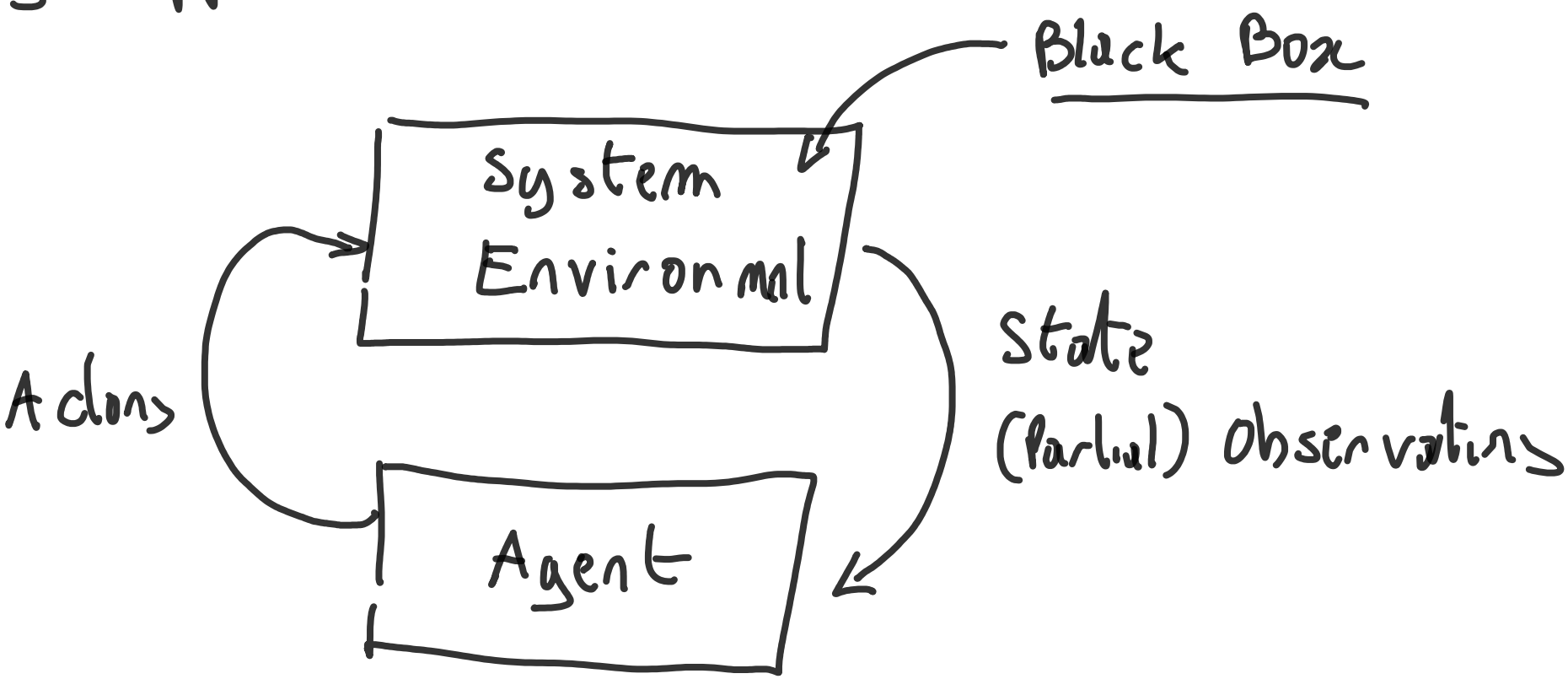
Reinforcement Learning for Selective Key Applications in Power Systems



⇒ Goal: Finding a good way to pick actions (policy) from the interactions with the system

S_t : State of the world

A_t : Actions

For sake of simplicity, we assume that | the states are finite
| the actions are finite

$$P(S_{t+1} \mid A_t, S_t, A_{t-1}, S_{t-1}, \dots, A_0, S_0)$$

$$= P(S_{t+1} \mid A_t, S_t) \leftarrow \text{Markovian Assumption}$$

R_t : Rewards (Instantaneous)

$$P(R_{t+1} \mid A_t, S_t, A_{t-1}, \dots) = P(R_{t+1} \mid A_t, S_t)$$

Policy : $\pi : s_t, a_{t-1}, s_{t-1}, \dots, a_0, s_0 \longmapsto a_t$

↳ Markovian Policy : $\pi : s_t \longrightarrow a_t$

Stochastic policy : $s_t \longrightarrow \pi(\cdot \mid s_t)$

Return (way of measuring the quality)

$$\max_{\pi} \mathbb{E}_{s_0} \mathbb{E}_{\pi} \left[\sum_{t \geq 0} \gamma^t R_t \right]$$

(Cumulative Reward)

- planning

discount factor $\gamma \in (0,1)$

$\Rightarrow$ Everything is well defined

but far away future is less important than close future

Markovian Decision Process (MDP) setting

$P(S_{t+1} | A_t, S_t)$ known

$$E_\pi\left(\sum \gamma^t R^t \mid S_0\right) = E_\pi(R_1) + \gamma E_\pi\left[E_\pi\left[\sum_{t \geq 1} \gamma^{t-1} R^t \mid S_1\right] S_0\right]$$

$$\Downarrow$$

value function → $V_\pi(S_0) = E_\pi(R_1) + \gamma E_\pi[V_\pi(S_1) \mid S_0]$

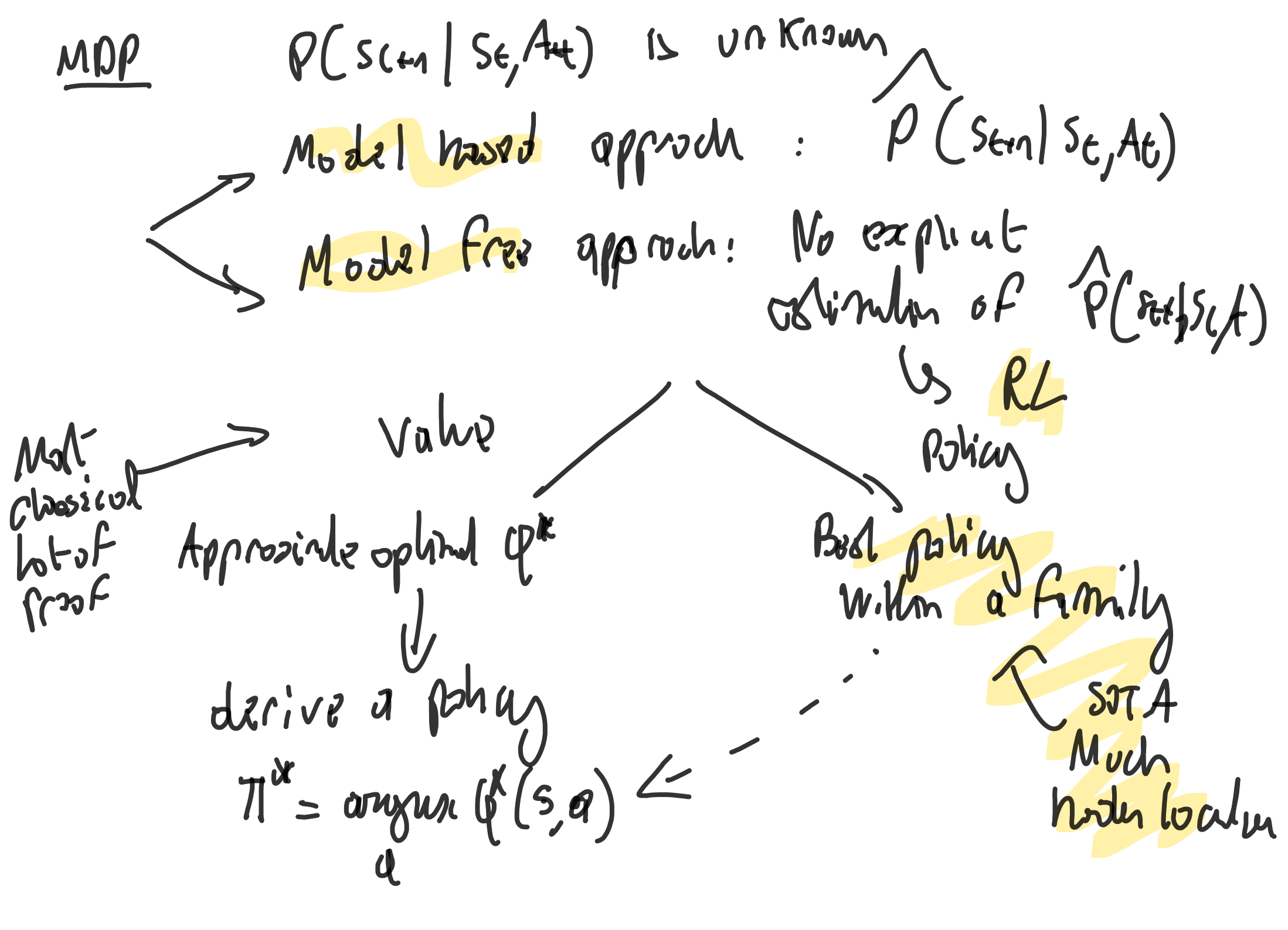
↳ V_π is the solution of a linear equation

Best Policy $Q^*(s,a) = r(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a')$

Solution: → solving this system

↘ Linear Programming

↘ Iterative schemes: Fixed Point, Banach, Riesz


MDP
$P(s_{t+1} \mid s_t, A_t)$ is unknown
Model based approch : $\hat{P}(s_{t+1} \mid s_t, A_t)$
Model free approch: No explicit obtention of $\hat{P}(s_{t+1} \mid s_t, A_t)$
RL
Policy
Value
Most classical but of work
Approximate optimal q^*
derive a policy
$\pi^* = \arg\max_a q^*(s,a)$
Best policy within a family
SOTA
Much harder to obtain