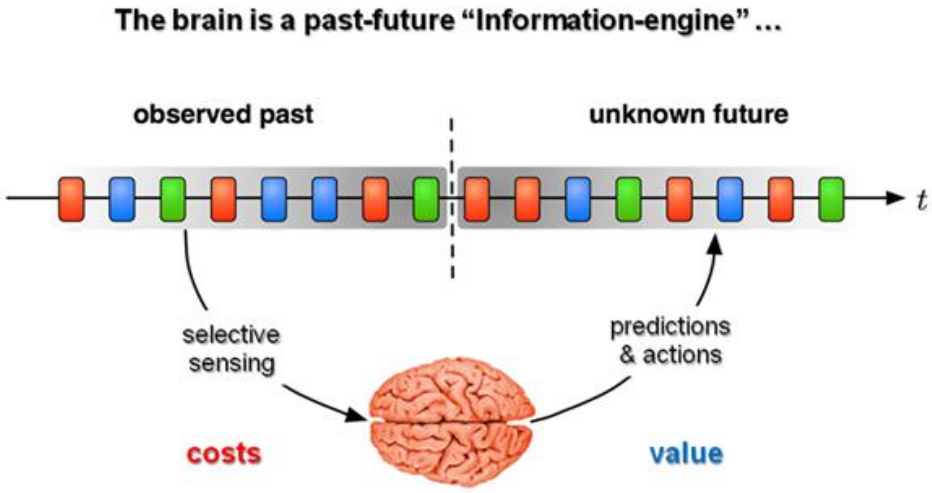
# On the balance of Information-flow and Information Discounting in I-RL

***Dagstuhl 2011***





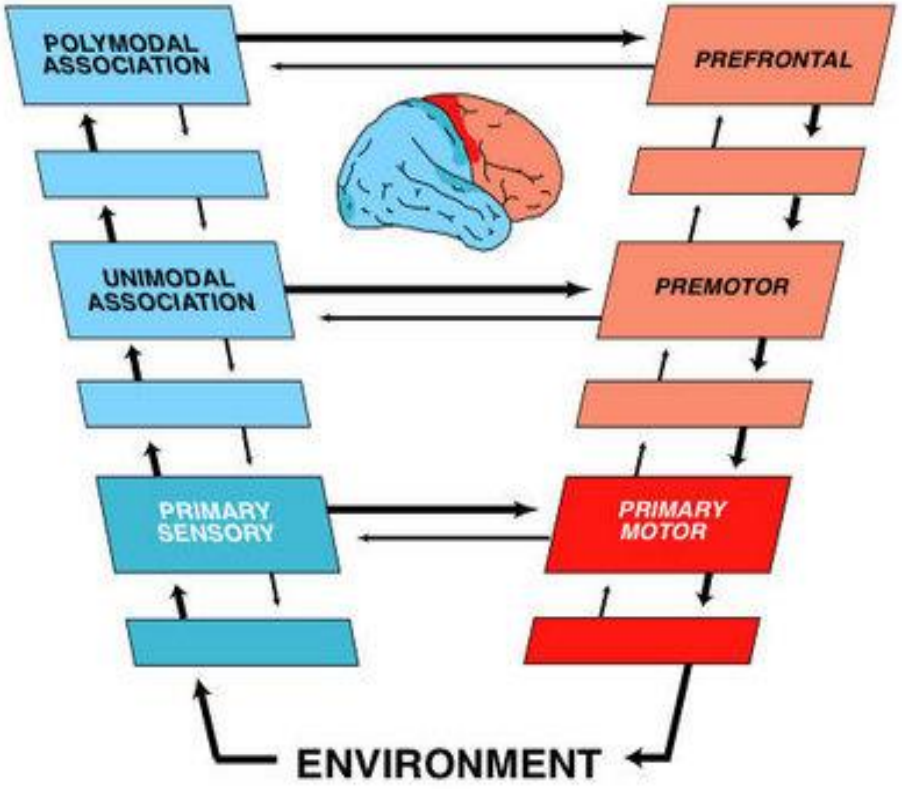
Naftali Tishby

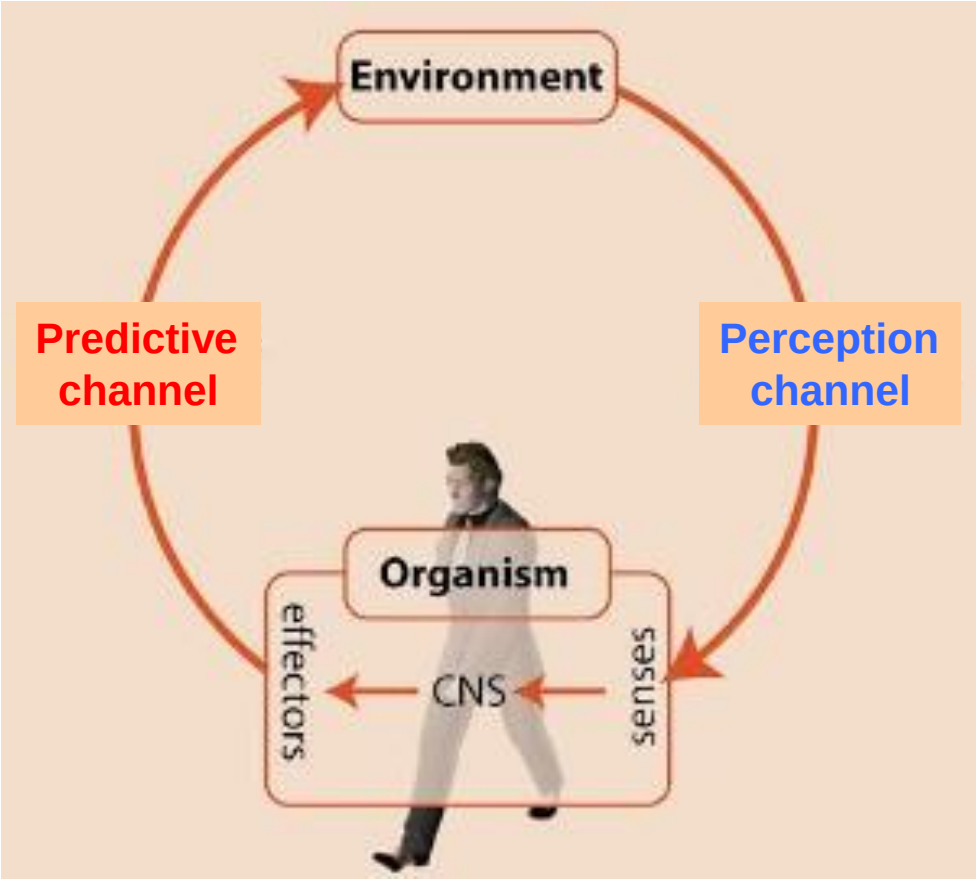
**Bill Bialek  Daniel Polani  Eli Nelken  Jonathan Rubin  Ohad Shamir**

**ELSC**
**Interdisciplinary Center for Neural Computation**
**School of Engineering and Computer Science**
**Hebrew University, Jerusalem, Israel**

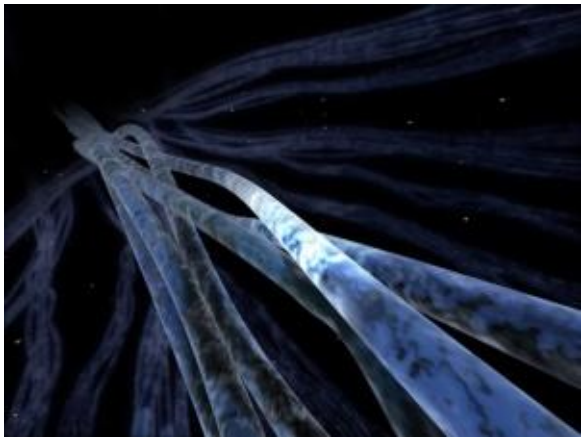
# The Perception-Action Cycle

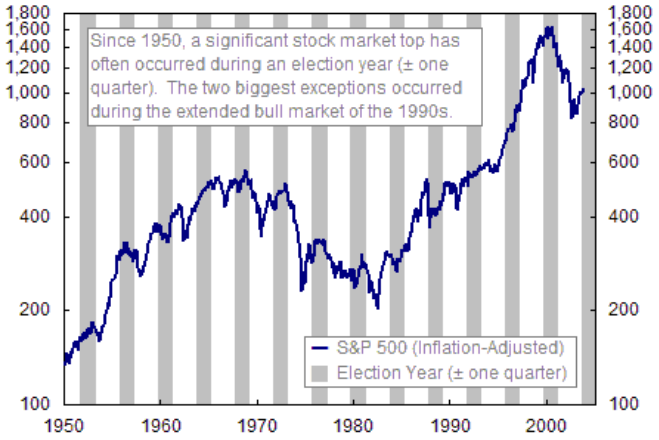
The **circular flow of information** that takes place **between the organism and its environment** in the course of a sensory-guided sequence of behavior towards a goal.
(JM Fuster)



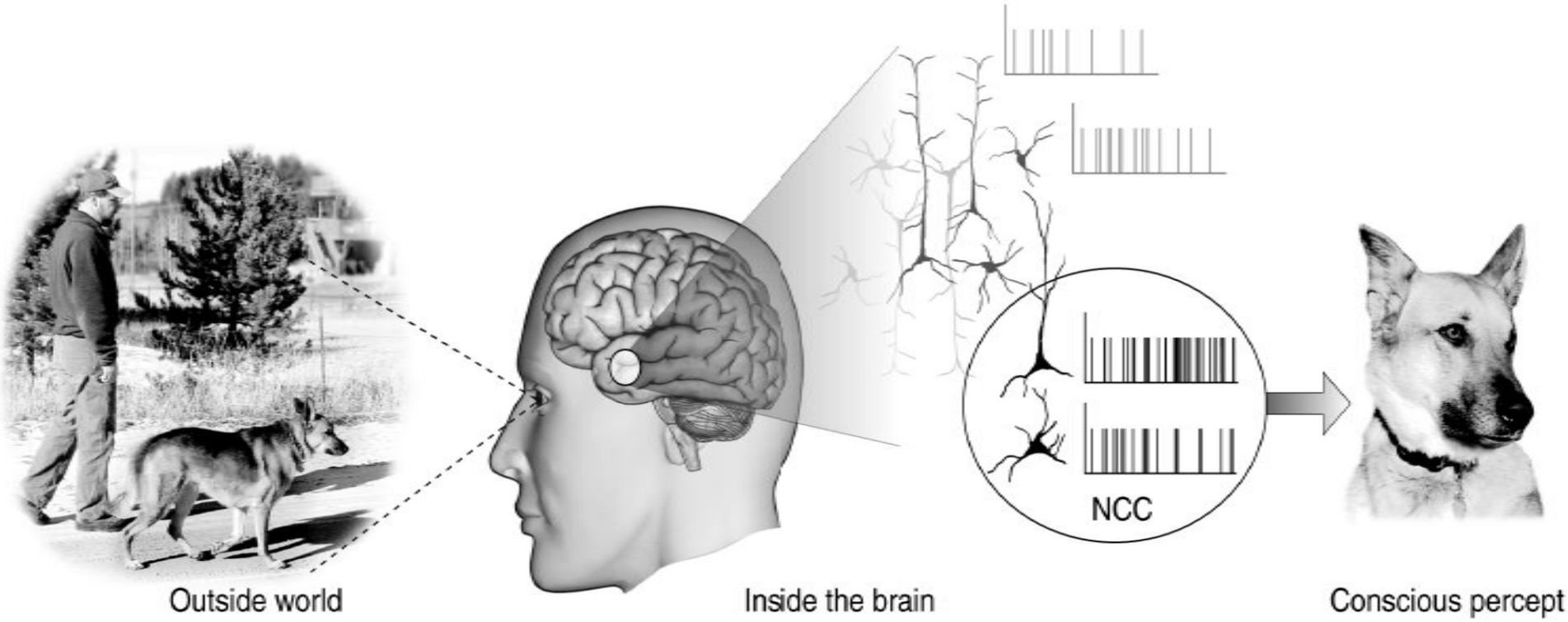
# Life is all about making valuable predictions…



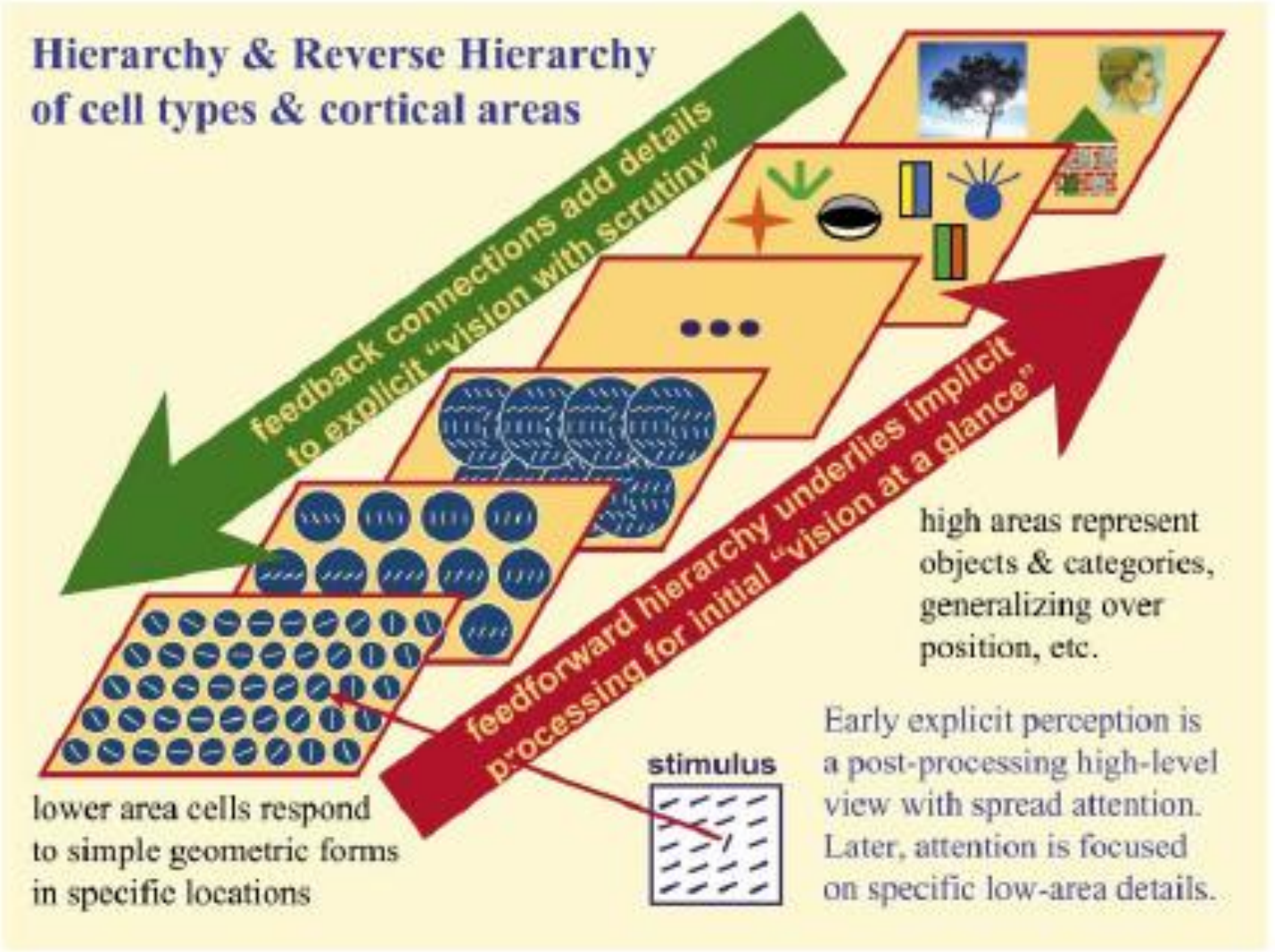

# The brain's primary task is making valuable predictions

**Perception** - goal oriented **hypotheses generation,** directed by **active predictions & useful decisions, tested** by [external or internal] **information gathering**



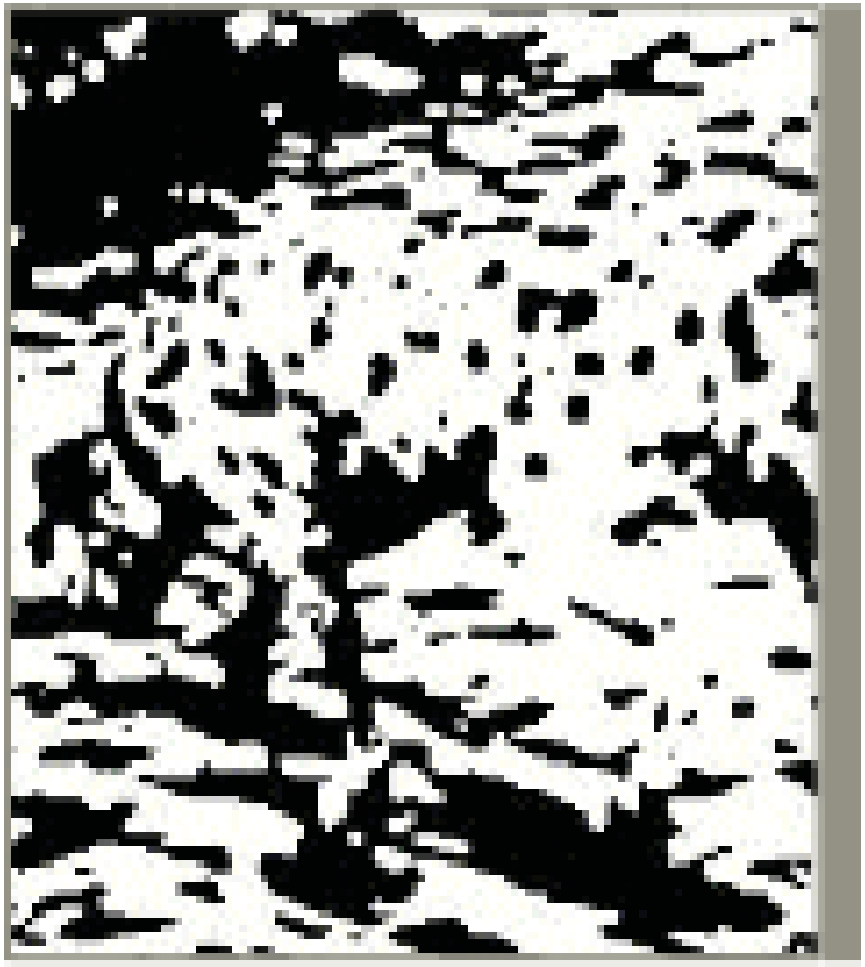

From C. Koch,

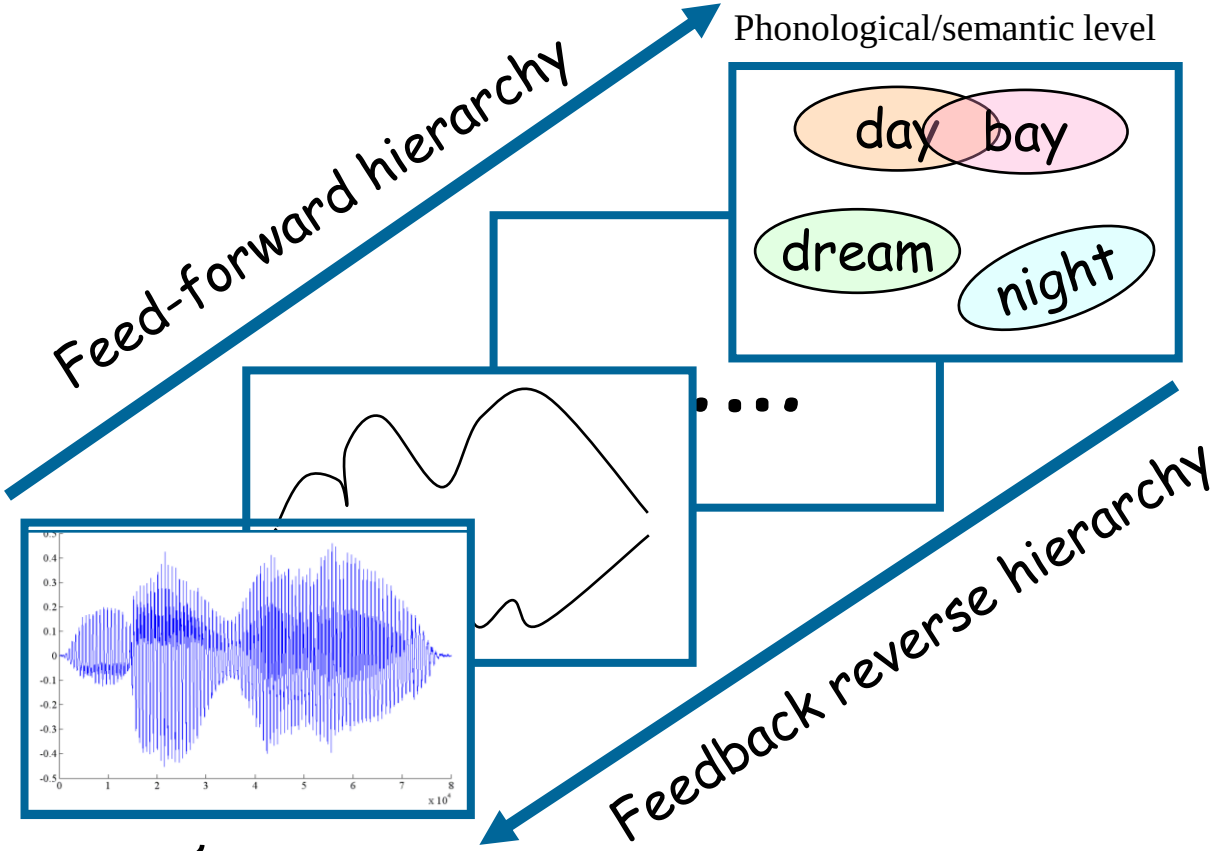
# Hierarchies and reverse hierarchies




Tsostos 1990; Hochstein and Ahissar 2002

# We see what we expect to see






Initial perception is based on high-level, phonological representations

Low level representations are sensitive to fine temporal cues, in a μs resolution

Nelken et al, 2005

# [Partially Observed] Markov Decision Processes

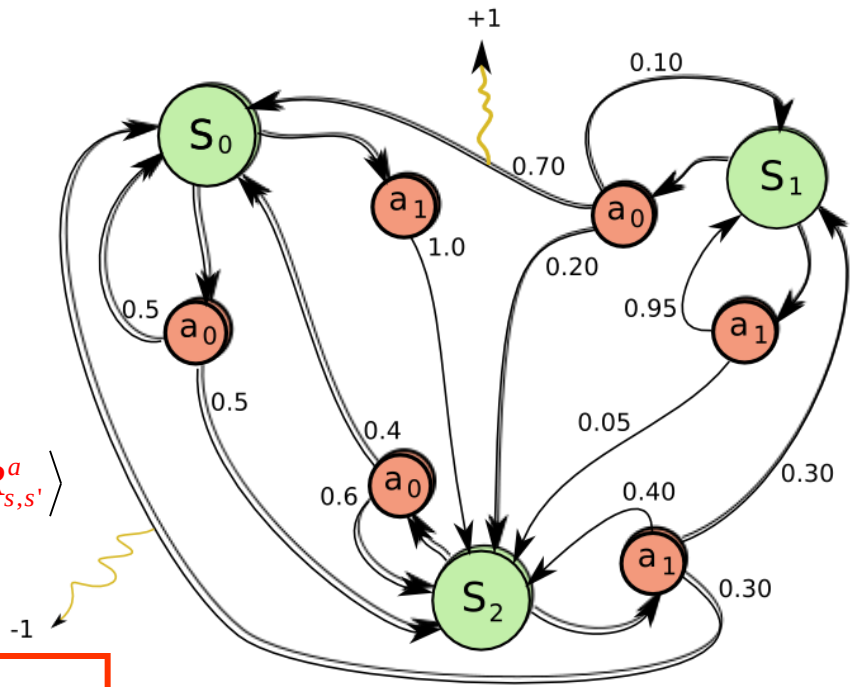
**An ergodic MDP :** four tuple of

$\langle s_t \in S, a_t \in A, P^a_{s,s'} = p(s' \mid s,a), R^a_{s,s'} \rangle$

states, actions, ergodic transition probabilities and rewards.

**An ergodic POMDP** also has stocahstic observations at each state (as in HMM)

$\langle s_t \in S, o_t \in O, \sigma(o \mid s), a_t \in A, P^a_{s,s'} = p(s' \mid s,a), R^a_{s,s'} \rangle$

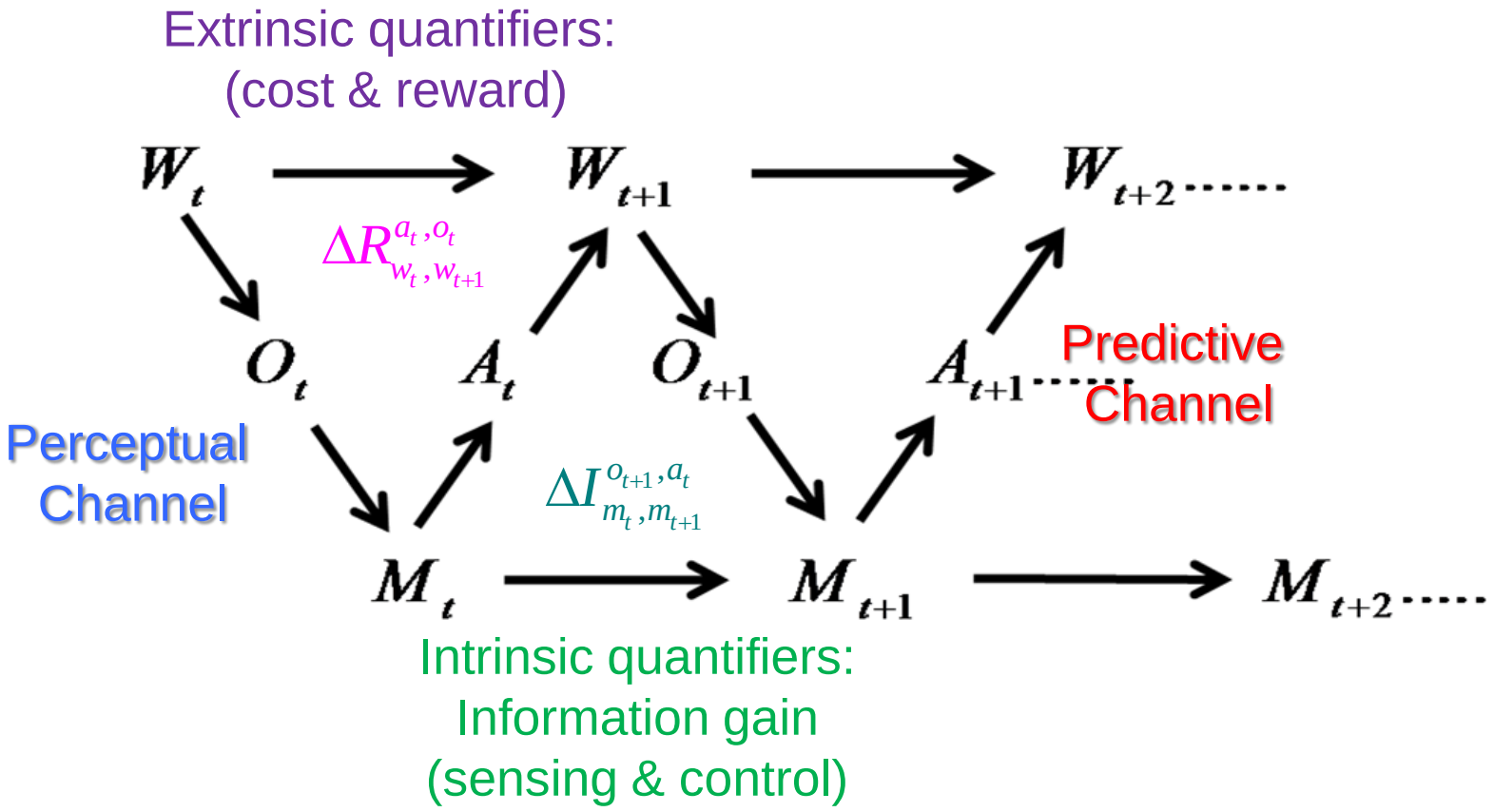



**The planning problem** :

Find the optimal [?] policy, $\pi(a \mid s)$ , that maximizes expected future reward: $E_{p(a_t, s_{t+1}, a_{t+1}, s_{t+2} \ldots \mid s_t)} \left[ \sum_{j=t}^{\infty} R^{a_j}_{s_j, s_{j+1}} \right]$

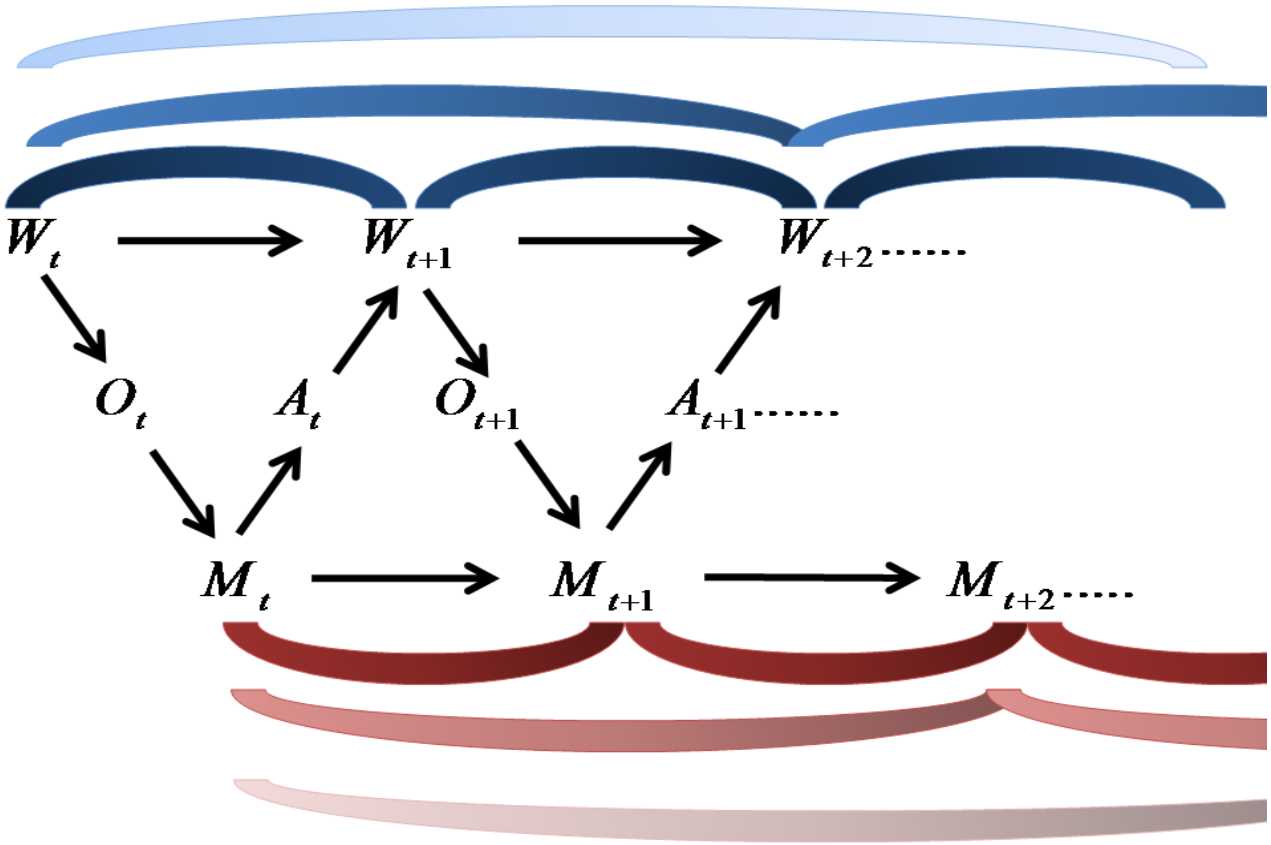
subject to various constraints.

# Graphical model for the perception-action-cycle




Both future extrinsic reward (Value) and intrinsic (Information to-go) are optimized together using Bellman-like equations w.r.t. to both channels.

# POMDPs, HMMs and Predictive Information



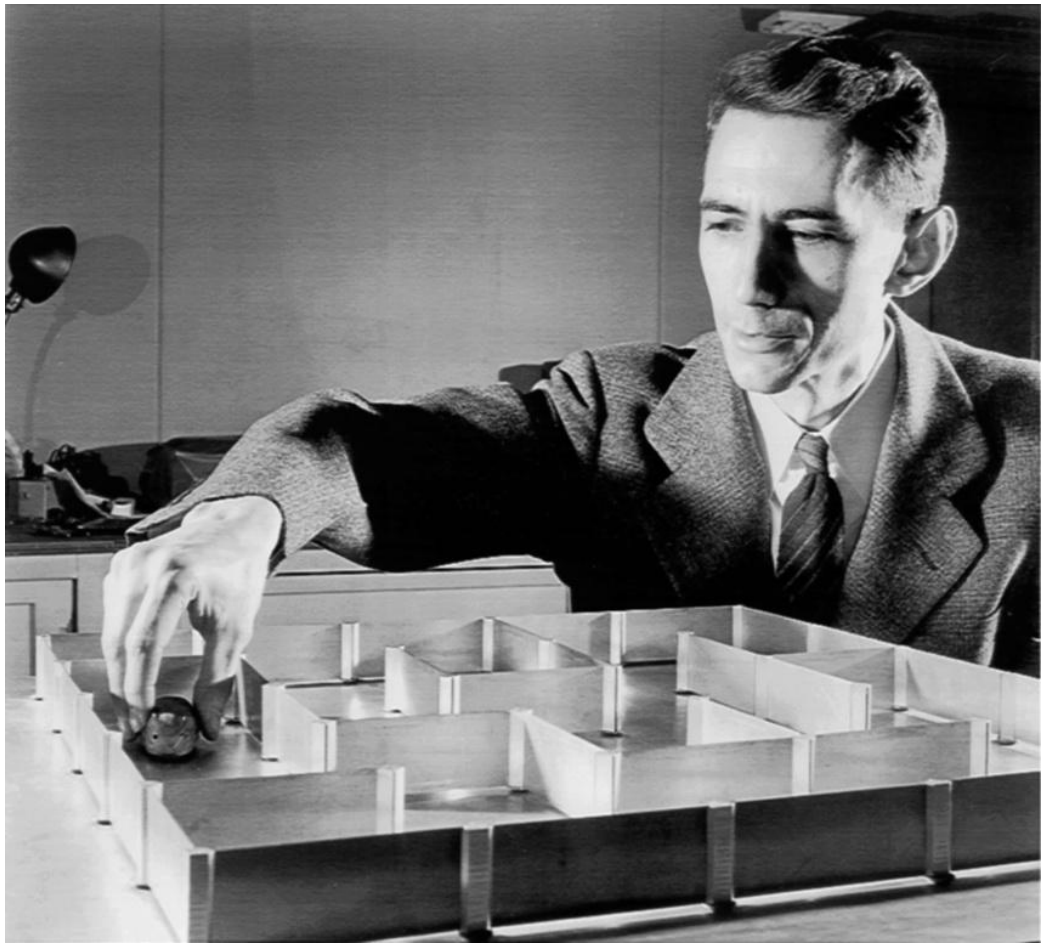

External time

Intrinsic
Information time

Top-down perceptual & planning hierarchy improves the information gain and the predictive horizon.
Correct discounting of past and future is a natural consequence.

# Bellman meets Shannon

**Richard Ernest Bellman**
(August 26, 1920 – March 19, 1984)



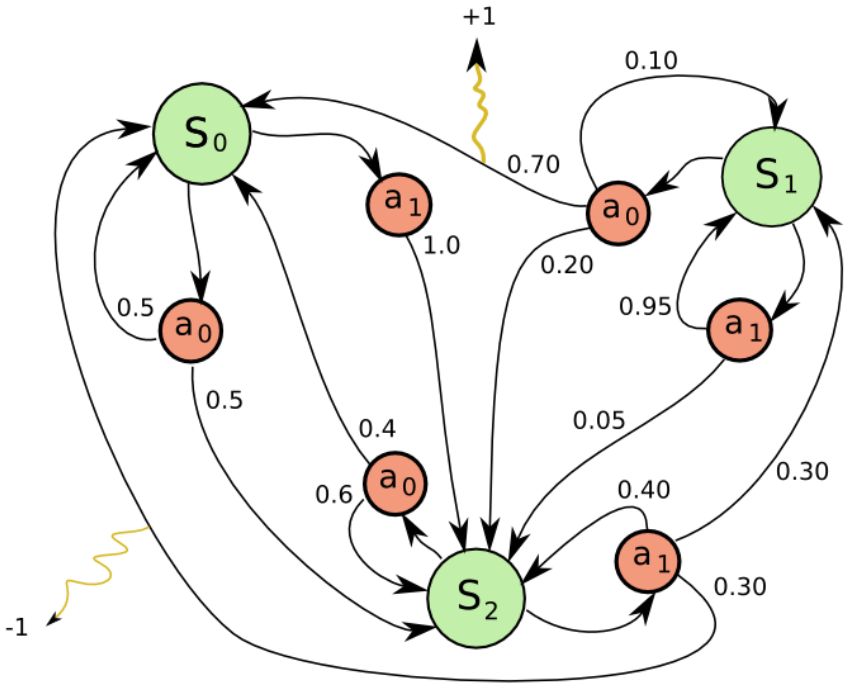
**Claude Elwood Shannon**
(April 30, 1916 – February 24, 2001)

# Decision-sequences and information

**Proposition 1 :**

state/decision sequences behave like codes:

- they can be concatenated to longer sequences
- sequence lengths obey a Kraft inequality
- number/complexity of (binary) decisions along a trajectory, is lower-bounded by the decision Entropy (as in source coding).




**Proposition 2 :**

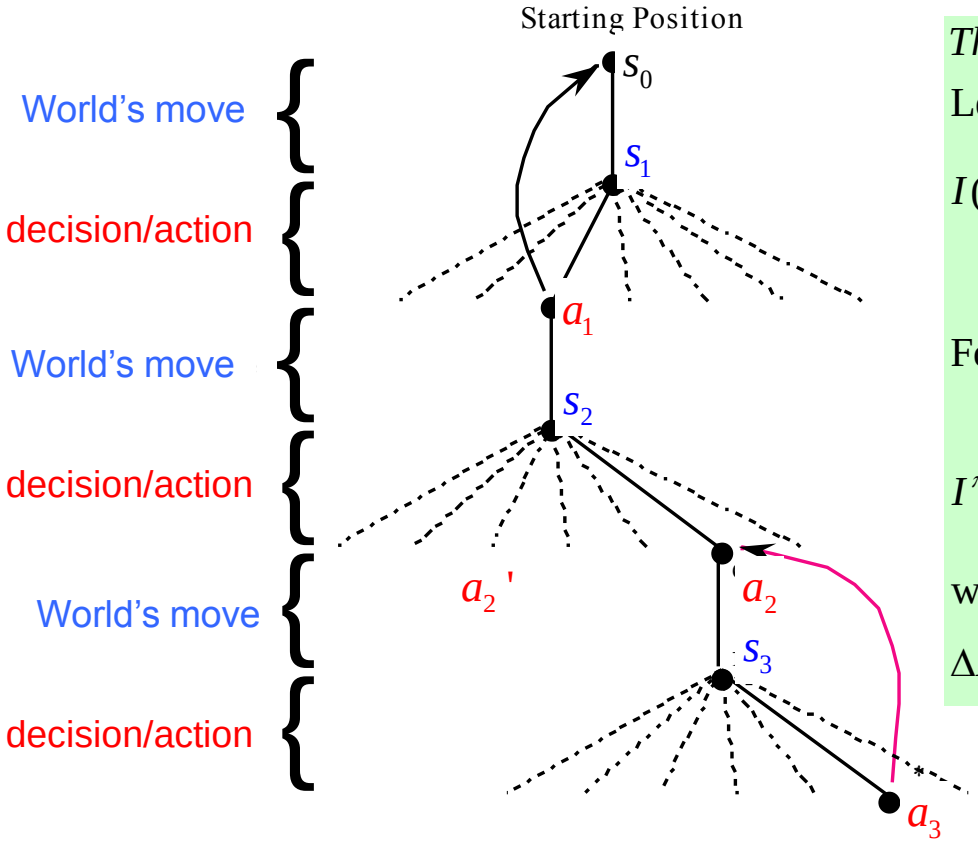
policies are like code nodes

subsequences of optimal decision sequences are also optimal - Bellman optimality condition.

**Proposition 3 :**

Entropy (and relative-entropy, KL-divergences) are the only (up to units) functionals of the state-decision sequence pdf that obey an additive Bellman-like equation on the simplex.

# Decision/action sequences and information




*Theorem*:

Let $G$ denote our target (relevant) variable

$$I(s;G) = E_g \log \frac{p(g \mid s)}{p(g)} = \sum_{g \in G} p(g,s) \log \frac{p(g \mid s)}{p(g)}$$

- the specific mutual information at state $s$ on $G$.

For an MDP $\langle S, A, P, \pi \rangle$ the following recursion holds:

$$I^{\pi}(s_t;G) = \sum_{a_t \in A} \pi(a_t \mid s_t) \sum_{s_{t+1}} P^{a_t}_{s_t,s_{t+1}} [\Delta I^{a_t}_{s_t,s_{t+1}} + I^{\pi}(s_{t+1};G)]$$

with

$$\Delta I^{a_t}_{s_t,s_{t+1}} \triangleq -I(G; s_{t+1} \mid s_t, a_t)$$

$s$ - the state before our action

$s'$ - the state after our action

# Proof idea: Recursive Information-chain rules…

The future actions-states in MDP sequence: $A_t, S_{t+1}A_{t+1}, S_{t+2}A_{t+2}, \ldots, S_{t+n}A_{t+n}, \ldots$

Let: $\mathcal{I}^{\pi}(s_t, a_t) \triangleq E_{p(s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots | s_t, a_t)} \log \frac{p(s_{t+1}a_{t+1}s_{t+2}a_{t+2}, \ldots | s_t, a_t)}{p(s_{t+1})\pi(a_{t+1})p(s_{t+2})\pi(a_{t+2}), \ldots}$

then:

$$\mathcal{I}^{\pi}(s_t, a_t) = E_{p(s_{t+1}, a_{t+1} | s_t, a_t)} \left[ \log \frac{p(s_{t+1} | s_t, a_t)}{p(s_{t+1})} + \log \frac{\pi(a_{t+1} | s_{t+1})}{\pi(a_{t+1})} + \mathcal{I}^{\pi}(s_{t+1}, a_{t+1}) \right]$$

With: $p(s_{t+1}, a_{t+1} | s_t, a_t) = p(s_{t+1} | s_t, a_t)\pi(a_{t+1} | s_{t+1})$

and: $\Delta I^{a_t}_{s_t, s_{t+1}} \triangleq \log \frac{p(s_{t+1} | s_t, a_t)}{p(s_{t+1})} + \log \frac{\pi(a_{t+1} | s_{t+1})}{\pi(a_{t+1})}$

□

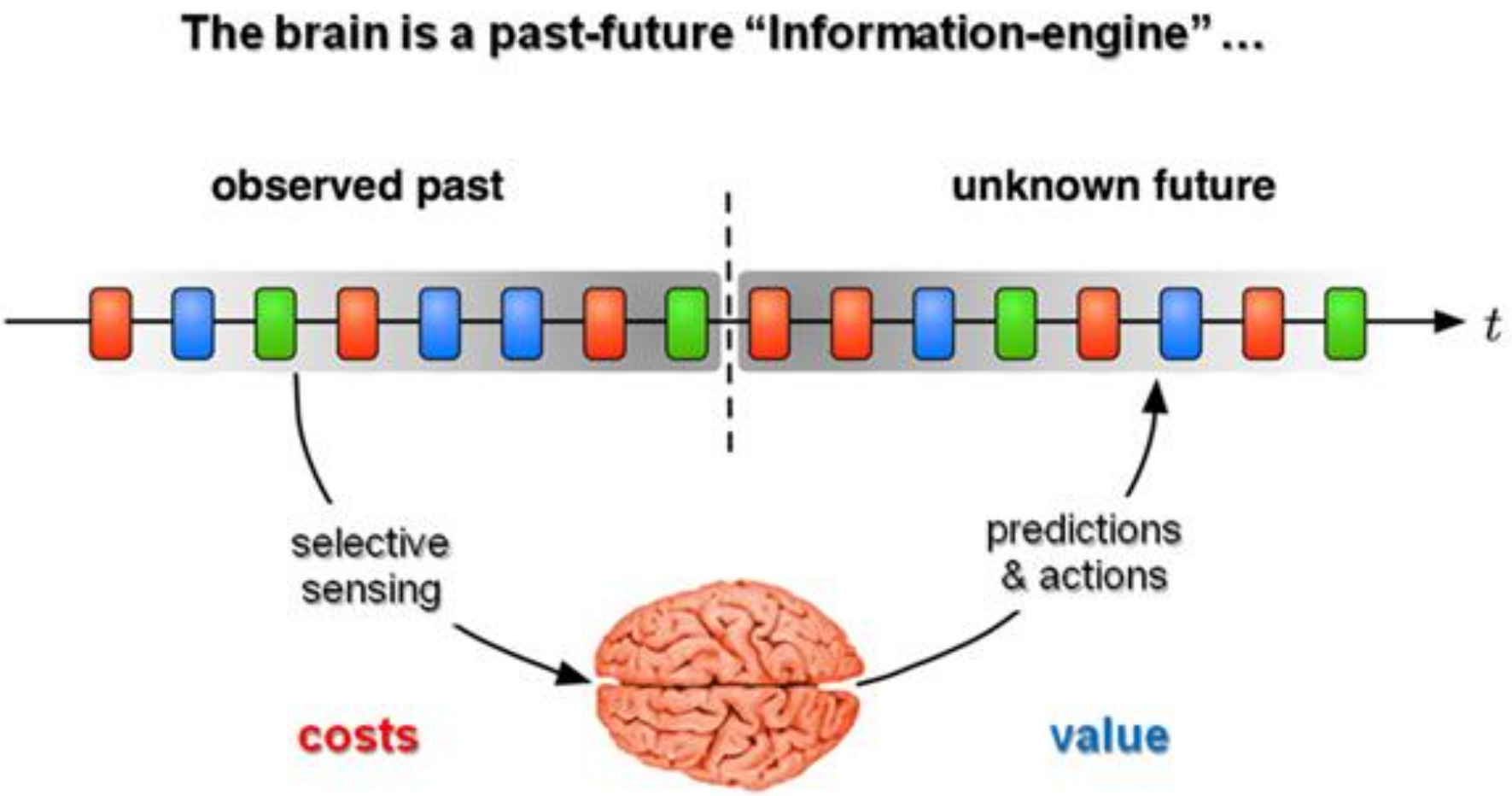
$\log \frac{p(s_{t+1} | s_t, a_t)}{p(s_{t+1})}$ - sensory information gain (perception capacity)

$\log \frac{\pi(a_{t+1} | s_{t+1})}{\pi(a_{t+1})}$ - required predictive capacity (control complexity)

**Predictive Information:**

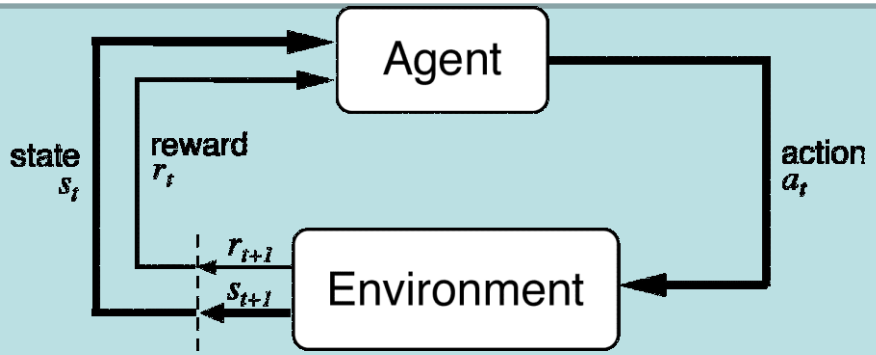
**Information in the past about the future**

# The essence of the cycle



# How much information is needed for valuable behavior?

# Value (extrinsic) and Information (intrinsic)…




Agent and environment interact at time steps: $t = 0,1,2,\ldots$

Agent observes TRUE state at step $t$: $s_t \in S$

produces action at step $t$: $a_t \in A(s_t)$ with $\pi(a \mid s)$

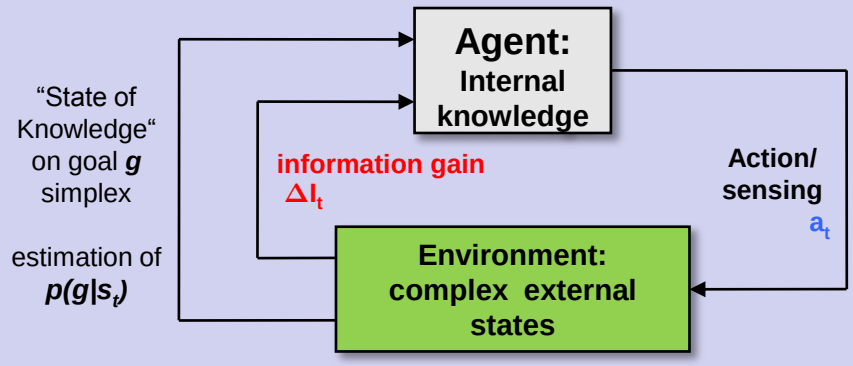
gets resulting reward: $r_{t+1} \in \Re$

resulting next state: $s_{t+1}$ with $p(s_{t+1} \mid s_t, a_t) = P^a_{s,s'}$

**Bellman equation for value $V^\pi$:**

$$V^\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} P^a_{ss'} \left[ R^a_{ss'} + \gamma V^\pi(s') \right]$$

solved for $V^\pi(s)$ by DP given $P^a_{ss'}, R^a_{ss'}$ and $\pi$




Agent has a goal variable $g \in G$

interacts with environment at time steps: $t = 0,1,2,\ldots$

estimates/infer an internal state: $\hat{s}_t \in \hat{S}$,

characterized by $p(\hat{s} \mid s)$, $p(g \mid \hat{s})$

produces action at step $t$: $a_t \in A(s_t)$ with $\pi(a \mid \hat{s})$

get/estimate information gain: $\Delta I^a_{s,s'} \in \Re$

resulting world next state: $s_{t+1}$ with $P^a_{s,s'}$

**Bellman equation for $I^\pi$:**

$$I^\pi(\hat{s}; g) = \sum_a \pi(a \mid \hat{s}) \sum_{s'} P^a_{ss'} \left[ \Delta I^a_{ss'} + I^\pi(\hat{s}'; g) \right]$$

solved for $I^\pi$ using DP and prob. inference

# Combining (future) Value and Information

In cases where information is free, we can maximize value irrespective of its information cost.
In general, however, we want

(1) to reduce decision complexity (get home in the simplest way)
(2) maximize the enviornment information gain (e.g. with the coins)
(3) increase robustness to model fluctuations

All three can be obtained by combining the Information and Value equations.

# Trading Value and (future) Information

$$\mathcal{I}^{\pi}(s_t,a_t) \triangleq E_{p(s_{t+1},a_{t+1},s_{t+2},a_{t+2},\ldots|s_t,a_t)} \log \frac{p(s_{t+1}a_{t+1}s_{t+1}a_{t+1},\ldots \mid s_t,a_t)}{p(s_{t+1})\pi(a_{t+1})p(s_{t+2})\pi(a_{t+2}),\ldots} \quad \text{then:}$$

$$\mathcal{I}^{\pi}(s_t,a_t) = E_{p(s_{t+1},a_{t+1}|s_t,a_t)}\left[\log \frac{p(s_{t+1}\mid s_t,a_t)}{p(s_{t+1})} + \log \frac{\pi(a_{t+1}\mid s_{t+1})}{\pi(a_{t+1})} + \mathcal{I}^{\pi}(s_{t+1},a_{t+1})\right]$$

With: $p(s_{t+1},a_{t+1}\mid s_t,a_t) = p(s_{t+1}\mid s_t,a_t)\pi(a_{t+1}\mid s_{t+1})$

We want: $\arg\min_{\pi}\ \mathcal{I}^{\pi}(s_t,a_t) - \beta Q^{\pi}(s_t,a_t) \triangleq \arg\min_{\pi} F^{\pi}(s_t,a_t,\beta)$

with: $Q^{\pi}(s_t,a_t) = \sum_{s_{t+1}} p(s_{t+1}\mid s_t,a_t)\left[R^{a_t}_{s_t s_{t+1}} + V^{\pi}(s_{t+1})\right]$

$$\mathcal{I}^{\pi}(s_t,a_t) - \beta Q^{\pi}(s_t,a_t) =$$

$$= E_{p(s_{t+1}|s_t,a_t)\pi(a_{t+1}|s_{t+1})}\left[\log \frac{p(s_{t+1}\mid s_t,a_t)}{p(s_{t+1})} + \log \frac{\pi(a_{t+1}\mid s_{t+1})}{\pi(a_{t+1})} - \beta R^{a_t}_{s_t s_{t+1}} + \mathcal{I}^{\pi}(s_{t+1},a_{t+1}) - \beta V^{\pi}(s_{t+1})\right]$$

$$= E_{p(s_{t+1}|s_t,a_t)\pi(a_{t+1}|s_{t+1})}\left[\log \frac{p(s_{t+1}\mid s_t,a_t)}{p(s_{t+1})} + \log \frac{\pi(a_{t+1}\mid s_{t+1})}{\pi(a_{t+1})} - \beta R^{a_t}_{s_t s_{t+1}} + \mathcal{I}^{\pi}(s_{t+1},a_{t+1}) - \beta Q^{\pi}(s_{t+1},a_{t+1})\right]$$

or

$$F^{\pi}(s_t,a_t,\beta) = E_{p(s_{t+1}|s_t,a_t)}\left[\log \frac{p(s_{t+1}\mid s_t,a_t)}{p(s_{t+1})} - \beta R^{a_t}_{s_t s_{t+1}} + E_{\pi(a_{t+1}|s_{t+1})}\left[\log \frac{\pi(a_{t+1}\mid s_{t+1})}{\pi(a_{t+1})} + F^{\pi}(s_{t+1},a_{t+1},\beta)\right]\right]$$

# Information bounded RL

define : $q(s'|s,a) = q^{a}_{s,s'} = \frac{p(s')}{Z(\beta,s,a)} \exp\left(\beta R^{a}_{s,s'}\right)$

the "optimal" (reward as sufficient statistic) transition probabilities,

and $p(s')$ the state prior,

$Z(\beta,s,a) = \sum_{s'} p(s') \exp\left(\beta R^{a}_{s,s'}\right)$, is the local partition function.

Then the state-action free energy, $F^{\pi}(s_t,a_t,\beta) \triangleq \mathfrak{I}^{\pi}(s_t,a_t) - \beta Q^{\pi}(s_t,a_t)$, Bellman equation is:

$$F^{\pi}(s_t,a_t,\beta) = \sum_{s_{t+1}} p(s_{t+1}|s_t,a_t) \sum_{a_{t+1}} \pi(a_{t+1}|s_{t+1}) \left[ \begin{array}{l} \log \frac{p(s_{t+1}|s_t,a_t)}{q(s_{t+1}|s_t,a_t)} + \log \frac{\pi(a_{t+1}|s_{t+1})}{\pi(a_{t+1})} \\ +F^{\pi}(s_{t+1},a_{t+1},\beta) - \log Z(\beta,s_t,a_t) \end{array} \right]$$

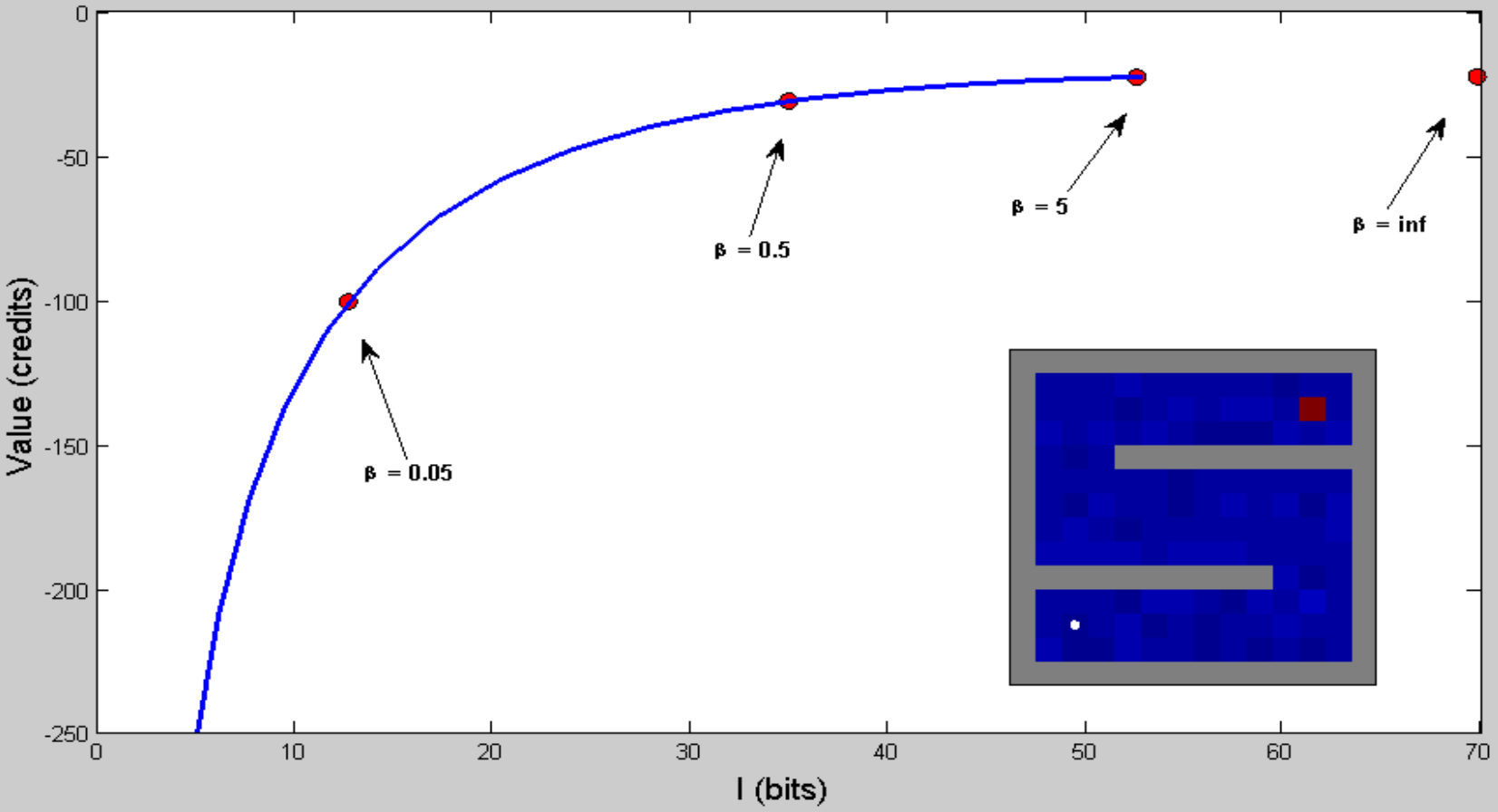
the desired optimal policy is (somewhat surprisingly):

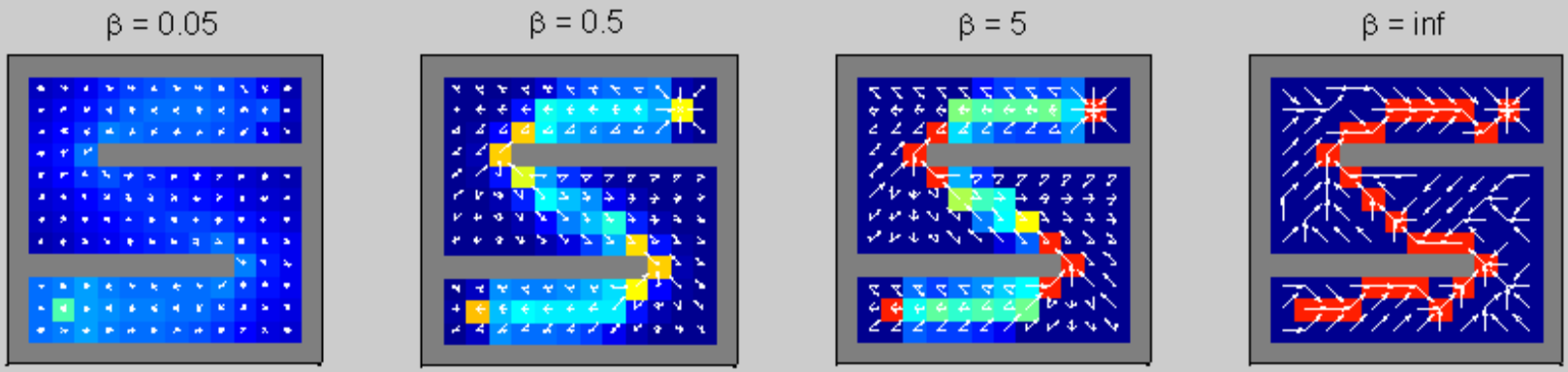
$$\pi(a|s) = \frac{\pi(a)}{Z(s,\beta)} \exp(F^{\pi}(s,a,\beta))$$

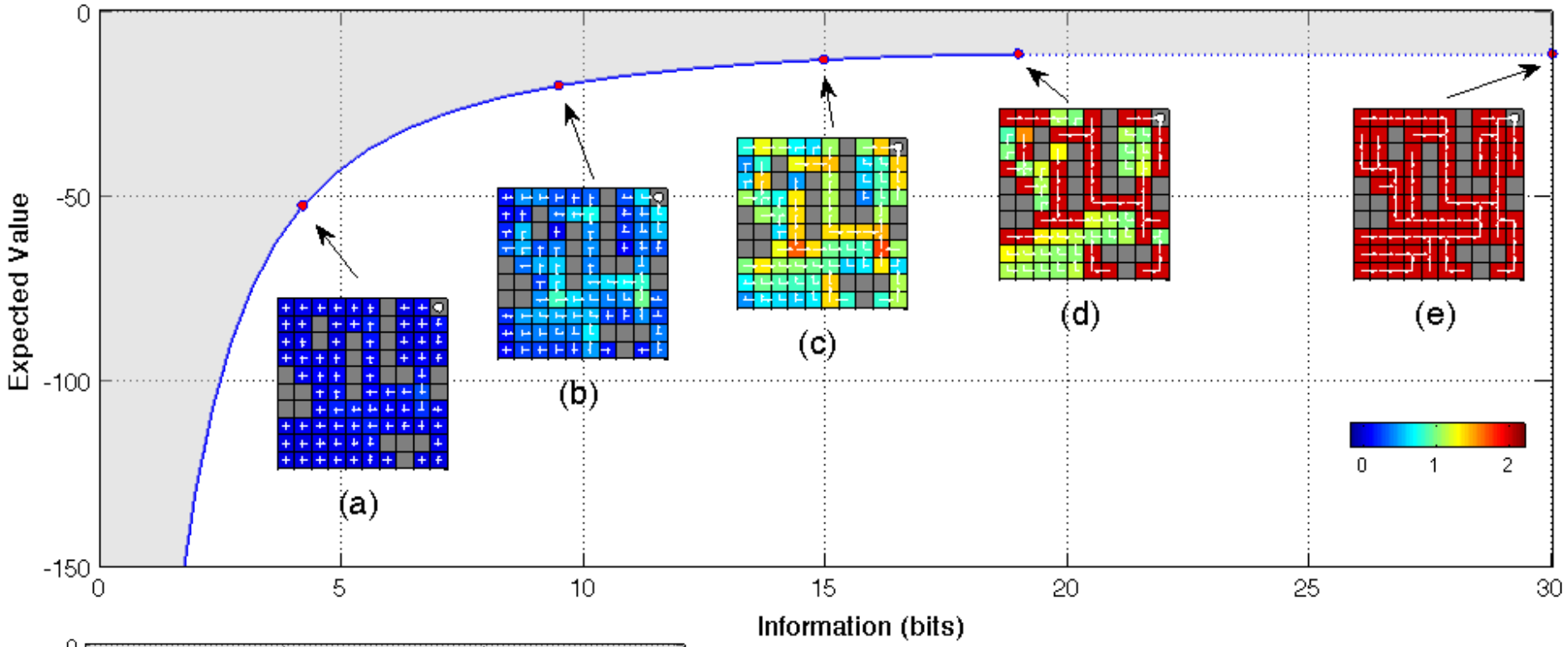
$$Z(s,\beta) = \sum_{a} \pi(a) \exp(F^{\pi}(s,a,\beta))$$

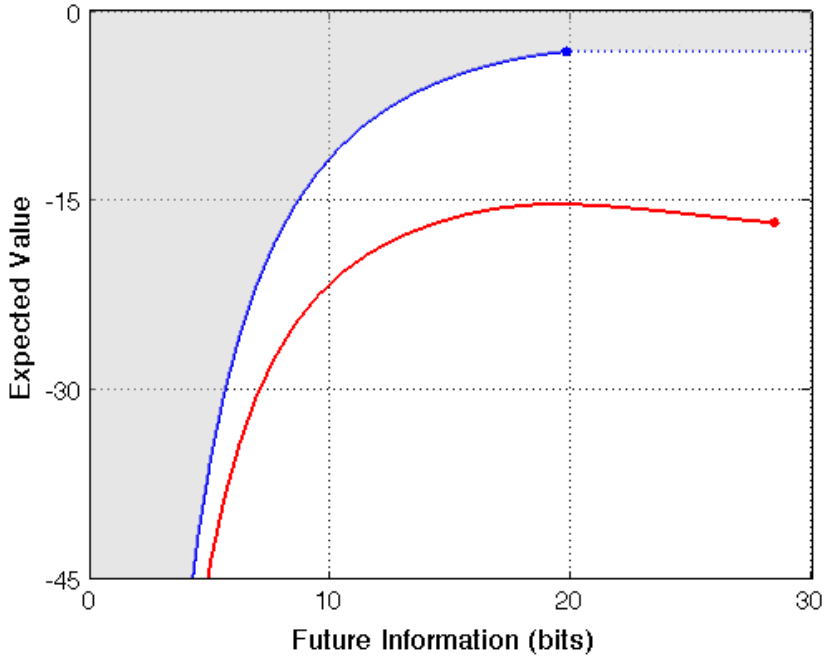
$$\pi(a) = \sum_{s} \pi(a|s) p(s)$$

These 3 equations should be iterated till convergence for every state (like Blahut Arimoto).


0
-50
-100
-150
-200
-250
0
10
20
30
40
50
60
70
Value (credits)
I (bits)
β = 0.05
β = 0.5
β = 5
β = inf


β = 0.05
β = 0.5
β = 5
β = inf



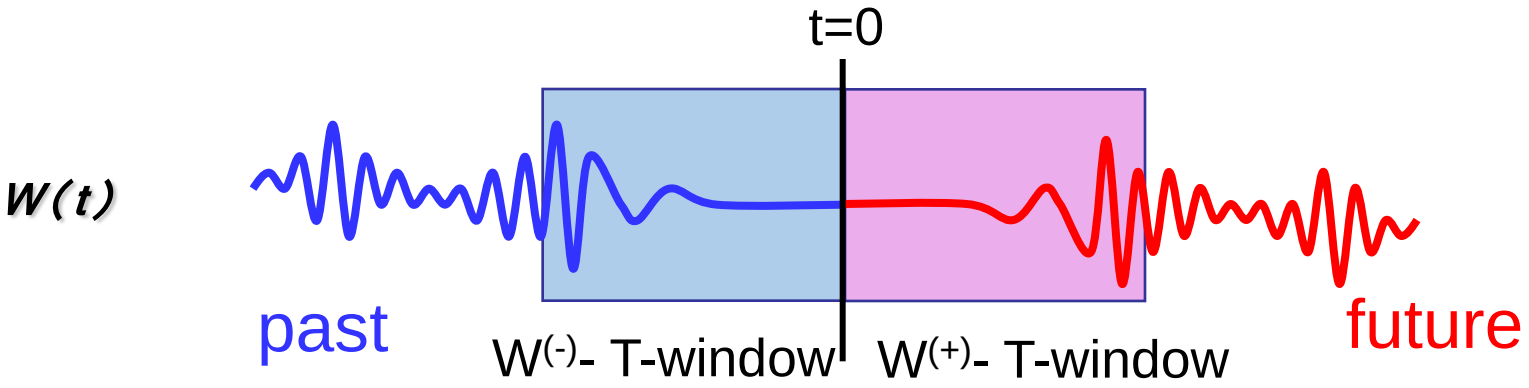




- The optimal tradeoff between Value and Future Information.
- Lower information → softer policy
- The optimal – out of training – value is obtained at a finite $\beta$

# How to discount Information (intrinsic) Rewards?

# 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

- Fast growth is a signature of infinite dimensional processes (e.g. speech)

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

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

# Information gain discounting

- Information gains should accumulate sub-linearly

  $\Rightarrow$ discounting by a factor $\gamma(t) = t^{-\eta}$ $, 0 \leq \eta \leq 1$

- $\eta = 1 \Rightarrow$ Logarithmic predictive information

  fixed intrinsic/extrinsic rewards ratio:

  $\Rightarrow$ exponential discounting of external rewards

  $\Rightarrow$ logarithmic "Intrinsic Time Fisheye"

- $\eta < 1 \Rightarrow$ Power-law predictive information

  $\Rightarrow$ hyperbolic discounting of external rewards

  $\Rightarrow$ hyperbolic intrinsic "Time Fisheye"


Time Fisheye
BEAUTY HAS A TAST
illy