AlphaGo Zero & AlphaZero

Mastering Go, Chess and Shogi without human knowledge

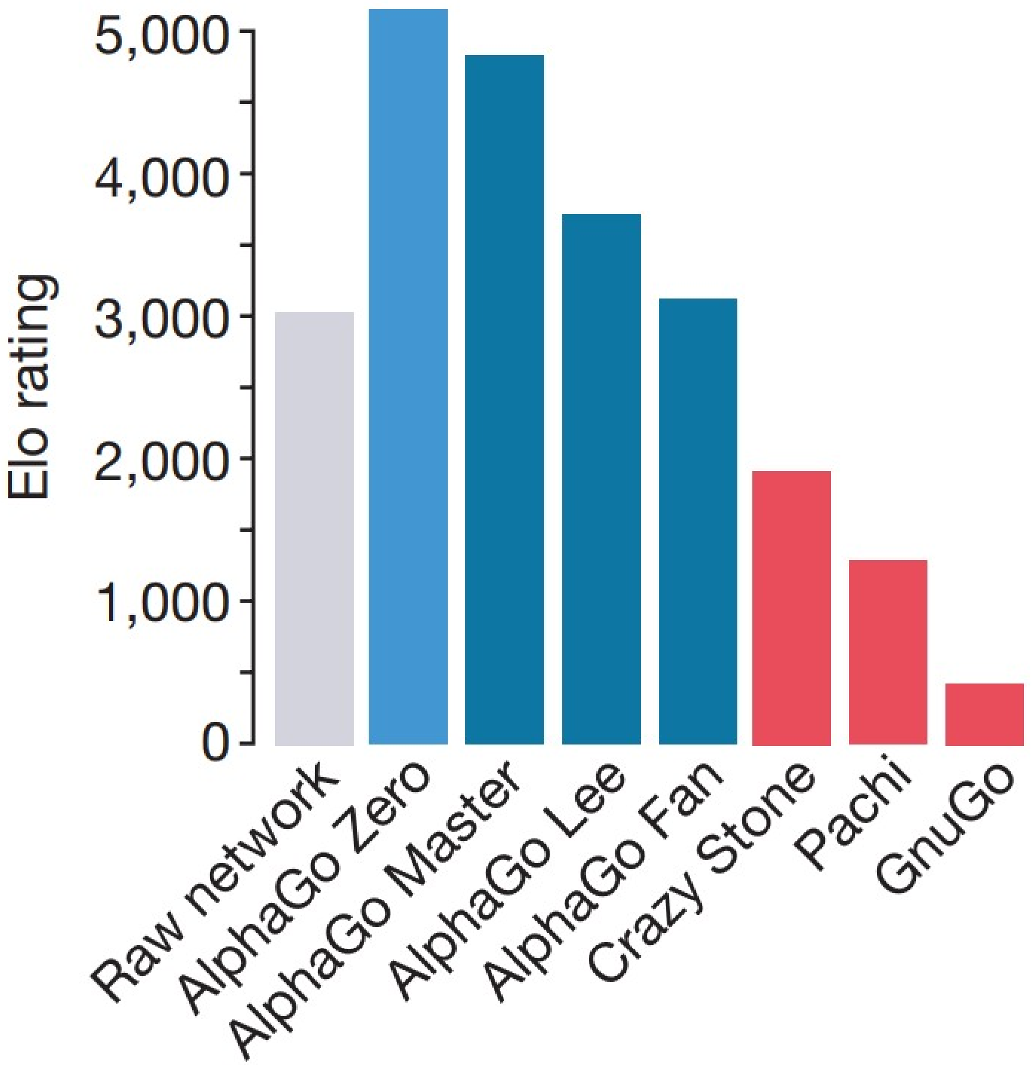
Silver et al. 2017-2018

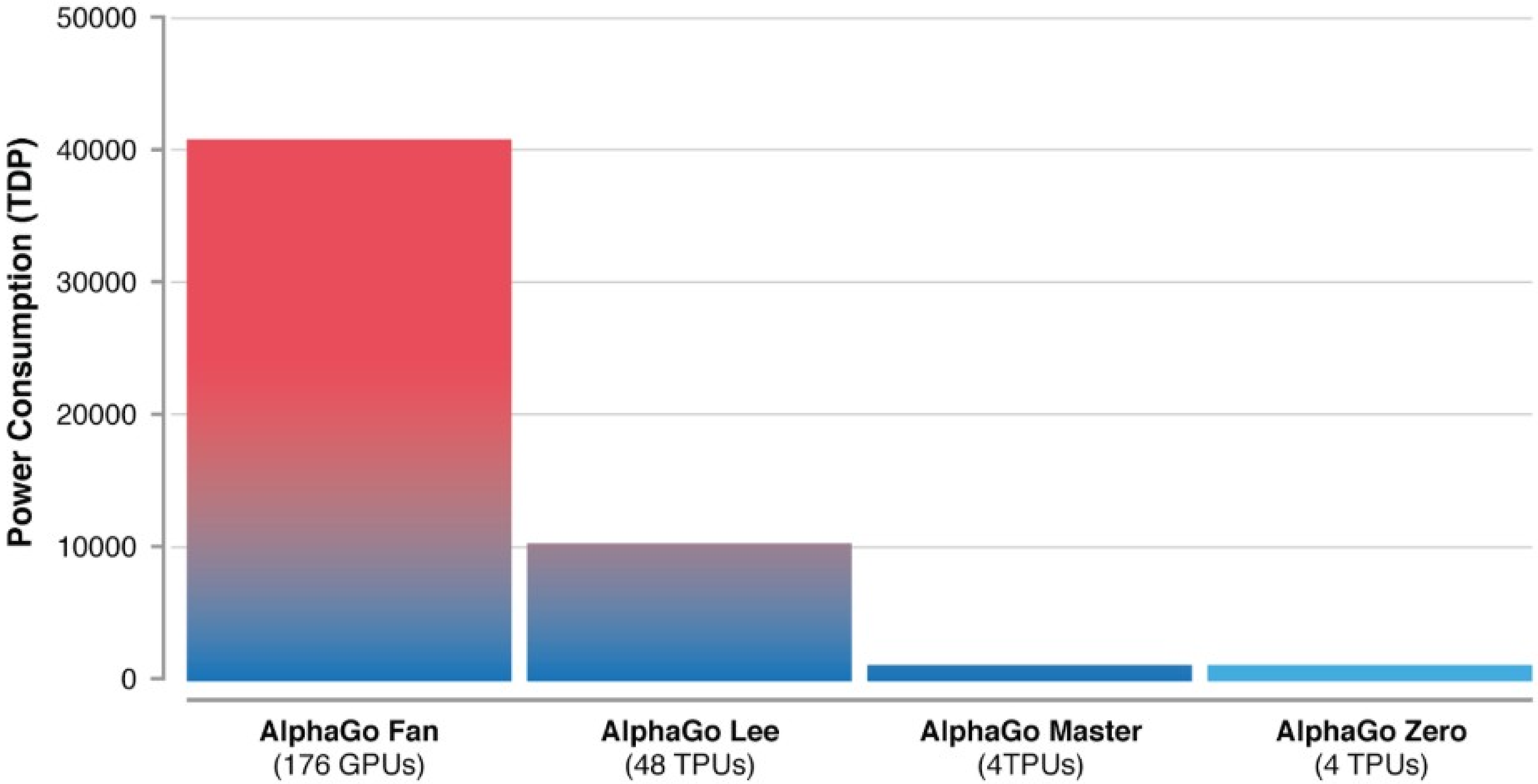
Presenter: Philipp Wimmer

Outline

- Timeline
- AlphaGo Zero
  - Training Pipeline
  - Modified MCTS
  - Reasons for better performance
- AlphaZero
  - Generalization to Chess/Shogi
  - AlphaZero vs Stockfish
- Conclusion

- 2012: Crazy Stone
  - MCTS search with handcrafted heuristic
  - Professional level play
- 2015: AlphaGo Fan
  - MCTS with Value and Policy Network
  - Defeated European Go champion Fan Hui (2-dan)
- 2016: AlphaGo Lee
  - Larger networks than AlphaGo Fan
  - Added selfplay of policy network
  - Won 3/1 against Lee Sedol (9-dan)
- 2017: AlphaGo Master
  - Use single network for both policy and value
  - Using ResNet instead of Convolutional NN
  - Won 60/0 against team of professional players
- 2018 AlphaGo Zero
  - Trained from scratch
- 2018 AlphaZero
  - Generalized to Chess & Shogi



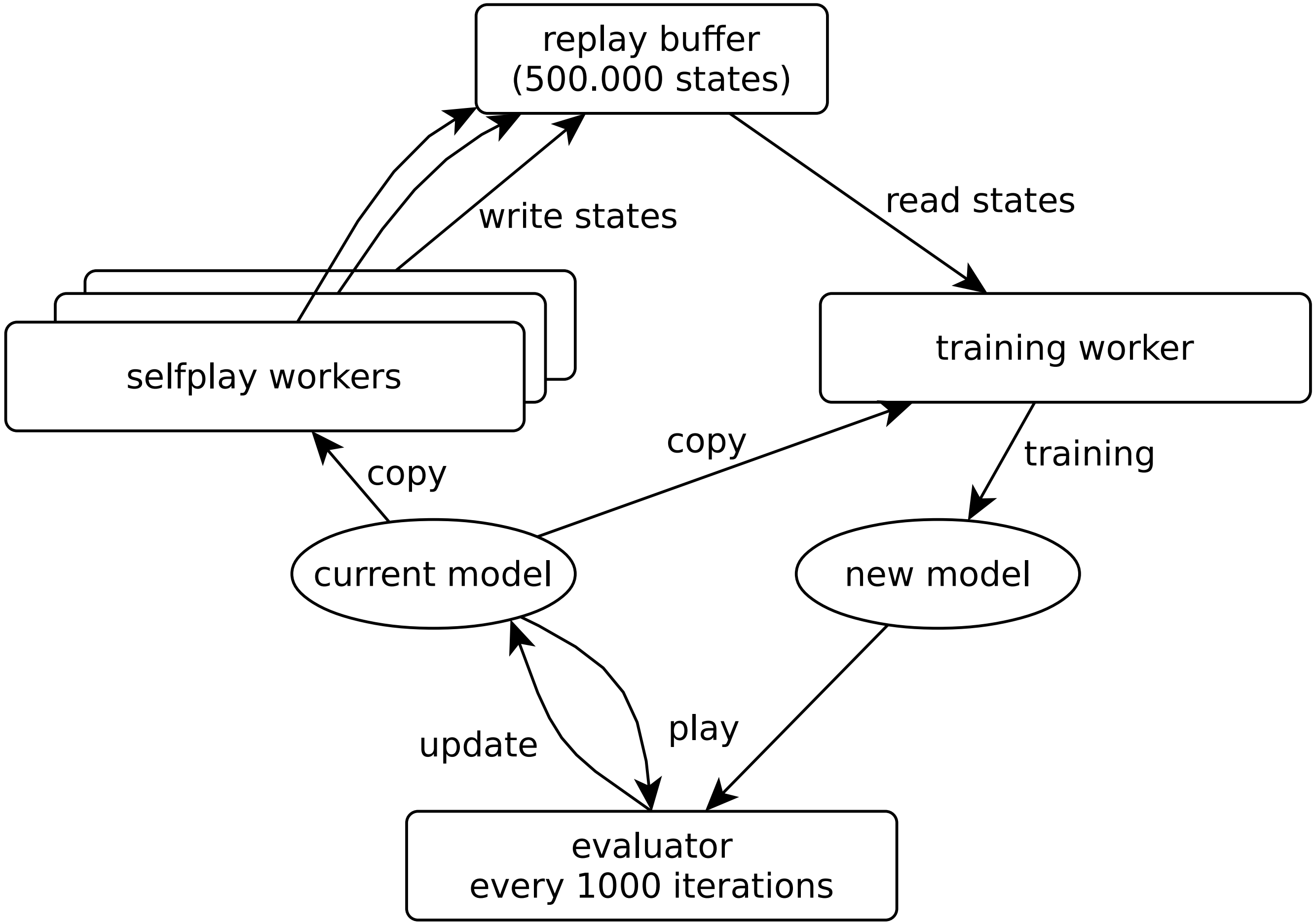

Power Consumption (TDP)
50000
40000
30000
20000
10000
0
AlphaGo Fan
(176 GPUs)
AlphaGo Lee
(48 TPUs)
AlphaGo Master
(4TPUs)
AlphaGo Zero
(4 TPUs)

AlphaGo Zero: learning from scratch

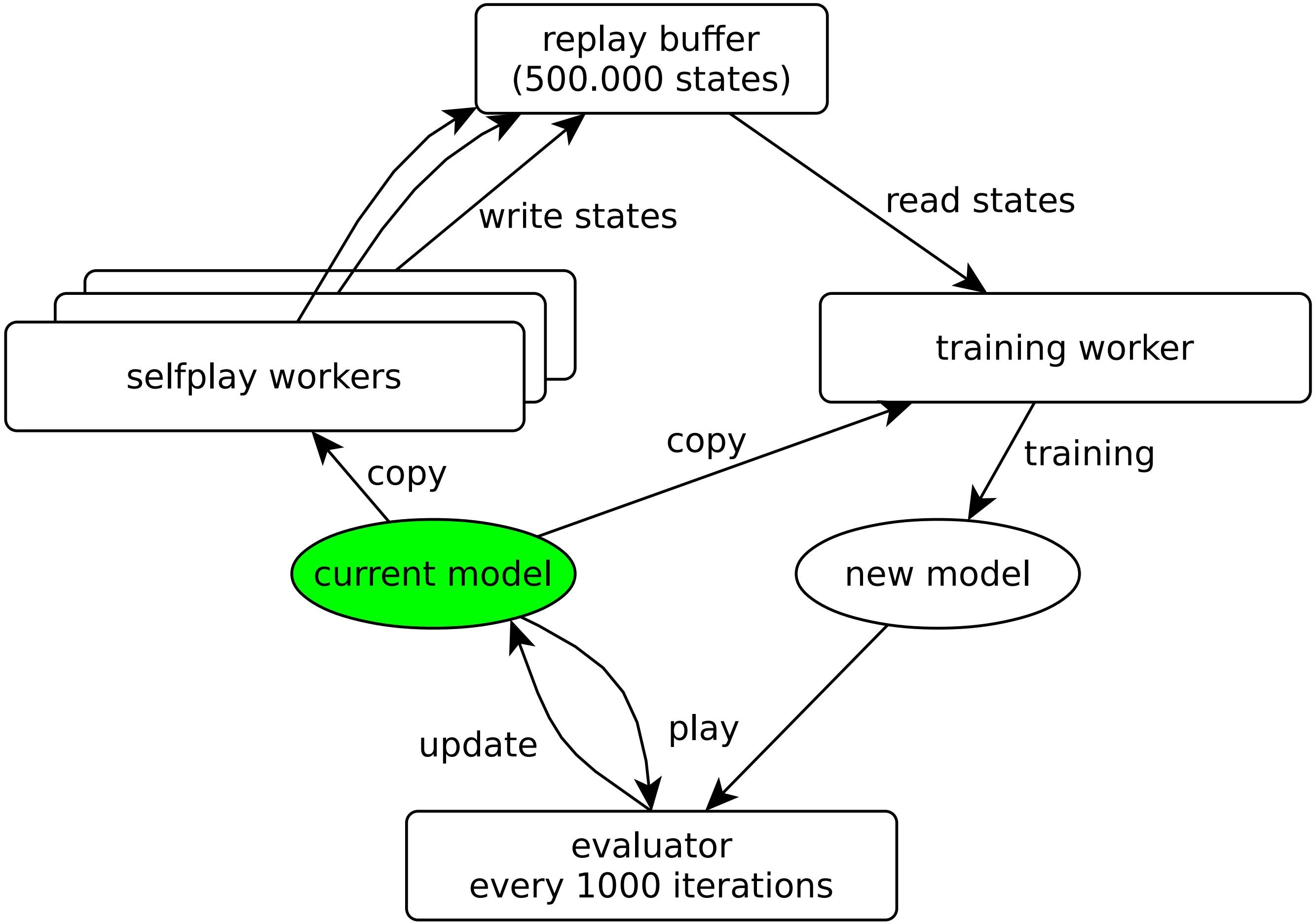
- No human knowledge
  - Trained by self-play reinforcement learning from scratch
  - Only raw board as input
- Single neural network
  - Policy and value networks are combined into single NN
- Simpler (cheaper) search during gameplay
  - Instead of Monte-Carlo rollouts, only uses NN to evaluate

Less complex and more general => AlphaZero (also plays Chess, Shogi, ...)

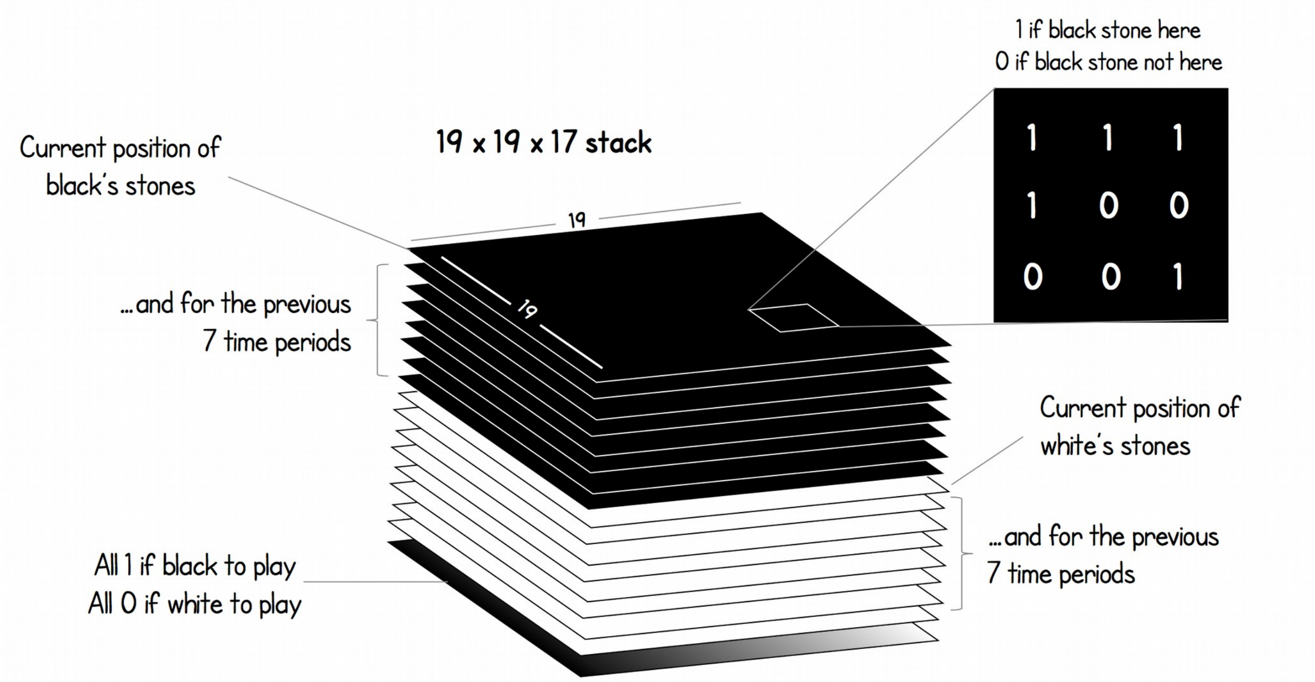
Learning Pipeline



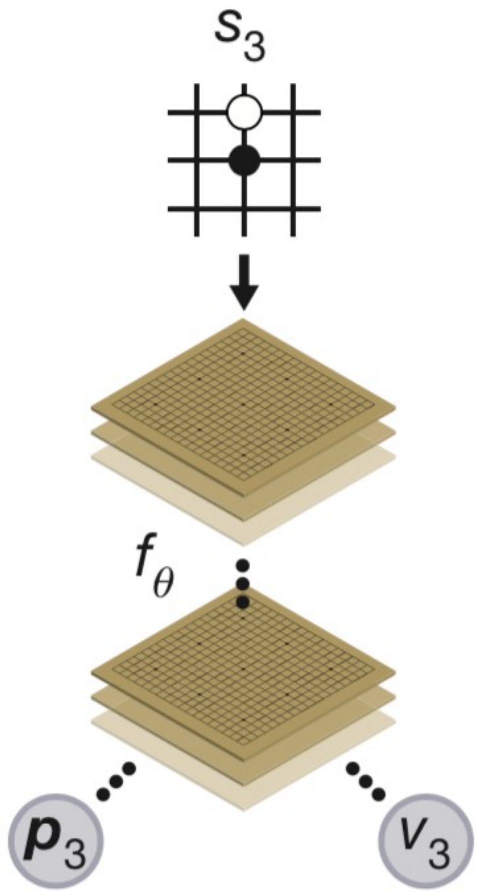
Learning Pipeline



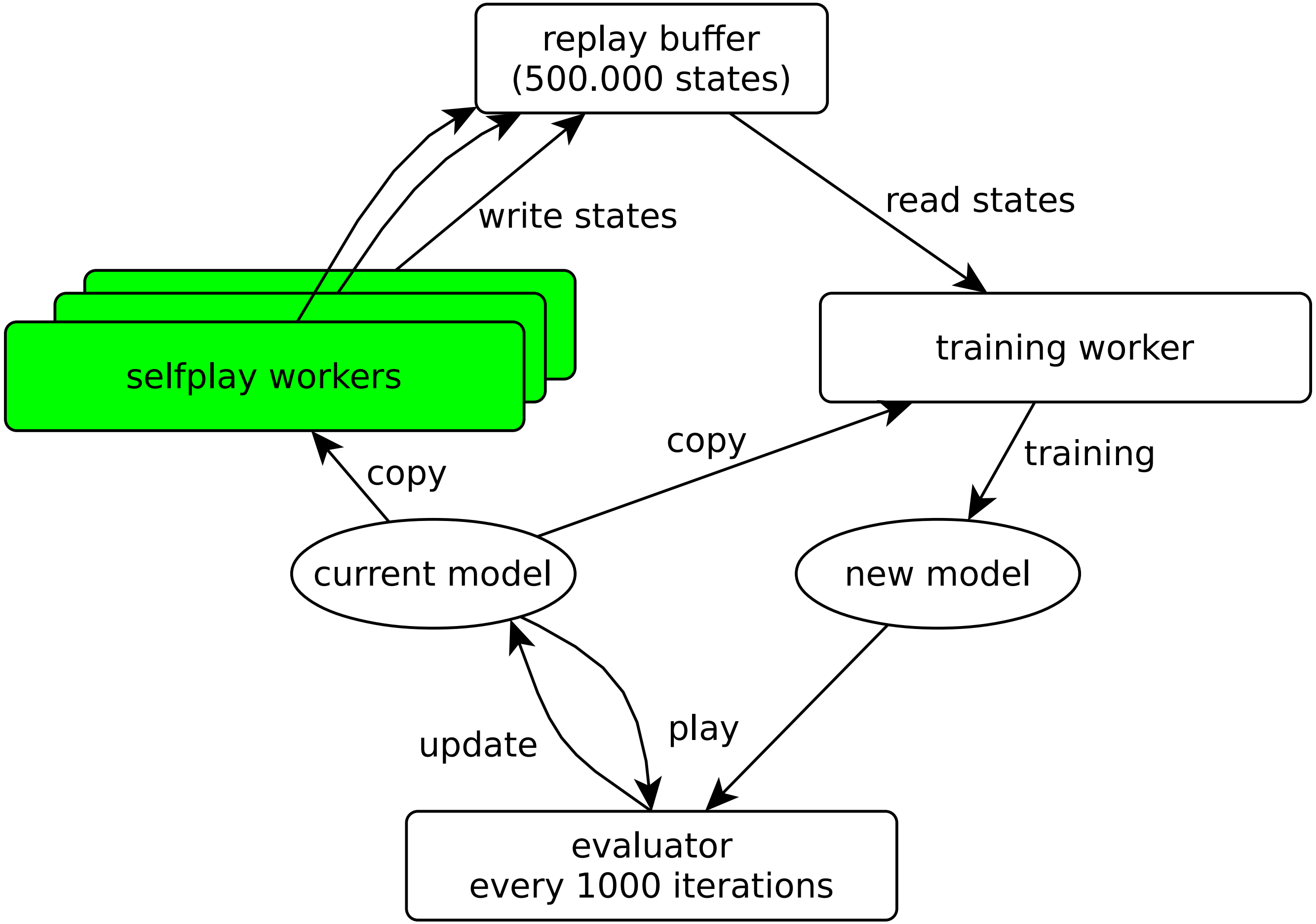
Policy/Value-Network



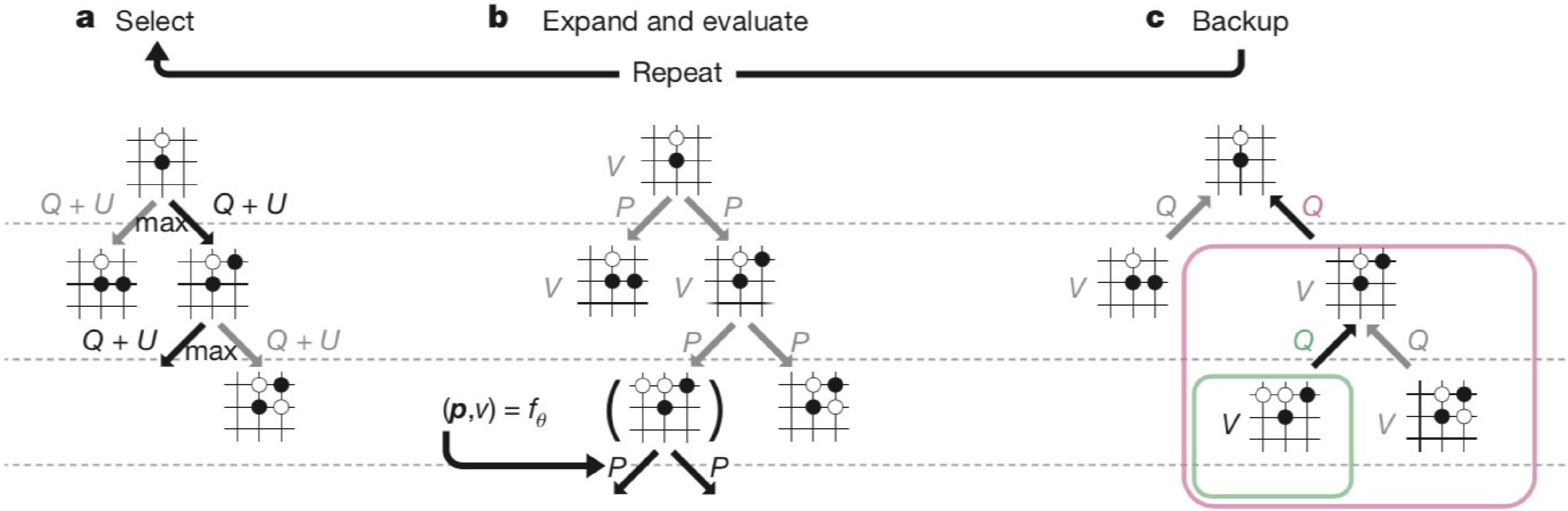
https://medium.com/applied-data-science/alphago-zero-explained-in-one-diagram-365f5abf67e0



Learning Pipeline

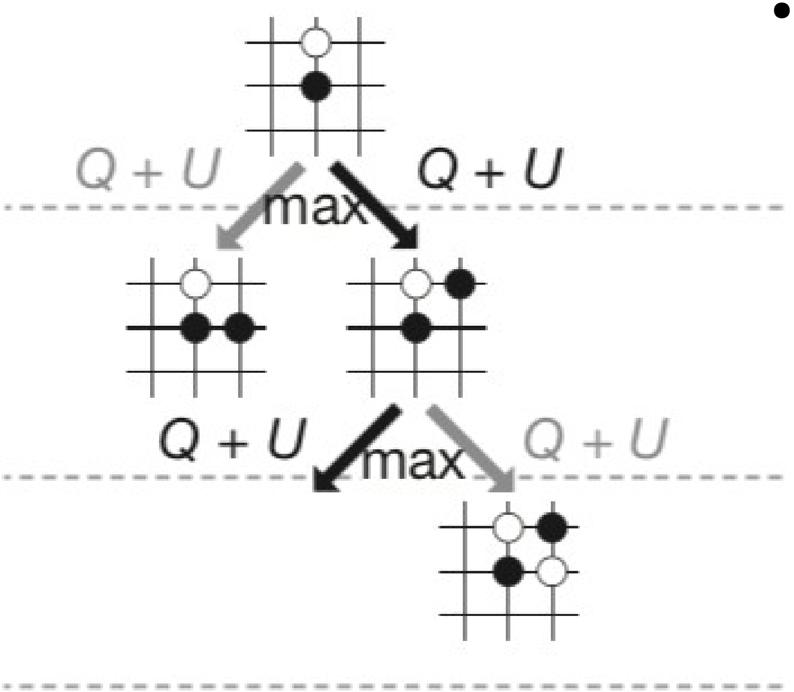


Modified MCTS



Each edge stores: $\{ \underbrace{N(s,a)}_{\text{visit count}}, \underbrace{W(s,a)}_{\text{total action value}}, \underbrace{Q(s,a)}_{\text{mean action value}}, \underbrace{P(s,a)}_{\text{prior probability}} \}$

Select



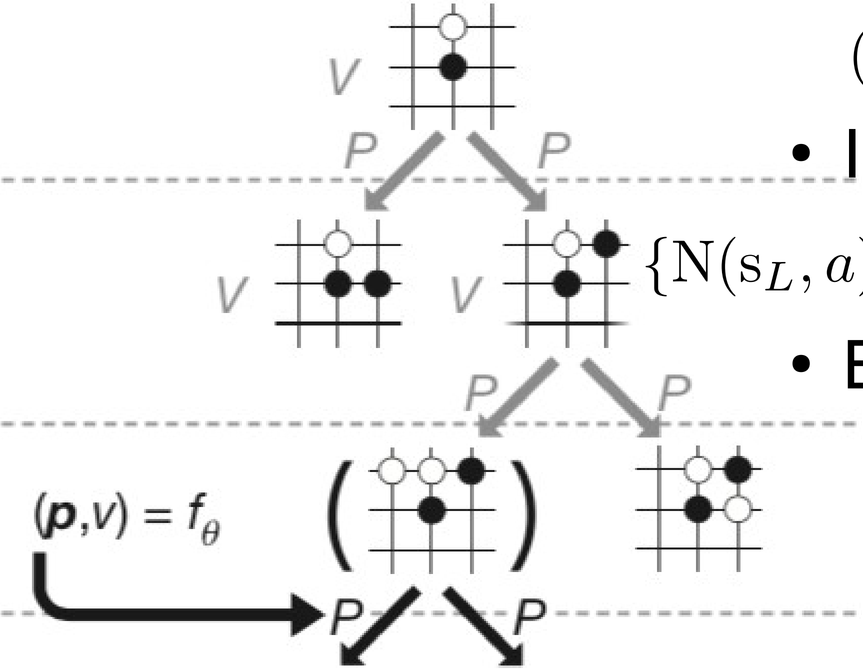
- Select action according to PUCT

$$\mathrm{U(s,a)} = \mathrm{c}_{puct} P(s,a) \frac{\sqrt{\sum_b N(s,b)}}{1+N(s,a)}$$

$$a_t = \underset{a}{\operatorname{argmax}}(Q(s_t,a) + U(s_t,a))$$

$$c_{puct} : \text{level of exploaration}$$

Expand



- Evaluate NN at leaf node:

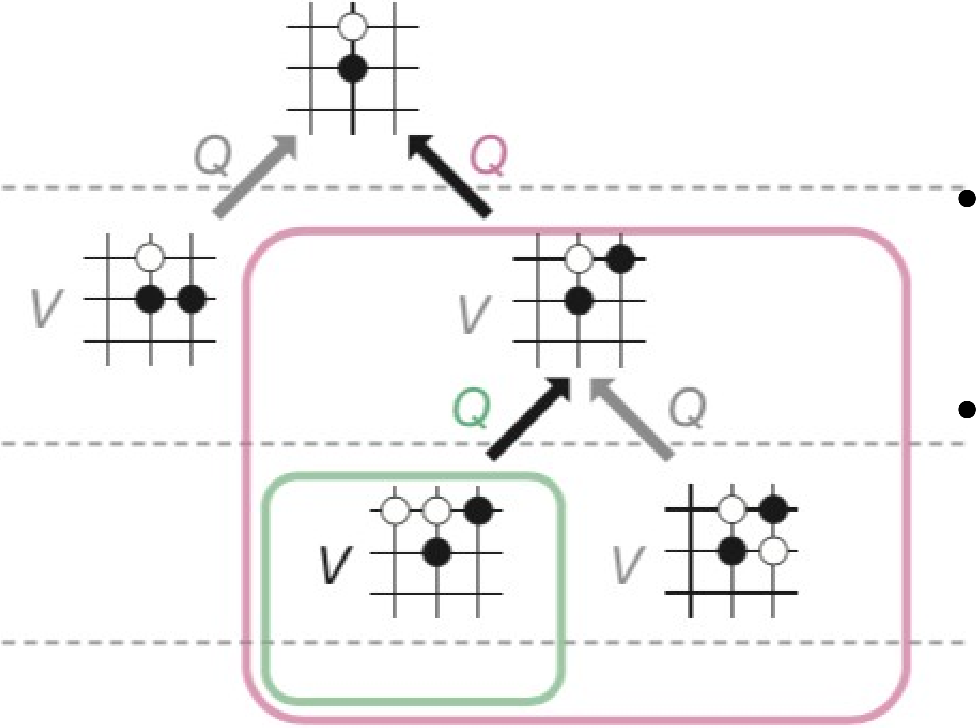
$$(d(\mathbf{p}), v) = f_\theta(d(s_L))$$

- Insert new edge:

$$\{\mathrm{N}(\mathrm{s}_L, a) = 0, W(s_L, a) = 0, Q(s_L, a) = 0, P(s_L, a) = p_a\}$$

- Backup value

Backup



- Increment visit counts

$$N(s_t, a_t) = N(s_t, a_t) + 1$$

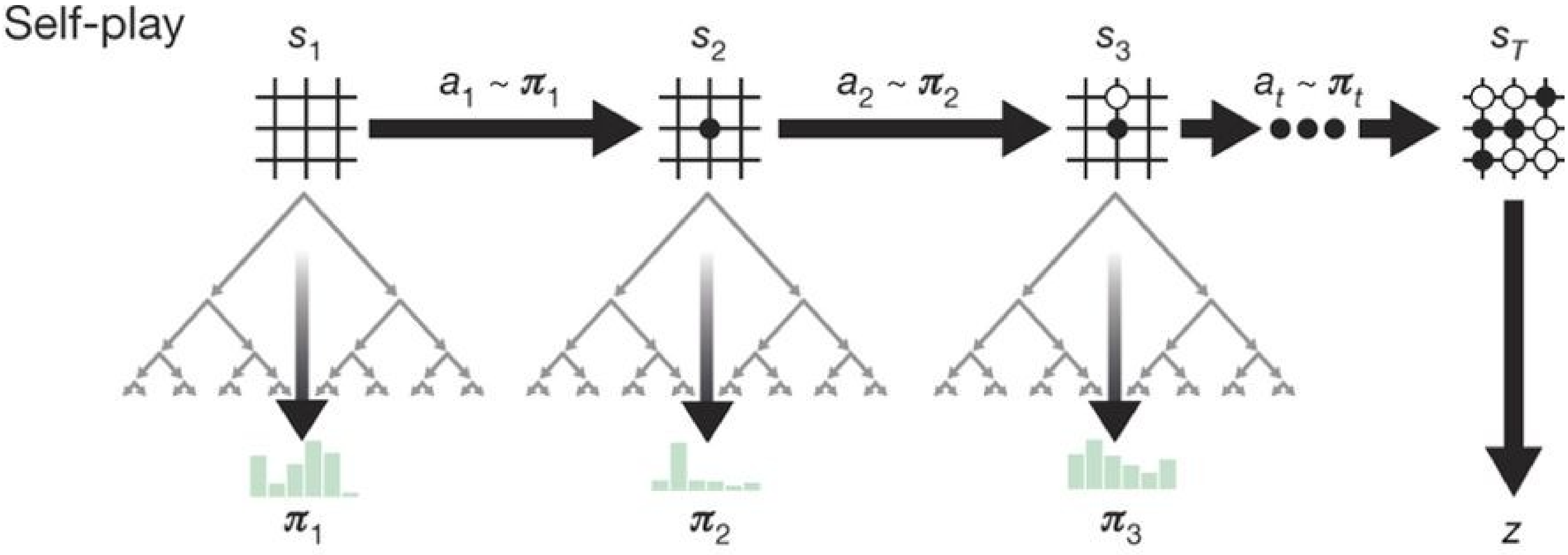
- Add value to action value

$$W(s_t, a_t) = W(s_t, a_t) + v$$

- Update Mean action value

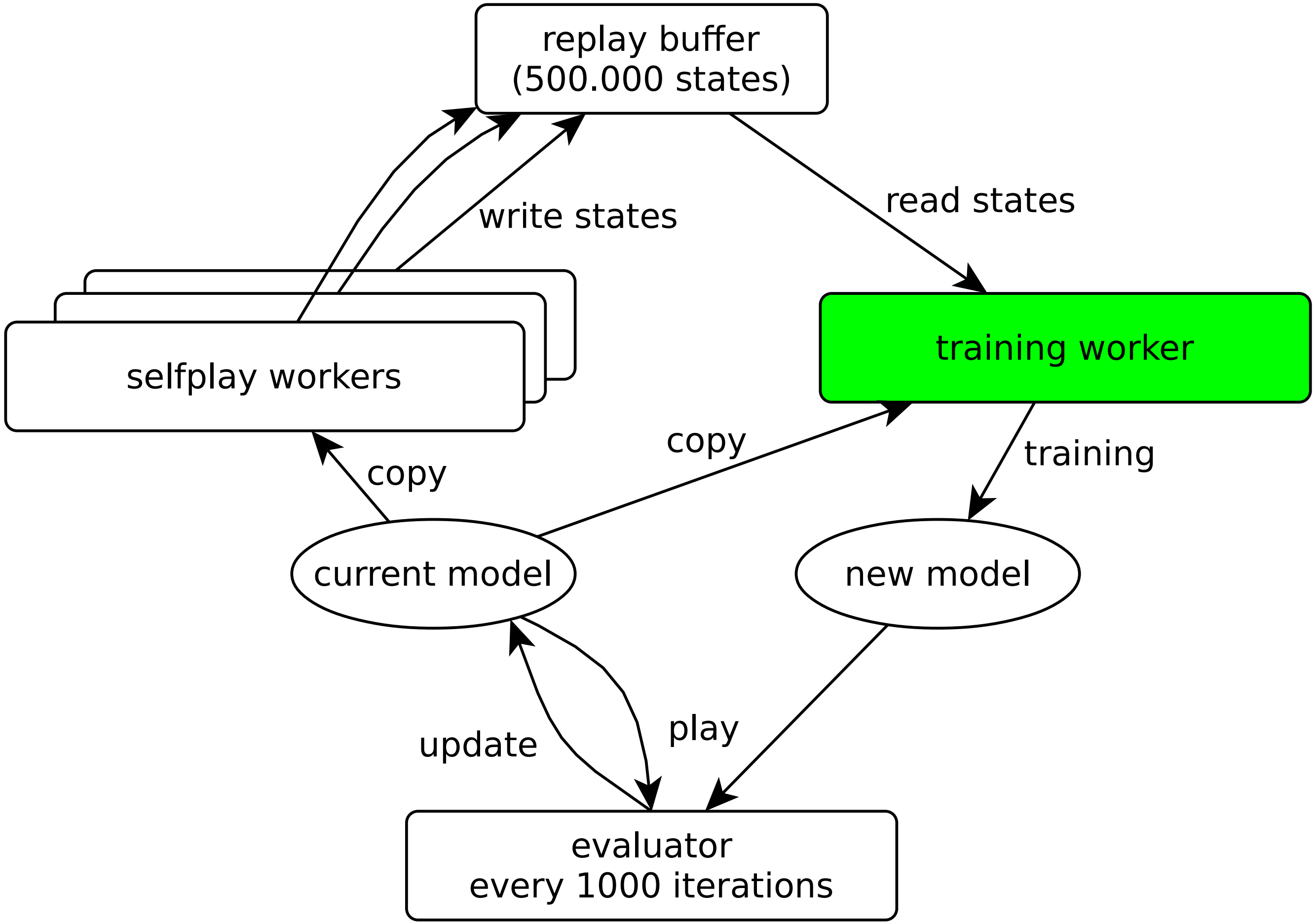
$$Q(s_t, a_t) = \frac{W(s_t, a_t)}{N(s_t, a_t)}$$

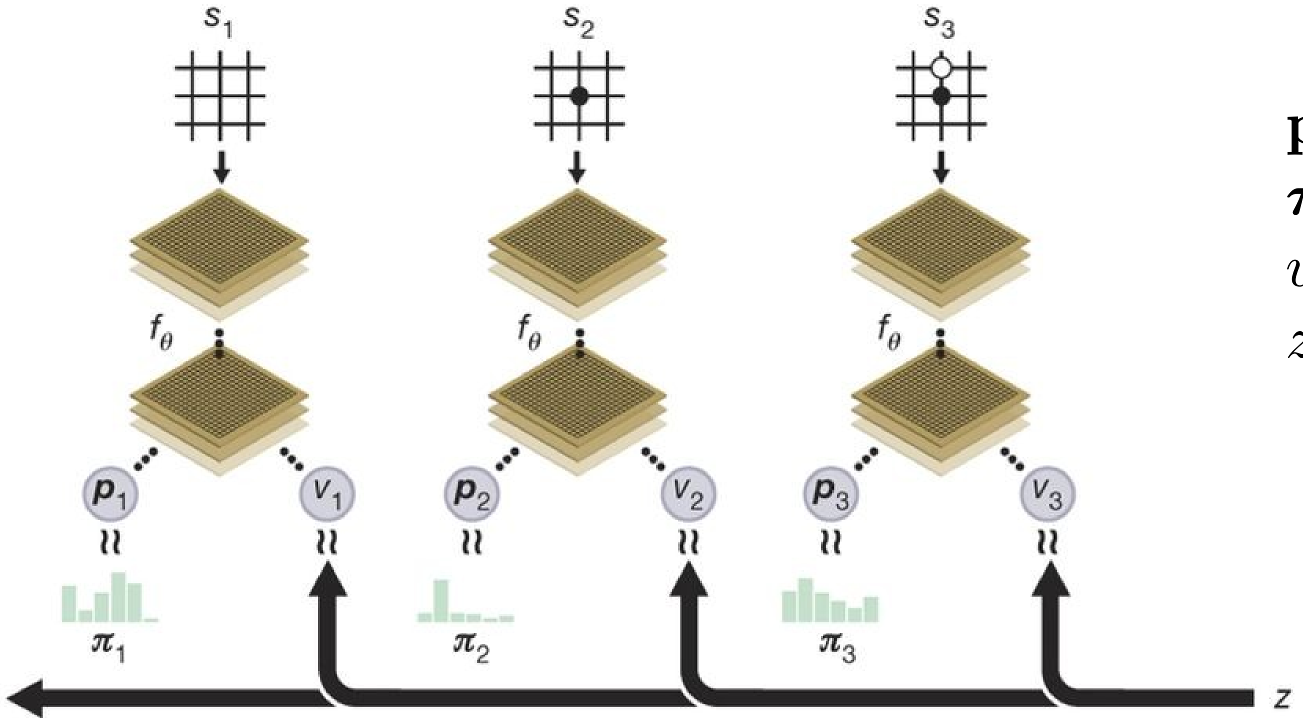
Play



$$\pi(a|s_t) = \frac{N(s_t, a)^{1/\tau}}{\sum_b N(s_t, b)^{1/\tau}}$$

Policy iteration





$(\mathrm{p},\mathrm{v}) = \mathrm{f}_\theta(s)$

$\mathbf{p}$: policy

$\boldsymbol{\pi}$: MCTS probabilities

v : value

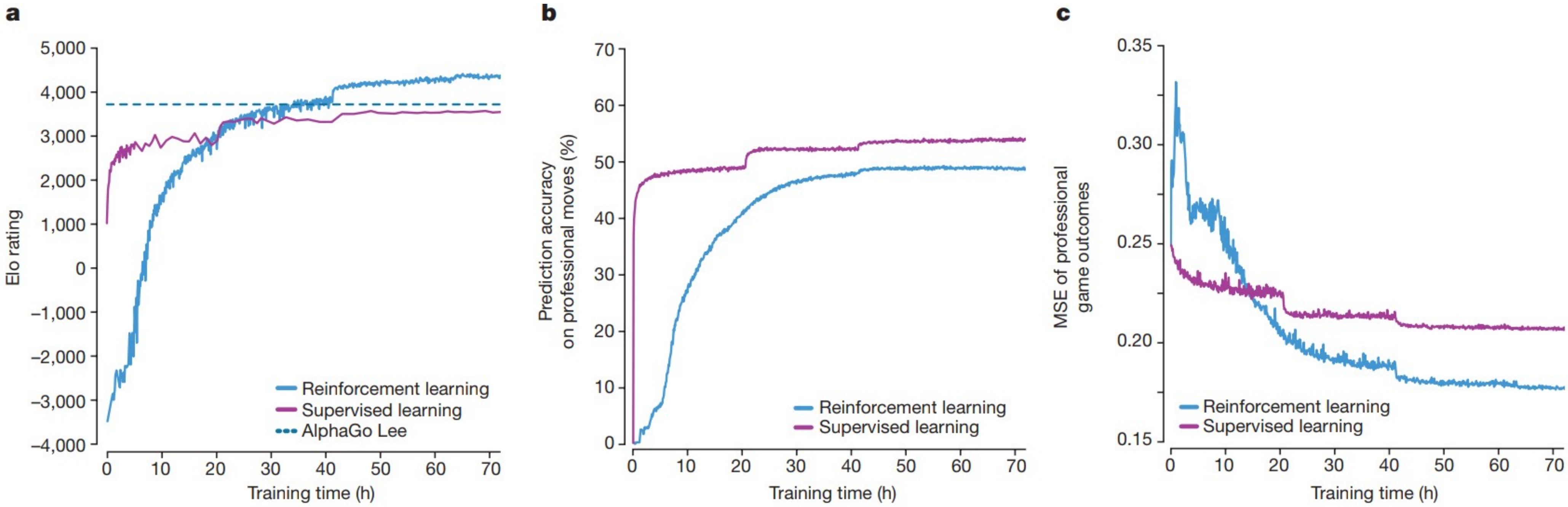
z : outcome of game

$$\underbrace{l}_{\text{loss-function}} = \underbrace{(z-v)^2}_{\ell^2-\text{loss}} - \underbrace{\boldsymbol{\pi}^T \log \mathbf{p}}_{\text{cross-entropy}} + \underbrace{c||\theta||^2}_{\ell^2-\text{weight normalization}}$$

Why is it better?

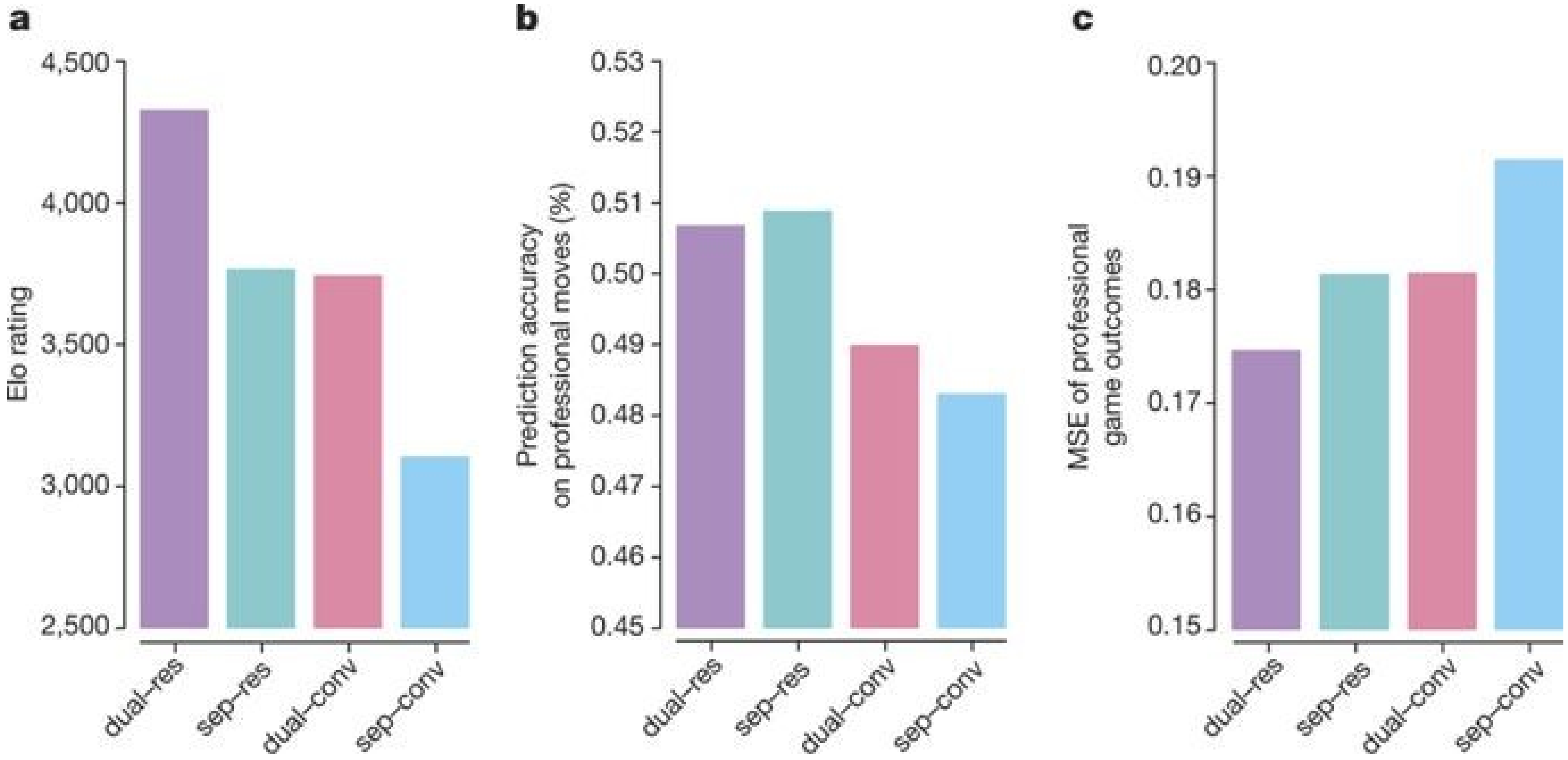
- MCTS search in training loop provides stable gradient for training
  - Augmented policy is always better at predicting the best move

Supervised vs Reinforcement learning



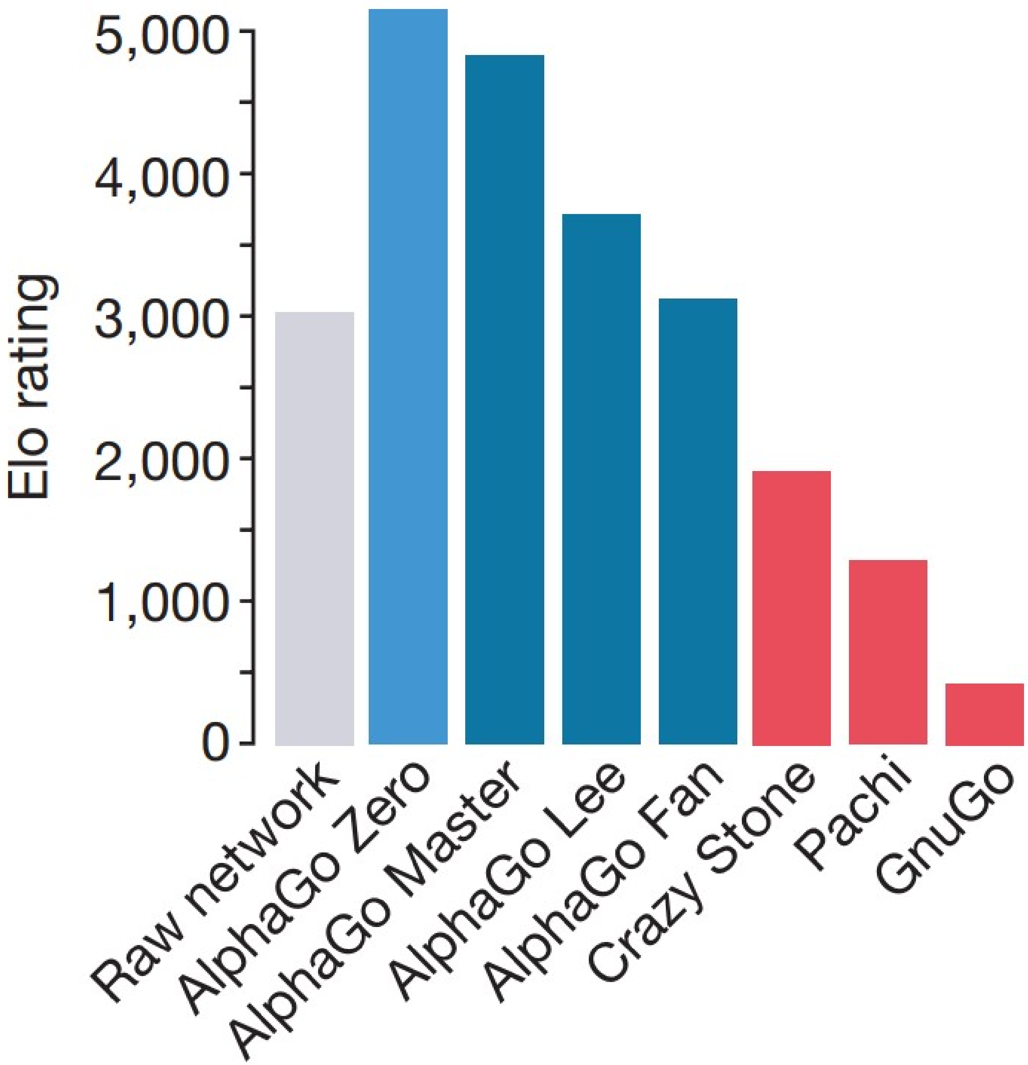
Why is it better?

- MCTS search in training loop provides stable gradient for training
  - Augmented policy is always better at predicting the best move
- ResNets instead of ConvNets
  - Ability to train even deeper models
- Same network for Policy and Value
  - Multi-task learning with hard parameter sharing regularizes training and prevents overfitting



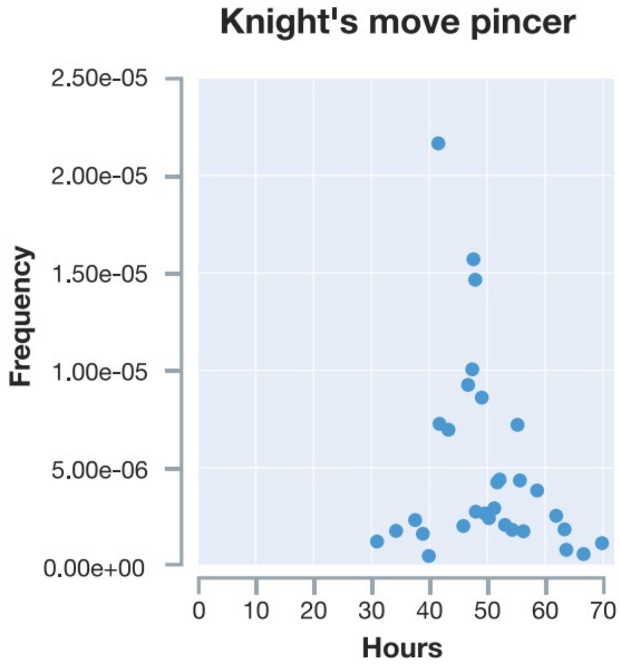
a
4,500
4,000
3,500
3,000
2,500
Elo rating
dual–res
sep–res
dual–conv
sep–conv
b
0.53
0.52
0.51
0.50
0.49
0.48
0.47
0.46
0.45
Prediction accuracy
on professional moves (%)
dual–res
sep–res
dual–conv
sep–conv
c
0.20
0.19
0.18
0.17
0.16
0.15
MSE of professional
game outcomes
dual–res
sep–res
dual–conv
sep–conv

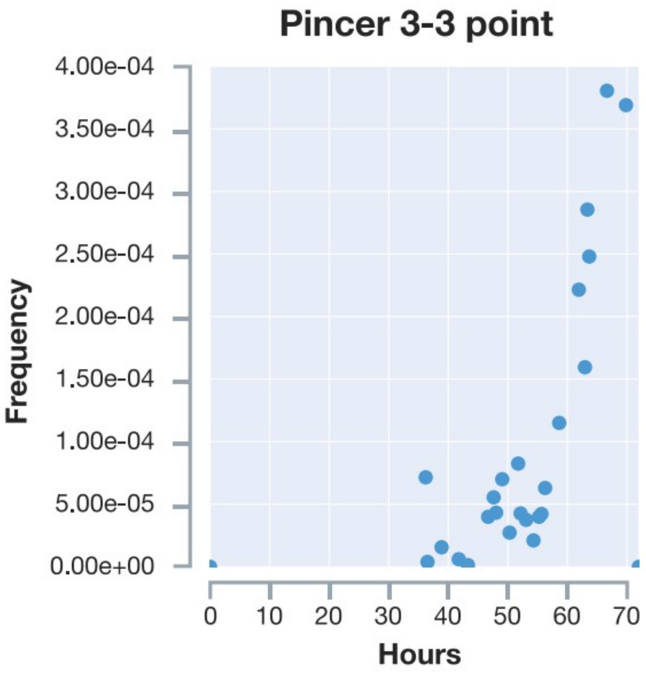
- 2012: Crazy Stone
  - MCTS search with handcrafted heuristic
  - Professional level play
- 2015: AlphaGo Fan
  - MCTS with Value and Policy Network
  - Defeated European Go champion Fan Hui (2-dan)
- 2016: AlphaGo Lee
  - Larger networks than AlphaGo Fan
  - Added selfplay of policy network
  - Won 3/1 against Lee Sedol (9-dan)
- 2017: AlphaGo Master
  - Use single network for both policy and value
  - Using ResNet instead of Convolutional NN
  - Won 60/0 against team of proffessional players
- 2018 AlphaGo Zero
  - Trained from scratch
- 2018 AlphaZero
  - Generalized to Chess & Shogi



Comparison to human play

- Superhuman performance
- Learned to play human Joseki





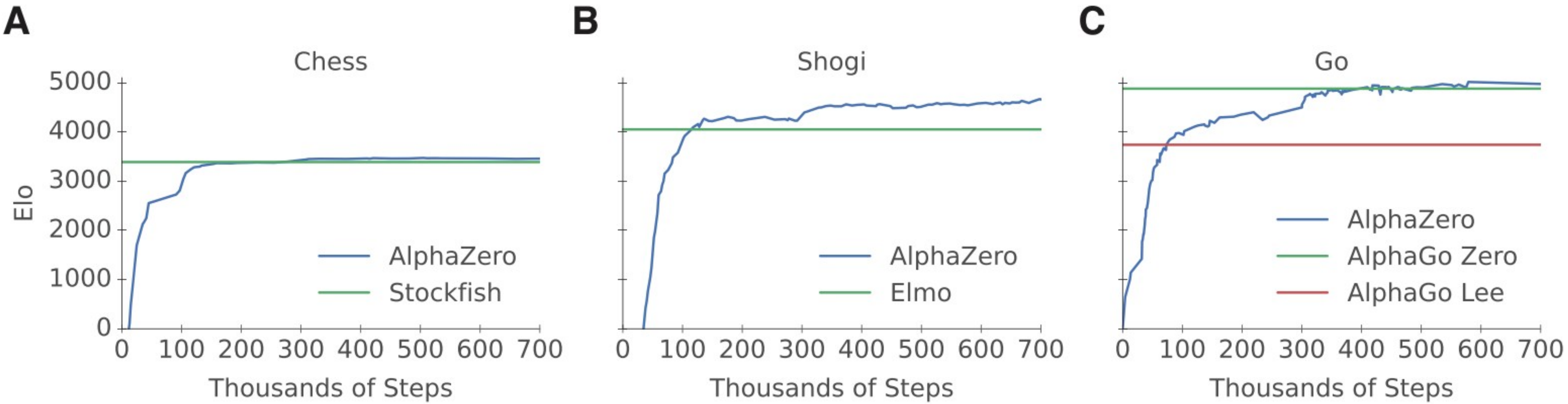
AlphaGo Zero vs AlphaZero

- Absence of human knowledge made transfer to Shogi and Chess very easy
- No change to NN architecture
- Only raw board states as input
- No evaluator → Continuous update of NN

Go vs Chess/Shogi

Rules	Go	Chess/Shogi
Translation invariance	Yes	Partial
Locality	Yes	No
Symmetry	Yes	No
Action space	Simple	Compound
Game outcomes	Probability of winning	Win / Loss / Draw

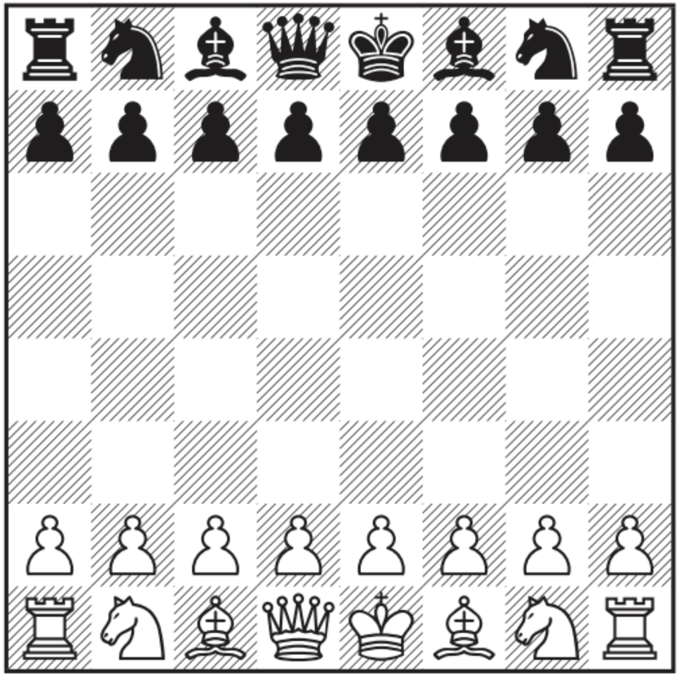
Performance of AlphaZero

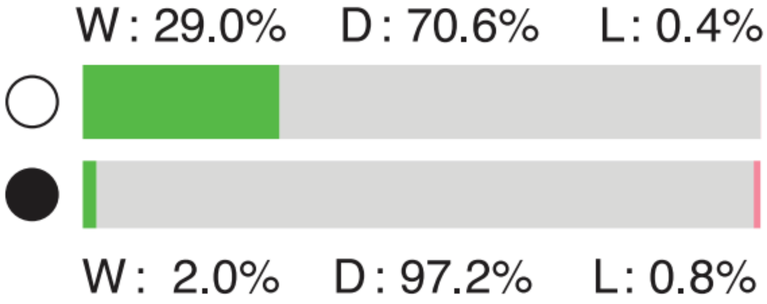


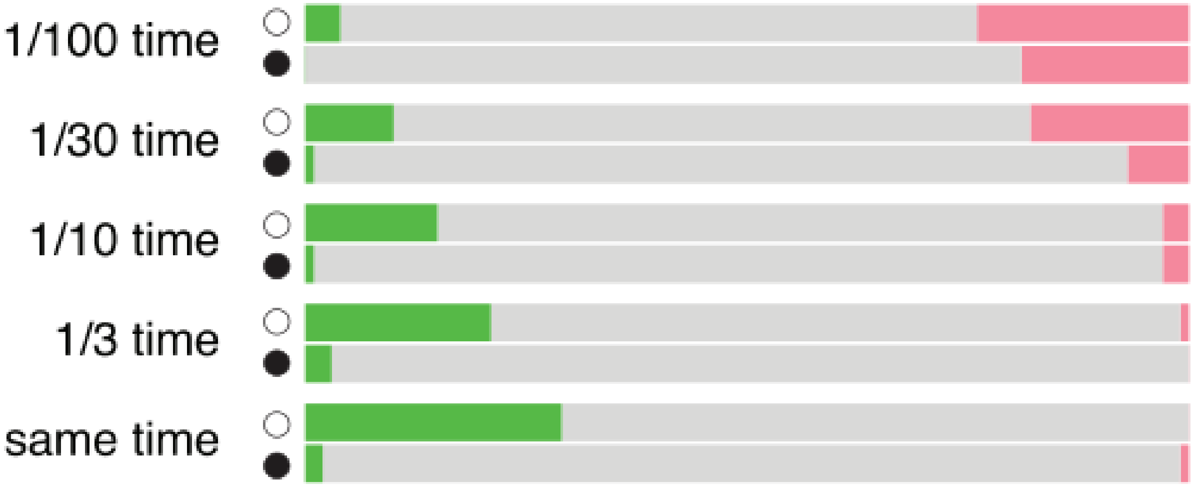
AlphaZero vs Stockfish

- Stockfish: Alpha Beta Pruning with handcrafted heuristics, endplay-tables, opening book, etc…
- Stockfish 60’000’000 Moves/Second
- AlphaZero 60’000 Moves/Second

AlphaZero vs. Stockfish







Conclusion

- Playing smart is better than brute-force
- Generality is better than handcrafting features
- Not injecting human knowledge promotes generality
- Multitask learning prevents overfitting

References

- Silver, David, et al. "Mastering the game of go without human knowledge." Nature 550.7676 (2017): 354.
- Silver, David, et al. "Mastering chess and shogi by self-play with a general reinforcement learning algorithm." arXiv preprint arXiv:1712.01815 (2017).
- Silver, David, et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science 362.6419 (2018): 1140-1144.
- AlphaGo Zero Cheat Sheet: https://medium.com/applied-data-science/alphago-zero-explained-in-one-diagram-365f5abf67e0