A Multi-Domain Evaluation of Scaling in Soar's Episodic Memory

Nate Derbinsky, Justin Li, John E. Laird

University of Michigan

Motivation

Prior Work

- Nuxoll & Laird (‘12): integration and capabilities
- Derbinsky & Laird (‘09): efficient algorithms

Core Question

To what extent is Soar’s episodic memory effective and efficient for real-time agents that persist for long periods of time across a variety of tasks?

Approach: Multi-Domain Evaluation

- Existing agents from diverse tasks (49)
  - Linguistics, planning, games, robotics
- Long agent runs
  - Hours-days RT (10^5 – 10^8 episodes)
- Evaluate at each X episodes
  - Memory consumption
  - Reactivity for >100 task relevant cues
    - Maximum time for cue matching $<^?$ 50 msec.

Outline

- Overview of Soar's EpMem
- Word Sense Disambiguation (WSD)
- Planning
- Video Games & Robotics

Episodic Memory
Problem Formulation

Representation

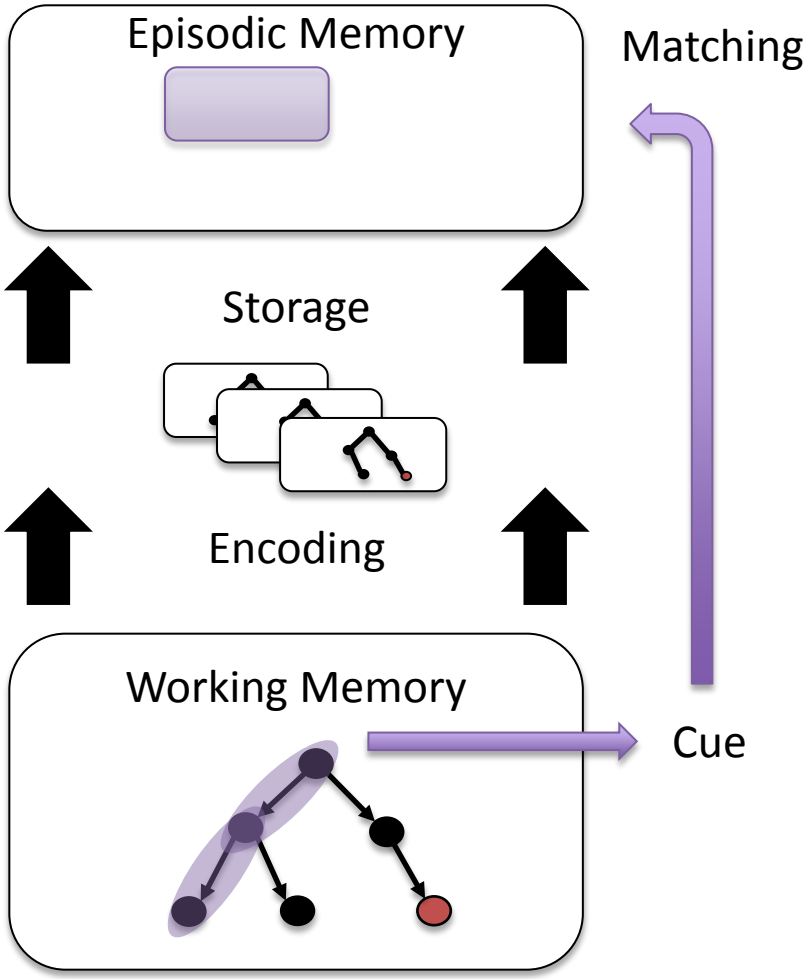
- Episode: connected di-graph
- Store: temporal sequence

Encoding/Storage

- Automatic
- No dynamics

Retrieval

- Cue: acyclic graph
- Semantics: desired features in context
- Find the most recent episode that shares the most leaf nodes in common with the cue



Episodic Memory
Algorithmic Overview

Storage

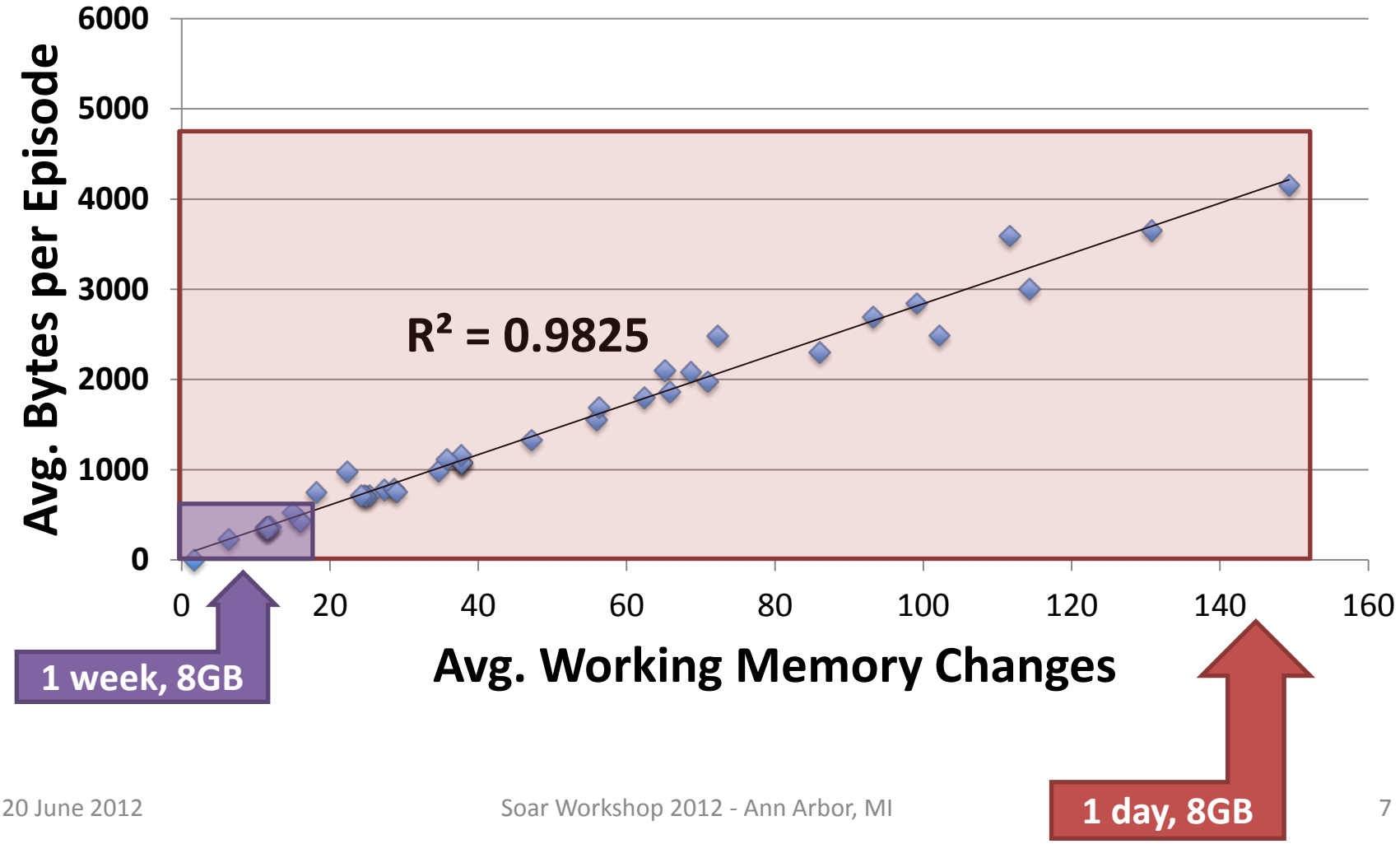
- Capture WM-changes as temporal intervals

Cue Matching (reverse walk of cue-relevant Δ's)

- 2-phase search
  - Only graph-match episodes that have all cue features independently
- Only evaluate episodes that have changes relevant to cue features
- Incrementally re-score episodes

Episodic Memory

Storage Characterization



Episodic Memory
Retrieval Characterization

Assumptions

- Few changes per episode (*temporal contiguity*)
- Representational re-use (*structural regularity*)
- Small cue

Scaling

- Search distance (# changes to walk)
  - *Temporal Selectivity*: how often does a WME change
  - *Feature Co-Occurrence*: how often do WMEs co-occur within a single episode (related to search-space size)
- Episode scoring (similar to rule matching)
  - *Structural Selectivity*: how many ways can a cue WME match an episode (i.e. multi-valued attributes)

Word Sense Disambiguation

Experimental Setup

- Input: <“word”, POS>; Output: sense #; Result
  - Corpus: SemCor (~900K eps/exposure)

- Agent
  - Maintain context as n-gram
  - Query EpMem for context
    - If success, get next episode, output result
    - If failure, *null*

Accuracy	First	Second
2-gram	14.57%	92.82%
3-gram	2.32%	99.47%

Word Sense Disambiguation

Results

Storage

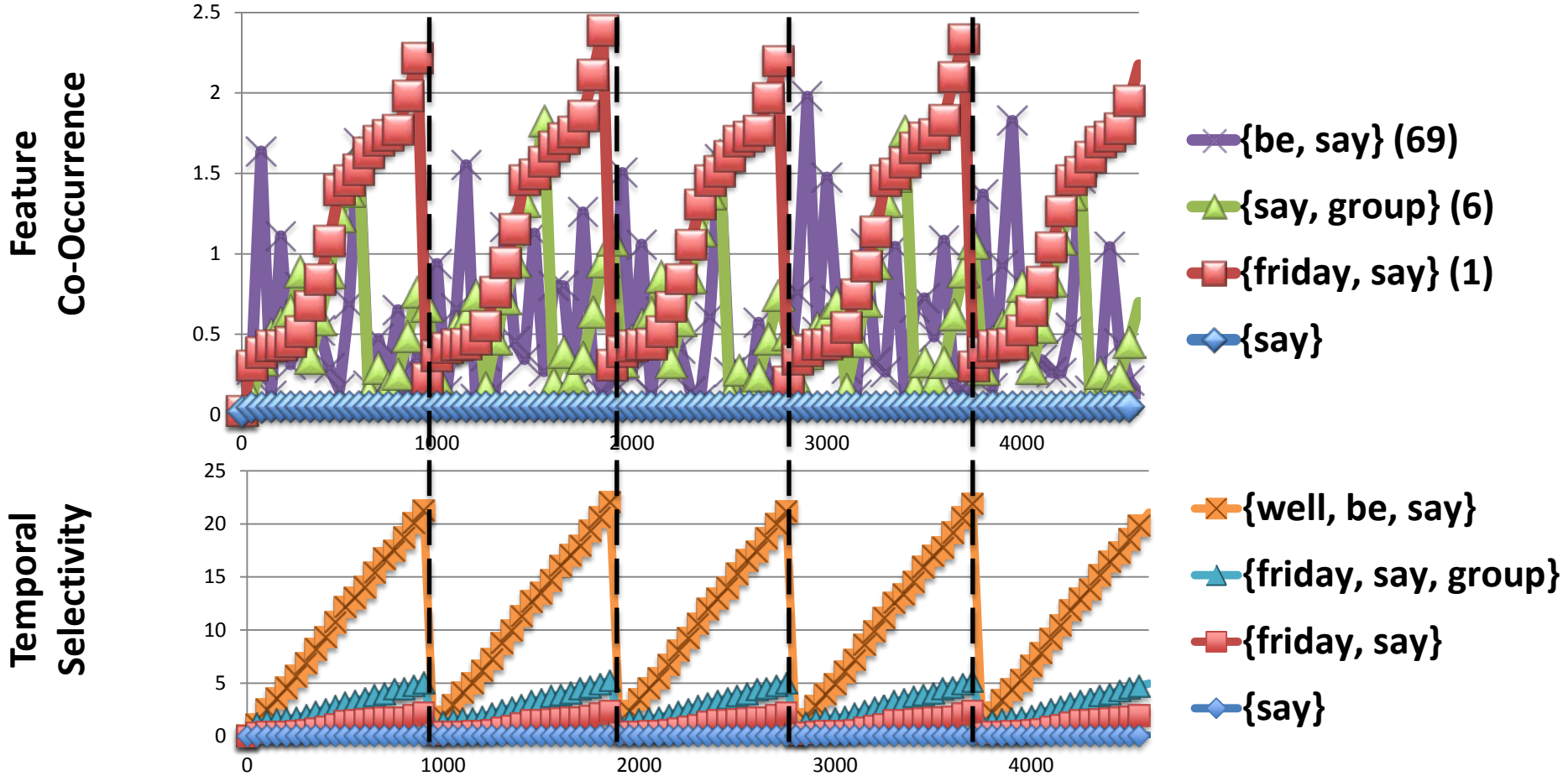
- Avg. 234 bytes/episode

Cue Matching

- All 1-, 2-, and 3-gram cues reactive
- 0.2% of 4-grams exceed 50msec.

N-gram Retrieval Scaling

Retrieval Time (msec) vs. Episodes (x1000)



Planning
Experimental Setup

- 12 automatically converted PDDL domains
  - Logistics, Blocksworld, Eight-puzzle, Grid, Gripper, Hanoi, Maze, Mine-Eater, Miconic, Mystery, Rockets, and Taxi
  - 44 distinct problem instances (e.g. # blocks)

- Agent: randomly explore state space
  - 50K episodes, measure every 1K

Planning
Results

Storage

- Reactive: <12.04 msec./episode
- Memory: 562 – 5454 bytes/episode

Cue Matching (reactive: < 50 msec.)

1. Full State: only smallest state + space size (12)
2. Relational: none
3. Schema: all (max = 0.08 msec.)

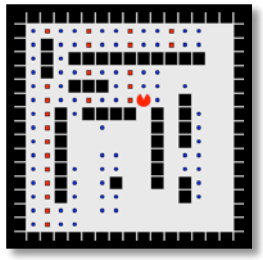
Video Games & Mobile Robotics

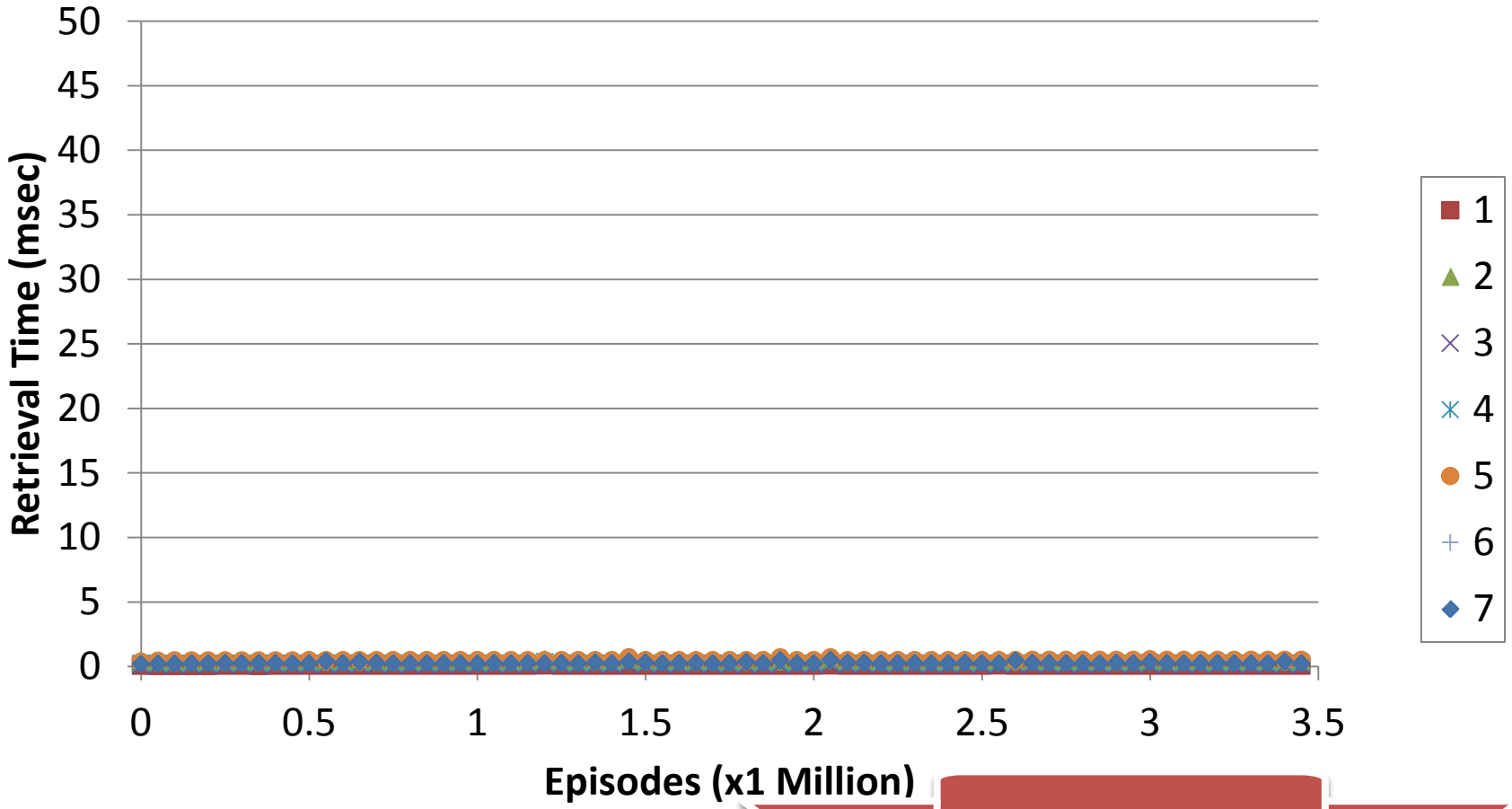
Experimental Setup

	Domain	Agent	Duration	Eval. Rate
	TankSoar	*mapping-bot*	3.5M	50K
	Eaters	*advanced-move*	3.5M	50K
	Infinite Mario	[Mohan & Laird '11]	3.5M	50K
	Rooms World	[Laird, Derbinsky & Voigt '11]	12 hours	300K

- Hand-coded cues (per domain)

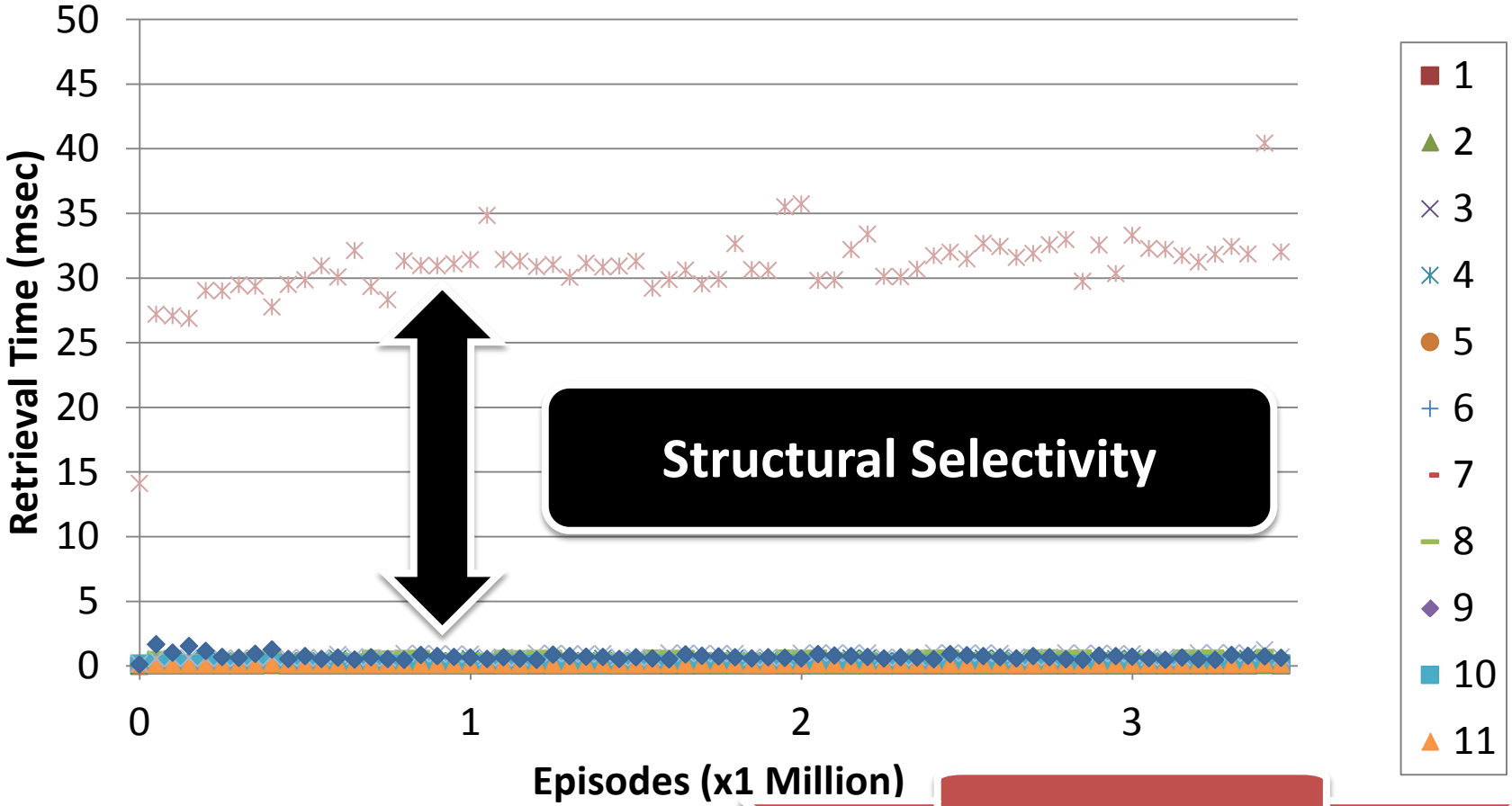
Data: Eaters



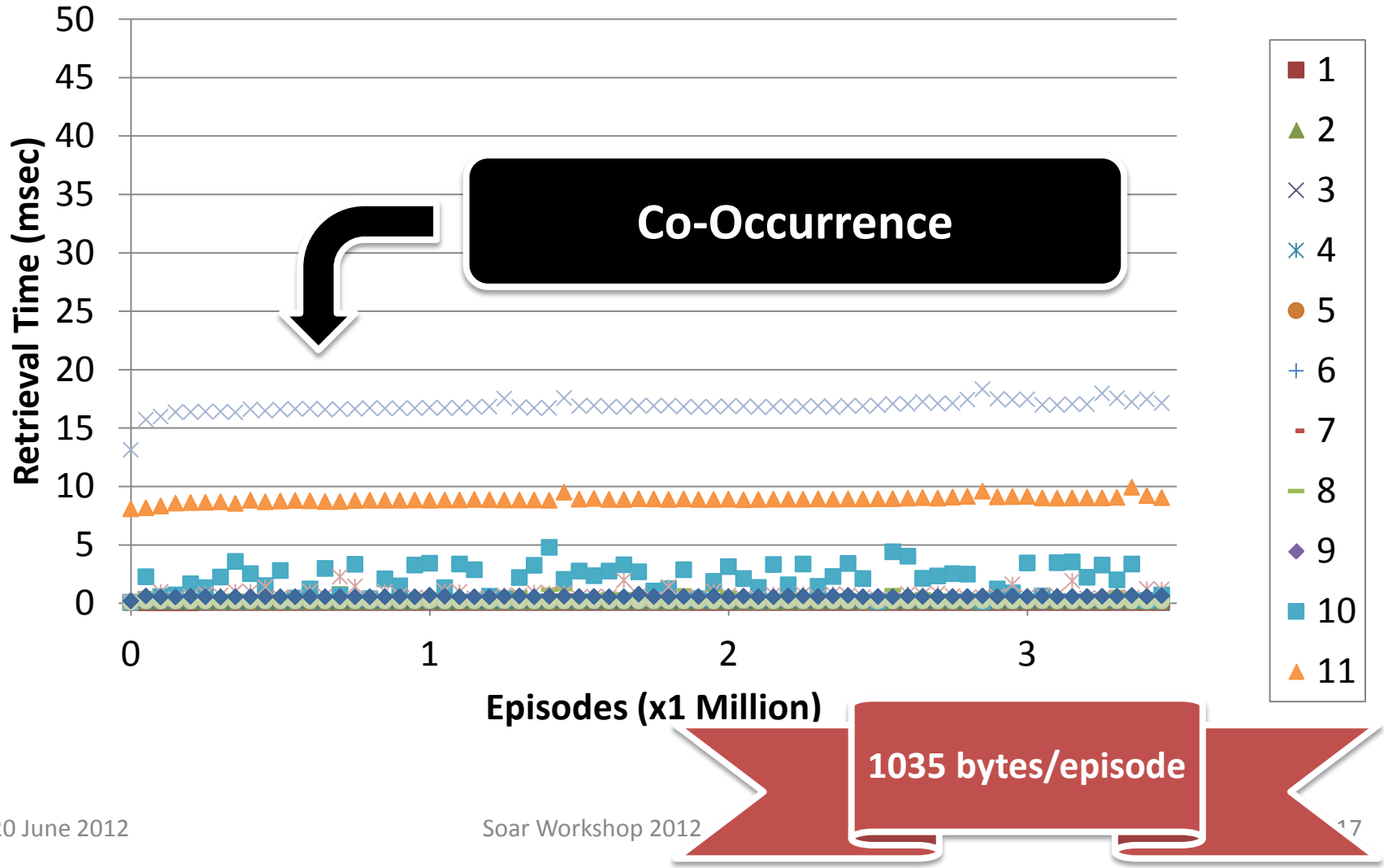


813 bytes/episode

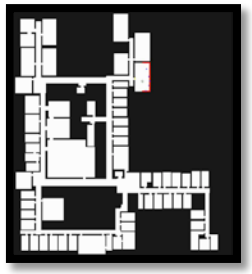
Data: Infinite Mario

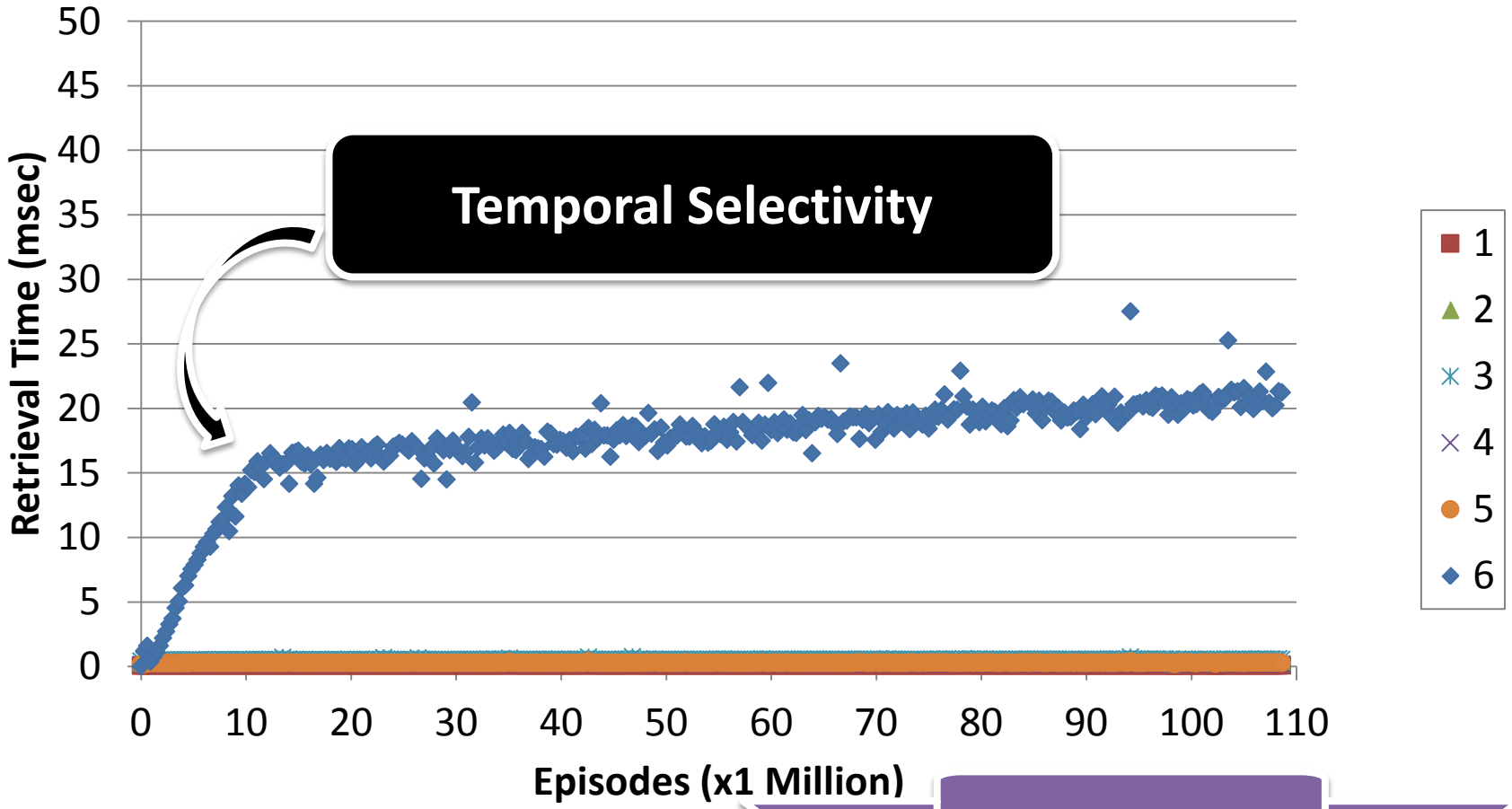


Data: TankSoar



Data: Mobile Robotics





113 bytes/episode

Summary of Results

Generality

- Demonstrated 7 cognitive capabilities
  - Virtual sensing, action modeling, long-term goal management, …

Reactivity

- <50 msec. storage time for all tasks (ex. temporal discontiguity)
- <50 msec. cue matching for many cues

Scalability

- No growth in cue matching for many cues (days!)
  - Validated predictive performance models
- 0.18 - 4 kb/episode (days – months)

Evaluation

Nuggets

- Unprecedented evaluation of general episodic memory
  - Breadth, temporal extent, analysis
- Characterization of EpMem performance via task-independent properties
- Soar's EpMem (v9.3.2) is effective and efficient for many tasks and cues!
- Domains and cues available

Coal

- Still easy to construct domain/cue that makes Soar unreactive
- Unbounded memory consumption (given enough time)

For more details, see paper in proceedings of AAAI 2012