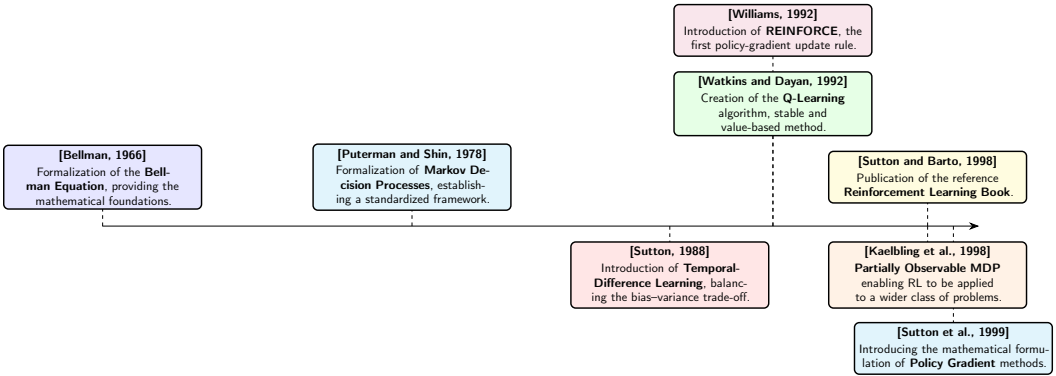
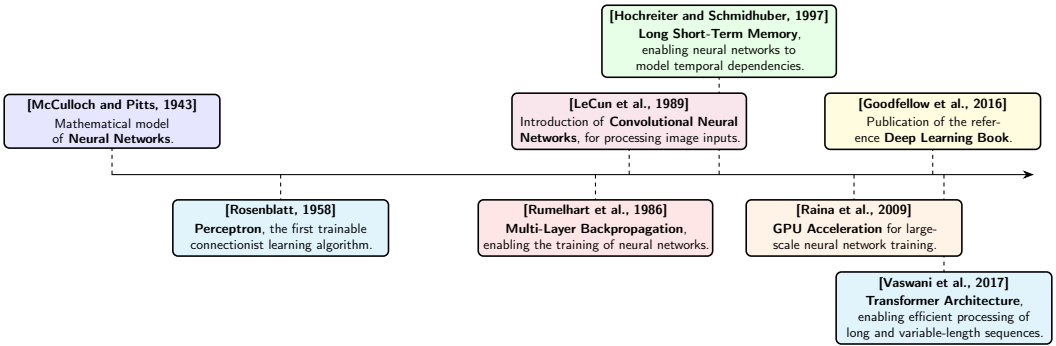
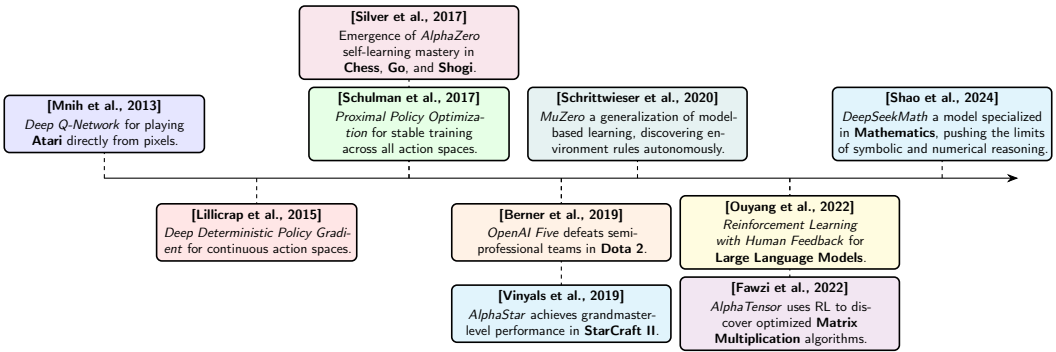
RLlib: Industry-Grade, Scalable Reinforcement Learning

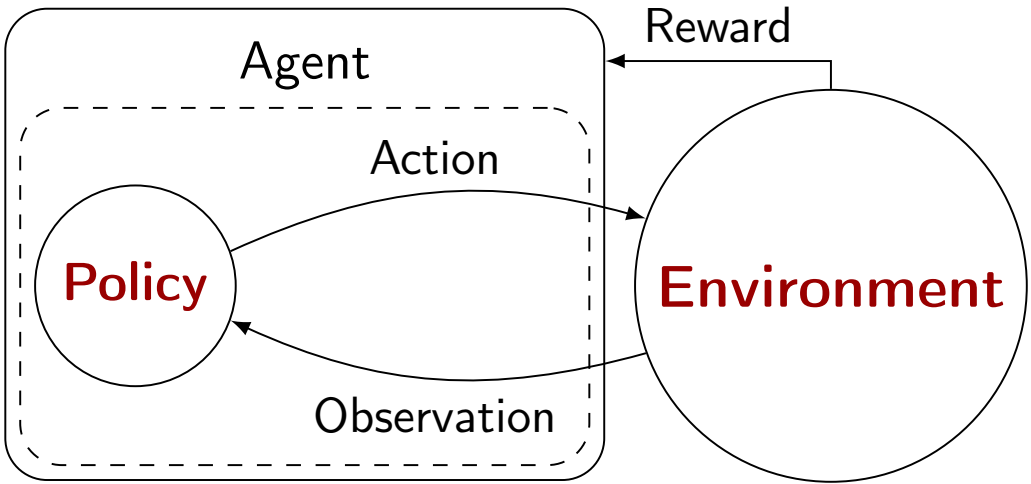
Maxime Alaarabiou


[Bellman, 1966]
Formalization of the **Bellman Equation**, providing the mathematical foundations.
[Puterman and Shin, 1978]
Formalization of **Markov Decision Processes**, establishing a standardized framework.
[Williams, 1992]
Introduction of **REINFORCE**, the first policy-gradient update rule.
[Watkins and Dayan, 1992]
Creation of the **Q-Learning** algorithm, stable and value-based method.
[Sutton, 1988]
Introduction of **Temporal-Difference Learning**, balancing the bias–variance trade-off.
[Sutton and Barto, 1998]
Publication of the reference **Reinforcement Learning Book**.
[Kaelbling et al., 1998]
Partially Observable MDP enabling RL to be applied to a wider class of problems.
[Sutton et al., 1999]
Introducing the mathematical formulation of **Policy Gradient** methods.


[Hochreiter and Schmidhuber, 1997]
Long Short-Term Memory, enabling neural networks to model temporal dependencies.
[McCulloch and Pitts, 1943]
Mathematical model of Neural Networks.
[LeCun et al., 1989]
Introduction of Convolutional Neural Networks, for processing image inputs.
[Goodfellow et al., 2016]
Publication of the reference Deep Learning Book.
[Rosenblatt, 1958]
Perceptron, the first trainable connectionist learning algorithm.
[Rumelhart et al., 1986]
Multi-Layer Backpropagation, enabling the training of neural networks.
[Raina et al., 2009]
GPU Acceleration for large-scale neural network training.
[Vaswani et al., 2017]
Transformer Architecture, enabling efficient processing of long and variable-length sequences.


[Silver et al., 2017]
Emergence of *AlphaZero* self-learning mastery in **Chess**, **Go**, and **Shogi**.
[Mnih et al., 2013]
Deep Q-Network for playing **Atari** directly from pixels.
[Schulman et al., 2017]
Proximal Policy Optimization for stable training across all action spaces.
[Schrittwieser et al., 2020]
MuZero a generalization of model-based learning, discovering environment rules autonomously.
[Shao et al., 2024]
DeepSeekMath a model specialized in **Mathematics**, pushing the limits of symbolic and numerical reasoning.
[Lillicrap et al., 2015]
Deep Deterministic Policy Gradient for continuous action spaces.
[Berner et al., 2019]
OpenAI Five defeats semi-professional teams in **Dota 2**.
[Ouyang et al., 2022]
Reinforcement Learning with Human Feedback for **Large Language Models**.
[Vinyals et al., 2019]
AlphaStar achieves grandmaster-level performance in **StarCraft II**.
[Fawzi et al., 2022]
AlphaTensor uses RL to discover optimized **Matrix Multiplication** algorithms.

RL Algorithm



Gymnasium:

```python
import gymnasium as gym

env = gym.make("ALE/SpaceInvaders-v5", render_mode="rgb_array")
obs, info = env.reset()
action = env.action_space.sample()
obs, reward, terminated, truncated, info = env.step(action)
```

PettingZoo:

```python
from pettingzoo.atari import space_invaders_v2

env = space_invaders_v2.env(render_mode="rgb_array")
env.reset()
for agent in env.agent_iter():
    obs, reward, term, trunc, info = env.last()
    action = env.action_space(agent).sample() if not term else None
    env.step(action)
```

Environment Configuration	
game_name	'SpaceInvaders-v5'
repeat_action_probability	0.05
frameskip	5
resize_observation_shape	(64, 64)
convert_to_grayscale	True
reward_scale_factor	0.05
frame_stack_len	4
normalize_observation	True
observation_numpy_type	np.float16
Policy Architecture Configuration	
architecture	CnnPPO
configuration_cnn	[(16,4,2),(32,4,2),(64,4,2),(128,4,2)]
configuration_hidden_layers	[512,256,128]
activation_function_class	LeakyReLU
use_layer_normalization_cnn	True
use_share_cnn	True
Reinforcement Learning Configuration	
algorithm_name	'PPO'
rollout_fragment_length	2048
train_batch_size	2048 * 8
minibatch_size	2048
lambda_gae	0.95
kullback_leibler_coefficient	0.5
clip_policy_parameter	0.1
clip_value_function_parameter	10
entropy_coefficient	0.01
number_epochs	10
learning_rate	0.00015
gradient_clip	100.0
gradient_clip_by	'global_norm'

Why choose RLlib?

Open-source and actively maintained

Scalable from laptop to large compute clusters

Supports many modern algorithms: **DQN, DDPG, PPO, Dreamer, ...**

Customizable Policy Architecture: Dense, CNN, LSTM, Attention

Compatible with both **PyTorch** and **TensorFlow**

Automatic Plotting of training curves

Automatic Checkpointing and Restart of training

Multi-agent and **Hierarchical** RL

Advanced Callback System for monitoring and customization

Modules for **Exploration**, **Curriculum Learning**, and **Custom RL Algorithms**

```python
from ray.rllib.algorithms.ppo import PPOConfig
from pprint import pprint

# Configure the algorithm.
config = (
    PPOConfig()
    .environment("ALE/SpaceInvaders-v5")
)

# Build the algorithm.
algo = config.build_algo()
# Train it for 5 iterations ...
for _ in range(5):
    pprint(algo.train())
# ... and evaluate it.

pprint(algo.evaluate())
# Release the algo's resources.
algo.stop()
```

Bellman, R. (1966).
Dynamic programming.
science, 153(3731):34–37.
Book.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., et al. (2019).
Dota 2 with large scale deep reinforcement learning.
arXiv preprint arXiv:1912.06680.
Article. **Video**.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R. Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. (2022).
Discovering faster matrix multiplication algorithms with reinforcement learning.
Nature, 610(7930):47–53.

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep Learning.
MIT Press.

Book.

Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.
Neural computation, 9(8):1735–1780.
Article.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998).
Planning and acting in partially observable stochastic domains.
Artificial intelligence, 101(1-2):99–134.
Article.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989).
Backpropagation applied to handwritten zip code recognition.
Neural computation, 1(4):541–551.
Article.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015).

Continuous control with deep reinforcement learning.
Article.


McCulloch, W. S. and Pitts, W. (1943).
A logical calculus of the ideas immanent in nervous activity.
The bulletin of mathematical biophysics, 5(4):115–133.
Article.


Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013).
Playing atari with deep reinforcement learning.
Article.


Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022).
Training language models to follow instructions with human feedback.
Advances in Neural Information Processing Systems, 35:27730–27744.
Article.


Puterman, M. L. and Shin, M. C. (1978).

Modified policy iteration algorithms for discounted markov decision problems.
Management Science, 24(11):1127–1137.
Article.


Raina, R., Madhavan, A., and Ng, A. Y. (2009).
Large-scale deep unsupervised learning using graphics processors.
In Proceedings of the 26th annual international conference on machine learning, pages 873–880.
Article.


Rosenblatt, F. (1958).
The perceptron: a probabilistic model for information storage and organization in the brain.
Psychological review, 65(6):386.
Article.


Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986).
Learning representations by back-propagating errors.
nature, 323(6088):533–536.
Article.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., and Silver, D. (2020).

Mastering atari, go, chess and shogi by planning with a learned model.
Nature, 588(7839):604–609.
Article.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017).
Proximal policy optimization algorithms.
arXiv preprint arXiv:1707.06347.
Article.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. (2024).
Deepseekmath: Pushing the limits of mathematical reasoning in open language models.
Article.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2017).
Mastering chess and shogi by self-play with a general reinforcement learning algorithm.
Article.

Sutton, R. S. (1988).
Learning to predict by the methods of temporal differences.
Machine learning, 3(1):9–44.
Article.

Sutton, R. S. and Barto, A. G. (1998).
Reinforcement Learning: An Introduction.
MIT Press.
Book.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999).
Policy gradient methods for reinforcement learning with function approximation.

Advances in neural information processing systems, 12.
Article.


Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
Advances in neural information processing systems, 30.
Article.


Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019).
Grandmaster level in starcraft ii using multi-agent reinforcement learning.
nature, 575(7782):350–354.


Watkins, C. J. and Dayan, P. (1992).
Q-learning.
Machine learning, 8(3):279–292.
Article.


Williams, R. J. (1992).

Simple statistical gradient-following algorithms for connectionist reinforcement learning.
Machine learning, 8(3):229–256.