物流大数据、优化问题和智能物流优化

姚尧　博士，副教授

地理与信息工程学院，地图制图学与地理信息工程

阿里巴巴集团，访问学者

Email: yaoy@cug.edu.cn

办公地点：未来城校区地信楼522办公室

主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

01 城市物流优化

我们的快递需要多久才可以到?
它们跨域了几个国家? 几个城市?
我们是否满意快递公司的服务?

能否思考一下它们的具体流程以及涉及的理论知识?

01 城市物流优化

往年“双十一”期间频繁爆仓？为什么现在很少听说“爆仓”？为什么？

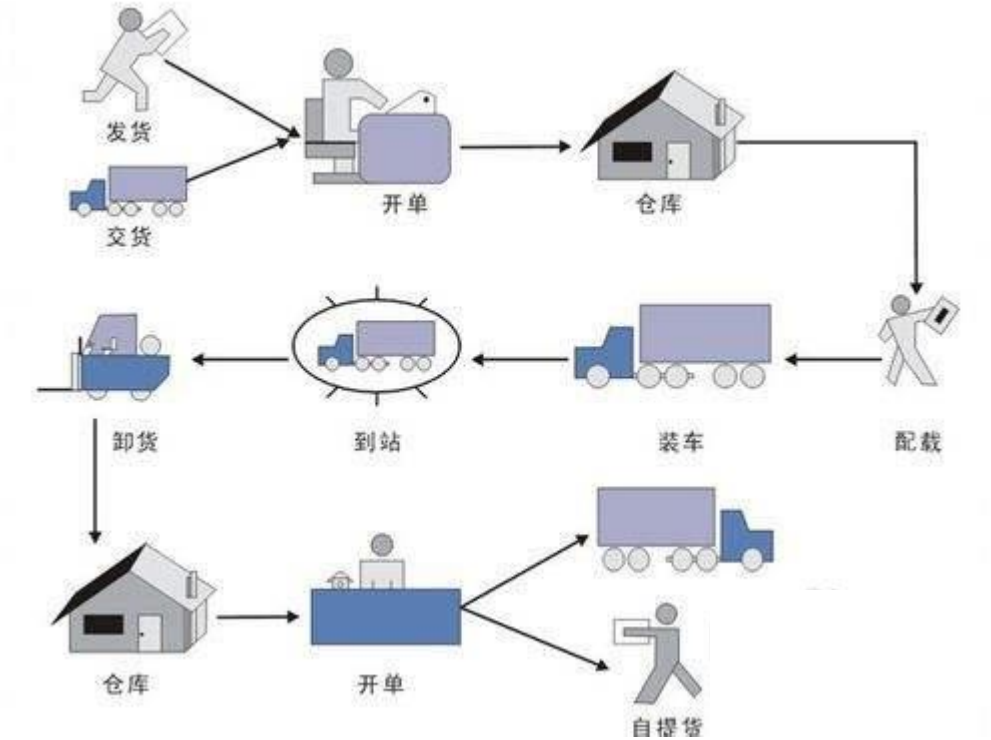
京东物流 JD Logistics

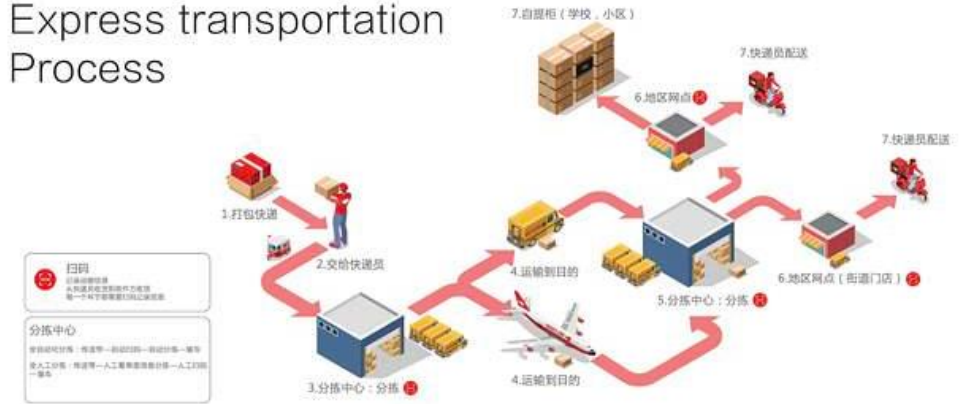
SF EXPRESS 顺丰速运

CAINIAO 菜鸟

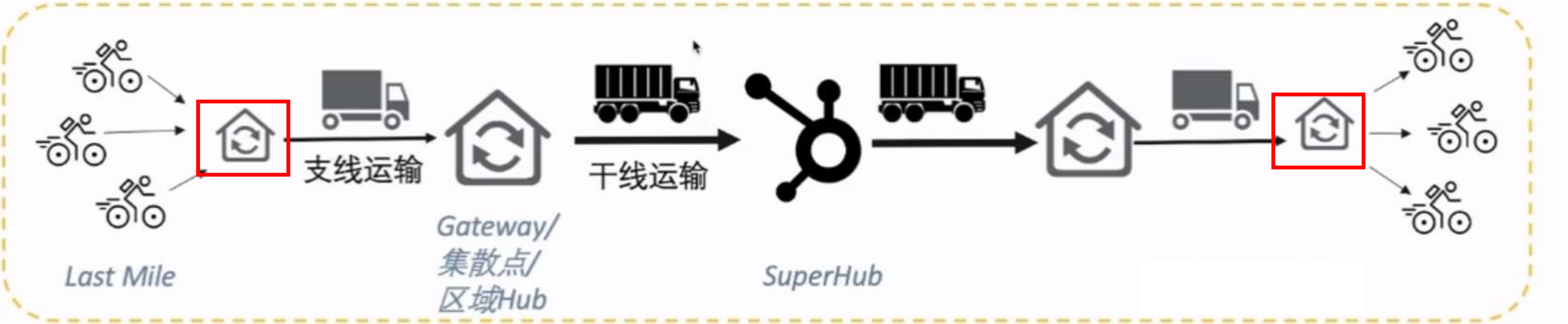
01 城市物流优化

城市物流配送：城市物流配送是指在城市内，对货物进行运输、存储、包装、流通加工、装卸搬运等作业，并按顾客的要求在正确的时间、正确的地点把正确的货物送到顾客手中的物流活动。





01 城市物流优化



城市物流基本形式

01 城市物流优化

小型“分拨中心”即“前置仓”连接着城市“物流中心”，随着城市物流网络的扩展，“前置仓”如何扩展呢？运输路线如何确定呢？

- **要求**

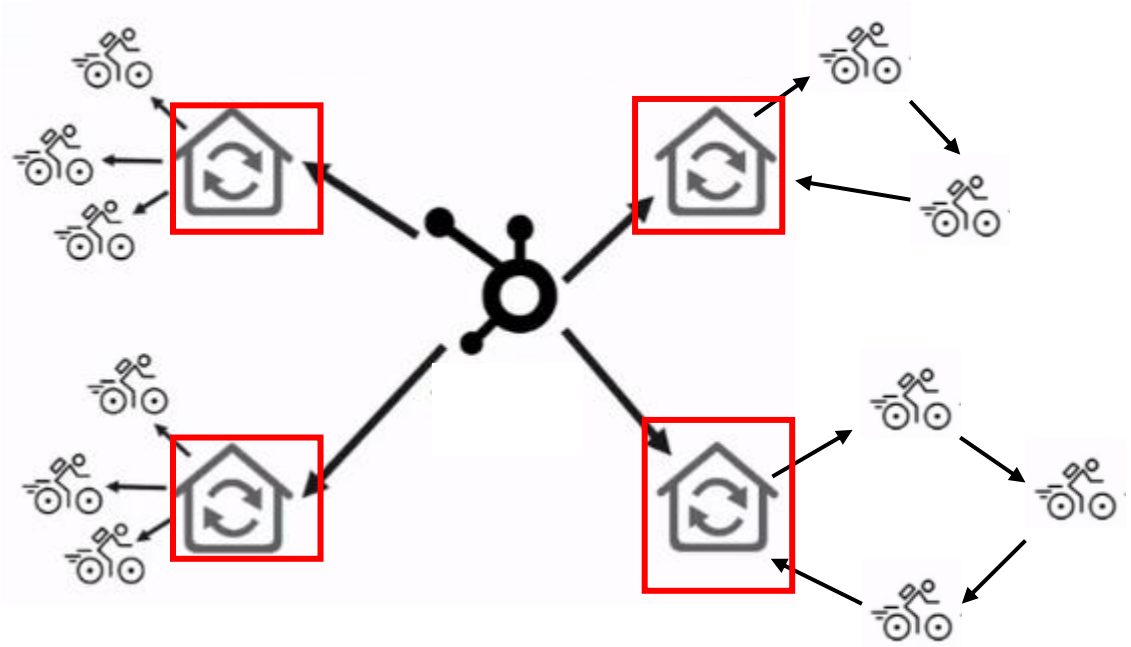
成本最小

路线最短

时效性强

交通便利

……



01 城市物流优化

城市物流优化：在市场经济的框架下，综合考虑城市内的交通环境、交通拥堵和能源消耗等因素，对城市内的物流和运输活动进行优化的过程。

- **优化算法**（在各种约束条件下，能够使目标达到最好）

  蚁群算法

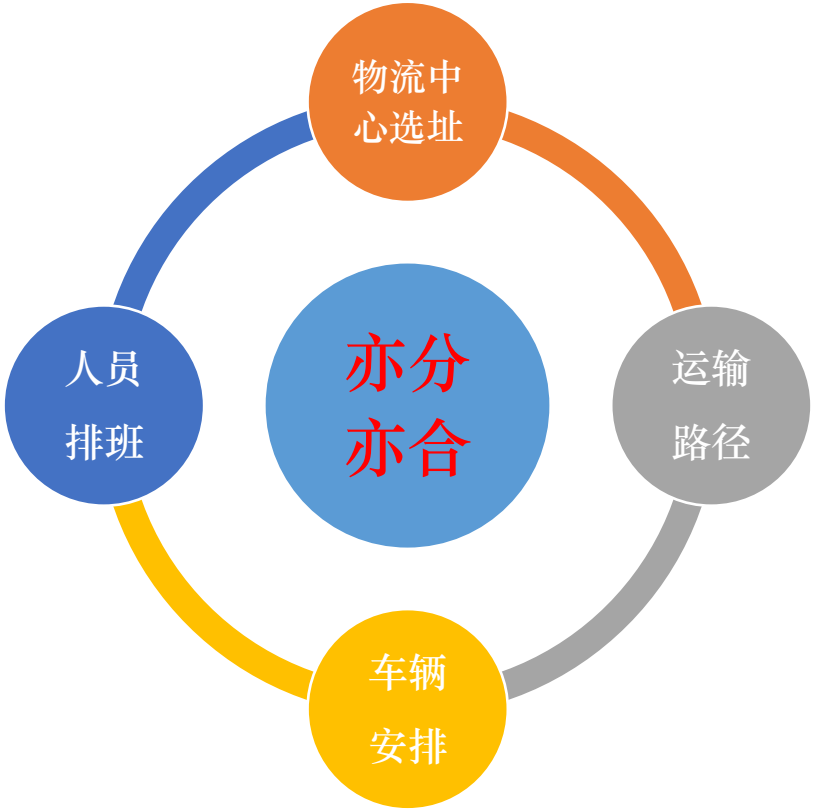
  遗传算法

  强化学习

  ……

01 城市物流优化

城市物流优化主要涉及：物流中心选址、车辆安排、运输路径、人员排班等主要问题



01 城市物流优化

- 具体怎么做?

效能评估

业务场景

运营数据

运营要求

中、长期资源与营运规划

短期资源与营运调度策略

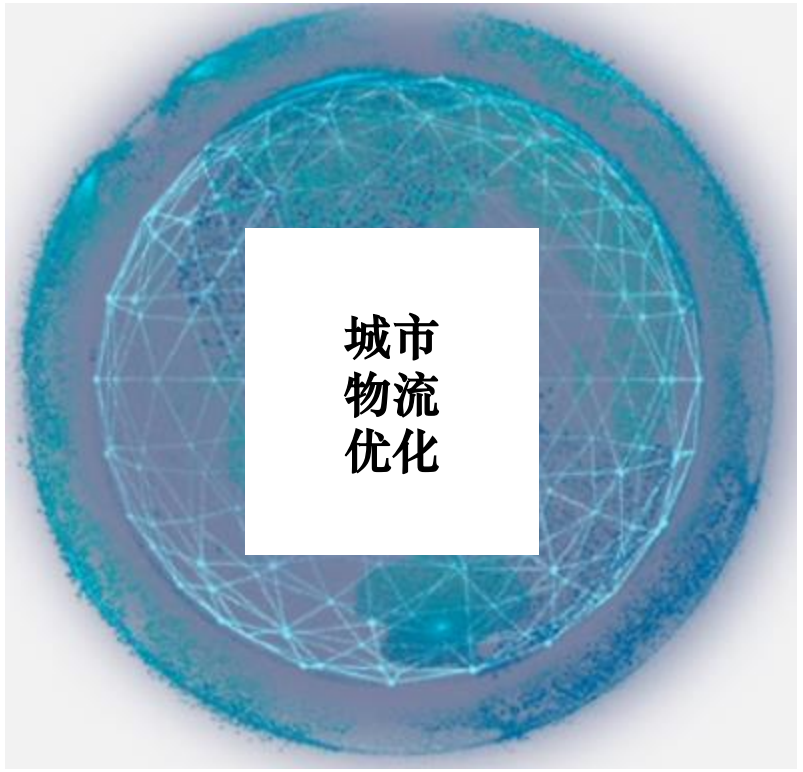
01 城市物流优化

➢ 具体怎么做?

业务场景:
全面而深入的理解

工程化:
打通数据、模型算法、落地应用全流程

两支撑



三要素

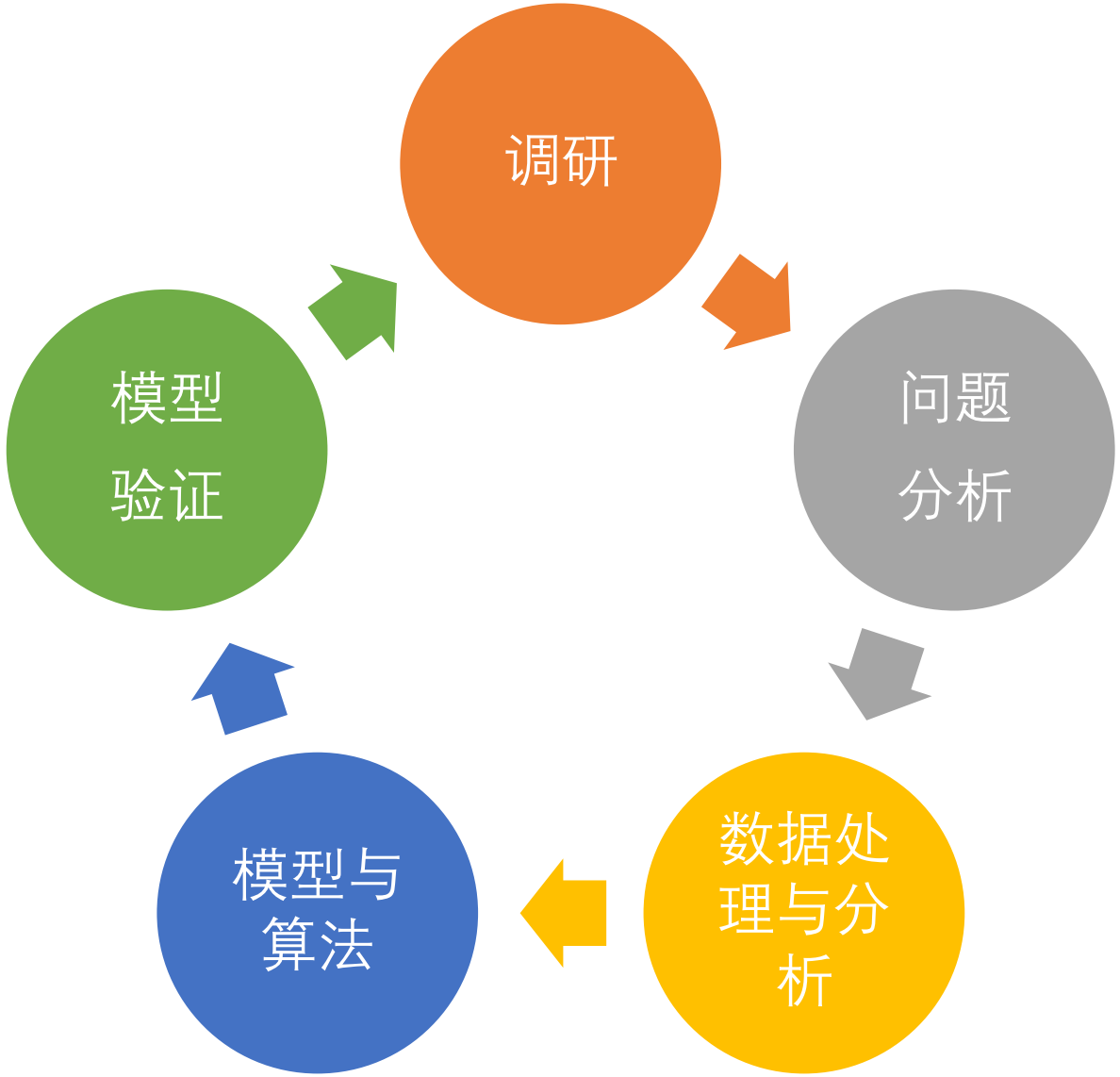
数据:
哪些数据、如何获取、数据校验

模型:
哪类模型、如何建立、模型校验

算法:
哪类算法、如何开发、有效性验证

01 城市物流优化

- 具体怎么做?



主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

02 物流大数据

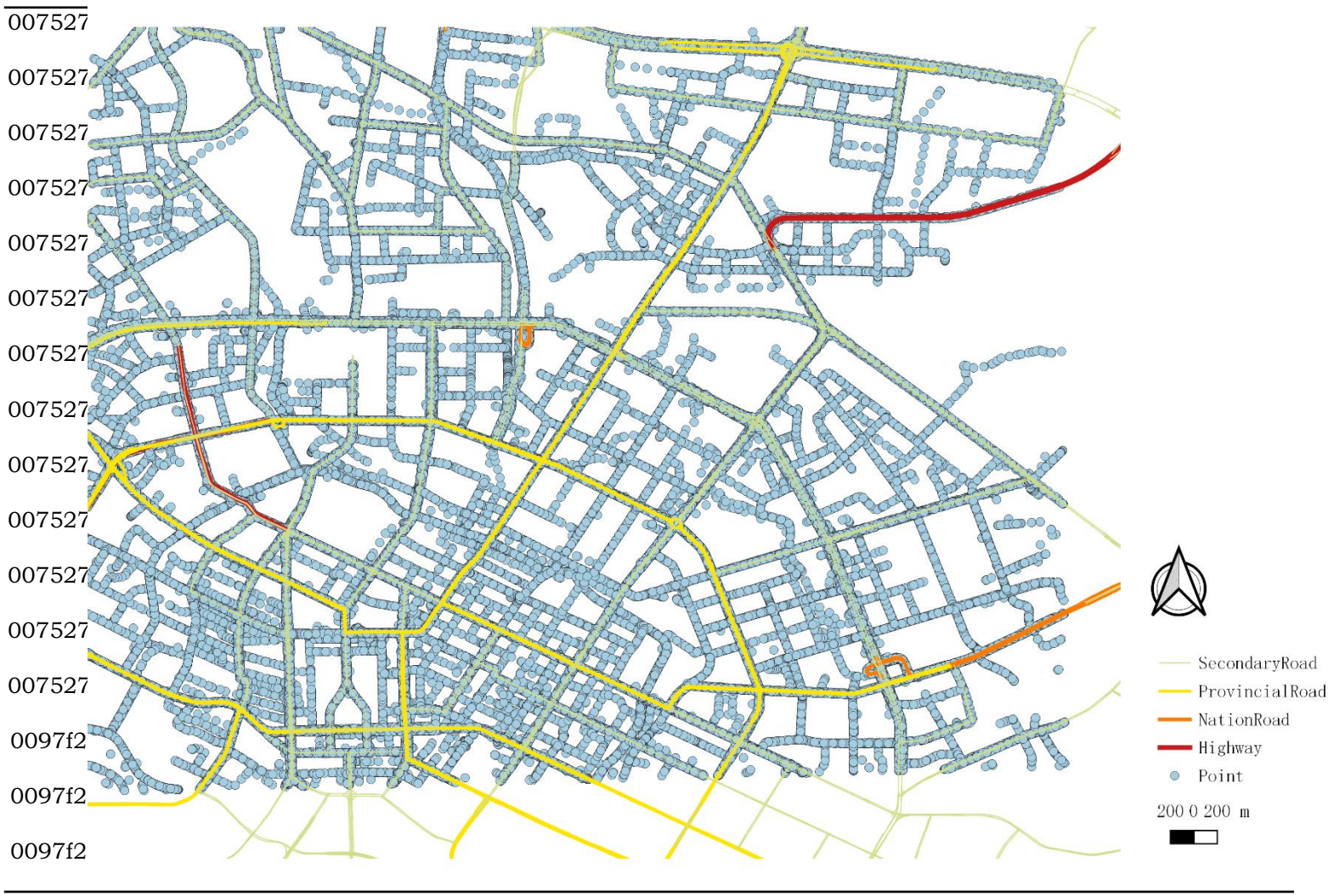
- **城市路网数据**（OSM:https://outreach.didichuxing.com/research/opendata/en/）

  - 高速公路、省道等多级路网；
  - 需要对路网进行简化、拓扑检查等；

02 物流大数据

➢ 出租车数据（滴滴打车:https://www.openstreetmap.org/）

- 每隔3-5s记录一次；
- 某市三环以内，每天12G的数据量；
- 后续根据任务需要进行抽稀等处理；



02 物流大数据

➢ 菜鸟驿站数据

provinceCode	cityCode	countyCode	corpCode	corpName	corpType	latitude	longitude	networkName	networkNo
420000	420100	420102	屏蔽		0	屏蔽			
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				
420000	420100	420102			0				

02 物流大数据

➢ 订单数据

- 订单具体需求；
- 订单地址等信息；

id	name	time	address	nums	weight	capitable
屏蔽				1	15	0.0756
				2	30	0.1512
				3	45	0.2268
				4	60	0.3024
				5	75	0.378
				2	30	0.1512
				3	45	0.2268
				4	60	0.3024
				5	75	0.378
				2	30	0.1512
				3	45	0.2268
				4	60	0.3024
				5	75	0.378

02 物流大数据

➢ 物流数据

仓库编码	运输地编码	件数	重量	体积	订单类型	日期
屏蔽				0.0756	LC	2016-06-11 18:00:00
				0.1512	LC	2016-06-11 18:30:00
				0.2268	LC	2016-06-11 18:30:00
				0.3024	LC	2016-06-11 18:30:00
				0.378	LC	2016-06-11 18:50:00
				0.1512	LC	2016-06-11 18:50:00
				0.2268	LC	2016-06-11 18:30:00
				0.3024	LC	2016-06-11 18:20:00
				0.378	LC	2016-06-11 18:10:00
				0.1512	LC	2016-06-11 18:30:00
				0.2268	LC	2016-06-11 18:10:00
				0.3024	LC	2016-06-11 18:40:00
				0.378	LC	2016-06-11 18:40:00

字段	说明
仓库编码	运输地中对应的仓库的编码
运输地编码	运输地中对应配送点的编码
件数	对应运输地对应的订单件数
重量	对应运输地对应的订单重量
体积	对应运输地对应的订单体积
未装载罚款	默认值，可不用修改
订单类型	对应的订单的类型(HW;恒温；LC：冷藏；LD：冷冻)
要求到达日期	订单最晚到达配送点的时间，非必填
备注	非必填
最早出发时间	订单最早可以从仓库发出的时间，非必填

主要内容

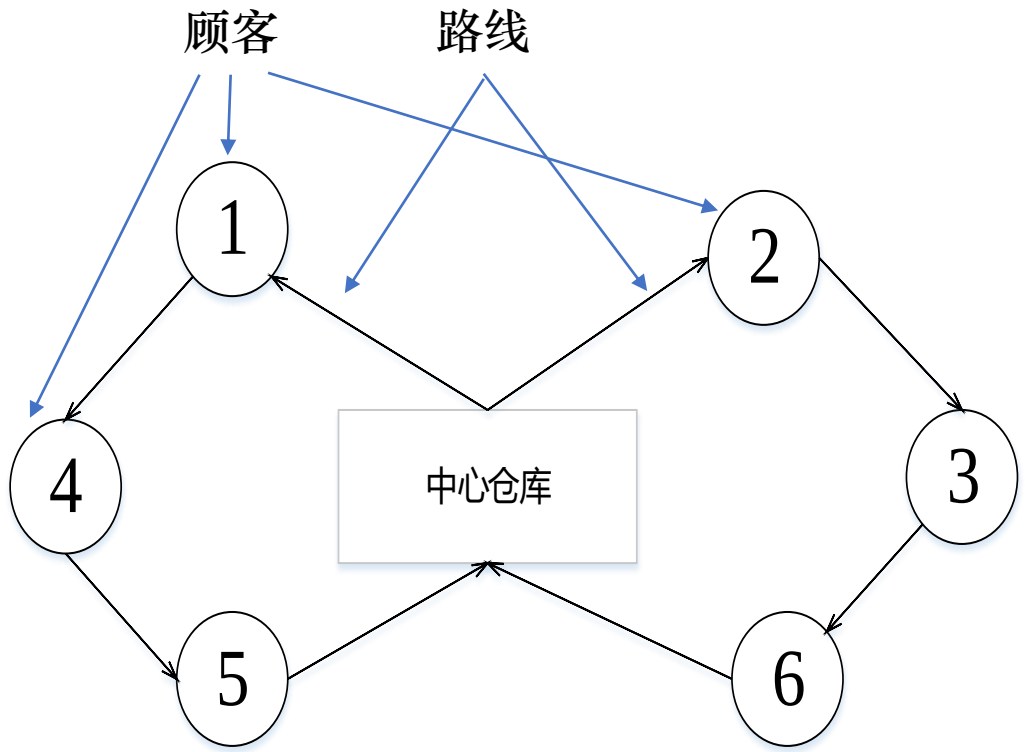
- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

03 优化问题

- **城市物流优化目标**
  - 运输车辆数目最小；
  - 出行距离最小；
  - ……
- **约束**
  - 物流中心；
  - 客户需求；
  - 车辆类型
  - ……

如何解决呢？

数学解决方法：可归纳为优化问题



03 优化问题

- **优化问题的一般形式**

$$\min \quad f(x) \tag{1.1}$$

(目标函数)

$$s.t. \quad h_i(x) = 0, i = 1,2,\dots,m \tag{1.2}$$

(等式约束)

$$g_j(x) \geq 0, j = 1,2,\dots,p \tag{1.3}$$

(不等式约束)

其中x是n维向量。

在实际应用中,可以将求最大值的目标函数取相反数后统一成公式中求最小值的形式。

03 优化问题

➢ 相关定义

可行解：满足约束条件（1.2）和（1.3）的称x为可行解,也称为可行点或容许点。

可行域：全体可行解构成的集合称为可行域，也称为容许集，记为D，即：

$$D = \{x | h_i(x), i = 1, \dots, m; g_j(x) \geq 0, j = 1, \dots, p, x \in R^n\}$$

若$h_i(x), g_j(x)$为连续函数，则D为闭集。

整体最优解：若$x^* \in D$,对于一切$x \neq x^*$,恒有$f(x^*) \leq f(x)$，则称x^*为最优化问题的整体最优解。

若$x^* \in D$,对于一切$x \neq x^*$,恒有$f(x^*) < f(x)$，则称x^*为最优化问题的严格整体最优解。

求解最优化问题，就是求目标函数$f(x)$在约束条件（1.2）和（1.3）下的极小点。

03 优化问题

➢ 城市物流优化的解决方法

解决方法

- ▲精确算法（二次型、分支定界算法、列生成算法等）
- ▲经典启发算法（节约算法、插入算法等）
- ▲智能启发算法（蚁群算法、模拟退火算法等）
- ▲强化学习（蒙特卡洛强化学习、时序差分强化学习等）

主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

04 启发算法

智能优化算法又称为现代启发式算法，是一种具有全局优化性能、通用性强、且适合于并行处理的算法。这种算法一般具有严密的理论依据，而不是单纯凭借专家经验，理论上可以在一定的时间内找到最优解或近似最优解。

04 启发算法

➢ 常见的智能启发算法

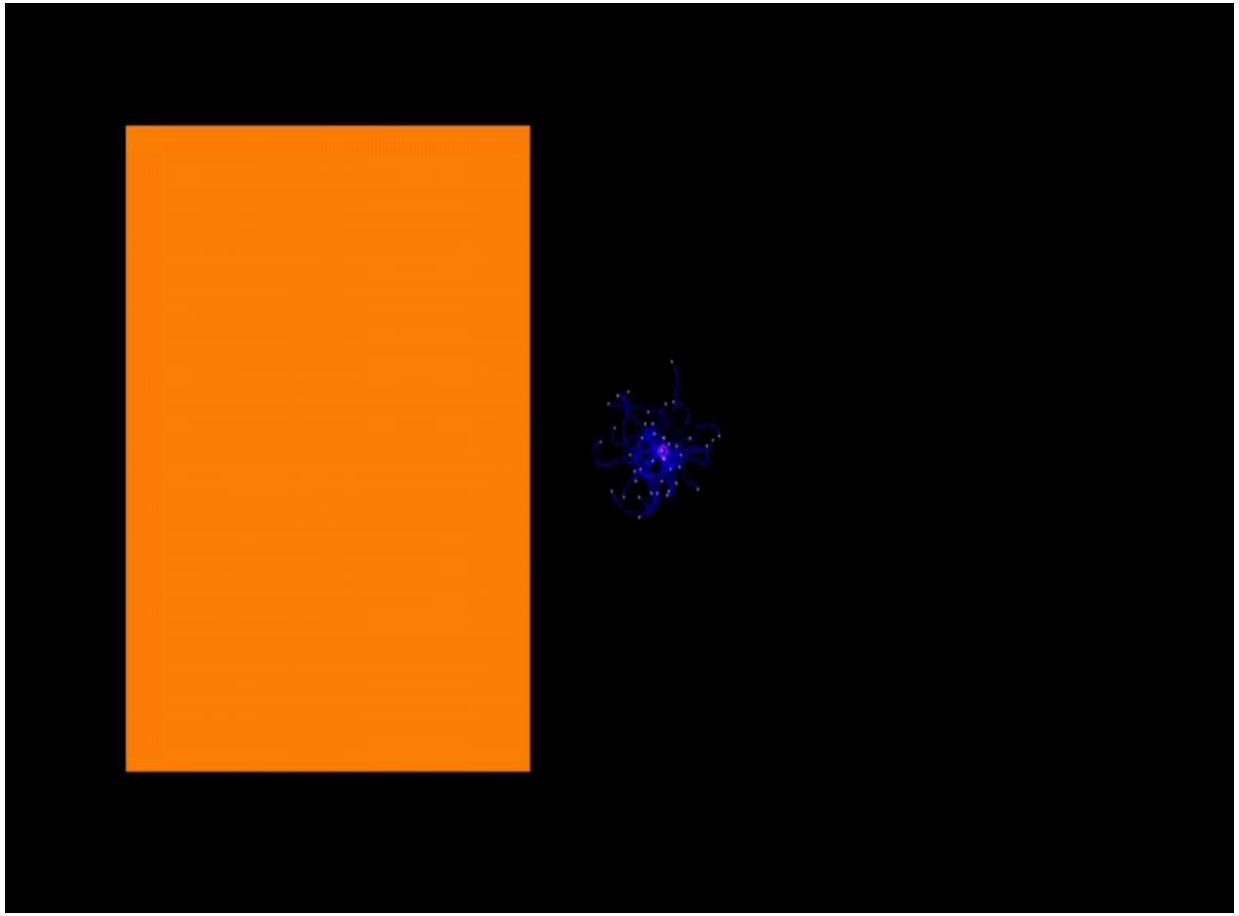
- 蚁群算法（Ant Colony Optimization, ACO）；
- 遗传算法（Genetic Algorithm, GA）；
- 模拟退火算法（Simulated Annealing, SA）；
- 禁忌搜索算法（Tabu Search, TS）；
- ……

04 蚁群算法

➢ 蚁群算法简介

- 蚁群算法(ant colony optimization, ACO)，又称蚂蚁算法，是一种用来在图中寻找优化路径的机率型算法。其灵感来源于蚂蚁在寻找食物过程中发现路径的行为。

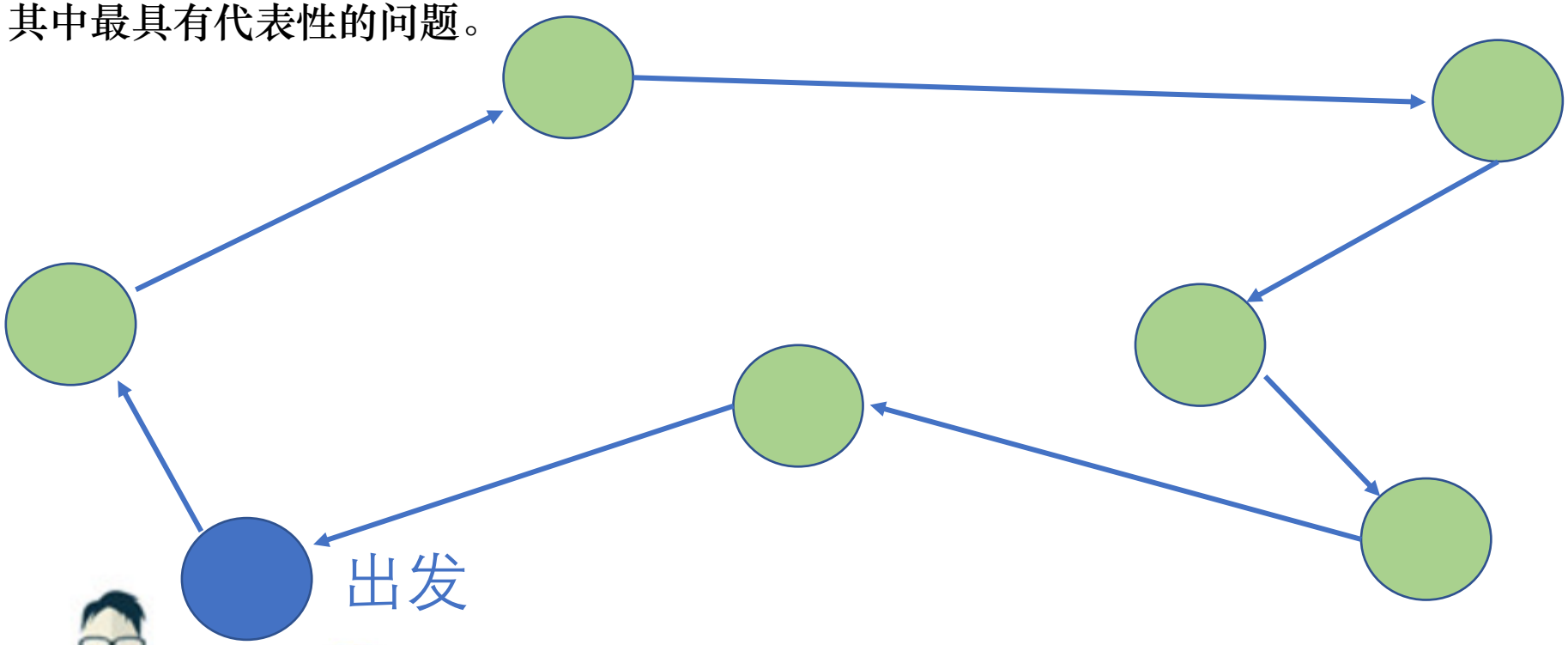
- 蚂蚁在行走过程中会释放一种称为“信息素”的物质，用来标识自己的行走路径。在寻找食物的过程中，根据信息素的浓度选择行走的方向，并最终到达食物所在的地方。



04 蚁群算法

➢ 蚂蚁系统（AS算法）——最早的ACO算法

意大利学者M.Dorigo于1991年在其博士论文中首次系统地提出一种基于蚂蚁种群的新型智能优化算法“蚂蚁系统（Ant system,简称AS）该算法能够帮助人类更好的解决路径优化问题，求解旅行商问题是其中最具有代表性的问题。



最短距离

➢ 算法流程

1、初始化参数包括：信息素、启发式因子等。

2、将各只蚂蚁放置各顶点，禁忌表为对应的顶点。

3、取1只蚂蚁，计算转移概率$P_{ij}^{k}(t)$，按轮盘赌的方式选择下一个顶点，更新禁忌表，再计算概率，再选择顶点，再更新禁忌表，直至遍历所有顶点1次。

4、计算该只蚂蚁留在各边的信息素量 $\Delta\tau_{ij}^{k}$，该蚂蚁死去。

5、重复3~4，直至 m只蚂蚁都周游完毕。

6、计算各边的信息素增量$\Delta\tau_{ij}$和信息素量$\Delta\tau_{ij}(t+n)$。

7、记录本次迭代的路径，更新当前的最优路径，清空禁忌表。

8、判断是否达到预定的迭代步数，或者是否出现停滞现象。若是，算法结束，输出当前最优路径；否，转2，进行下一次迭代。

➢ 蚂蚁系统（AS算法）—构建路径

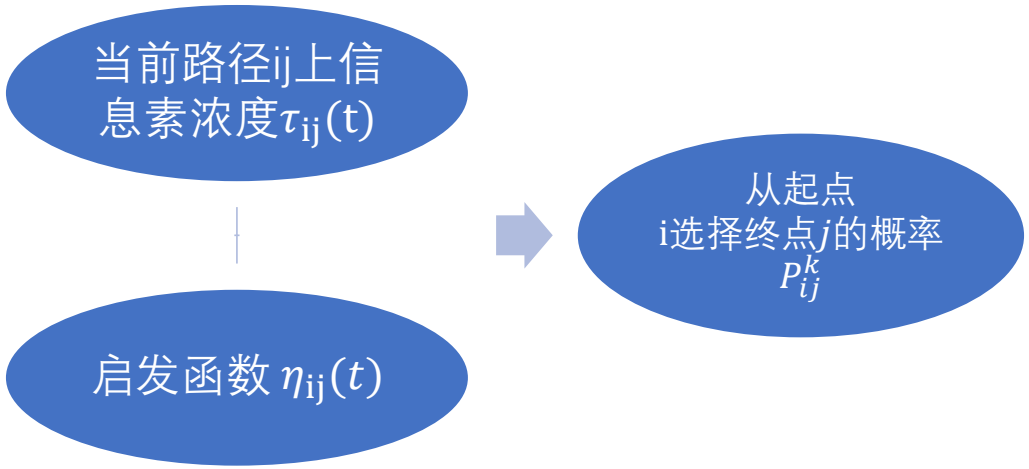
$$P_{ij}^{k}=\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)*\eta_{ij}^{\beta}(t)}{\sum_{s\in allowed_k}\tau_{ij}^{\alpha}(t)*\eta_{ij}^{\beta}(t)} & j\in allowed_k\\ 0 & \text{其他}\end{cases}$$

i, j分别为起点和终点。

$allowed_k$为尚未访问过的节点的集合。

$\eta_{ij}(t)=1/d_{ij}$是两点i,j路径距离的倒数。

$\tau_{ij}(t)$为时间t时由i到j的信息素浓度。



04 蚁群算法

➢ 蚂蚁系统（AS算法）— 模拟计算过程

- 蚂蚁随机生成在各个城市
- 根据概率公式确定下一目标城市
- 完成一次迭代，更新信息素浓度
- 达到设置迭代次数，完成计算，推荐最优路径

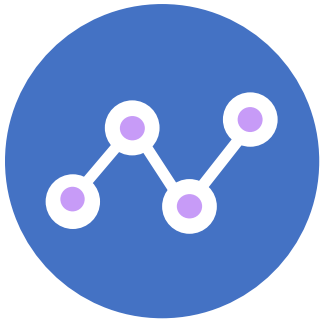
04 蚁群算法

➢ 蚁群算法（ACO）特点

自组织的算法

自组织：组织力或组织指令是来自于系统的内部。

在抽象意义上讲，自组织就是在没有外界作用下使得系统熵减小的过程(即是系统**从无序到有序**的变化过程)。



并行的算法

每只蚂蚁搜索的过程彼此**独立**，仅通过信息激素进行通信。

在问题空间的多点同时开始进行独立的解搜索，不仅增加了算法的可靠性，也使得算法具有较强的全局搜索能力。

正反馈的算法

蚂蚁能够最终找到最短路径，直接依赖于最短路径上**信息激素的堆积**，而信息激素的堆积却是一个正反馈的过程。

正反馈是蚂蚁算法的重要特征，它使得算法演化过程得以进行。

➢ AS算法的优点与不足

优点

- 较强的鲁棒性——稍加修改即可应用于其他问题。（鲁棒性就是系统的健壮性，用以表征控制系统对特性或参数摄动的不敏感性。）
- 分布式计算——本质上具有并行性。
- 易于与其他启发式算法结合。

不足

- 一般需要较长的搜索时间。
- 容易出现停滞现象。

04 蚁群算法

➢ 改进的蚁群优化算法

- ▲最优解保留策略蚂蚁系统（带精英策略的蚂蚁系统ASelite)
- ▲蚁群系统（ACS)
- ▲最大-最小蚂蚁系统（MMAS)
- ▲基于优化排序的蚂蚁系统（ASrank)
- ▲最优最差蚂蚁系统（BWAS)
- ▲一种新的自适应蚁群算法（AACA)
- ▲基于混合行为的蚁群算法（HBACA)

➢ 带精英策略的蚂蚁系统（ASelite）

特点——在信息素更新时给予当前最优解以额外的信息素量，使最优解得到更好的利用。找到全局最优解的蚂蚁称为“精英蚂蚁”。

$$\tau_{ij}(t+n)=(1-\rho)\cdot\tau_{ij}(t)+\Delta\tau_{ij}+\Delta\tau_{ij}^{*}$$

$$\Delta\tau_{\mathrm{ij}}^{*}=\begin{cases}\sigma\dfrac{Q}{L^{gb}} & \text{若边ij是当前最优解的一部分}\\ 0 & \text{否则}\end{cases}$$

$\Delta\tau_{ij}^{*}$ ——精英蚂蚁在边 ij 上增加的信息素量；

σ —— 精英蚂蚁个数；

L^{gb} ——当前全局最优解路径长度。

➢ 蚁群系统（ACS）

特点

1、状态转移规则——伪随机比率规则

设$q_0 \in (0,1)$为常数，$q_0 \in (0,1)$为随机数，如果 $q \le q_0$，则蚂蚁转移的下一座城市是使$[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}$取最大值的城市；若$q > q_0$，仍按转移概率确定。

2、全局更新规则——只有精英蚂蚁才允许释放信息素，即只有全局最优解所属的边才增加信息素。

3、局部更新规则——蚂蚁每次从城市i转移到城市j后，边(i,j)上的信息素适当减少。

规则 1 和 2 都是为了使搜索过程更具有指导性，即使蚂蚁的搜索主要集中在当前找出的最好解邻域内。规则 3 则是为了使已选的边对后来的蚂蚁具有较小的影响力，以避免蚂蚁收敛到同一路径。

➢ 最大最小蚂蚁系统（MMAS）

特点

1、每次迭代后，只对最优解所属路径上的信息素更新。

2、对每条边的信息素量限制在范围 $[\tau_{min},\tau_{max}]$ 内，目的是防止某一条路径上的信息素量远大于其余路径，避免过早收敛于局部最优解。

关于 τ_{min},τ_{max} 的取值，没有确定的方法，有的书例子中取为0.01，10；有的书提出一个在最大值给定的情况下计算最小值的公式。

➢ 基于优化排序的蚂蚁系统（ASrank）

特点：每次迭代完成后，蚂蚁所经路径由小到大排序，并根据路径长度赋予不同的权重，路径越短权重越大。信息素更新时对 $\Delta\tau_{ij}^{k}$ 考虑权重的影响。

➢ 最优最差蚂蚁系统（BWAS）

特点：主要是修改了ACS中的全局更新公式，增加对最差蚂蚁路径信息素的更新，对最差解进行削弱，使信息素差异进一步增大。

➢ 一种新的自适应蚁群算法（AACA）

特点：将ACS中的状态转移规则改为自适应伪随机比率规则，动态调整转移概率，以避免出现停滞现象。

说明：在ACS的状态转移公式中，q_0 是给定的常数；在AACA中，　是q_0随平均节点分支数ANB而变化的变量。ANB较大，意味着下一步可选的城市较多，q_0 也变大，表示选择信息素和距离最好的边的可能性增大；反之减小。

➢ 基于混合行为的蚁群算法（HBACA）

特点：按蚂蚁的行为特征将蚂蚁分成4类，称为4个子蚁群，各子蚁群按各自的转移规则行动，搜索路径，每迭代一次，更新当前最优解，按最优路径长度更新各条边上的信息素，如此直至算法结束。

蚂蚁行为——蚂蚁在前进过程中，用以决定其下一步移动到哪个状态的规则集合。

蚂蚁行为

1、蚂蚁以随机方式选择下一步要到达的状态。

2、蚂蚁以贪婪方式选择下一步要到达的状态。

3、蚂蚁按信息素强度选择下一步要到达的状态。

4、蚂蚁按信息素强度和城市间距离选择下一步要到达的状态。

主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

05 强化学习概述

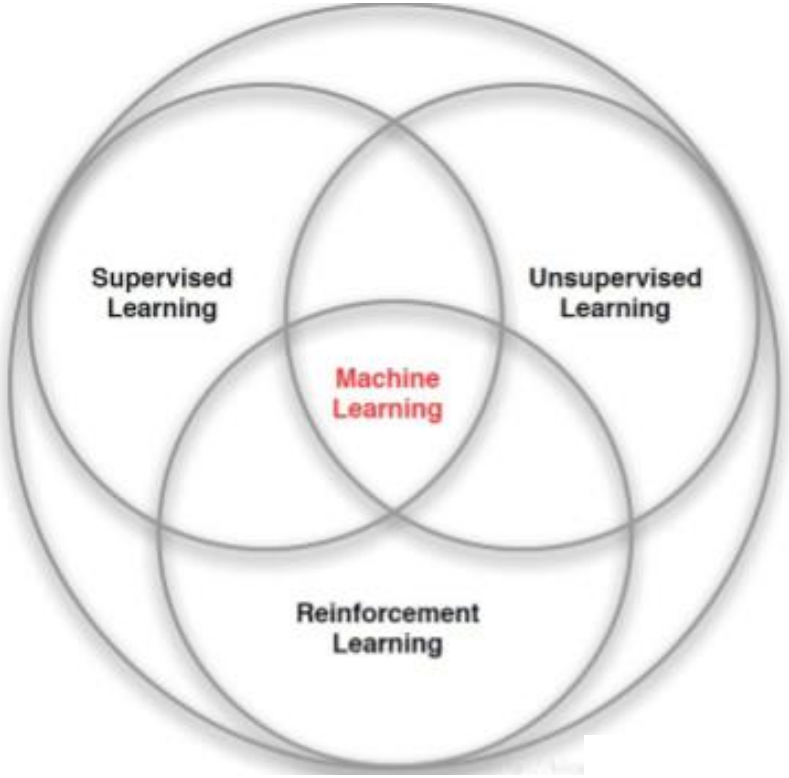
基本概念：强化学习又称为增强学习、加强学习、再励学习或激励学习，是一种从环境状态到行为映射的学习，目的是使动作从环境中获得的累积回报值最大。强化学习是机器学习分支之一，介于监督学习和无监督学习之间。

机器学习三大分支：

无监督学习

监督学习

强化学习

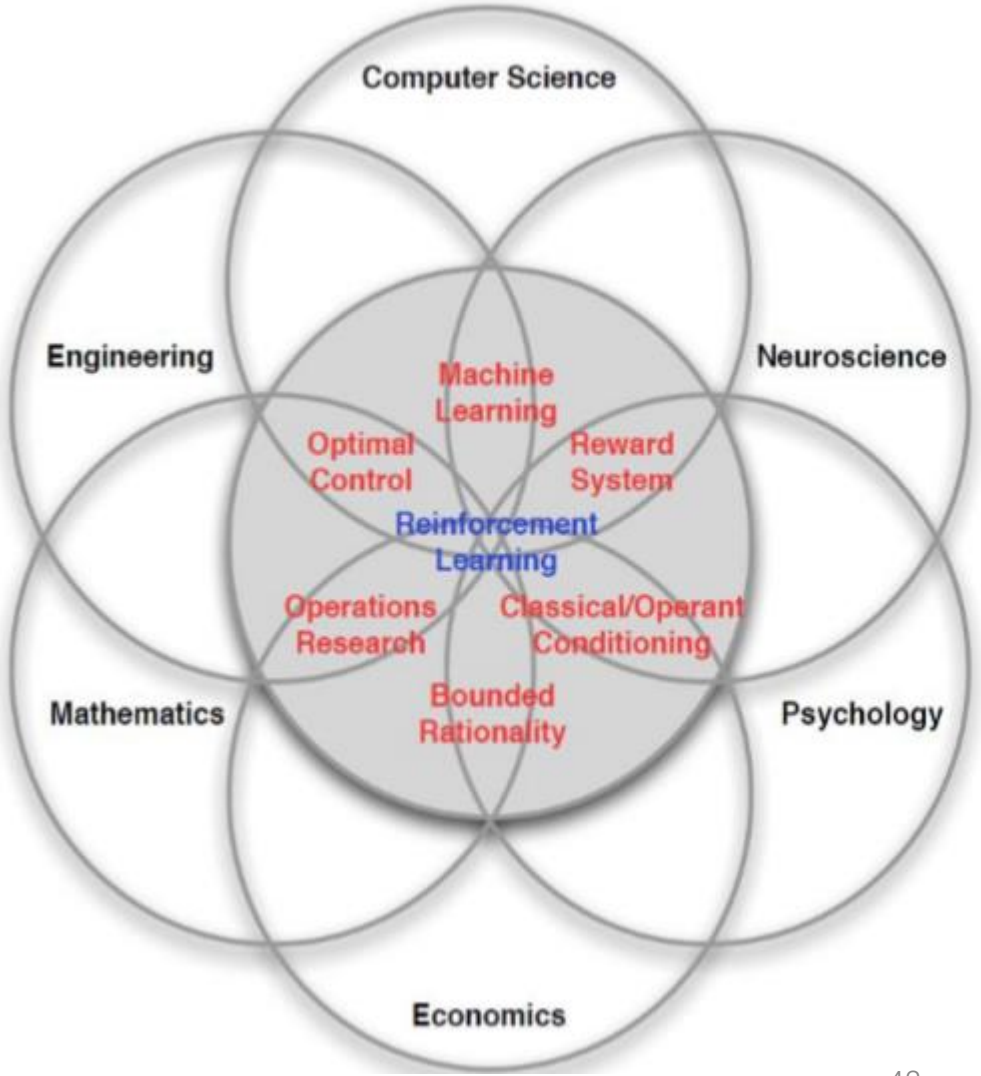


05 强化学习概述

- 强化学习技术是从控制理论 、统计学、心理学等相关学科发展而来。

- 在人工智能、机器学习和自动控制等领域中得到广泛研究和应用，并被认为是设计智能系统的核心技术之一。

- 随着强化学习的数学基础研究取得突破性进展后，强化学习成为机器学习领域研究热点之一。



05 强化学习概述

➢ 强化学习发展历史

- 1954年Minsky首次提出“强化”和“强化学习”的概念；
- 1953到1957年，Bellman提出了求解最优控制问题的一个有效方法---动态规划，同年，还提出了最优控制问题的随机离散版本，就是著名的马尔可夫决策过程 ；
- 1960年Howard提出马尔可夫决策过程的策略迭代方法，这些都成为现代强化学习的理论基础；
- 1972年，Klopf把试错学习和时序差分结合在一起；
- 1988年 Sutton提出了TD算法；
- 1989年 Watkins提出了Q学习算法；
- 1994年 Rummery等提出了SARSA学习算法；
- 2015 Google DeepMind公司提出了深度强化学习DRL。

➢ 强化学习的特点

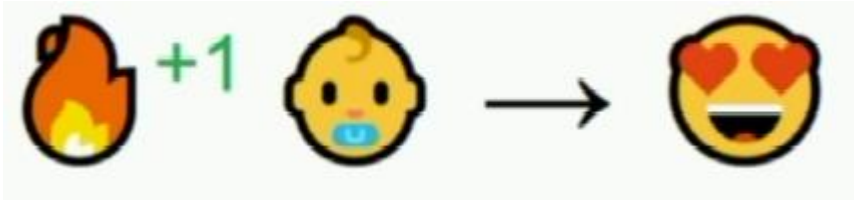
强化学习围绕着如何与环境交互学习，在行动—评价的环境中获得改进的行动方案，以适应环境达到预想的目的。学习者并不会被告知采取哪个动作，而只能通过尝试每个动作，获得环境对所采取动作的反馈信息，从而指导以后的行动。因此，强化学习主要特点包括：

- 试错搜索： Agent通过尝试多个动作，搜索最优策略；
- 延迟回报：其反馈信号是延迟的而非瞬间的；
- 适应性：Agent不断利用环境中的反馈信息来改善其性能；
- 不依赖外部教师信号：因为Agent只根据反馈信号进行学习， 因此不需要外部教师信号。

05 强化学习概述

➢ **核心思想：智能体 agent 在环境 environment 中学习，根据环境的状态 state，执行动作 action，并根据环境的反馈 reward（奖励）来指导更好的动作。**

- **一个孩子，第一次看到了火**
  - 来到了火边，感受到温暖，正反馈（+1）



  - 用手去触摸火，被烫到，负反馈（-1）

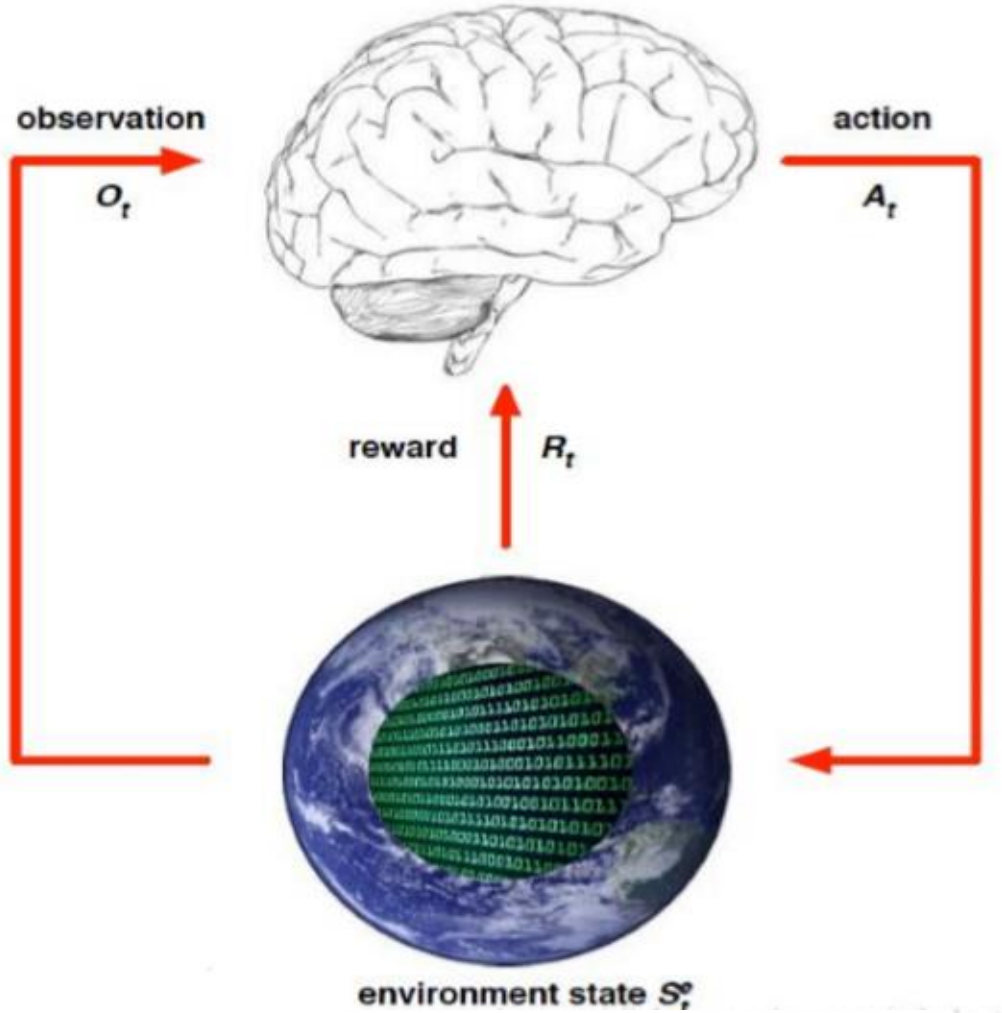
  - 结论：在火旁边是好的（温暖+1），但不能靠太近（被烫-1）

05 强化学习概述

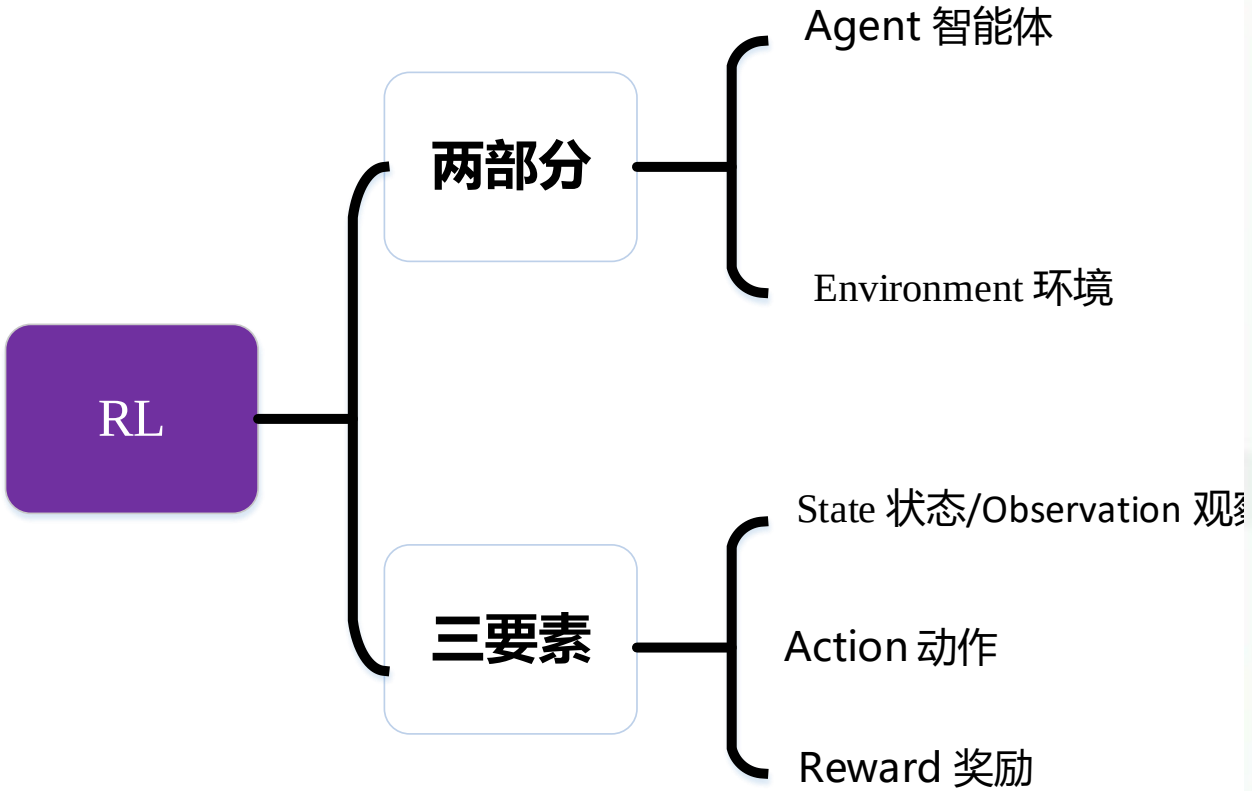
➢ 强化学习基本模型

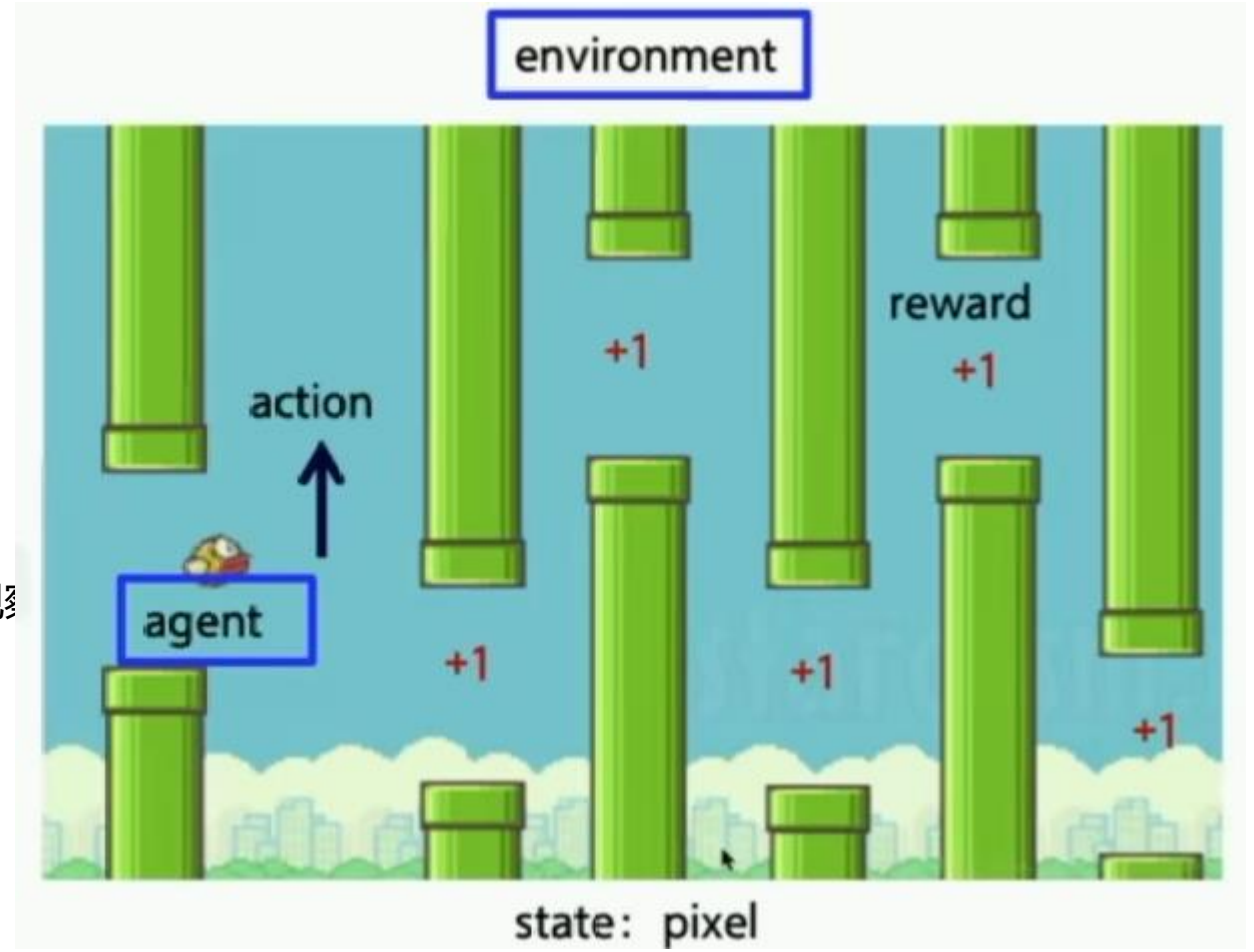
- 在强化学习中，Agent选择一个动作a作用于环境；
- 环境接收该动作后发生变化，同时产生一个强化信号 Reward（奖或罚）反馈给Agent；
- Agent再根据强化信号和环境的当前状态s再选择下一个动作，选择的原则是使受到奖赏值的概率增大。

强化学习的目的就是寻找一个最优策略，使得Agent在运行中所获得的累计期望回报最大。



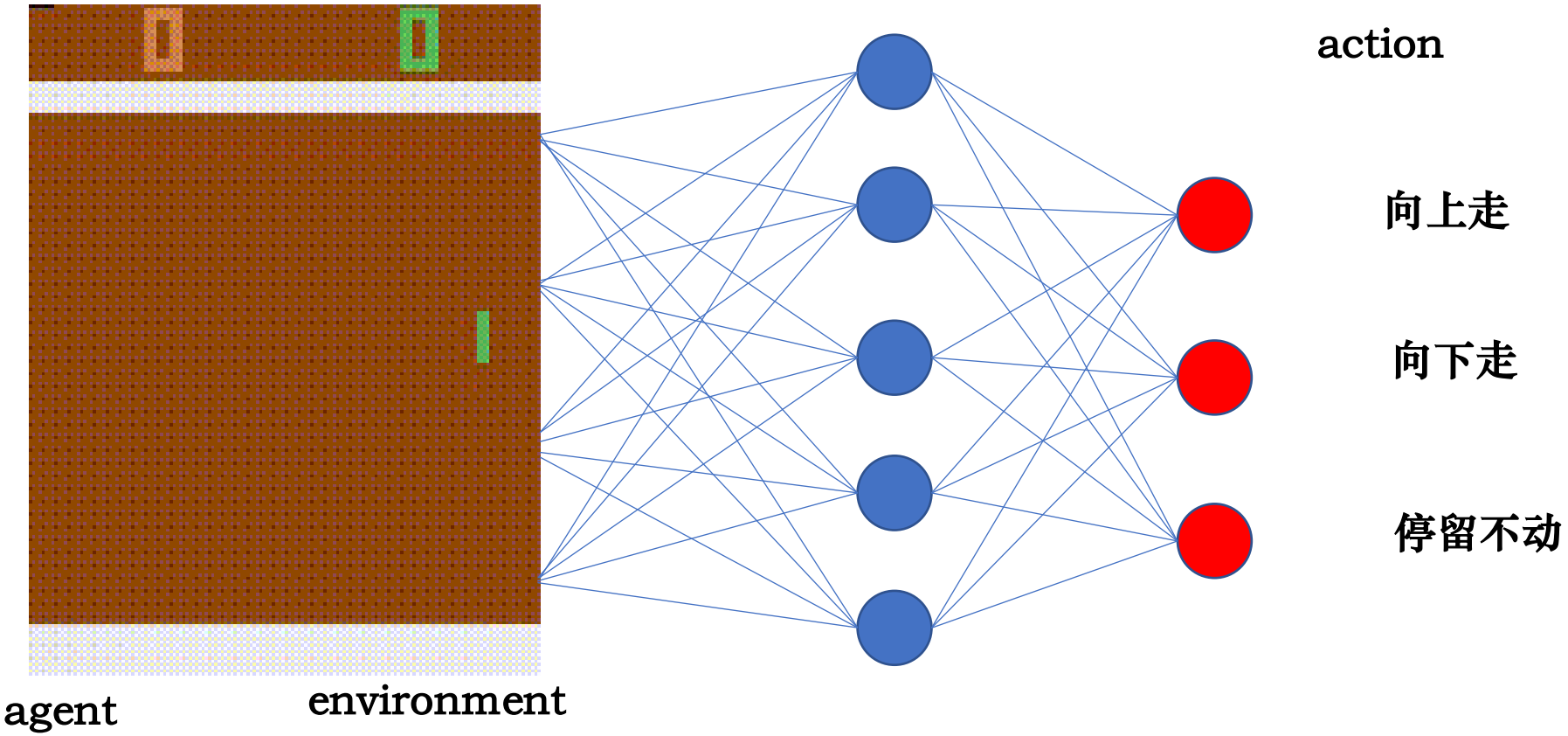
➢ 强化学习基本模型


RL
两部分
Agent 智能体
Environment 环境
三要素
State 状态/Observation 观
Action 动作
Reward 奖励


environment
action
agent
reward
+1
+1
+1
+1
+1
+1
state: pixel

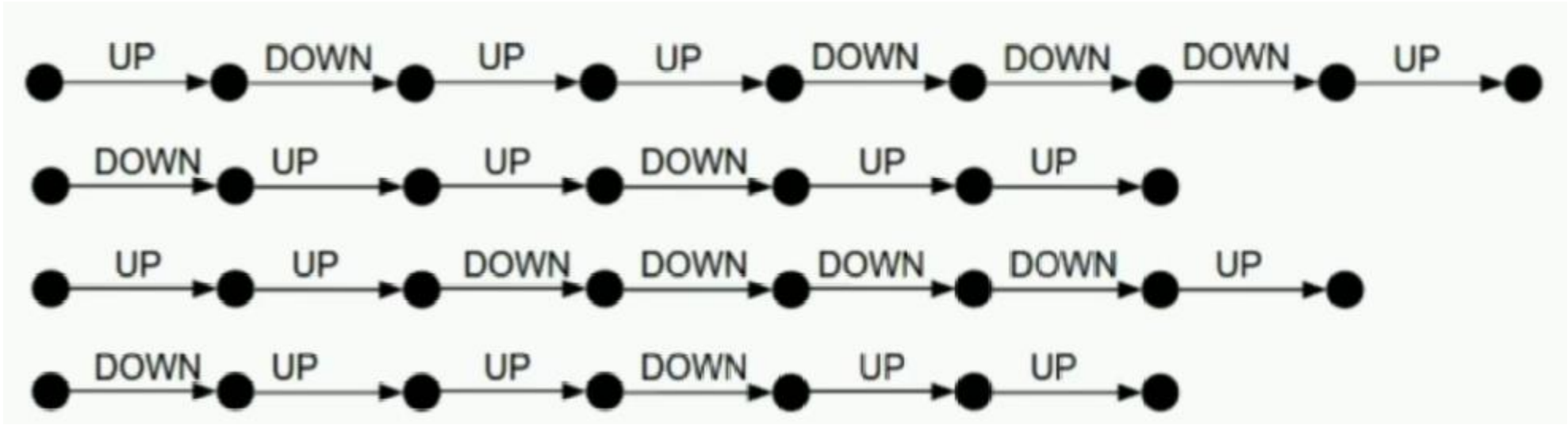
- 游戏：乒乓球

state:像素级别图像



强化学习概述

- 游戏：乒乓球

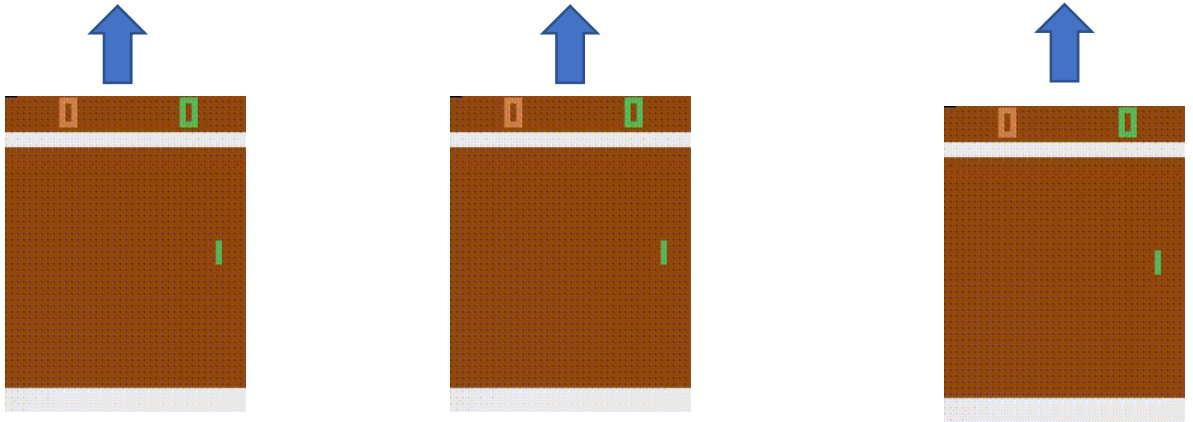


WIN

LOSE

LOSE

WIN



05 强化学习概述

- 游戏：走迷宫

state:当前位置

出发点

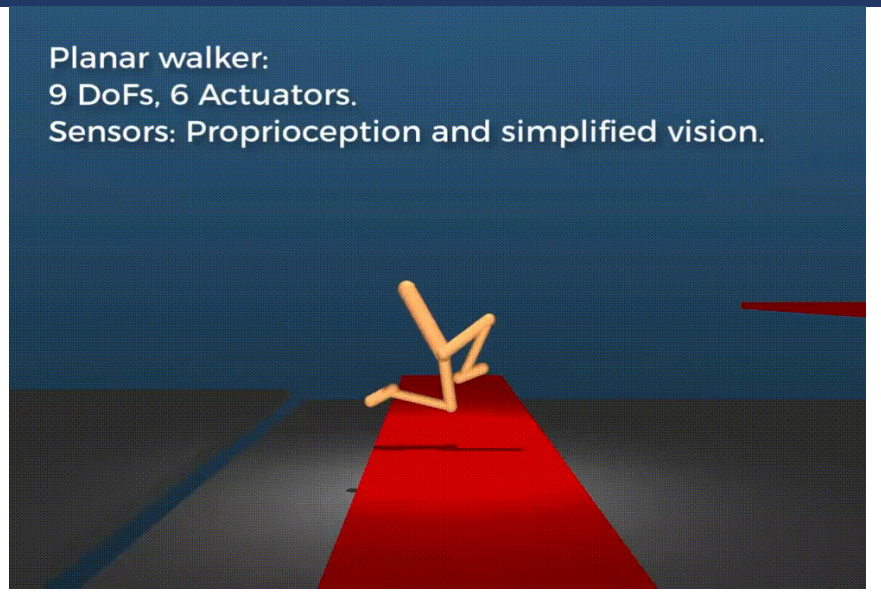
environment:掉进洞里游戏结束

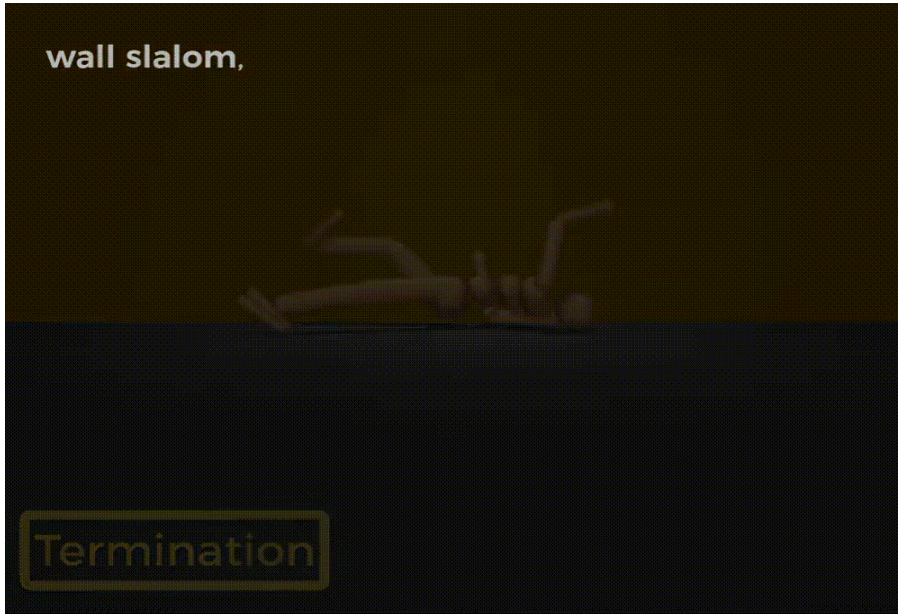
reward

目的地

05 强化学习概述

- 运动与平衡





environment: 环境和自己

agent: 仿真机器人

state: 骨骼、关节、肌肉的状态等

action: 肌肉收缩信号

reward: 髋关节高于某一高度（站立），速度（奔跑）

➢ 个性化与推荐

environment:
- 可用新闻列表
- 以及手机前面的你

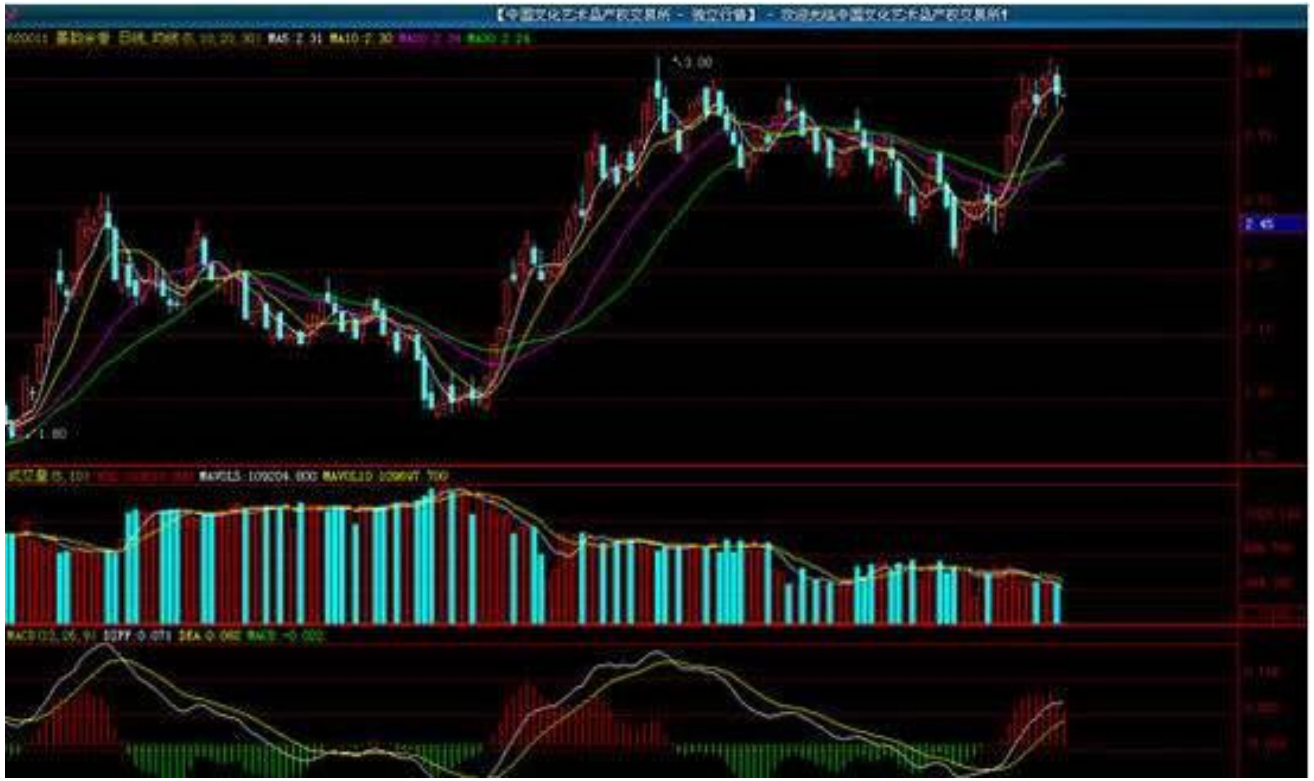
agent: 百度app

用户点击: +reward
用户跳过: -reward
用户离开: -reward

state: 当前推荐列表用户理解

action

- 股票：思考题，什么是 action, state，reward



environment: 股票市场

agent: 控制器

state: 股票历史曲线

action：买入金额、卖出金额

reward: 股票累积收益

- **交通治理**

environment: 交通状况

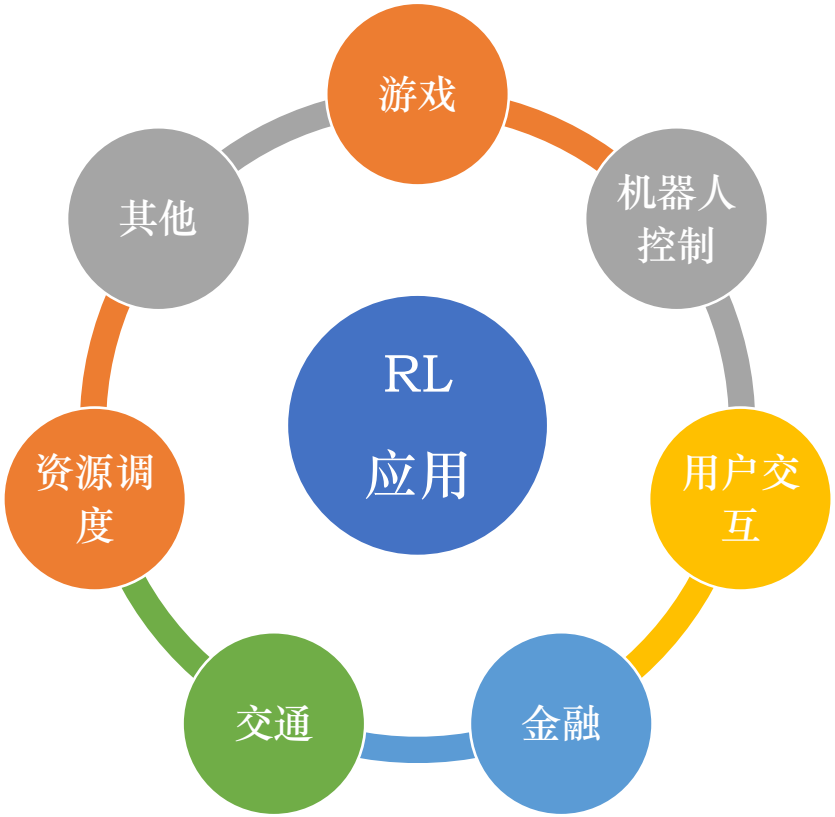
agent: 交通控制器

state: 各个路口摄像头输入图像

action: 红绿灯的亮灭

reward: 拥堵情况（路口滞留的车辆数）

➢ 强化学习应用



➢ 强化学习的分类

从广义上讲，强化学习是解决序贯决策问题的方法之一，将强化学习纳入马尔科夫决策过程的框架后，可以分为基于模型的动态规划方法和基于无模型的强化学习方法。

05 强化学习

➢ 数学基础

- 高等数学；
- 线性代数（向量空间的变换思维）；
- 概率与数理统计（期望、方差）；

➢ 编程基础

- Python: 基本语法、numpy、pandas；

➢ 机器学习基础

- 神经网络（FC、CNN）；

05 强化学习

➢ 理论

- 书籍

  经典书：《Reinforcement Learning: An Introduction （强化学习导论）》（强化学习教父 Richard Sutton的教材）

- 视频

  理论课：2015 David Silver经典强化学习公开课、UC Berkeley CS285、斯坦福CS234

➢ 经典论文（进阶）

- DQN. “Playing Atari with deep reinforcement learning.” https://arxiv.org/pdf/1312.5602.pdf
- A3C. “Asynchronous methods for deep reinforcement learning.” http://www.jmlr.org/proceddings/papers/v48/mniha16.pdf
- PPO. “Proximal policy optimization algorithms.” https://arxiv.org/pdf/1207.0634

➢ 前沿研究方向

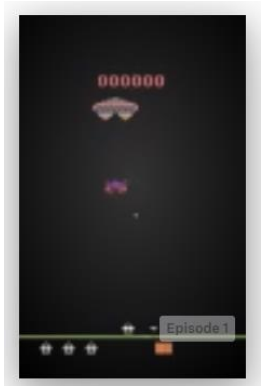
- Model-base RL、Hierarchical RL、Multi Agent RL、Meta Learning

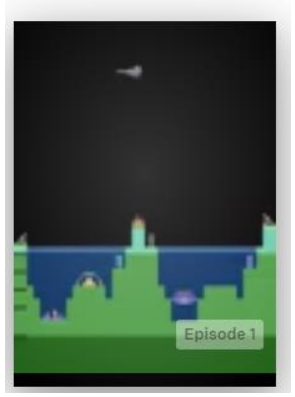
➢ 编程实践：GYM

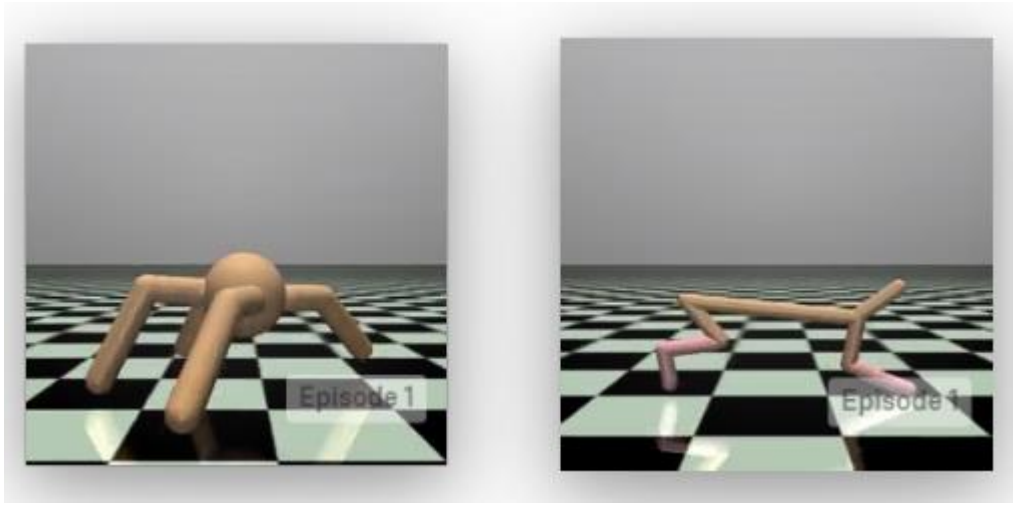
- GYM
  仿真平台、python开源库、RL测试平台
- 官网
  https://gym.openai.com/

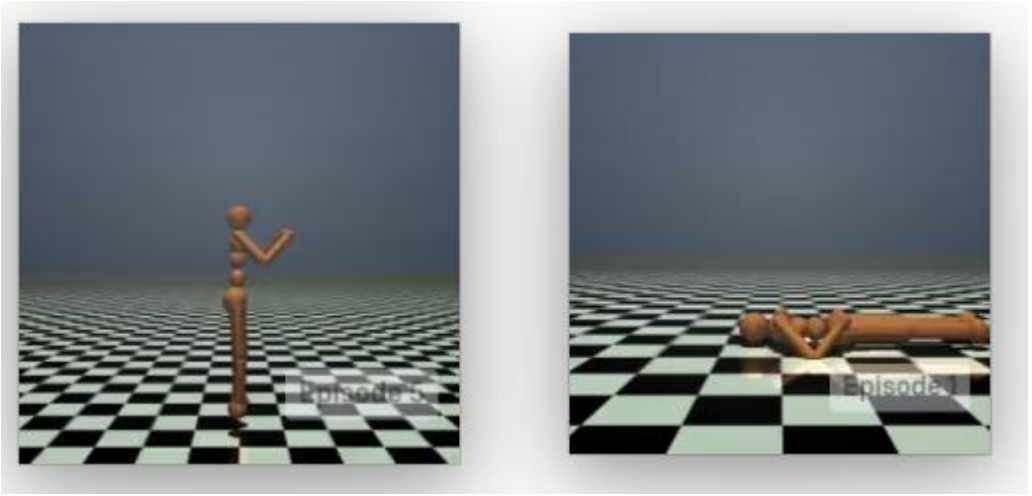
➢ 控制场景

- 离散控制场景：一般使用atatri环境评估
- 连续控制场景：一般使用mujoco环境游戏评估









05 强化学习模型

- 5.1 有模型学习
- 5.2 无模型学习
- 5.3 案例介绍

5.1 有模型学习

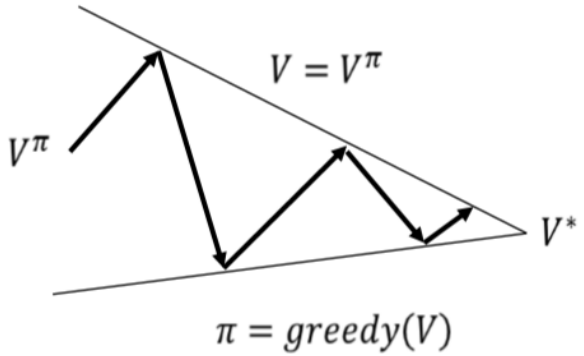
定义：在已知模型的环境中学习，称为“有模型学习”，也即，对于多步强化学习任务，其对应的马尔可夫决策过程四元组 表示<S,A,R,P>均为已知，称为“模型已知”。

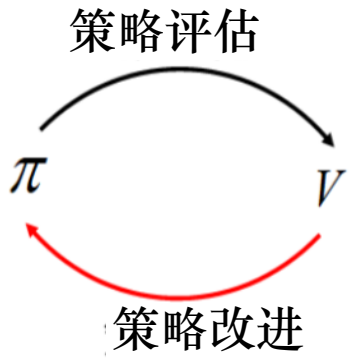
- S：环境的状态空间;
- A：agent可选择的动作空间 ；
- R(s,a)：奖励函数，返回的值表示在状态下执行a动作的奖励 ；
- P(s’|s,a)：状态转移概率函数，表示从s状态执行a动作后环 境转移至s′状态的概率。

目标：找到一个策略π能够最大化我们的对未来的奖励的期望$E(\sum_{t=0}^{n}\gamma^{t}R_{t})$，$R_t$为$t$时刻的奖励$\gamma$为折扣因子，代表距离现在遥远的奖励不如现在的奖励大。

➢ 策略迭代算法—流程

- 某一个随机策略作为初始策略；
- 策略评价+策略改进+策略评价+策略改进+……；
- 若满足收敛条件，则退出，否则，转入②；

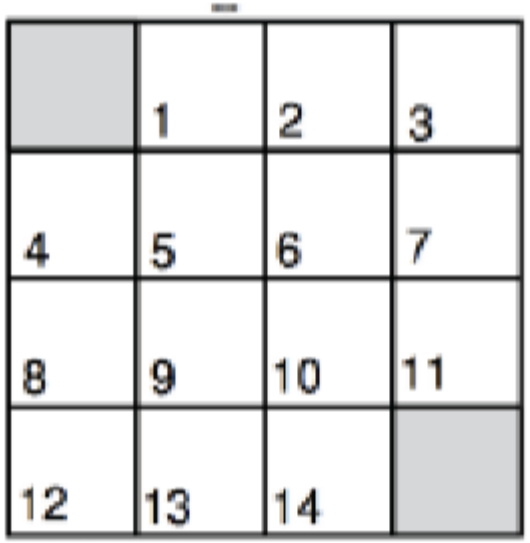




策略迭代算法的缺点在于：每次改进策略后都需要重新进行策略评价，计算比较耗时。

➢ 策略迭代算法—策略评价举例

- 即时奖励：左图是一个九宫格，左上角和右下角是终点，它们的reward是0，其他的状态reward都是-1；
- 状态空间：除了灰色两个格子，其他都是非终点状态；
- 动作空间：在每个状态下，都有四种动作可以执行，分别是上下左右(东西南北)；
- 转移概率：任何想要离开大正方形的动作将保持其状态不变，也就是原地不动。其他时候都是直接移动到 下一个状态。所以状态转移概率是确定性的
- 折扣因子：γ=1；
- 当前策略：在任何状态下，agent都采取均匀随机策略 ，也就是它的动作是随机选择的，即：π(e|*)=π(w|*)=π(s|*)=π(n|*)=0.25；



问题：评价均匀随机策略。也就是说，求解均匀随机策略下所有状态的V值

5.2 无模型学习

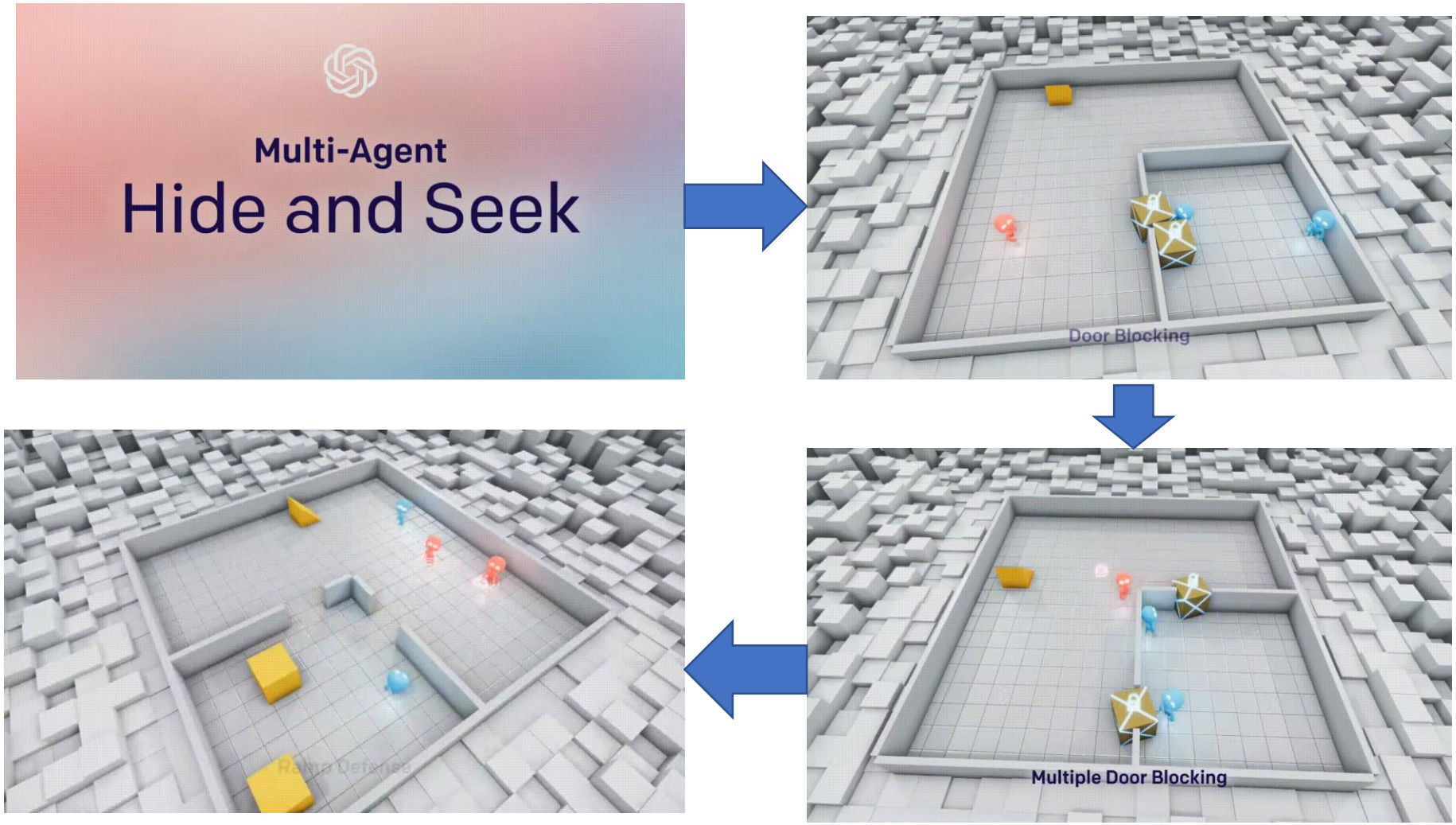
- 当模型未知，即状态转移概率、奖赏函数往往我们是不知道的，甚至很难知道环境中一共有多少状态。此时我们无法直接利用Bellman方程来求解得到最优策略；

- 若学习算法不依赖环境建模，则称为“无模型学习”，或称模型无关的学习（Model-free Learning）；

- 模型无关的强化学习，是在不知道马尔科夫决策过程的情 况下学习到最优策略。模型无关的策略学习主要有两种算法：蒙塔卡洛强化学习，时序差分强化学习。而时序差分强化学习又包括SARSA 和 Q-learning两种算法。

5.2 无模型学习

- 游戏


Multi-Agent
Hide and Seek
Door Blocking
Multiple Door Blocking

5.2 无模型学习

➢ 蒙特卡洛采用

- MDP是通过5元组：<S,P,A,R,γ>来做决策的。对于这种已知模型的情况，也就是知道了这个5元组，我们可以通过求解贝尔曼方程获得奖赏最大化；

- 但是，在现实世界中，我们无法同时知道这个5元组。比如状态转移概率就很难知道，我们无法使用bellman方程来求解V和Q值；

- 一个想法是，虽然我不知道状态转移概率P，但是这个概率是真实存在的。我们可以直接去尝试，不断采样，然后会得到奖赏，通过奖赏来评价值函数。

5.2 无模型学习

- （同策略）蒙特卡洛强化学习

周志华 著. 机器学习，北 京：清华大学出版社，2016 年1月，pp:384

输入： 环境 E;
动作空间 A;
起始状态 x_0;
策略执行步数 T.

过程：

$Q(x,a)=0,\ \mathrm{count}(x,a)=0,\ \pi(x,a)=\frac{1}{|A(x)|}$;
for $s=1,2,\ldots$ **do**
  在 E 中执行策略 π 产生轨迹
  $< x_0,a_0,r_1,x_1,a_1,r_2,\ldots,x_{T-1},a_{T-1},r_T,x_T >$;
  for $t=0,1,\ldots,T-1$ **do**
    $R=\frac{1}{T-t}\sum_{i=t+1}^{T} r_i$;
    $Q(x_t,a_t)=\frac{Q(x_t,a_t)\times\mathrm{count}(x_t,a_t)+R}{\mathrm{count}(x_t,a_t)+1}$;
    $\mathrm{count}(x_t,a_t)=\mathrm{count}(x_t,a_t)+1$
  end for
  对所有已见状态 x :

$$\pi(x,a)=\begin{cases}\arg\max_{a'} Q(x,a'), & \text{以概率 } 1-\epsilon;\\ \text{以均匀概率从 } A \text{ 中选取动作}, & \text{以概率 } \epsilon .\end{cases}$$

end for

输出： 策略 π

5.2 无模型学习

- （Sarsa）时序差分强化学习

周志华 著. 机器学习，北 京：清华大学出版社，2016 年1月，pp:388

输入： 环境 E;
动作空间 A;
起始状态 x_0;
奖赏折扣 γ;
更新步长 α.

过程：

1: $Q(x,a)=0$, $\pi(x,a)=\frac{1}{|A(x)|}$;
2: $x=x_0$, $a=\pi(x)$;
3: **for** $t=1,2,\ldots$ **do**
4: $r,x'=$ 在 E 中执行动作 a 产生的奖赏与转移的状态;
5: $a'=\pi^{\epsilon}(x')$;
6: $Q(x,a)=Q(x,a)+\alpha\big(r+\gamma Q(x',a')-Q(x,a)\big)$;
7: $\pi(x)=\arg\max_{a''} Q(x,a'')$;
8: $x=x', a=a'$
9: **end for**

输出： 策略 π

5.3 案例介绍

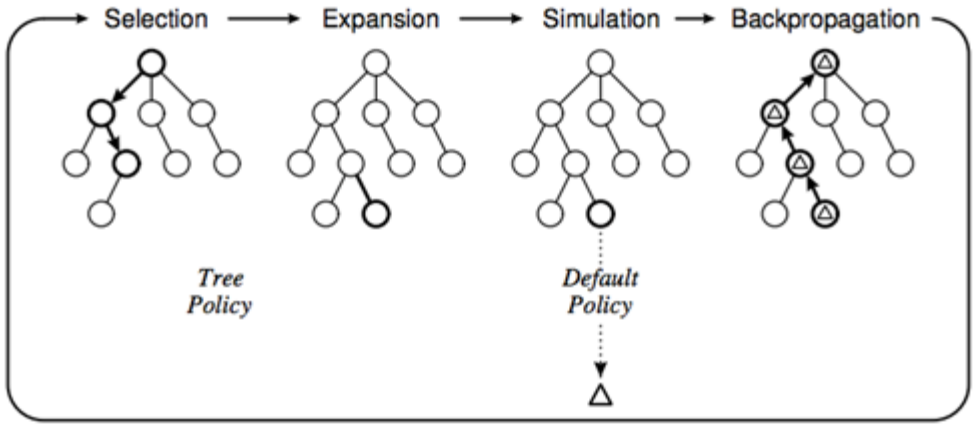
- **AlphaGo VS 柯洁**

2017年5月，AlphaGo和中国棋手柯洁在浙江乌镇，进行了一场“人机大战”，最后以3:0打败柯洁。

5.3 案例介绍

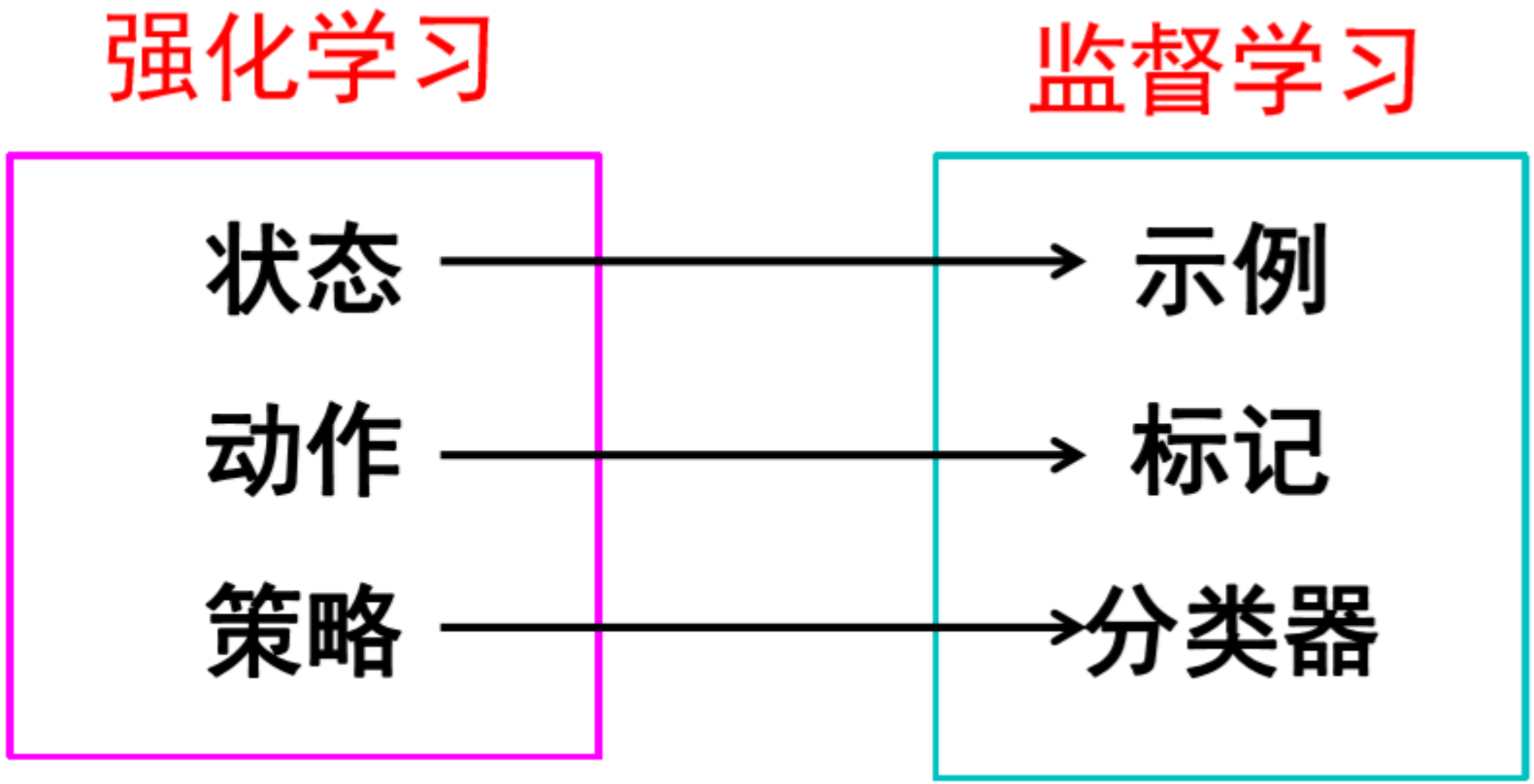
➢ AlphaGo工作原理

阿尔法围棋（AlphaGo）为了应对围棋的复杂性，结合了监督学习和强化学习的优势。它通过训练形成一个策略网络（policynetwork），将棋盘上的局势作为输入信息，并对所有可行的落子位置生成一个概率分布。然后，训练出一个价值网络（valuenetwork）对自我对弈进行预测，以-1（对手的绝对胜利）到1（AlphaGo的绝对胜利）的标准，预测所有可行落子位置的结果。



05 总结

- 强化学习在某种意义上可看作具有“延迟标记信息”的监督学习问题



05 总结

- 深度学习（DL）技术和强化学习（RL）的结合，形成了深度强化学习（DRL），迅速成为人工智能界的焦点；
- 在视频游戏、棋类游戏、机器人控制等领域取得了巨大成功。

可能面临的一些问题：

- 难以平衡“探索”和“利用”，以致算法陷入局部极小；
- 样本利用较低；
- 对环境容易出现过拟合；
- 灾难性的不稳定性。

……

05 总结

潜在的研究方向：

- 提高无模型方法的数据利用率和扩展性；
- 设计高效的探索策略。平衡“探索”与“利用”；
- 与模仿学习结合，既能更快地得到反馈、又能更快地收敛；
- 探索好的奖励机制。奖励机制对强化学习算法性能的影响是巨大的，因此该方向一直是强化学习的研究热点；
- 混合迁移学习和多任务学习。当前强化的采样效率较低，而且学到的知识不通用，迁移学习与多任务学习可以有效解决这些问题；

……

主要内容

- 1 城市物流优化
- 2 物流大数据
- 3 优化问题
- 4 启发算法
- 5 强化学习
- 6 智能物流优化

06 城市物流配送中心选址-路径研究

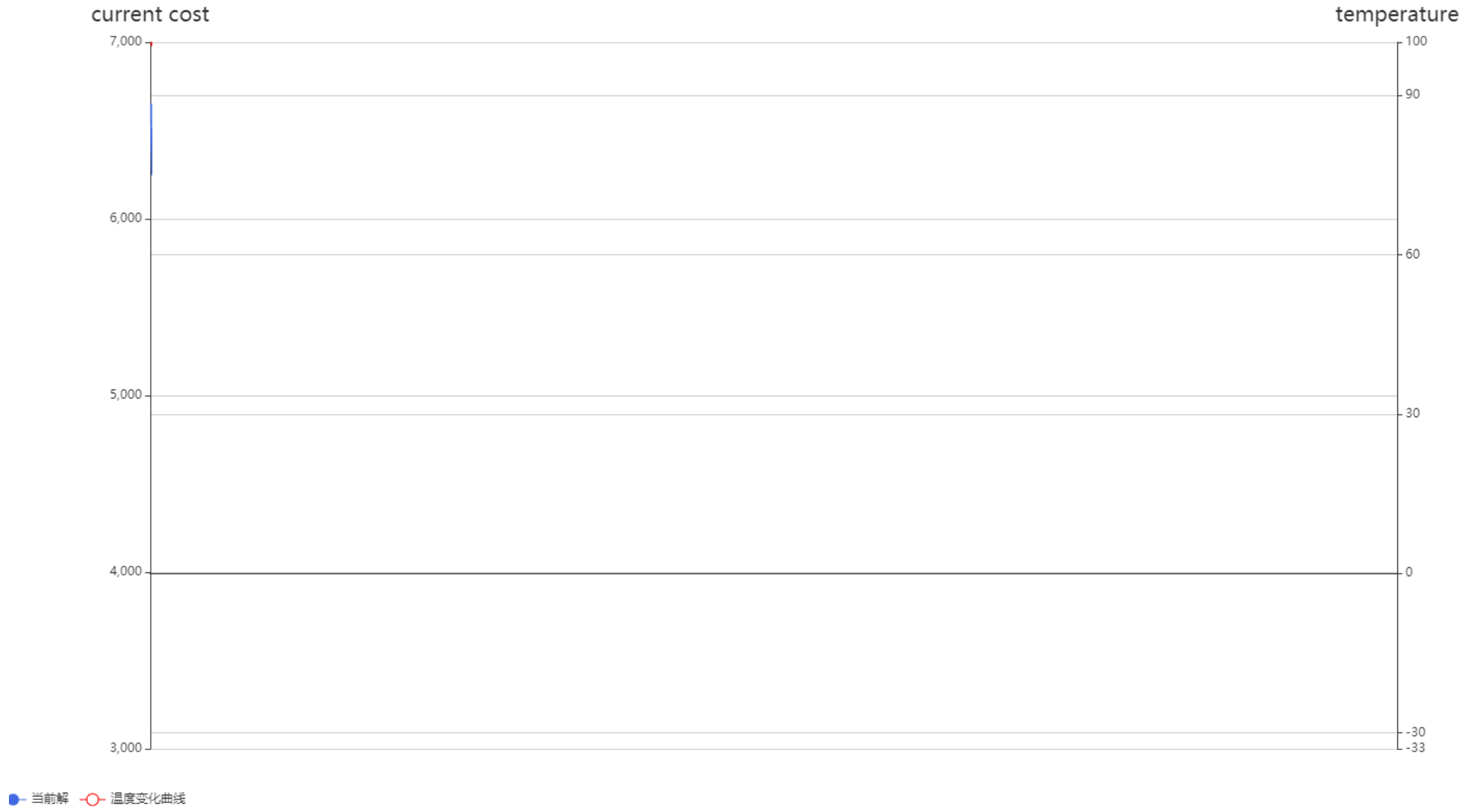
- 研究问题
  - 如何扩展“前置仓”？
  - 如何布局“前置仓”？
  - 如何安排行驶路线？

➢ 数据处理

- **浮动车GPS轨迹数据：**
  - 去除数据中存在的异常数据点和运营状态为空的点以及删除位置记录出错的点；
  - 基于OSM路网进行地图匹配。
- **订单信息：**
  - 剔除掉无效订单；
  - 提取物流中心。
- **OSM路网数据：**
  - 剔除掉步行的道路；
  - 提取道路交叉点，建立邻接矩阵。

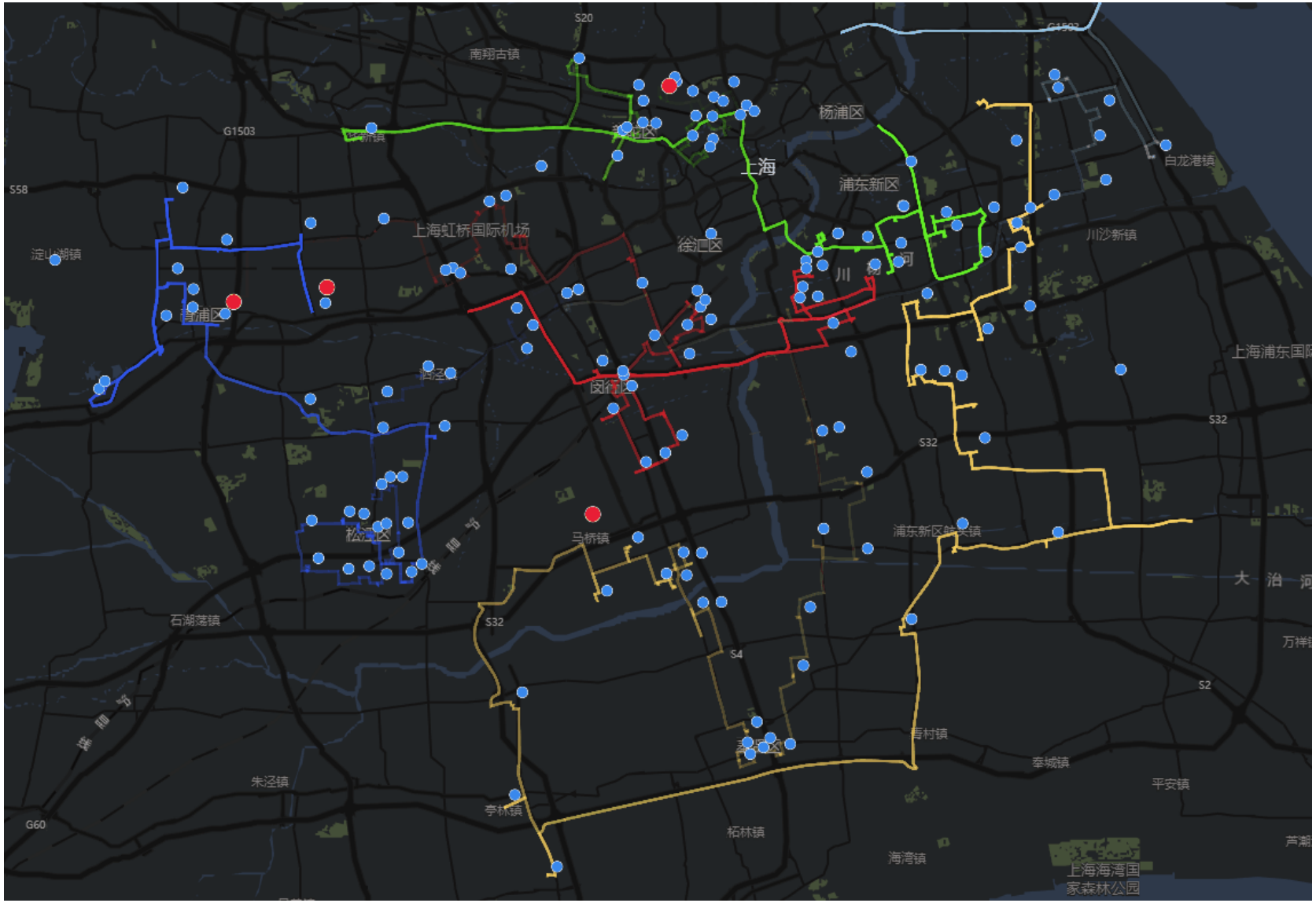
06 城市物流配送中心选址-路径研究

- 结果



06 城市物流配送中心选址-路径研究

- 结果



总结

本章介绍了物流大数据的数据来源、数据预处理过程以及常用的轨迹大数据分析方法以及其在物流优化等方面的应用。

物流大数据主要形式为浮动车数据、路网数据、物流中心数据等。

数据预处理轨迹清洗的过程中主要是对轨迹、订单中无效数据与重复数据进行清除。

强化学习、蚁群算法等方法是常用的分析方法，对路线的优化，有利于节省成本和城市发展等。

全局最优难寻，根据环境和当前变化趋势先确定局部最优。