Gradient Expectations: Predictive Neural Networks

Keith L. Downing

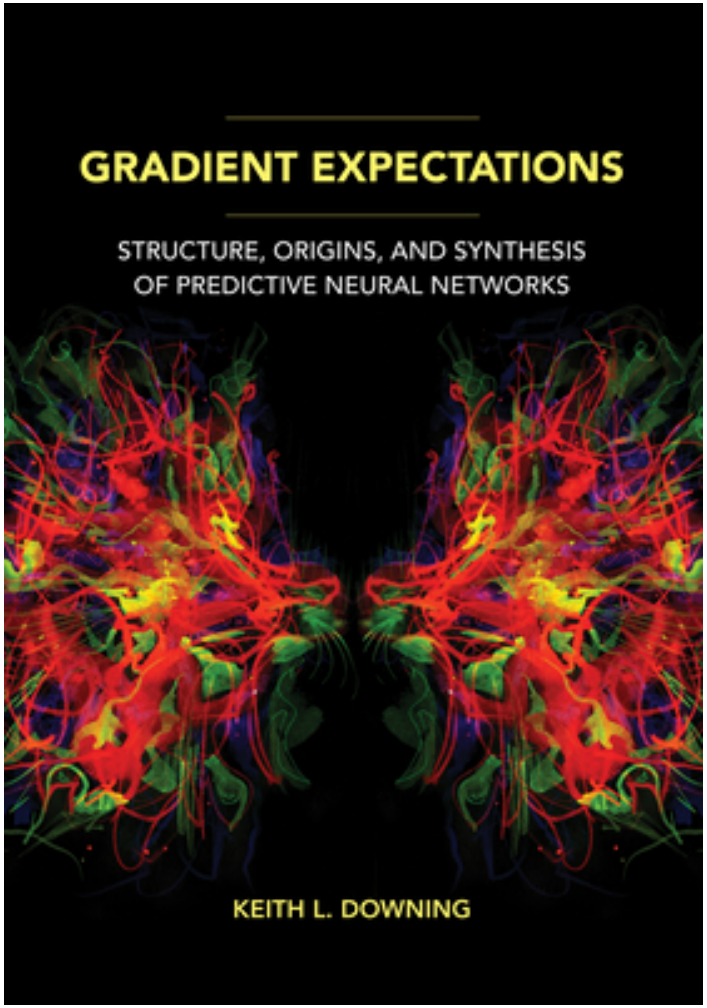
Department of Computer Science
The Norwegian University of Science and Technology (NTNU)
Trondheim, Norway
keithd@ntnu.no

June 3, 2024

Charting a path from early artificial neural networks to the contemporary vision of the predictive brain, with rich forays into biology and evolution, this book explains the buzz about brains as engines of prediction.
(Andy Clark, University of Sussex)

MIT Press
2023

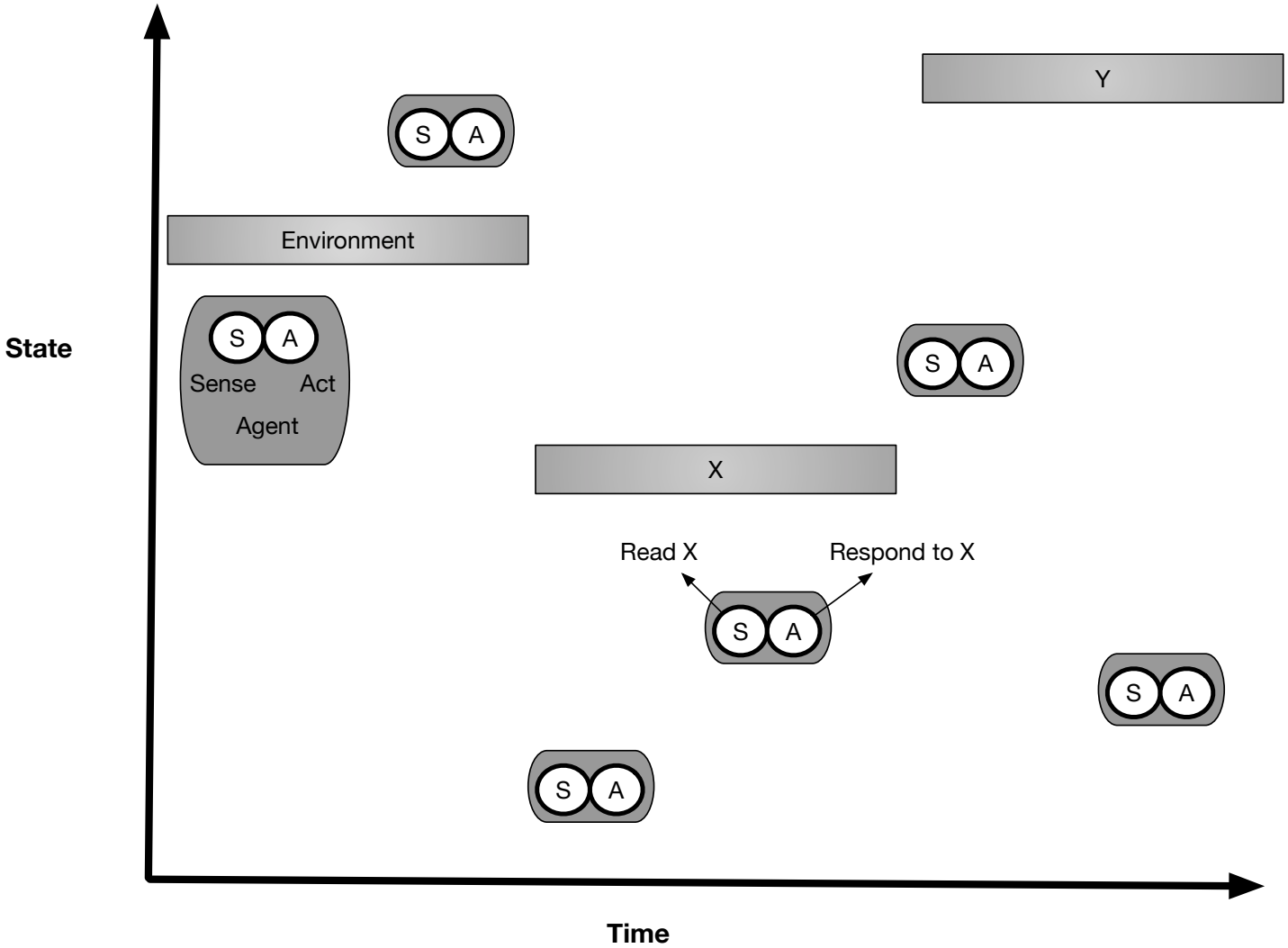
Downing's reach is omnidirectional. He connects the roots and new growth of deep learning with math, neuroscience, and evolutionary biology, ethology and computer science to show how intelligence emerged in animals and is emerging in machines.
(Josh Bongard, University of Vermont)



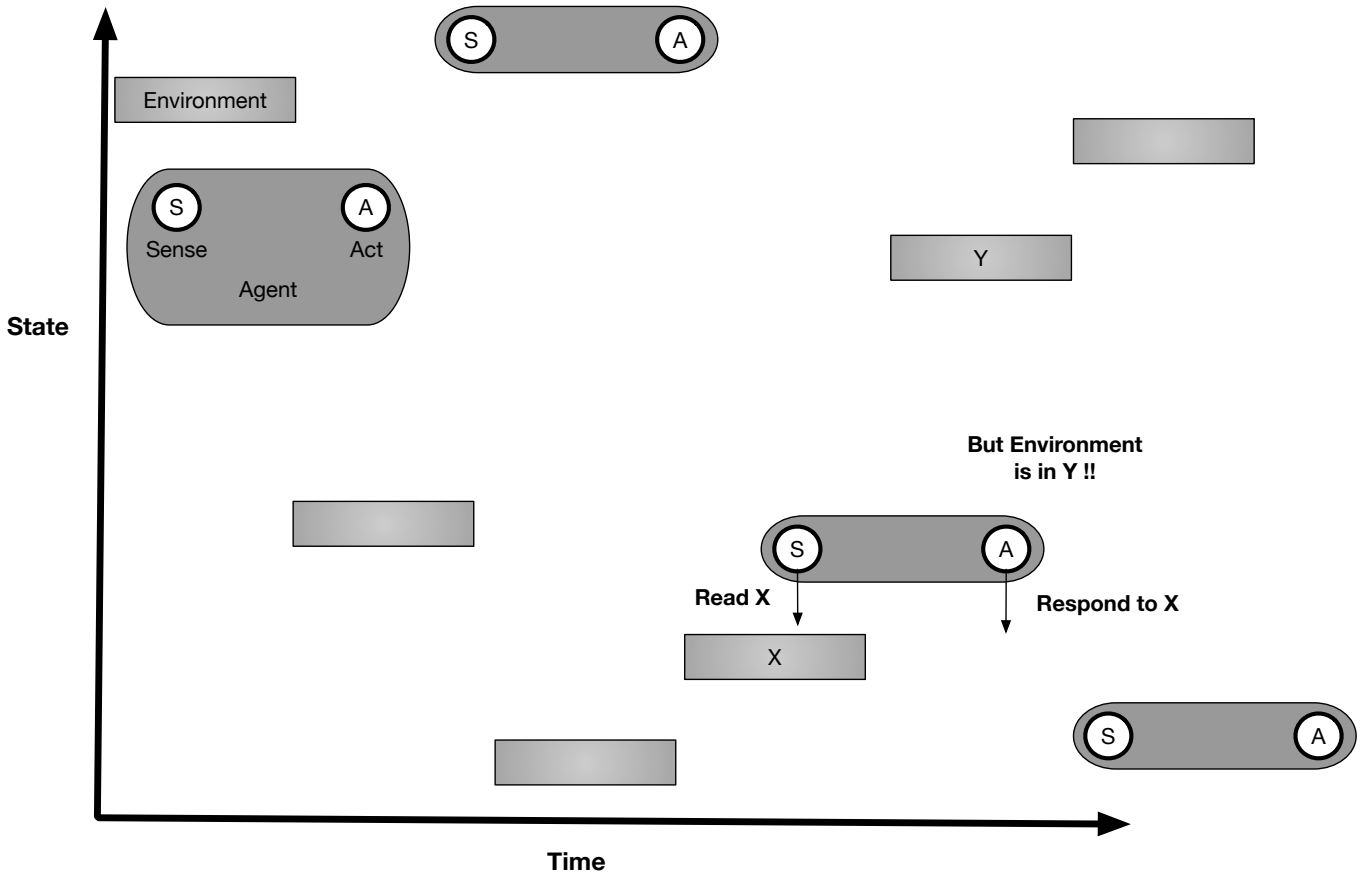
The Brain = A Prediction Machine ??

- The short punch line of this book is that brains are **foretelling devices**, and their **predictive** powers emerge from the various rhythms they perpetually generate...*Rhythms of the Brain* (Buzsaki, 2006)
- The capacity to **predict** the outcome of future events – critical to successful movement – is likely, the ultimate and most common of all global brain functions...*i of the Vortex* (Llinas, 2001)
- The mystery is, and remains, how mere matter manages to give rise to thinking, imagining, dreaming, and the whole smorgasboard of mentality, emotion and intelligent action....But there is an emerging clue...The clue can be summed up in a single word: **prediction**. To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become **masters of prediction**...*Surfing Uncertainty* (Clark, 2016)
- ...the core task of all brains... is to regulate the organism's internal milieu – by responding to needs and, **better still, by anticipating needs and preparing to satisfy them before they arise**... "Anticipatory regulation" replaces the more familiar "homeostatic regulation"...*Principles of Neural Design* (Sterling & Laughlin, 2015)

Sensing and Acting in a Slow World

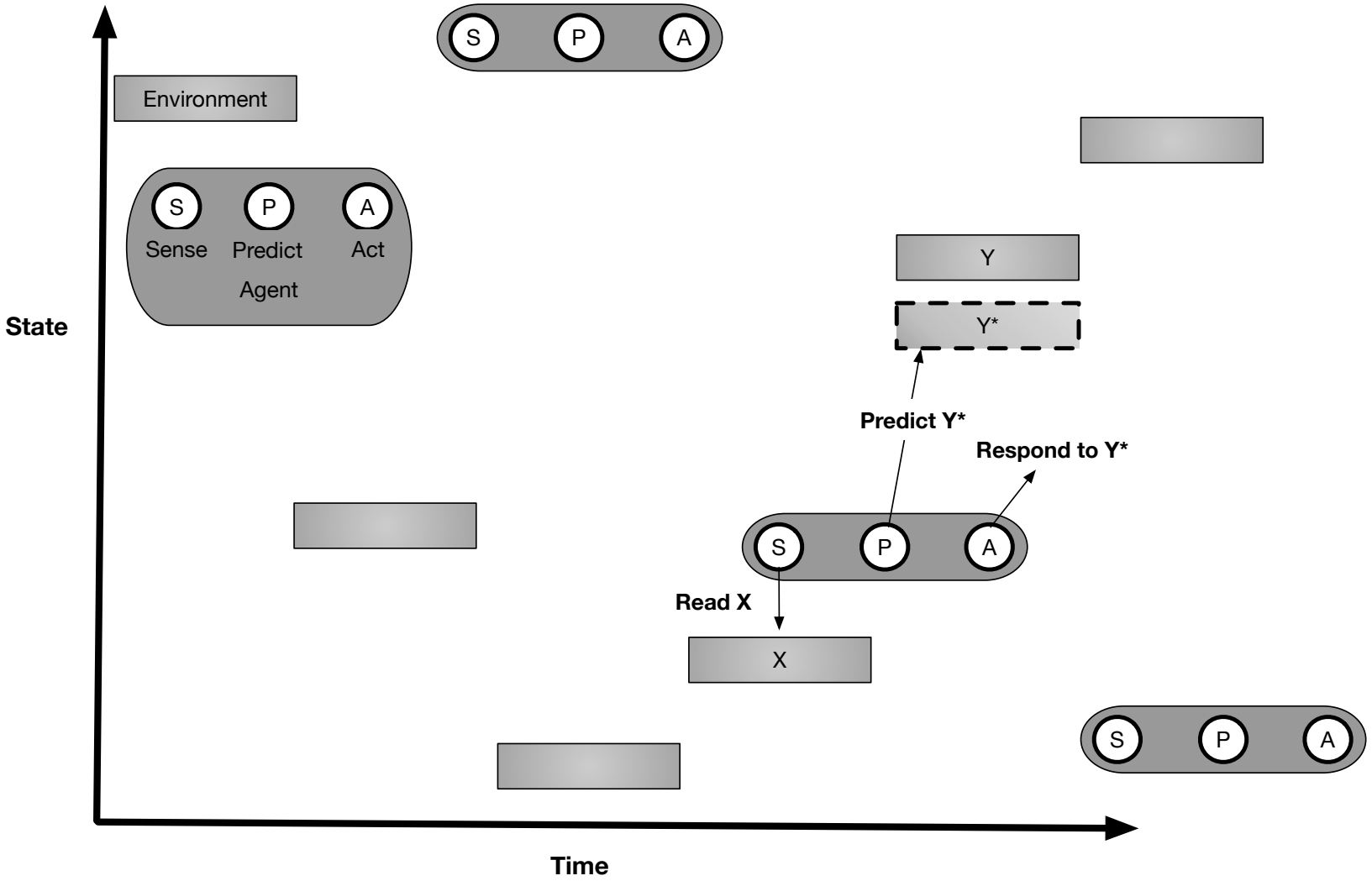


Life (and DEATH) in the Fast Lane



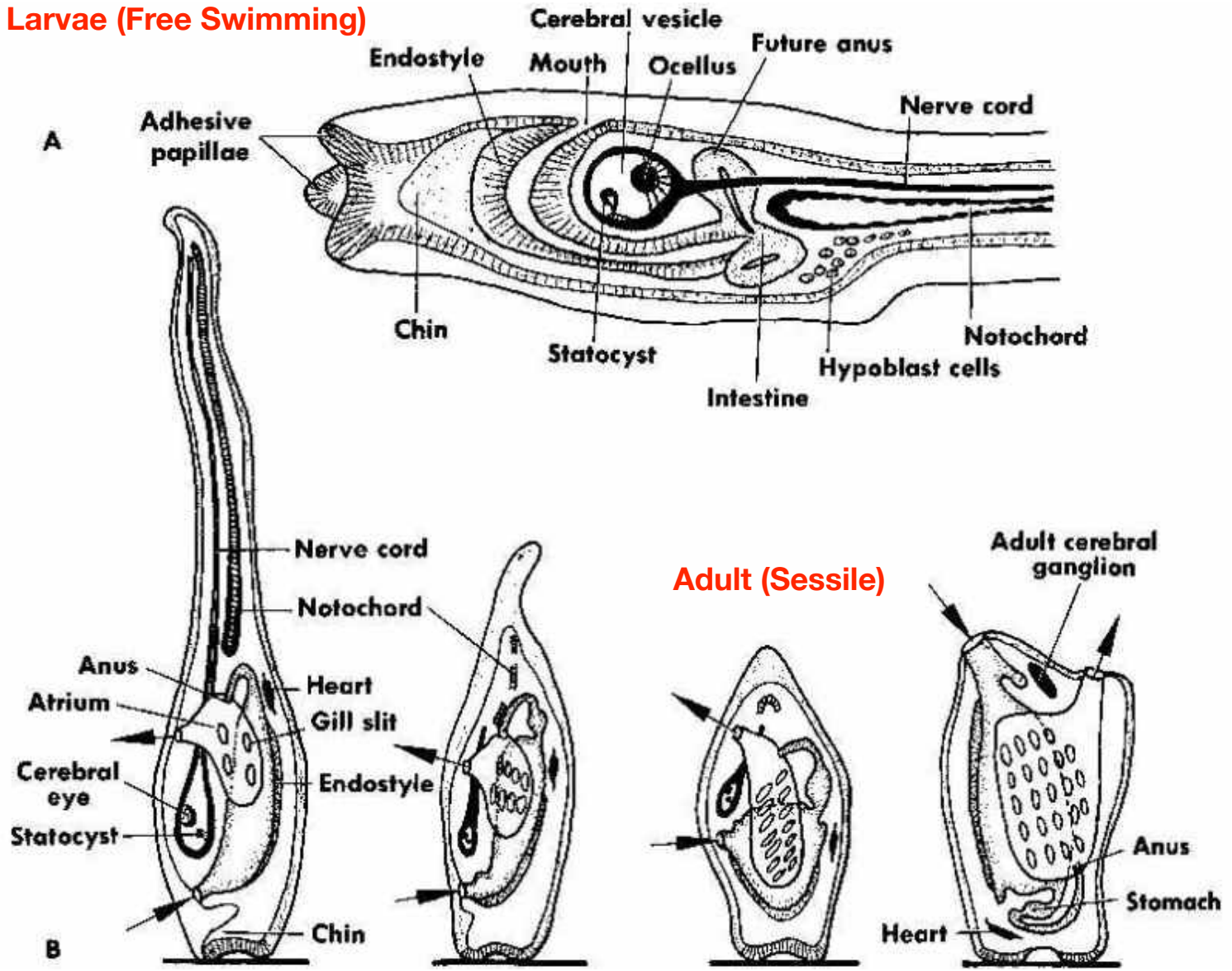
* For a **mobile** agent, relative frequency of environmental change increases.

Predict to Survive



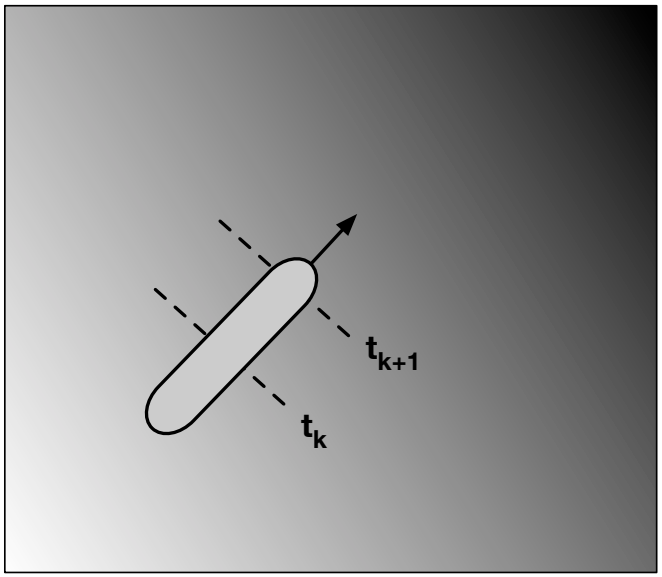
* Mobile agents need prediction.

The Brain-Eating Sea Squirt

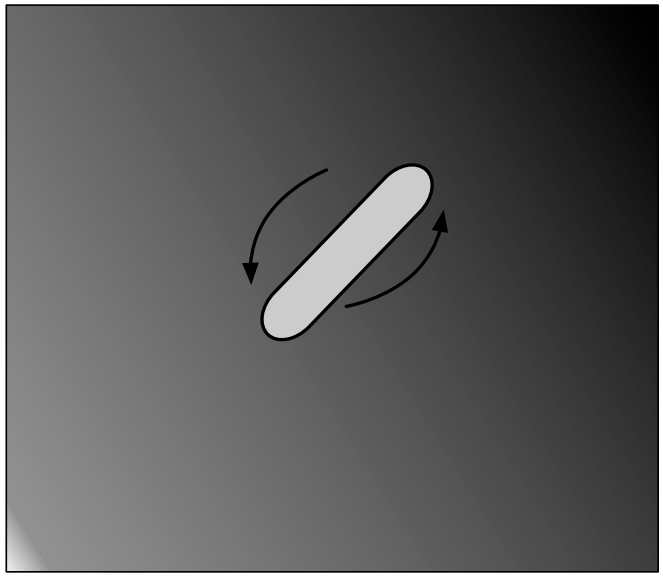


Implicit Prediction using Gradients

Swimming up a **strong** attractant gradient



Tumbling in a **weak** attractant gradient

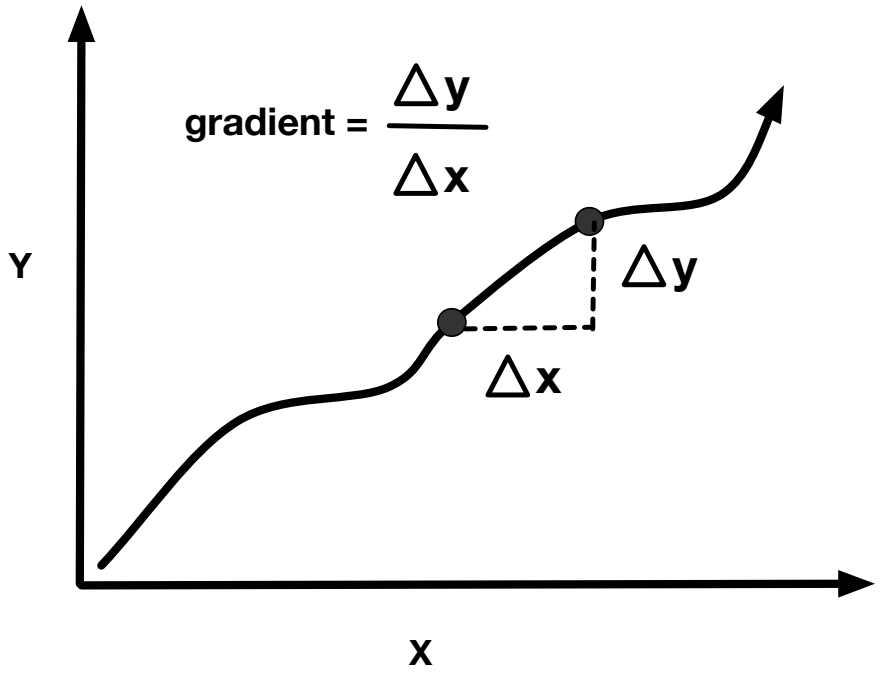


Gradient = $A(t_{k+1}) - A(t_k)$

A(t) = attractant read by head sensor at time t.

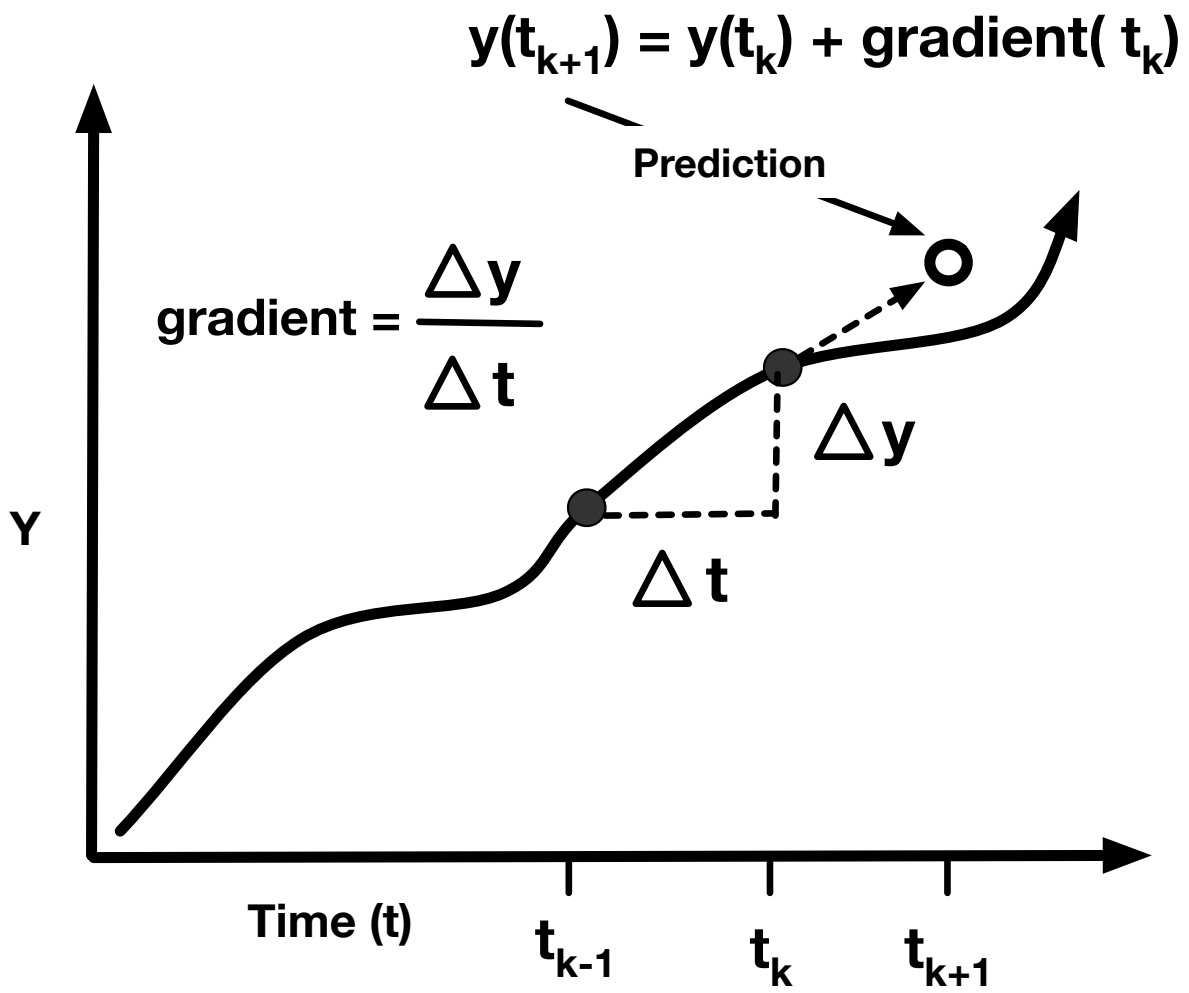
Gradients are simple, cheap (and amazingly accurate) predictors of future states: Future = Present + Gradient

Gradient = Derivative

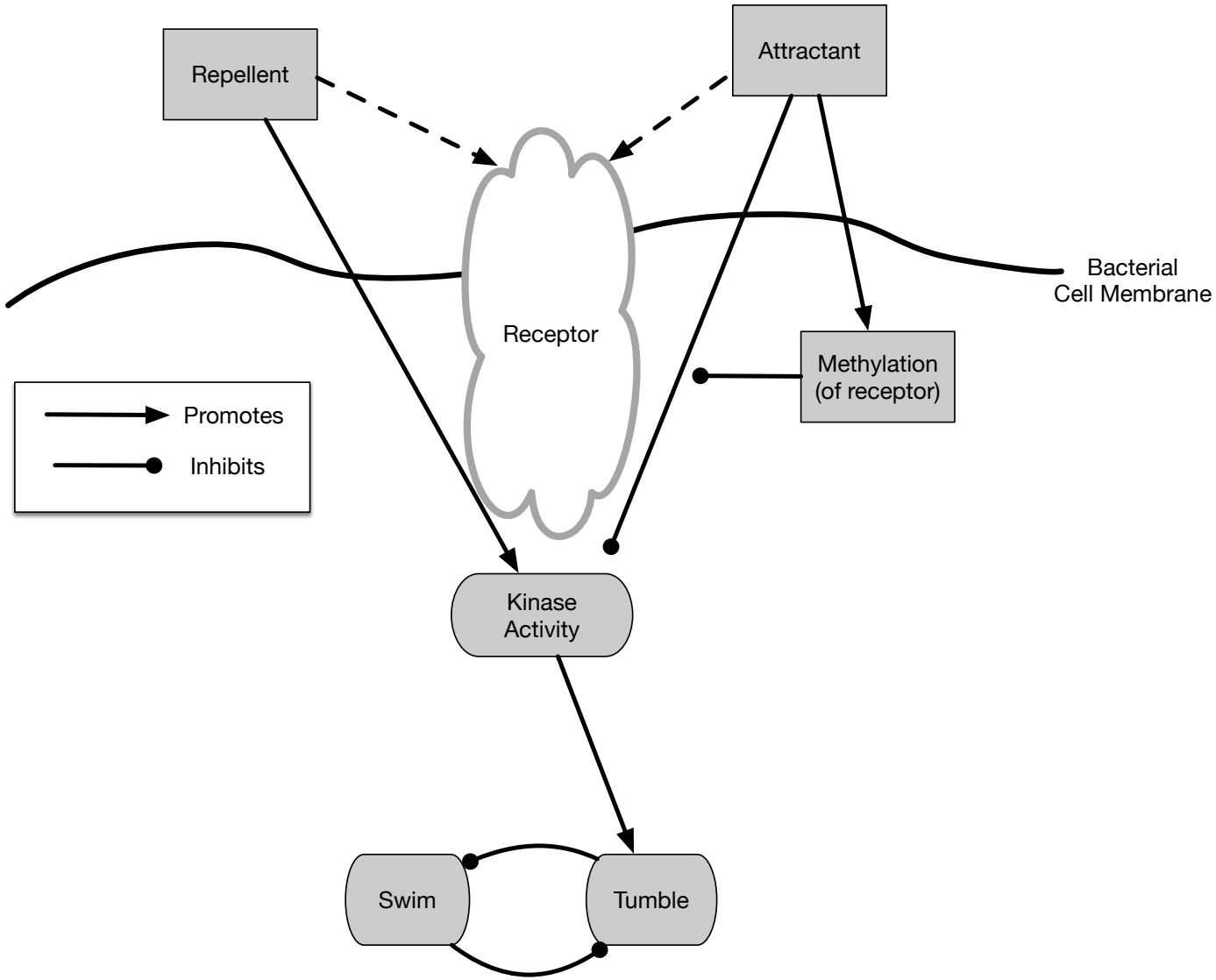


Situation	X	Y
Bacterial Foraging	Location	Nutrients
Finance	Time	Stock Price
Thermostat	Heat	Temperature
Deep Learning	Connection Weights	Output Error
Evolutionary Computation	Genotype	Fitness

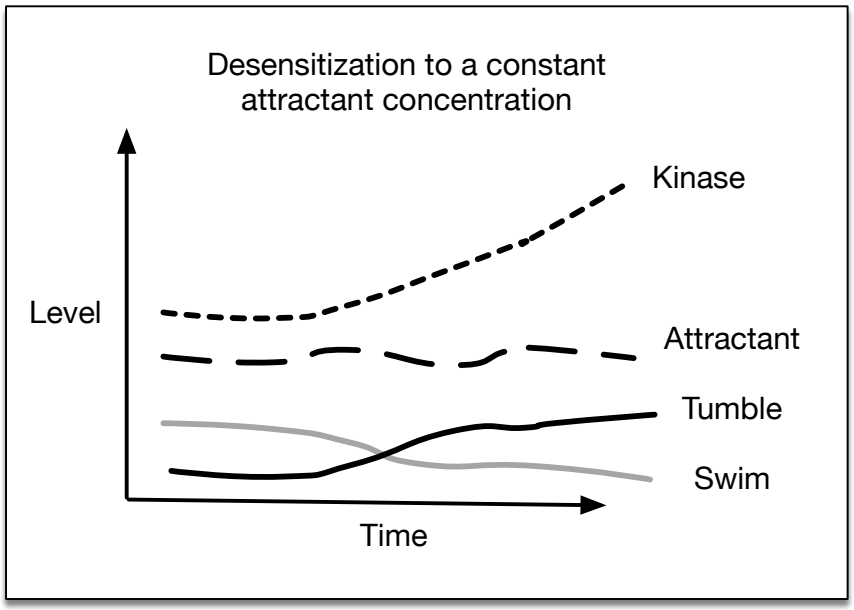
Prediction via Gradients

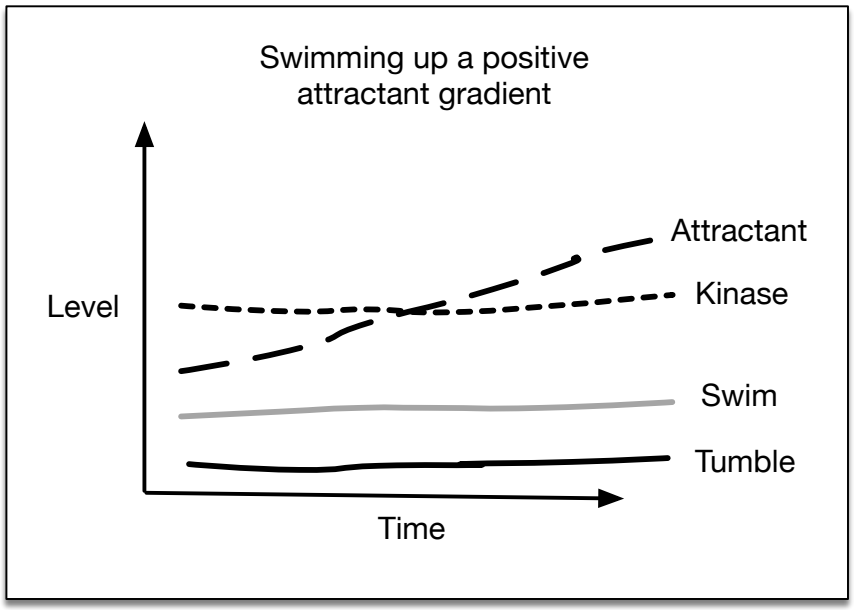


Gradient-Driven Behavior via Chemistry



Bacterial Response to the Gradient

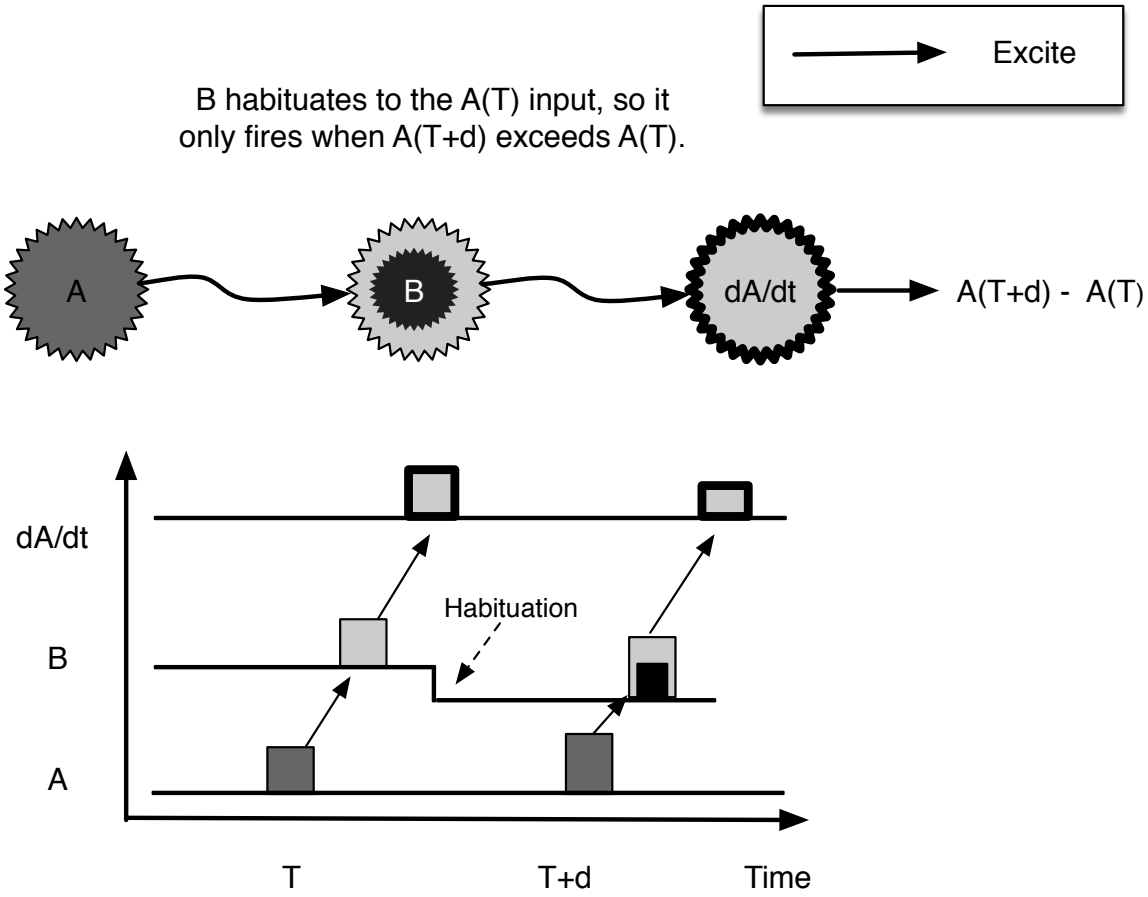




- Bacterium responds to the gradient of the attractant, not simply its current value.
- By moving, the bacterium encodes a spatial gradient as a temporal gradient, which it can then detect using receptors in a single location.

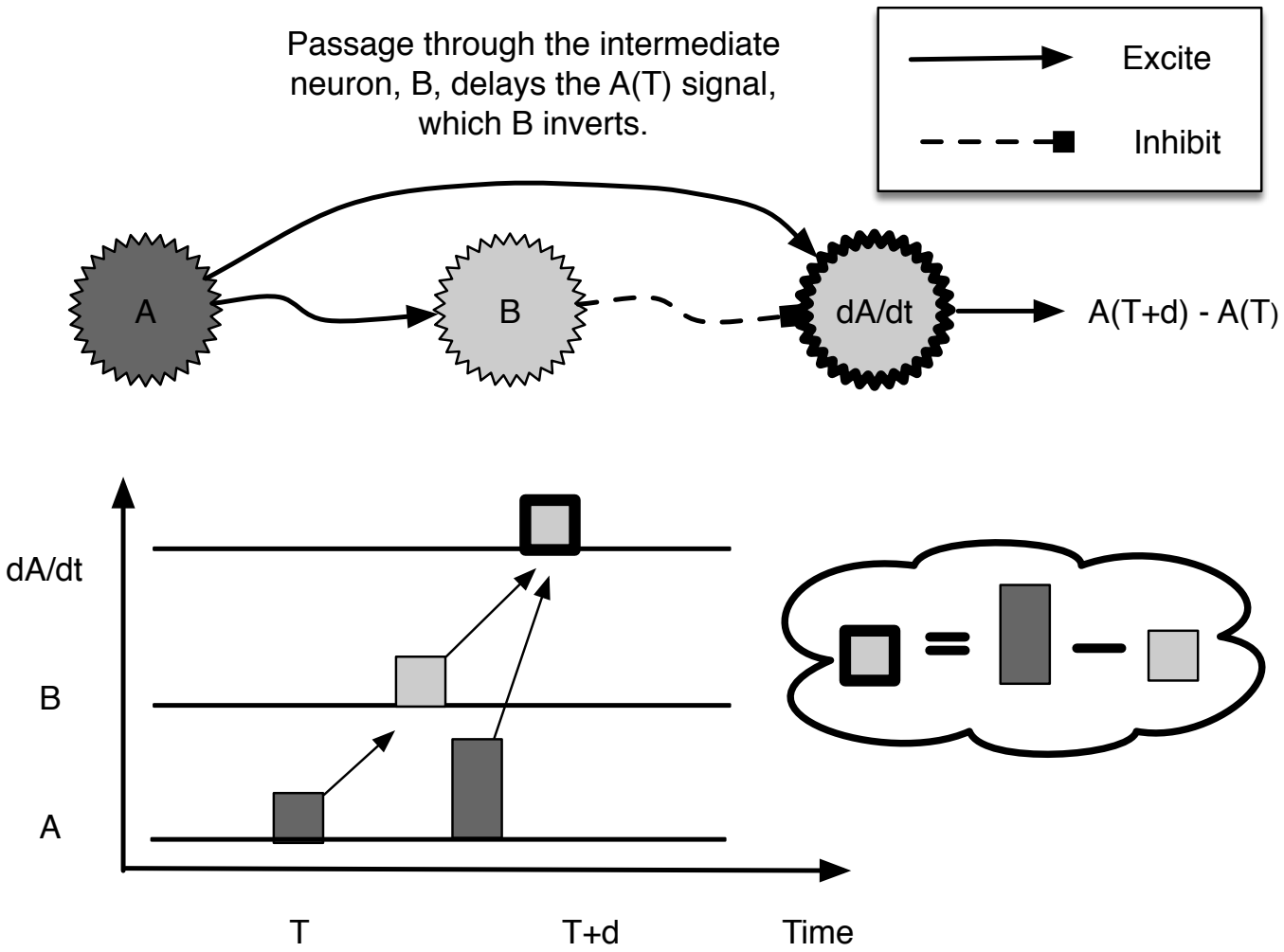
Temporal Differentiation via Habituation

Neurons can detect gradients too!

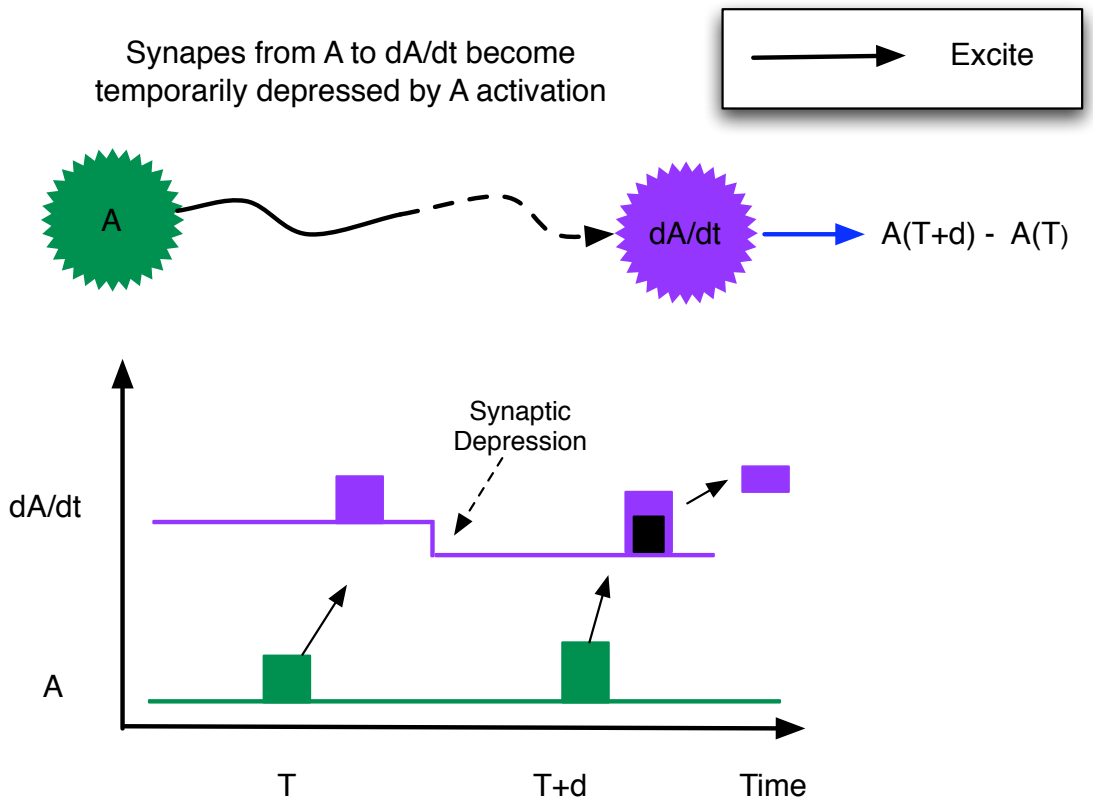


This is how (nematode worm) C. Elegans navigates. (Larsch et. al., 2015)

Temporal Differentiation via Delayed Inhibition



Temporal Differentiation via Synaptic Depression

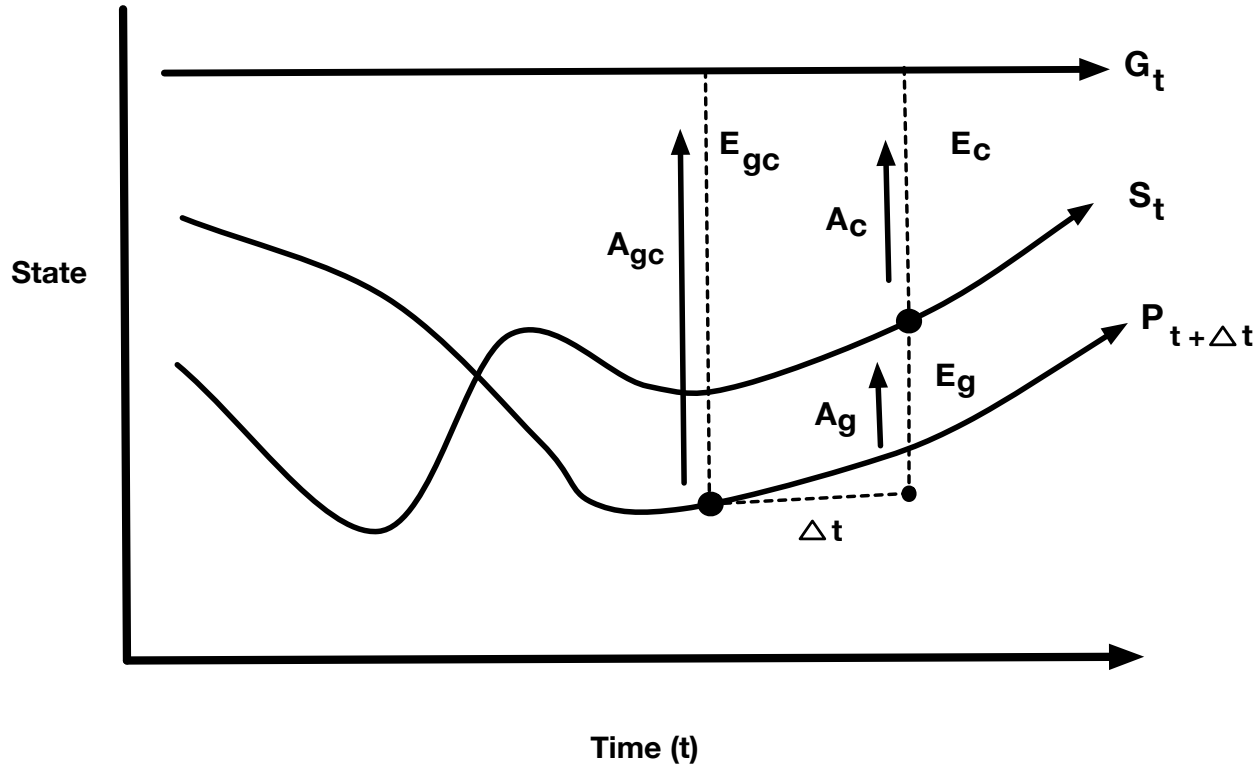


Tripp, B. and Eliasmith, C. (2010), Population Models of Temporal Differentiation, *Neural Computation*, 22, pp. 621-659.

Prediction via Averages

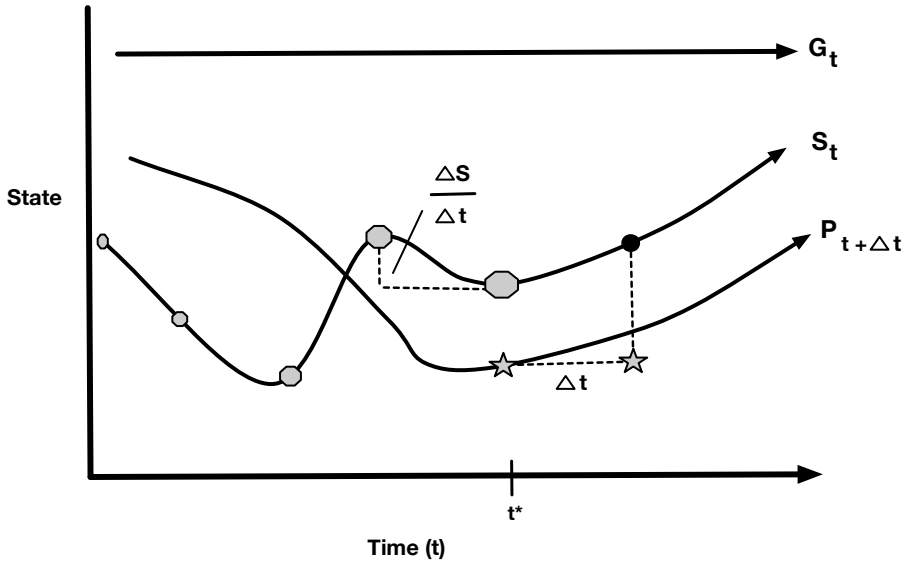
- A simple average (over time or space) can give a good prediction of a variable's next state ($S_{t+\triangle}$).
  - Time $\rightarrow$ History: $S_{t-k}, S_{t-k+1}, \ldots S_{t-1}, S_t$
  - Space $\rightarrow$ current value of the same variable in nearby regions, e.g. concentrations of a particular chemical in neighboring cells.
- Averaging $\rightarrow$ summing, integrating.
- Averages in space or time can also determine the **current** value S_t if it's unknown. These are also called *predictions* despite having a present rather than future tense.
- Neurons often average: they aggregate and scale signals over space and time (remember), and also leak (forget).
- Averages can both contribute to gradients and combine with gradients to support prediction.

Predictions -vs- Goals



- G_t = Goal at time t.
- S_t = System state at time t.
- $P_{t+\triangle t}$ = Prediction (guess) at time t of state at time $t + \triangle t$.
- E = error of guess (g), control (c), control relative to guess (gc).
- A = actions to reduce errror of guess (g), control (c), both (gc).

Prediction ≈ Control



Prediction

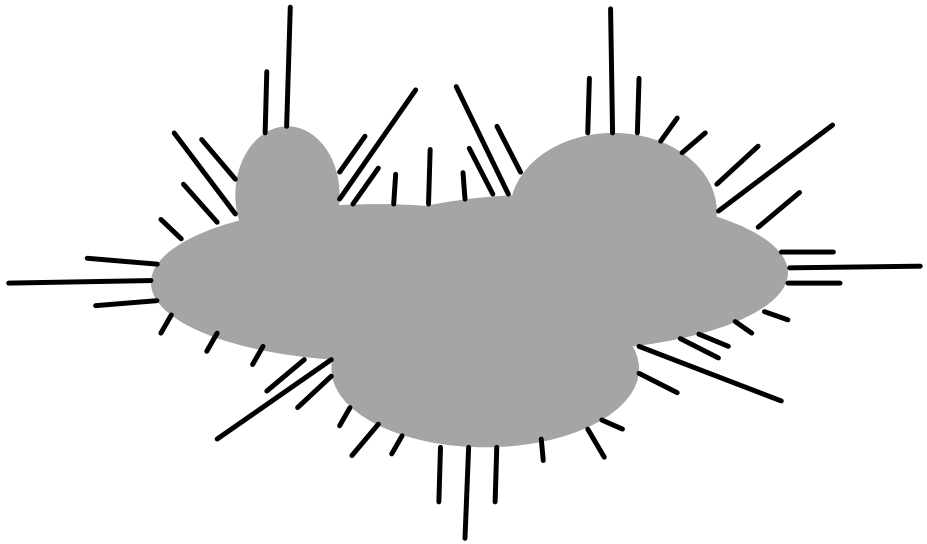
$$E_{t+\triangle t} = \Gamma(G \vec{-} S_t) = \underbrace{k_g(G - S_t) + k_g \frac{\triangle(G - S_t)}{\triangle t}}_{gradient-based} + \underbrace{k_a \sum_{j=0}^{M} w_j (G - S_{t-j\triangle t})}_{average-based} \quad (1)$$

PID Control

$$u_t = k_p e_t + k_d \frac{\triangle e_t}{\triangle t} + k_i \sum_{j=0}^{t} e_j \quad (2)$$

Same neural circuits evolved for control could be reused for prediction.

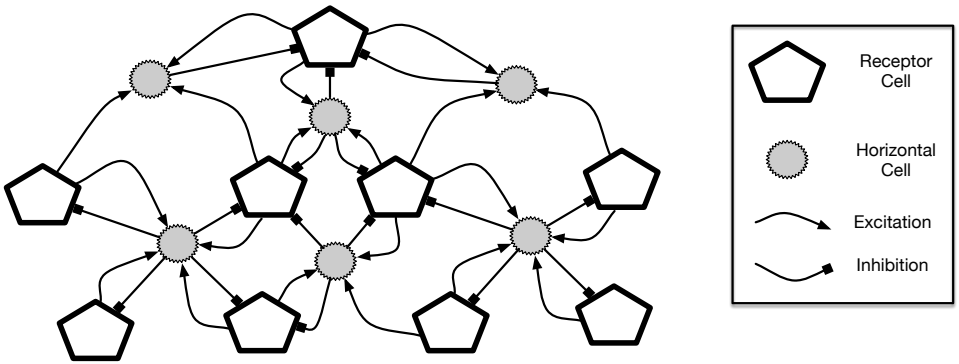
Redundancy Reduction in Visual Perception



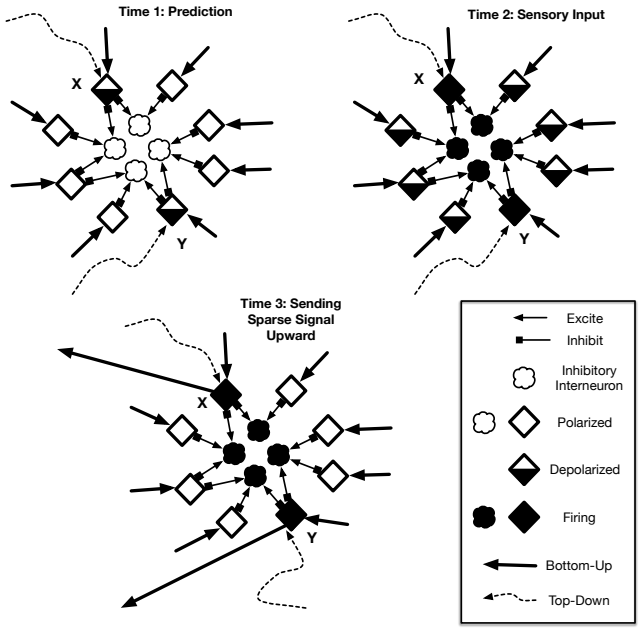
When we begin to consider perception as an information-handling process, it quickly becomes clear that much of the information received by any higher organism is redundant. Sensory events are highly interdependent in both space and time..... we can make better-than-chance inferences with respect to the prior and subsequent states of these receptors.. (Fred Attneave, 1954)

- *Efficient Coding* - Oliver(1952) - similar, for telecom.
- *Predictive coding* coined - Srinivasan et. al.(1982)

Predictive Coding in Neural Circuits

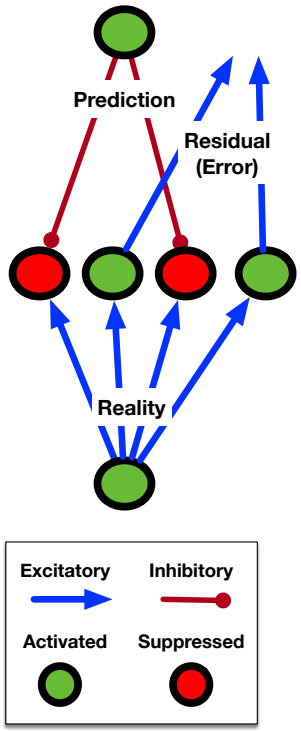


Srinivasan et. al. (1982)

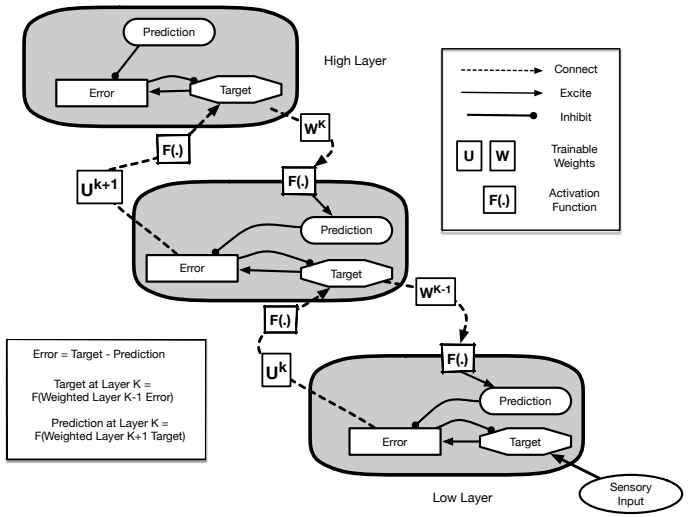


Hawkins(2015)

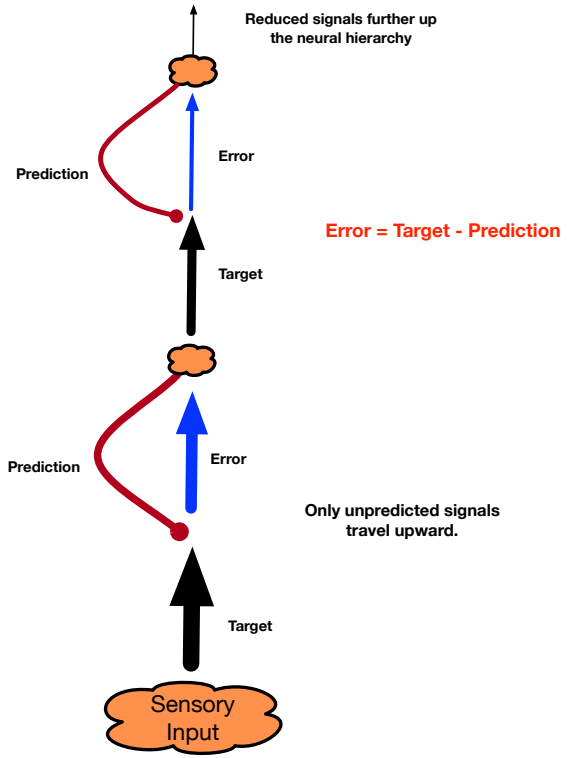
Predictive Coding in Artificial Neural Networks



Neural Essence

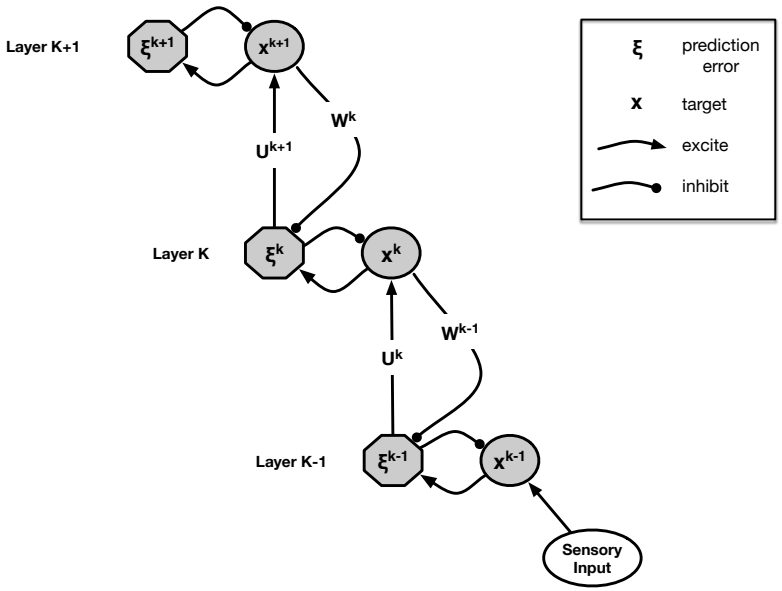


Rao + Ballard (1999)

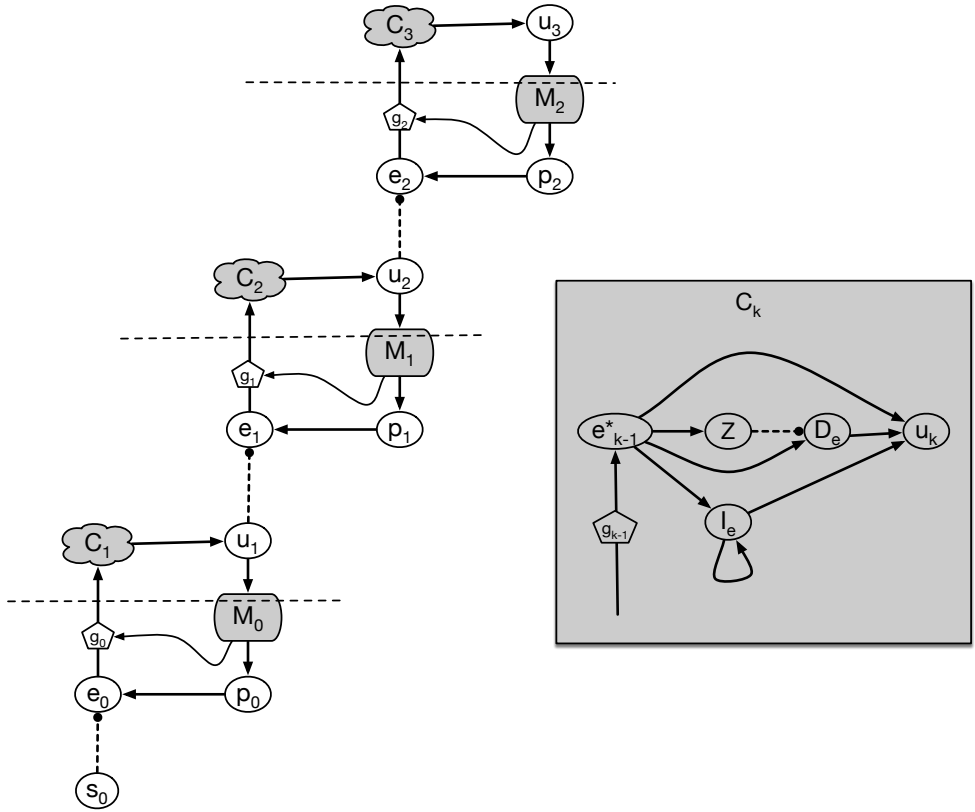


General Model

Predictive Coding Networks and Control

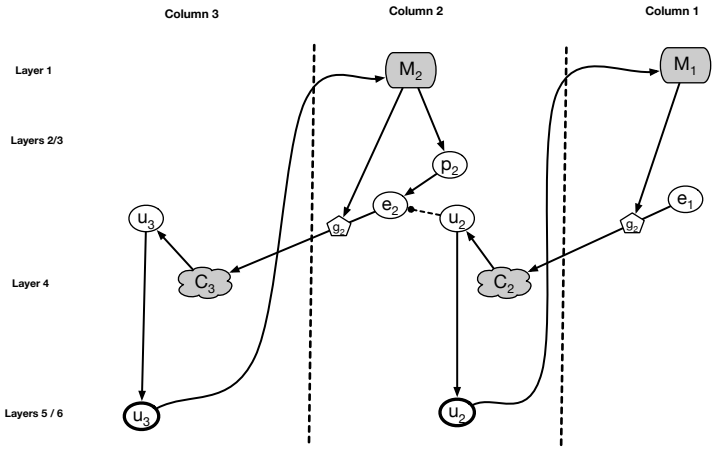


Layered Predictors

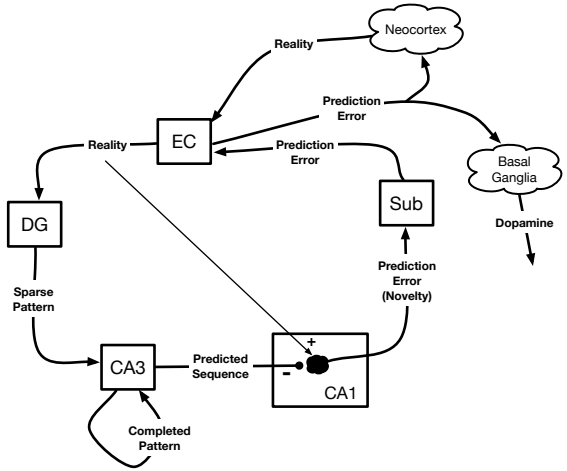


Hierarchy of PID Controllers

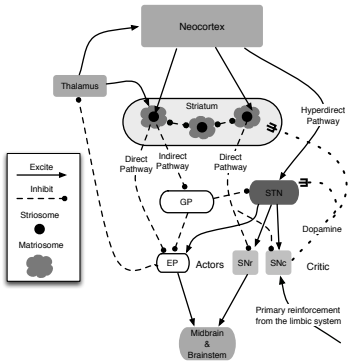
Prediction in Diverse Brain Regions



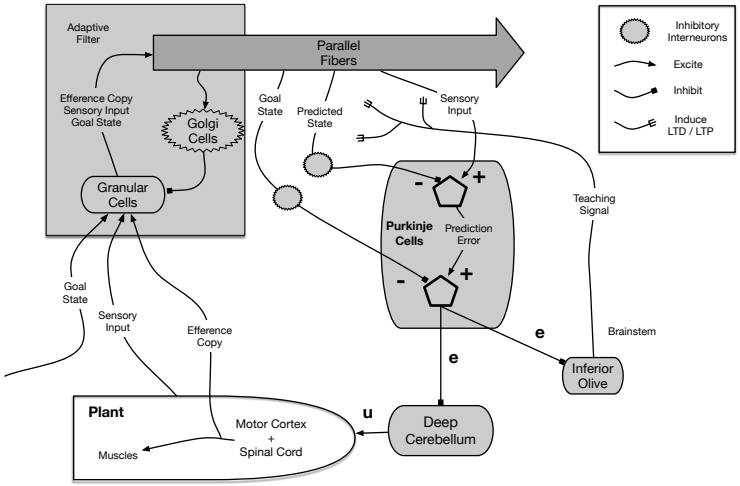
Neocortex



Hippocampus

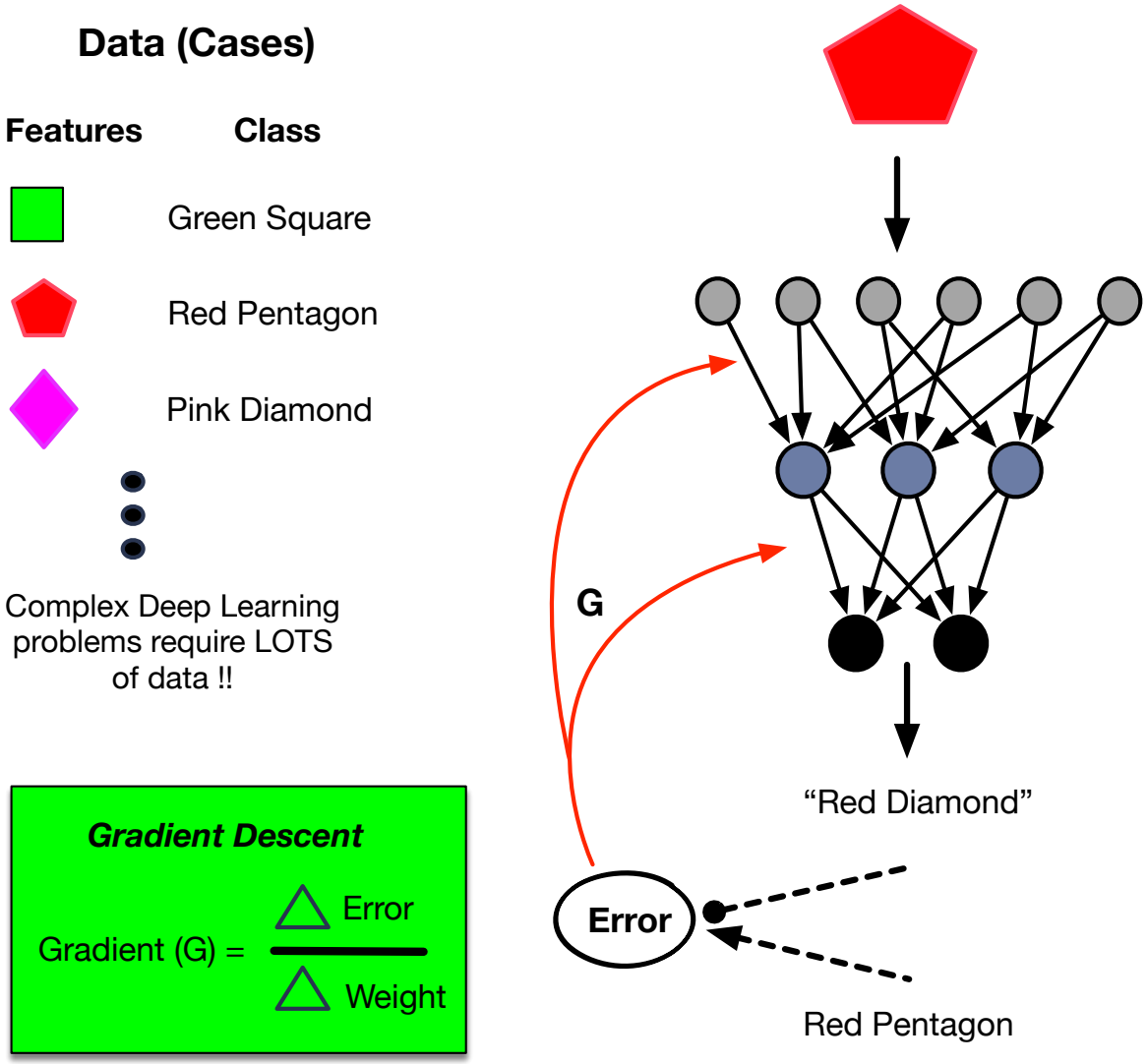


Basal Ganglia



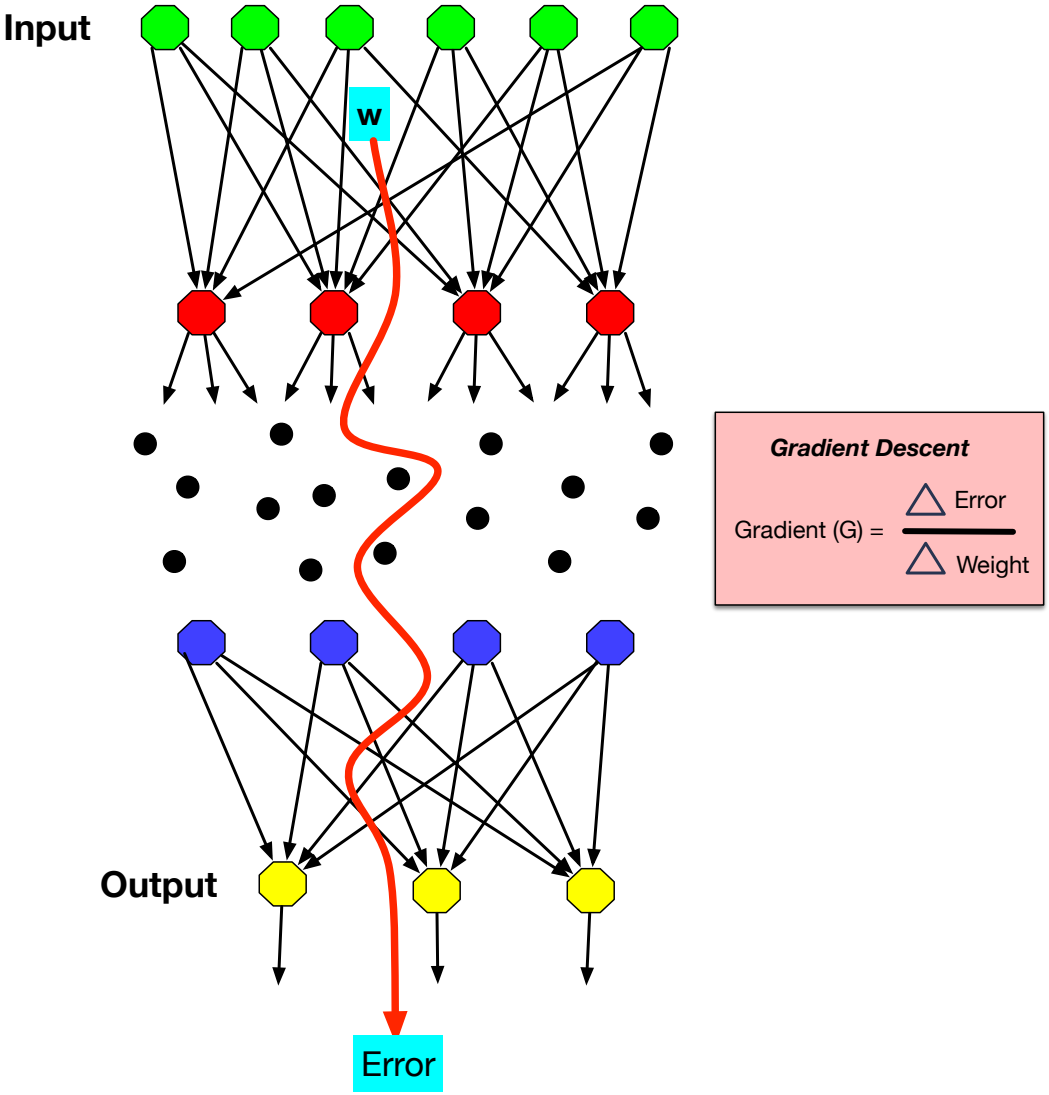
Cerebellum

Backpropagation: Cornerstone of Deep Learning



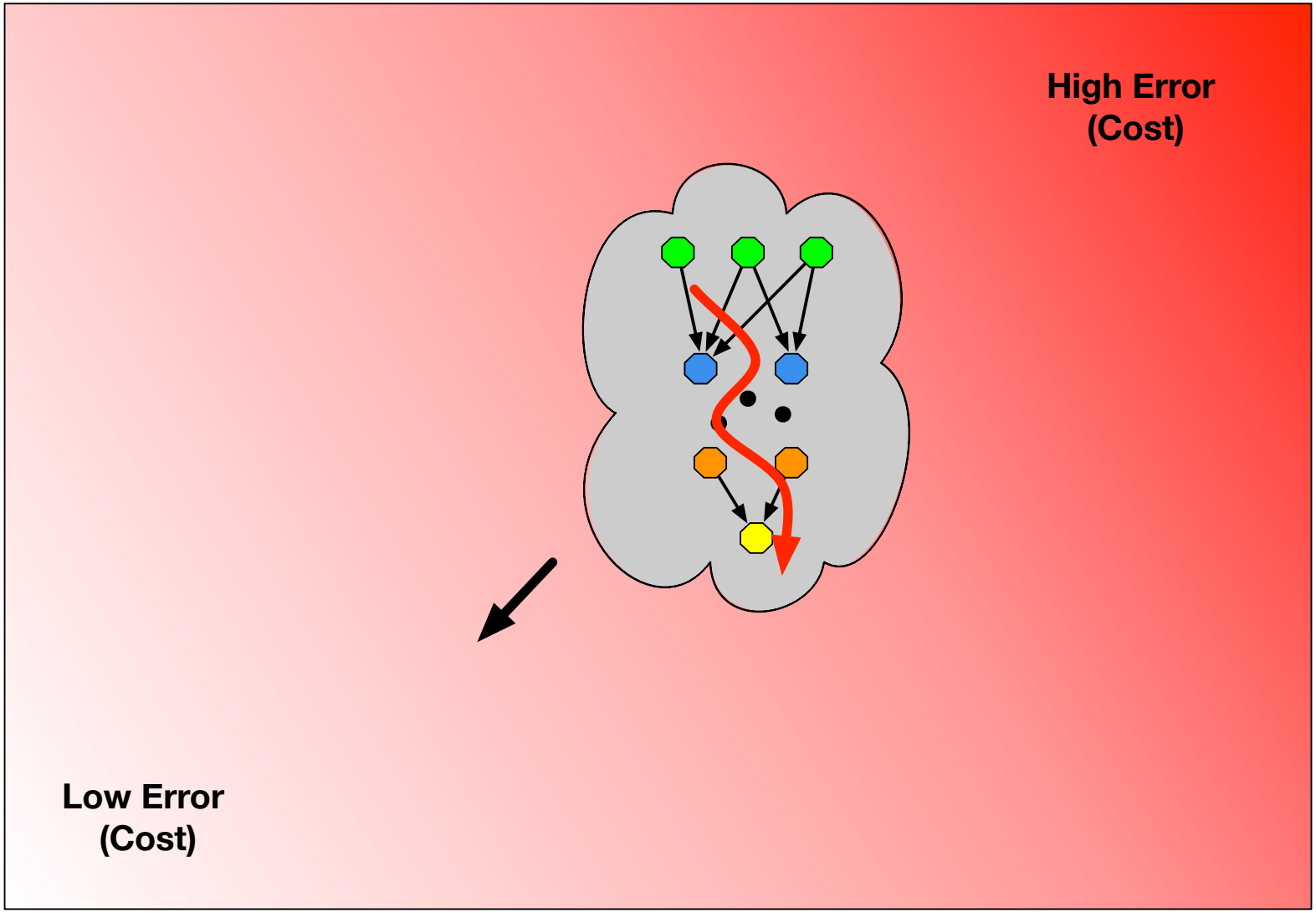
Gradient enables **prediction** of future error resulting from a weight change.

Gradients: Global and Ubiquitous



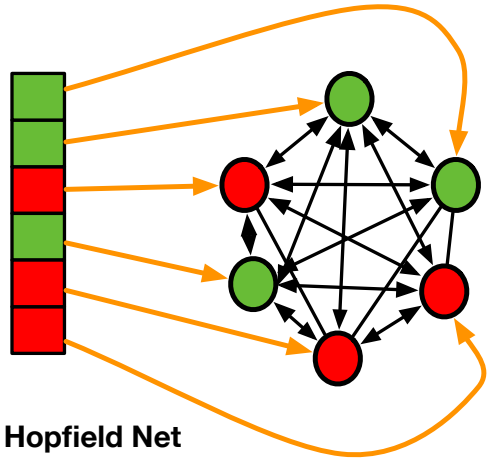
Learning based on these long-distance gradients $\rightarrow$ recent AI success. But not biologically plausible.

Descending the Error Gradient



Weight modifications based on **global** (distant) network relationships $\longrightarrow$
System descends global error gradient. **Metric = Total Error**

Connectionism and Prediction: Energy Nets

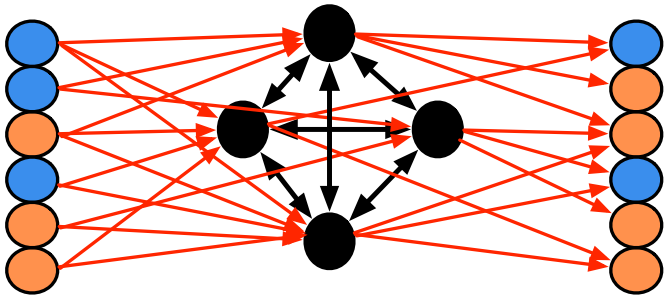


Hopfield Net

+ Stochasticity

+ Wake - Sleep

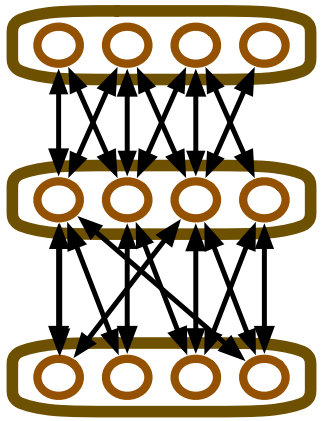
+ Contrastive Divergence



Boltzmann Machine

+ Many Hidden Layers

- Intralayer Links

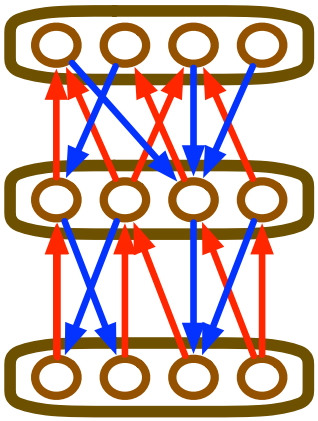


Restricted
Boltzmann Machine

+ Unidirectional Links

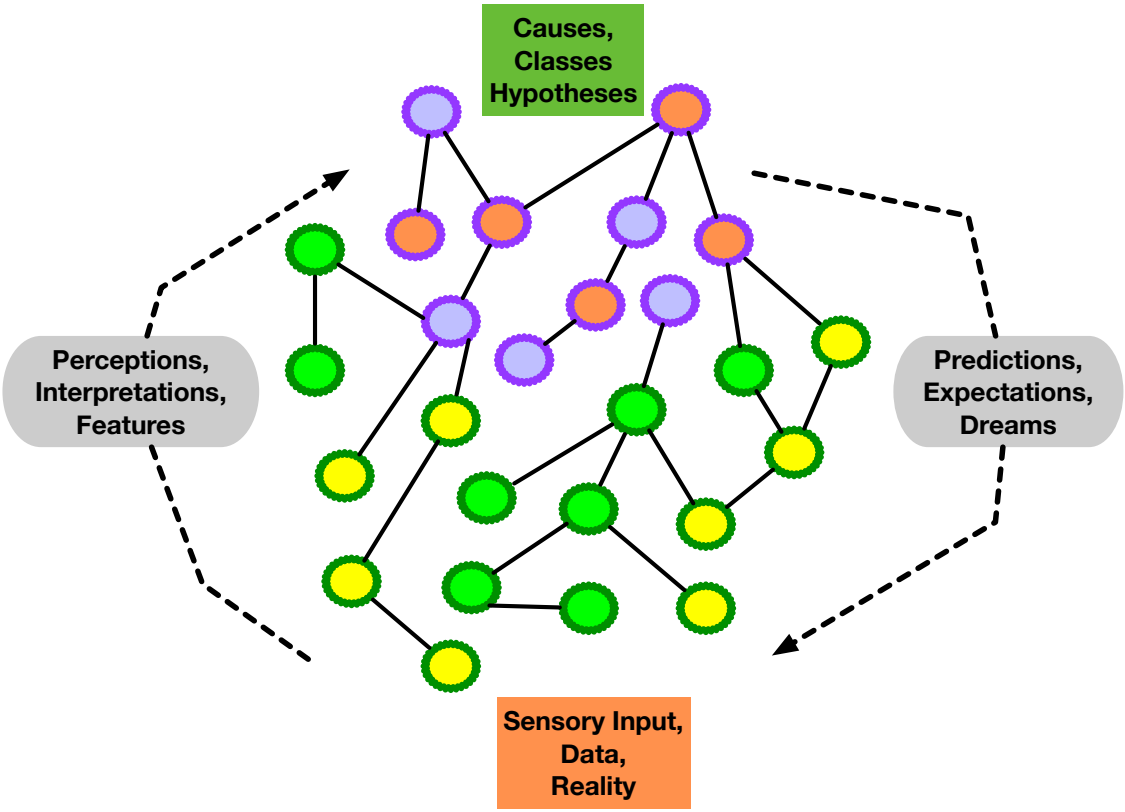
+ Recog - vs- Gen Links

+ Free Energy



Helmholtz Machine

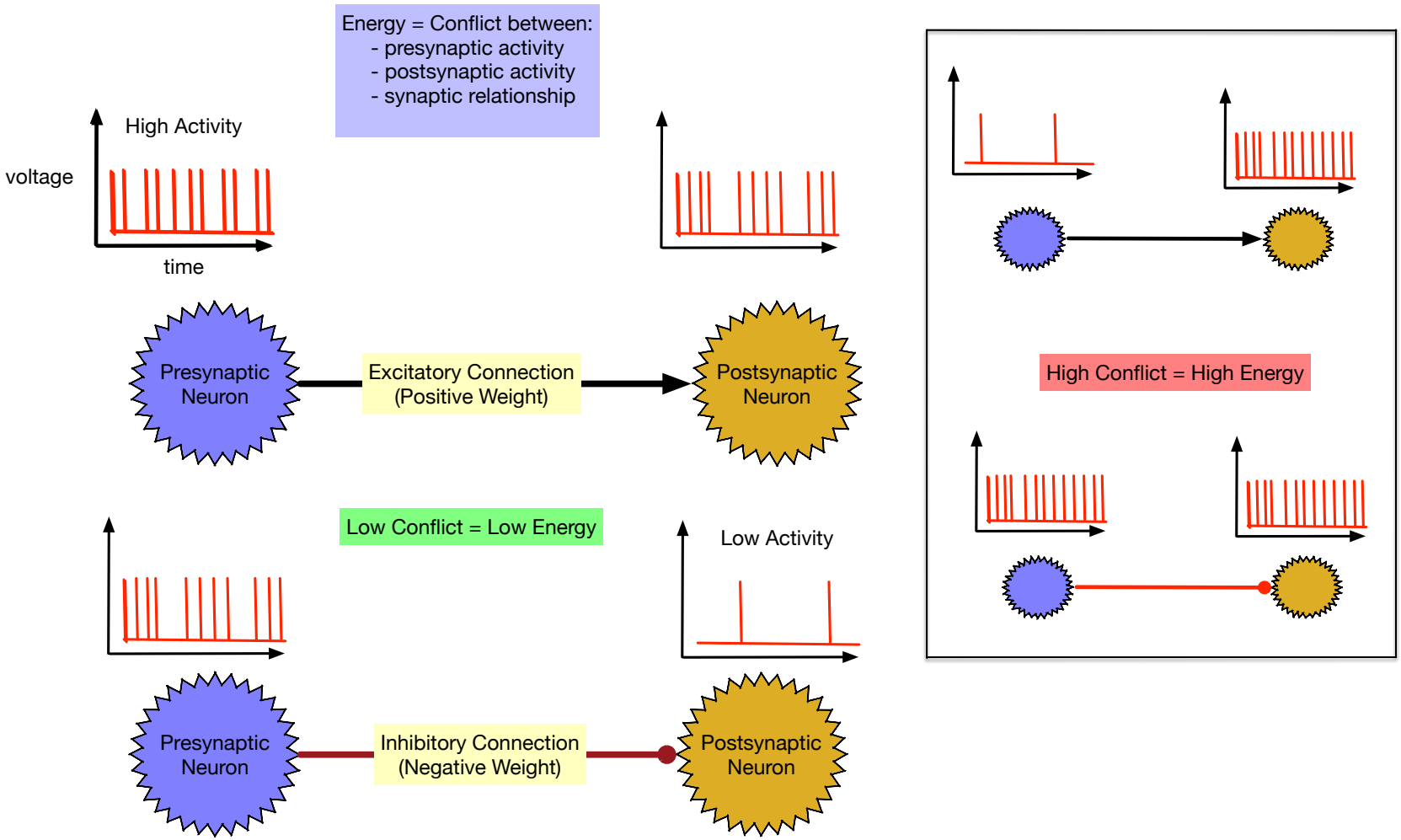
Alternating Phases: Interpretation + Prediction



Wake-Sleep Training

- Wake phase (based on data)
- Sleep / Dream phase (based on model-generated patterns = predictions)
- Different variations in Boltzmann, Restricted Boltzmann and Helmholtz Machines.

Energy Metrics for Neural Networks



$$Energy = \sum_{pre,post} -X_{pre} X_{post} W_{pre \to post}$$

Energy Gradients

- Minimize energy instead of error.
- Each weight's contribution to energy is **only local**:

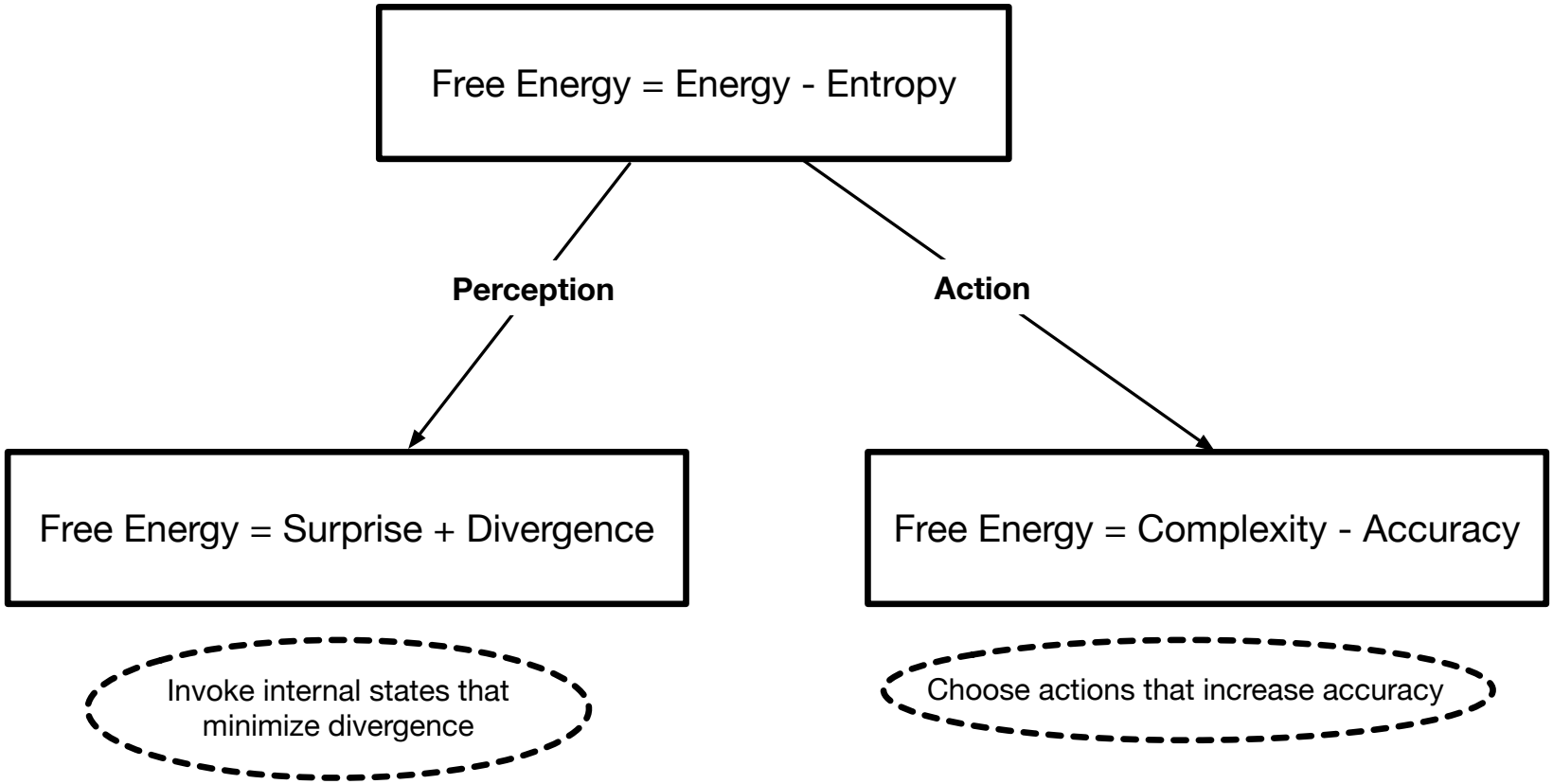
$$\frac{\partial Energy}{\partial W_{ab}} = \frac{\partial}{\partial W_{ab}} \sum_{j,k} -X_j X_k W_{jk} = -X_a X_b$$

  where $X_j, X_k, W_{j,k} \in [-1, 1]$
- Learning = Adjusting weights to reduce energy.
- Learning = Descending the energy gradient.

$$\triangle W_{ab} = -\lambda \frac{\partial Energy}{\partial W_{ab}} = \lambda X_a X_b$$

  where λ = learning rate
- This is very Hebbian, very biological..very ALIFE !!
- X_a, X_b and Energy take many forms (in different models), but learning remains Hebbian.

Karl Friston's Free Energy Principle (FEP)



- Basis = Variational Free Energy = $F_g^r(d) = \langle E_g(s;d) \rangle_r - H_r(s|\Theta)$
- Use following starting version for perception and action derivations:

$$F_g^r(d) = \langle -ln[p_g(s,d)] \rangle_r + \langle ln[p_r(s|\Theta)] \rangle_r$$

Perception and The Free Energy Principle

- Use this definition of Free Energy:

$$F_g^r(d) = \underbrace{-ln[p_g(d)]}_{Surprise} + \underbrace{D_{KL}(p_r(s|\Theta), p_g(s|d))}_{Divergence}$$

- Perception = Run system in recognition mode given sensory input, d.
- Goal: Produce a distribution of internal states that best matches the distribution of potential causes of d, (i.e. environmental states that generate d).
- Use system's complete state distribution (s,d) during generative mode as basis for the target distribution of causes: $p_g(s|d)$
- Focus on Θ in $p_r(s|\Theta)$, where Θ = parameters of the system such as synaptic strengths, neuromodulators.
- Perception = choosing Θ to reduce divergence and thus reduce variational free energy.
- Bayesian Brain Hypothesis: Perception = modifying system parameters to move the recognition distribution $p_r(s|\Theta)$ closer to the posterior distribution $p_g(s|d)$ found by inverting the generative model.

Action and The Free Energy Principle

- Use this definition of Free Energy:

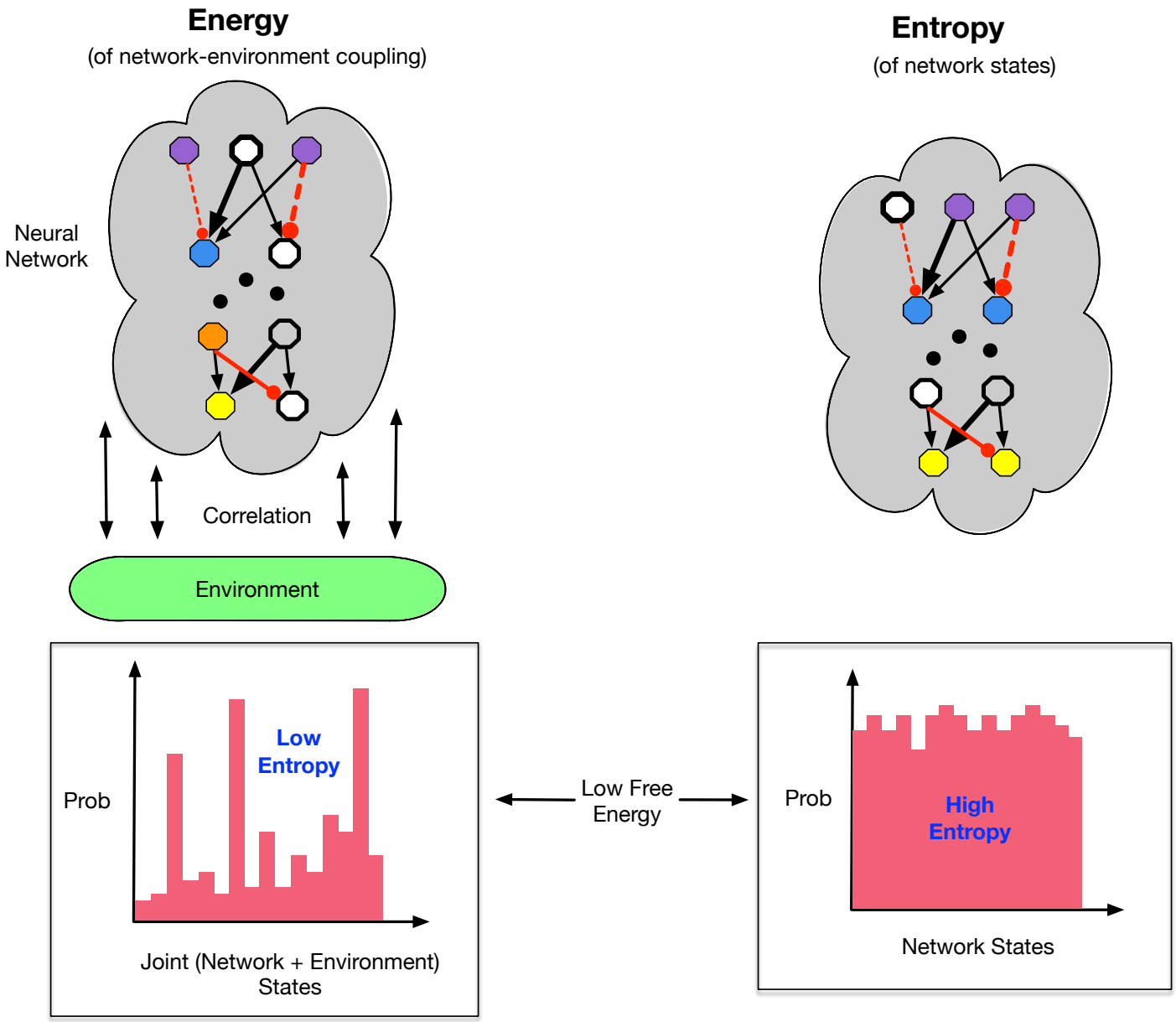
$$F_g^r(d) = \underbrace{D_{KL}(p_r(s|\Theta), p_g(s))}_{Complexity} - \underbrace{\sum_s p_r(s|\Theta) ln[p_g(d|s)]}_{(Predictive)Accuracy}$$

- Action = Active Inference = Choose activities producing sensory inputs (d) that are consistent with the current representation of the world.
- Representation defined by $p_r(s|\Theta)$ and $p_g(d|s)$.
- Now, the input sensory state (d) is a function of the action (α):

$$F_g^r(d) = \underbrace{D_{KL}(p_r(s|\Theta), p_g(s))}_{Complexity} - \underbrace{\sum_s p_r(s|\Theta) ln[p_g(d(\alpha)|s)]}_{(Predictive)Accuracy}$$

- Goal: Increase Predictive Accuracy by choosing actions that produce sensory states that are highly probable under p_r and p_g.
- In other words: Reduce Free Energy by reducing prediction error.

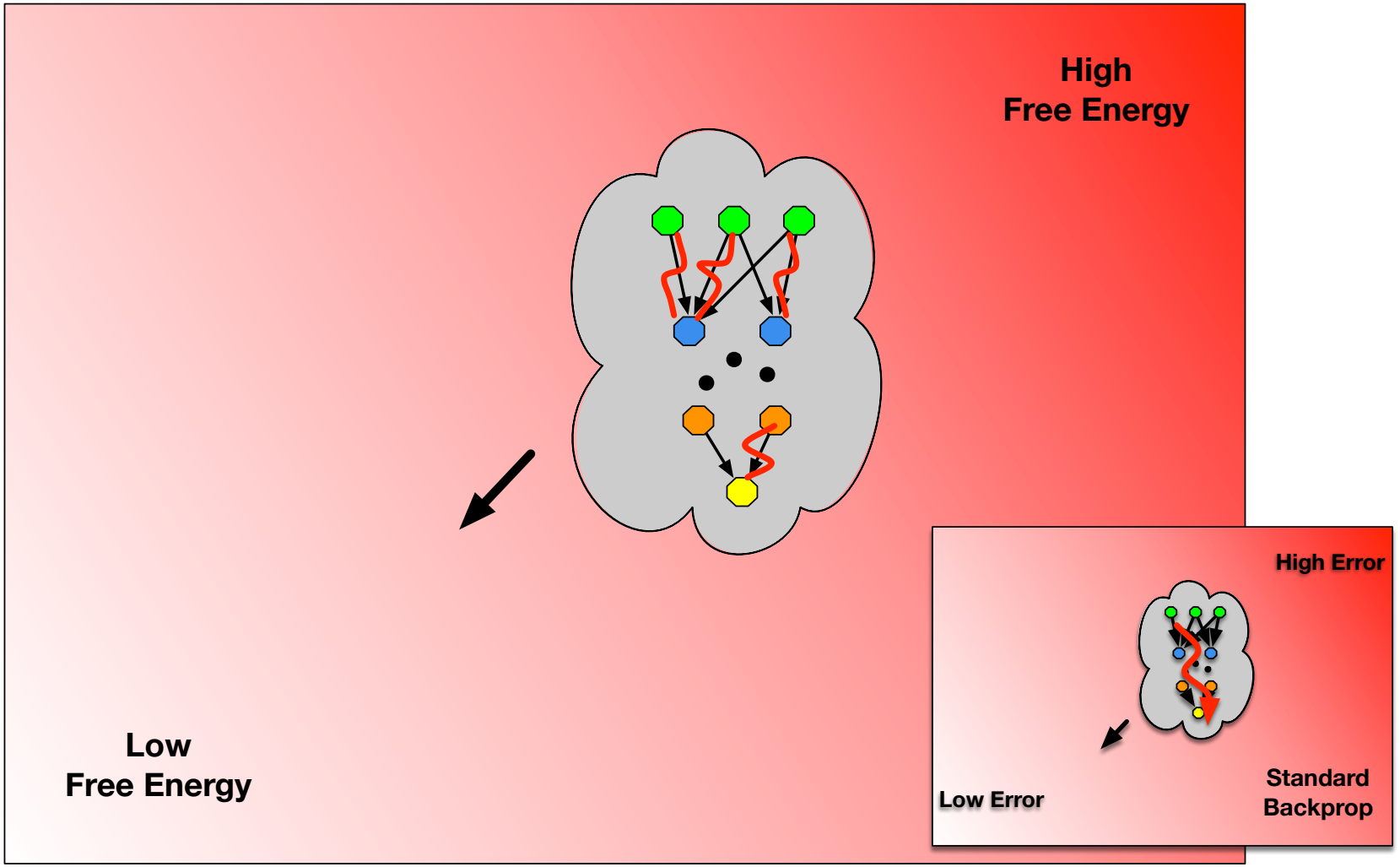
Goal: Free Energy ↓ ↦ Energy ↓ - Entropy ↑



Free Energy in Neural Networks

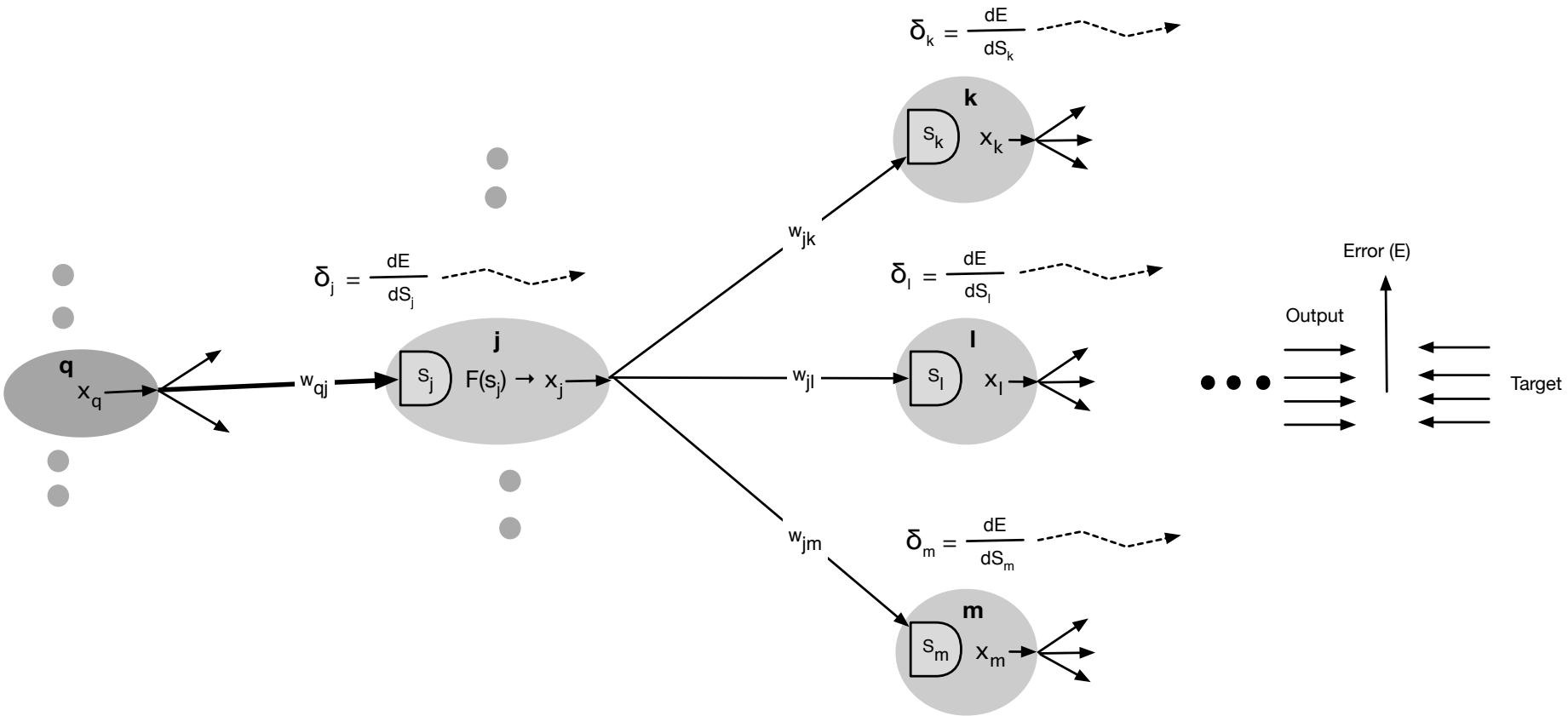
- NN adapts to achieve useful mappings between internal states (S) and environmental states (D).
- Mapping success = similarity between the prob. distr. of outputs produced by the NN ($p_g(D)$) and D's natural prob. distr : p(D).
- These probabilities stem from a measure of energy based directly on the concept of **surprisal** from information theory.
- This relationship between prob and energy is exactly the same as given by the Boltzmann distribution.
- The process of making $p_g(D)$ similar to p(D) = minimizing Kullback-Leibler divergence: $D_{KL}(p_g(D), p(D))$.
- This turns out to be equivalent to minimizing $-ln(p_g(D) = -lnZ$, where again, Z is from Boltzmann distr., and -lnZ = free energy.
- Thus, adapting an NN = minimizing free energy.

Descending the Free-Energy Gradient



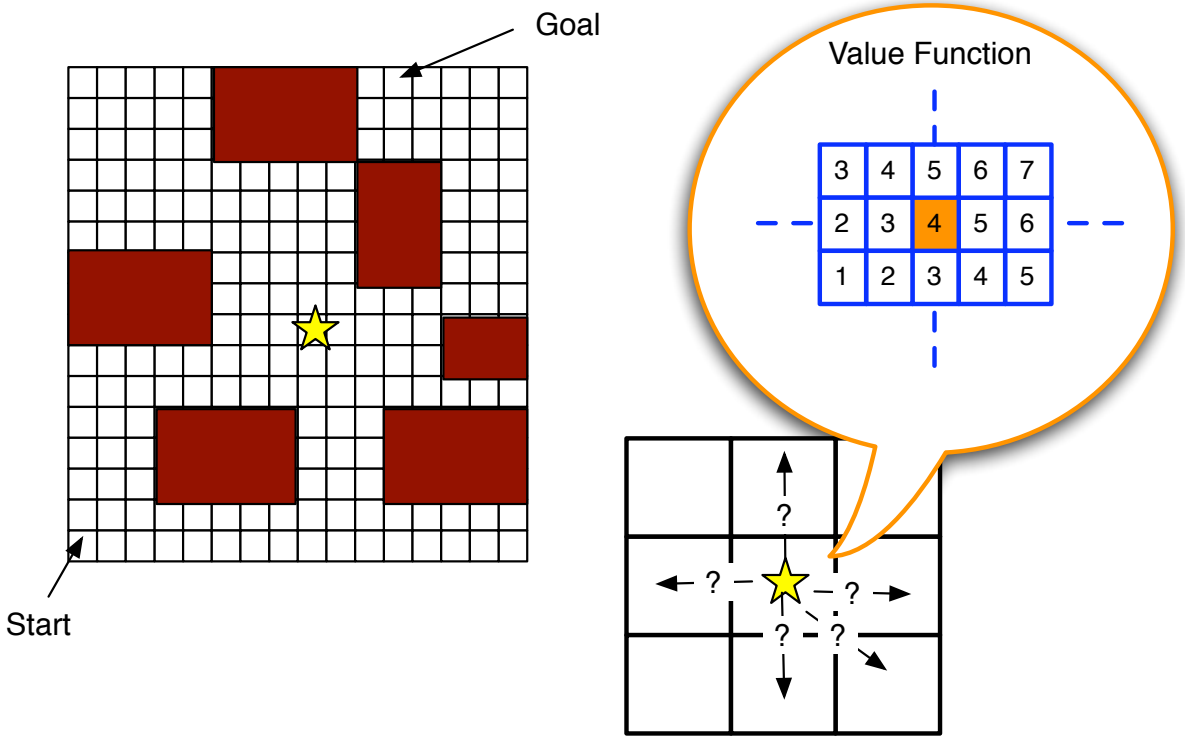
Weight modifications based on **local** network relationships $\longrightarrow$ System moves down the global free-energy gradient.

Predictive Coding instead of Backprop



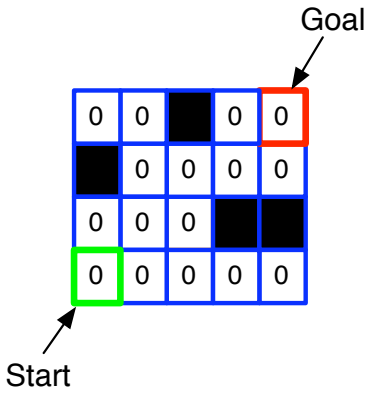
- Replacing δ (a long-distance gradient) with ξ (a prediction error) yields an energy network with respectable classification abilities.
- Rafal Bogacz et. al. (2017, 2019, 2021) - they do the math...and code.

Reinforcement Learning



- Do series of actions (strategy) to get from start to goal.
- Receive **intermittent** feedback (i.e. reward)
- Over **many trials**, learn a good strategy.

Prediction of Future Reward



Reach Goal + Backup Reward

0	0		0	9
	5	6	7	8
0	4	0		
0	3	0	0	0

Reach Deadend + Backup Penalty

-9	-5		0	9
	3	6	7	8
0	2	0		
0	1	0	0	0

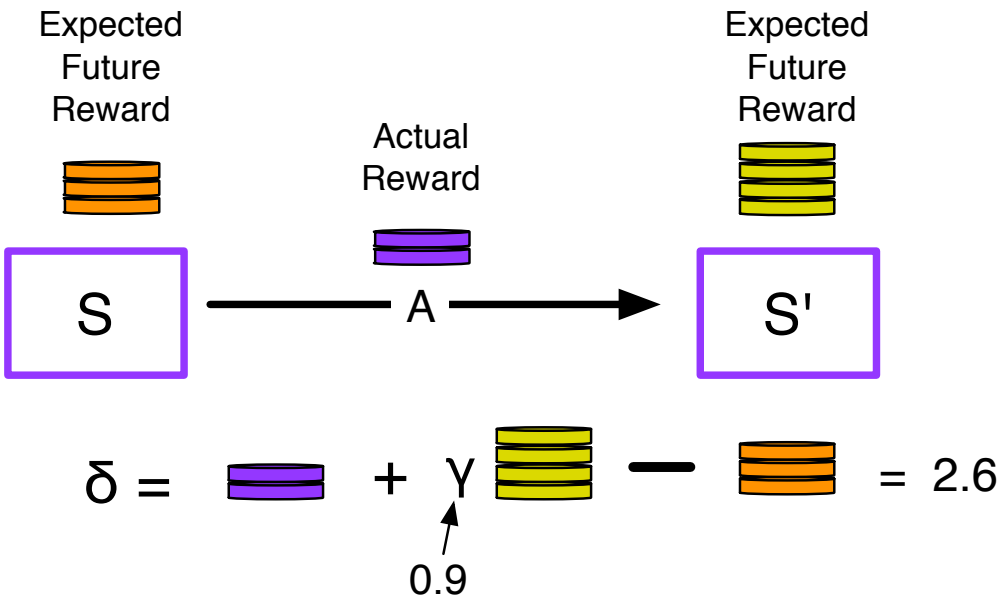
After Many Exploratory Rounds

-9	-5		7	9
	3	4	5	7
1	2	3		
0	1	2	-5	-9

V(s) = Value of state s

= Predicted cumulative reward from s to a goal state.

Temporal Differencing (RL Variant)



$$\triangle \mathbf{V}(\mathbf{S}) = \lambda \delta$$

where $\lambda =$ learning rate, $\gamma =$ discount factor, and $\delta =$ TD Error

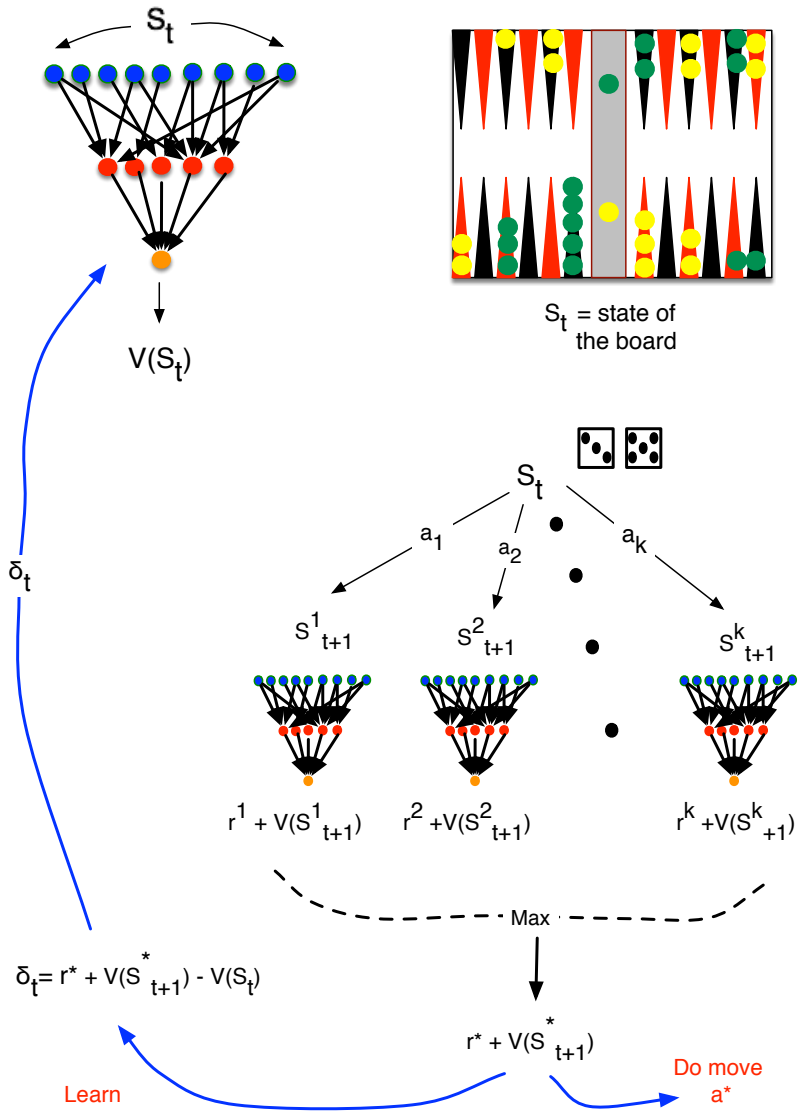
Bootstrapped Prediction Improvement

Prediction of Sum Reward($t_k \rightarrow t_{Final}$) updated by

Prediction of Sum Reward($t_{k+1} \rightarrow t_{Final}$)

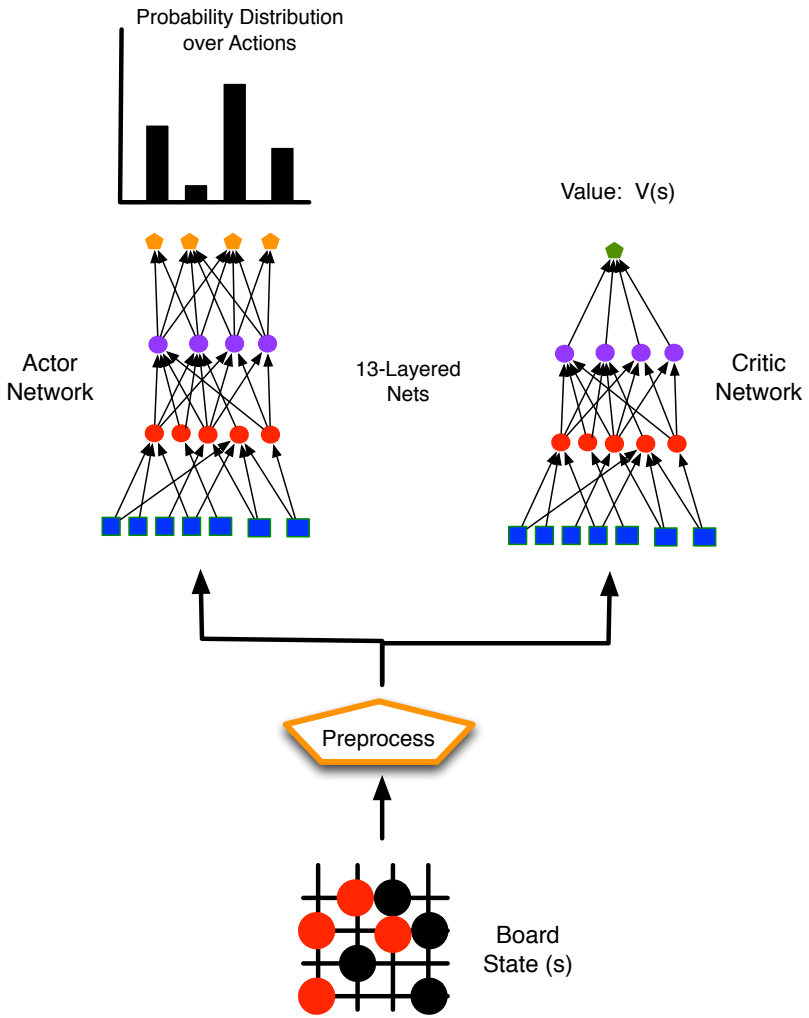
Adaptation driven by **gradients of predictions**.

TD-Gammon (Tesauro, 1995): RL + NN



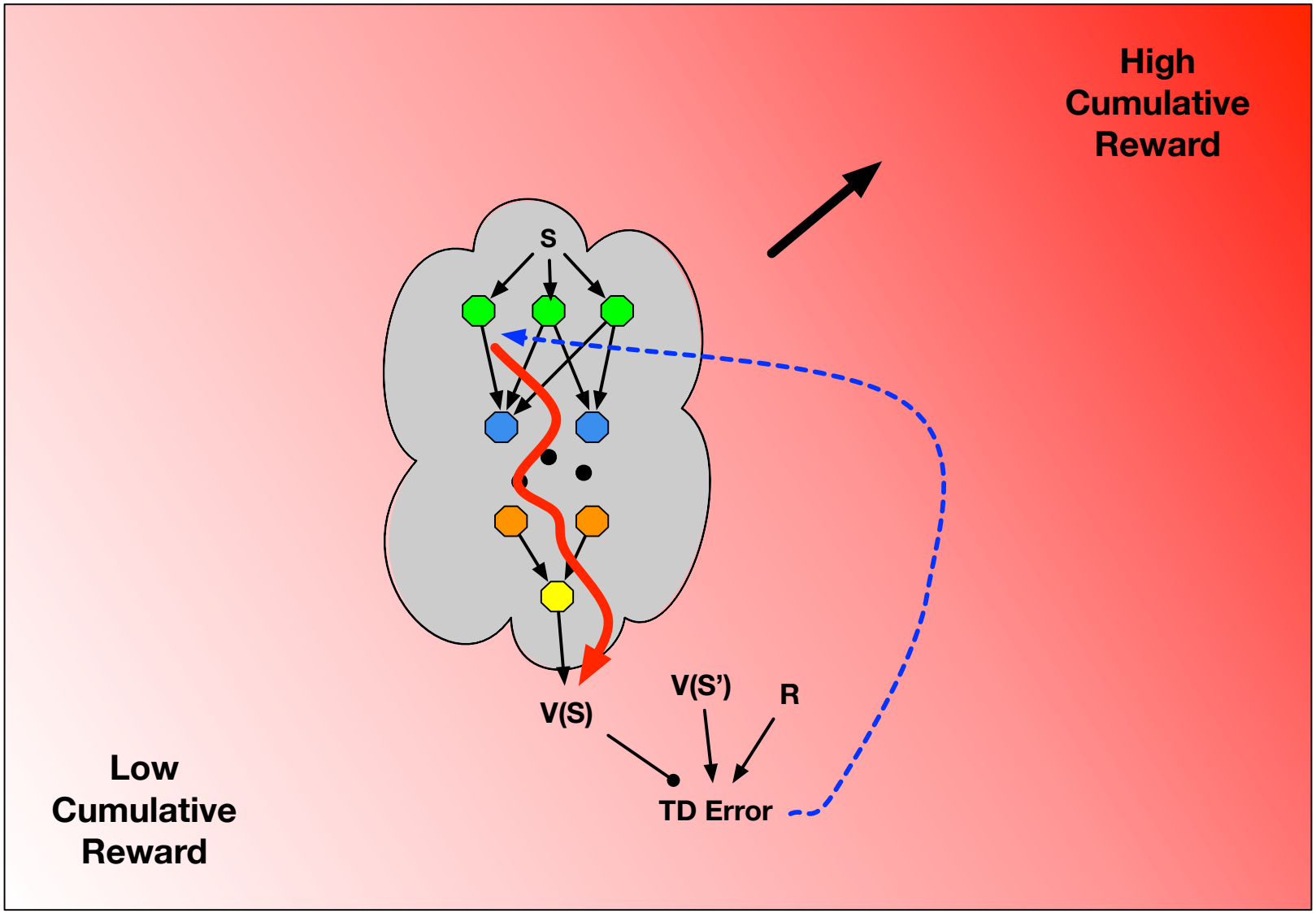
Net learns value function, V(s), using **Predicted-reward gradient:** $\frac{\triangle V(s)}{\triangle w}$

AlphaGo (DeepMind, 2016): RL + Deep Nets



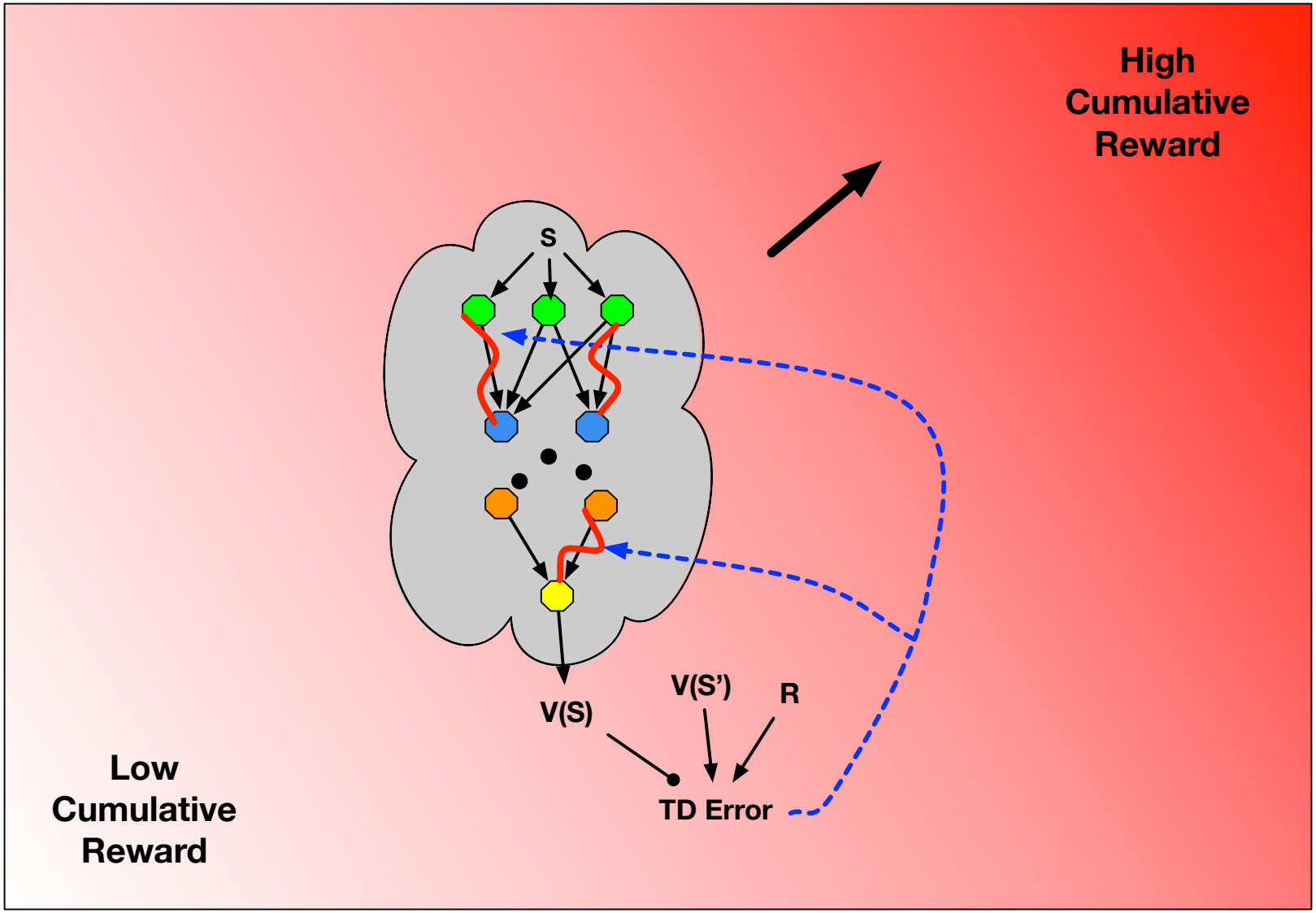
Taking Tesauro's (1995) work to a higher level with deep nets.

Gradients of Deep Reinforcement Learning



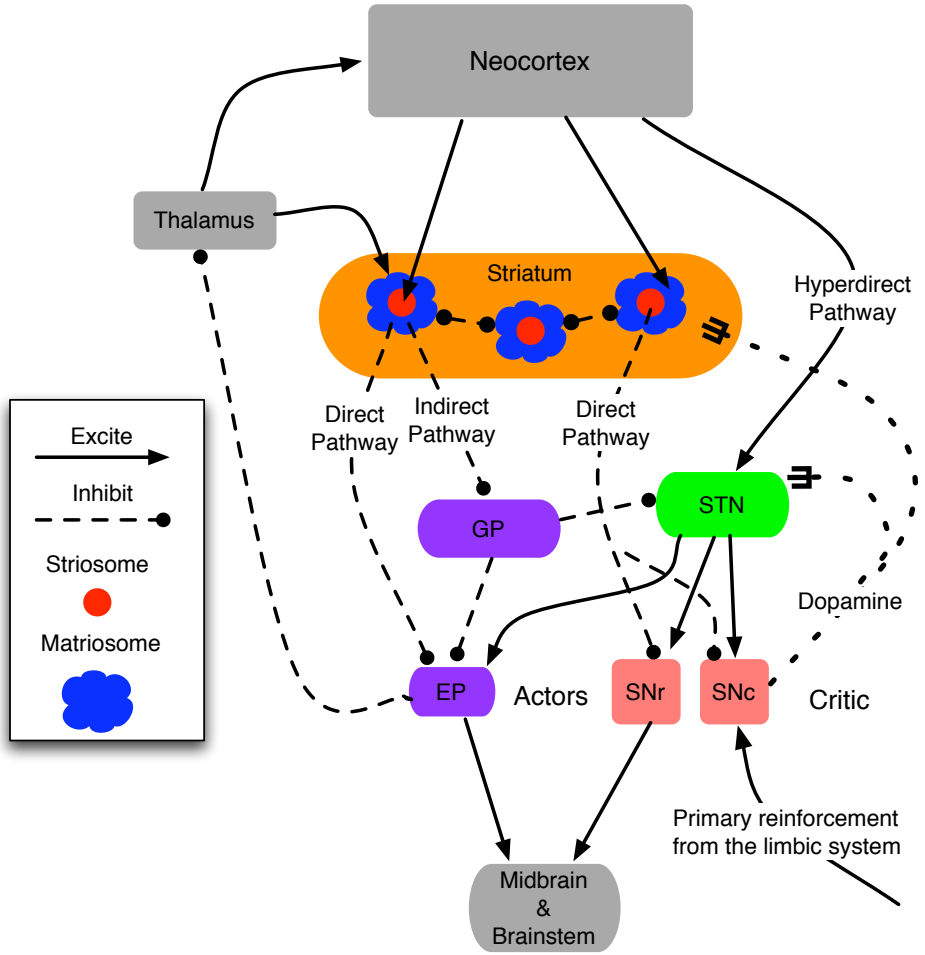
Learning via long gradients ($\frac{\triangle V(s)}{\triangle w}$) **modulated by** TD Error.

Biologically Realistic Deep RL



Hebbian learning modulated by TD Error.

RL in the Basal Ganglia

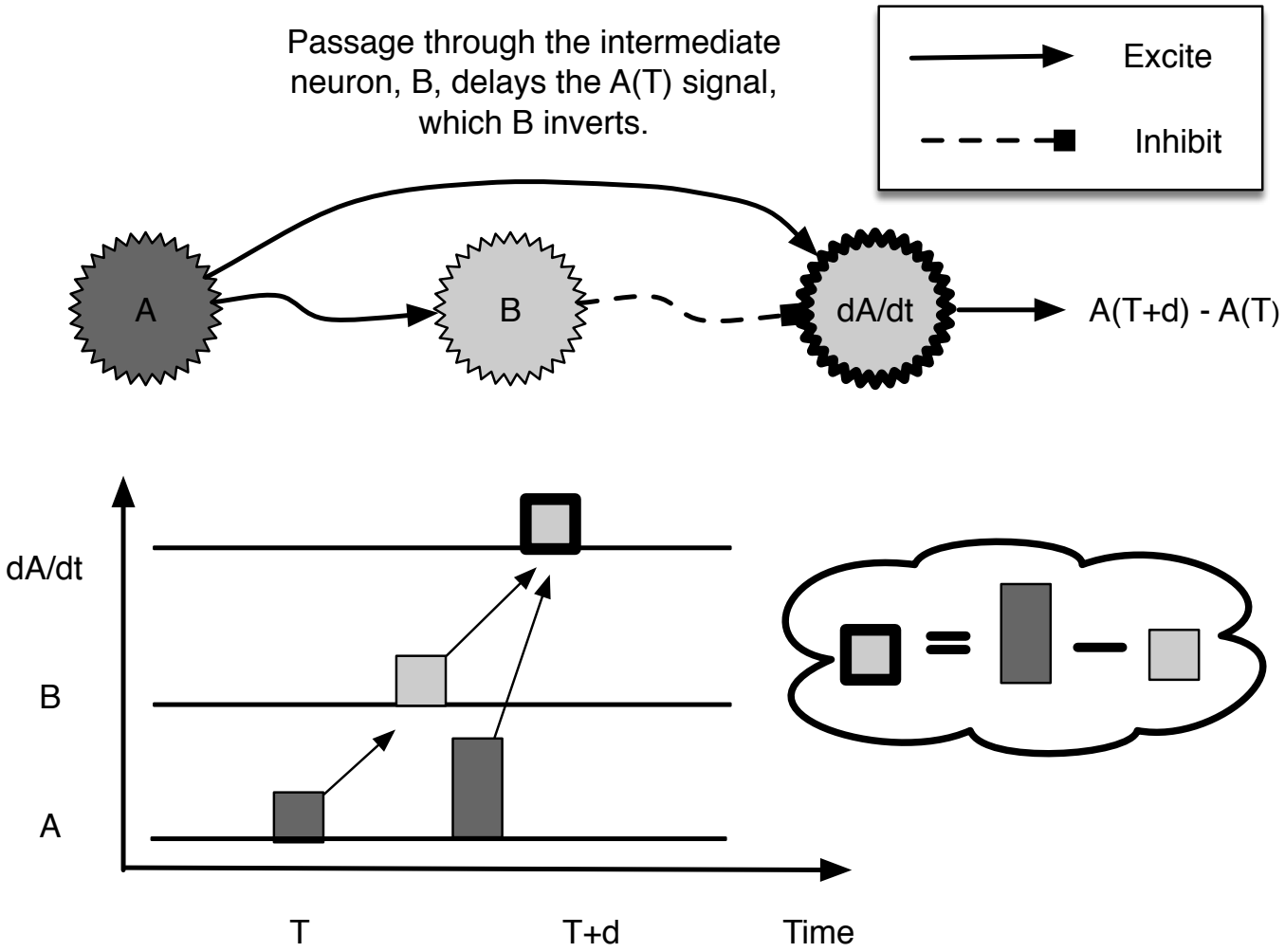


Neural computation of TD Error: $\delta = V(S_t) + R_t - V(S_{t-1})$

- Excitatory Inputs to SNc: V(S_t) (hyperdirect path) + R_t (limbic system)
- Inhibitory Inputs to SNc: V(S_{t-1}) via direct path.

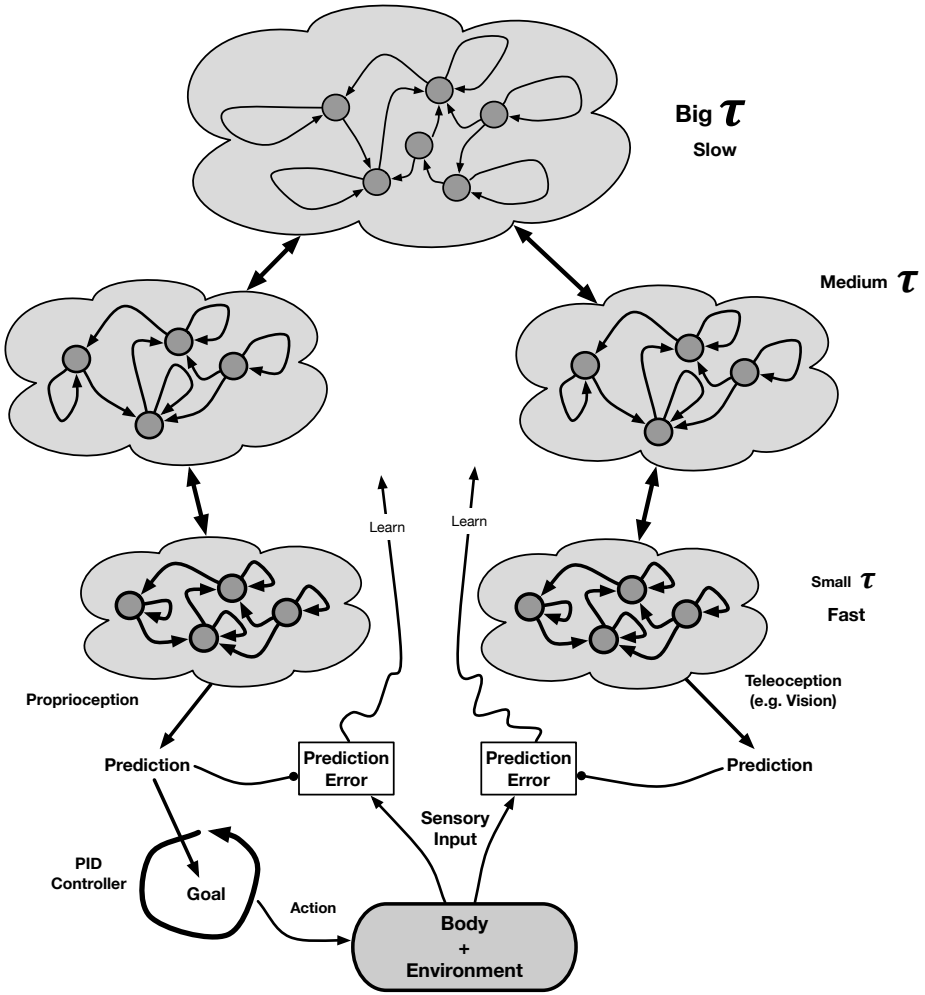
Temporal Differentiation via Delayed Inhibition

As shown earlier....



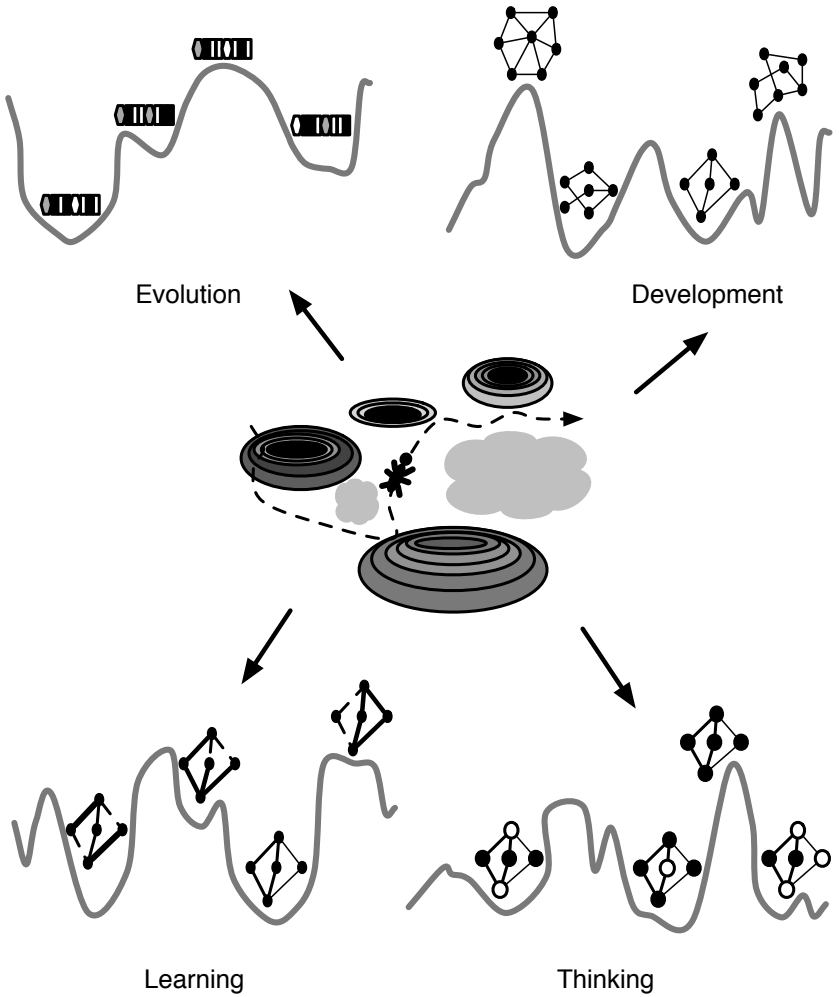
CTRNNs and MTRNNs

Continuous-time recurrent networks with multiple timescales.



Tani et. al., 2016,2017,2020; Beer et. al., 1982,2003,2015

Multiple Levels of Search and Emergence



Search and emergence at one timescale support and constrain more search and emergence at other scales.

Simulation of Emergent Prediction

Goal: Simulate the emergence of predictive networks via evolution, development and learning.

Primitives

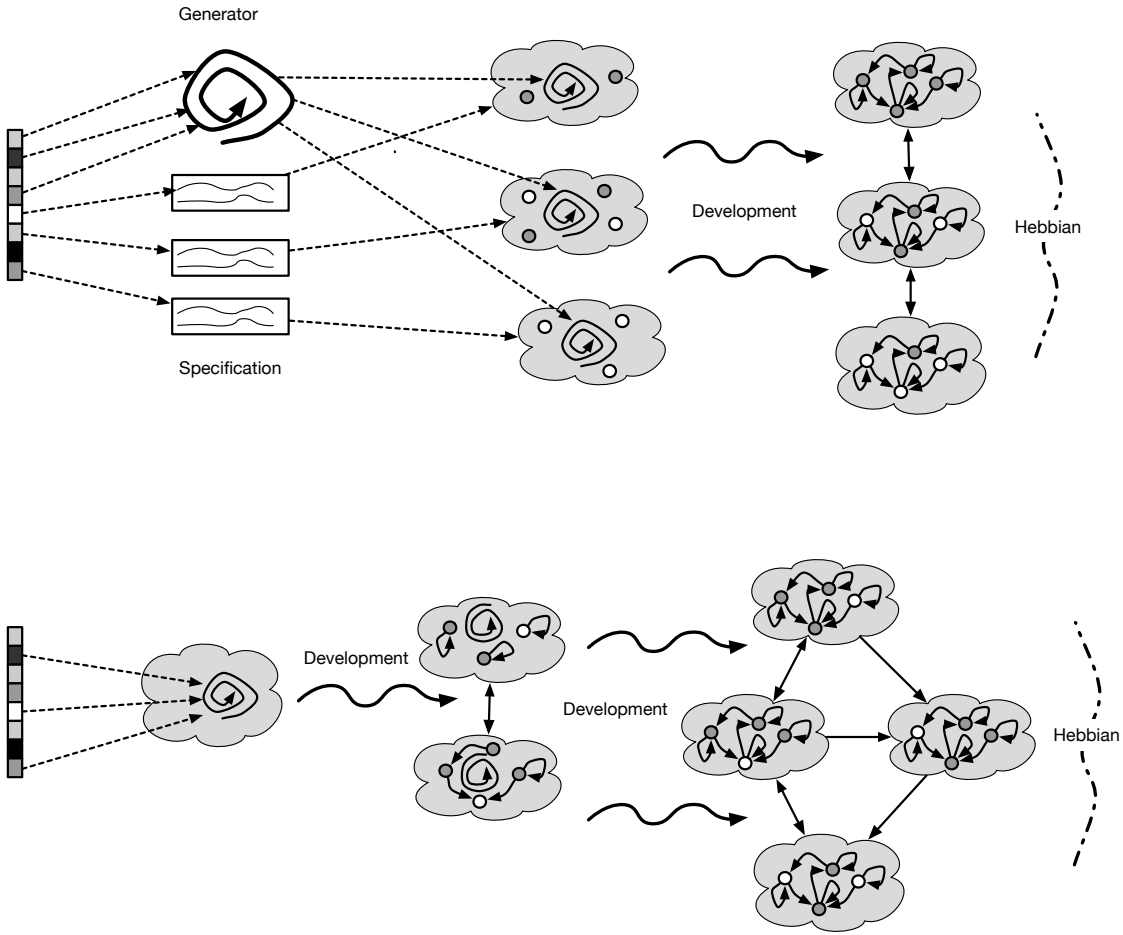
- Local gradients (derivatives)
- Integration
- Inhibition
- Comparators
- Multiple timescales
- Modular mechanisms

Motivation for a relatively unrestricted design space

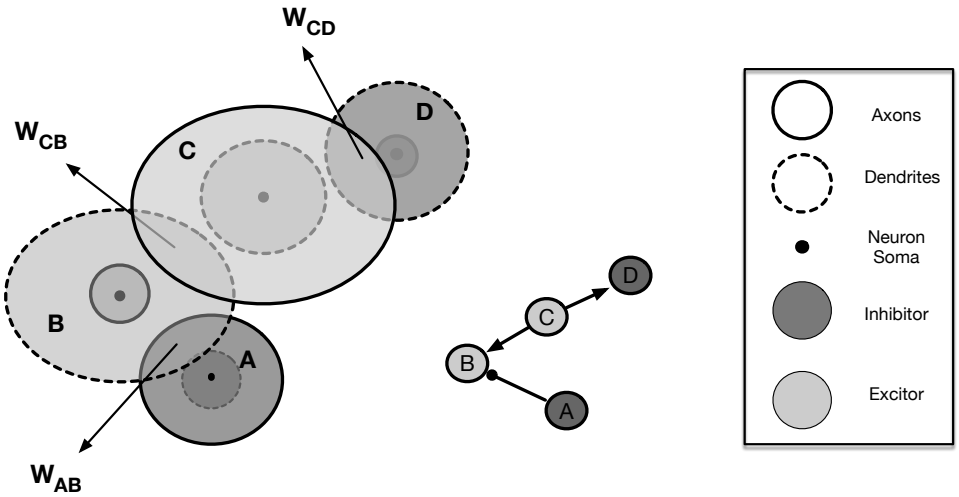
- The brain shows that there are many, diverse designs for predictors. An emergent AI system needs to freely explore.
- The mixture of competition and cooperation in evolution, development and learning may underlie the mixture of excitation and inhibition in predictive circuits.

POE Networks

POE = Phylogenetic (Evolving), Ontogenetic (Developing), Epigenetic (Learning)



The D'Arcy System



- Extension of van Ooyen et. al.'s (2003) model of activity-dependent neurite growth to embody two key mechanisms: Neural Darwinism (Edelman, 1987) and Displacement Theory (Deacon, 1998).
- Overlapping development and learning with critical periods for each.
- Evolution (for each neurite) of 2-d location, time constant, influence (excite or inhibit), axonal and dendritic growth limits, neural density, etc.
- Neural dynamics = CTRNN
- Neuromodulatory neurites affect neuron-level learning updates.

Predicting Prediction's Future in AI

Yogi Berra - American baseball legend

- It's tough to make predictions...especially about the future.
- The future ain't what it used to be.

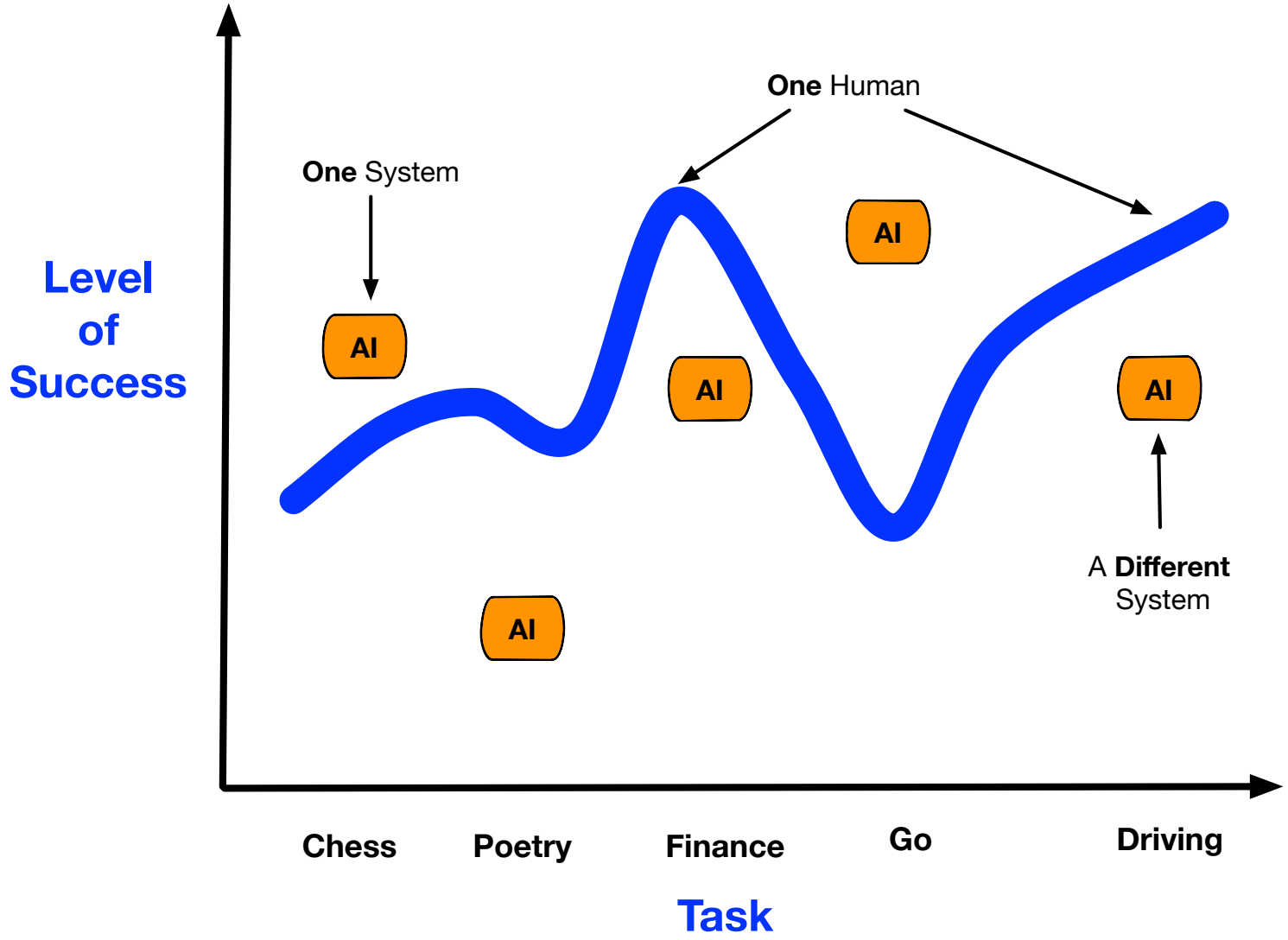
- Everything is not prediction, but prediction is everywhere.
- Can we exploit this in Bio-Inspired AI systems?
  - Energy nets (from 1980's and 90's) + Free-Energy Principle (Friston, 2010) give theoretical optimism.
  - Bogacz et. al. give practical optimism.
  - Deep Learning's success comes from big data, GPUs, etc.; POE systems now exploit these as well.
- Or is the emergent commonsense in Generative AI (e.g. ChatGPT) the best route to AGI?

Emergent Intelligence in Large Language Models

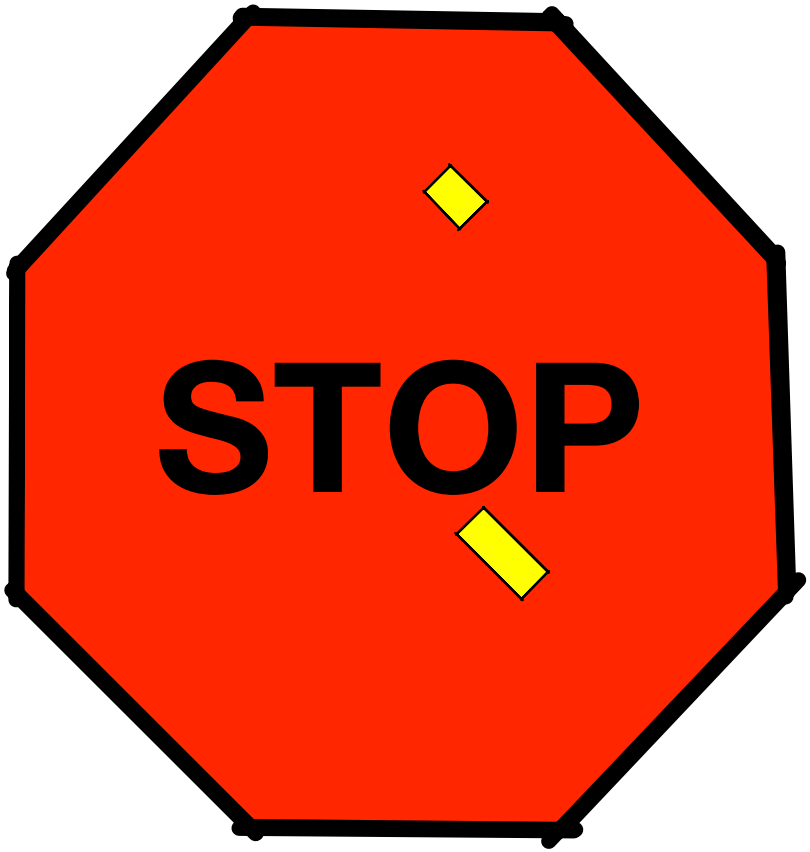
- LLMs are trained to do a predictive task.
- Prediction is a cheap way to do supervised learning, since target = next word or token.
- From this, sophisticated, intelligent behavior (and deep understanding??) seems to have emerged!
- Is prediction the foundation of emergent intelligence?
- How far can we take this prediction-based emergent intelligence toward AGI?

Goal: Artificial General Intelligence (AGI)

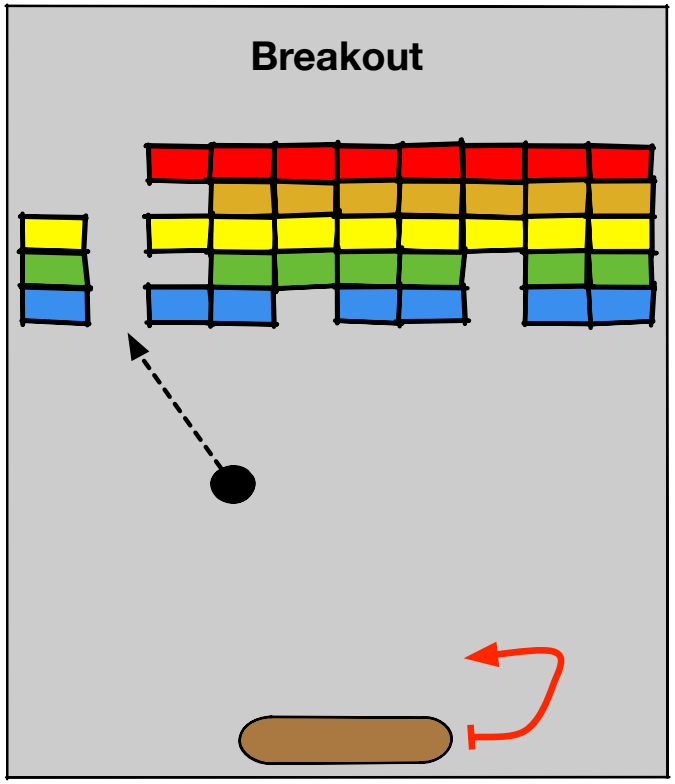
Modern AI systems are very **specific** and brittle.



Brittleness of DL and DRL

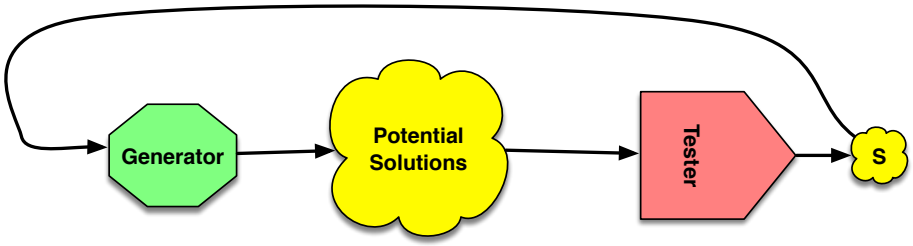


Deep Learning



Deep Reinforcement Learning

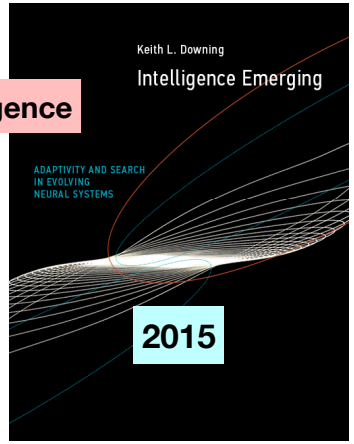
Problem-Solving Search = Generation + Testing



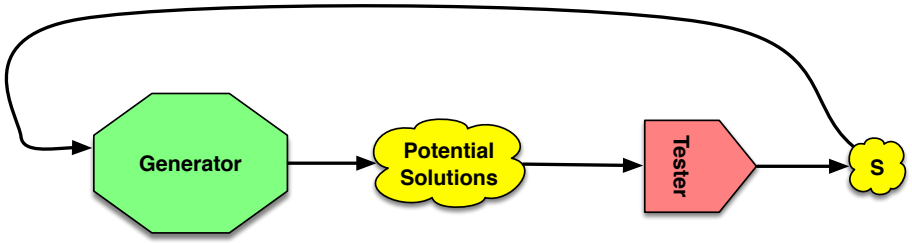
Nature and Bio-Inspired AI

Random generation but ruthless testing.

Emergence



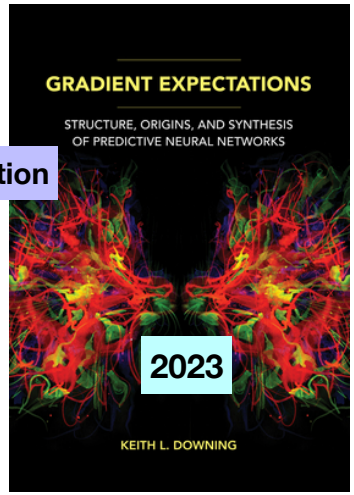
* Size of generator and tester indicates amount of *intelligence/constraint* in them.



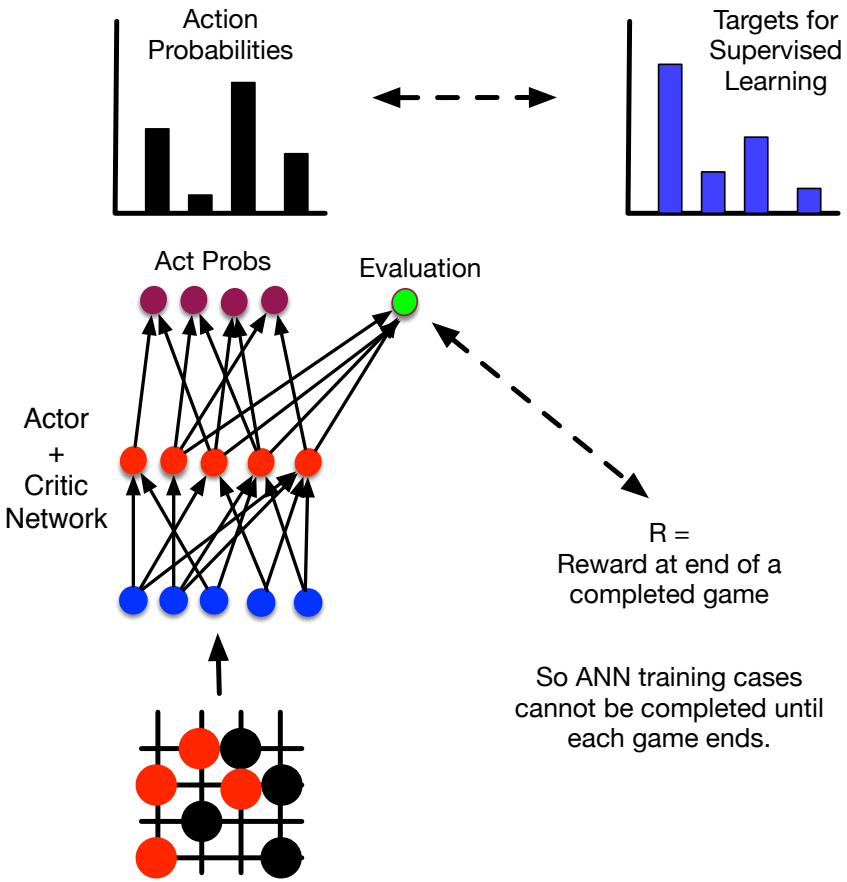
Standard AI and OR Problem Solvers

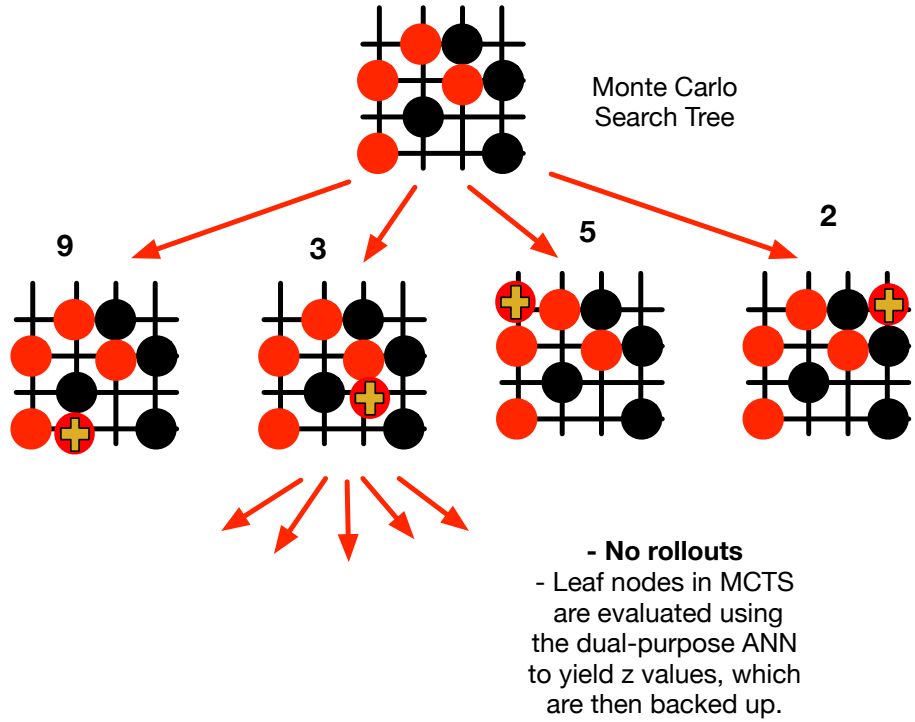
Adaptive predictions enable intelligent generation of alternatives.

Prediction



Monte Carlo Tree Search + Gradient Descent





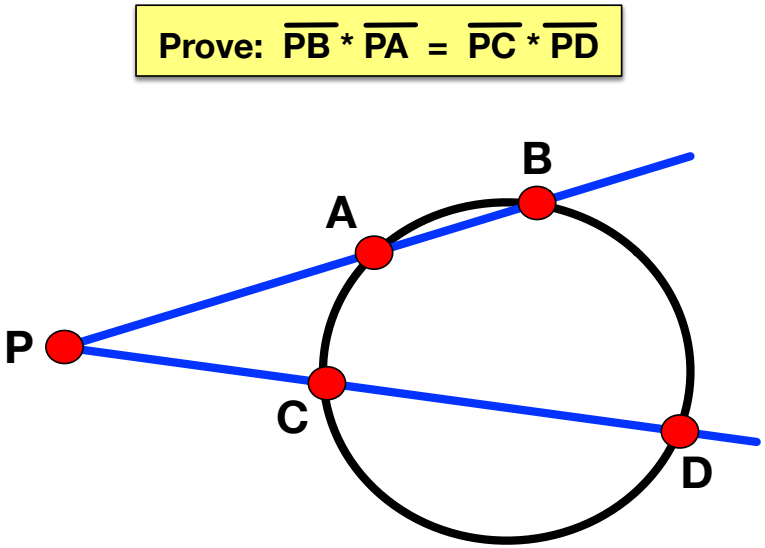
AlphaGo Zero = GOFAI + Neural Nets !!

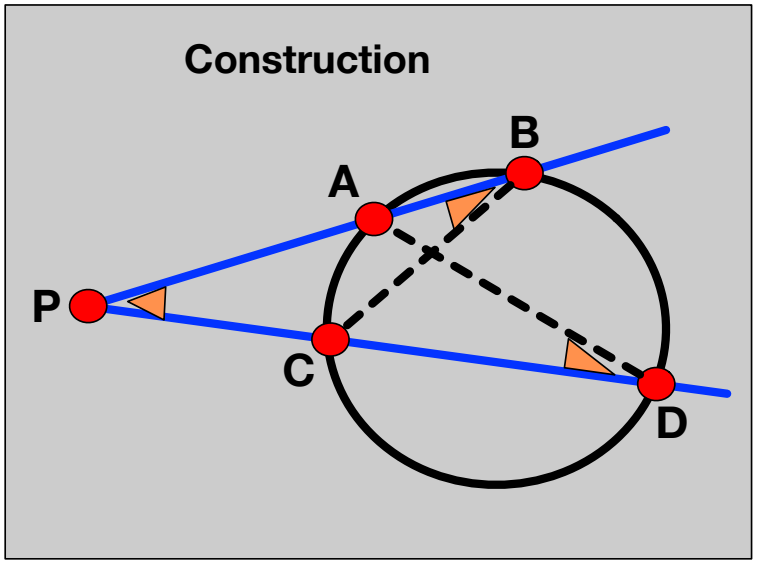
Google DeepMind's AI Breakthroughs

- **AlphaGo** - MCTS + Several NN's (some trained on expert game data). Bootstrapped intelligence via self-play. Beat world champion.
- **AlphaGo Zero** - MCTS + one main NN. No expert data, only bootstrapping from random to world-champion play. Beat AlphaGo 100 games to 0.
- **AlphaZero** - Extended AlphaGo Zero to other games and became world champion at all of them.
- **DeepNash** - Deep Reinforcement Learning to play Stratego (an imperfect information game) at world-class level. Only uses self-play.
- **AlphaFold** - Essentially solved the protein-folding problem using deep convolutional networks. Possibly AI's greatest contribution to science !!
- **AlphaGeometry** - Combines traditional AI geometry theorem proving with LLMs to achieve near gold-medalist performance in International Mathematics Olympiad.

** Most of these combine GOFAI with Deep Learning

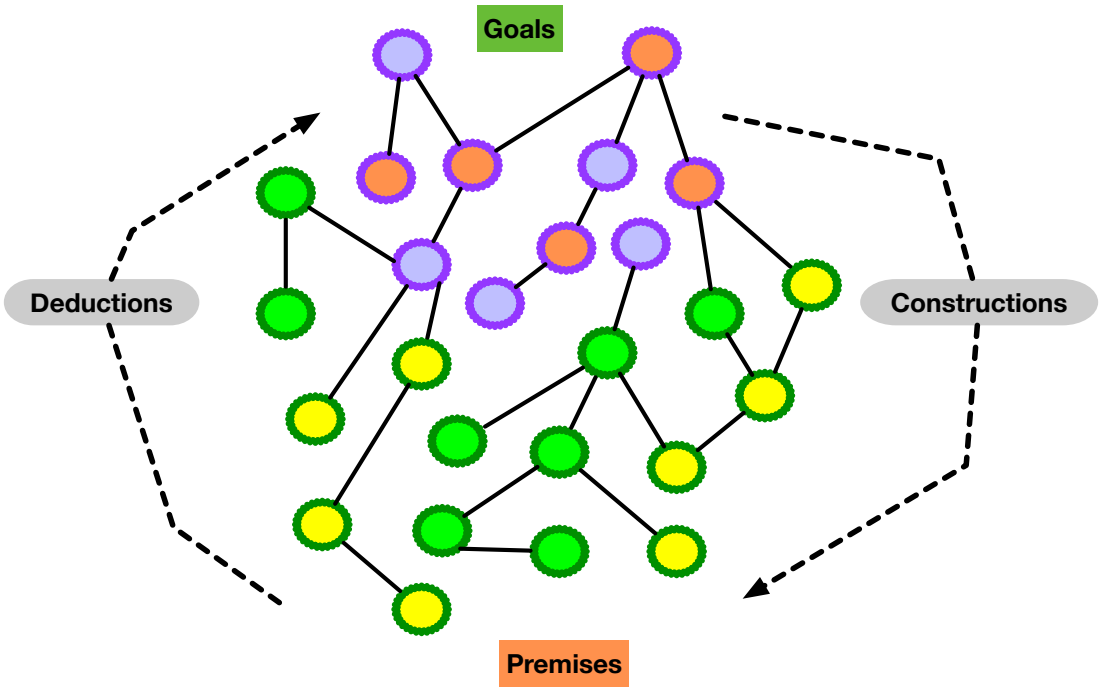
Geometry Theorem Proving: Construction = Creativity





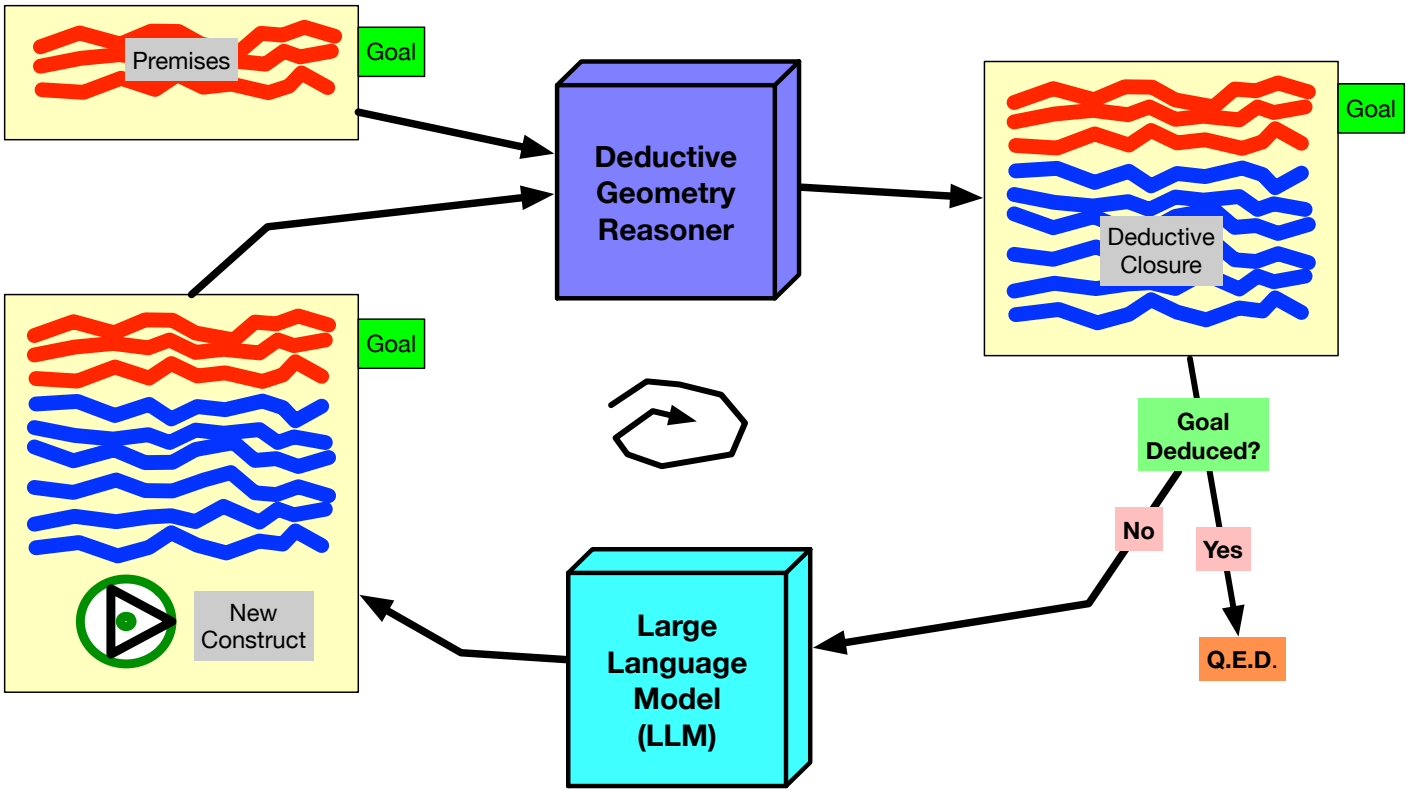
- $\measuredangle PDA \cong \measuredangle PBC$ (same subtended arc)
- $\triangle PDA \sim \triangle PBC$ (3 equal angles)
- $\frac{PB}{PC} = \frac{PD}{PA}$ (since similar triangles)
- $PB \star PA = PC \star PD$ (rearrangement) Q.E.D.

Dual-Phase Problem Solving in Mathematics



- Deduction ∼ Recognition: Infer likely consequences of the data.
- Construction ∼ Prediction: Make intelligent guesses as to which actions (changes to the data) will help achieve the goal.

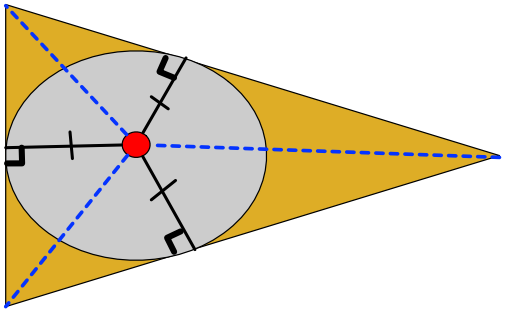
AlphaGeometry: Deduction + Construction

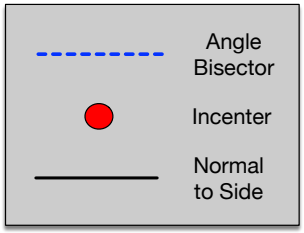


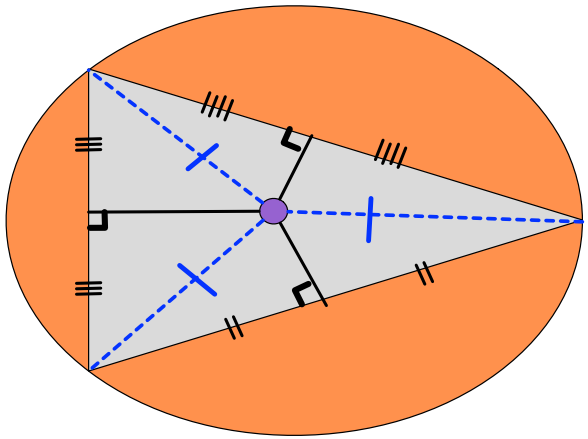
- Computing deductive closure $\approx$ a classification task: recognizing all consequences of the given facts.
- Construction = A Generating process.
- Intelligent generation = a clear sign of understanding.
- Friston's Free Energy principles: construction == prediction == action(s) to influence future observations so that they align with dominant causal hypotheses. Now, it's to align with the goal statement (to be proved).

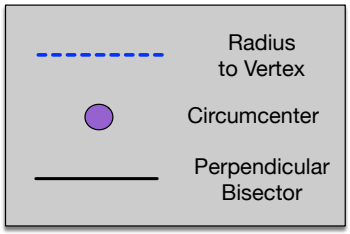
Construction -vs- Recognition

Concept	Recognition	Construction
InCenter (of InCircle)	Equidistant Sides	Angle Bisectors
Circumcenter (of CircumCircle)	Equidistant vertices	Perpendicular Bisectors









When construction replaces prediction, the two phases may use different terms.

$\rightarrow$ One phase is no longer the reverse of the other.