On Prediction and Planning in Partially Observable Markov Decision Processes with Large Observation Sets

Pablo Samuel Castro
pcastr@cs.mcgill.ca
McGill University
Joint work with: Doina Precup and Prakash Pananganden

Motivation

- Interested in sequential decision making under uncertainty
- Agent must infer its "state" based on observations of environment
- A larger observation space gives more information, but increases complexity of problem
- Hardware is cheap and small => many sensors/observations!

Our contribution

- Allow subsets of observation space to be specified for planning/learning.
- Provide theoretical foundations when planning/learning using this idea.
- Will address questions such as:
  - How is agent's behaviour affected by using only a subset of all observations?
  - How are agent's predictions affected?

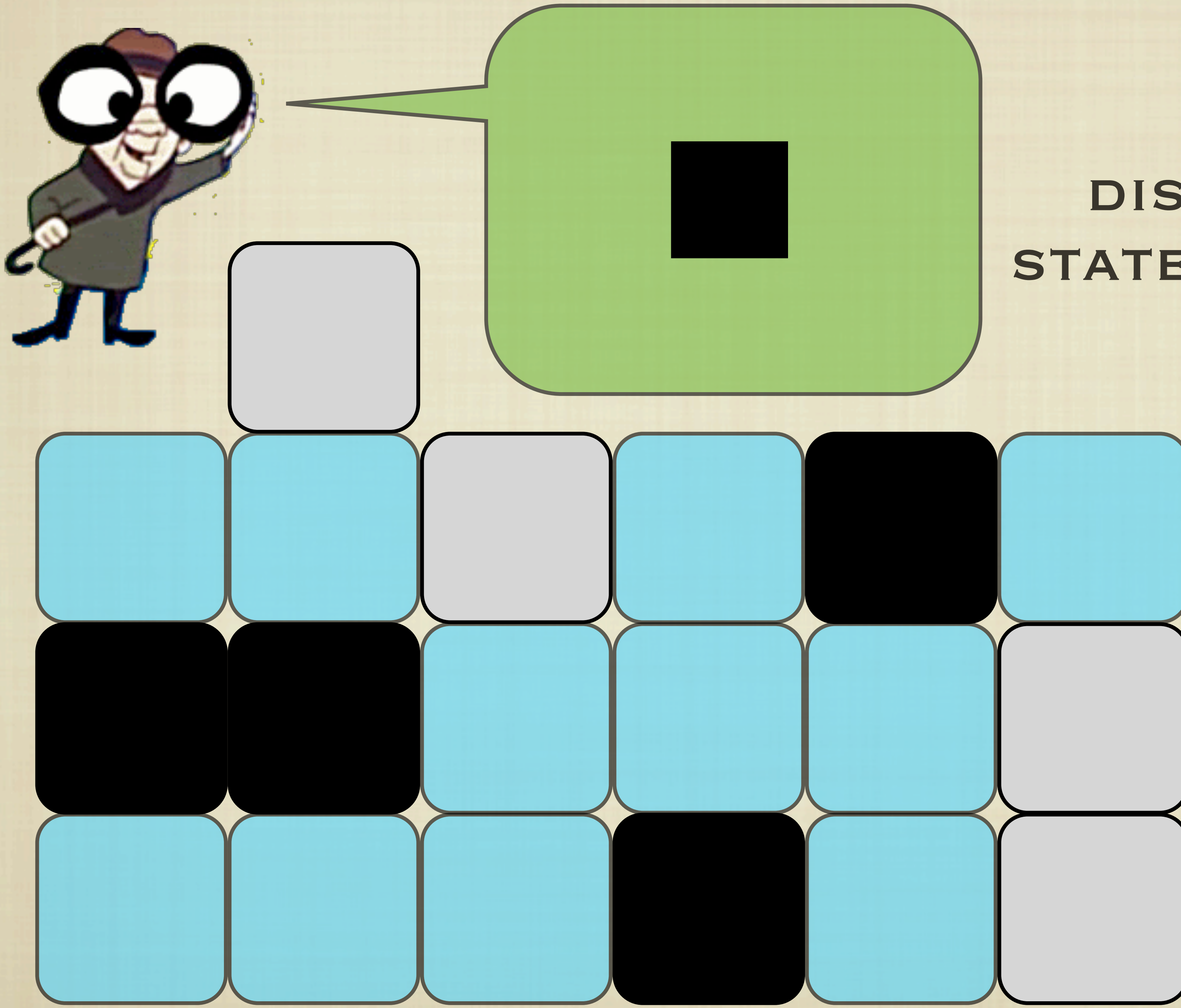
Outline

- POMDP review
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

Outline

- **POMDP review**
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

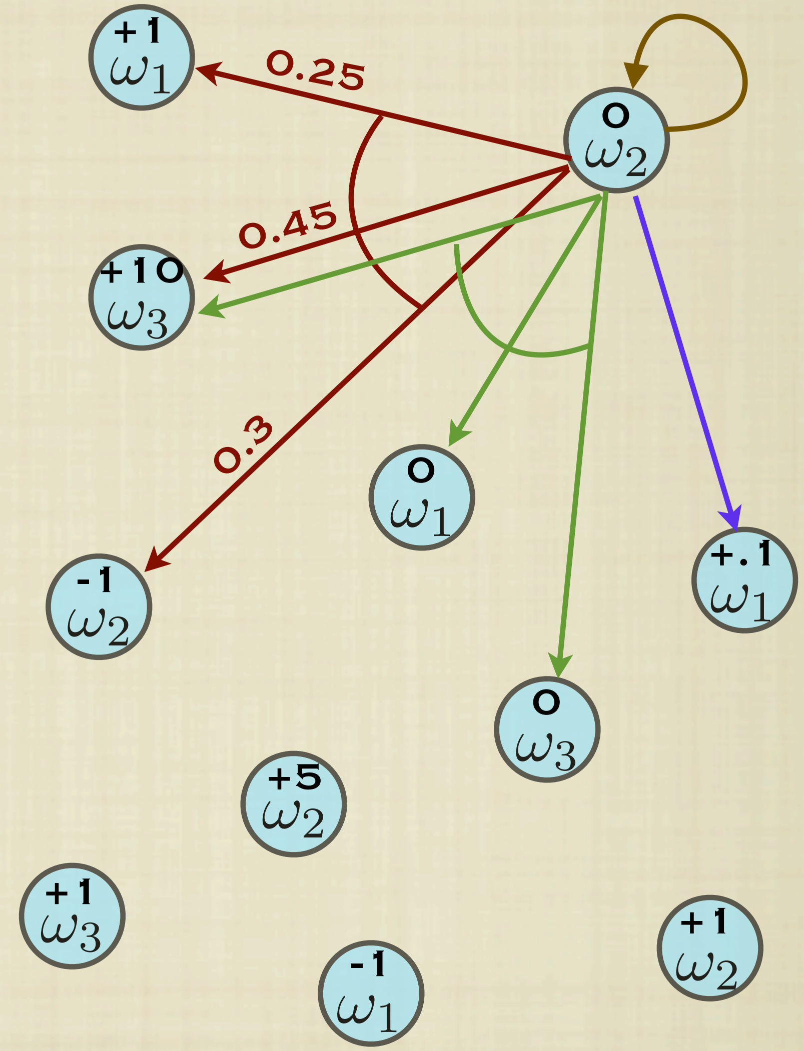
Partially observable MDPs (POMDPs)



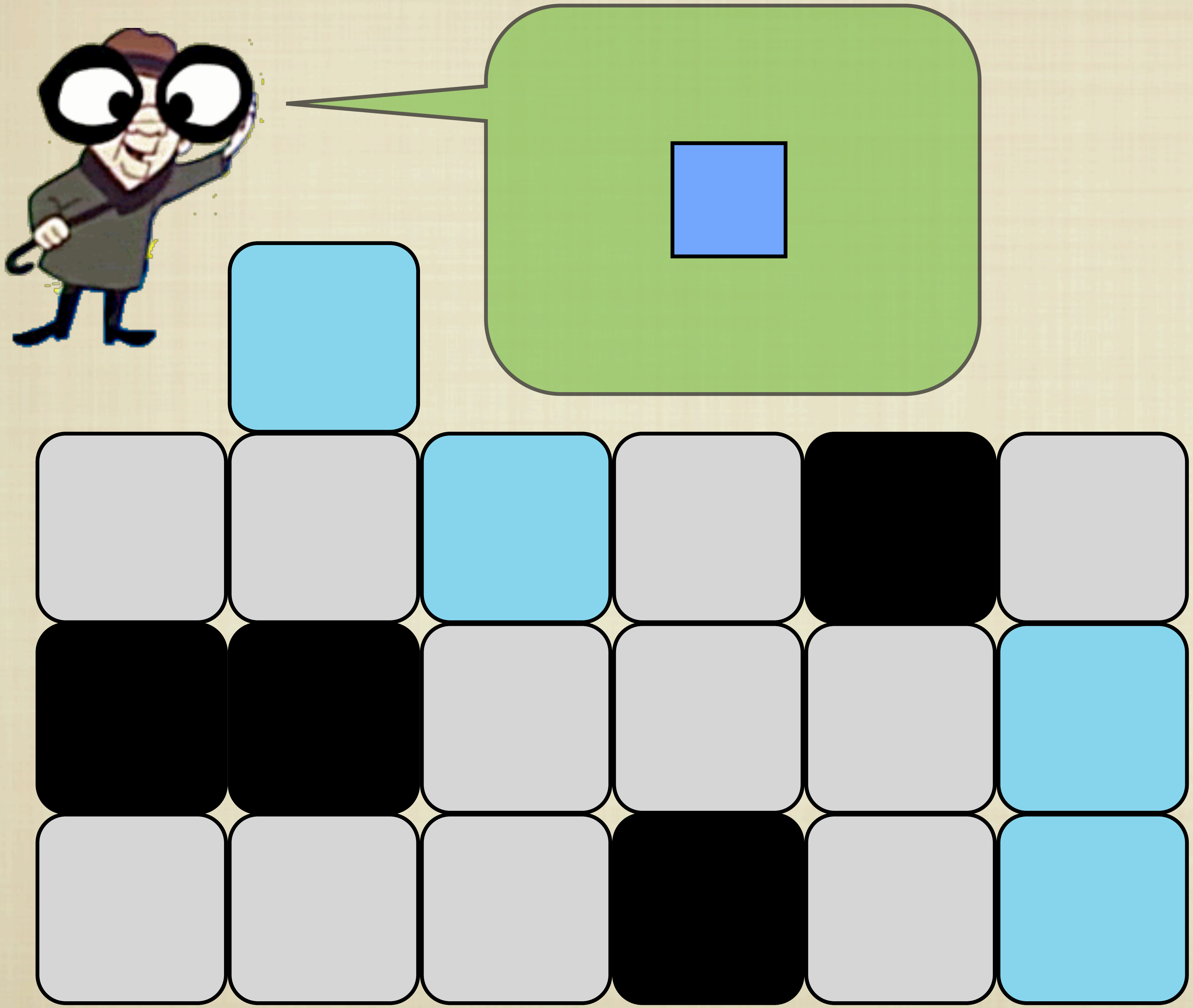
Maintain a distribution over states based on clues

Standard POMDPs

- 6-tuple $\langle S, A, P, R, \Omega, O \rangle$ consisting of
  - Set of states S, $(s, s', t, \ldots)$
  - Set of actions A, $(a, b, \ldots)$
  - Probabilistic transition function $P(s,a)(s')$
  - Bounded reward function $R(s,a)$
  - Set of observations Ω
  - Observation function $O(a,s)(\omega)$
  - Discount factor $0 \leq \gamma < 1$



Belief states



Move forward

Belief states

- A belief state μ is a distribution over S.
- Given μ, action a and observation ω, there is a unique next belief state $\tau(\mu, a, \omega)$.
- Can also compute probability of next observations.

Outline

- POMDP review
- **New POMDP formulation**
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

POMDPs

- 5-tuple $\langle S, A, P, \Omega, O \rangle$ consisting of
  - Set of states S, $(s, s', t, \ldots)$
  - Set of actions A, $(a, b, \ldots)$

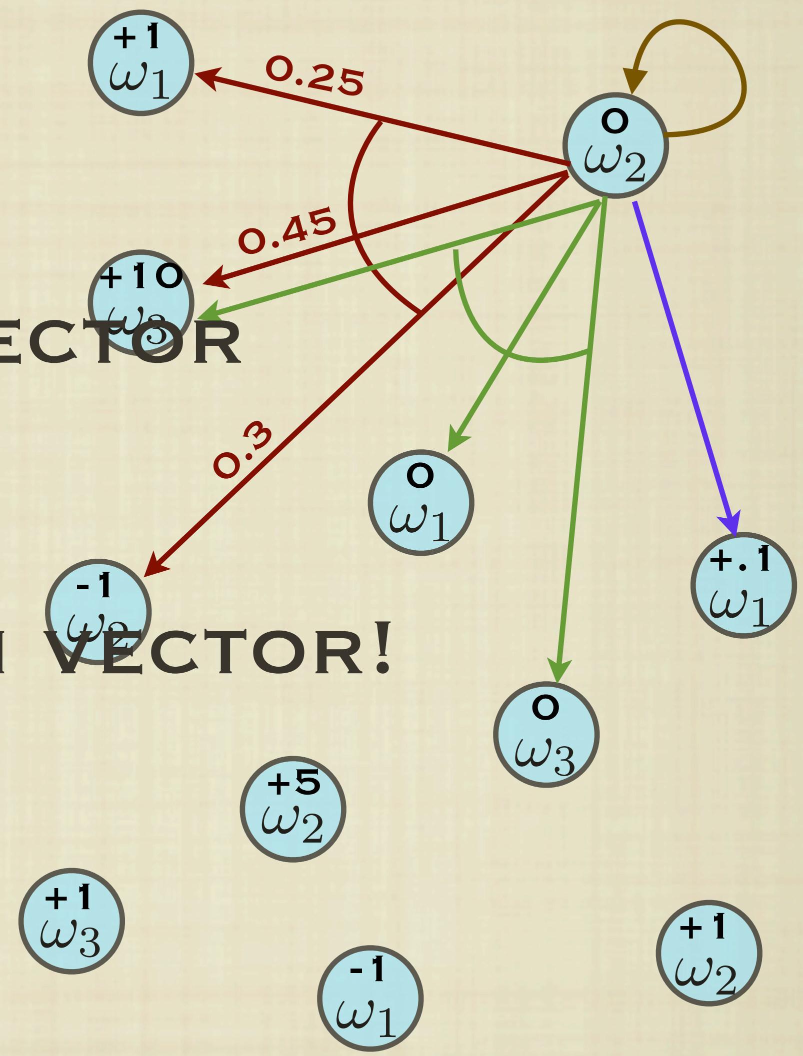
K-DIMENSIONAL OBSERVATION VECTOR

  - Probabilistic transition function $P(s, a)(s')$

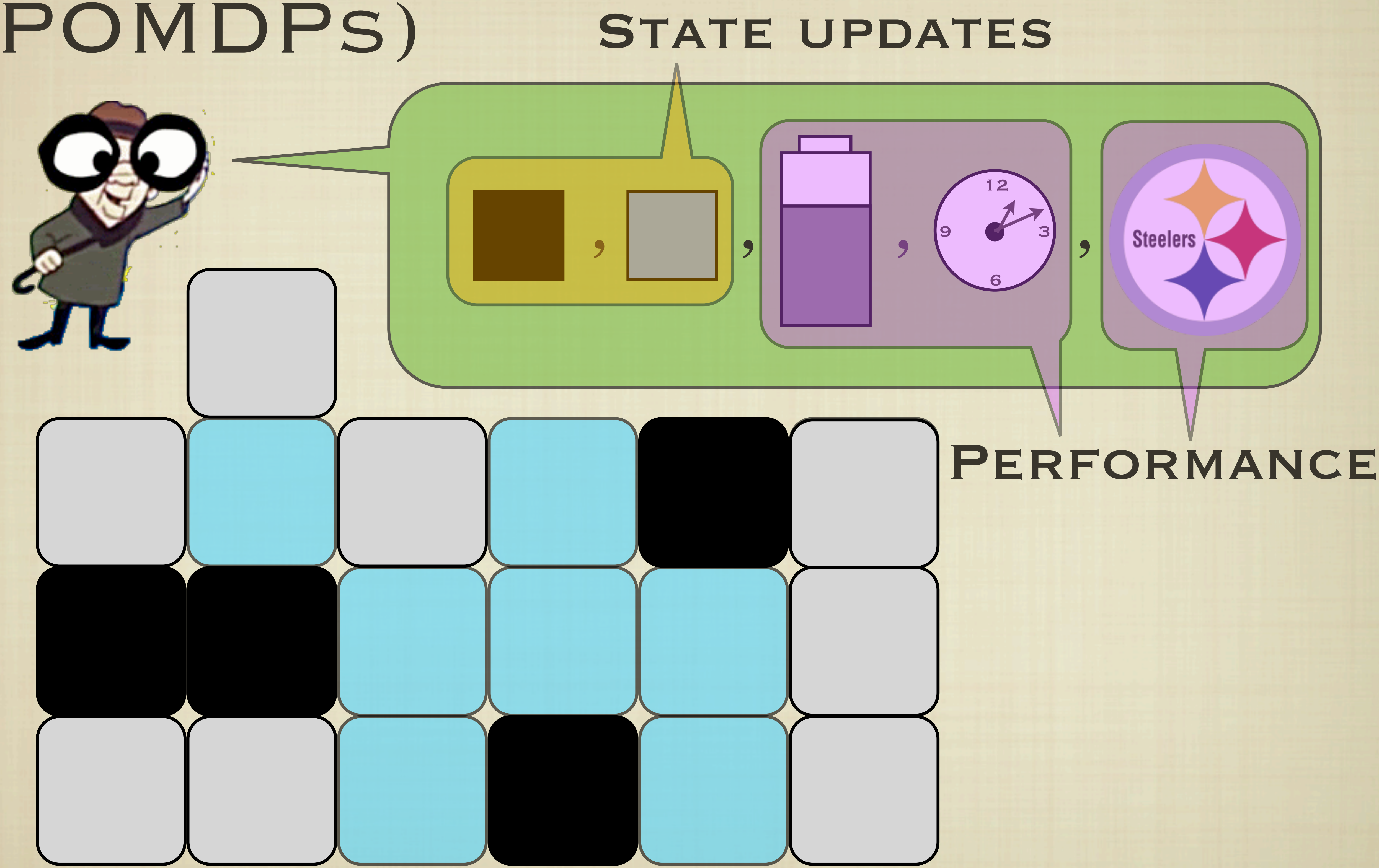
$$= \Omega_1 \times \Omega_2 \times \cdots \times \Omega_k$$

Rewards part of observation vector!

  - Set of observations Ω
  - Observation function $O(a, s)(\omega)$ $O_{s,\omega}$
  - Discount factor $0 \leq \gamma < 1$



Partially observable MDPs (POMDPs)



Specifying data and interest

- Let $\mathcal{D} \subseteq \{1, 2, \ldots, k\}$ be indices of observation coordinates used for belief updates
- Let $\mathcal{I} \subseteq \{1, 2, \ldots, k\}$ be indices of observation coordinates that are observables of interest for planning/ prediction.
- Let $\Omega_{\mathcal{D}}$ be set of observations containing only observations from $\mathcal{D}$. similarly for $\Omega_{\mathcal{I}}$.

New POMDP dynamics

- We project observation functions with binary projection matrices $\Phi_{\mathcal{D}}$: $O_{\mathcal{D}} = O\Phi_{\mathcal{D}}$
- Unique next beliefs specific to choice of $\mathcal{D}$: $$\tau_{\mathcal{D}}(\mu, a, \omega) = \frac{\mu P^a O^{\omega}_{\mathcal{D}}}{\boxed{\mu P^a O^{\omega}_{\mathcal{D}} \mathbf{e}^T}}$$
- Can define a transition fn. between belief states $T_{\mathcal{D}}(\mu, a)(\mu')$.

Measuring performance

- Elements from $\Omega_{\mathcal{I}}$ may be of many different types.
- Need a way to quantify an agent's performance.
- We assume a function $f : \Omega_{\mathcal{I}} \to \mathbb{R}$ that maps observations of interest to a real number.

Policies and value functions

- Closed-loop policies: Map belief states to actions ($\pi \in \Pi$)
- Value of a belief state μ under π: $\mathbb{E}^{\pi}\left[\sum_{i=0}^{H} \gamma^i r_i | \mu\right]$

$$V^{\pi}_{\mathcal{D},\mathcal{I}} = \sum_{\omega_{\mathcal{I}} \in \Omega_{\mathcal{I}}} Pr(\omega_{\mathcal{I}} | \mu, \pi(\mu)) f(\omega_{\mathcal{I}}) + \gamma \sum_{\mu' \in \mathcal{B}} T_{\mathcal{D}}(\mu, \pi(\mu))(\mu') V^{\pi}_{\mathcal{D},\mathcal{I}}(\mu')$$

- Optimal value function:

$$V^{*}_{\mathcal{D},\mathcal{I}}(\mu) = \max_{a \in A} \left\{ \sum_{\omega_{\mathcal{I}} \in \Omega_{\mathcal{I}}} Pr(\omega_{\mathcal{I}} | \mu, a) f(\omega_{\mathcal{I}}) + \gamma \sum_{\mu' \in \mathcal{B}} T_{\mathcal{D}}(\mu, a)(\mu') V^{*}_{\mathcal{D},\mathcal{I}}(\mu') \right\}$$

Is new POMDP definition suitable?

- Are fully observable MDPs still expressible?
- Does the definition properly follow intuition? (e.g. do larger observation subsets yield improved performance)?

MDPs

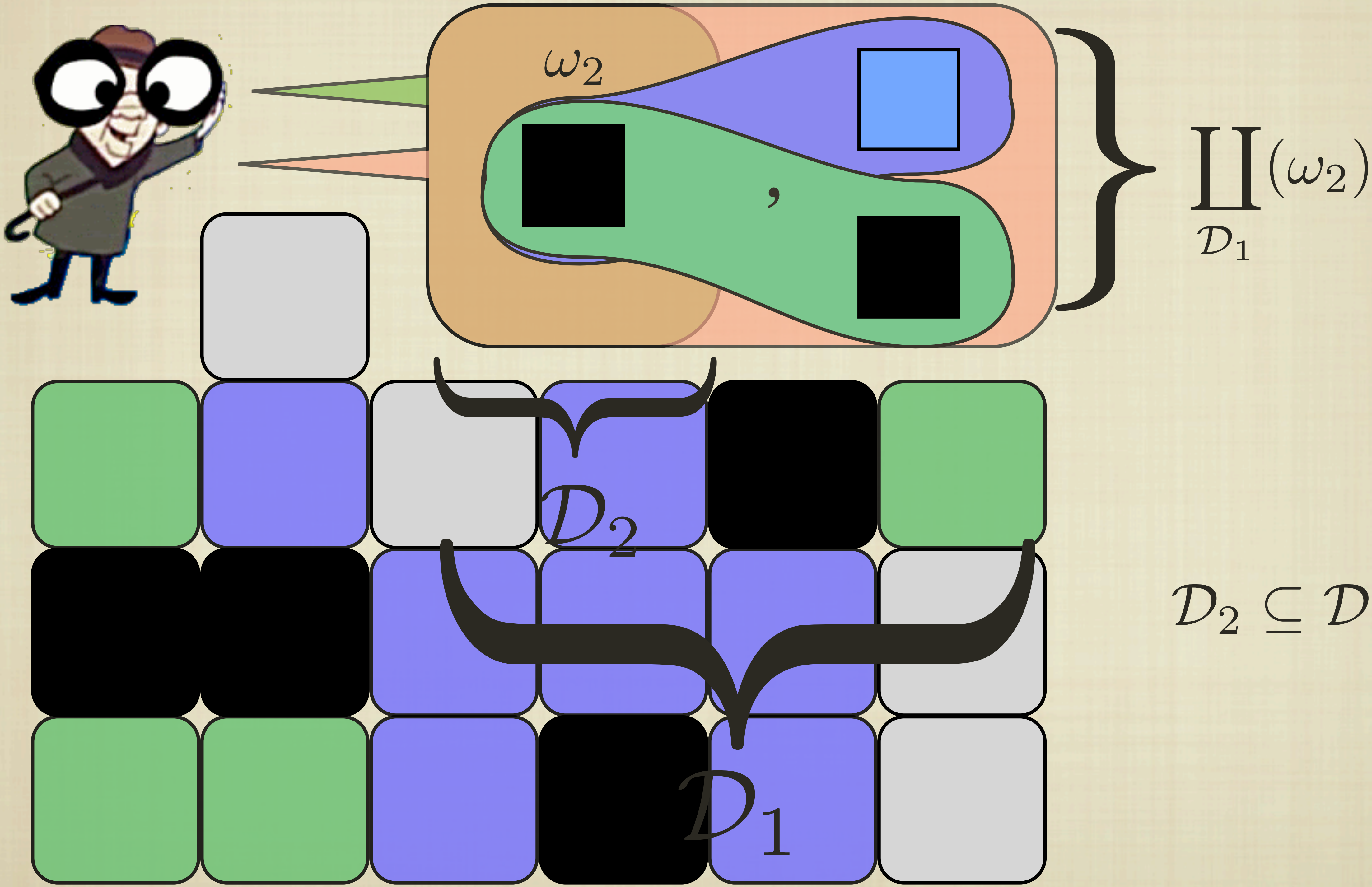
- Assume the observations are just $S \times \mathbb{R}$
- $\mathcal{D}$ points to S and $\mathcal{I}$ points to $\mathbb{R}$.

Optimal value functions

- **Proposition:** Given indexing sets $\mathcal{D}_2 \subseteq \mathcal{D}_1$ and $\mathcal{I}$, then

$$V^*_{\mathcal{D}_2,\mathcal{I}} \leq V^*_{\mathcal{D}_1,\mathcal{I}}$$

Convexity of beliefs

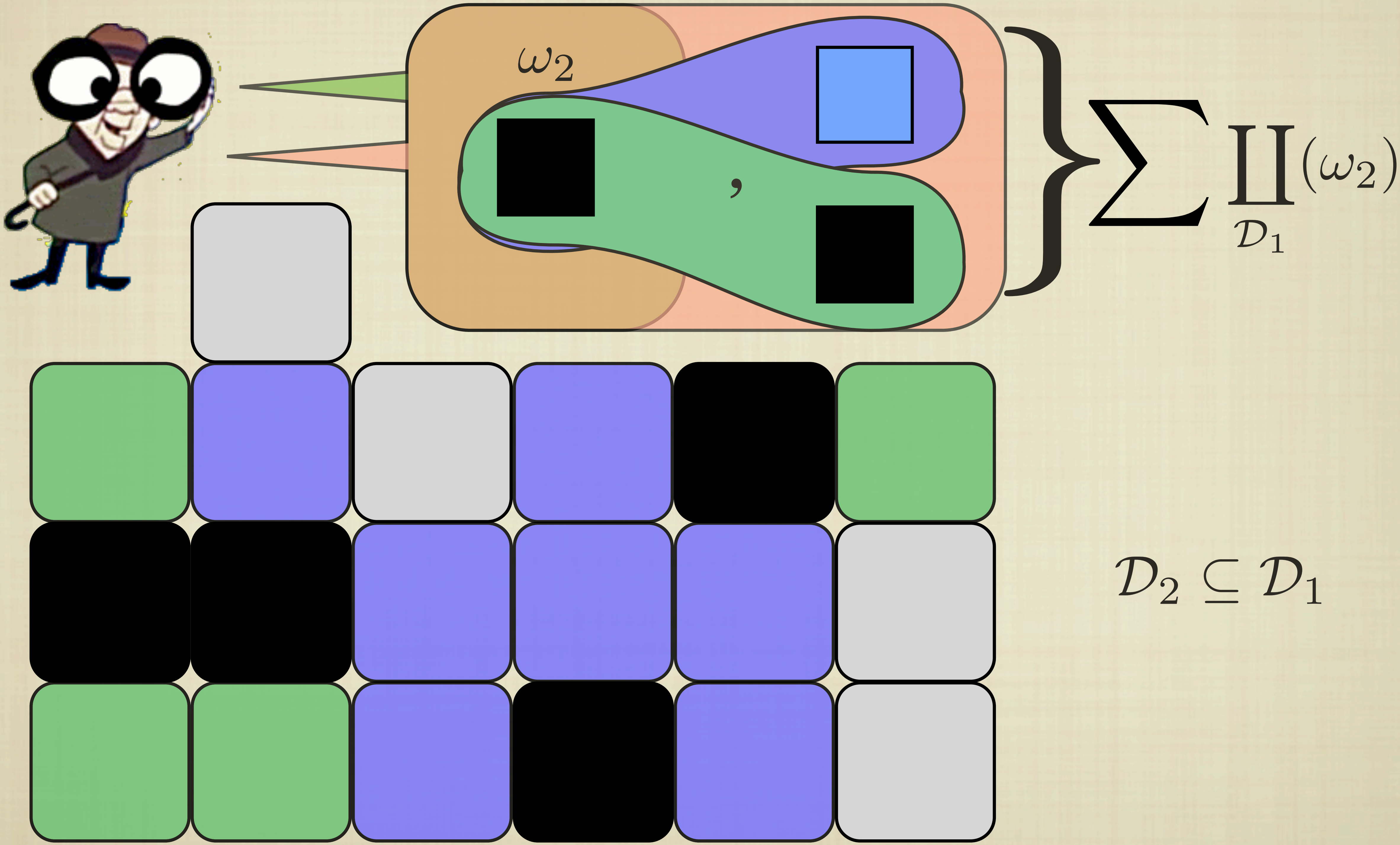


$\mathcal{D}_2 \subseteq \mathcal{D}_1$

Convexity of beliefs

Theorem: Given a belief μ, an action a, $\mathcal{D}_2 \subseteq \mathcal{D}_1$, and observation $\omega_2 \in \Omega_{\mathcal{D}_2}$, the unique next belief $\tau_{\mathcal{D}_2}(\mu, a, \omega_2)$ can be expressed as a convex combination of the belief states $\{\tau_{\mathcal{D}_1}(\mu, a, \omega_1)\}_{\omega_1 \in \coprod_{\mathcal{D}_1}(\omega_2)}$.

Convexity of beliefs



$\mathcal{D}_2 \subseteq \mathcal{D}_1$

Outline

- POMDP review
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

Equivalence relations

- Partition belief space into equivalence classes
- Capture some form of behavioural equivalence
- Two beliefs in same equivalence are behaviourally indistinguishable

Outline

- POMDP review
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

Value function equivalences

- For all belief states μ, ν let $\Pi_{\mu,\nu}$ be the set of all policies $\pi \in \Pi$ where $\pi(\mu) = \pi(\nu)$
- Belief states μ, ν are $(\mathcal{D}, \mathcal{I})$-**closed value equivalent** if for all $\pi \in \Pi_{\mu,\nu}$,

$$V^{\pi}_{\mathcal{D},\mathcal{I}}(\mu) = V^{\pi}_{\mathcal{D},\mathcal{I}}(\nu)$$

- Belief states μ, ν are $(\mathcal{D}, \mathcal{I})$-**optimal value equivalent** if

$$V^{*}_{\mathcal{D},\mathcal{I}}(\mu) = V^{*}_{\mathcal{D},\mathcal{I}}(\nu)$$

Closed and optimal value equivalences

Theorem: If two states are closed value equivalent, then they are necessarily optimal value equivalent.

$$
\begin{array}{c}
V^{\pi} \text{ equivalence} \\
\downarrow \\
V^{*} \text{ equivalence}
\end{array}
$$

Closed and optimal value equivalences

V^π equivalence

$\downarrow$

V^* equivalence

Lemma: If s_0, t_0 are V^π equivalent and $V^*(s_0) > V^*(t_0)$, then prob. of reaching t_0 from s_0 under π^* is strictly positive.

Let Π_{CV} be set of all policies π constructed from some optimal policy π^* as follows:

$$\pi(s') = \pi^*(s_0) \text{ if } s' = t_0$$

$$\pi(s') = \pi^*(s') \text{ otherwise}$$

Closed and optimal value equivalences

$$V^\pi \text{ equivalence} \\ \downarrow \\ V^* \text{ equivalence}$$

- B is set of bounded functions $V : S \times \Pi_{CV} \to [0,1]$
- $\mathcal{R} \in B$, $\mathcal{R}(s,\pi) = R(s,\pi(s))$
- $\Upsilon : B \to B$, $\Upsilon(V)(s,\pi) = \gamma \sum_{s' \neq t_0} P(s,\pi(s))(s') V(s',\pi) + P(s,\pi(s))(t_0) V(t_0,\pi)$
- $\tau(e) = \mathcal{R} + \Upsilon(e)$ has least fixed pt $e^*(s,\pi) = V^\pi(s)$

Theorem (based on (Kozen, 2007))

Define $\varphi \subseteq B$ as $V \in \varphi \Rightarrow \forall \pi \in \Pi_{CV} . V(s,\pi) \geq V^*(s)$, then if $\varphi \neq \emptyset$ and $e \in \varphi \Rightarrow \tau(e) \in \varphi$, then $e^* \in \varphi$.

Closed and optimal value equivalences

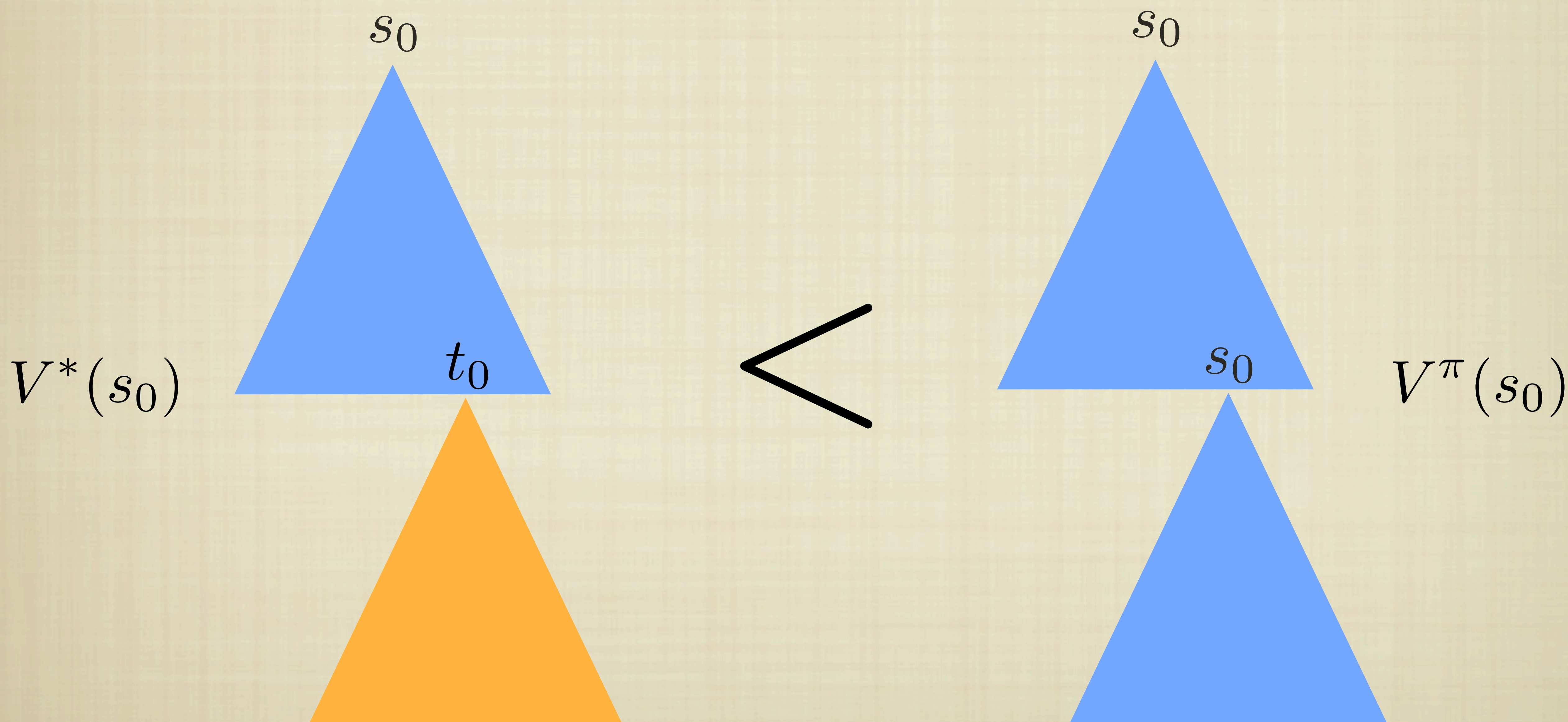
V^π equivalence

$\downarrow$

V^* equivalence

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \leq V^*(t_0)$.

Proof: Assume $V^*(s_0) > V^*(t_0)$



Closed and optimal value equivalences

$$V^\pi \text{ equivalence} \downarrow V^* \text{ equivalence}$$

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \leq V^*(t_0)$.

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \geq V^*(t_0)$.

Closed and optimal value equivalences

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \leq V^*(t_0)$.

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \geq V^*(t_0)$.

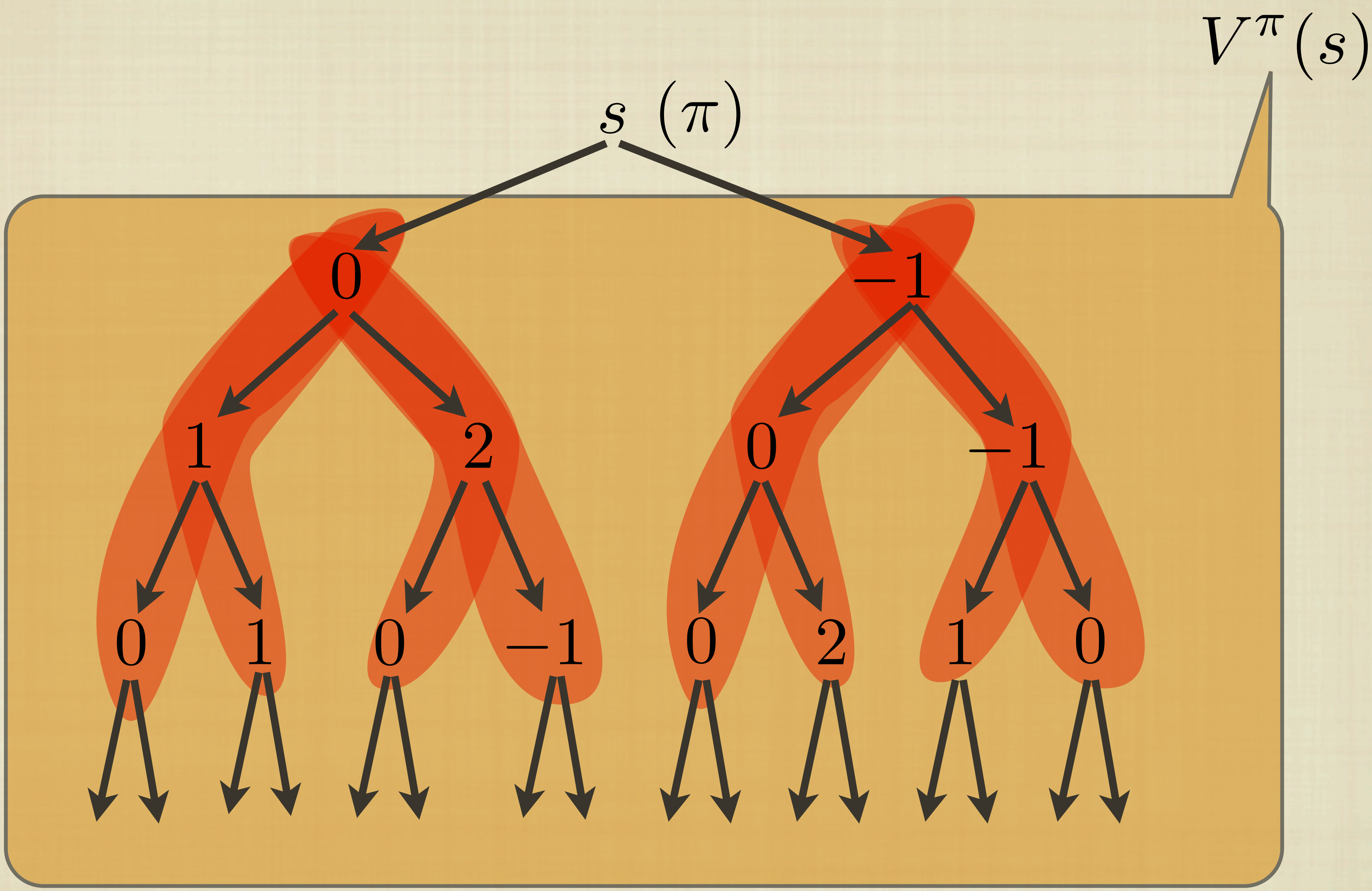
Theorem: If s_0 and t_0 are V^π equivalent then $V^*(s_0) = V^*(t_0)$.

$$\begin{array}{c} V^\pi \text{ equivalence} \\ \downarrow \\ V^* \text{ equivalence} \end{array}$$

Outline

- POMDP review
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

Trajectory equivalences



Trajectory equivalence

- Two belief states μ, ν are $\mathcal{I}$-closed trajectory equivalent if for all $\pi \in \Pi_{\mu,\nu}$ and all finite observation trajectories, $\alpha = \langle \omega_1, \omega_2, \ldots, \omega_n \rangle \in \Omega_{\mathcal{I}}^*$

$$Pr(\alpha|\mu, \pi) = Pr(\alpha|\nu, \pi)$$

Trajectory equivalence

- Open-loop policies $\theta \in \Theta$ map time steps to actions
- Two belief states μ, ν are $\mathcal{I}$-open trajectory equivalent if for all $\theta \in \Theta$ and all finite observation trajectories $\alpha = \langle \omega_1, \omega_2, \ldots, \omega_n \rangle \in \Omega_{\mathcal{I}}^*$,

$$Pr(\alpha|\mu, \theta) = Pr(\alpha|\nu, \theta)$$

- A trajectory α and open loop policy θ constitute a PSR test (Littman et al., 2002)!

$$\langle a_1, \omega_1, a_1, \omega_2, \ldots, a_n, \omega_n \rangle$$

Hierarchy

- If $\mathcal{D} \subseteq \mathcal{I}$, then the following hierarchy is obtained

$$\begin{tikzcd}
\mathcal{I}-open\ trajectory \arrow[ddr] & & \mathcal{I}-closed\ trajectory \arrow[d] \\
& & (\mathcal{D},\mathcal{I})-closed\ value \arrow[dl] \\
& (\mathcal{D},\mathcal{I})-optimal\ value &
\end{tikzcd}$$

Hierarchy

- If $\mathcal{D} \not\subseteq \mathcal{I}$, then the following hierarchy is obtained

$$
\begin{tikzcd}
\mathcal{I} - open\ trajectory \arrow[ddr] & & \mathcal{I} - closed\ trajectory \arrow[d] \\
& & (\mathcal{D},\mathcal{I}) - closed\ value \arrow[dl] \\
& (\mathcal{D},\mathcal{I}) - optimal\ value &
\end{tikzcd}
$$

Hierarchy

$$\mathcal{I} - open\ trajectory \not\longrightarrow (\mathcal{D}, \mathcal{I}) - optimal\ value$$

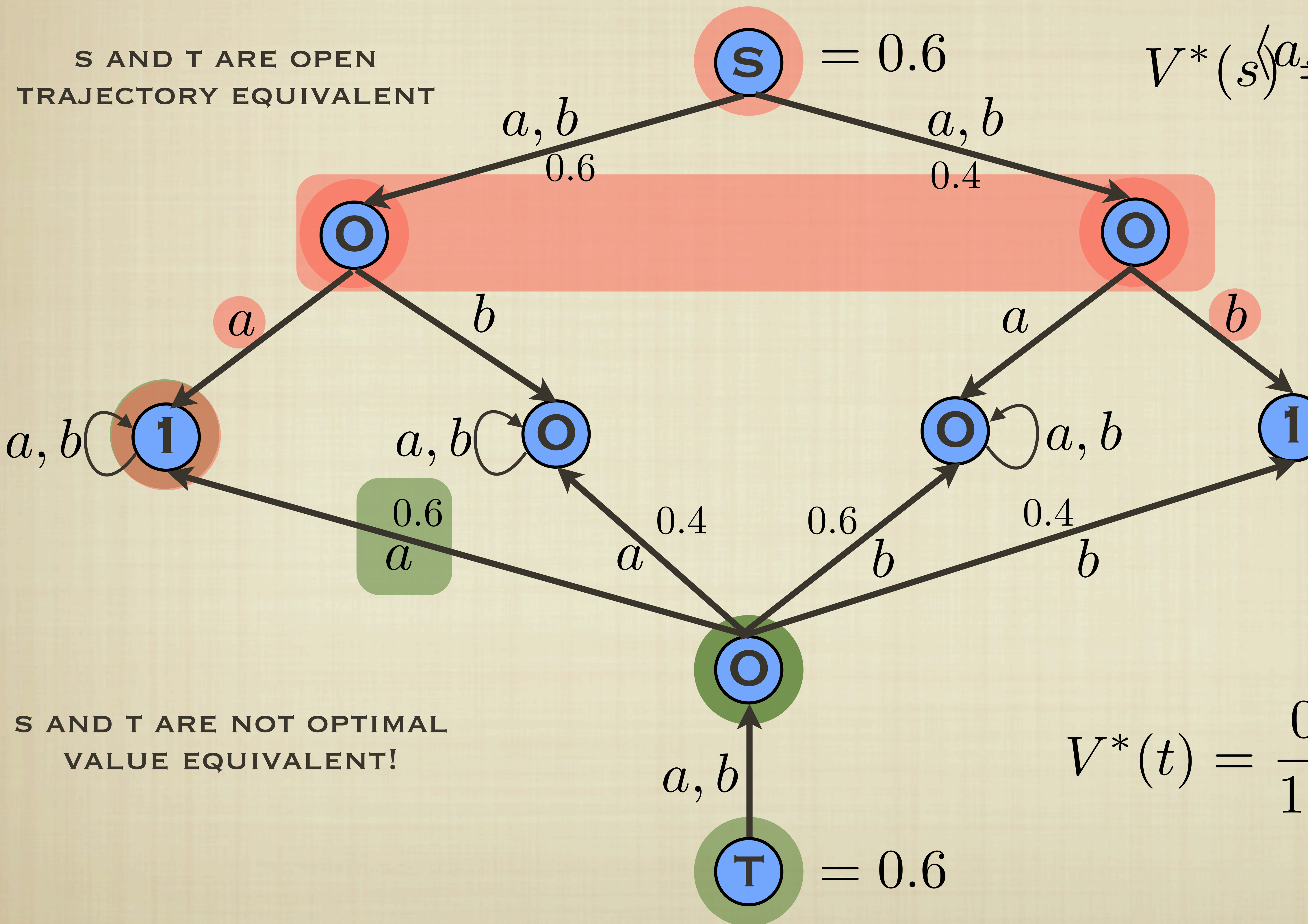
$\mathcal{I}$ – *open trajectory*

$\not\downarrow$

$(\mathcal{D}, \mathcal{I})$ – *optimal value*

Hierarchy

S AND T ARE OPEN TRAJECTORY EQUIVALENT



$$V^*(s) = \frac{\langle a, 0, a, 1 \rangle}{1-\gamma}$$

γ

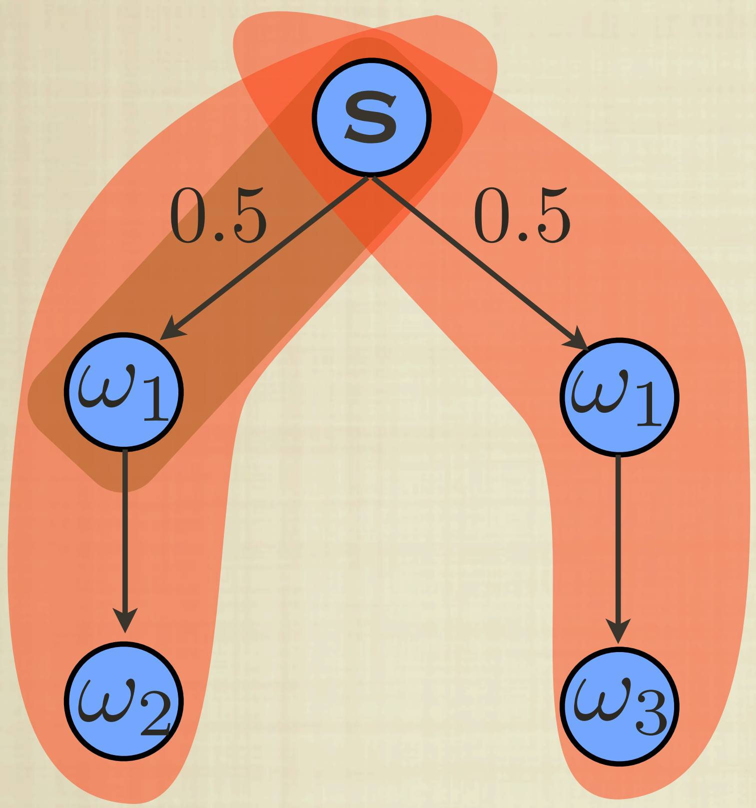
S AND T ARE NOT OPTIMAL VALUE EQUIVALENT!

$$V^*(t) = \frac{0.6\gamma}{1-\gamma}$$

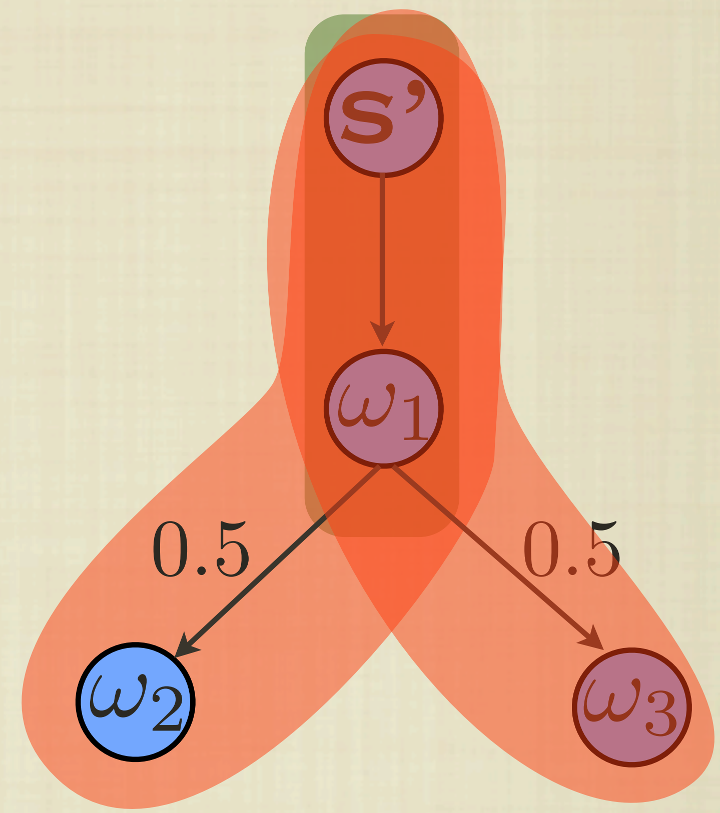
Outline

- POMDP review
- New POMDP formulation
- Equivalence relations
  - Value functions
  - Trajectory predictions
  - Bisimulation
- Conclusions and future work

Bisimulation



0.5



Bisimulation

- An equivalence relation E is a $(\mathcal{D},\mathcal{I})$-bisimulation relation if whenever μ,ν are $(\mathcal{D},\mathcal{I})$-bisimilar then
  - For all $\omega \in \Omega_{\mathcal{I}}, a \in A$, $Pr(\omega|\mu,a) = Pr(\omega|\nu,a)$
  - For all $c \in \mathcal{B}/_E, a \in A$,

$$\sum_{\mu' \in c} T_{\mathcal{D}}(\mu,a)(\mu') = \sum_{\mu' \in c} T_{\mathcal{D}}(\nu,a)(\mu')$$

- If μ and ν are $(\mathcal{D},\mathcal{I})$-bisimilar we will write $\mu \sim \nu$.

Deterministic Bisimulation

- An equivalence relation E is a deterministic $(\mathcal{D}, \mathcal{I})$-bisimulation relation if whenever μ, ν are deterministic $(\mathcal{D}, \mathcal{I})$-bisimilar then
  - For all $\omega \in \Omega_{\mathcal{I}}, a \in A$, $Pr(\omega|\mu, a) = Pr(\omega|\nu, a)$
  - For all $\omega \in \Omega_{\mathcal{D}}$, $a \in A$, $\tau_{\mathcal{D}}(\mu, a, \omega) E \tau_{\mathcal{D}}(\nu, a, \omega)$
- If μ and ν are deterministic $(\mathcal{D}, \mathcal{I})$-bisimilar we will write $\mu \simeq \nu$.

Hierarchy

$$\mathcal{D} \subseteq \mathcal{I}$$

$$Deterministic$$

$$(\mathcal{D}, \mathcal{I}) - bisimulation$$

$$=$$

$$\begin{tikzcd}
\mathcal{I} - open\ trajectory \arrow[d] & & \mathcal{I} - closed\ trajectory \arrow[d] \\
(\mathcal{D}, \mathcal{I}) - bisimulation \arrow[dr] & & (\mathcal{D}, \mathcal{I}) - closed\ value \arrow[dl] \\
& (\mathcal{D}, \mathcal{I}) - optimal\ value &
\end{tikzcd}$$

Hierarchy

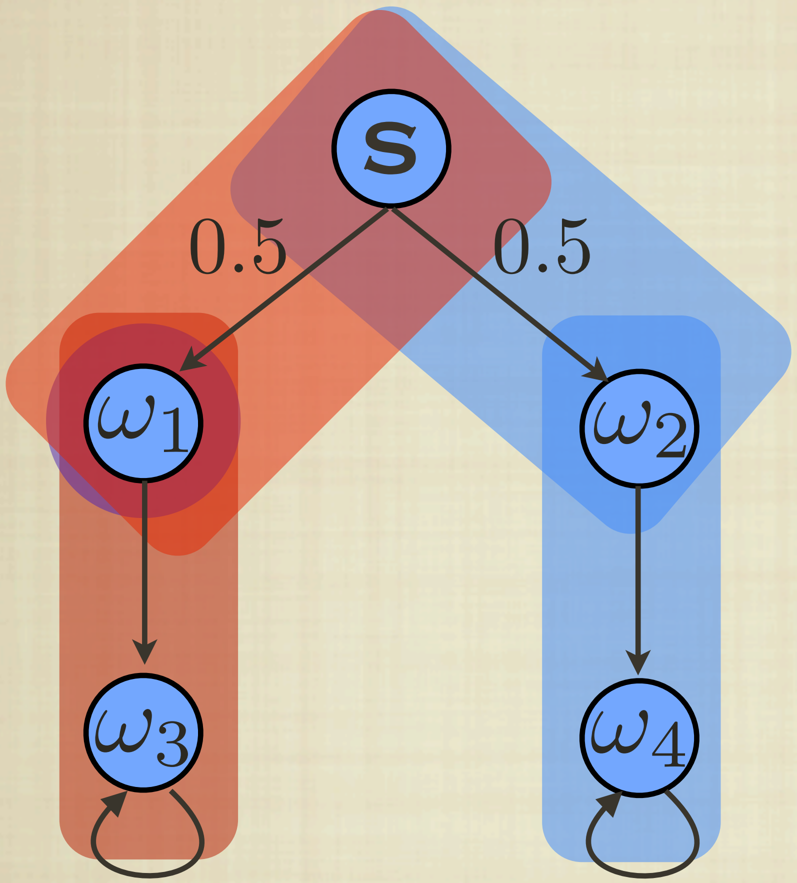
$$\begin{array}{c} Deterministic \\ (\mathcal{D},\mathcal{I})-bisimulation \end{array} \not\longleftrightarrow (\mathcal{D},\mathcal{I})-bisimulation$$

$(\mathcal{D},\mathcal{I}) - bisimulation$

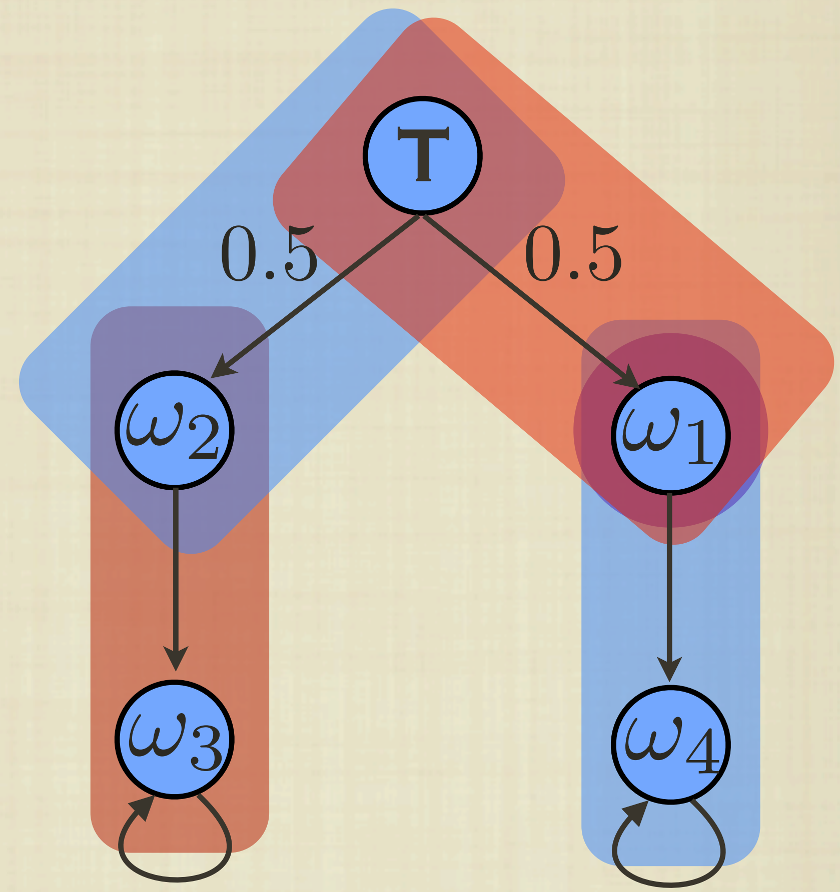
$\not\downarrow$

$Deterministic$
$(\mathcal{D},\mathcal{I}) - bisimulation$

Hierarchy



$$s \sim t$$
$$s \not\approx t$$



Hierarchy

$\mathcal{D} \not\subseteq \mathcal{I}$

$$Deterministic$$

$$(\mathcal{D},\mathcal{I}) - bisimulation$$

$$=$$

$$\begin{tikzcd}
\mathcal{I} - open\ trajectory \arrow[d] & & \mathcal{I} - closed\ trajectory \arrow[d] \\
(\mathcal{D},\mathcal{I}) - bisimulation \arrow[dr] & & (\mathcal{D},\mathcal{I}) - closed\ value \arrow[dl] \\
& (\mathcal{D},\mathcal{I}) - optimal\ value &
\end{tikzcd}$$

Hierarchy

$\mathcal{D} \not\subseteq \mathcal{I}$

$$\textit{Deterministic}$$

$$(\mathcal{D},\mathcal{I}) - bisimulation$$

$$\begin{tikzcd}
(\mathcal{D},\mathcal{I}) - bisimulation \arrow[d] \arrow[dr] & & (\mathcal{D},\mathcal{I}) - closed\ value \arrow[dl] \\
\mathcal{I} - open\ trajectory & (\mathcal{D},\mathcal{I}) - optimal\ value &
\end{tikzcd}$$

Hierarchy

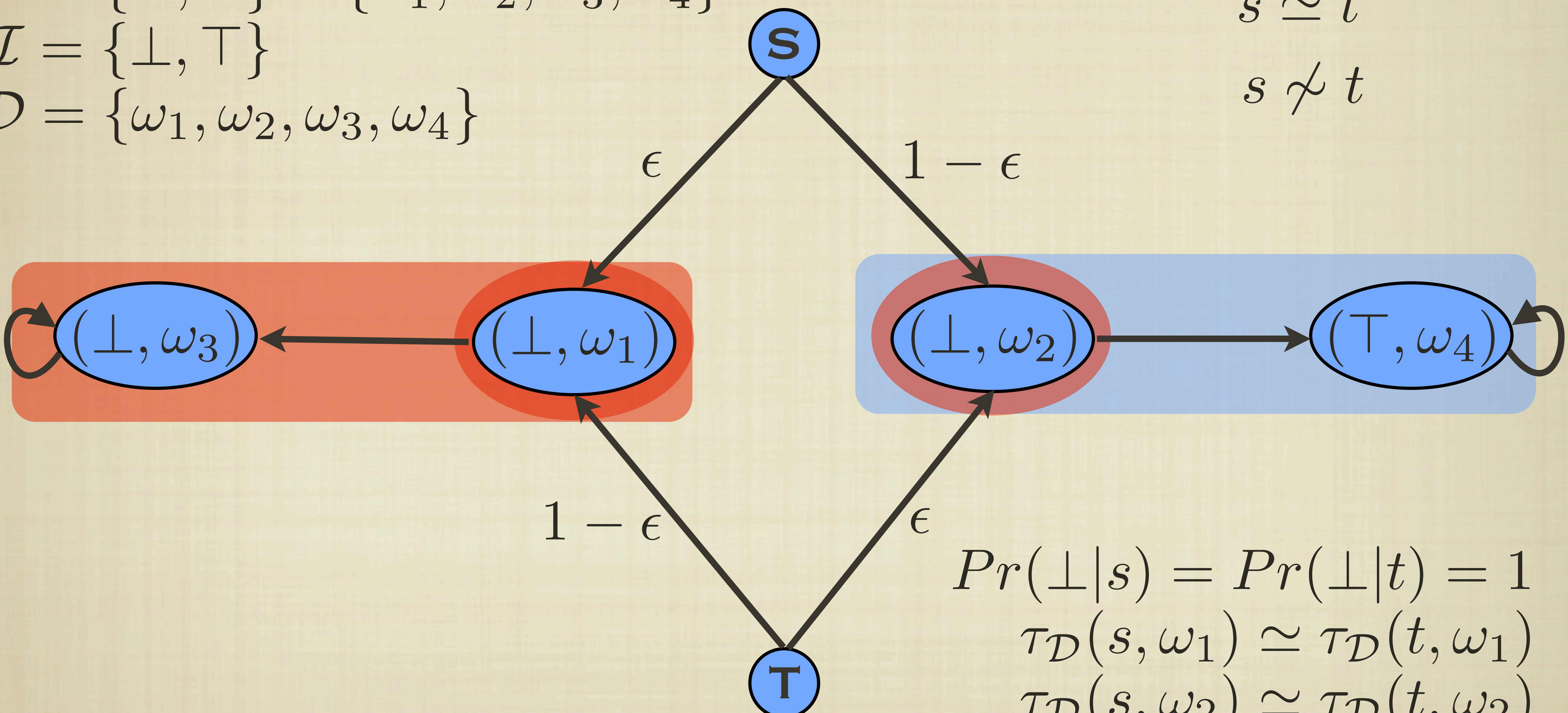
$$\begin{array}{c} Deterministic \\ (\mathcal{D},\mathcal{I}) - bisimulation \end{array} \not\longrightarrow (\mathcal{D},\mathcal{I}) - bisimulation$$

Hierarchy

$$\begin{array}{c} \textit{Deterministic} \\ (\mathcal{D},\mathcal{I}) - \textit{bisimulation} \\ \not\downarrow \\ (\mathcal{D},\mathcal{I}) - \textit{bisimulation} \end{array}$$

$$\Omega = \{\bot, \top\} \times \{\omega_1, \omega_2, \omega_3, \omega_4\}$$
$$\mathcal{I} = \{\bot, \top\}$$
$$\mathcal{D} = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

$$s \simeq t$$
$$s \not\simeq t$$



$$Pr(\bot|s) = Pr(\bot|t) = 1$$
$$\tau_{\mathcal{D}}(s, \omega_1) \simeq \tau_{\mathcal{D}}(t, \omega_1)$$
$$\tau_{\mathcal{D}}(s, \omega_2) \simeq \tau_{\mathcal{D}}(t, \omega_2)$$

Hierarchy

$\mathcal{D} \not\subseteq \mathcal{I}$

Deterministic

$(\mathcal{D},\mathcal{I}) - bisimulation$

$$\begin{tikzcd}
(\mathcal{D},\mathcal{I}) - bisimulation \arrow[d] \arrow[dr] & & (\mathcal{D},\mathcal{I}) - closed\ value \arrow[dl] \\
\mathcal{I} - open\ trajectory & (\mathcal{D},\mathcal{I}) - optimal\ value &
\end{tikzcd}$$

Strengthening open trajectory

$\mathcal{D} \not\subseteq \mathcal{I}$

$$\begin{tikzcd}
\Delta^* = (\mathcal{D},\mathcal{I}) - \mathit{bisimulation} \arrow[ddr] & \\
\Delta \arrow[u] \arrow[loop left] & \\
 & \mathcal{I} - \mathit{open\ trajectory} \arrow[ul]
\end{tikzcd}$$

From (Castro et al., 2009)

Conclusions

- Subsets must be chosen with care to avoid sub-optimal performance
- Open trajectory equivalence is closely related to PSRs; we showed this is not appropriate with respect to bad choices of $\mathcal{D}$ and $\mathcal{I}$.
- In most situations we would require $\mathcal{D} \subseteq \mathcal{I}$.
- $(\mathcal{D}, \mathcal{I})$-bisimulation is robust even when $\mathcal{D} \not\subseteq \mathcal{I}$.

Conclusions

$\mathcal{D} \subseteq \mathcal{I}$

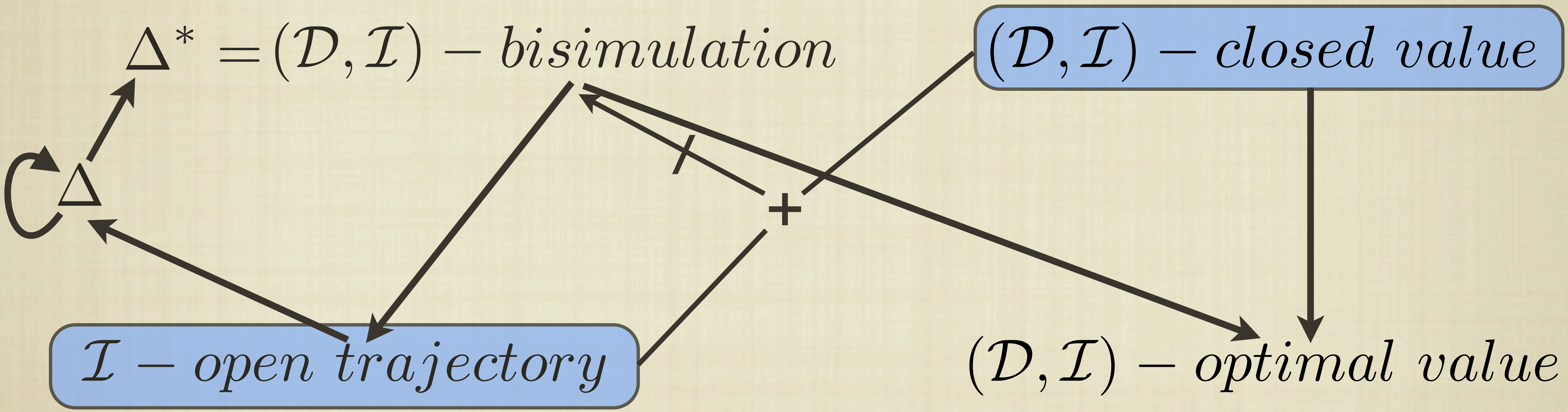
$$\begin{array}{ccc}
\begin{array}{c} Deterministic \\ (\mathcal{D},\mathcal{I})-bisimulation \\ = \\ \mathcal{I}-open\ trajectory \end{array} & & \mathcal{I}-closed\ trajectory \\
\downarrow & & \downarrow \\
(\mathcal{D},\mathcal{I})-bisimulation & & (\mathcal{D},\mathcal{I})-closed\ value \\
& \searrow \quad \swarrow & \\
& (\mathcal{D},\mathcal{I})-optimal\ value &
\end{array}$$

Conclusions

$\mathcal{D} \not\subseteq \mathcal{I}$

$Deterministic$

$(\mathcal{D},\mathcal{I}) - bisimulation$



Current work

- We are currently working on learning algorithms for determining $\mathcal{D}$, assuming $\mathcal{I}$ is known.
  - Start with a small $\mathcal{D}$, incrementally add more observations.
  - Start planning/learning with a small $\mathcal{D}$, use an expert/oracle to determine whether more observations are necessary

Future work

- We project Ω onto $\Omega_{\mathcal{D}}$ and $\Omega_{\mathcal{I}}$ using binary projection matrices.

  - If we allow general projection matrices, does open trajectory equivalence yield something similar to TPSRs (Rosencratz & Gordon, 2004; Boots et al., 2010).

- Life-long learning: Many tasks to solve, different choices of $\mathcal{D}$ and $\mathcal{I}$, depending on task.

- ranking of observations to dynamically set $\mathcal{D}$ based on time requirements.

References

- Kozen, D. (2007). Coinductive proof principles for stochastic processes. Logical Methods in Computer Science 3(4:8). DOI: 10.2168/LMCS-3 (4:8) 2007.

- Littman, M., R. Sutton, and S. Singh (2002). Predictive representations of state. In Proceedings of the 14th Conference on Advances in Neural Information Processing Systems (NIPS-02), pp. 1555–1561.

- Castro, P. S., P. Panangaden, and D. Precup (2009). Notions of state equivalence under partial observability. In Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09), pp. 1653–1658.

- Boots, B., S. M. Siddiqi, and G. J. Gordon (2010). Closing the learning-planning loop with predictive state representations. In Proc. Robotics: Science and Systems VI.

- Rosencrantz, M. and G. Gordon (2004). Learning low dimensional predictive represen- tations. In Proceedings of the International Conference on Machine Learning (ICML- 04).

Closed and optimal value equivalences

V^π equivalence
$\downarrow$
V^* equivalence

Lemma: If s_0 and t_0 are V^π equivalent then $V^*(s_0) \le V^*(t_0)$.

Proof: Assume $V^*(s_0) > V^*(t_0)$.

$\exists V. \quad V(s,\pi) \ge V^*(s)?$ Yes! Just take $V \equiv 1$

$V(s,\pi) \ge V^*(s) \Rightarrow \tau(V)(s,\pi) \ge V^*(s)?$ Yes!

Take any $s \neq t_0$ and $\pi \in \Pi_{CV}$

$$\tau(V)(s,\pi) = R(s,\pi(s)) + \gamma \sum_{s' \neq t_0} P(s,\pi(s))(s')V(s',\pi) + \gamma P(s,\pi(s))(t_0)V(s_0,\pi)$$

$$= R(s,\pi^*(s)) + \gamma \sum_{s' \neq t_0} P(s,\pi^*(s))(s')V(s',\pi) + \gamma P(s,\pi^*(s))(t_0)V(s_0,\pi)$$

$\pi(s') = \pi^*(s_0)$ if $s' = t_0$
$\pi(s') = \pi^*(s')$ otherwise

I.H.

$$\ge R(s,\pi^*(s)) + \gamma \sum_{s' \neq t_0} P(s,\pi^*(s))(s')V^*(s') + \gamma P(s,\pi^*(s))(t_0)V^*(s_0)$$

$$\ge R(s,\pi^*(s)) + \gamma \sum_{s' \neq t_0} P(s,\pi^*(s))(s')V^*(s') + \gamma P(s,\pi^*(s))(t_0)V^*(t_0)$$

$$= V^*(s)$$

We've shown that for any $s \neq t_0$ and $\pi \in \Pi_{CV}$, $V^\pi(s) \ge V^*(s)$ with strict inequality if $P(s,\pi^*(s))(t_0) > 0$

By last corollary we know $\exists s' \neq t_0.\ P(s',\pi^*(s'))(t_0) > 0$

Thus, $V^\pi(s') > V^*(s')$, contradicting optimality of $V^*(s')$

By contradiction, $V^*(s_0) \le V^*(t_0)$

Q.E.D.

Specifying data and interest

- Let $\Phi_{\mathcal{D}}$ be a projection matrix used to compute $O_{\mathcal{D}} : n \times |\Omega_{\mathcal{D}}|$:

$$O_{\mathcal{D}} = O\Phi_{\mathcal{D}}$$

- If we have $\mathcal{D}_2 \subseteq \mathcal{D}_1$, the projection Φ_{12} yields the following:

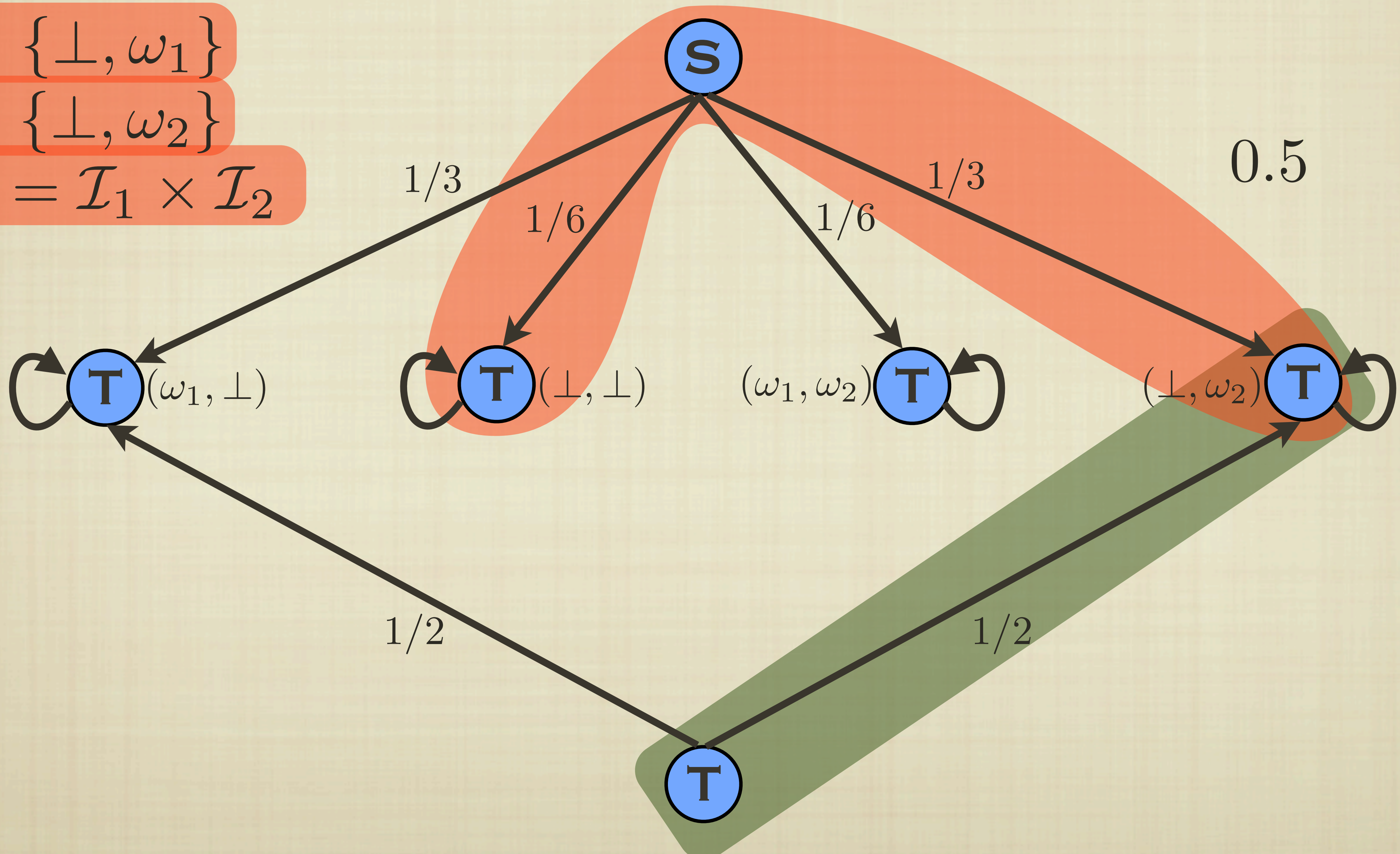
$$\Phi_{\mathcal{D}_2} = \Phi_{\mathcal{D}_1}\Phi_{12}$$

$$O_{\mathcal{D}_2} = O_{\mathcal{D}_1}\Phi_{12}$$

Approximating bisimulation

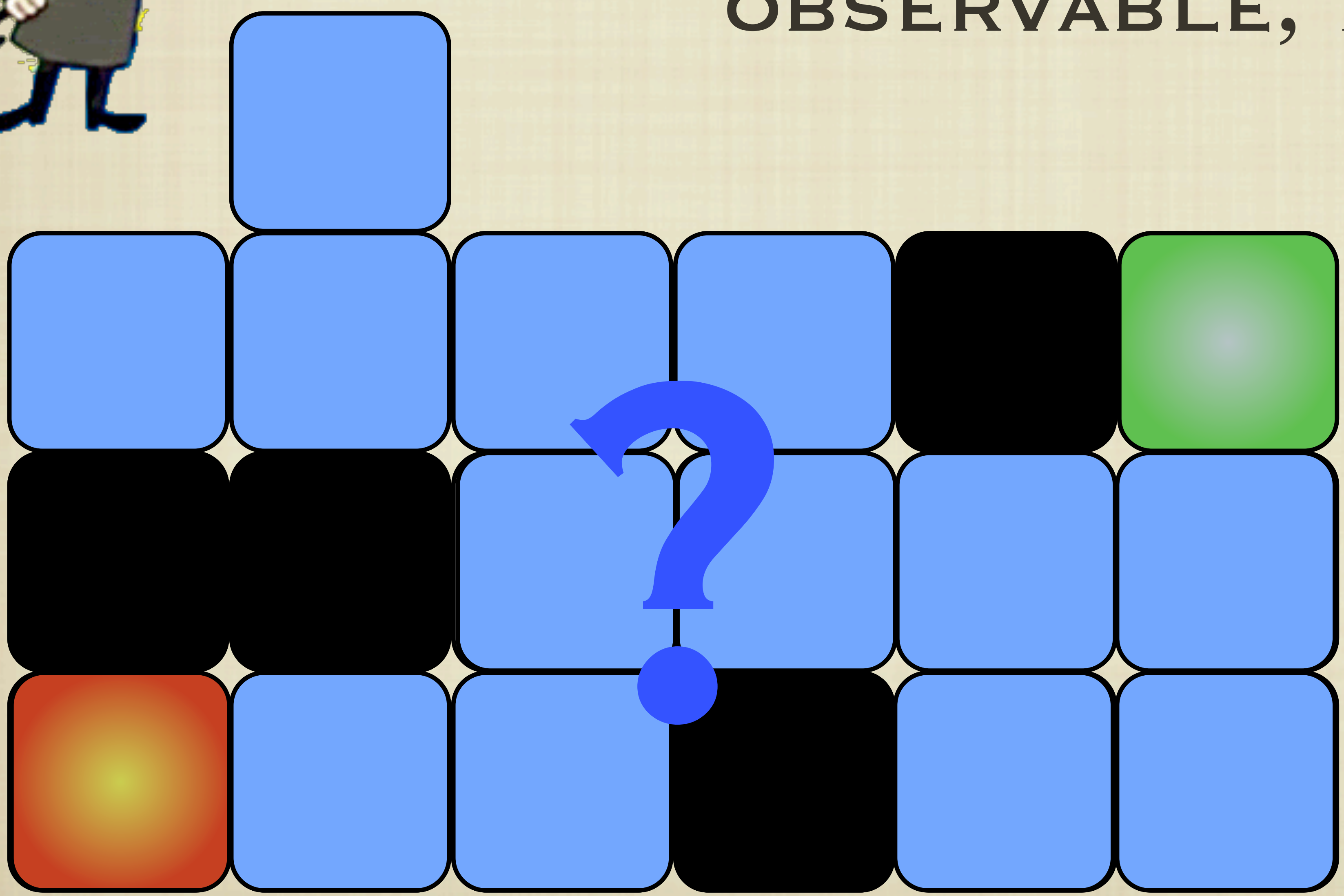
- **Proposition: Given $\mathcal{D}, \mathcal{I}, \mu, \nu$ may be $(\mathcal{D}, \mathcal{I}_i)$-bisimilar for all $\mathcal{I}_i \subset \mathcal{I}$, but fail to be $(\mathcal{D}, \mathcal{I})$-bisimilar.**

$$\mathcal{I}_1 = \{\perp, \omega_1\}$$
$$\mathcal{I}_2 = \{\perp, \omega_2\}$$
$$\mathcal{D} = \mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2$$



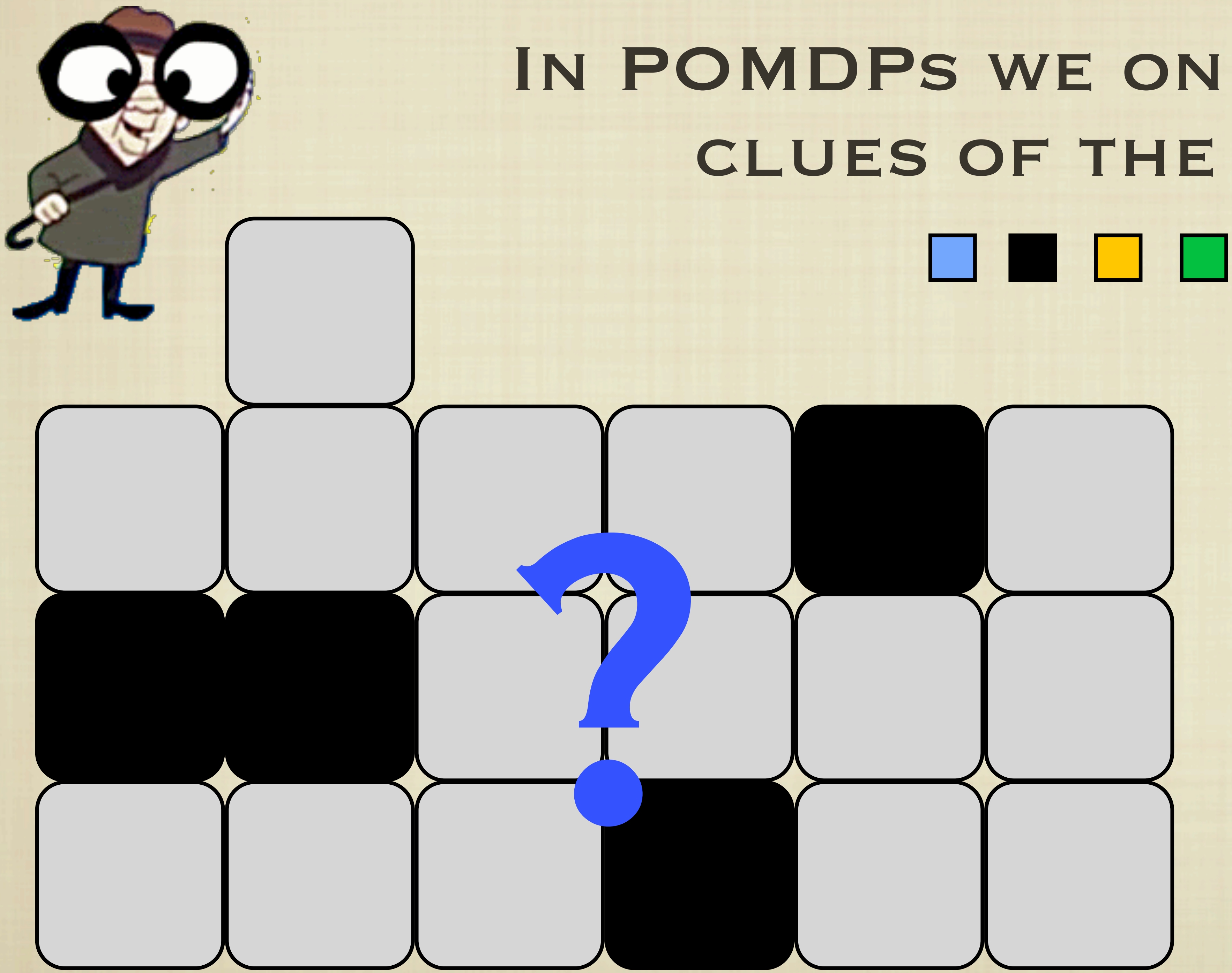
Partially observable MDPs (POMDPs)

If state is fully observable, it is a MDP

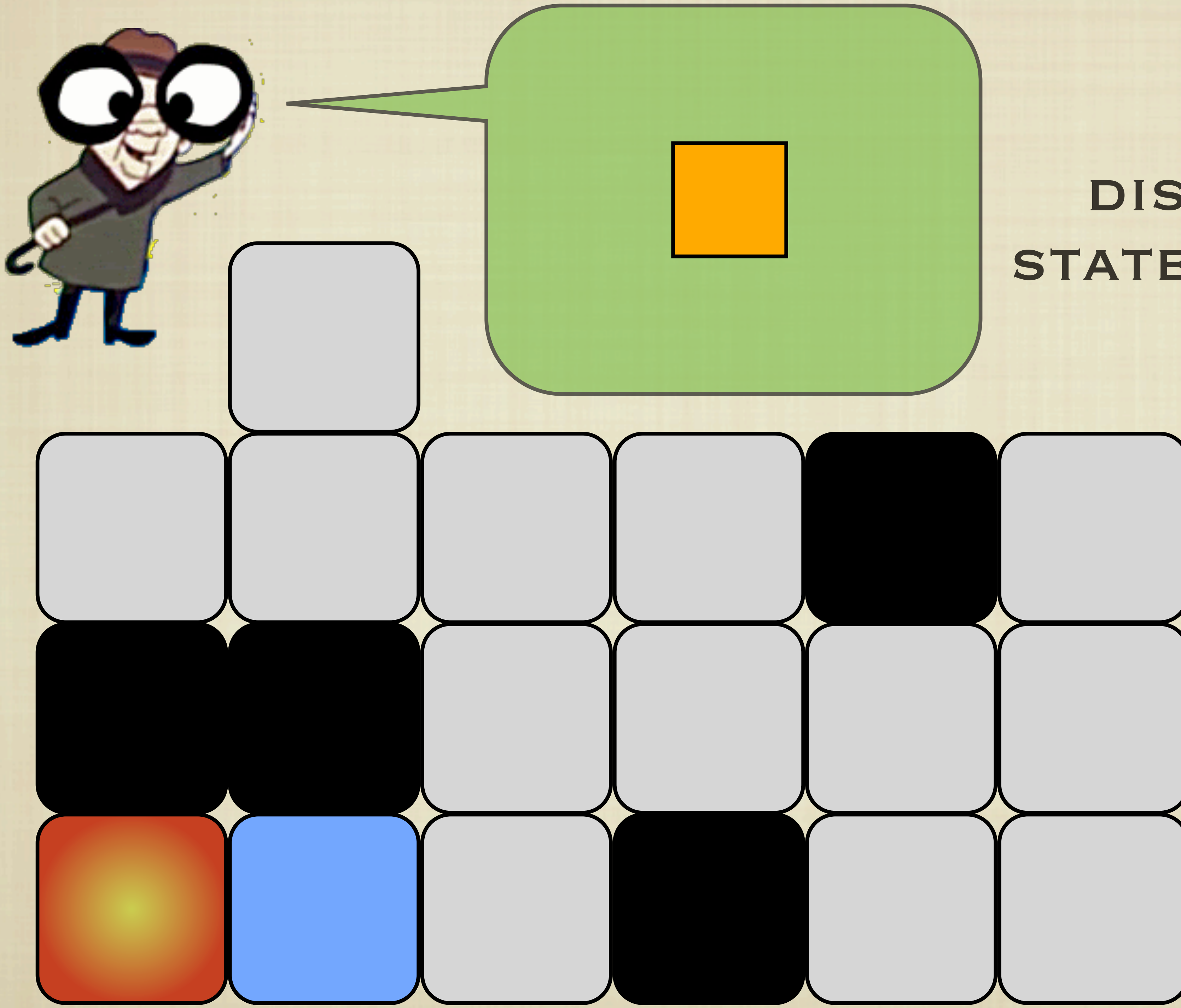


Partially observable MDPs (POMDPs)

In POMDPs we only receive clues of the state



Partially observable MDPs (POMDPs)



Maintain a distribution over states based on clues