bigdata

BIG DATA CONFERENCE

26-28 NOVEMBER 2019, VILNIUS

Dr. Andreas Bühlmeier

www.buhlmeier.com

BREAKTHROUGHS & FUTURE OF (DEEP) REINFORCEMENT LEARNING

FOUNDATION, IMPLEMENTATION, APPLICATIONS & TRENDS

dbc ENTERPRISE INTELLIGENCE

bigdata

THIS -> (rather self.):

- 01 WHY Reinforcement Learning Matters
- 02 WHO invented it? RL roots
- 03 HOW does it work?
- 04 LIVE: Q-Learning in action
- 05 NOW: Methods & Milestones
- 06 NEXT: Future outlook

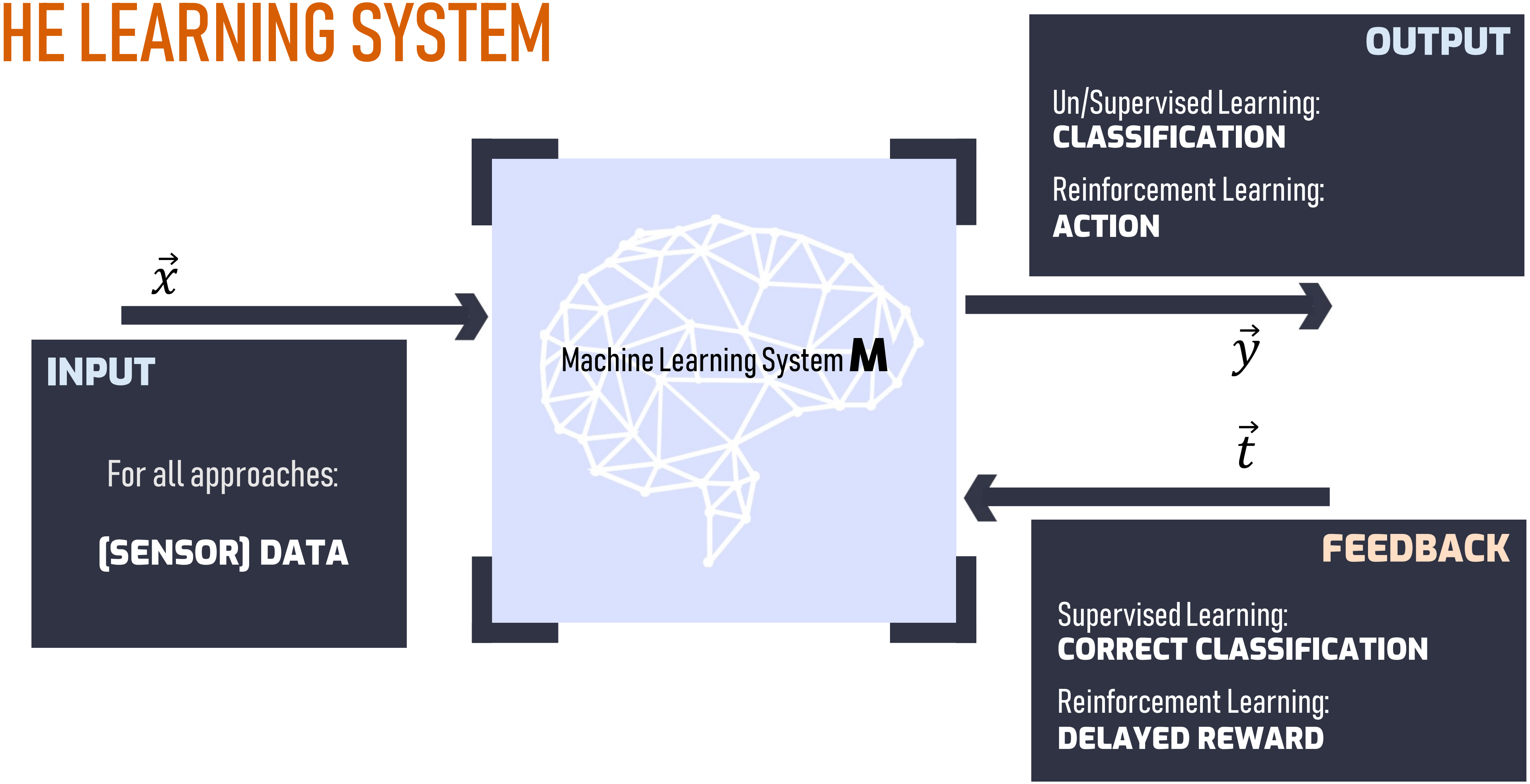
01

WHY

RL MATTERS

...it is ***different.***

THE LEARNING SYSTEM



INPUT

For all approaches:

(SENSOR) DATA

$\vec{x}$

Machine Learning System **M**

OUTPUT

Un/Supervised Learning:
CLASSIFICATION

Reinforcement Learning:
ACTION

$\vec{y}$

$\vec{t}$

FEEDBACK

Supervised Learning:
CORRECT CLASSIFICATION

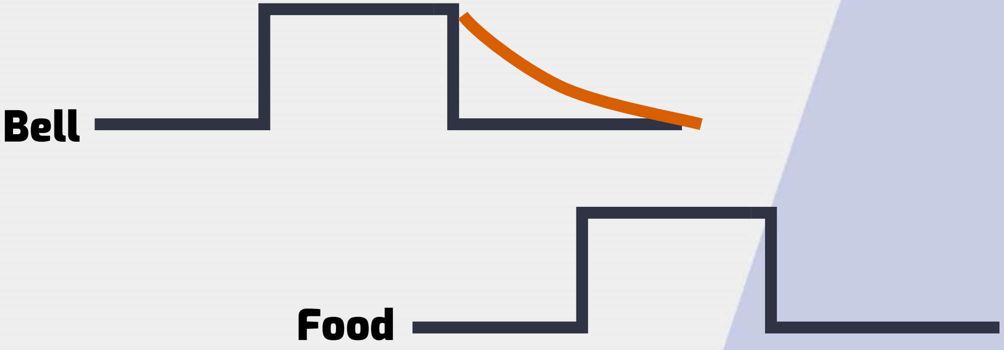
Reinforcement Learning:
DELAYED REWARD

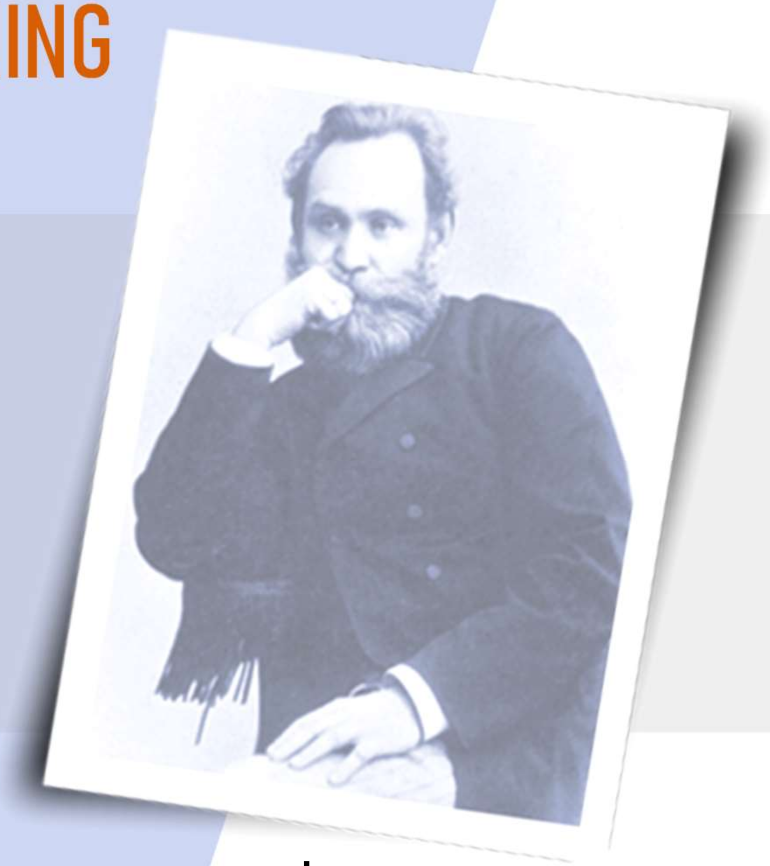
02 WHO

INVENTED IT

ROOTS of Reinforcement Learning Theories

A: PSYCHOLOGY, CLASSICAL CONDITIONING





- Pavlov (1927): … every stimulus must leave a **TRACE** in the nervous system…

See also: Barto and Sutton: Reinforcement Learning – An Introduction, 2018, MIT Press

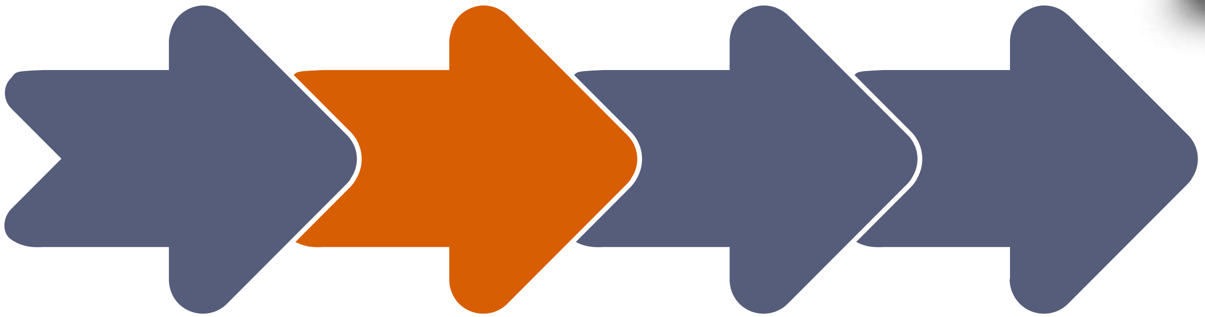
B: DYNAMIC PROGRAMMING

- Optimal Control, Bellmann (1952):

... **sequence of operations** ...

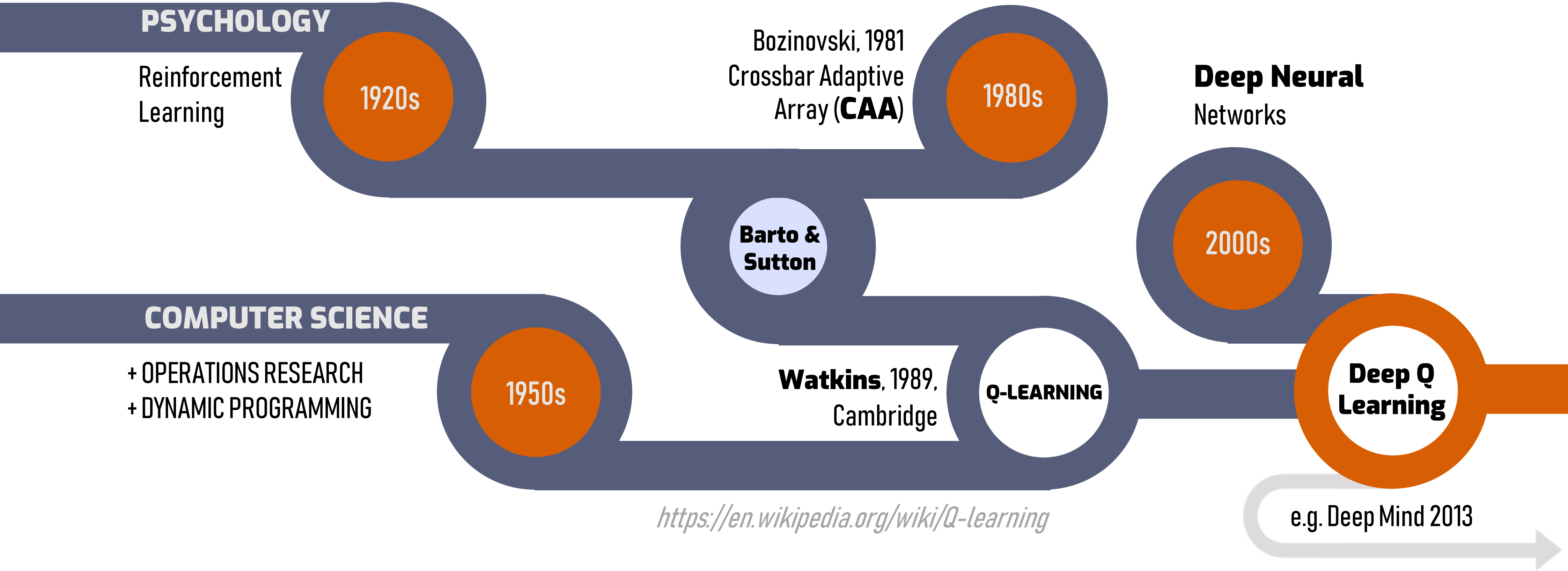
for the purpose of achieving

a desired **result**...

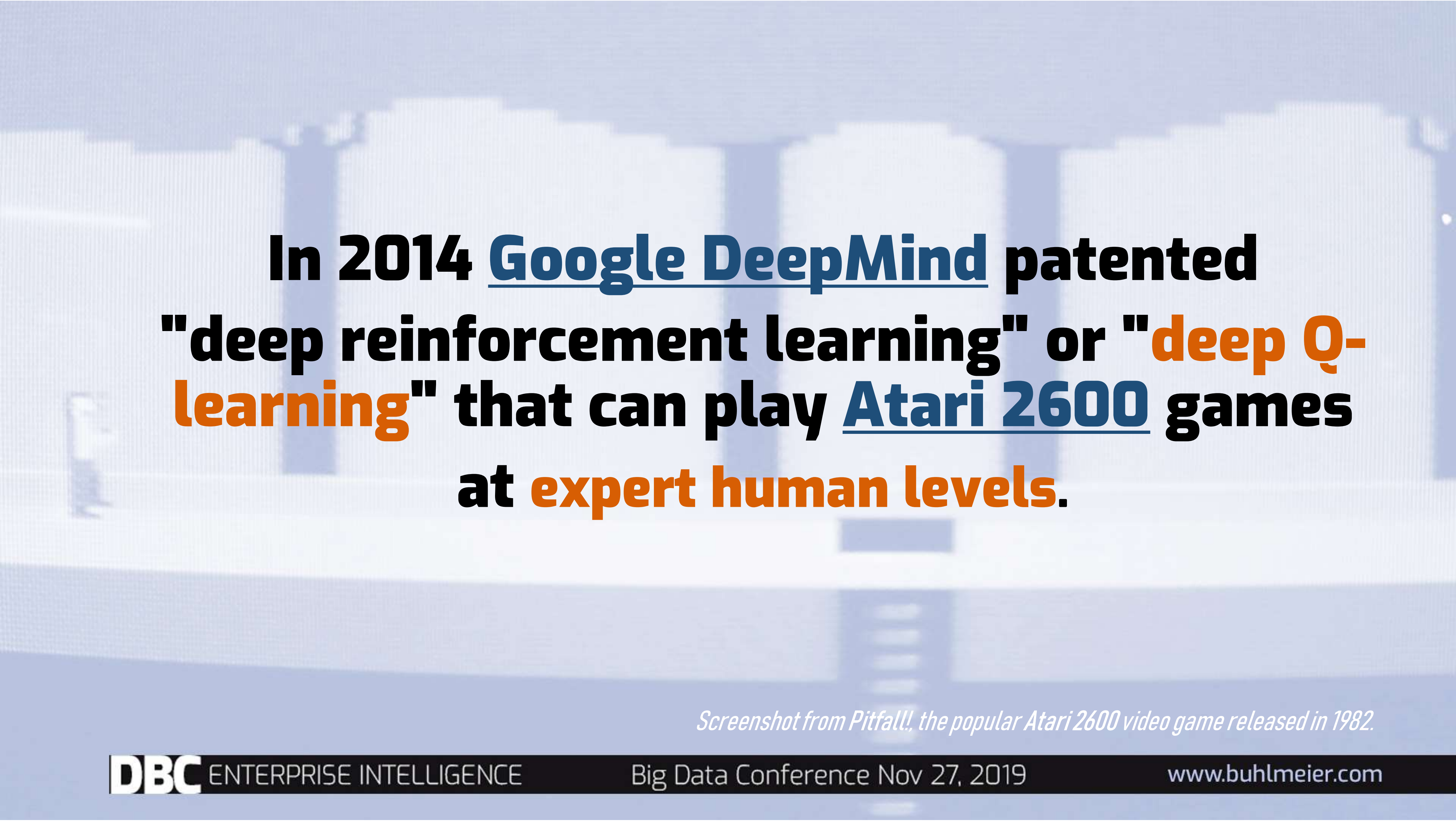


R Bellman, *On the Theory of Dynamic Programming*, Proceedings of the National Academy of Sciences

C: BRINGING IT TOGETHER: A TIMELINE OF RL



https://en.wikipedia.org/wiki/Q-learning



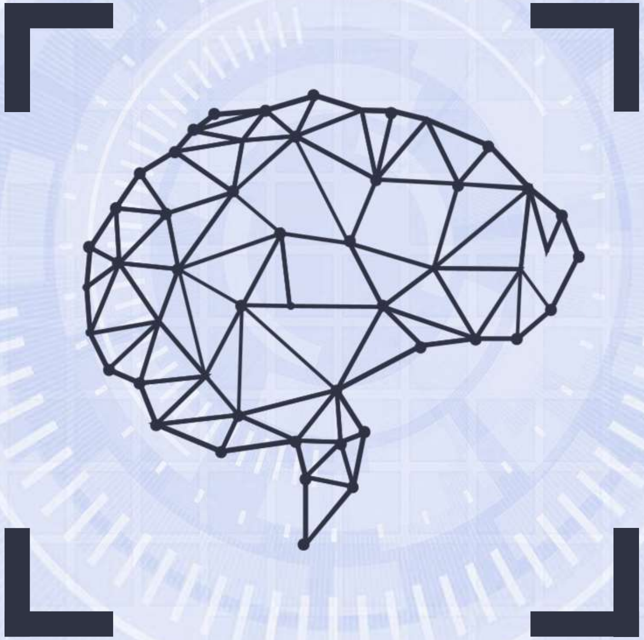
In 2014 Google DeepMind patented "deep reinforcement learning" or "deep Q-learning" that can play Atari 2600 games at expert human levels.

Screenshot from Pitfall!, the popular Atari 2600 video game released in 1982.

WHY & WHO: SUMMARY I

01 Q-Learning rooted in Psychology and Computer Science

02 Several decades of development



Main differences to other machine learning algorithms:

03 Concept of an agent that senses and acts

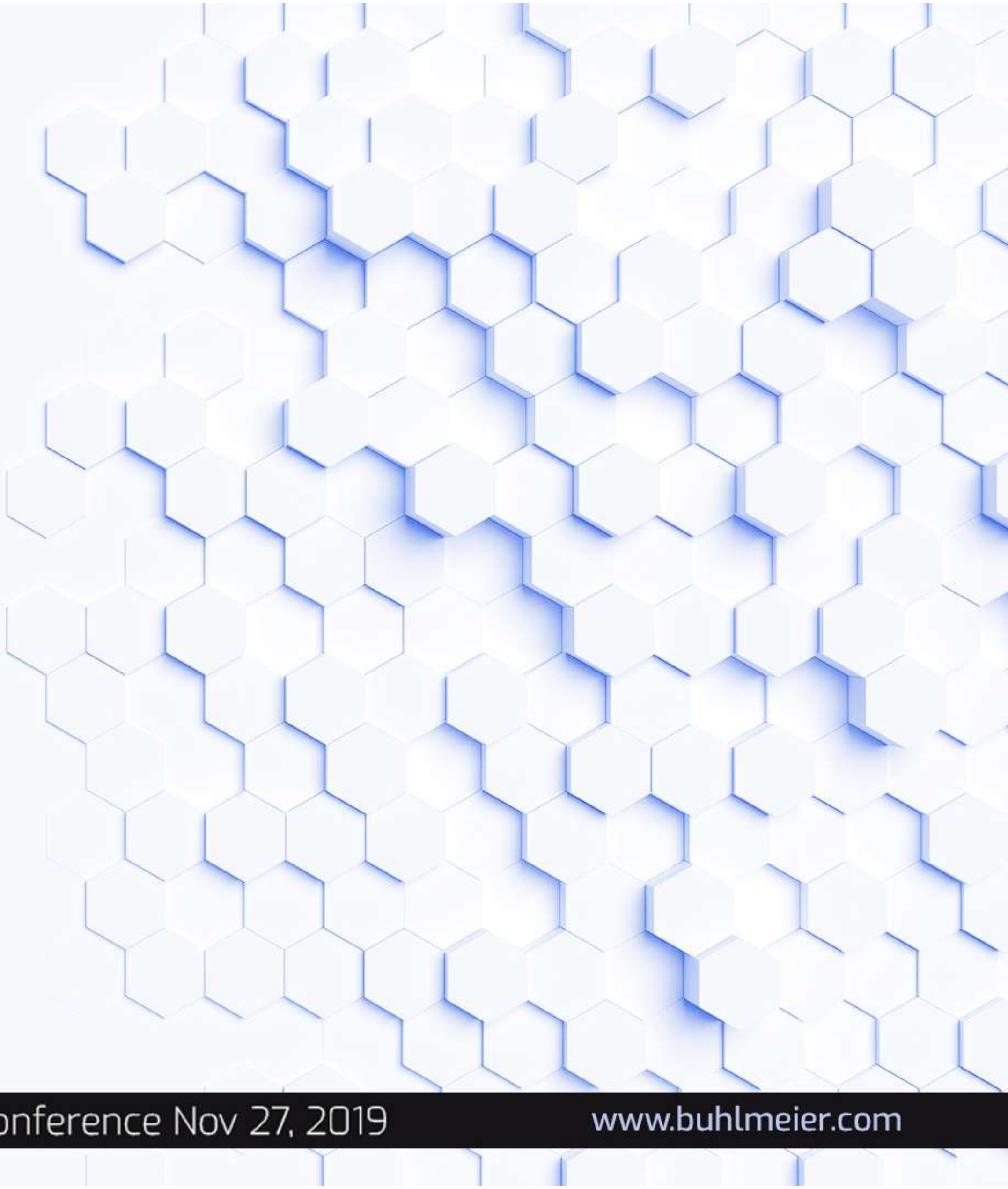
04 feedback only after a sequence of actions

03

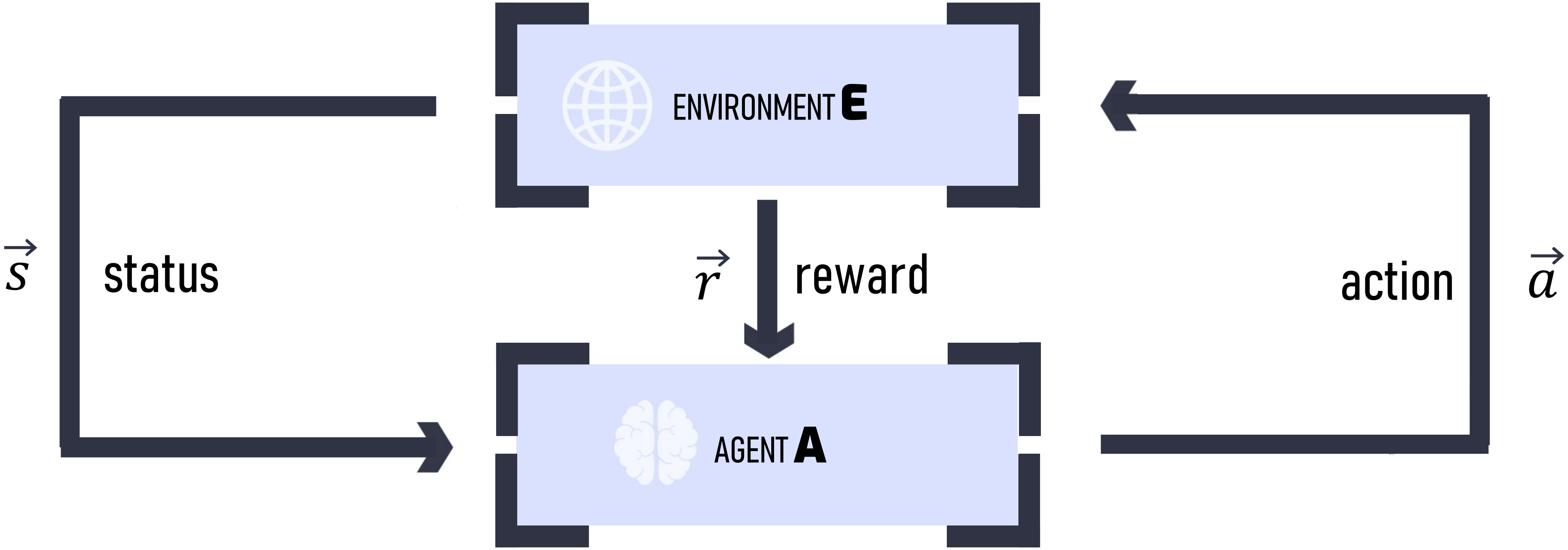
HOW

RL WORKS

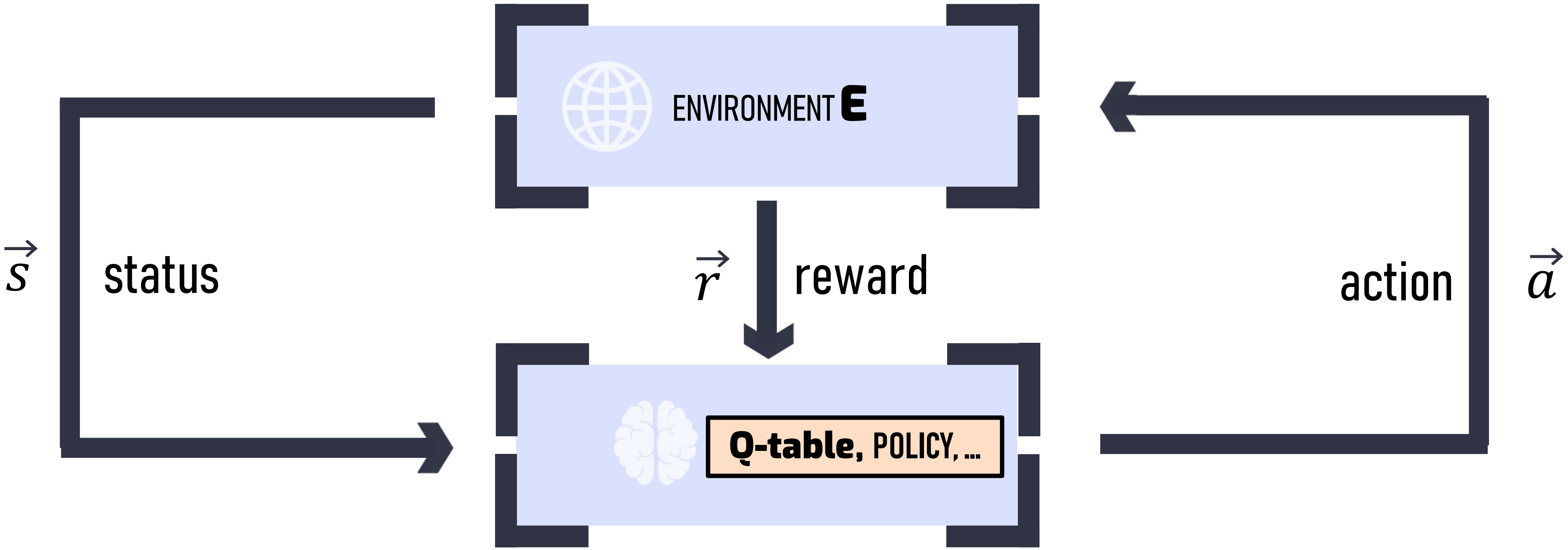
...and first successes.



ADDING AN ENVIRONMENT



ADDING AN ENVIRONMENT

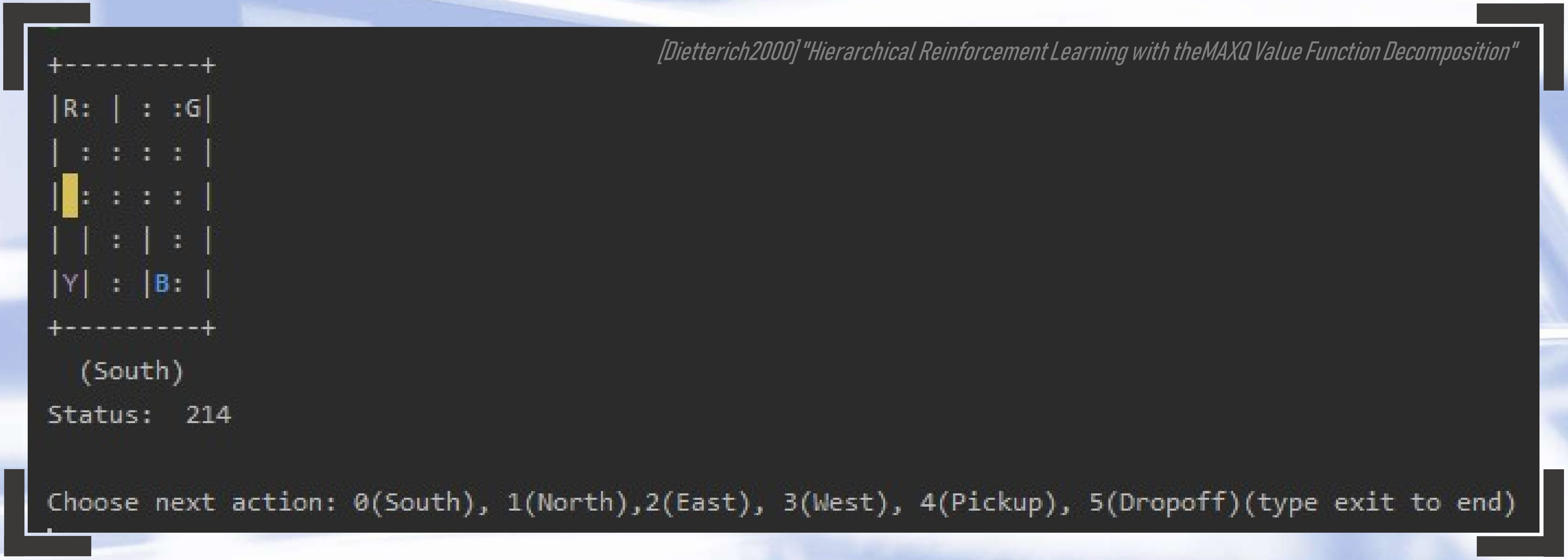


04 LIVE

Q-LEARNING in action

TAXI ENVIRONMENT FOR REINFORCEMENT LEARNING

TAXI ENVIRONMENT TEST: **SWITCHING TO LIVE**

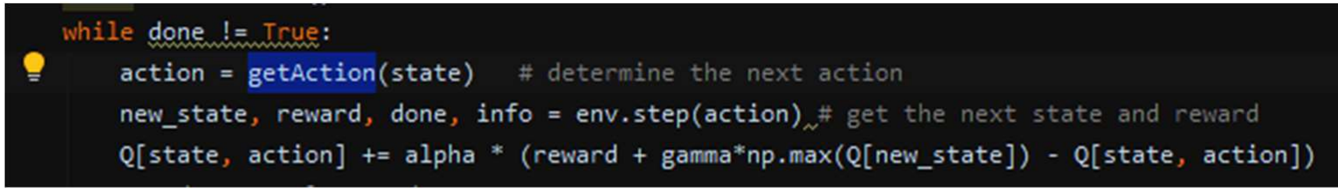


[Dietterich2000] "Hierarchical Reinforcement Learning with theMAXQ Value Function Decomposition"

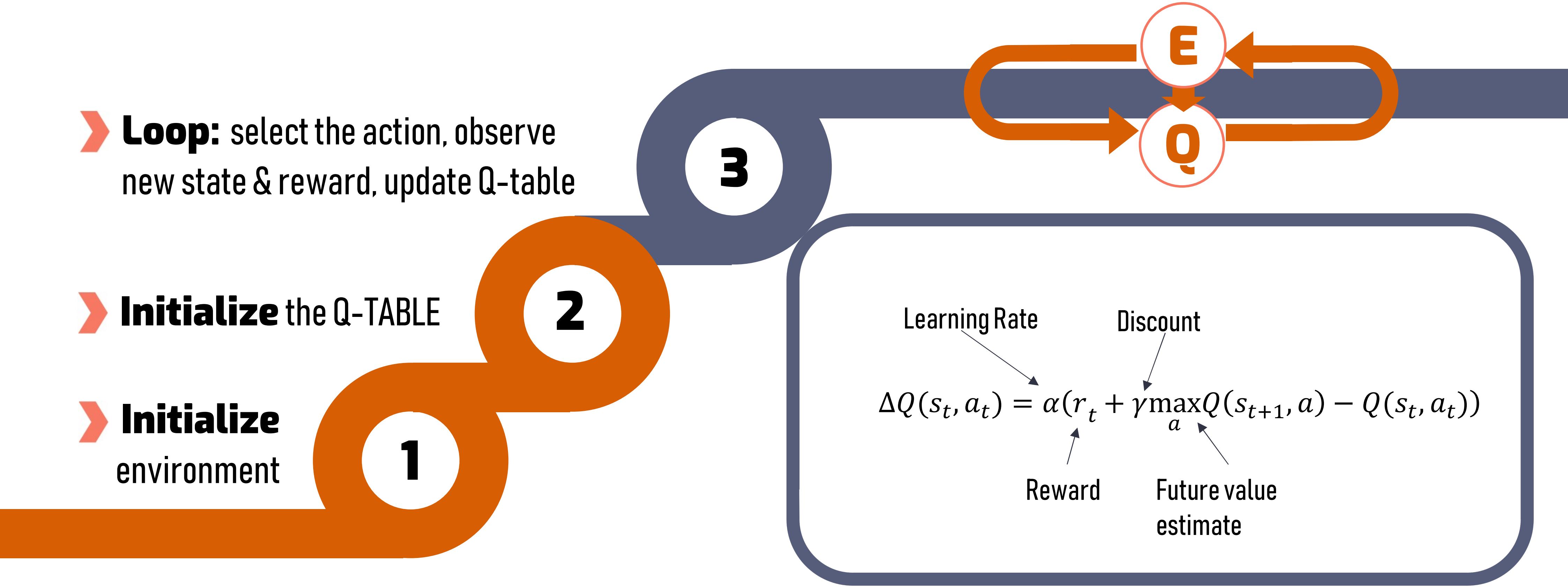
```
+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
  (South)
Status:  214


Choose next action: 0(South), 1(North),2(East), 3(West), 4(Pickup), 5(Dropoff)(type exit to end)
```

4 locations | Pick up at **blue**, drop off at **purple** | Free taxi is **yellow**, with passenger **green** | Successful drop-off +20 pts | Each timestep: -1 pt | Pick-up/drop-off penalties: -10 pts|

STEPS IN Q-LEARNING



```
while done != True:
    action = getAction(state)    # determine the next action
    new_state, reward, done, info = env.step(action) # get the next state and reward
    Q[state, action] += alpha * (reward + gamma*np.max(Q[new_state]) - Q[state, action])
```



- **Loop:** select the action, observe new state & reward, update Q-table
- **Initialize** the Q-TABLE
- **Initialize** environment

Learning Rate

Discount

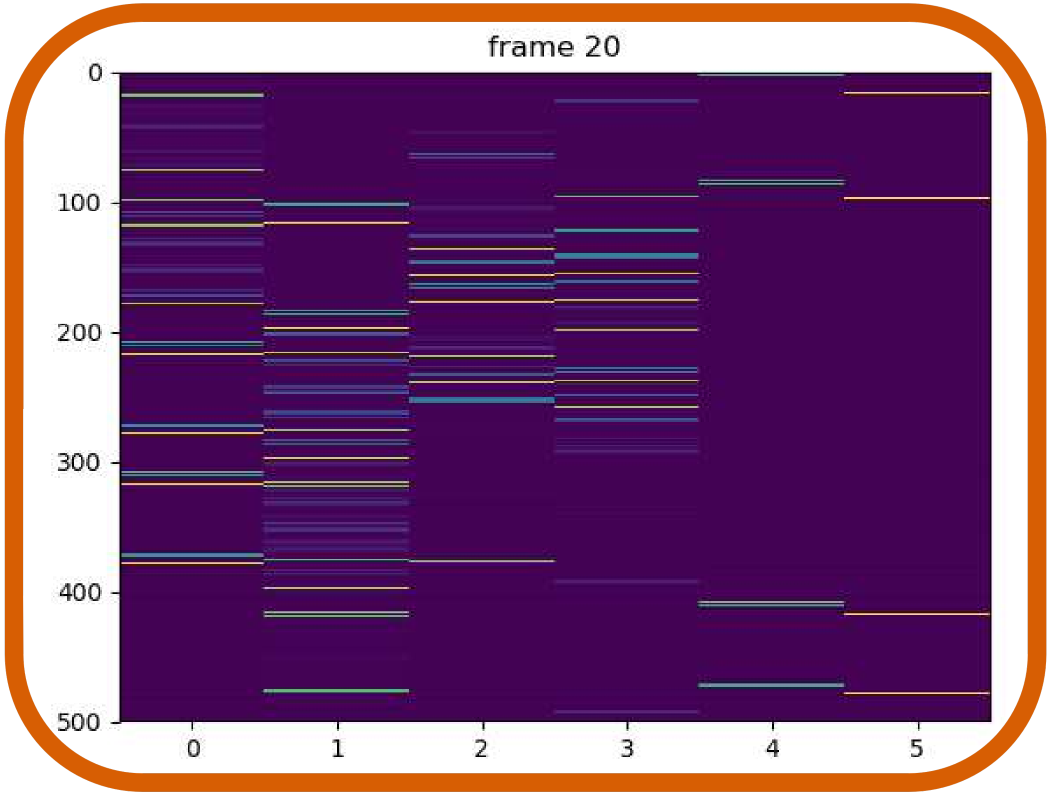
$$\Delta Q(s_t, a_t) = \alpha(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

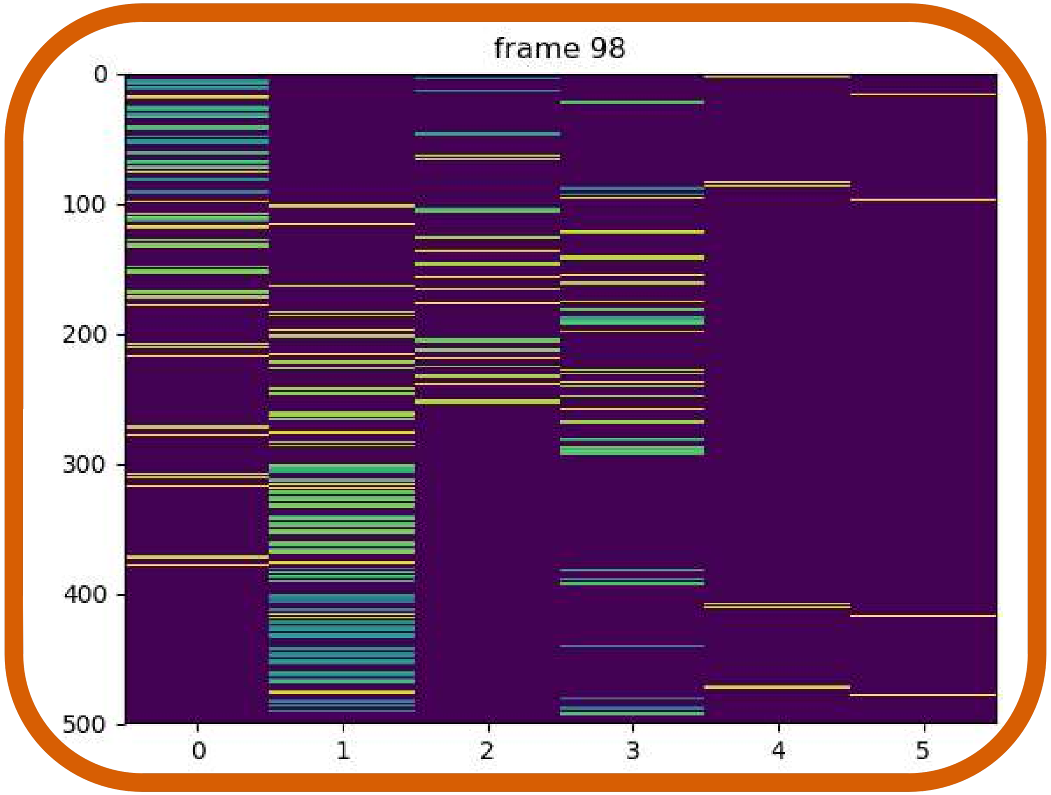
Reward

Future value estimate

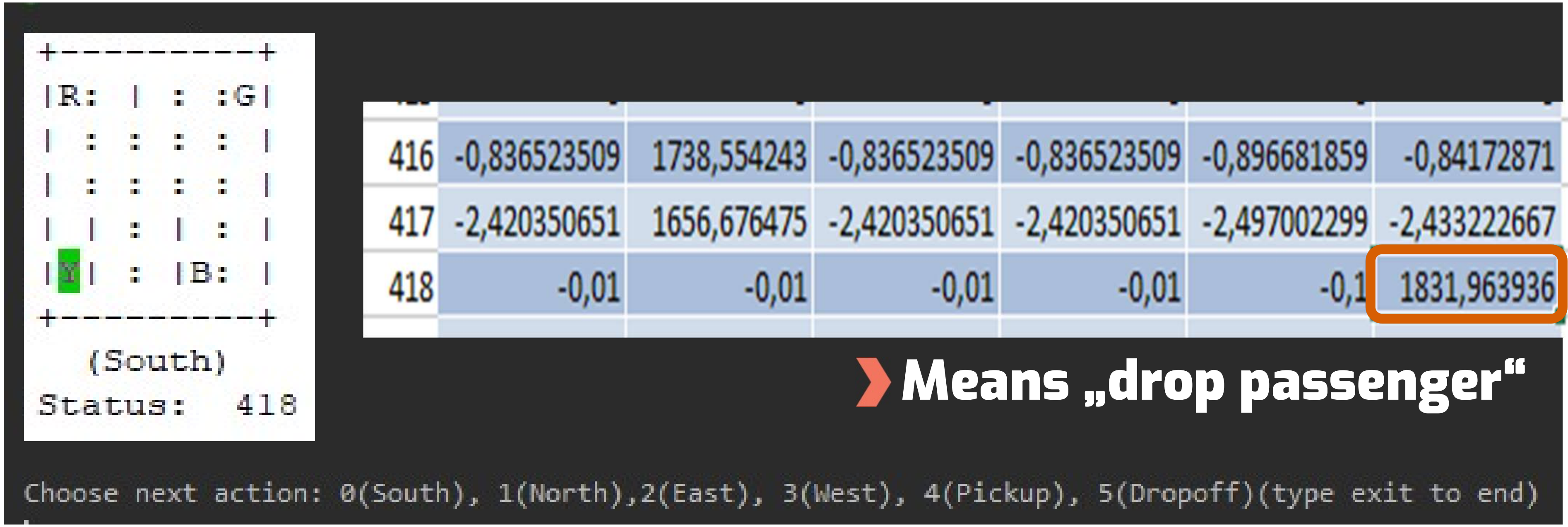
FILLING THE Q-TABLE

500 states: 25 squares, 5 locations for the passenGer, 4 destinations | **6 actions:** 4 directions, pick up, drop





DOWN TO THE NUMBERS:



```
+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
  (South)
Status:  418
```

416	-0,836523509	1738,554243	-0,836523509	-0,836523509	-0,896681859	-0,84172871
417	-2,420350651	1656,676475	-2,420350651	-2,420350651	-2,497002299	-2,433222667
418	-0,01	-0,01	-0,01	-0,01	-0,1	1831,963936

❯ Means „drop passenger“

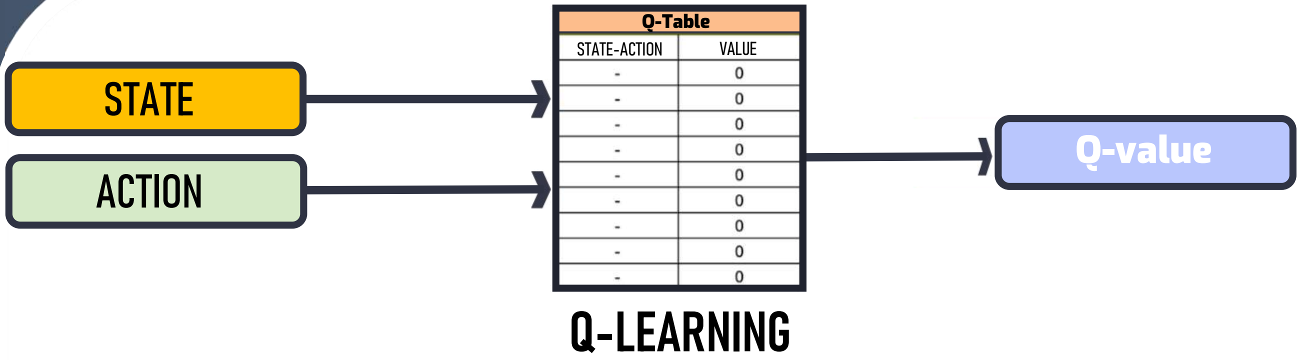
```
Choose next action: 0(South), 1(North),2(East), 3(West), 4(Pickup), 5(Dropoff)(type exit to end)
```

4 locations | Pick up at **blue**, drop off at **purple** | Free taxi is **yellow**, with passenger **green** | Successful drop-off +20 pts | Each timestep: -1 pt | Pick-up/drop-off penalties: -10 pts|

CURSE OF DIMENSIONALITY

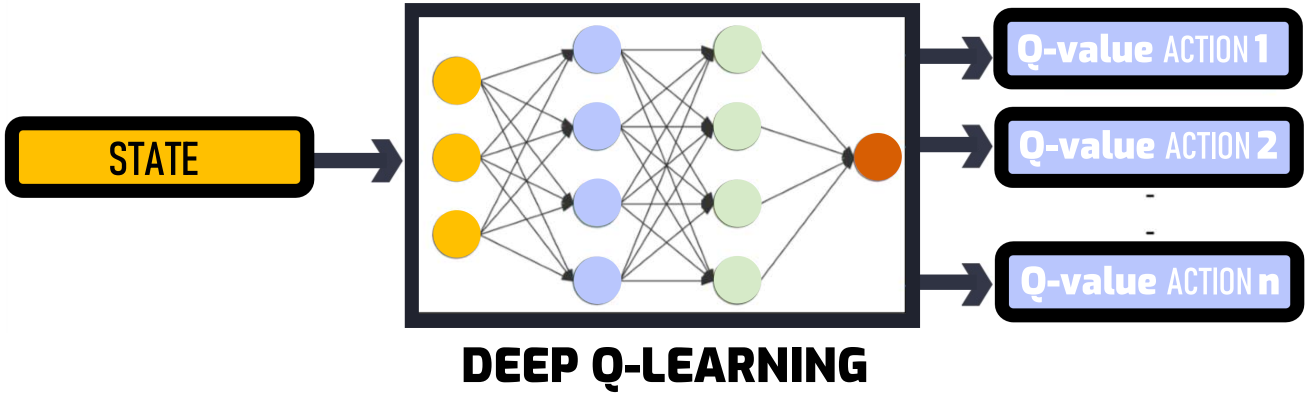
- Taxi game has 500 states, 6 possible actions = 3000 values
- In more realistic scenarios, the dimensionality explodes
  - Camera with 1M pixel * 256 color values… etc

better replace the table



Q-Table

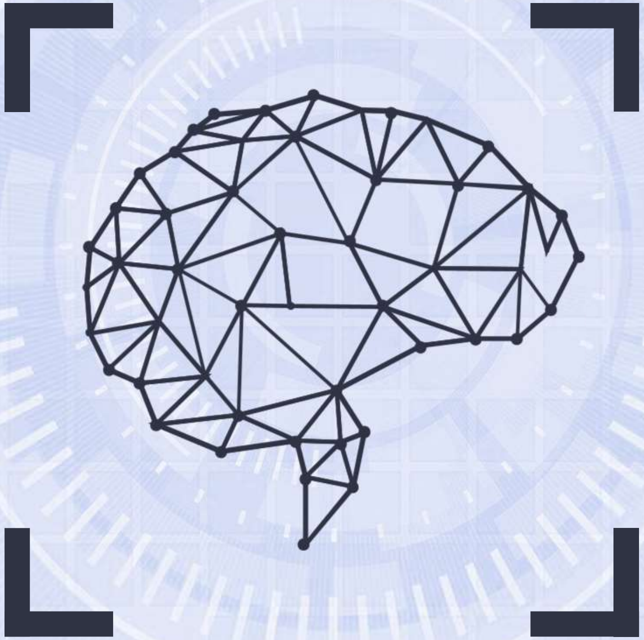
STATE-ACTION	VALUE
-	0
-	0
-	0
-	0
-	0
-	0
-	0
-	0
-	0



HOW / IN ACTION: SUMMARY II

01 Q-Learning is a relatively simple algorithm

02 Basic Q Learning stores values in a table



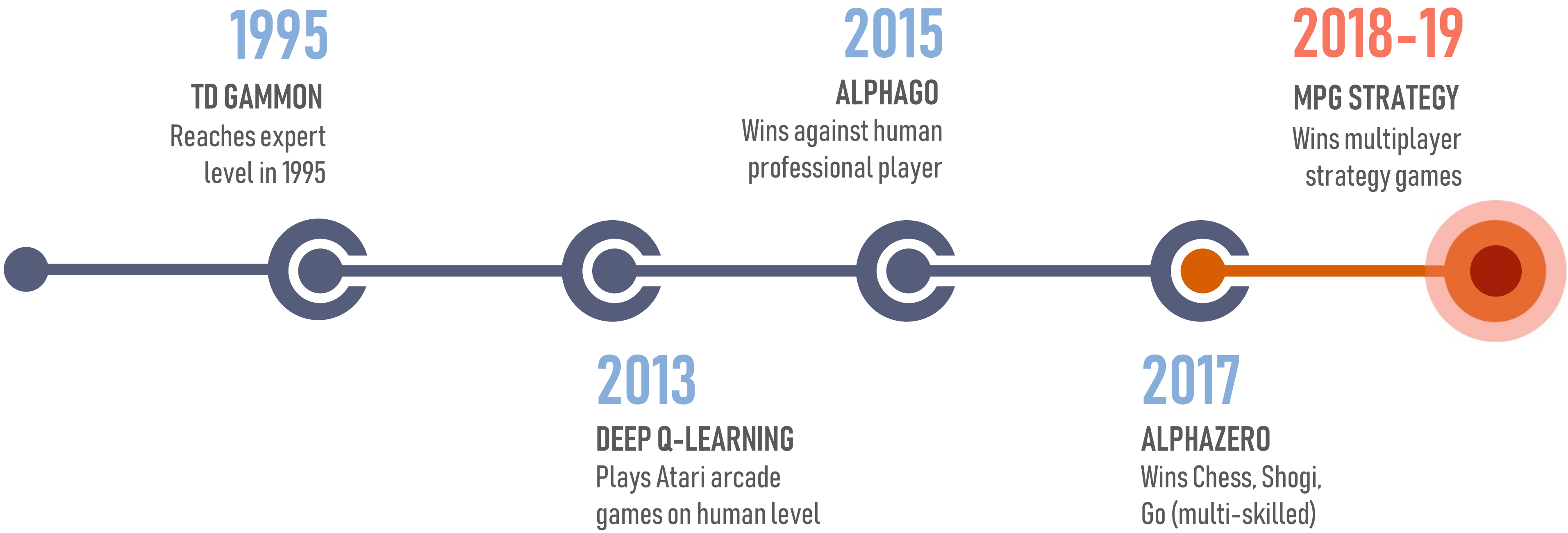
03 (Deep) NN overcomes the curse of dimensionality

04 Q-Learning is ‚model free‘ (no idea what the next state will be)

05 NOW

Methods & Milestones

MILESTONES: WINNING GAMES



1995

TD GAMMON
Reaches expert level in 1995

2013

DEEP Q-LEARNING
Plays Atari arcade games on human level

2015

ALPHAGO
Wins against human professional player

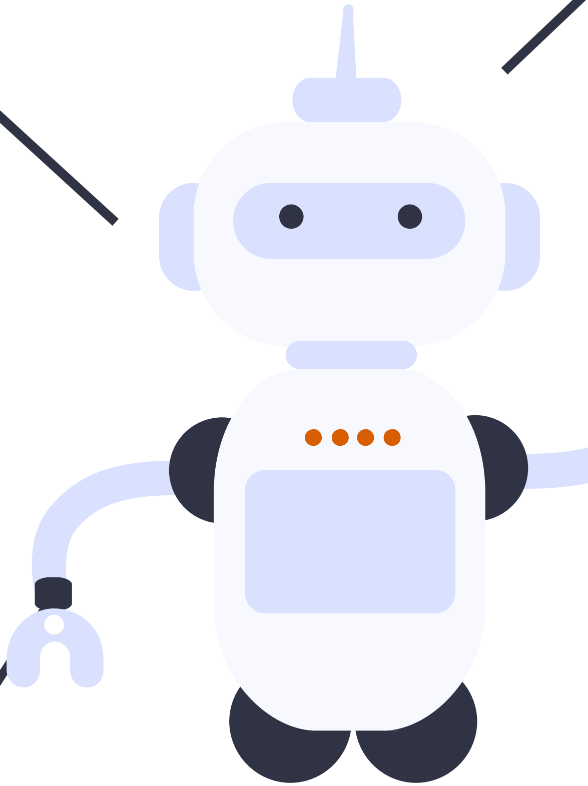
2017

ALPHAZERO
Wins Chess, Shogi, Go (multi-skilled)

2018-19

MPG STRATEGY
Wins multiplayer strategy games

„TRICKS“



PRE-PROCESS

Preprocessing, e.g. number of frames, reduces size, color

HARDWARE

Special hardware: 5000 Google TPUs, 44 CPUs (but works with less)

REPLAY

Experience replay using:

$$e_t = (s_t, a_t, r_t, s_{t+1})$$

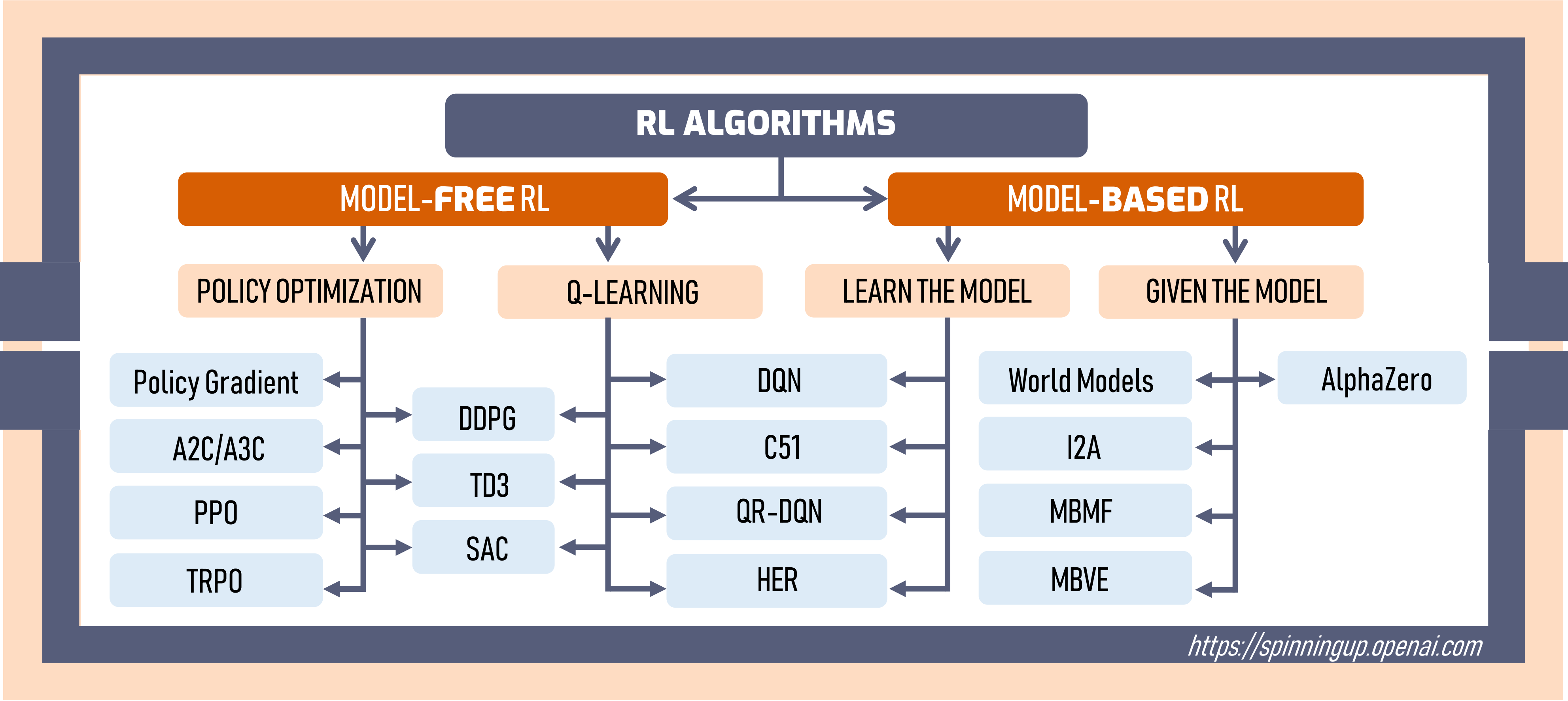
UPDATE

Batch update, adaptive learning rate

NETWORK

Divide the network into two parts

TAXONOMY OF RL METHODS



RL ALGORITHMS

MODEL-**FREE** RL

MODEL-**BASED** RL

POLICY OPTIMIZATION

Q-LEARNING

LEARN THE MODEL

GIVEN THE MODEL

Policy Gradient

A2C/A3C

PPO

TRPO

DDPG

TD3

SAC

DQN

C51

QR-DQN

HER

World Models

I2A

MBMF

MBVE

AlphaZero

https://spinningup.openai.com

FUTURE OUTLOOK

06 NEXT

Optimization of Molecules via Deep Reinforcement Learning

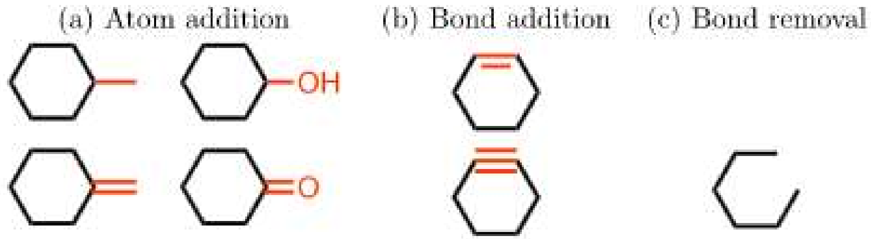


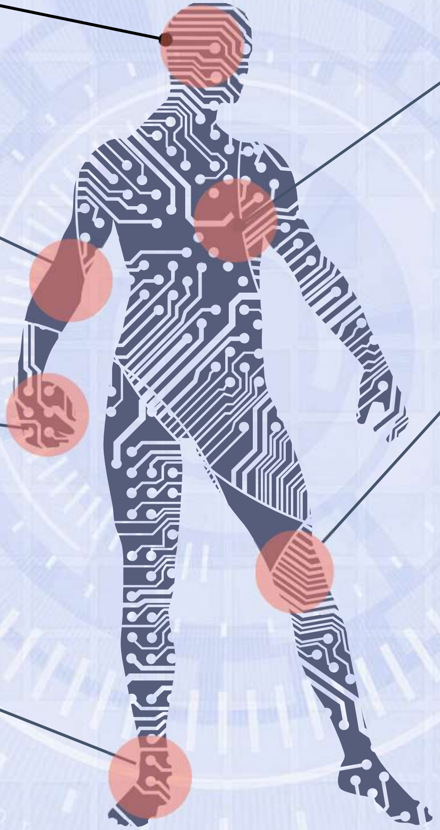
Figure 1. Valid actions on the state of cyclohexane. Modifications are shown in red. Invalid bond additions which violate the heuristics explained in Section 2.1 are not shown.

https://www.nature.com/articles/s41598-019-47148-x

GYM UPDATE FROM 2013 TO 2019 & BEYOND

https://clvrai.github.io/furniture/

KEY TAKE HOME CONCEPTS



Reinforcement Learning has a **long history**

Markovian Decision Process

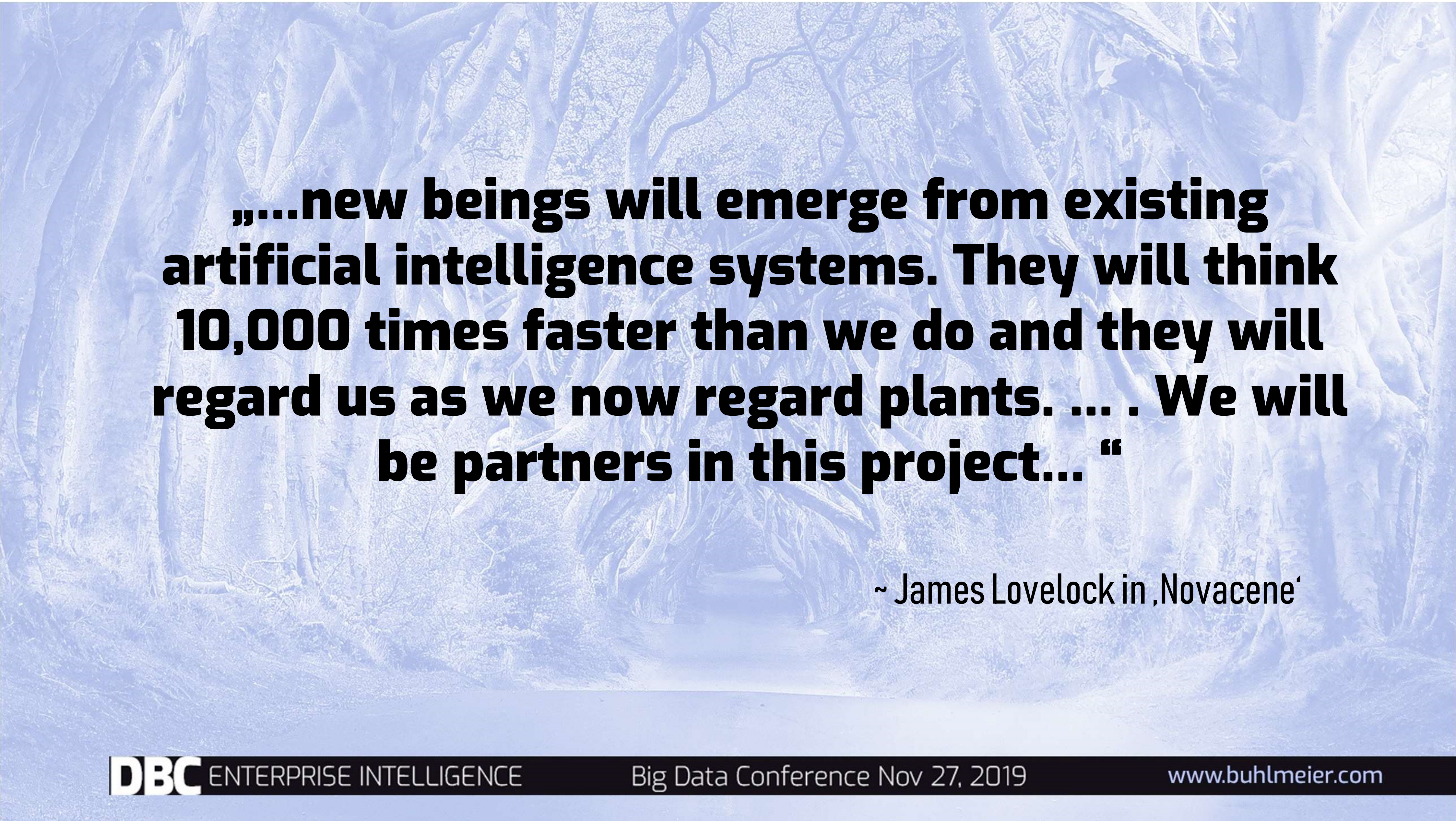
Q-Learning:
exploration vs exploitation

Deep Learning

Future Applications:
what to expect

New results expected

- in all areas that can be simulated (can do many trials)
- from new combinations of the many approaches



„…new beings will emerge from existing artificial intelligence systems. They will think 10,000 times faster than we do and they will regard us as we now regard plants. … . We will be partners in this project… “

~ James Lovelock in ‚Novacene‘

PRODUCTION & DESIGN CREDITS

- MARIA JC MONTEIRO: slideshow concept, design, layout, production
- ALLPPT.COM: free powerpoint templates
- WIKIPEDIA.ORG: Ivan Pavlov & Richard Ernest Bellman images
- GOOGLE IMAGES: „Pitfall“game image, usage rights „labelled for reuse with modification“
- PIXABAY.COM: „free for commercial use, no attribution required“ images
  - Thanks to (in order of image first appearance) Svajoklis1 | Svajunas Kraujalis, Pexels, PourquoiPas, TheDigitalArtist | Pete Linforth, CreativeMagic | Magic Creative , Free-Photos, B_Me | Brian Merrill

bigdata

BIG DATA CONFERENCE

26-28 NOVEMBER 2019, VILNIUS

BREAKTHROUGHS & FUTURE OF (DEEP) REINFORCEMENT LEARNING

Dr. ANDREAS BÜHLMEIER

DBC ENTERPRISE INTELLIGENCE

WWW.BUHLMEIER.COM

Labai ačiū!

HAPPY TO ANSWER YOUR QUESTIONS NOW ☺

dbc ENTERPRISE INTELLIGENCE