Information Theory and The Perception-Action Cycle

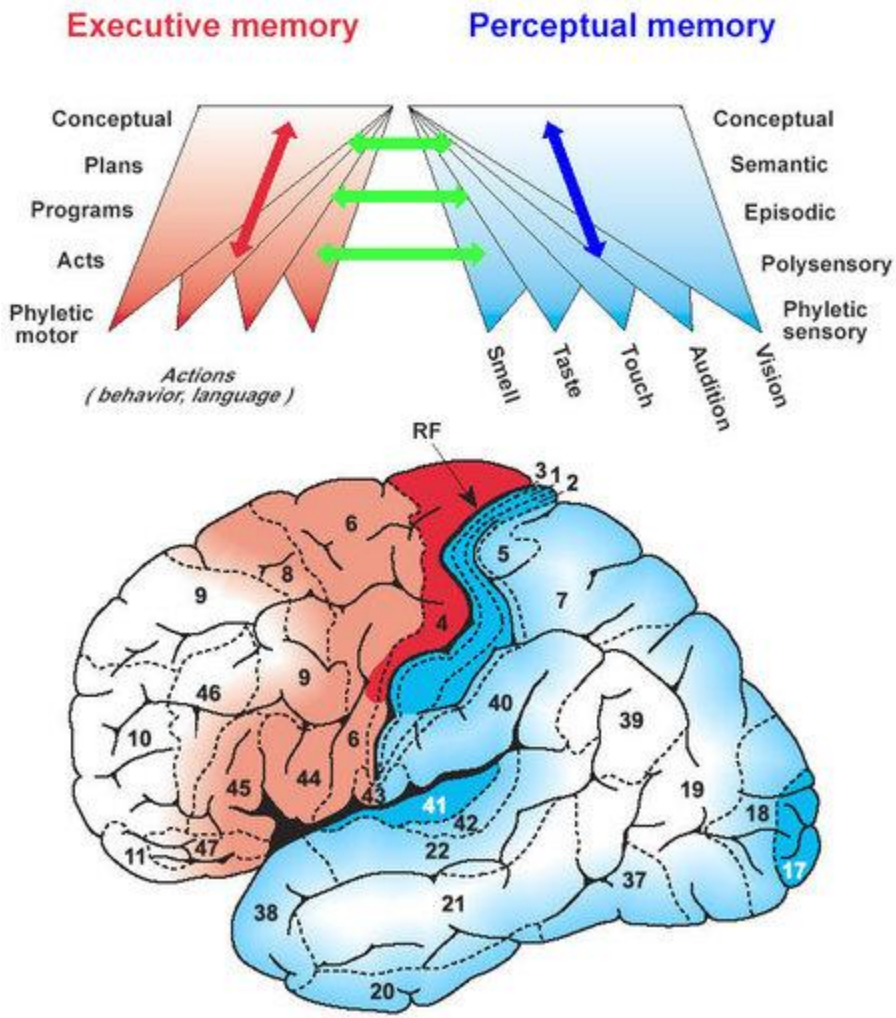
Naftali Tishby

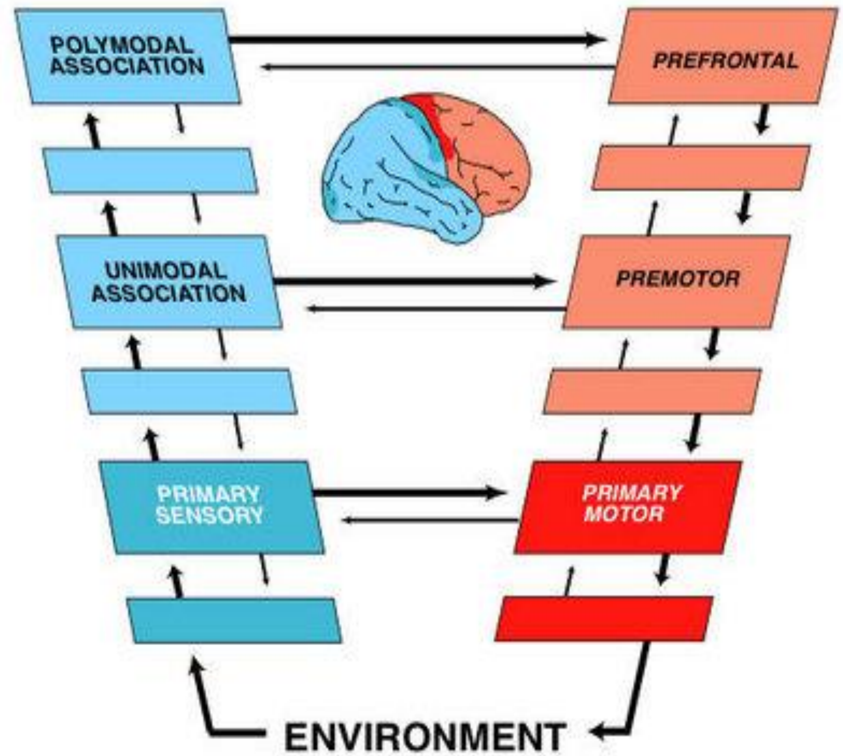
Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem

Information Beyond Shannon

Venice - Italy, December 29-30, 2008,

Perception-Action Cycles

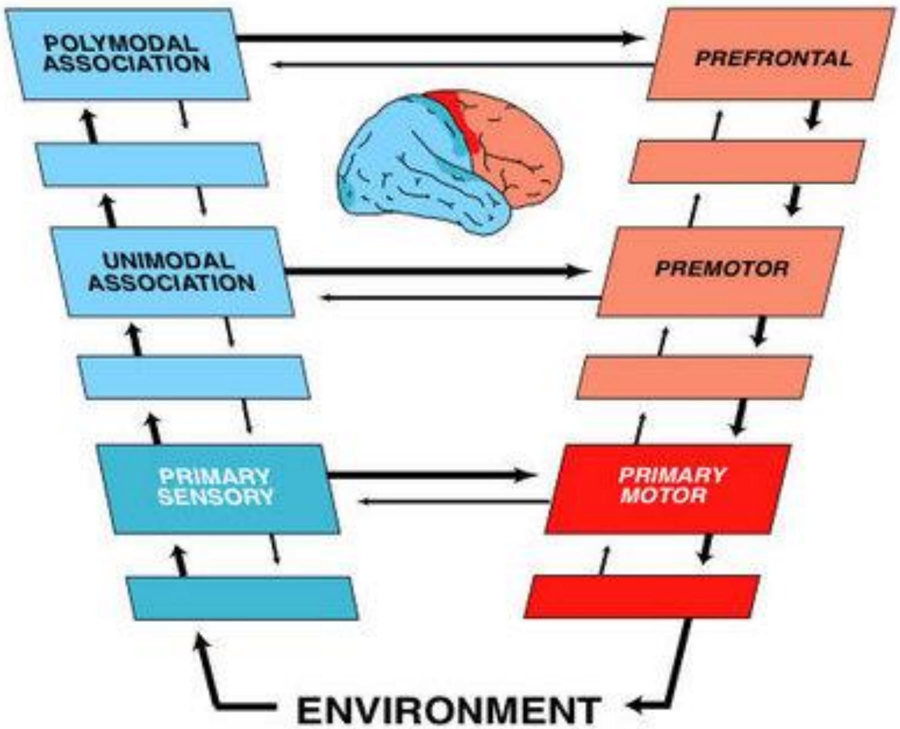


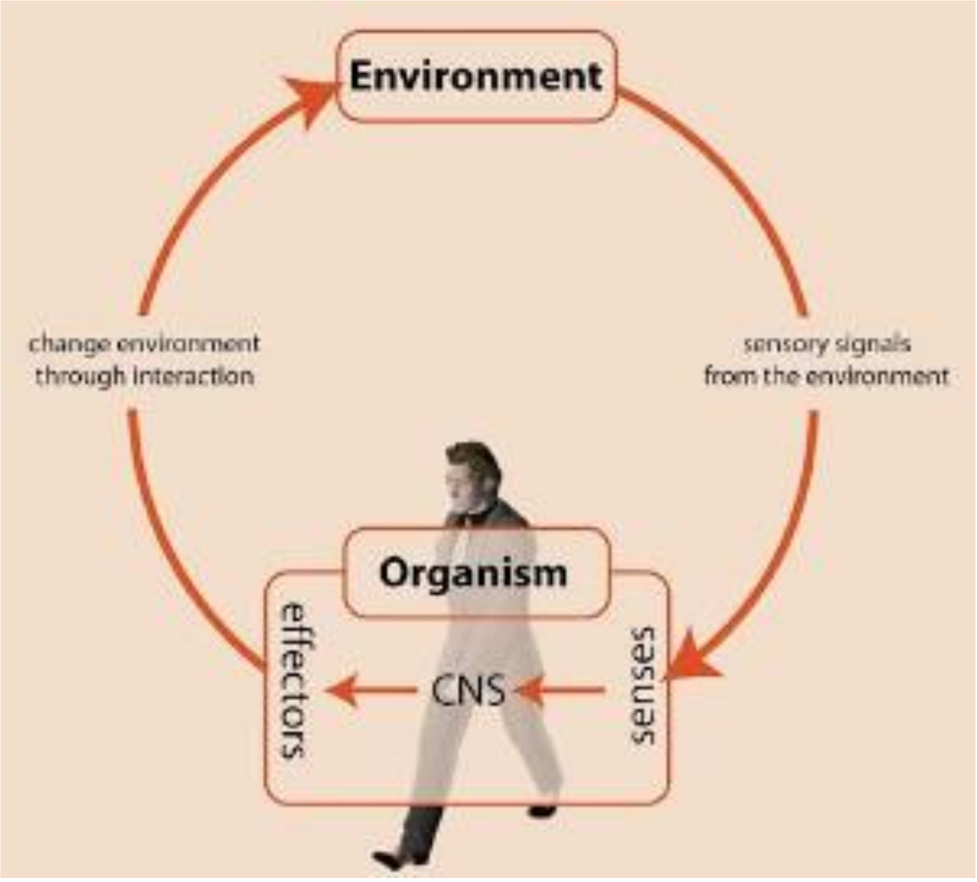


Multiple cycles with Multiple time scales!

The Perception-Action Cycle

The **circular flow of** **information** that takes place **between the organism and its environment** in the course of a sensory-guided sequence of behavior towards a goal.
(JM Fuster)



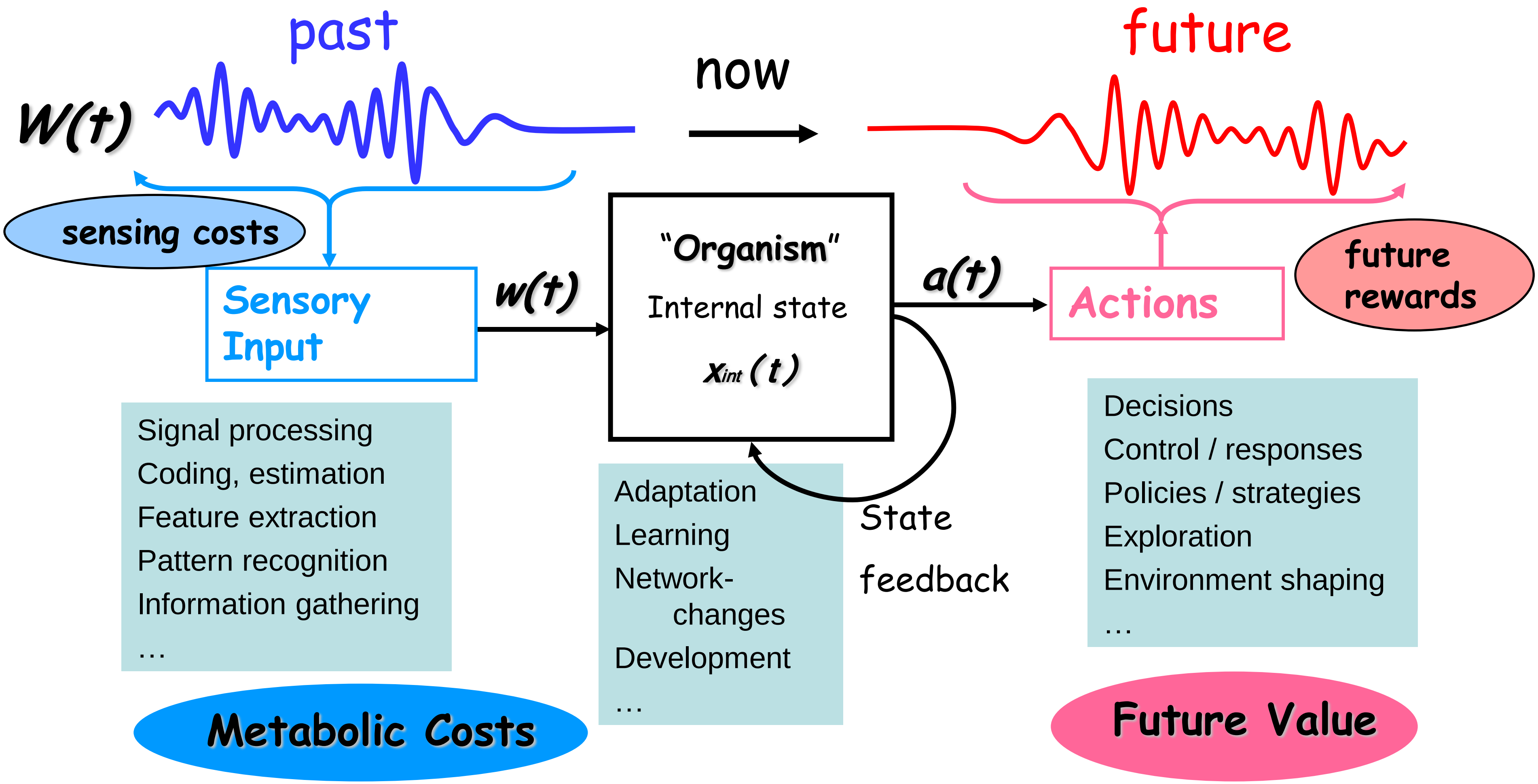


Outline

- Predictive information and the perception-action cycle
  - A model for the circular flow of information in the cycle(s)
  - The analogy with Shannon's Information Theory
  - The unknown **future** as the **channel input**
  - The future-past channel capacity: **Predictive Information**
- Two solvable examples
  - Gambler in a binary world
    - Optimal solution: the **Past-Future Information Bottleneck**
  - A linear system in a Gaussian environment
    - Optimal (Kalman-Ho) dimension reduction in LQR control
- Application to neuroscience
  - Surprise in Auditory Perception
    - Or why do we enjoy music?

A conceptual framework

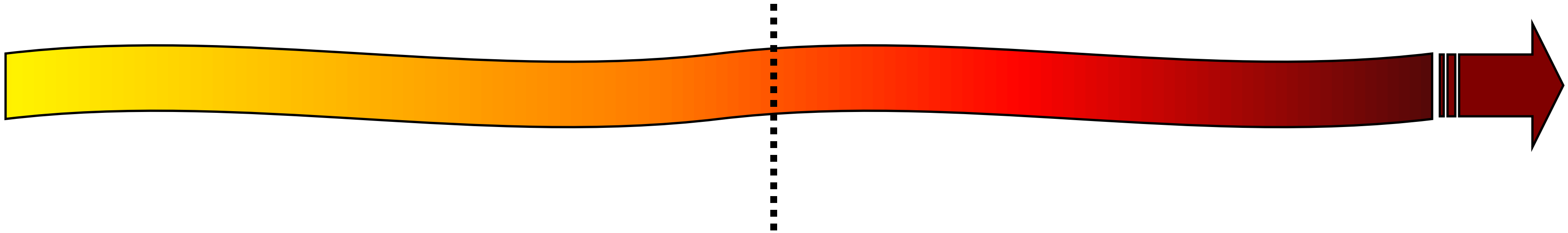
The "**Environment**": Partially observed, (stationary?) stochastic process



We must simplify …

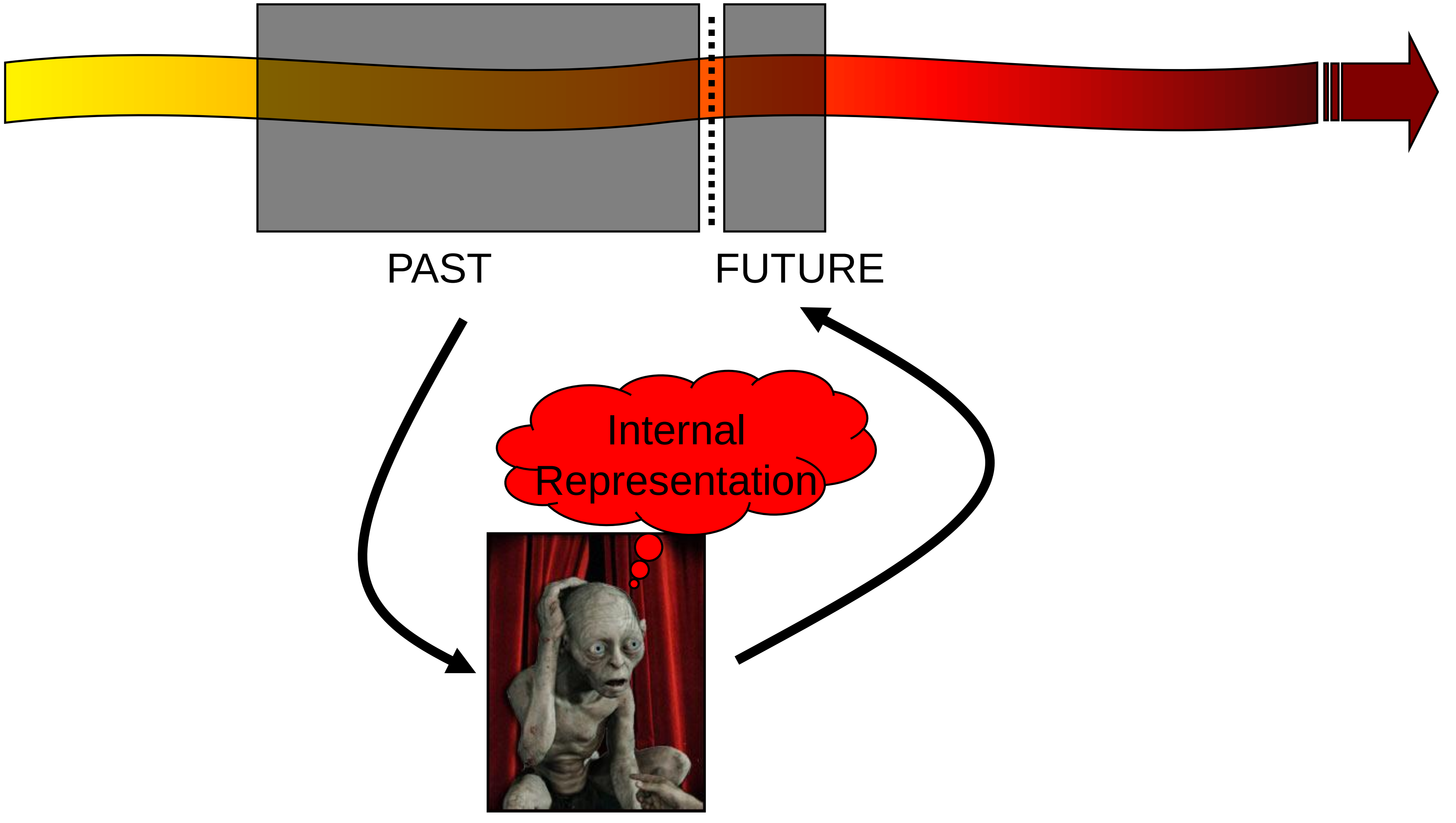
(…hopefully not oversimplify…)

Internal Representations

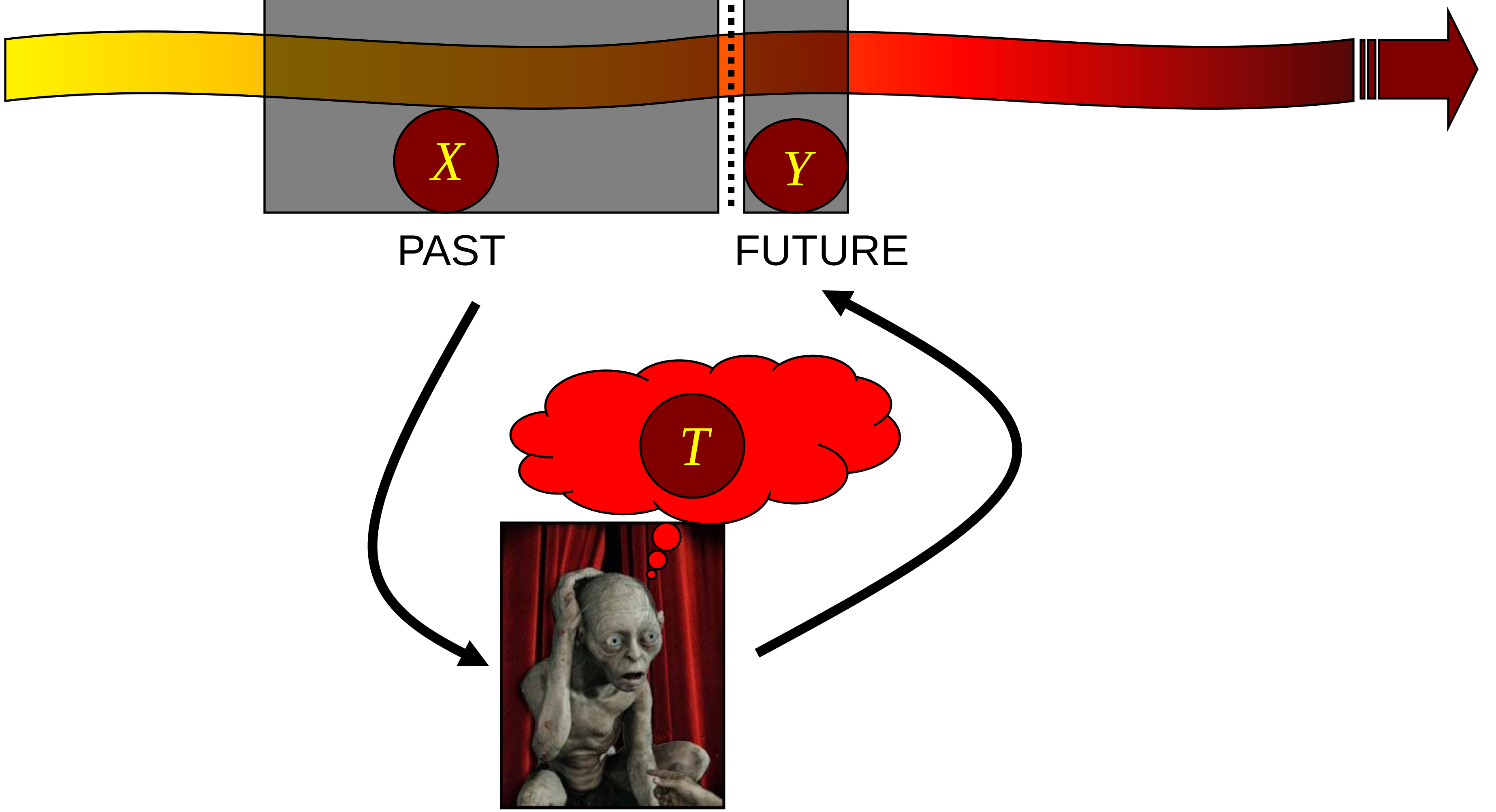


The Environment: stationary stochastic process

Internal Representations

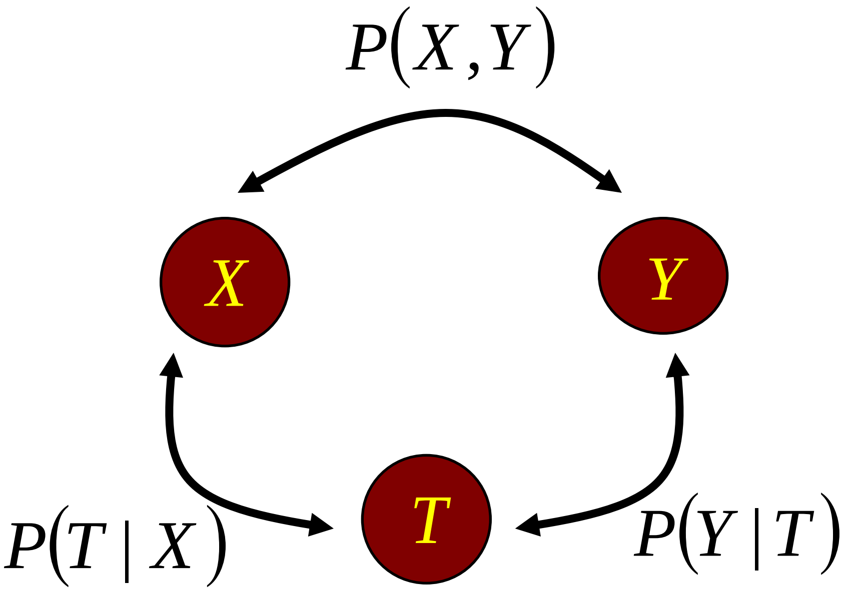


Internal Representations



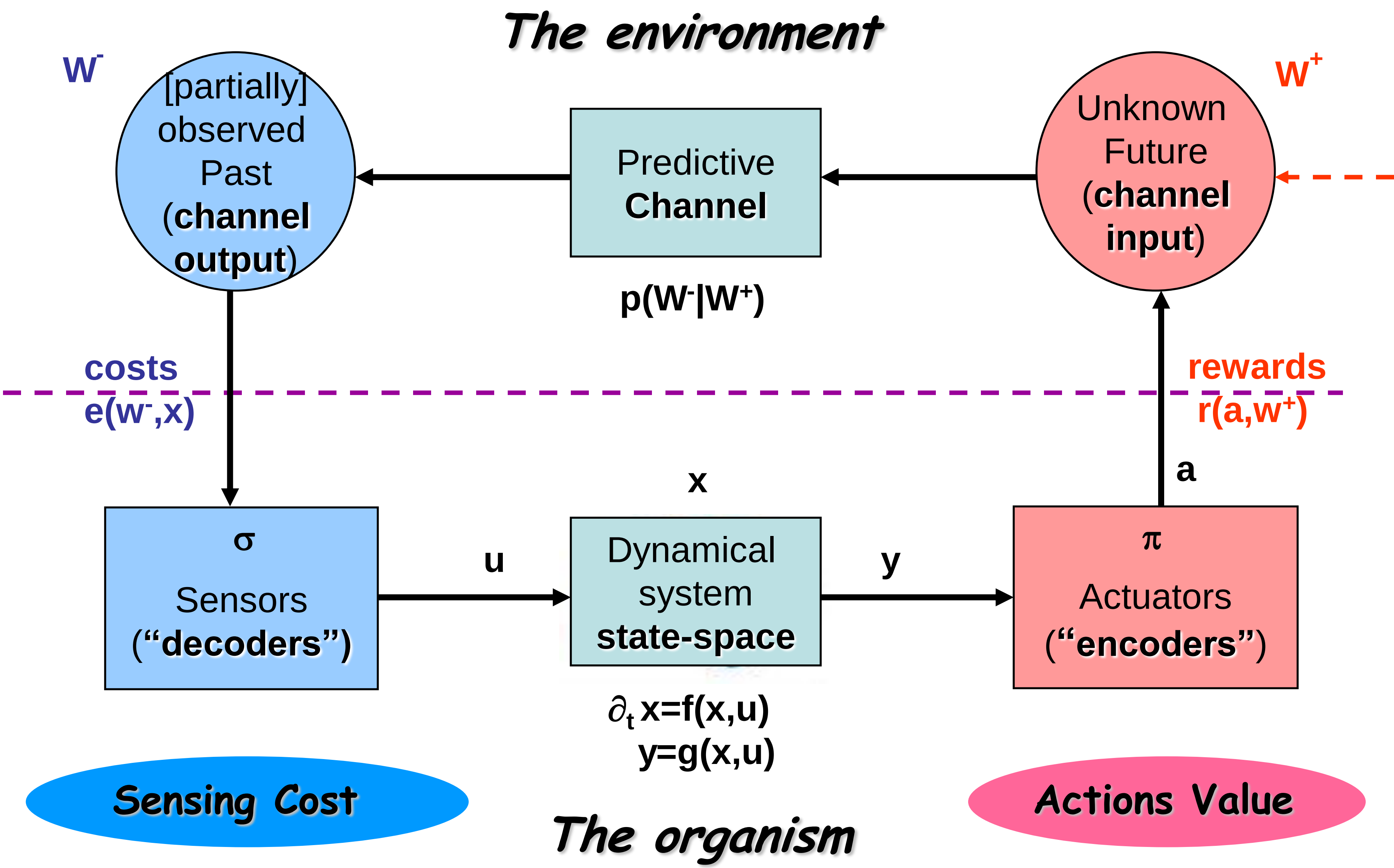
(Optimal) Internal Representations

we like to think probabilistically

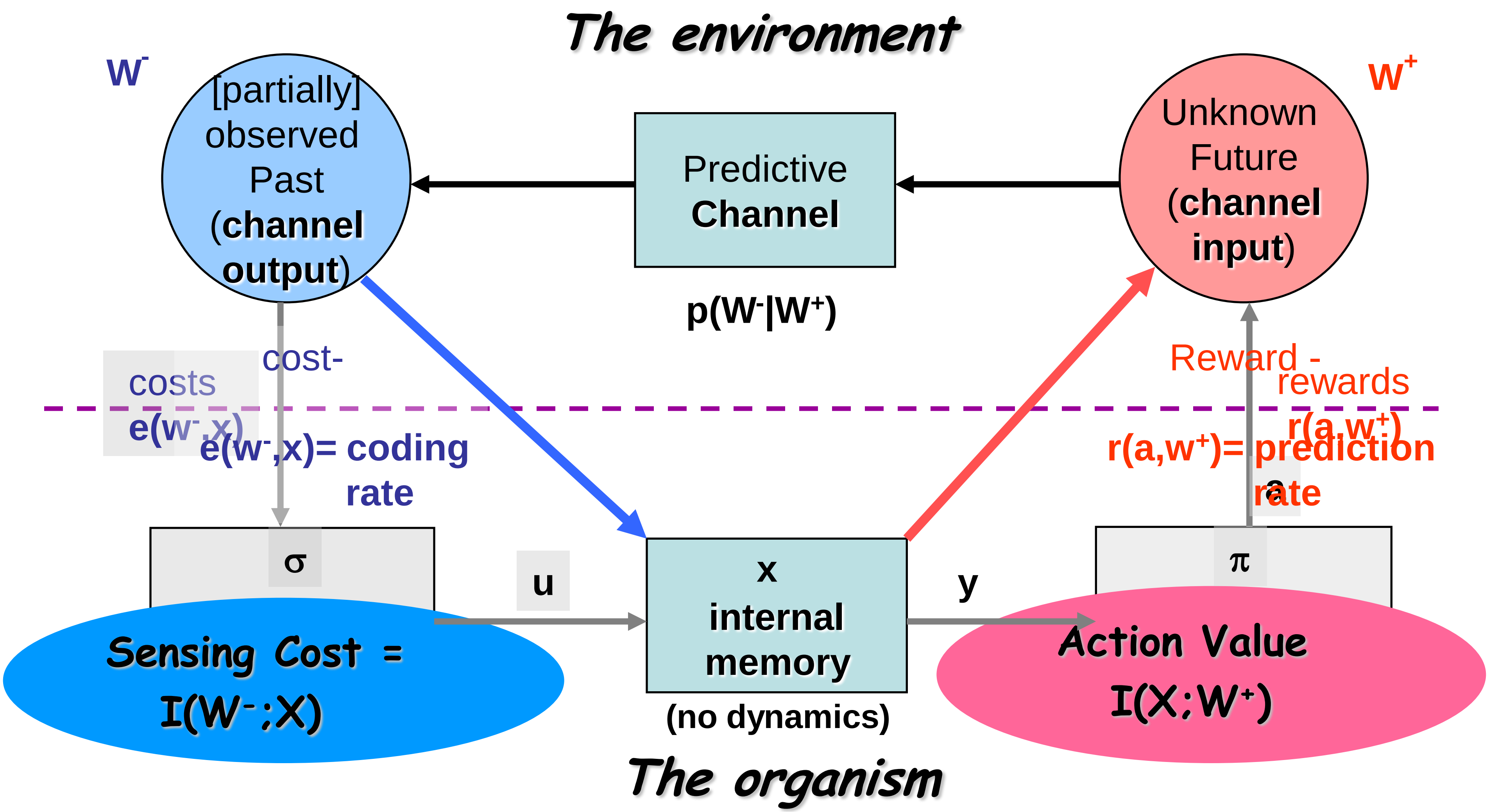


- Environment: P(X,Y)
- Internal representation: P(T|X), P(Y|T)

Information Theoretic view of The Perception-Action Cycle



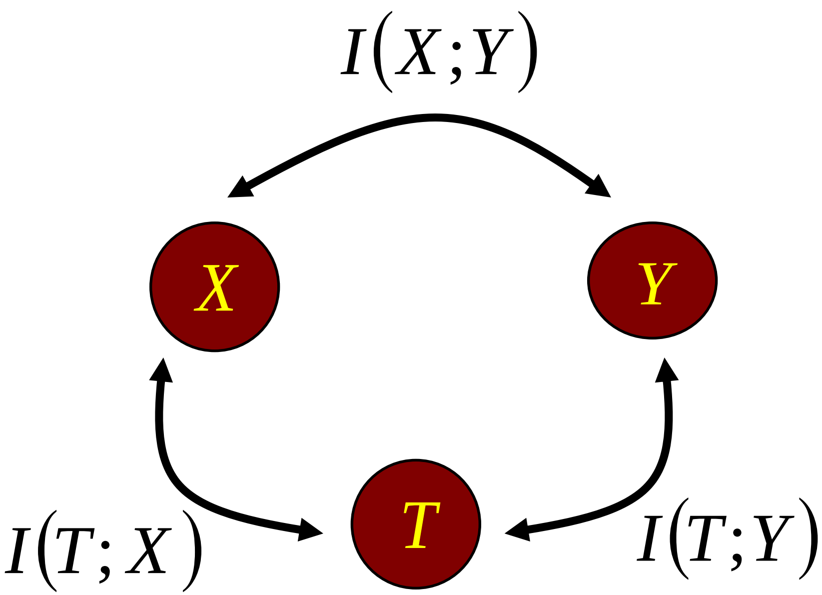
Simpler
Perception-Action Cycle



Optimum: The Information Bottleneck optimal decoders/predictors

(Optimal) Internal Representations

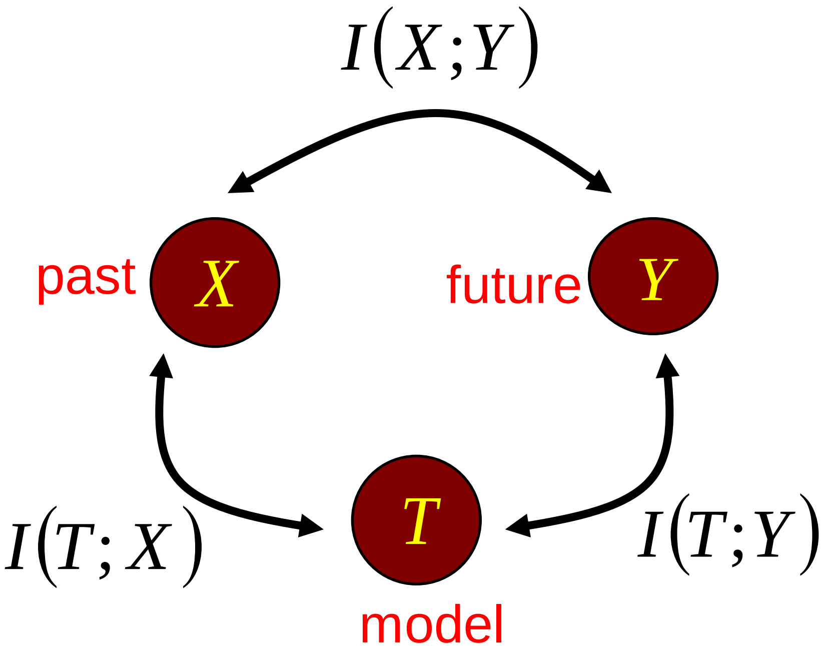
and we want a computational principle…



- Environment: I(X;Y) – predictive information
- Internal representation: I(T;X) , I(T;Y) - compression & prediction

(Optimal) Internal Representations

and a computational principle…



Model Quantifiers:

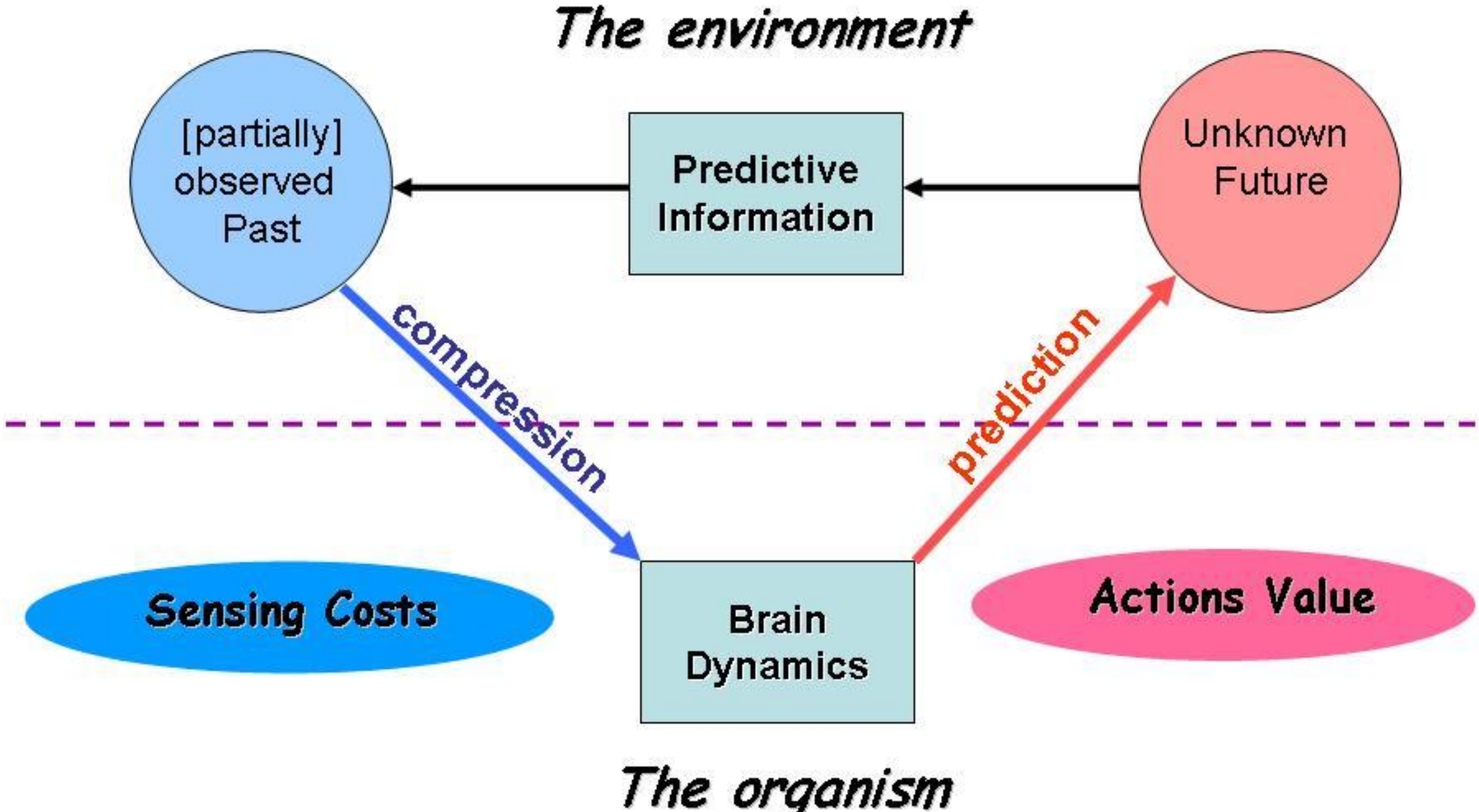
- Complexity ("cost"): I (T;X)
- Predictive Info ("value"): I(T;Y)

Optimality Trade-off:

- minimize complexity
- maximize predictive-info

- Environment: I(X;Y) – predictive information
- Internal representation: I(T;X) , I(T;Y) - compression & prediction

Perception-Prediction-Action Cycle



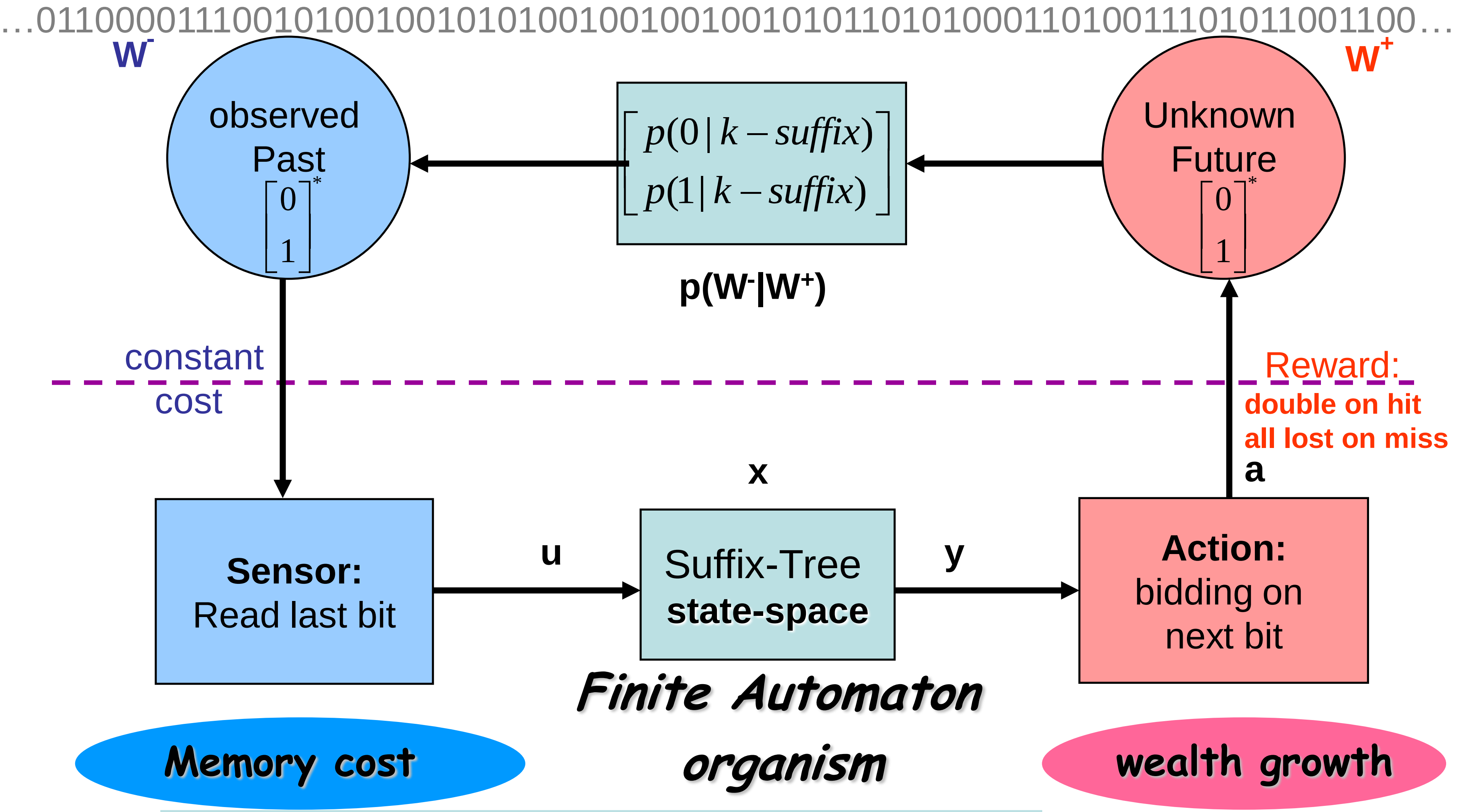
The Past-Future Information Bottleneck

A simple example:

The compulsive gambler in a binary world

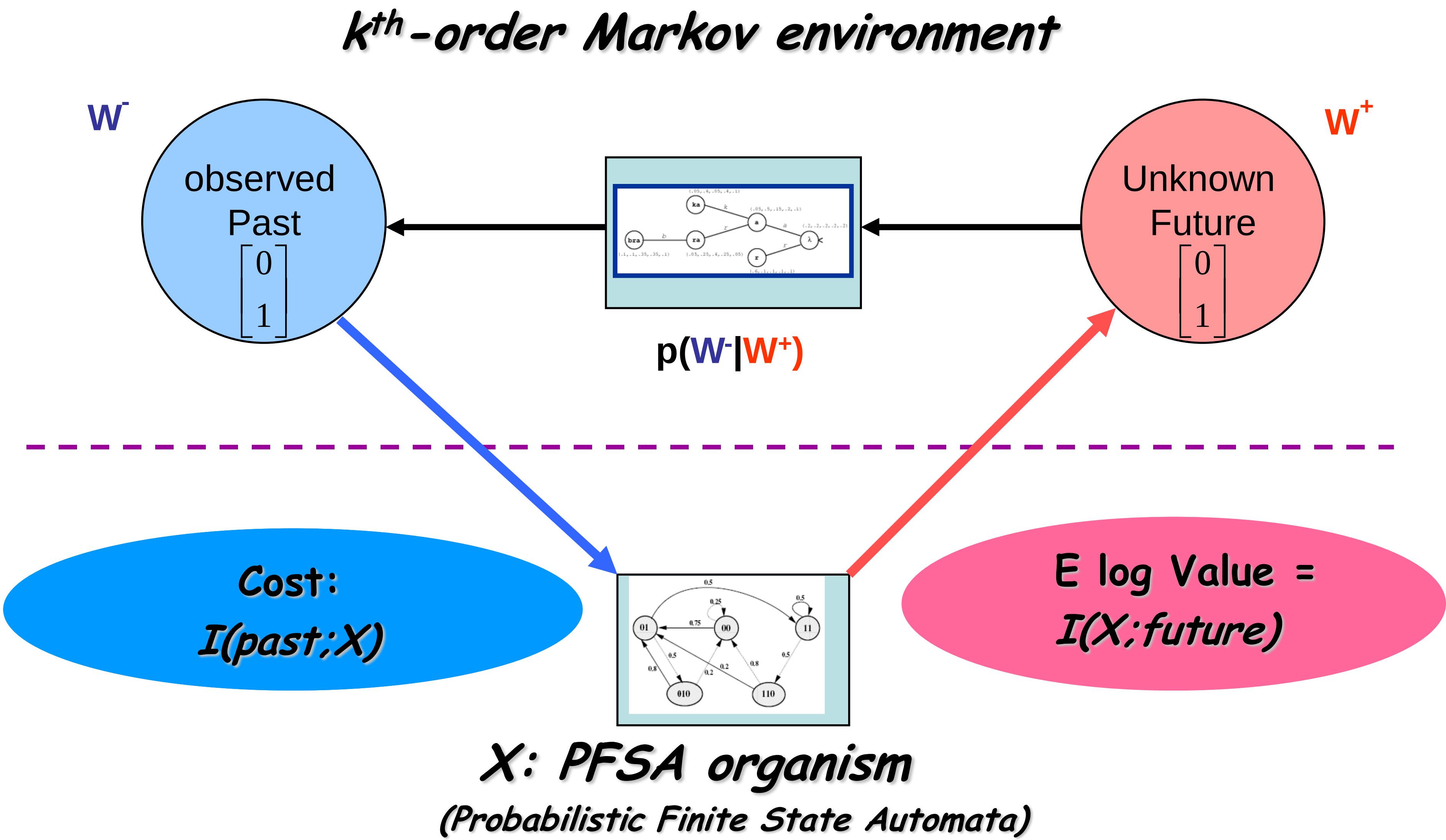
A solvable example

A Gambler in a k-order Markov environment



Optimum: any side-information helps

The optimal compulsive gambler

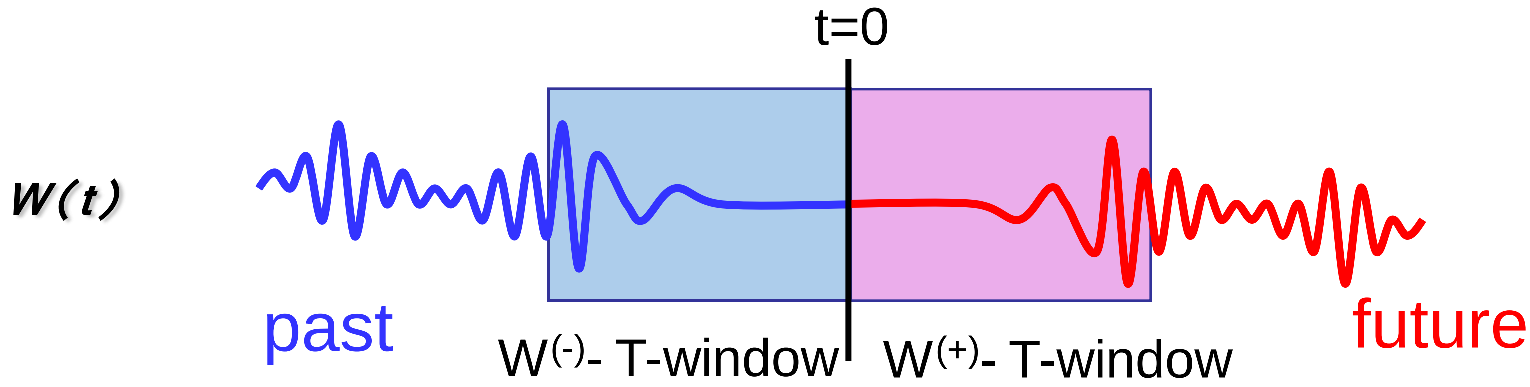


Optimum: proportional bidding with IB predictive information

The Predictive Channel

Predictive Information: The Capacity of the Future-Past Channel

(with Bialek and Nemenman, 2001)



- *Estimate* $P^T(W^{(-)},W^{(+)})$: *T- past-future distribution*

$$I_{pred}[T] = \left\langle \log \frac{p(W^T{}_{future} \mid W^T{}_{past})}{p(W^T{}_{future})} \right\rangle_{p(W_{past},W_{future})}$$

Logarithmic growth for finite dimensional processes

- Finite parameter processes (e.g. Markov chains)

$$I_{pred}(T \to \infty) \approx \frac{\dim(\theta)}{2} \log T$$

- Similar to stochastic complexity (MDL)

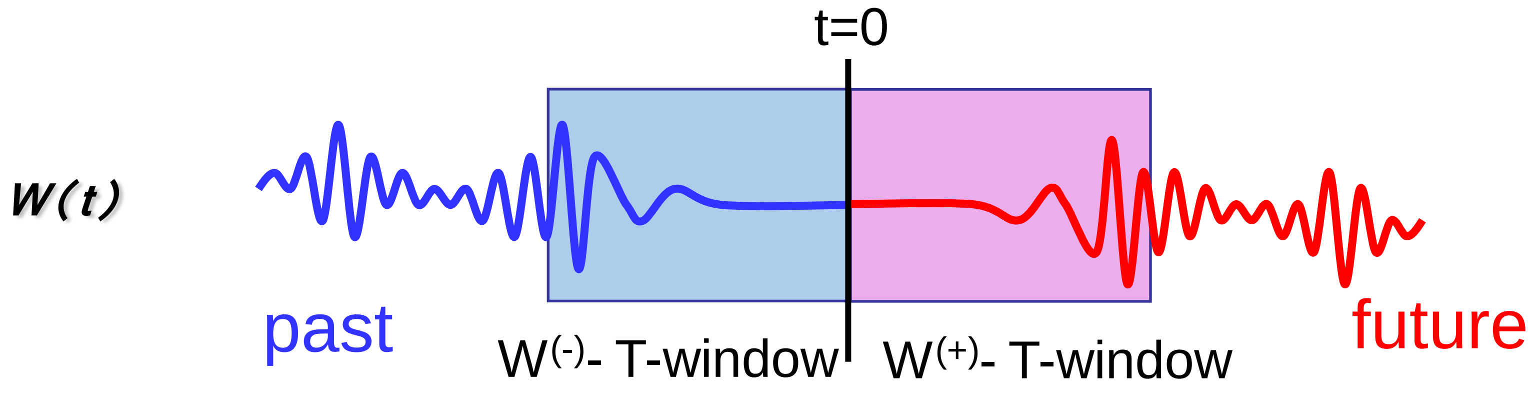
Power law growth

- Such fast growth is a signature of infinite dimensional processes

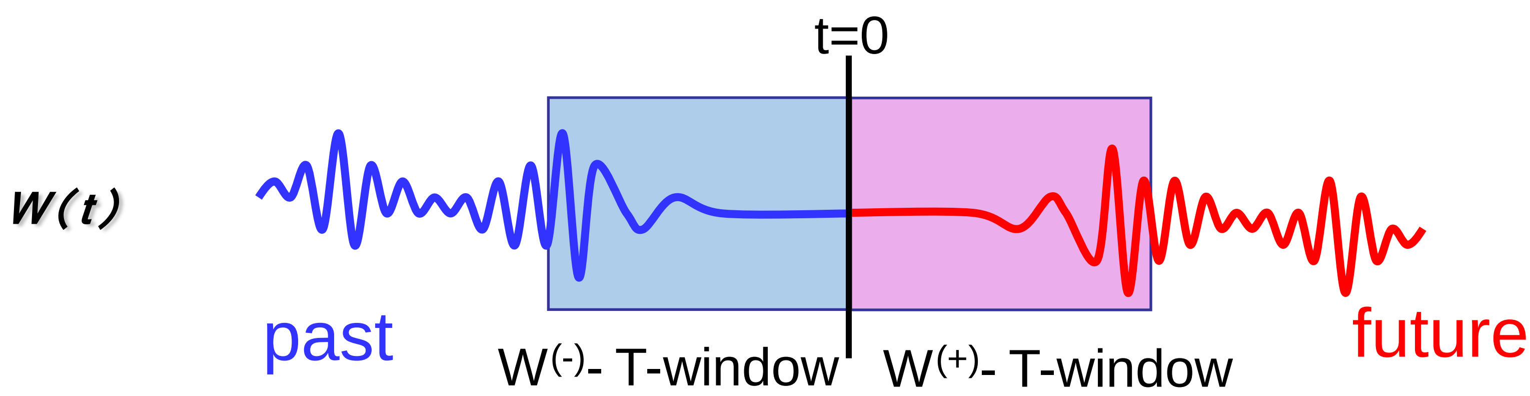
$$I_{pred}(T \to \infty) \approx T^{\alpha} \qquad \alpha < 1$$

- Power laws emerges in cases where the interactions/correlations have long range

But WHAT - in the past - is predictive ?



The predictive capacity has multiple scales



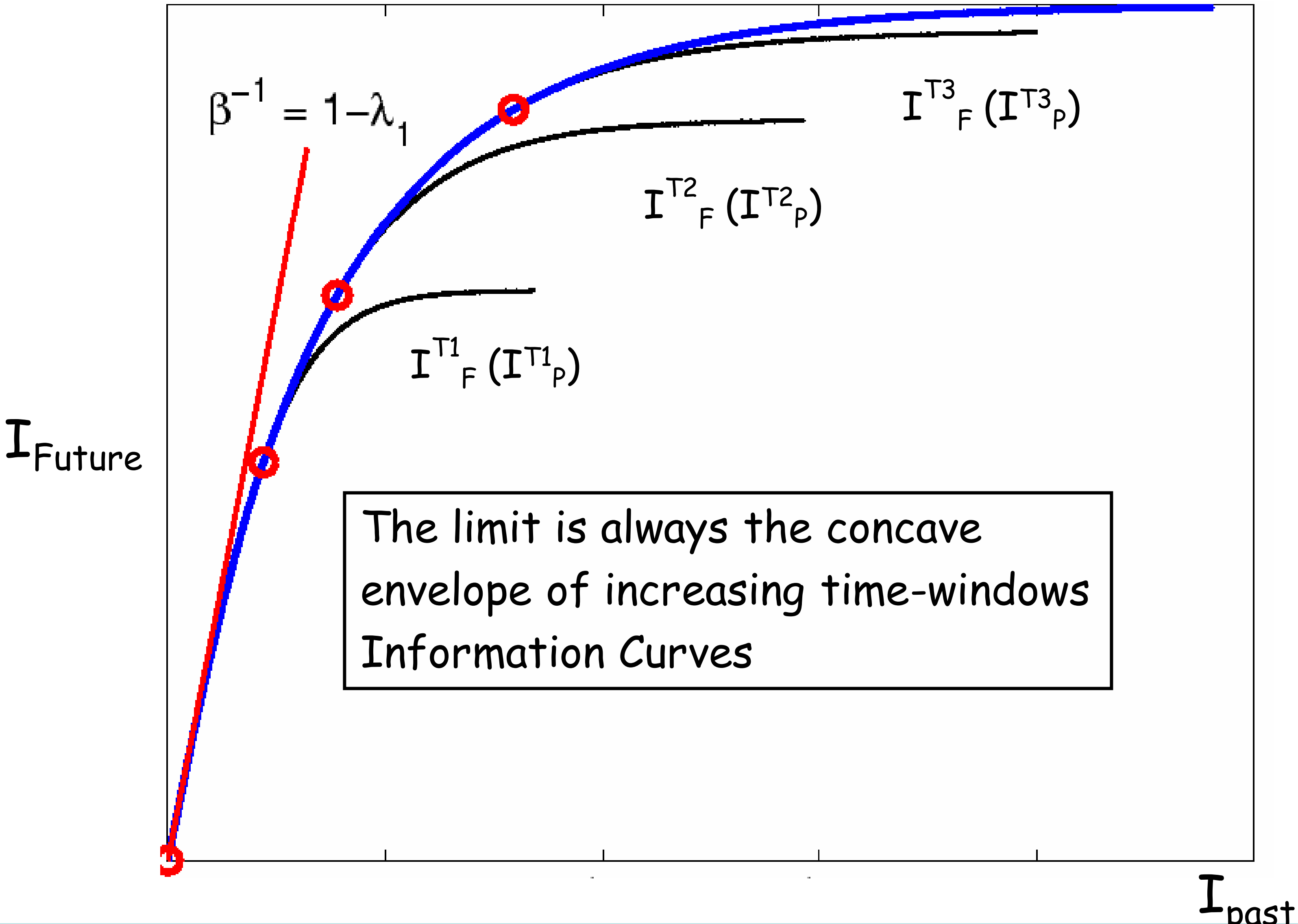
- *Find the "relevant part" of the past w.r.t. future…*

Solve:

$$\text{Min}_Z \; I(W^{(-)};Z) - \beta I(W^{(+)};Z) \quad \textit{for all } \beta > 0$$

⟹ *T- past-future information curve:* $I^T_F(I^T_P)$

- $$I_{Future}(I_{Past}) = \lim_{T\to\infty} I^T_F(I^T_P)$$



The environment's Predictive Information bounds the cycle's efficiency and the **Perception-Action Capacity**

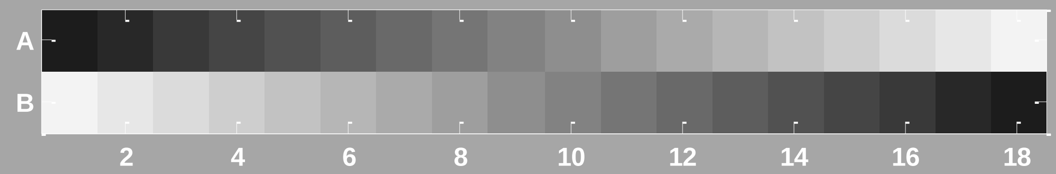
.

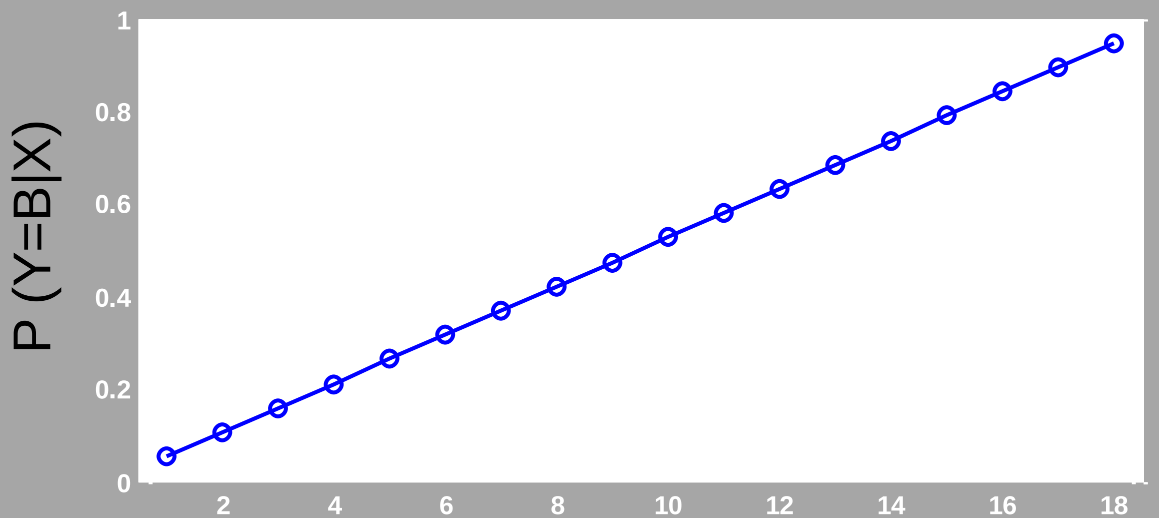
A simple illustration

$$x \in \{1,2,\ldots,18\} \quad , \quad |X| = 18$$

$$y \in \{A, B\} \quad , \quad |Y| = 2$$

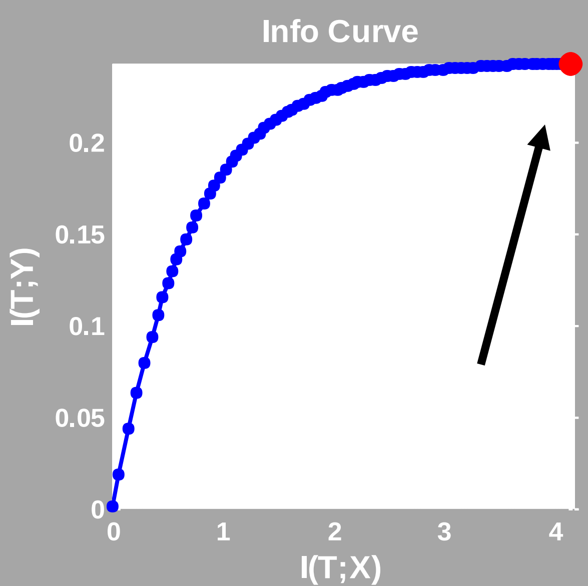
$$P(X,Y)$$

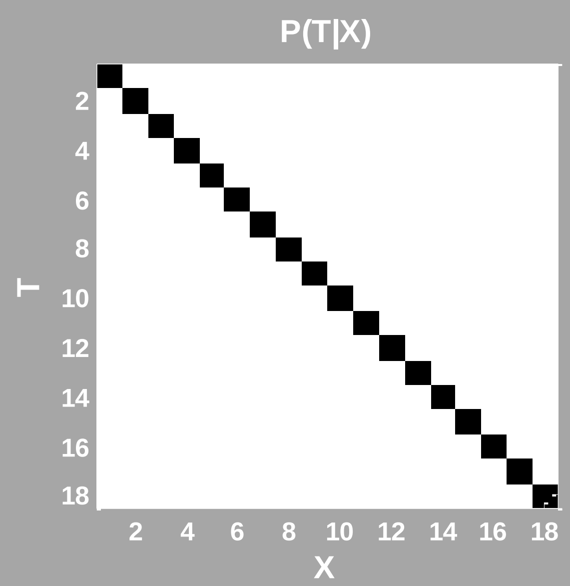


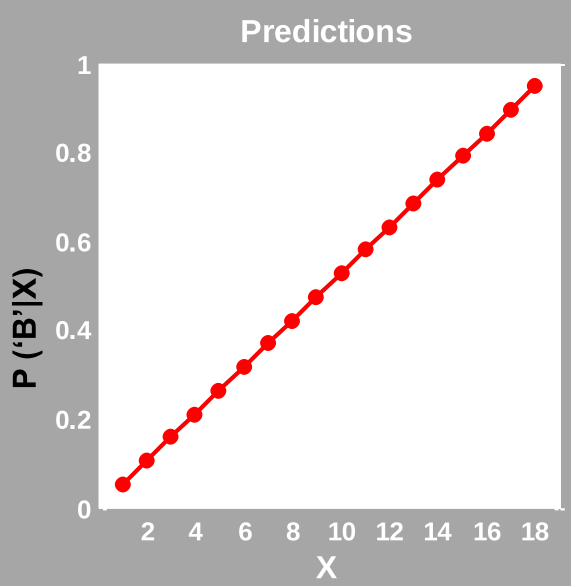


A simple illustration

(most complex) (perfect copy) (perfect predictions)

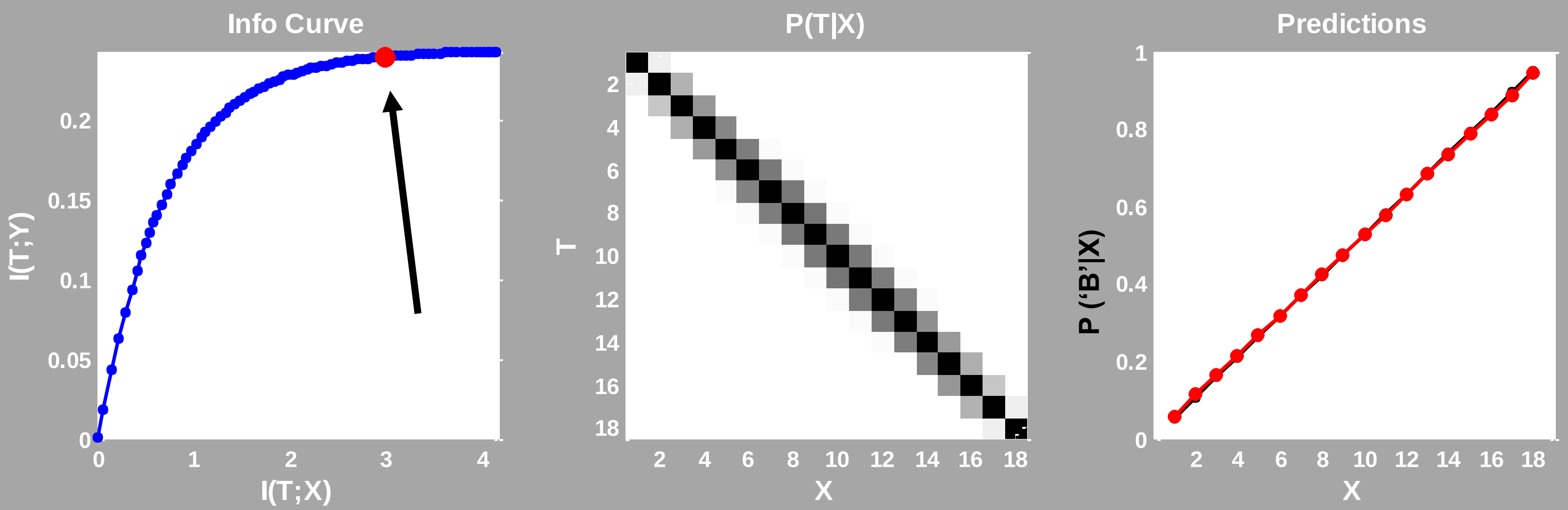






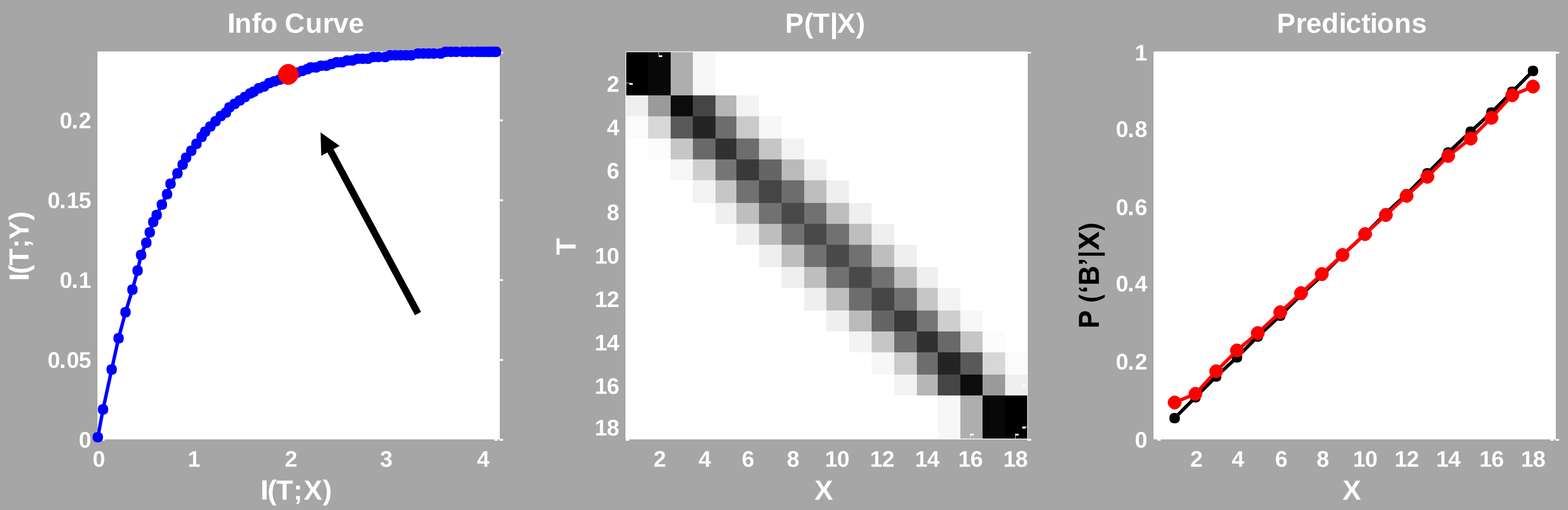
$$T = X \quad , \quad I(T;X) = H(X)$$

A simple illustration



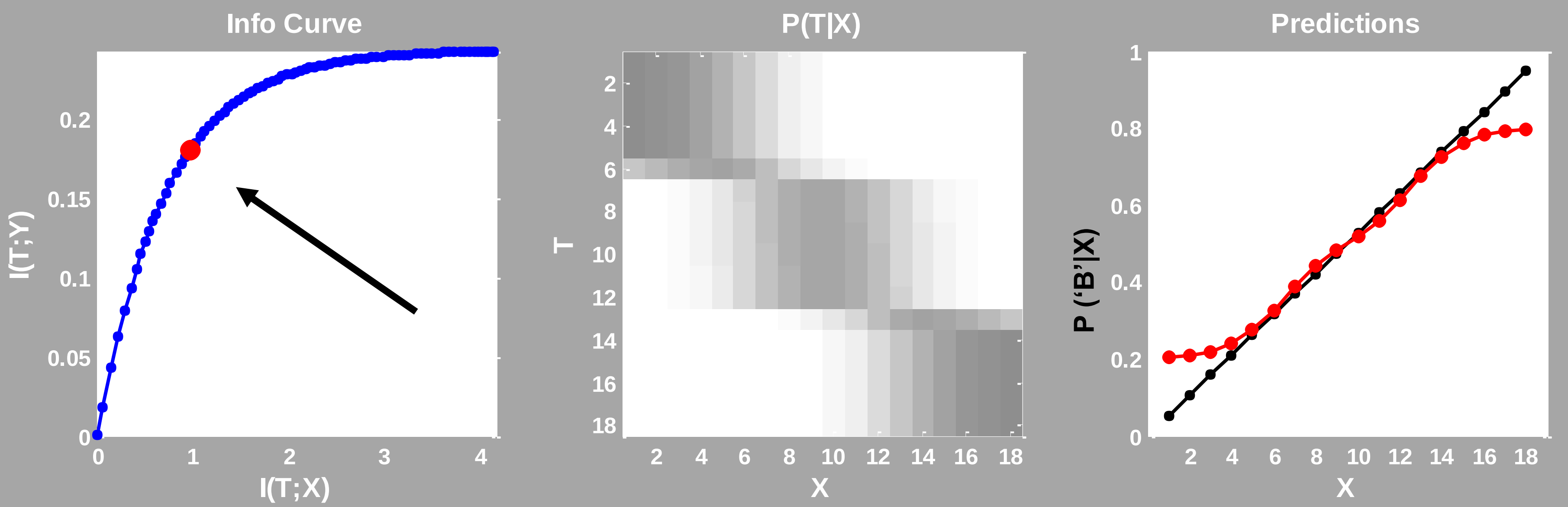
$$I(T;X)=3\,\text{bit}$$

A simple illustration



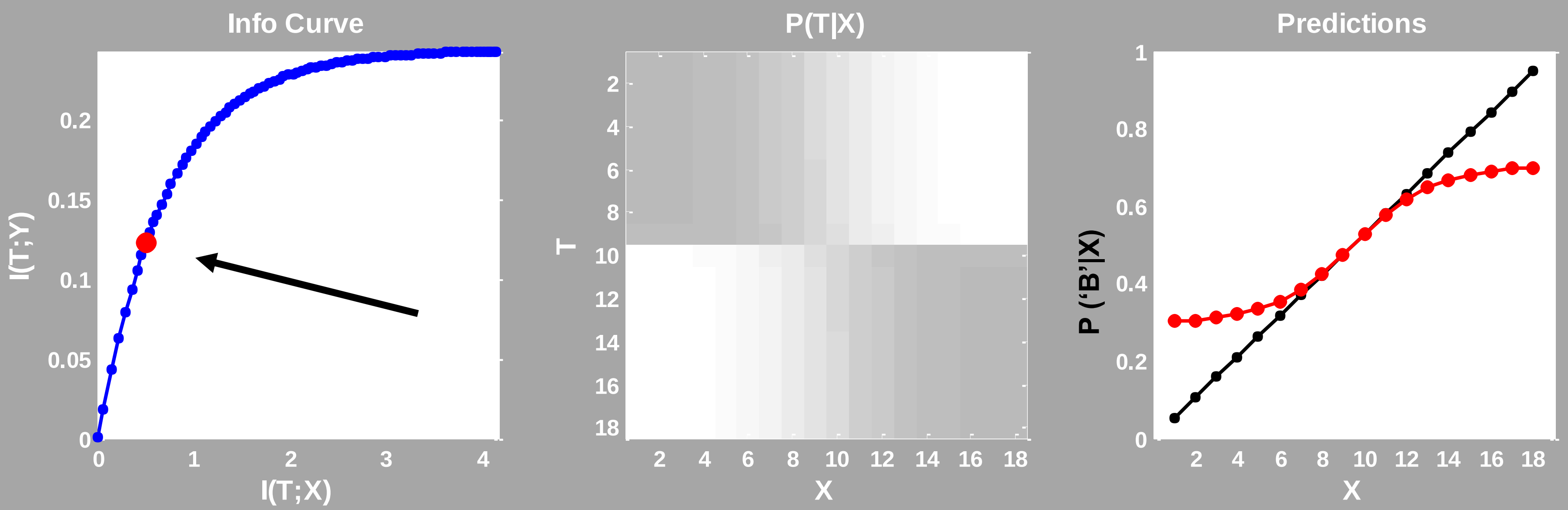
$$I(T;X) = 2\,\text{bit}$$

A simple illustration



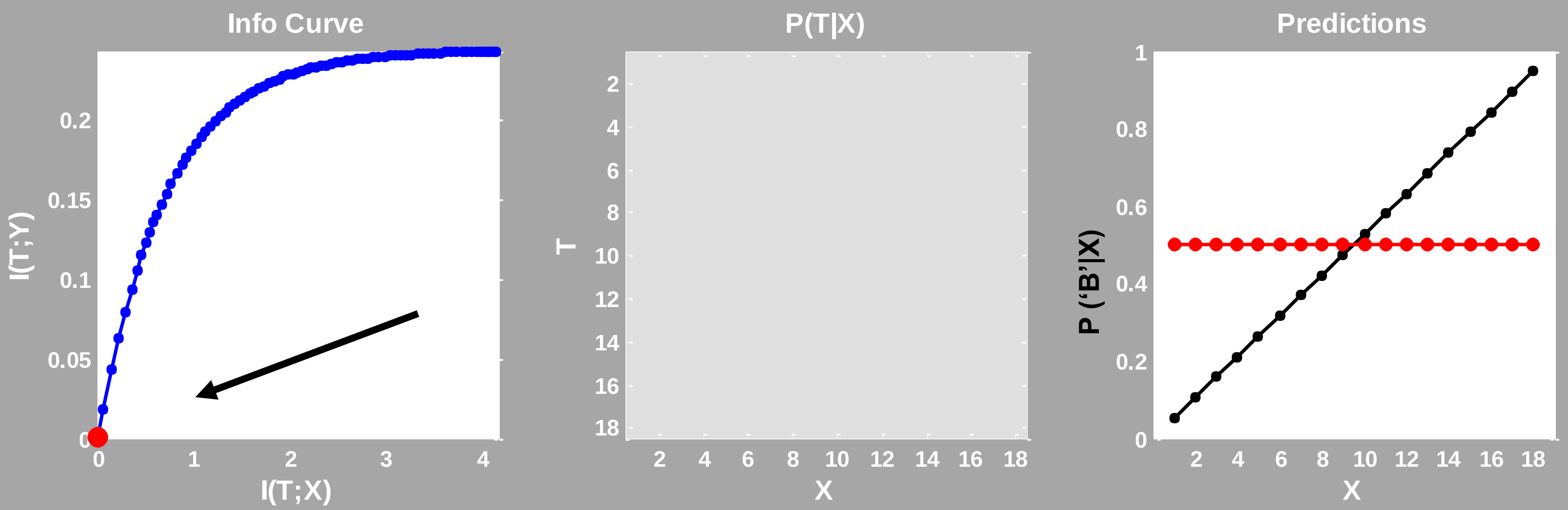
$$I(T;X)=1\text{bit}$$

A simple illustration



$$I(T;X) = 0.5\,\text{bit}$$

A simple illustration



$$I(T;X) = 0\,\text{bit}$$

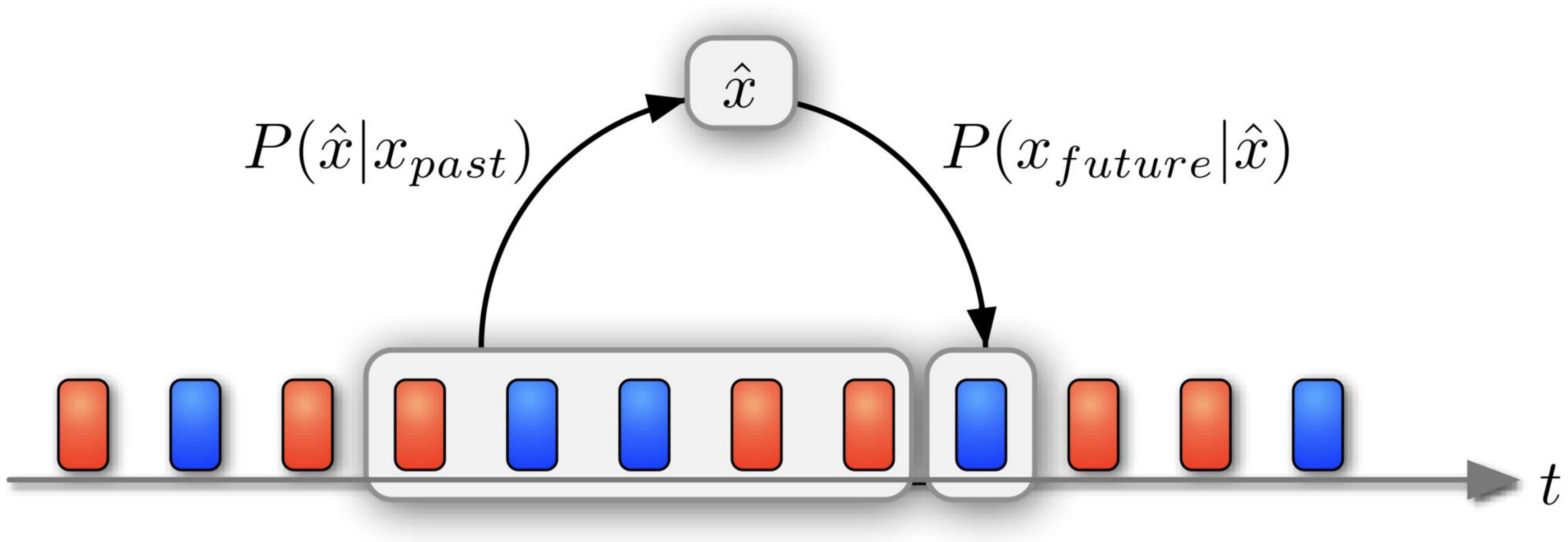
Application to neuroscience:

Auditory cortex encodes surprise

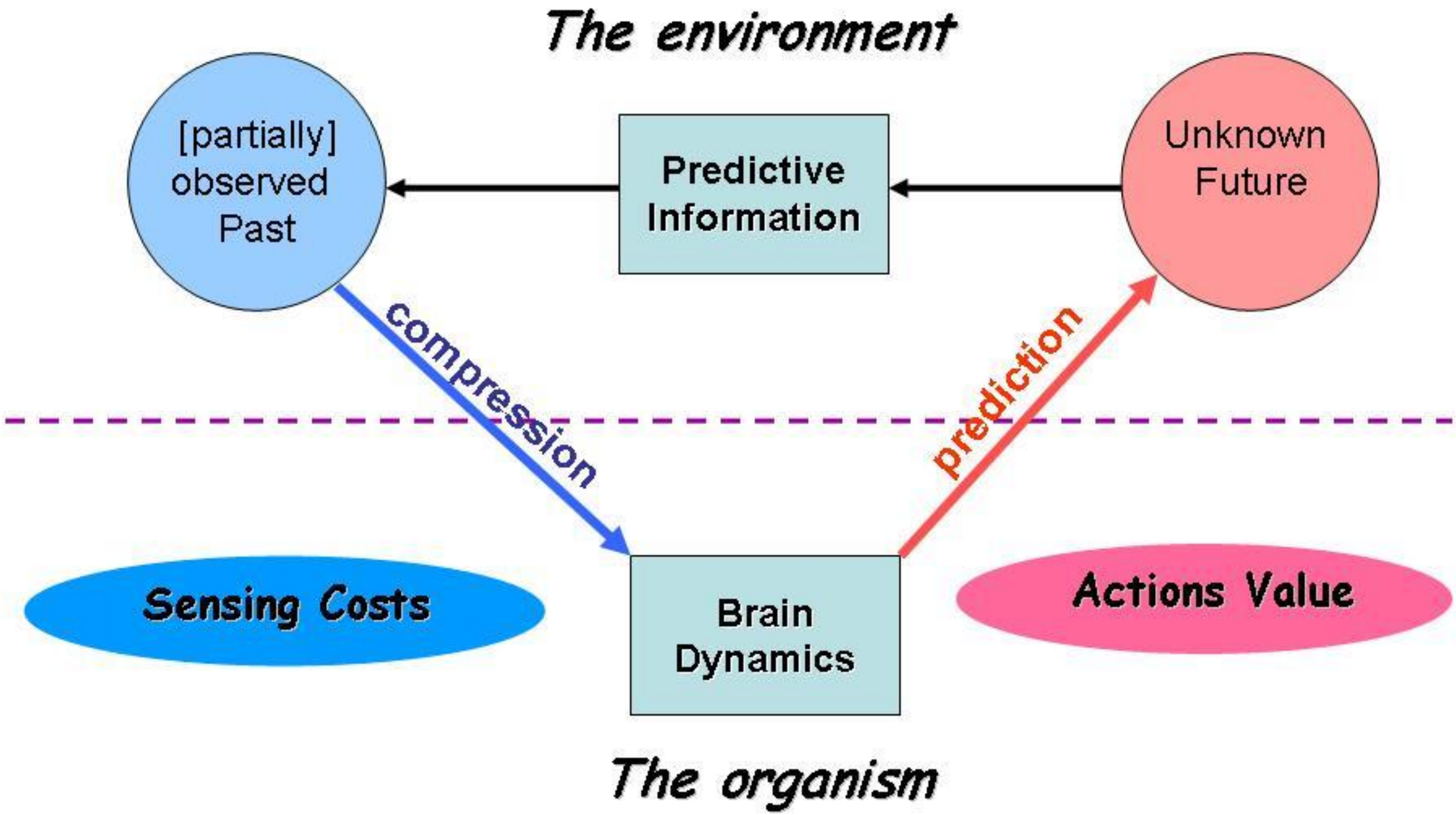
(or why do we enjoy music?)

(with Israel Nelken and Jonathan Rubin, Shlomo Dubnov)

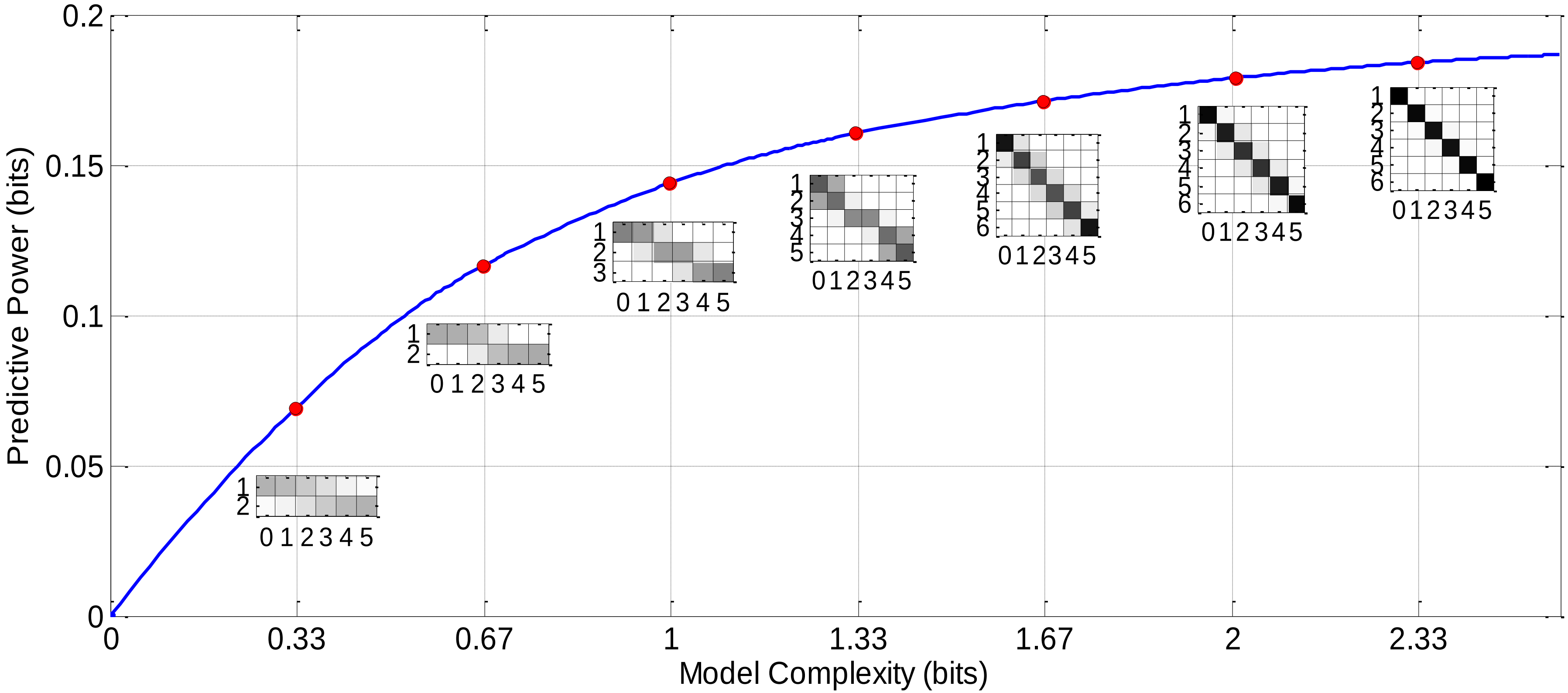
The predictive bottleneck



Perception-Prediction-Action Cycle

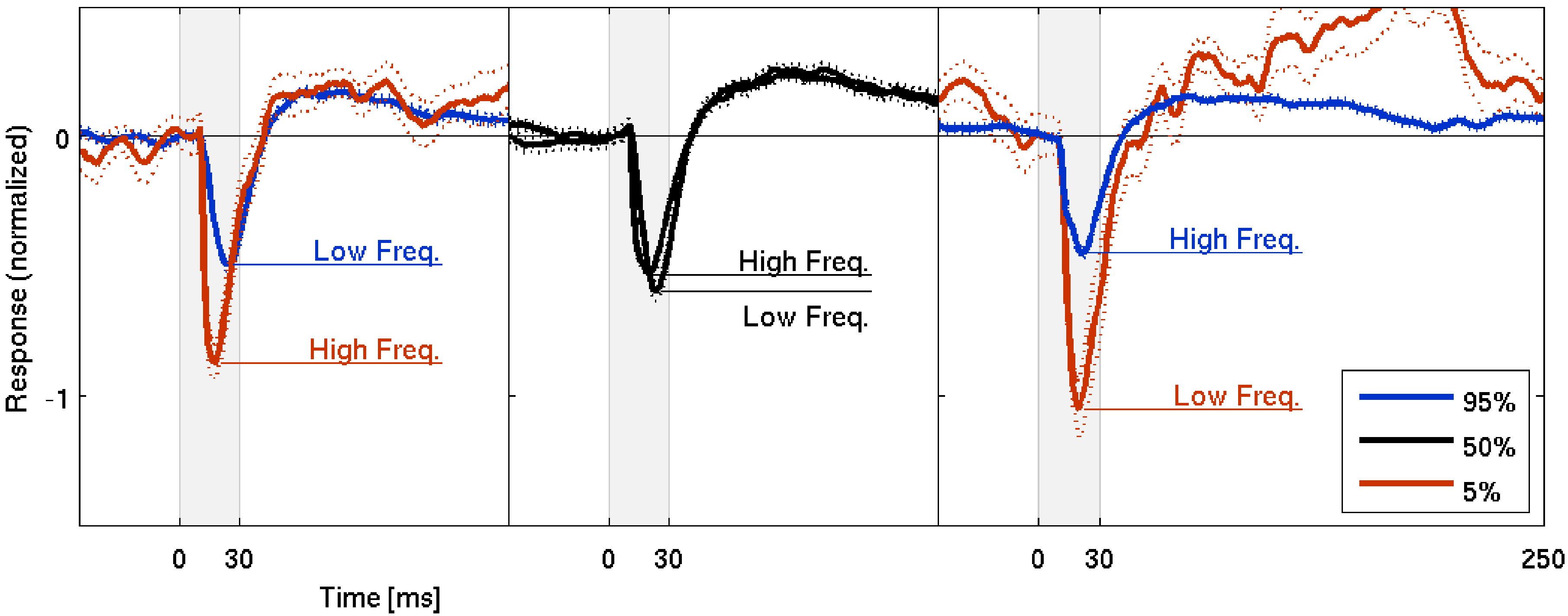


The Past-Future Information Bottleneck

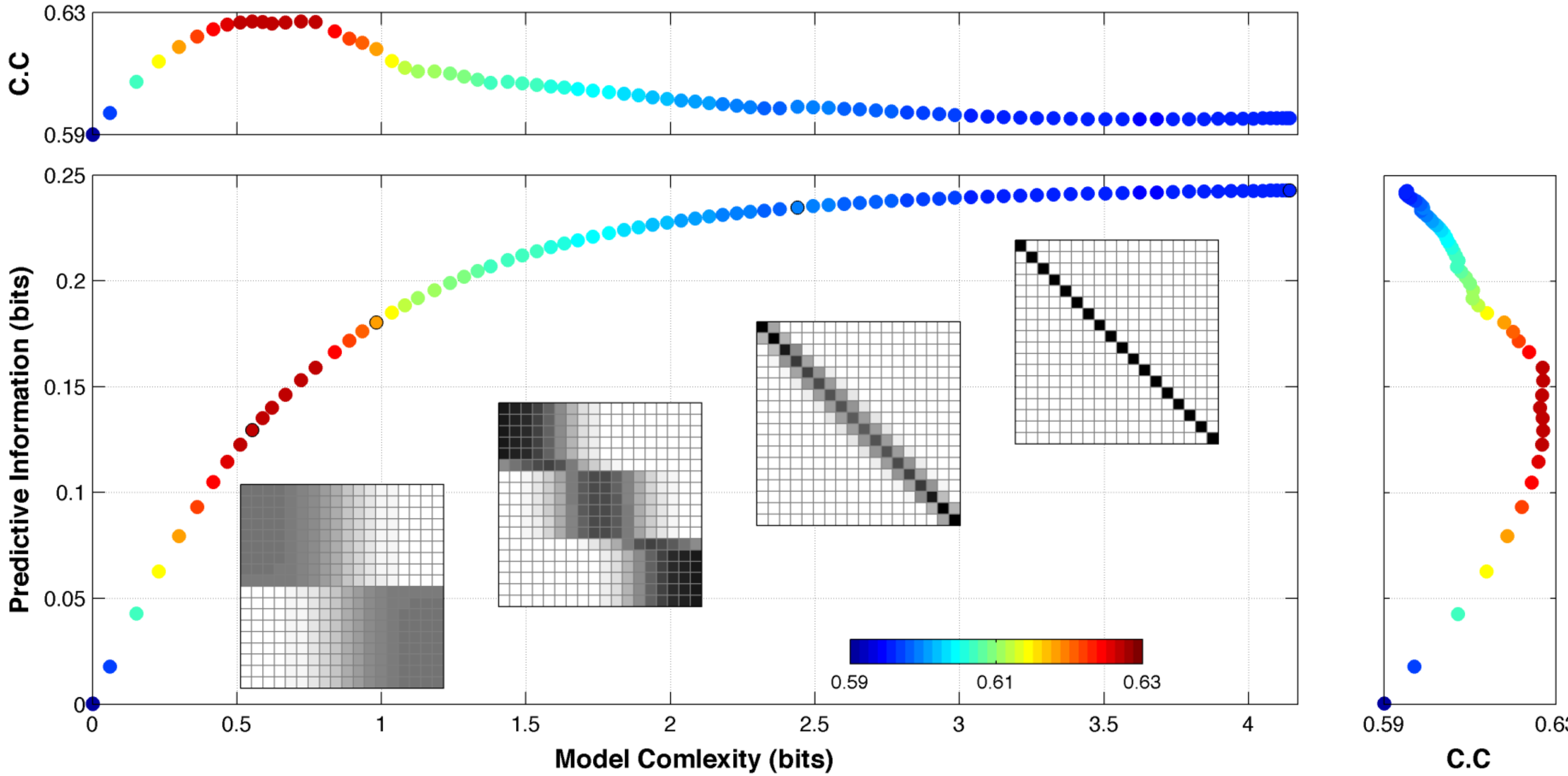


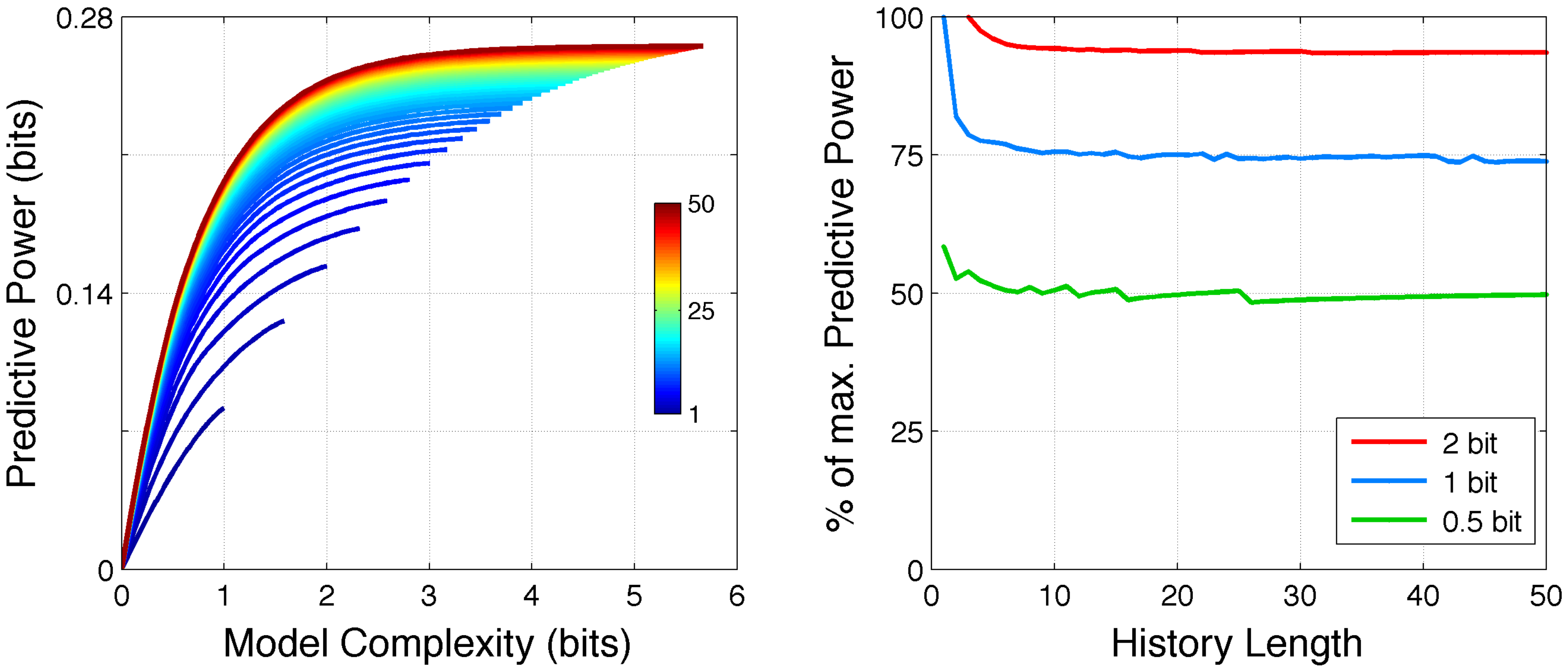
Information curve showing the optimal predictive information (surprise) as a function of the complexity of the internal model (memory bits) for the next-tone prediction of oddball sequences using a memory duration of 5 tones back.

The physiological surprise

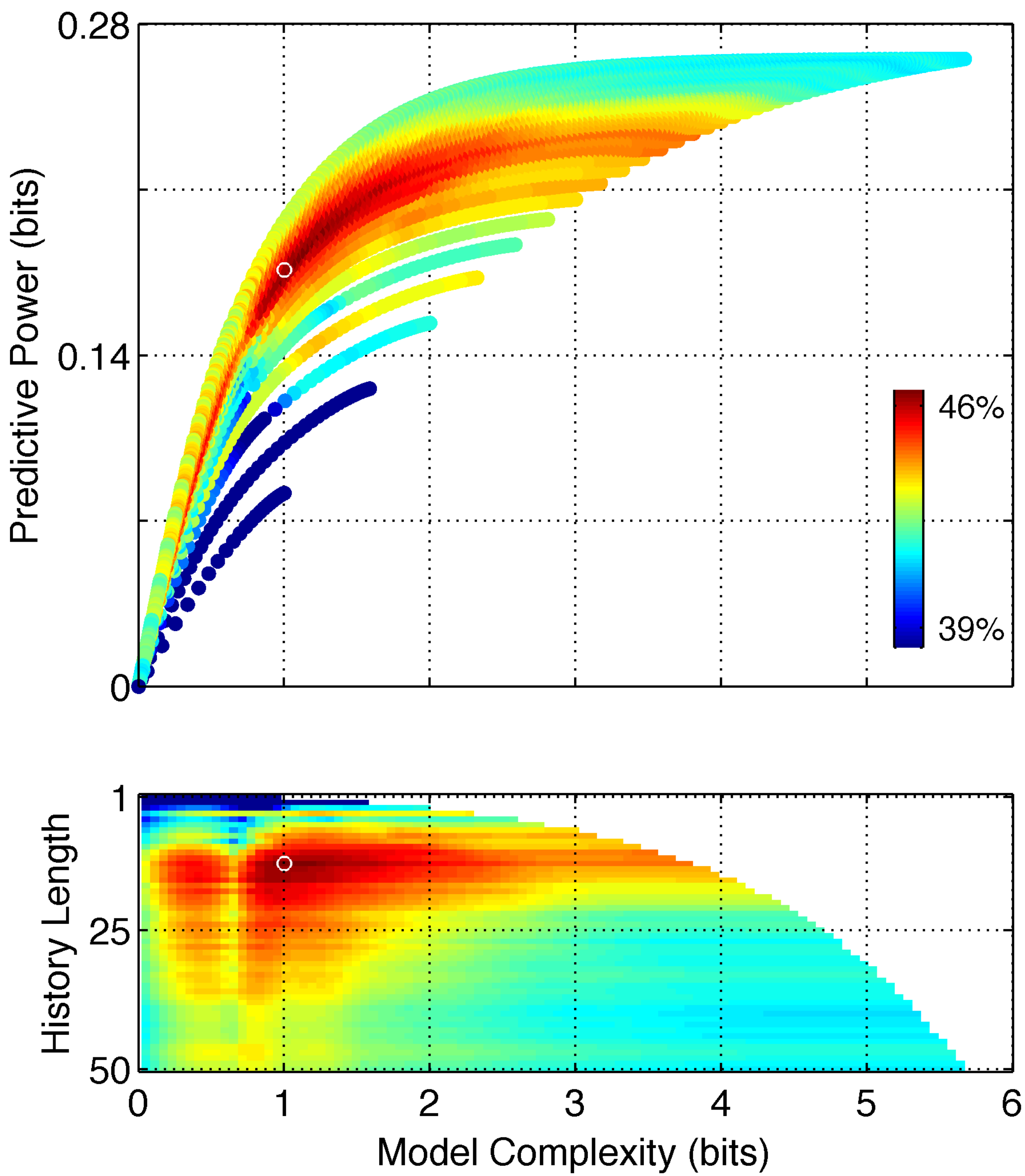


Quantifying the complexity of neural representations

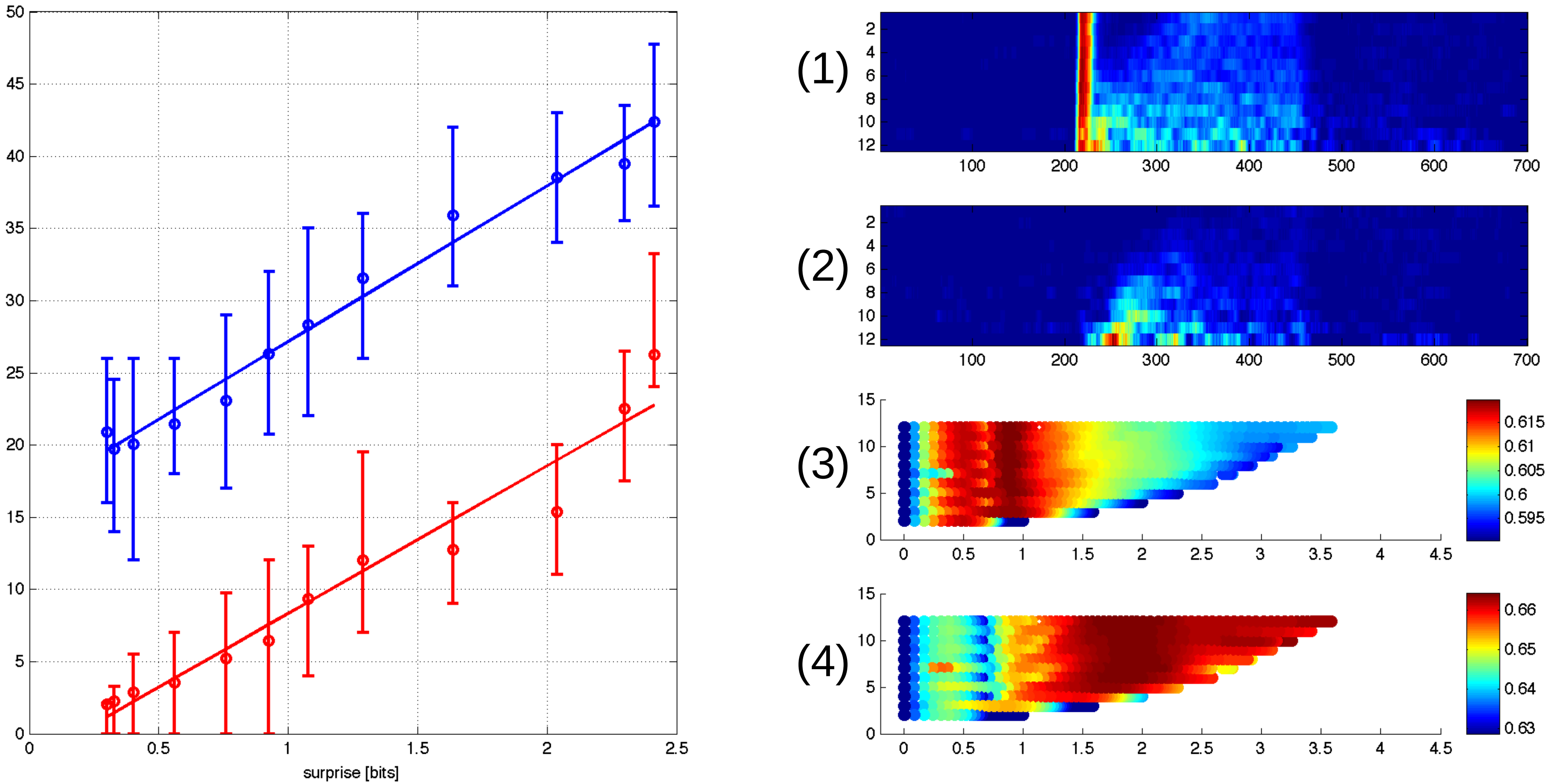




0.28
0.14
0
0
1
2
3
4
5
6
Predictive Power (bits)
Model Complexity (bits)
50
25
1
100
75
50
25
0
0
10
20
30
40
50
% of max. Predictive Power
History Length
2 bit
1 bit
0.5 bit

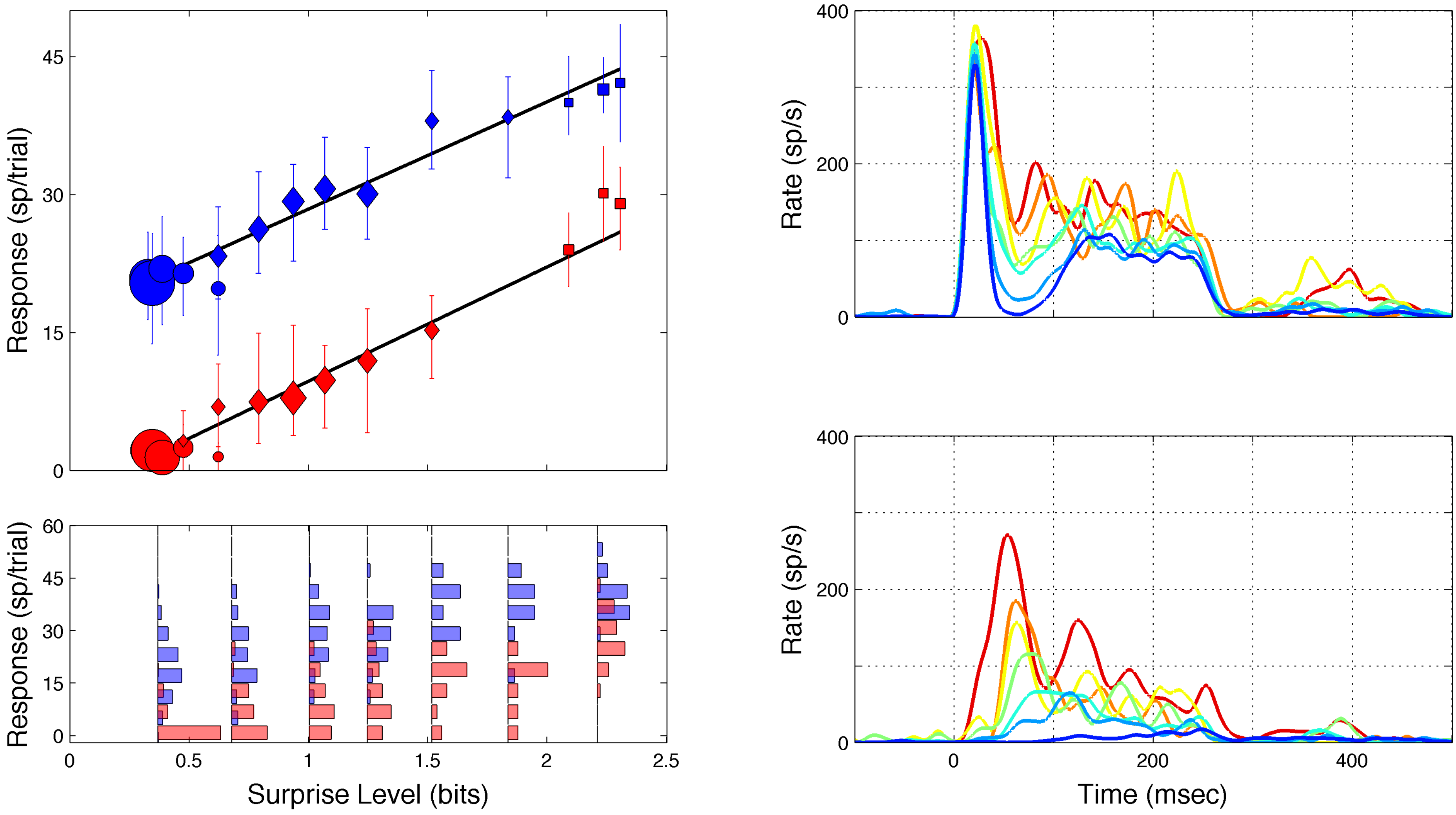


Predictive Power (bits)
0.28
0.14
0
46%
39%
History Length
1
25
50
0
1
2
3
4
5
6
Model Complexity (bits)



Left: scatter plots of the neural responses to either ‘A’ (blue) or ‘B’ (red) and the surprise values calculated for a specific model. Dots mark the mean response at a given surprise level, and the error-bars represent 25 and 75 percentile of the data. Right: (1) PSTH for stimulus ‘A’, each row is the averaged PSTH corresponding to a single point in the scatter-plot, sorted from low to high surprise level. (2) PSTH for stimulus ‘B’. (3) Correlations for ‘A’ (as explained before). (4) Correlations for ‘B’.

The PSTH plots help to see what part of signal is correlated with the surprise. For instance the onset seems pretty constant (and absent in the responses to ‘B’), where the sustained part seems to be very correlated with the surprise.



Cortical representation of (optimal) auditory surprise

Summary

- The Perception-Action Cycles have an intriguing analogy with Shannon's model of communication, which suggests asymptotic bounds on the optimal cycle's efficiency

- This model extends old results on optimal gambling to a much more general optimal value-cost tradeoff with long sensing-decision-action sequences

- Crucial quantities are the "environment's predictive capacity" and the "perception-action-capacity".

- While obviously still rudimentary, the model provides new ways for analyzing neuroscience data and new insights on motor control and deficiencies.

Many Thanks to…

- Bill Bialek
- Amir Globerson
- Ilya Nemenman
- Eli Nelken
- Jonathan Rubin
- Gal Chechik
- Shlomo Dubnov
- Ohad Shamir
- Naama Parush
- Felix Creutzig
- Roi Weiss