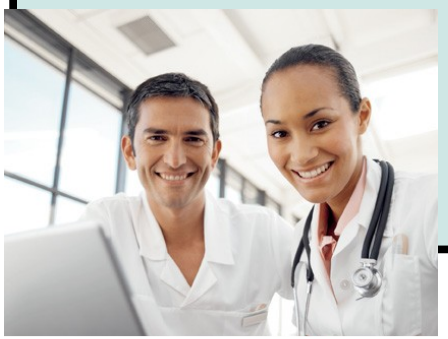
Partially-Observable Markov Decision Processes as Dynamical Causal Models

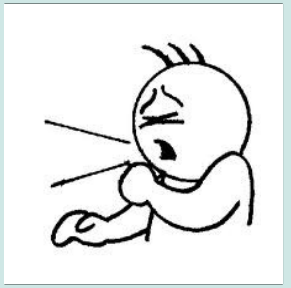
Finale Doshi-Velez
NIPS Causality Workshop 2013

The POMDP Mindset



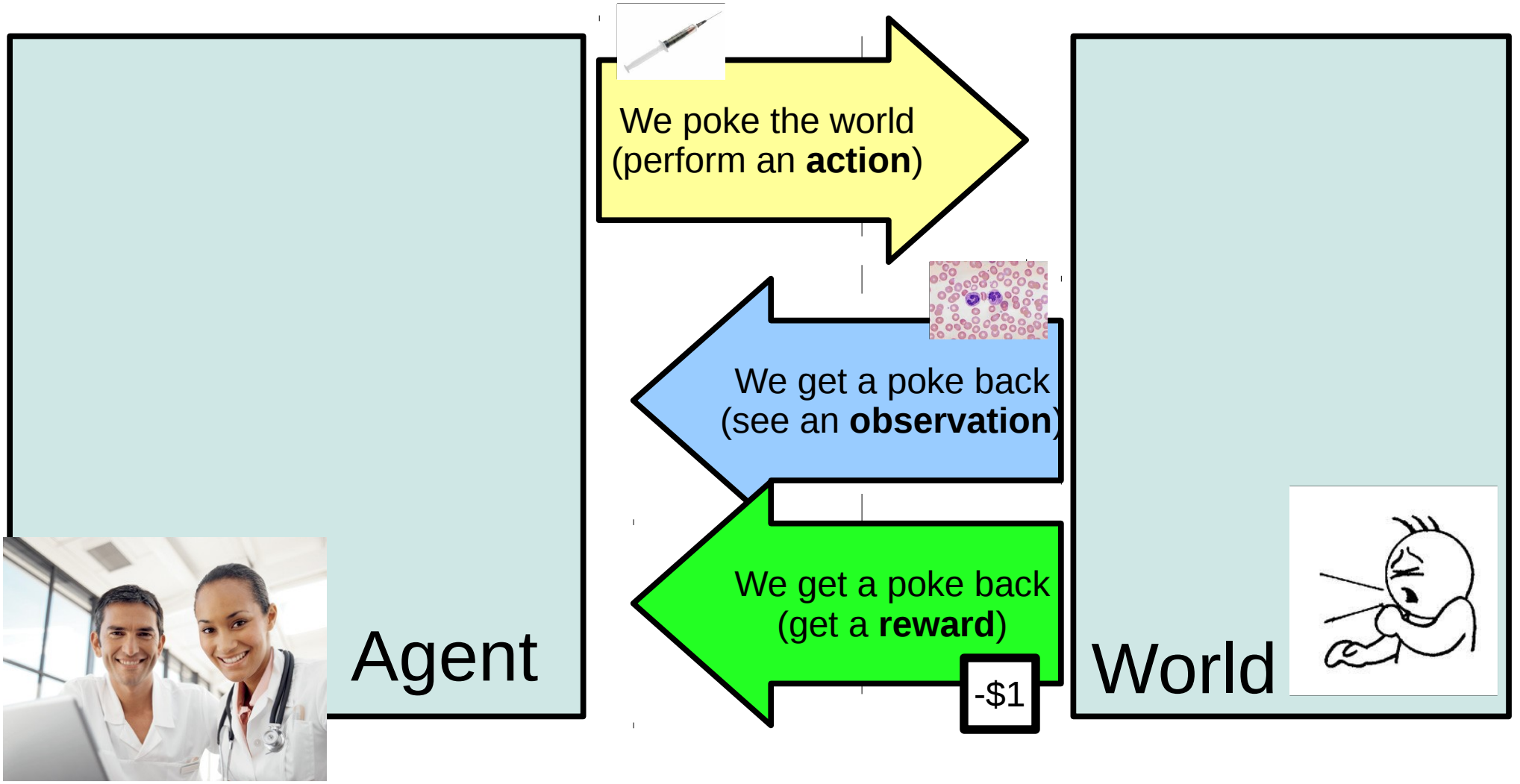
Agent

We poke the world
(perform an **action**)

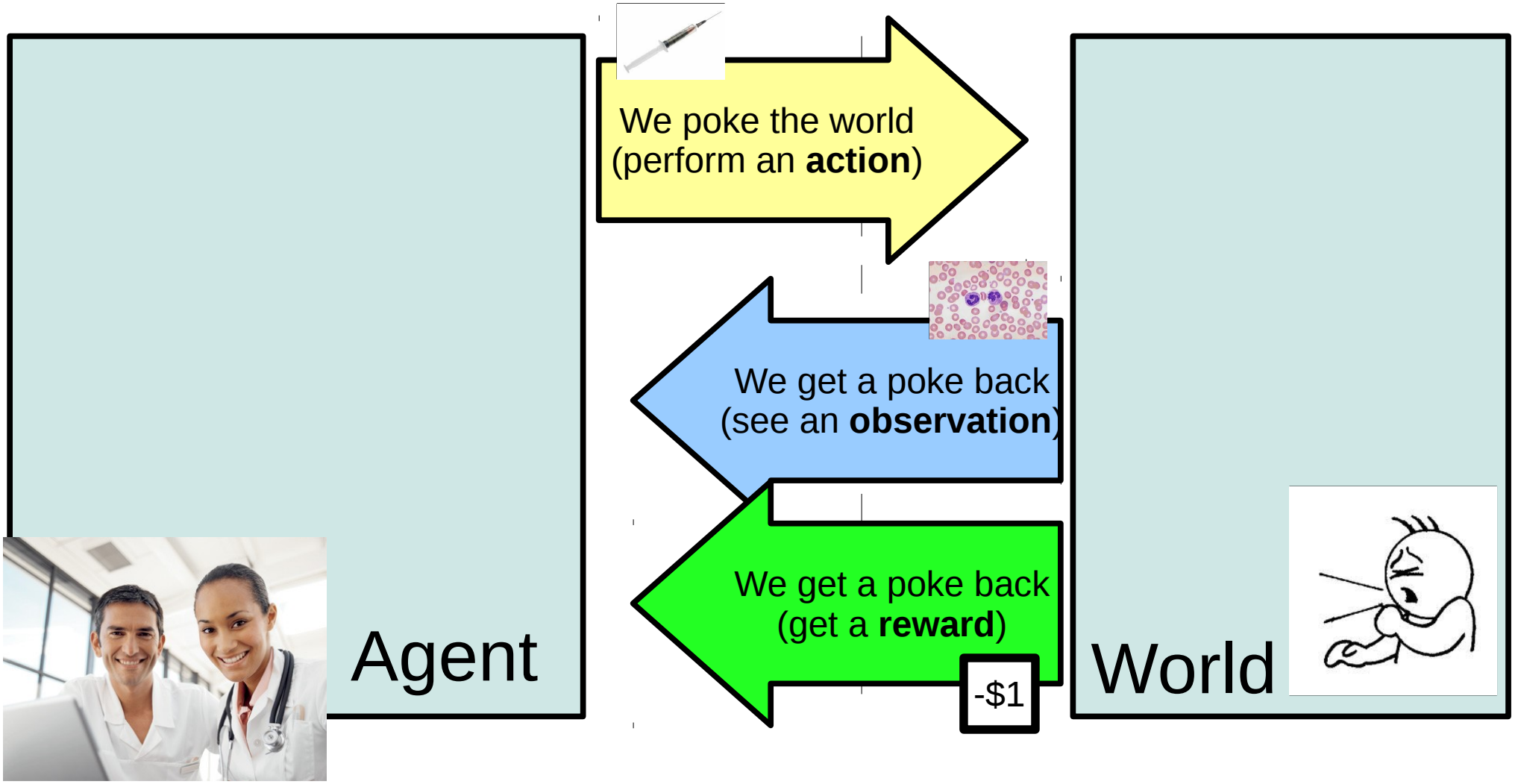


World

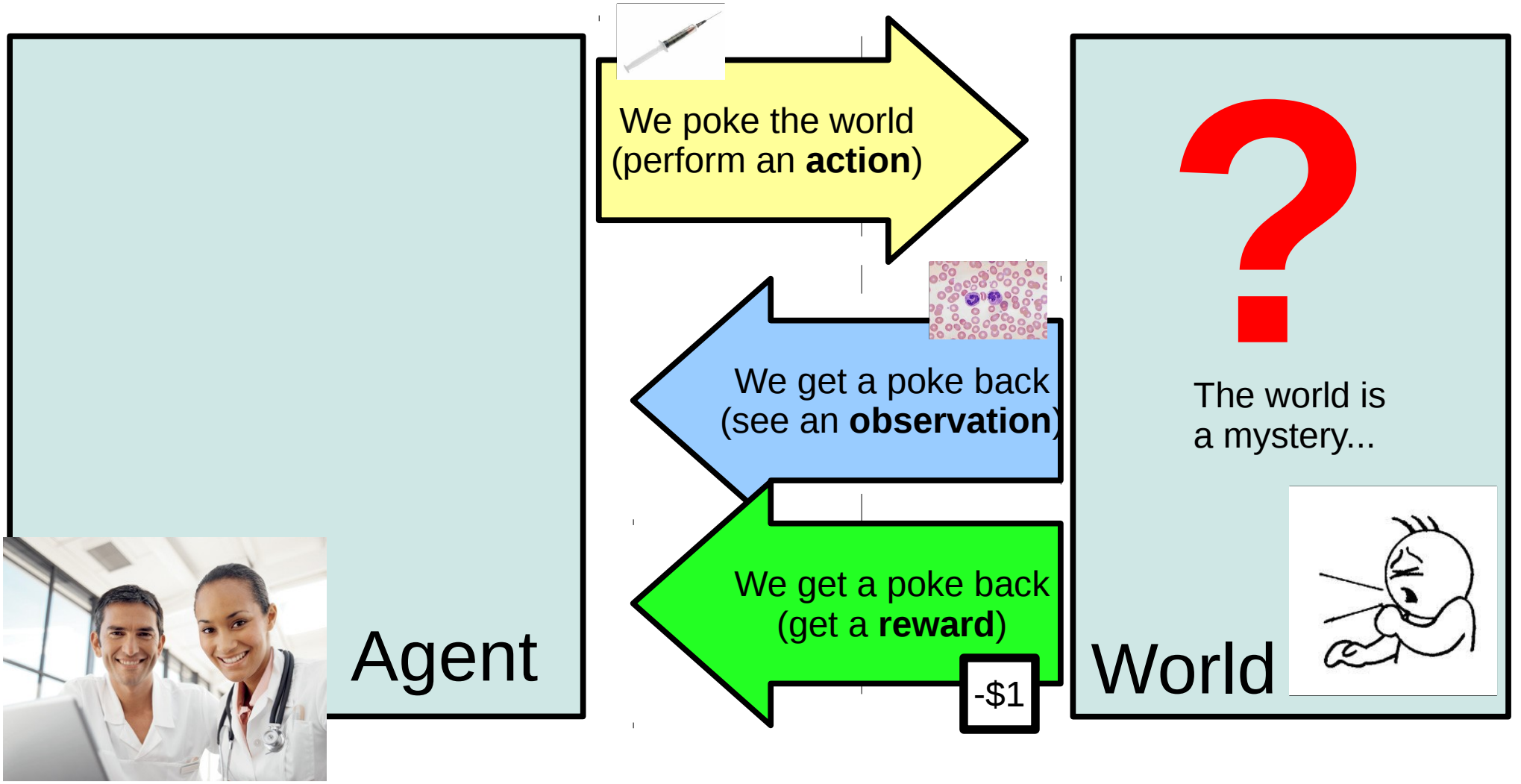
The POMDP Mindset



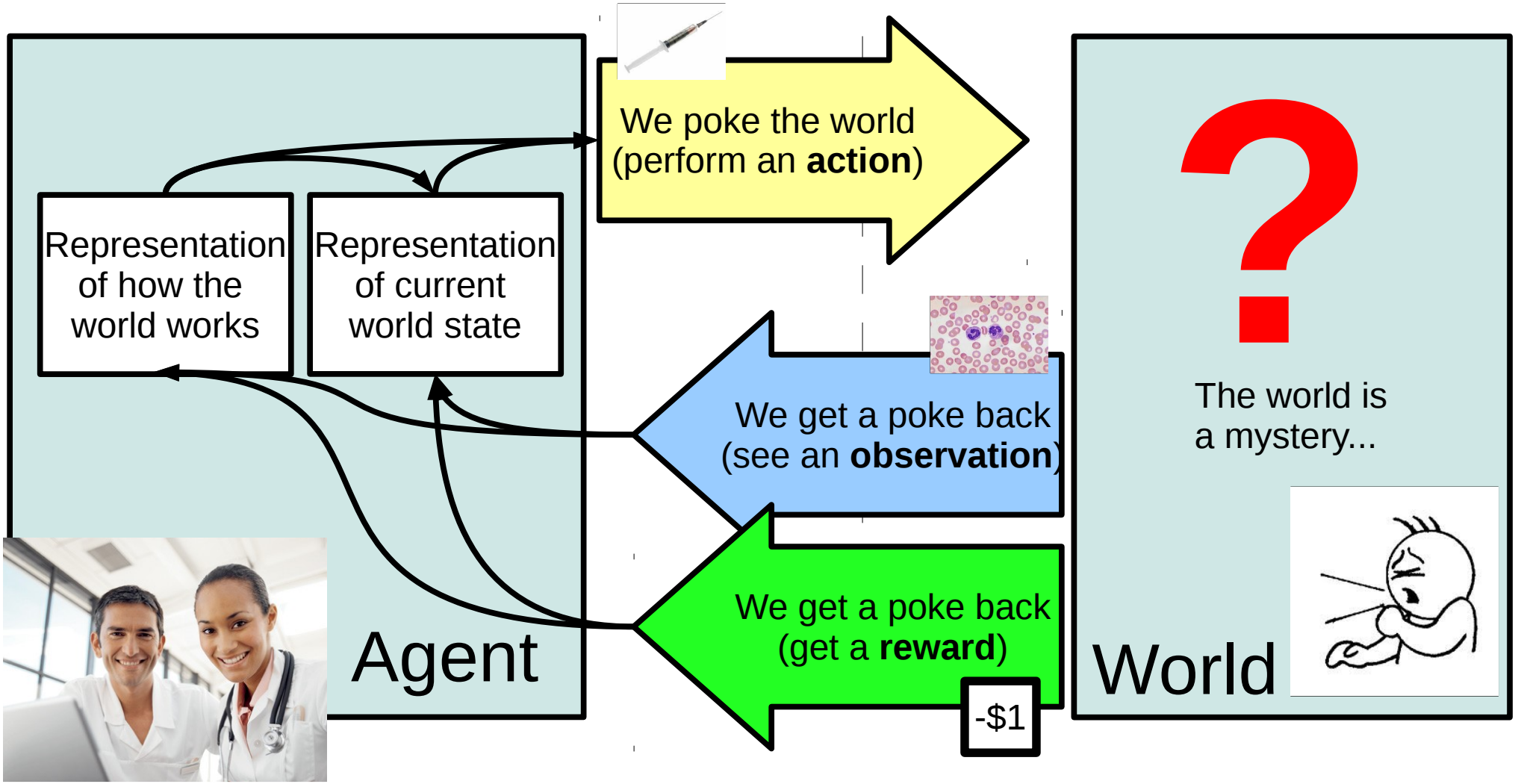
What next?


We poke the world
(perform an **action**)
We get a poke back
(see an **observation**)
We get a poke back
(get a **reward**)
-$1
Agent
World

What next?


We poke the world
(perform an **action**)
We get a poke back
(see an **observation**)
We get a poke back
(get a **reward**)
-$1
Agent
The world is
a mystery...
World

The agent needs a representation to use when making decisions

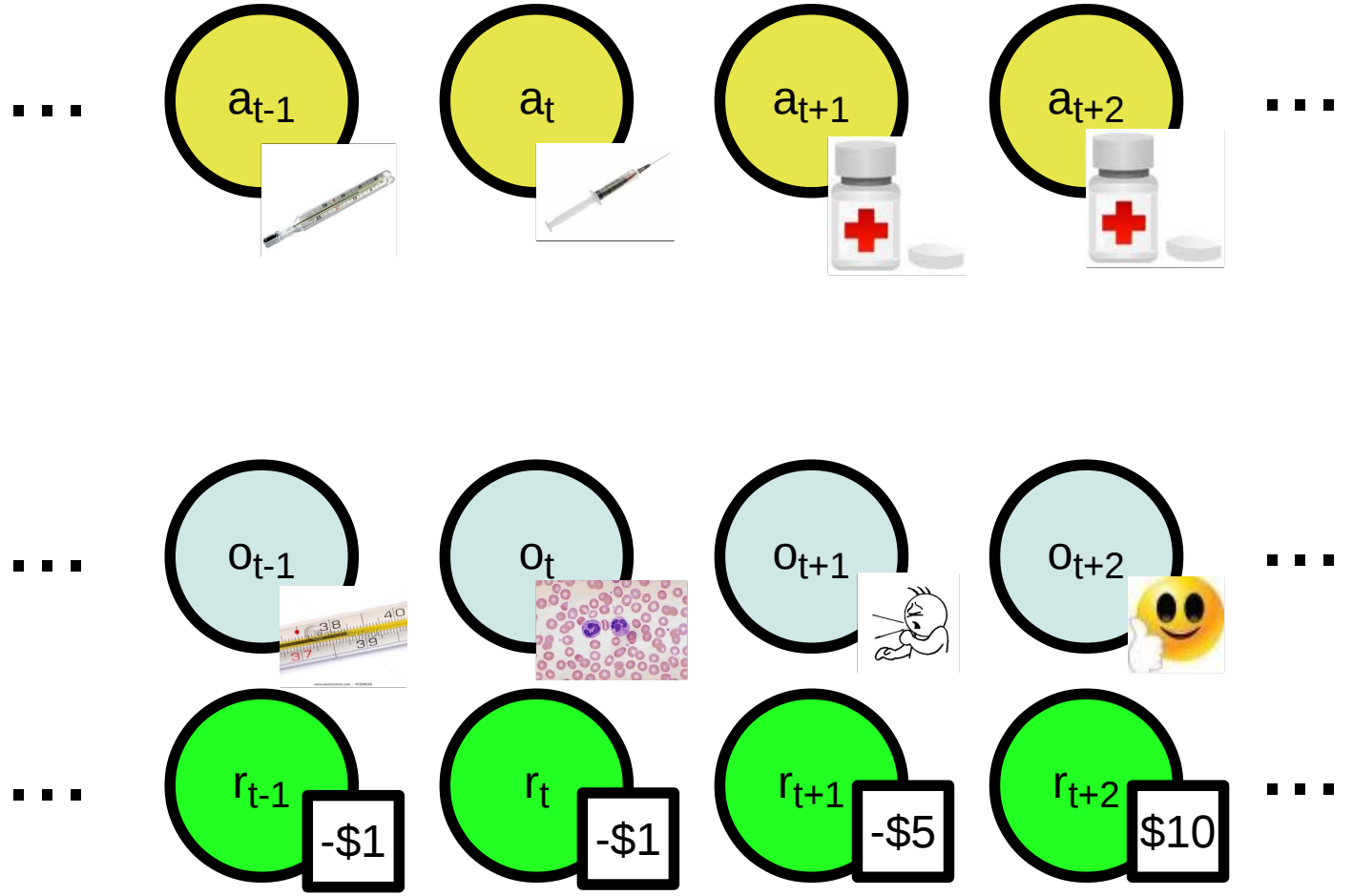


Many problems can be framed this way

- Robot navigation (take movement actions, receive sensor measurements)
- Dialog management (ask questions, receive answers)
- Target tracking (search a particular area, receive sensor measurements)

… the list goes on ...

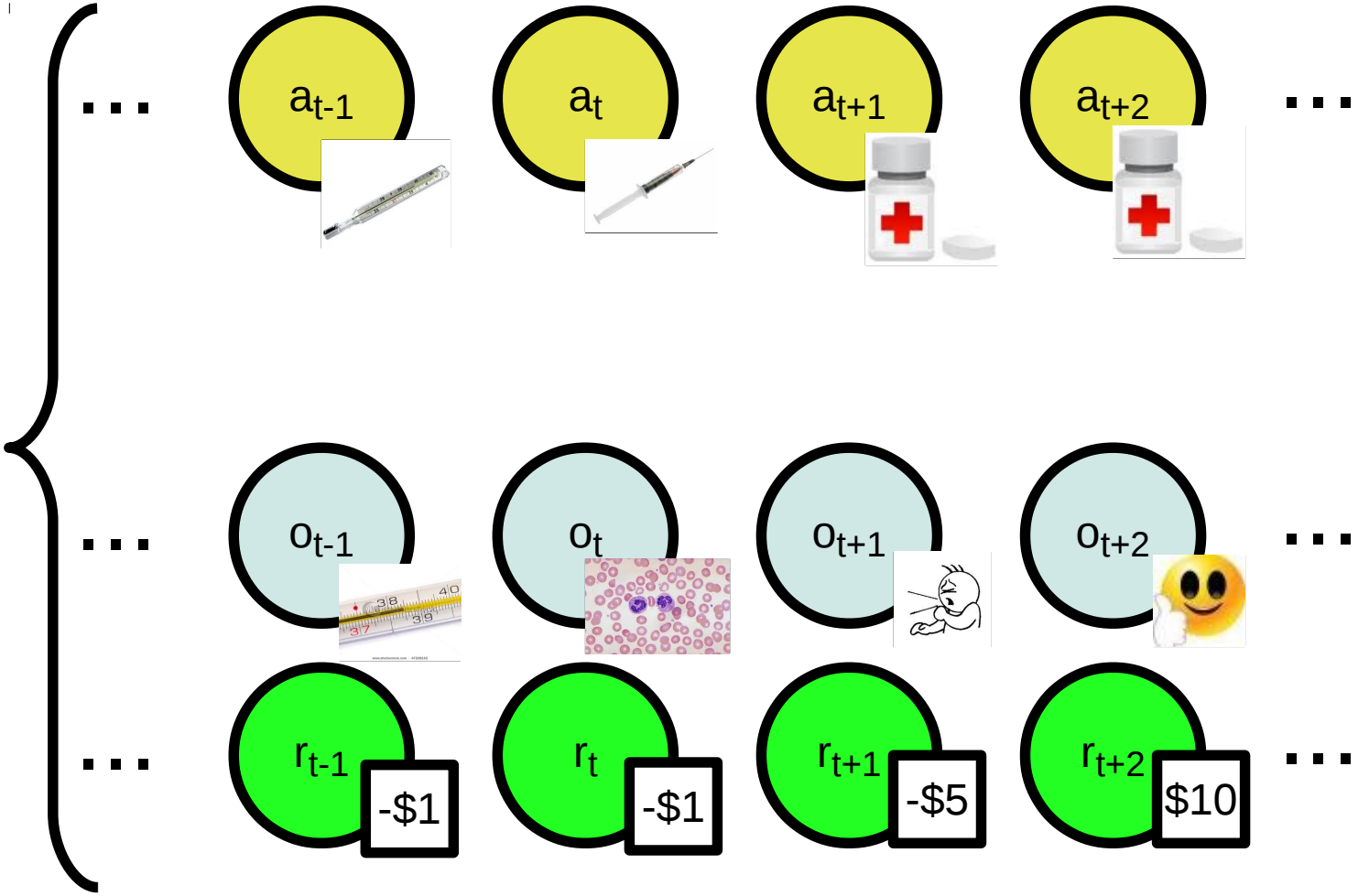
The Causal Process, Unrolled



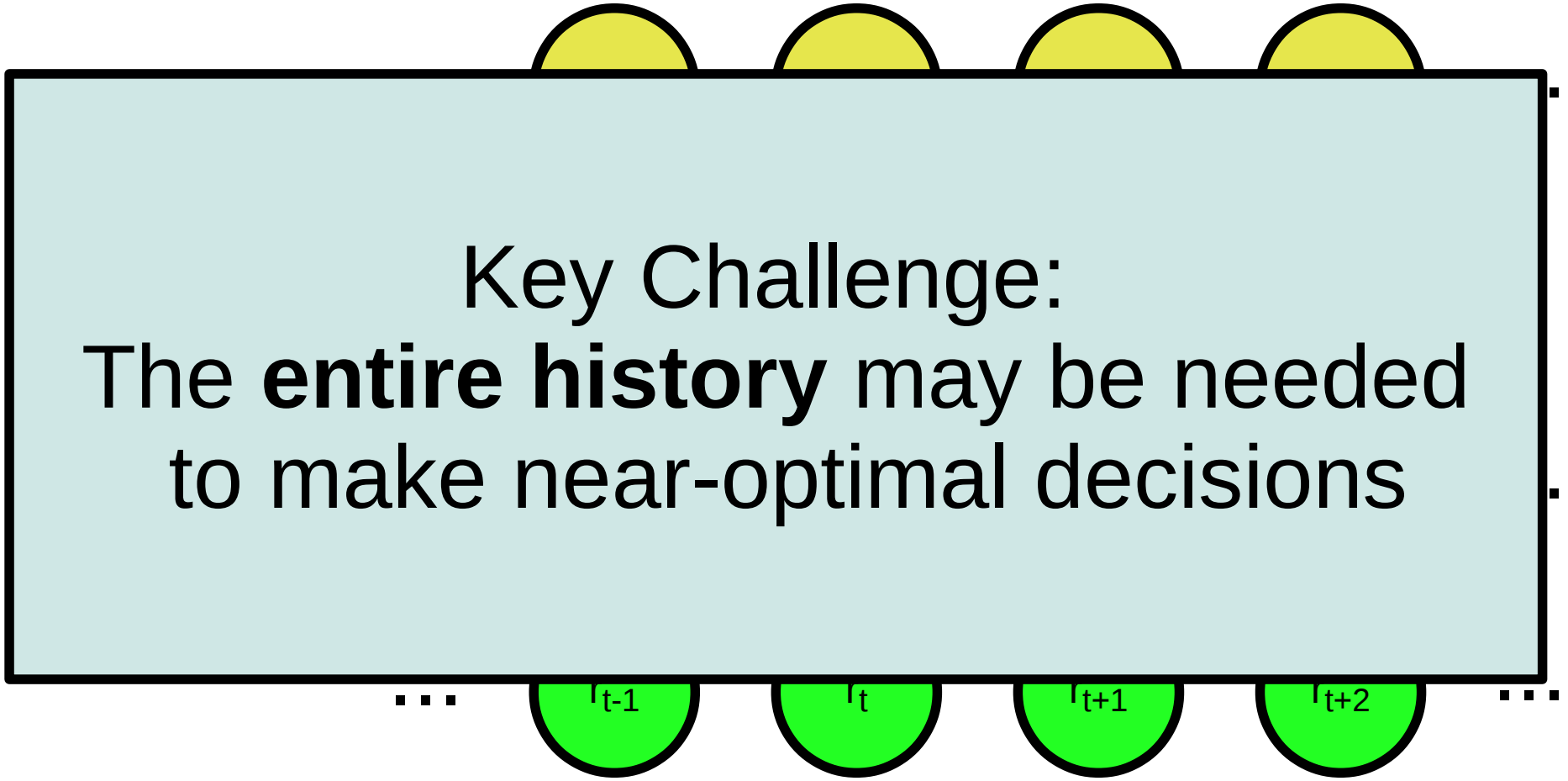
The Causal Process, Unrolled

Given a **history** of actions, observations, and rewards

How can we act in order to **maximize long-term future rewards**?

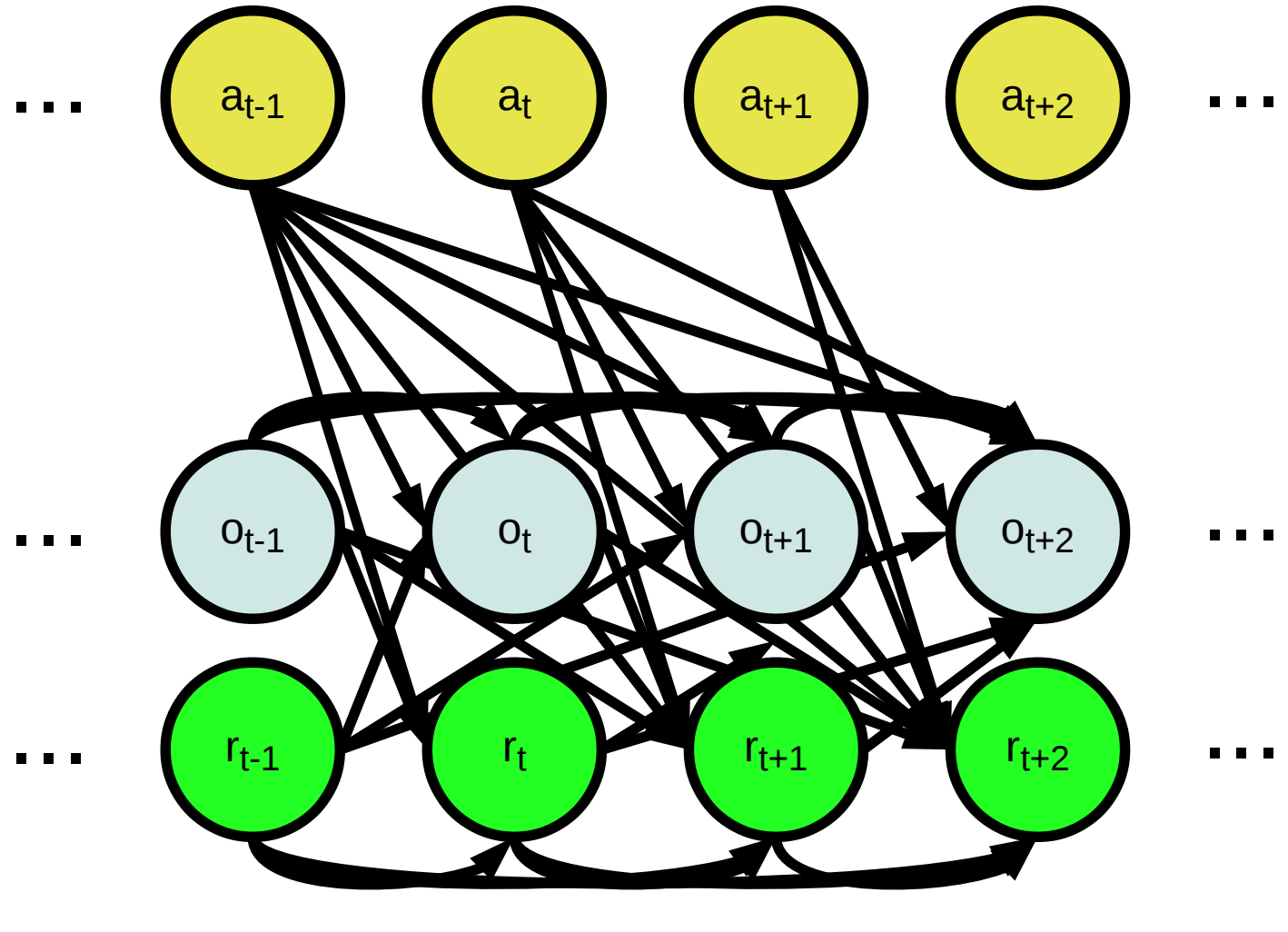


The Causal Process, Unrolled



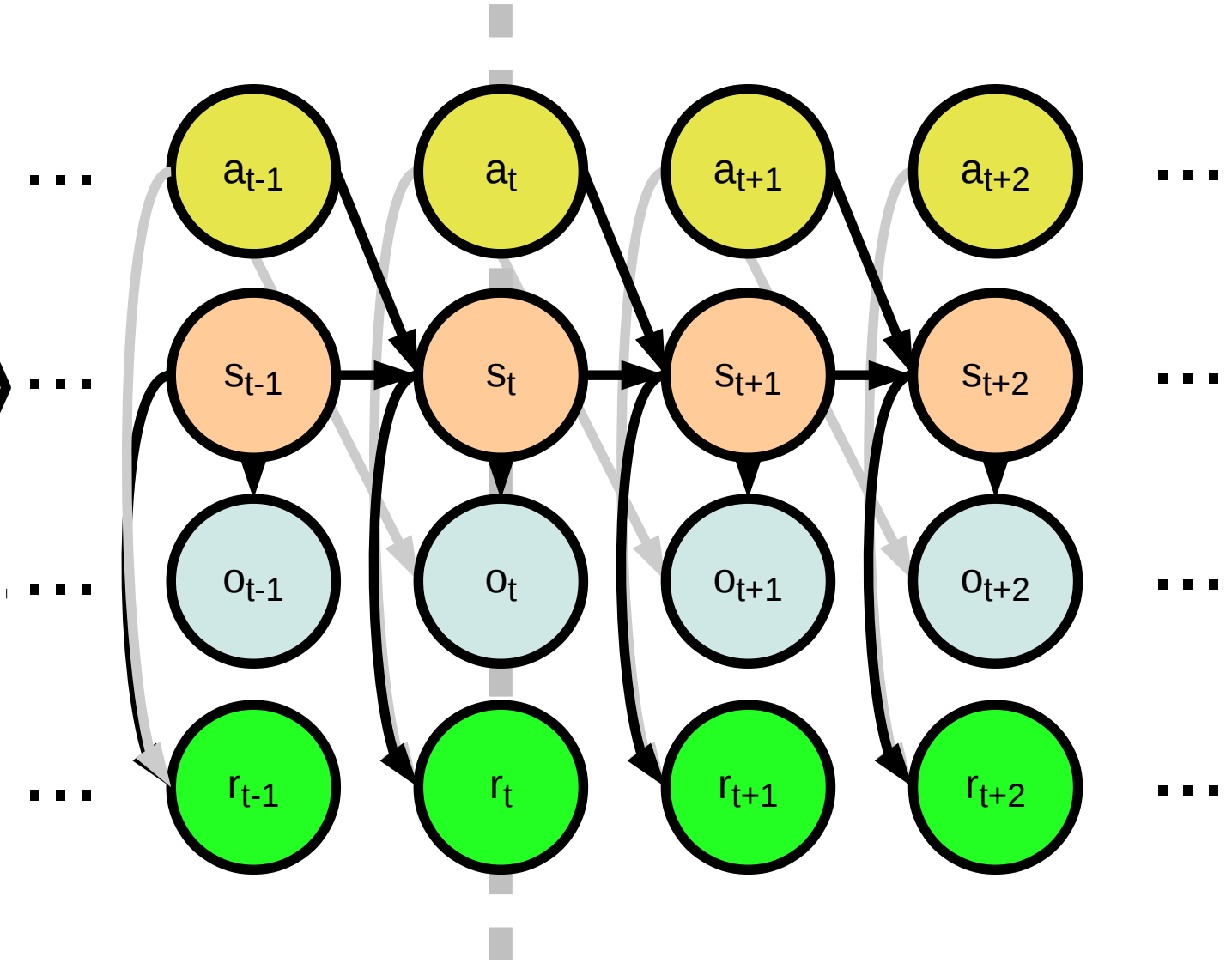
The Causal Process, Unrolled

All past events are needed to predict future events

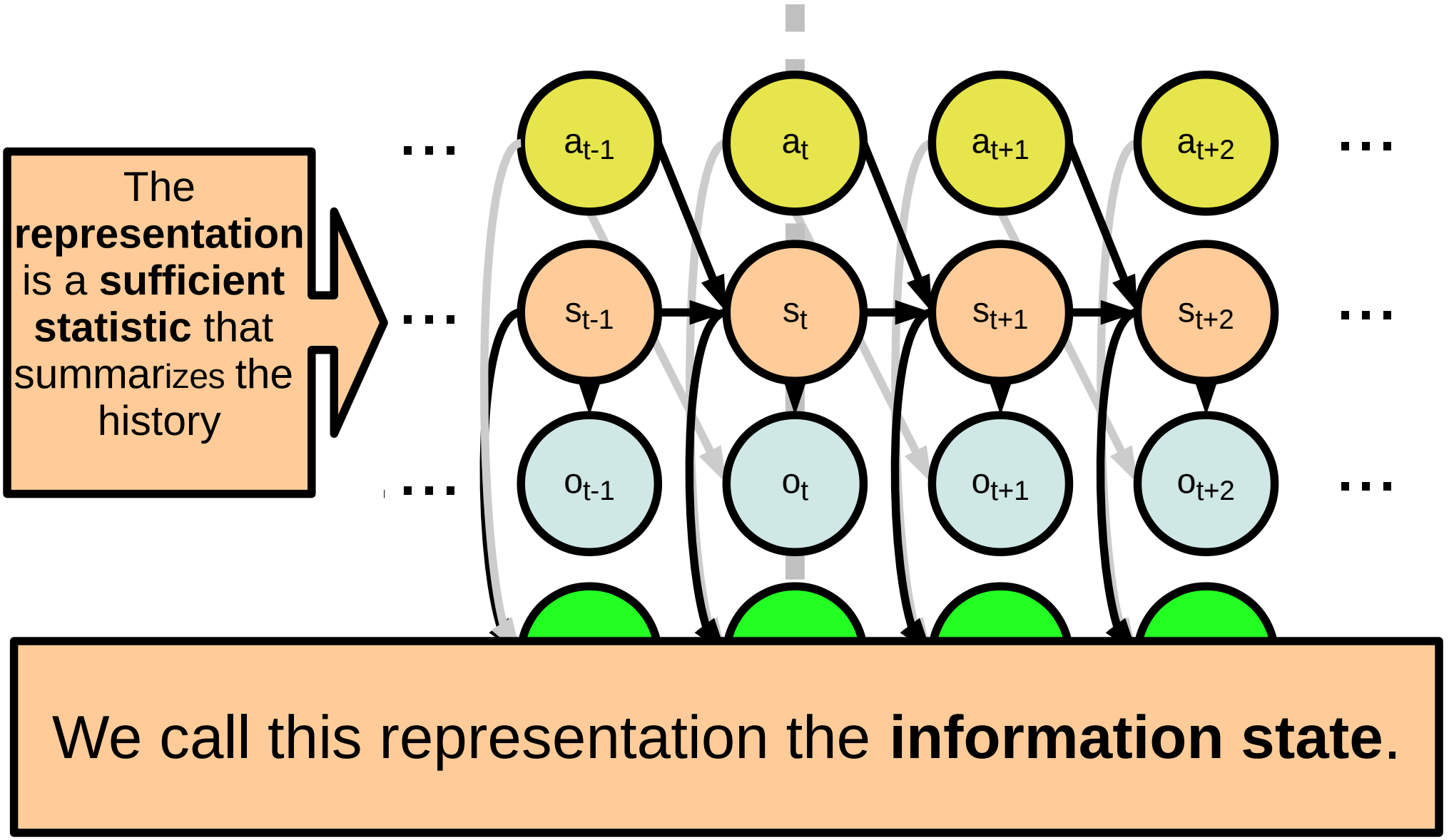


The Causal Process, Unrolled

The **representation** is a **sufficient statistic** that summarizes the history



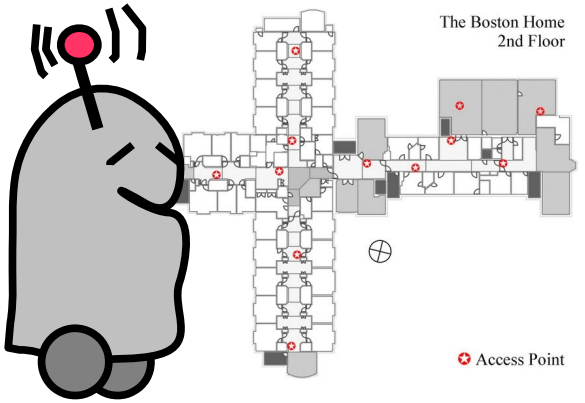
The Causal Process, Unrolled



We call this representation the **information state**.

What is **state**?

- Sometimes, there exists an obvious choice for this hidden variable (such as a robot's true position)
- At other times, learning a representation that makes the system Markovian may provide insights into the problem.



Formal POMDP definition

A POMDP consists of

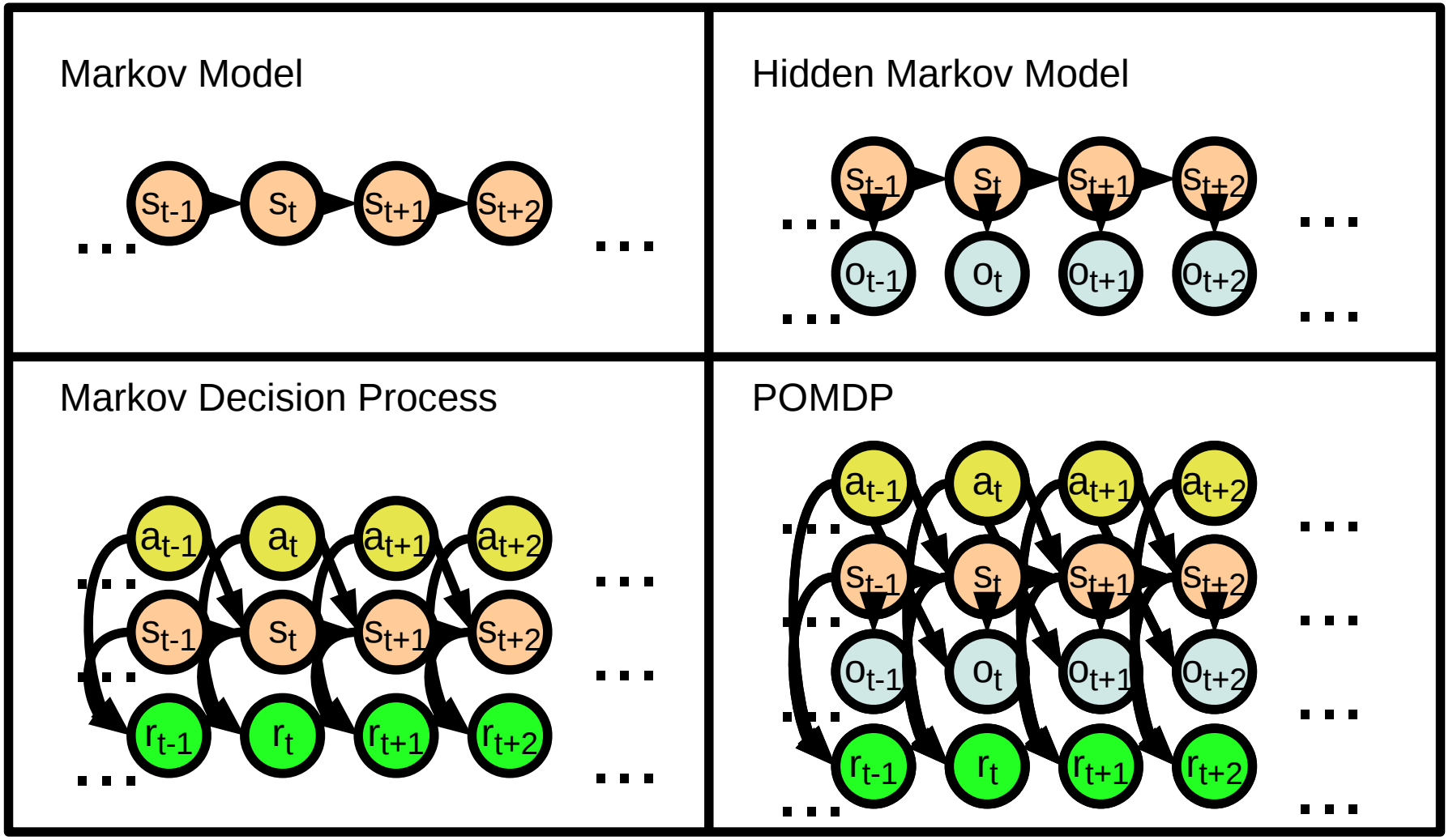
- A set of states S, actions A, and observations O
- A transition function T(s' | s , a)
- An observation function O(o | s , a)
- A reward function R(s , a)
- A discount factor γ

The goal is to maximize E[$\sum_{t=1}^{\infty} \gamma^t R_t$], the expected long-term discounted reward.

Relationship to Other Models

Hidden State?

Decisions?



Formal POMDP definition

A POMDP consists of

- A set of states S, actions A, and observations O
- A transition function T(s' | s , a)
- An observation function O(o | s , a)
- A reward function R(s , a)
- A discount factor γ

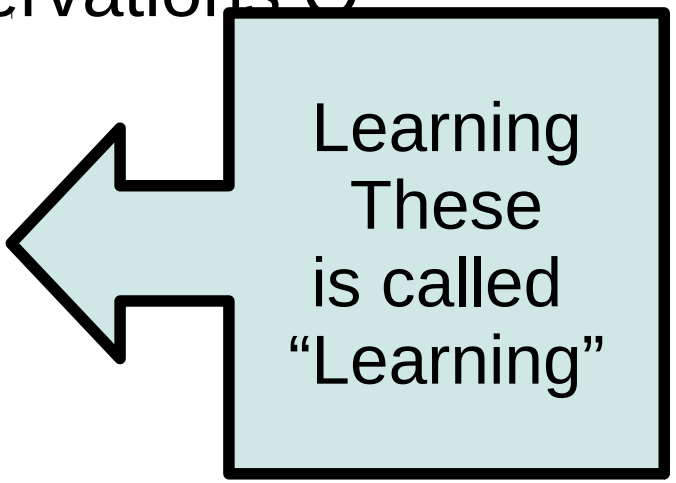
The goal is to maximize $E[\sum_{t=1}^{\infty} \gamma^t R_t]$, the term discounted reward.

This optimization is called "Planning"

Formal POMDP definition

A POMDP consists of

- A set of states S, actions A, and observations O
- A transition function **T(s' | s , a)**
- An observation function **O(o | s , a)**
- A reward function **R(s , a)**
- A discount factor γ



The goal is to maximize E[$\sum_{t=1}^{\infty} \gamma^t R_t$], the expected long-term discounted reward.

Planning

Bellman Recursion for the **value** (long-term expected reward)

$$V(b) = max\, E[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b]$$

$$. = max_a [R(b,a) + \gamma (\sum_{o \in O} P(b_{oa}|b) V(b_{oa}))]$$

State and *State*, a quick aside

- In the POMDP literature, the term “state” usually refers to the hidden state (i.e. the robot's true location).
- The posterior distribution of states s is called the “belief” b(s). It is a sufficient statistic for the history, and thus the information state for the POMDP.

Planning

Bellman Recursion for the **value** (long-term expected reward)

$$V(b) = max\, E[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b]$$

$$. = max_a [R(b,a) + \gamma (\sum_{o \in O} P(b_{oa}|b) V(b_{oa}))]$$

Planning

Bellman Recursion for the **value** (long-term expected reward)

$$V(b) = max\, E\left[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b\right]$$

$$= max_a \left[R(b,a) + \gamma\left(\sum_{o \in O} P(b_{oa}|b)\, V(b_{oa})\right)\right]$$

Belief b
(sufficient statistic/
information state)

Planning

Bellman Recursion for the **value** (long-term expected reward)

$$V(b) = max\, E[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b]$$

$$. = max_a [\underbrace{R(b,a)} + \gamma(\sum_{o \in O} P(b_{oa}|b) V(b_{oa}))]$$

Immediate reward for taking action a in belief b

Planning

Bellman Recursion for the **value** (long-term expected reward)

$$V(b) = max\, E[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b]$$

$$. = max_a [R(b,a) + \underbrace{\gamma(\sum_{o \in O} P(b_{oa}|b) V(b_{oa}))}_{\text{Expected future rewards}}]$$

Planning

Bellman Recursion for the **value** (long-term expected reward)

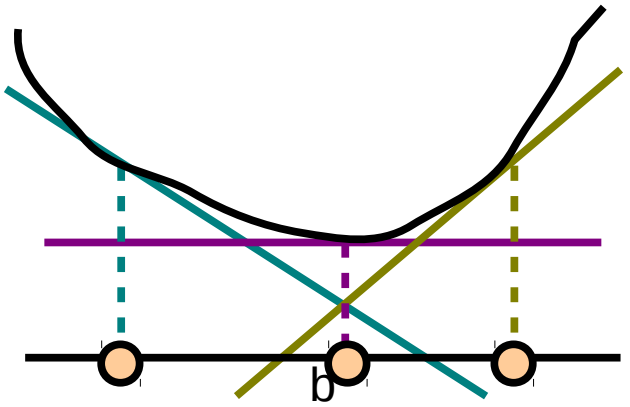
$$V(b) = max\, E[\sum_{t=1}^{\infty} \gamma^t R_t | b_0 = b]$$

$$. = max_a [R(b,a) + \gamma (\sum_{o \in O} P(b_{oa}|b) V(b_{oa}))]$$

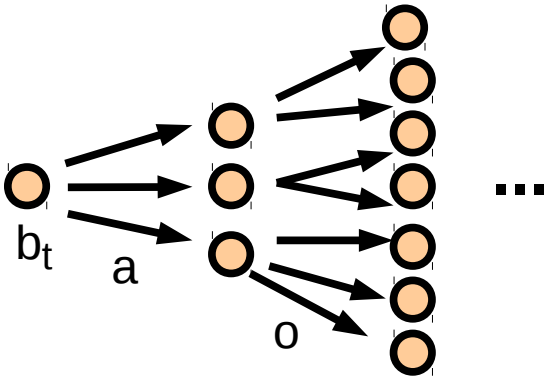
… especially when b is high-dimensional, solving for this continuous function is not easy (PSPACE-HARD)

Planning: Yes, we can!

- Global: Approximate the entire function V(b) via a set of support points b'.
  - e.g. SARSOP



- Local: Approximate a the value for a particular belief with forward simulation
  - e.g. POMCP



Learning

- Given histories

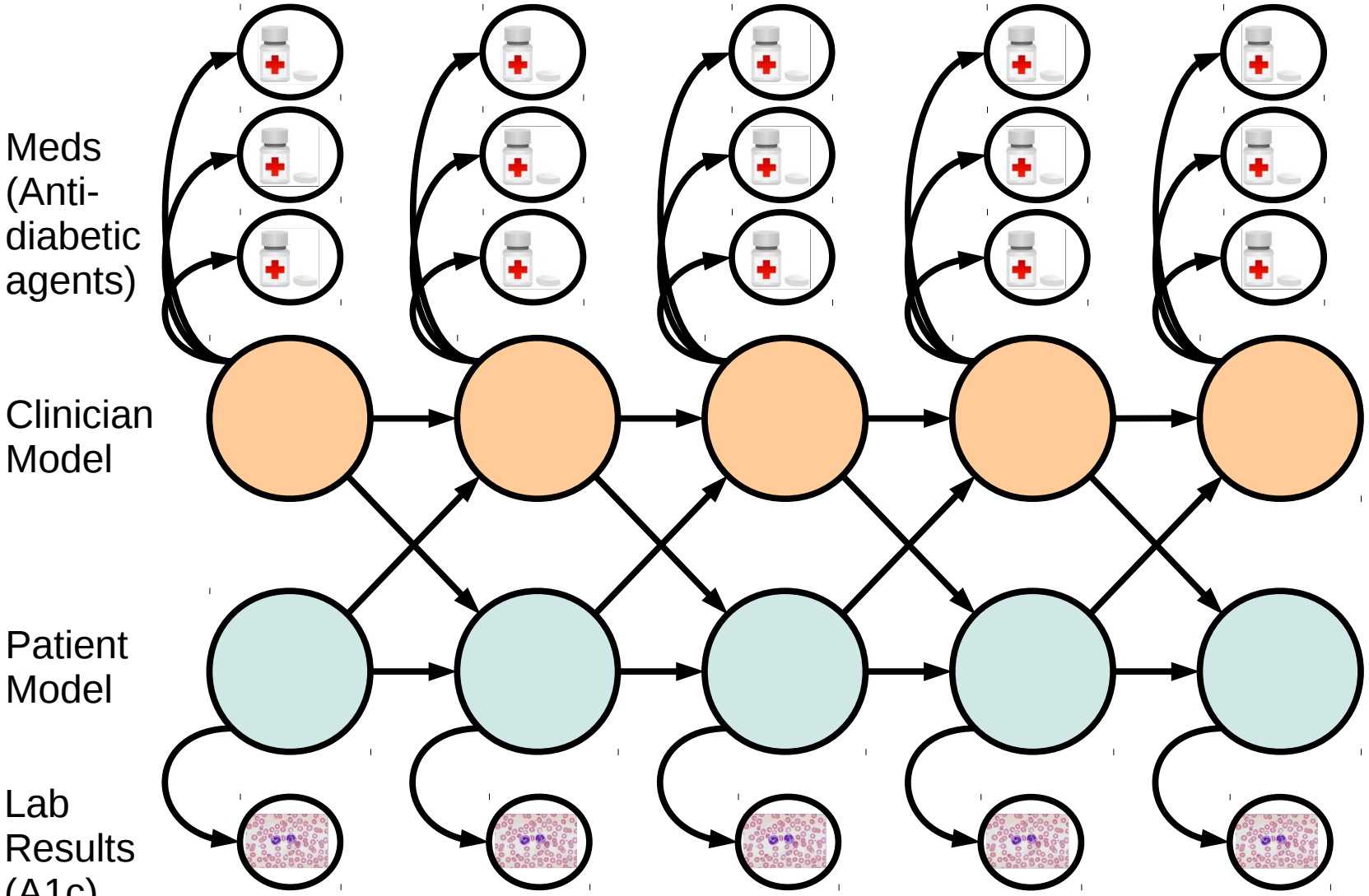
$$h=(a_1,r_1,o_1,a_2,r_2,o_2,\ldots,a_T,r_T,o_T)$$

  we can learn T, O, R via forward-filtering/backward-sampling or <fill in your favorite timeseries algorithm>

- Two principles usually suffice for exploring to learn:
  - Optimism under uncertainty: Try actions that might be good
  - Risk control: If an action seems risky, ask for help.

Example: Timeseries in Diabetes

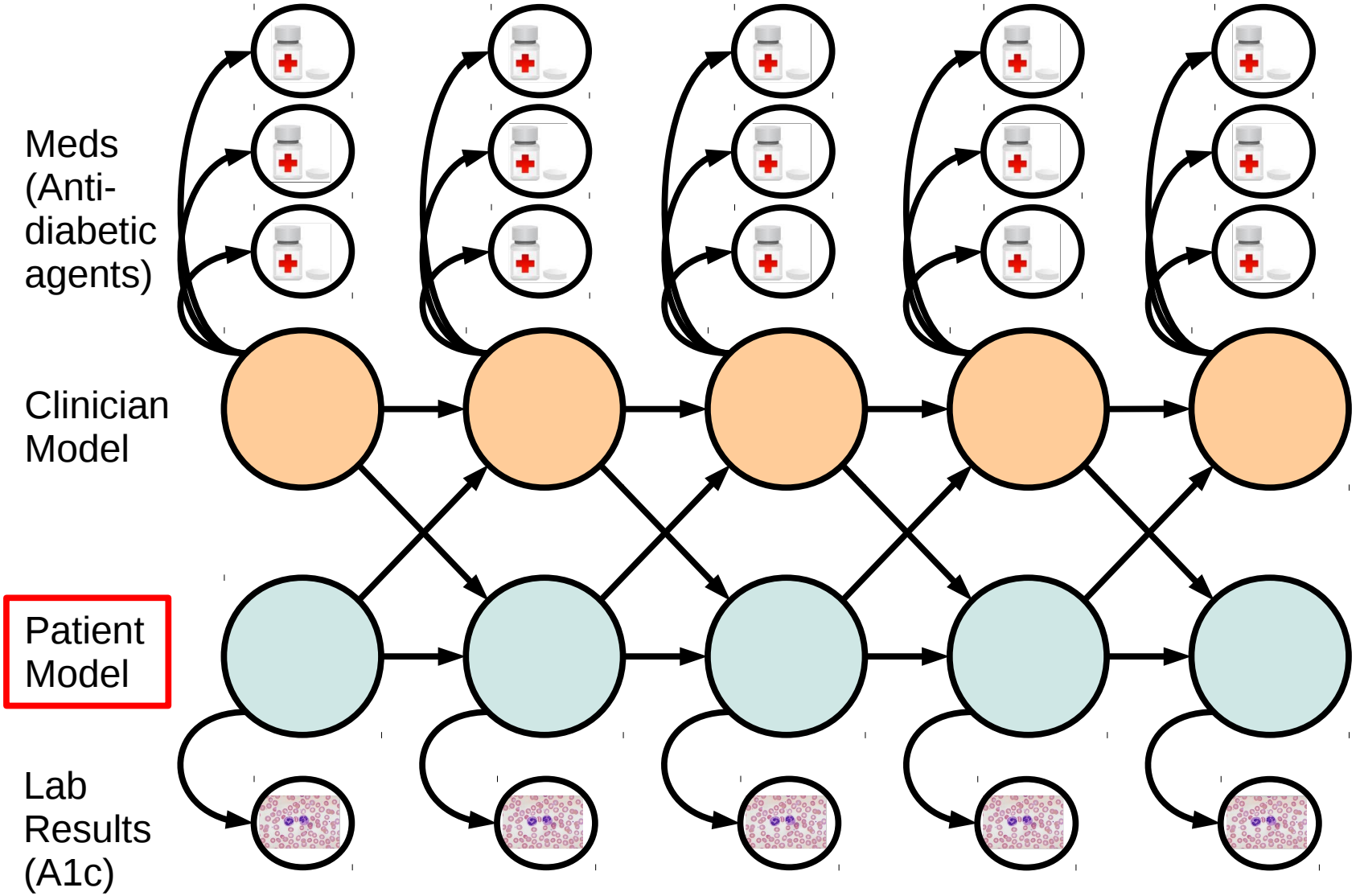
Data: Electronic health records of ~17,000 diabetics with 5+ A1c lab measurements and 5+ anti-diabetic agents prescribed.



Collaborators: Isaac Kohane, Stan Shaw

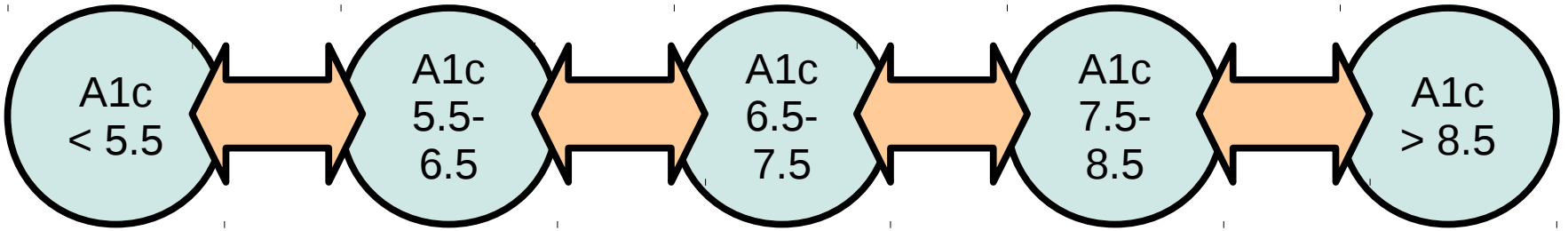
Example: Timeseries in Diabetes

Data: Electronic health records of ~17,000 diabetics with 5+ A1c lab measurements and 5+ anti-diabetic agents prescribed.



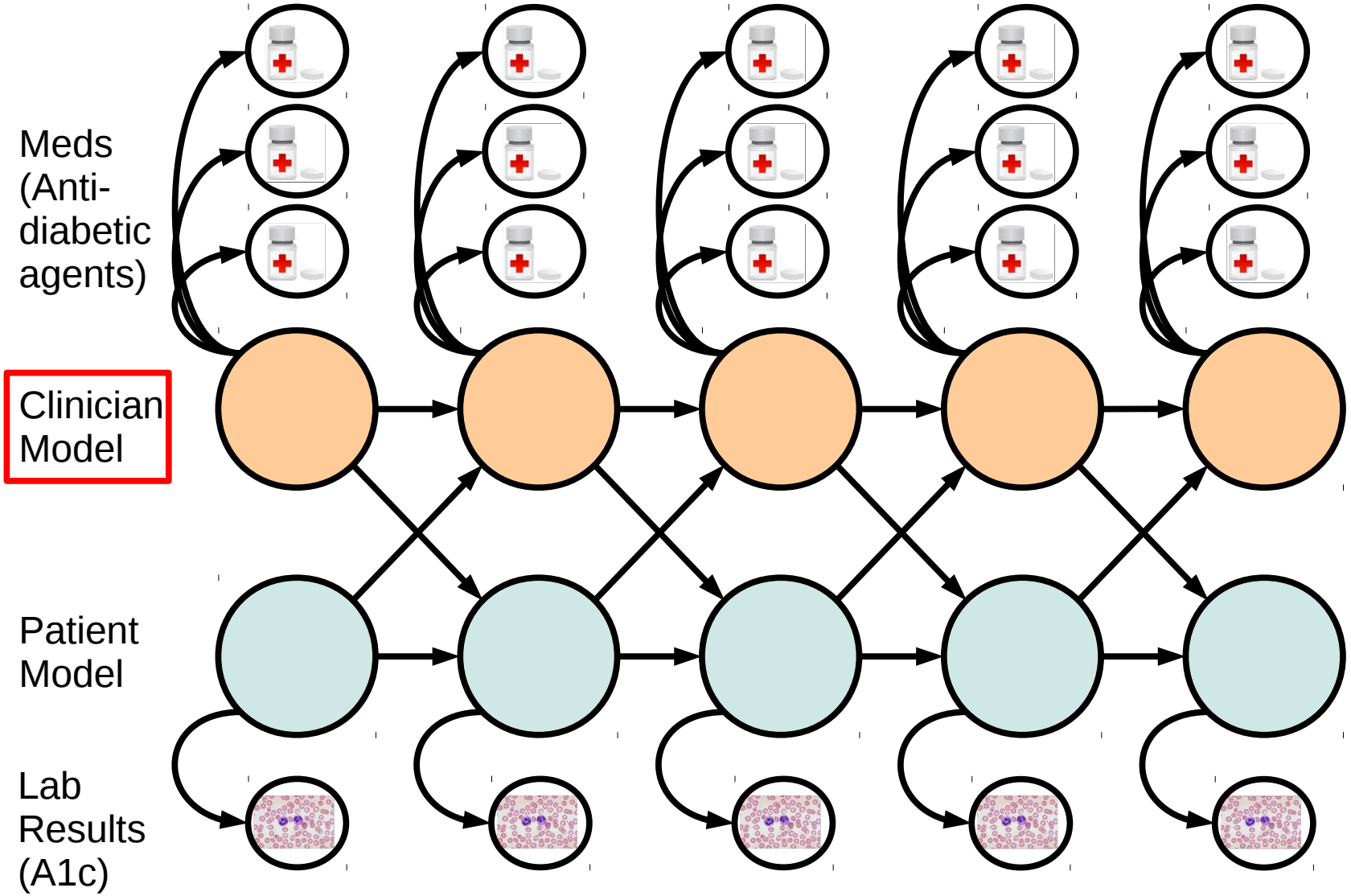
Discovered Patient States

- The “patient states” each correspond to a set of A1c levels (unsurprising)



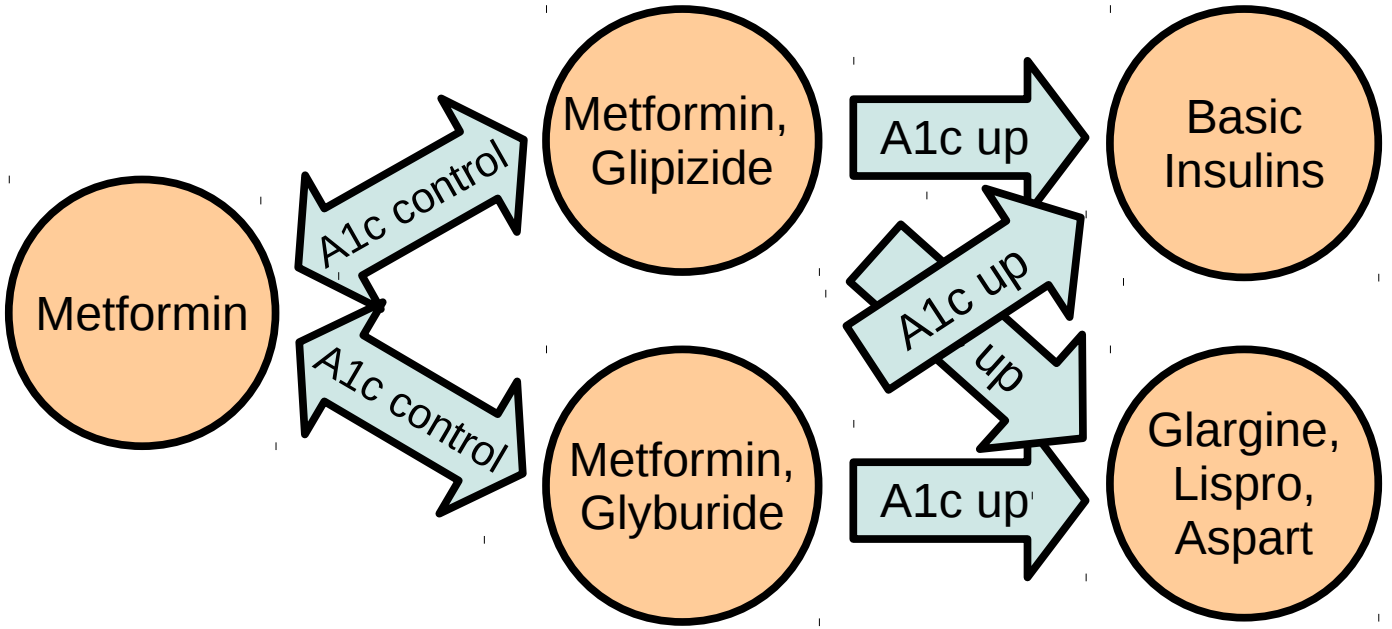
Example: Timeseries in Diabetes

Data: Electronic health records of ~17,000 diabetics with 5+ A1c lab measurements and 5+ anti-diabetic agents prescribed.



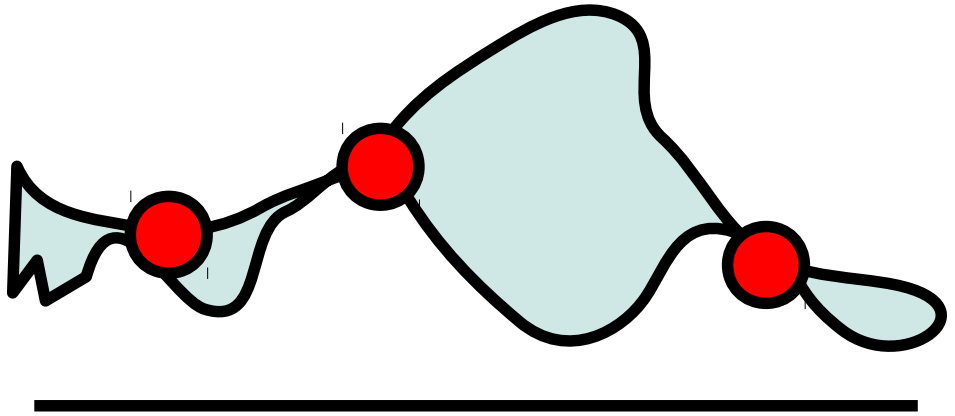
Discovered Clinician States

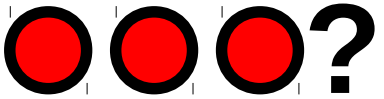
- The “clinician states” follow the standard treatment protocols for diabetes (unsurprising, but exciting that we discovered this is a completely unsupervised manner)
- Next steps: Incorporate more variables; identify patient and clinician outliers (quality of care)



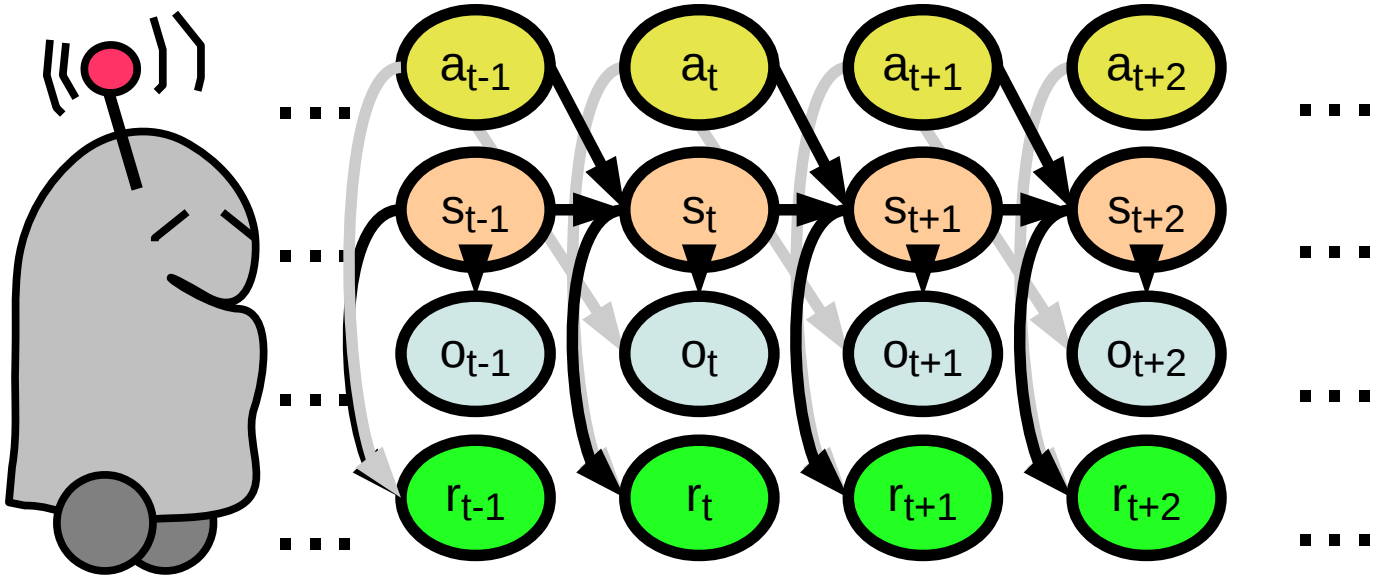
Example: Experimental Design

- In a very general sense:
  - Action space: all possible experiments + “submit”
  - State space: which hypothesis is true
  - Observation space: results of experiments
  - Reward: cost of experiment
- Allows for non-myopic sequencing of experiments.
- Example: Bayesian Optimization





Joint with: Ryan Adams/HIPS group

Summary

- POMDPs provide a framework for
  - modeling causal dynamical systems
  - making optimal sequential decisions
- POMDPs can be learned and solved!