

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

Sequential Medical Decision-Making: Sepsis Management

An agent

takes actions

based on the state of a system

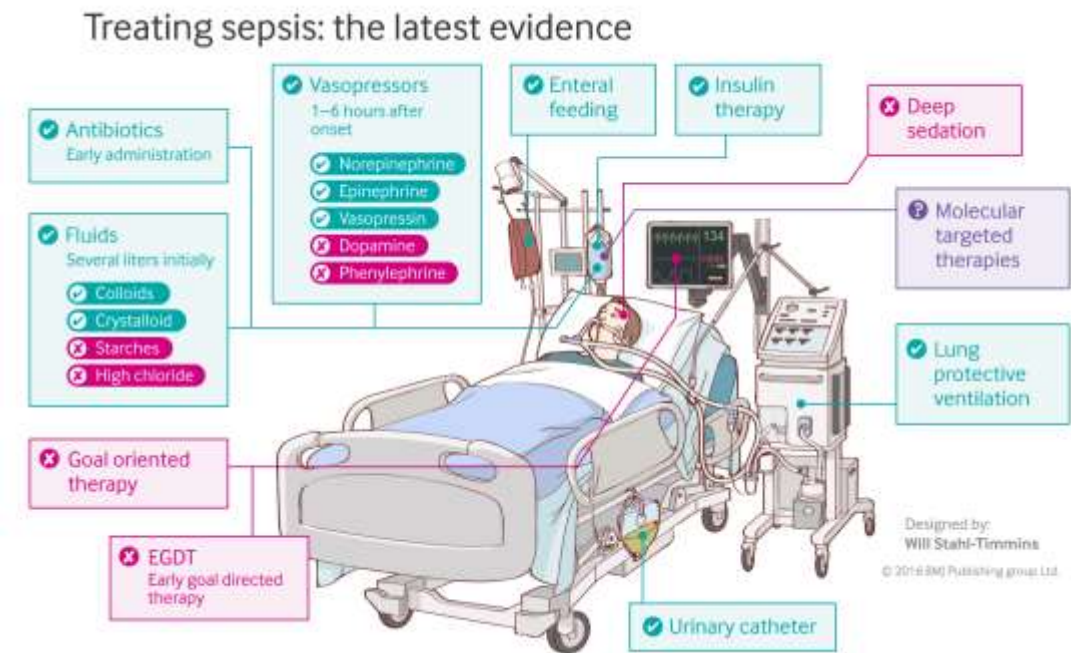
to maximize reward

A clinician

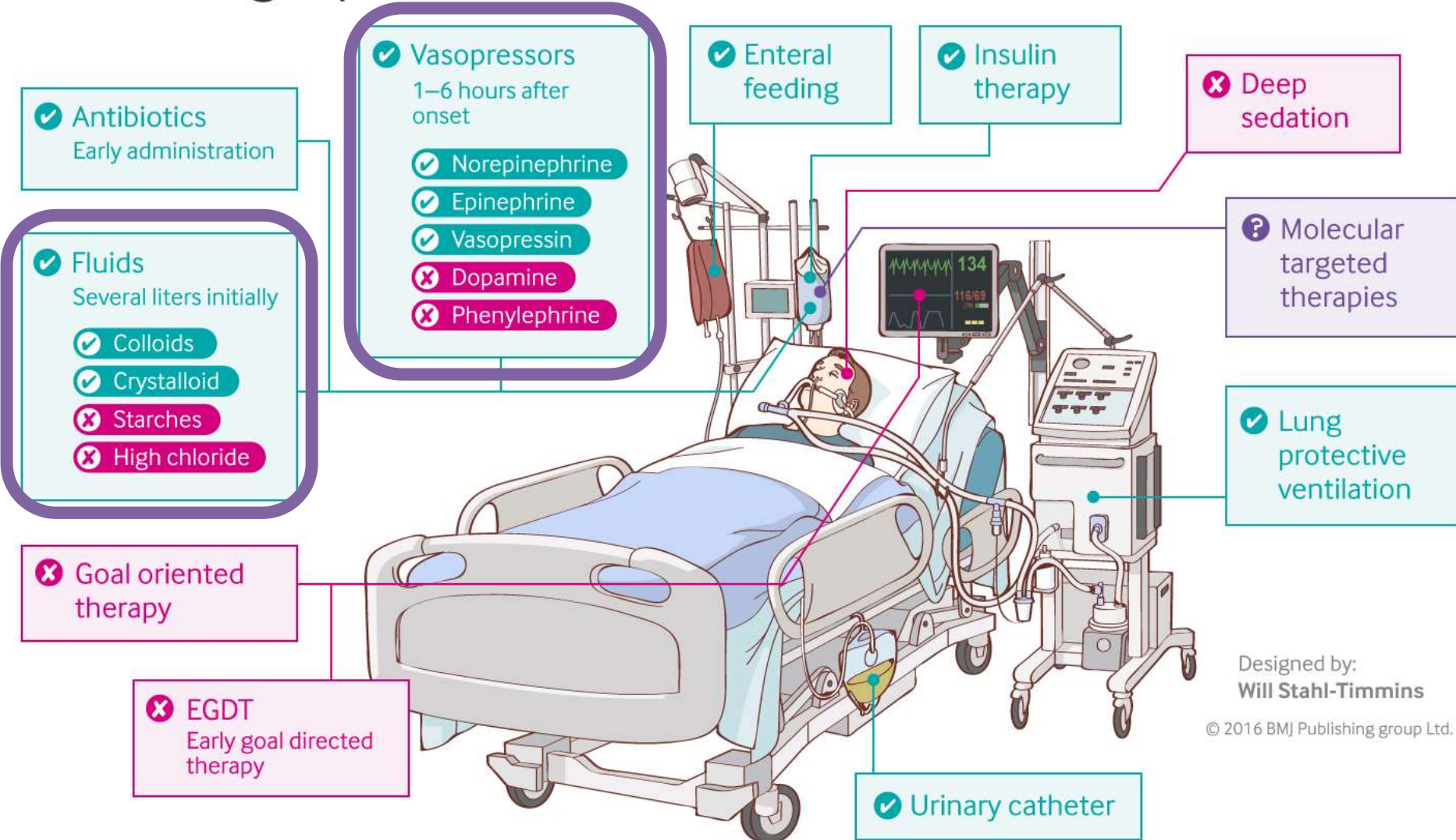
gives fluid and/or vasopressor

based on the patient's physiologic status

to maximize chance of survival



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).

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



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

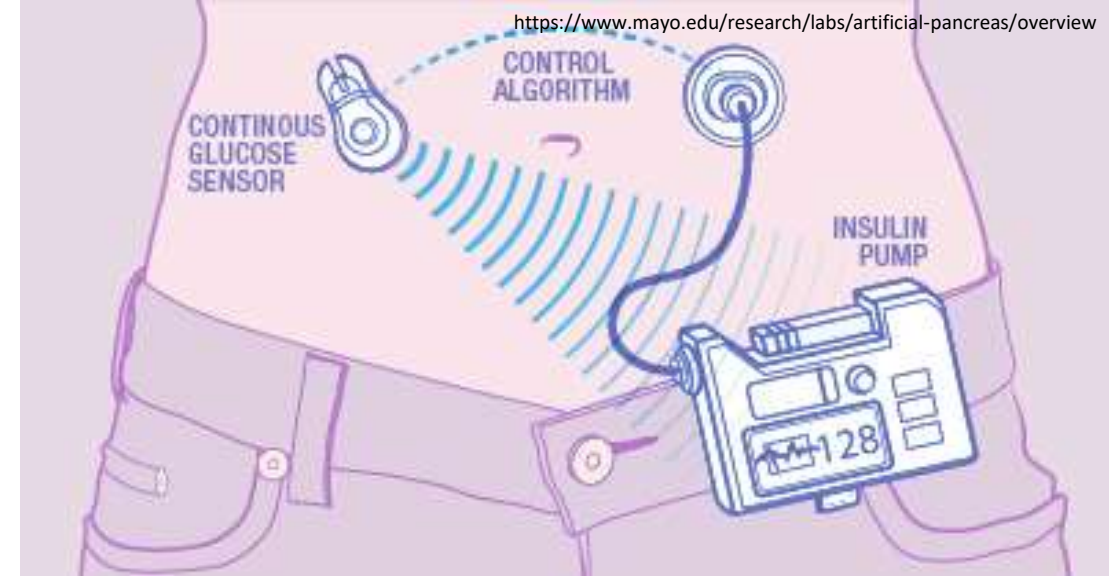
Sequential Medical Decision-Making: Artificial Pancreas

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**

 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 PROBLEM AS REINFORCEMENT LEARNING

- 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} ...

- 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

Defining Actions

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

Actions: Sepsis Management



How much fluid?



Vasopressor dose?

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	...	s^n

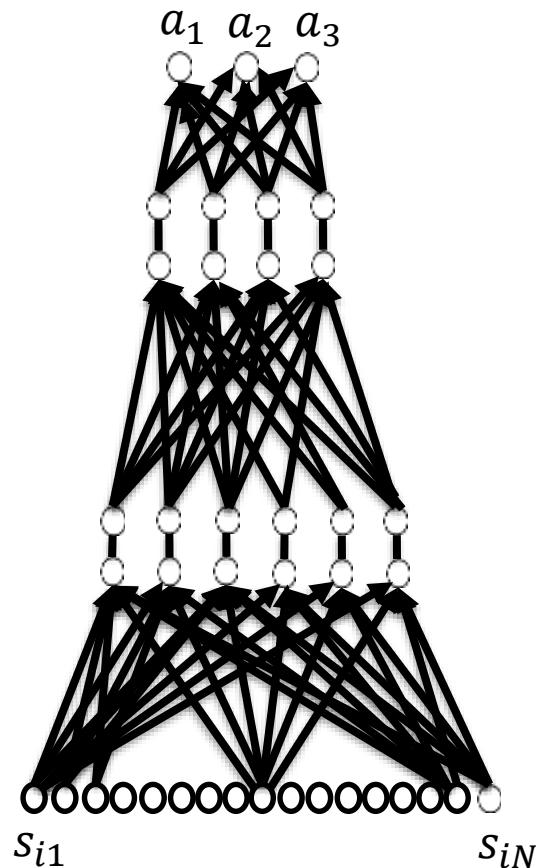
Q Learning:

- 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 $a \in A$



Actions a_t : Sepsis Management

Five IV fluid quantities:
{0, Quartile1, Q2, Q3, Q4}

×

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

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/>

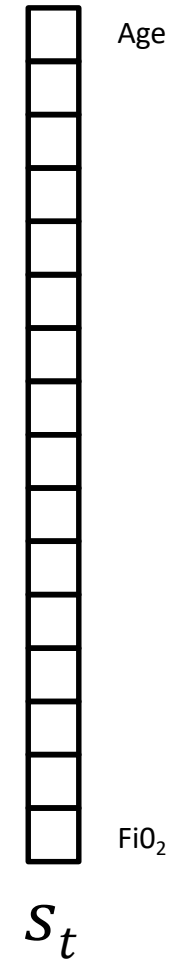
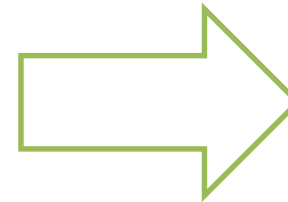
- A set of **states**: $S = \{s^1, s^2, \dots, s^n\}$
- A set of possible **actions**: $A = \{a^1, a^2, \dots, a^m\}$
- A **reward** $r(s, a, s')$ for reaching state s' from state s after taking action a
- (sometimes) A model $P(s, a, s')$ that describes the **probability** of reaching state s' from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

Defining States

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

State s_t : Sepsis

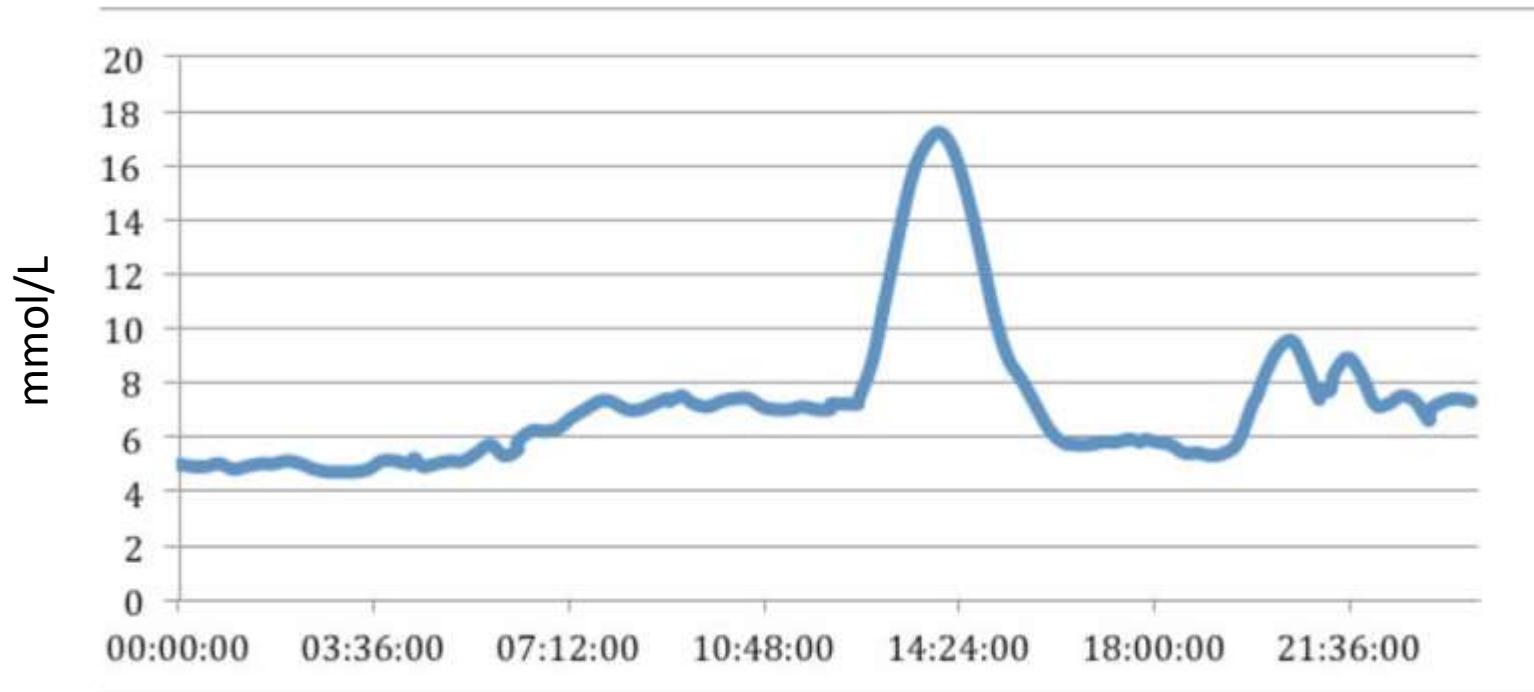
Category	Items	Type
Demographics	Age	Continuous
	Gender	Binary
	Weight	Continuous
	Readmission to intensive care	Binary
	Elixhauser score (premorbid status)	Continuous
Vital signs	Modified SOFA*	Continuous
	SIRS	Continuous
	Glasgow coma scale	Continuous
	Heart rate, systolic, mean and diastolic blood pressure, shock index	Continuous
	Respiratory rate, SpO2	Continuous
	Temperature	Continuous
Lab values	Potassium, sodium, chloride	Continuous
	Glucose, BUN, creatinine	Continuous
	Magnesium, calcium, ionized calcium, carbon dioxide	Continuous
	SGOT, SGPT, total bilirubin, albumin	Continuous
	Hemoglobin	Continuous
	White blood cells count, platelets count, PTT, PT, INR, pH, PaO2, 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

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

Blood Glucose summary statistics

- time in normo-, hypo-, and hyper-glycemia
- measures of BG variability

Recent patient behaviors

- physical activity
- eating

- 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

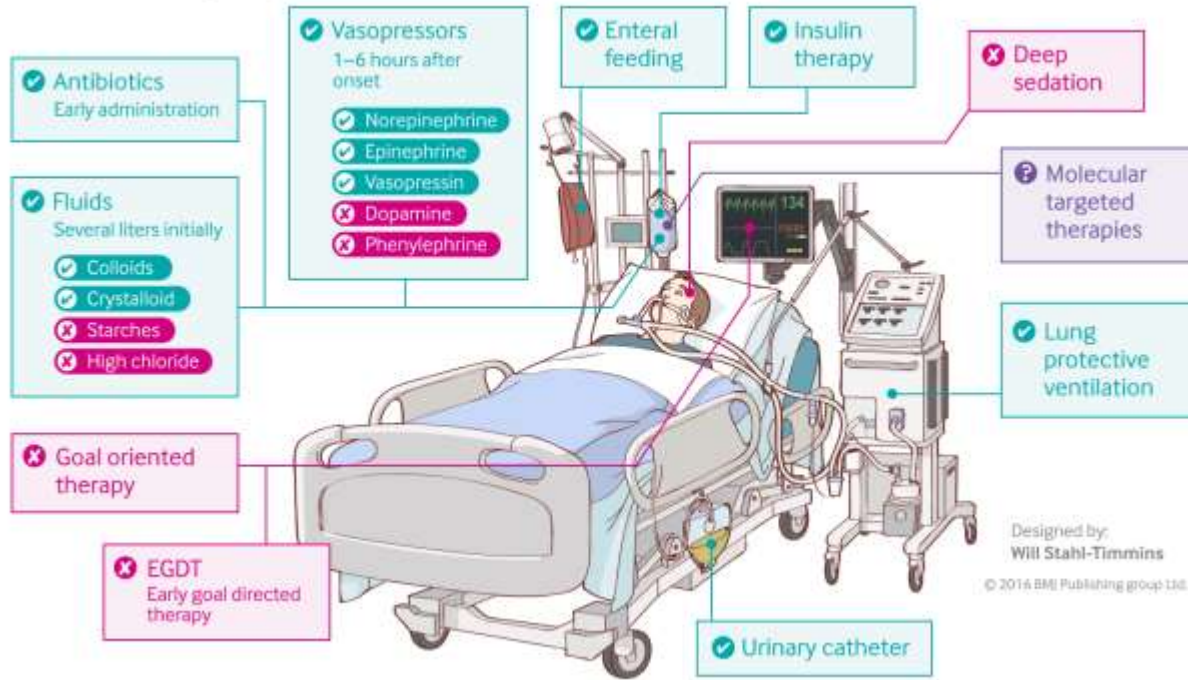
The RL Paradigm

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

The reward quantifies our objectives



Treating sepsis: the latest evidence



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

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

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 ×10 ³ /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 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

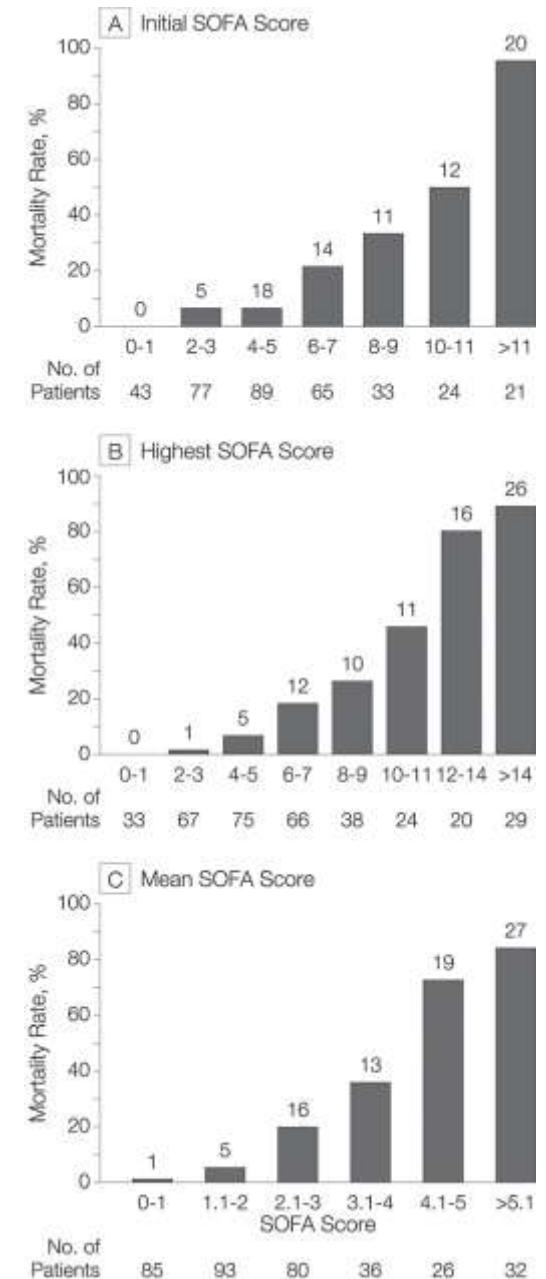
*Norepi indicates norepinephrine; Dob, dobutamine; Dop, dopamine; Epi, epinephrine; and Fio₂, fraction of inspired oxygen.

†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

$$r(s_t, s_{t+1}) = \underbrace{C_0 \mathbb{1}((s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}) \& (s_{t+1}^{\text{SOFA}} > 0))}_{\text{If SOFA score is unchanged and greater than zero, receive penalty}} + \underbrace{C_1 (s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}})}_{\text{Receive reward/penalty up to } |C_1| \text{ for decrease/increase in SOFA}} + \underbrace{C_2 \tanh(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}})}_{\text{Receive reward/penalty up to } |C_2| \text{ for decreases/increases in lactate}}$$

If SOFA score is unchanged and greater than zero, receive penalty
 Receive reward/penalty up to $|C_1|$ for decrease/increase in SOFA
 Receive reward/penalty up to $|C_2|$ for decreases/increases in lactate

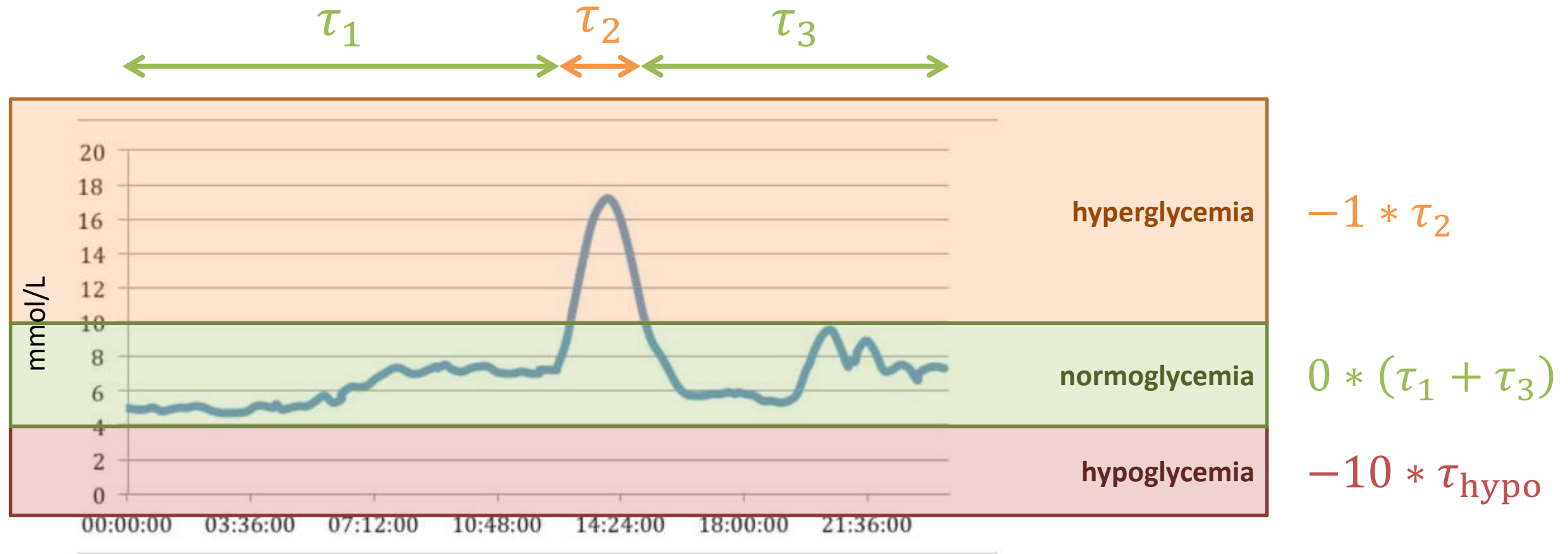
* C_0 , C_1 , and C_2 are negative, so they can be viewed as penalties

Reward r_t : Sepsis Management

$$r(s_t, s_{t+1}) = C_0 \mathbb{1} \left((s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}) \& (s_{t+1}^{\text{SOFA}} > 0) \right) + C_1 (s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}) + C_2 \tanh(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}})$$

+15 if a patient survived their ICU stay. -15 if not.

Reward r_t : Artificial Pancreas



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

RL: LEARNING THROUGH TRIAL AND ERROR

In RL, we typically learn “from scratch”:

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



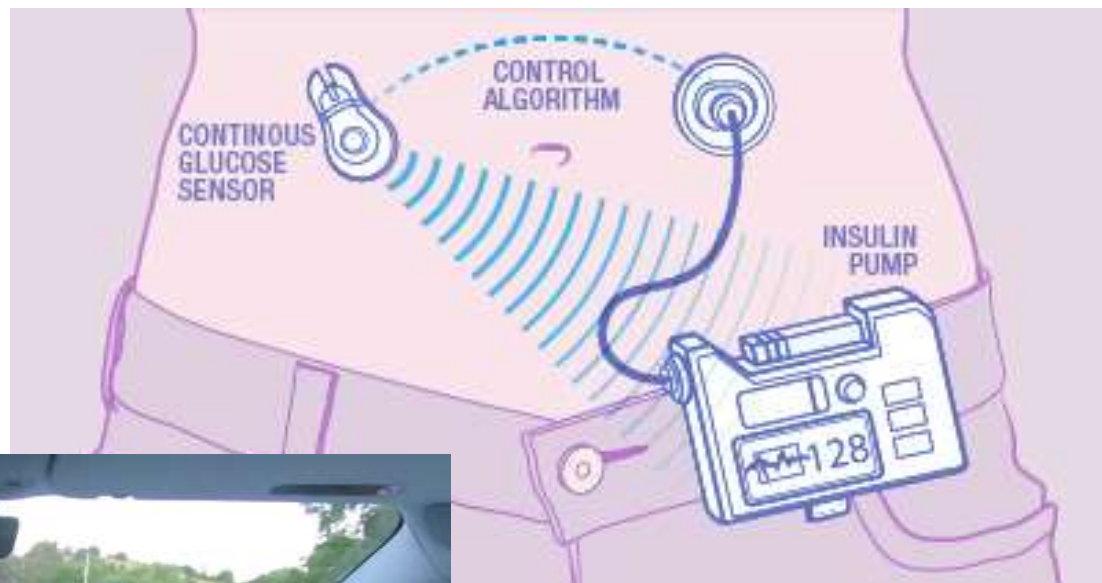
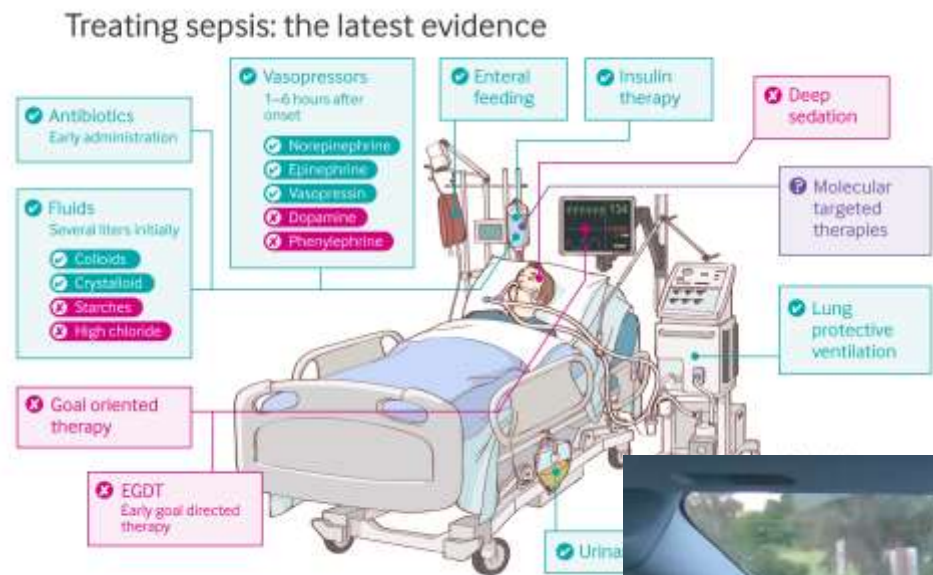
Drone Uses AI and 11,500 Crashes to Learn How to Fly

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

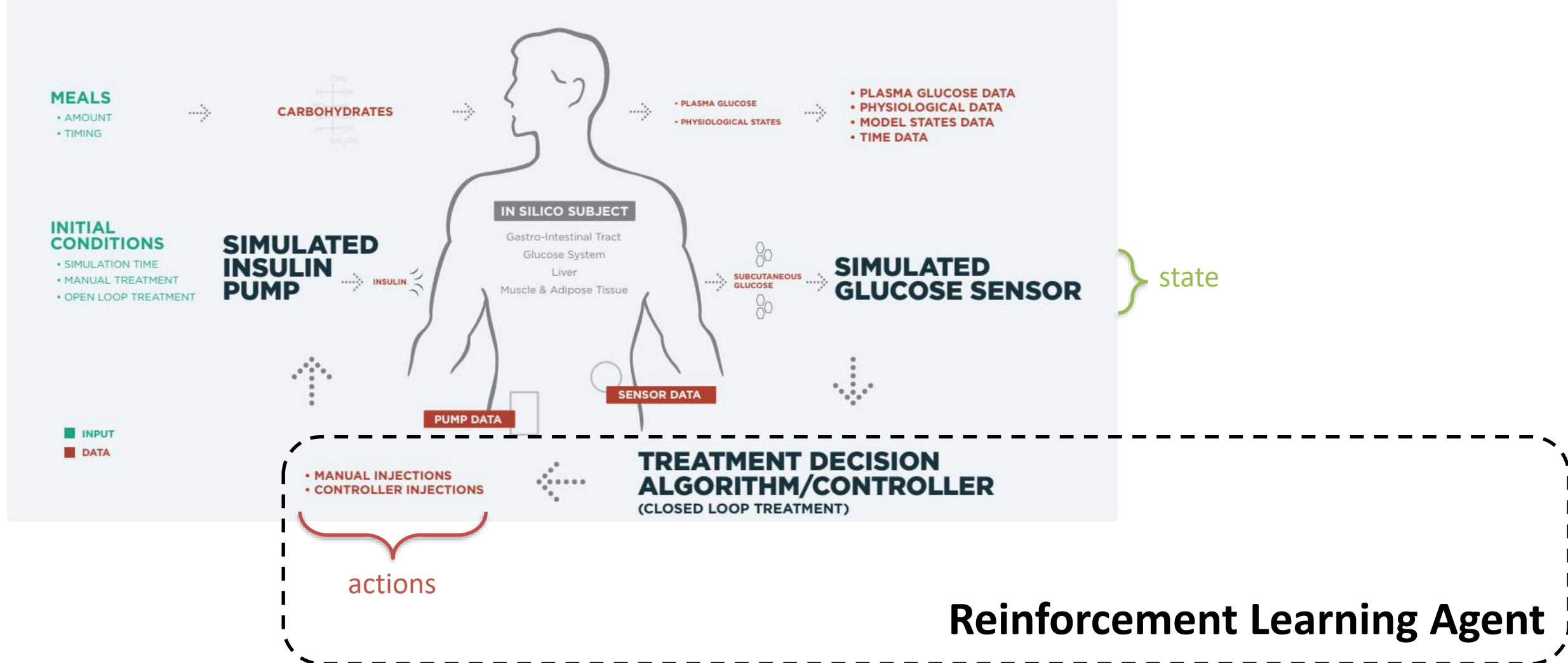


What are the alternatives?

1. Learn from observational data
 - cannot control which actions are taken
2. Learn in a simulated environment
 - state transitions may not match reality
3. Learn with expert oversight
 - high cost (e.g. clinician time)
 - may prevent algorithm from learning a better policy

Learn Policy from Simulated Environment

T1DMS MODEL



Reinforcement Learning Agent

Trial to evaluate learned policy



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?

Sepsis Management and Artificial Pancreas

EVALUATING PERFORMANCE

Evaluation Measures

- Classification performance?
- Mean Square Error?

GOAL:

Learn a policy $\pi: S \rightarrow A$

that assigns each state s to the action a that
maximizes expected reward over time

HOW?

- experience/try things and record what happens
- reinforce actions that are rewarding
- discourage actions that are expensive or lead to poor outcomes

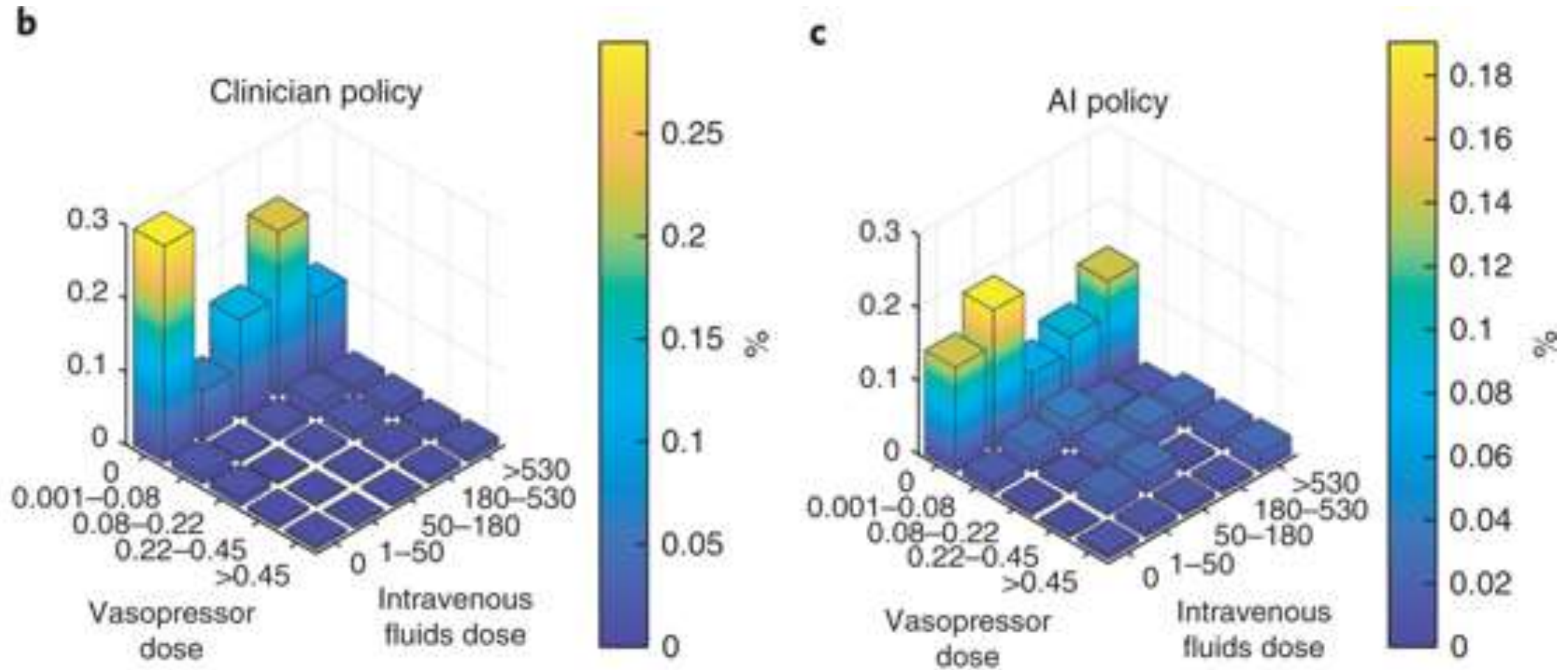
The RL Paradigm

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

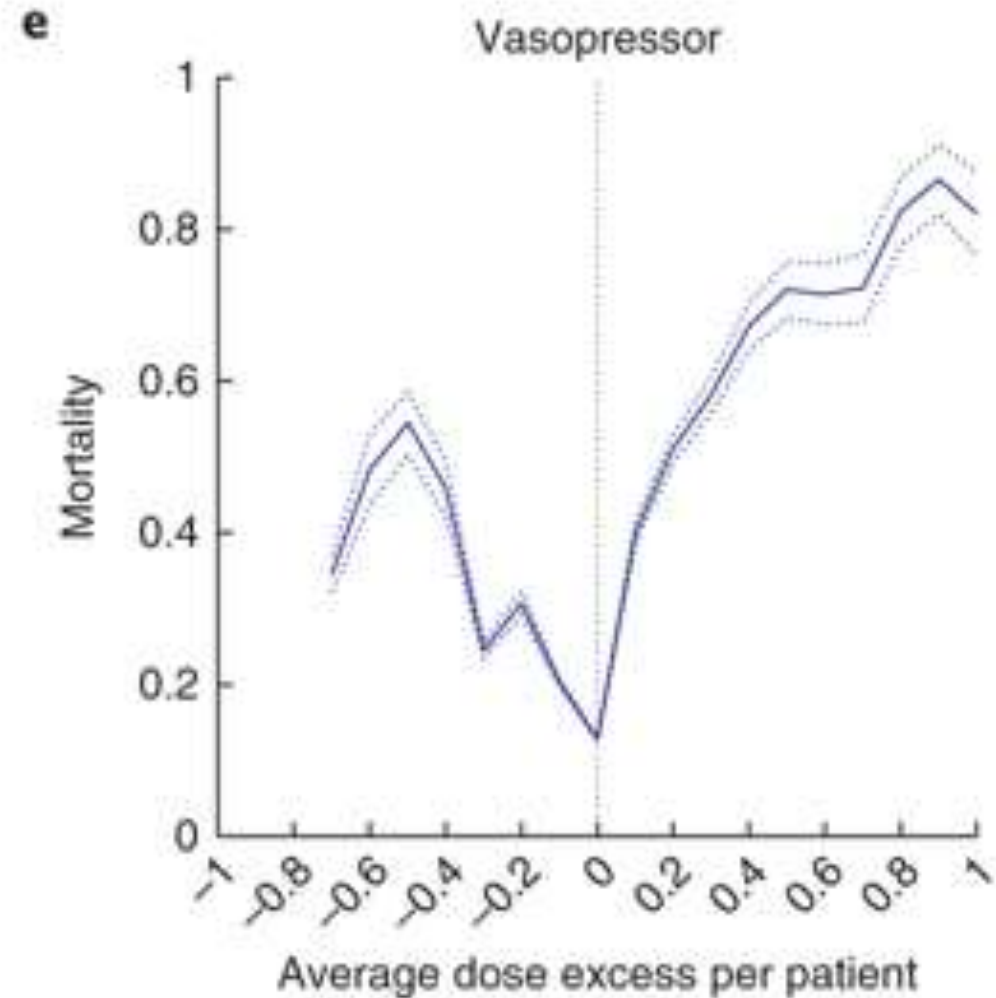
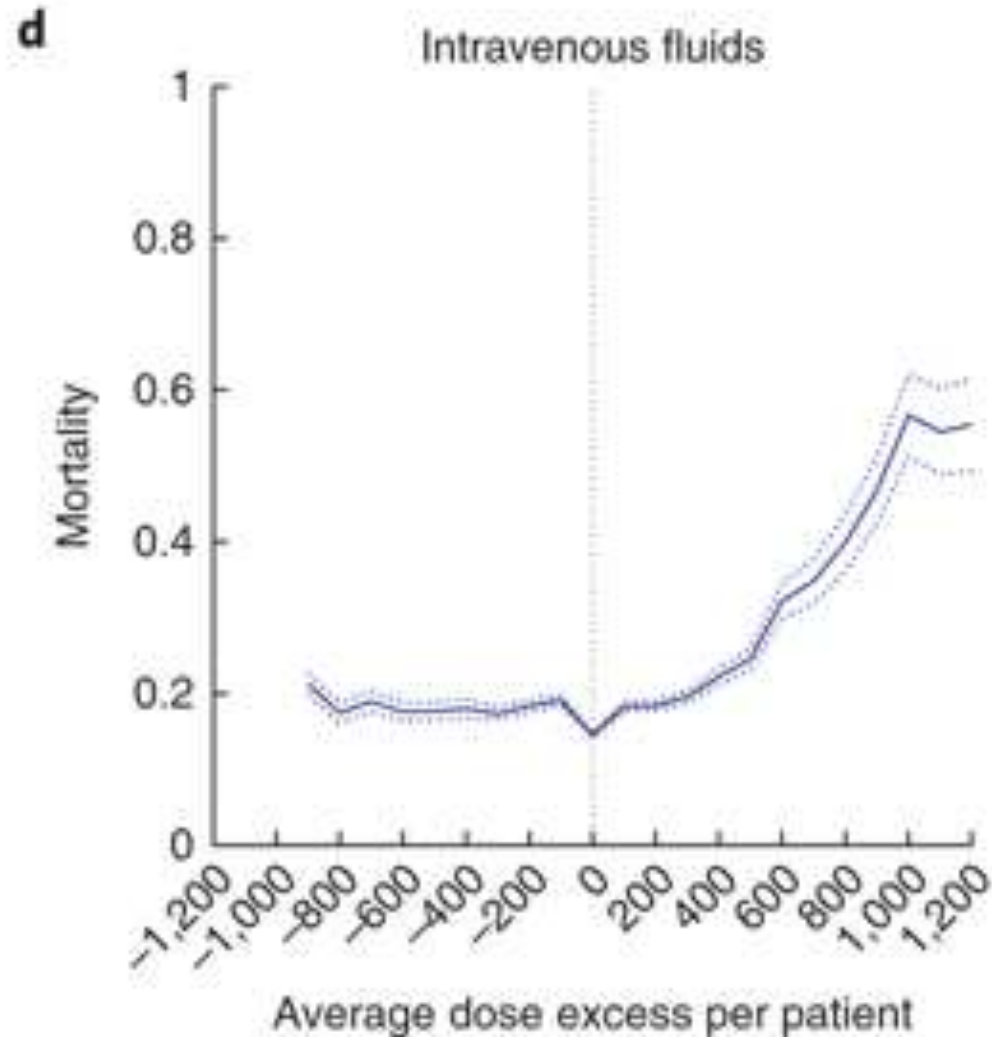
Evaluation Measures

- Classification performance?
- Mean Square Error?
- Ideally, we compare outcomes between competing policies
- In this case, the policy the algorithm is competing against is real physician decision-making
- Physician policy can be directly evaluated, but the learned policy – since it was not actually followed – can only be indirectly evaluated (“off-policy evaluation”)

Sepsis Results: Learned Policy vs Physician Policy

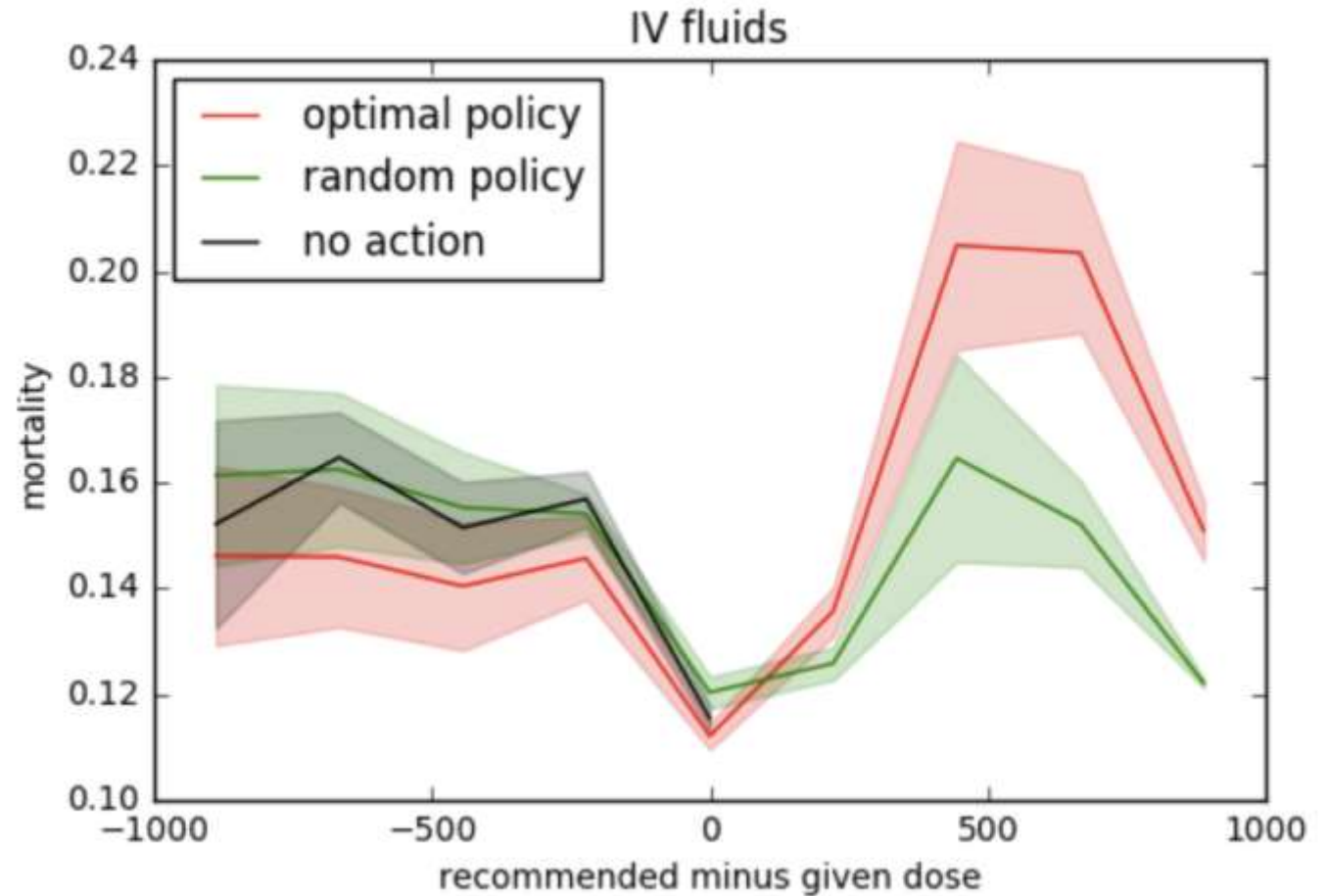


Sepsis Results: Observed Mortality



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)

Clinical Trials and Self-Experimentation

OTHER DIRECTIONS IN RL FOR MEDICINE

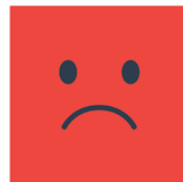
Sequential Decision-Making Problems are Everywhere in Medicine

**A reinforcement
learning approach to
weaning of mechanical
ventilation in intensive
care units.**

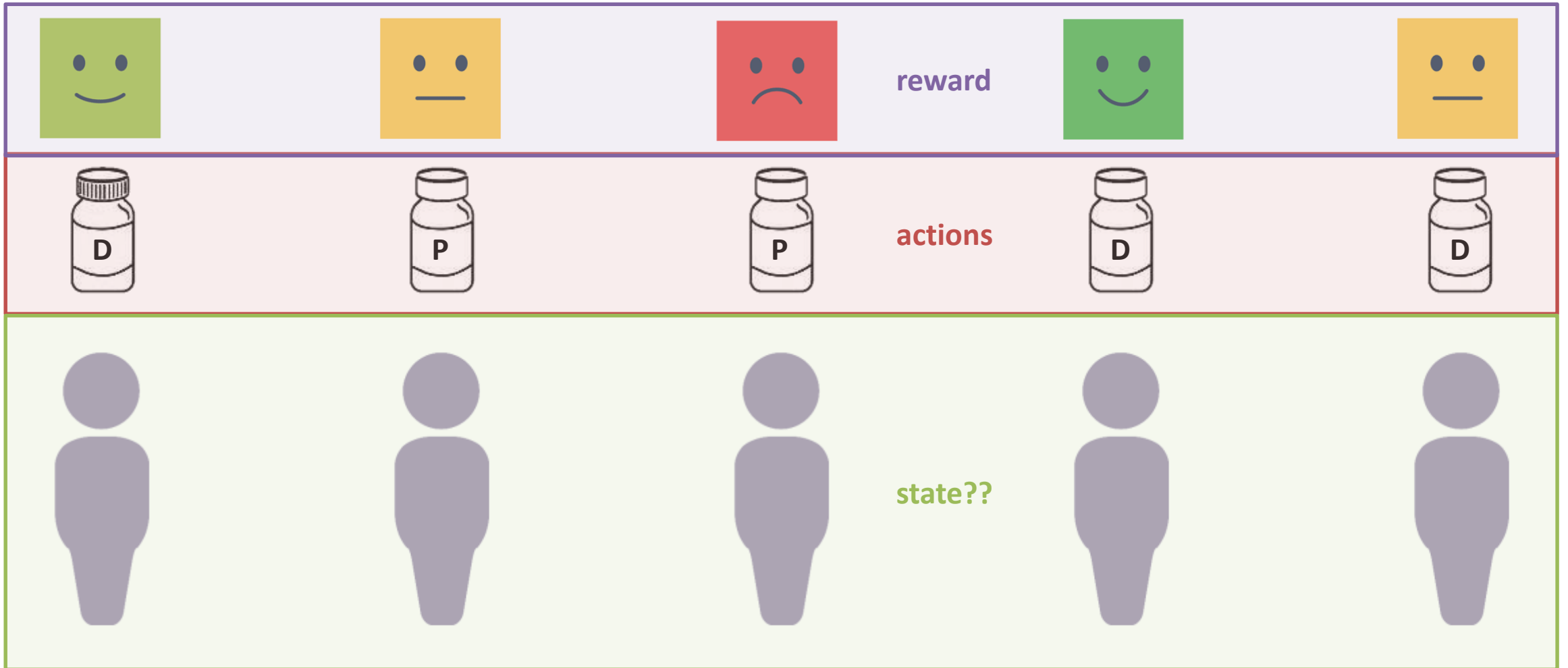
Prasad, Niranjani, et al.
arXiv:1704.06300 (2017).



Suppose we are evaluating a new drug...



A special case of RL

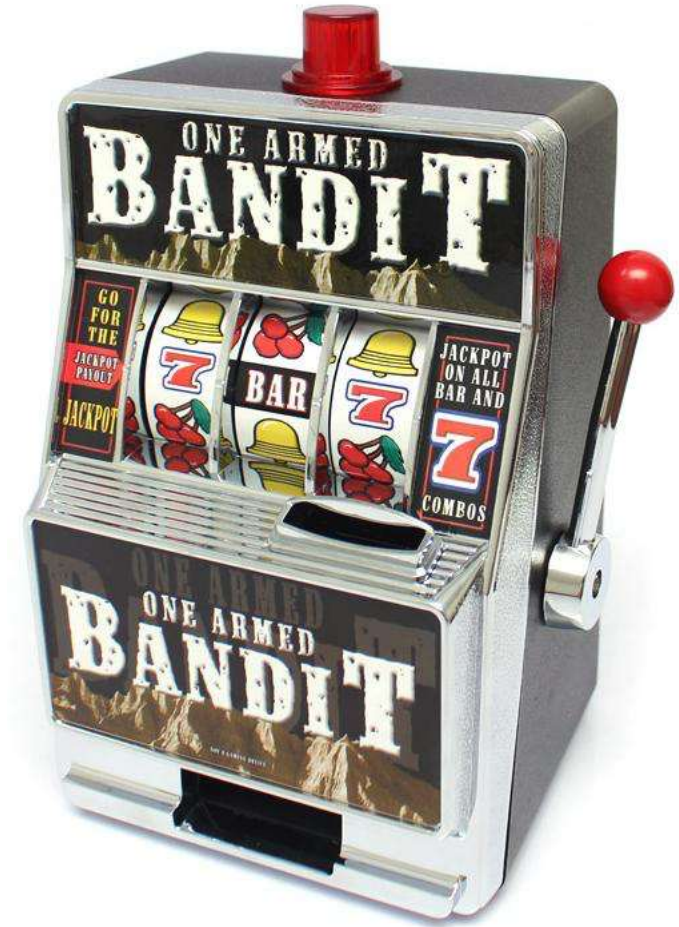




Application: Optimal Allocation of Clinical Trial Participants

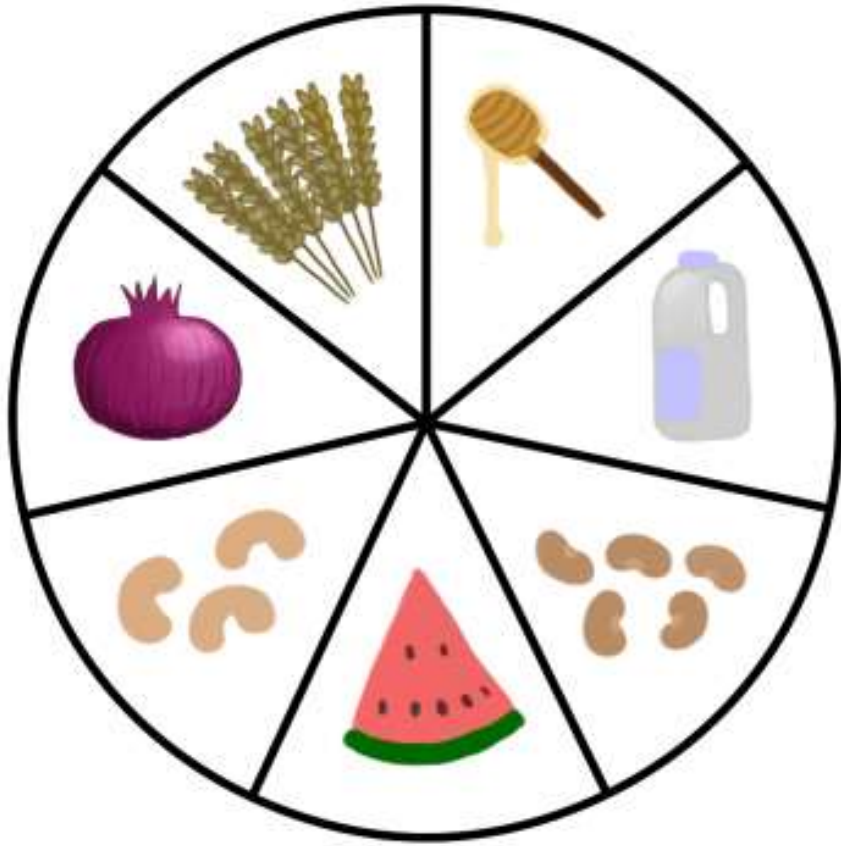
“An explicit assumption is the goal to treat patients effectively, in the trial as well as out. That is controversial (...)”

(Stangl, Inoue and Irony, 2012)



N-of-1 Trial: Identify IBS Triggers

IBS Trigger Foods

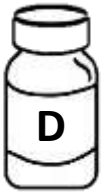
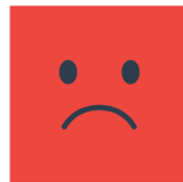


Find foods (i.e. “actions”) that minimize IBS symptoms (i.e. “reward”)

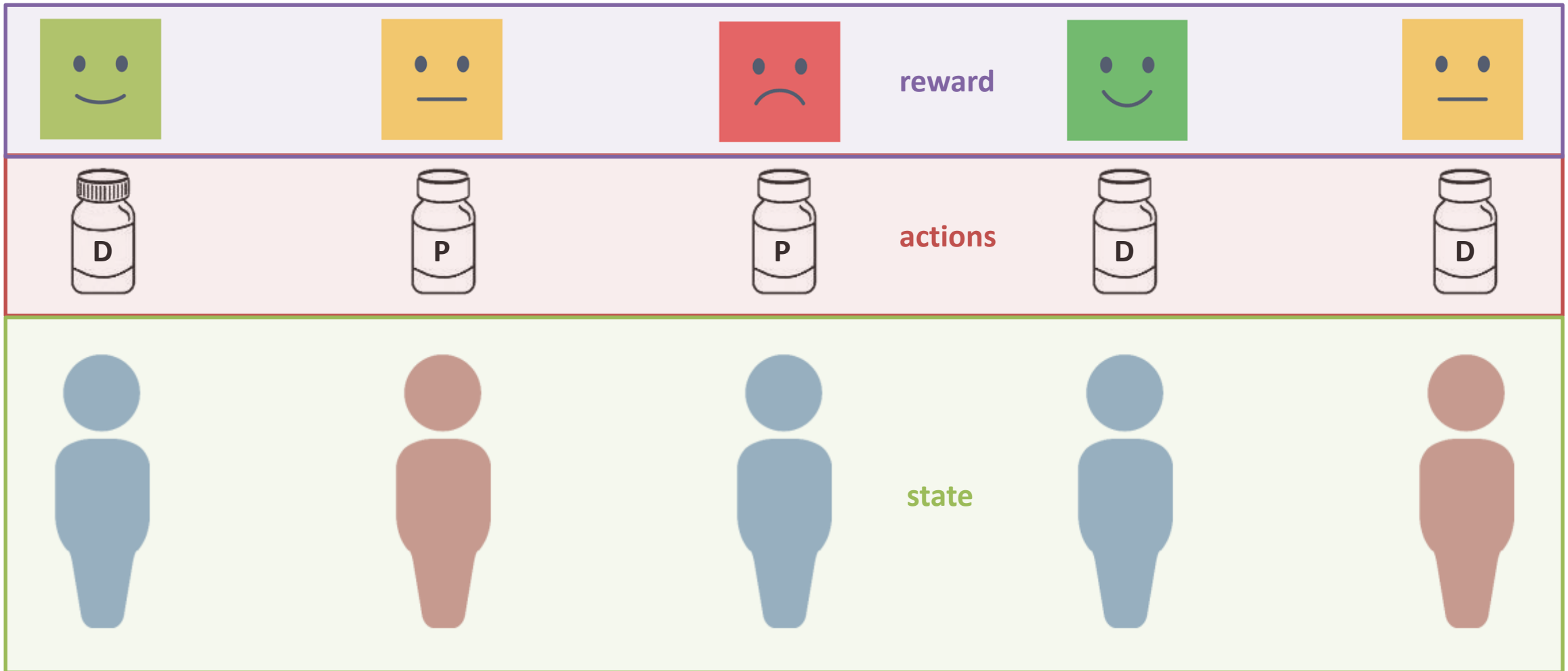
TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers.

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

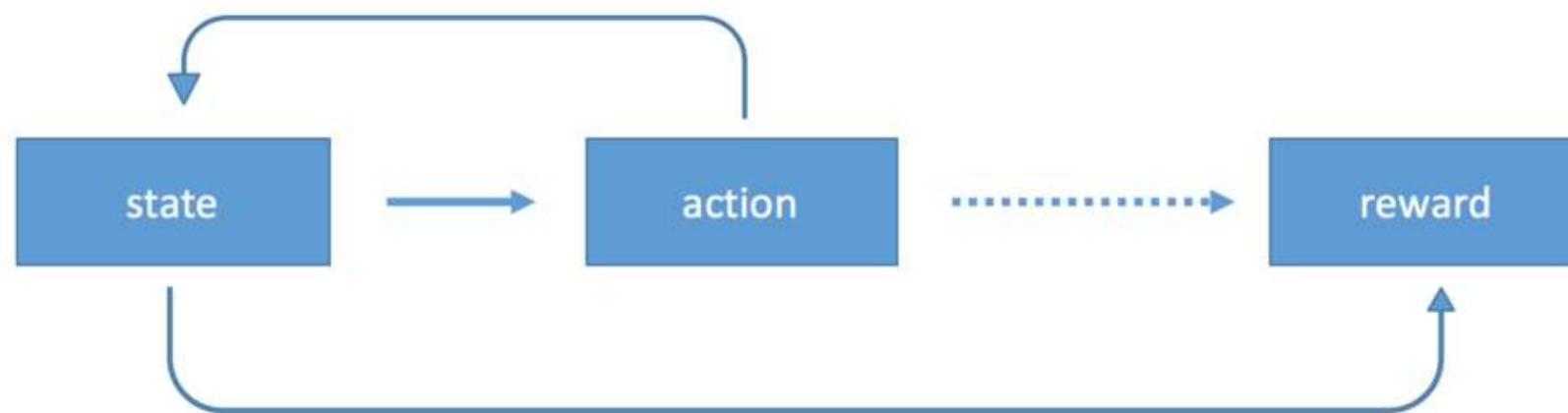




Multi-armed Bandit



Contextual Bandit



Full RL Problem

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

THANK YOU!

Questions or ideas? Please contact me at m.engelhard@duke.edu