Research Project Proposal:

Sample complexity of transfer learning in reinforcement learning under a parametrized setting

Riccardo Poiani
riccardo.poiani@mail.polimi.it
CSE Track

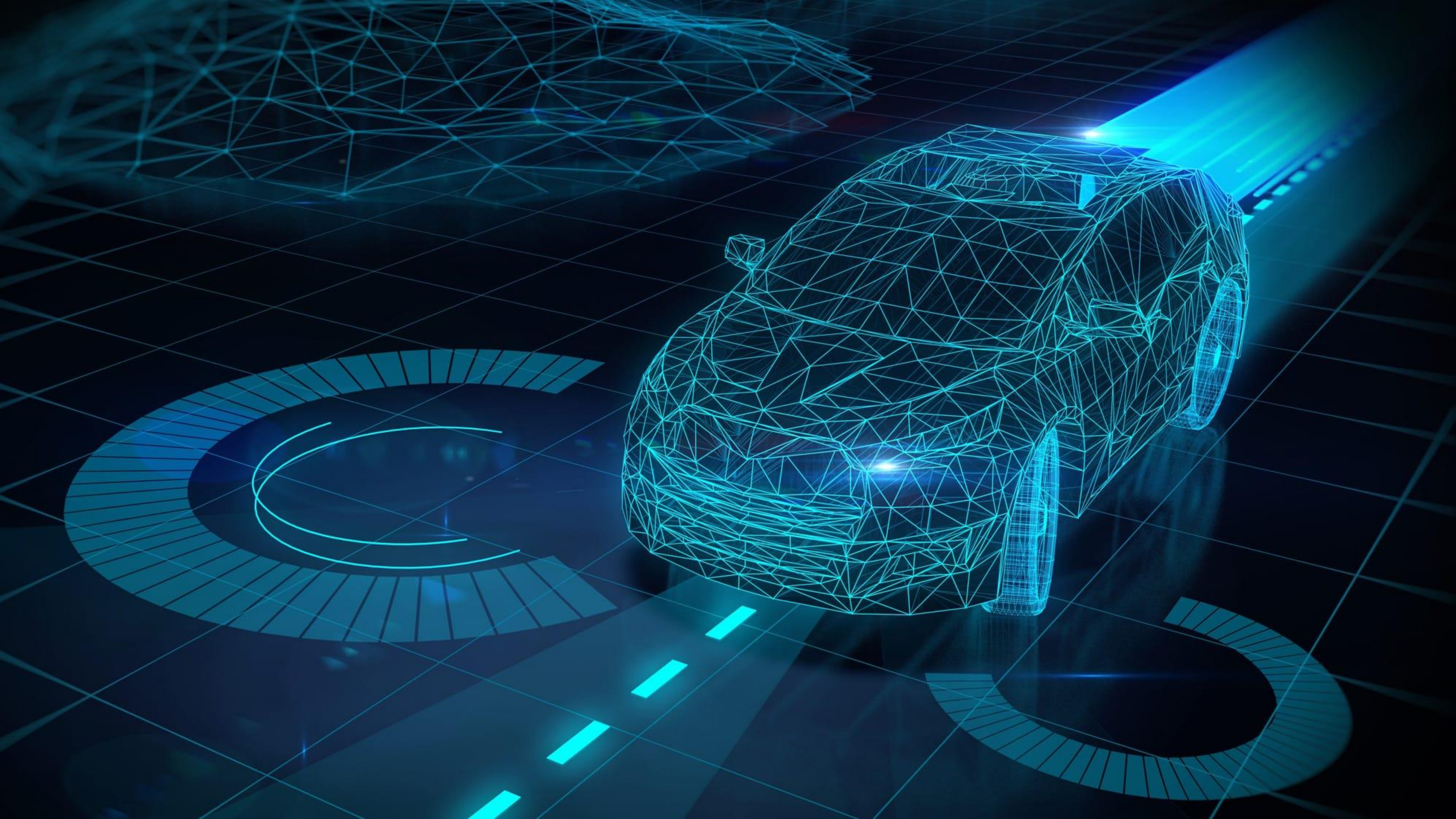
POLITECNICO
MILANO 1863

HONOURS PROGRAMME
HP-SR
in Information Technology

Outline

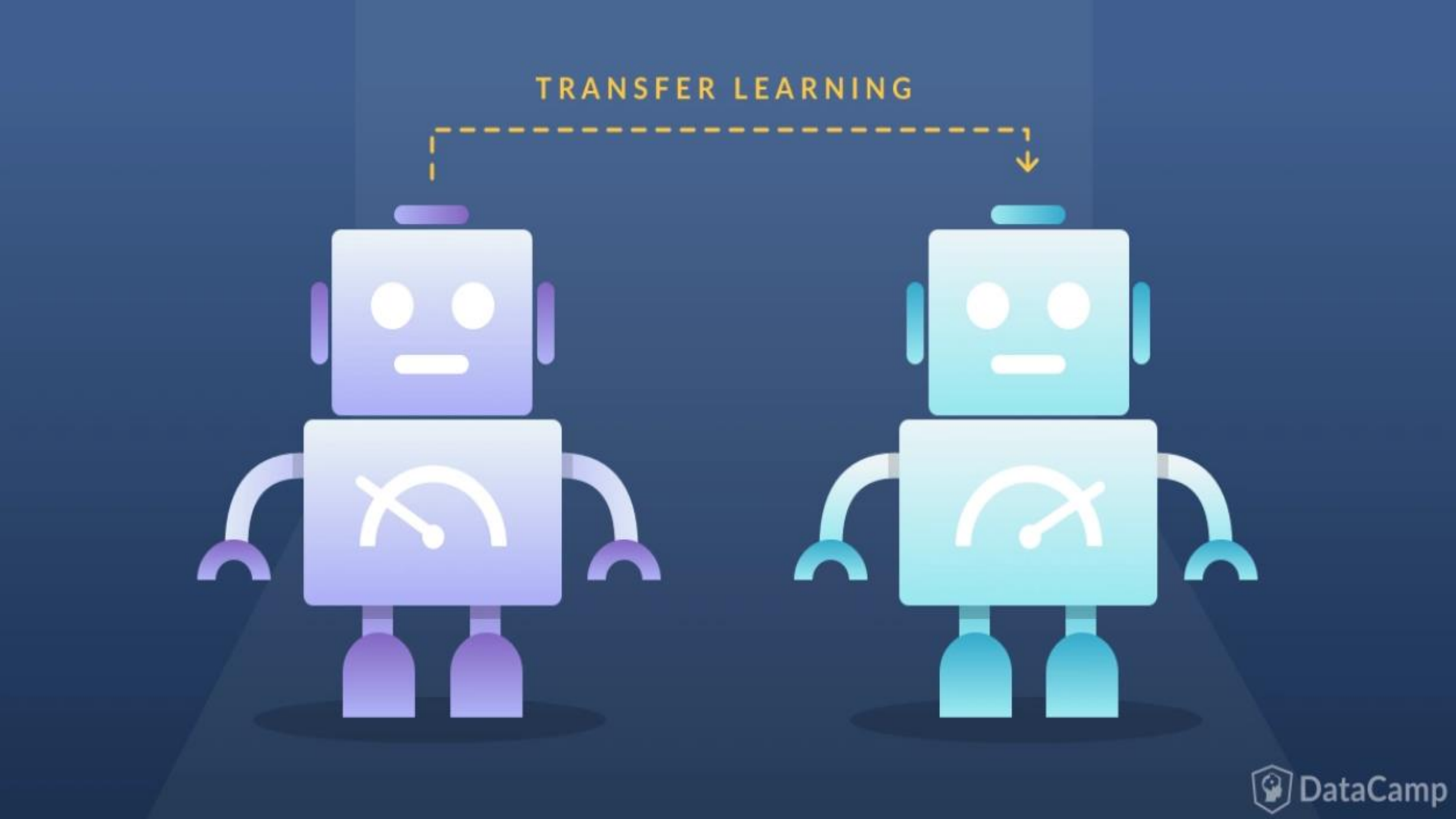
- Motivation
- State of the art
  - Generative setting
  - Non-generative setting
- Research idea and plan

- **Motivation**
- State of the art
  - Generative setting
  - Non-generative setting
- Research idea and plan

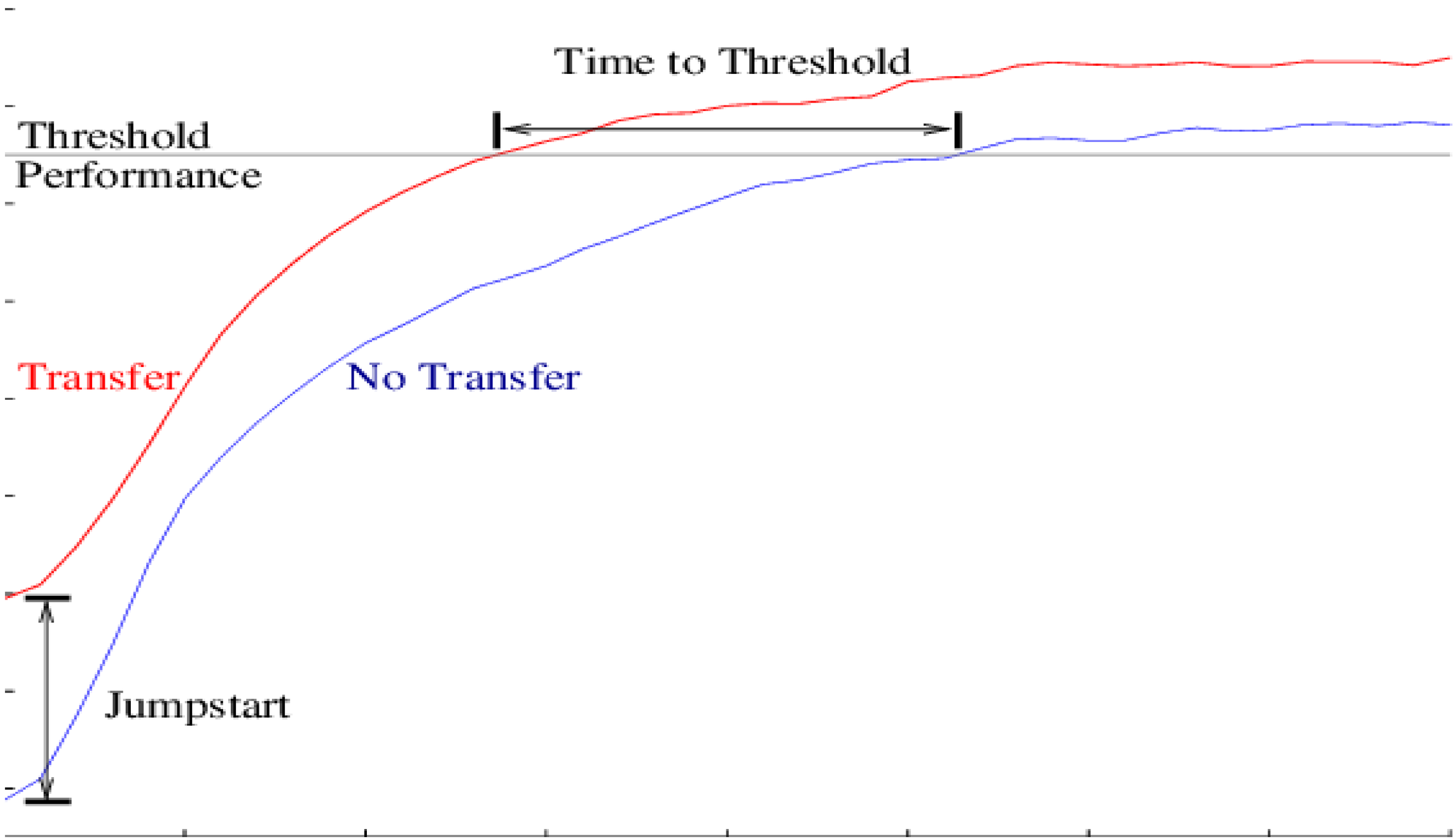


Problems and challanges

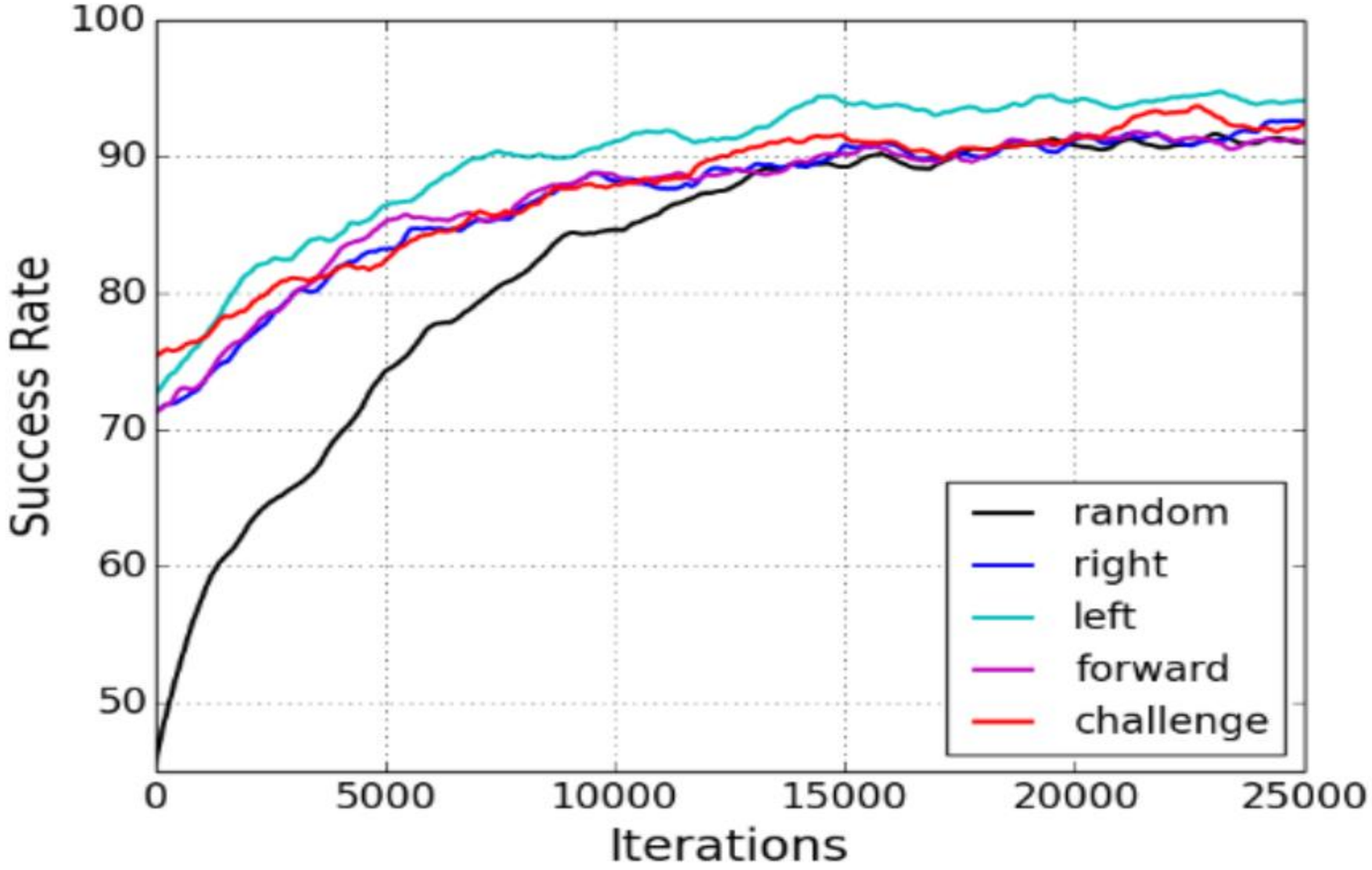
- **Superhuman** achievements in some problems but…
- Training costs **money**
- Training is **slow**
- Training can be **dangerous**


TRANSFER LEARNING
DataCamp

Transfer: benefits



Transfer: an example



Isele et. al [2017]

Reinforcement learning (RL)

An **agent** acts in an **environment** in order to maximize a **reward signal**.

The problem is usually formalized as a Markov Decision Process:

- States: S
- Actions: A
- Initial state distributions
- Reward function
- Transition distribution
- Discount factor:
  It encodes information about
  horizon H

RL: goal

- A **policy** π is a distribution over the actions, given the state
- The goal is to learn an **optimal policy** (up to some required accuracy)
    - the policy that **maximizes** the **expected cumulated discounted** reward
    - Often expressed in term of $\boldsymbol{V^{\pi}(s)}$ or $\boldsymbol{Q^{\pi}(s,a)}$
- Many algorithms exist: SARSA, Q-learning, Delayed Q-learning…

RL: sample complexity

Number of timestamps in which the policy is sub-optimal w.r.t. a fixed quantity ϵ

RL: PAC-MDP efficient algorithm

- **Probabilistic correct** with confidence at least $1 - \delta$
- **Polynomial sample complexity** in the relevant quantities $\left(S, A, \frac{1}{\epsilon}, \frac{1}{\delta}, H\right)$

Setting and goal of the project

- Typical transfer setting
- The agent acts in an **enviroment** whose dynamics are characterized by some **unknown parameter** $\theta \in \Theta$
- Understanding how to **exploit transferred knowledge** to reduce sample complexity
  - Generative case
  - Non generative case
- Research objective: algorithms with **theoretical** guarantees; **experiments**

- Motivation
- **State of the art**
  - Generative setting
  - Non-generative setting
- Research idea and plan

RL: Transfer

Paper	Allowed Differences	Knowledge Transferred	Metric
Abel et. al [2018]	Reward	V(s) / Q(s,a)	Jumpstart and Sample complexity
Azar et al. [2013]	Transitions and Rewards	Policy	Cumulated reward
Tirinzoni et al [2019]	Transitions and rewards	Samples	Cumulated reward
Ammar et al. [2015]	All	Samples	Cumulated reward
Tirinzoni et al. [2018]	Transitions and rewards	V(s)	Sample complexity
Many others…			

- Motivation
- **State of the art**
  - **Generative setting**
  - Non-generative setting
- Research idea and plan

Generative settings

- The analysis of the transfer case is currently missing
- Classical RL cases
  - A typical lower bound of the problem: $\tilde{O}\left(\frac{|S||A|H^3}{\epsilon^2}\right)$
  - **Uniform** sampling approach (Azar et al. [2013])
    - match lower bound under some assumptions
  - **Variance reduced** approach (Sidford et al. [2019])

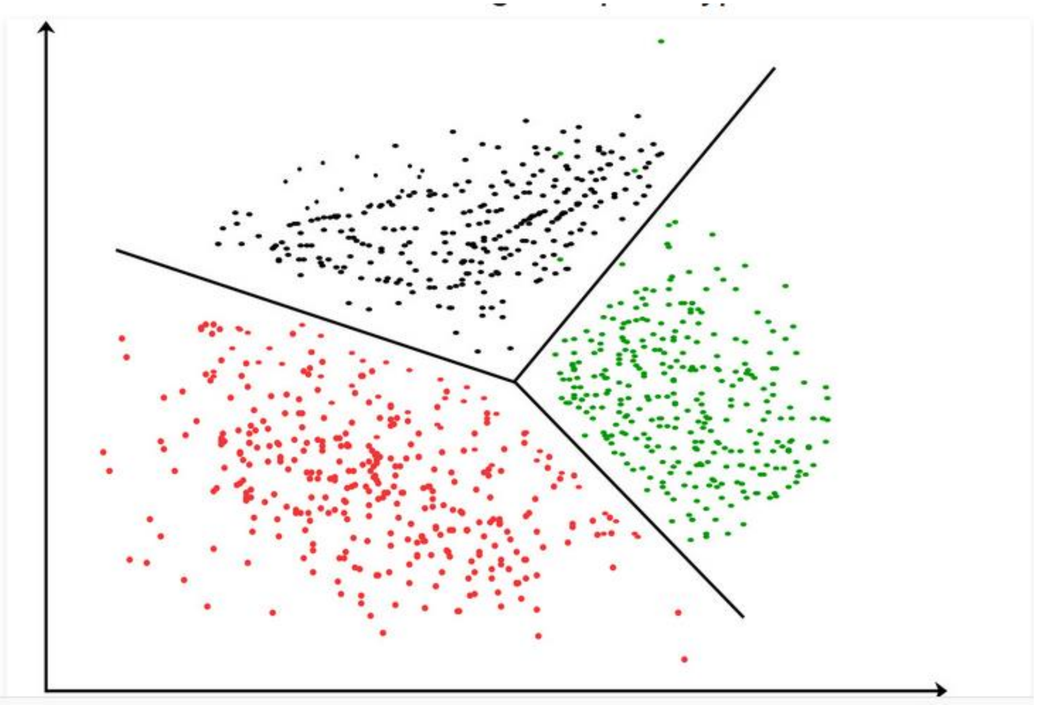
- Motivation
- **State of the art**
  - Generative setting
  - **Non-generative setting**
- Research idea and plan

Non-generative setting

- Maximum Exploration Reinforcement Learning (MERL) (Lattimore et al. [2013])
  - $\tilde{O}\left(\frac{|\Theta| H^3}{\epsilon^2}\right)$ **match a lower bound** up to a log factor
  - **Impractical** algorithm
- **Parameter elimination** method (PEL) (Dyagilev et al. [2008])
  - $\tilde{O}\left(\frac{|\Theta| H^6}{\epsilon^3}\right)$
  - Sequential probability ratio test

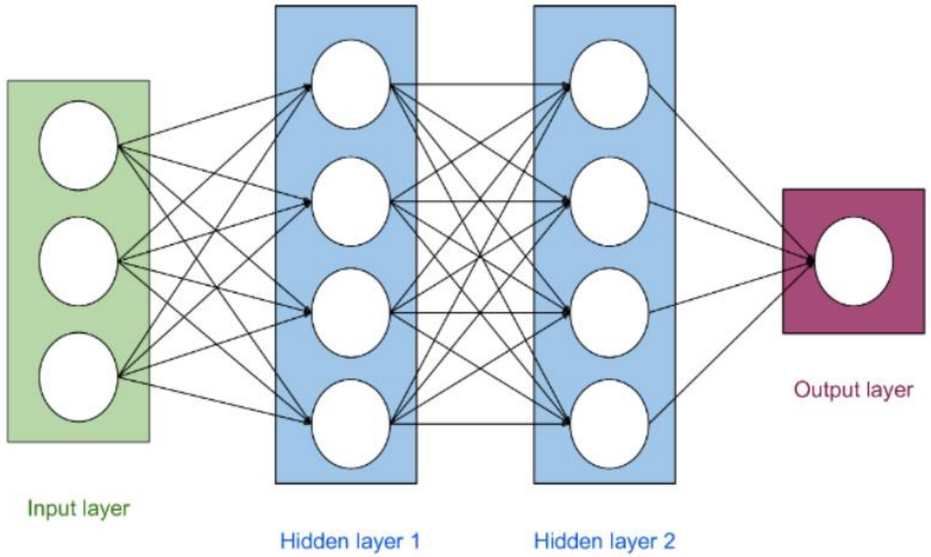
Non-generative setting

- On the sample complexity of Multi-task RL (Brunskill et. Al [2013])
  - Multi-task setting
  - **Clustering** approach
  - **Theoretical** bounds
  - Trade-off between structure **exploitation and exploration**



Non-generative setting

- **Hidden** parameter MDPs (Killian et. Al [2017])
  - **Complex** solution that works very well in practice
  - No theoretical guarantees
- **Contextual** MDPs (Modi et. Al [2017])
  - Continuous space for the context
  - **Known** context



- Motivation
- State of the art
  - Generative setting
  - Non-generative setting
- **Research idea and plan**

How to tackle the problem

- The problem of sample complexity in the transfer learning setting is hard
- There is little understanding so far in the literature
- We can take advantage of a **generative model** to better understand the problem
- From this simplified case, take insight for more practical algorithms

Desired achievements

- Generative case [65% completed]
  - **Online** algorithm with **theoretical** guarantees [85%]
  - **Better bounds** than the classical RL case by exploiting the structure [95%]
  - Propose a **real setting** when the algorithm can be used [10%]

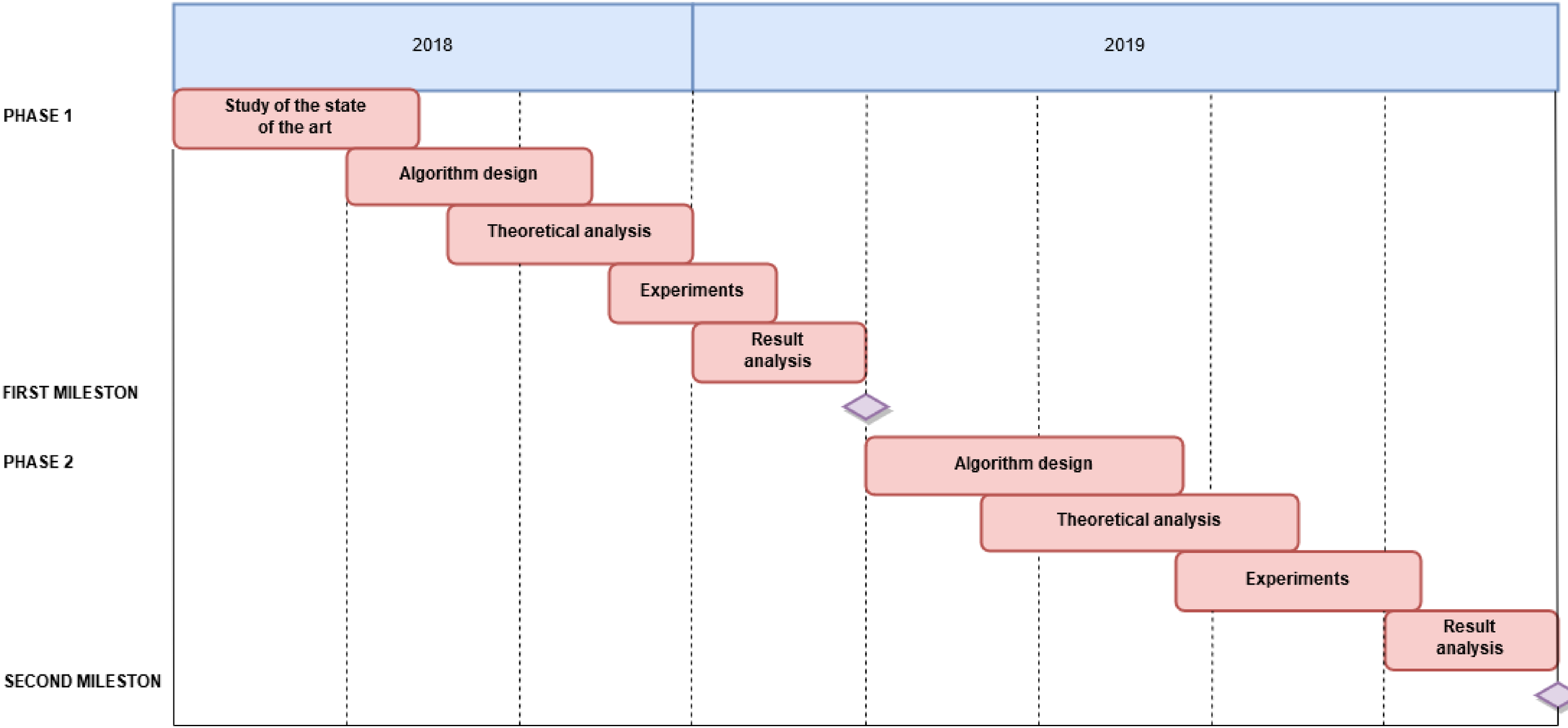
Desired achievements

- Non-generative case [0% completed]
  - **Online** algorithm with **theoretical** guarantees
  - **Experiment** to compare against state-of-the-art algorithms

Milestones

- **ICML** 2020 7 February
- **NeurIPS** 2020: around the end of May

Research plan



Thanks for your attention!