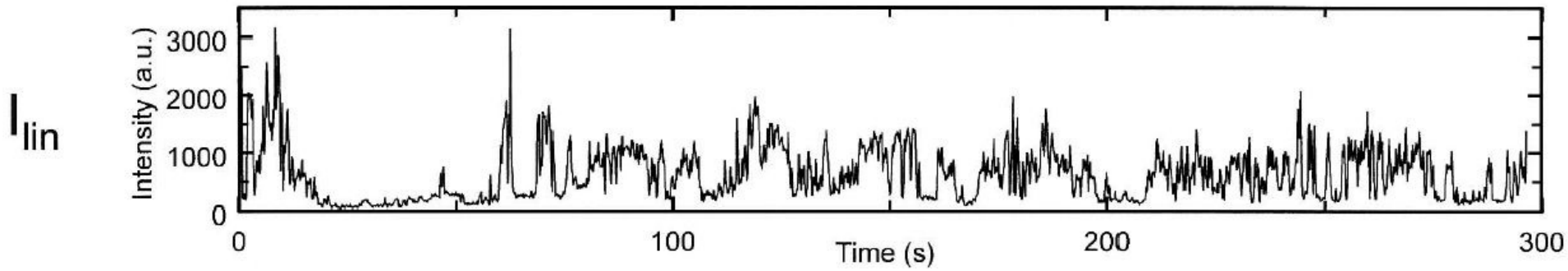
- collect natural scenes to get stimulus pdf
- 15,000 readings
- use linear scans: 10, 25, or 50°
- calc contrast within each scan: $\Delta I / \langle I \rangle$
- measure actual response of LMC to varying contrasts

$\Rightarrow$ fly LMC transmits information efficiently



Coding a natural intensity time series

van Hateren and Snippe (2001)

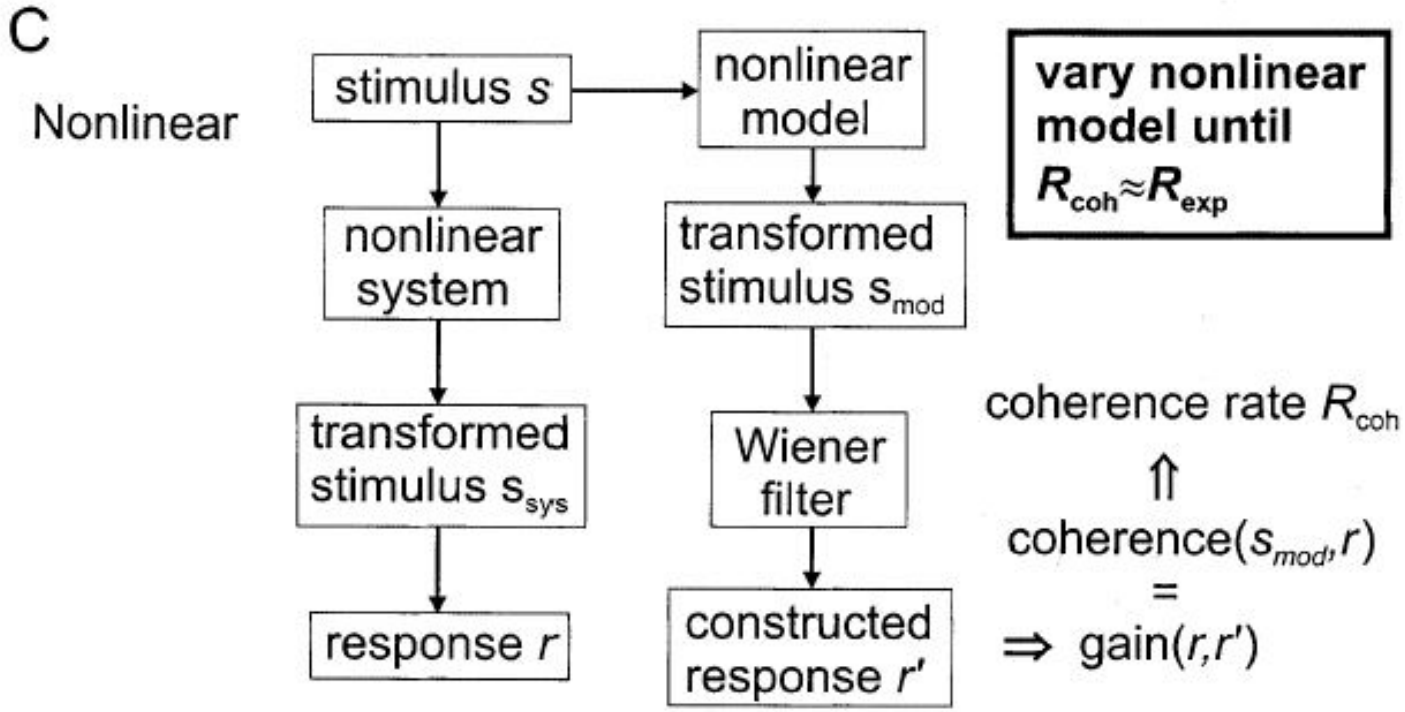


- recorded with a photometer, walking around outdoors
- dynamic range of intensity is much larger than that of photoreceptors
- large changes can occur on short time-scale
- Most, if not all, species can quickly adjust their gain to changing light levels.

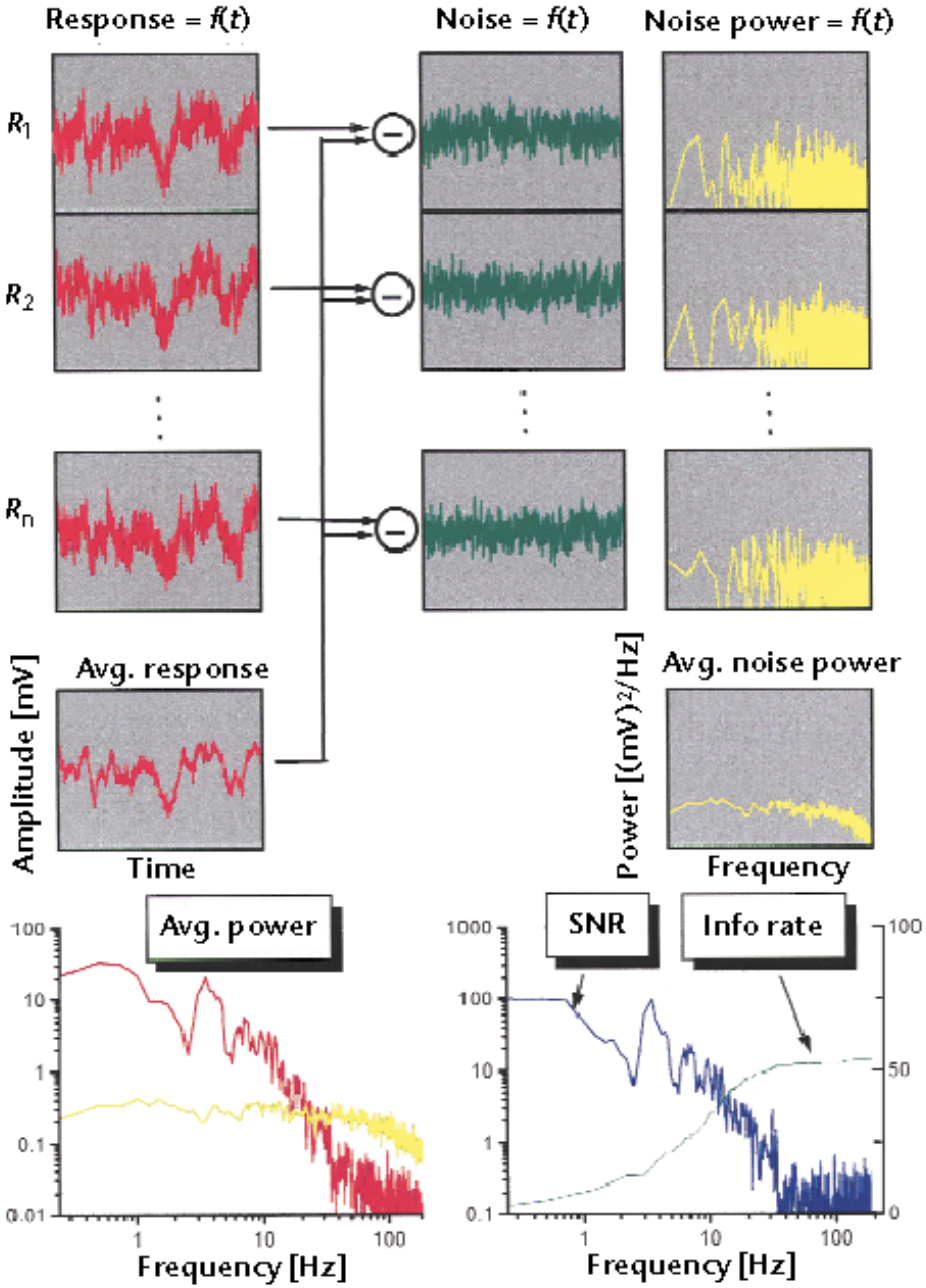
Questions:

- How should the signal be transformed?
- What gain control model should be used?
- How should the optimality the system (the fly in this case) be evaluated?

An evaluation method for non-linear encoding models



Measuring the capacity of the system

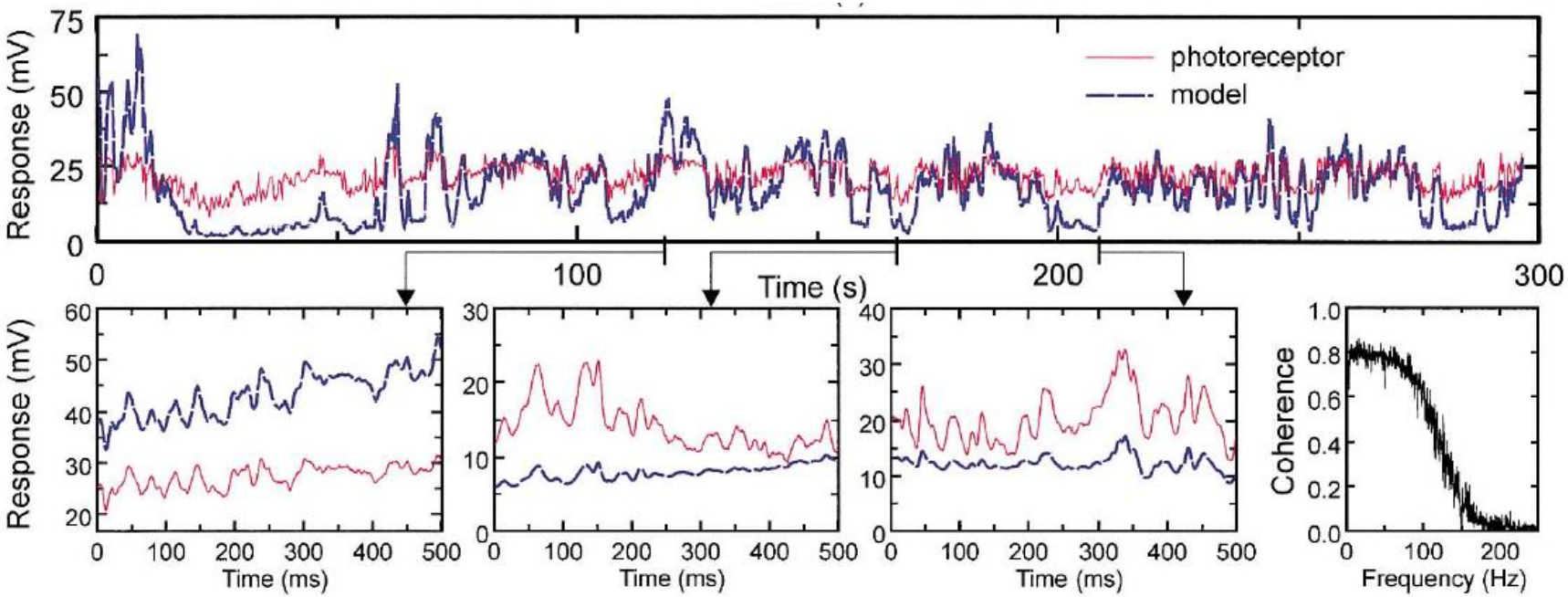


The noise is given by: $N = S - S_{est}$.
The signal to noise ratio is

$$SNR = \frac{\langle S_{est} {S_{est}}^* \rangle}{\langle N N^* \rangle}$$

A linear model

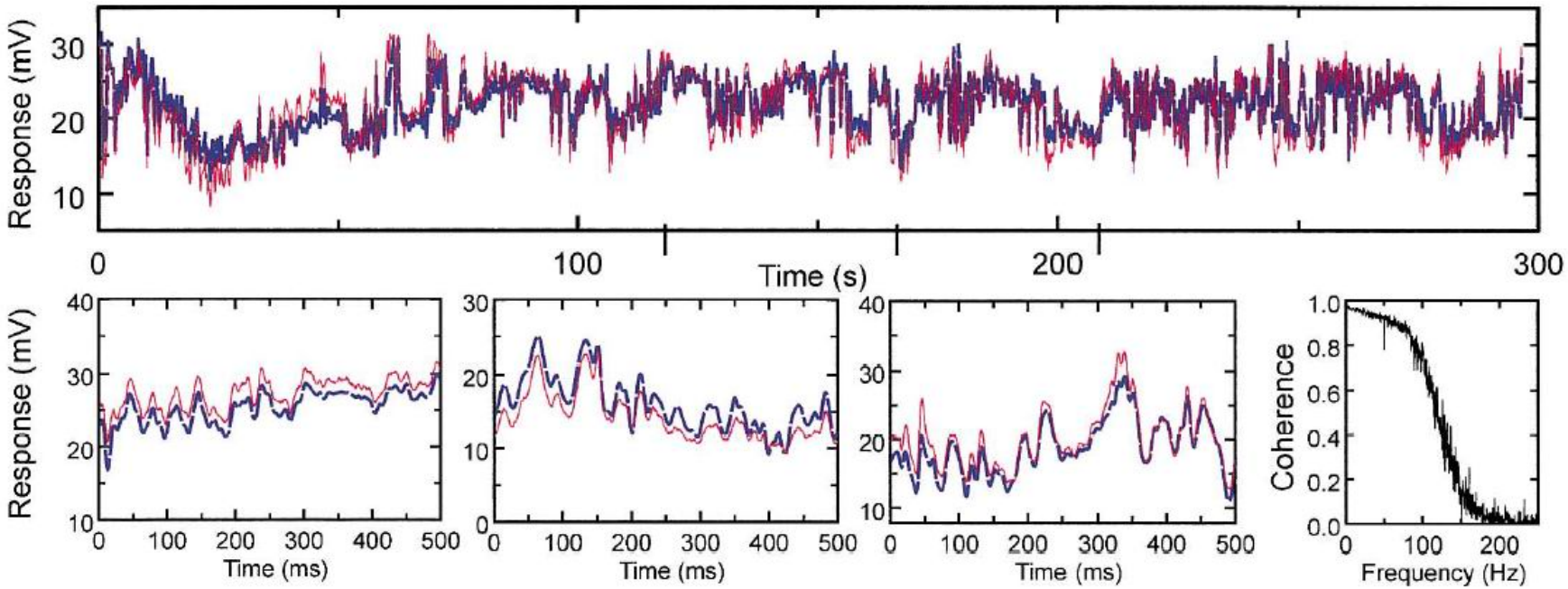
M_{lin}



- prediction of neural response of linear model is poor
- For linear model, coherence is sub-optimal at all frequencies.

Gain model with a static non-linearity: log

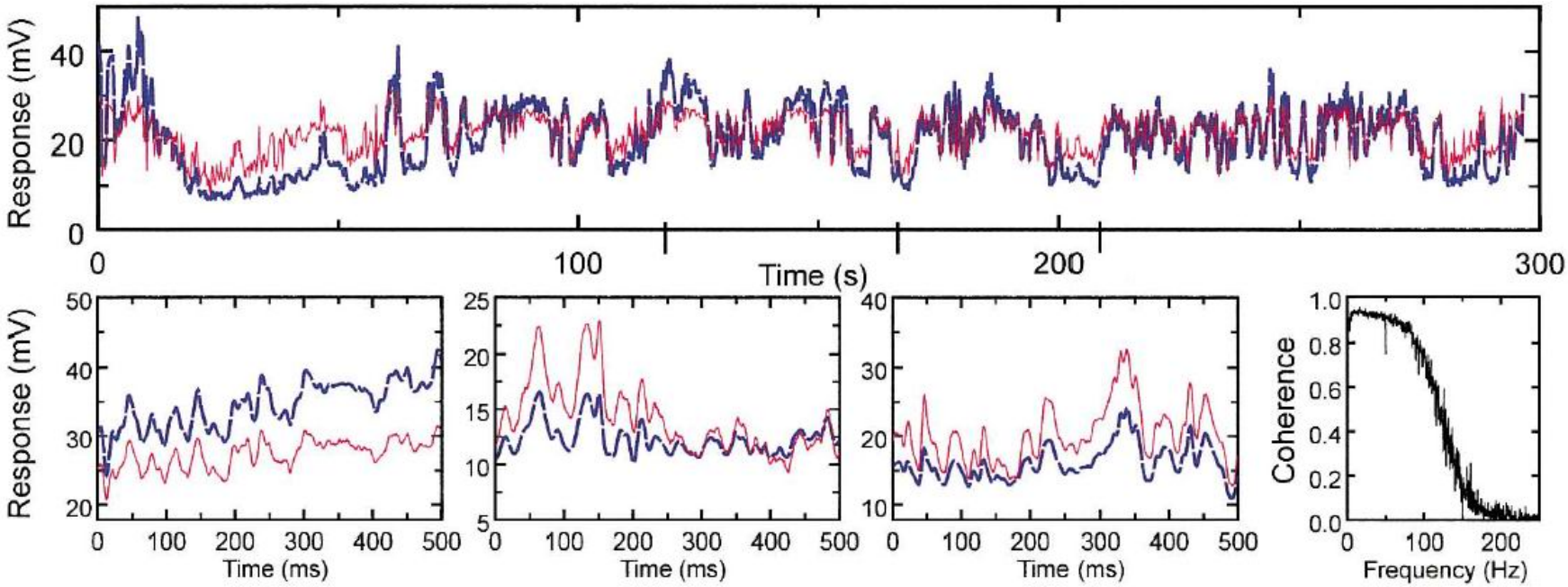
M_{log}



Coherence at low frequencies is improved, but coding is not perfect.

Gain model with a static non-linearity: sqrt

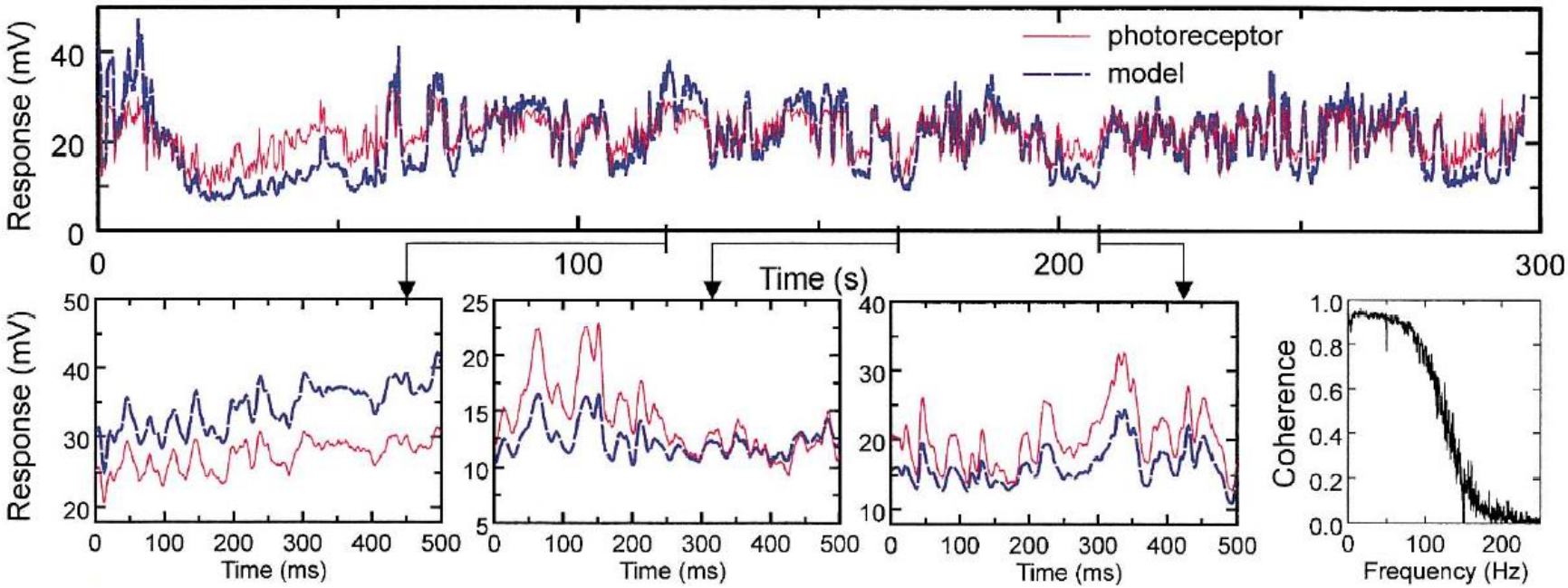
M_{sqrt}



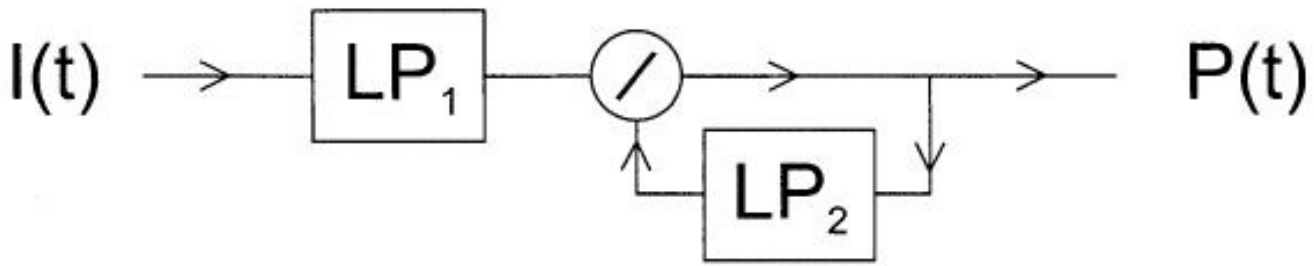
sqrt is slightly worse than log.

Gain model with a Dynamic non-linearity

M_D

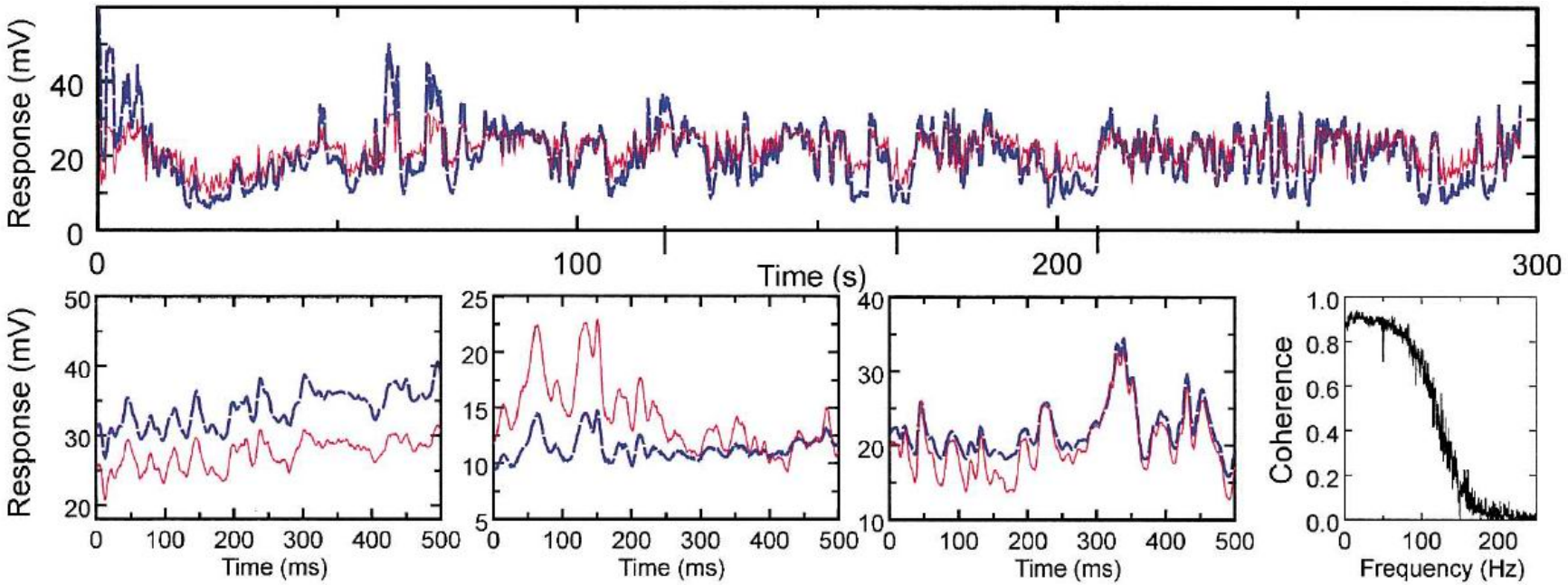


M_D: De Vries-Rose

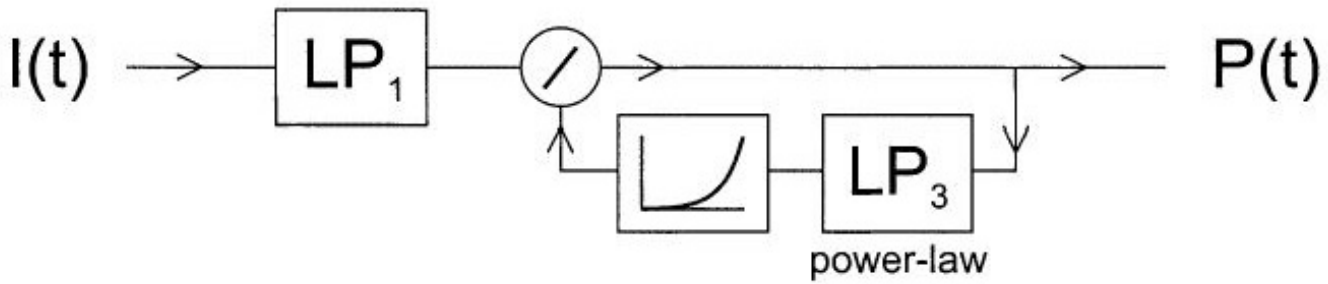


Gain model with a Dynamic non-linearity

M_W

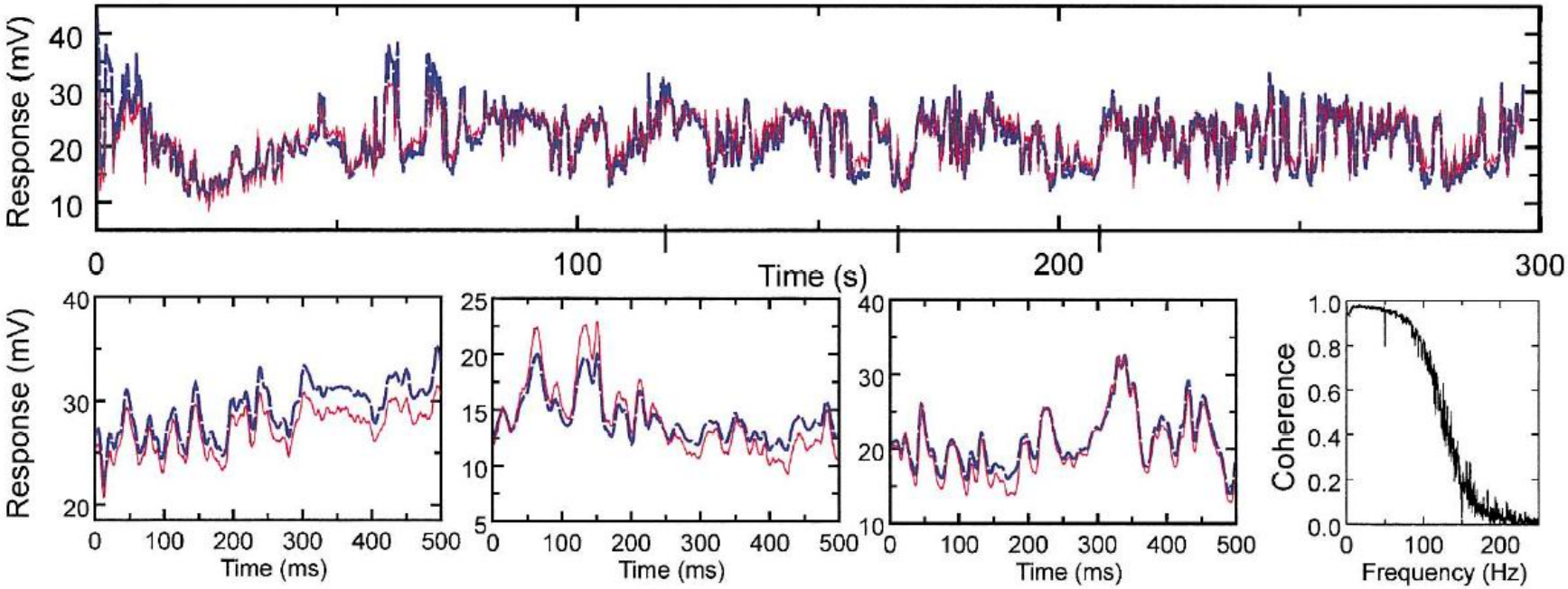


M_W: Weber

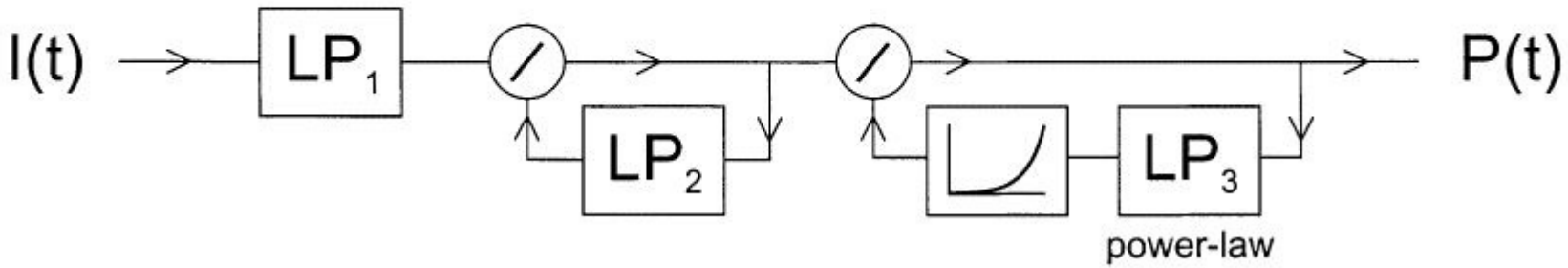


Gain model with a Dynamic non-linearity

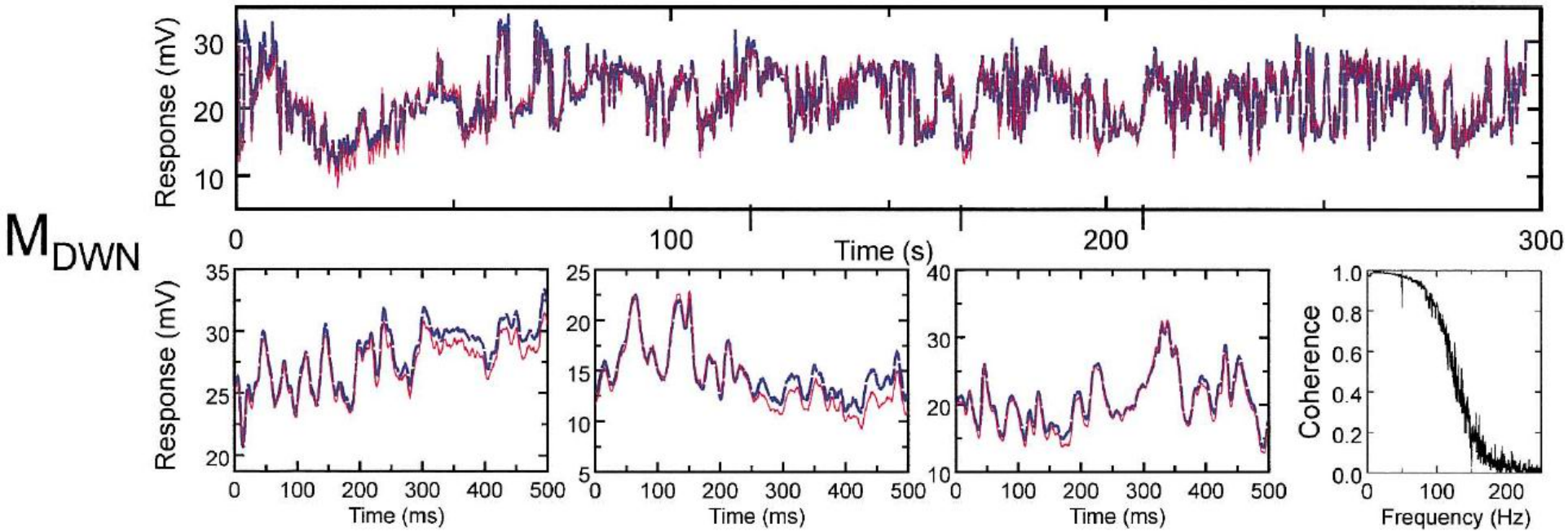
M_{DW}



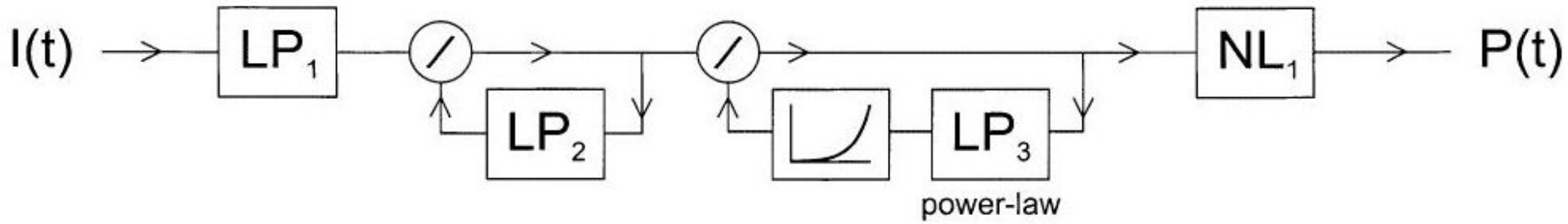
M_{DW}: De Vries-Rose & Weber



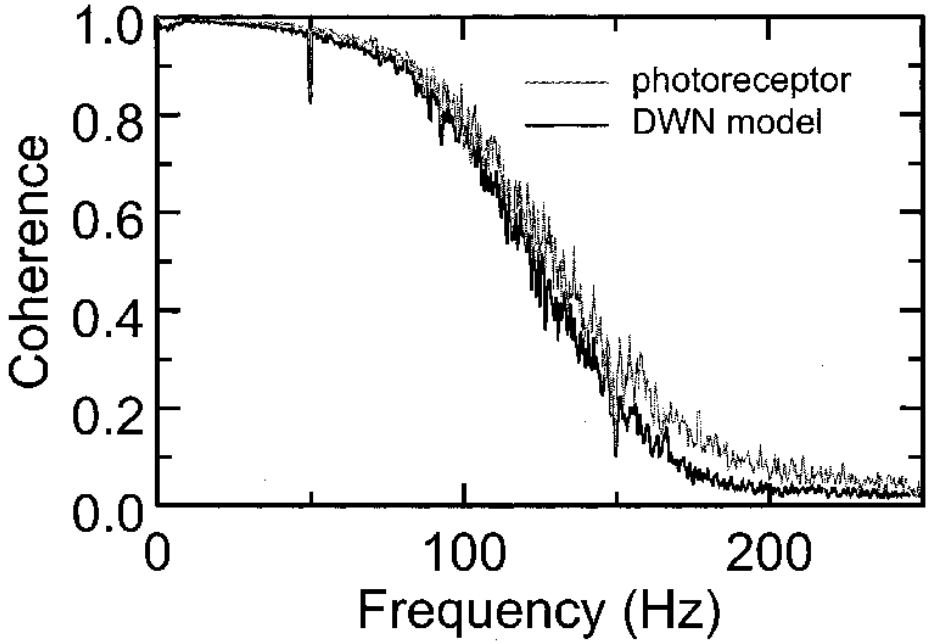
Best model requires several stages

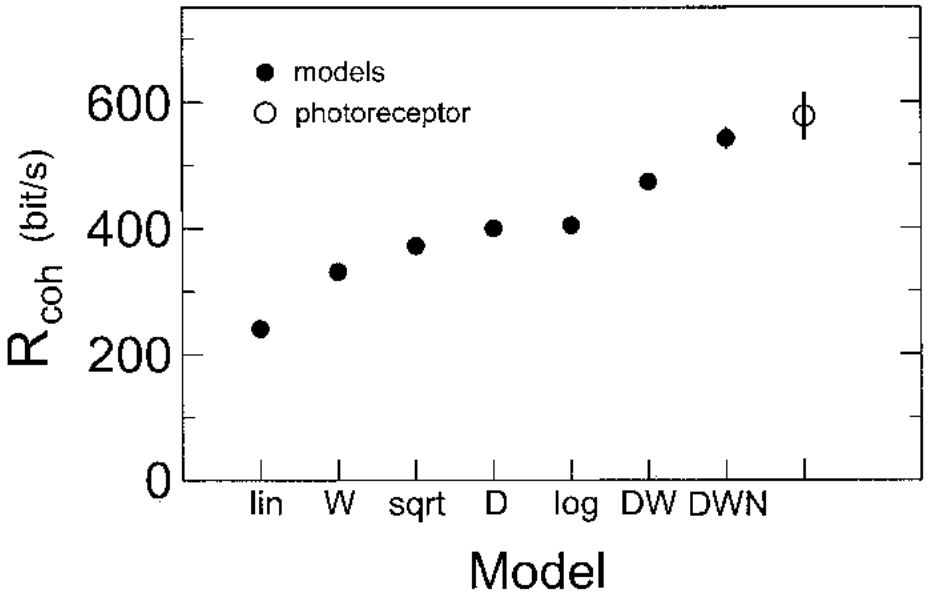


M_{DWN}: De Vries-Rose & Weber & NL



Coherence rates of the different models





Upper bound on capacity of fly photoreceptor is measured by estimating the variability in the response to the same stimulus.