Sequential Medical Decision-Making with RL

August 2, 2019

MMCi Applied Data Science
Block 5, Lecture 2

Matthew Engelhard

Sepsis Management and Artificial Pancreas

TWO ILLUSTRATIVE EXAMPLES

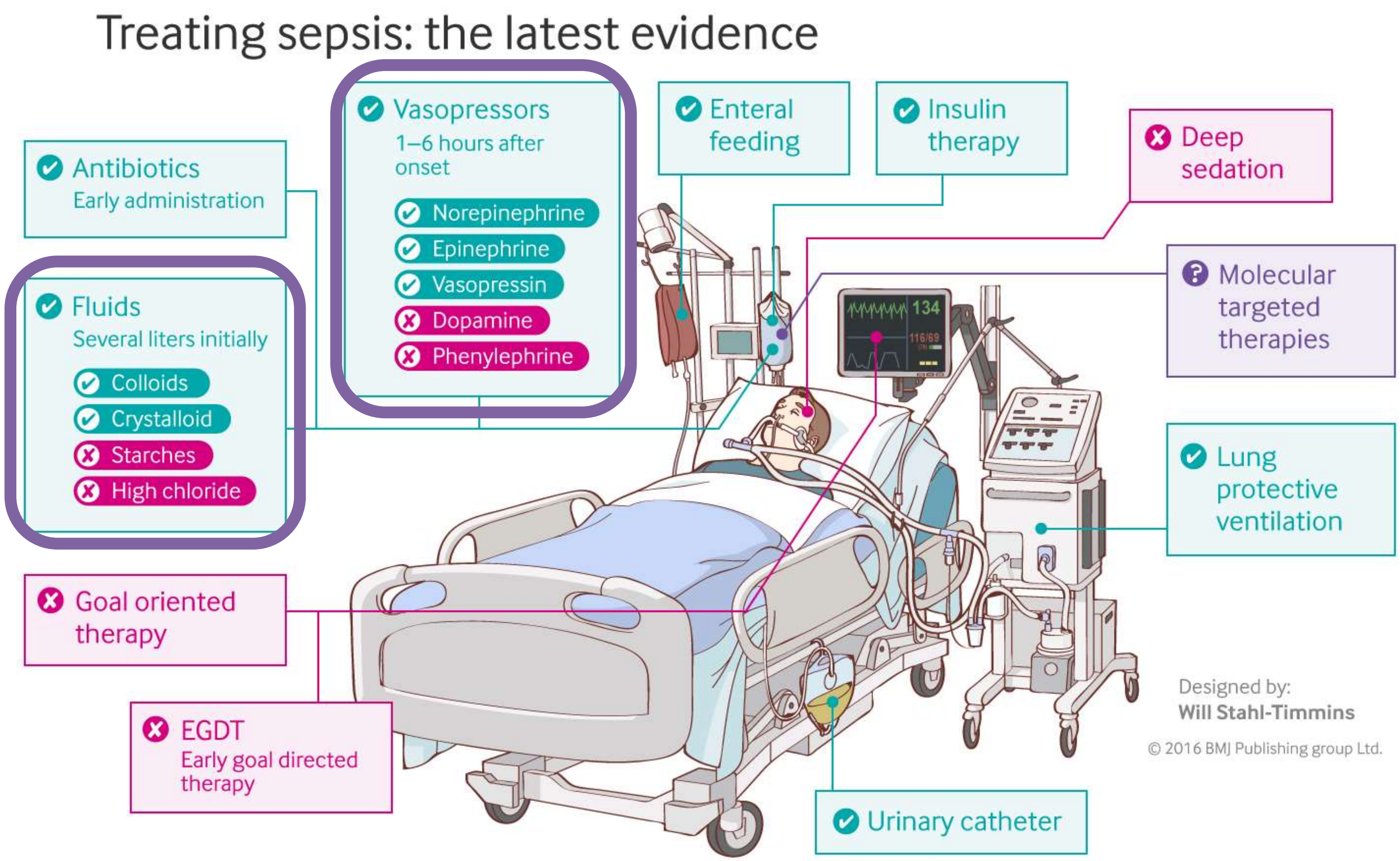
Sequential Medical Decision-Making

An agent

takes actions

based on the state of
a system

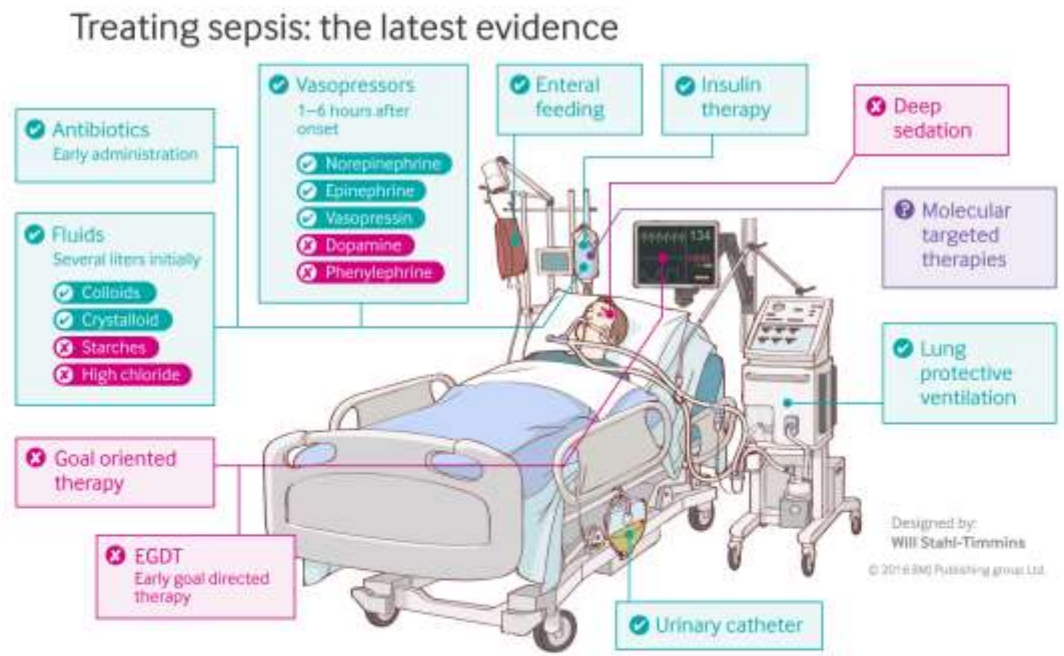
to maximize reward



“Uncertainties still exist regarding the optimal type of fluid, the optimal volume, and the best way to monitor the response to therapy.”

Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. bmj. 2016 May 23;353(i1585).

Sequential Medical Decision-Making: Sepsis Management



An agent	A clinician
takes actions	gives fluid and/or vasopressor
based on the state of a system	based on the patient's physiologic status
to maximize reward	to maximize chance of survival

Deep Reinforcement Learning for Sepsis Treatment

Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M.

arXiv:1711.09602. 2017 Nov 27

- Policy via Deep Q-Learning
- 17,898 patients from MIMIC-III

MENU

nature medicine

Article | Published: 22 October 2018

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon & A. Aldo Faisal

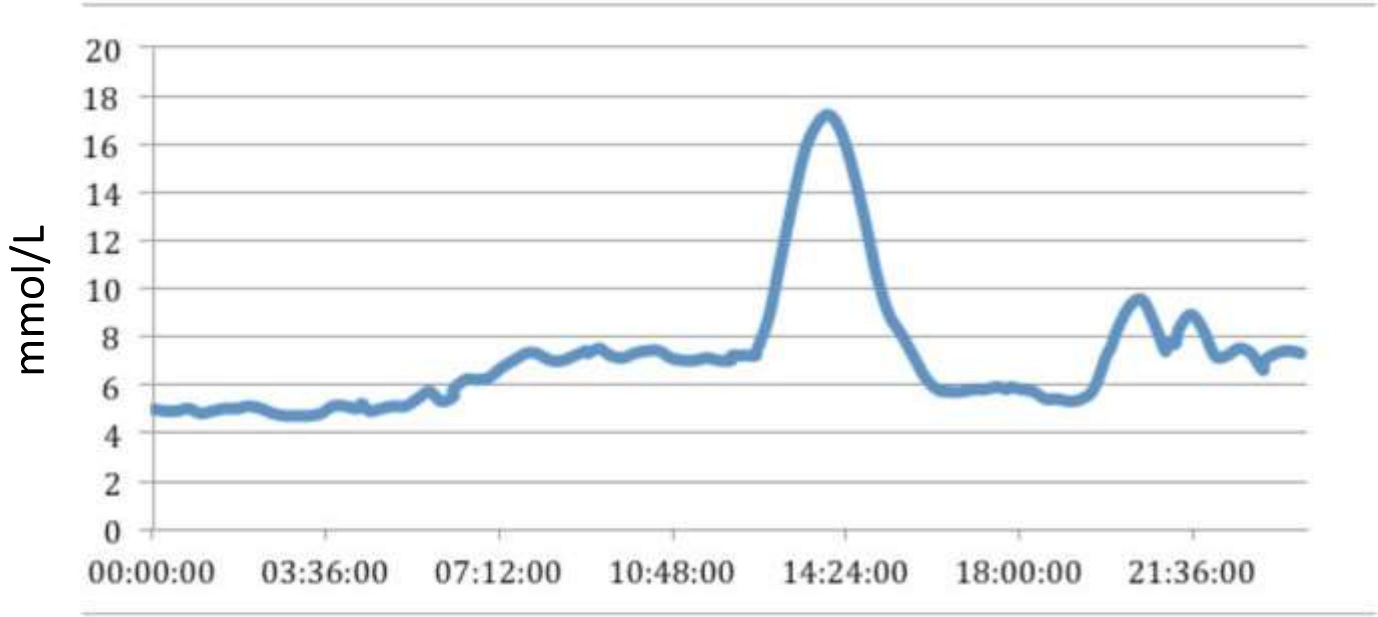
Nature Medicine **24**, 1716–1720 (2018) | Download Citation

Independent validation on the Philips eICU Research Institute Database:

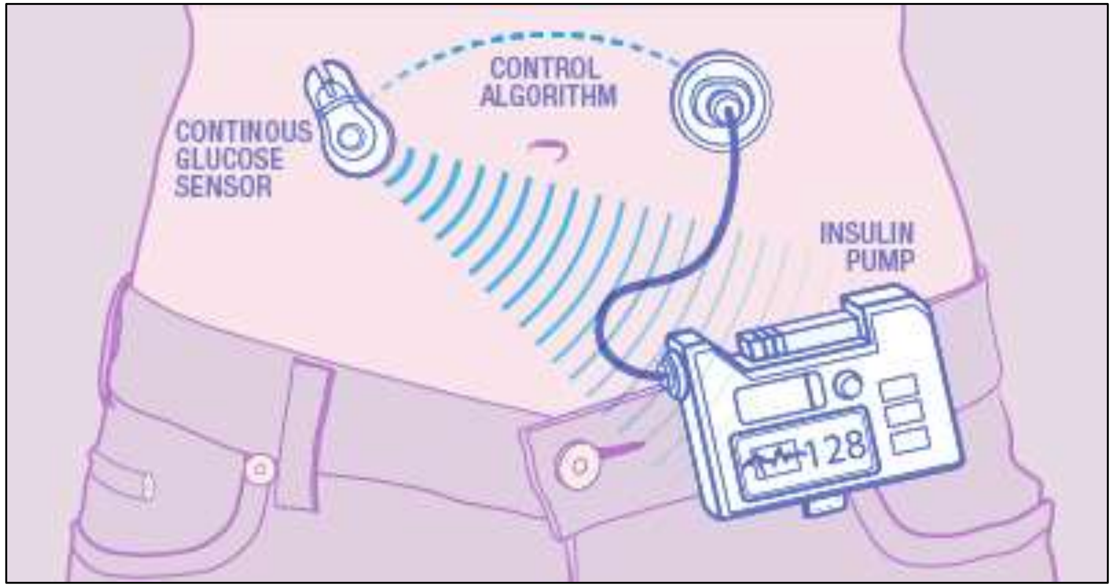
>3.3 million admissions from 2003–2016 in 459 ICUs across the US

Closed-Loop Blood Glucose Control (artificial pancreas)

Blood Glucose Readings from Continuous Glucose Monitor



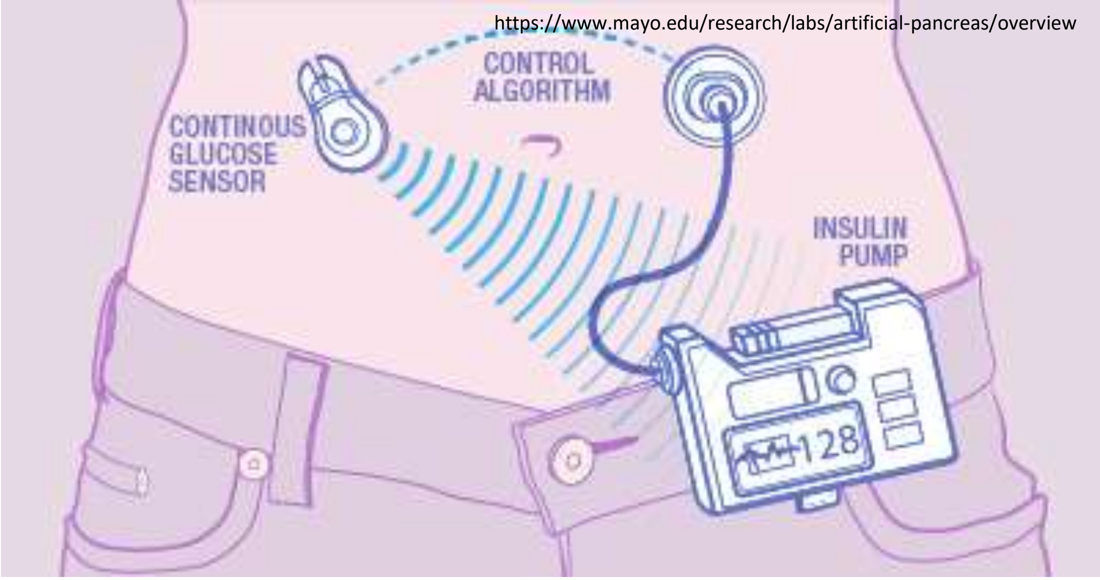
https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e



Sequential Medical Decision-Making:

Artificial Pancreas

https://www.mayo.edu/research/labs/artificial-pancreas/overview



An agent	A computer program
takes actions	administers insulin
based on the state of a system	based on blood glucose and recent patient behaviors
to maximize reward	to maintain normoglycemia

PLOS | ONE

PUBLISH ABOUT BROWSE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Model-Free Machine Learning in Biomedicine: Feasibility Study in Type 1 Diabetes

Elena Daskalaki, Peter Diem, Stavroula G. Mougiakakou

Published: July 21, 2016 • https://doi.org/10.1371/journal.pone.0158722

FORMULATING THE RL PROBLEM

The RL Paradigm

- A set of states: $S = \{s^1, s^2, \dots, s^n\}$
- A set of possible actions: $A = \{a^1, a^2, \dots, a^m\}$
- A reward $r(s, a, s')$ for reaching state s' from state s after taking action a

GOAL:
Learn a <u>policy</u> $\pi: S \rightarrow A$
that assigns each state s to the action a that **maximizes expected reward over time**

$$\dots \quad s_{t-1} \quad a_{t-1} \quad r_{t-1} \quad s_t \quad a_t \quad r_t \quad s_{t+1} \quad \dots$$

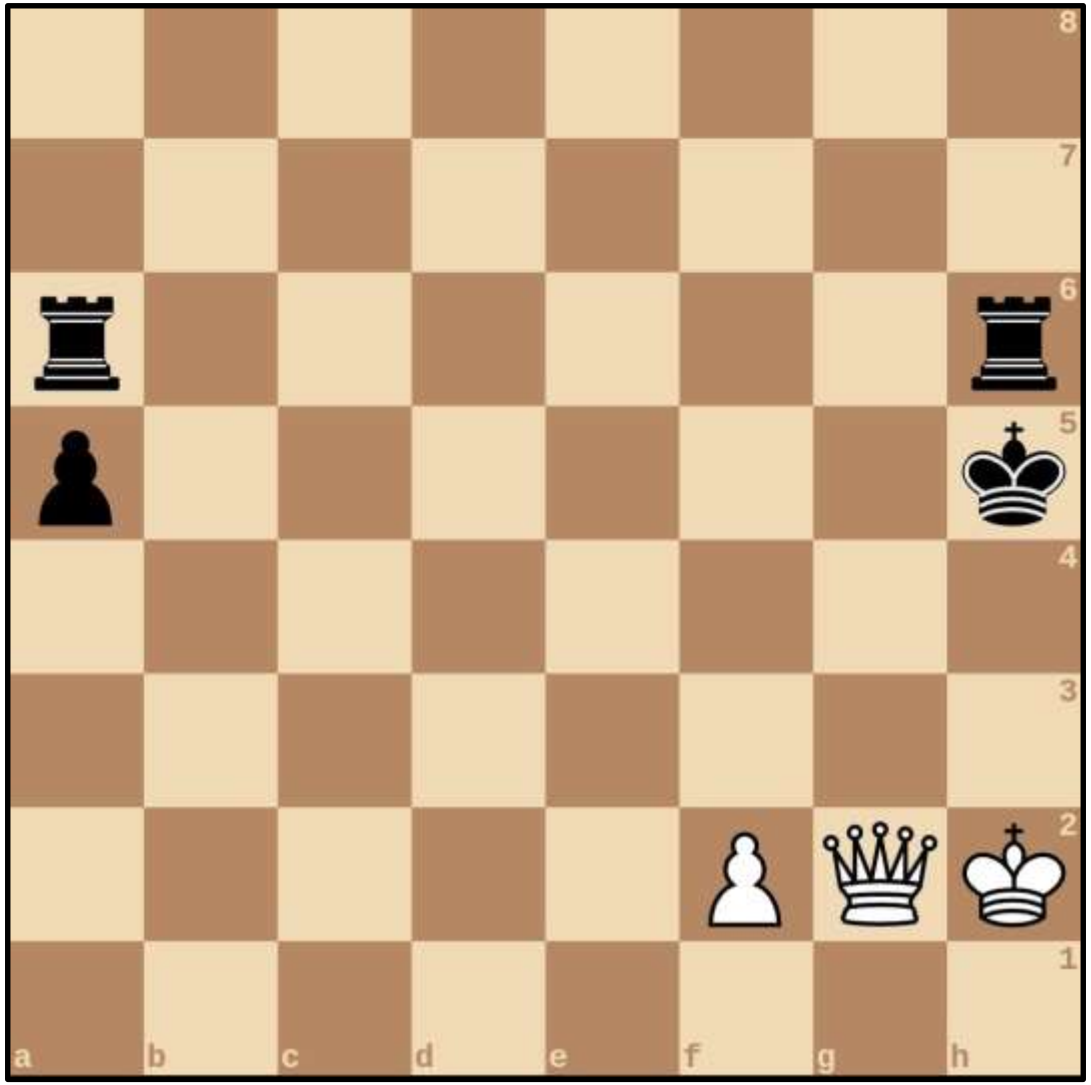
The RL Paradigm

- **A set of states: $\boldsymbol{S} = \{\boldsymbol{s}^1, \boldsymbol{s}^2, \dots, \boldsymbol{s}^n\}$**
- A set of possible actions: $A = \{a^1, a^2, \dots, a^m\}$
- A reward $r(s, a, s')$ for reaching state *s'* from state *s* after taking action *a*

GOAL:
Learn a policy $\pi: S \rightarrow A$
that assigns each state *s* to the action *a* that **maximizes expected reward over time**

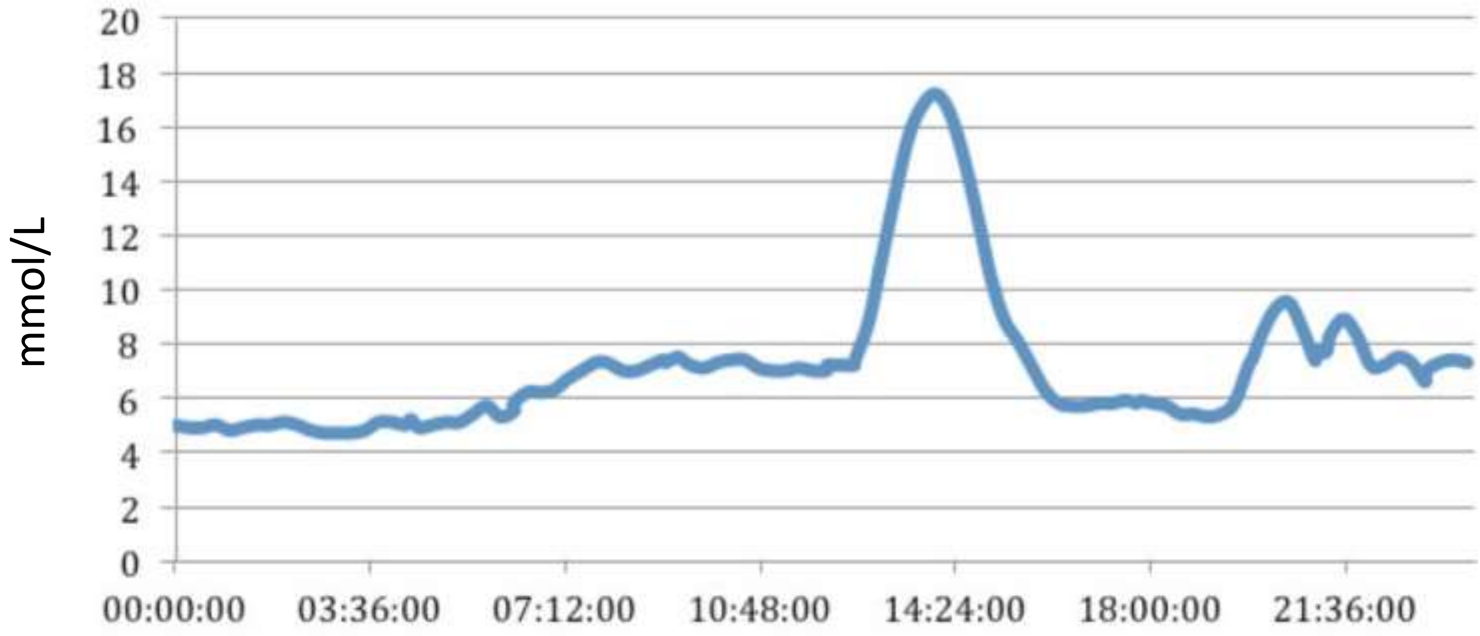
… s_{t-1} a_{t-1} r_{t-1} s_t a_t r_t s_{t+1} …

Chess or Go: the state is what you see on the board



State s_t: Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



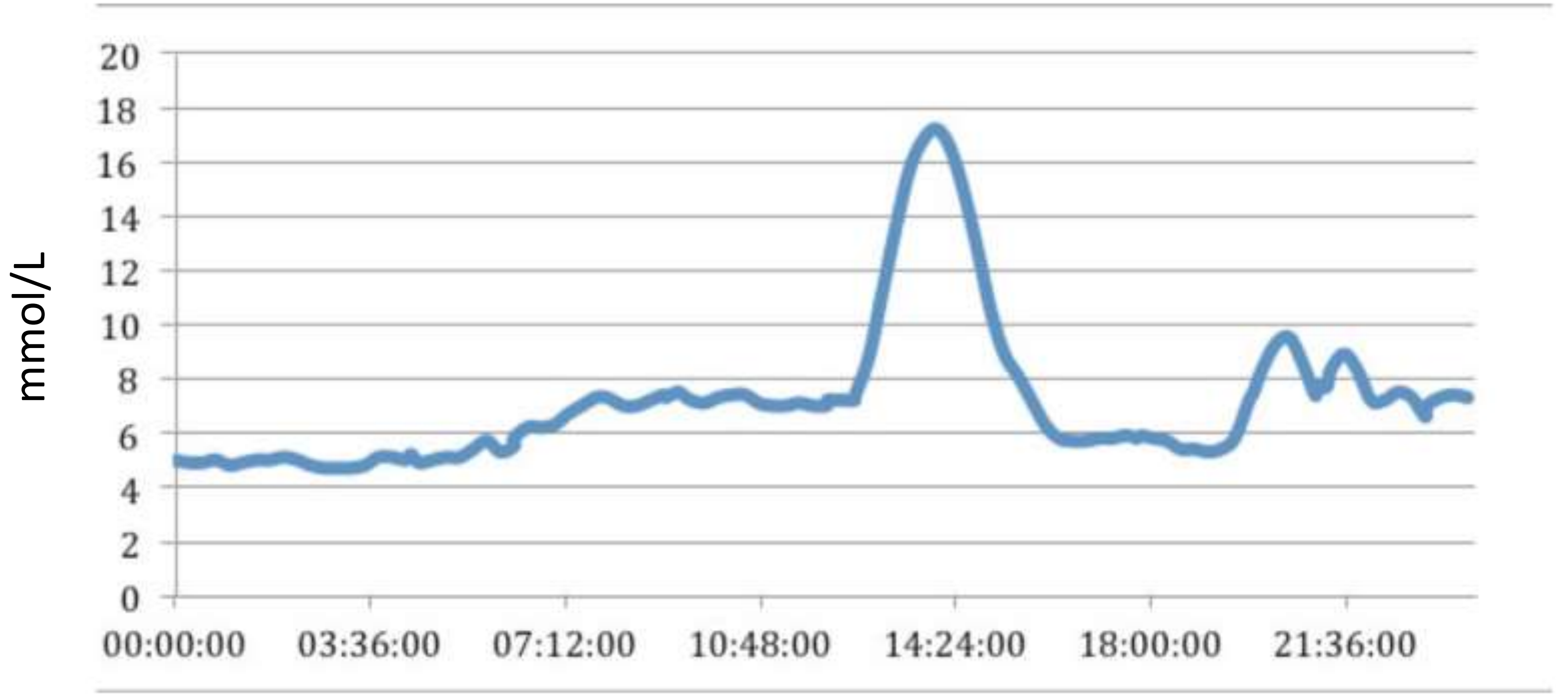
https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e

<u>Idea 1:</u>

The state is the current blood glucose value

State s_t: Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



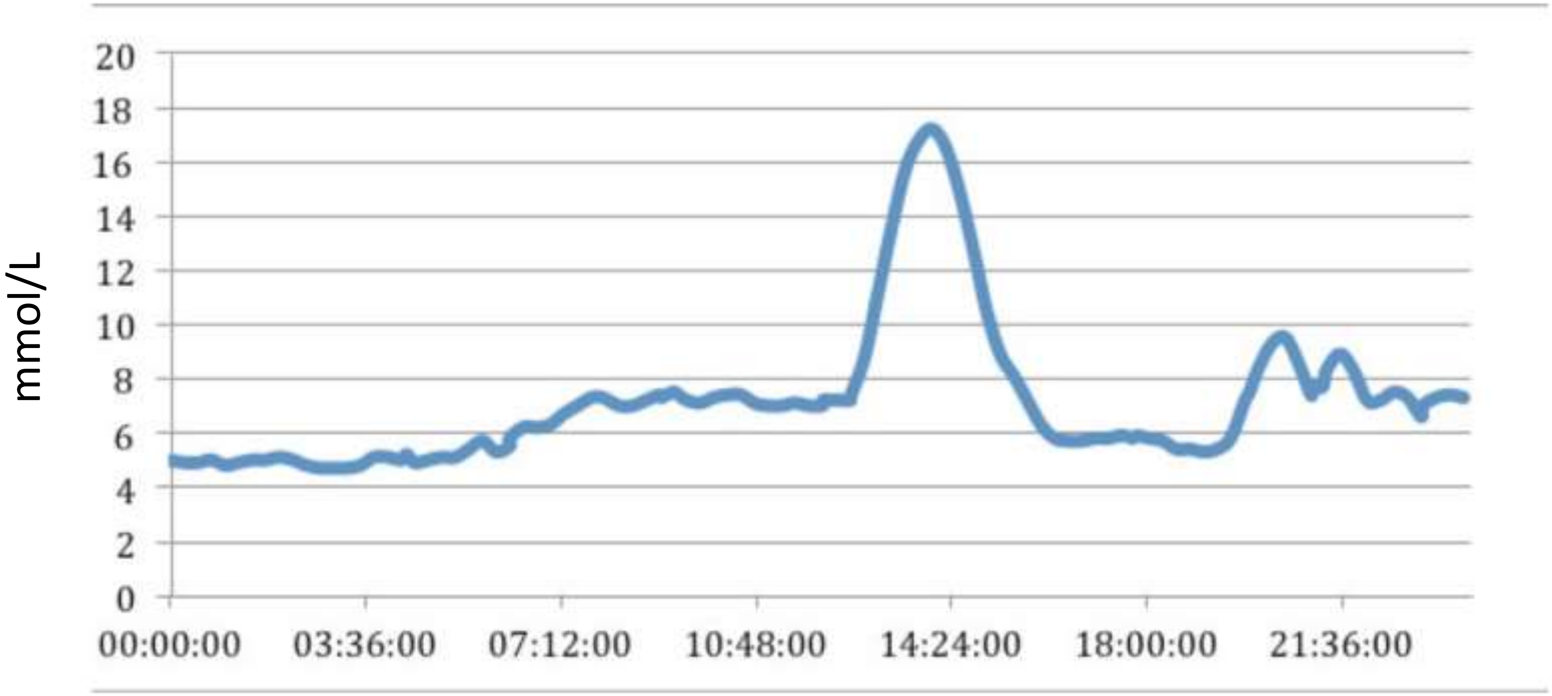
https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e

<u>Idea 2:</u>

The state is the current blood glucose value plus recent trends

State s_t: Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



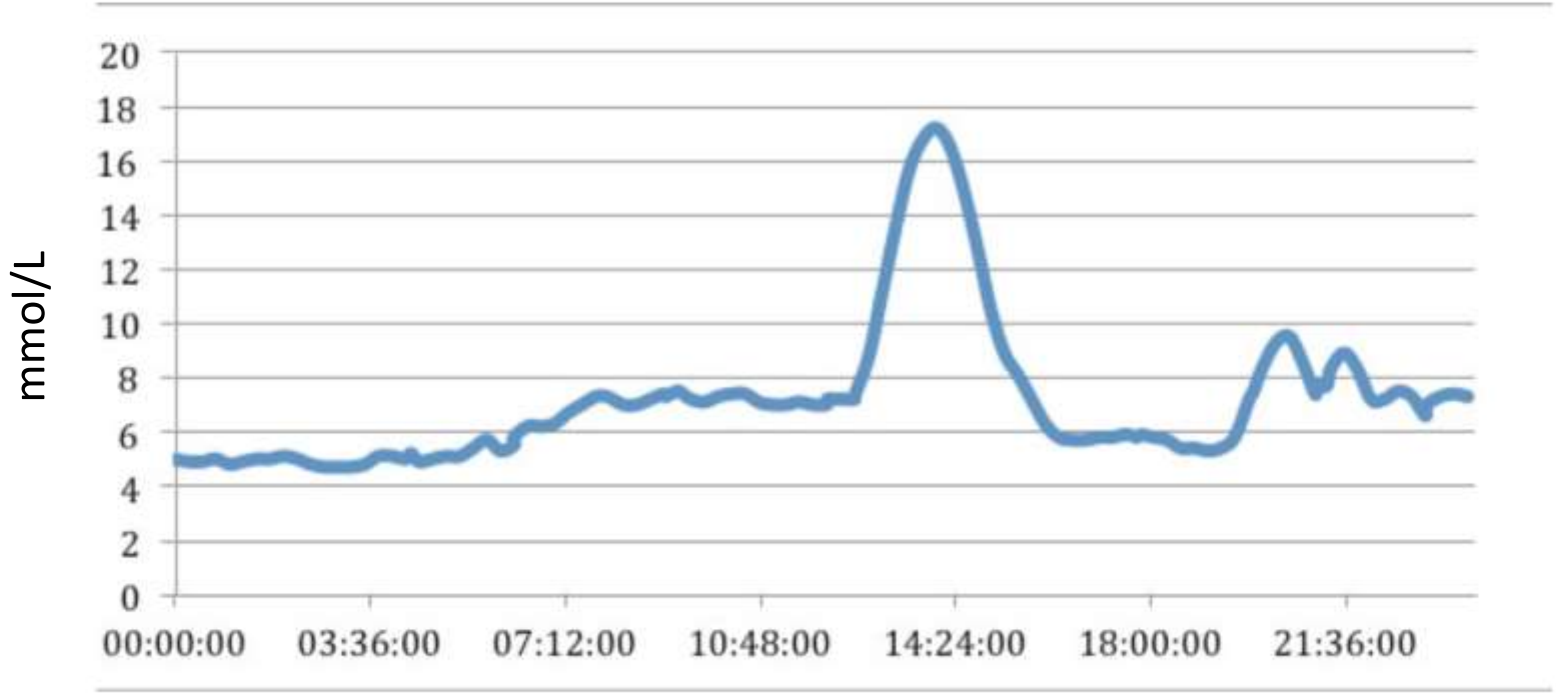
https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e

Idea 3:

The state is the current blood glucose value, recent trends, and the patient's insulin sensitivity

State s_t: Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



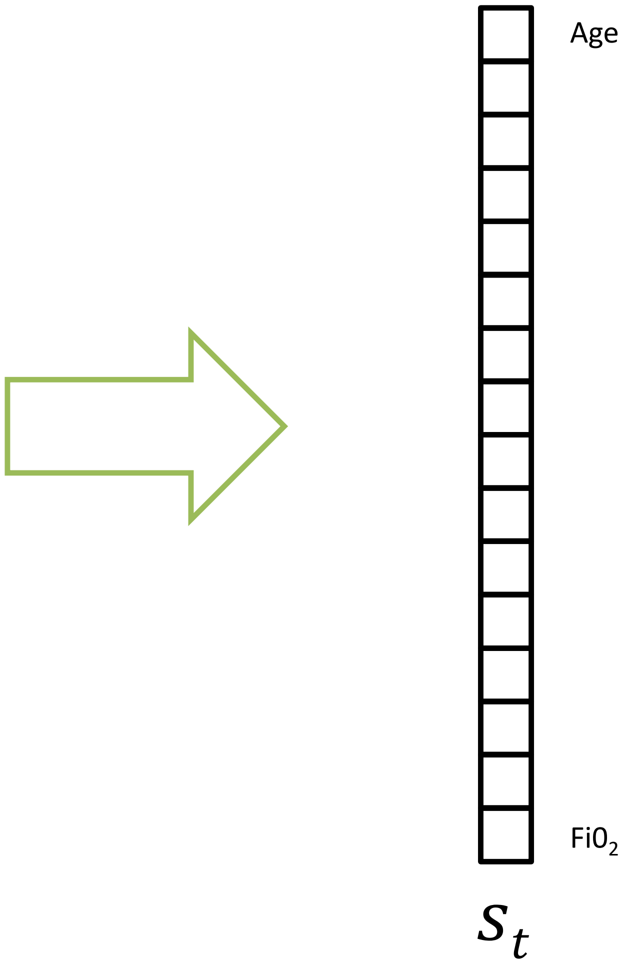
https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e

<u>The state: all information relevant to our decision</u>

- BG trends
- Previous insulin doses
- Patient physiology
- Recent behaviors (e.g. eating, physical activity)

State s_t: Sepsis

Category	Items	Type
Demographics	Age	Continuous
	Gender	Binary
	Weight	Continuous
	Readmission to intensive care	Binary
	Elixhauser score (premorbid status)	Continuous
Vital signs	Modified SOFA*	Continuous
	SIRS	Continuous
	Glasgow coma scale	Continuous
	Heart rate, systolic, mean and diastolic blood pressure, shock index	Continuous
	Respiratory rate, SpO2	Continuous
	Temperature	Continuous
Lab values	Potassium, sodium, chloride	Continuous
	Glucose, BUN, creatinine	Continuous
	Magnesium, calcium, ionized calcium,	Continuous
	carbon dioxide	Continuous
	SGOT, SGPT, total bilirubin, albumin	Continuous
	Hemoglobin	Continuous
	White blood cells count, platelets count, PTT, PT, INR, pH, Pa02, PaCO2, base excess, bicarbonate, lactate, PaO2/FiO2 ratio	Continuous
Ventilation parameters	Mechanical ventilation	Binary
	FiO2	Continuous
Medications and fluid balance	Current IV fluid intake over 4h	Continuous
	Maximum dose of vasopressor over 4h	Continuous
	Urine output over 4h	Continuous
	Cumulated fluid balance since admission (includes preadmission data when available)	Continuous



- Static features
- Time-varying features

The RL Paradigm

- A set of states: $S = \{s^1, s^2, \dots, s^n\}$
- **A set of possible actions: $\boldsymbol{A = \{a^1, a^2, \dots, a^m\}}$**
- A reward $r(s, a, s')$ for reaching state *s'* from state *s* after taking action *a*

GOAL:
Learn a policy $\pi: S \rightarrow A$
that assigns each state *s* to the action *a* that
maximizes expected reward over time

$$\dots \quad s_{t-1} \quad a_{t-1} \quad r_{t-1} \quad s_t \quad a_t \quad r_t \quad s_{t+1} \quad \dots$$

Actions a_t : Sepsis Management

How much fluid?

Vasopressor dose?

Actions a_t : Artificial Pancreas

Insulin Basal Rate

$\times$

Insulin Bolus Amount

https://time.com/4703099/continuous-glucose-monitor-blood-sugar-diabetes/

Actions a_t : Artificial Pancreas

Insulin Basal Rate

$\times$

Insulin Bolus Amount

$\times$

Glucagon Bolus Amount

Must be discretized

https://time.com/4703099/continuous-glucose-monitor-blood-sugar-diabetes/

We need a finite set of actions to choose from

a^1					
a^2			$Q(s^3, a^2)$		
a^3					
…					
a^m					
	s^1	s^2	s^3	…	s^n

Q Learning:

- Q function implemented as a look-up table
- <u>Input</u>: state s
- <u>Output</u>: $Q(s, a)$ for each action $a \in A$

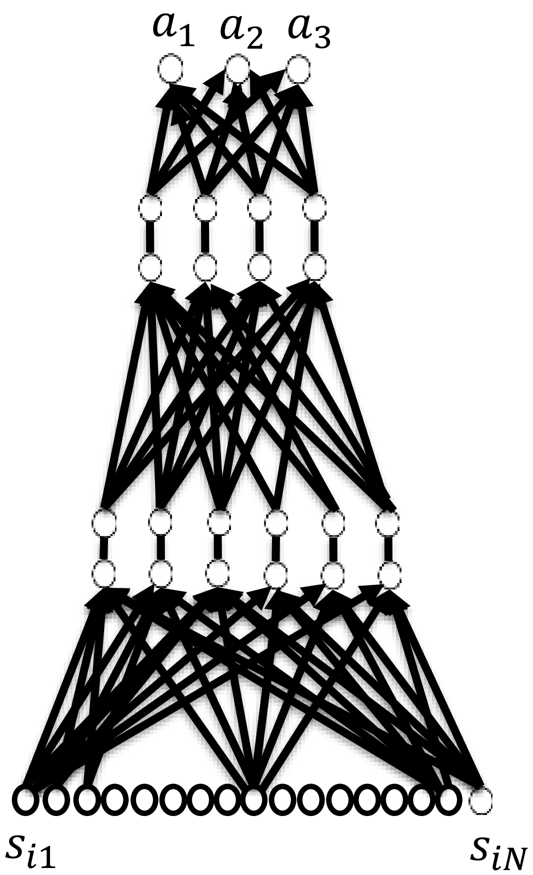
We need a finite set of actions to choose from

	s^1	s^2	s^3	…	s^n
a^1					
a^2			$Q(s^3, a^2)$		
a^3					
…					
a^m					

Q Learning:

- Q function implemented as a look-up table
- <u>Input</u>: state s
- <u>Output</u>: $Q(s, a)$ for each action $a \in A$

We need a finite set of actions to choose from



Deep Q Learning:

- Q function implemented as a deep neural network
- Input: state s
- Output: $Q(s, a)$ for each action $a \in A$

Tradeoff: more actions means greater flexibility, but also makes the problem more complex

Actions a_t: Sepsis Management

Five IV fluid quantities:

{0, Quartile1, Q2, Q3, Q4}

$\times$

Five vasopressor doses

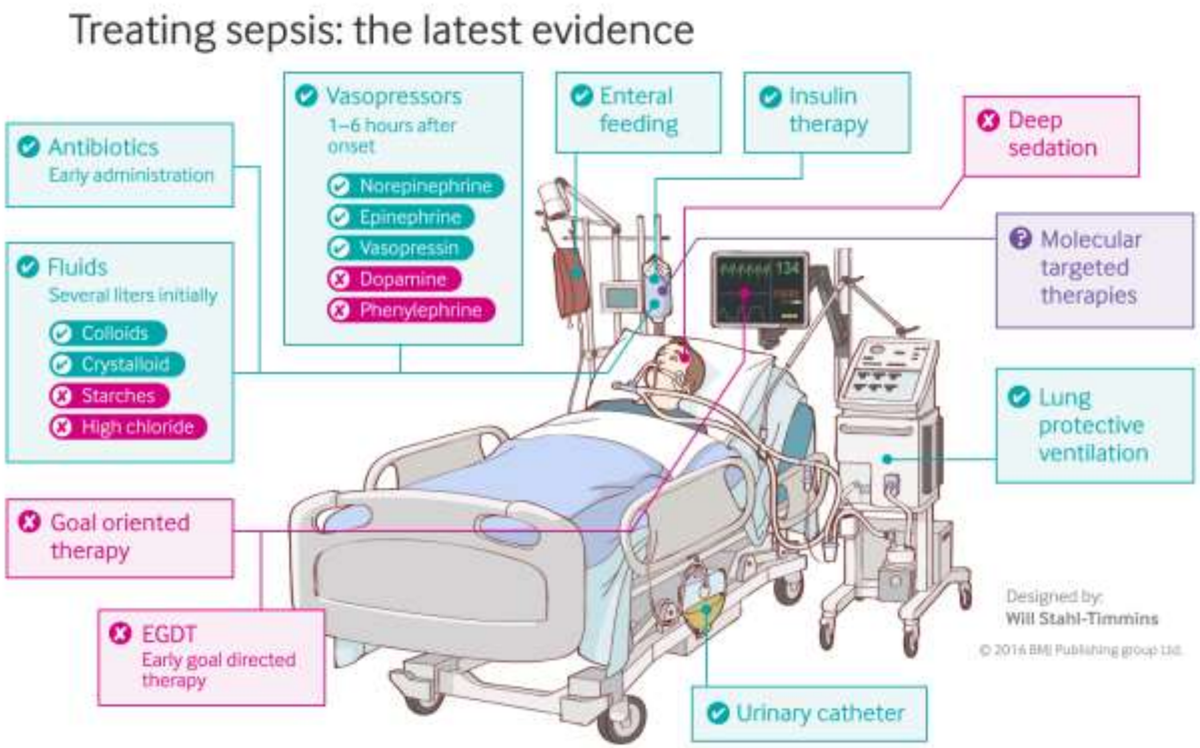
{0, Quartile1, Q2, Q3, Q4}

- Quartiles are defined based on the training dataset
- Example action:
  (Q2 fluid, 0 vasopressor)
- 25 possible actions

The RL Paradigm

- A set of states: $S = \{s^1, s^2, \ldots, s^n\}$
- A set of possible actions: $A = \{a^1, a^2, \ldots, a^m\}$
- **A reward $\boldsymbol{r(s, a, s')}$ for reaching state *s'* from state *s* after taking action *a***
- (sometimes) A model $P(s, a, s')$ that describes the probability of reaching state *s'* from state *s* after taking action *a*

$$\ldots \quad s_{t-1} \quad a_{t-1} \quad r_{t-1} \quad s_t \quad a_t \quad r_t \quad s_{t+1} \quad \ldots$$



“Uncertainties still exist regarding the optimal type of fluid, the optimal volume, **and the best way to monitor the response to therapy**.”

Clinician goals: keep the patient stable.

- central venous pressure (8-12 mm Hg)
- mean arterial pressure (65-90 mm Hg)
- urine output (0.5 mL/kg/h)
- central venous oxygen saturation (70%)

RL goals: optimize the outcome

- prevent death
- prevent organ damage

-> The RL algorithm chooses actions that maximize expected reward over time

Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. bmj. 2016 May 23;353(i1585).

Primary objective: prevent mortality

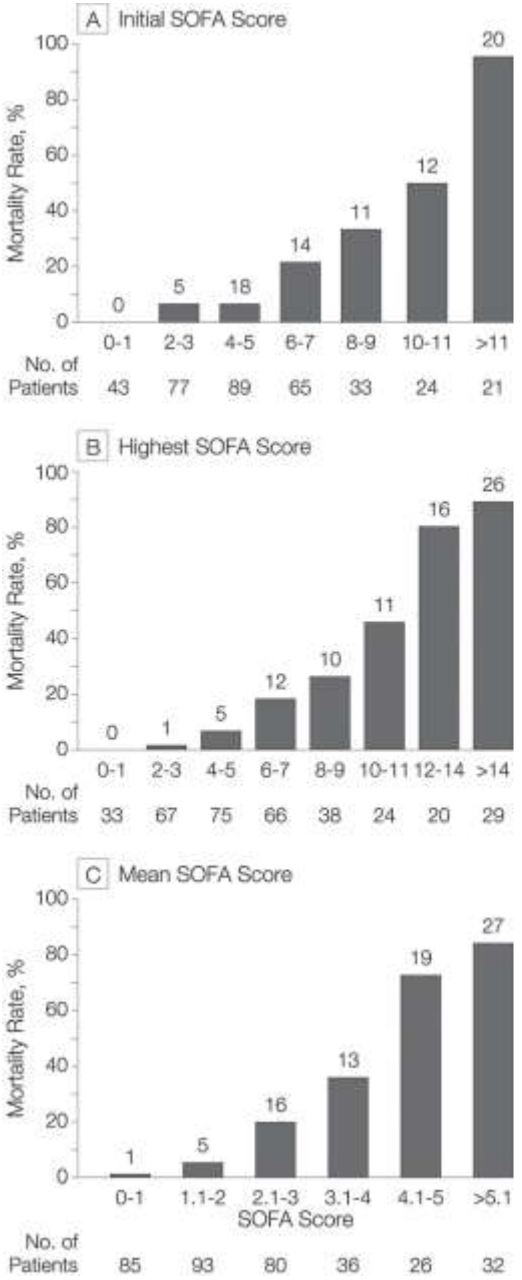
- Should this be the only goal?
- We want to optimize patient outcomes, but follow-up data is not available

Additional Goal: Prevent Organ Dysfunction -> SOFA Score (0-24)

Table 1. The Sequential Organ Failure Assessment (SOFA) Score*

Variables	SOFA Score 0	1	2	3	4
Respiratory Pao_2/Fio_2, mm Hg	>400	≤400	≤300	≤200†	≤100†
Coagulation Platelets $\times 10^3/\mu L$‡	>150	≤150	≤100	≤50	≤20
Liver Bilirubin, mg/dL‡	<1.2	1.2-1.9	2.0-5.9	6.0-11.9	>12.0
Cardiovascular Hypotension	No hypotension	Mean arterial pressure <70 mm Hg	Dop ≤5 or dob (any dose)§	Dop >5, epi ≤0.1, or norepi ≤0.1§	Dop >15, epi >0.1, or norepi >0.1§
Central nervous system Glasgow Coma Score Scale	15	13-14	10-12	6-9	<6
Renal Creatinine, mg/dL or urine output, mL/d‖	<1.2	1.2-1.9	2.0-3.4	3.5-4.9 or <500	>5.0 or <200

*Norepi indicates norepinephrine; Dob, dobutamine; Dop, dopamine; Epi, epinephrine; and Fio_2, fraction of inspired oxygen.
†Values are with respiratory support.
‡To convert bilirubin from mg/dL to μmol/L, multiply by 17.1.
§Adrenergic agents administered for at least 1 hour (doses given are in μg/kg per minute).
‖To convert creatinine from mg/dL to μmol/L, multiply by 88.4.

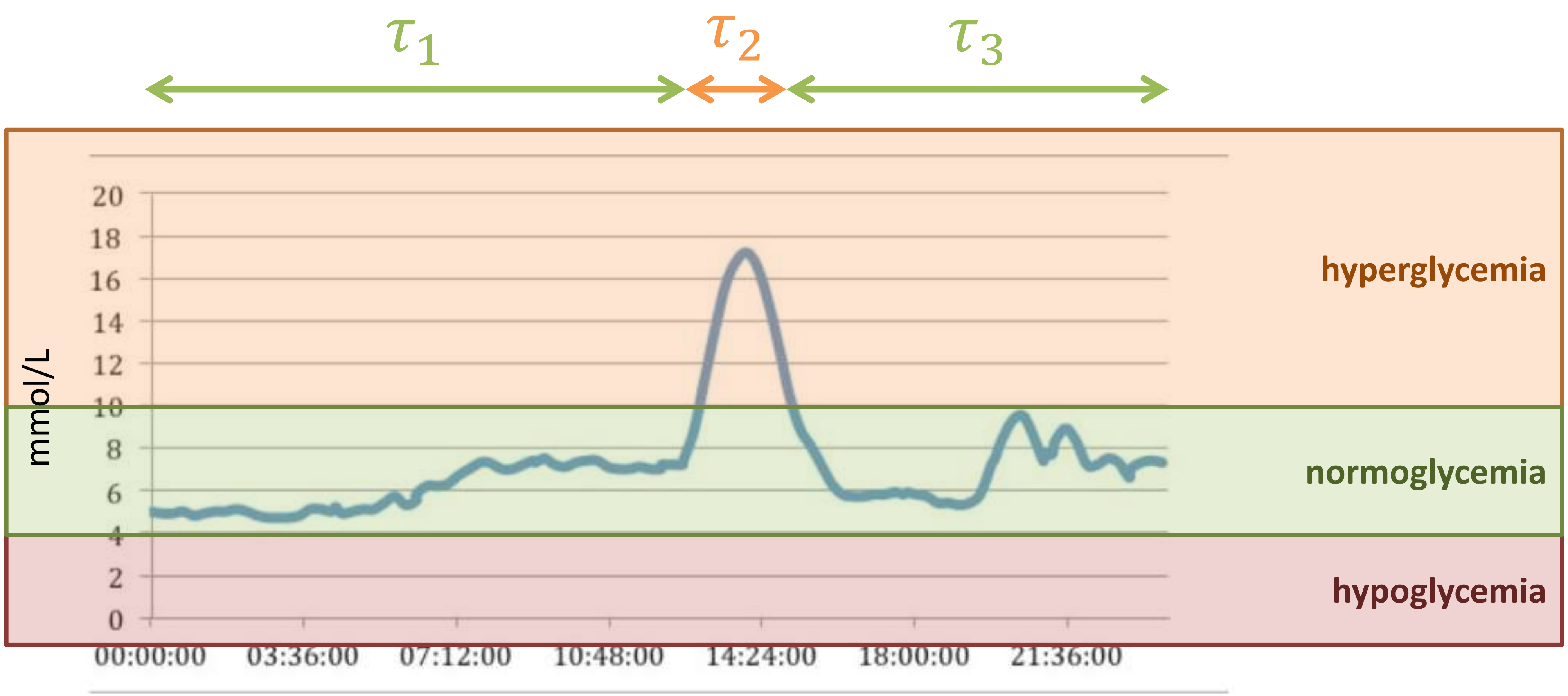


Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *JAMA.* 2001;286(14):1754–1758.

Reward r_t: Sepsis Management

- Receive a large reward if the patient survives, and a large penalty if they die
- Receive a penalty if the SOFA score remains greater than zero
- Receive an additional penalty/reward proportional to the increase/decrease in the SOFA score
- Receive an additional penalty/reward proportional to the increase/decrease in lactate

Reward r_t: Artificial Pancreas



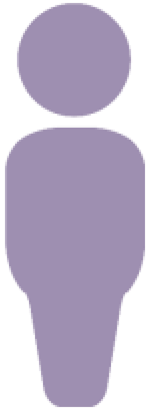
$-1 * \tau_2$

$0 * (\tau_1 + \tau_3)$

$-10 * \tau_{\text{hypo}}$

https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e

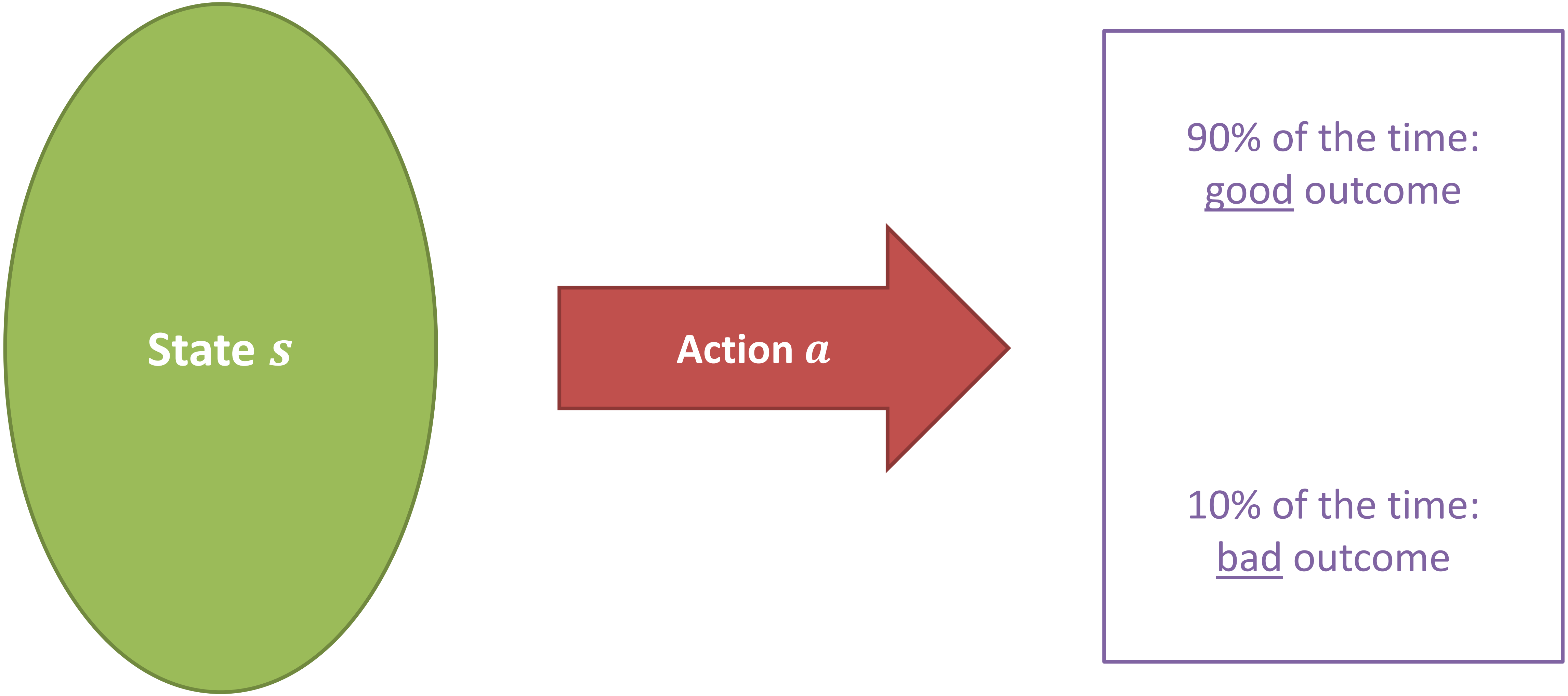
The reward quantifies our objectives



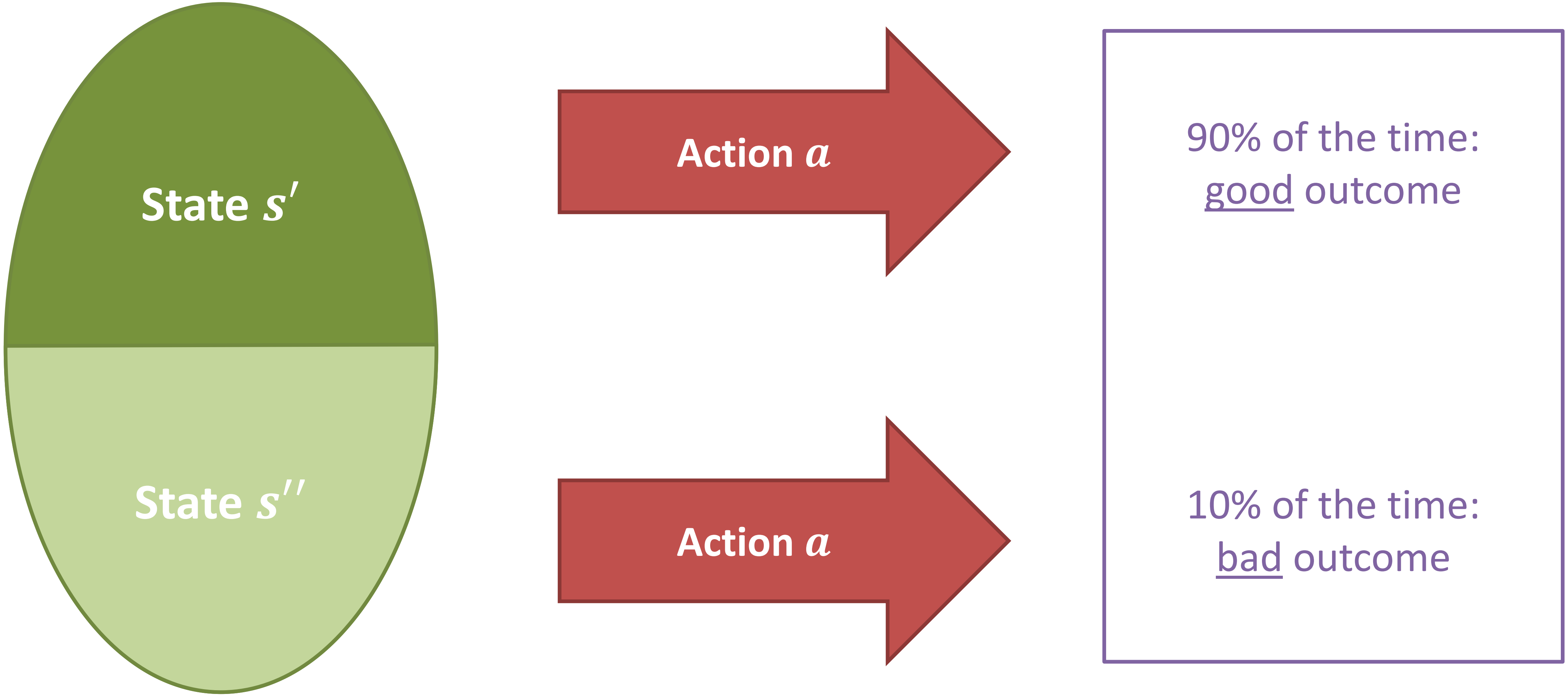
Sepsis Management and Artificial Pancreas

CHALLENGE 1: MISSING INFORMATION

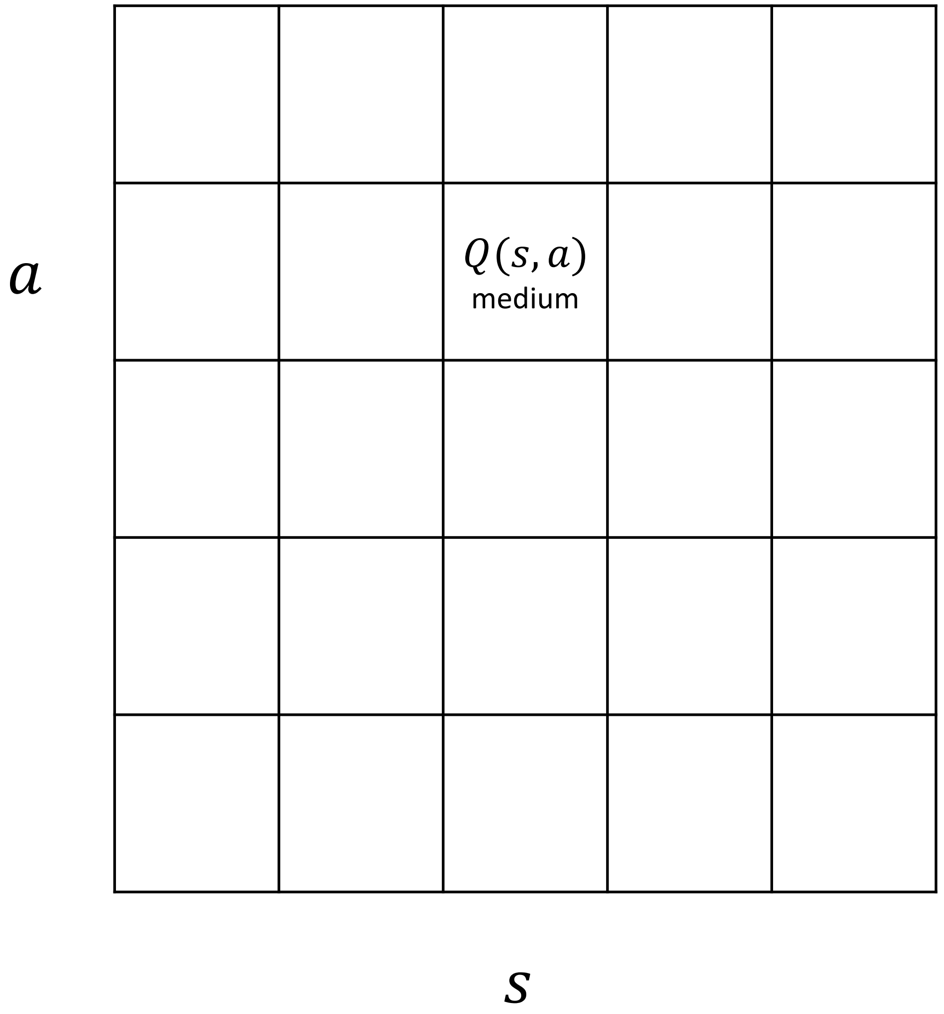
What happens when important information is missing?



What happens when important information is missing?



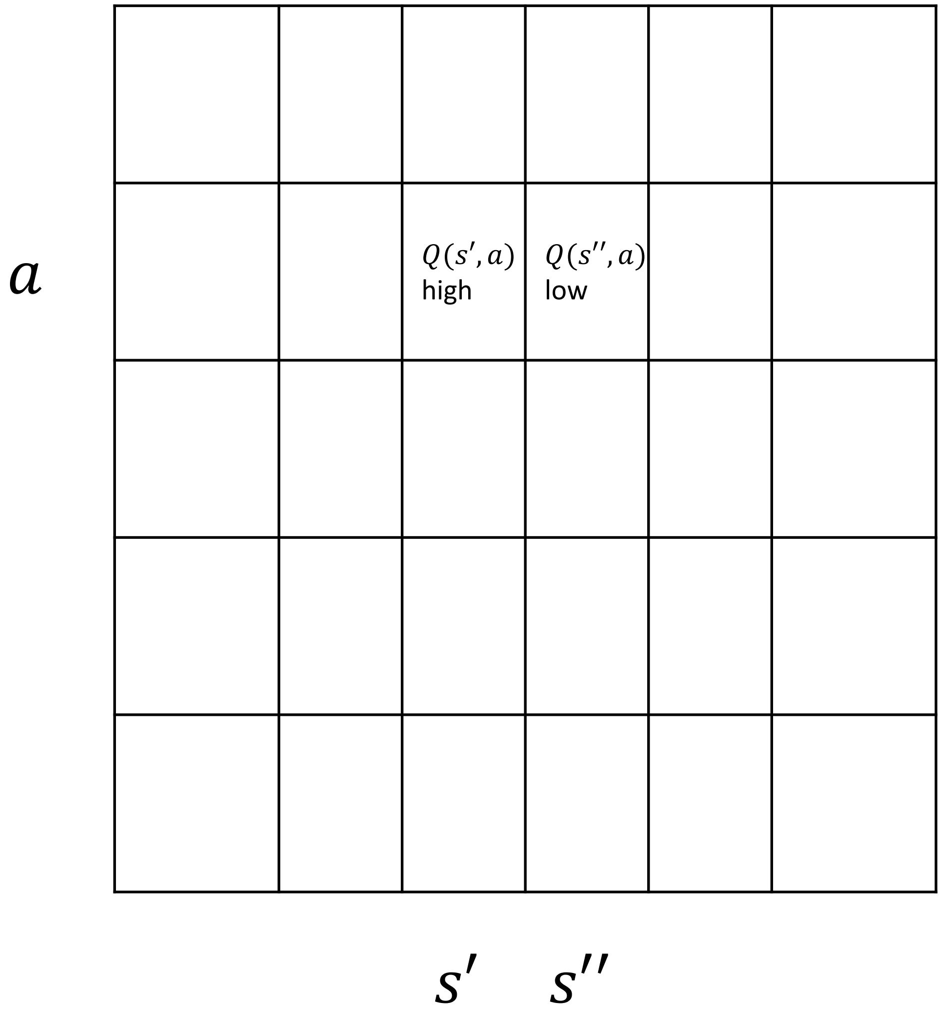
Negative impact of missing state information



Q Learning:

- The expected reward for choosing action a while in state s is moderate

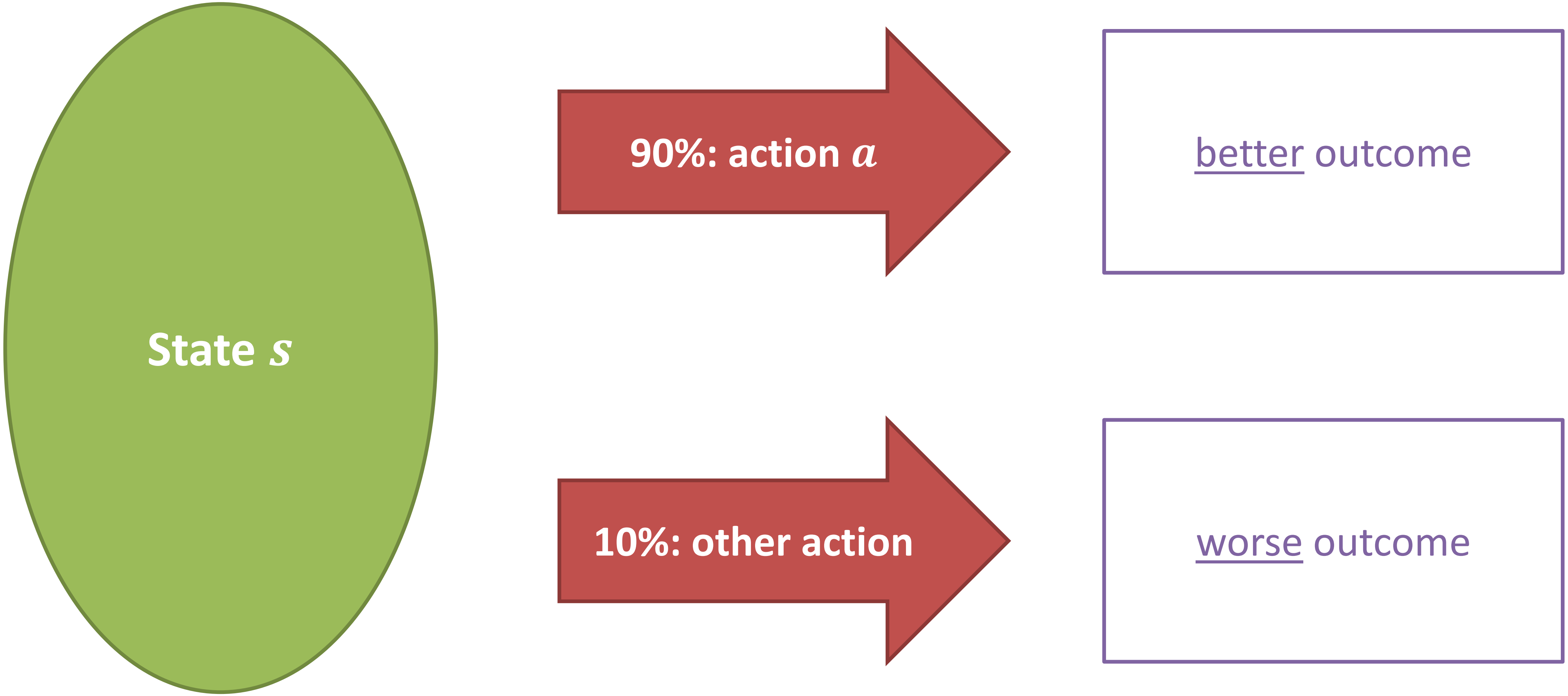
Negative impact of missing state information



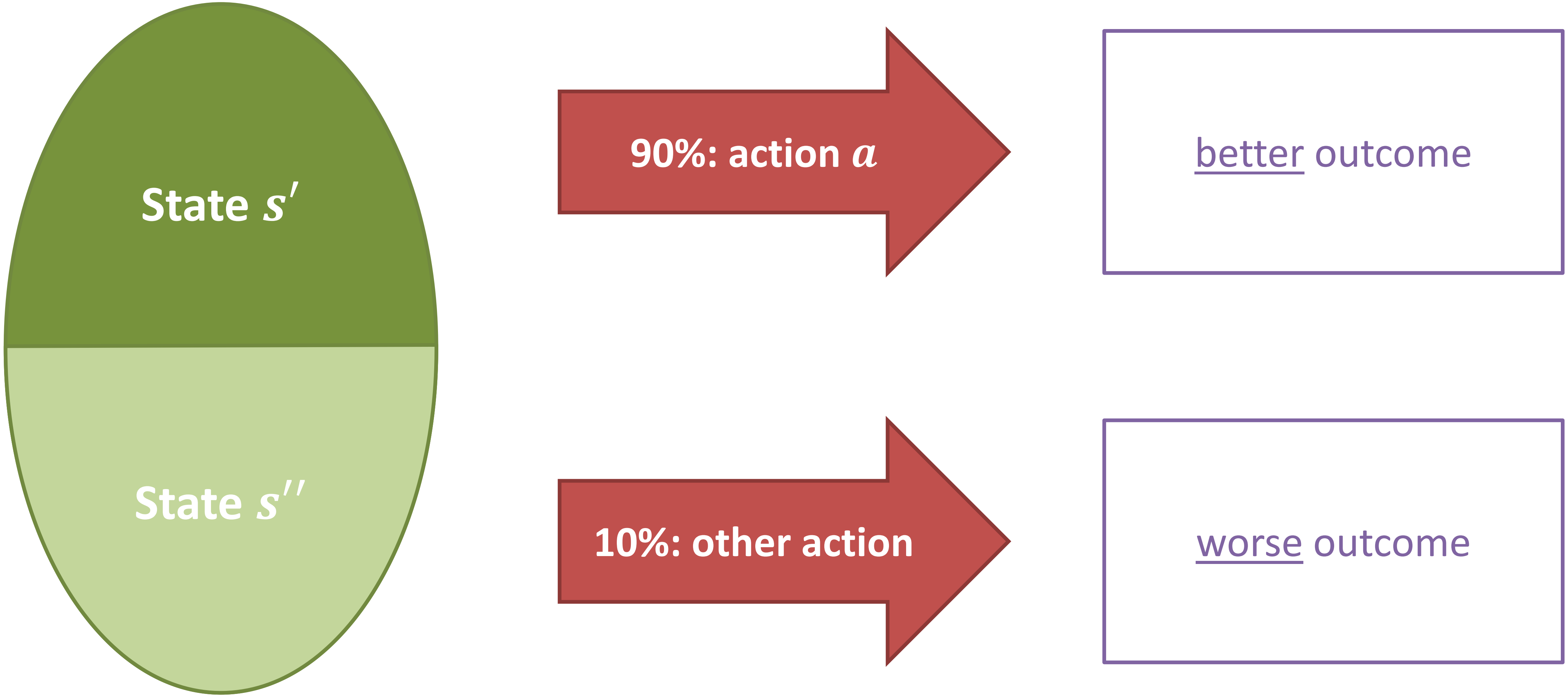
Q Learning:

- The expected reward for choosing action a while in state s' is high
- The expected reward for choosing action a while in state s'' is low

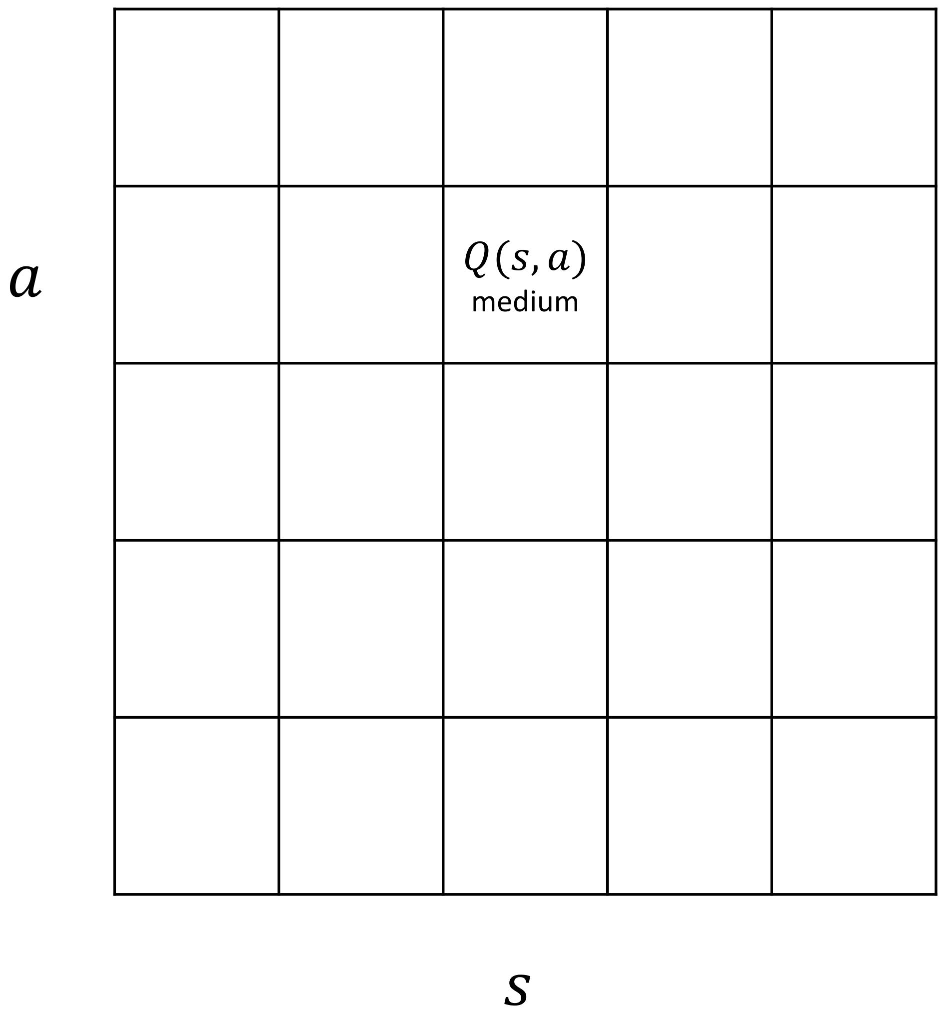
Learning from observational data: where things can REALLY go wrong



What's really happening…



Learning from observational data: where things can REALLY go wrong

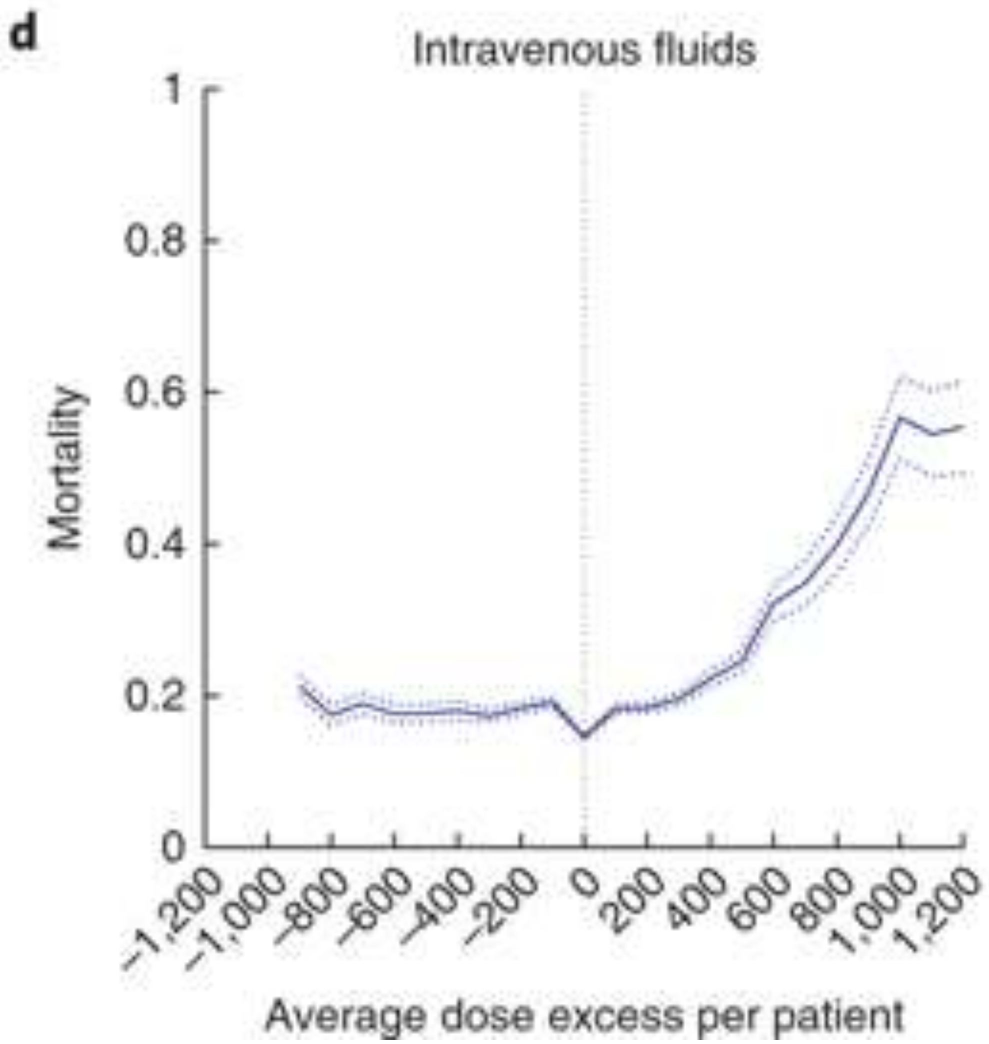


Q Learning:

- The expected reward for choosing action a while in state s <u>appears to be high</u>

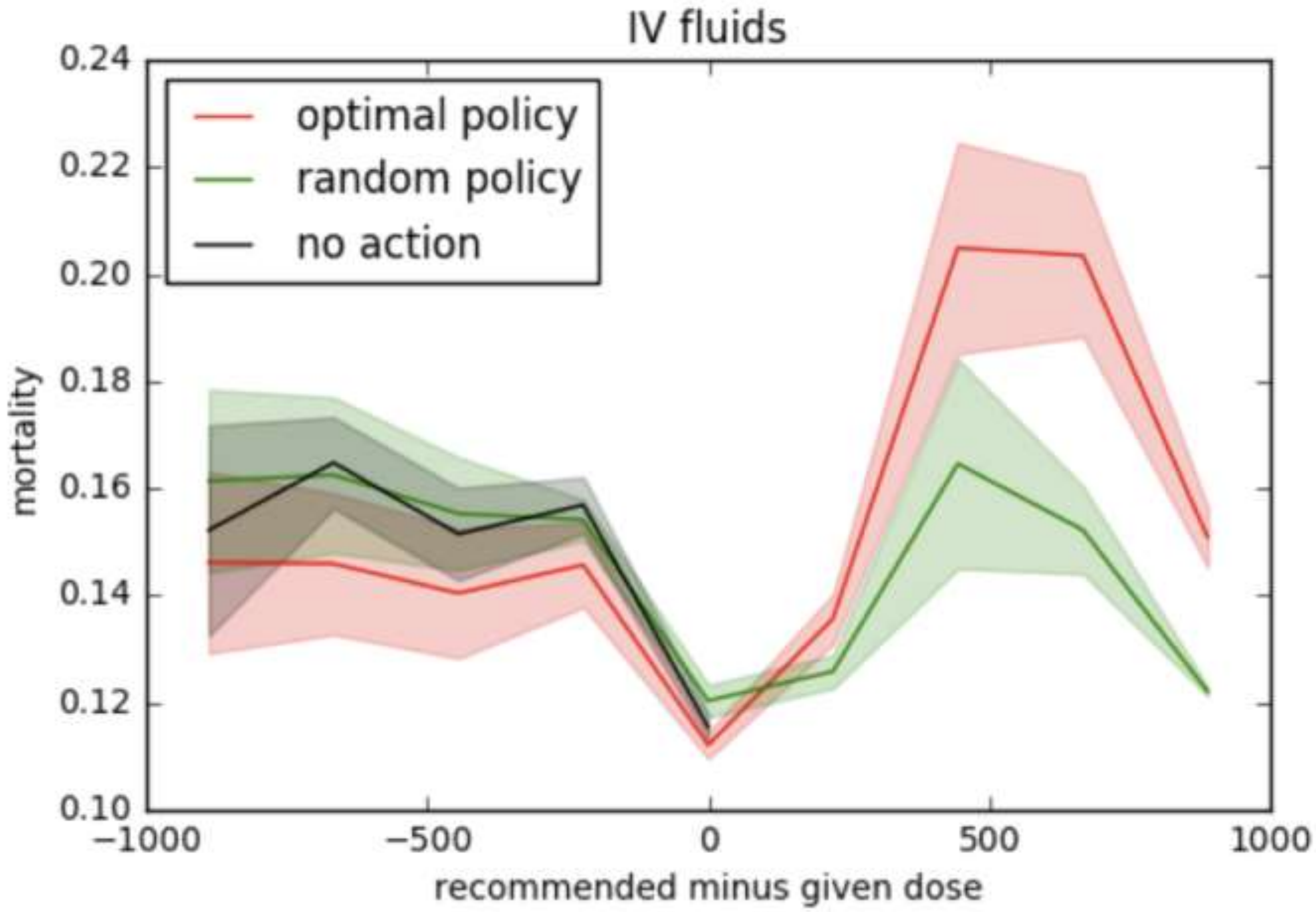
- In fact, this is a very bad action for some patients

Sepsis Results: Observed Mortality



Concerns about Off-Policy Evaluation

U curve with naïve baselines



1) Sicker patients get higher dosages!
2) Discretizing dosages by quantile bad.

Slide Credit: Michael Hughes (michaelchughes.com)

RL: LEARNING THROUGH TRIAL AND ERROR

In RL, we typically learn “from scratch”:

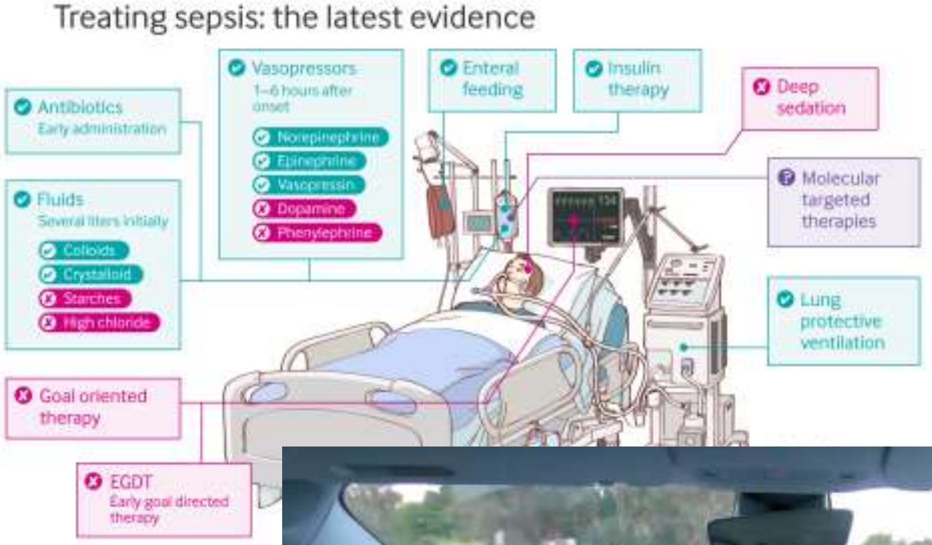
- Try things and see what works
- Initially our actions are random

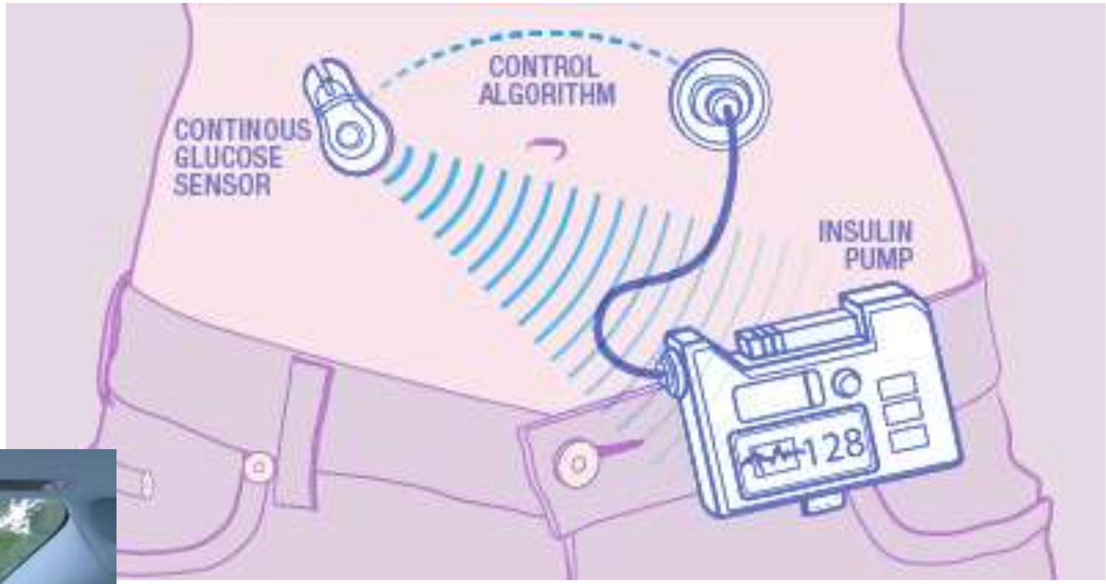
Drone Uses AI and 11,500 Crashes to Learn How to Fly

Crashing into objects has taught this drone to fly autonomously, by learning what not to do

By [Evan Ackerman]

Failing 11,500 times isn't always an option





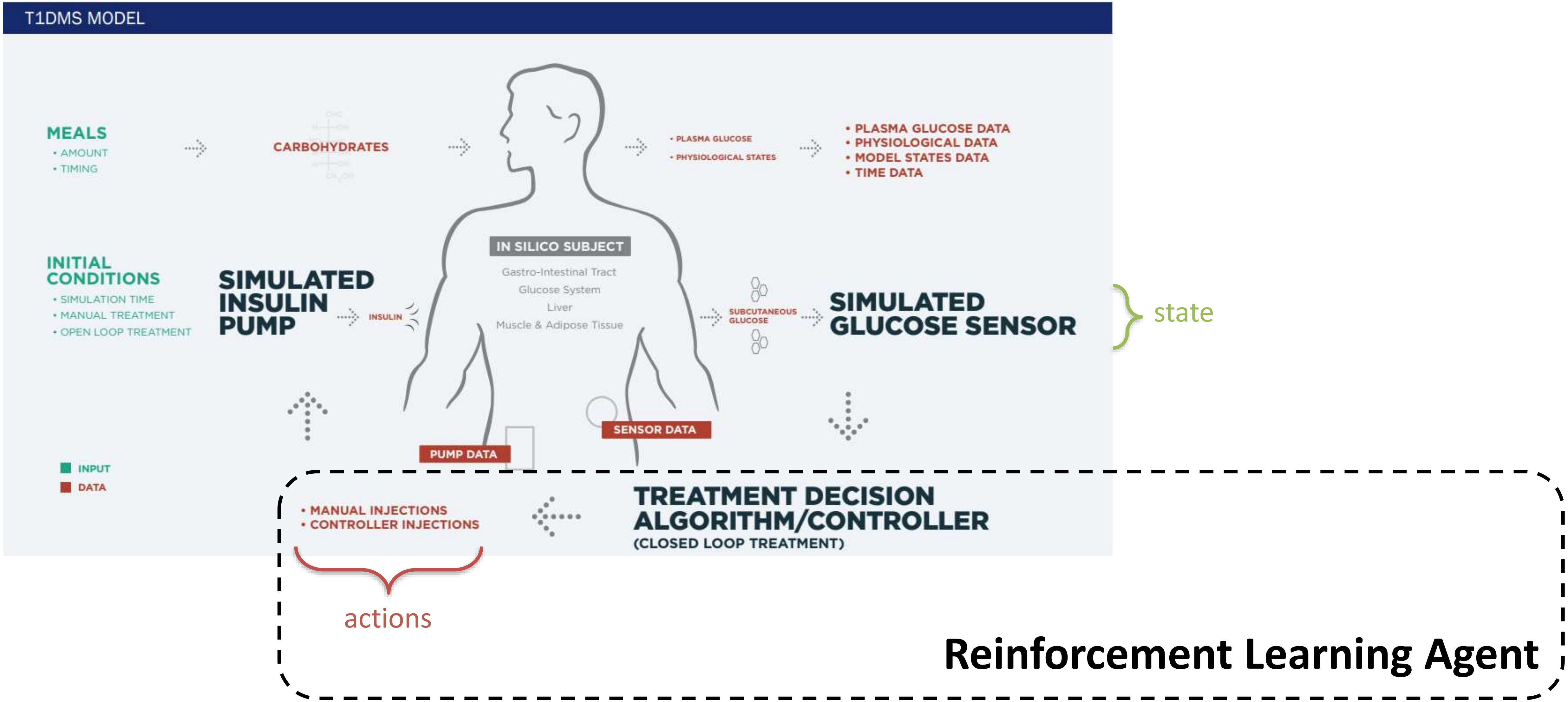
-> What are the alternatives?

A1. Learn from Observational Data

The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data (2018).

- provides a large number of (state, action, reward) examples

A2. Learn Policy from Simulated Environment



A3. Learn with Expert Oversight

physiologic state

algorithm recommendation

physician approval

Open Questions for RL

1. How can we incorporate existing knowledge to avoid “starting from scratch”?

2. *Should we* avoid starting from scratch?

Clinical Trials and Self-Experimentation

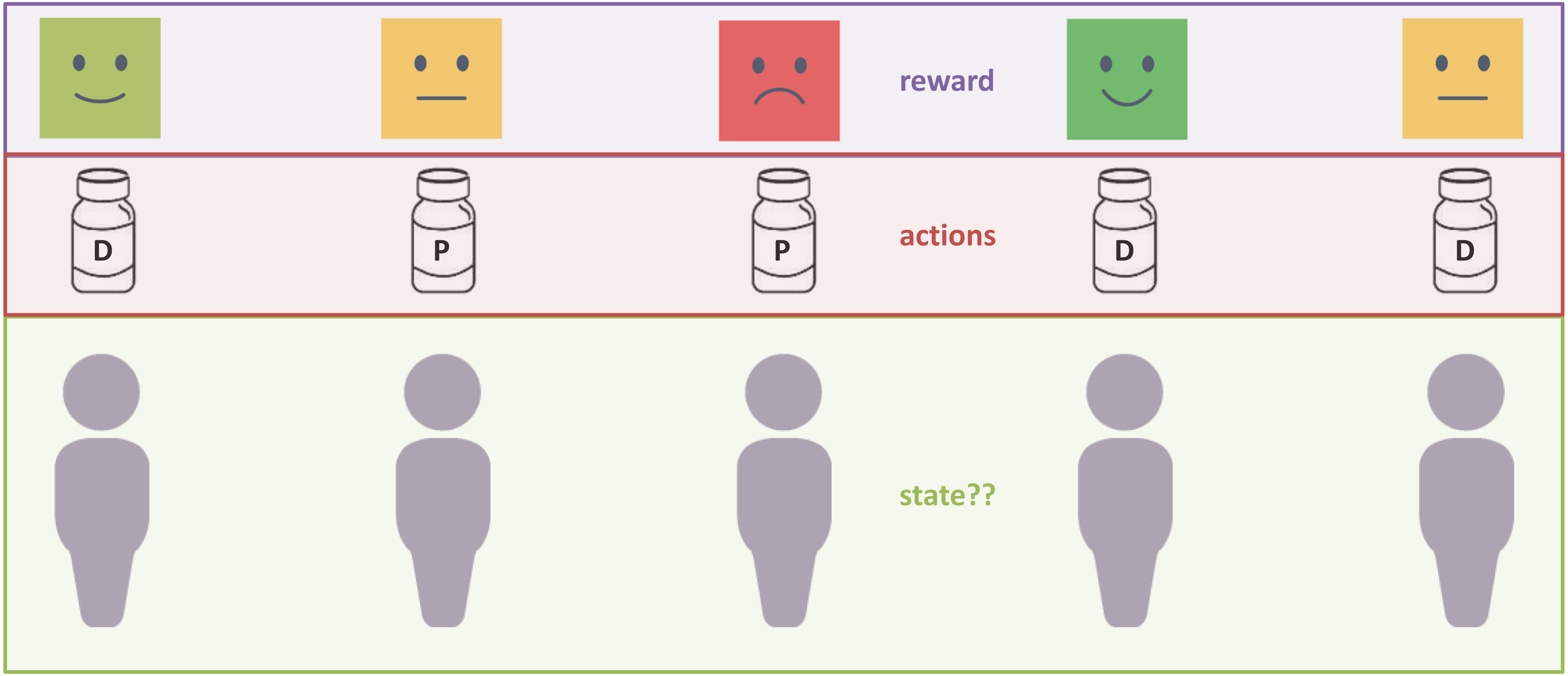
OTHER DIRECTIONS IN RL FOR MEDICINE

Sequential Decision-Making Problems are Everywhere in Medicine

A reinforcement learning approach to weaning of mechanical ventilation in intensive care units.
Prasad, Niranjani, et al. arXiv:1704.06300 (2017).

Suppose we are evaluating a new drug…

A special case of RL



action → reward

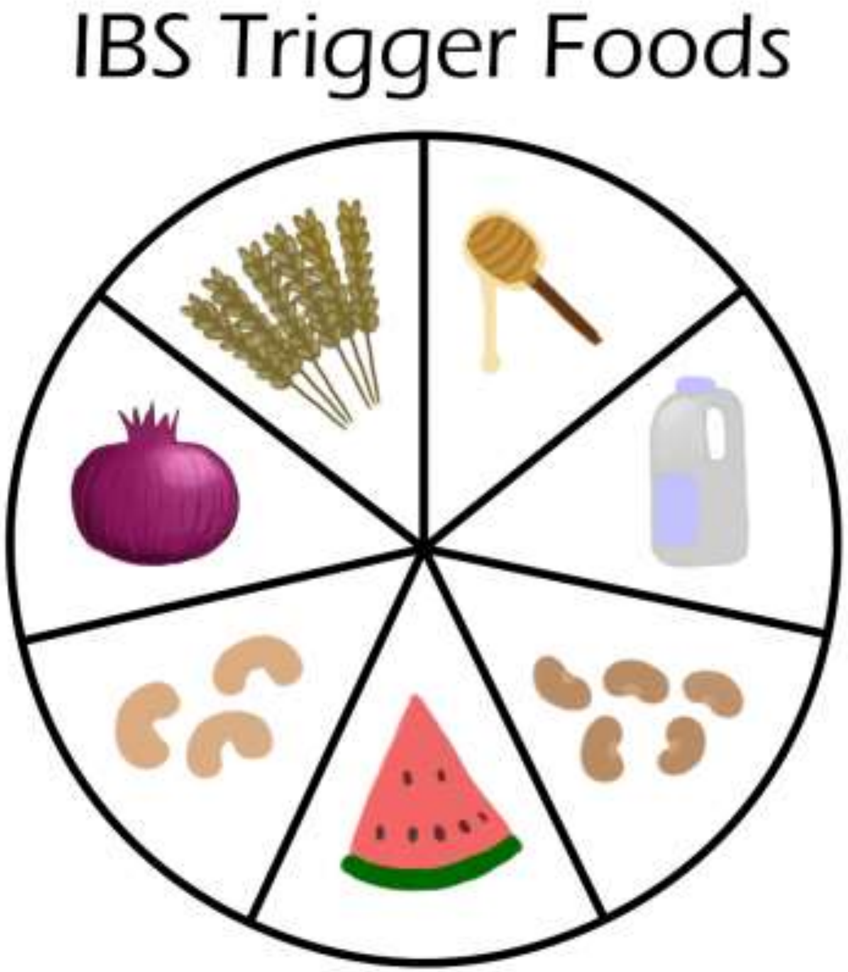
Multi-armed Bandit

Application: Optimal Allocation of Clinical Trial Participants

"An explicit assumption is the goal to treat patients effectively, in the trial as well as out. That is controversial (…)"

(Stangl, Inoue and Irony, 2012)

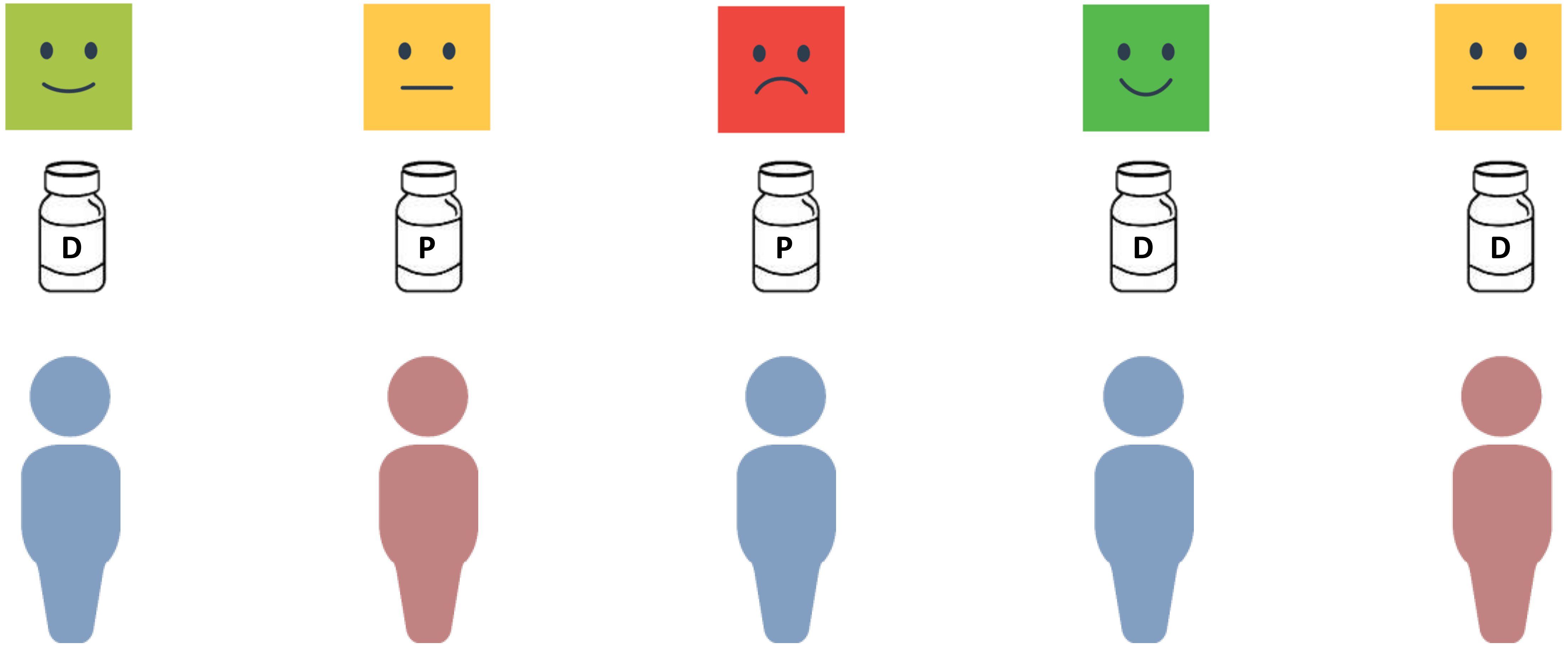
N-of-1 Trial: Identify IBS Triggers



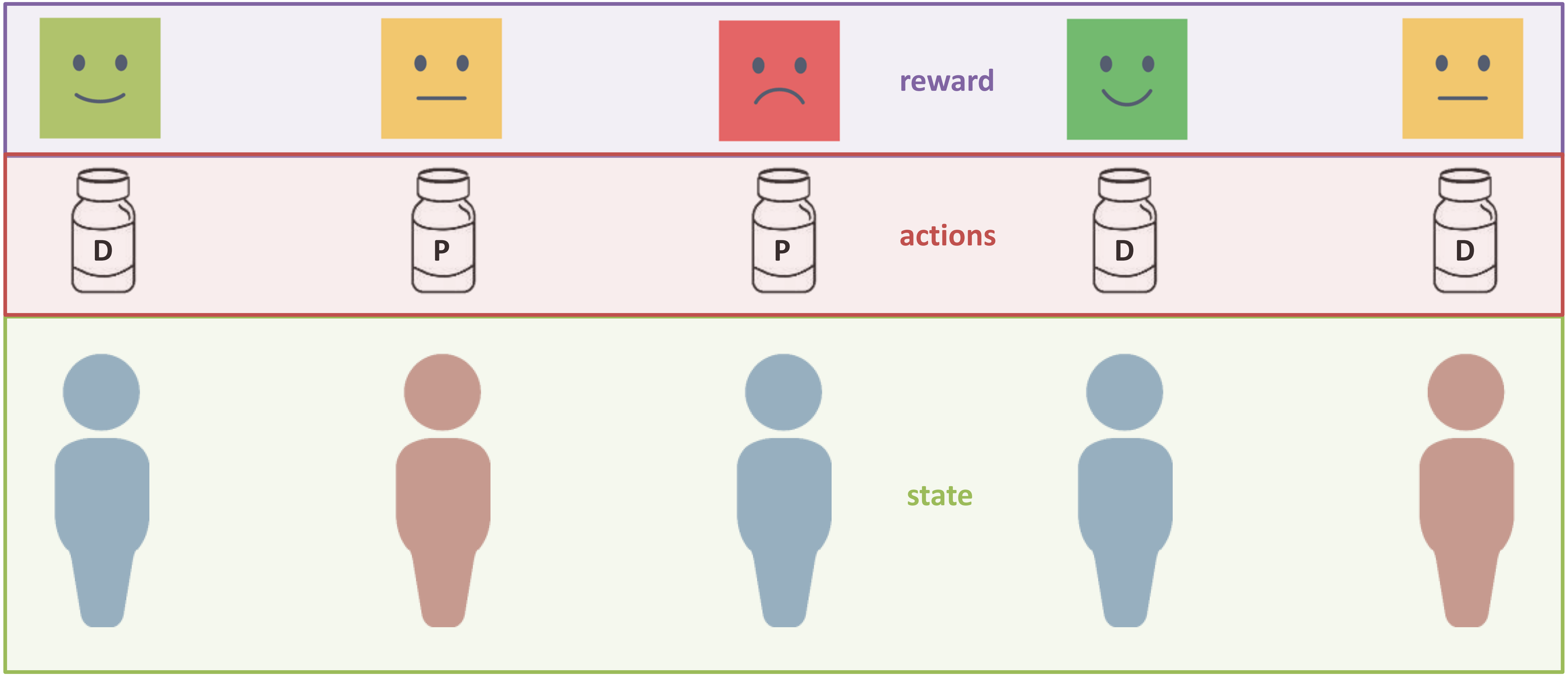
Find foods (i.e. “actions”) that minimize IBS symptoms (i.e. “reward”)

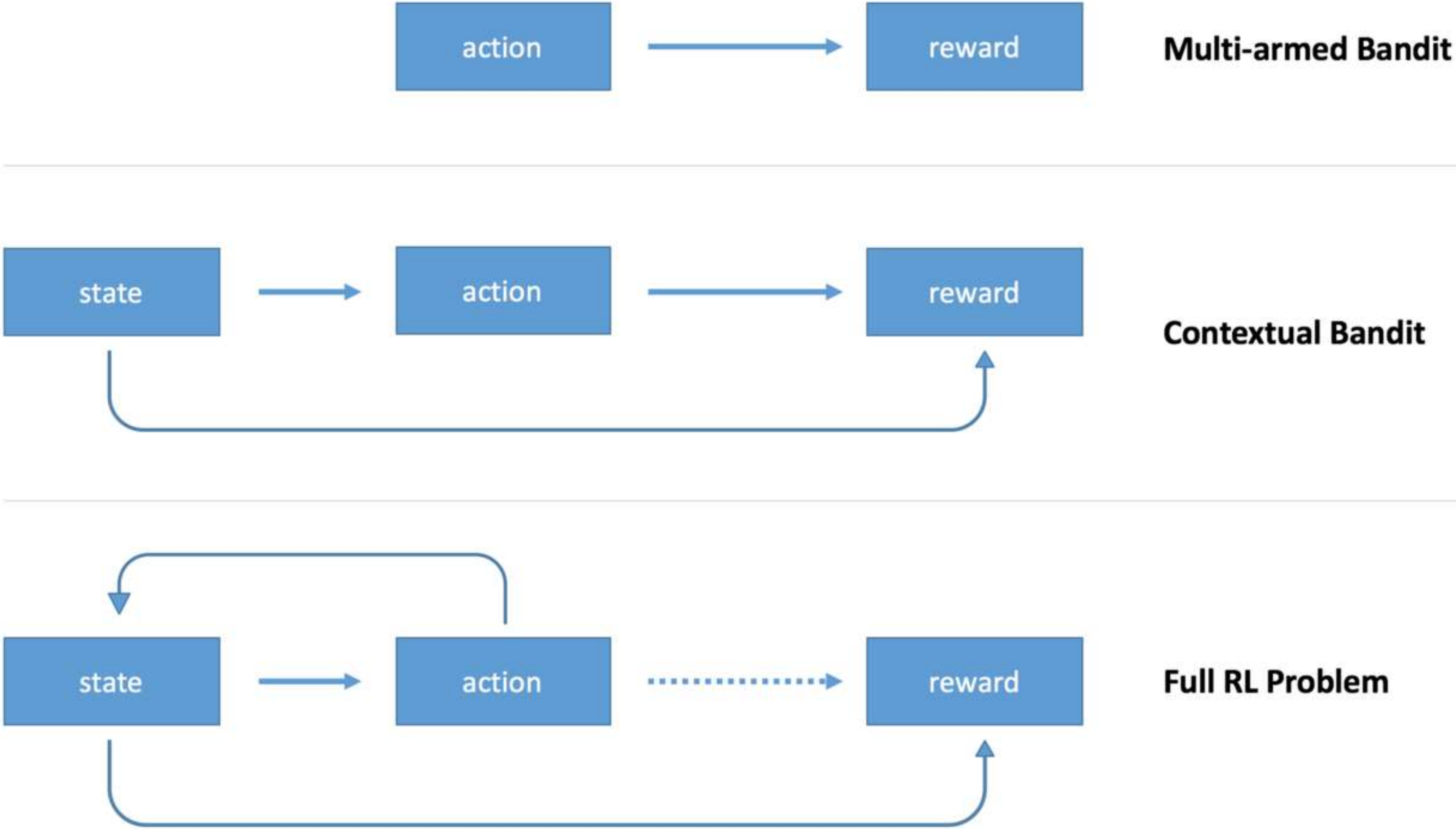
TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers.
Karkar R, Schroeder J, Epstein DA, et al.
SIGCHI Conference 2017;2017:6850-6863.

This time, we track what works for men versus women



Personalized clinical trials can be formulated as RL





From https://medium.com/emergent-future/