Reinforcement Learning Introduction

Odalric-Ambrym Maillard

Dec. 2021

Inria Scool

What is R.L.?

Robots?

Set of tools?

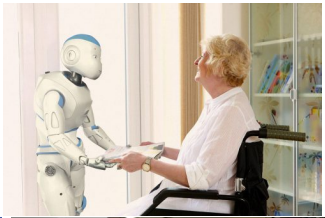
Content Recommendation?

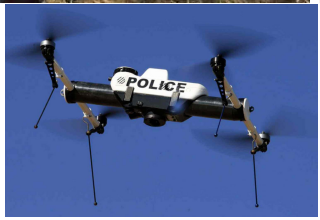
+ Fears: HAL, Terminator, Big-Brother?

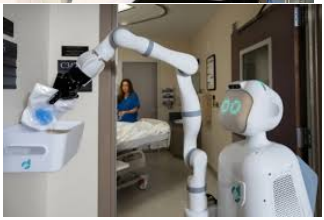
Robots everywhere?

Human bias!







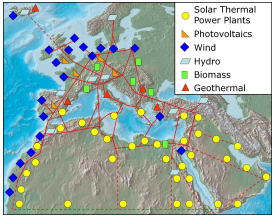


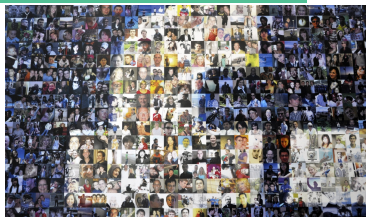
- +: Repetitive tasks, accurate.
- -: Expensive, **local** perception, energy, maintenance costs, ethics?

Mimic/replace human expertise vs assisting their limited abilities?

RL: Robots and games

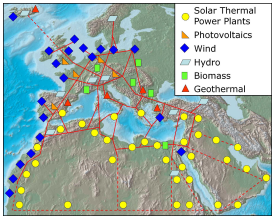
Global sensing and acting

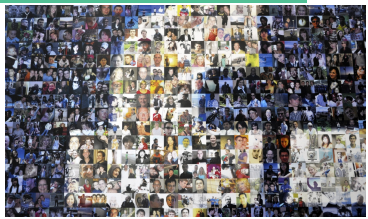




- Making sense of millions of data.
- Identifying similar contexts.
- Global testing, (safe) trial-and-error.

Global sensing and acting





- Making sense of millions of data.
- Identifying similar contexts.
- Global testing, (safe) trial-and-error.

+10 million users
smartphone app.
Help identify pathogens.
Suggest preventive/curative actions.

Plantix

Go

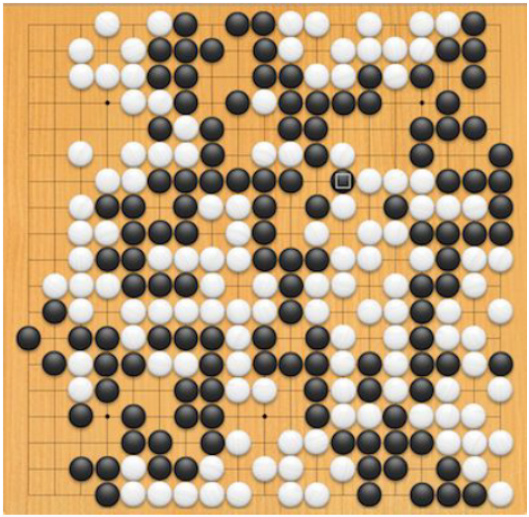
10^{100} board configurations

Atari

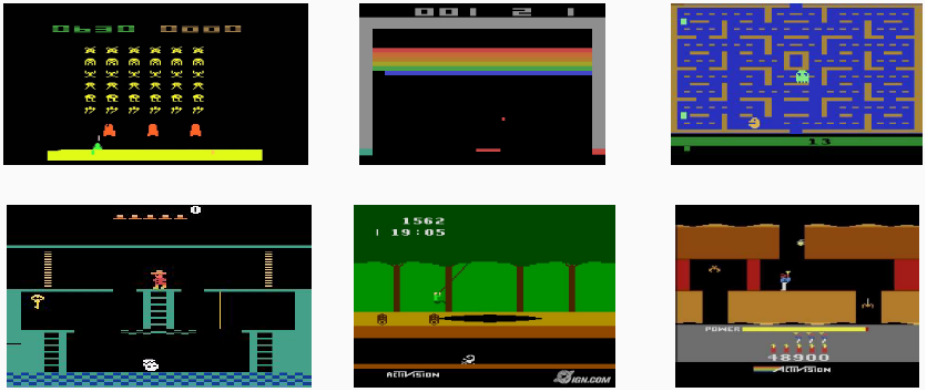
Complex visual input
Very different games

Starcraft

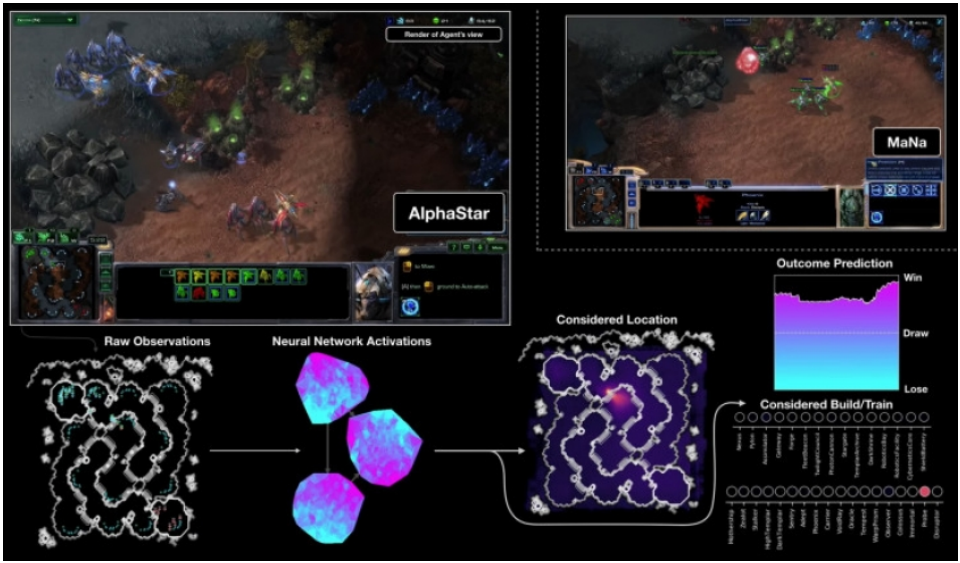
Complex observations and Planning



19×19 possible actions, 10^{100} board configurations



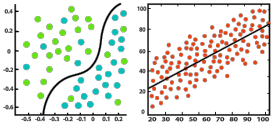
Complex visual inputs, very different games



Complex observations and planning

Supervised ("Learn to predict")
- Classify, Regress

Perception

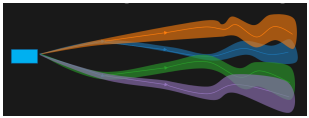


Unsupervised ("Learn to represent")
- Cluster, Transform

Representations

Reinforced ("Learn to interact")
- Act, Plan

Decisions



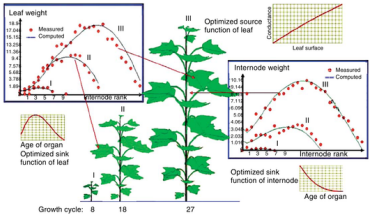
Reinforcement Learning is about automating **Decision** making.

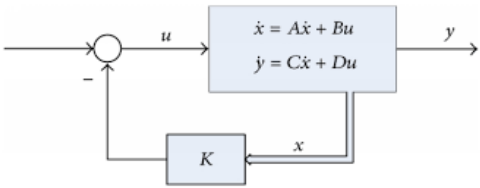
RL builds the **brain**.

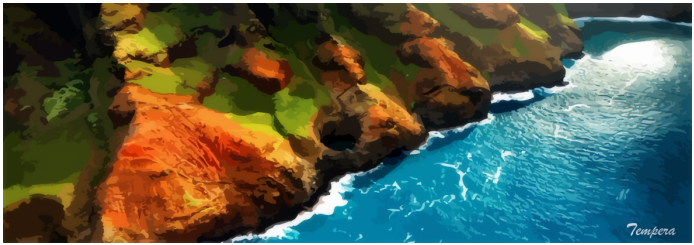
When? **Uncertain** world (black-box, unknown, noisy)
How? Statistics, optimization, control theory.

Reinforcement Learning: The agent does *not* know the dynamics of the system beforehand (Black-box). Learned by **trial and error**.

Control theory: The system's dynamics is perfectly known beforehand (white-box).







Sustainable Decision Making

Sample-efficient, energy-efficient, for **real-world** challenges

Societal Decision Making

Decision companions, ethics (no human farming!), **collaborative** decisions.

Ethics

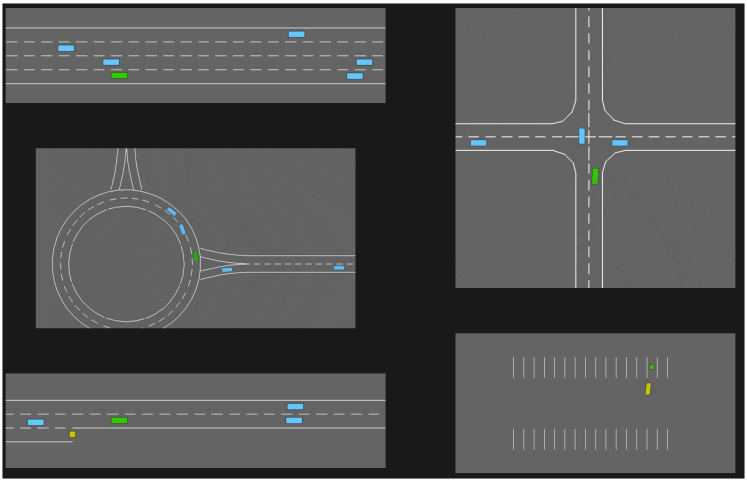
- Assist vs Replace?
- Recommend vs Influence ?
- Crowd-source vs Enslave?
- Fairness, gender/etc. bias in data.
- Energetic, Environmental costs (total/per user)?

- B4H (Bandits For Health): Suggest medical consultation or treatment based on smart meters and many patients.

- What sequence of exercises will maximize the learning progress of this student?

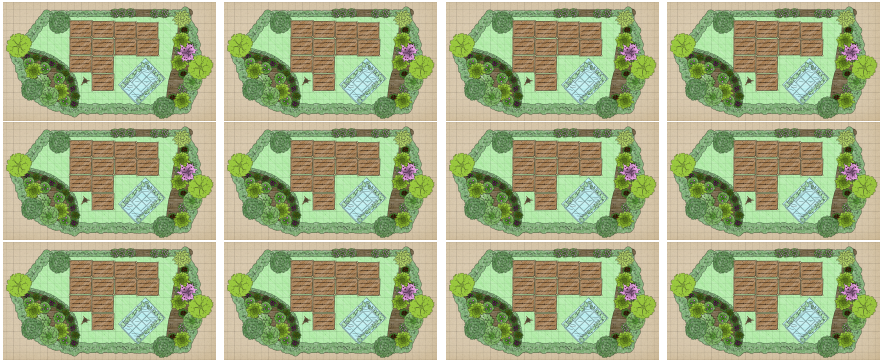
- pix.fr, lelivrescolaire.fr

- **Robust** driving in the presence of other vehicles (Renault).



- Propagation of uncertainty, obstacle avoidance, efficiency vs safety.

- SR4SG: Sequential Recommendation for Sustainable Gardening.



- Recommend **good practice** between farms, share/build knowledge.
- Towards Massively Collaborative Agriculture.
  - (Re-)discover local practice and expertise.
  - Experiment/Propagate novel practices (invasive species, diseases, etc.)
  - Transfer practice between regions (climate evolution).

Continuous interaction.
Agent does not know environment: *learning*.

System, called the *Environment*, with which we interact:

- we get **observations**,

System, called the *Environment*, with which we interact:

- we get **observations**,
- we output **decisions**,

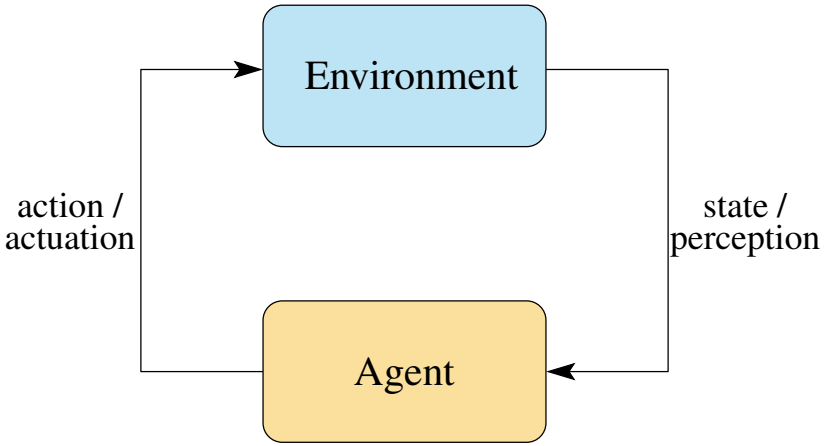
System, called the *Environment*, with which we interact:

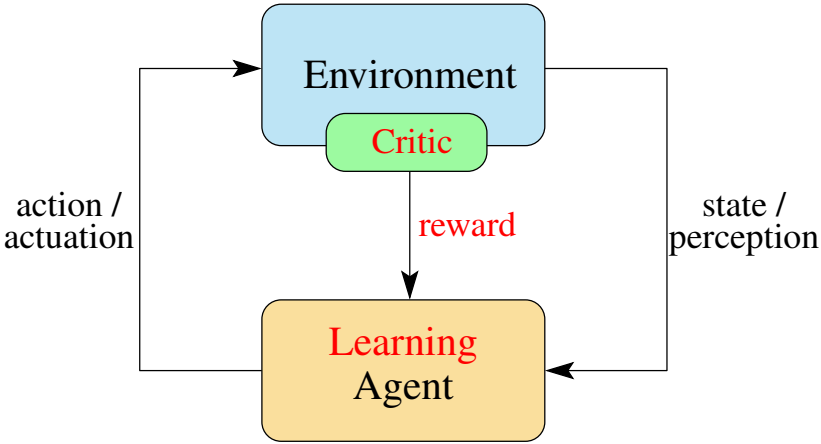
- we get **observations**,
- we output **decisions**,
- we get **new** observations,

System, called the *Environment*, with which we interact:

- we get **observations**,
- we output **decisions**,
- we get **new** observations,
- and so on and so forth.

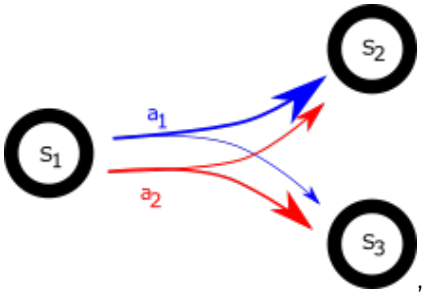
We receive a signal telling how good/bad are decisions: **reward**.


Environment
action /
actuation
state /
perception
Agent


Environment
Critic
reward
action /
actuation
state /
perception
Learning
Agent

for $t = 1, \ldots, n$ **do**
    The agent perceives state s_t
    The agent performs action a_t
    The environment evolves to s_{t+1}
    The agent receives reward r_t
end for

- *States* $s_1, s_2, \ldots$: "How you describe your system from observations"
- *Actions* $a_1, a_2, \ldots$: "What you are allowed to do"
- *Transitions* (unknown): "how the system changes when you choose some action"

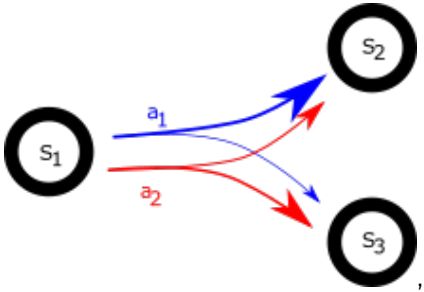


,

- *Initial* states: "Where you start from". Final state (optional): "Where you stop"
- *Reward* (unknown): "How good is an action". Linked to desirable states.

Simplified example 1: Clinical trials

- Environment is a human
- *States*: $s_1 = \texttt{sick}, s_2 = \texttt{cured}, s_3 = \texttt{uncured}$
- *Actions*: $a_1 =$  , $a_2 =$  , etc.
- *Transitions*:



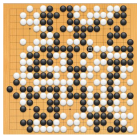
,

- *Initial* state: `sick`. Final states: `cured` and `uncured`.
- *Reward*:+1 if cured, 0 else.

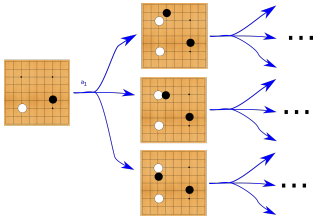
Actions have different success probabilities. (States do not really matter here.)

Simplified example 2: Game of go

- Environment is board.



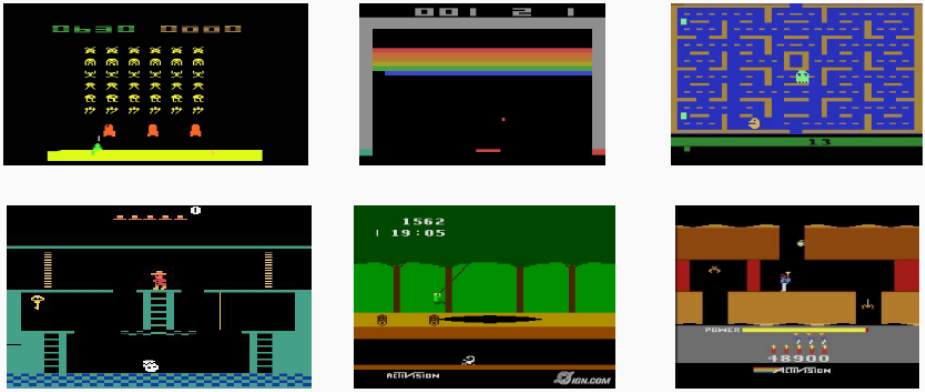
- *States*: board (goban) configurations.
- *Actions*: all possible moves allowed by game rules.
- *Transitions*: e.g. play white at 2×2, then opponent plays, then ...



- *Initial* state: Empty board. *Final* states: end game configurations.
- *Reward*:$+1$ if win, 0 else.

Many possible states ($3^{19\times19}$), trajectories of (states, actions, next states).

- Atari games



- *Observations*: Screen pixels *Actions*:  *Rewards*: Score update
- *Transitions*: internal rules of each game (very different).

Complex visual inputs ($256^{640\times480}$ possible screens).

Games with Nature?

Diverse situations: *Diverse games*

Q: Actions? Observations? Rewards?

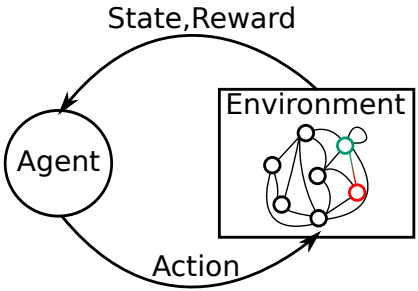
- The system is **partially** known (*Uncertainty*)

- The system is **partially** known (*Uncertainty*)
- We get to know it thanks to **observations** only. (*Statistics*)

- The system is **partially** known (*Uncertainty*)
- We get to know it thanks to **observations** only. (*Statistics*)
- Observations are collected **actively**. (*Algorithm*)

- The system is **partially** known (*Uncertainty*)
- We get to know it thanks to **observations** only. (*Statistics*)
- Observations are collected **actively**. (*Algorithm*)
- Decisions may have **effect** on next possible observations. (*Dynamics*)

- The system is **partially** known (*Uncertainty*)
- We get to know it thanks to **observations** only. (*Statistics*)
- Observations are collected **actively**. (*Algorithm*)
- Decisions may have **effect** on next possible observations. (*Dynamics*)
- Possibly noisy, high dimensional, structured, with risk-aversion, not so clear reward signal, dependencies, etc. (*Models*)



Sequential Decision Making under Uncertainty

Next observation ?

- May depend on all past history of interaction: unfathomable?

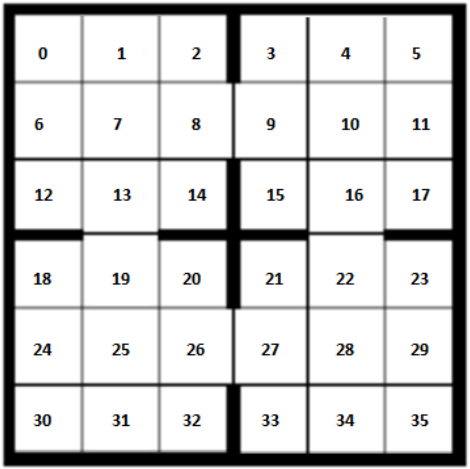
$$\mathbb{P}\Big(O_t \Big| \underbrace{o_1, a_1, r_1, \dots, o_{t-1}}_{\text{exponentially many!!}}, a_{t-1}\Big)$$

- State = **Summary** of history

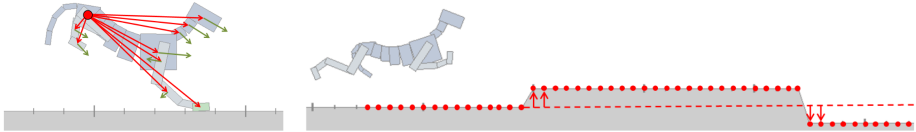
$$= \mathbb{P}(O_t \Big| s_{t-1}, a_{t-1})$$

"All you need to know to perform action"

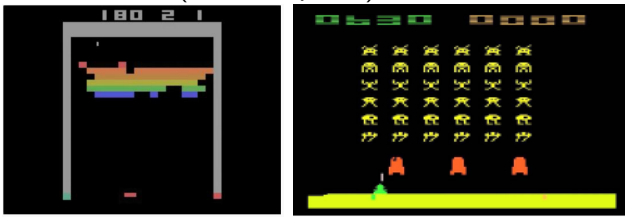
State= position



State= relative positions, velocities, and relative heights



State= frame (640x480 pixels) and last few frame differences



Medical trials


: $\mathcal{A} = \{$

,

,

,

,

$\}$

Ad-placement industry


: $\mathcal{A} = \{$


,


,


,

$\}$

Plant-health care


: $\mathcal{A} = \{$

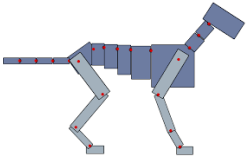
,

,

,

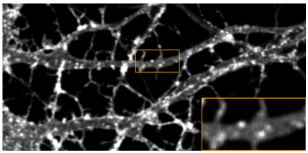
$\}$

Motion-planning:



: $\mathcal{A} = \left\{ \text{accelerations at each joint} \right\}$

High-resolution microscopy:



: $\mathcal{A} = \left\{ \text{exposure time, dosage.} \right\}$

The goal of an agent is to **maximize the rewards** accumulated over time:

$$\sum_{t=1}^{T} r_t$$

r_t : reward obtained in state s_t when playing action a_t.

Either by:

- **Finding a policy** that accumulates rewards when played (e.g. Simulator).
- **Interacting directly** with the real-world (no simulator available).

"*The more applied you go, the stronger theory you need*"

Merci

odalricambrym.maillard@inria.fr
odalricambrymmaillard.wordpress.com