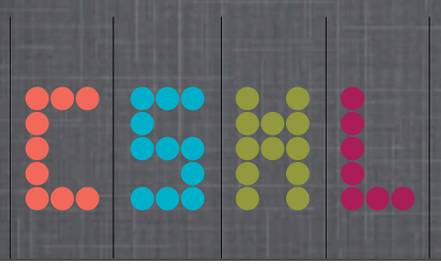
Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search

Arthur Guez, David Silver°, Peter Dayan**

** Gatsby Unit, UCL* *° Dept. of CS, UCL*

Large-scale Online Learning and Decision Making
2012 - Windsor



MOTIVATION:

Solve large tasks with uncertainty about the dynamics.

Want to take advantage of structured prior knowledge.

All rewards (and costs) count from the start, with discounting.

MOTIVATION:

Solve large tasks with uncertainty about the dynamics.

Want to take advantage of structured prior knowledge.

All rewards (and costs) count from the start, with discounting.

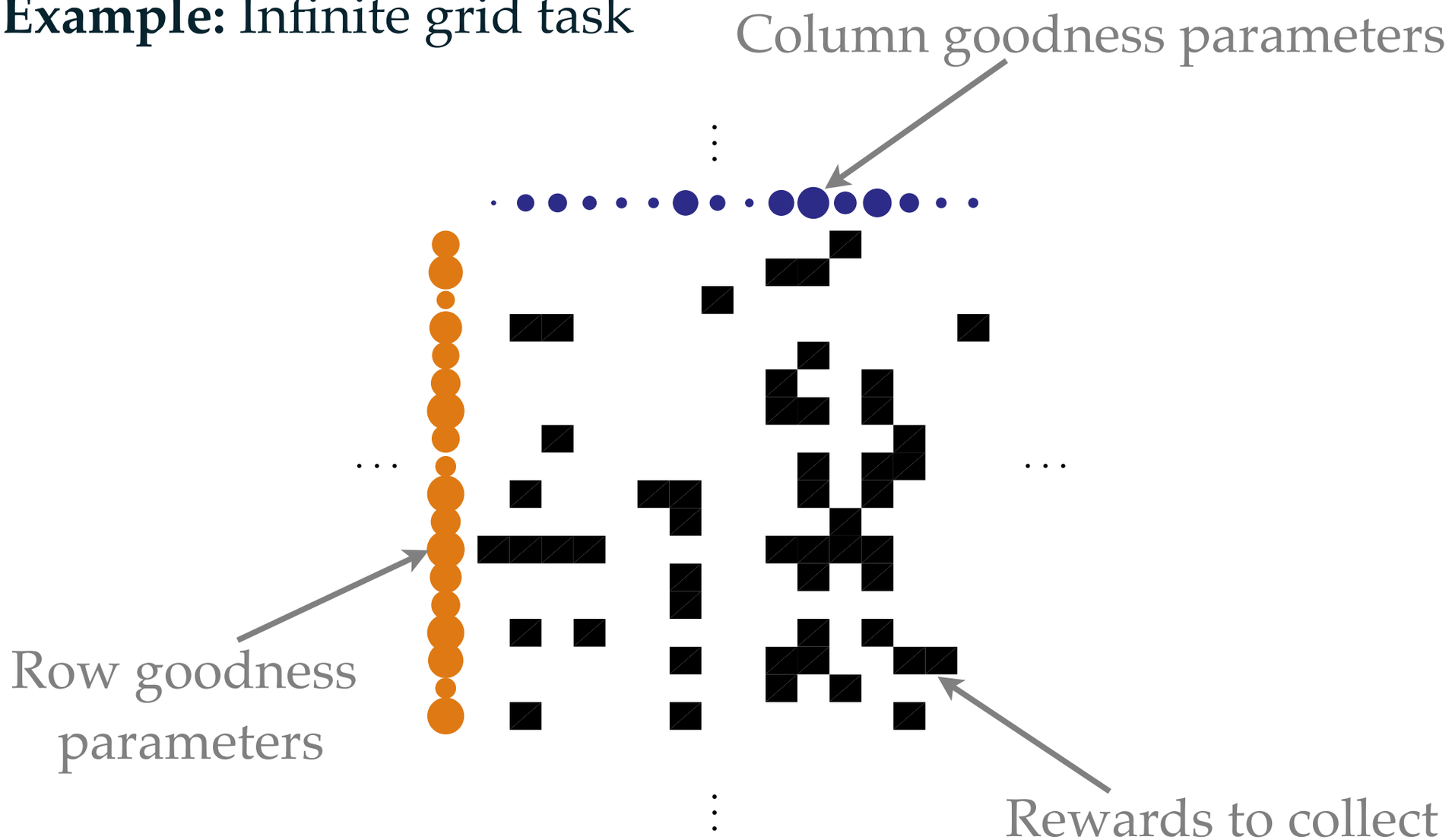
➡ **A natural formalism: Bayesian model-based RL.**

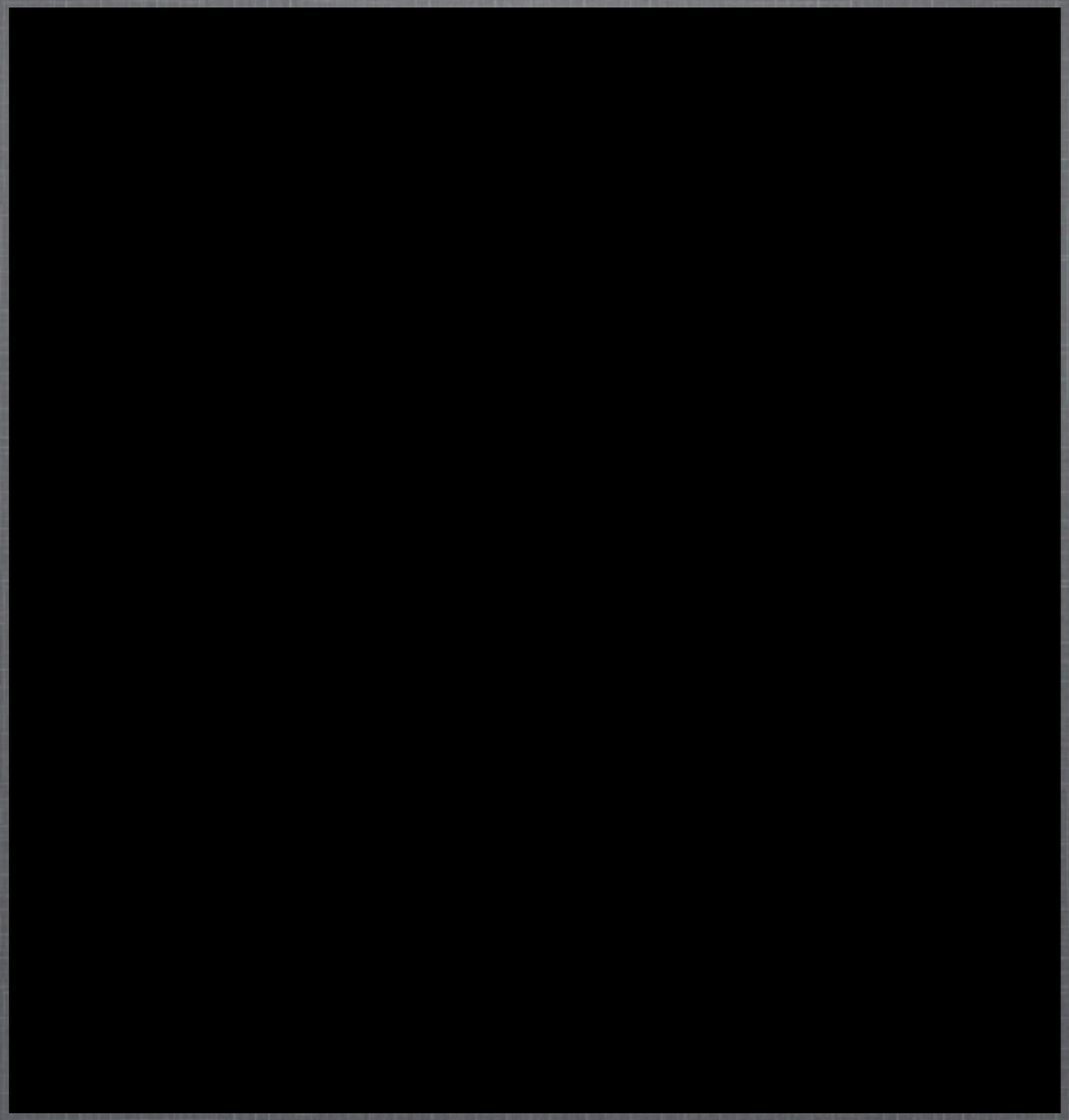
MOTIVATION:

Solve large tasks with uncertainty about the dynamics.
Want to take advantage of structured prior knowledge.
All rewards (and costs) count from the start, with discounting.

➡ **A natural formalism: Bayesian model-based RL.**

Example: Infinite grid task





OUR APPROACH:

A tailored Monte-Carlo Tree Search algorithm for Bayes-Adaptive MDPs.

BAYESIAN MODEL-BASED RL

- Typical MDP description: $M = \langle S, A, \mathcal{P}, \mathcal{R}, \gamma \rangle$ but $\mathcal{P}$ is a latent variable with prior $P(\mathcal{P})$.
- **Goal:** Find exploration policy that maximizes expected sum of *discounted* rewards: $\mathbb{E}_{M(\mathcal{P})}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, h_0 = s]$
- Equivalent to solving an augmented MDP in **belief space with known dynamics**, the Bayes-Adaptive MDP (BAMDP):

BAYESIAN MODEL-BASED RL

- Typical MDP description: $M = \langle S, A, \mathcal{P}, \mathcal{R}, \gamma \rangle$ but $\mathcal{P}$ is a latent variable with prior $P(\mathcal{P})$.
- **Goal:** Find exploration policy that maximizes expected sum of *discounted* rewards: $\int_{\mathcal{P}} P(\mathcal{P}) \ \mathbb{E}_{M(\mathcal{P})}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, h_0 = s]$
- Equivalent to solving an augmented MDP in **belief space with known dynamics**, the Bayes-Adaptive MDP (BAMDP):

BAYESIAN MODEL-BASED RL

- Typical MDP description: $M = \langle S, A, \mathcal{P}, \mathcal{R}, \gamma \rangle$ but $\mathcal{P}$ is a latent variable with prior $P(\mathcal{P})$.
- **Goal:** Find exploration policy that maximizes expected sum of *discounted* rewards: $\int_{\mathcal{P}} P(\mathcal{P}) \; \mathbb{E}_{M(\mathcal{P})}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, h_0 = s]$
- Equivalent to solving an augmented MDP in **belief space with known dynamics**, the Bayes-Adaptive MDP (BAMDP):

$$\mathcal{P}^+(\langle s, h \rangle, a, \langle s', h' \rangle) = \mathbb{1}[h' = has'] \int_{\mathcal{P}} \mathcal{P}(s, a, s') P(\mathcal{P}|h) \, \mathrm{d}\mathcal{P}$$

BAYESIAN MODEL-BASED RL

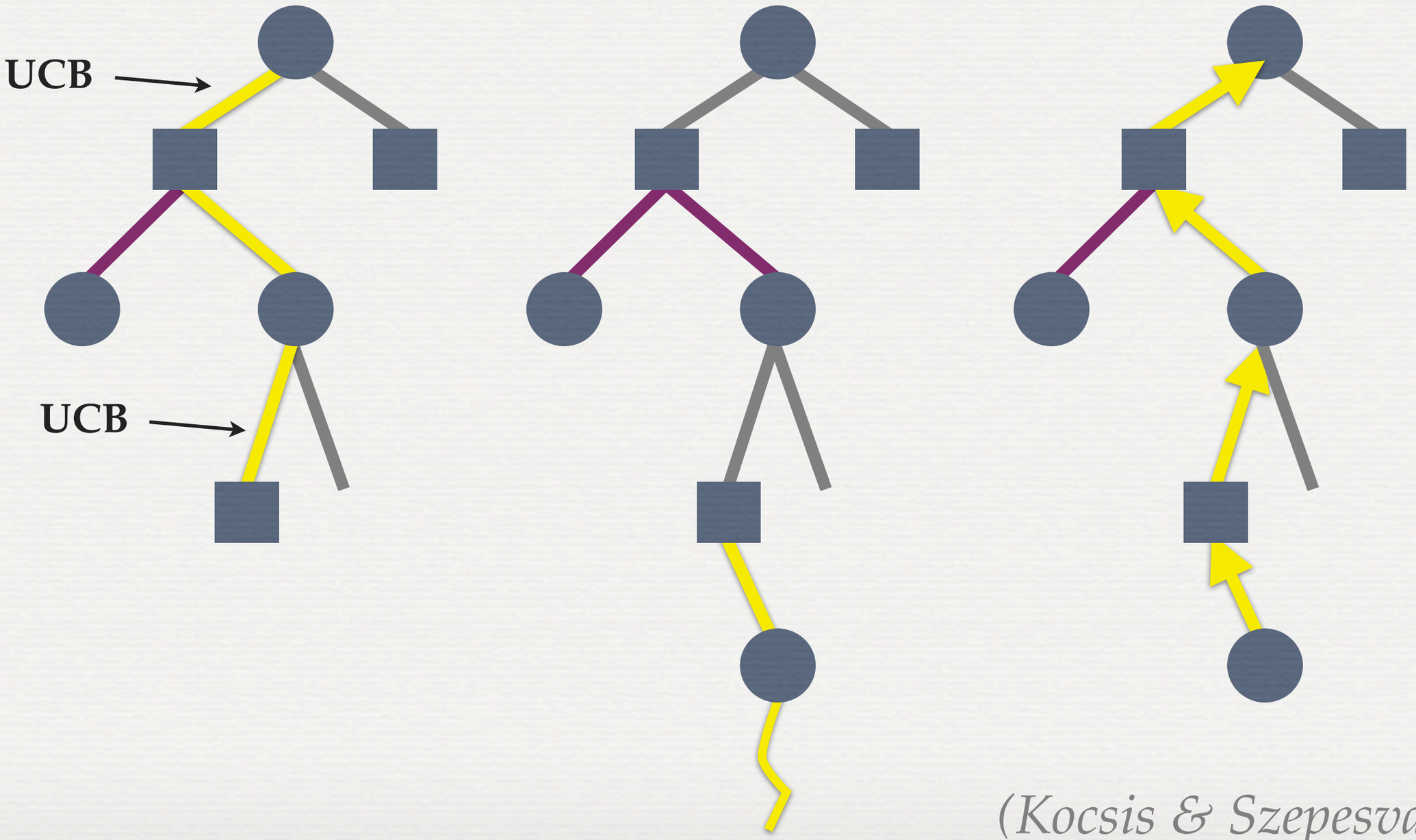
- Typical MDP description: $M = \langle S, A, \mathcal{P}, \mathcal{R}, \gamma \rangle$ but $\mathcal{P}$ is a latent variable with prior $P(\mathcal{P})$.
- **Goal:** Find exploration policy that maximizes expected sum of *discounted* rewards: $\int_{\mathcal{P}} P(\mathcal{P}) \; \mathbb{E}_{M(\mathcal{P})}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, h_0 = s]$
- Equivalent to solving an augmented MDP in **belief space with known dynamics**, the Bayes-Adaptive MDP (BAMDP):

$$\mathcal{P}^+(\langle s, h \rangle, a, \langle s', h' \rangle) = \mathbb{1}[h' = has'] \int_{\mathcal{P}} \mathcal{P}(s, a, s') P(\mathcal{P}|h) \, \mathrm{d}\mathcal{P}$$

- **Major obstacle:** Computationally intractable to solve exactly.

QUICK REMINDER: MCTS/UCT

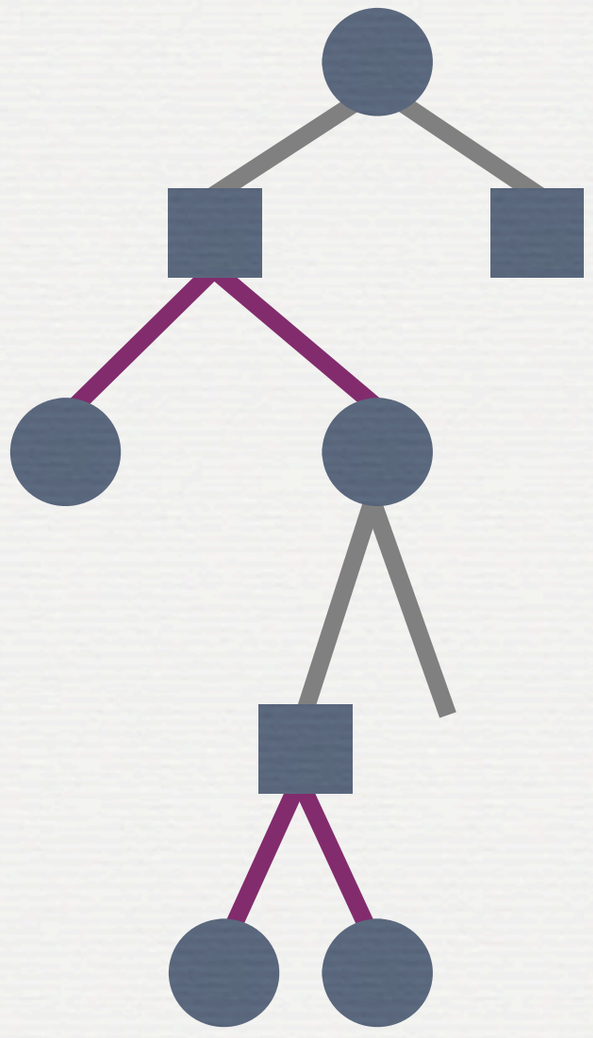
Selection ⟶ Expansion+Rollout ⟶ Update



(Kocsis & Szepesvári 2006)

OUR APPROACH: SOLUTION 1

- **BA-UCT**: Leverage modern sample-based MDP solver: MCTS/UCT applied to BAMDP.



✓ Can handle large state spaces.

OUR APPROACH: SOLUTION 1

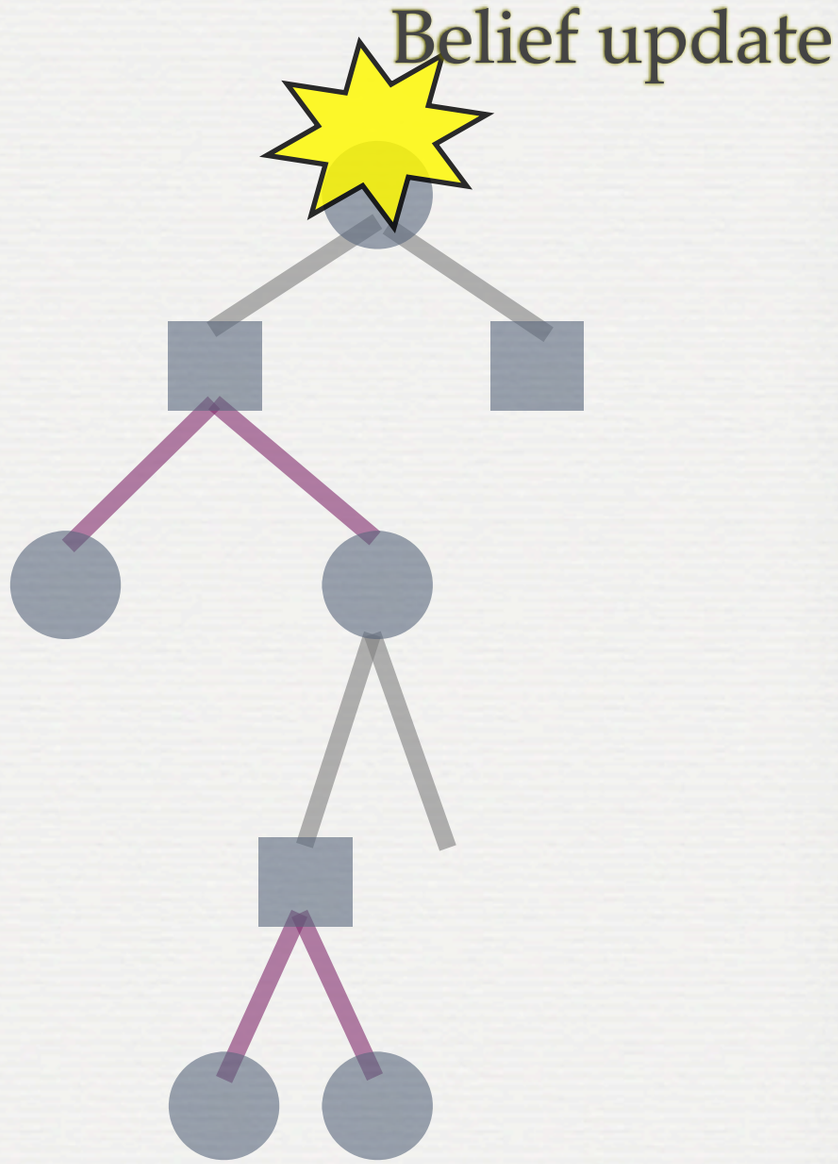
- **BA-UCT**: Leverage modern sample-based MDP solver: MCTS/UCT applied to BAMDP.

✓ Can handle large state spaces.

- **Issue:** Expensive belief updates at every tree node.

OUR APPROACH: SOLUTION 2

- **BA-UCT + Root Sampling (RS)**: Take advantage of tree to filter the posterior distribution, as in POMCP *(Silver & Veness 2010)*.
  - Sample $\mathcal{P}$ at root for each sim.



✓ One belief update per step.

✓ Still Bayes-optimal in the limit.

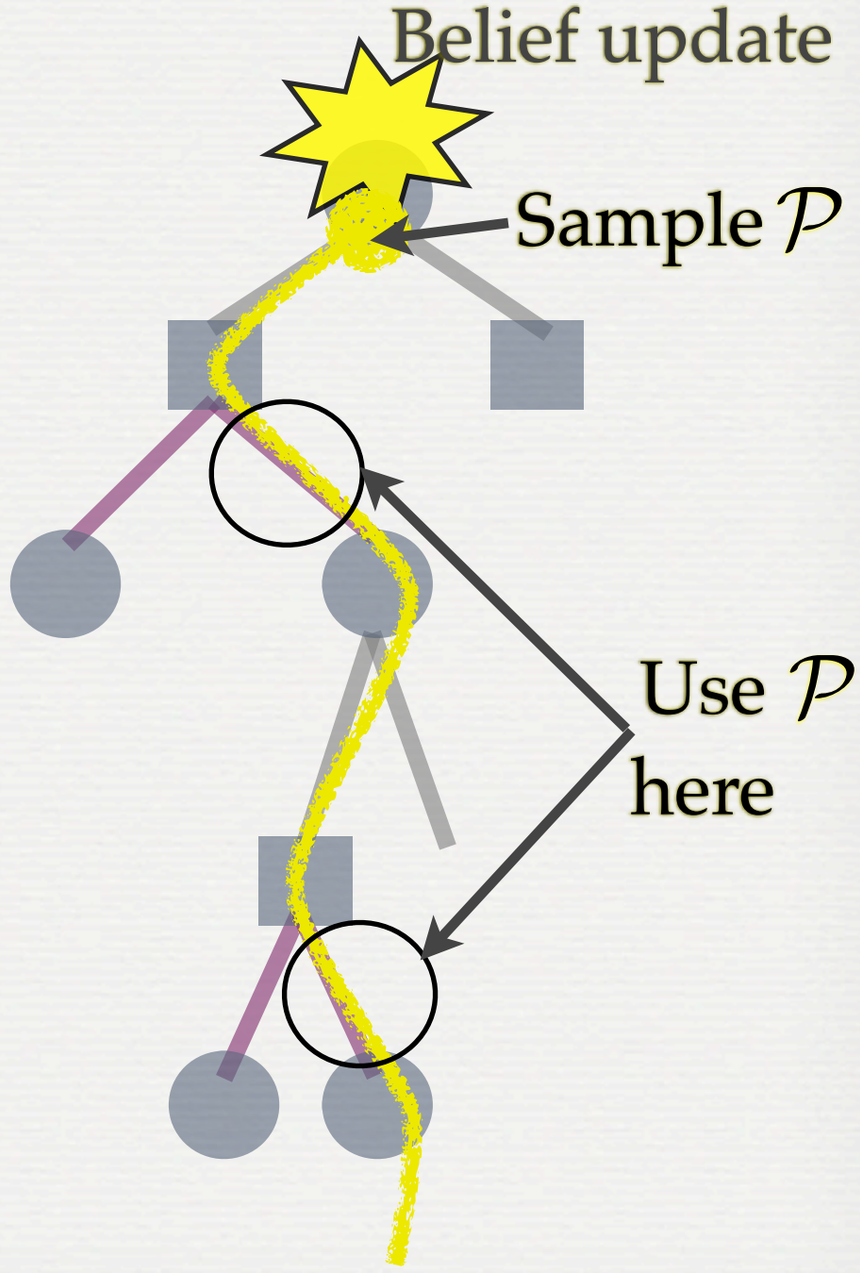
OUR APPROACH: SOLUTION 2

- **BA-UCT + Root Sampling (RS)**: Take advantage of tree to filter the posterior distribution, as in POMCP *(Silver & Veness 2010)*.
  - Sample $\mathcal{P}$ at root for each sim.

✓ One belief update per step.

✓ Still Bayes-optimal in the limit.

- **Issue:** Wasteful sampling, each simulation only requires a small subset of all parameters.

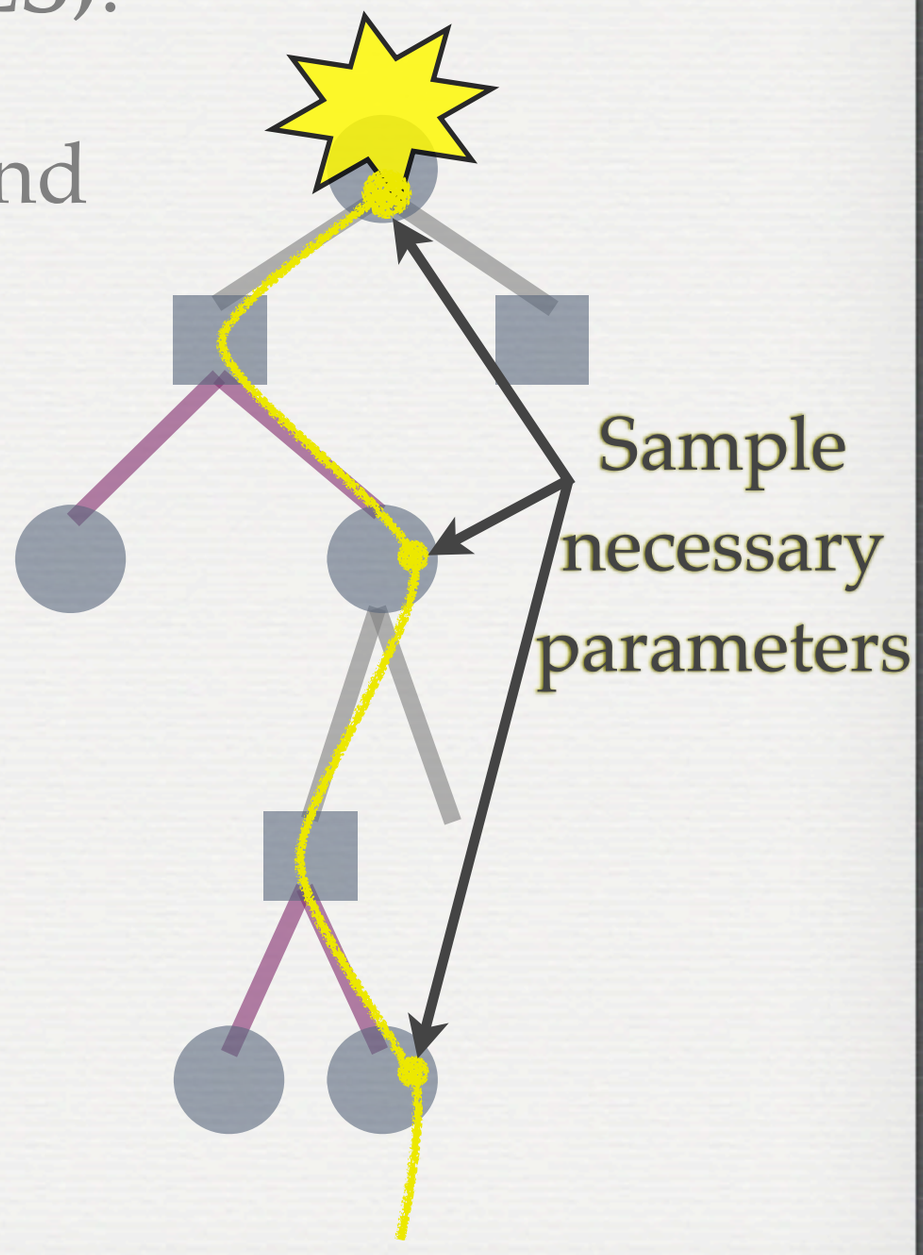


OUR APPROACH: SOLUTION 3

- **BA-UCT + RS + Lazy Sampling (LS)**:

  Sample MDP parameters on demand based on current simulation.

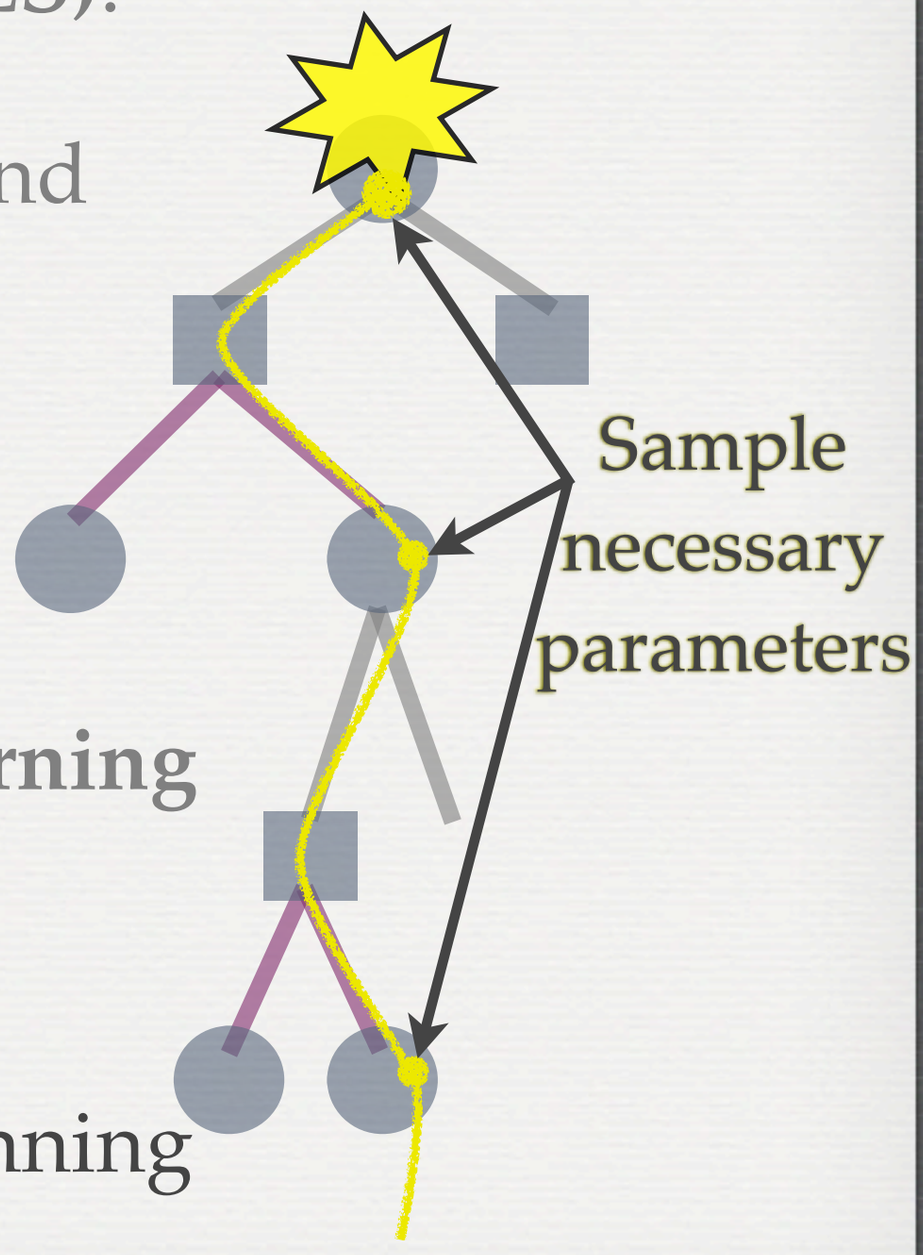
  ✓ Minimize sampling overhead, put more effort on search.



OUR APPROACH: SOLUTION 3

- **BA-UCT + RS + Lazy Sampling (LS)**:

  Sample MDP parameters on demand based on current simulation.



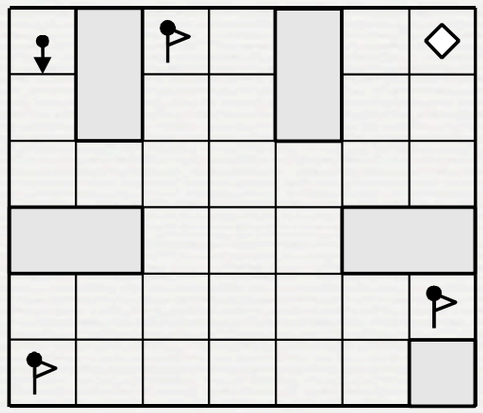
  ✓ Minimize sampling overhead, put more effort on search.

- **BA-UCT + RS + LS + Rollout Learning**

  Learn better rollout policy online.

➡ **B**ayes-**A**daptive **M**onte-**C**arlo **P**lanning (**BAMCP**)

DOES IT MATTER?



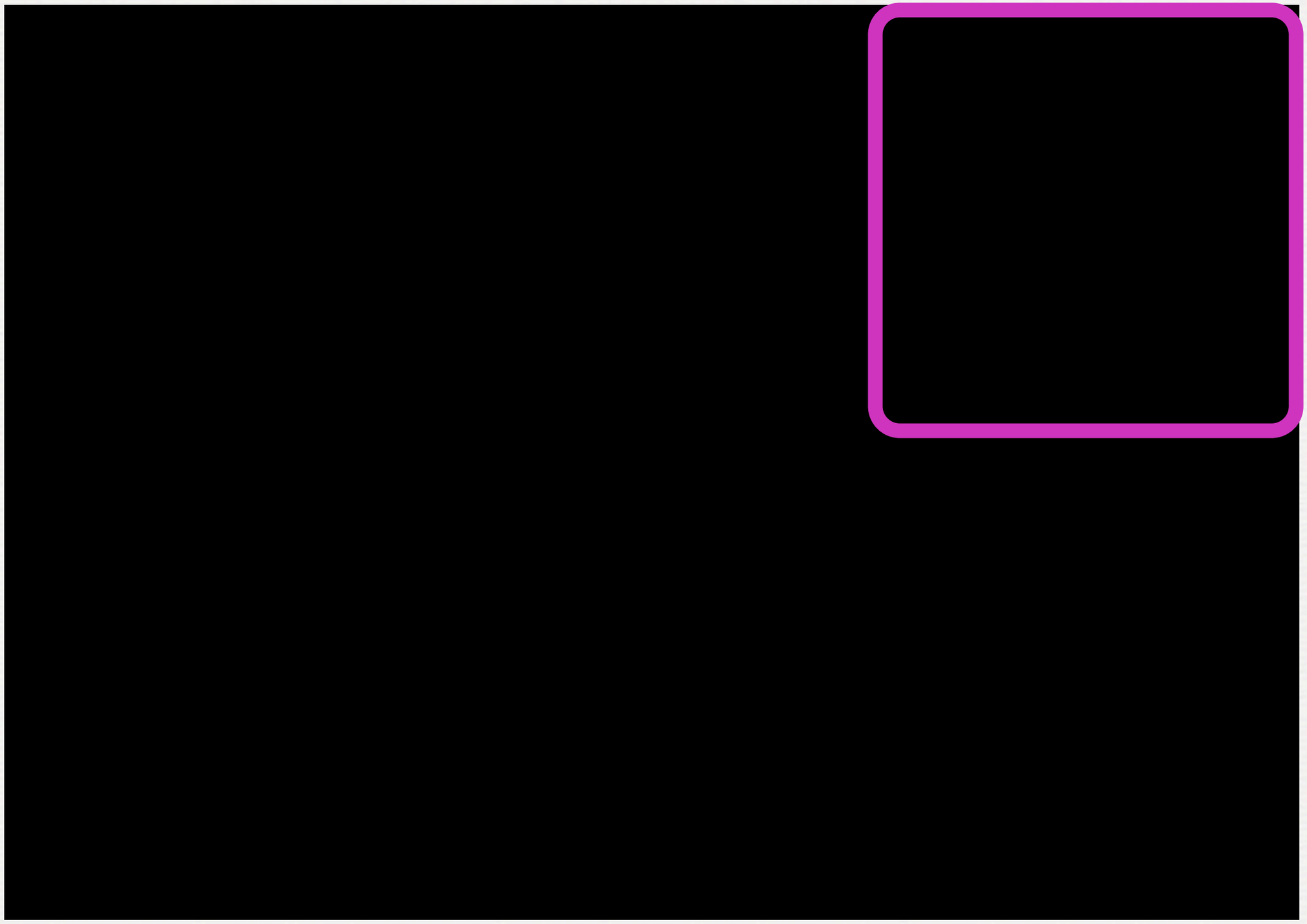
Compare performance on Dearden's Maze.

- 264 states,
- Sparse Dirichlet-Multinomial prior.

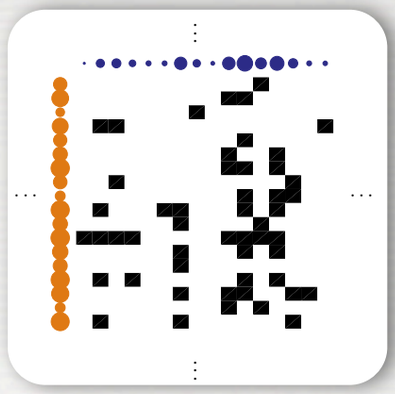
Yes! Even on simple tasks,
each improvement independently improves performance.

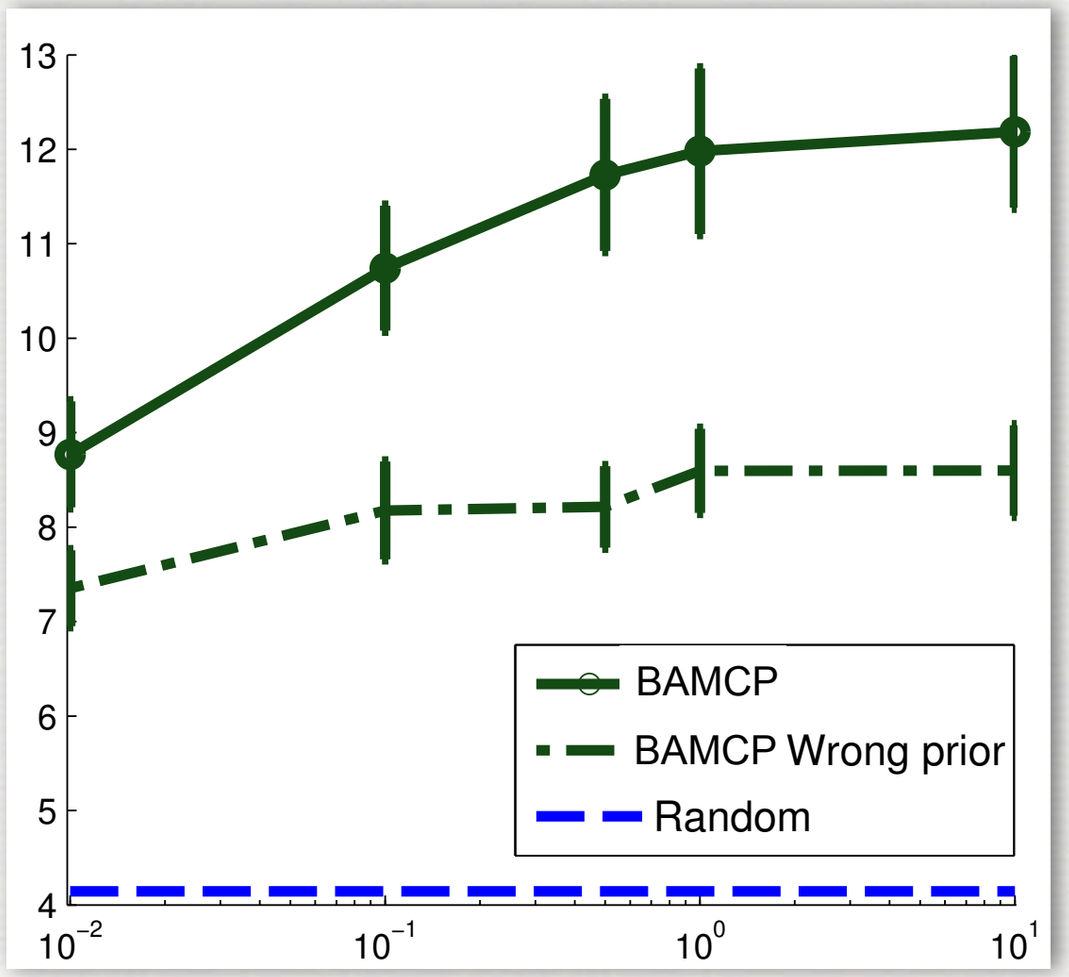
BAMCP on the infinite 2D grid task

Lazy posterior samples



BAMCP on the infinite 2D grid task





Prior affects behavior and performance

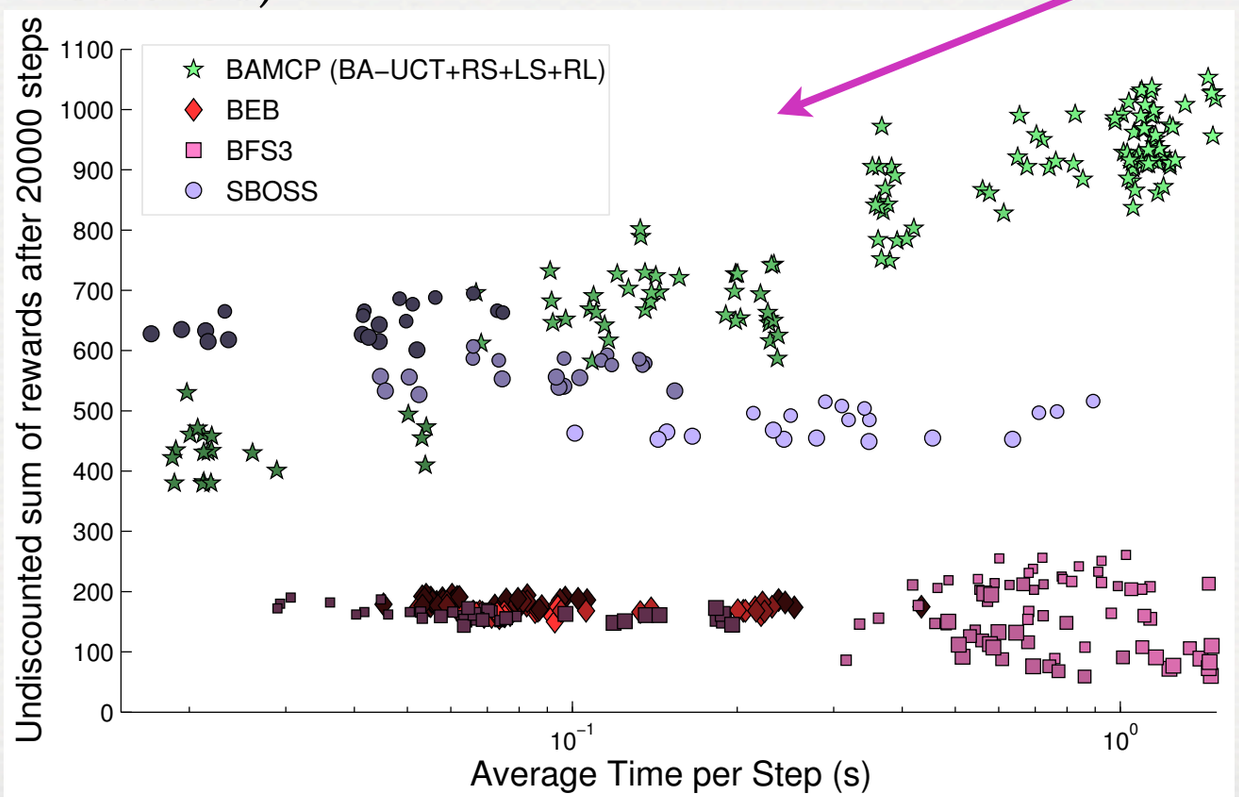
Sample 50 grids, run BAMCP on each, average results.

RESULTS: STANDARD DOMAINS

Our approach

	Double-loop	Grid5	Grid10	Dearden Maze
BAMCP	**387.6 ± 1.5**	**72.9 ± 3**	**32.7 ± 3**	**965.2 ± 73**
BFS3	382.2 ± 1.5	66 ± 5	10.4 ± 2	240.9 ± 46
SBOSS	371.5 ± 3	59.3 ± 4	21.8 ± 2	671.3 ± 126
BEB	386 ± 0	67.5 ± 3	10 ± 1	184.6 ± 35
Bayesian DP*	377 ± 1	-	-	-
Bayes VPI+MIX*	326 ± 31	-	-	817.6 ± 29
IEQL+*	264 ± 1	-	-	269.4 ± 1
QL Boltzmann*	186 ± 1	-	-	195.2 ± 20

(Sum of rewards)



CONCLUSION

- Introduce tractable sample-based algorithm for Bayesian RL,
- State-of-the-art performance results on standard domains,
- Scales to large tasks,
- Can exploit structured priors.

Acknowledgements: Gatsby Charitable Foundation, Natural Sciences and Engineering Research Council of Canada.