Optimization Principles in Neural Coding and Computation

NIPS 2004 Tutorial
Monday, December 13, 2004

William Bialek

Joseph Henry Laboratories of Physics, and
Lewis-Sigler Institute for Integrative Genomics
Princeton University

http://www.princeton.edu/~wbialek/wbialek.html

what is at stake in a discussion of optimization?

the classic example: photon counting in vision

more examples, especially optimal coding in spike trains

do we have one principle or many?

Where vision begins (at night) ...

Rod photoreceptor cell in the retina

outer segment:
packed with ~ 1 billion
molecules of rhodopsin

outer segment membrane:
ion channels close in response to light,
electrical current is decreased

inner segment
membrane:
current is shaped
into a voltage signal

~ 25 microns
(1/1000 of an inch)

inner segment:
basic biology of
the cell

synaptic ending:
connect to other cells,
voltage causes release
of neurotransmitter

images from MJ Berry & FM Rieke

classical evidence for photon counting (~1940s):

probability of seeing = probability of > K photons absorbed, K = 5-7
photons spread over ~500 rod cells, so double hits negligible
-> each rod must give a reliable response to single photons
-> summation must be near ideal

why K=5 (1950s-60s)?

actually, sometimes K=2
think of detection as discrimination against "dark noise"
-> trading of sensitivity vs reliability, could get to K=1
-> effective dark noise ~1 event/minute/rod
-> lifetime of rhodopsin ~ 1000 years!

detecting single photon responses from rods (1970s)

single photon signal ~ 10 x (continuous background noise)
highly reproducible from photon to photon
observation of dark noise at level predicted from behavior
-> reliable macroscopic signal to single molecular event
-> "molecule multiplication" with small variance
-> "toad cooling" experiments

These observations pose challenges at many levels:

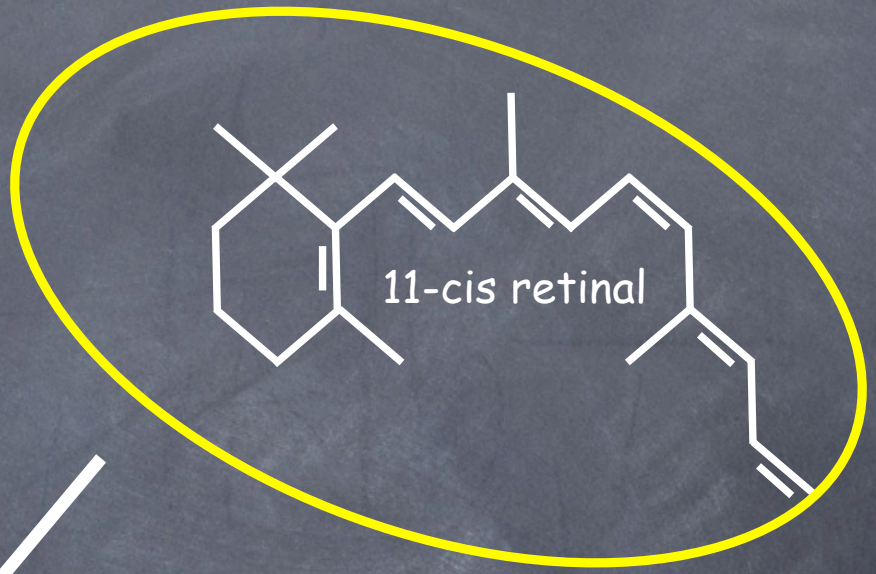
- Dynamics of the rhodopsin molecule itself
- Dynamics of the biochemical network for amplification of single molecular events
- Filtering and nonlinearity in the synaptic network of the retina
- Learning

In each case there are notions of optimization ...

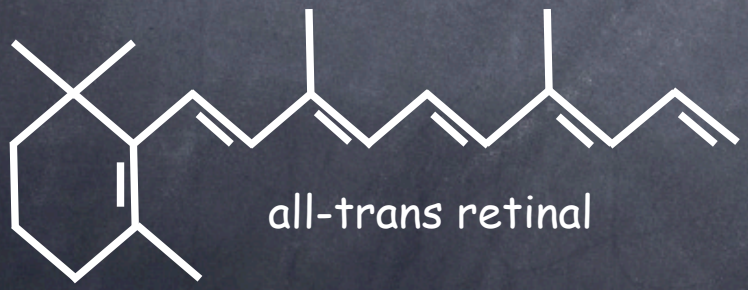
rhodopsin = organic pigment retinal covalently bound to (and enveloped by) the protein opsin

reminder of chemists' conventions:

- lines show bonds between carbon atoms
- hydrogen atoms are not shown explicitly
- single/double bond alternation is schematic



absorbing a photon triggers the 11-cis -> all-trans "isomerization"



the protein responds to isomerization with structural changes that start a cascade of biochemical events

opsin acts as an electronic state-selective catalyst

	retinal alone	rhodopsin
photon triggered isomerization	~ 1 nanosecond	~ 200 femtoseconds
thermally activated (spontaneous) isomerization	~ 1 year	~ 1000 years

~ 1 nanosecond: so slow that fluorescence is serious competition

~ 200 femtoseconds: so fast that it competes with loss of quantum coherence (!)

can't go faster than loss of coherence (Schrodinger's cat)

thermal isomerization = minimum dark noise level

~ 1 year -> ~ 1000 years reduces dark noise by more than 1000 times

everything else works so well that behavioral noise level ~ thermal isomerization

if rhodopsin had the properties of retinal, night vision would be nearly impossible

we don't really understand all these numbers *sigh*

Once the photon triggers a structural change in rhodopsin, the cell needs to detect this single molecular event and amplify it to a macroscopic level

macroscopic output = current from closing ion channels in the cell membrane

already a problem: single photon current ~ 1 picoamp close to typical current through just one open channel ... but we don't see the single channel noise

channel noise (which would be dominant) is suppressed by having the light channels flicker rapidly between closed and open states ...

the fixed and inescapable noise variance is spread over huge (> 1000 Hz) bandwidth, lowering density in relevant (~ 1 Hz) range

the output actually is a change in probability of being open for many channels

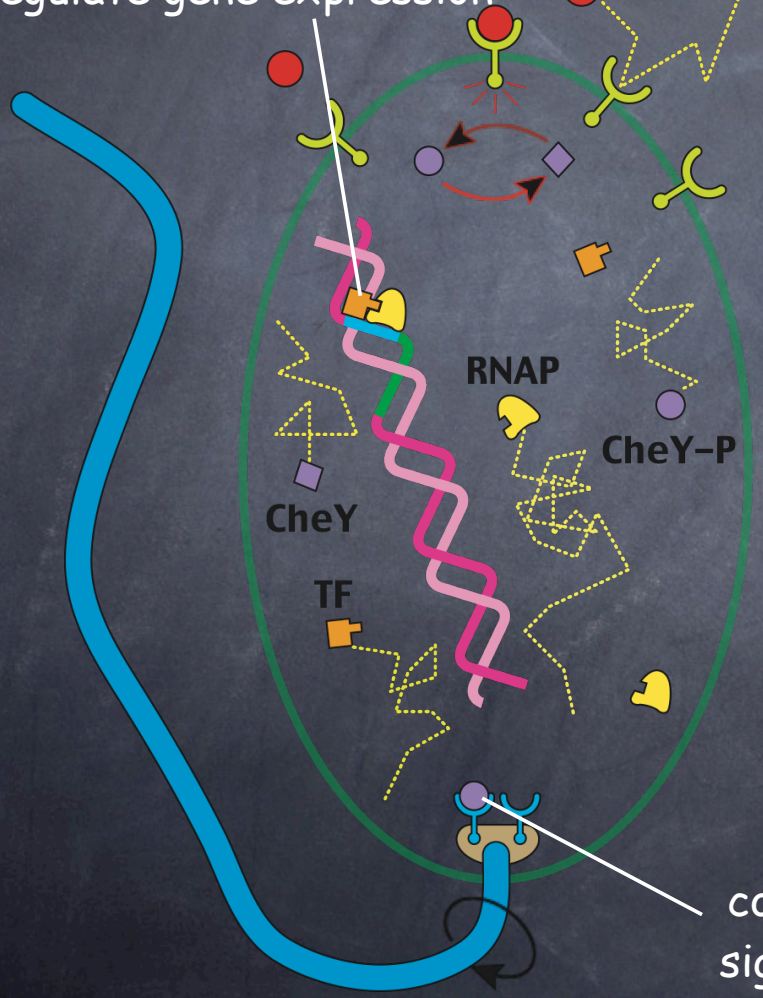
Many more "design features" in the biochemical amplifier

Also, problems of molecule counting are universal ...

Bacteria as molecule counters

count the interesting
molecules in the environment

count the molecules that
regulate gene expression



count the internal
signaling molecules

How accurately can they count?

Physical limits from:

- diffusion
- small size of receptor sites
- short time to average

E Coli perform close to these physical limits

Physics of chemoreception.
HC Berg & EM Purcell, Biophys J 20, 193-219 (1977).

Physical limits to biochemical signaling.
W Bialek & S Setayeshgar, arXiv.org/physics/0301001.

rod cells produce a current I(t) ...

although this is a model, you can measure everything (!)

$$I(t) = \sum_i I_0(t - t_i) + \text{noise}$$

responses to single photons at times t_i

responses to thermal isomerization events + continuous (~ Gaussian) background noise

what does the brain want to do with these currents?

"noise" obviously is irrelevant and should be suppressed

more subtly, even photon arrival times are not relevant in themselves

because statistics of photon arrivals are Poisson, the only things of possible relevance to behavior are functions of the photon rate r(t)

tempting to formulate statistical inference problem:
rod currents I(t) -> photon arrival rate r(t)

but this isn't right ... presumably brain isn't interested in r(t) = light intensity, but only in some features.
which features?

at low light intensities it doesn't matter because we have sufficient statistics:

P[r(t)|I(t)] depends only on a filtered version of the rod current
filter F(t) = matched filter for $I_0(t)$ against spectrum N(ω) of "noise"

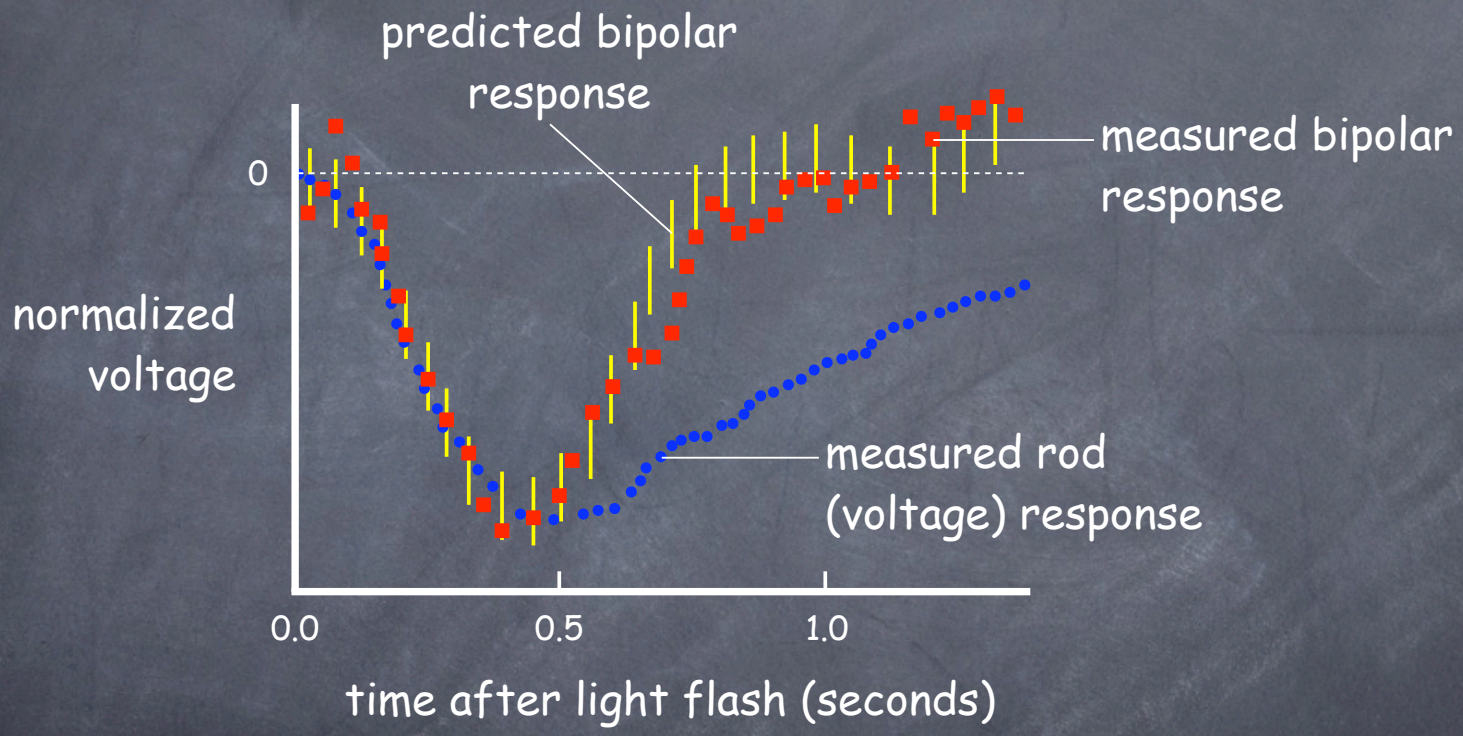
$$\tilde{F}(\omega) = \frac{\tilde{I}_0(\omega)}{N(\omega)}$$

again, these quantities are measured

this filter should be implemented at the first stage of visual processing

Temporal filtering in retinal bipolar cells:
Elements of an optimal computation?
W Bialek & WG Owen,
Biophys J 58, 1227-1233 (1990).

Testing the theory... no free parameters



Optimal filtering in the salamander retina.
F Rieke, WG Owen & W Bialek,
in Advances in Neural Information Processing 3,
R Lippman, J Moody & D Touretzky, eds, pp 377-383
(Morgan Kaufmann, San Mateo CA, 1991).

(remember, predictions come from measured rod signals and noise, not a model of synaptic mechanisms)

more to say about this synapse:

adding up signals from many rods not so easy ... SNR for one photon in one cell ~10, but if you look for 1 photon in >100 cells ...

need nonlinearity before summation
there is an optimal setting for the "threshold" of the nonlinearity

seems to be right!

Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity.
GD Field & F Rieke, Neuron 34, 773-785 (2002).

and much more about photon counting ...

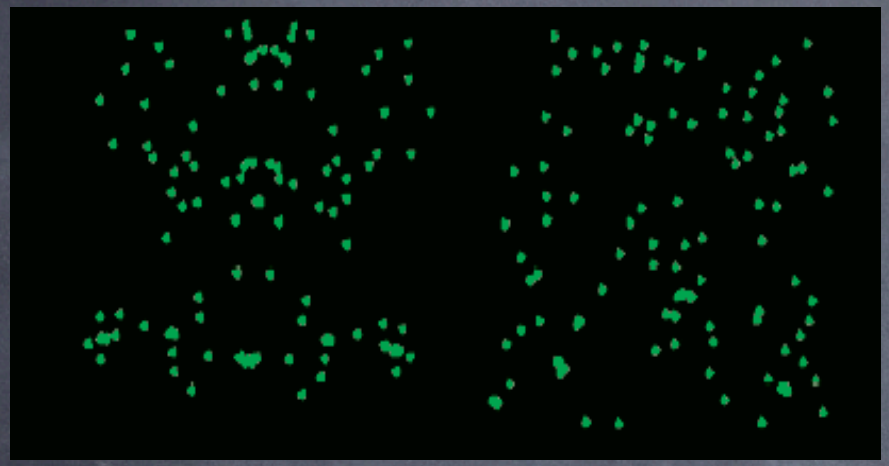
delays in retinal processing depend on intensity, so can generate illusions by mixing movement with intensity changes

our brain falls for these illusions (eg, Pulfrich effect), and so do toads in striking at dimly illuminated targets

delays so long in photon counting regime that toads would strike behind moving targets ... so they learn to compensate (!)

Visual performance of the toad (Bufo bufo) at low light levels: Retinal ganglion cell responses and prey-catching accuracy. AC Aho et al. J Comp Physiol A 172, 671-682 (1993).

Seeing the whole from (slightly) random parts



Symmetry is one of the "gestalt" percepts ... a property of the whole, not the parts of an object

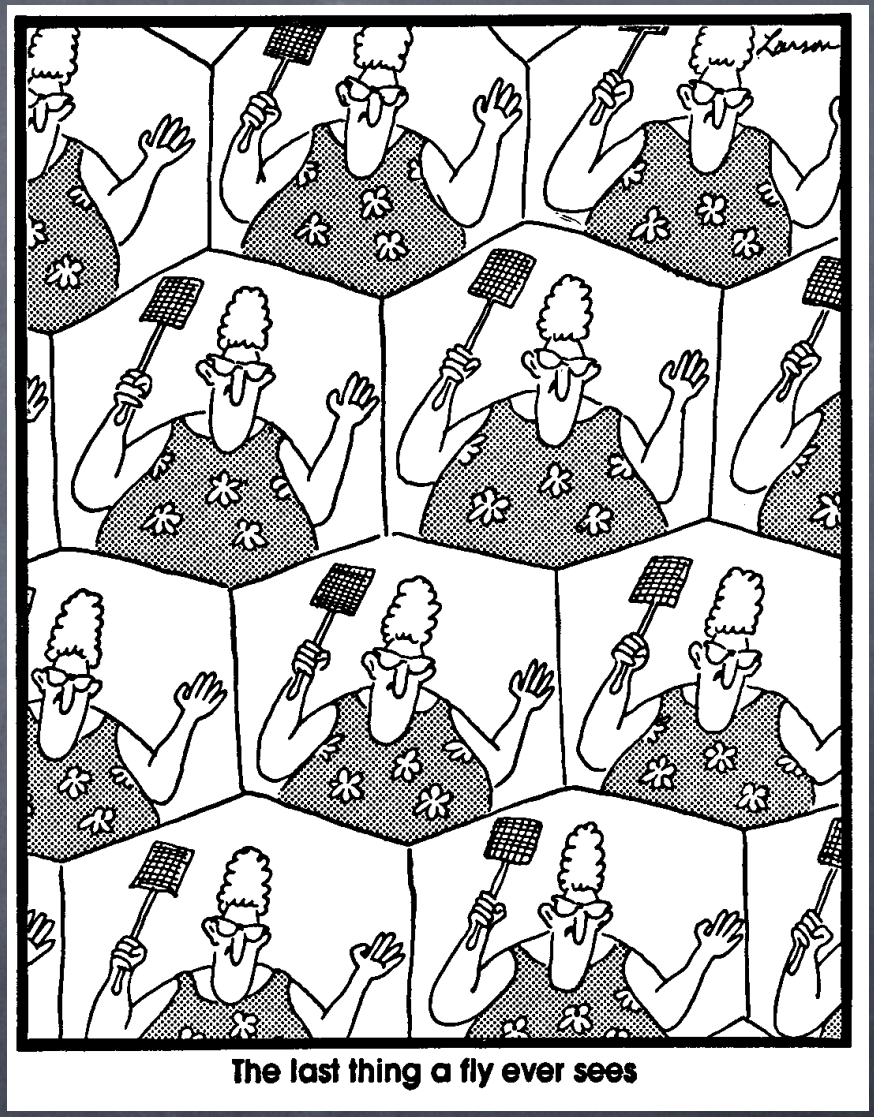
How well can we distinguish a real tendency toward symmetry from a random, statistical coincidence?

Almost as well as possible given the rules of probability

Related ideas in pitch perception ...

The absolute efficiency of perceptual decisions.
HB Barlow, Philosophical Transactions of the Royal Society of London Series B, 290, 71-82 (1980).

Vision, but in a different animal ...

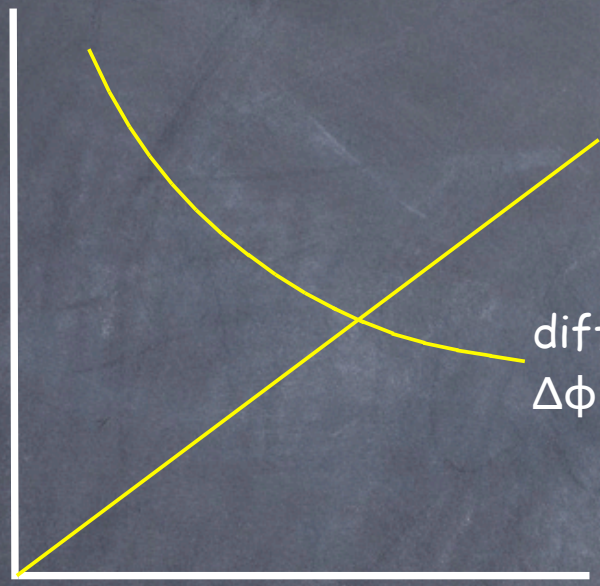


The last thing a fly ever sees

not as different as Mr Larson thinks

how big should we make the pixels of the compound eye?

angular resolution $\Delta\phi$



geometrical optics:
$\Delta\phi \sim d/R$
(R = radius of head)

diffraction:
$\Delta\phi \sim \lambda/d$

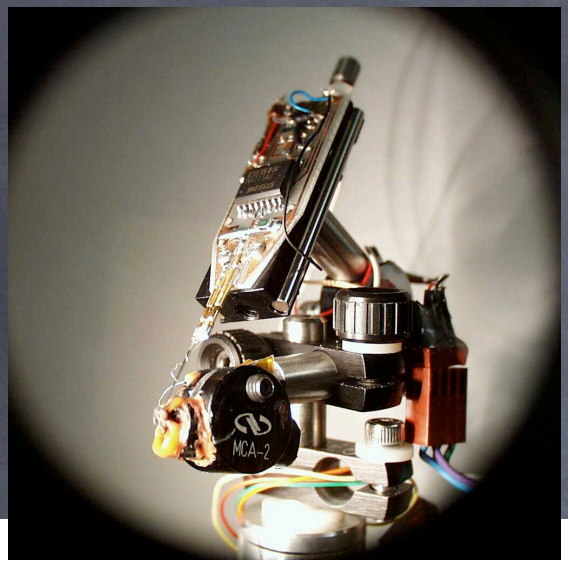
pixel size d

optimum: min $\Delta\phi$ at $d \sim (\lambda R)^{1/2}$

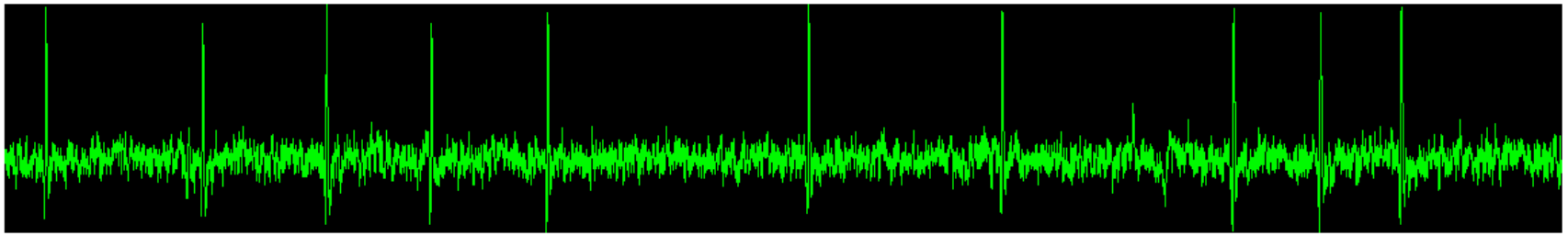
agrees with exp't on many insects with different R
not right when SNR is low ...
need to optimize information not resolution

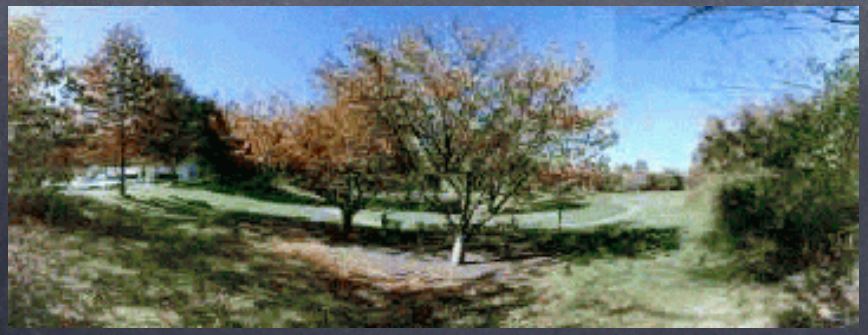
The size of ommatidia in compound eyes.
HB Barlow, J Exp Biol 29, 667-674 (1952).

see also the Feynman lectures!



place a small wire in the back of the fly's head
to "listen in" on the electrical signals from nerve cells
that respond to movement





Experiments from R de Ruyter van Steveninck & GD Lewen

The fly solves (at least) two problems:

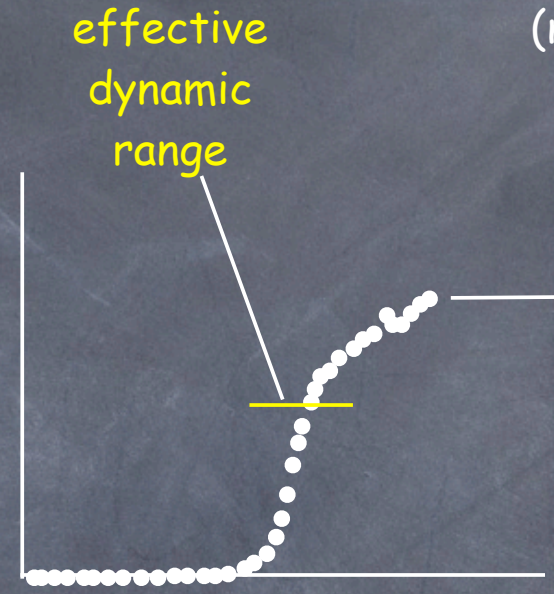
coding the trajectory of motion in sequences of spikes, and
computing motion from signals in the retina

optimization in computation:
1. do motion sensitive neurons allow discrimination and estimation with precision close to limits set by noise at the sensory input?
2. what function(al) of the inputs provides the best estimate?
3. can we analyze the response of these neurons to 'dissect' the computation and test the predictions from (2)?

optimization in coding:
1. do these neurons use most of their dynamic range for real signals, or irrelevant variability?
2. what are the symbols in the code?
3. what events do these symbols signify in the motion signal?
4. **is there an optimal choice of this mapping?**

What determines the structure of input/output relations?

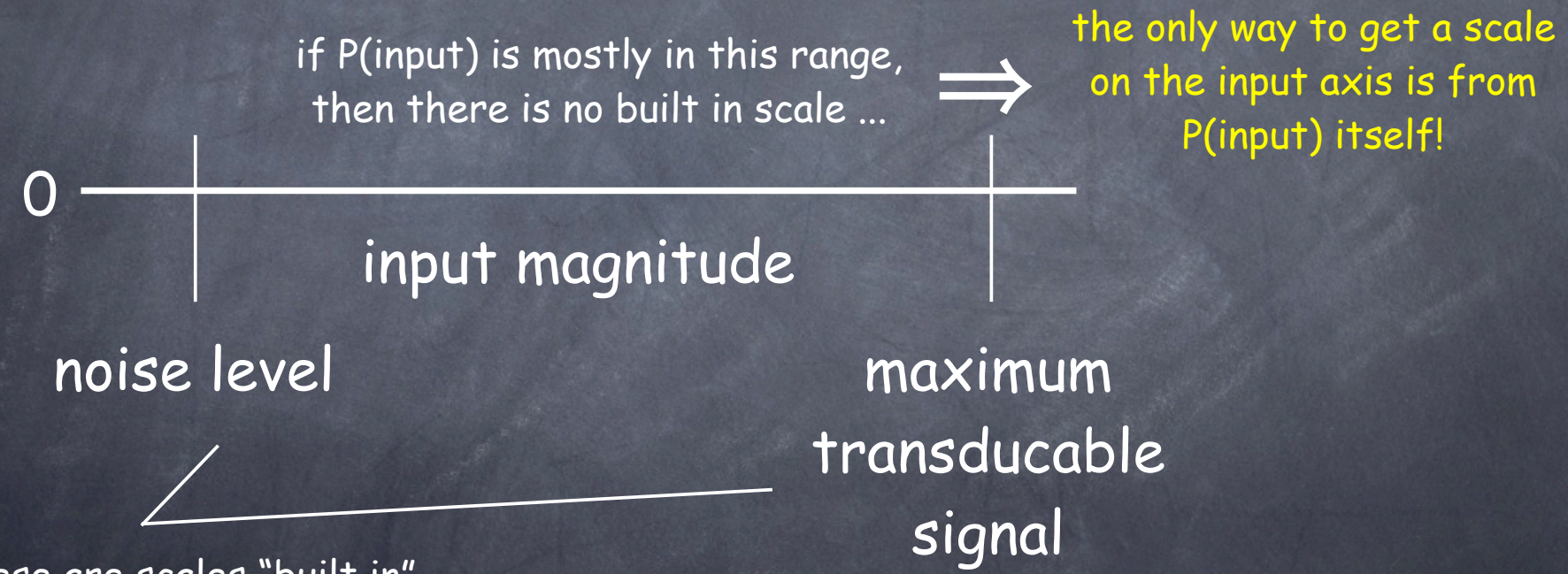
(not just a question about neurons!)



In particular, what sets the scale along the input axis?

Suppose we chose the input/output relation to maximize the information I(input; output) ...

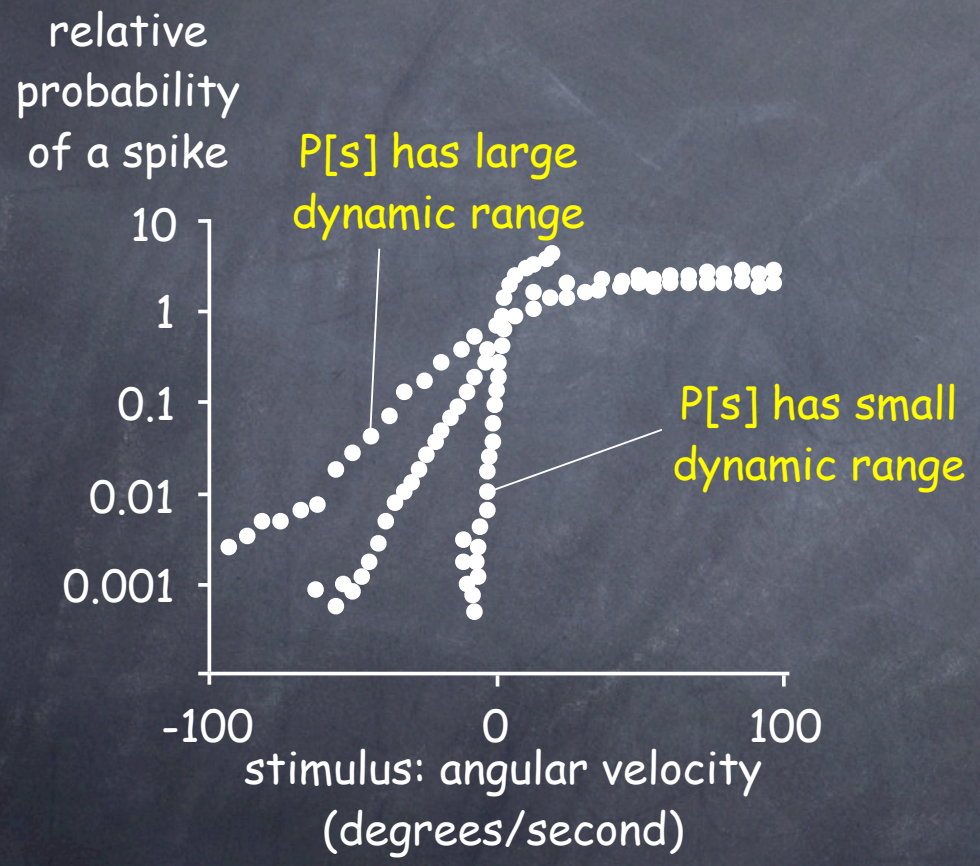
Because mutual information is context dependent, the optimal input/output relation is matched to P(input)

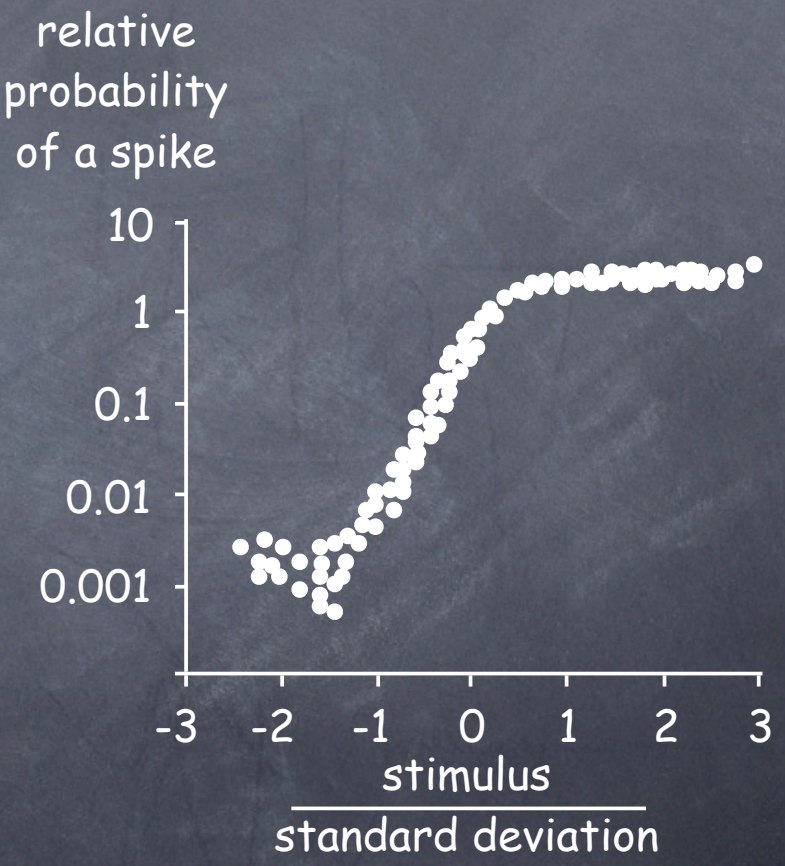


(somewhat embarrassingly, equations don't add much to this picture)

Measure input/output relations when inputs are drawn from different distributions P[s]

(important technical question of how to do this!)



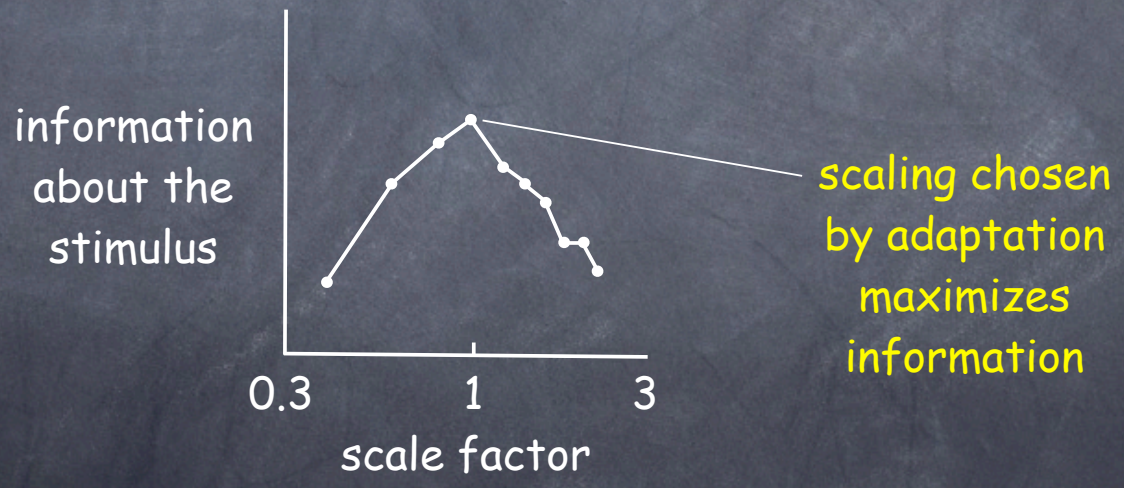


Adaptive rescaling optimizes information transmission.
N Brenner, W Bialek & RR de Ruyter van Steveninck,
Neuron 26, 695-702 (2000).

the code adapts to the distribution of inputs, and

the form of adaptation is consistent with the an optimization principle, but ...

how do we know that information is optimized?

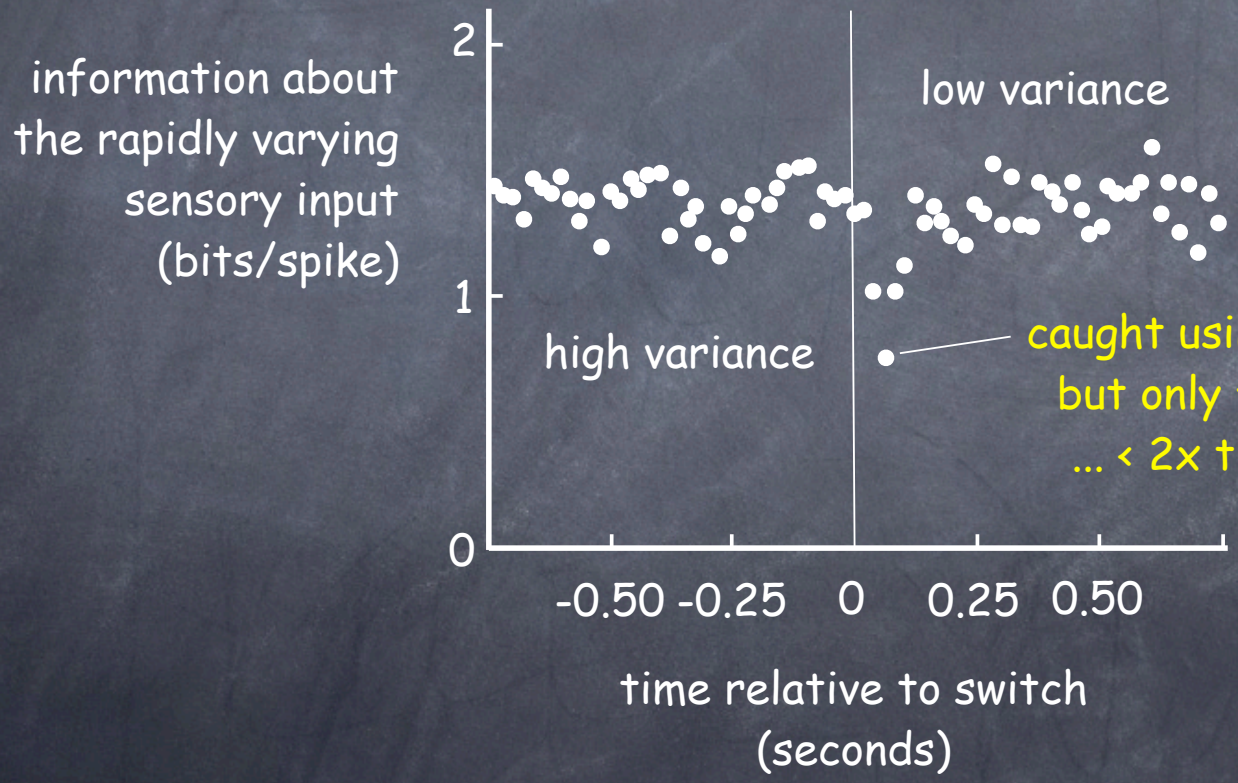


calculate the information that would be transmitted if P[s] is fixed and the cell chose different rescalings of the input/output relation ...

How quickly can the system adjust?

(interestingly difficult to measure)

how long does it take to be sure
that we are seeing a new distributions vs.
outliers in the old distribution?



Efficiency and ambiguity in an adaptive neural code.
AL Fairhall, GD Lewen, W Bialek & RR de Ruyter van Steveninck,
Nature 412, 787-792 (2001).

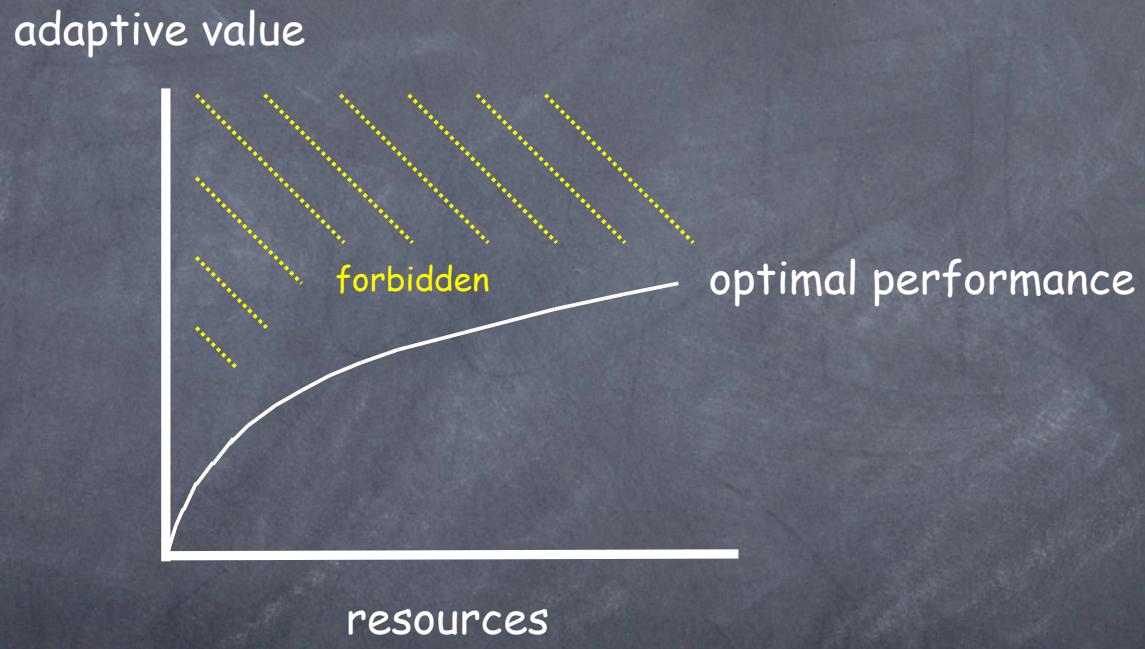
do we have one principle or many?

optimal estimation sounds unified, but comes out differently for each thing we want to estimate. how do we choose?

information theoretic principles have even more generality. but information always is about something. information about what?

(these are polite versions of questions asked by biologists)

Can we find a general, but still biologically meaningful notion of optimal performance?

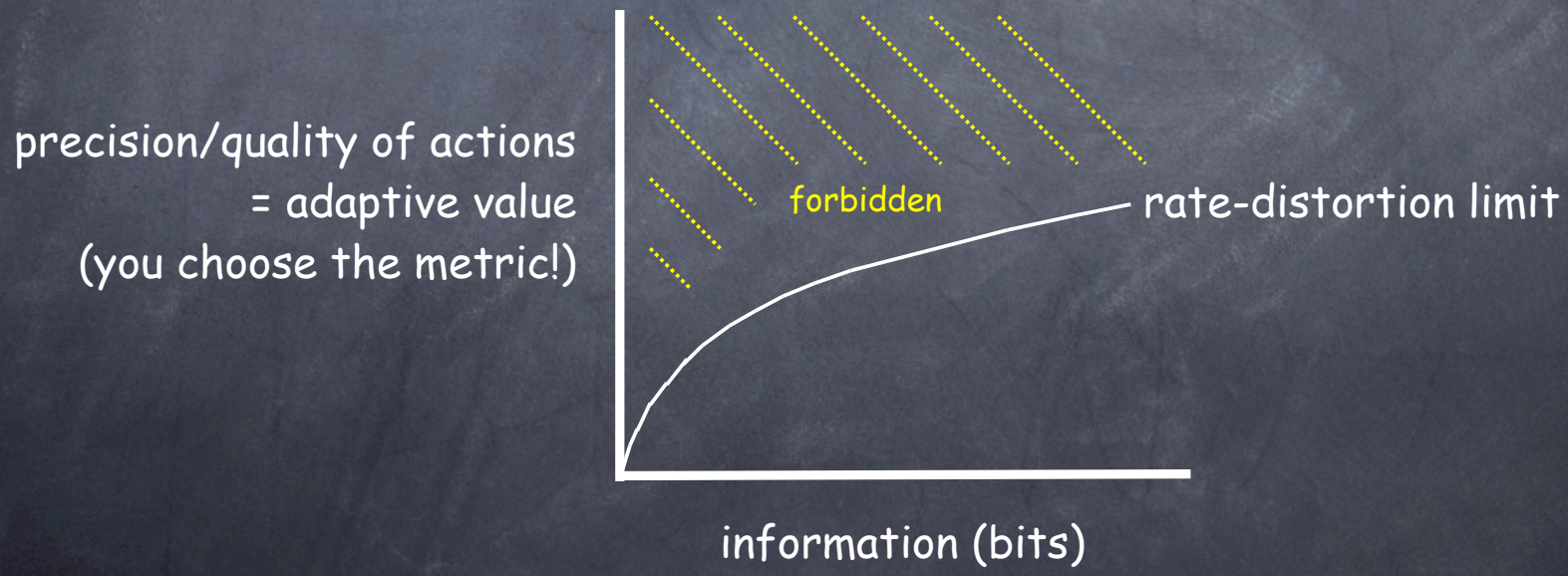


(note that we would still will have a family of solutions, but this is a good thing!)

these arguments based on unpublished work with N Tishby

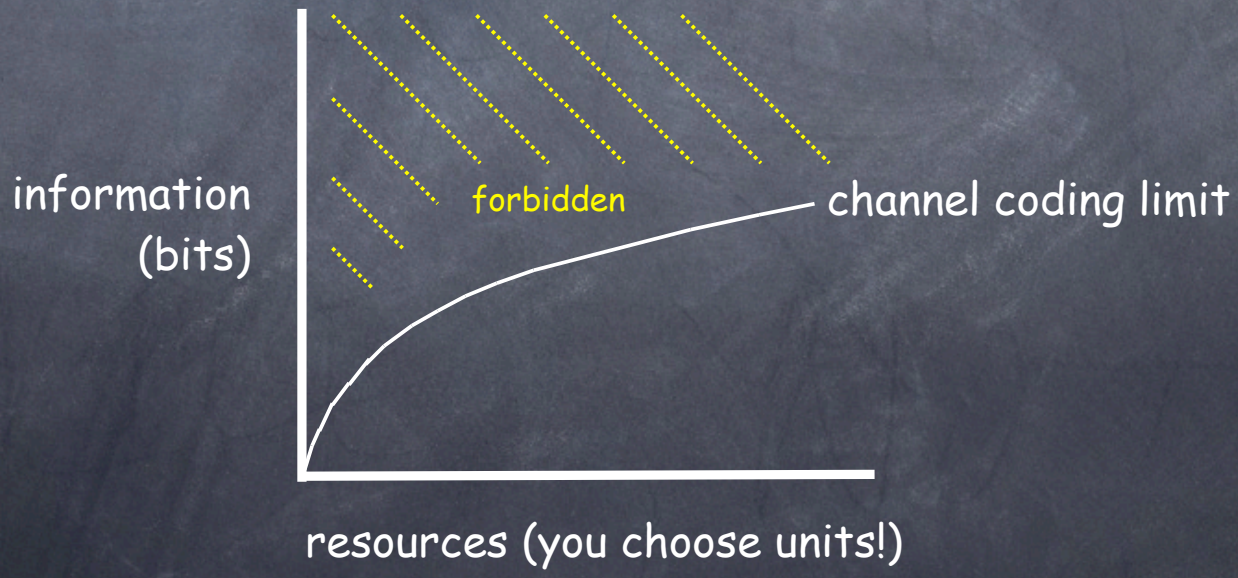
Organisms are not rewarded for maximizing information; they are rewarded for appropriate actions.

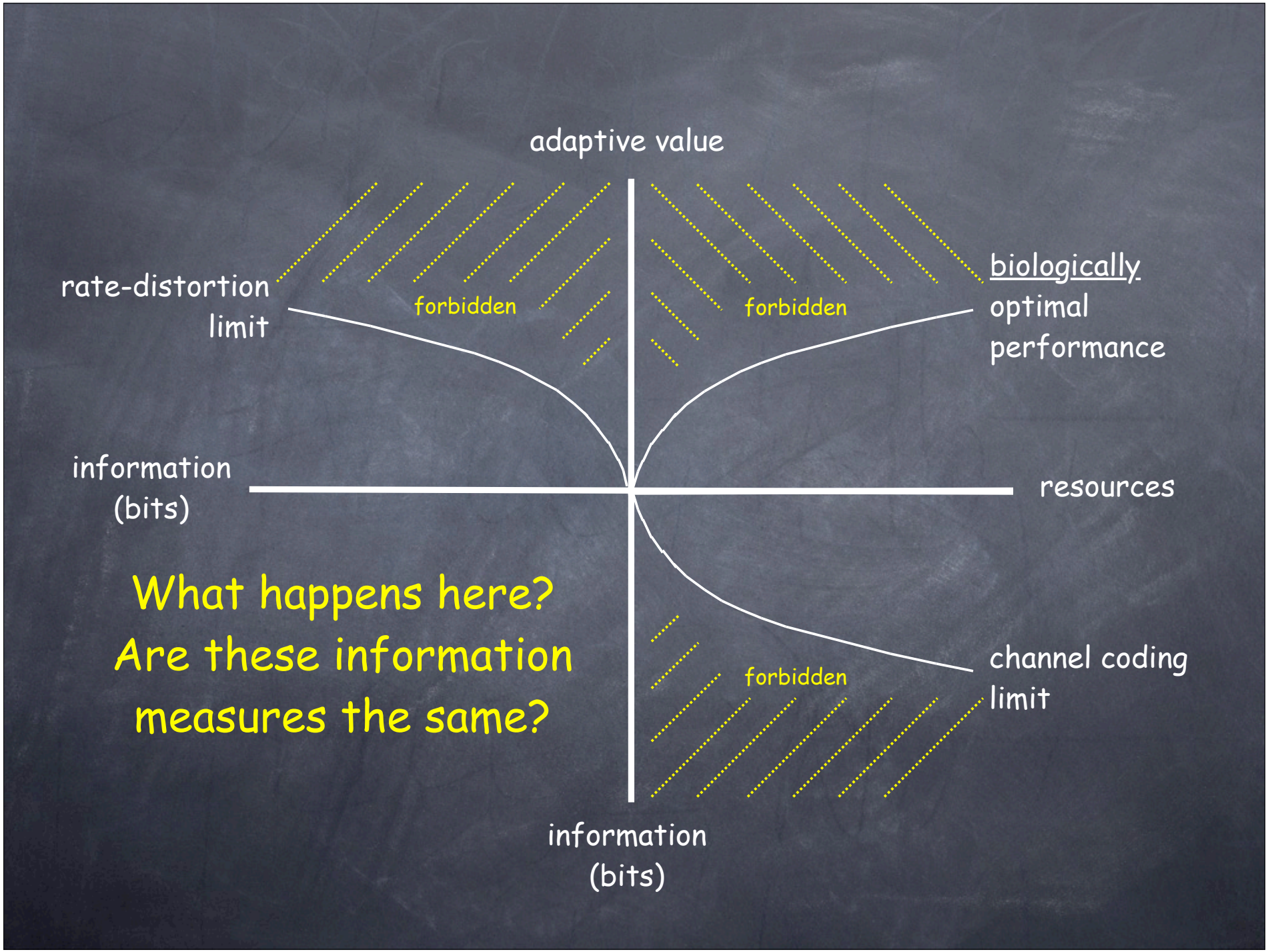
But to act with some level of precision or effectiveness requires a minimum amount of information ... this is the content of rate-distortion theory



Organisms are not just under pressure to act, they have to do so with limited resources.

But given fixed resources (e.g., energy), there is a limit to how much information we can represent ... this the content of channel coding theory

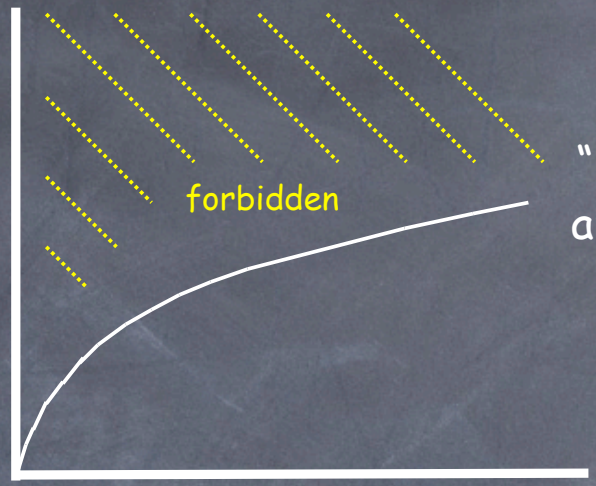



adaptive value
rate-distortion limit
forbidden
forbidden
biologically optimal performance
information (bits)
resources
What happens here? Are these information measures the same?
forbidden
channel coding limit
information (bits)

the information we extract from sensory inputs must be information about the past (causality)

but information that has a chance of having adaptive value must be predictive:
information about the future

given the statistical structure of the world, a certain amount of information about the past provides only a limited amount of information about the future

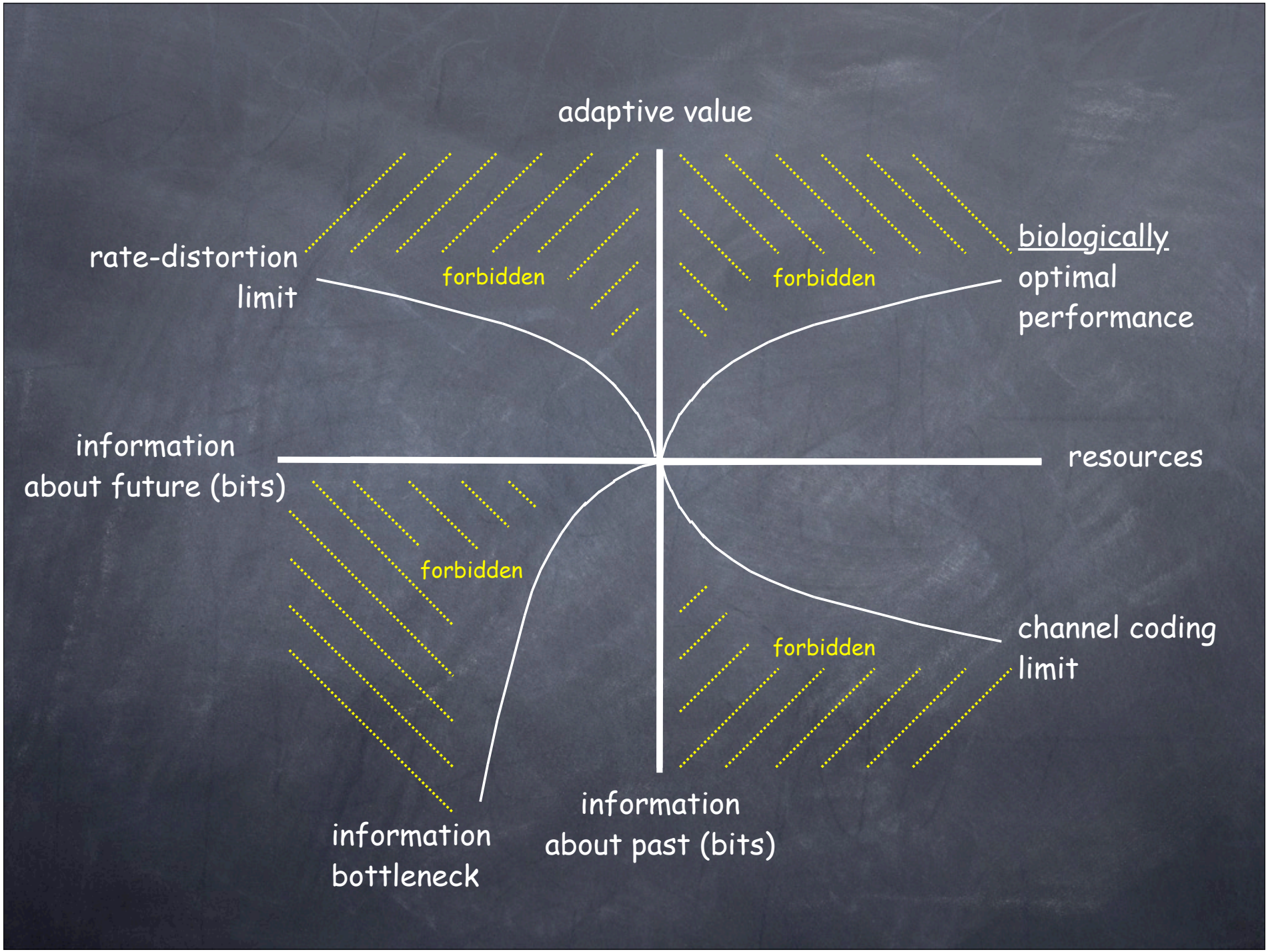


"information bottleneck"
applied to predictive info

efficient representation of predictive information contains many other problems:
filtering to separate signal from noise
estimating parameters of a model to be learned from the data
...

Predictability, complexity and learning.
W Bialek, I Nemenman & N Tishby,
Neural Comp 13, 2409-2463 (2001).

The information bottleneck method.
N Tishby, FC Pereira & W Bialek,
in Proceedings of the 37th Allerton Conference, B Hajek & RS Sreenivas,
eds, pp 368-377 (University of Illinois, 1999).


adaptive value
rate-distortion limit
forbidden
forbidden
biologically optimal performance
information about future (bits)
resources
forbidden
forbidden
channel coding limit
information bottleneck
information about past (bits)