Introduction

Fundamentals of Reinforcement Learning

Institut für Nachrichtentechnik
Fachgebiet Kommunikationstechnik
Prof. Dr.-Ing. Anja Klein,
Dr. Sabrina Klos & Dr. Andrea Ortiz

Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

- You can describe the characteristics and main elements of Reinforcement Learning and identify examples of Reinforcement Learning tasks.
- You can explain the main components of Reinforcement Learning agents.
- You can explain the main problems within Reinforcement Learning.

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- Characteristics of RL
- Components of RL Agents
- Problems within RL
- Lecture Overview

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Motivation**
- Characteristics of RL
- Components of RL Agents
- Problems within RL
- Lecture Overview

Idea of Reinforcement Learning (RL)

Core idea of RL is the fundamental way humans learn

https://pixabay.com/

https://pixabay.com/

https://pixabay.com/

Learning by interacting with the environment

- There is no explicit teacher.
- Learner has direct sensorimotor connection to the environment.

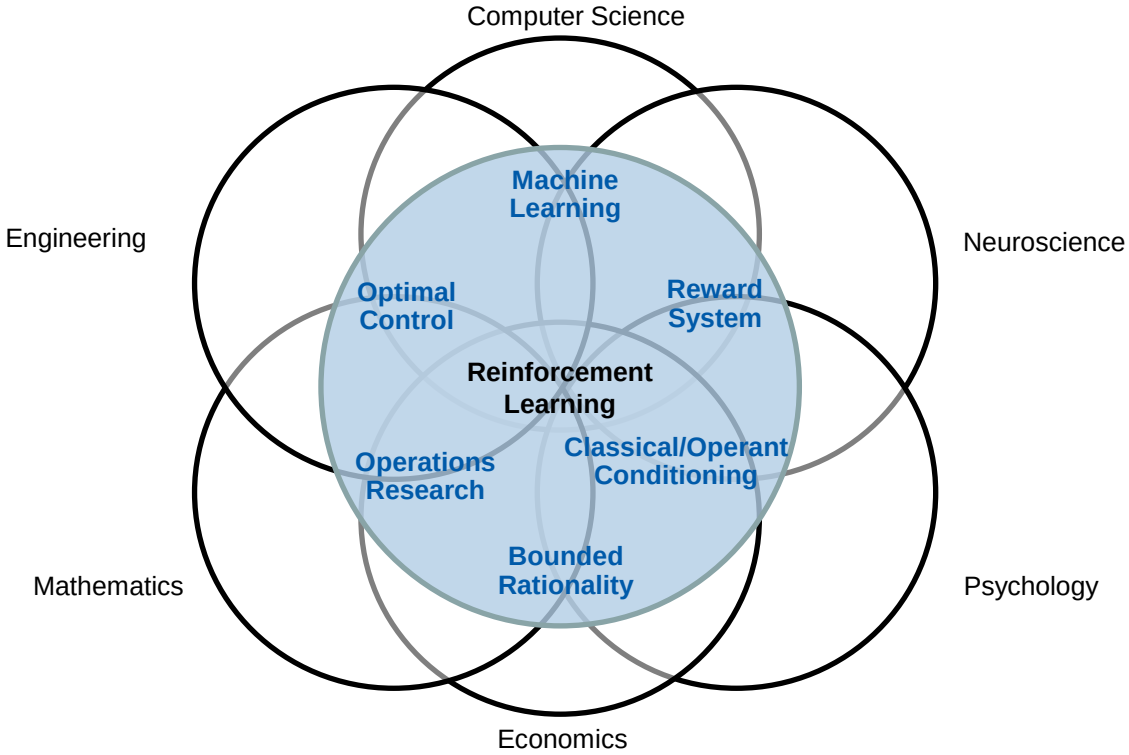
Idea of RL

RL is a computational approach to goal-directed learning from interaction

- RL explores idealized learning situations where an agent learns to map situations to actions in order to achieve some goal.
- RL deals with
  - How to design algorithms for machines that solve learning problems.
  - How to evaluate such designed algorithms through mathematical analysis or numerical evaluation.

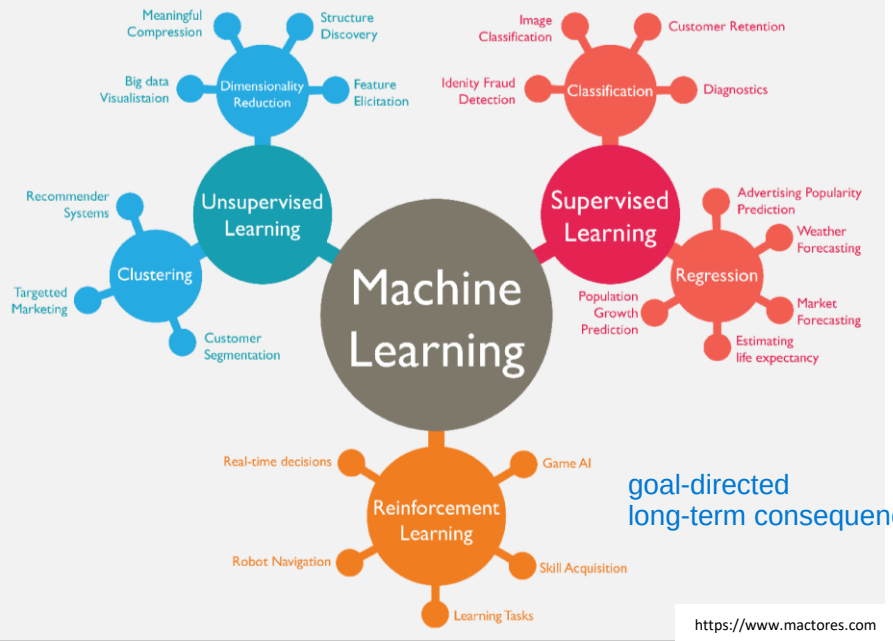
Idea of RL

RL relates to several scientific and engineering disciplines



Idea of RL

RL is a sub-category of Machine Learning (ML)



goal-directed
long-term consequences

https://www.mactores.com

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- **Characteristics of RL**
- Components of RL Agents
- Problems within RL
- Lecture Overview

Characteristics of RL

RL is different from other ML paradigms

TECHNISCHE UNIVERSITÄT DARMSTADT

Characteristics of RL	
 **Evaluative Feedback** != instructive	There is no supervisor, only a reward signal, i.e., trial-and-error search needed.
 **Delayed Feedback**	Reward feedback may be delayed, not instantaneous.
 **Sequential and Associative Setting** associada a uma observação	Time really matters, i.e., sequential non i.i.d data, and best action depends on situation.
 **Influence on Environment**	Actions may affect subsequent situations and rewards, i.e., actions may have long term consequences.

Examples of RL

How to make a robot pick pins from a bin

FANUC‘s bin-picking robot

https://www.youtube.com/watch?v=ydh_AdWZflA

Examples of RL

How to make an artificial system master the game of chess

https://deepmind.com

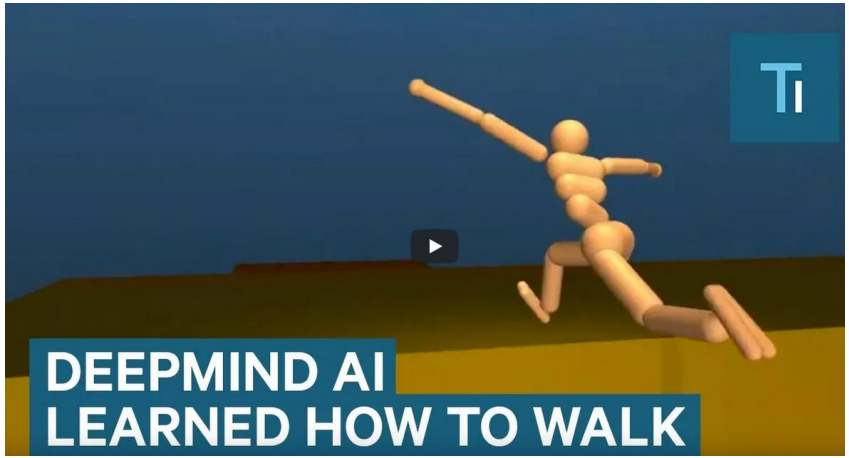
Google DeepMind‘s algorithm AlphaZero

https://www.youtube.com/watch?v=7L2sUGcOgh0

https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Examples of RL

How to make a virtual robot walk



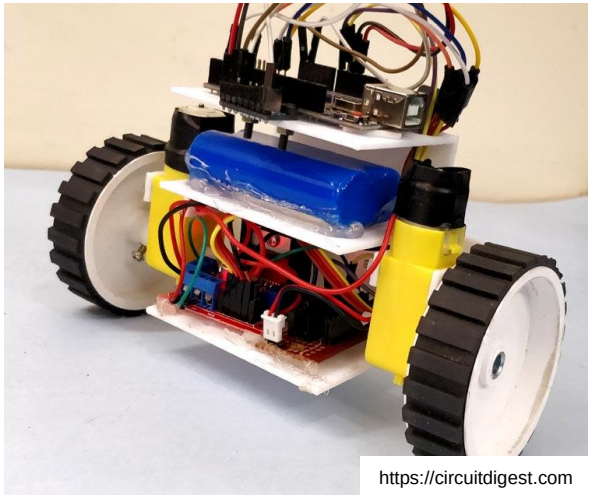
Google DeepMind‘s AI walkers

https://www.youtube.com/watch?v=gn4nRCC9TwQ

https://deepmind.com/blog/article/producing-flexible-behaviours-simulated-environments

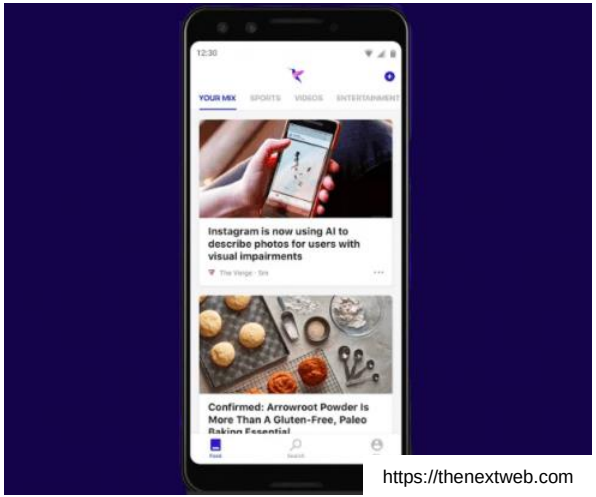
Question

Which of these examples are potential RL tasks?



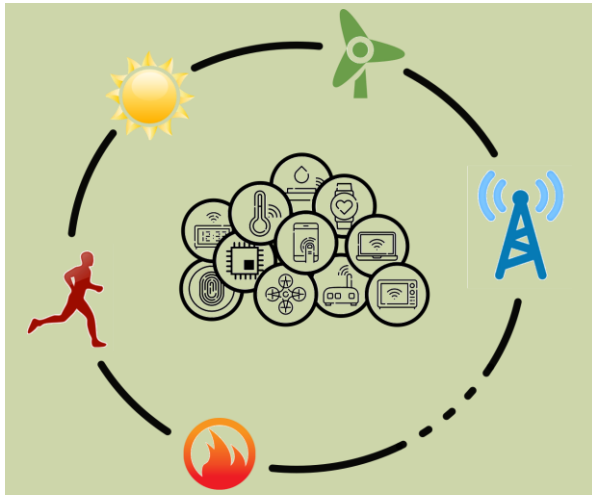
https://circuitdigest.com

Build a self-balancing robot



https://thenextweb.com

Personalize a news feed

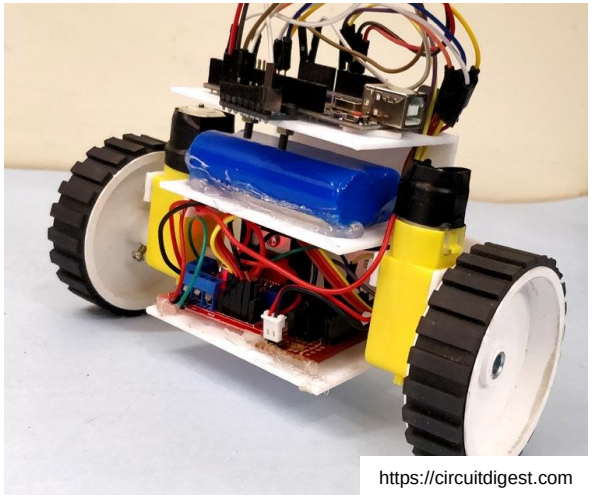


Optimize an energy harvesting communication system

tentar diferentes estratégias de envio e ver qual funciona melhor

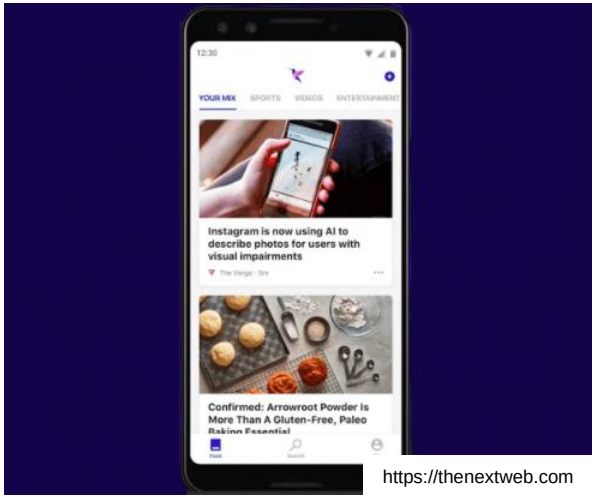
Answer

All of these examples are potential RL tasks!



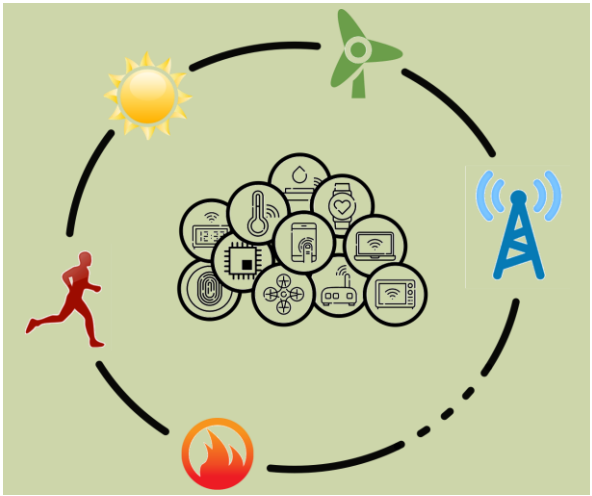
https://circuitdigest.com

Build a self-balancing robot



https://thenextweb.com

Personalize a news feed



Optimize an energy harvesting communication system

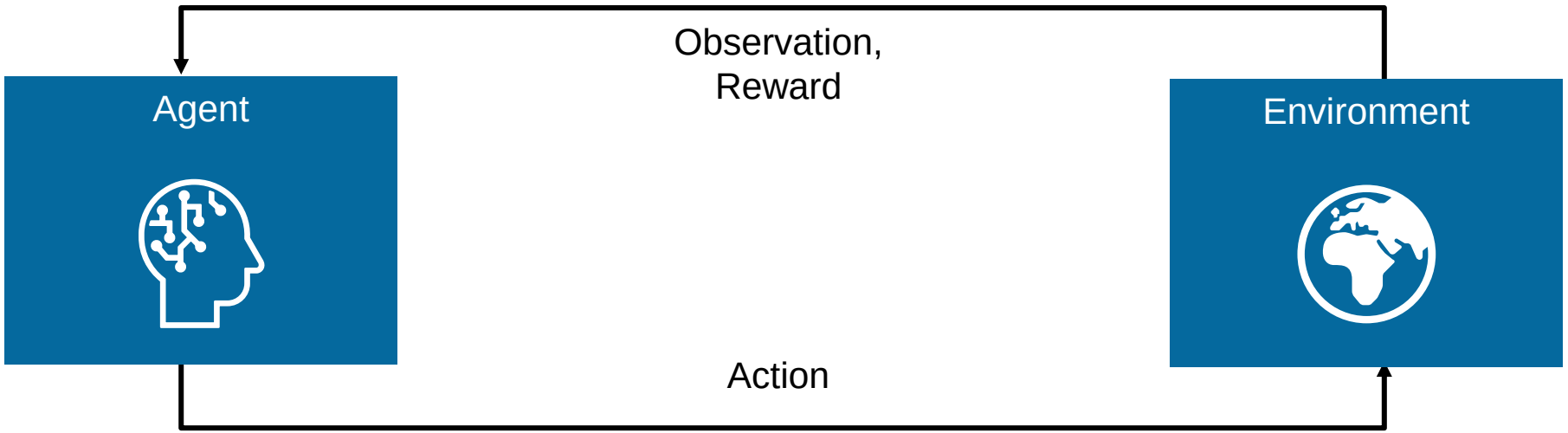
These are all potential tasks for an active decision-making agent interacting with its environment, seeking to achieve a goal despite uncertainty about its environment.

Agent and Environment

We can visualize this interaction in a diagram

TECHNISCHE UNIVERSITÄT DARMSTADT

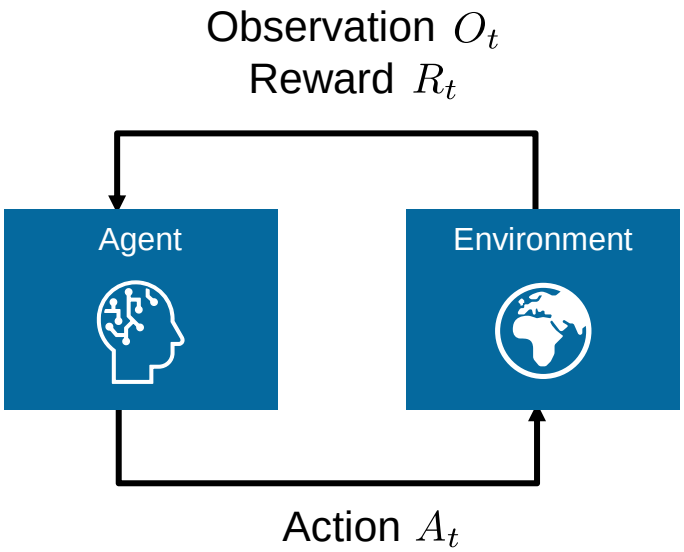
Agent–environment interaction



por ex, no robo do slide anterior: Envinroment vira qualquer coisa que ele não possa controlar totalmente e arbitráriamente

Agent and Environment

The agent and the environment interact sequentially



At each time step t:

The agent

- Receives observation O_t
- Executes action A_t
- Receives scalar reward R_t

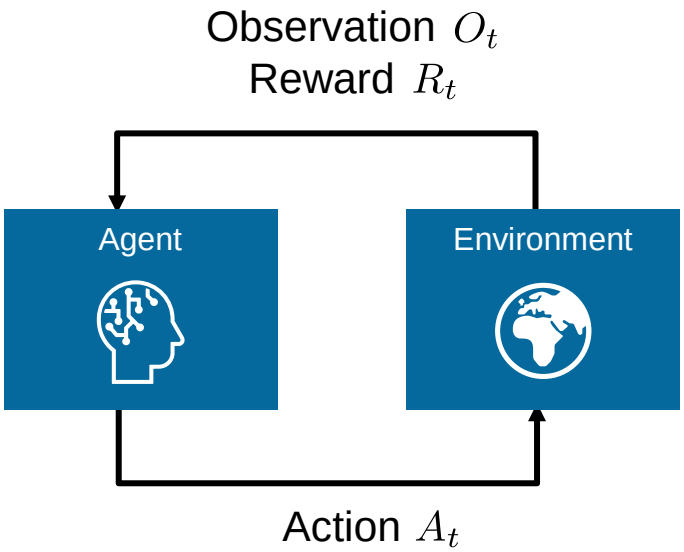
The environment

- Emits observation O_t
- Receives action A_t
- Emits scalar reward R_t

History

The history is the sequence of all observable variables up to time t

tentar maximizar soma das recompensas ou a recompensa final total



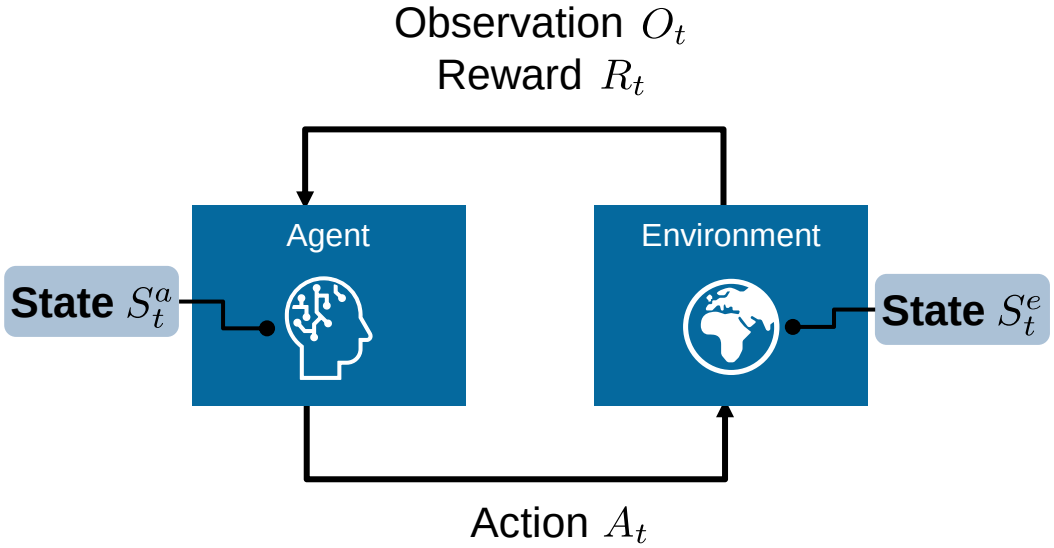
- **History** H_t**:** The sequence of observations, actions, rewards

  $H_t = O_1, A_1, R_1, ..., O_{t-1}, A_{t-1}, R_{t-1}.$
- Which observation O_t the environment selects in time step t, depends on H_t.
- Which action A_t the agent selects in time step t, depends on H_t and O_t.
- Which reward R_t the environment selects in time step t, depends on H_t, O_t and A_t.

Ponto muito interessante: Rt pode ser determinado por Ht que é determinado por ações passadas, ou seja, ações passadas podem influenciar a recompensa la na frente

State

A state is the information used to determine what happens next



"summary of history"

- States are functions of the history.
- **Environment state S_t^e:** The (private) data used by the environment to pick the next observation/reward.
- The environment state is in many cases not visible to the agent or if so, it may contain irrelevant information.
- **The agent's state S_t^a:** The (internal) data used by the agent (i.e., its RL algorithm) to pick the next action.

State

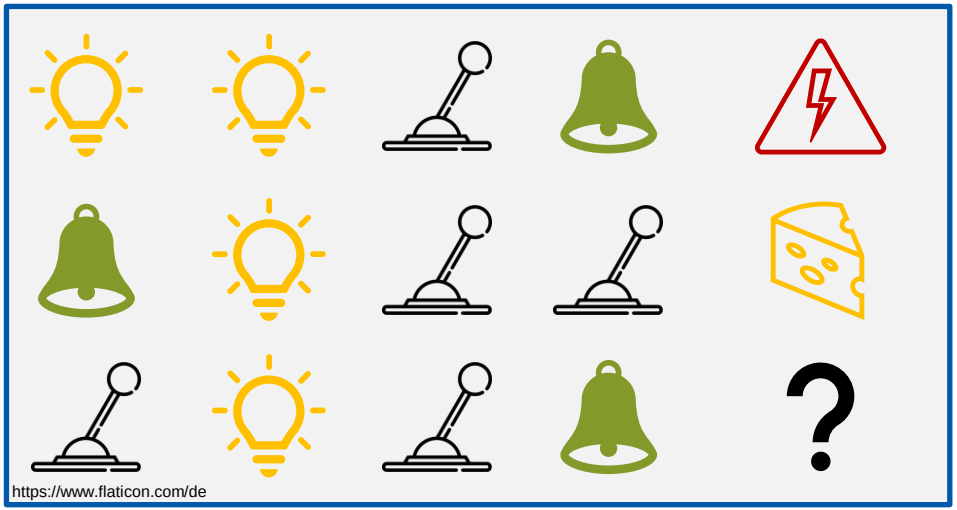
Agent‘s state can take a variety of forms

	form	example
State	High-level sensations	 Symbolic description of objects, e.g., object positions and orientation.
	Mid-level sensations	 Pre-processed sensor readings, e.g., after applying object detection.
	Low-level sensations	 Direct sensor readings, e.g. photo of bin.

Picture source: Lee, J.; Kang, S. and Park, S. “3D Pose Estimation of Bin Picking Object using Deep Learning and 3D Matching.” In *Proc. International Conference on Informatics in Control, Automation and Robotics,* 2018.

Question

Which reward to expect in round no. 3?



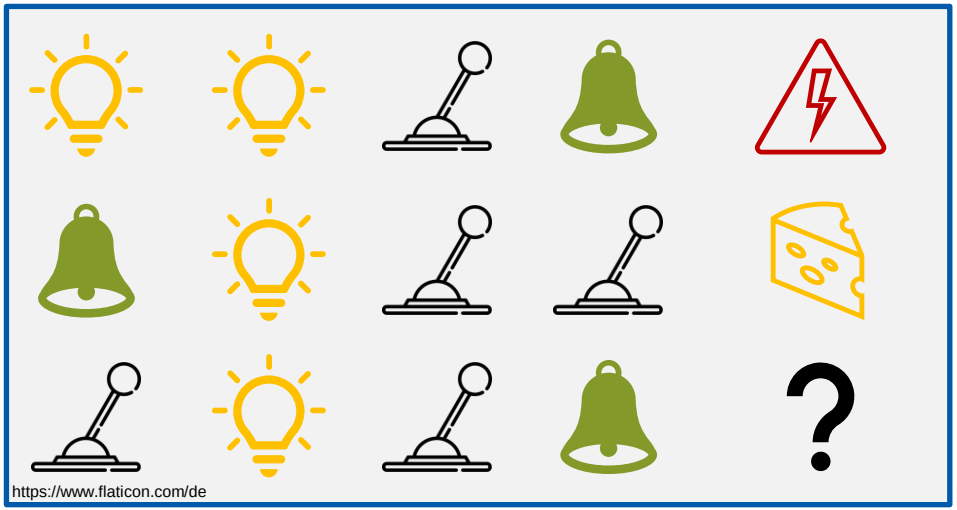
https://www.flaticon.com/de

o OTARIO do rato é ELETROCUTADO se ele errar

https://www.chrissanders.org

Answer

Prediction of expected reward depends on the choice of agent state



https://www.flaticon.com/de

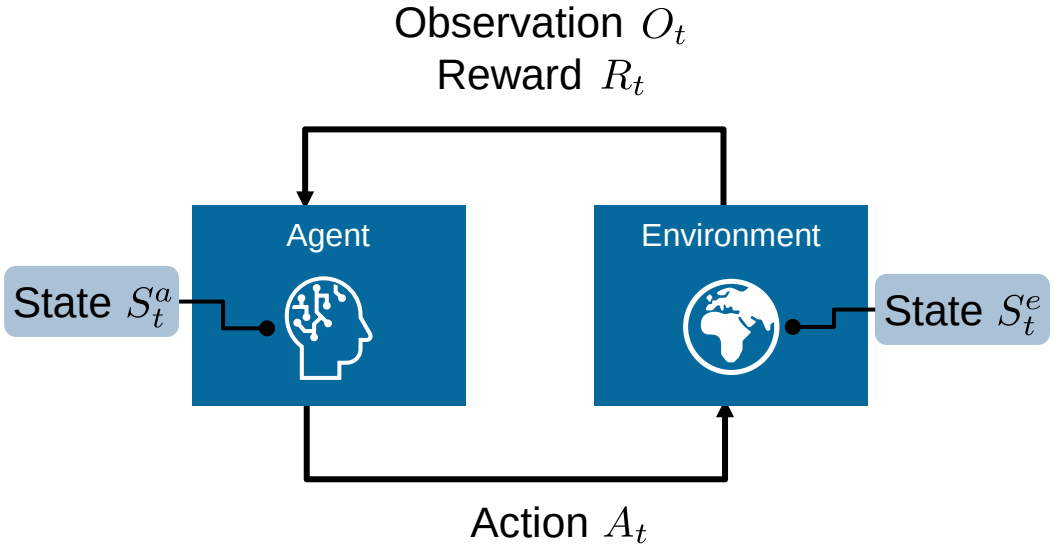
Which reward to expect in round 3 depends on choice of agent state, e.g.

- Last 3 items in sequence;
- Counts for lights, bells and levers;
- Complete sequence of items.

https://www.chrissanders.org

State

Under full observability, the agent directly observes the environment

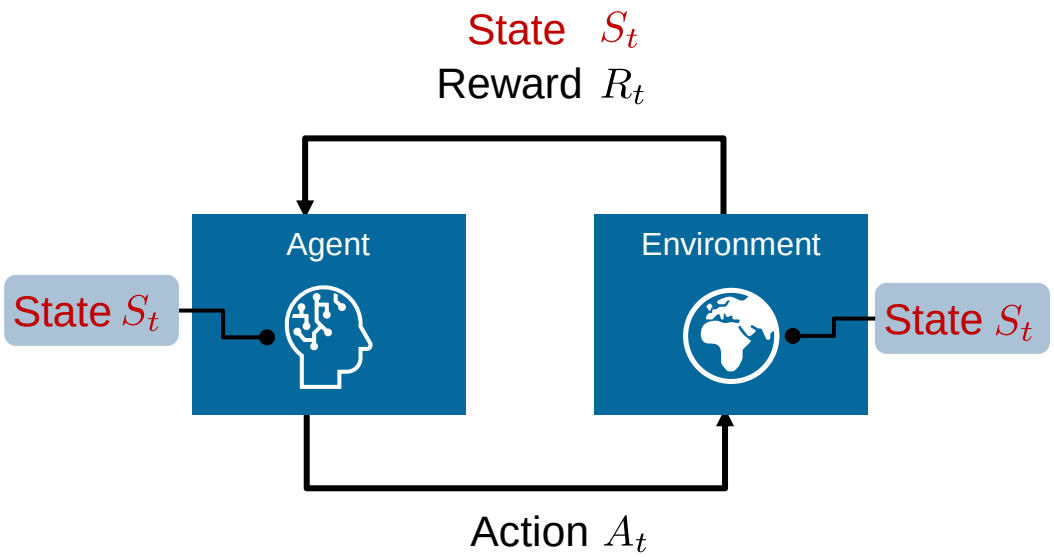


- **Full observability:** The agent directly observes the environment's state, i.e.,

$$O_t = S_t^a = S_t^e.$$

State

Under full observability, the agent directly observes the environment



- **Full observability:** The agent directly observes the environment's state, i.e.,

$$O_t = S_t^a = S_t^e.$$

- **Notation:** We denote this state by S_t.
- In this case, the state is a **Markov state**.
- The problem can be modelled by a **Markov Decision Process (MDP)**.

→ **Covered mostly in this course.**

State

A Markov state contains all useful information from the history

TECHNISCHE UNIVERSITÄT DARMSTADT

We are interested in states that contain all useful information from the history.

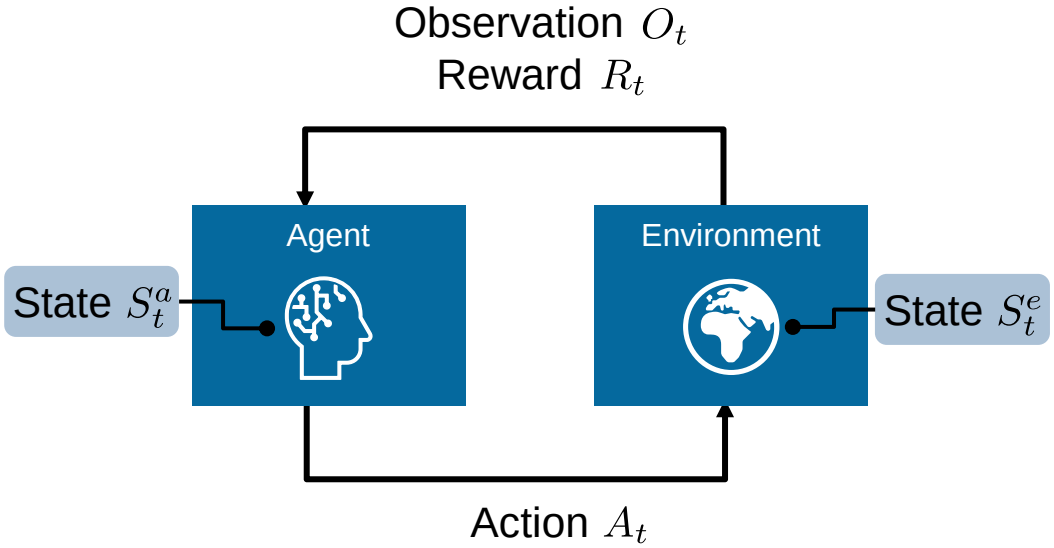
Definition (Markov State)

A state is **Markov** if it includes information about all aspects of the past agent-environment interaction that make a difference for the future.

- A Markov state is a sufficient statistic of the future.
- Once the Markov state is known, the history may be thrown away.
- The history H_t is Markov. embora isso n seja mt util, pq é mt pesado
- The environment state S_t^e is Markov. → Under full observability, the agent state is Markov!

State

Under partial observability, the agent indirectly observes the environment



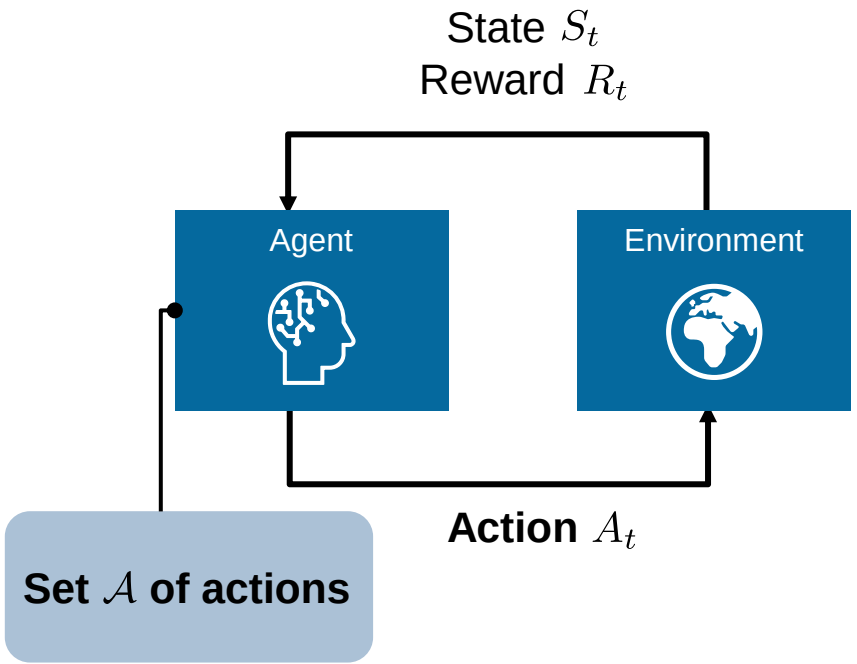
- **Partial observability:** The agent state and the environment state are not identical, i.e.
$$S_t^a \neq S_t^e.$$
- The agent must construct its own state representation S_t^a.
- The problem can be modelled by a partially observable Markov decision process (POMPD).

→ **Touched briefly at end of semester.**

por ex, um robo com uma camera. Ele não tem informação completa sobre o estado de envinroment

Action

Actions can be any decisions the agent wants to learn how to make

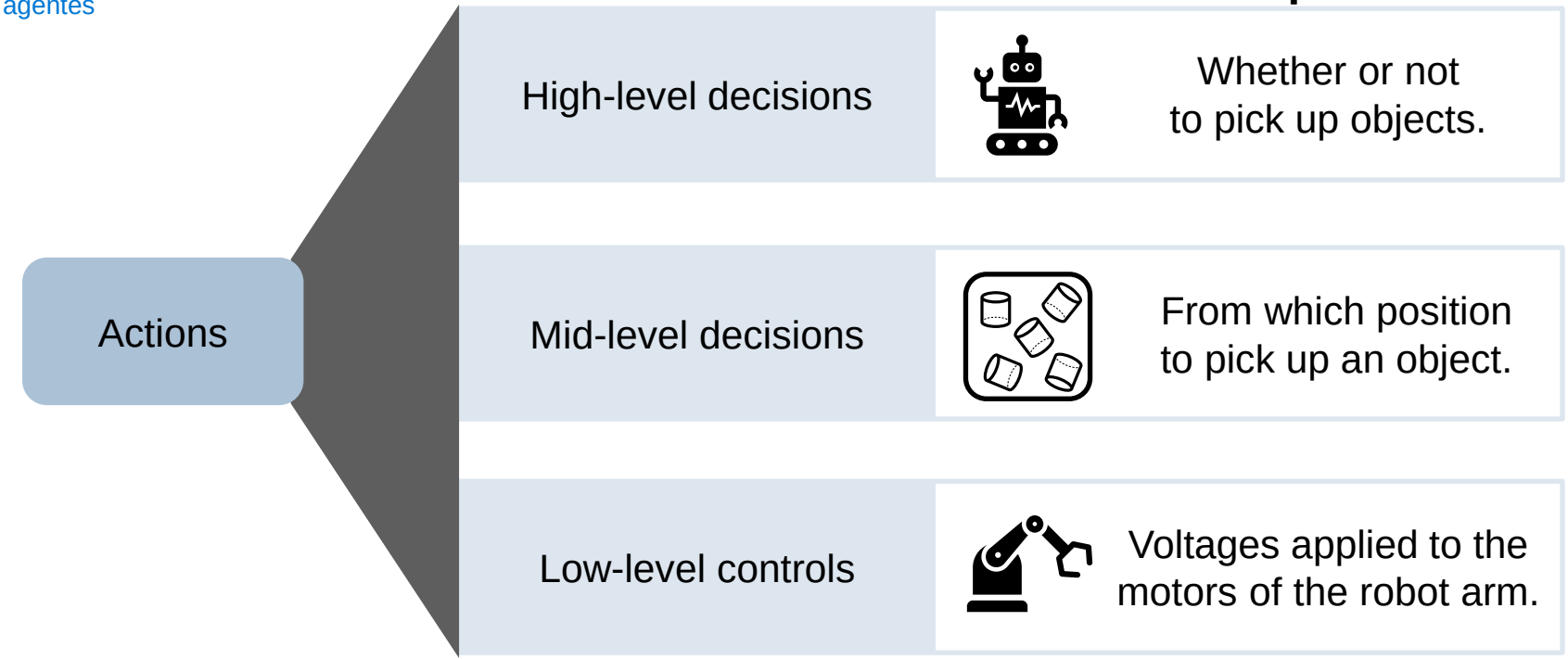


- Actions are the decisions the agent wants to learn how to make.
- **Set $\mathcal{A}$ of actions:** In the simplest case, the agent selects an action from the same set in each time step t, i.e.,
$$A_t \in \mathcal{A}.$$
- If the set of actions depends on the state, we write $\mathcal{A}(s)$ for the set of actions available in state s.

Action

Actions can take a variety of forms

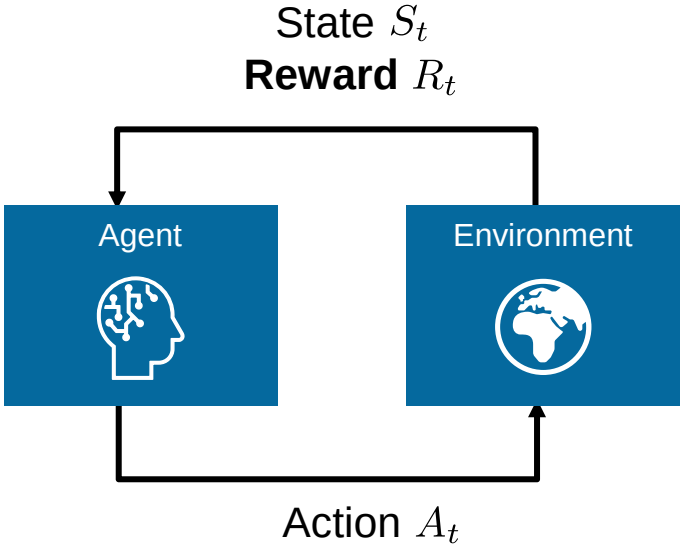
dentro de um mesmo robô podem haver diferntes agentes



Reward

A reward is a scalar feedback signals that defines the goal of RL

In general, Rt is a stochastic function in the State of the system



- Rewards indicate how well agent is doing in selecting actions.
- **Reward** R_t: Scalar feedback signal received by the agent in part as a consequence of its action in time slot t.
- **Agent‘s goal**: Select actions to maximize the cumulative reward.

soma

- RL is based on the reward hypothesis.

Definition (Reward Hypothesis)

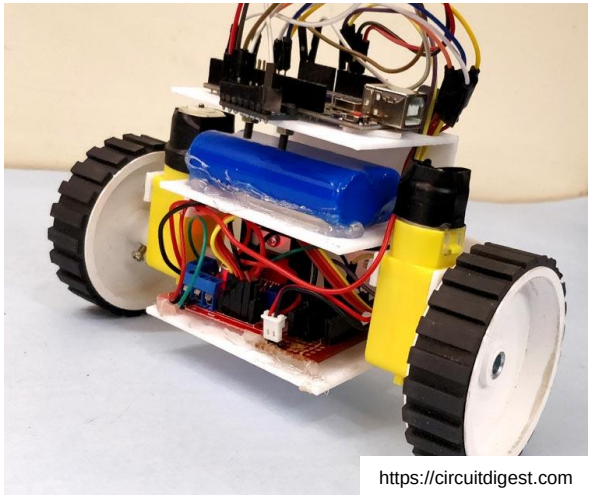
All goals can be described by the maximization of expected cumulative reward.

Reward

Examples of reward

maximizar essa reward é aumentar a prob de chegar a mensagem no destino

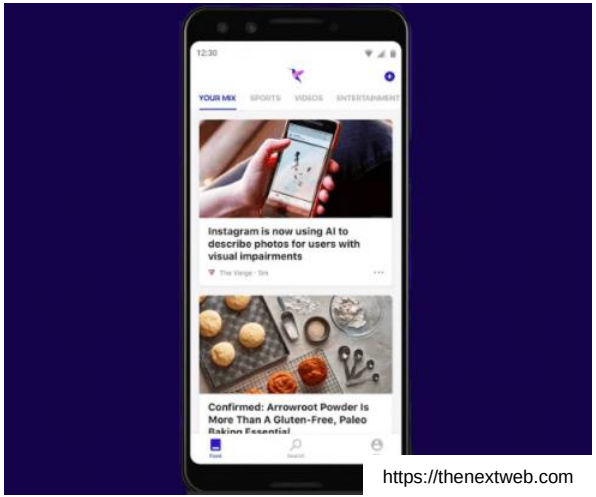
TECHNISCHE UNIVERSITÄT DARMSTADT



https://circuitdigest.com

Build a self-balancing robot

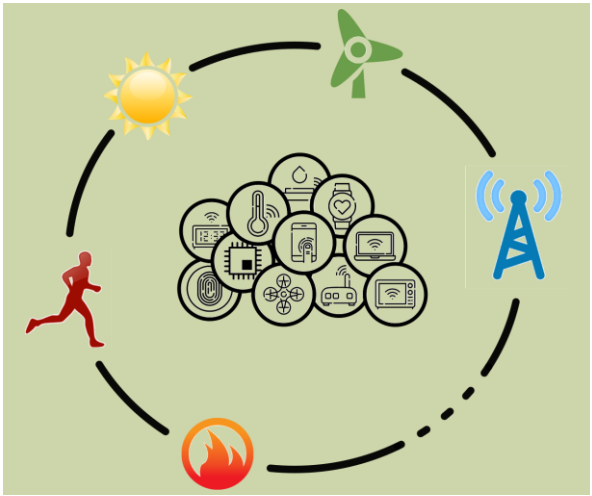
Possible reward:
-1 on each failure;
0 otherwise.



https://thenextweb.com

Personalize a news feed

Possible reward:
+1 if user clicks displayed news;
0 otherwise.

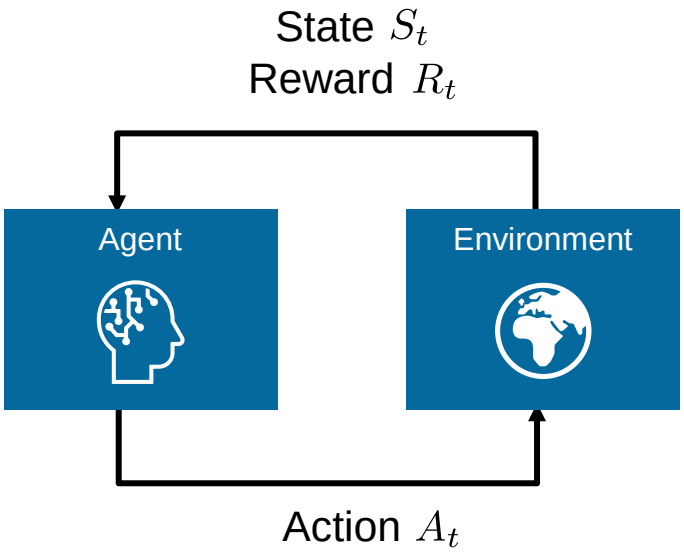


Optimize an energy harvesting communication system

Possible reward:
Throughput achieved during one interval of data transmission.

Reward

Achieving agent‘s goal is challenging due to characteristics of RL



- **Agent‘s goal**: Select actions to maximize the cumulative reward.

normalmente não tem o conhecimento da dinamica do estado/recomepensa a prior

- **Challenges**

não consegue prever

  - State and reward dynamics are unknown to the agent.
  - Actions may have long term consequences.
  - Reward may be delayed.
  - Sacrificing immediate reward may lead to more long-term reward.

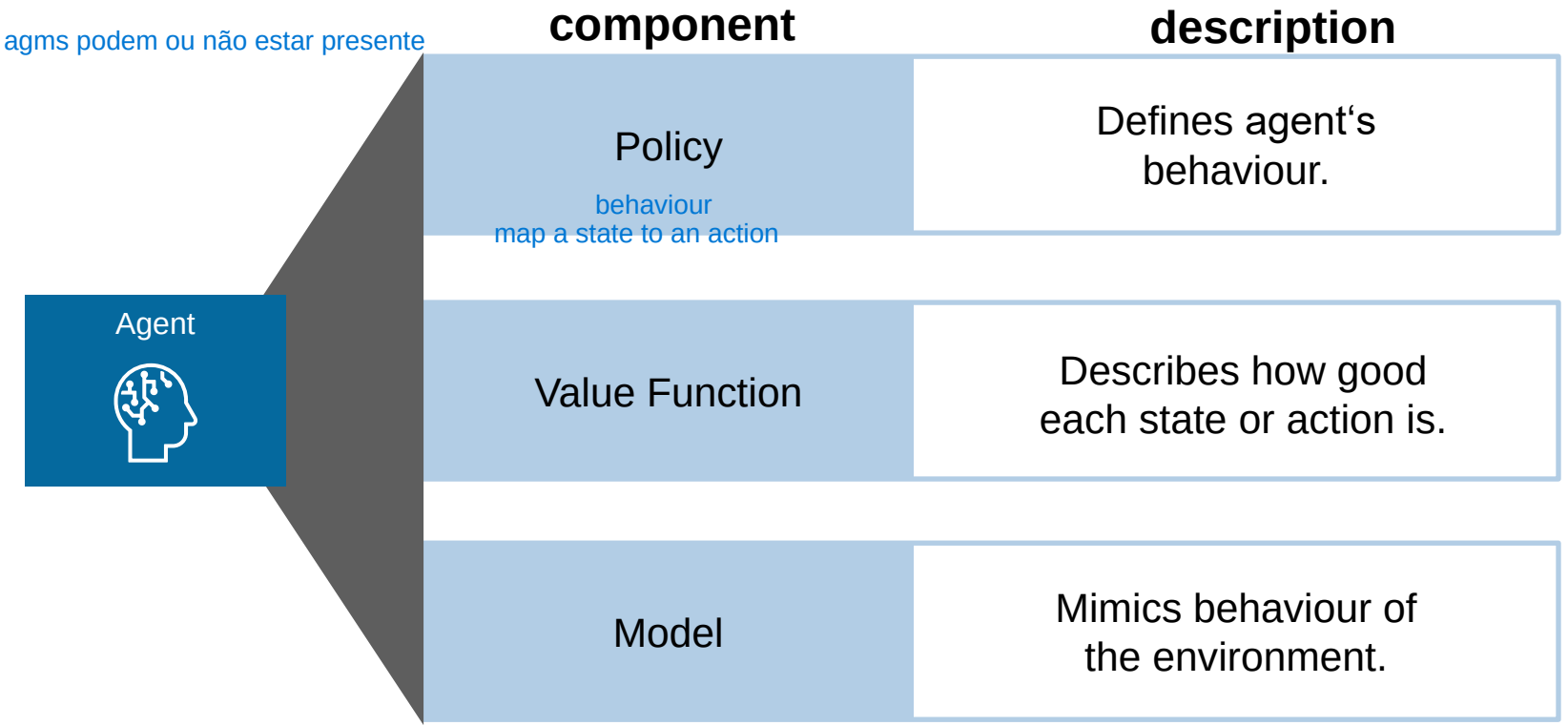
Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- Characteristics of RL
- **Components of RL Agents**
- Problems within RL
- Lecture Overview

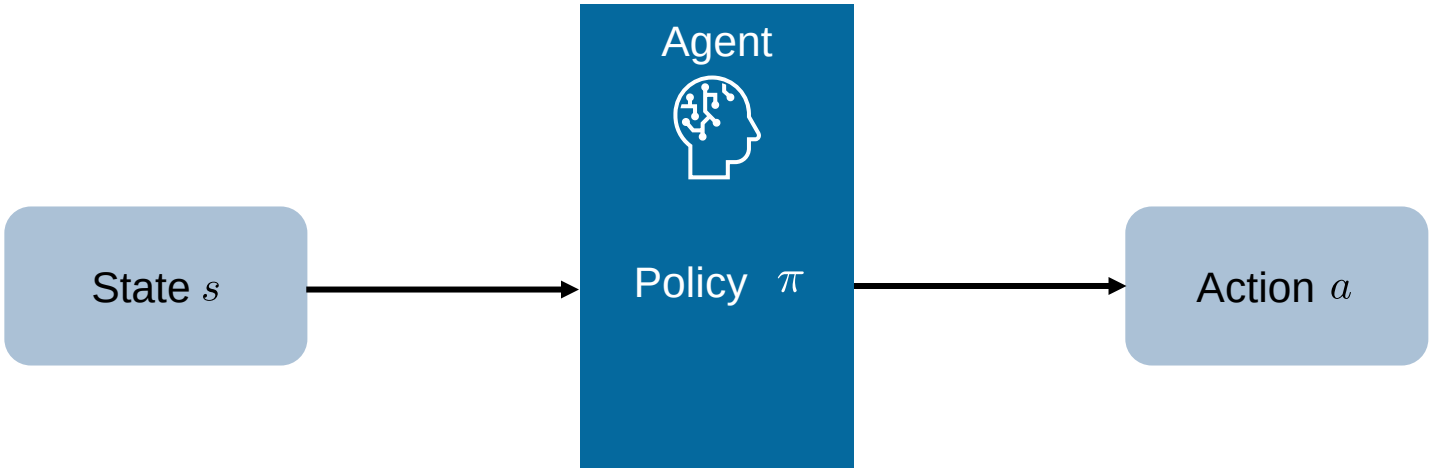
Components of RL agents

An RL agent may include one or more of the following major components



Policy

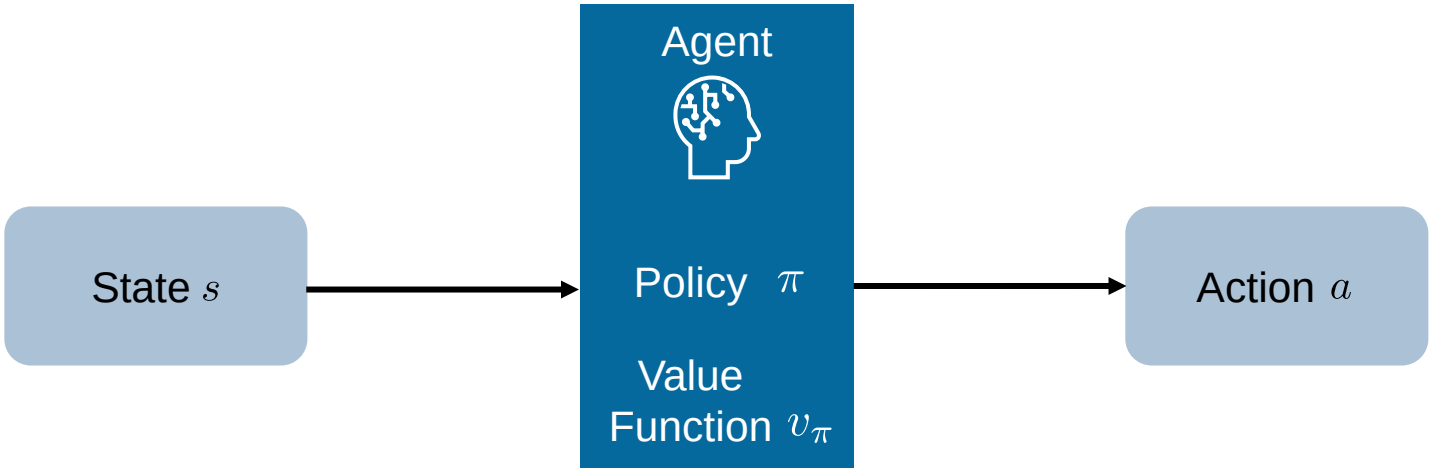
A policy determines the agent‘s behavior



- A **policy** determines the agent‘s way of behaving at a given time.
- It is a mapping from state to action.
- A policy can be deterministic (i.e., $\pi(s) = a$) or stochastic (i.e., $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$).

Value Function

A value function is used to evaluate the goodness of states

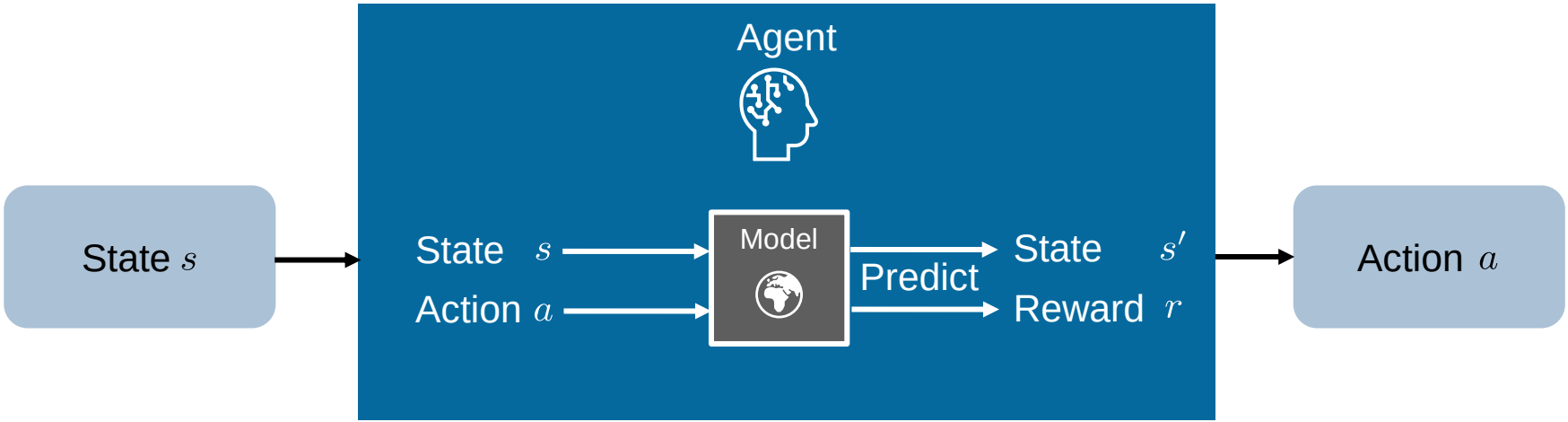


- A **value function** predicts the future reward of a state in the long run.
- It can be used to evaluate how good a state is (e.g., $v_\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s]$).
- Desirable actions are those that lead to states of highest values.

Model

vamos focar em model-free methods

A model predicts what the environment will do next

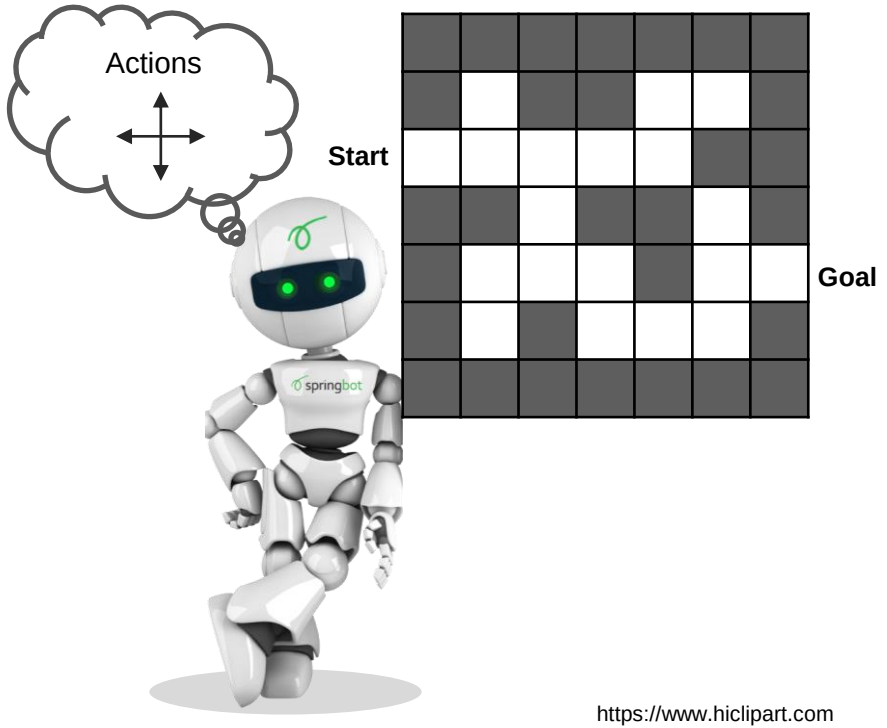


- A **model** mimics the behavior of the environment.
- It can be used to predict the next state (e.g., $p(s'|s,a) := \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$).
- It can be used to predict the next reward (e.g., $r(s,a) := \mathbb{E}[R_t|S_t = s, A_t = a]$).

Components of RL agents

Example: How to make a robot solve a maze as quickly as possible?

- **States:** Agent's location
- One terminal state is the goal.
- **Actions:** N, E, S, W
- Actions out of the grid do not have any effect.
- **Reward:** $r = -1$ per time-step until terminal state is reached.

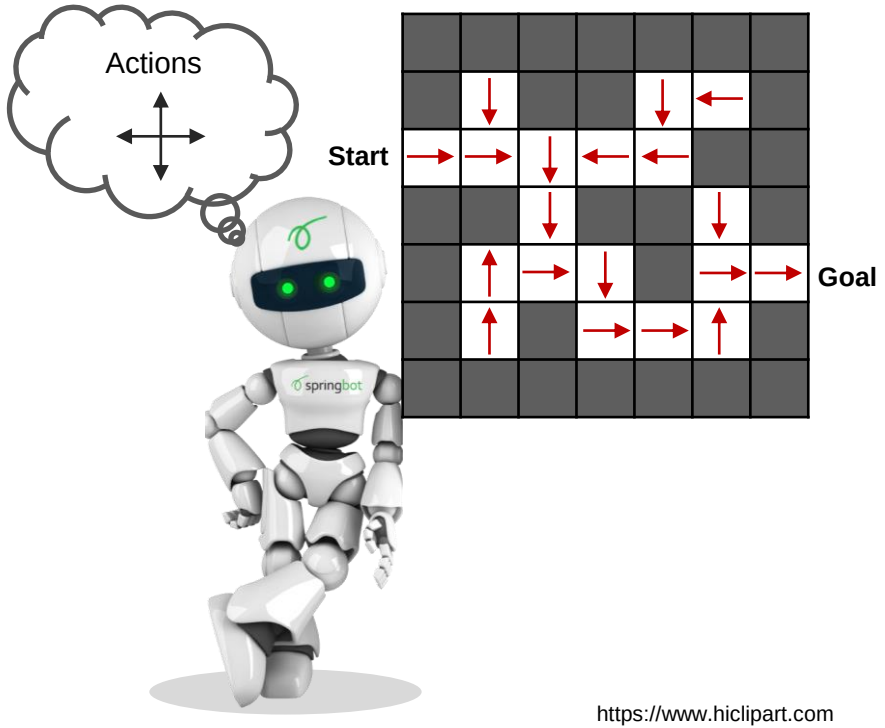


https://www.hiclipart.com

Components of RL agents

Example: The agent‘s policy

- **Policy:** Red arrows represent an exemplary determinstic policy.
- Policy determines the agent‘s behaviour in each state.

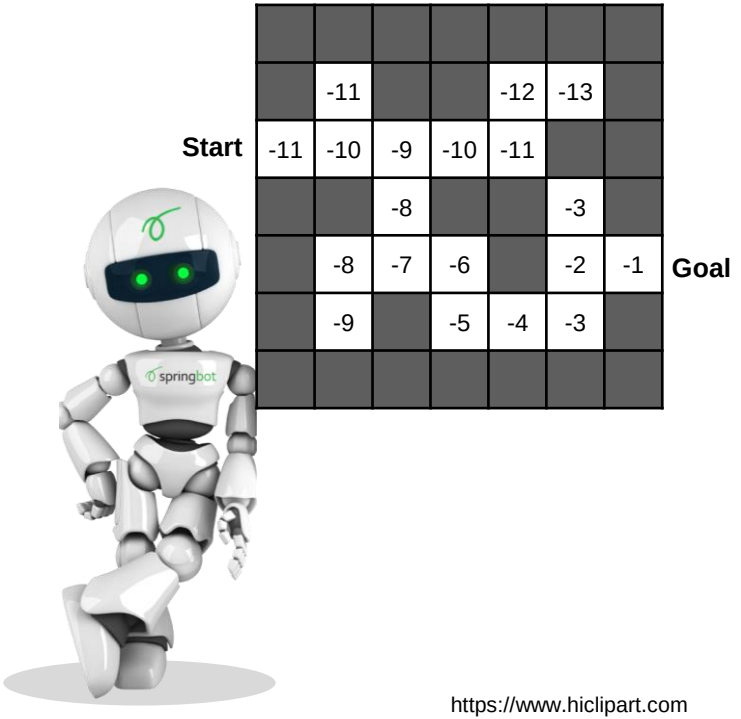


https://www.hiclipart.com

Components of RL agents

Example: The agent‘s value function

- **Value function:** Numbers represent the value function $v_\pi(s)$ for each state s.
- The values give an idea of optimal behaviour.

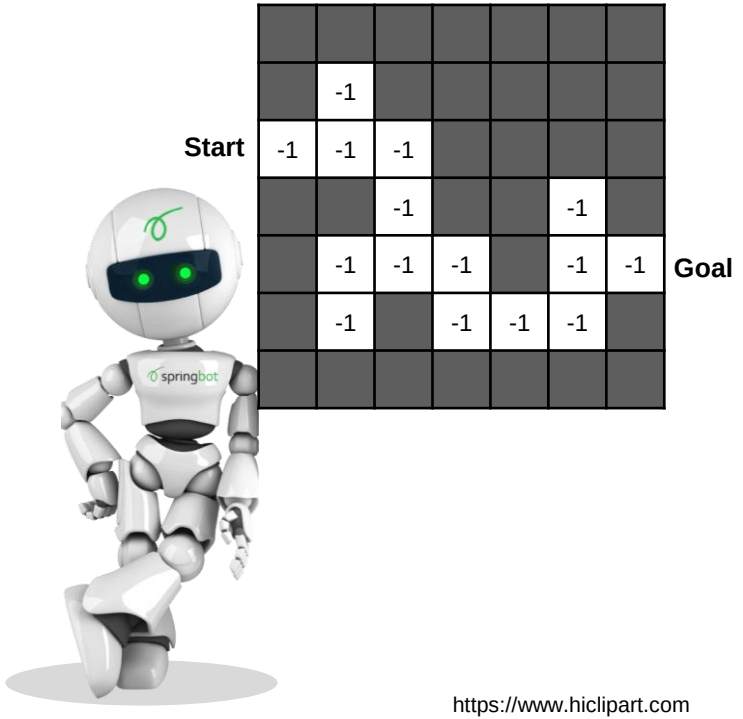


https://www.hiclipart.com

Components of RL agents

Example: The agent‘s model of reality

- **Model:** Agent may have a (possibly imperfect) internal model of the environment.
- Shows what agent has understood of the environment so far regarding
  - Dynamics: How actions change the state.
  - Rewards: How much reward from each state.
- Grid layout represents transition model $p(s'|s,a)$.
- Numbers represent immediate reward $r(s,a)$ from each state s.



https://www.hiclipart.com

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- Characteristics of RL
- Components of RL Agents
- **Problems within RL** CHALLENGES
- Lecture Overview

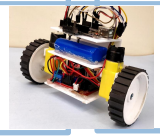
Learning and Planning

Two types of sequential decision making problems are learning and planning

TECHNISCHE UNIVERSITÄT DARMSTADT

Sequential Decision Making

não temos modelo do ambiente

	Reinforcement Learning	Planning
Goal	Improve policy, when environment initially unknown.	Improve policy, when model of environment is known.
Approach	Agent-environment interaction.	Agent performs computations with its model.
Example	Robot learns self-balancing directly from trial-and-error balancing. 	Robot gets perfect model of its movements and its environment; can plan ahead to find optimal self-balancing policy.

Exploration and Exploitation

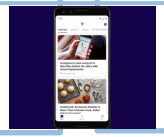
A challenge in RL is how to balance exploration and exploitation

TECHNISCHE UNIVERSITÄT DARMSTADT

Reinforcement Learning

Balance

	Exploration giveup what we already now	**Exploitation** exploração do conhecido
Goal	Gather more information to achieve high reward in future.	Exploit current information to achieve high reward now.
Approach	Select undiscovered actions.	Select best previous actions.
Example	Show new article in news feed.	Show the so far most clicked article in news feed.



Prediction (or Evaluation) and Control

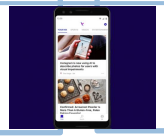
Solving an RL problem requires to solve two types of sub-problems

TECHNISCHE UNIVERSITÄT DARMSTADT

Reinforcement Learning

(precisa ter um evaluation mais ou menos bom)

	Prediction (or Evaluation)	Control
Goal	Evaluate the future given a policy.	Optimize the future.
Approach	Determine value function for given policy.	Find the best policy.
Example	Determine expected no. of clicks of “uniform random news display“ policy.	Determine the best news display policy which maximizes the expected no. of clicks.



Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

- You can describe the characteristics and main elements of Reinforcement Learning and identify examples of Reinforcement Learning tasks.
  - → Goal-directed learning from interaction; agent, environment, state, action, reward.
- You can explain the main components of Reinforcement Learning agents.
  - → Policy, value function, model.
- You can explain the main problems within Reinforcement Learning.
  - → Learning and planning; exploration and exploitation; prediction/evaluation and control.

Outline

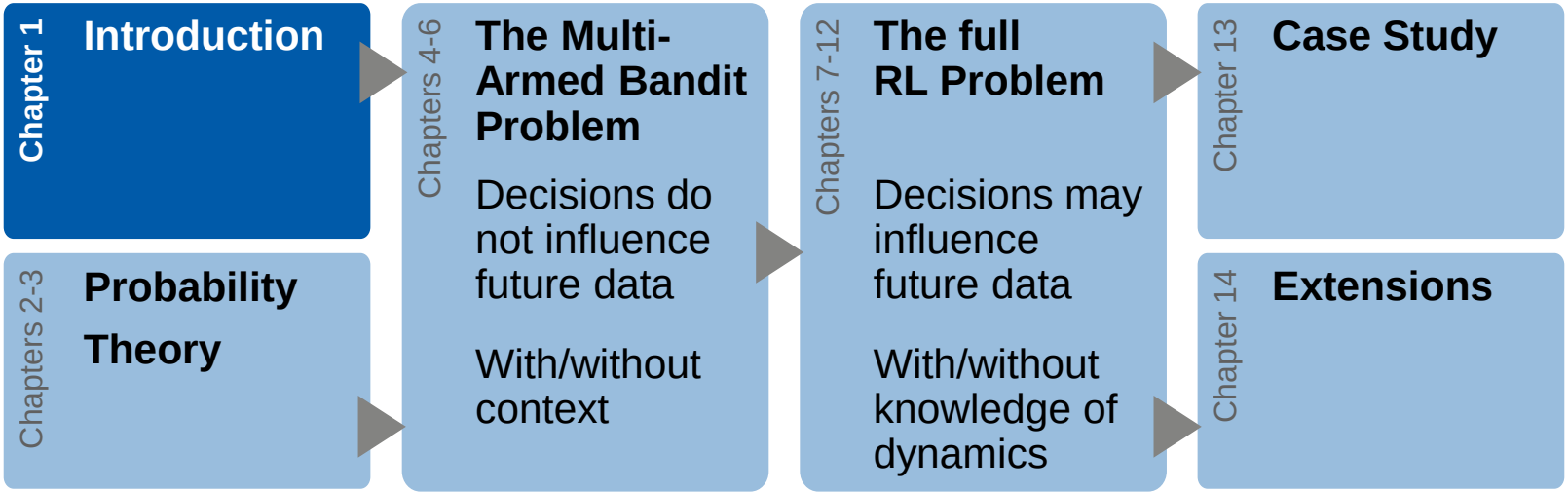
TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- Characteristics of RL
- Components of RL Agents
- Problems within RL
- **Lecture Overview**

Lecture Overview

We study the main methods from RL and apply them to engineering problems

TECHNISCHE UNIVERSITÄT DARMSTADT



Outro

Yet another example of RL…☺

Positive Reinforcement - The Big Bang Theory

https://www.youtube.com/watch?v=JA96Fba-WHk