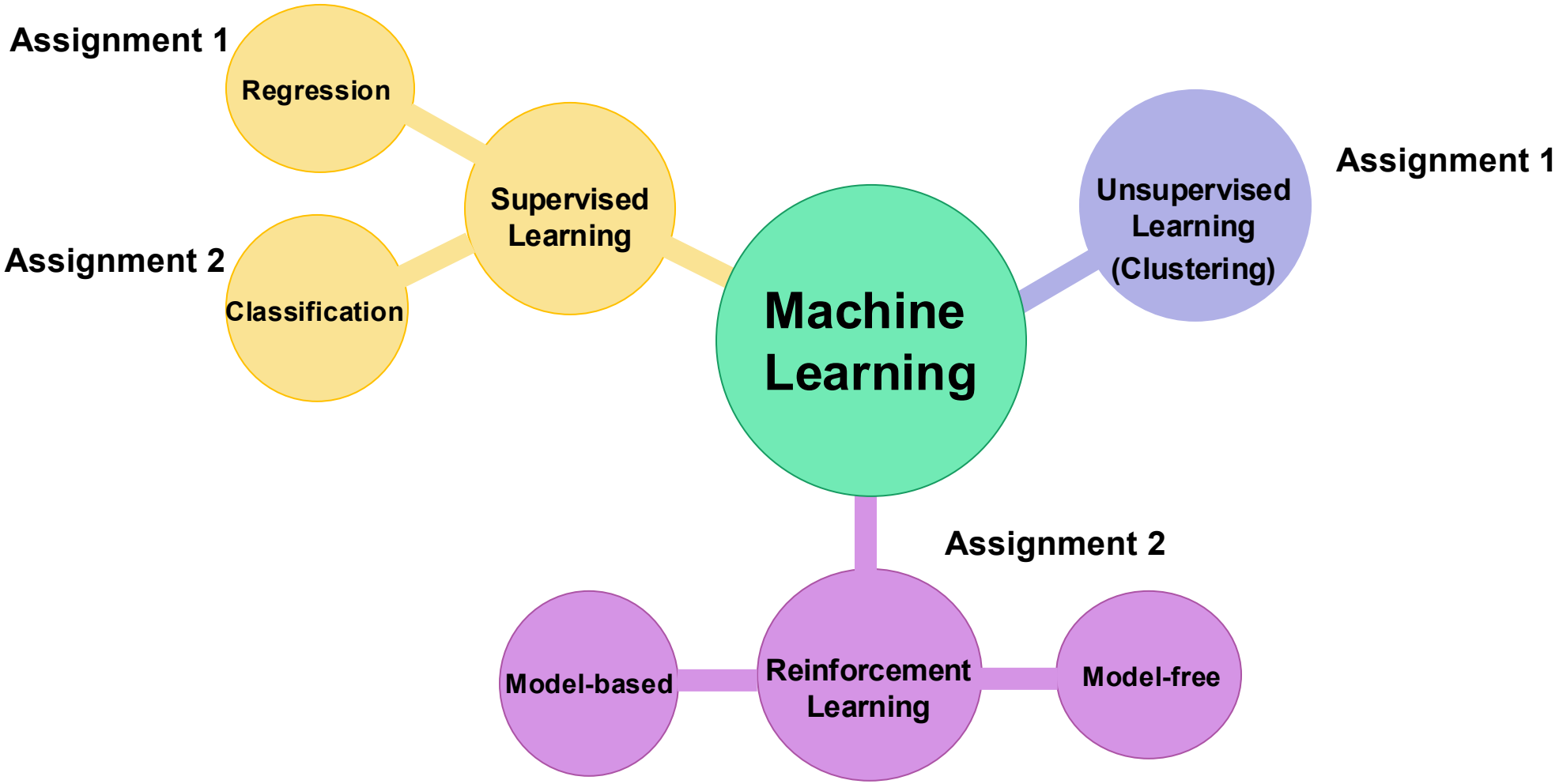
46765 – Machine learning for energy systems

Reinforcement learning for real-time control of an asset

Farzaneh Pourahmadi

Recap of the course!

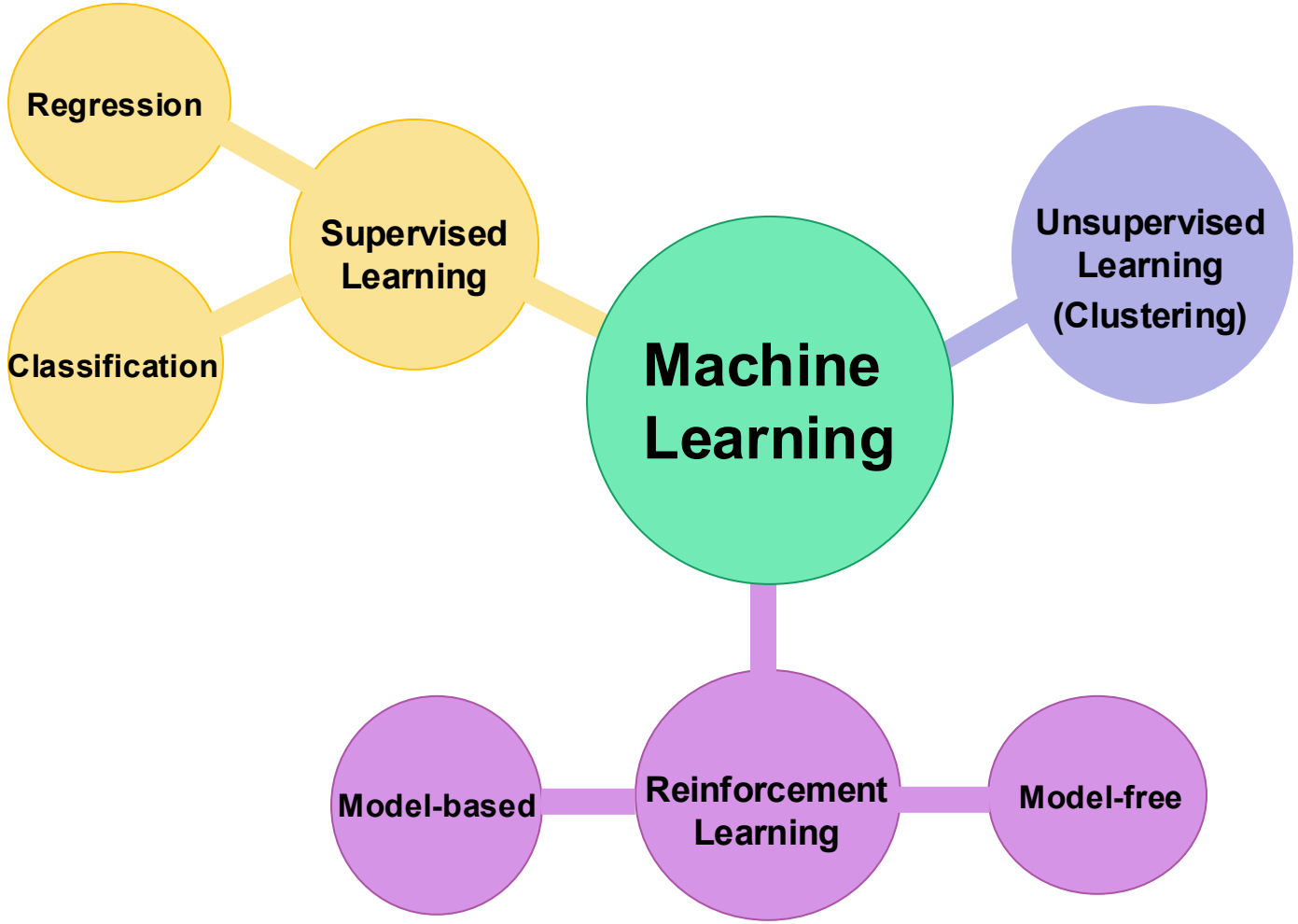


Recap of the course so far!

Wind power prediction

Market price prediction

Load prediction



Recap of the course!

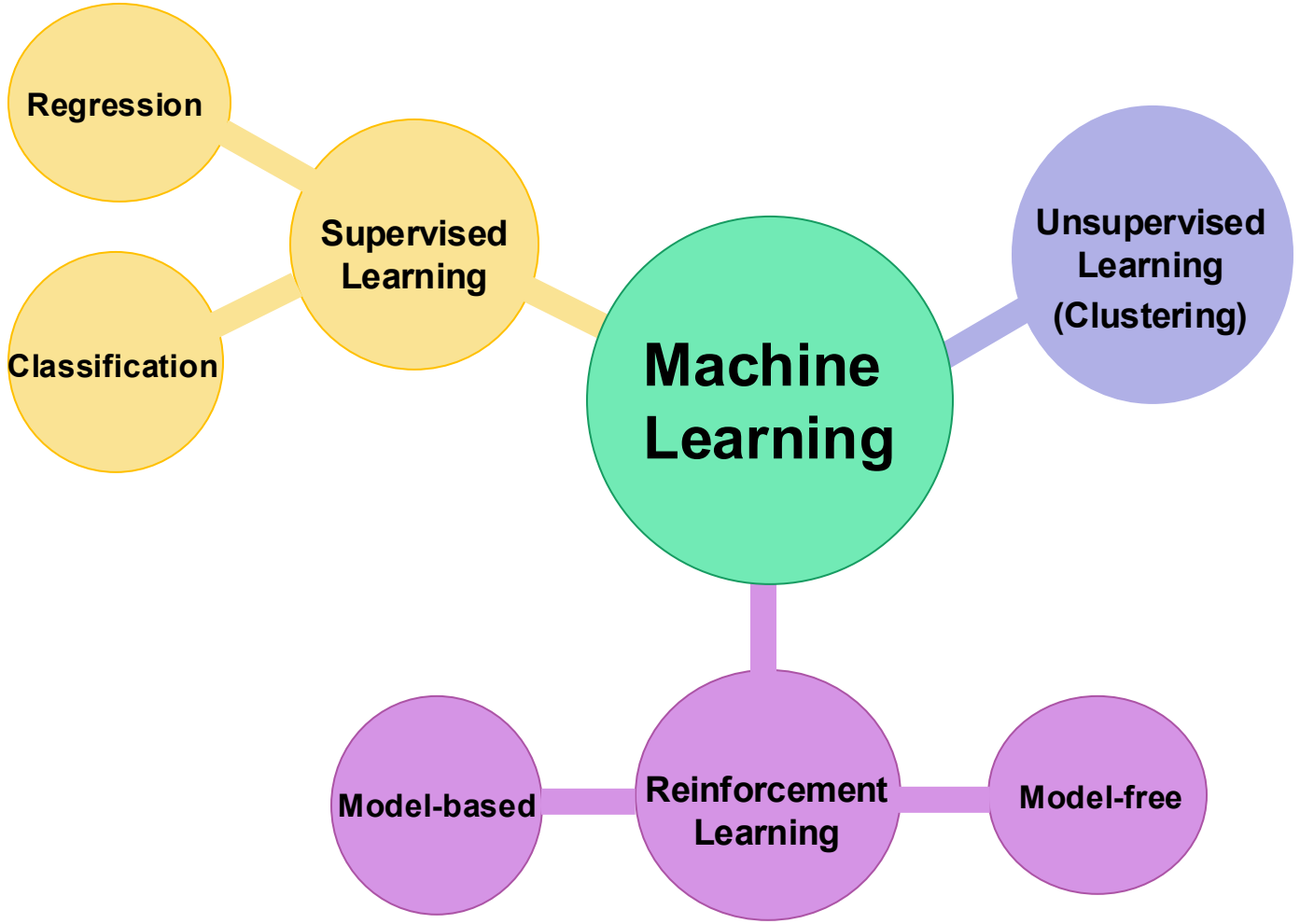
Wind power prediction

Market price prediction

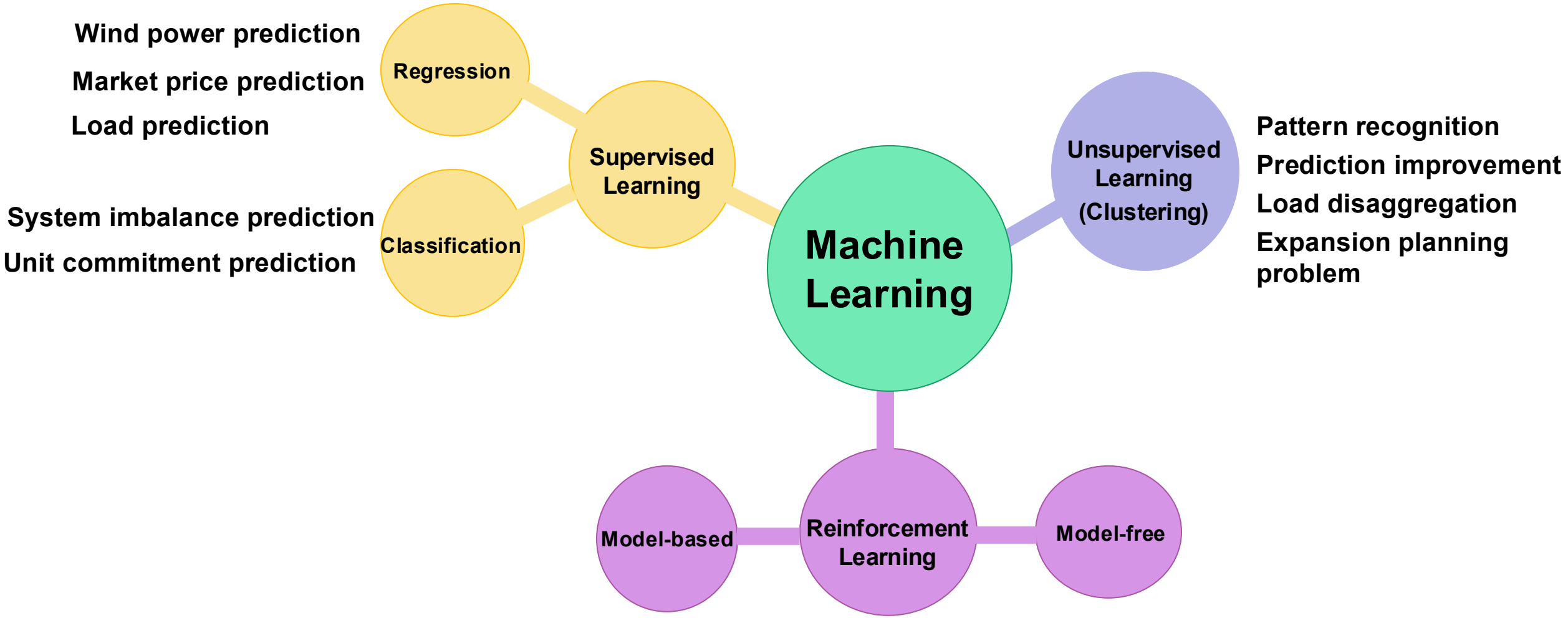
Load prediction

System imbalance prediction

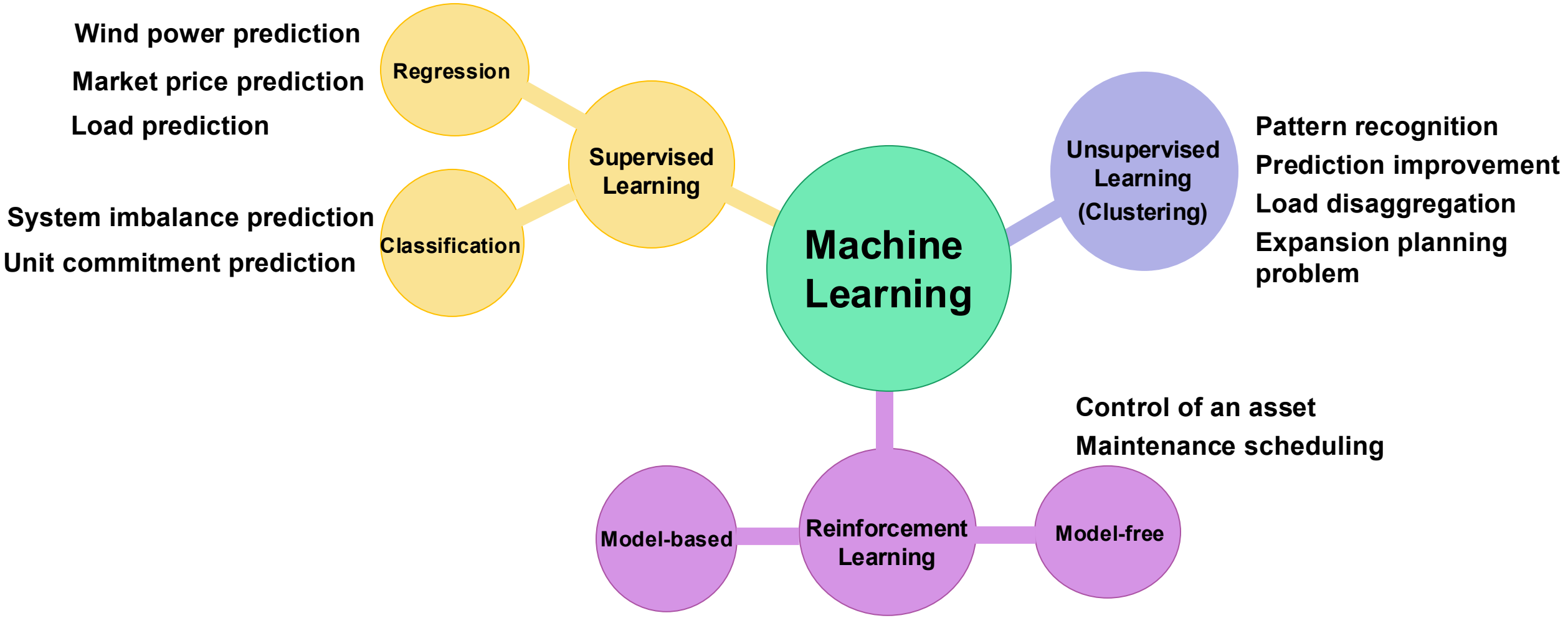
Unit commitment prediction



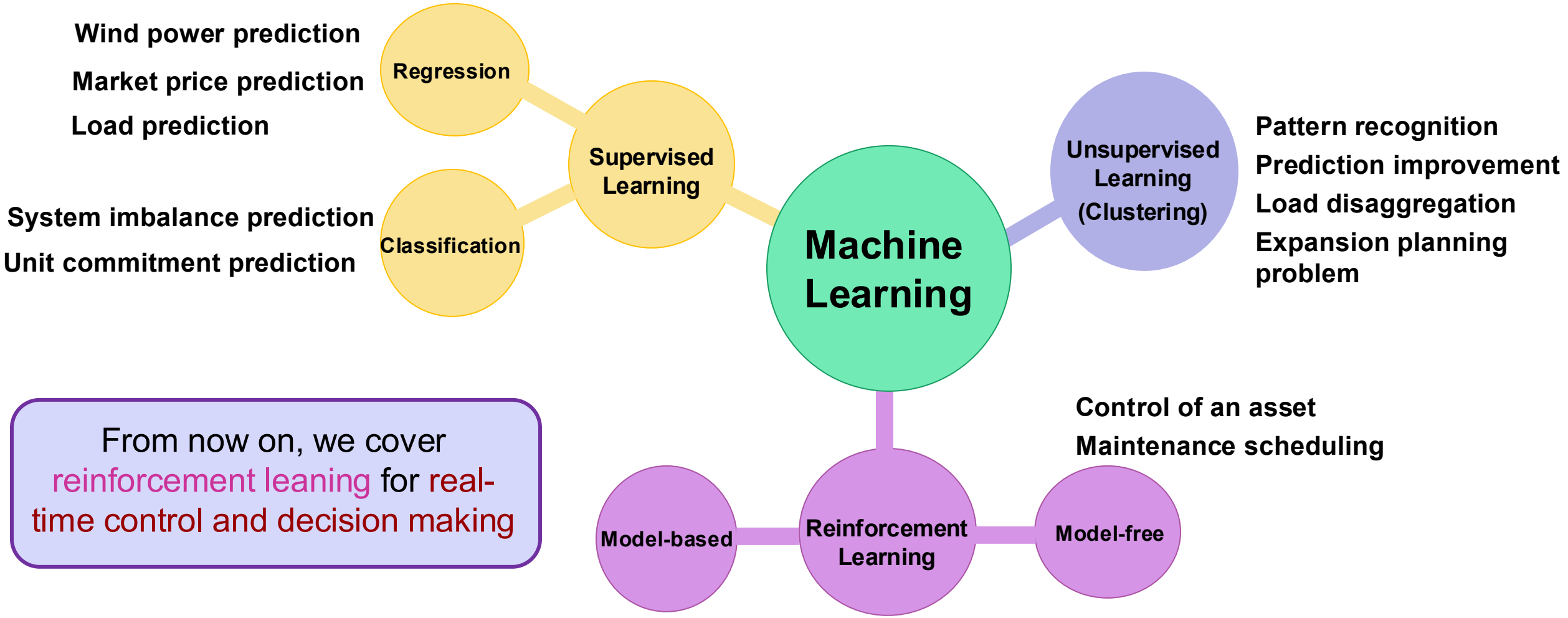
Recap of the course so far!



Recap of the course so far!



Recap of the course so far!



From now on, we cover reinforcement leaning for real-time control and decision making

Learning objectives

Through this lecture, it is aimed for you to be able to

- Explain when RL is preferable to supervised/unsupervised learning in energy systems
- Figure out how to describe agent and environment for different energy applications
- Describe the fundamentals of Markov decision process

Discussion:

Reinforcement learning

supervised and unsupervised learning

Reinforcement learning vs. supervised and unsupervised learning

Data Type:

- **Supervised Learning:** Requires a labeled dataset with paired input-output examples for training.
- **Unsupervised Learning:** Operates on unlabeled data, focusing on finding intrinsic structures or patterns.
- **Reinforcement Learning:** We do not necessarily have a dataset in advance. By interacting with an environment and receiving feedback in the form of rewards or penalties, we learn the optimal actions.

Temporal Aspect:

- **Supervised and Unsupervised Learning:** Typically involve a static dataset where each data point is independent of others.
- **Reinforcement Learning:** Inherently involves a dynamic dataset (a temporal aspect), as the agent's actions influence subsequent states and rewards.

Reinforcement learning vs. supervised and unsupervised learning

Data Type:

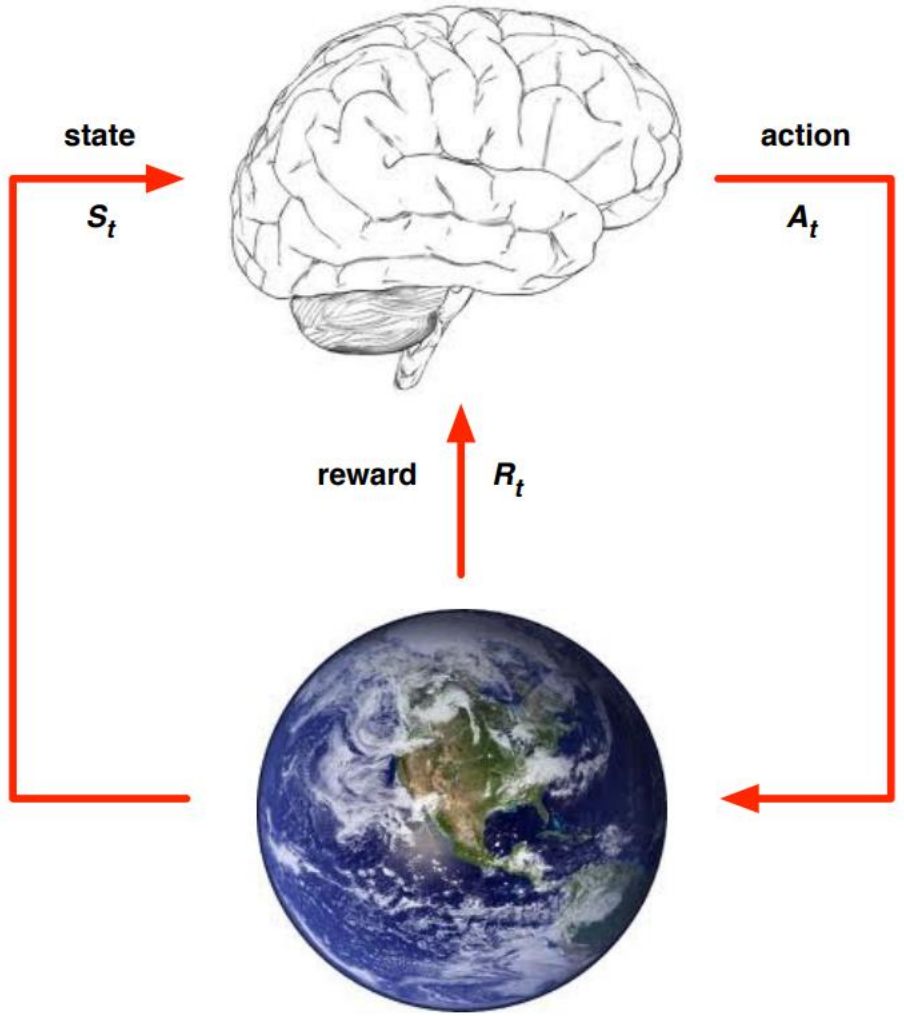
- **Supervised Learning:** Requires a labeled dataset with paired input-output examples for training.
- **Unsupervised Learning:** Operates on unlabeled data, focusing on finding intrinsic structures or patterns.
- **Reinforcement Learning:** We do not necessarily have a dataset in advance. By interacting with an environment and receiving feedback in the form of rewards or penalties, we learn the optimal actions.

Temporal Aspect:

- **Supervised and Unsupervised Learning:** Typically involve a static dataset where each data point is independent of others.
- **Reinforcement Learning:** Inherently involves a dynamic dataset (a temporal aspect), as **the agent's actions influence subsequent states and rewards**.

Reinforcement learning: Agent and environment

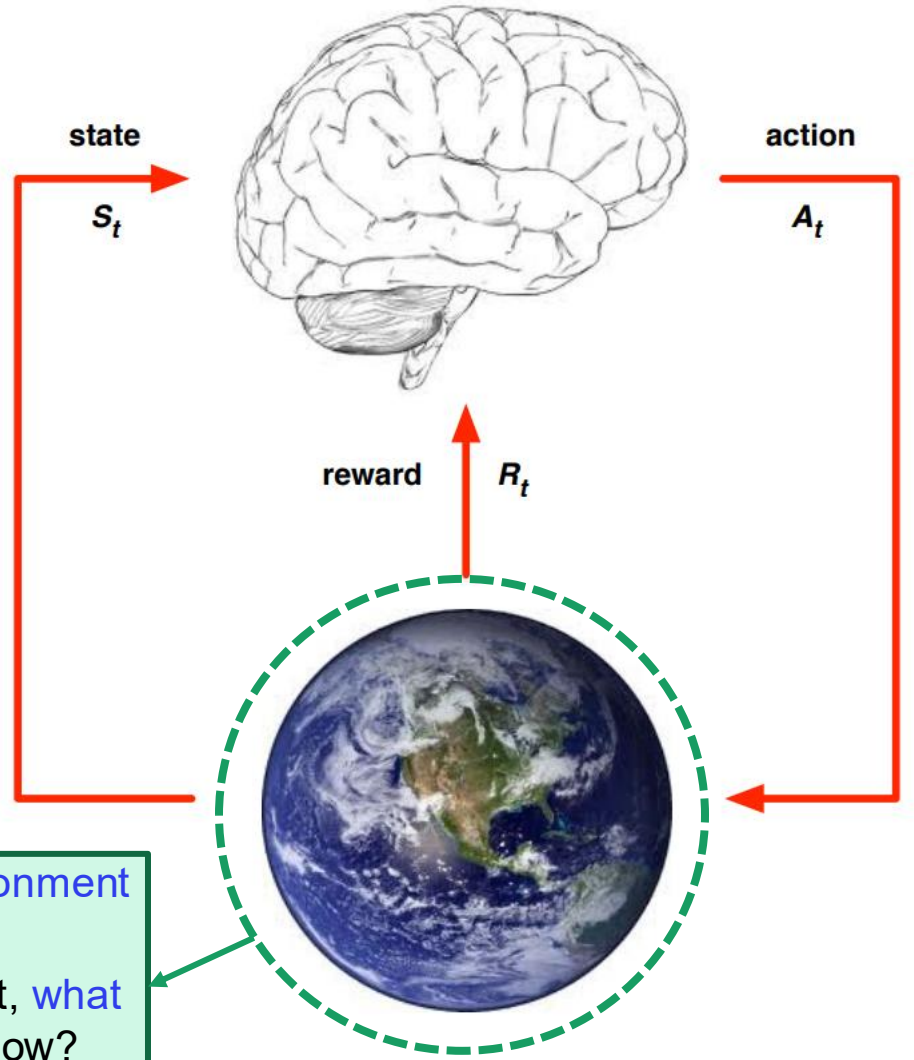
- At each step *t* the agent:
  - Executes action A_t
  - Receives observation S_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation S_{t+1}
  - Emits scalar reward R_{t+1}
- *t* increments at environment step



The fundamental RL loop

Reinforcement learning: Agent and environment

- At each step *t* the agent:
  - Executes action A_t
  - Receives observation S_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation S_{t+1}
  - Emits scalar reward R_{t+1}
- *t* increments at environment step



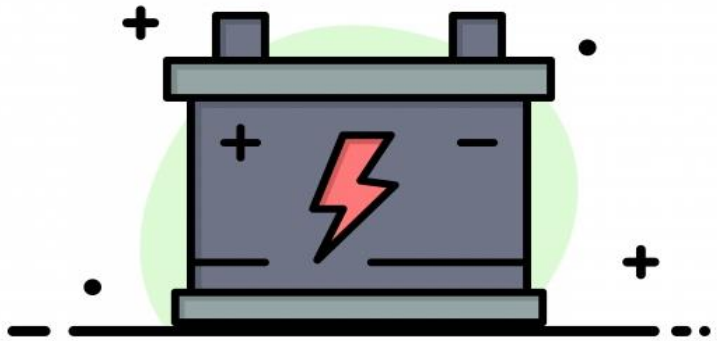
- How can we model the environment to find the optimal action?
- To represent the environment, what information do we need to know?


example

Reinforcement learning: Battery example

Problem description:

Imagine you are the owner of a battery storage system, and you want to participate in the electricity markets to generate revenue while also contributing to the stability of the power grid. The goal is to use reinforcement learning to develop an intelligent control strategy that optimizes the operation of your battery storage system.

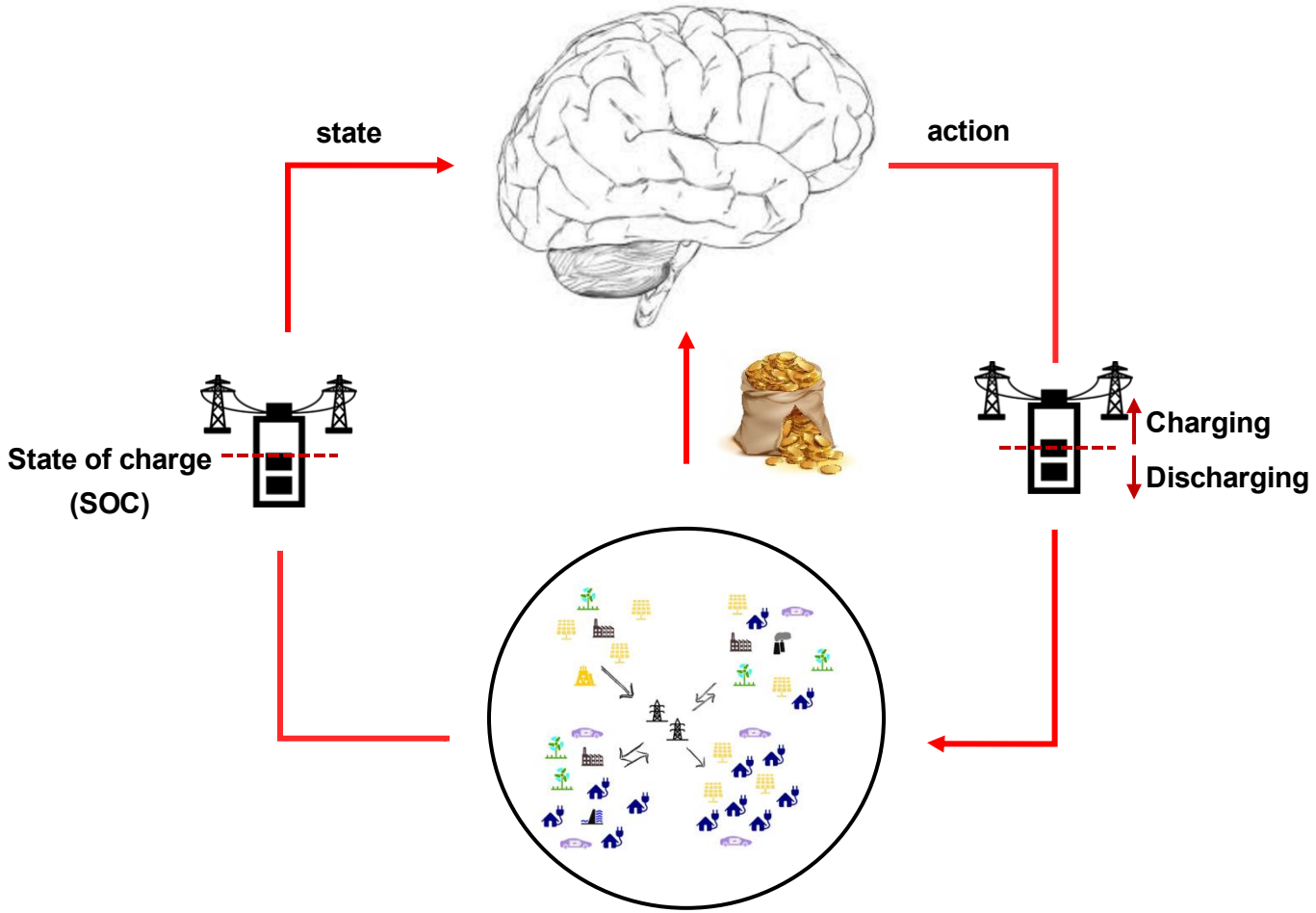


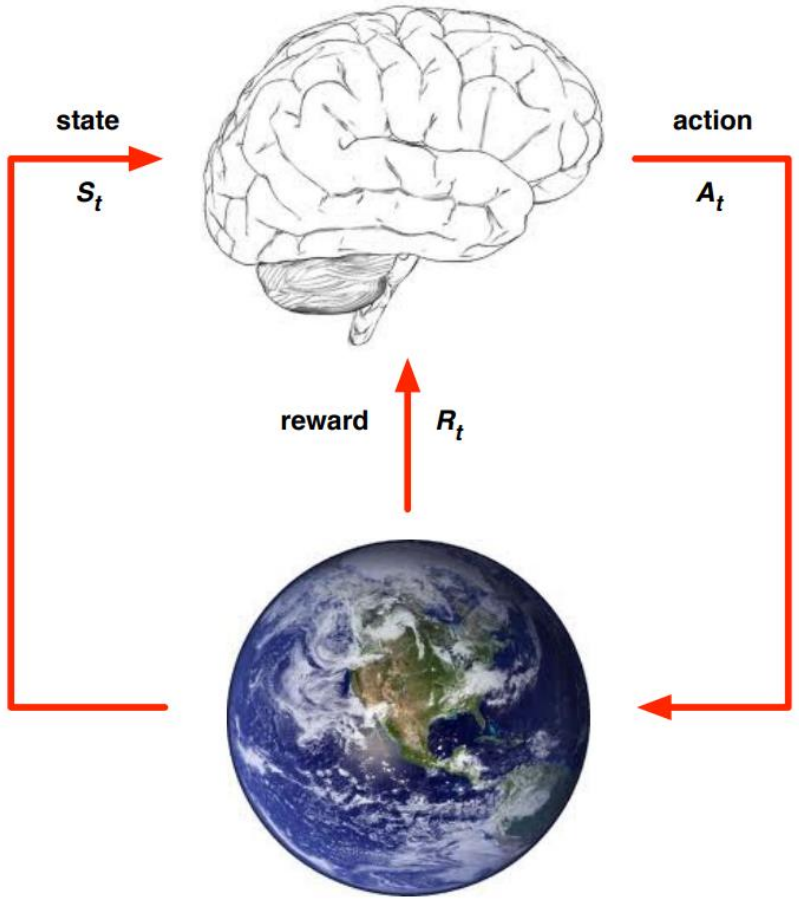
Reinforcement learning: Battery example

How can we model this case as a reinforcement learning problem?

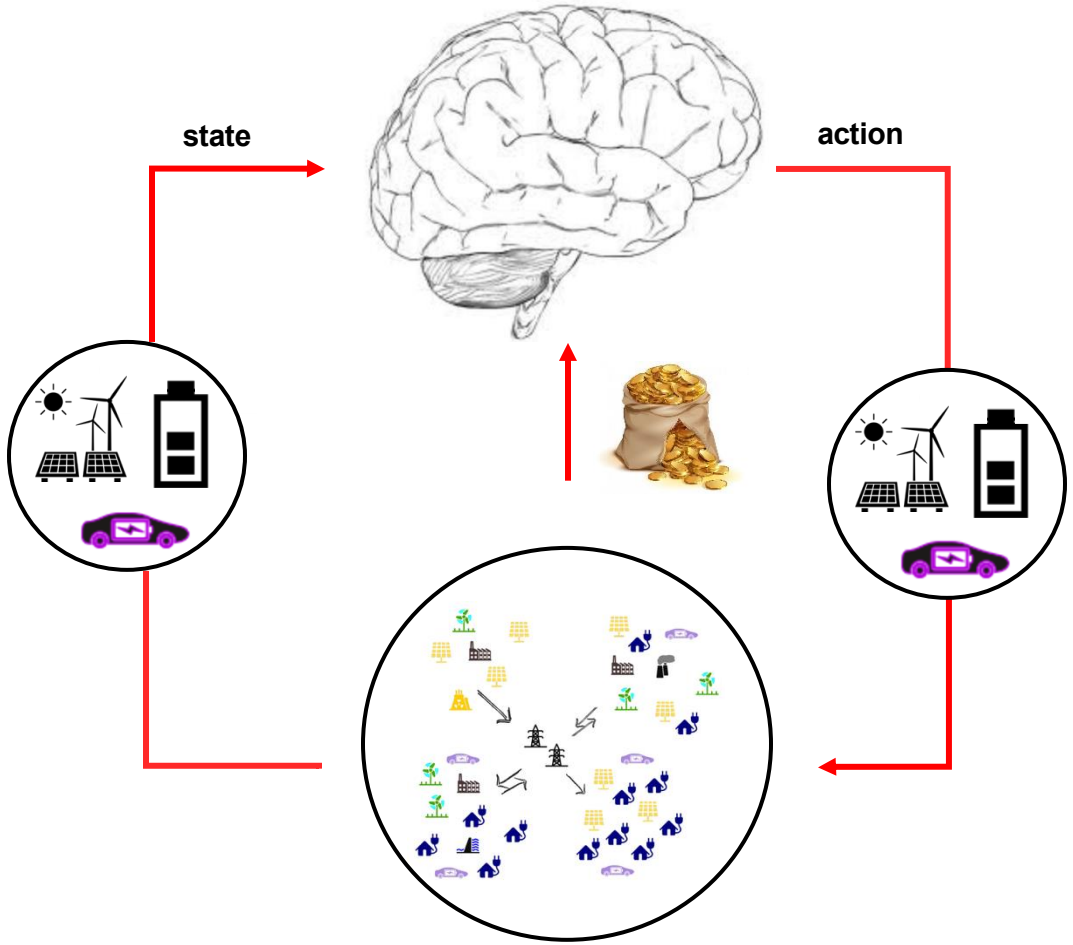
Describe the agent, environment, state and action!

Reinforcement learning: Battery example



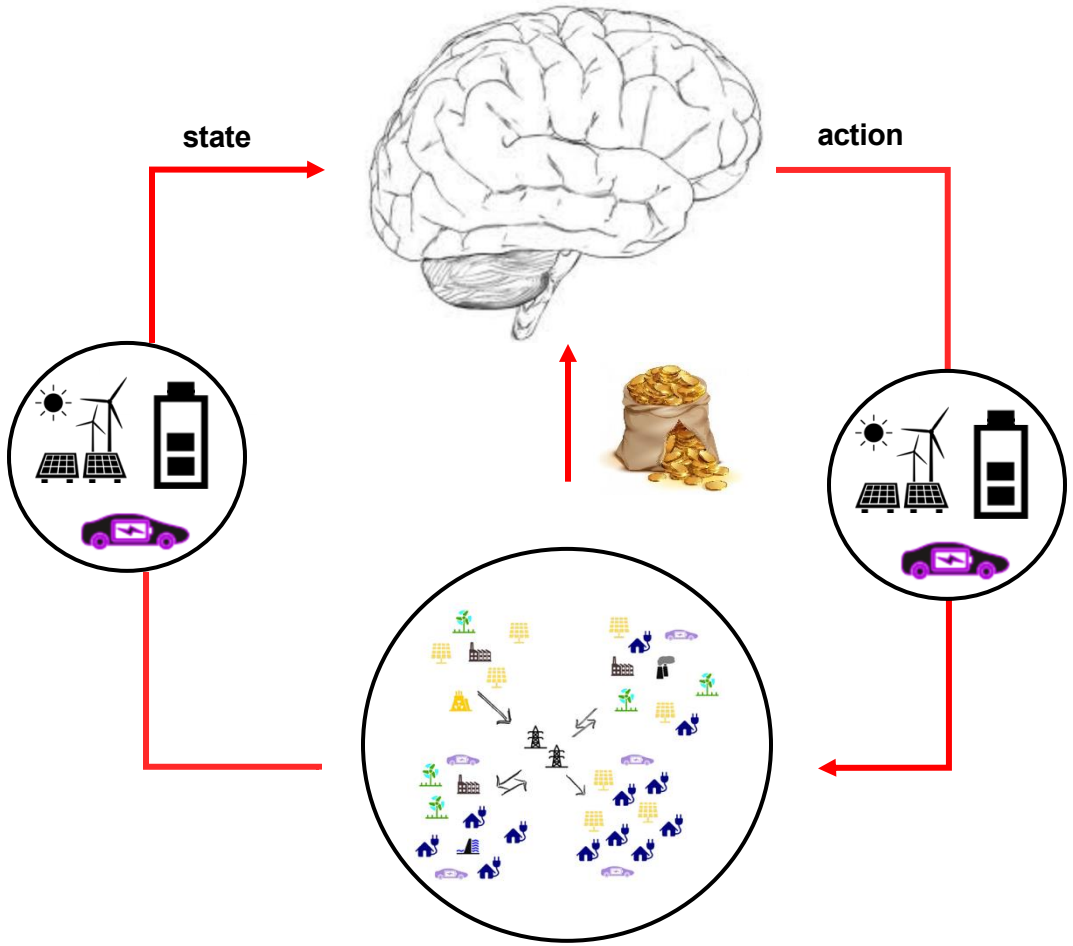


Reinforcement learning: Another example



Reinforcement learning: Another example

- SOC of the battery storage
- SOC of the electric vehicle
- Temprature
- Price
- Weather condition
- …



- How much to bid into electricity markets
- How much to store wind and solar production
- How much to charge/discharge the battery storage
- How much to charge the electric vehicle
- …

For the aforementioned examples, why is it not efficient to use supervised or unsupervised learning? Why reinforcement learning?

Why reinforcement learning in energy systems?

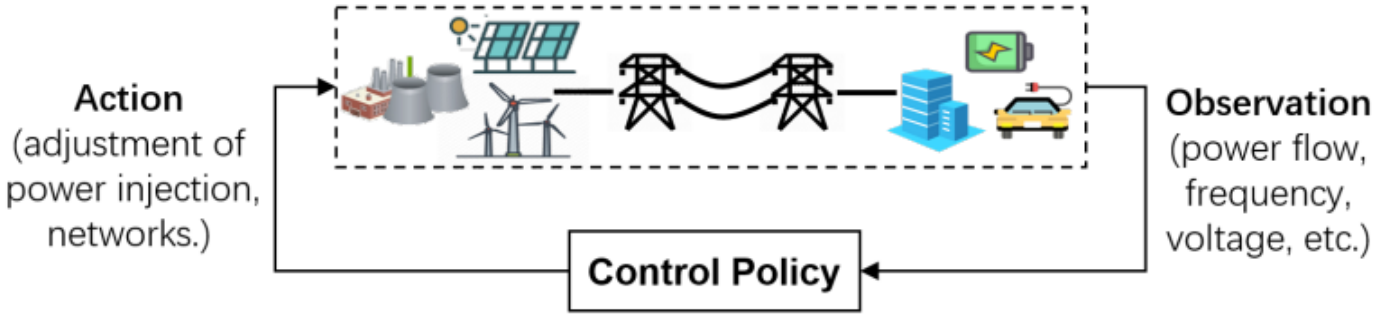
- The most appealing virtue of RL is **model-free**, i.e., it makes decisions without explicitly estimating the underlying models.
- Therefore, RL has the potential to capture **hard-to-model dynamics (the models are too complex to be useful)** and could outperform model-based methods in highly complex tasks.
- For instance, when the problems are intrinsically hard to model, such as the **human-in-loop control (e.g., in demand response),** RL and other data-driven methods are promising.
- Moreover, the data-driven nature of RL allows it to **adapt** to real-time observations and perform well in *uncertain dynamical environments*.

Do you think you could use RL for Assignment 1?

Applications of reinforcement learning in energy systems

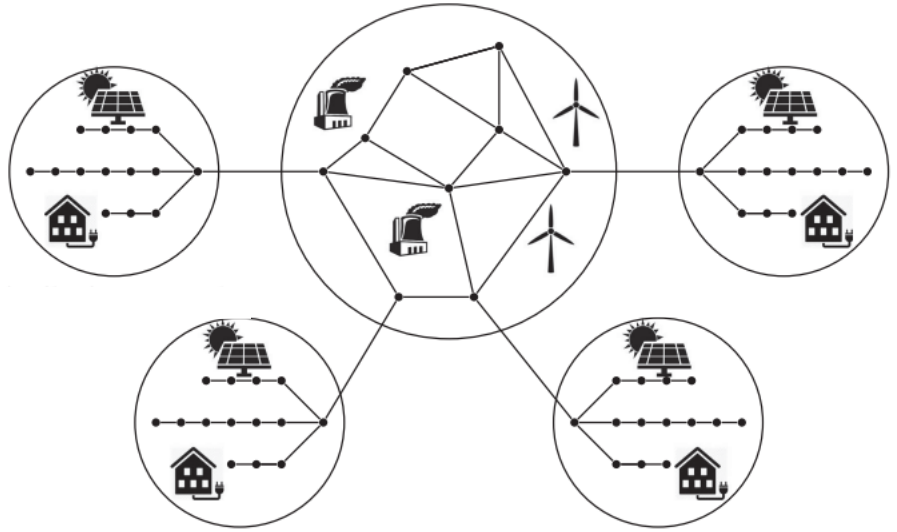
Applications of reinforcement learning

- For power system operation, frequency level and voltage profile are the two most critical indicators of system operating status, whilst reliable and efficient energy management is the core task.
- Therefore, the three key applications are frequency regulation, voltage control, and energy management.



Applications of reinforcement learning

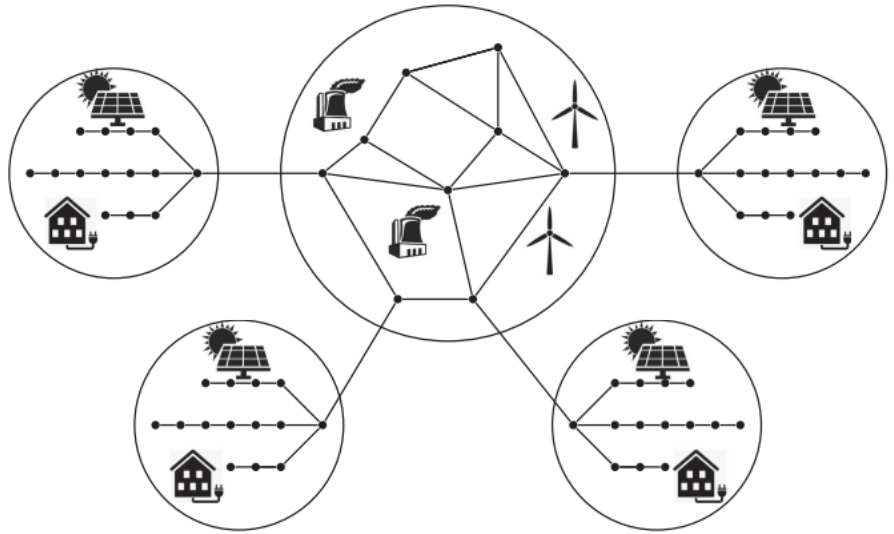
In energy management, there exist significant uncertainties that result from unknown models, parameters of networks, DER facilities, uncertain customer behaviors, unpredictable weather conditions, and etc



Applications of reinforcement learning

The energy management schemes using RL are mainly classified as two cases:

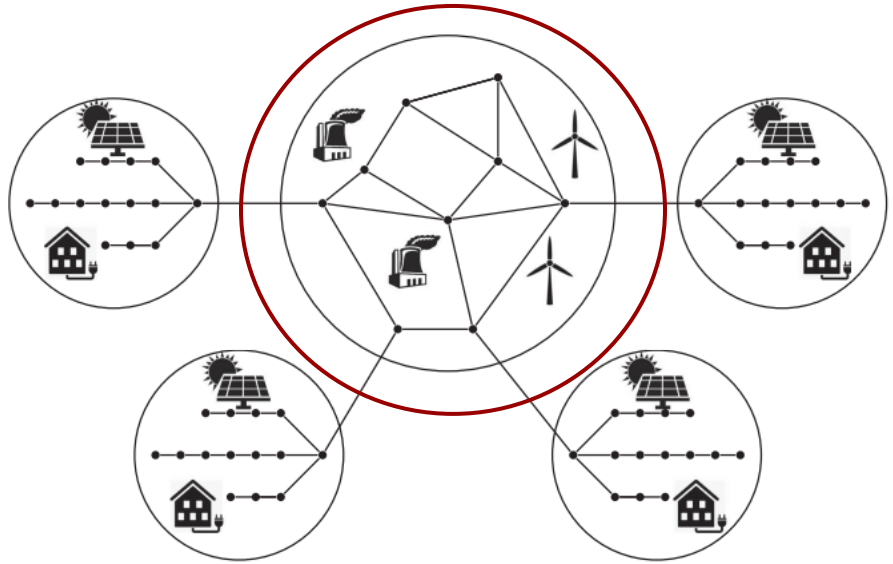
1) one is to **design the energy pricing policies** to guide and influence the power consumption or generation of end users from the perspective of system operators or load service entities, where the electricity price is defined as the action



Applications of reinforcement learning

The energy management schemes using RL are mainly classified as two cases:

1) one is to **design the energy pricing policies** to guide and influence the power consumption or generation of end users from the perspective of system operators or load service entities, where the electricity price is defined as the action

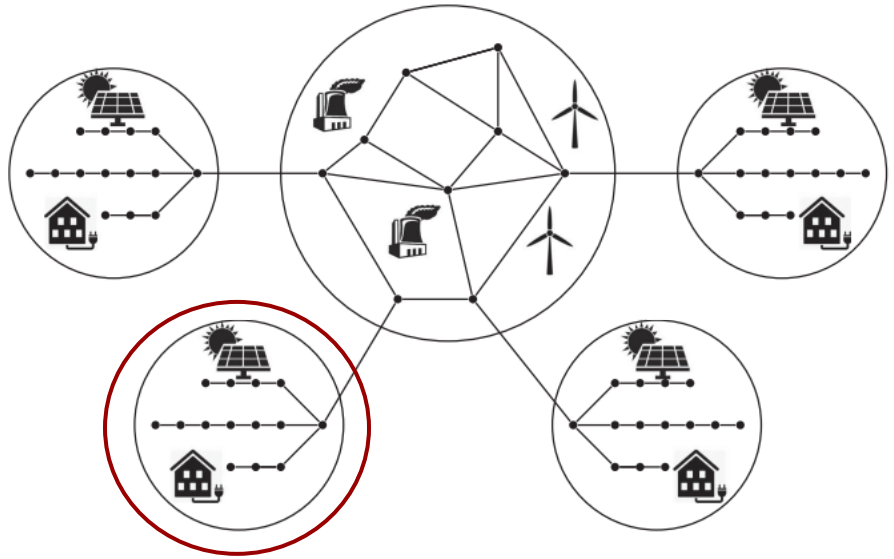


Applications of reinforcement learning

The energy management schemes using RL are mainly classified as two cases:

1) one is to **design the energy pricing policies** to guide and influence the power consumption or generation of end users from the perspective of system operators or load service entities, where the electricity price is defined as the action

2) the other **directly schedules the operation** of DERs, the heating, ventilation, and air conditioning (HVAC) systems, and other adjustable loads from the perspective of a stake holder to maximize the total net profit as a price taker



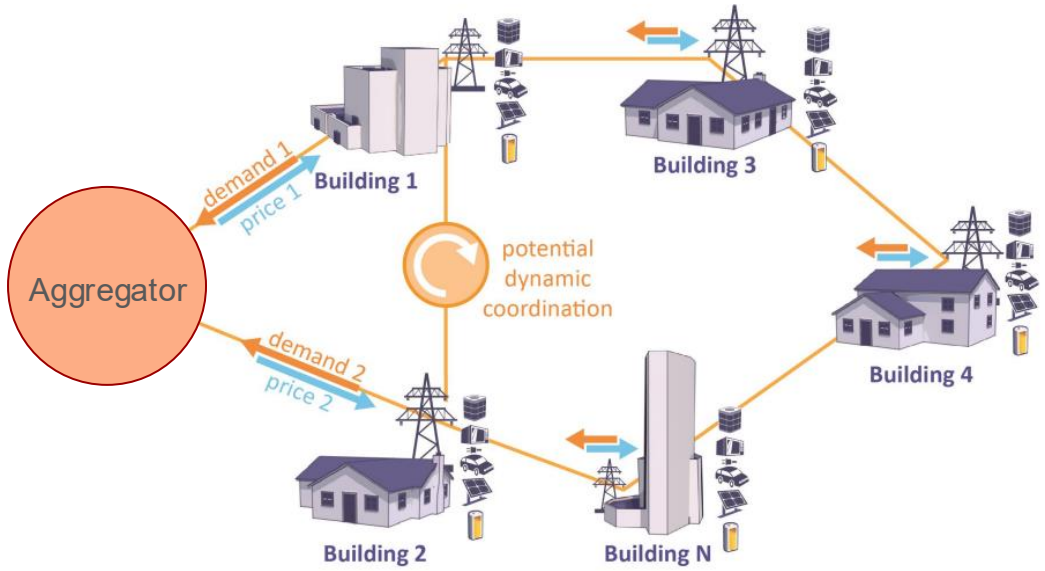
Applications of reinforcement learning

Let us see how to use RL to schedule the operation of

1) **Distributed energy resources**

2) **Building HVAC**

3) **Residential load**



Applications of reinforcement learning

1) Distributed Energy Resources: Consider a bundle of several typical DERs, including a photovoltaics (PV) panel, a battery unit, and an electric vehicle (EV). Describe an action vector and a state vector that both have dimensions exceeding 3.

Applications of reinforcement learning

1) Distributed Energy Resources: Consider a bundle of several typical DERs, including a photovoltaics (PV) panel, a battery unit, and an electric vehicle (EV). Describe an action vector and a state vector that both have dimensions exceeding 3.

Power output of PV, battery, and EV

$$\boldsymbol{a}_t = (p_t^{PV}, p_t^{Bat}, p_t^{EV})$$

$$\boldsymbol{s}_t = (E_t^{Bat}, E_t^{EV}, x_t^{EV}, \bar{p}_t^{PV})$$

Associated state of charge (SOC) levels

Other related states of EV, e.g. current location (at home or outside), travel plan, etc.

Applications of reinforcement learning

2) Building HVAC: Buildings account for a large share of the total energy usage, about half of which is consumed by the heating, ventilation, and air conditioning systems. Smartly scheduling HVAC operation has huge potential to save energy cost, but the building climate dynamics is intrinsically hard to model and affected by many environmental factors. Describe an action vector and a state vector that both have dimensions exceeding 3.

Applications of reinforcement learning

2) Building HVAC: Buildings account for a large share of the total energy usage, about half of which is consumed by the heating, ventilation, and air conditioning systems. Smartly scheduling HVAC operation has huge potential to save energy cost, but the building climate dynamics is intrinsically hard to model and affected by many environmental factors. Describe an action vector and a state vector that both have dimensions exceeding 3.

A building is divided into multiple thermal zones

Conditioned air temprature

Supply air temprature

Supply air flow at different zones

$$\boldsymbol{a}_t = \left(T_t^C, T_t^S, \left(m_{it}\right)_{i \in N}\right)$$

$$\boldsymbol{s}_t = \left(T_t^{Out}, \left(T_{it}^{In}, e_{it}\right)_{i \in N}\right)$$

Outside temprature

Inside temprature and occupancy rate at different zones

Applications of reinforcement learning

3) Residential Loads: Residential demand response motivates changes in electric use by end-use customers in response to time-varying electricity price or incentive payments. The domestic electric appliances are classified as 1) non-adjustable loads, e.g., computers, refrigerators, which are critical and must be satisfied; 2) adjustable loads, e.g., air conditioner, washing machine, whose operating power or time can be tuned. Describe an action vector and a state vector.

Applications of reinforcement learning

3) Residential Loads: Residential demand response motivates changes in electric use by end-use customers in response to time-varying electricity price or incentive payments. The domestic electric appliances are classified as 1) non-adjustable loads, e.g., computers, refrigerators, which are critical and must be satisfied; 2) adjustable loads, e.g., air conditioner, washing machine, whose operating power or time can be tuned. Describe an action vector and a state vector.

$z_{it} \in \{0,1\}$ denotes whether switching the on/off mode (equal 1) or keeping unchanged (equal 0).

The power consumption of the load, which can be adjusted either discretely or continuously depending on the load characteristics

$$\boldsymbol{a}_t = ((z_{it}, p_{it})_{i \in N})$$

For each load

$$\boldsymbol{s}_t = ((\alpha_{it}, x_{it})_{i \in N})$$

$\alpha_{it} \in \{0,1\}$ equals 0 for the off status and 1 for the on status.

Indoor and outdoor temperatures are contained if load is an air condition; or captures the task progress and remaining time to the deadline for a washing machine load.

Applications of reinforcement learning

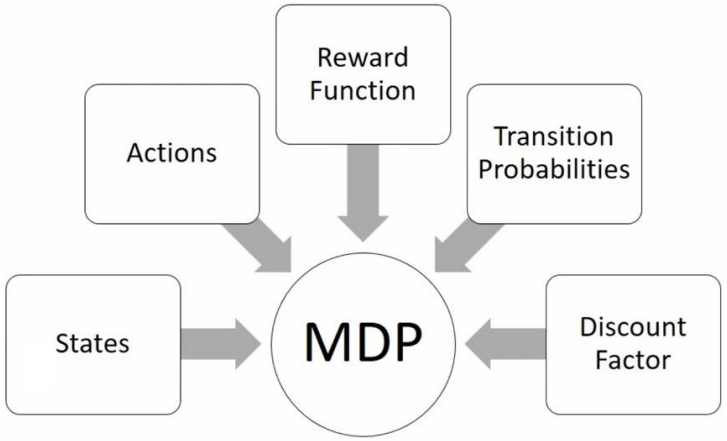
In addition to the operational states above, there are some critical system states including:

- The current time
- Past values and future predictions e.g. electricity prices from past K_p time steps to future K_f time predictions
- Previous actions may also be considered as components of the state

Markov decision process

Markov decision process

Markov decision processes (MDP) is fundamental to reinforcement learning, since it provides a systematic way to represent the interaction between an agent and environment.



Introduction to MDPs

- Markov decision processes formally describe an environment for reinforcement learning
- Where the environment is fully observable
- Markov property: a state $\boldsymbol{s}_t$ is Markov if and only if:

$$\mathbb{P}[\boldsymbol{s}_{t+1}|\boldsymbol{s}_t] = \mathbb{P}[\boldsymbol{s}_{t+1}|\boldsymbol{s}_1, \dots, \boldsymbol{s}_t]$$

- The state captures all relevant information from the history
- The future is independent of the past given the present

Introduction to MDPs

A Markov decision process is a tuple $(S, A, \{p_{sa}\}, \gamma, R)$.

- S is a set of states ($\boldsymbol{s} \in S$).
- A is a set of actions ($\boldsymbol{a} \in A$).

→ You have seen how to define states and actions!

- $\{p_{sa}\}$ are the state transition probabilities. For each state $\boldsymbol{s} \in S$ and action $\boldsymbol{a} \in A$, p_{sa} is a distribution over the state space. In other words, p_{sa} gives the distribution over what states we will transition to if we take action $\boldsymbol{a}$ in state $\boldsymbol{s}$. For problems that actions will not impact the uncertainty of next sate, we only have $\{p_s\}$.
- $R(\boldsymbol{s})$ or $R(\boldsymbol{s}, \boldsymbol{a}) \in \mathbb{R}$ is the reward function.
- $\gamma \in [0,1]$ is called the discount factor.

State transition matrix without action

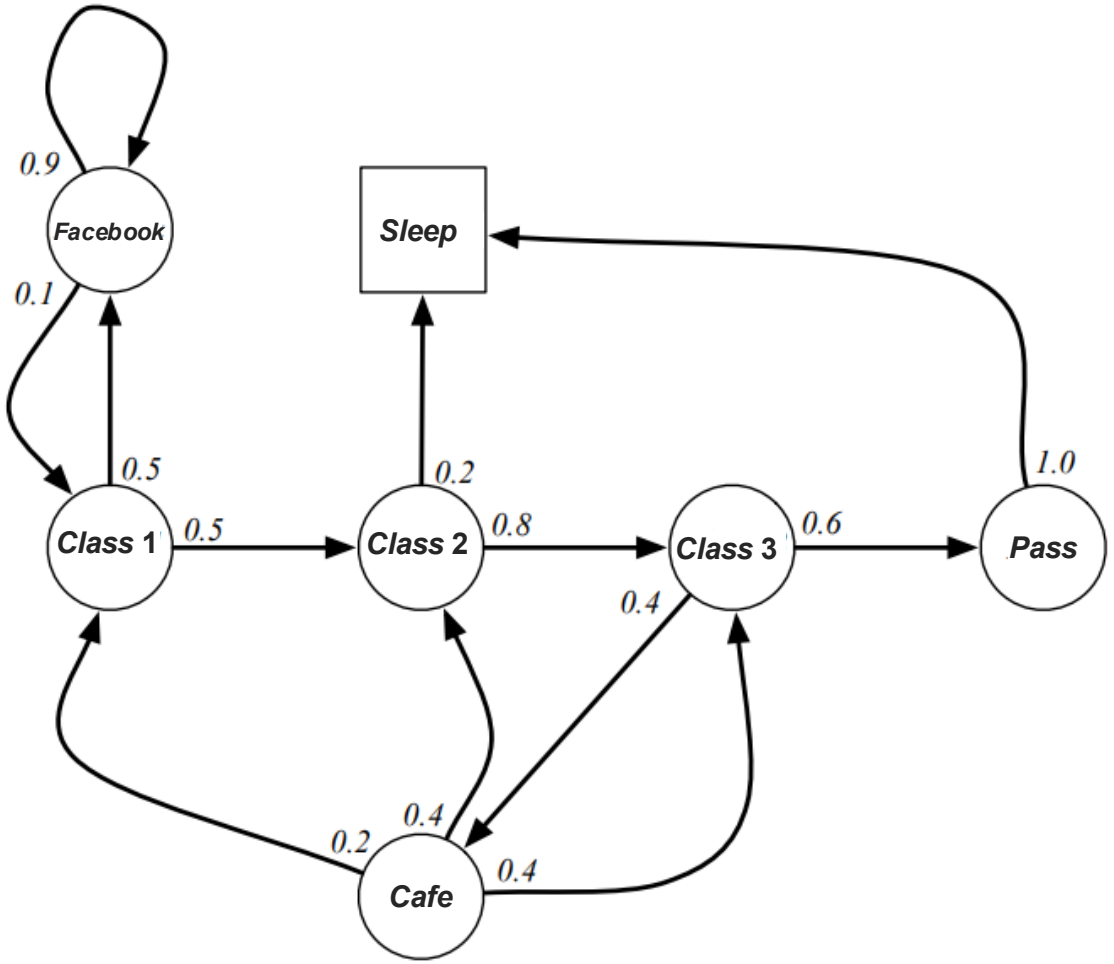
For a Markov state $\boldsymbol{s}$ and successor state $\boldsymbol{s}'$, the state transition probability is defined by

$$p_s(\boldsymbol{s}') = \mathbb{P}[\boldsymbol{s}_{t+1} = \boldsymbol{s}' | \boldsymbol{s}_t = \boldsymbol{s}]$$

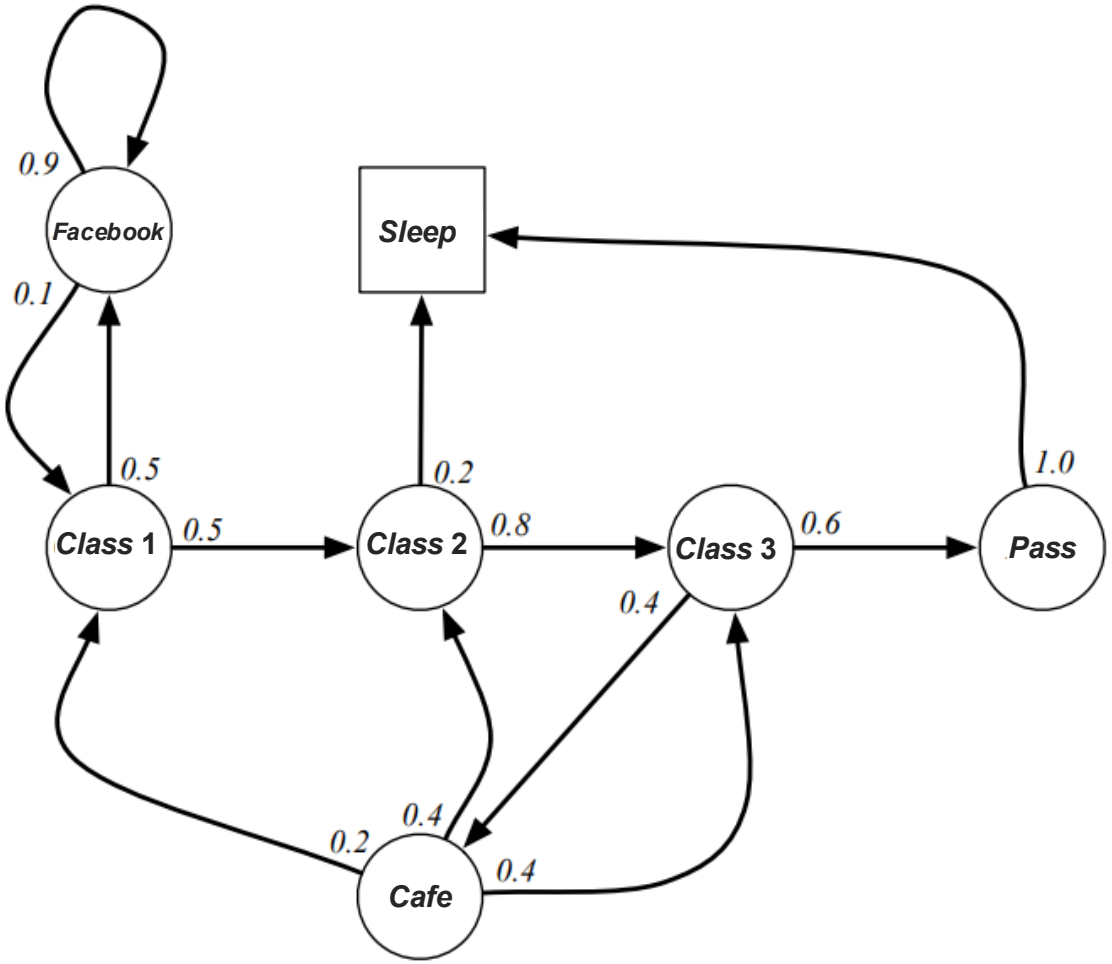
State transition matrix P defines transition probabilities from all states $\boldsymbol{s}$ to all successor states $\boldsymbol{s}'$,

$$P = \begin{bmatrix} p_{\boldsymbol{s_1}}(\boldsymbol{s}_1) & \cdots & p_{\boldsymbol{s_1}}(\boldsymbol{s}_n) \\ \vdots & & \vdots \\ p_{\boldsymbol{s_n}}(\boldsymbol{s}_1) & \cdots & p_{\boldsymbol{s_n}}(\boldsymbol{s}_n) \end{bmatrix}$$

Example: student Markov chain

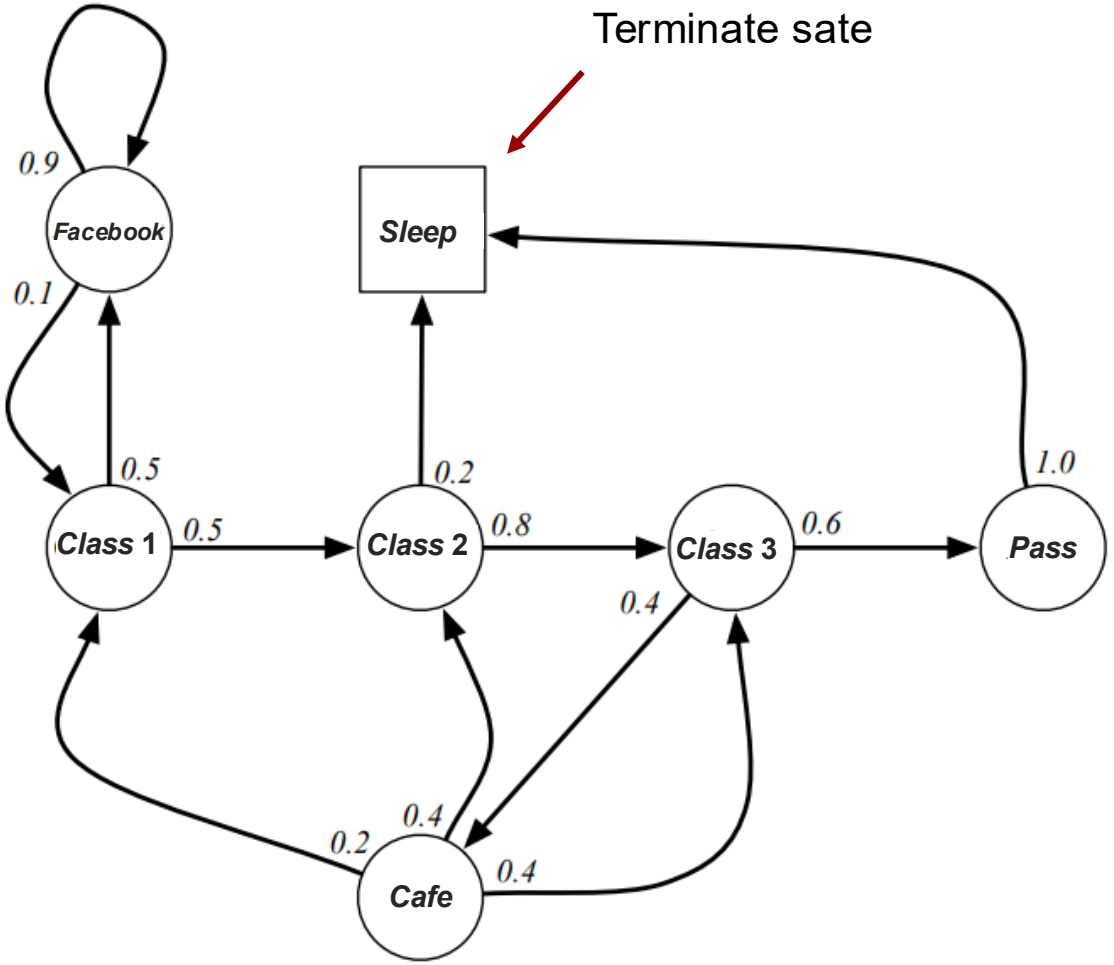


Example: student Markov chain



Once we got this Markov chain, let us think what it means to take samples of this?

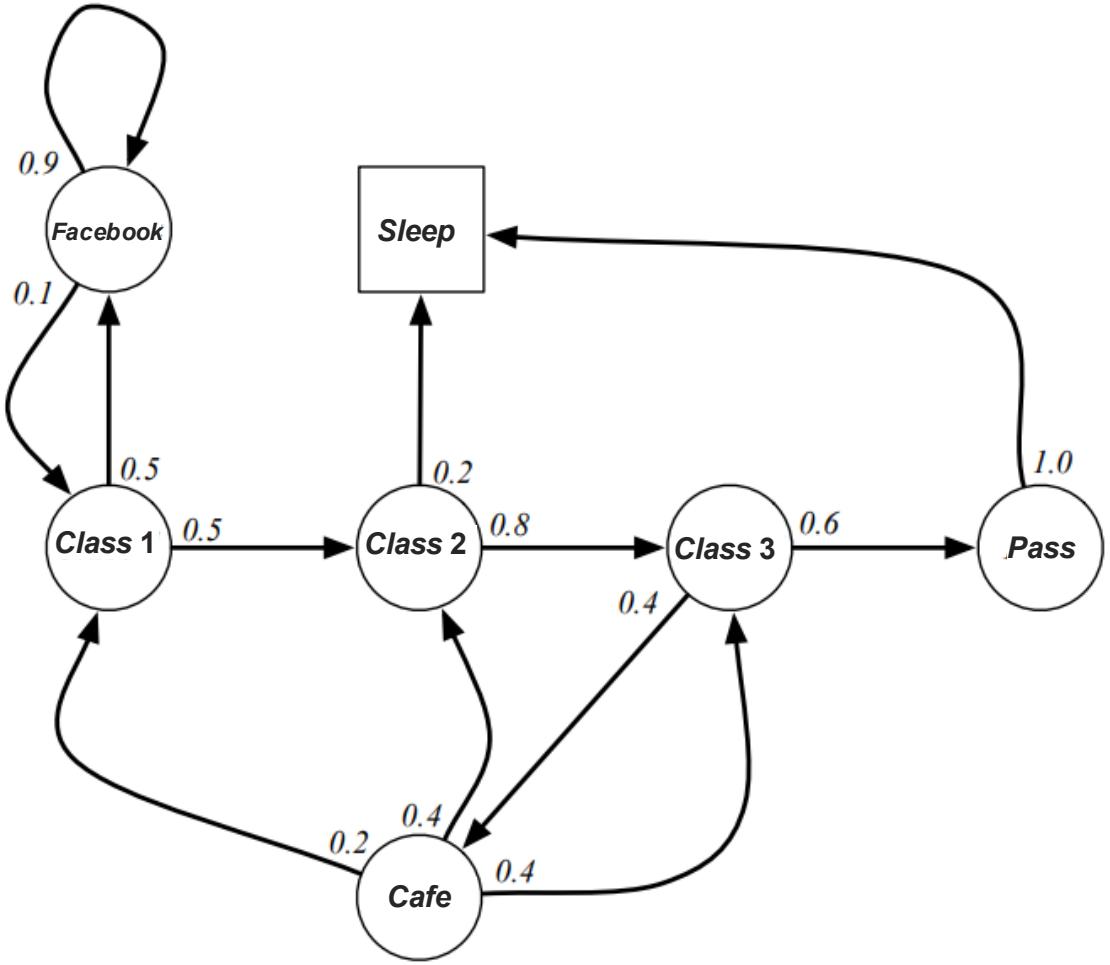
Example: student Markov chain episodes



Sample episodes for Student Markov Chain starting from $s_1 =$ Class 1

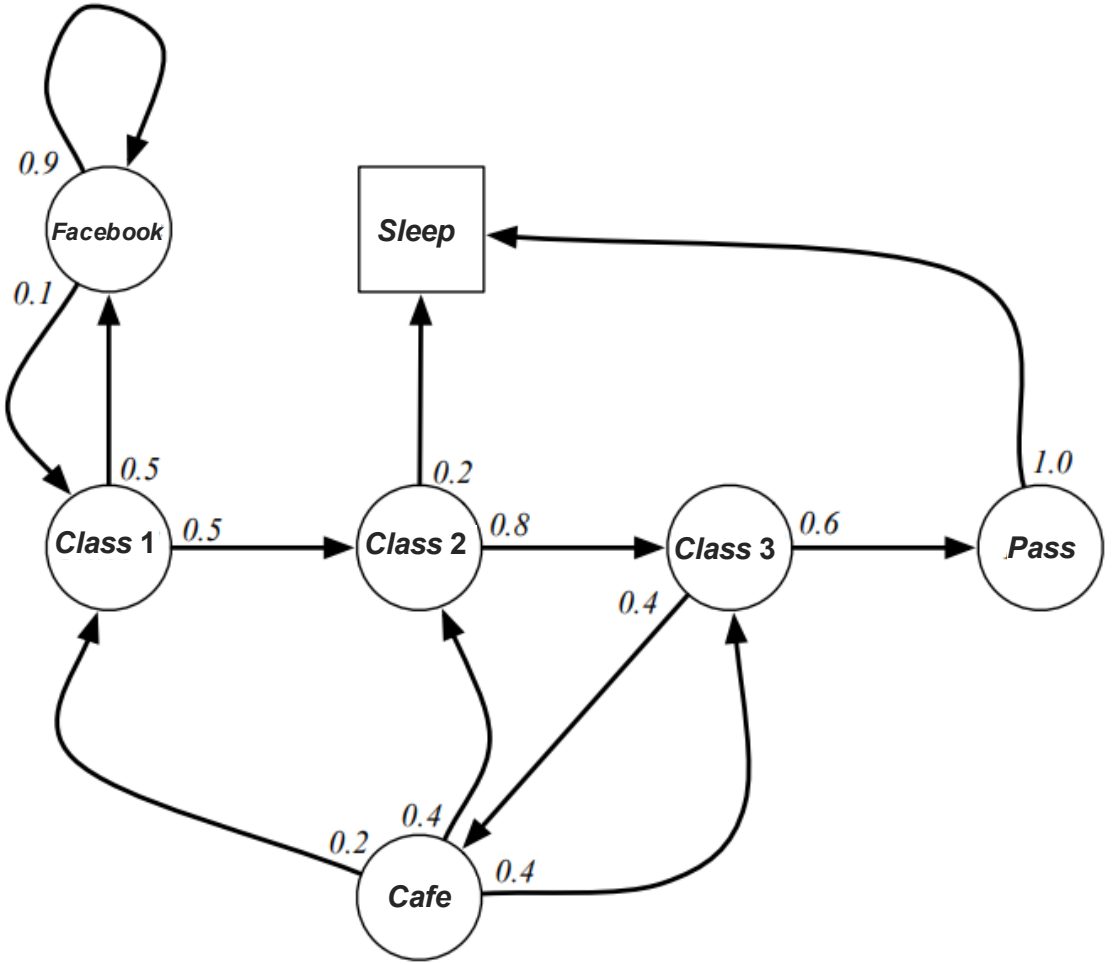
- Class 1, Class 2, Class 3, Pass, Sleep
- Class 1, FB, FB, Class 1, Class 2, Sleep
- Class 1, Class 2, Class 3, Café, Class 2, Class 3, Pass, Sleep
- …

Example: student Markov transition matrix



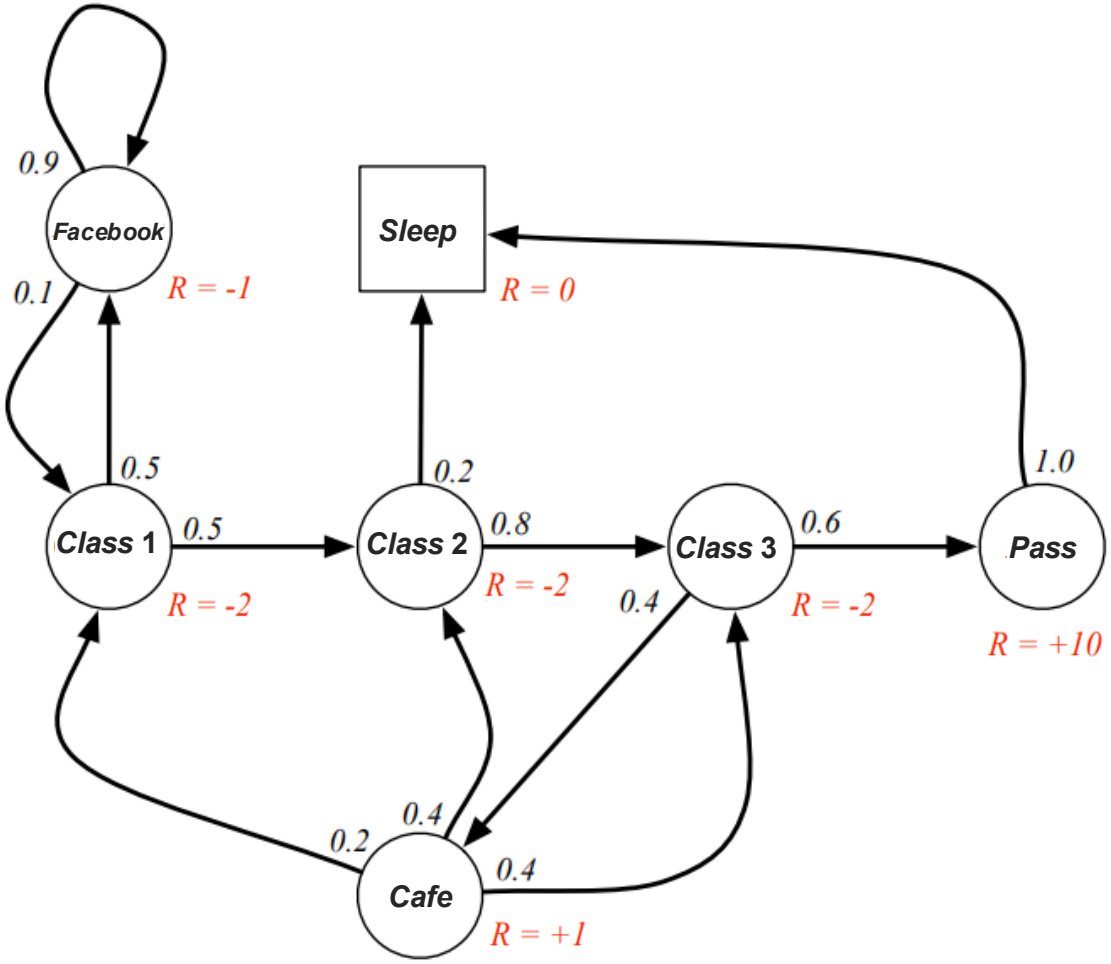
Can you write the transition matrix for this example?

Example: student Markov transition matrix



$$P = \begin{array}{c} \\ \text{C1} \\ \text{C2} \\ \text{C3} \\ \text{Pass} \\ \text{Cafe} \\ \text{FB} \\ \text{Sleep} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{C1} & \text{C2} & \text{C3} & \text{Pass} & \text{Cafe} & \text{FB} & \text{Sleep} \end{array} \\ \begin{bmatrix} & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{bmatrix} \end{array}$$

Example: student Markov chain with rewards



Value function

The value function $V(\boldsymbol{s})$ is the expected return starting from state s giving the long-term value of state $\boldsymbol{s}$

$$V(\boldsymbol{s}) = E[R(\boldsymbol{s}_0) + \gamma R(\boldsymbol{s}_1) + \gamma^2 R(\boldsymbol{s}_2) + \cdots | \boldsymbol{s}_0 = \boldsymbol{s}]$$

Most Markov reward and decision processes are discounted. Why?

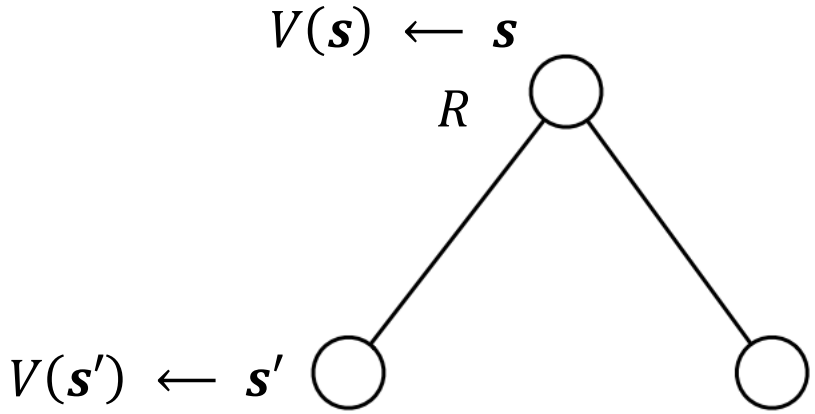
Discount factor

- Mathematically convenient to discount rewards
- Avoids infinite returns in cyclic Markov processes
- Uncertainty about the future may not be fully represented
- If the reward is financial, immediate rewards may earn more interest than delayed rewards Animal/human behaviour shows preference for immediate reward
- In economic applications where reward is the amount of money made, γ also has a natural interpretation in terms of the interest rate (where a dollar today is worth more than a dollar tomorrow)
- It is sometimes possible to use undiscounted Markov reward processes (i.e. $\gamma = 1$), e.g. if all sequences terminate.

Bellman equation

$$V(\boldsymbol{s}) = E[R(\boldsymbol{s}_t) + \gamma V(\boldsymbol{s}_{t+1}) | \boldsymbol{s}_t = \boldsymbol{s}]$$

$$V(\boldsymbol{s}) = R(\boldsymbol{s}) + \gamma \sum_{\boldsymbol{s}' \in S} P_s(\boldsymbol{s}') V(\boldsymbol{s}')$$



Bellman equation

The Bellman equation can be expressed concisely using matrices,

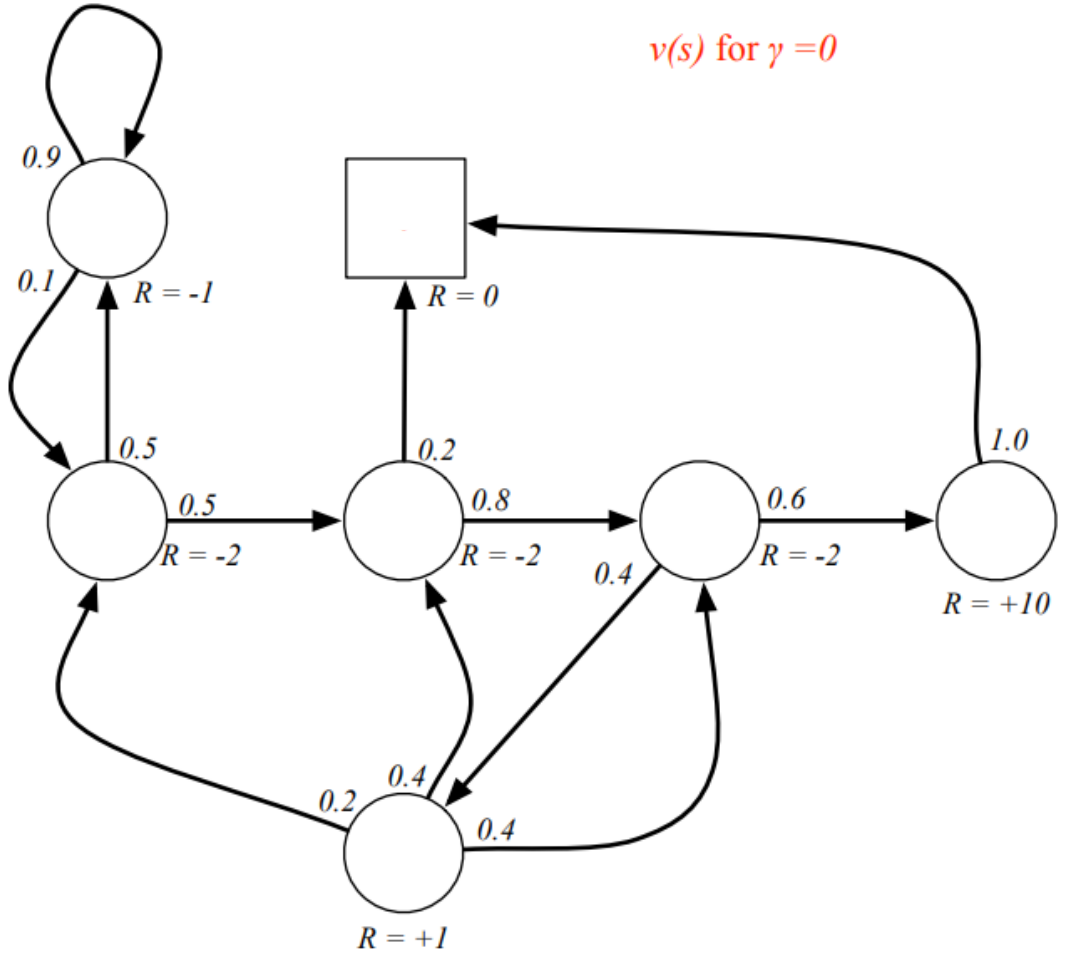
$$V = R + \gamma PV$$

The Bellman equation is a linear equation. Hence, it can be solved directly:

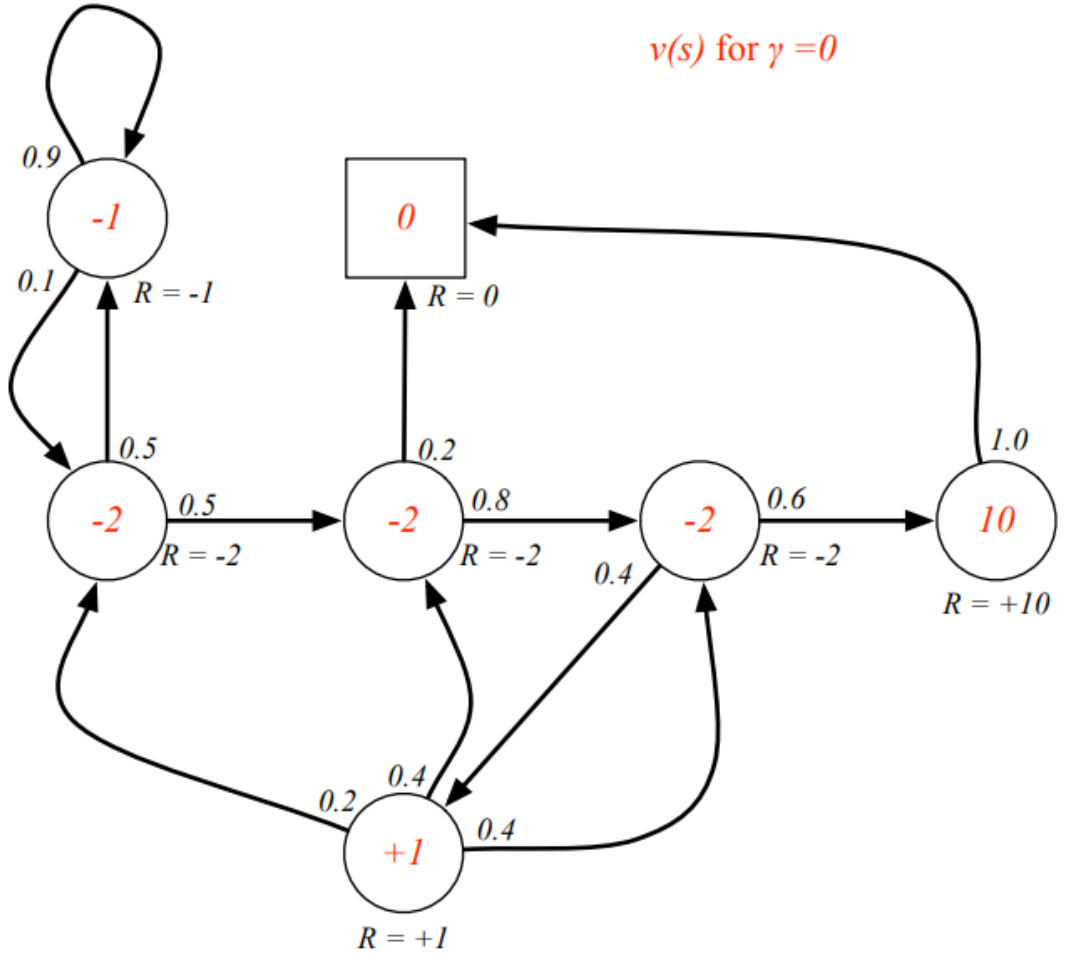
$$\boxed{V = (I - \gamma P)^{-1} R}$$

Direct solution only possible for small MRPs.

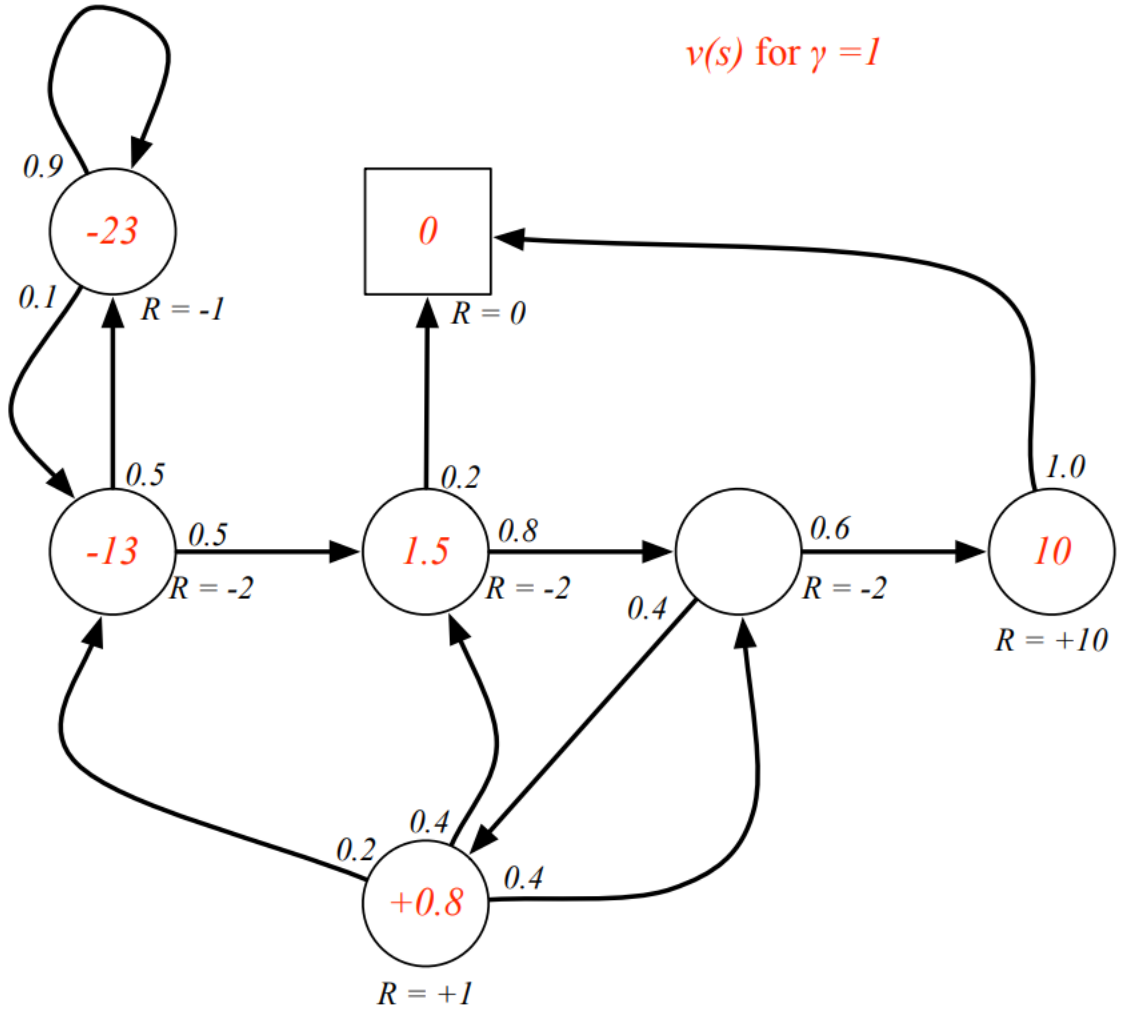
Example: value function



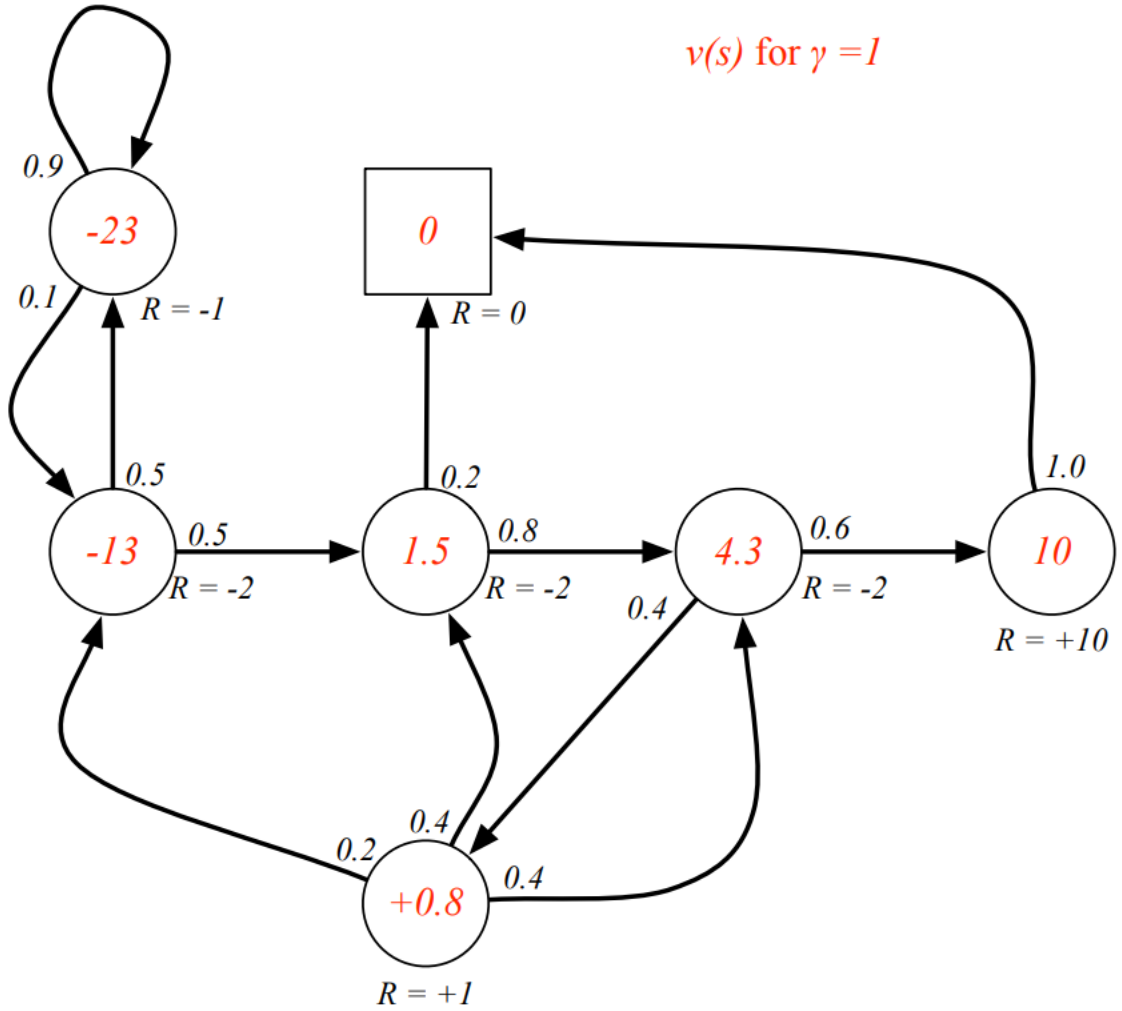
Example: value function

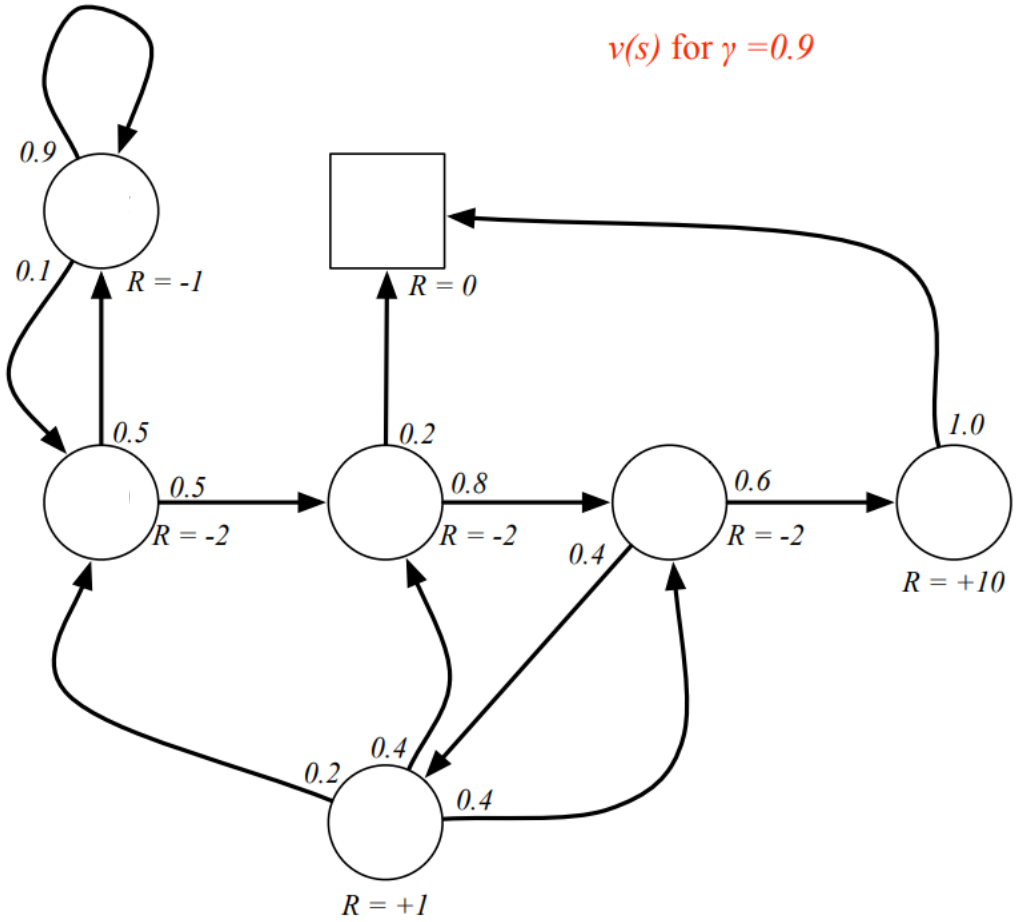


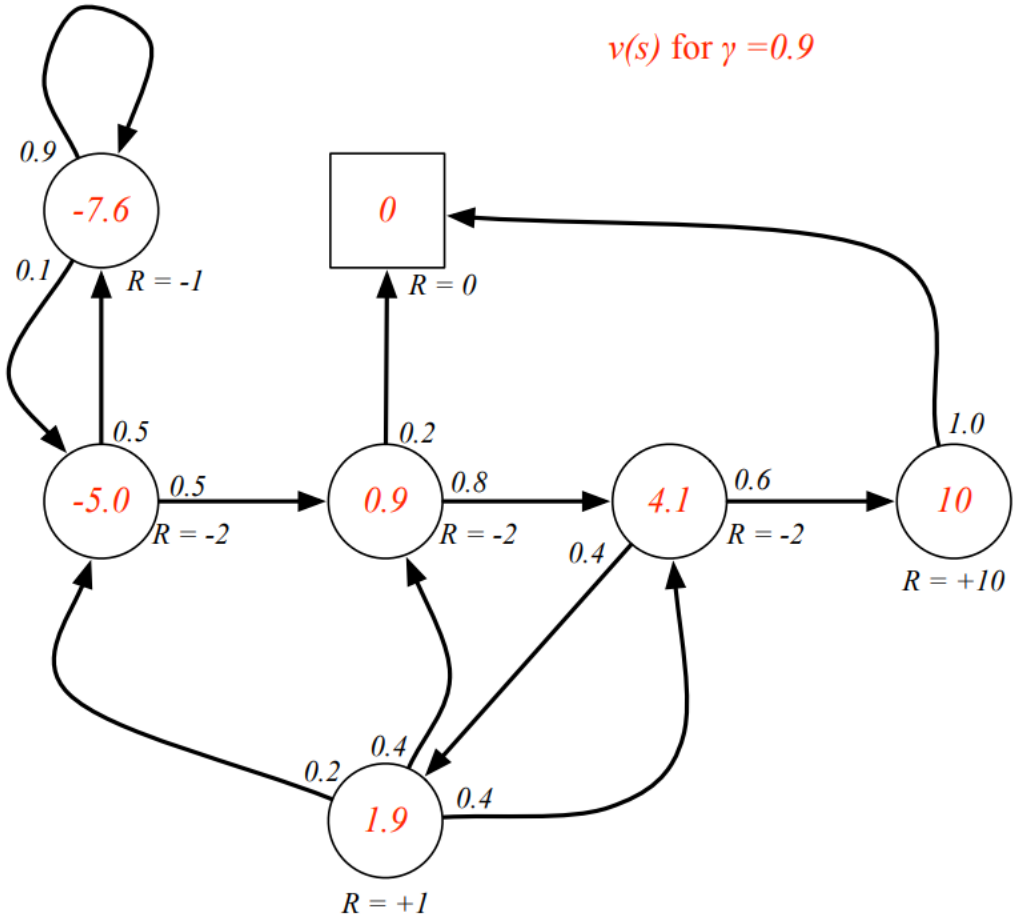
Example: value function



Example: value function




v(s) for γ =0.9
0.9
0.1
R = -1
R = 0
0.5
0.2
1.0
0.5
0.8
0.6
R = -2
R = -2
R = -2
0.4
R = +10
0.2
0.4
0.4
R = +1


v(s) for γ =0.9
0.9
-7.6
0.1
R = -1
0
R = 0
0.5
0.2
1.0
-5.0
0.5
R = -2
0.9
0.8
R = -2
4.1
0.6
R = -2
10
R = +10
0.4
0.4
0.2
0.4
1.9
R = +1

Learning objectives

Through this lecture, it is aimed for you to be able to

- Explain when RL is preferable to supervised/unsupervised learning in energy systems
- Figure out how to describe agent and environment for different energy applications
- Describe the fundamentals of Markov decision process

Thanks for your attention!

Email: farpour@dtu.dk