# On optimal dynamic treatment regimes (full reinforcement learning)

Nilanjana Laha

TAMIDS

# Broad goal

**Month 1**

These images are generated by Dall-E

**Month 2**

**Month 3**

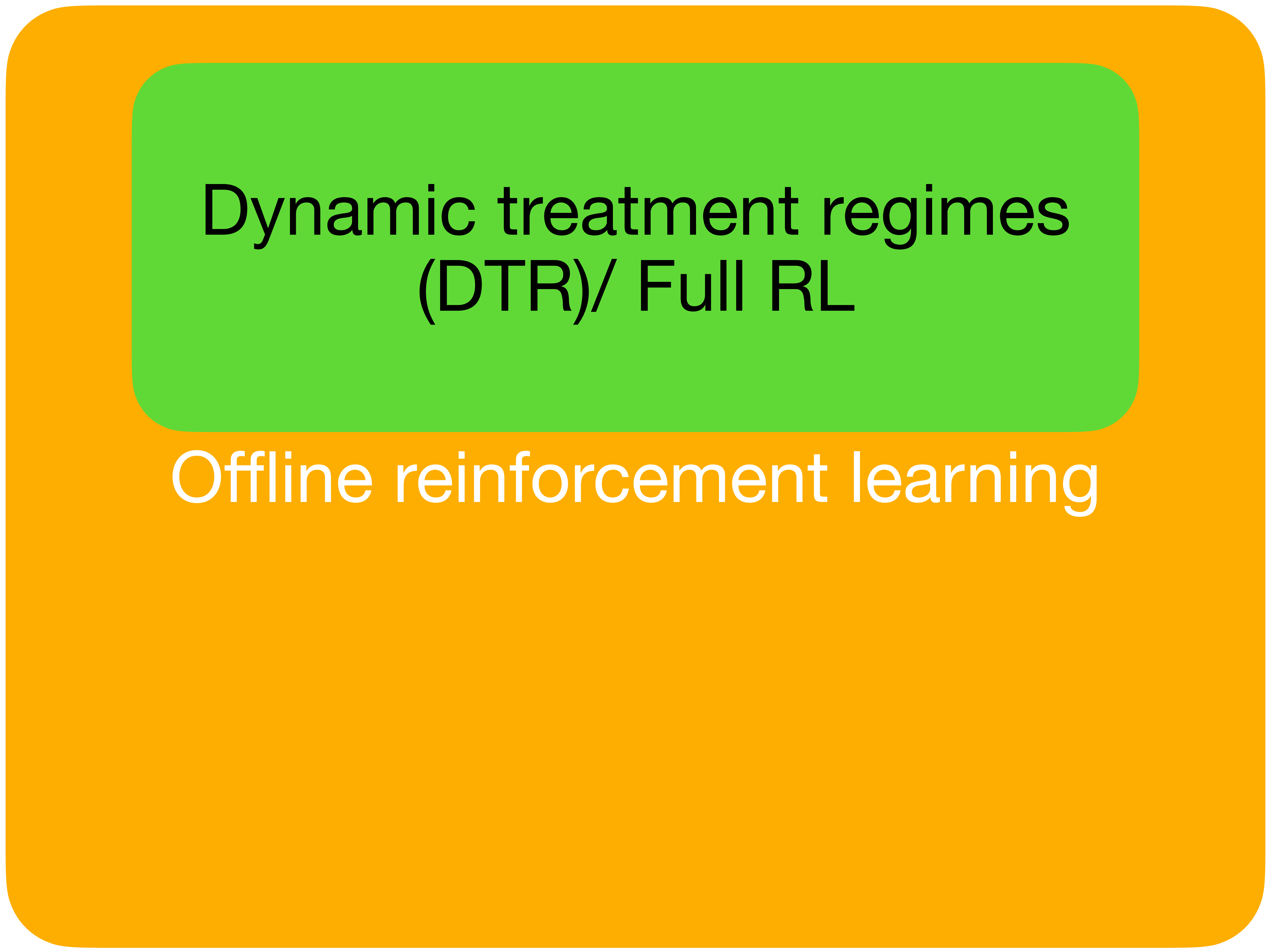
Chronic diseases demand ongoing treatments. Can we apply reinforcement learning for optimal, **patient-specific**, data-driven treatment policy?

# Where does it stand as an area?

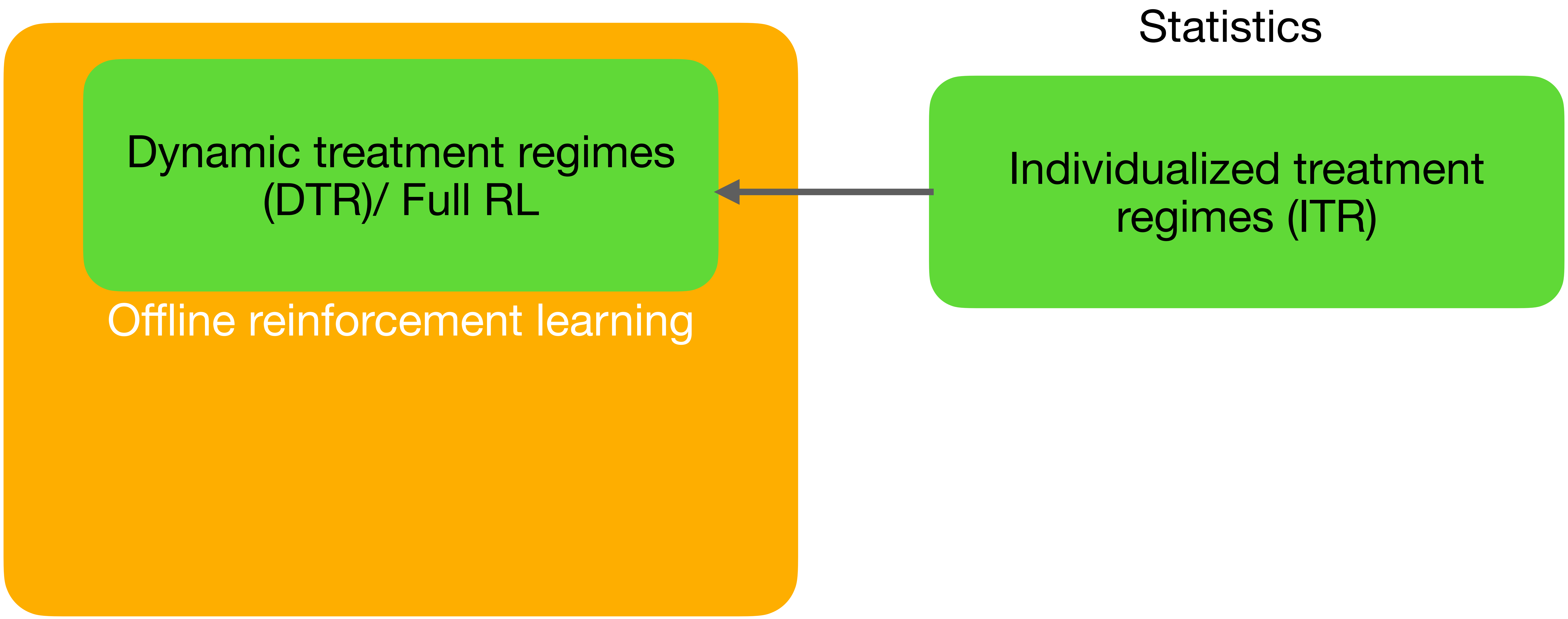
# Where does it stand as an area?

Dynamic treatment regimes (DTR)/ Full RL

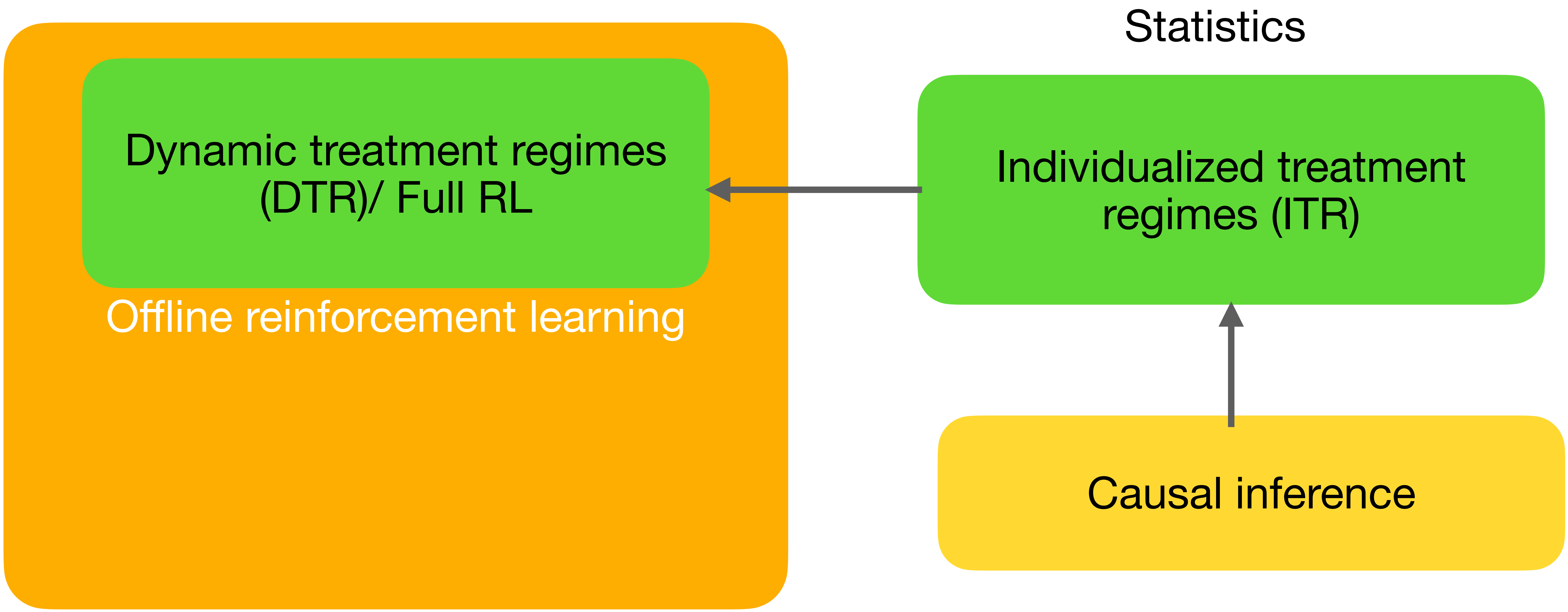
# Where does it stand as an area?



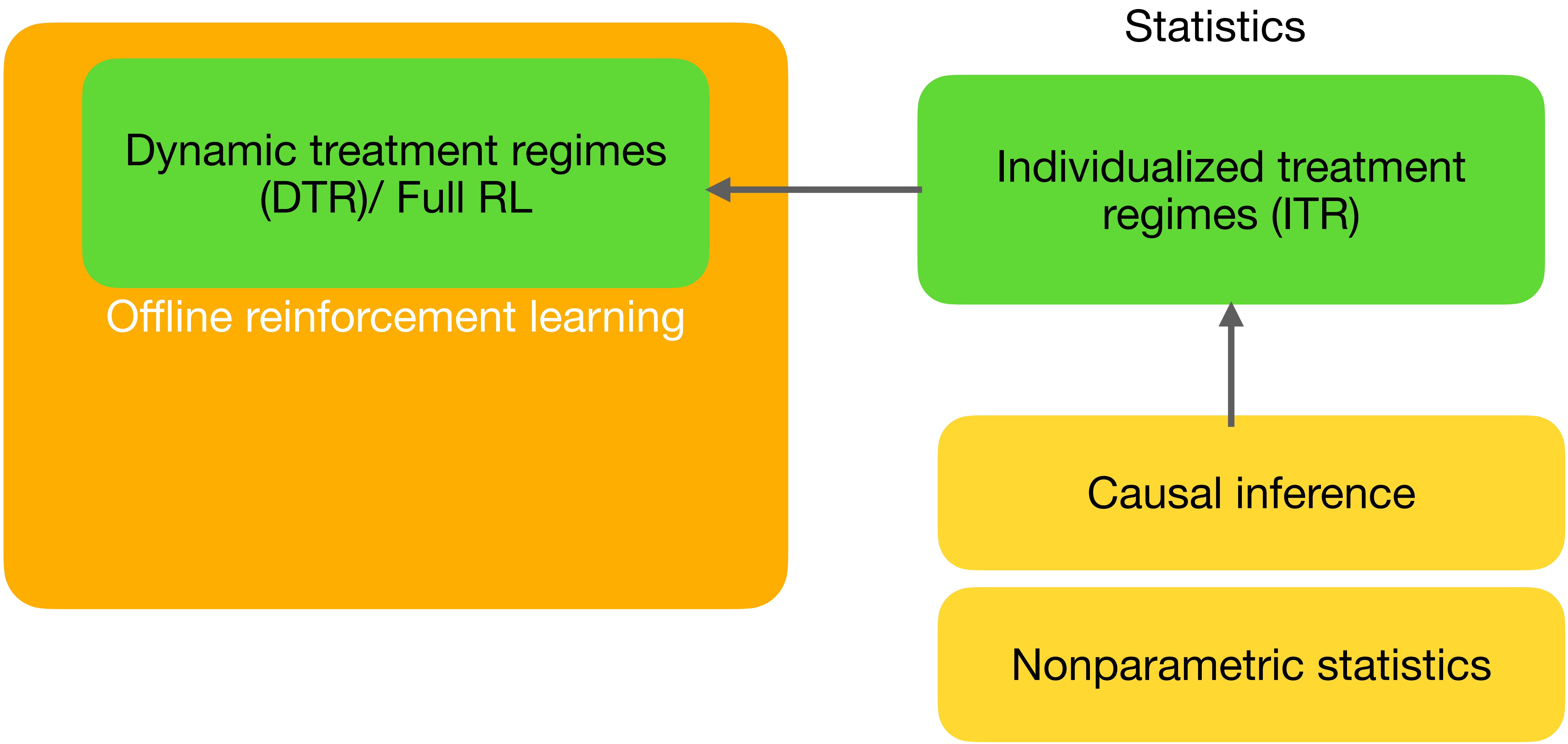
# Where does it stand as an area?



# Where does it stand as an area?



# Where does it stand as an area?



# Outline

- Example: sepsis
- Problem formulation
- Proposed method
- Open questions

# Example: sepsis

# Sepsis

# Sepsis

Cause: Body's response to infection injures own tissues, organs.

Image source: MedicineNet

# Sepsis

Cause: Body's response to infection injures own tissues, organs.

Image source: MedicineNet

Expensive

In-patient cost > $22 billion

# Sepsis

Cause: Body's response to infection injures own tissues, organs.

Image source: MedicineNet

**Expensive**

In-patient cost > $22 billion

**Challenging**

Fatality 30 %

# Sepsis

Cause: Body's response to infection injures own tissues, organs.

Image source: MedicineNet

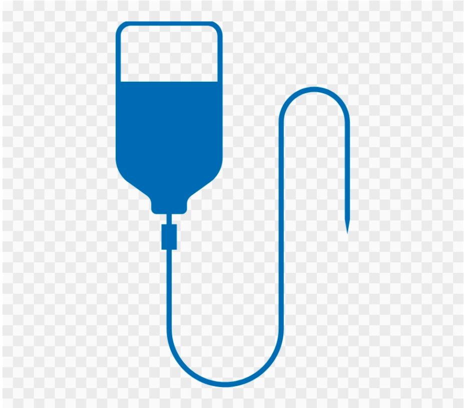
**Expensive**

In-patient cost > $22 billion

**Challenging**

Fatality 30 %

**Popular treatment:**



IV-fluid administration

# Sepsis

Cause: Body's response to infection injures own tissues, organs.

Image source: MedicineNet

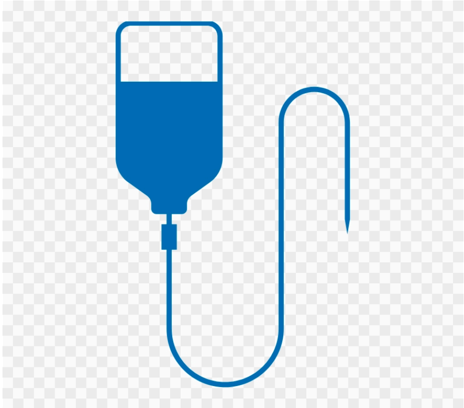
**Expensive**

In-patient cost > $22 billion

**Challenging**

Fatality 30 %

**Popular treatment:**



IV-fluid administration

Goal: policy learning for IV-fluid administration

# Sepsis-3 data (Beth Israel Hospital, Boston)

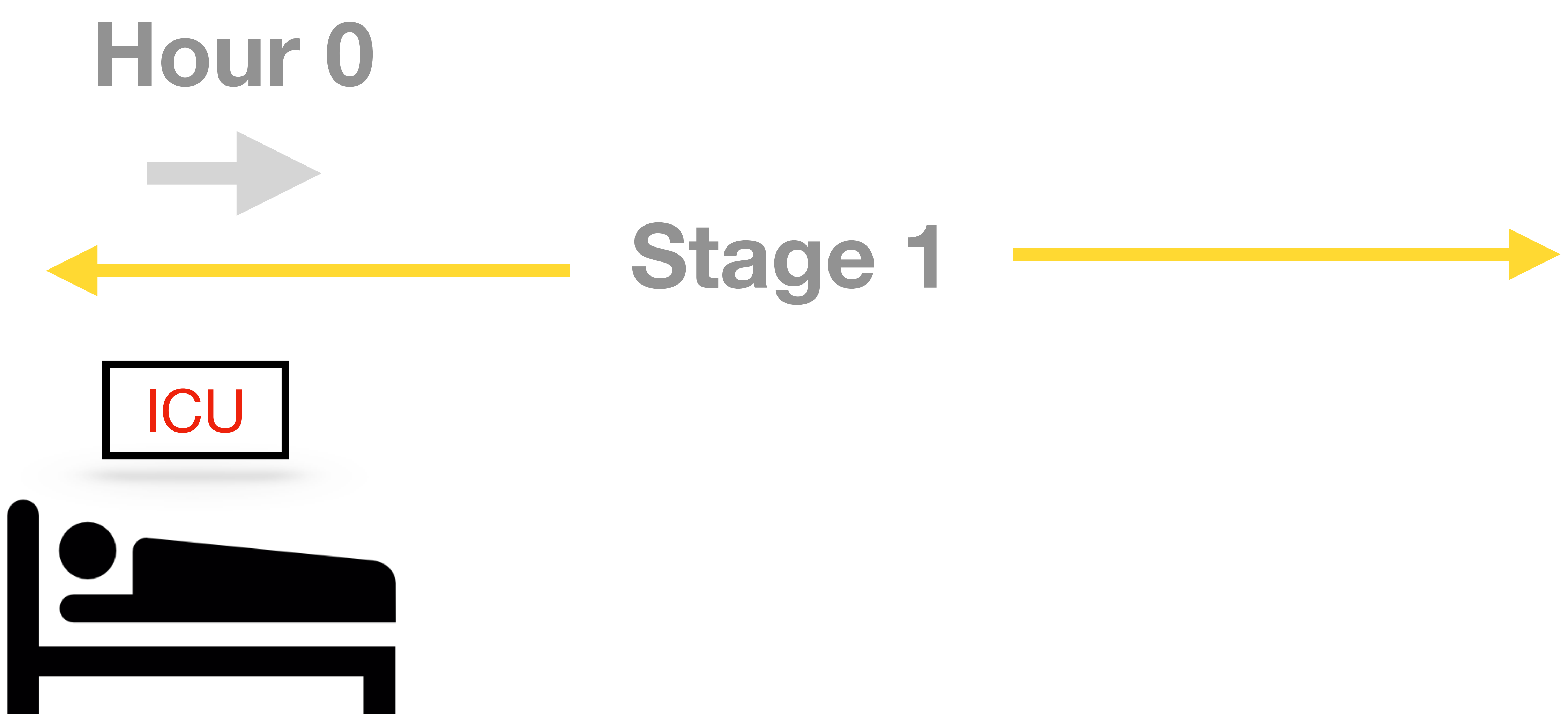
# Sepsis-3 data (Beth Israel Hospital, Boston)

## Hour 0

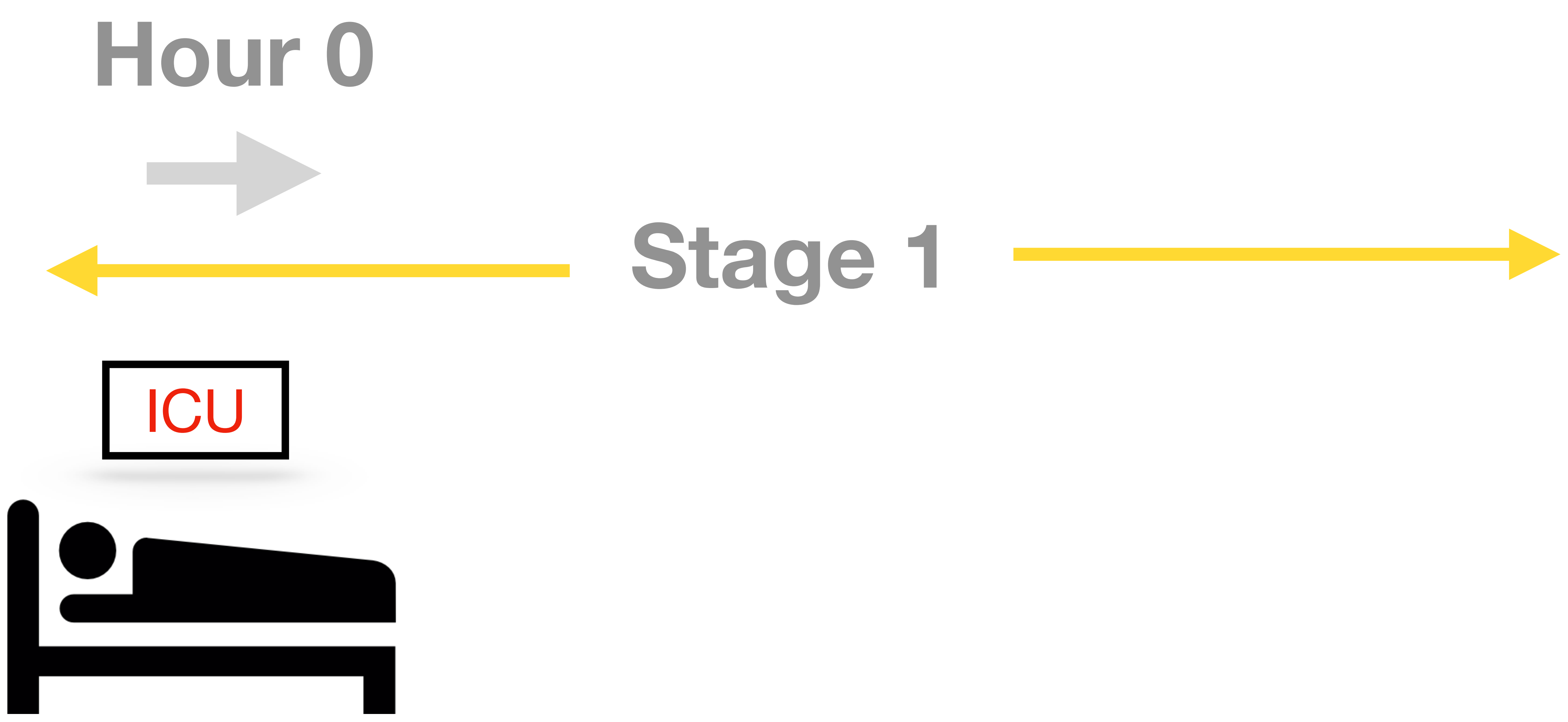
# Sepsis-3 data (Beth Israel Hospital, Boston)

**Hour 0**

# Sepsis-3 data (Beth Israel Hospital, Boston)



# Sepsis-3 data (Beth Israel Hospital, Boston)



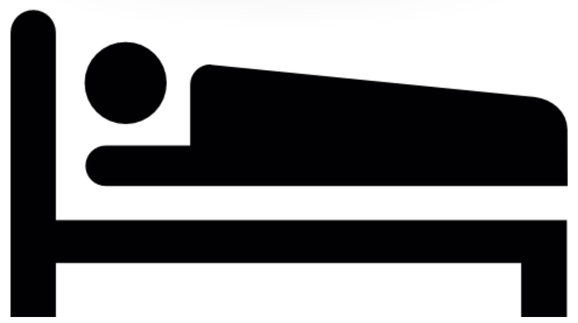

Baseline covariates: $S_1$

# Sepsis-3 data (Beth Israel Hospital, Boston)

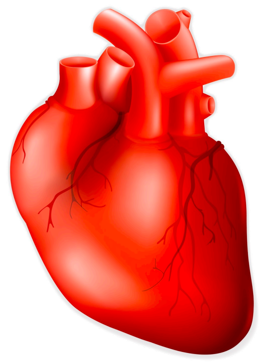
Hour 0

Stage 1

ICU



Baseline covariates: $S_1$

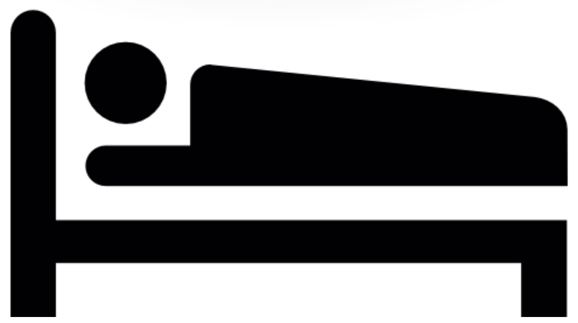
 ,

# Sepsis-3 data (Beth Israel Hospital, Boston)

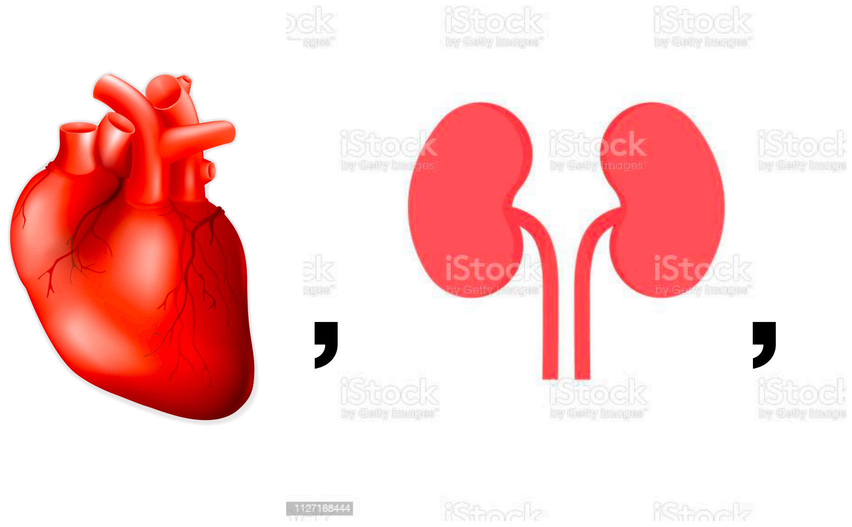
Hour 0

Stage 1

ICU



Baseline covariates: $S_1$

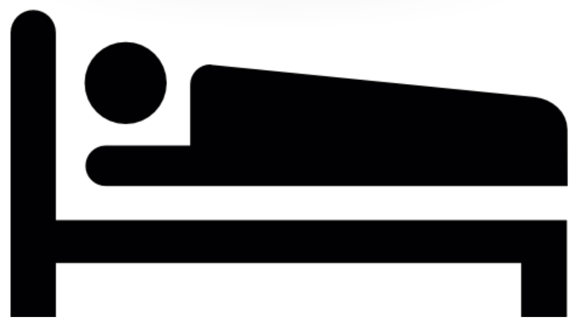


## Sepsis-3 data (Beth Israel Hospital, Boston)

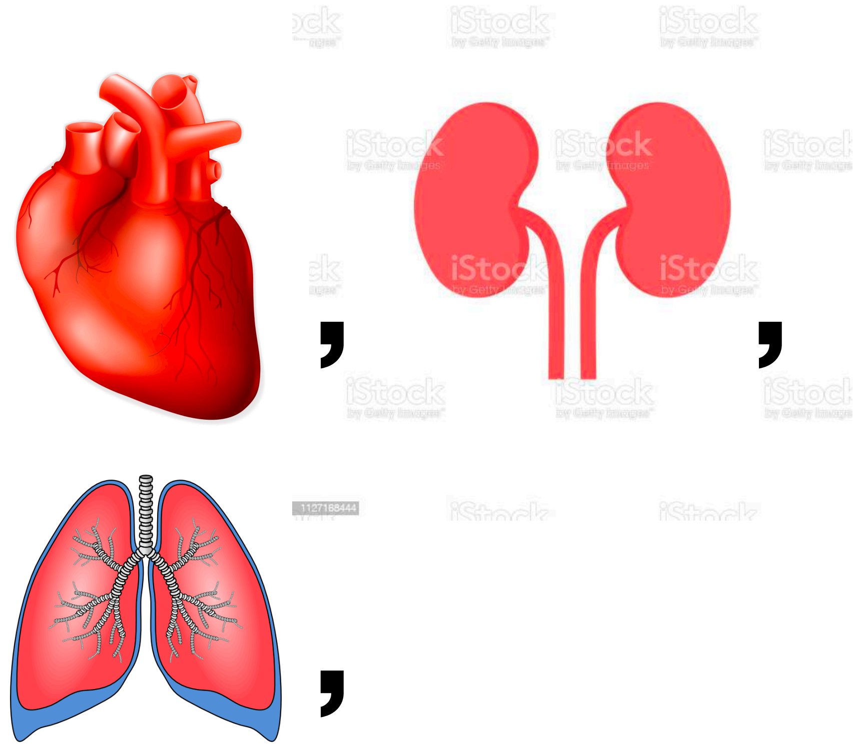
**Hour 0**

**Stage 1**

ICU



Baseline covariates: $S_1$

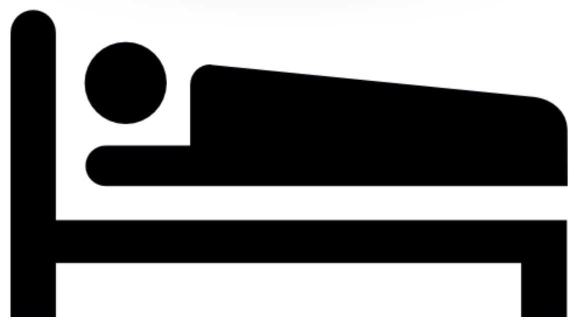


# Sepsis-3 data (Beth Israel Hospital, Boston)

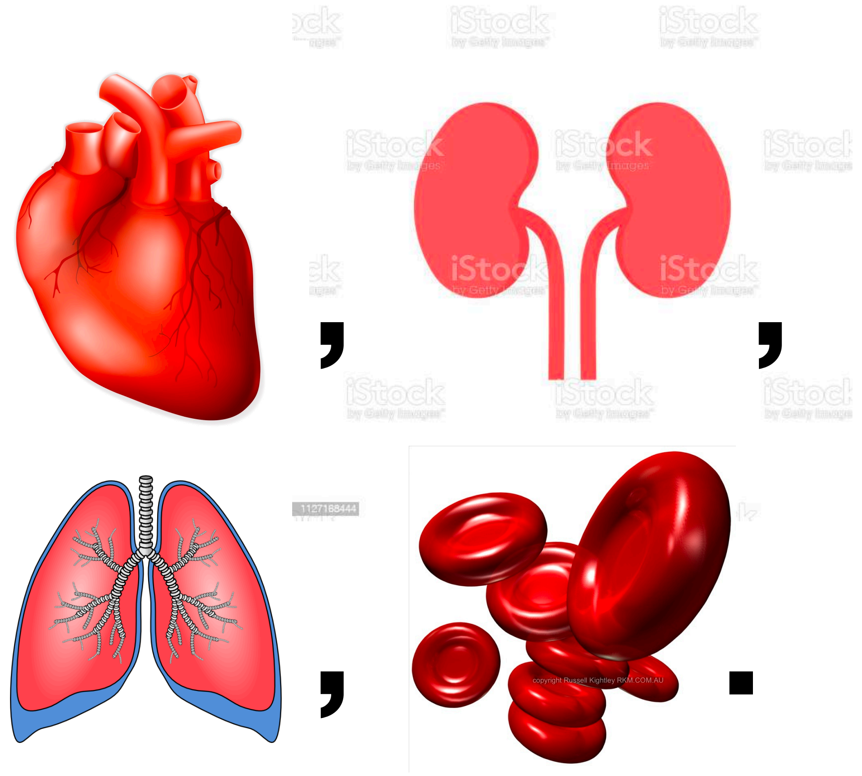
Hour 0

Stage 1

ICU



Baseline covariates: $S_1$



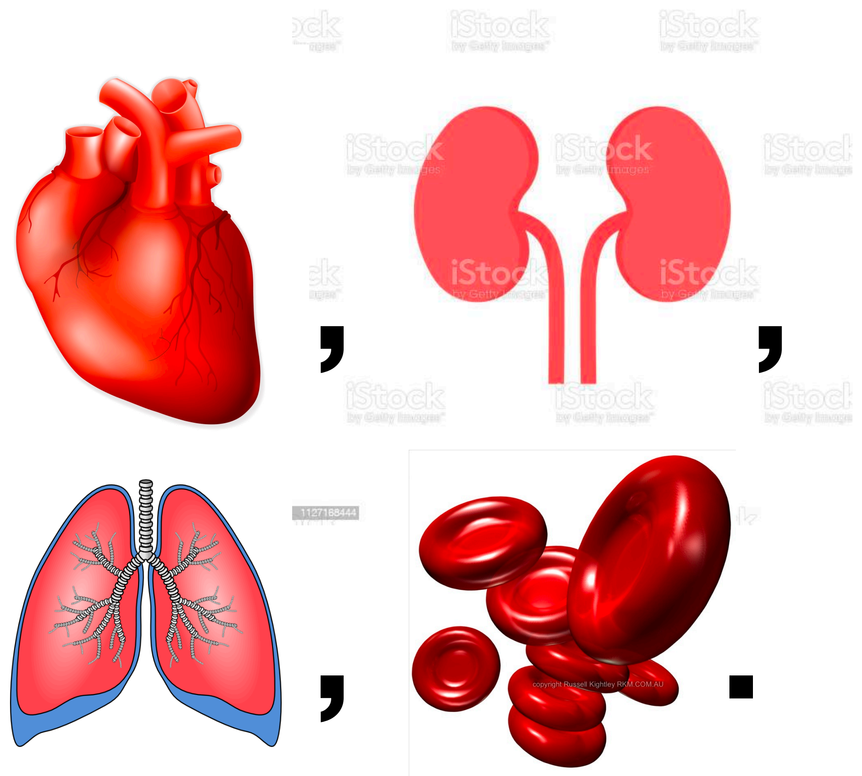
# Sepsis-3 data (Beth Israel Hospital, Boston)

**Hour 0**

**Stage 1**


Baseline covariates: $S_1$

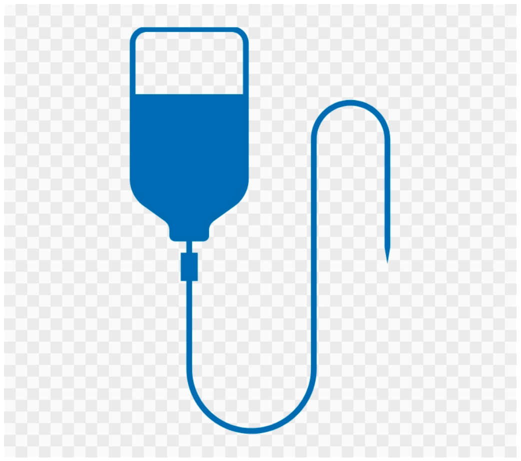
Treatment $A_1$



# Sepsis-3 data (Beth Israel Hospital, Boston)

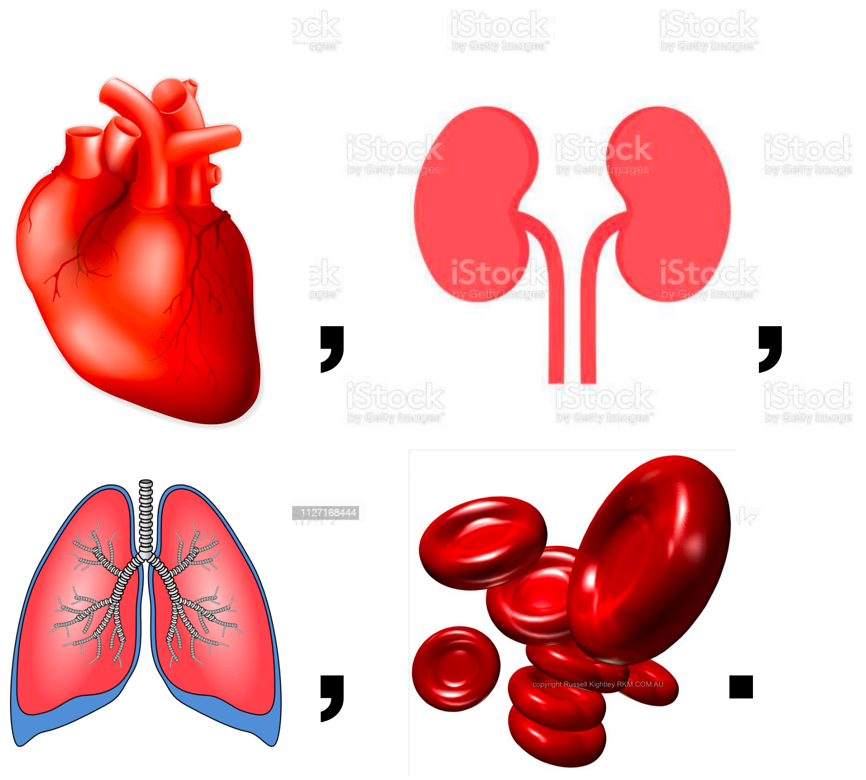
Hour 0

Stage 1

Baseline covariates: $S_1$

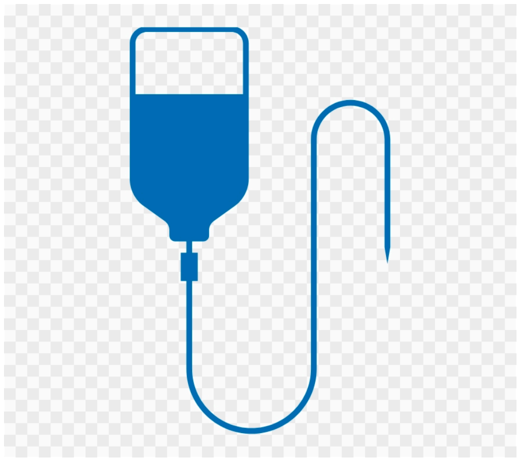
Treatment $A_1$



## Sepsis-3 data (Beth Israel Hospital, Boston)

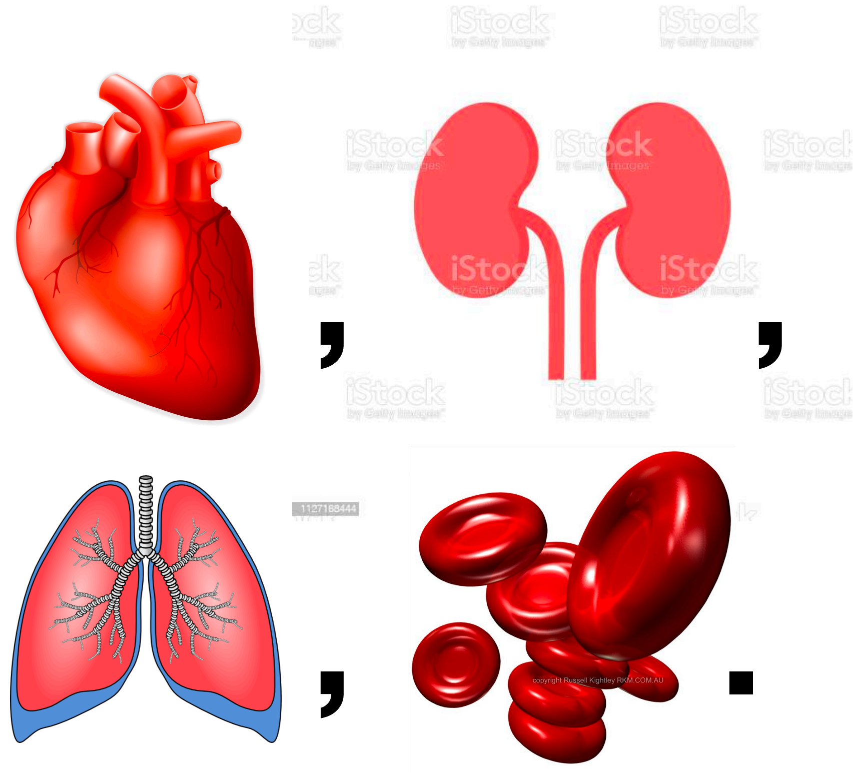
Hour 0

Stage 1

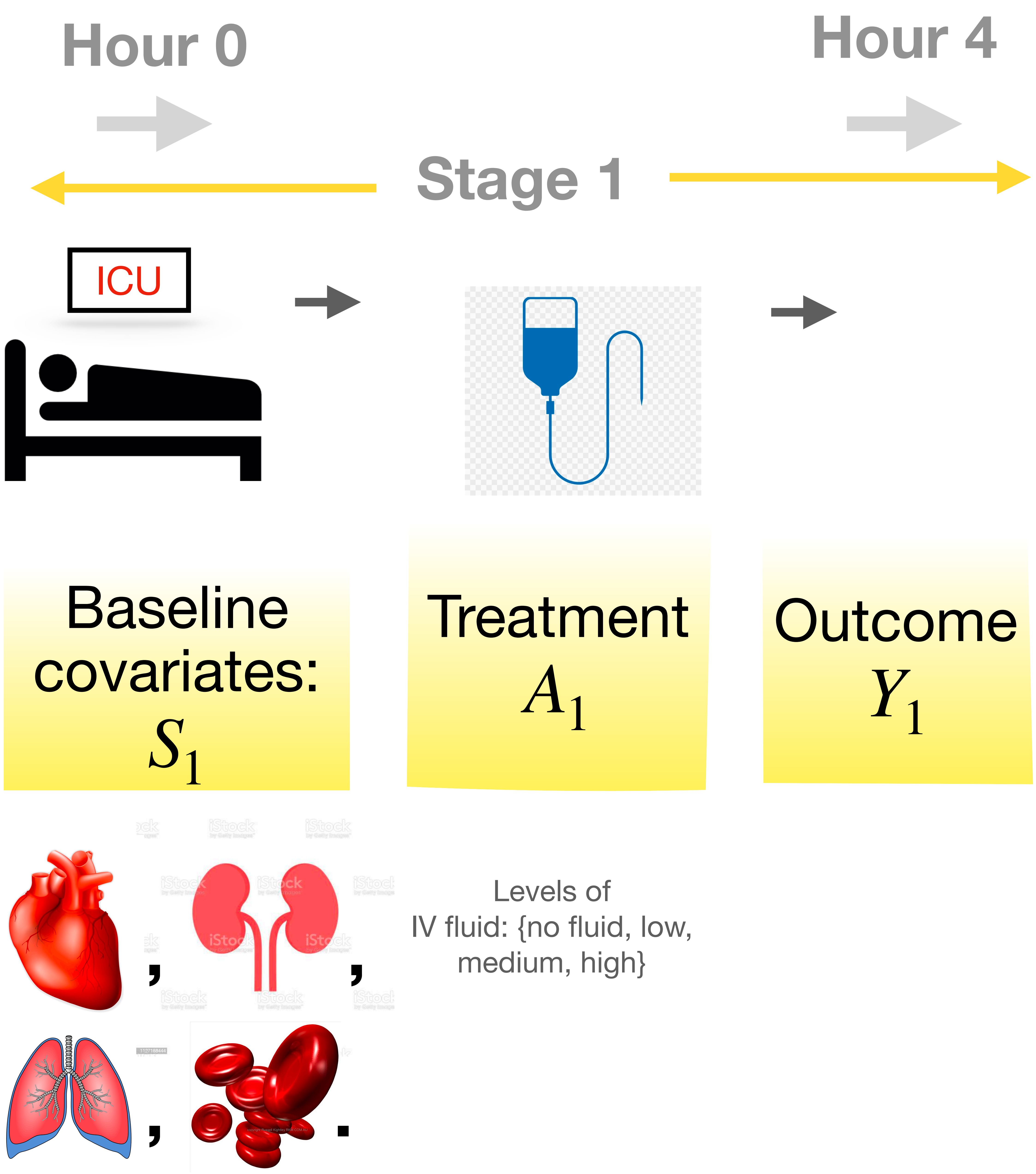
Baseline covariates: $S_1$

Treatment $A_1$

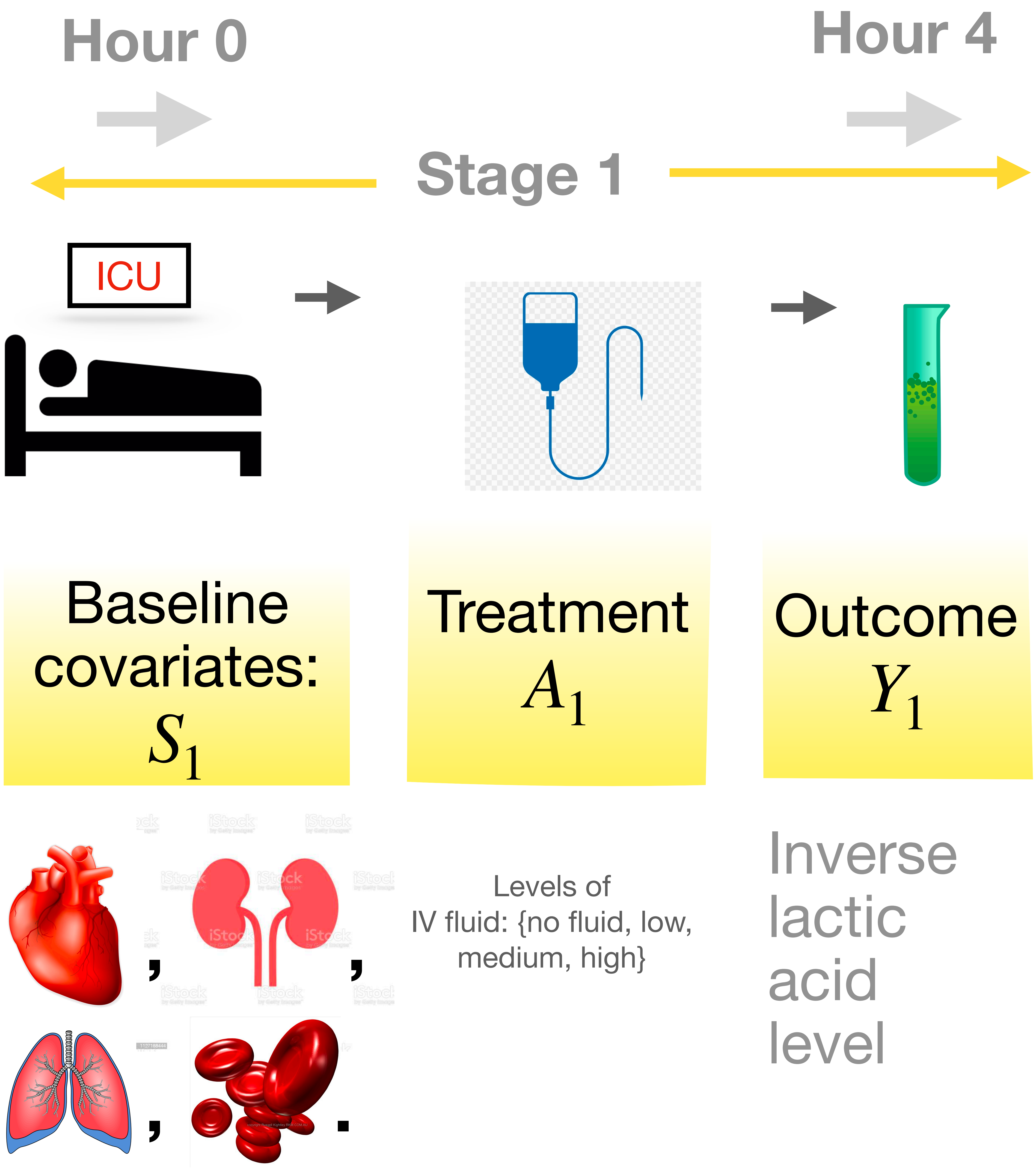


Levels of IV fluid: {no fluid, low, medium, high}

# Sepsis-3 data (Beth Israel Hospital, Boston)



# Sepsis-3 data (Beth Israel Hospital, Boston)

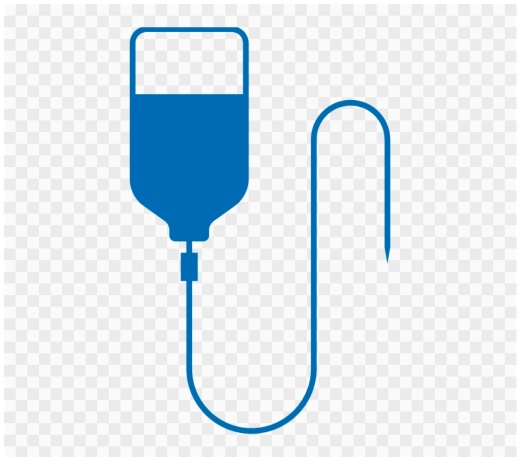


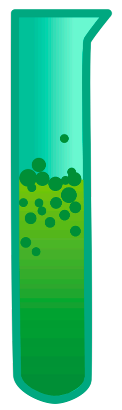
## Sepsis-3 data (Beth Israel Hospital, Boston)

Hour 0

Hour 4

Stage 1

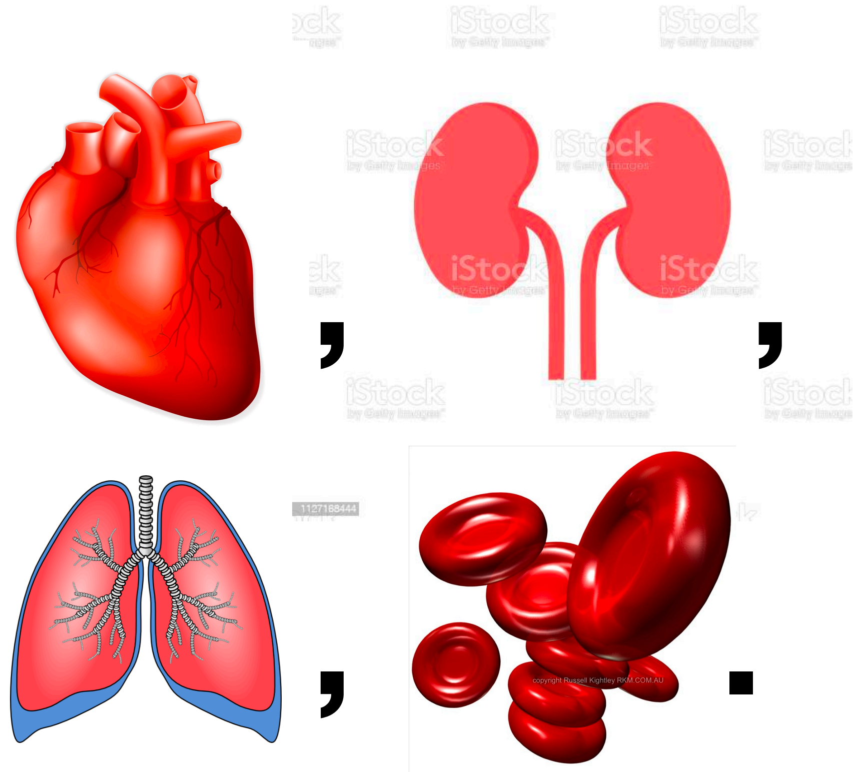





Baseline covariates: $S_1$

Treatment $A_1$

Outcome $Y_1$



Levels of IV fluid: {no fluid, low, medium, high}

Inverse lactic acid level

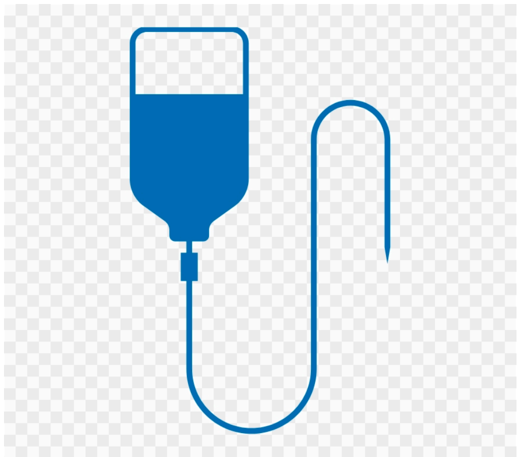
Inverse lactate level

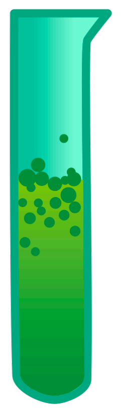
## Sepsis-3 data (Beth Israel Hospital, Boston)

Hour 0

Hour 4

Stage 1

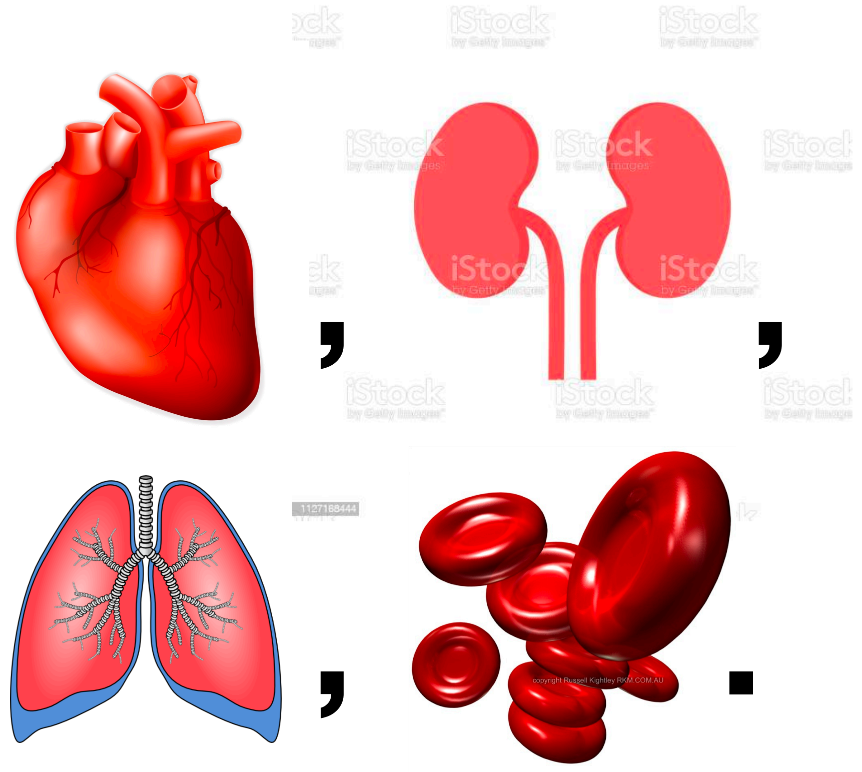





Baseline covariates: $S_1$

Treatment $A_1$

Outcome $Y_1$



Levels of IV fluid: {no fluid, low, medium, high}

Inverse lactic acid level

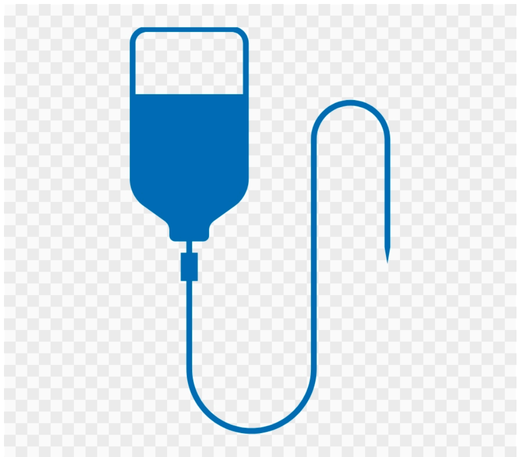
Inverse lactate level

## Sepsis-3 data (Beth Israel Hospital, Boston)

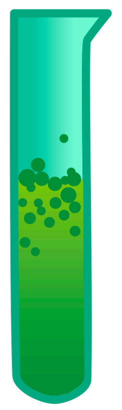
**Hour 0** → **Hour 4**

← **Stage 1** →


→



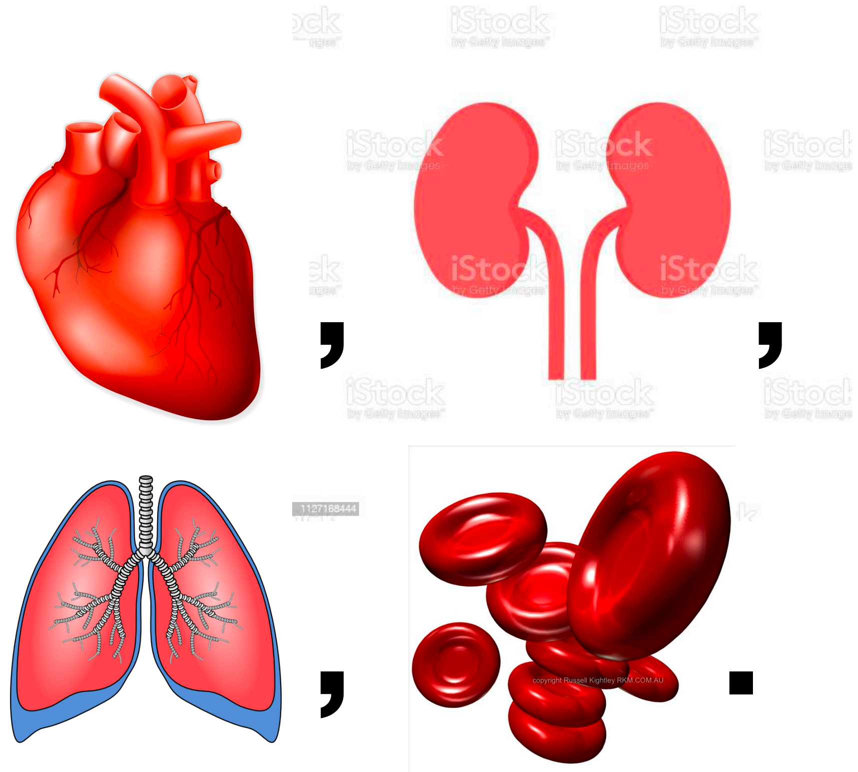
→



Baseline covariates: $S_1$

Treatment $A_1$

Outcome $Y_1$

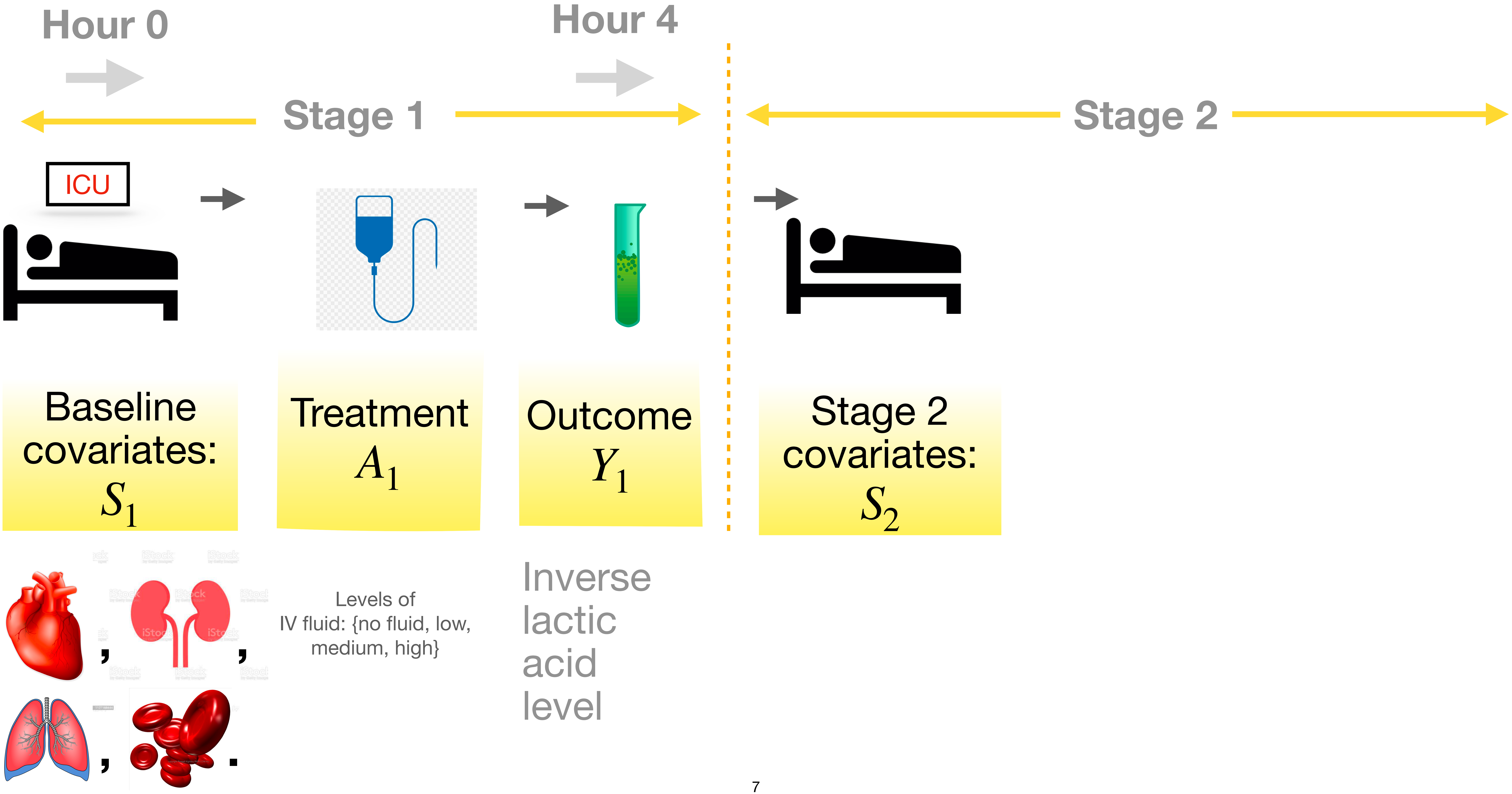


Levels of
IV fluid: {no fluid, low, medium, high}

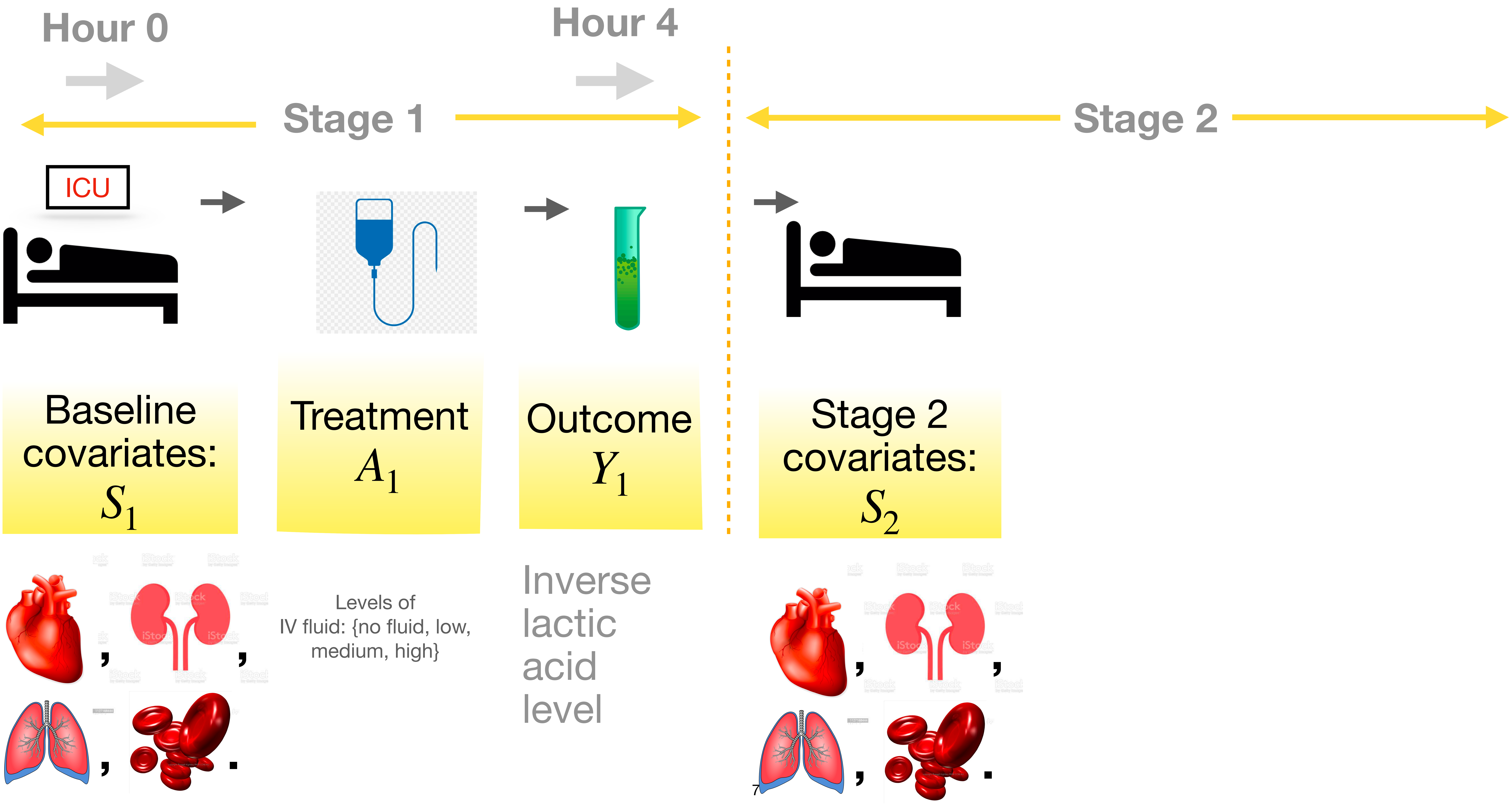
Inverse lactic acid level

Inverse lactate level ↑ Treatment working

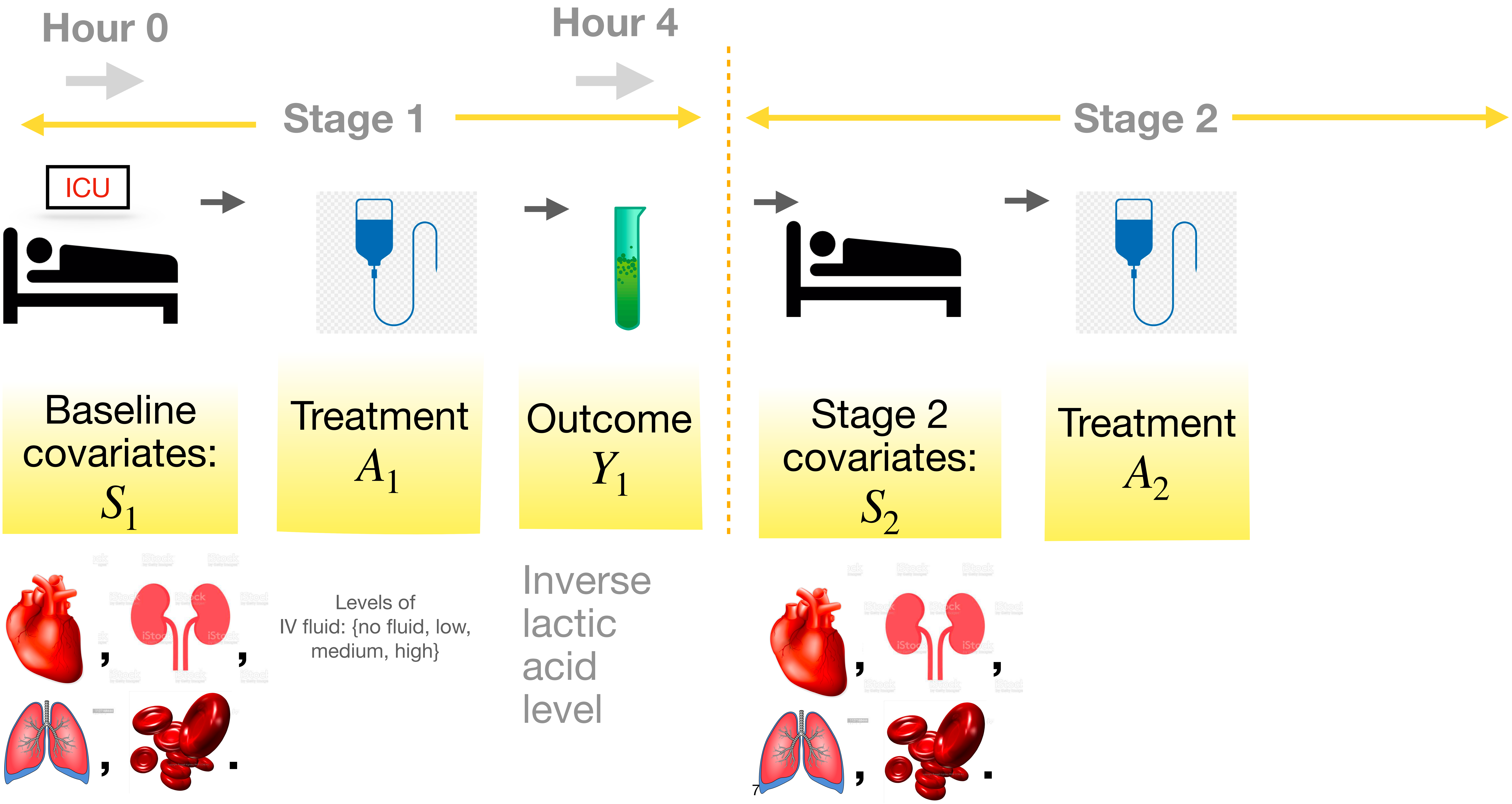
# Sepsis-3 data (Beth Israel Hospital, Boston)



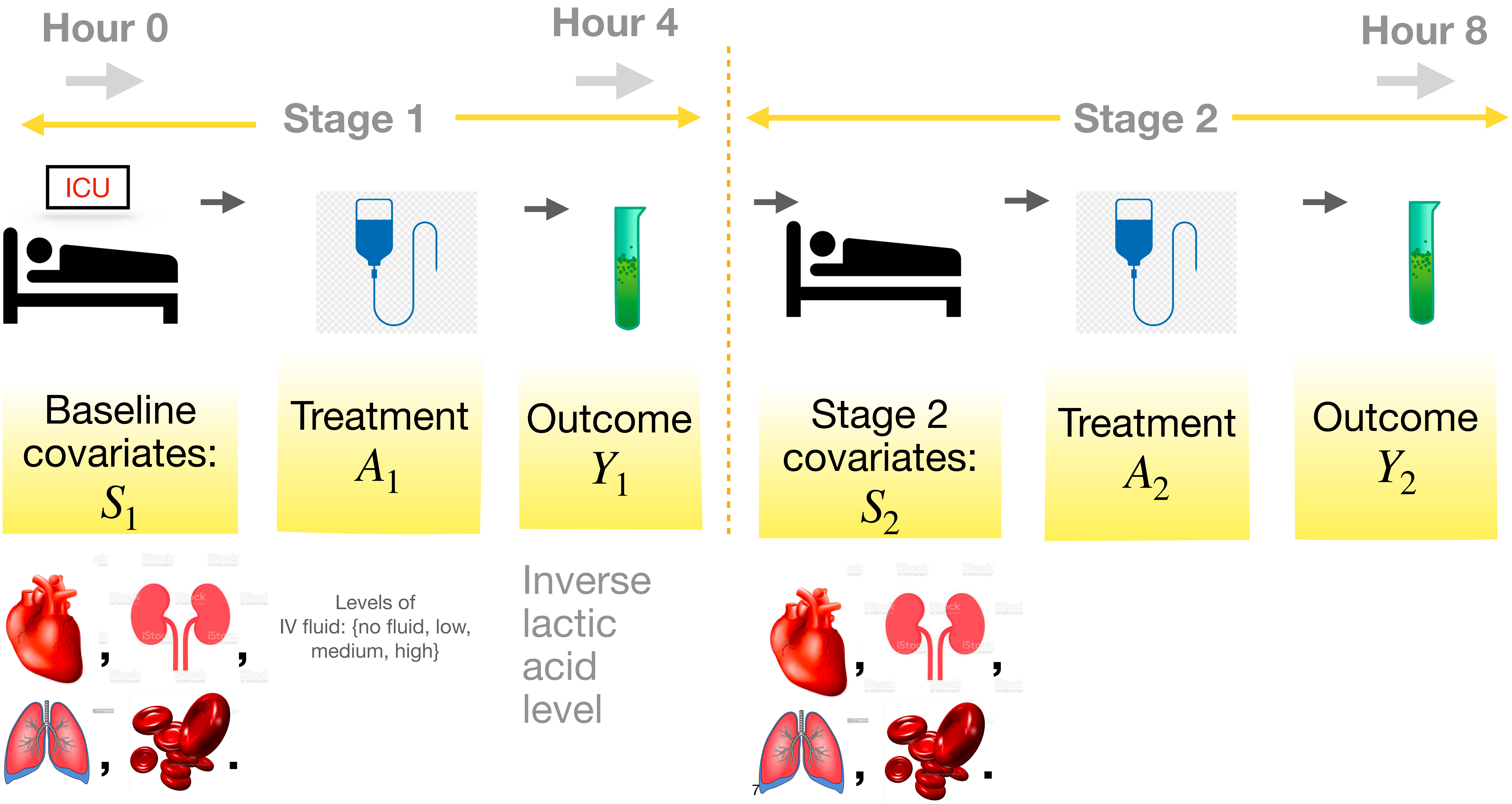
## Sepsis-3 data (Beth Israel Hospital, Boston)



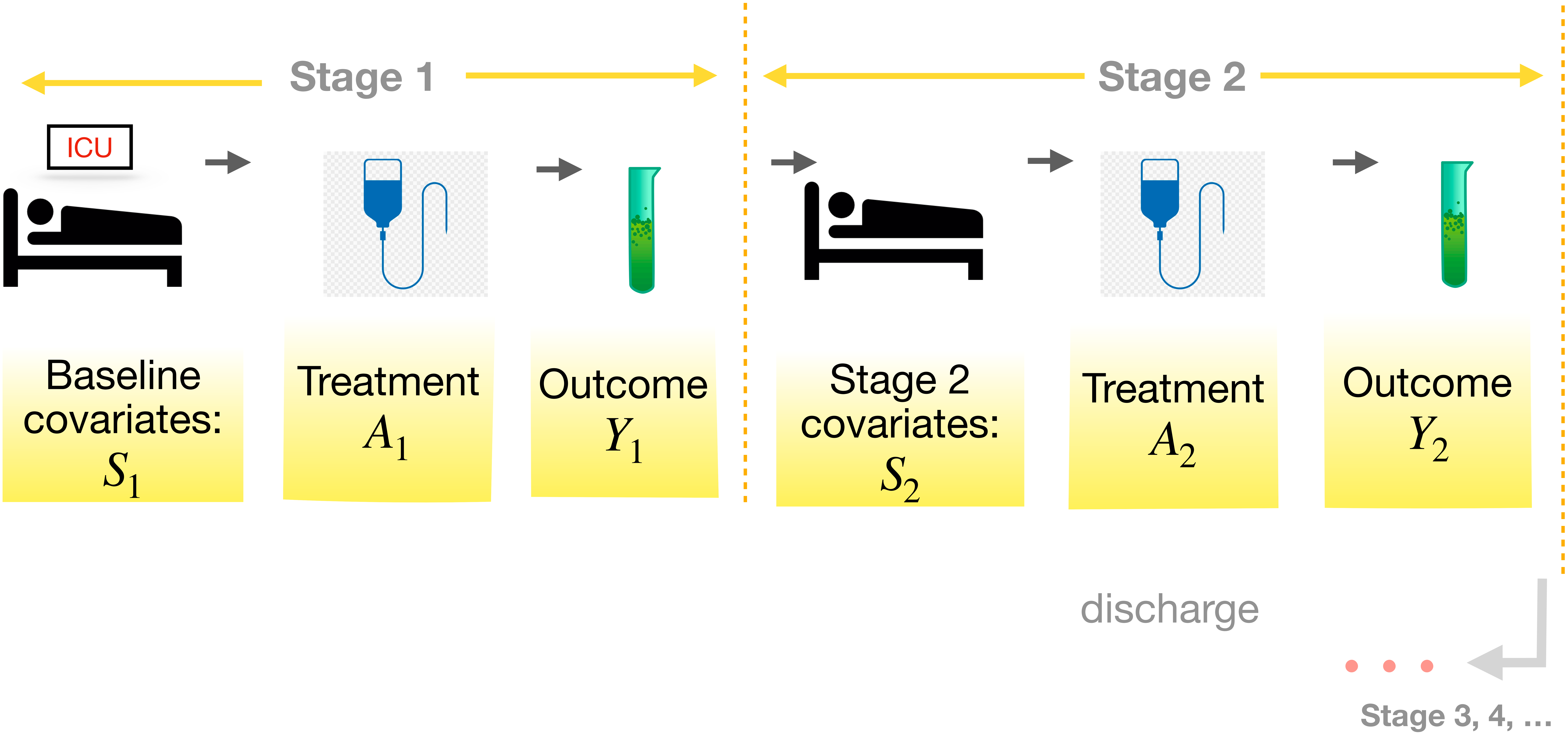
## Sepsis-3 data (Beth Israel Hospital, Boston)



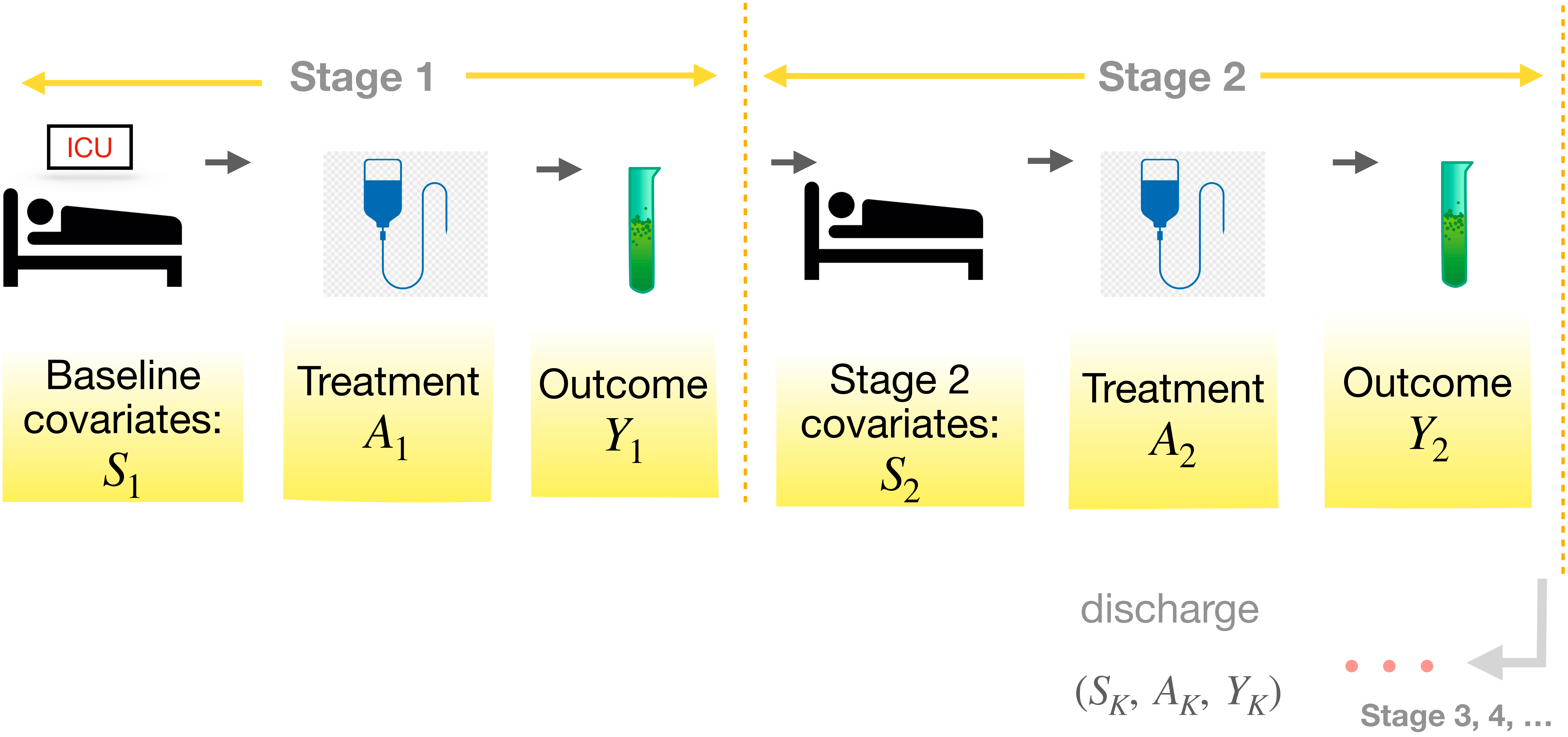
# Sepsis-3 data (Beth Israel Hospital, Boston)



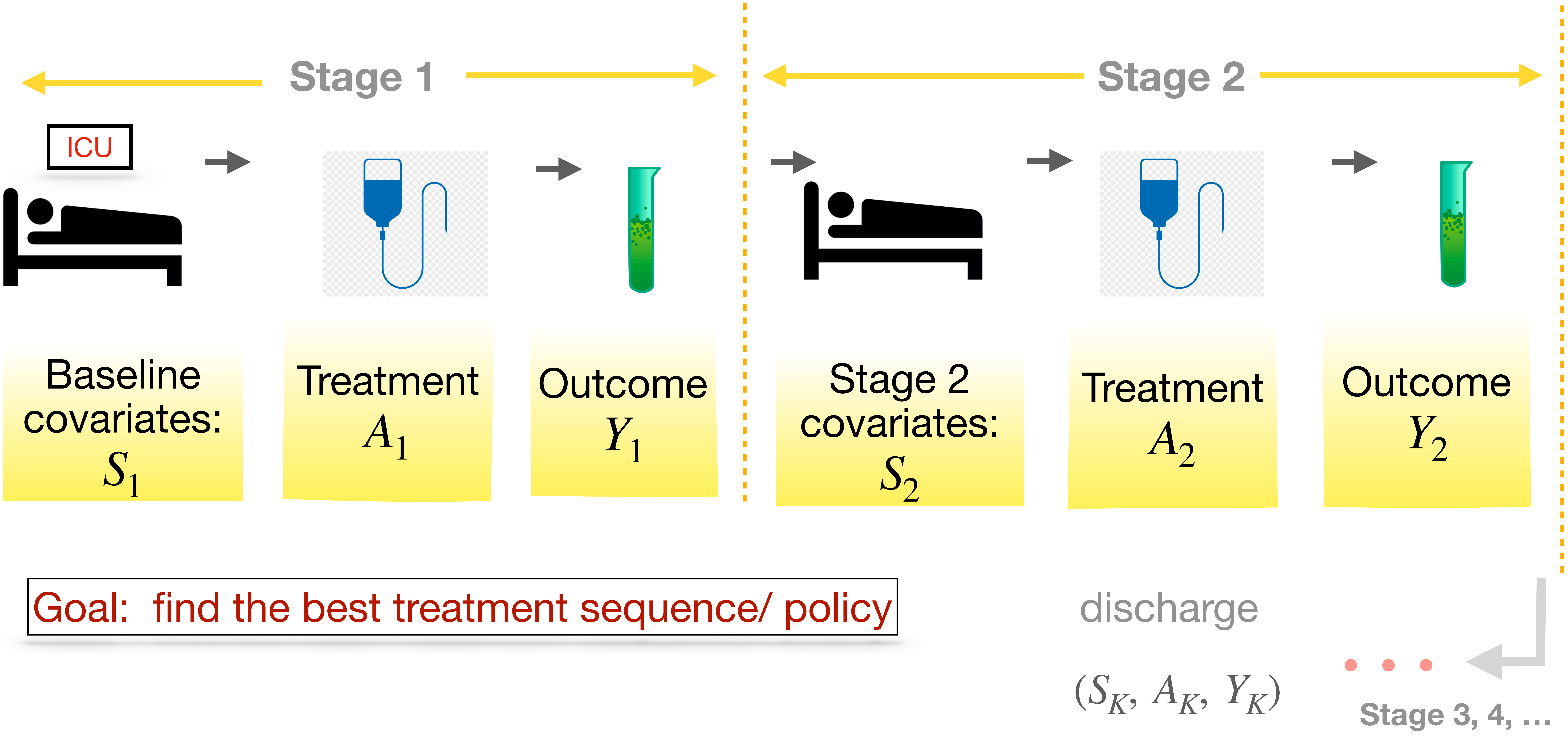
# Sepsis-3 data (Beth Israel Hospital, Boston)



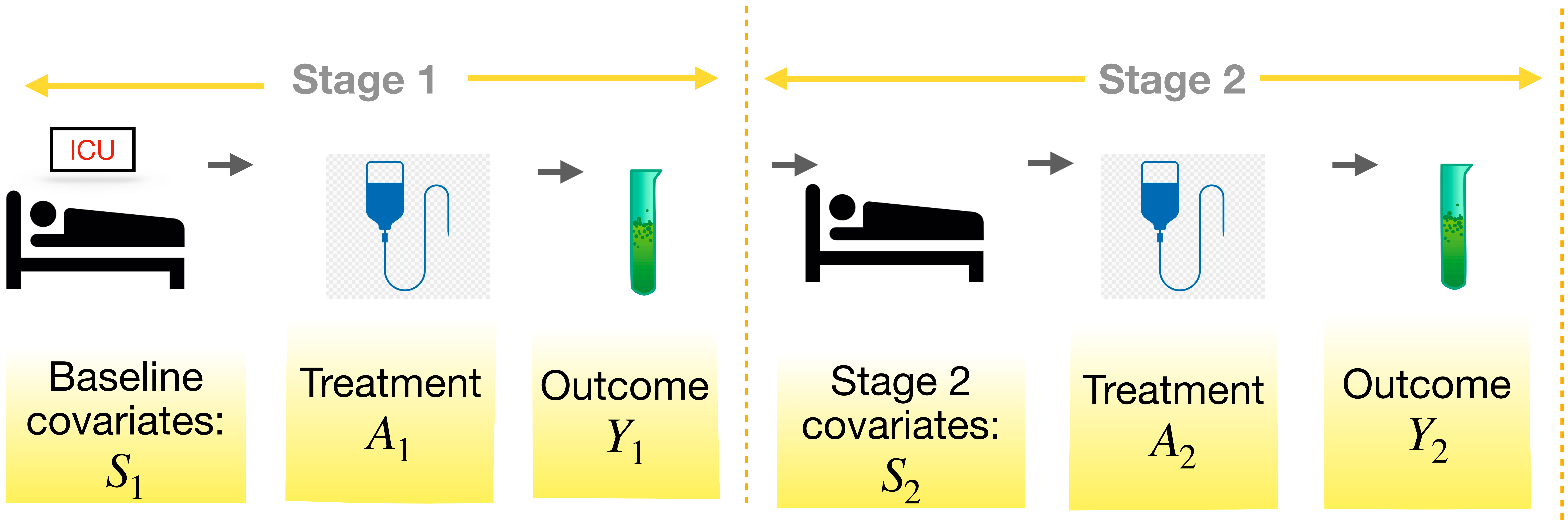
# Sepsis-3 data (Beth Israel Hospital, Boston)

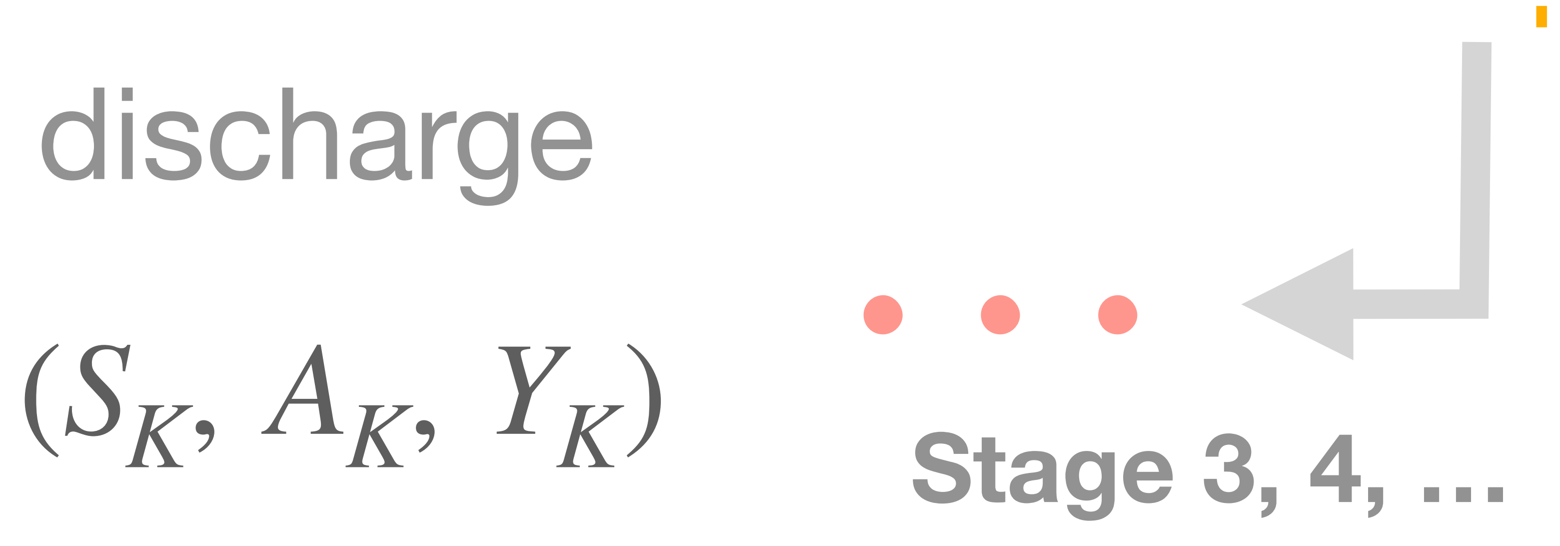


# Sepsis-3 data (Beth Israel Hospital, Boston)



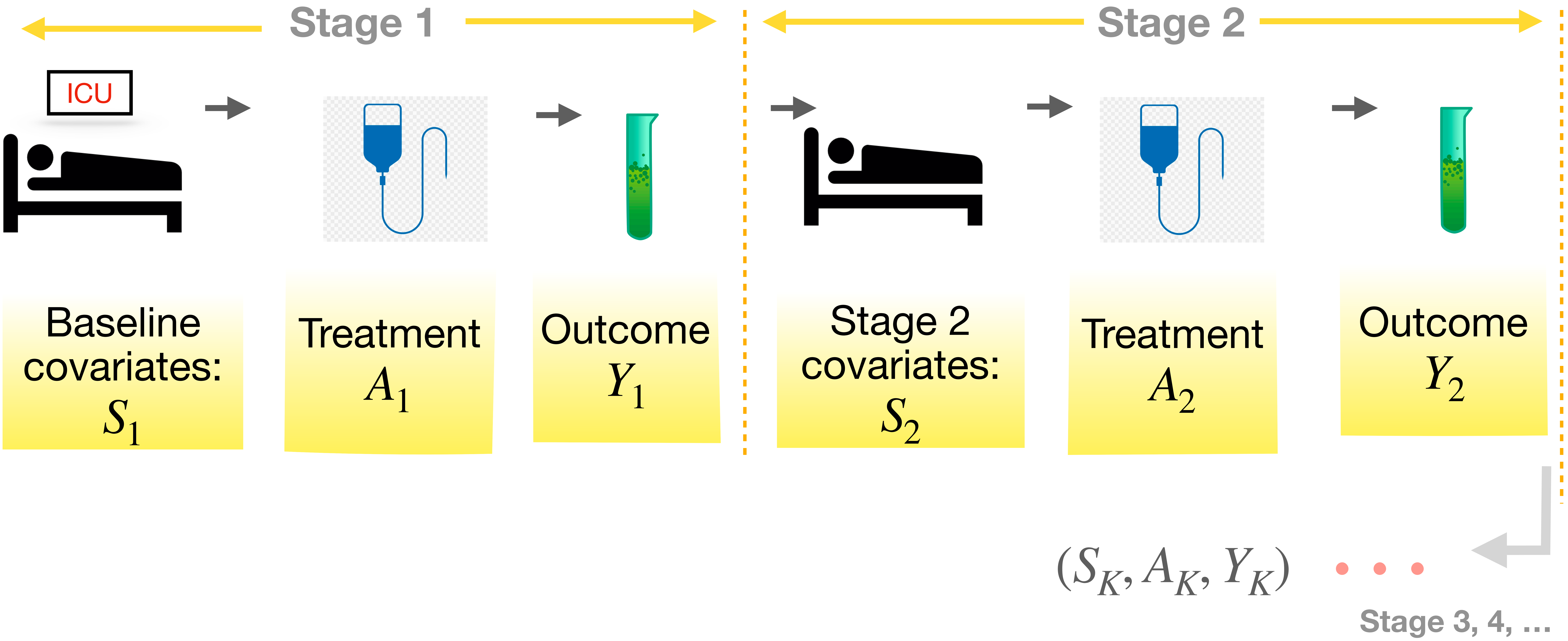
## Sepsis-3 data (Beth Israel Hospital, Boston)



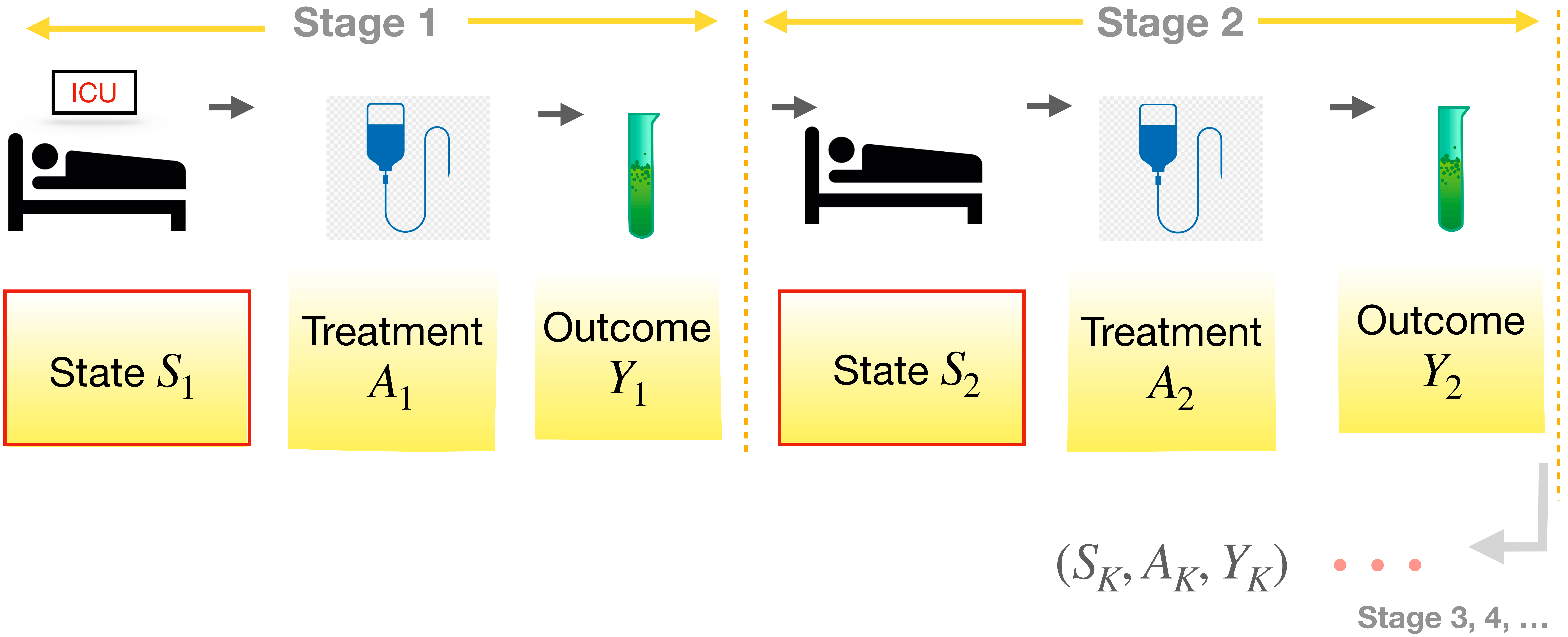




best treatment policy $\pi^* \Longrightarrow \sum_{i=1}^{K} Y_i$ : maximized

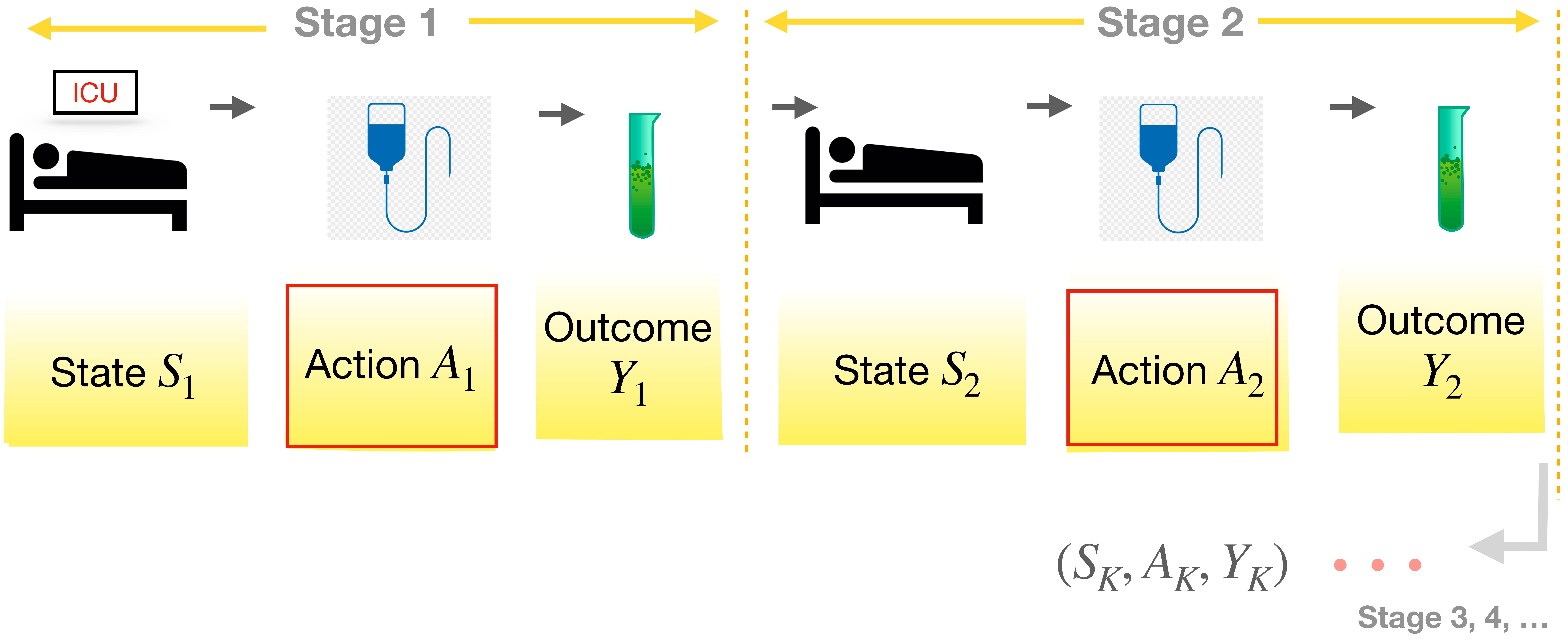
# Offline Reinforcement Learning



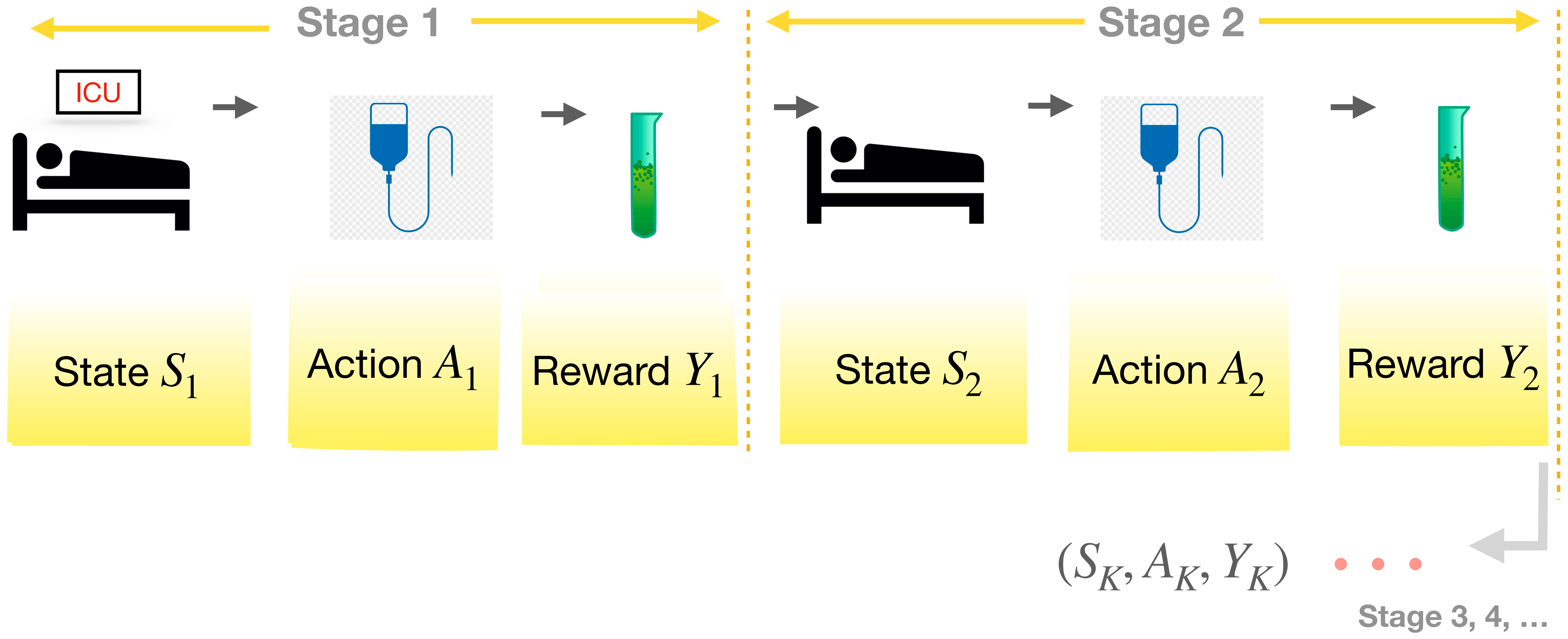
# Offline Reinforcement Learning



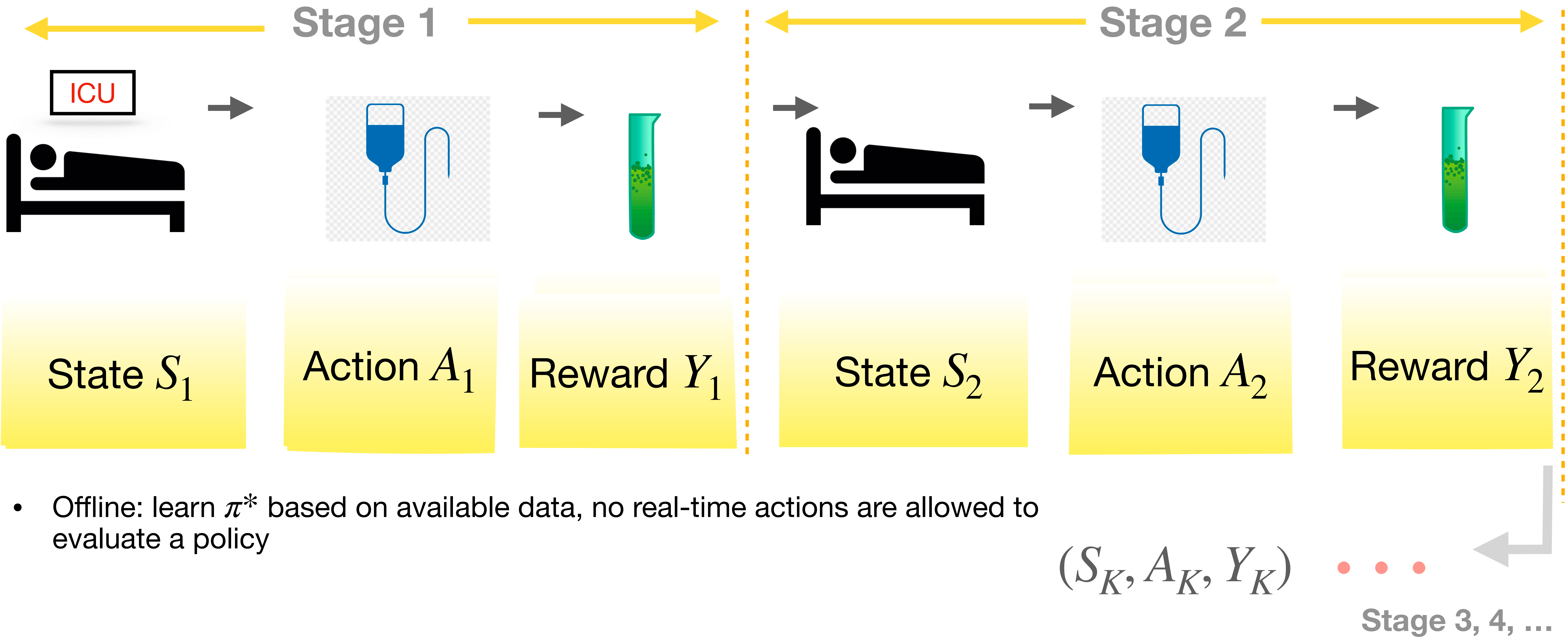
# Offline Reinforcement Learning



# Offline Reinforcement Learning

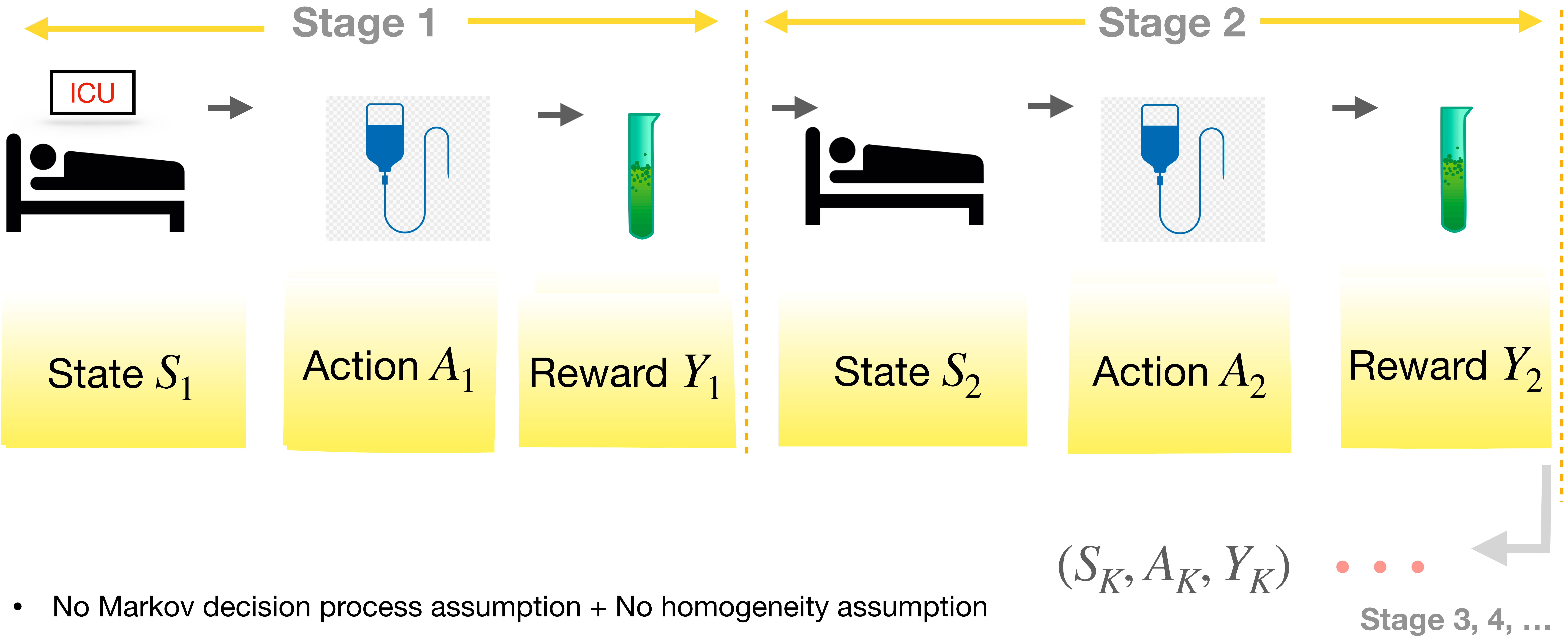


# Offline Reinforcement Learning



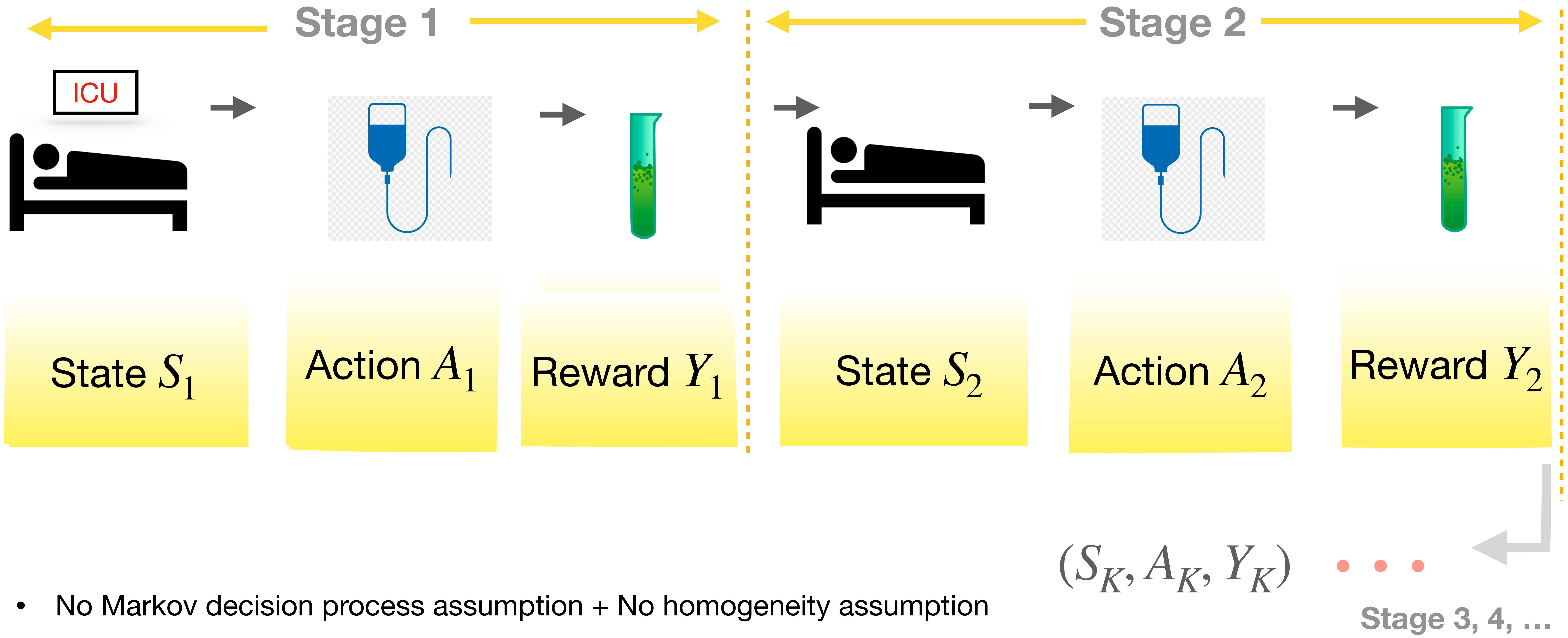

- Offline: learn $\pi^*$ based on available data, no real-time actions are allowed to evaluate a policy

# Offline Reinforcement Learning




- No Markov decision process assumption + No homogeneity assumption

# Offline Reinforcement Learning




- No Markov decision process assumption + No homogeneity assumption
- Hence called Full reinforcement learning

# Outline

- Example: sepsis
- Problem formulation

  A. Mathematical formulation

  B. Existing approaches
- Proposed method
- Open questions

# Outline

- Example: sepsis
- Problem formulation

  A. Mathematical formulation

  B. Existing approaches
- Proposed method
- Open questions

# Mathematical formulation

# History

$$S_1 \rightarrow A_1 \rightarrow Y_1 \rightarrow S_2 \rightarrow A_2 \rightarrow Y_2 \cdots \rightarrow A_K \rightarrow Y_K$$

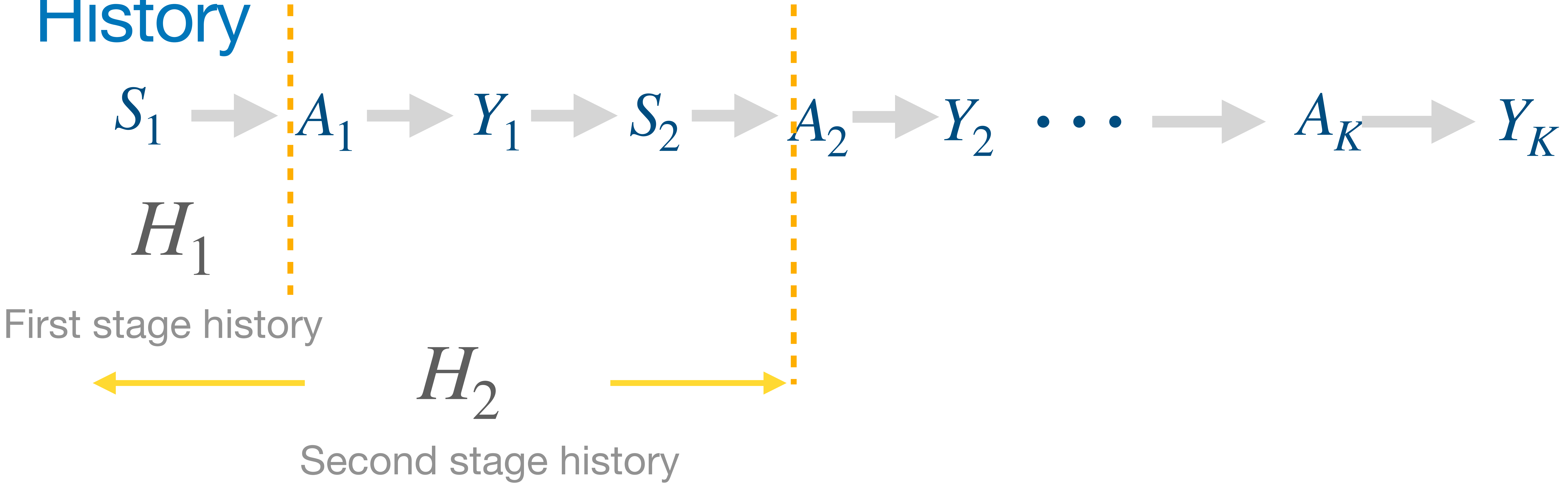
# History

$$S_1 \rightarrow A_1 \rightarrow Y_1 \rightarrow S_2 \rightarrow A_2 \rightarrow Y_2 \cdots \rightarrow A_K \rightarrow Y_K$$

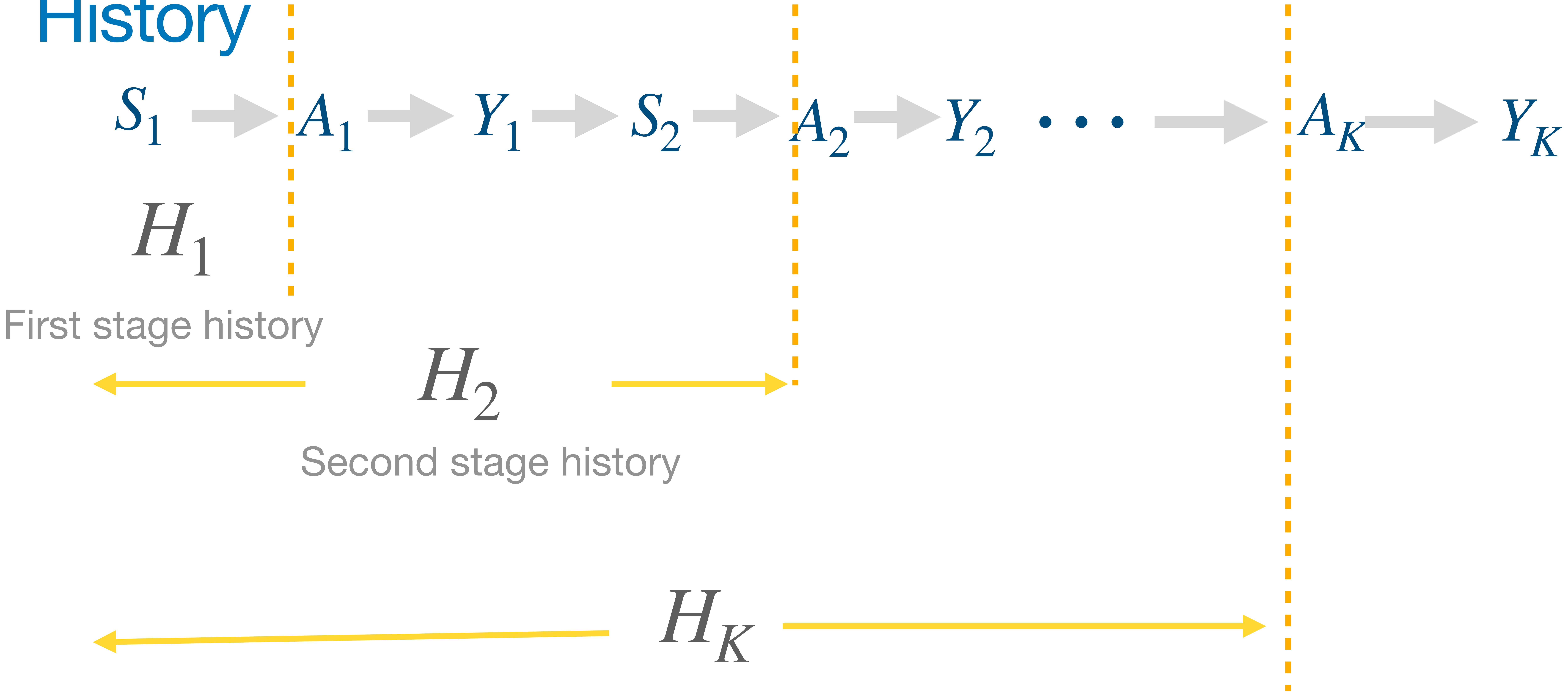
$H_1$

First stage history

# History


$S_1$ $A_1$ $Y_1$ $S_2$ $A_2$ $Y_2$ $\cdots$ $A_K$ $Y_K$
$H_1$
First stage history
$H_2$
Second stage history

# History

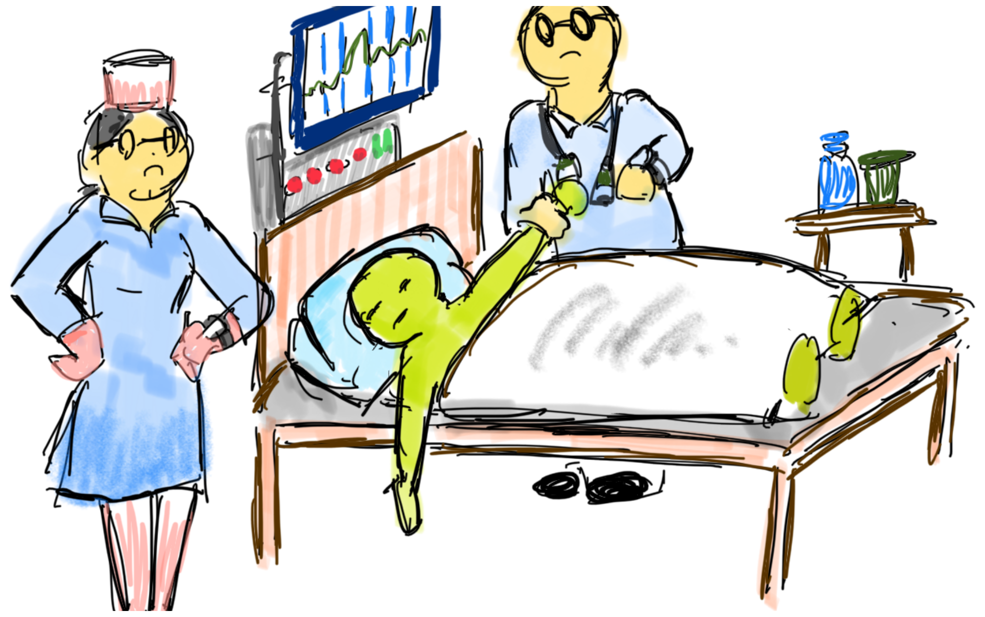



$$H_k = (S_1, A_1, Y_1, \ldots, S_{K-1}, A_{K-1}, Y_{K-1}, S_K)$$

# Policy

# Policy

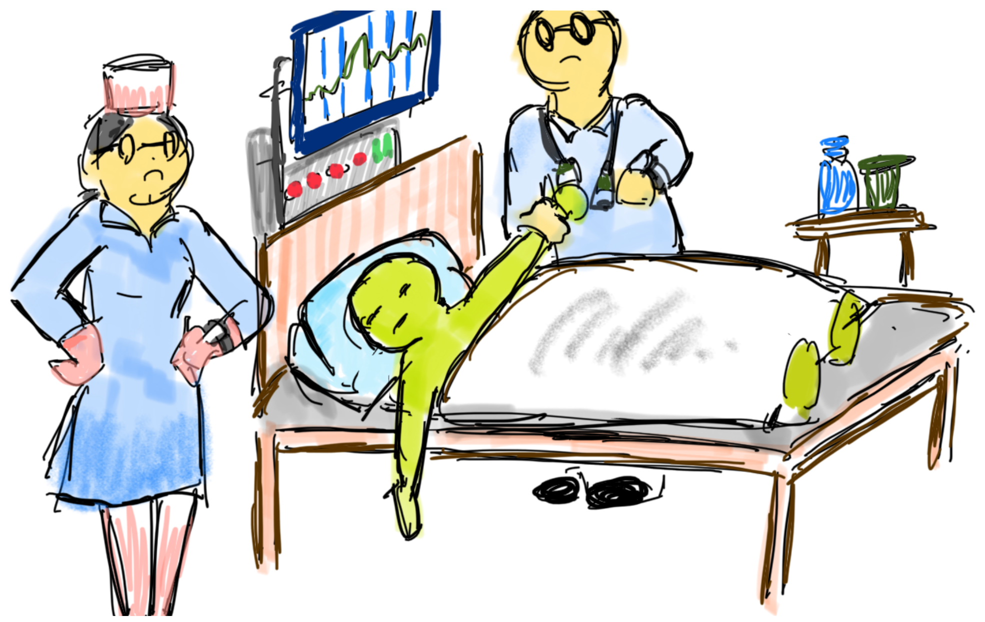
Stage 1



Look at $H_1$

# Policy

Stage 1



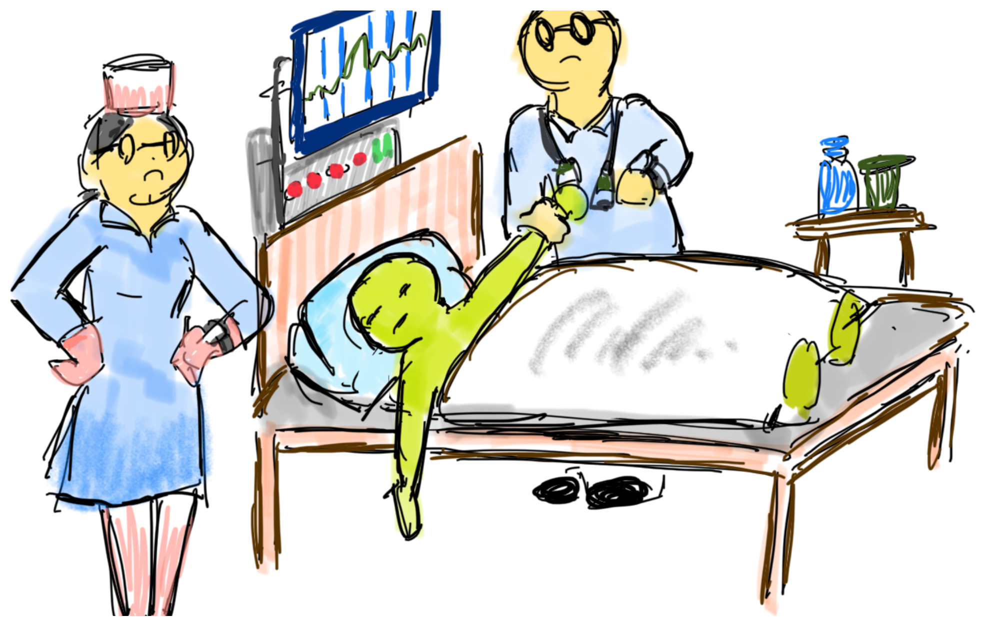
Look at $H_1$

Choose IV-fluid level

$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

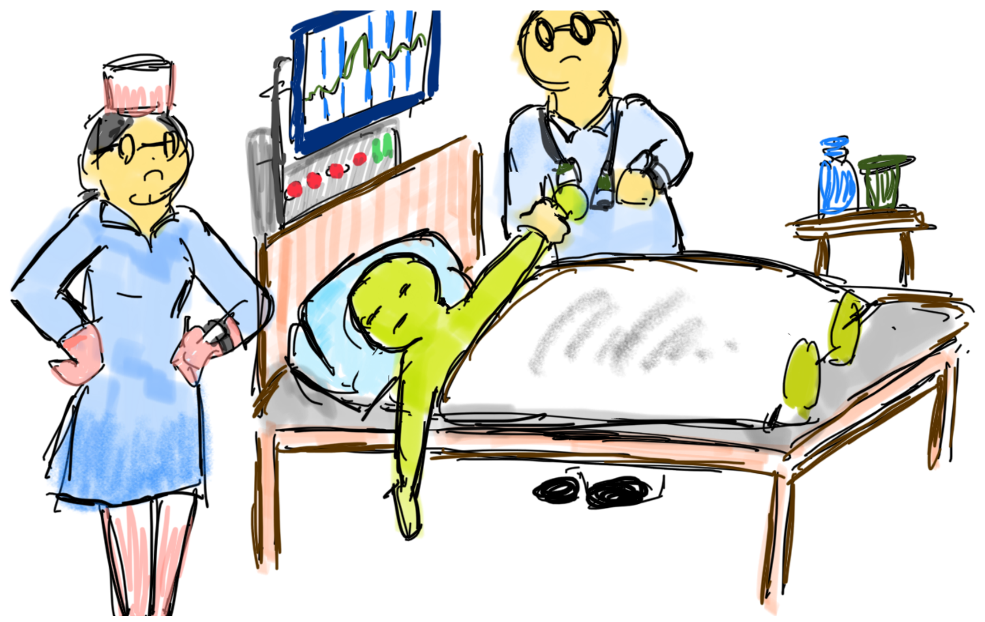
$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

Treatment Assignments

$\pi_1 : H_1 \mapsto \mathscr{A}$

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

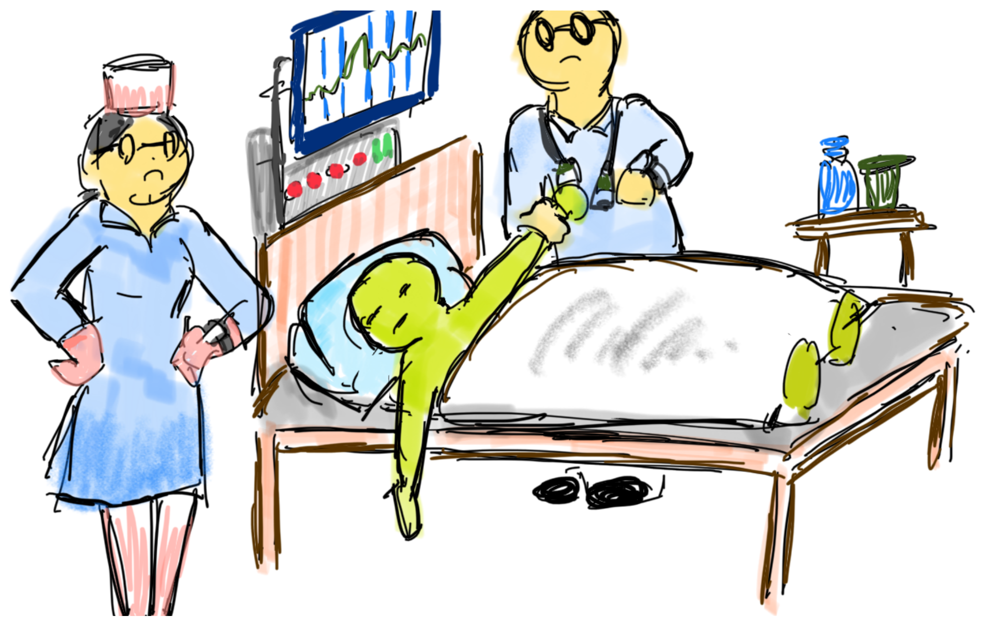
Treatment Assignments

$$\pi_1 : H_1 \mapsto \mathscr{A}$$

**Stages 2, 3, 4, …**

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

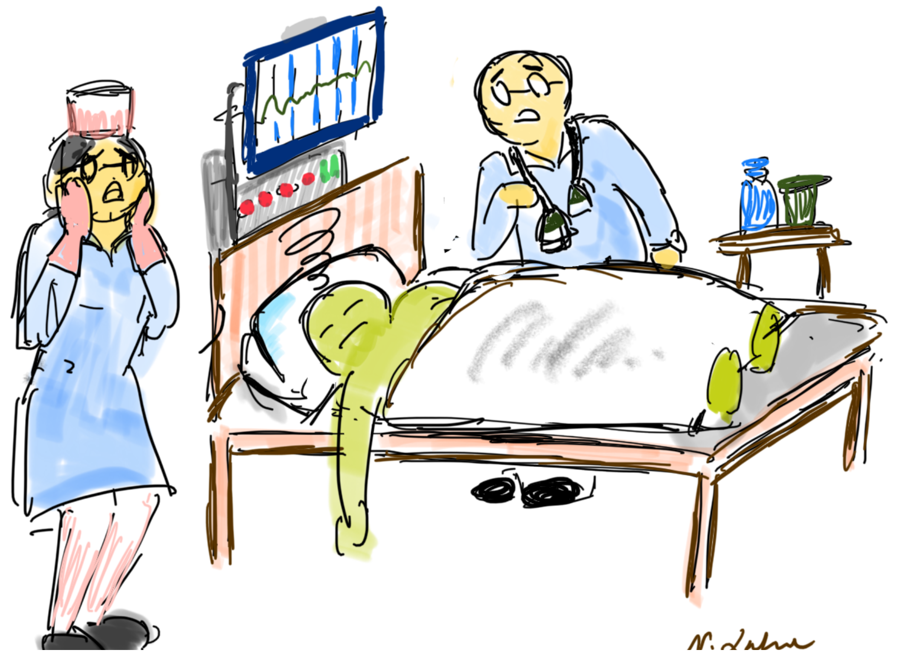
$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

Treatment Assignments

$$\pi_1 : H_1 \mapsto \mathscr{A}$$

Stages 2, 3, 4, ...

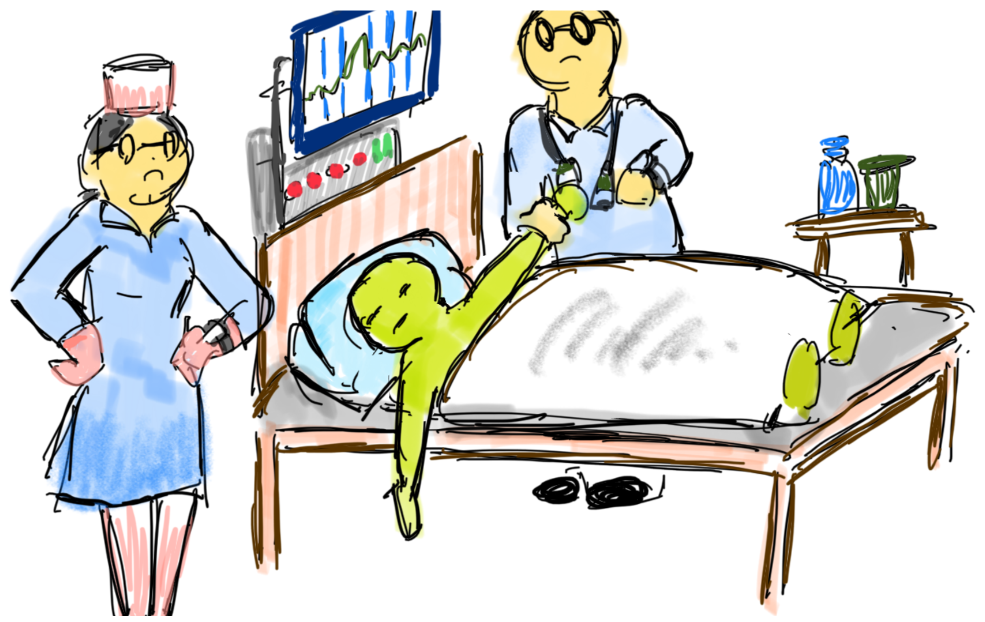
Stage $k$



Look at $H_k$

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

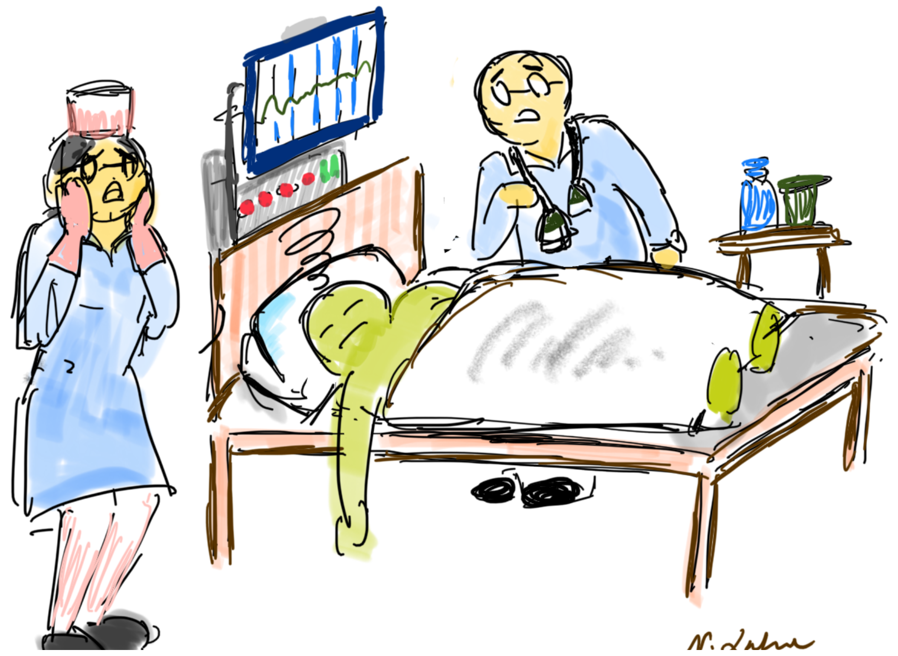
$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

Treatment Assignments

$$\pi_1 : H_1 \mapsto \mathscr{A}$$

**Stages 2, 3, 4, …**

Stage $k$



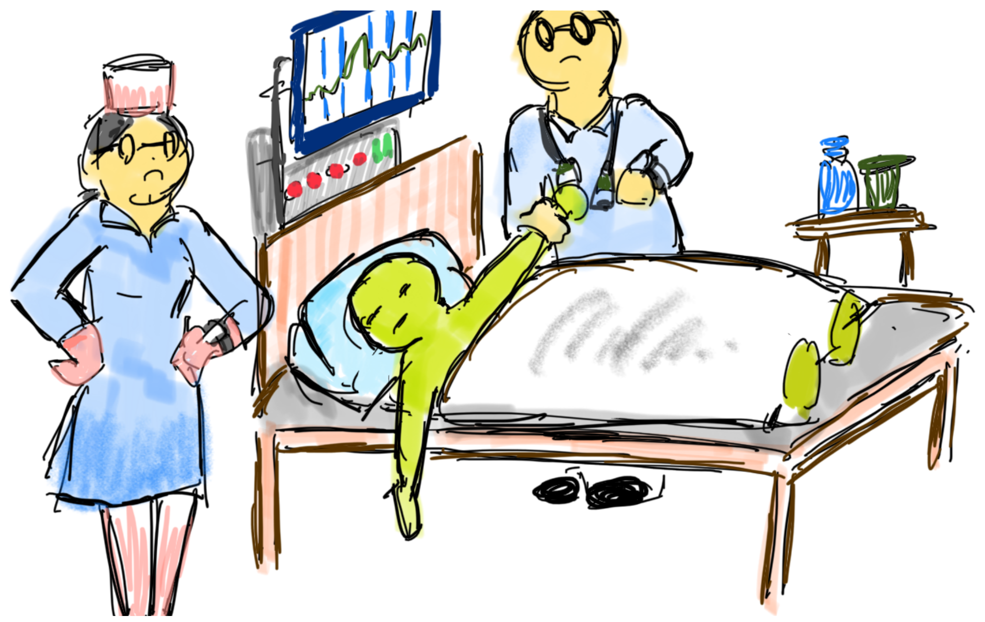
Look at $H_k$

Choose IV-fluid level

$\pi_k(H_k) \in$ {no fluid, low, mid, high}

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

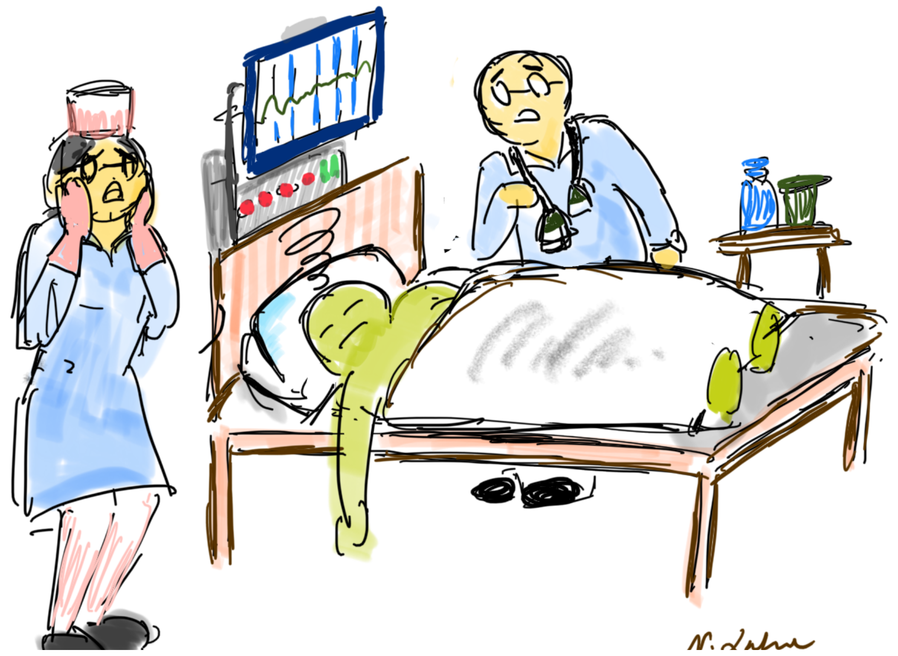
$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

Treatment Assignments

$\pi_1 : H_1 \mapsto \mathscr{A}$

**Stages 2, 3, 4, …**

Stage $k$



Look at $H_k$

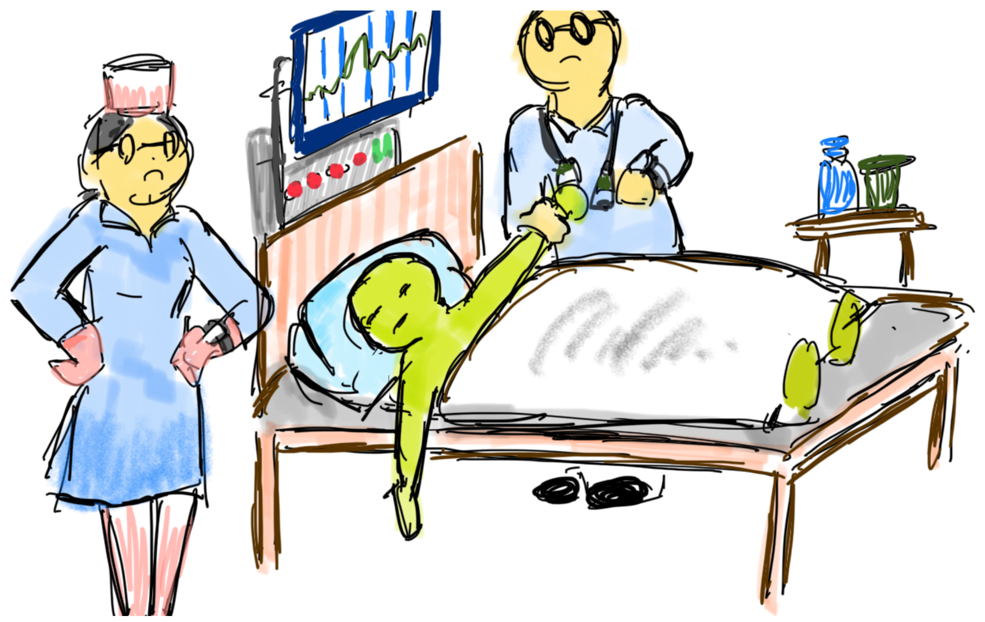
Choose IV-fluid level

$\pi_k(H_k) \in$ {no fluid, low, mid, high}

$\pi_k : H_k \mapsto \mathscr{A}$

# Policy

Stage 1



Look at $H_1$

Choose IV-fluid level

$\pi_1(H_1) \in$ $\mathscr{A}$ = {no fluid, low, mid, high}

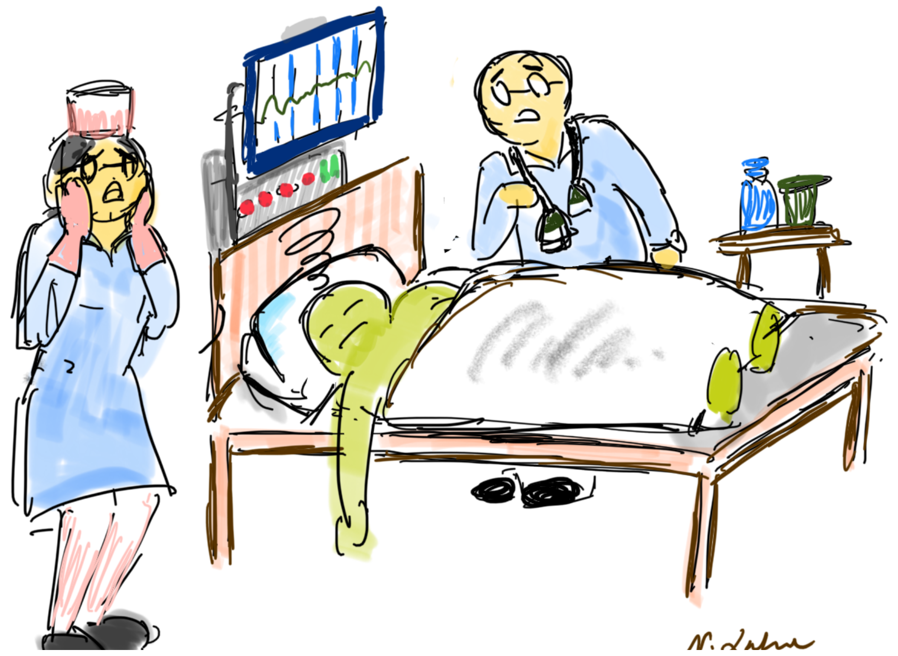
Treatment Assignments

$$\pi_1 : H_1 \mapsto \mathscr{A}$$

Stages 2, 3, 4, …

Policy $\pi = (\pi_1, \ldots, \pi_K)$

Stage $k$



Look at $H_k$

Choose IV-fluid level

$\pi_k(H_k) \in$ {no fluid, low, mid, high}

$$\pi_k : H_k \mapsto \mathscr{A}$$

# Example of policy in full RL

# Example of policy in full RL

Stage 1:

# Example of policy in full RL

Stage 1:

$$\pi_1$$

# Example of policy in full RL

Systolic blood pressure $\leq$ 90 mm Hg

Stage 1:

$$\pi_1$$

# Example of policy in full RL

Systolic blood pressure $\leq$ 90 mm Hg

Yes

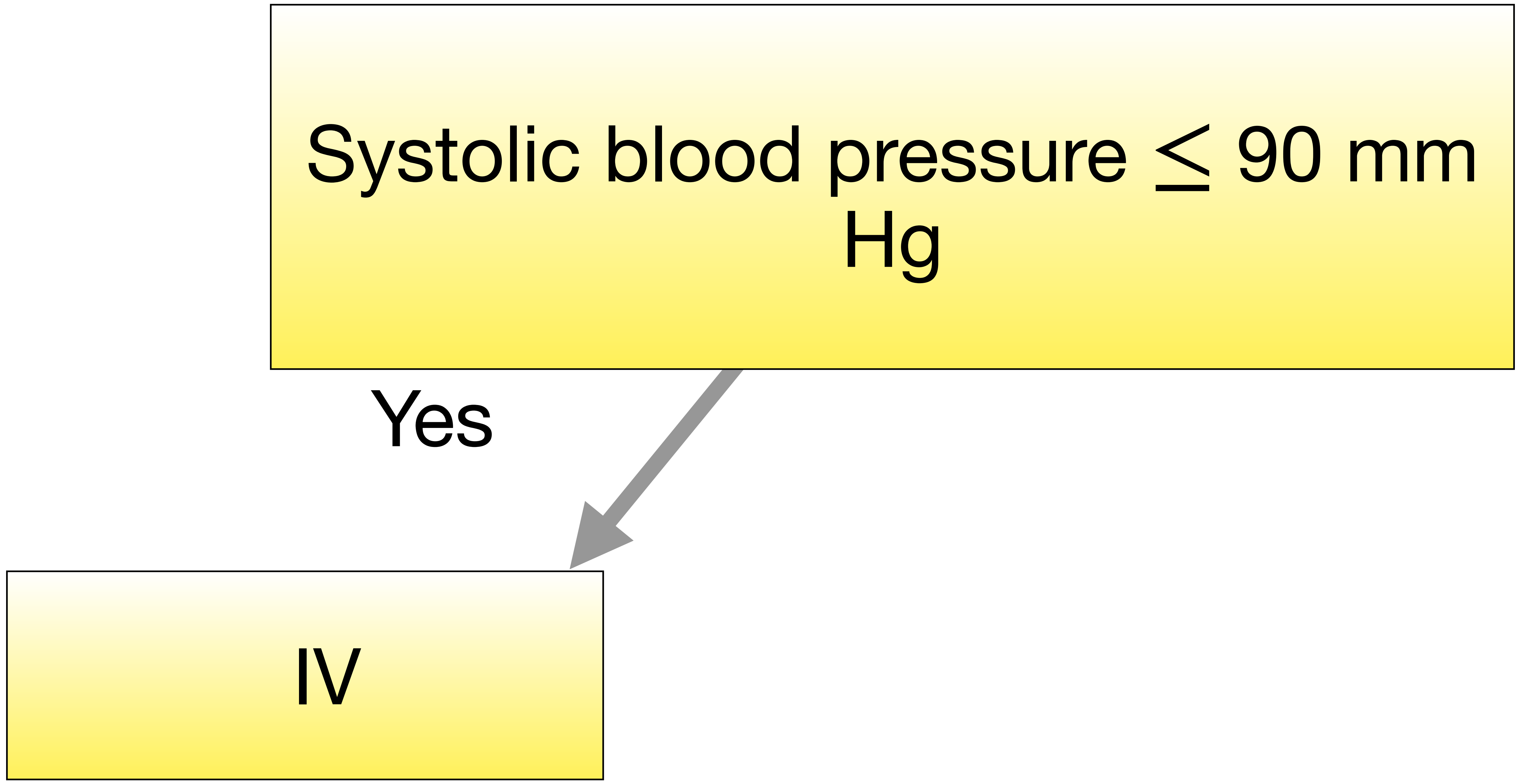
Stage 1:

$\pi_1$

# Example of policy in full RL

Stage 1:

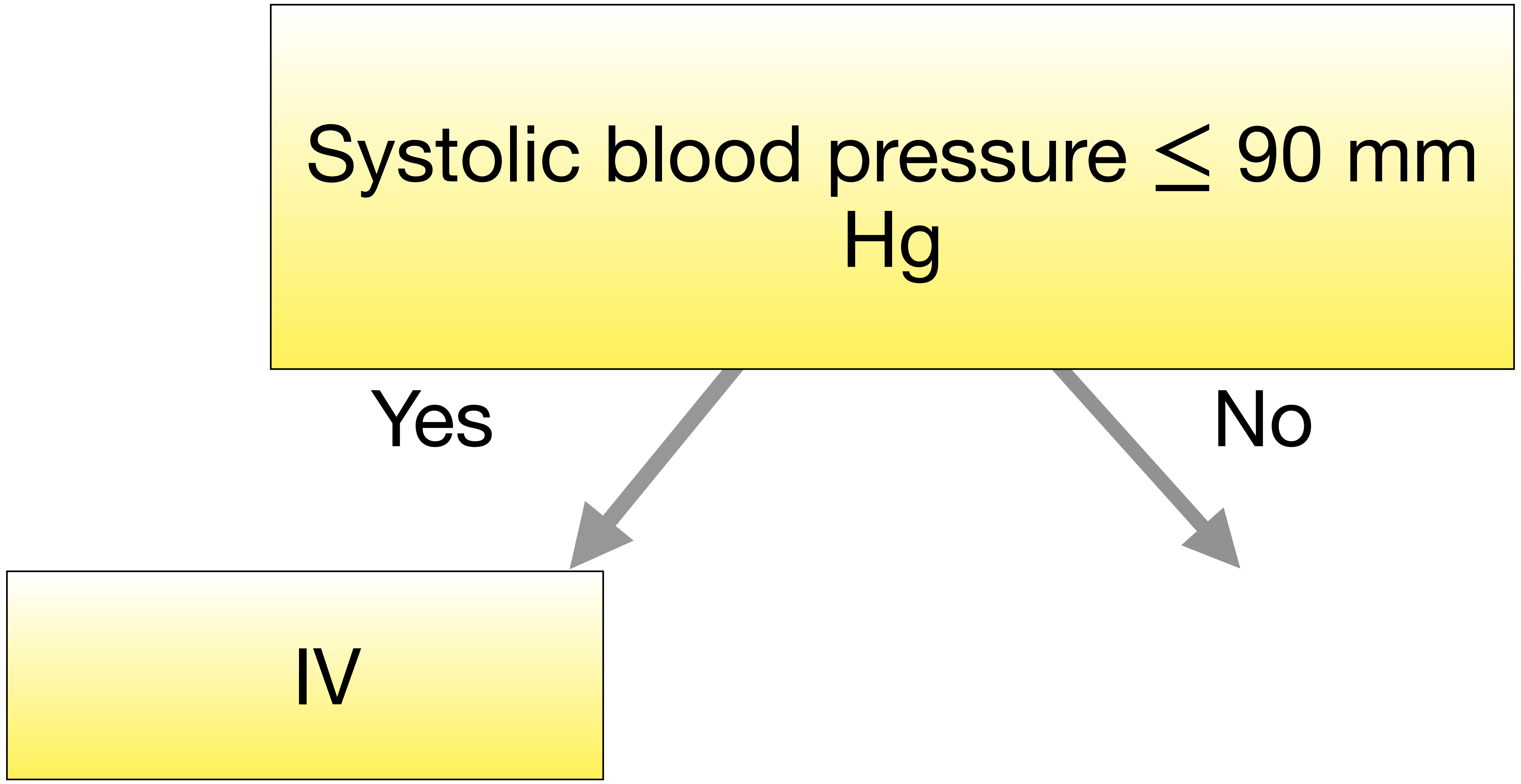
$\pi_1$



# Example of policy in full RL

Stage 1:

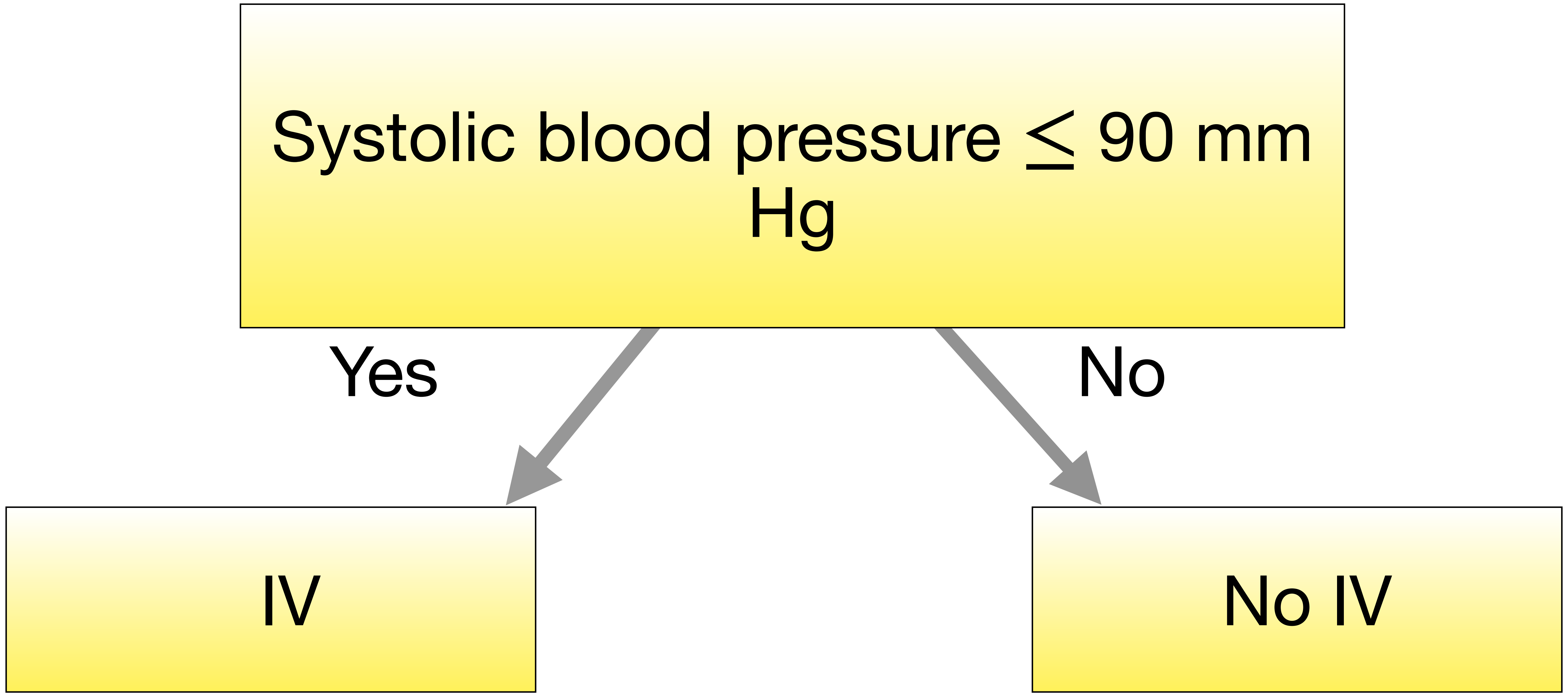
$\pi_1$



# Example of policy in full RL

Stage 1:

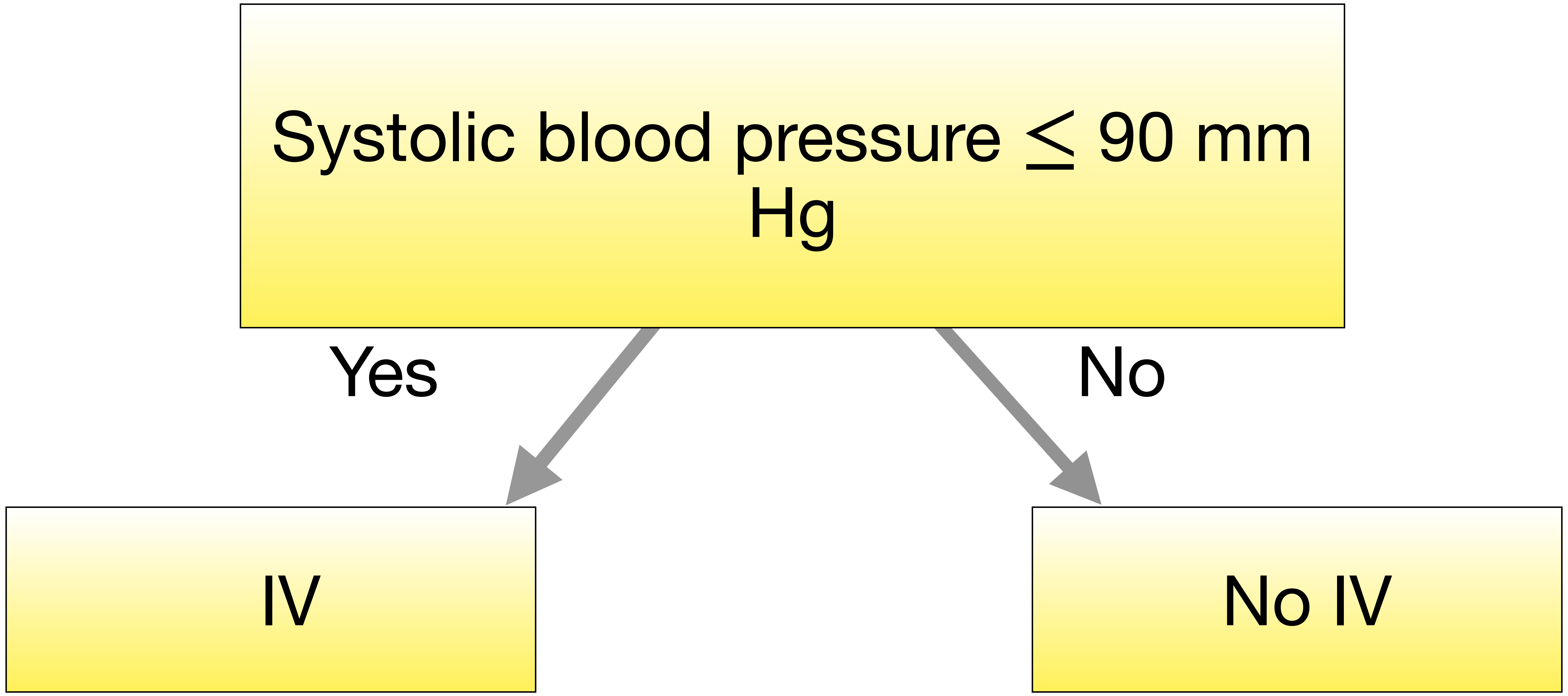
$\pi_1$



# Example of policy in full RL

Stage 1:

$\pi_1$



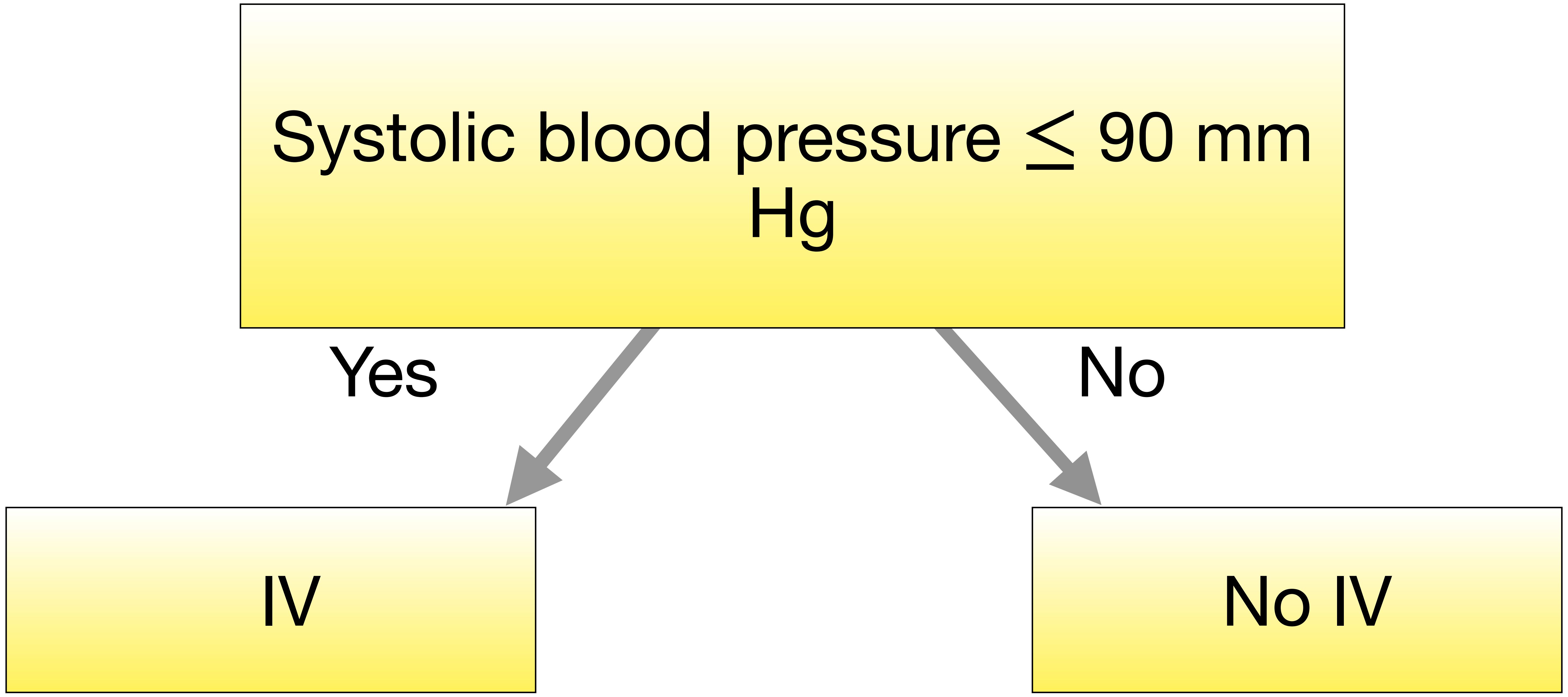

Stage $k$:

$t = 2, \ldots, K$.

# Example of policy in full RL

Stage 1:

$\pi_1$




Stage $k$:

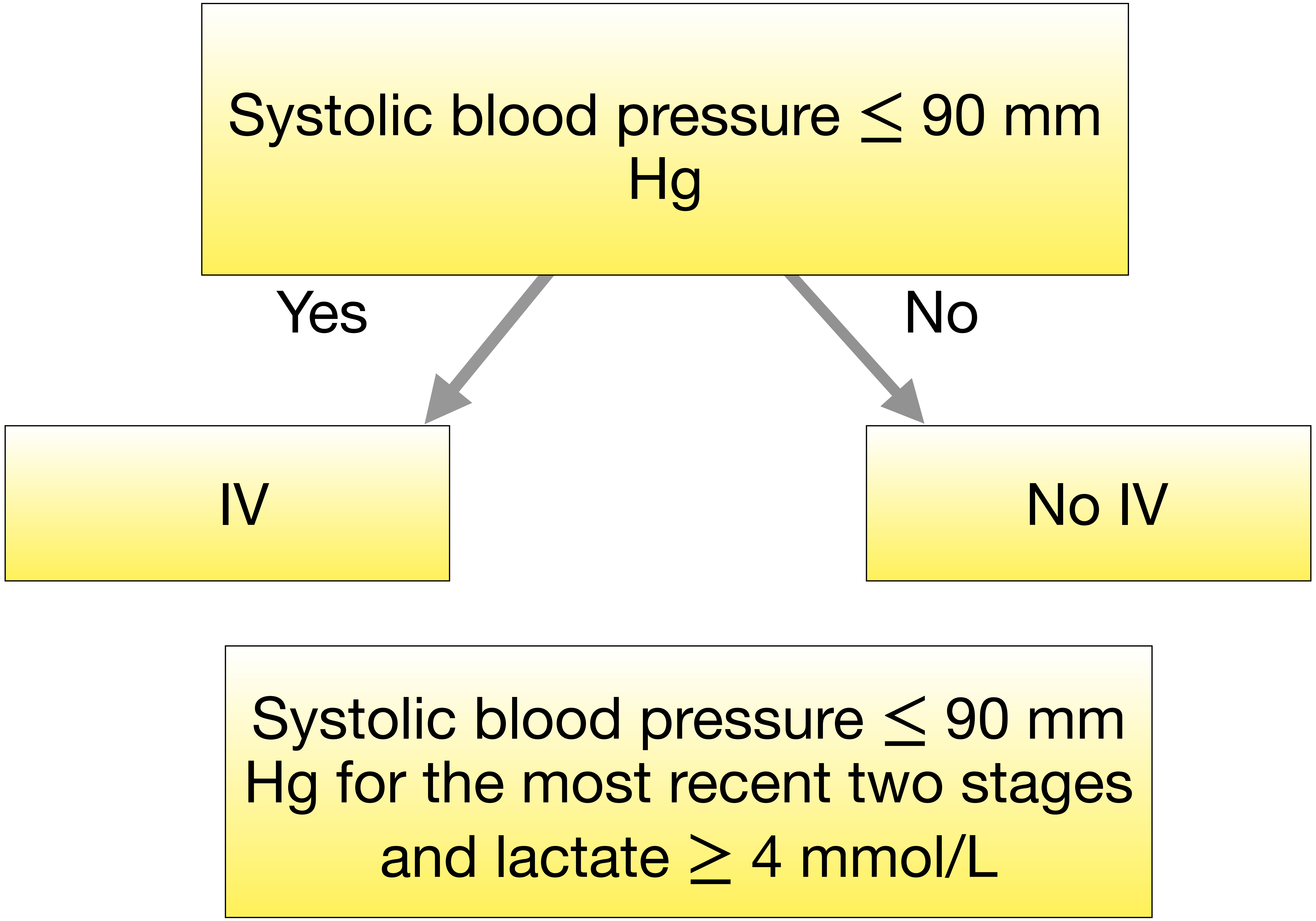
$t = 2, \ldots, K$.

$\pi_k$

# Example of policy in full RL

Stage 1:

$\pi_1$




Stage $k$:

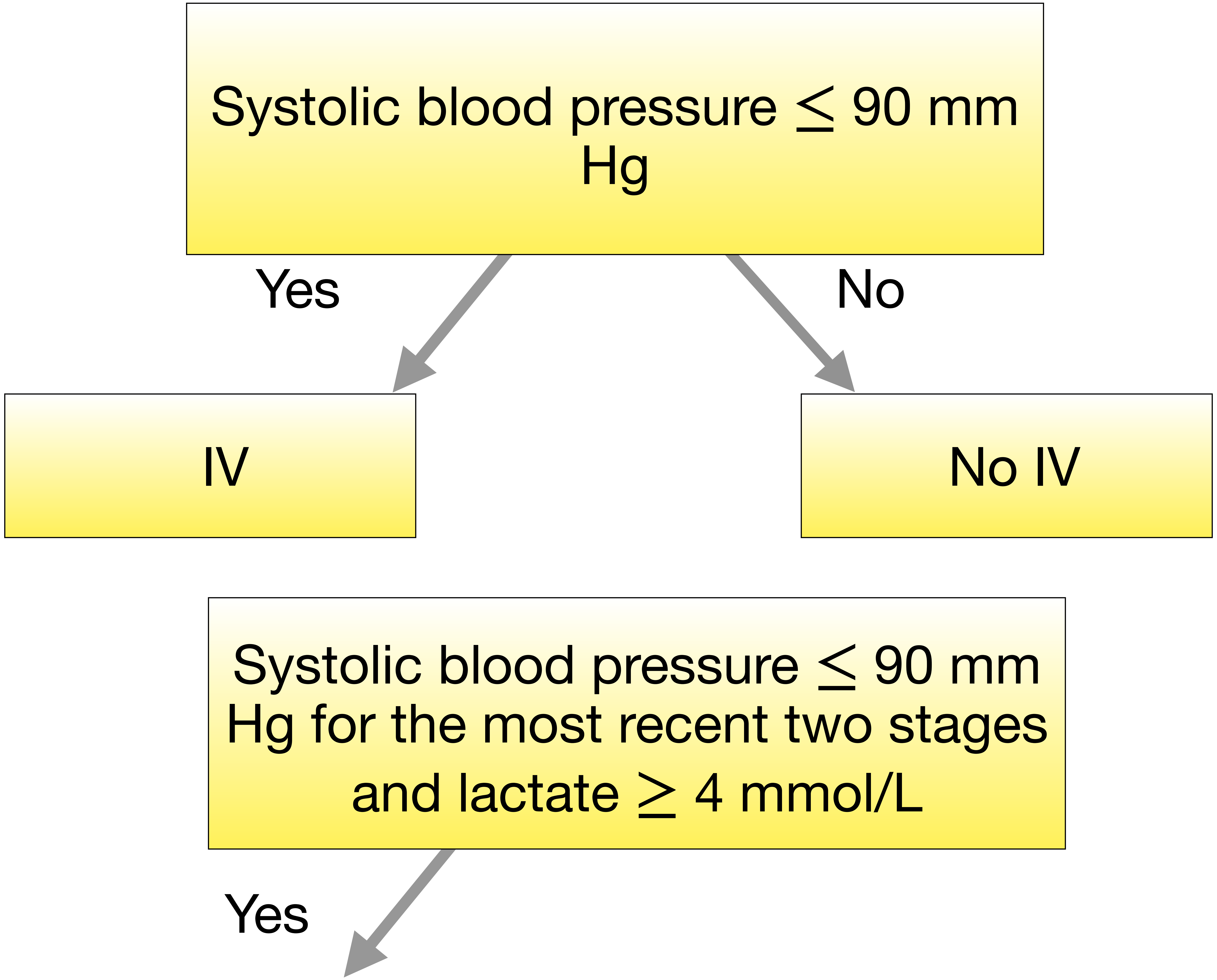
$t = 2, \ldots, K$.

$\pi_k$

# Example of policy in full RL

Stage 1:

$\pi_1$




Stage $k$:

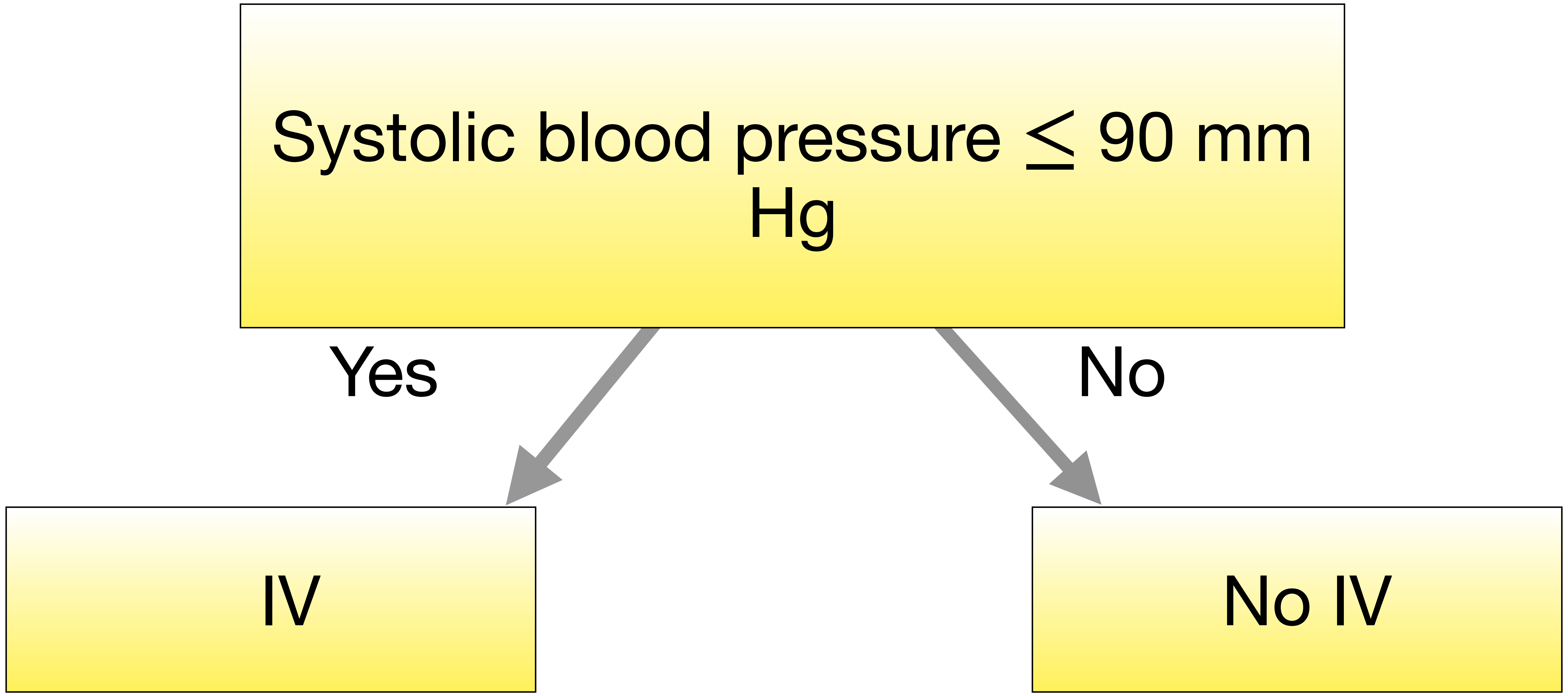
$t = 2, \ldots, K$.

$\pi_k$

# Example of policy in full RL

Stage 1:

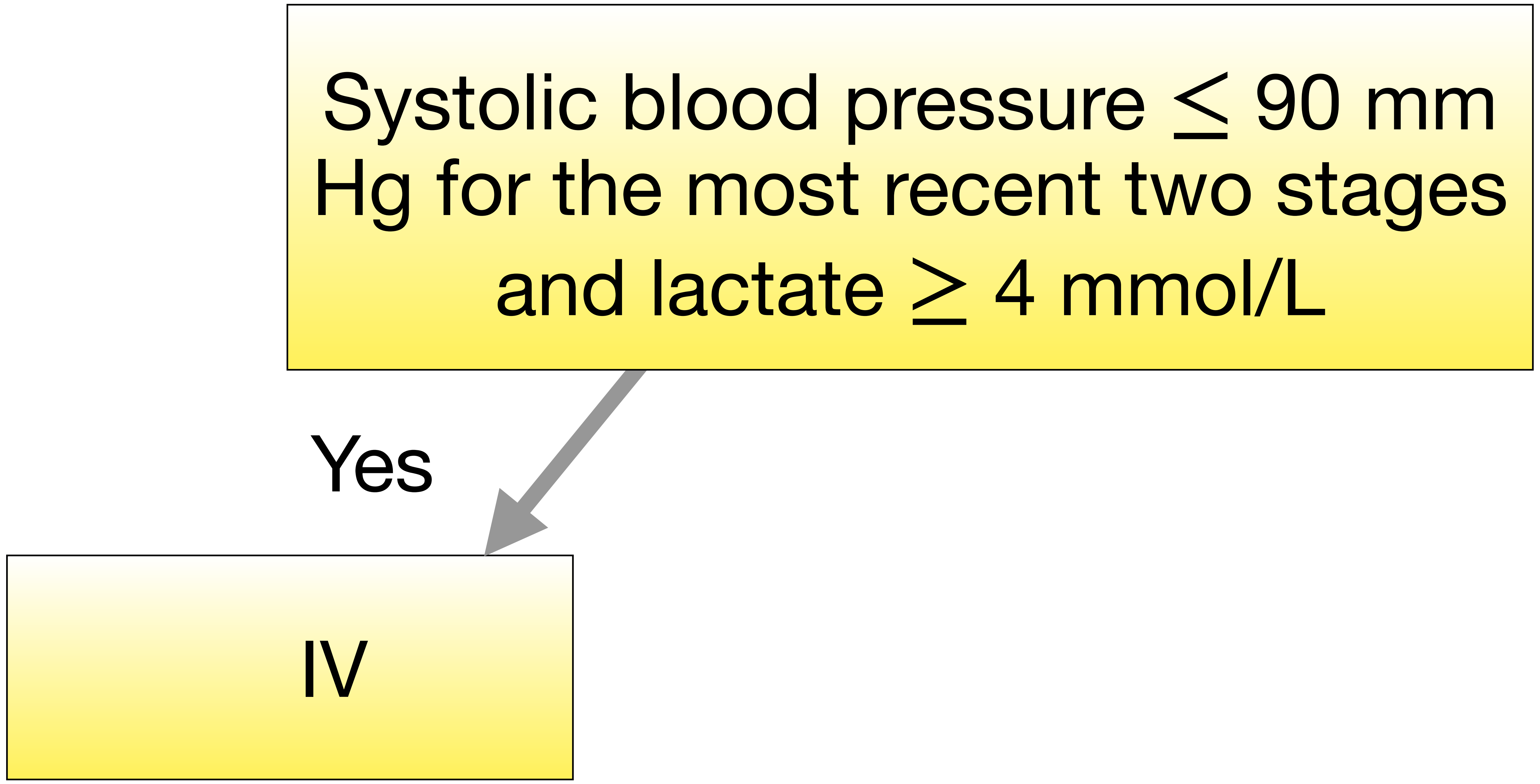
$\pi_1$




Stage $k$:

$t = 2, \ldots, K$.

$\pi_k$

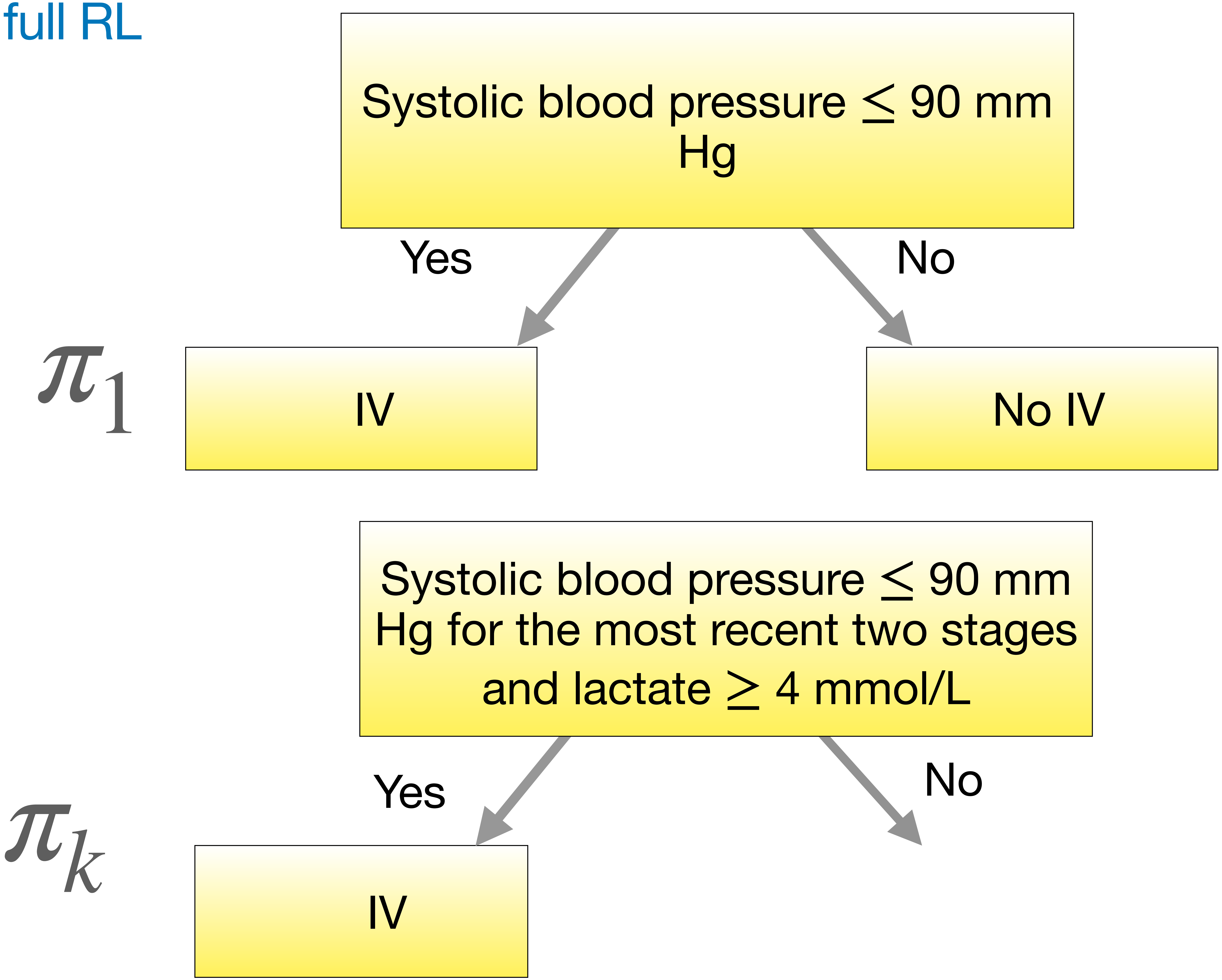


# Example of policy in full RL

Stage 1:

Stage $k$:

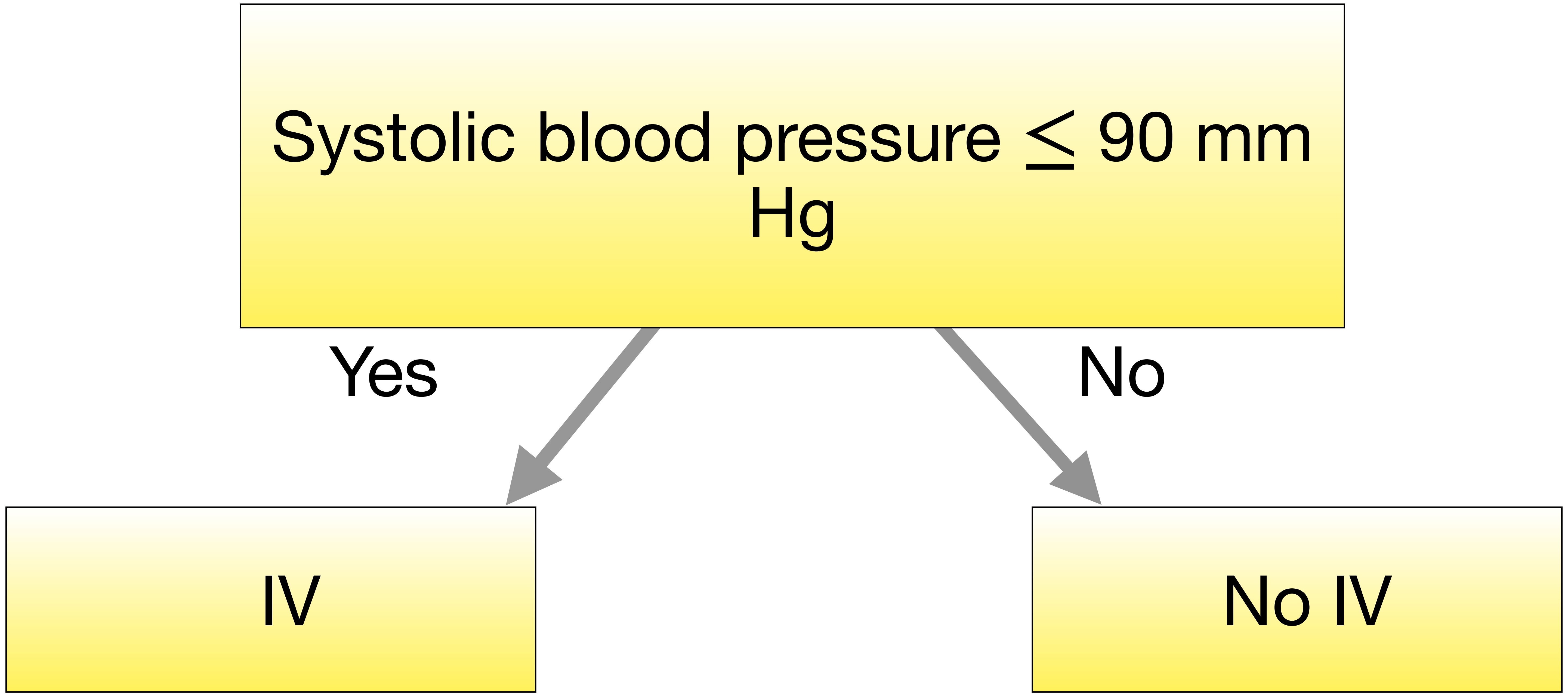
$t = 2, \ldots, K$.



# Example of policy in full RL

Stage 1:

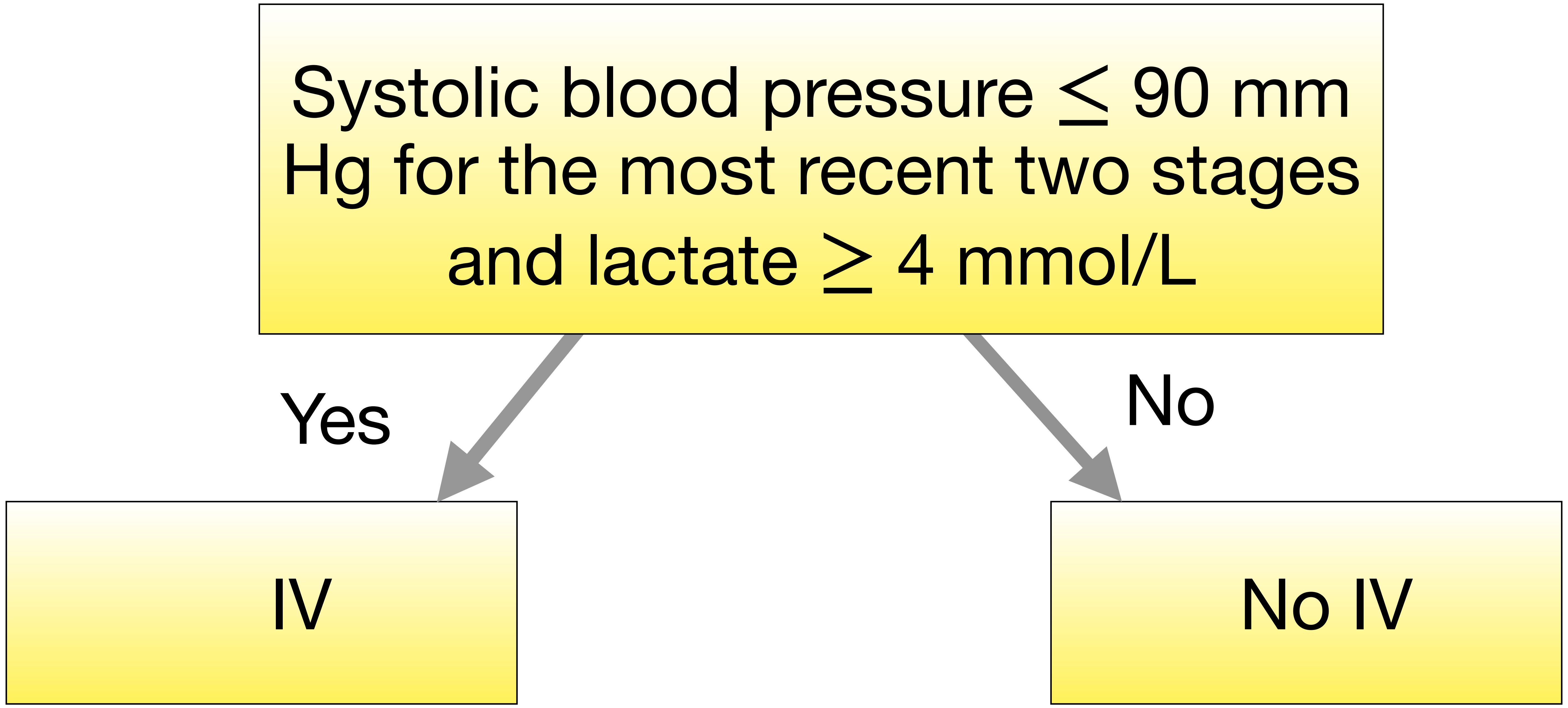
$\pi_1$




Stage $k$:

$t = 2, \ldots, K$.

$\pi_k$



# Optimal treatment policy: value function

# Optimal treatment policy: value function

$Y_k(\pi)$ : potential outcome at stage $k$ had policy $\pi$ been followed

# Optimal treatment policy: value function

$Y_k(\pi)$ : potential outcome at stage $k$ had policy $\pi$ been followed

Value function of $\pi$ :

$$V^{\pi} = E[Y_1(\pi)\ldots + Y_K(\pi)]$$

# Optimal treatment policy: value function

$Y_k(\pi)$ : potential outcome at stage $k$ had policy $\pi$ been followed

Value function of $\pi$ :

$$V^{\pi} = E[Y_1(\pi) \ldots + Y_K(\pi)]$$

$$\pi^* = \text{argmax}_{\pi} V^{\pi}$$

# Optimal treatment policy: value function

$Y_k(\pi)$ : potential outcome at stage $k$ had policy $\pi$ been followed

Value function of $\pi$ :

$$V^{\pi} = E[Y_1(\pi) \ldots + Y_K(\pi)]$$

$$\pi^* = \text{argmax}_{\pi} V^{\pi}$$

Optimal policy

# Optimal treatment policy: value function

$Y_k(\pi)$ : potential outcome at stage $k$ had policy $\pi$ been followed

Value function of $\pi$ :

$$V^{\pi} = E[Y_1(\pi) \ldots + Y_K(\pi)]$$

Not observed random variables

$$\pi^* = \text{argmax}_{\pi} V^{\pi}$$

# Outline

- Example: sepsis
- Problem formulation

  A. Mathematical formulation

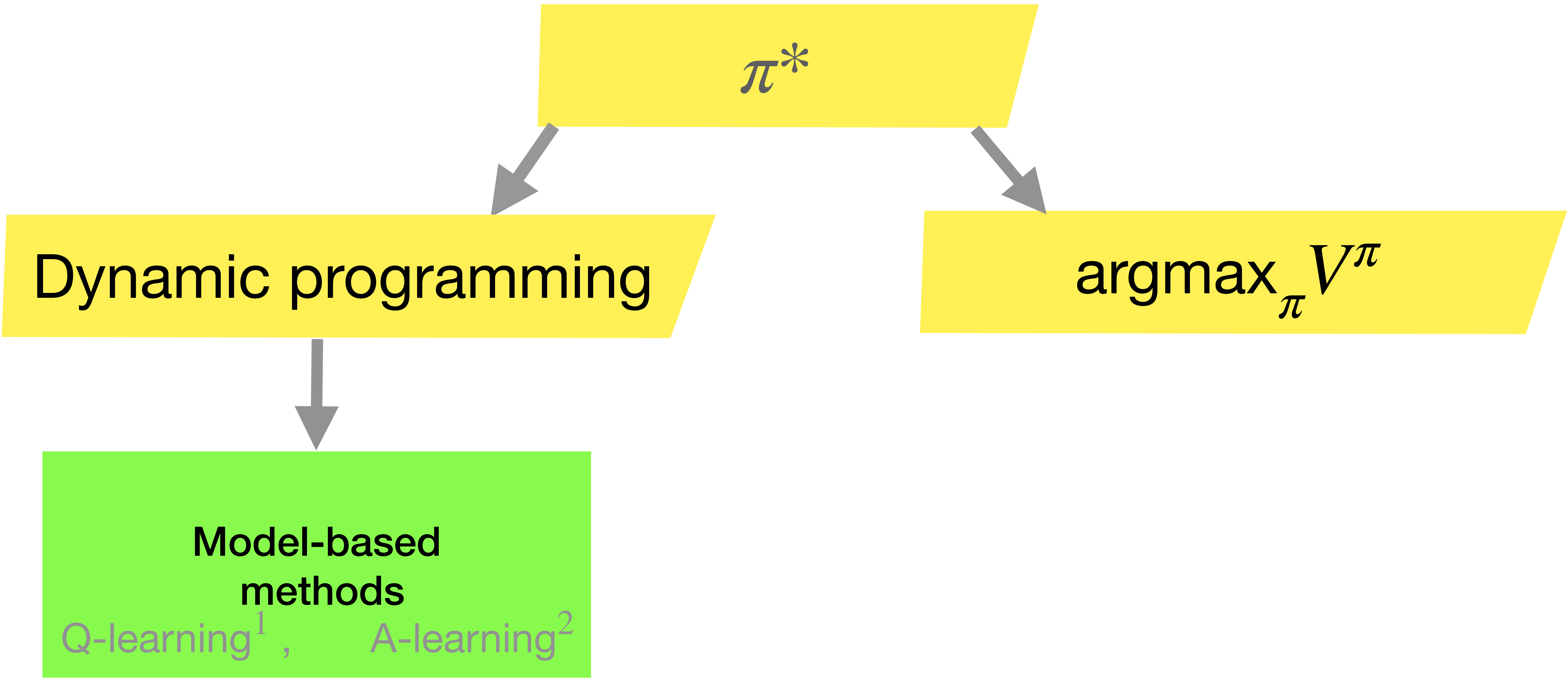
  B. Existing approaches
- Proposed method
- Open questions

# Existing approaches

# Estimation of $\pi^*$

# Estimation of $\pi^*$

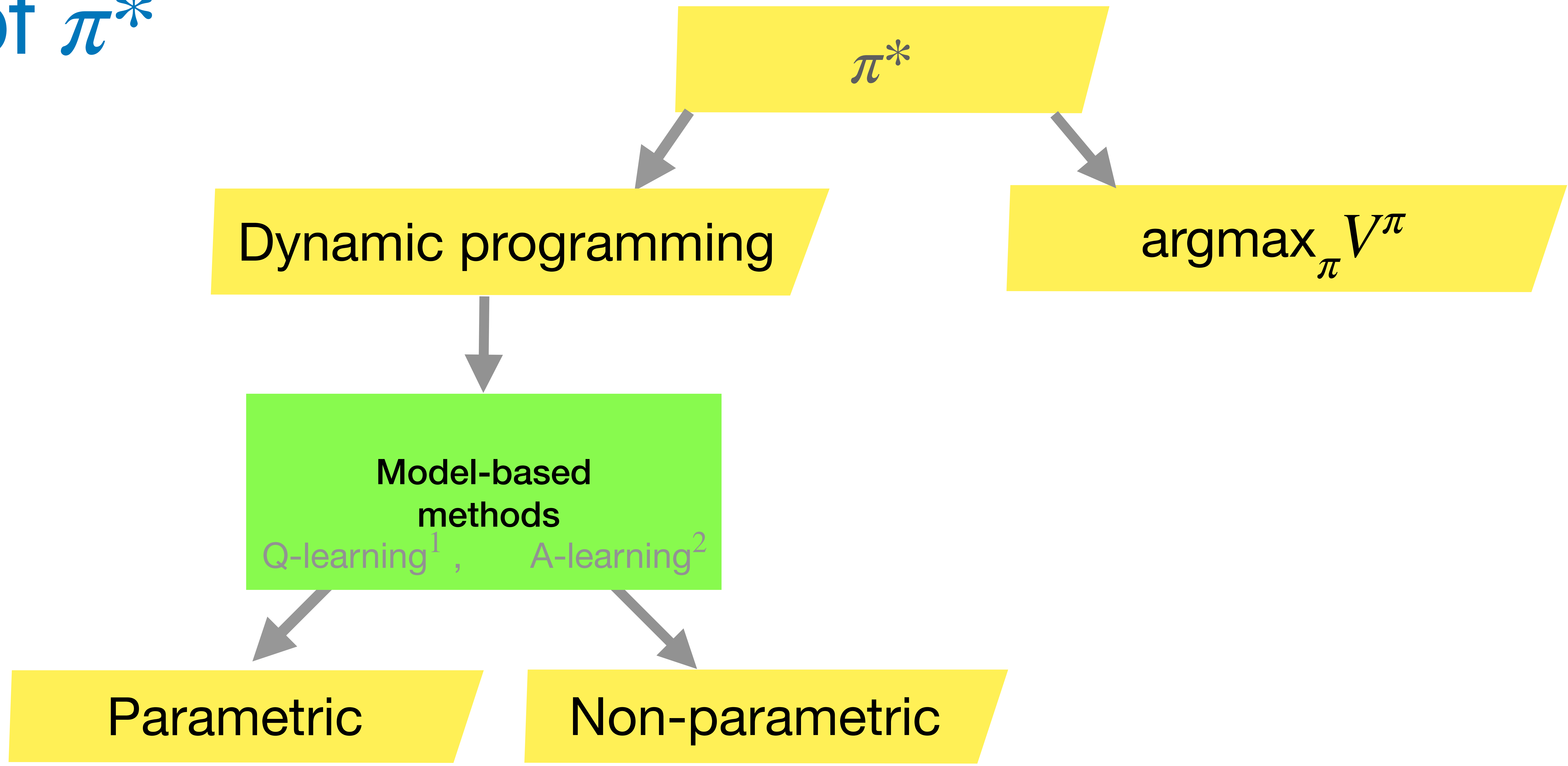
# Estimation of $\pi^*$





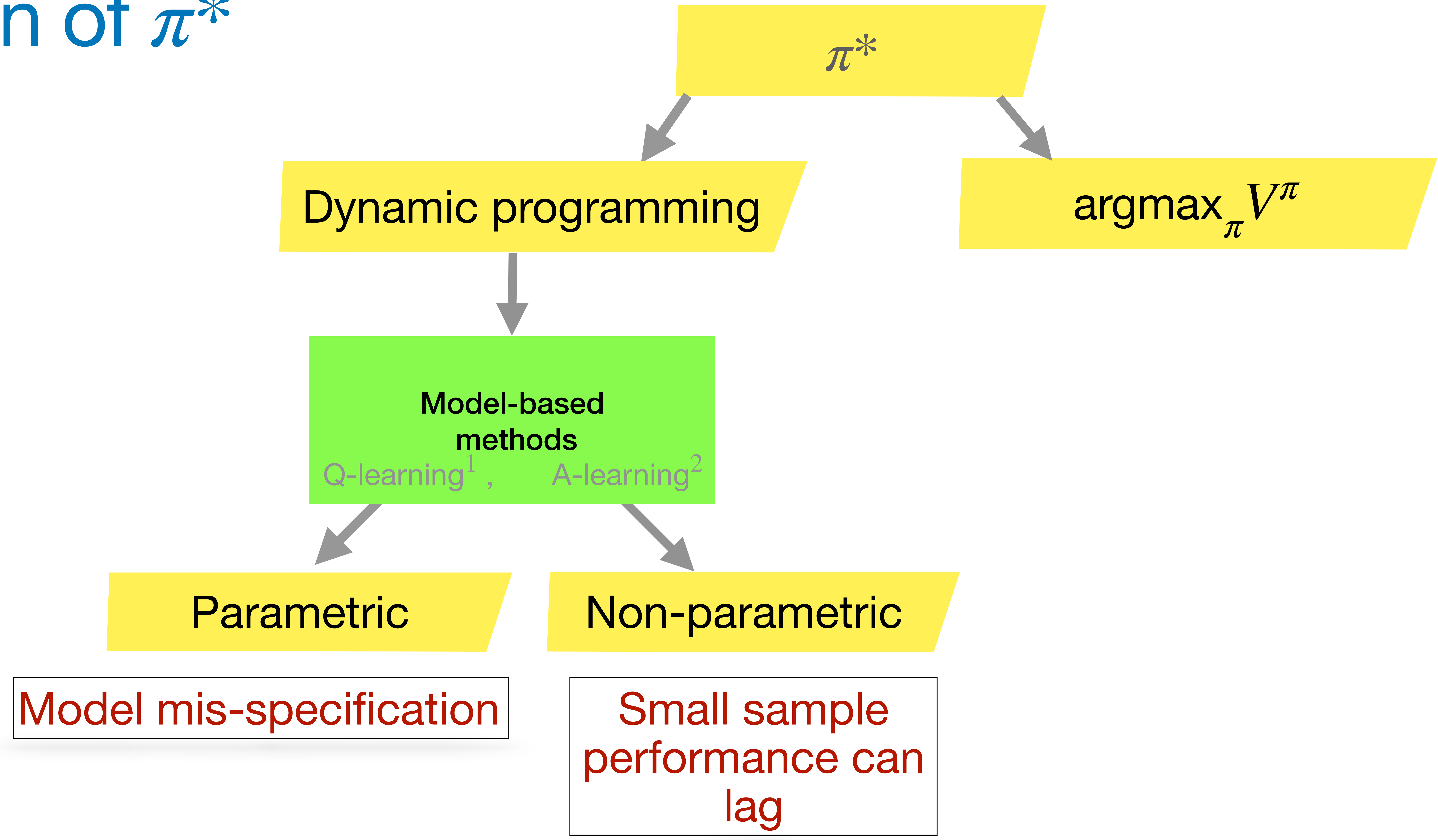
1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004

# Estimation of $\pi^*$



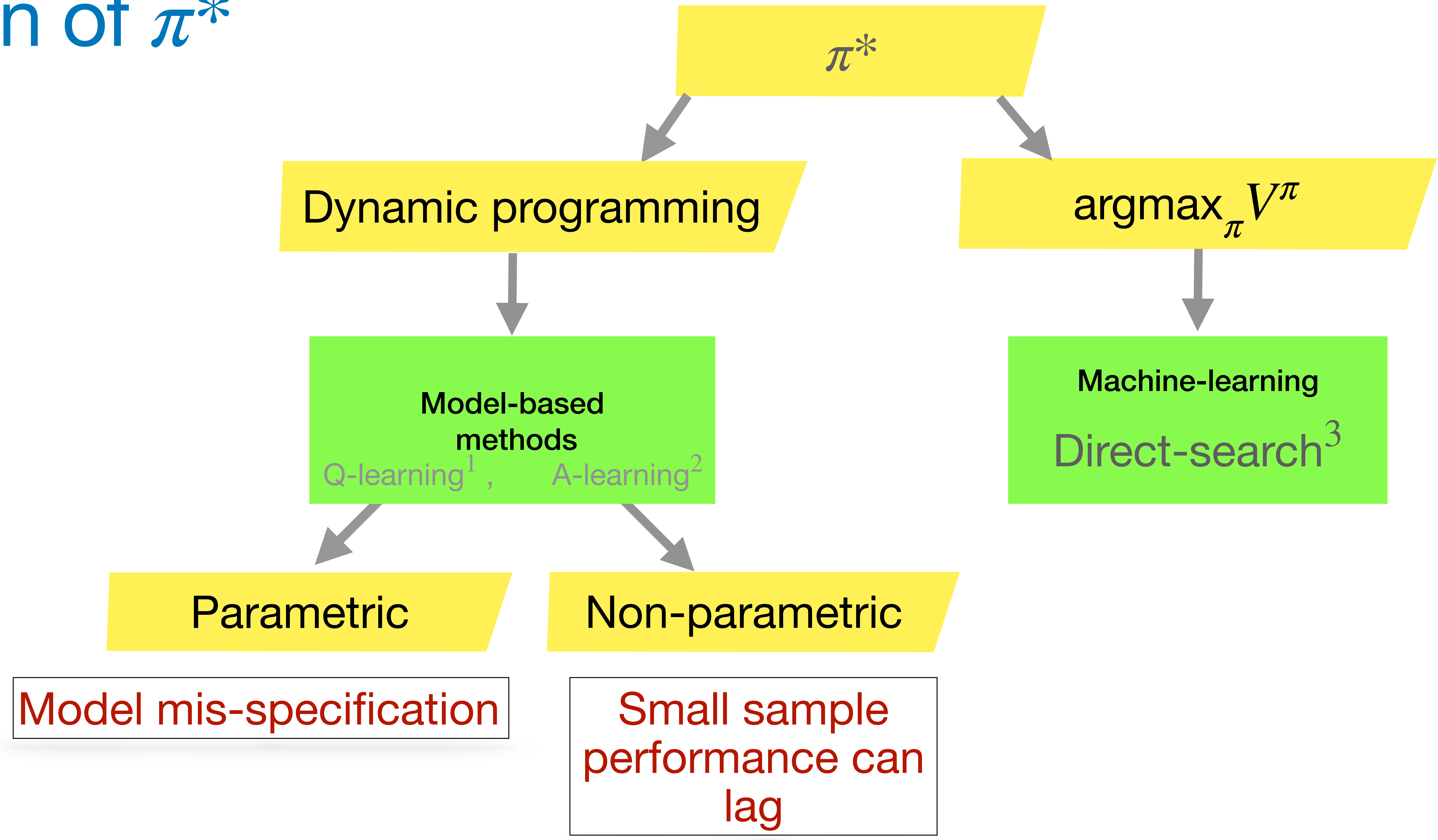

1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004

# Estimation of $\pi^*$



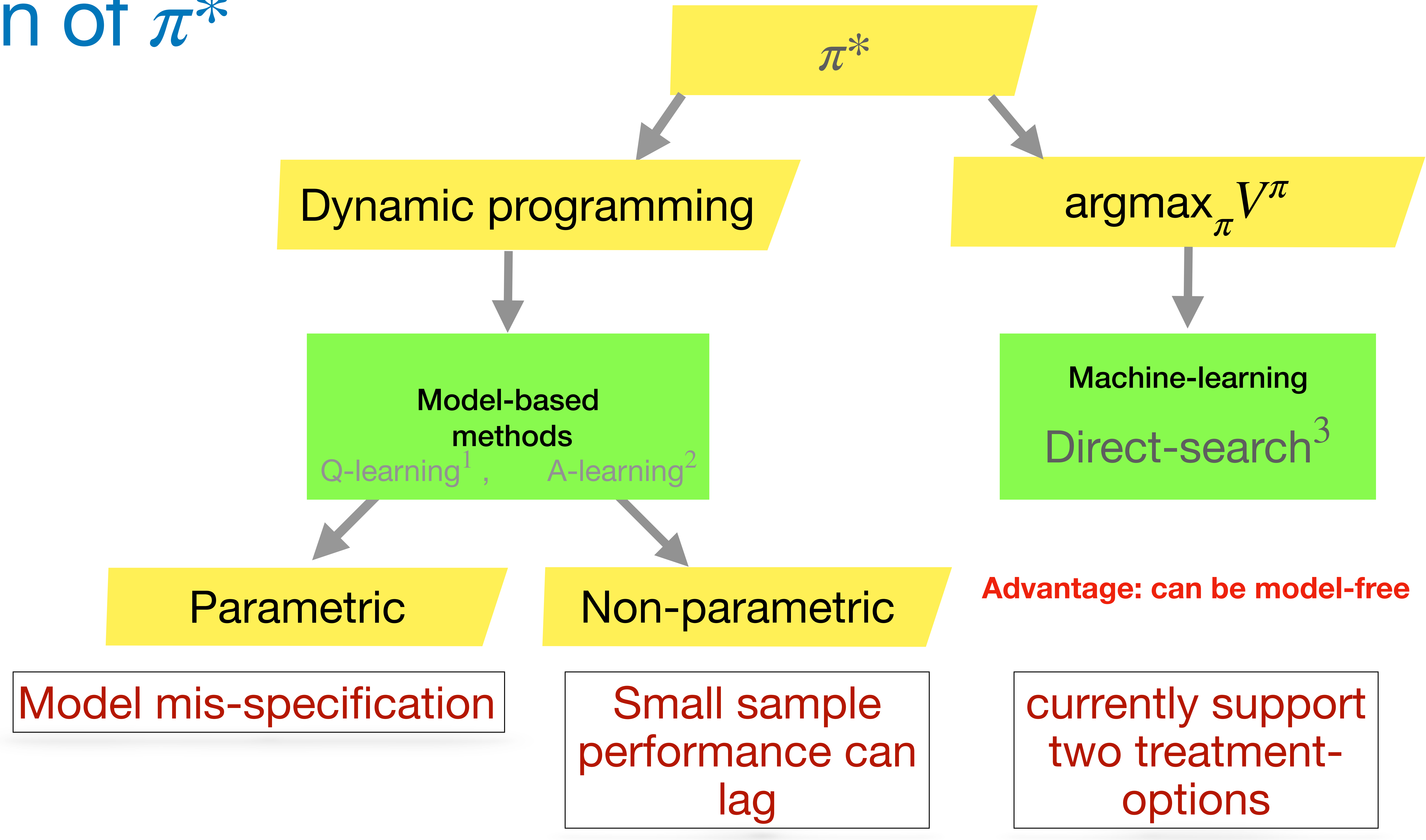

1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004

# Estimation of $\pi^*$




1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2022

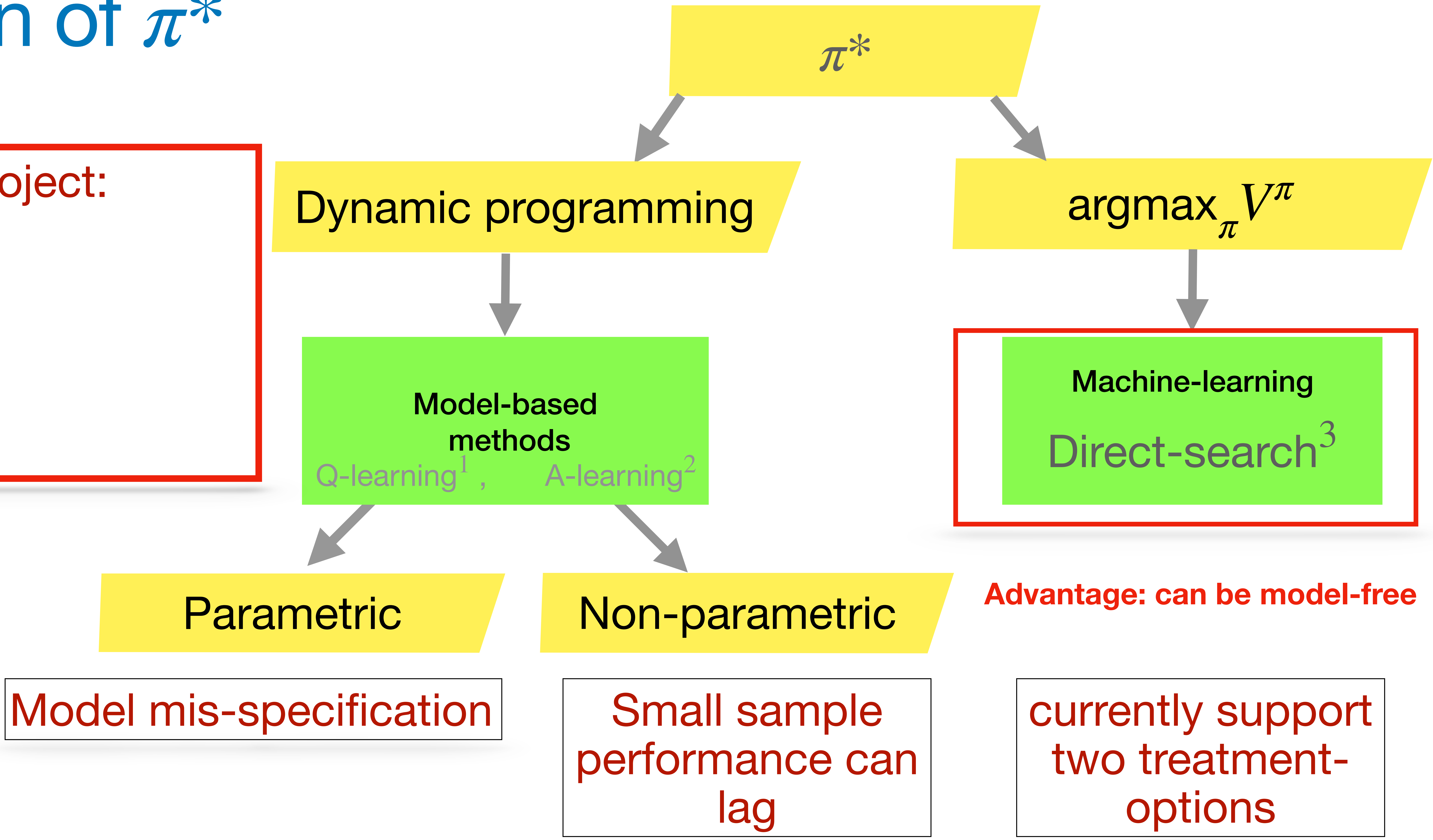
# Estimation of $\pi^*$




1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2022

# Estimation of $\pi^*$

Goal of the project:

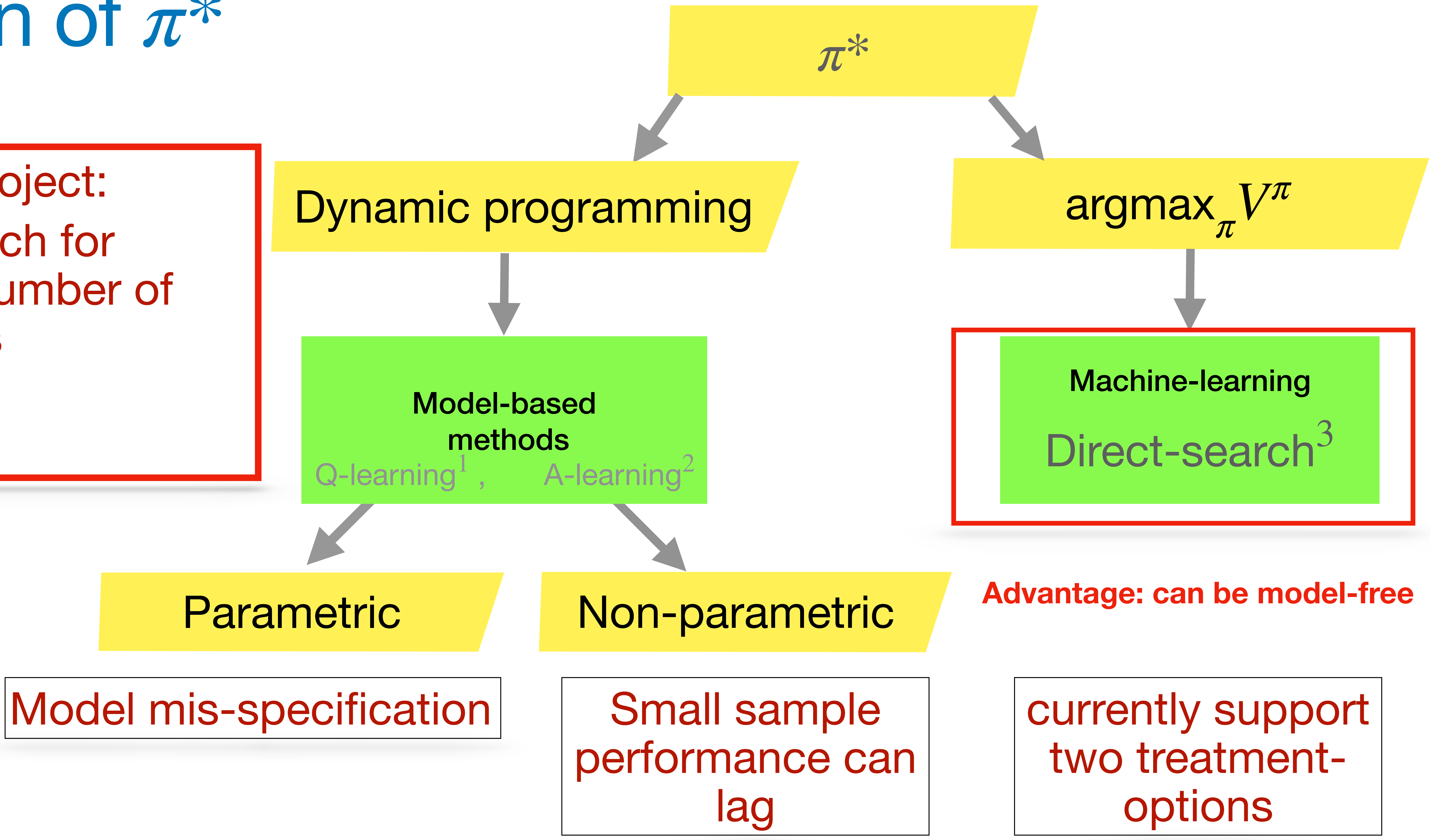



1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2022

# Estimation of $\pi^*$

Goal of the project:

1. direct search for arbitrary number of treatments

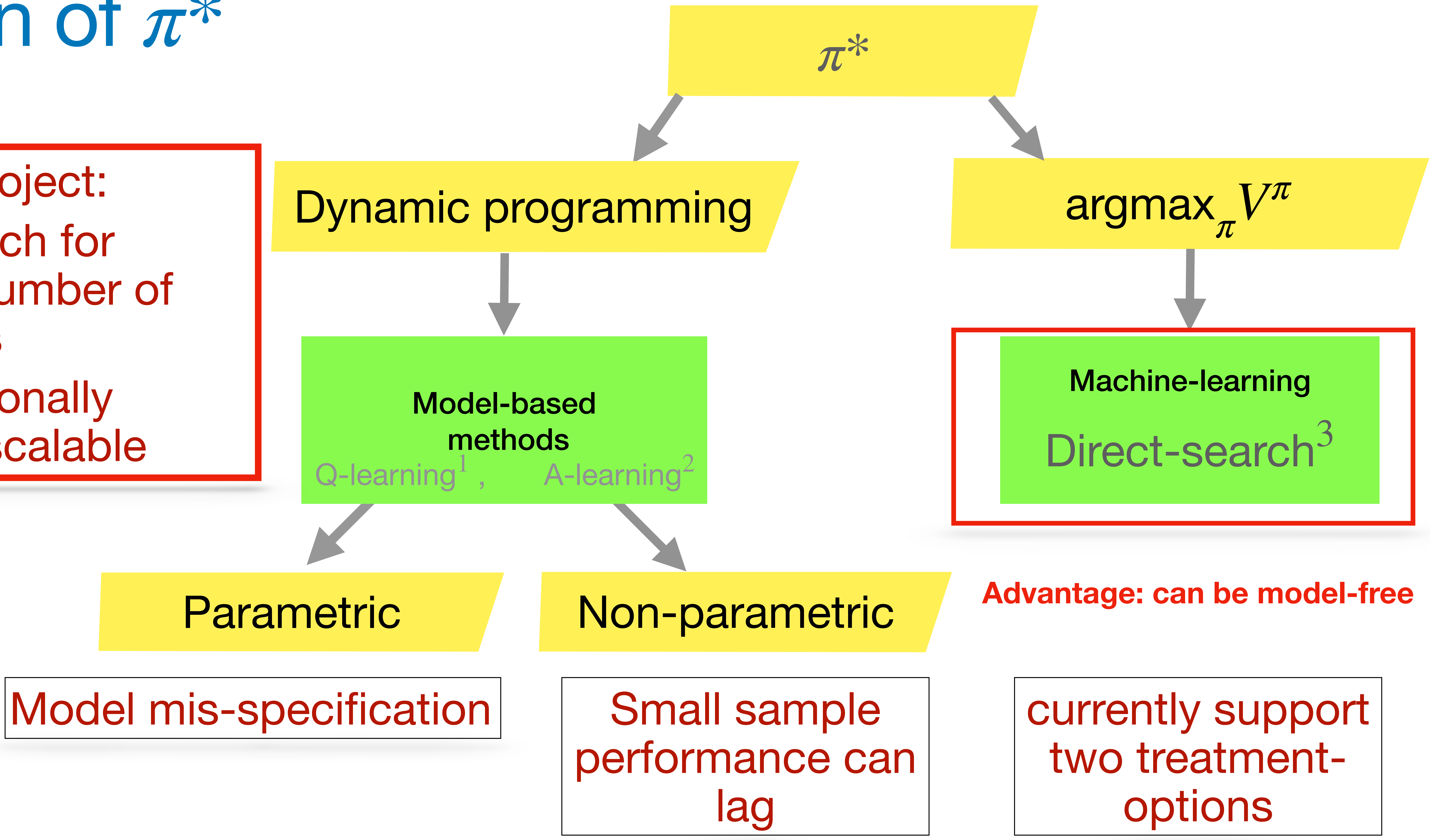



1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2022

# Estimation of $\pi^*$

Goal of the project:

1. direct search for arbitrary number of treatments
2. Computationally efficient and scalable




1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2022

# Outline

- Example: sepsis
- Problem formulation
- Proposed method

A. Methodology

B. Example on a toy data

- Open questions

# Outline

- Example: sepsis
- Problem formulation
- Proposed method

A. Methodology

B. Example on a toy data

- Open questions

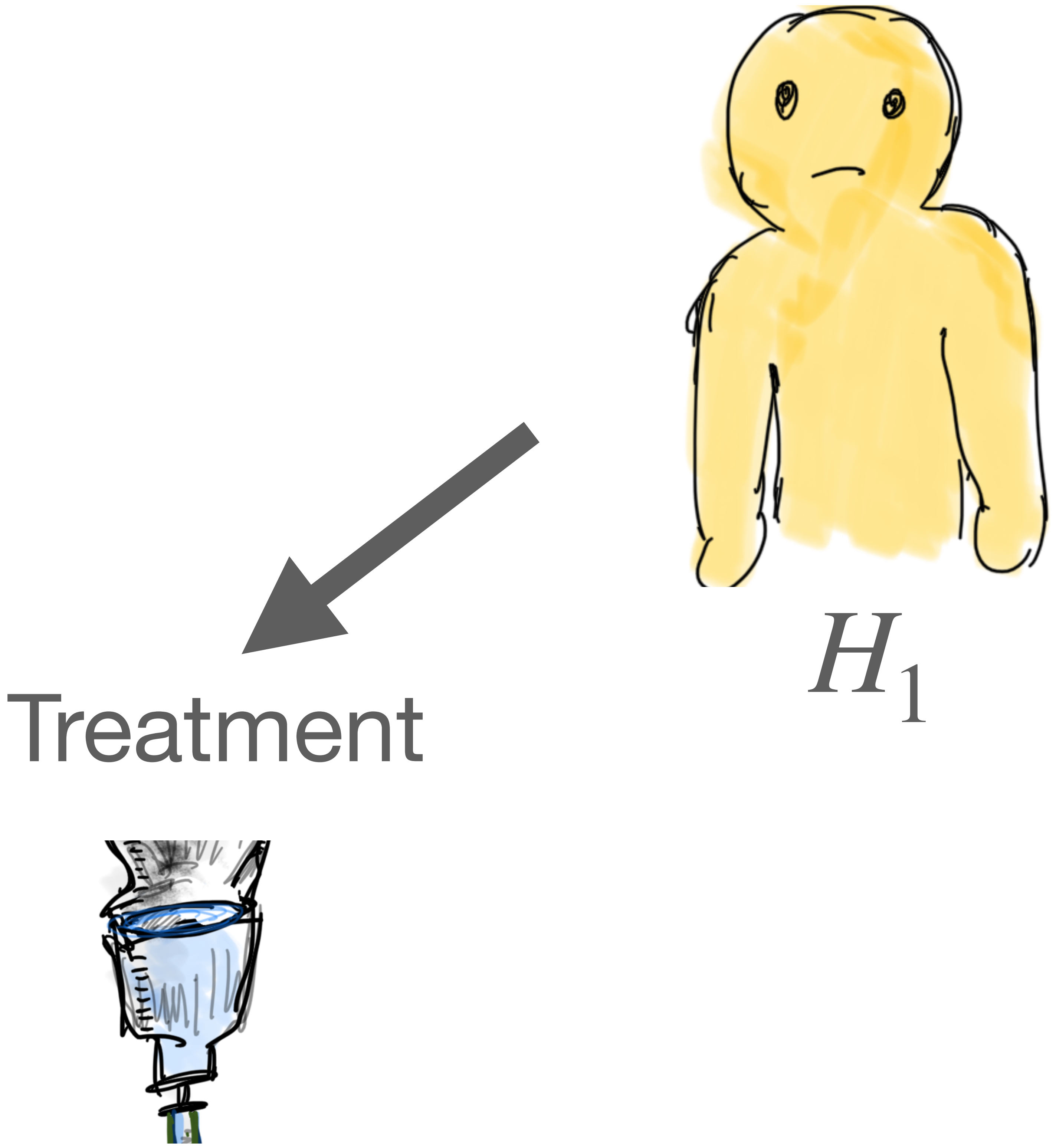
# Proposed method

# Only two treatment options

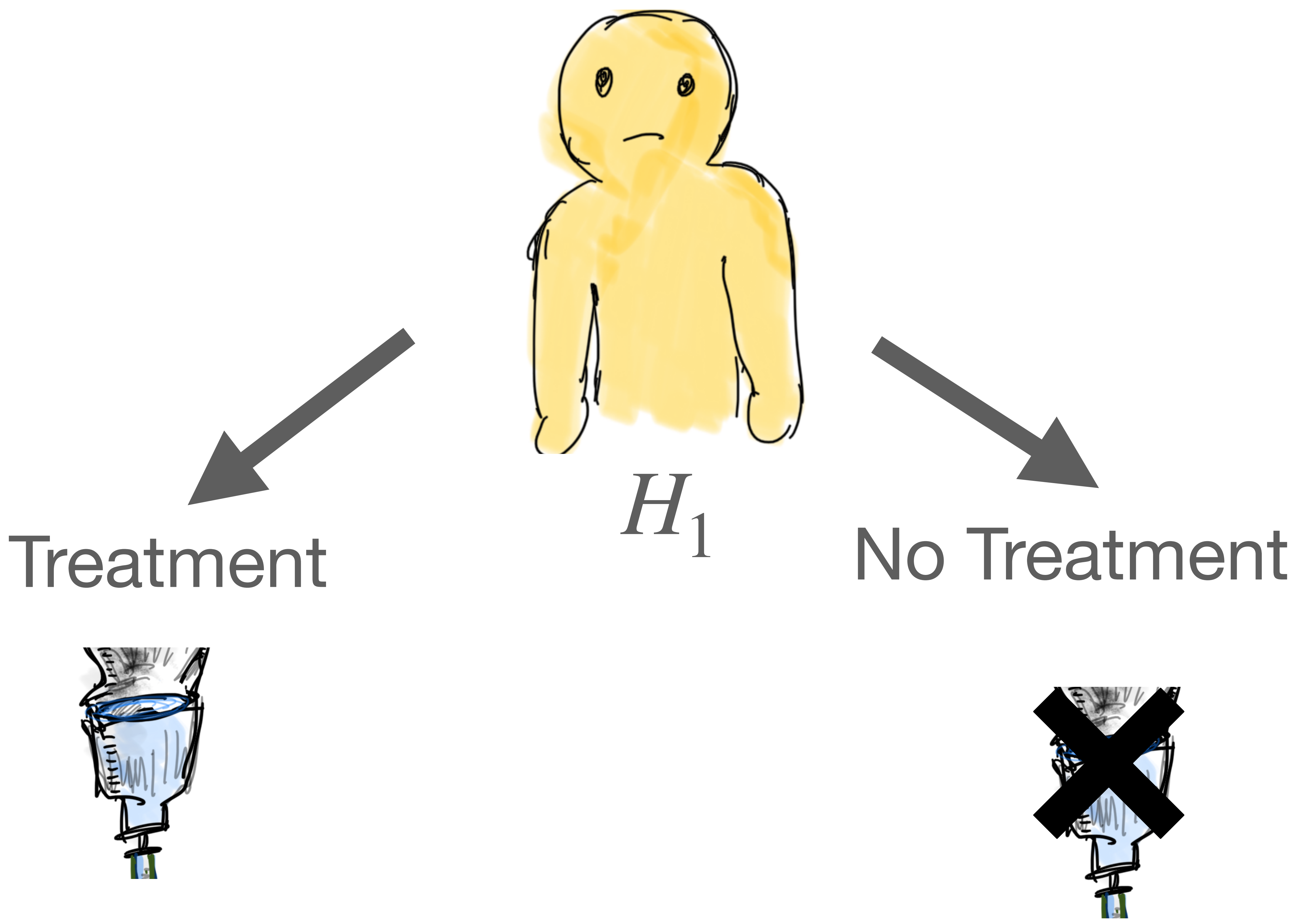
# Only two treatment options

$H_1$

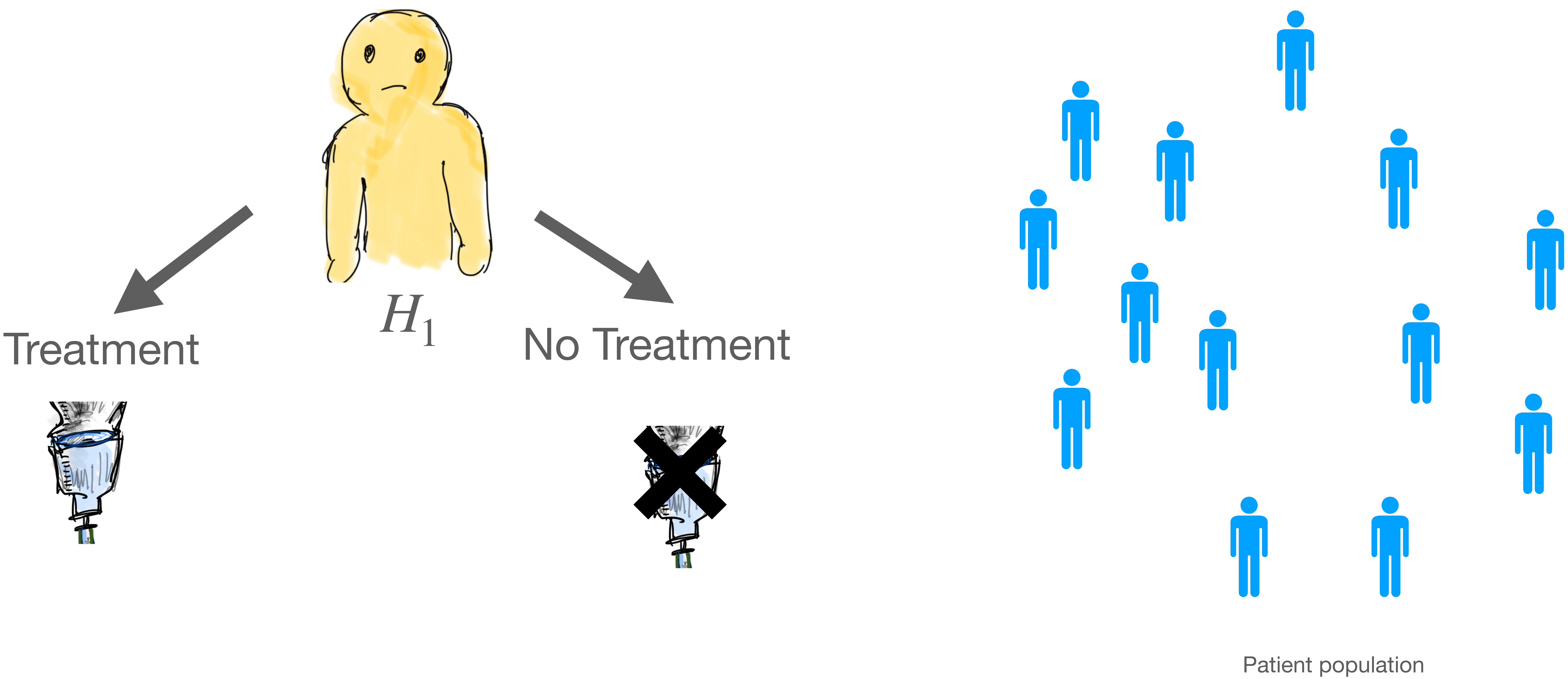
# Only two treatment options



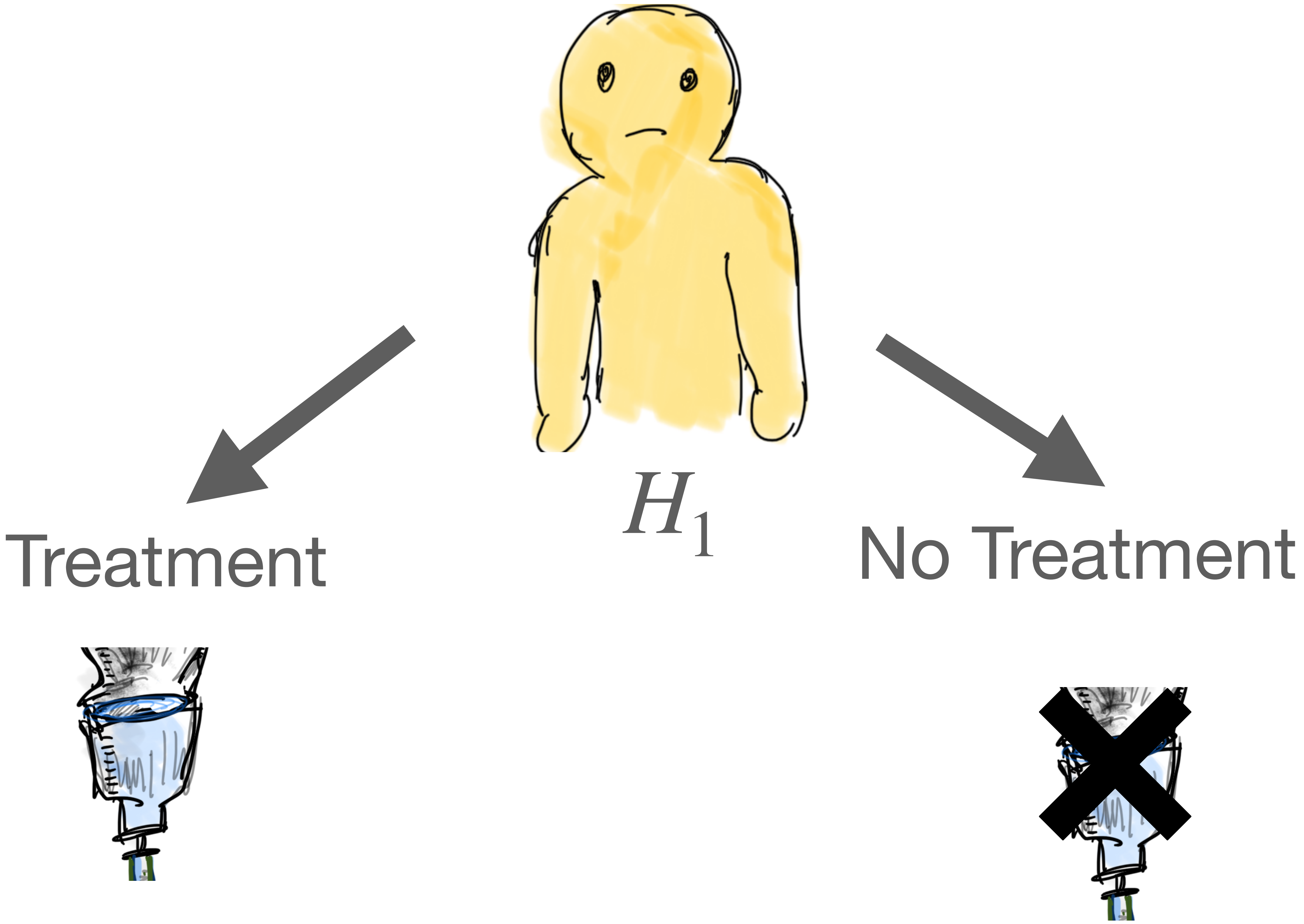
# Only two treatment options

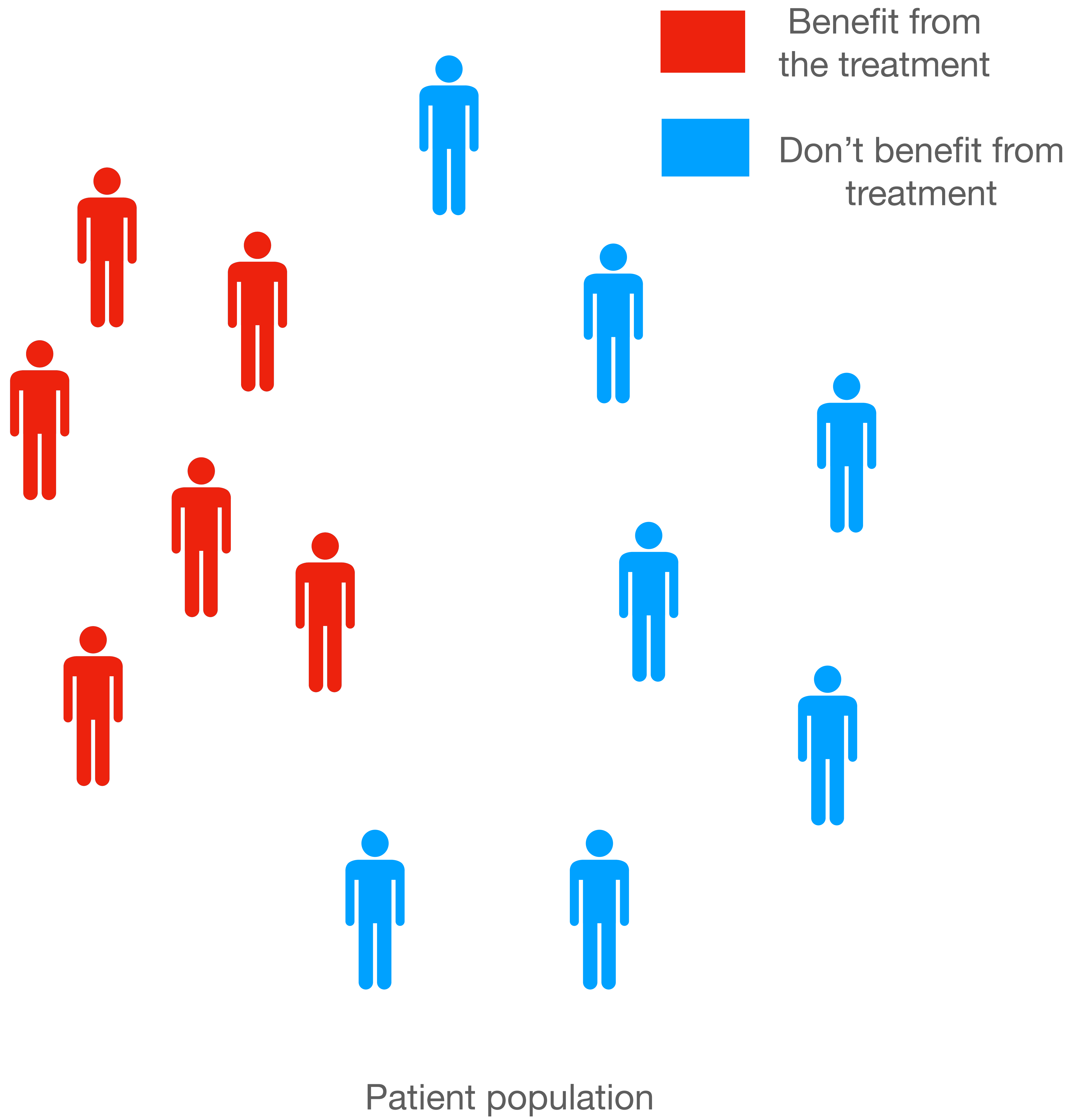


# Only two treatment options



# Only two treatment options



# Only two treatment options

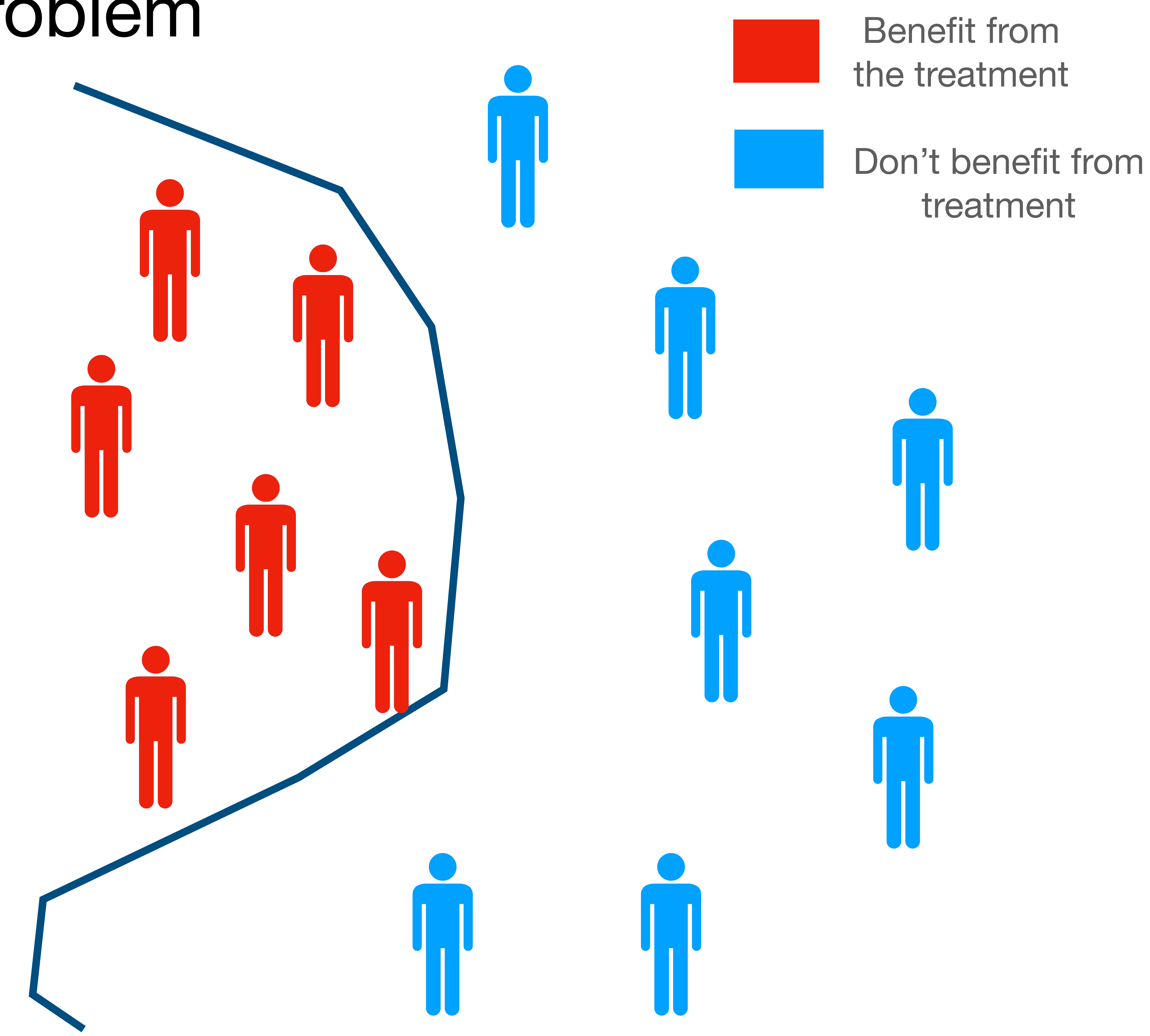
Treatment assignment at each stage:

binary classification problem

$H_1$

Treatment

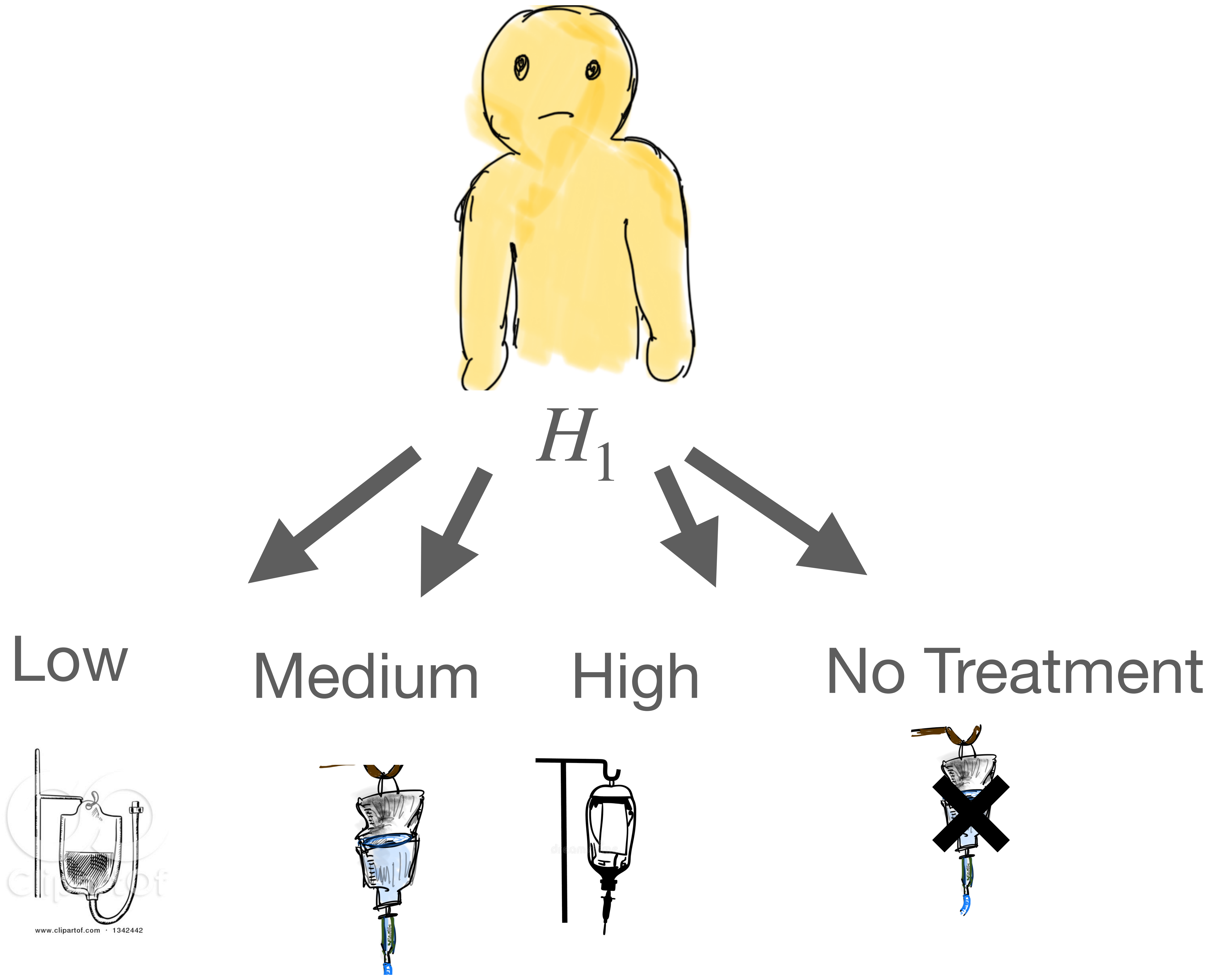
No Treatment



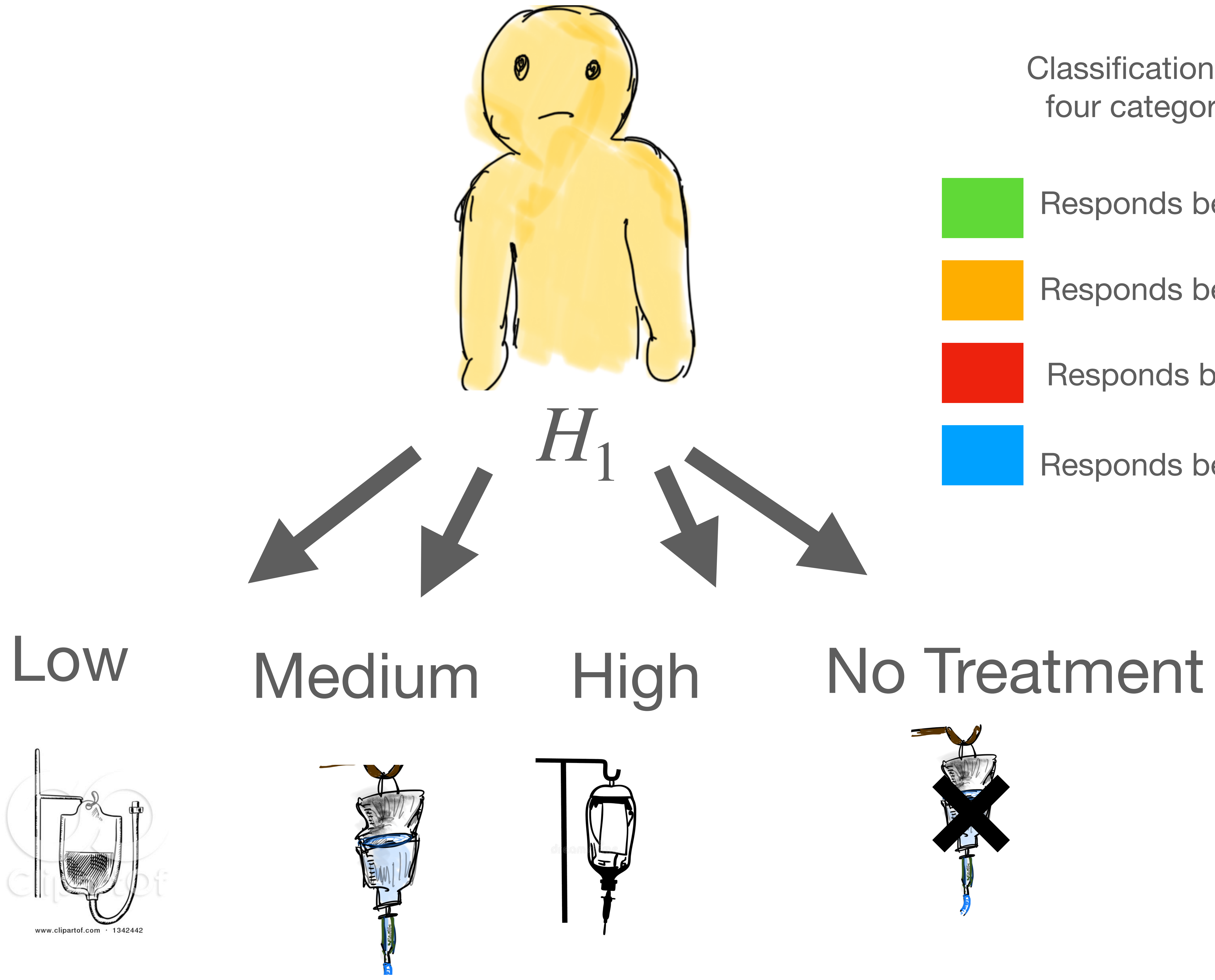

Patient population

# More than two treatment option

# More than two treatment option

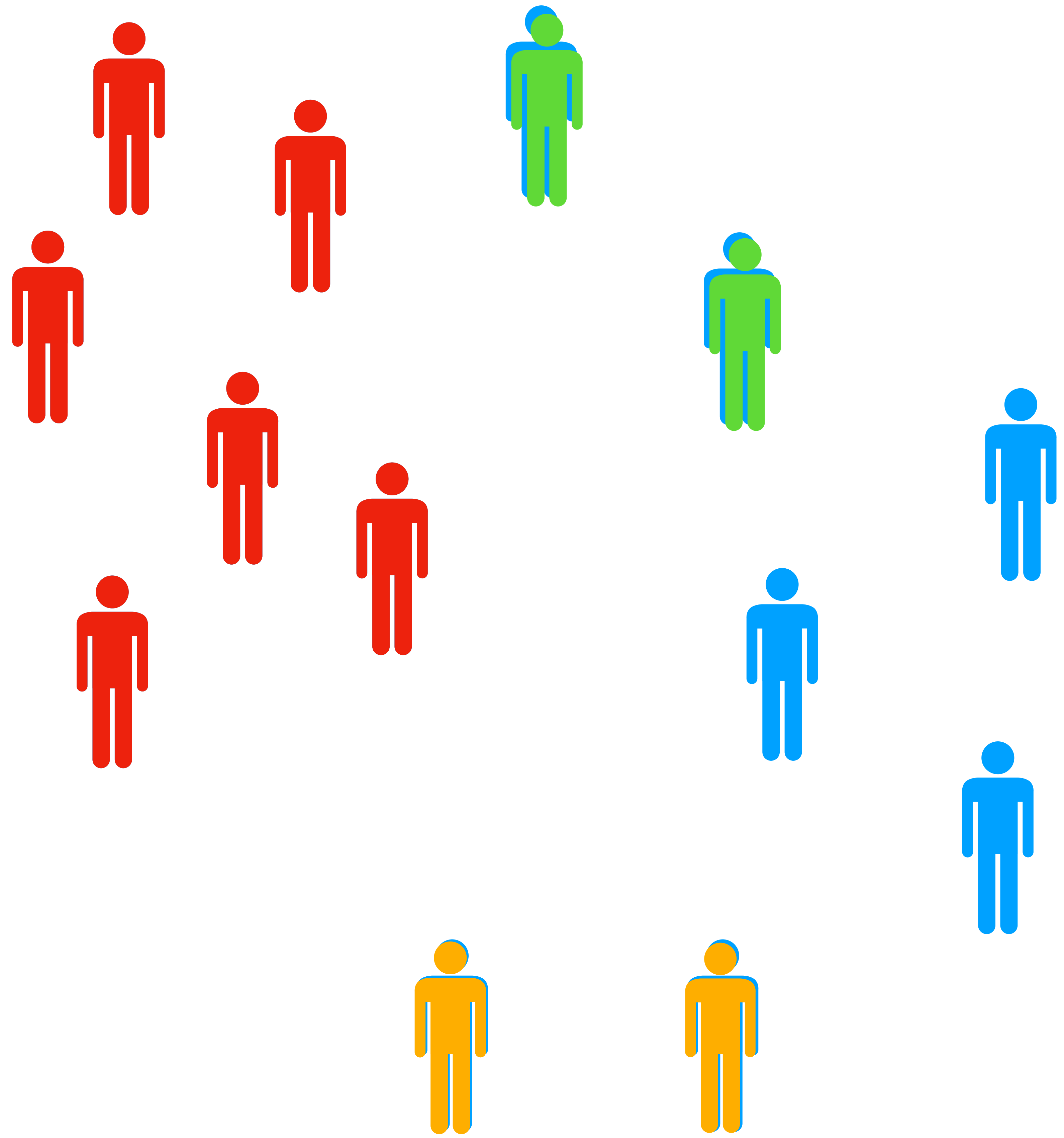


# More than two treatment option




Classification into
four categories

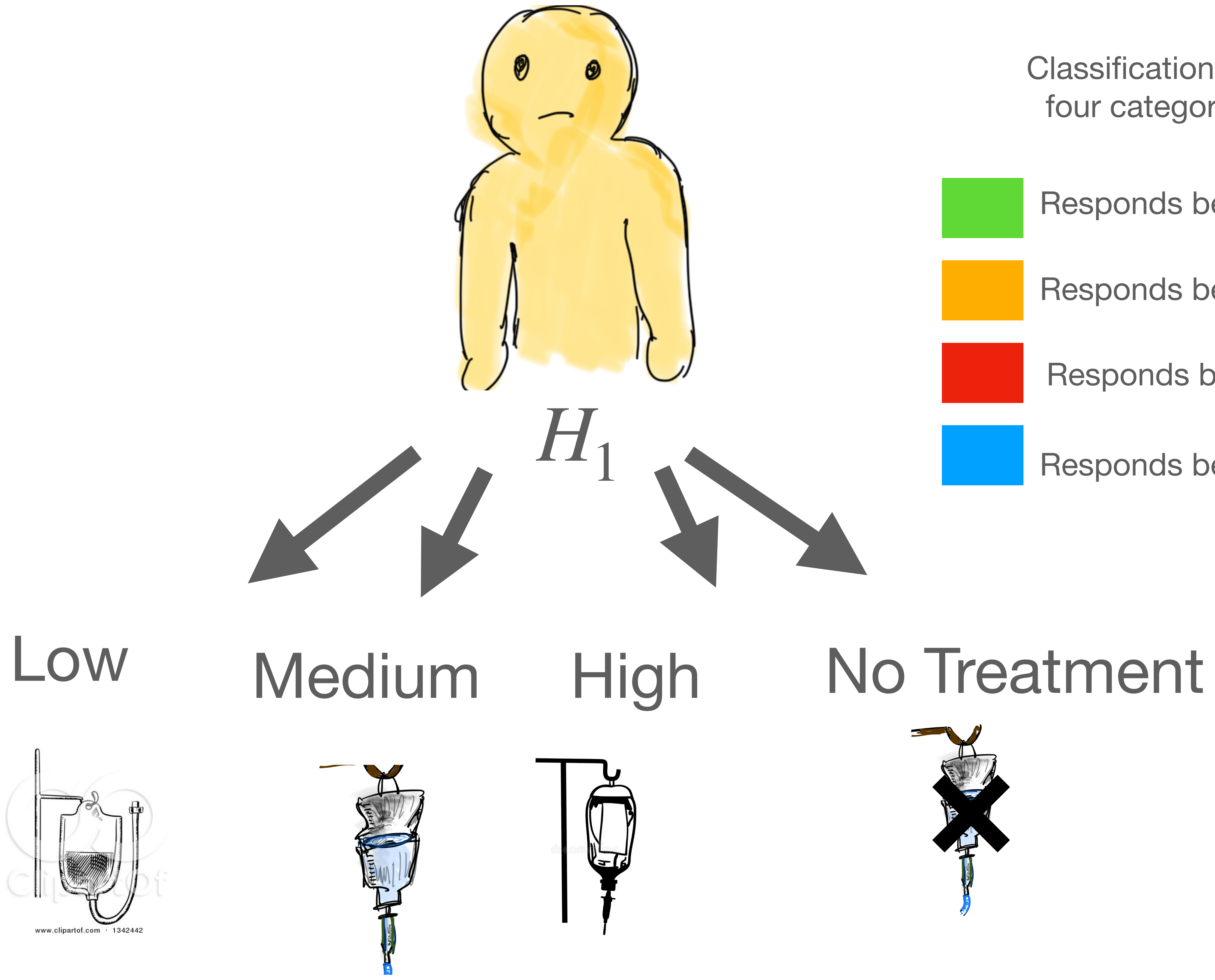
- Responds best to low
- Responds best to medium
- Responds best to high
- Responds best tono trt



Patient population

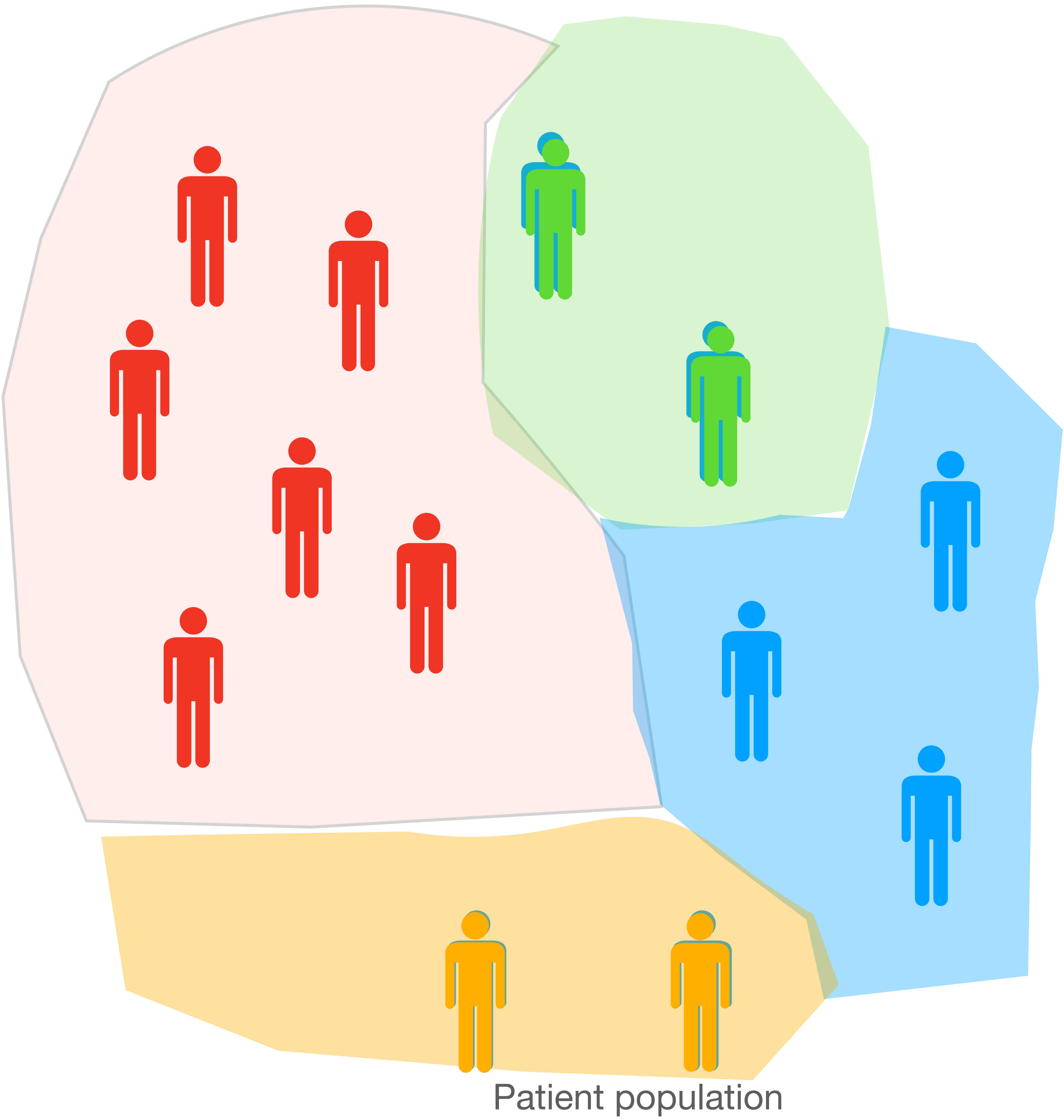
# More than two treatment option

Treatment assignment at each stage: multicategory classification problem



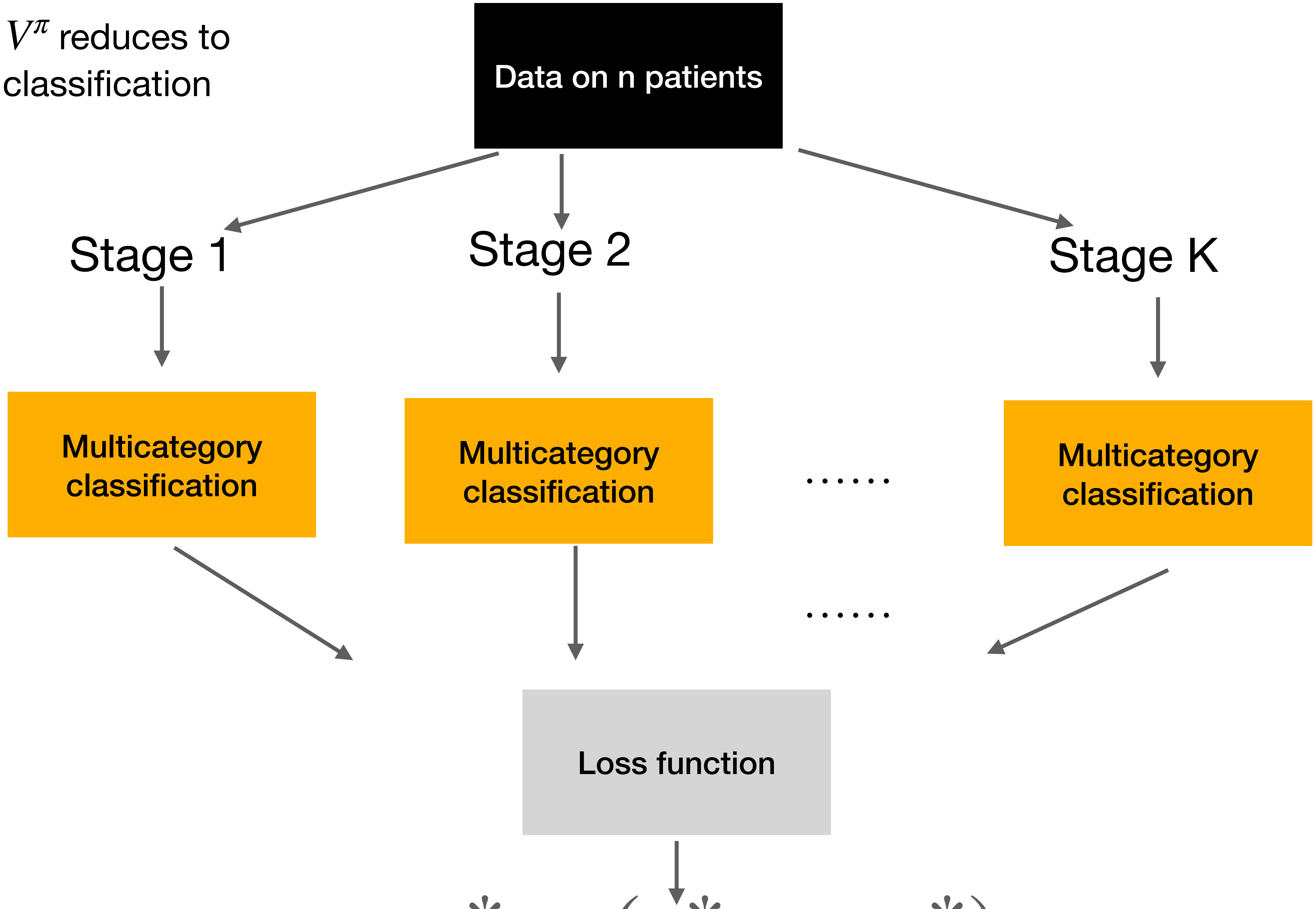

Classification into
four categories

- Responds best to low
- Responds best to medium
- Responds best to high
- Responds best tono trt



# Proposed method

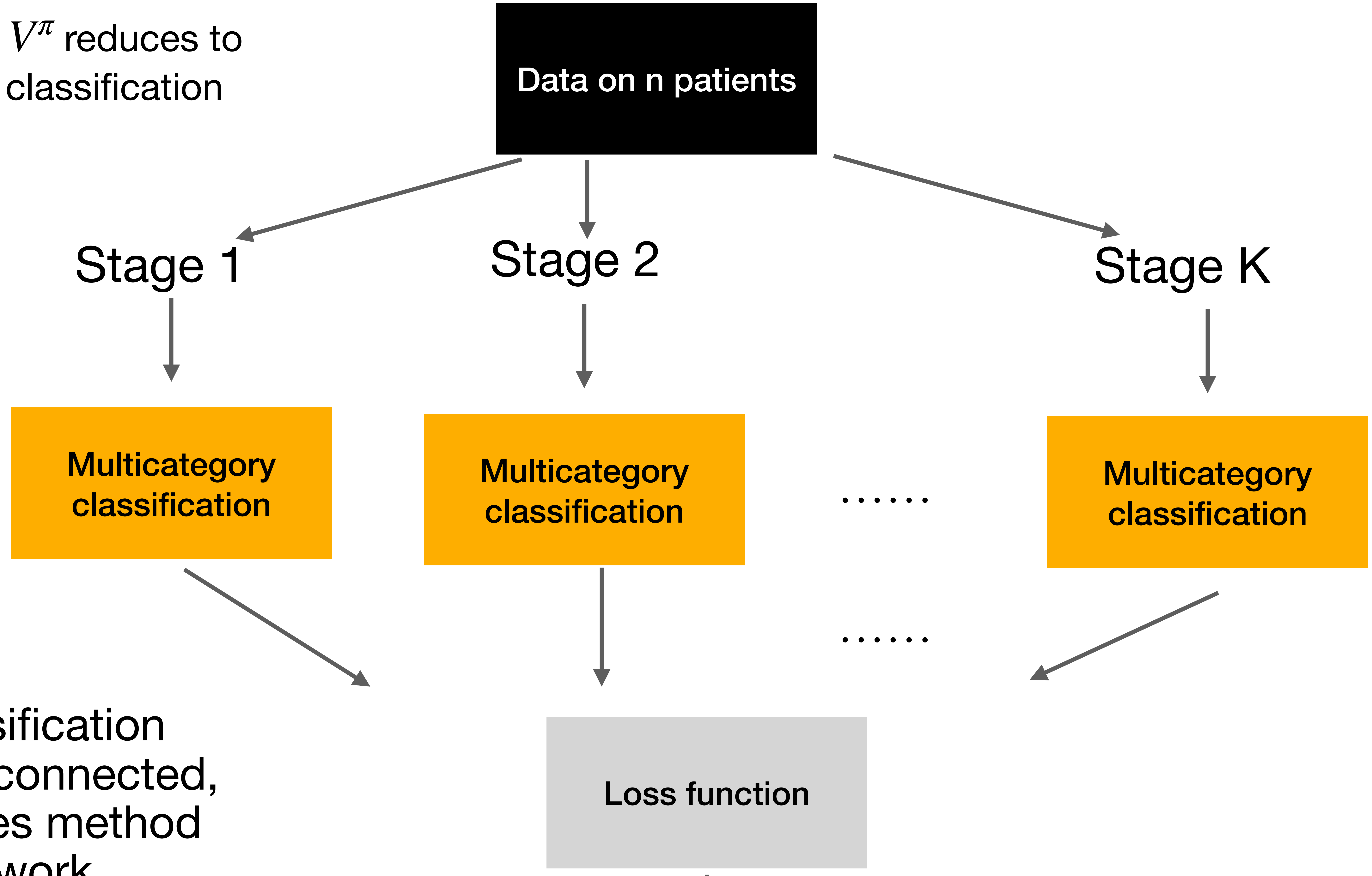
Maximization of $V^{\pi}$ reduces to simultaneous $K$ classification problems




$$\pi^* = (\pi_1^*, \ldots, \pi_K^*)$$

# Proposed method

Maximization of $V^{\pi}$ reduces to simultaneous $K$ classification problems

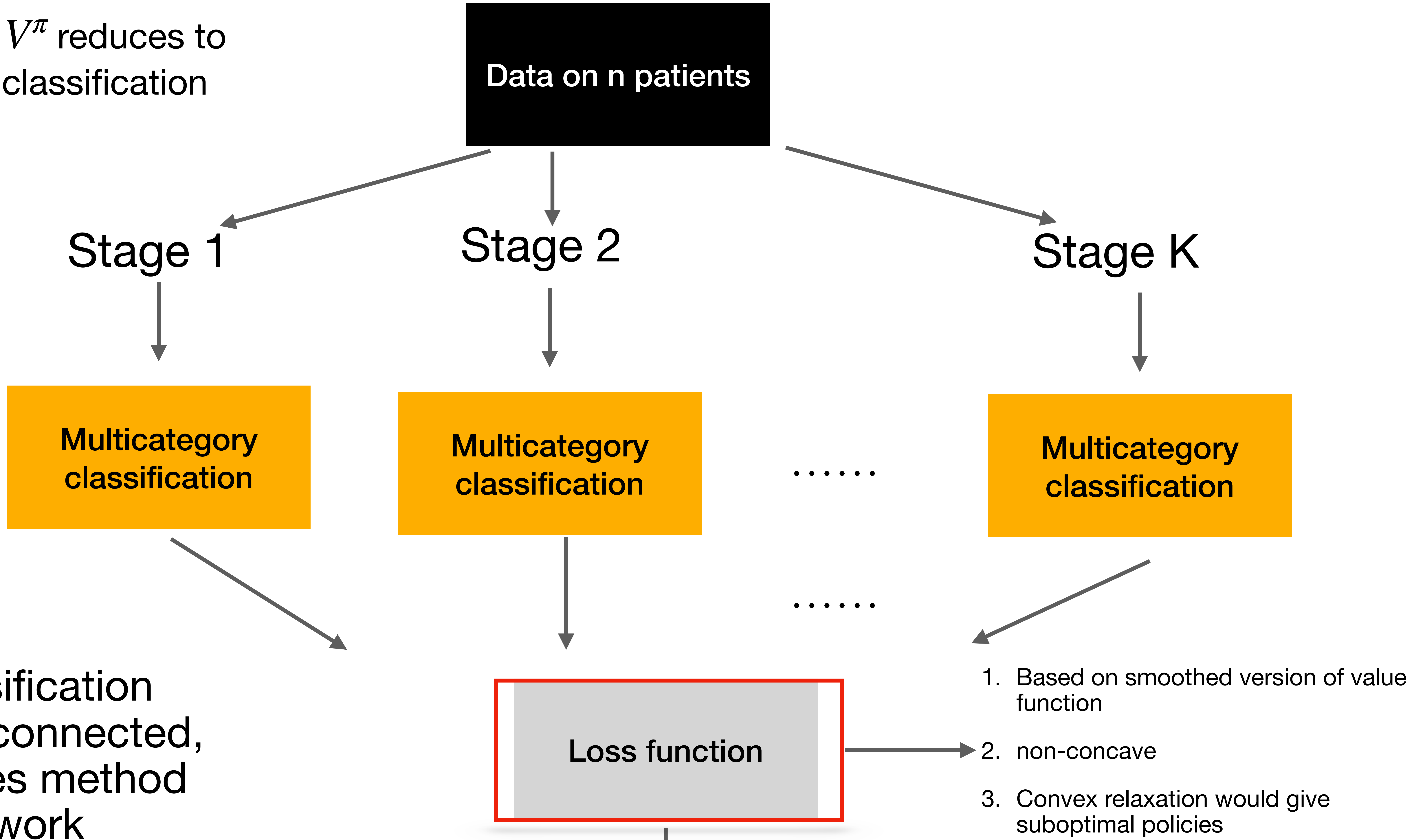



$$\pi^* = (\pi_1^*, \ldots, \pi_K^*)$$

The $K$ classification problems are connected, off-the-shelves method will not work

# Proposed method

Maximization of $V^{\pi}$ reduces to simultaneous $K$ classification problems




The $K$ classification problems are connected, off-the-shelves method will not work

$$\pi^* = (\pi_1^*, \ldots, \pi_K^*)$$

# Proposed method

Maximization of $V^{\pi}$ reduces to simultaneous $K$ classification problems

Data on n patients

Stage 1 → Multicategory classification

Stage 2 → Multicategory classification

……

Stage K → Multicategory classification

The $K$ classification problems are connected, off-the-shelves method will not work

Loss function

1. Based on smoothed version of value function
2. non-concave
3. Convex relaxation would give suboptimal policies

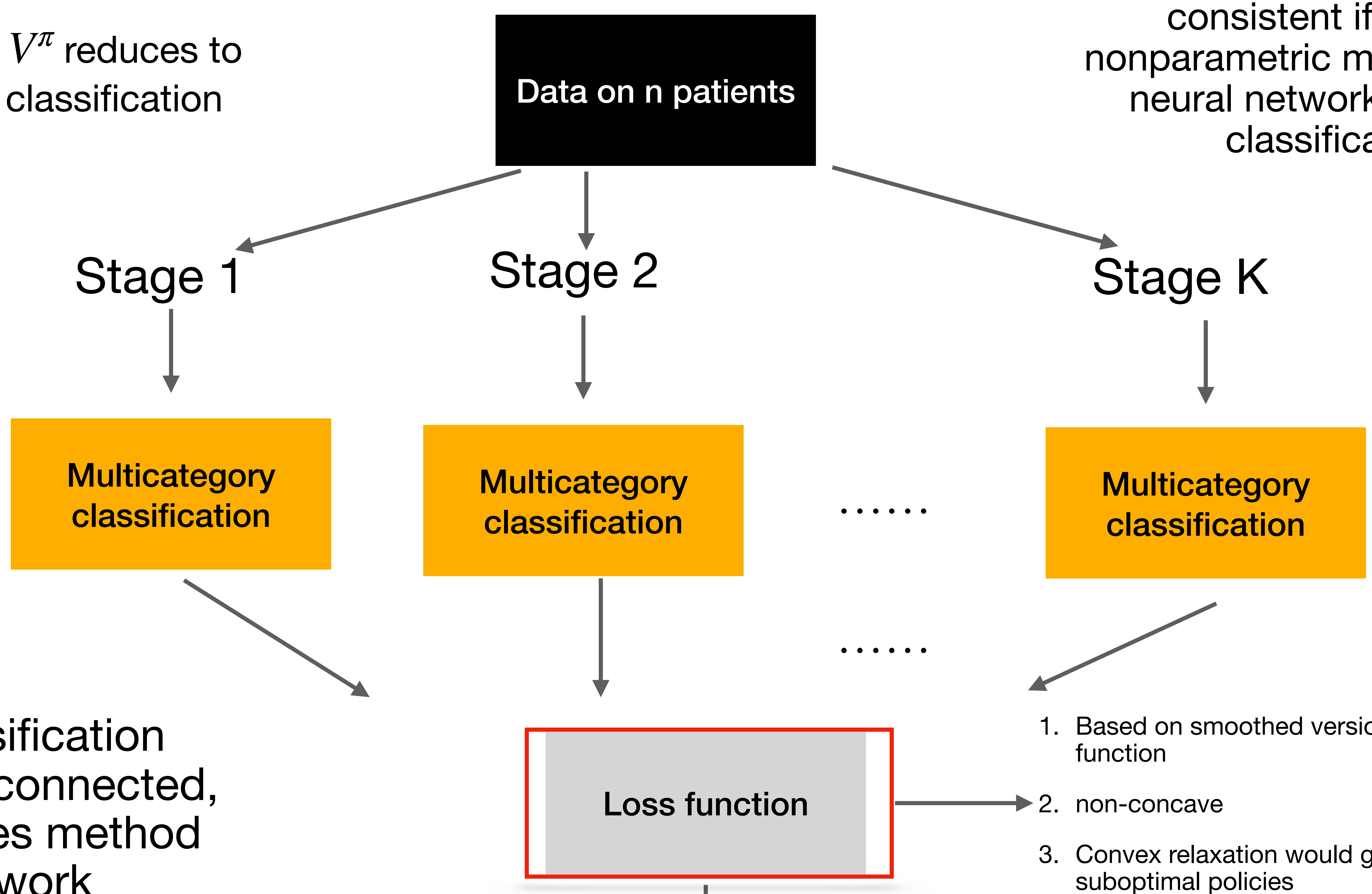
Can use stochastic gradient descent (SGD)

$$\pi^* = (\pi_1^*, \ldots, \pi_K^*)$$

# Proposed method

Maximization of $V^\pi$ reduces to simultaneous $K$ classification problems

Estimated policy will be consistent if we use nonparametric methods, e.g., neural networks, for the classification




The $K$ classification problems are connected, off-the-shelves method will not work

1. Based on smoothed version of value function
2. non-concave
3. Convex relaxation would give suboptimal policies

Can use stochastic gradient descent (SGD)

Population level solution $\pi^* = (\pi_1^*, \ldots, \pi_K^*)$

# Outline

- Example: sepsis
- Problem formulation
- Proposed method

A. Methodology

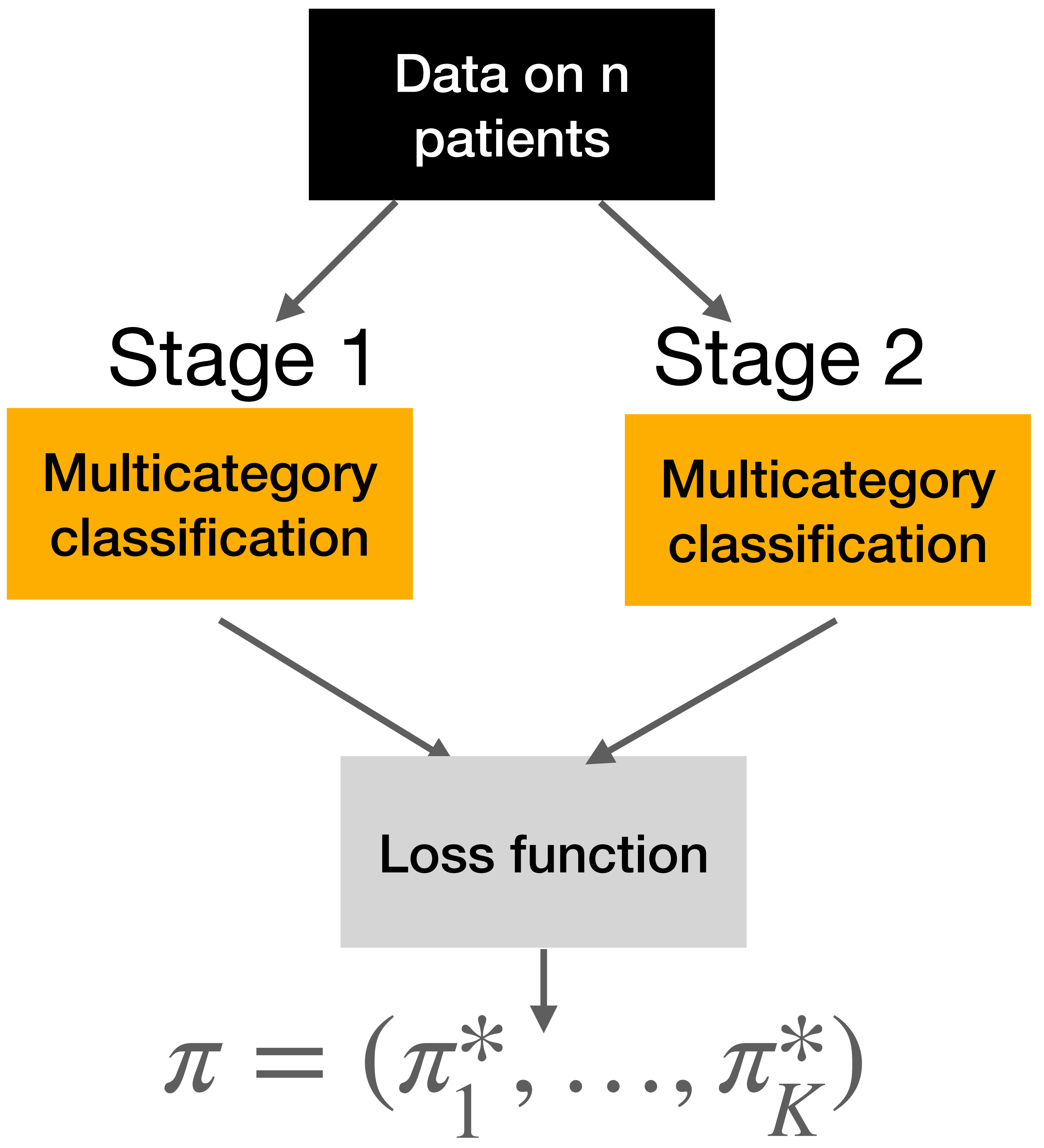
B. Example on a toy data

- Open questions

# Example with toy data

*This work is by Sneha Mishra, my former summer RA

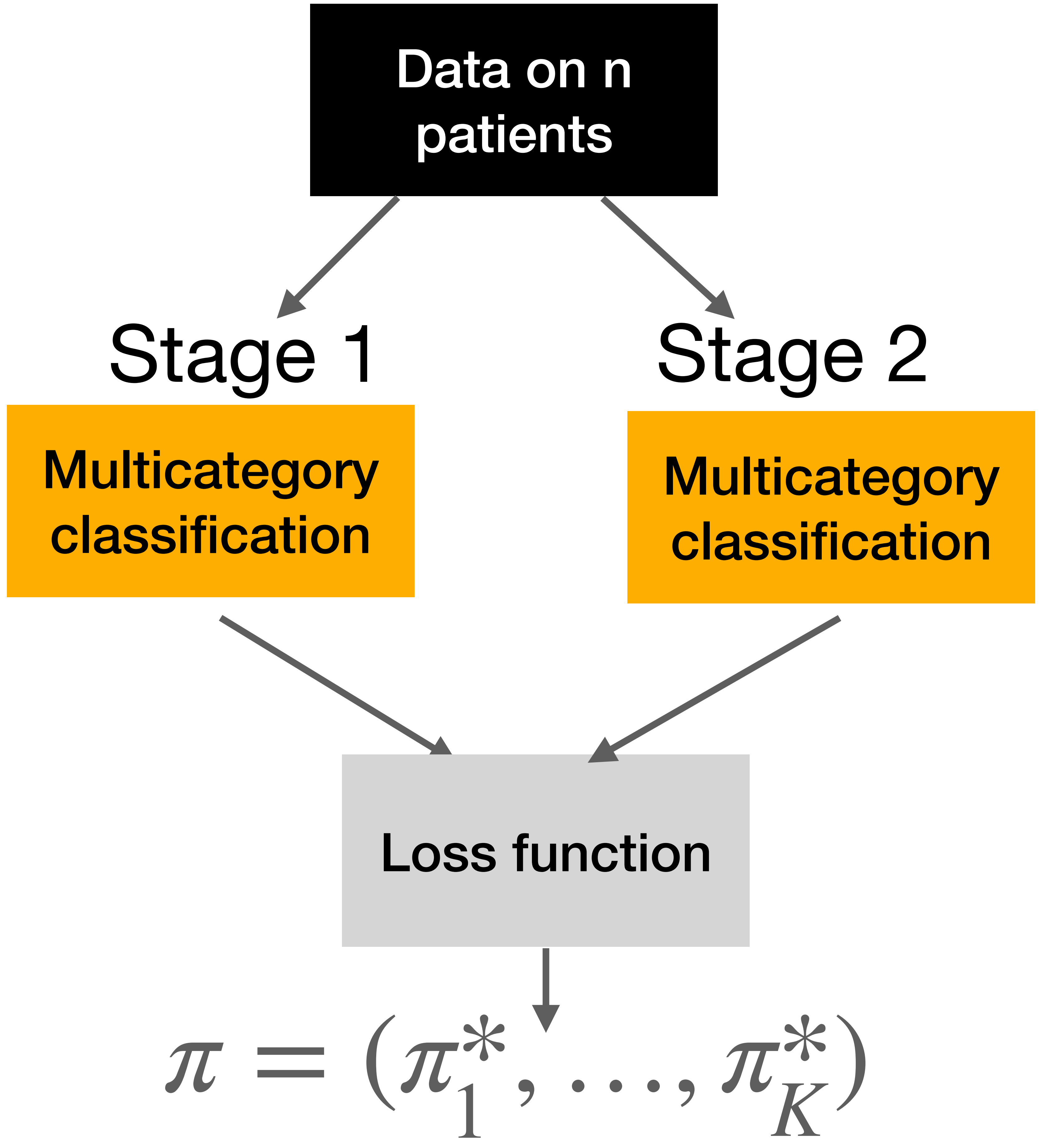
# One example



Data on n patients
Stage 1
Stage 2
Multicategory classification
Multicategory classification
Loss function
$\pi = (\pi_1^*, \ldots, \pi_K^*)$

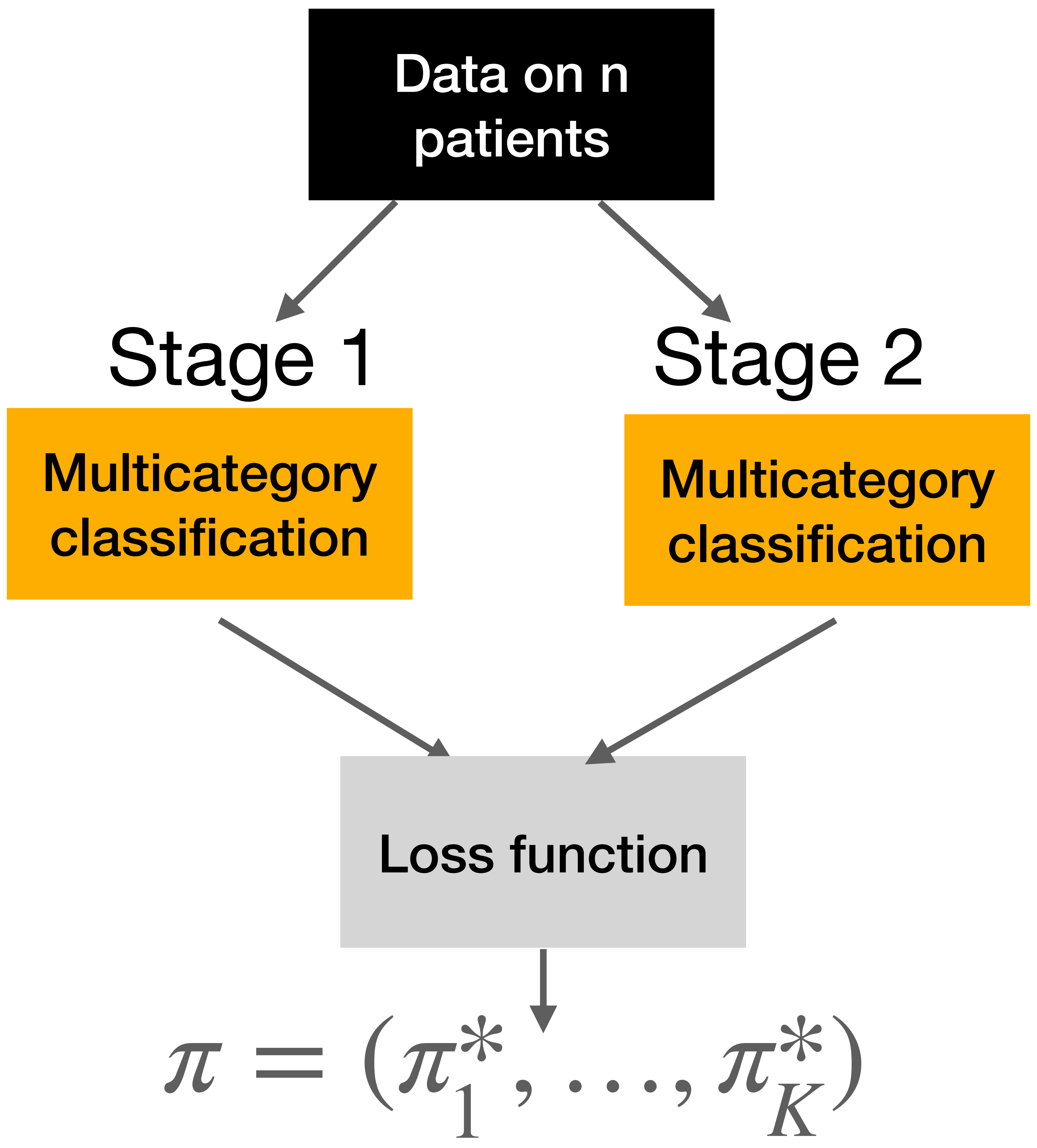
# One example

- Suppose number of stages, i.e., $K = 2$



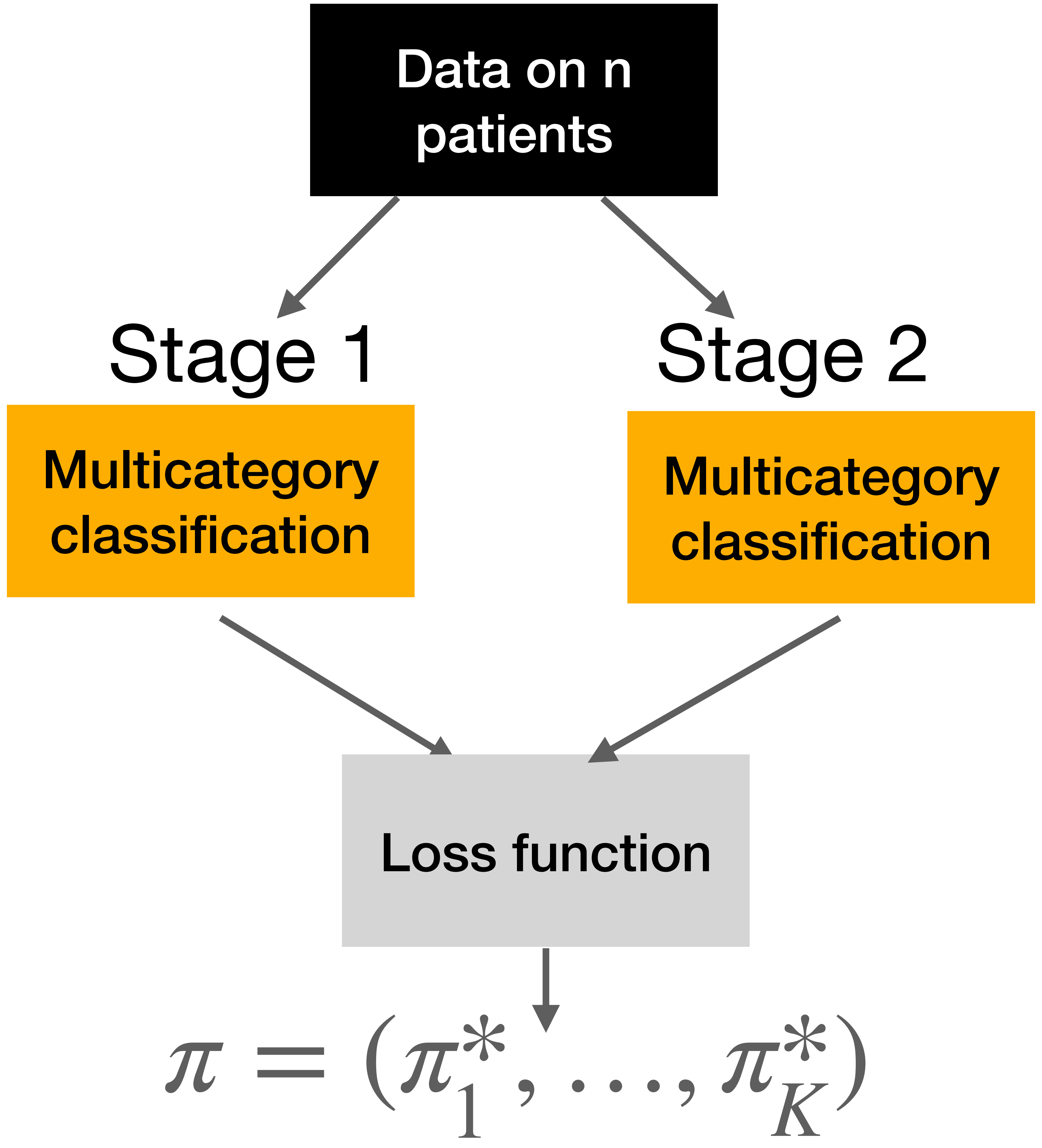
# One example

- Suppose number of stages, i.e., $K = 2$
- Number of treatments at each stage: 3.



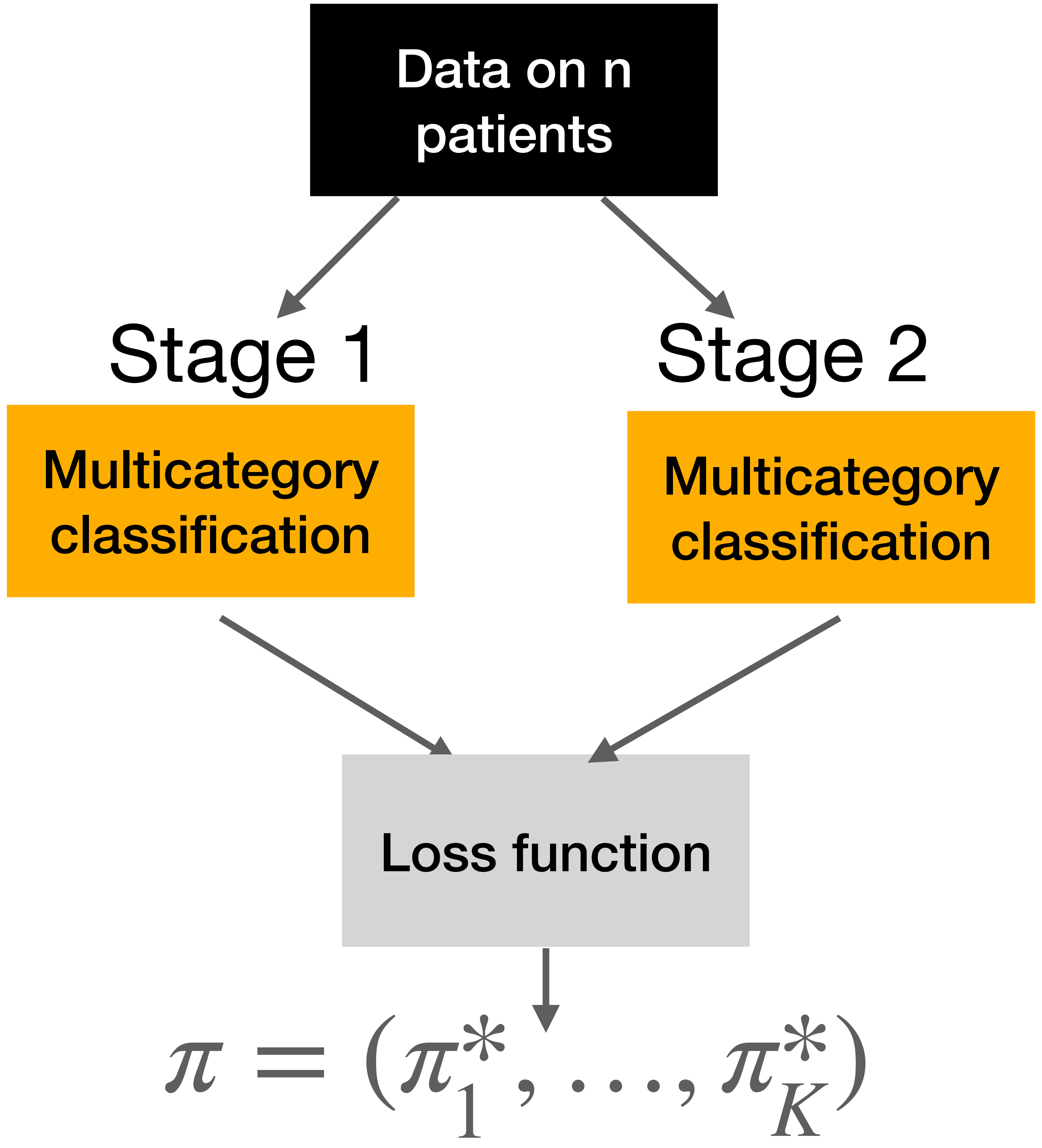
# One example

- Suppose number of stages, i.e., $K = 2$
- Number of treatments at each stage: 3.
- Use neural network classifiers



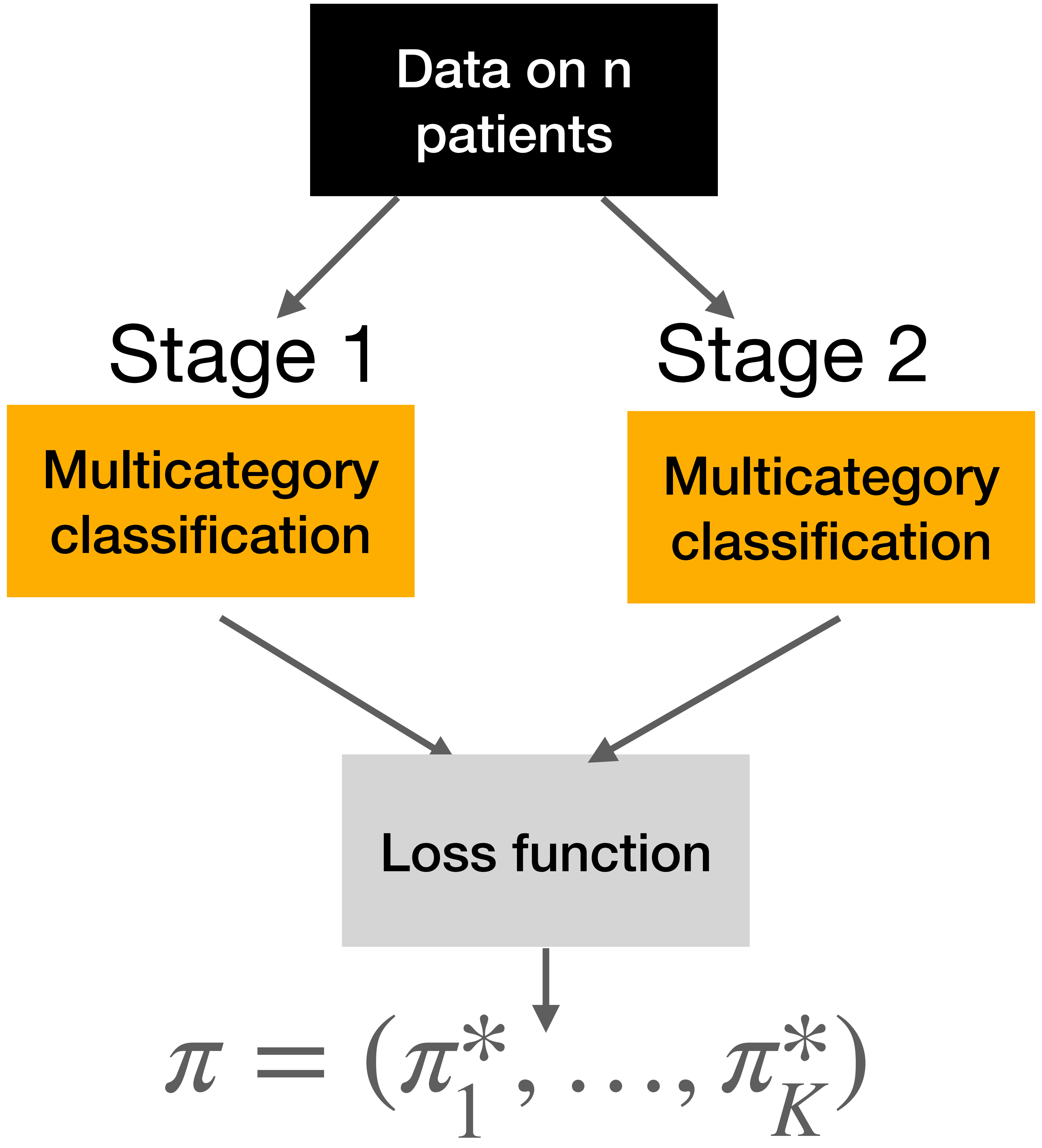
# One example

- Suppose number of stages, i.e., $K = 2$
- Number of treatments at each stage: 3.
- Use neural network classifiers
- No. Of covariates: 3

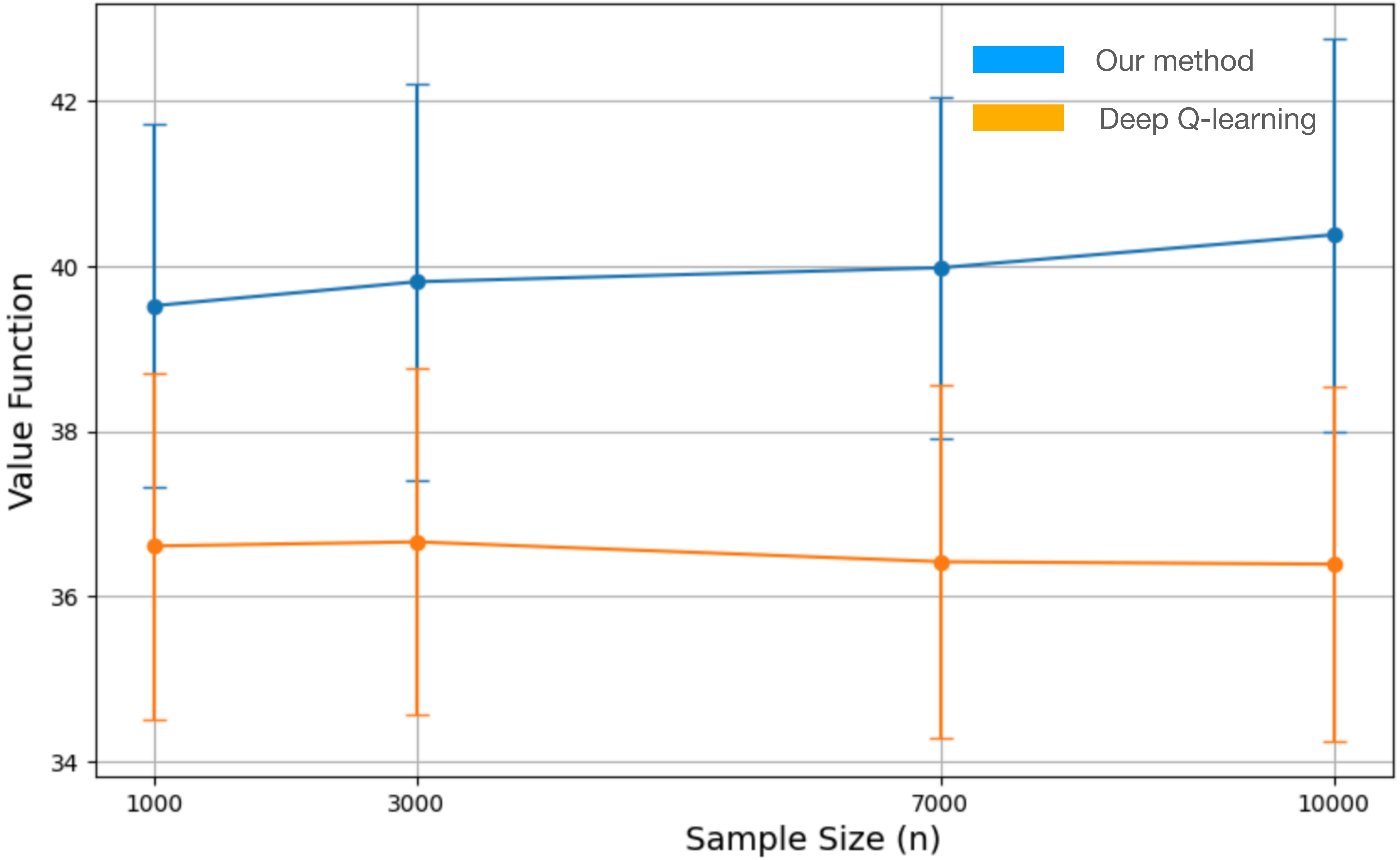


# One example

- Suppose number of stages, i.e., $K = 2$
- Number of treatments at each stage: 3.
- Use neural network classifiers
- No. Of covariates: 3
- The covariates and rewards were Gaussian, and the rewards were generated by a linear model.



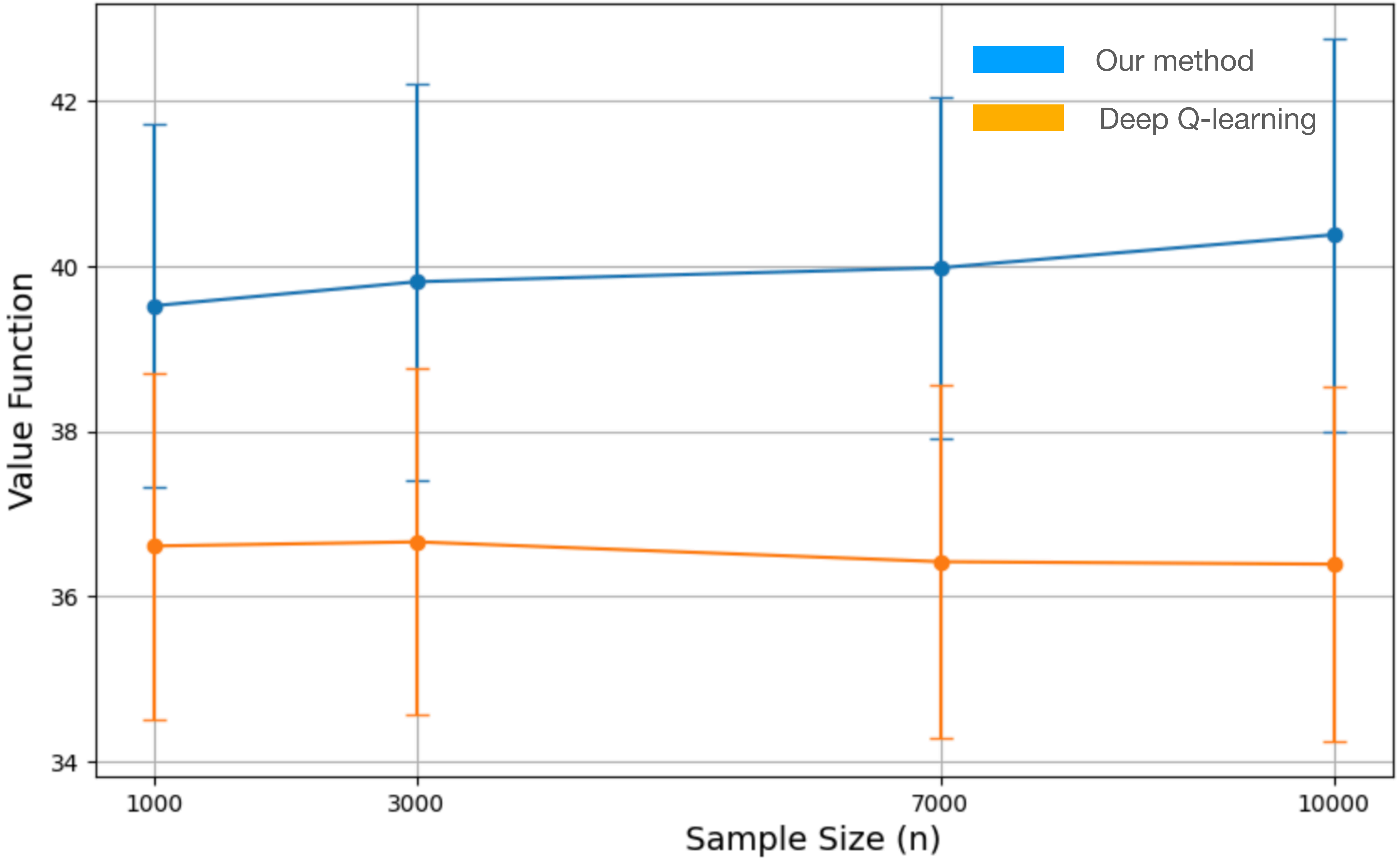
# Plot of the population-level value functions




Plot of the value function for our method and deep Q-learning based on the toy data

The deep Q-learning line represents the optimal policy generated by deep Q-learning method for DTR — that is current gold standard

# Plot of the population-level value functions




Plot of the value function for our method and deep Q-learning based on the toy data

The deep Q-learning line represents the optimal policy generated by deep Q-learning method for DTR — that is current gold standard

# Outline

- Example: sepsis
- Problem formulation
- Proposed method
- Open questions

A. Implementation and optimization

B. Regret decay rate

C. Doubly robust learning

# Outline

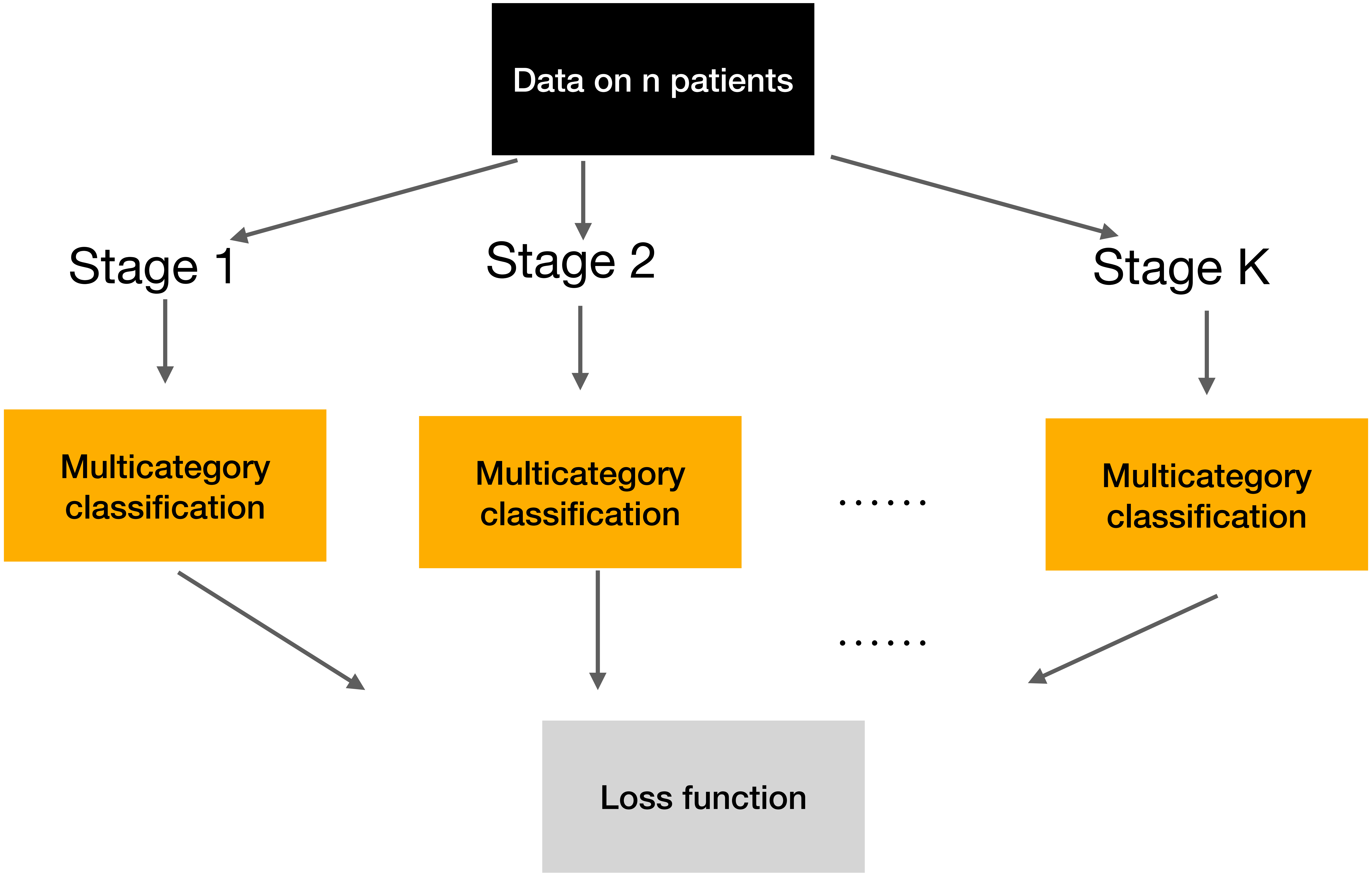
- Example: sepsis
- Problem formulation
- Proposed method
- Open questions

A. Implementation and optimization

B. Regret decay rate

C. Doubly robust learning

# Implementation

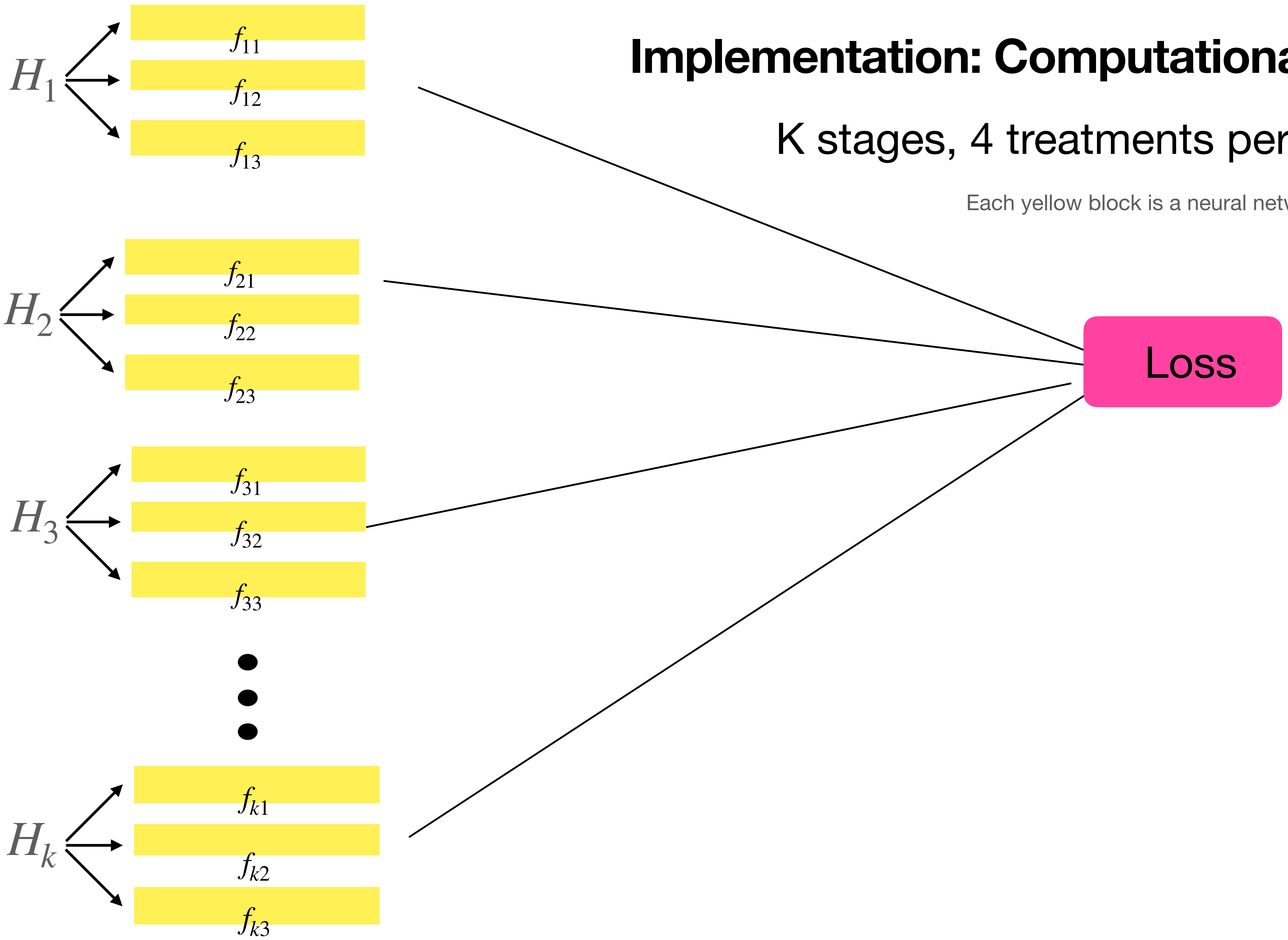



$$\pi^* = (\pi_1^*, \ldots, \pi_K^*)$$

# Implementation: Computational challenge

K stages, 4 treatments per stage

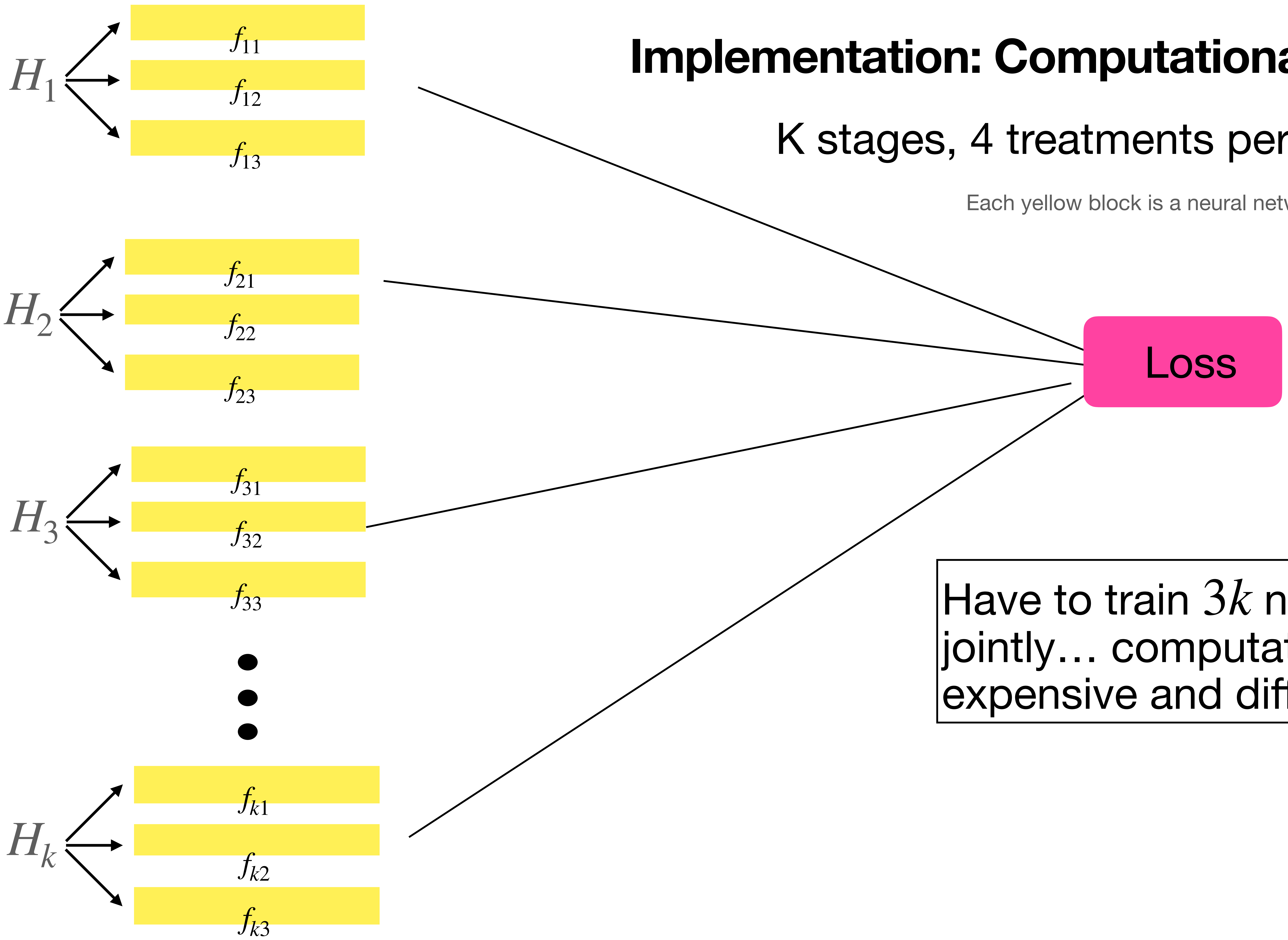
Each yellow block is a neural network



# Implementation: Computational challenge

K stages, 4 treatments per stage

Each yellow block is a neural network

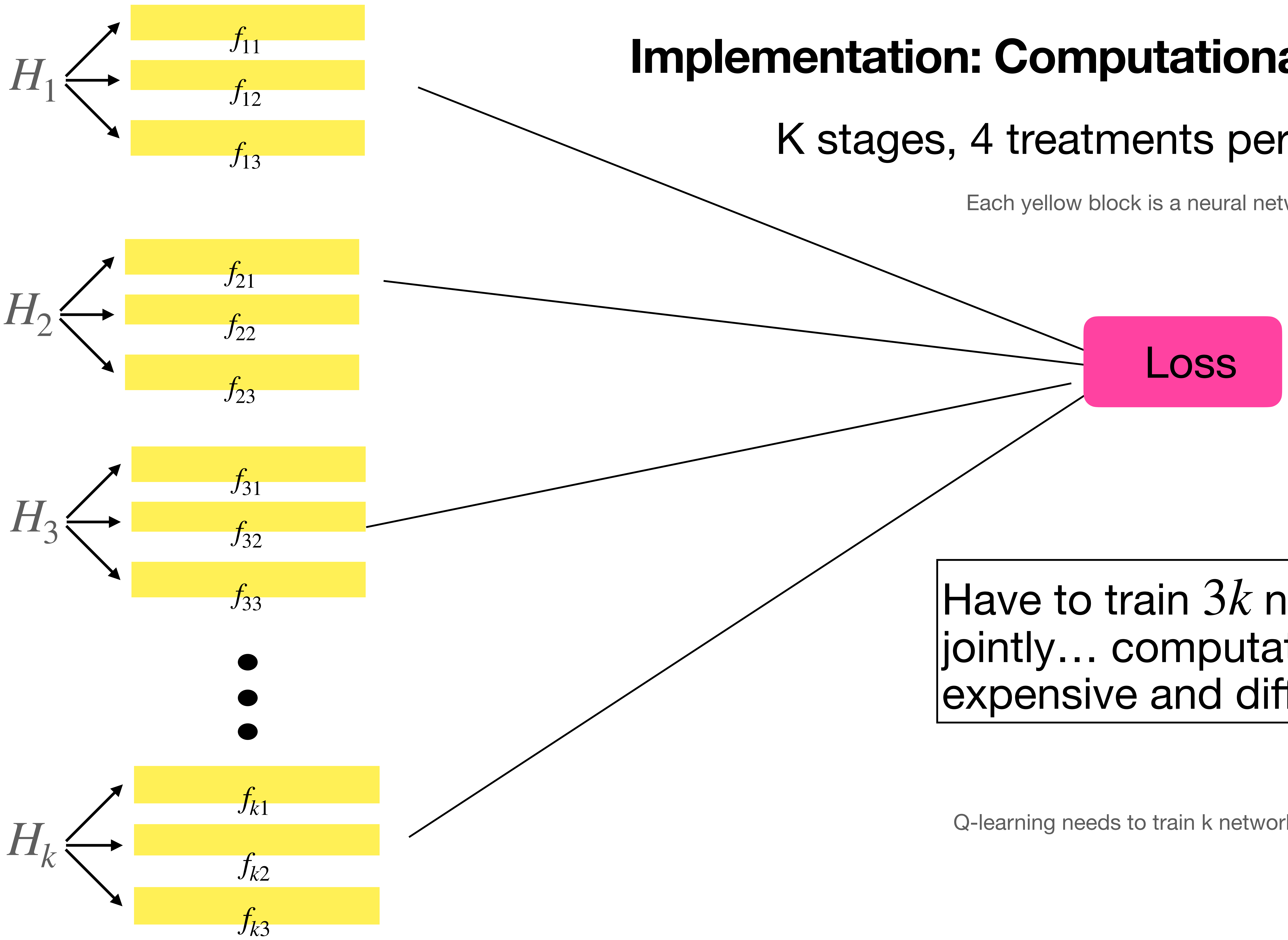



Have to train $3k$ networks jointly… computationally expensive and difficult to tune

# Implementation: Computational challenge

K stages, 4 treatments per stage

Each yellow block is a neural network



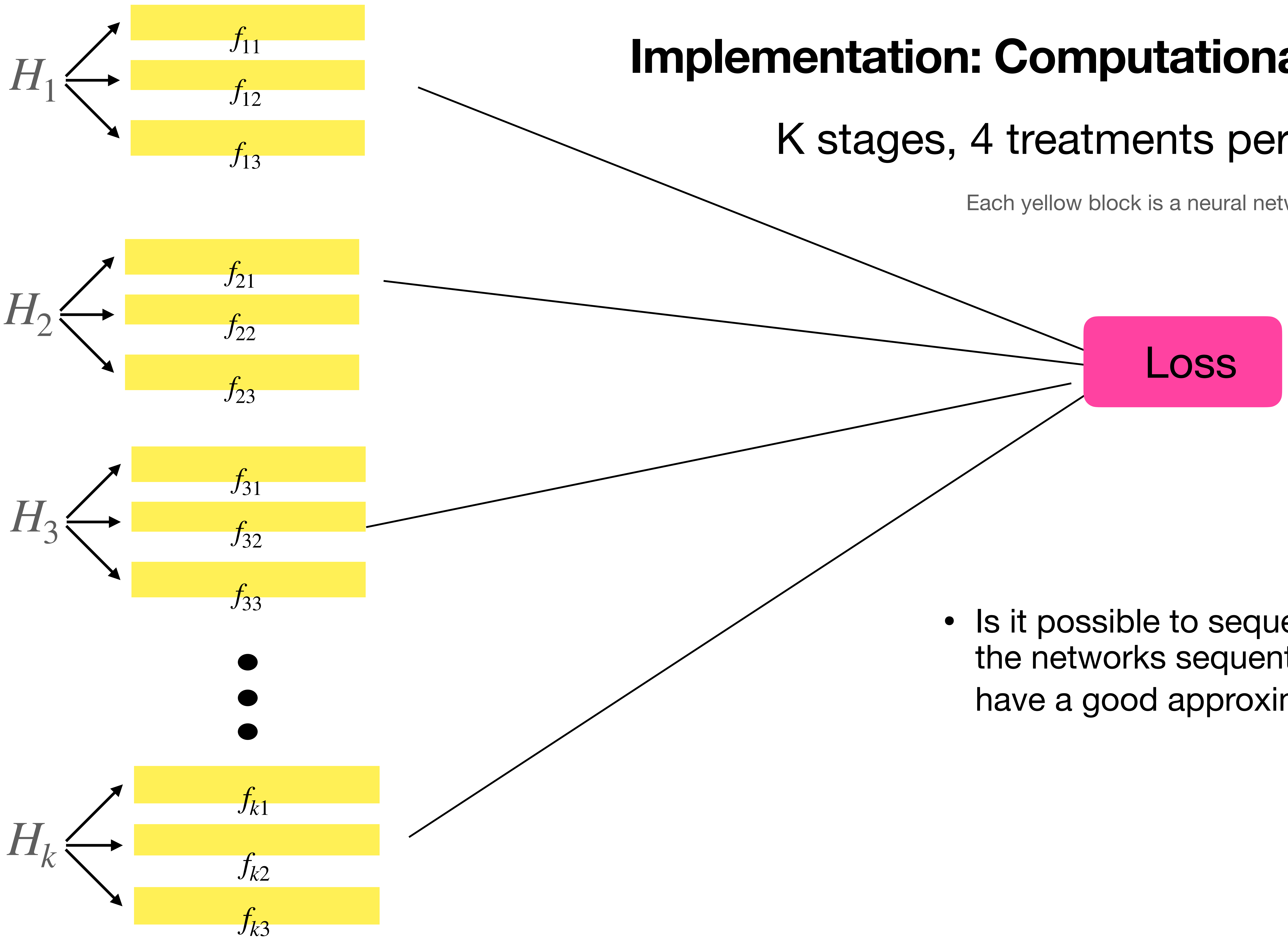

Have to train $3k$ networks jointly… computationally expensive and difficult to tune

Q-learning needs to train k networks sequentially

# Implementation: Computational challenge

K stages, 4 treatments per stage

Each yellow block is a neural network

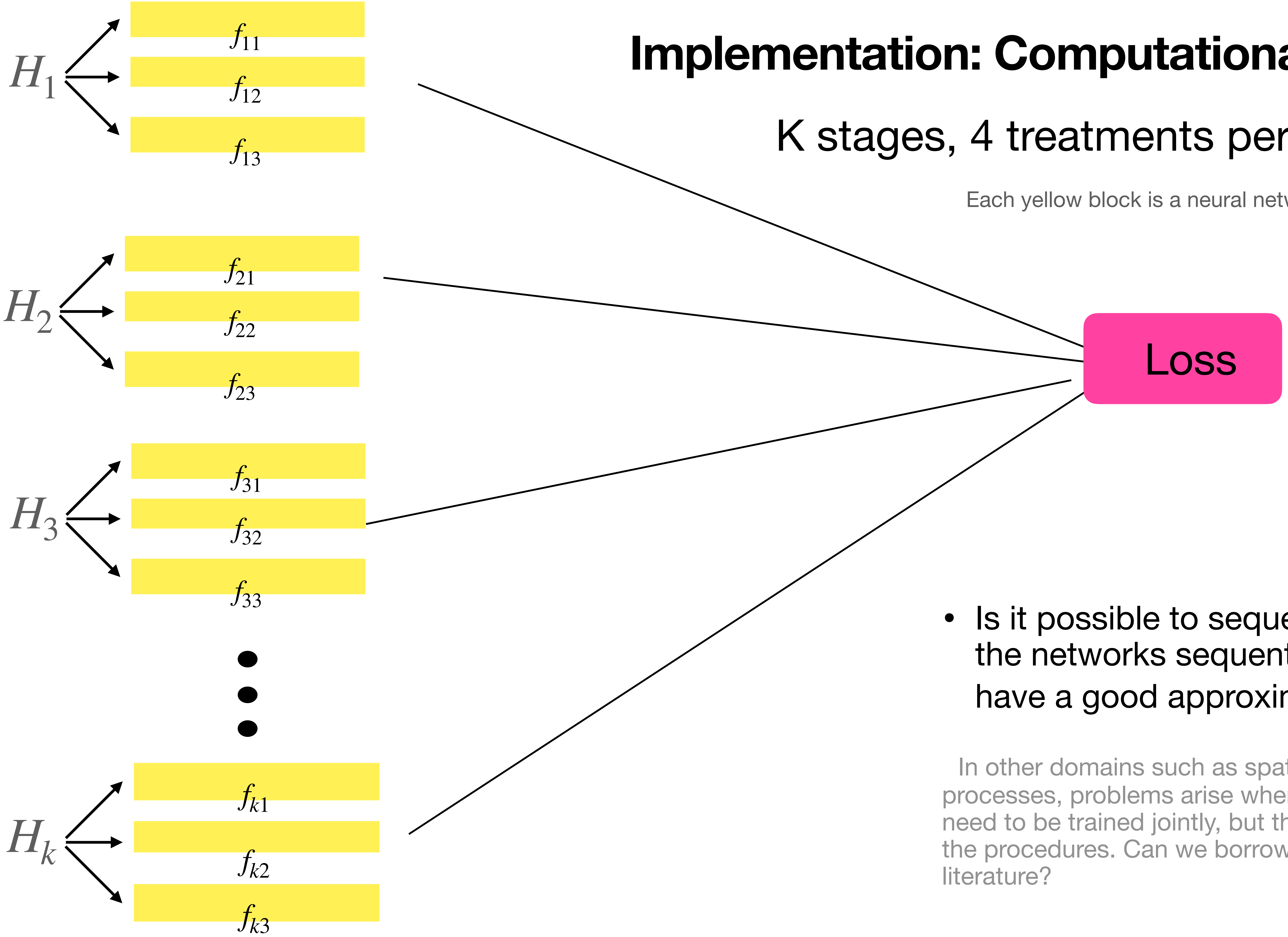



- Is it possible to sequentialize: train the networks sequentially but still have a good approximation of $\pi^*$?

# Implementation: Computational challenge

K stages, 4 treatments per stage

Each yellow block is a neural network




- Is it possible to sequentialize: train the networks sequentially but still have a good approximation of $\pi^*$?

In other domains such as spatio-tempral processes, problems arise where neural networks need to be trained jointly, but they sequentialize the procedures. Can we borrow ideas from that literature?

# Optimization

# Optimization

# Optimization

- Currently using stochastic gradient descent (SGD) for optimization — too general. We have a specific problem — can we tailor an optimization method?

# Optimization

- Currently using stochastic gradient descent (SGD) for optimization — too general. We have a specific problem — can we tailor an optimization method?

- Feng et al. (2022) used similar  loss function for another machine learning problem called ‘maximum score estimation”, and tailored an optimization method for their problem. Can we do something similar?

# Optimization

- Currently using stochastic gradient descent (SGD) for optimization — too general. We have a specific problem — can we tailor an optimization method?
- Feng et al. (2022) used similar  loss function for another machine learning problem called ‘maximum score estimation”, and tailored an optimization method for their problem. Can we do something similar?

Will require analysis of the optimization landscape

# Optimization landscape

1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Optimization landscape

1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Optimization landscape

- Neural network classifiers:

1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Optimization landscape

- Neural network classifiers:

  Existing deep learning results: can be used[1].

1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Optimization landscape

- Neural network classifiers:

Existing deep learning results: can be used[1].

**Challenges:** loss non-standard, existing results not directly applicable

1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Optimization landscape

- Neural network classifiers:

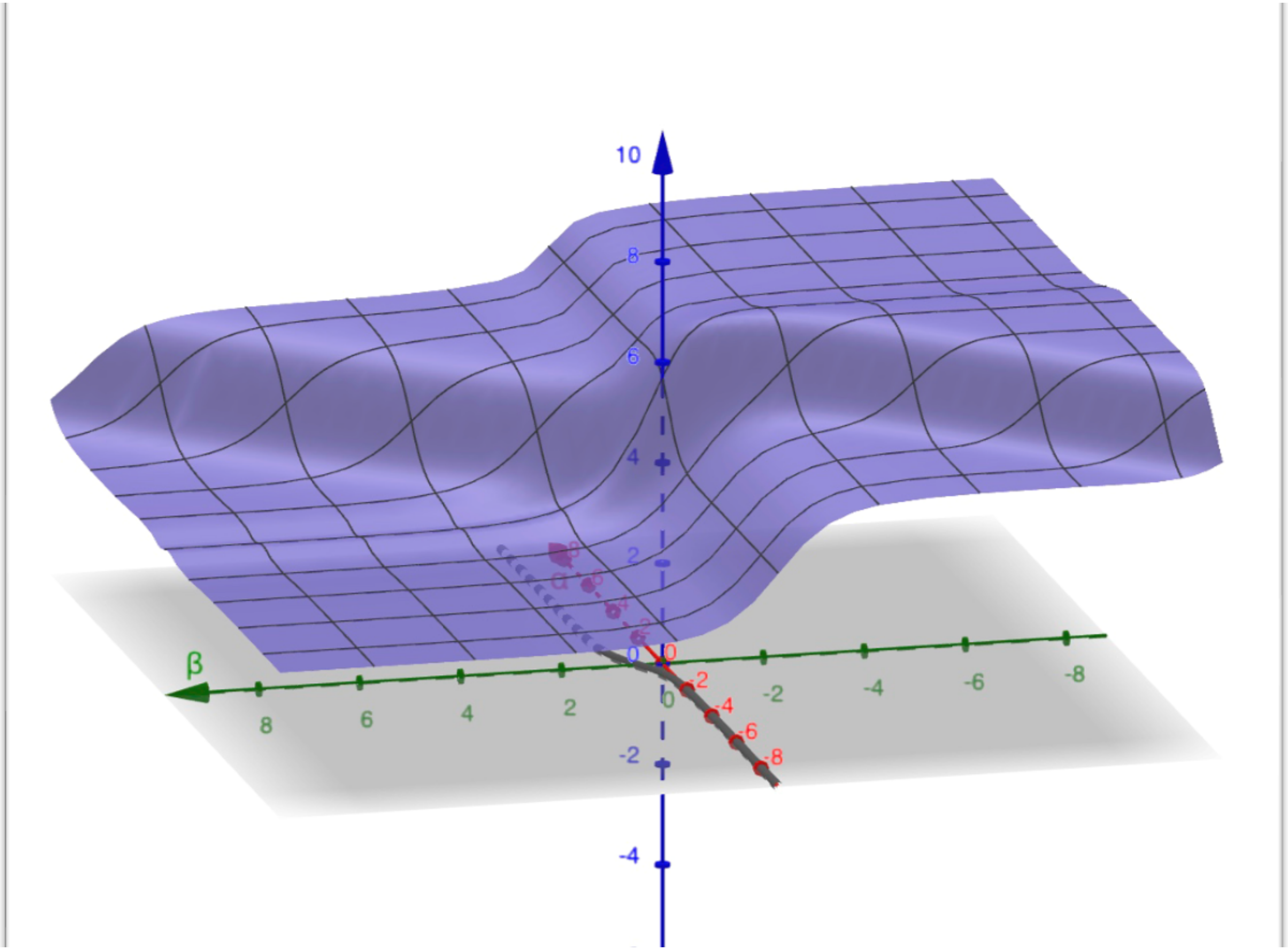
  Existing deep learning results: can be used[1].

  **Challenges:** loss non-standard, existing results not directly applicable

- Linear classifiers:

  optimization surface — specific properties: No local minima + regions with small gradient[2]

**K=1, 3 treatments, one covariate ($S_1 \in \mathbb{R}$), linear classifier**




1. Nguyen et al., 2017 and 2019
2. Laha et al., 2022

# Skills you will learn

1. Bottou (2018)

# Skills you will learn

1. Bottou (2018)

# Skills you will learn

DTR

1. Bottou (2018)

# Skills you will learn

DTR

PyTorch

1. Bottou (2018)

# Skills you will learn

DTR

PyTorch

Working with deep neural nets

1. Bottou (2018)

# Skills you will learn

DTR

PyTorch

Working with deep neural nets

Convergence of optimization-methods for non-convex problems[1]

1. Bottou (2018)

# Skills you will learn

DTR

PyTorch

Working with deep neural nets

Convergence of optimization-methods for non-convex problems[1]

1. Bottou (2018)

# Outline

- Example: sepsis
- Problem formulation
- Proposed method
- Open questions

A. Implementation and optimization

B. Regret decay rate

C. Doubly robust learning

# Open questions

## Regret decay

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$.

What is the rate of decay of regret?

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't be very different from the 2-treatments case (Laha et al., 2022).

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't be very different from the 2-treatments case (Laha et al., 2022).

Skills you will learn:

# Open questions
## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't  be very different from  the 2-treatments case (Laha et al., 2022).

Skills you will learn:

1. DTR

# Open questions
## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't be very different from the 2-treatments case (Laha et al., 2022).

Skills you will learn:

1. DTR
2. Empirical risk minimization theory

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't  be very different from  the 2-treatments case (Laha et al., 2022).

Skills you will learn:

1. DTR
2. Empirical risk minimization theory
3. Some theory on multicategory classification

# Open questions

## Regret decay

- Regret: $V^{\pi^*} - V^{\widehat{\pi}}$ measures how well we approximated $\pi^*$ using $\hat{\pi}$ .

What is the rate of decay of regret?

Probably won't be very different from the 2-treatments case (Laha et al., 2022).

Skills you will learn:

1. DTR
2. Empirical risk minimization theory
3. Some theory on multicategory classification
4. Some theory on policy learning in offline RL

# Outline

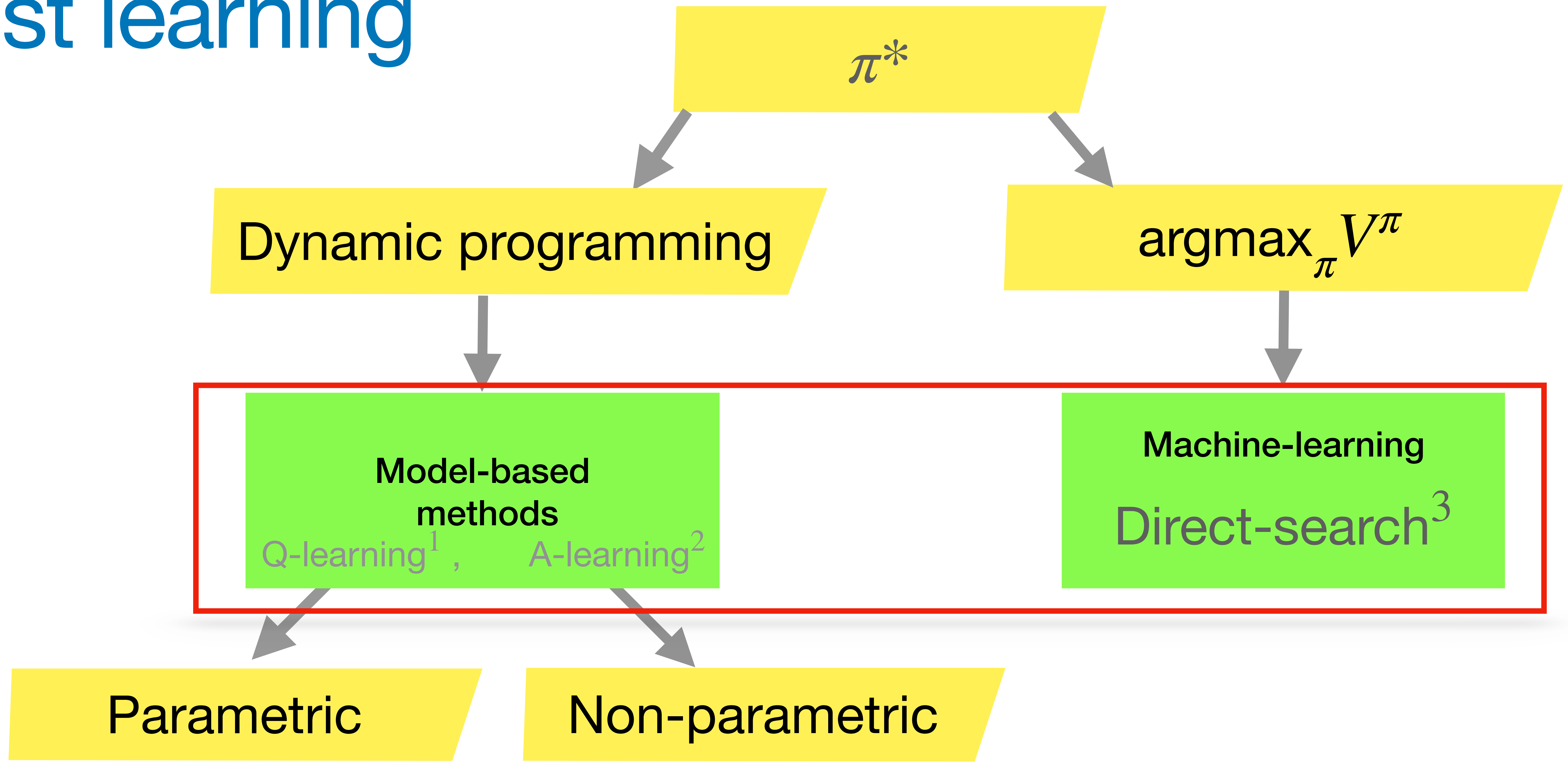
- Example: sepsis
- Problem formulation
- Proposed method
- Open questions

A. Implementation and optimization

B. Regret decay rate

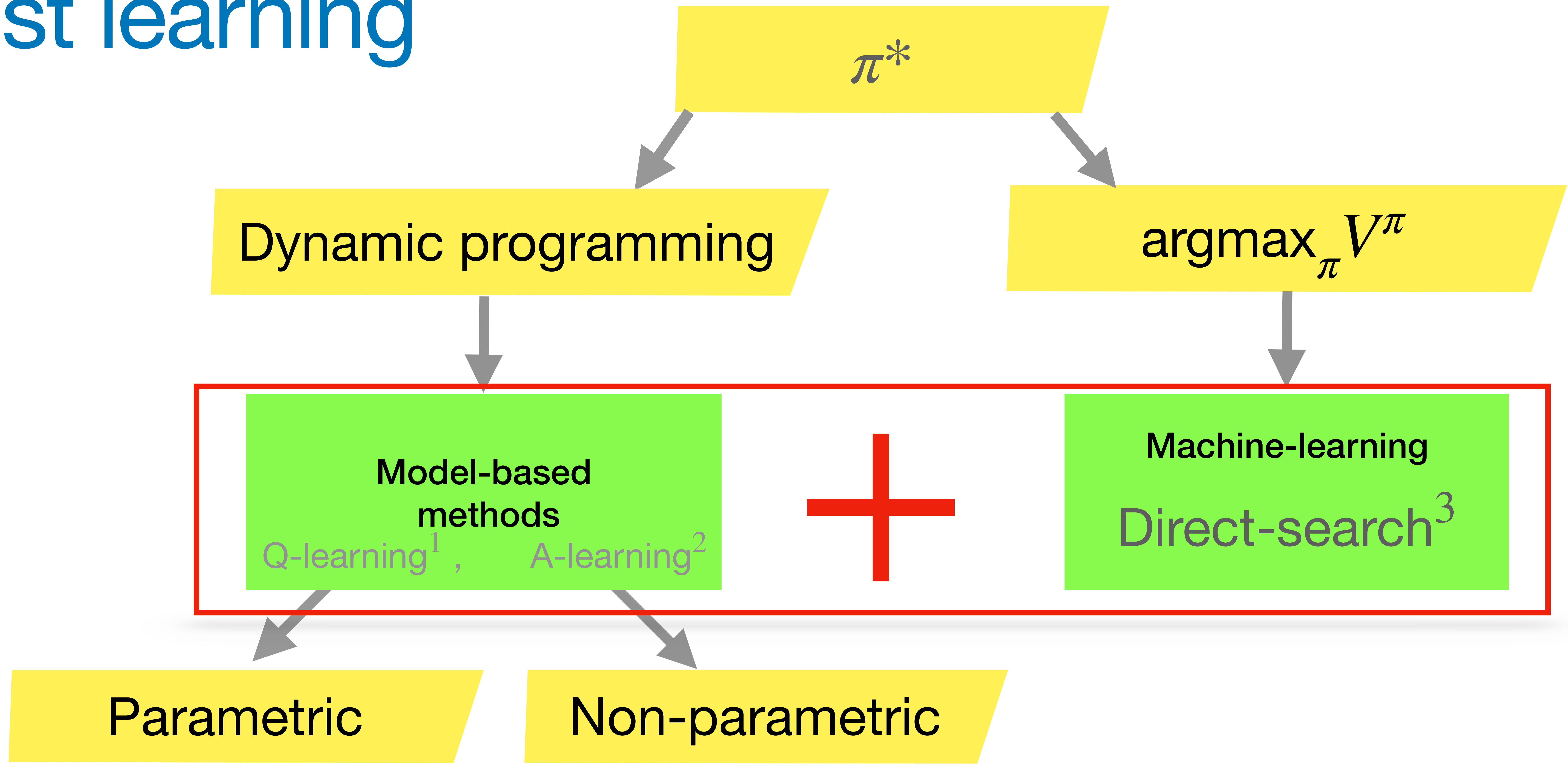
C. Doubly robust learning

# Doubly robust learning




1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2023

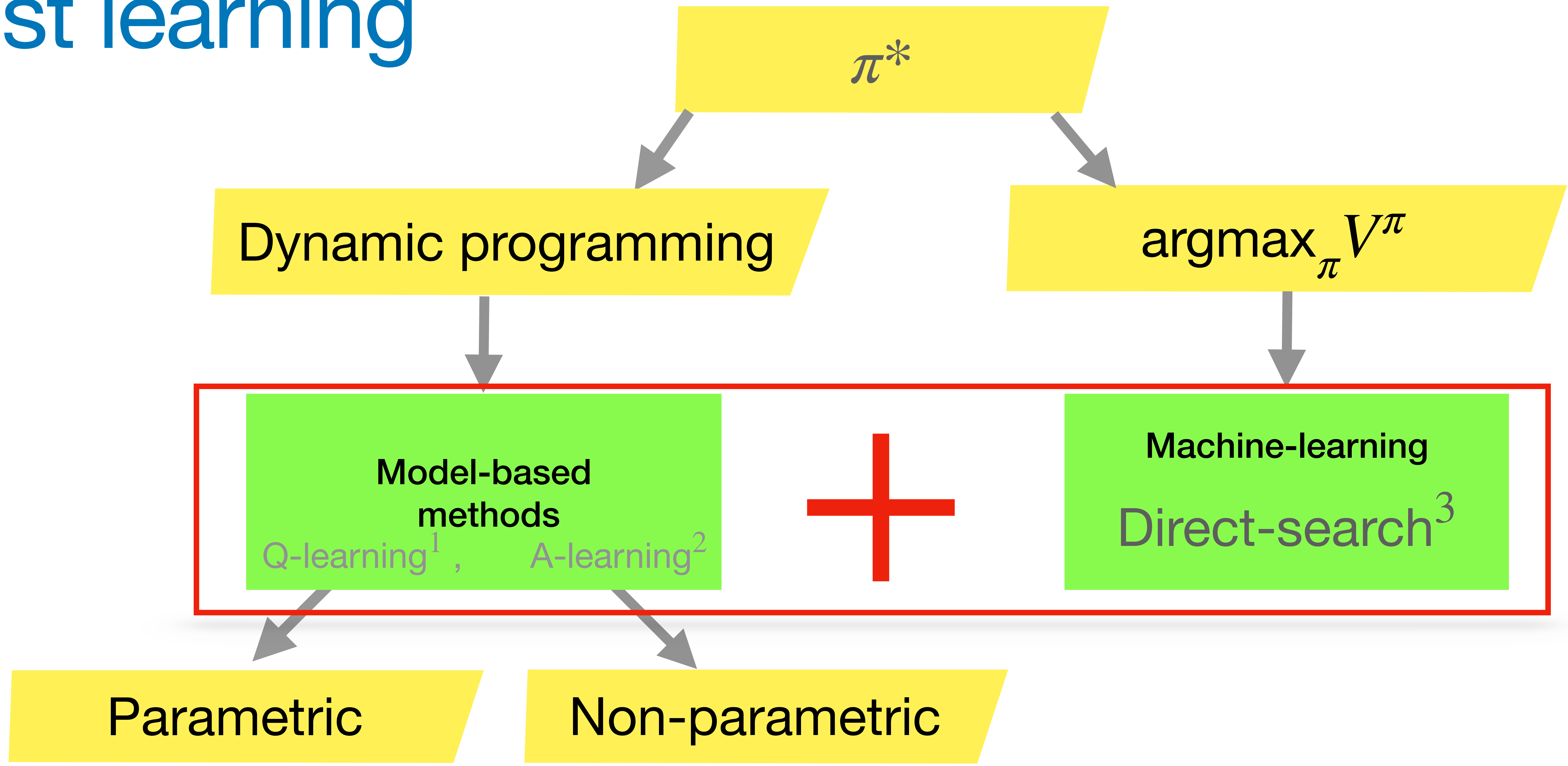
# Doubly robust learning




Hybrid method (idea taken from offline RL)

1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2023

# Doubly robust learning

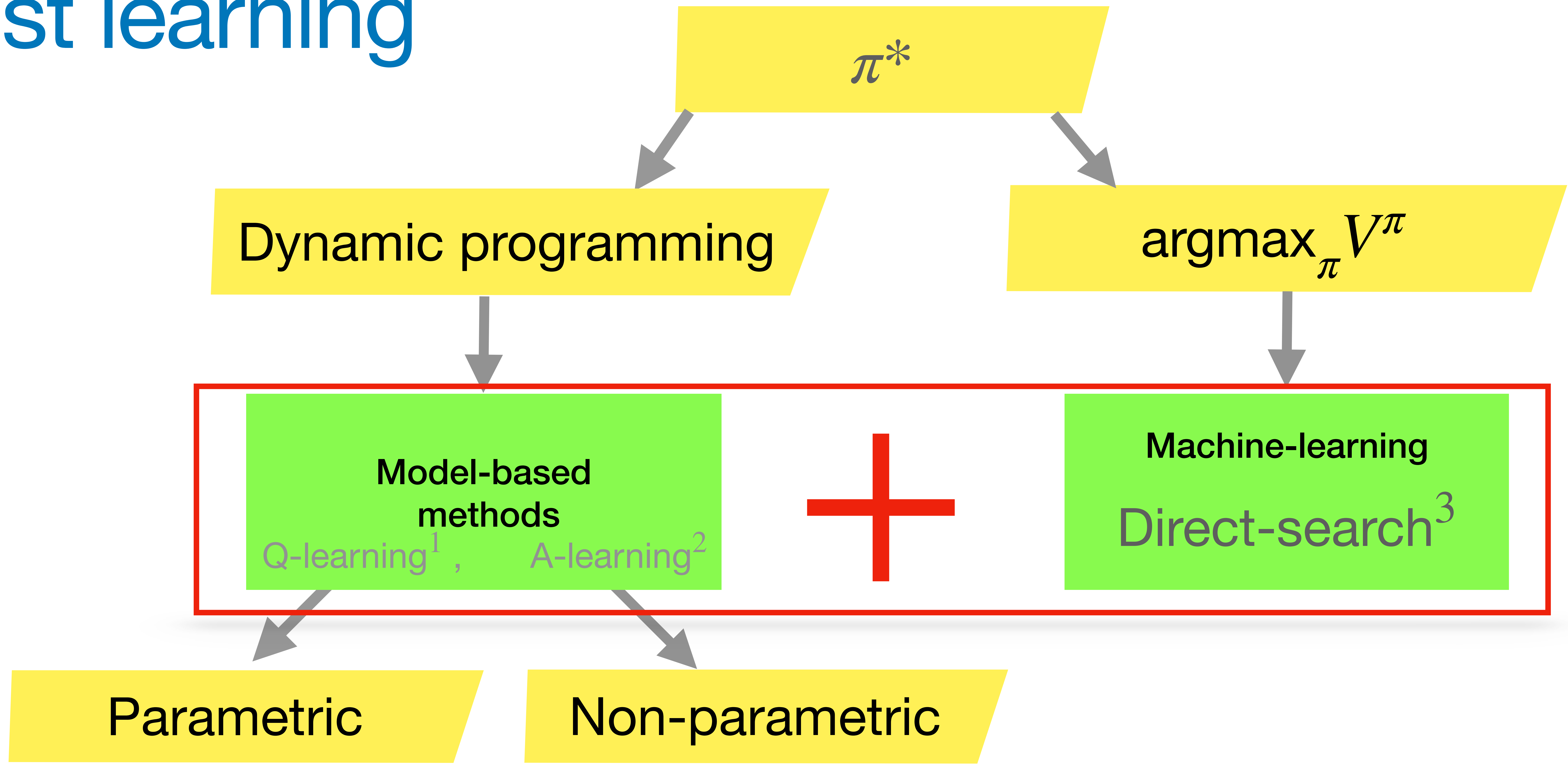



Hybrid method (idea taken from offline RL)

If either the Q-learning model assumptions or the estimation of treatment assignment probabilities correct, then $\pi^*$ consistently estimated

1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2023

# Doubly robust learning




Hybrid method (idea taken from offline RL)

If either the Q-learning model assumptions or the estimation of treatment assignment probabilities correct, then $\pi^*$ consistently estimated

1. Watkins, 1989; Schulte et al. 2014
2. Murphy, 2003; Robins, 2004
3. Zhao et al. 2012; 2015, Laha et al. 2023

# Open questions

# Open questions

1. I already have the method, but same questions on implementation

# Open questions

1. I already have the method, but same questions on implementation
2. regret decay: $\sqrt{n}-$consistent

# Open questions

1. I already have the method, but same questions on implementation

2. regret decay: $\sqrt{n}-$consistent

Skills you will learn:

# Open questions

1. I already have the method, but same questions on implementation
2. regret decay: $\sqrt{n}-$consistent

Skills you will learn:

1. DTR

# Open questions

1. I already have the method, but same questions on implementation
2. regret decay: $\sqrt{n}-$consistent

Skills you will learn:

1. DTR
2. Q-learning

# Open questions

1. I already have the method, but same questions on implementation
2. regret decay: $\sqrt{n}-$consistent

Skills you will learn:

1. DTR
2. Q-learning
3. doubly robust offline RL

# Open questions

1. I already have the method, but same questions on implementation

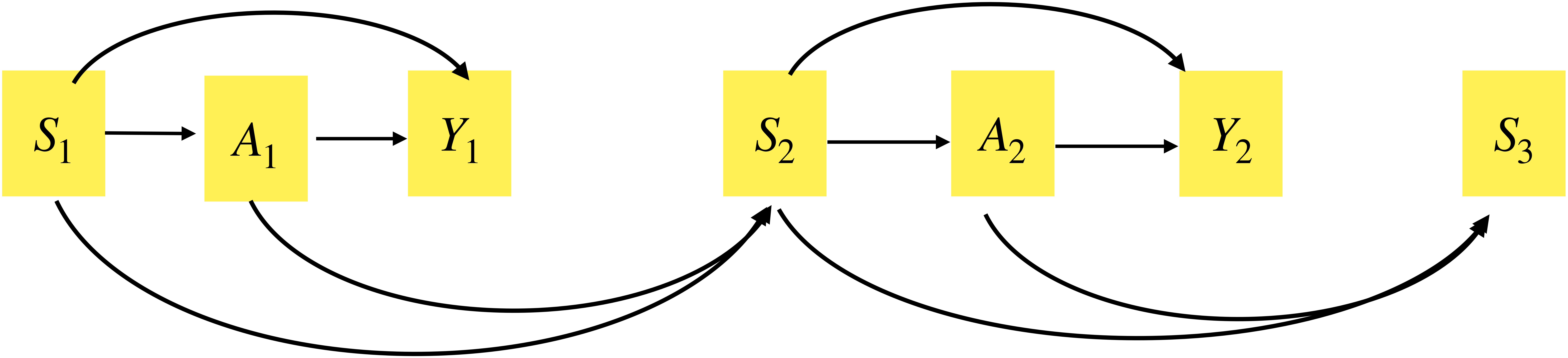
2. regret decay: $\sqrt{n}-$consistent

Skills you will learn:

1. DTR
2. Q-learning
3. doubly robust offline RL
4. Some doubly robust literature in causal inference

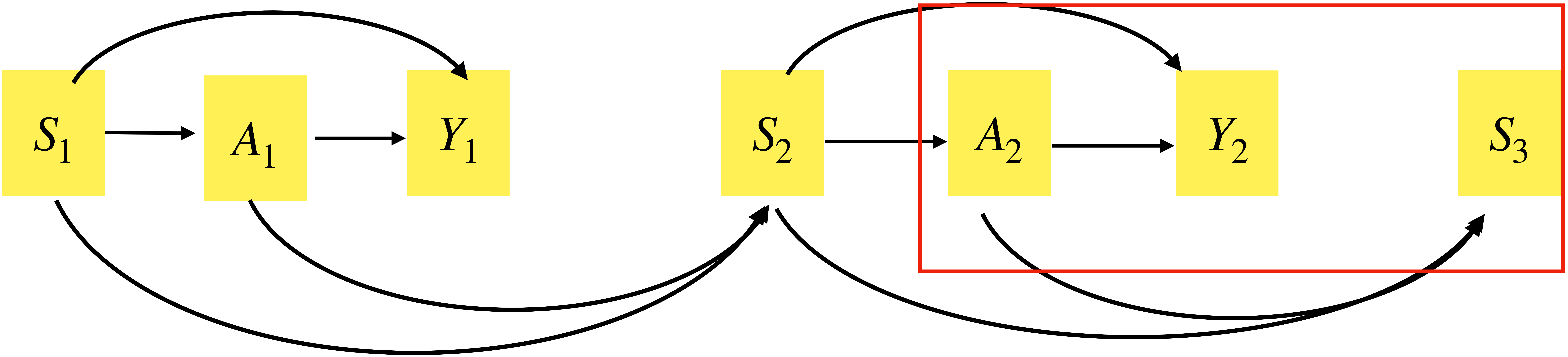
# Full reinforcement learning

We do not make Markov decision process (MDP) assumption



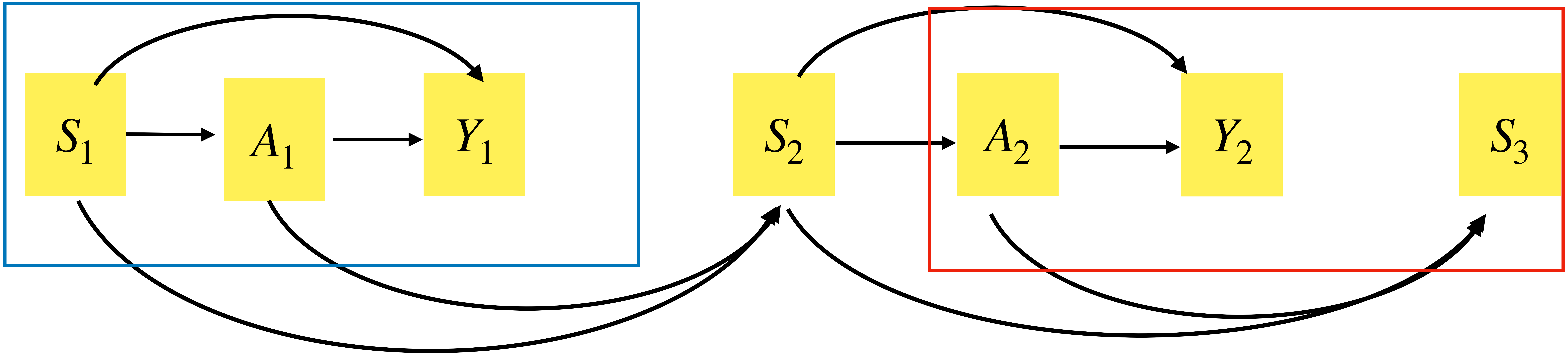
# Full reinforcement learning

We do not make Markov decision process (MDP) assumption

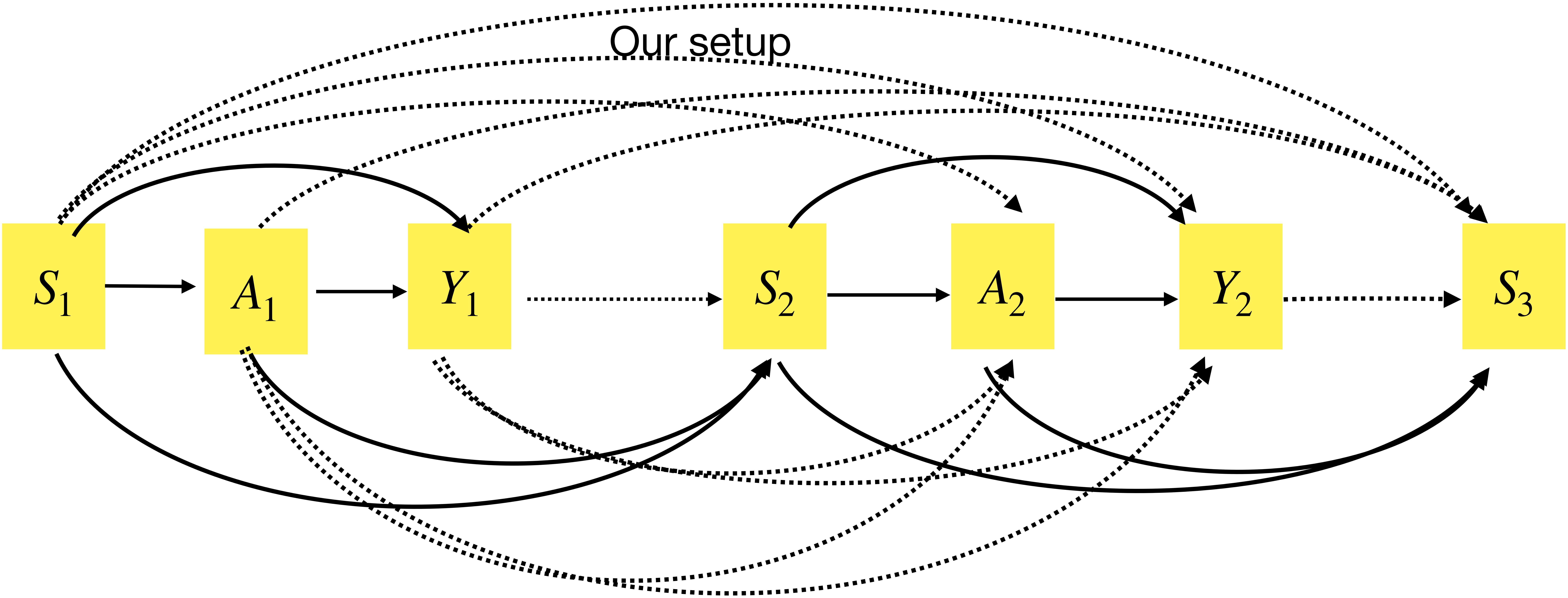


# Full reinforcement learning

We do not make Markov decision process (MDP) assumption

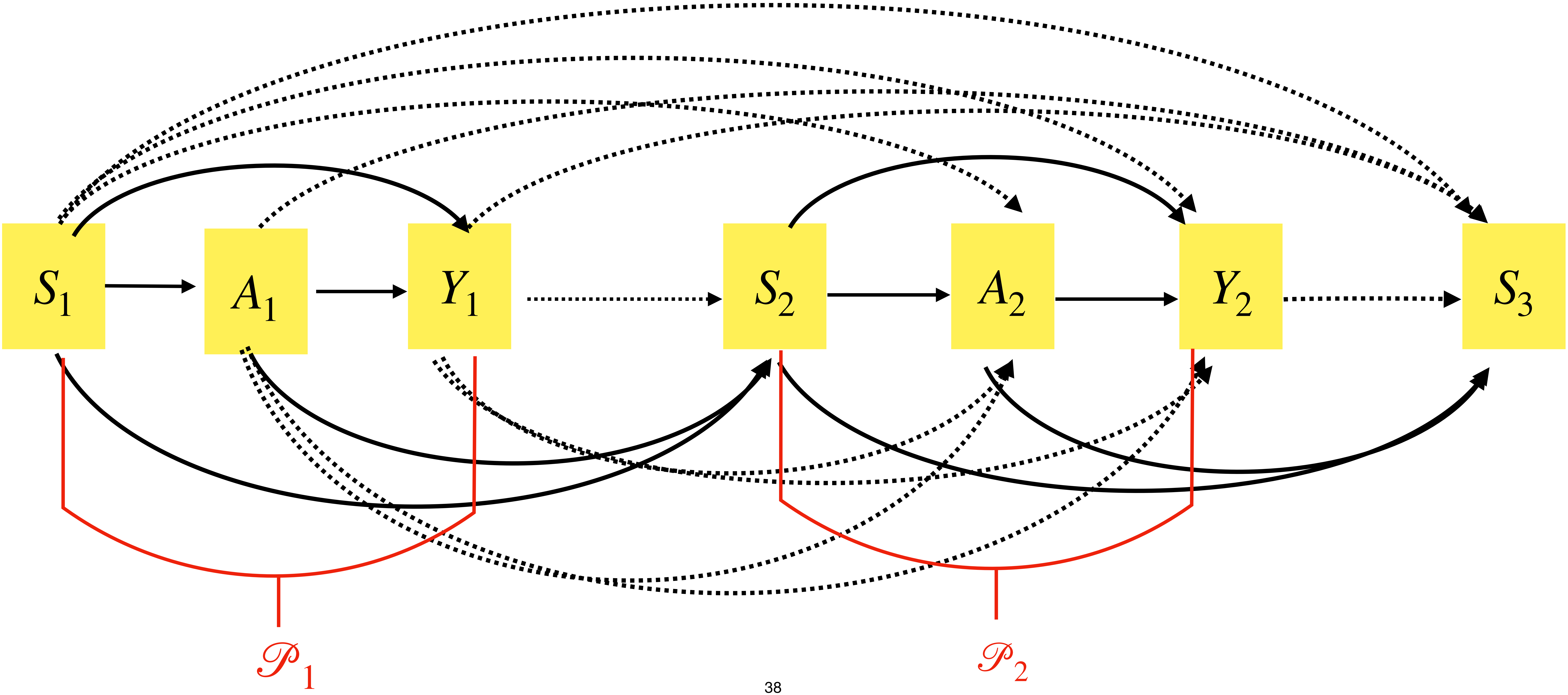


# Full reinforcement learning

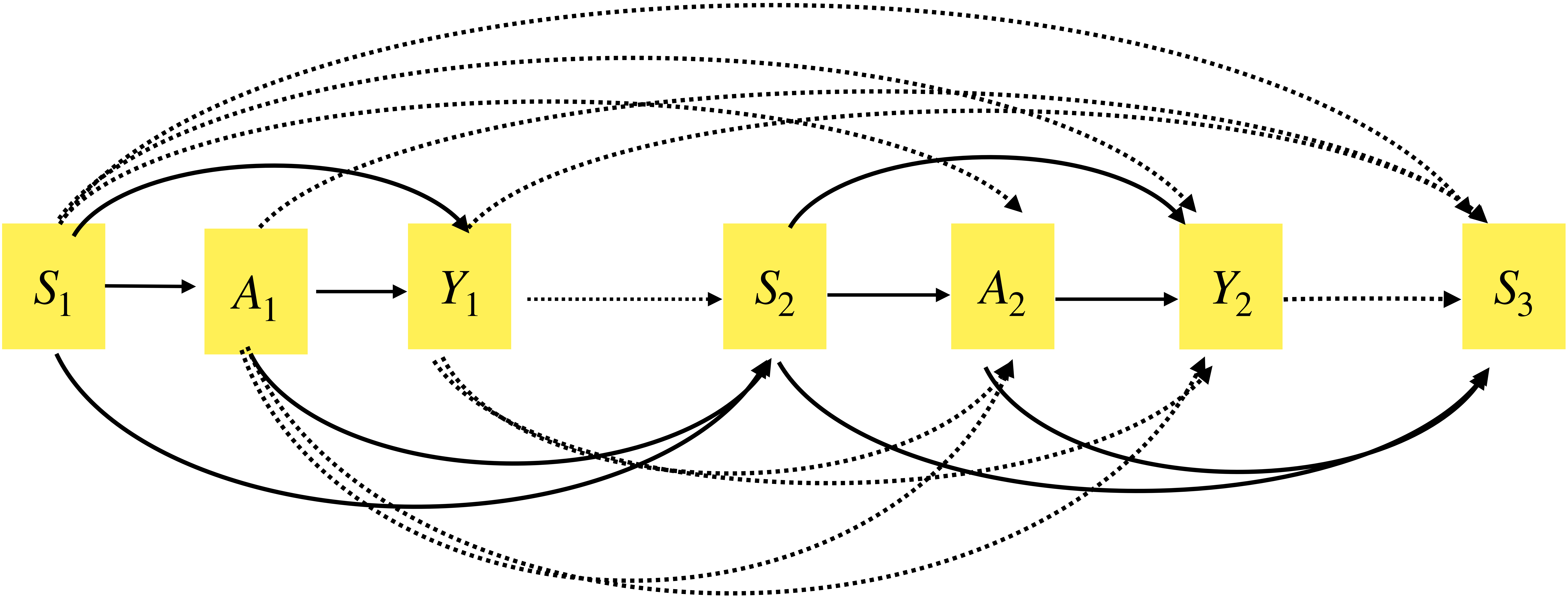


# Full reinforcement learning

No stationarity



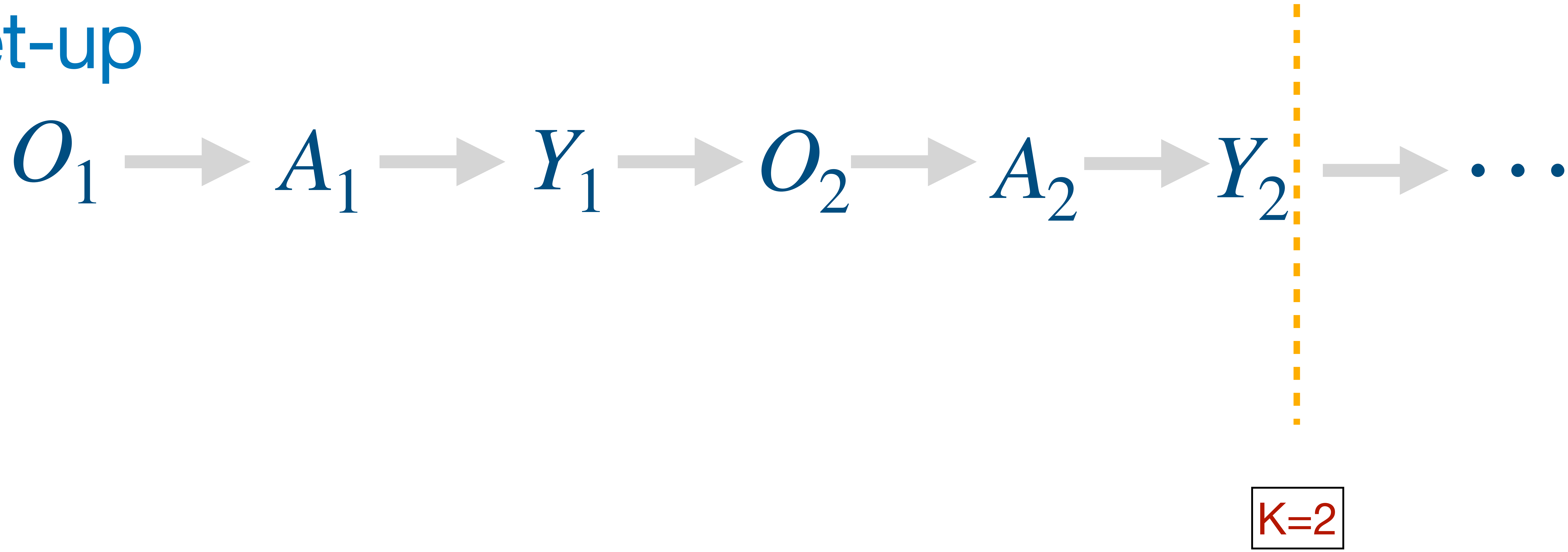
# Full reinforcement learning



# Set-up

$$O_1 \rightarrow A_1 \rightarrow Y_1 \rightarrow O_2 \rightarrow A_2 \rightarrow Y_2 \rightarrow \cdots$$

# Set-up

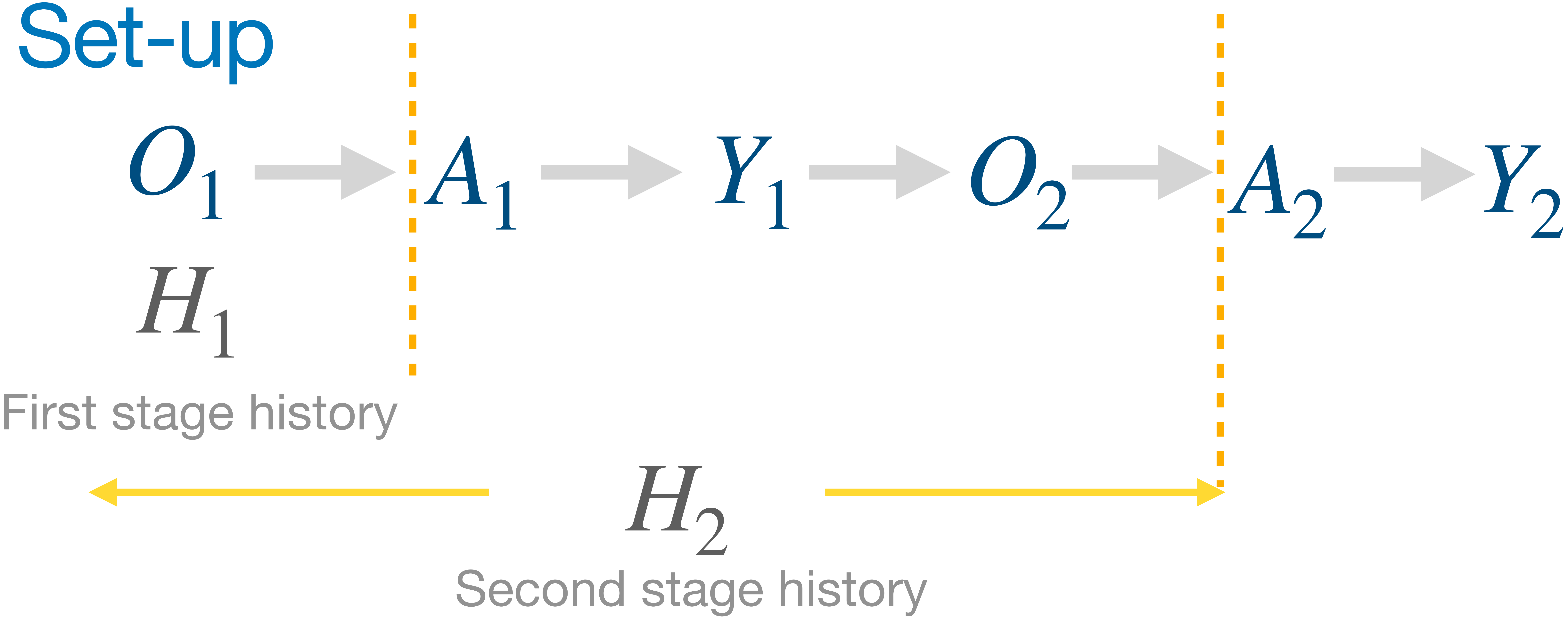


# Set-up

$$O_1 \rightarrow A_1 \rightarrow Y_1 \rightarrow O_2 \rightarrow A_2 \rightarrow Y_2$$

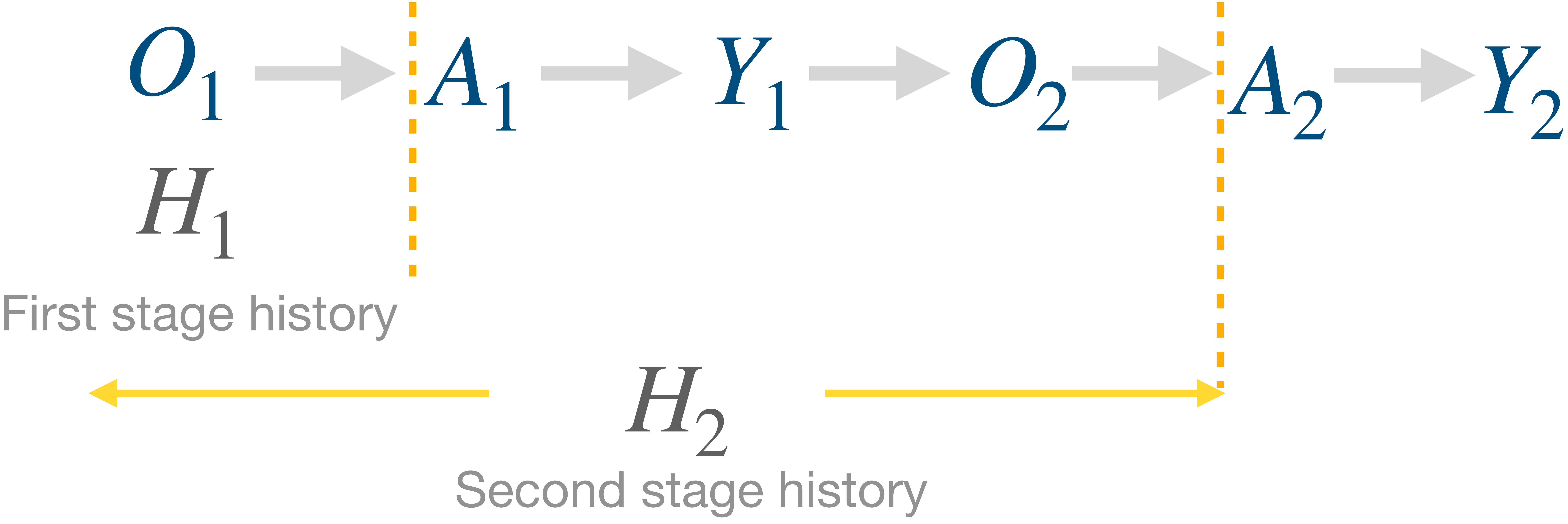
K=2

# Set-up



$O_1$
$A_1$
$Y_1$
$O_2$
$A_2$
$Y_2$
$H_1$
First stage history
$H_2$
Second stage history

# Set-up




Treatment policy $\pi = (\pi_1, \pi_2)$

# Value function estimation

$$V^{\pi} = \mathbb{E}[Y_1(\pi) + \ldots + Y_K(\pi)]$$

Optimal treatment assignment $\pi^* = \text{argmax}_{\pi} V^{\pi}$

# Value function estimation

$$V^{\pi} = \mathbb{E}[Y_1(\pi) + \ldots + Y_K(\pi)]$$

Potential outcomes

Optimal treatment assignment $\pi^* = \text{argmax}_{\pi} V^{\pi}$

# Value function estimation

Under standard identifiability assumptions*,

$$V^{\pi} = \mathbb{E}\left[(Y_1 + \ldots + Y_K)\frac{\pi_1(A_1 \mid H_1)\ldots\pi_K(A_K \mid H_K)}{\pi_{b,1}(A_1 \mid H_1) \quad \ldots\pi_{b,K}(A_K \mid H_K)}\right]$$

$\pi_{b,k}$'s behavior policy: ratio called inverse probability weights

Optimal treatment assignment $\pi^* = \text{argmax}_{\pi}\ V^{\pi}$

*Orellana et al., 2010

# Value function estimation

Under standard identifiability assumptions*,

$$V^{\pi} \approx \mathbb{P}_n \left[ (Y_1 + \ldots + Y_K) \frac{\pi_1(A_1 \mid H_1) \ldots \pi_K(A_K \mid H_K)}{\pi_{b,1}(A_1 \mid H_1) \ \ldots \pi_{b,K}(A_K \mid H_K)} \right]$$

$\mathbb{P}_n$ : empirical distribution function

Optimal treatment assignment $\pi^* = \text{argmax}_{\pi} \, V^{\pi}$

*Orellana et al., 2010

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi V^\pi$

$$V^\pi = \mathbb{E}[Y_1(\pi) + Y_2(\pi)]$$

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi \, V^\pi$

$$V^\pi = \mathbb{E}[Y_1(\pi) + Y_2(\pi)]$$

Potential outcomes

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi \; V^\pi$

Under standard identifiability assumptions*,

$$V^\pi = \mathbb{E}\left[ (Y_1 + Y_2) \frac{1\{\pi_1(H_1) = A_1\} \; 1\{\pi_2(H_2) = A_2\}}{P(A_1 \mid H_1) \; P(A_2 \mid H_2)} \right]$$

observed random variables

*Orellana et al., 2010

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi V^\pi$

Under standard identifiability assumptions*,

$$V^\pi \approx \frac{1}{n}\sum_{i=1}^{n}\left((Y_{1i} + Y_{2i})\frac{1\{\pi_1(H_{1i}) = A_{1i}\}\ 1\{\pi_2(H_{2i}) = A_{2i}\}}{P(A_{1i} \mid H_{1i})\ P(A_{2i} \mid H_{2i})}\right)$$

*Orellana et al., 2010

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi V^\pi$

Under standard identifiability assumptions*,

$$V^\pi \approx \frac{1}{n}\sum_{i=1}^{n}\left((Y_{1i} + Y_{2i})\frac{1\{\pi_1(H_{1i}) = A_{1i}\}\ 1\{\pi_2(H_{2i}) = A_{2i}\}}{P(A_{1i} \mid H_{1i})\ P(A_{2i} \mid H_{2i})}\right)$$

Maximize $V^\pi$ over a. Class of policies

*Orellana et al., 2010

# Value function

Optimal treatment policy $\pi^* = \text{argmax}_\pi \, V^\pi$

Under standard identifiability assumptions*,

$$V^\pi \approx \frac{1}{n}\sum_{i=1}^{n}\left( (Y_{1i} + Y_{2i}) \frac{1\{\pi_1(H_{1i}) = A_{1i}\}\ 1\{\pi_2(H_{2i}) = A_{2i}\}}{P(A_{1i} \mid H_{1i})\ P(A_{2i} \mid H_{2i})} \right)$$

**Discontinuous + non, convex**

Direct optimization not computationally feasible

*Orellana et al., 2010

# Shortcomings of previous method

- $\min_{f: H \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}(f(H_1)) \neq A_1]\right]$

# Shortcomings of previous method

- $$\min_{f: H \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}(f(H_1)) \neq A_1]\right]$$

$$C(H_1, Y_1) = \frac{Y_1}{P(A_1 \mid H_1)}$$

# Shortcomings of previous method

- $$\min_{f: H \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}(f(H_1)) \neq A_1]\right]$$

$$C(H_1, Y_1) = \frac{Y_1}{P(A_1 \mid H_1)}$$

If I don't know what doctors were thinking, need to model the probabilities

# Shortcomings of previous method

- $$\min_{f:H \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}(f(H_1)) \neq A_1]\right]$$

$$C(H_1, Y_1) = \frac{Y_1}{P(A_1 \mid H_1)}$$

If I don't know what doctors were thinking, need to model the probabilities

Bad estimation

$\widehat{\pi}$ bad estimator of $\pi^*$

# Shortcomings of previous method

- $$\min_{f: H \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}(f(H_1)) \neq A_1]\right]$$

$$C(H_1, Y_1) = \frac{Y_1}{P(A_1 \mid H_1)}$$

If I don't know what doctors were thinking, need to model the probabilities

$P(A_1 \mid H_1)$ is small $\Longrightarrow$ the estimator of $C(H_1, A_1)$ can be highly variable

# Loss function when stage $K = 1$

# Classifiers for stage 1

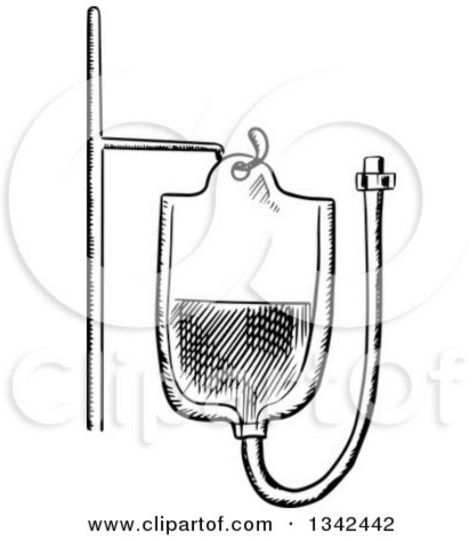
# Classifiers for stage 1

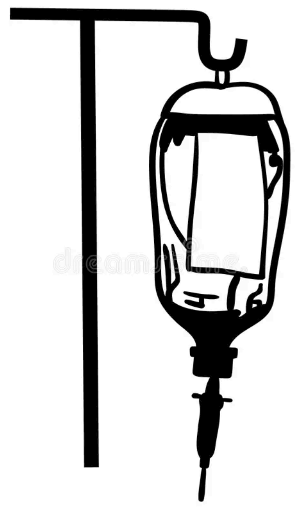
$H_1$

# Classifiers for stage 1

$H_1$

## Possible categories



Low

Medium

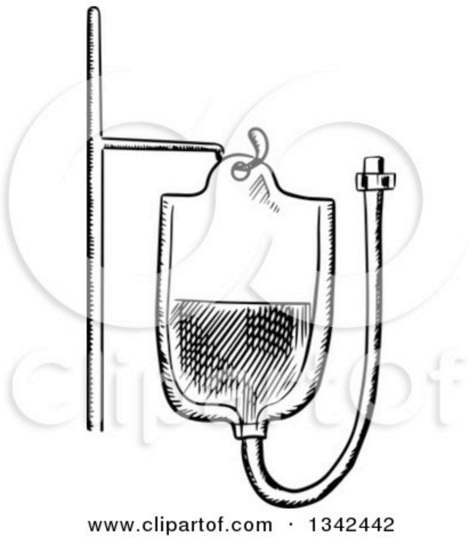
High

No IV

# Classifiers for stage 1

## Possible categories

$H_1$

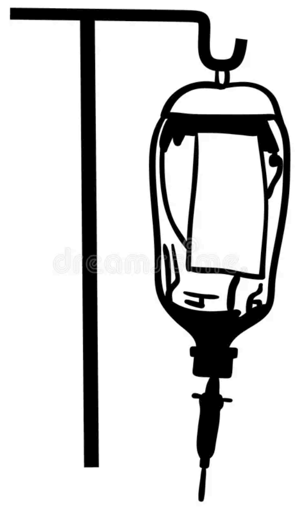


Low

$f_1(H_1)$

Medium

$f_2(H_1)$



High

$f_3(H_1)$

No IV

$f_4(H_1)$

Classifier:

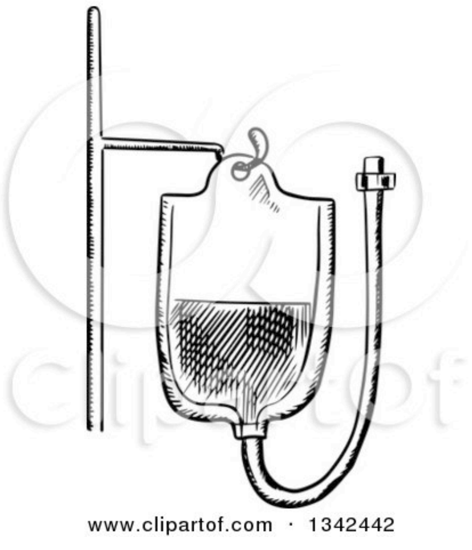
$$f = (f_1, \ldots, f_4)$$

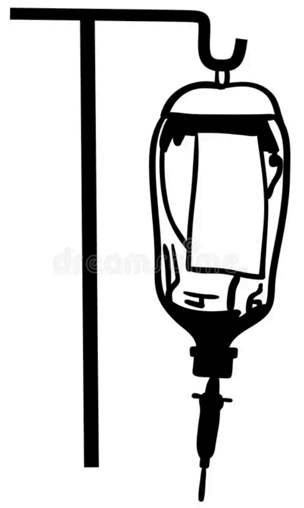
$f_i : H_1 \mapsto \mathbb{R} \quad i = 1,\ldots,4$

# Classifiers for stage 1

## Possible categories

$H_1$





| Low | Medium | High | No IV |
|---|---|---|---|
| $f_1(H_1)$ | $f_2(H_1)$ | $f_3(H_1)$ | $f_4(H_1)$ |

Classifier:

$$f = (f_1, \ldots, f_4)$$

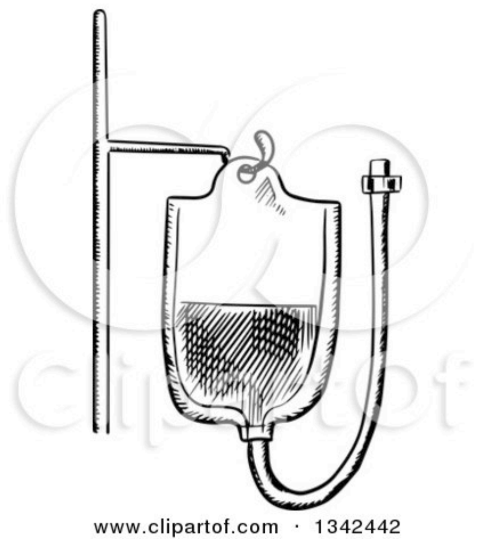
$$f_i : H_1 \mapsto \mathbb{R} \quad i = 1,\ldots,4$$

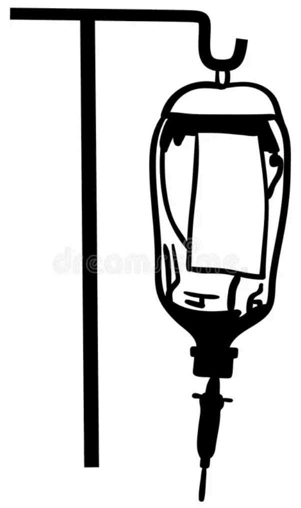
**Maximum**

# Classifiers for stage 1

$H_1$

## Possible categories



Low | Medium | High | No IV

Classifier:

$f_1(H_1)$ $f_2(H_1)$ $f_3(H_1)$ $f_4(H_1)$

$$f = (f_1, \ldots, f_4)$$

$f_i : H_1 \mapsto \mathbb{R} \quad i = 1,\ldots,4$

**Maximum**

$$\pi_1(H_1) = \text{argmax}_i \; f_i(H_1)$$

# The loss function

**Case** $T = 1$

- $\max_{f: H_1 \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}_i \ f_i(H_1) \neq A_1]\right]$

# The loss function

## Case $T = 1$

- $\max_{f: H_1 \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}_i\ f_i(H_1) \neq A_1]\right]$

In practice search over a smaller class, currently we consider neural network classes

# The loss function

## Case $T = 1$

- $$\max_{f: H_1 \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}_i \ f_i(H_1) \neq A_1]\right]$$

Depends on data

# The loss function

## Case $T = 1$

- $\max_{f: H_1 \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}_i\, f_i(H_1) \neq A_1]\right]$

Discontinuity

# The loss function

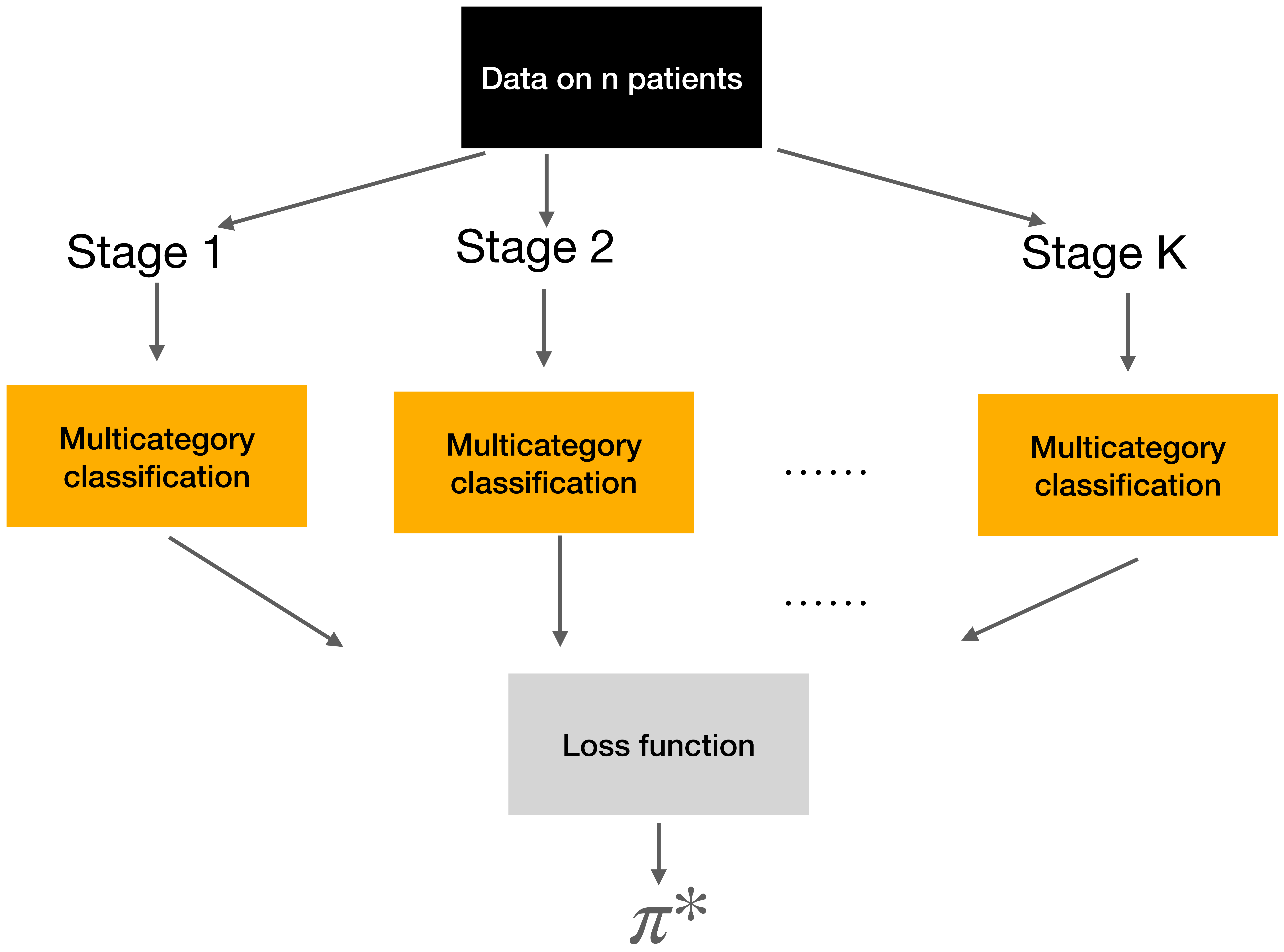
## Case $T = 1$

- $$\max_{f: H_1 \mapsto \mathbb{R}^4} E\left[C(H_1, Y_1) \times 1[\text{argmax}_i\ f_i(H_1) \neq A_1]\right]$$

Discontinuity

**Our proposal:** smooth out the sources for discontinuity at each step

# Smoothed loss function



# Smoothed loss function

# Smoothed loss function

# Smoothed loss function