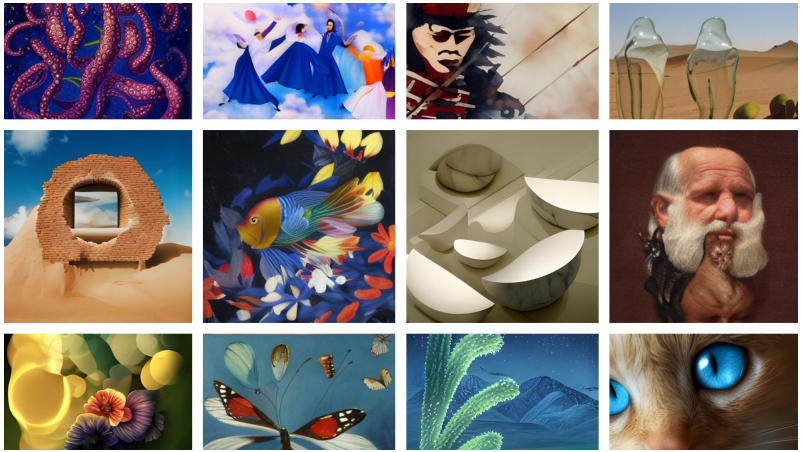
# AIL 722: Reinforcement Learning

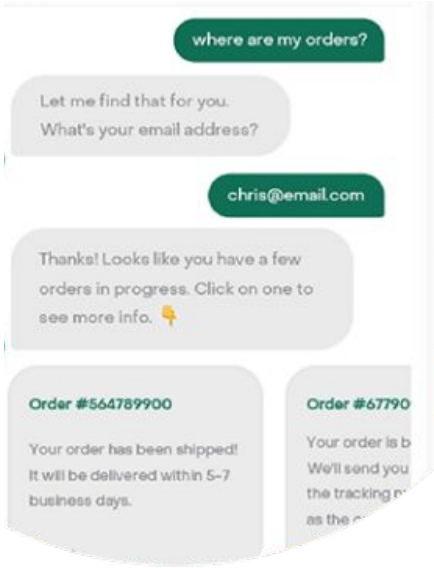
## Lec 1: Course Introduction

Raunak Bhattacharyya

# Recent Advances in AI



Source: Meta-AI

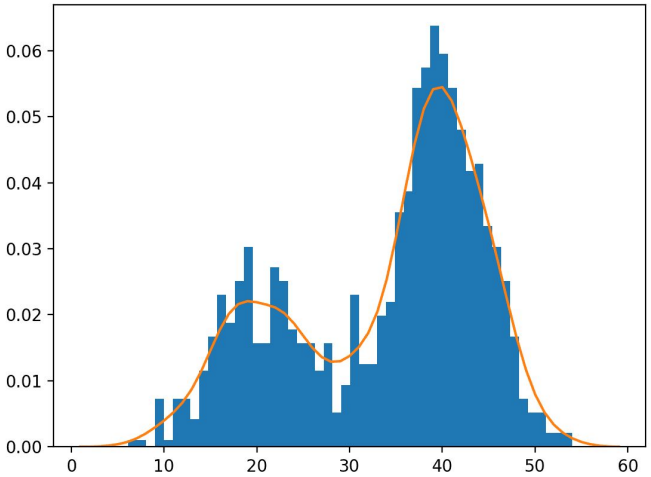



Source: Hootsuite

# Core Idea

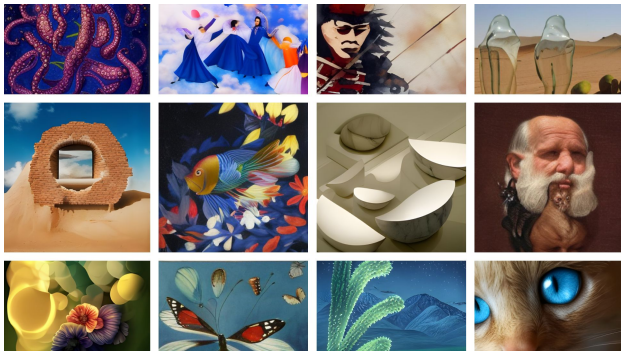
Source: Adobe

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

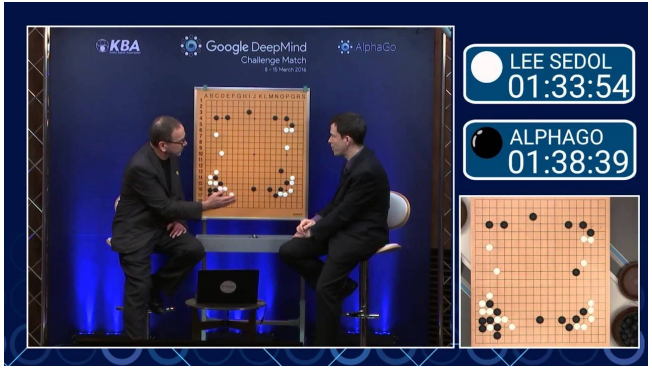



Source: MachineLearningMastery

# RL: Discovery



**Looks like something a person might draw!**




Source: Deepmind, DQN

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

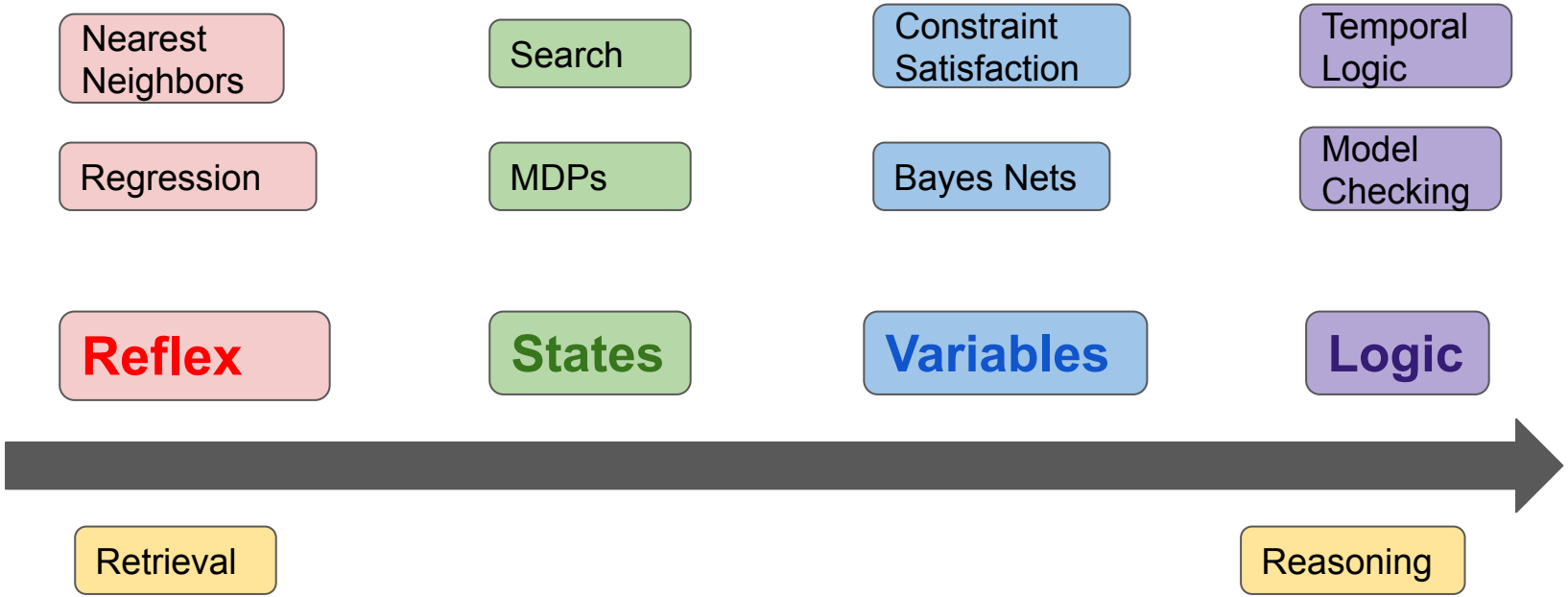
Adapted from Berkeley CS 285, Instructor: Sergey Levine

# What Is Reinforcement Learning

Mathematical formalism for learning-based decision making

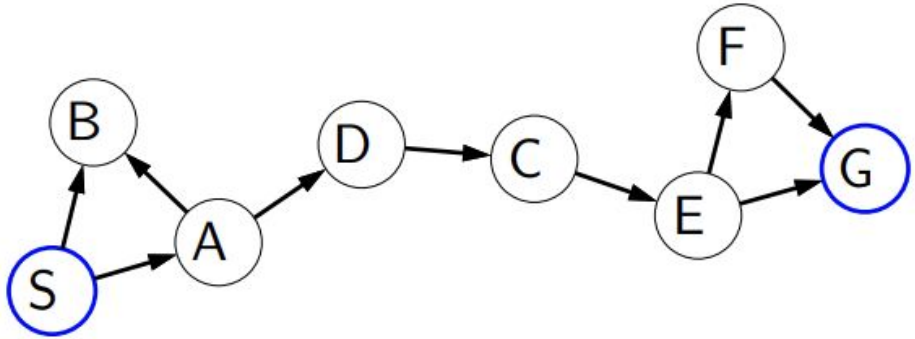
Approach for learning decision making and control from experience

# Contextualizing RL




Adapted from Stanford CS 221, Instructor: Percy Liang

# Search Problems



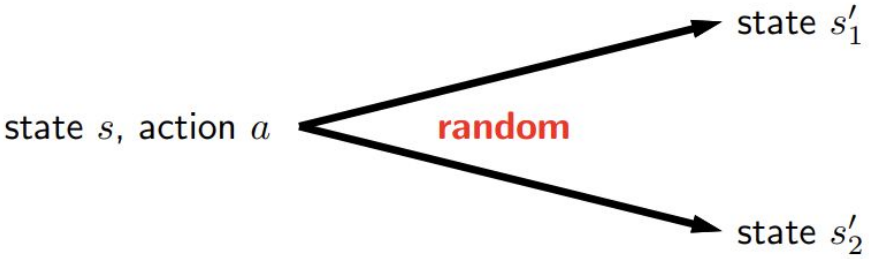

**deterministic**

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

# Uncertainty in the Real World

How other agents might behave

Source: istockphoto



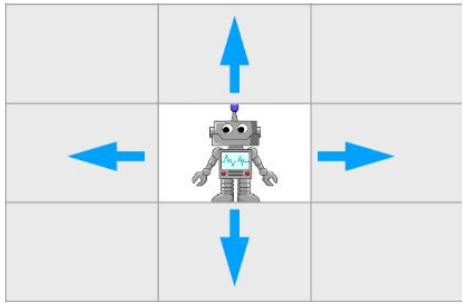
# Applications

Sensors

Demand

Weather

# Motivating Example



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

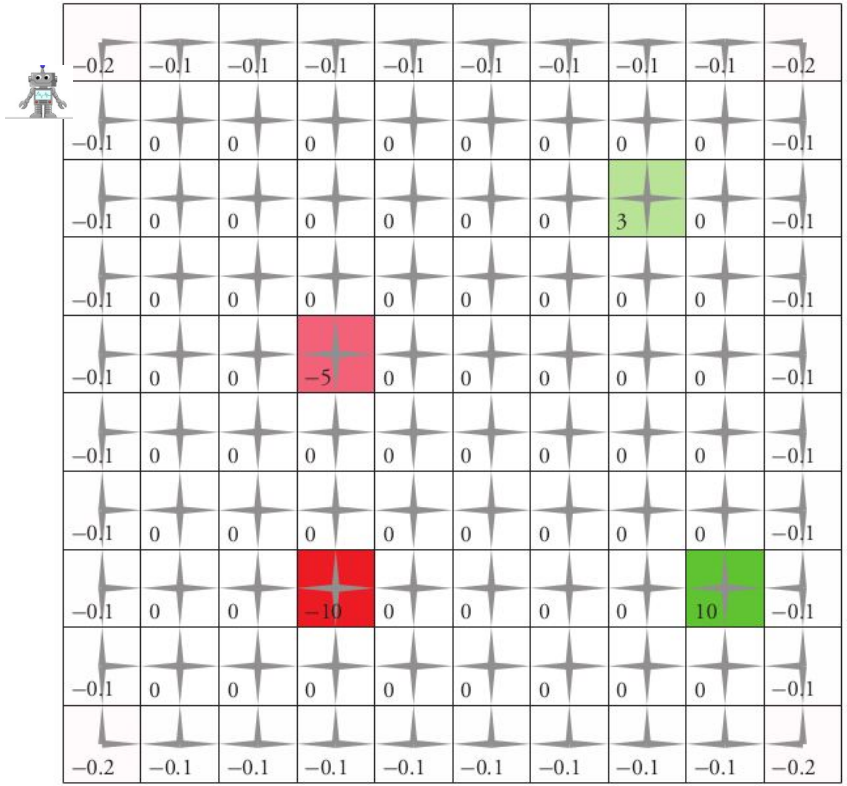
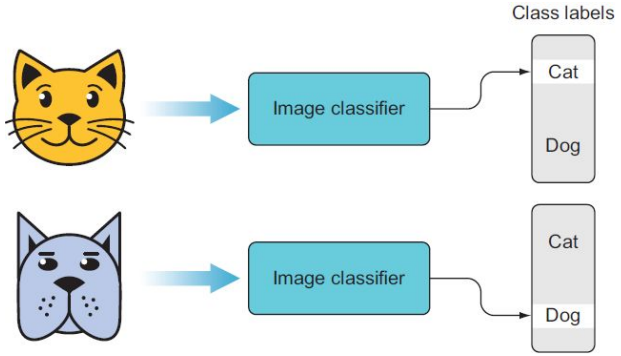
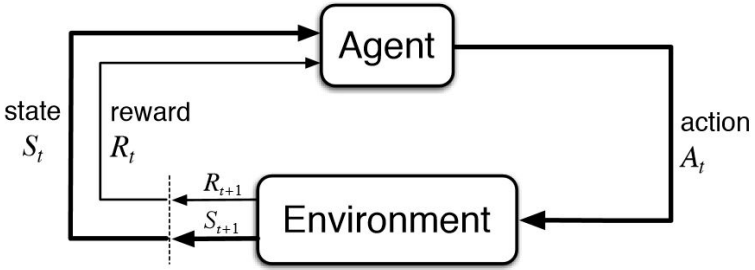



Figure from Algorithms for Decision Making, Mykel Kochenderfer

# 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

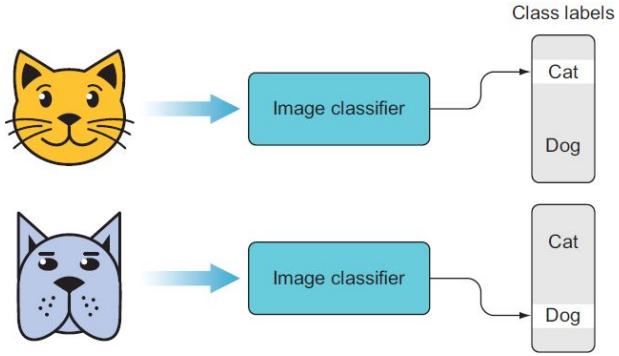
**Data:** $(s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_T, a_T, r_T)$

**Goal:** Learn policy $\pi_\theta : s_t \rightarrow a_t$

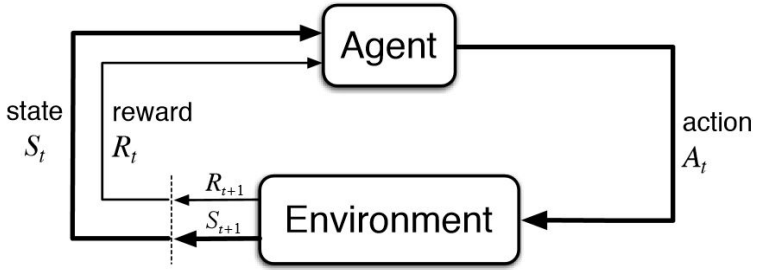
to maximize total reward obtained

**Pick your own action**

# Contrast to Supervised Learning




Source: Medium

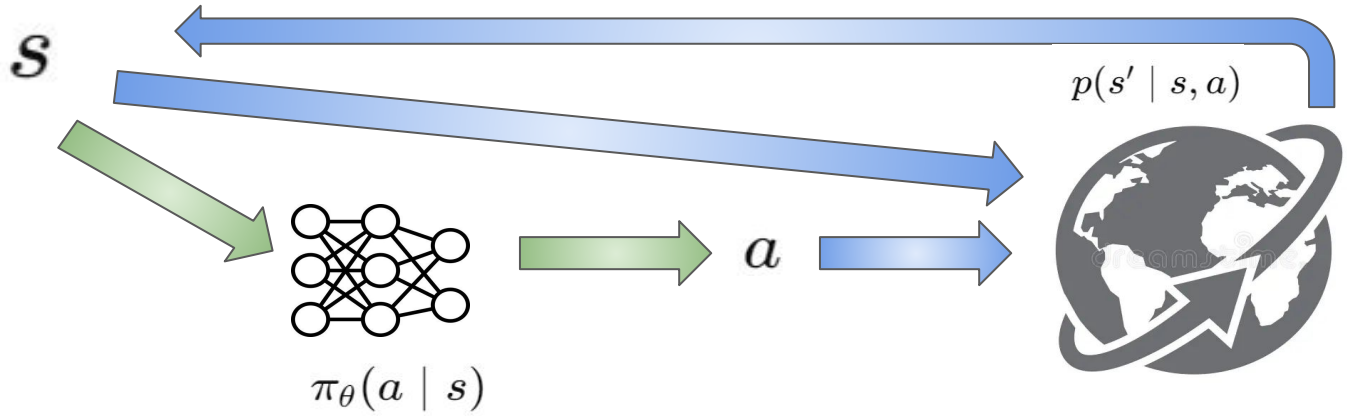



Source: Sutton & Barto

- 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




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

$p_\theta(\tau)$

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

# Learning Objectives

- Ability to recognize the applicability of RL, formulate problems as RL problems, choose the right algorithm, and implement said algorithm
- Get a broad perspective on RL
- Understand the ‘why’ of RL algorithms
- Exposure to standard RL software and benchmarks
- Ability to implement RL algorithms