

# Sequential Medical Decision-Making with RL

12 July, 2019

Duke/Duke-NUS  
Machine Learning Summer School

Matthew Engelhard

Sepsis Management and Artificial Pancreas

# **TWO ILLUSTRATIVE EXAMPLES**

# Sequential Medical Decision-Making

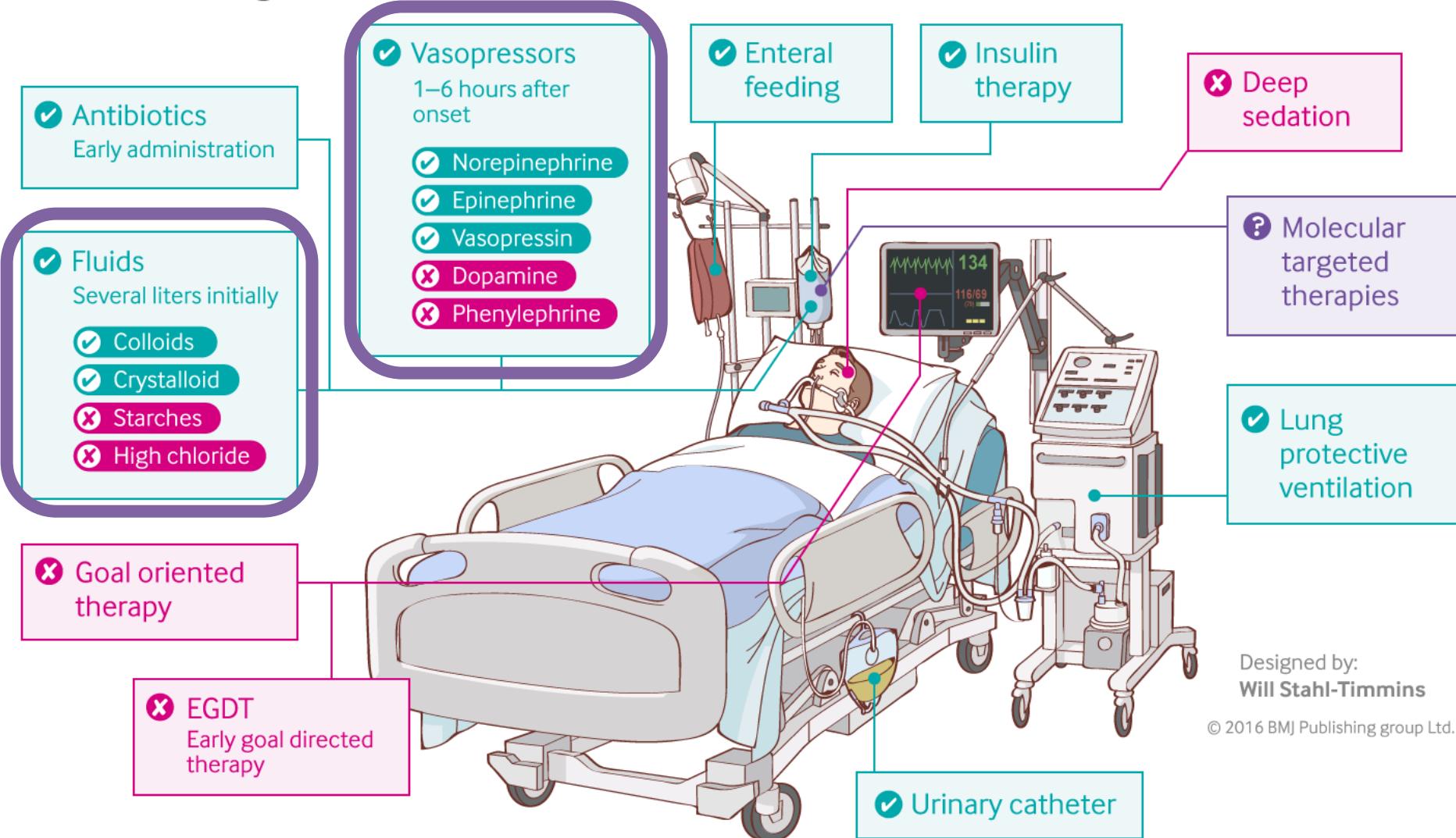
An agent

takes actions

based on the state of  
a system

to maximize reward

# Treating sepsis: the latest evidence



Designed by:  
Will Stahl-Timmins

© 2016 BMJ Publishing group Ltd.

**“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

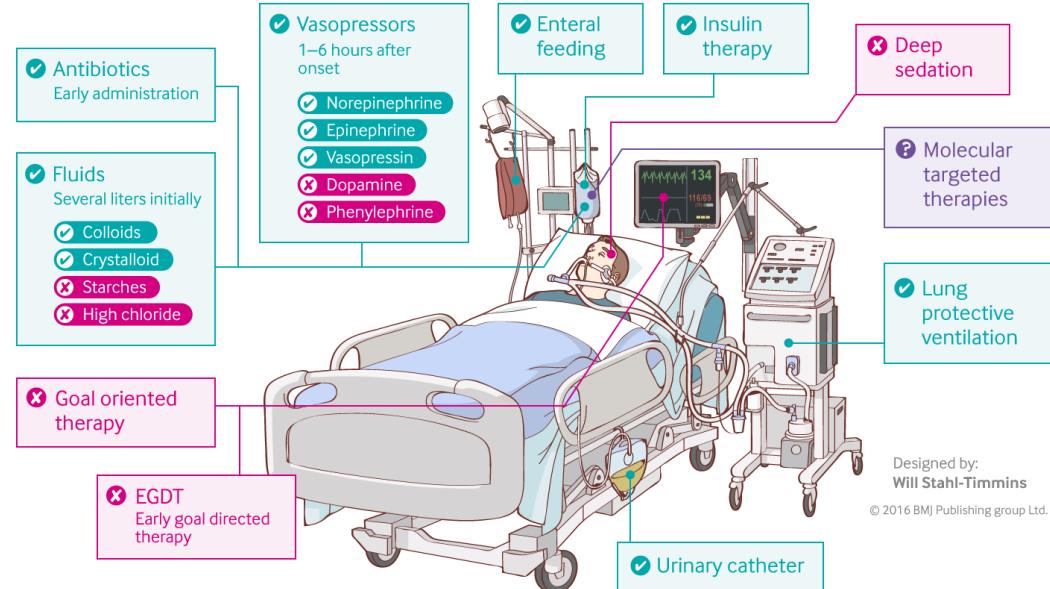
An agent

takes actions

based on the state of a system

to maximize reward

### Treating sepsis: the latest evidence



Designed by:  
Will Stahl-Timmins  
© 2016 BMJ Publishing group Ltd.

A clinician

gives fluid and/or vasopressor

based on the patient's physiologic status

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

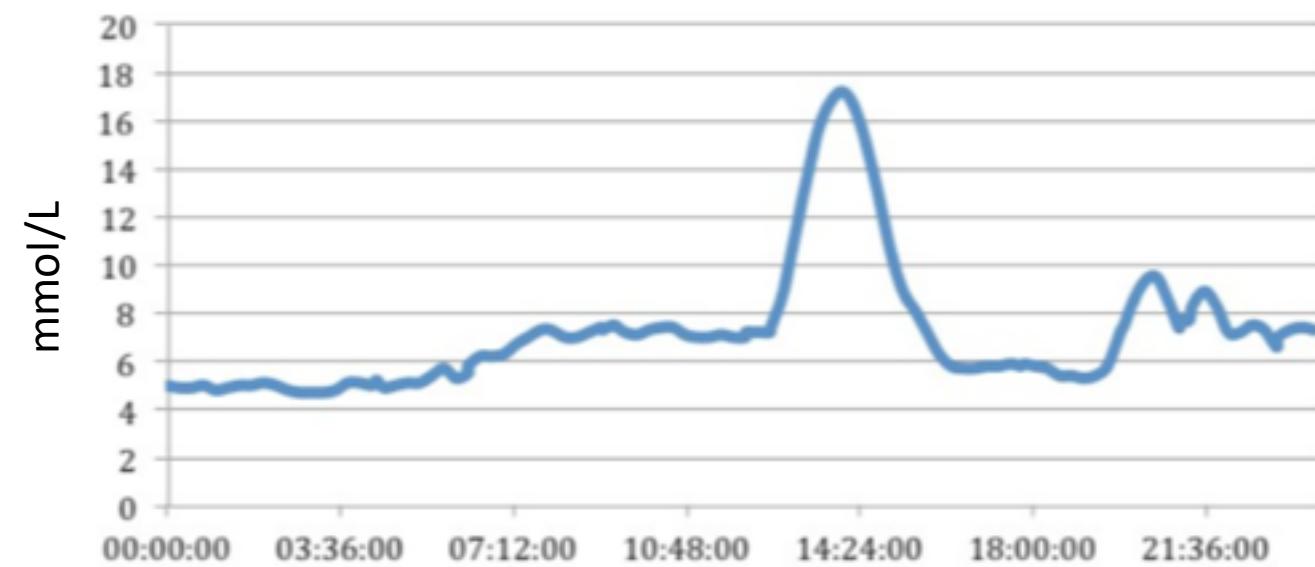


The image shows a screenshot of a Nature Medicine article page. At the top left is a 'MENU' button. To its right is the 'nature medicine' logo. Below the logo, the word 'Article' is followed by 'Published: 22 October 2018'. The main title of the article is 'The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care'. Below the title, the authors' names are listed: Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, & A. Aldo Faisal. At the bottom of the article summary, it says 'Nature Medicine 24, 1716–1720 (2018) | Download Citation'.

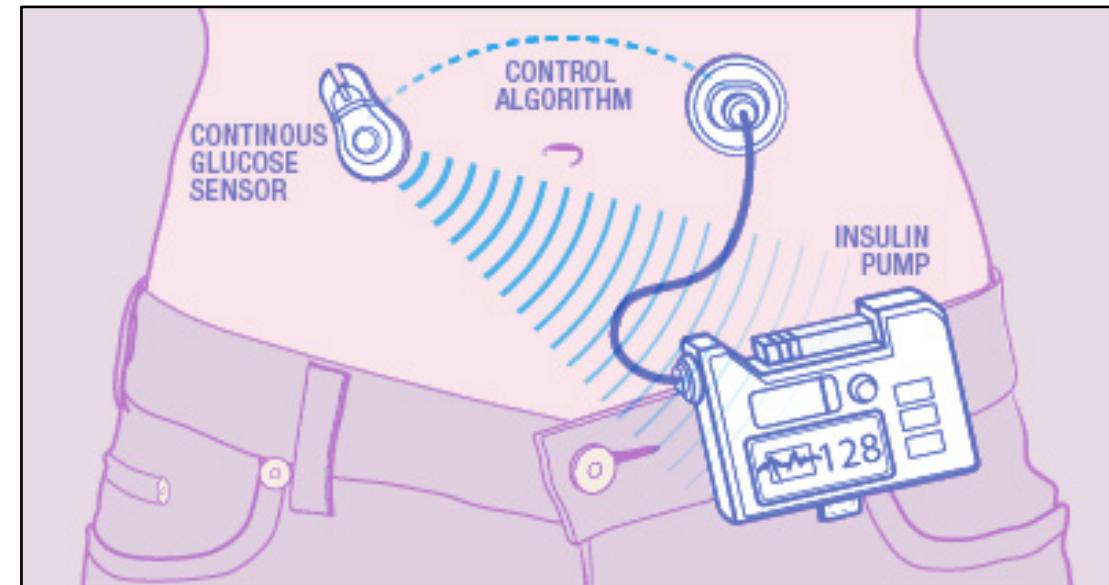
Independent validation on the Philips eICU Research Institute Database:  
>3.3 million admissions from 2003–2016 in 459 ICUs across the US

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



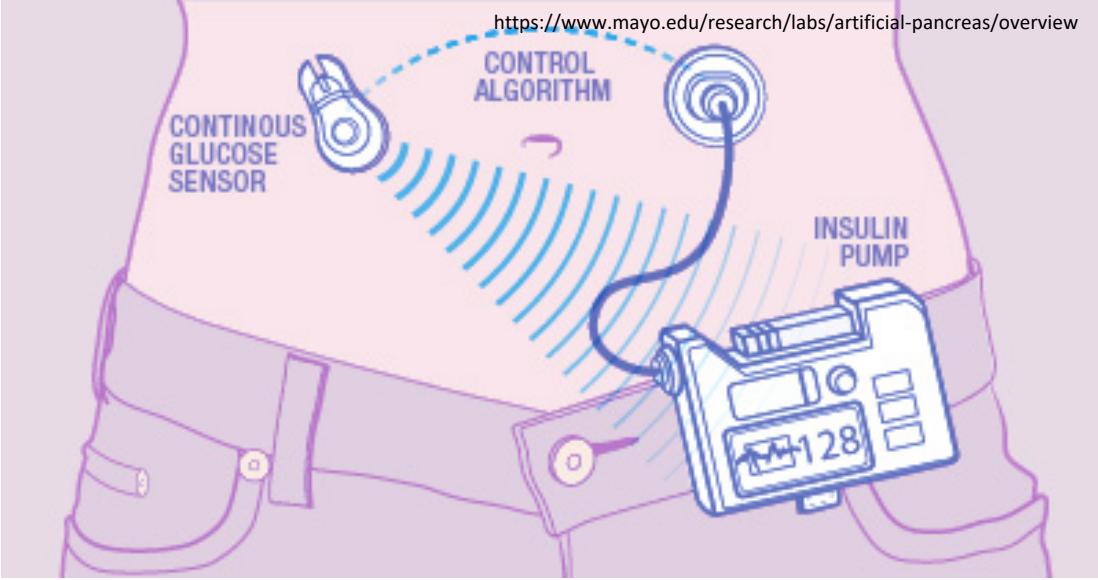
# Sequential Medical Decision-Making: Artificial Pancreas

An **agent**

takes **actions**

based on the **state** of  
a system

to maximize reward



A **computer program**

administers insulin

based on **blood glucose** and  
**recent patient behaviors**

to maintain **normoglycemia**



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

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

# The RL Paradigm

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

- 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  $a$

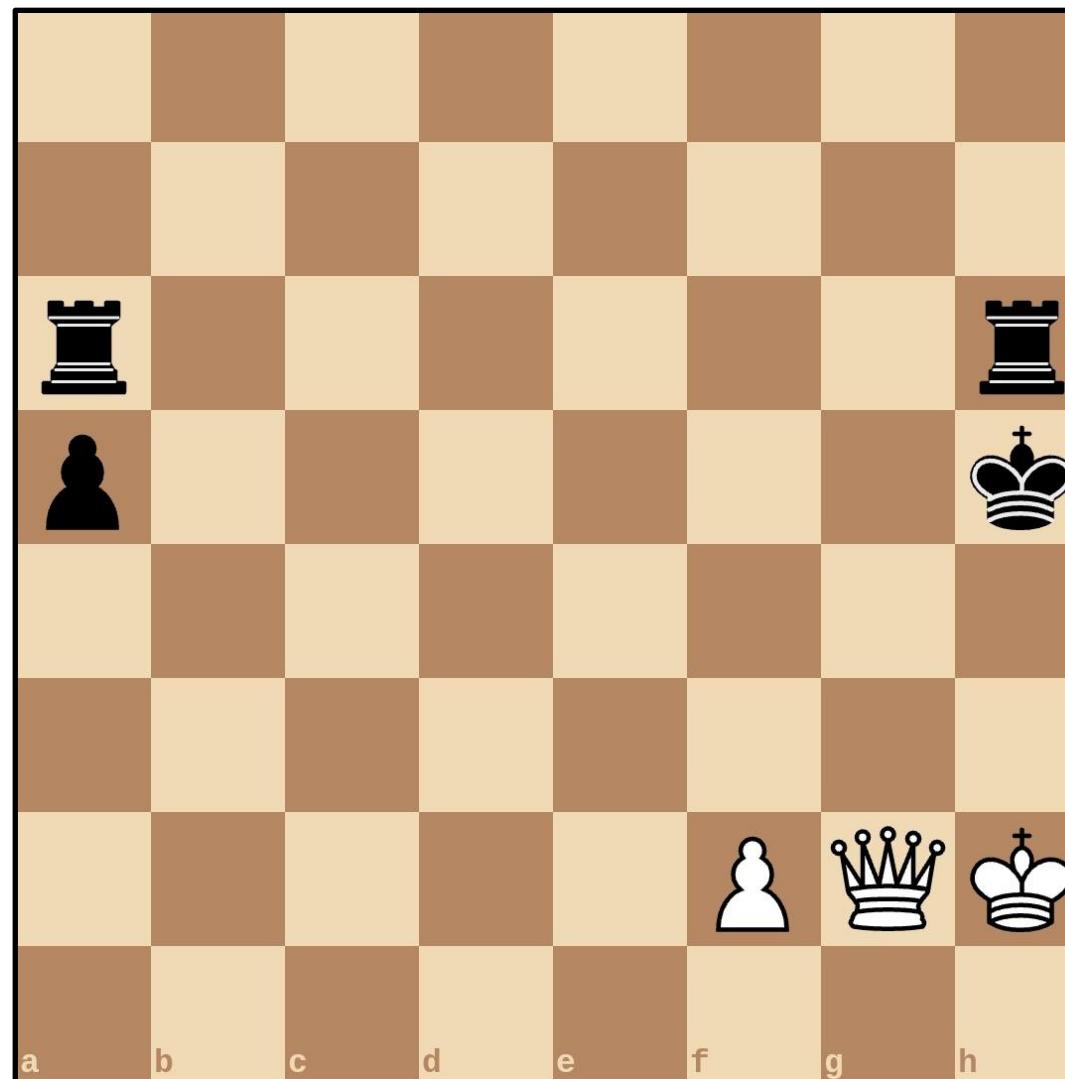
**GOAL:**

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

# The RL Paradigm

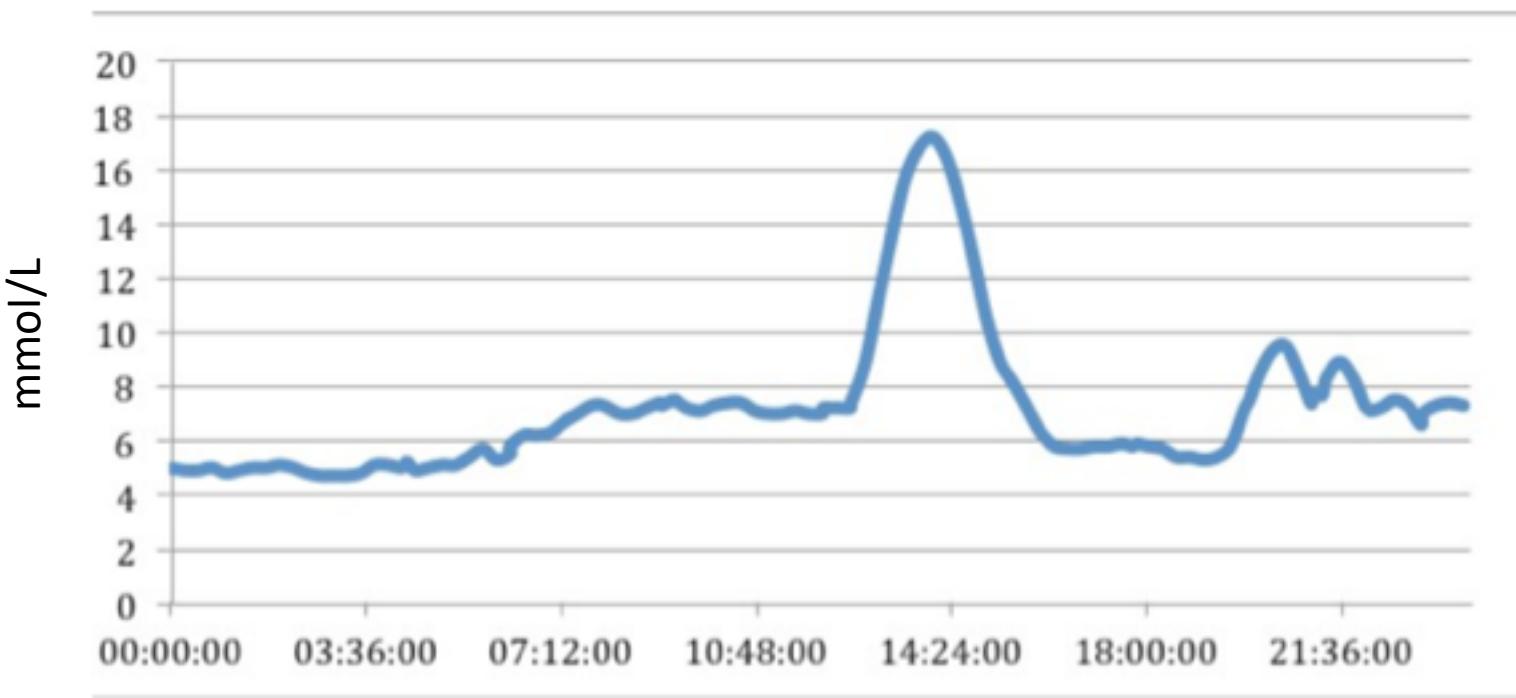
...  $s_{t-1}$   $a_{t-1}$   $r_{t-1}$   $s_t$   $a_t$   $r_t$   $s_{t+1}$  ...

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



# State $s_t$ : Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



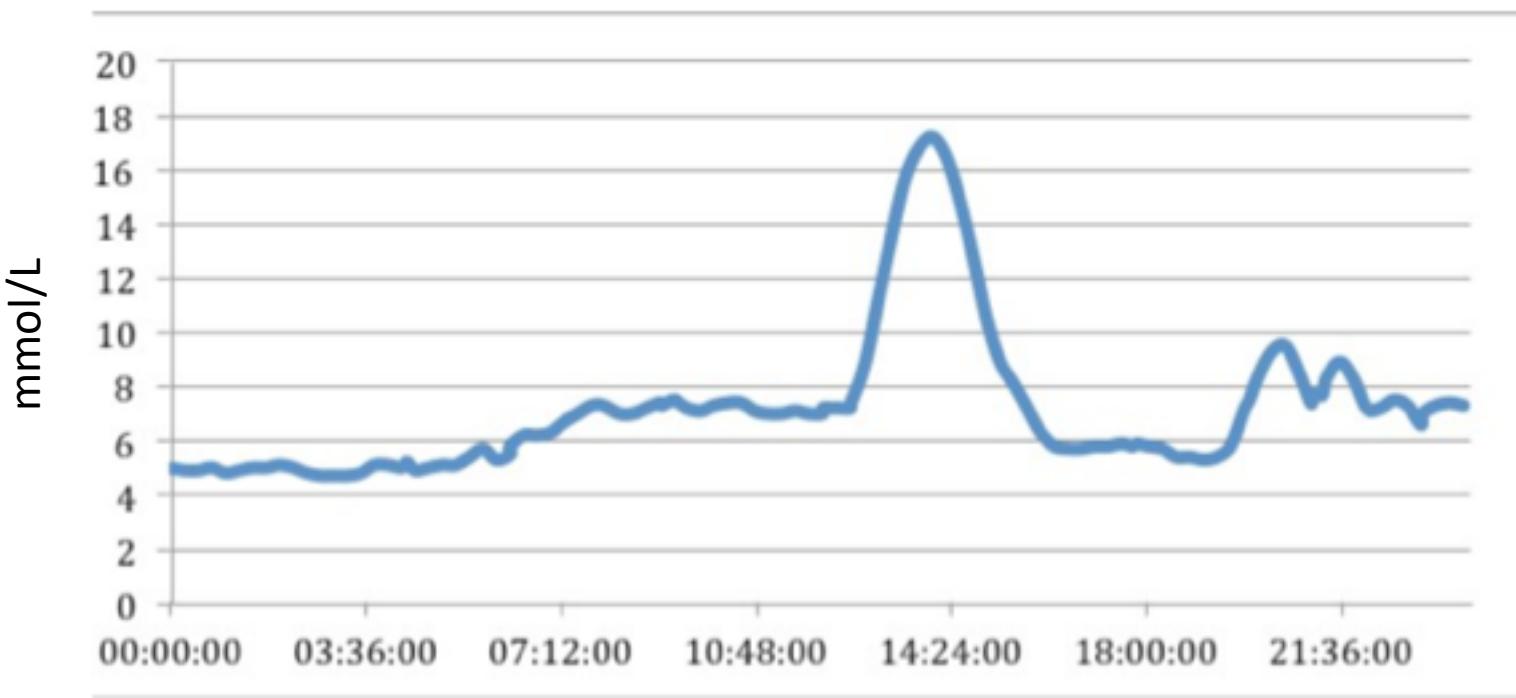
Idea 1:

The state is the current blood glucose value

[https://medium.com/@justin\\_d\\_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e](https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e)

# State $s_t$ : Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



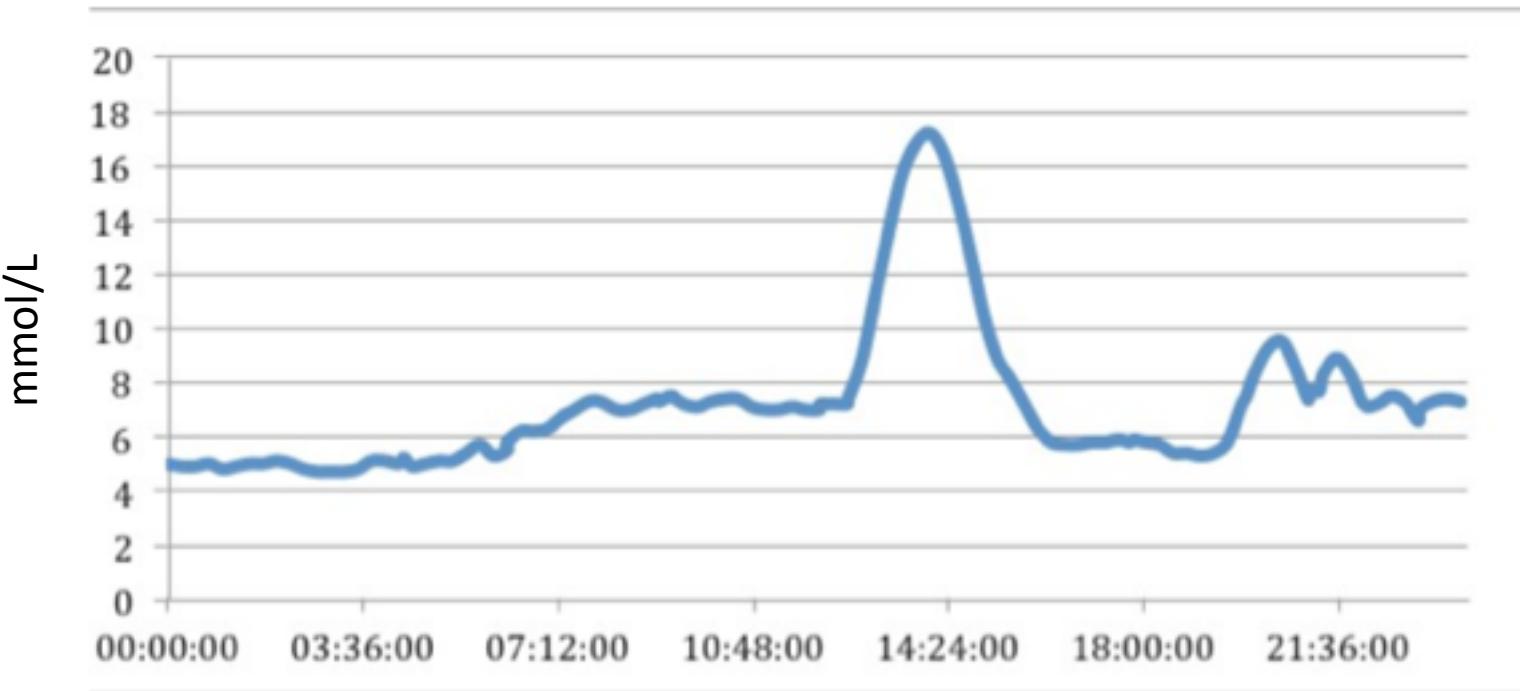
## Idea 2:

The state is the current blood glucose value plus recent trends

[https://medium.com/@justin\\_d\\_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e](https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e)

# State $s_t$ : Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



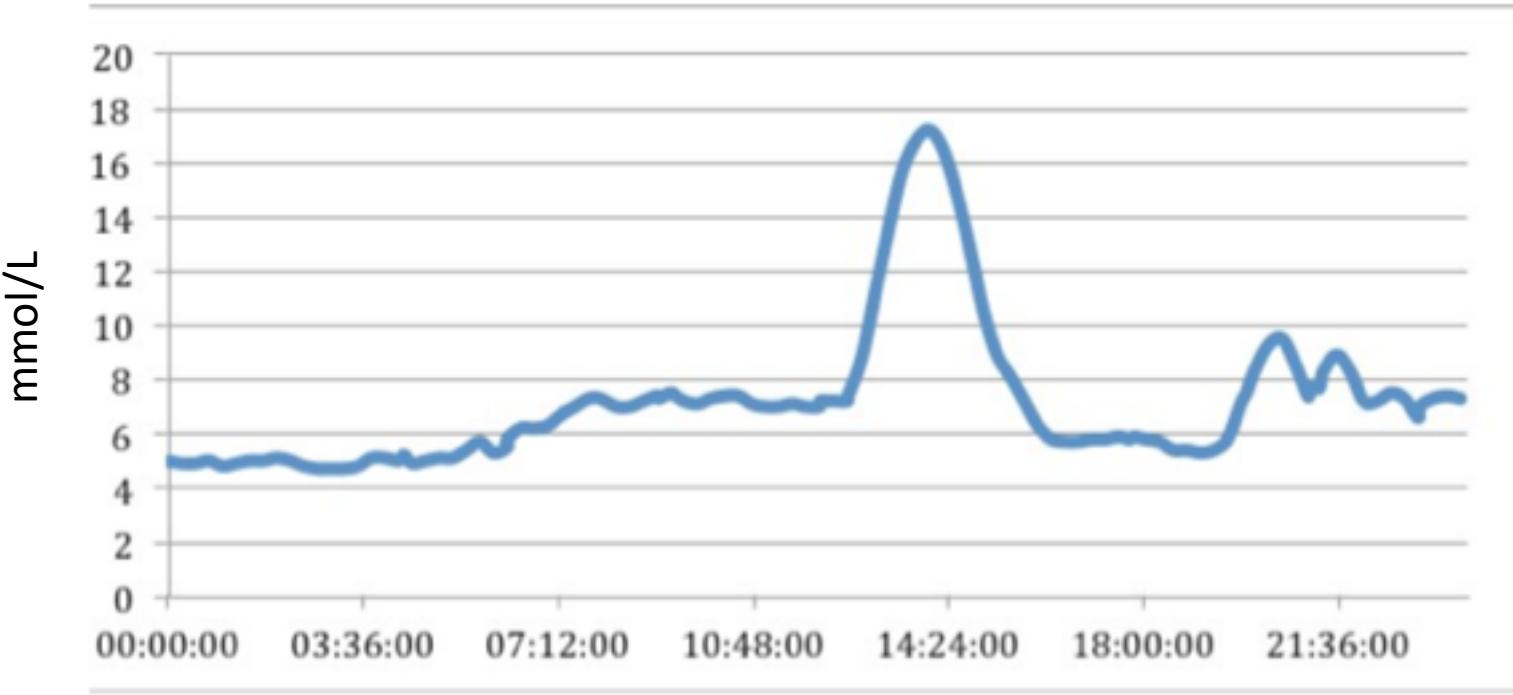
## Idea 3:

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

[https://medium.com/@justin\\_d\\_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e](https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e)

# State $s_t$ : Artificial Pancreas

Blood Glucose Readings from Continuous Glucose Monitor



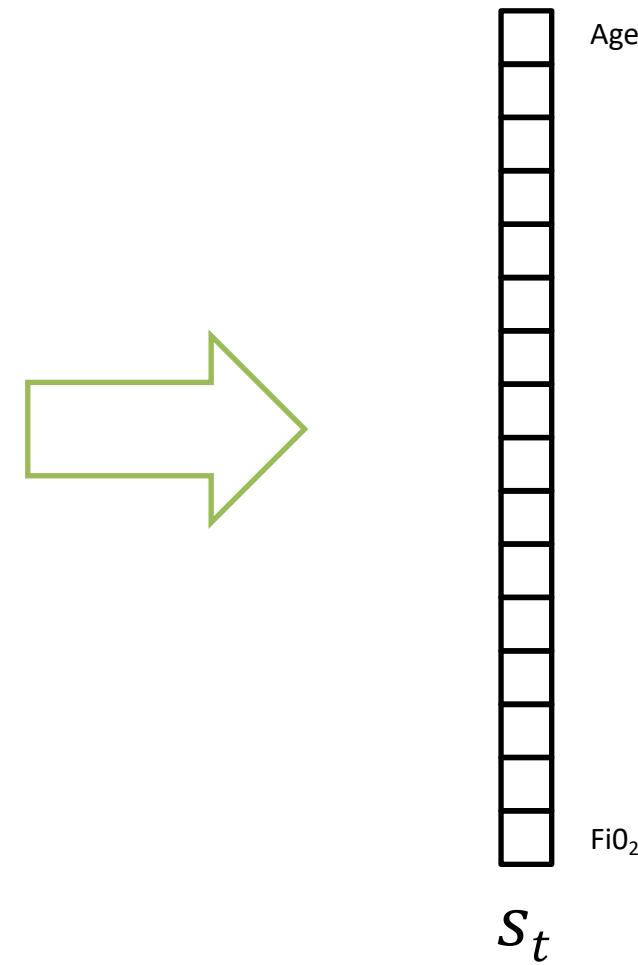
[https://medium.com/@justin\\_d\\_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e](https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e)

The state: all information relevant to our decision

- 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
Lab values	Temperature	Continuous
	Potassium, sodium, chloride	Continuous
	Glucose, BUN, creatinine	Continuous
	Magnesium, calcium, ionized calcium, carbon dioxide	Continuous
	SGOT, SGPT, total bilirubin, albumin	Continuous
Ventilation parameters	Hemoglobin	Continuous
	White blood cells count, platelets count, PTT, PT, INR, pH, PaO2, PaCO2, base excess, bicarbonate, lactate, PaO2/FiO2 ratio	Continuous
	Mechanical ventilation	Binary
	FiO2	Continuous
	Current IV fluid intake over 4h	Continuous
Medications and fluid balance	Maximum dose of vasopressor over 4h	Continuous
	Urine output over 4h	Continuous
	Cumulated fluid balance since admission (includes preadmission data when available)	Continuous



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

# The RL Paradigm

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

# Actions $a_t$ : Sepsis Management



How much fluid?



Vasopressor dose?

# Actions $a_t$ : Artificial Pancreas

Insulin Basal Rate

×

Insulin Bolus Amount



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

# Actions $a_t$ : Artificial Pancreas

Insulin Basal Rate

×

Insulin Bolus Amount

×

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
- Input: state  $s$
- Output:  $Q(s, a)$  for each action  $a \in A$

# We need a finite set of actions to choose from

$a^1$				
$a^2$			$Q(s^3, a^2)$	
$a^3$				
...				
$a^m$				
	$s^1$	$s^2$	$s^3$	$\dots$
				$s^n$

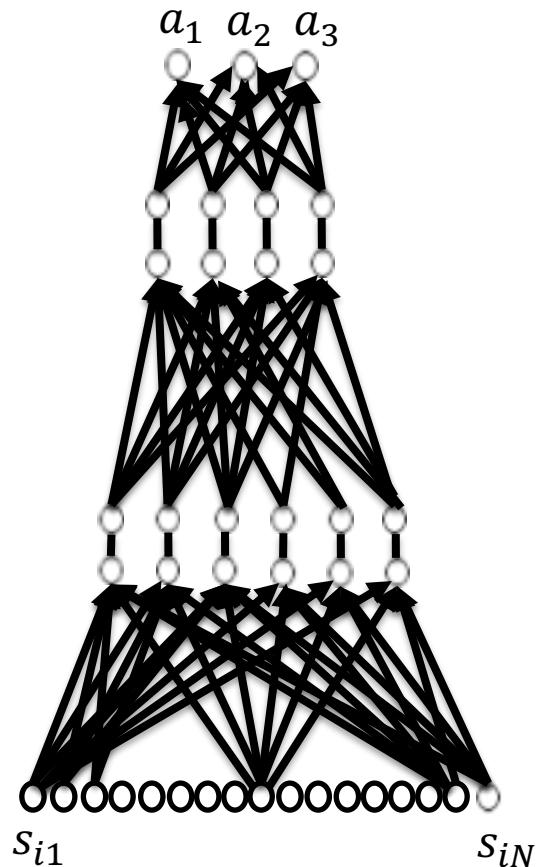
Q Learning:

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

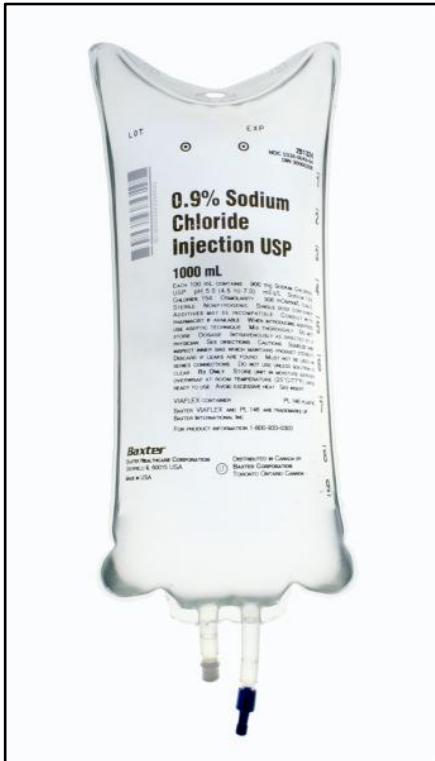
# We need a finite set of actions to choose from

## 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:  
 $\{0, \text{Quartile1}, \text{Q2}, \text{Q3}, \text{Q4}\}$

×

Five vasopressor doses  
 $\{0, \text{Quartile1}, \text{Q2}, \text{Q3}, \text{Q4}\}$

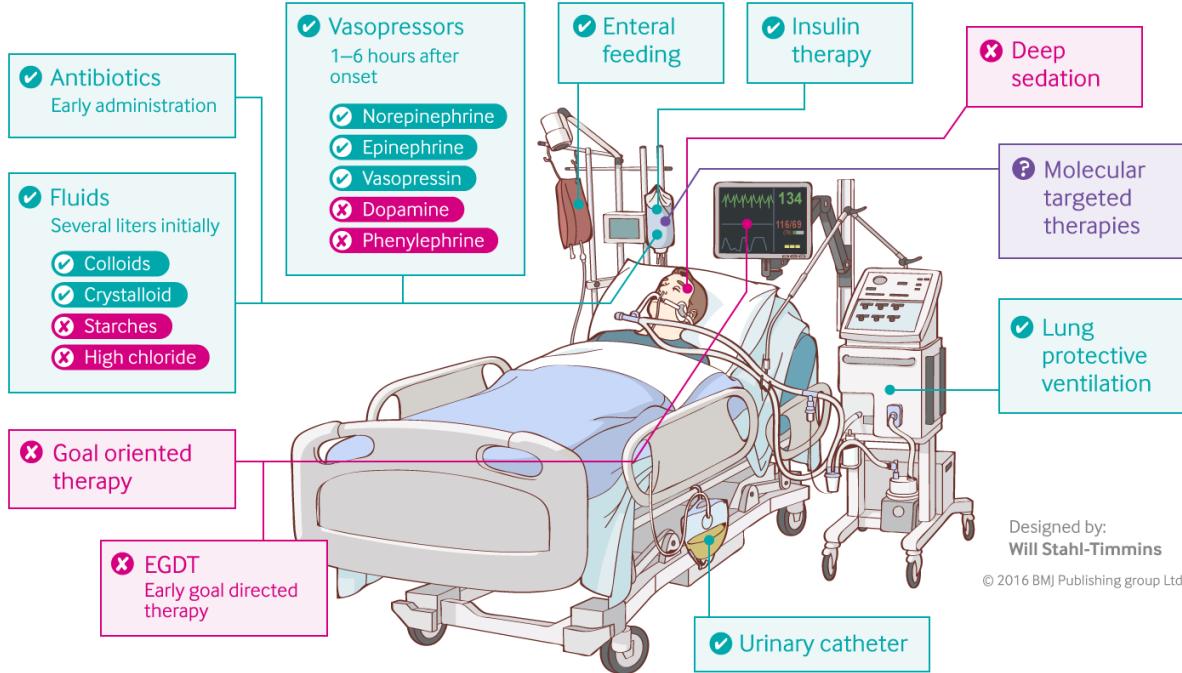
- Quartiles are defined based on the training dataset
- Example action:  
(Q2 fluid, 0 vasopressor)
- 25 possible actions

- 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  $a$
- (sometimes) A model  $P(s, a, s')$  that describes the **probability** of reaching state  $s'$  from state  $s$  after taking action  $a$

# The RL Paradigm

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

## Treating sepsis: the latest evidence



“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**

# 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 <sub>2</sub> /Fio <sub>2</sub> , mm Hg	>400	≤400	≤300	≤200†	≤100†
Coagulation Platelets × 10 <sup>3</sup> /µL‡	>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 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

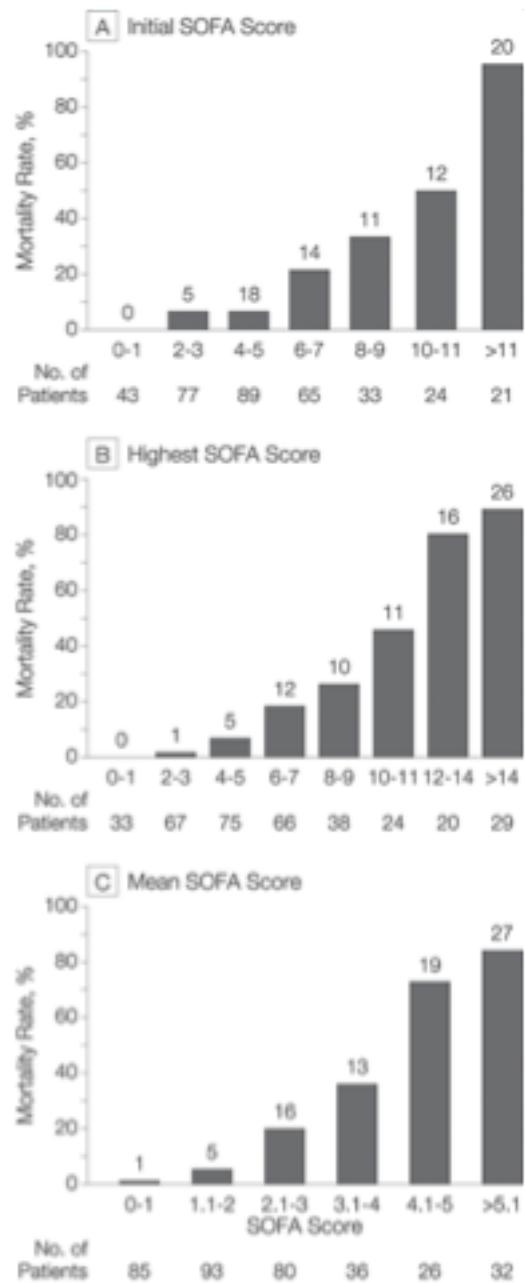
\*Norepi indicates norepinephrine; Dob, dobutamine; Dop, dopamine; Epi, epinephrine; and Fio<sub>2</sub>, fraction of inspired oxygen.

†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.

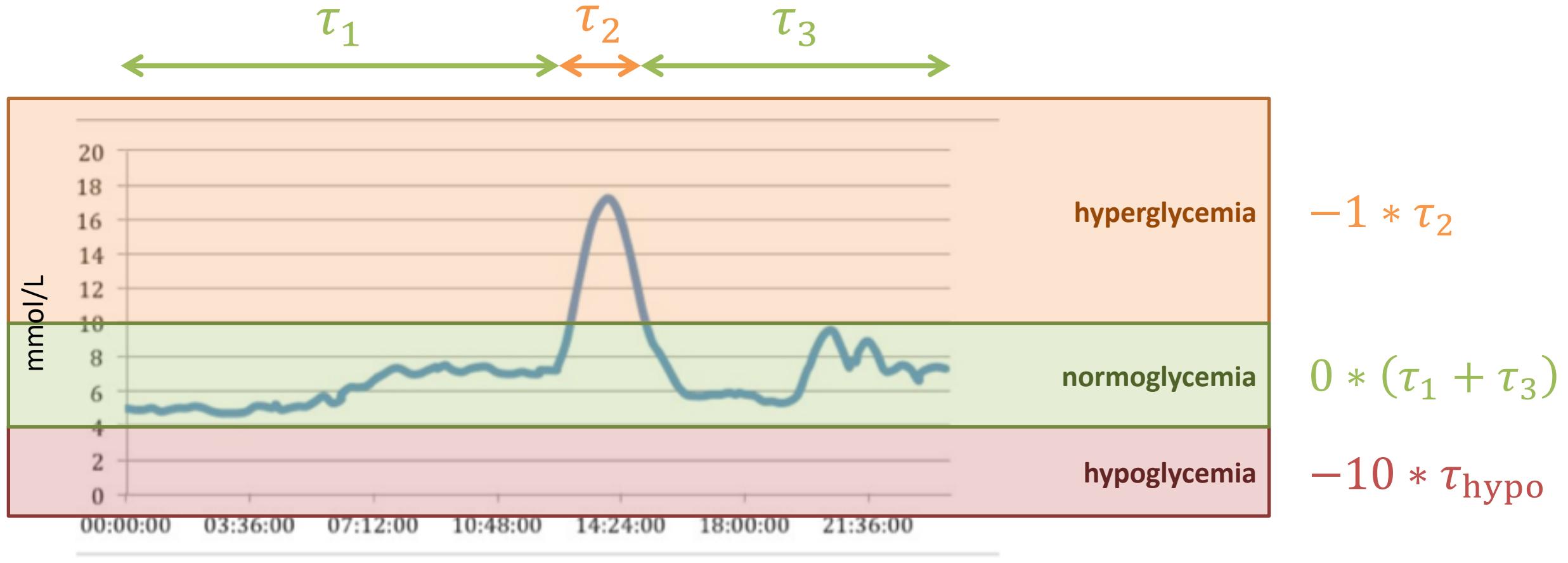


Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *JAMA*. 2001;286(14):1754–1758.

# 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



[https://medium.com/@justin\\_d\\_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e](https://medium.com/@justin_d_lawler/continuous-glucose-monitoring-the-first-four-weeks-7a6aa5fdb06e)

# The reward quantifies our objectives



*Defining Tomorrow's Medicine*



Duke Health



MEDISAVE



MEDISHIELD



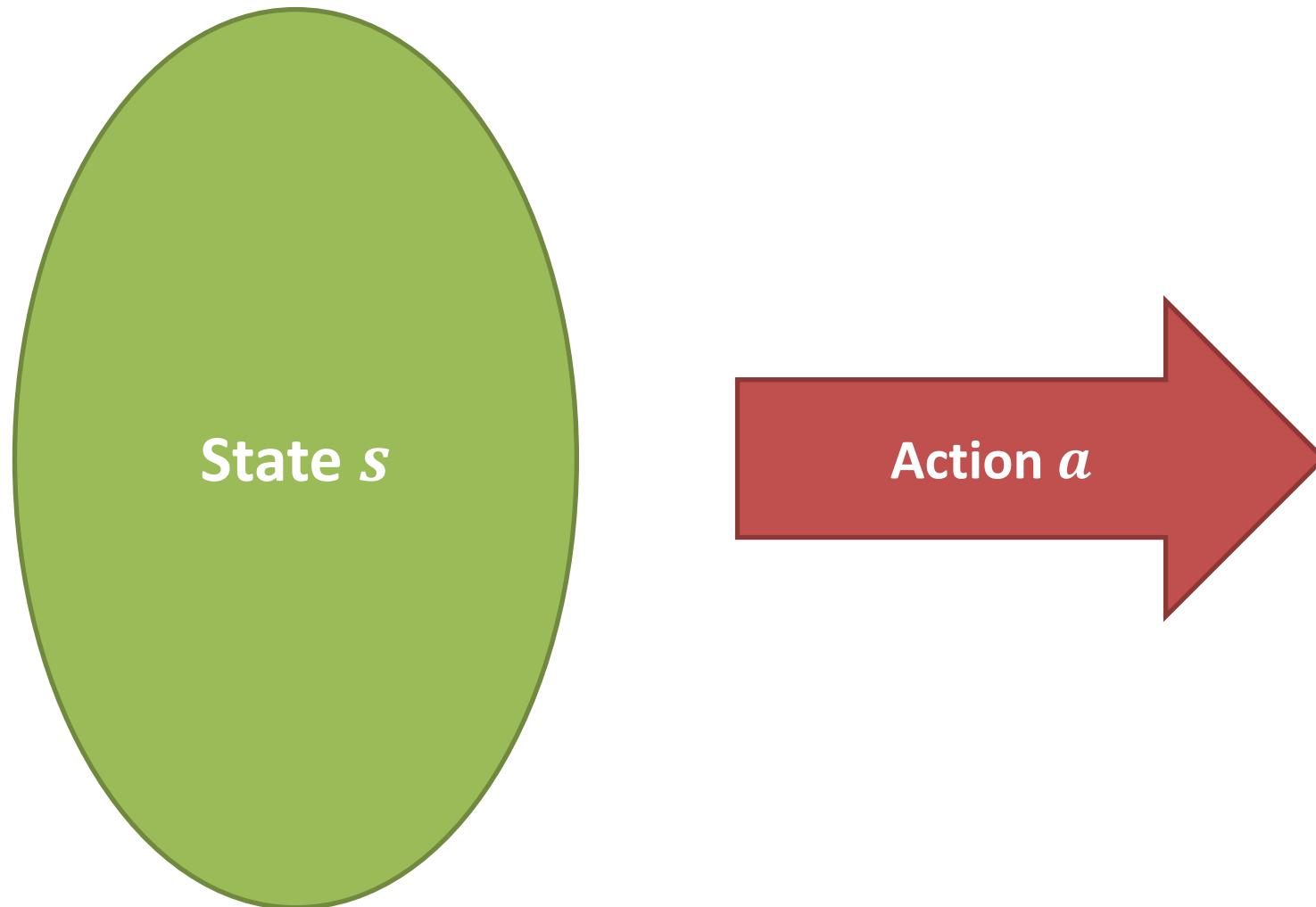
MEDIFUND



Sepsis Management and Artificial Pancreas

# **CHALLENGE 1: MISSING INFORMATION**

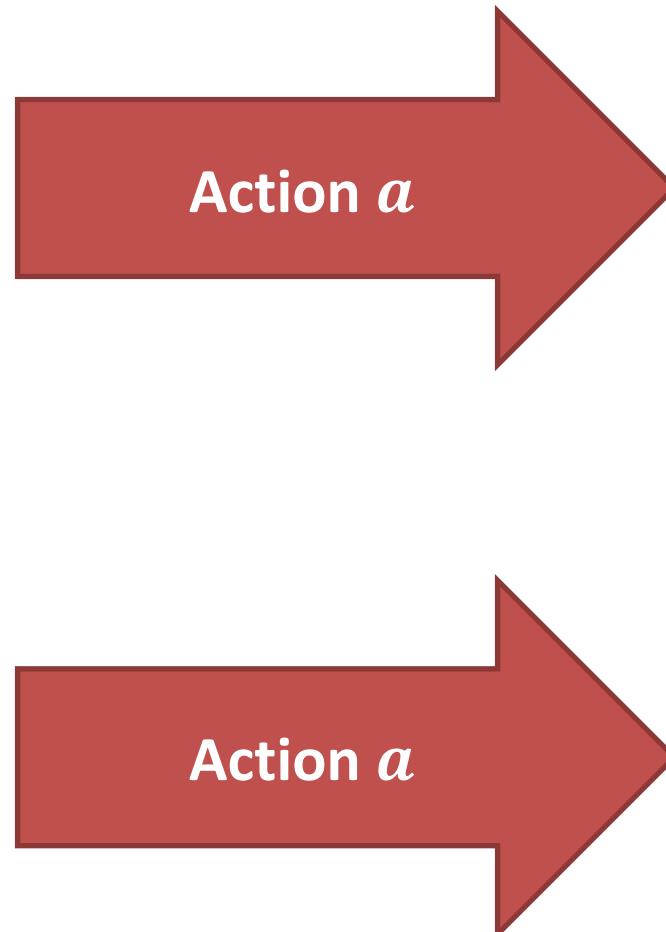
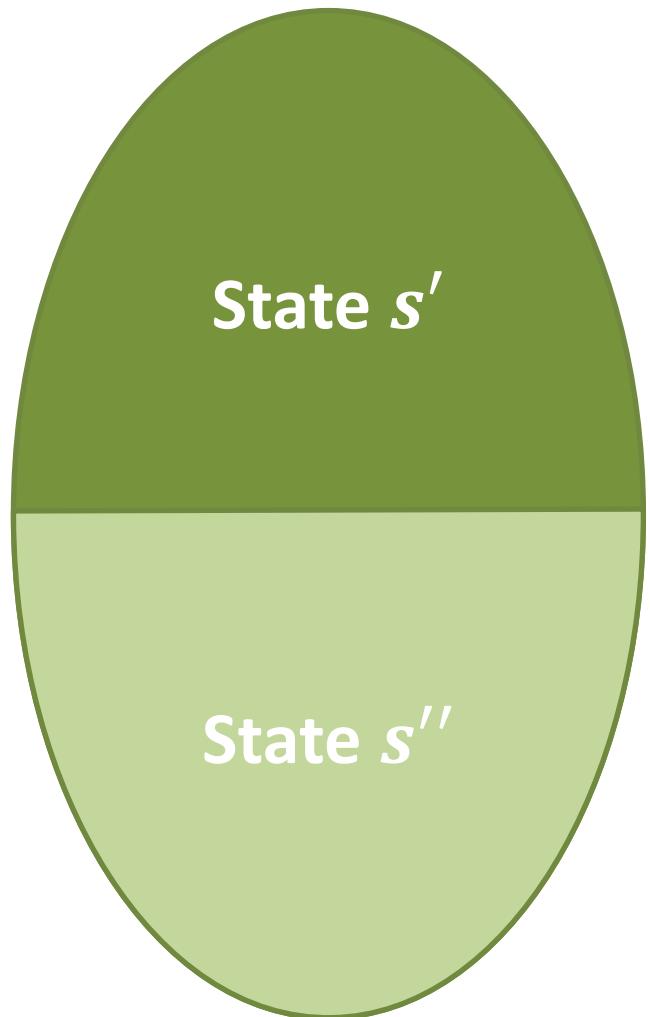
# What happens when important information is missing?



90% of the time:  
good outcome

10% of the time:  
bad outcome

# What happens when important information is missing?



90% of the time:  
good outcome

10% of the time:  
bad outcome

# Negative impact of missing state information

$a$		$Q(s, a)$ medium		

$s$

Q Learning:

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

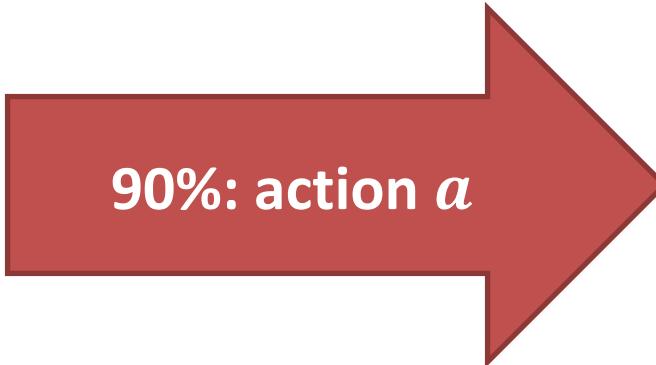
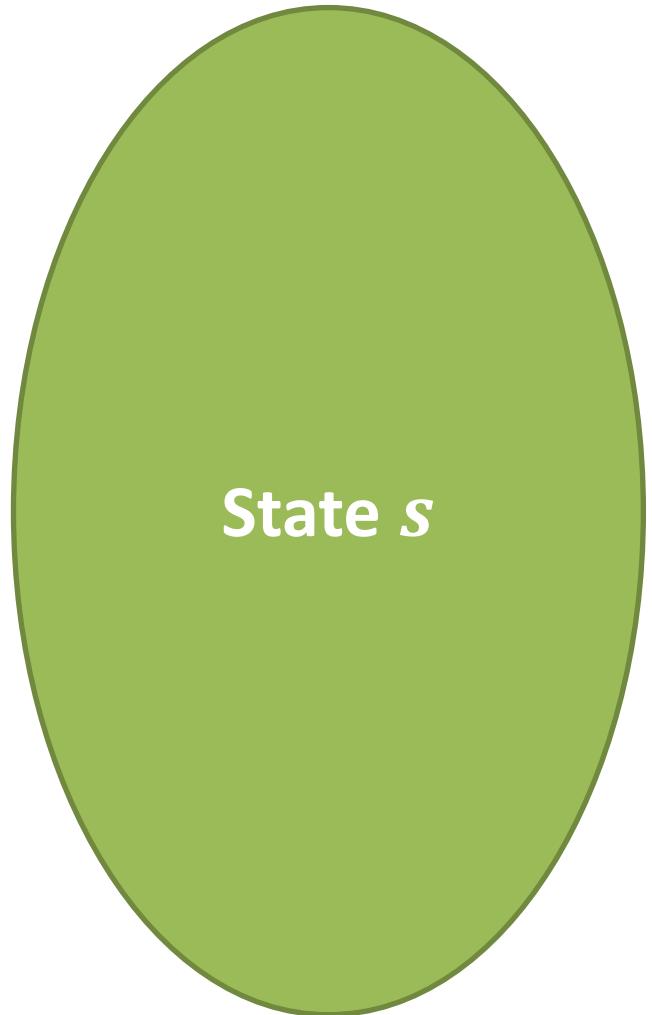
# Negative impact of missing state information

$a$		$Q(s', a)$ high	$Q(s'', a)$ low		
		$s'$	$s''$		

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

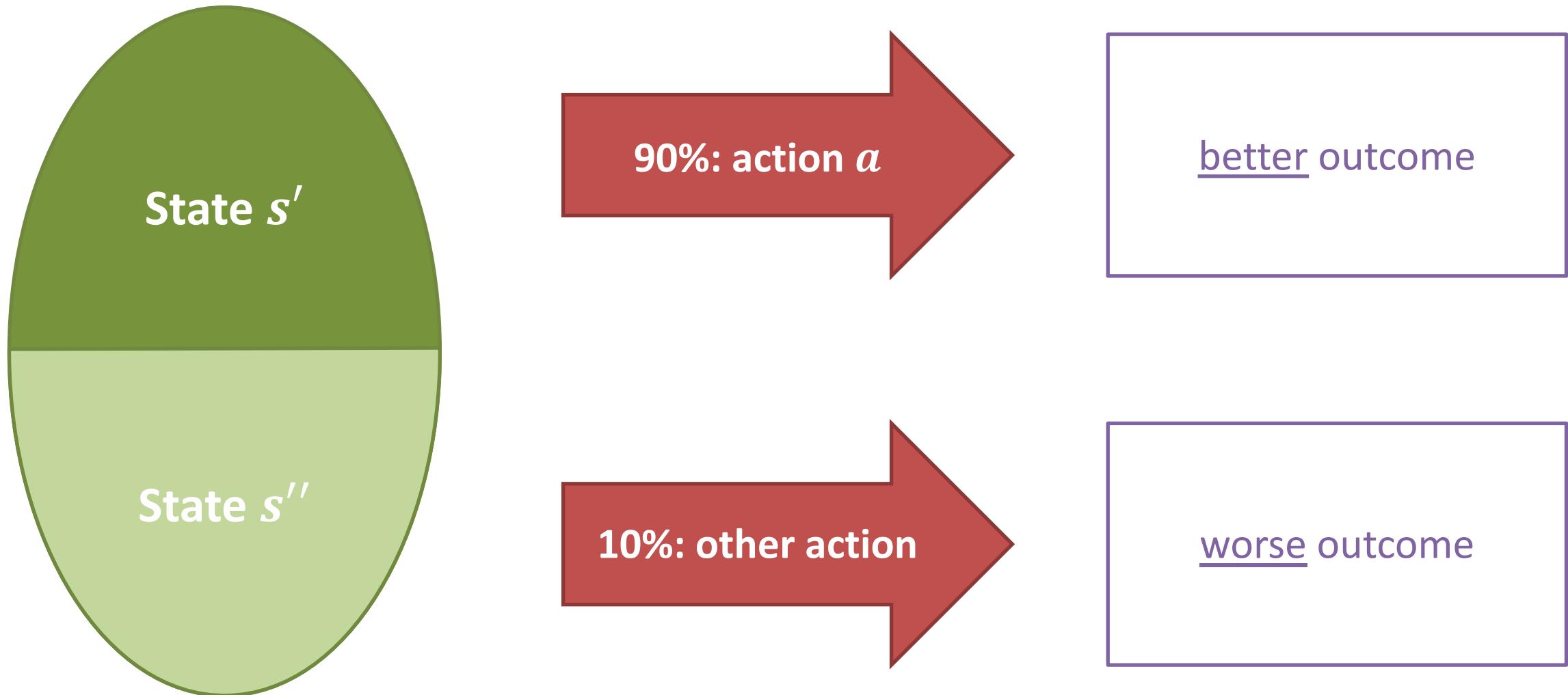


better outcome



worse outcome

# What's really happening...



# Learning from observational data: where things can REALLY go wrong

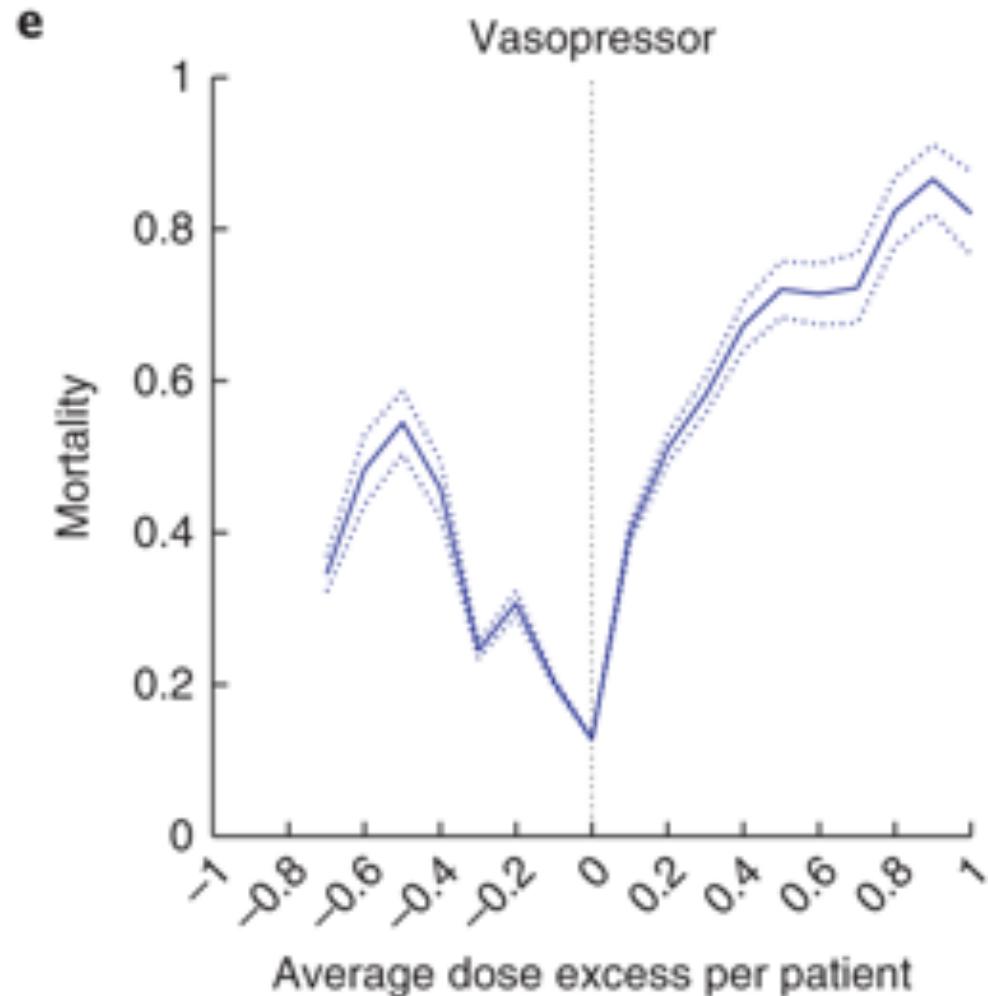
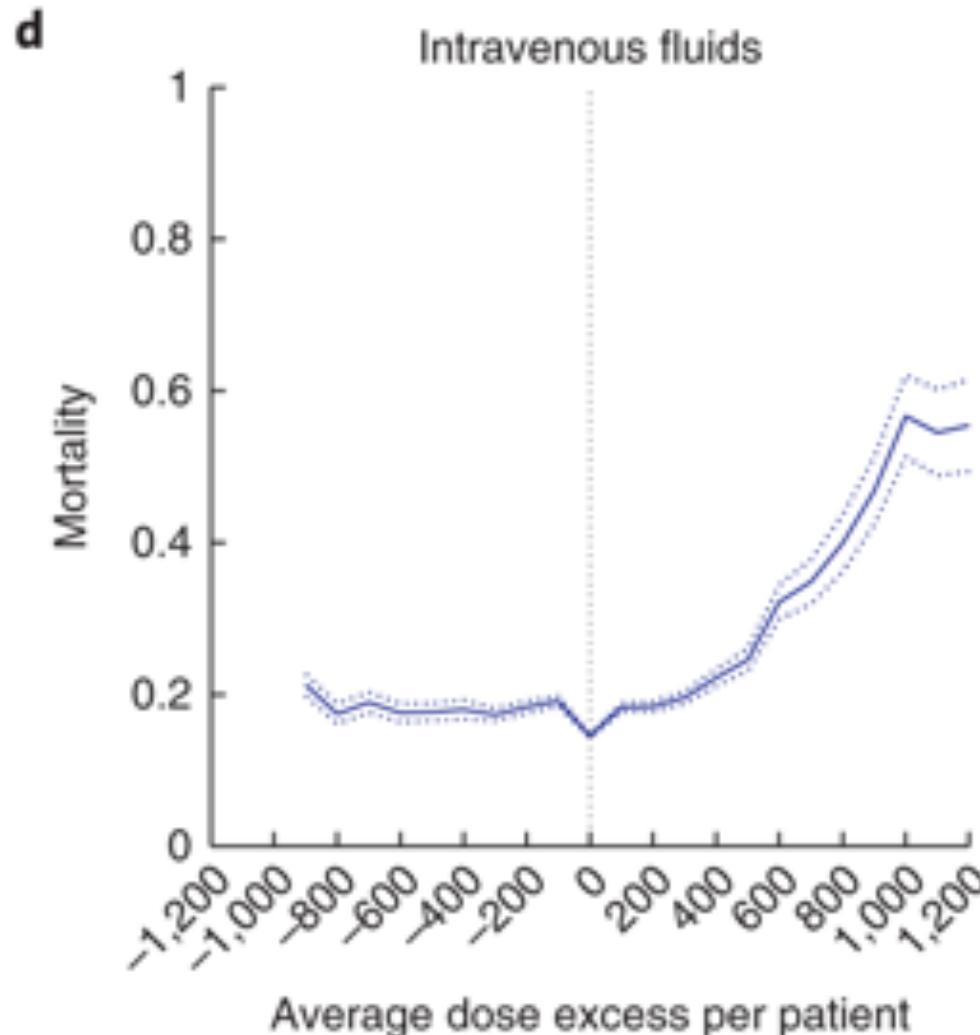
$a$		$Q(s, a)$ medium		

$s$

## Q Learning:

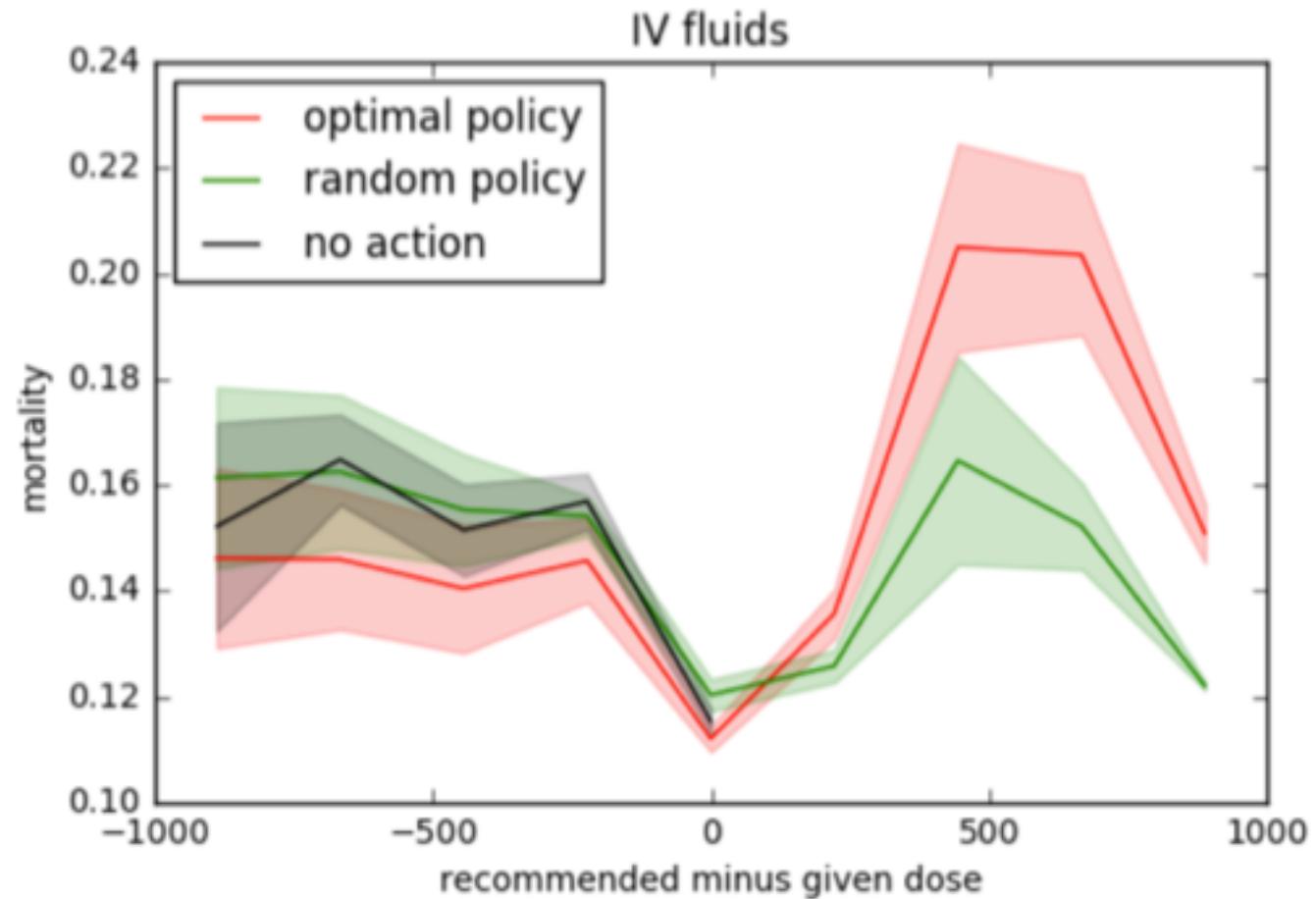
- The expected reward for choosing action  $a$  while in state  $s$  appears to be high
- In fact, this is a very bad action for some patients

# 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](http://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

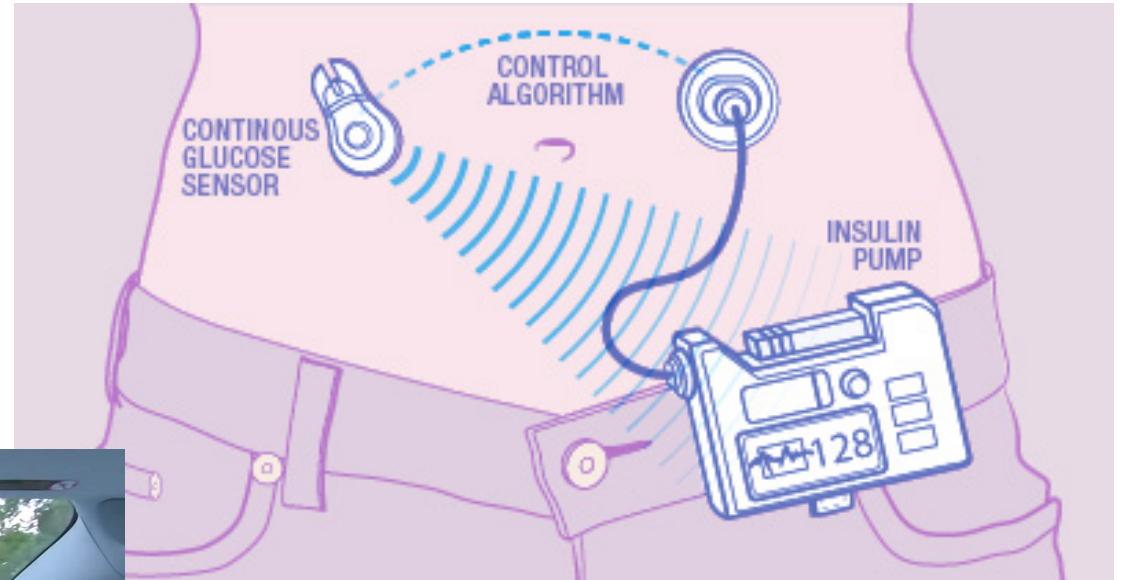
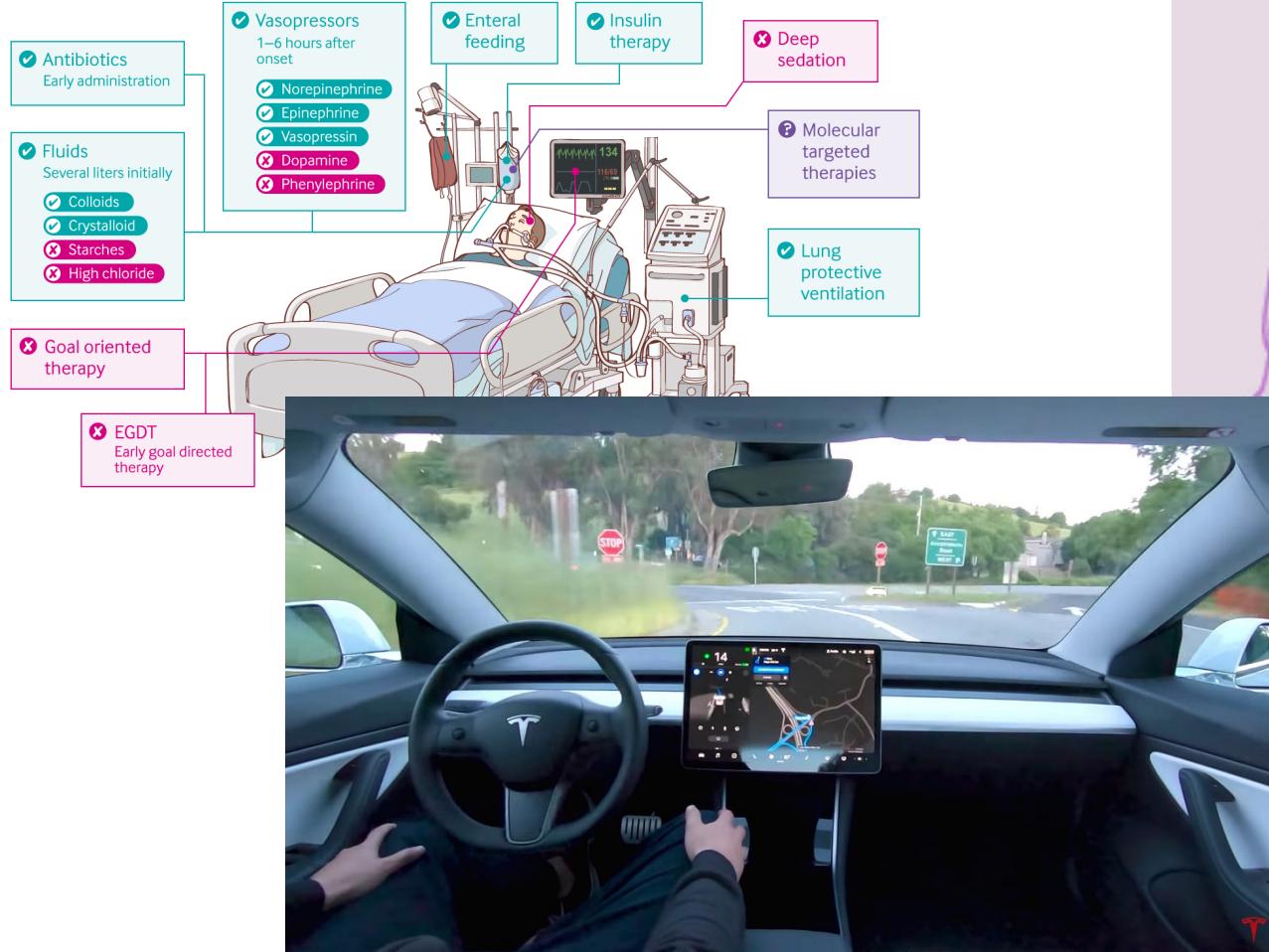
Crashing into objects has taught this drone to fly autonomously, by learning what not to do

By [Evan Ackerman](#)



# Failing 11,500 times isn't always an option

Treating sepsis: the latest evidence



-> What are the alternatives?

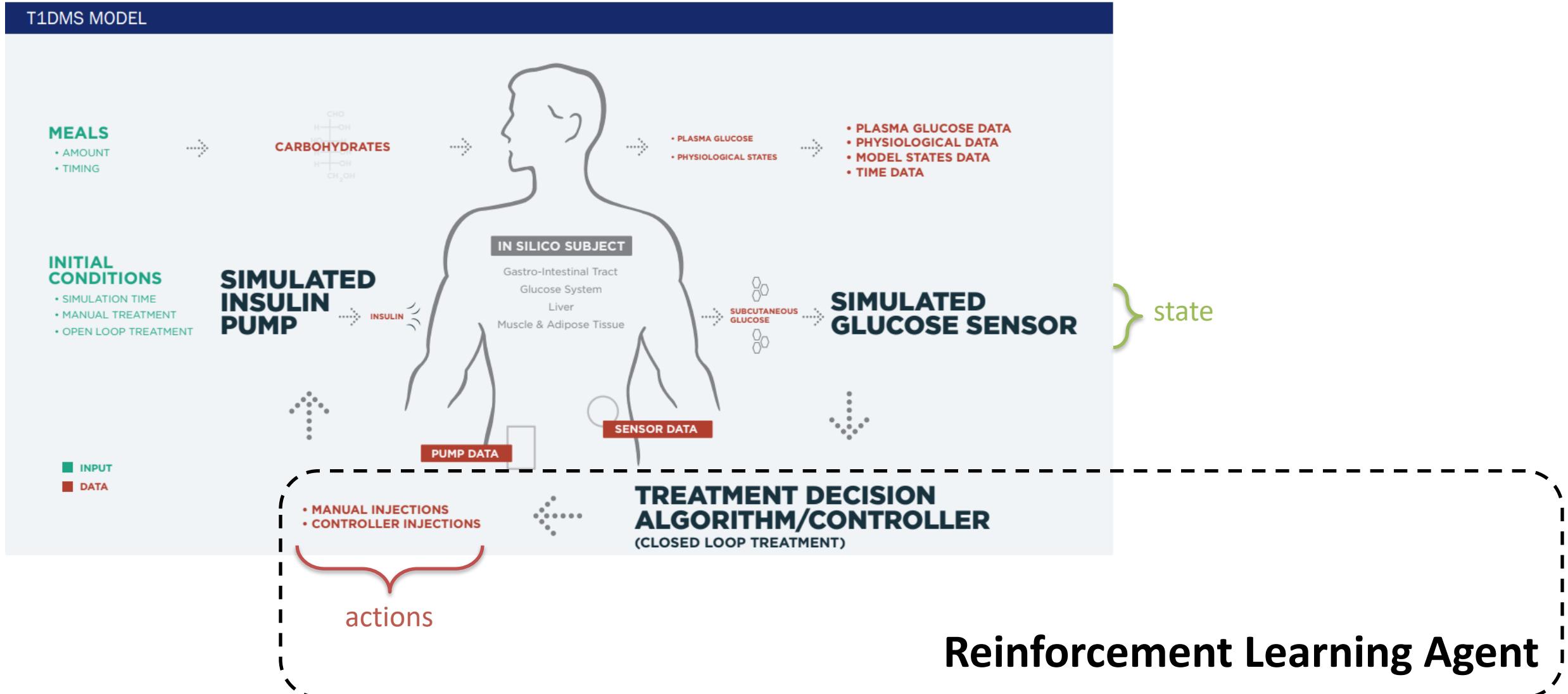
# A1. Learn from Observational Data



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

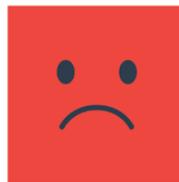
# **OTHER DIRECTIONS IN RL FOR MEDICINE**

# Sequential Decision-Making Problems are Everywhere in Medicine

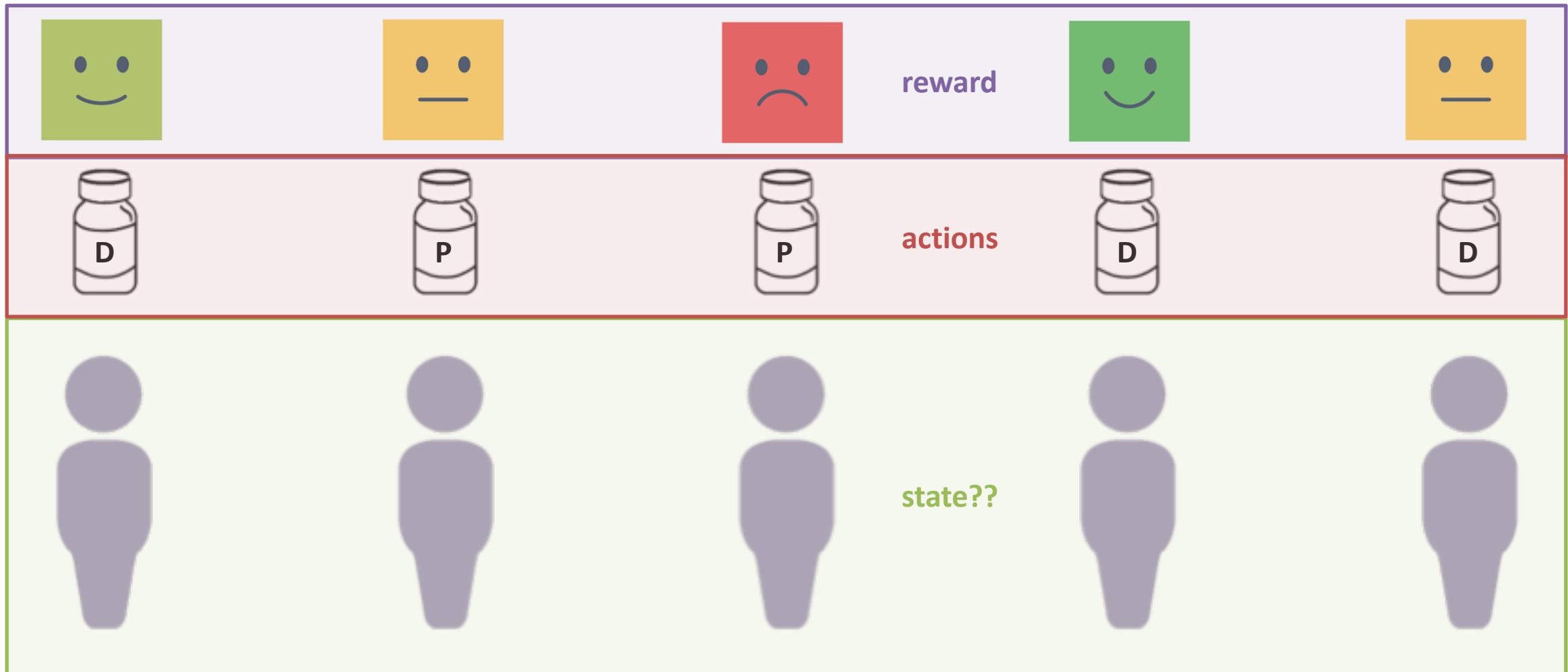
A reinforcement learning approach to weaning of mechanical ventilation in intensive care units.  
Prasad, Niranjani, et al.  
arXiv:1704.06300 (2017).



# Suppose we are evaluating a new drug...



# A special case of RL





## Application: Optimal Allocation of Clinical Trial Participants

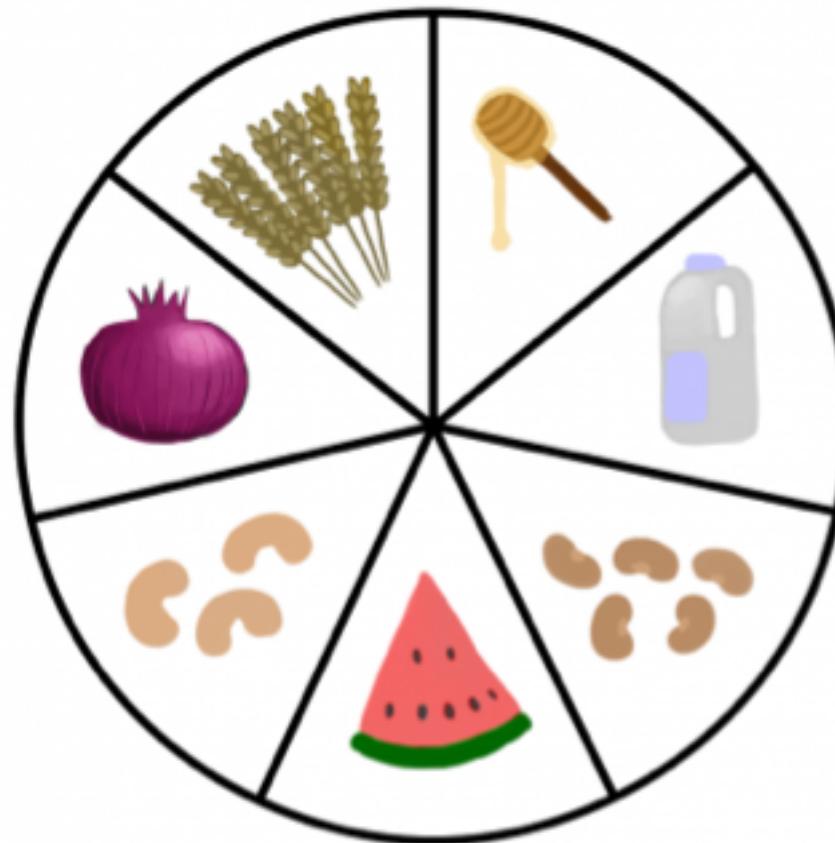
*“An explicit assumption is the goal to treat patients effectively, in the trial as well as out. That is controversial (...)"*

(Stangl, Inoue and Irony, 2012)



# N-of-1 Trial: Identify IBS Triggers

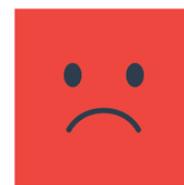
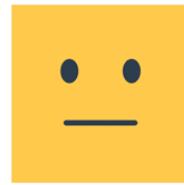
## IBS Trigger Foods



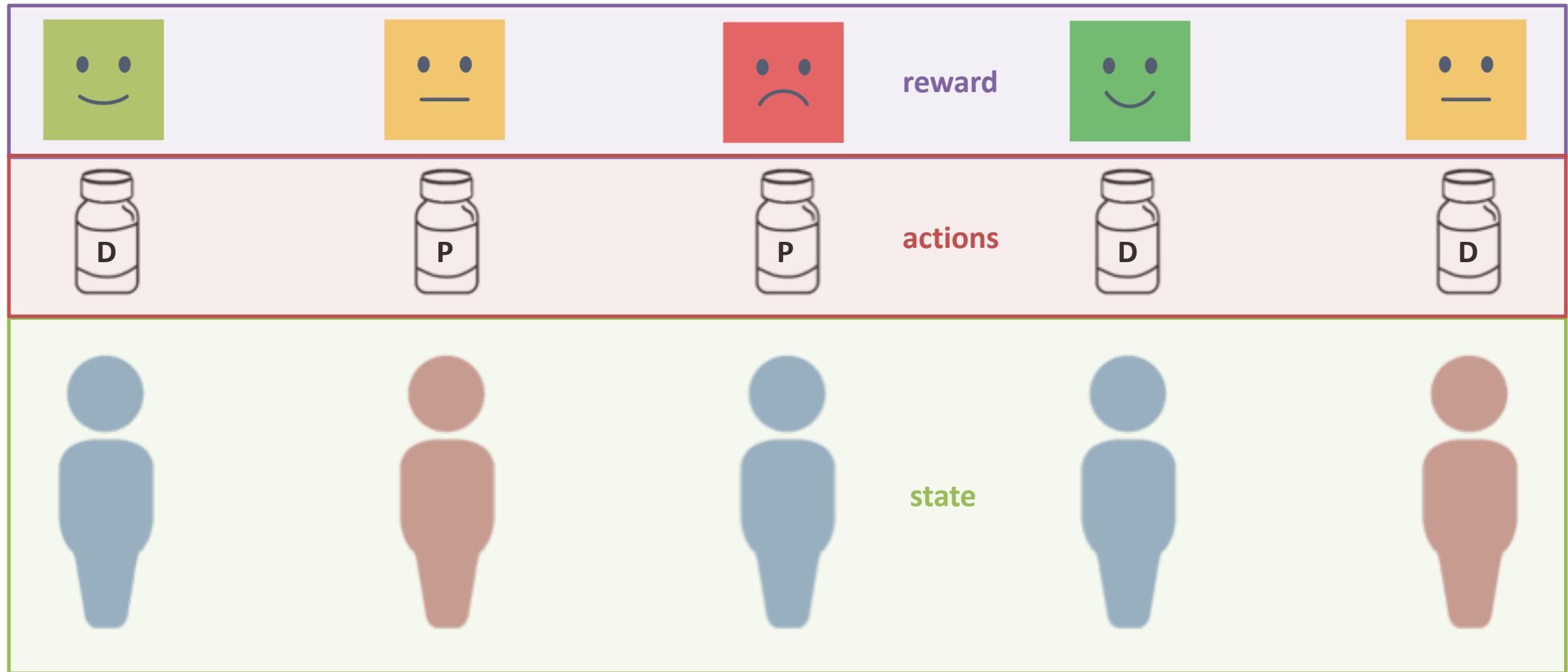
Find foods (i.e. “actions”) that minimize IBS symptoms (i.e. “reward”)

**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

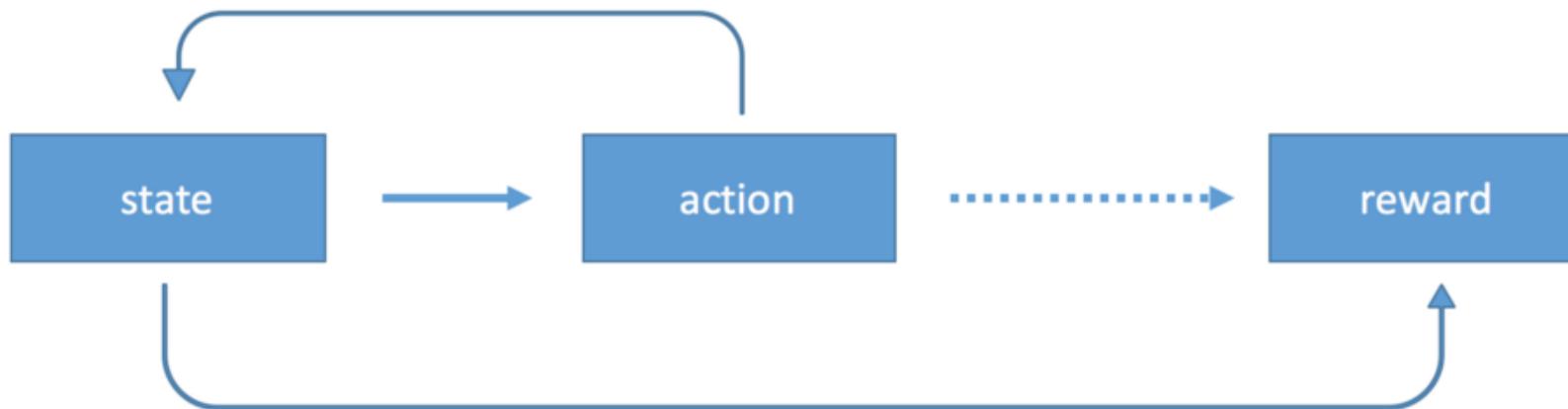




**Multi-armed Bandit**



**Contextual Bandit**



**Full RL Problem**

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

# THANK YOU!

Questions or ideas? Please contact me at [m.engelhard@duke.edu](mailto:m.engelhard@duke.edu)