Cognitive Architecture: An Approach to AGI

Nate Derbinsky
Associate Teaching Professor
Northeastern University

Outline

- Why **AGI**?
  - Research questions/goals
- What is **Cognitive Architecture**?
  - Prototypical assumptions, structures
  - Representative snapshots
- An Example of **Research in Soar**?
  - Human inspiration -> what to remember/forget
- Where to **Learn More**?

A Rough Definition of AGI

- Understanding/development of systems that exhibit "human-level intelligence"

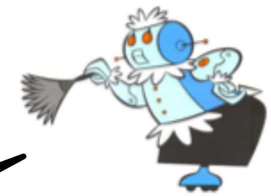
- Agents that…
  - **persist** for long periods of time,
  - exhibiting **robust and adaptive behavior**
  - in a **variety of tasks** and situations

AGI & Me

Expectations Meet (Current) Reality

"Alexa, please write me an `rsync` script."

"Sorry, I don't know that one."



"Do you have time to teach me?"

Common Motivations

(Existential) Curiosity

- Abstract knowledge creation
- Answering challenging questions

Cognitive Modeling

- Understanding how a (human) brain/mind functions
- Applications in medicine, HCI/HRI, simulation, …

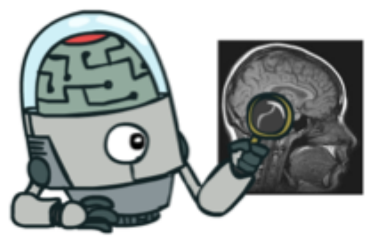
Systems Development

- Build more capable hardware/software for replacing and/or augmenting human performance
- *When designing an artifact, look to examples*

Motivations/Questions Dictate Approach

Ground Truth

What to Judge	Humanly	Rationally
Thinking	Cognitive Modeling	"Laws of Thought"
Acting	Turing Test	Rational Agent



Theories of Cognition

Without implementation and integration, it can be **difficult to synthesize** and generalize from **diverse findings** on intelligence

Fitts' Law

$$ID = \log_2 \frac{2A}{W}$$

+

Power Law of Practice

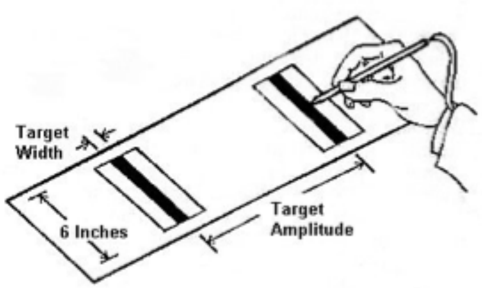
$$RT = aP^{-b} + c$$

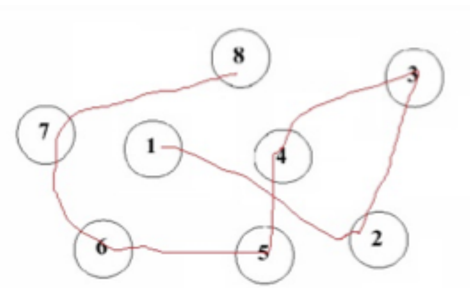
+

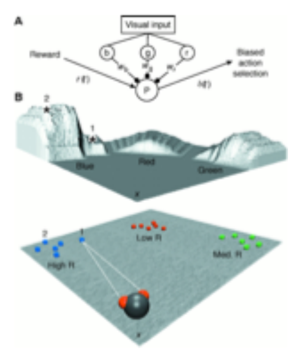
TD Learning

$$V(s) \mathrel{+}= \alpha(r + \gamma V(s') - V(s))$$

?
=





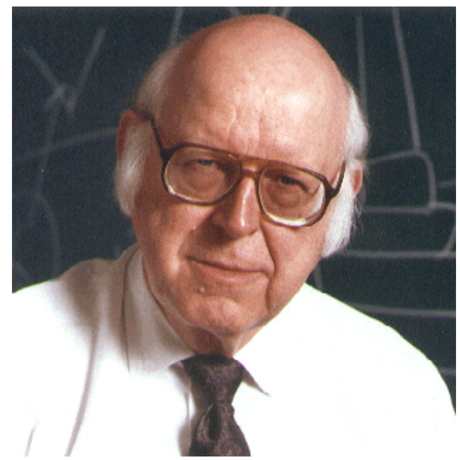


Unified Theories of Cognition

[Newell 1990]

Cognitive Architecture specifies those aspects of cognition that remain constant across the lifetime of an agent

- Memory systems of agent's beliefs, goals, experience
- Knowledge representation
- Functional processes that lead from perception through to behavior
- Learning mechanisms

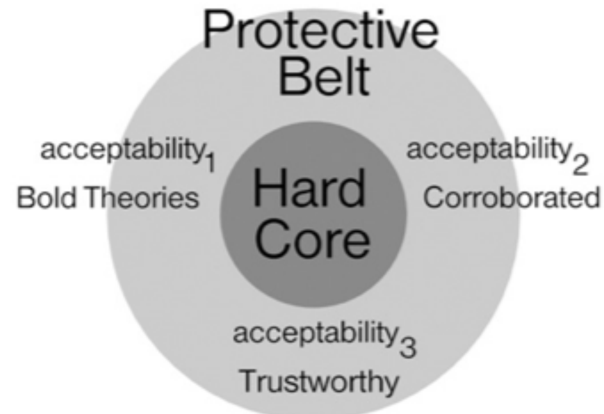


Goal. Understand and exhibit intelligence across a diverse set of tasks and domains

Making (Scientific) Progress

[Lakatos 1970]

- Research in cognitive architecture often resembles a Lakatosian “research programme”
  - A hard core of “central tenets”
  - A “protective belt” of assumptions
- As discoveries are made, the belt is amended and the core expanded
  - The size of the core and the breadth of the tasks leads to desirable **constraints** that increasingly **limit the design space**
- Let’s now consider some of these core assumptions…



Time Scales of Human Action

[Newell 1990]

Scale (sec)	Time Units	System	World (theory)
10^7	Months		
10^6	Weeks		Social Band
10^5	Days		
10^4	Hours	Task	
10^3	10 min	Task	Rational Band
10^2	Minutes	Task	
10^1	10 sec	Unit Task	
10^0	1 sec	Operations	Cognitive Band
10^{-1}	100 ms	Deliberate act	
10^{-2}	10 ms	Neural circuit	
10^{-3}	1 ms	Neuron	Biological Band
10^{-4}	100 µs	Organelle	

Core Takeaways

- There exist **regularities at multiple time scales** that are productive for understanding the mind
- There exist **useful layers of abstraction** between bands, roughly…
  - Biological: neuroscience
  - Cognitive/Rational: psychology, cognitive science
  - Social: economics, political science, sociology
- Cognitive Architectures typically focus on the deliberative act (though some model lower)
  - **Processing at the higher levels then amounts to sequences of these interactions over time**

Bounded Rationality

[Simon 1957]

- Agent rationality is limited by…
  - **tractability** of the decision problem
  - **cognitive limitations** of the mind
  - **time available** to make the decision

- *"Decision makers can **satisfice** either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world. Neither approach, in general, dominates the other, and both have continued to co-exist"*

Physical Symbol System Hypothesis

[Newell & Simon 1976]

- A **Physical Symbol System** takes physical patterns (**symbols**), combines them into structures (**expressions**), and manipulates them (using **processes**) to produce new expressions

- *A physical symbol system has the necessary and sufficient means for general intelligent action*
  - Likely requires non-symbolic representation(s) and processes (e.g. statistical, spatial)

Active Architectures by Focus

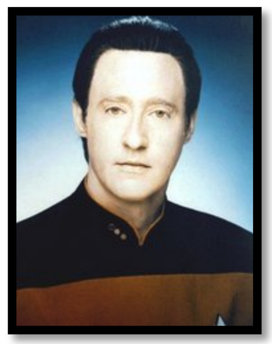
Biological Modeling

Leabra
SPAUN

Psychological Modeling

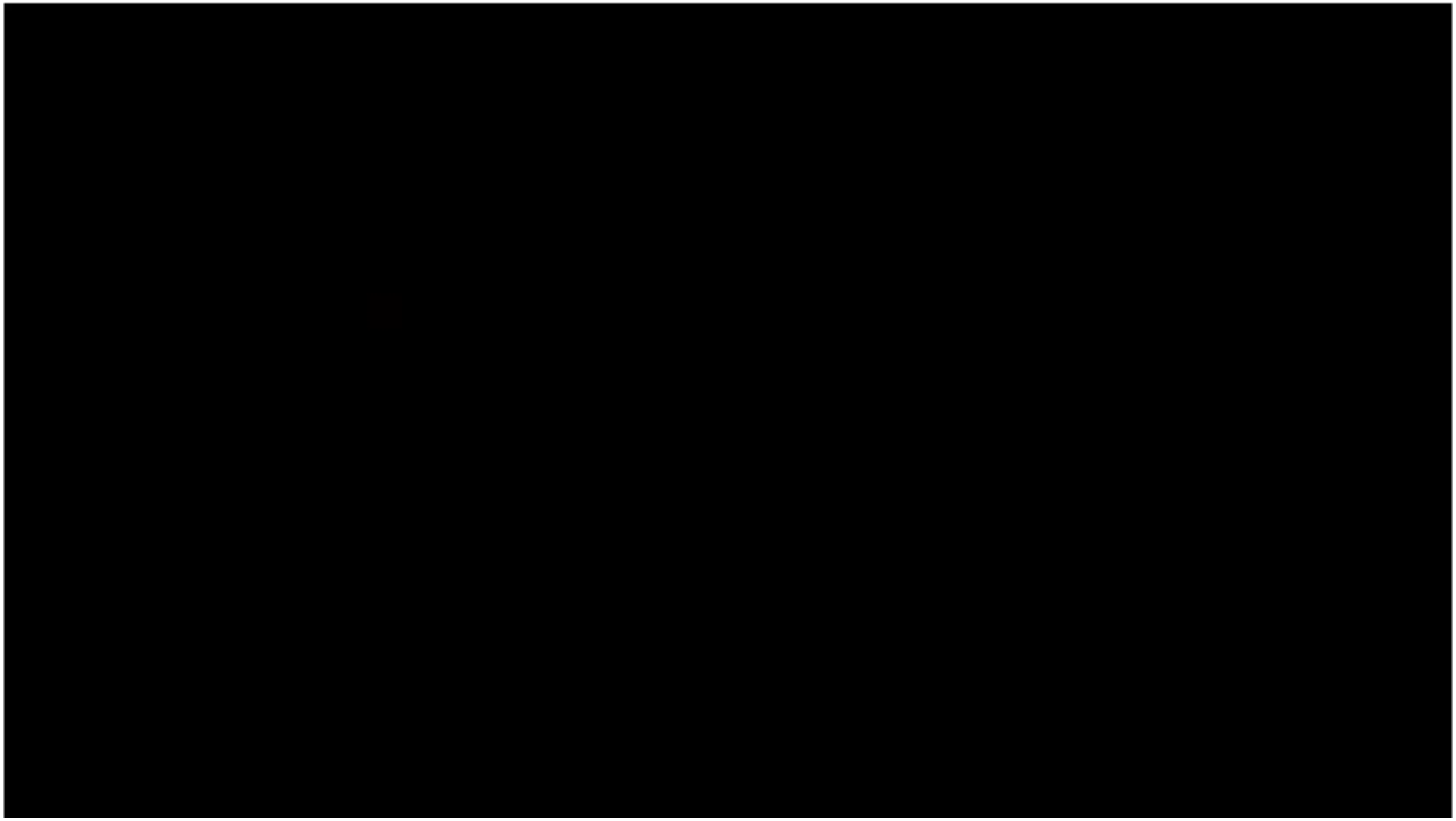
ACT-R
CLARION
EPIC

Agent Functionality

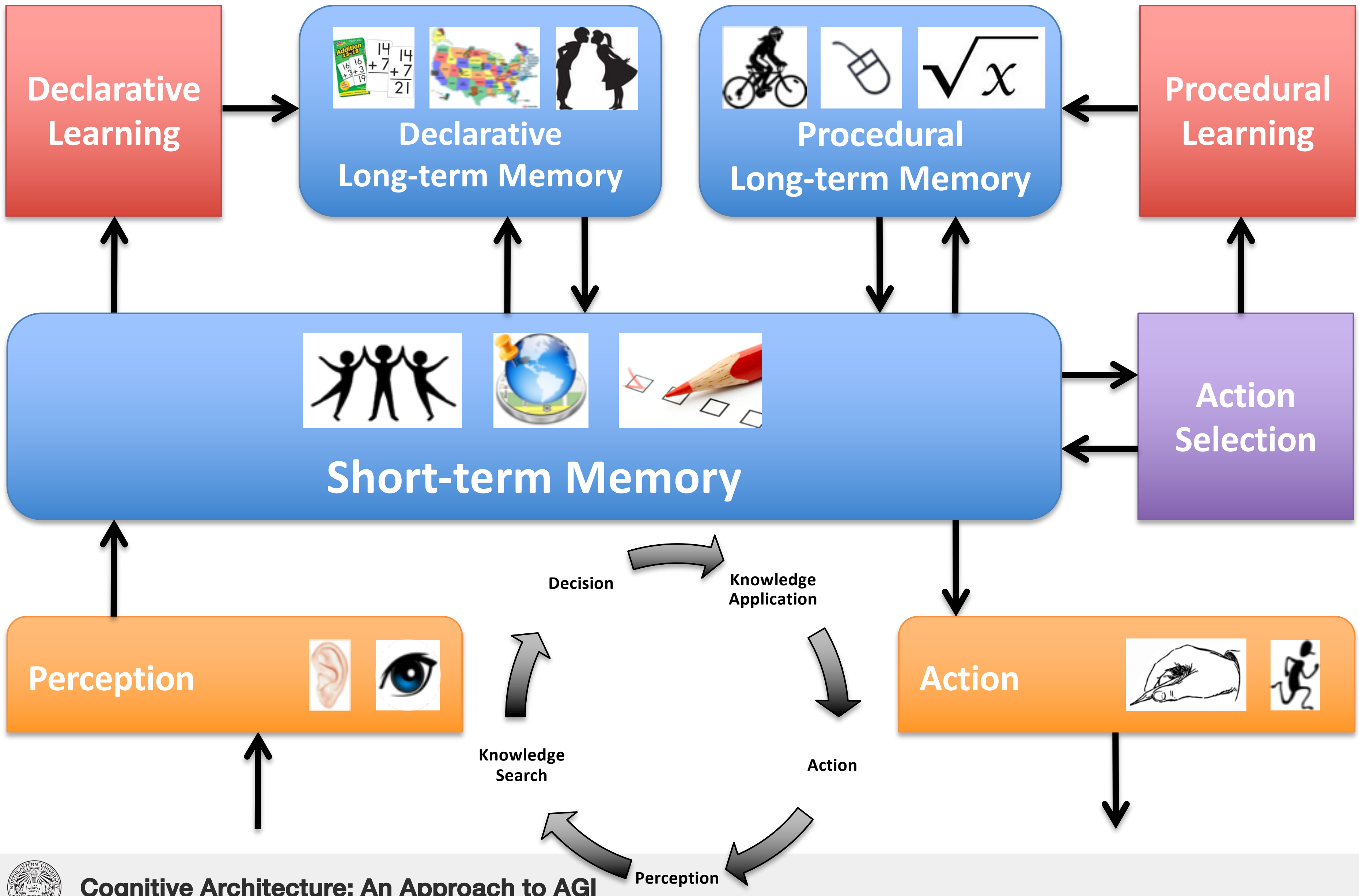


Companions
ICARUS
LIDA
Sigma
Soar

Semantic **P**ointer **A**rchitecture **U**nified **N**etwork



Prototypical Architecture



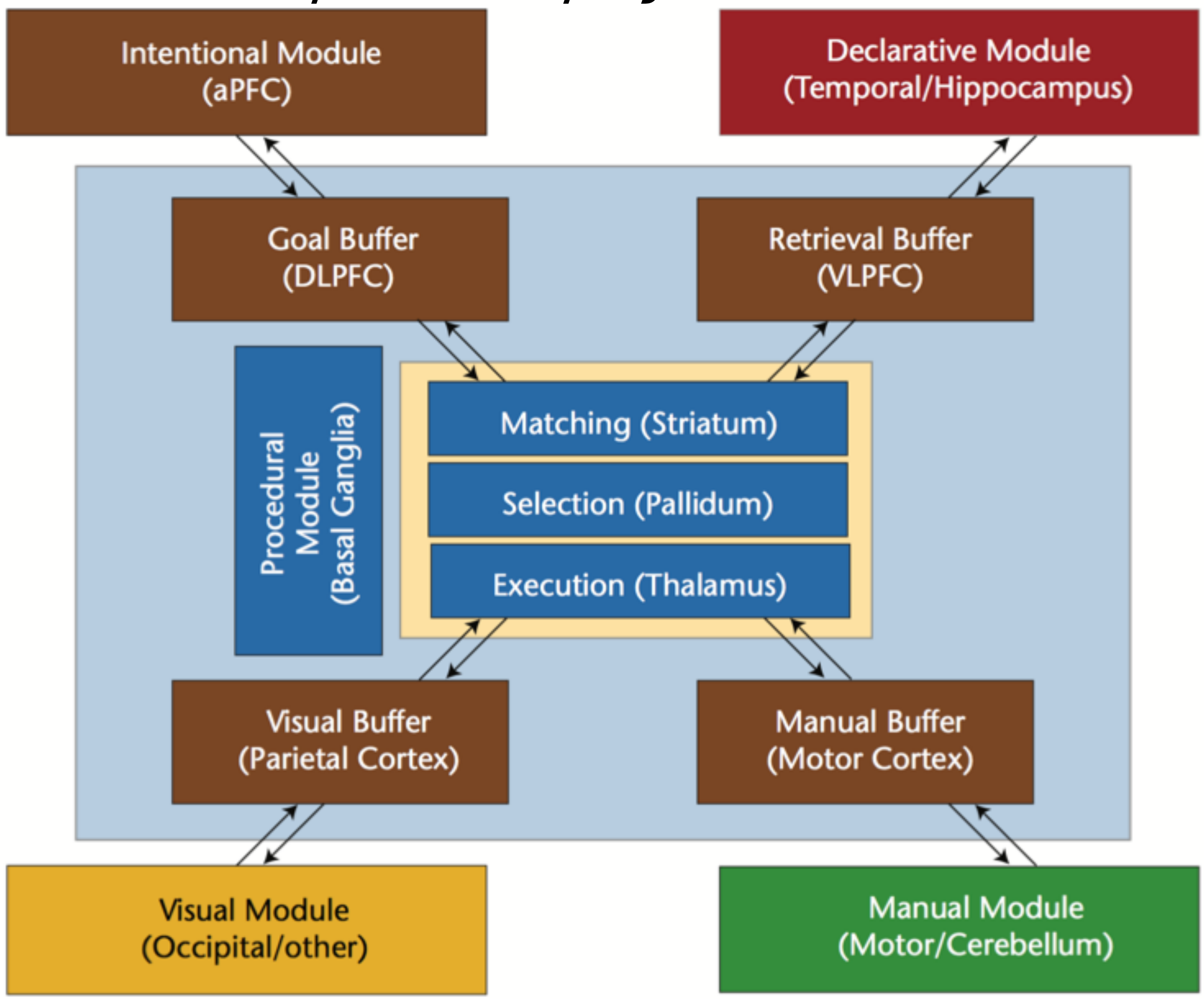
Defining an Agent

- **Agent = Architecture + Knowledge**
  - Knowledge can be task-specific/general
  - In this context, "architecture" encompasses both fixed processes and tuned parameters

- It is typical for the architecture to structure behavior around a cognitive **cycle**, whereby complex behavior arises out of sequences of primitive decisions

ACT-R

http://actr.psy.cmu.edu



ACT-R People

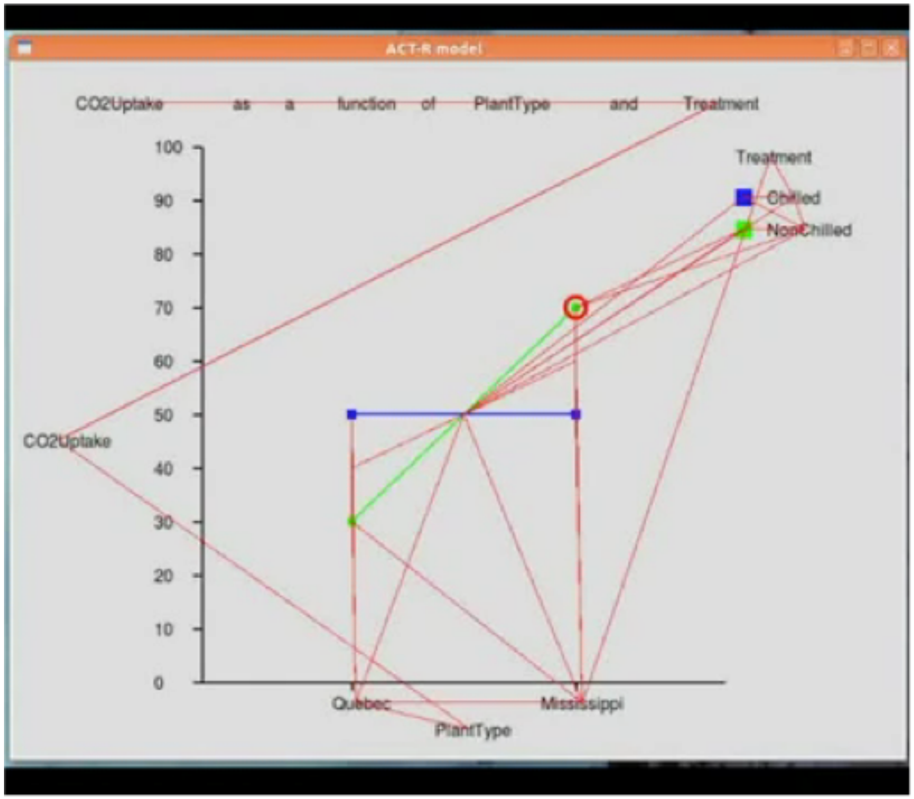
John R. Anderson
Professor of Psychology, CS @ CMU

Christian Lebiere
Research Scientist @ CMU

ACT-R Notes

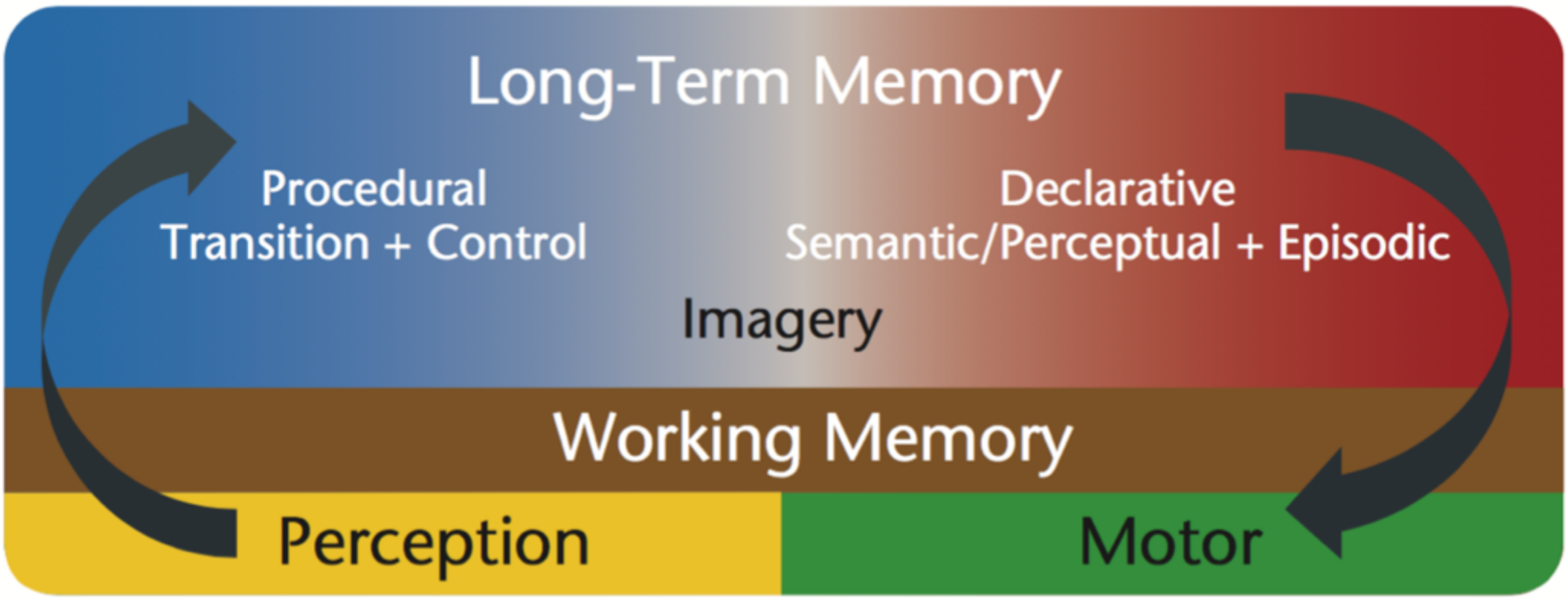
David Peebles
Reader, Cognitive Science @ U. of Huddersfield

- Over 1100 related publications
- Main version in LISP, ported to at least 2 other languages/platforms
- Makes detailed predictions about decision times, error rates, learning, etc. in a variety of architectural processes
- Annual Workshop, Summer School



Sigma (Σ)

http://cogarch.ict.usc.edu



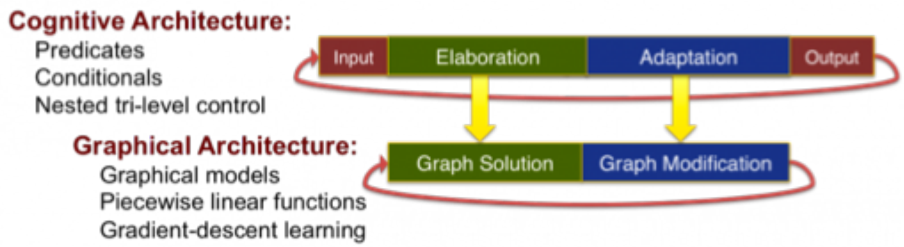
Sigma (Σ)

Paul Rosenbloom
Professor of CS @ USC
Director of Cog Arch @ ICT

- Created originally to explore a uniform substrate (factor graphs) to reproduce Soar

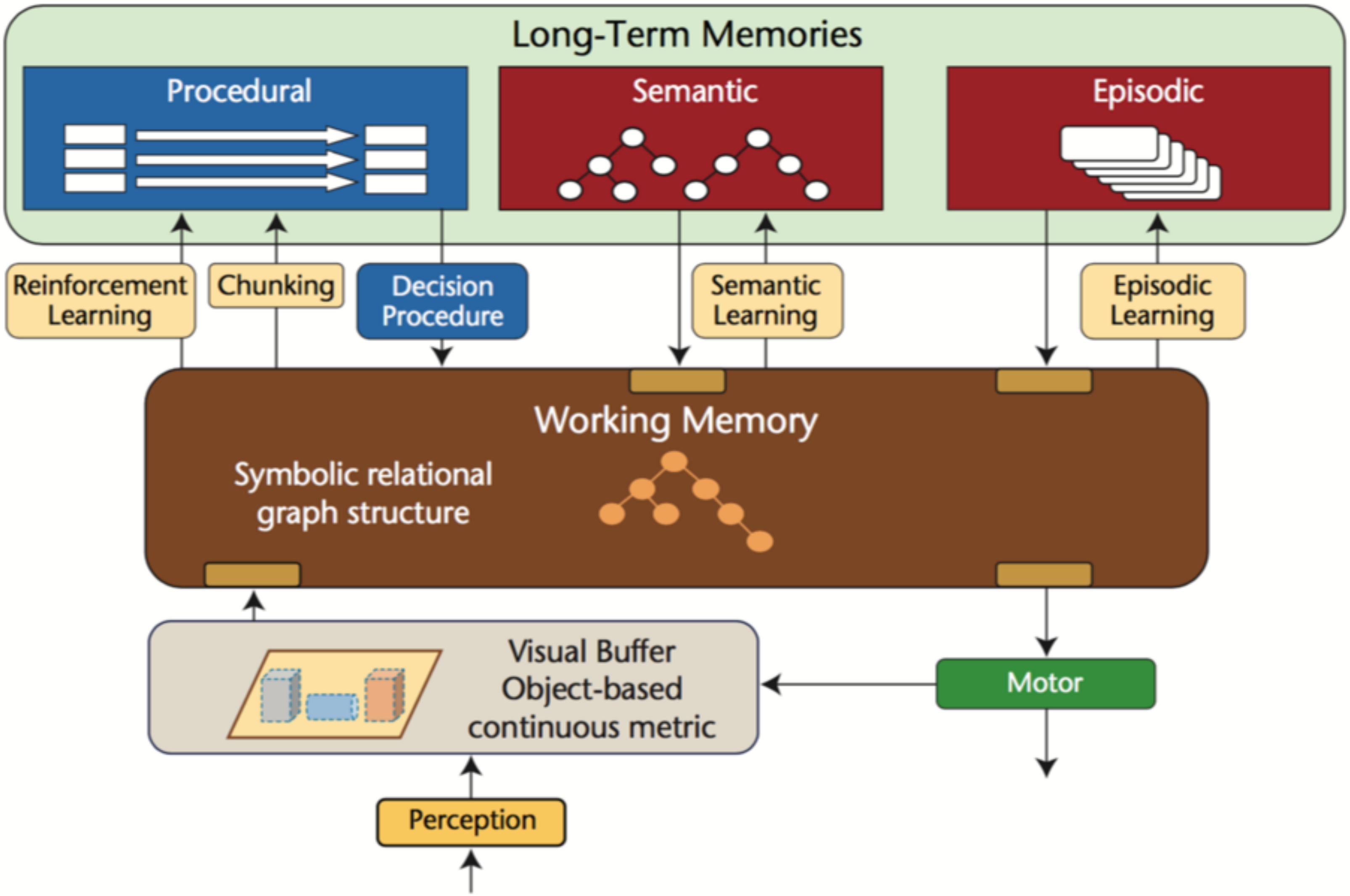
- Now integrates multiple modern forms of representation/learning

- Basis for future Virtual Humans projects @ ICT



Soar

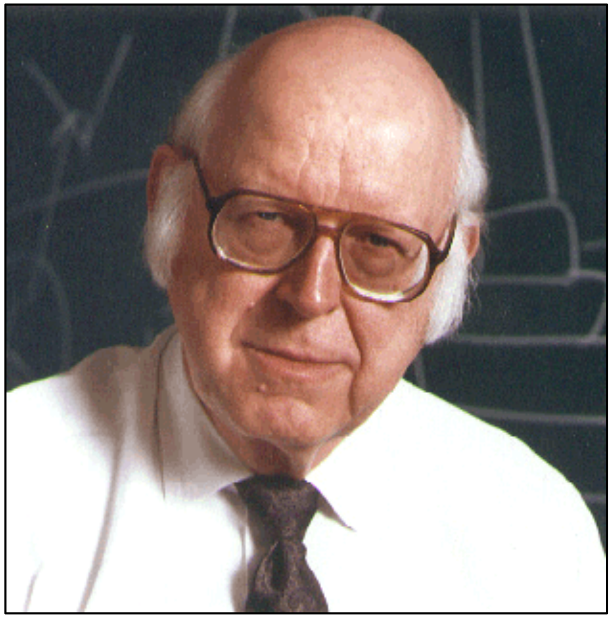
https://soar.eecs.umich.edu



Soar People

John Laird
Professor, CS @ U. of Michigan
Co-Founder @ Soar Technology

Allen Newell
Researcher in CS/Psych @ RAND, CMU
Turing Award, Nat. Medal of Science



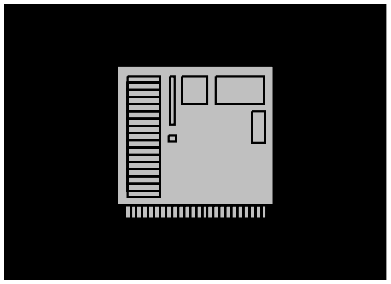
Paul Rosenbloom
Professor of CS @ USC
Director of Cog Arch @ ICT

Soar Notes

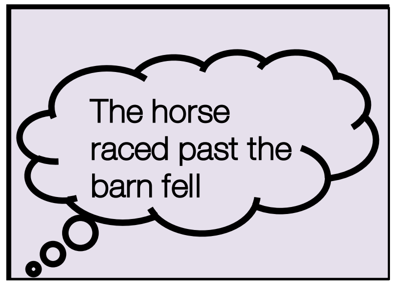
- Focus on efficiency
  - Goal: each decision takes *at most* 50 ms (most agents take much less than 1 ms)
- Public distribution and documentation
  - Major OSs (Windows, macOS, Linux)
  - Many languages (C++, Java, Python, …)
- Annual Workshop

Soar

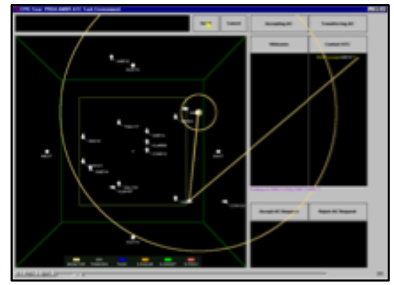
Select Applications (1)



R1-Soar
Computer Configuration

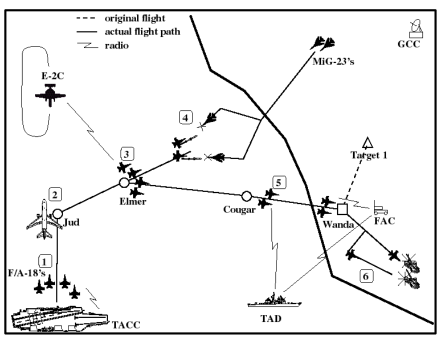


NL-Soar
Language Processing

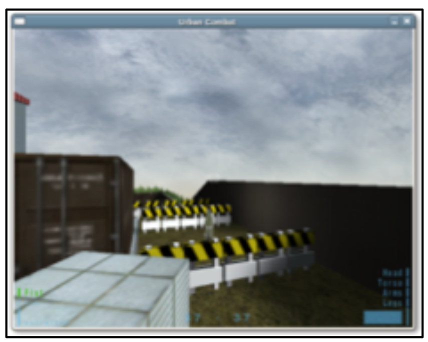


Amber EPIC-Soar
Modeling HCI

ICT Virtual Human
Natural Interaction, Emotion



TacAir-Soar
Complex Doctrine & Tactics



Urban Combat
Transfer Learning

Soar Quakebot
Anticipation

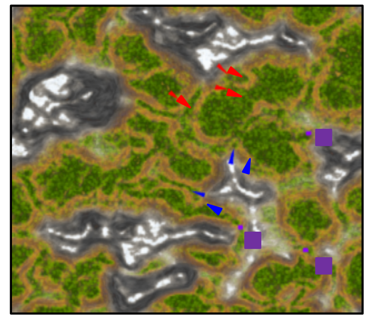
Haunt
Actors and Director

Soar

Select Applications (2)

MOUTbot
Team Tactics & Unpredictable Behavior

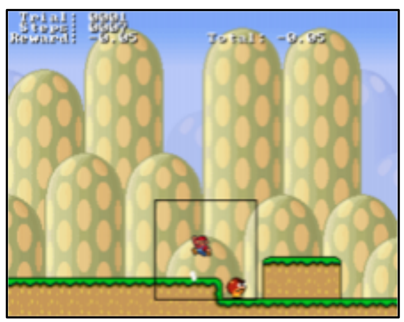
SORTS
Spatial Reasoning & Real-time Strategy



Simulated Scout
Mental Imagery

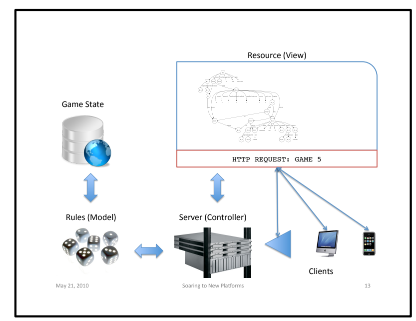
Splinter-Soar
Robot Control

ReLAI
Mental Imagery & Reinforcement Learning



Infinite Mario
Hierarchical Reinforcement Learning

iSoar
Mobile Reinforcement Learning



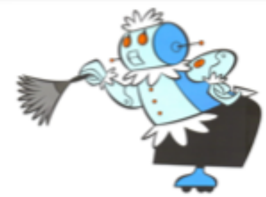
RESTful Soar
Web-based Gameplay, Probabilistic Learning

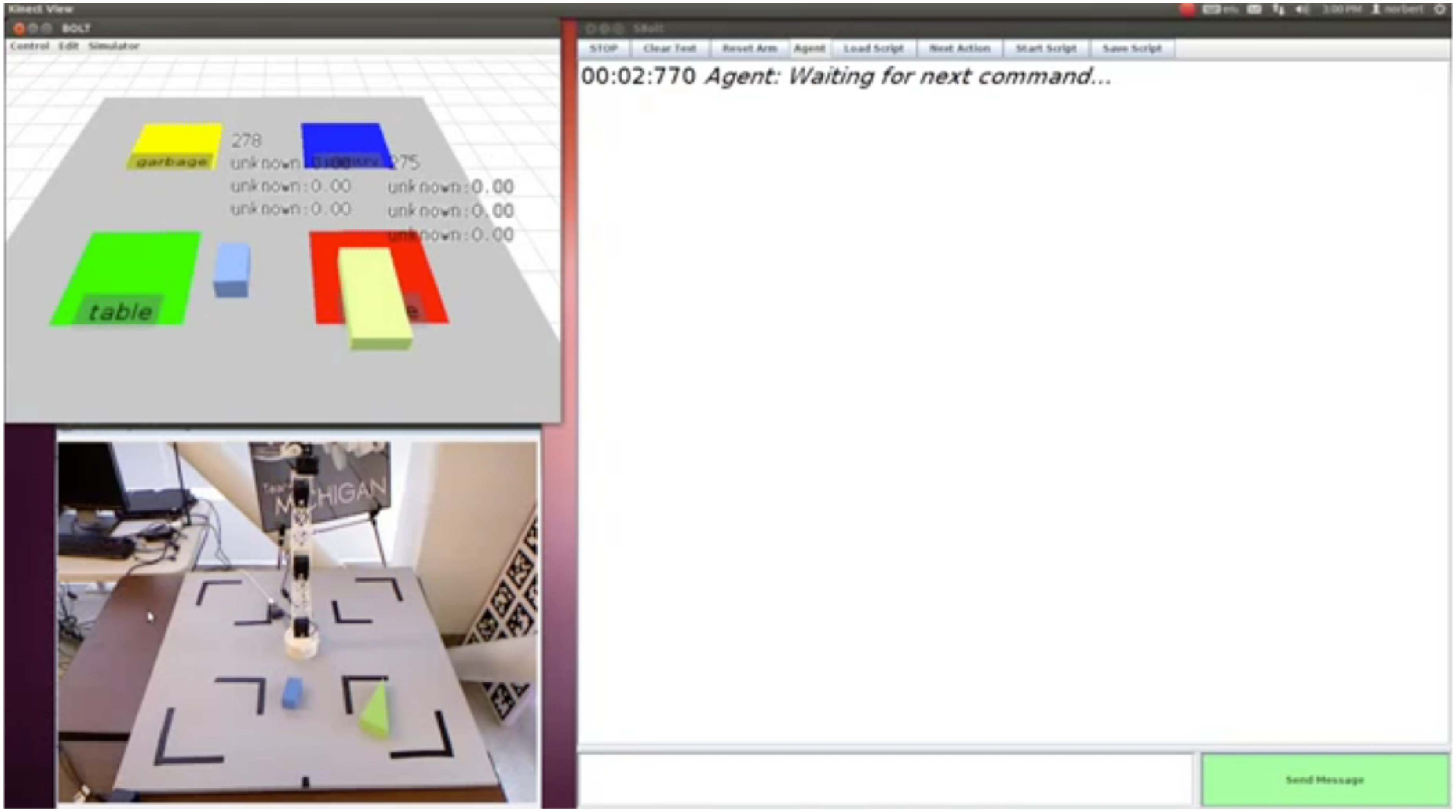
LuminAI

ADAM Lab @ GATech

Rosie

Soar Group @ UMich





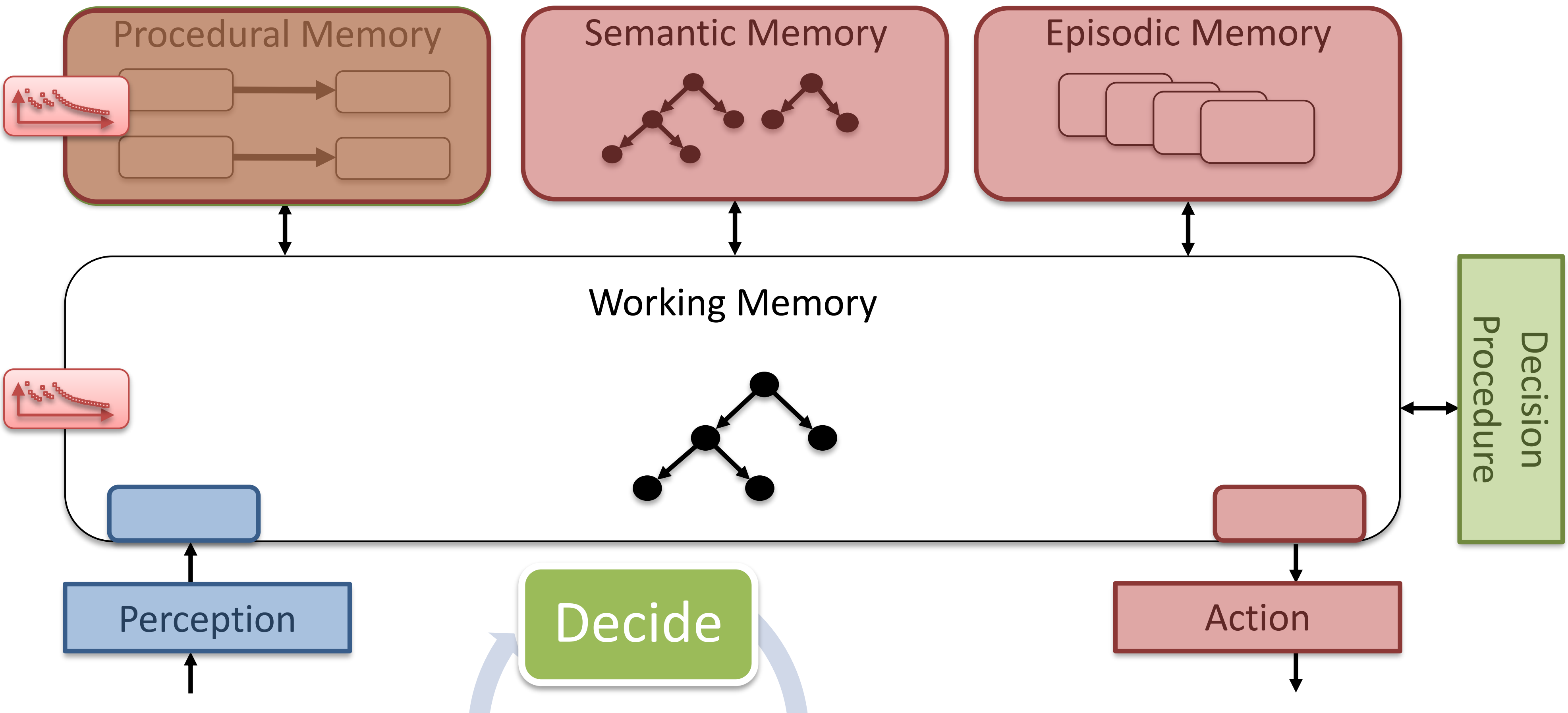
Dueling Research Foci in Soar

- **Architectural enhancement,** must be…
  - useful across a wide variety of agents
  - task-independent
  - efficient

- **Agent development**, to…
  - explore the bounds of architectural commitments/integration
  - solve interesting problems

Soar 9 [Laird 2012]

Memory Integration



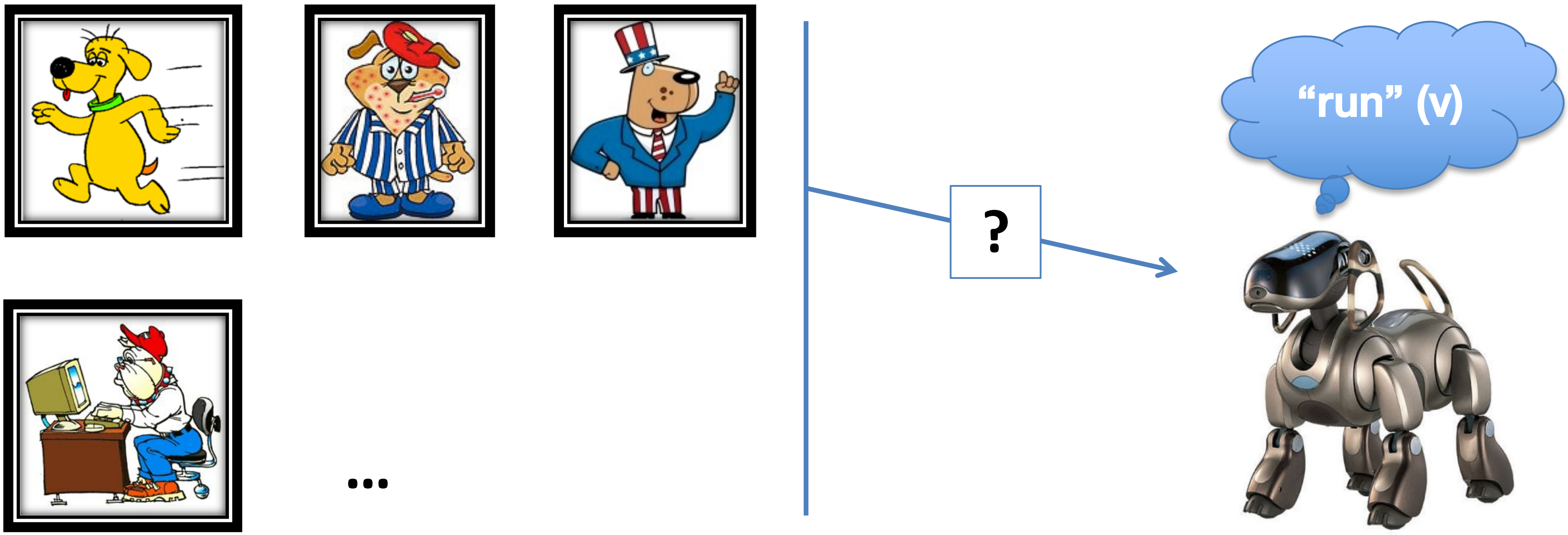
Agent **reactivity** is the time required to execute each decision cycle, including LT memory access ∴ only bounded search!

One Research Path

- **Problem.** Given knowledge + ambiguous cue, what should a fixed LTM mechanism return?
- Clue via "Rational Analysis of Memory" [Anderson et al. 2004]: frequency + recency of use (*Base-Level Activation*)
- Analysis: works well in WSD **[AAAI 2011]**
- Efficiency: new algorithms to scale **[ICCM 2010]**
- Found empirically that the approach yielded beneficial behavior across architectural mechanisms & tasks **[ACS 2012] [CSR 2013]**
  - Semantic LTM Retrieval: WSD
  - WM Decay: Robotic navigation
  - Procedural Decay: RL value-function representation in dice game

Problem ala Word-Sense Disambiguation

Memory **Agent**



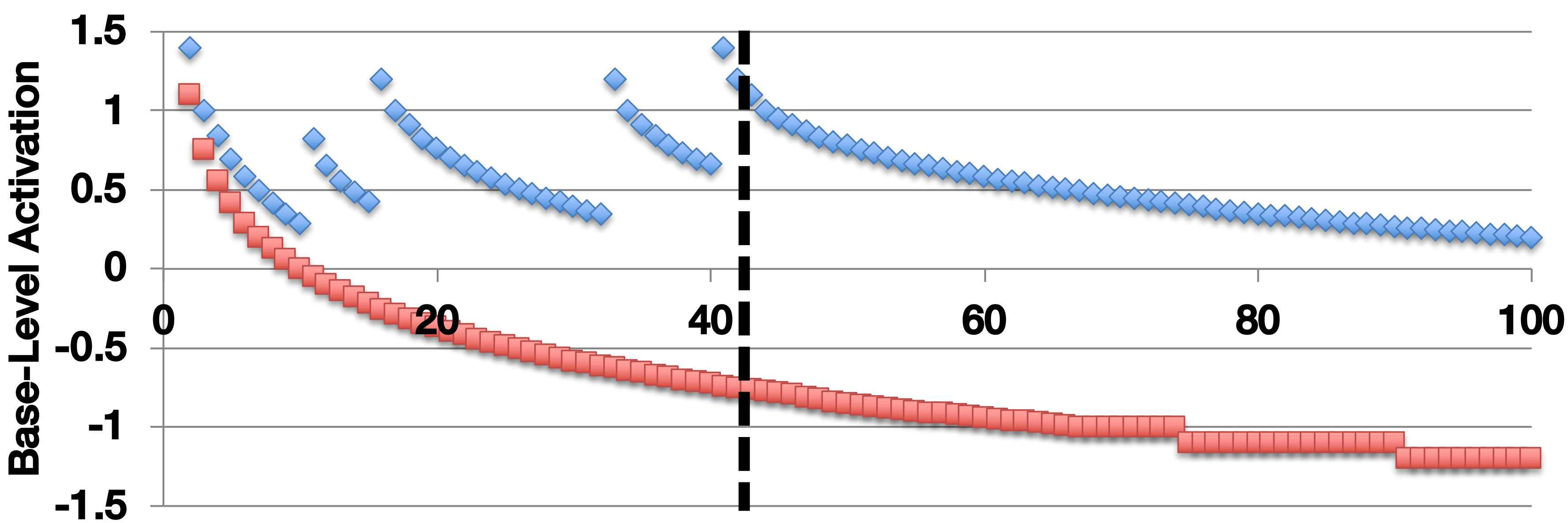
Base-Level Decay

[Anderson et al. 2004]

Predict future usage via history

Used to model human retrieval bias, errors, and forgetting via failure

$$\ln\left(\sum_{j=1}^{n} t_j^{-d}\right)$$

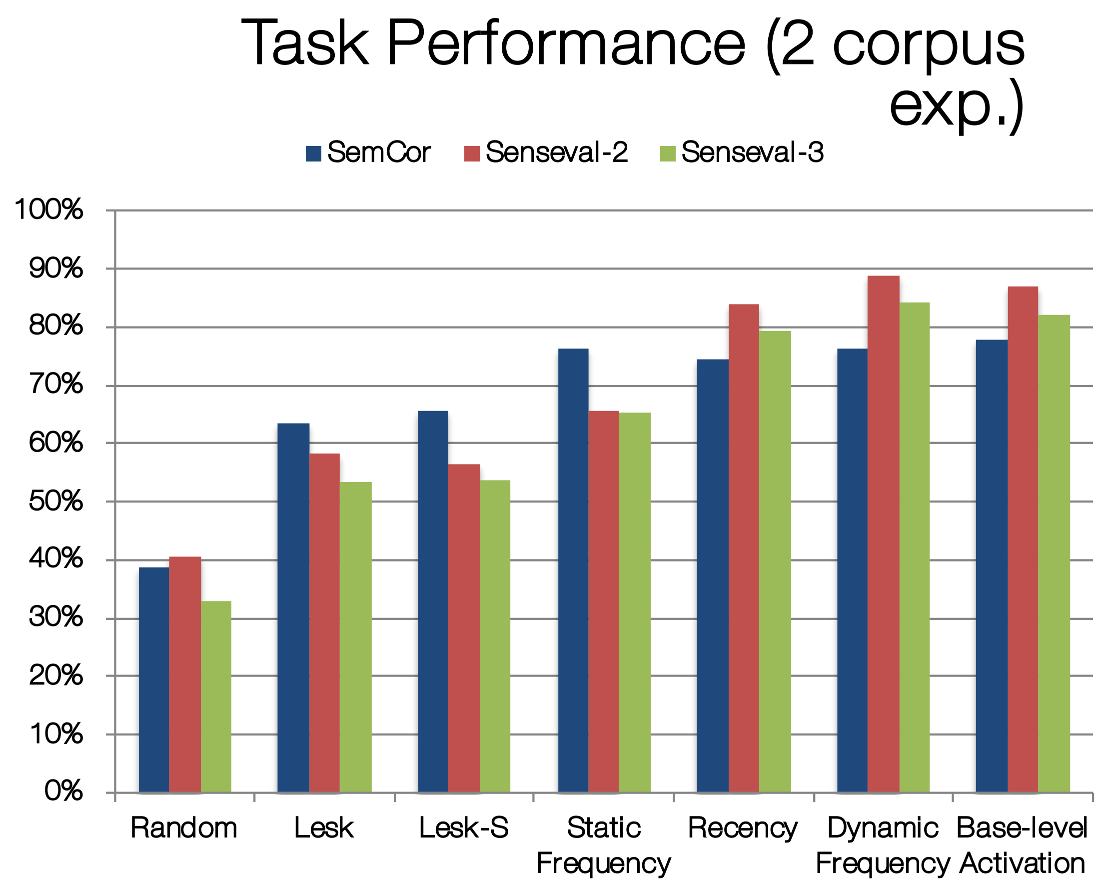


WSD Evaluation

Historical Memory Retrieval Bias

Experimental Setup

- Input: "word", POS
- Given: WordNet v3
  - Correct sense(s) after each attempt



Efficiency

Approach

- Bound *n*, approximate tail effects [Petrov 2006]
- Since present over-estimates future, and only need relative ranking, re-compute on update

$$\ln\left(\sum_{j=1}^{n} t_j^{-d}\right)$$

	SemCor	Senseval-2	Senseval-3
Max. Query Time	1.34 msec	1.00 msec	0.67 msec
Task Perf. Difference	0.82%	-0.56%	-0.72%
Minimum Model Fidelity*	90.30%	95.70%	95.09%

* The smallest proportion of senses that the approximation selected within a corpus exposure that matched those of the base-level activation model.

Related Problem: Memory Size

Mobile Robotics

- Long-lived system
- Building a map in working memory for planning/navigation
- Large WM = large episodes = long time to reconstruct experience

Liar's Dice

- RL: many games
- Sparsely representing an enormous value function
- Large procedural memory = limiting agent lifetime

Hypothesis: Rational to forget a memory if…

1. not useful (via base-level activation) &
2. likely can reconstruct if necessary

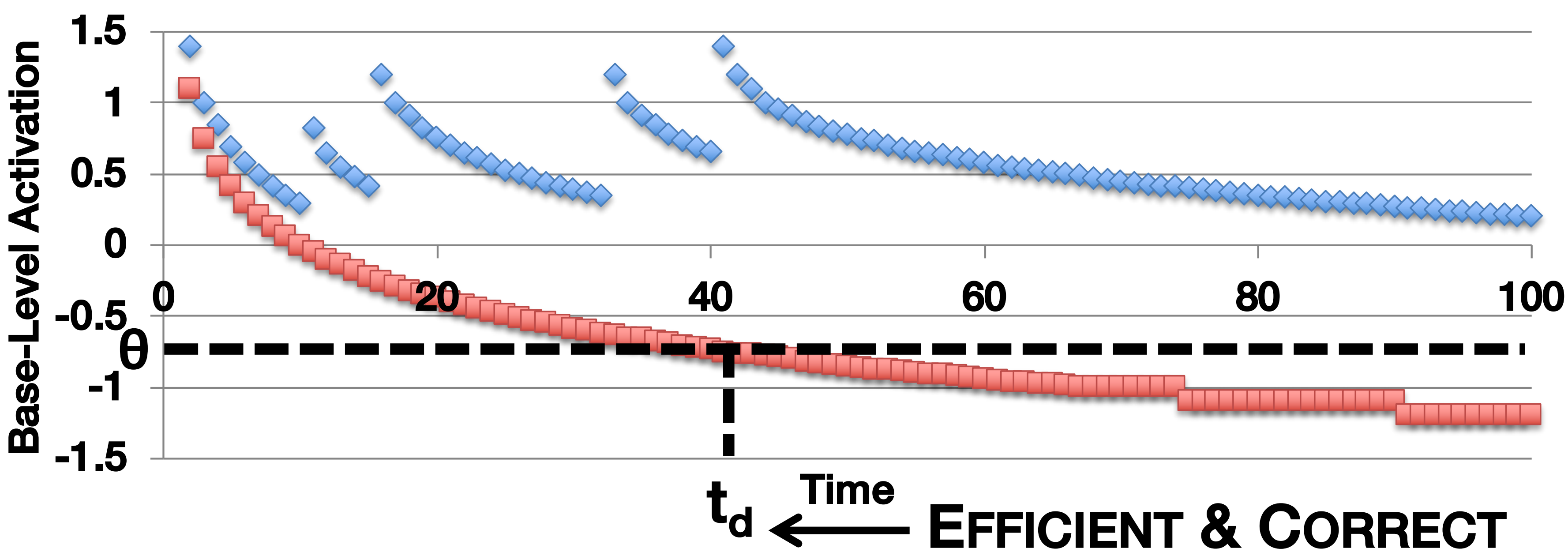
Base-Level Decay

[Anderson et al. 2004]

Predict future usage via history

- Used to model human retrieval bias, errors, and forgetting via failure

$$\ln\left(\sum_{j=1}^{n} t_j^{-d}\right)$$



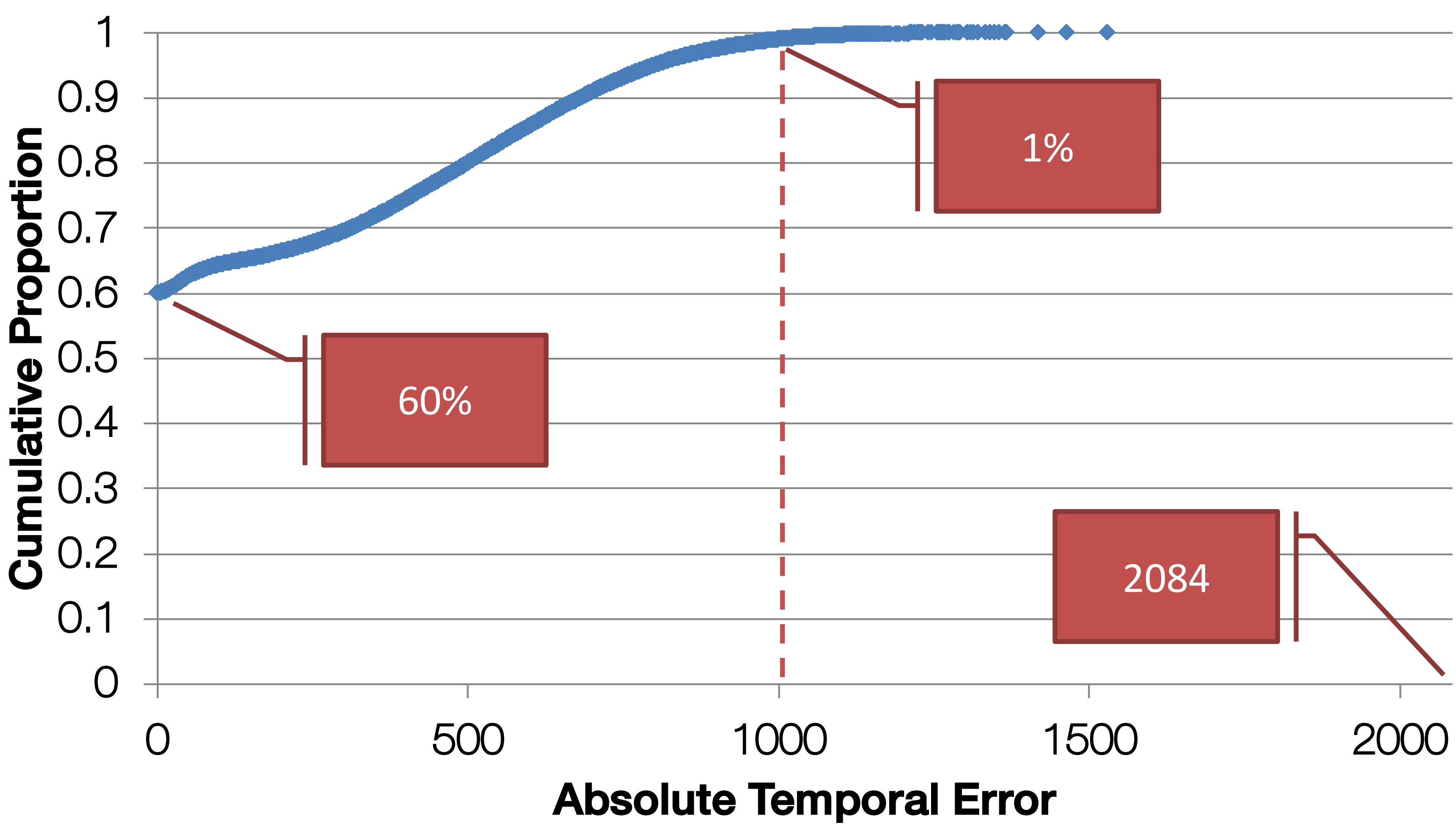
EFFICIENT & CORRECT

~~O(# memories)~~

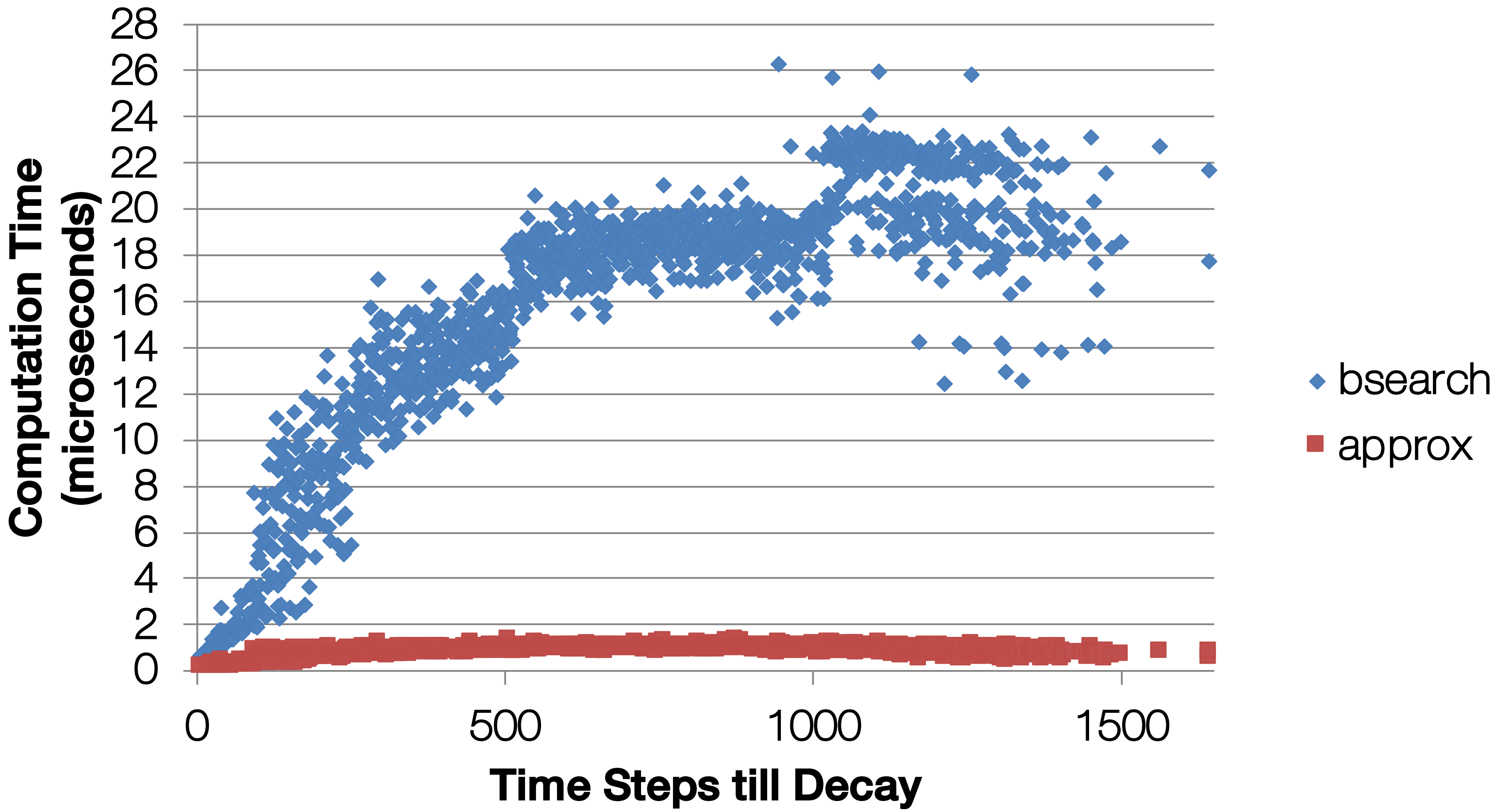
Efficient Implementation

- Underestimate time of decay
  - Approximate as sum of the decay of individual accesses (can compute exactly)
- At predicted time…
  - If below threshold, forget
  - Otherwise, re-estimate

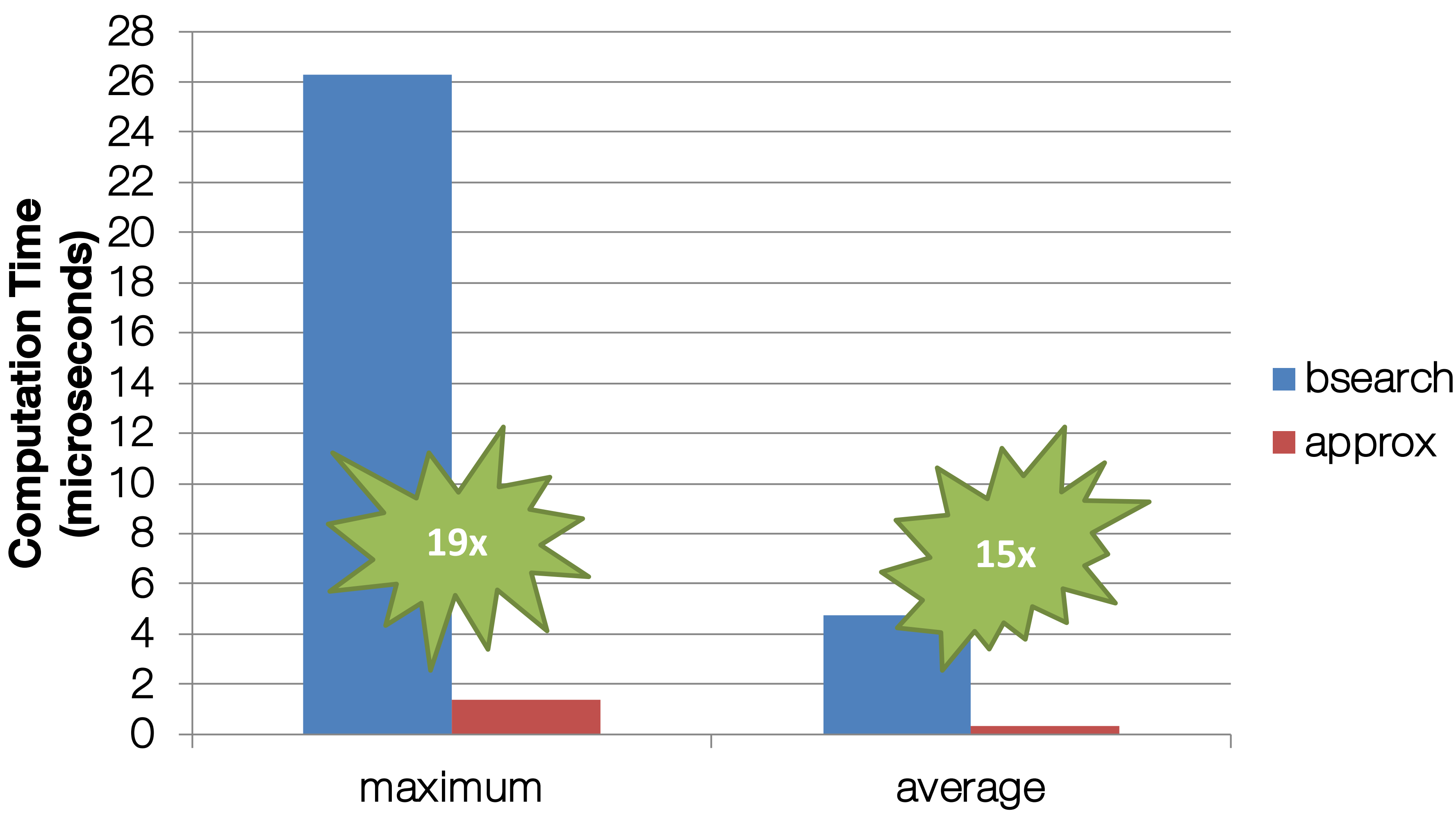
Approximation Quality



Prediction Complexity



Prediction Computation

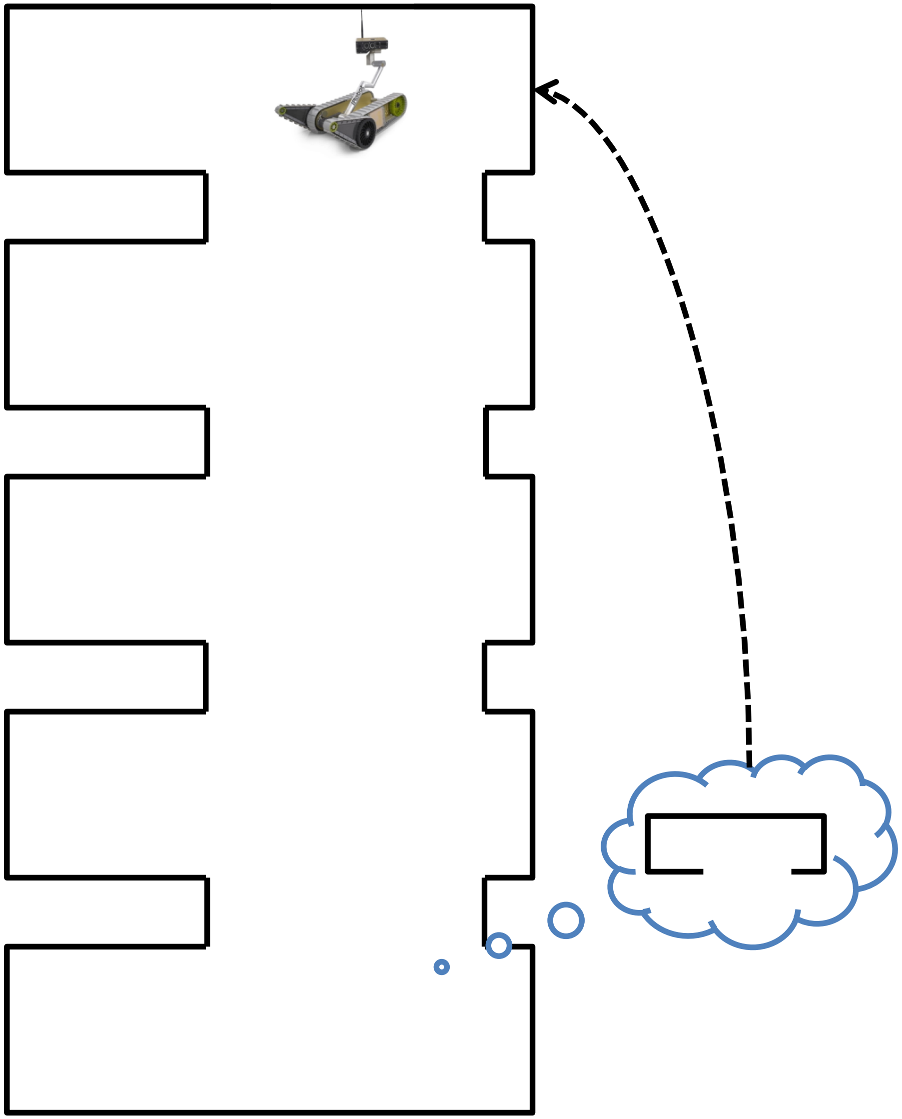


Task #1: Mobile Robotics

Simulated Exploration & Patrol

- 3rd floor, BBB Building, UM
  - 110 rooms
  - 100 doorways
- Builds map in memory from experience

Map Knowledge



Room Features

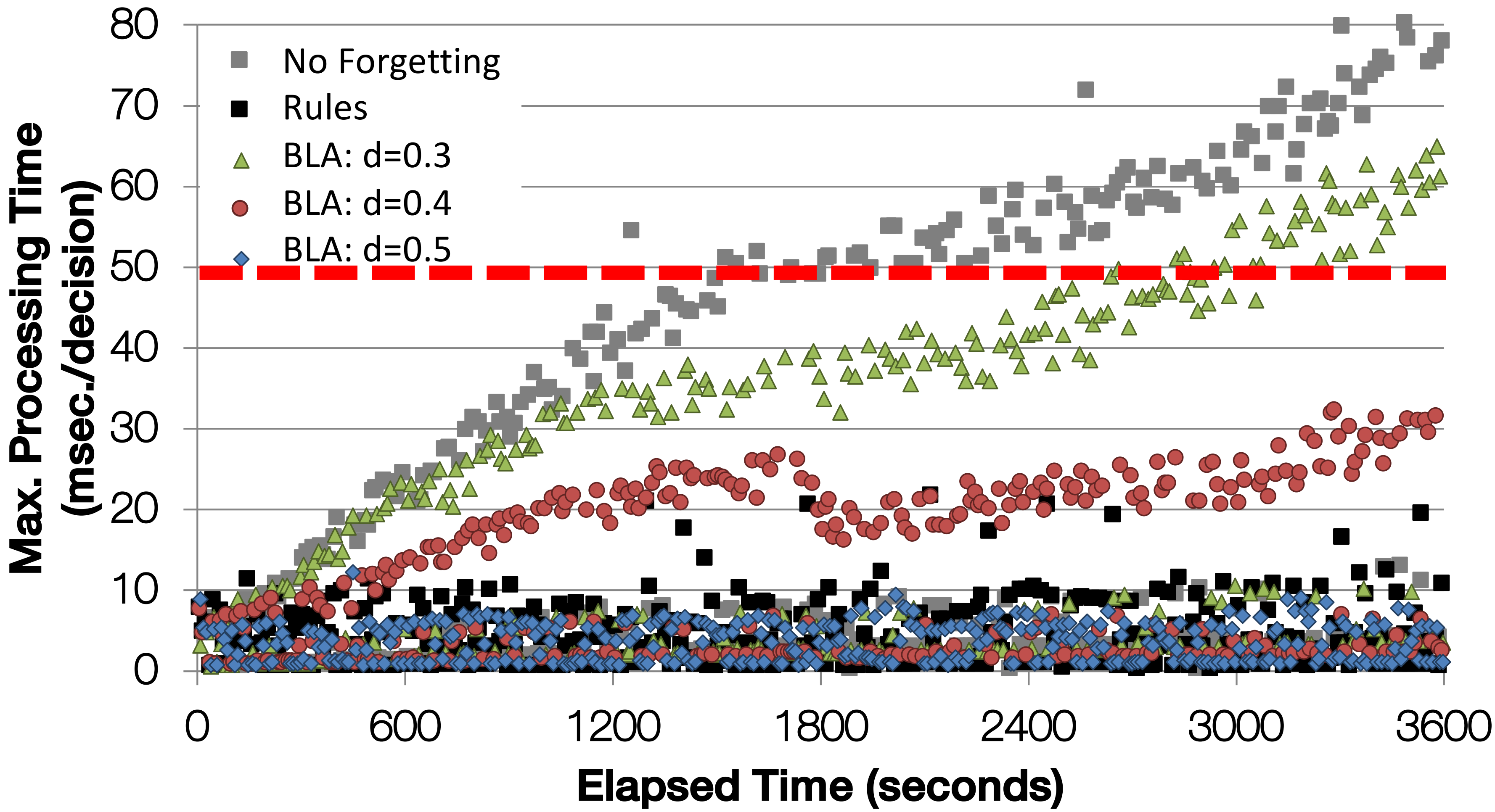
- Position, size
- Walls, doorways
- Objects
- Waypoints

Usage

- Exploration (-->SMem)
- Planning/navigation (<--SMem)

  Reconstruction

Results: Decision Time

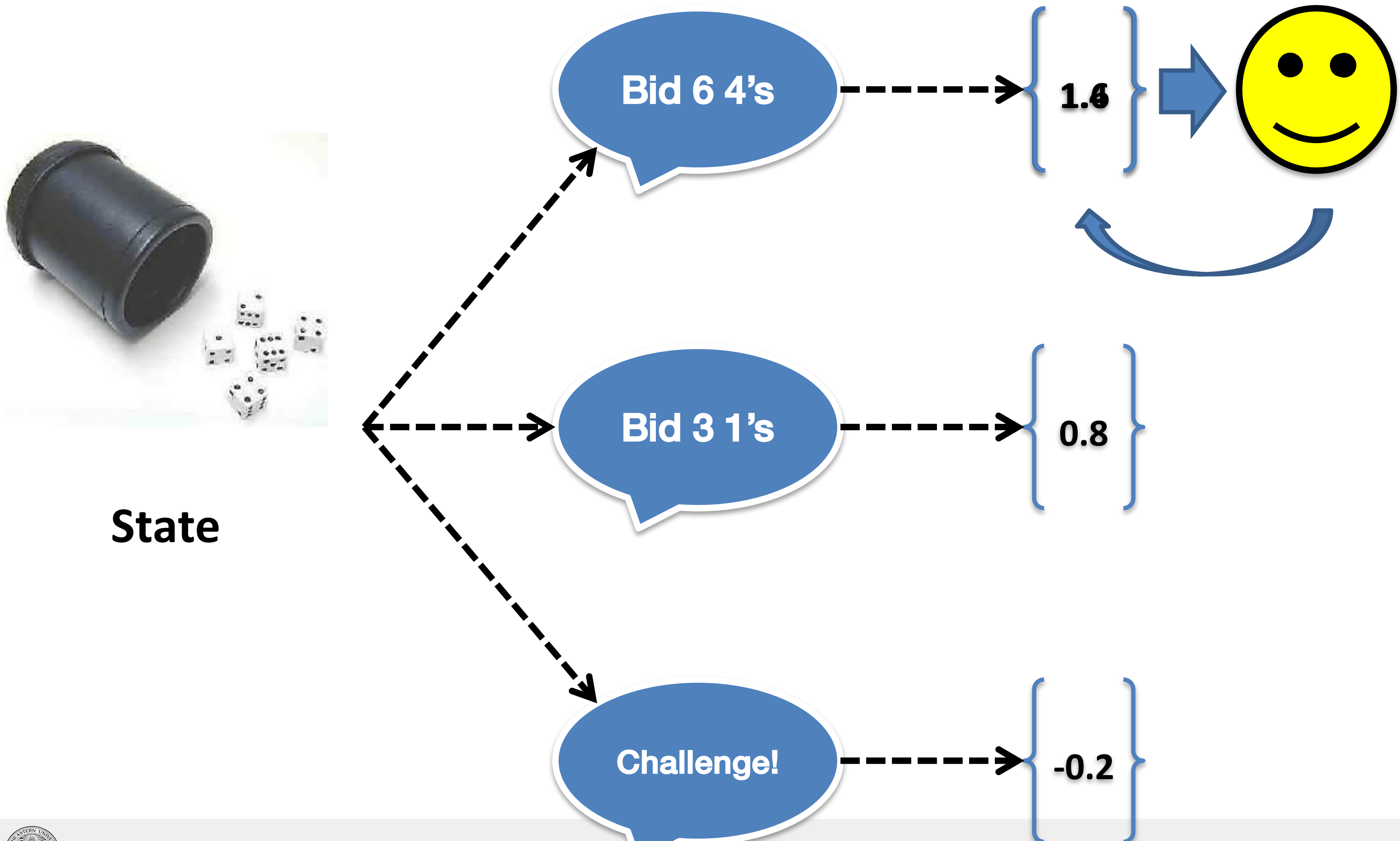


Task #2: Liar's Dice

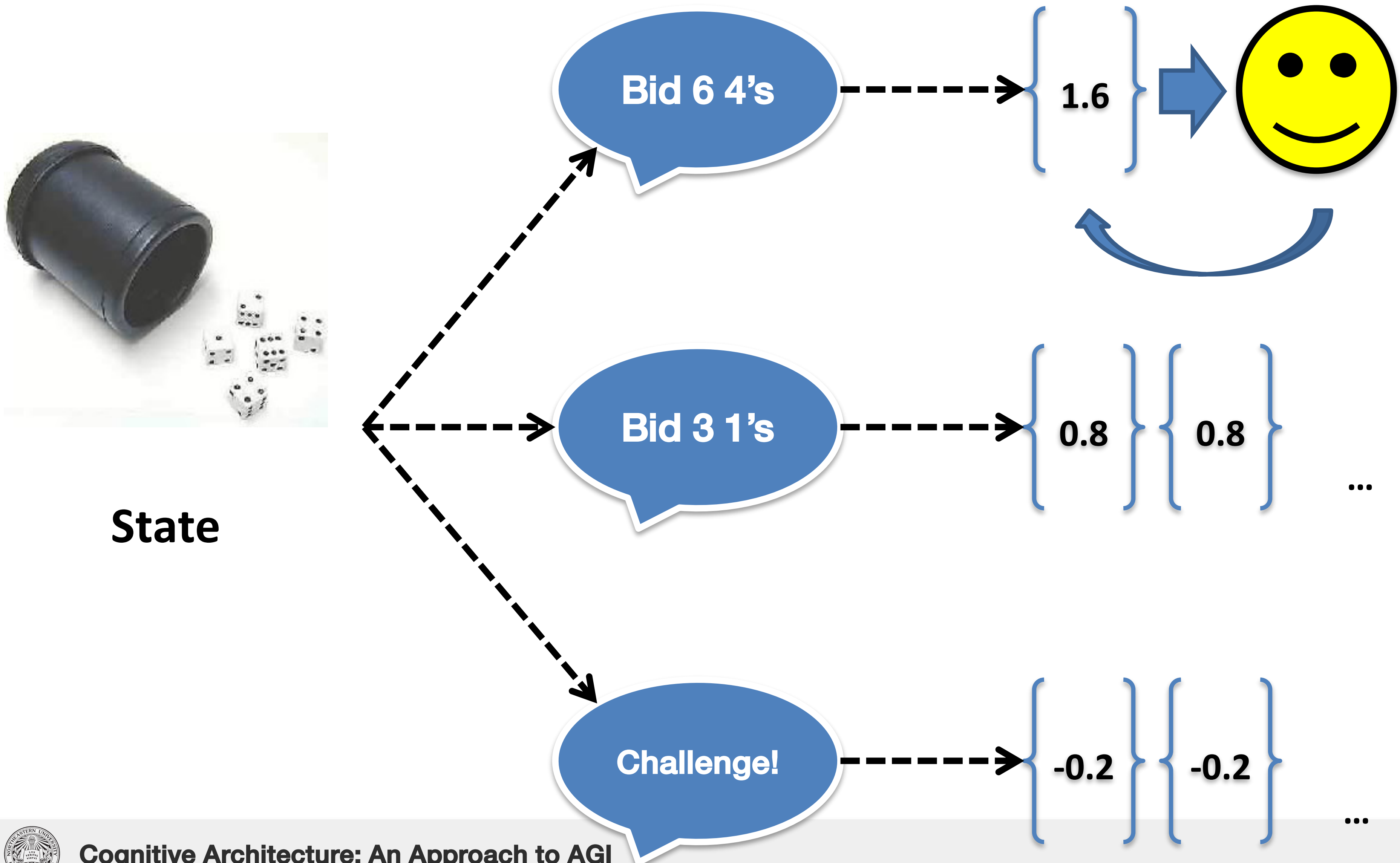
Michigan Liar's Dice

- Complex rules, hidden state, stochasticity
  - Rampant uncertainty
- Agent learns via reinforcement (RL)
  - Large state space (10^6-10^9 for 2-4 players)

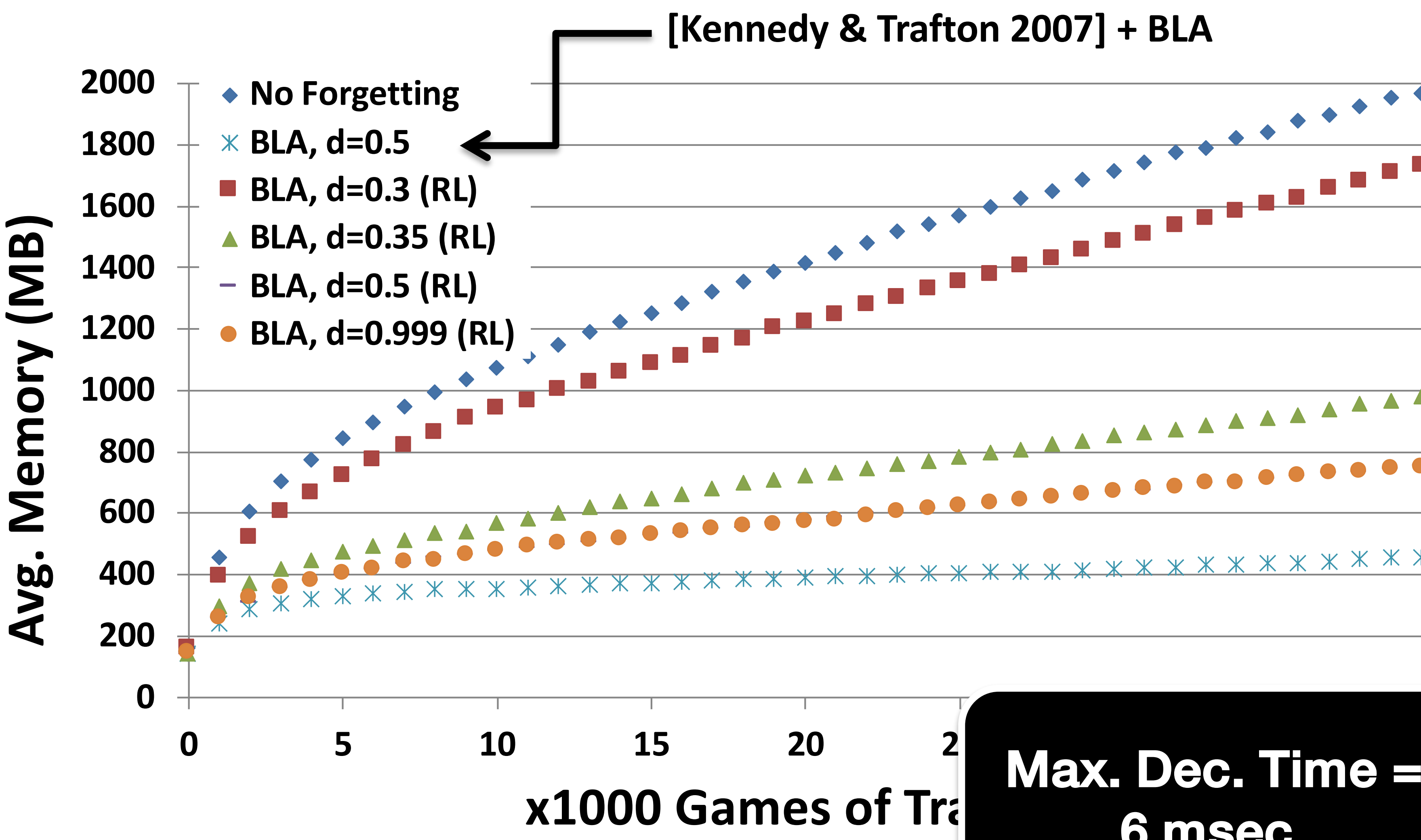
Reasoning --> Action Knowledge



Forgetting Action Knowledge

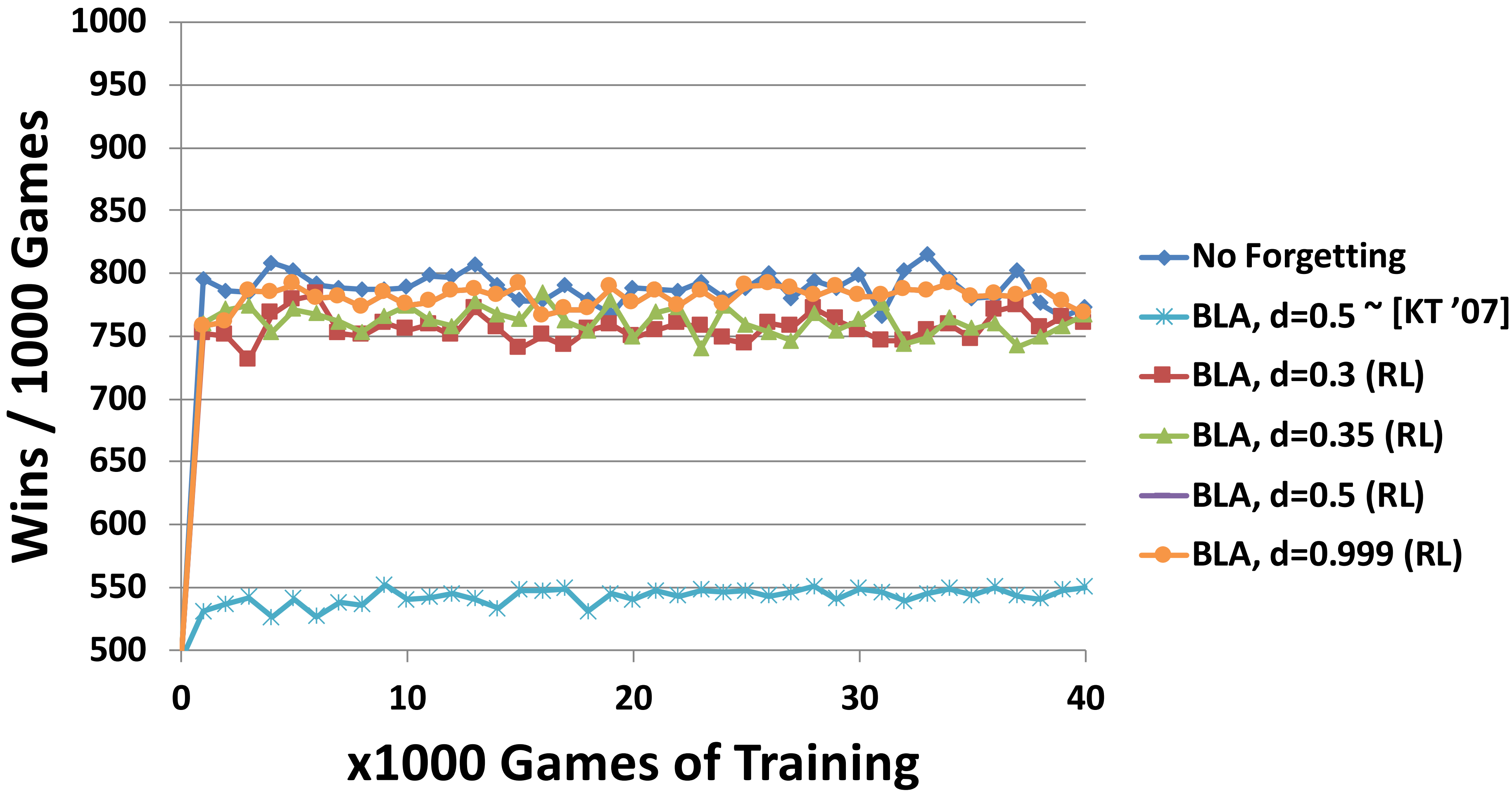


Results: Memory Usage



Max. Dec. Time = 6 msec.

Results: Competence



Conclusions

- Human-inspired estimate of future need (i.e. Base-Level Activation) served as a useful heuristic for memory ranking and forgetting signal in multiple tasks
- Novel algorithms to efficiently implement these as fixed, task-general mechanisms within Soar

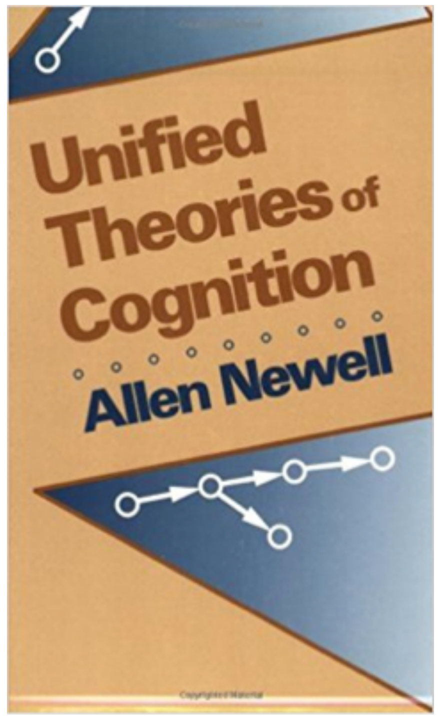
Some CogArch Open Issues

- Integration of models/agents
  - Transfer learning
  - Cross-architectural comparisons
- Multi-modal representations, memory, processing
  - Related: symbol grounding
- Meta-cognition
  - Self-monitoring of agent's own cognitive processes, goal setting
- Ethical (i.e. *What if we succeed?*)

CogArch “vs” (Deep) Machine Learning

- Often tackling different problems
  - And that’s a good thing! (right tool for the right job)
- Can be complimentary
  - ML integration for perceptual processing, feature extraction, learning, actuation, …
  - CogArch for naturally encoding known processes in an associative fashion

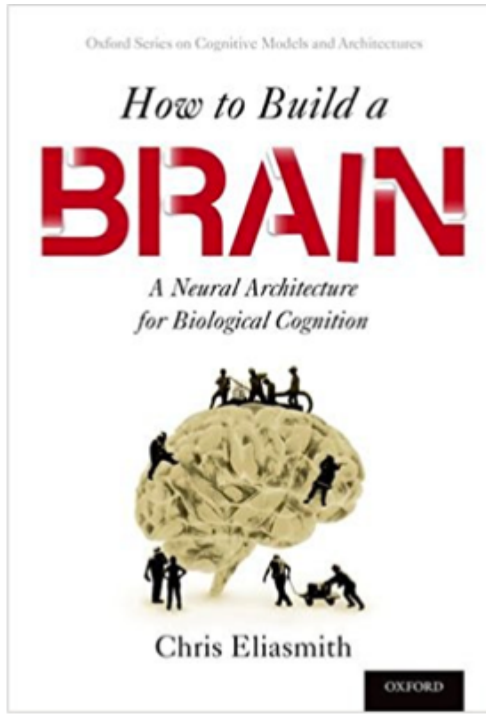
Recommended Reading (1)



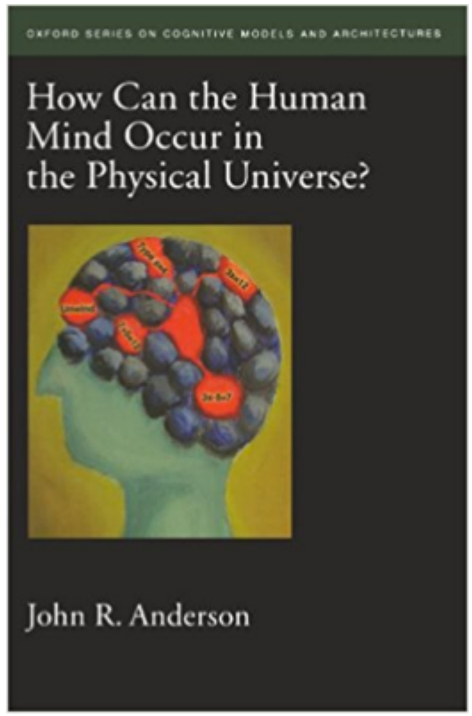
Recommended Reading (2)

Full disclosure: I am an author, but all proceeds have been donated to the Soar group at the University of Michigan.

Recommended Reading (3)



Recommended Reading (4)



Recommended Reading (5)

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research xxx (2008) xxx–xxx

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Cognitive architectures: Research issues and challenges

Action editor: Ron Sun

Pat Langley [a,*], John E. Laird [b], Seth Rogers [a]

[a] *Computational Learning Laboratory, Center for the Study of Language and Information, Stanford University, Stanford, CA 94305, USA*

[b] *EECS Department, The University of Michigan, 1101 Beal Avenue, Ann Arbor, MI 48109, USA*

Received 19 April 2006; accepted 8 July 2006

Abstract

In this paper, we examine the motivations for research on cognitive architectures and review some candidates that have been explored in the literature. After this, we consider the capabilities that a cognitive architecture should support, some properties that it should exhibit related to representation, organization, performance, and learning, and some criteria for evaluating such architectures at the systems level. In closing, we discuss some open issues that should drive future research in this important area.

Keywords: Cognitive architectures; Intelligent systems; Cognitive processes

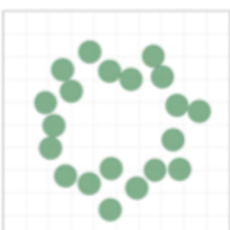
Recommended Reading (6)

Recommended Venues

- AAAI
  - Cognitive Systems Track
- ICCM
  - Cognitive Modeling
- CogSci
  - Cognitive Science
- ACS
  - Advances in Cognitive Systems
- Cognitive Systems Research
- AGI, BICA
- Soar Workshop, ACT-R Workshop

Thank You :)

Questions?



Nate Derbinsky
Associate Teaching Professor
Northeastern University
https://derbinsky.info