AIL 7022: Reinforcement Learning

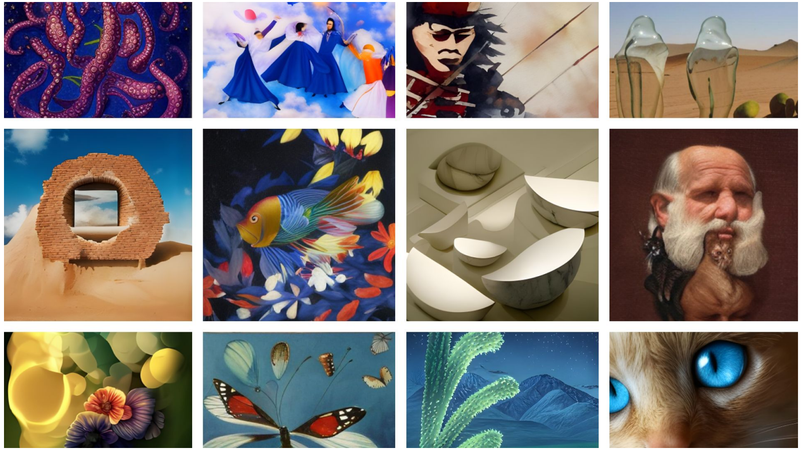
Lecture 1: Overview & Warm-up

Instructor: Raunak Bhattacharyya

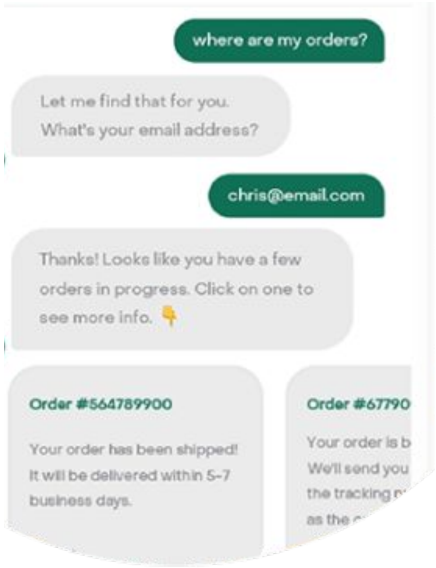
YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Course Webpage

Recent Advances in AI



[Source: Meta-AI]

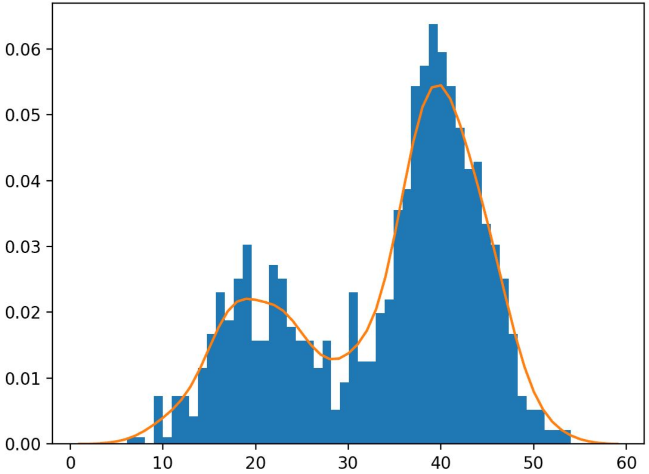


[Source: Hootsuite]

Core Idea

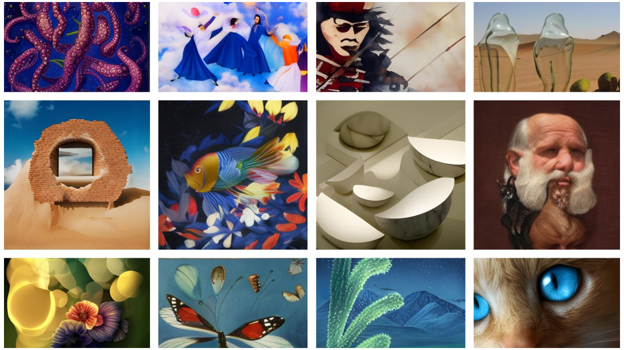
Source: Adobe

$p_\theta(\mathbf{x})$ $p_\theta(\mathbf{y}|\mathbf{x})$

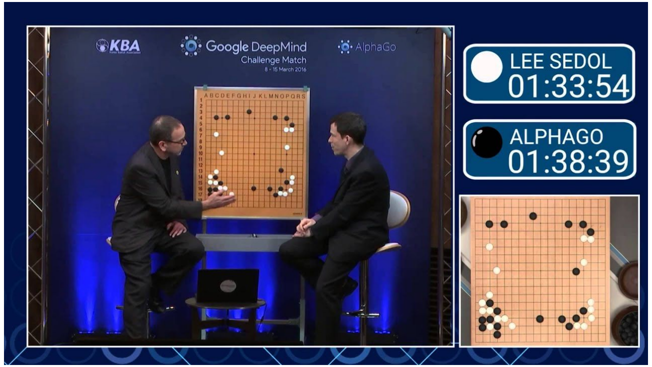


Source: MachineLearningMastery

RL: Discovery



Looks like something a person might draw!



[Source: Deepmind, DQN]

Unexpected: sometimes better than what a human may have done!

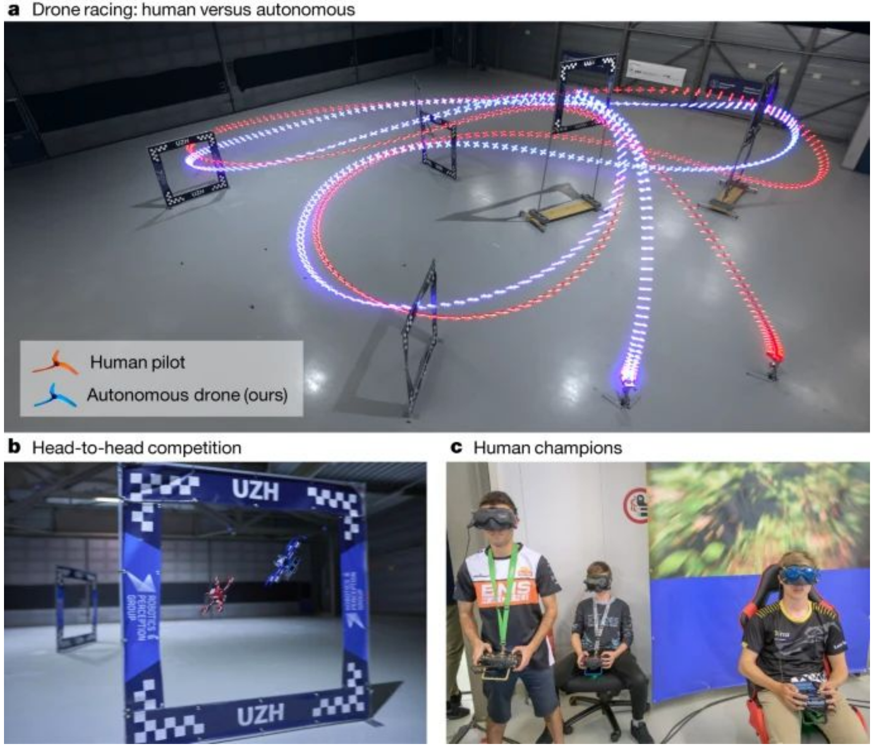
Adapted from Sergey Levine

Why this Course

- Exciting recent breakthroughs

Breakthroughs

Gran Turismo, Source: [GT Sophy]



Champion-level drone racing, Source: [Nature]

Superhuman GT

Superhuman Drone racing

Emergence of Locomotion

Manipulation

Why this Course

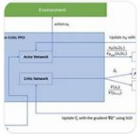
- Exciting recent breakthroughs

- Based on derived foundational approaches

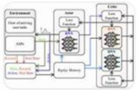
SOTA RL algorithms

most common baseline rl algorithms used in research papers starting from 2022

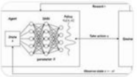
Proximal Policy Optimization (PPO)
PPO is renowned for its balance between implementation simplicity and performance. It has been widely adopted as a standard baseline in various RL research contexts.



Soft Actor-Critic (SAC)
SAC is an off-policy algorithm that integrates maximum entropy principles, enhancing exploration and learning stability. It is particularly effective in continuous action spaces.



Deep Q-Network (DQN)
DQN combines Q-learning with deep neural networks, enabling agents to learn policies directly from high-dimensional inputs like images. It remains a foundational baseline in RL research.



Twin Delayed Deep Deterministic Policy Gradient (TD3)
TD3 addresses overestimation biases in the Deep Deterministic Policy Gradient (DDPG) algorithm, offering improved performance in continuous control tasks.

Generated by ChatGPT

Why this Course

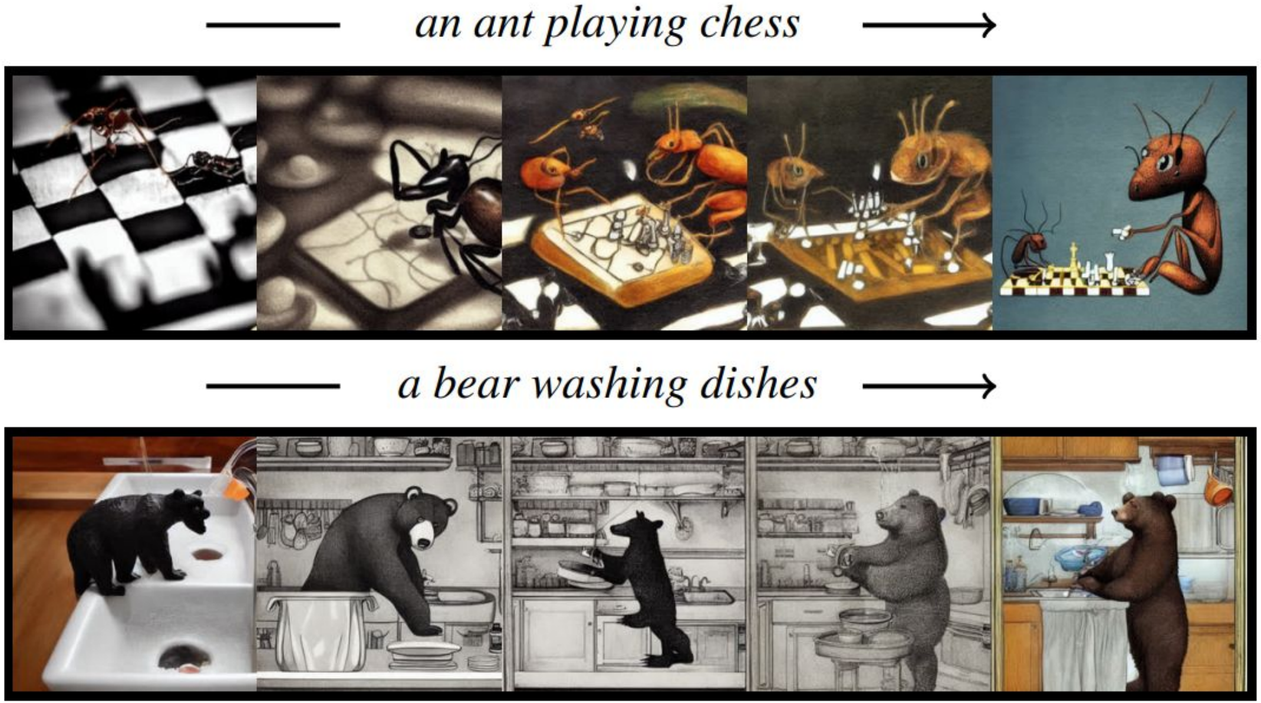
- Exciting recent breakthroughs
- Based on derived foundational approaches
- Benefits of scale and generalisation (think: CV in 2013)

Mobile Manipulation

Why this Course

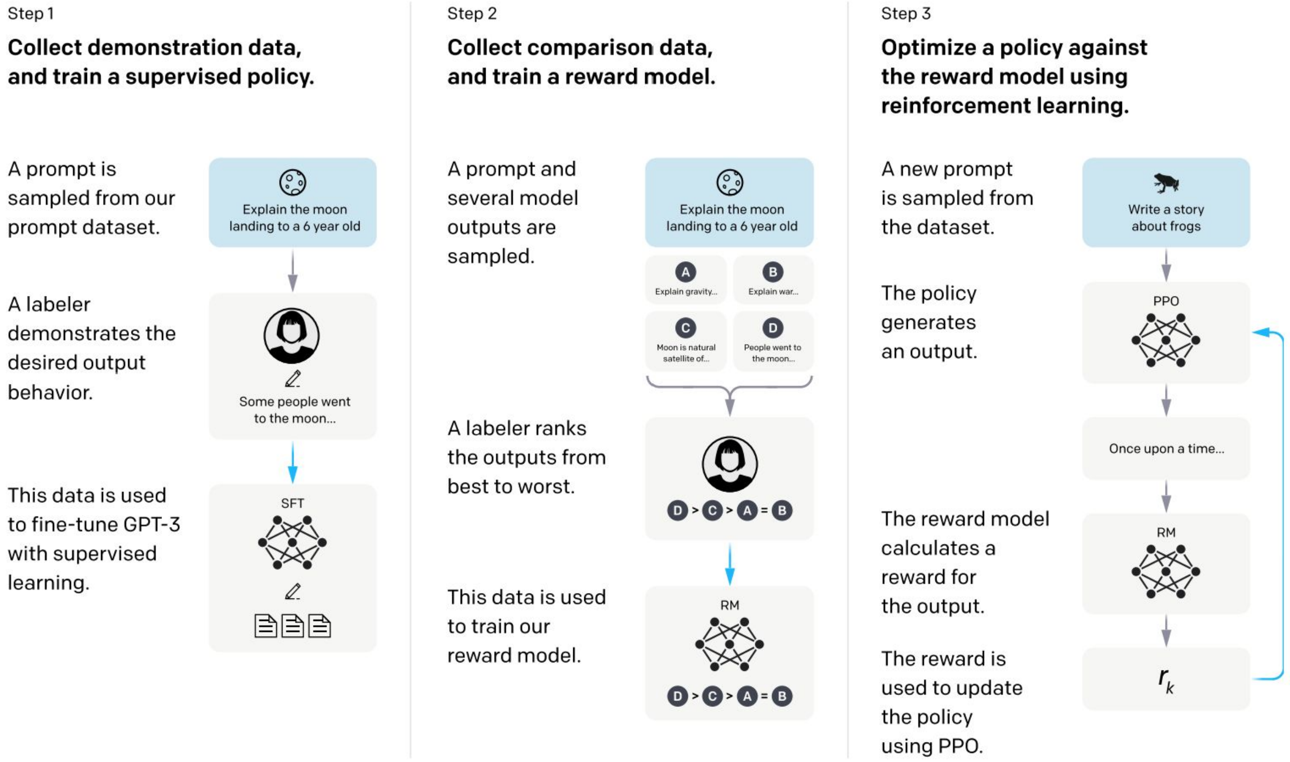
- Exciting recent breakthroughs
- Based on derived foundational approaches
- Benefits of scale and generalisation (think: CV in 2013)
- Being used in other cutting edge AI research & technologies

Optimising Image Generation



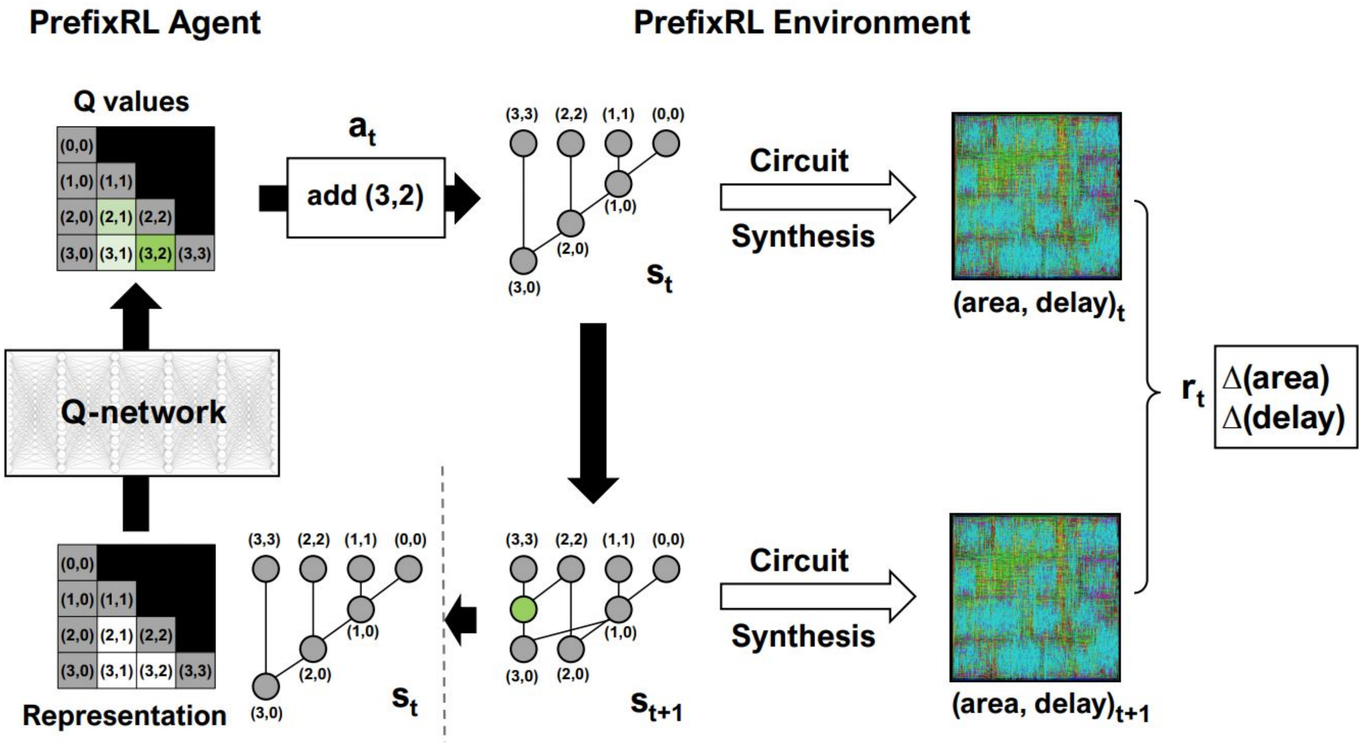
Source: DDPO

Language


Step 1
Collect demonstration data, and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain the moon landing to a 6 year old
A labeler demonstrates the desired output behavior.
Some people went to the moon...
This data is used to fine-tune GPT-3 with supervised learning.
SFT
Step 2
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled.
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
A labeler ranks the outputs from best to worst.
D > C > A = B
This data is used to train our reward model.
RM
D > C > A = B
Step 3
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
Write a story about frogs
The policy generates an output.
PPO
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

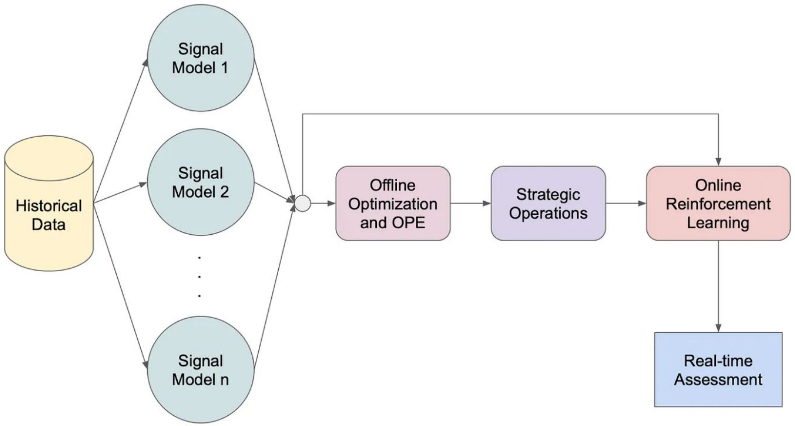
Language

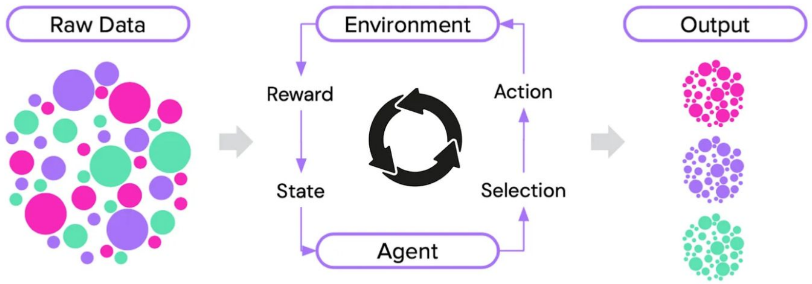
Circuit



PrefixRL, NVIDIA

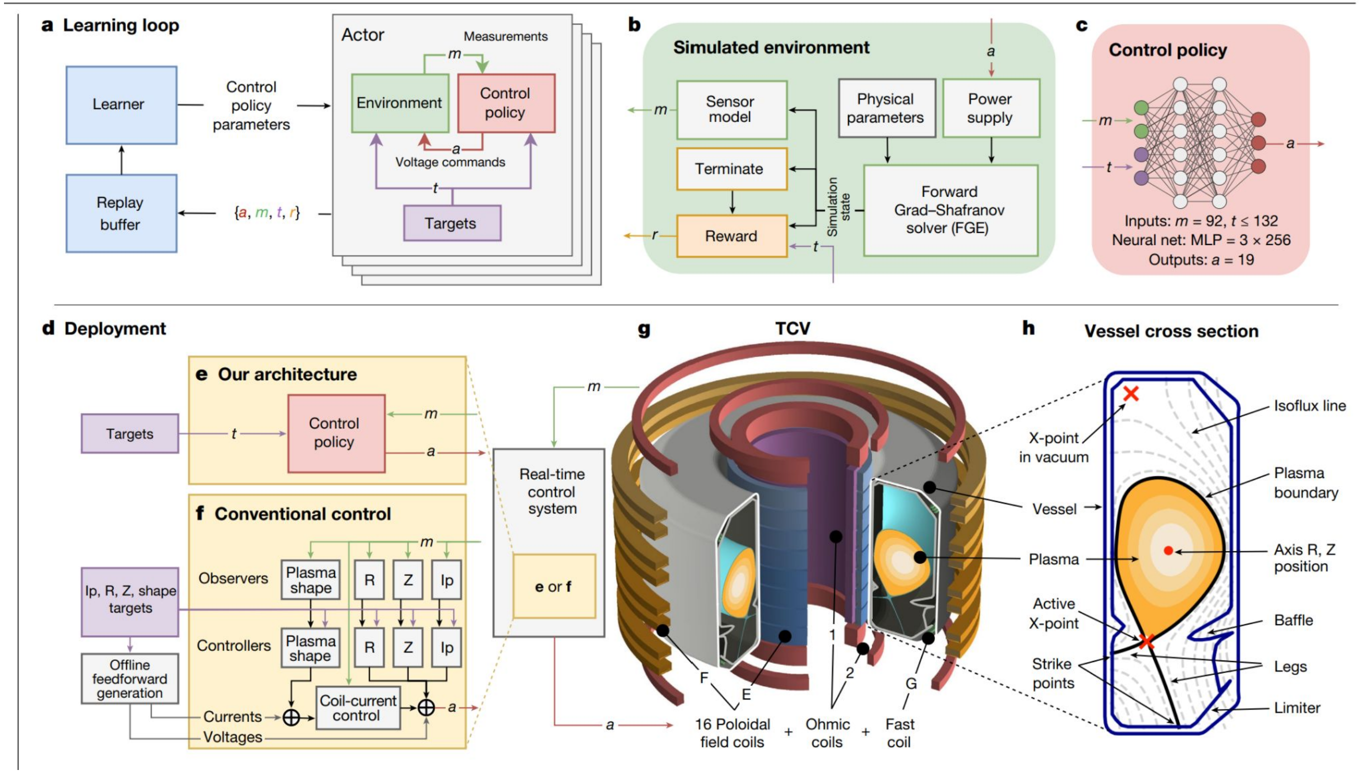
Pricing at Lyft





Algorithmic Pricing, Lyft

Nuclear Fusion Research



Why this Course

- Exciting recent breakthroughs
- Based on derived foundational approaches
- Benefits of scale and generalisation (think: CV in 2013)
- Being used in other cutting edge AI research & technologies
- Open questions

Open Questions

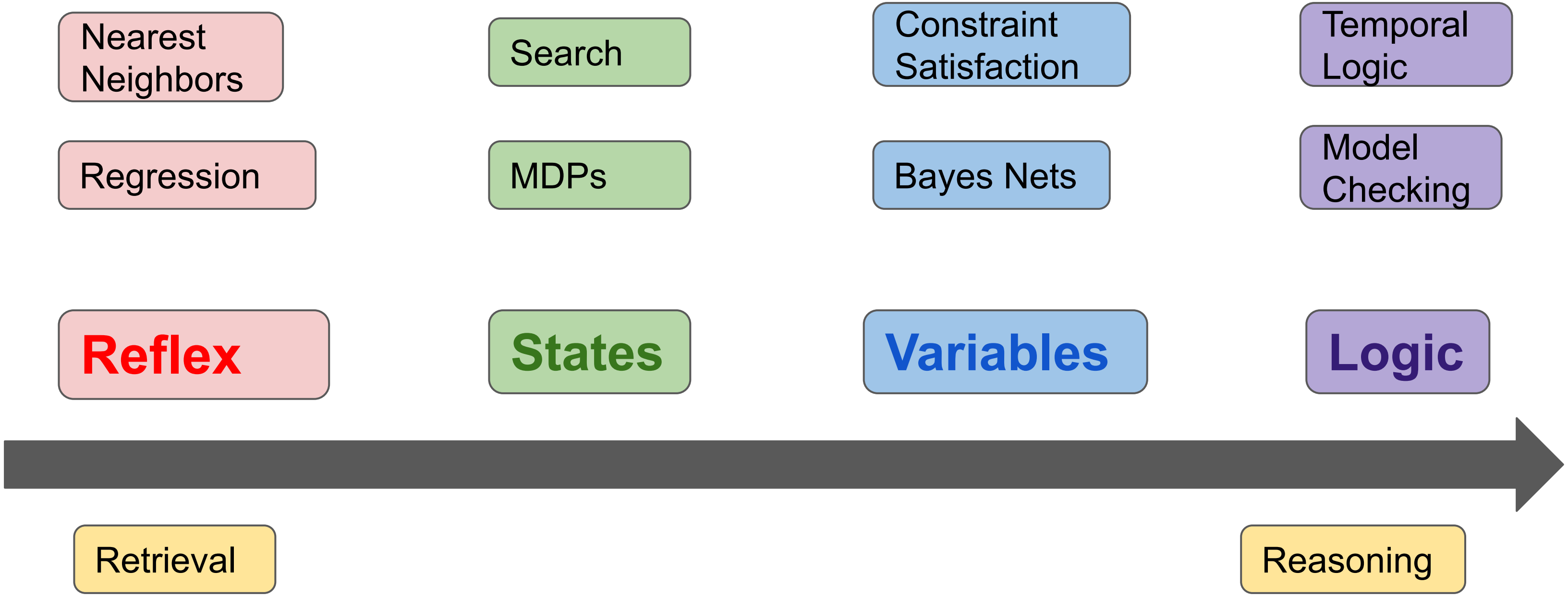
Theory often Follows Invention

- *Telescope [1608]*
- *Steam engine [1695-1715]*
- *Electromagnetism [1820]*
- *Sailboat [???]*
- *Airplane [1885-1905]*
- *Compounds [???]*
- *Feedback amplifier [1927]*
- *Computer [1941-1945]*

- *Optics [1650-1700]*
- *Thermodynamics [1824-….]*
- *Electrodynamics [1821]*
- *Aerodynamics [1757]*
- *Wing theory [1907]*
- *Chemistry [1760s]*
- *Electronics [….]*
- *Computer Science [1950-1960]*

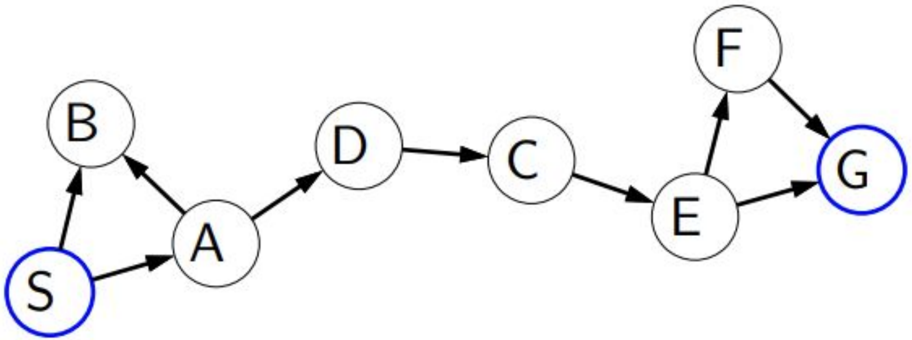
Slide from Yann Lecun

Contextualizing RL



Adapted from Percy Liang

Search Problems



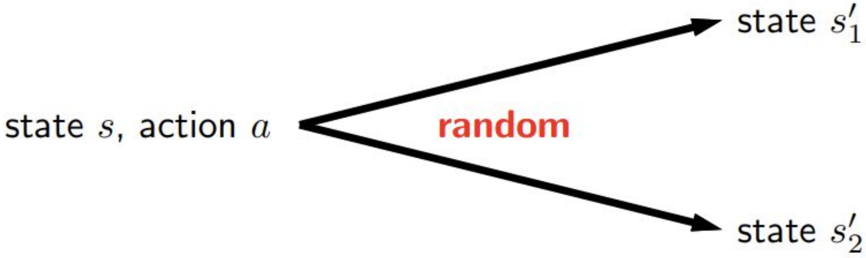
deterministic

state s, action a $\longrightarrow$ state $\text{Succ}(s, a)$

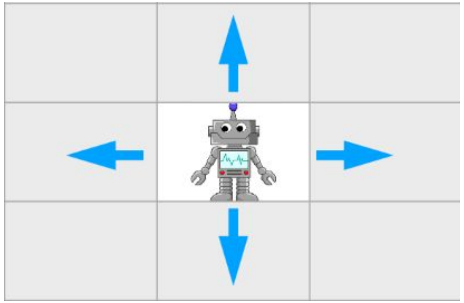
Uncertainty in the Real World

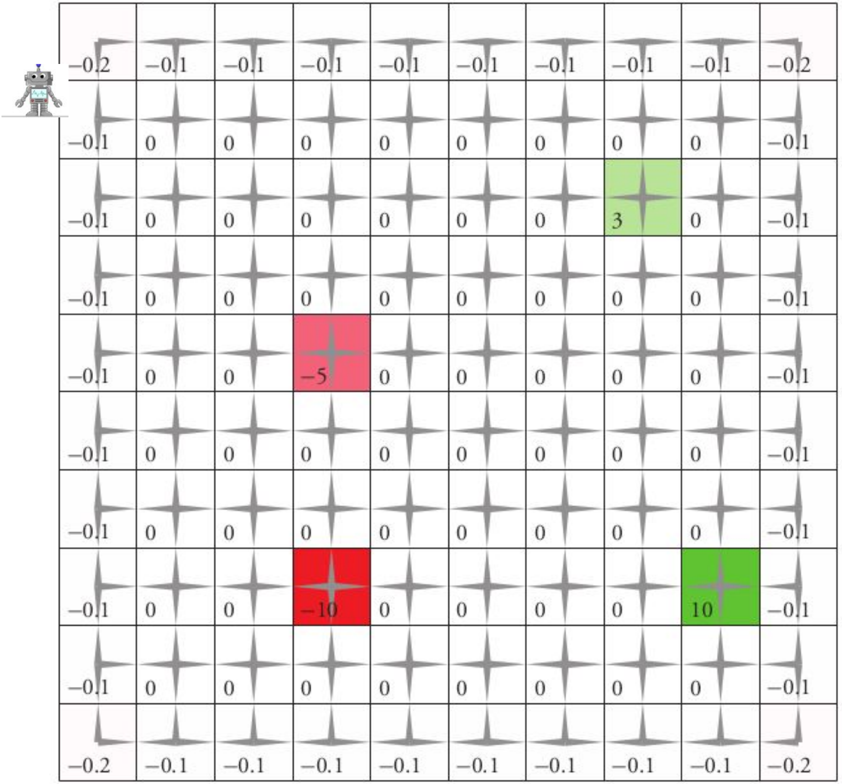
How other agents might behave

[Source: istockphoto]



Motivating Example

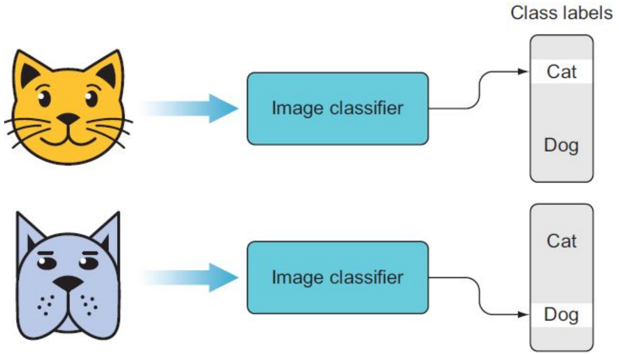




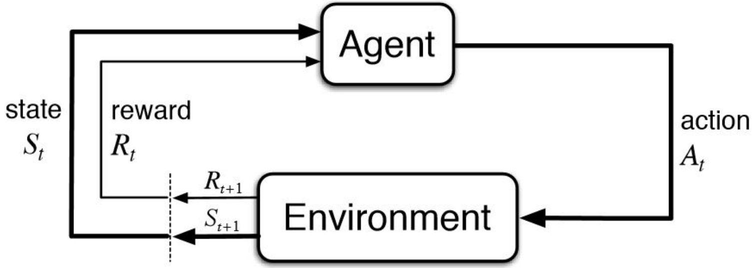
- 10x10 grid
- Up, down, left, right
- 0.7 **correct** dir (as instructed), 0.1 rest
- Green cells are absorbing (end state)

Figure from [Algorithms for Decision Making](), Mykel Kochenderfer

Contrast to Supervised Learning



[Source: Medium]



[Source: Sutton & Barto]

Input: x

Output: y

Data: $D = \{(x_i, y_i)\}$

Goal: $f_\theta(x_i) \approx y_i$

Someone gives you the labels

Input: State s_t at each time step

Output: Action a_t at each corresponding time step

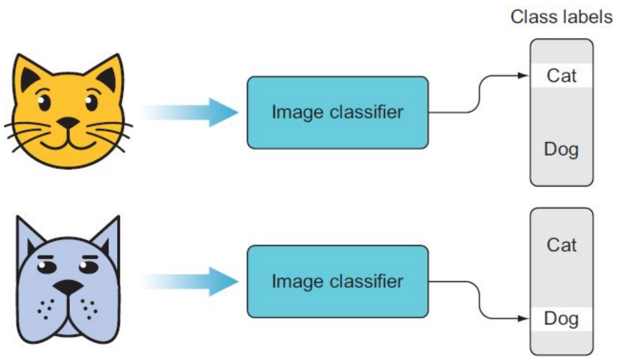
Data: $(s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T)$

Goal: Learn policy $\pi_\theta : s_t \to a_t$

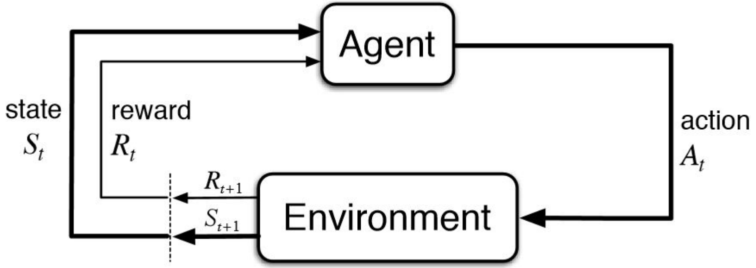
to maximize total reward obtained

Pick your own action

Contrast to Supervised Learning



Source: Medium

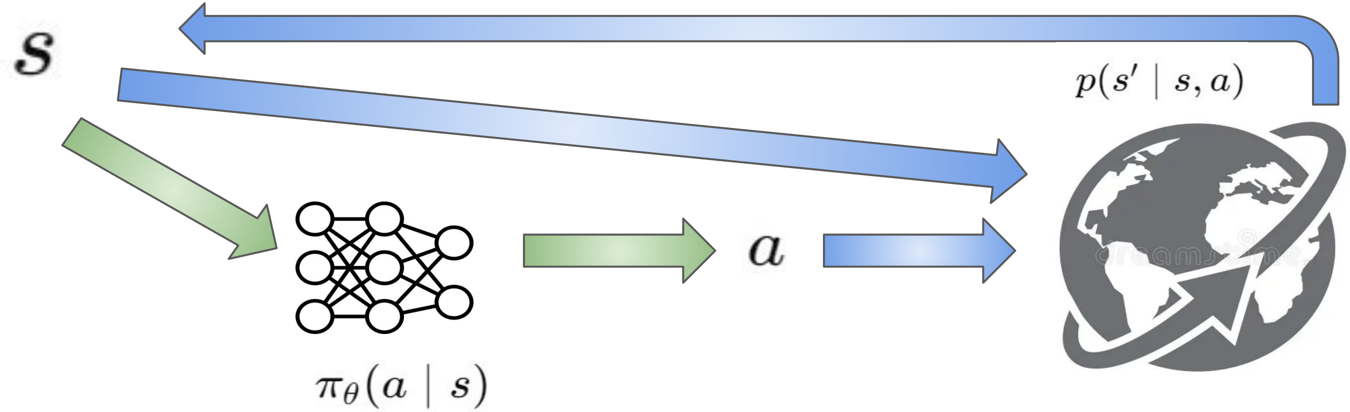


Source: Sutton & Barto

- i.i.d. data
- Known ground truth labels in training

- Data is not i.i.d.
  - Previous outputs influence future inputs
- No ground truth labels
  - We know the reward

RL Objective

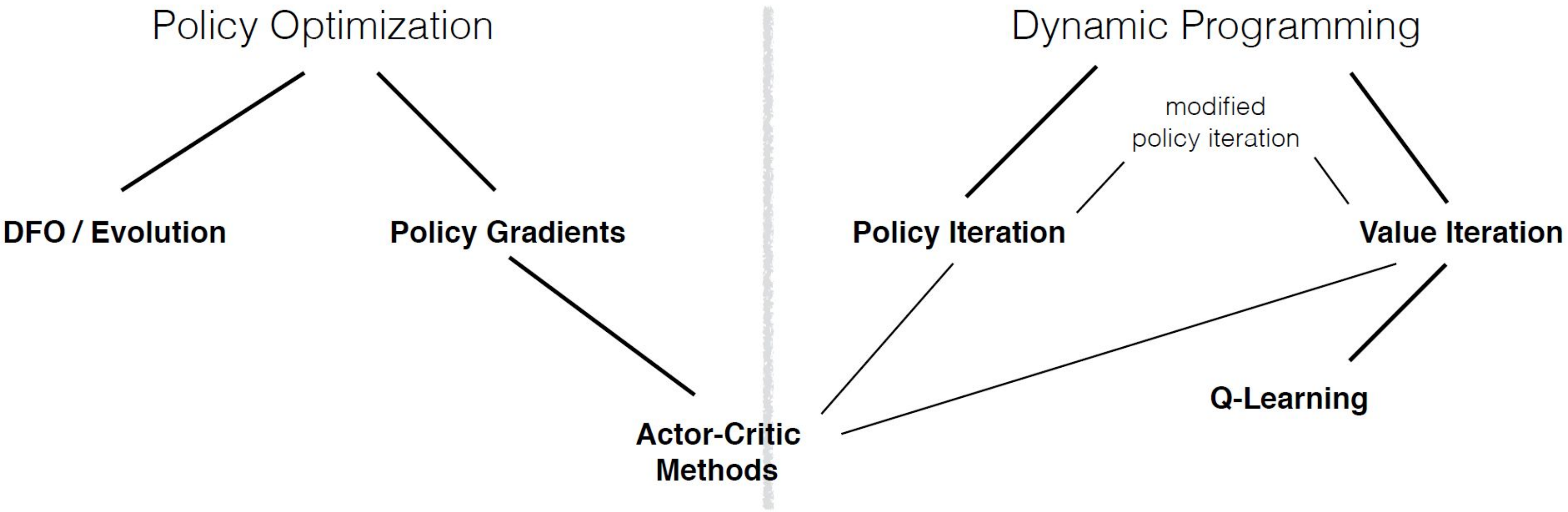


$$p_\theta(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t)$$

$$p_\theta(\tau)$$

$$\boxed{\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_t r(s_t, a_t) \right]}$$

Approaches to RL



Slide from John Schulman

What we will cover

Module 1

- What types of problems has RL been used for
- Representing said problems
- Notation and mathematical formulation

Module 2

- Solution approach
- Key constructs
- Approximations

Module 3

- What if we don’t have the full model?
- Monte Carlo methods
- Temporal difference approaches

Module 4

- Flagship model-free methods
- Q-learning
- SARSA
- DQN
- Double Q learning

Module 5

- How do we explore?
- Bandits
- UCB

Module 6

- A paradigm shift: Alternate approach
- Search problem
- Optimisation-based approaches
- Actor-critic methods

Module 7

- Learning models
- Model-based RL
- Dyna

Module 8

- How do humans do RL?
- Soft optimality
- Maximum entropy

Module 9

- Can agents learn from humans?
- Imitation learning
- Combining with powerful generative models

Module 10

- Learning reward functions
- Inverse RL

Module 11

- RL without trial and error
- Offline RL