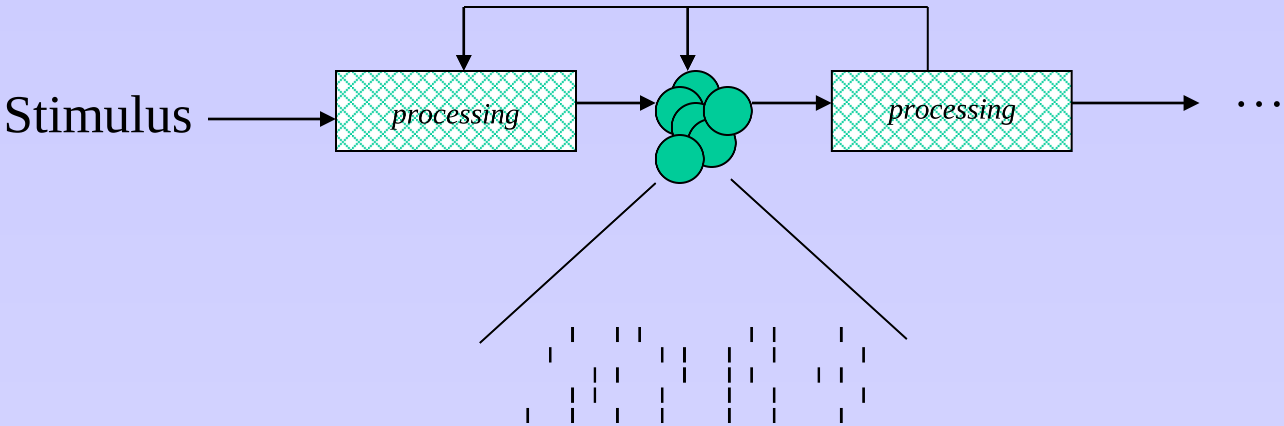
Class 13

Information Theory (intro)

- Final: Wednesday May 11th
- Projects write ups due Friday May 13th, noon.

What is **information**?



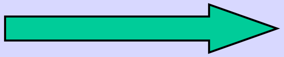
What does the response of a neuron tell us about a stimulus?

(e.g. orientation, color, facial identity…)

Vs.

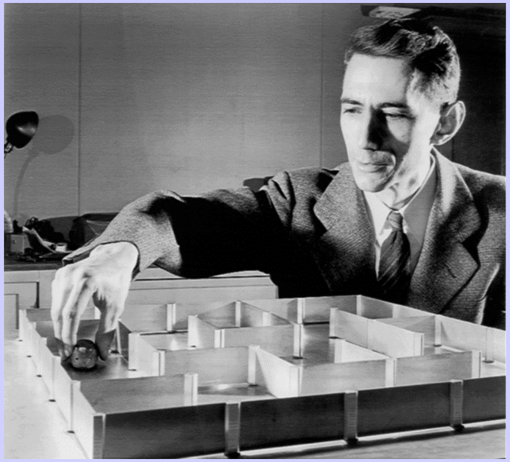
How much does the response of a neuron tell us about a stimulus?

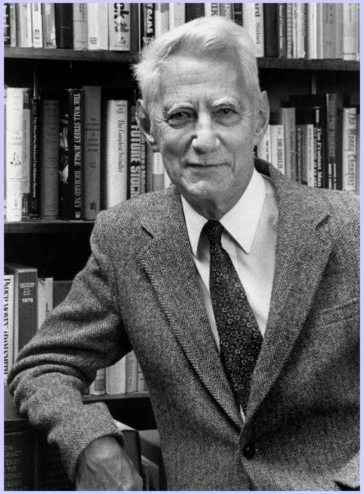
(e.g. 20%, 50%, 3 bits …'information capacity')



Need family of stimuli, many trials

Information Theory





Claude Elwood Shannon
1916-2001

- A Mathematical Theory of Communication (1948). Bell Labs.

Information ⇔ Communication

- "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point".
- The information content of a message consists of the number of 1s and 0s it takes to transmit it.
- But…: The goal of the nervous system is **not just** to transmit information

Entropy

Hypothesis: Neural response (spike train) constitute a (noisy) code.

Entropy: measure of the capacity of the 'code'.

- Response characterized by firing rate (e.g. Nb spikes/Trial length)
- Shannon Entropy = measure of how 'surprising/interesting' a response is.

P(r)= probability of getting response r

h(P(r)) = entropy of r = measure of 'surprise/interest'

Properties:

- h(1) → 0, h(0) → large positive
- Surprise/interest for 2 independent neurons: h(p1.p2)=h(p1)+h(p2)

$$h(P(r)) = -\log_2(P(r))$$

???

$$H = \sum_r W_r h(P(r)) = -\sum_r W_r \log_2(P(r))$$

(across a set of responses)

Entropy

$$H = \sum_r W_r h(P(r)) = -\sum_r W_r \log_2(P(r))$$

Constraints on W_r:

- Responses with very small (0) probability should contribute 0 'surprise'.
- Responses with very large probability (1) should contribute 0 'surprise'.

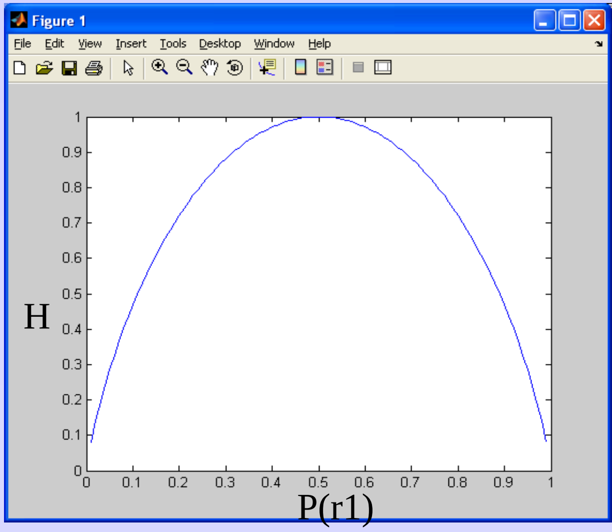
$$H = -\sum_r P(r)\log_2(P(r))$$

- If a neuron responds reliably only 1 way with rate r: H=0
- If a neuron responds in only one of 2 ways r1, r2:

p(r1)+p(r2)=1 and

$$H = -(1-p(r1))\log_2(1-p(r1)) - p(r1)\log_2(p(r1)) = H(x=r1)$$

A code consisting of only 2 responses has maximum entropy when both responses are equally likely: *1 bit entropy*.



Mutual Information

- Entropy is a measure of general response variability (all stimuli/response together).
- There is information about a particular stimulus if the variability in response to repeated presentation of that stimulus is smaller than the variability in response to repeated presentations of all-different stimuli.

Entropy of the responses due to s (only):

$$H_s = -\sum_r P(r \mid s) \log_2 (P(r \mid s))$$

Need to measure 'surprise/variability' **not due** to stimulus variation? :

$$\sum_s P(s) H_s = \text{noise entropy} = \mathrm{H_n}$$

About r About s

$$M = H - H_n$$

$$H = -\sum_r W_r \log_2 (P(r))$$

M= *Mutual Information*: How much entropy is actually used. How much knowing *r* reduces the uncertainty about *s* having occurred.

Mutual Information

$$M = H - H_n$$

$$M = -\sum_r P(r)\log(P(r)) - \sum_s P(s)H_s$$

$$= -\sum_r P(r)\log(P(r)) + \sum_{s,r} P(s)P(r|s)\log(P(r|s))$$

By definition of conditional probability $P(r) = \sum_s P(s)P(r|s)$

$$\boxed{M = \sum_{r,s} P(s)P(r|s)\log\left(\frac{P(r|s)}{P(r)}\right)}$$

Mutual Information

$$P(r) = \sum_s \underbrace{P(s)P(r \mid s)}$$

P(r,s) = ‘joint probability’ = P(r)P(s|r)

P(r,s) = probability of stimulus *s* appearing and response *r* being evoked.

$$\boxed{M = \sum_{r,s} P(r,s) \log_2 \left(\frac{P(r,s)}{P(r)P(s)} \right)}$$

Note: Information that a set of responses conveys about a set of stimuli = Information that a set of stimuli conveys about a set of responses.

Mutual Information: Fun facts

- If responses are unrelated to the identity of the stimulus

$$M = \sum_{r,s} P(s)P(r\,|\,s)\log_2\left(\frac{P(r\,|\,s)}{P(r)}\right)$$

$$M = \sum_{r,s} P(r,s)\log_2\left(\frac{P(r,s)}{P(r)P(s)}\right)$$

$\Rightarrow$ P(r|s)=P(r) $\Rightarrow$ M=0

- If each stimulus *s* reliably produces a different response r_s

$\Rightarrow$ $P(r_s)=P(s)$ $\Rightarrow$ P(r|s)=1 only if $r=r_s$.

$$M = \sum_{r,s} P(s)P(r\,|\,s)\log_2\left(\frac{P(r\,|\,s)}{P(r)}\right) = P(s)\log_2\left(\frac{1}{P(r_s)}\right)$$

$M = -P(s)\log_2\big(P(s)\big)$ $\Leftrightarrow$ M = Entropy of the stimulus

Mutual Information: Fun facts

- Case when there are only 2 responses (r1 and r2) to 2 stimuli (s1 and s2). The probability of incorrect response is P_i (< 0.5). If s1 and s2 are presented with equal probability.

Prob to be correct: $P(r1|s1)=P(r2|s2)=1-P_i$

Prob to be wrong: $P(r1|s2)=P(r2|s1)=P_i$

$P(s1)=P(s2)=1/2$

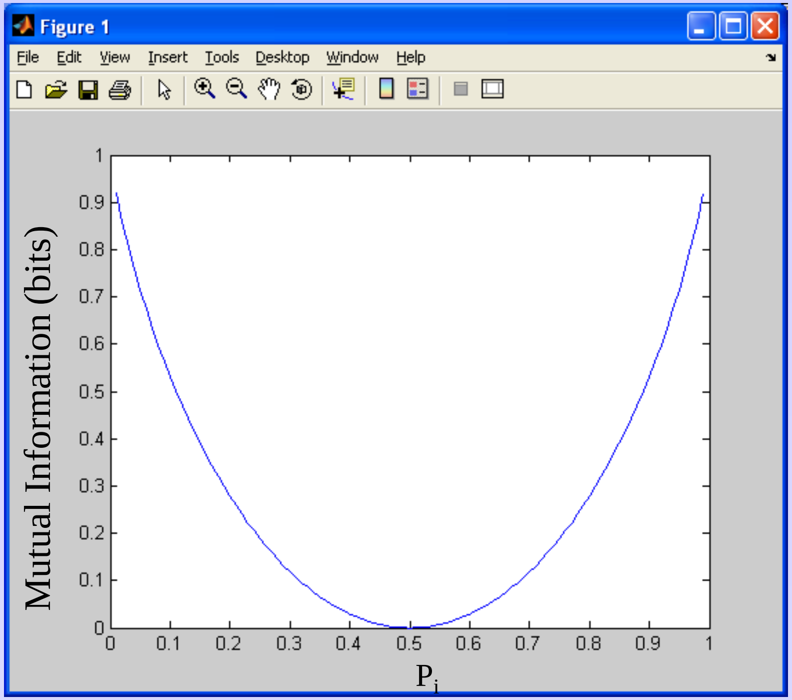
$$\boxed{M=1+(1-P_i)\log_2(1-P_i)+P_i\log_2(P_i)}$$

and …

$P_i=0 \rightarrow M=1$ bit

$P_i=1/2$ (random) $\rightarrow M=0$

($P_i=1 \rightarrow M=1$bit)



(Note: P_i > ½ : swap 1 and 2!)

Mutual Information

- Neurons are used for decoding: what is the probability of having s1 if r1 is observed?

$$P(s1|r1) = P(r1|s1)P(s1)/P(r1) \quad \longleftarrow \quad \textbf{Bayes theorem}$$

$$= 1-P_i$$

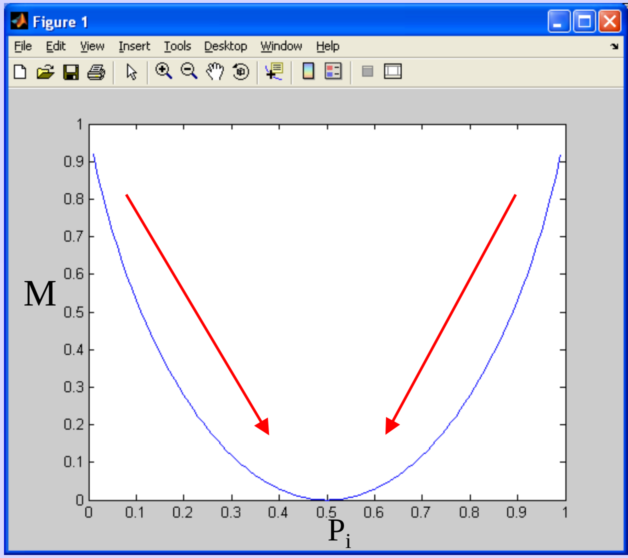
Before measurement: the expectation of getting s1 is 1/2. After measurement the expectation becomes $1-P_i$.

$$1/2 \longrightarrow 1-P_i$$

There is an increase in probability.

There is an increase in certainty.

There is a decrease in uncertainty = **M**.



Mutual Information: KL

- Addendum

Kullback-Leibler (**KL**) divergence is a kind of statistical 'distance' between 2 distributions:

$$\boxed{D(P,Q) = \sum_r P(r) \log_2 \left(\frac{P(r)}{Q(r)} \right)}$$

but

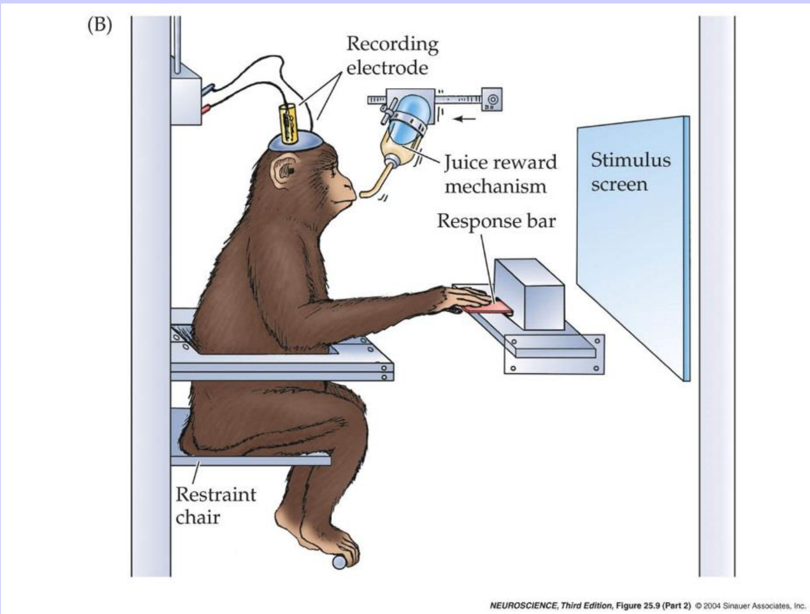
$$M = \sum_{r,s} P(r,s) \log_2 \left(\frac{P(r,s)}{P(r)P(s)} \right)$$

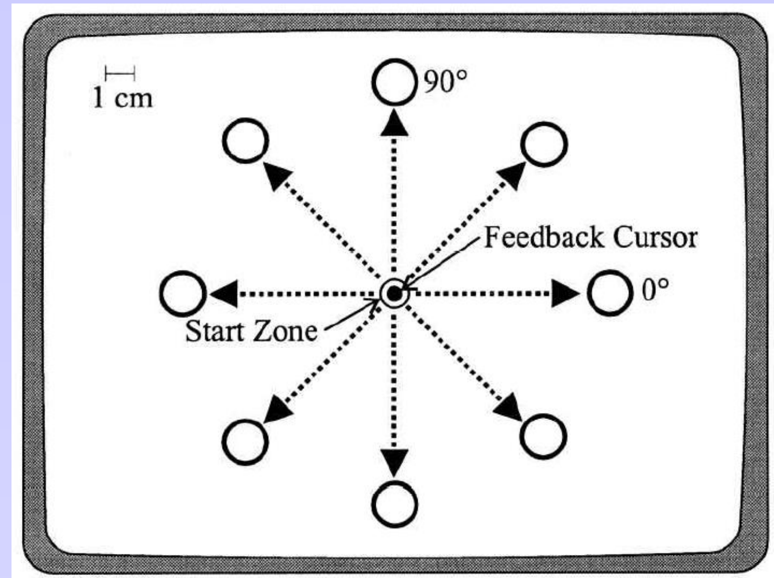
$\Rightarrow$ M=D(P(r,s),P(s)P(r))

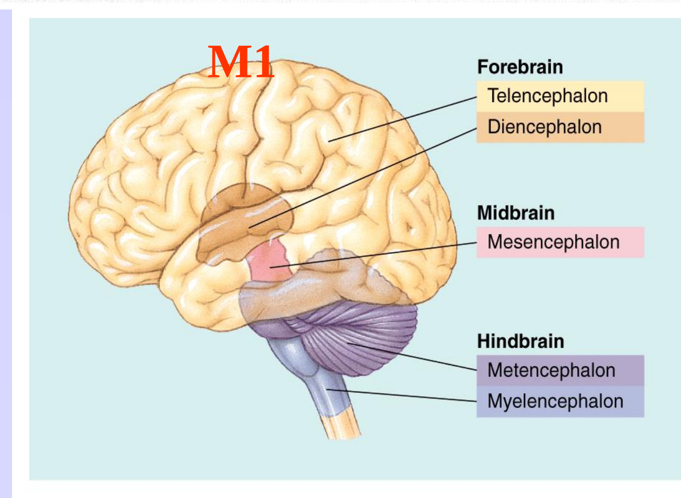
M is the KL divergence ('distance') between the actual probability distribution and the probability distribution if stimuli and responses were independent from each other.

Synchrony and information

- Can synchrony (potentially) carry information about movement direction? Is it related to firing rate?







Experiment timeline

CHT Delay Movement

~0.5s 1-1.5 s

(Hastopoulos et al, 1998)

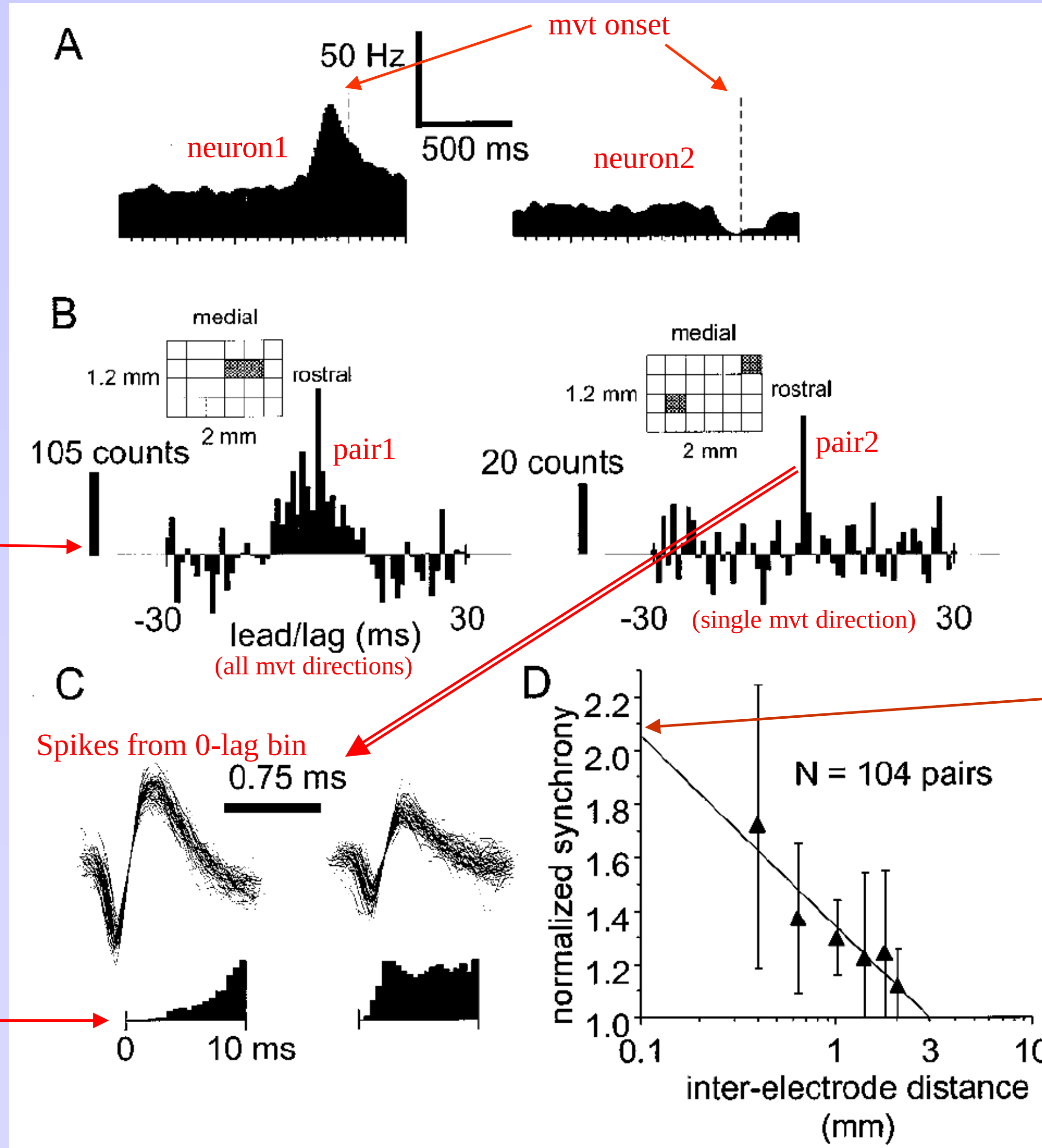
Synchrony and information

- Synchrony log-linearly decreases with cortical distance

PSTH

Cross-correlograms
(1ms bin)

Autocorrelograms
(1ms bin)



$$\frac{\text{Data 0-lag}}{\text{Shuffled 0-lag}}$$

>1

(Hastopoulos et al, 1998)

Information and Synchrony

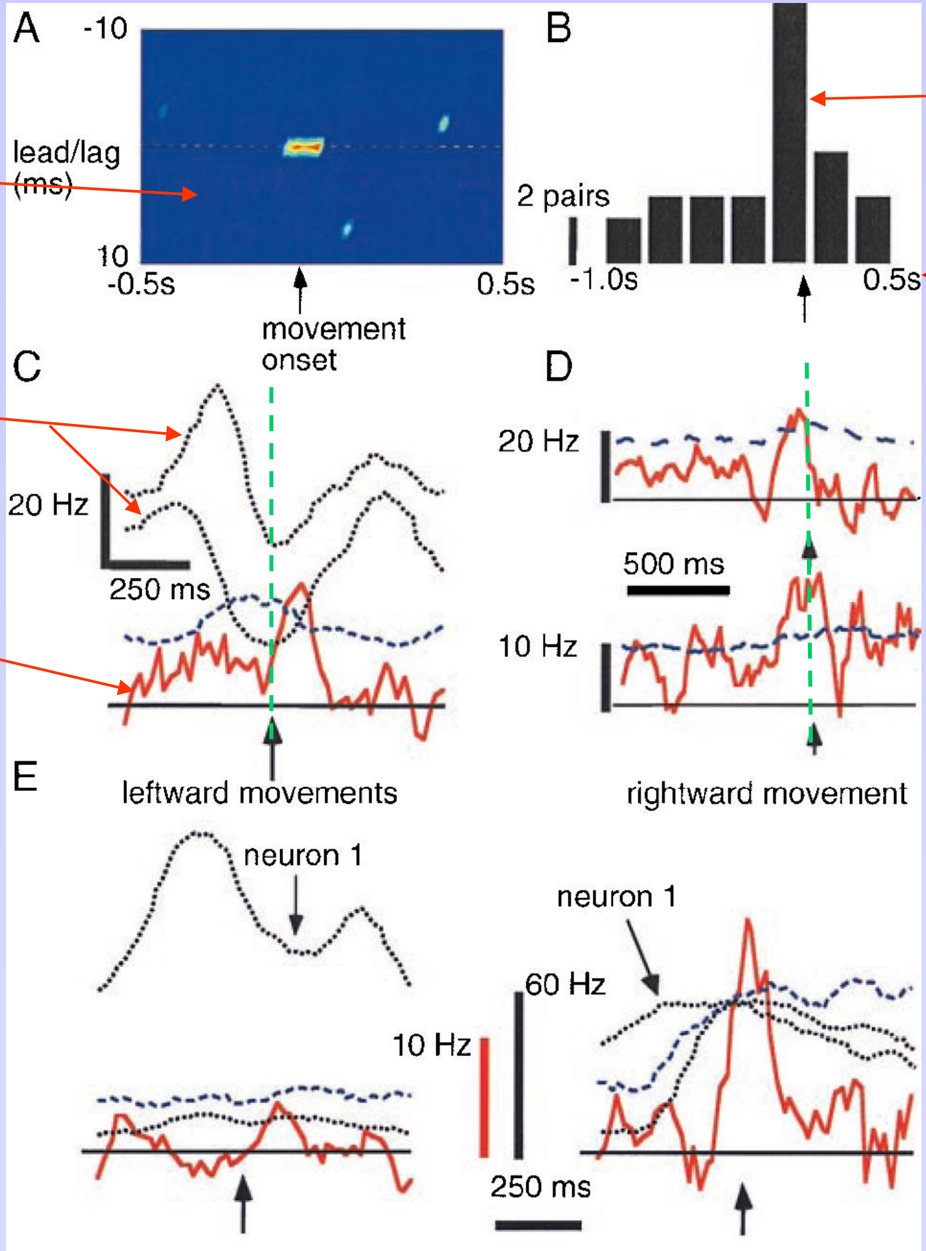
X-correlogram Vs. time

Synchrony does not depend on firing rate

Firing rates

Sliding (100 ms) x-correlations at 0-lag

Synchrony varies with movement direction: i.e. potentially carries information



Nb pairs with synchrony peaks

Time when synchrony occurs (/mvt onset)

Pair 1-2

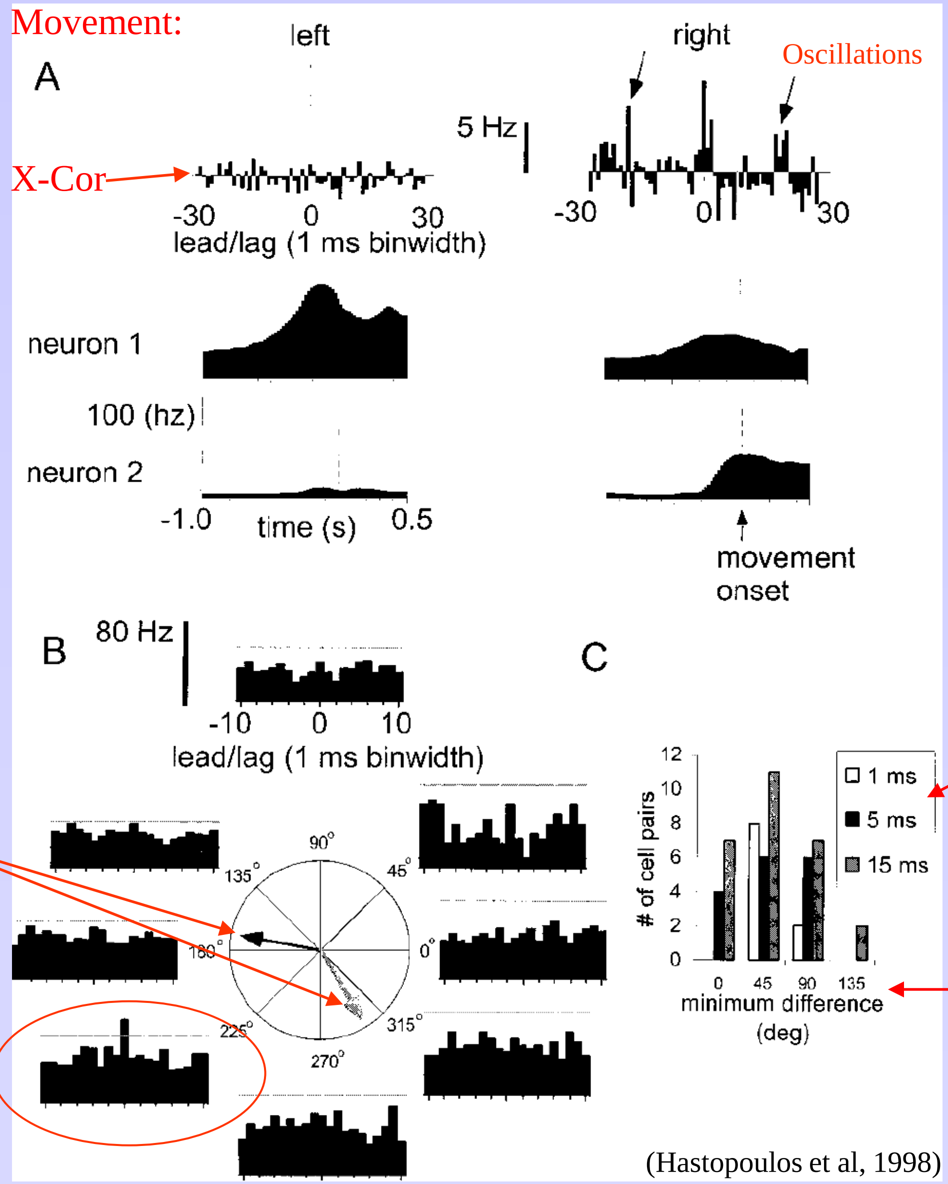
Same cell, different pairs → Different patterns of synchrony

Pair 1-3

(Hastopoulos et al, 1998)

Information and Synchrony

- Directional tuning in synchrony is different from direction tuning in firing rate



Firing rate direction tuning for n1 and n2

Significant Synchrony tuning 0-lag

Bin=Synchrony 'time scale'

'synchrony tuning' – 'firing rate tuning'

Information and Synchrony

$$M = \sum_{r,s} P(s) P(r \mid s) \log_2 \left(\frac{P(r \mid s)}{P(r)} \right)$$

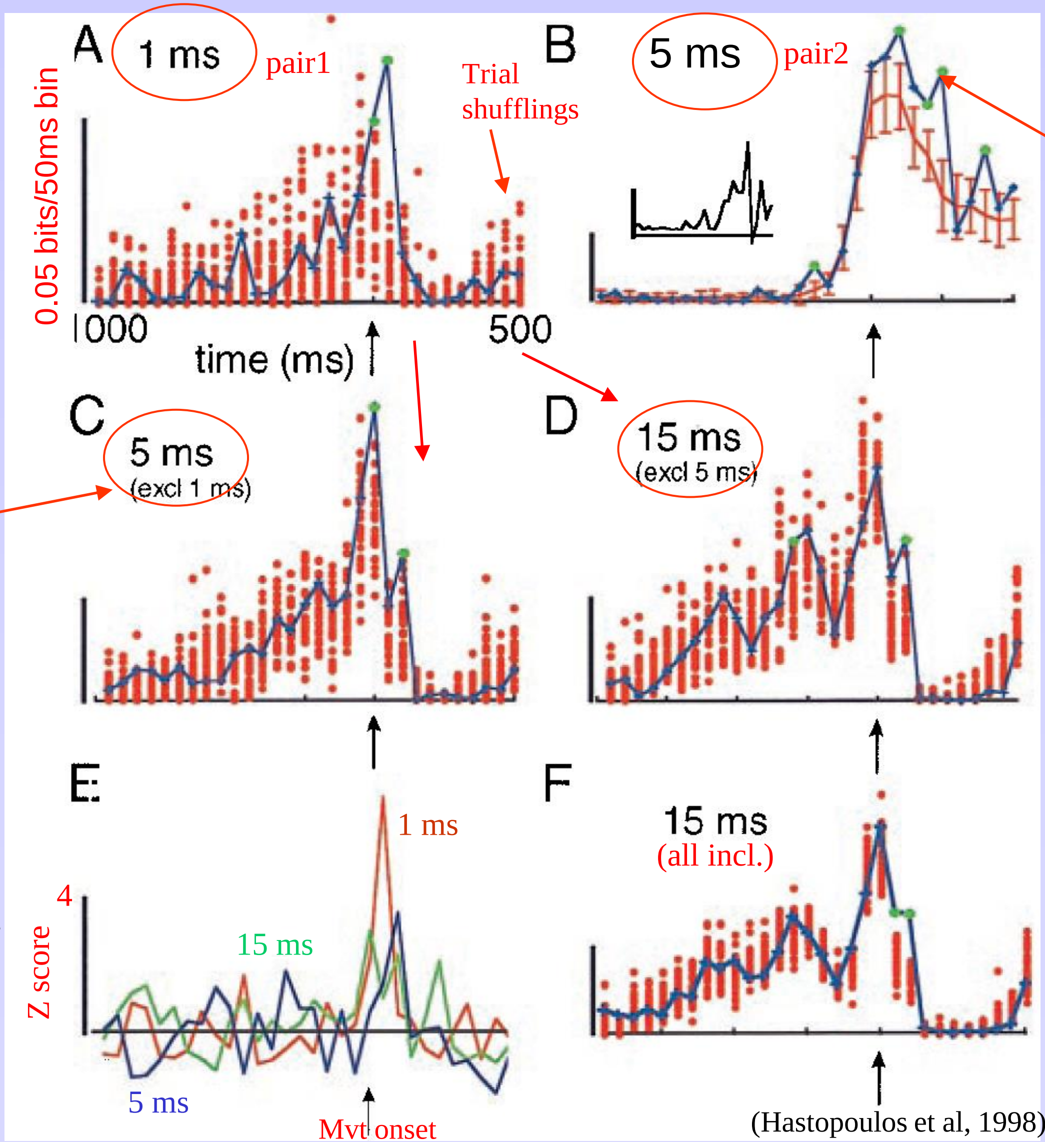
- Mutual information between synchronous (coincident) neurons and movement direction

$$M = \sum_{dir} P(dir) \sum_{coinc} P(coinc \mid dir) \log_2 \left(\frac{P(coinc \mid dir)}{P(coinc)} \right)$$

P(dir) = probability of a movement direction = set by experimenter (= stimulus).
P(coinc) = probability of finding coincident spikes (= response).

Information and Synchrony

- Temporal variations of mutual information at multiple 'time scales'



green = significant

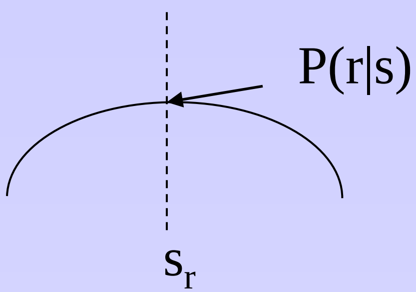
Synchrony time scale

(Hastopoulos et al, 1998)

➔ Increase in mutual information after mvt onset is robust with respect to time scale

Fisher Information and Accuracy

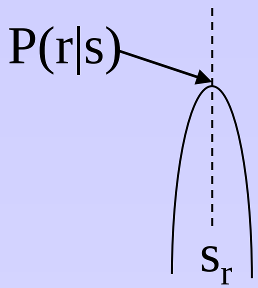
- Case where the stimulus (may) vary continuously
- $p(r|s)$ is a continuous function of the stimulus.
- $p(r|s)$ is maximum at the value of s_r that gives the response **r.**



P(r|s) NOT very selective
(small variation in s → small variations in p)

Small curvature

'poor' information about s



P(r|s) very selective
(small variation in s → large variations in p)

Large curvature

'rich' information about s

Fisher Information

- Definition

$$F(s) = \left\langle -\frac{\partial^2 \log(p(r\,|\,s))}{\partial^2 s} \right\rangle_r$$

in most conditions, *F(s)* can also be written:

$$F(s) = \left\langle \left(\frac{\partial \log(p(r\,|\,s))}{\partial s} \right)^2 \right\rangle_r = \sum_r p(r\,|\,s) \left(\frac{\partial \log(p(r\,|\,s))}{\partial s} \right)^2$$

Note: $F(s) \geq 0$

Fisher Information and accuracy

- Imagine a stimulus is presented many times (i.e. multiple trials).

S(s) = Estimation of a stimulus, given the responses (whatever the algorithm!)

b(s)= 'bias'=$\langle \mathbf{S} \rangle_{\text{trials}}$- s

σ(s) = variance (**S**) = 'how good one is at estimating the stimulus'

$$\sigma(s) \geq \frac{1}{F(s)}$$

'Cramer-Rao bound'

'=' if *S(s)* is the optimal estimator

- **Fisher Information**: A measure of encoding accuracy: limit to the accuracy with which any decoding scheme can extract information about a stimulus.
- Fisher Information is used in 'estimation theory'.
- See also Kanitscheider et al. (2015).

Fisher Information and discriminability

- Fisher Information can also be used to measure discriminability

High estimation accuracy

??

High discriminability

$$d' = \frac{\Delta\mu}{\sigma}$$

If unbiased estimator:

$$\Delta\mu = \Delta s_{est} = \Delta s$$

If optimal estimator:

$$\sigma(s) = \frac{1}{F(s)}$$

$$\Rightarrow \quad d' = \Delta s \sqrt{F(s)}$$

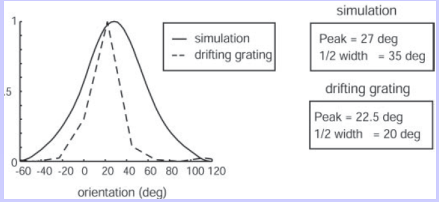
➔ The larger the Fisher information, the larger the (potential) discriminability

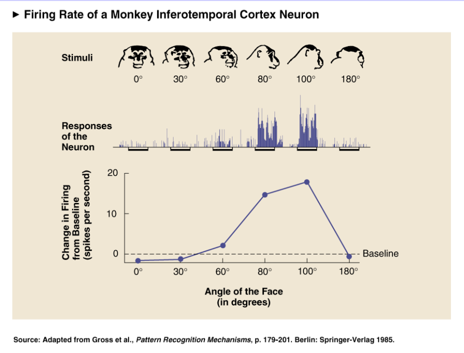
Fisher Information of a population of neurons

- Fisher information is **additive**

For N independent neurons:

$$F_{tot}(s) = \sum_{i=1}^{N} F_i(s)$$



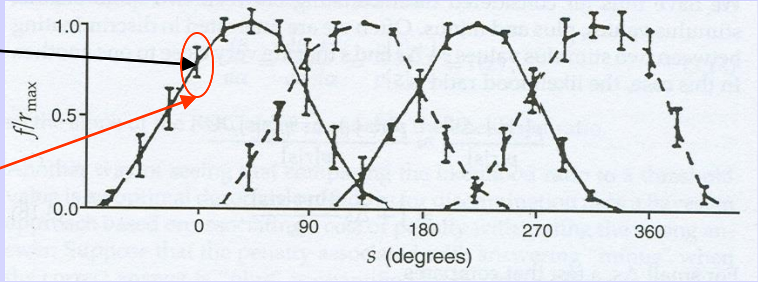


- Case where neurons have a tuning curve

Slope of the rate function

$$F_{tot}(s) = T \sum_{i=1}^{N} \frac{\left(f_i'(s)\right)^2}{\sigma_i^2(s)}$$

$f_1(s)$



Variance of the spike count of neuron i in response to s

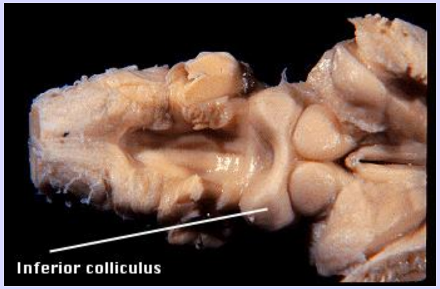
→ A neuron contributes the most to the information of a population of neurons for stimuli that make its firing rate change significantly (*not* for stimuli that elicit maximal firing rates), **and/or** when spike count variance is small.

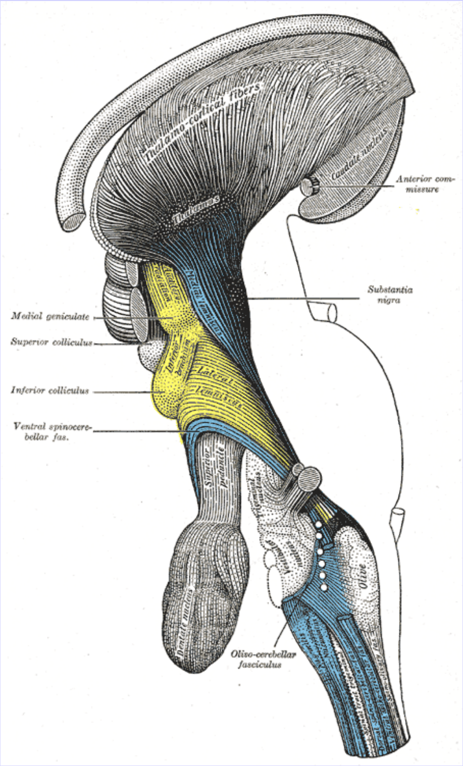
Fisher Information

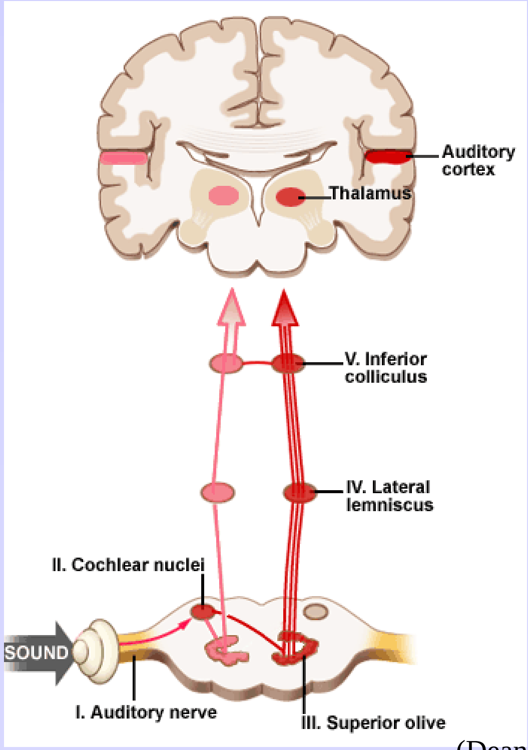
Fun facts:

- Our ability to discriminate sounds is not sensitive to overall sound intensity
- Our encoding of sounds is 'efficient', no matter what sound intensity
- The change in *firing rate* of Inferior Collicullus neuron is limited to 35 dB (hearing spans 0-120 dB)

How is efficiency achieved?







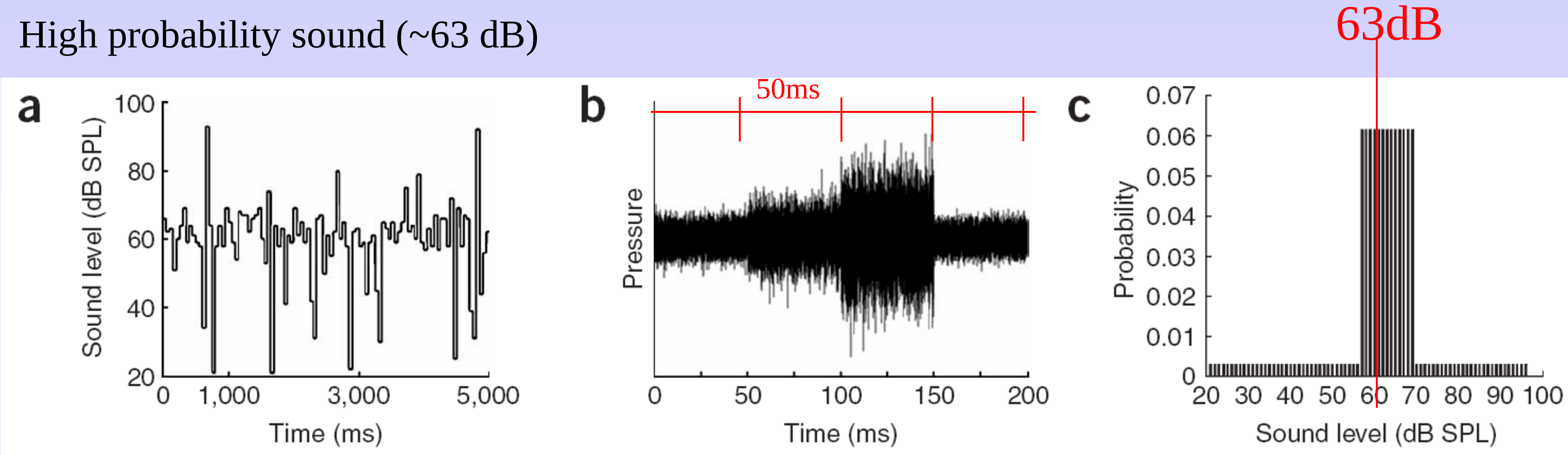
(Dean et al.2005)

Fisher Information

- Anesthetized guinea pig with earphones, inferior colliculus.
- Stimuli: 7 min trains of 50 ms white noise bursts sequence of ~X dB each.

Stimuli:

High probability sound (~63 dB)



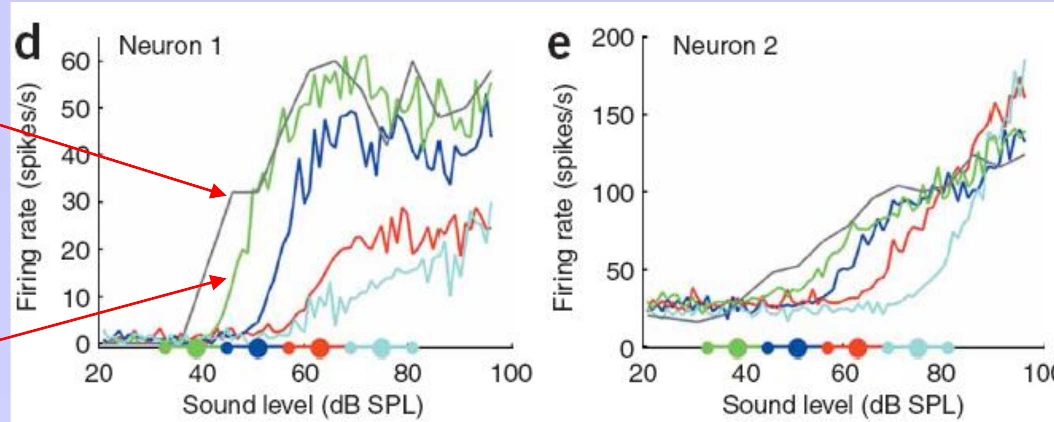
(Dean et al.2005)

Fisher Information

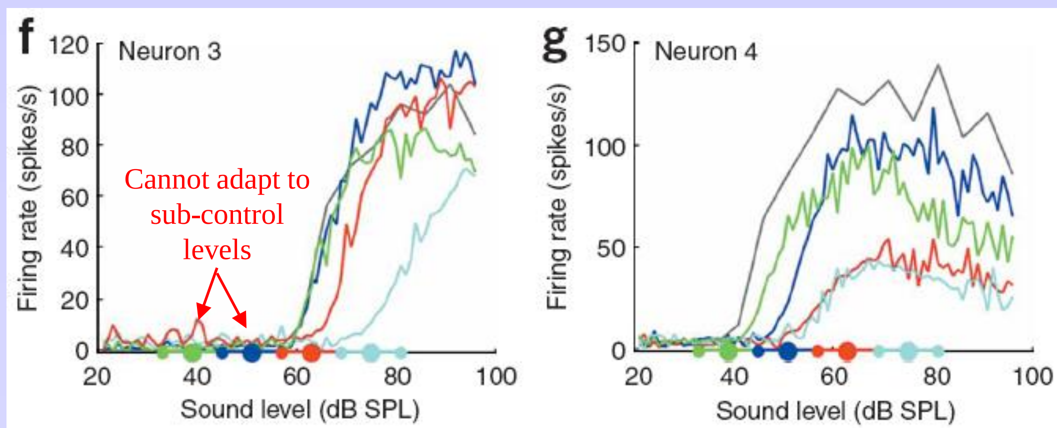
Equiprobable (mixed) sound levels
(control rate function)

Unimodal
Sound level distributions

39 dB
51 dB
63 dB
75 dB



→ Neuron FR curve adapts to the statistics of the stimulus



(Dean et al.2005)

- Firing rate function shifts towards most probable sound level (never below control).
- Reduction in slope with high sound levels.

Fisher Information

- Do the shift and slope changes improve ‘coding accuracy’.
- Nothing is known of the actual way sounds are coded…

Use information theory!

Accuracy = ‘variance of spike count of the estimate’.

Bounded by 1/F(s)

The bound can *in principle* be reached (Max Likelihood estimator)

Use Fisher Information as a measure of ‘accuracy’

$$f_a(s) = \sum_r P_a[r|s] \left(\frac{d \ln P_a[r|s]}{ds} \right)^2$$

s = sound level

r = spike count in 50 ms (8 ms delay)

$$F(s) = \sum_a f_a(s)$$

Fisher Information

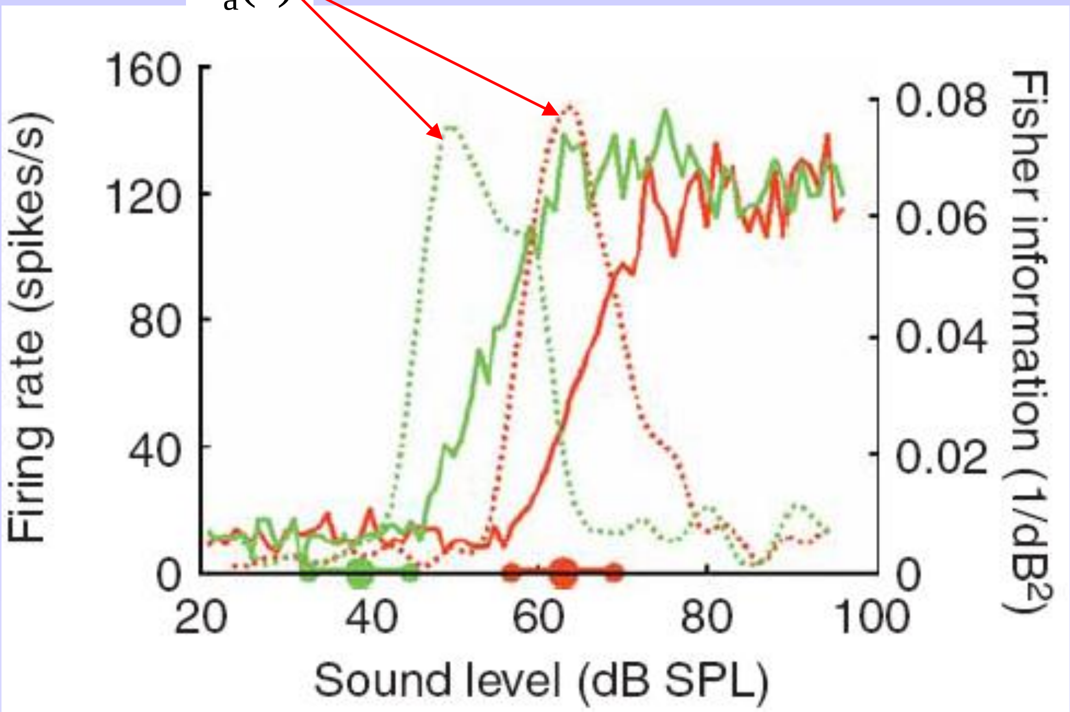
- Peak in Fisher information is at or near the mean stimulus intensity

→ Highest stimulus coding accuracy

$$f_a(s) = \sum_r P_a[r|s]\left(\frac{d\ln P_a[r|s]}{ds}\right)^2$$

$$F_{tot}(s) = T\sum_{i=1}^{N}\frac{\left(f_i'(s)\right)^2}{\sigma_i^2(s)}$$

$f_a(s)$



(Dean et al.2005)

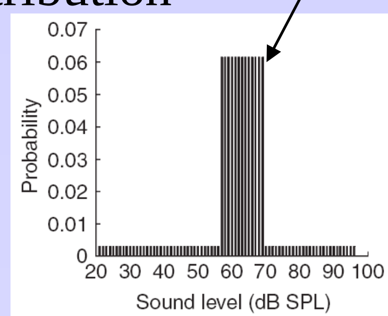
Rate-level function shift → $f_a(s)$ shift

Fisher Information

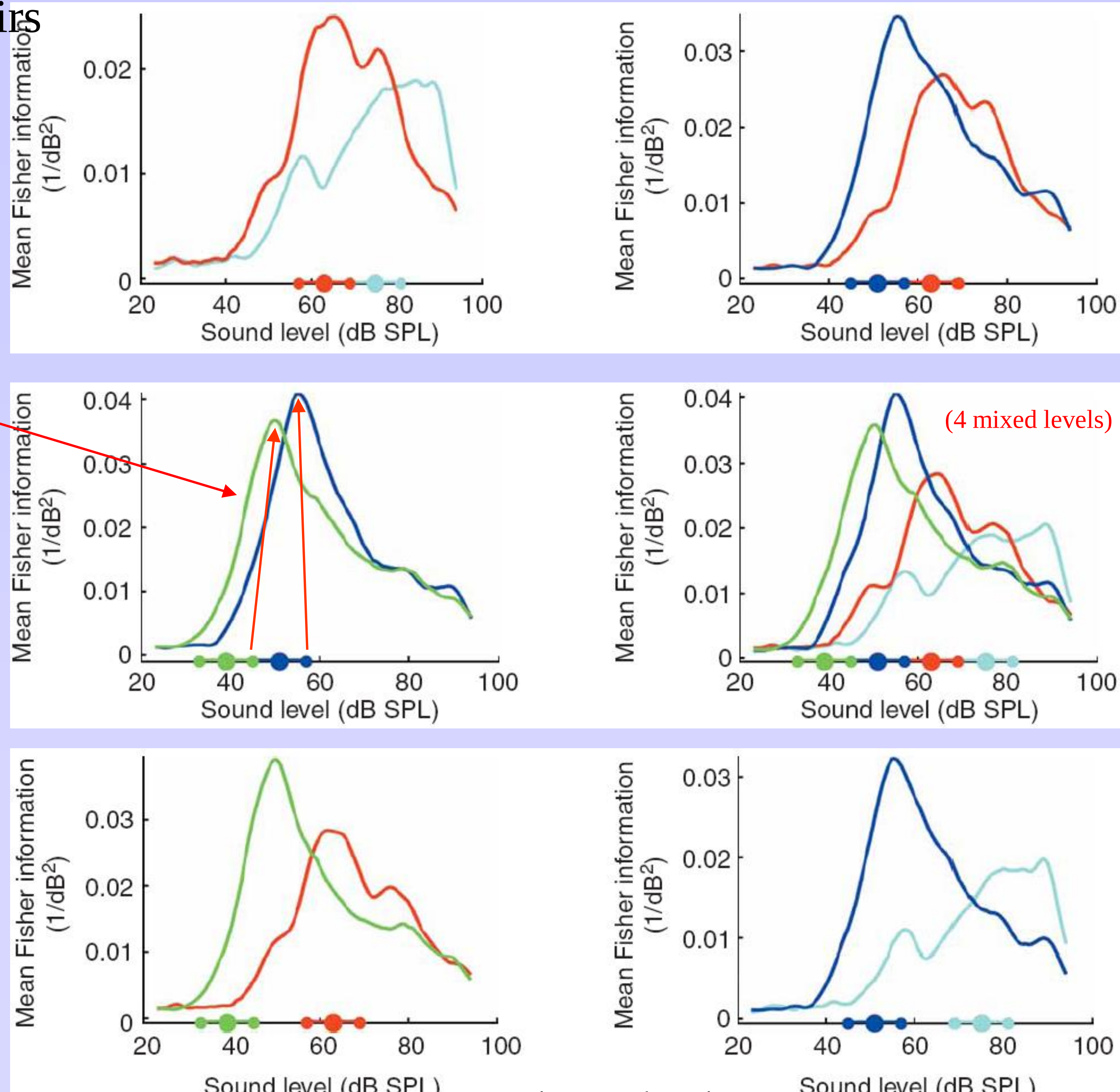
- Mixed presentations of pairs of sound level distributions
- Population Fisher information

$$F(s) = \sum_a f_a(s)$$

Peak ‘accuracy’ at upper boundary of probability distribution



$F_{tot}(s)$

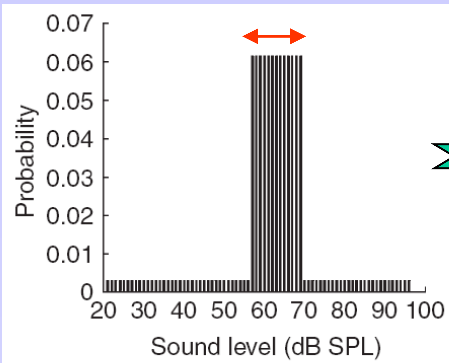


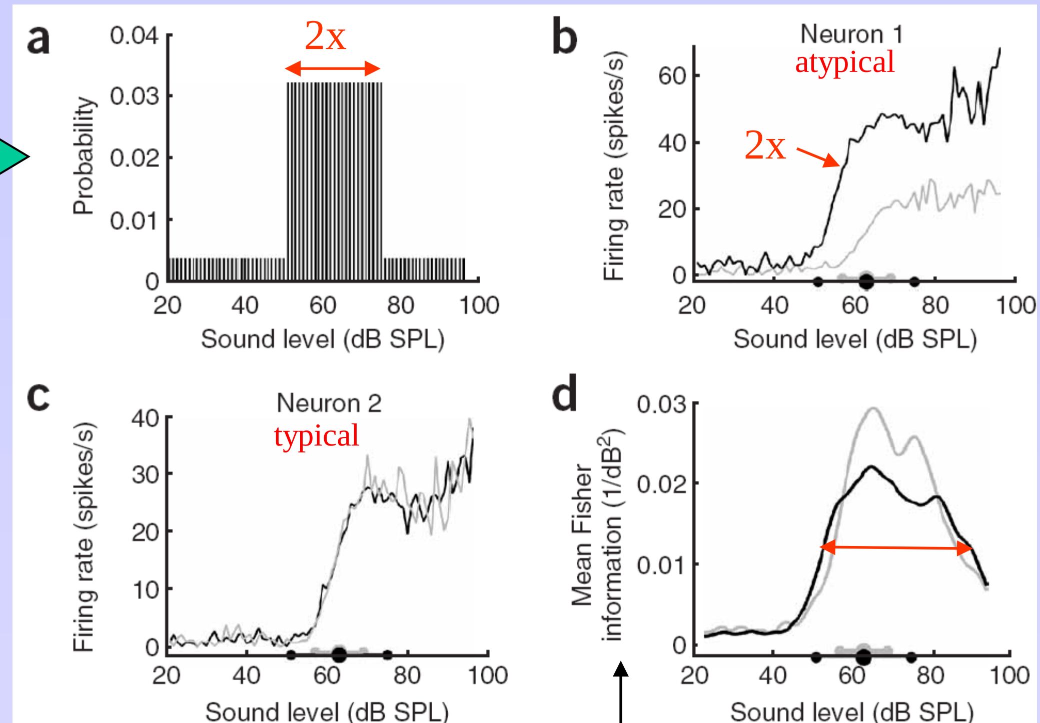
(Dean et al.2005)

➔ Adaptation to stimulus level for (potential) maximal accuracy

Fisher Information

- Adaptation to stimulus variance?





$F_{tot}(s)$

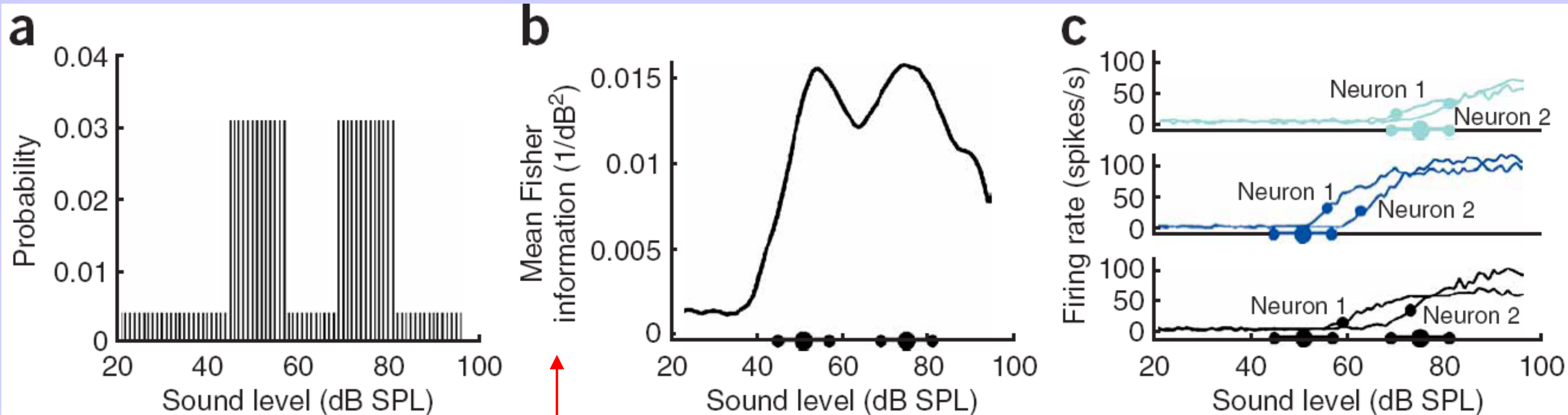
(Dean et al.2005)

→ Slight adaptation of accuracy to stimulus variance in spite of lack of firing rate adaptation

Fisher Information

- Adaptation to stimulus bimodality?

(Dean et al.2005)



$F_{tot}(s)$

No firing rate-level adaptations

Accuracy adaptation of the *population*

Note: in general…

high threshold neurons → shift towards the high sound-level probability peak
low threshold neurons → shift towards the low sound-level probability peak