A Bayesian Estimator of Entropies in a Severely Undersampled Regime: Theory and Applications to the Neural Code

Ilya Nemenman
LANL/CCS-3

http://nsb-entropy.sf.net

Entropy (unique measure of randomness, in bits)

$$S[X] = -\sum_{x=1}^{K} p_x \log p_x = -\langle \log p_x \rangle$$

$$0 \le S[X] \le \log K$$ (number of "bins")

$$N(x_0, \sigma^2) \;\Rightarrow\; S[X] = \frac{1}{2}\log(2\pi e \sigma^2)$$

Why knowing entropy is interesting?

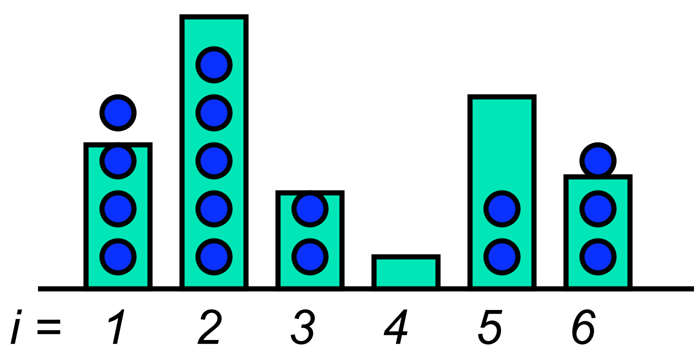
- Information content of symbolic sequences
  - Spike trains
  - Bioinformatics
  - Linguistics
- Dynamical systems
  - Complexity of dynamics
  - Dimensions of strange attractors
- Rare events statistics
- …

Why is this a difficult problem?

Maximum likelihood (plug-in) estimation:

$p_i, \quad i = 1 \ldots K$

(*K* - # of bins)



$$p_i^{ML} = \frac{n_i}{N}$$

(*N* - sample size)

$$S_{ML} = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N}$$

!

$$\langle S_{ML} \rangle \leq -\sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S$$

Why is this a difficult problem?

$$\langle S_{ML} \rangle \leq -\sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S$$

log *K*

$$\text{bias} \propto -\frac{2^S}{N} \gg (\text{variance})^{1/2} \propto \frac{1}{\sqrt{N}}$$

Fluctuations underestimate entropies

(and usually overestimate mutual informations)

(Need smoothing)

Why is this a difficult problem?

- Events of negligible probability may contribute a lot to entropy due to log (not true for high order entropies, such as Renyi ≥2)

$$R_\alpha = \frac{1}{1-\alpha}\log\sum p_i^\alpha$$

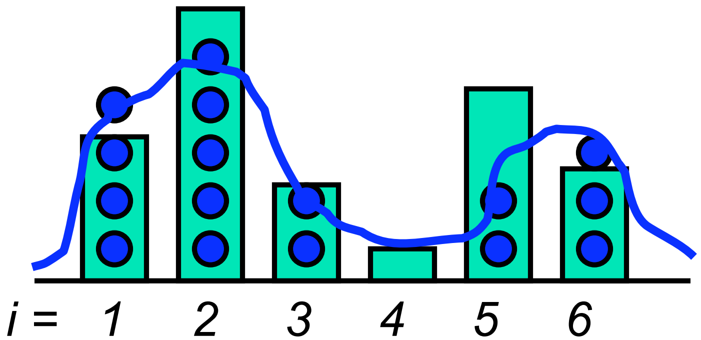
- Small errors in *p* --> large errors in *S*
- *S*(best *p*) ≠ best *S*(*p*)
- But can use *R* to bound *S*

Why is this a difficult problem
No go theorems

For N i.i.d. samples from a distribution on K (countable or $>>N$) bins (note that non-i.i.d is the same as $K \dashrightarrow \infty$):

- No universal rates of convergence exist for LZ, plug-in, and other estimators (Antos & Kontoyiannis, 2002; Wyner & Foster, 2003)
- For and universal estimator, there is always a bad distribution with bias ~1/log *N.*
- No finite variance unbiased entropy estimators (Grassberger 2003, Paninski 2003)
- No universally consistent multiplicative estimator (Rubinfeld et al, 2002)
- Universal consistent estimators only possible for *N/K*-->const (Paninski, 2003)

In other words: Correct smoothing possible only for…



$$S \le \log N$$

(often not enough)

Incorrect smoothing = over- or underestimation.

Developed for problems ranging from mathematical finance to computational biology.

For estimation of entropy at $K/N \le 1$ see:
Grassberger 1989, 2003, Antos and Kontoyiannins 2002, Wyner and Foster 2003, Batu et al. 2002, Paninski 2003, Panzeri and Treves 1996, Strong et al. 1998

What if $S > \log N$?

But there is hope (Ma, 1981):

For uniform K-bin distribution the first coincidence occurs for

$$N_c \sim \sqrt{K} = \sqrt{2^S}$$

$$S \sim 2 \log N_c$$

Time of first coincidence

Can make estimates for square-root-fewer samples!

Can this be extended to nonuniform cases?

- Assumptions needed (won't work always)
- Estimate entropies without estimating distributions (good entropy estimator $\neq$ good distribution estimator).

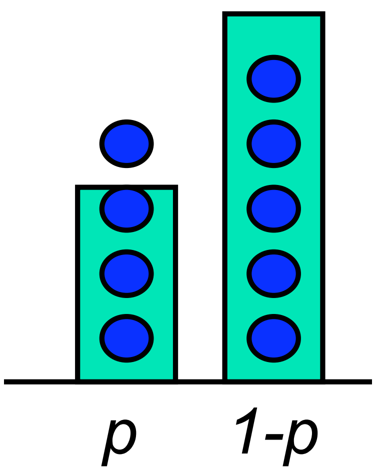
What if $S > \log N$?

- Imagine sampling sequences of length m from N_c samples with replacements.
- $\sim N_c^m$ different sequences
- Uniformely distributed due to equipartition $\log p = -mS$
- Thus using Ma: $mS = 2 \log N_c^m$, and $S = 2 \log N_c^m$
- What happens earlier: non-independence of sequences, or equipartition?
- Sometimes may estimate entropies with little bias using coincidences (LZ) even for non-uniform distributions.

What is unknown?

Binomial distribution:

$$S = -p \log p - (1-p)\log(1-p)$$

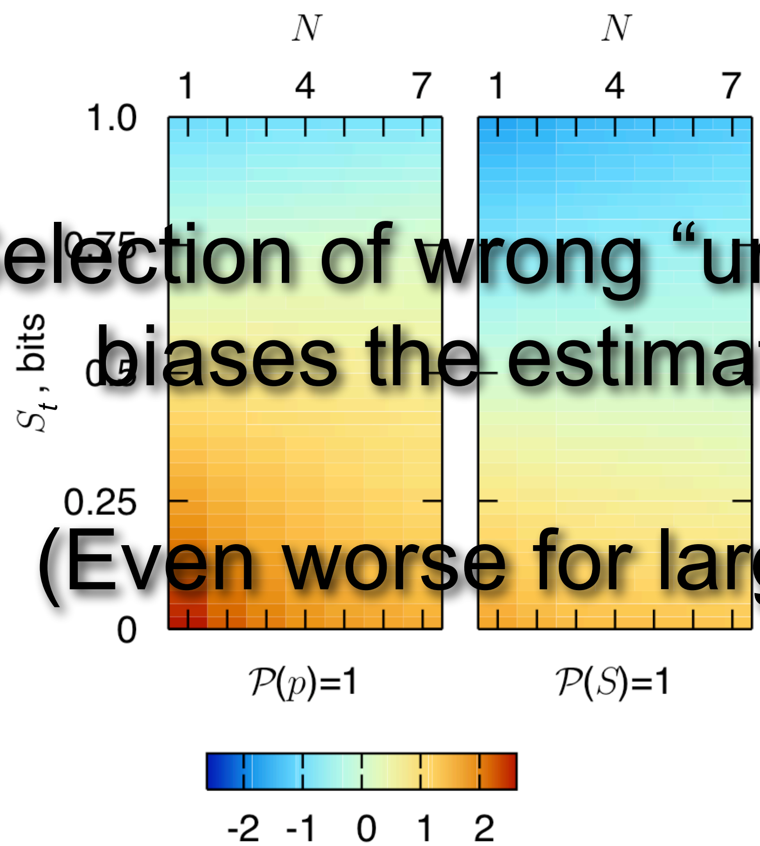


Assume (Bayes)

uniform (no assumptions)

p S

What is unknown?



$$\varepsilon = \left\langle \frac{S_{est} - S_{true}}{\delta S_{est}} \right\rangle$$

Selection of wrong "unknown" biases the estimation.

(Even worse for large *K*.)

For large *K*

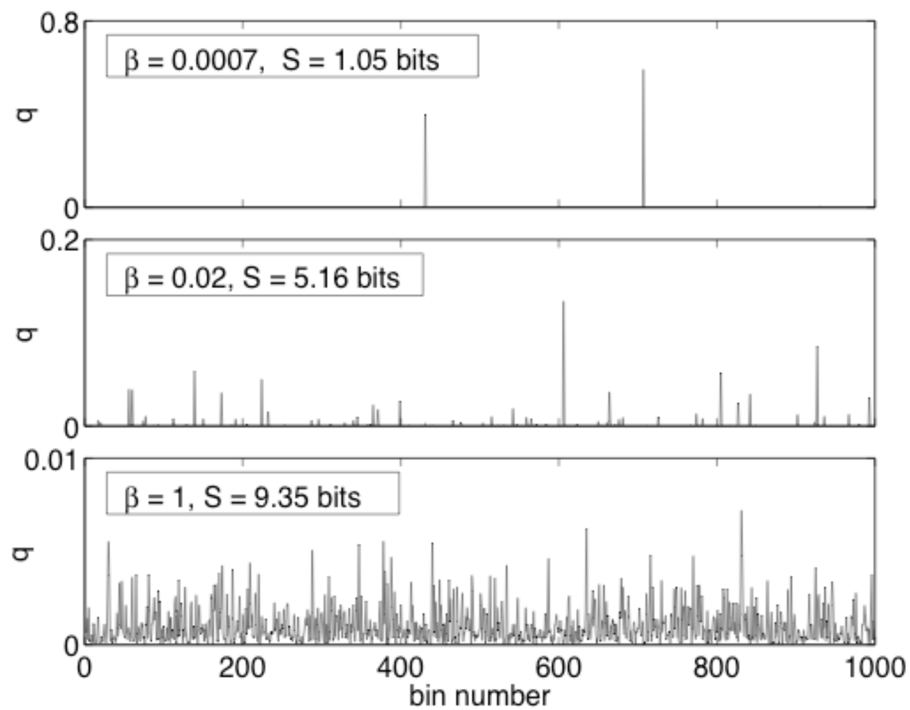
- The problem is more severe.
- Uniformize on *S* (approximately).
- Will work for a certain type of distributions only.

For large K the problem is extreme (S known a priori)

$$P_\beta(\{q_i\}) = \frac{1}{Z(\beta)}\,\delta\left(1 - \sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1}$$

Dirichlet priors, a.k.a., adding pseudocounts (include the uniform prior, the ML prior, and others).

Inference is analytic

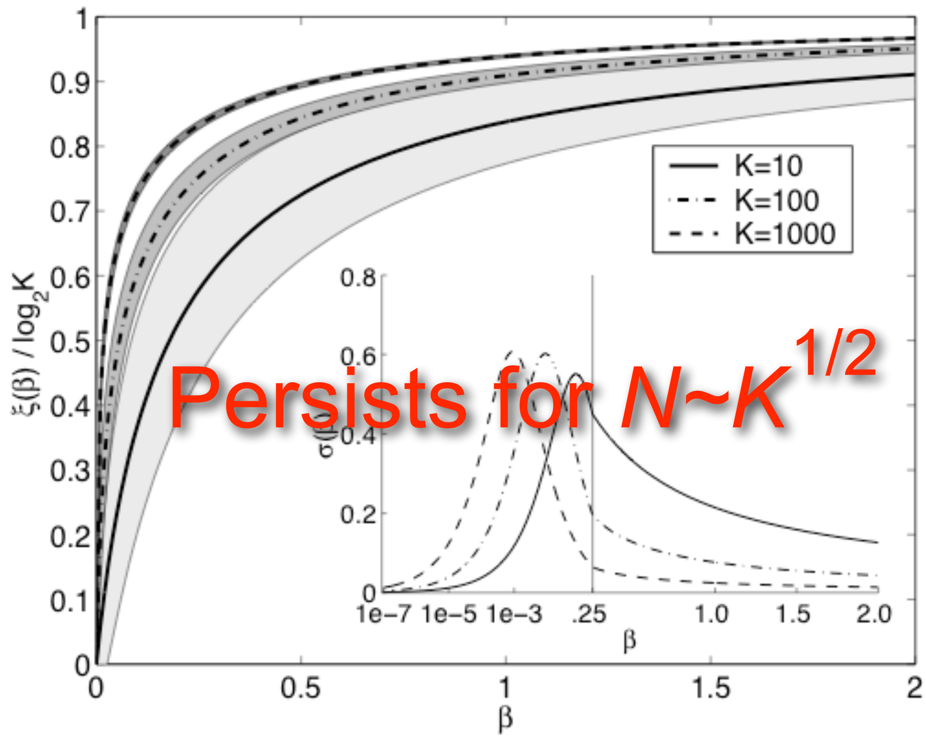


For large K the problem is extreme (S known a priori)

$$\xi(\beta) = \langle S_\beta(0) \rangle = \psi_0(K\beta + 1) - \psi_0(\beta + 1)$$

$$\sigma^2(\beta) = \langle \delta S_\beta^2(0) \rangle = \frac{\beta + 1}{K\beta + 1}\psi_1(\beta + 1) - \psi_1(K\beta + 1)$$

But a priori entropy distribution is narrow; need $N>K$ to overcome the bias.

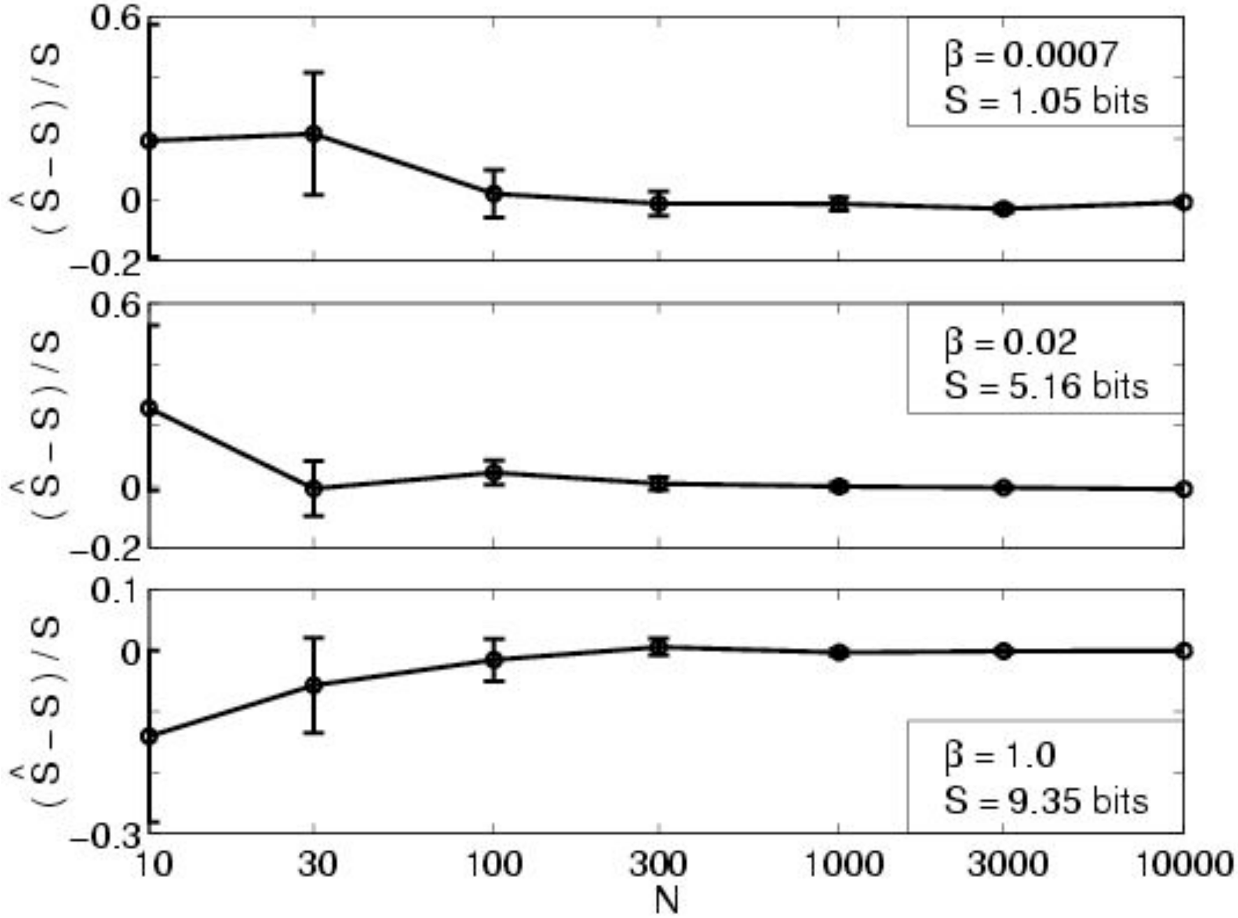


Uniformize on *S*

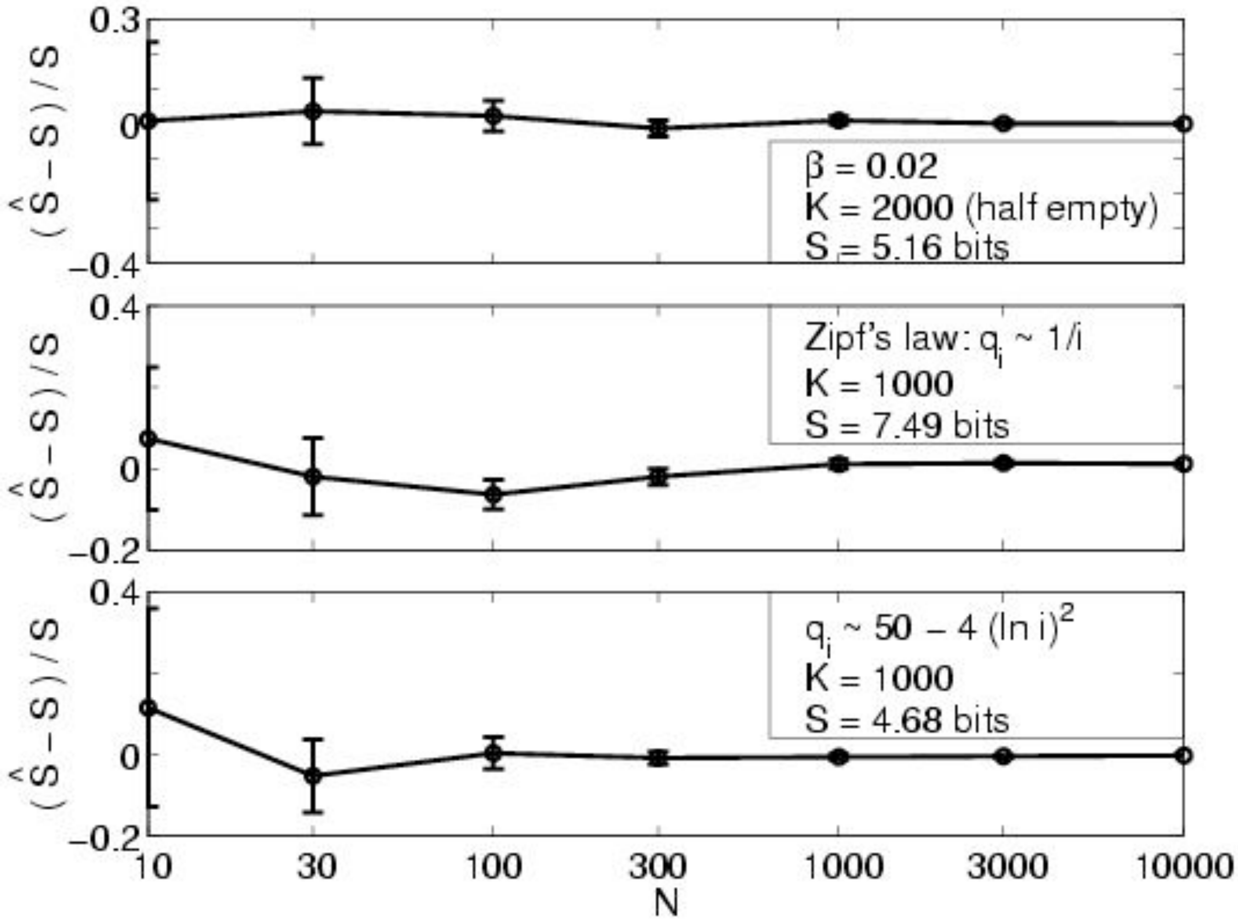
$$P_\beta(\{q_i\},\beta) = \frac{1}{Z}\,\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^\beta\,\left.\frac{dS}{d\beta}\right|_{N=0} P(S|_{N=0})$$

- A delta-function sliding along the a priori entropy expectation.
- This is also Bayesian model selection (small β large phase space)
- Have error bars (dominated by posterior variance in β, not at fixed β).

Typical cases (correct prior)



Atypical cases (incorrect prior)



For NSB solution

- Posterior variance scales as $(N - K_1)/K$
- Little bias, except for distribution with long rank-order tails.
- Counts coincidences and works in Ma regime (if works, see above).
- Is consistent.
- Allows infinite *K*

$$\hat{S} = \left(C_\gamma - \ln 2\right) + 2\ln N - \psi_0\left(\frac{N - K_1}{N}\right) + O(1/K, 1/N)$$

$$\delta\hat{S}^2 = \psi_1\left(\frac{N - K_1}{N}\right) + O(1/K, 1/N)$$

(Nemenman et al. 2002, Nemenman 2003)

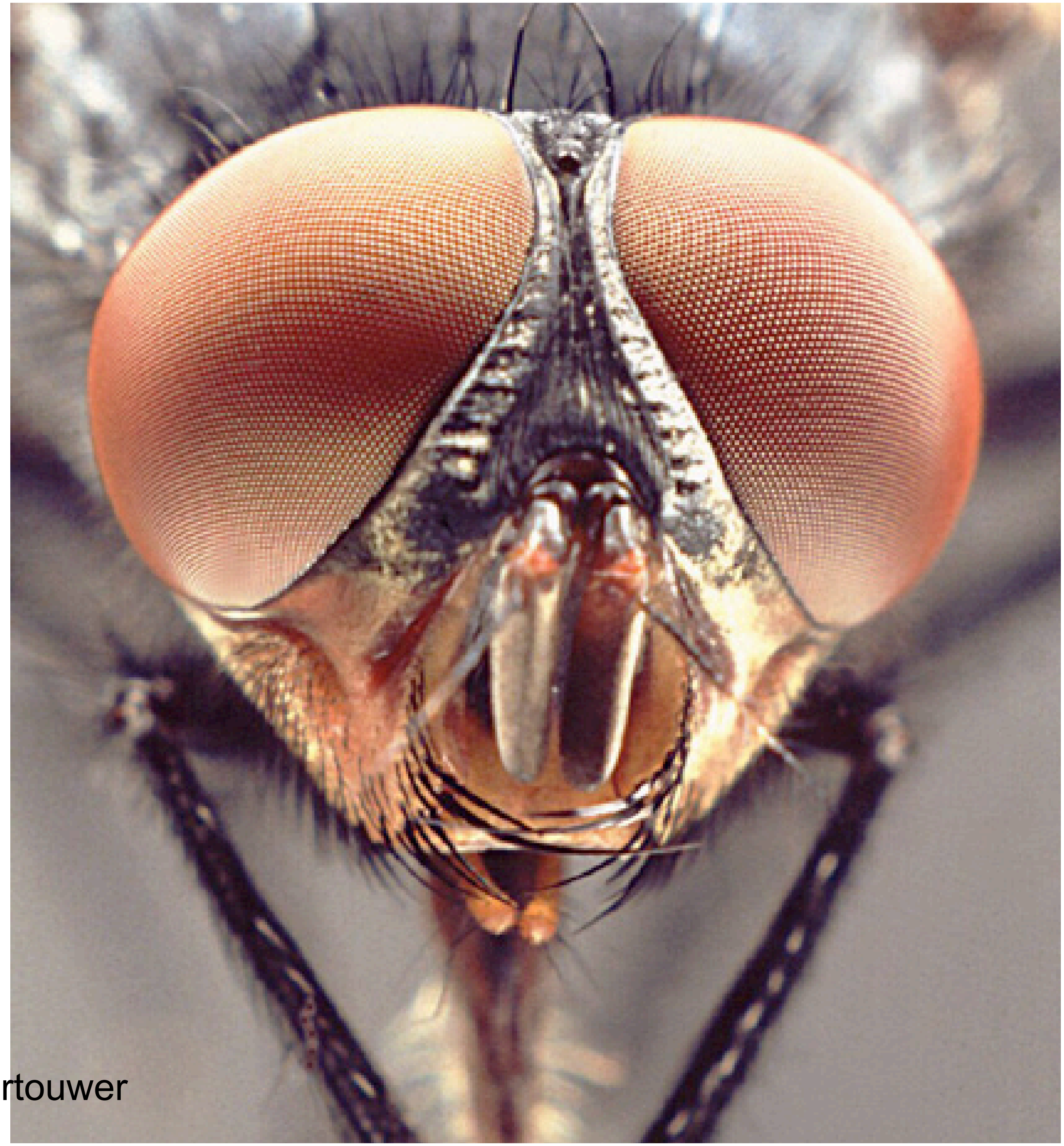
General principle?

Priors uniforms on quantities of interest

Software implementation

…and many other details:

http://nsb-entropy.sf.net

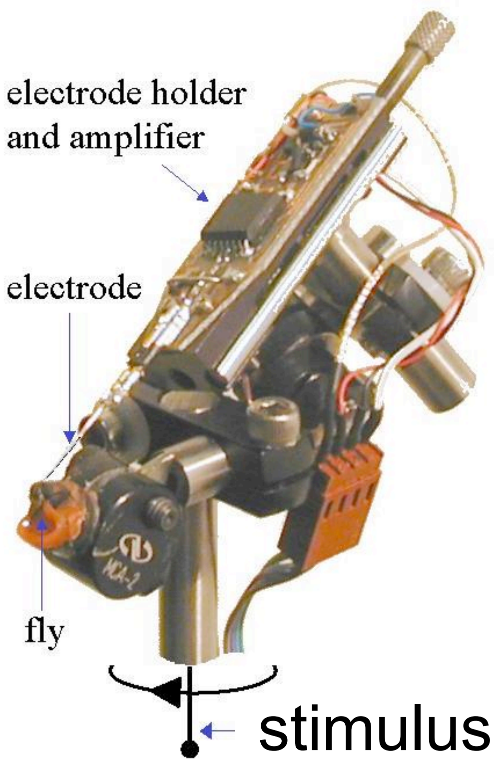


H. L. Leertouwer

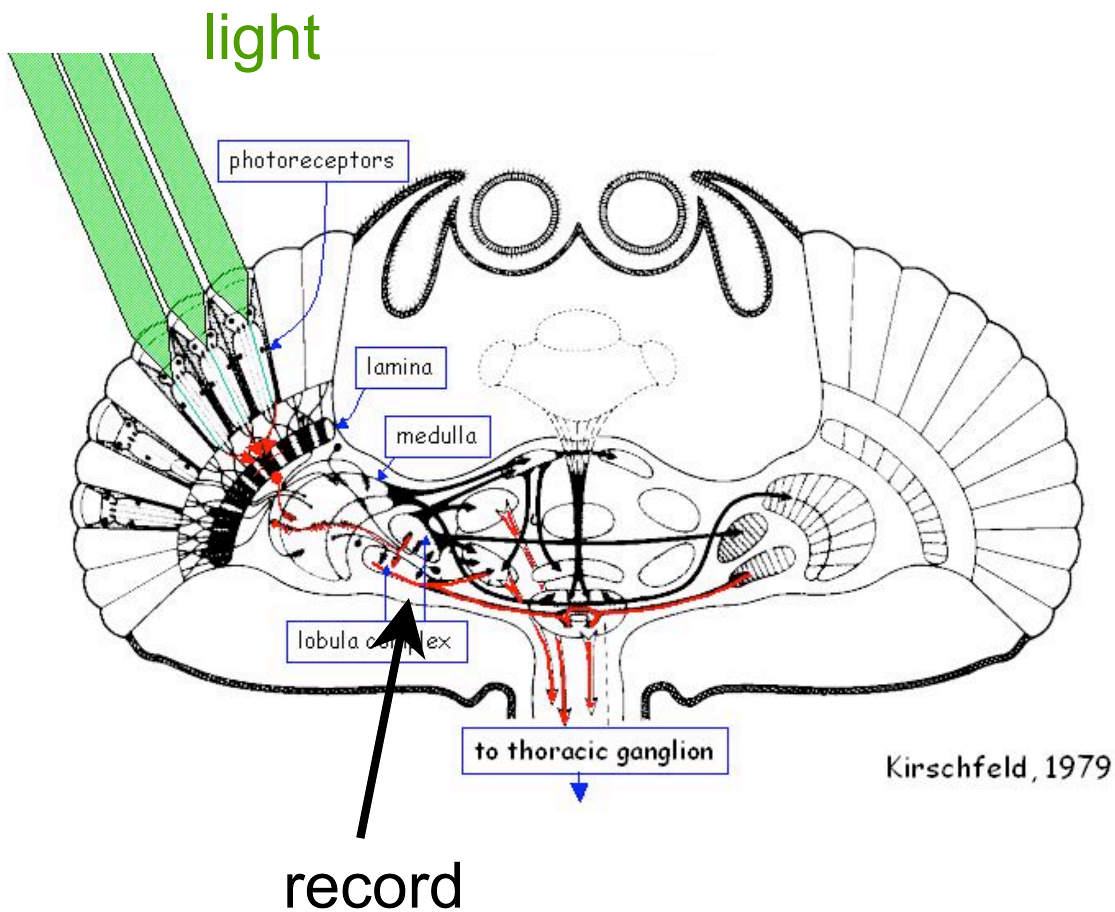
Questions

- Can we understand the code?
- Which features of it are important?
  - Rate of precise timing (how precise)?
  - Synergy between spikes?
- What/how much does the fly know?
- Is there an evidence for optimality?

Recording from fly's H1

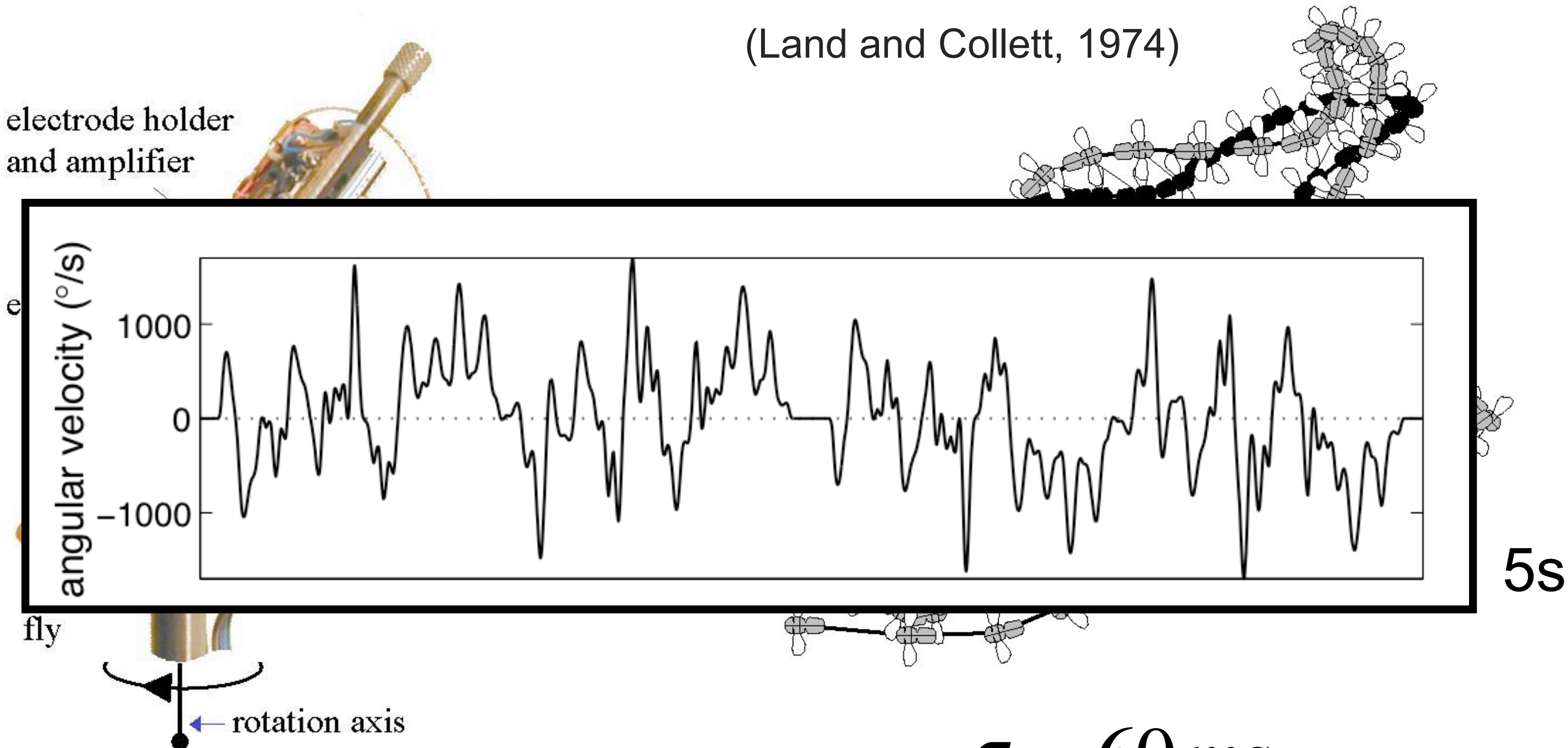


(Lewen et al, 2001)



Kirschfeld, 1979

Natural stimuli



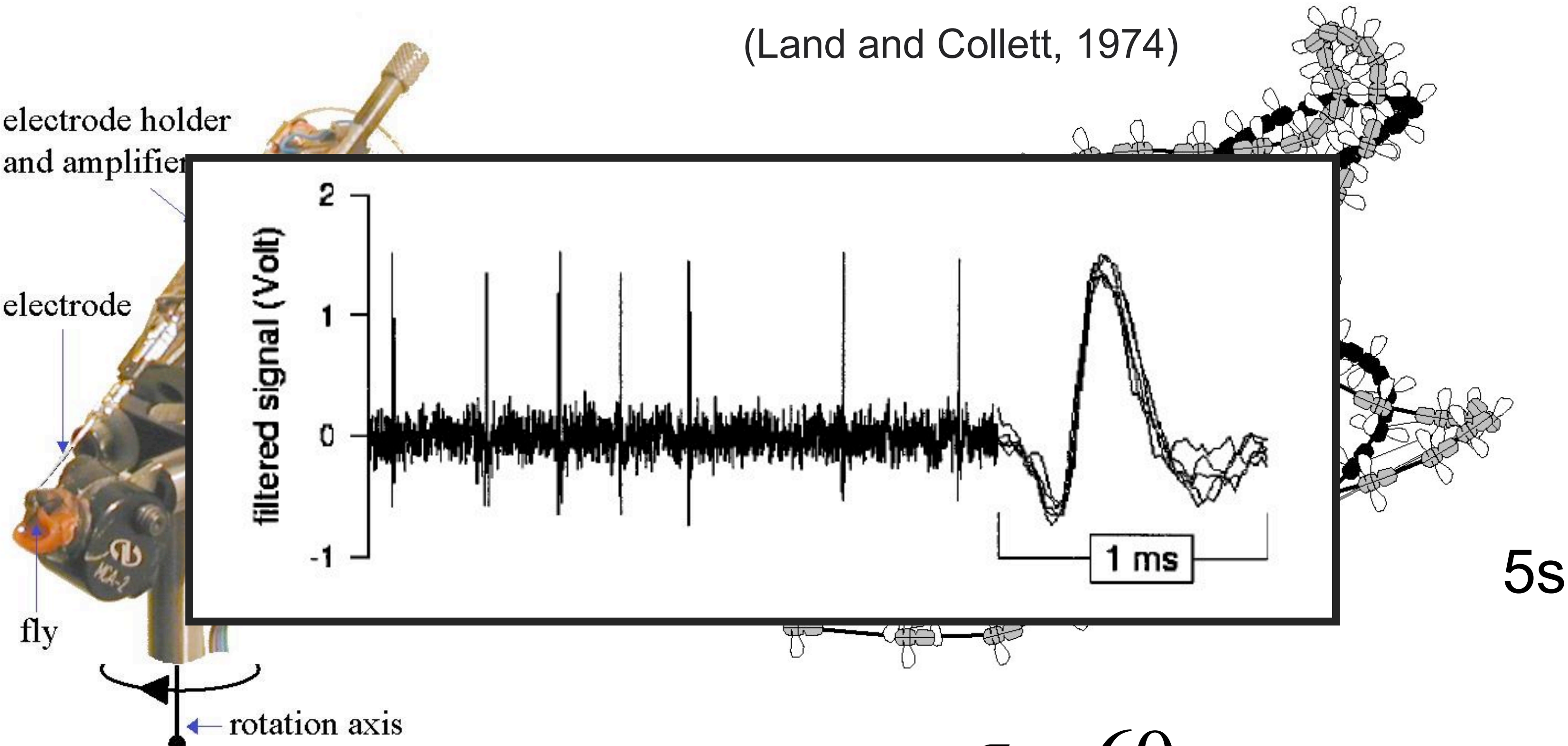
(Lewen et al, 2001)

$$\tau = 60ms$$

$$\text{response} = 30ms$$

Natural stimuli

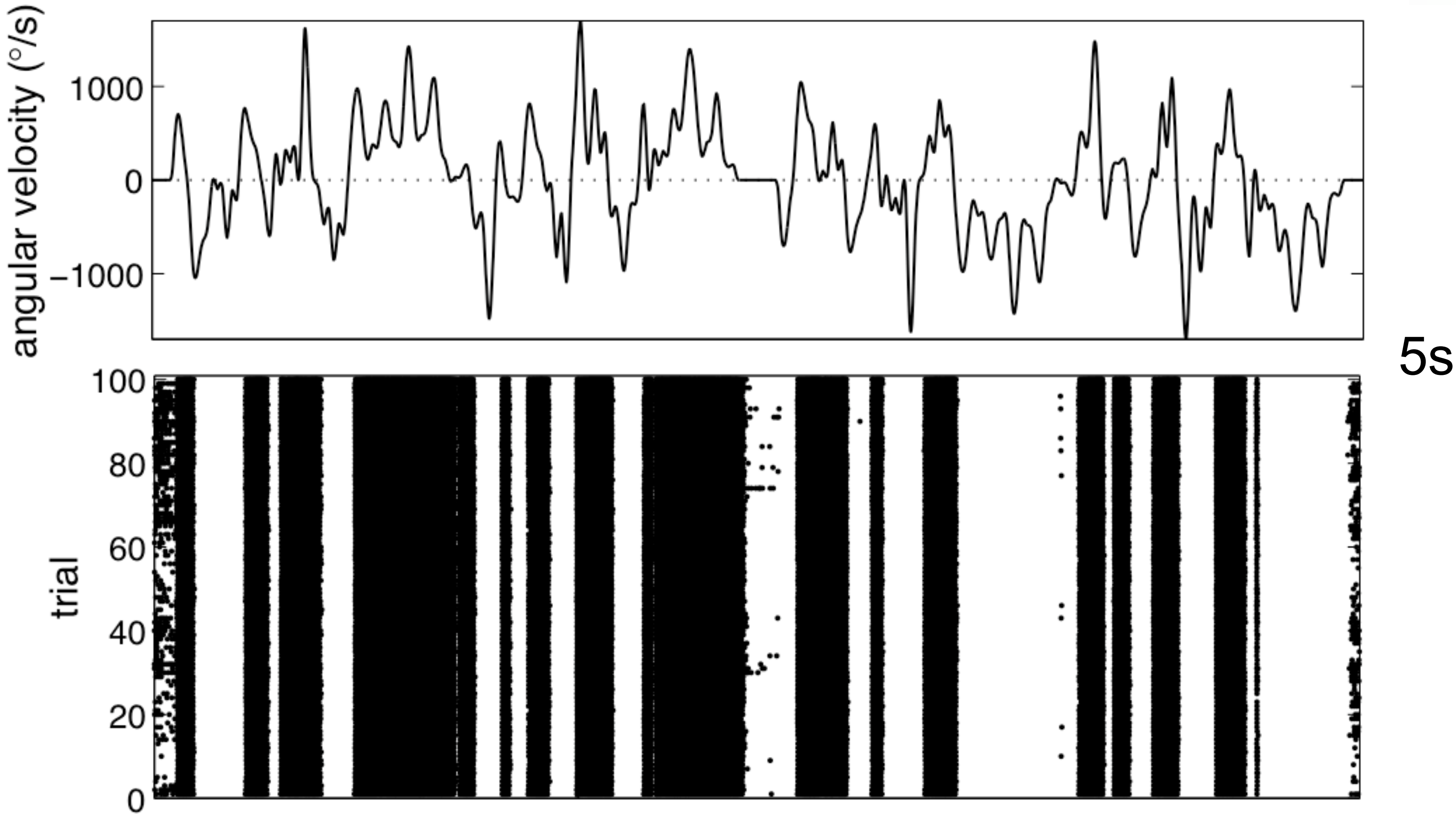
(Land and Collett, 1974)



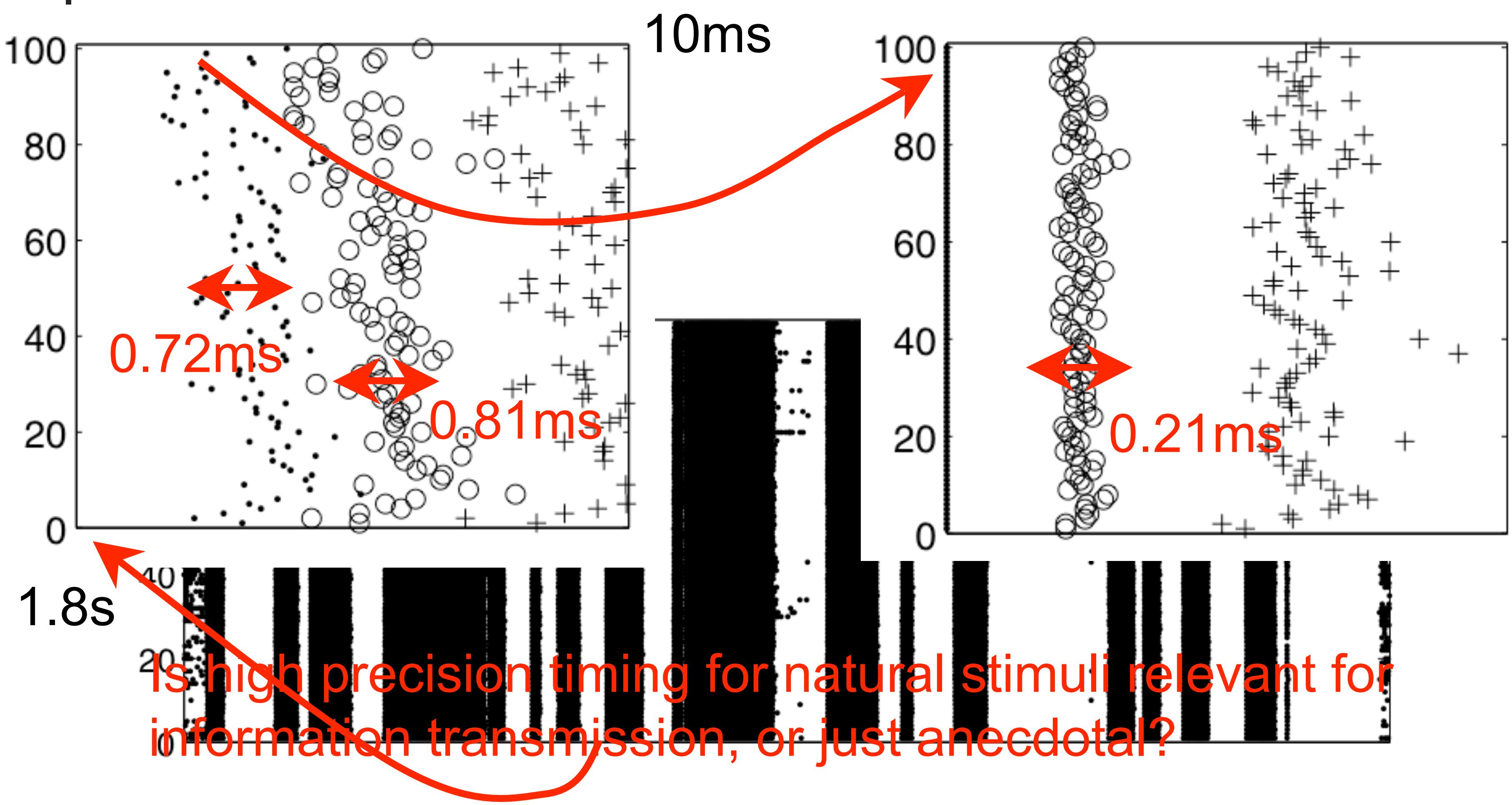
(Lewen et al, 2001)

$$\tau = 60ms$$
$$\text{response} = 30ms$$

Natural stimulus and response

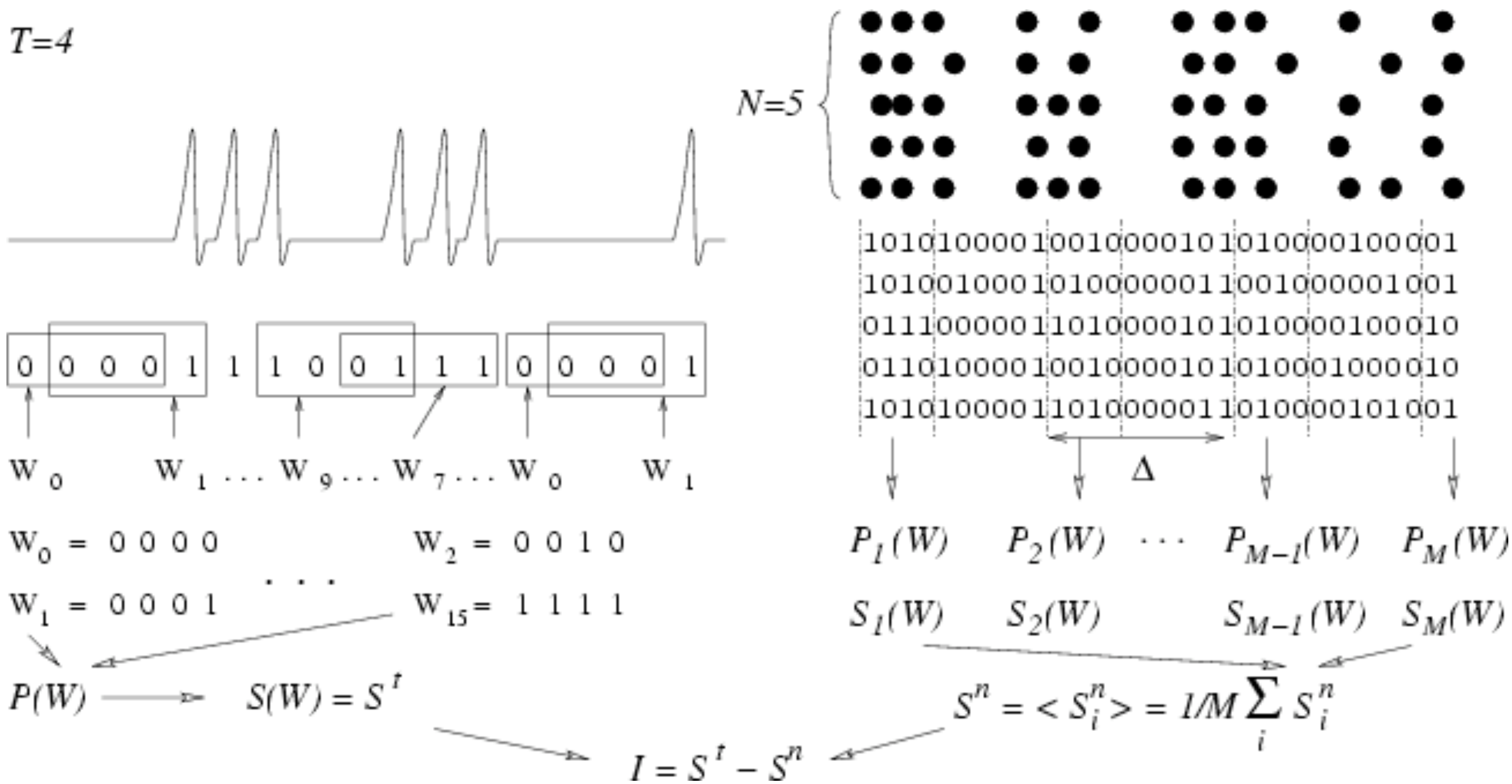


Highly repeatable spikes (not rate coding)



Is high precision timing for natural stimuli relevant for information transmission, or just anecdotal?

Experiment design



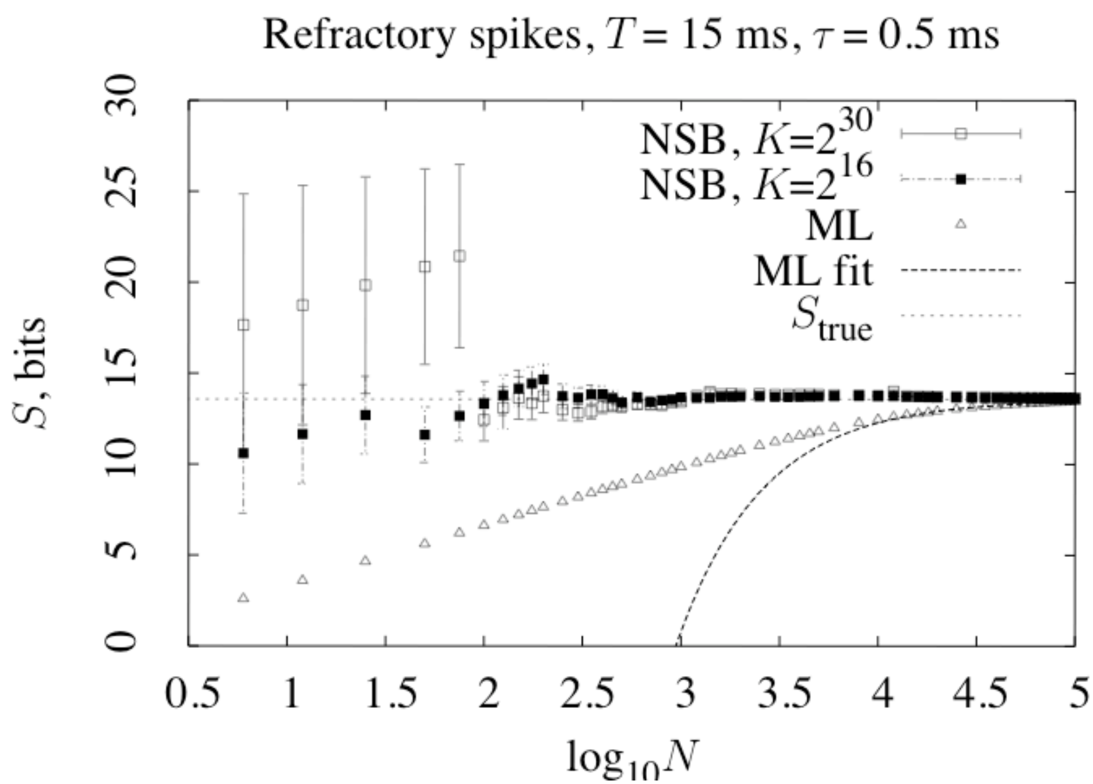
(Strong et al., 1998)

Problems

- Total of about 10-15 min of recordings (limited by stationarity of the outside world)
- At most 200 repetitions
- Stimulus correlation of 60ms: only 10000 independent samples (repeated or nonrepeated)
- Need to sample words of length 30 ms (behavioral) to 60 ms (stimulus) at resolution down to 0.2 ms (binary words of length up to ~100).

Synthetic test of NSB

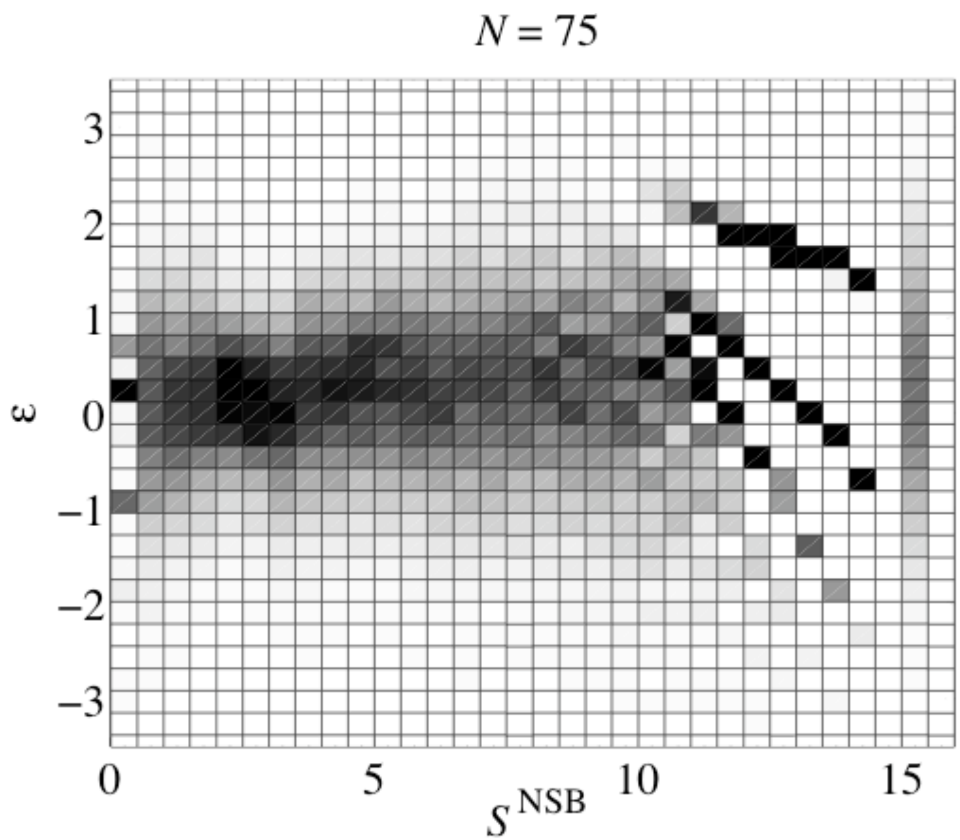
Refractory Poisson, rate 0.26 spikes/ms, refractory period 1.8 ms, T=15ms, discretization 0.5ms, true entropy 13.57 bits.



- Estimator is unbiased if consistent and self-consistent.
- Always do this check.

(Nemenman et al. 2004)

Natural data (all *S*)



$$\varepsilon = \frac{S^{NSB}(N) - S}{\delta S^{NSB}(N)}$$

$$\approx \frac{S^{NSB}(N) - S(N = \max)}{\delta S^{NSB}(N)}$$

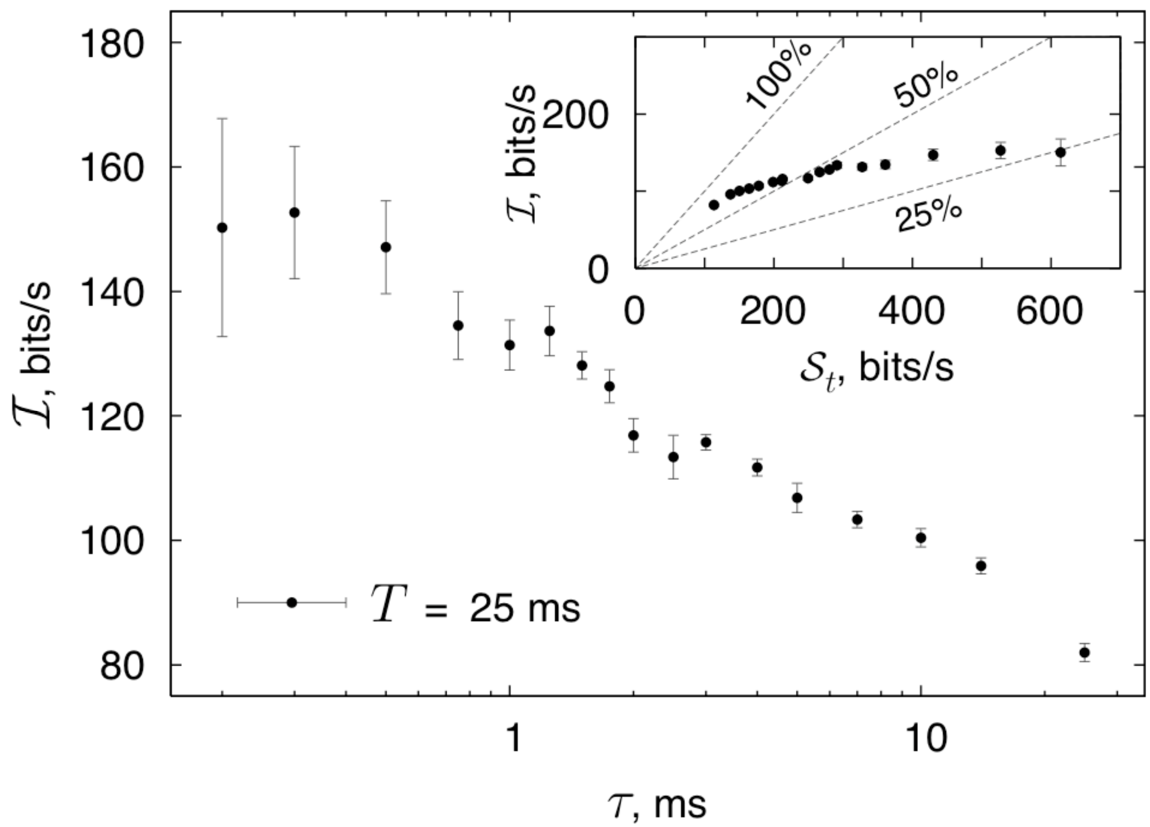
Max=196 repeats

(Nemenman et al. 2004)

Neural code: What remains hidden?

- Given entropy of slices, find the mean noise entropy with error bars (slice entropies are correlated and bimodal).
- Samples for total entropy are also correlated and have long tailed Zipf plots.
- For very fine discretizations and T~30ms need extrapolation.

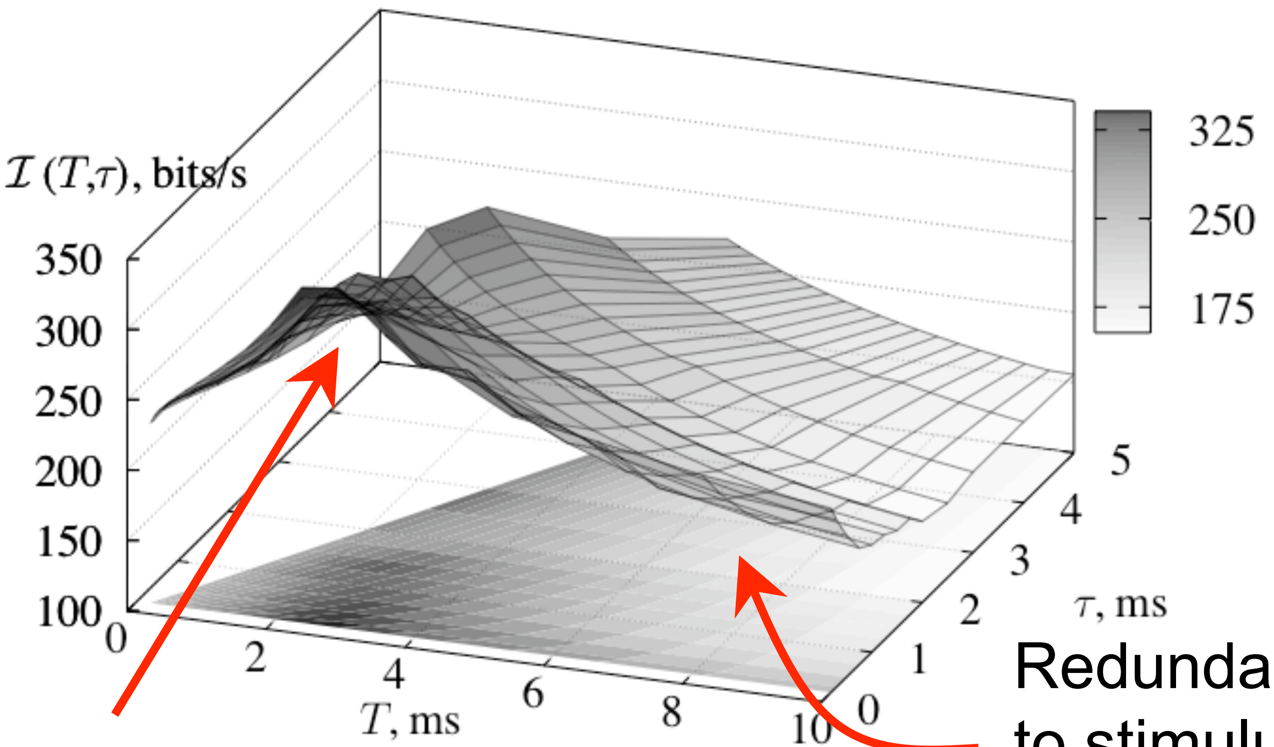
Information rate at *T*=25ms



0.2 ms -- comparable to channel opening/ closing noise and experimental noise.

- Information present up to τ =0.3 ms
- 30% more information at τ<1ms. Encoding by refractoriness?
- ~1 bit/spike at 150 spikes/s and low-entropy correlated stimulus. Design principle?
- Efficiency >50% for τ >1ms, and ~75% at 25ms. Optimized for natural statistics?

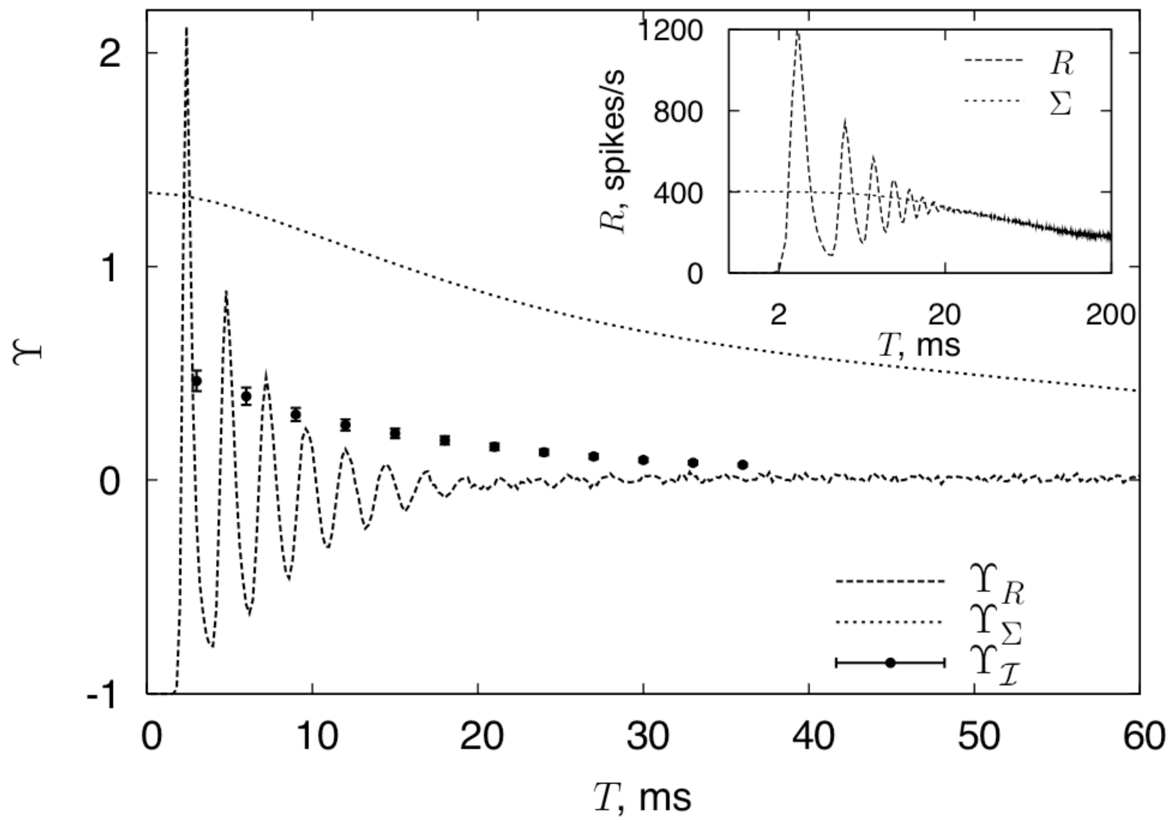
Synergy from spike combinations



Spike pairs

Redundancy due to stimulus

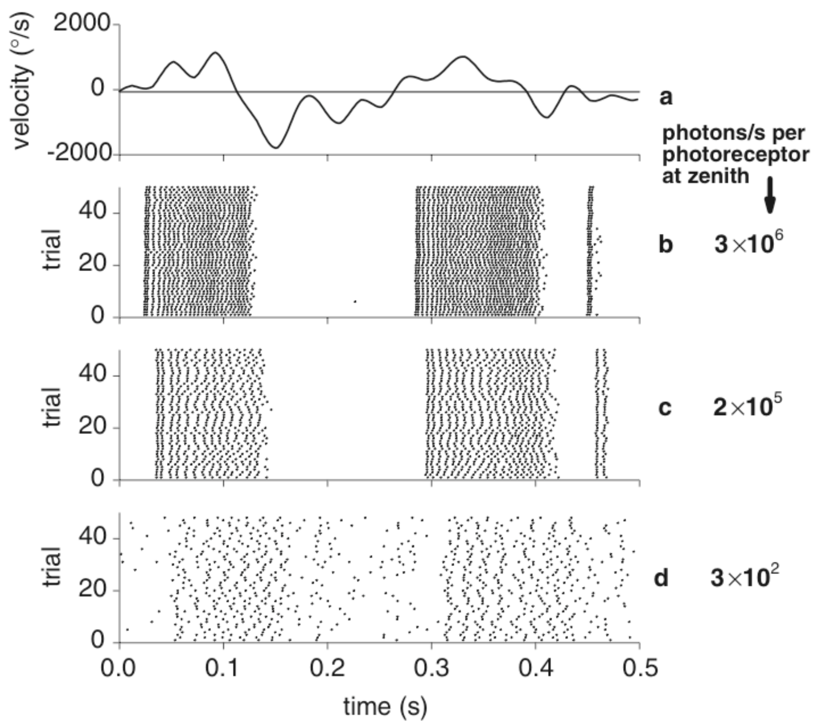
New bits (optimized code)



- Spikes are very regular (>10 beats) WKB decoder? Interspike potential?
- CF at half its value, but fly gets new bits every 25 ms
- Independent info (even though entropies are *T* dependent).

Behaviorally optimized code!

Precision is limited by physical noise sources



(Lewen, et al 2001)

$$T = 6 \text{ ms}$$

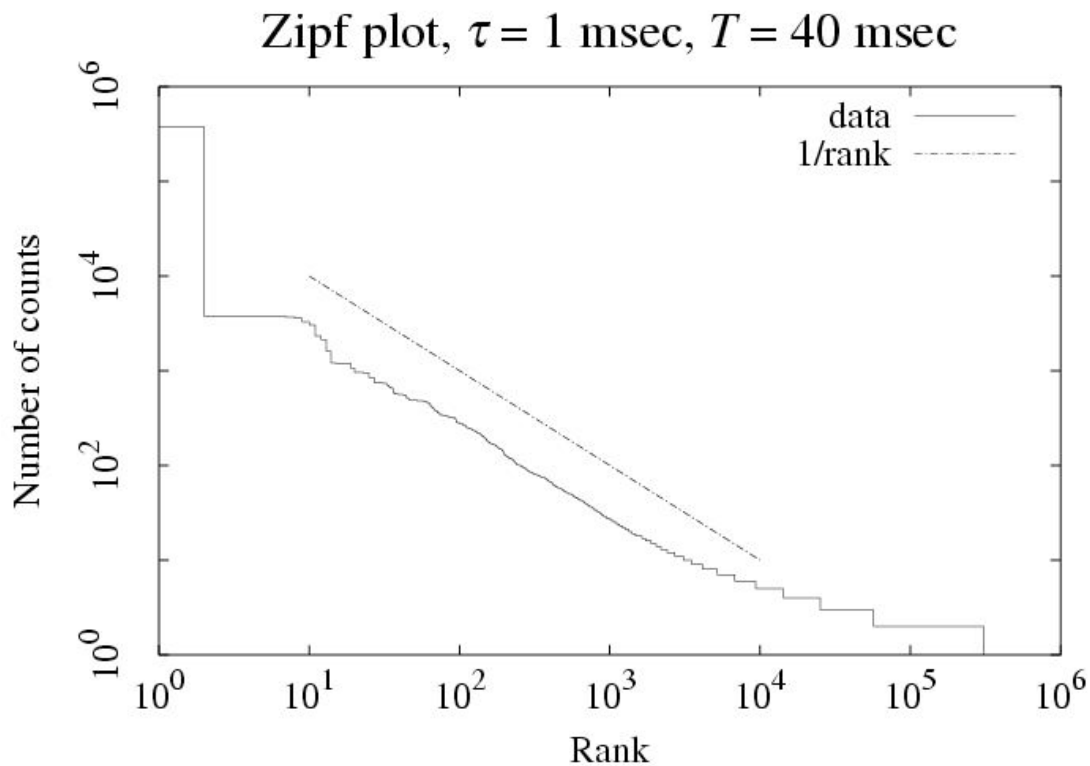
$$\tau = 0.2 \text{ ms}$$

$$1.49 \text{ vs. } 1.61 \cdot 10^6 \text{ph/(s} \cdot \text{rec)}$$

$$I^+ - I^- = 0.020 \pm 0.011 \text{ bits}$$

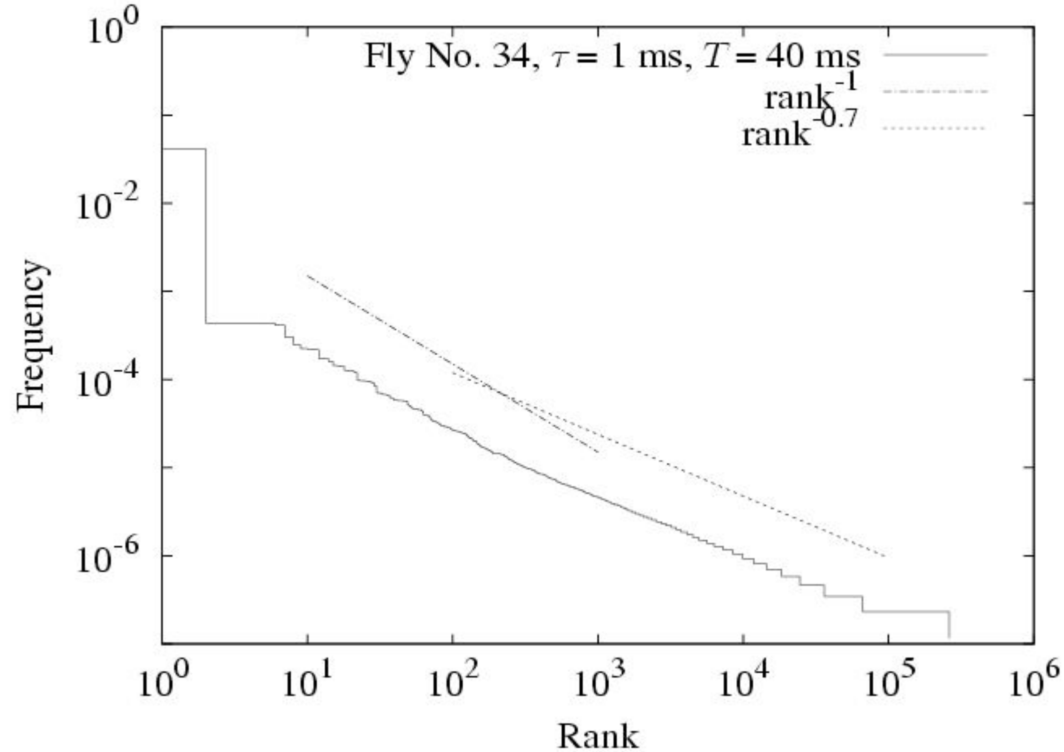
A very intelligent fly

- One often considers a 1/f rank-order plot as a sign of intelligence.
- But…



A very intelligent fly

- One often considers a 1/f rank-order plot as a sign of intelligence.
- But…



Zipf law may be a result of complexity of the world, not the language.