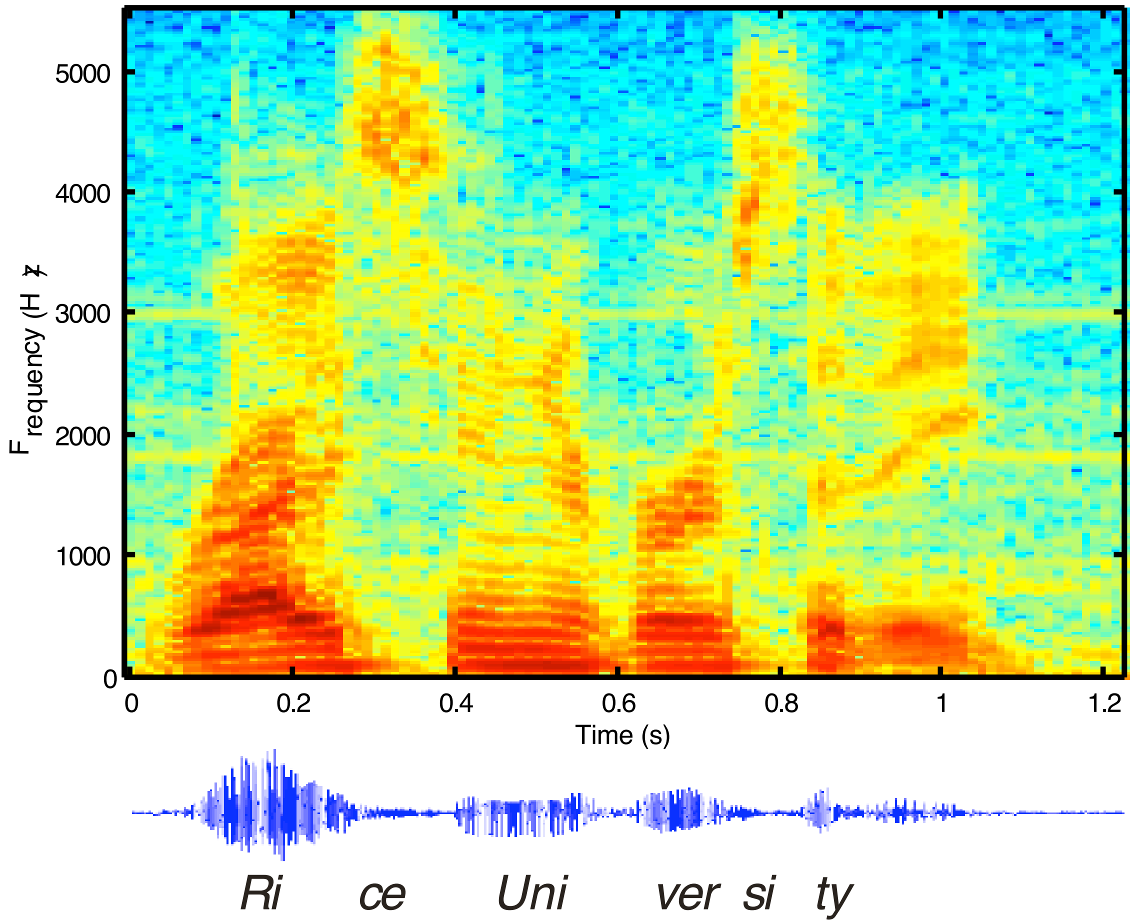
From Signal to Information Processing

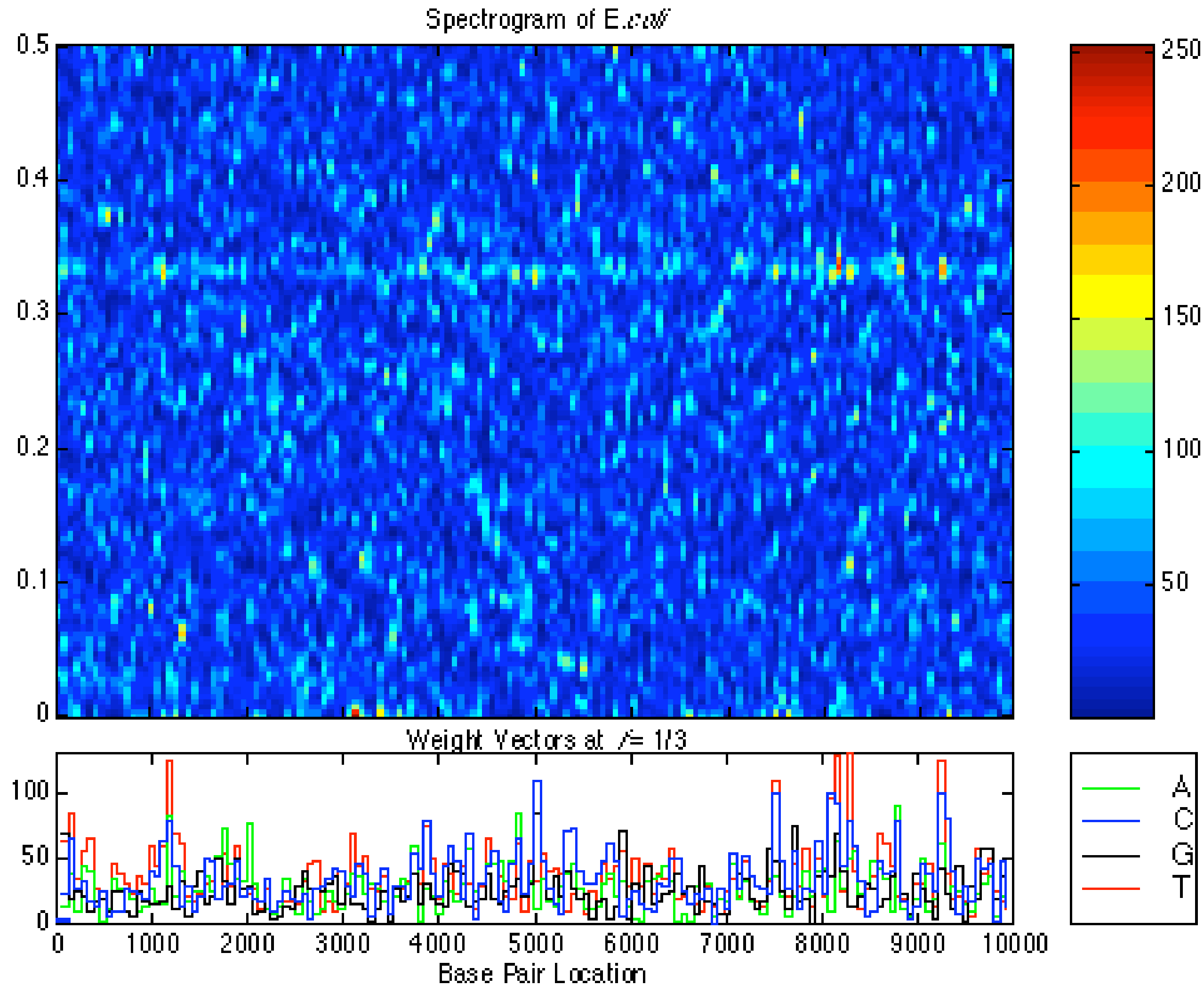
Don H. Johnson
Computer and Information Technology Institute
Department of Electrical & Computer Engineering
Rice University
Houston, Texas

What's the problem?

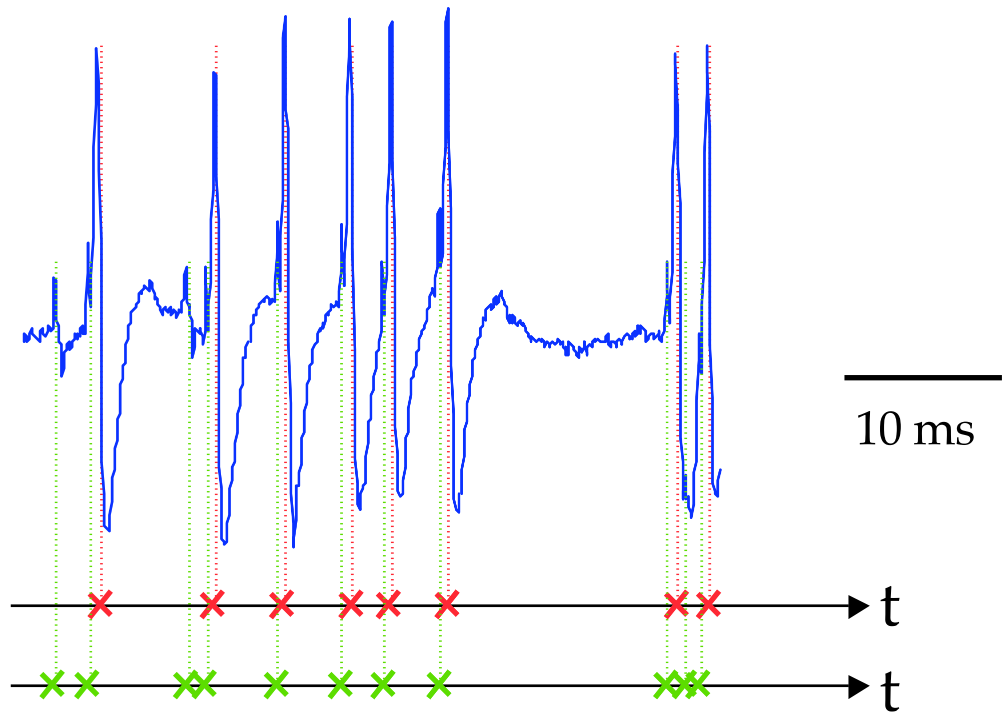
Signal processing has been concerned with form, not
what the signal represents



Not all signals are so easy to analyze

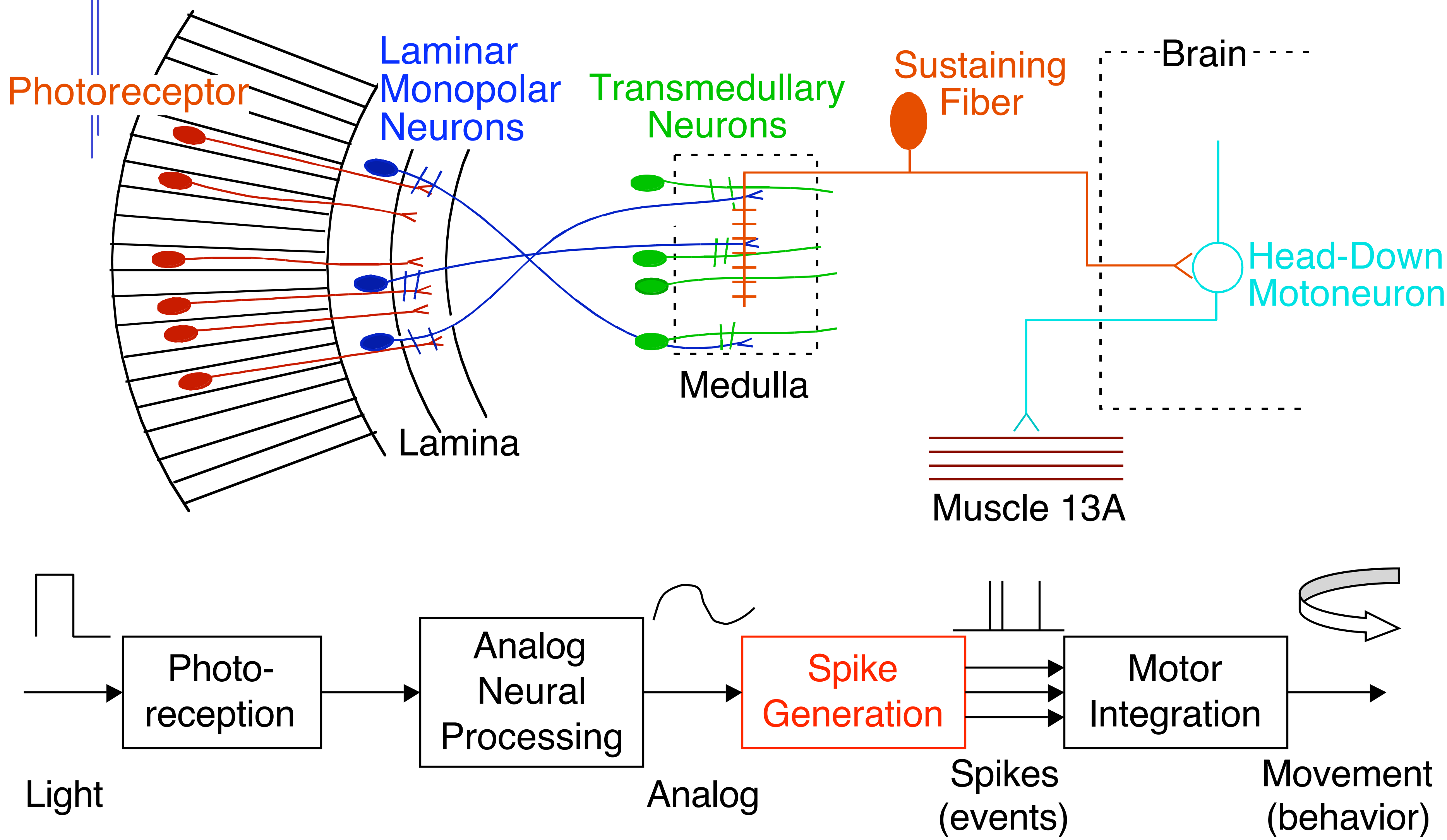


Neural representation of information



Information represented by *when* spikes occur either in single neuron responses

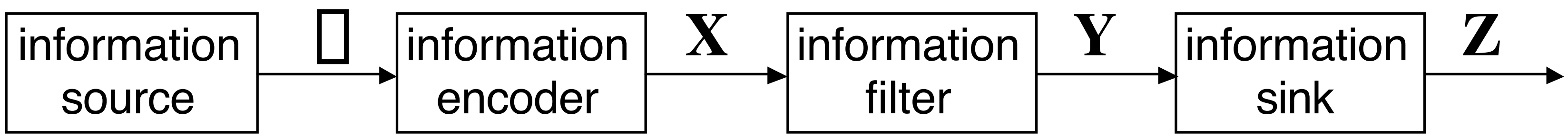
Crayfish dorsal light reflex pathway



Beginnings of information processing

- Information is "in the eye of the beholder"
  - Cellular telephony example (interference to one is information to another)
  - Without interacting with information encoded by a signal, examining signals won't reveal how well (or if) information is represented
- Signals convey information, but how *effectively* to they do so?
- Systems process information, selectively suppressing irrelevant information and accentuating important information by acting on signals (*information filters*)
- System design is usually signal-based, not information based. *What effect does system design have on information processing?*

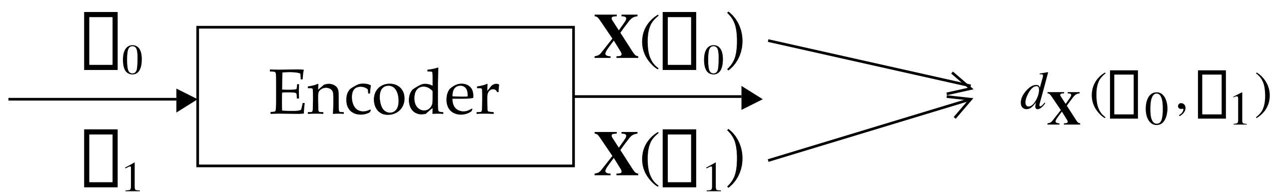
Canonic information processing structure



- Information □ is *always* represented—encoded—by signals
- Systems “process information” indirectly by acting on signals
- Result **Z** is an action or a behavior (i.e., a measurable quantity)
- Any viable information processing theory *must* encompass a variety of signals
- Here, all signals are assumed to be stochastic

Signals represent information

- Let $\square$ represent the information encoded in a signal $\mathbf{X}(\square)$
- Quantify how accurately information changes $\square_0$ $\square$ $\square_1$ are represented by signals with a *distance measure* $d_{\mathbf{X}}(\square_0, \square_1)$



How to choose a distance?

- Calculate distance between the probability distributions $p_{\mathbf{X}}(\mathbf{x}; \square_0)$, $p_{\mathbf{X}}(\mathbf{x}; \square_1)$ characterizing the signal
- Because $p_{\mathbf{X}}(\mathbf{x}; \bullet)$ maps the signal domain to the real-line, we can calculate distances *regardless* of the kind of signal
- Information extraction systems—determining $\square$ from $\mathbf{X}(\square)$—fall into two categories
  - Classification: Which of several values of $\square$ occurred
    Optimal classifier is the likelihood ratio test
    No general formula for performance is known
  - Estimation: Determine $\square$ from a continuum of values
    Mean-squared error a frequently used performance measure

Distances and optimal processing

- The optimal classifier that tries to determine whether θ_0 or θ_1 was encoded will have an error probability of the form $$P_e \sim 2^{-d_{\mathbf{X}}(\theta_0,\theta_1)}$$
- Cramér-Rao lower bound on the mean-square error incurred by *any* (unbiased) estimator

$$E[\varepsilon^2] \geq \frac{1}{F(\theta)} \quad \text{(scalar } \theta\text{)} \qquad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \geq [\mathbf{F}(\boldsymbol{\theta})]^{-1} \quad \text{(vector } \boldsymbol{\theta}\text{)}$$

$$[\mathbf{F}(\boldsymbol{\theta})]_{ij} = E\left[\frac{\partial \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \ln p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_j}\right]$$

Fisher information matrix

- Fisher information matrix related to distance induced by small information changes (locally Gaussian property)

$$d_{\mathbf{X}}(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0+\delta\boldsymbol{\theta}) \propto K \cdot (\delta\boldsymbol{\theta})^T \mathbf{F}(\boldsymbol{\theta}_0)\,\delta\boldsymbol{\theta}$$

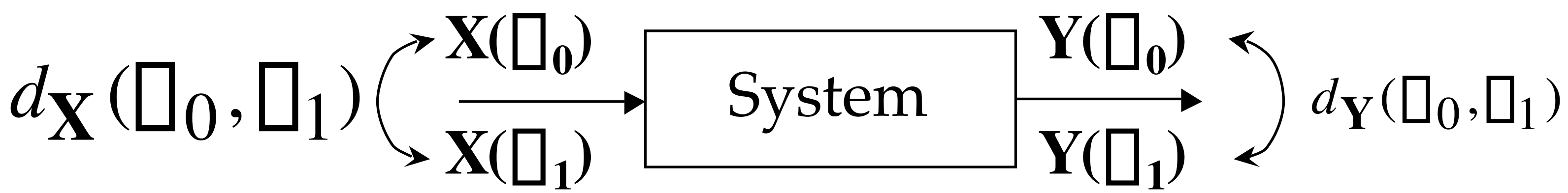
- *With one distance, we can quantify how well information is represented from both classification and estimation viewpoints*

Information processing fundamental

- Information-theoretic distance measures obey the Data Processing Theorem:

$$d_{\mathbf{X}}(\square_0, \square_1) \geq d_{\mathbf{Y}}(\square_0, \square_1)$$

Systems cannot increase how well information is represented by their inputs



Choosing a distance measure

- Many information theoretic distances have the locally Gaussian property
- Only two are known to be related to optimal classifier performance
- We choose distance measures related to the Kullback-Leibler distance

$$D_{\mathbf{X}}(\theta_1 \| \theta_0) = \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x};\theta_1) \log \frac{p_{\mathbf{X}}(\mathbf{x};\theta_1)}{p_{\mathbf{X}}(\mathbf{x};\theta_0)}$$

- Choose base-2 logarithms, which gives distance “units” of bits.

Properties of K-L distance

- $D_{\mathbf{X}}(\theta_1 \| \theta_0) \geq 0$ Equality only when $p_{\mathbf{X}}(\mathbf{x};\theta_1) = p_{\mathbf{X}}(\mathbf{x};\theta_0)$
- $D_{\mathbf{X}}(\theta_1 \| \theta_0) \neq D_{\mathbf{X}}(\theta_0 \| \theta_1)$ (K - L "distance" is not necessarily symmetric)
- If $\mathbf{X}(\theta)$ has statistically independent components,

$$D_{\mathbf{X}}(\theta_1 \| \theta_0) = \sum_n D_{X_n}(\theta_1 \| \theta_0)$$

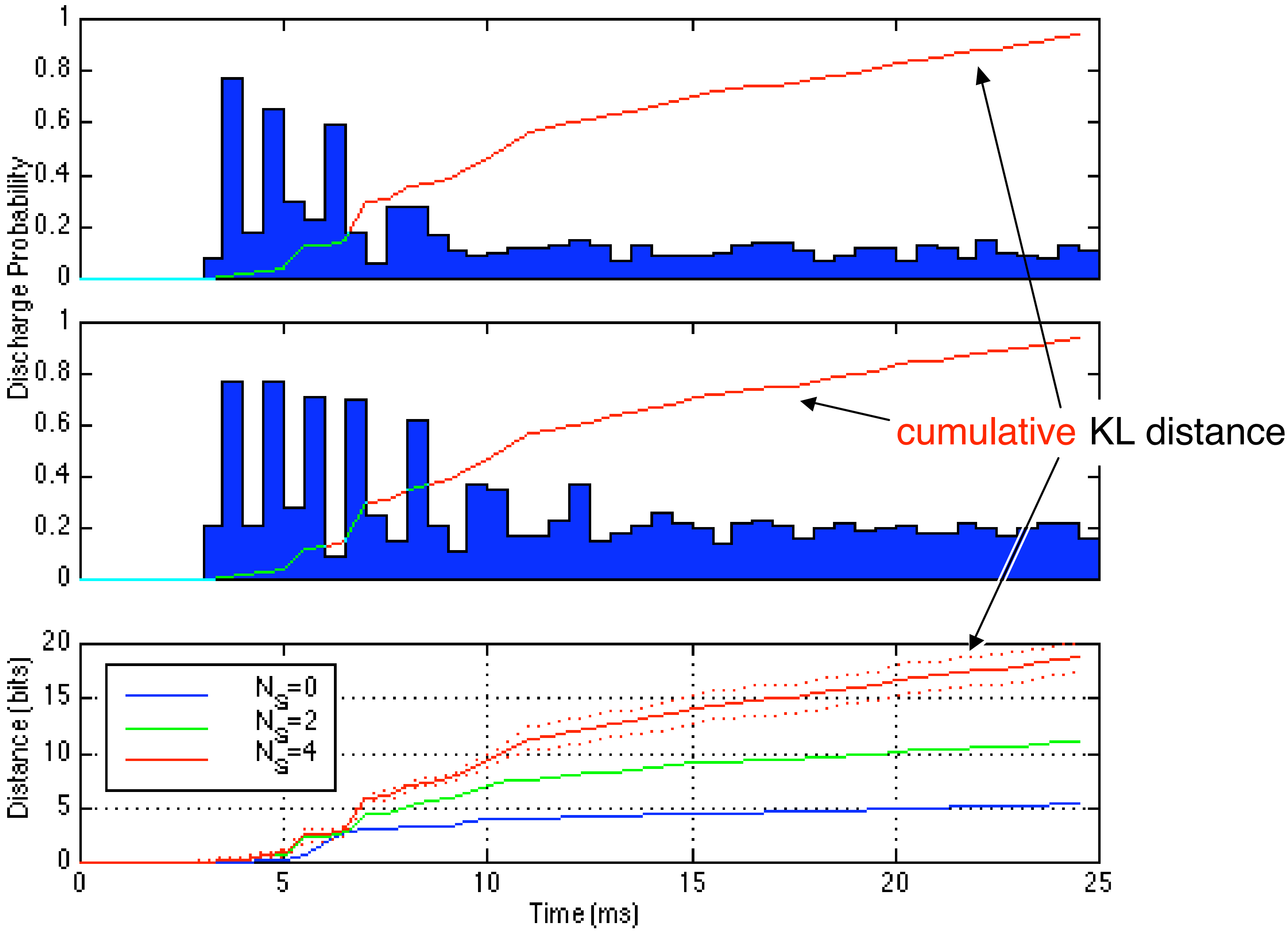
- K-L distance is the "exponential rate" of Neyman-Pearson detector's false-alarm probability

$$P_F \sim 2^{-N D_X(\theta_1 \| \theta_0)} \text{ for fixed } P_M$$

- Distance resulting from information perturbations is "proportional" to Fisher information

$$D_{\mathbf{X}}(\theta_0 + \delta\theta \,\|\, \theta_0) \approx \frac{\delta\theta^{T}\,\mathbf{F}(\theta_0)\,\delta\theta}{2\ln 2}$$

Distance between LSO response patterns



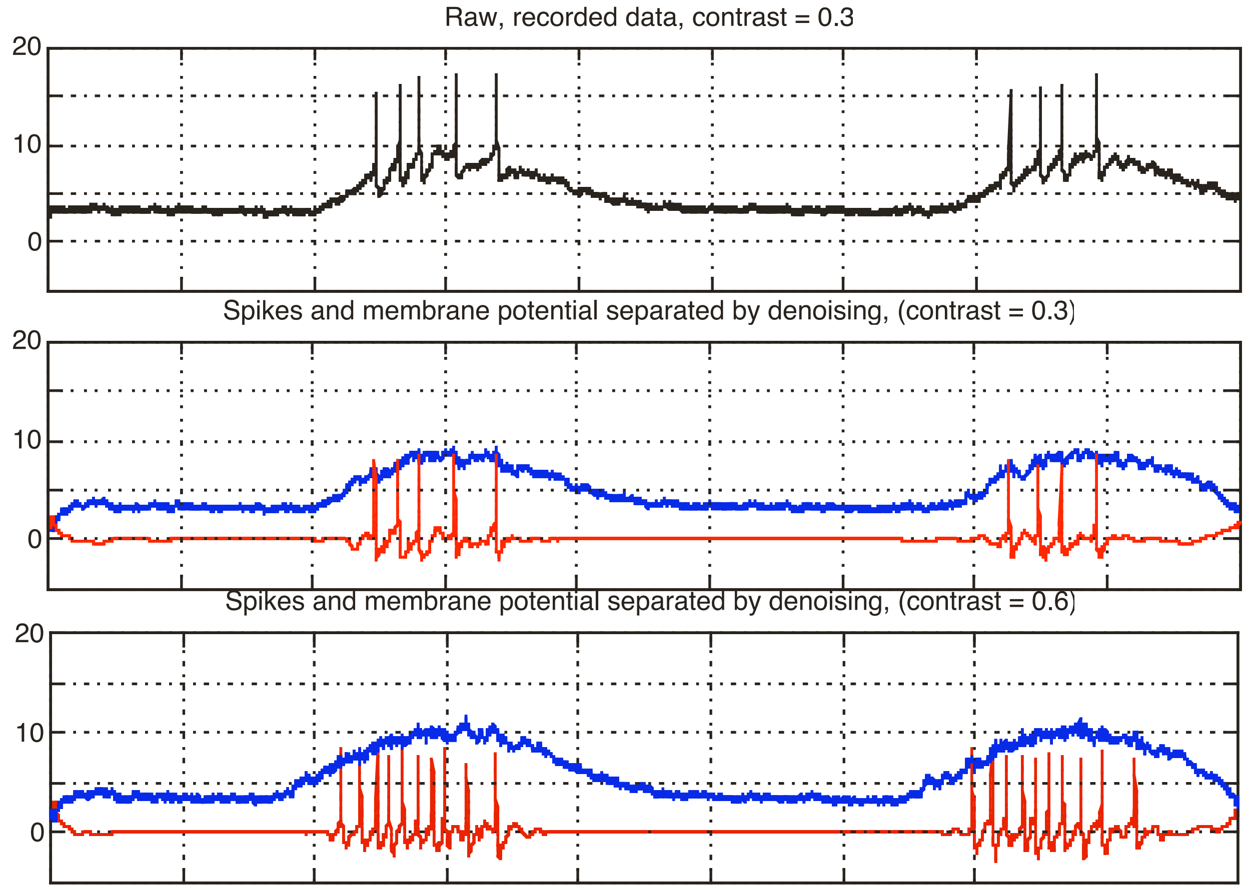
Analyzing system performance

- Quantify a system's information processing performance with the information transfer ratio

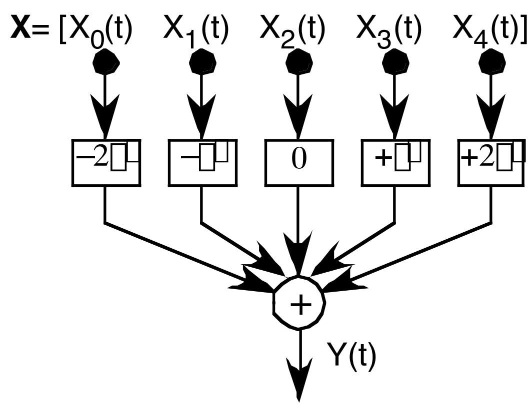
$$\square_{\mathbf{X},\mathbf{Y}}(\square_0,\square_1) = \frac{d_{\mathbf{Y}}(\square_0,\square_1)}{d_{\mathbf{X}}(\square_0,\square_1)}$$

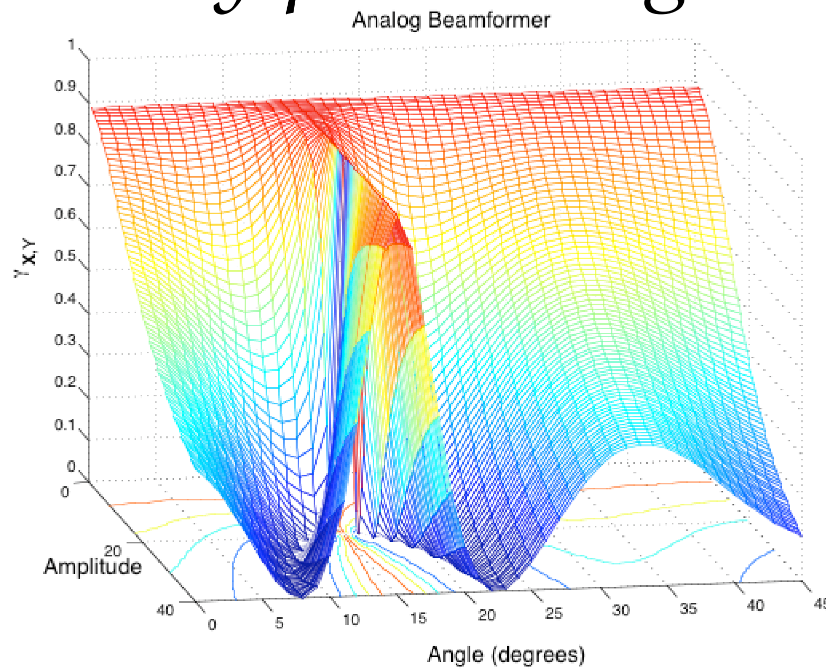
  - $0 \leq \square_{\mathbf{X},\mathbf{Y}}(\square_0 , \square_1) \leq 1$
  - If $\square_{\mathbf{X},\mathbf{Y}}(\square_0 , \square_1) = 1$, the information change is well encoded in the output signal.
  - If $\square_{\mathbf{X},\mathbf{Y}}(\square_0 , \square_1) << 1$, the information change is poorly encoded in the output signal
- Choose a reference $\square_0$; explore how $\square$ varies about this point
- *Information filtering*

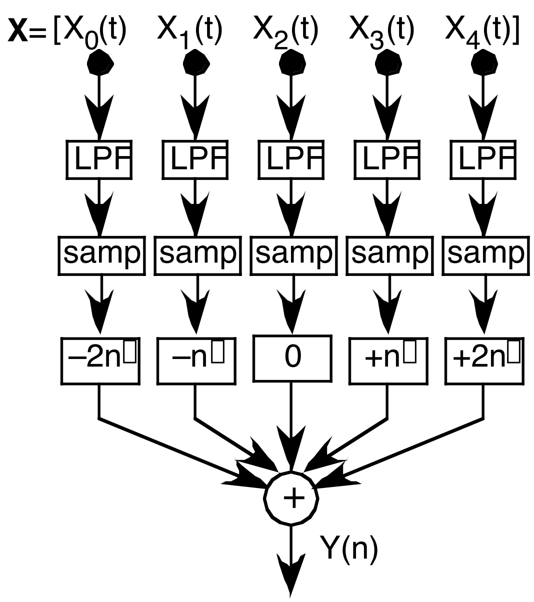
Information transfer across a synapse

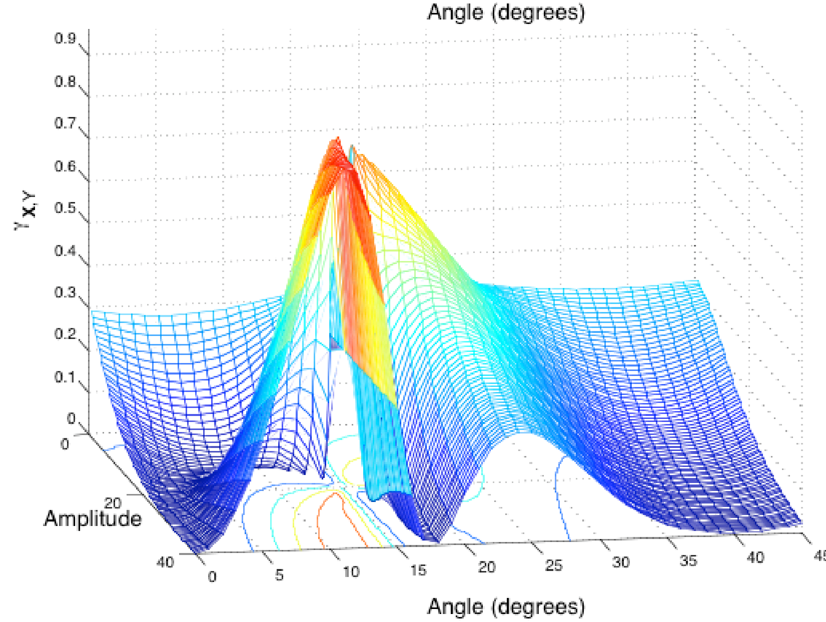


Information filtering: Array processing



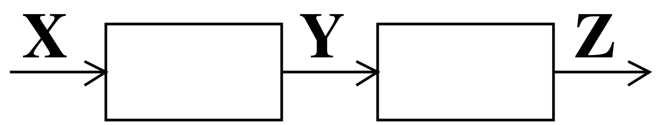






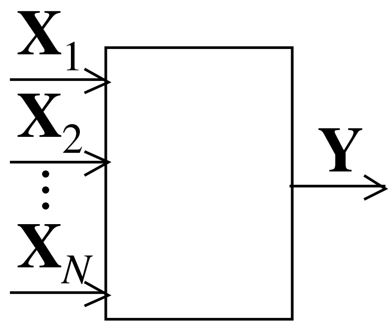
System theory of information processing

Cascade of systems



$$\gamma_{\mathbf{X},\mathbf{Z}} = \gamma_{\mathbf{X},\mathbf{Y}} \cdot \gamma_{\mathbf{Y},\mathbf{Z}}$$

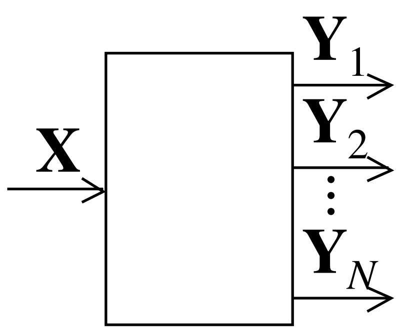
Multiple input systems



If inputs are independent,

$$\frac{1}{\gamma_{\mathbf{X},\mathbf{Y}}} = \sum_n \frac{1}{\gamma_{X_n,\mathbf{Y}}} \Rightarrow \gamma_{\mathbf{X},\mathbf{Y}} \le \min_n \{\gamma_{X_n,\mathbf{Y}}\}$$

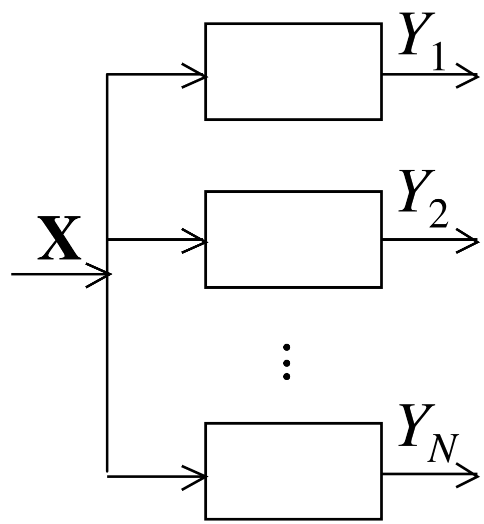
Multiple output systems (*e.g.*, neural populations)



$$\gamma_{\mathbf{X},\{Y_1,\ldots,Y_N\}} = \gamma_{\mathbf{X},Y_1} + \sum_{n=2}^{N} \gamma_{\mathbf{X},\{Y_n | Y_1,\ldots,Y_{n-1}\}}$$

Non-cooperative populations

- The non-cooperative structure defines a baseline for multi-output systems



- The outputs are *conditionally* independent, *not* statistically independent

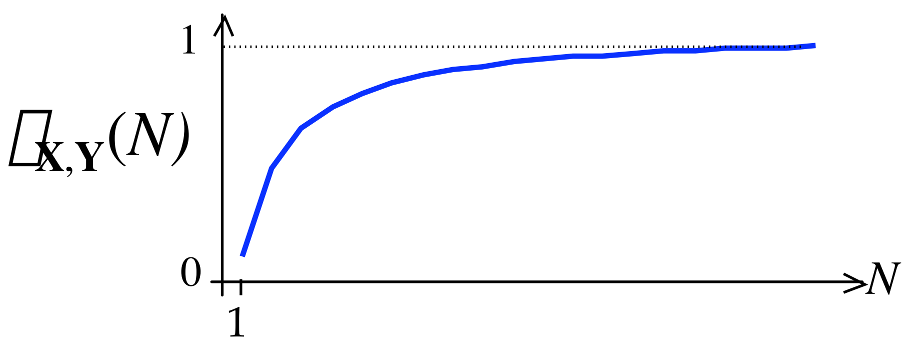
$$p(Y_1, Y_2, \ldots, Y_N; \theta) = \int p(Y_1|\mathbf{x})\, p(Y_2|\mathbf{x}) \cdots p(Y_N|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x}; \theta)\, d\mathbf{x}$$

- The outputs contain only input-induced dependence

Non-cooperative population theory

- Assume each system is not too noisy ($\square_n \geq \square_{\min} > 0$)
- As the population size N increases, the population can represent the information expressed by its input without loss, regardless of the information representation

$$\lim_{N \square} \square_{\mathbf{X},\mathbf{Y}}(N) = 1$$



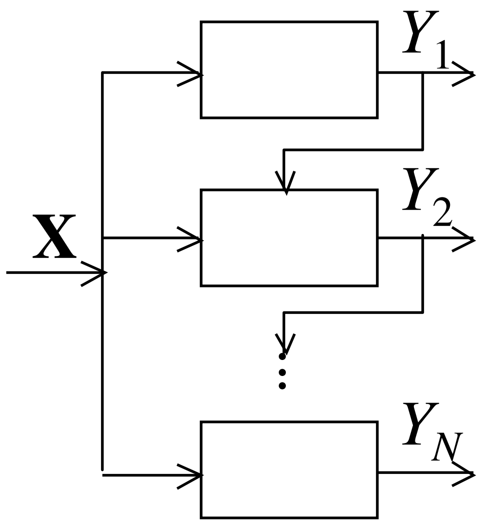
$$\square_{\mathbf{X},\mathbf{Y}}(N) \square 1 \square \frac{k}{N}$$ Continuous code

or

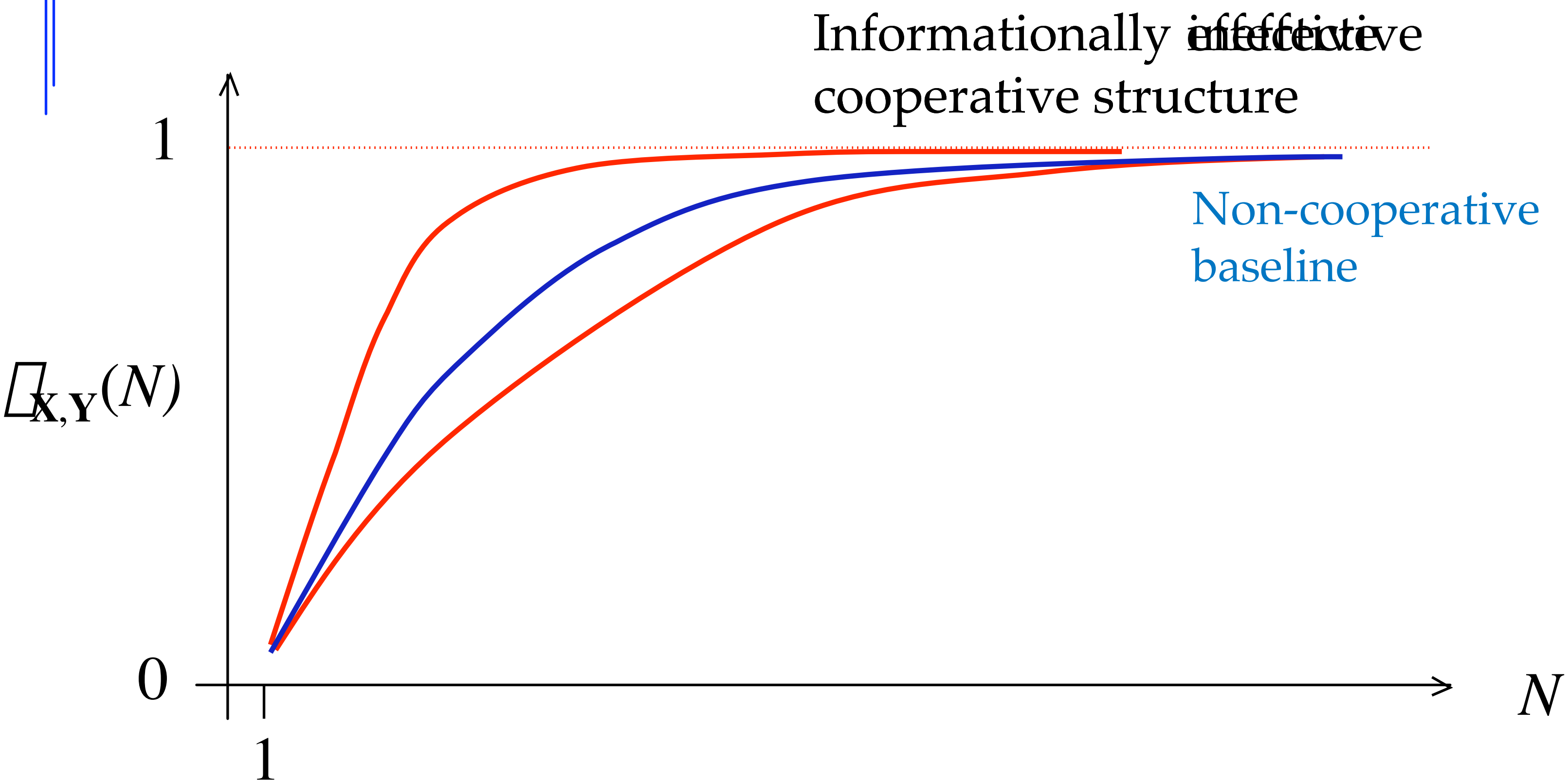
$$\square_{\mathbf{X},\mathbf{Y}}(N) \square 1 \square k_1 e^{\square k_2 N}$$ Discrete code

Cooperative populations

If the cooperation among systems involves output feedback to a limited number of other systems, *the asymptotics of noncooperative systems apply as well*.

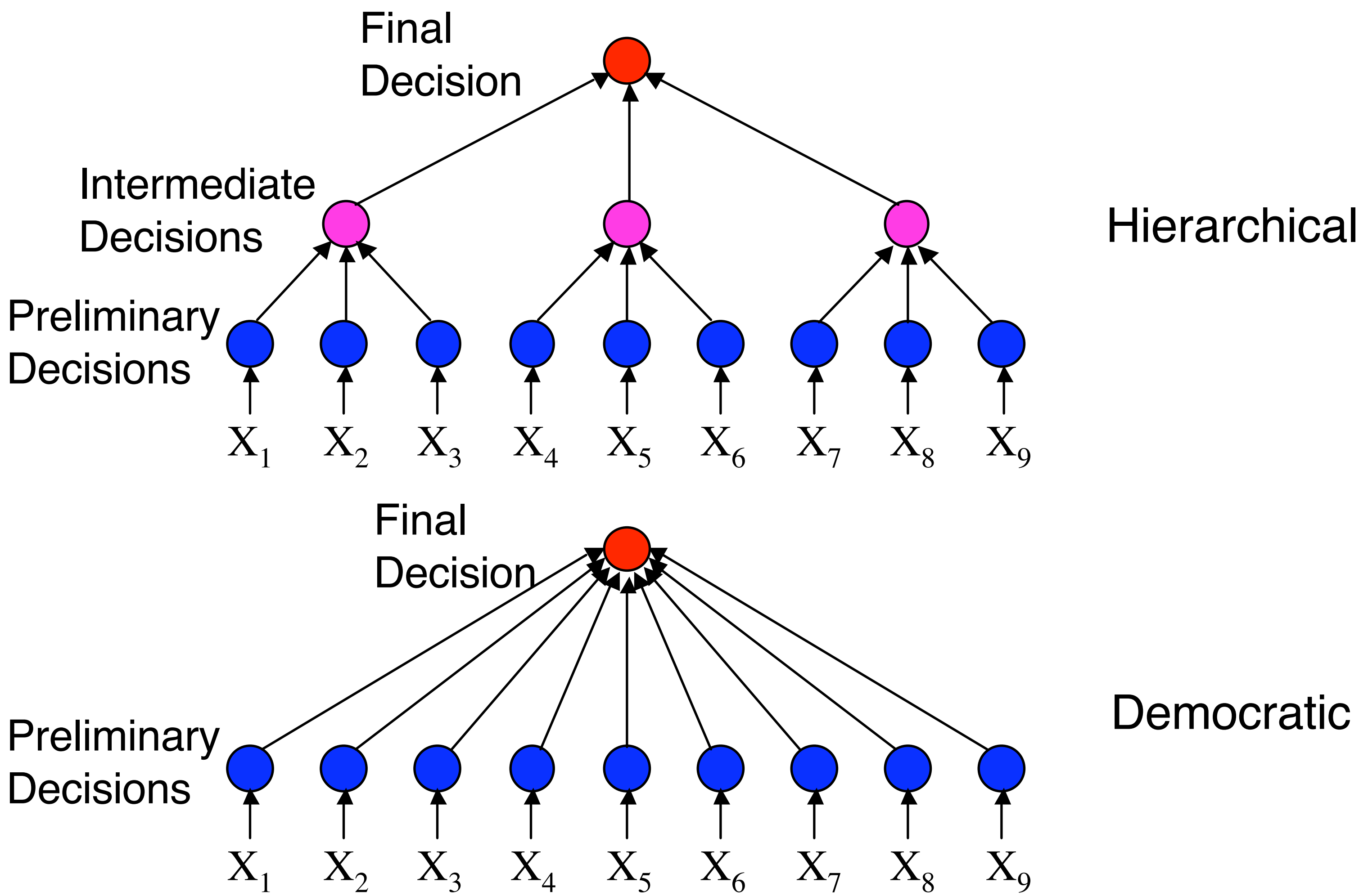


Population coding performance limits

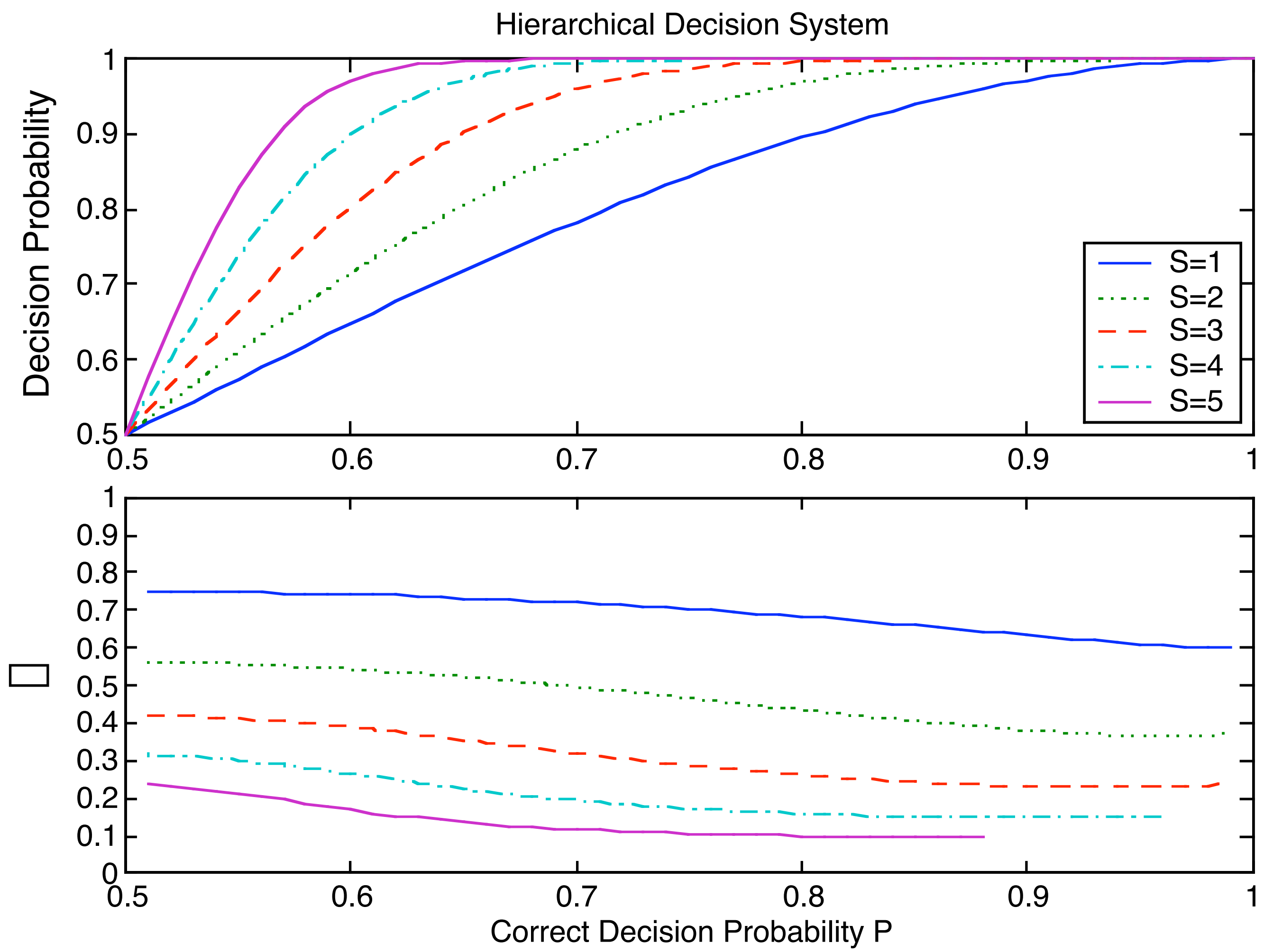


Distributed decision systems

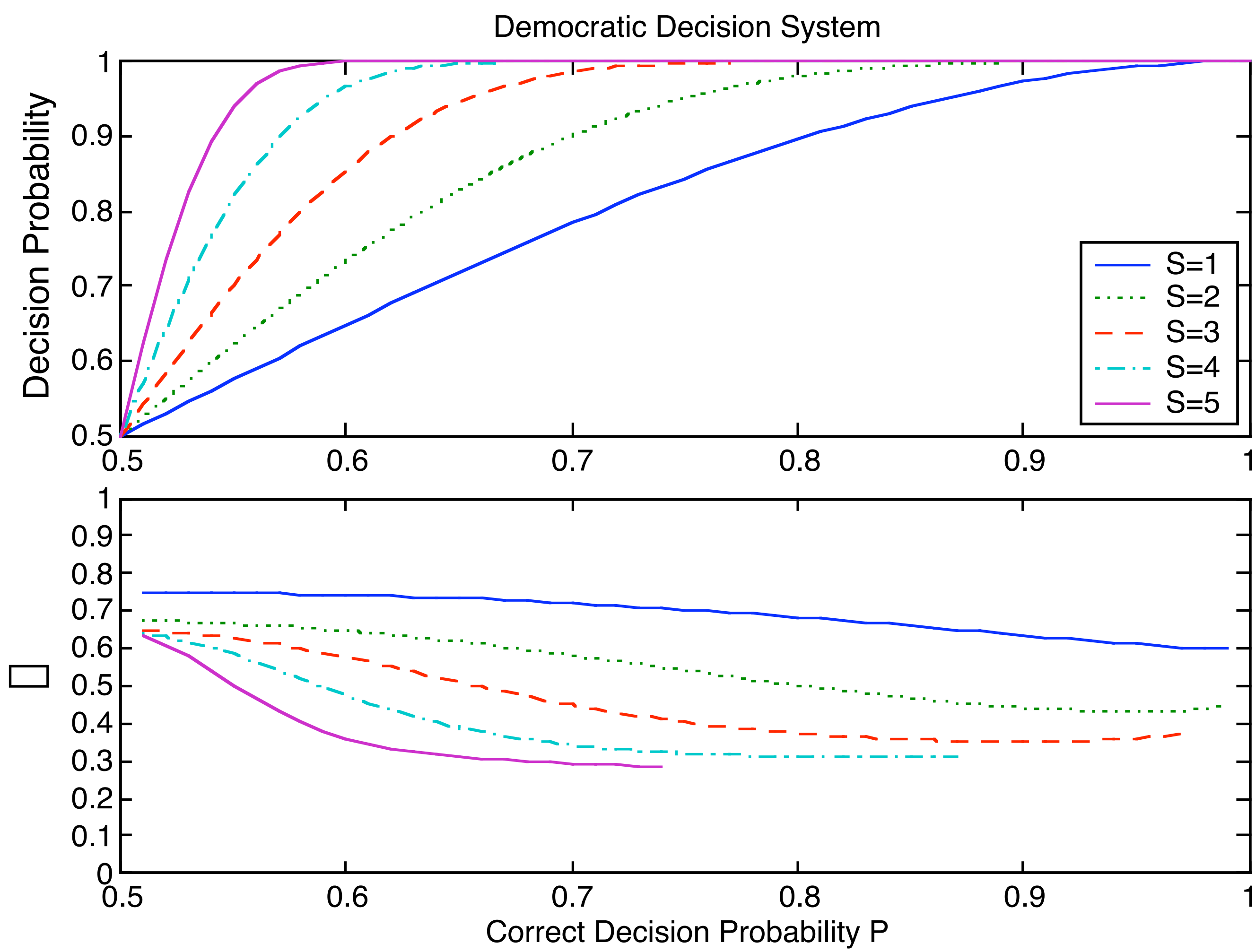
What is the most effective way to integrate individual decisions into a global decision?



Results: Hierarchical structure



Results: Democratic structure



Summary

- A *theory of information processing* must not depend on the nature of the signals representing information
- The theory presented here uses information theoretic distances, particularly the Kullback-Leibler distance, as the primary tool
- Data Processing Theorem is a *fundamental* result that can be widely applied
- Information processing *structures* have fundamental properties regardless of…
  - the information being processed
  - the signals representing the information
- We can assess signal encoding and system processing, hopefully leading to better designs that focus on the *information*, not the signal

Collaborators

Co-Investigators

Keith Baggerly
Raymon Glantz

Graduate students

Michael Lexa
Chris Rozell
Sinan Sinanovic

Undergraduates

Michelle Lloyd

Post-Docs

Charlotte Gruner