From Backpropagation to Brain-Like CyberInfrastructure: A Ladder of Universal Designs

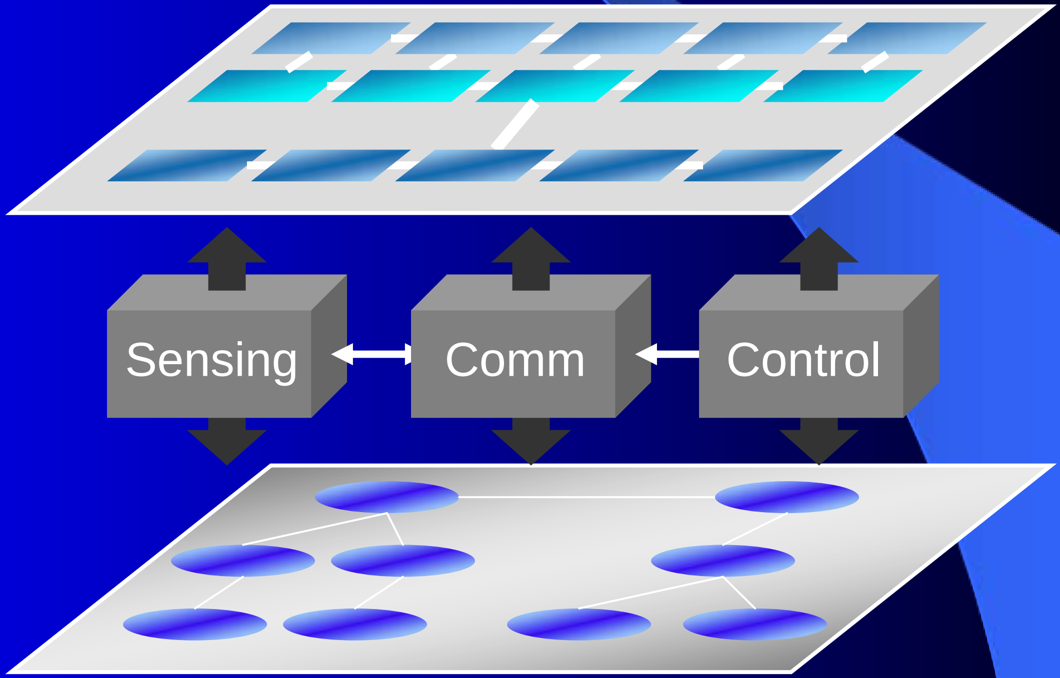
- Brain-like Cyberinfrastructure: What and Why?
- Backpropagation – Story of a Universal Tool
- A roadmap for developing mathematical designs/models but also a conceptual theory already
- Levels of Intelligence from Minsky to global mind –
  - Emergence of the 1^{st} Generation ADP Theory of Mammal Brain with two connected ladders – PREDICTION and CONTROL
  - From Two-brain theory to "3 brain" to advanced ADP

For details & equations: www.werbos.com

Cyberinfrastructure: The Entire Web From Sensors To Decisions/Actions/Control For Max Performance, “Nervous System of Global Economic Infrastructure”

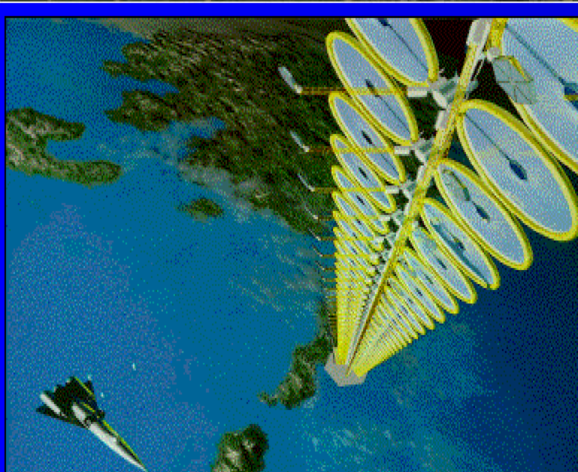
Brain-like =
General-purpose,
Adaptive,
Resilient (≠ robust),
Optimize performance
with **FORESIGHT**

Self-Configuring HW Modules



Coordinated SW Service Components

Why It is a Life-or-Death Issue



HOW?

- www.ieeeusa.org/policy/energy_strategy.ppt
- Photo credit IEEE Spectrum

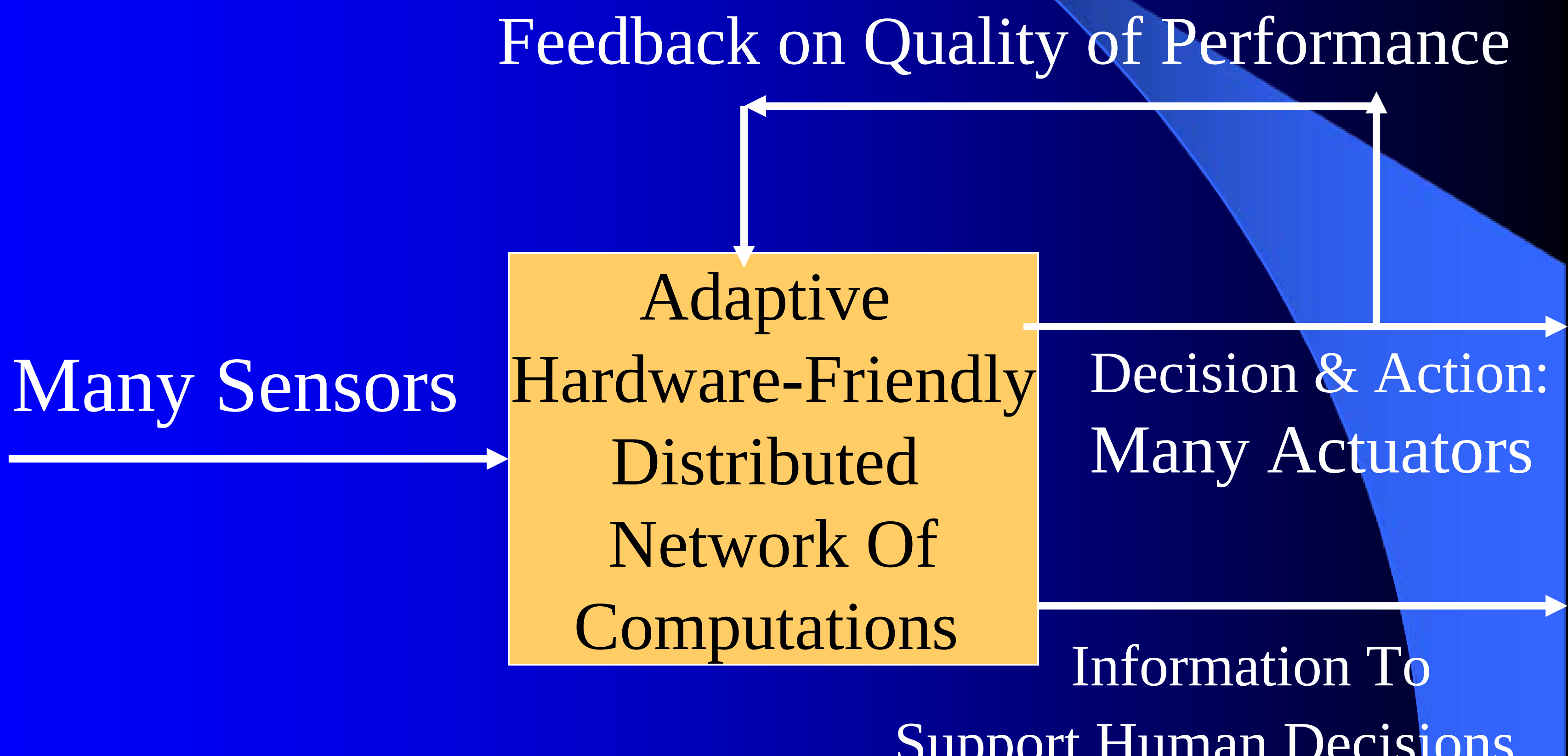
As Gas Prices ⇑ Imports ⇑ & Nuclear Tech in unstable areas ⇑, human extinction is a serious risk. Need to move faster.

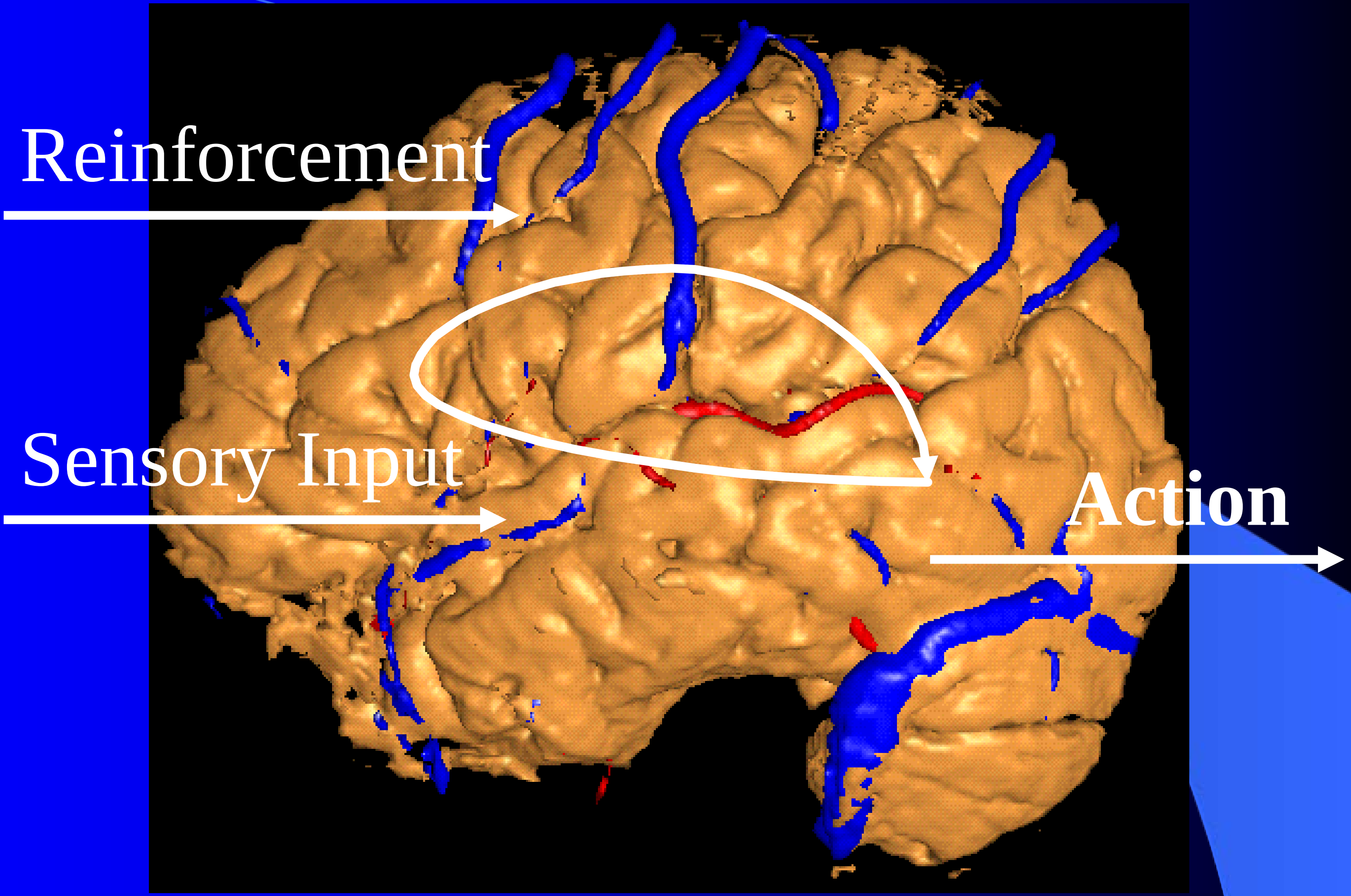
Optimal time-shifting – big boost to rapid adjustment, $

Why It Requires Artificial Neural Networks (ANNs)

- For optimal performance in the general nonlinear case (nonlinear control strategies, state estimators, predictors, etc…), we need to adaptively estimate nonlinear functions. Thus we must use universal nonlinear function approximators.
- Barron (Yale) proved basic ANNs (MLP) much better than Taylor series, RBF, etc., to approximate smooth functions of many inputs. Similar theorems for approximating dynamic systems, etc., especially with more advanced, more powerful, MLP-like ANNs.
- ANNs more “chip-friendly” by definition: Mosaix chips, CNN here today, for embedded apps, massive thruput

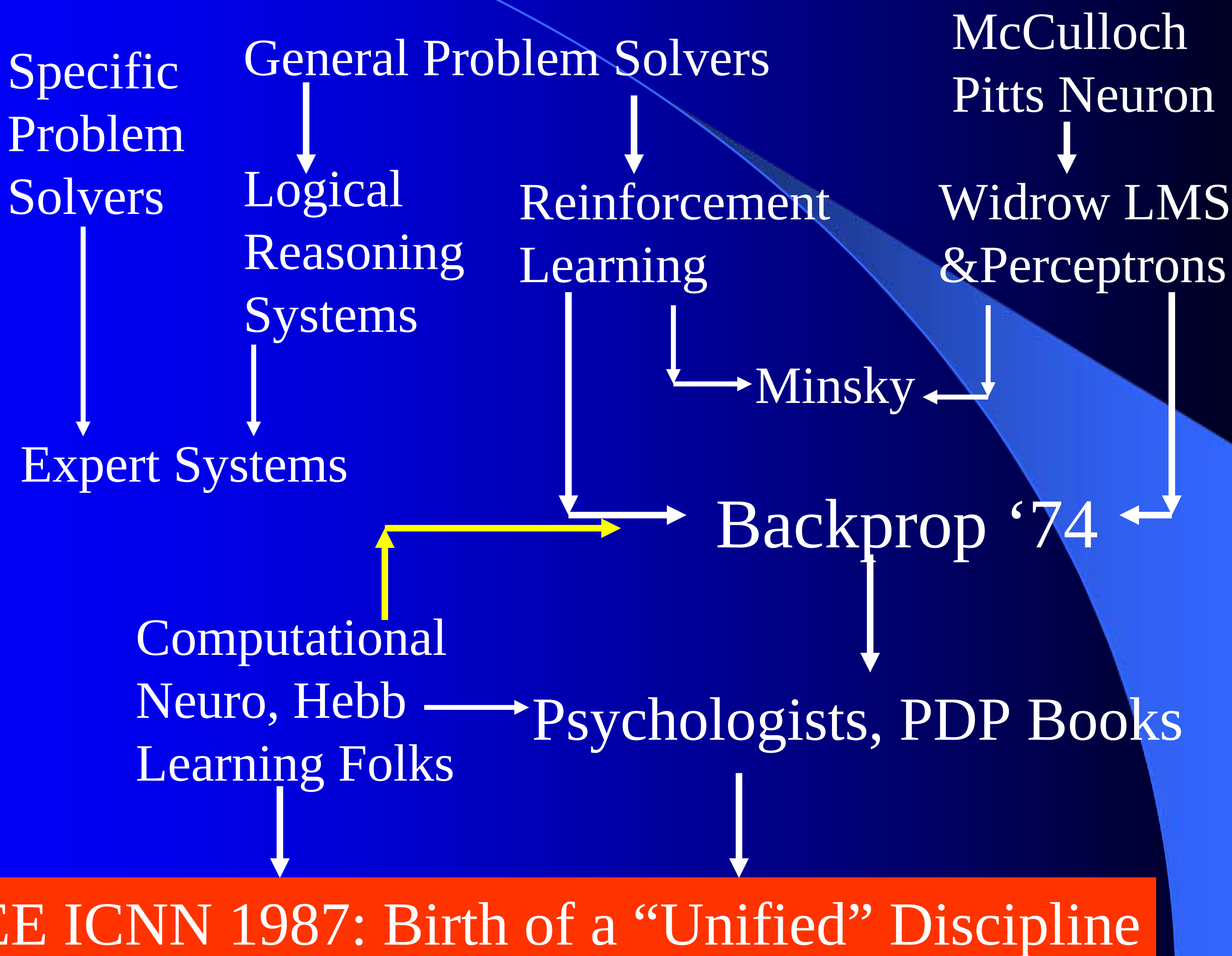
Main Goal for Neural Networks In Future Research: Unified General-Purpose Intelligence



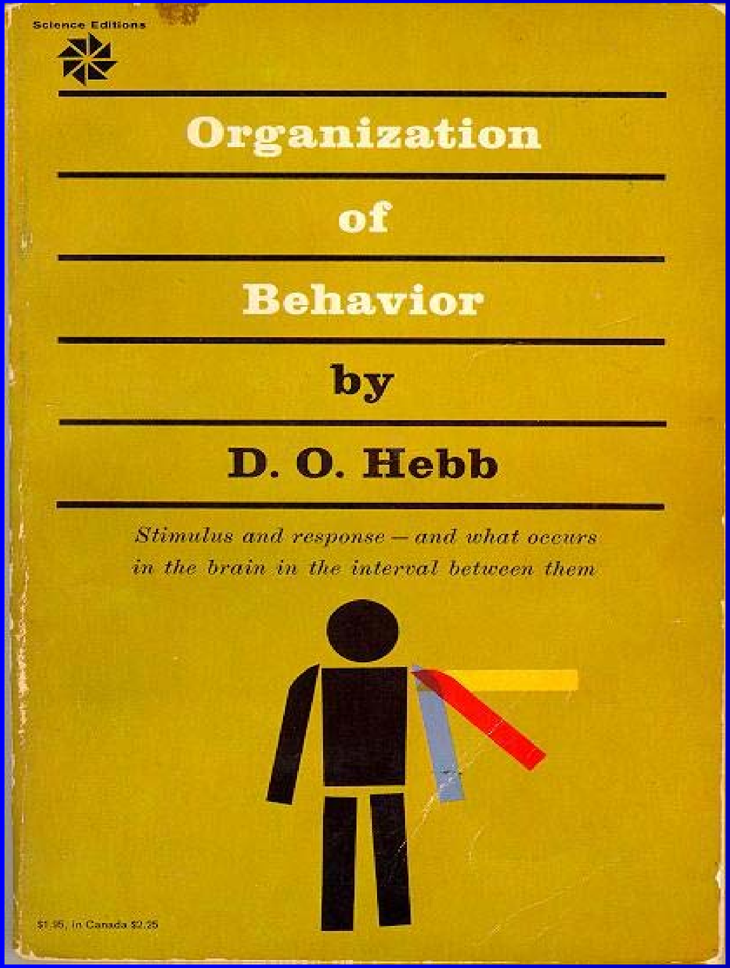


The Brain As a Whole System Is an Intelligent Controller

Where Did ANNs Come From?



Hebb 1949: Intelligence As An Emergent Phenomenon or Learning



"The general idea is an old one, that any two cells or systems of cells that are especially active at the same time will tend to become 'associated,' so that activity in one facilitates activity in the other" -- p.70 (Wiley 1961 printing)

The search for the General Neuron Model (of Learning)

"Solves all problems"

Claim (1964) : Hebb's Approach Doesn't Quite Work As Stated

- Hebbian Learning Rules Are All Based on Correlation Coefficients
- Good Associative Memory: one component of the larger brain (Kohonen, ART, Hassoun)
- Linear decorrelators and predictors
- Hopfield f(u) minimizers never scaled, but:
  - Gursel Serpen and SRN minimizers
  - Brain-Like Stochastic Search (Needs R&D)

Understanding Brain Requires Models Tested/Developed Using Multiple Sources of Info

- Engineering: Will it work? Mathematics understandable, generic?
- Psychology: Connectionist cognitive science, animal learning, folk psychology
- Neuroscience: computational neuroscience
- AI: agents, games (backgammon, go), etc.
- LIS and CRI

Maximizing utility over time

Model of reality

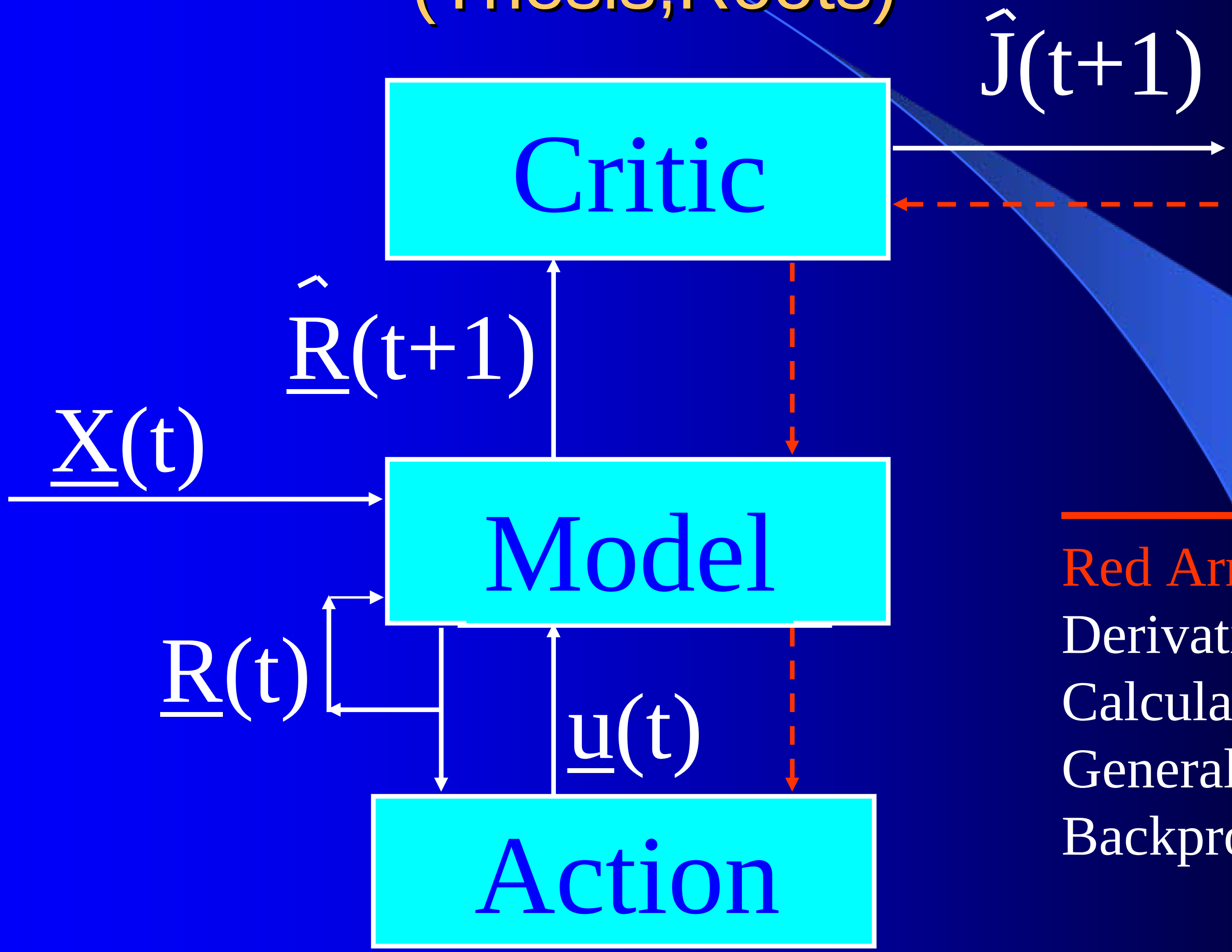
Utility function U

Dynamic programming

$$J(\mathbf{x}(t)) = \underset{\mathbf{u}(t)}{\mathbf{Max}} \langle U(\mathbf{x}(t), \mathbf{u}(t)) + J(\mathbf{x}(t+1)) \rangle / (1+r)$$

Secondary, or strategic utility function J

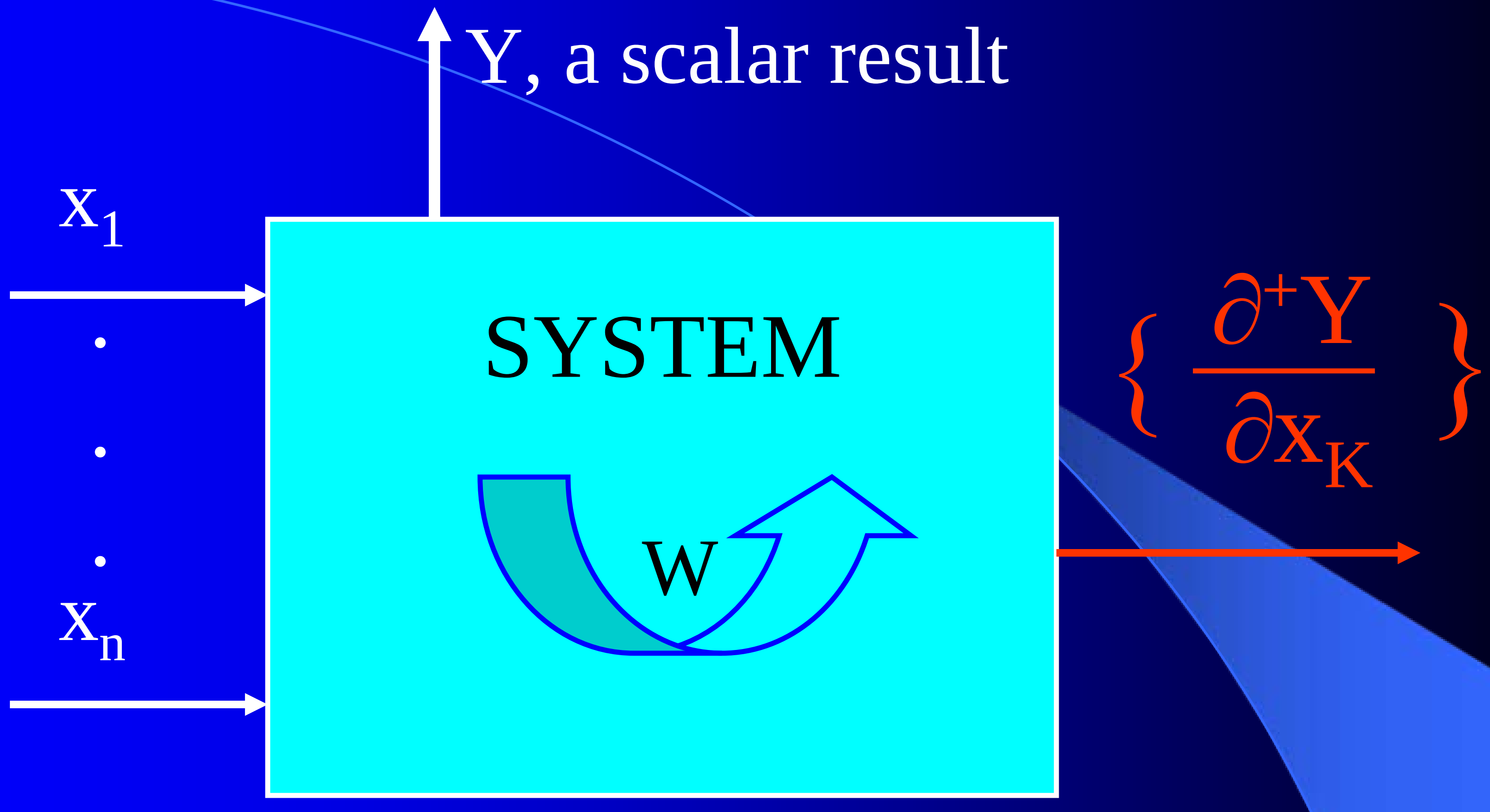
1971-2: Emergent Intelligence Is Possible If We Allow Three Types of Neuron (Thesis,Roots)



Red Arrows: Derivatives Calculated By Generalized Backpropagation

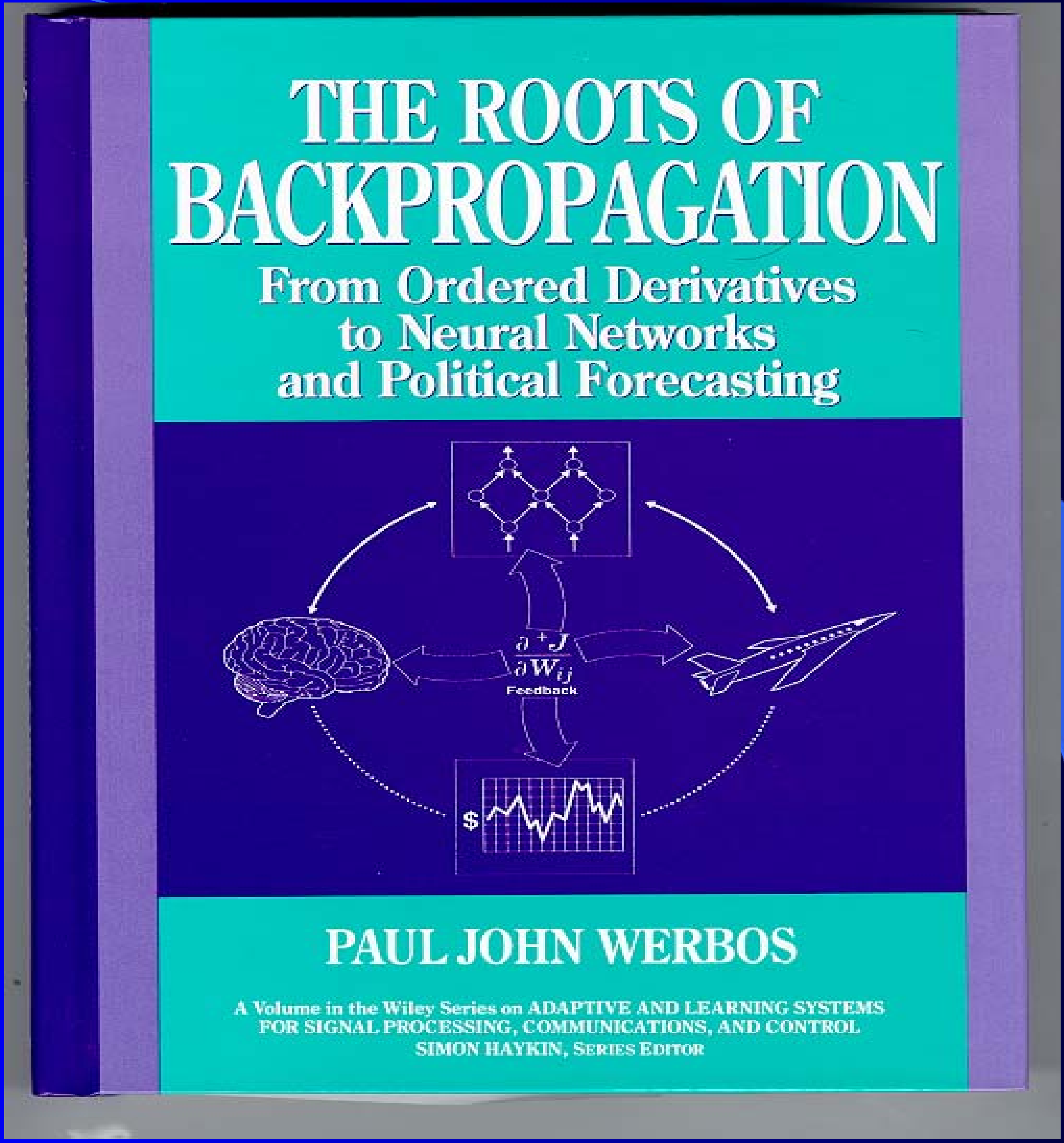
Harvard Committee Response

- We don’t believe in neural networks – see Minsky (Anderson&Rosenfeld, Talking Nets)
- Prove that your backwards differentiation works. (That is enough for a PhD thesis.) The critic/DP stuff published in ’77,’79,’81,’87..
- Applied to affordable vector ARMA statistical estimation, general TSP package, and robust political forecasting

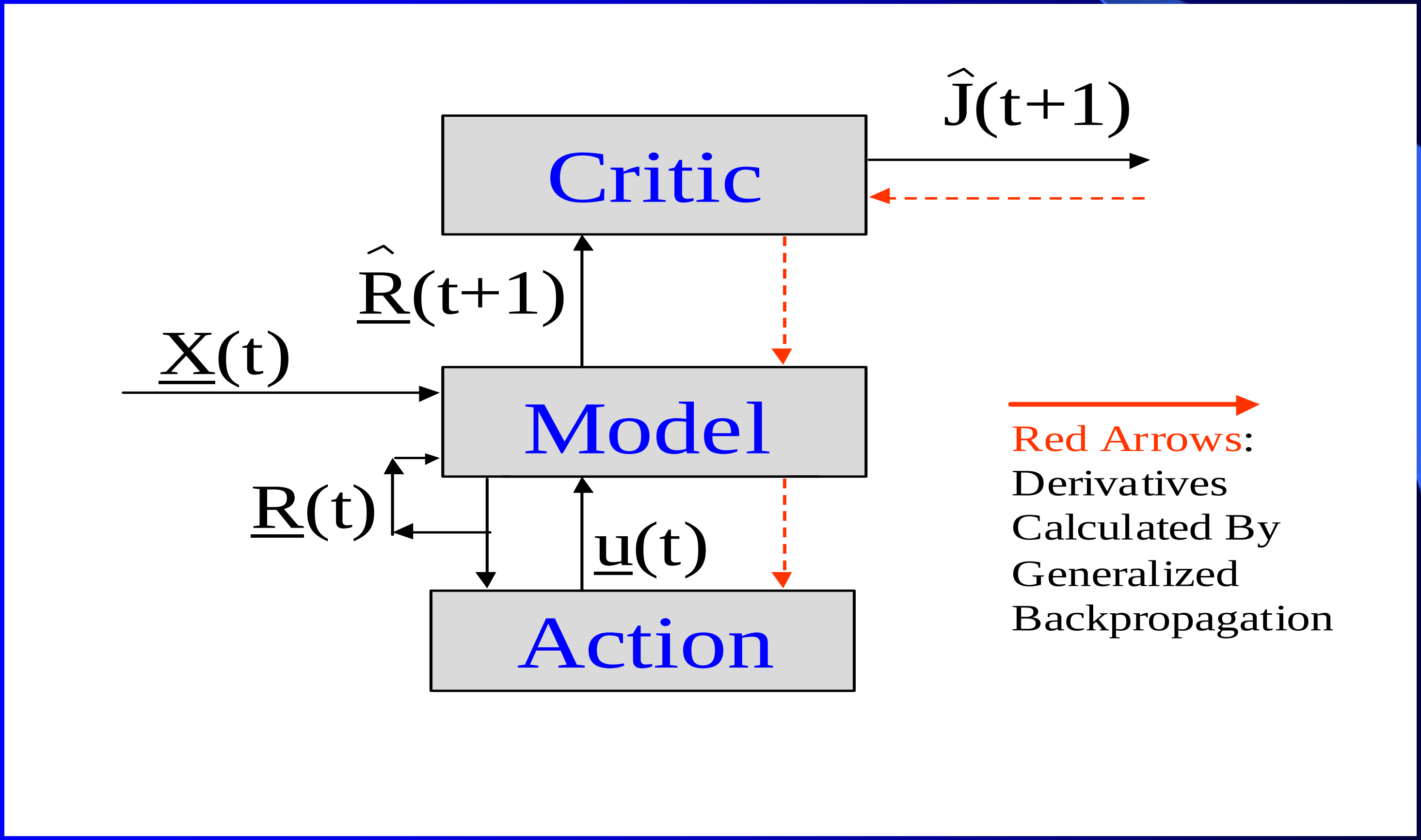


(Inputs x_k may actually come from many times)

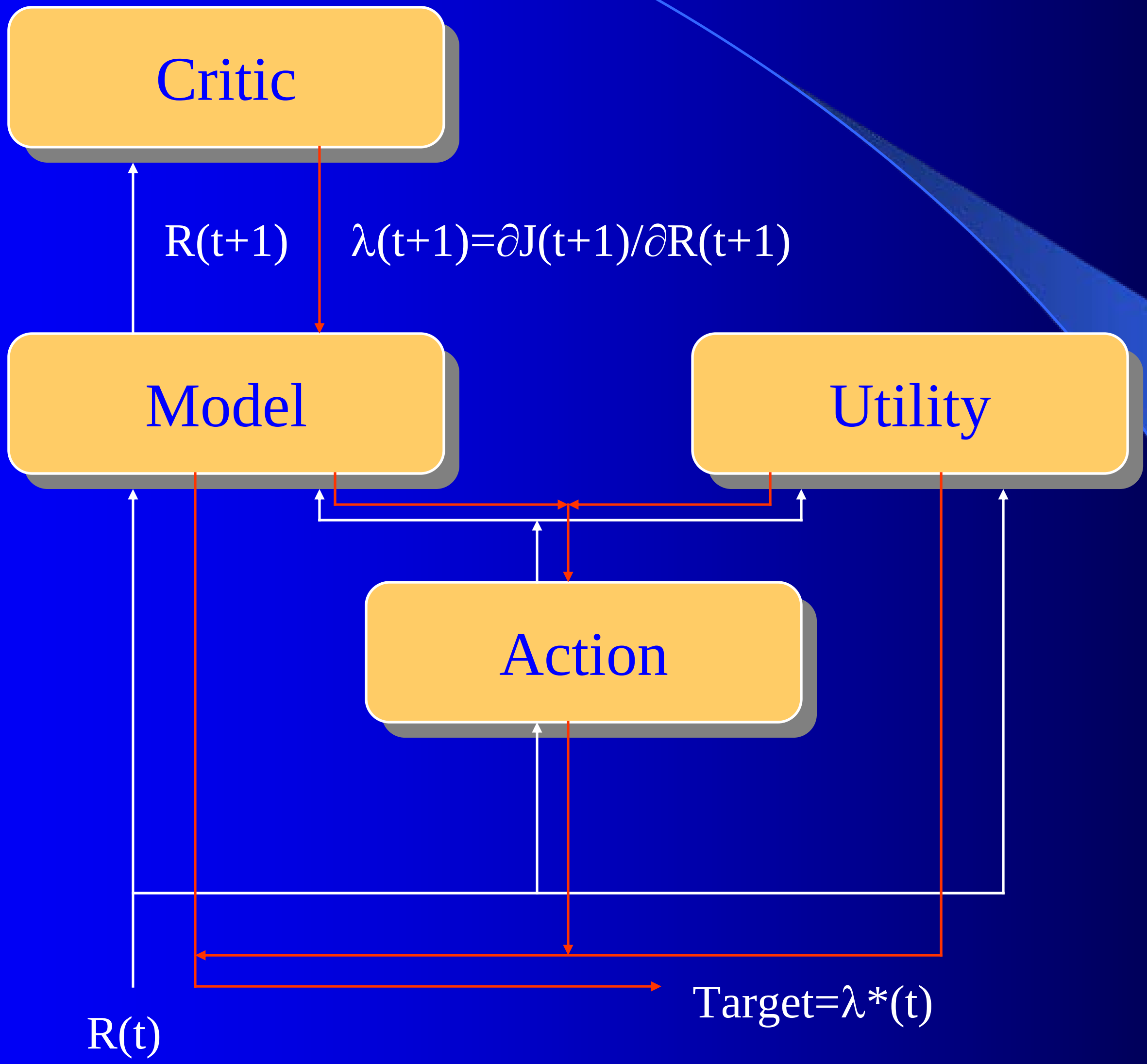
Backwards Differentiation: But what kinds of SYSTEM can we handle? See details in AD2004 Proceedings, Springer, in press.


THE ROOTS OF
BACKPROPAGATION
From Ordered Derivatives
to Neural Networks
and Political Forecasting
$\frac{\partial^+ J}{\partial W_{ij}}$
Feedback
$
PAUL JOHN WERBOS
A Volume in the Wiley Series on ADAPTIVE AND LEARNING SYSTEMS
FOR SIGNAL PROCESSING, COMMUNICATIONS, AND CONTROL
SIMON HAYKIN, SERIES EDITOR

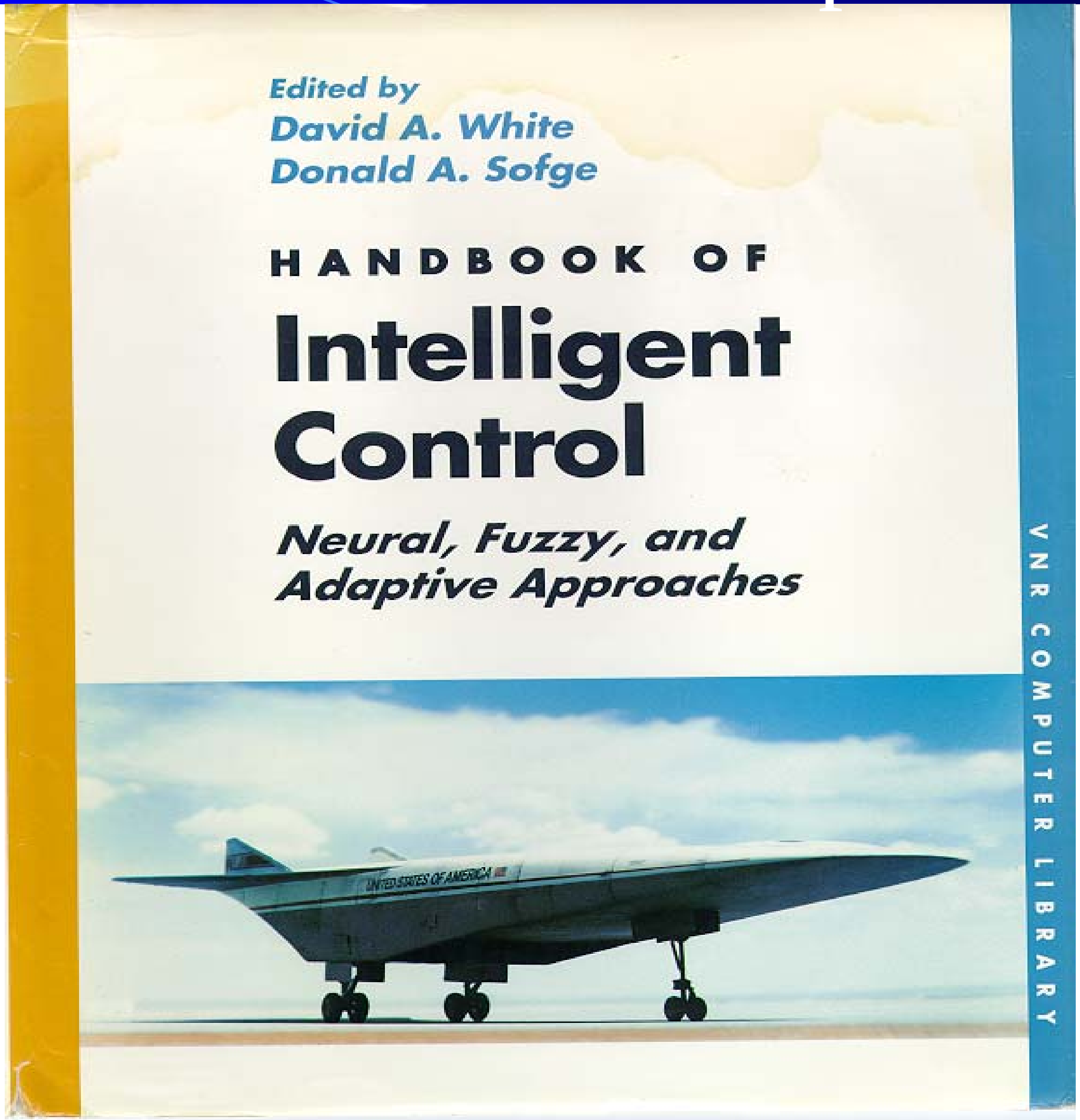
To Fill IN the Boxes:
(1) NEUROCONTROL, to Fill in Critic or Action;
(2) System Identification or Prediction (Neuroidentification) to Fill In Model



Dual Heuristic Programming (DHP)



NSF/McAir Workshop 1990



White and Sofge eds, Van Nostrand, 1992

1st Generation Theory of Mammal Brain

- As in 71-72 proposal, brain has 3 main parts:
  - Cortex+thalamus: Model to predict/impute reality. See Nicolelis&Chapin, Science, rat whisker work.
  - Limbic system: Critic gives "emotional" assessment of what Freud called "objects" (Papez, James Olds)
  - Brain-stem: action or "motor" system (and inherited fixed preprocessors/postprocessors)
  - Clock signals from extracortical sources (Foote, Llinas)
  - Backprop unavoidable. (Bliss, Spruston, Sejnowski)
- Technical level improvements and big runs enough to span gap form 1971-72 to mammal brain:
  - Fill in "Model" with hybrid Simultaneous/Time-Lagged Recurrent Network trained by Error Critic (fully specified in Handbook of Intelligent Control)
  - Critic is sum of multiple "HDP" components each trained by GDHP, which gives power of DHP for continuous variables but handles continuous/discrete mix.
  - In each box, faster learning, per robust statistics, learning from memory, etc.
- BUT IS IT ENOUGH? For what?

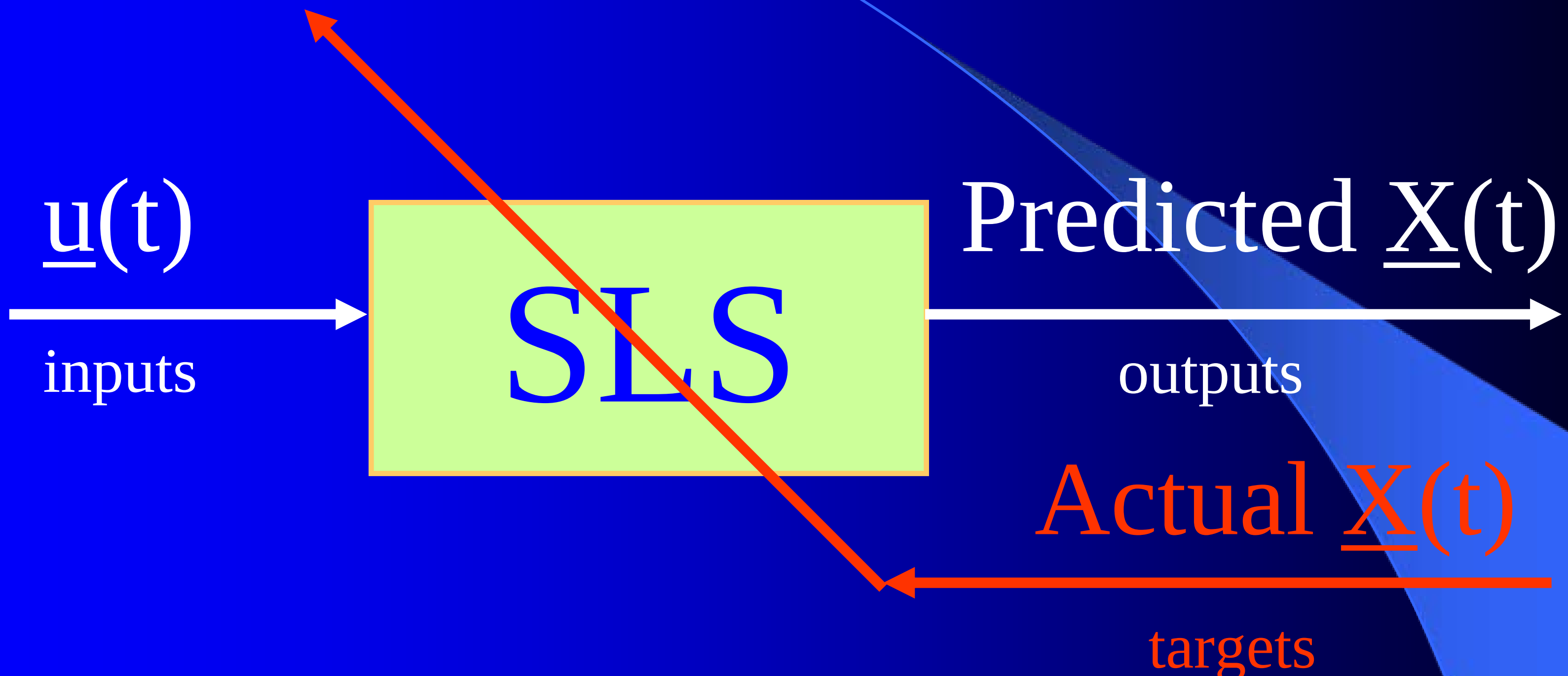
Neural Networks That Actually Work In Diagnostics, Prediction & Control: Common Misconceptions Vs. Real-World Success (excerpts from tutorial at www.werbos.com)

- Neural Nets, A Route to Learning/Intelligence
  - goals, history, basic concepts, consciousness
- State of the Art -- Working Tools Vs. Toys and Fads
  - static prediction/classification
  - dynamic prediction/classification
  - control: cloning experts, tracking, optimization
- Advanced Brain-Like Capabilities & Grids

3 Types of Diagnostic System

- All 3 train predictors, use sensor data $\underline{X}(t)$, other data $\underline{u}(t)$, fault classifications F_1 to F_m
- Type 1: predict $F_i(t)$ from $\underline{X}(t)$, $\underline{u}(t)$, MEMORY
- Others: first train to predict $\underline{X}(t+1)$ from $\underline{X}$,$\underline{u}$,MEM
  - Type 2: when actual $\underline{X}(t+1)$ 6σ from prediction, ALARM
  - Type 3: if prediction net predicts BAD $\underline{X}(t+T)$, ALARM
- Combination best. See PJW in Maren, ed, *Handbook Neural Computing Apps*, Academic, 1990.

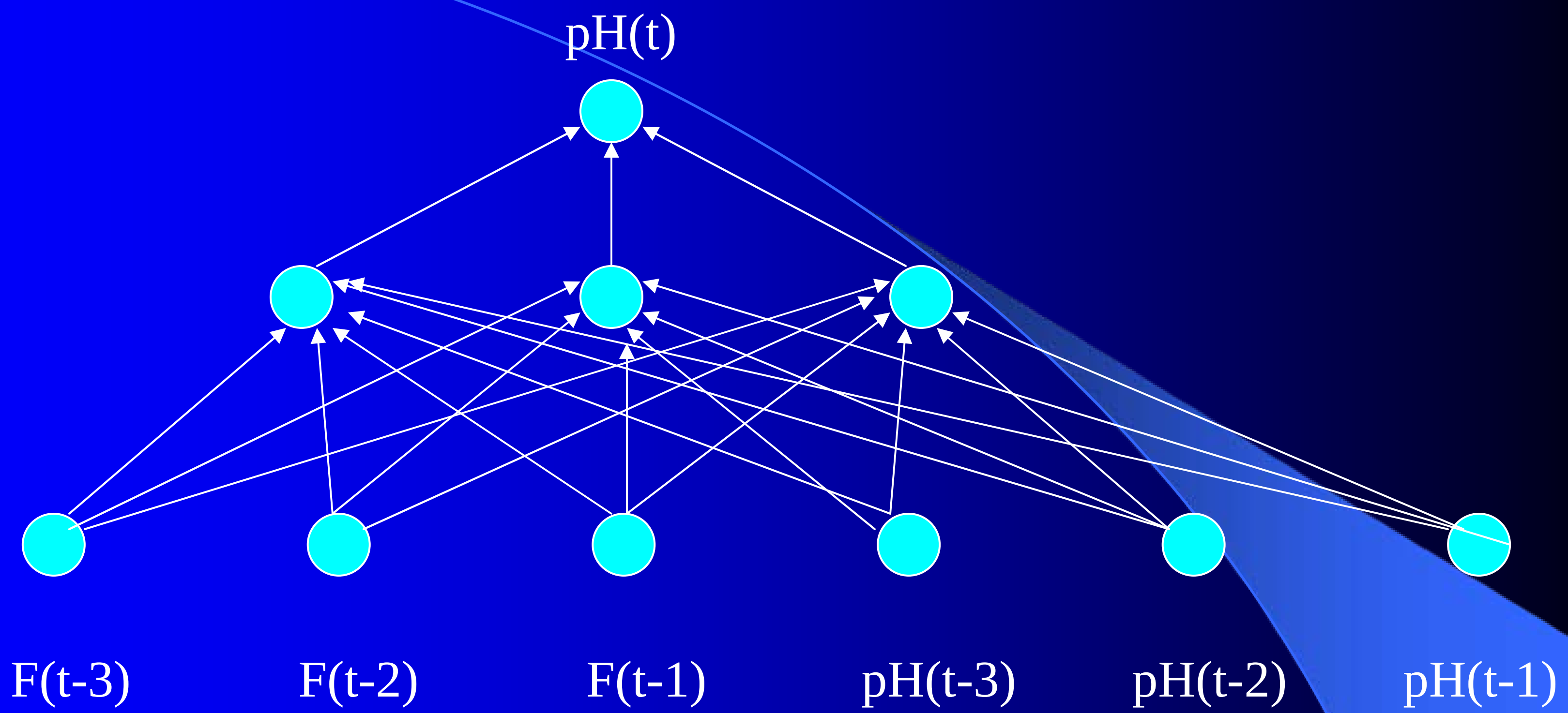
Supervised Learning Systems (SLS)



SLS may have internal dynamics but no “memory” of times t-1, t-2...

Brain-Style Prediction Is NOT Just Time-Series Statistics!

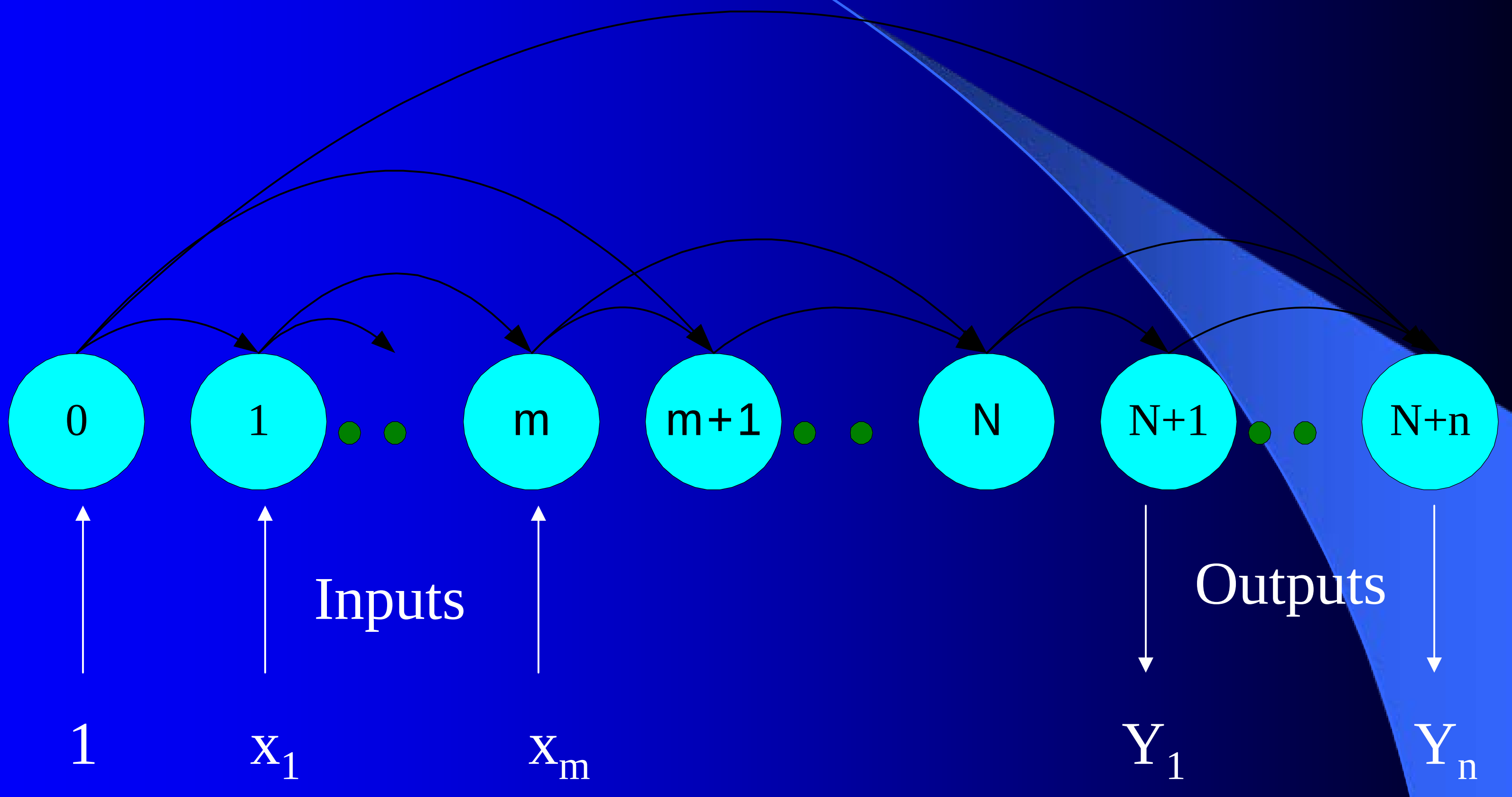
- One System does it all -- not just a collection of chapters or methods
- Domain-specific info is 2-edged sword:
  - – need to use it; need to be able to do without it
- Neural Nets demand/inspire new work on general-purpose prior probabilities and on dynamic robustness (See HIC chapter 10)
- SEDP&Kohonen: general nonlinear stochastic ID of partially observed systems



Example of TDNN used in HIC, Chapter 10

TDNNs learn NARX or FIR Models, not NARMAX or IIR

Generalized MLP



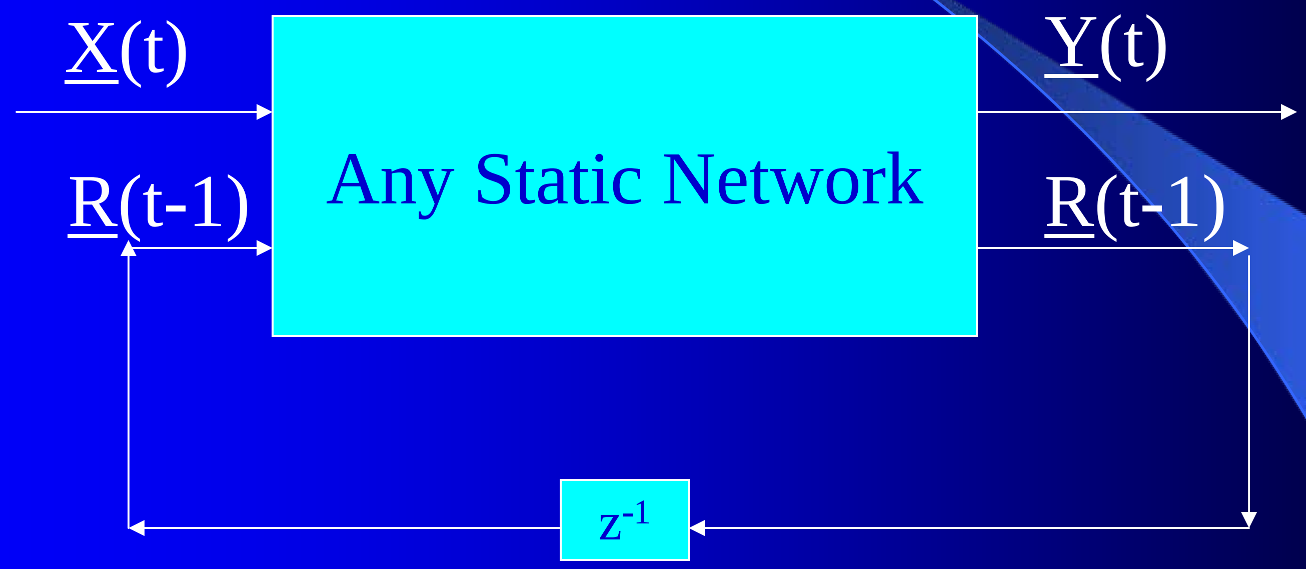
No feedforward or associative memory net can give brain-like performance! Useful recurrence--

- For short-term memory, for state estimation, for fast adaptation – time-lagged recurrence needed. (TLRN = time-lagged recurrent net)
- For better Y=F(X,W) mapping, Simultaneous Recurrent Networks Needed. For large-scale tasks, SRNs WITH SYMMETRY tricks needed – cellular SRN, Object Nets
- For robustness over time, “recurrent training”

Why TLRNs Vital in Prediction: Correlation ≠ Causality!

- E.g.: law X sends extra $ to schools with low test scores
- Does negative correlation of $ with test scores imply X is a bad program? No! Under such a law, negative correlation is hard-wired. Low test scores cause $ to be there! No evidence + or – re the program effect!
- Solution: compare $ at time t with performance changes from t to t+1! More generally/accurately: train dynamic model/network – essential to any useful information about causation or for decision!

The Time-Lagged Recurrent Network (TLRN)



$\underline{Y}(t)=\underline{f}(\underline{X}(t), \underline{R}(t-1)); \underline{R}(t)=\underline{g}(\underline{X}(t), \underline{R}(t-1))$

$\underline{f}$ and $\underline{g}$ represent 2 outputs of one network

All-encompassing, NARMAX(1 ≡ n)

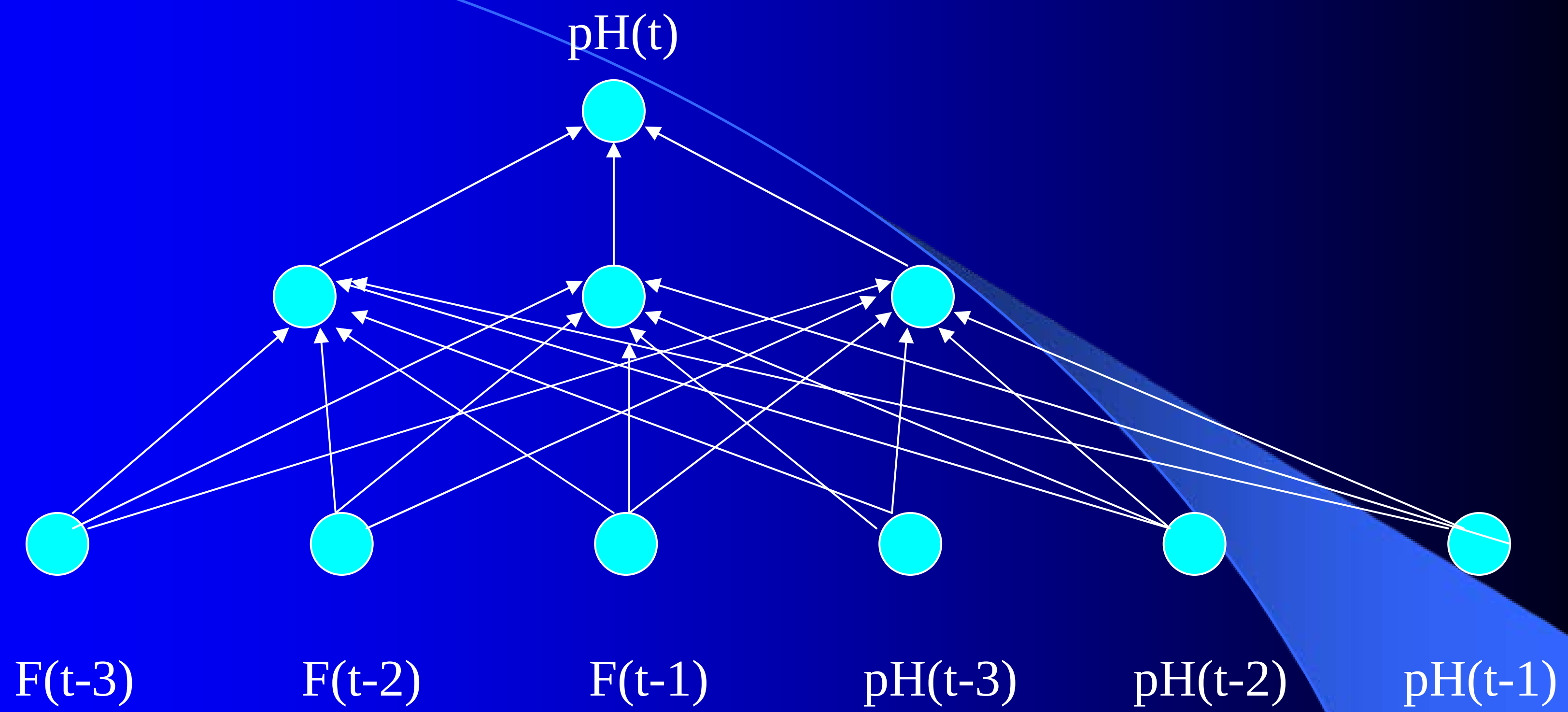
Felkamp/Prokhorov Yale03: >>EKF,≈ hairy

Training: Brain-Style Prediction Is NOT Just Time-Series Statistics!

- One System does it all -- not just a collection of chapters or methods
- Domain-specific info is 2-edged sword:
  - need to use it; need to be able to do without it
- Neural Nets demand/inspire new work on general-purpose prior probabilities and on dynamic robustness (See HIC chapter 10)
- SEDP&Kohonen: general nonlinear stochastic ID of partially observed systems

Three Approaches to Prediction

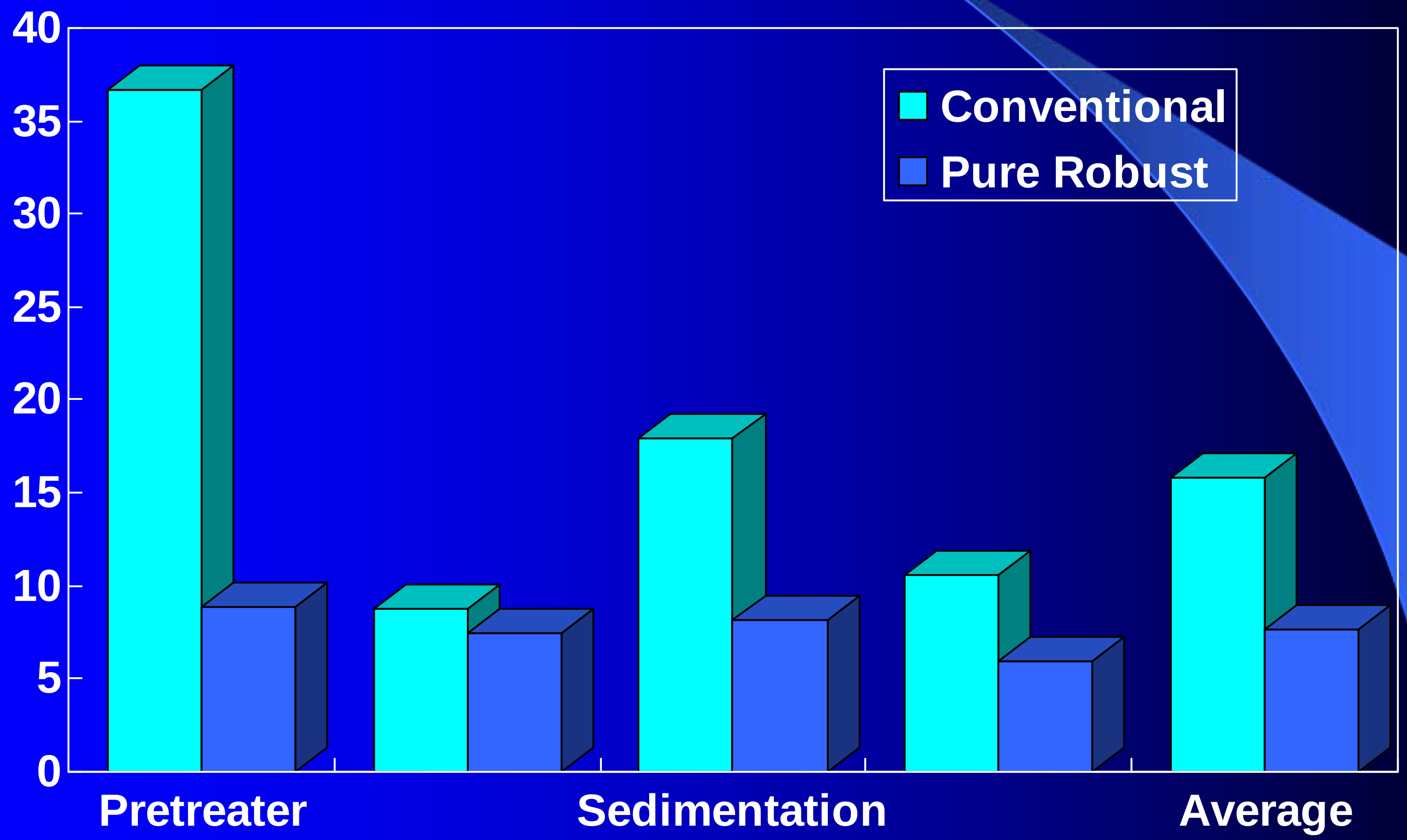
- Bayesian: Maximize Pr(Model|data)
  - "Prior probabilities" essential when many inputs
- Minimize "bottom line" directly
  - Vapnik: "empirical risk" static SVM and "sytructural risk" error bars around same like linear robust control on nonlinear system
  - Werbos '74 thesis: "pure robust" time-series
- Reality: Combine understanding and bottom line.
  - Compromise method (Handbook)
  - Model-based adaptive critics
- Suykens, Land????



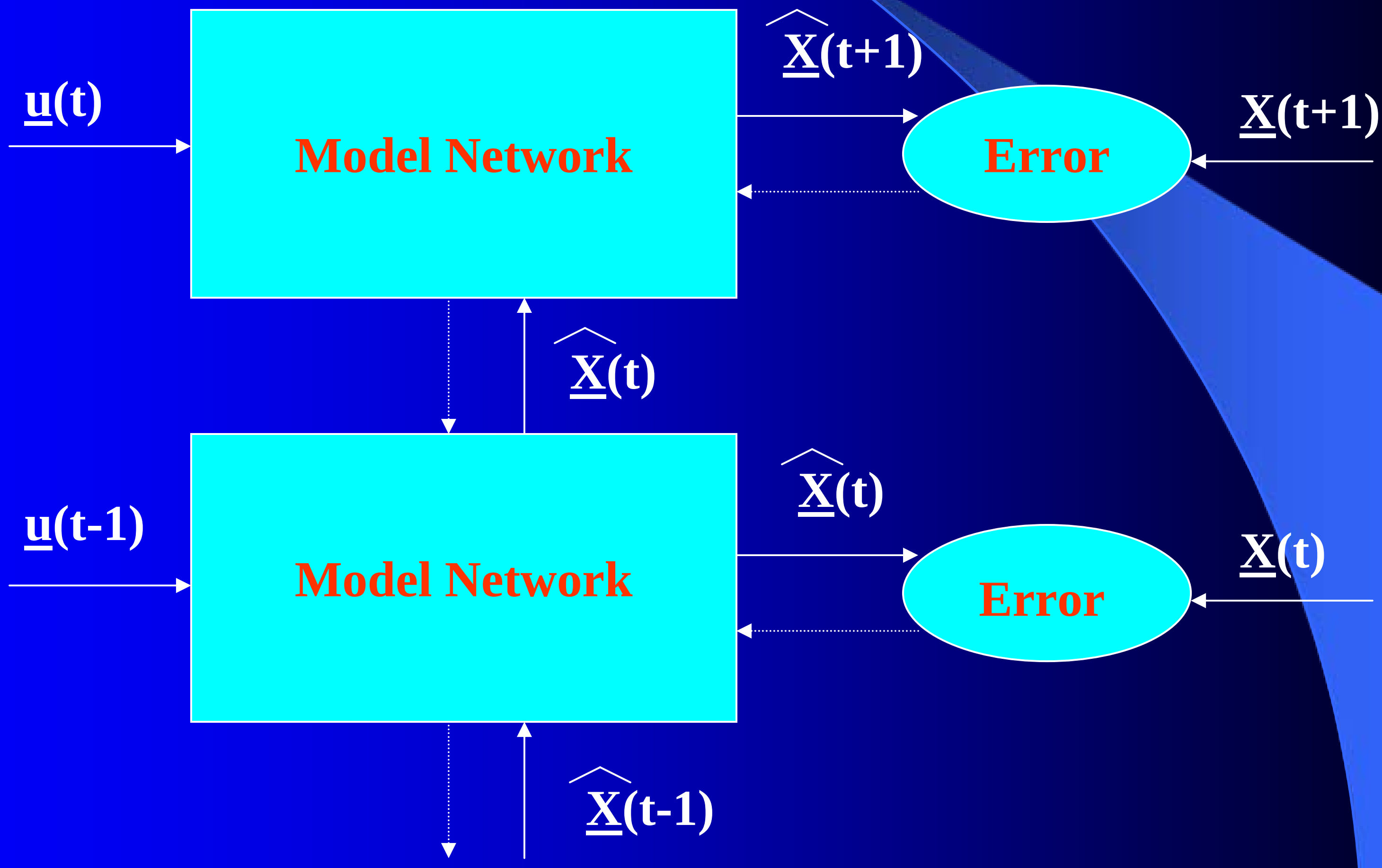
Example of TDNN used in HIC, Chapter 10

TDNNs learn NARX or FIR Models, not NARMAX or IIR

Prediction Errors (HIC p.319)



PURE ROBUST METHOD



Beyond Bellman: Learning & Approx-imation for Optimal Management of Larger Complex Systems www.eas.asu.edu/~nsfadp

- Basic thrust is scientific. Bellman gives exact optima for 1 or 2 continuous state vars. New work allows 50-100 (thousands sometimes). Goal is to scale up in space and time -- the math we need to know to know how brains do it. And unify the recent progress.
- Low lying fruit -- missile interception, vehicle/engine control, strategic games
- Workshops: ADP02 & Dynamic Stochastic Grid testbed; ADP06 April 2006

Wunsch/venayagamoorthy/Harley ADP Turbogenerator Control

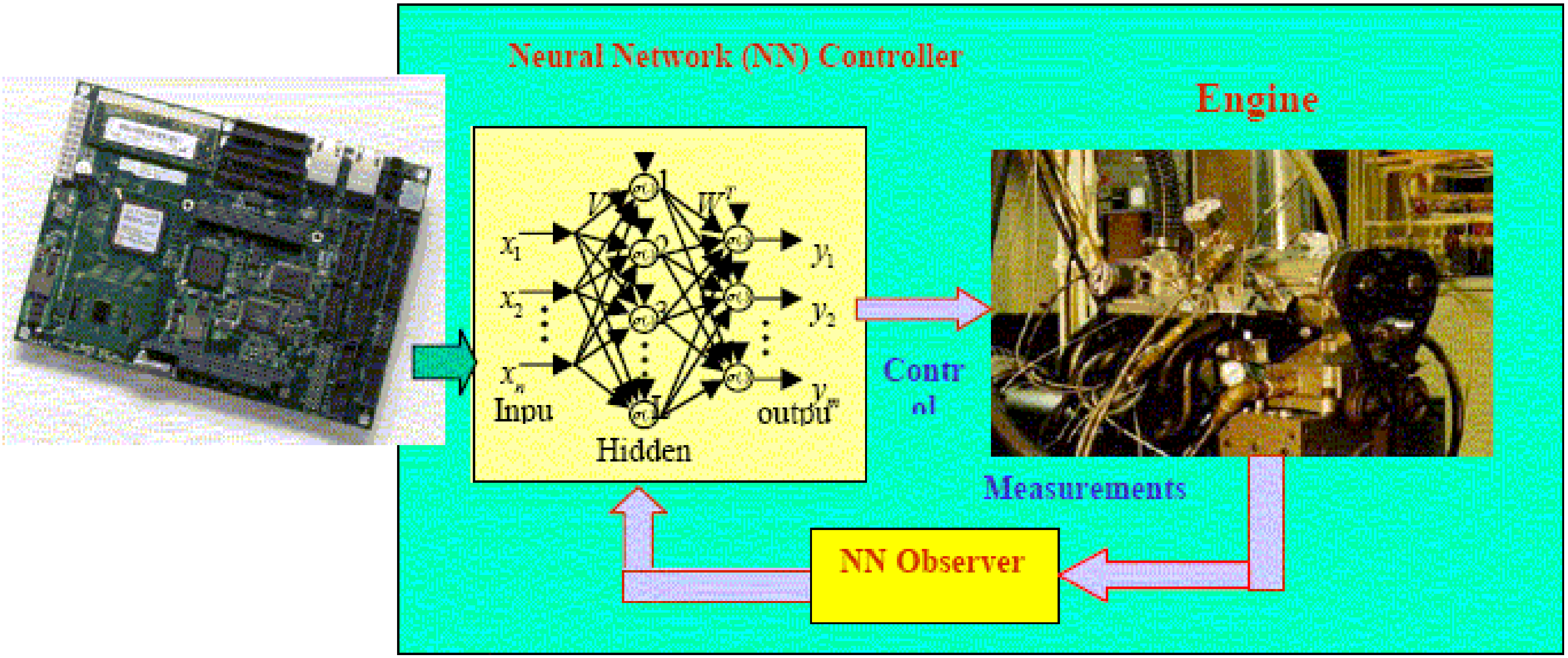
- Stabilized voltage & reactance under intense disturbance where neuroadaptive & usual methods failed
- Being implemented in full-scale experimental grid in South Africa
- Best paper award IJCNN99
- 1st of many, being deployed

Human mentors robot and then robot improves skill

Schaal, Atkeson
NSF ITR project

Learning allowed robot to quickly learn to imitate human, and then improve agile movements (tennis strokes). Learning many agile movements quickly will be crucial to enabling >80% robotic assembly in space.

ADP Controller Cuts NOx emissions from Diesel Engines by 98%

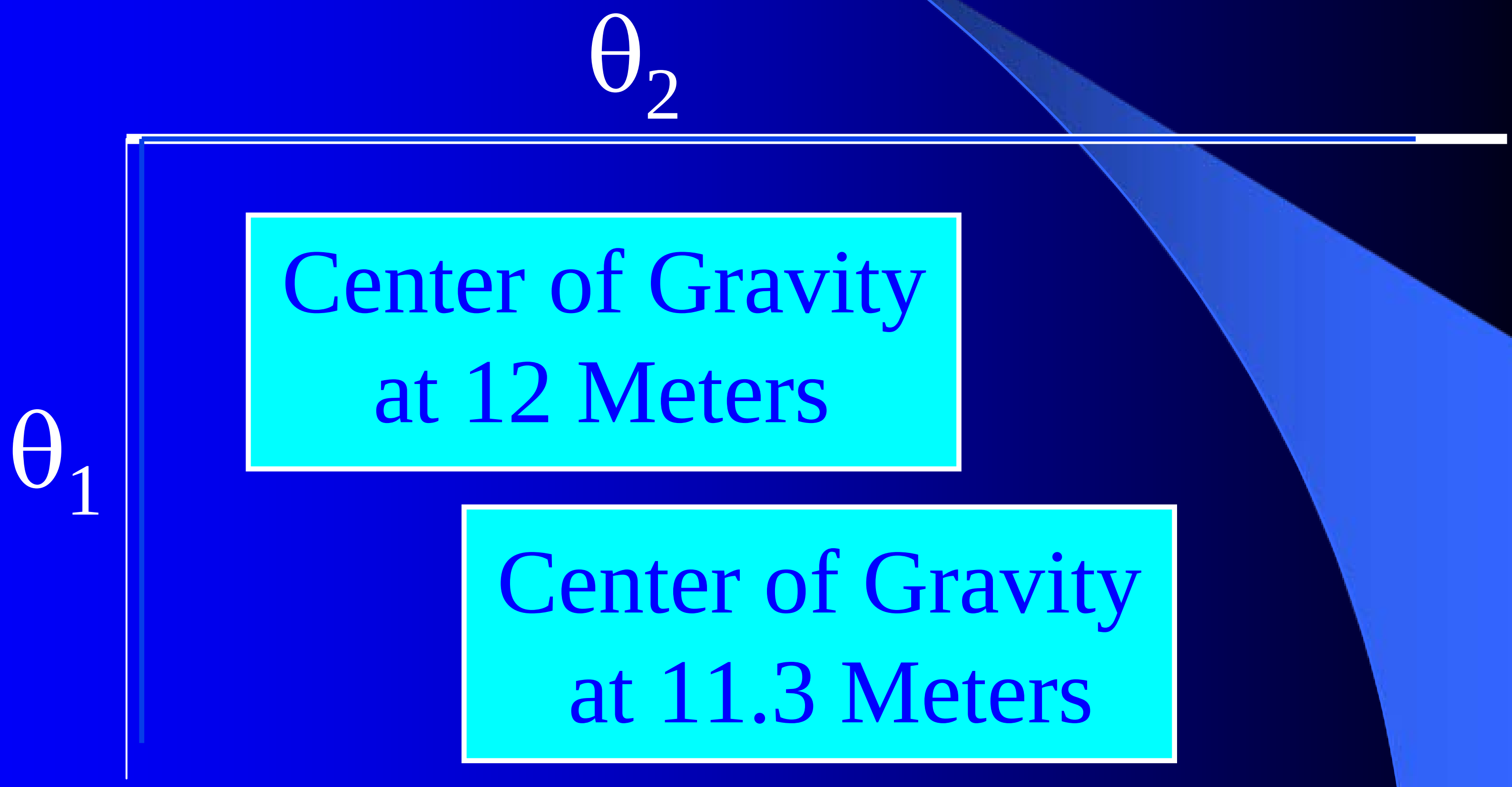


- Sarangapani UMR NSF grant

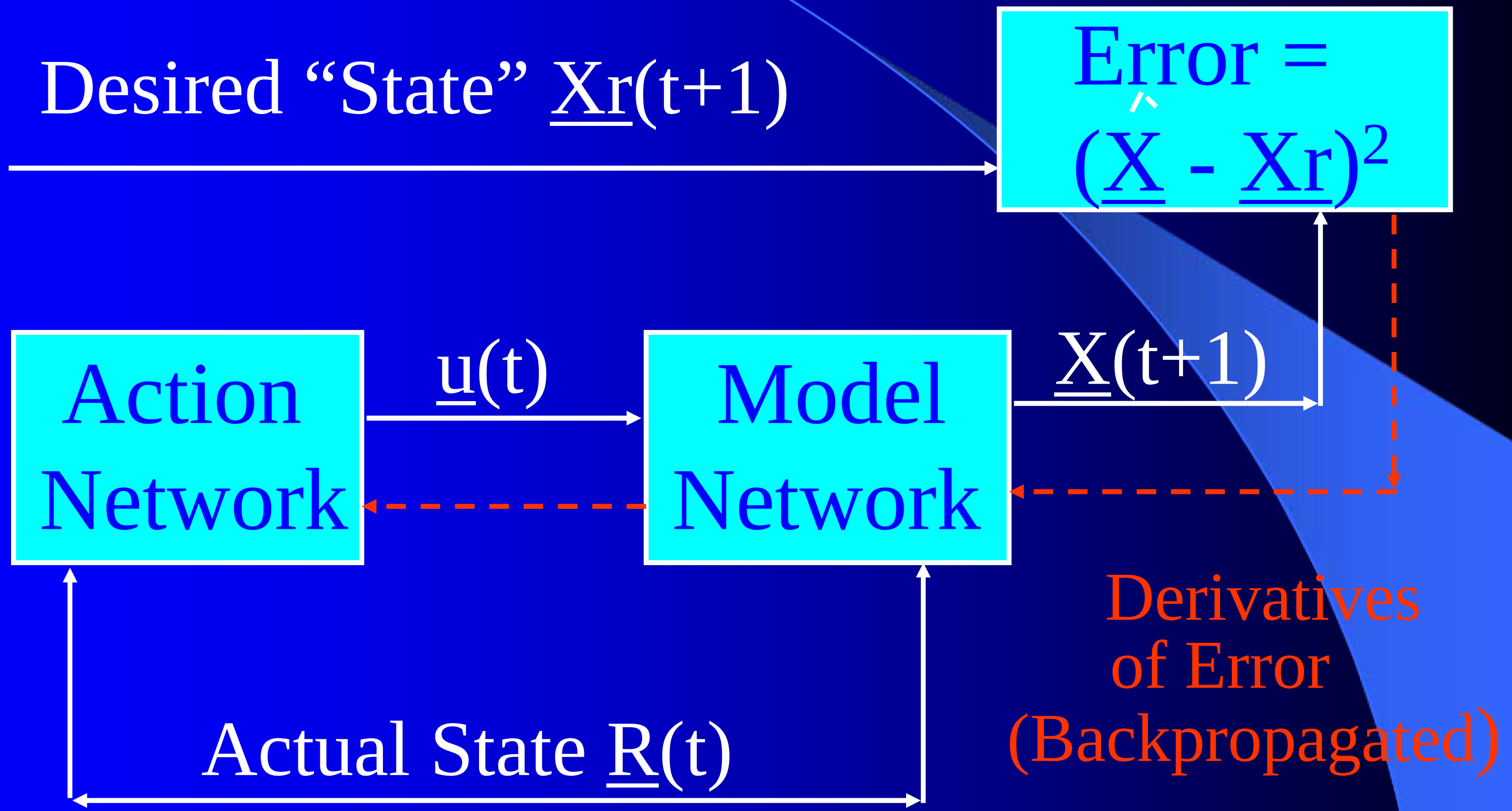
Three Ways To Get Stability

- Robust or H Infinity Control (Oak Tree)
- Adaptive Control (Grass)
- Learn Offline/Adaptive Online (Maren 90)
  - “Multistreaming” (Ford, Felkamp et al)
  - Need TLRN Controller, Noise Wrapper
  - ADP Versions: Online or “Devil Net”

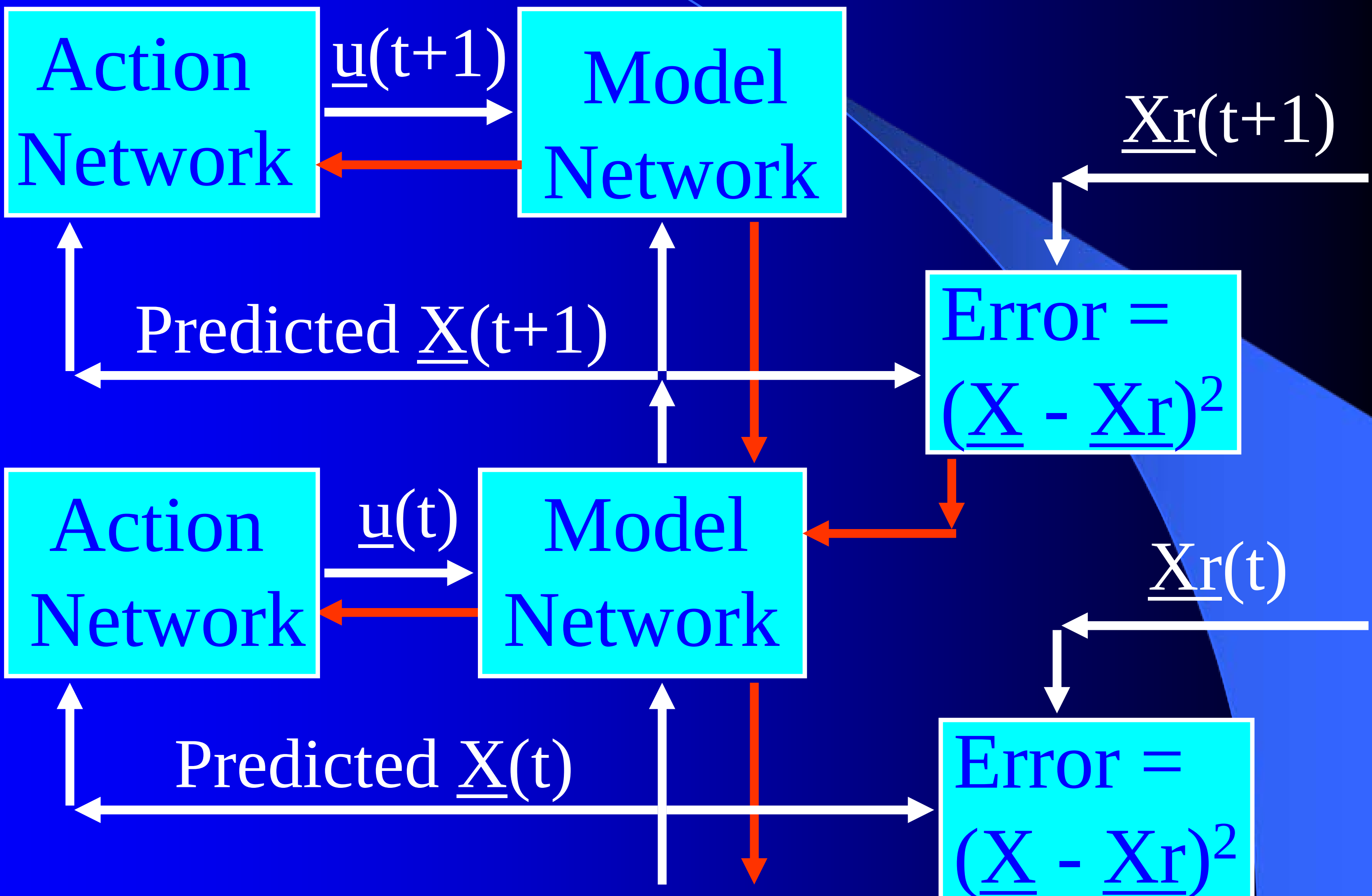
Example from Hypersonics: Parameter Ranges for Stability (H∞)



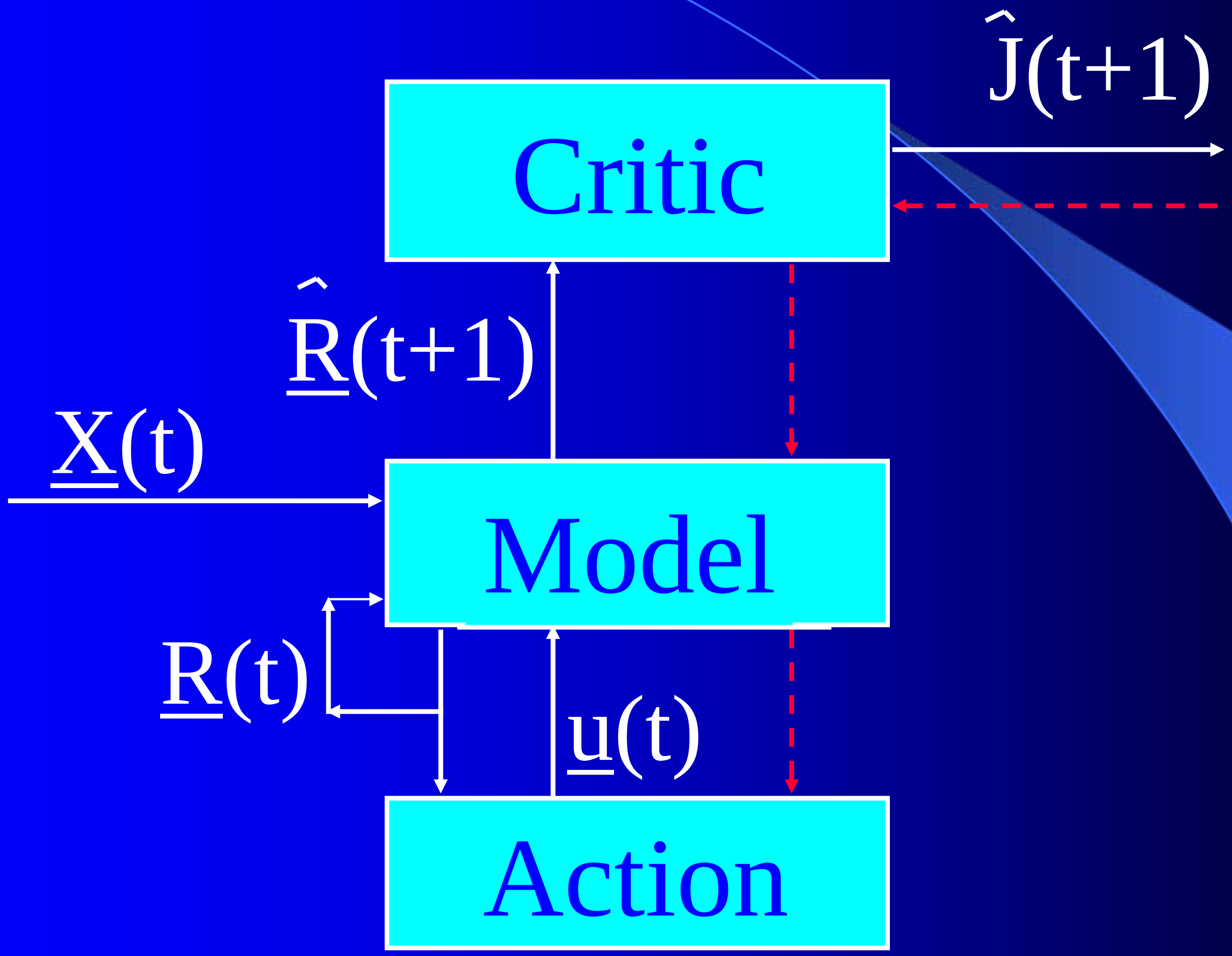
Idea of Indirect Adaptive Control



Backpropagation Through Time (BTT) for Control (Neural MPC)



Level 3 (HDP+BAC) Adaptive Critic System



Gaps in the "SOA" level of ADP Proper: Where Is…?

- Whole system universal stability proof for linear MIMO adaptive control using HDPG, DHPG, GDHPG? (See arxiv.org 1998..)
- General-purpose tools in MatLab, etc.?
- Community knowledge, unification, tools?
- ADP linked to good observers like TLRN? (e.g. see Feldkamp/Prokhorov paper posted at…)
- Good balance of online/offline iteration/learning, of model use vs robustness, discrete/continuous (e.g. GDHP)?
- Good "competition" example?
- Followup on best big application demonstrations?

2^{nd} Generation "Two Brains in One Model"

Upper Level Critic Network

Upper Level Model Network

4-8 hertz

J(t+1)-J(t) from Upper System

U(t) for
Lower System

Additional Local
Utility Components

Lower Level
Adaptive Critic System
Inf. Olive + Cerebellum

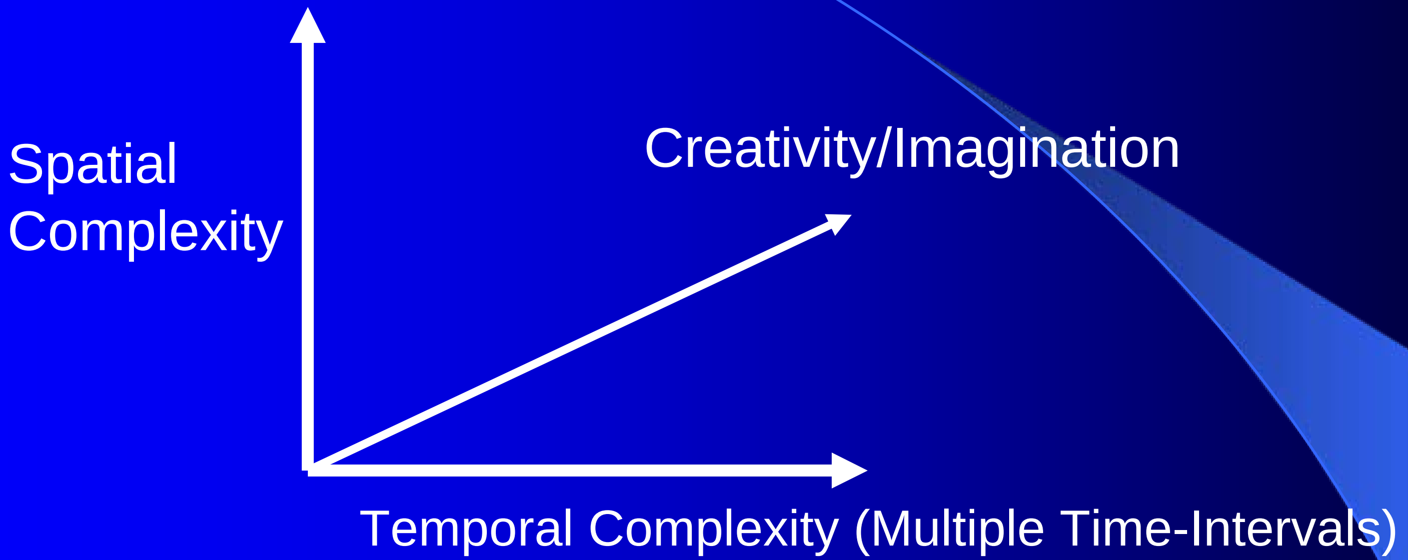
100-200 hertz

Concept in "Statistical/Numerical…", Trans. SMC, 1987 (on web)
Joint papers with Pellionisz (experimental follow-on still warranted)
See equations in Handbook of Intelligent Control, Ch. 13 & Prokhorov

3rd Gen: 3 Brains in 1?

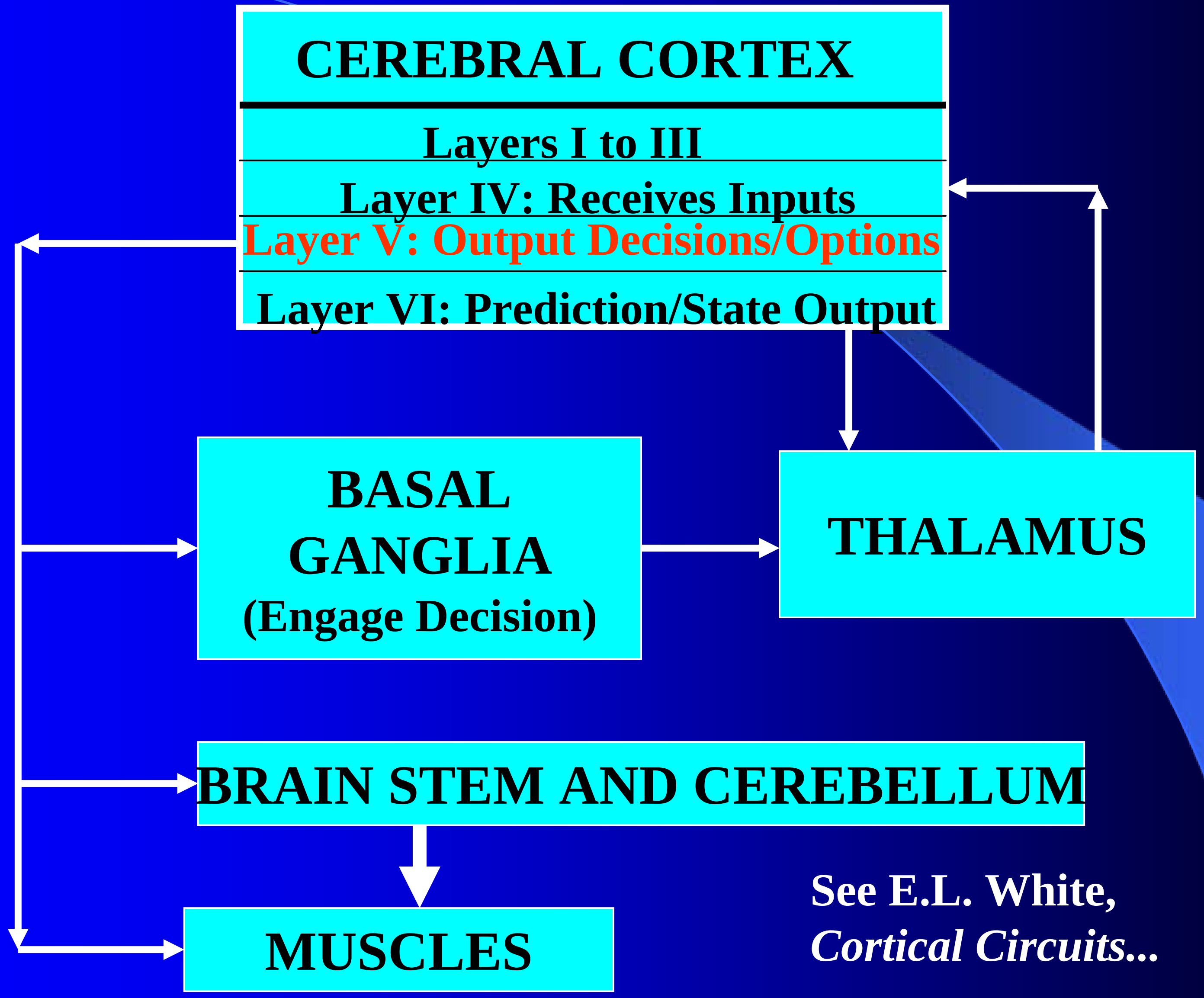
- Upper Brain: Values, Noise, Limbic Critic and Neocortex
- Middle: Basal Ganglia, AI-Like, Tasks, Mishkin, Houk, Brooks, Landing Intent
- Lower: Smoothing/Speed/LQG Like, Olive Critic and Cerebellum
- Complex 3rd Generation Theory over-responsive to AI (Albus) sketched in 1997 paper in Karny et al.

Key Issues in 3rd Generation Model



- Can we (and do brains) do better than 2nd gen brain in handling greater spatial & temporal complexity, by new designs & exploiting unspecialized but structured prior information (Kant) to get faster/better learning?
- What is our answer to AI's "spatial/temporal chunking" & stochastic search?
- All 3 demand more attention and work!!!

3^{rd} Generation View of Creativity/Imagination: Layer V = "Option Networks"



See E.L. White, *Cortical Circuits...*

- Challenge: www.werbos.com/WerbosCEC99.htm.
- Important work by Serpen, Pelikan, Wunsch, Thaler, Fu – but still wide open. Widrow testbed.

3rd Generation View of Time

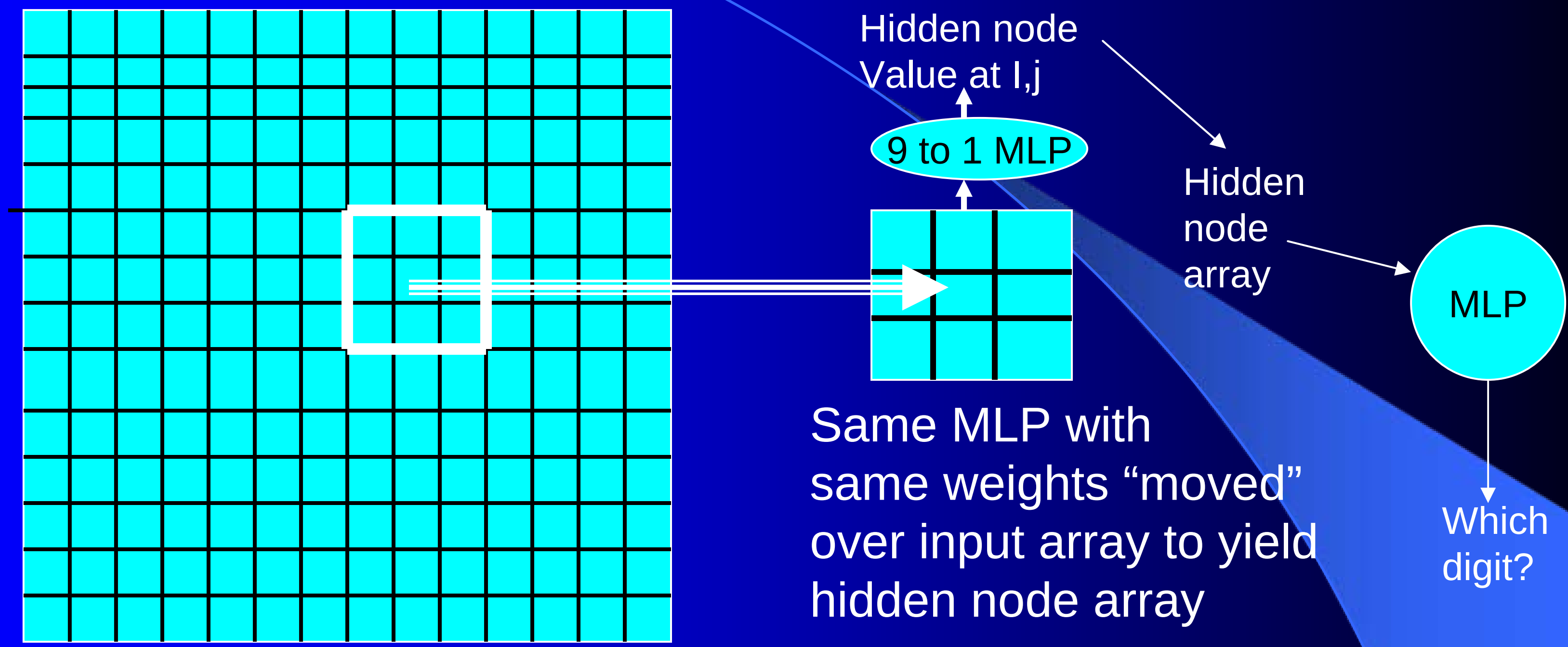
- Before 1997, under NSF$, Sutton had modified "Bellman equation" for idea of "options" – chunks of action over time optimized at low level, to be selected by ADP at high level as discrete choices. (No object, no parameters)
- In 1987 paper, I reported more general Bellman equations for time structuring, e.g.:

$$J_i^T = (J_i^A)^T + \text{SUM (over j in N(i))} \; J_J^T (J^B)_i^J$$

where JA represents utility within valley i before exit, and JB works back utility from the exits in New valleys j within the set of possible next valleys N(i). Leads directly to a neural net approximator using "decision blocks" similar to then-current ideas re basal ganglia and "tasks".

•Despite many discussions, no apps except options in robotics "behavior libraries" (e.g. Schaal) yet! Barriers: politics; my time; presence of spatial complexity also in many potential apps! Most "context" better handled by TLRNs.

Moving Window Net: Clue Re Complexity



Large pixel array input for Zip Code Digit

- Best ZIP Code Digit Recognizer Used "Moving Window" or "conformal" MLP! (Guyon, LeCun, AT&T story, earlier…)
- Exploiting symmetry of Euclidean translation crucial to reducing number of weights, making large input array learnable, outcomes.

Cellular SRN: The Recurrent (SRN) Generalization of “Conformal MLP”

GENERALIZED MAZE PROBLEM

Jhat(ix,iy) for all $0<ix,iy<N+1$

(an N by N array)

↑

NETWORK

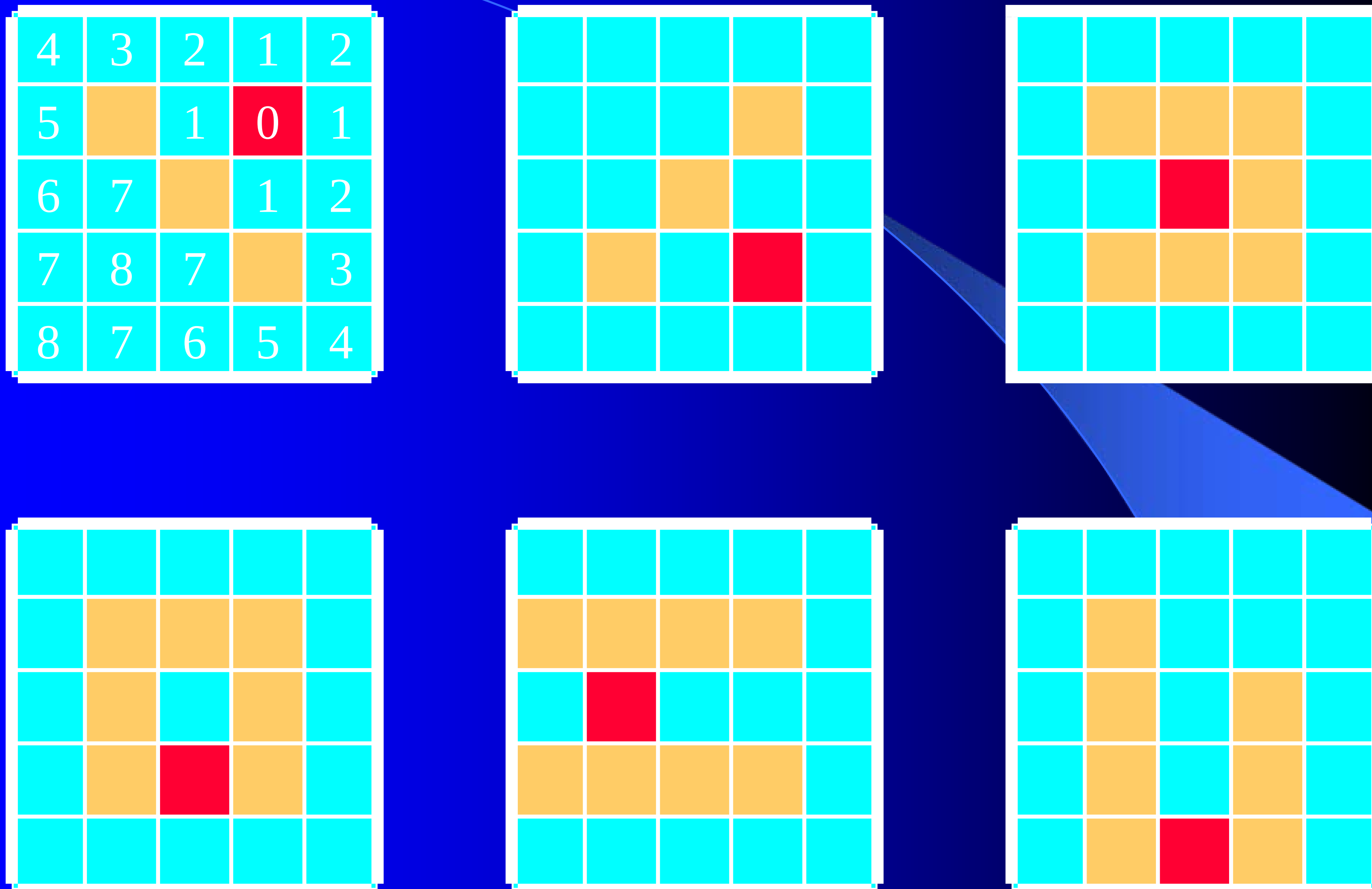
↑

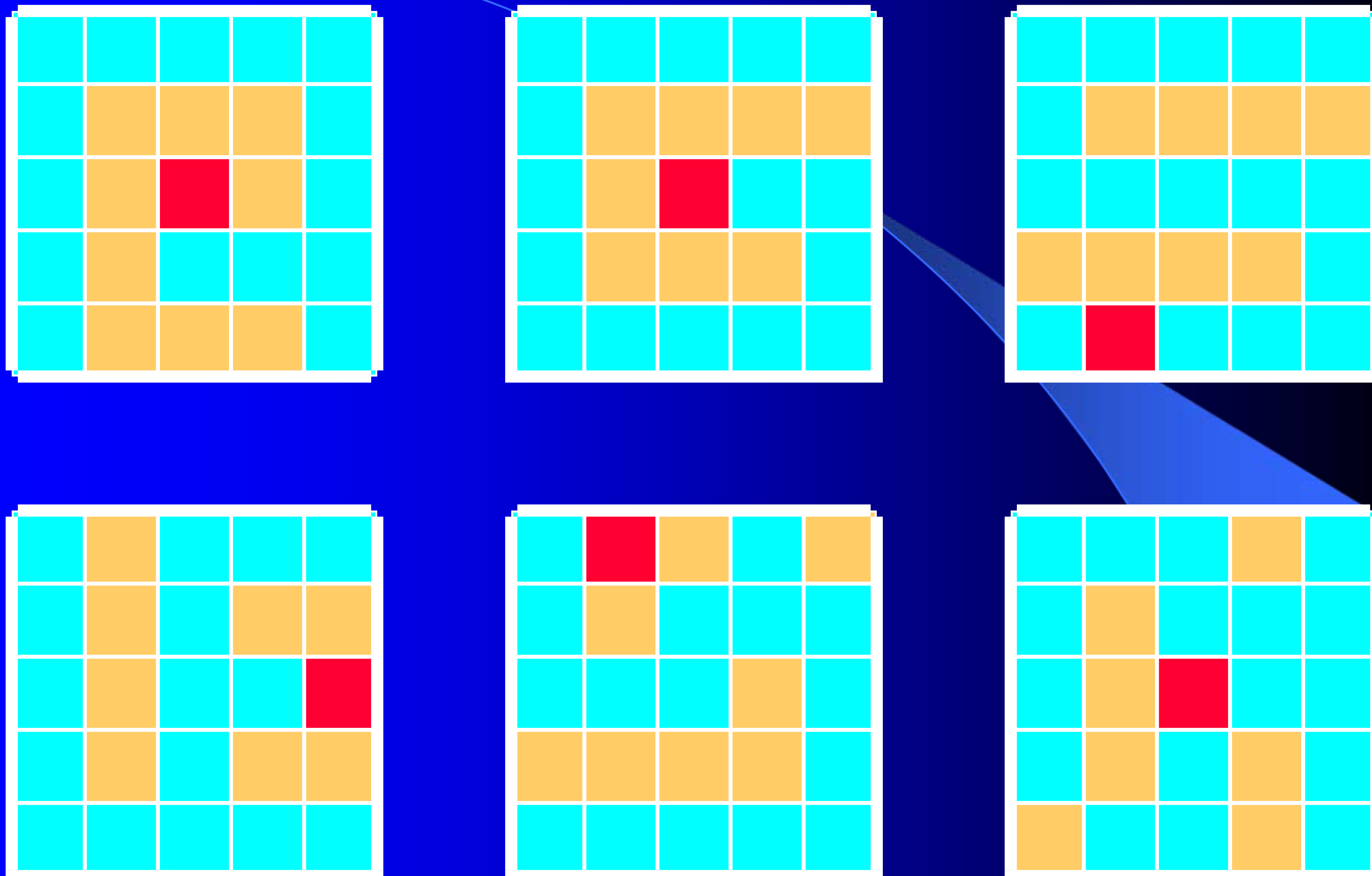
Maze Description

- Obstacle (ix,iy) all ix,iy
- Goal (ix,iy) all ix,iy

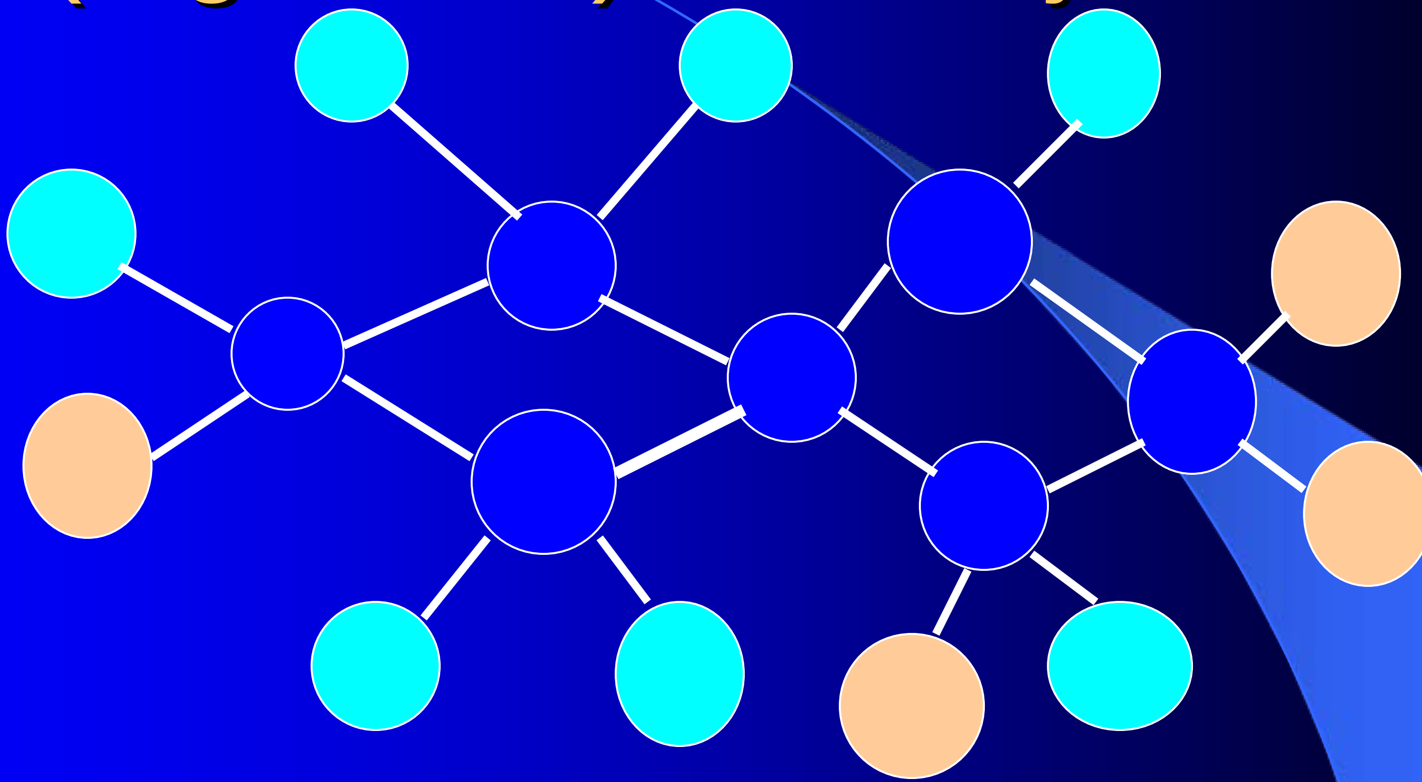
At arXiv.org, nlin-sys, see adap-org 9806001

For rapid practical learning, Ilin, Kozma


4 3 2 1 2
5 1 0 1
6 7 1 2
7 8 7 3
8 7 6 5 4



Spatial Symmetry in the General Case (e.g. Grids): the Object Net



- 4 General Object Types (busbar, wire, G, L)
- Net should allow arbitrary number of the 4 objects
- How design ANN to input and output FIELDS -- variables like the SET of values for current ACROSS all objects?
- Great preliminary success (Fogel's Master Class Chess player; U. Mo. Power)
- But how learn the objects and the symmetry transformations???? (Brain and images!!)

From Neural Networks to the Intelligent Power Grid: What It Takes to Make Things Work

- What is an Intelligent Power Grid, and why do we need it?
- Why do we need neural networks?
- How can we make neural nets really work here, & in diagnostics/"prediction"/"control" in general?

Paul J. Werbos, pwerbos@nsf.gov

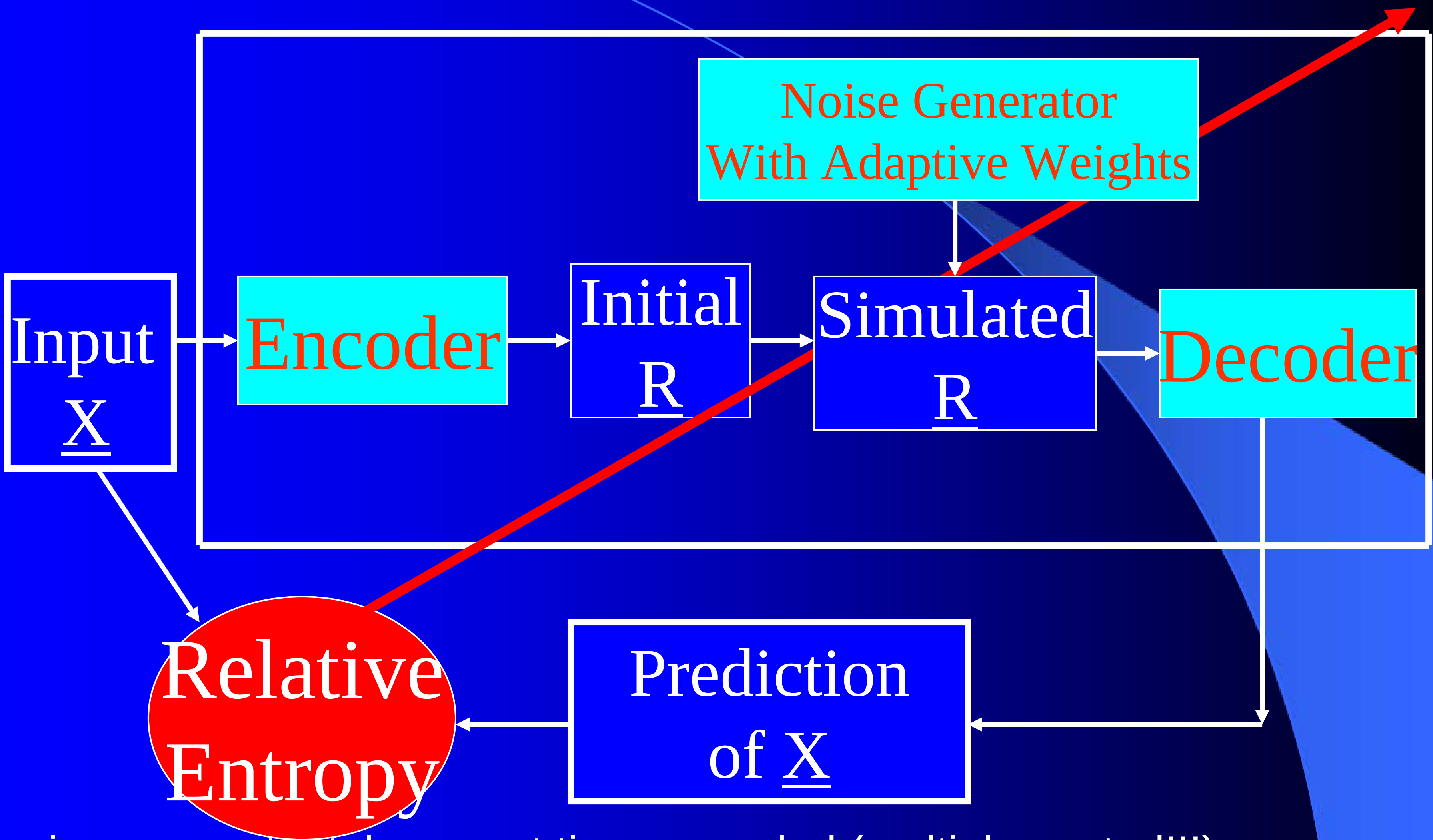
Quick Review of www.face-rec.org

- See Chellappa et al for review
- 3 best recognizers use ANNs, learning
- Wechsler, von der Malsburg: need to learn elastic symmetry transformations (e.g. curl up mouth), not just Euclidean
- Low-lying fruit: use CSRN or Object Net to learn elastic symmetry transformations, but how does brain do it? Foveal vision doesn’t have Euclidean metric symmetry. (Though topology helps, connection learning.)

Fresh Look: Initial Approach to Brain-Like Symmetry Learning and Use

- First learn a family of vector maps $\underline{f}_\alpha$ such that:
  - $\Pr(\underline{f}_\alpha(\underline{x}(t+1)|\ \underline{f}_\alpha(\underline{x}(t)) = \Pr(\underline{x}(t+1)|\underline{x}(t))$ for the same conditional probability distribution Pr and all α.
- Exploit these symmetries via:
  - "Reverberatory generalization": after observing or remembering the pair $\{x(t+1),x(t)\}$, also train on $\{\underline{f}_\alpha(\underline{x}(t+1),\underline{f}_\alpha\ (\underline{x}(t))\}$.
  - "Multiplex gating": after inputting x(t), pick α to map x to $\underline{f}_\alpha(\underline{x}(t)$, and use that as input to a "universal" canonical prediction model. (e.g. Olshausen. Not the same as spontaneous or affective or salience gating.)
  - "Multimodular gating": like multiplexed, but implement parallel (coordinated) copies of the canonical model to allow use on multiple objects in parallel at the same time.
- Human brains seem to exploit the first two (or second), but how are the symmetry transforms learned? How far can a purely emergent kind of design get by learning?

Solution Exists Off-the Shelf! (SEDP, HIC Chapter 13)

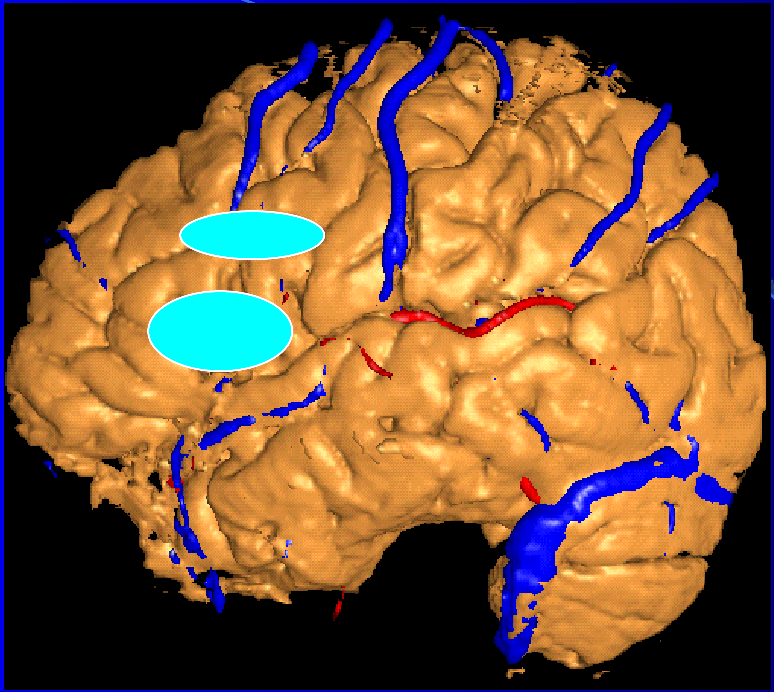


Learning symmetry takes most time; encoded (multiplex gated!!!) image allows fast learning of objects, faces, etc., as if brute force gating or transformation encoding!!

To get from SEDP to full Mammal Brain Like Spatial Complexity:

- Work to improve learning speed, robustness & generalization in SRN, TLRN, CSRN, Object Net, GDHP and SEDP – including memory-based learning as discussed often, & analysis of mathematical properties, toolkits, etc.
- Active control of saccade & efferent copy to encoder
- Test short-term object permanence (automatic), and augment long-term memory I/O interface for "object identity" and "world modeling."

New Data on Complexity in the Brain



Petrides (IJCNN06) shows that dorsolateral (DL) and orbitofrontal (OF) prefrontal cortex – our "highest" brain centers – answer two basic questions:
OF: Where did I leave my car this time in the parking lot? (space?)
DL: What was I trying to do anyway? (time?)

- BUT: even bird brains (no neocortex) handle great spatial complexity & have big basal ganglia!!
- Hypothesis: SEDP fits pyramid cell geometry very well but is already be in old cortex (bird!)
- Neocortex (mouse) harnesses/alters stochastic mechanism in SEDP for creativity.
- OF strengthens object identity & world modeling & object-oriented action. (Test birds, lizards!)
- Temporal aggregation is by "re-entrant" mechanism, not explicit temporal hierarchy.