SMASH

machine learning for science and humanities postdoctoral program

Swarnendu Ghosh

Postdoctoral Researcher & Marie-Sklowdoska Curie Fellow

Faculty of Computer & Information Science, University of Ljubljana, Slovenia

Erasmus Mundus Fellow, University of Evora, Portugal

M.E. & Ph.D, Jadavpur University, India

UNIVERSIDADE DE ÉVORA

UNIVERZA V LJUBLJANI

UNIVERZA V NOVI GORICI

Jožef Stefan Institute

IZUM Institute of Information Science, Maribor

REPUBLIC OF SLOVENIA
MINISTRY OF THE ENVIRONMENT, CLIMATE AND ENERGY
SLOVENIAN ENVIRONMENT AGENCY

This is co- funded by the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355- SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund.

I FEEL SLOVENIA

Co-funded by the European Union

Vehicular Automation with Reinforcement Learning Agents

I FEEL SLOVENIA

Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

Outline

The Foundation - Reinforcement Learning Foundations (25 mins)
Establishing the mathematical and conceptual foundation

The System – Vehicular Autonomy (10 mins)
The "Physical" system and the traditional software stack

The Bridge – RL in AV (15 mins)
Implementation, Simulation, and Training

The Horizon – Gaps and Opportunities (10 mins)
Current bottlenecks and future research directions

Key Learning Objective: *By the end of this session, participants will understand how to transition from theoretical MDPs to safety-critical, autonomous navigation in simulated chaotic environments.*

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

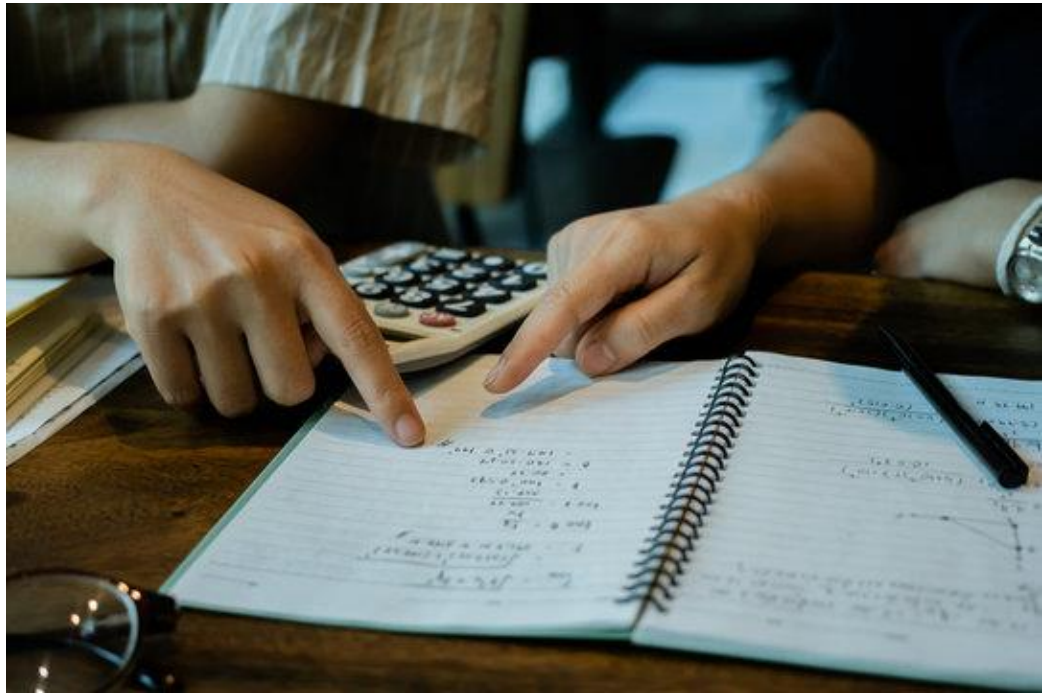
The Foundation - Reinforcement Learning Foundations

Establishing the mathematical and conceptual foundation

- Types of Learning
- Reinforcement Learning
- Terminologies
- Fundamental Computations
- Major types of RL

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

What is Learning ?



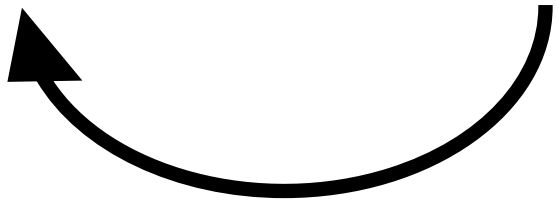
MINIMIZE ERROR
NEED DESIRED OUTPUT

MAXIMIZE SIMILARITY
NEED FAMILIAR PATTERNS

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

What is Learning ?

TRY → FAIL → ADAPT



I FEEL SLOVENIA

Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

What is Learning ?

- **I WILL TELL YOU THE RULES**
  → **JUST FOLLOW THE RULES**

- **I WILL SHOW YOU SOME DATA PATTERNS AND CORRECT REPONSES**
  → **TRY TO FIND THE RELATION BETWEEN THE PATTERN AND THE OUTCOME**

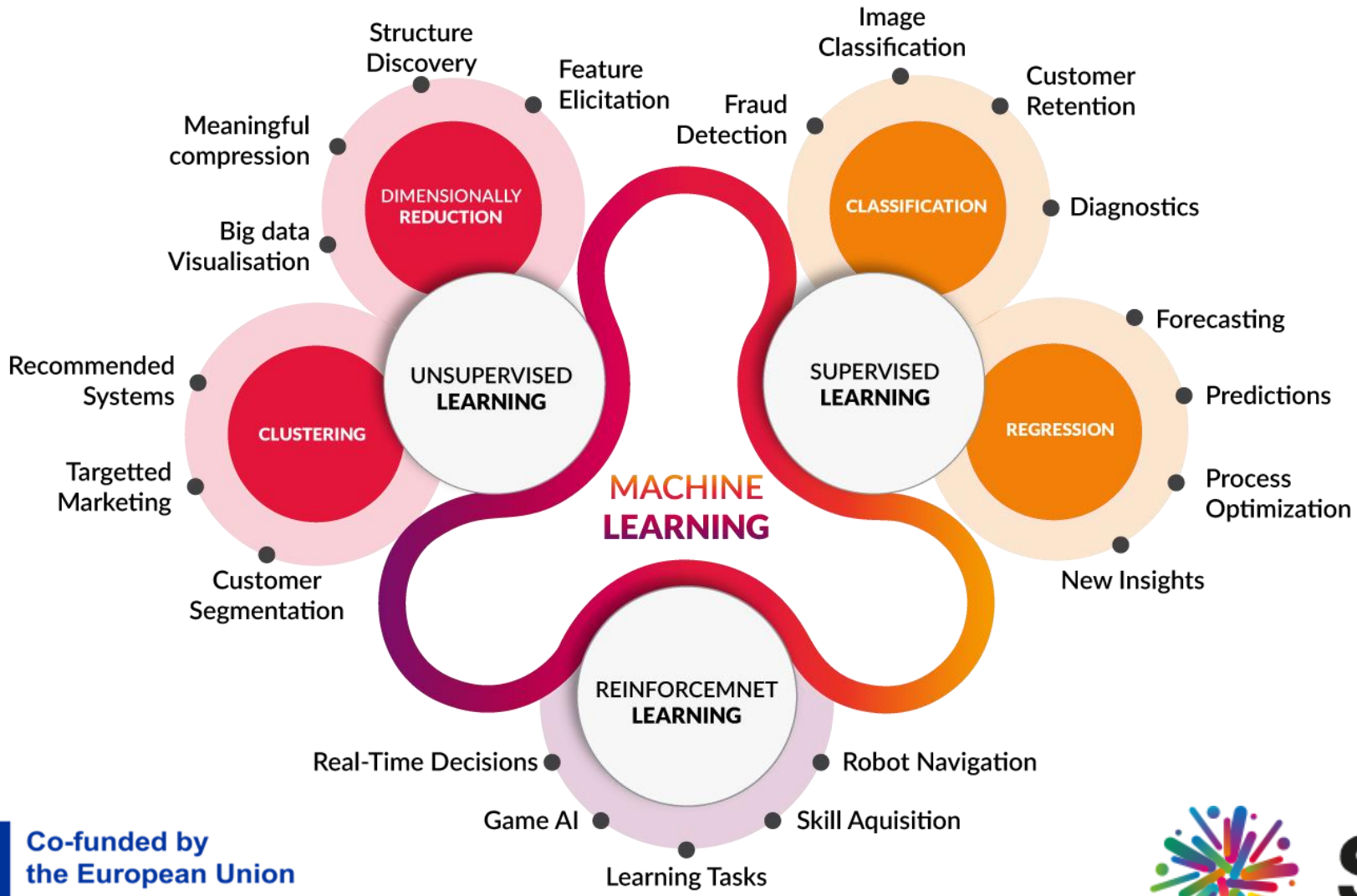
- **I WILL JUST GIVE YOU RAW DATA, I DON'T KNOW THE PATTERNS BUT I HAVE SOME SAMPLE OUTPUTS**
  → **SEE IF YOU CAN EXTRACT THE PATTERNS AND RELATE WITH THE OUTPUT**

- **I WILL JUST GIVE YOU THE DATA**
  → **FIND SOME PATTERNS IN THE DATA THAT MAKES SENSE**

- **I HAVE NEITHER THE DATA NOR THE OUTPUT .. BUT I HAVE PLACE FOR YOU TO WORK**
  → **TRY TO FIND OUT WHATS THE BEST WAY TO WORK HERE**

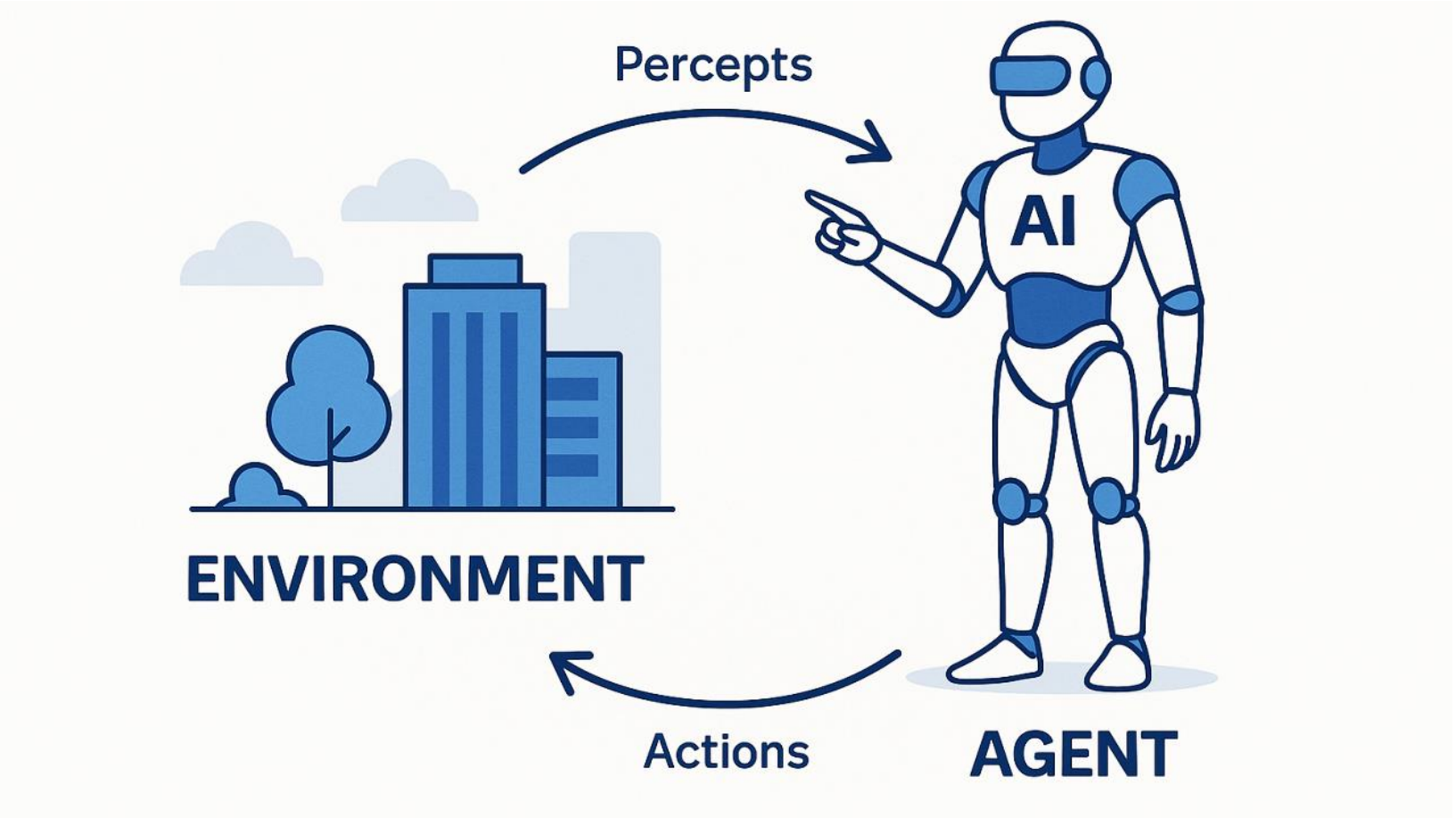
I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

The ML World



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Agents and Environments

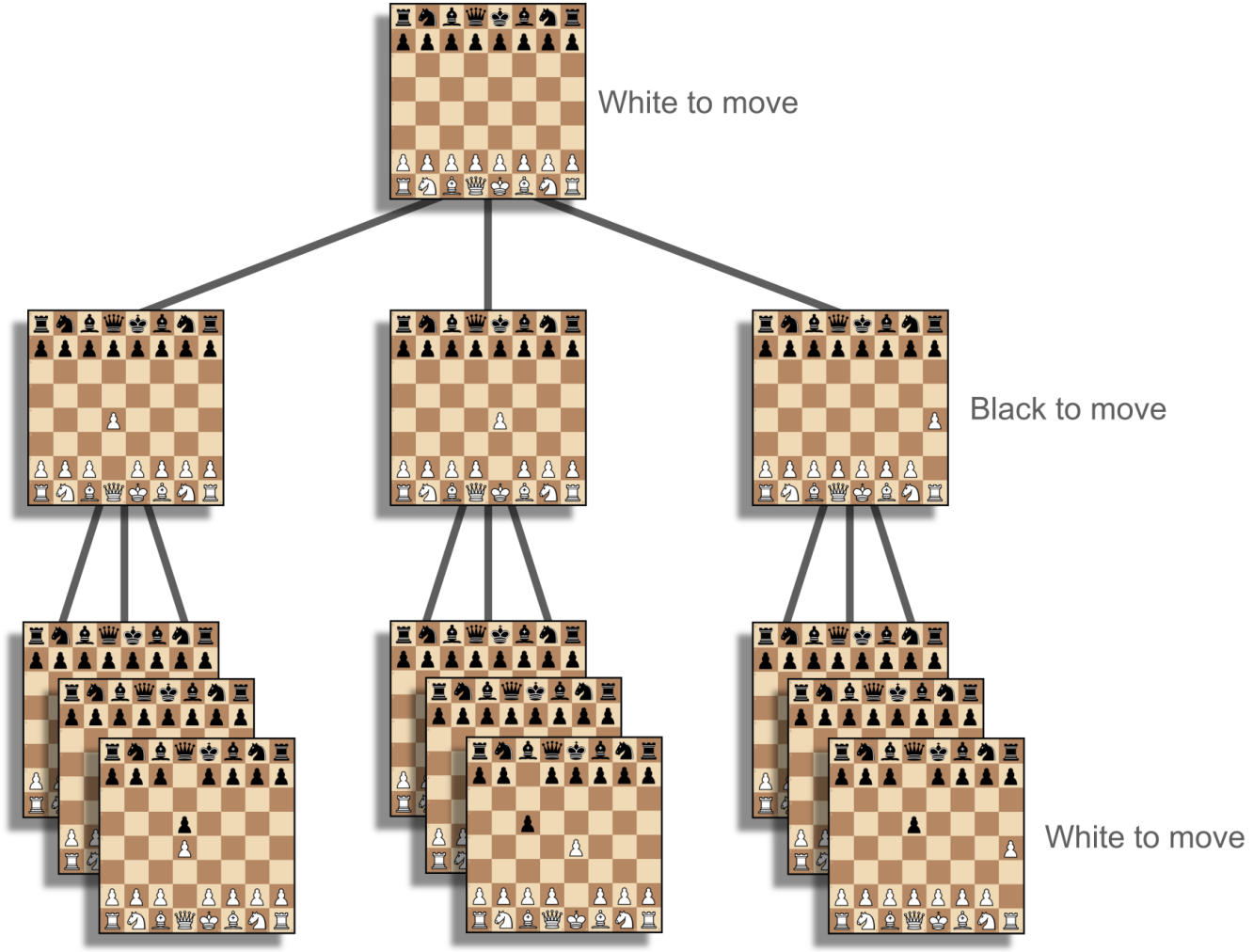


I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Markov Decision Process

The defining characteristic of an MDP is the **Markov Property**:

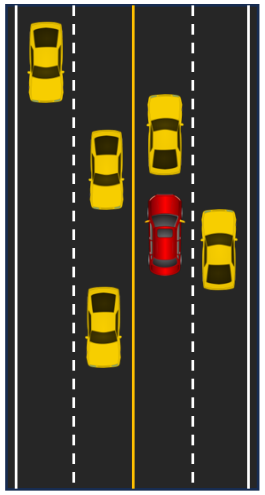
The future is independent of the past, given the present.



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

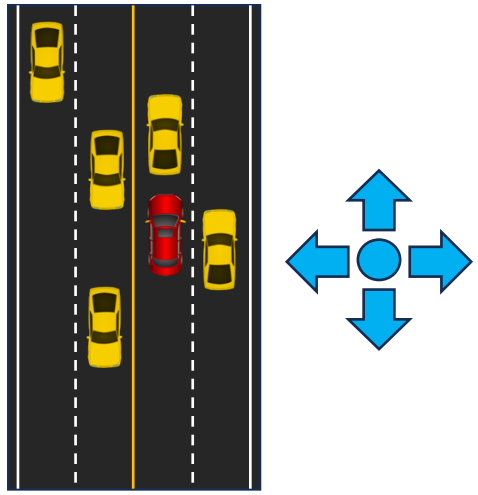
State

- A mathematical "snapshot" representing all relevant information about the environment at time t.
- A data vector containing the ego-vehicle's coordinates, current velocity, and the relative positions/trajectories of nearby pedestrians and vehicles detected via sensor fusion.



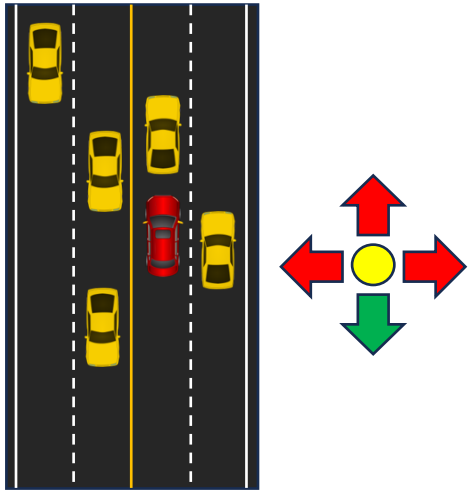
Action

- The decision or command the agent executes to transition between states.
- High-level maneuvres like **"Perform Lane Change"** or low-level control signals such as applying **15% braking pressure** or a **5° steering adjustment**.



Reward

- A scalar feedback signal (+/−) that evaluates the immediate quality of an action.
  - **Positive:** Progress toward the destination and maintaining lane centre.
  - **Negative:** Penalties for collisions, traffic light violations, or high "traffic entropy" (unsafe environmental complexity).

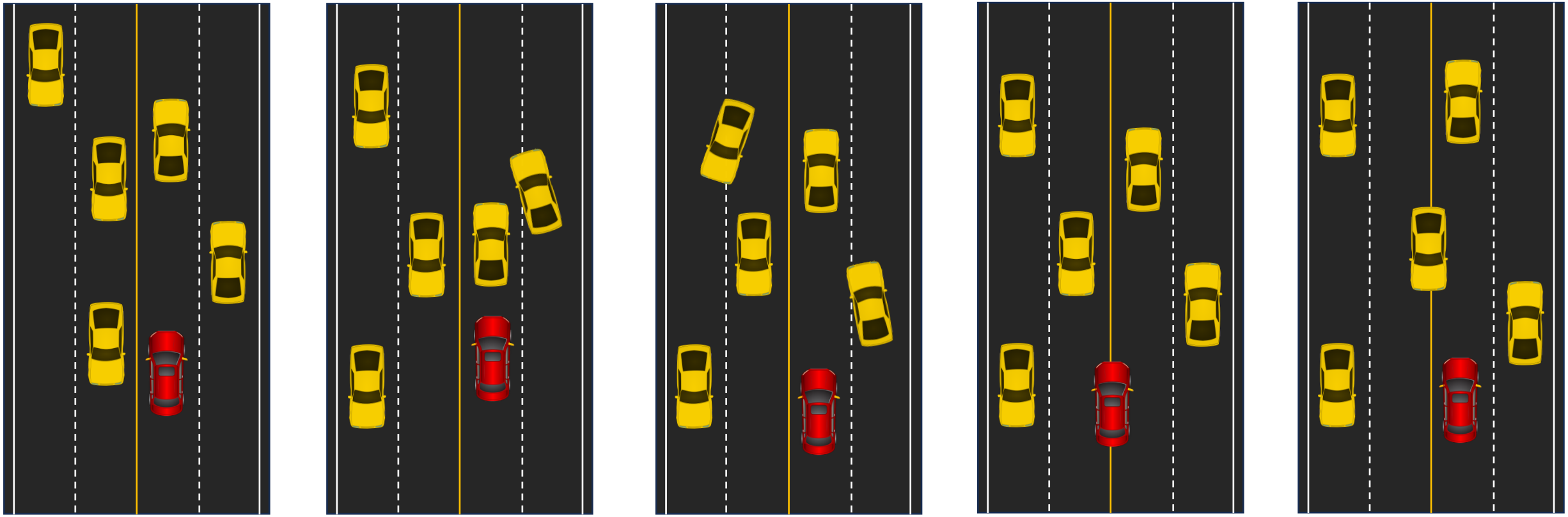


I FEEL SLOVENIA

Co-funded by the European Union

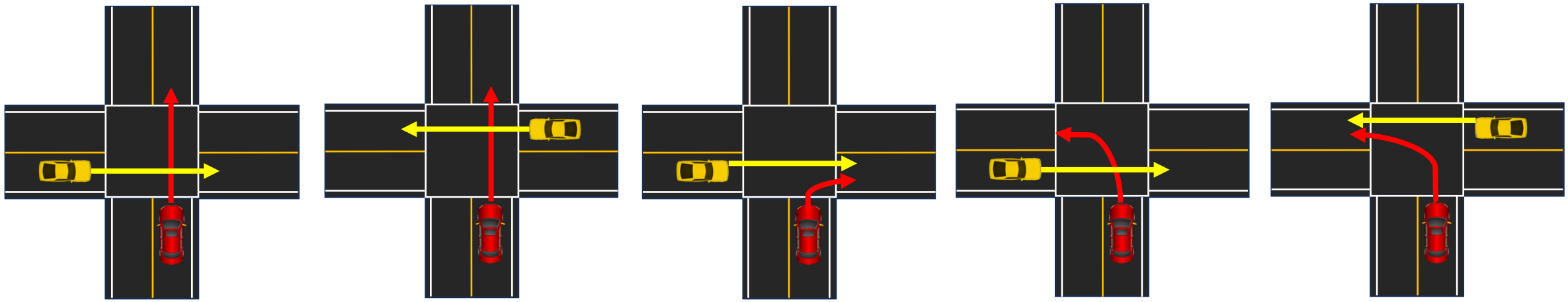
SMASH
machine learning for science and humanities postdoctoral program

What is Value ?



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

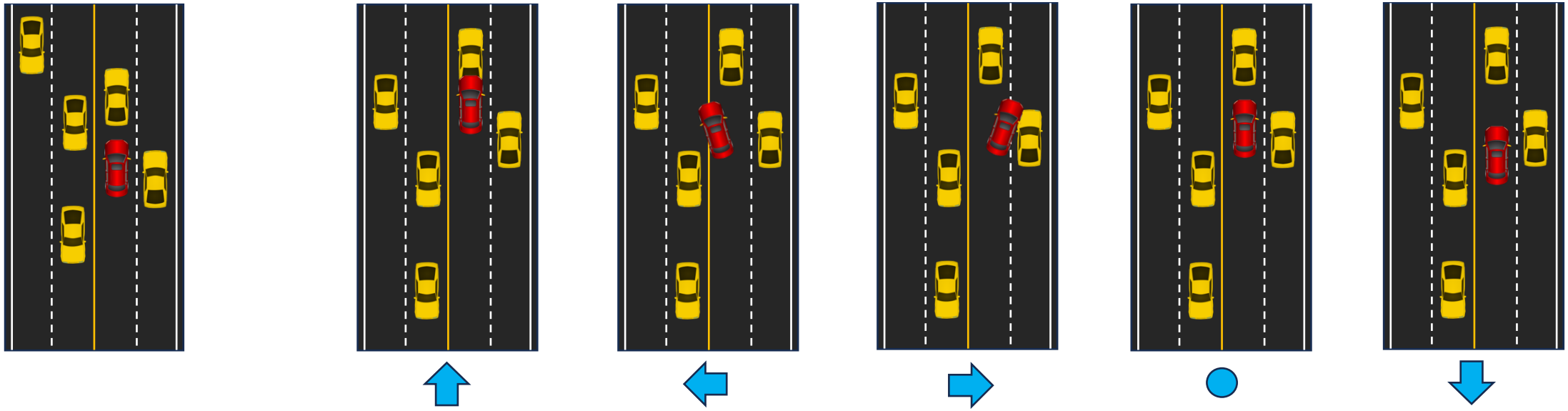
What is Policy ?



SHOULD THE RED CAR STOP OR GO ?

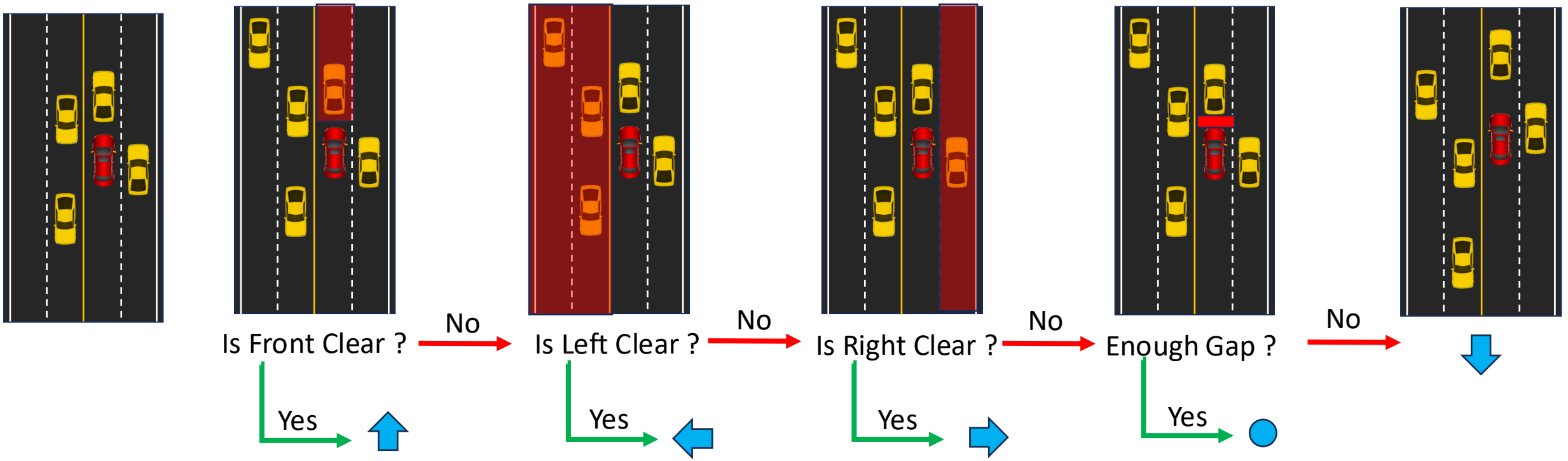
I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Value based Learning



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Policy based Learning



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Bellman's (Optimality) Equation

STOCHASTIC OUTPUT

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \right]$$

DETERMINISTIC OUTPUT

$$V^*(s) = \max_a [R(s,a) + \gamma V(s')]$$

- $V^*(s)$: The **Value** of the current state.
- $\max_a$: The assumption that the agent will choose the **best possible action**.
- $R(s,a)$: The **Immediate Reward** received for taking action a in state s.
- γ $(Gamma)$: The **Discount Factor** (e.g., 0.9), which devalues rewards the further they are in the future.
- $\sum_{s'} P(s' \mid s,a)\, V(s')$: The **Expected Future Value**. We multiply the probability of landing in a next state s' by the value of that state $V(s')$.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Bellman's (Expectation) Equation

Because we don't know the correct policy, let us consider a policy π.
So at any point t, we would like to maximize the EXPECTED REWARD.

$$V^{\pi}(s_t) = \mathbb{E}_{\pi}[R_{t+1}(s_t, a_{\pi}) + \gamma V^{\pi}(s_{t+1})]$$

- $V^{\pi}(s_t)$: The value of the state at time t under policy π.
- $\mathbb{E}_{\pi}$: The expectation (average) taken over the probabilities defined by policy π and the environment transitions.
- $R_{t+1}(s_t, a_{\pi})$: The reward depends on the current state and the action chosen by the policy.
- $\gamma\, V^{\pi}(s_{t+1})$: The discounted value of the next state.

$$V^{\pi}(s) \rightarrow V^{*}(s)$$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Bellman's (Expectation) Equation

State-Value Function: $V^{\pi}(s_t)$

$$
\begin{aligned}
&= \mathbb{E}_{\pi}[R_{t+1} + \gamma V^{\pi}(s_{t+1})] \\
&= \mathbb{E}_{\pi}[R_{t+1} + \gamma (R_{t+2} + \gamma V^{\pi}(s_{t+2}))] \\
&= \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V^{\pi}(s_{t+2})] \\
&= \mathbb{E}_{\pi}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 \left(R_{t+3} + \gamma V^{\pi}(s_{t+3})\right)\right] \\
&= \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V^{\pi}(s_{t+3})] \\
&\dots \\
&= \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots + \gamma^n R_{t+n+1} + \dots]
\end{aligned}
$$

$$
V^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\right]
$$

SMASH
machine learning for science and humanities postdoctoral program

Bellman Expectation Equation is typically implemented through **Iterative Policy Evaluation**. This is the process of computing the state-value function $V(s)$ for a fixed policy π

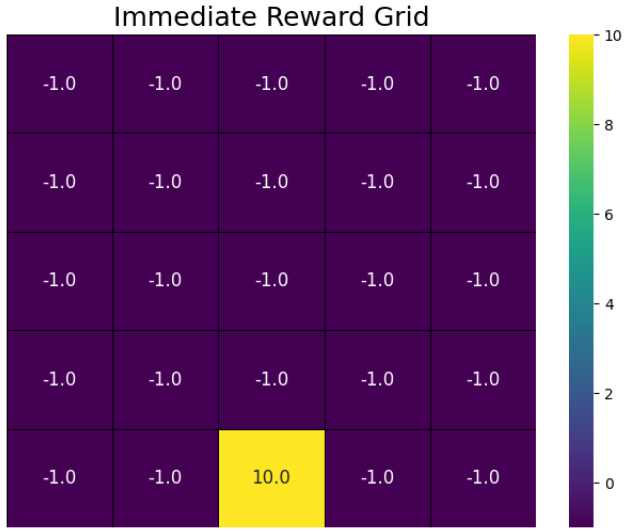
The Pseudocode

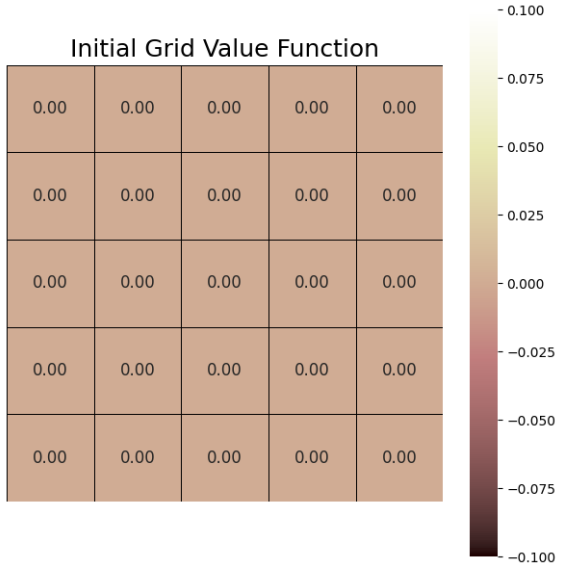
$\pi = Policy$
$\gamma = Discount\ Factor$
$\theta = Convergence\ threshold$
$\mathcal{S} = State\ Space$
$\mathcal{A} = Action\ Space$

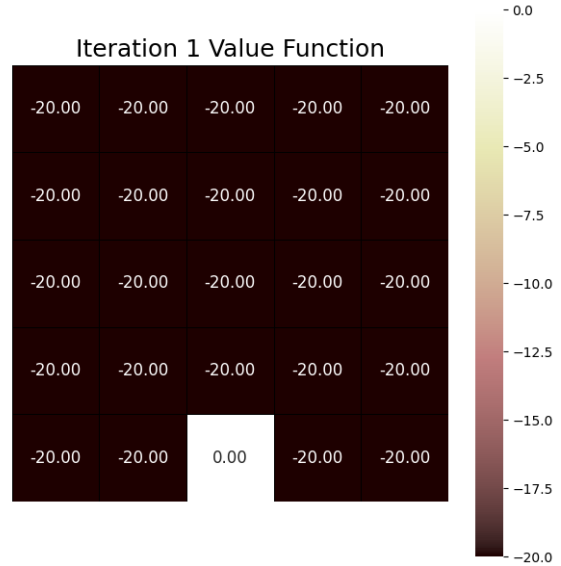
Function $(\pi, \gamma, \theta, \mathcal{S}, \mathcal{A})$**:**
  Initialize:
    - $V(s) \leftarrow 0, \forall s \in \mathcal{S}$ # Set initial value as 0
  Loop θ:
    - $\Delta \leftarrow \theta$ # Set Threshold as θ
    - $V_{old} \leftarrow V$ # Create backup values
    - **for each** $(s \in \mathcal{S})$**:**
      - **for each** $(a \in \mathcal{A})$: # For each state − action pair
        - $v_{temp} \leftarrow 0$. # temporary value
        - **for each** $(s', r\,|(s, a))$:
          - $v_{temp} \leftarrow v_{temp} + \pi(a|s) \cdot p(\langle s', r\rangle|\langle s, a\rangle) \cdot [\, r + \gamma\, V_{old}(s')]$
          # accumulate rewards for temporary value
      - $V(s) \leftarrow v_{temp}$. # update new value
    - $\Delta = \max\left(\theta, \frac{1}{|\mathcal{S}|}\sum_{i=0}^{|\mathcal{S}|-1} V(s_i)\right)$. # check threshold

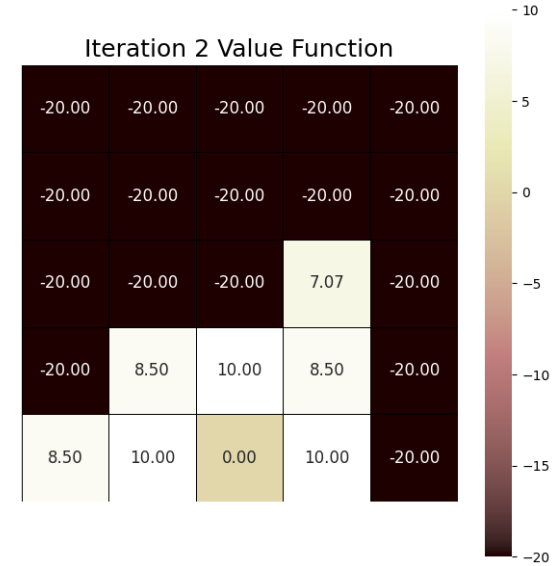
  until $\Delta < \theta$
return V

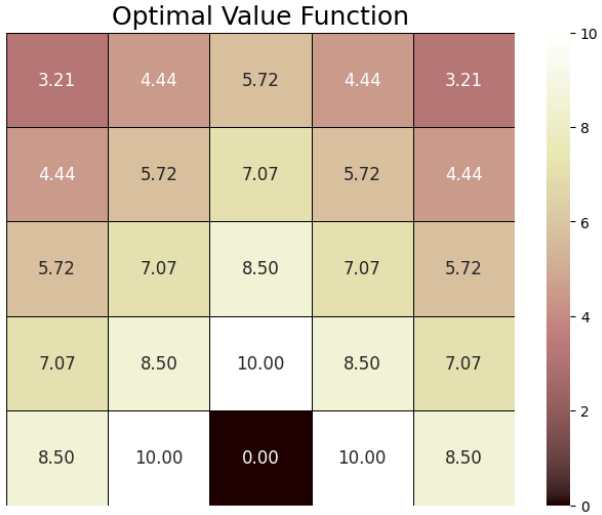
SMASH
machine learning for science and humanities postdoctoral program

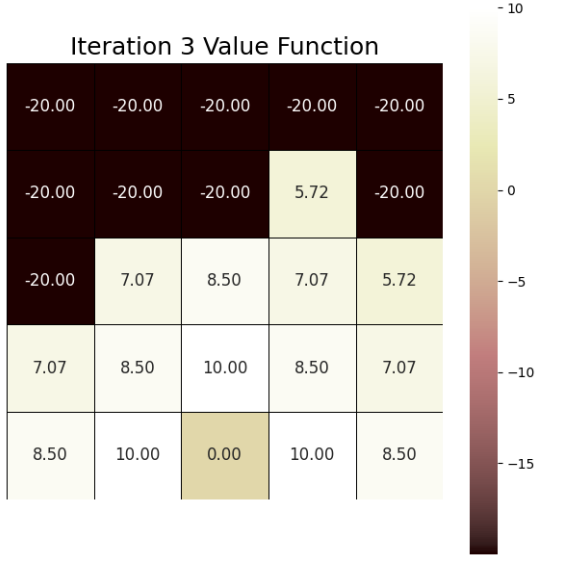


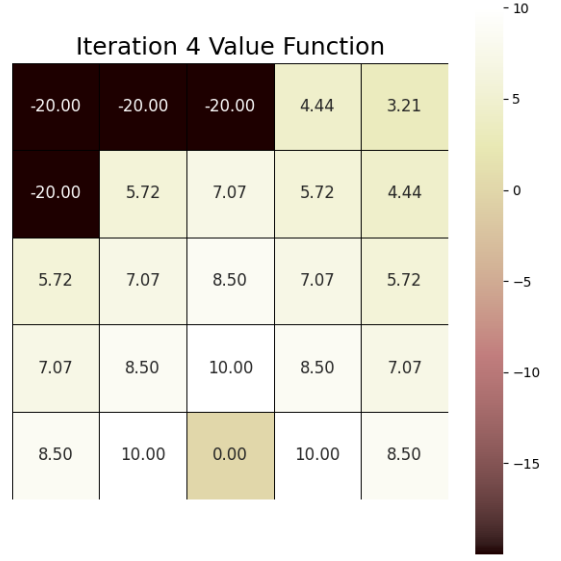


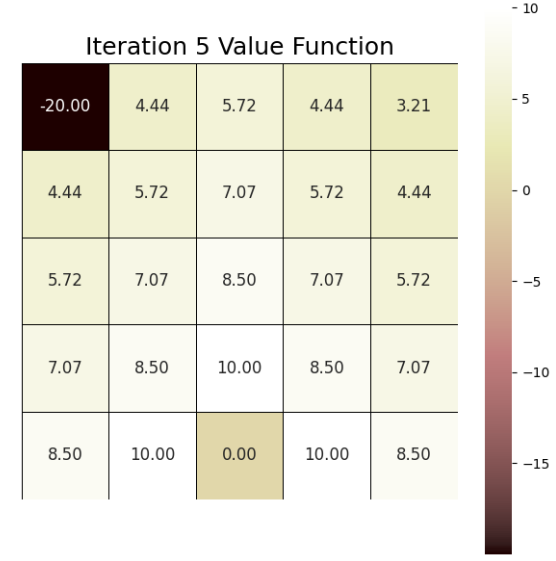












[GOOGLE COLAB LINK]

SMASH
machine learning for science and humanities postdoctoral program

Considerations for choosing RL Methods

Value-Based vs Policy Based vs Hybrid	**On-Policy vs Off-Policy**	**Model-Free vs Model-Based**	**Discrete vs Continuous**
Exploration vs Exploitation	**Stability vs Plasticity**	**Episodic vs Active**	**Stationary vs. Non-Stationary**

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

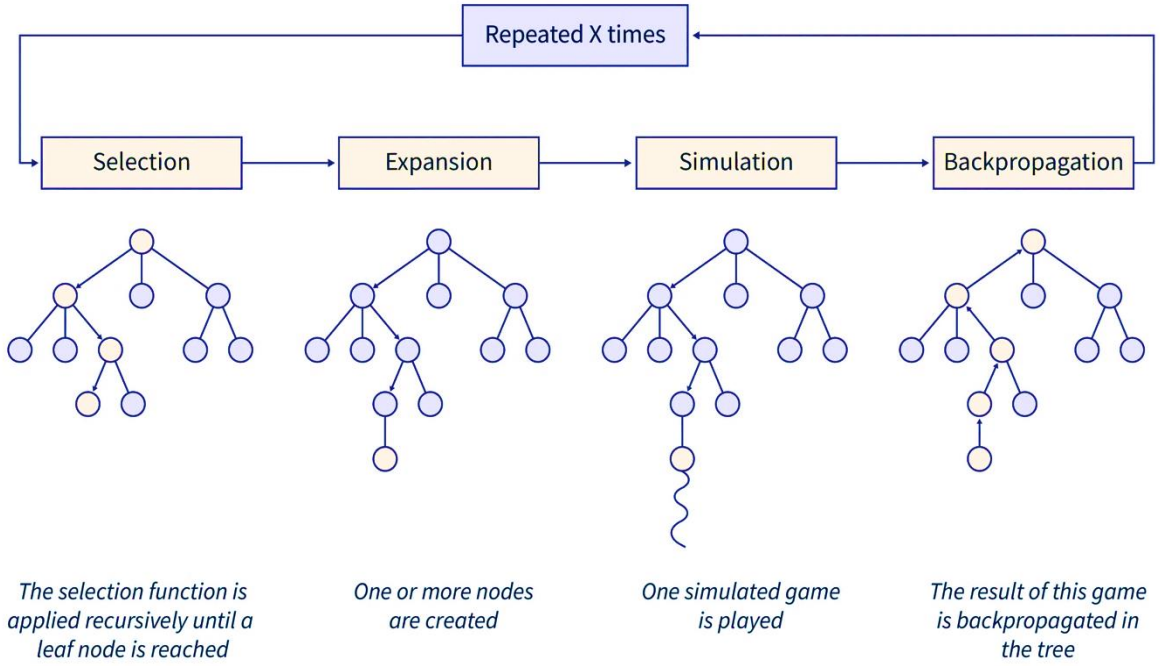
Considerations for choosing RL Methods

Algorithm	Category	On-Policy?	Model-Free?	Discrete Action?	Continuous Action?	Key Mechanism
Q-Learning	Value-Based	N	Y	Y	N	Tabular Q-estimates
SARSA	Value-Based	Y	Y	Y	N	Q(s,a) updates using next action
DQN	Deep Value	N	Y	Y	N	Replay Buffer & Target Network
Double DQN	Deep Value	N	Y	Y	N	Decouples selection from evaluation
REINFORCE	Policy Gradient	Y	Y	Y	Y	Monte Carlo policy updates
A2C / A3C	Actor-Critic	Y	Y	Y	Y	Synchronous/Asynchronous updates
TRPO	Policy Gradient	Y	Y	Y	Y	KL-Divergence trust regions
PPO	Policy Gradient	Y	Y	Y	Y	Clipped surrogate objective
DDPG	Actor-Critic	N	Y	N	Y	Deterministic Policy + Replay Buffer
TD3	Actor-Critic	N	Y	N	Y	Addresses overestimation in DDPG
SAC	Actor-Critic	N	Y	Y*	Y	Maximum Entropy framework
Dyna-Q	Hybrid	N	N	Y	N	Learns model + simulated planning
AlphaZero	Model-Based	N	N	Y	N	MCTS + Policy/Value Networks
PETS	Model-Based	N	N	N	Y	Probabilistic Ensembles with MPC
Dreamer (v1-3)	Model-Based	N	N	Y	Y	Latent World Model planning

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Monte Carlo Tree Search

MCTS is a heuristic search algorithm that builds a search tree asynchronously. It balances **exploration** (trying new moves) and **exploitation** (focusing on moves that yielded high rewards in simulations) to find the best move in complex, high-dimensional spaces without requiring a full state-space evaluation.



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Q-Learning

Q-Learning is an **off-policy** reinforcement learning algorithm that aims to find the best action to take given a current state. It is considered "off-policy" because it learns the value of the **optimal policy** independently of the agent's actual actions (which might include random exploration).

UPDATE RULE : $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$

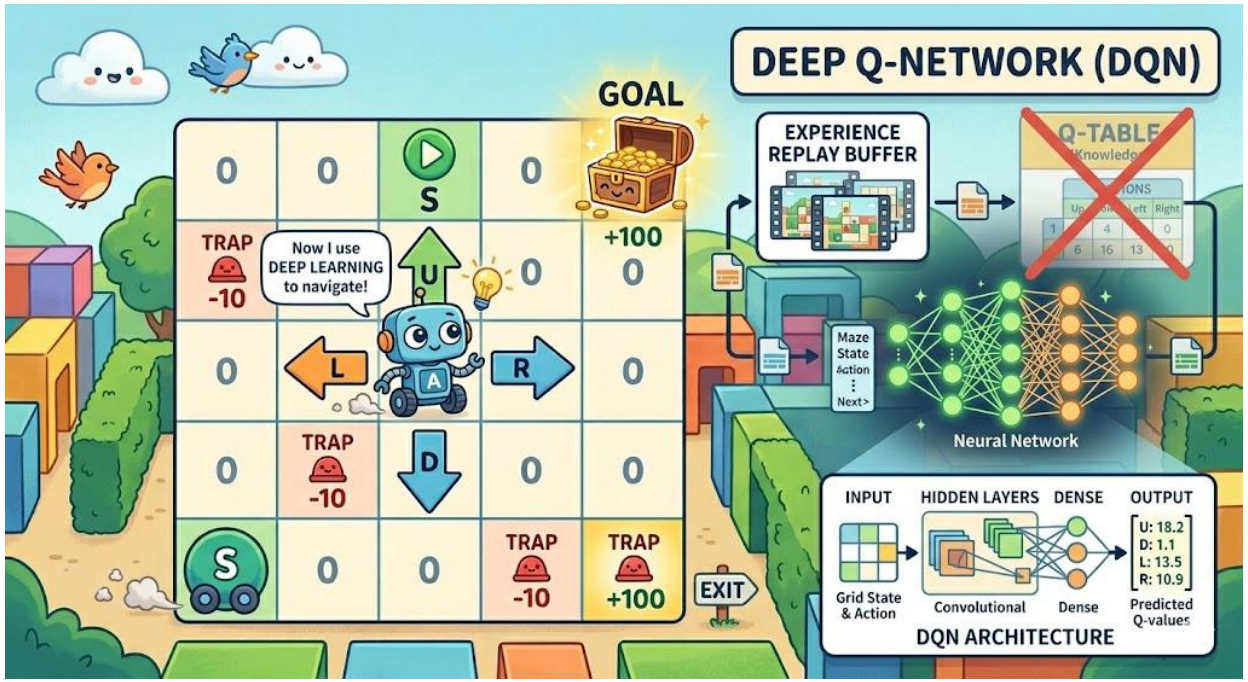
$\max_{a'} Q(s',a')$**:** The algorithm "searches" through all possible actions in the next state and picks the highest value.

α, γ : Learning rate and discount factor

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

DQN

DQN was the first algorithm to successfully combine **Q-Learning** with deep neural networks. It solves the "curse of dimensionality" by using a CNN or MLP to approximate the Q-table, allowing agents to handle complex, high-dimensional inputs (like raw pixels from **CARLA** or Atari games).



UPDATE RULE :

$$Q_{TARGET} = r + \gamma \max_{a'} Q(s', a'; \theta^{-}), \quad \text{\# based on frozen weights } (\theta^{-})$$

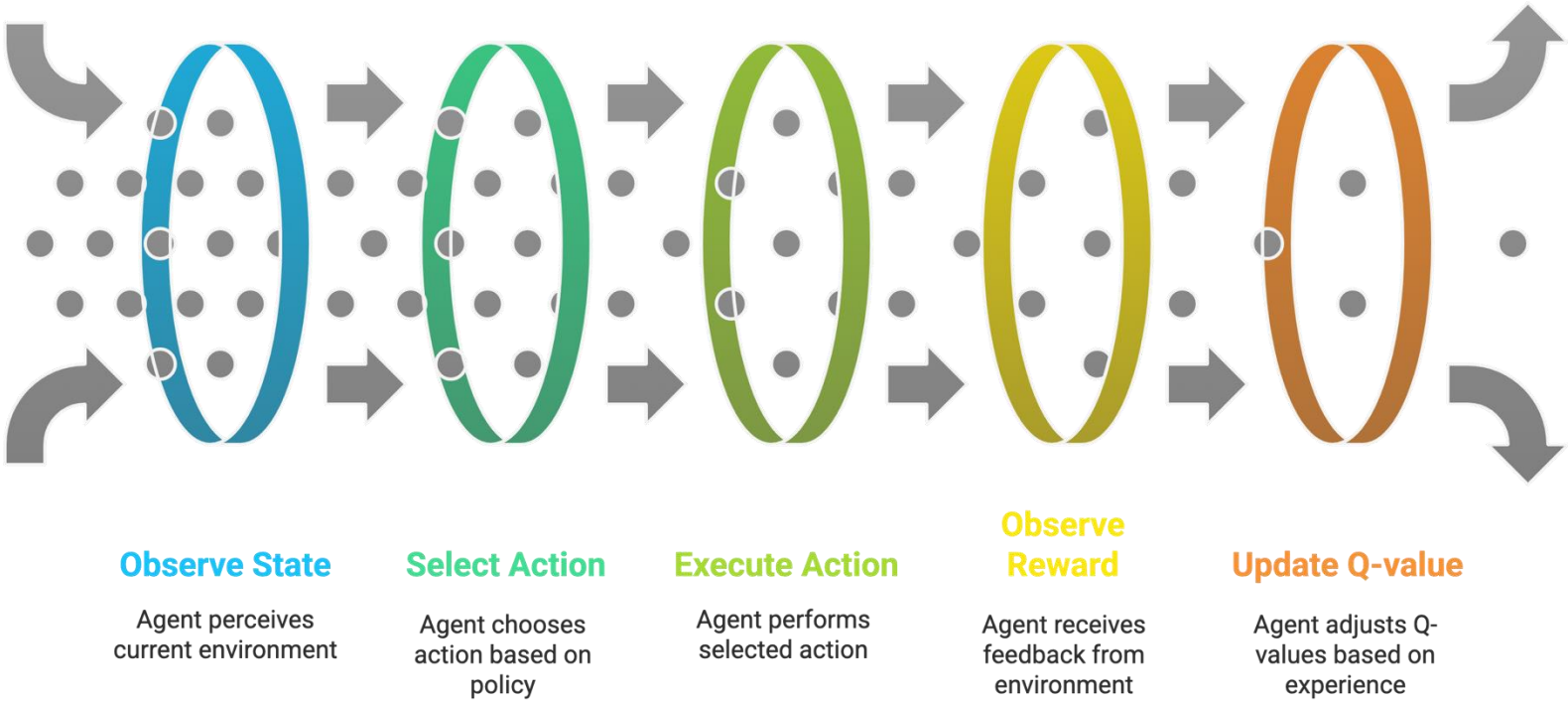
$$Q_{ACTUAL} = Q(s, a; \theta), \quad \text{\# based on current weights } (\theta)$$

$$L(\theta) = (Q_{TARGET} - Q_{ACTUAL})^2$$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

SARSA

SARSA is an **on-policy** reinforcement learning algorithm that estimates the value of a state-action pair (Q-value). Unlike Q-Learning, which assumes the agent will take the absolute best possible future action, SARSA learns based on the **actual** action the agent takes, including its mistakes or exploration steps.



UPDATE RULE : $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$

α : Learning rate.
γ : Discount factor for future rewards.
$Q(s',a')$: The value of the action actually chosen for the next step.

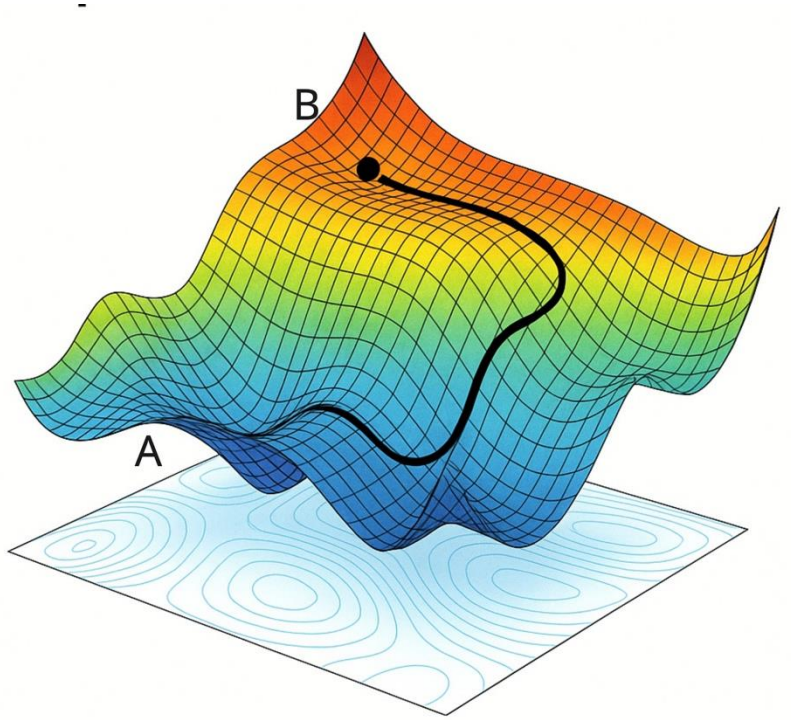
I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

REINFORCE – Policy Gradient

REINFORCE is a fundamental **Policy Gradient** algorithm that maps states directly to actions without using a Q-table. It updates the weights of a neural network θ by increasing the probability of actions that lead to high cumulative rewards and decreasing the probability of those that lead to poor outcomes.

UPDATE RULE : $\Delta\theta = \alpha \cdot G_t \cdot \nabla_\theta \ln \pi_\theta(a_t|s_t)$

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$$



G_t : The "Return" (cumulative reward). This acts as a **scalar signal**:
if G_t is high, the gradient is pushed strongly;
if G_t is negative, the action becomes less likely.

$\nabla_\theta \ln \pi_\theta(a_t|s_t)$: The "**Score Function.**" It tells us the direction in which to move the network weights to increase the probability of action a_t.

α : Learning rate.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Actor-Critic Learning

The agent is divided into two distinct functional components generally two neural networks:
Actor Network - $\mathcal{N}_\theta$, **Critic Network** - $\mathcal{N}_\phi$
often sharing a common feature-extraction backbone

In **REINFORCE** the iterative update was dependent on **Cumulative Reward**

$$\boldsymbol{G_t} = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$$

In **Actor-Critic** the iterative update is dependent on **Advantage** $\widehat{\boldsymbol{A_t}} = G_t - V_\phi(s_t)$
where, $V_\phi(s_t)$ is the estimated value of the state.

ACTOR (Policy-Based)
Role: Learn Optimal Policy π_θ
Output: Action Probabilities $\pi(\mathrm{a}|\mathrm{s})$
Weight Update: $\Delta\theta = \alpha \cdot \widehat{A_t} \cdot \nabla_\theta \ln \pi_\theta(a_t|s_t)$

CRITIC (Value-Based)
Role: Learns to estimate the value of states.
Output: A scalar value $V(s)$: expected cumulative reward.
Weight Update: $\Delta\phi = \beta \cdot \widehat{A_t} \cdot \nabla_\phi V_\phi(s_t)$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH machine learning for science and humanities postdoctoral program

PPO – Proximal Policy Optimization

Standard Actor-Critic updates can be too aggressive. If the Advantage $\widehat{A_t}$ is large, the weight update can move the policy π_θ into a "collapsed" state from which the agent cannot recover.
Instead of the log-probability gradient, PPO tracks how much the **new** policy deviates from the **old** policy

Probability Ratio $$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

Clipped Probability Ratio

$$r_t^{clipped}(\theta) = \min\left(\max(r_t(\theta), 1-\epsilon), 1+\epsilon\right)$$

This always between $(1-\epsilon\ , 1+\epsilon)$

Generalized Advantage Estimation $\left(\widehat{A_t}^{GAE}\right)$

$$\widehat{A_t}^{GAE} = \delta_t + (\gamma\lambda)\delta_{t+1} + (\gamma\lambda)^2\delta_{t+2} + \cdots$$

where, $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$

ACTOR Weight Update :

$$\Delta\theta = \alpha \cdot \nabla_\theta \min\left(r_t(\theta) \cdot \widehat{A_t}^{GAE}, r_t^{clipped}(\theta) \cdot \widehat{A_t}^{GAE}\right)$$

CRITIC (Value-Based)

Weight Update: $\Delta\phi = \beta \cdot \widehat{A_t} \cdot \nabla_\phi V_\phi(s_t)$

where, $\widehat{A_t} = G_t - V_\phi(s_t)$,

and, $G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH machine learning for science and humanities postdoctoral program

Variants

- **Monte Carlo Tree Search (MCTS)** — *Look-ahead Search*
  - **AlphaZero:** Combines MCTS with Neural Networks for intuition.
  - **MuZero:** Learns the environment rules while searching.
- **Q-Learning / DQN** — *Value-Based*
  - **Double DQN:** Fixes "over-optimism" in reward estimation.
  - **Dueling DQN:** Separates state value from action advantage.
  - **Rainbow DQN:** An ensemble of seven major DQN improvements.
- **SARSA** — *On-Policy Learning*
  - **Expected SARSA:** Improves stability by averaging next-step actions.
  - **SARSA(λ):** Bridges the gap between short-term and long-term rewards.
- **REINFORCE:** The foundation; optimizes the policy via trial and error.
  - **TRPO (Trust Region Policy Optimization)** Uses KL-divergence to keep updates within a "safe" range.
- **PPO:** Focus on small updates
  - **PPO-Lagrangian:** Uses a Lagrangian multiplier to enforce "hard" constraints
- **Actor-Critic** — *The Hybrid Architecture*
  - **A3C / A2C:** Runs multiple agents in parallel for faster data collection.
  - **DDPG / TD3:** The go-to for **continuous** control (e.g., robotic joints).
  - **Soft Actor-Critic (SAC):** Maximizes "Entropy" to prevent the agent from getting stuck in one strategy.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

The System – Vehicular Autonomy

The "Physical" system and the traditional software stack

- SAE
- Levels of Autonomy
- SENSE
- PLAN
- ACT
- SCOPE

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

The Society of Automotive Engineers

AE International (formerly the Society of Automotive Engineers) is a global association of more than 128,000 engineers and related technical experts in the aerospace, automotive, and commercial-vehicle industries.

Core Mission & Influence

- **Standards Development:** Maintains over 10,000 active standards that ensure safety, quality, and interoperability across global transport sectors.
- **Professional Development:** Provides lifelong learning resources, certifications, and technical publications for mobility professionals.
- **STEM Outreach:** Encourages future talent through programs like the Collegiate Design Series (e.g., Formula SAE).

I FEEL SLOVENIA
Co-funded by the European Union
SMASH machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

“On road” refers to publicly accessible roadways (including parking areas and private campuses that permit public access) that collectively serve all road users, including cyclists, pedestrians, and users of vehicles with and without driving automation features.

- The (human) user,
- The driving automation system, and
- Other vehicle systems and components.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

DDT (Dynamic Driving Task) ✓

The DDT is generally split into two main categories of functions:

Operational Functions (Physical Control):

- **Lateral Motion Control:** Steering the vehicle.
- **Longitudinal Motion Control:** Accelerating and braking (managing speed).

Tactical Functions (Planning and Response):

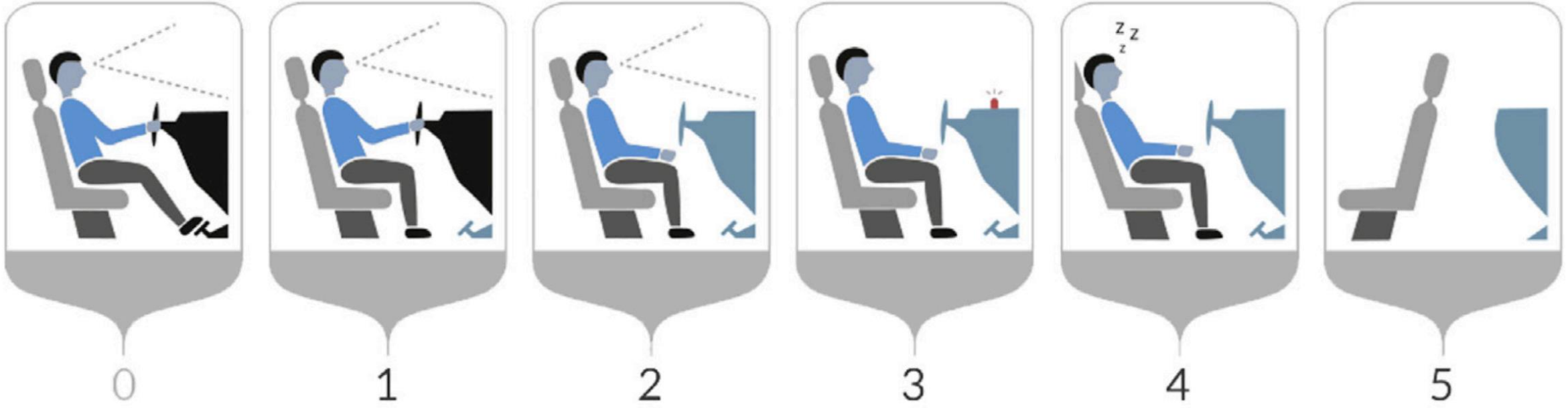
- **OEDR (Object and Event Detection and Response):** Monitoring the environment (detecting pedestrians, other cars, traffic lights) and reacting to them.
- **Maneuver Planning:** Deciding to change lanes, turn at an intersection, or signal a turn.
- **Enhancing Conspicuity:** Using lights, signals, or the horn to communicate with other road users.

✗

Trip Scheduling: Deciding what time to leave.
Destination Selection: Deciding where to go.
Waypoint Selection: Choosing which route to take (e.g., "Take the highway instead of the scenic route").

I FEEL SLOVENIA
Co-funded by the European Union
SMASH machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

Level 0: No Driving Automation

The human driver performs the entire DDT. There is no sustained control of steering or acceleration by the vehicle.
System Role: May provide momentary alerts or emergency intervention.
Example: A 2005 sedan with traditional cruise control (which only holds speed) or a car with blind-spot warnings and emergency braking.

Level 1: Driver Assistance

The system can handle **either** steering or acceleration/braking, but not both at the same time. The driver must remain fully engaged.
System Role: Assists with one specific task.
Example: Adaptive Cruise Control (handles speed but you steer) or **Lane Keep Assist** (nudges steering but you handle pedals).

Level 2: Partial Driving Automation

The system handles **both** steering and acceleration/braking simultaneously. However, the driver is still "driving" and must supervise the system constantly.
System Role: Combined lateral and longitudinal control.
Example: Tesla Autopilot, **GM Super Cruise**, or **Ford BlueCruise**. The car stays in the lane and maintains distance, but the human must keep eyes on the road.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

Level 3: Conditional Driving Automation

This is the "tipping point." The system performs the **entire DDT**, including monitoring the environment (OEDR). The human is no longer "driving," but must be ready to take over if the system requests it.
System Role: The car drives itself under specific conditions (e.g., traffic jams on highways).
Example: Mercedes-Benz DRIVE PILOT. In certain slow-moving highway traffic in Germany or Nevada/California, the driver can legally look away from the road to watch a movie until the system beeps.

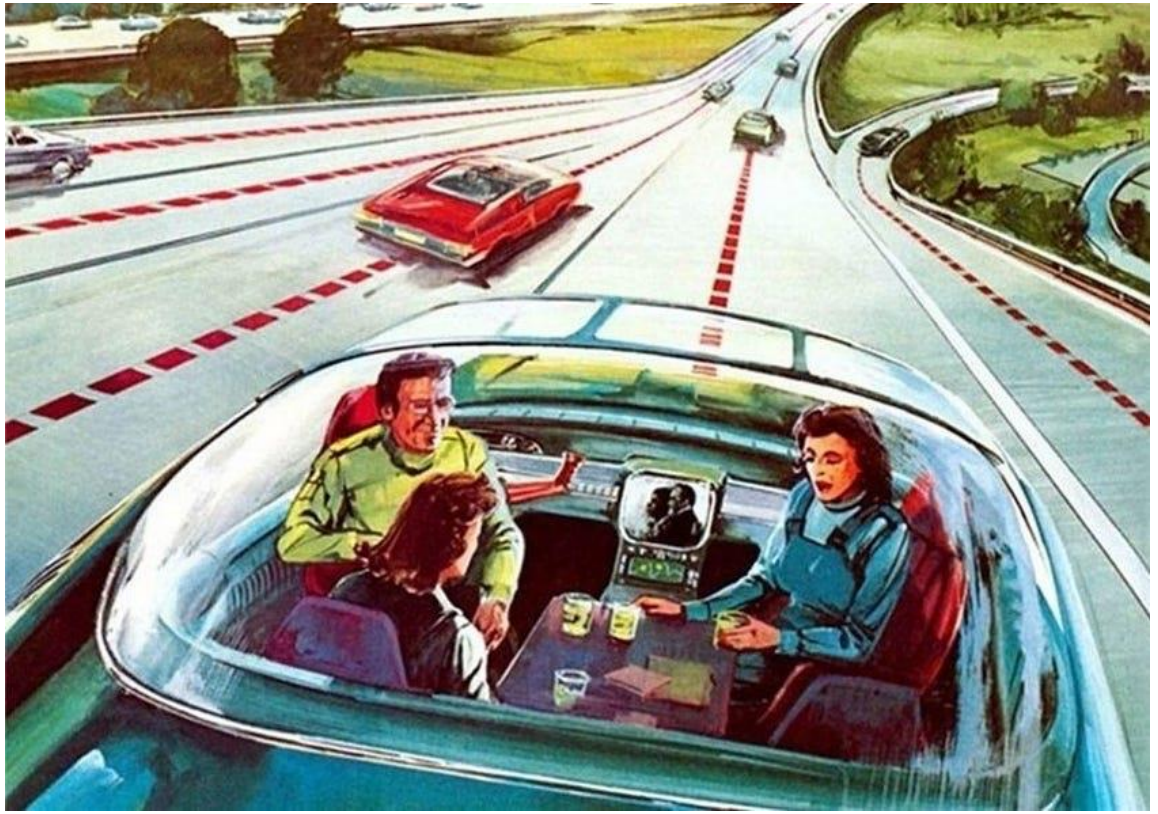
Level 4: High Driving Automation

The system performs the entire DDT and can handle its own "fallback" if something goes wrong. It does not require a human to take over. However, it is limited to a specific area (geofencing) or specific weather conditions.
System Role: Fully autonomous within a restricted domain.
Example: Waymo or **Cruise** robotaxis operating in specific cities like Phoenix or San Francisco. They have no driver, but they cannot drive outside their mapped territory.
Level 5: Full Driving Automation

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

Level 5: Full Driving Automation
The system can perform all driving tasks under all conditions that a human driver could manage. There are no geographical or environmental restrictions.
System Role: Operates everywhere, anytime, with no human intervention needed.
Example: Theoretical only. No production vehicle or service currently meets Level 5. A vehicle at this level might not even have a steering wheel or pedals.



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

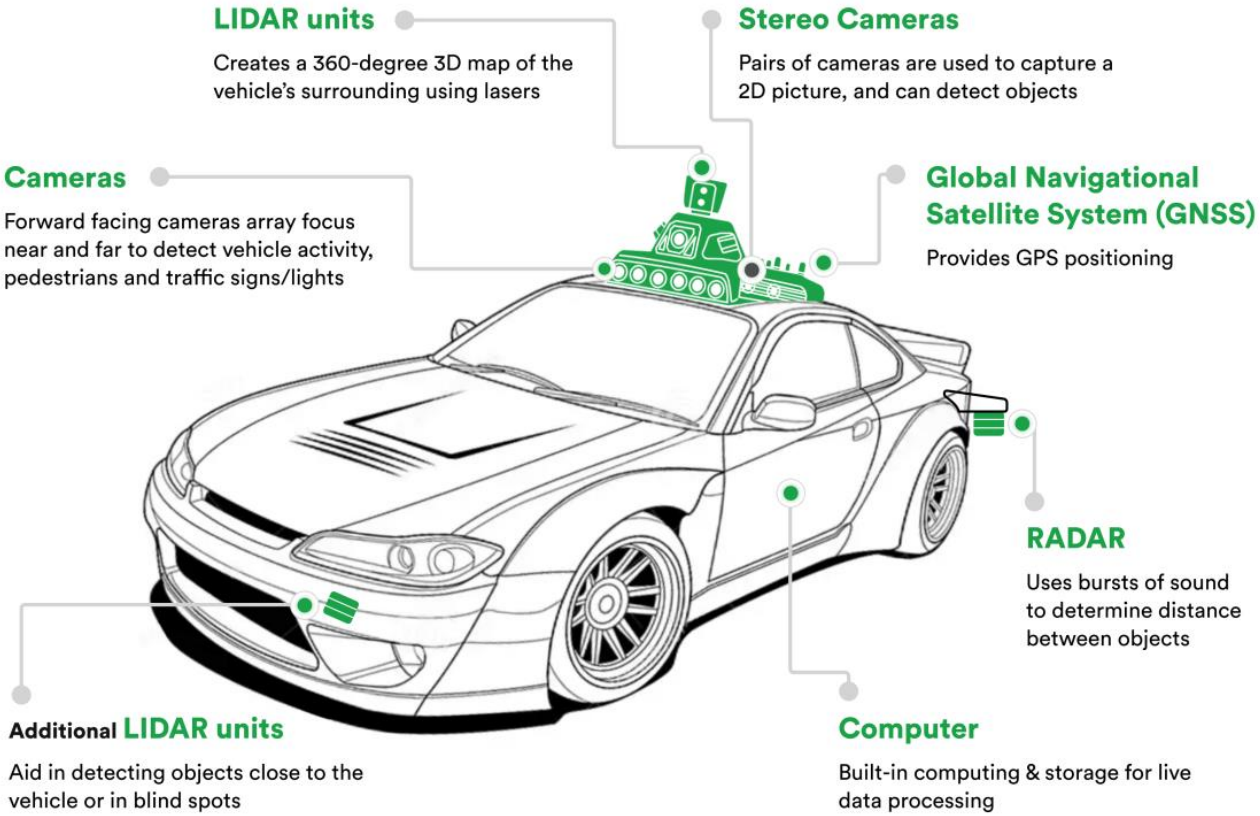
The Anatomy of an Autonomous Vehicle Hardware Stack

The Three Pillars:

- **Sensing (Exteroceptive):** Gathering environmental data.
- **Processing (Compute):** The AI brain for perception, planning, and localization.
- **Actuation (Control):** Converting digital commands into physical movement.

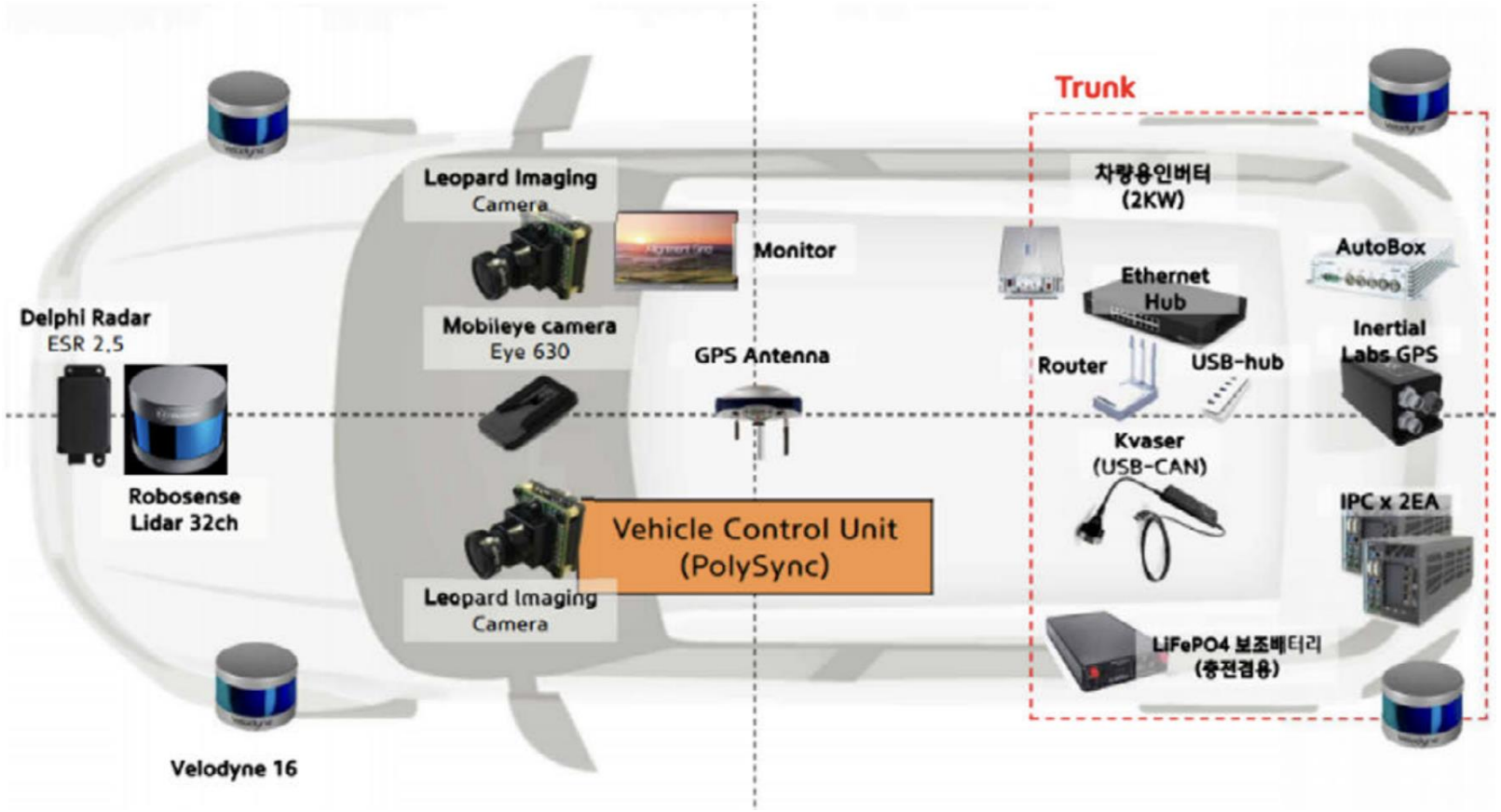
System Concept: A real-time, modular **Sense–Plan–Act** pipeline.

Connectivity: The "Centralized Electronic Architecture" (replacing hundreds of small ECUs with a few high-performance controllers).



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

The Anatomy of an Autonomous Vehicle Hardware Stack



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

SENSE – The Perception Layer

Goal: To convert the physical environment into digital data through a suite of diverse sensors.

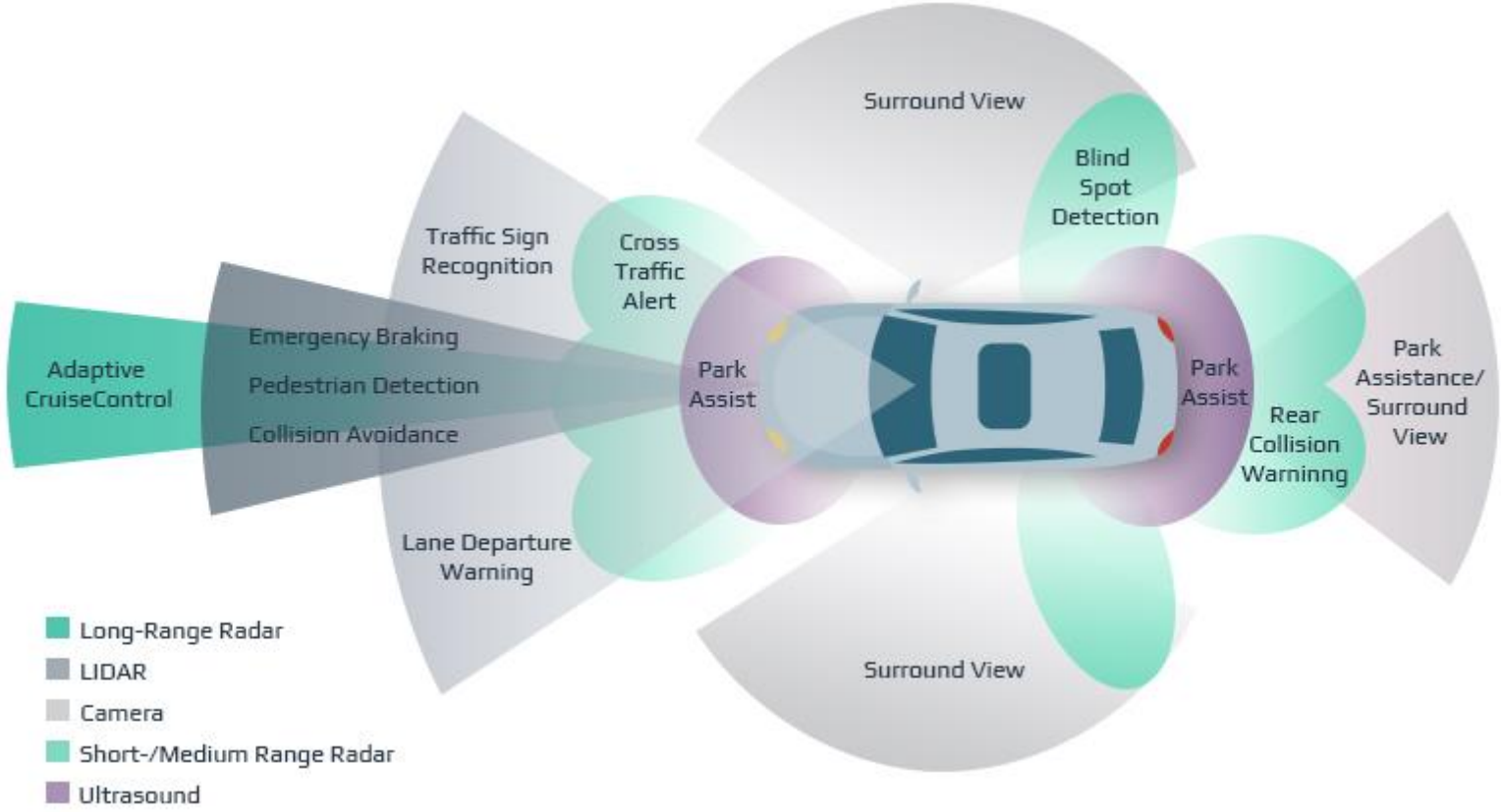
Exteroceptive Sensors (External Environment):

- **LiDAR:** Uses laser pulses to create a high-density 3D point cloud; essential for precise distance and shape detection.
- **Cameras (Vision):** Captures texture and color; critical for reading traffic signs, lane markings, and traffic light states.
- **Radar:** Uses radio waves to detect object distance and velocity; excels in adverse weather (fog, rain) where others might fail.
- **Ultrasonic:** Short-range sensors for low-speed maneuvers like parking or blind-spot detection.

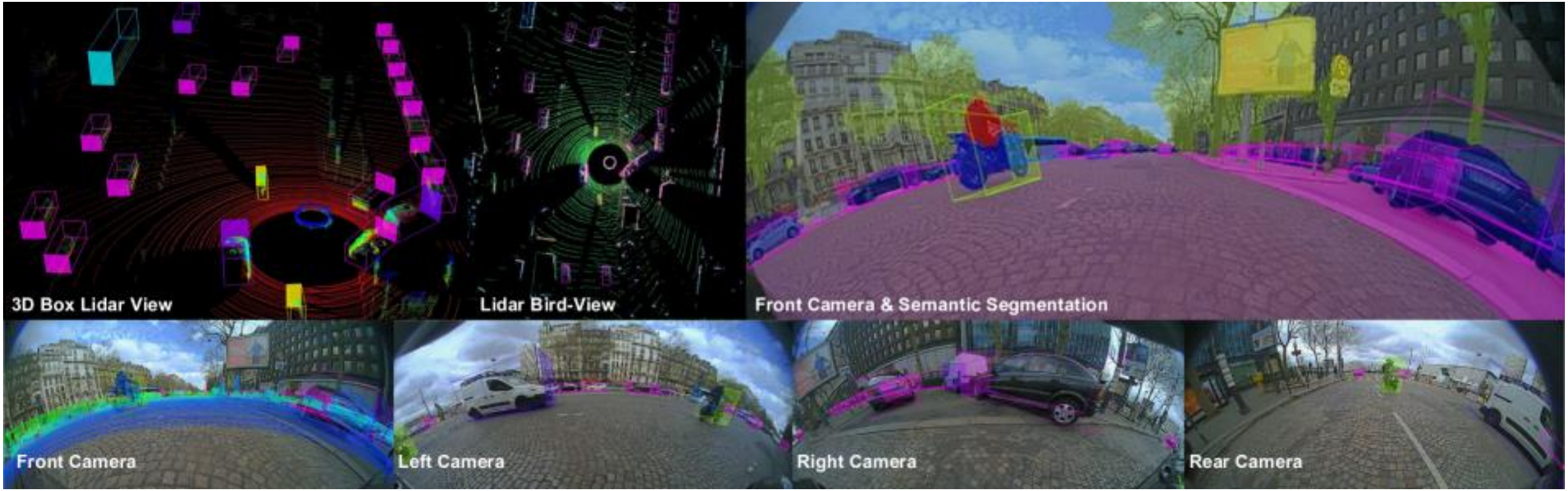
Proprioceptive Sensors (Internal State):

- **GNSS/GPS:** Provides absolute global positioning.
- **IMU (Inertial Measurement Unit):** Tracks the vehicle's acceleration and orientation (pitch, roll, yaw).

Connectivity (V2X): Hardware for communicating with infrastructure or other vehicles to "see" beyond line-of-sight.



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program



I FEEL SLOVENIA

Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

PLAN – The Intelligence Layer

Goal: To process sensor data, localize the vehicle, and determine the safest, most efficient path.

High-Performance Computing (AI Brain):

- **GPUs & TPUs:** Dedicated hardware accelerators (e.g., NVIDIA DRIVE, Tesla FSD Computer) designed for massive parallel processing of neural networks.
- **FPGA/ASIC:** Custom silicon optimized for low-latency inference and energy efficiency.

Central Computing Architecture:

- Moving from distributed ECUs (Electronic Control Units) to a **Centralized Domain Controller** that handles sensor fusion (combining LiDAR, camera, and radar data).

Redundancy & Safety:

- Secondary "fail-operational" processors that can safely pull the car over if the primary computer fails.
- Real-Time Operating Systems (RTOS) that guarantee the computer responds to hazards within milliseconds.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

ACT – The Control Layer

Goal: To translate digital commands from the "Plan" stage into physical movement via mechanical actuators.

Drive-by-Wire Systems:

- **Steer-by-Wire:** Replaces the mechanical steering column with electrical motors that turn the wheels based on digital signals.
- **Brake-by-Wire:** Uses electronic actuators to apply hydraulic pressure to the brakes, allowing for smoother and faster automated deceleration.
- **Electronic Throttle Control:** Manages engine torque or electric motor output for acceleration.

Human-Machine Interface (HMI):

- Displays and haptic feedback systems (like steering wheel vibrations) to alert the driver when they need to resume control (especially at SAE Levels 2 and 3).

Communication Bus (CAN/Ethernet):

- The high-speed physical "nervous system" (Automotive Ethernet or CAN FD) that carries signals from the computer to the actuators with zero lag.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Scope for Informatics Students

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

The Bridge – RL in AV

Implementation, Simulation, and Training

- RL for AV
- Optimal RL Configurations
- Control Hierarchies
- Observation Space
- Action Space
- Reward Modelling
- Simulations
- Tools

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Why RL for Autonomous Driving

Traditional robotics relies on **explicit programming**, but driving is an **implicit negotiation**.

- **Handling the "Long Tail" of Edge Cases:**
  - Classic algorithms struggle with rare, unpredictable scenarios (e.g., a distracted pedestrian, erratic lane changes).
  - RL generalizes better by learning *policies* rather than hard-coded rules.
- **Adaptive Decision Making in Dynamic Traffic:**
  - Driving isn't just about lane keeping; it’s about interacting with other "agents."
  - RL is naturally suited for multi-agent environments where the ego-vehicle must predict and react to the intent of others.
- **Optimization of Conflicting Objectives:**
  - Drivers constantly balance **Safety**, **Efficiency**, and **Comfort**.
  - Through the **Reward Function**, RL can find a mathematical "sweet spot" between these trade-offs that is difficult to tune manually.
- **Reduction of Engineering Complexity:**
  - Replaces thousands of "if-else" statements with a unified learning framework.
  - Reduces the need for hand-engineered heuristics in motion planning

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Why RL for Autonomous Driving

We don't want to tell the car *how* to drive

We want to tell it what a *good* driving looks like

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Classification of RL Methods

- World Knowledge
  - Model-based – MCTS, Dyna-Q, WorldModels
  - Model-Free - Q-Learning, SARSA, PPO.
- Optimization Method
  - Value-Based – Q-Learning, DQN
  - Policy-Based – REINFORCE, Policy Gradient
  - Hybrid – ActorCritic, PPO, A3C, SAC
- Nature of Learning
  - On-Policy – SARSA, PPO, SAC
  - Off-Policy - Q-Learning, DQN, SAC, DDPG.
- Scope
  - Tabular - Classic Q-Learning, SARSA.
  - Deep RL - DQN, PPO, SAC

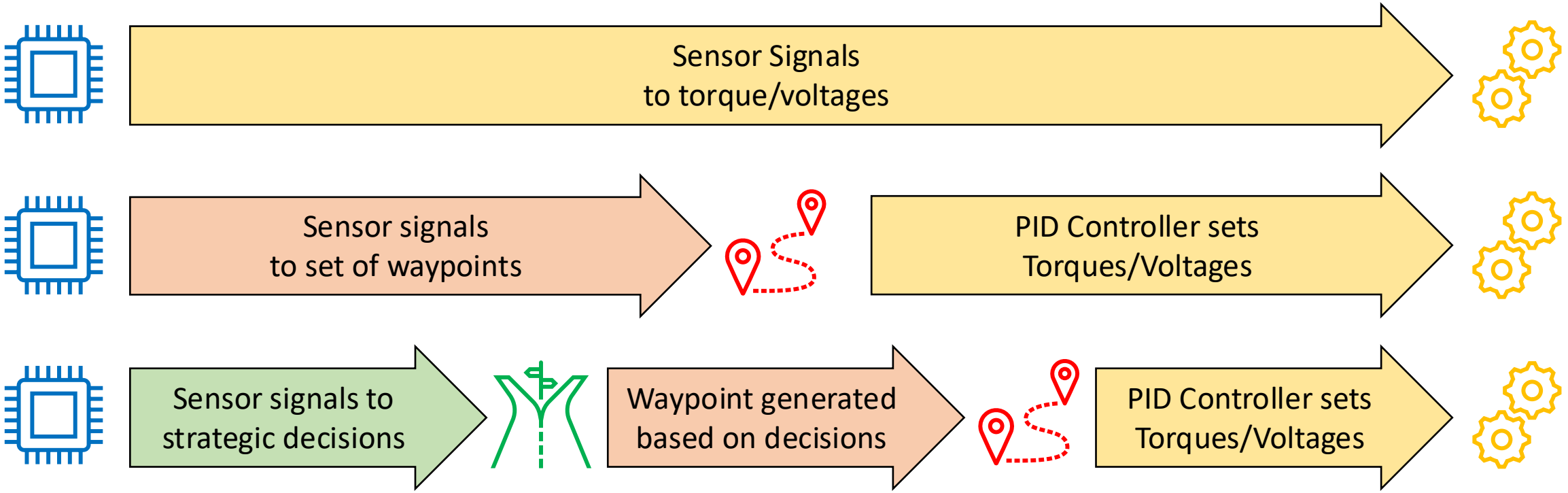
I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

How to choose ? RL for Autonomous Driving ?

- Value-based vs Policy based vs Actor-Critic
- On-Policy vs Off-Policy
- Discrete vs Continuous
- Model-Free vs Model-based
- Exploration vs Exploitation
- Stability vs Plasticity
- Episodic vs Continuous
- Stationary vs Non-Stationary

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Control Hierarchies in RL for Driving



I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

OBSERVATION SPACE

Raw Sensor Information

- RGB Sensors
- LIDARs
- RADARs
- IMU/GNSS/GPS etc

Derived Information

- Road Markings
- Pedestrian Position & Vectors
- Vehicle Position & Vectors
- Self Position & Vectors
- Anticipated Future Positions
- Drivable vs Non-Drivable Area
- Traffic Signals and Signs

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

ACTION SPACE

The Control Hierarchy
The action space determines how the agent interacts with the environment. In AV research, we generally categorize these into two levels:

Continuous Action Space (The Standard)
- **Actions:** Throttle $[0, 1]$, Brake $[0, 1]$, and Steering $[-1, 1]$.
- **Pros:** Allows for high-precision, smooth driving maneuvres.
- **Cons:** Harder to explore during training; the agent might "vibrate" the steering wheel initially as it learns.

Discrete Action Space (The Simplified)
- **Actions:** Preset modes: [Stay in Lane, Lane Change Left, Lane Change Right, Accelerate, Brake].
- **Pros:** Simplifies the learning problem significantly; great for high-level tactical decision-making.
- **Cons:** Results in "jerky" movements; cannot handle fine-grained emergency steering.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

ACTION SPACE

Low-Level vs. Mid-Level Control

- **Direct Control (Low - Level):** The RL agent outputs raw torque/voltage commands. This requires the agent to learn basic physics (like friction) alongside traffic rules.
- **Waypoint Tracking (Mid - Level):** The RL agent outputs a **Target Waypoint** or a **Target Speed**, which is then fed into a traditional controller (like a **PID** or **MPC**).
- **Strategic Decision Making (High - Level):** Agent chooses from a set of commands [Stay in Lane, Lane Change Left, Lane Change Right, Accelerate, Brake]

Safety Constraints

- **Action Clipping:** Ensuring the agent doesn't "snap" the steering wheel 90 degrees at 100 km/h.
- **Temporal Consistency:** Penalizing the agent for rapid, oscillating changes in action to ensure passenger comfort.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Reward Modelling

Episodic Rewards	Episodic Penalties
Reach Destination	Collision

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Reward Modelling

Episodic Rewards Reach Destination	**Episodic Penalties** Collision
Immediate Rewards Progress to Destination Compliance with Rules Withing Contraints Good Behaviour Smooth Driving	**Immediate Penalties** Missing Waypoints Breaking Rules Overspeeding Unnecessary Behaviours Rough Jerky Movements

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Handling Sparse Rewards

Problem: The vehicle only gets a "goal" reward at the end of a long drive.
Solution: Reward Shaping – providing incremental rewards for staying in the lane and moving forward.

Feature	Long-Term Episodic Reward	Short-Term / Immediate Reward
Timing	Delivered only at the end of the episode.	Delivered at every single time step (t).
Objective	Mission success or failure.	Behavioral alignment and safety.
Example 1	**Completion:** +100 for reaching the final GPS destination safely.	**Progress:** +0.5 for every meter traveled along the center of the lane.
Example 2	**Survival:** +50 for finishing a 5-minute drive without any collisions.	**Safety:** -1.0 for every frame where the car is too close to a pedestrian (Time-to-Collision).
Example 3	**Efficiency:** A bonus based on the total time taken for the whole trip.	**Comfort:** -0.1 for high lateral G-forces or sudden "jerky" braking.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Reward Shaping – Positive Rewards

Reward Category	Specific Signal	Description	Example Reward Calculation
Mission Success	**Reaching Destination**	The primary goal; delivered only upon completing the entire planned route.	$R_{final} = +1000$
Progress	**Reaching Waypoints**	Interim bonuses for staying on the intended path.	$R_{waypoint} = +50\ (upon\ passing\ a\ waypoint)$
Progress	**Reduced Distance to Target**	A continuous signal based on Euclidean or road distance reduction.	$R_{distance} = \alpha \times (d_{t-1} - d_t)\ (if\ d_t < d_{t-1})$
Compliance	**Responding to Signals**	Correct behavioral response to dynamic traffic elements.	$R_{compliance} =$ $+10\ (for\ each\ step\ stopped\ at\ a\ red\ light)$
Constraint	**Maintaining Target Speed**	Rewards keeping pace with the flow of traffic within legal limits.	$R_{speed} = \beta \times e^{-(v_{target} - v_{current})^2}$
Behavioral	**Maintaining Lane Center**	Promotes stable, predictable, and safe lateral positioning.	$R_{lane} = +1.0\ \ if\ (\lvert car_{center} - lane_{center} \rvert < t)$
Comfort	**Lower G-Forces**	Encourages smooth, non-aggressive acceleration, braking, and steering.	$R_{comfort} = \gamma \times (G_{safe} - G_{current})$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Reward Shaping – Negative Rewards or Penalties

Penalty Category	Specific Signal	Description	Example Penalty Calculation
Efficiency	**Time Consumed**	A continuous penalty to encourage the agent to find the quickest route.	$P_{time} = -0.1\ (per\ unit\ time)$
Safety	**Collision**	A high negative reward for if vehicle collides with another object	$P_{collission} = -1000$
Navigation	**Missing Waypoints**	A major penalty for straying off the designated route or missing a turn.	$P_{route} = -500$
Navigation	**Increased Distance from Target**	A continuous penalty when the agent moves away from its goal.	$P_{distance} = \delta \times (d_t - d_{t-1})$
Compliance	**Failing Traffic Signals or Crossing Speed Limit**	A significant safety penalty for running red lights, ignoring stop signs, overspeeding etc.	$P_{violation} = -200\ or$ $P_{overspeed} = \epsilon \times (speed_{limit} - speed_{current})$
Behavioral	**Non-mandatory Lane-Changing**	Minimizes unnecessary lane-weaving, which increases collision risk.	$P_{lane\_change} =$ $-50\ (per\ illegal\ lane\ change)$ or $= -2\ (legal\ but\ non-mandatory)$
Comfort	**Higher G-Forces or Jerks**	Penalizes aggressive inputs that degrade passenger comfort.	$R_{comfort} = \gamma \times (G_{safe} - G_{current})$

I FEEL SLOVENIA
Co-funded by the European Union
SMASH machine learning for science and humanities postdoctoral program

The Role of High-Fidelity Simulation

- **Safety (Zero-Risk Training):** Training an RL agent involves thousands of crashes. In a simulator, a "collision" is just a reset button; in reality, it is a catastrophic liability.
- **Cost Efficiency:** Maintaining a fleet of sensor-heavy vehicles is expensive. Running 100 parallel instances of a simulator on a server cluster costs a fraction of the price.
- **Time Scaling:** Simulators can run **faster than real-time**. While a human driver is limited by the 24-hour day, an agent can "experience" years of driving data in a matter of weeks.
- **Edge Case Generation:** It is nearly impossible to find a specific triple-lane-change accident in the real world on demand. In simulation, you can script "The Long Tail" of rare, dangerous events repeatedly until the agent masters them.
- **Curriculum Learning:** Start the agent in a very simple environment (e.g., a straight, empty road with the goal only 5 meters away). As the agent masters the simple task, gradually increase the distance and complexity (traffic, curves, intersections). This ensures the agent "hits" the reward frequently enough to start learning.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

Core Requirements for High Fidelity

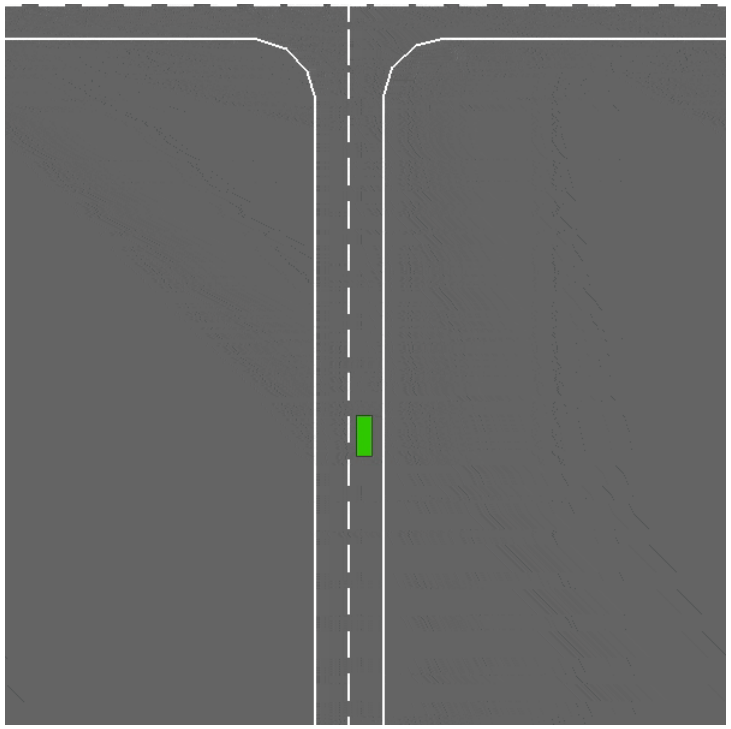
To ensure **Sim-to-Real Transfer**, where a model trained in simulation actually works on a real car, the environment must meet three criteria:

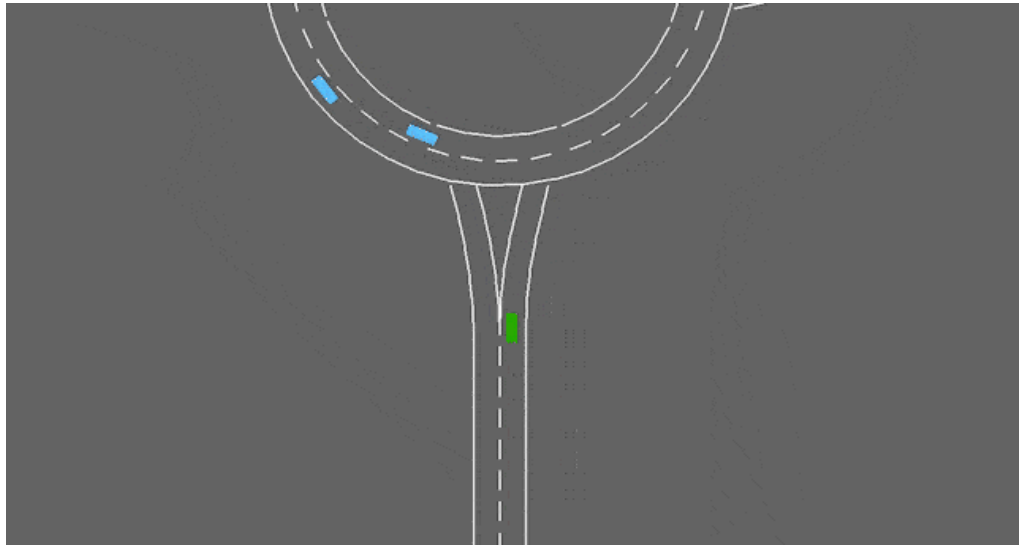
Requirement	Description
Realistic Physics	Must accurately model tire-road friction, mass distribution, aerodynamics, and inertia. If the "math" of the car is wrong, the RL agent will learn "cheats" that don't work in reality.
Photorealistic Rendering	Crucial for agents using **Camera/Vision sensors**. The lighting, shadows, and textures must mimic the real-world so the Neural Network generalizes to actual sunlight and rain.
Programmable Traffic	The environment cannot be static. It requires dynamic "Actors" (pedestrians, cyclists, and other cars) with varied behaviours to test the agent's tactical decision-making.

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

SOME TOOLS

https://github.com/Farama-Foundation/HighwayEnv





I FEEL SLOVENIA

Co-funded by the European Union

SMASH
machine learning for science and humanities postdoctoral program

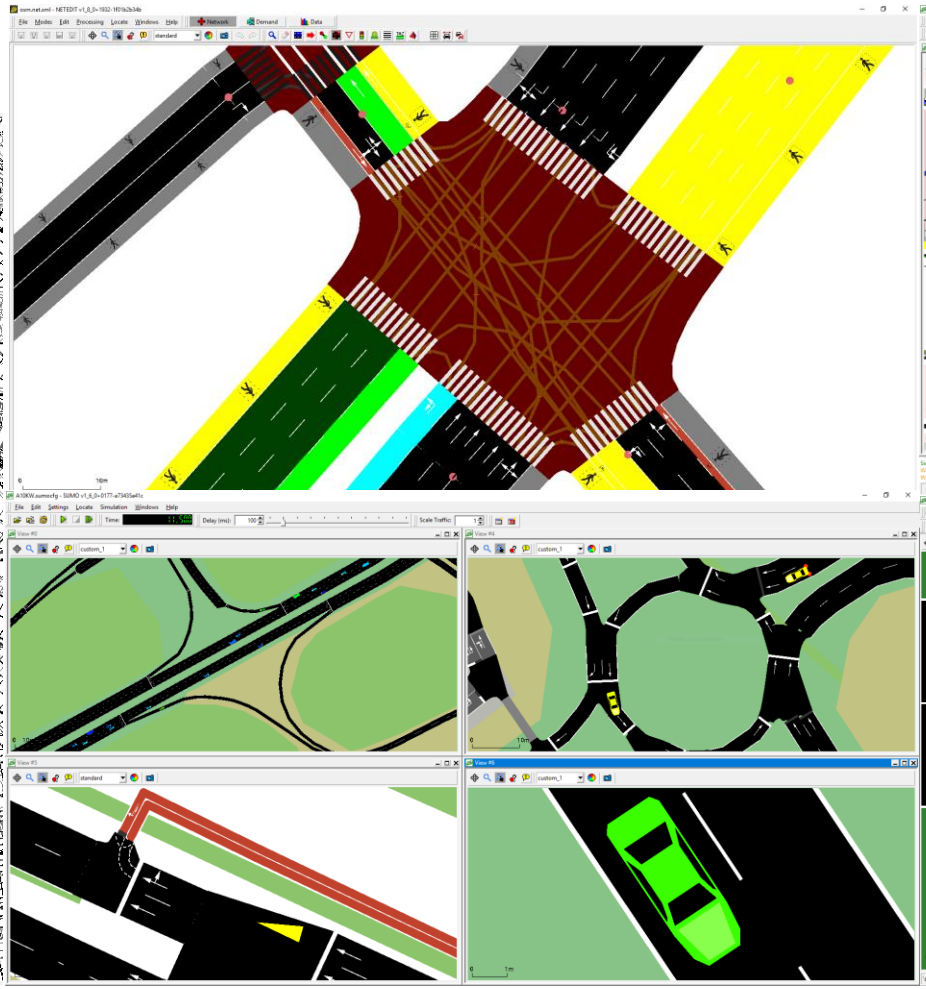
SOME TOOLS

https://carla.org/

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

SOME TOOLS

https://eclipse.dev/sumo/



SMASH

machine learning for science and humanities postdoctoral program

SOME TOOLS

https://metadriverse.github.io/

SMASH
machine learning for science and humanities postdoctoral program

The Horizon – Gaps and Opportunities

Current bottlenecks and future research directions

- GAPS
- OPPURTUNTIES
- THREATS

I FEEL SLOVENIA
Co-funded by the European Union
SMASH
machine learning for science and humanities postdoctoral program

GAPS

TECHNICAL

- Long-tailed Events
- Extreme weathers
- Entropy in Traffic
- Data Scarcity
- Reaction Time
- Communication

COGNITIVE

- Ability of Computer Vision
- Semantic Intelligence
- Social Norms
- Emotional Aspects of Traffic

INFRASTRUCTURE

- Road Qualities
- Signs are markings built for Humans
- V2X communication
- VIoT infrastructure
- 5G anc connectivity

REGULATORY

- Needs new policies
- Uniformity of Traffic Laws
- AI Certification

SMASH

machine learning for science and humanities postdoctoral program

OPPURTUNITIES

SERVICE

- Mobility as a Service
- From Ownership to Subscription on demand
- Reduction in overall number of vehicles, parking spaces, increased availability
- **Inclusivity** – Better mobility solutions for physically impaired or elder individuals

LOGISTICS

- AV supply chains
- AV trucks can operate 20+ hours a day
- AV Platoons can operate more optimally and consume lesser fuel
- Personnel Costs will be lowered
- Reduced journey times due to V2X capabilities

SAFETY

- AVs will have lesser accident rates
- Avs have a more diverse set of sensors
- Avs are not emotional, always rational
- No loss of focus due to fatigue
- V2X communication can improve holistic traffic quality.

MARKET

- Biggest Consumer Market in Densest Areas
- Most complex traffic, most frustration, most loss of time and fuel in big cities
- Human drivers in dense traffic can cause chaos.
- Shift from personal cars to shared AVs.
- Optimized fuel usage, No Honking.

SMASH
machine learning for science and humanities postdoctoral program

THREATS

SECURITY

- Adversarial Attacks
- Denial of Service
- Hijacking
- Jamming

TRUST

- AI Mistakes won't be forgiven
- Public reactions
- Lack of AI knowledge will prevent trust building

LABOR

- AVs will eventually replace drivers one day.
- AVs will need personnel with new skill sets
- Labor Unions and Politics
- Transition will be messy

ETHICS

- Choosing the lesser evil
- Moral Frameworks (Human vs AI)
- Passenger vs Pedestrian
- Who will be accused / defended ?

SMASH
machine learning for science and humanities postdoctoral program

ITS NOT A QUESTION OF “IF”
IT’S A QUESTION OF “WHEN”

SMASH
machine learning for science and humanities postdoctoral program

THANK YOU FOR THE OPPURTUNITY

DHONYOBAD - SHUKRIYA - OBRIGADO - HVALA

SMASH

machine learning for science and humanities postdoctoral program

I FEEL SLOVENIA

Co-funded by the European Union

UNIVERZA V LJUBLJANI

Jožef Stefan Institute

IZUM

Institute of Information Science, Maribor

REPUBLIC OF SLOVENIA
MINISTRY OF THE ENVIRONMENT, CLIMATE AND ENERGY
SLOVENIAN ENVIRONMENT AGENCY

This is co- funded by the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355- SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund.
However, Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.