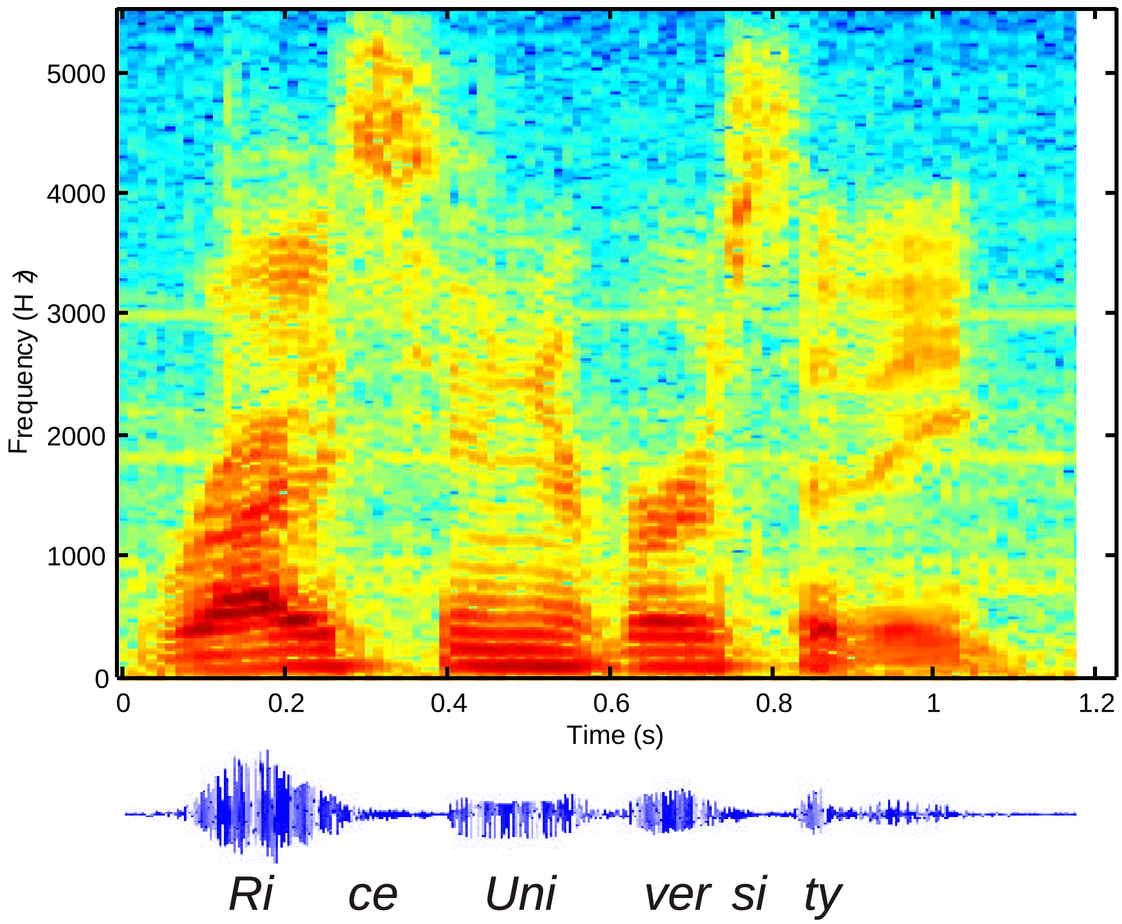
A Theory of Information Processing

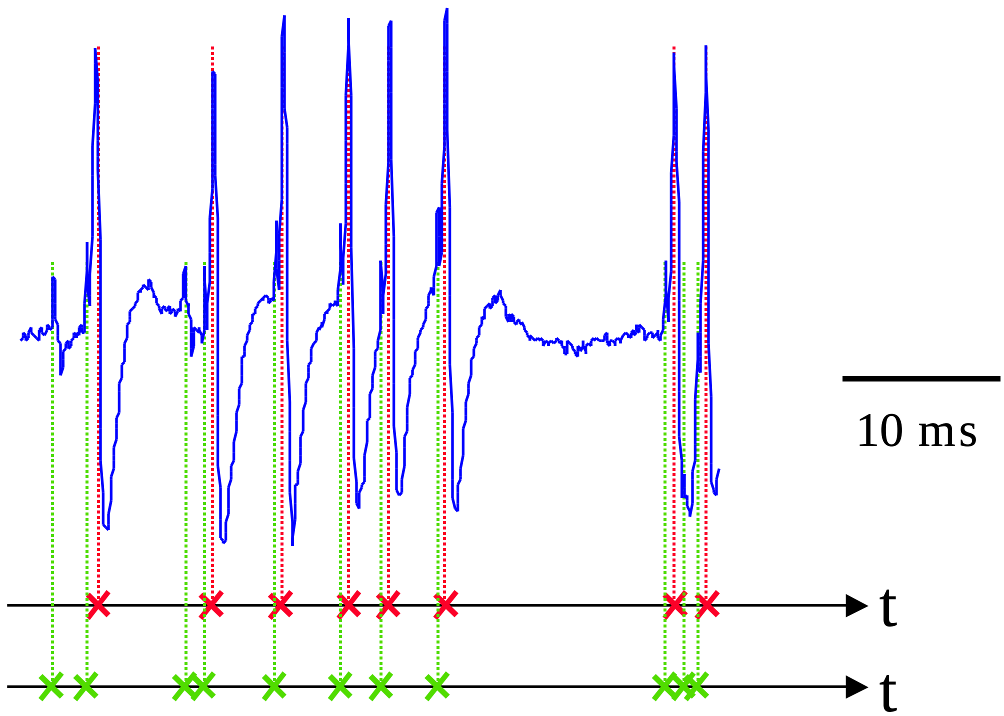
Don H. Johnson
Center for Multimedia Communication
Department of Electrical & Computer Engineering
Rice University
Houston, Texas

What's the problem?

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

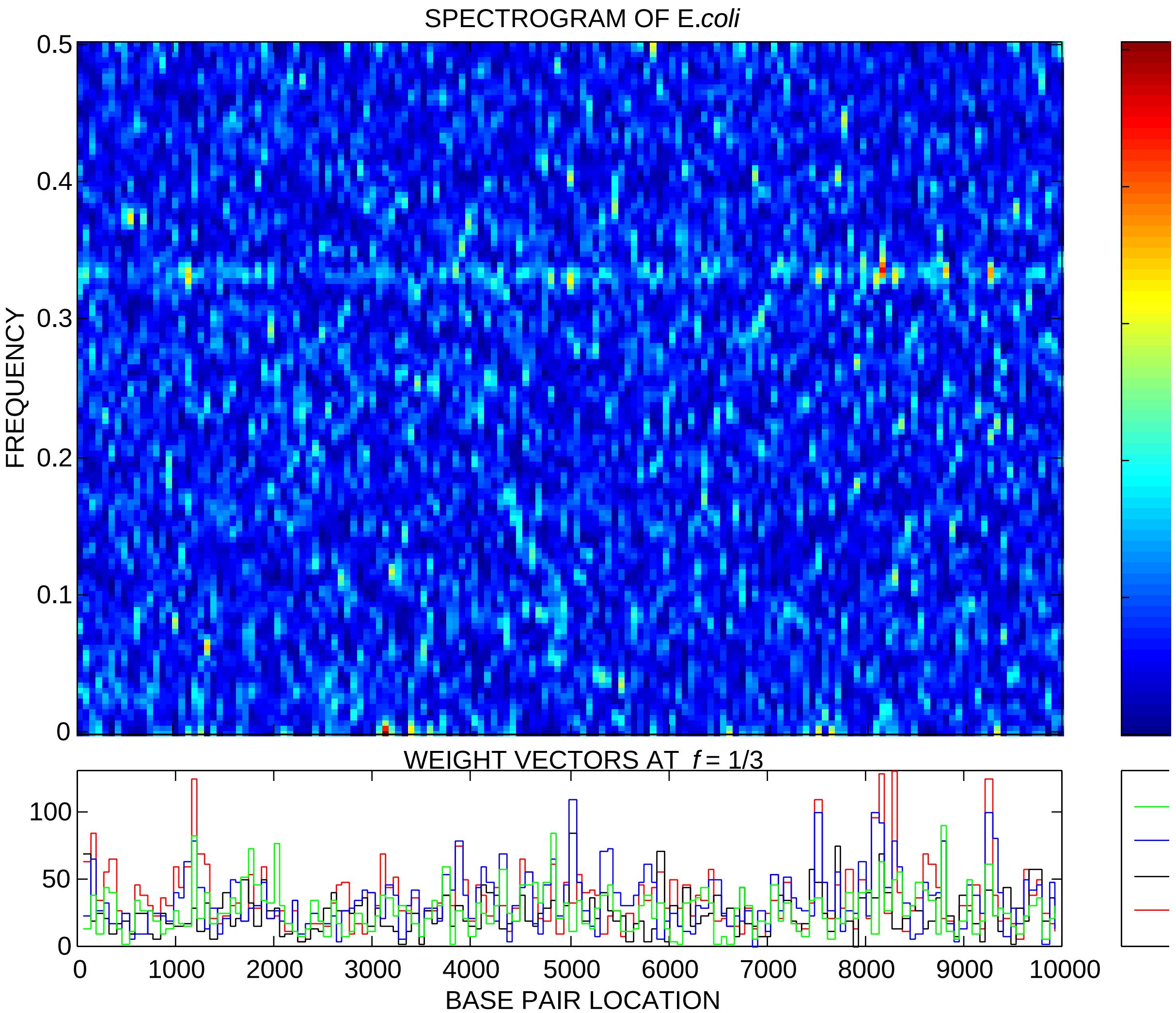


Neural representation of information



Information represented by *when* spikes occur either in single neuron responses or, more importantly, jointly in population (ensemble) neural responses

Some signals aren't numeric



A quote from Warren Weaver

"In communication there seems to be problems at three levels: 1) technical, 2) semantic, and 3) influential.

"The technical problems are concerned with the accuracy of transference of information from sender to receiver.…The semantic problems are concerned with the interpretation of meaning by the receiver, as compared with the intended meaning of the sender.…The problems of influence or effectiveness are concerned with the succe with which the meaning conveyed to the receiver leads to the desired conduct on his part.

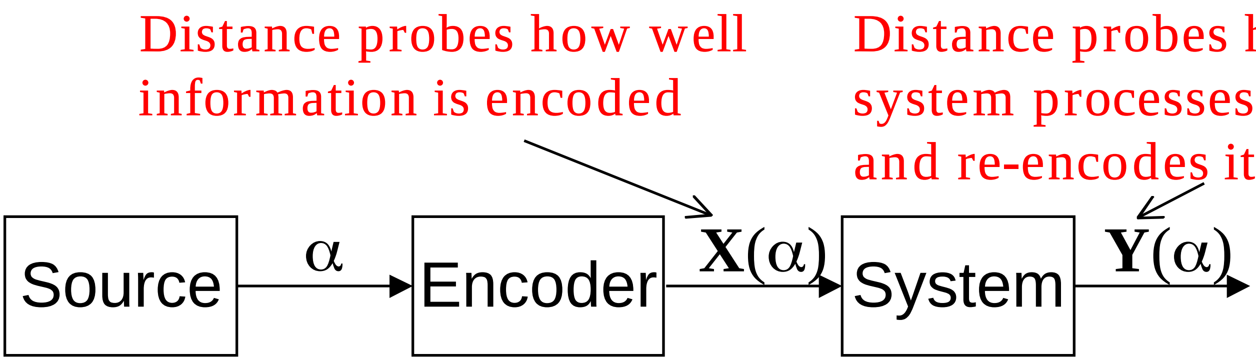
"…The concept of information developed in this theory a first seems disappointing and bizarre—disappointing because it has nothing to do with meaning, and bizarre because it deals not with a single message but rather with the statistical character of a whole ensemble of messages, bizarre also because in these statistical terms the words information and uncertainty find themselves partners."

Toward a theory of information processing

- Tenets
  - What is information is "in the eye of the beholder"
  - Information does not exist physically
  - Information is conveyed by signals
  - Systems process information *indirectly* by acting on signals
- Analysis techniques must not depend on the kind of signal being analyzed (numeric, symbolic, mixed)
- Issues a viable theory must address
  - *Effectiveness of representation*
    How effectively is information represented by signals?
  - *Information processing*
    How well do systems extract/suppress information

Signals represent information

- Let α represent the information encoded in a signal $\mathbf{X}($
- Investigate how accurately information changes $\alpha_0 \rightarrow$ are represented by signals with a *distance measure* $d_{\mathbf{X}}(\alpha_0,\alpha_1)$
- Calculate distance between the probability distributio $p_{\mathbf{X}}(\mathbf{x},\alpha_0)$, $p_{\mathbf{X}}(\mathbf{x},\alpha_1)$ 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 sign
- We choose a distance measure related the performance of *optimal* information processing systems



Optimal information processing

- Information extraction systems—determining α from $\mathbf{X}(\alpha)$—fall into two categories
  - Classification: Which of several values of α occurred
    Optimal classifier is the likelihood ratio test
    No general formula for performance is known
  - Estimation: Estimate α from a continuum of values
    Optimal estimator depends on error criterion
    Cramér-Rao lower bound on the mean-squared error incurred by *any* (unbiased) estimator

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

$$[\mathbf{F}(\alpha)]_{ij} = E\left[\frac{\partial \ln p_{\mathbf{X}}(\mathbf{x},\alpha)}{\partial \alpha_i} \frac{\partial \ln p_{\mathbf{X}}(\mathbf{x},\alpha)}{\partial \alpha_j}\right]$$

Fisher information matrix

Distances and optimal processing

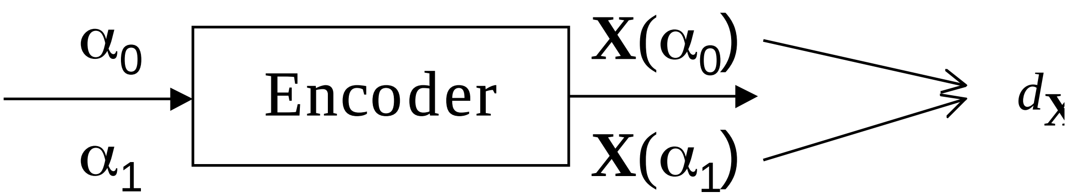
- The optimal classifier that, based on the observation of (having independent, identically distributed components), tries to determine whether α_0 or α_1 was encoded will have an error probability of the form

$$P_e \sim 2^{-N d_X(\alpha_0, \alpha_1)}$$

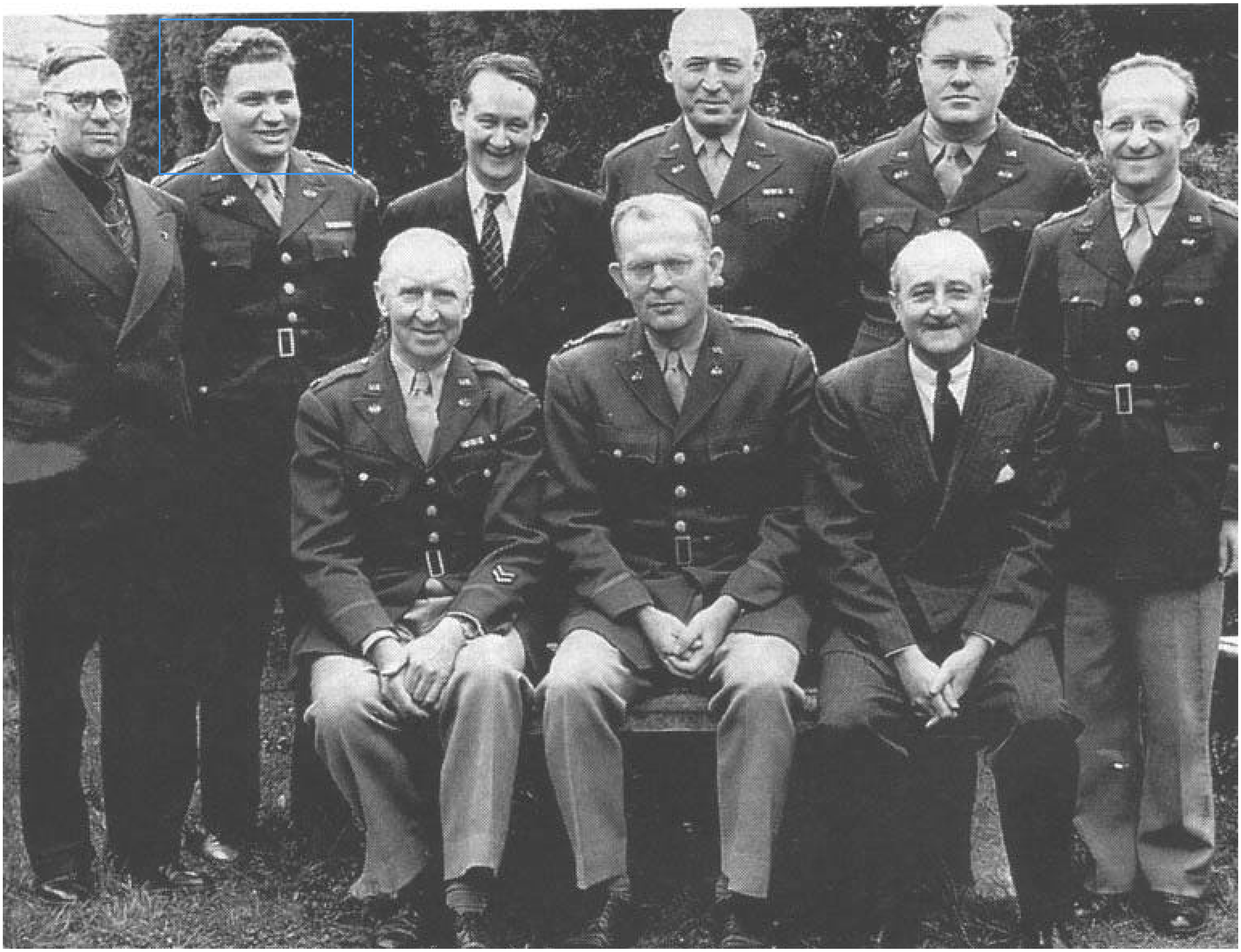
- Fisher information matrix related to distance induced by small information changes

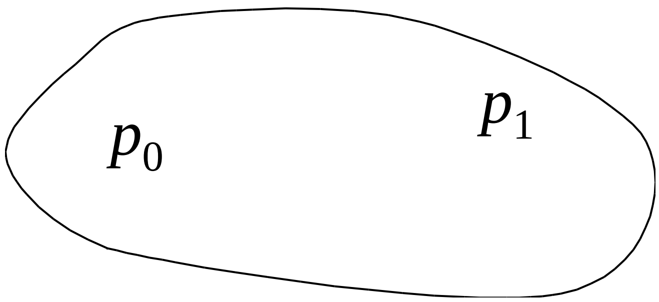
$$d_{\mathbf{X}}(\alpha_0, \alpha_0 + \delta\alpha) \approx K \cdot \delta\alpha' \mathbf{F}(\alpha_0) \delta\alpha$$

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



Solomon Kullback





p_0
p_1

Symmetric distances

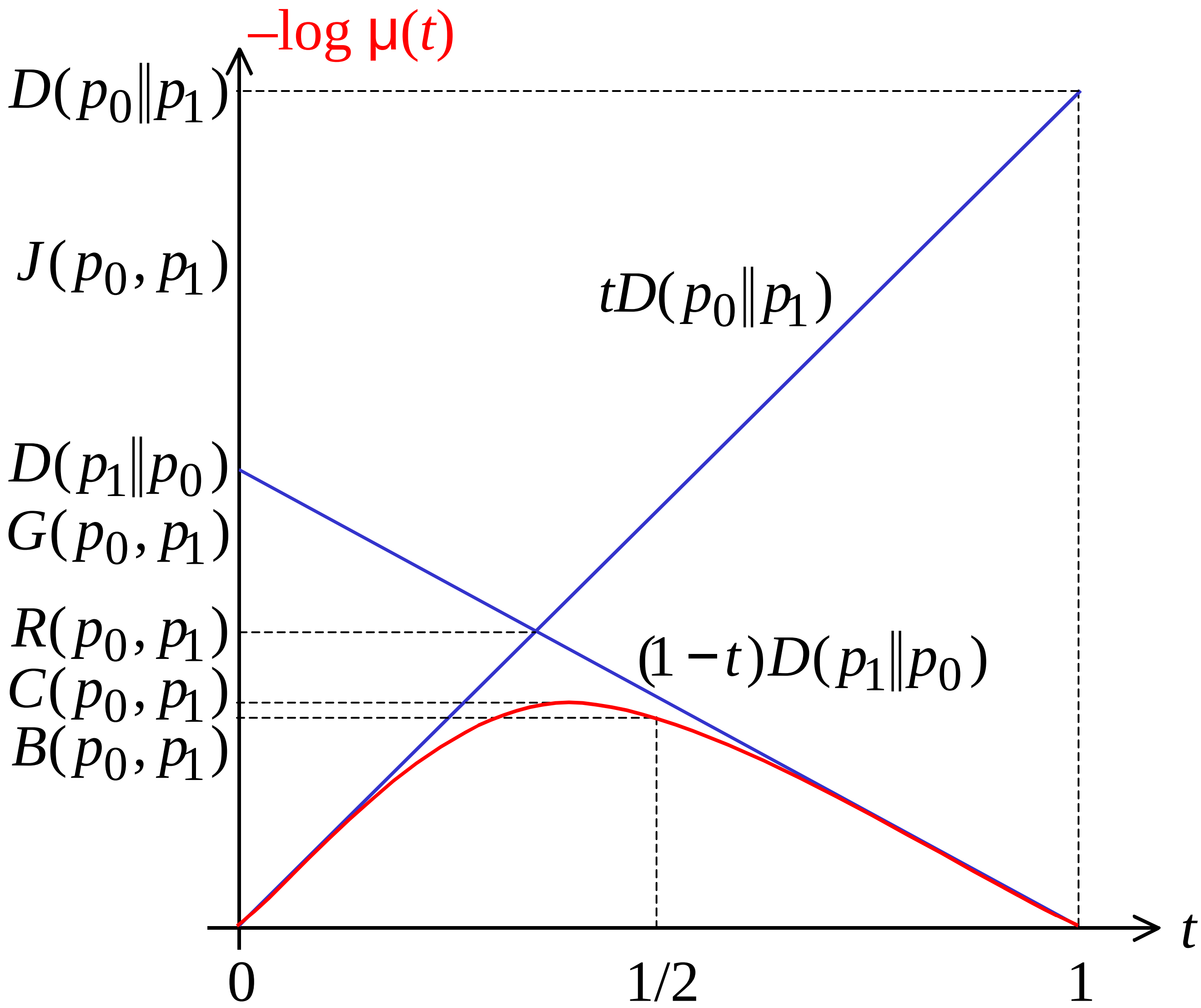
- Only other known distance measure related to classification is the Chernoff distance

$$C(\alpha_0,\alpha_1) = \max_{0 \le t \le 1} - \log \mu(t), \quad \mu(t) = \int [p(\mathbf{x},\alpha_0)]^{1-}$$

- A symmetric, easily computed alternative is the so-called resistor-average distance

$$R_{\mathbf{X}}(\alpha_0,\alpha_1) = \frac{D_{\mathbf{X}}(\alpha_1 \| \alpha_0) D_{\mathbf{X}}(\alpha_0 \| \alpha_1)}{D_{\mathbf{X}}(\alpha_1 \| \alpha_0) + D_{\mathbf{X}}(\alpha_0 \| \alpha_1)}$$

Symmetrizing the Kullback-Leibler distan



Nuance #1 in estimating K-L distance

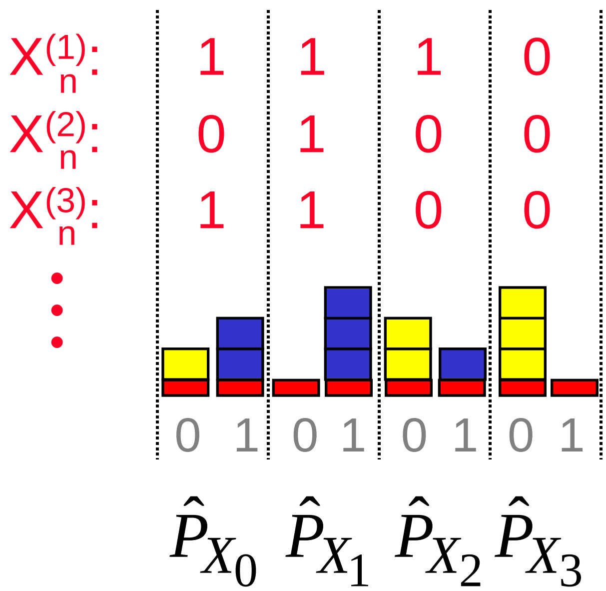
- K-L distance estimate is biased
- Use the bootstrap to estimate (and remove) bias and to estimate confidence interval
  - $\mathbf{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(R)}\}$ original dataset
  - Randomly select $\mathbf{X}^{(r)}$ to create a new "
    $\mathbf{X}^* = \{\mathbf{X}^{(2)}, \mathbf{X}^{(5)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(1)}\}$ (R compo
  - Use each bootstrap dataset to estimate probabilities and distance
  - Repeat the bootstrap estimates hundreds of time to sample the distance estimate's distribution
  - Use this distribution to estimate bias and confidenc intervals

Nuance #2 in estimating K-L distance

- K-L distance estimate subject to "infinities" if estimate of second distribution equals zero

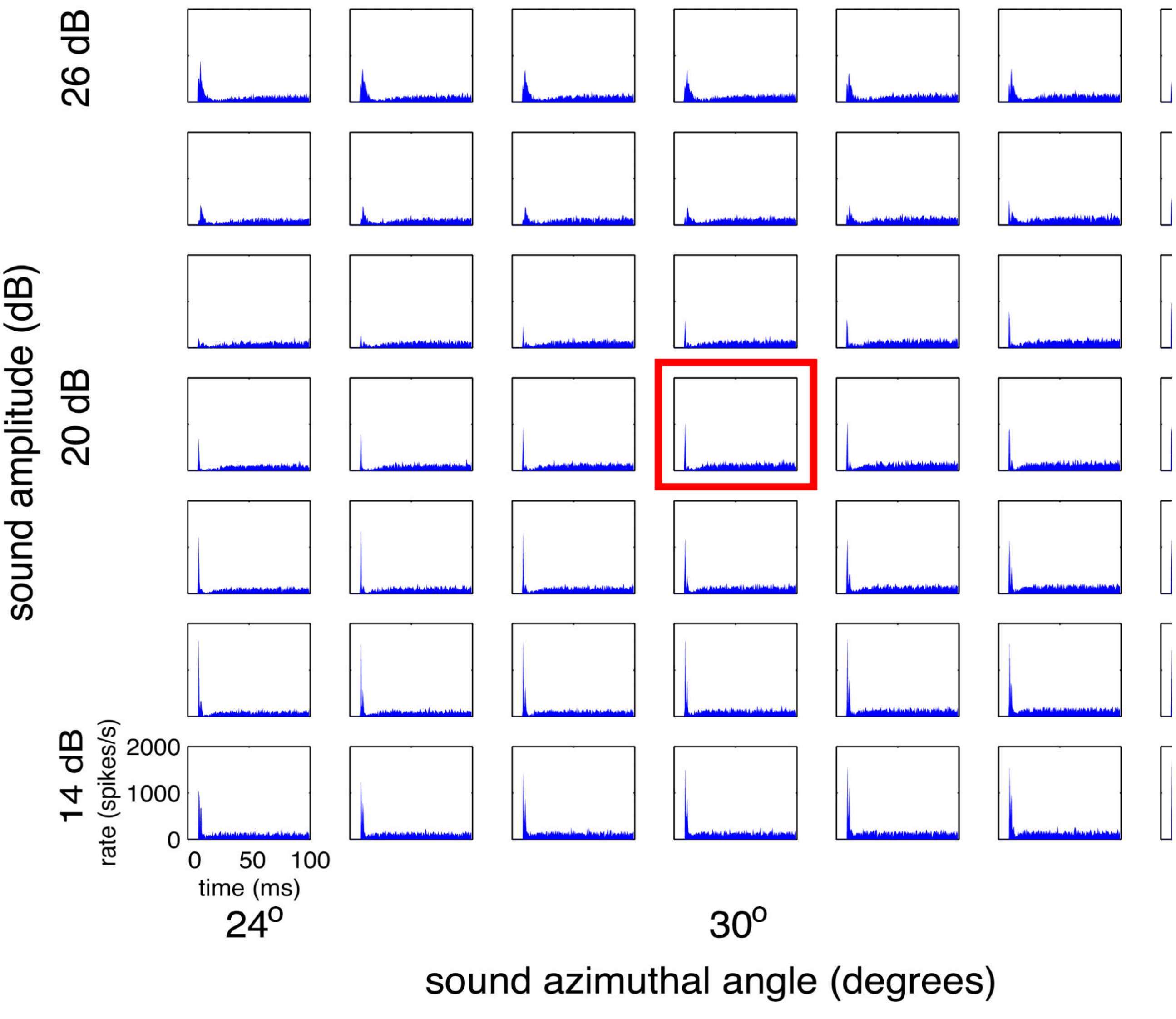
$$D_{\mathbf{X}}(\alpha_1 \| \alpha_0) = \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}, \alpha_1) \log \frac{p_{\mathbf{X}}(\mathbf{x}, \alpha_1)}{p_{\mathbf{X}}(\mathbf{x}, \alpha_0)}$$

- Modify probability estimate by "preloading" bins with a nonzero value

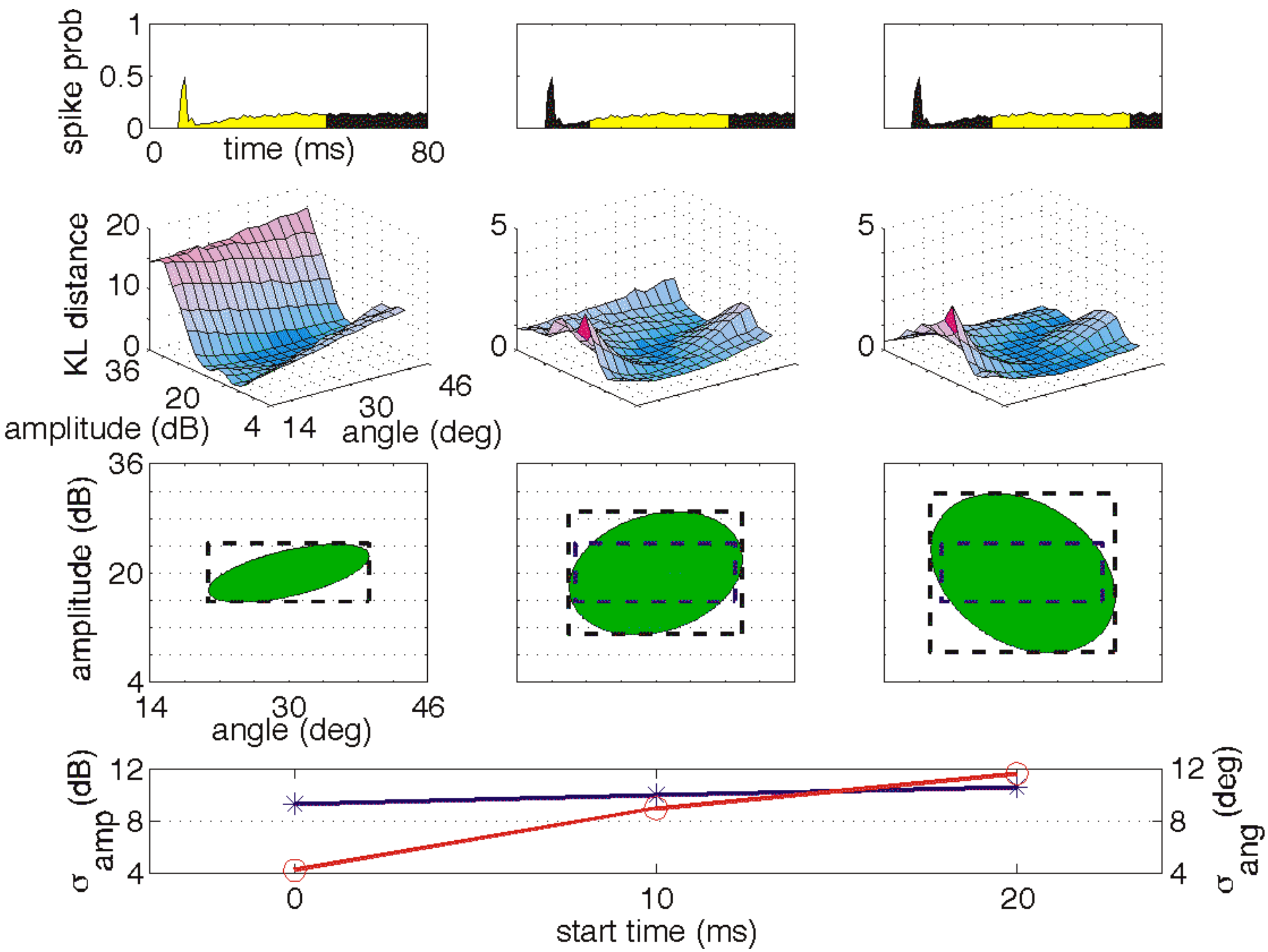


Results from universal source coding give optimal preloading value of 1/2

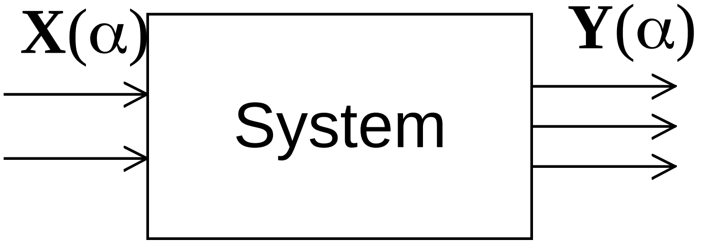
LSO response to binaural stimuli



LSO amplitude-azimuth processing



Systems and information processing



- Measure the distance between $\mathbf{X}(\alpha_0)$ and $\mathbf{X}(\alpha_1)$ and between $\mathbf{Y}(\alpha_0)$ and $\mathbf{Y}(\alpha_1)$.
- Information theoretic distance measures obey the Data Processing Theorem:

$$d_{\mathbf{X}}(\alpha_0,\alpha_1) \geq d_{\mathbf{Y}}(\alpha_0,\alpha_1)$$

  Systems cannot increase how well information is represented by their inputs
- All information-theoretic distances reflect estimation performance (mean-squared error)
- Only a few can be related to classification

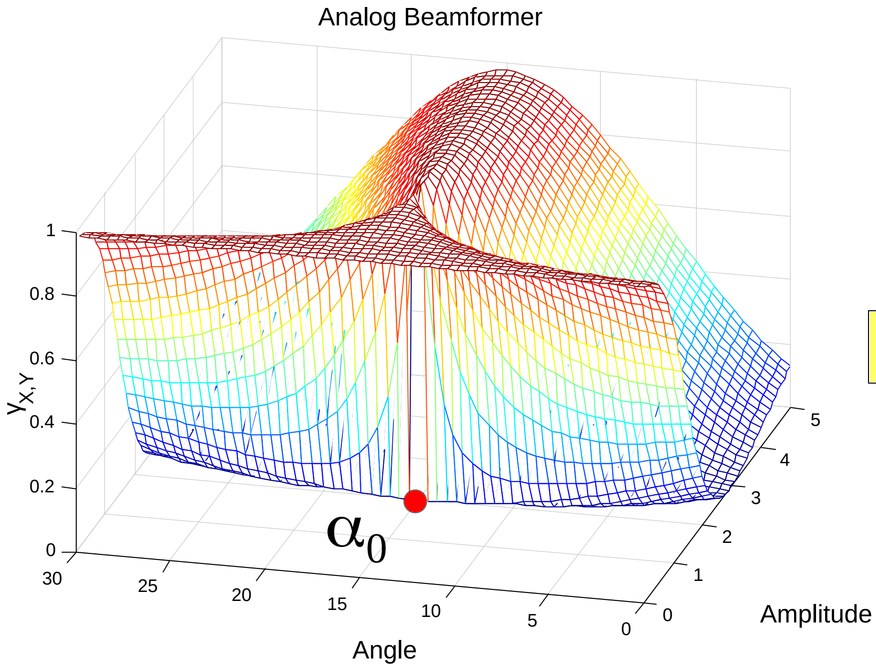
Systems and information processing

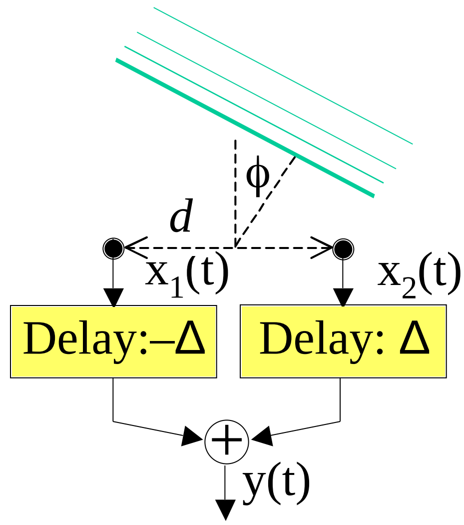
- Quantify a system's information processing performance with the information transfer ratio

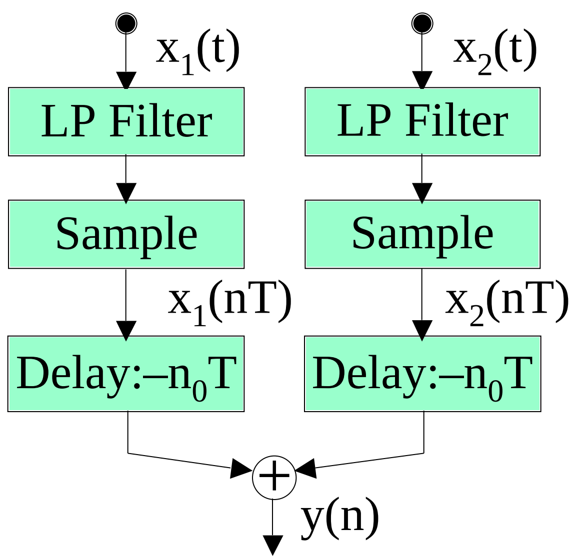
$$\gamma_{\mathbf{X},\mathbf{Y}}(\alpha_0,\alpha_1) = \frac{d_{\mathbf{Y}}(\alpha_0,\alpha_1)}{d_{\mathbf{X}}(\alpha_0,\alpha_1)}$$

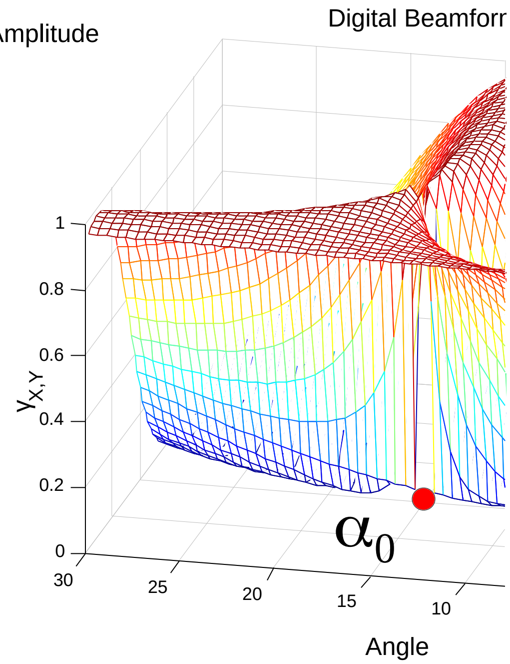
- Properties
  - $\gamma_{\mathbf{X},\mathbf{Y}} \leq 1$ (Data Processing Theorem)
    - If $\gamma_{\mathbf{X},\mathbf{Y}}(\alpha_0, \alpha_1) \approx 1$, the information change is well encoded in the output signal.
    - If $\gamma_{\mathbf{X},\mathbf{Y}}(\alpha_0, \alpha_1) \ll 1$, the information change is poorly encoded in the output signal
  - $\gamma_{\mathbf{X},\mathbf{Y}}(\alpha_0, \alpha_0+\delta\alpha)$ is *locally invariant* to the choice of distance measure

Information filtering: Array processing



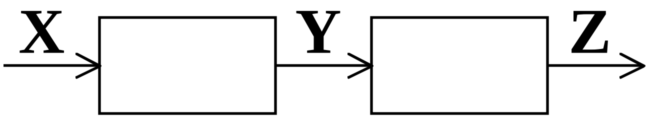






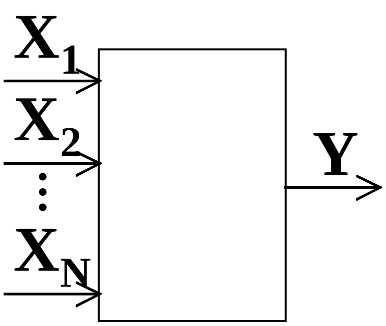
System theory of information processing

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,

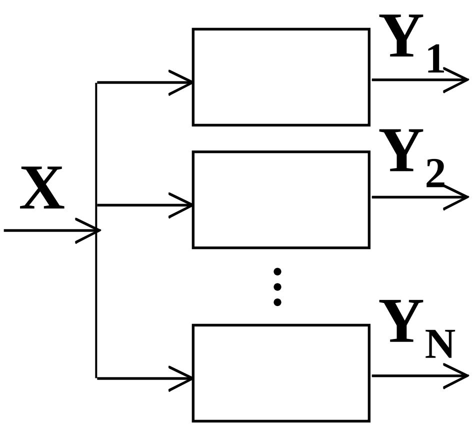
$$\frac{1}{\gamma_{\mathbf{X},Y}} = \sum_{n} \frac{1}{\gamma_{X_n}}$$

$$\gamma_{\mathbf{X},Y} \le \min_{n}$$

Limits of population coding

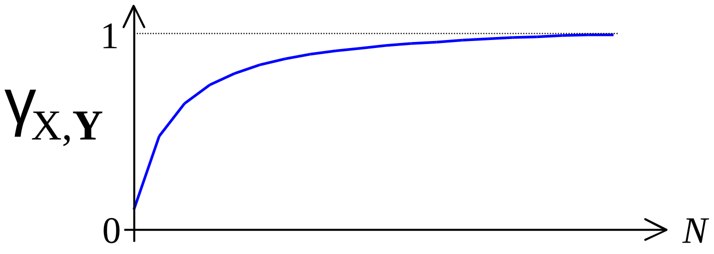
- *New result*

  A set of identical, independent neurons receive the same input. If each output reflects the information change, as the size of the population increases the population represents the information expressed by th input without loss. This result applies regardless of th processing or the code.



Conjecture: $\gamma_{X,\mathbf{Y}}(N) \approx$

or $\gamma_{X,\mathbf{Y}}(N) \approx$



Summary

- A *theory of information processing* must not depend on th nature of the signals representing information
- Use information-theoretic distance measures that reflec optimal system performance
- Studying how systems and signals respond to information (semantic) changes essential