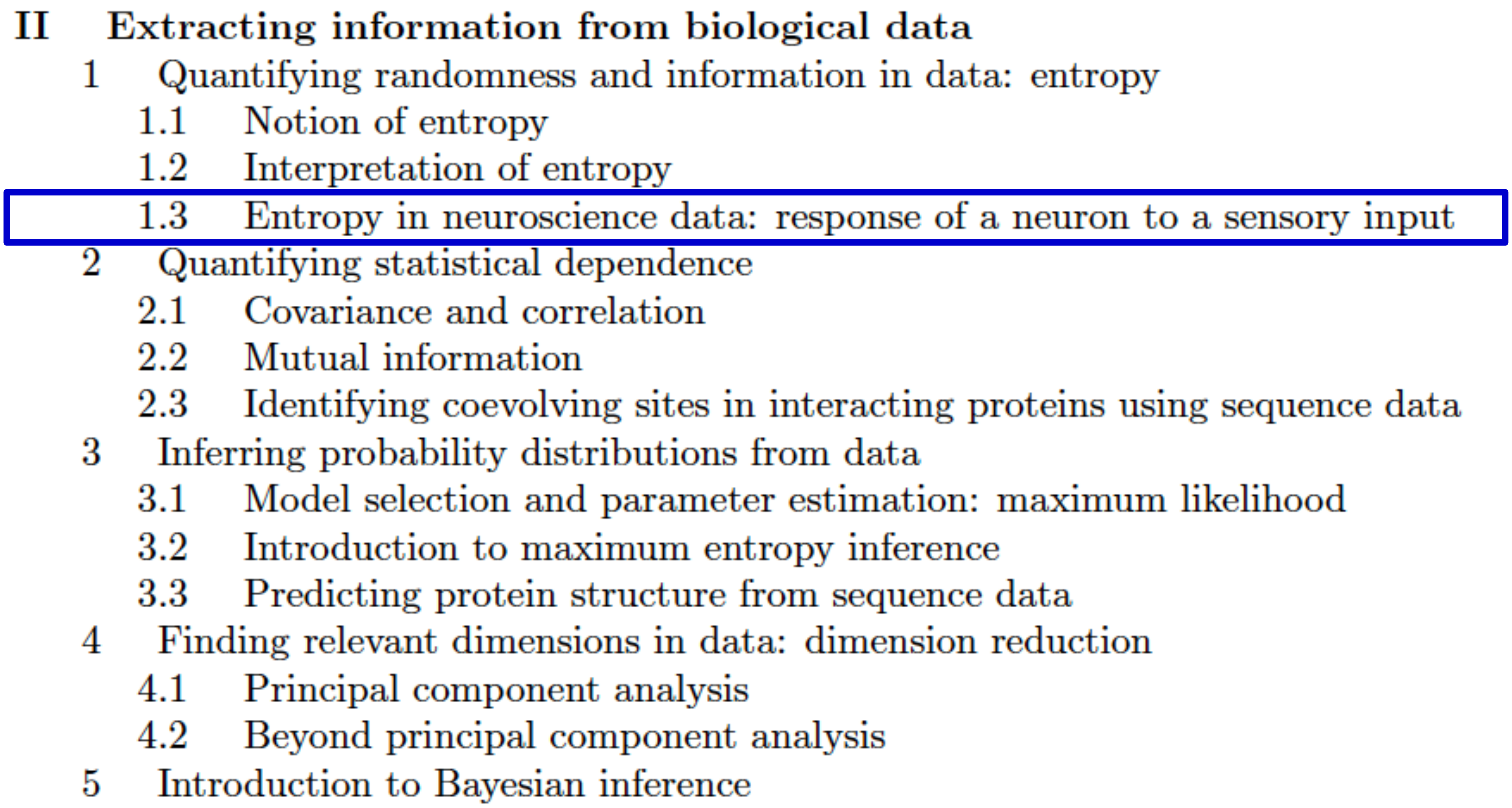
Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol

Lecture 7

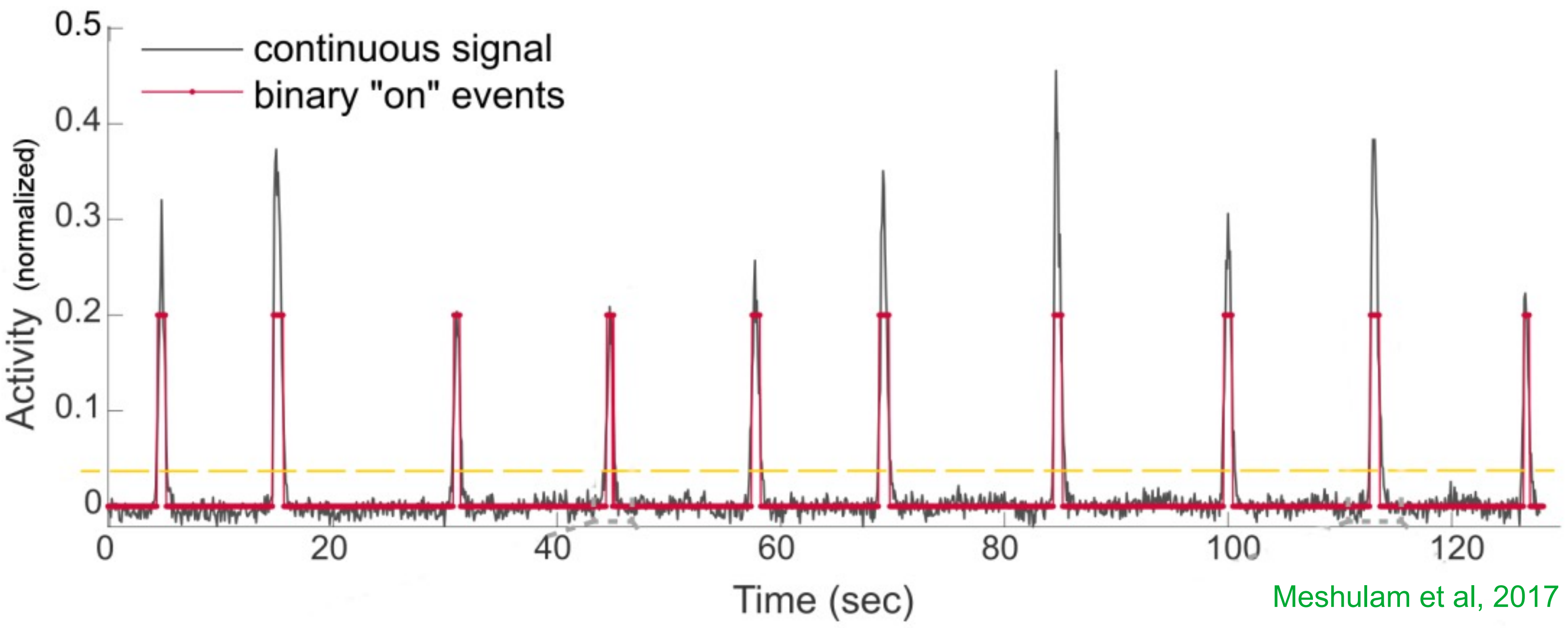
Outline of the course



II Extracting information from biological data

- 1 Quantifying randomness and information in data: entropy
  - 1.1 Notion of entropy
  - 1.2 Interpretation of entropy
  - 1.3 Entropy in neuroscience data: response of a neuron to a sensory input
- 2 Quantifying statistical dependence
  - 2.1 Covariance and correlation
  - 2.2 Mutual information
  - 2.3 Identifying coevolving sites in interacting proteins using sequence data
- 3 Inferring probability distributions from data
  - 3.1 Model selection and parameter estimation: maximum likelihood
  - 3.2 Introduction to maximum entropy inference
  - 3.3 Predicting protein structure from sequence data
- 4 Finding relevant dimensions in data: dimension reduction
  - 4.1 Principal component analysis
  - 4.2 Beyond principal component analysis
- 5 Introduction to Bayesian inference

Neuroscience data

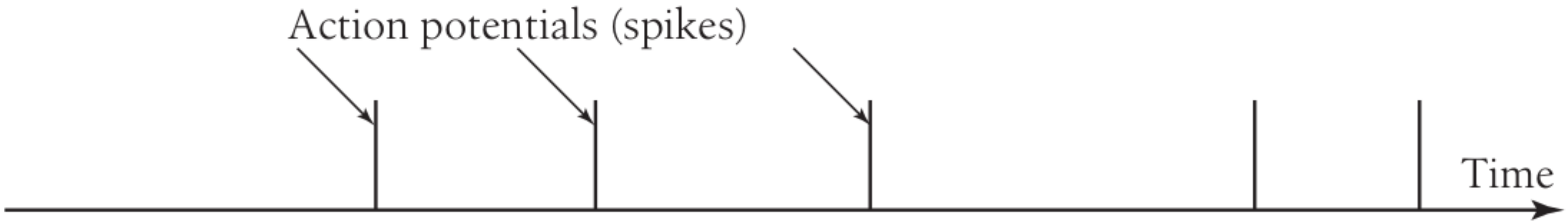


Meshulam et al, 2017

Raw signal from a neuron in units of normalized fluorescence
The series of neuron action potentials is converted to binary on/off activity levels

How does this signal convey information from the stimulus seen by the animal?

From neuroscience data to binary words



- Duration of a spike ~ 0.5 ms
- Minimum time between spikes ~ 3 ms
- Construct bins of duration $\Delta\tau$ in which you count the number of spikes

→ How should we choose the time $\Delta\tau$ in order to get binary (0 or 1) counts?

How should we choose the time $\Delta\tau$ in order to get binary (0 or 1) counts?

0% A. Smaller than 0.5 ms

0% B. Larger than 0.5 ms

0% C. Smaller than 3 ms

0% D. Larger than 3 ms

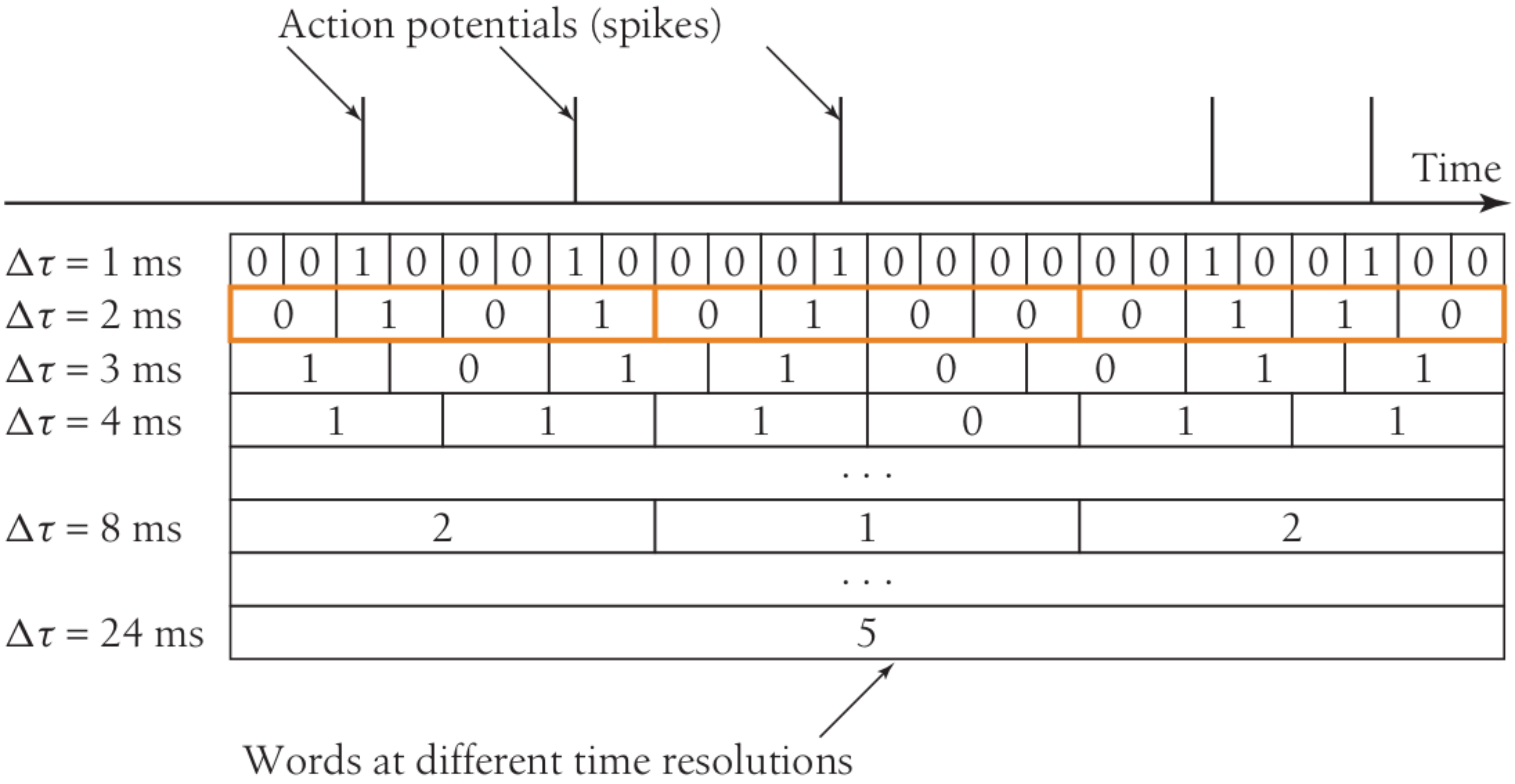
0% E. Between 0.5 ms and 3 ms

- Duration of a spike ~ 0.5 ms
- Minimum time between spikes ~3 ms
- Construct bins of duration $\Delta\tau$ in which you count the number of spikes

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

From neuroscience data to binary words



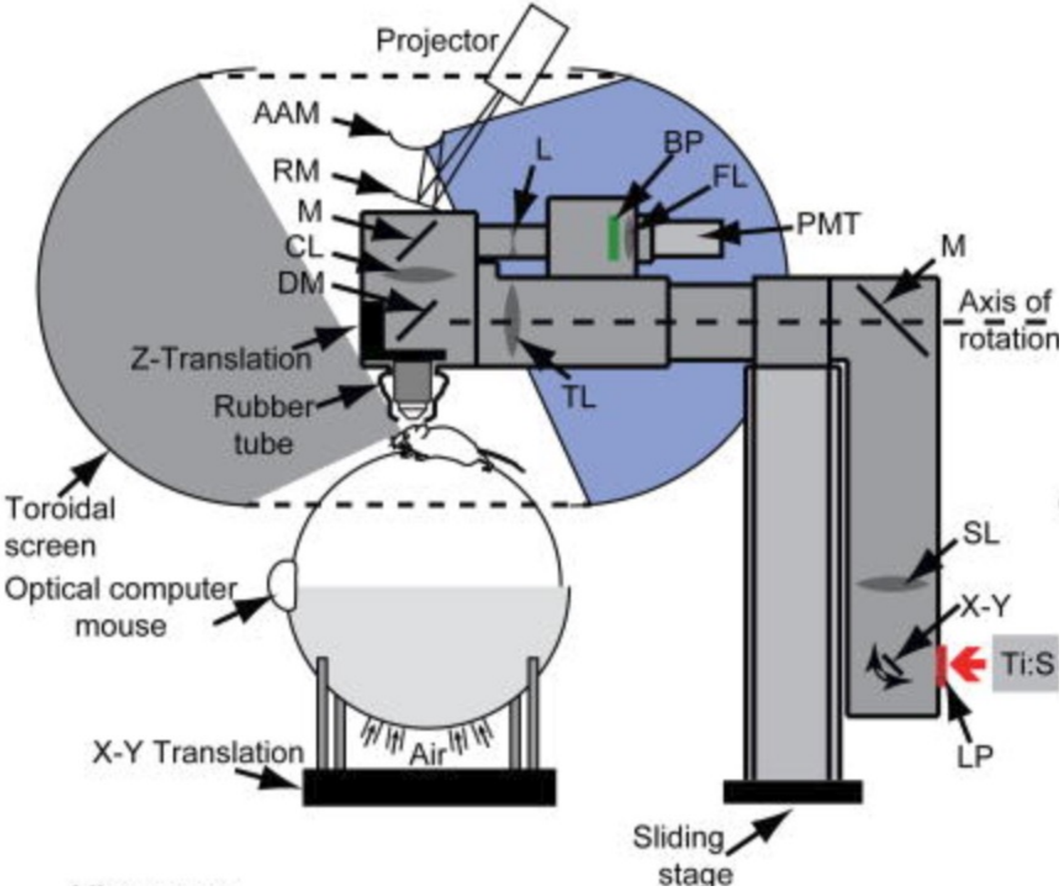
Action potentials (spikes)
Time
Words at different time resolutions

$\Delta\tau = 1$ ms	0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0
$\Delta\tau = 2$ ms	0 1 0 1 0 1 0 0 0 1 1 0
$\Delta\tau = 3$ ms	1 0 1 1 0 0 1 1
$\Delta\tau = 4$ ms	1 1 1 0 1 1
	. . .
$\Delta\tau = 8$ ms	2 1 2
	. . .
$\Delta\tau = 24$ ms	5

Minimum time between spikes ~3 ms → for sufficiently small $\Delta\tau$, the words are binary

Orange: $\Delta\tau$ = 2 ms and τ = 8 ms (word duration) → 3 successive 4-letter words in 24 ms, namely 0101, 0100, 0110

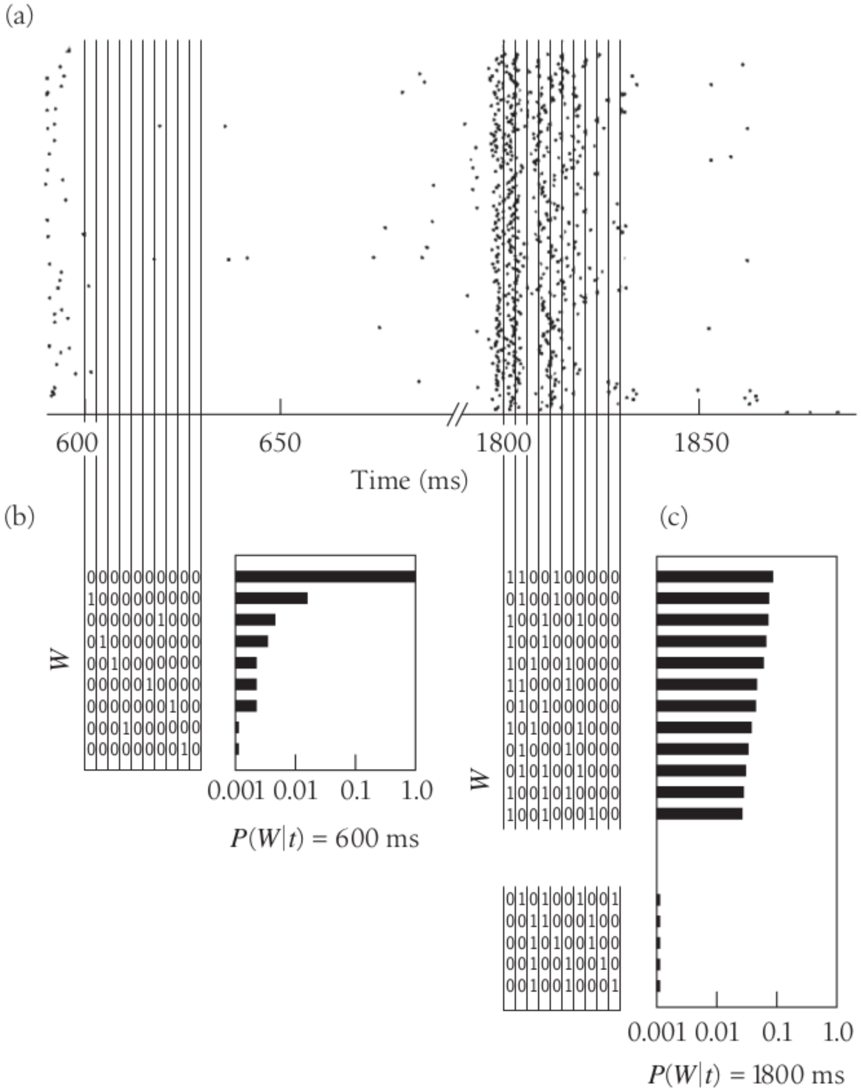
Neuroscience data



Dombeck et al, 2010

Experimental setup: spherical treadmill + virtual reality apparatus (projector, RM: reflecting mirror, AAM: angular amplification mirror, toroidal screen and a optical computer mouse to record ball rotation)+ custom two-photon microscope for *in vivo* microscopy

Information in neuroscience data



How do sequences of spikes represent the sensory world?

Study of the motion-sensitive neuron H1 in the fly's visual system while a fly is shown the same movie several times (a pattern of random bars that moves across the visual field at variable velocity)

"Words" of 10 characters correspond to τ = 30 ms (~ behavior reaction time), with each binary character corresponding to $\Delta\tau$ = 3 ms

The distribution of words that occur at a particular moment in the movie, P(W|t), is shown for t = 600 ms and t = 1800 ms

de Ruyter et al, 1997

What is the number of different possible sequences of 10 binary digits?

A. 10

B. 20

C. 100

D. 1024

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

What is the maximum possible entropy of an ensemble of sequences of 10 binary digits?

A. $\log_2(10)$ bits
B. 10 bits
C. 2 bits
D. $\log_2(10)/10$ bits
E. 1024 bits

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Do you expect the entropy of the signal knowing the time at which we observe it, $H(S|t)$, to be:

A. Larger than $H(S)$
B. Smaller than $H(S)$
C. It depends on t

To answer, please:

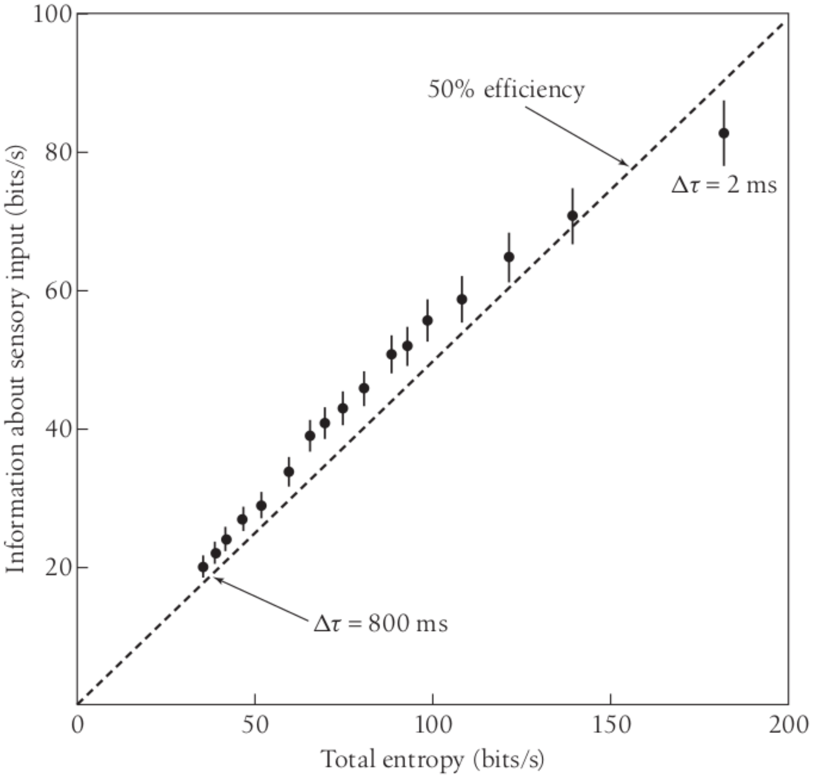
- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Information in neuroscience data

- **Impact of word duration τ and of time resolution Δτ:**

If correlations in the spike train have finite range, then for long τ, $S(\tau, \Delta\tau) \propto \tau$

In the long-τ limit, varying Δτ and reporting $\lim_{\tau \to \infty} \dfrac{S(\tau, \Delta\tau)}{\tau}$ yields:



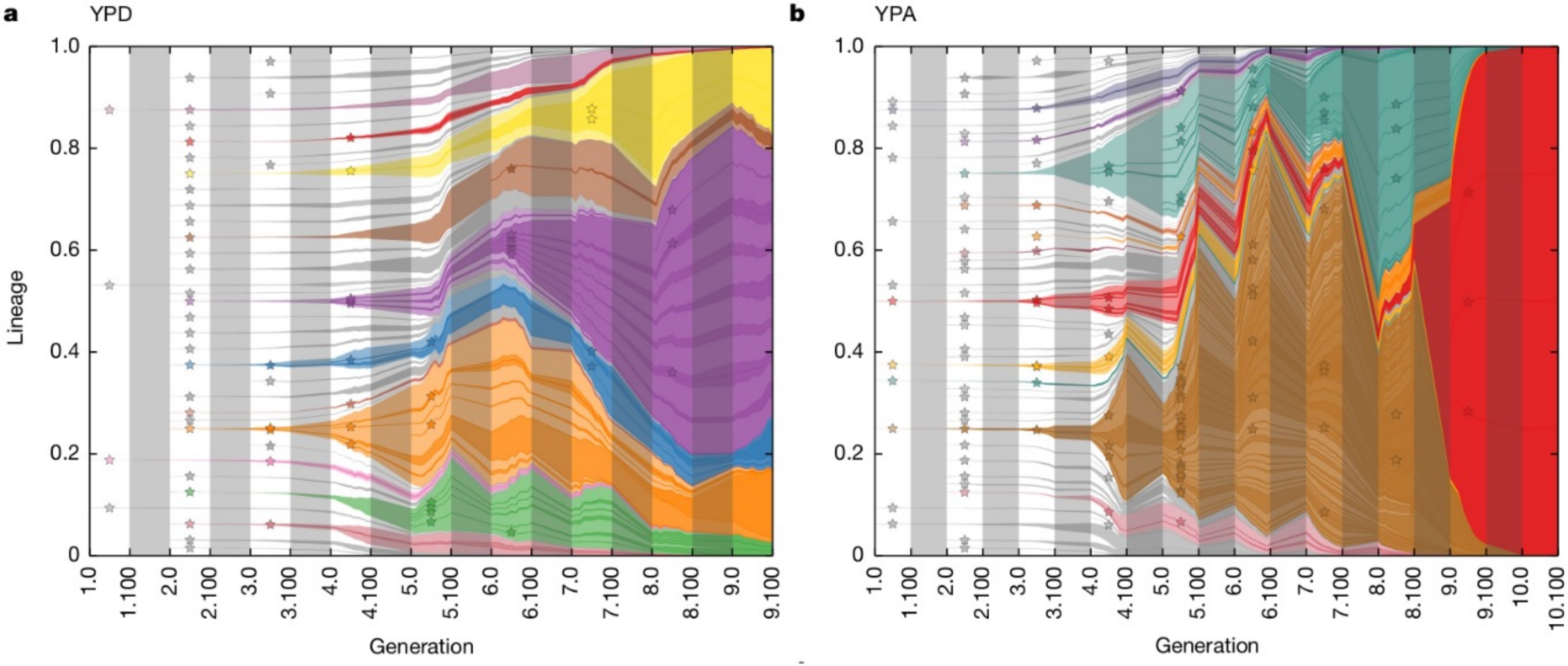
As the time resolution is varied from Δτ = 800 ms down to Δτ = 2 ms, finer details of the neural response are distinguished
→ increasing entropy

Across this range, almost constant efficiency

Strong et al, 1998

Entropy in ecology and evolution

- **Diversity of a yeast population**



Nguyen Ba et al, 2019

Stacked frequencies of barcoded lineages in a population of yeast versus time. New barcodes are added in the gray phases.

(a): Yeast in a rich medium (YPD). (b): Same rich medium + added acetic acid (YPA).

How would the observed diversity evolve in a population if there was no rebarcoding?

How would the observed diversity (i.e. the number of different barcodes) evolve in a constant-size population if there was no rebarcoding?

A. It would decrease

B. It would increase

C. It would stay constant

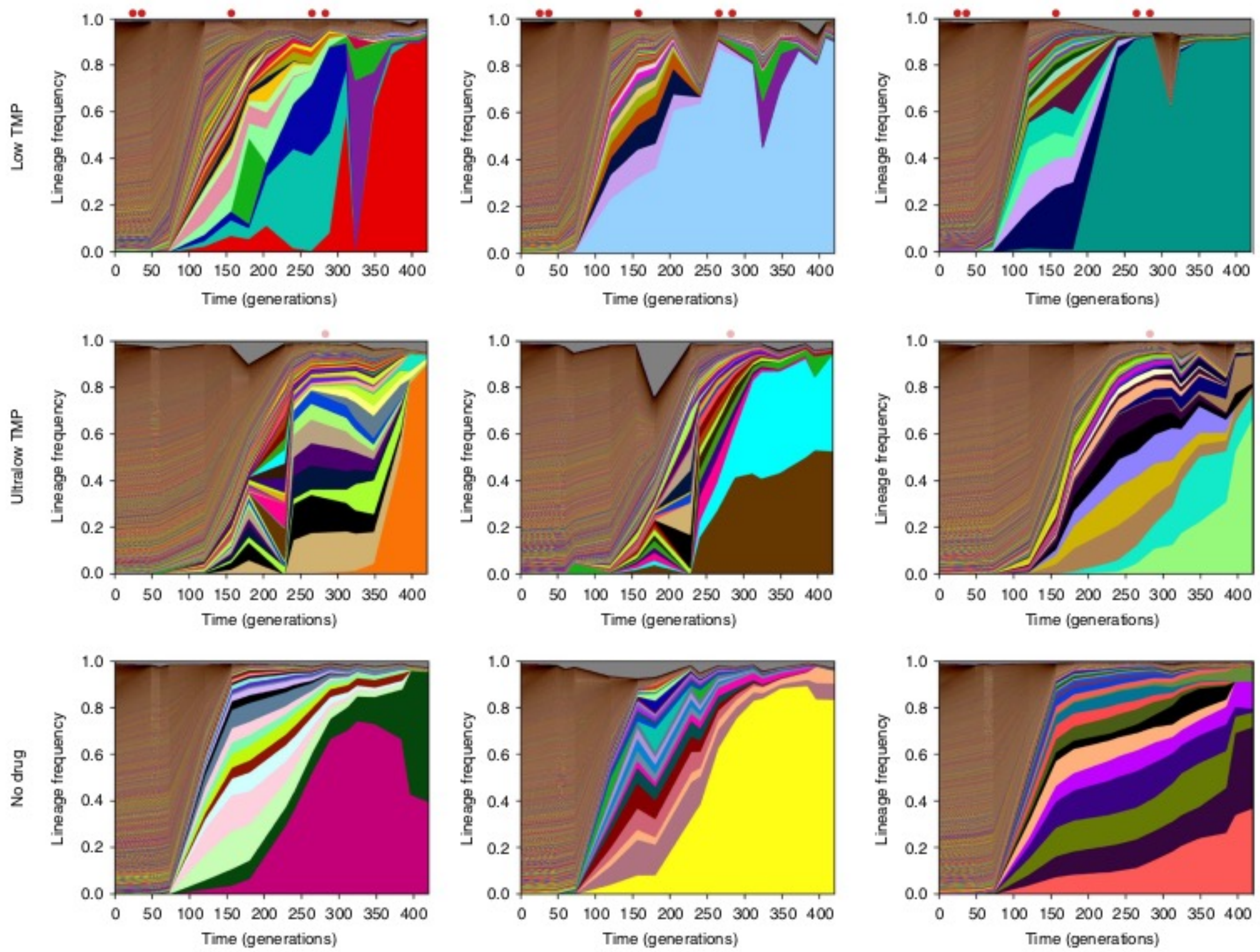
D. It depends

To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Entropy in ecology and evolution

- **Diversity of a bacterial population**



Barcoded lineage frequencies → decay of the initial diversity of the population (no rebarcoding)

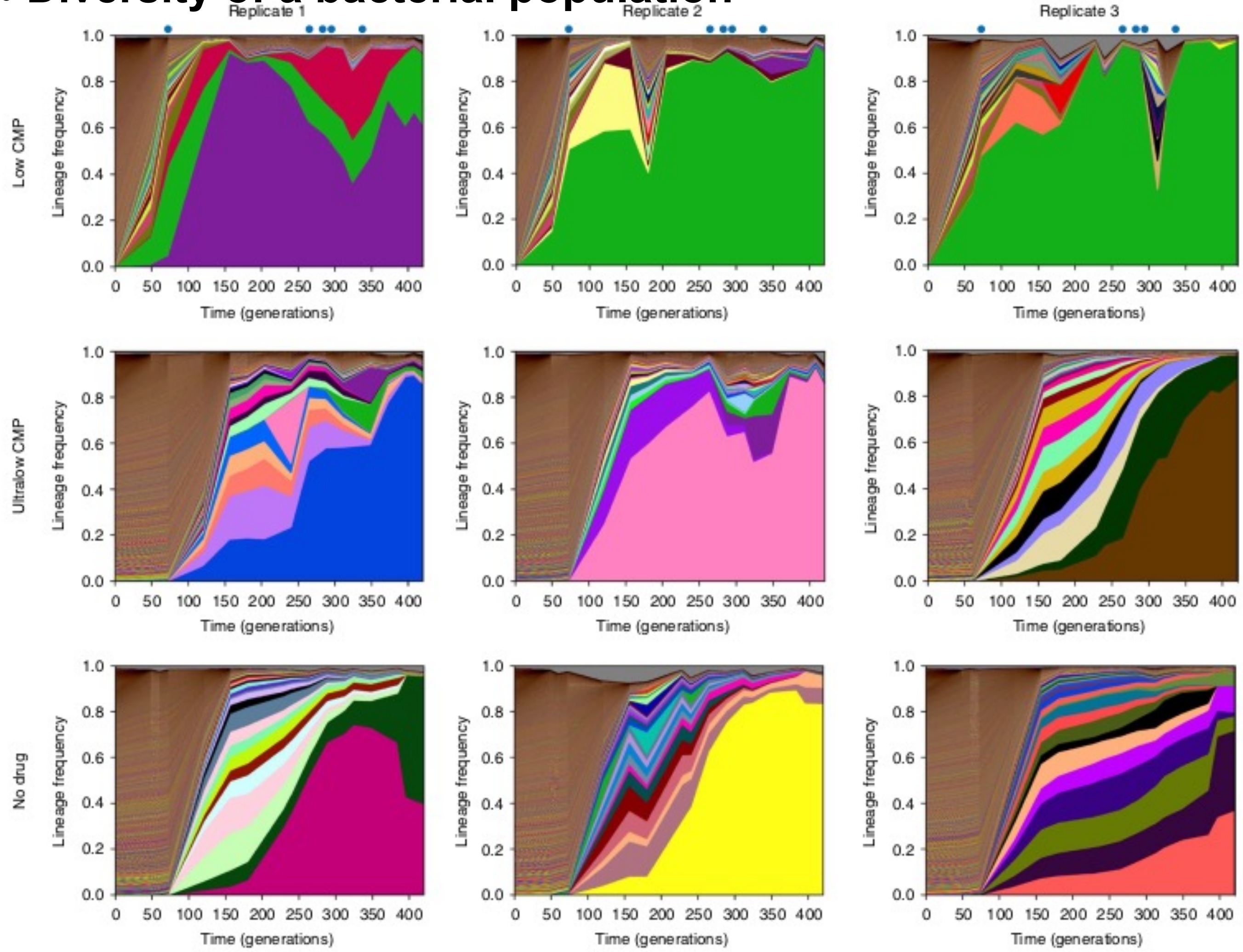
Each color corresponds to a distinct abundant lineage; vertical width indicating its frequency

Dots above each plot mark times at which the drug concentration changed

Jasinska et al, 2020

Entropy in ecology and evolution

- **Diversity of a bacterial population**



Barcoded lineage frequencies → decay of the initial diversity of the population (no rebarcoding)

Each color corresponds to a distinct abundant lineage; vertical width indicating its frequency

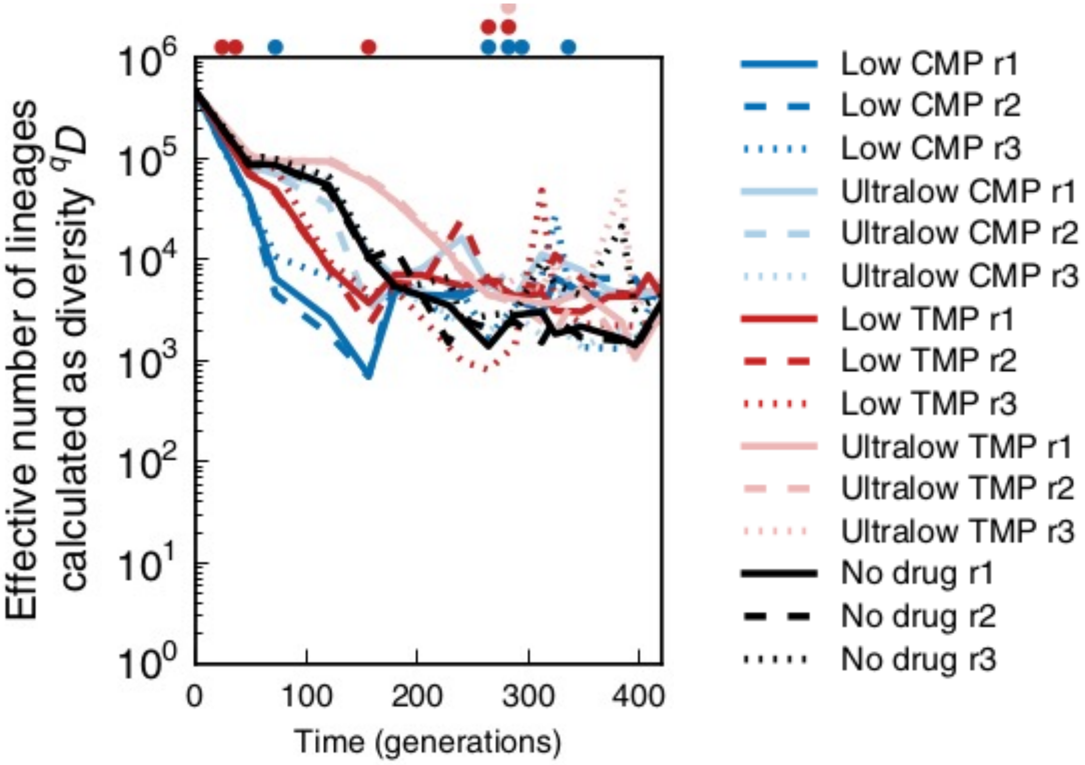
Dots above each plot mark times at which the drug concentration changed

Jasinska et al, 2020

Entropy in ecology and evolution

- **Diversity of a bacterial population**

Number of unique barcoded lineages:



Issue: gives the same importance to rare and frequent lineages

… and there are many rare ones

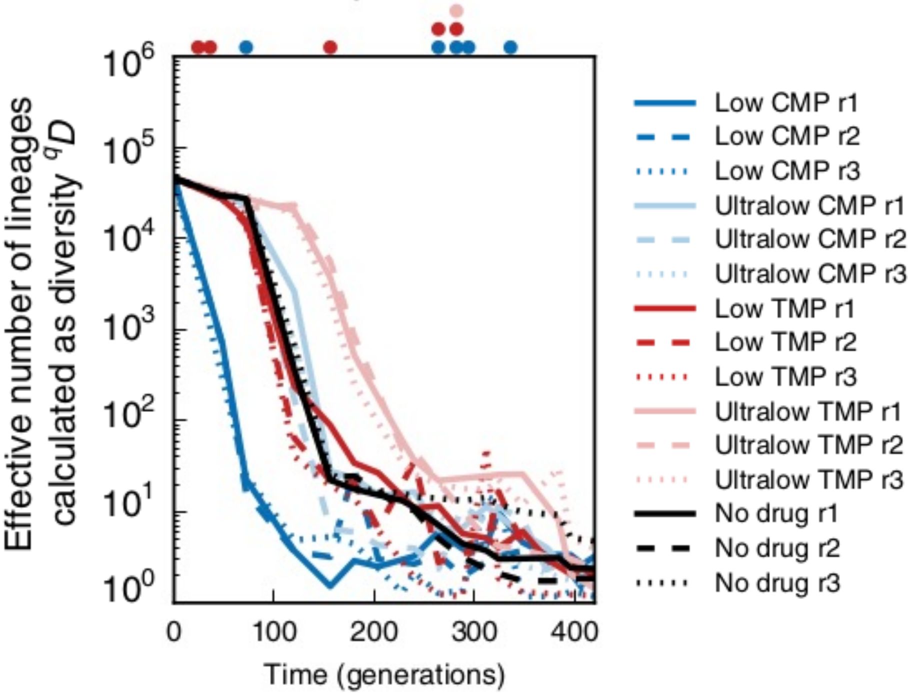
Jasinska et al, 2020

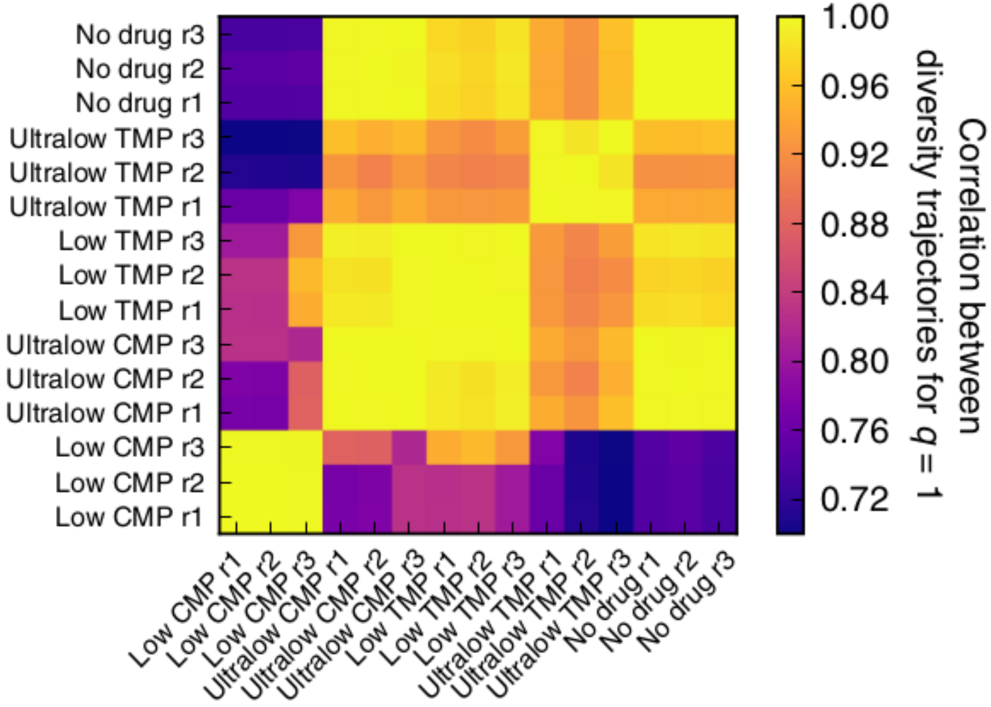
How can we define an effective number of lineages using entropy?

Entropy in ecology and evolution

- **Diversity of a bacterial population**

Effective number of lineages = exponential of the entropy





Jasinska et al, 2020