به نام خدا

Iran University
of Science and
Technology

یادگیری تقویتی در کنترل

دکتر سعید شمقدری

دانشکده مهندسی برق
گروه کنترل

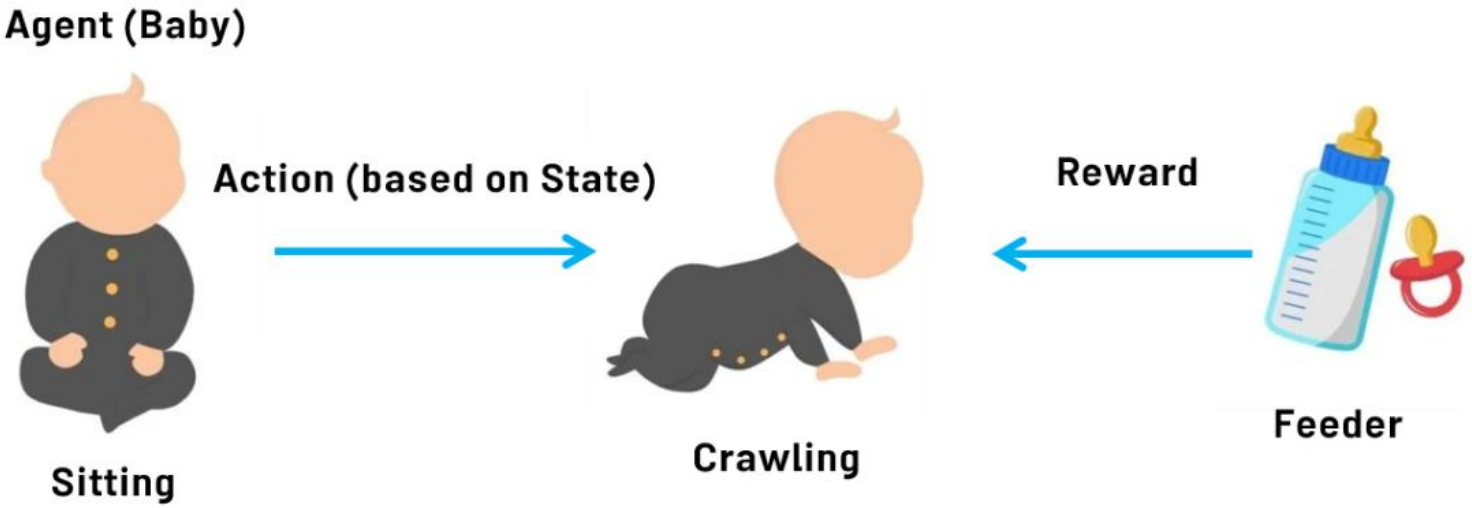
نیمسال اول ۱۴۰۵-۱۴۰۴

Introduction

یادگیری تقویتی: یادگیری از طریق تعامل

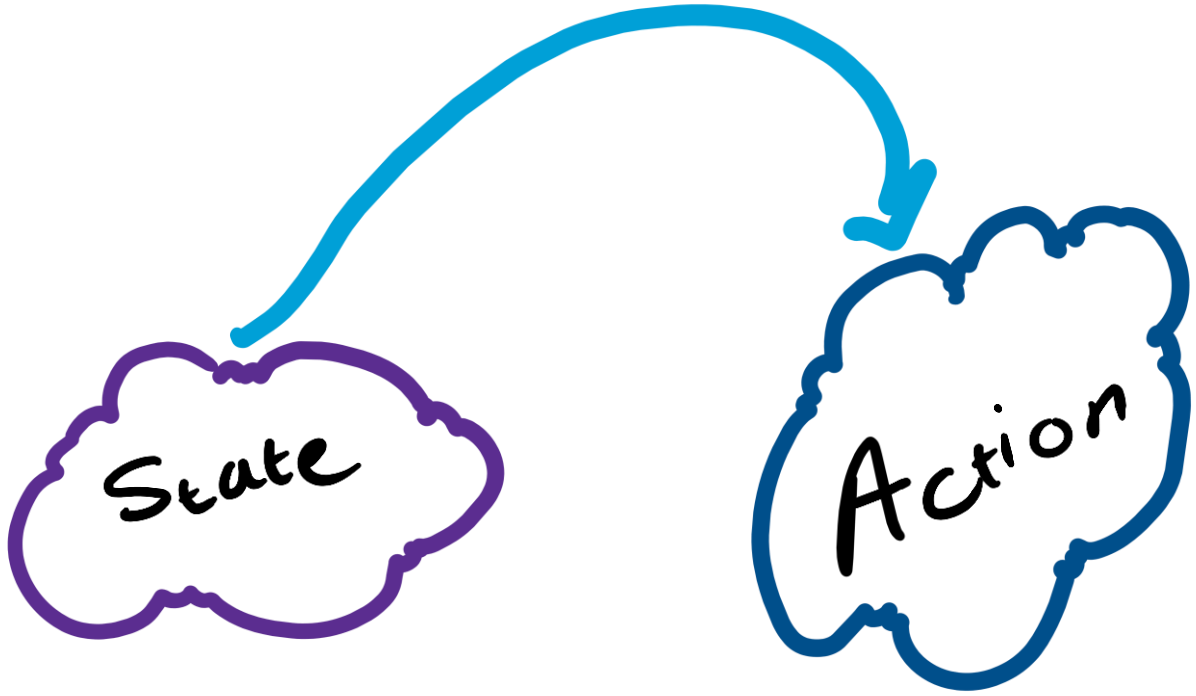
طبیعی‌ترین نوع یادگیری: یادگیری از طریق تجربه

راه‌رفتن چهارپایان
مذاکره با یک فرد
رانندگی با یک خودرو (قدیمی/مدرن)
بازی کردن نوزاد



جمع آوری اطلاعات: آگاهی از نتیجه یک سری Action ها
یادگیری: برای رسیدن به **هدف** چه Action ای باید انجام داده شود؟

RL: Goal Directed Learning



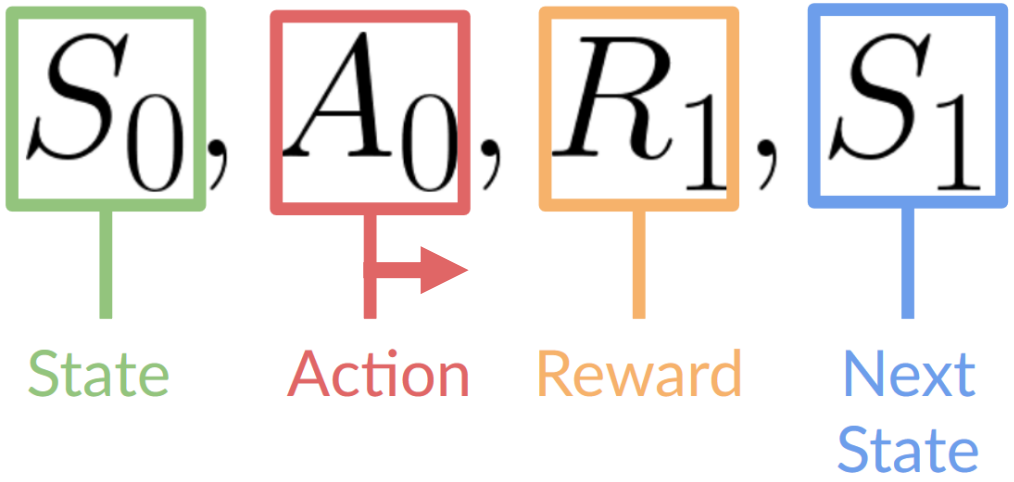
یادگیری تقویتی: نگاشتی از State به Action

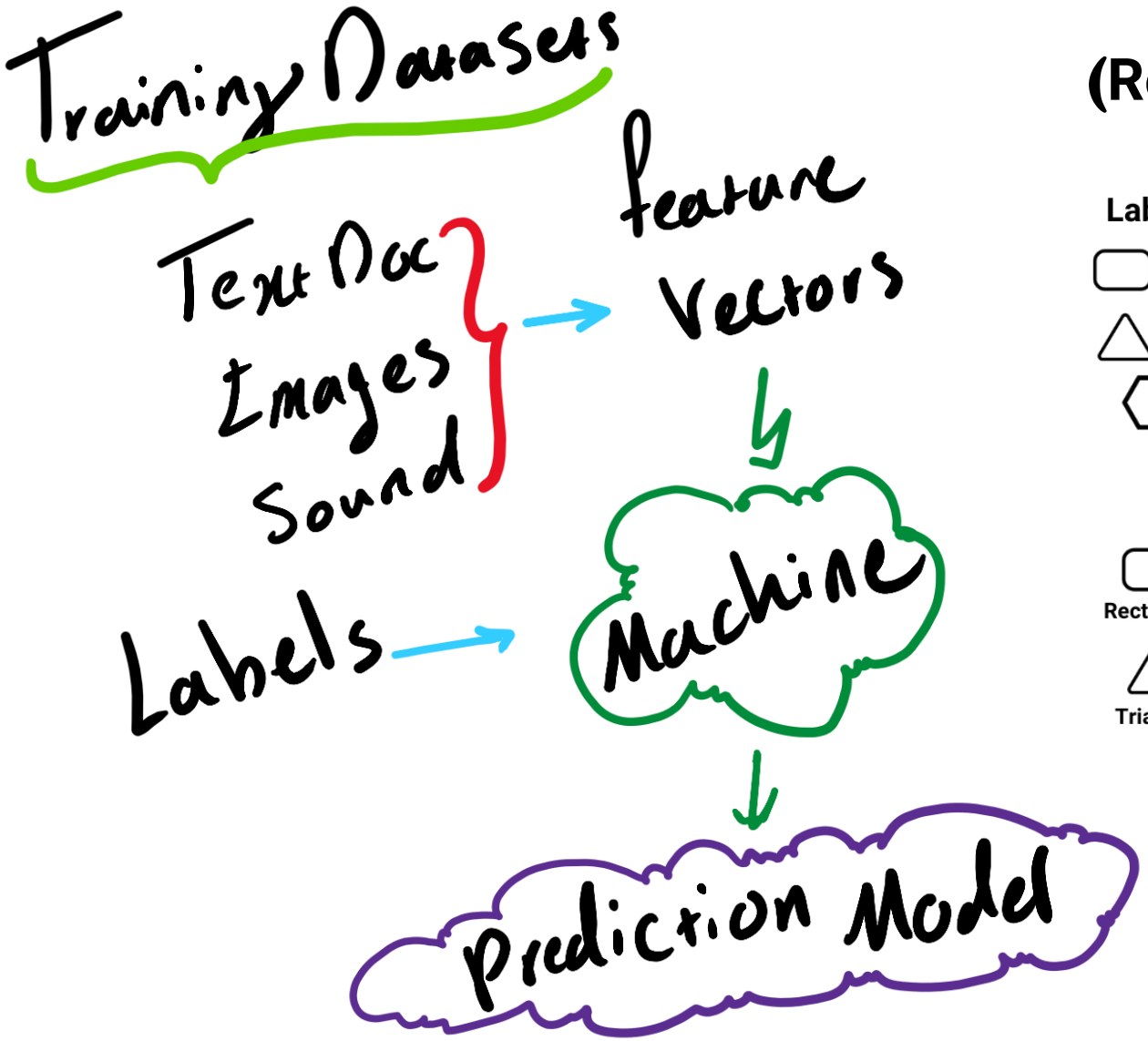
هدف: ماکزیمم کردن پاداش در دراز مدت

حل با ایده های کنترل بهینه

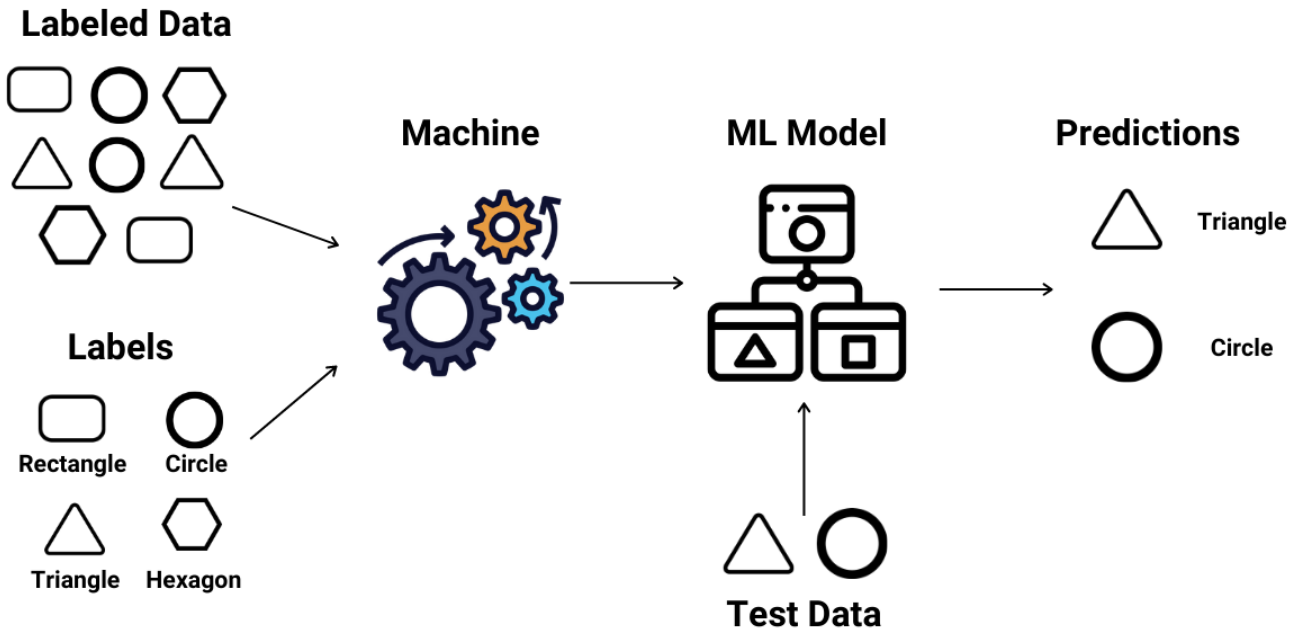
تاثیر انتخاب Action:

reward لحظه بعد، state لحظه بعد و ....





یادگیری نظارتی (Regression, Classification)

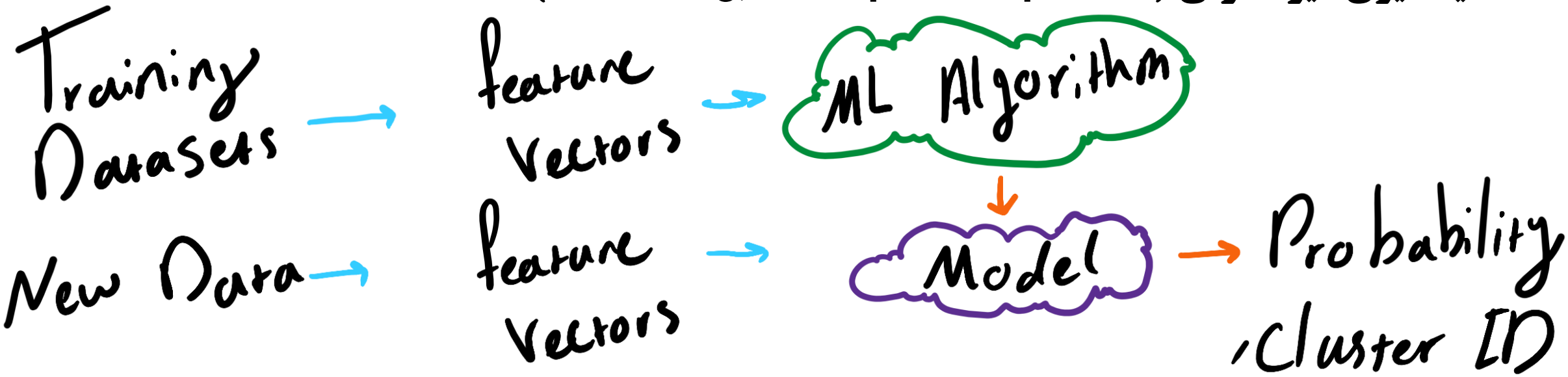


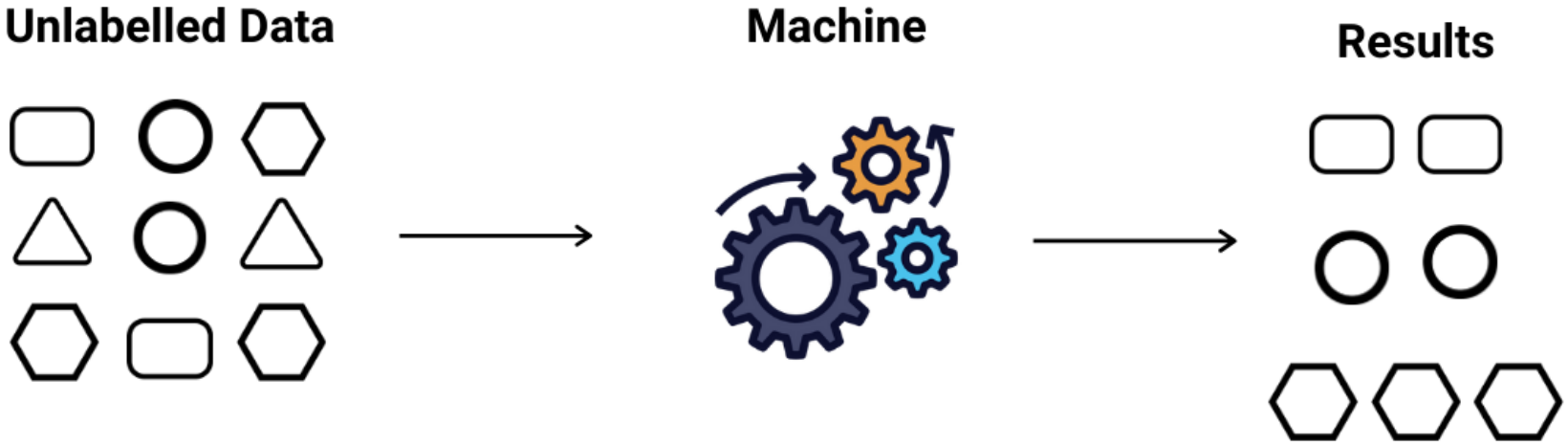
RL، متفاوت با یادگیری نظارتی

عدم نیاز به دیتاست لیبل دار

عدم دسترسی به پاسخ صحیح

یادگیری غیرنظارتی (Clustering, Principal Component)





RL، متفاوت با یادگیری غیرنظارتی

عدم نیاز به یافتن ساختار

بیشینه کردن پاداش بجای توجه به ساختار دیتای سیستم

جمع‌بندی ...

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data	Labeled data	Unlabeled data	Environment and feedback
Goal	Learn mapping between input data and output labels	Discover patterns, relationships, or groupings	Learn policy to maximize cumulative reward

Reinforcement Learning: Life-Long Learning

الزامات Agent در یادگیری تقویتی

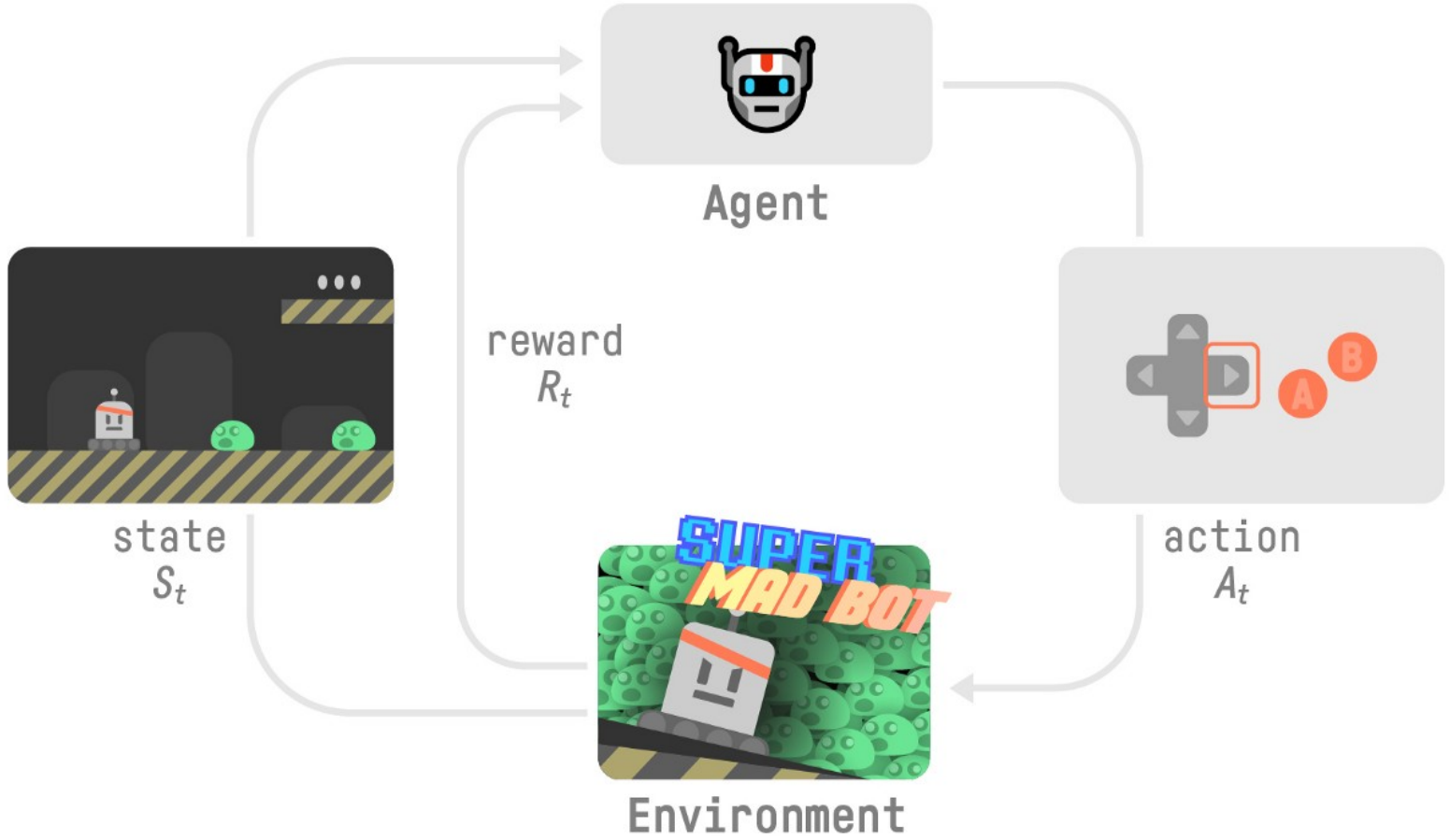
Sense

↓

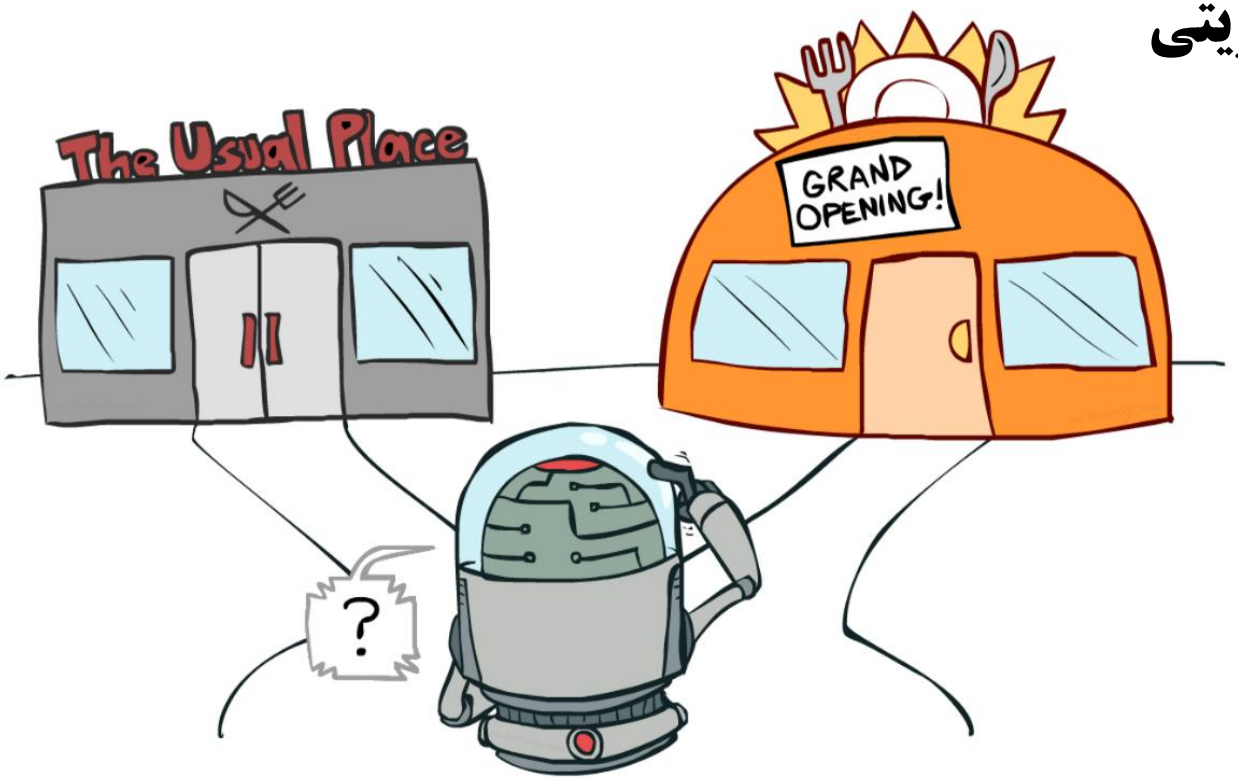
Act

↓

Goal



چالش‌های یادگیری تقویتی



Exploitation
Exploration

حرکت به سمت بهینه برای بیشینه کردن reward بر اساس شناخت حاصل تا لحظه جاری

شناخت بهتر سیستم (Stochastic/deterministic) در طول زمان

چالش‌های یادگیری تقویتی

Exploitation حرکت به سمت بهینه برای بیشینه کردن reward بر اساس شناخت حاصل تا لحظه جاری

Exploration شناخت بهتر سیستم (Stochastic/deterministic) در طول زمان

انتخاب Action های بهتر از قبل

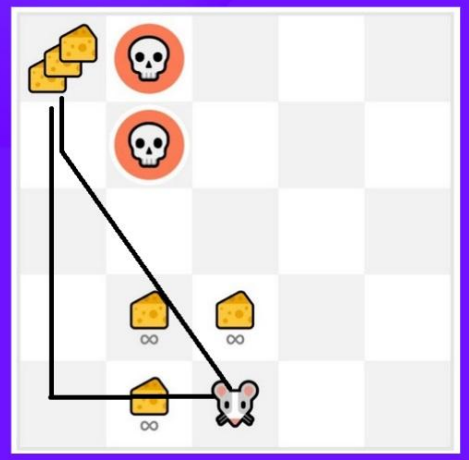
در سیستم‌های Stochastic:

استفاده از یک Action چندین بار: تخمین Expected Reward

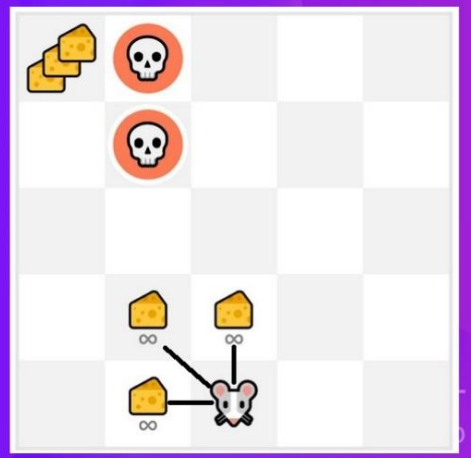
جمع‌بندی ...

Exploration/ Exploitation tradeoff

Exploration: trying **random actions** in order to find **more information about the environment.**

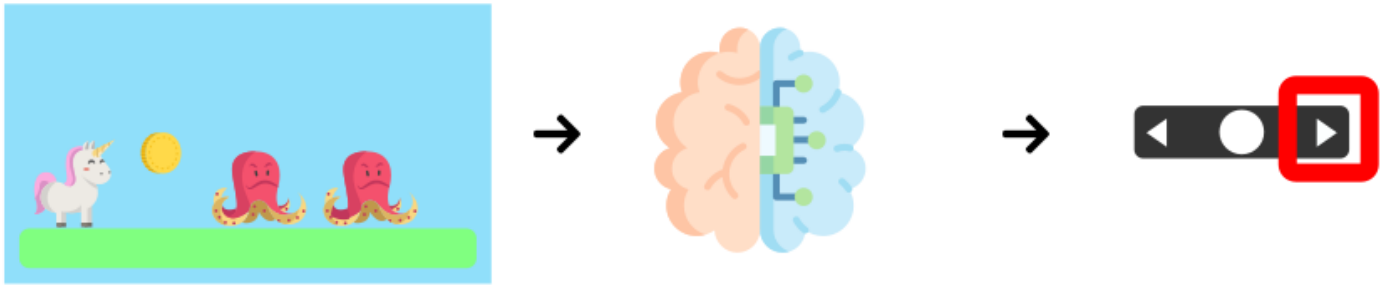


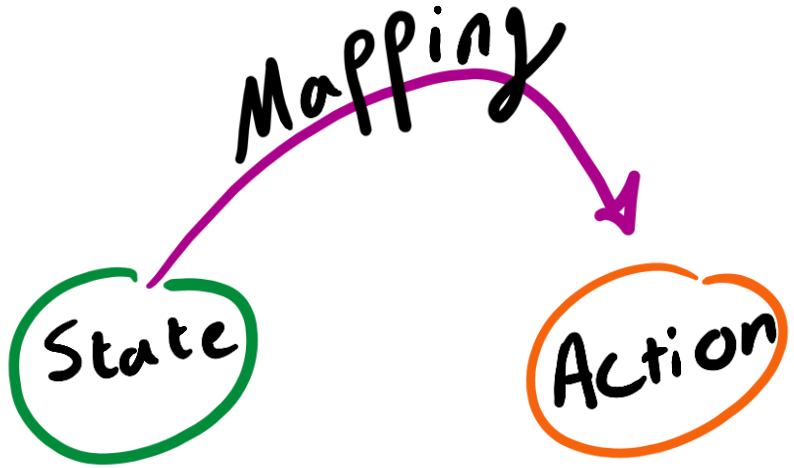
Exploitation: using known information to **maximize the reward.**



چهار المان اصلی یادگیری تقویتی

1. Policy





State → π(State) → Action

- Function
- Look-up Table
- Search

Policy

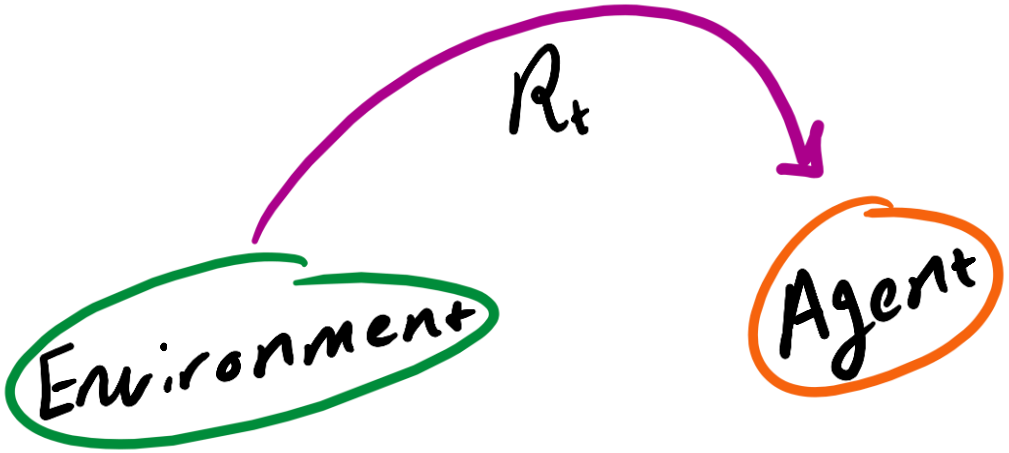
- Deterministic
- Stochastic

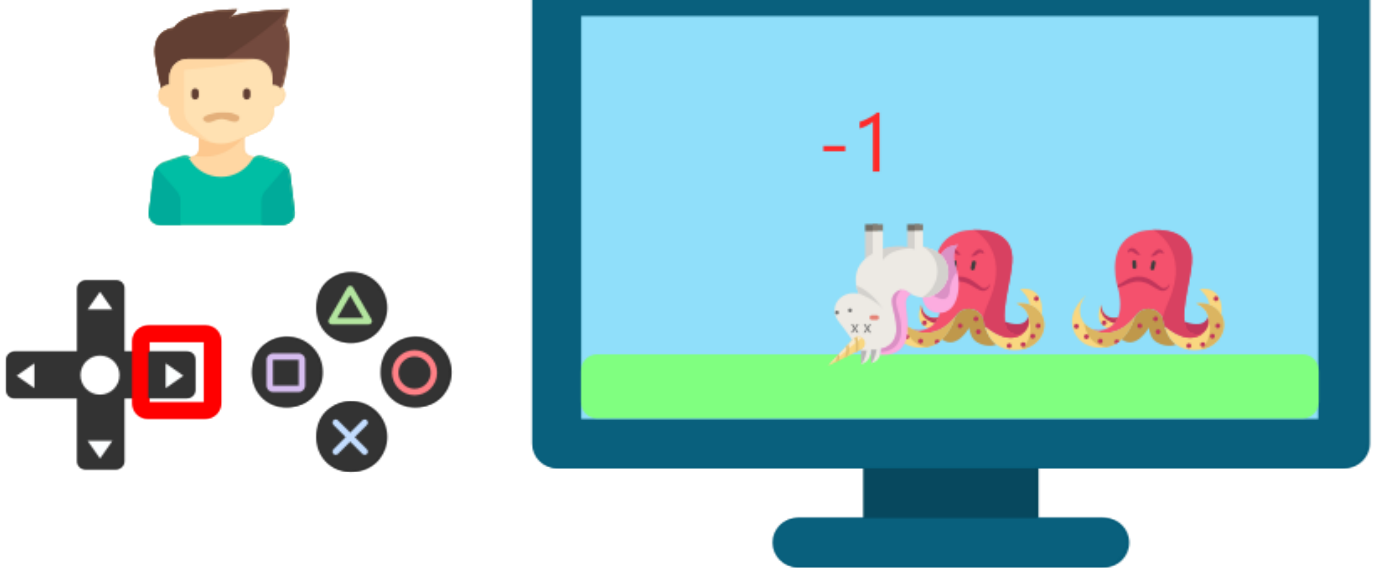
چهار المان اصلی یادگیری تقویتی

2. Reward Signal

هدف Agent:

بیشینه کردن Total Reward





Reward
- Deterministic
- Stochastic

چهار المان اصلی یادگیری تقویتی

3. Value Function

Reward: پاداش لحظه‌ای (خوب لحظه‌ای)

Value: پاداش درازمدت (خوب دراز مدت)

Expected Rewards

انسان

Reward: لذت یا ناراحتی لحظه‌ای

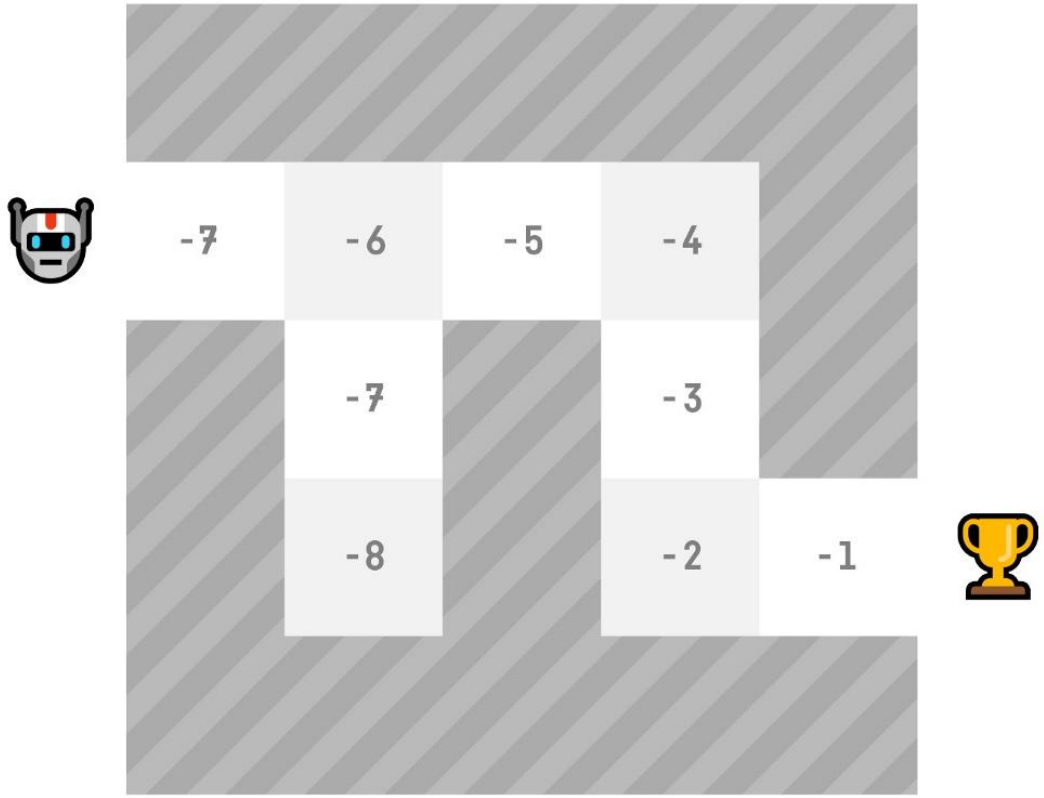
Value: قضاوت بلندمدت از رضایت/عدم رضایت

وَعَسَى أَنْ تَكْرَهُوا شَيْئًا وَهُوَ خَيْرٌ لَكُمْ وَعَسَى أَنْ تُحِبُّوا شَيْئًا وَهُوَ شَرٌّ لَكُمْ

Yet it may be that you dislike something, which is good for you, and it may be that you love something, which is bad for you.

چهار المان اصلی یادگیری تقویتی

3. Value Function



معیار انتخاب Action، Reward یا Value؟

چالش Value:

روش محاسبه یا تخمین

چهار المان اصلی یادگیری تقویتی

4. Model

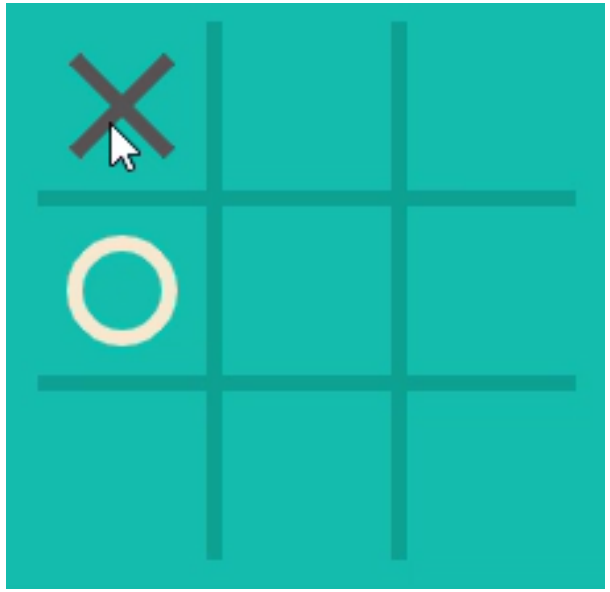
Deterministic
Stochastic
} بیان رفتار محیط {
Known
Unknown

$$S_t \rightarrow A_t \xrightarrow{\text{Model}} \begin{cases} S_{t+1} \\ A_{t+1} \end{cases}$$

چهار المان اصلی یادگیری تقویتی

یادگیری تقویتی و روش‌های تکاملی

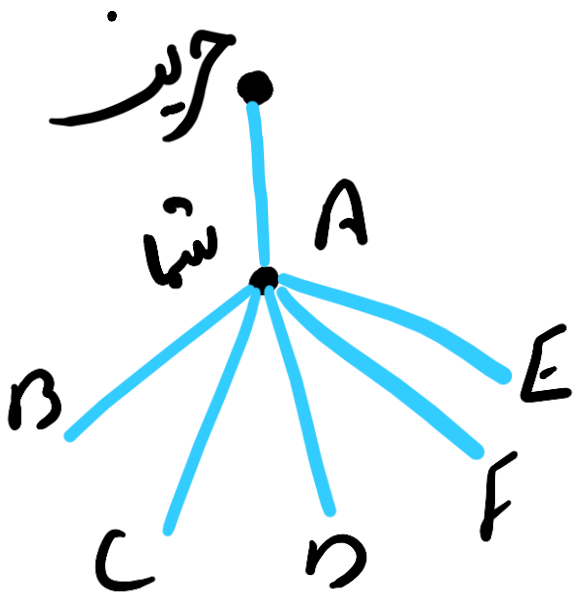
عدم توجه به جزئیات پالیسی در روش‌های تکاملی



عملکرد بازیکن ماهر؟
براساس Game Theory

فرض: رقیب غیر حرفه‌ای

تعریف State در بازی X-O؟
وضعیت مهره ها + نوبت کی؟



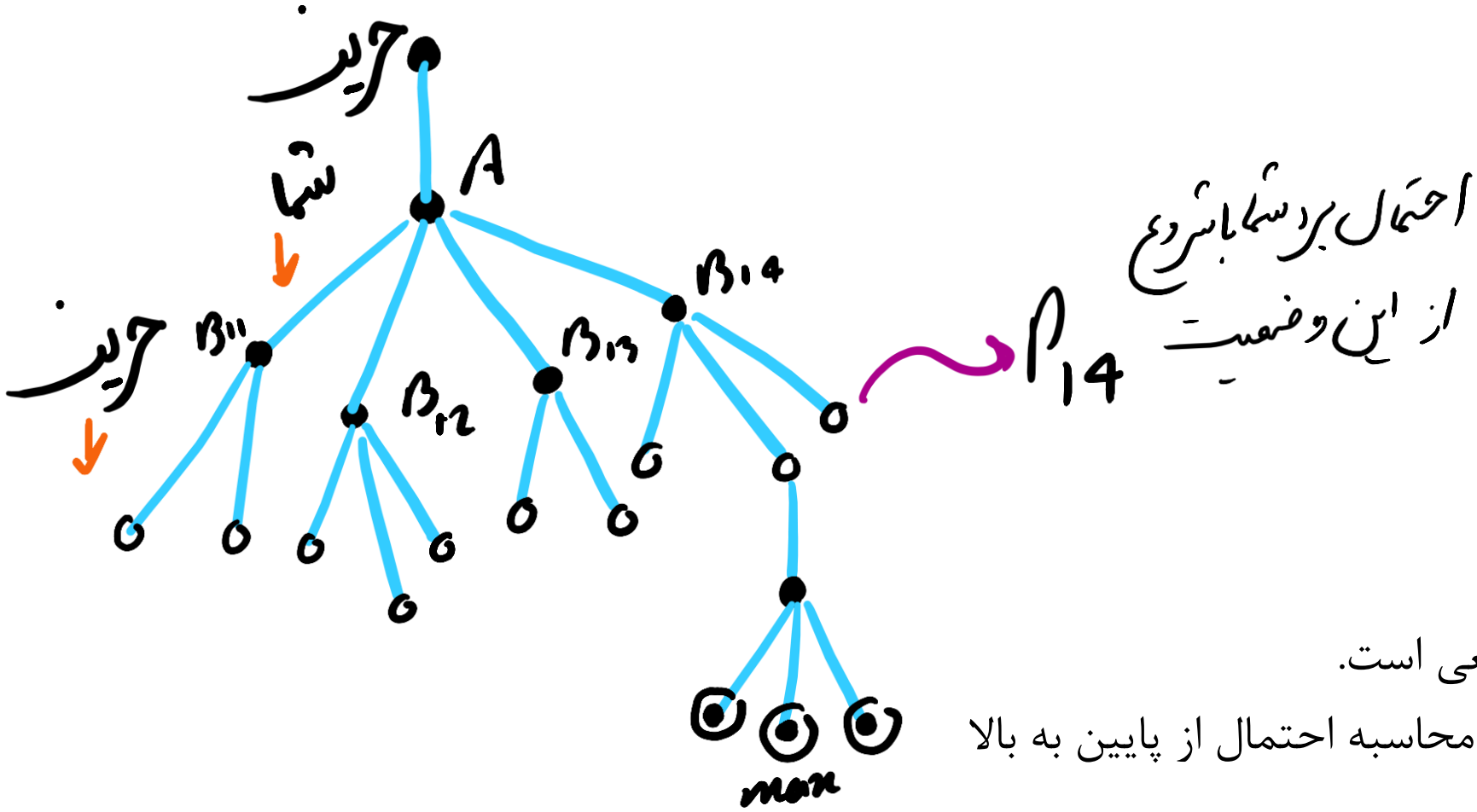
تشکیل درخت بازی

یکی از B تا F برنده باشد : A برنده
همه B تا F بازنده باشند : A بازنده

روش Dynamic Programming:
نگاه به sate نهایی در درخت بازی و حرکت از پایین به بالا

تعیین انتخاب مناسب برای برنده شدن

تشکیل درخت بازی



توجه: آخرین ردیف احتمال برد قطعی است.

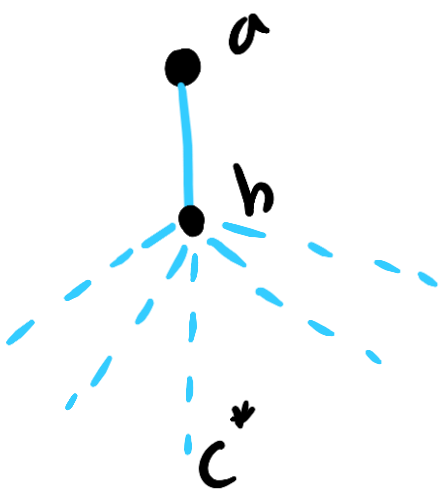
Dynamic Programming: محاسبه احتمال از پایین به بالا

حل با یادگیری تقویتی

Temporal Difference

تهیه جدول Value که هر سطر آن یک State است
Init کردن جدول (Value)
(مثال :
یک سطر O : Pr=0
یک سطر X : Pr=1
سایر : Pr=0.5)

انتخاب یک پالیسی
به روز رسانی Value با مشاهدات



الگوریتم بازی براساس TD:

حرکت رقیب از a به b
تخمین تابع Value برای حرکت از b
انتخاب حریصانه (Greedy) : حرکت از b به c*
دریافت Reward جدید و محاسبه V(S(t+1))
بروز رسانی جدول با مشاهدات بازی:

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[V(S_{t+1}) - V(S_t)\Big]$$

انجام Exploration: انتخاب حرکت های غیر بهینه به صورت رندوم
مثال : (رفتن به رستوران – اکتشاف نفت)
توجه : عدم به روز رسانی جدول

اثبات همگرایی؟

جمع‌بندی: حل با یادگیری تقویتی