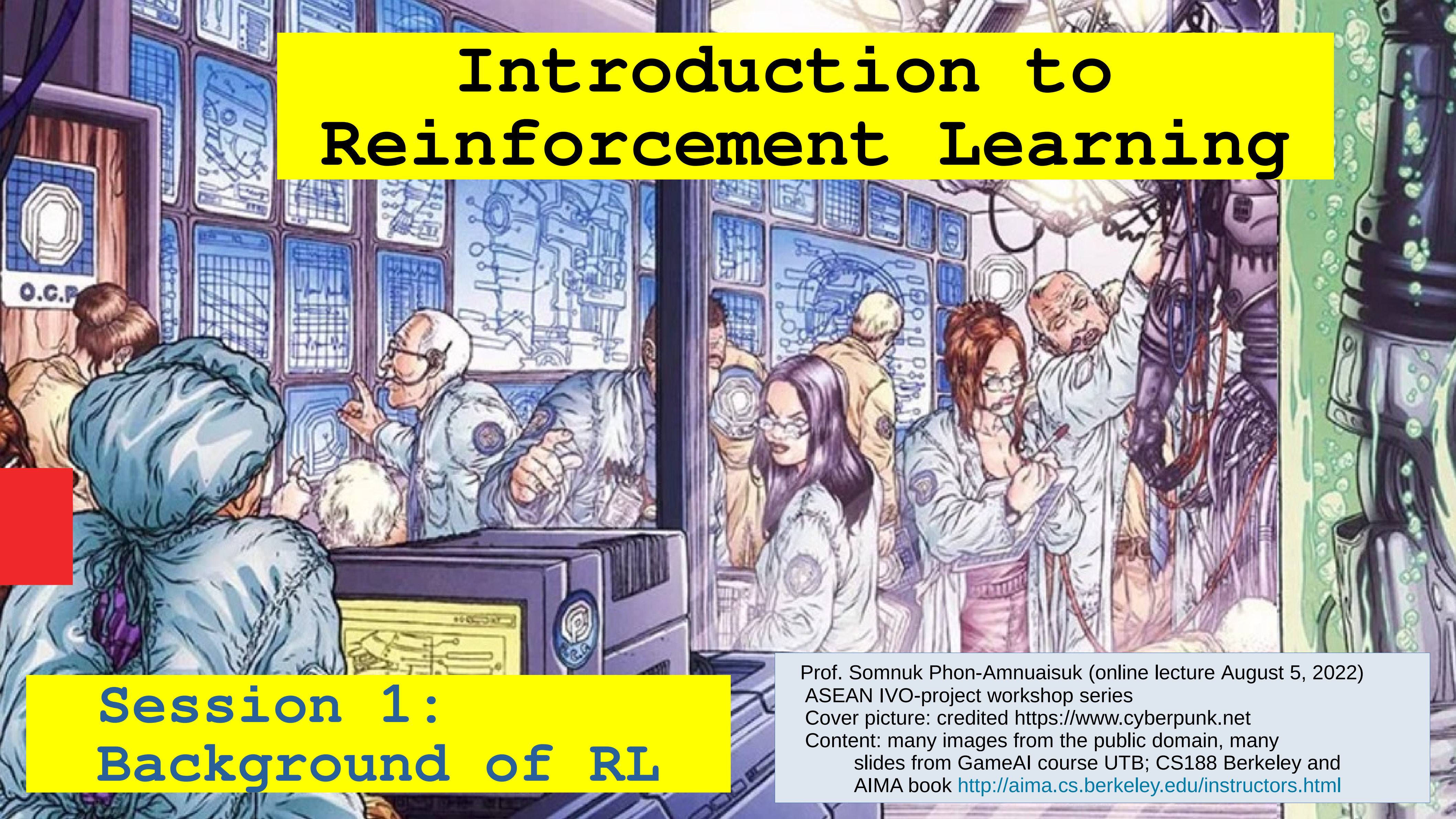
Introduction to Reinforcement Learning



Session 1: Background of RL

Prof. Somnuk Phon-Amnuaisuk (online lecture August 5, 2022)
ASEAN IVO-project workshop series
Cover picture: credited https://www.cyberpunk.net
Content: many images from the public domain, many slides from GameAI course UTB; CS188 Berkeley and AIMA book http://aima.cs.berkeley.edu/instructors.html


meme
A meme (/ˈ mi: m/ meem), a neologism coined by Richard Dawkins, is "an idea, behavior, or style that spreads from person to person within a culture". A meme acts as a unit for carrying cultural ideas, symbols, or practices that can be transmitted from one mind to another through writing, speech, gestures, rituals, or other imitable phenomena with a mimicked theme.

Outline

- Briefly about this ASEAN-IVO workshop series
  - UG students who have not done ANN and RL but attempting FYPs with these components
- What is artificial intelligence?
- A brief history and development of AI.
- An introduction to Reinforcement Learning.
- Development of enabling technology.
- Technology trends.

What is artificial intelligence?

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

What is artificial intelligence?

A Proposal for the

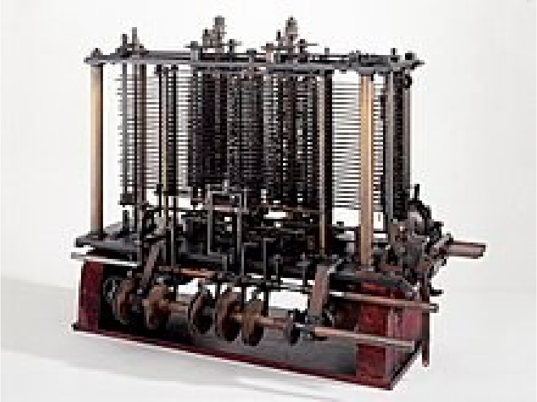
DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

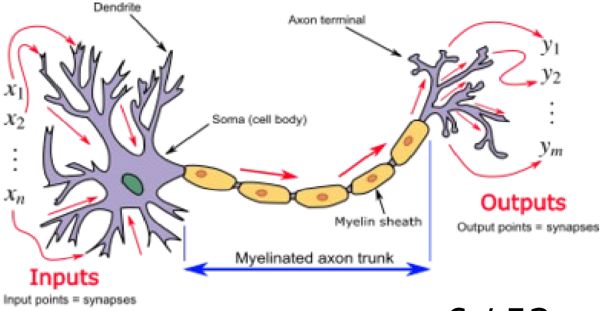
June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

Challenges in AI

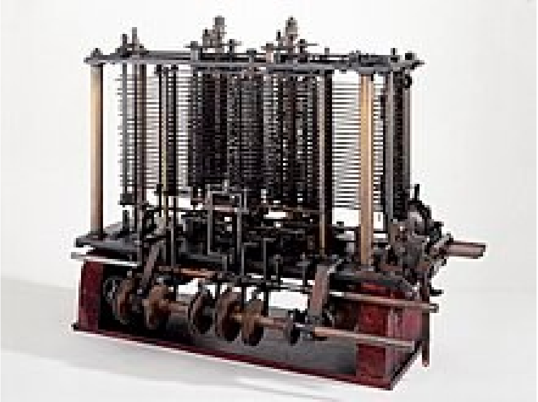
- Crunching numbers
- Infer, deduce
- Muscle skills
- Musical intelligence
- Emotional intelligence
- Spatial intelligence
- Linguistic intelligence
- etc.

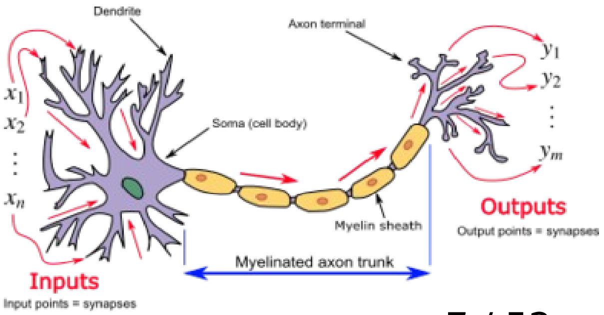


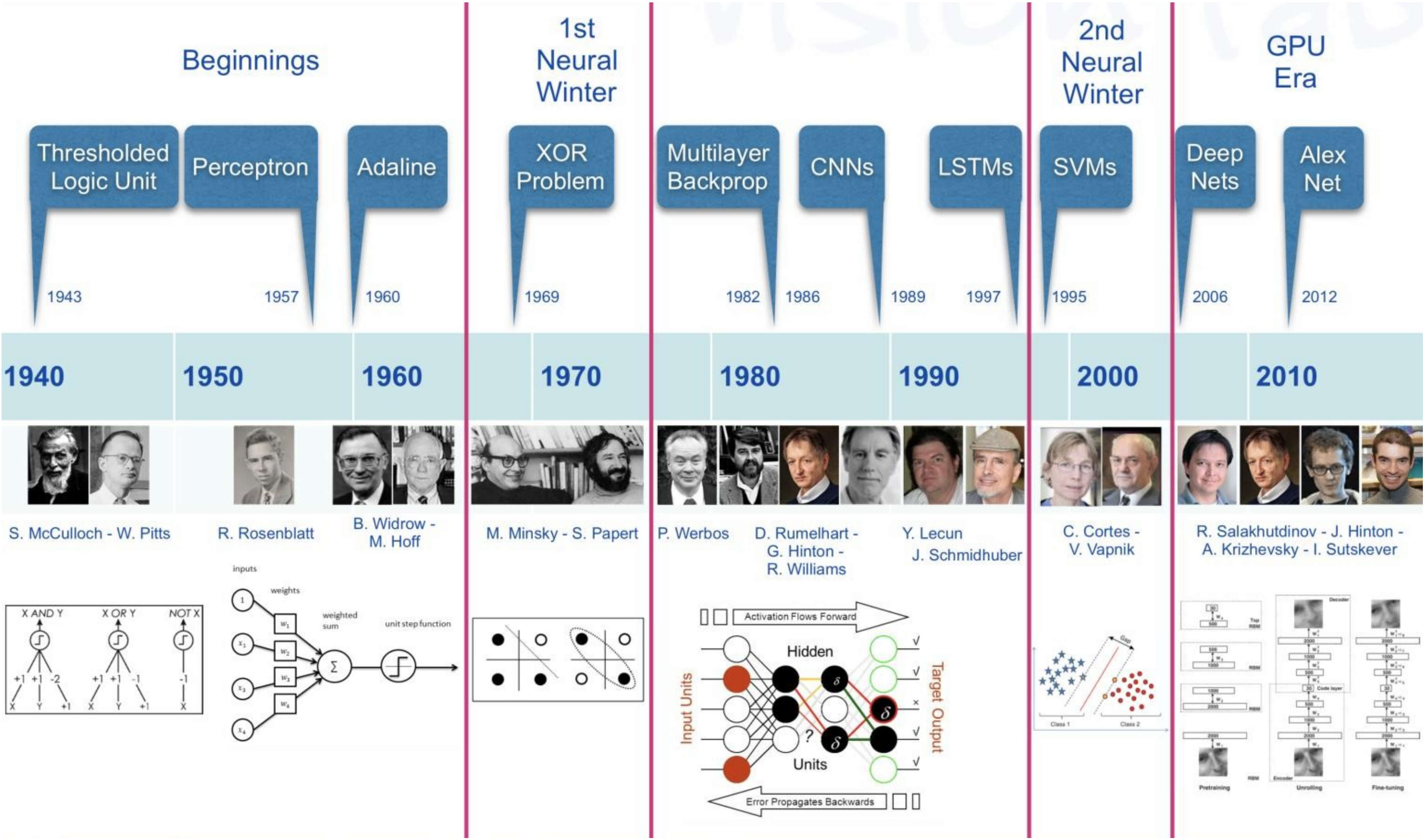


Challenges in AI

- Knowledge representation (KR)
- Compute – quantitative reasoning
- Qualitative reasoning
- Uncertainty
- Changes in beliefs
- Common sense
- Incremental learning

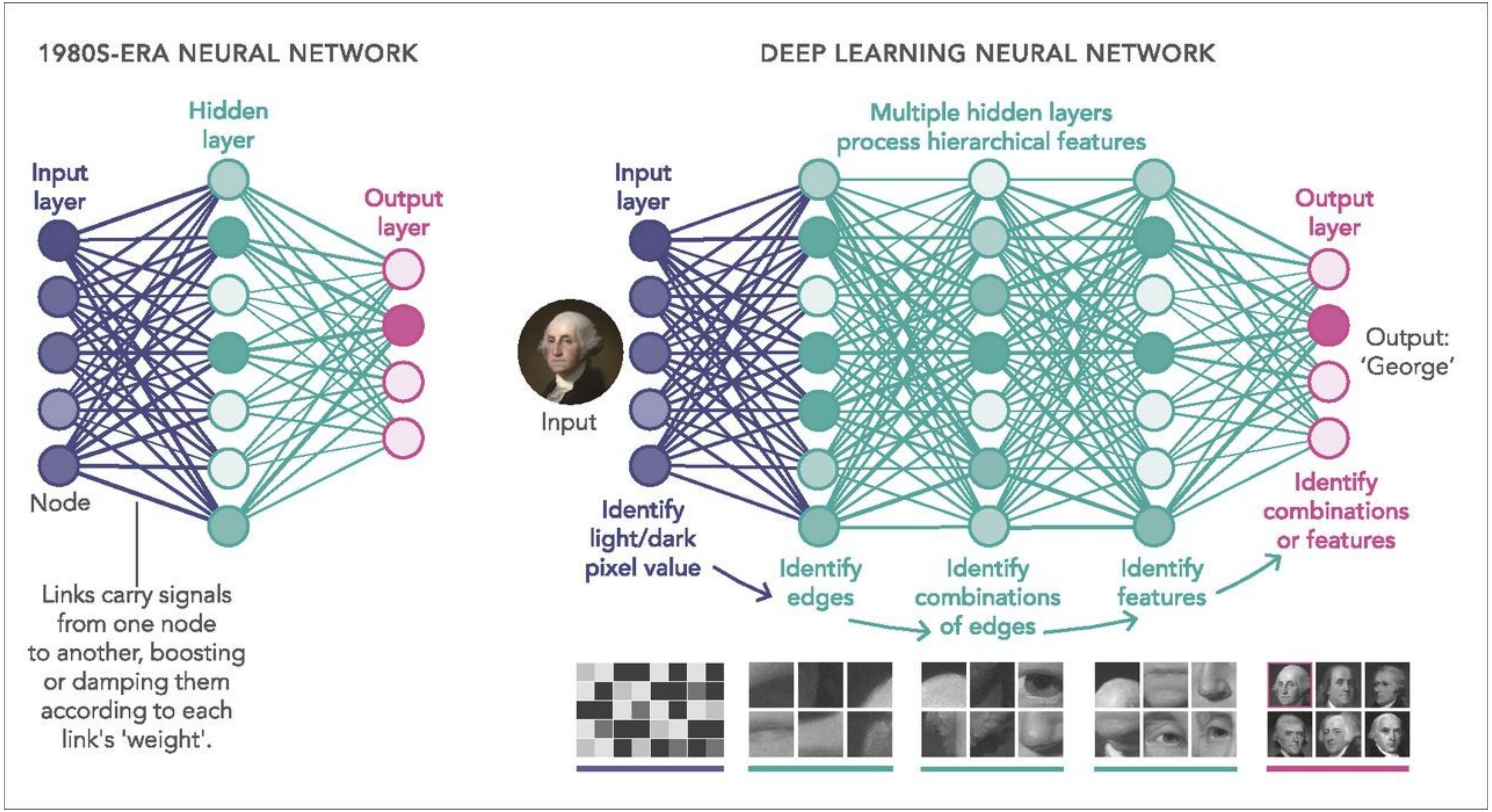




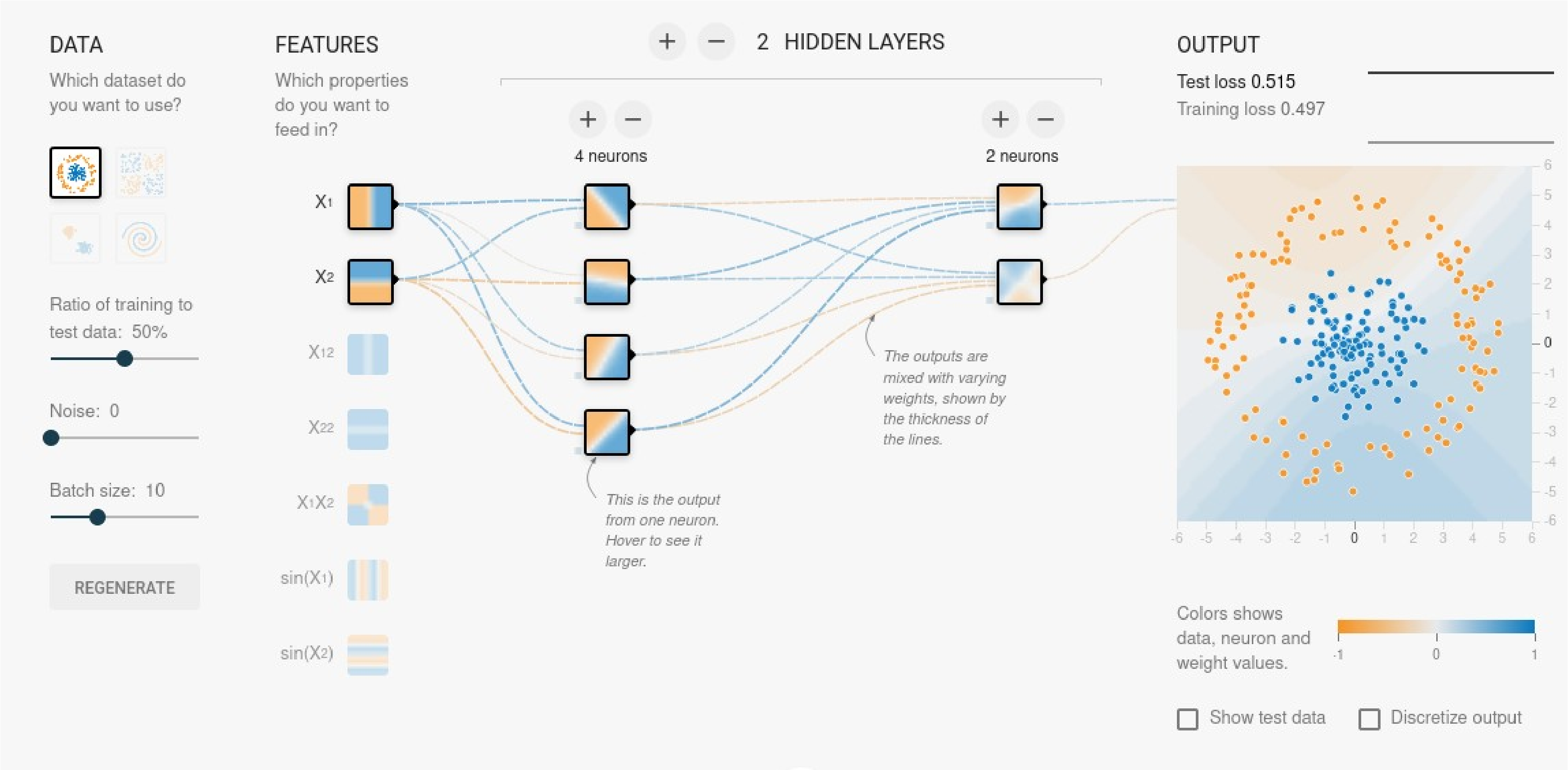


credit: pinterest

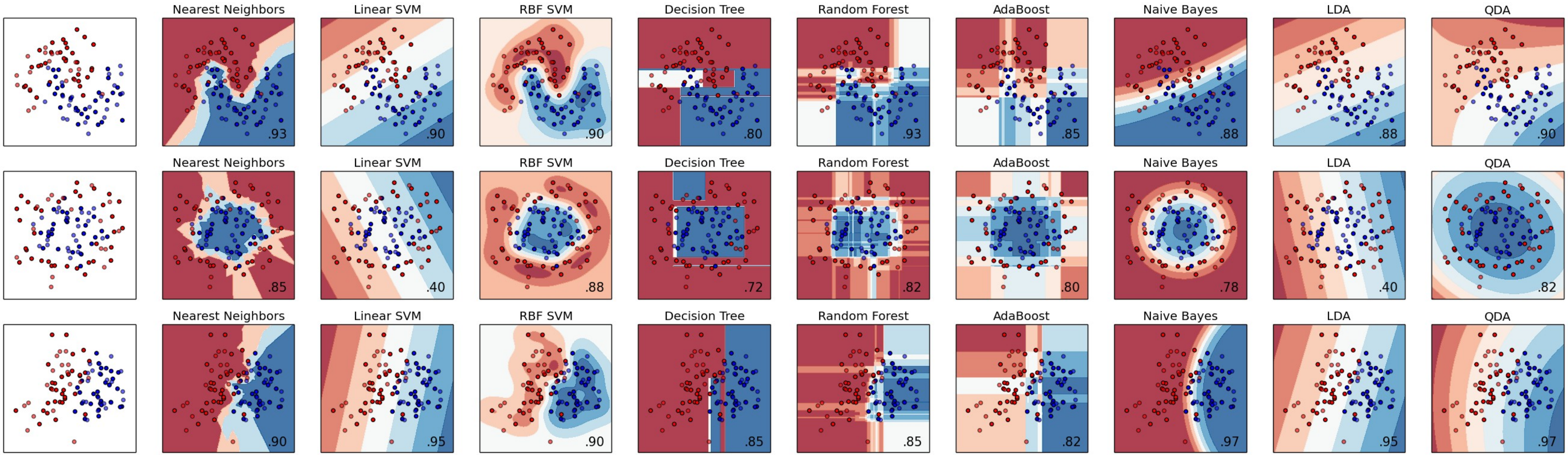
Deep Learning



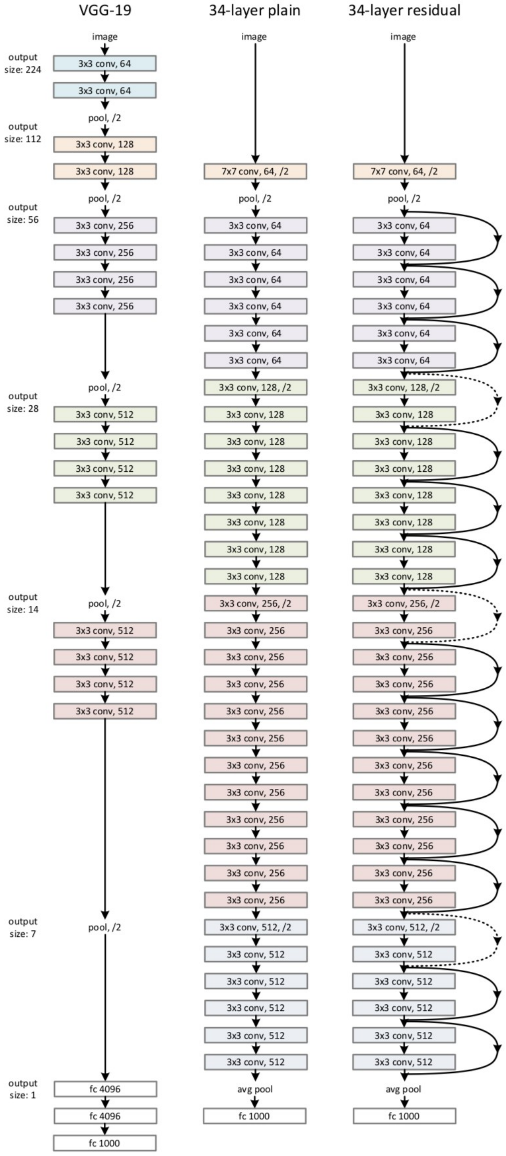
credit: Lucy Reading-Ikkanda (artist). https://www.pnas.org/doi/10.1073/pnas.1821594116 https://playground.tensorflow.org/



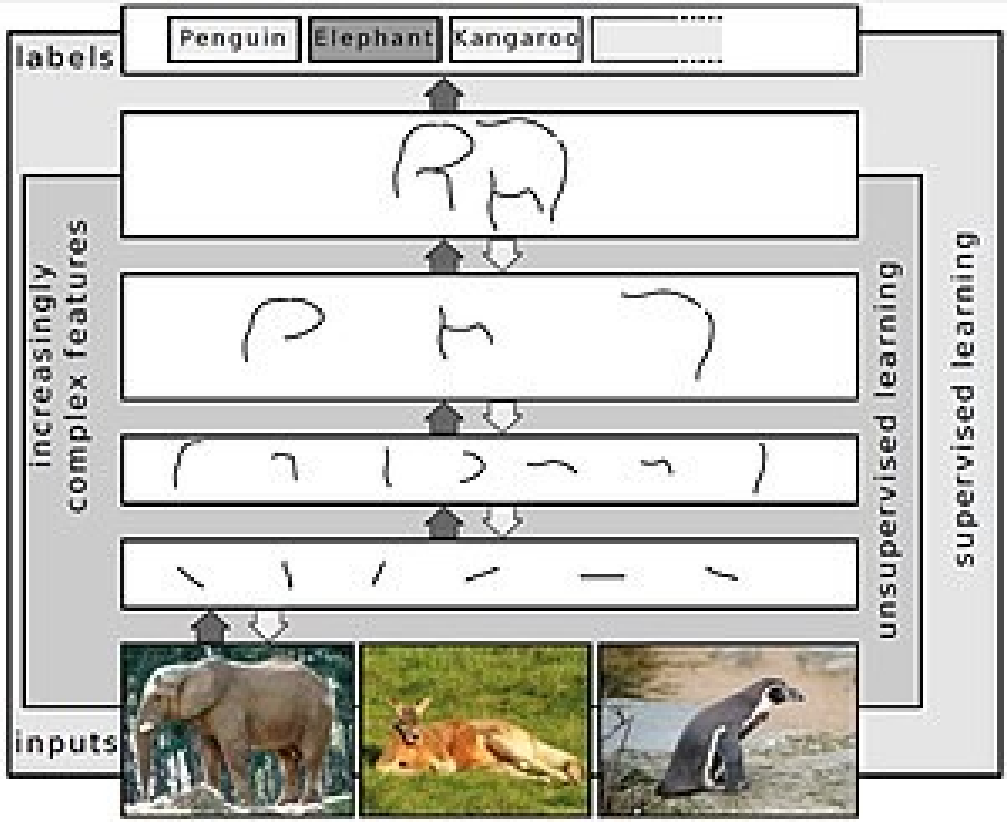
Decision Boundary

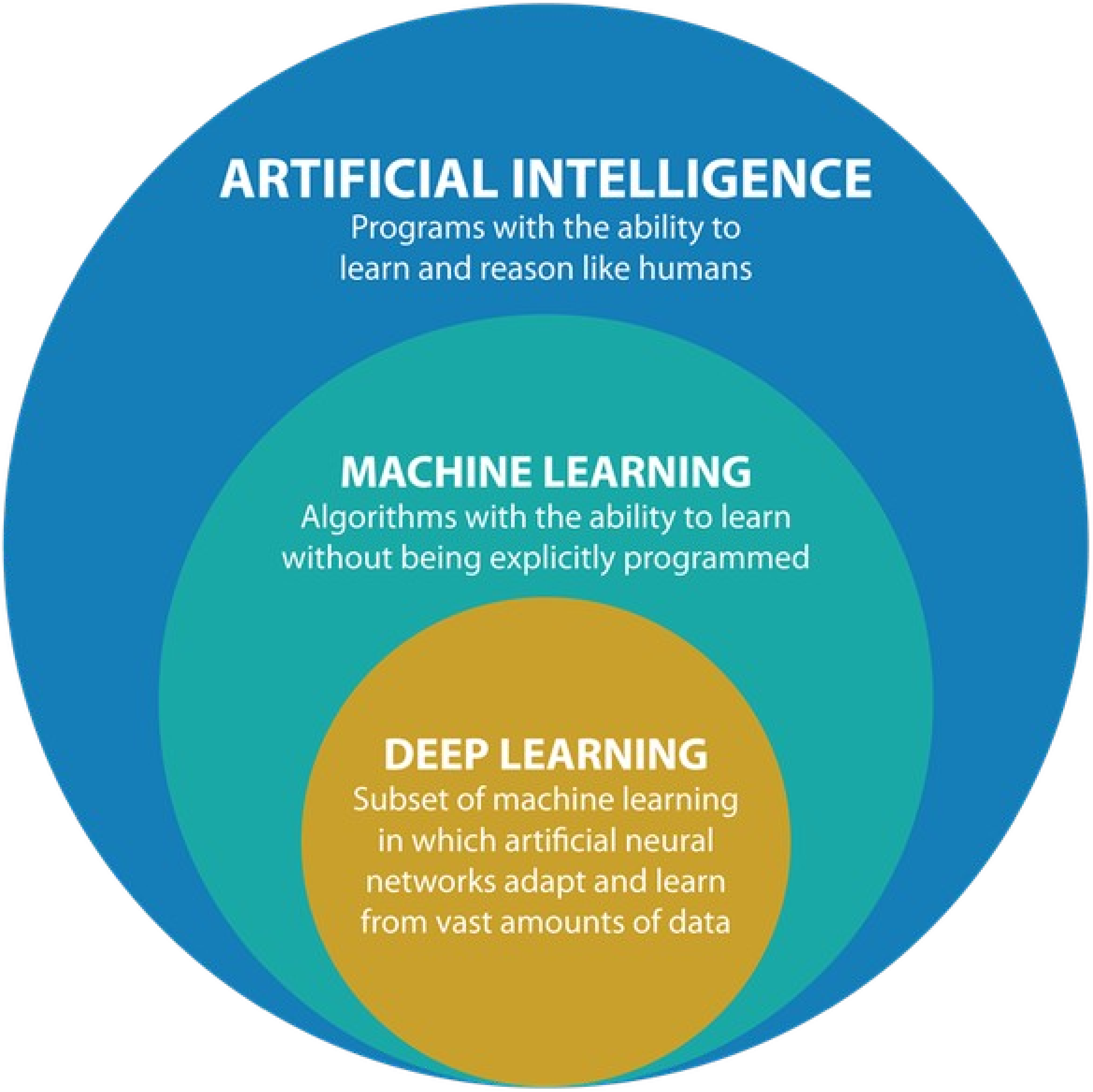


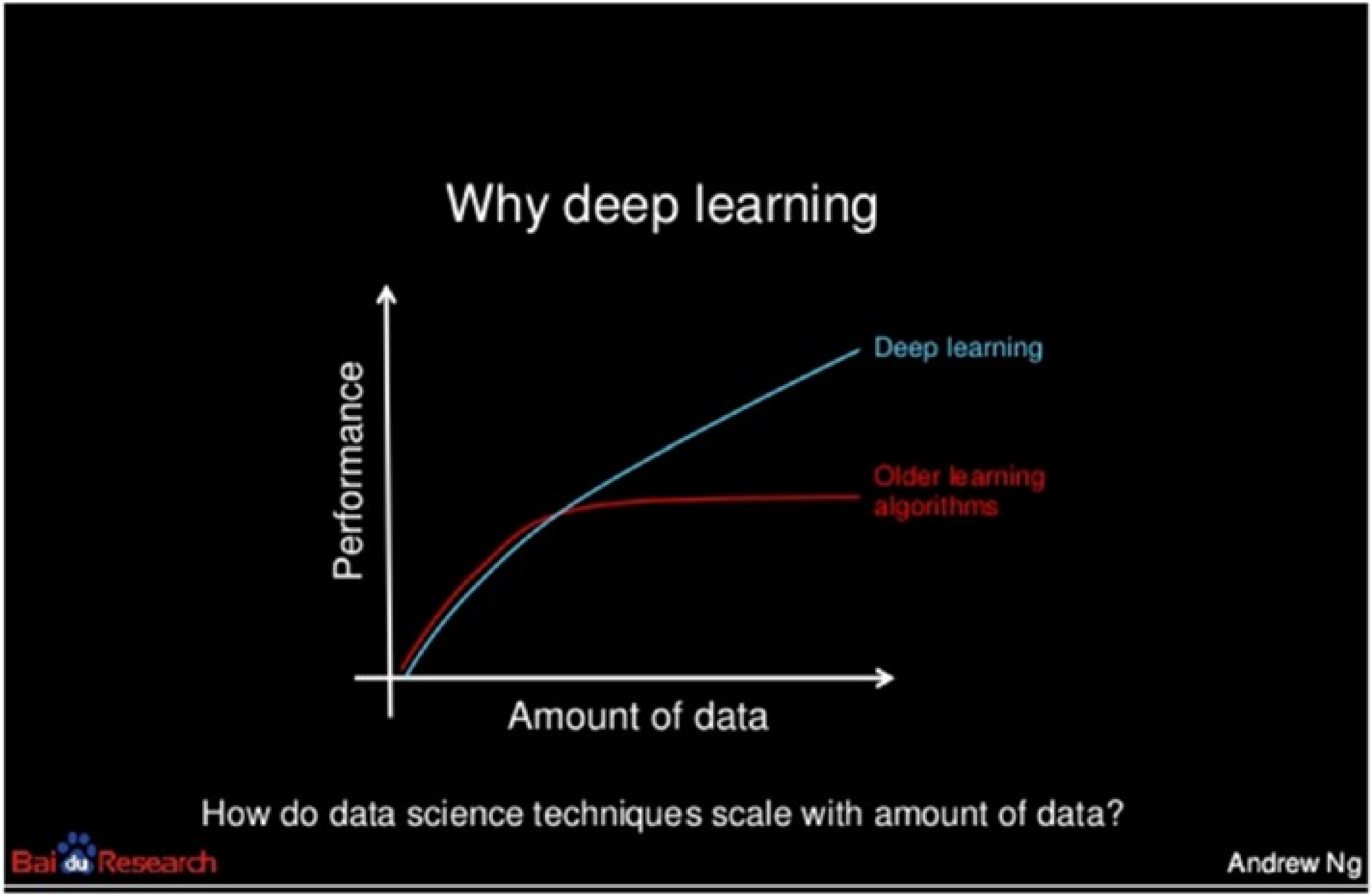
Credit: scikitlearn



Deep Learning

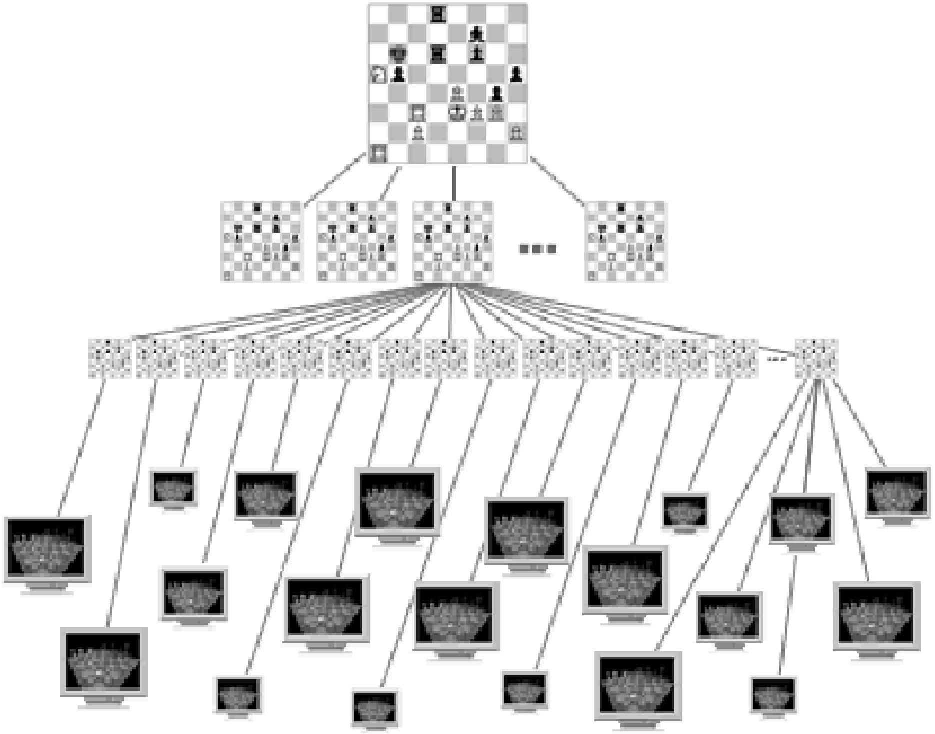



ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data



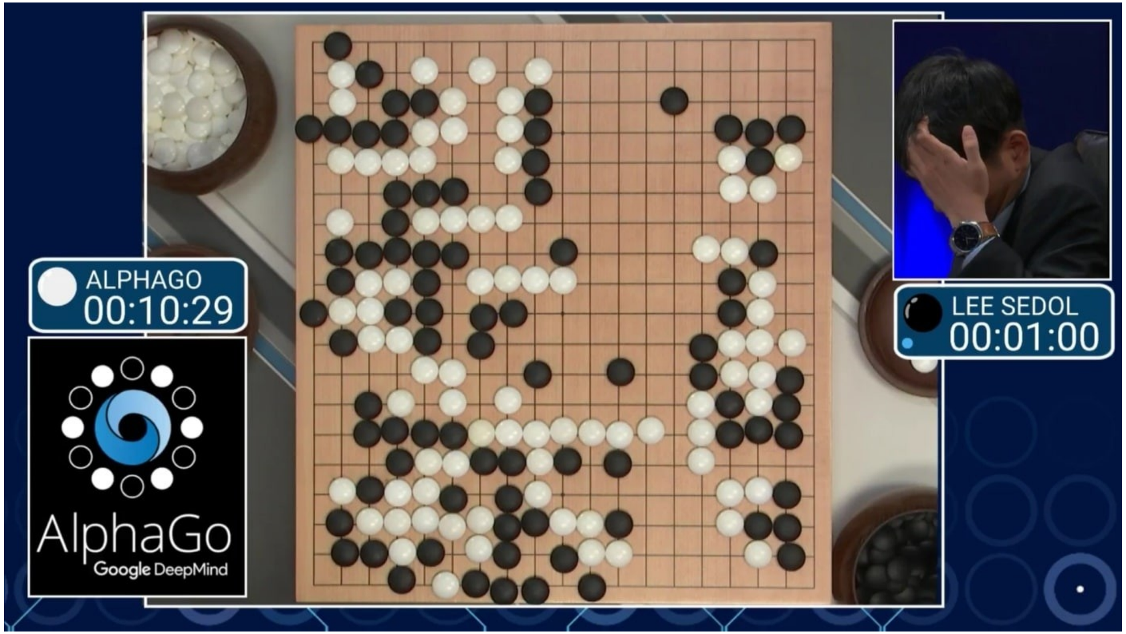
Credit: Baidu research

Notable AI Applications



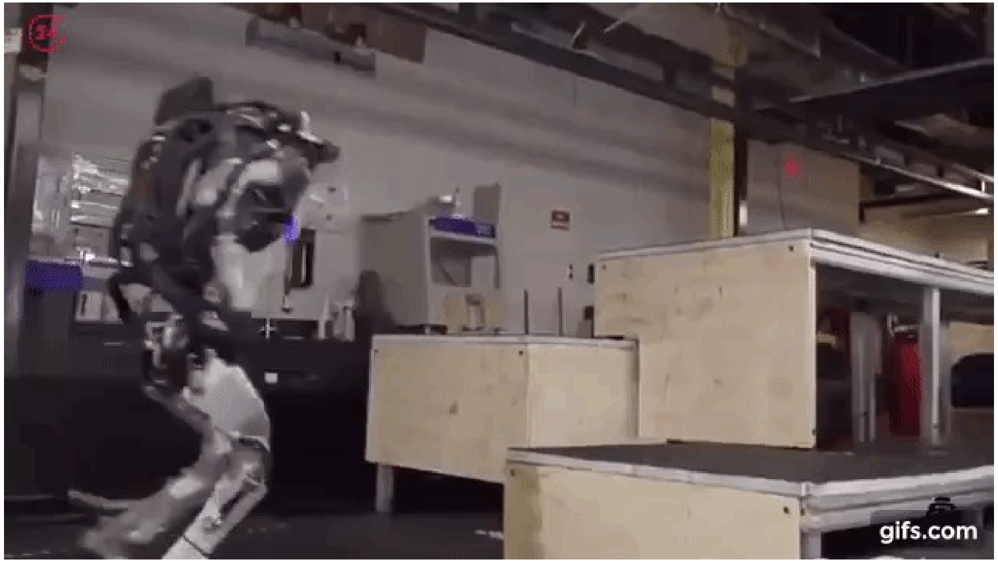
https://www.researchgate.net/figure/Distributed-chess-tree-search_fig1_224056308

Notable AI Applications



Game	Board size	State space	Game tree size
Go	19 x 19	10^{172}	10^{360}
Chess	8 x 8	10^{50}	10^{123}
Checkers	8 x 8	10^{18}	10^{54}

Notable AI Applications



Notable AI Applications

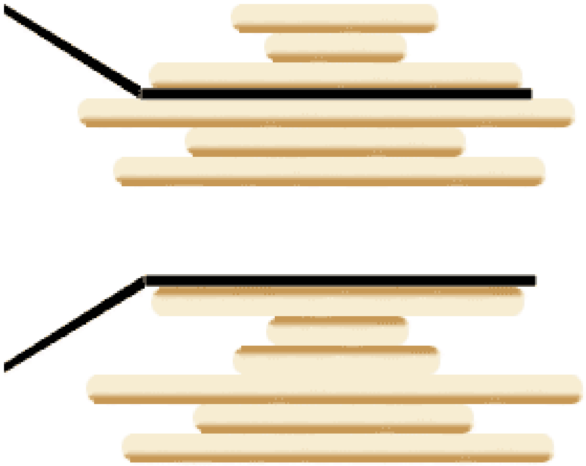
Example: Pancake sorting problem

- Pancake sorting is the colloquial term for the mathematical problem of sorting a disordered stack of pancakes in order of size when a spatula can be inserted at any point in the stack and used to flip all pancakes above it.

- In 1979, Bill Gates and Christos Papadimitriou gave an upper bound of (5n+5)/3. This was improved, thirty years later, to 18n/11 by a team of researchers at the University of Texas at Dallas.

- The minimum number of flips required to sort any stack of n pancakes has been shown to lie between 15/14n and 18/11n (approximately 1.07n and 1.64n,) but the exact value is not known.

(Adapt from Wikipedia)



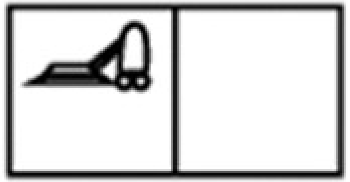
State Space Search

- State space search is a process used in the field of computer science, in which successive configurations or states of an instance are considered, with the intention of finding a goal state with a desired property.
- The typical state space graph is much too large to generate and store in memory. Hence, nodes are generated as they are explored, and typically discarded thereafter.
- A state space is formally represented as a tuple (S, A, B(.,.), eval(.), ,
- S is a set of all possible states, A is a set of possible actions
- B: (s(t),a(t)) --> s(t+1), and eval: [s(t-n),...,s(t)] --> reward

(Adapt from Wikipedia)

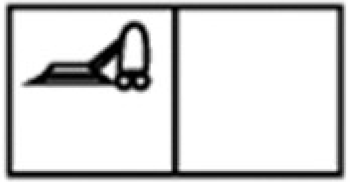
State Space Search: A Vaccum-World

- The world has 2 grids and a vacuum agent lives in this world.

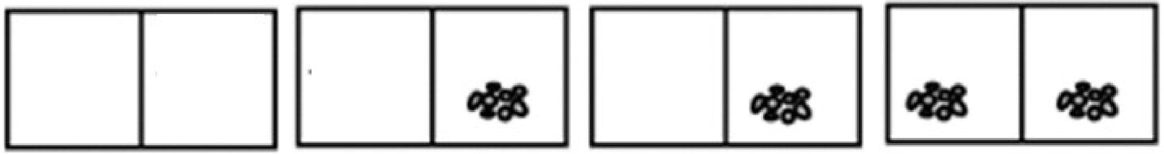


State Space Search: A Vaccum-World

- The world has 2 grids and a vacuum agent lives in this world.



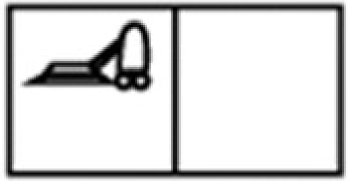
- Each grid may be clean or dirty.



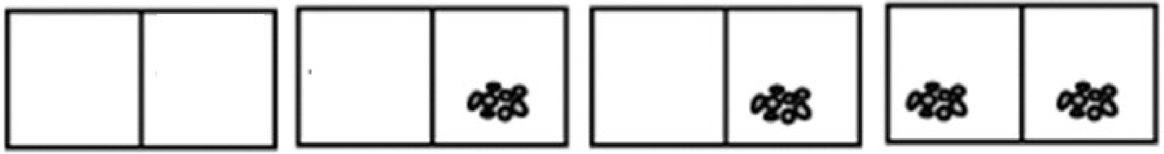
- **How many possible states are there?**

State Space Search: A Vaccum-World

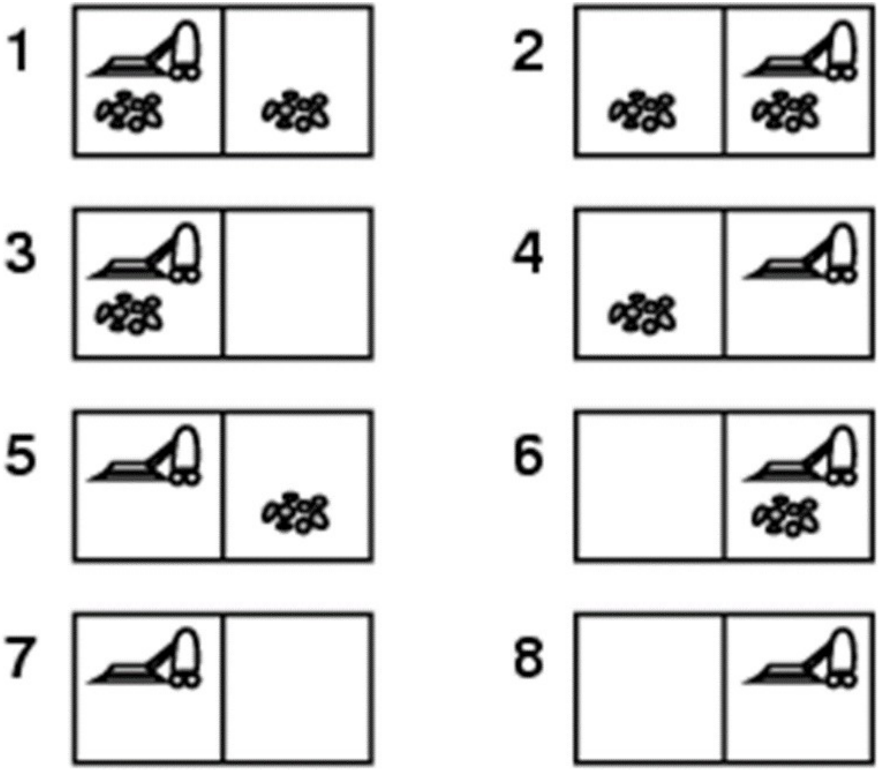
- The world has 2 grids and a vacuum agent lives in this world.



- Each grid may be clean or dirty.



- How many possible states are there?

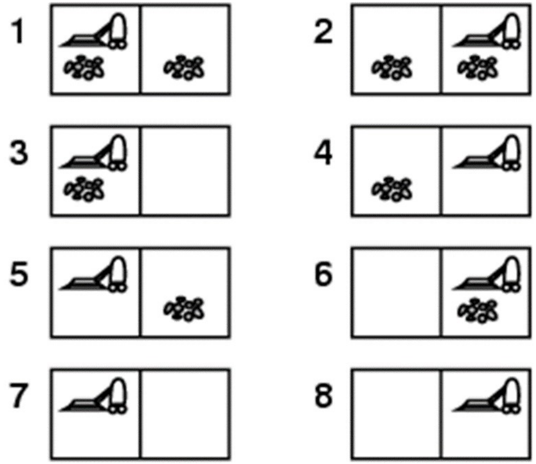


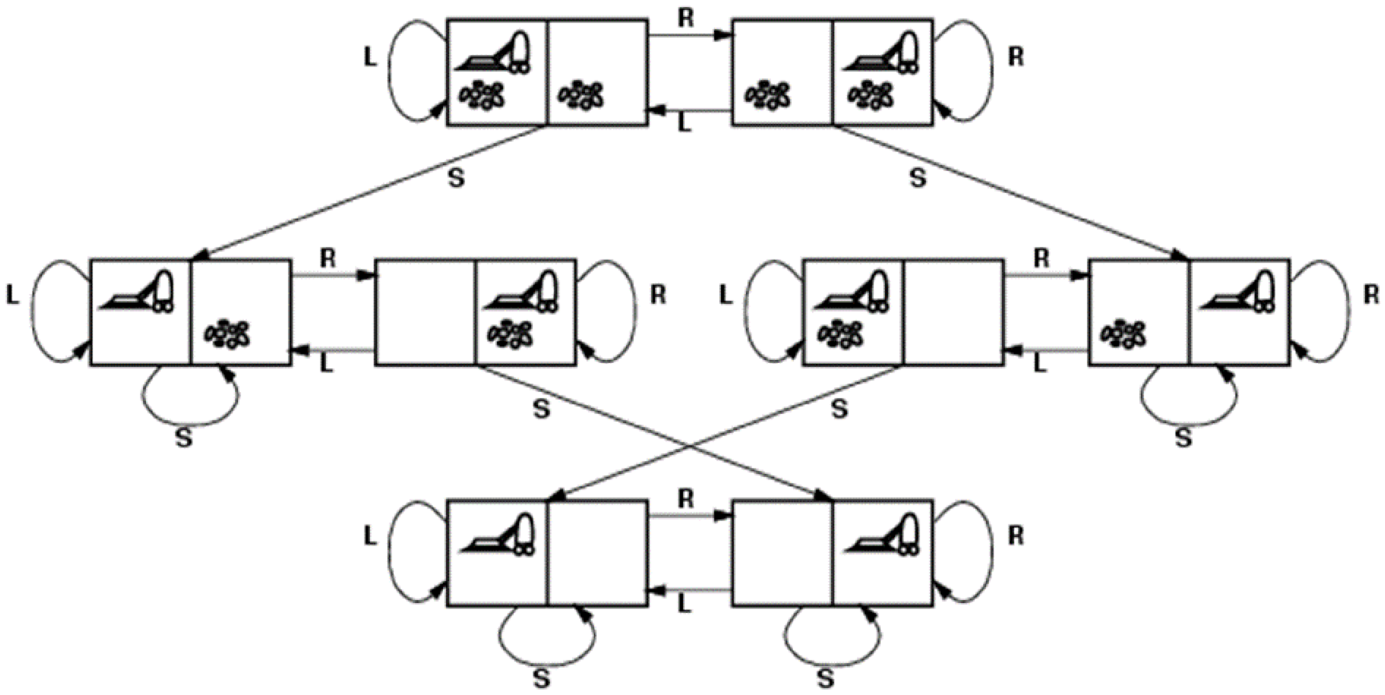
State Space Search: A Vaccum-World

S = {1,2,3,4,5,6,7,8}   A = {S, L, R}

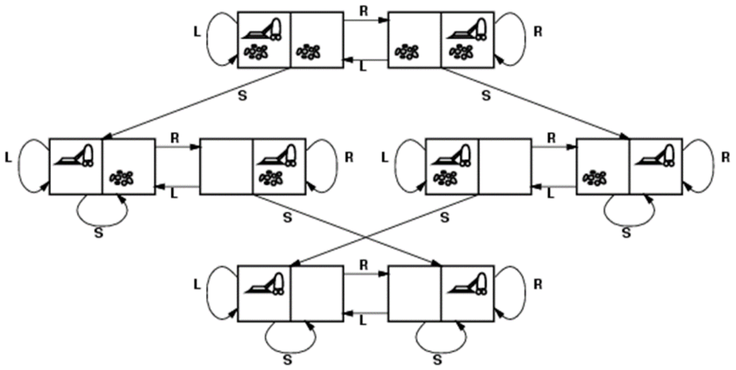
ex. B(1,R) --> 2

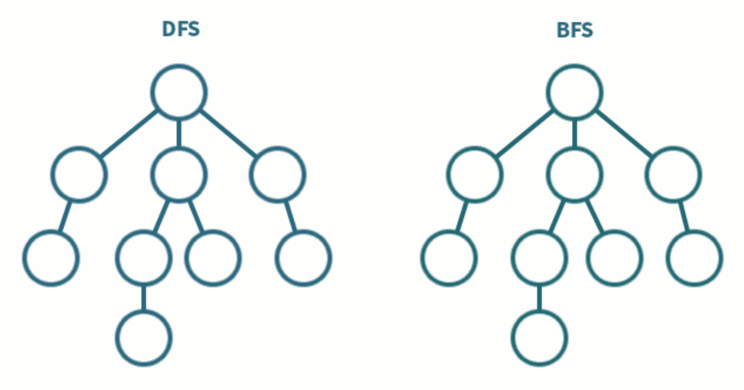
eval([1,5,6,8]) --> reward signal   eval([1,2,4,3,7]--> reward signal





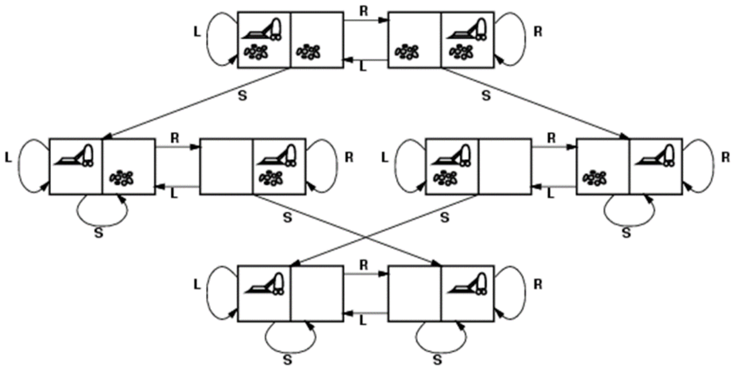
State Space Search: Branch and Bound

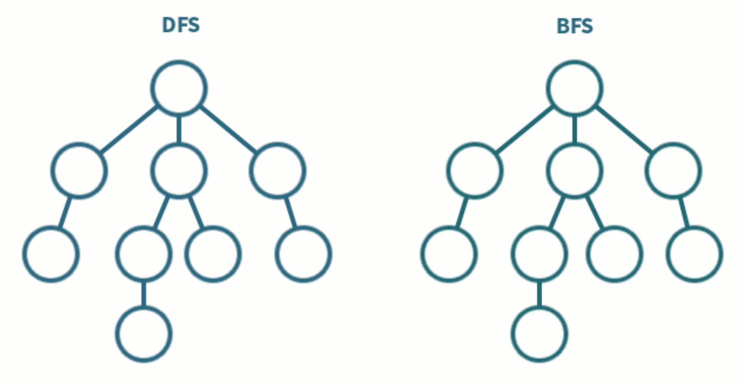




- Branching strategy- Example of the two simplest strategies are known as depth-first search and breadth-first search.
- Depth-first search (also known as branch and backtrack) moves straight down a sequence of branches until a terminal node is reached before backtracking up to the nearest junction.
- Breadth-first search, on the other hand, enumerates all the branches at one level before moving on to the next level.
- The bounding process allows us to prune out some partial solutions that cannot lead to optimal solutions. Cutting them and their descendants from the search tree improves search efficiency.

State Space Search: Branch and Bound





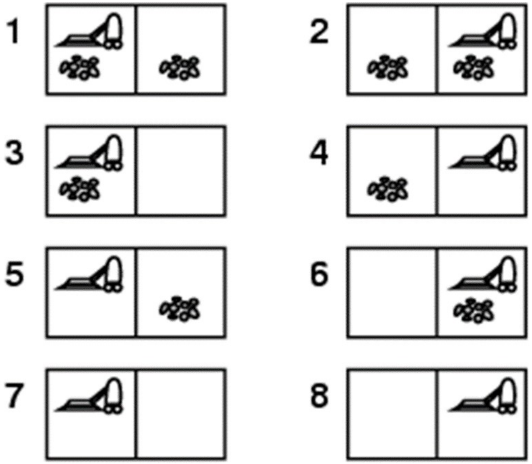
- Depth-first search – inserts branched states in front of a list
- Breadth-first search – inserts branched states to the back of a list
- Uninformed search: the tree is traversed based on the way a node is expanded and how subsequet nodes are explored
- Informed search: exploits information about the goal node's location in the form of a heuristic function
  - $f(n) = g(n) + h(n)$
  - Path cost: $g(n)$
  - Heuristic: $h(n)$

State Space Search: Lessons Learned

S = {1,2,3,4,5,6,7,8} A = {S, L, R}

ex. B(1,R) --> 2

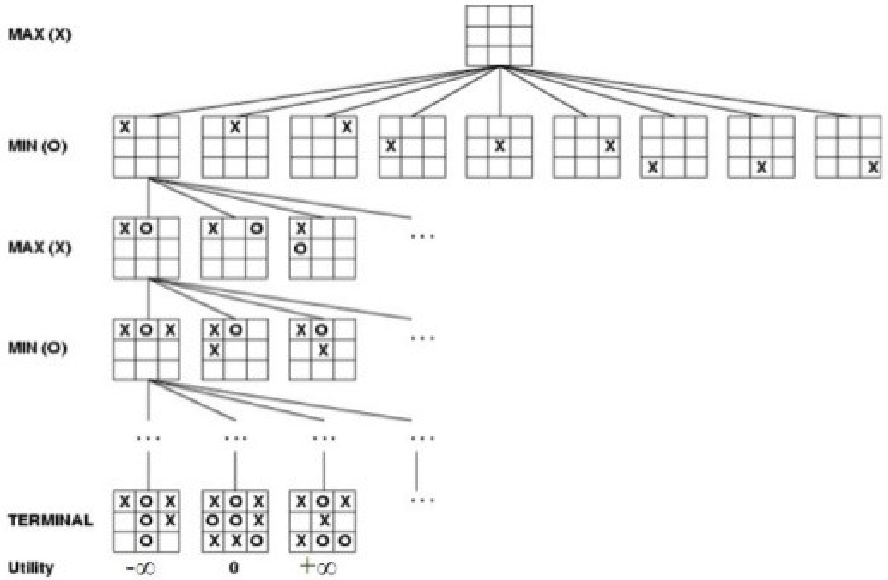
eval([1,5,6,8]) --> reward signal eval([1,2,4,3,7]--> reward signal



- When faced with the problem of finding an optimum over a finite set of alternatives.
- Hooray... we can enumerate all the alternatives and then select the best.
- However, for anything other than the smallest problems, such an approach is computationally infeasible.
- It is not a trivial task to knowledge engineer the problem solving approach.

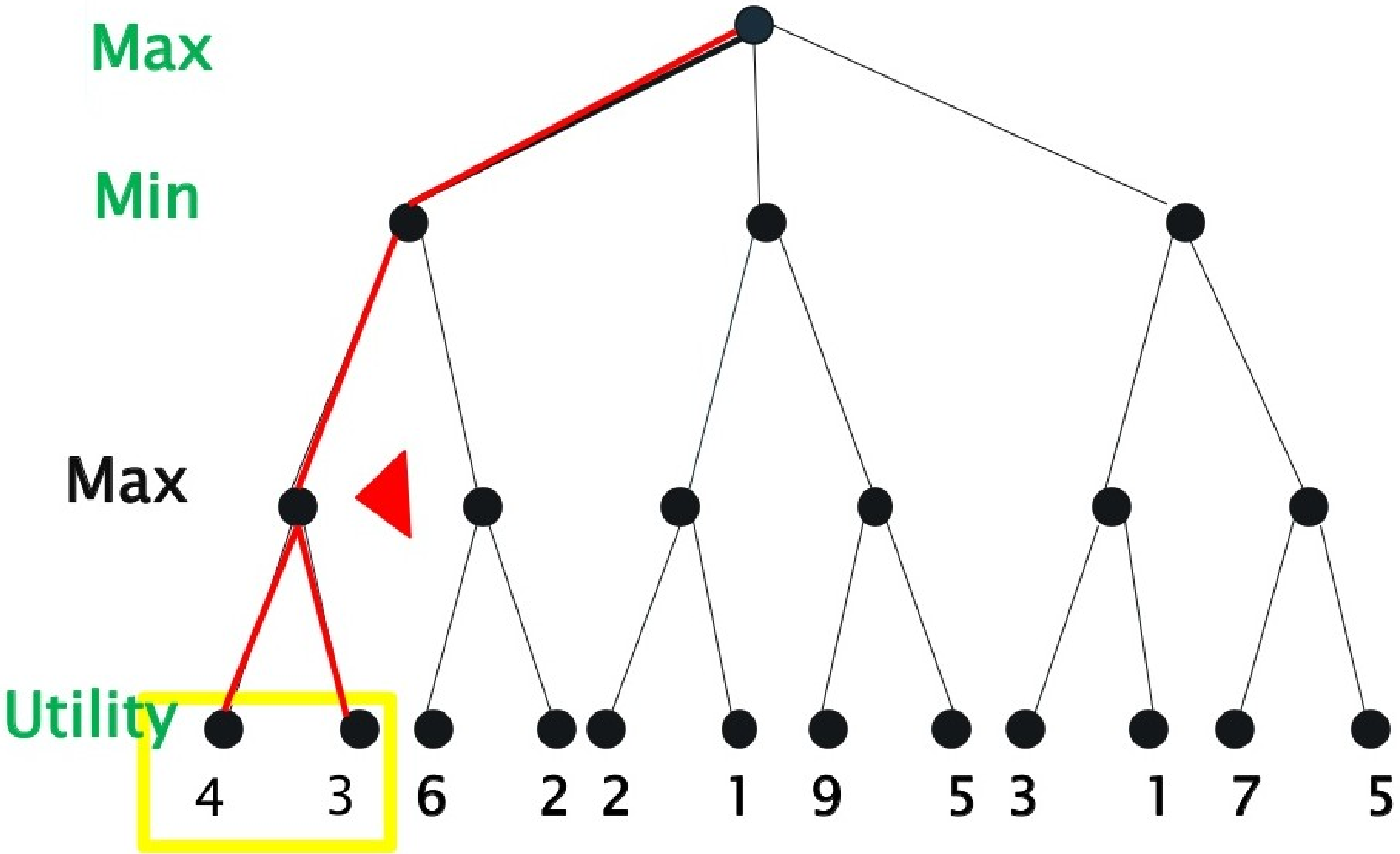
State Space Search: Adversary Search

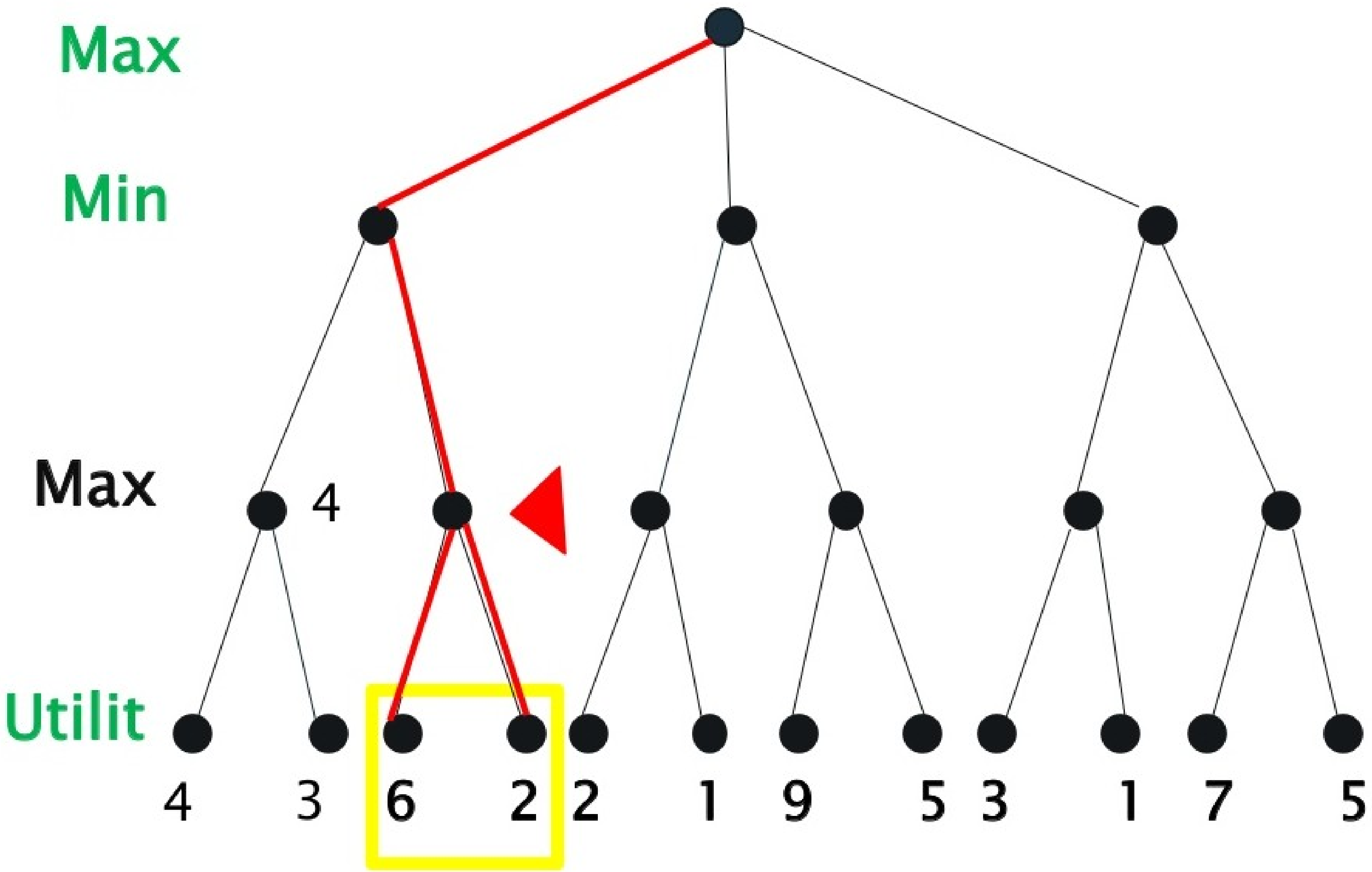
Game tree for Tic-Tac-Toe

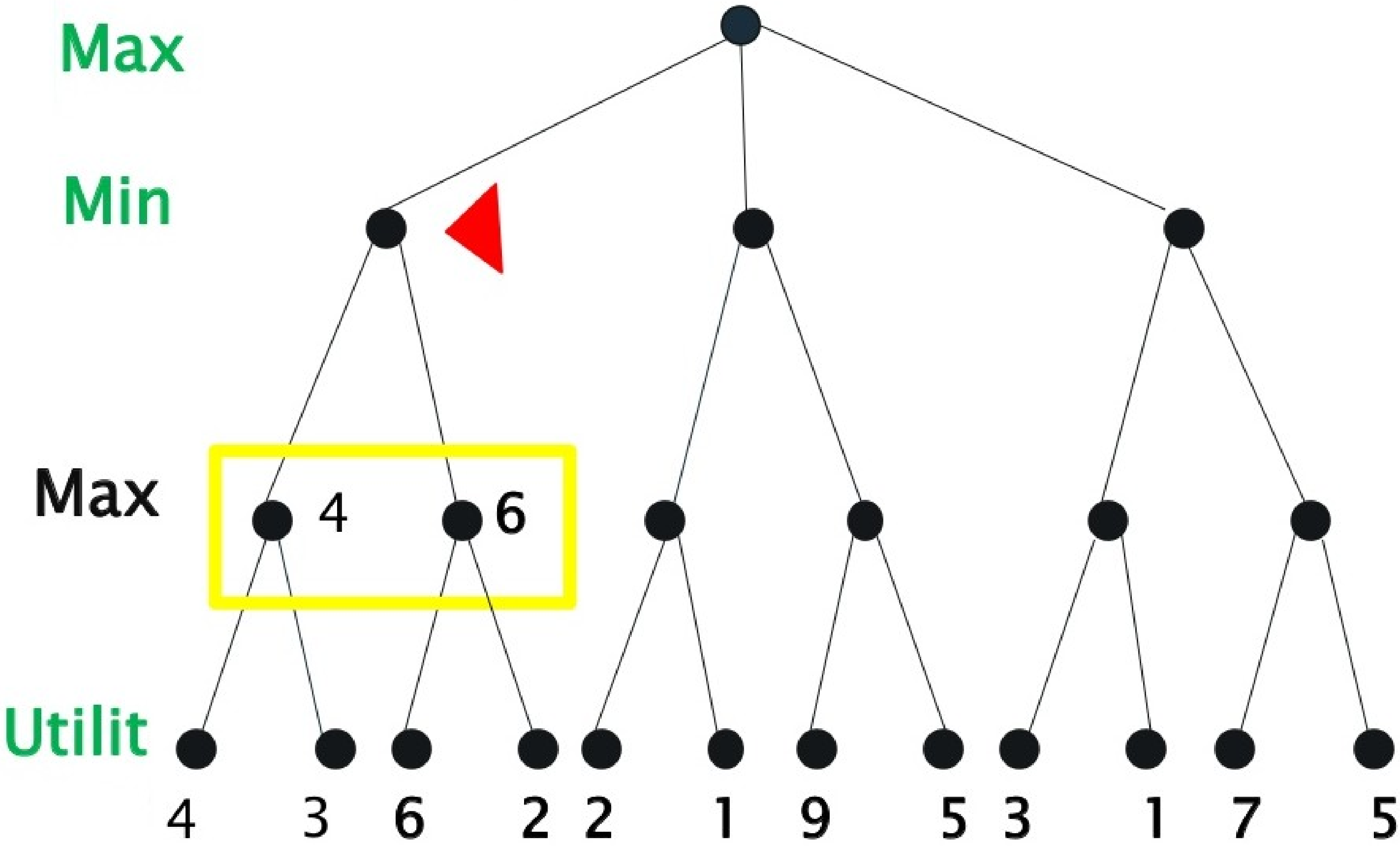


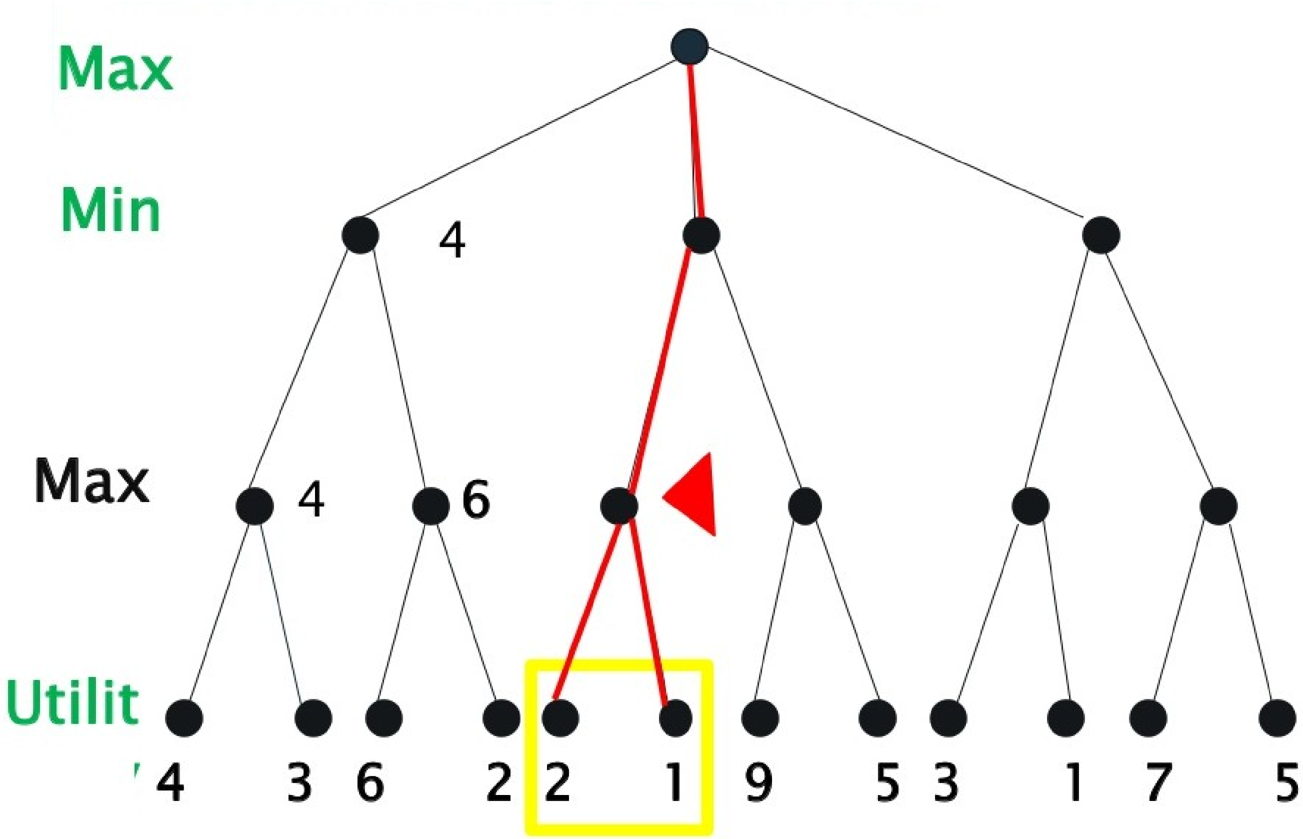
Courtesy : Artificial Intelligence and Soft Computing. Behavioural and Cognitive Modelling of the Human Brain

- A two player game
- Each player will try to maximize one own profit
- The player x will plan the search by assuming that the player o will play his/her best, hence Max(x), Min(o)

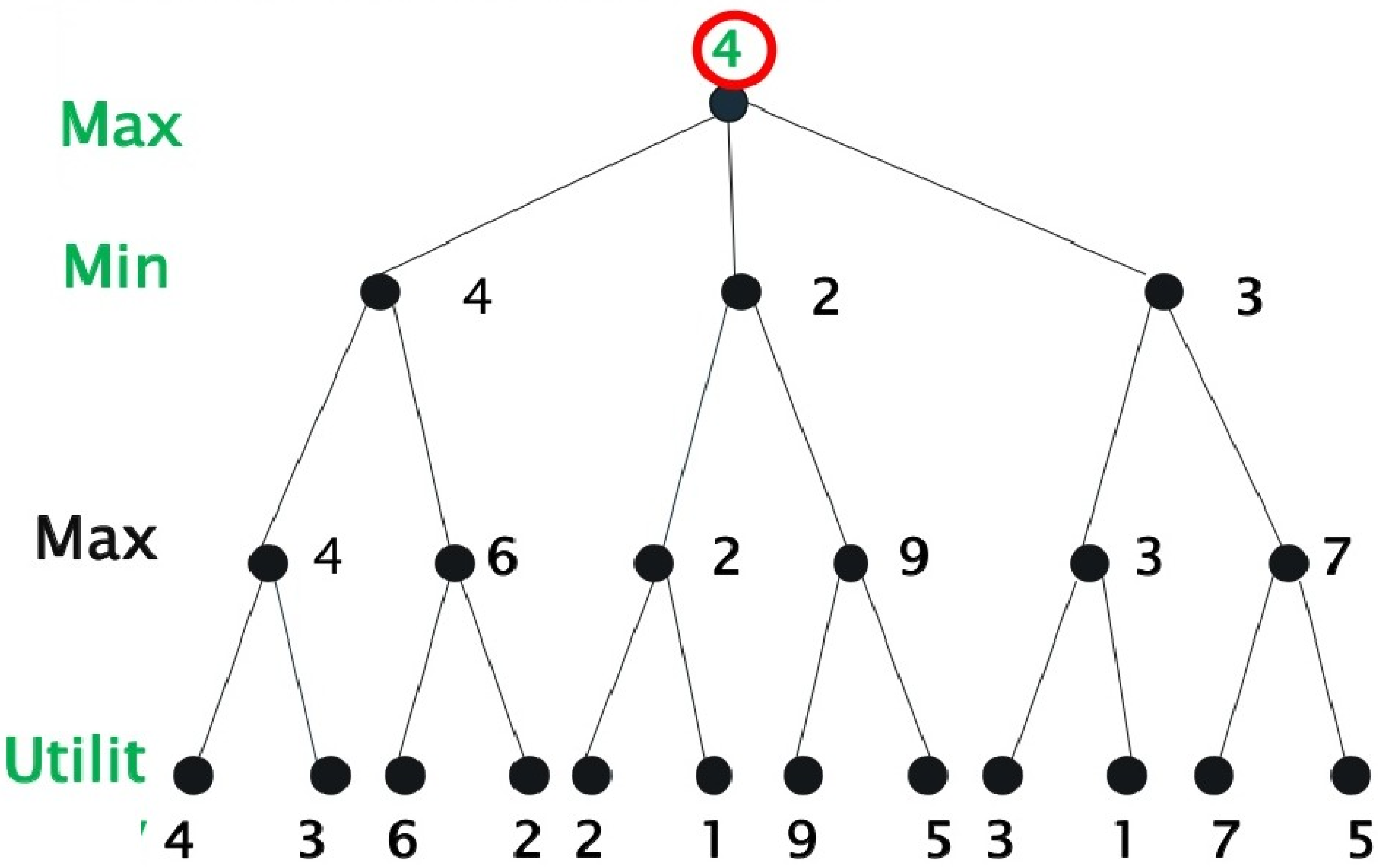

Max
Min
Max
Utility
4 3 6 2 2 1 9 5 3 1 7 5

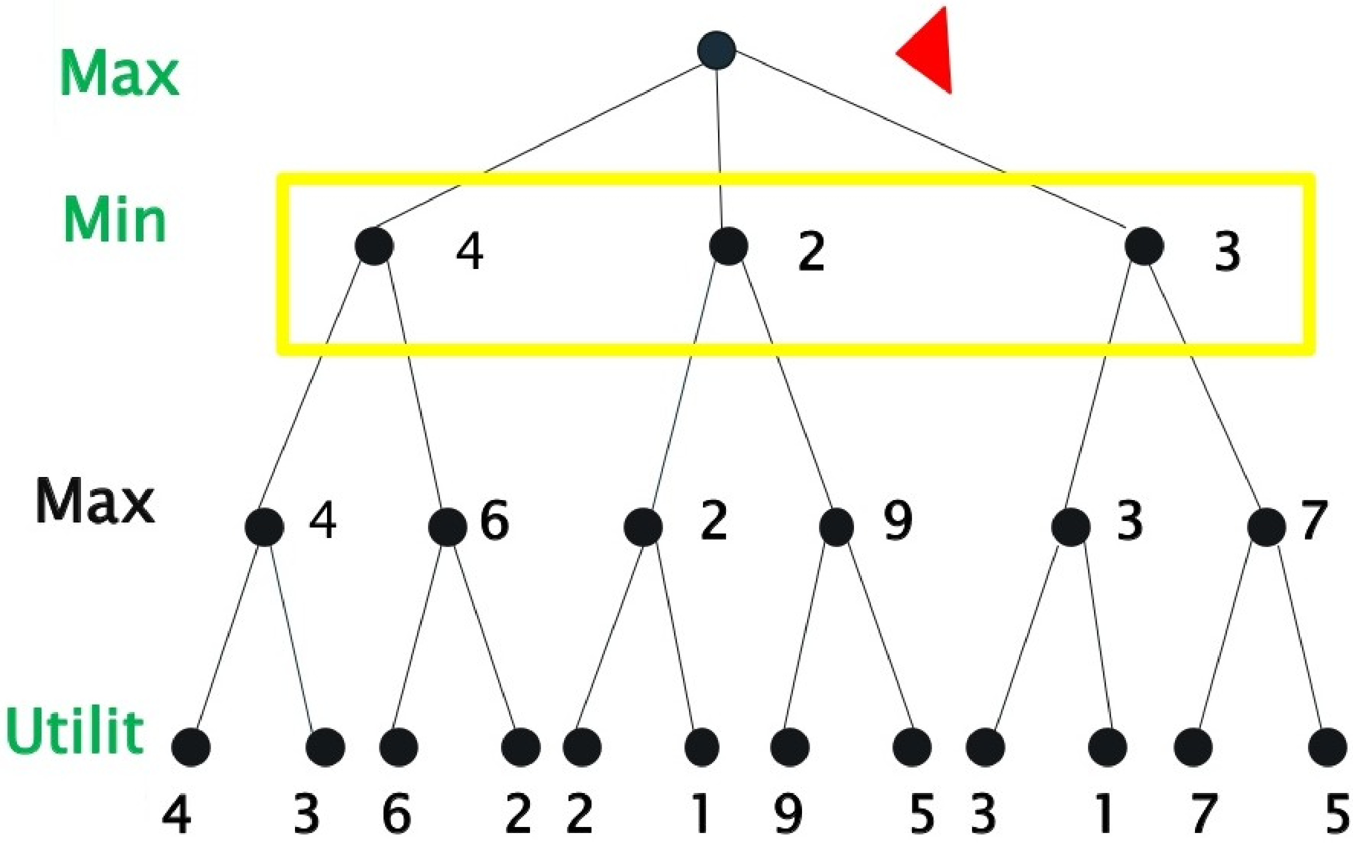

Max
Min
Max
4
Utilit
4
3
6
2
2
1
9
5
3
1
7
5


Max
Min
Max
4
6
Utilit
4
3
6
2
2
1
9
5
3
1
7
5

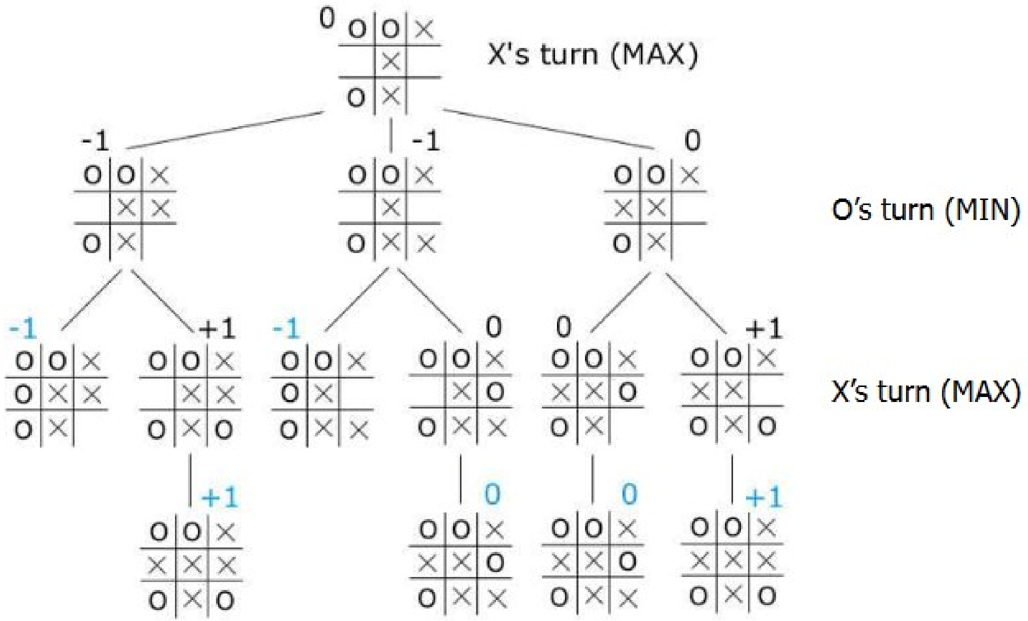


Max
Min
4
Max
4
6
Utilit
4
3
6
2
2
1
9
5
3
1
7
5


4
Max
Min
4
2
3
Max
4
6
2
9
3
7
Utilit
4
3
6
2
2
1
9
5
3
1
7
5


Max
Min
4
2
3
Max
4
6
2
9
3
7
Utilit
4
3
6
2
2
1
9
5
3
1
7
5

State Space Search: Adversary Search

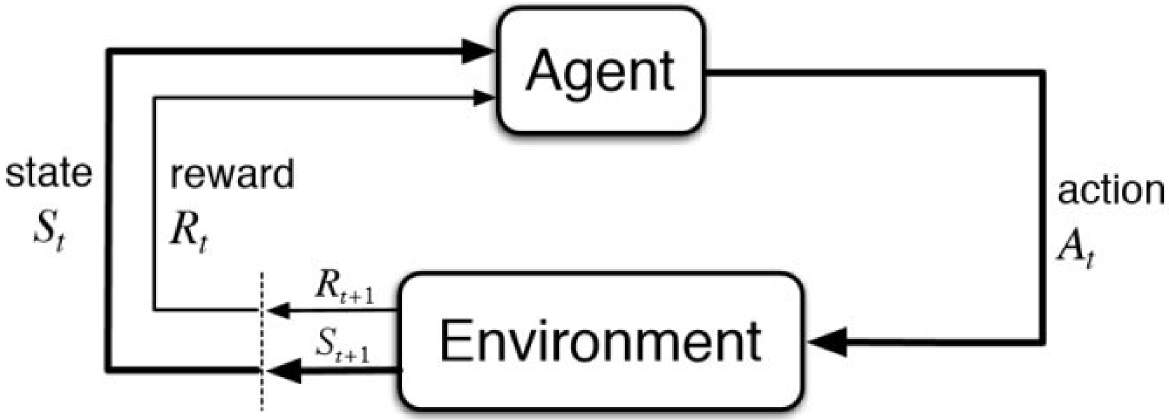


- A two player game where each player try to maximize one own profit.
- The player x will plan the search by assuming that the player o will play his/her best, hence Max(x), Min(o)
- We think for our opponent, why don't we just observe and exploit that observations?

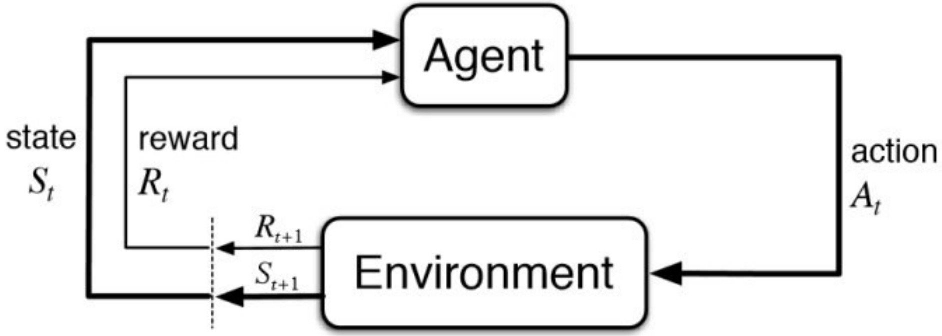
Reinforcement Learning

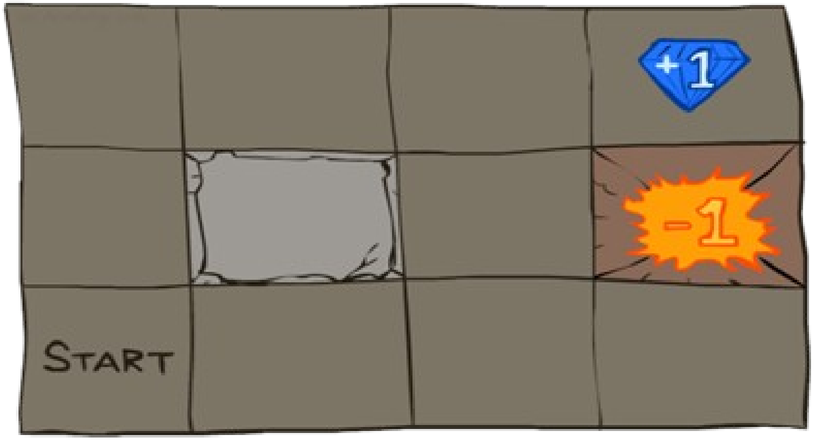
- Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

  (Wikipedia)



Reinforcement Learning





Reinforcement Learning

- Reinforcement learning (RL) is one of three basic machine learning paradigms.
- It learns to behave intelligently from reward signals which implicitly teaching the agent about the environment.
- The environment is typically formulated as a Markov decision process (MDP)
  - S finite states,
  - A finite actions
  - T(s,a,s’) transition probability
  - R(s,a,s’) transition reward

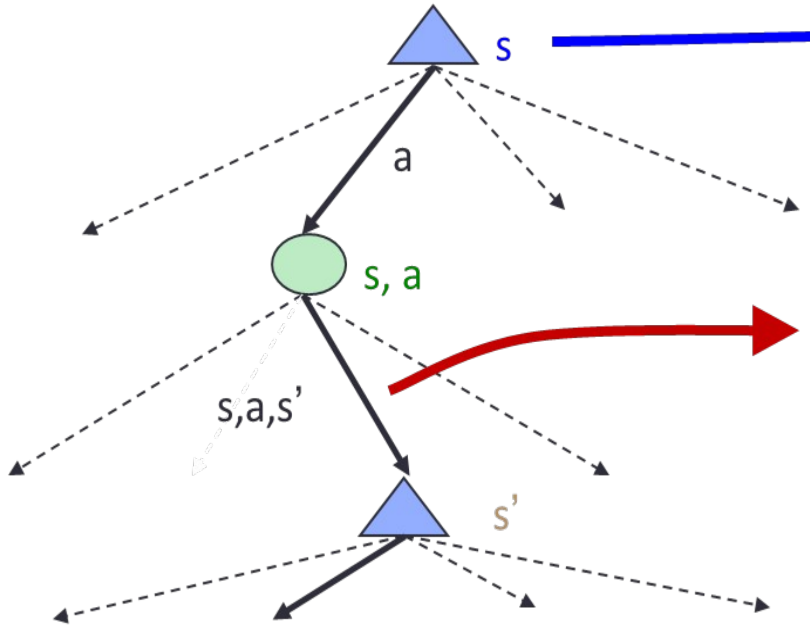
Markov Decision Process

- A Markov decision process is a 5-tuple (S,A,T(.,.,.),R(.,.,.),gamma) where
- S is a finite set of states
- A is a finite set of actions
- T (s,a,s') = P(s(t+1) = s' | s(t) = s, a(t) = a)

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \ldots S_0 = s_0)$$
$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t)$$

- R (s,a,s') is the expected reward after taking action a
- Gamma in [0,1] is a discount factor

Reinforcement Learning

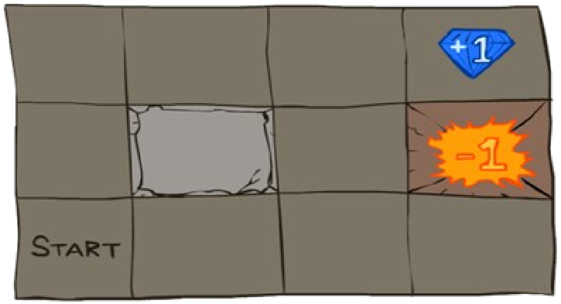


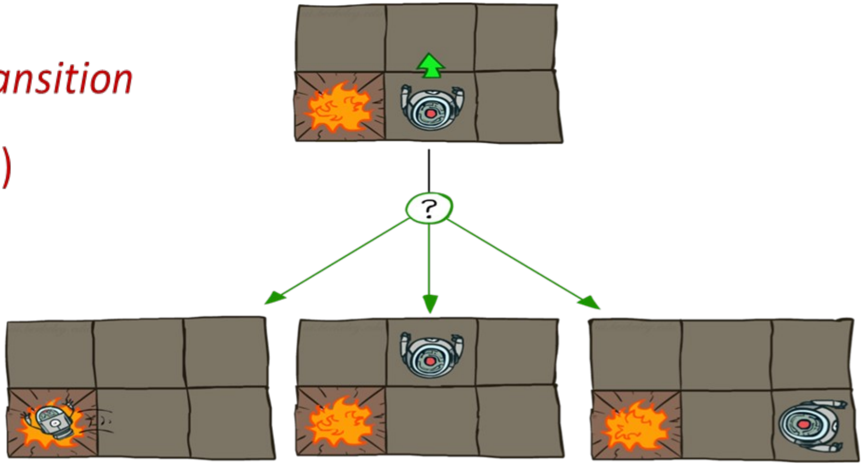
s is a *state*

(s,a,s') called a *transition*

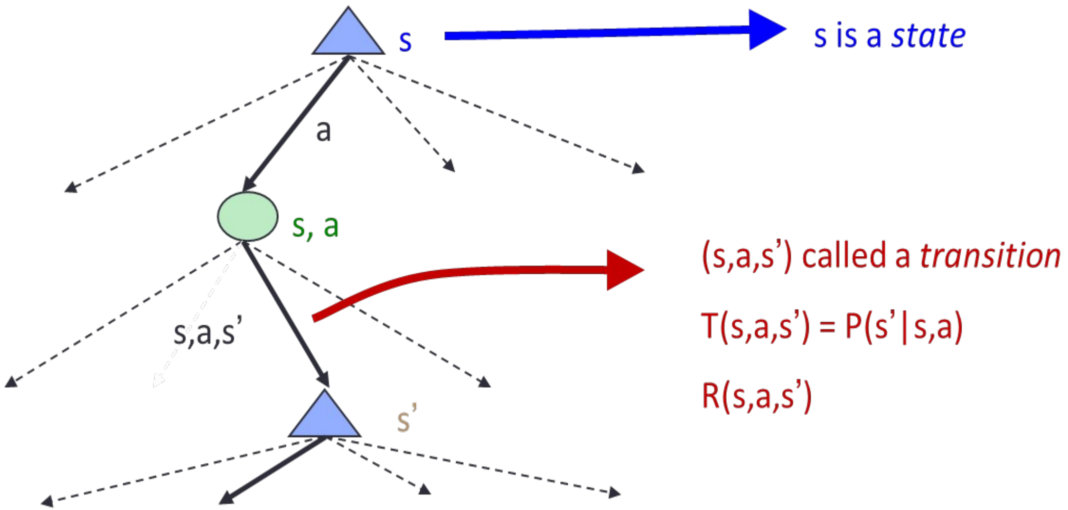
T(s,a,s') = P(s'|s,a)

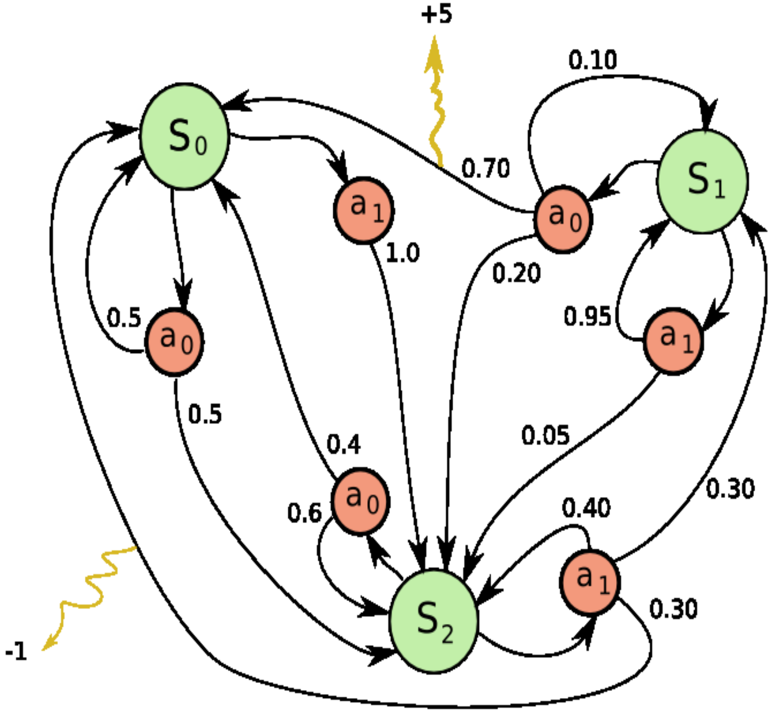
R(s,a,s')





Reinforcement Learning





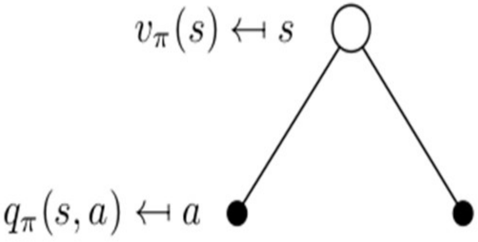
$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$

$$q_\pi(s,a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$$

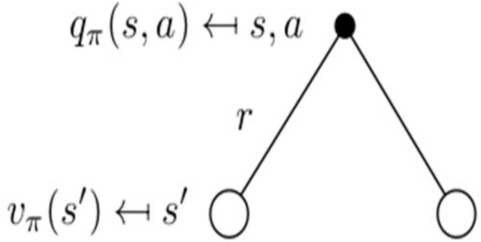
Bellman Equation - compute V(s)

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_t = s\right] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+2} \,\middle|\, S_{t+1} = s'\right]\right] \\
&= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a)\Big[r + \gamma v_\pi(s')\Big], \quad \forall s \in \mathcal{S},
\end{aligned}
$$

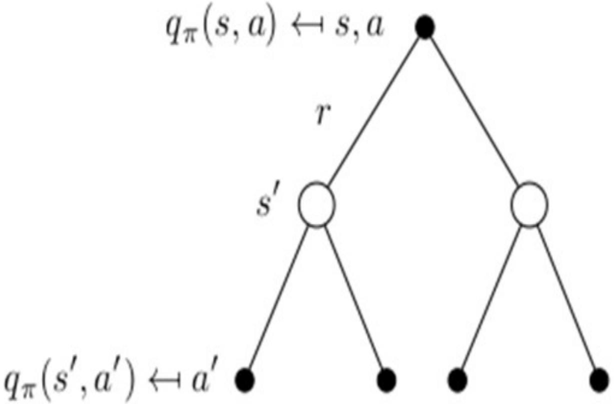
Reinforcement Learning



$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$



$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$



$$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s',a')$$

Reinforcement Learning

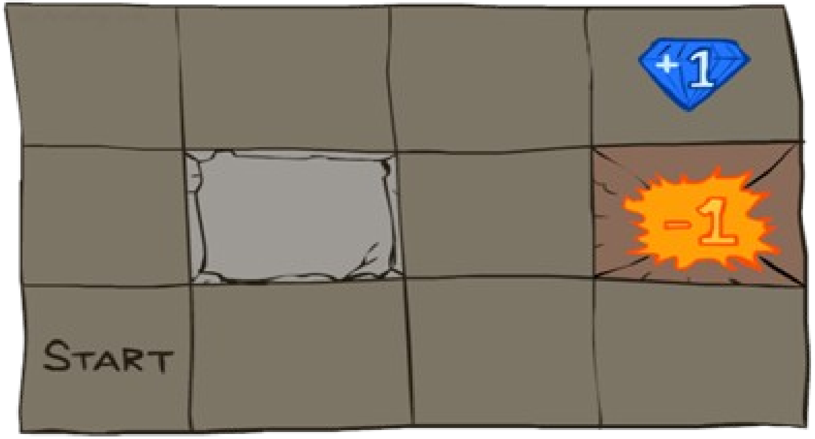
- Below are the Bellman equations, and they characterize optimal values.
- V*(s) = expected utility starting in s and acting optimally
- Q*(s,a) = expected utility starting out having taken action a from state s and (thereafter) acting optimally

$$V^*(s) = \max_a Q^*(s,a)$$

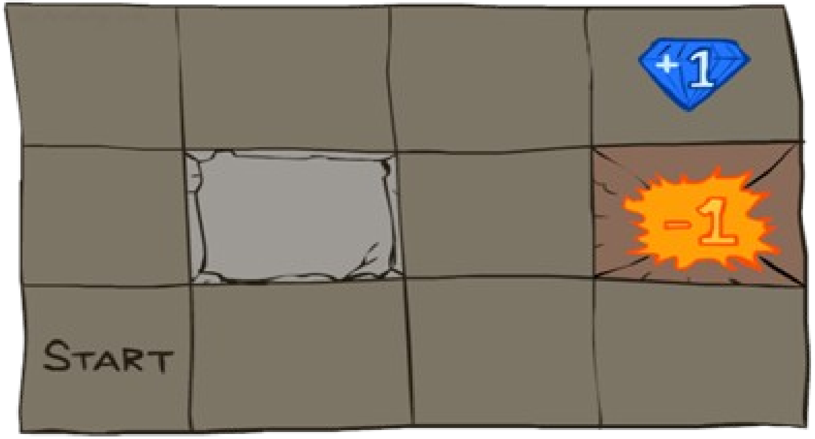
$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$$

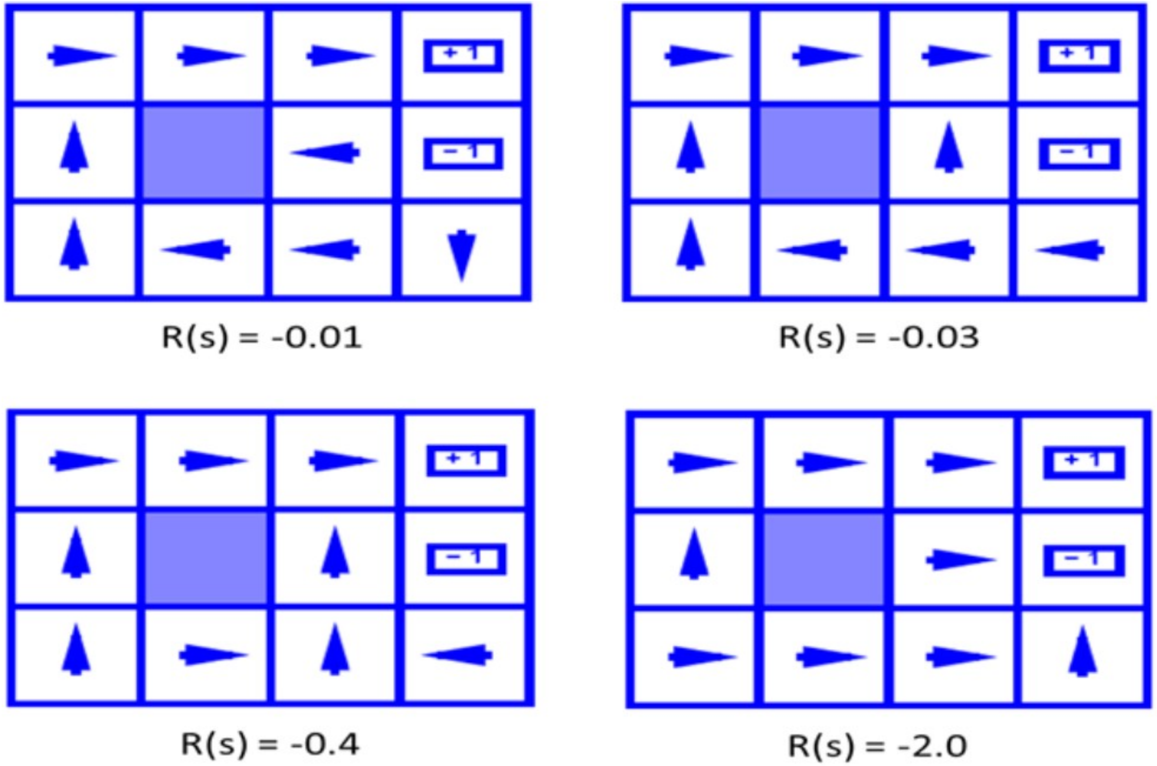
$$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$$

Reinforcement Learning



Reinforcement Learning





Reinforcement Learning

Algorithm	Description	Model	Policy	Action Space	State Space	Operator
Monte Carlo	Every visit to Monte Carlo	Model-Free	Off-policy	Discrete	Discrete	Sample-means
Q-learning	State–action–reward–state	Model-Free	Off-policy	Discrete	Discrete	Q-value
SARSA	State–action–reward–state–action	Model-Free	On-policy	Discrete	Discrete	Q-value
Q-learning - Lambda	State–action–reward–state with eligibility traces	Model-Free	Off-policy	Discrete	Discrete	Q-value
SARSA - Lambda	State–action–reward–state–action with eligibility traces	Model-Free	On-policy	Discrete	Discrete	Q-value
DQN	Deep Q Network	Model-Free	Off-policy	Discrete	Continuous	Q-value
DDPG	Deep Deterministic Policy Gradient	Model-Free	Off-policy	Continuous	Continuous	Q-value
A3C	Asynchronous Advantage Actor-Critic Algorithm	Model-Free	On-policy	Continuous	Continuous	Advantage
NAF	Q-Learning with Normalized Advantage Functions	Model-Free	Off-policy	Continuous	Continuous	Advantage
TRPO	Trust Region Policy Optimization	Model-Free	On-policy	Continuous	Continuous	Advantage
PPO	Proximal Policy Optimization	Model-Free	On-policy	Continuous	Continuous	Advantage
TD3	Twin Delayed Deep Deterministic Policy Gradient	Model-Free	Off-policy	Continuous	Continuous	Q-value
SAC	Soft Actor-Critic	Model-Free	Off-policy	Continuous	Continuous	Advantage

Technology Trends:

Chainer

mxnet

Caffe2

Microsoft CNTK

TensorFlow

Keras

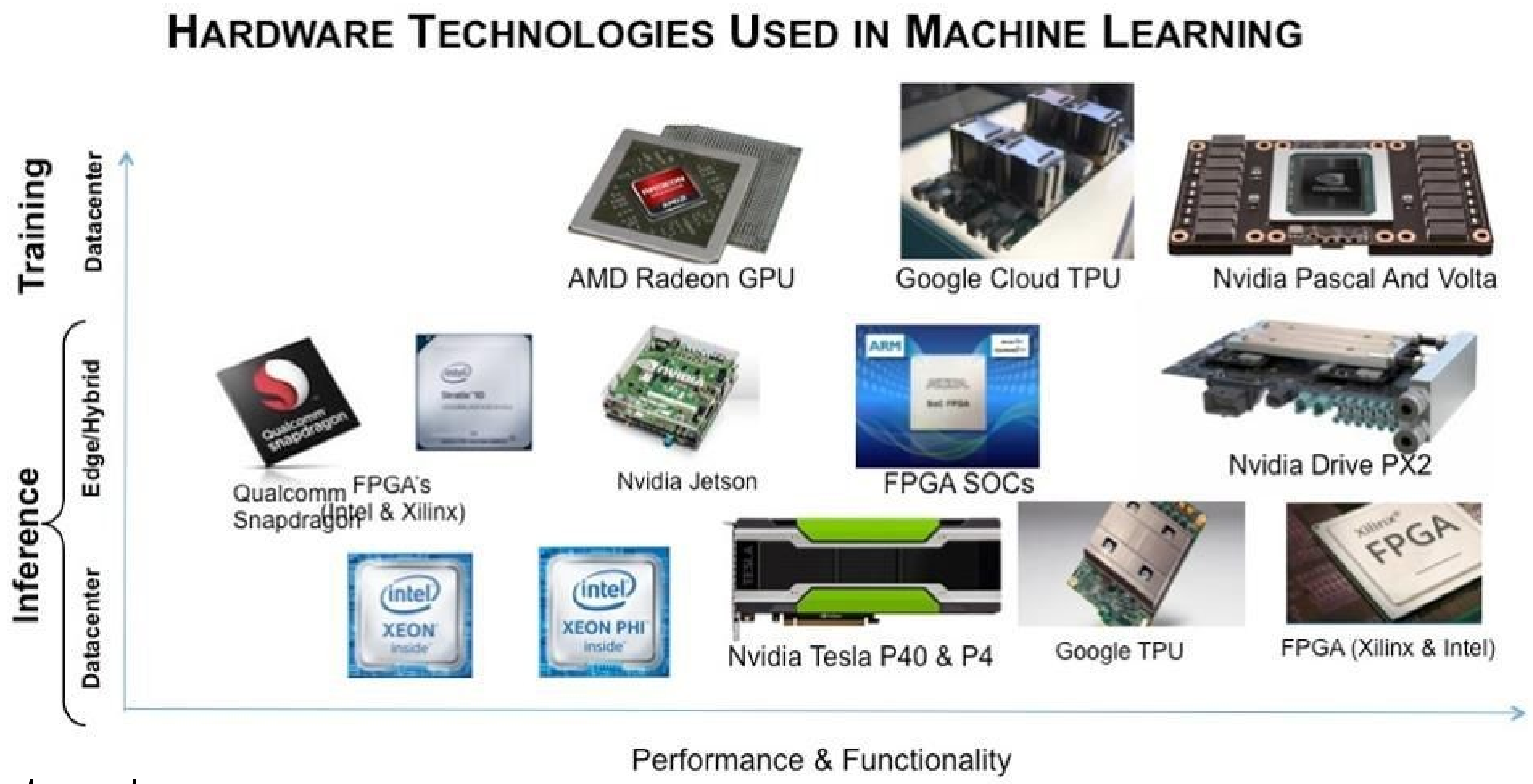
GLUON

PYTORCH

theano

Internet

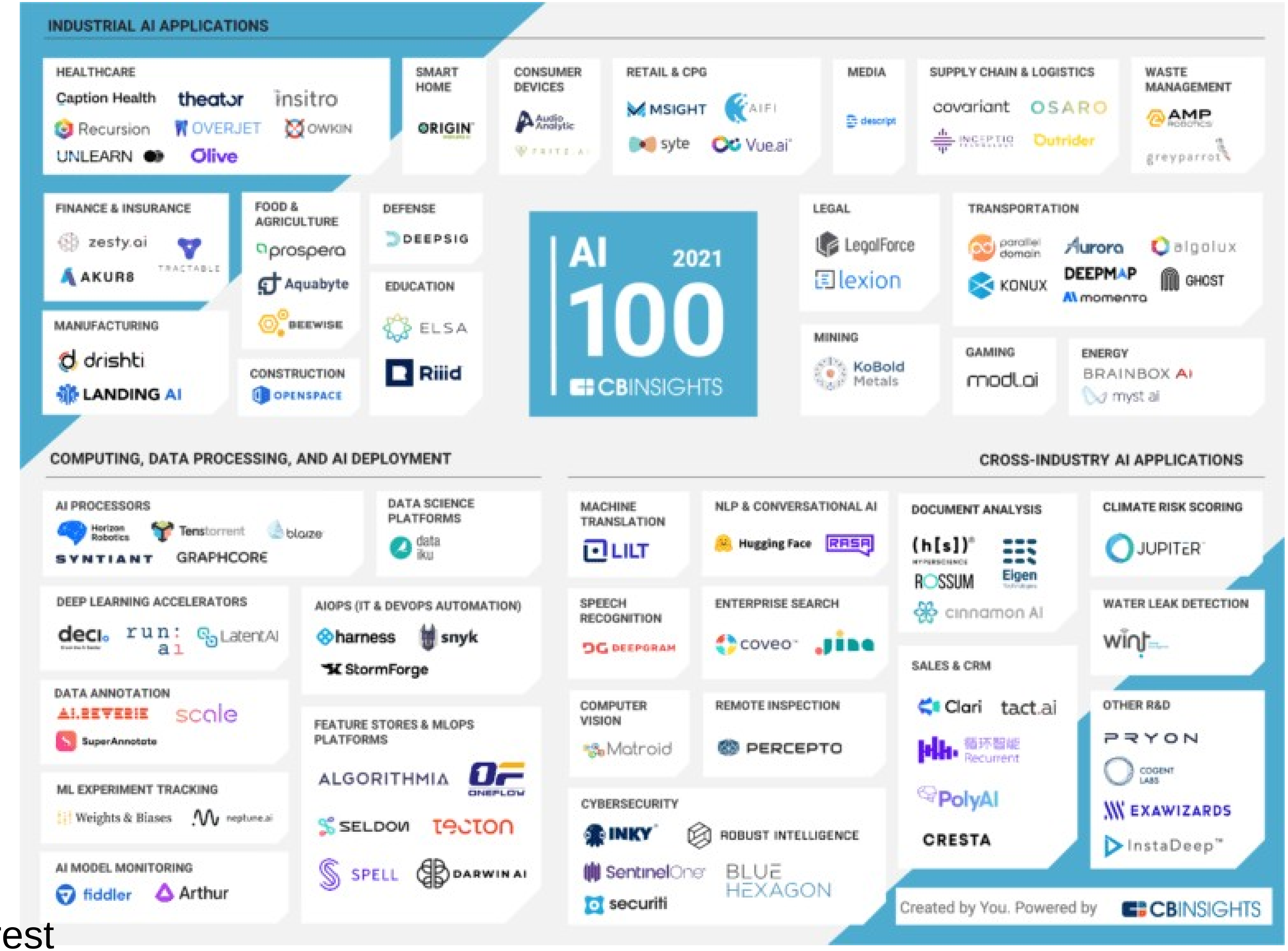
Technology Trends:



Internet

Technology Trends:

- TensorFlow, Keras, Pytorch, Caffe, Chainer
- Python, Java, Javascript, Swift, C++, Matlab
- Google AutoML, Microsoft Azure, AWS, IBM Watson
- Intel OpenVINO, NVIDIA Omniverse, Facebook Meta, OpenAI, DeepMind, HuggingFace, Rasa, etc.
- Generative Pre-trained Transformer (GPT)
- AlphaGo (2014), AlphaZero, AlphaStar: Games
- AlphaFold (2016): Protein fold prediction
- WaveNet (2016) Text to speech system
- etc



Credit Pinterest

Q & A