Entropy and Information in the Neuroscience Laboratory

G80.3042.002:
Statistical Analysis and Modeling of Neural Data

Jonathan D. Victor

Department of Neurology and Neuroscience
Weill Medical College of Cornell University

November 2007

Thanks

WMC Neurology and Neuroscience

Keith Purpura
Ferenc Mechler
Anita Schmid
Ifije Ohiorhenuan
Qin Hu

Former students

Dmitriy Aronov
Danny Reich
Michael Repucci
Bob DeBellis

Collaborators

Dan Gardner (WMC)
Bruce Knight (Rockefeller)
Patricia DiLorenzo (SUNY Binghamton)

Support

NEI
NINDS
WMC Neurology and Neuroscience

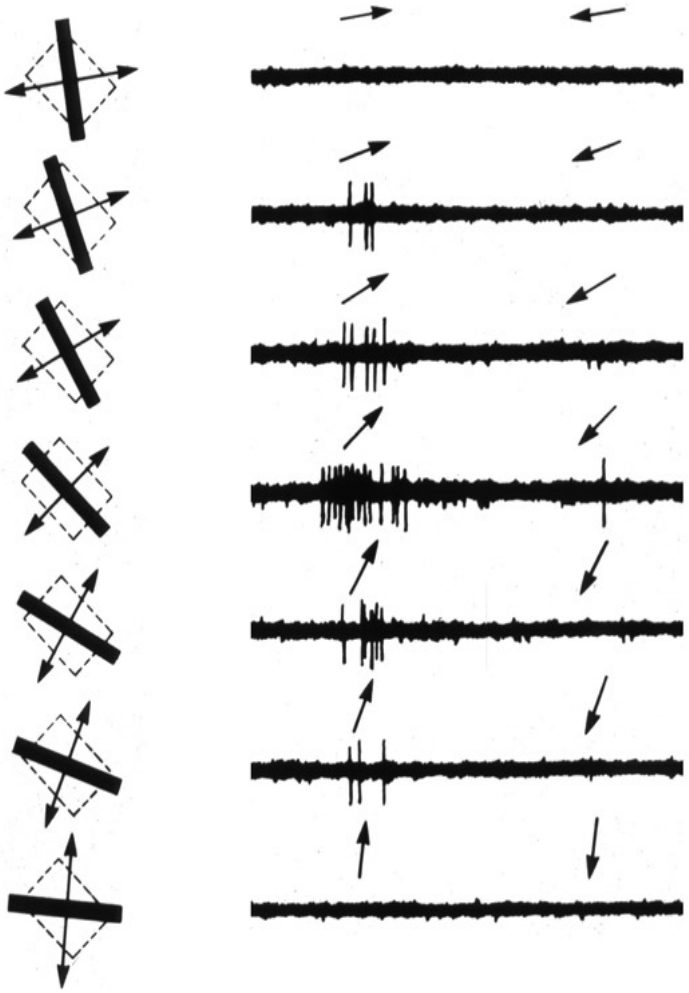
Outline

- Information
  - Why calculate information?
  - What is it?
  - Why is it challenging to estimate?
  - What are some useful strategies for neural data?
  - Examples
- Maximum entropy methods
  - Data analysis
  - Stimulus design
- Homework

Information: Why Calculate It?

- An interesting, natural quantity
  - Compare across systems (e.g., “one spike per bit”)
  - Determine the constraints on a system (e.g., metabolic cost of information)
  - See where it is lost (e.g., within a neuron)
  - Insight into what the system is “designed” to do
  - A non-parametric measure of association
- To evaluate candidates for neural codes
  - What statistical features are available?
    - Can precise spike timing carry information?
    - Can neuronal diversity carry information?
  - What codes can be rigorously *ruled out*?

Even in visual cortex, the neural code unknown



Hubel and Wiesel 1968

What physiologists say:

- Neurons have definite selectivities (“tuning”)
- Tuning properties can account for behavior

What physiologists also know:

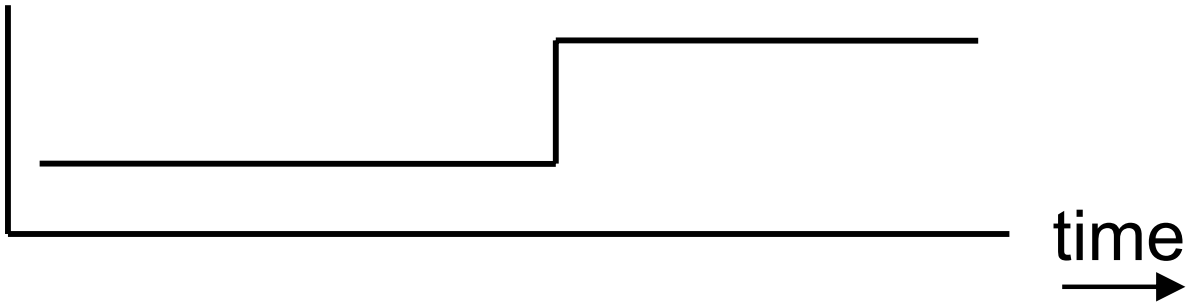
- Responses depend on multiple stimulus parameters
- Response variability (number of spikes, and firing pattern) is substantial and complicated

Some coding hypotheses

- At the level of individual neurons
  - Spike count
  - Firing rate envelope
  - Interspike interval pattern, e.g., bursts
- At the level of neural populations
  - Total population activity
  - Labeled lines
  - Patterns across neurons
  - Synchronous spikes
  - Oscillations

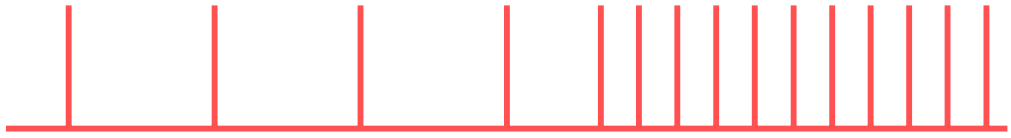
Coding by intervals can be faster than coding by count

Signaling a step change in a sensory input



coding by count (rate averaged over time)

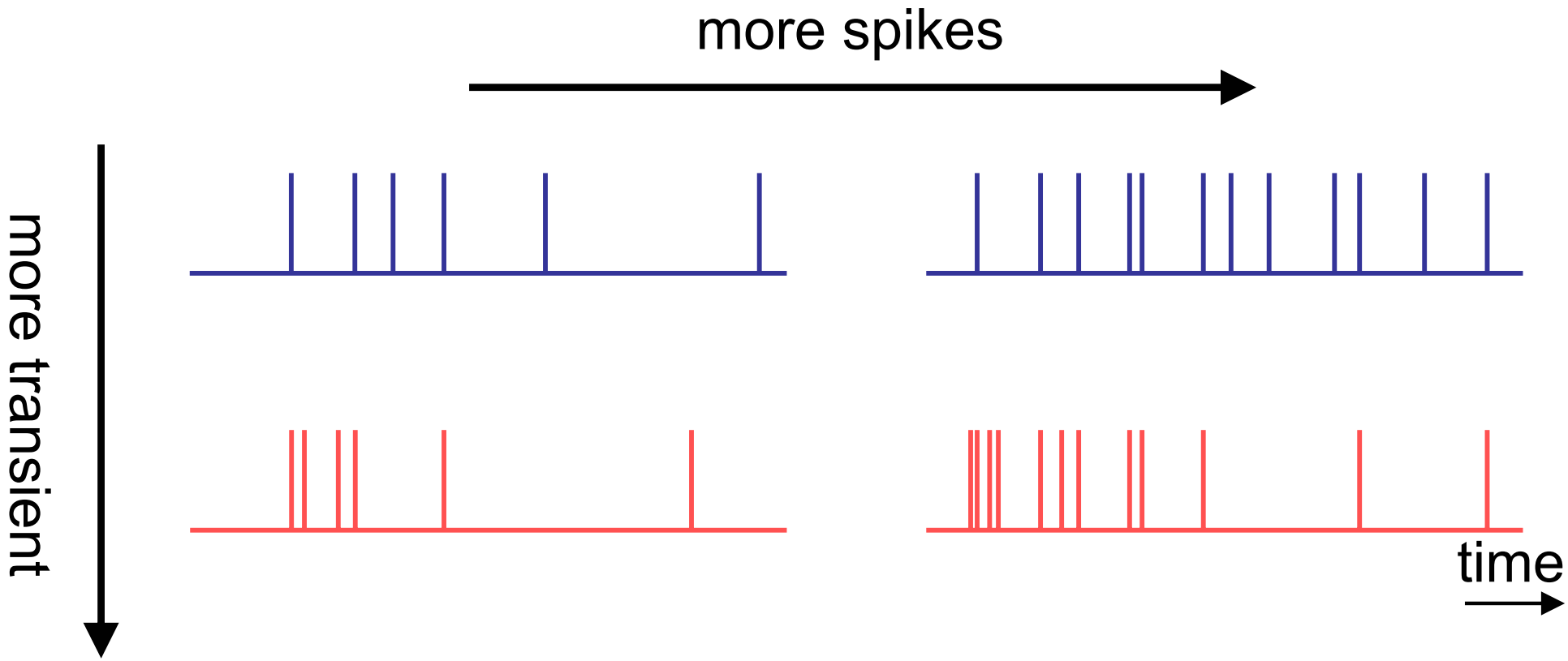
coding by interval pattern



One short interval indicates a change!

Tradeoff: firing must be regular

Coding by rate envelope supports signaling of multiple attributes



Codes based on spike patterns can also support signaling of multiple attributes.

A direct experimental test of a neural coding hypothesis is difficult

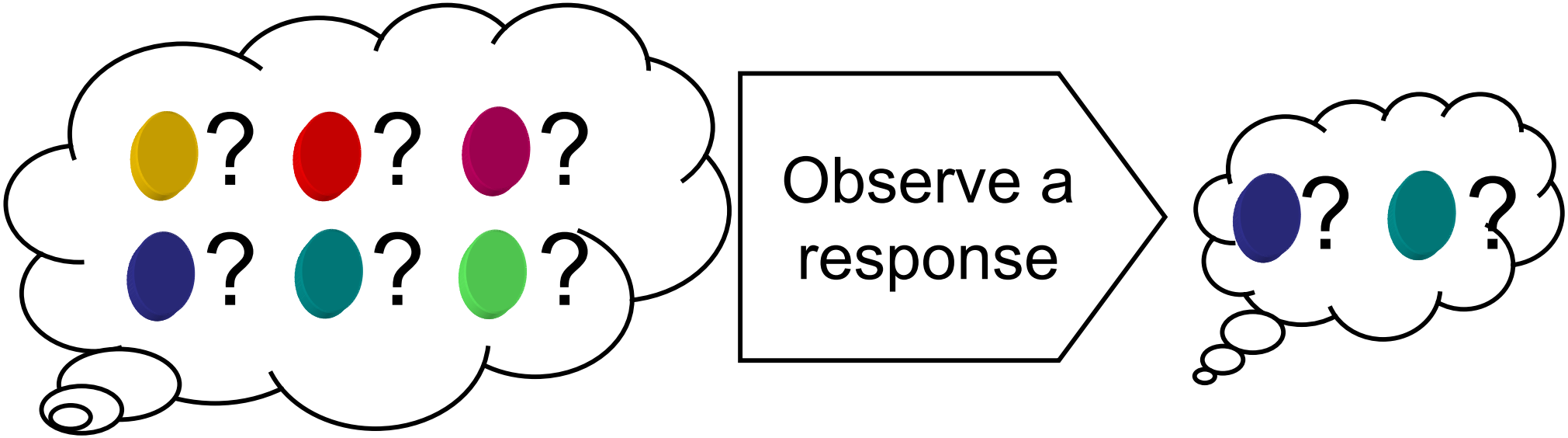
- Count, rate, and pattern are interdependent

  "Time is that great gift of nature which keeps everything from happening at once." (C.J. Overbeck, 1978)

- We'd have to manipulate count, rate, and pattern selectively *AND observe an effect on behavior*
- So, we need some guidance from theory

Information = Reduction in Uncertainty

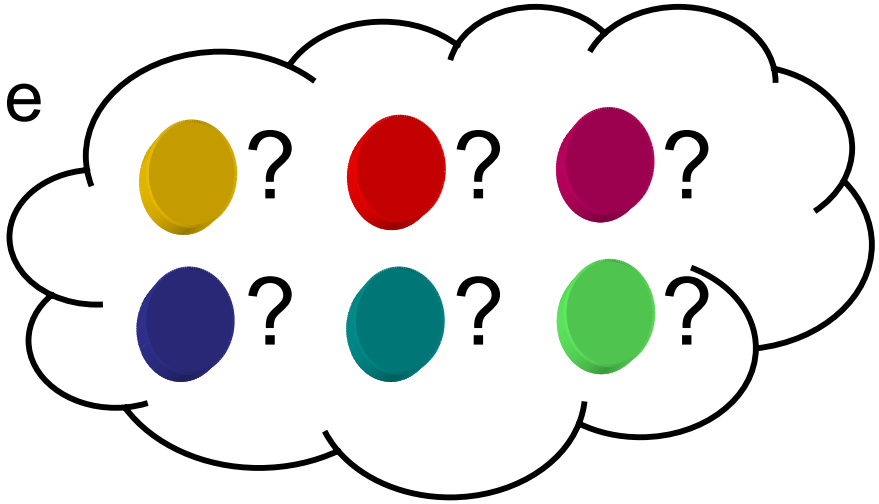
(Claude Shannon, 1948)



- Reduction in uncertainty from 6 possibilities to 2
- Information = log(6/2)

In a bit more detail:

A priori knowledge



Observe a response

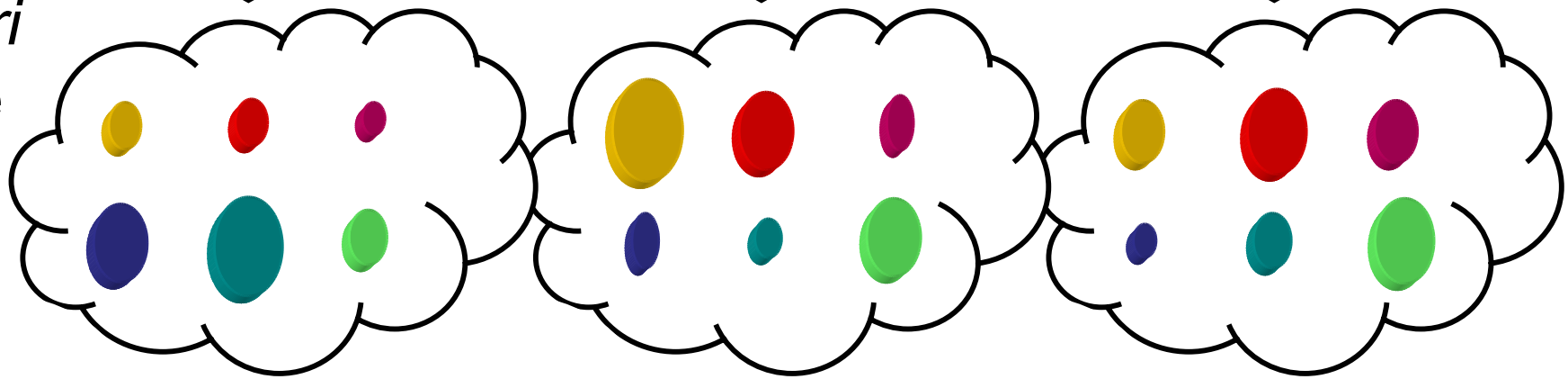
No spikes

One spike

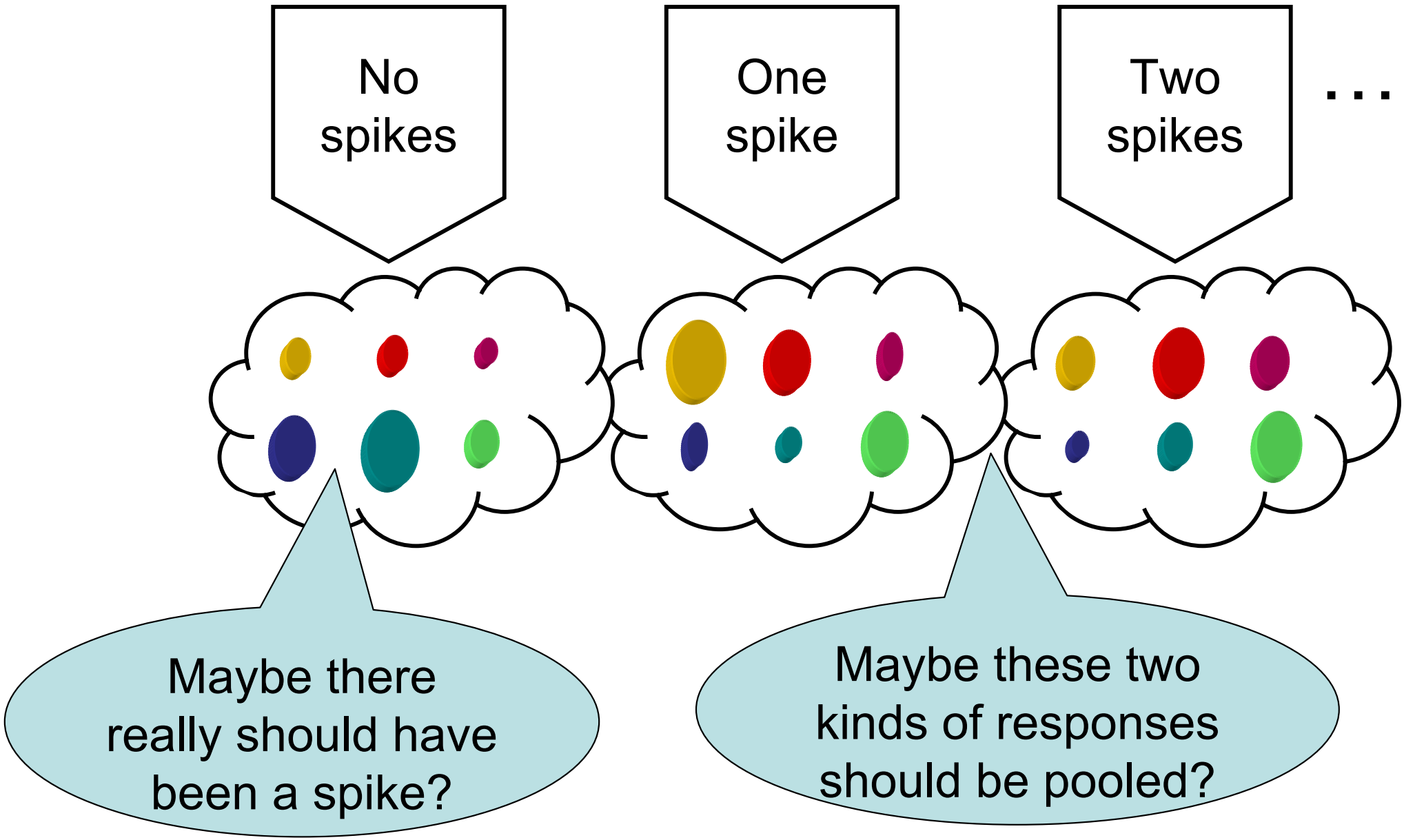
Two spikes

...

A posteriori knowledge



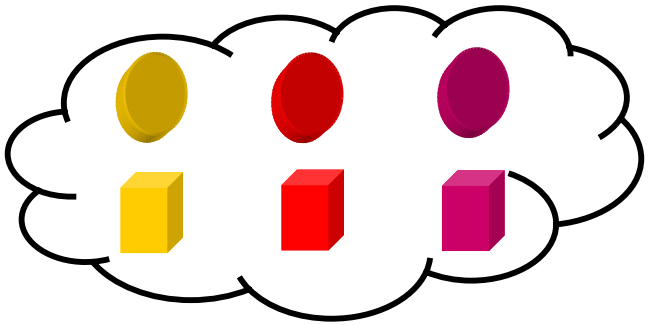
Second-guessing shouldn't help



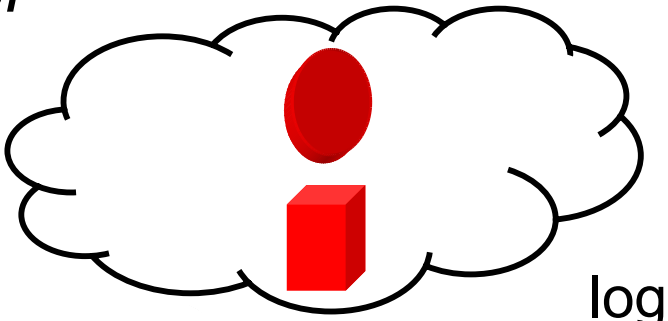
The "Data Processing Inequality": information cannot be increased by re-analysis.

Information on independent channels should add

Color channel



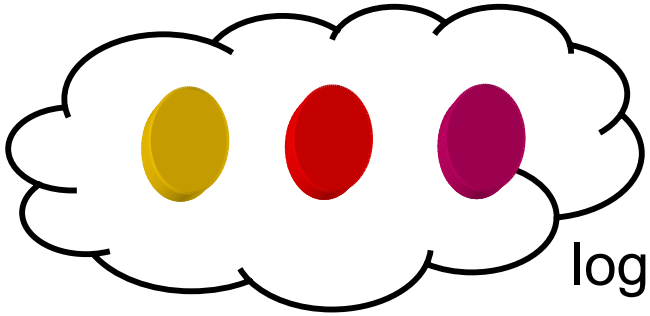
Observe a response



log(6/2) = log(3)

Shape channel

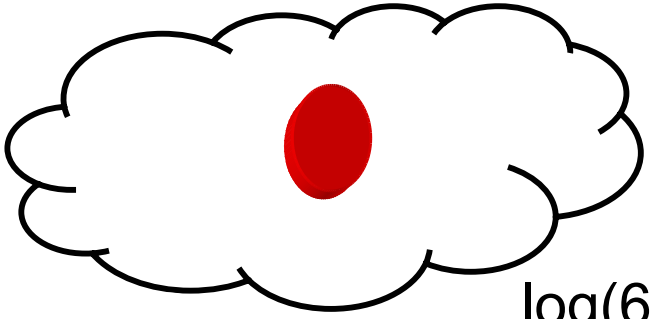
Observe a response



log(6/3) = log(2)

Both channels

Observe a response



log(6/1) = log(6)

$$\log(6) = \log(3 \times 2) = \log(3) + \log(2)$$

Surprising Consequence

Data Processing Inequality

+

Independent channels combine additively

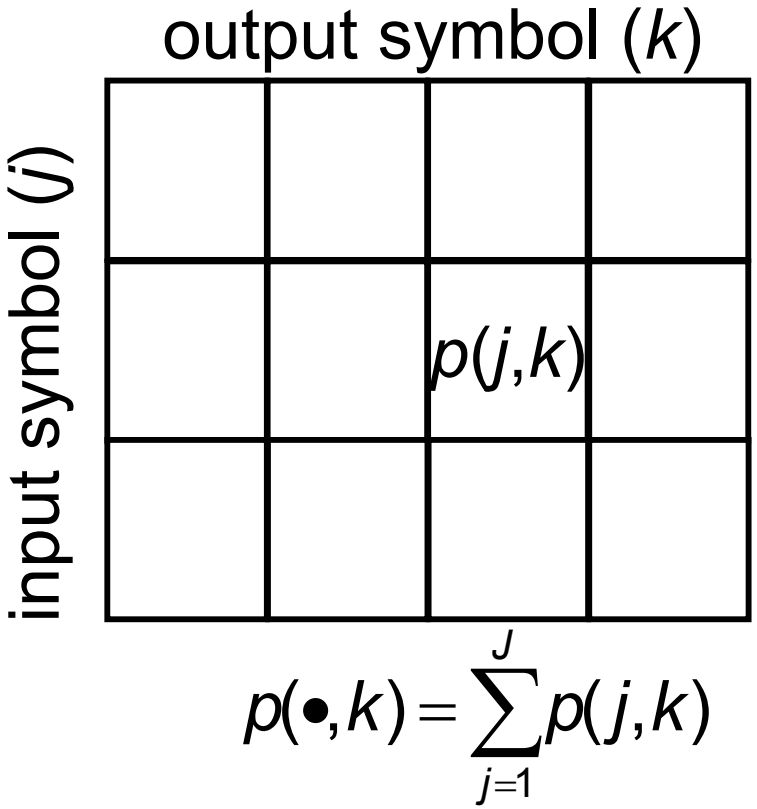
+

Continuity

=

Unique definition of information, up to a constant

Information: Difference of Entropies



$$p(j,\bullet) = \sum_{k=1}^{K} p(j,k)$$

$$p(\bullet,k) = \sum_{j=1}^{J} p(j,k)$$

a priori distribution:

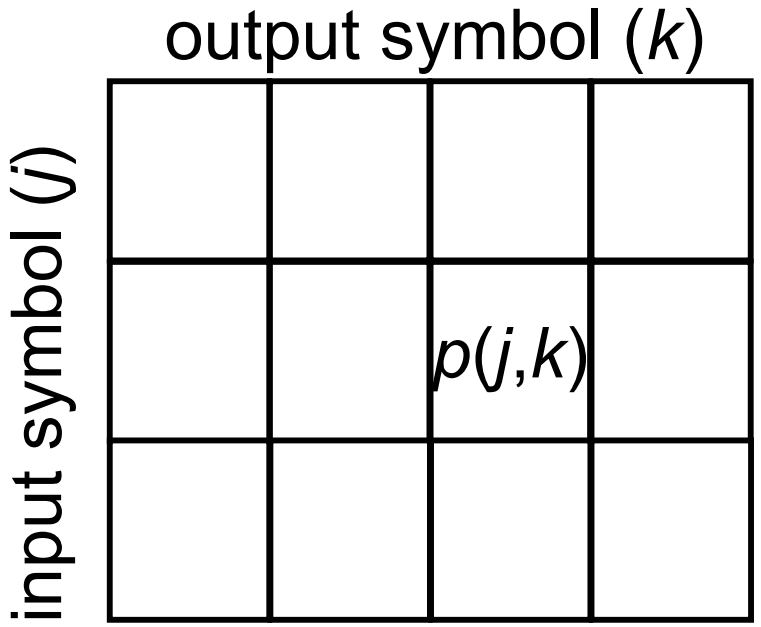
$$p(j,\bullet)$$

a posteriori distribution, given output symbol *k*:

$$p(j \mid k) = p(j,k) / p(\bullet,k)$$

Information ={Entropy of the *a priori* distribution of input symbols}
minus
{Entropy of *a posteriori* distribution of input symbols, given the observation *k*, averaged over all *k*}

Information: Symmetric Difference of Entropies



Information =

{entropy of output}+{entropy of input} -{entropy of table}

$$I = H\{p(j,\bullet)\} + H\{p(\bullet,k)\} - H\{p(j,k)\}$$

Information: Properties

$$I = H\{p(j,\bullet)\} + H\{p(\bullet,k)\} - H\{p(j,k)\}$$

I is symmetric in input and output

I is independent of labeling within input and output

$I \geq 0$, and $I = 0$ if and only if $p(j,k) = p(j,\bullet)p(\bullet,k)$

$I \leq H\{p(j,\bullet)\}$ and $I \leq H\{p(\bullet,k)\}$

Data Processing Inequality: if Y determines Z, then

$$I(X,Y) \geq I(X,Z)$$

Information: Related Quantities

Channel capacity

maximum information for any input ensemble

Efficiency

{Information}/{Channel capacity}

Redundancy

$$1 - \frac{\{\text{Information from all channels}\}}{\{\text{sum of informations from each channel}\}}$$

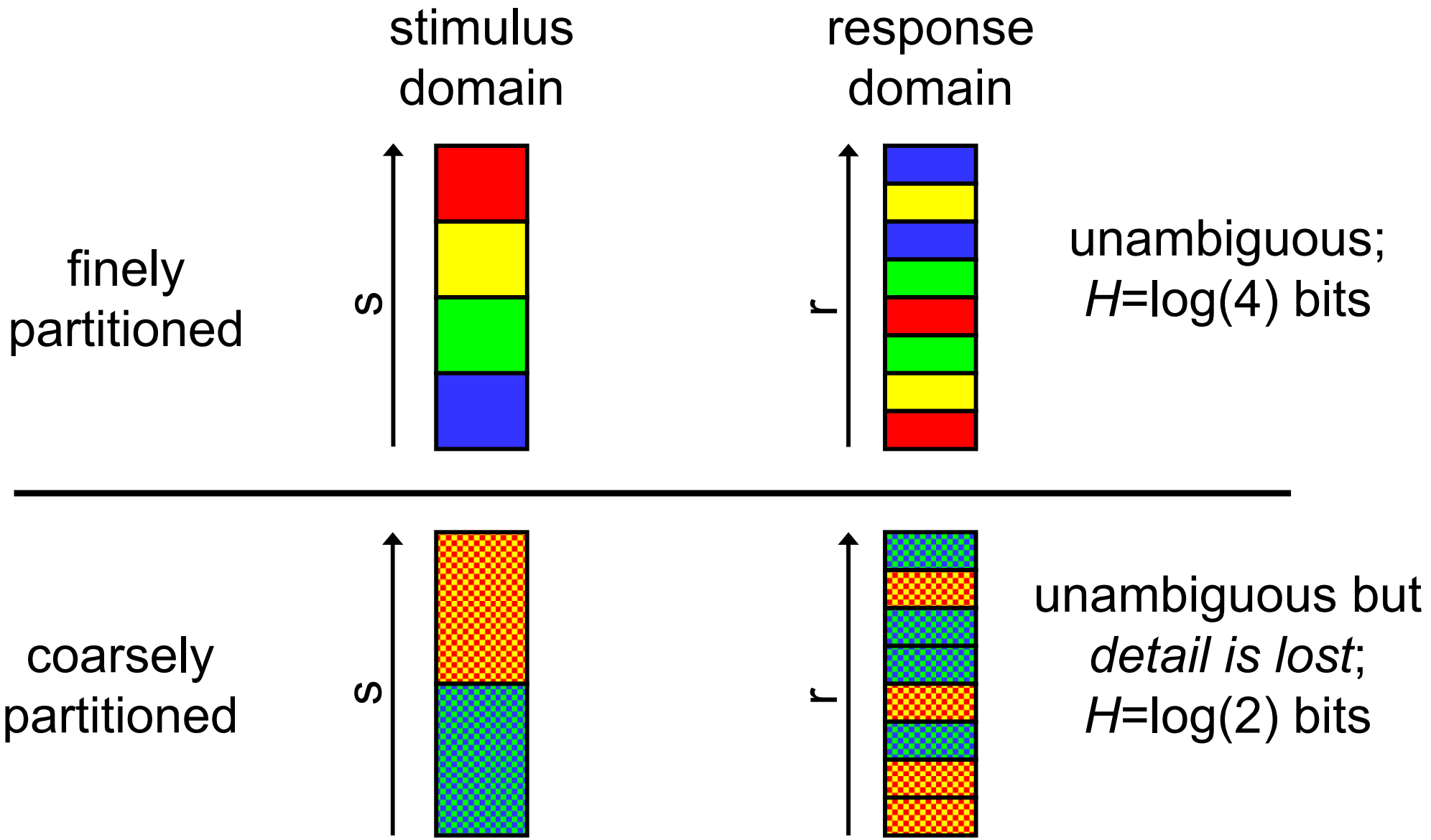
Redundancy Index

$$\frac{1 - \dfrac{\{\text{Information from all channels}\}}{\{\text{sum of informations from each channel}\}}}{1 - \dfrac{\{\text{Information from all channels}\}}{\{\text{maximum of informations from each channel}\}}}$$

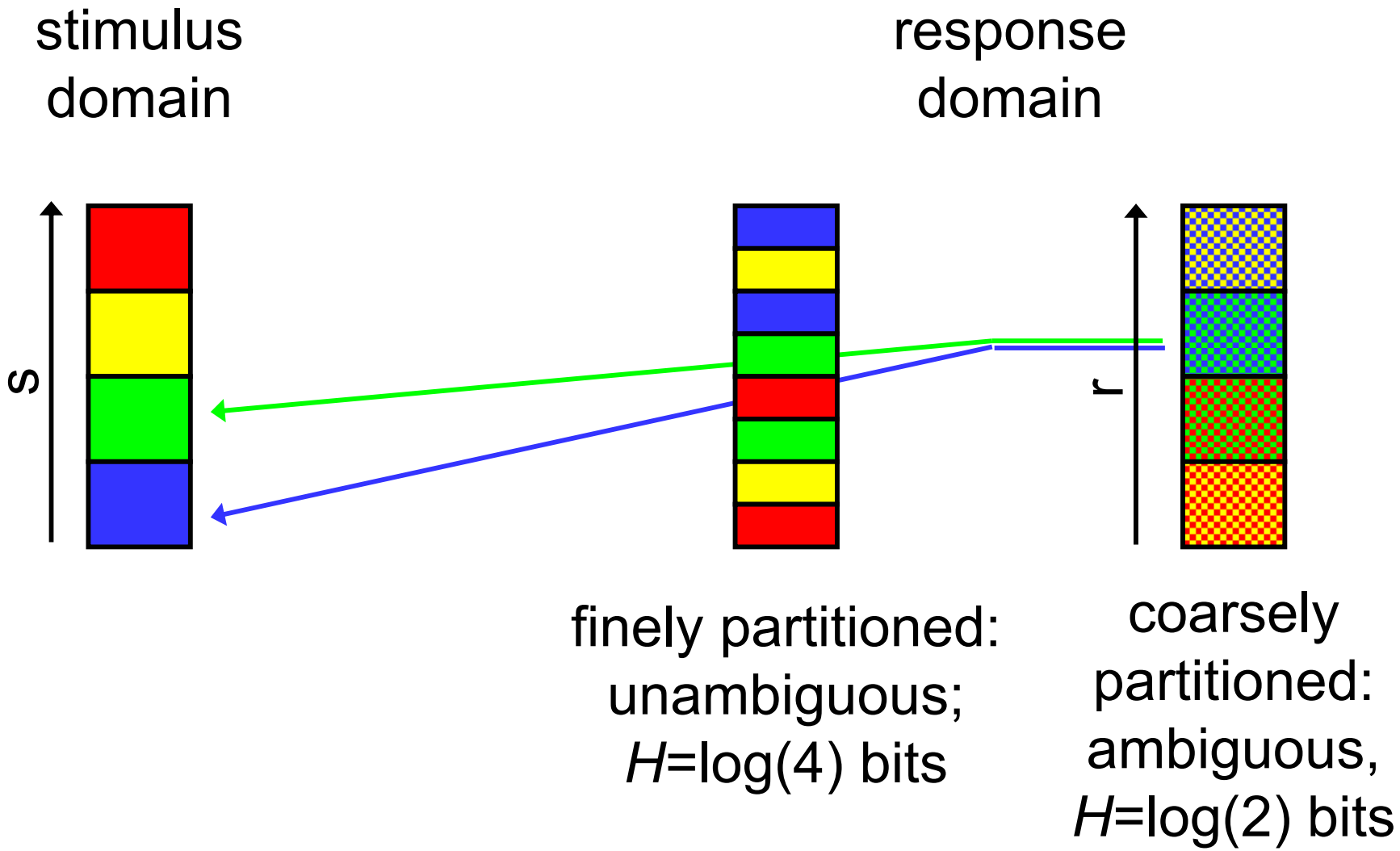
Investigating neural coding: not Shannon's paradigm

- Shannon
  - symbols and codes are known
  - joint (input/output) probabilities are known
  - what are the limits of performance?
- Neural coding
  - symbols and codes are not known
  - joint probabilities must be measured
  - ultimate performance often known (behavior)
  - what are the codes?

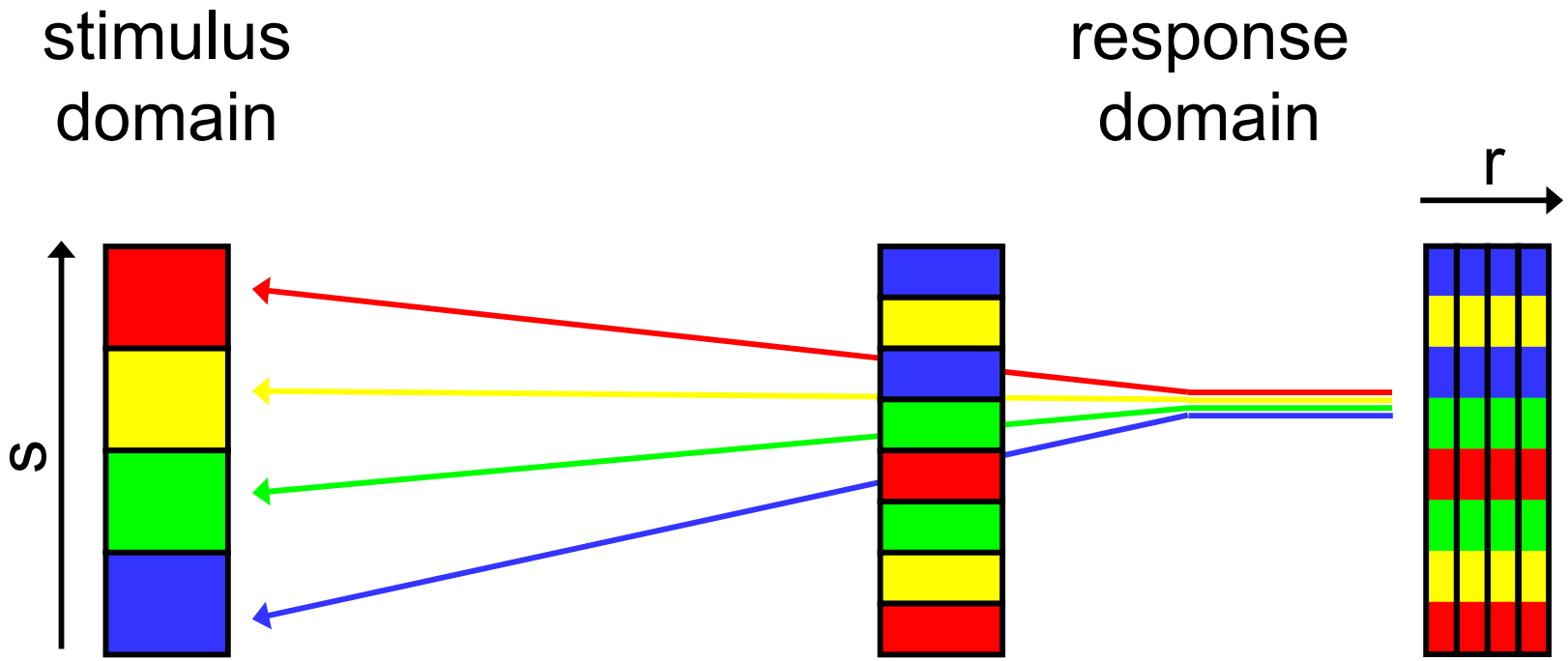
Information estimates depend on partitioning of stimulus domain



Information estimates depend on partitioning of response domain



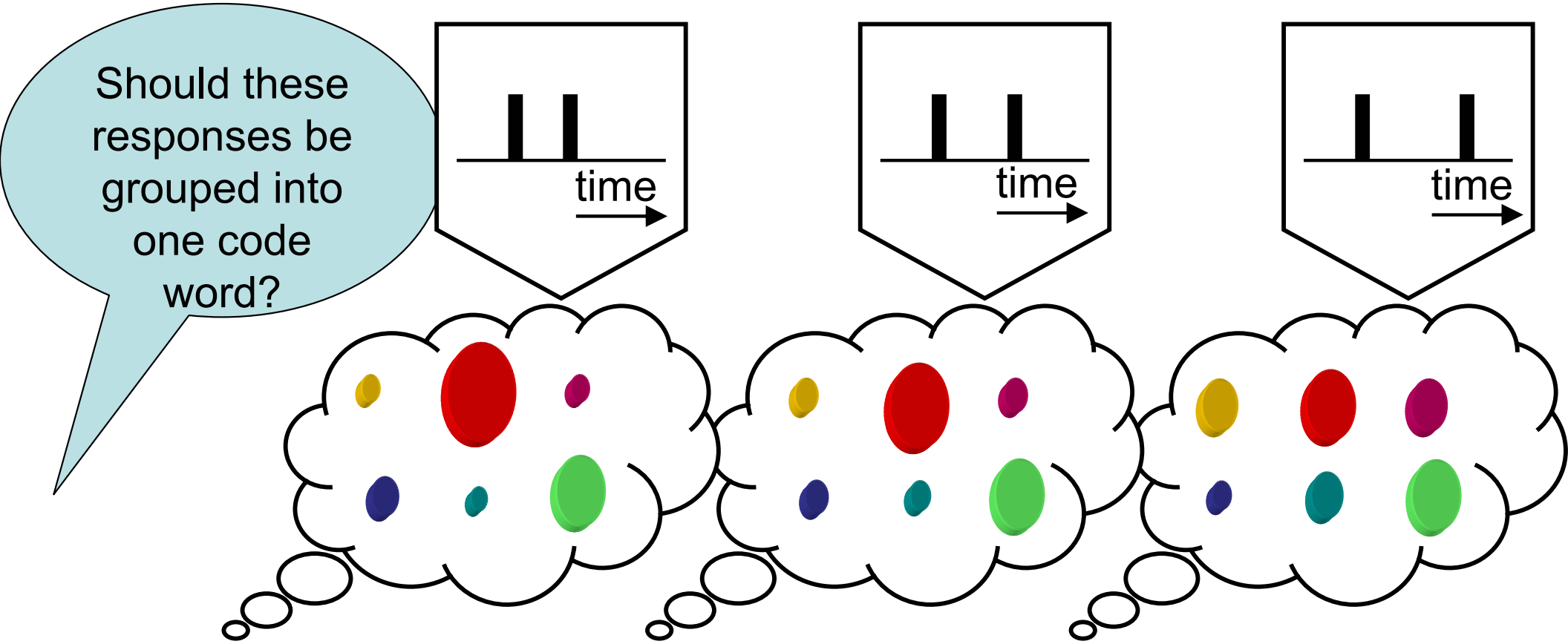
Information estimates depend on partitioning of response domain, II



finely partitioned: unambiguous; H=log(4)=2 bits

wrongly partitioned: ambiguous, H=log(1)=0 bits

Revenge of the Data Processing Inequality



Data Processing Inequality says ***NO***: If you group, you underestimate information

The Basic Difficulty

We need to divide stimulus and response domains finely, to avoid underestimating information ("Data Processing Theorem").

We want to determine $\langle p \log p \rangle$, but we only have an estimate of p, not its exact value. Dividing stimulus and response domains makes p small. This increases the **variability** of estimates of p.

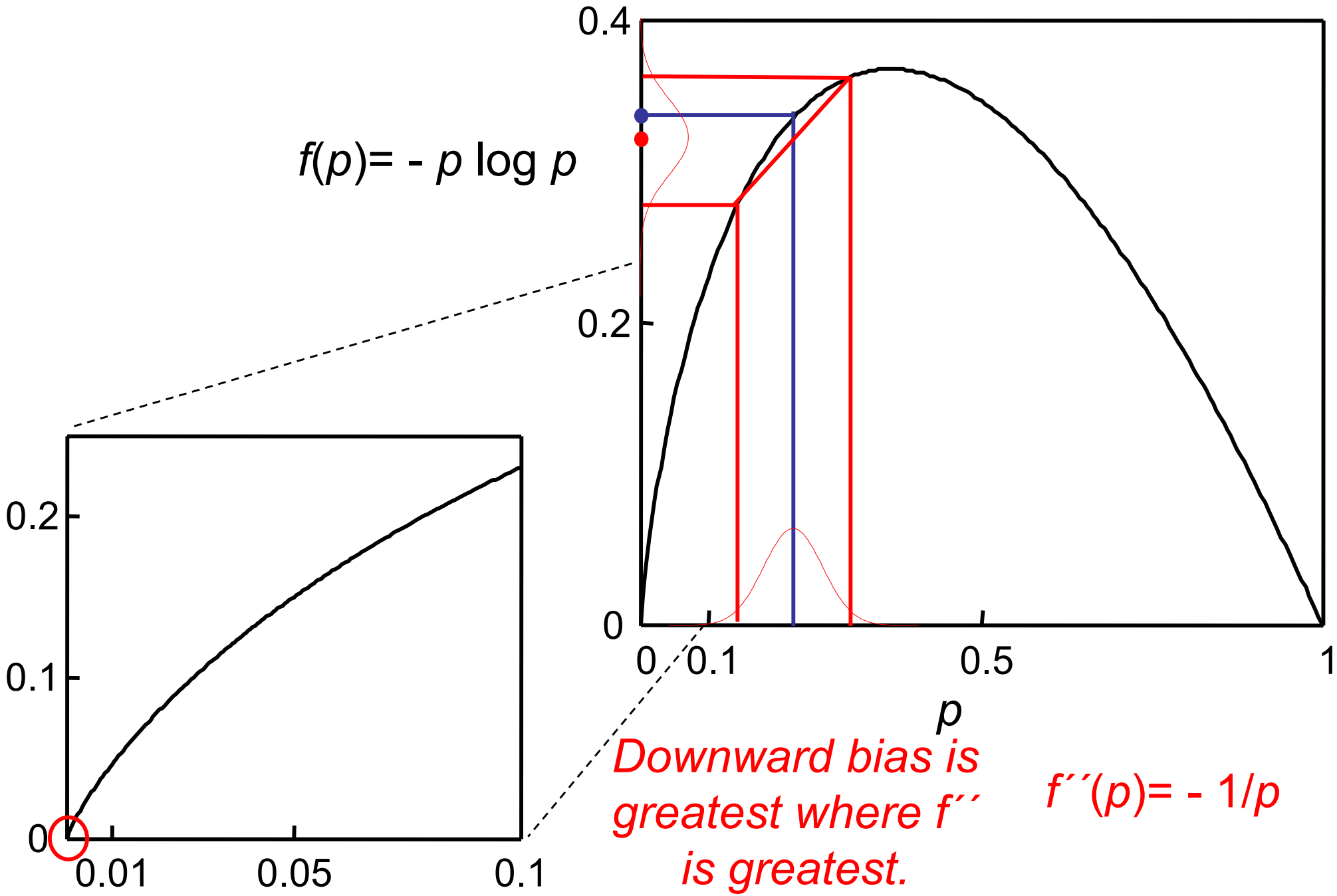
But that's not all…

$p \log p$ is a nonlinear function of p.

Replacing $\langle p \log p \rangle$ with $\langle p \rangle \log \langle p \rangle$ incurs a bias.

How does this **bias** depend on p?

Biased Estimates of -p log p



The Classic Debiaser: Good News/ Bad News

We don't have to debias every $p \log p$ term, just the sum.

The good news (for entropy of a discrete distribution):

> The plug-in entropy estimate has an asymptotic bias proportional to $(k-1)/N$, where N is the number of samples and k is the number of different symbols (Miller, Carlton, Treves, Panzeri).

The bad news:

> Unless $N>>k$, the asymptotic correction may be worse than none at all.

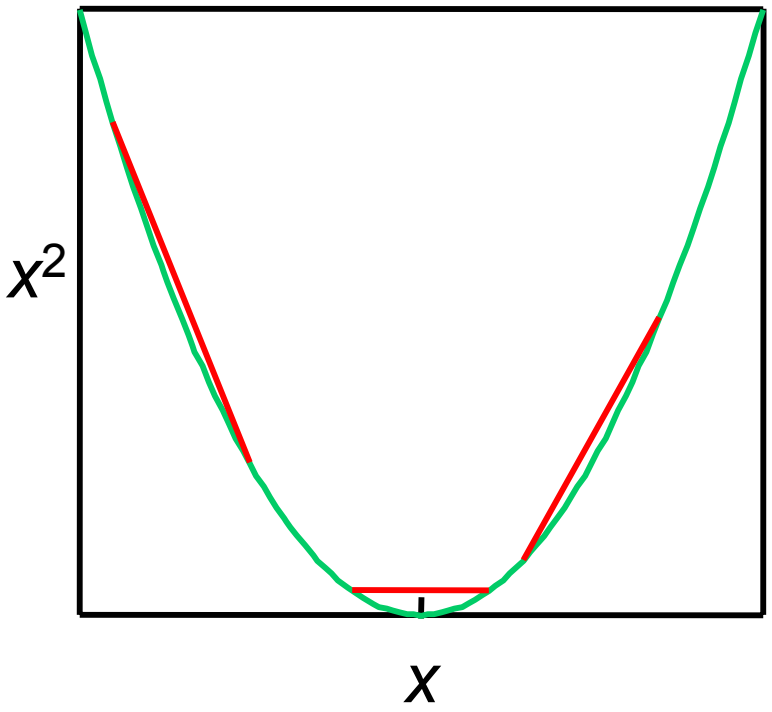
More bad news:

> We don't know what k is.

Another debiasing strategy

Toy problem:

$\langle x^2 \rangle \neq \langle x \rangle^2$



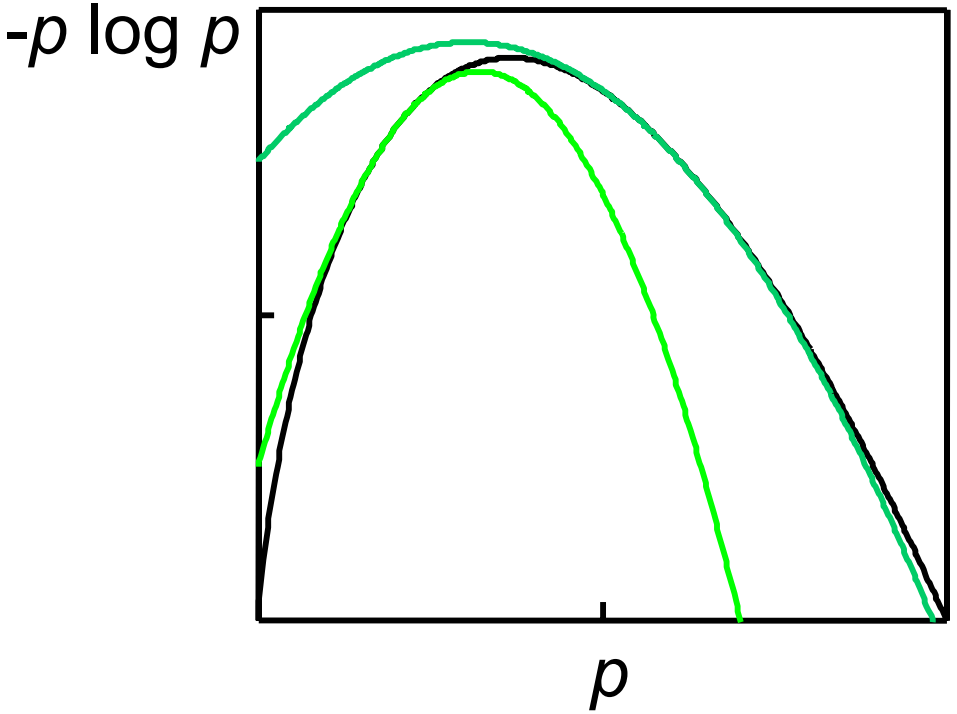
For a parabola, bias is constant.

This is why the naïve estimator for variance can be simply debiased:

$\sigma^2_{est} = \langle (x - \langle x \rangle)^2 \rangle / (N-1)$

Our problem:

$\langle -p \log p \rangle \neq -\langle p \rangle log \langle p \rangle$



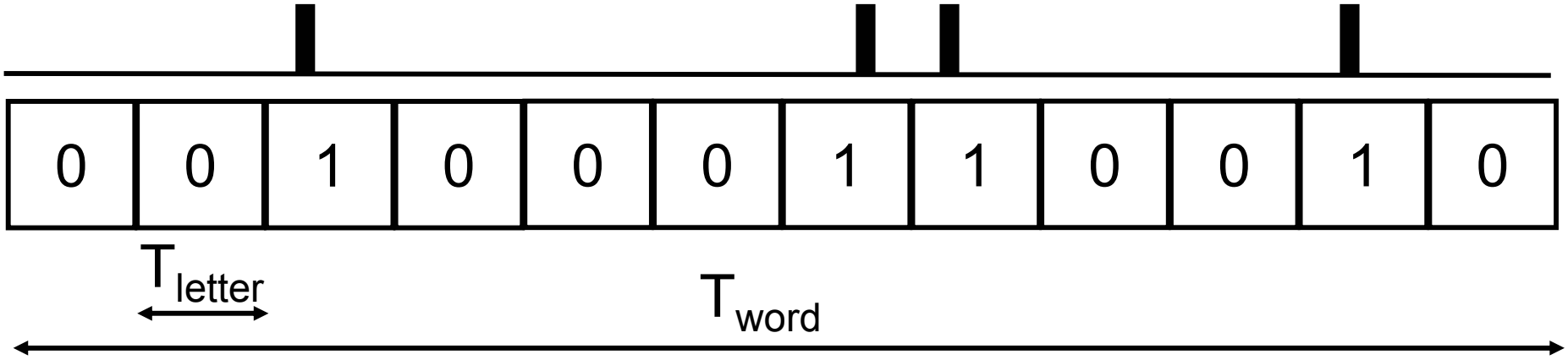
Bias depends on the best local parabolic approximation. This leads to a polynomial debiaser. (Paninski)

Better than classical debiaser, but p=0 is still worst case. And it still fails in the extreme undersampled regime.

The “Direct Method”

(Strong, de Ruyter, Bialek, et al. 1998)

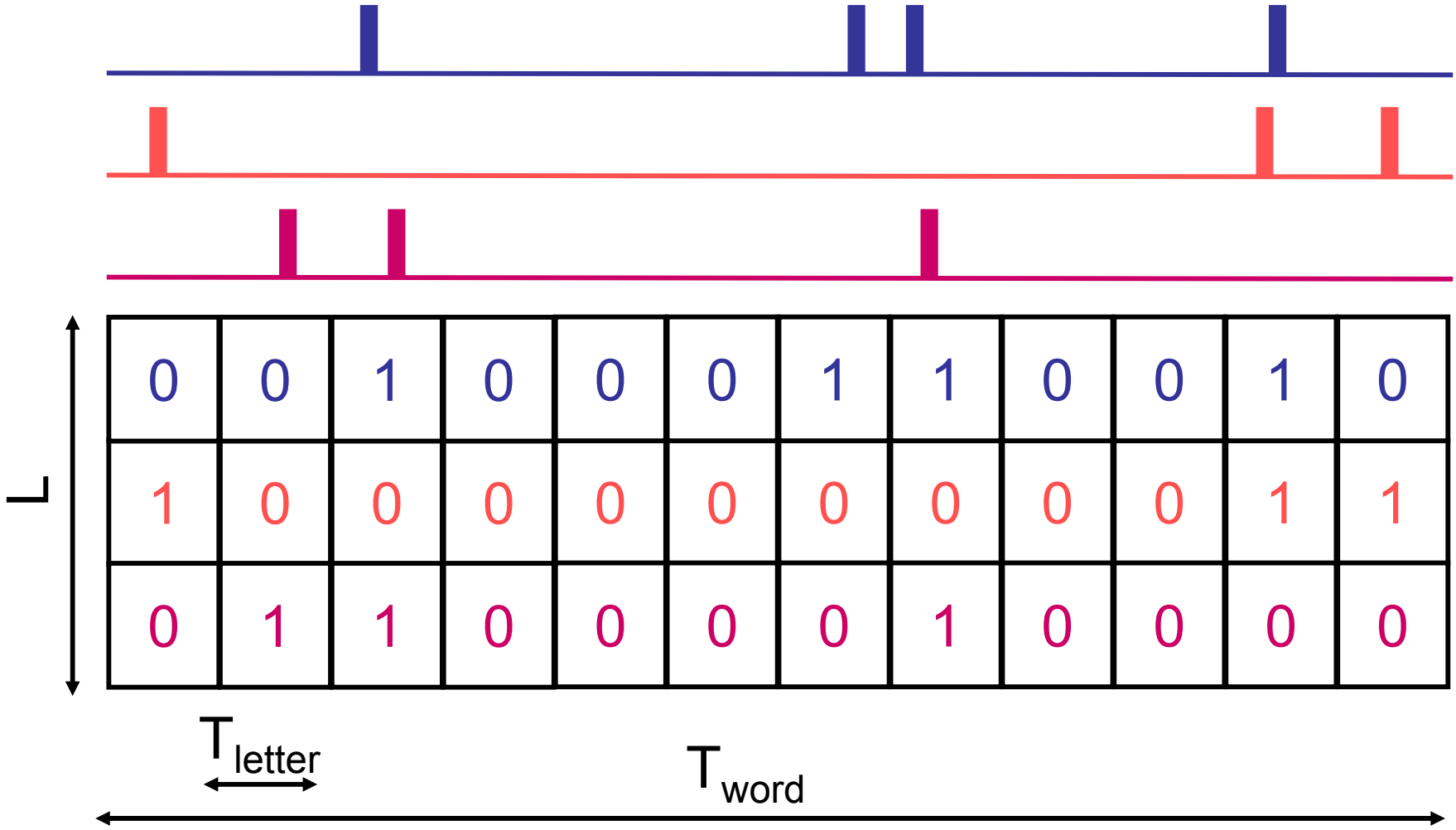
- Discretize the response into binary “words”



- T_{letter} must be small to capture temporal detail
  - timing precision of spikes: <1 ms
- T_{word} must be “large enough”
  - insect sensory neuron: 12 -15 ms may be adequate
  - Vertebrate CNS: 100 ms at a minimum
- Up to $2^{(T_{word}/T_{letter})}$ probabilities to be estimated

Multiple neurons: a severe sampling problem

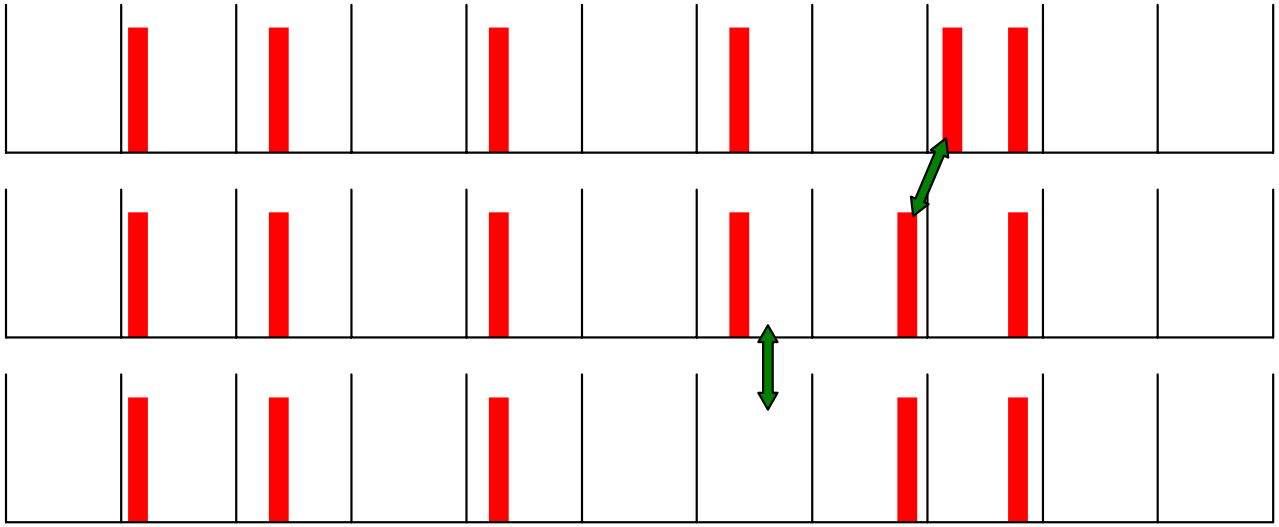
- One dimension for each bin and each neuron



- 2^[L(T_{word} / T_{letter})] probabilities must be estimated.

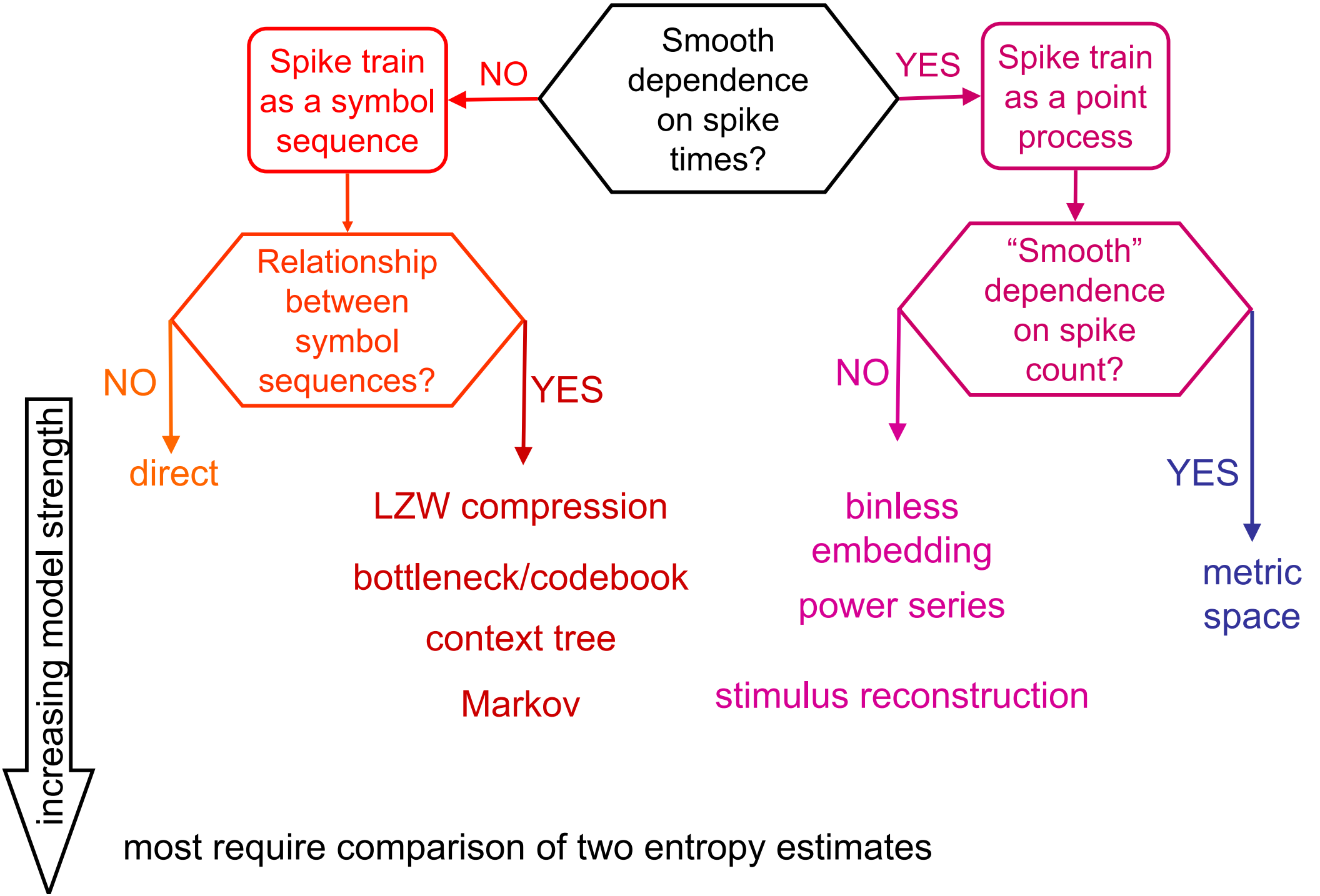
What else can we do?

- Spike trains are events in time
- So there are relationships between them:
  - a continuous topology
  - discrete relationships: how many spikes? (and on which neurons?)



- How can we exploit this?

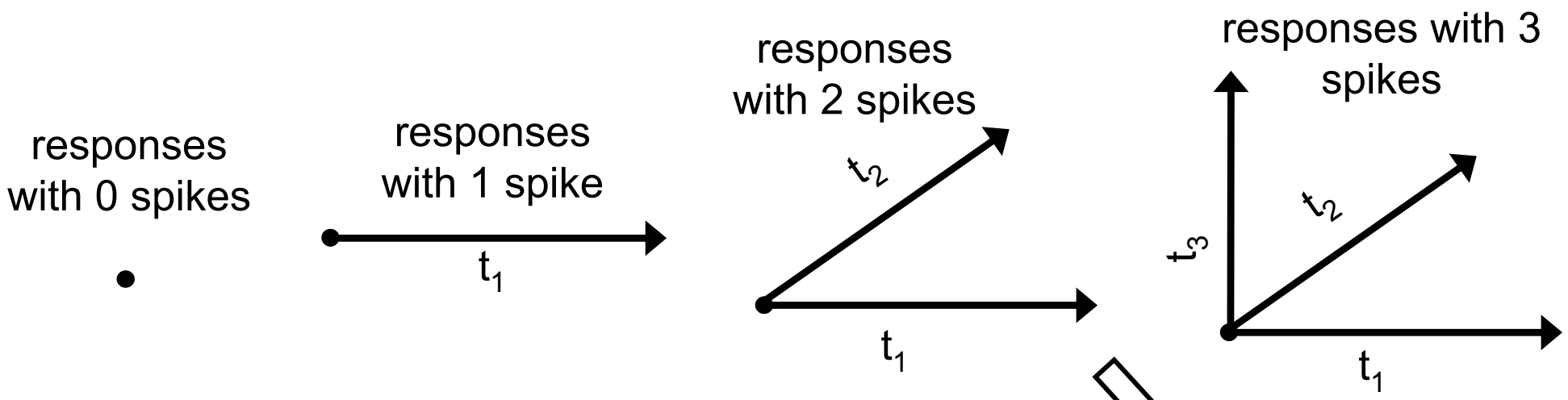
Strategies for Estimating Entropy and Information



most require comparison of two entropy estimates

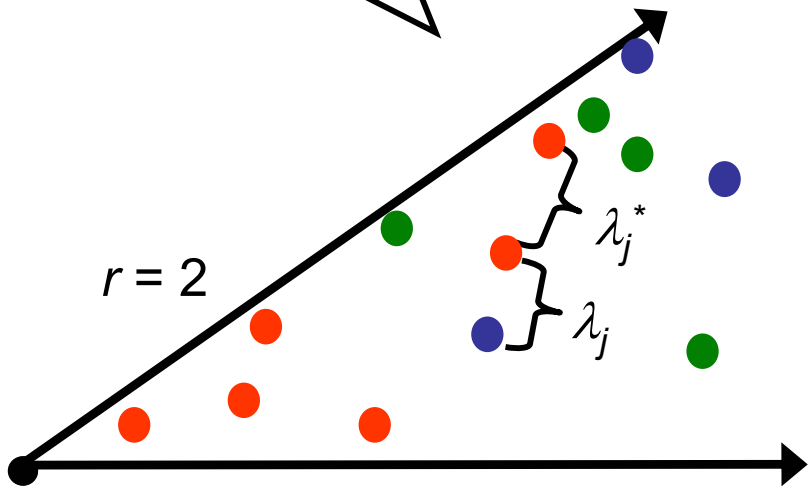
Binless Embedding Method in a nutshell

Embed responses with *r* spikes as points in an *r*-dimensional space



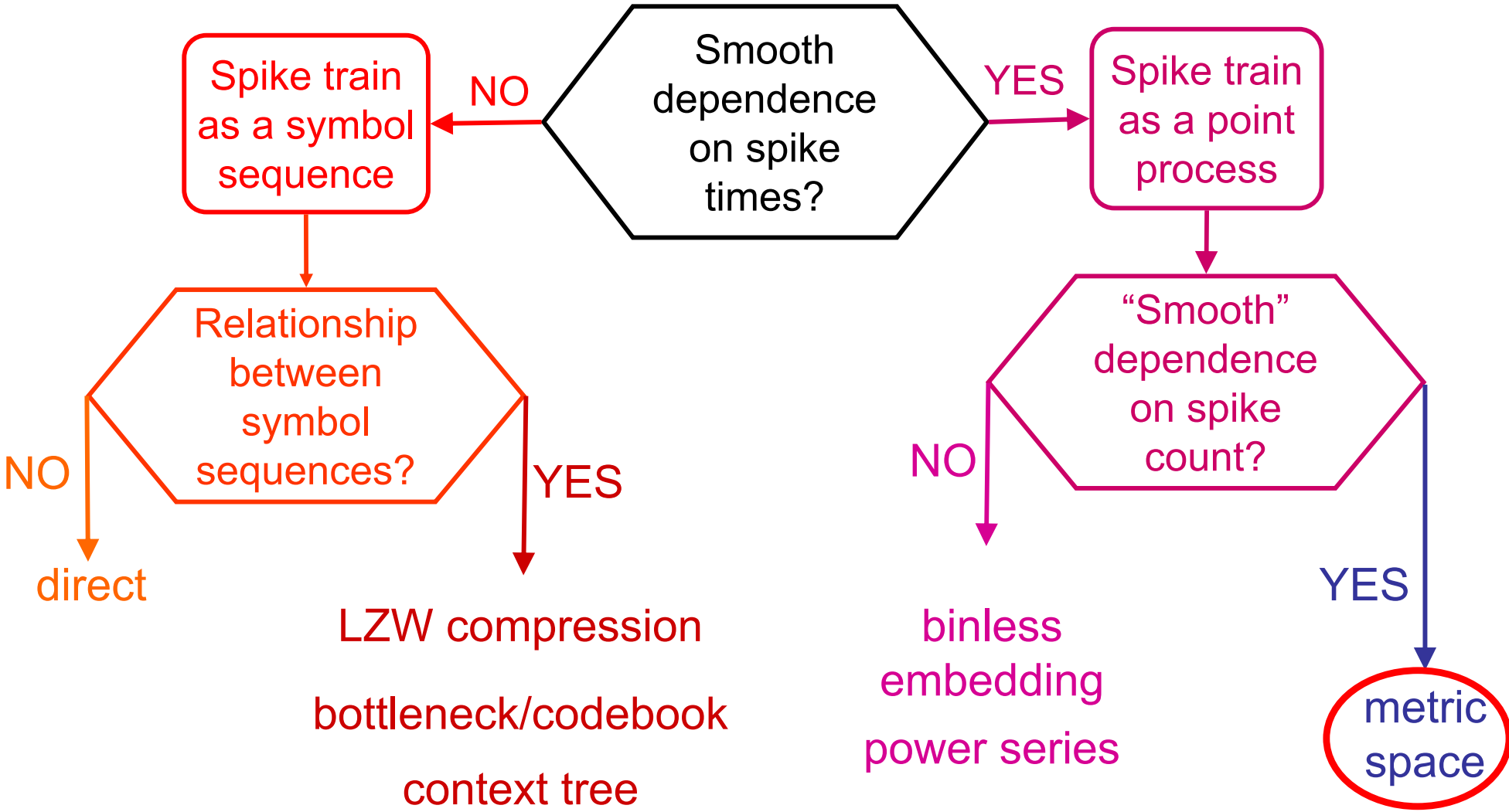
In each r-dimensional stratum, use Kozachenko-Leonenko (1987) nearest-neighbor estimate

$$I \approx \frac{r}{N \ln 2} \sum_j \ln\left(\frac{\lambda_j}{\lambda_j^*}\right) - \frac{1}{\ln 2} \sum_k \frac{N_k}{N} \ln \frac{N_k - 1}{N - 1}$$

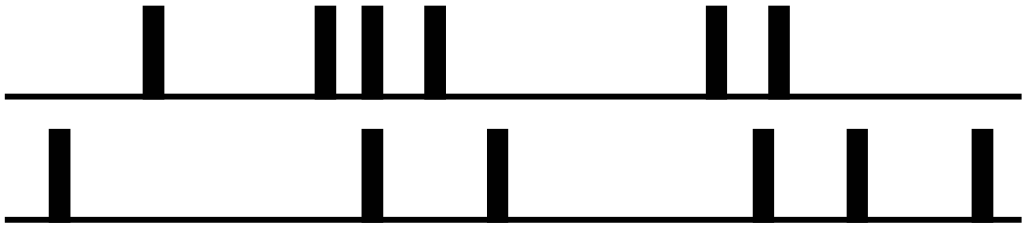


Victor, Phys Rev E (2002)

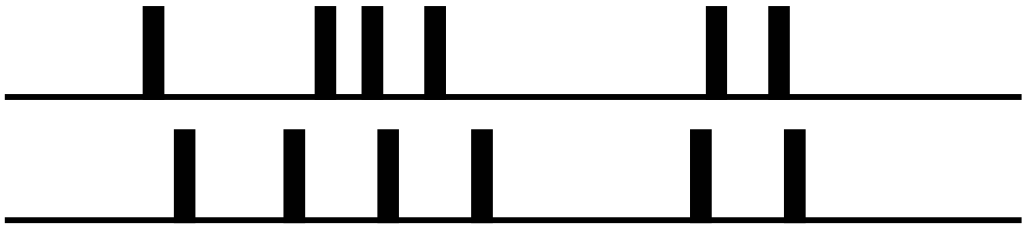
Strategies for Estimating Entropy and Information



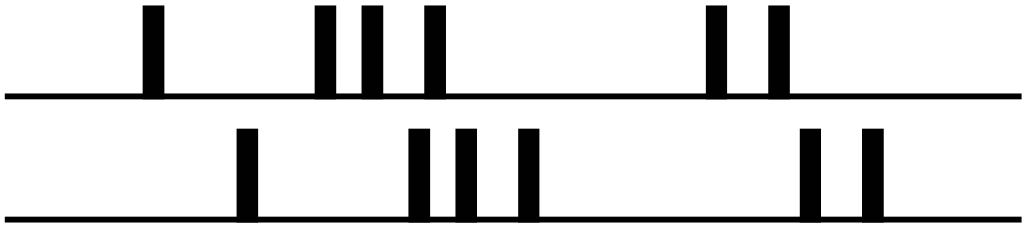
Coding hypotheses: in what ways can spike trains be considered similar?



Similar *spike counts*



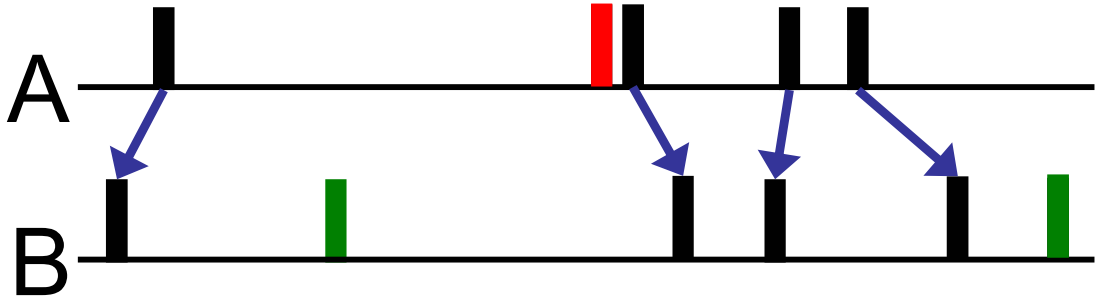
Similar *spike times*



Similar *interspike intervals*

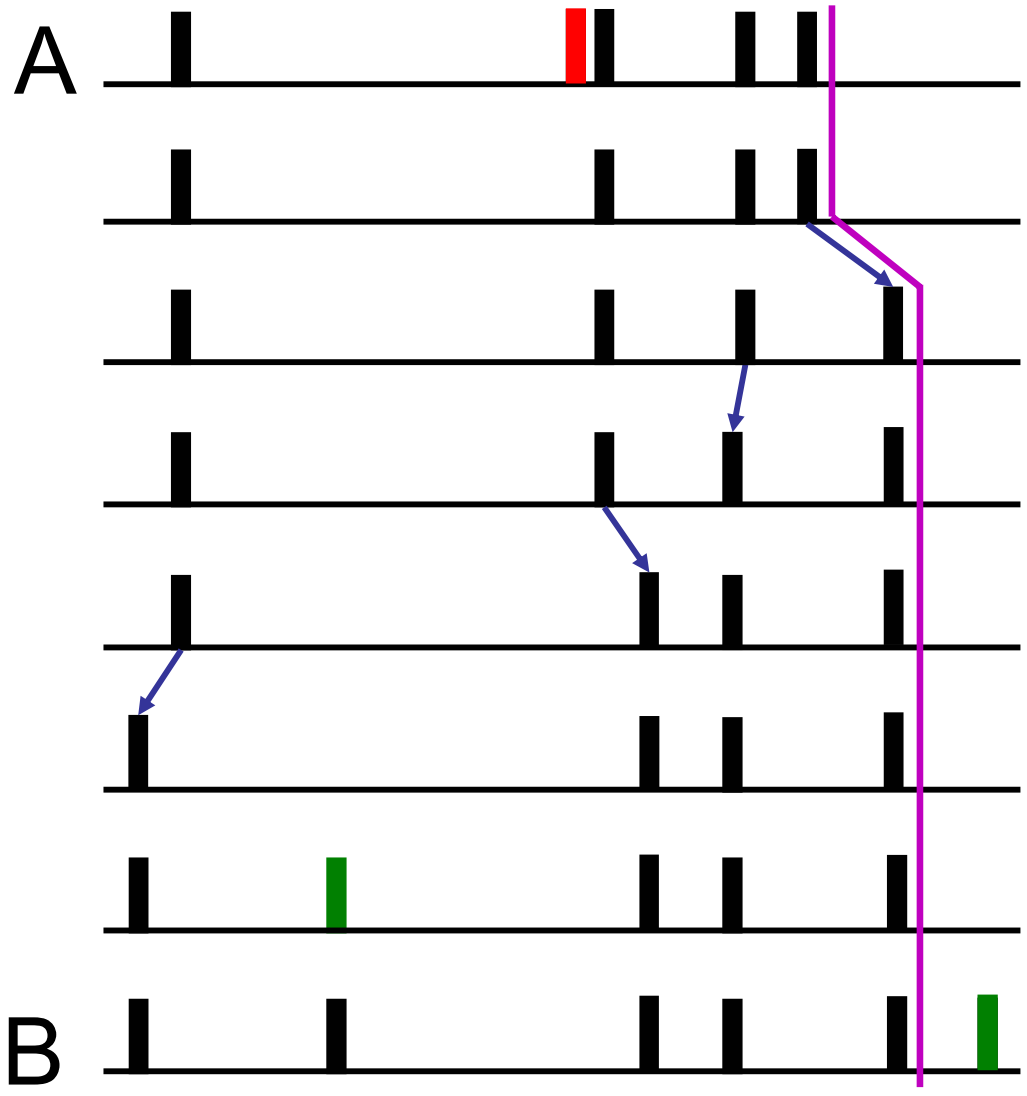
Measuring similarity based on spike times

- Define the “distance” between two spike trains as the simplest morphing of one spike train into the other by inserting, deleting, and moving spikes



- Unit cost to insert or delete a spike
- We don’t know the relative importance of spike timing, so we make it a parameter, q: shift a spike in time by ΔT incurs a cost of $q\,\Delta T$
- Spike trains are similar only if spikes occur at similar times (i.e., within 1/q sec), so q measures the informative precision of spike timing

Identification of Minimal-Cost Paths

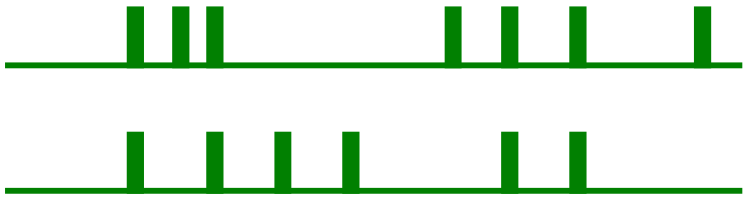


"World lines" cannot cross.
So, either
(i) The last spike in A is deleted,
(ii) The last spike in B is inserted
(iii) The last spike in A and the last spike in B must correspond via a shift

The algorithm is closely analogous to the Needleman-Wunsch & Sellers (1970) dynamic programming algorithms for genetic sequence comparisons.

Distances between all pairs of responses determine a response space

responses to stimulus 1

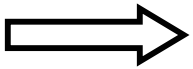


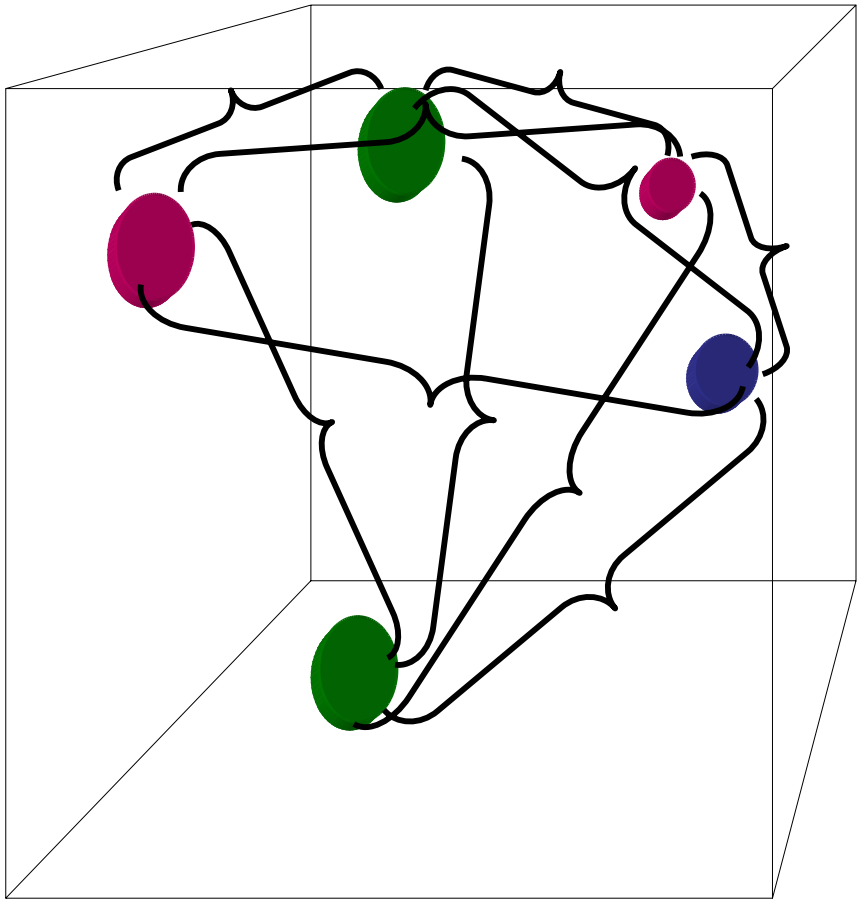
responses to stimulus 2

responses to stimulus 3

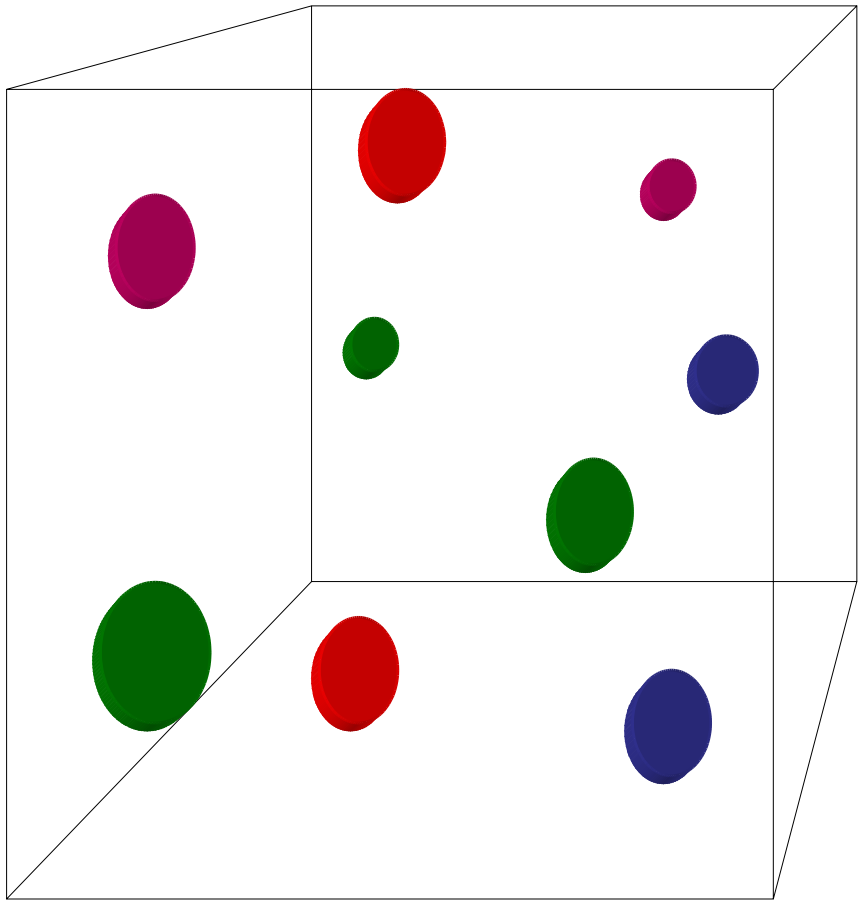
etc.

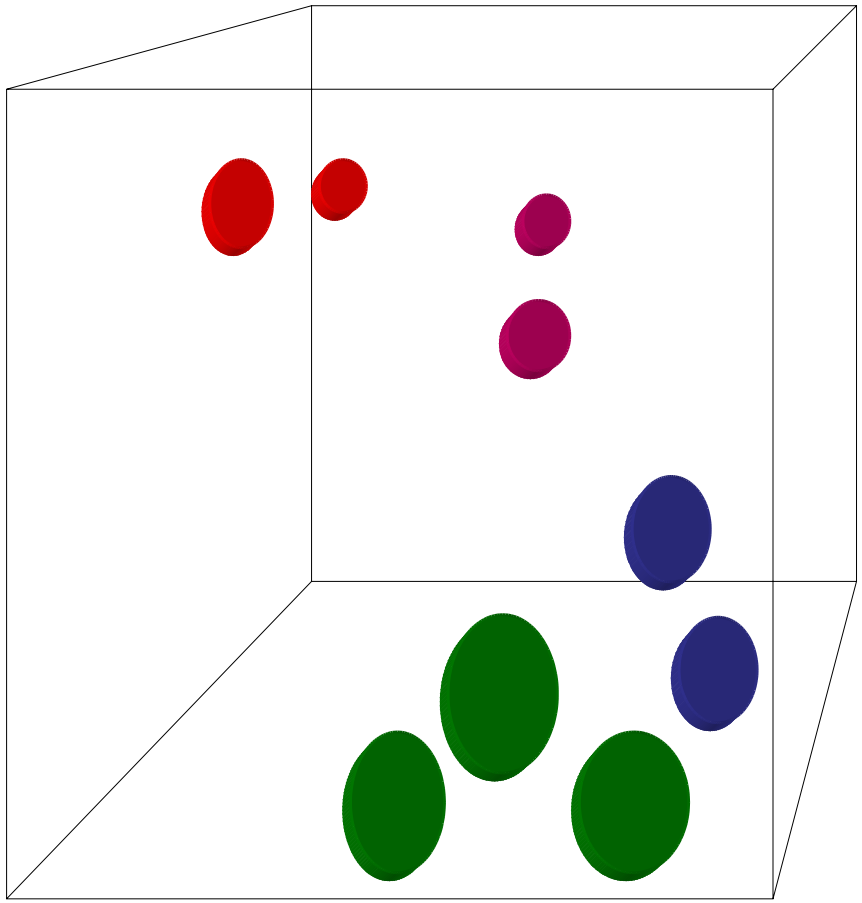
calculate all pairwise distances





Configuration of the response space tests whether a hypothesized distance is viable



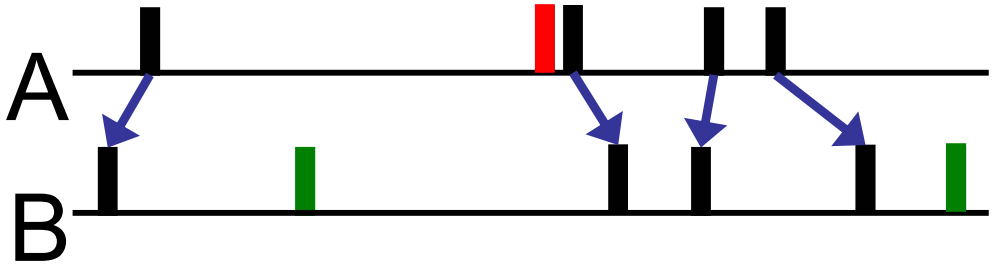


Random: responses to the four stimuli are interspersed

Systematic clustering:
responses to the stimuli are grouped
and
nearby groups correspond to similar stimuli

Metric Space Method in a nutshell

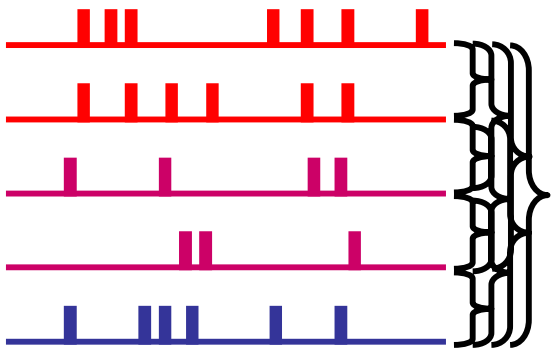
Postulate a parametric family of edit-length metrics (distances") between spike trains

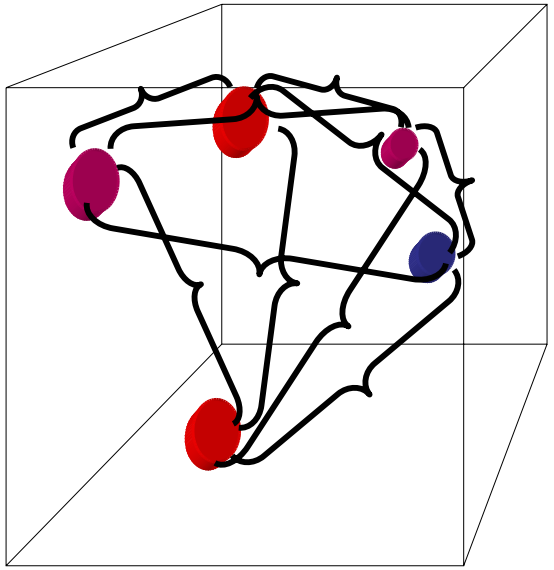


Allowed elementary transformations:

- insert or delete a spike: unit cost
- shift a spike in time by ΔT: cost is $q\,\Delta T$

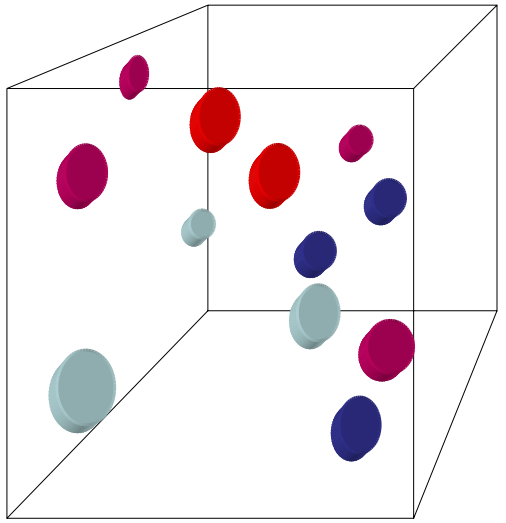
Create a "response space" from the distances

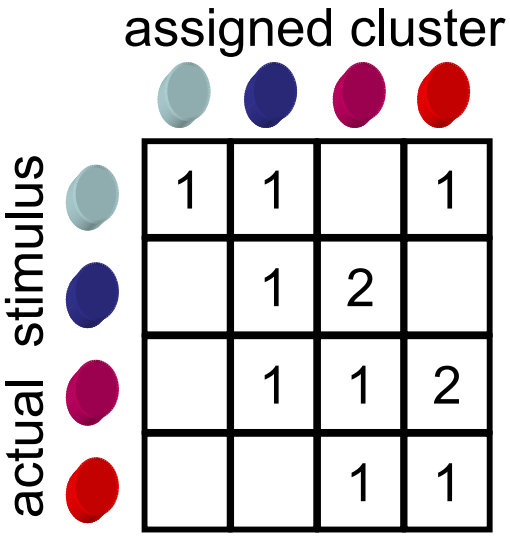




Victor and Purpura, Network (1997)

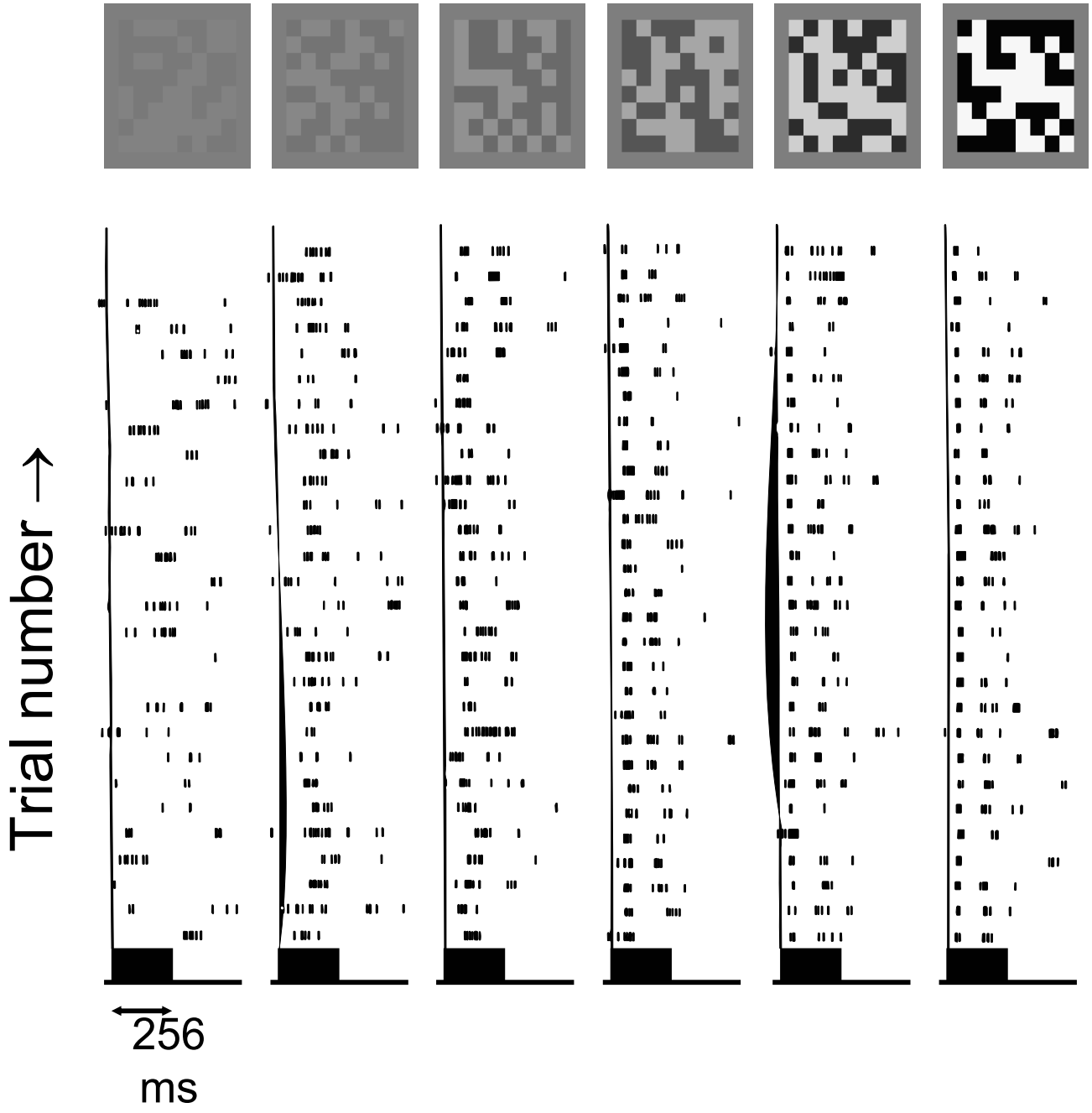
Cluster the responses



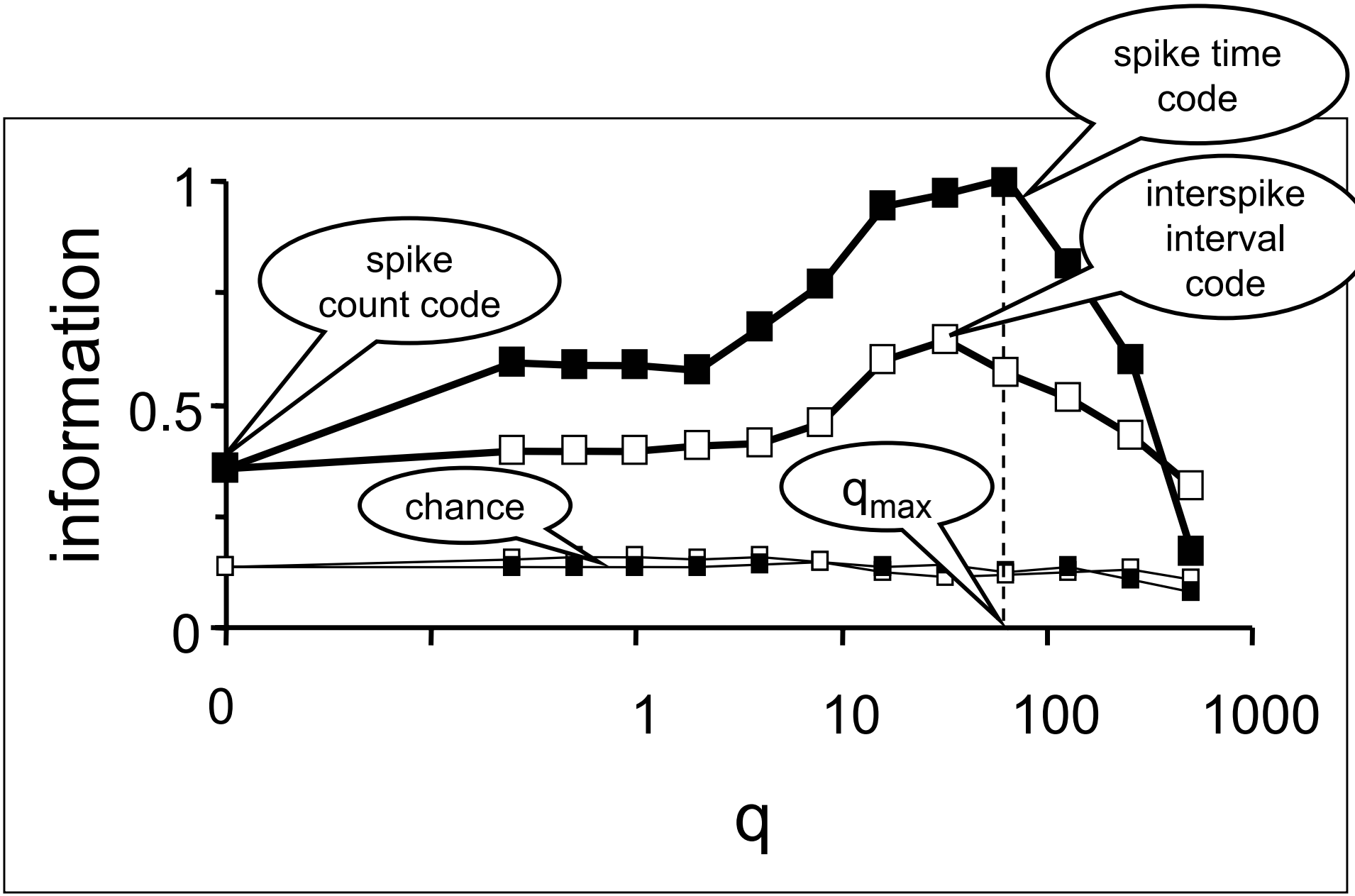


Information = row entropy + column entropy - table entropy

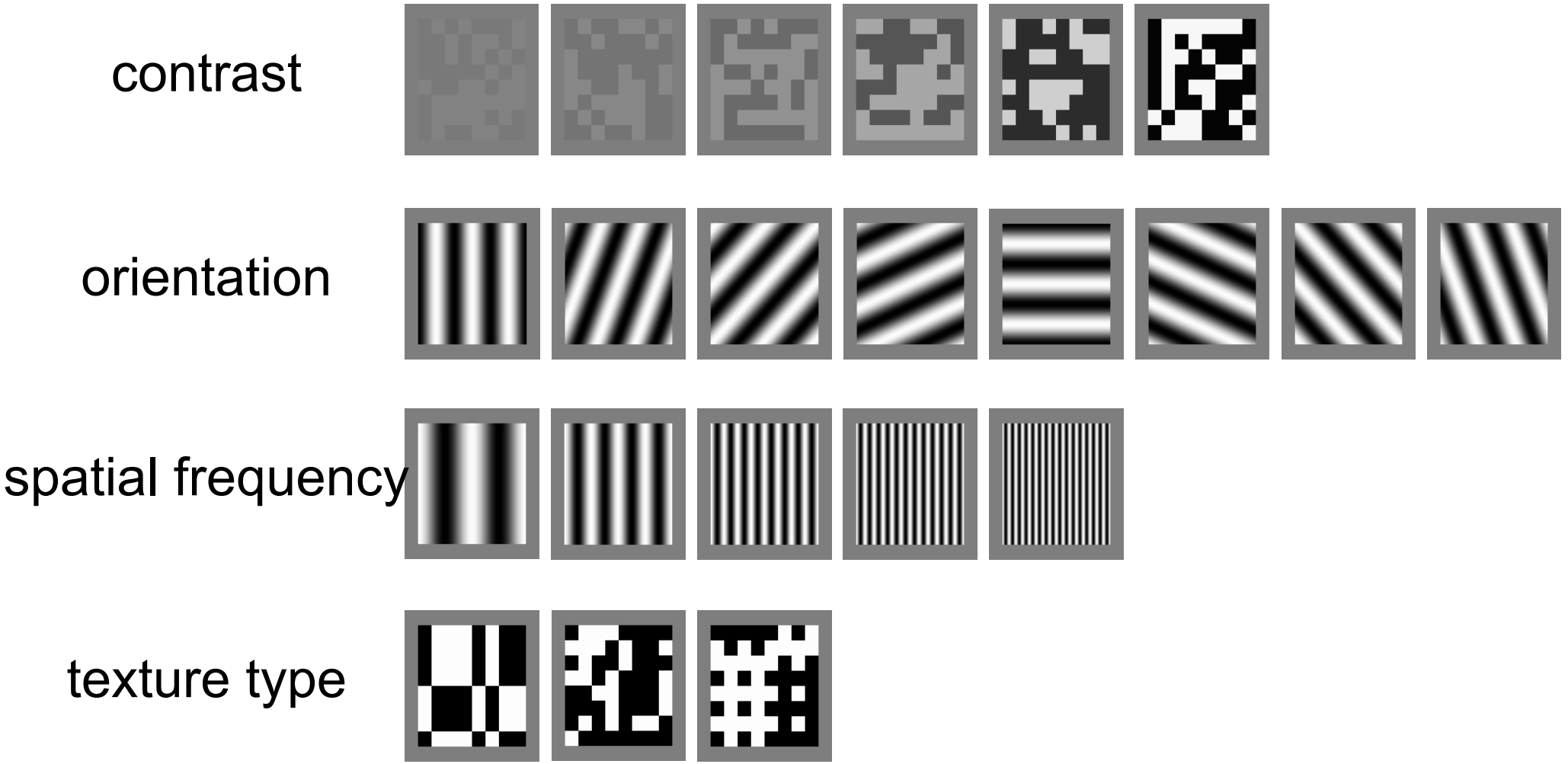
Visual cortex: contrast responses



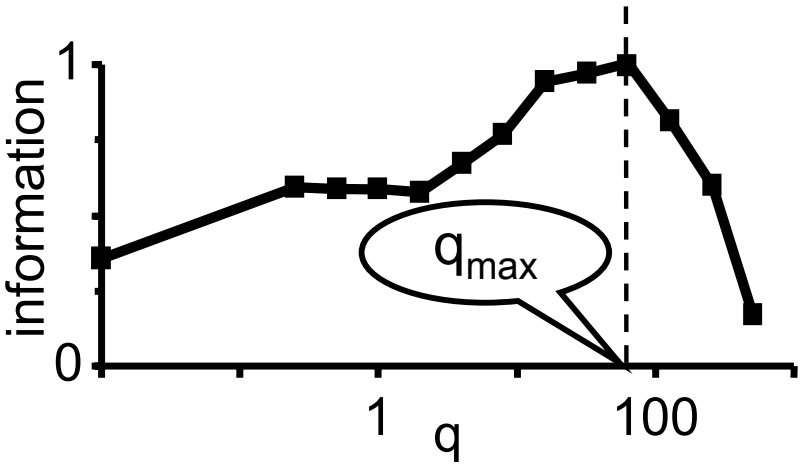
Visual cortex: contrast coding

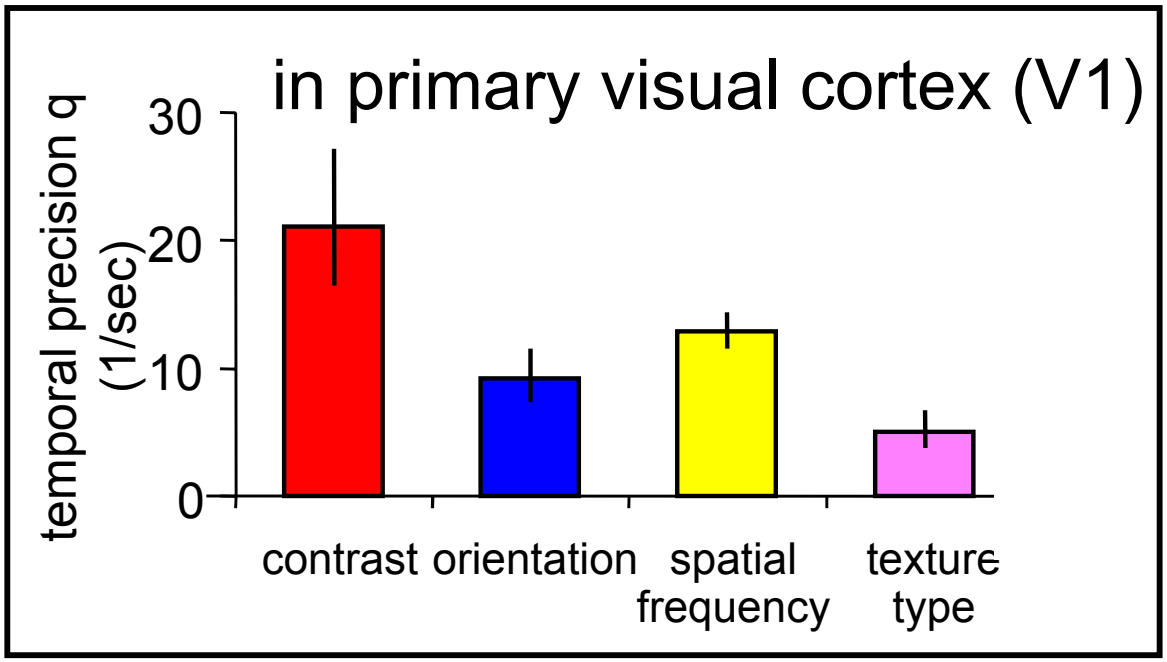


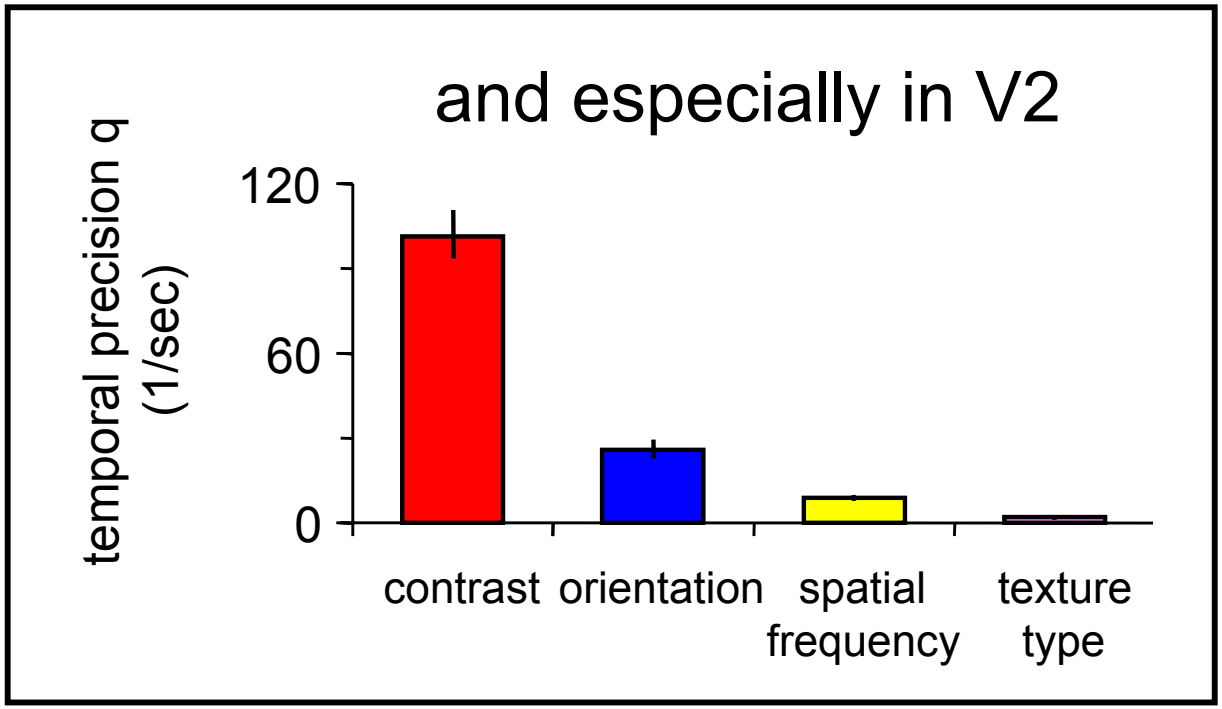
Multiple visual sub-modalities



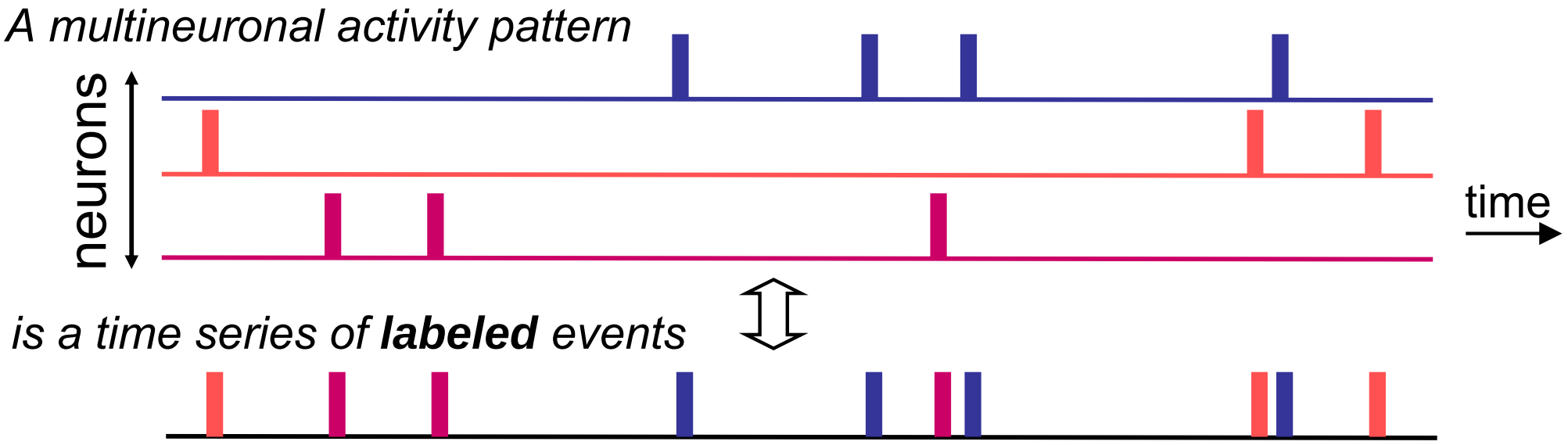
Attributes are coded in distinct ways







Analyzing coding across multiple neurons



Distances between labeled time series can also be defined as the minimal cost to morph one into another, with one new parameter:

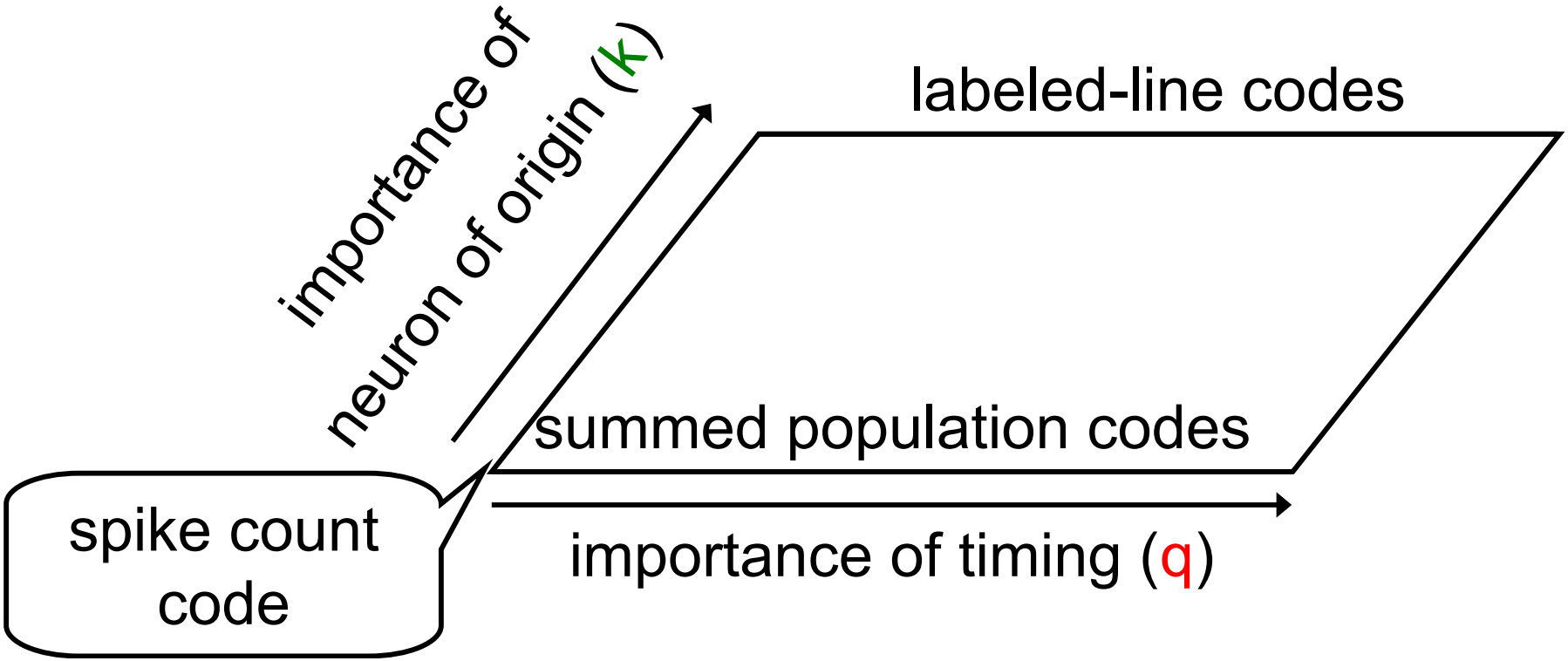
- Cost to insert or delete a spike: 1
- Cost to move a spike by an amount ΔT: $q\,\Delta T$
- Cost to change the label of a spike: k

k determines the importance of the neuron of origin of a spike.

- k=0: summed population code
- k large: labeled line code

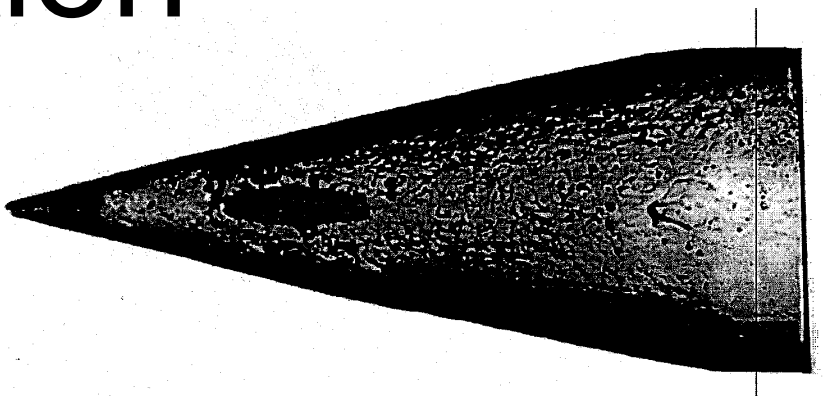
Multineuronal Analysis via the Metric-Space Method: A two-parameter family of codes

- Change the time of a spike: cost/sec = q
  - q=0: spike count code
- Change the neuron of origin of a spike: cost = k
  - k=0: summed population code : (neuron doesn't matter)
  - k=2: labelled line code (neuron matters maximally)

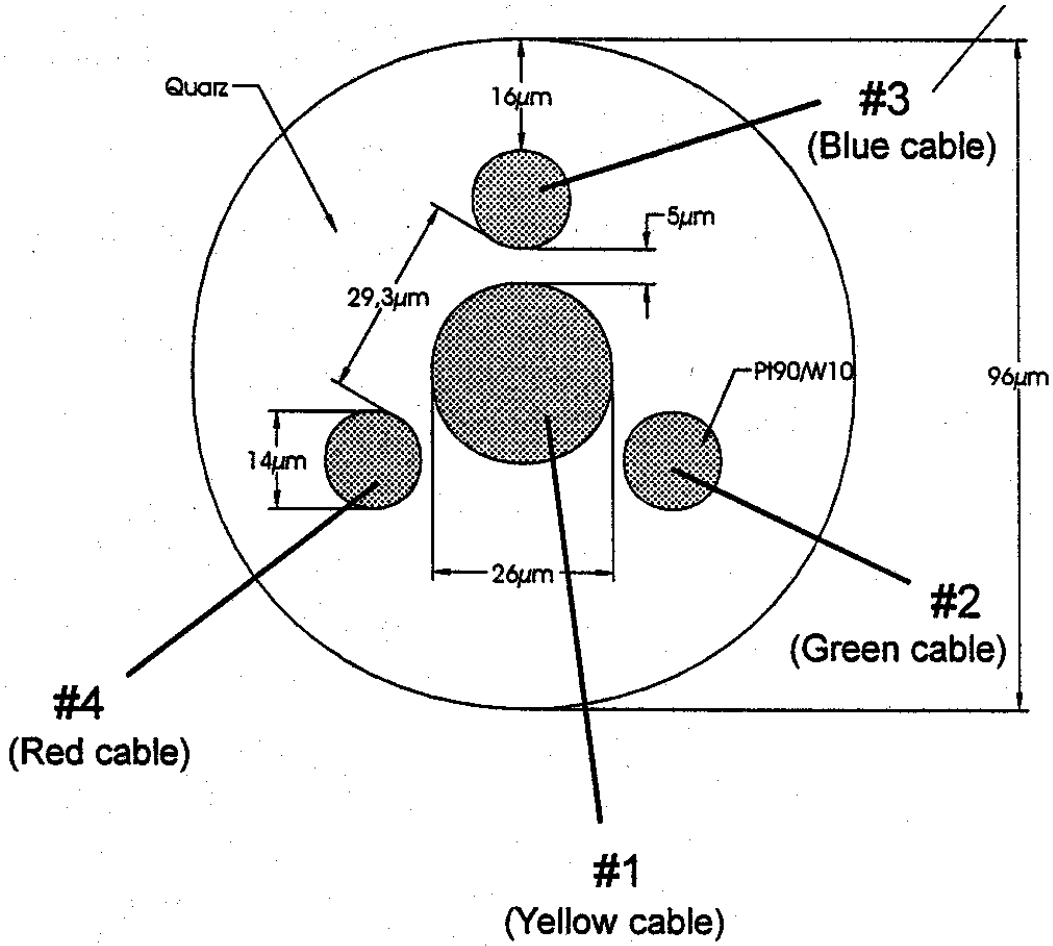


Preparation

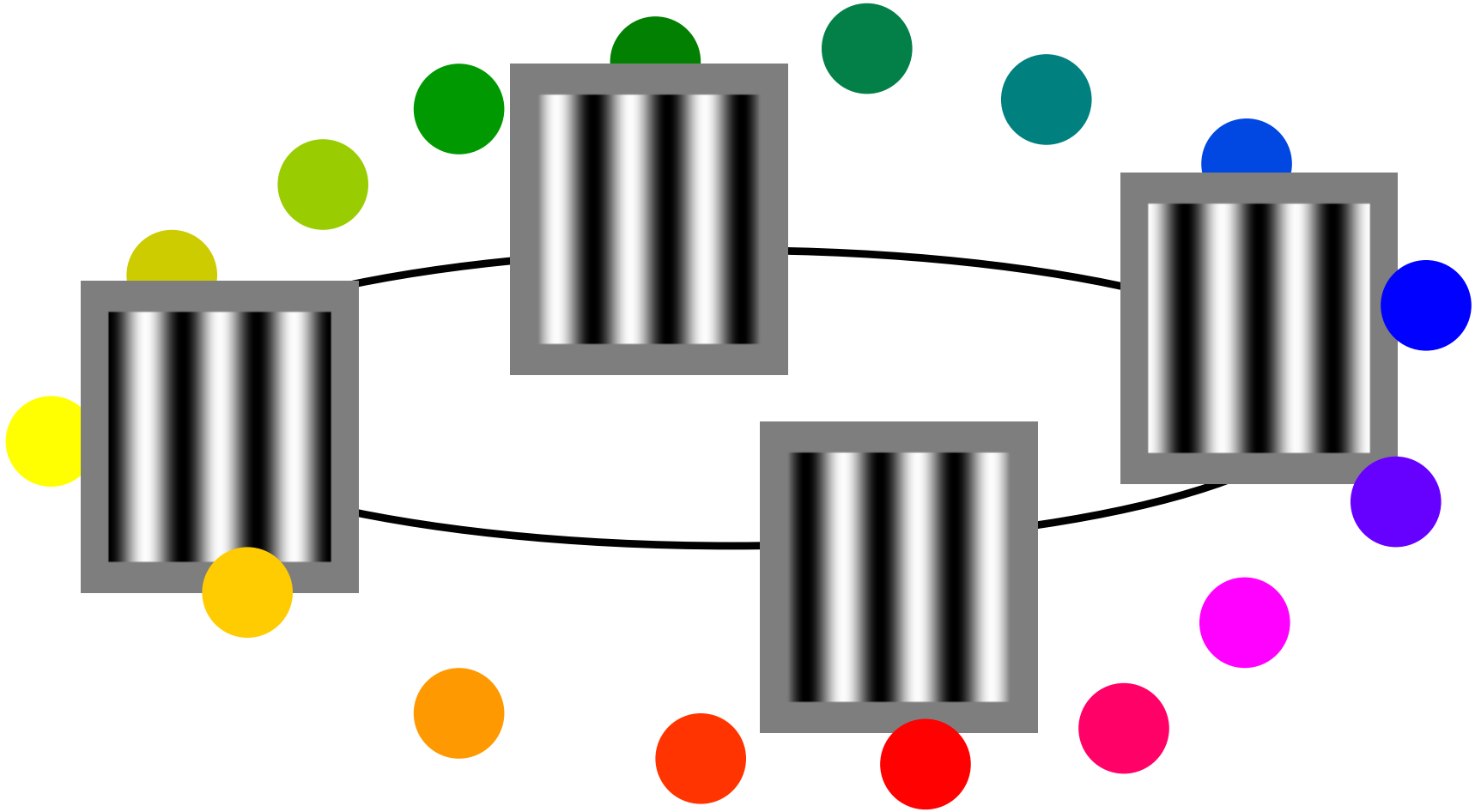
- Recordings from primary visual cortex (V1) of macaque monkey



- Multineuronal recording via tetrodes
  - ensures neurons are neighbors (ca. 100 microns)

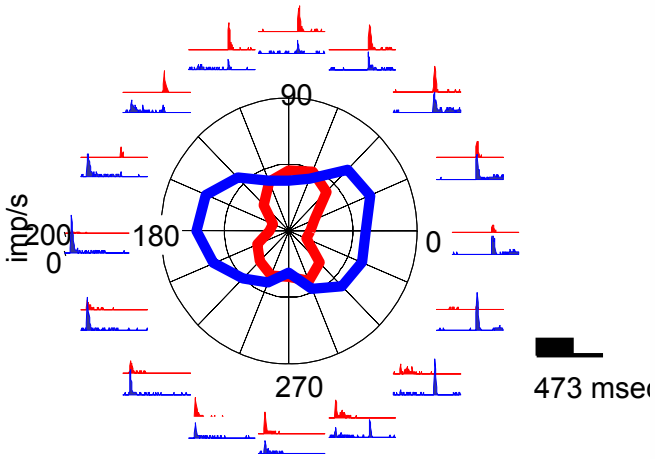


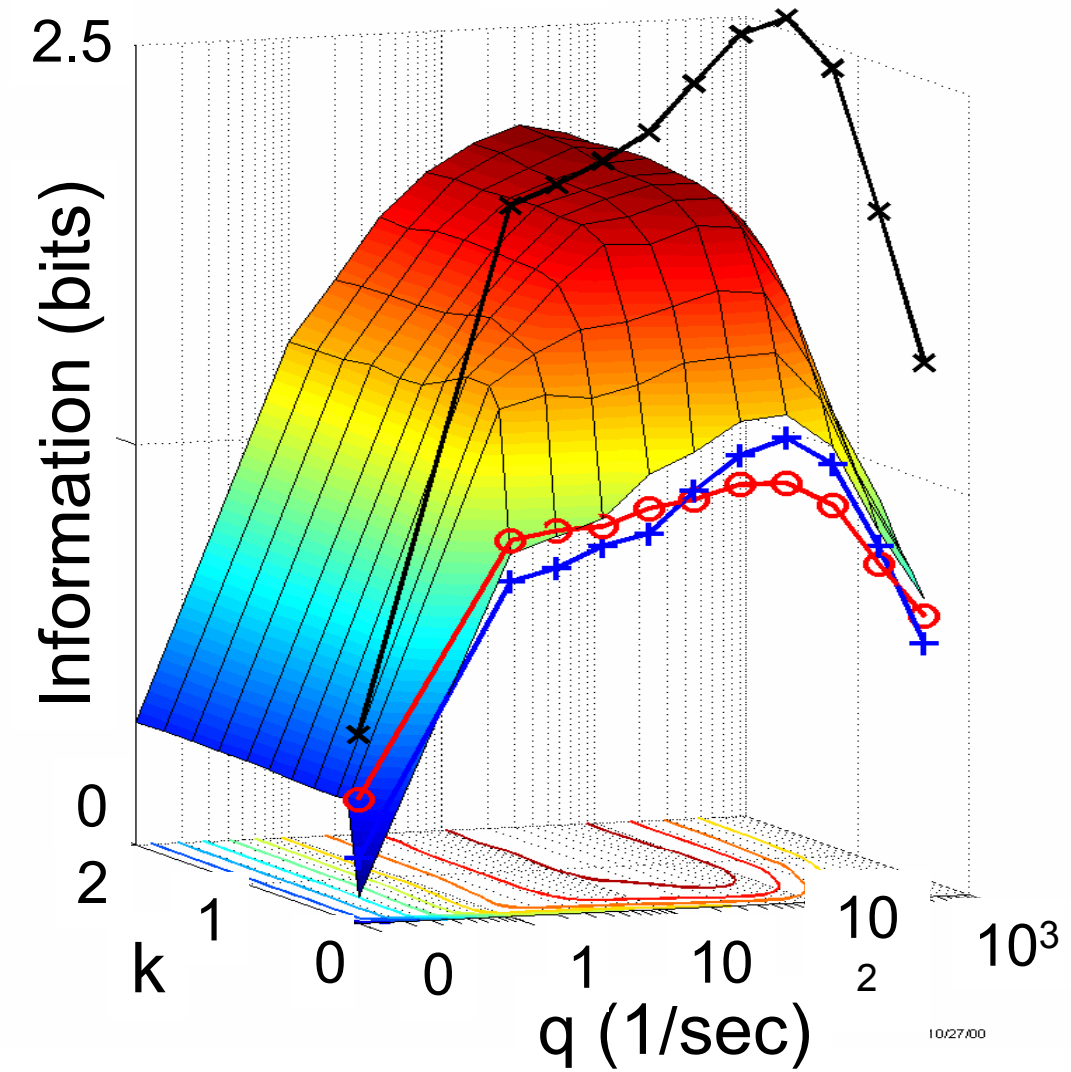
The stimulus set: a cyclic domain



16 kinds of stimuli in the full stimulus set

Spatial phase coding in two simple cells





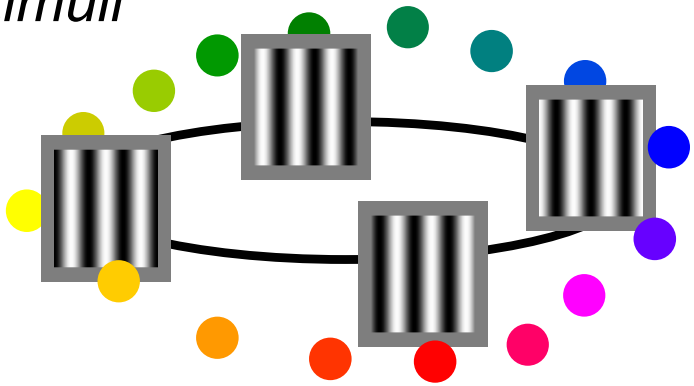
Discrimination is not enough

- *Information* indicates discriminability of stimuli, but not whether stimuli are *represented* in a manner amenable to later analysis

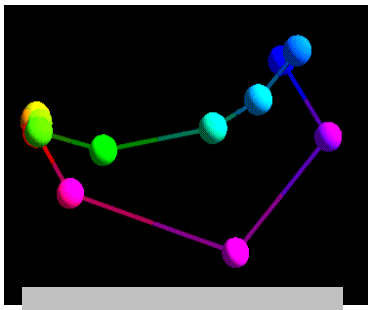
- Do the distances between the responses recover the geometry of the stimulus space?

Representation of phase by a neuron pair

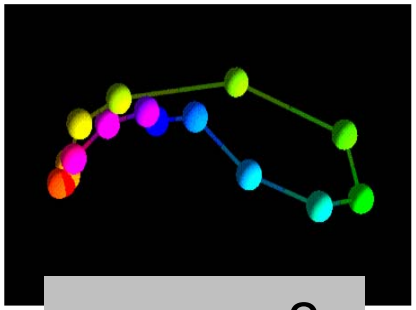
stimuli



reconstructed response space: each neuron considered separately

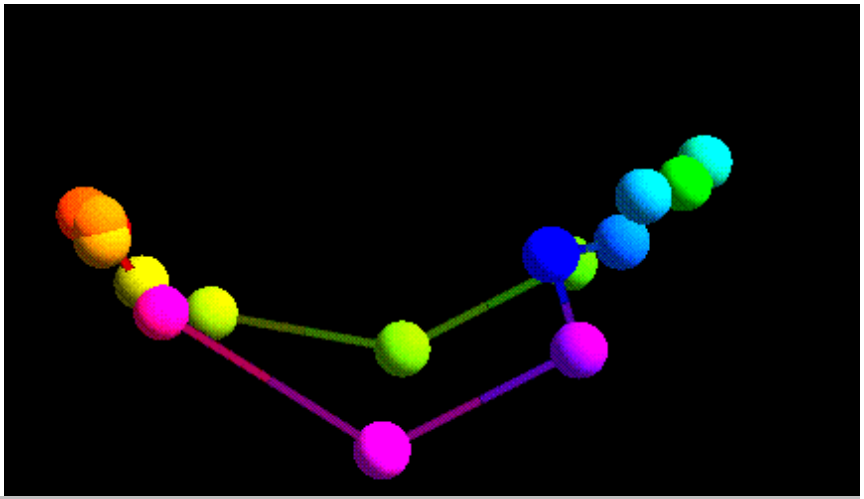


neuron 1

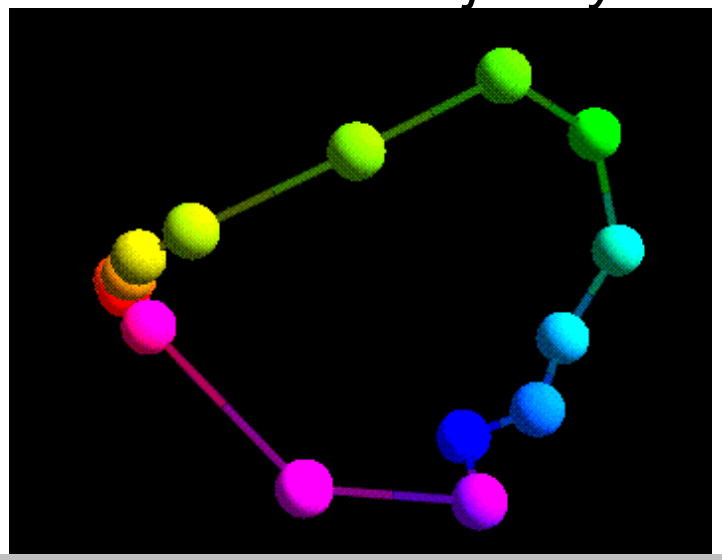


neuron 2

reconstructed response space: two neurons considered jointly



summed population code (k=0)



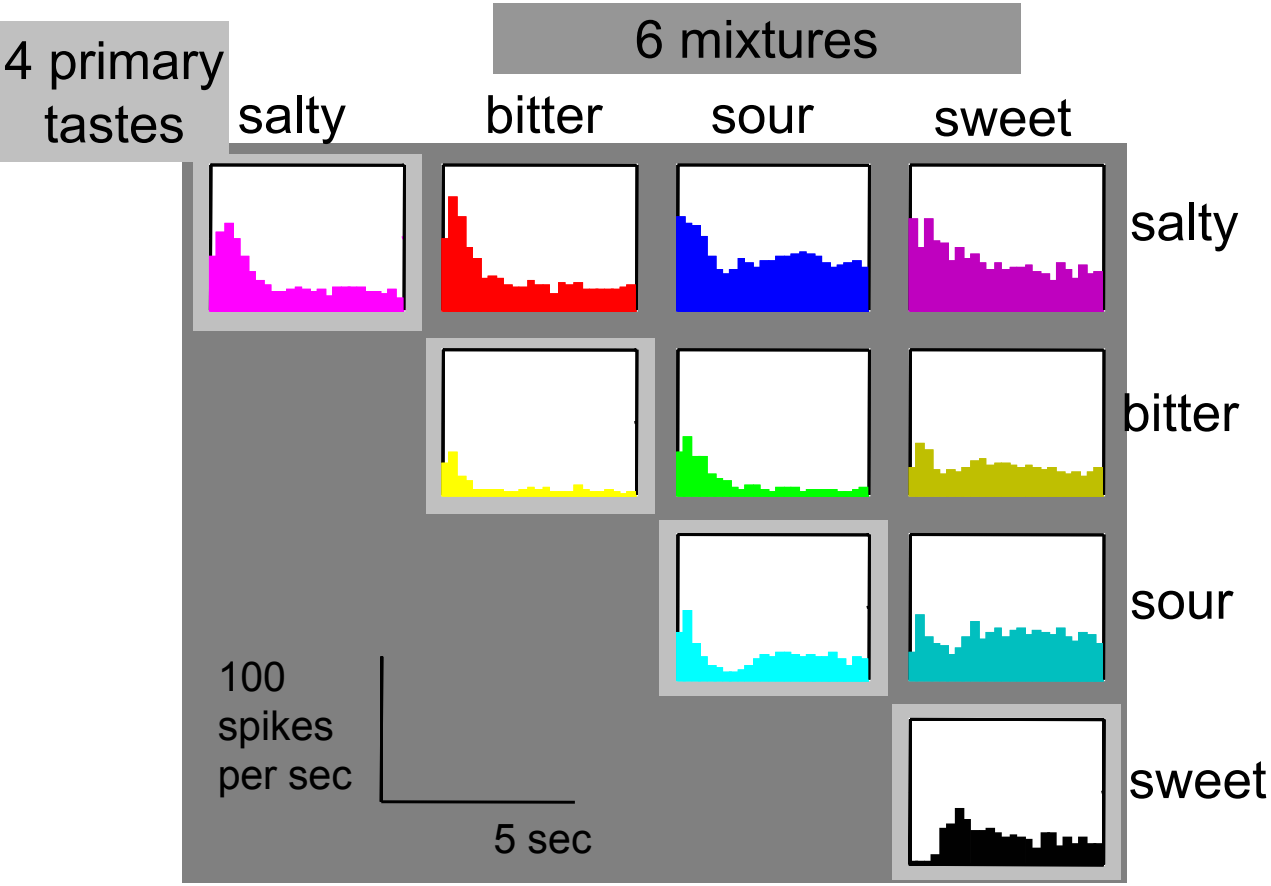
respect neuron of origin (k=1)

Representation of stimulus space is more faithful when neuron-of-origin of each spike is respected.

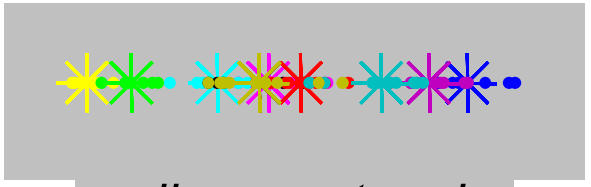
Representing Taste

responses of rat solitary tract neuron

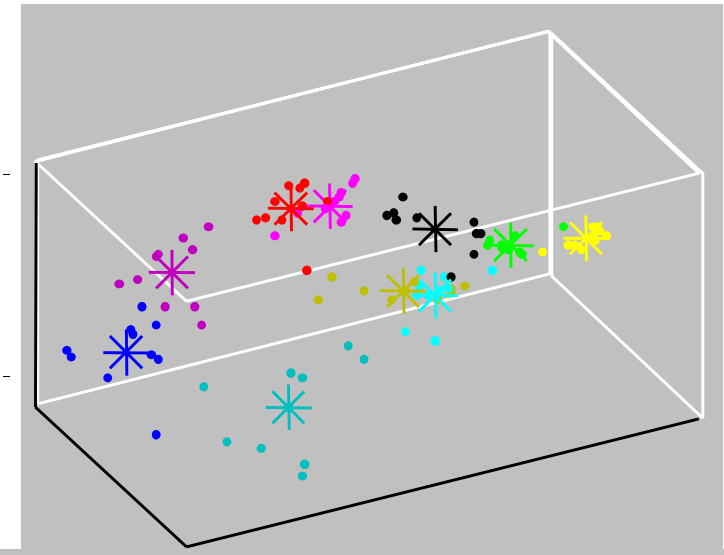
P. DiLorenzo, J.-Y. Chen, SfN 2007



reconstructed response spaces



spike count code



spike timing code,
informative precision q ~ 200 msec

Temporal pattern supports full representation of the 4 primary tastes and their 6 mixtures

Summary: Estimation of Information-Theoretic Quantities from Data

- It is challenging but approachable
- There is a role for multiple methodologies
  - Methodologies differ in assumptions about how the response space is partitioned
  - The way that information estimates depend on this partitioning provides insight into neural coding
- It can provide some biological insights
  - The temporal structure of spike trains is informative
  - (Adding parametric tools): temporal structure can represent the geometry of a multidimensional sensory space

Maximum-Entropy Methods

- The maximum-entropy criterion chooses a specific distribution from an incomplete specification
  - Typically unique: the most random distribution consistent with those specifications
  - Often easy to compute (but not always!)
- Maximum-entropy distributions
  - General characteristics
  - Familiar examples
  - Other examples
- Applications

General Characteristics of Maximum-Entropy Distributions

- Setup
  - Seek a probability distribution $p(x)$ on a specified domain
  - Specify constraints $C_1, \ldots, C_M$ of the form $\Sigma C_m(x)p(x)=b_m$
  - Domain can be continuous; sums become integrals
  - Constraints are linear in p but may be nonlinear in x
    - $C(x)=x^2$ constrains the variance of p
  - Maximize $H(p)=-\Sigma p(x)\log p(x)$ subject to these constraints
- Solution
  - Add a constraint $C_0=1$ to enforce normalization of p
  - Lagrange multipliers λ_m for each C_m
  - Maximize $-\Sigma p(x)\log p(x)+\Sigma\lambda_m\Sigma C_m p(x)$
  - Solution: $p(x)=\exp\{\Sigma\lambda_m C_m(x)\}$, with the multipliers λ_m are determined by $\Sigma C_m(x)p(x)=b_m$
    - These equations are typically nonlinear, but occasionally can be solved explicitly
    - The solution is always unique (mixing property of entropy)

Examples of Maximum-Entropy Distributions

- Univariate
  - On [a, b], no constraints: uniform on [a, b]
  - On [-∞,∞], variance constrained: $K\exp(-cx^2)$
  - On [0, ∞], mean constrained: $K\exp(-cx)$
  - On [0, 2π], Fourier component constrained: Von Mises distribution, $K\exp\{-c\cos(x-\phi)\}$
- Multivariate
  - Mean and covariances constrained: $K\exp\{(\boldsymbol{x}-\boldsymbol{m})^{\mathsf{T}}\boldsymbol{C}(\boldsymbol{x}-\boldsymbol{m})\}$
  - Independent, with marginals $P(x)$, $Q(y)$: $P(x)Q(y)$
- Discrete
  - On {0,1} on a 1-d lattice, with m-block configurations constrained: the mth-order Markov processes
  - On {0,1} on a 2-d lattice, with adjacent pair configurations constrained: the Ising problem

Wiener/Volterra and Maximum-Entropy

- Wiener-Volterra systems identification: Determine the nonlinear functional F in a stimulus-response relationship $r(t)=F[s(t)]$
  - Discretize prior times: $s(t)=(s(t-\Delta t), \ldots, s(t-L\Delta t))= (s_1, \ldots, s_L)$
  - Consider only the present response: $r=r(0)$
  - F is a multivariate Taylor series for r in terms of $(s_1, \ldots, s_L)$
  - Measure low-order coefficients, assume others are 0
- Probabilistic view
  - Same setup as above, find $P(r,s)=P(r, s_1, \ldots, s_L)$
  - Measure some moments
    - Mean and covariance of s: $\langle s_j \rangle$, $\langle s_j s_k \rangle$,…
    - Cross-correlations: $\langle r s_j \rangle$, $\langle r s_j s_k \rangle$,…
    - Mean of r: $\langle r \rangle$
    - Variance of r: $\langle r^2 \rangle$
  - Find the maximum-entropy distribution constrained by the above
  - This is $K\exp\{-c(r-F(s_1,\ldots,s_L))^2\}$

- The Wiener-Volterra series with additive and Gaussian noise *is* the maximum-entropy distribution consistent with the above constraints

- What happens if the noise is not Gaussian (e.g., spikes?)

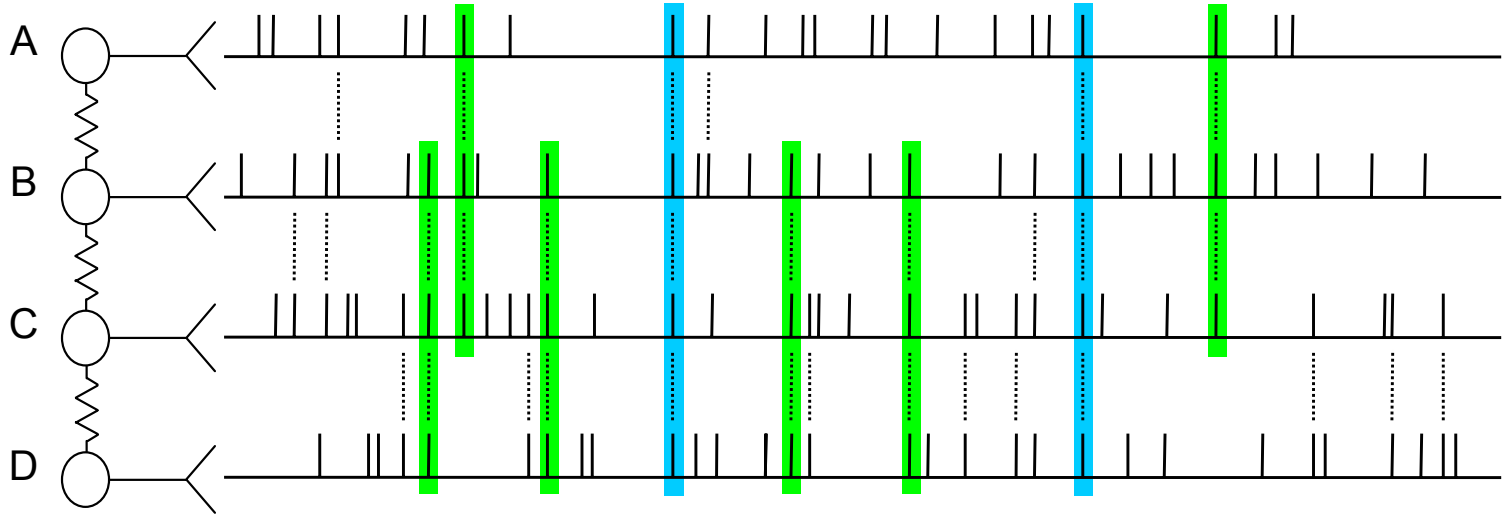
Victor and Johanessma, 1986

Some Experiment-Driven Uses of Maximum-Entropy Methods

- Rationale: since it is the most random distribution consistent with a set of specifications, it can be thought of as an automated way to generate null hypotheses
- Modeling multineuronal activity patterns
  - Spontaneous activity in the retina: Shlens et al. 2006, Schneidman et al. 2006
  - Spontaneous and driven activity in visual cortex: Ohiorhenuan et al. 2007
- Designing stimuli

Analyzing Multineuronal Firing Patterns

Consider three neurons A, B, C. If all pairs are uncorrelated, then we have an obvious prediction for the triplet firing event, namely $p(A,B,C)=p(A)p(B)p(C)$. But if we observe pairwise correlations, what do we predict for $p(A,B,C)$, and higher-order patterns?



Use constrained maximum-entropy distributions, of course.

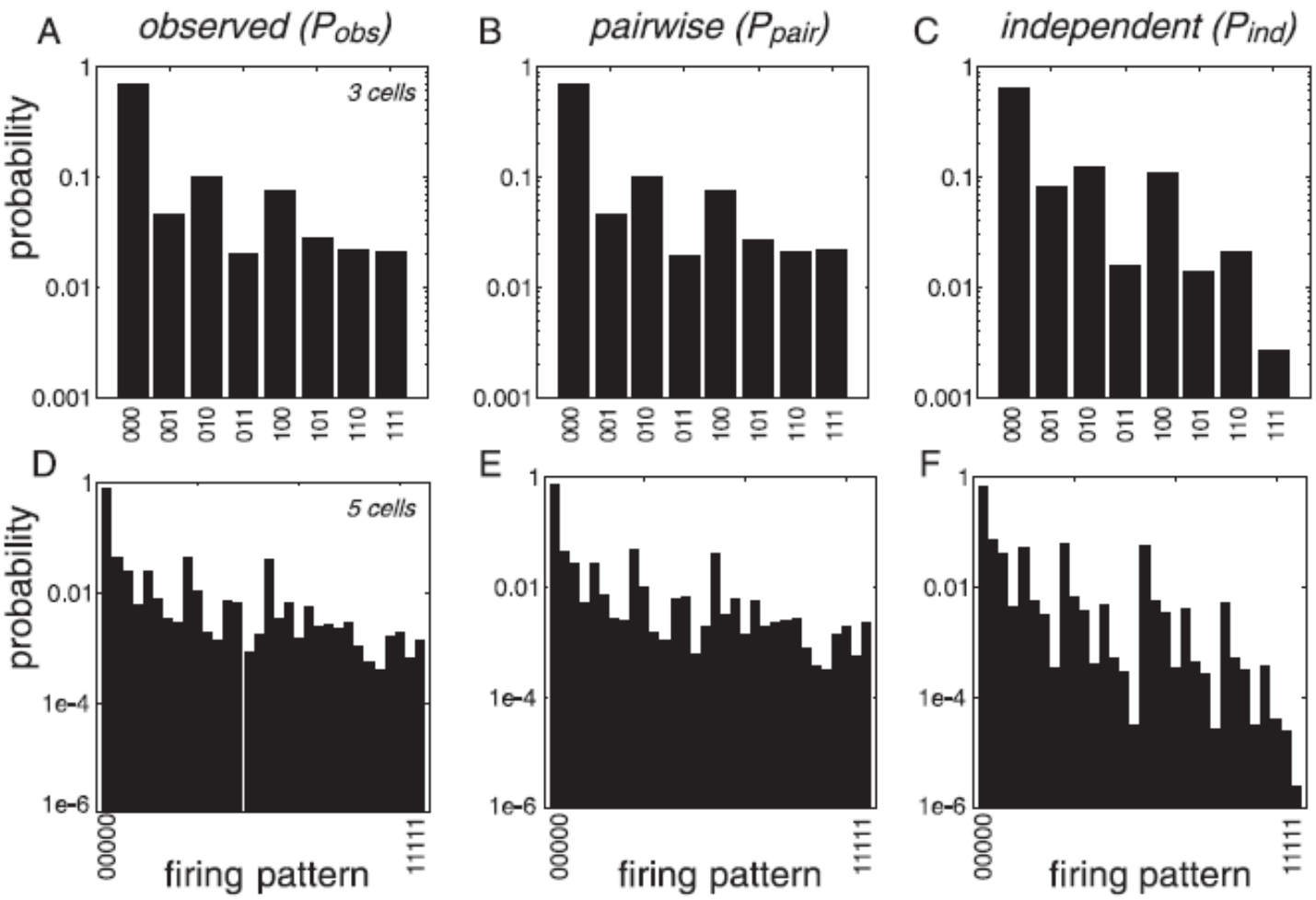
Measure $p(\begin{smallmatrix}1&\\&\end{smallmatrix}), p(\begin{smallmatrix}&1\\&\end{smallmatrix}), p(\begin{smallmatrix}&\\1&\end{smallmatrix}), p(\begin{smallmatrix}&\\&1\end{smallmatrix}), p(\begin{smallmatrix}1&1\\&\end{smallmatrix}), p(\begin{smallmatrix}1&\\1&\end{smallmatrix}), \ldots$

Find the maximum-entropy distribution, and use to predict

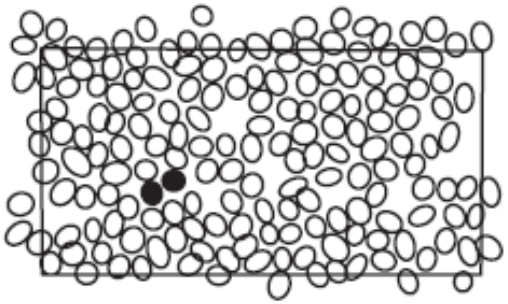
$p(\begin{smallmatrix}1&1\\1&\end{smallmatrix})$, $p(\begin{smallmatrix}1&\\1&1\end{smallmatrix})$, $p(\begin{smallmatrix}1&1\\&1\end{smallmatrix})$, $p(\begin{smallmatrix}&1\\1&1\end{smallmatrix})$, $p(\begin{smallmatrix}1&1\\1&1\end{smallmatrix})$.

Multineuronal Firing Patterns: Retina

Maximum-entropy distributions built from pairs account for joint activity of clusters of up to 7 cells (macaque retina, Shlens et al., J. Neurosci. 2006)



And only nearest-neighbor pairs matter.

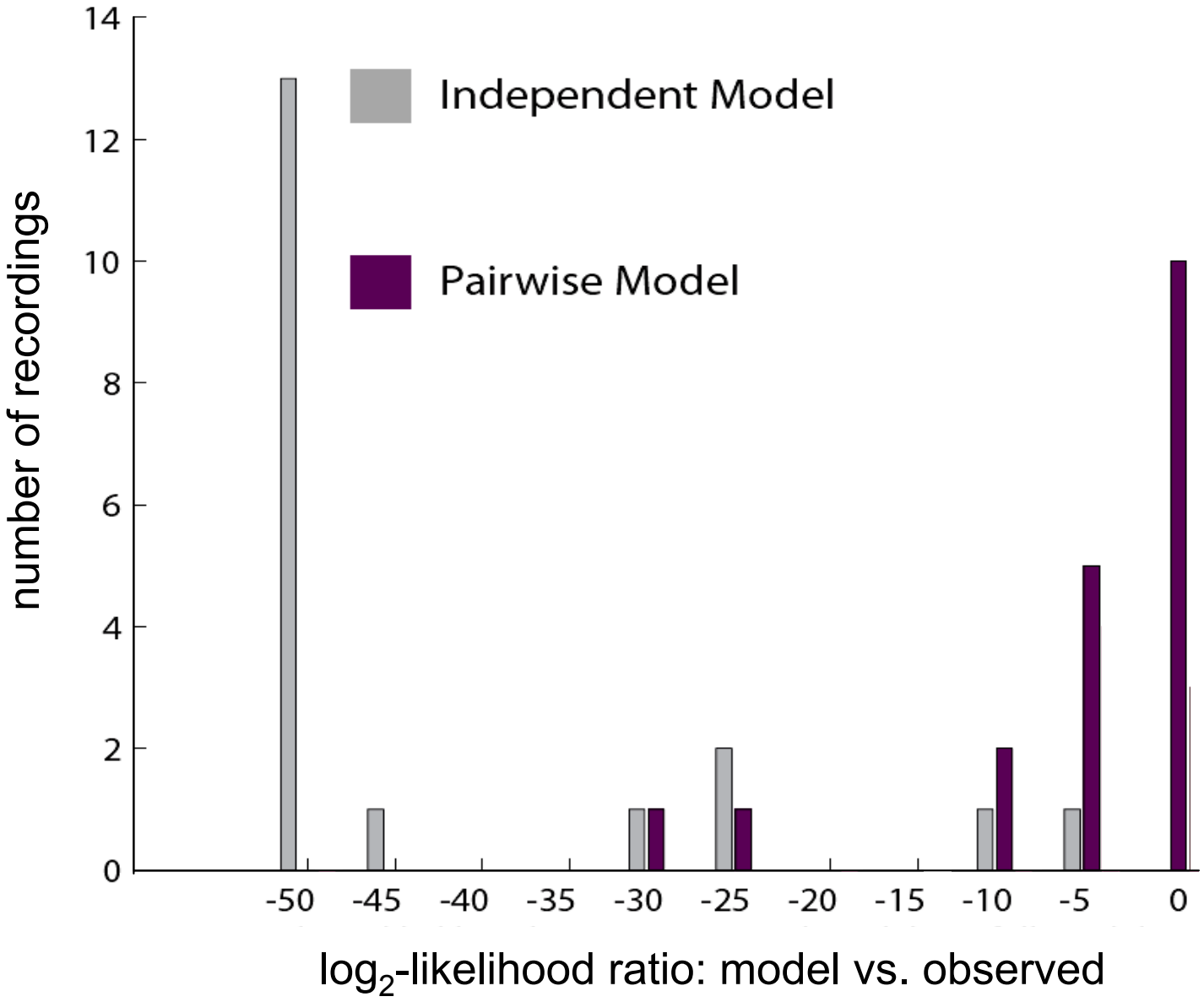


Also see Schneidmat et al., Nature 2006 (15 cells), and review in Current Opinion In Neurobiology, 2007 (Nirenberg and Victor)

Significance of Maximum-Entropy Analysis

- The pairwise model dramatically simplifies the description of multineuronal firing patterns
  - Reduction from 2^N parameters to $N(N-1)/2$ parameters
  - Further reduction to ~$4N$ parameters if only nearest neighbors matter
  - It makes sense in terms of retinal anatomy
- Limitations
  - What about firing patterns across time?
  - What about stimulus driving?
  - What about cortex (connectivity is more complex)

Multineuronal Firing Patterns: Visual Cortex



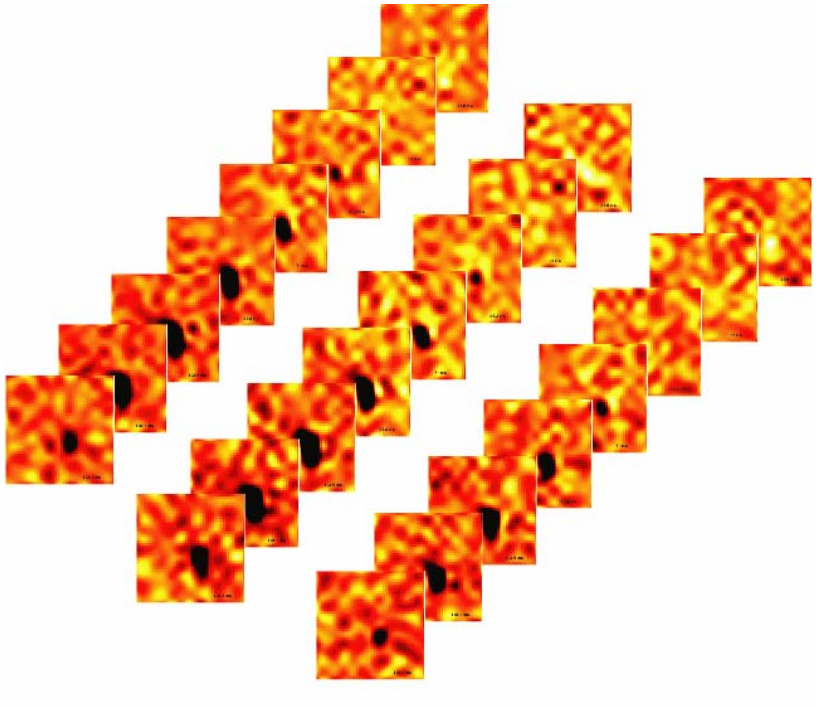
Maximum-entropy distributions built from pairs ***do not*** account for joint activity of clusters of 3-5 cells (macaque V1, Ohiorhenuan et al., CoSyNe 2006)

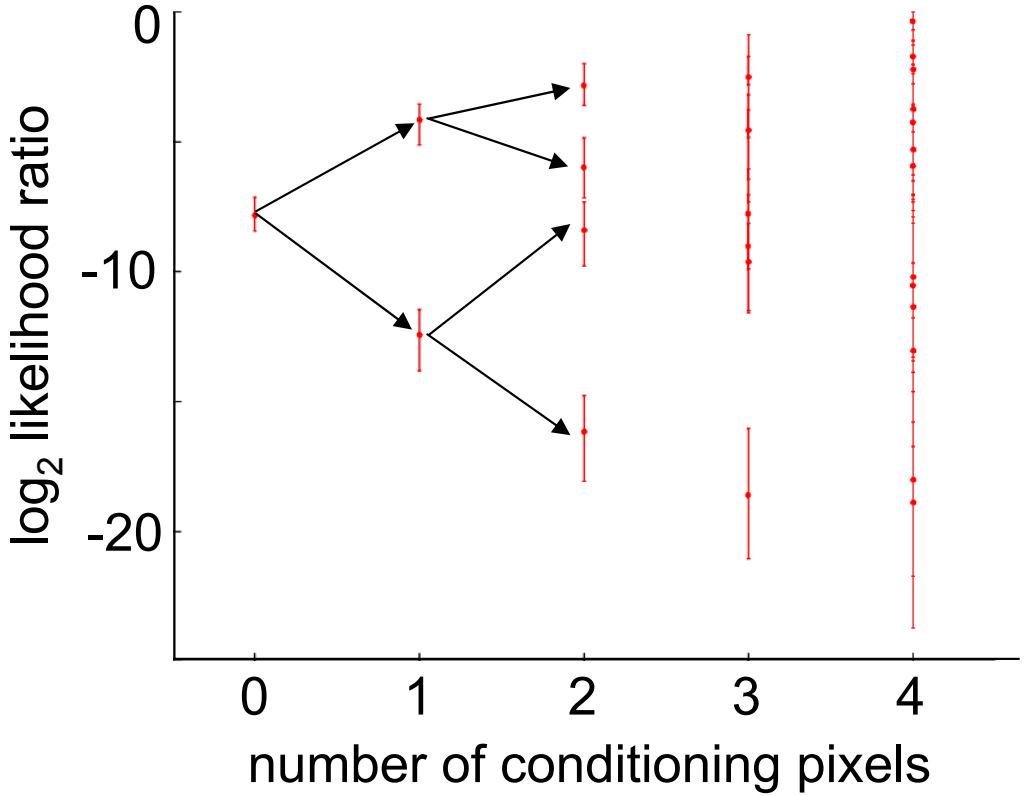
Multineuronal Firing Patterns: Visual Cortex

Analyzing stimulus-dependence

reverse correlation with pseudorandom checkerboard (m-sequence) stimuli

construct a pairwise maximum entropy model conditioned on one or more stimulus pixels that modulate the response probability distribution





Higher-order correlations are stimulus-dependent!
(Ohiorhenuan et al., SfN 2007)

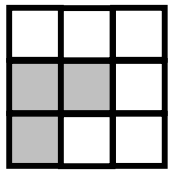
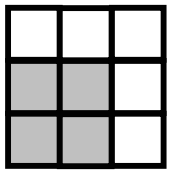
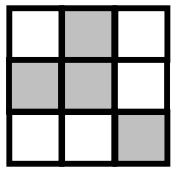
Maximum Entropy and Stimulus Design

- We'd like to understand how cortical neurons respond to natural scenes
- How can we determine what aspects of natural scenes make current models fail?
- Maximum-entropy approach: incrementally add constraints corresponding to natural-scene image statistics
  - Gaussian white noise or binary noise: resulting models are OK for retinal ganglion cells, poor in V1
  - Gaussian $1/f^2$ noise: in V1, >50% of variance not explained
  - How about adding higher-order statistical constraints?
    - Which ones to add?
    - How do they interact?

Dimensional Reduction of Image Statistics

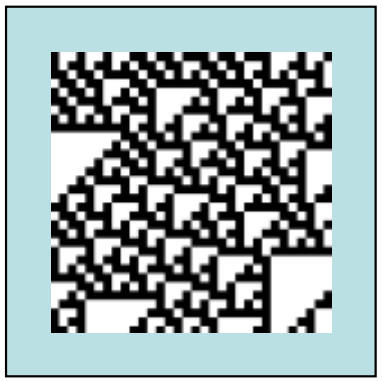
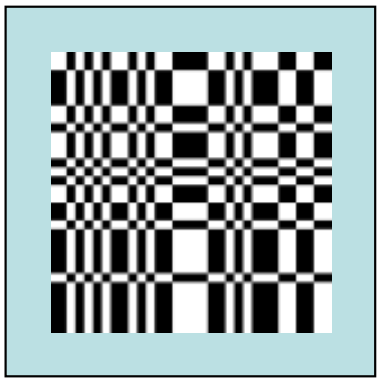
Image statistics can be studied in isolation by creating binary images with a single statistic constrained:

the mean value of the product of luminance values (+1 or -1) in a "glider" of several cells

some are visually salient

others are not


When two image statistics are constrained, simple combination rules account for the visual salience of the resulting texture.

Victor and Conte 1991, 2006

Homework

- *Debiasing entropy estimates:* Given an urn with an unknown but finite number of kinds of objects k, and an unknown number n_k of each kind of object, and N samples (with replacement) from the urn, estimate k.
  - Does it help to use a jackknife debiaser?
  - Does it help to postulate a distribution for n_k?

  (Treves and Panzeri, Neural Computation 1995)

- *The Data Processing Inequality:* Shannon's mutual information is just one example of a function defined on joint input-output distributions.
  - Are there any others for which Shannon information holds?
  - If so, are they useful?

  (Victor and Nirenberg, submitted 2007)

- *Maximum-entropy distributions:* Consider a Poisson neuron driven by a Gaussian noise. Build the maximum-entropy distribution constrained by the overall mean firing rate, input power spectrum, and spike-triggered average. Fit this with a linear-nonlinear (LN) model.
  - What is the nonlinearity's input-output function?
  - Is the spike-triggered covariance zero?

  (?? and Victor)