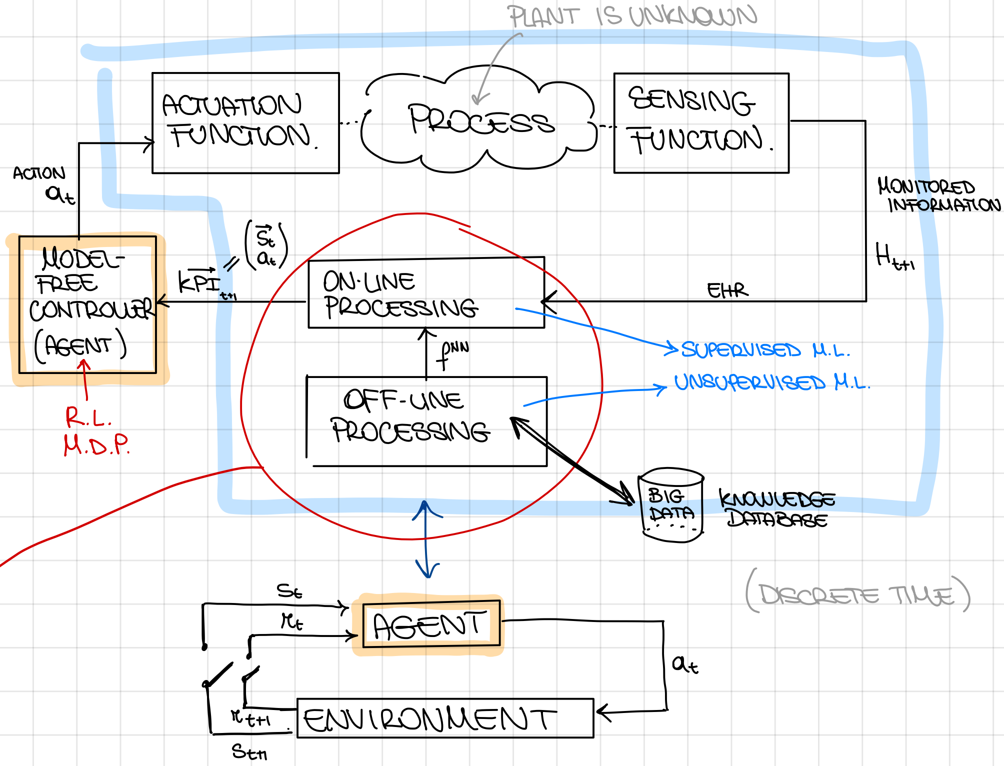
# CCEN

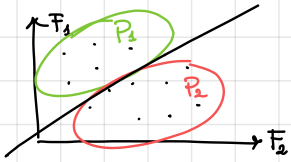
prof Della Pruscola

# INTRODUCTION




- MACHINE LEARNING TECHNIQUES: { UNSUPERVISED → PROFILING (OFF-LINE)
  SUPERVISED → NEURAL NETWORKS (ON-LINE)
  REINFORCEMENT

→ PROFILING: (es/ K-MEANS)




(K=2)

→ NEURAL NETWORKS: GIVEN THE TABLE (INPUT-OUTPUT)

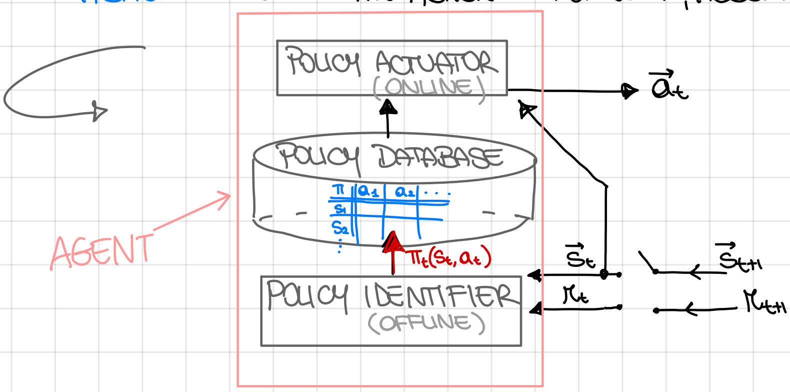
- R.L. = REINFORCEMENT LEARNING → NO KNOWLEDGE OF THE PROCESS IS NEEDED TO CONTROL IT.
- M.D.P. = MARKOV DECISION PROBLEM → KNOWLEDGE NEEDED

STATE → MARKOV PROPERTY: $\boxed{P\{S_{t+1} | S_t, a_t\} = P\{S_{t+1} | S_t, S_{t-1}, \dots; a_t, a_{t-1}, \dots\}}$

REWARD → MAXIMIZE THE EXPECTED VALUE OF THE LONG TERM RETURN $\boxed{\max E_\pi\left[\sum_{k=0}^{T} \gamma^k \cdot r_{t+k+1}\right]}$ $R_t$

→ YOU CAN ALSO CONSIDER $P\{r_{t+1} | S_{t+1}, S_t, a_t\} = P\{r_{t+1} | S_{t+1}, S_t, S_{t-1}, \dots; a_t, a_{t-1}, \dots\}$

ACTION → SELECT THE ACTION TO PERFORM, ACCORDING TO A POLICY $\pi_t(S_t, a_t)$




- APPLY THE ACTION AND THEN MONITOR ⇒ INCREASE / DECREASE VALUES ON π-DATABASE
- IF INCREASES IN DIMENSIONALITY ⇒ CAN BE REPLACED BY NEURAL NETWORK.

# MARKOV DECISION PROCESS → YOU KNOW THE PROCESS THROUGH:

- TRANSITION PROBABILITIES: $P_{ss'}^{a} = P\{S_{t+1}=s' \mid S_t=s, a_t=a\}$
- EXPECTED VALUE OF THE NEXT REWARD: $R_{ss'}^{a} = E[r_{t+1} \mid S_t=s, a_t=a, S_{t+1}=s']$

→ TARGET: FIND THE OPTIMAL POLICY $\pi^*$ WHICH MAXIMIZE $E_\pi[R_t]$

$$\max \quad E_\pi[R_t] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right] = \sum_{s\in S} E_\pi[R_t \mid S_t=s]\, P\{S_t=s\}$$

$= V_\pi^0(s)$

- ## STATE VALUE FUNCTION FOR POLICY $\pi$:

$$V^\pi(s) = E_\pi[R_t \mid S_t=s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1} \,\Big|\, S_t=s\right]$$

→ EXPECTED LONGTERM REWARD WHEN STARTING IN S AND FOLLOWING THE POLICY $\pi$ THEREAFTER.

→ BELLMAN EQUATION:

$t \to t+1$ $\quad$ $t+1 \to \infty$

$$V^\pi(s) = \underbrace{\sum_{a\in A(s)} \pi(s,a) \sum_{s'\in S} P_{ss'}^{a}\left[R_{ss'}^{a}\right.}_{\text{KNOWN}} + \gamma \cdot \underbrace{\left.V^\pi(s')\right]}_{\text{FIXED}} \quad \rightsquigarrow \max$$

→ THE VALUE OF THE STARTING STATE = WEIGHTED SUM OF THE POSSIBLE ARRIVAL STATES
+ THE REWARDS ASSOCIATED TO THE TRANSITION FROM THE START TO THE POSSIBLE ARRIVALS STATES.

→ THE WEIGHT OF EACH POSSIBLE ARRIVAL STATES = PROBABILITY OF ARRIVING AT SUCH STATE.

→ HOW TO GET AN OPTIMAL POLICY? → $\pi^*$ SUCH THAT: $V^{\pi^*}(s) \geq V^\pi(s)$, $\forall s, \forall \pi$.

(TH. → THERE ALWAYS EXISTS.)

- ## ACTION-VALUE FUNCTION FOR POLICY $\pi$:

$$Q^\pi(s,a) = E_\pi[R_t \mid S_t=s, a_t=a] \rightsquigarrow = \sum_{s'\in S} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma \cdot \max_{a'\in A(s')} Q^{\pi^*}(s',a')\right]$$

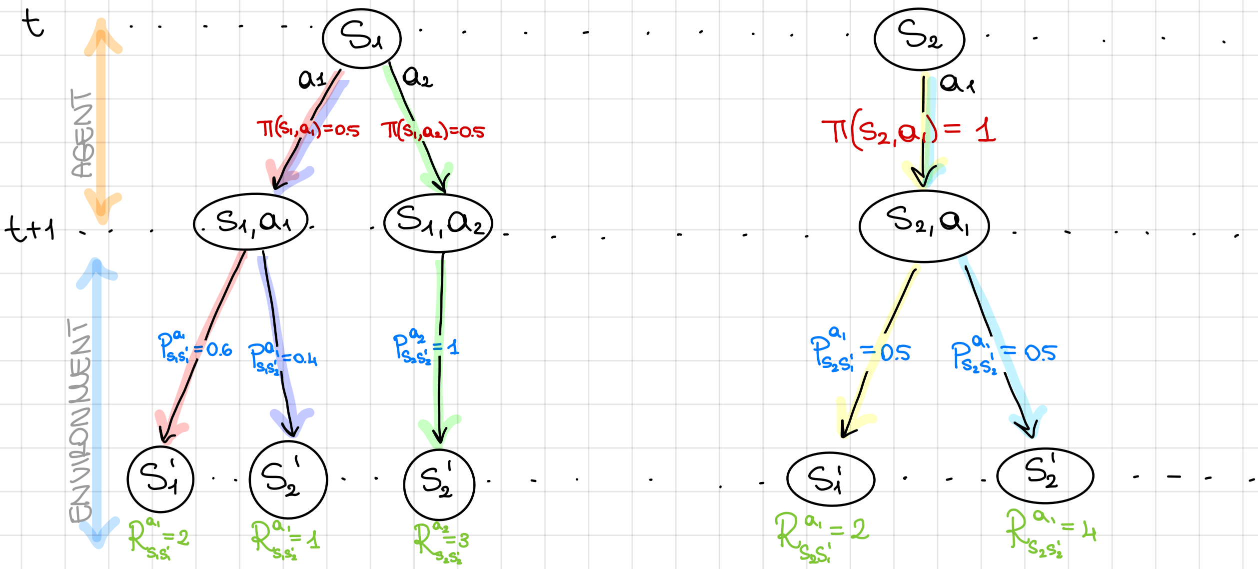
→ EXPECTED LONG TERM RETURN WHEN TAKING THE ACTION $a$ IN STATE $s$ AND FOLLOWING THE POLICY $\pi$.

$$V^\pi(s) = \dots = \sum_{a\in A(s)} \underbrace{E_\pi[R_t \mid S_t=s, a_t=a]}_{=Q^\pi(s,a)} \cdot \underbrace{P\{a_t=a \mid S_t=s\}}_{=\pi} = \sum_{a\in A(s)} Q^\pi(s,a) \cdot \pi(s,a)$$

# BACKUP DIAGRAM:

es/ CAN ROBOT → ACTIONS: $A=\{a_1,a_2\}$

→ STATES: $S=\{s_1,s_2\}$




t→t+1    t+1→∞ → $\gamma^1$ BECAUSE I STEP

$$\Rightarrow \begin{cases} V^\pi(s_1) = 0{,}5\cdot 0{,}6\left[2+\gamma\, V^\pi(s_1')\right] + 0{,}5\cdot 0{,}6\left[1+\gamma\cdot V^\pi(s_2')\right] + 0{,}5\cdot 1\left[3+\gamma\cdot V^\pi(s_2')\right] \\ V^\pi(s_2) = 1\cdot 0{,}5\left[2+\gamma\, V^\pi(s_1')\right] + 1\cdot 0{,}5\left[4+\gamma\cdot V^\pi(s_2')\right] \end{cases}$$

→ YOU CAN SOLVE IT AS $\begin{cases} V^\pi(s_1) = V^\pi(s_1') = x \\ V^\pi(s_2) = V^\pi(s_2') = y \end{cases}$

$$\begin{cases} x = (0{,}6+0{,}27x) + (0{,}2+0{,}18y) + (1{,}5+0{,}45y) \\ y = (1+0{,}45x) + (2+0{,}45y) \end{cases} \Rightarrow \begin{cases} 0{,}73x - 0{,}63y = 2{,}3 \\ -0{,}45x + 0{,}55y = 3 \end{cases} \Rightarrow \begin{cases} x = 26{,}66 \\ y = 27{,}27 \end{cases}$$

$$\Rightarrow E_\pi[R_t] = 26{,}66\cdot P\{s_t=s_1\} + 27{,}77\cdot P\{s_t=s_2\}$$

THE OPTIMAL WILL BE:

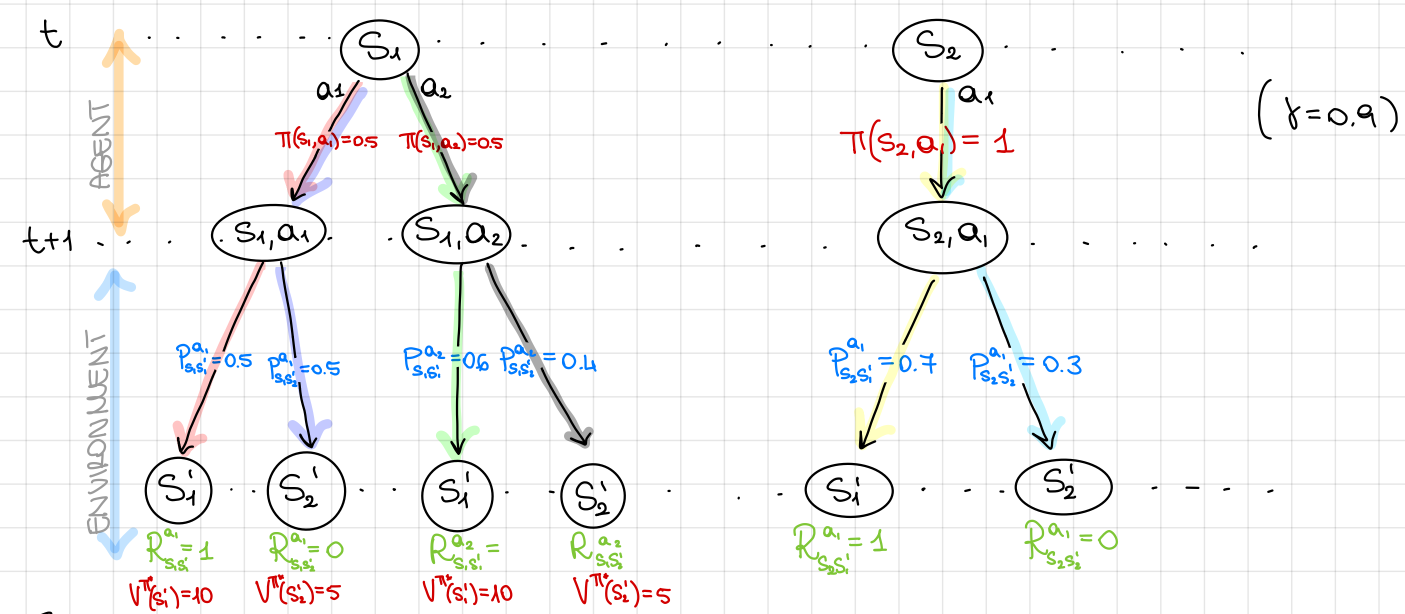
- $\boxed{V^{\pi^*}(s) = \max\limits_{a\in A(s)} \sum\limits_{s'\in S} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma \cdot V^{\pi^*}(s')\right]}$ ← OPTIMAL BELLMAN EQUATION

- $\boxed{a^*(s) = \arg\max\limits_{a\in A(s)} \sum\limits_{s'\in S} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma \cdot V^{\pi}(s')\right]}$ ← OPTIMAL ACTION TO PERFORME

- $\boxed{Q^{\pi^*}(s,a) = \sum\limits_{s'\in S} P_{ss'}^{a}\left[R_{ss'}^{a} + \gamma \cdot \max\limits_{a'\in A(s')} Q^{\pi^*}(s',a')\right]}$

NB: When you select the next optimal action it is not optimal just for the next step but it is optimal in a future prospective.

ES/ BACKUP DIAGRAM:




$S_1$:
$$\begin{cases} \text{PATH (red)}: & 1+0{,}9\cdot 10 = 10 \\ \text{PATH (blue)}: & 0+0{,}9\cdot 5 = 4{,}5 \\ \text{PATH (green)}: & 1+0{,}9\cdot 10 = 10 \\ \text{PATH (grey)}: & 0+0{,}9\cdot 5 = 4{,}5 \end{cases}$$

- $V^{\pi^*}(s_1) = \max\limits_{a_1,a_2}\left\{\overbrace{0{,}5(1+0{,}9\cdot 10) + 0{,}5(0+0{,}9\cdot 5)}^{\text{SELECT } a_1} ; \overbrace{0{,}6(1+0{,}9\cdot 10) + 0{,}4(0+0{,}9\cdot 5)}^{\text{SELECT } a_2}\right\}$

  $= \max\{7{,}25 ; 7{,}8\} \Rightarrow 7{,}8 \Rightarrow \underline{\underline{a_2}}$ → SO IF I AM IN $S_1$ I HAVE TO SELECT $a_2$.

- $a^*(s_1) = \arg\max\limits_{a_1,a_2}\{7{,}25 ; 7{,}8\} = a_2$

CALL: $\begin{cases} V^{\pi^*}(s_1) = V^{\pi^*}(s_1') = x_1 \\ V^{\pi^*}(s_2) = V^{\pi^*}(s_2') = x_2 \end{cases}$

$$\begin{cases} x_1 = \max\left\{\overbrace{0{,}5(1+0{,}9\cdot x_1) + 0{,}5(0+0{,}9\cdot x_2)}^{a_1} ; \overbrace{0{,}6(1+0{,}9\cdot x_1) + 0{,}4(0+0{,}9\cdot x_2)}^{a_2}\right\} & S_1 \\ x_2 = 0{,}7(1+0{,}9\cdot x_1) + 0{,}3(0+0{,}9\cdot x_2) & S_2 \end{cases}$$

. . .

$$\begin{cases} X_1 = \max\{0{,}45\,X_1 + 0{,}45\,X_2 + 0{,}5\ ;\ 0{,}54\,X_1 + 0{,}36\,X_2 + 0{,}6\} \\ X_2 = 0{,}63\,X_1 + 0{,}27\,X_2 + 0{,}7 \end{cases}$$

↓ TO SOLVE

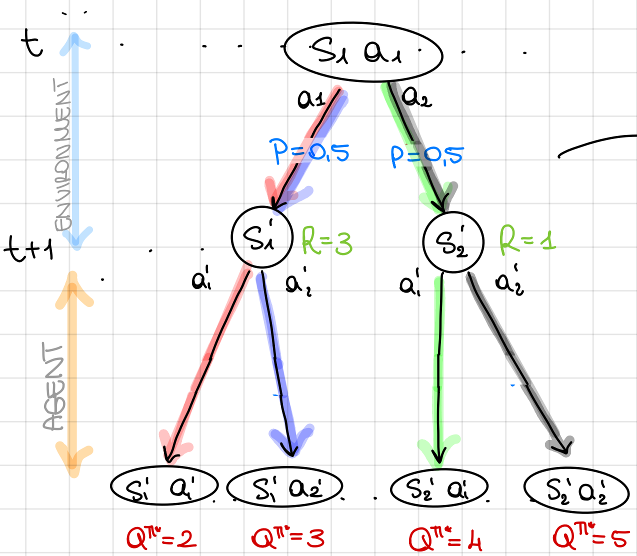
- HYP 1 → $a_1$ WINS ⇒ $\begin{cases} X_1 = 0{,}45\,X_1 + 0{,}45\,X_2 + 0{,}5 = 5{,}76 \\ X_2 = \ \ '' \qquad = 5{,}93 \end{cases}$
- HYP 2 → $a_2$ WINS ⇒ $\begin{cases} X_1 = 0{,}54\,X_1 + 0{,}36\,X_2 + 0{,}6 = 6{,}33 \\ X_2 = \ \ '' \qquad = 6{,}42 \end{cases}$

HIGHER!

⇒

| $\pi^*$ | $a_1$ | $a_2$ |
|---|---|---|
| $S_1$ | 0 | 1 |
| $S_2$ | 1 | X |

REMARK → I CAN ALSO CONSIDER Q(S):




$R + \gamma \cdot Q^{\pi^*}(s,a)$

$$\begin{cases} \text{PATH}: 3 + 0{,}9 \cdot 10 = 10 \\ \text{PATH}: 3 + 0{,}9 \cdot 5 = 4{,}5 \\ \text{PATH}: 1 + 0{,}9 \cdot 10 = 10 \\ \text{PATH}: 1 + 0{,}9 \cdot 5 = 4{,}5 \end{cases}$$

⇒

| $Q^{\pi^*}$ | $a_1$ | $a_2$ |
|---|---|---|
| $S_1$ | 24 | 52 |
| $S_2$ | 30 | 18 |

$$\begin{cases} Q^{\pi^*}(S_1 a_1) = Q^{\pi^*}(S_1'\ a_1') \\ Q^{\pi^*}(S_1 a_2) = Q^{\pi^*}(S_1'\ a_2') \\ Q^{\pi^*}(S_2 a_1) = Q^{\pi^*}(S_2' a_1') \\ Q^{\pi^*}(S_2 a_2) = Q^{\pi^*}(S_2'\ a_2') \end{cases}$$

→ N UNKNOWN

$N = |S| \cdot |A| = 4$

---

NB → DIFFERENCE BETWEEN → BELLMAN EQ.:

| $\pi$ | $a_1$ | $a_2$ |
|---|---|---|
| $S_1$ | 0.5 | 0.5 |
| $S_2$ | 1 | / |

→ OPTIMAL BELLMAN EQ.:

| $\pi^*$ | $a_1$ | $a_2$ |
|---|---|---|
| $S_1$ | 0 | 1 |
| $S_2$ | 1 | / |

# DYNAMIC PROGRAMMING (ch. 4)

↳ HOW TO SOLVE IN A SMARTER WAY → TWO METHODS

## ① POLICY ITERATION

→ EVALUATION + IMPROVEMENT PROCESS

$$\pi \xrightarrow{\text{EVALUATION}} V^{\pi} \xrightarrow{\text{IMPROVEMENT}} \pi' \xrightarrow{\text{EVAL}} V^{\pi'} \xrightarrow{\text{IMP}} \pi'' \xrightarrow{\text{EVAL}} \cdots \xrightarrow{\text{IMP}} \pi^*$$

BELLMAN EQ.    $\pi' \geq \pi$, $V^{\pi'}(s) \geq V^{\pi}(s)$    $\pi'' \geq \pi' \geq \pi$ ...    WHEN $V^{\bar{\pi}}(s) = V^{\bar{\bar{\pi}}}(s)$

### EVALUATION:

OR
- BY HAND (AS WE ALREADY KNOW)
- COMPUTING $\boxed{V^{\pi}_{k+1}(s) := \sum_{a \in A(s)} \pi(s,a) \cdot \sum_{s' \in S} P^{a}_{ss'} \left[ R^{a}_{ss'} + \gamma \cdot V^{\pi}_{k}(s) \right]}$, $k = 0, 1, \ldots$

↳ AS $k \to \infty \Rightarrow$ IT CONVERGES TO $V^{\pi}(s)$

es/ $V^{\pi}_{0}(s_1) = V_0(s_1') = 0$, $V^{\pi}_{0}(s_2) = V^{\pi}_{0}(s_2') = 0$

$\underline{k=0} \to V^{\pi}_{1}(s_1) = 0.55$

$V^{\pi}_{1}(s_2) = 1.085$

| $V(s_1)$ | $V(s_2)$ |
|---|---|
| $\emptyset$ | $\emptyset$ |
| 0.55 | 1.085 |
| ⋮ | ⋮ |

↓

STOP WHEN $|V_{k+1}(s) - V_k(s)| < \Delta$   PREFIXED

### IMPROVEMENT:

$$\pi' \geq \pi \quad \begin{cases} \pi'(s, \bar{a}(s)) = 1 \\ \pi'(s,a) = 0 \quad \forall a \neq \bar{a}(s), \forall s \end{cases}$$

↳ COMPUTE $\boxed{\bar{a}(s) = \arg\max_{a} Q^{\pi}(s,a) = \arg\max_{a \in A(s)} \sum P^{a}_{ss'} \cdot \left[ R^{a}_{ss'} + \gamma \cdot V^{\pi}(s) \right]}$

es/ $\bar{a}(s_1) = \arg\max_{a_1, a_2} \{ Q^{\pi}(s_1, a_1) ; Q^{\pi}(s_1, a_2) \} = a_2$

$\bar{a}(s_2) = \cdots = a_1$

$\Rightarrow$

| $\pi'$ | $a_1$ | $a_2$ |
|---|---|---|
| $s_1$ | 0 | 1 |
| $s_2$ | 1 | / |

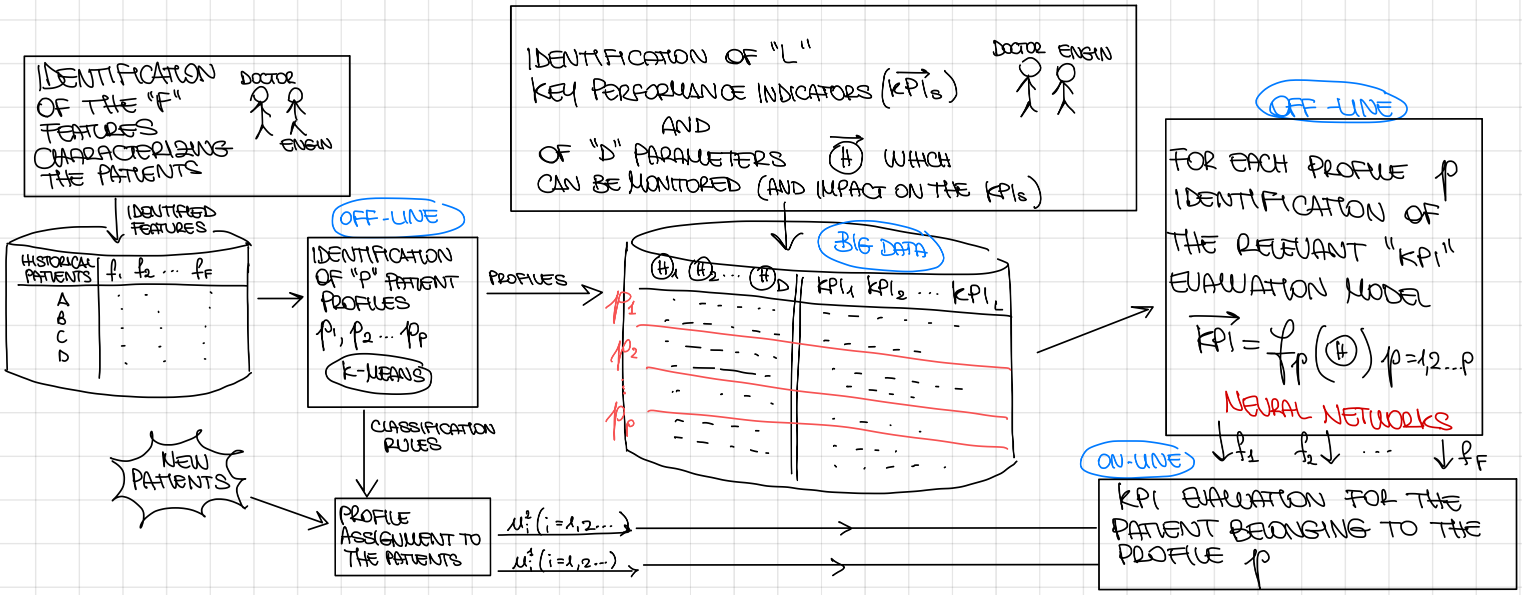
→ SINCE $\pi'' = \pi' \Rightarrow \pi' = \pi^*$

## ② VALUE ITERATION

UPDATE THE OPTIMAL BELLMAN EQUATION: $\boxed{V^{*}_{k+1}(s) := \max_{a \in A(s)} \sum_{s \in S} P^{a}_{ss'} \cdot \left[ R^{a}_{ss'} + \gamma \cdot V^{*}_{k}(s') \right]}$, $k = 0, 1, \ldots$

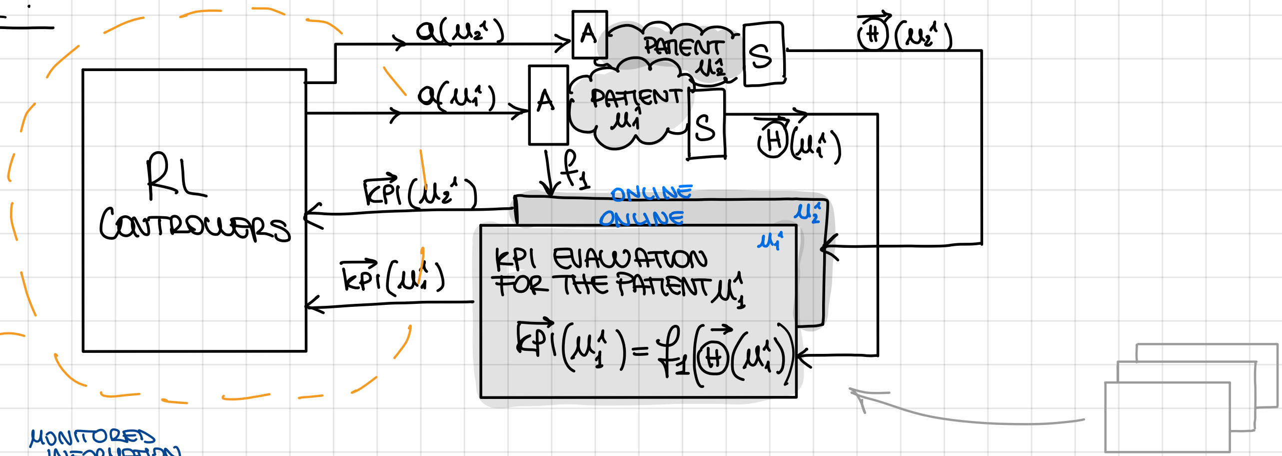
↳ WHEN $|V_{k+1}(s) - V_k(s)| < \Delta$ end.

# -CAMS-




PROFILE 1:

- $f_1$
- $\mu_1^1$, $\mu_2^1$

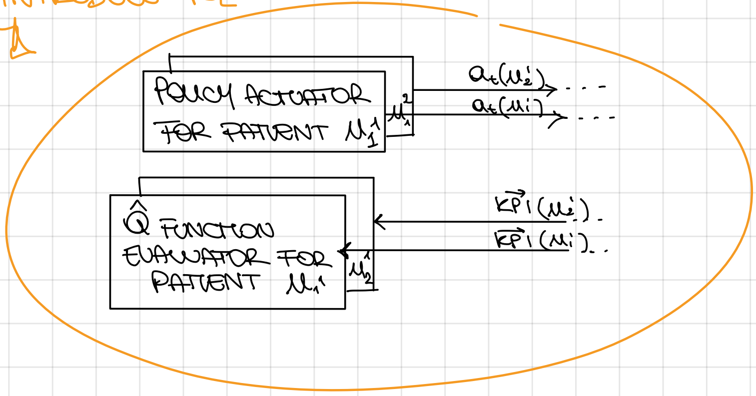



MONITORED INFORMATION

$$\vec{KPI} = f_i(\vec{H}) \qquad i = 1,2,\dots p$$

- TRAINING PHASE → DEDUCE $W_i$ TO REDUCE $f_1$
- OPERATIONAL PHASE → USE $KPI(\mu_i^1) = f_i(\vec{H})$ [NN]

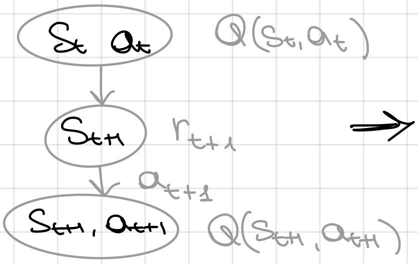
INTRODUCE RL :




WITH $\vec{KPI} = \begin{pmatrix} \vec{S} \\ r \end{pmatrix}$

- $Q_\pi(S_t, a_t) = E_\pi\Big[\underbrace{\sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k+1}}_{R_t} \Big| S_t, a_t\Big] \longrightarrow Q^*(S_t, a_t) = \max_\pi Q_\pi(S, a)$
- $r_t = -\|\vec{S_t} - S_{target}\|^2$

RECALL:

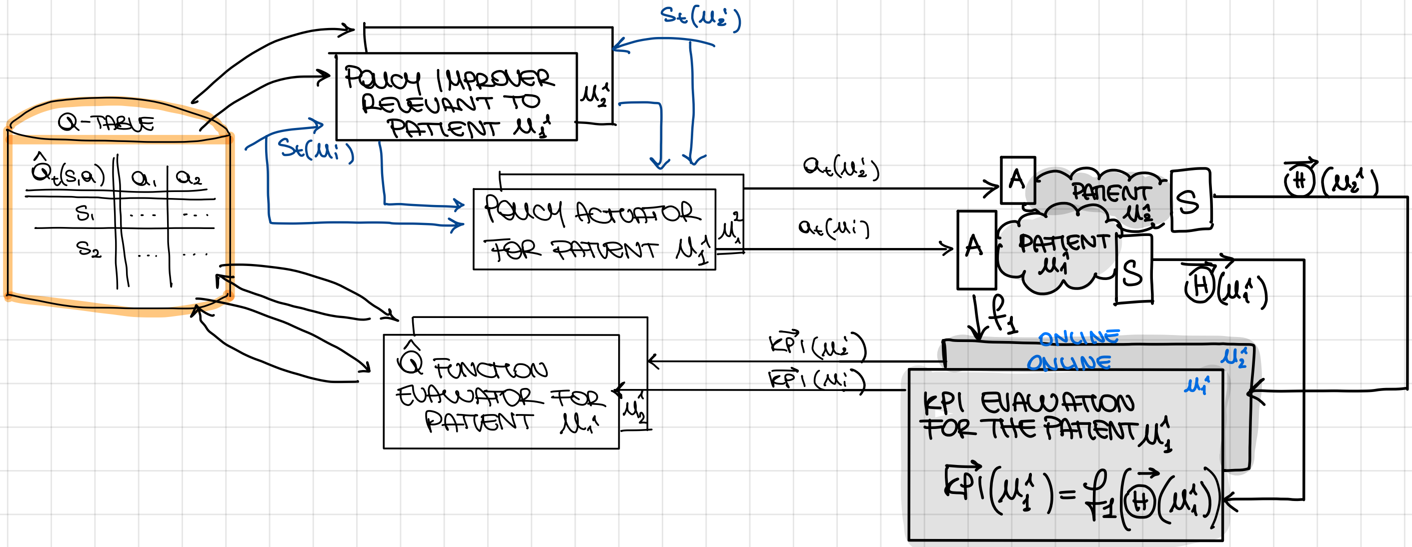



$\Longrightarrow$ Q - LEARNING

$$\alpha_t = \frac{1}{1+k(s,a)}$$

$$\boxed{\underbrace{\hat{Q}^*(S_t, a_t)}_{\text{NEW}} := \underbrace{\hat{Q}^*(S_t, a_t)}_{\text{OLD}} + \underbrace{\alpha_t}_{\text{STEP SIZE}} \Big[ \underbrace{r_{t+1} + \gamma \cdot \max_{a' \in A(S_{t+1})} \hat{Q}^*(S_{t+1}, a')}_{\text{NEW SAMPLE}} - \underbrace{\hat{Q}^*(S_t, a_t)}_{\text{OLD}} \Big]}$$

$\hookrightarrow$ TO BUILD THE Q-TABLE




- POLICY IMPROVER COMPUTE THE PREFERRED ACTION $= \hat{a}_t = \underset{a_t \in A(S_t)}{\operatorname{argmax}}\ Q^*(S_t, a)$

$$\begin{cases} \pi(S_t, \hat{a}_t) = 1 - \varepsilon + \dfrac{\varepsilon}{|A(S_t)|} \\ \pi(S_t, a) = \dfrac{\varepsilon}{|A(S_t)|} \end{cases}, \text{ WITH } \begin{matrix} a_t \neq \hat{a}_t \\ a_t \in A(S_t) \end{matrix}, \quad \varepsilon = \frac{1}{t} \qquad (\varepsilon\text{-GREEDY})$$